The Electromagnetic Radiation Spectrum

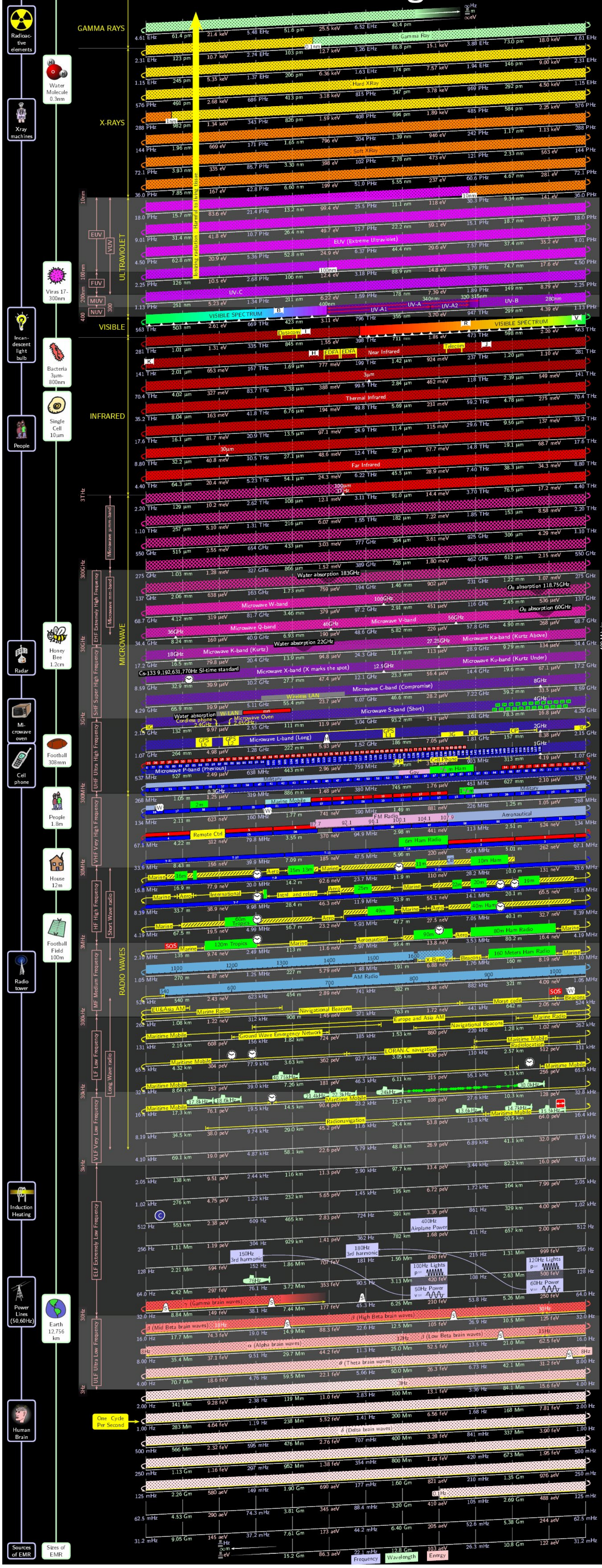

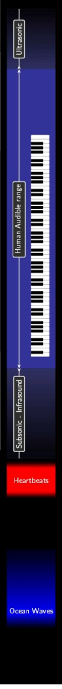

How to read this chart

- This chart is organized in octaves (frequency doubling/halving) starting at 1Hz and going higher (2,4,8, etc) and lower (1/2, 1/4, etc). The octave is a natural way to represent frequency.
- Frequency increases on the vertical scale in the upward direction.
- The horizontal bars wrap around from far right to far left as the frequency increases upwards.
- There is no limit to either end of this chart, however, due to limited space, only the "known" items have been shown here. A frequency of 0Hz is the lowest possible frequency but the method of depicting octaves used here does not allow for ever reaching 0Hz, only approaching it. Also, by the definition of frequency (Cycles per second), there is no such thing as negative frequency.
- Values on the chart have been labelled with the following colours: Frequency measured in Hertz, Wavelength measured in meters, Energy measured in electronVolts.

Ultraviolet Light

- Ultraviolet light is beyond the range of human vision.
- Physicists have divided ultraviolet light ranges into Vacuum Ultraviolet (VUV), Extreme Ultraviolet (EUV), Far Ultraviolet (FUV), Medium Ultraviolet (MUV), and Near Ultraviolet (NUV).
- UV-A, UV-B and UV-C were introduced in the 1930's by the Commission Internationale de l'Éclairage (CIE, International Commission on Illumination) for photobiological spectral bands.
- Short-term UV-A exposure causes sun-tanning which helps to protect against sunburn. Exposure to UV-B is beneficial to humans by helping the skin produce vitamin D. Excessive UV exposure causes skin damage. UV-C is harmful to humans but is used as a germicide.
- The CIE originally divided UVA and UVB at 315nm, later some photo-dermatologists divided it at 320nm.
- UVA is subdivided into UVA1 and UVA2 for DNA altering effects at 340nm.
- The sun produces a wide range of frequencies including all the ultraviolet light, however, UVB is partially filtered by the ozone layer and UVC is totally filtered out by the earth's atmosphere.
- A bumblebee can see light in the UVA range which helps them identify certain flowers.

Emission and Absorption

- As EMR passes through elements, certain wavelength bands get absorbed and some new ones get emitted. This absorption and emission produces characteristic spectral lines for each element which are useful in determining the makeup of distant stars. These lines are used to prove the red-shift amount of distant stars.
- When a photon hits an atom it may be absorbed if the energy is just right. The energy level of the electron is raised – essentially holding the radiation. A new photon of specific wavelength is created when the energy is released. The jump in energy is a discrete step and many possible levels of energy exist in an atom.
- Johann Balmer created this formula defining the photon emission wavelength (λ); where m is the initial electron energy level and n is the final electron energy level:

$$\lambda = 364.56nm\left(\frac{m^2}{m^2 - n^2}\right)$$

- Much of the interstellar matter is made of the simplest atom hydrogen. The hydrogen visible-spectrum emission and absorption lines are shown below:

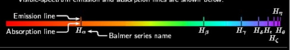

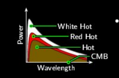

- Max Planck determined the relationship between the temperature of an object and its radiation profile; where R_λ is the radiation power, λ is the wavelength, T is the temperature:

$$R_\lambda = \frac{37418}{\lambda^5 \left(e^{\frac{14388}{\lambda T}} - 1\right)}$$

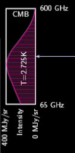

Cosmic Microwave Background Radiation

- CMB radiation is the leftover heat from the hot early universe, which last scattered about 400,000 years after the Big Bang.
- CMB permeates the entire universe at a temperature of 2.725 ± 0.001K.
- CMB was predicted in the 1940's by Ralph Alpher, George Gamow and Robert Herman.
- Arno Penzias and Robert Wilson accidentally discovered CMB while working for Bell Telephone Laboratories in 1965.
- The intensity is measured in Mega Jansky (Jy) per steradian. $1Jy = 10^{-26}W/m^2/Hz$

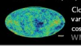

Close examination of slight CMB intensity variations in different parts of the sky help cosmologists study the formation of galaxies.

WMAP photo by NASA

Television

- Television is transmitted in the VHF and UHF ranges (30MHz - 3GHz).
- TV channels transmitted over the air are shown as TV.
- TV channels transmitted through cable (CATV) are shown as TV. CATV channels starting with "T-" are channels fed back to the cable TV station (like news feeds).
- Air and cable TV stations are broadcast with the separate video, colour, and audio frequency carriers grouped together in a channel band as follows:

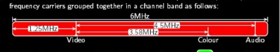

- Satellite channels broadcast in the C-Band are depicted as TV. These stations are broadcast in alternating polarities (Ex. Ch 1 is vertical and 2 is horizontal and vice versa on neighbouring satellites).
- The 15.7 kHz horizontal sweep signal produced by a TV can be heard by some young people. This common contaminant signal to VLF spectra listening is depicted as ↔.

Radio Bands

- The radio spectrum (ELF to EHF) is populated by many more items than can be shown on this chart, only a small sampling of bands used around the world have been shown.
- Communication using EMR is done using either:
  - Amplitude Modulation (AM)

    OR
  - Frequency Modulation (FM)

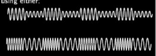

- Each country has its own rules and regulations for allotting bands in this region. For more information, look up the radio communications authority in your area (Ex. FCC in the US, DOC in Canada).
- Not all references agree on the ULF band range, the HAARP range is used here.
- **RA**dio **D**etecting **A**nd **R**anging (RADAR) uses EMR in the microwave range to detect the distance and speed of objects.
- **C**itizens **B**and Radio (CB) contains 40 stations between 26.965-27.405MHz.
- Schumann resonance is produced in the cavity between the Earth and the ionosphere. The resonant peaks are depicted as S
- Hydrogen gas emits radio band EMR at 21cm H
- Some individual frequencies are represented as icons:

xxHz	Submarine communications
🕒	Time and frequency standards
xxm	Ham radio and international meter bands
▨	Miscellaneous short wave radio
W	Weather stations
CP	Cellular and PCS Phones (including; FDMA, TDMA, CDMA ranges)

Sound

- Although sound, ocean waves, and heartbeats are not electromagnetic, they are included on this chart as a frequency reference. Other properties of electromagnetic waves are different from sound waves.
- Sound waves are caused by an oscillating compression of molecules. Sound cannot travel in a vacuum such as outer space.
- The speed of sound in air is 1240kph (770mph).
- Humans can only hear sound between ≈20Hz to ≈20kHz.
- Infrasound (below 20Hz) can be sensed by internal organs and touch. Frequencies in the 0.2Hz range are often the cause of motion sickness.
- Bats can hear sound up to ≈50kHz.
- The 88 piano keys of the Equal Temperament scale are accurately located on the frequency chart.
- Over the ages people have striven to divide the continuous audio frequency spectrum into individual musical notes that have harmonious relationships. Microtonal musicians study various scales. One recent count lists 4700 different musical scales.
- Middle C is depicted on the chart as C

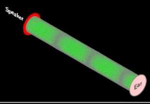

This image depicts air being compressed as sound waves in a tube from a speaker and then travelling through the tube towards the ear.

Gravity Waves

- Gravity is the mysterious force that holds large objects together and binds our planets, stars and galaxies together. Many people have unsuccessfully theorized about the details of gravity and its relationship to other forces. There have been no links between gravity waves and electromagnetic radiation.
- Gravity is theorized to warp space and time. In fact, gravity is responsible for bending light as observed by the gravity-lens example of distant galaxies.
- "Gravity waves" would appear as ripples in space-time formed by large objects moving through space that might possibly be detected in the future by very sensitive instruments.
- The speed that gravity propagates through space has been theorized to be the same as the speed of light.

Brain Waves

- By connecting electrodes from the human head to an electroencephalograph (EEG), it is possible to measure very small cyclic electrical signals.
- There has been much study on this topic, but like all effects on humans, the science is not as exact as the science of materials.
- Generally, lower brain wave frequencies relate to sleep, and the higher frequencies relate to alertness.
- Devices have been made for measuring and stimulating brain waves to achieve a desired state.

© unihedron.com 2004-12-12

Electromagnetic Radiation (EMR)

- EMR is emitted in discrete units called photons but has properties of waves as seen by the images below. EMR can be created by the oscillation or acceleration of electrical charge or magnetic field. EMR travels through space at the speed of light (2.997 924 58 $\times 10^8$ m/s). EMR consists of an oscillating electrical and magnetic field which are at right angles to eachother and spaced at a particular wavelength. There is some controversy about the phase relationship between the electrical and magnetic fields of EMR, one of the theoretical representations is shown here:

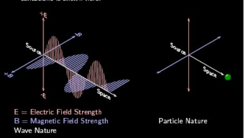

- The particle nature of EMR is exhibited when a solar cell emits individual electrons when struck with very dim light.
- The wave nature of EMR is demonstrated by the famous double slit experiment that shows cancelling and addition of waves.
- Much of the EMR properties are based on theories since we can only see the effects of EMR and not the actual photon or wave itself.
- Albert Einstein theorized that the speed of light is the fastest that anything can travel. So far he has not been proven wrong.
- EMR can have its wavelength changed if the source is receding or approaching as in the red-shift example of distant galaxies and stars that are moving away from us at very high speeds. The emitted spectral light from these receding bodies appears more red than it would be if the object was not moving away from us.
- We only have full electronic control over frequencies in the microwave range and lower. Higher frequencies must be created by waiting for the energy to be released from elements as photons. We can either pump energy into the elements (ex. heating a rock with visible EMR and letting it release infrared EMR) or let it naturally escape (ex. uranium decay).
- We can only see the visible spectrum. All other bands of the spectrum are depicted as hatched colours ▨.

Système International d'unité prefixes (SI unit prefixes)

Symbol	Name	Exp.	Multiplier
Y	yotta	10^{24}	1,000,000,000,000,000,000,000,000
Z	zetta	10^{21}	1,000,000,000,000,000,000,000
E	exa	10^{18}	1,000,000,000,000,000,000
P	peta	10^{15}	1,000,000,000,000,000
T	tera	10^{12}	1,000,000,000,000
G	giga	10^{9}	1,000,000,000
M	mega	10^{6}	1,000,000
k	kilo	10^{3}	1,000
		10^{0}	1
m	milli	10^{-3}	0.001
μ	micro	10^{-6}	0.000 001
n	nano	10^{-9}	0.000 000 001
p	pico	10^{-12}	0.000 000 000 001
f	femto	10^{-15}	0.000 000 000 000 001
a	atto	10^{-18}	0.000 000 000 000 000 001
z	zepto	10^{-21}	0.000 000 000 000 000 000 001
y	yocto	10^{-24}	0.000 000 000 000 000 000 000 001

Measurements on this chart

Symbol	Name	Value
c	Speed of Light	2.997 924 58 $\times 10^{8}$ m/s
h	Planck's Constant	6.626 1 $\times 10^{-34}$ J·s
$\hbar$	Planck's Constant (freq)	1.054 592 $\times 10^{-34}$ J·s
f	Frequency (cycles / second)	Hz
λ	Wavelength (meters)	m
E	Energy (Joules)	J

Conversions

$E = h \cdot f$

$\lambda = \frac{c}{f}$

1Å = 0.1nm

1nm = 10Å

1Joule = 6.24 $\times 10^{18}$ eV

Gamma Rays

- Gamma radiation is the highest energy radiation (up to ≈ 10^{20} eV) that has been measured. At this energy, the radiation could be from gamma-rays, protons, electrons, or something else.
- Alpha, beta, and delta radiation are not electromagnetic but are actually parts of the atom being released from a radioactive atom. In some cases this can cause gamma radiation. These are not to be confused with brain waves of similar names.

Visible Spectrum

- The range of EMR visible to humans is also called "Light". The visible spectrum also closely resembles the range of EMR that filters through our atmosphere from the sun.
- Other creatures see different ranges of visible light, for example bumble-bees can see ultraviolet light and dogs have a different response to colours than do humans.
- The sky is blue because our atmosphere scatters light and the shorter wavelength blue gets scattered the most. It appears that the entire sky is illuminated by a blue light but in fact that light is scattered from the sun. The longer wavelengths like red and orange move straight through the atmosphere which makes the sun look like a bright white ball containing all the colours of the visible spectrum.
- Interestingly, the visible spectrum covers approximately one octave.
- Astronomers use filters to capture specific wavelengths and reject unwanted wavelengths, the major astronomical (visual) filter bands are depicted as X

Infrared Radiation

- Infrared radiation (IR) is sensed by humans as heat and is below the range of human vision. Humans (and anything at room temperature) are emitters of IR.
- IR remote control signals are invisible to the human eye but can be detected by most camcorders.
- Night vision scopes/goggles use a special camera that senses IR and converts the image to visible light. Some IR cameras employ an IR lamp to help illuminate the view.
- IR LASERs are used for burning objects.
- A demonstration of IR is to hold a metal bowl in front of your face. The IR emitted by your body will be reflected back using the parabolic shape of the bowl and you will feel the heat.
- Fiber-optic based infrared communication signals are sometimes amplified with Erbium-Doped Fiber Amplifiers EDFA

LASER

- LASER is an acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation.
- A LASER is a device that produces monochromatic EMR of high intensity.
- With proper equipment, any EMR can be made to operate like a LASER. For example, microwaves are used to create a MASER.

Polarization

- As a photon (light particle) travels through space, its axis of electrical and magnetic fluctuations does not rotate. Therefore, each photon has a fixed linear polarity of somewhere between 0° to 360°. Light can also be circularly and elliptically polarized.
- Some crystals can cause the photon to rotate its polarization.
- Receivers that expect polarized photons will not accept photons that are in other polarities. (ex. satellite dish receivers have horizontal and vertical polarity positions).
- A polarized filter (like Polaroid™ sunglasses) can be used to demonstrate polarized light. One filter will only let photons that have one polarity through. Two overlapping filters at right angles will almost totally block the light that exits, however, a third filter inserted between the first two at a 45° angle will rotate the polarized light and allow some light to come out the end of all three filters.
- Light that reflects off an electrical insulator becomes polarized. Conductive reflectors do not polarize light.
- Perhaps the most reliably polarized light is a rainbow.
- Moonlight is also slightly polarized. You can test this by viewing the moonlight through a Polaroid™ sunglass lens, then rotate that lens, the moonlight will dim and brighten slightly.

Refraction

- Refraction of EMR is dependent on wavelength as can be seen by the prism example below.

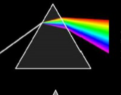

By using a glass prism, white light can be spread by refraction into a spectrum of its composite colours. All wavelengths of EMR can be refracted by using the proper materials.

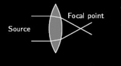

Convex lenses make objects appear closer and are used to correct far-sitedness.

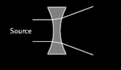

Concave lenses make objects appear farther away and are used to correct near-sitedness.

Photo by STScI

Heavy objects like dense galaxies and large planets cause light to bend due to gravitational lensing.

Reflection

- Reflection of EMR is dependent on wavelength as demonstrated when visible light and radio waves bounce off objects that X-Rays would pass through. Microwaves, which have a large wavelength compared to visible light, will bounce off metal mesh in a microwave oven whereas visible light will pass through.

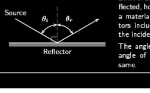

EMR of any wavelength can be reflected, however, the reflectivity of a material depends on many factors including the wavelength of the incident beam.

The angle of incidence (θ_i) and angle of reflection (θ_r) are the same.

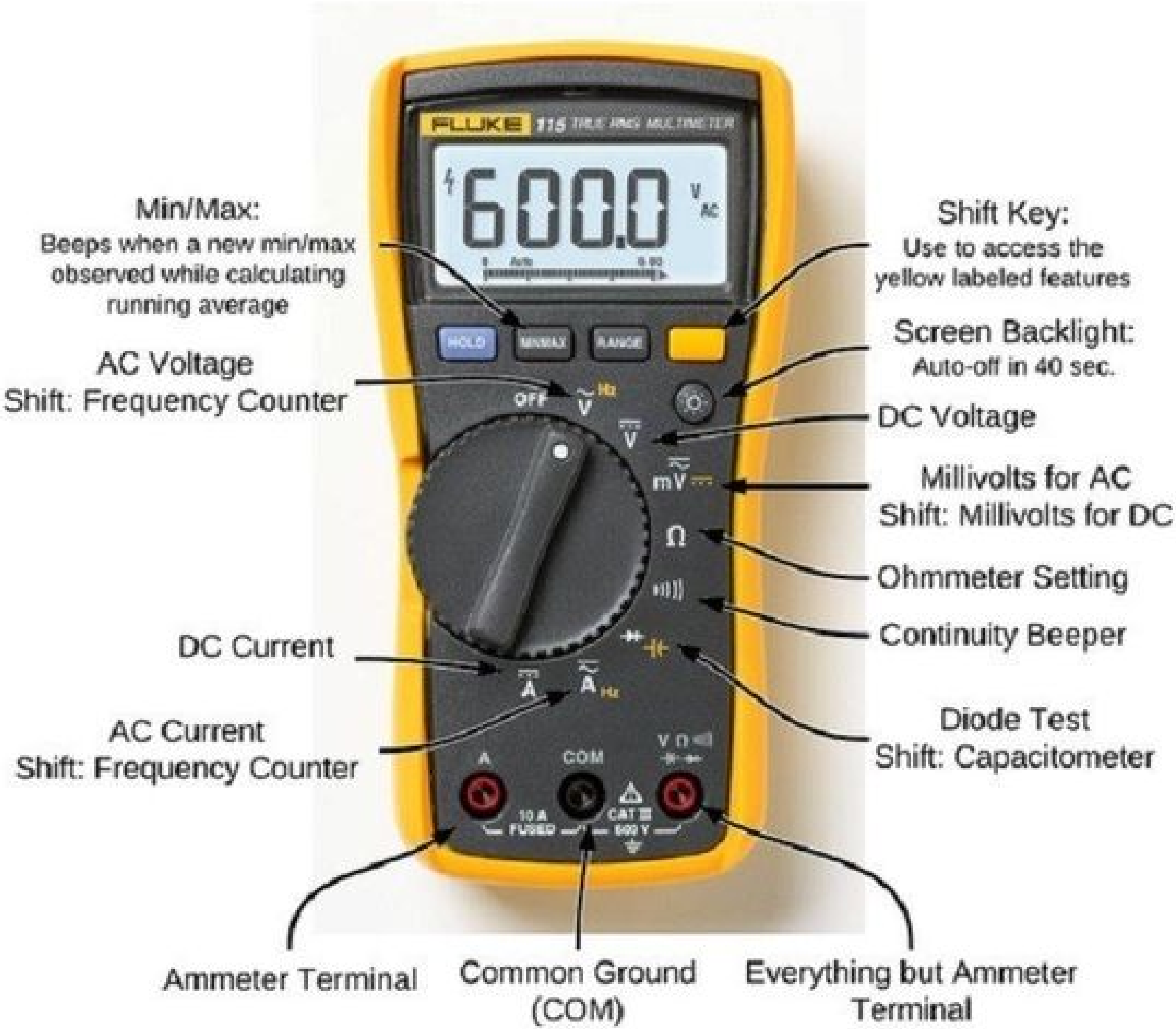

FLUKE
115 TRUE RMS MULTIMETER
600.0
V AC
Auto
HOLD
MIN MAX
RANGE
OFF
V Hz
V
mV
Ω
A
A Hz
A
COM
V Ω
10 A FUSED
CAT III 600 V

Min/Max:
Beeps when a new min/max observed while calculating running average

Shift Key:
Use to access the yellow labeled features

Screen Backlight:
Auto-off in 40 sec.

AC Voltage
Shift: Frequency Counter

DC Voltage

Millivolts for AC
Shift: Millivolts for DC

Ohmmeter Setting

Continuity Beeper

DC Current

AC Current
Shift: Frequency Counter

Diode Test
Shift: Capacitometer

Ammeter Terminal

Common Ground
(COM)

Everything but Ammeter
Terminal

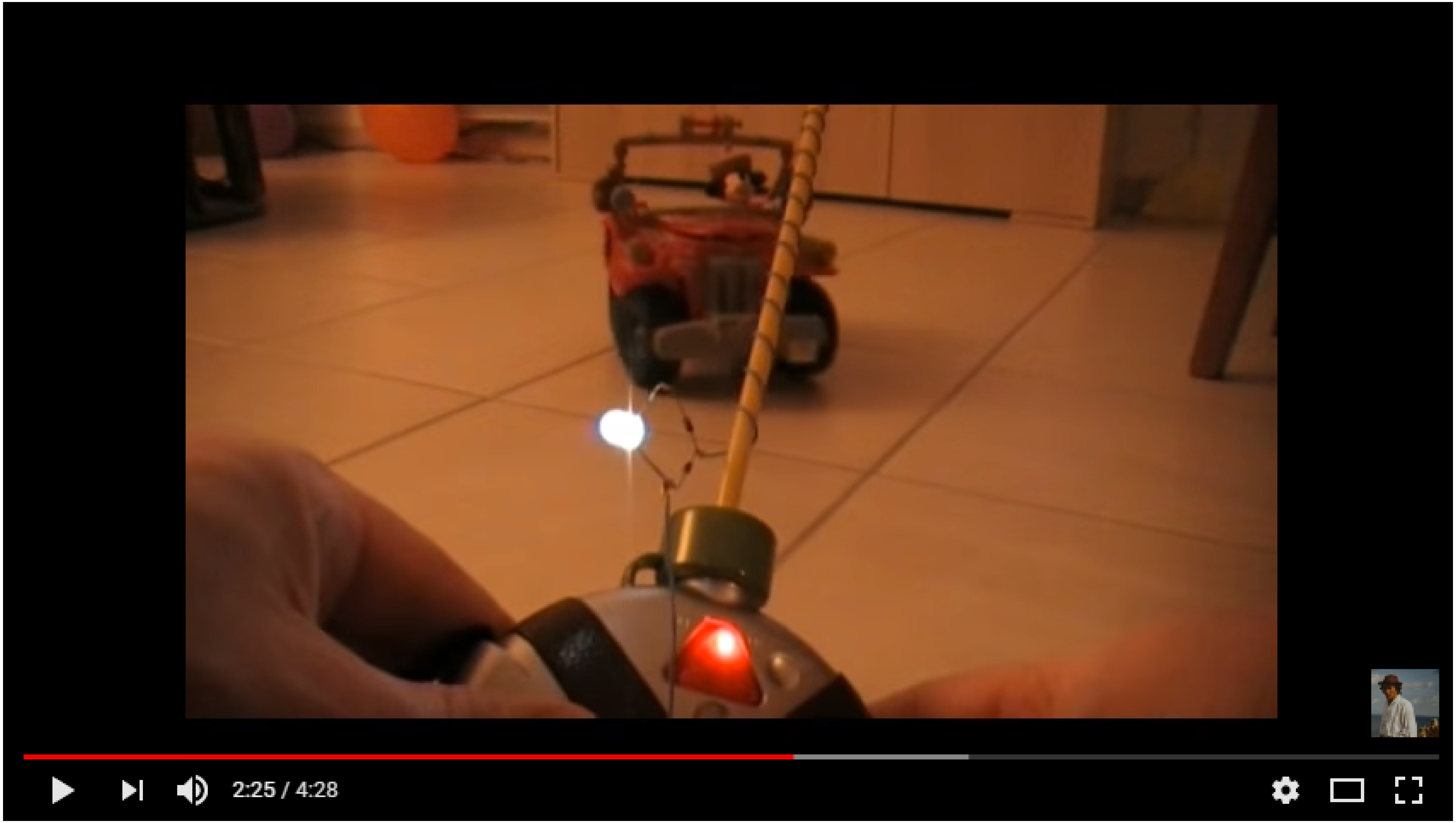

electricity from radiowaves 3

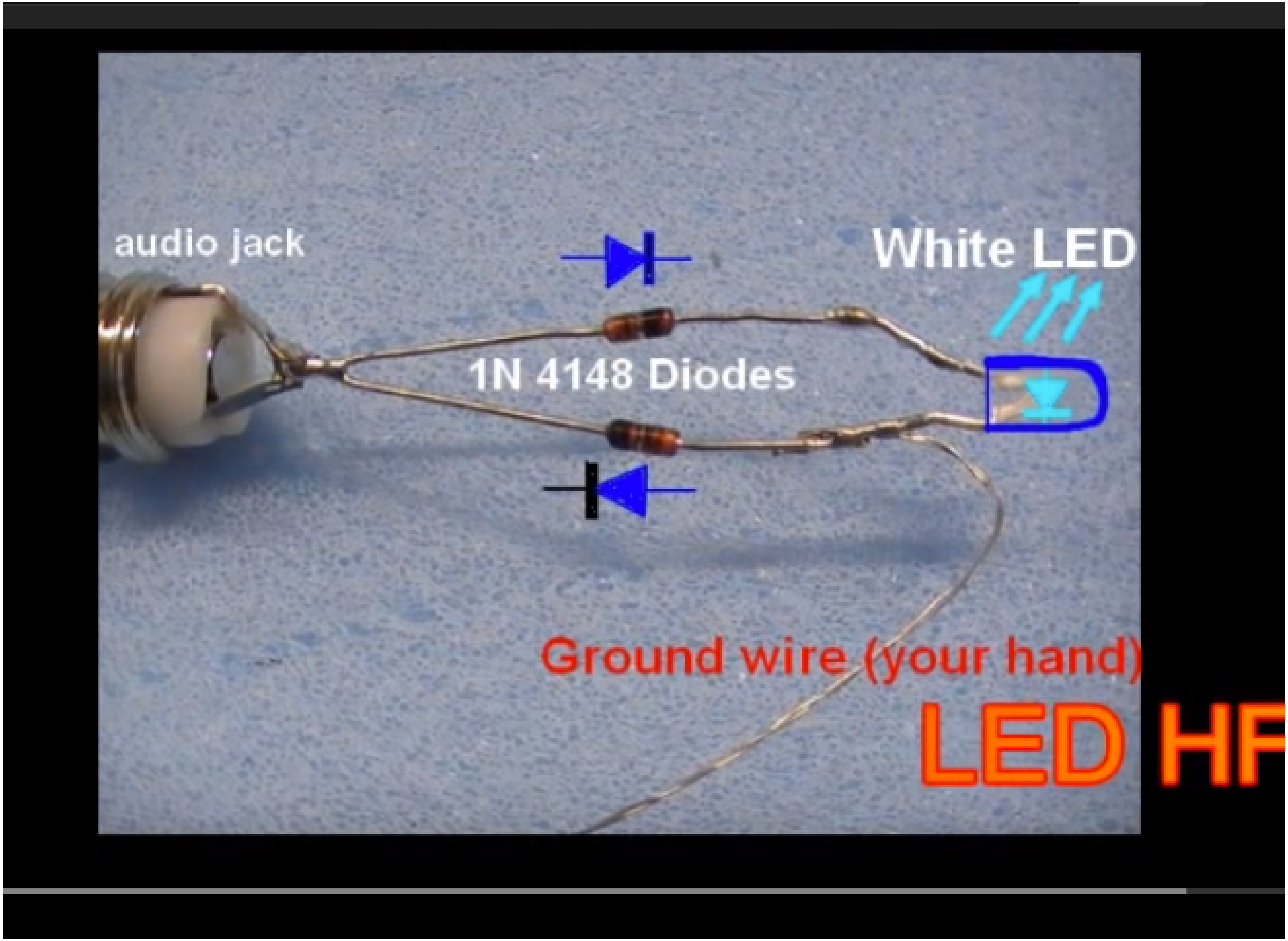
audio jack
White LED
1N 4148 Diodes
Ground wire (your hand)
LED HF

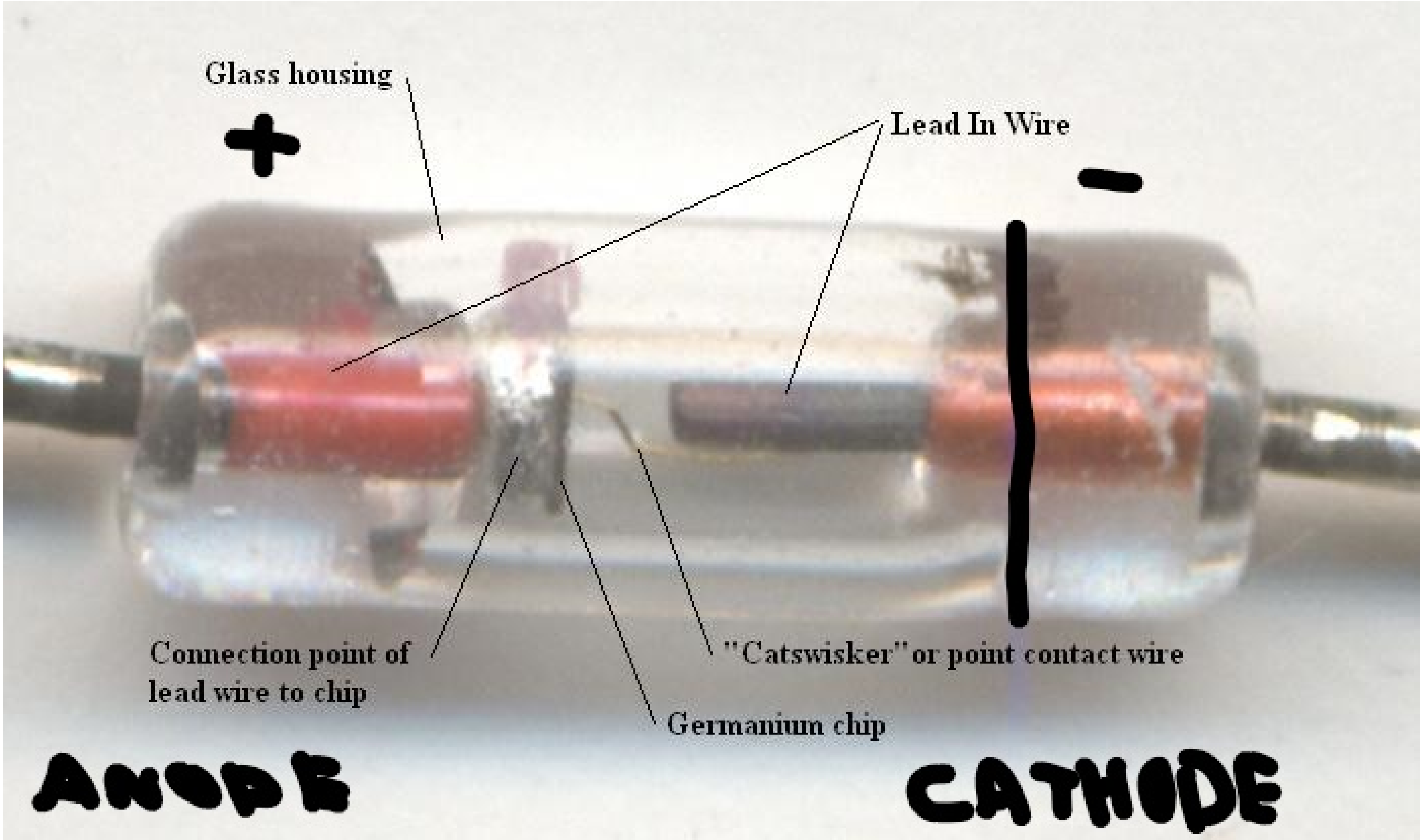
Glass housing
+
Lead In Wire
-
Connection point of
lead wire to chip
"Catswisker" or point contact wire
Germanium chip
ANODE
CATHODE

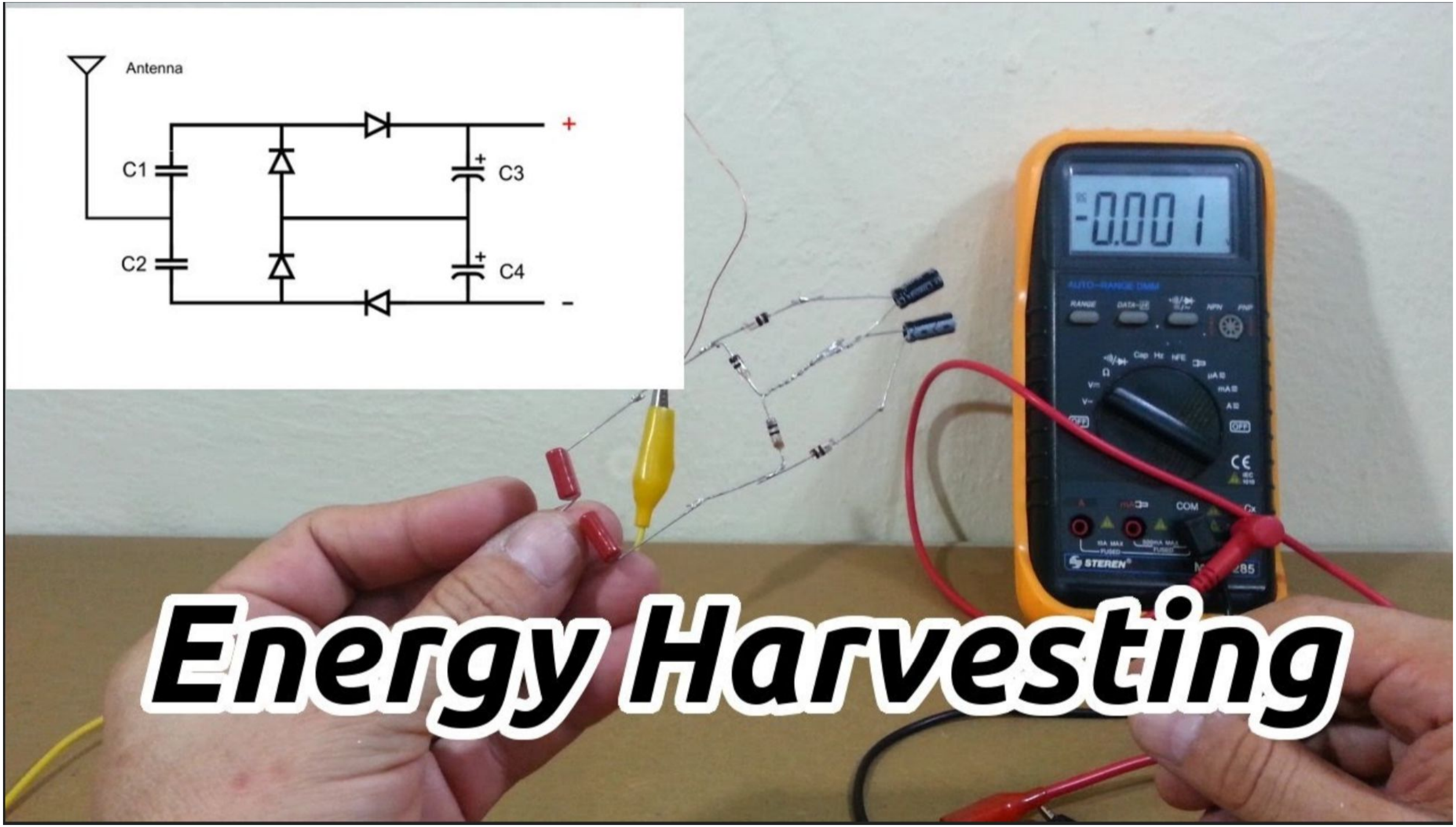
Antenna
+
C1
C3
C2
C4
-
-0.001
Energy Harvesting

feluri (fig. III.9), schemele fiind echivalente cu reprezentarea de bază din fig. III.8. Diodele sînt legate în serie, în formă de patrulater, două avînd comun anodul (punctul *2*), iar celelalte două catodul (punctul *4*). Tensiunea alternativă de intrare se aplică pe diagonala *1—3*, iar consumatorul se conectează pe diagonala *2—4*.

Pentru a urmări funcționarea punții, să presupunem că prima alternanță sosită în nodul *1* este pozitivă. Ea blochează dioda D_2 și o deschide pe D_1, debitînd prin R_S un curent I_1 (săgețile pline), care se întoarce la

tice sau cu parametri cît mai apro-

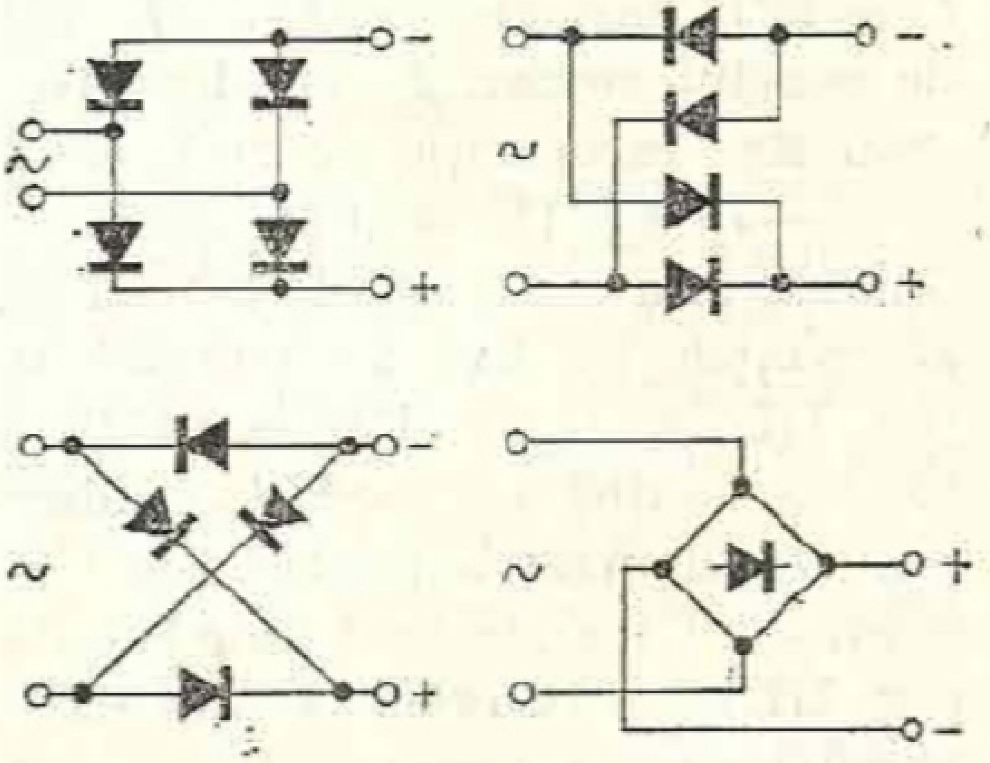

III.9. Puntea redresoare în diferite reprezentări.

circuits used on LC "RadioWaves"

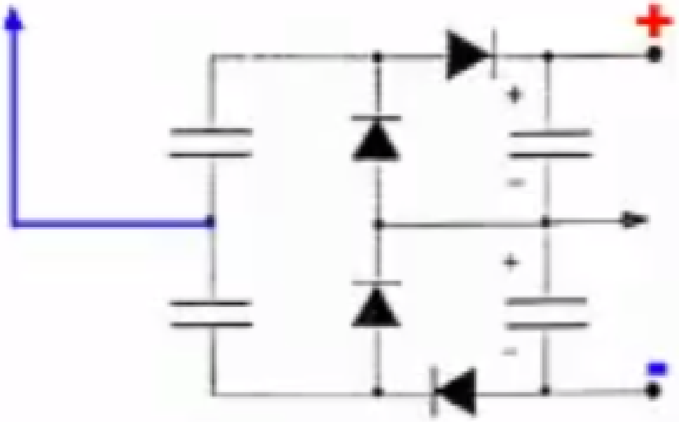

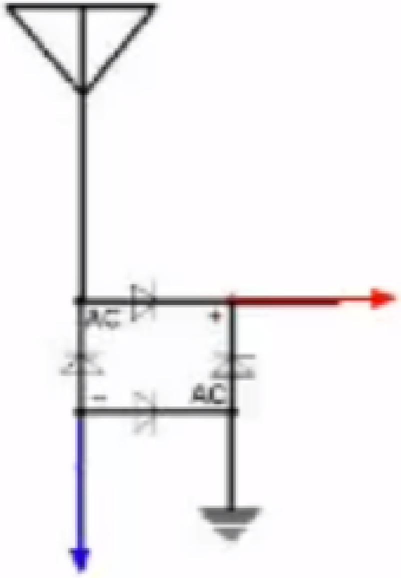

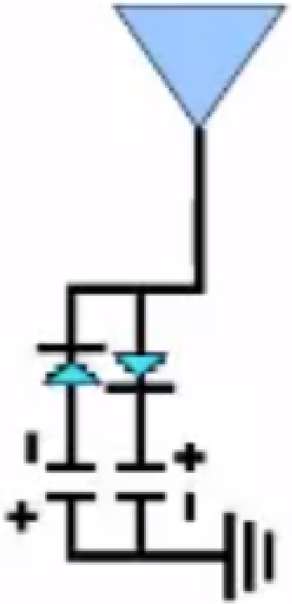

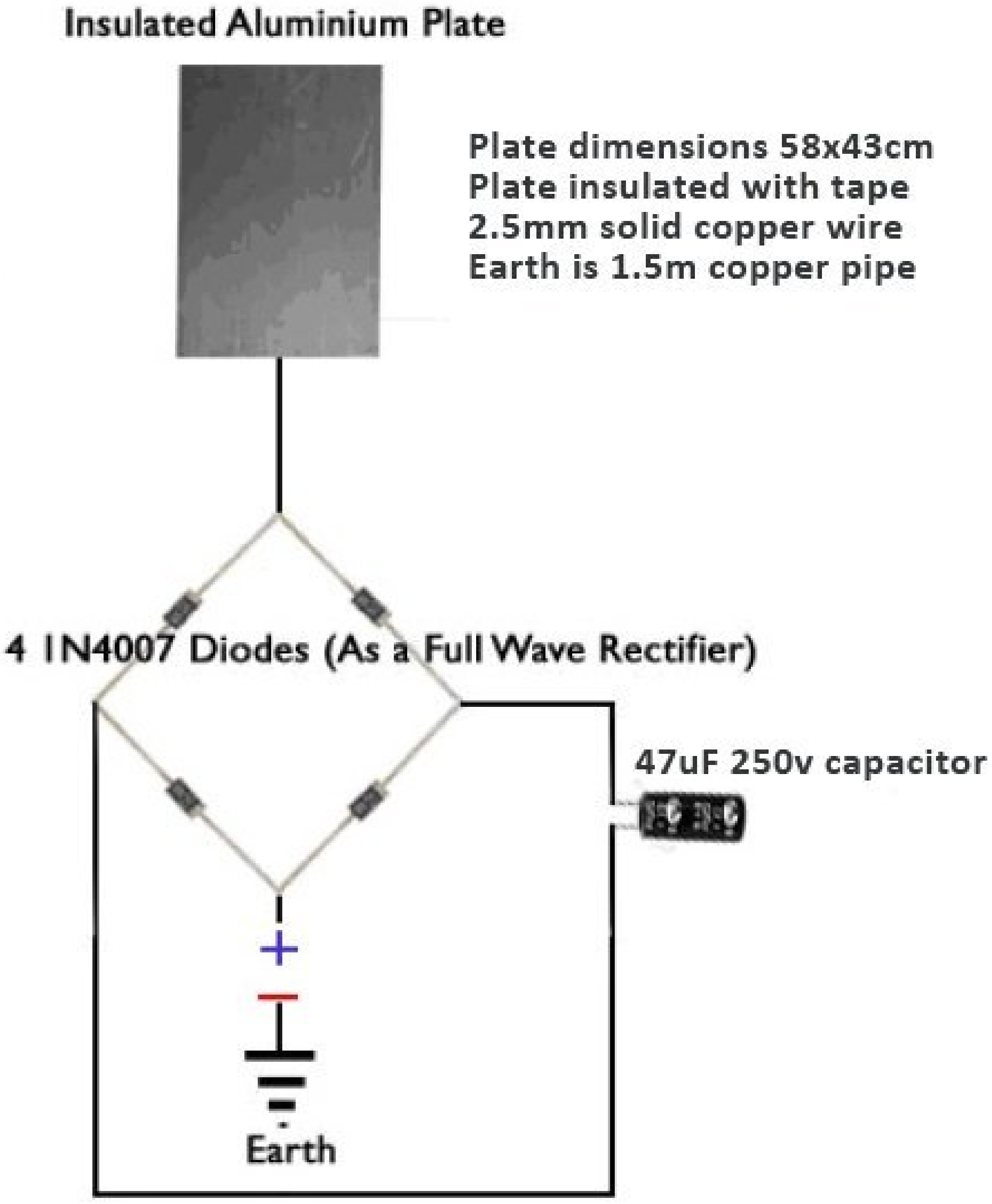
Insulated Aluminium Plate
Plate dimensions 58x43cm
Plate insulated with tape
2.5mm solid copper wire
Earth is 1.5m copper pipe
4 1N4007 Diodes (As a Full Wave Rectifier)
47uF 250v capacitor
+
−
Earth

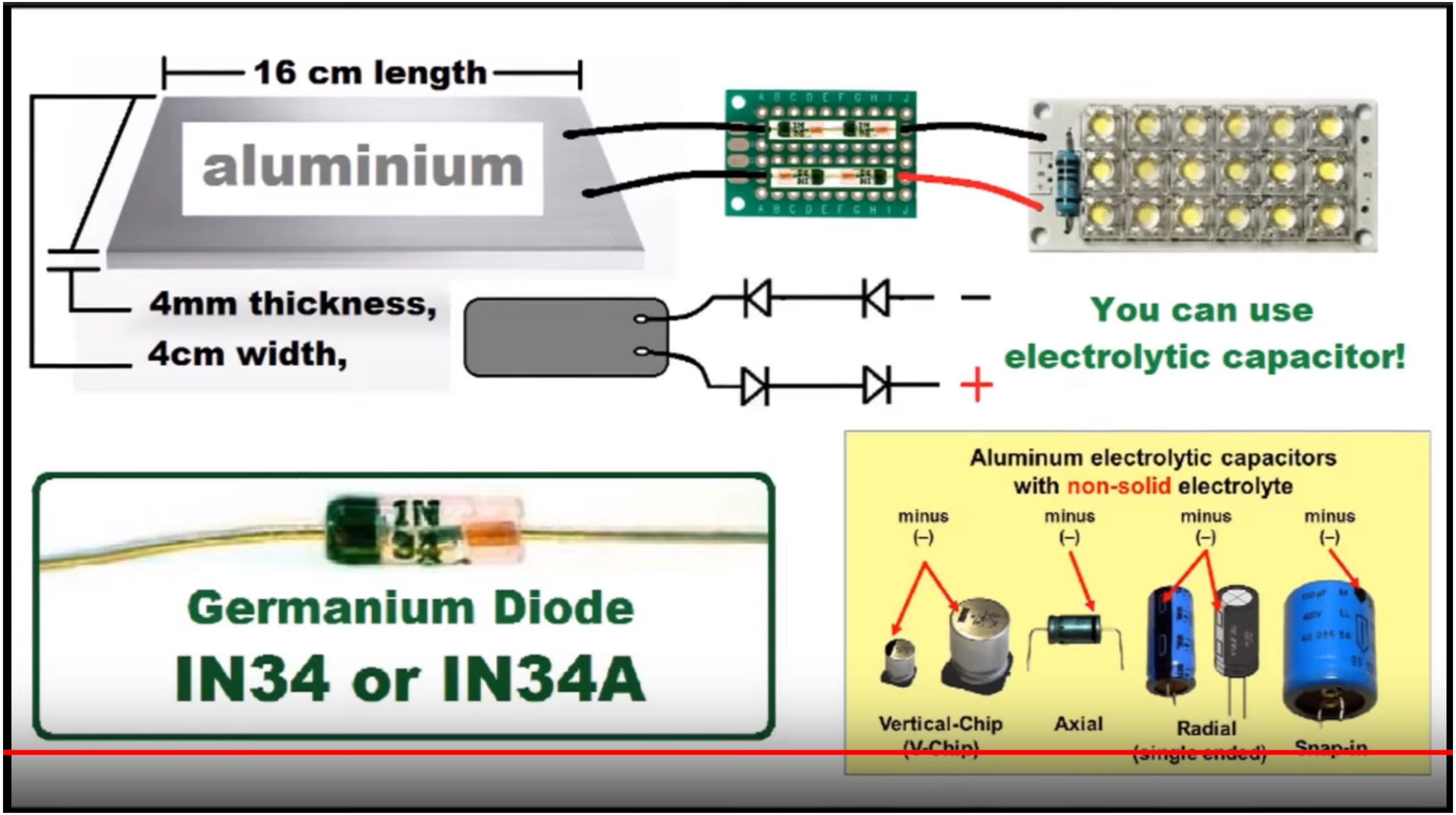

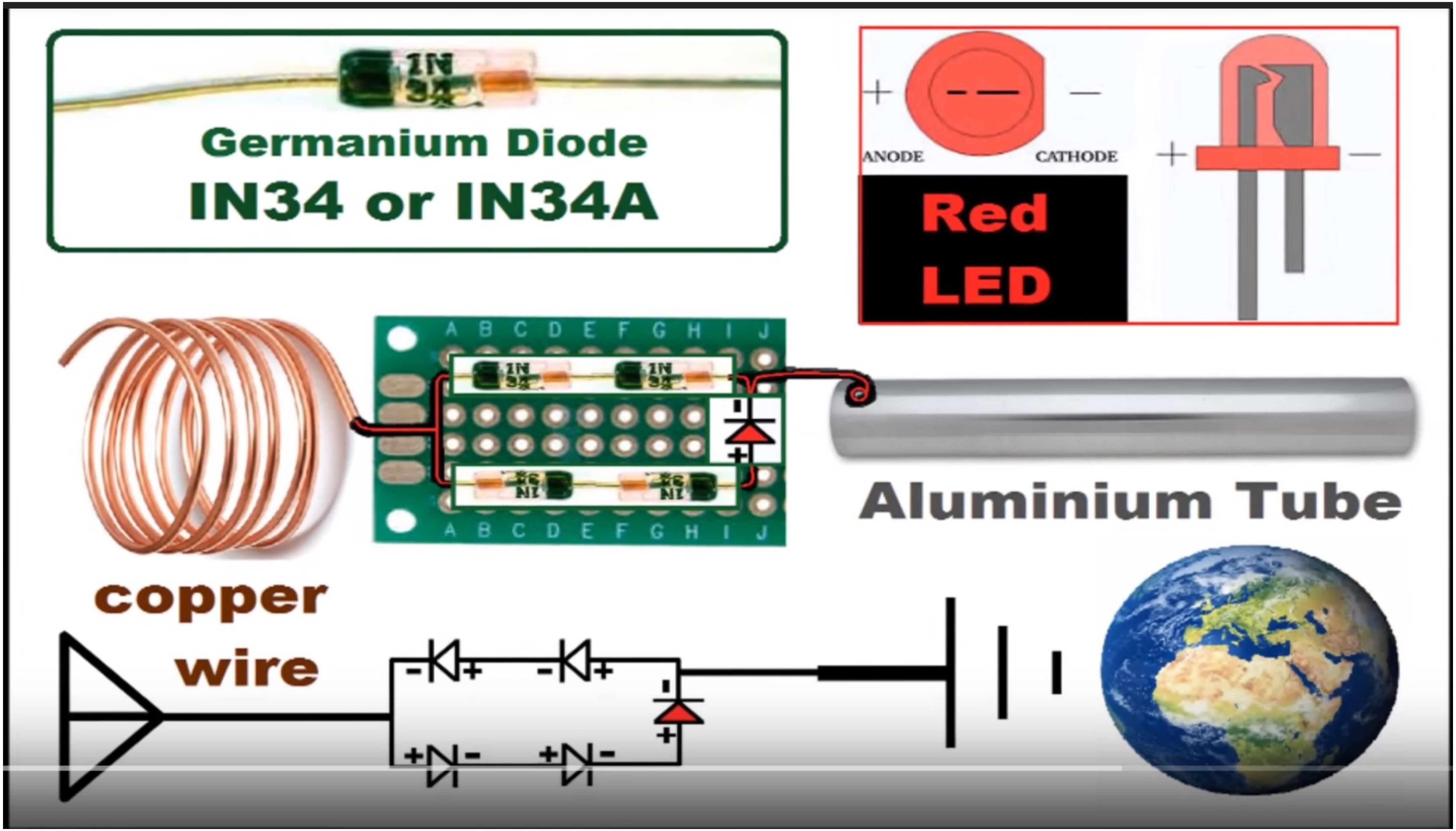
1N
34
Germanium Diode
IN34 or IN34A
+
−
ANODE
CATHODE
Red
LED
+
−
A B C D E F G H I J
1N
34
-
+
Aluminium Tube
copper
wire
-
+
-
+
-
+
+
-
+
-

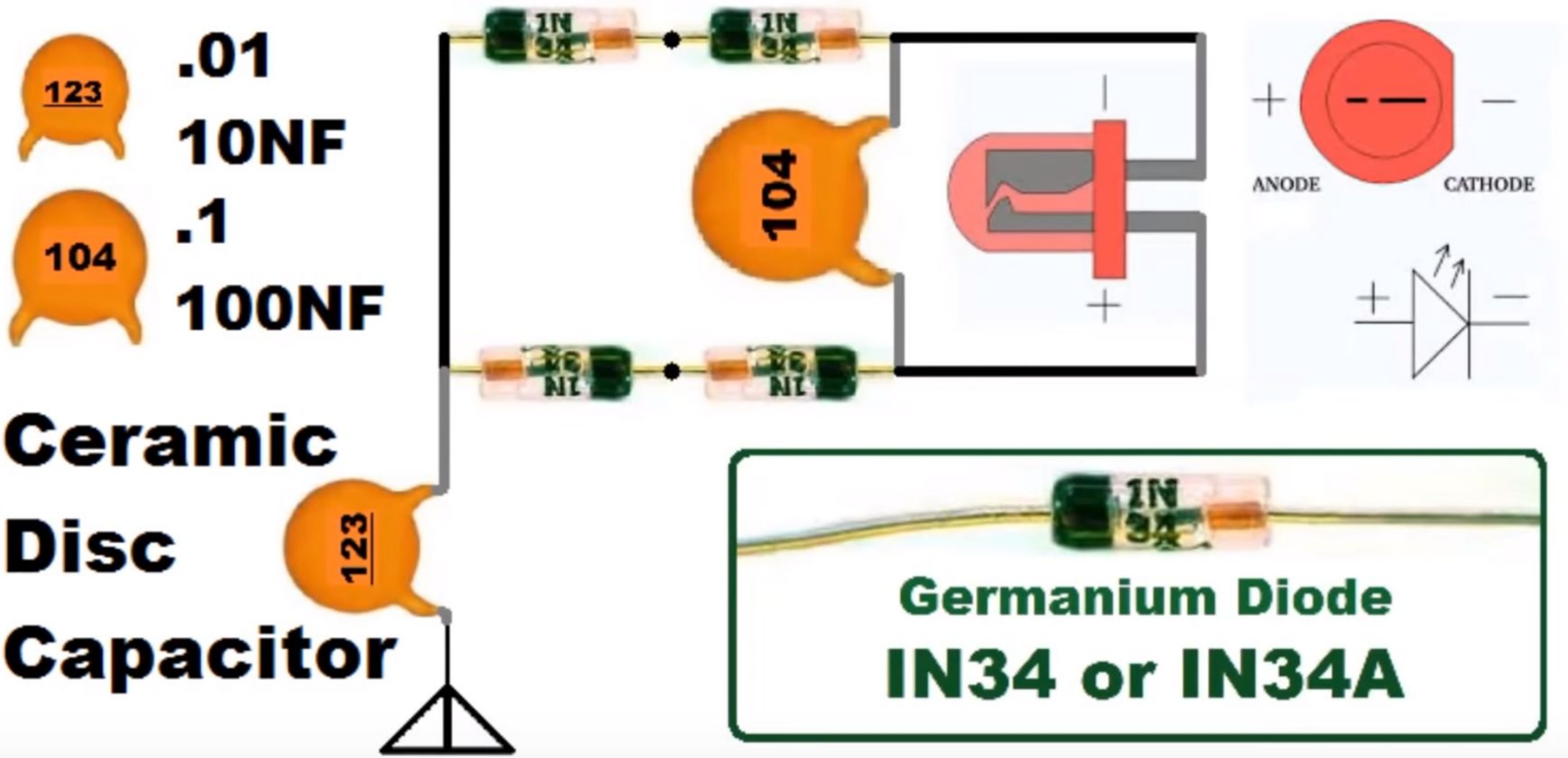
123
104
1N 34
104
+
−
ANODE
CATHODE
123

.01
10NF

.1
100NF

Ceramic
Disc
Capacitor

Germanium Diode
IN34 or IN34A

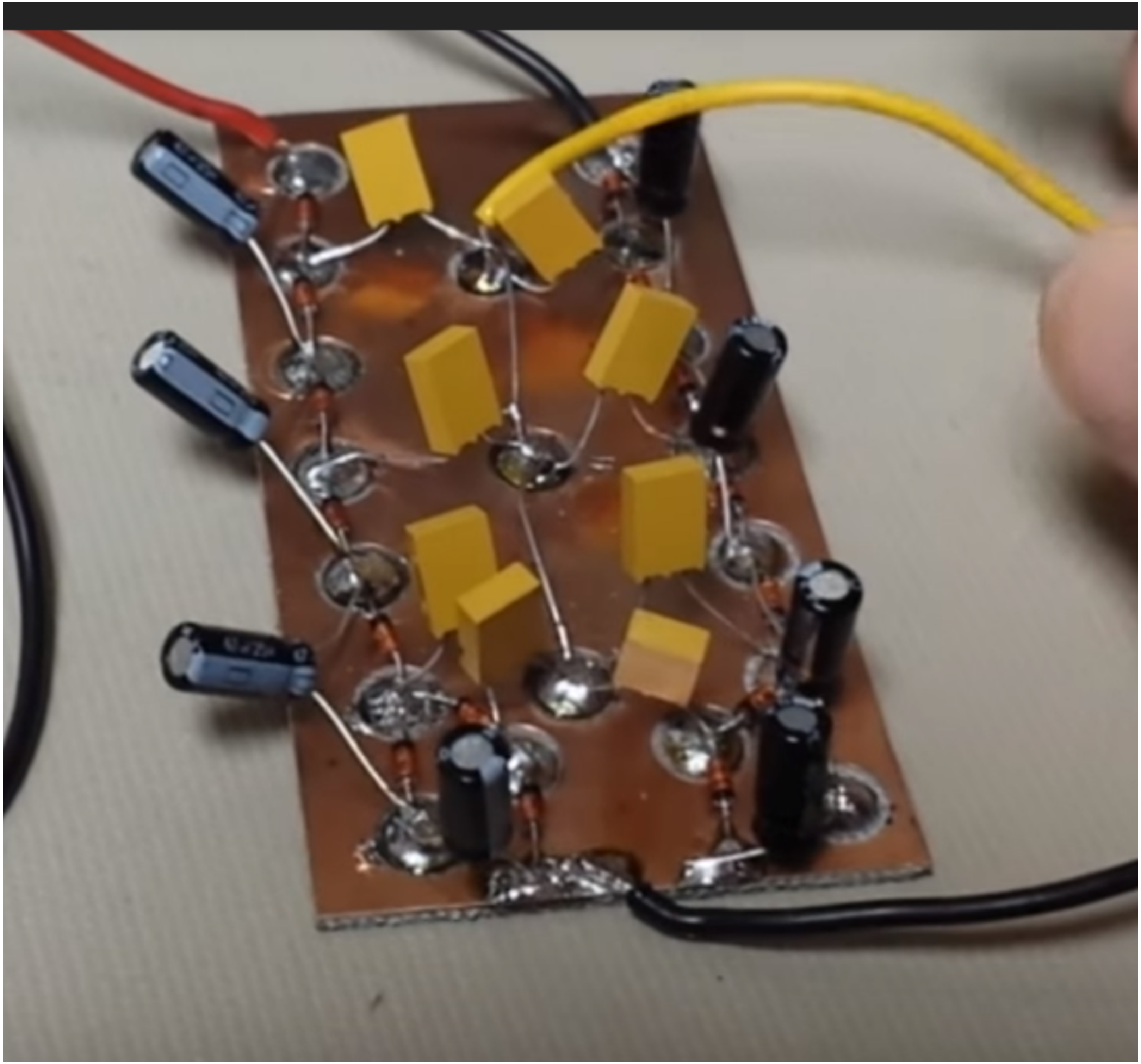

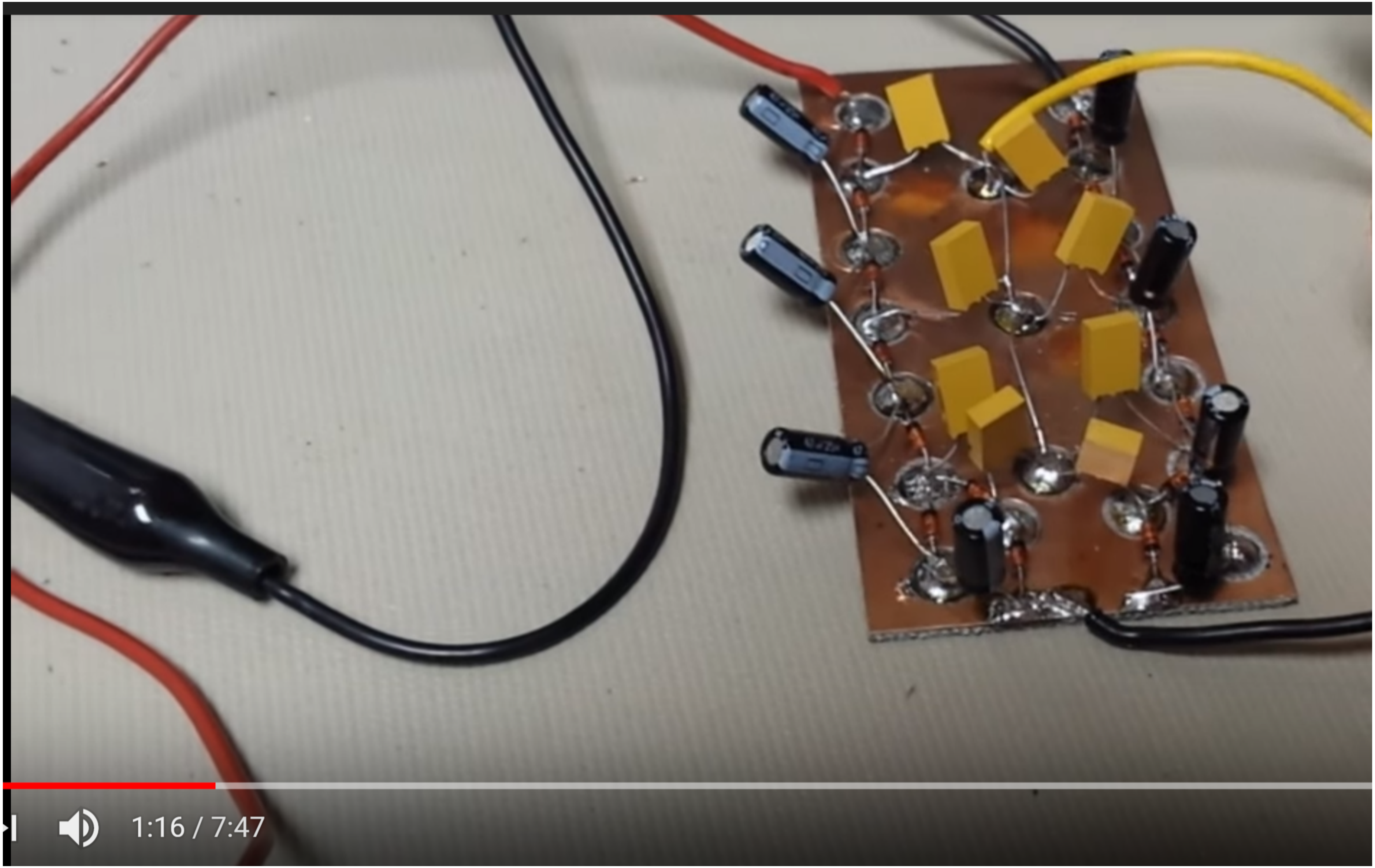

1:16 / 7:47

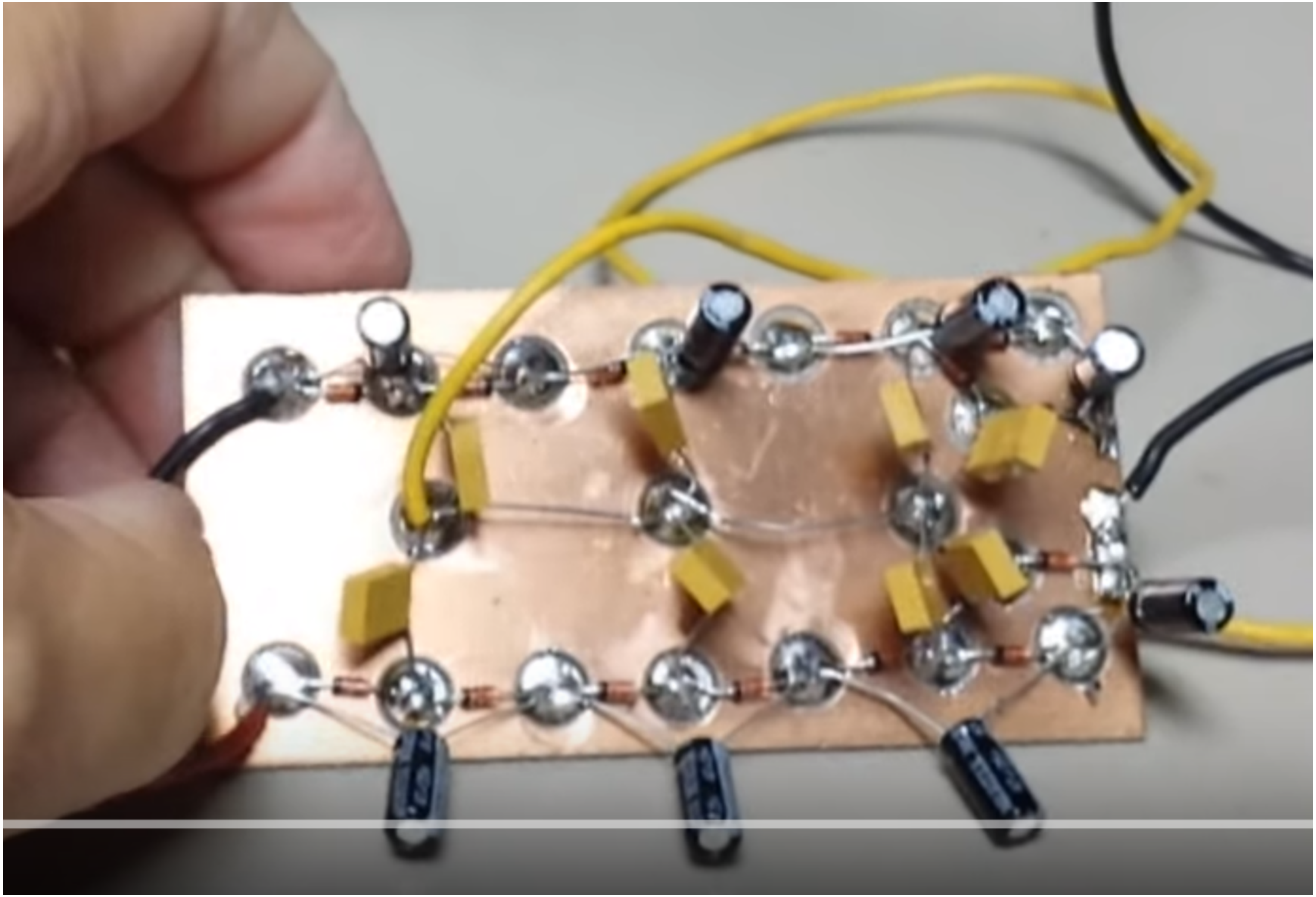

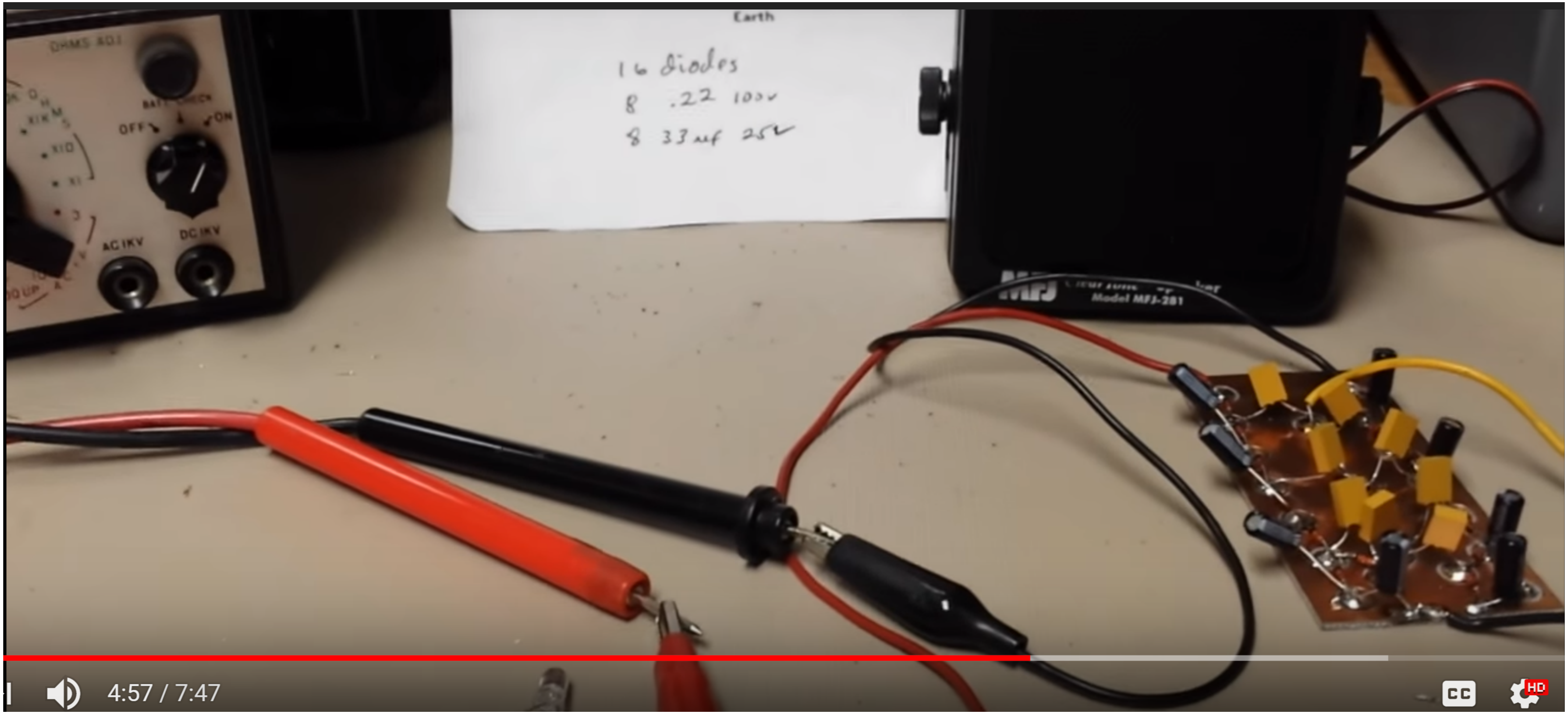
Earth
16 diodes
8 .22 100v
8 33uf 25v
OHMS ADJ
BATT CHECK
OFF
ON
AC IKV
DC IKV
MFJ
Model MFJ-281
4:57 / 7:47
CC
HD

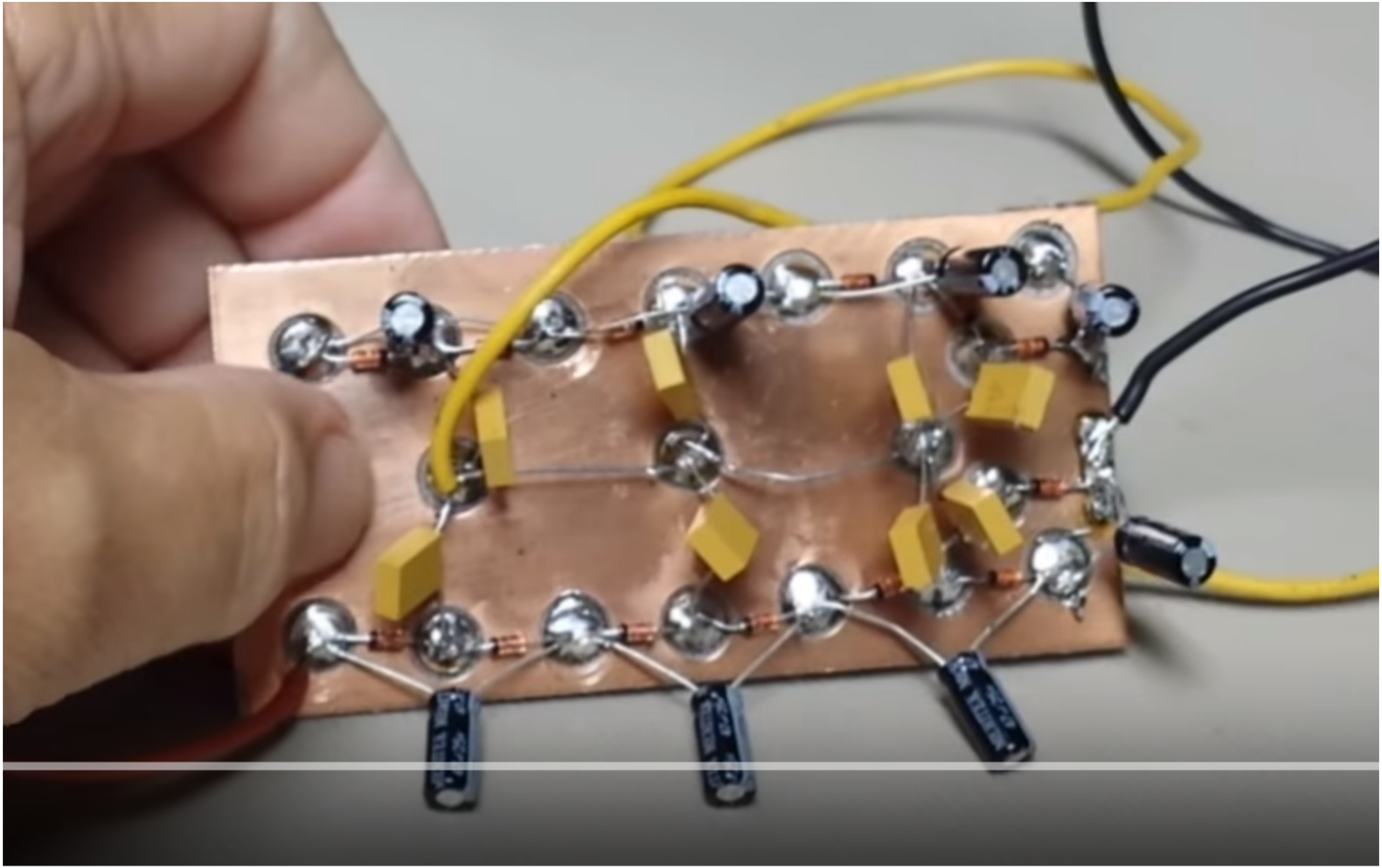

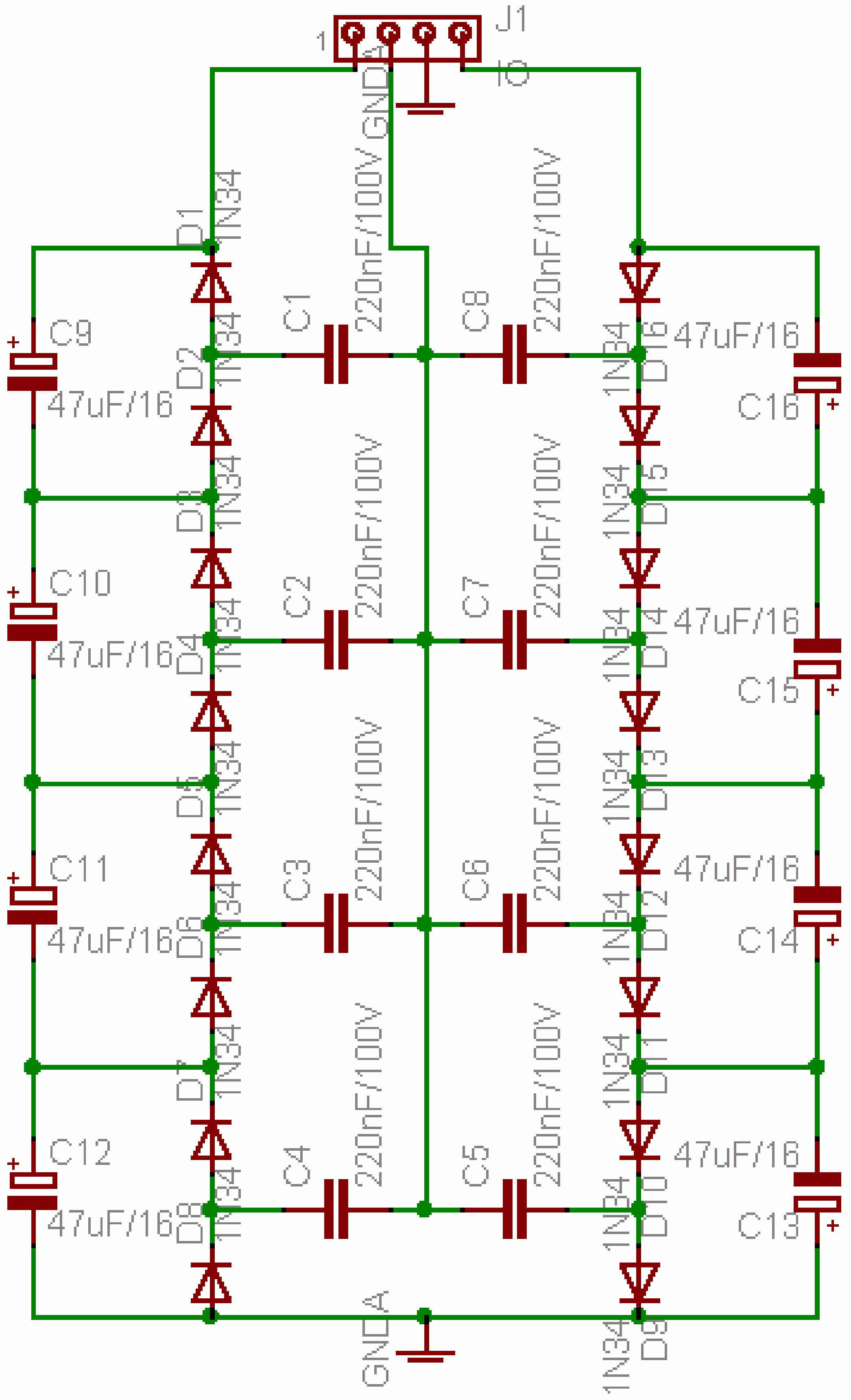
J1
1
IO
GNDA
D1 1N34
D2 1N34
D3 1N34
D4 1N34
D5 1N34
D6 1N34
D7 1N34
D8 1N34
D9 1N34
D10 1N34
D11 1N34
D12 1N34
D13 1N34
D14 1N34
D15 1N34
D16 1N34
C1 220nF/100V
C2 220nF/100V
C3 220nF/100V
C4 220nF/100V
C5 220nF/100V
C6 220nF/100V
C7 220nF/100V
C8 220nF/100V
C9 47uF/16
C10 47uF/16
C11 47uF/16
C12 47uF/16
C13 47uF/16
C14 47uF/16
C15 47uF/16
C16 47uF/16
+
GNDA

AMBIENT POWER MODULE

(C) - GROUNDLOOP 2012

Circuit Diagram

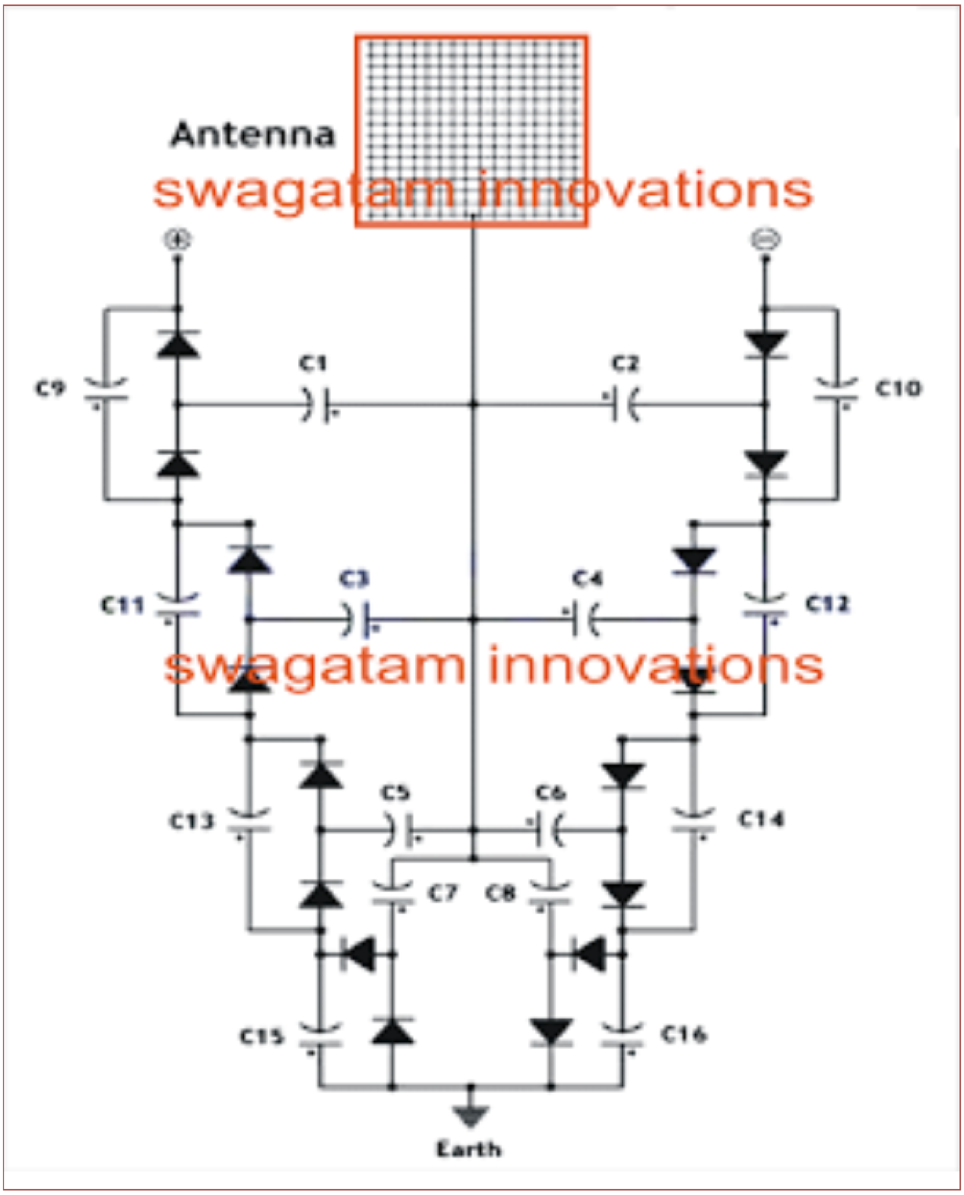
Antenna
swagatam innovations
⊕
⊖
C1
C2
C9
C10
C3
C4
C11
C12
swagatam innovations
C5
C6
C13
C14
C7
C8
C15
C16
Earth

Parts List

All Diodes are 1N4148

C1---C8 = 0.22uF/100V mylar

C9----C16 = 33uF/25V electrolytic

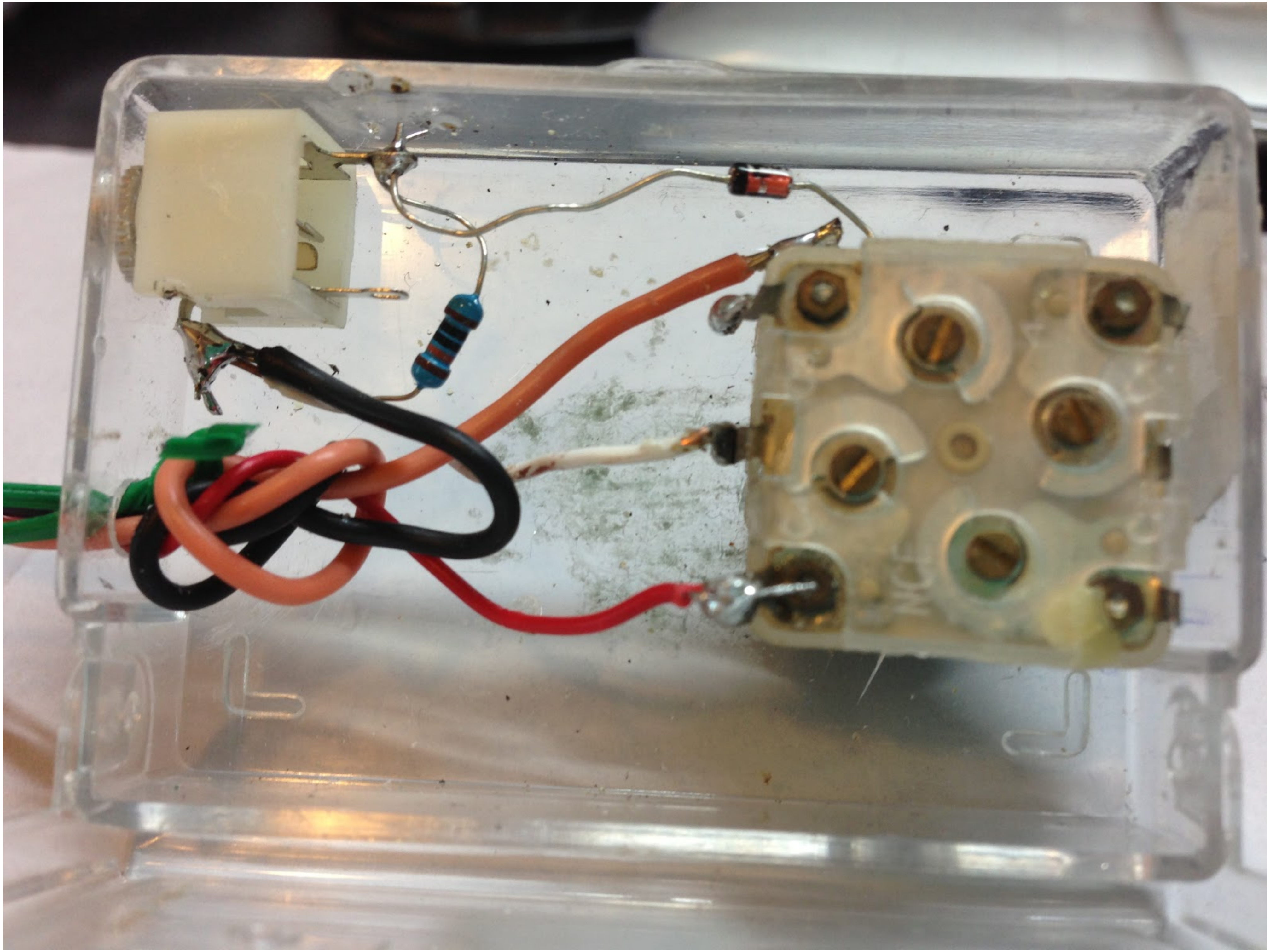

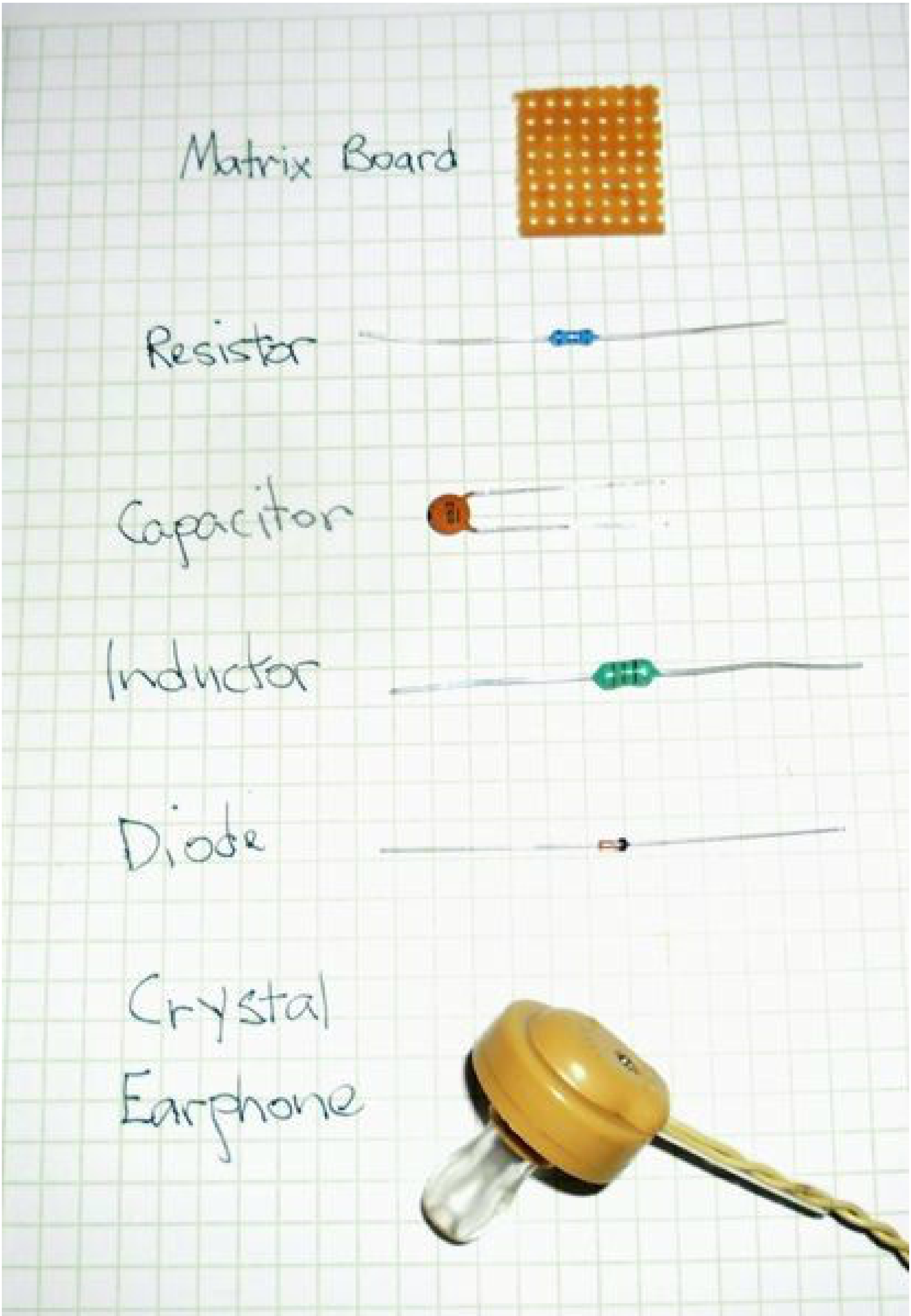
Matrix Board
Resistor
Capacitor
Inductor
Diode
Crystal Earphone

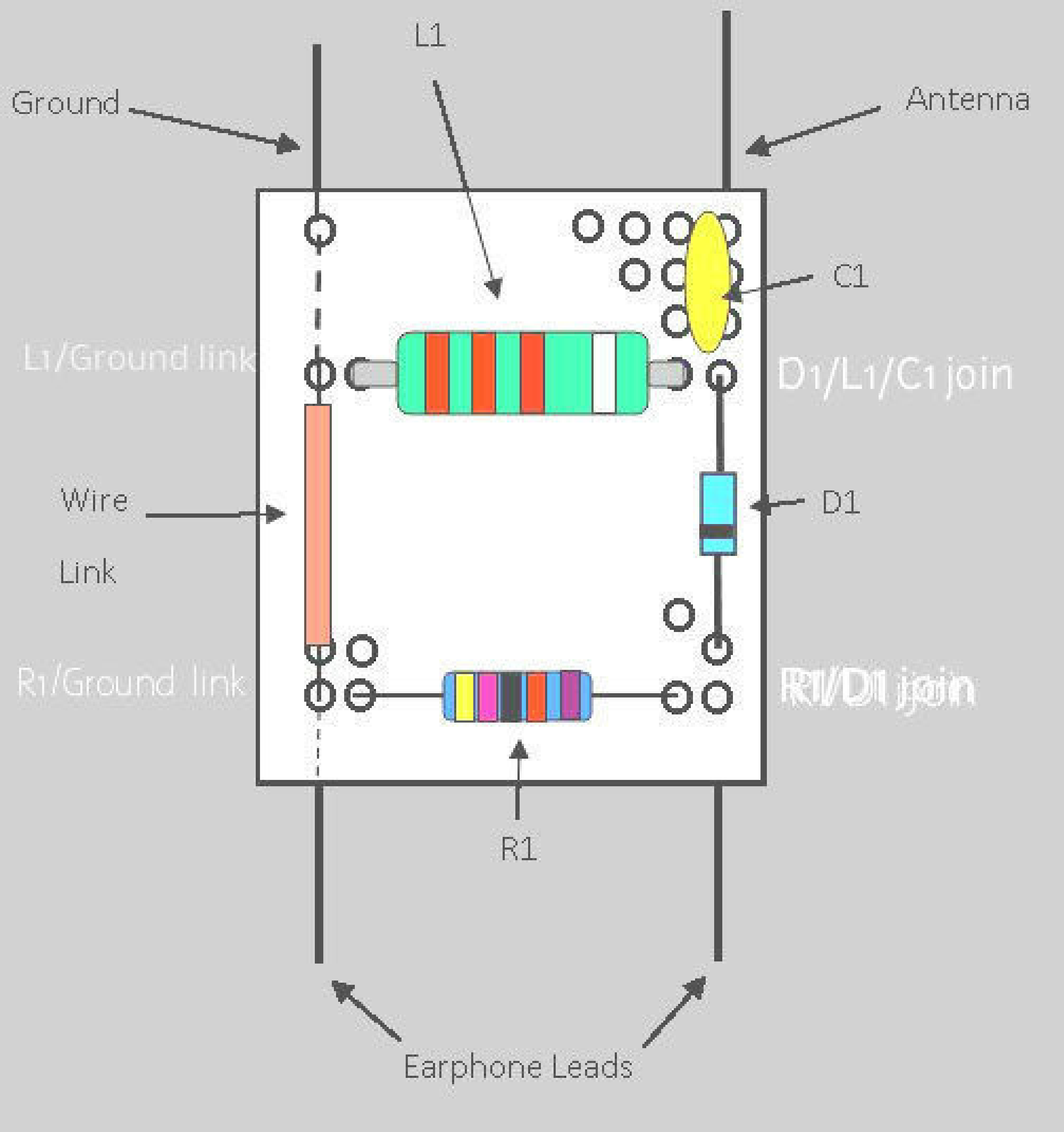
L1
Ground
Antenna
C1
L1/Ground link
D1/L1/C1 join
Wire
Link
D1
R1/Ground link
R1/D1 join
R1
Earphone Leads

24g NET
FRESHNESS IN 2 CALORIES
tic
tac
Strawberry
Fields

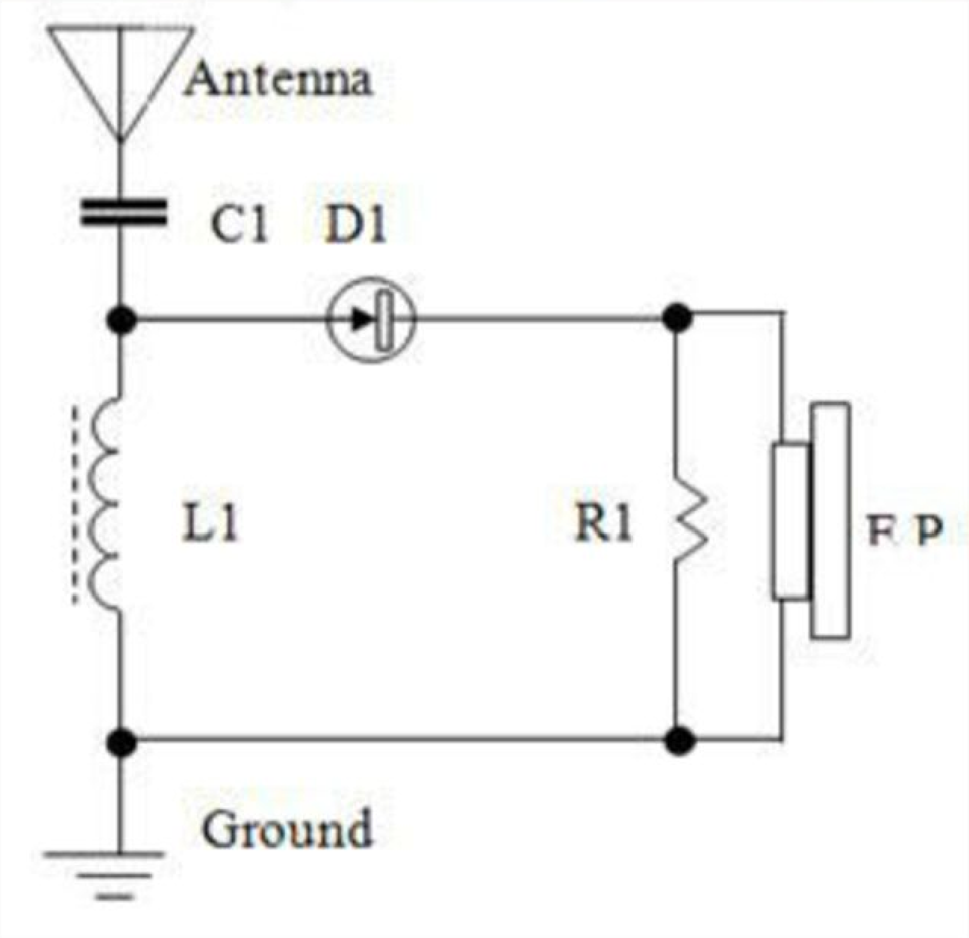
Antenna
C1
D1
L1
R1
E P
Ground

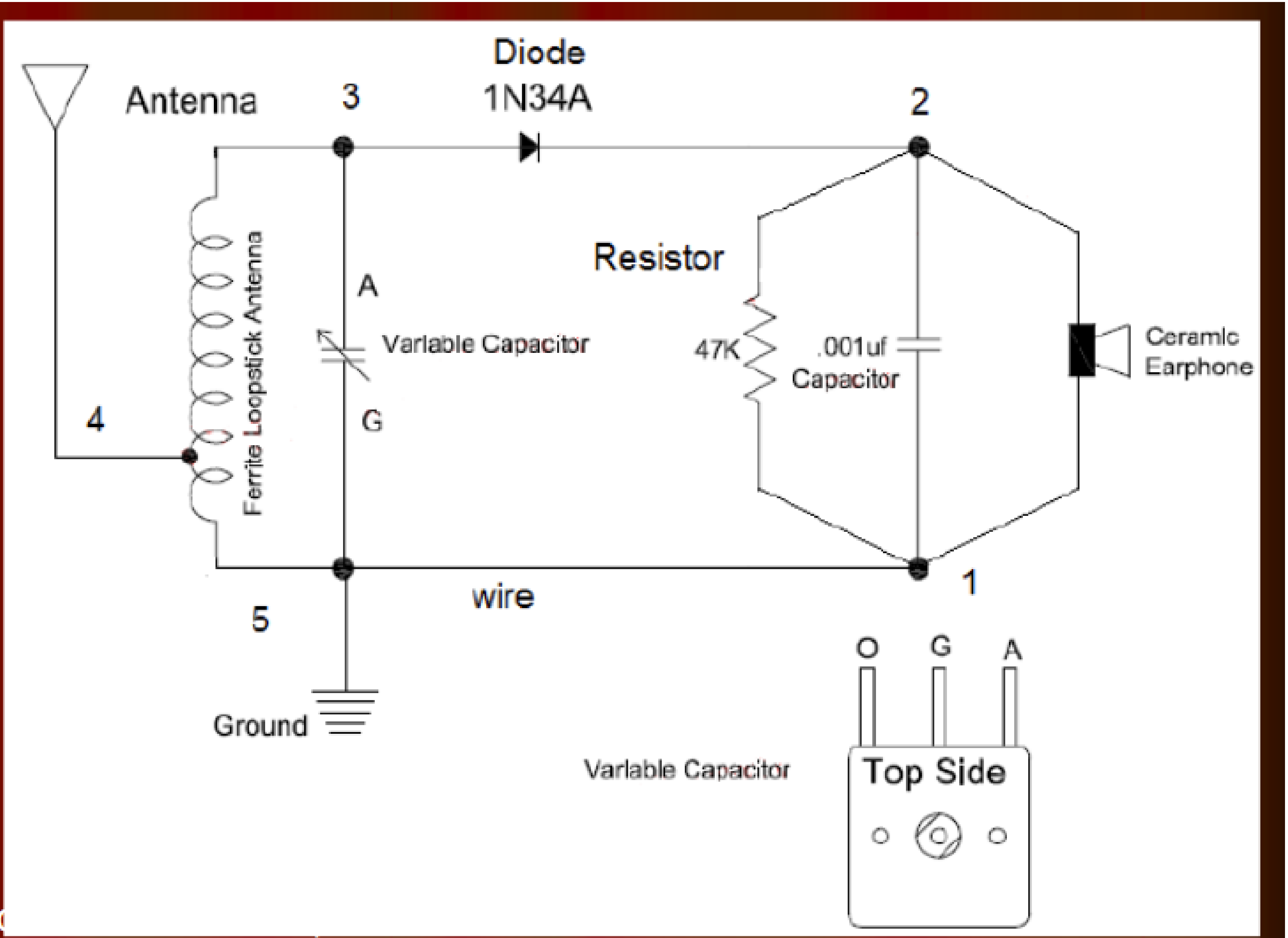
Diode
1N34A
Antenna
3
2
Resistor
A
Variable Capacitor
47K
.001uf
Capacitor
Ceramic
Earphone
Ferrite Loopstick Antenna
G
4
1
wire
5
Ground
O
G
A
Variable Capacitor
Top Side

[Sc
http://www.angelfire.com/electronic2/index1/DmCrystalSet.html]

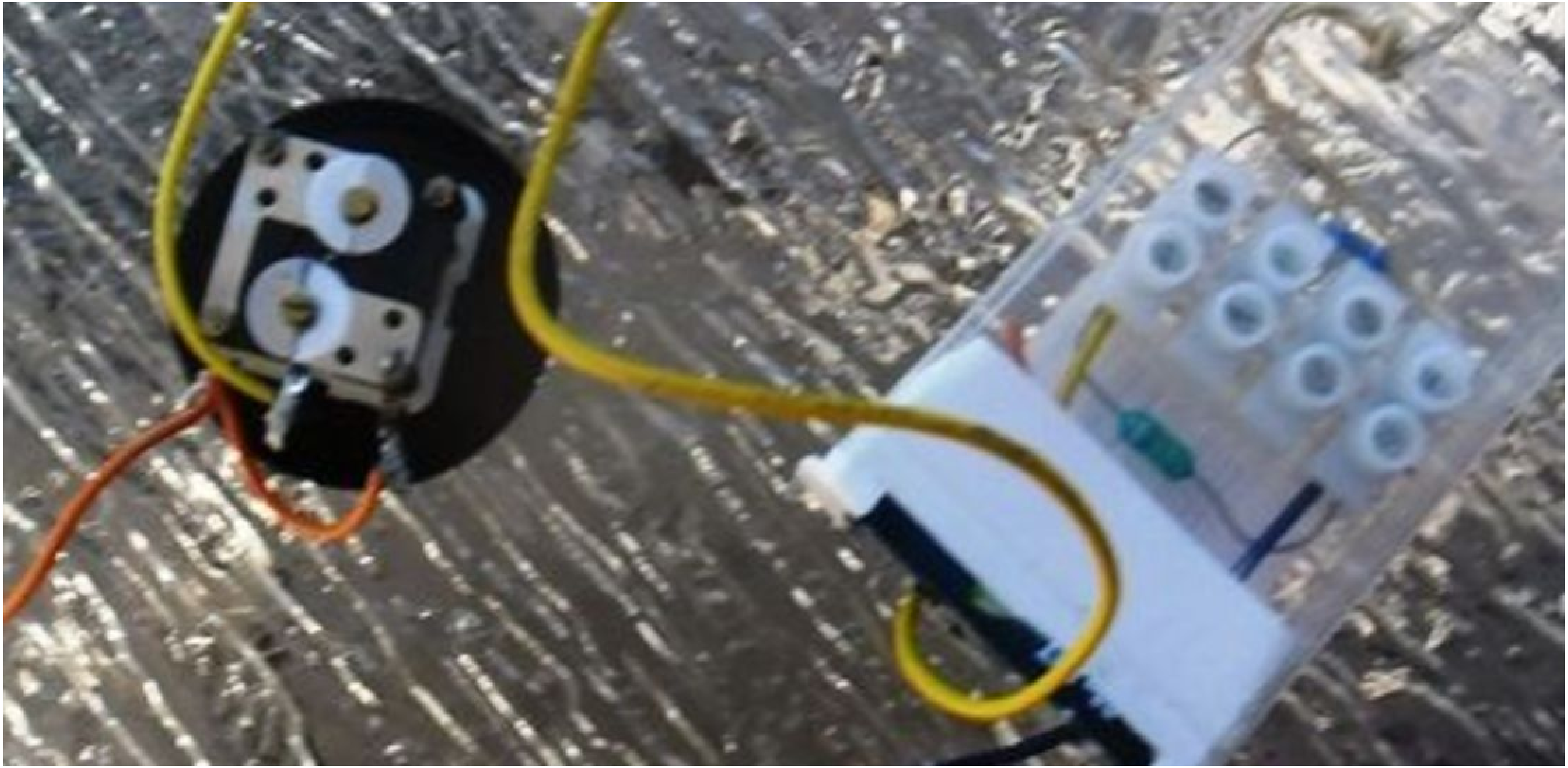

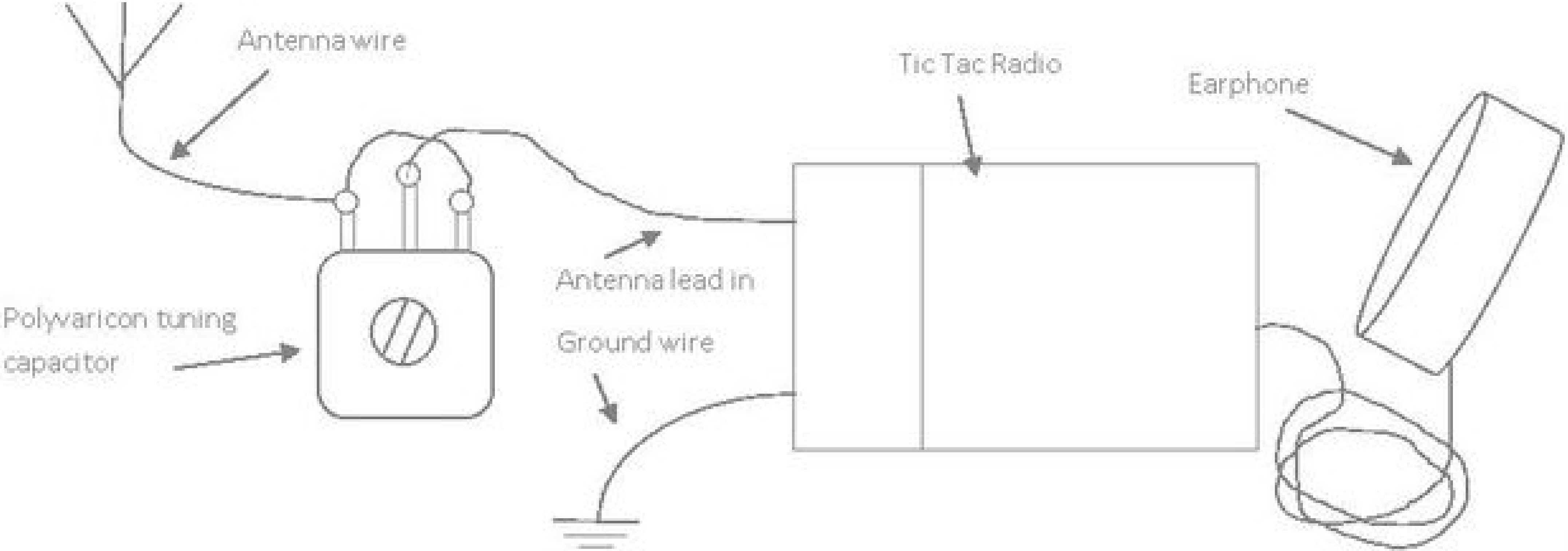
Antenna wire
Tic Tac Radio
Earphone
Antenna lead in
Polyvaricon tuning capacitor
Ground wire

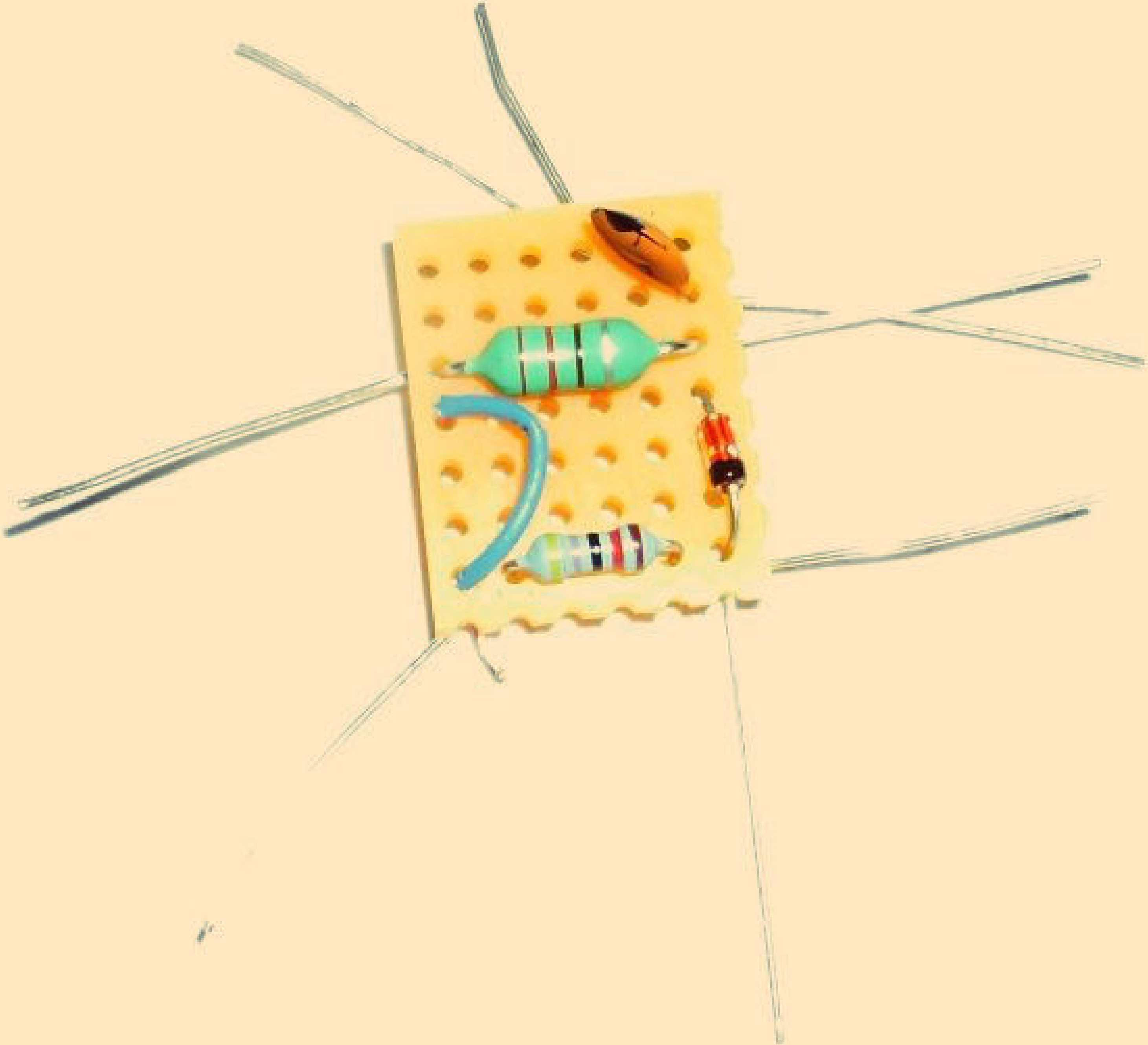

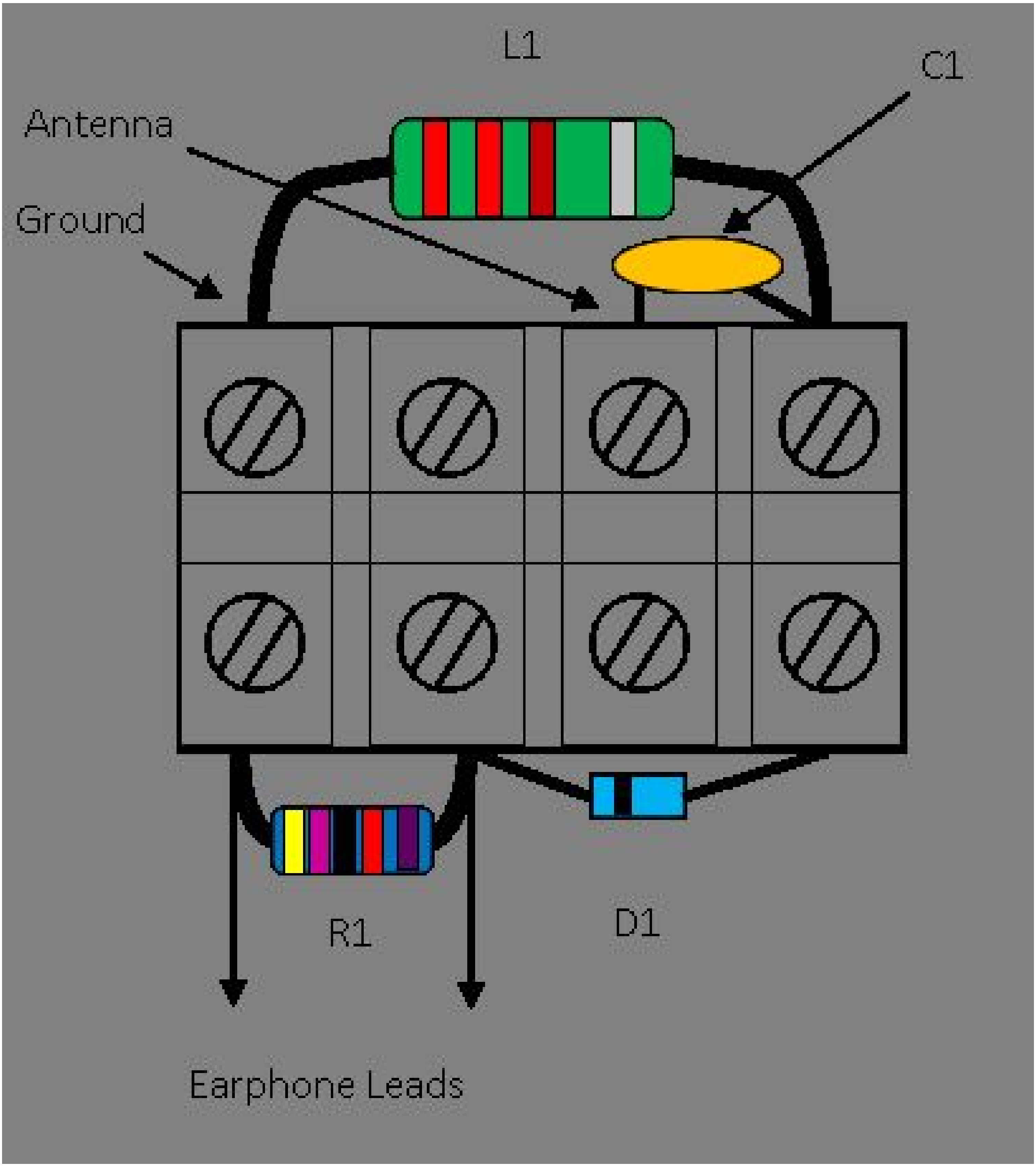
L1
C1
Antenna
Ground
R1
D1
Earphone Leads

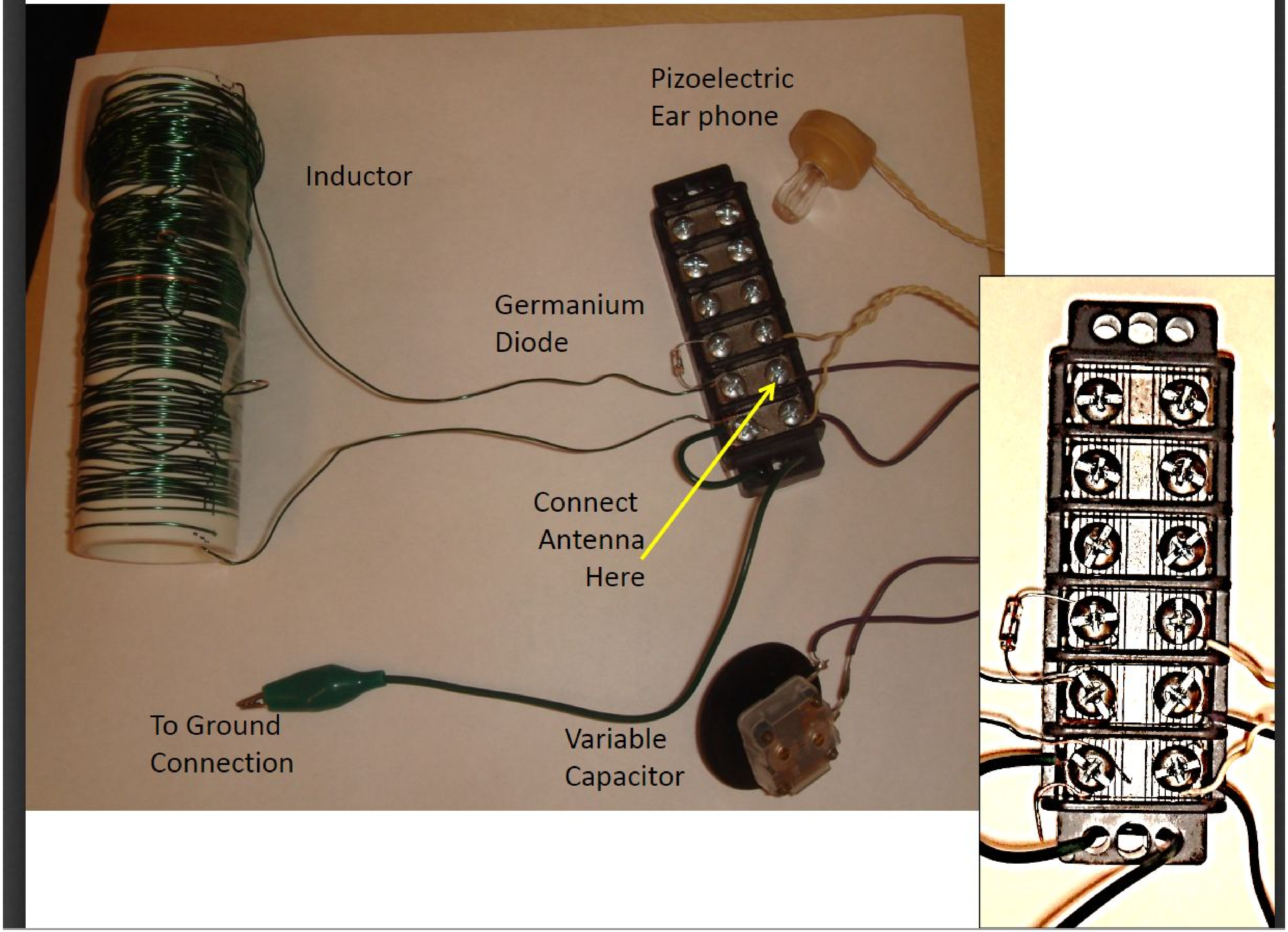

Pizoelectric
Ear phone

Inductor

Germanium
Diode

Connect
Antenna
Here

To Ground
Connection

Variable
Capacitor

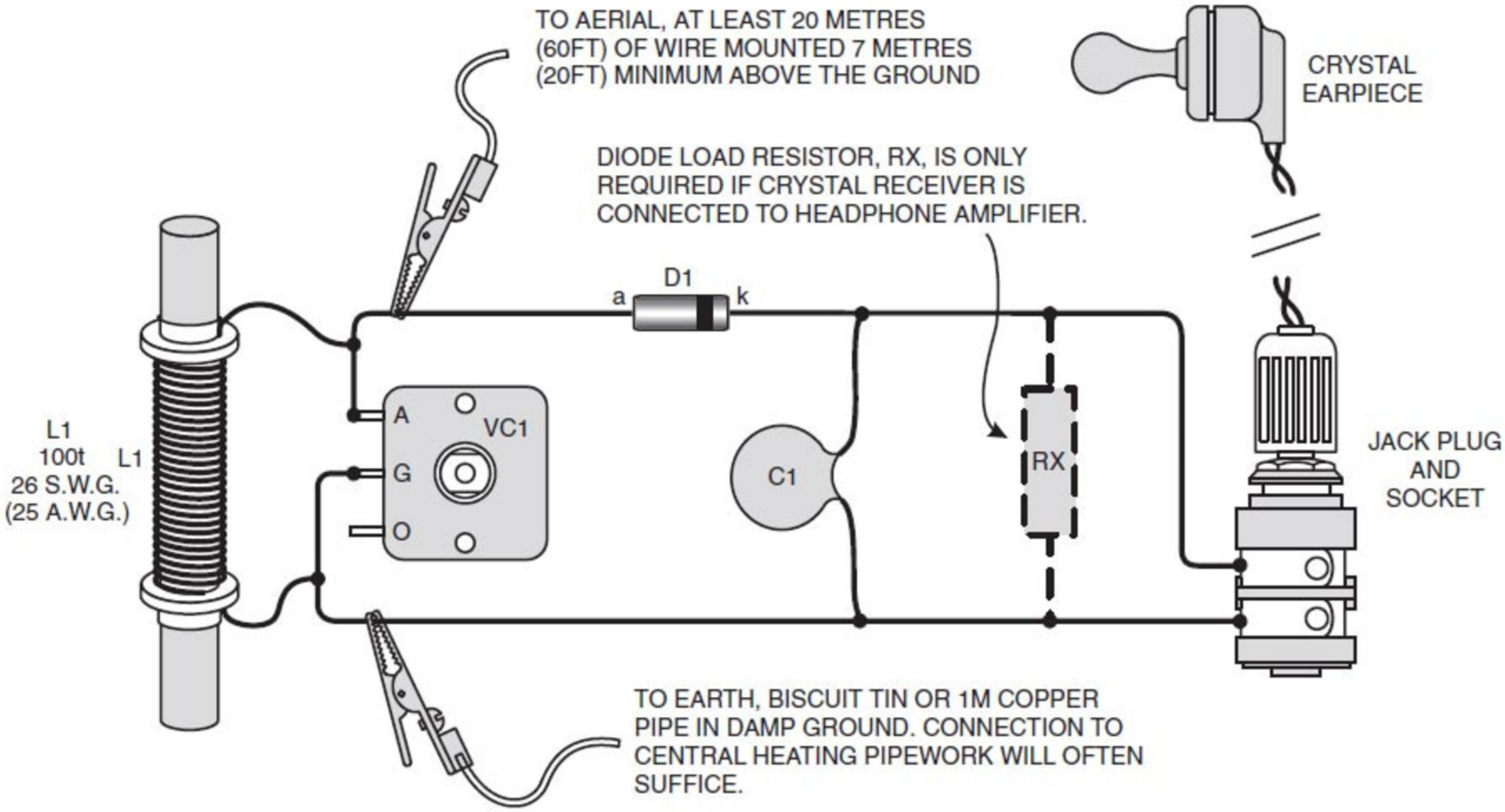
CRYSTAL
EARPIECE
D1
a
k
A
G
O
VC1
C1
RX
L1
100t
26 S.W.G.
(25 A.W.G.)
L1
JACK PLUG
AND
SOCKET

TO AERIAL, AT LEAST 20 METRES (60FT) OF WIRE MOUNTED 7 METRES (20FT) MINIMUM ABOVE THE GROUND

DIODE LOAD RESISTOR, RX, IS ONLY REQUIRED IF CRYSTAL RECEIVER IS CONNECTED TO HEADPHONE AMPLIFIER.

TO EARTH, BISCUIT TIN OR 1M COPPER PIPE IN DAMP GROUND. CONNECTION TO CENTRAL HEATING PIPEWORK WILL OFTEN SUFFICE.

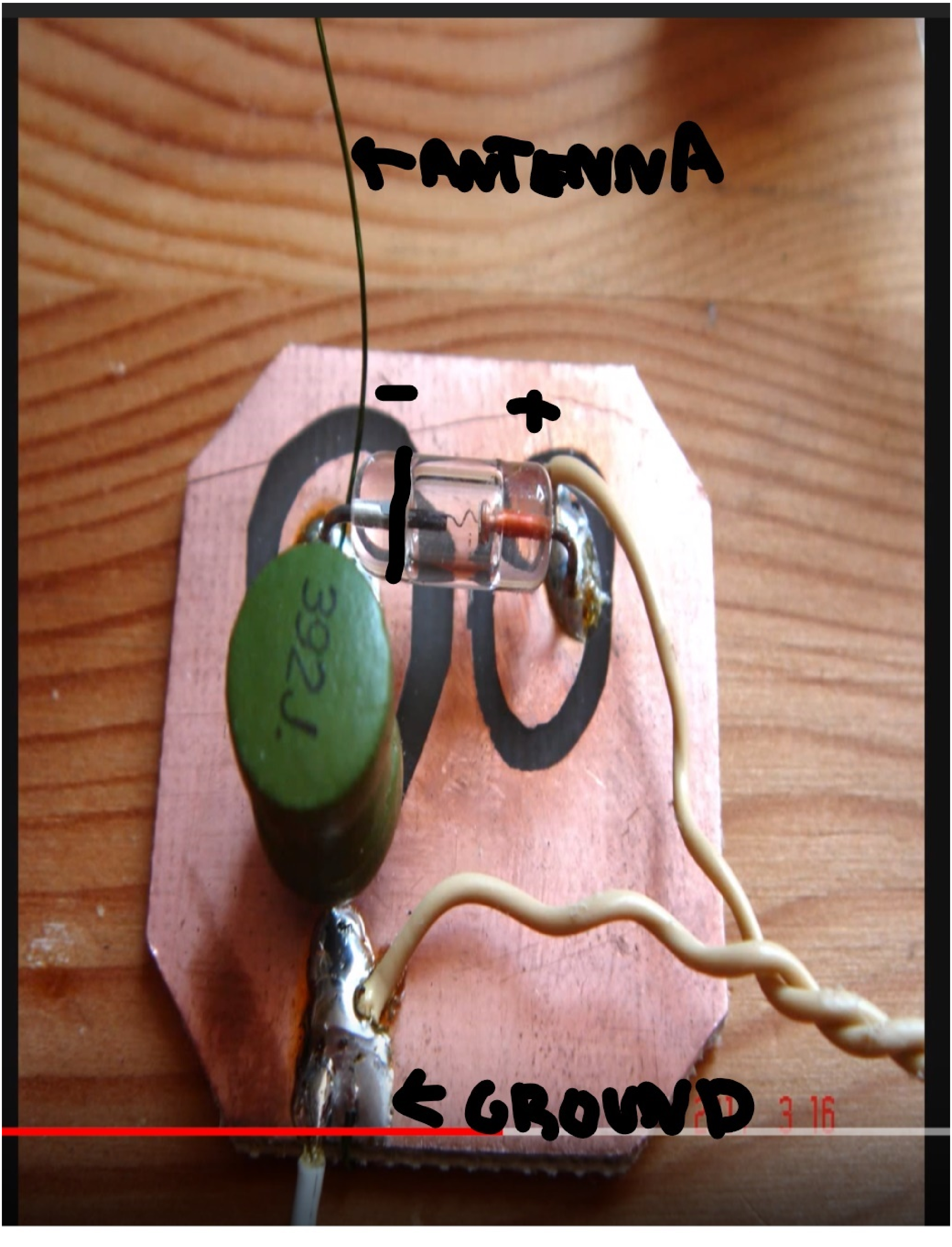
← ANTENNA
−
+
392J.
← GROUND
3 16

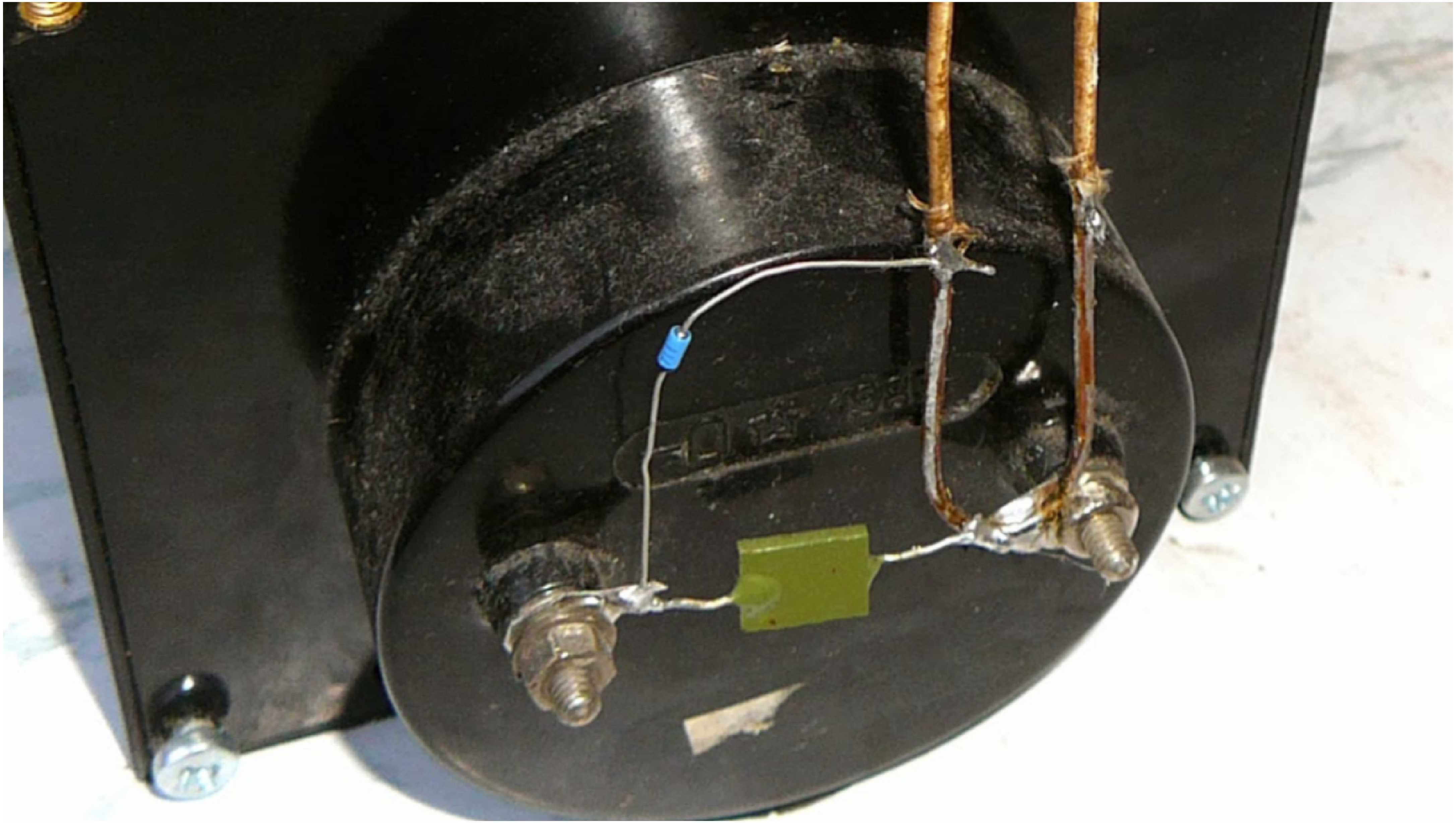

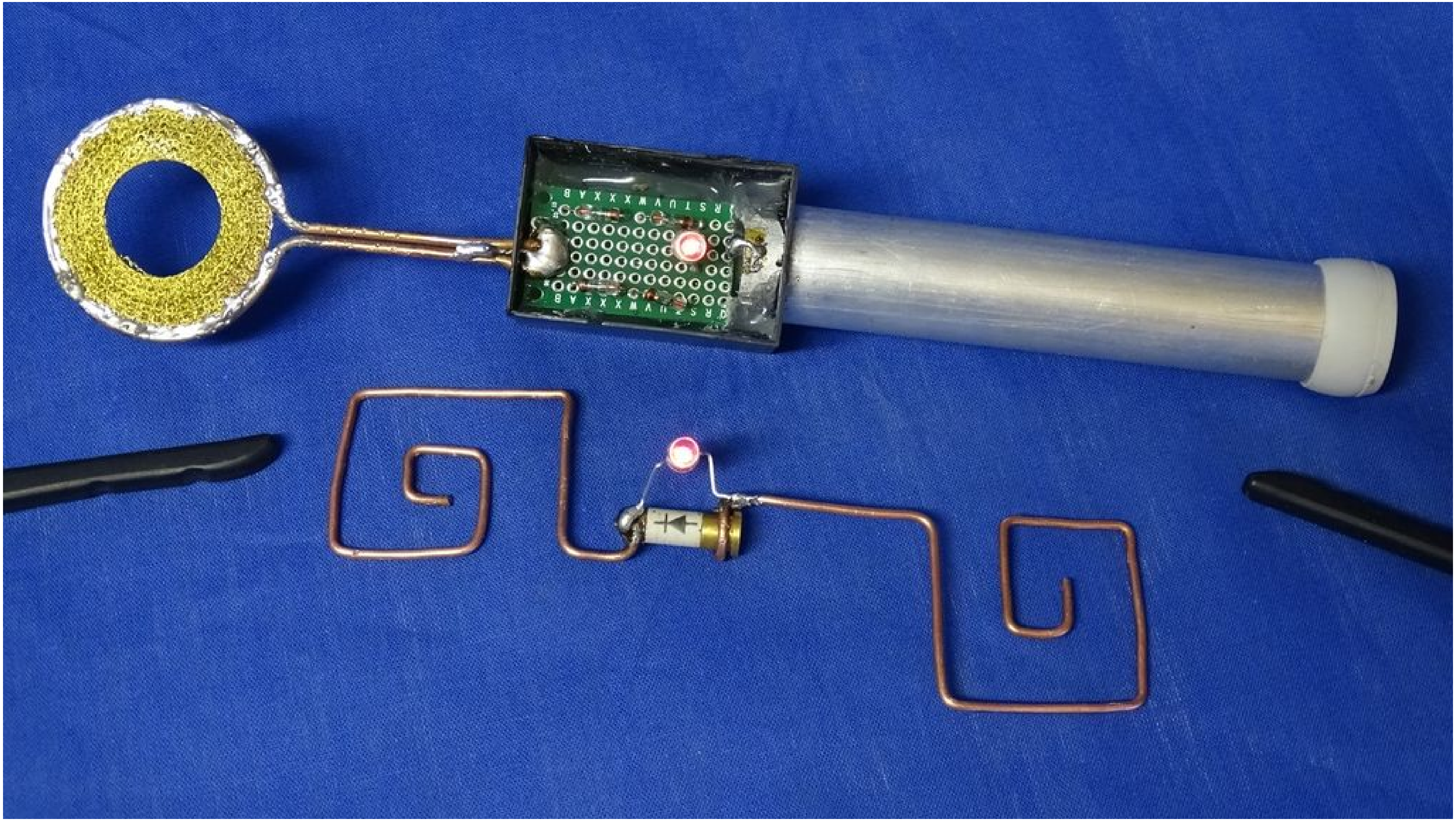

Step 1: LED+RF Diode

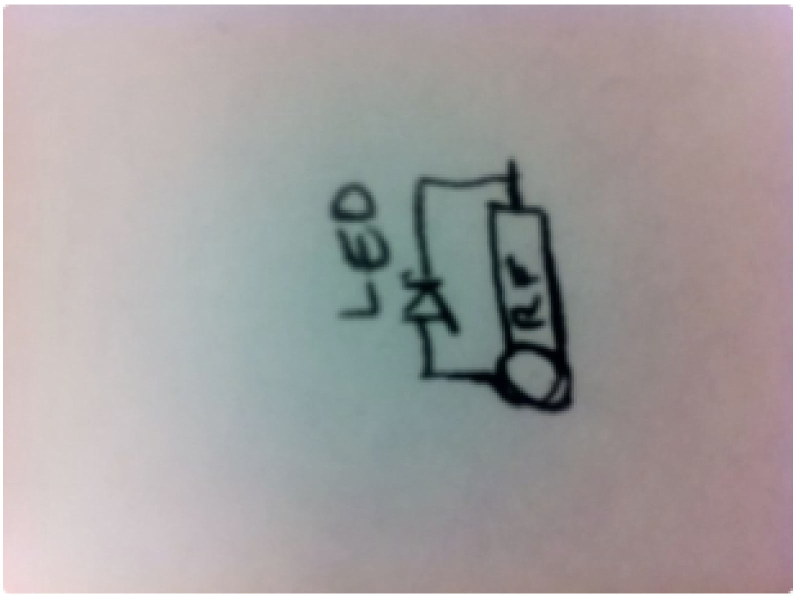

First Solder the led parallel to the Rf diode

Add Tip Ask Question Comment Download

Step 2: RF Diode+ LED+ Wires

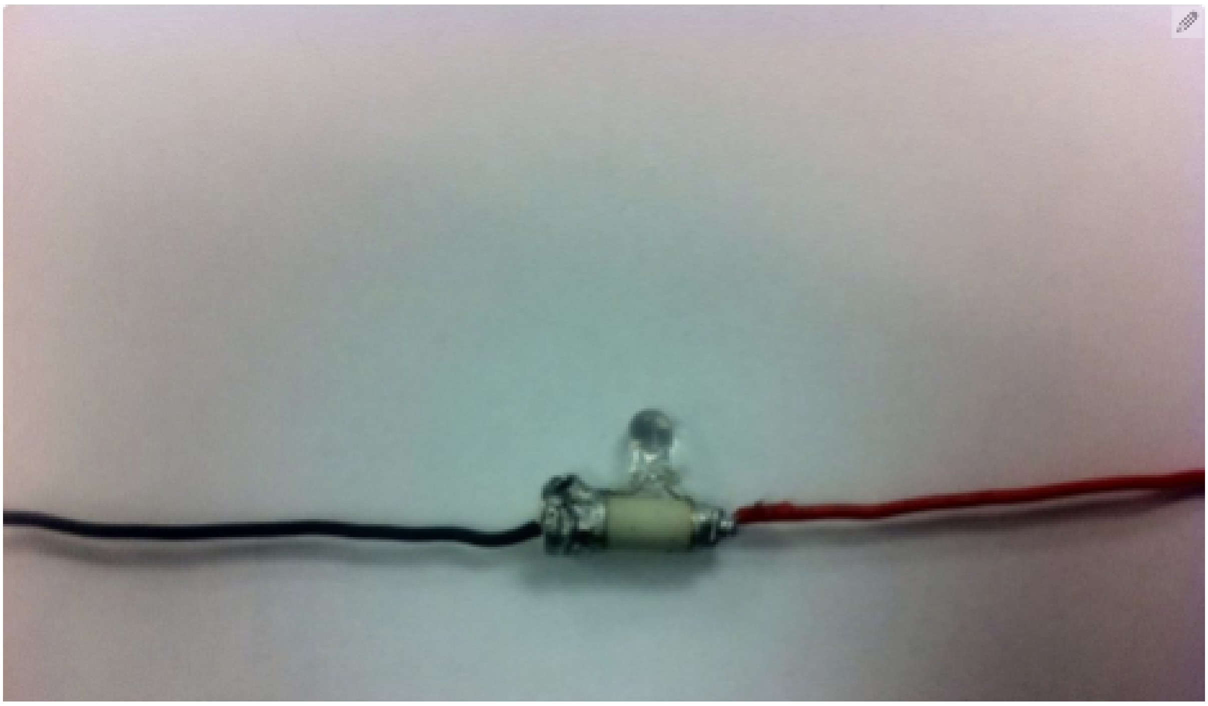

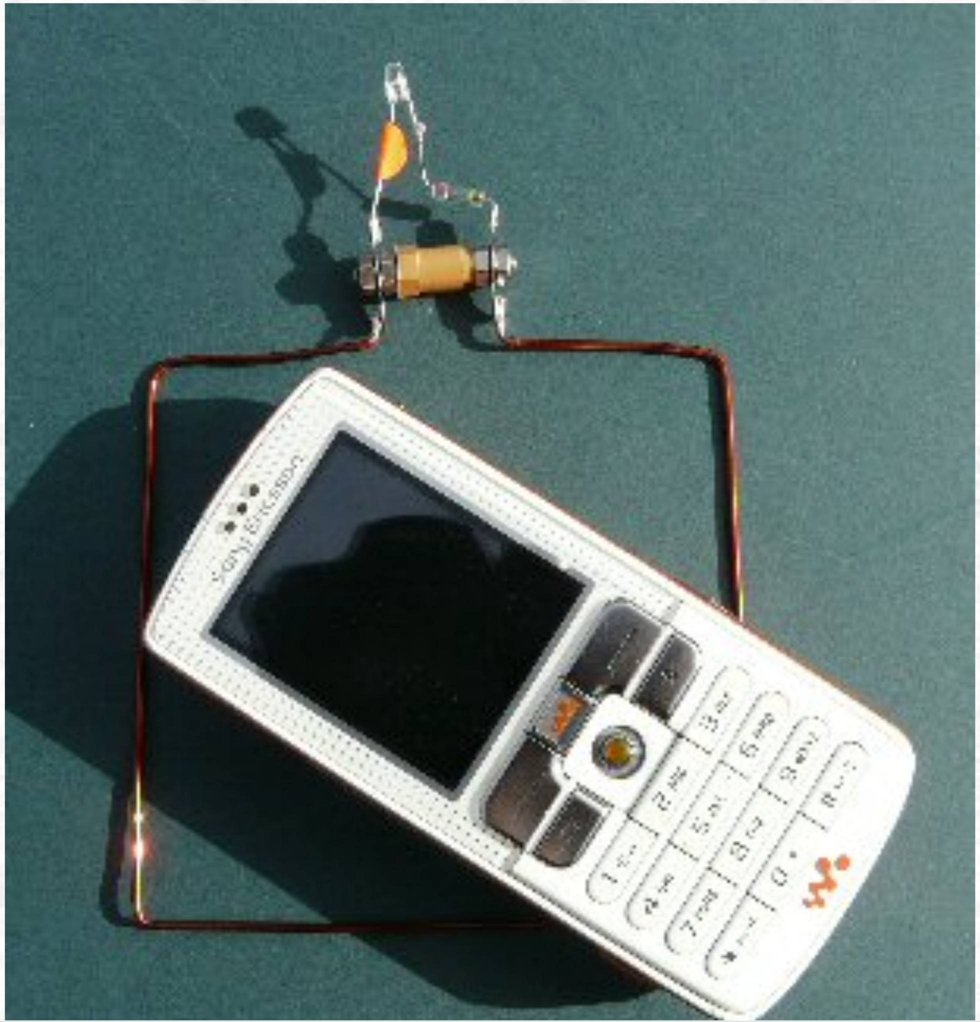

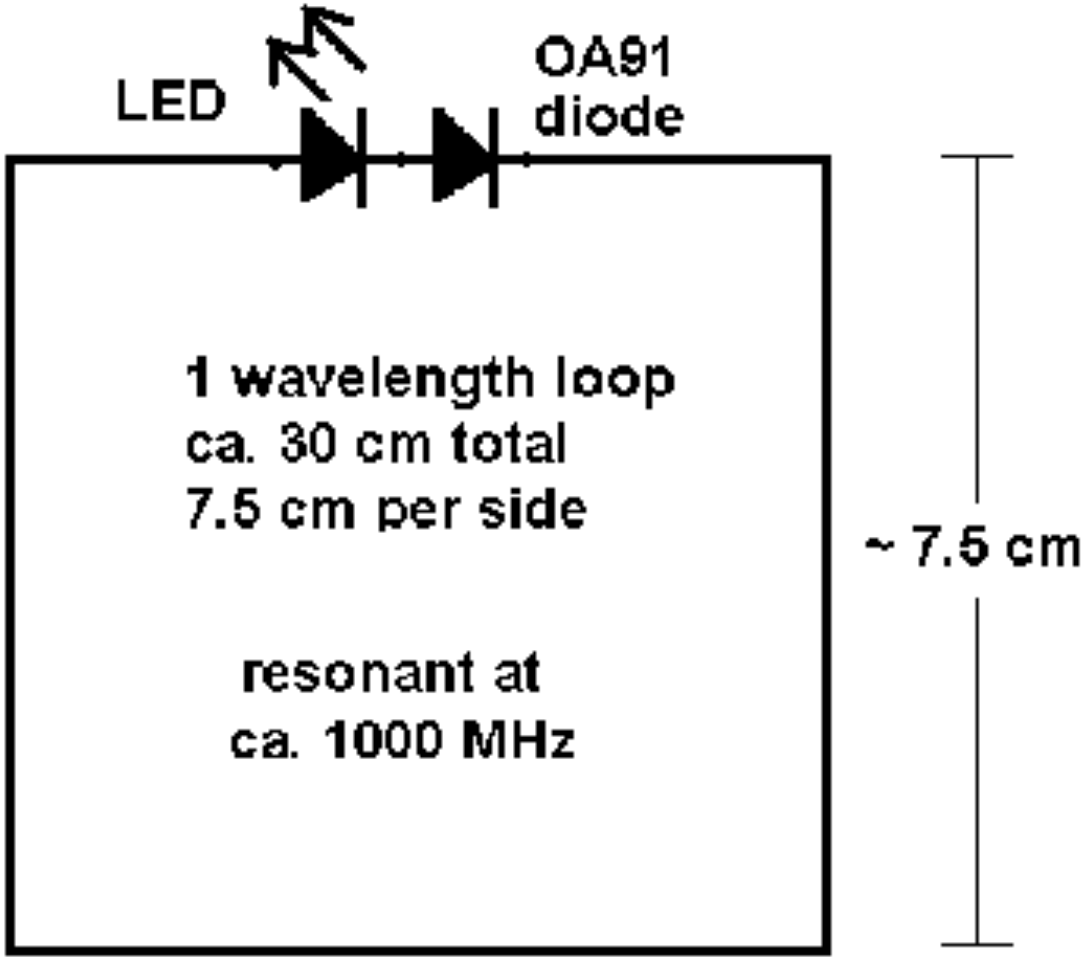
LED
OA91
diode
~ 7.5 cm

1 wavelength loop
ca. 30 cm total
7.5 cm per side

resonant at
ca. 1000 MHz

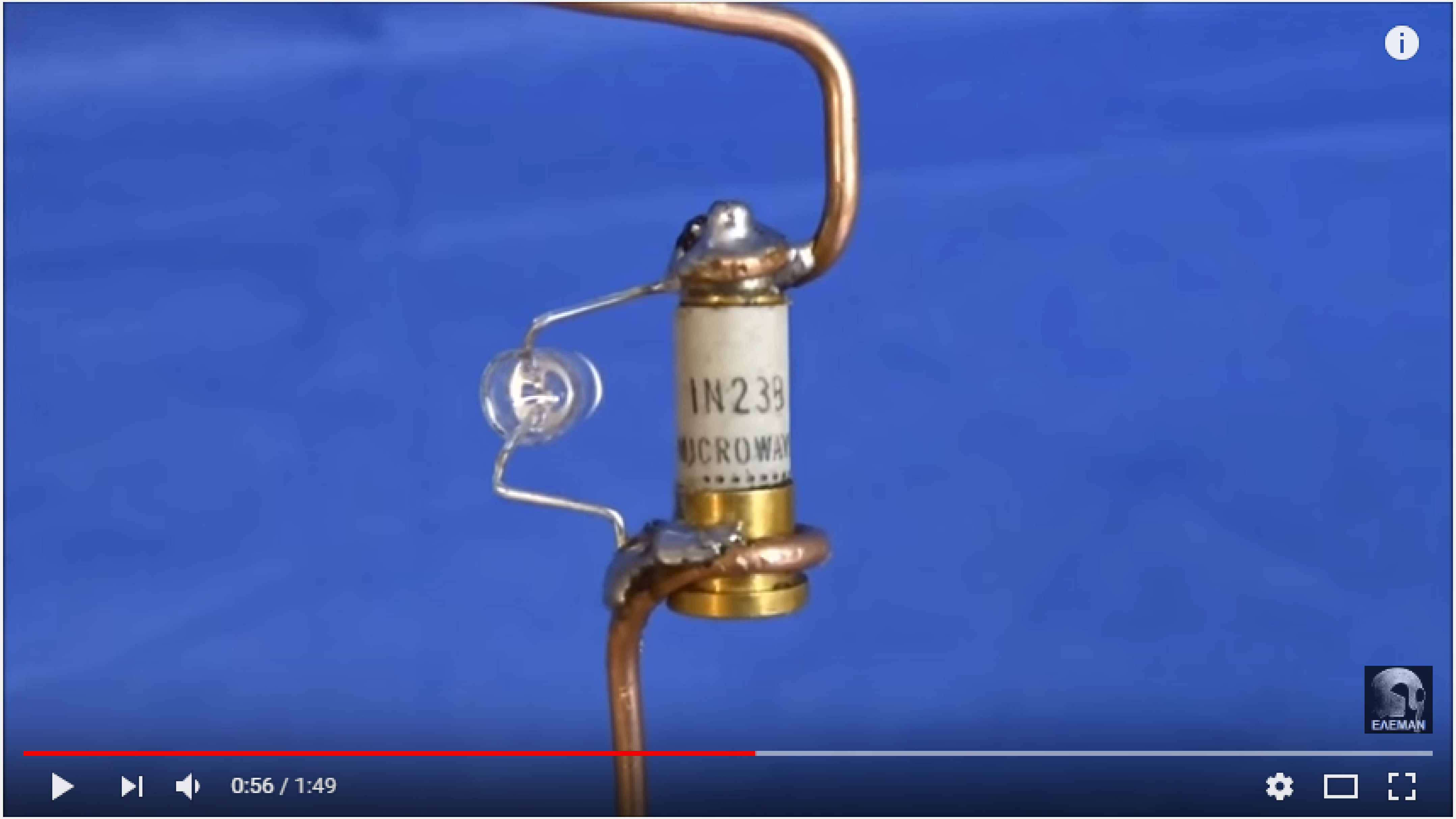
IN23B
MICROWA
ΕΛΕΜΑΝ
0:56 / 1:49

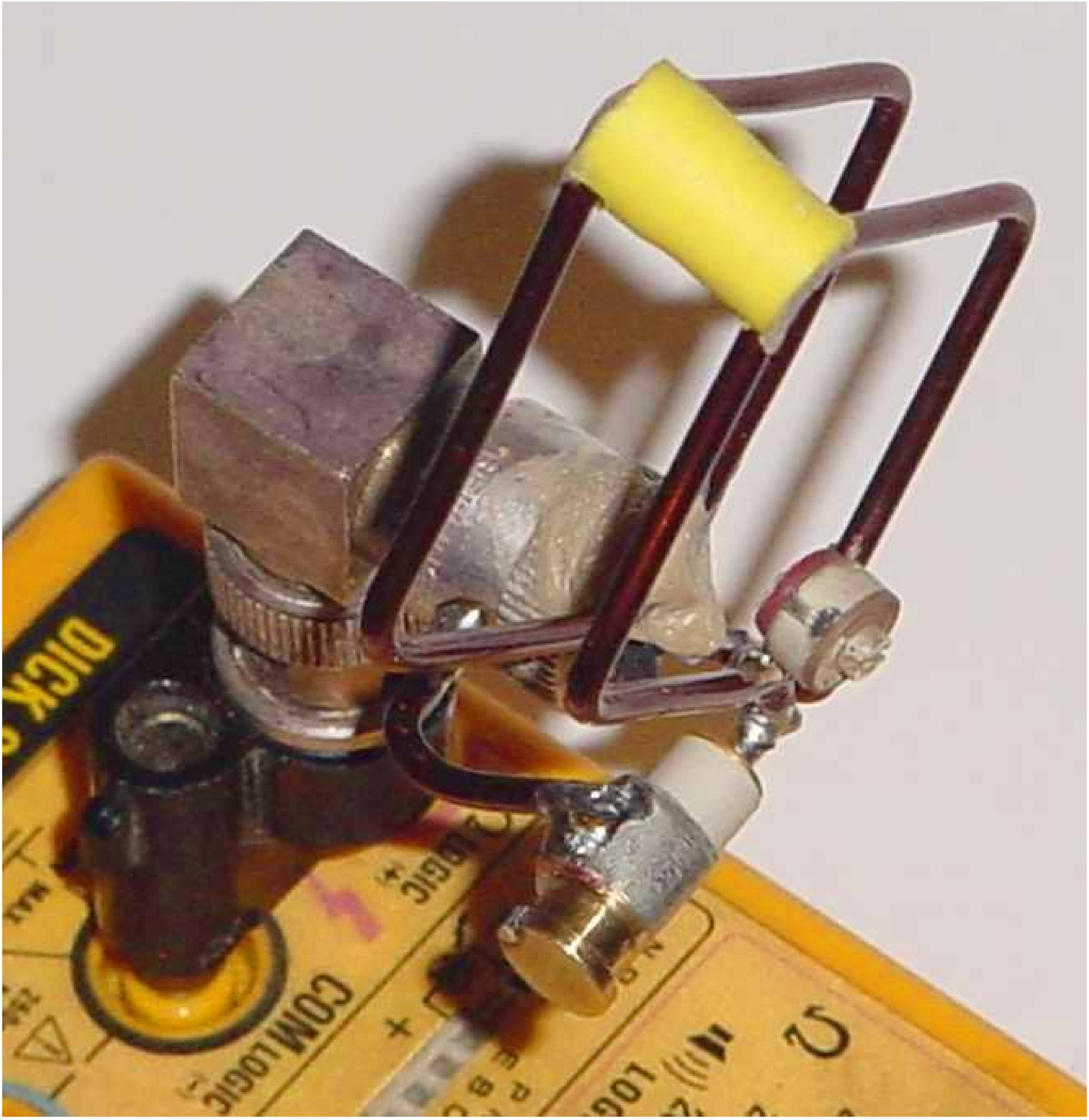

COM LOGIC

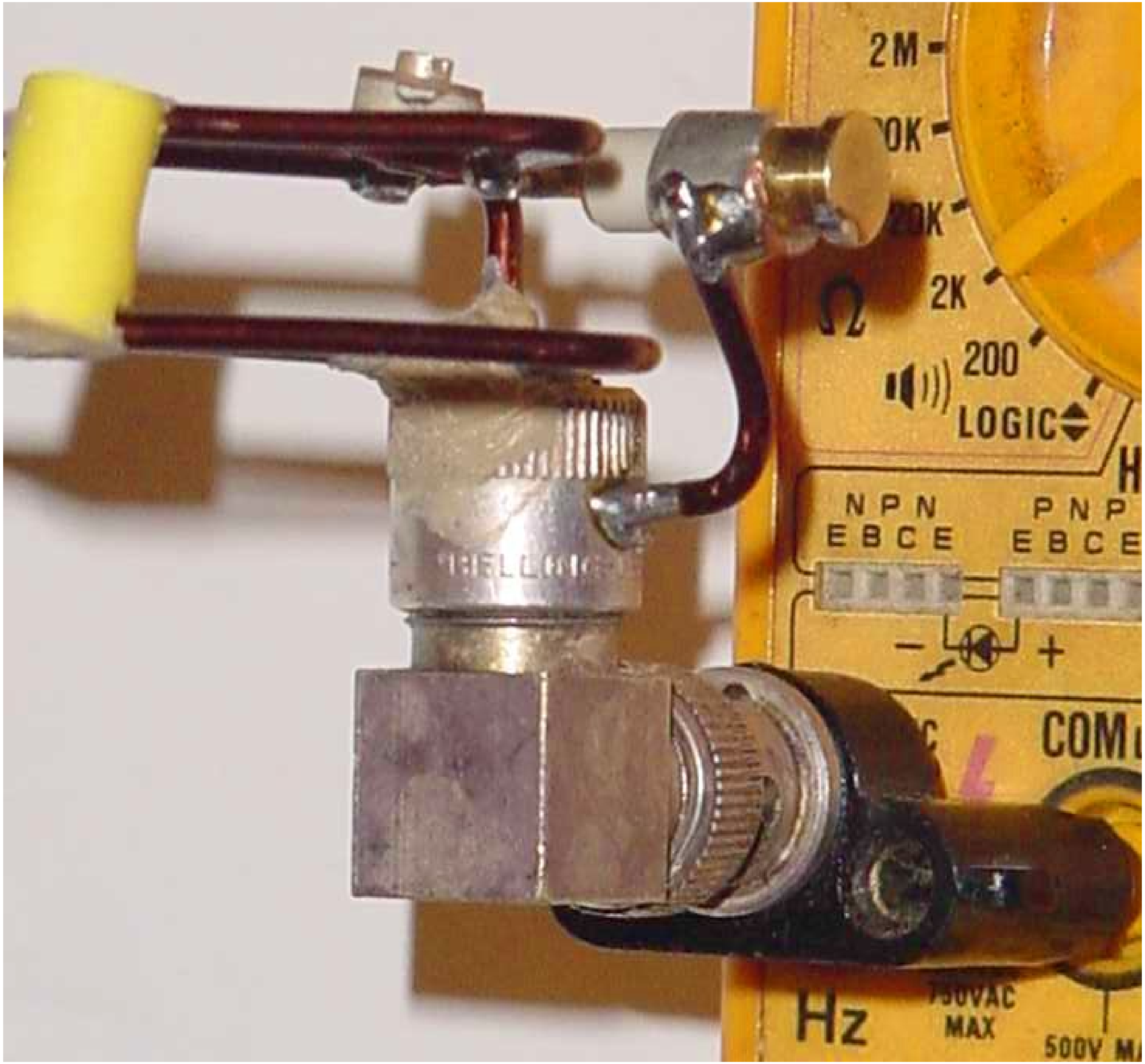
2M
2K
200
Ω
LOGIC
NPN
EBCE
PNP
EBCE
COM
Hz
MAX

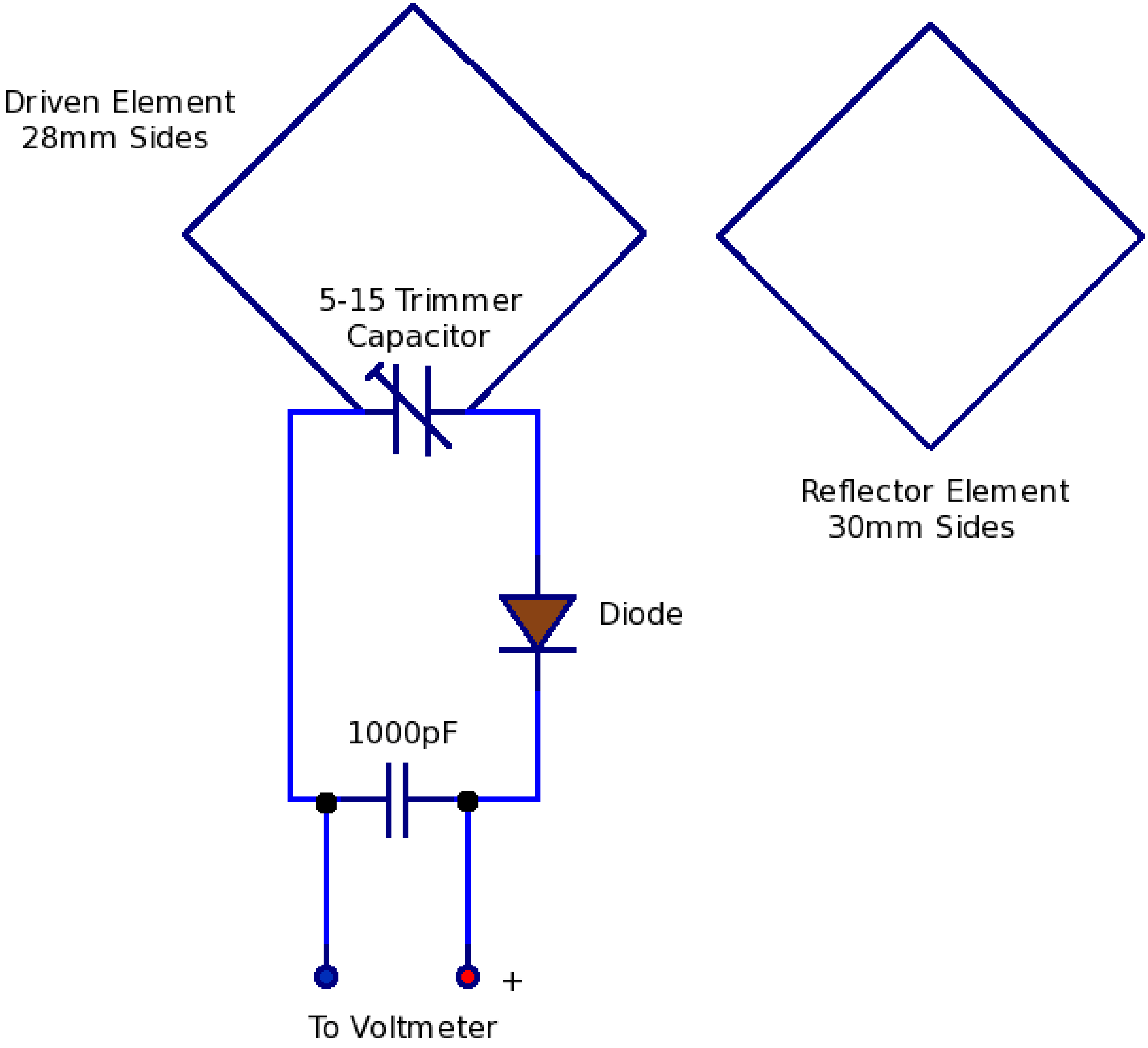
Driven Element
28mm Sides
5-15 Trimmer
Capacitor
Reflector Element
30mm Sides
Diode
1000pF
+
To Voltmeter

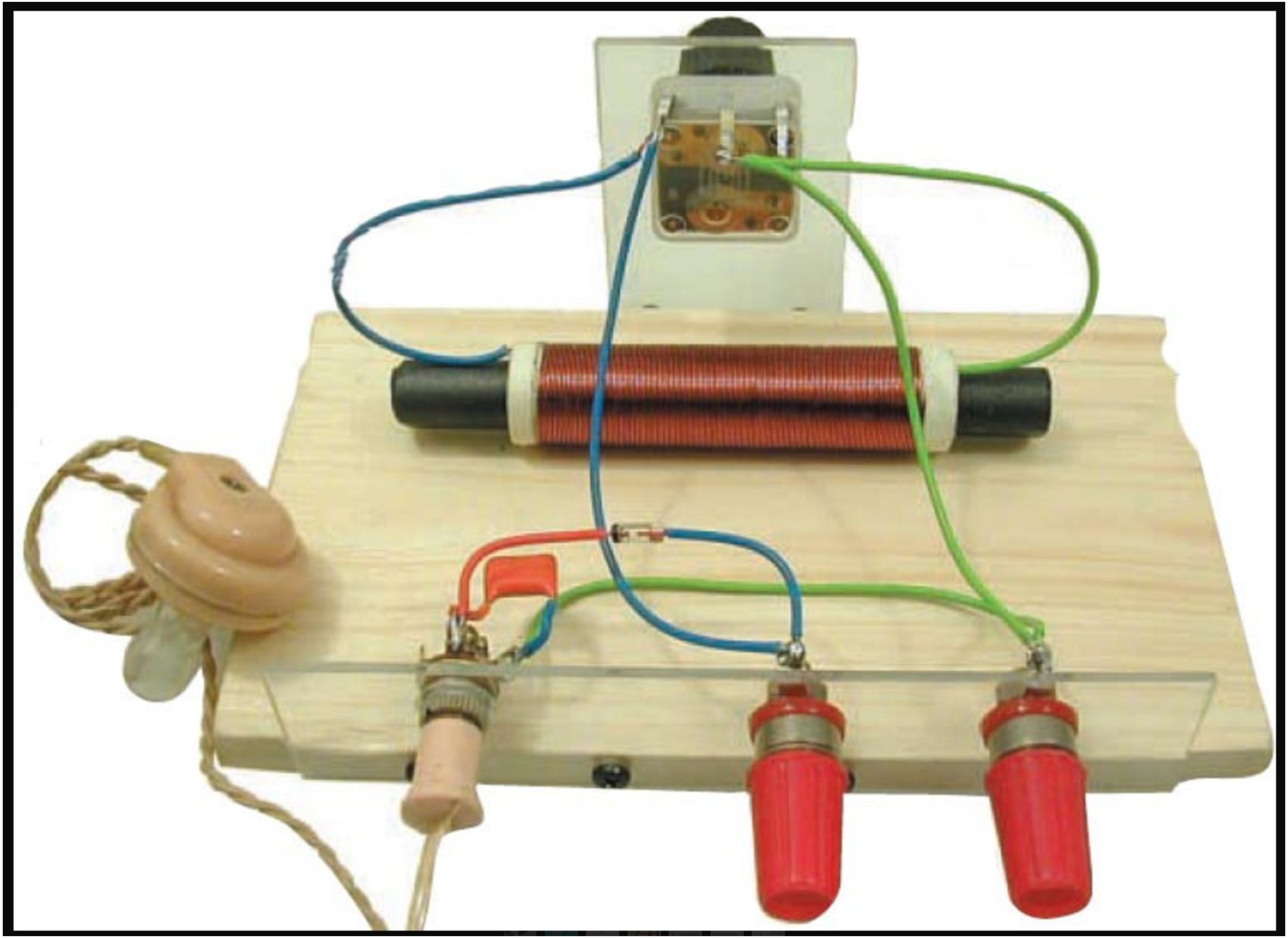

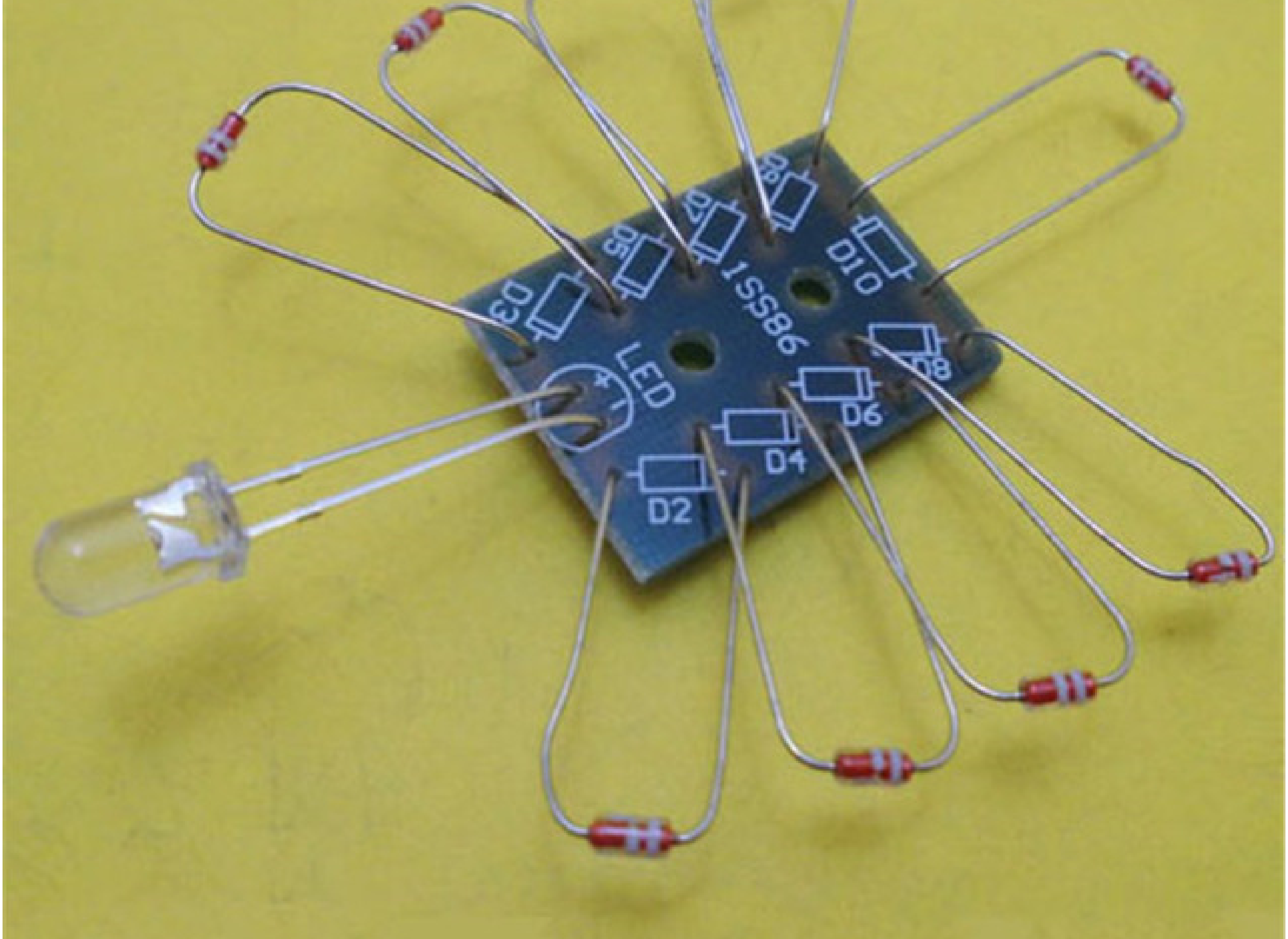
D3
D5
1SS86
D10
LED
D8
D6
D4
D2

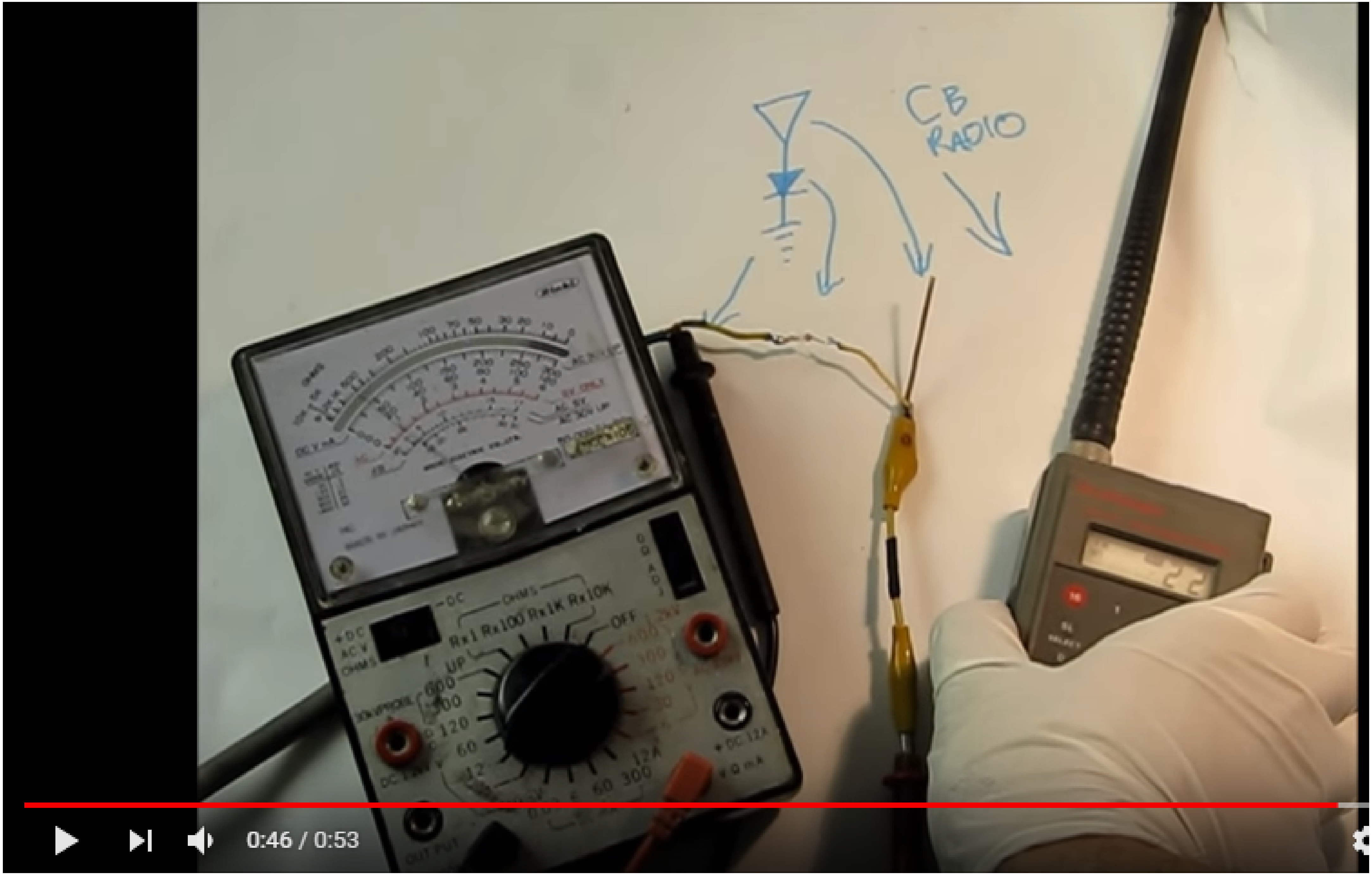
CB
RADIO
0:46 / 0:53

DIY Wireless Power - Part 2: Simple Wireless Power Transmission! (CB)

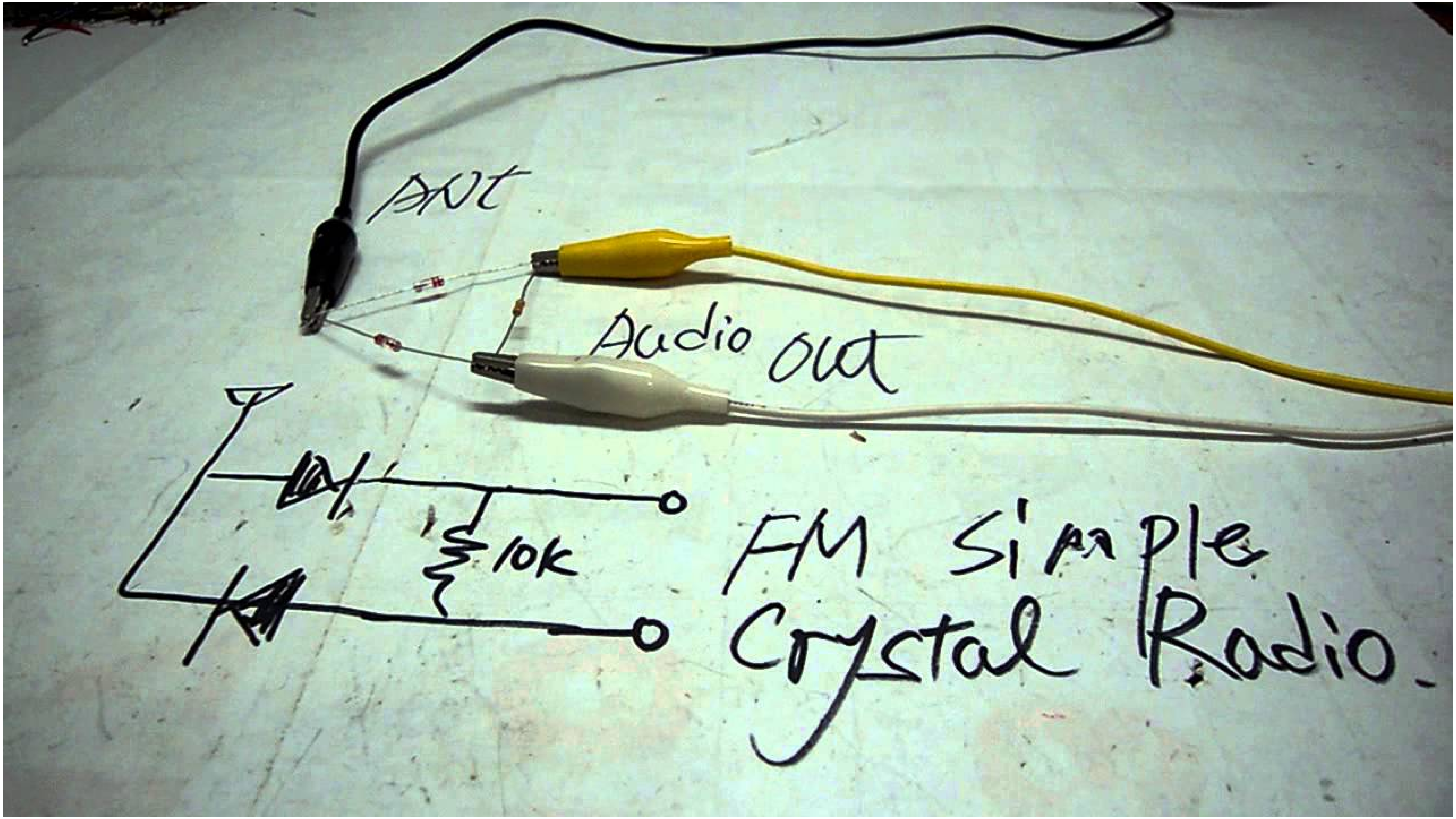
ANt
Audio out
10K
FM simple
Crystal Radio.

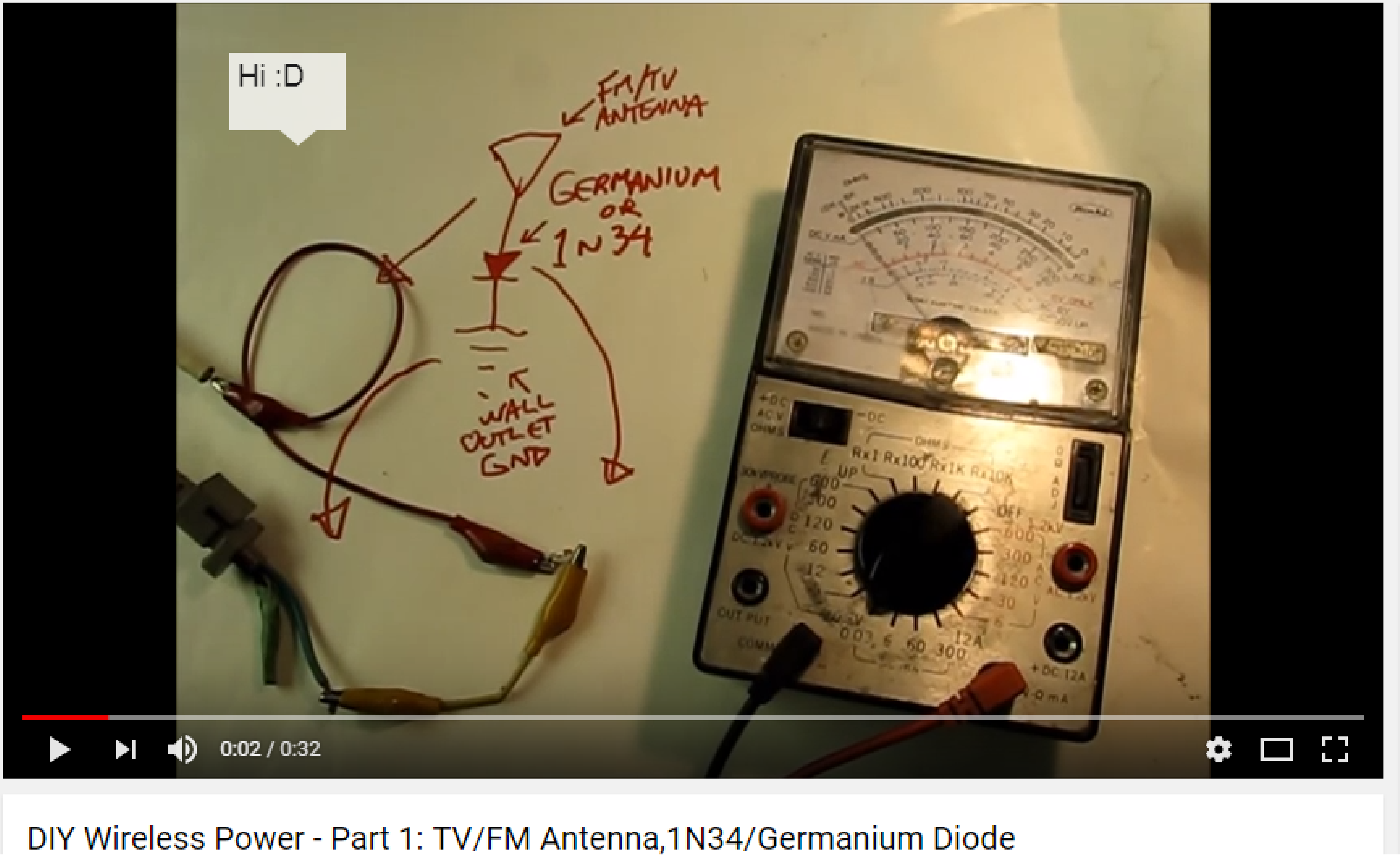
Hi :D
FM/TV ANTENNA
GERMANIUM OR 1N34
WALL OUTLET GND
0:02 / 0:32

DIY Wireless Power - Part 1: TV/FM Antenna,1N34/Germanium Diode

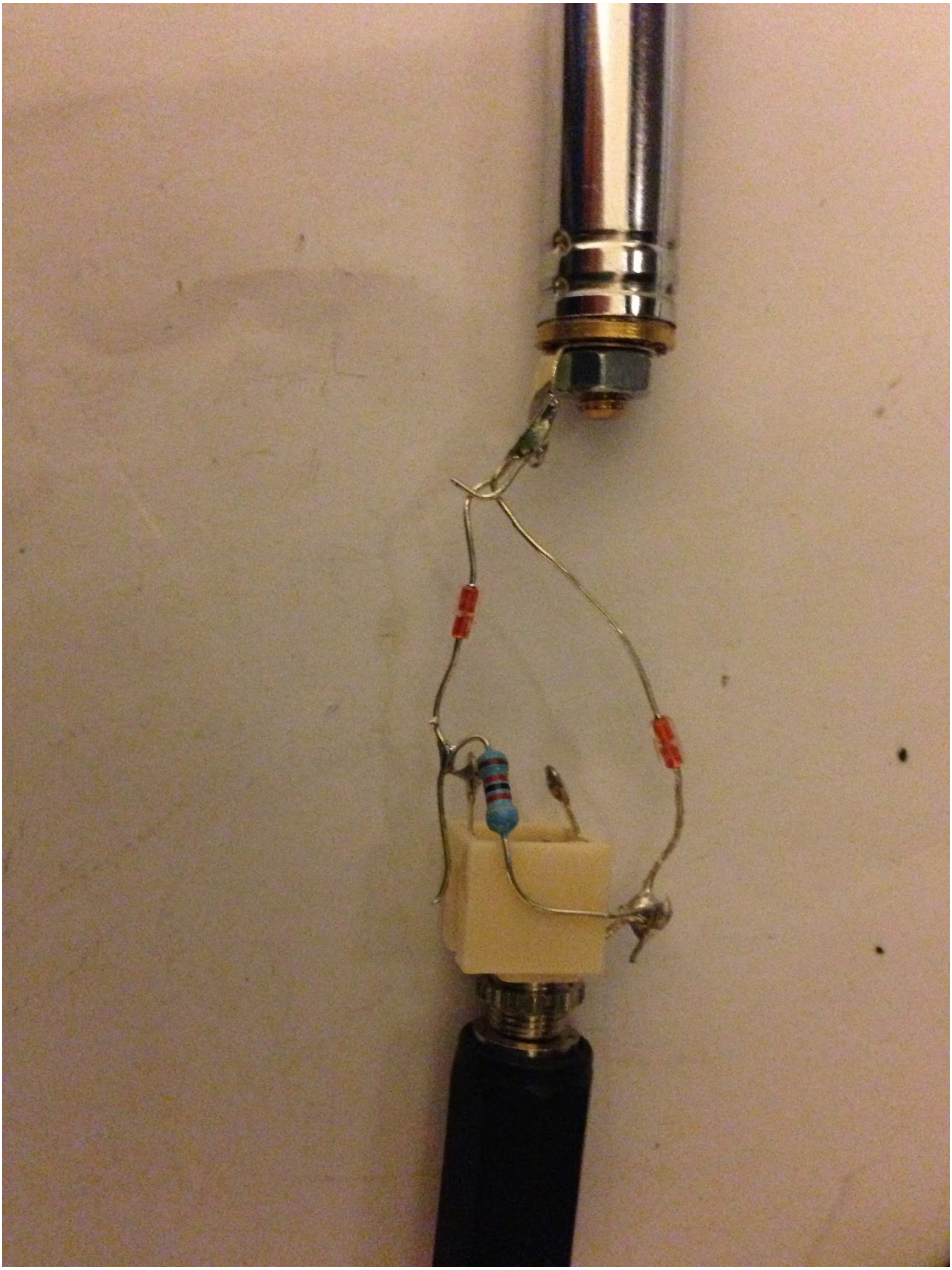

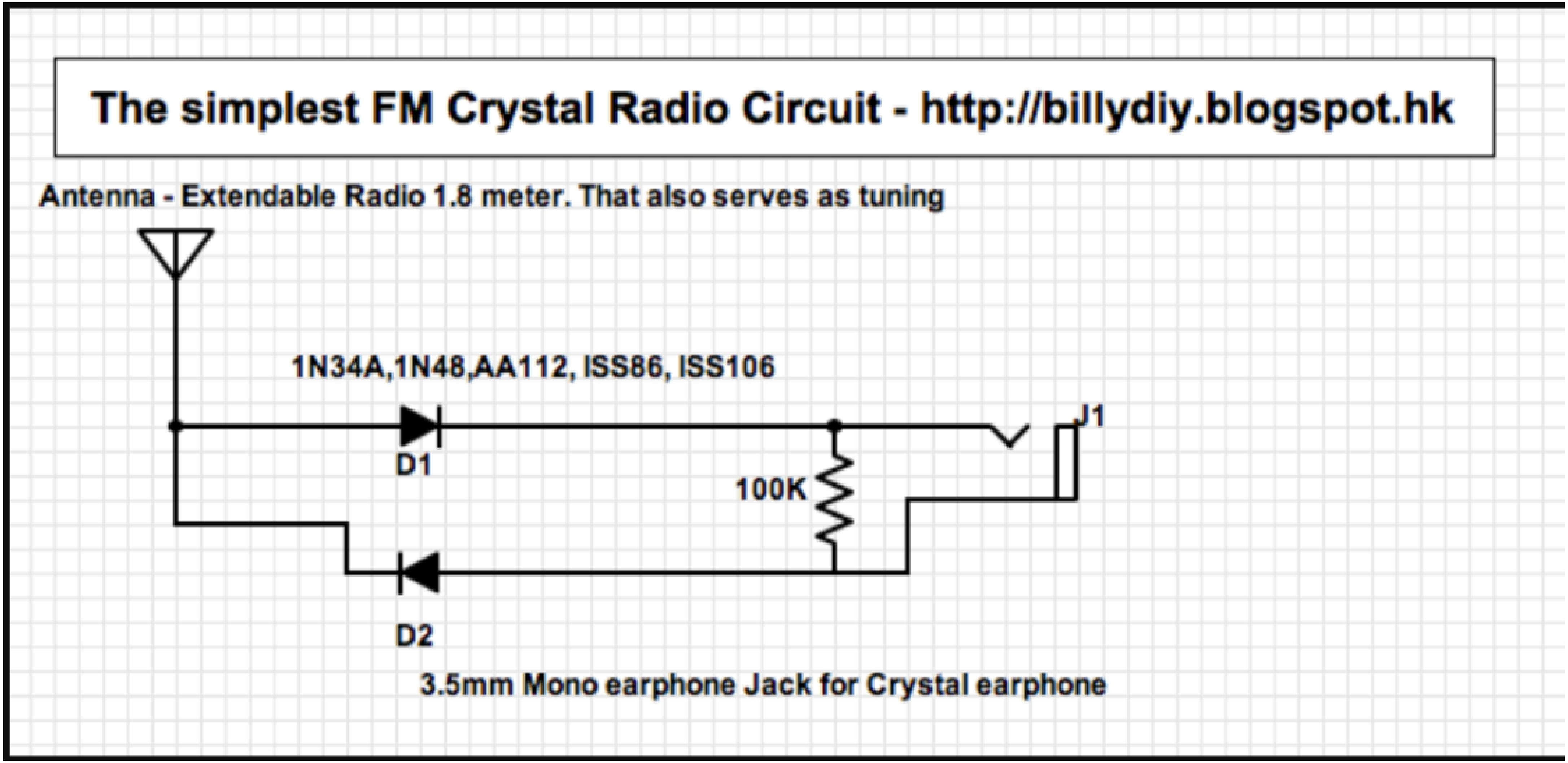

The simplest FM Crystal Radio Circuit - http://billydiy.blogspot.hk

Antenna - Extendable Radio 1.8 meter. That also serves as tuning

3.5mm Mono earphone Jack for Crystal earphone

1N34A,1N48,AA112, ISS86, ISS106
D1
100K
J1
D2

Simple crystal receiver for FM

Carlo Bramanti

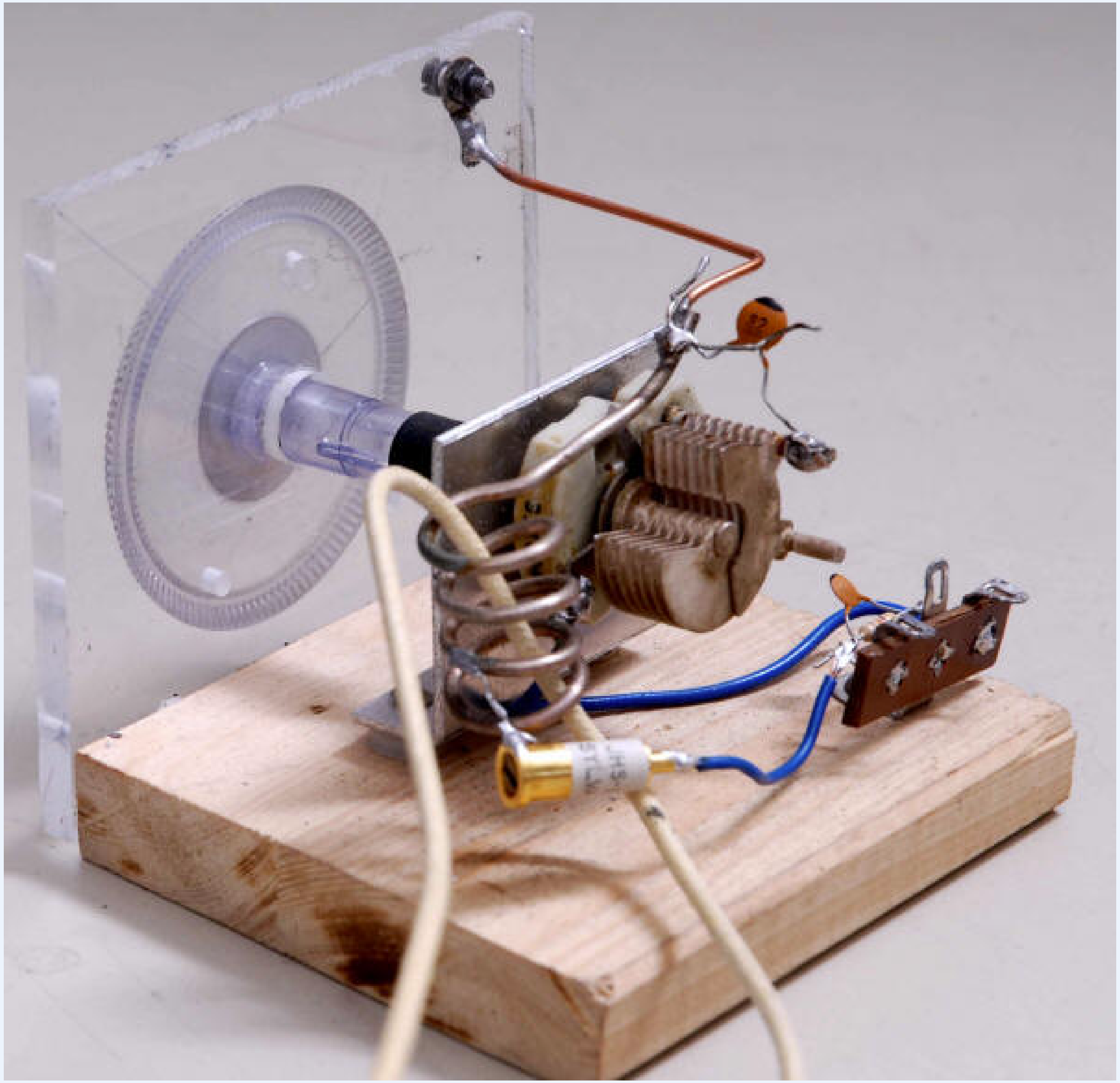

FM Crystal Radio Circuit

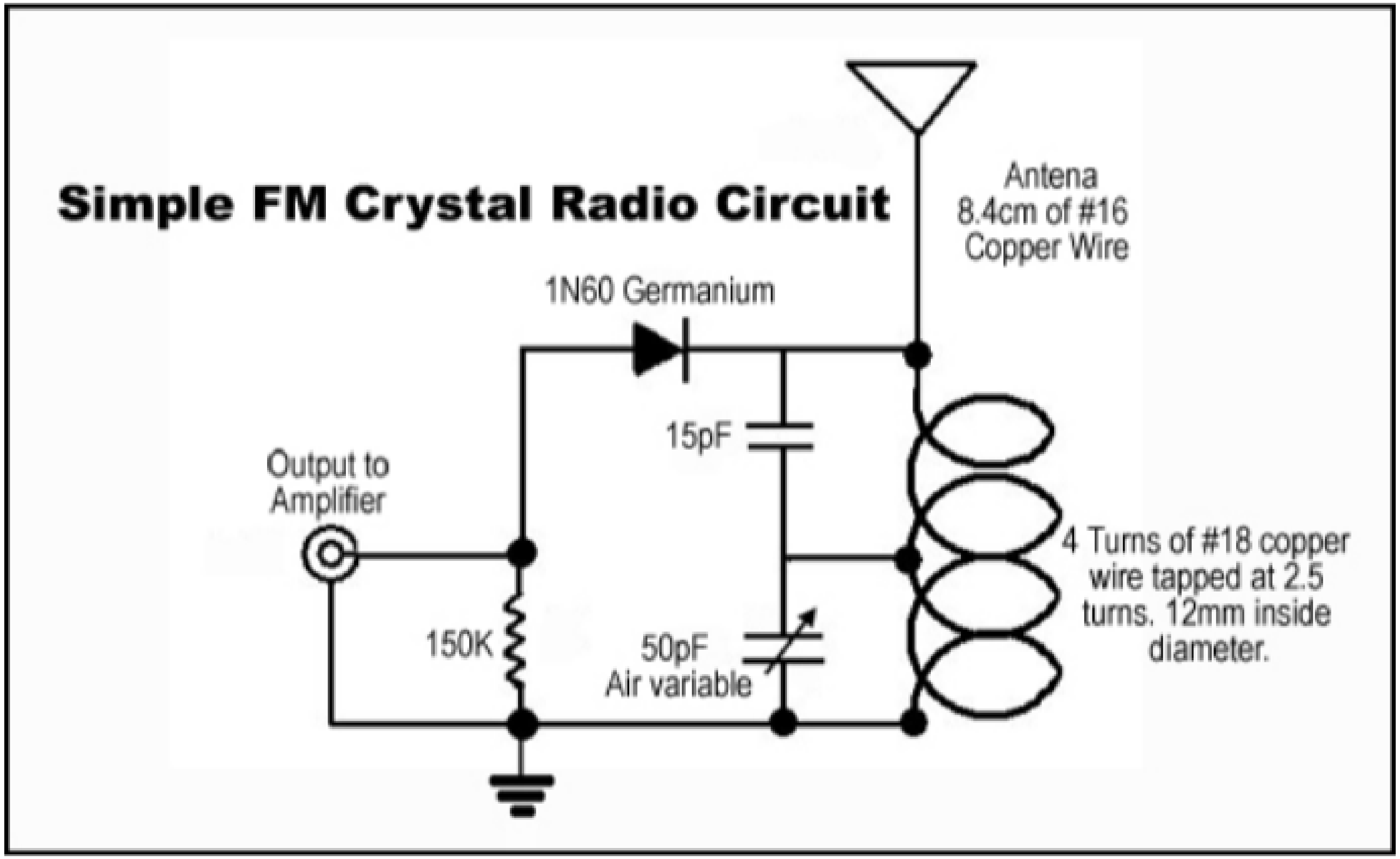
Simple FM Crystal Radio Circuit
Antena
8.4cm of #16
Copper Wire
1N60 Germanium
15pF
Output to
Amplifier
4 Turns of #18 copper
wire tapped at 2.5
turns. 12mm inside
diameter.
150K
50pF
Air variable

Parts List (some of these parts you can buy from our online store):

- 1N60 Germanium Diode
- 15pF Ceramic Capacitor
- 50pF Variable Capacitor
- 150K Ohm Resistor
- #16 & #18 Copper wires

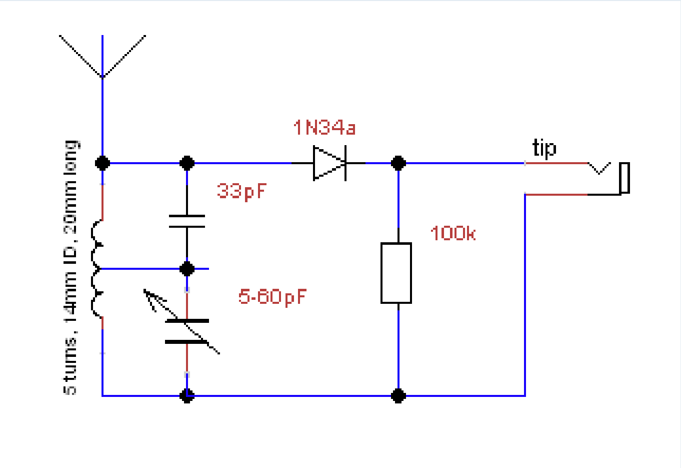
1N34a
tip
33pF
100k
5-60pF
5 turns, 14mm ID, 20mm long

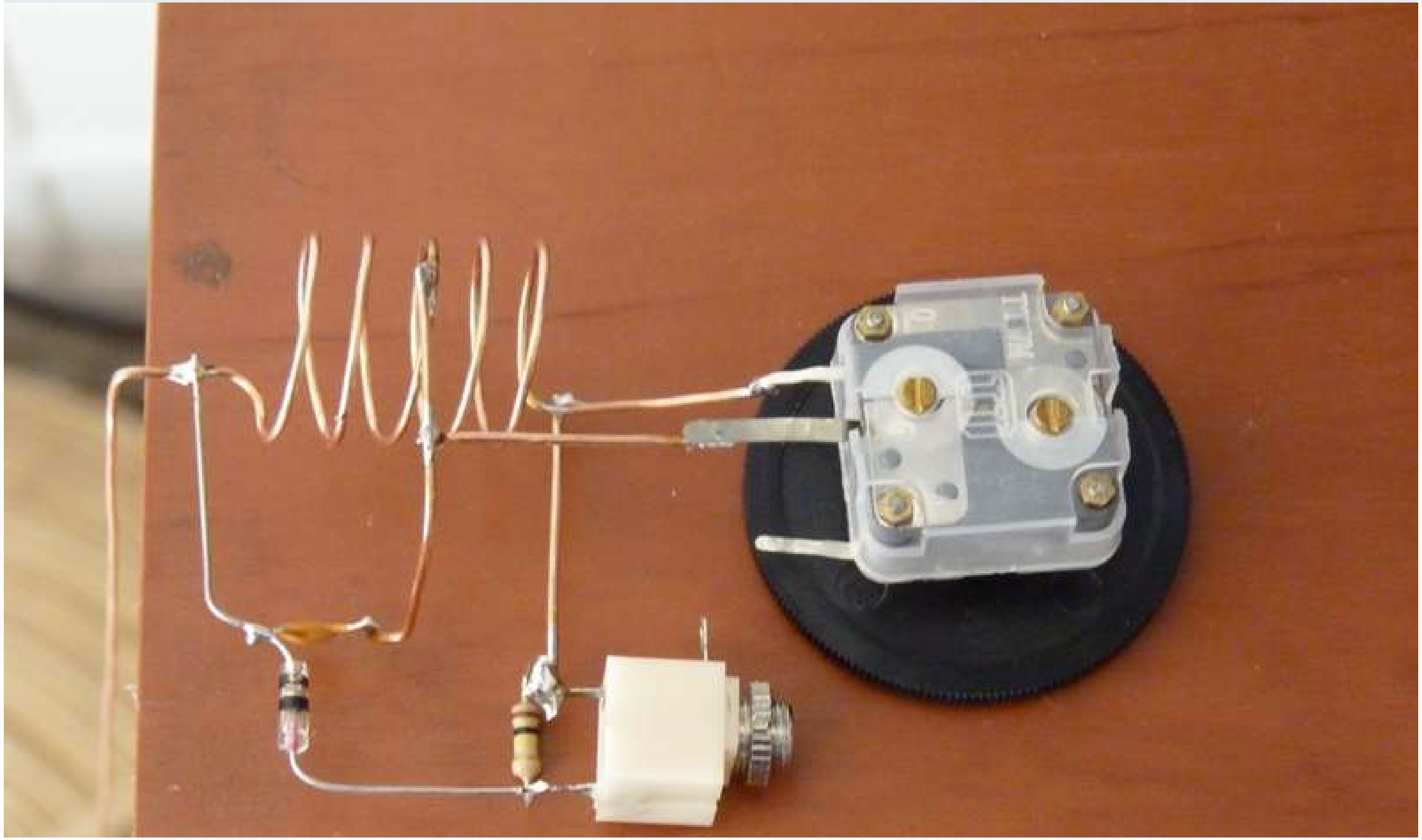

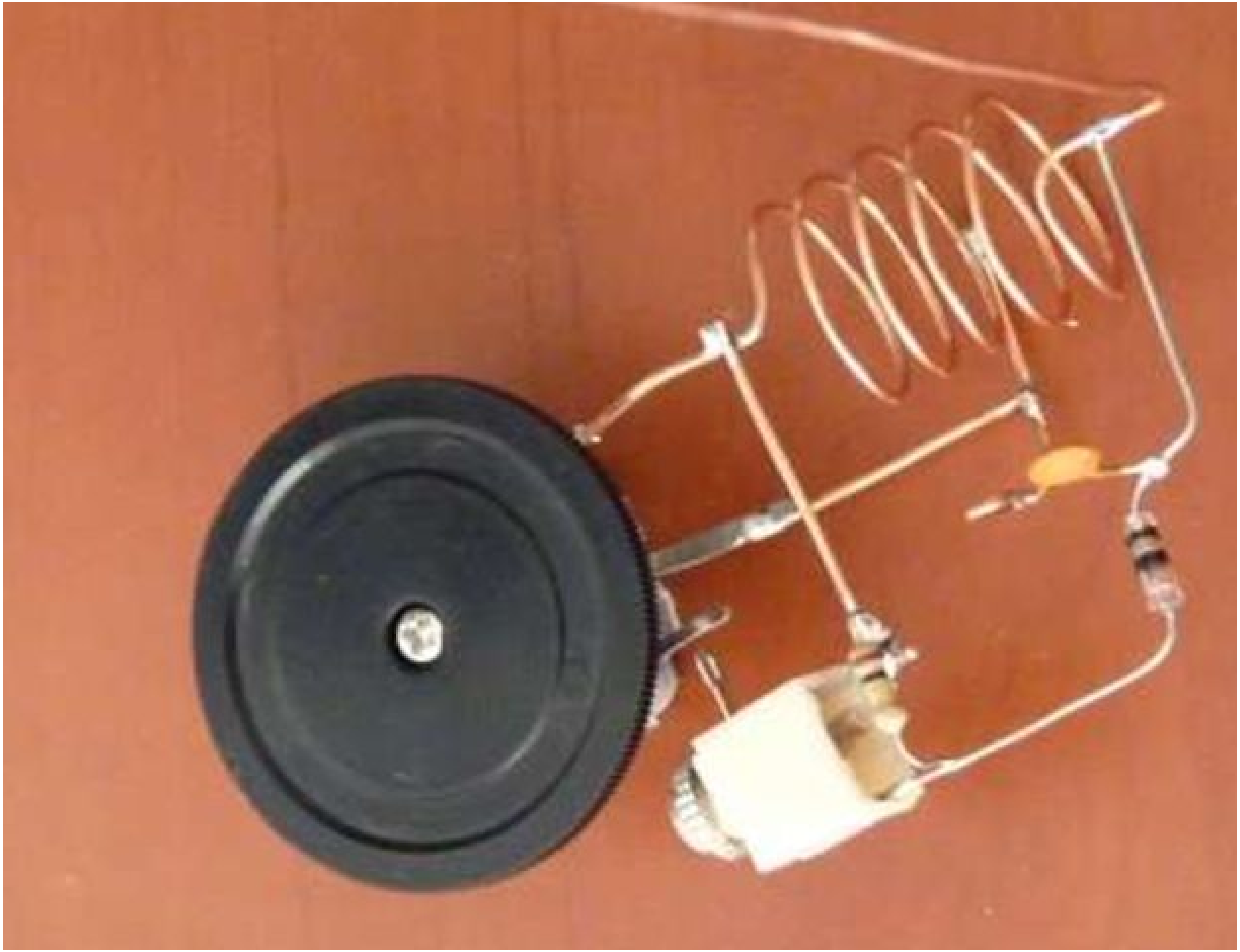

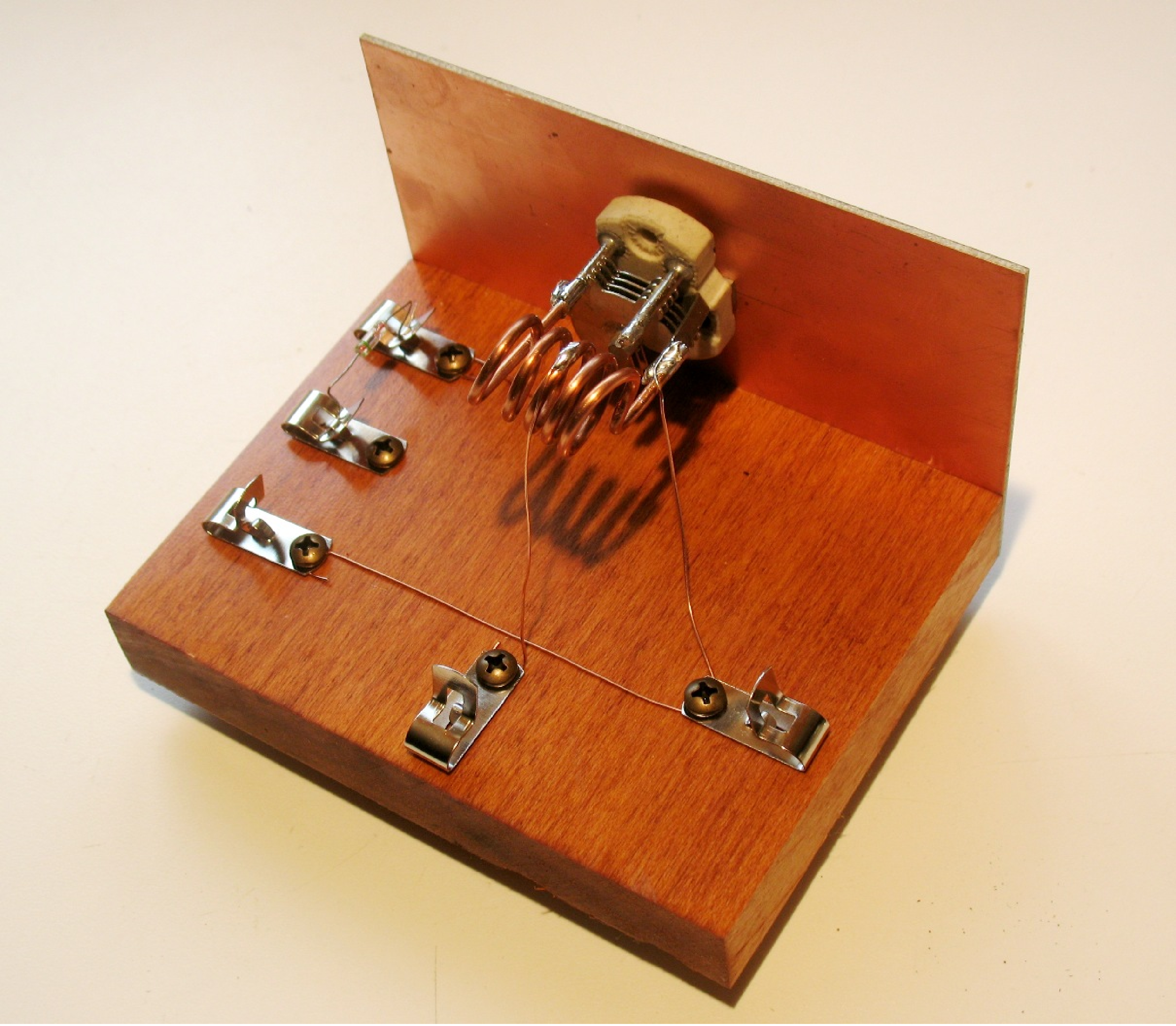

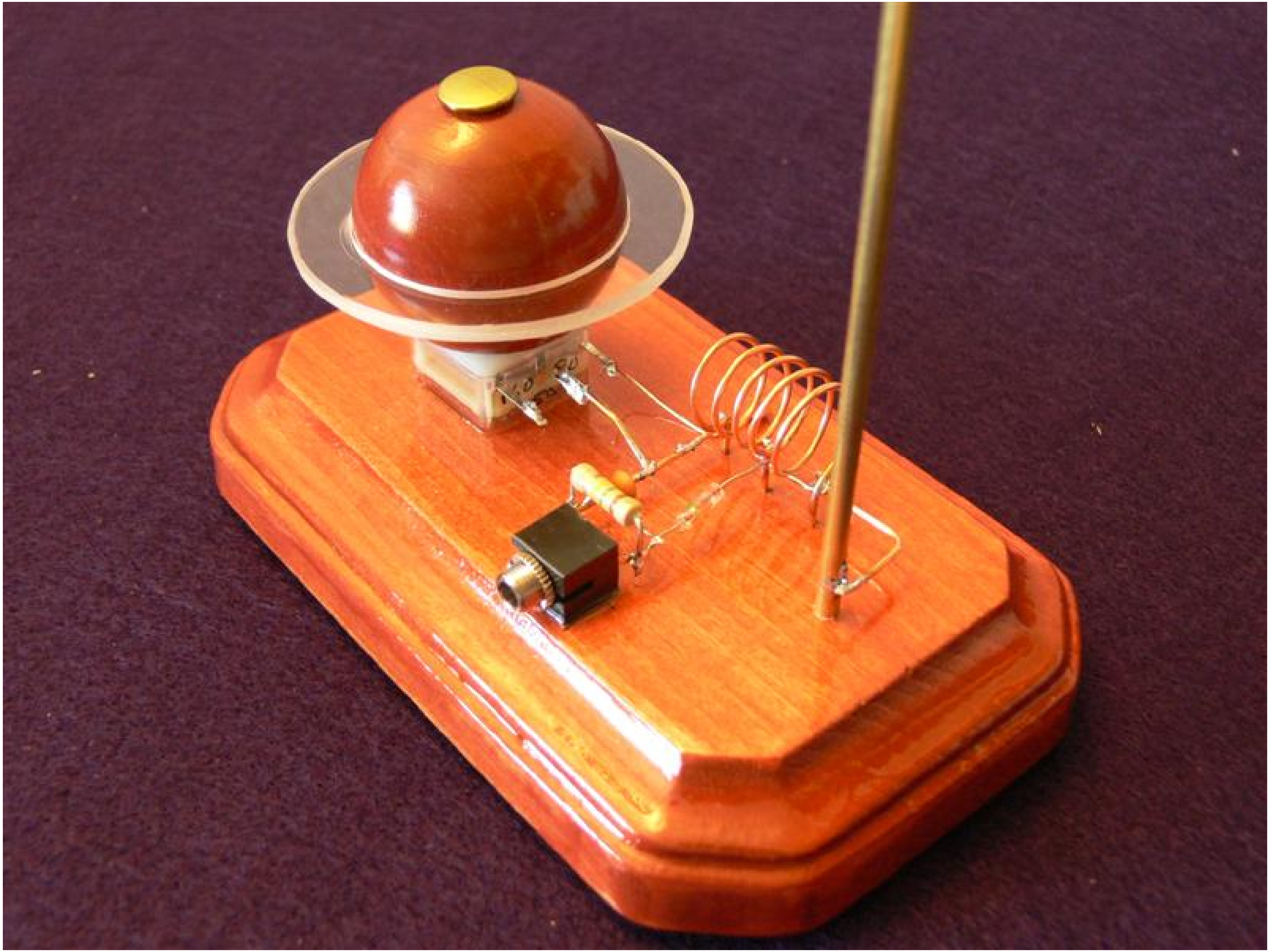

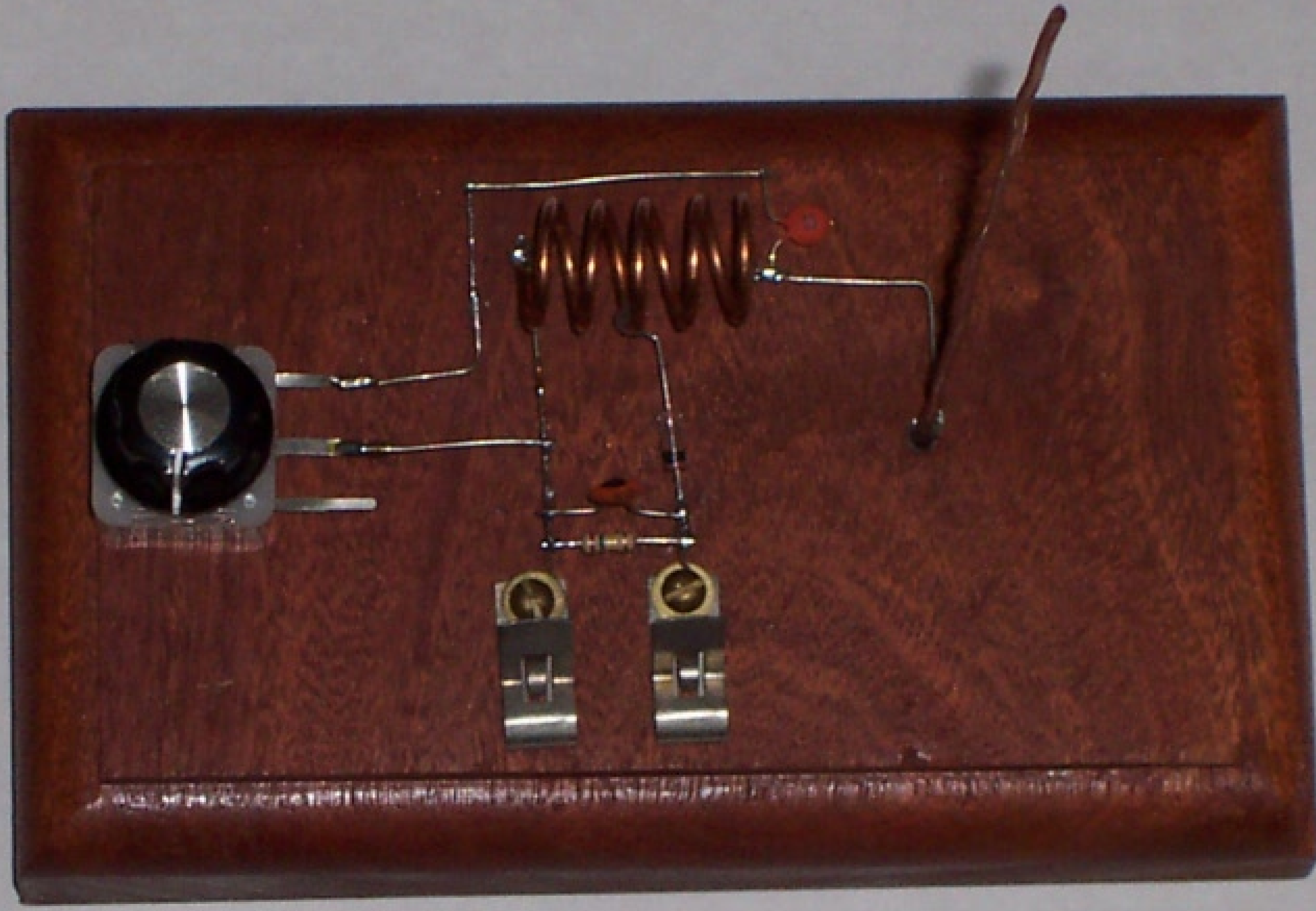

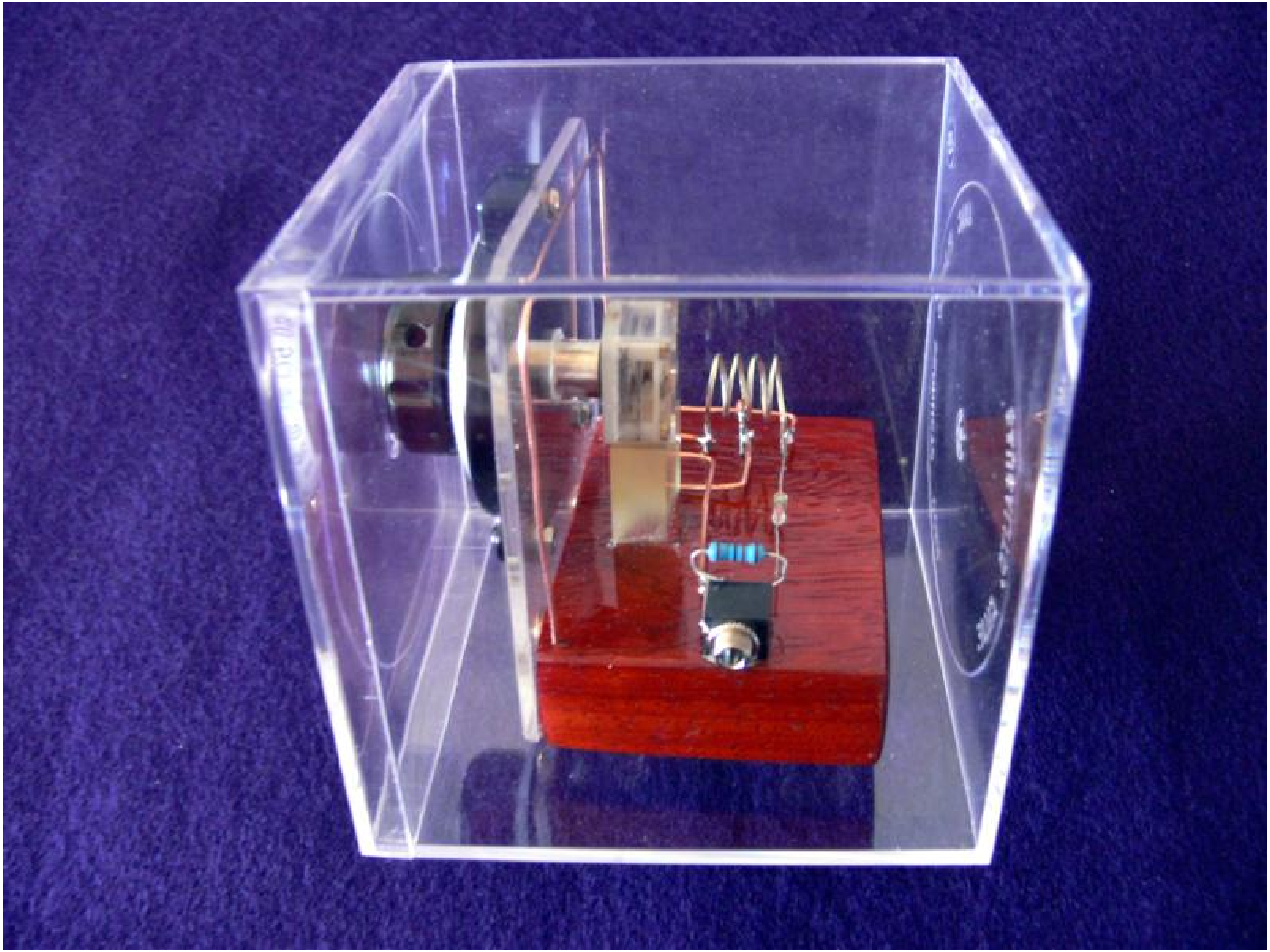

L - 4 turns #18 copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
Ant - 7 inches of #18 bare copper wire
C1 - 18 pf ceramic capacitor
C2 - 50 pf air variable capacitor
D - 1N34 diode or rock crystal
R - 150K resistor

All passive components

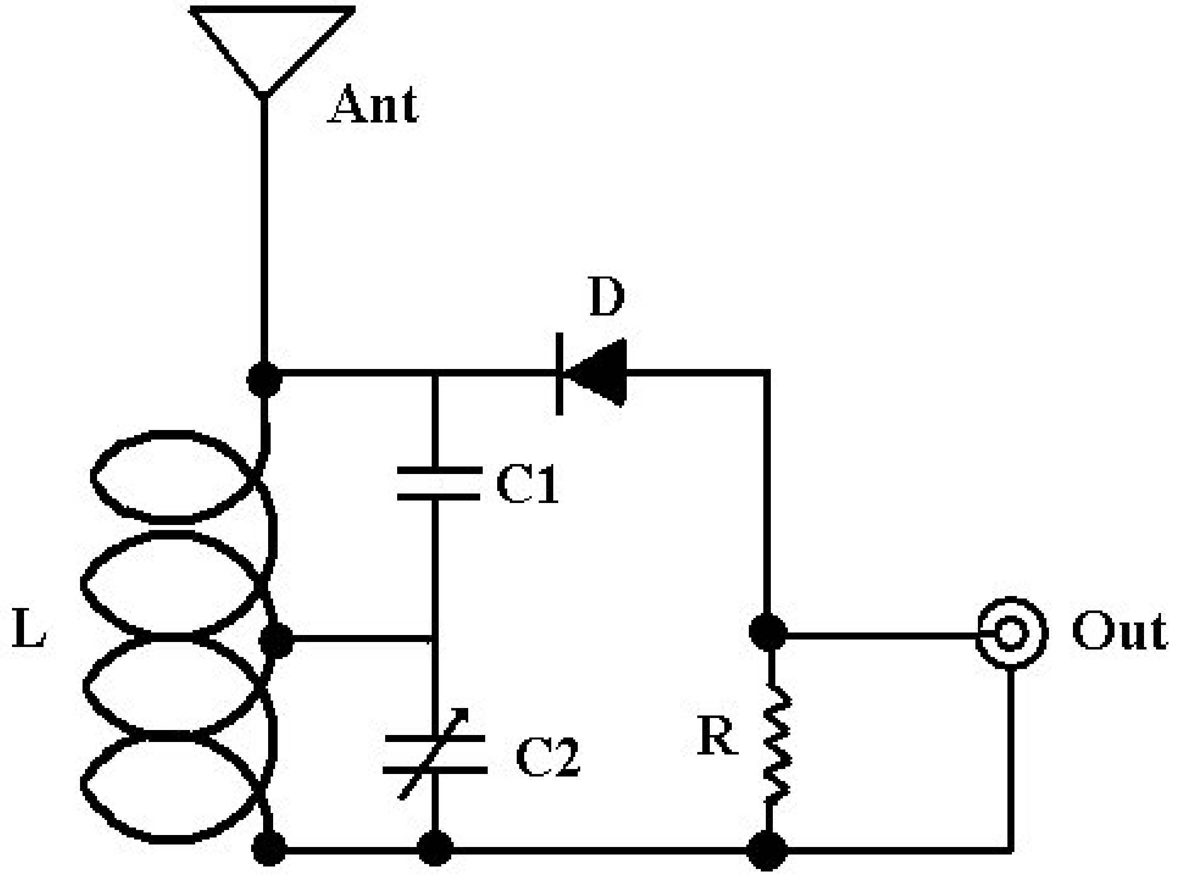

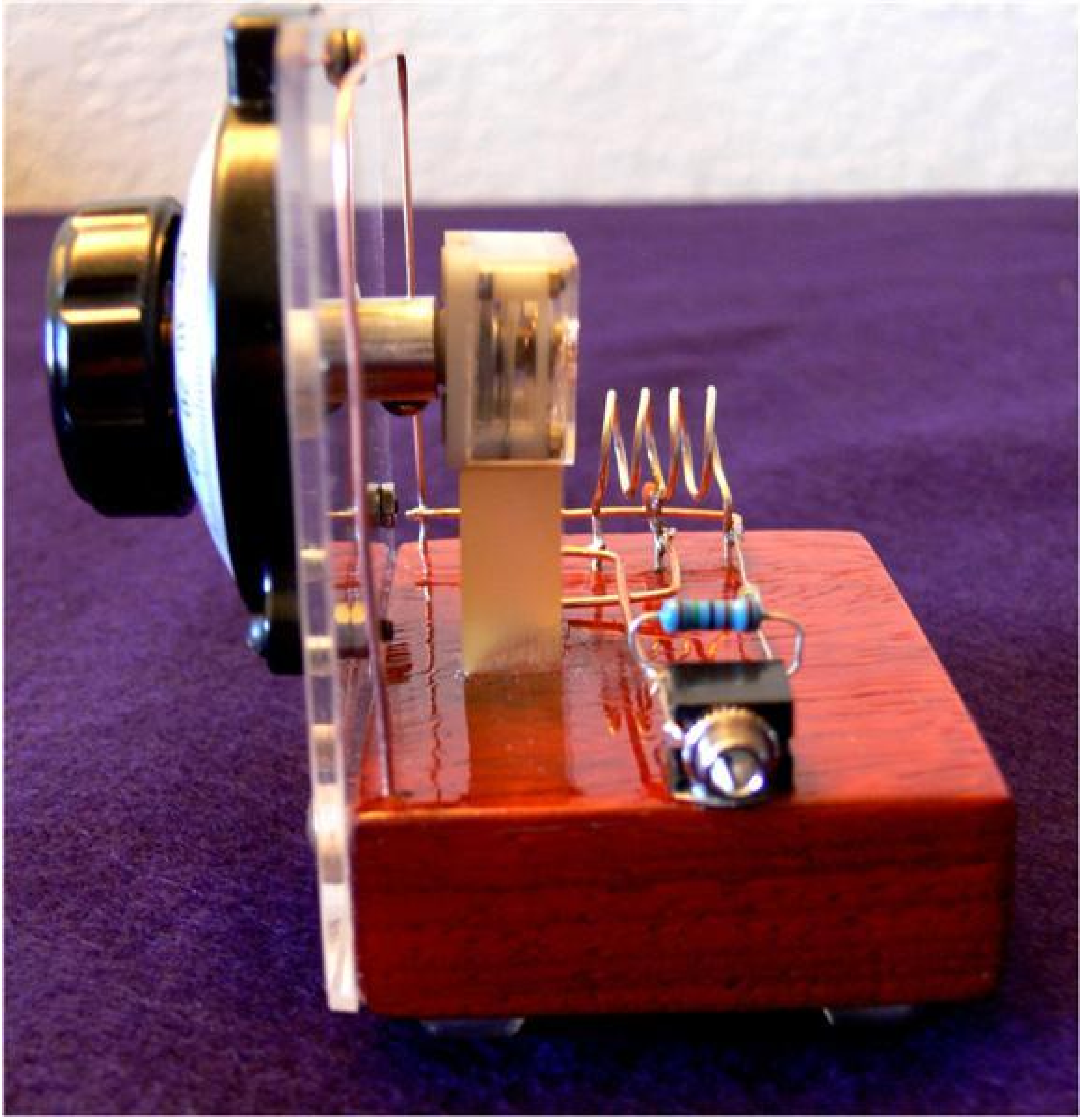

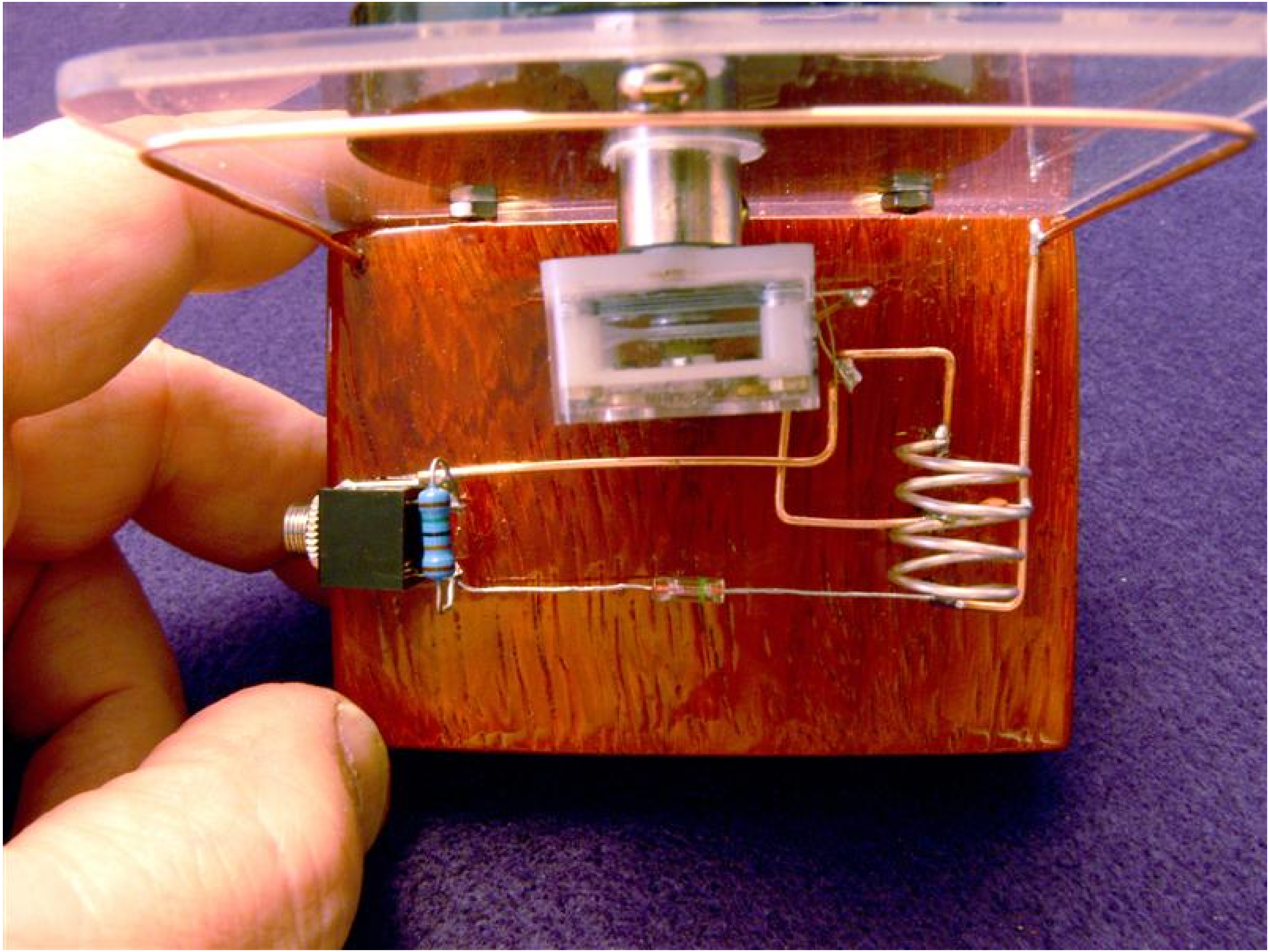

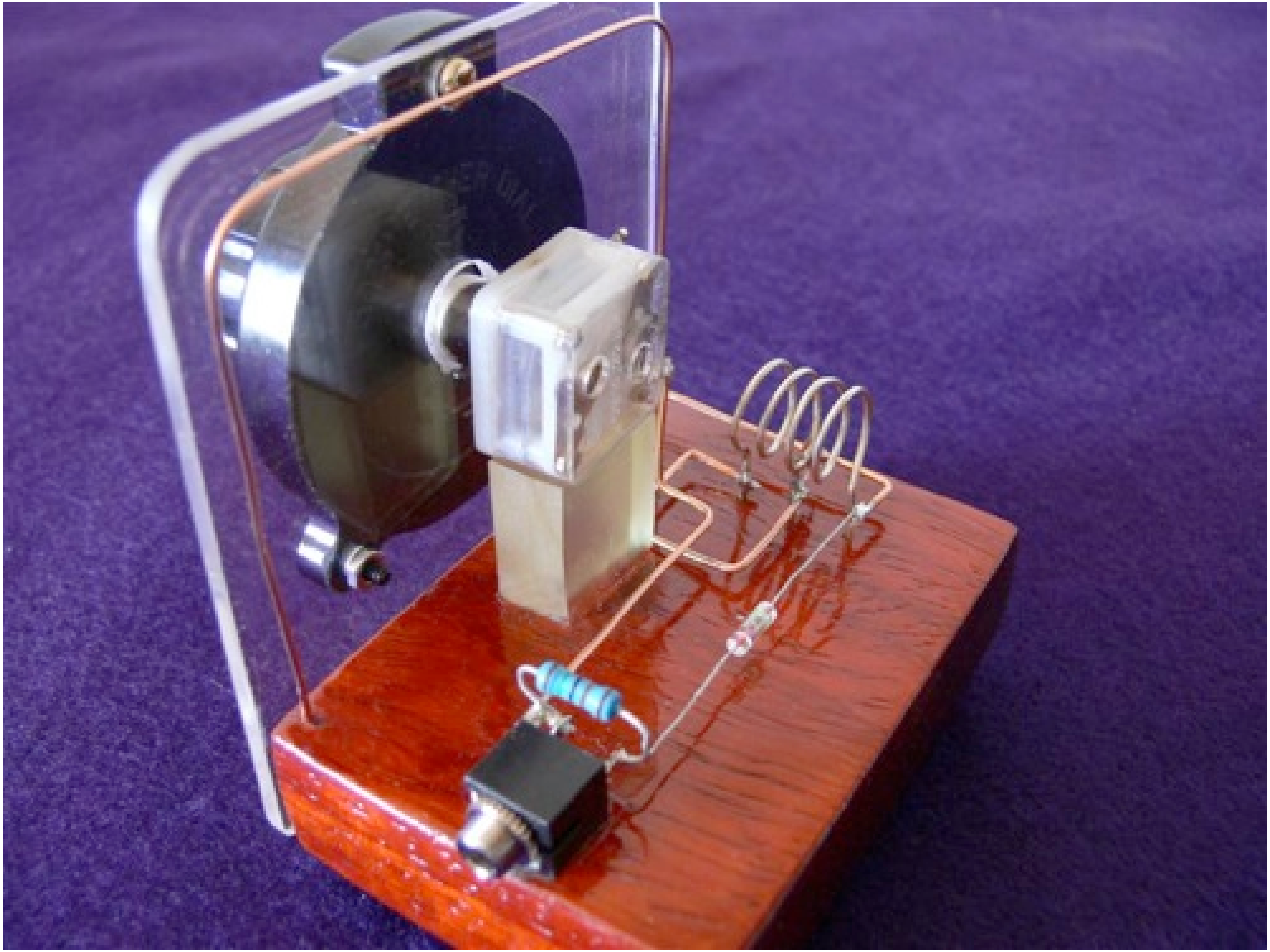

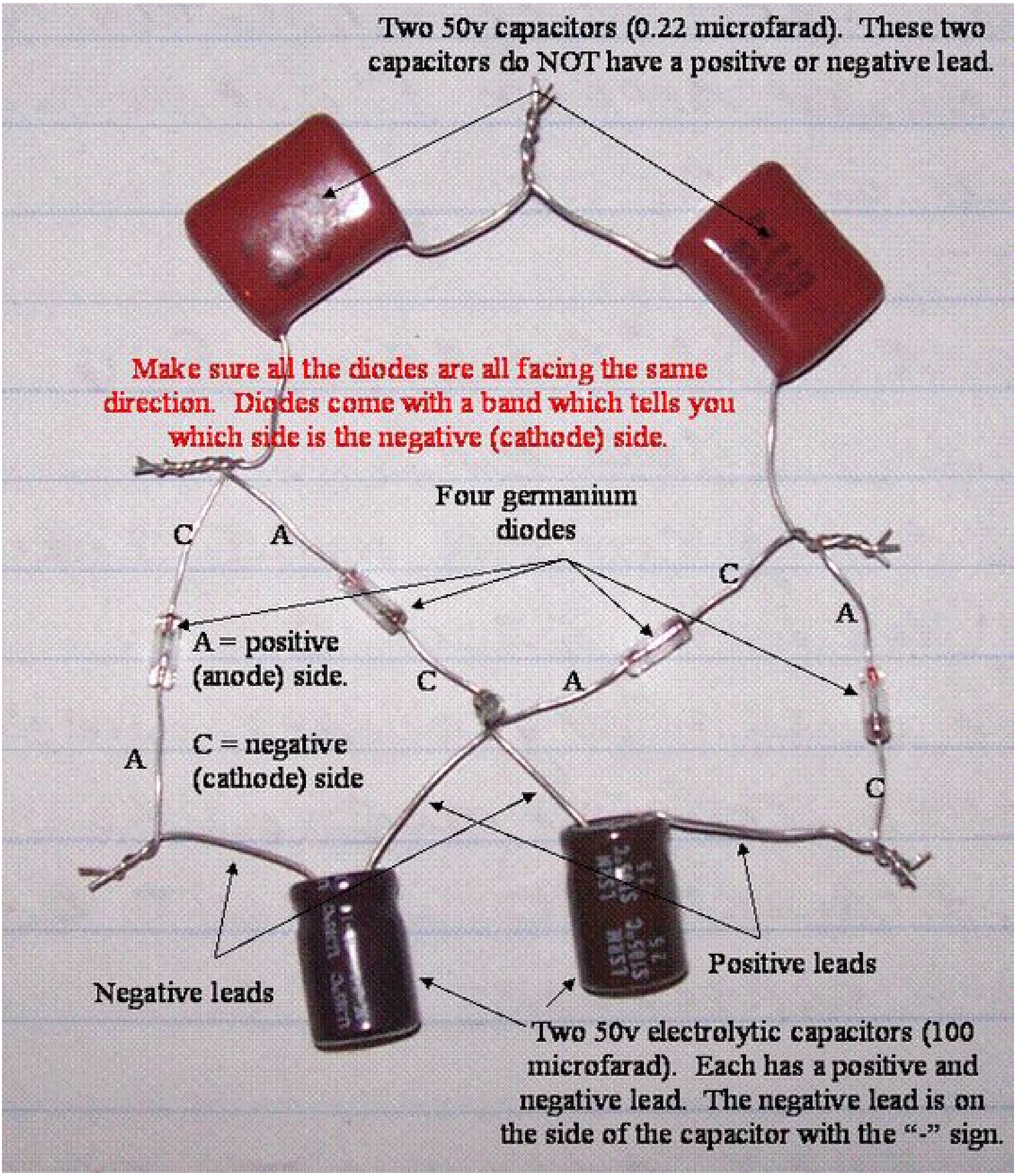

Two 50v capacitors (0.22 microfarad). These two capacitors do NOT have a positive or negative lead.

Make sure all the diodes are all facing the same direction. Diodes come with a band which tells you which side is the negative (cathode) side.

Four germanium diodes

A = positive (anode) side.

C = negative (cathode) side

Negative leads

Positive leads

Two 50v electrolytic capacitors (100 microfarad). Each has a positive and negative lead. The negative lead is on the side of the capacitor with the "-" sign.

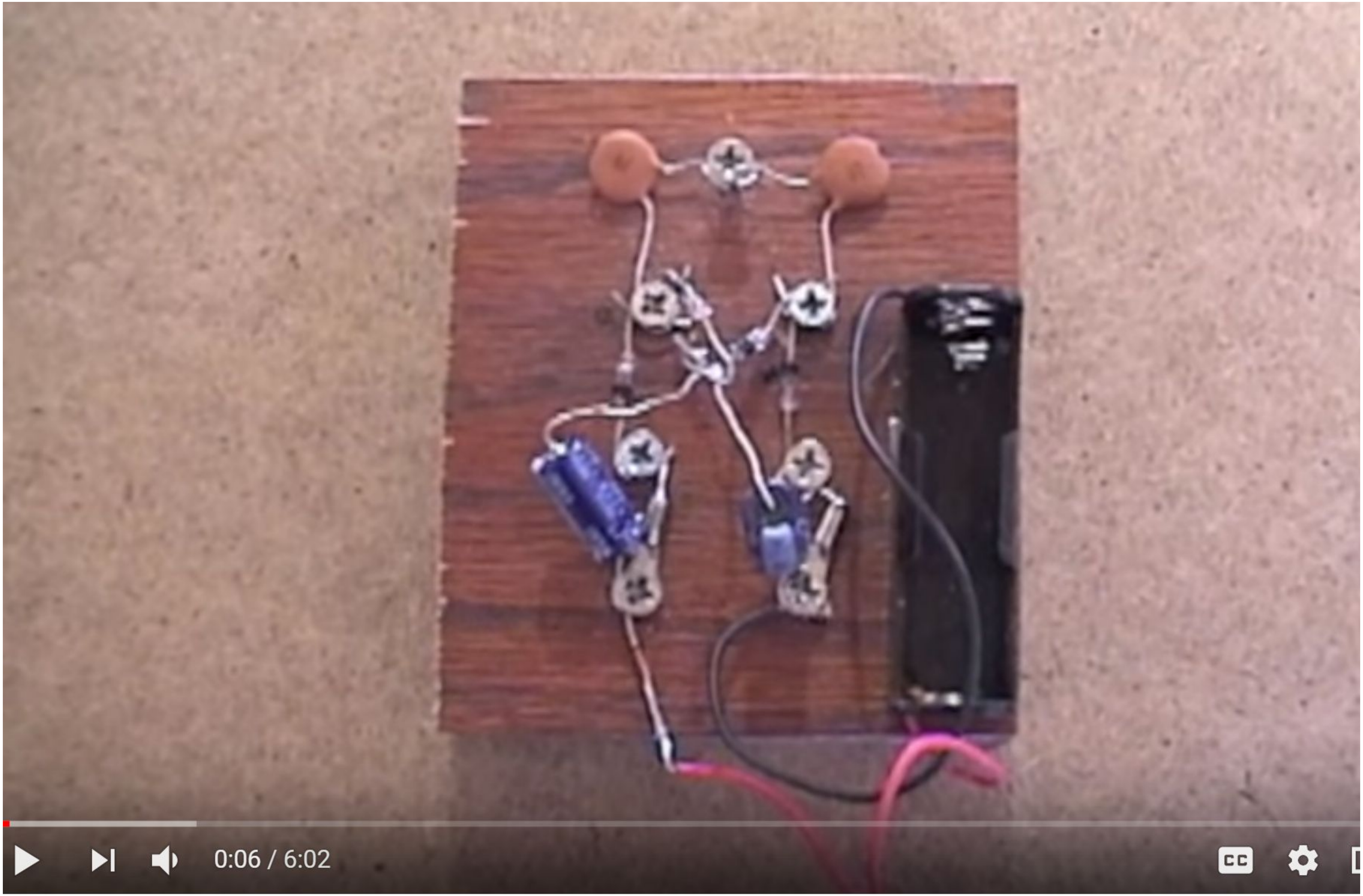
0:06 / 6:02

FIG. 10 BREADBOARD

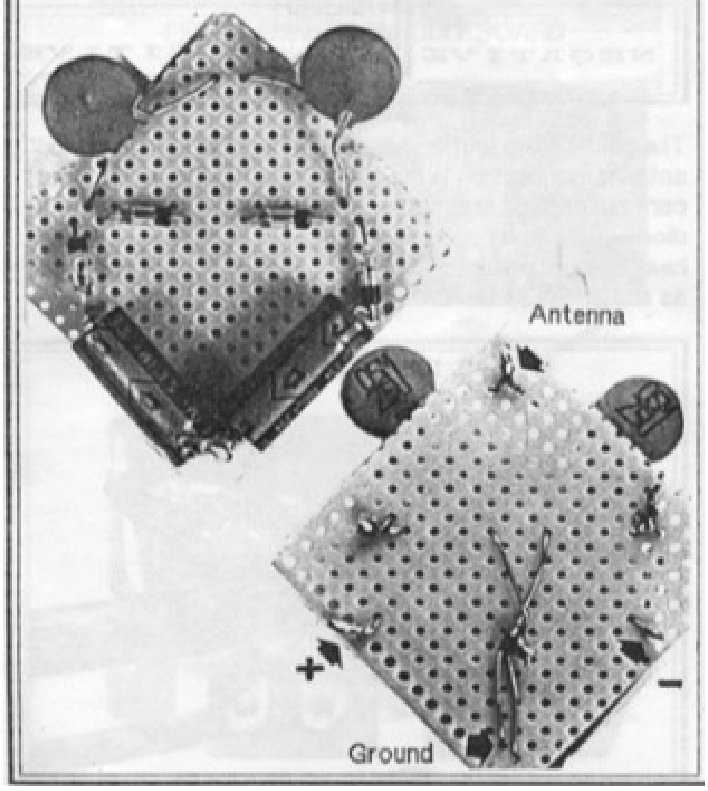
Antenna
+
−
Ground

the board as shown in Figure 11. This unit is powering a small red LED indicated by the arrow.

FIG. 11 EXPERIMENT BOARD

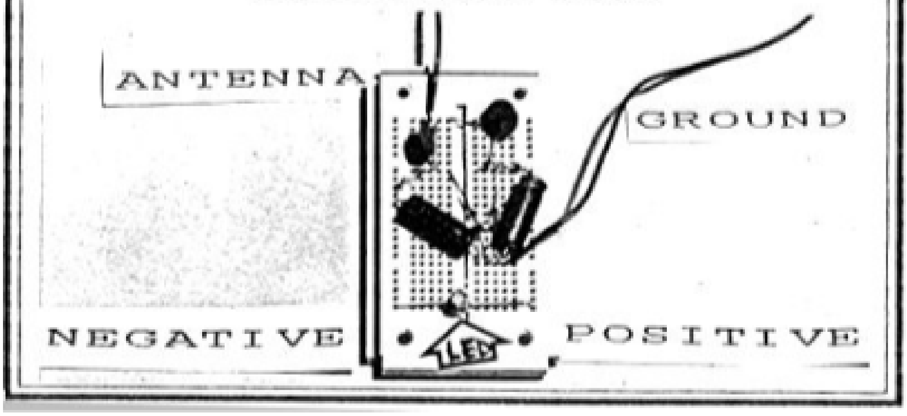
ANTENNA
GROUND
NEGATIVE
LED
POSITIVE

ade to the twisted ends of the ceramic capacitors. When soldering the leads of the 1N34 diodes, care must be taken

FIG. 9 TERMINAL STRIP

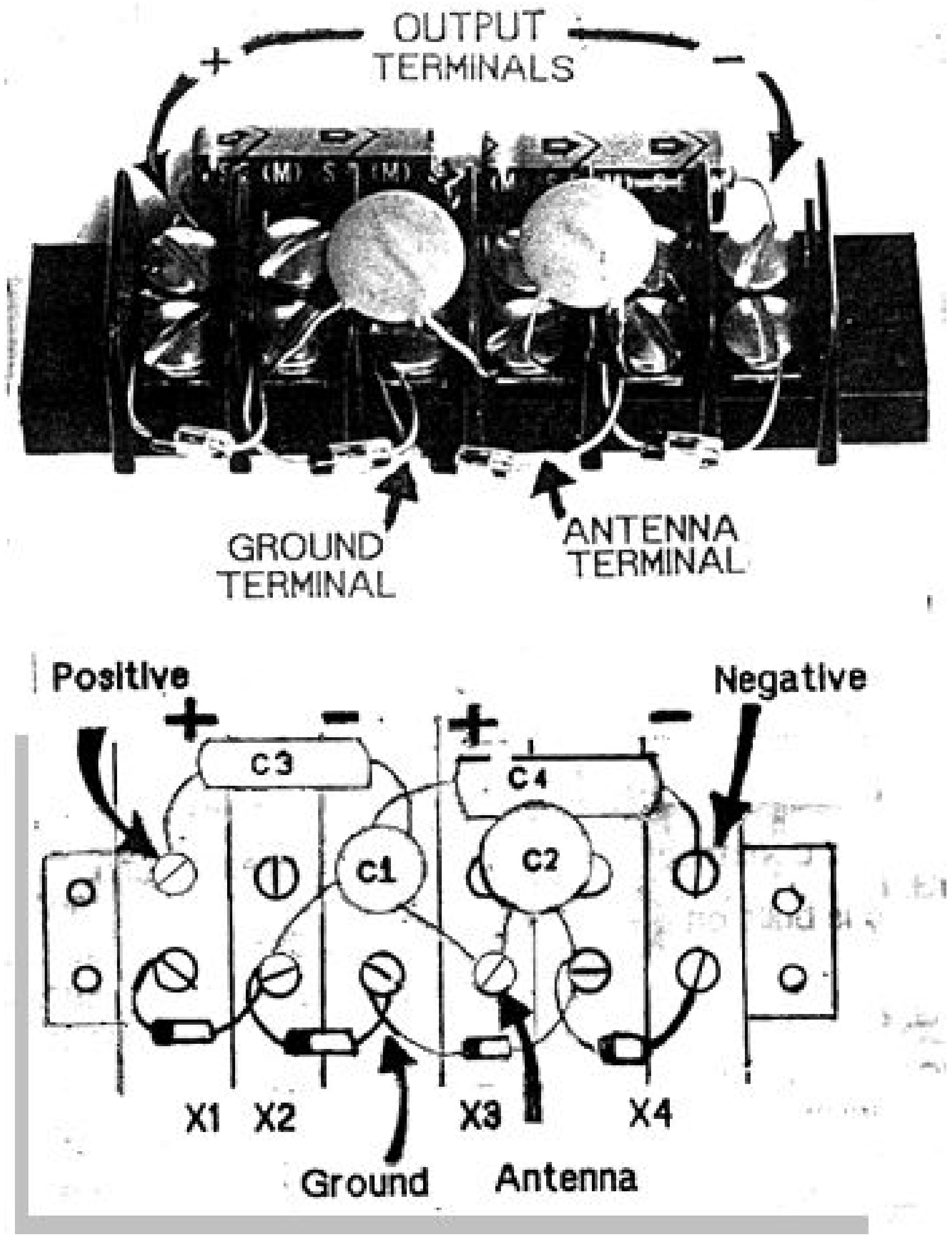
OUTPUT TERMINALS
+
−
GROUND TERMINAL
ANTENNA TERMINAL
Positive
Negative
+
−
+
−
C3
C4
C1
C2
X1
X2
X3
X4
Ground
Antenna

Amplifier circuit. The crystal radio used for example purposes is the crystal radio made from scraps here.

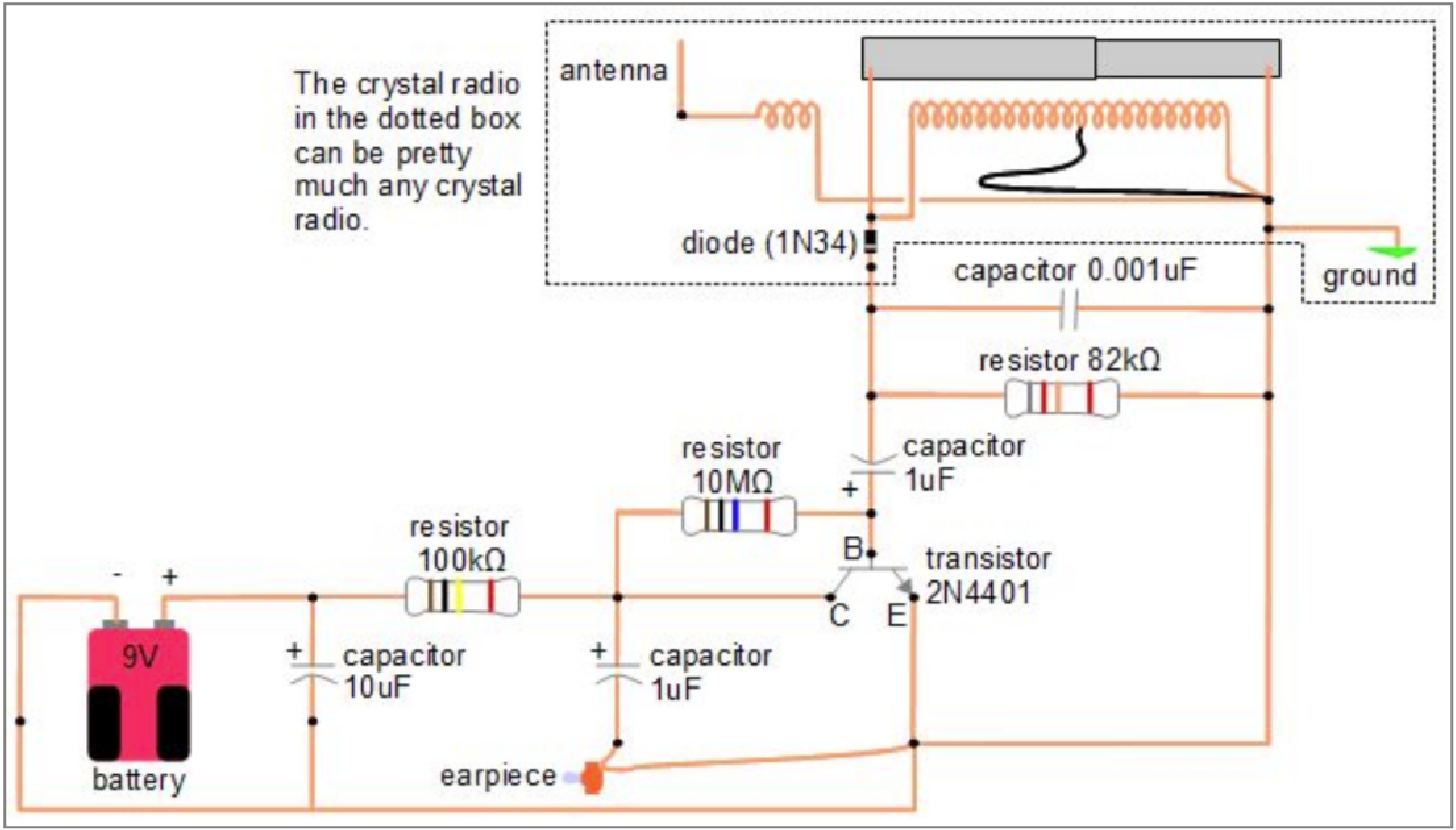
The crystal radio in the dotted box can be pretty much any crystal radio.
antenna
diode (1N34)
ground
capacitor 0.001uF
resistor 82kΩ
capacitor 1uF
+
resistor 10MΩ
B
transistor 2N4401
C
E
resistor 100kΩ
-
+
9V
battery
+
capacitor 10uF
+
capacitor 1uF
earpiece

SATURDAY, NOVEMBER 13, 2010

SIMPLE FIELD STRENGTH METER

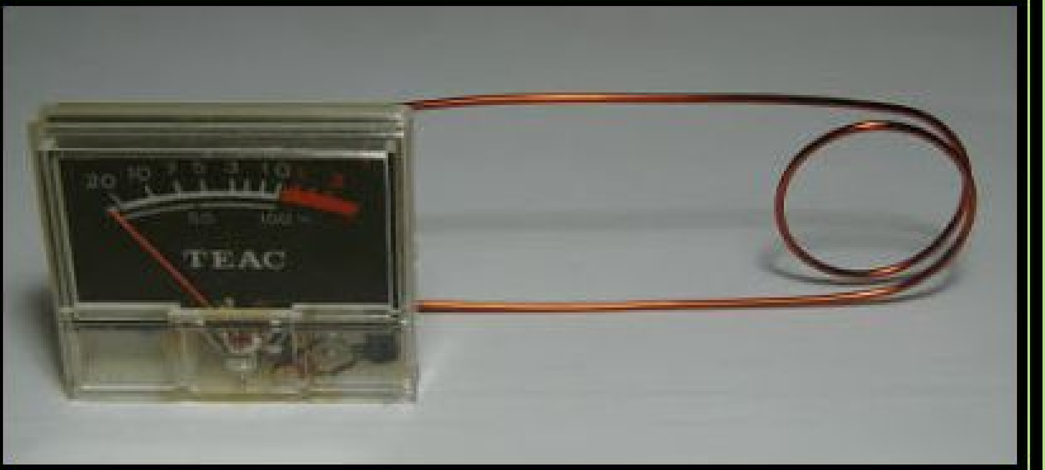
TEAC
20
10
7
5
3
1
0
50
100%

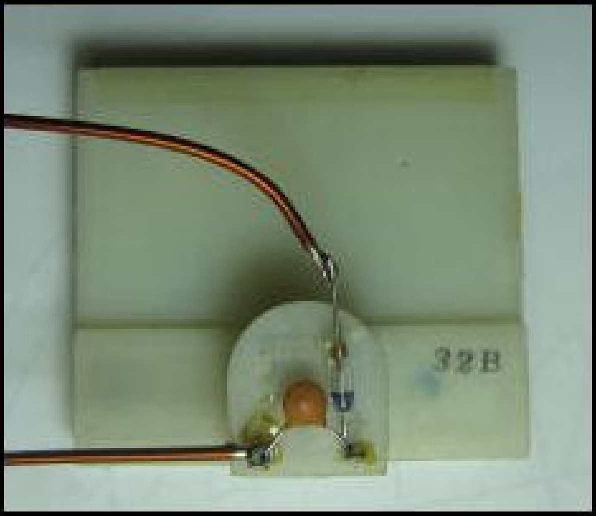

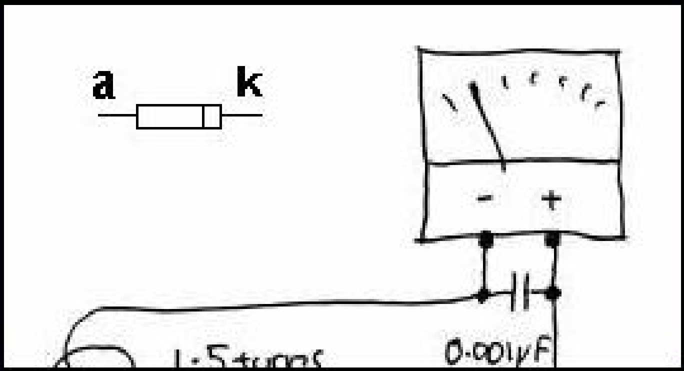
a
k
-
+
0.001μF
1.5 turns

MY

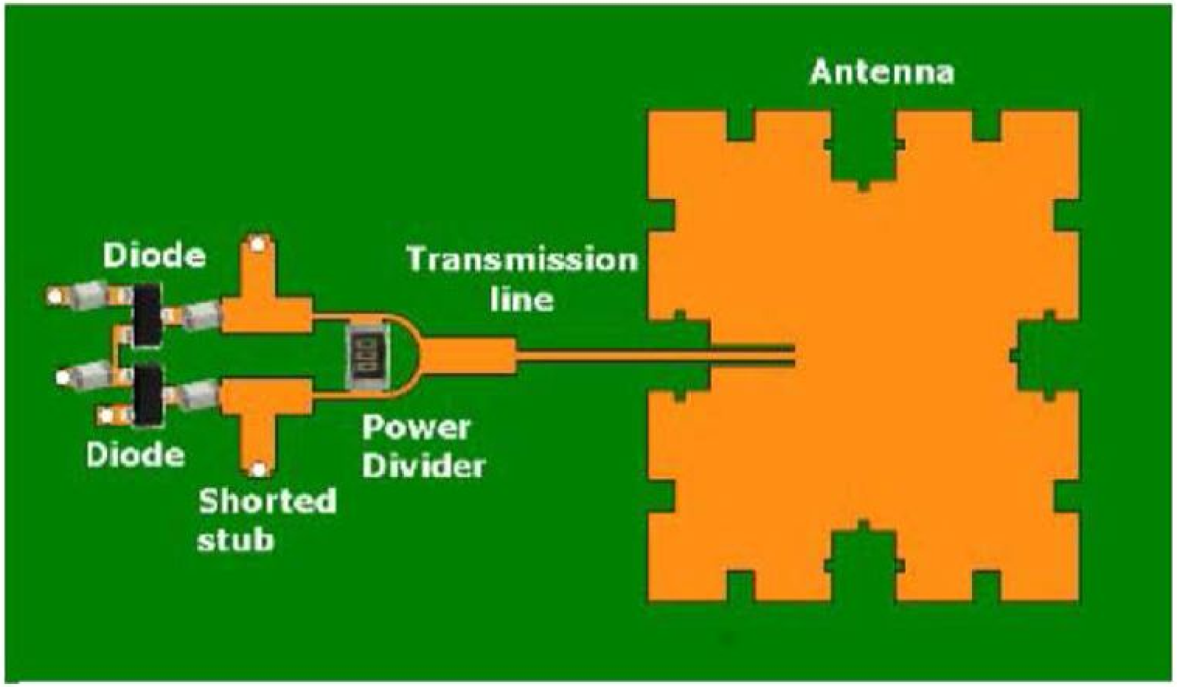
Antenna
Diode
Transmission line
Diode
Shorted stub
Power Divider

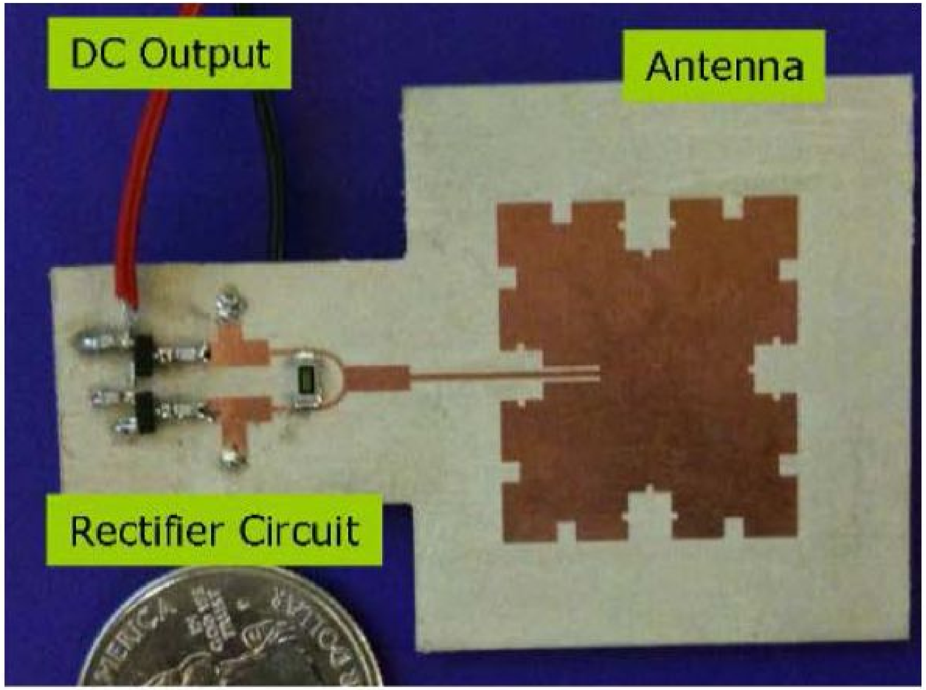
DC Output
Antenna
Rectifier Circuit

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

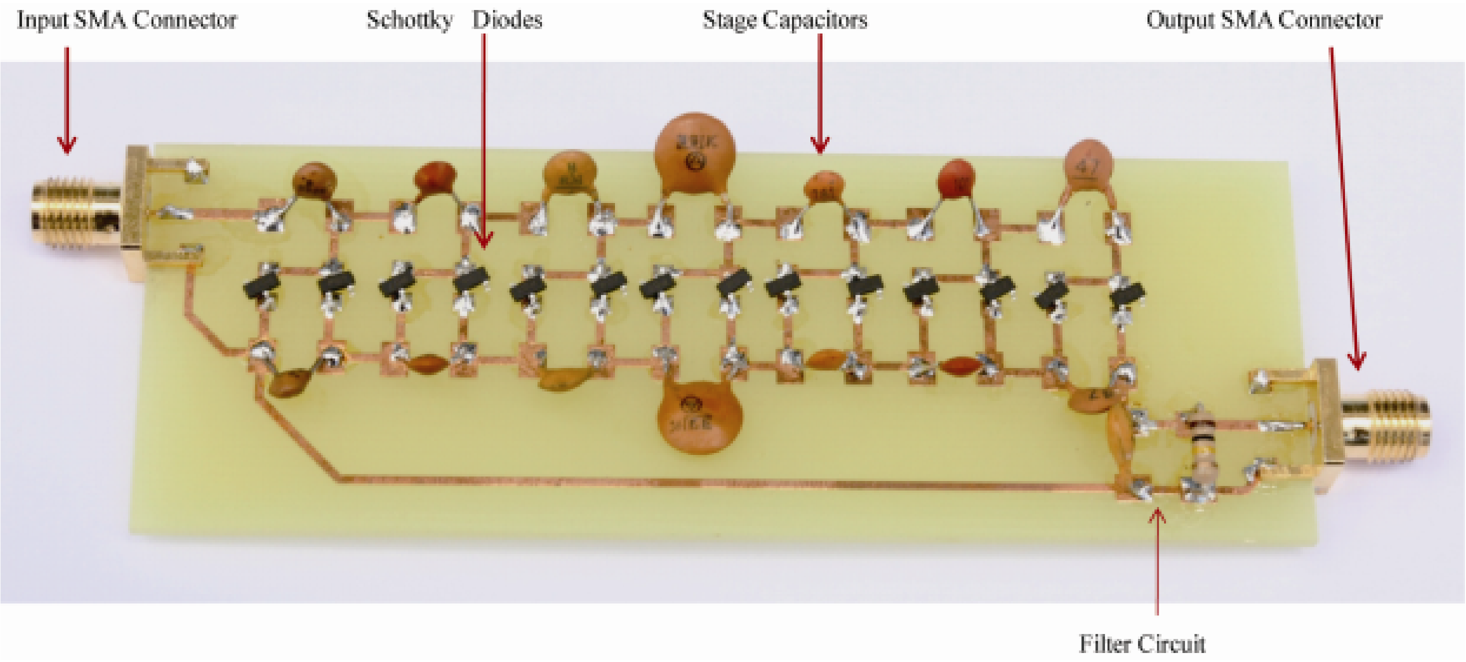

Figure 7. Photograph of assembled circuit board.

sign in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

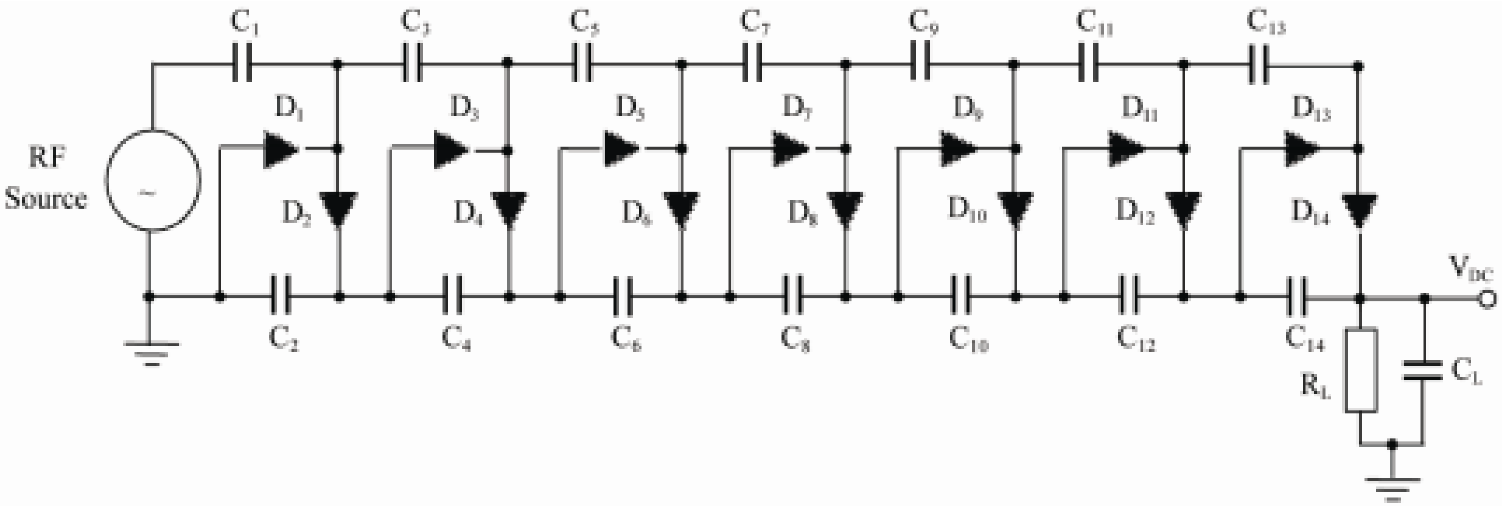

Figure 6. Schematic of 7 stage voltage multiplier.

Copyright © 2012 SciRes.

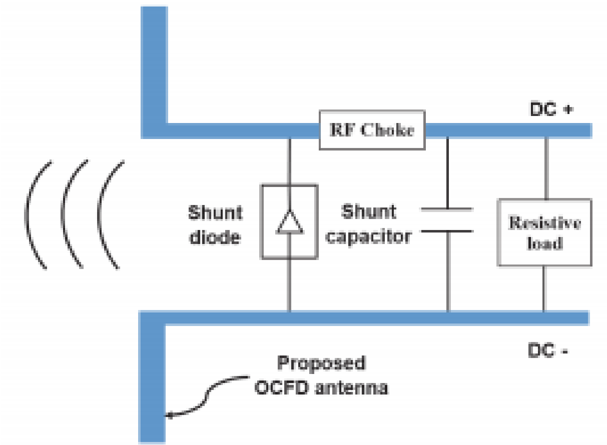

Fig. 9. Configuration of a single shunt diode (Class F) rectifier with a dipole antenna.

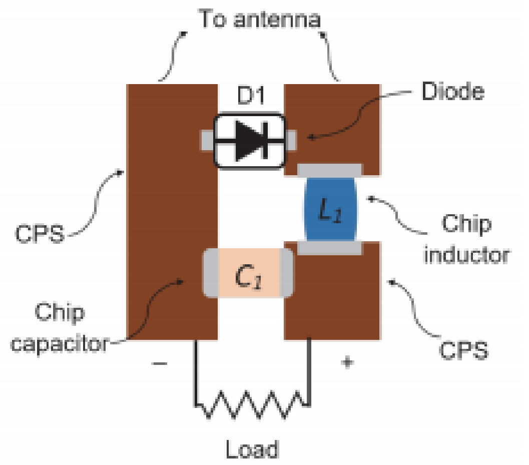

Fig. 10. Configuration of the proposed rectifier on coplanar striplines (CPS).

TABLE III
CIRCUIT COMPONENTS USED IN THE DESIGN

Component name	Nominal Value	Part number and supplier
D1	Schottky diode	SMS7630-079LF, Skyworks
L1	47 nH chip inductor	0603HP47N, Coilcraft
C1	100 nF chip capacitor	GRM188R71H104JA93D, Murata

antenna have a radius of 50 mm and a circumference angle of

respectively. While the imaginary part of the proposed OCFD is around 0 Ω at resonant frequencies 0.6 GHz, 1.2 GHz and 2.4 GHz, which are f_0, $2f_0$, and $4f_0$ respectively. These results have demonstrated that the simulated results agree with the OCFD theory as discussed in Section III–A. Furthermore, the imaginary part of the impedance of the antenna over the resonant frequency band from 1.4 to 2 GHz turns from negative values (for the reference antenna) to positive values (for the proposed antenna). As shown in Fig. 7(b), the value of the imaginary part of the proposed antenna impedance varies between 0 and 300 Ω over the desired frequency band. This feature could help the proposed antenna to produce a better conjugate matching with the rectifier, since the imaginary part of the impedance of the rectifier normally varies between -700 and 0 Ω as we discussed earlier. The simulated 3D radiation patterns of the proposed antenna at the frequencies of interest are depicted in Fig. 8. The 2D polar plots of antenna patterns in *E-plane* and *H-plane* are shown as well. Here we have only showed the directivity (maximum gain) of the antenna (without taking the mismatch loss into account). From Fig. 8, it can be seen that the antenna has symmetrical patterns about YOZ plane with a maximum directivity of 1.8 dBi at 0.9 GHz, 3.5 dBi at 1.8 GHz and 3.3 dBi at 2.4 GHz. The antenna is more directive towards the long arm direction at 1.8 GHz and 2.4 GHz with the half-power beam-widths (HPBW) of around 174° and 185° respectively. The HPBW is about 96° at 0.9 GHz.

Therefore, the proposed broadband OCFD antenna has obtained high impedance over a wide frequency range. The proposed design is just an example to illustrate the proposed new method. The details of the dipole could be modified according to the frequency of interest.

IV. RECTENNA INTEGRATION

A. Rectifier Configuration

The proposed high impedance OCFD antenna may directly conjugate match with the input impedance of a rectifier over a wide frequency band. The rectifier should only consist of few circuit components for rectification, DC storage and output. A single shunt diode rectifier is selected due to its very simple structure and high conversion efficiency [33]. The configuration of the single shunt diode rectifier with a dipole

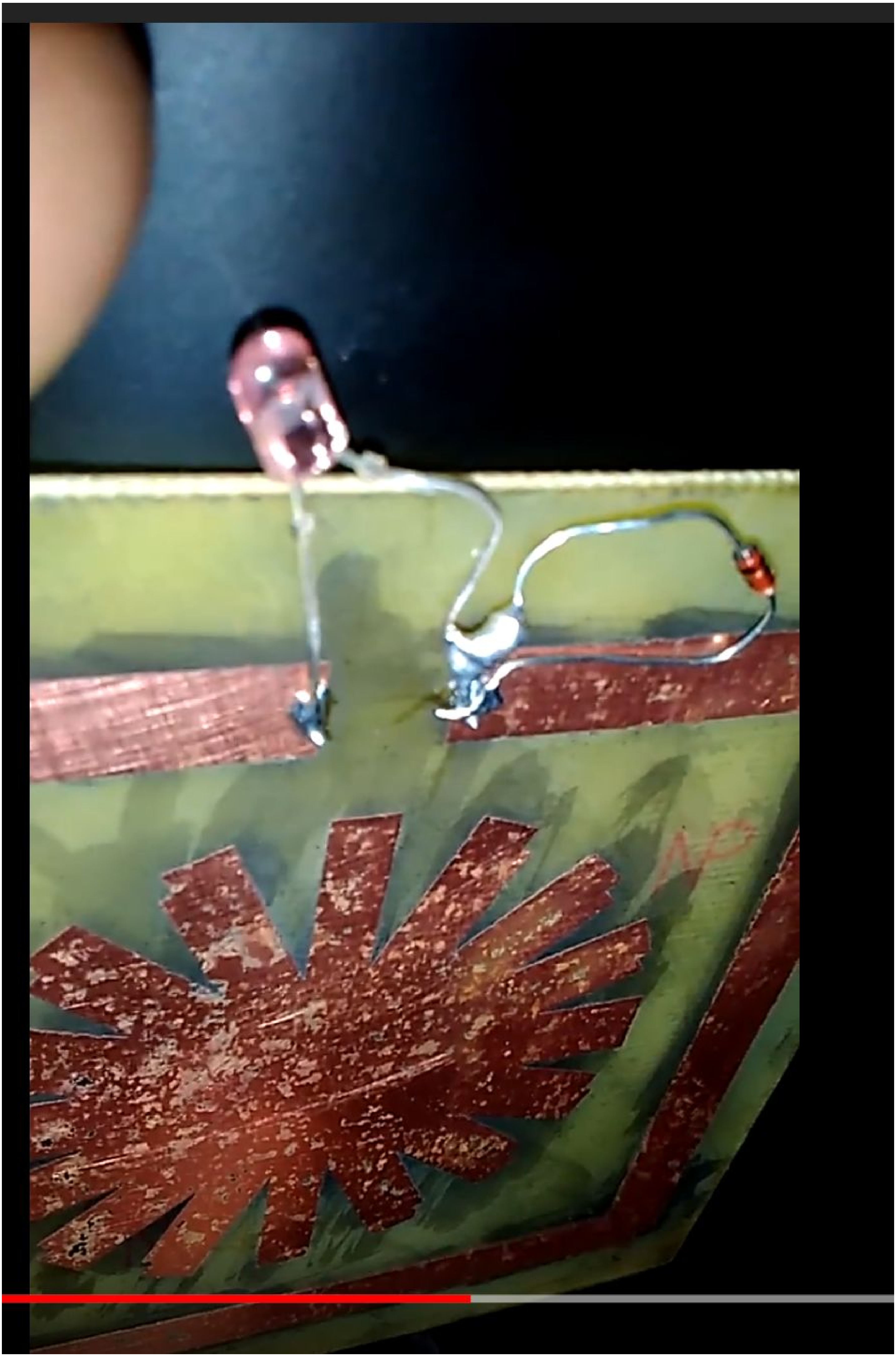
AP

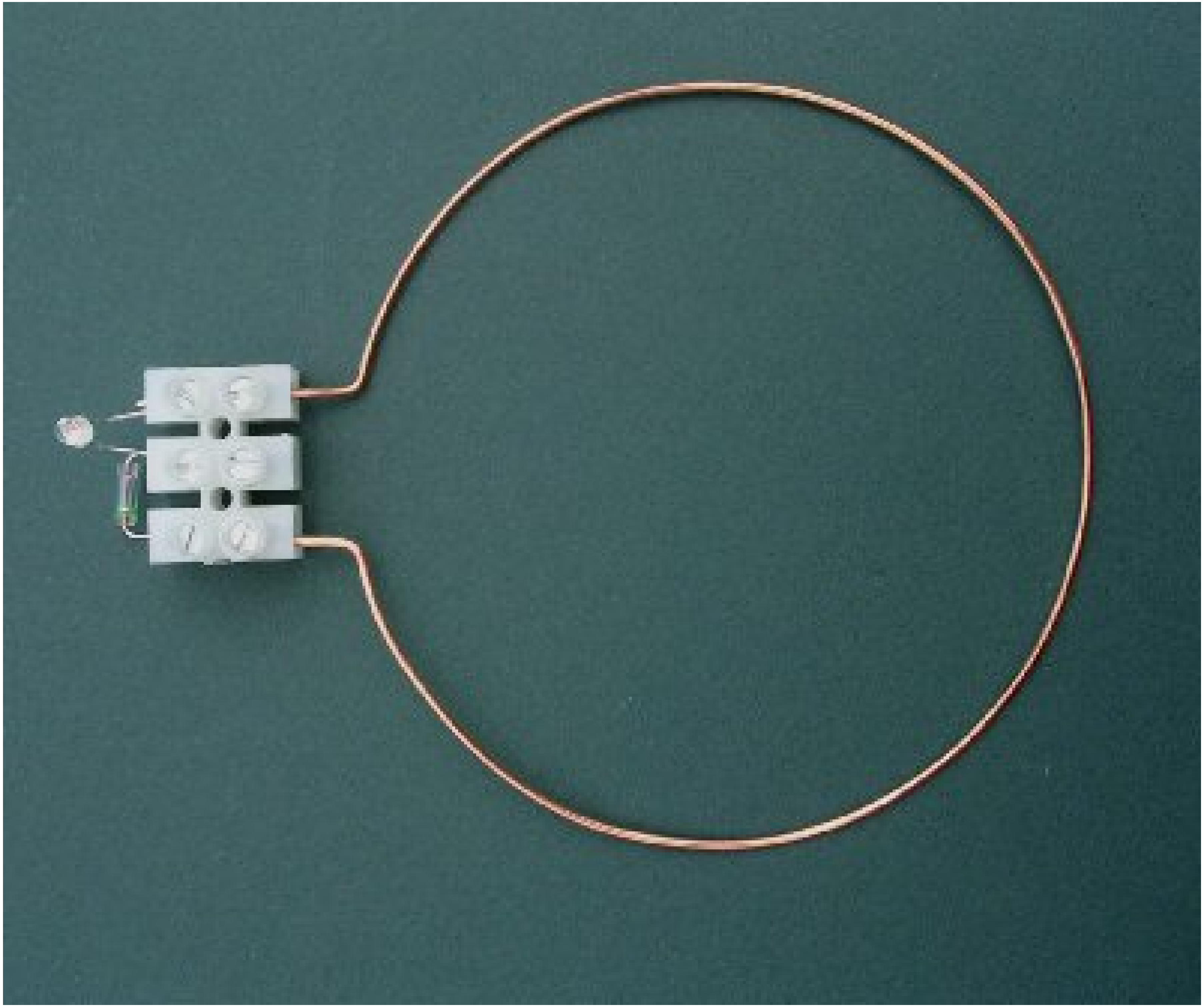

Sony Ericsson

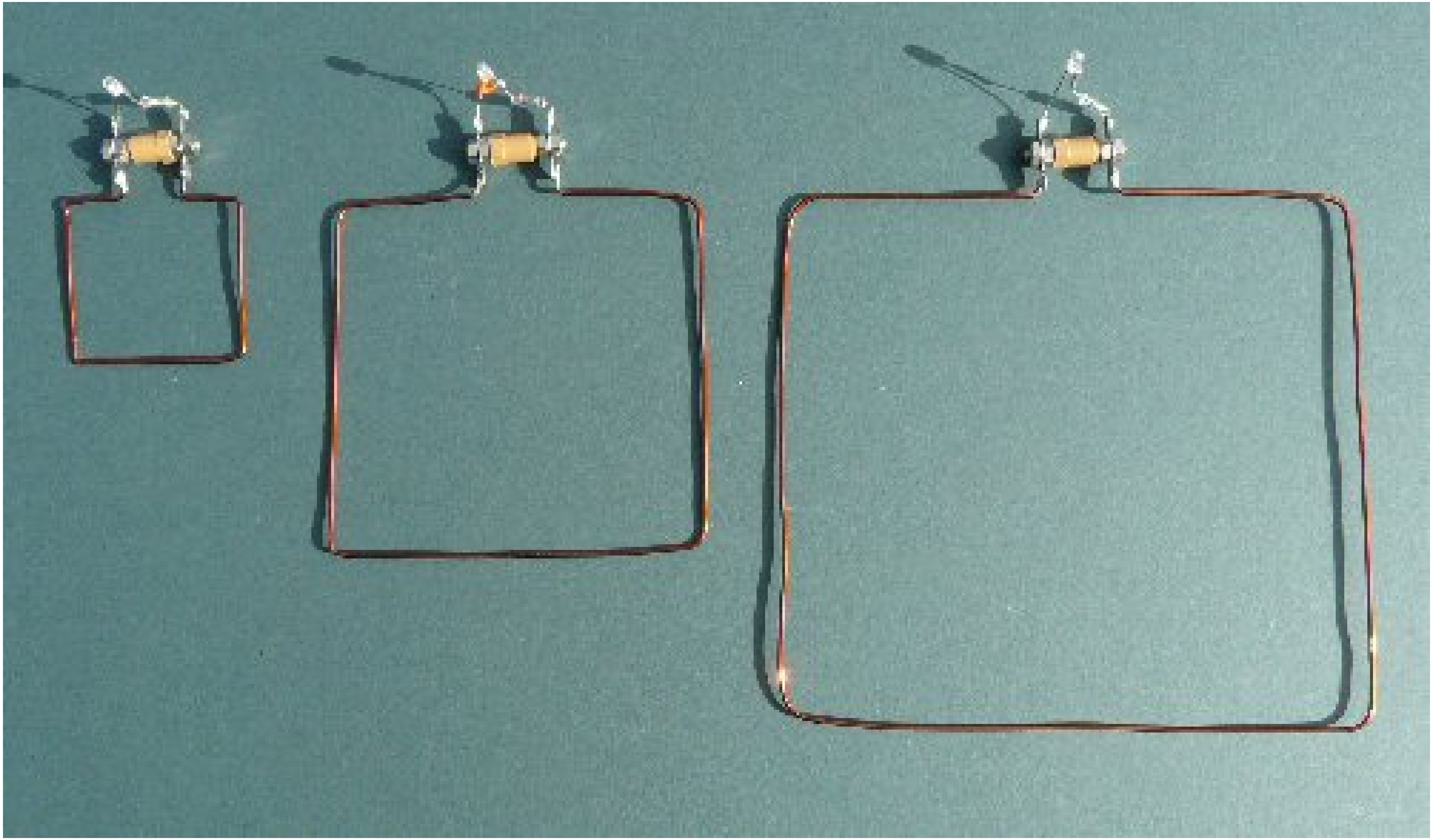

ake and it can be sent in an envelope. I made tens of them and sent them to politicians, newspapers, universities... I gave some to local people, together with a user
later version, that can be rolled up in an envelope that fits the conditions to be sent with only one stamp.

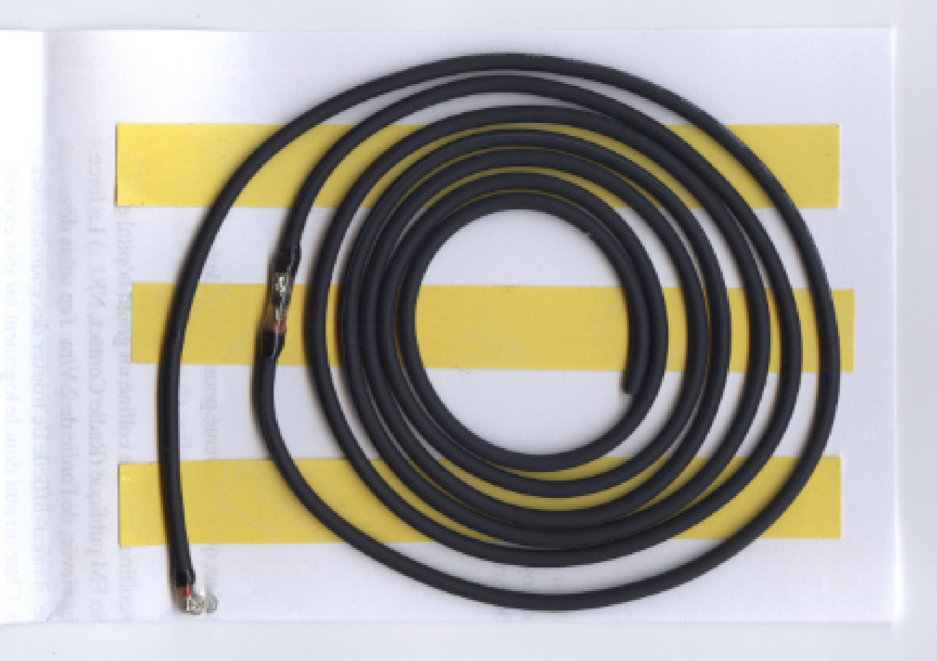

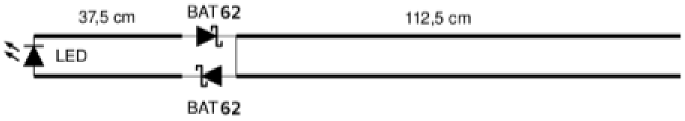

BAT62 detection diodes are no more produced. **BAT15** diodes work fine but they wear out; after a few months they stop functioning. **SMS7630** diodes are great but very little and mechanically fragile. **MMSD701T1G** diodes are sturdy and powerful; an excellent choice for a beginner. Such SMD diodes do also work for cell phone frequencies, which allows to test out a snake with a calling cell phone pushed against it. But any detection diodes that can manage 100 MHz will do.

The LED I'm currently using is the **L-7113SEC-H** . It lights up with a low tension and a very low current (the bluer a LED, the more tension it needs). Its color is red yet close to orange hence it is easily seen by the human eye (the eye is most sensitive to green, yellow and orange). The beam is quite narrow so when the LED is directed towards somebody's eyes it will appear quite bright.

For the lengths of 37.5 and 112.5 centimeters, any electric wire with two copper conductors will do. Audio signal wire is a practical solution. Use the shielding as one of the two conductors. The lengths of the two segments must not be precise. What matters is that the total length of the snake be 1.5 meters. Do not hesitate to try out if a little longer or shorter snake gives better results.

A schematic of my current probes, that I connect to a standard multimeter, measuring Volts DC. The measure displayed by the multimeter must be multiplied by 10. When using a 200.0 mV scale, just read while forgetting the dot:

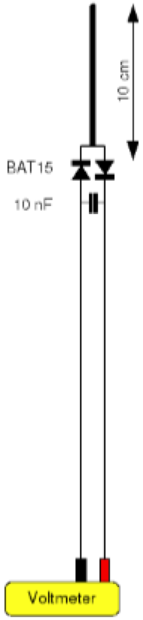

Lead Free Status	Lead Free
RoHS Status	RoHS Compliant

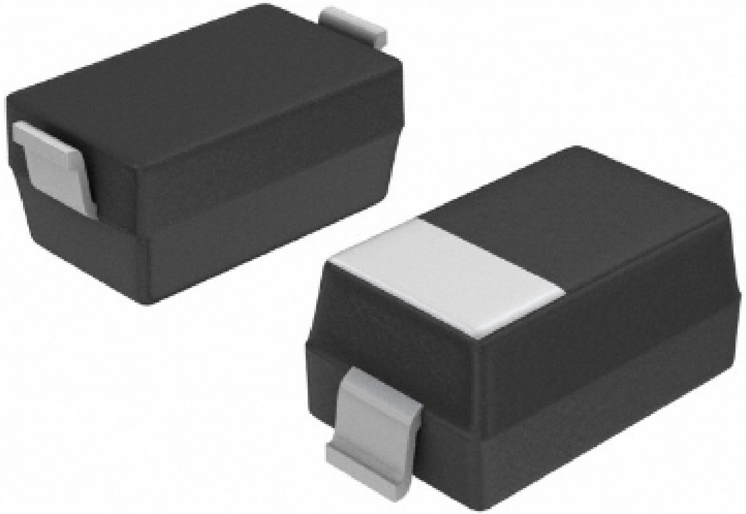

Features, Applications

The MMSD301T1, and MMSD701T1 devices are spin-offs of our popular MMBD301LT1, and MMBD701LT1 SOT-23 devices. They are designed for high-efficiency UHF and VHF detector applications. Readily available to many other fast switching RF and digital applications.

Extremely Low Minority Carrier Lifetime Very Low Capacitance Low Reverse Leakage AEC Qualified and PPAP Capable S Prefix for Automotive and Other Applications Requiring Unique Site and Control Change Requirements These Devices are Pb-Free, Halogen Free/BFR Free and are RoHS Compliant*

XXX G = Specific Device Code SMMSD701T1G = Date Code = Pb-Free Package
Rating Reverse Voltage MMSD701T1G, SMMSD701T1G Forward Current (DC) Continous Forward Power Dissipation = 25C Junction Temperature Storage Temperature Range Symbol VR Value to +150 Unit Vdc M G

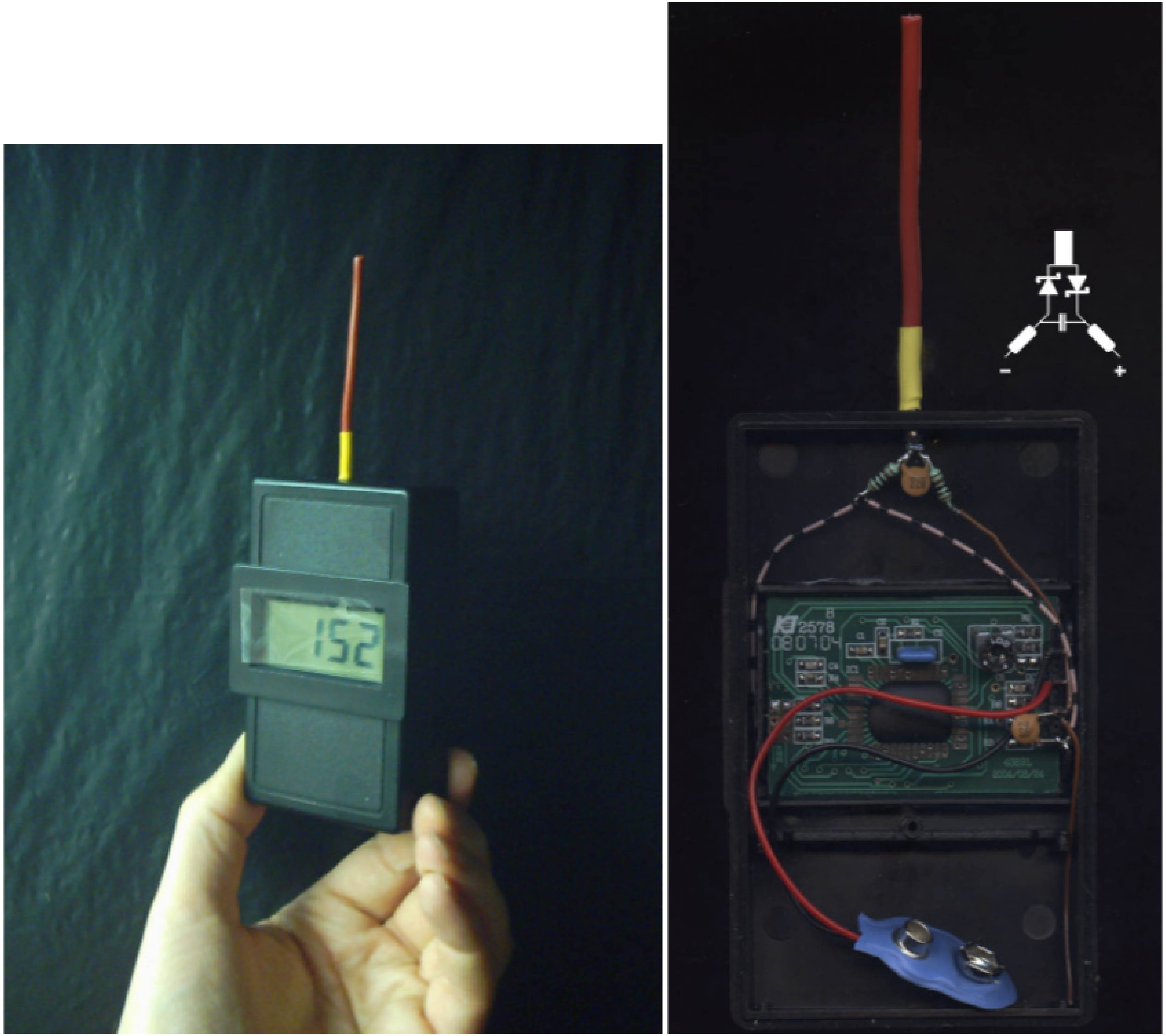
152
2578
080704
-
+

Parts Required:

1. 100nF & 100pF
2. n4148 Diode x(2)
3. A bright good quality LED

Circuit Diagram:

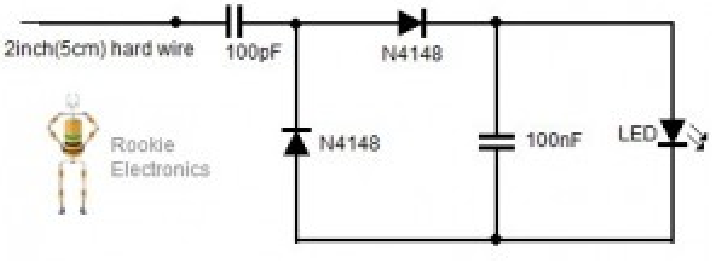
2inch(5cm) hard wire
100pF
N4148
N4148
100nF
LED
Rookie Electronics

Bread board Arrangement:

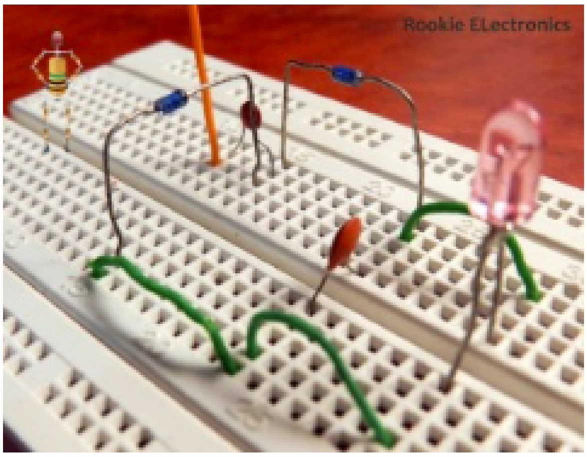
Rookie ELectronics

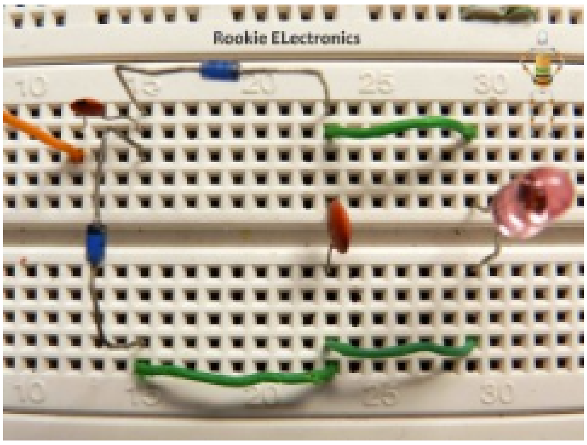
Rookie ELectronics

Strip Board:

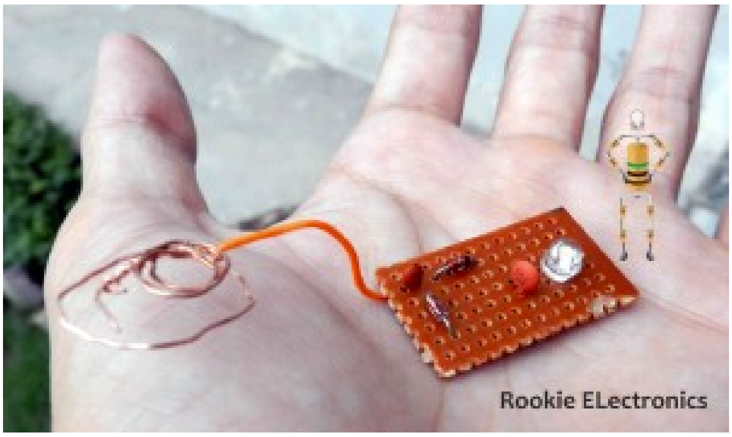
Rookie ELectronics

Rookie ELectronics

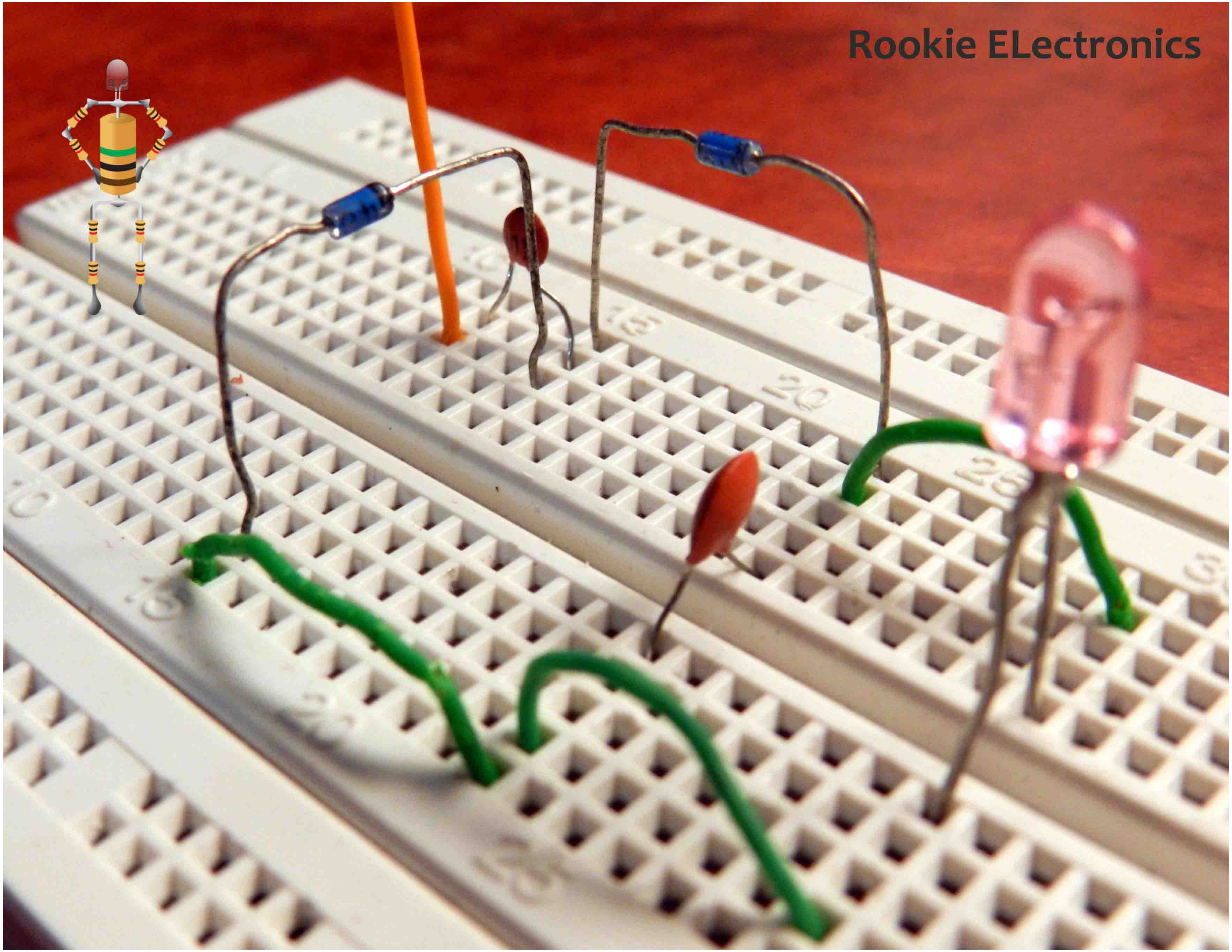

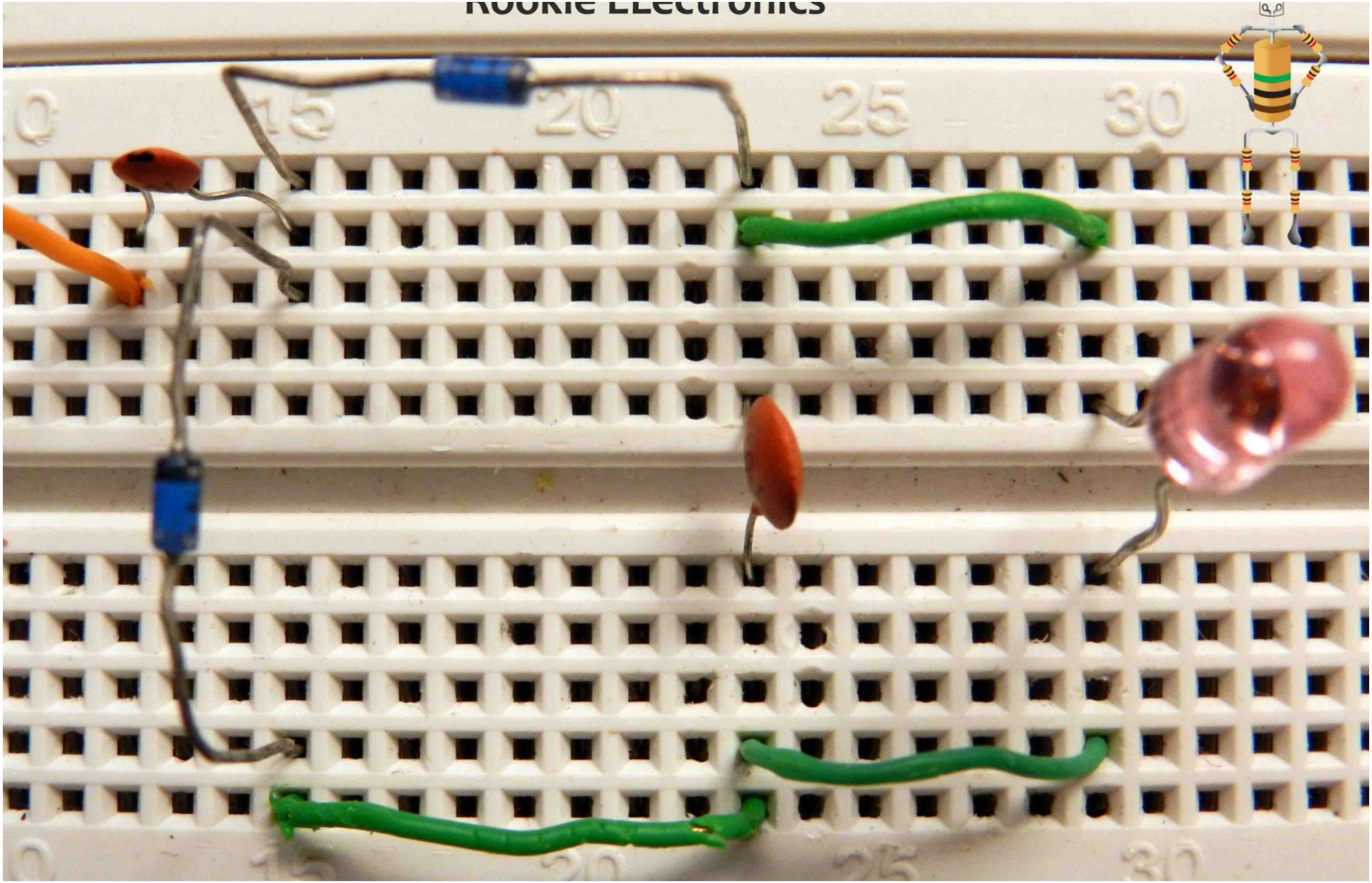

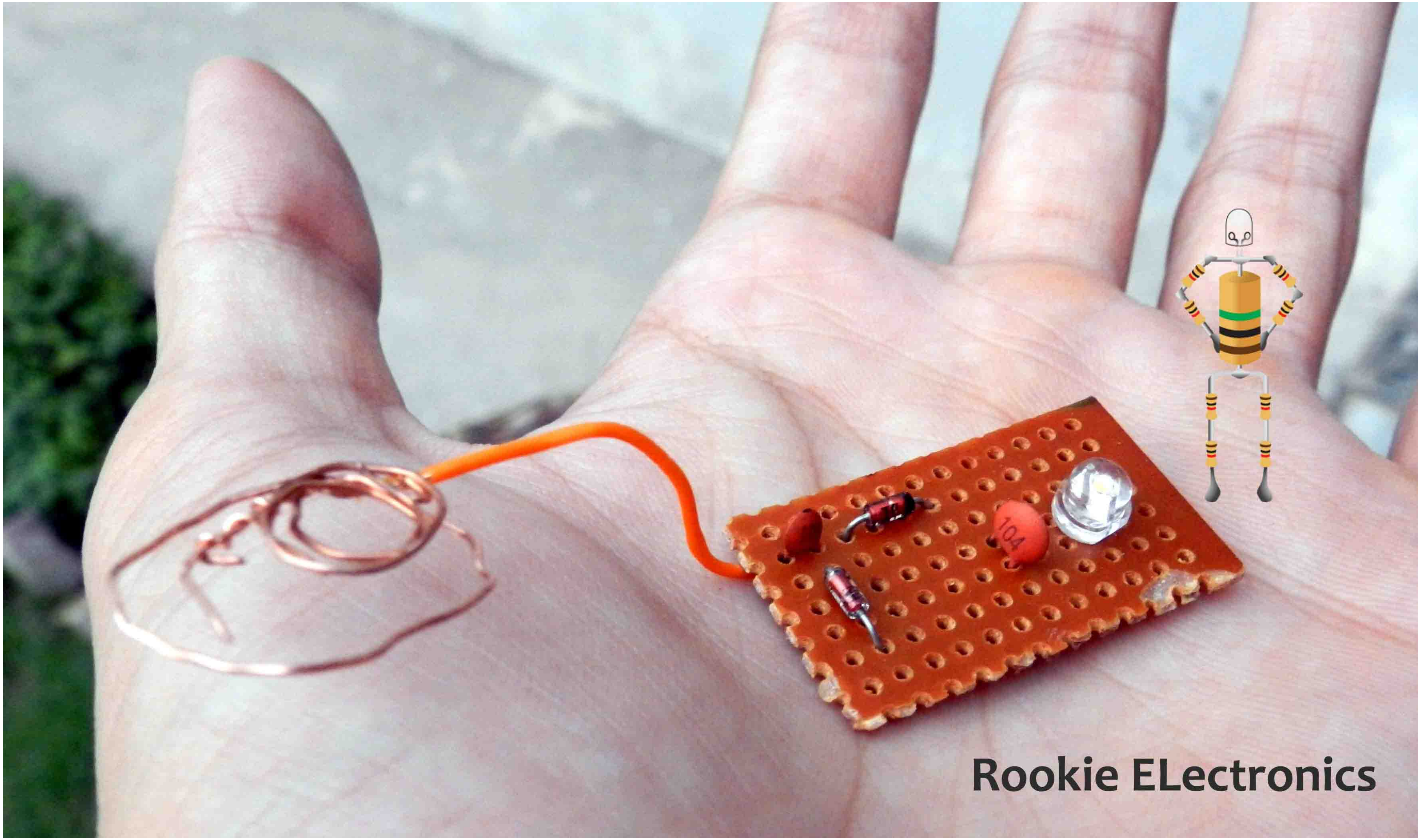

Rookie ELectronics

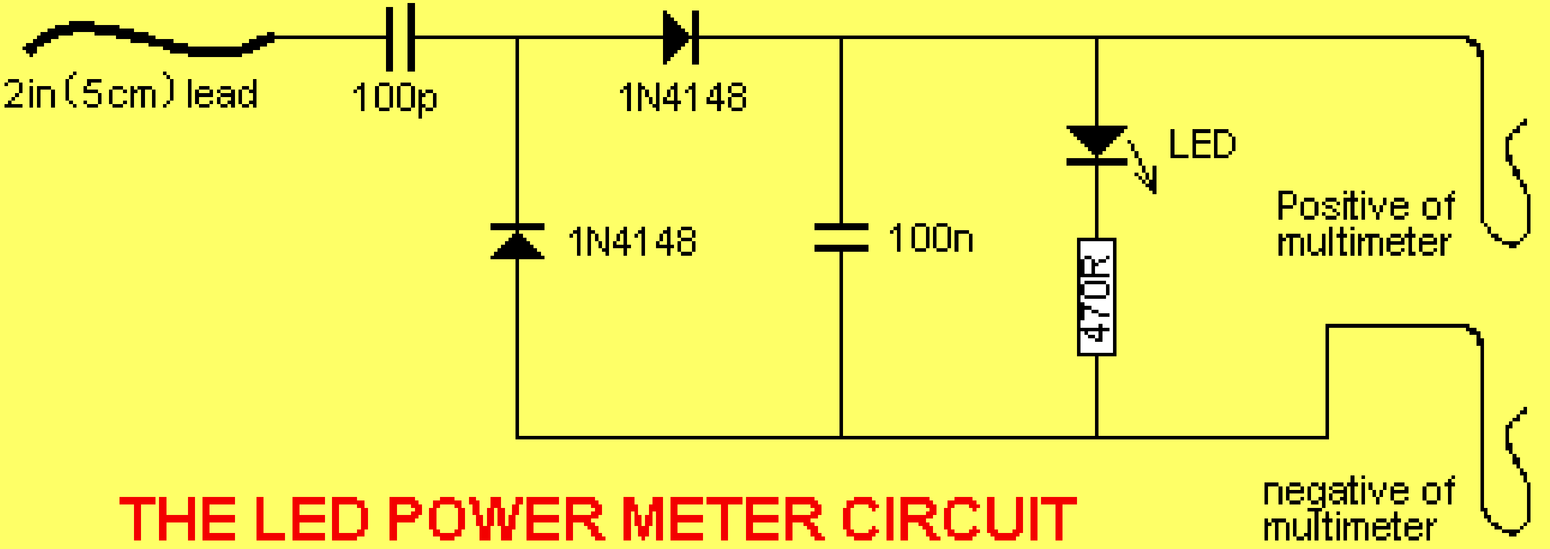
2in(5cm) lead
100p
1N4148
1N4148
100n
LED
470R
Positive of multimeter
negative of multimeter

THE LED POWER METER CIRCUIT

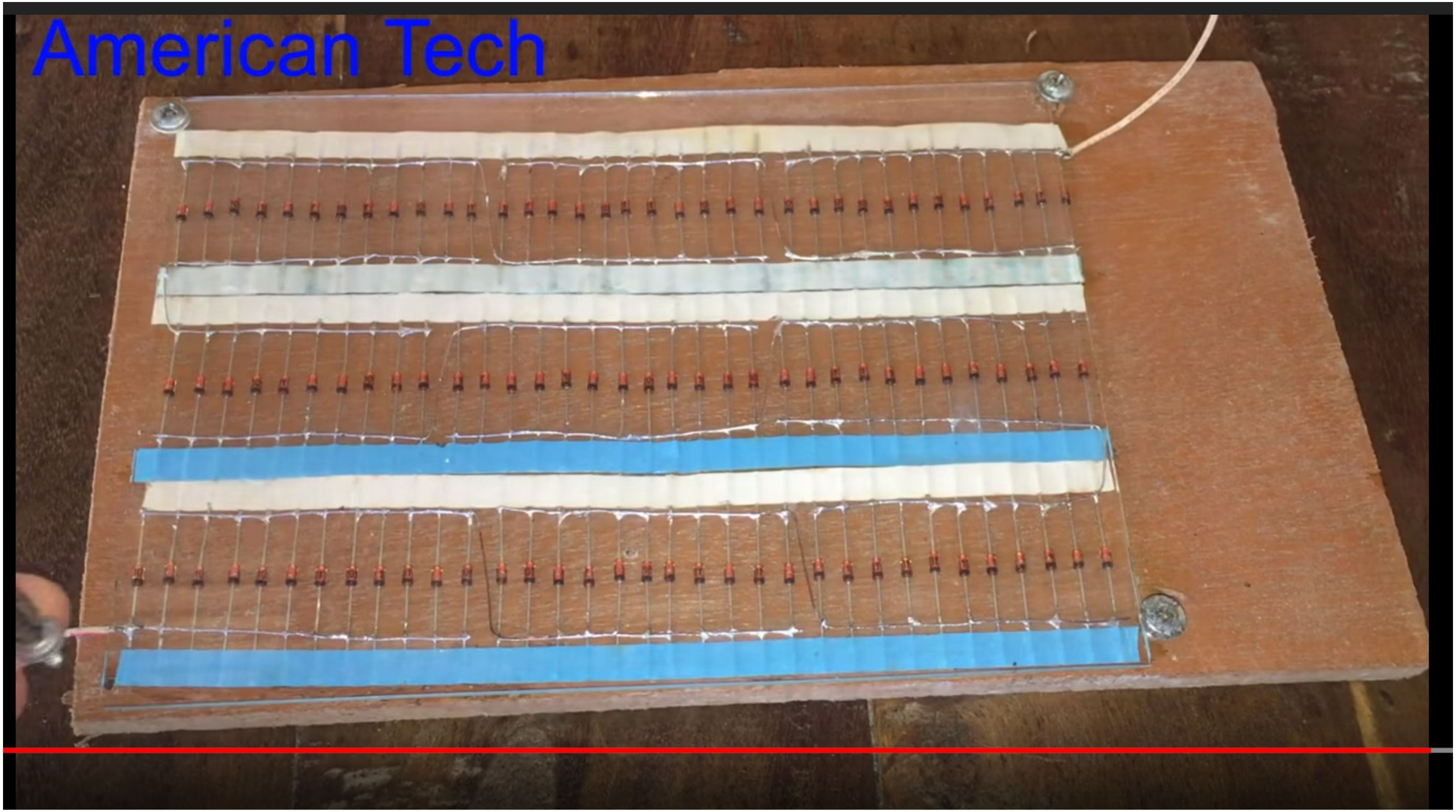

American Tech

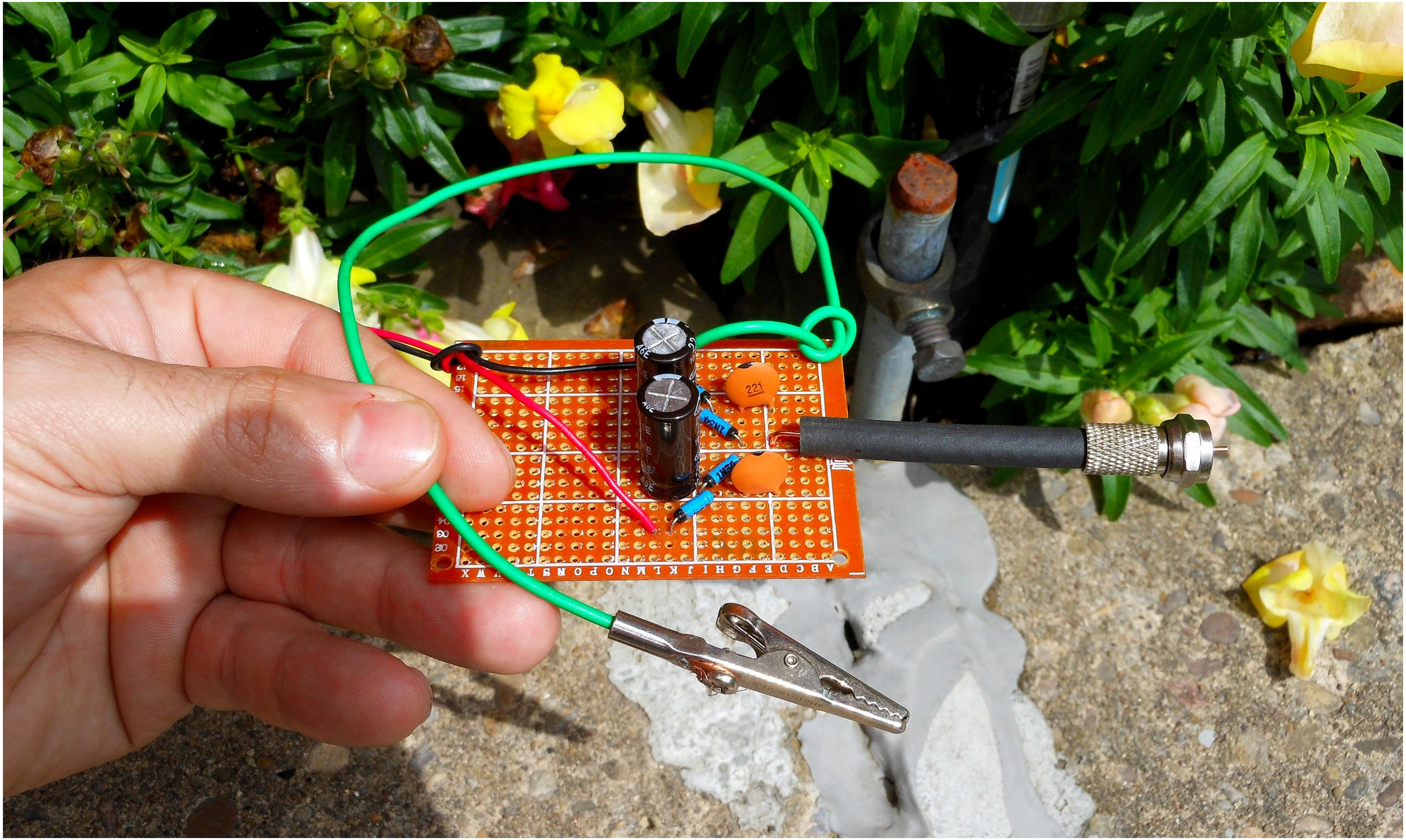
221

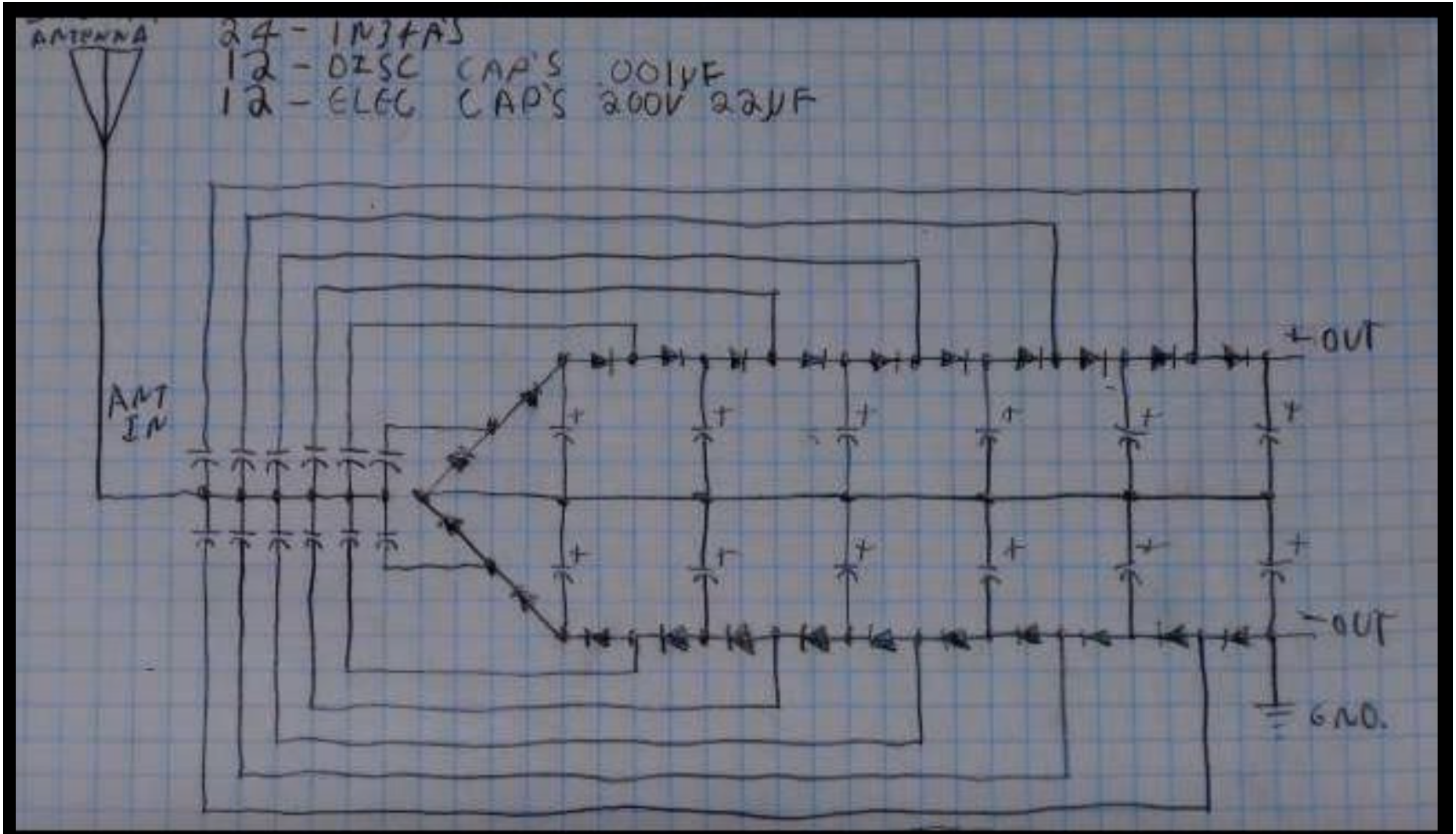
ANTENNA
ANT IN
+OUT
-OUT
GND.
+

24 - 1N34A'S
12 - DISC CAP'S .001µF
12 - ELEC CAP'S 200V 22µF

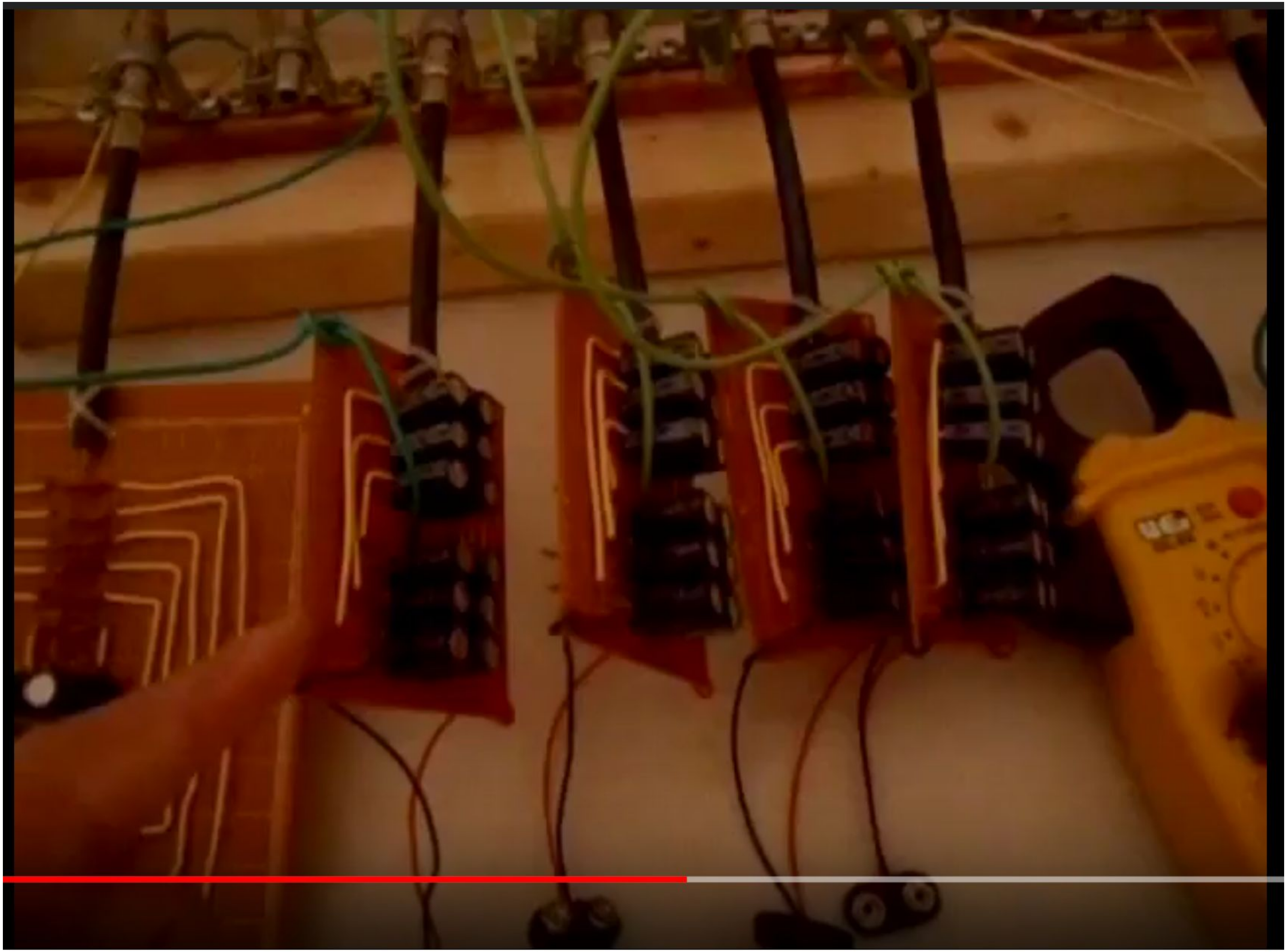

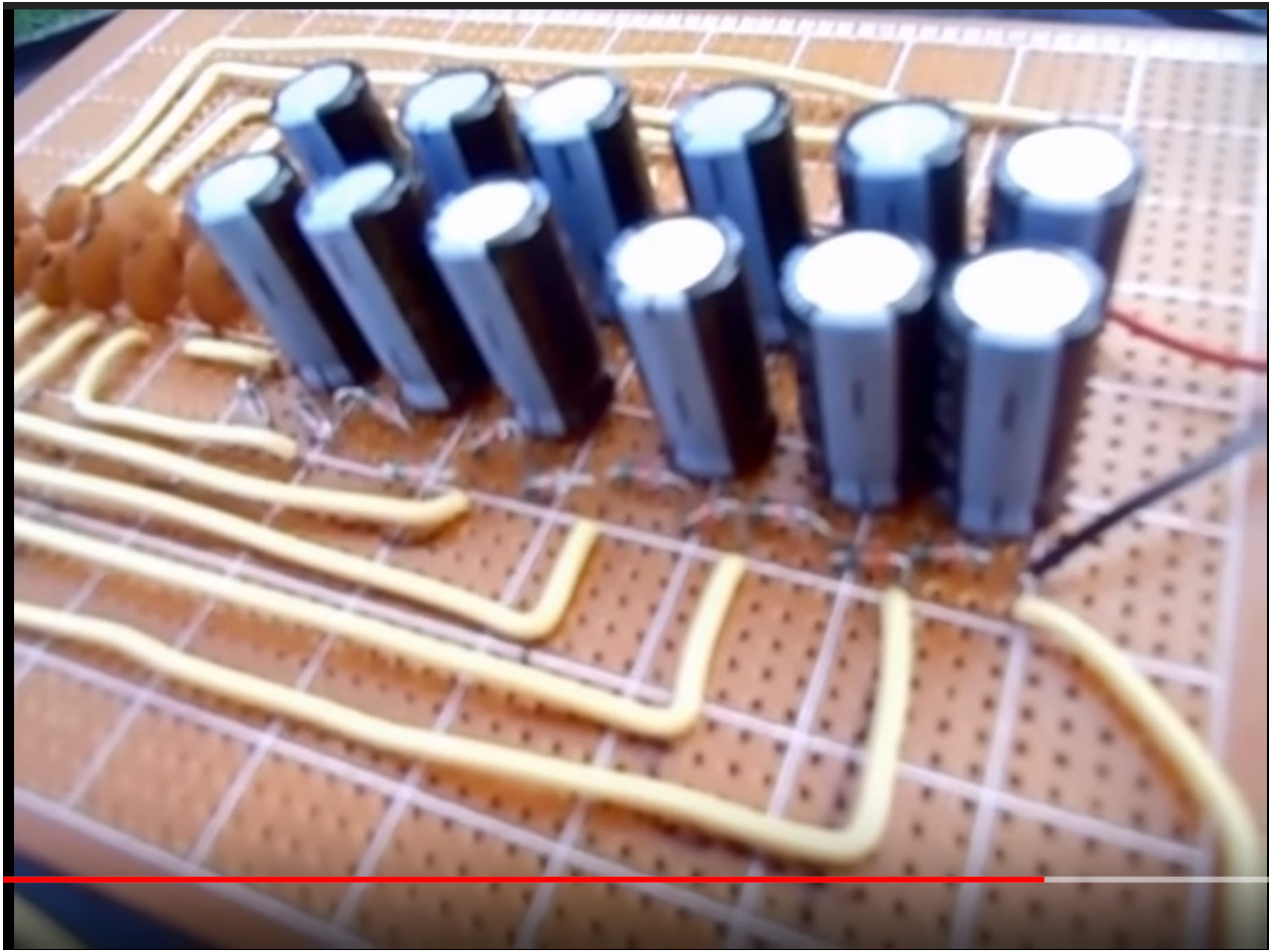

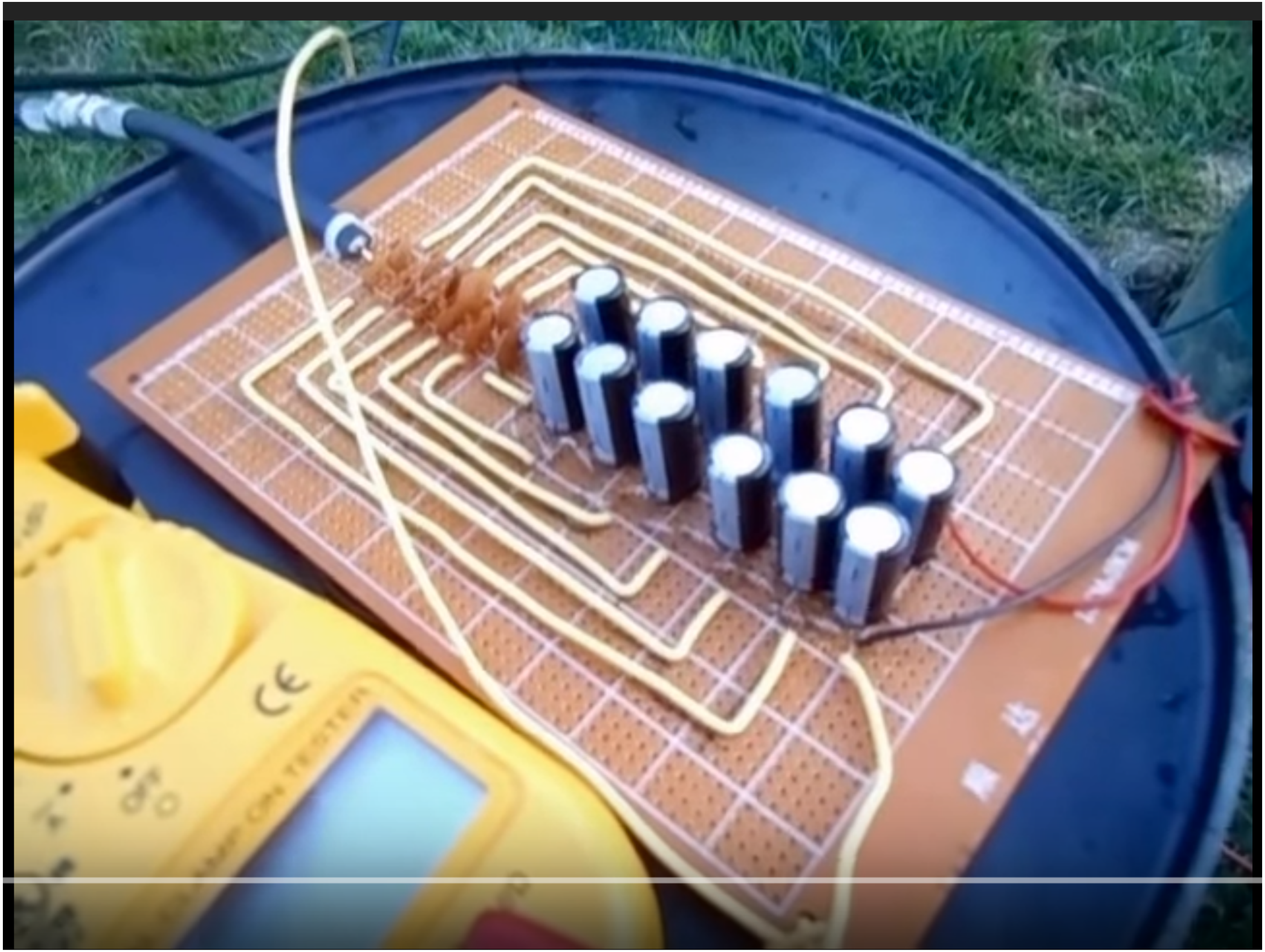
CE

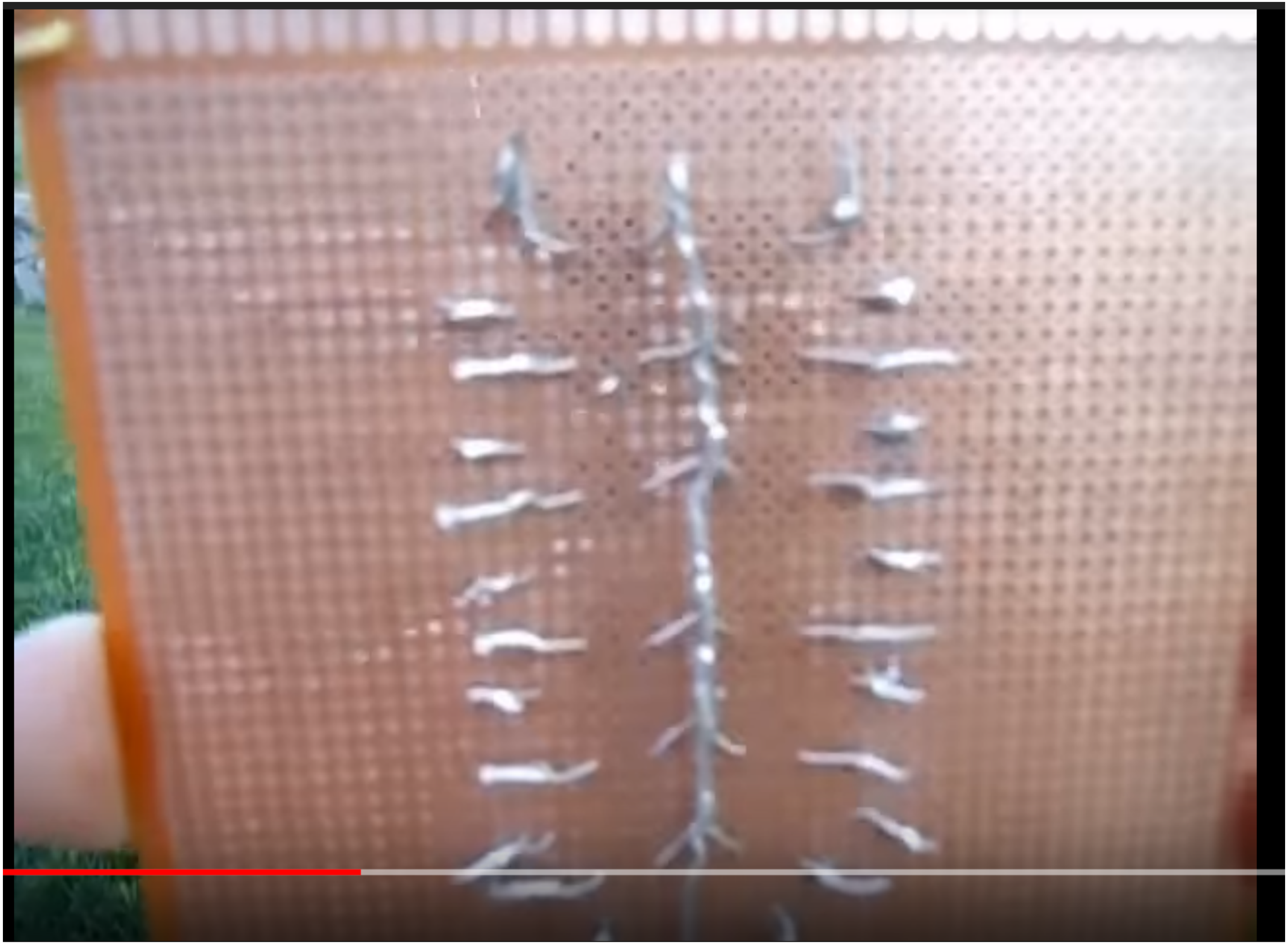

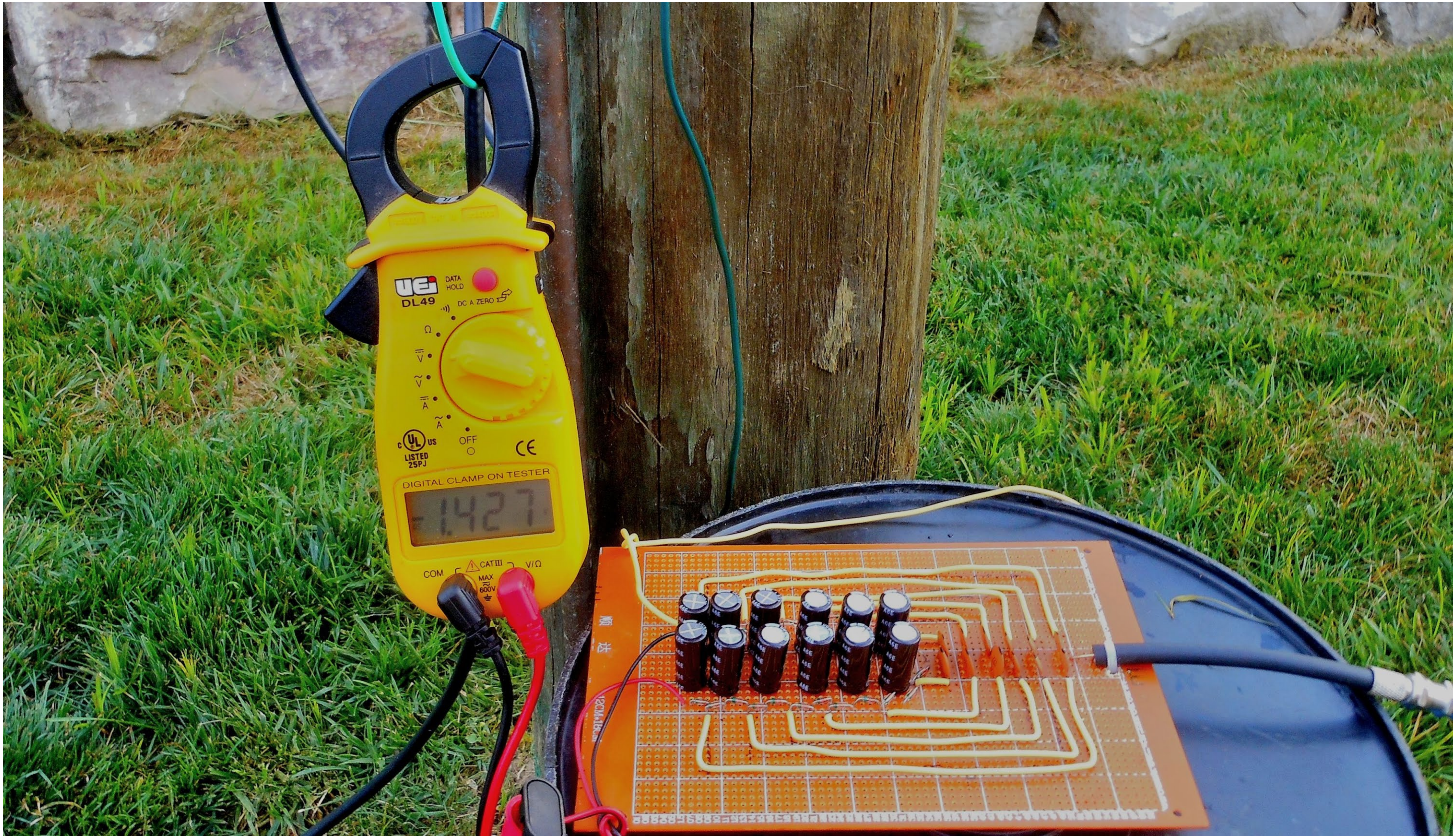
UEi
DL49
DATA HOLD
DC A ZERO
OFF
LISTED 25PJ
CE
DIGITAL CLAMP ON TESTER
-1.427
COM
CAT III
MAX 600V
V/Ω

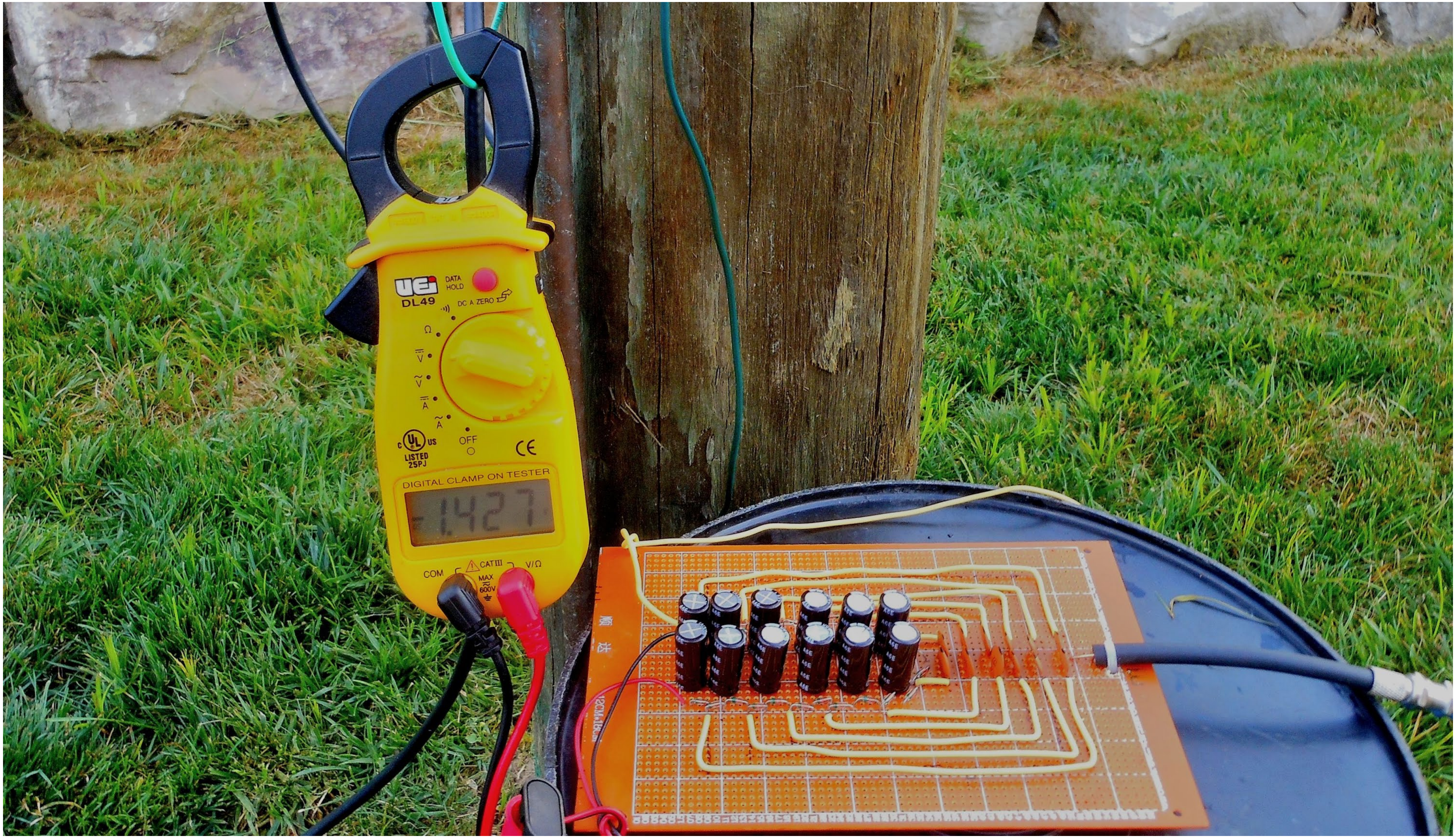
UEi
DL49
DATA HOLD
DC A ZERO
OFF
LISTED 25PJ
DIGITAL CLAMP ON TESTER
-1.427
COM
CAT III
MAX 600V
V/Ω

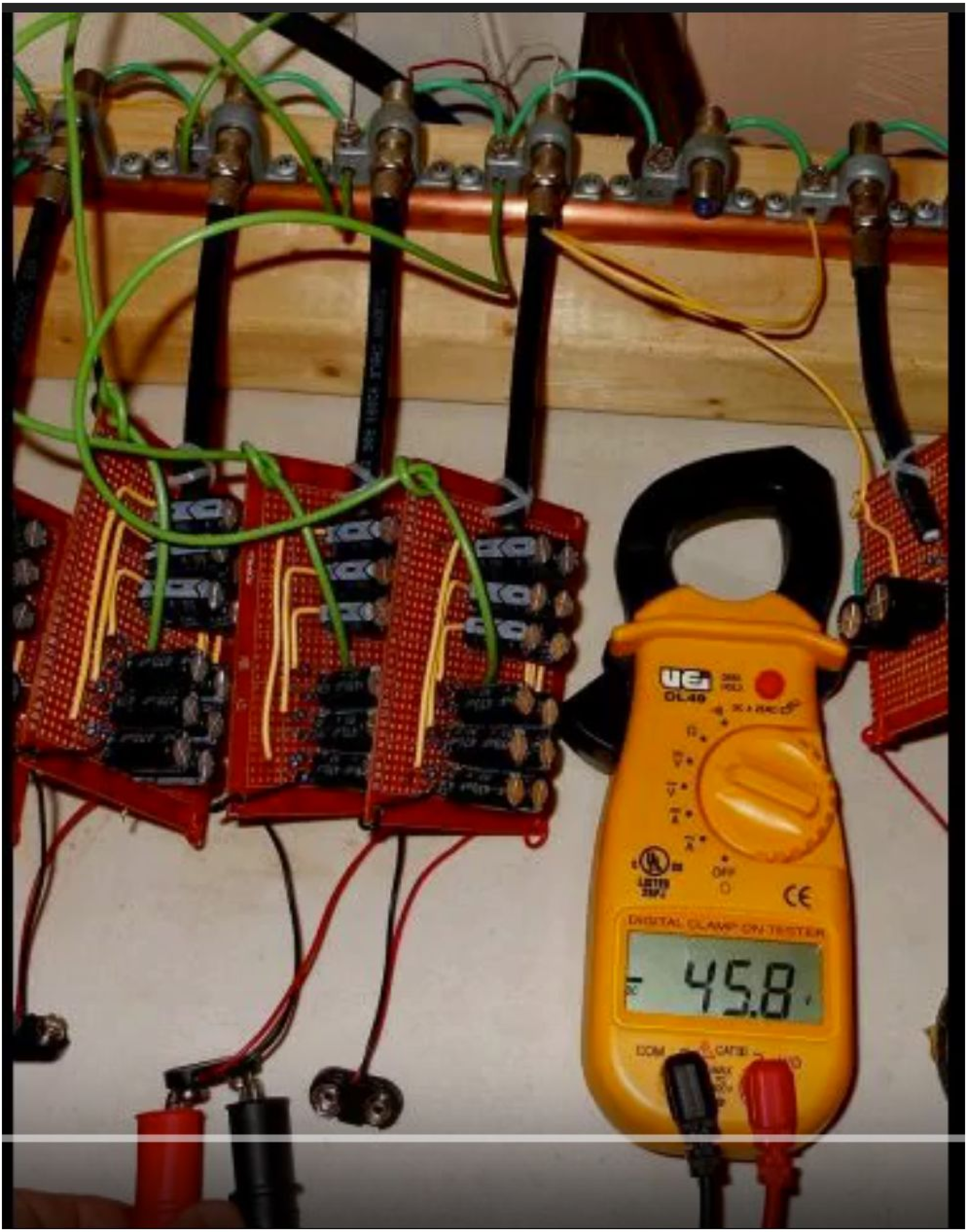
UEi
DL49
OFF
CE
DIGITAL CLAMP ON TESTER
45.8
COM

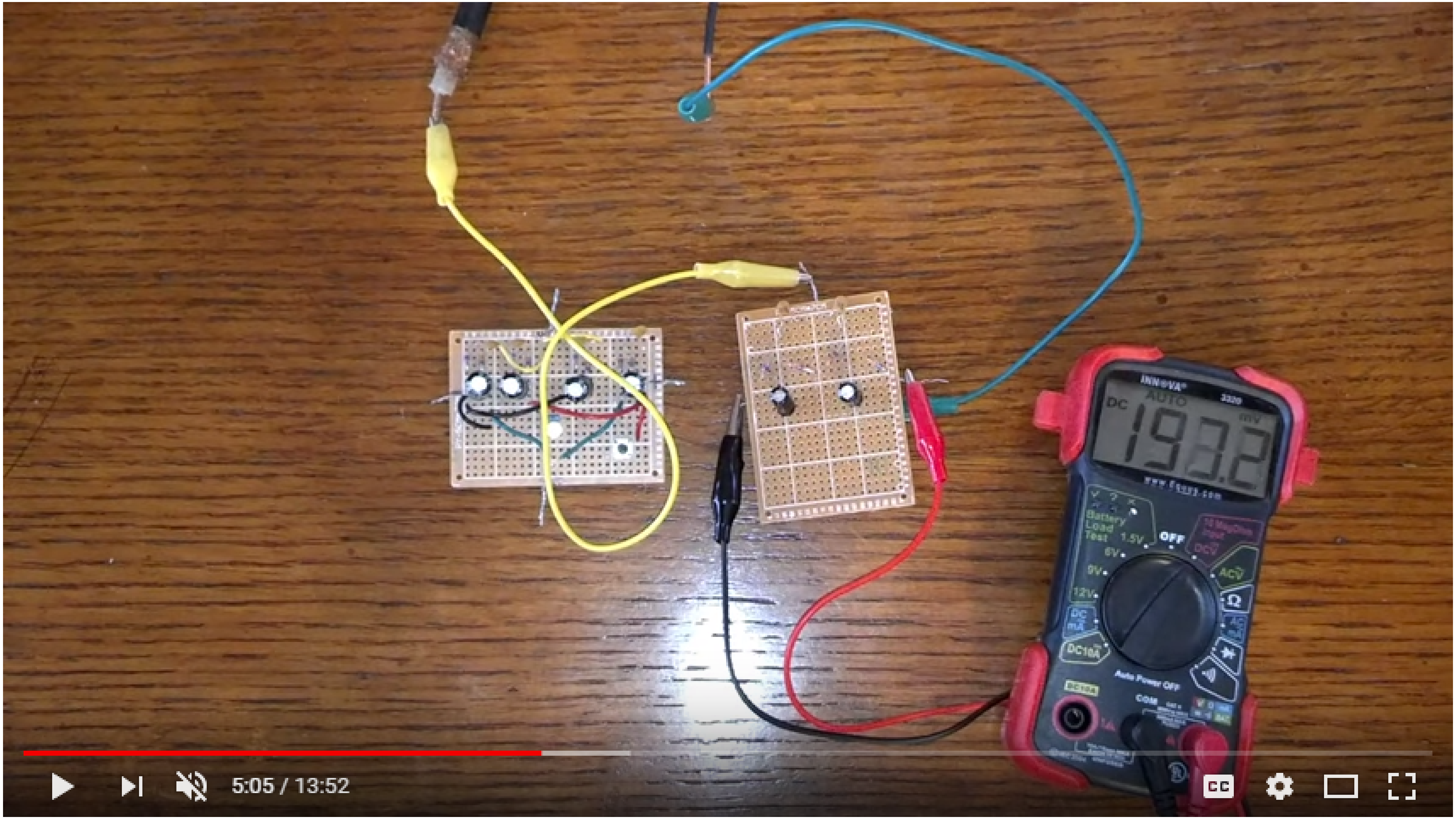
INNOVA
3320
DC
AUTO
193.2
mV
OFF
5:05 / 13:52

Tesla Free Energy Air Circuit

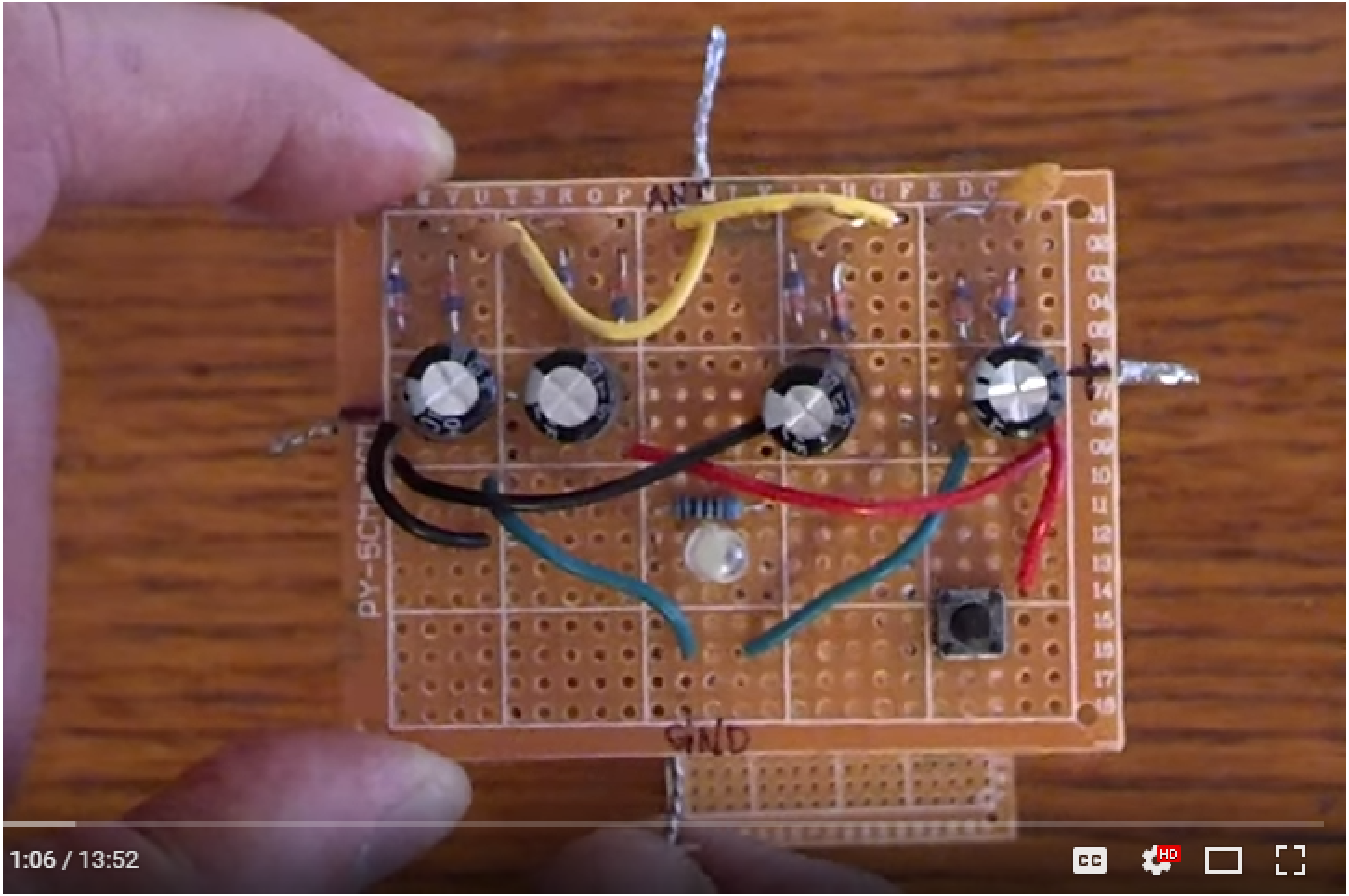
ANT
+
GND
1:06 / 13:52
CC
HD

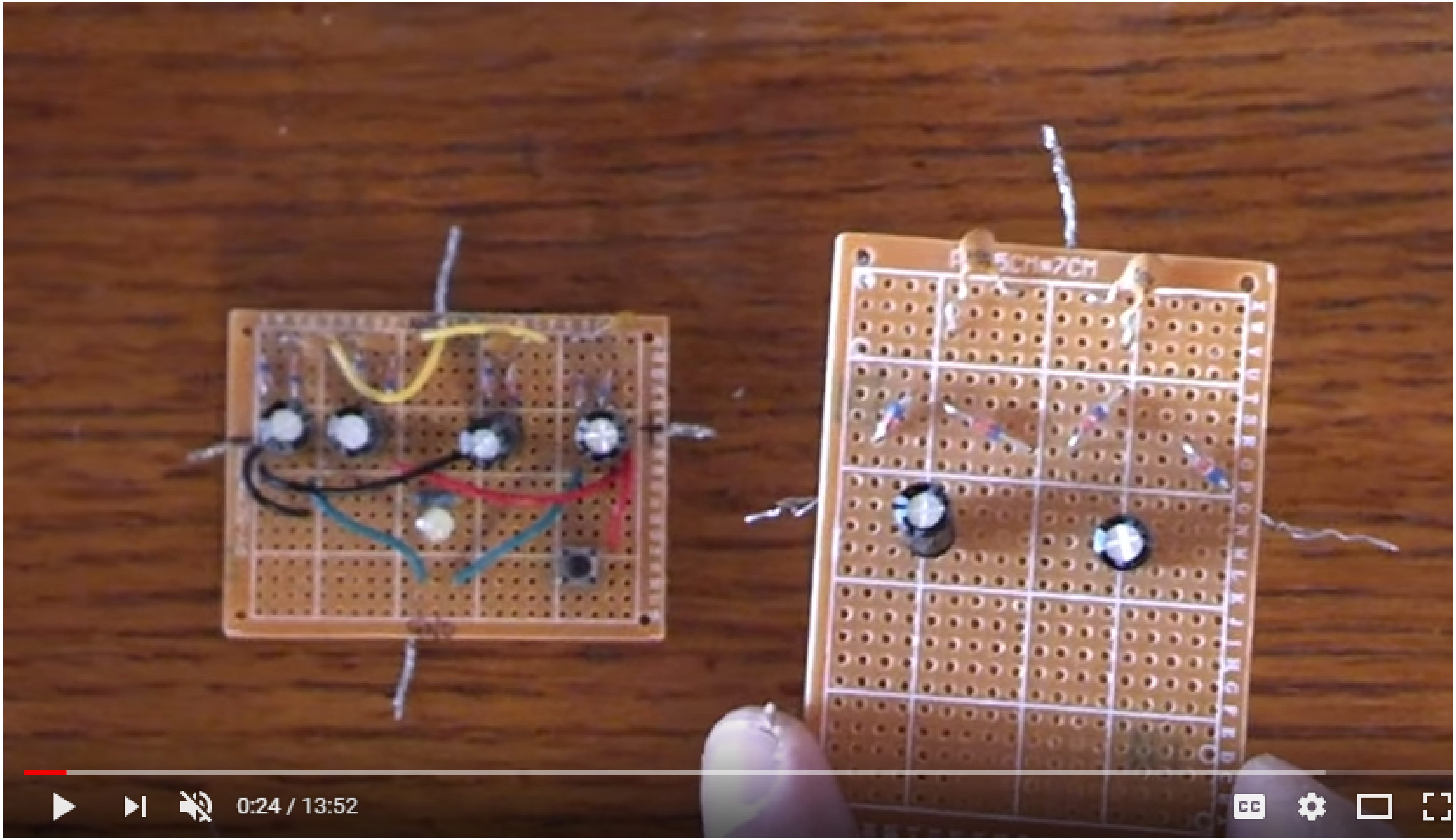
5CM*7CM
0:24 / 13:52

Tesla Free Energy Air Circuit

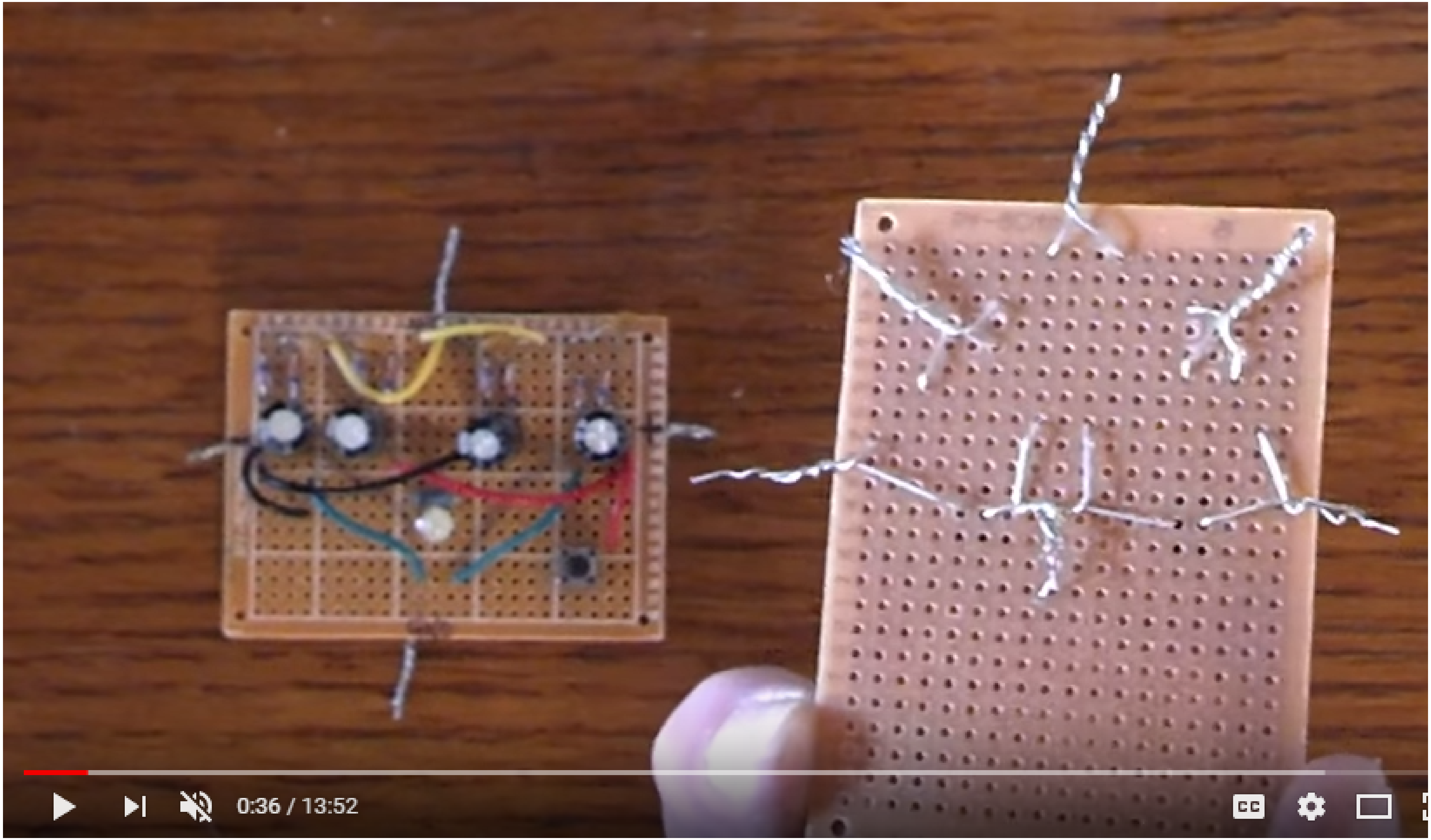

Tesla Free Energy Air Circuit

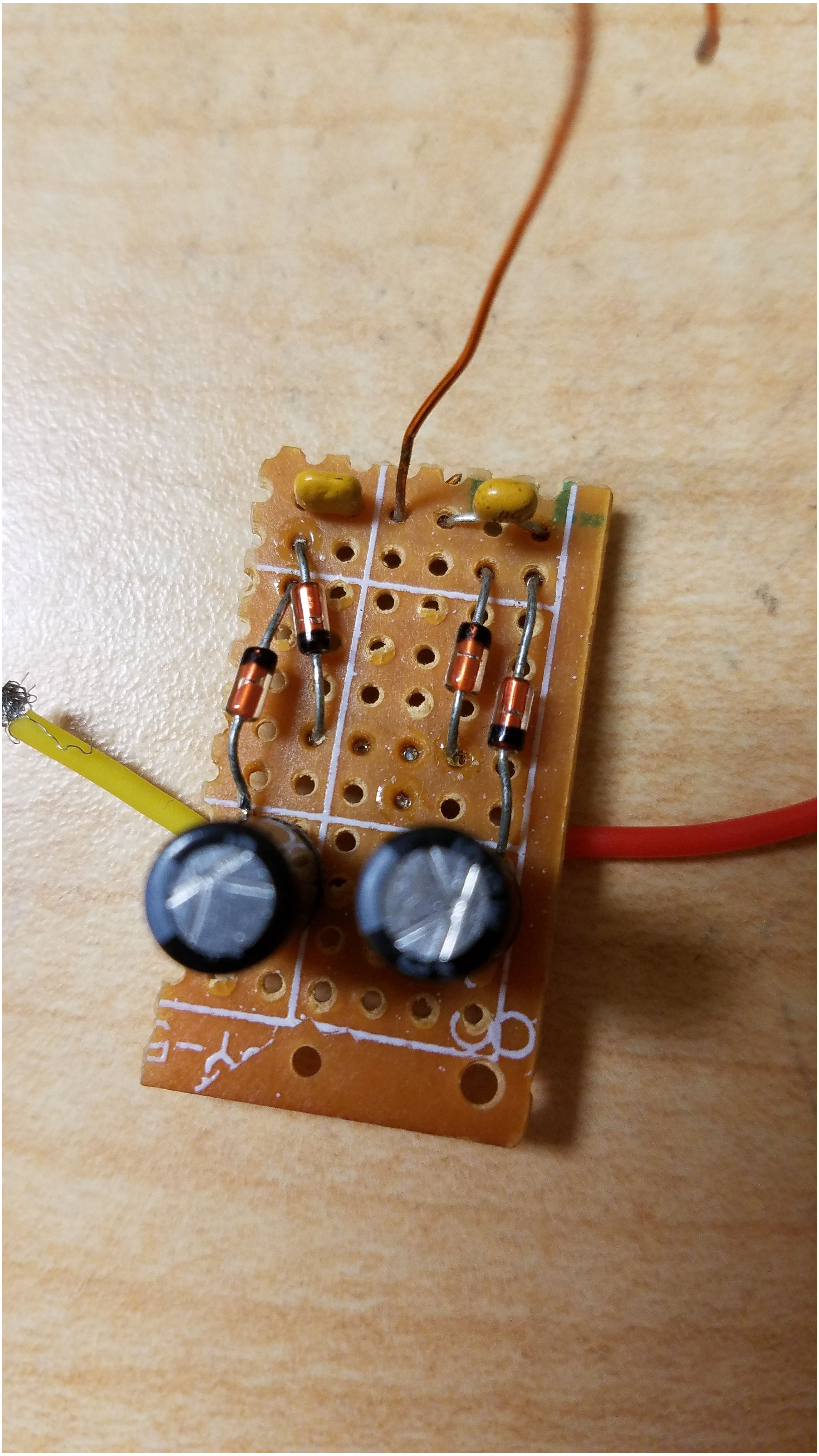

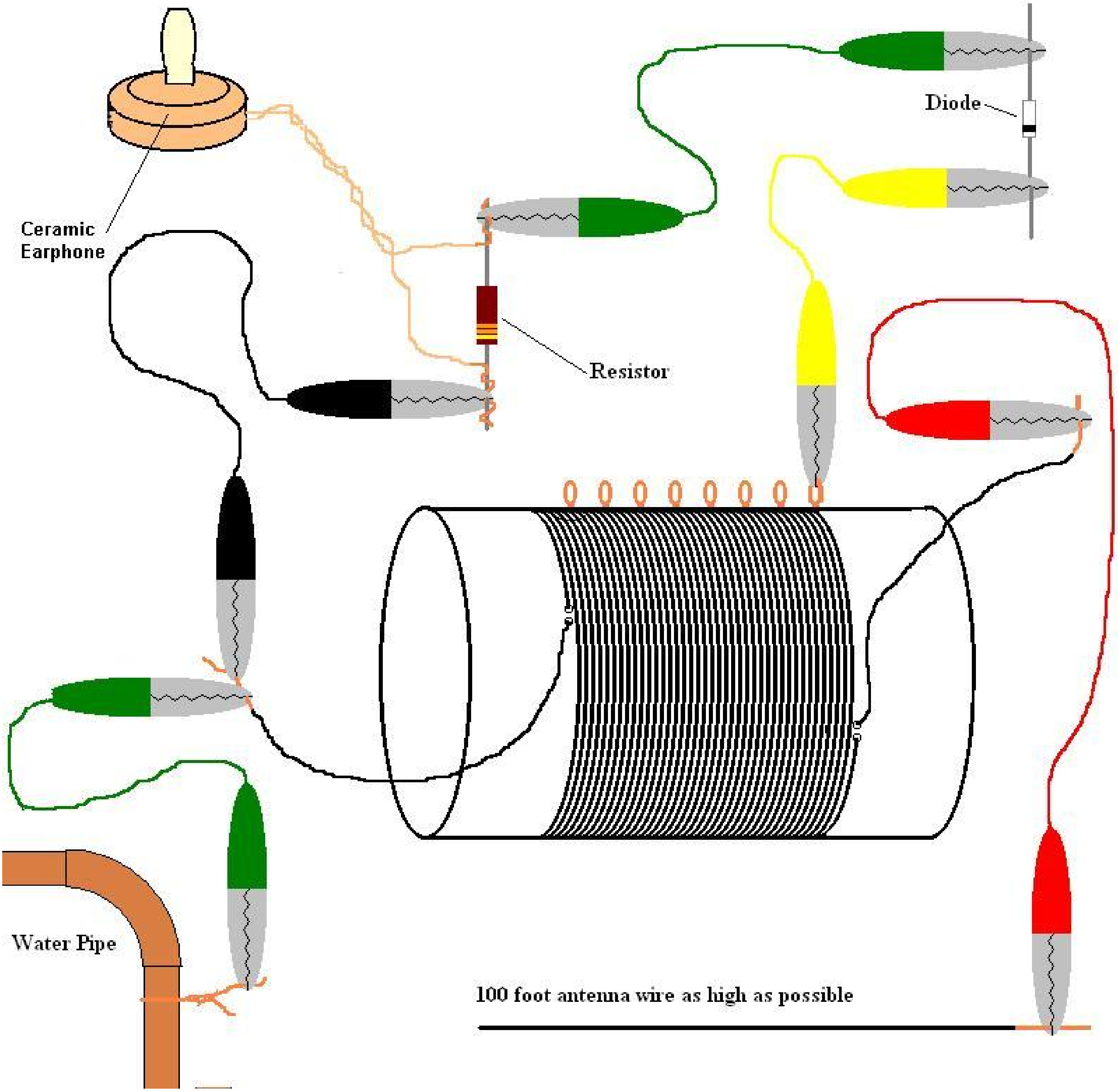
Diode
Ceramic
Earphone
Resistor
Water Pipe
100 foot antenna wire as high as possible

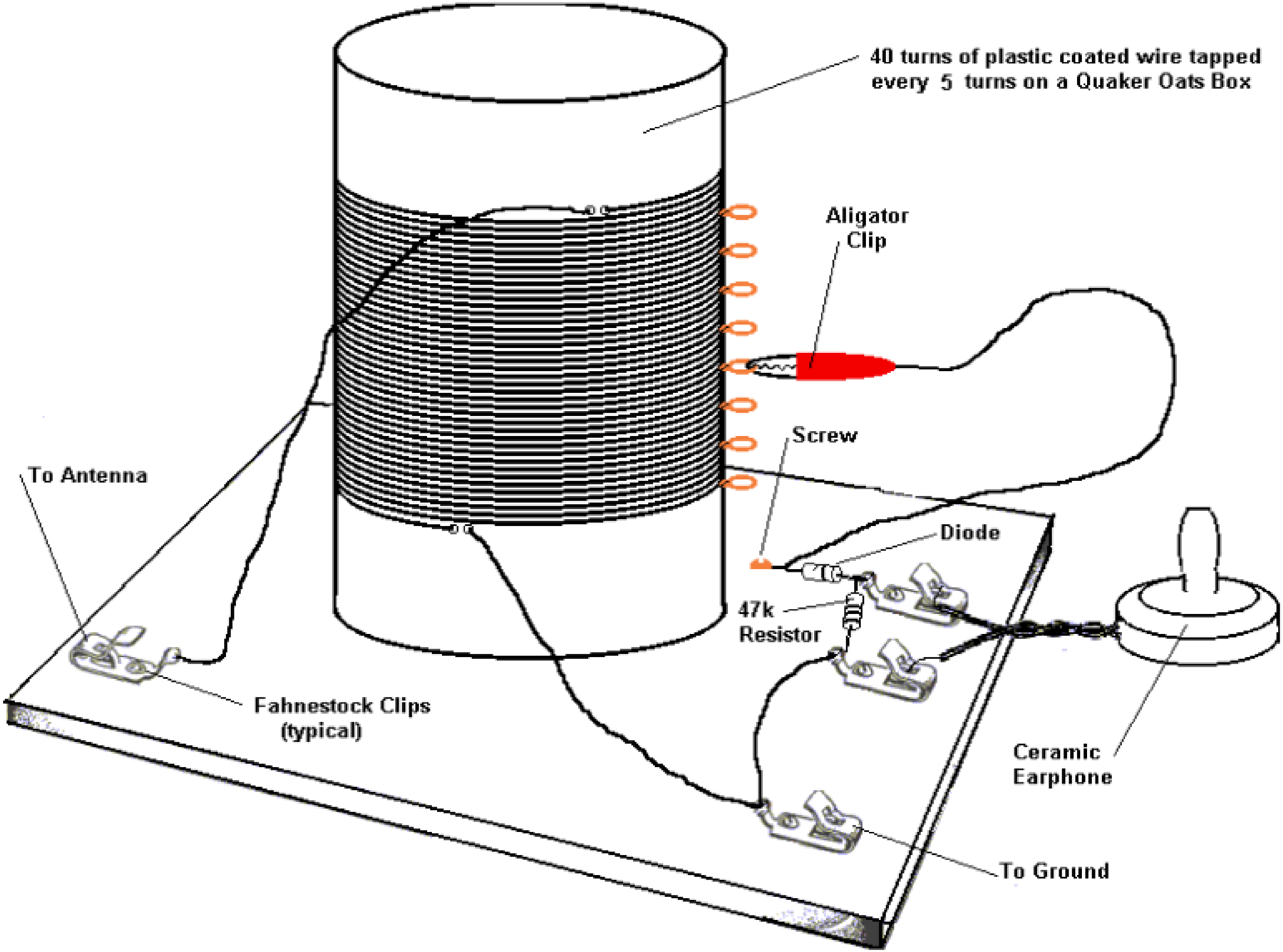
40 turns of plastic coated wire tapped every 5 turns on a Quaker Oats Box
Aligator Clip
Screw
Diode
47k Resistor
To Antenna
Fahnestock Clips (typical)
Ceramic Earphone
To Ground

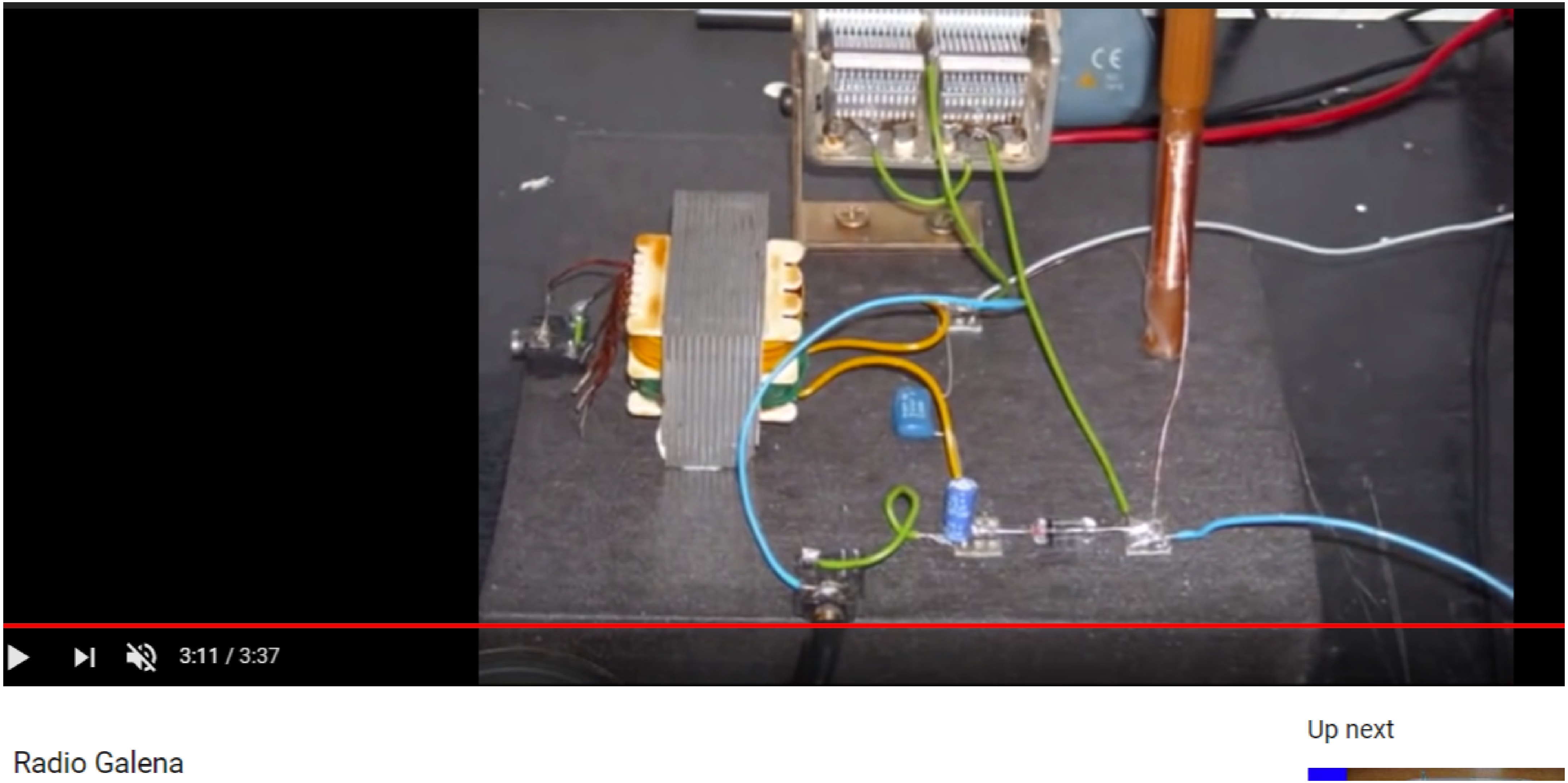
3:11 / 3:37

Radio Galena

Up next

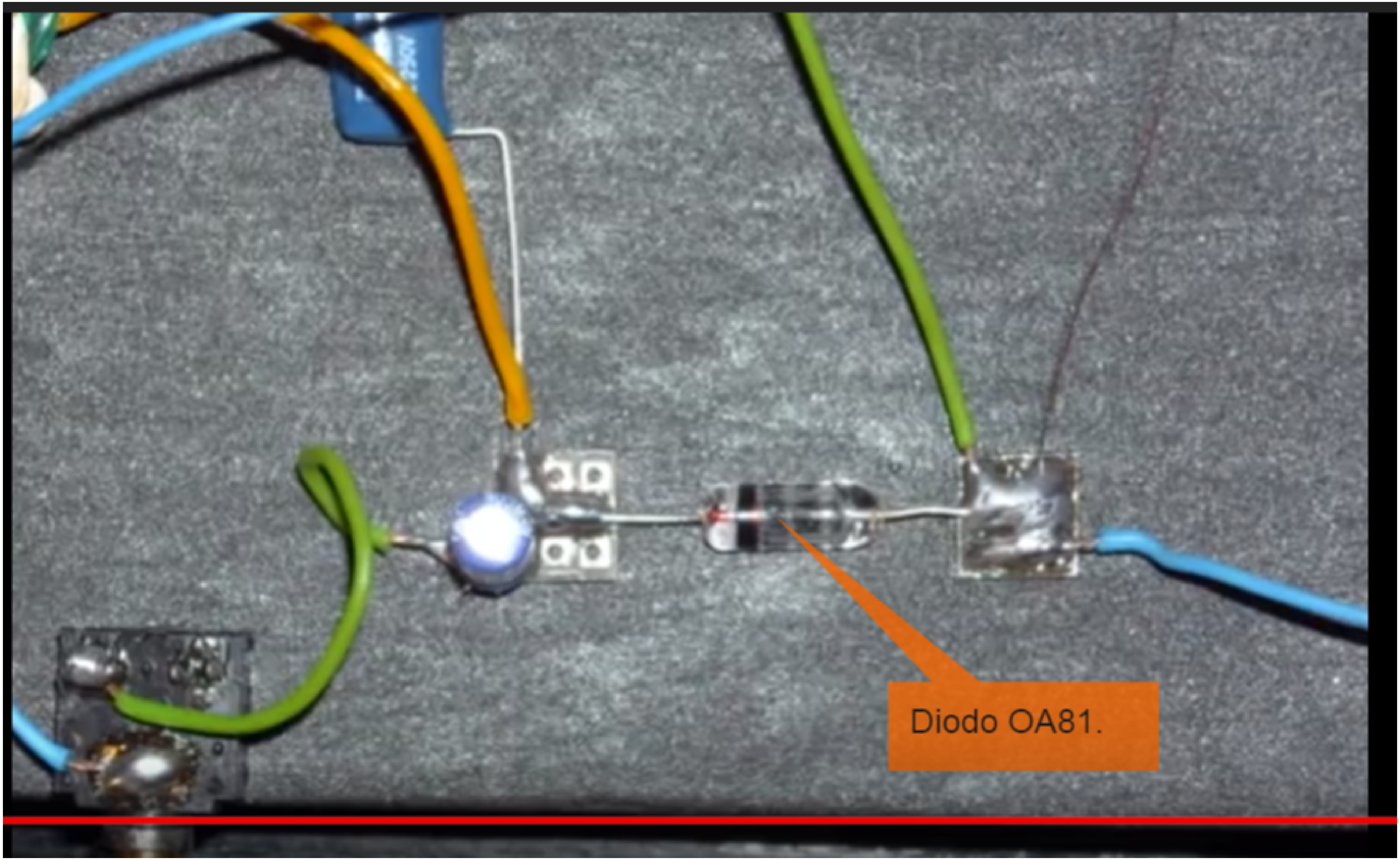
Diodo OA81.

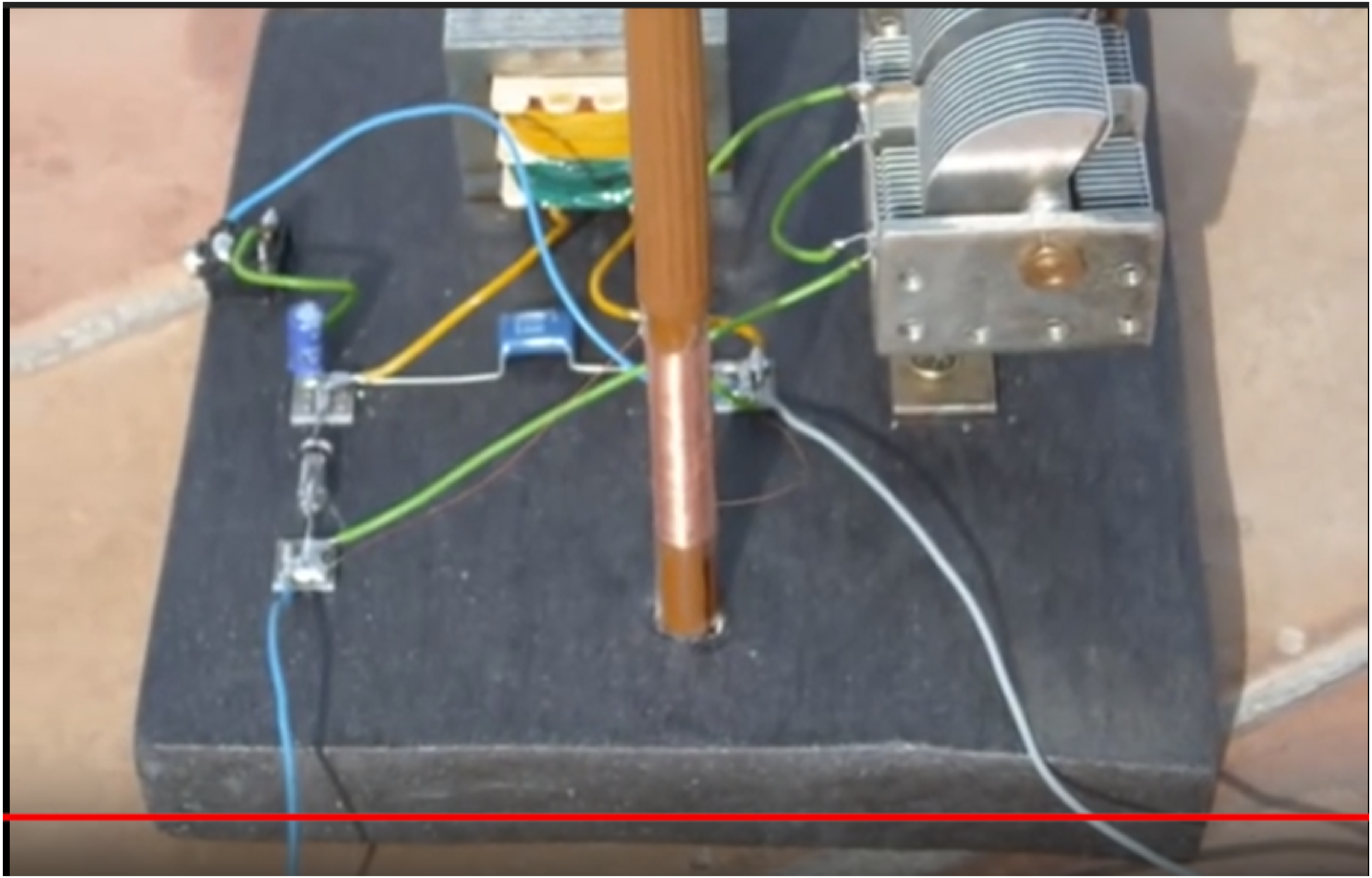

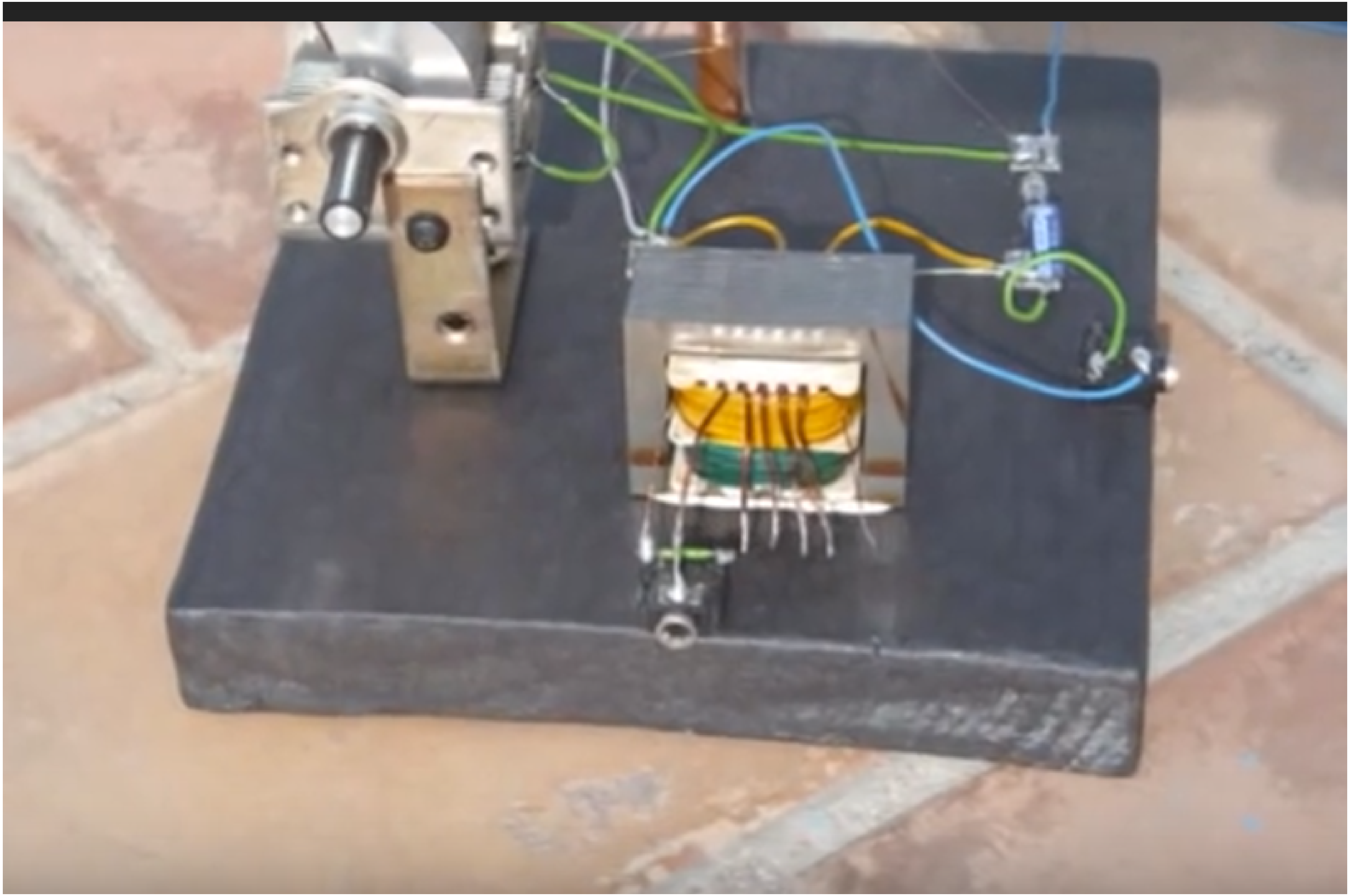

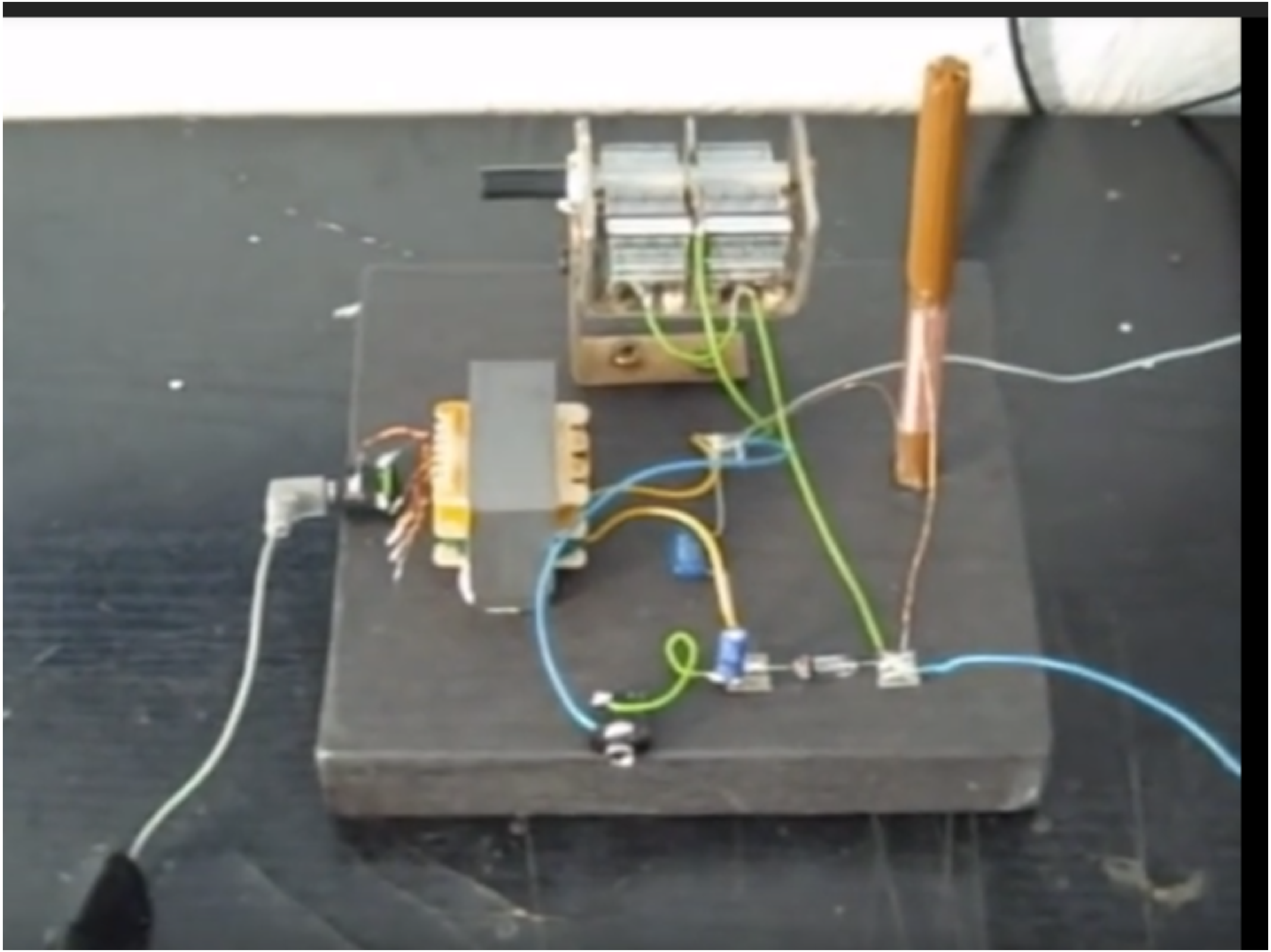

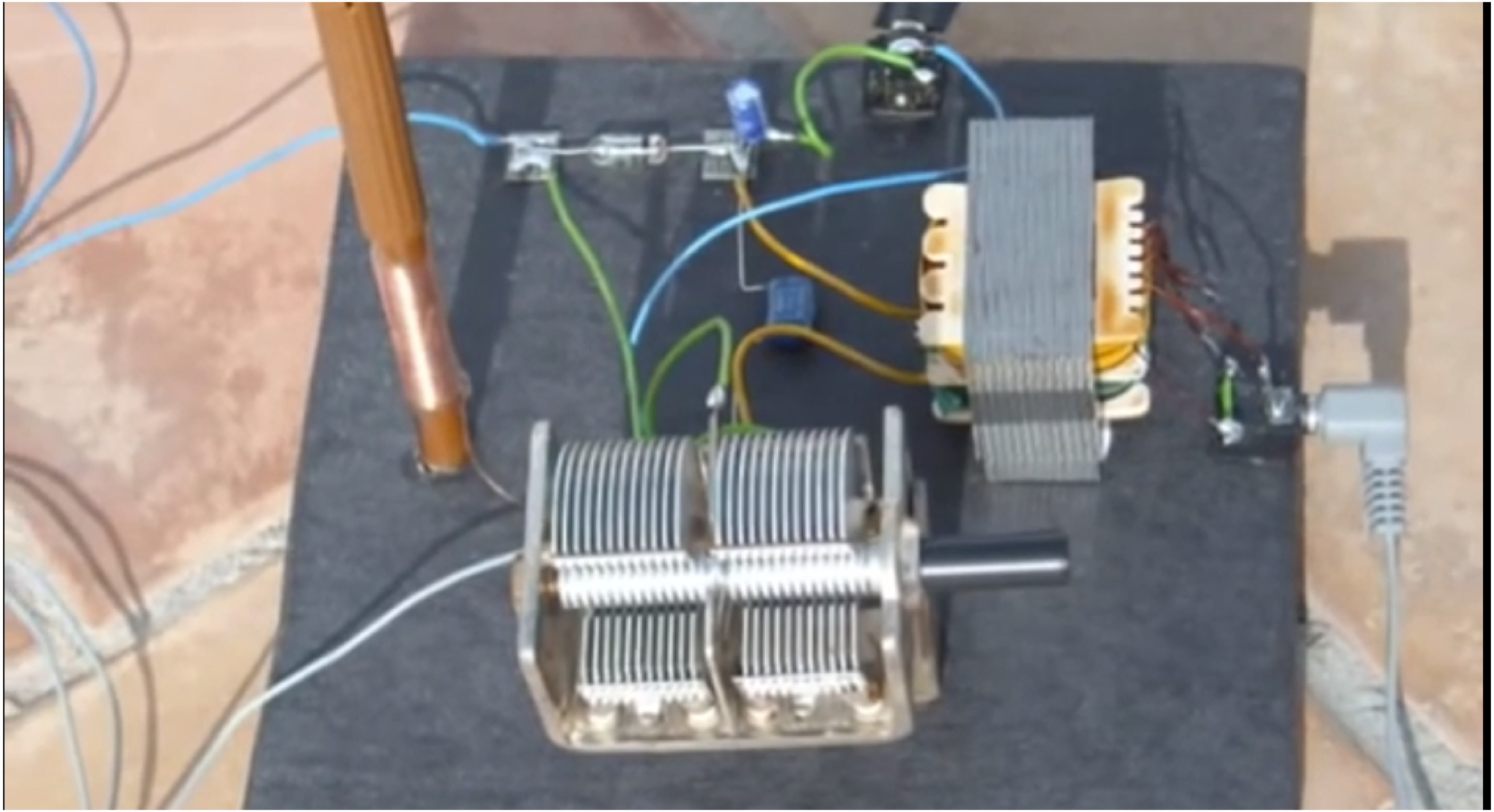

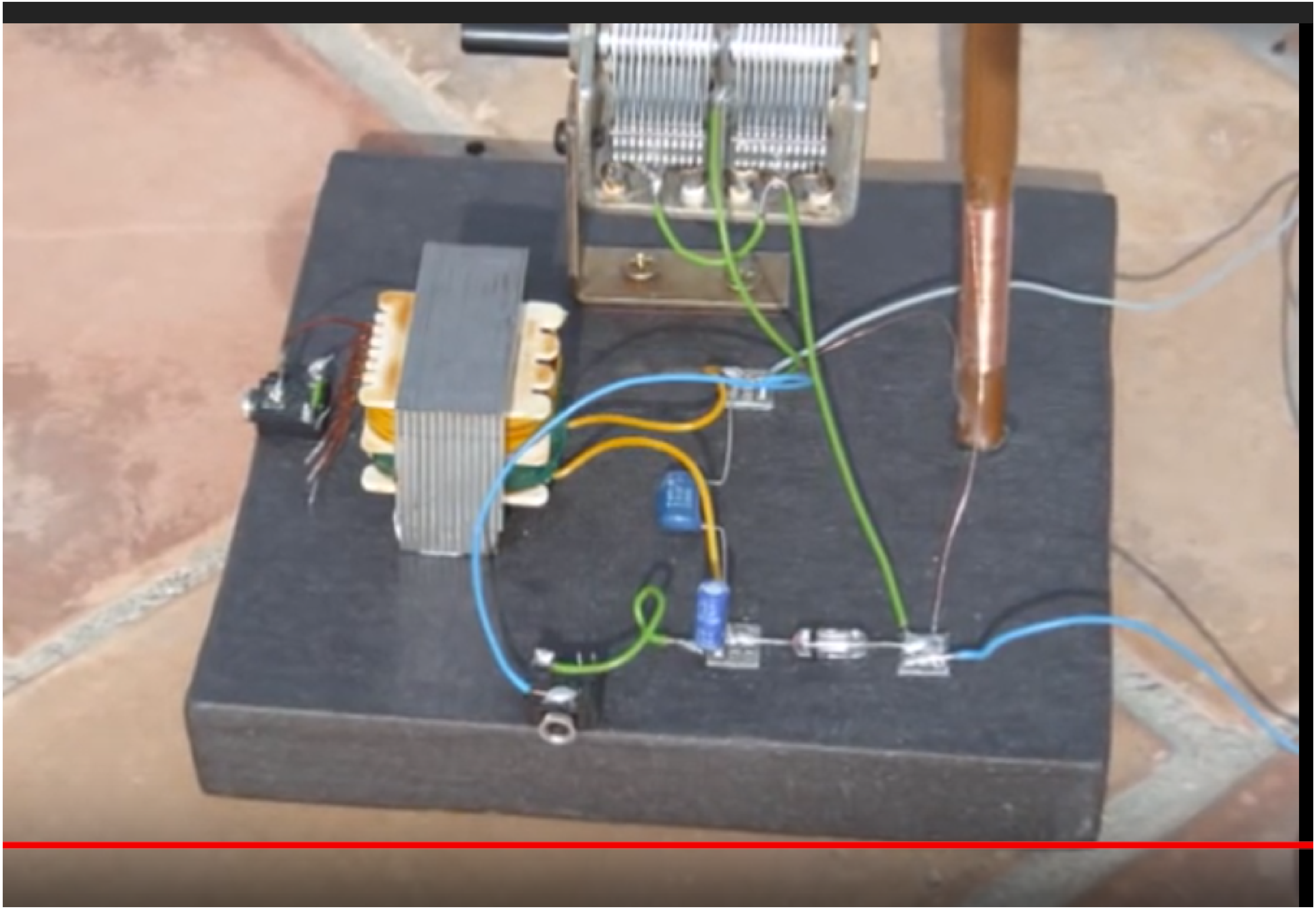

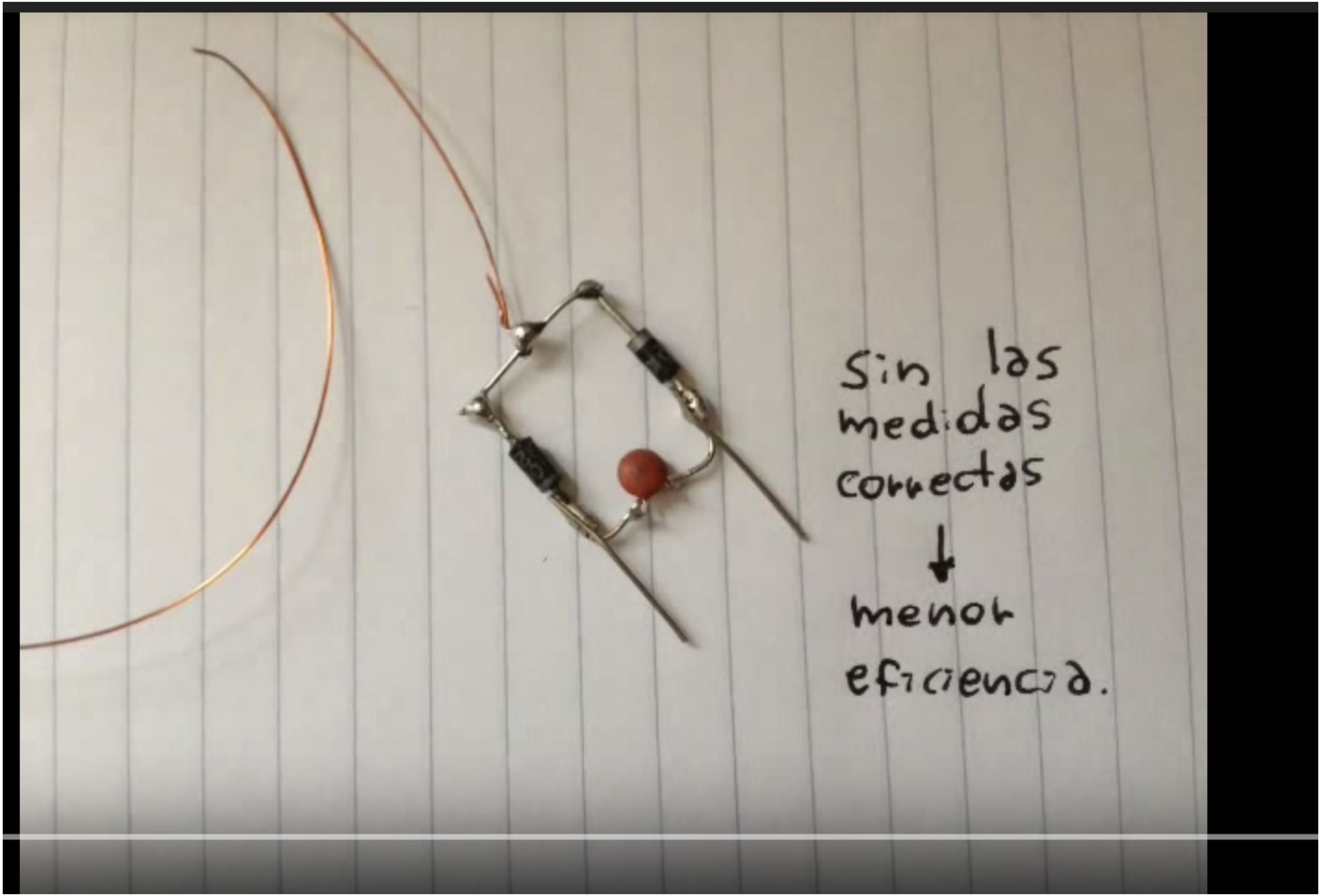

Sin las
medidas
correctas
↓
menor
eficiencia.

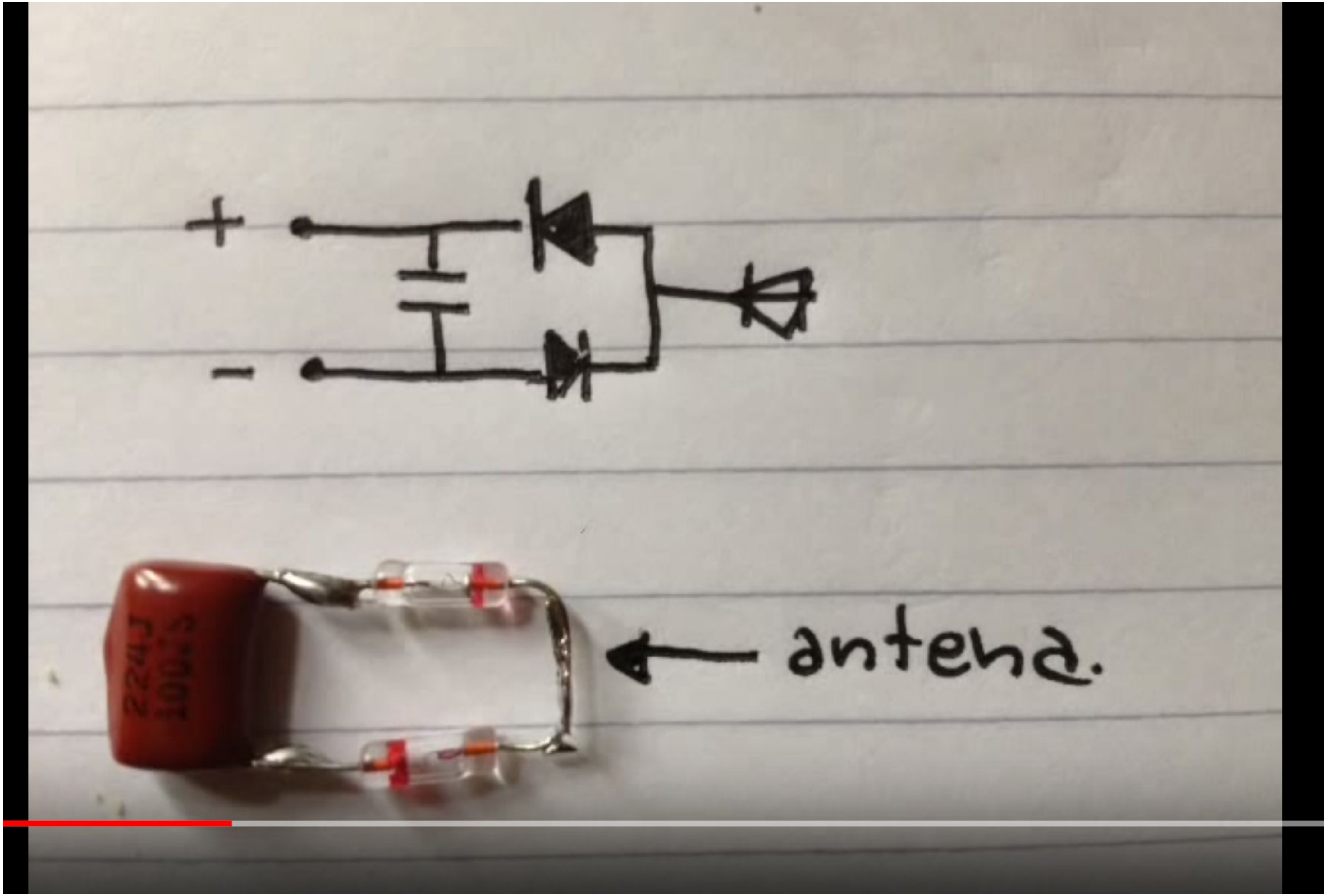
+
-
antena.
224J

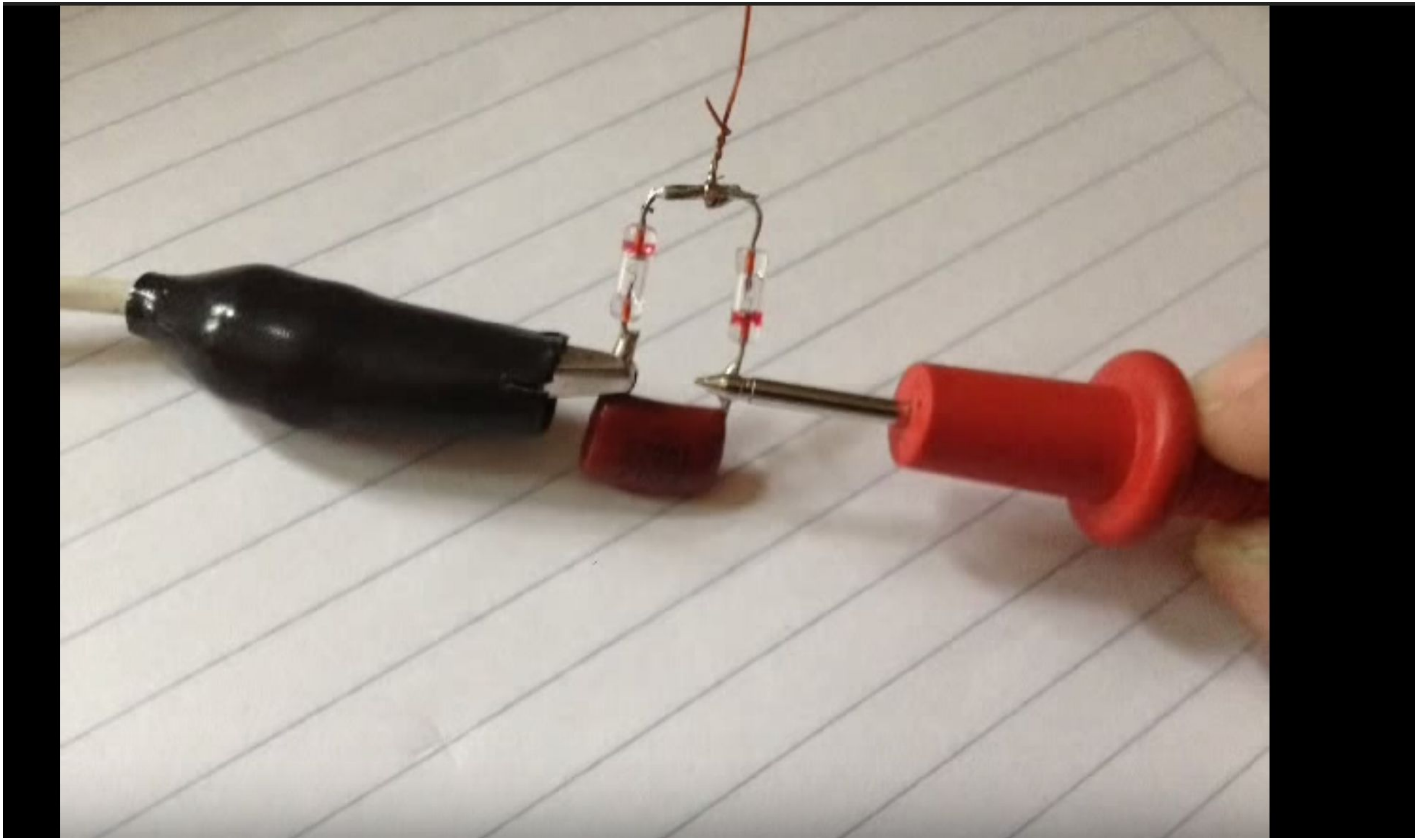

Ricevitore a cristallo per FM

Un progetto di Giacomo Cavuoti

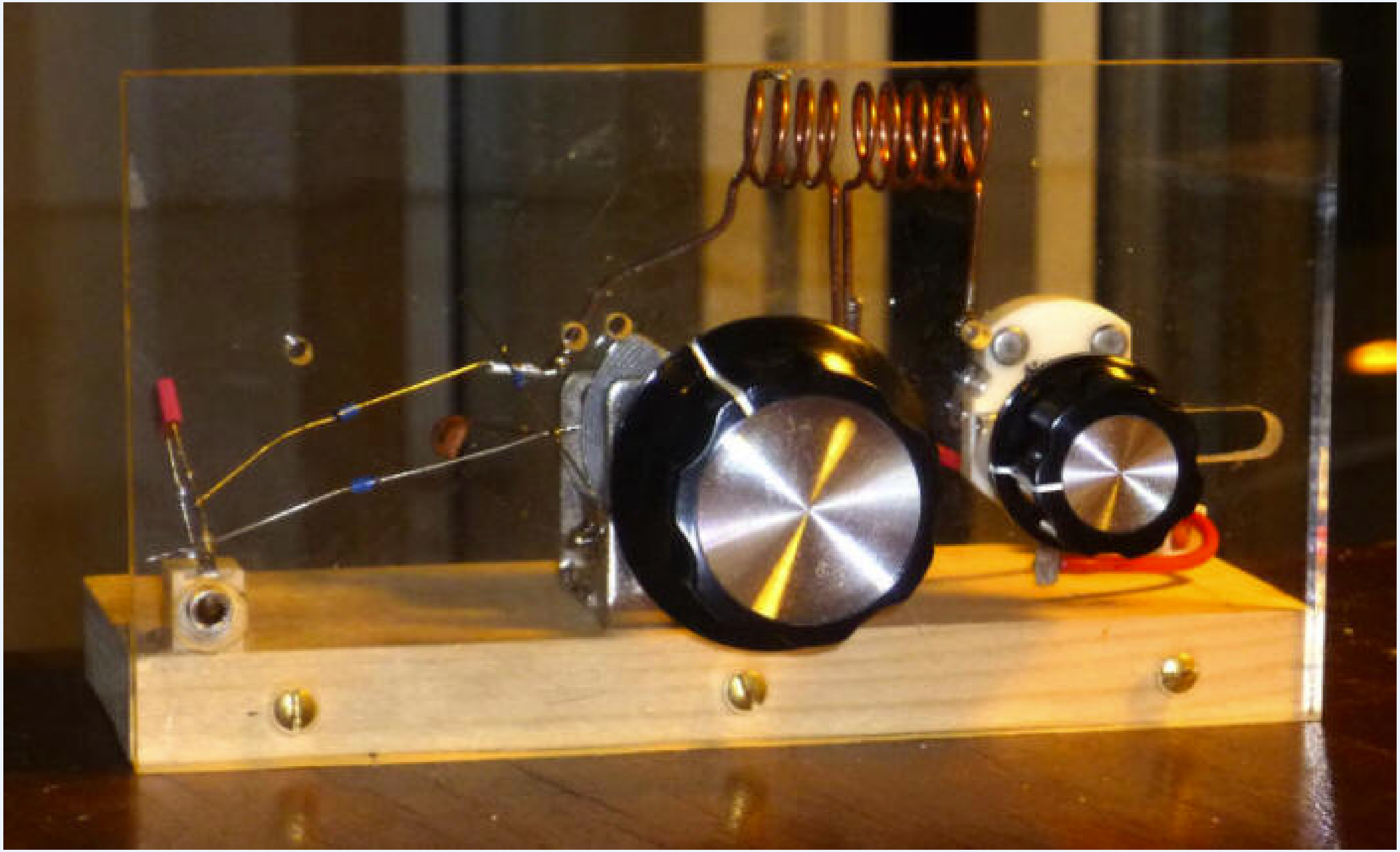

Seguendo l'invito di Leonardo ho realizzato questo semplice ricevitore a cristallo per FM (fare clic sullo schema qua sotto per vederlo ingrandito).

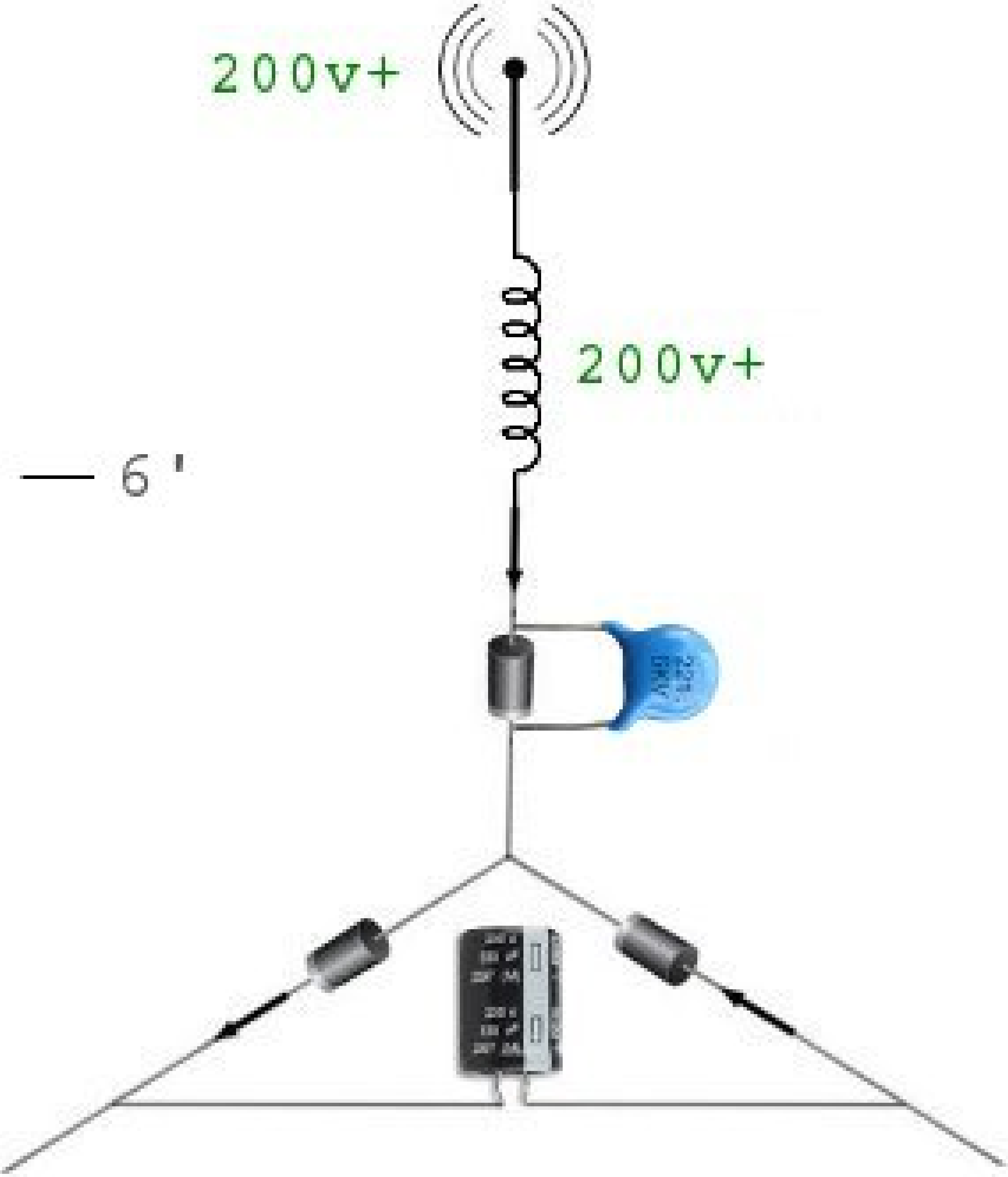
200v+
200v+
— 6 '

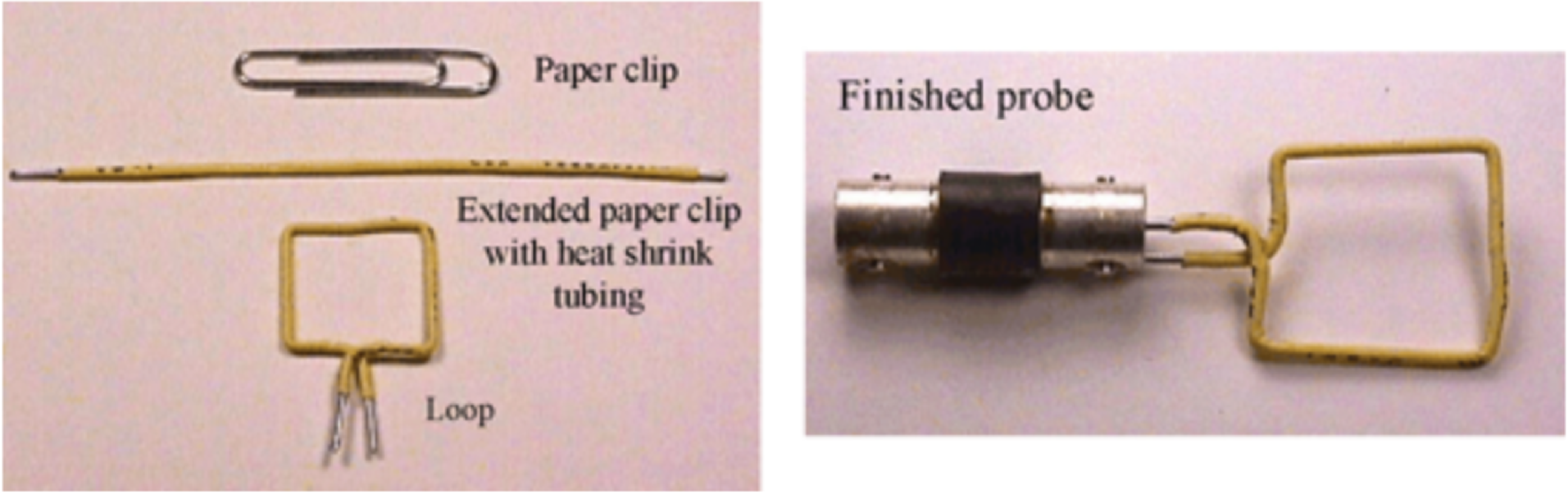
Paper clip
Extended paper clip with heat shrink tubing
Loop
Finished probe

Figure 1: Magnetic field probe build from a paper clip.

YAESU
HOME
SEL
CLAR
VFOa CW
14.054.00 R
SW
A
B
C
F
V/M
LOCK
PWR
AF SQL/RF
ANT
C-POISE
1:54 / 2:18
HD

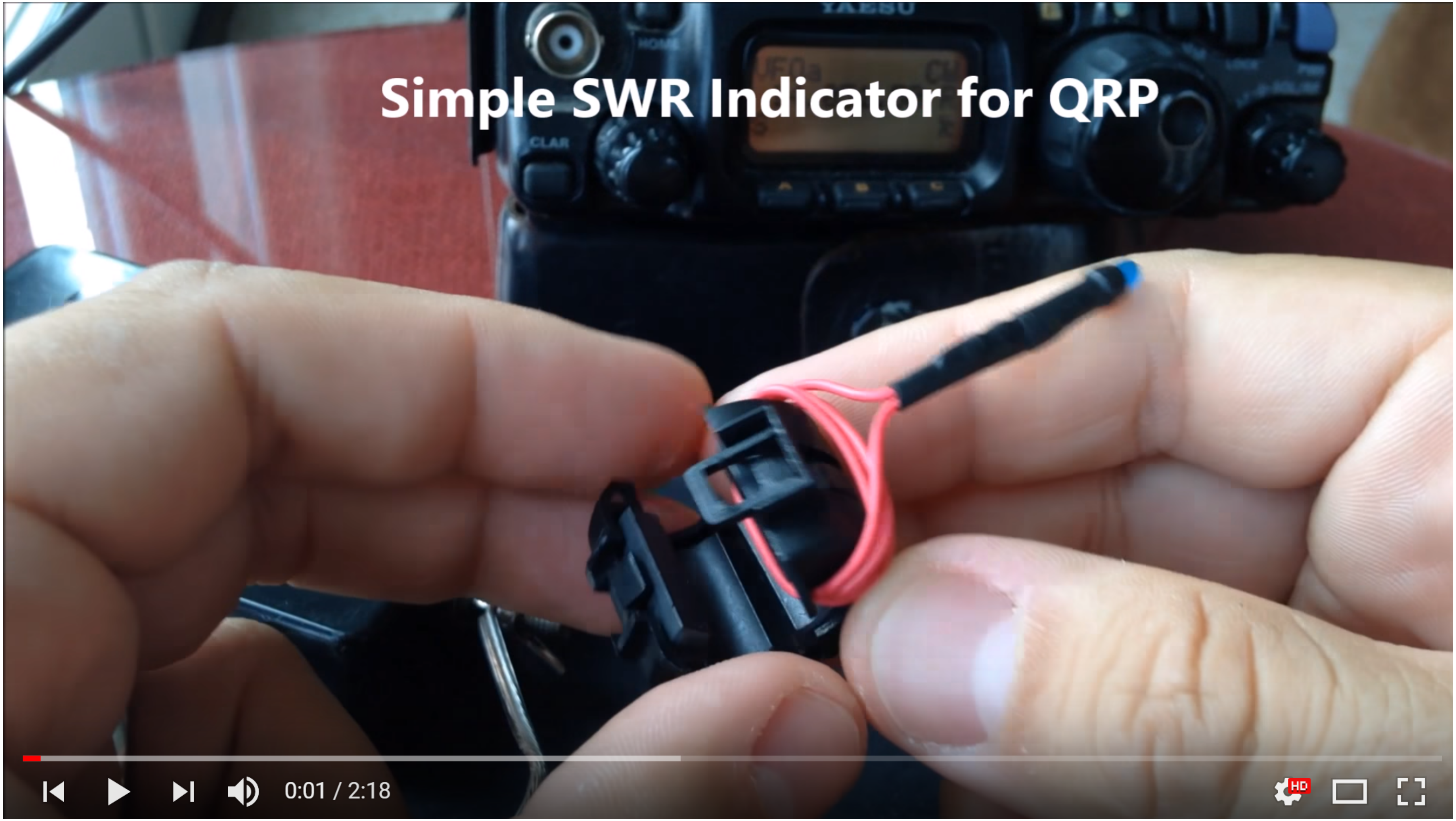

YAESU
Simple SWR Indicator for QRP
CLAR
0:01 / 2:18
HD

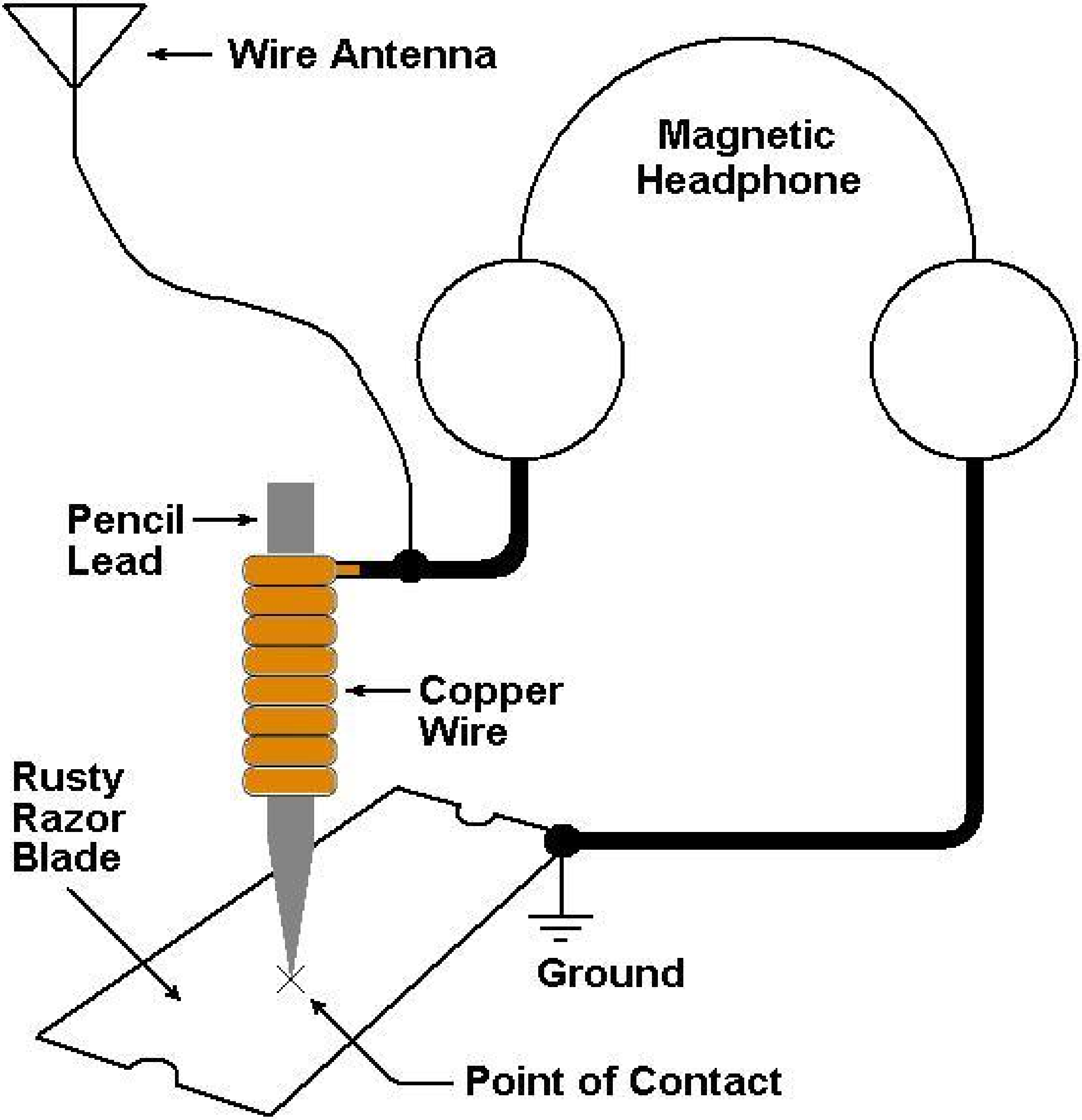
Wire Antenna
Magnetic Headphone
Pencil Lead
Copper Wire
Rusty Razor Blade
Ground
Point of Contact

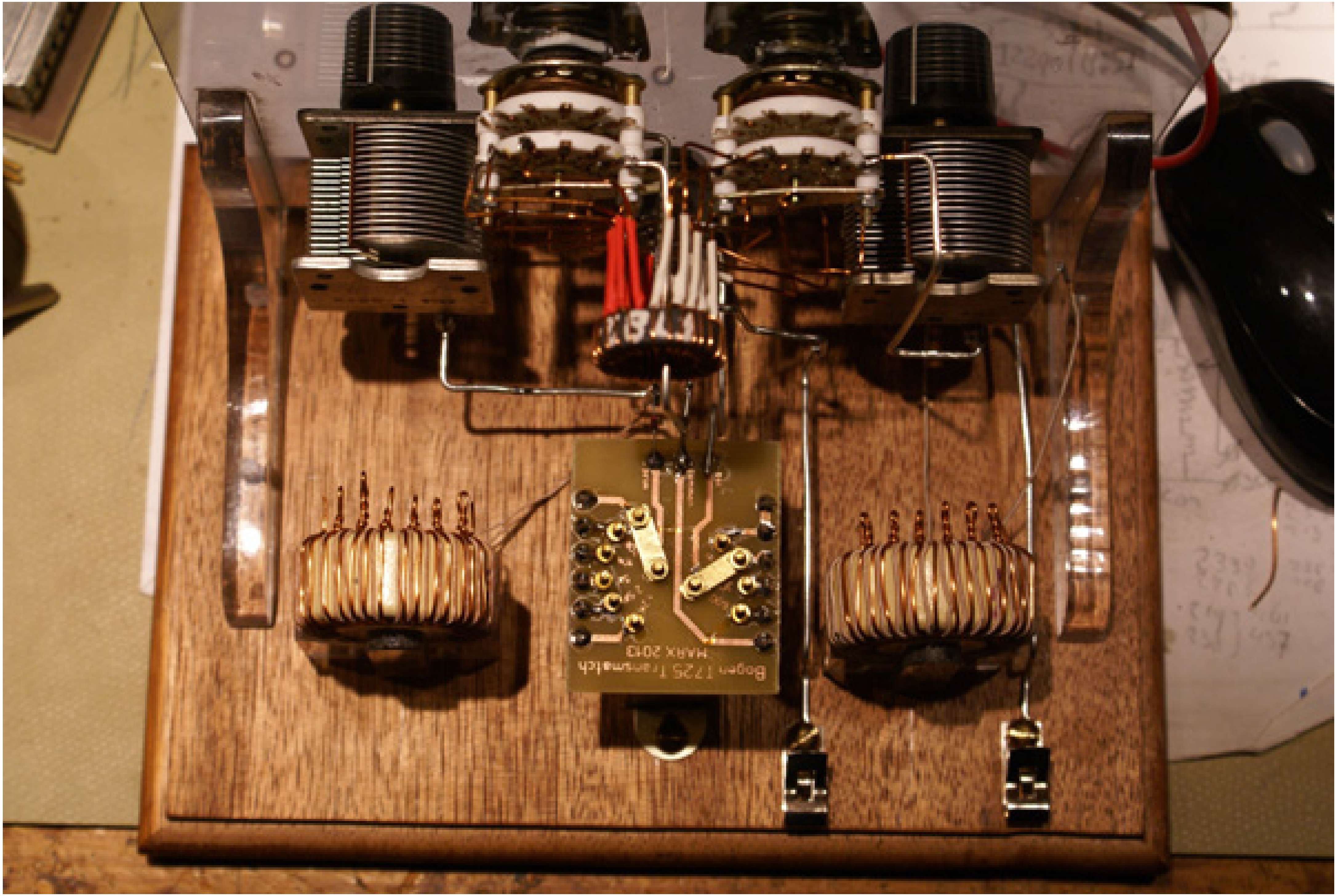
Bogen T725 Transmatch
MARX 2013

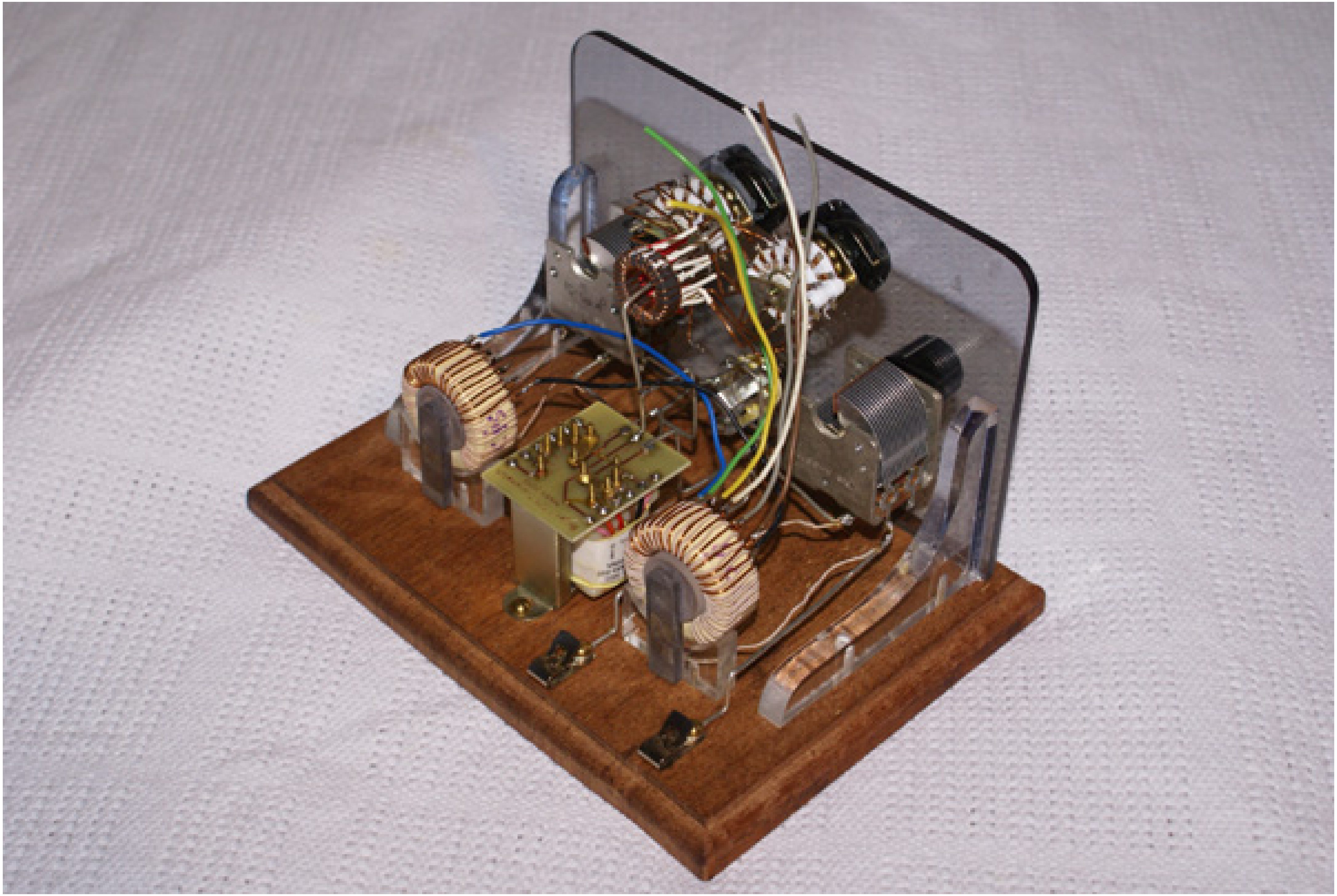

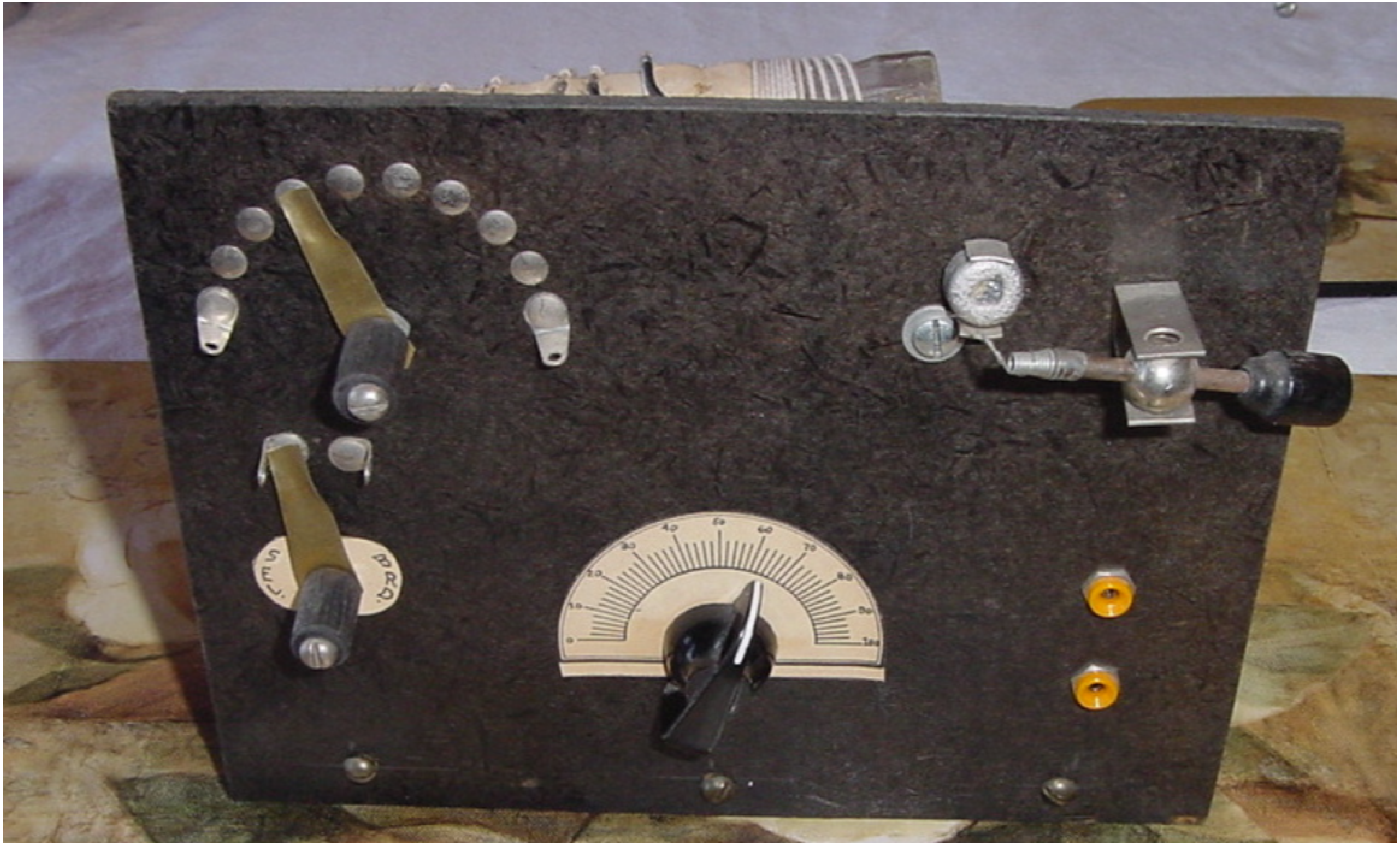
SEL.
BRD.
0
10
20
30
40
50
60
70
80
90
100

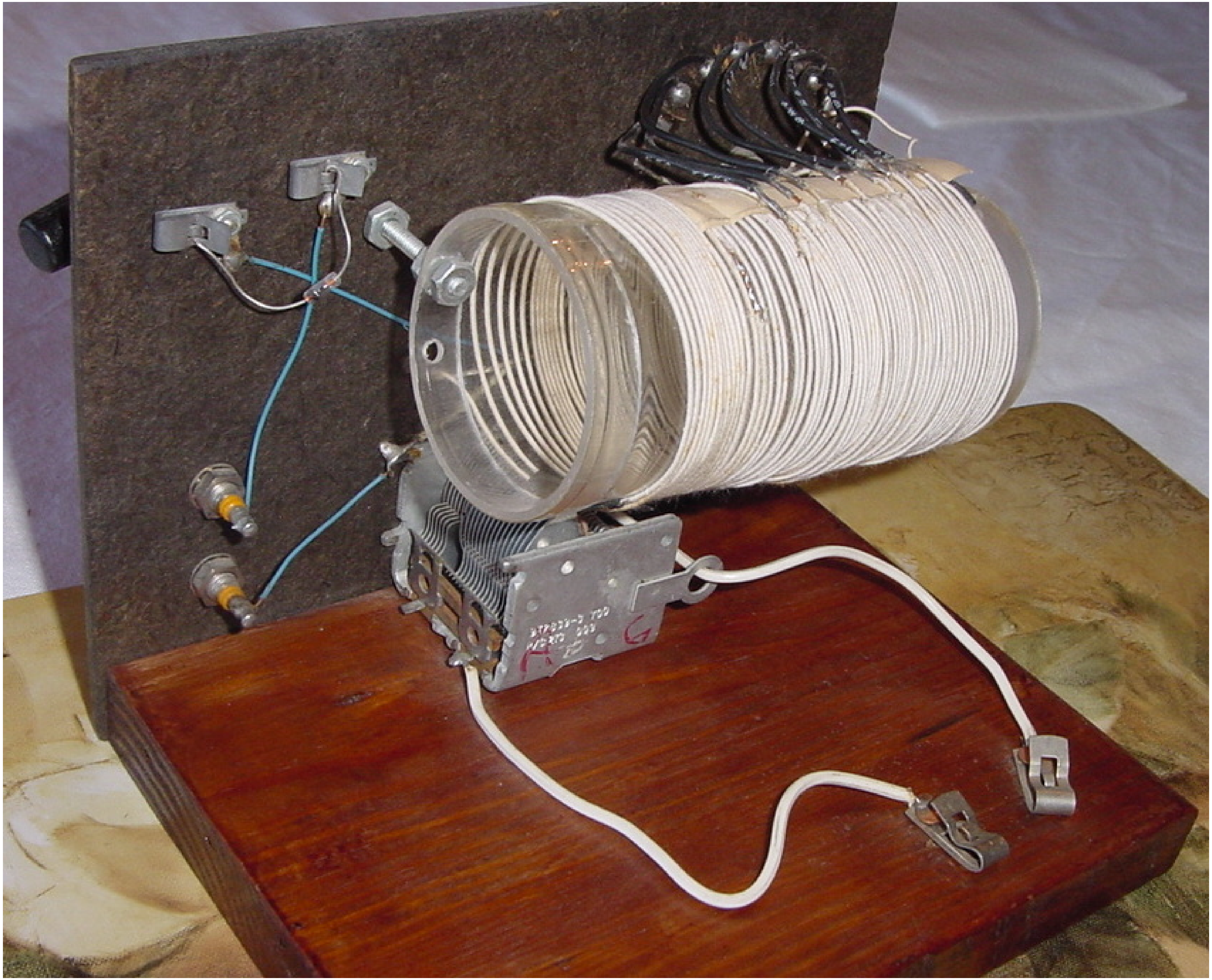
A
G

Another Prototype similar to the Heathkit using toroid coil

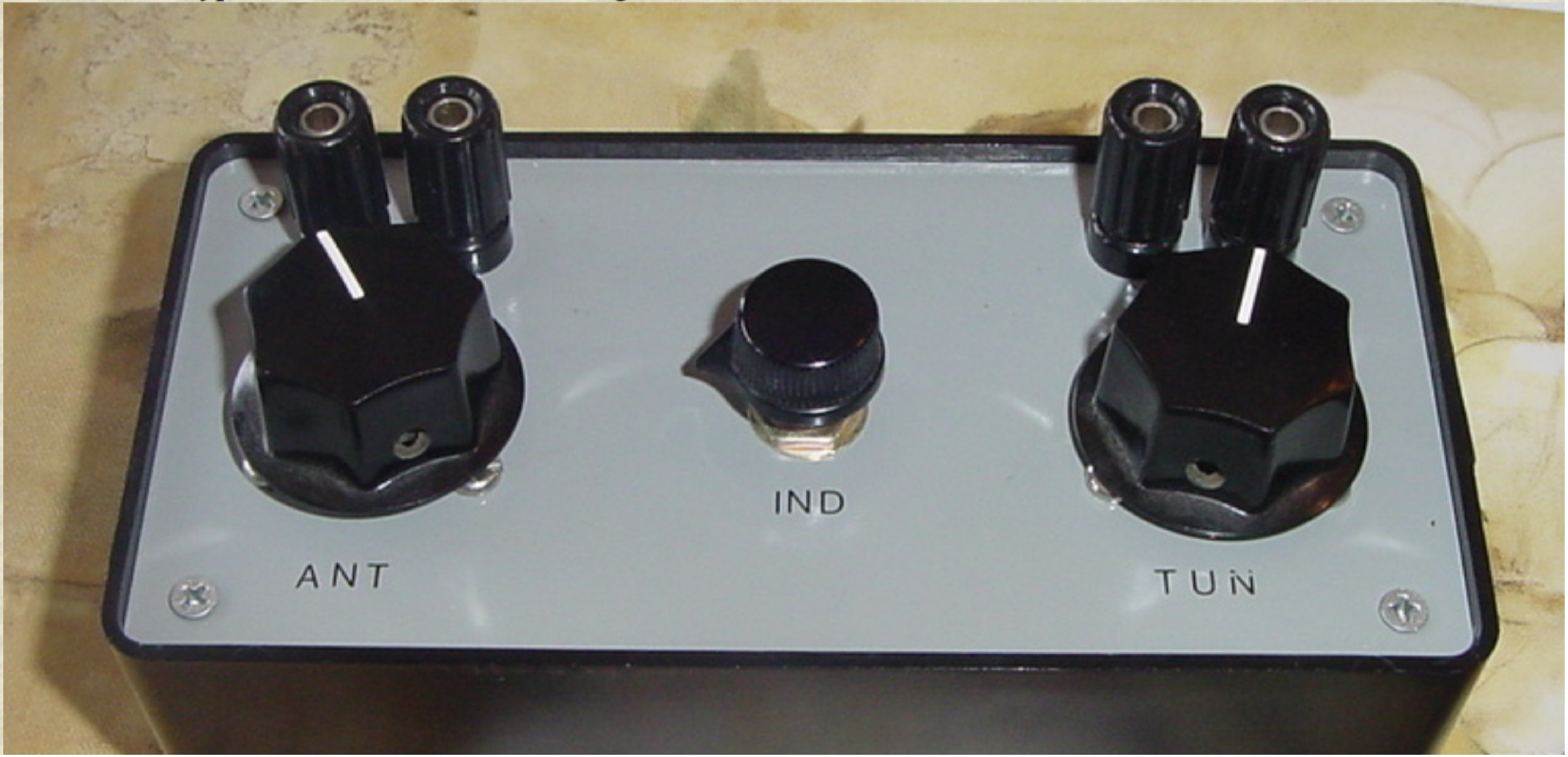

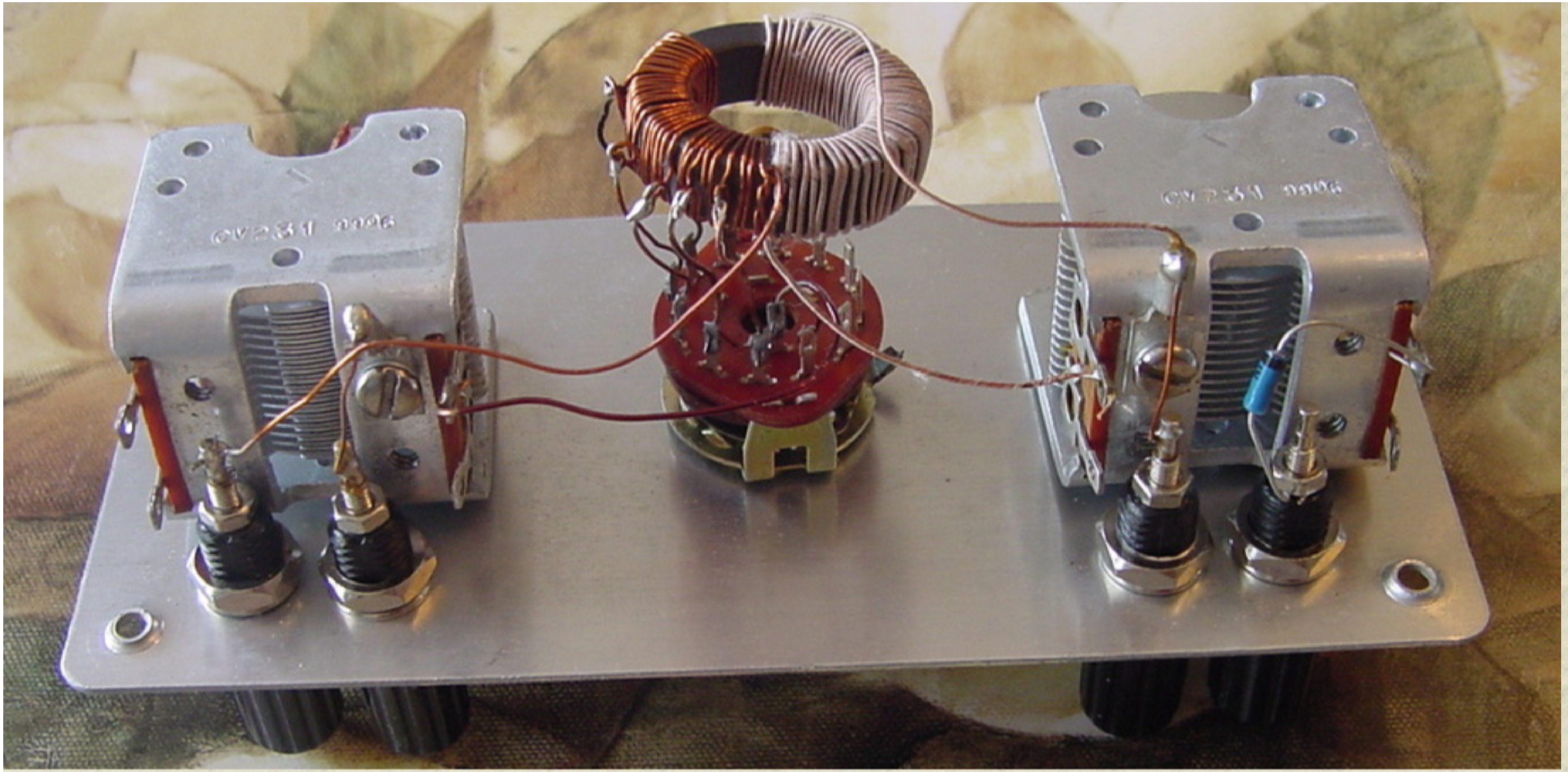

This is a dual tuned crystal set made in a box like the CR-1, just a fun prototype I made a few years back. It pulls in AM stations loud and clear.

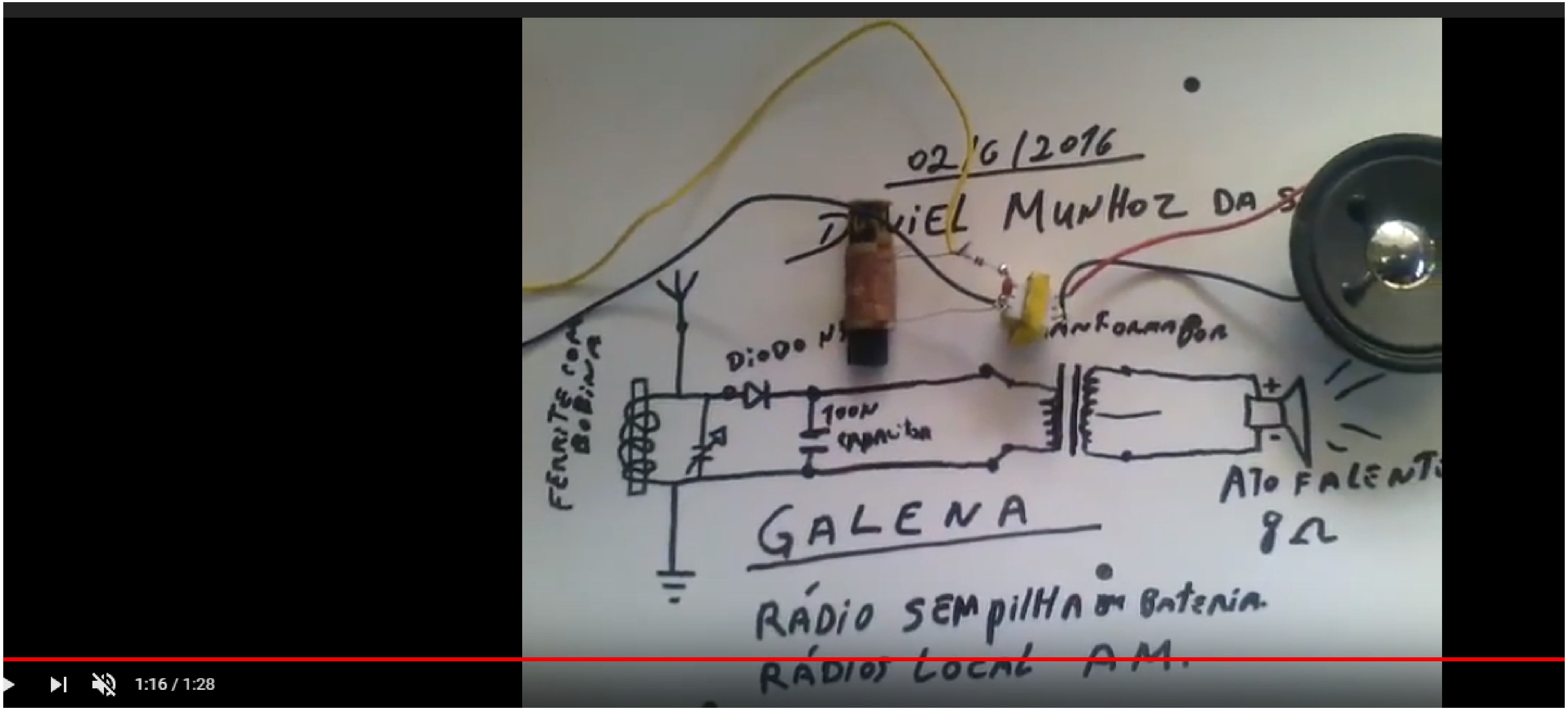
02/6/2016
DANIEL MUNHOZ DA S
FERRITE COM BOBINA
DIODO N
100N capacitor
TRANFORMADOR
+
-
ALTO FALENTE
8Ω
GALENA
RÁDIO SEM pilHA ou Bateria.
RÁDIOS LOCAL AM.
1:16 / 1:28

GALENA , Rádio AM ,sem pilha.

Up next

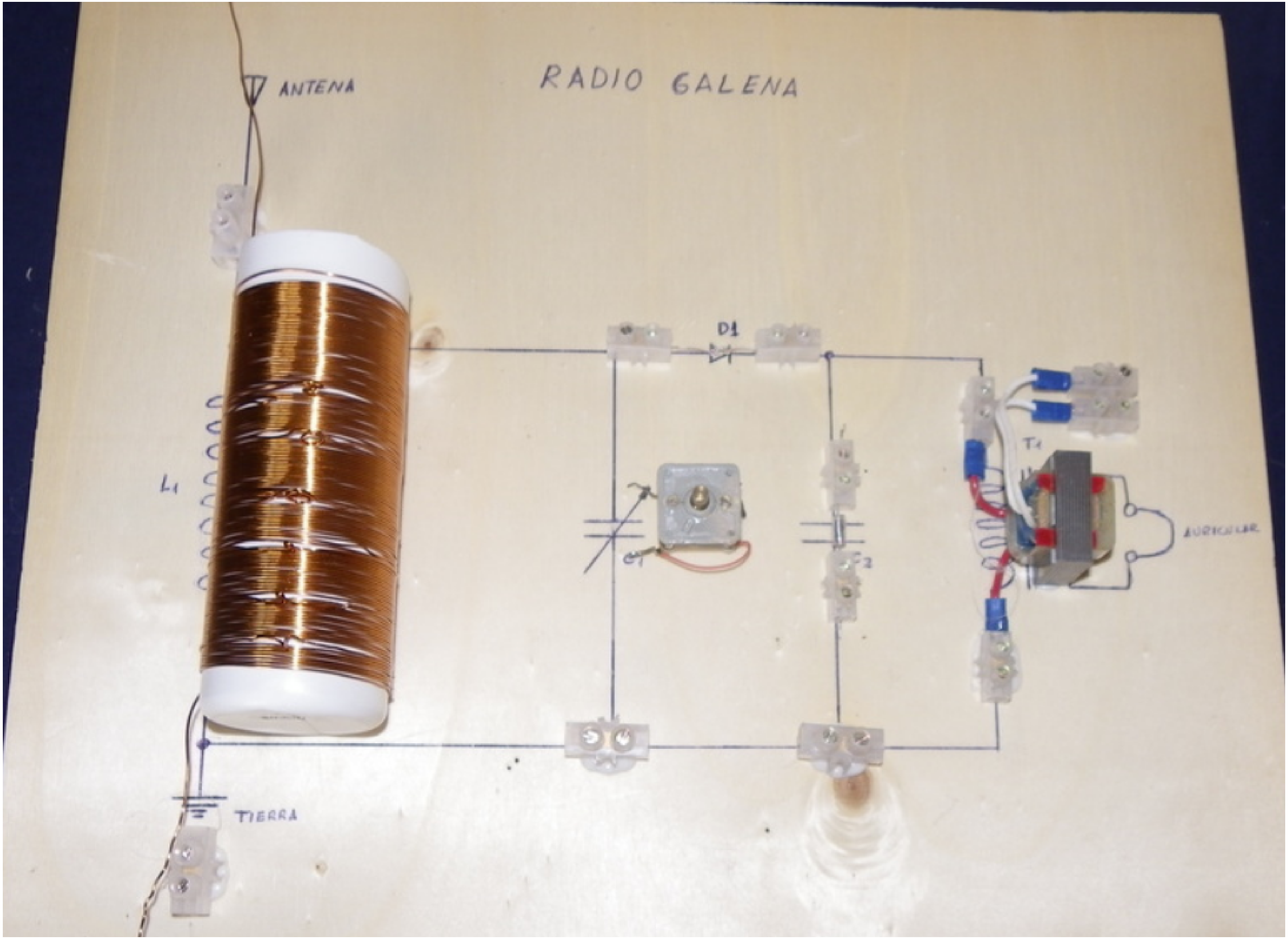
RADIO GALENA
ANTENA
L1
D1
C1
C2
T1
AURICULAR
TIERRA

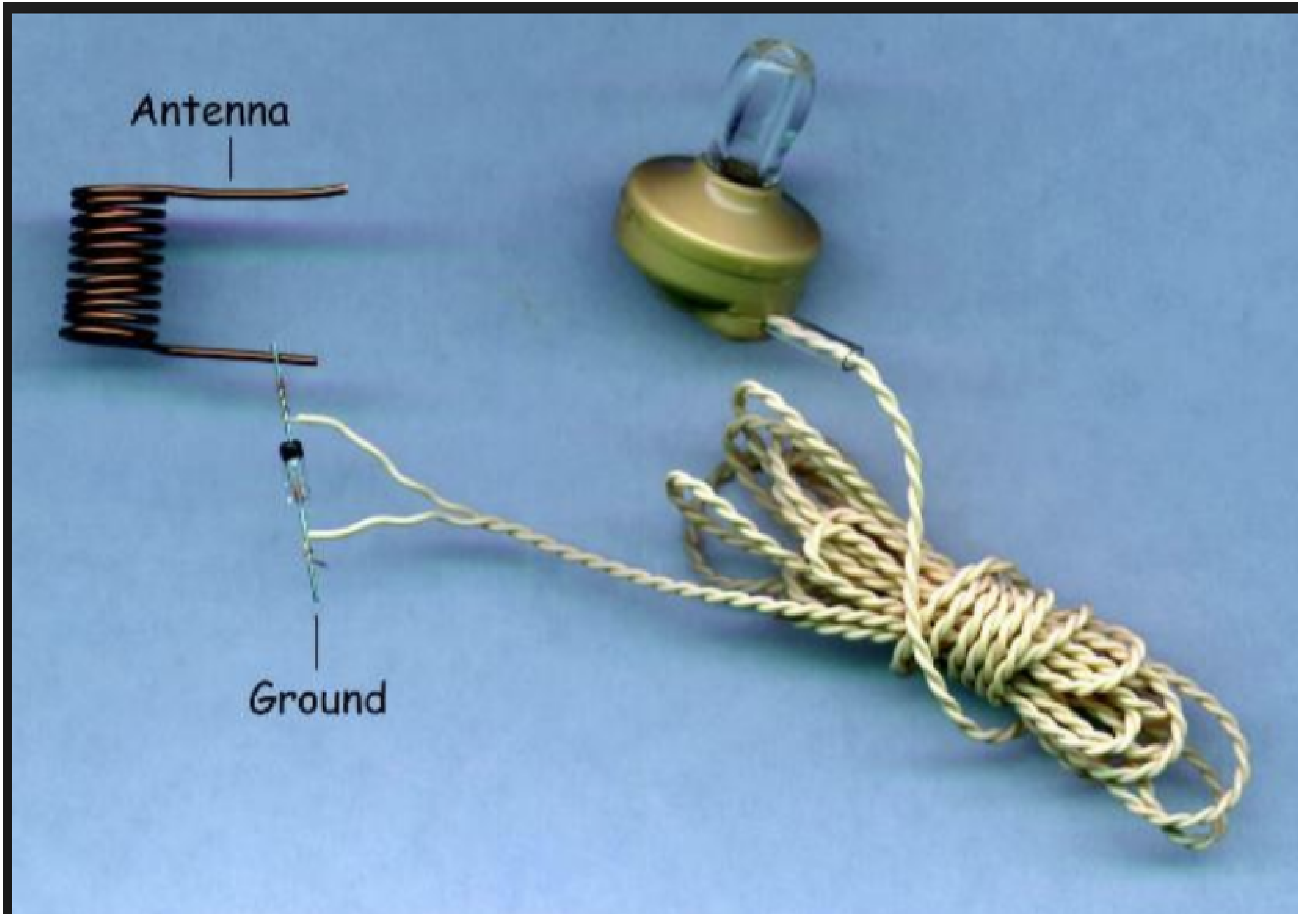
Antenna
Ground

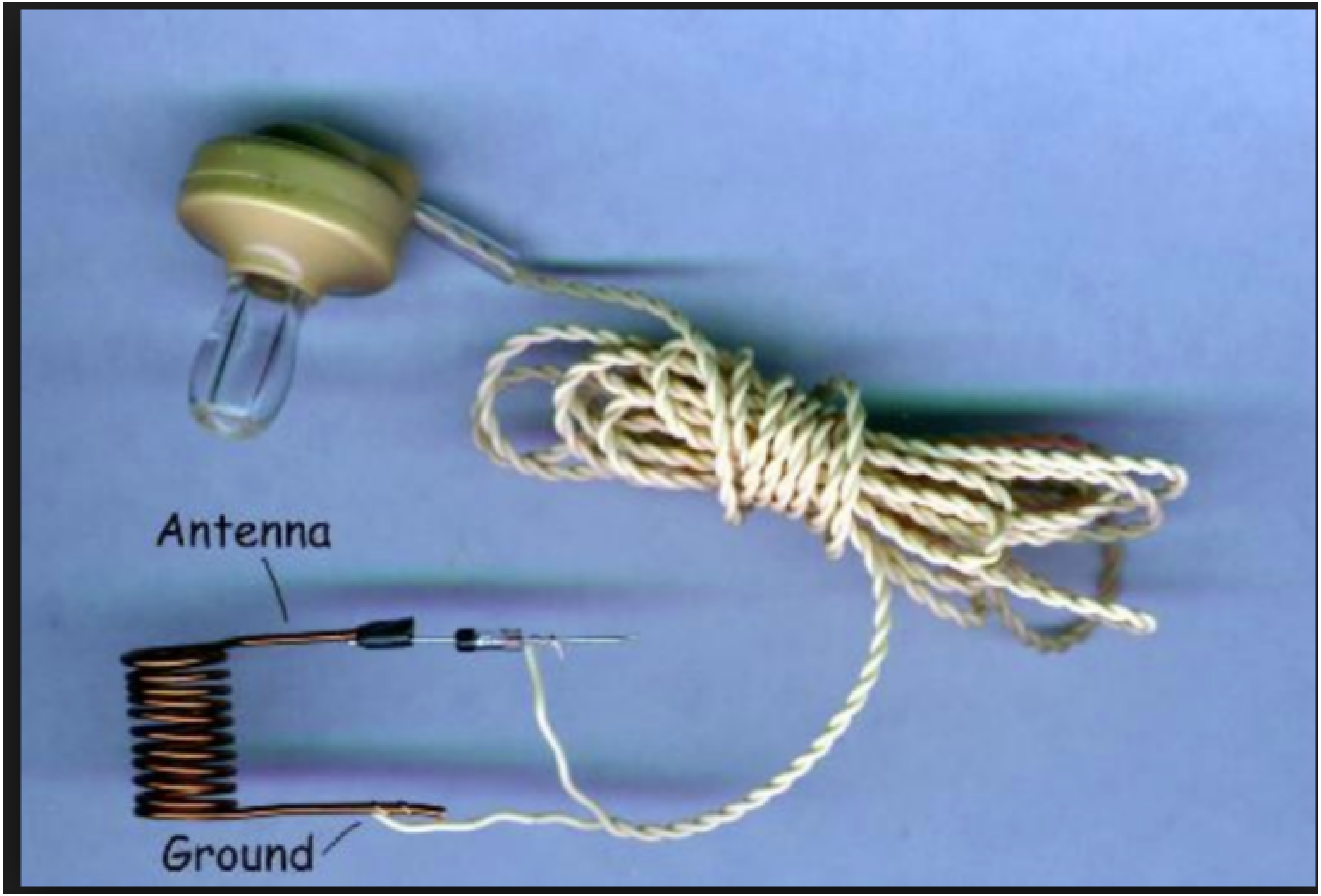
Antenna
Ground

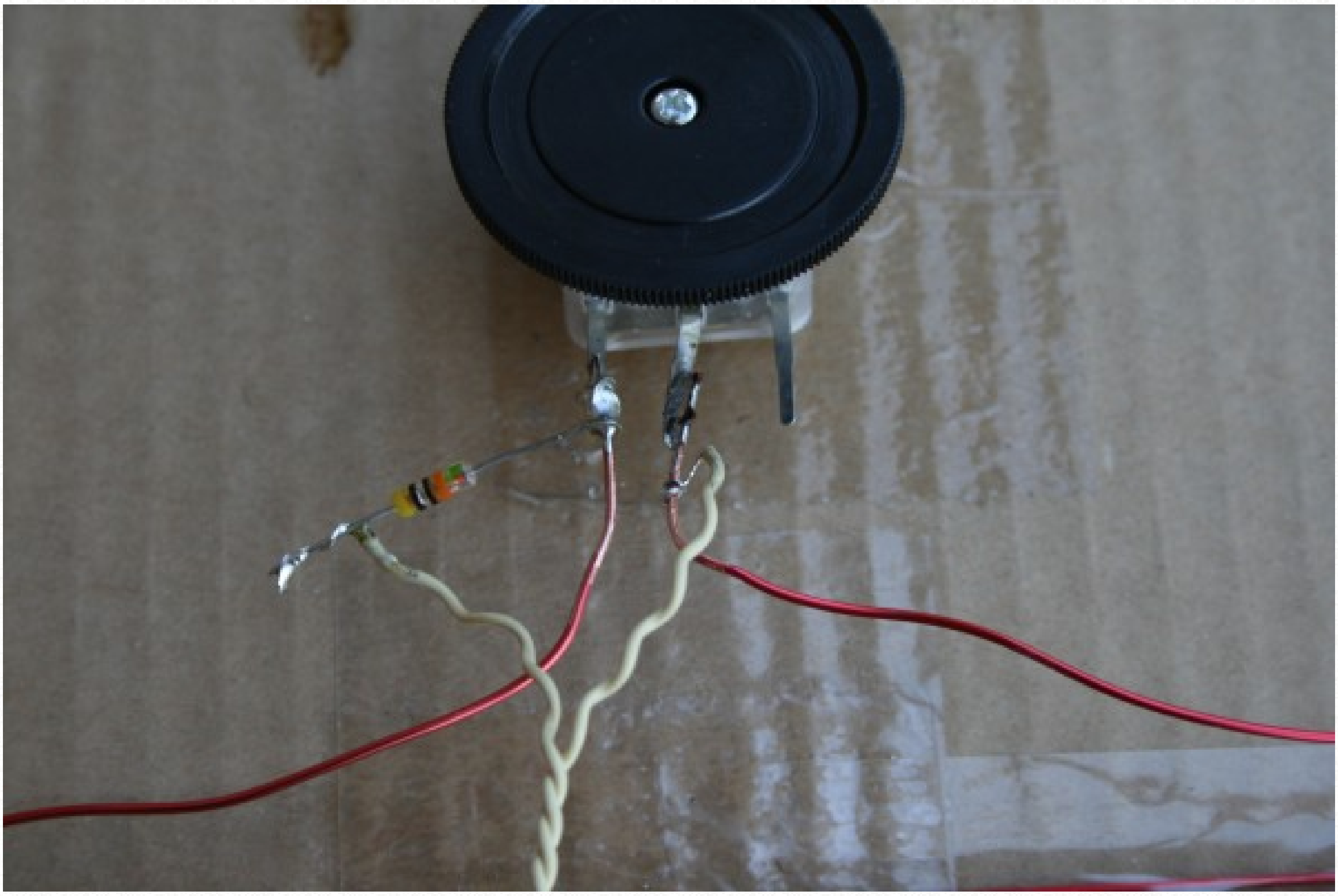

Click on photo for a larger picture

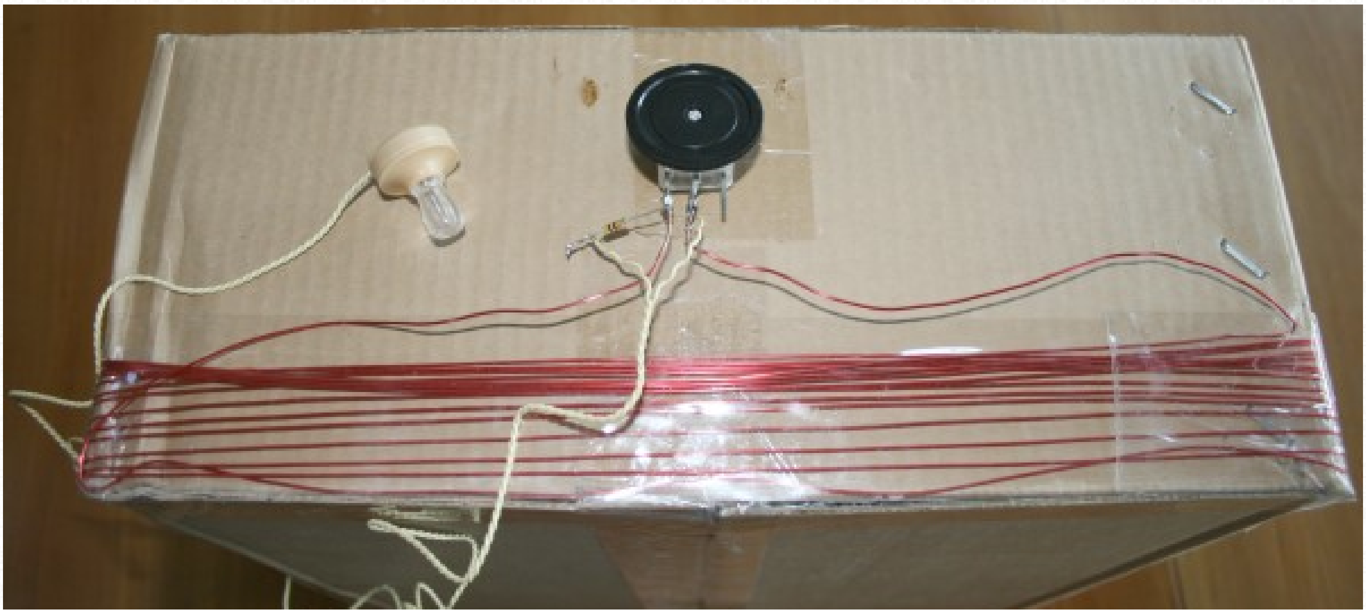

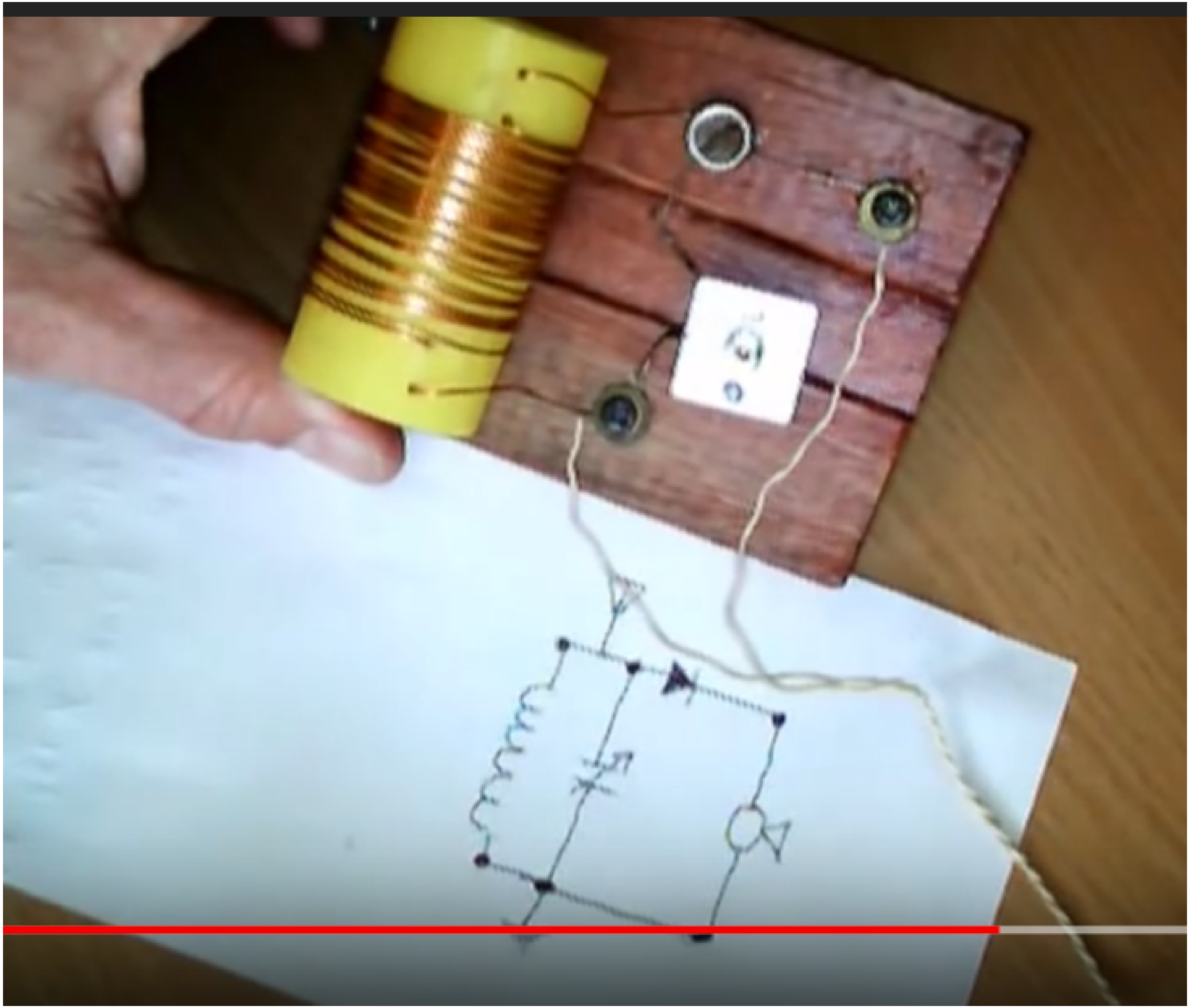

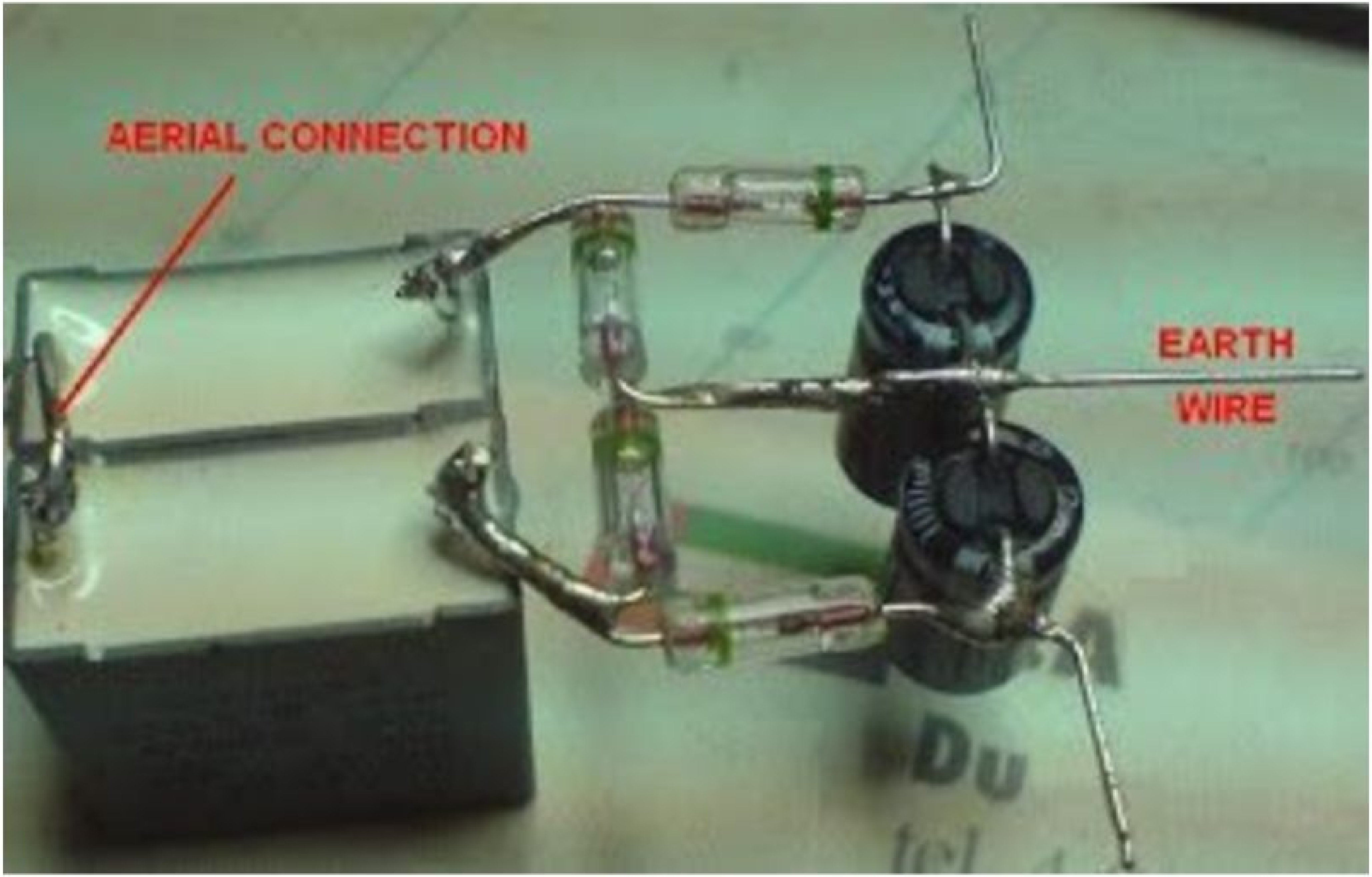
AERIAL CONNECTION
EARTH WIRE

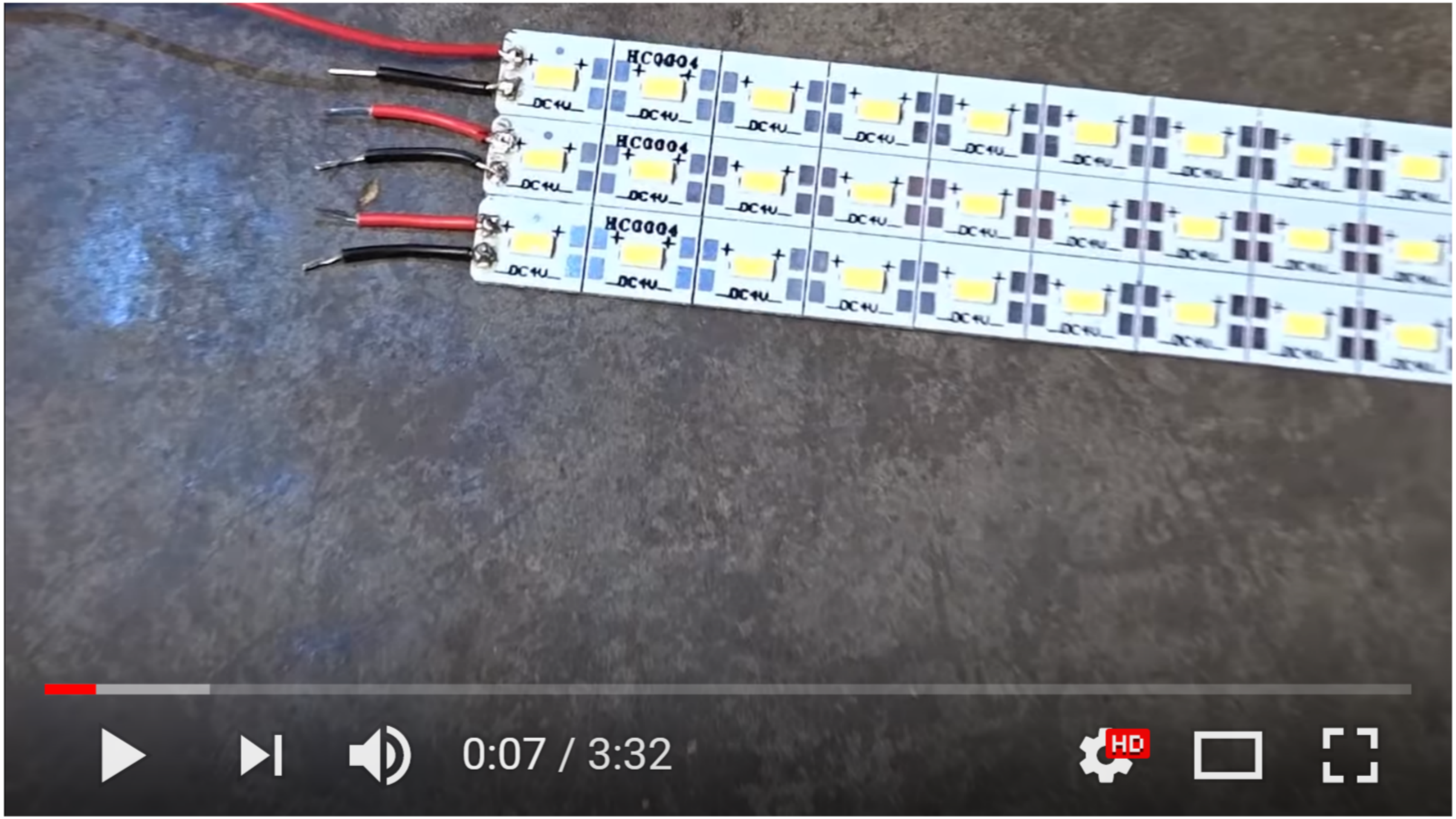
HC0004
DC4V
0:07 / 3:32
HD

New idea - You can make Solar Cell from LED (Free Energy)

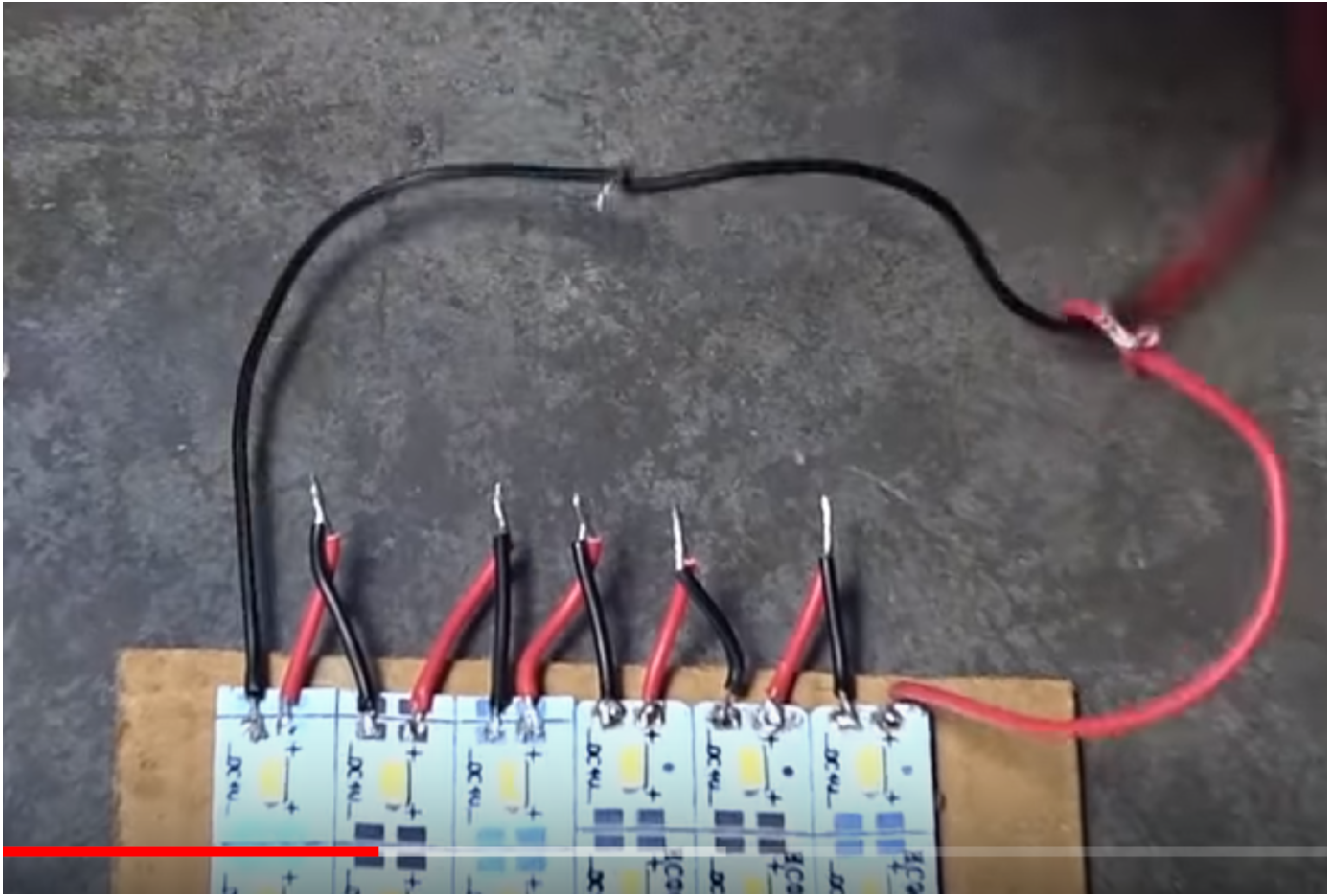

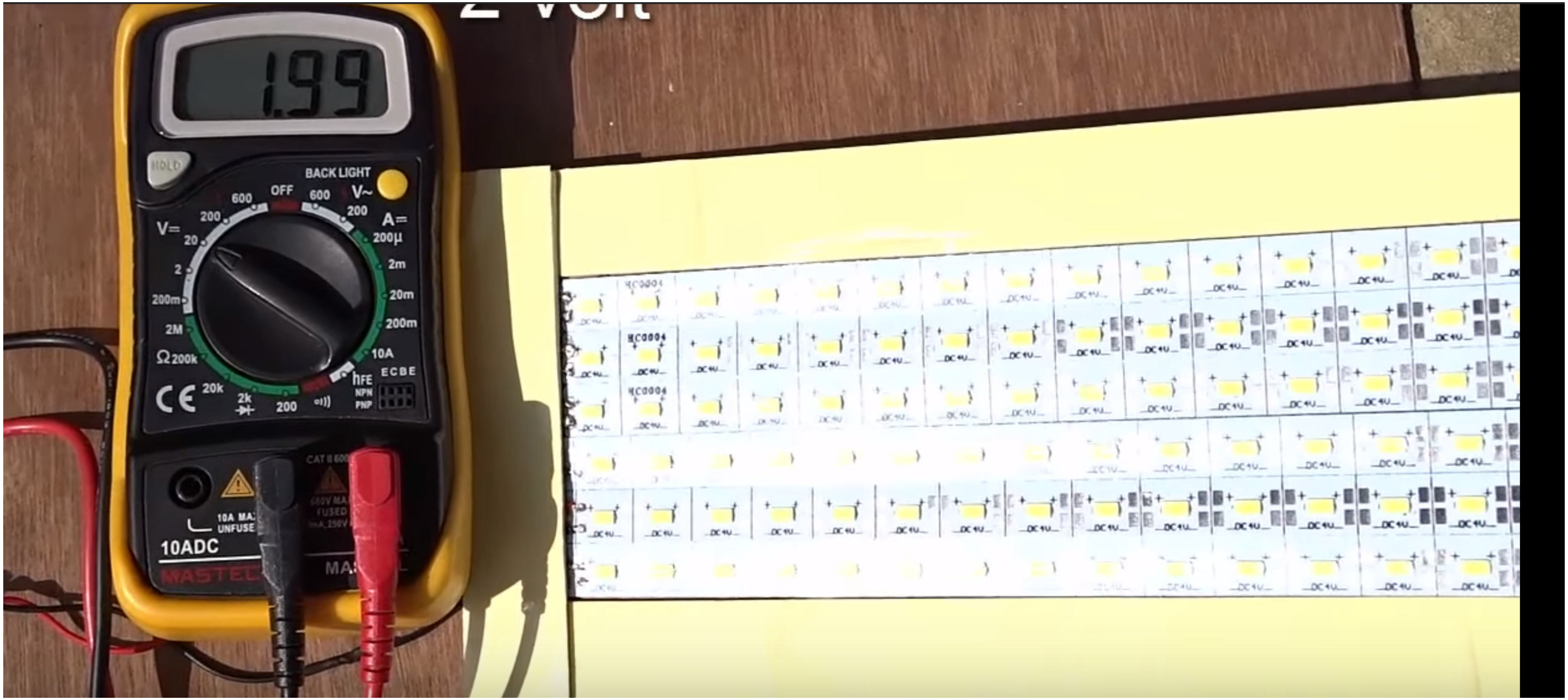
2 volt
1.99
HOLD
BACK LIGHT
OFF
V=
V~
A=
10A
hFE
NPN
PNP
ECBE
10ADC
MASTECH
DC 4V

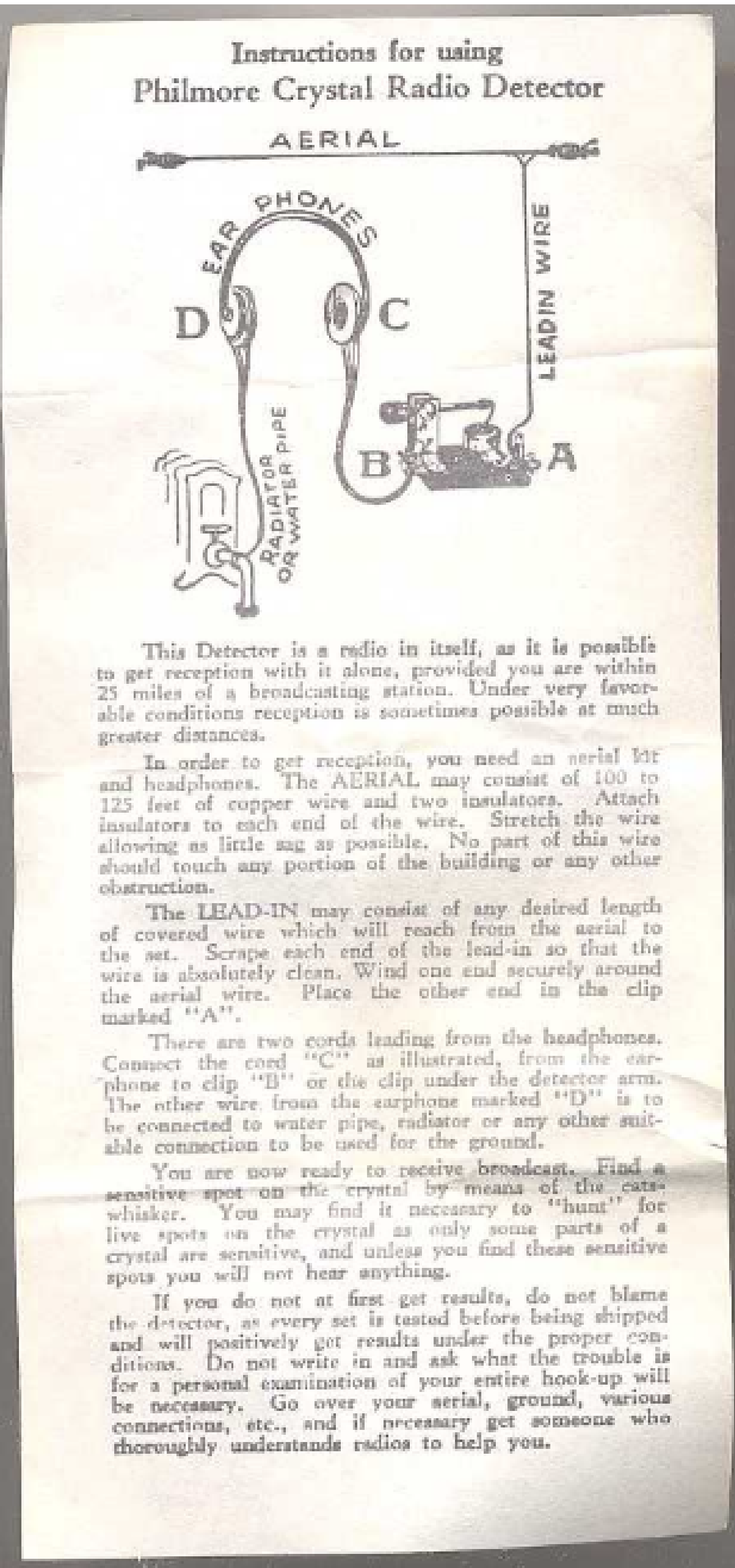

Instructions for using Philmore Crystal Radio Detector

AERIAL
EAR PHONES
LEADIN WIRE
D
C
B
A
RADIATOR OR WATER PIPE

This Detector is a radio in itself, as it is possible to get reception with it alone, provided you are within 25 miles of a broadcasting station. Under very favorable conditions reception is sometimes possible at much greater distances.

In order to get reception, you need an aerial kit and headphones. The AERIAL may consist of 100 to 125 feet of copper wire and two insulators. Attach insulators to each end of the wire. Stretch the wire allowing as little sag as possible. No part of this wire should touch any portion of the building or any other obstruction.

The LEAD-IN may consist of any desired length of covered wire which will reach from the aerial to the set. Scrape each end of the lead-in so that the wire is absolutely clean. Wind one end securely around the aerial wire. Place the other end in the clip marked "A".

There are two cords leading from the headphones. Connect the cord "C" as illustrated, from the earphone to clip "B" or the clip under the detector arm. The other wire from the earphone marked "D" is to be connected to water pipe, radiator or any other suitable connection to be used for the ground.

You are now ready to receive broadcast. Find a sensitive spot on the crystal by means of the cats-whisker. You may find it necessary to "hunt" for live spots on the crystal as only some parts of a crystal are sensitive, and unless you find these sensitive spots you will not hear anything.

If you do not at first get results, do not blame the detector, as every set is tested before being shipped and will positively get results under the proper conditions. Do not write in and ask what the trouble is for a personal examination of your entire hook-up will be necessary. Go over your aerial, ground, various connections, etc., and if necessary get someone who thoroughly understands radios to help you.

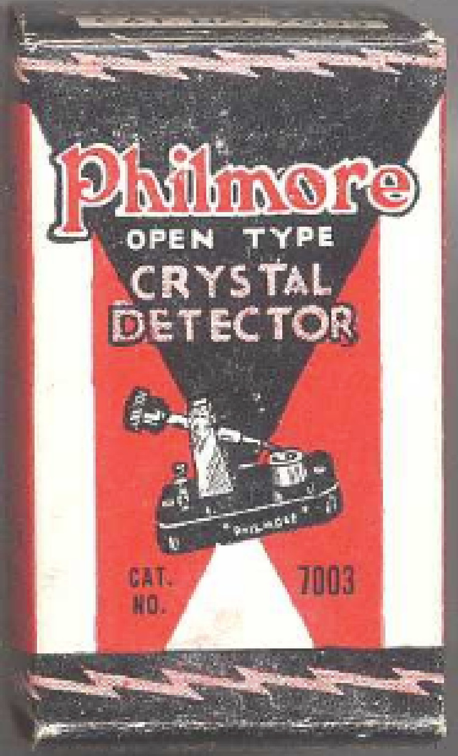

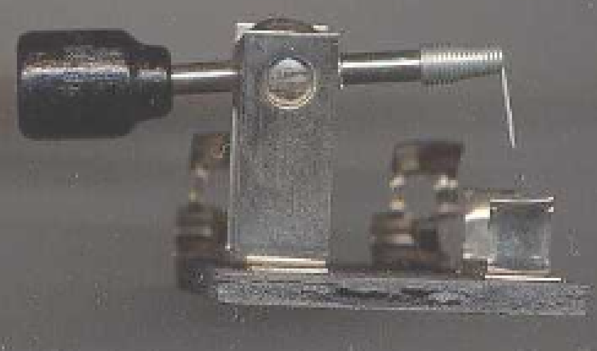

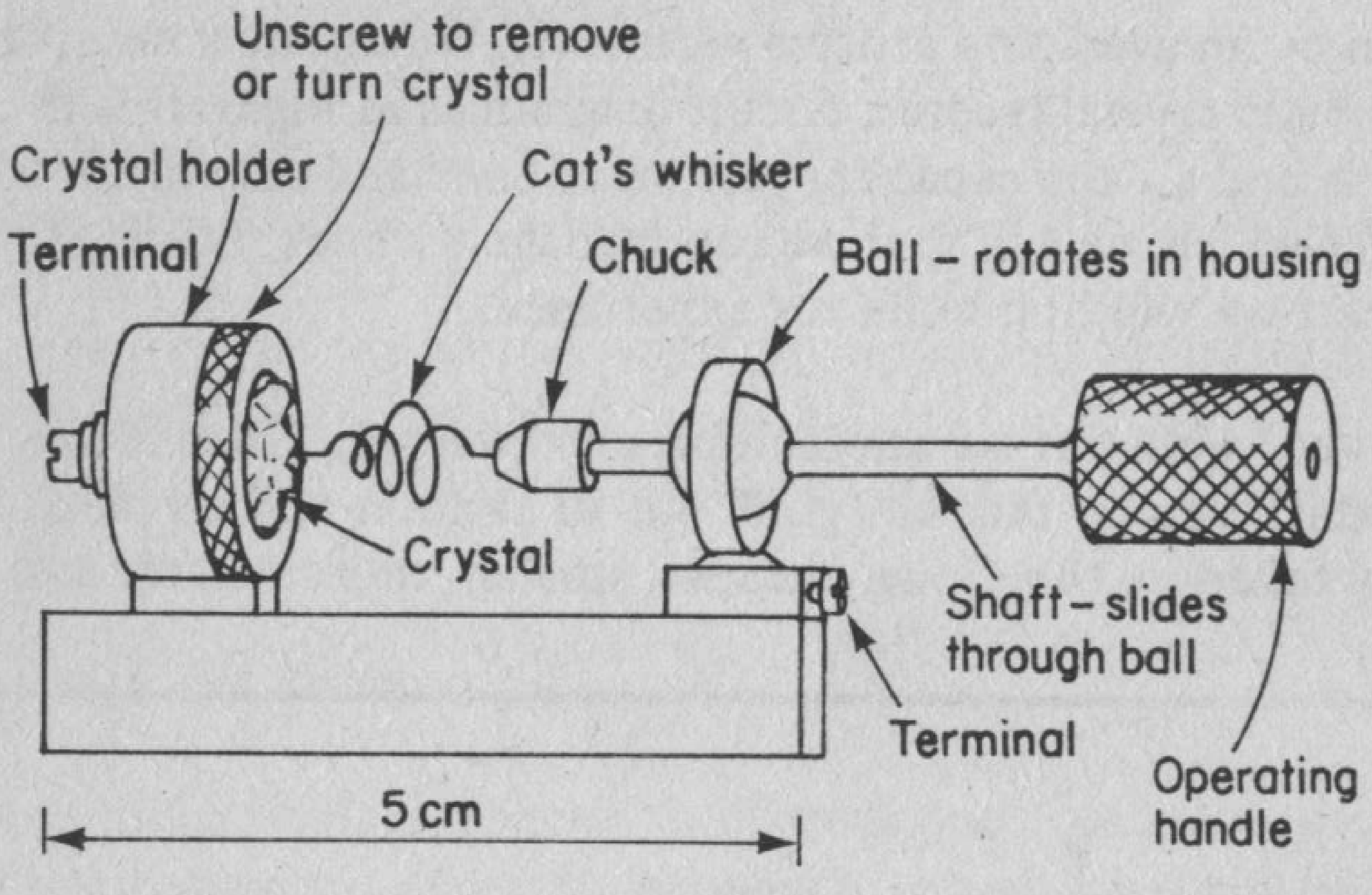
Unscrew to remove or turn crystal
Crystal holder
Cat's whisker
Terminal
Chuck
Ball – rotates in housing
Crystal
Shaft – slides through ball
Terminal
Operating handle
5 cm

Modern diode (7·5 X 2·5 mm)

Majoritatea componentelor active folosite în circuitele electronice moderne sînt dispozitivele bazate pe semiconductoare.

Cel mai simplu dispozitiv este *dioda* punctiformă cu germaniu. Ea are proprietatea de bază de a se comporta ca și cum este conectată direct la o sursă electrică de curent continuu (plusul sursei la plusul diodei) și ca un izolator, cînd este conectată invers la aceeași sursă (plusul sursei la minusul diodei), ca în figura 2.9.

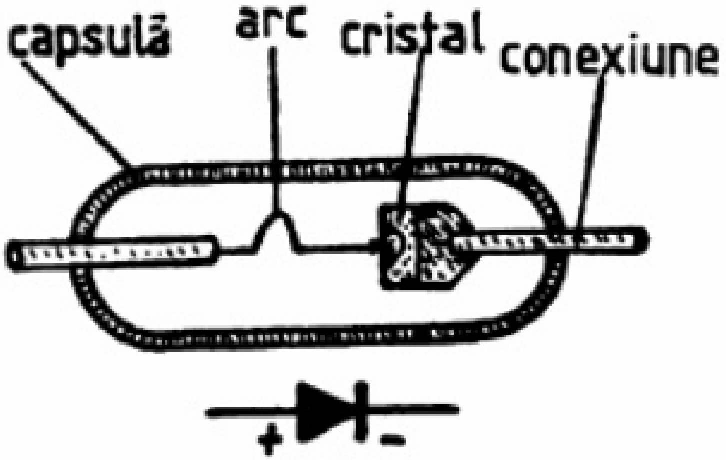
capsulă
arc
cristal
conexiune
+
-

FIG. 2.9.

Dioda punctiformă cu germaniu

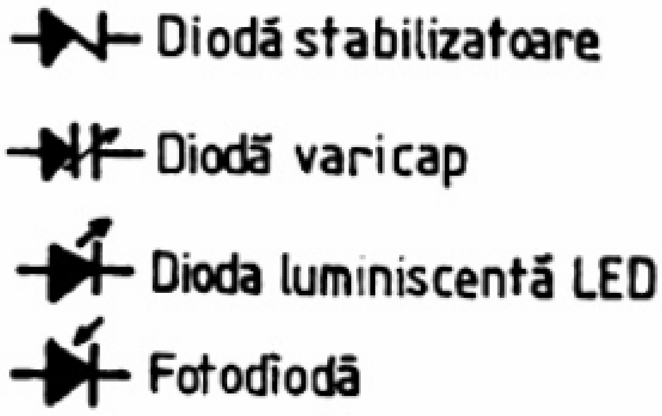
Diodă stabilizatoare
Diodă varicap
Dioda luminiscentă LED
Fotodiodă

FIG. 2.10.

Tipuri de diode semiconductoare

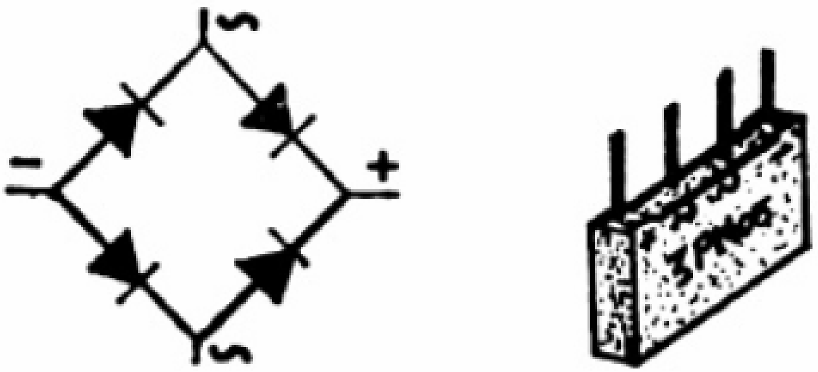

FIG. 2.11.

Puntea redresoare

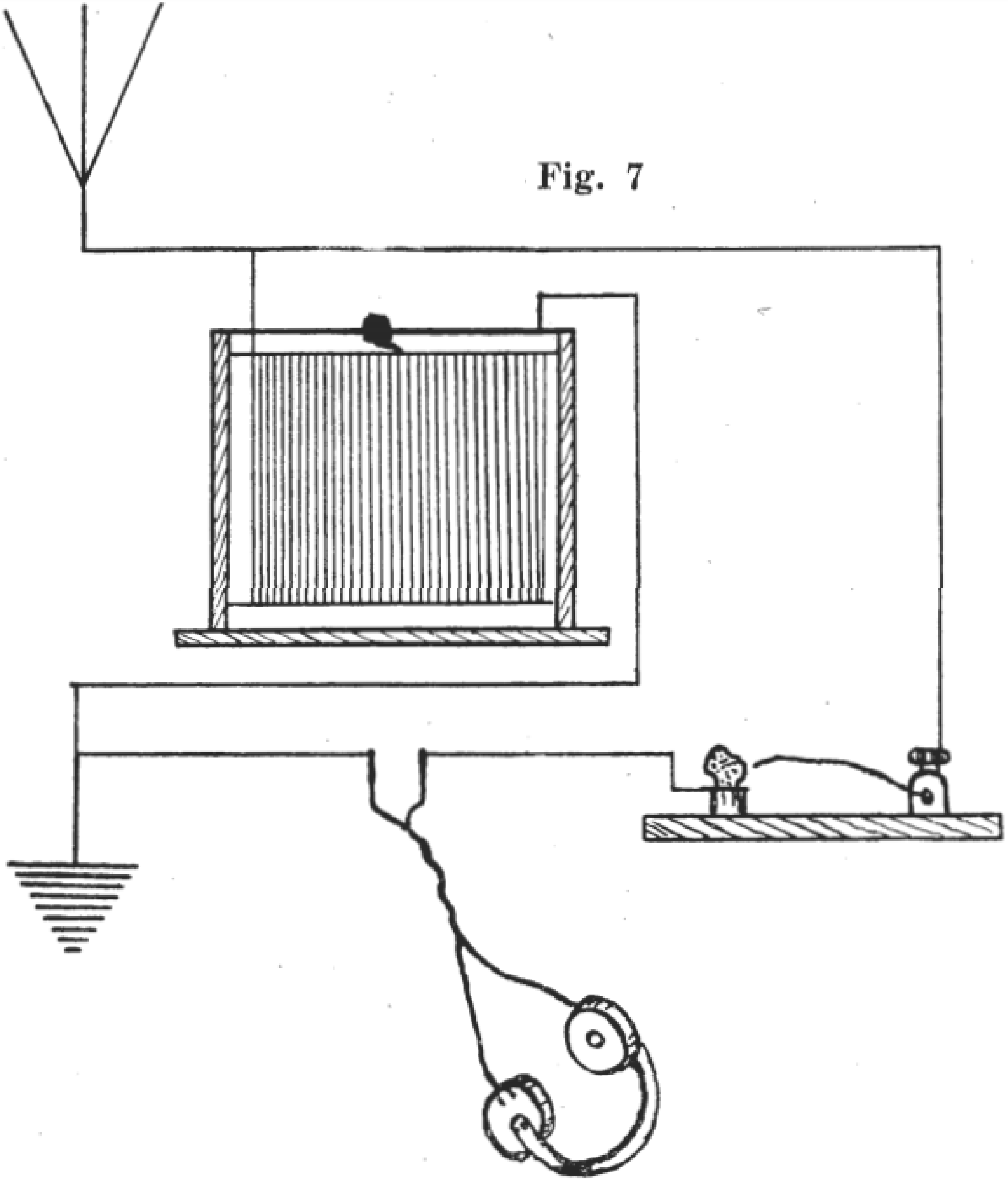

Fig. 7

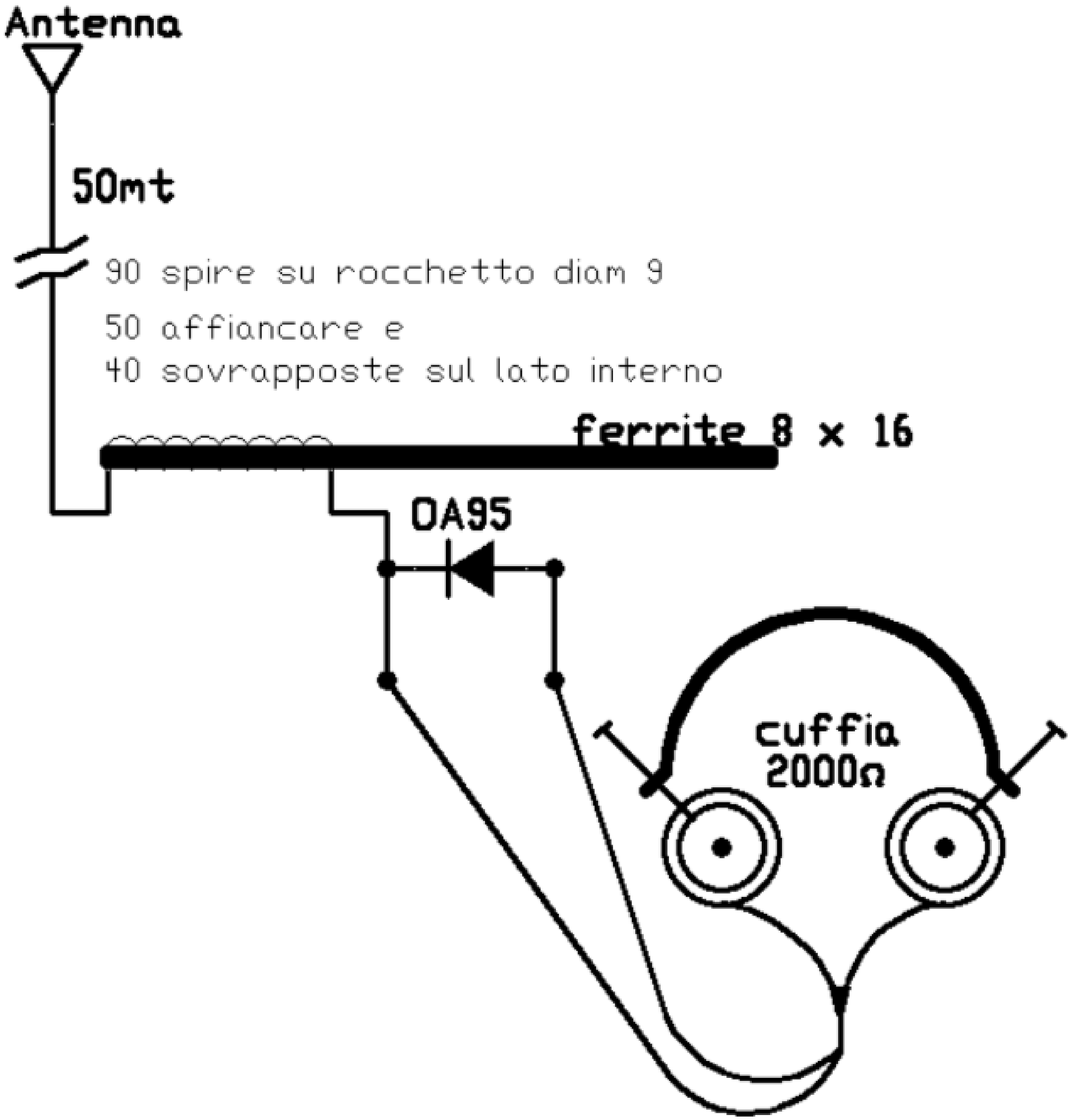
Antenna
50mt
90 spire su rocchetto diam 9
50 affiancare e
40 sovrapposte sul lato interno
ferrite 8 x 16
OA95
cuffia
2000Ω

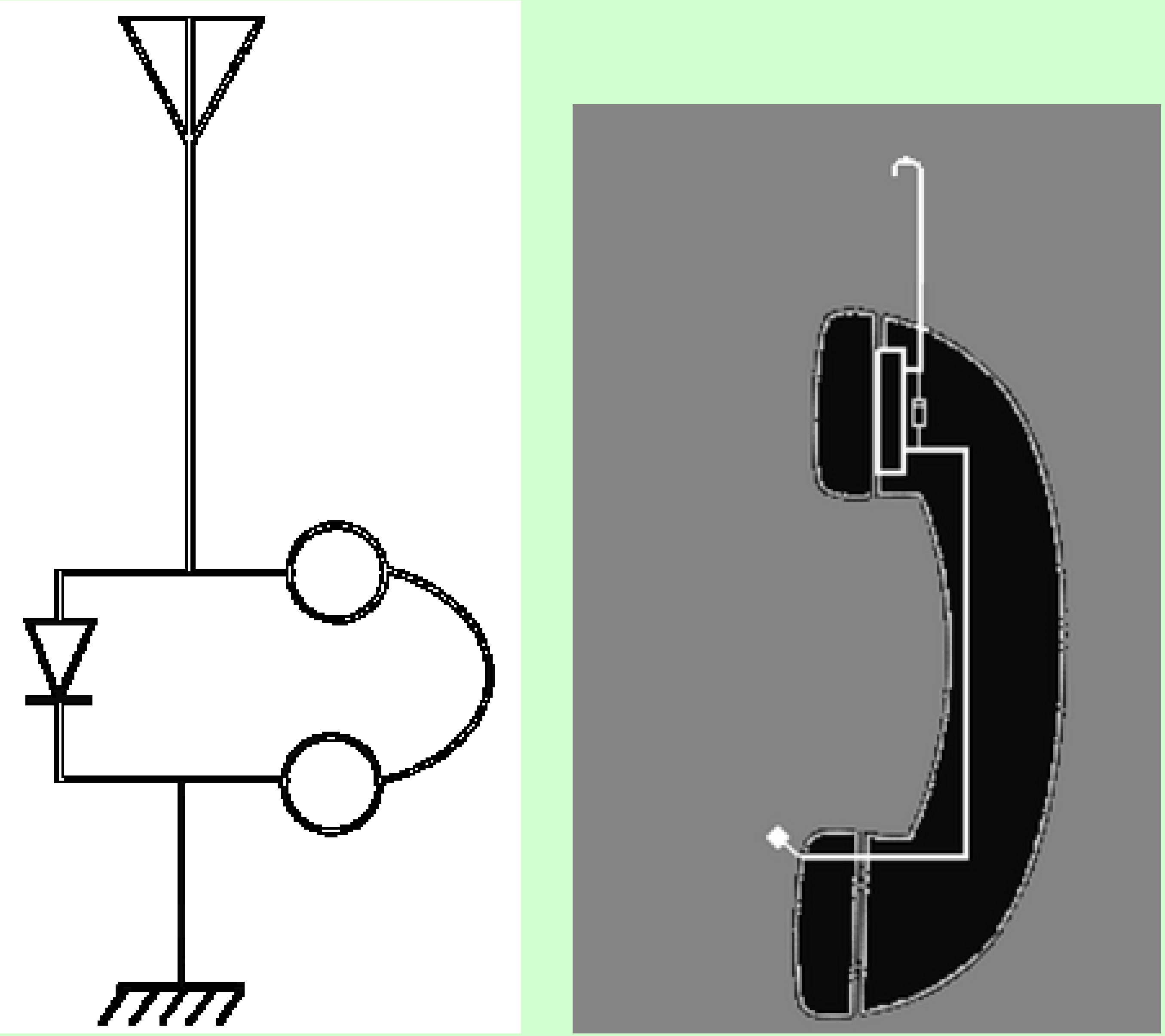

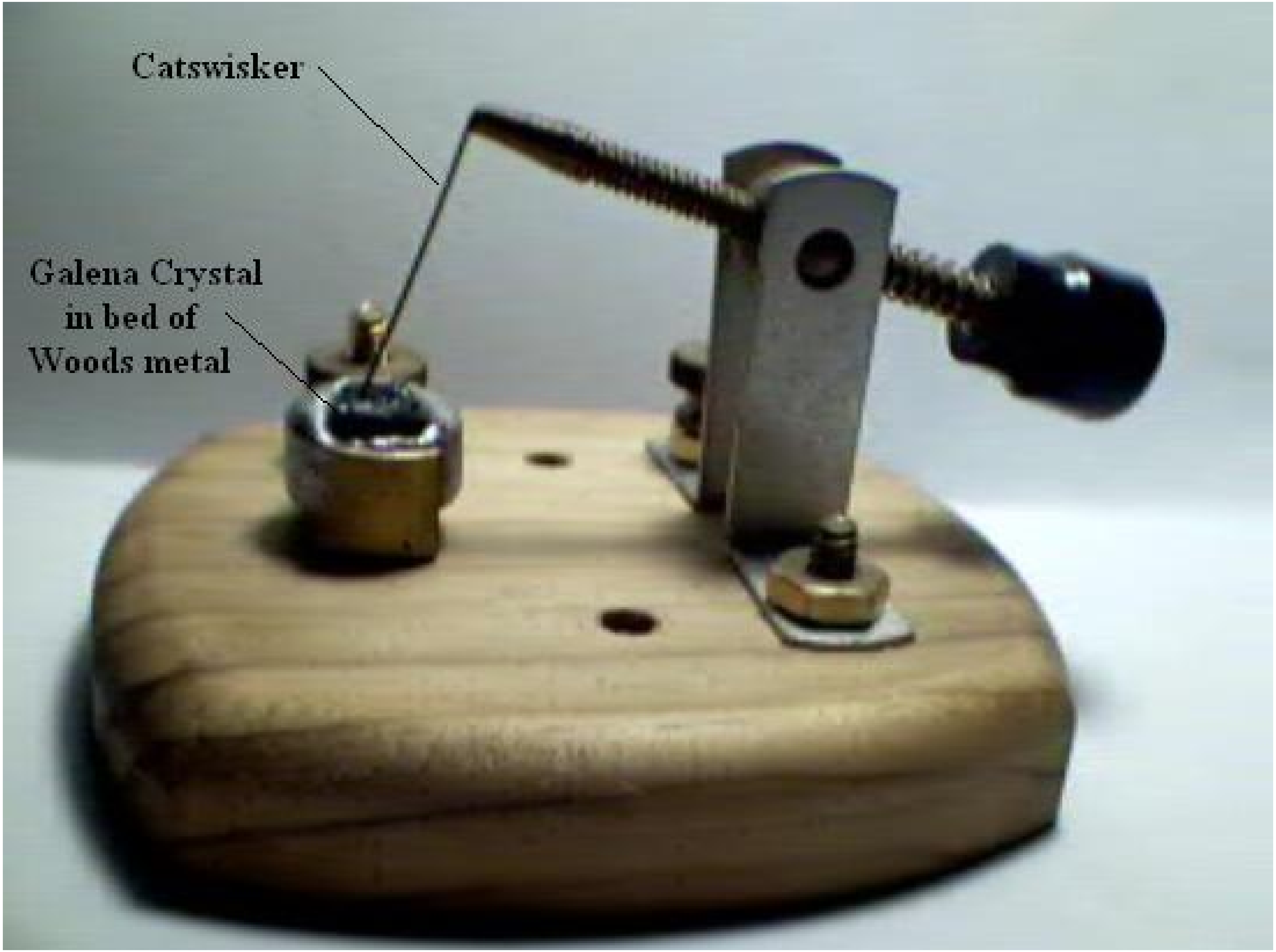
Catswisker
Galena Crystal in bed of Woods metal

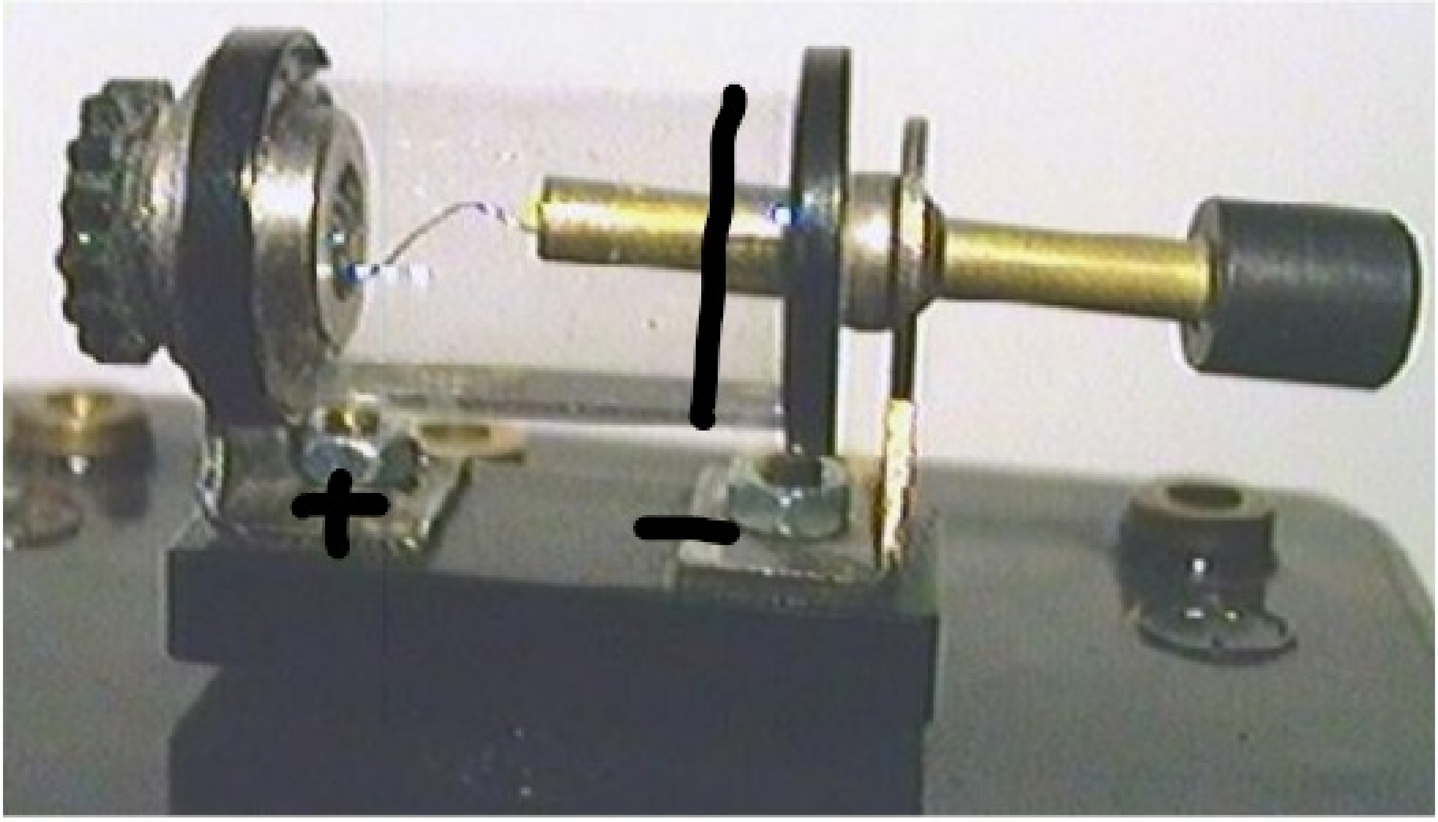
+
−

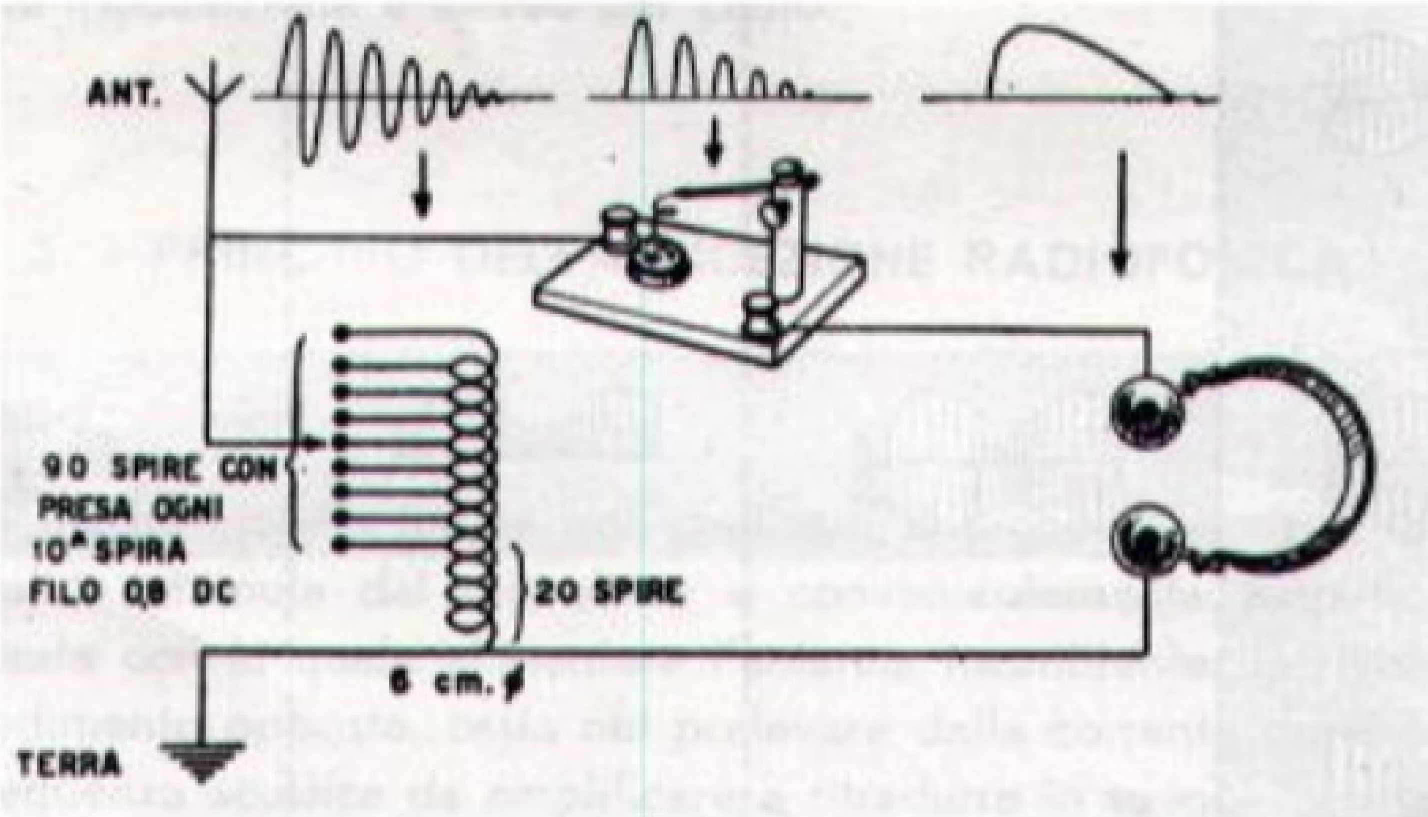
ANT.
90 SPIRE CON PRESA OGNI 10ª SPIRA FILO 0,8 DC
20 SPIRE
6 cm. ∅
TERRA

Lead sulfide crystals, also called "galena"

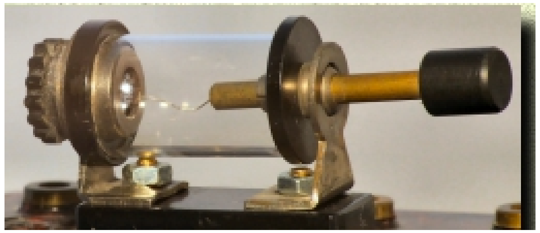

Crystal detector, the lever was used to position the spring contact on a crystal face capable of performing the radio signal detection.

Yeah , but how does a galena radio work? Here is the constructive scheme of a receiver of this kind, obtained from a book of popular radio commu
to the knowledge of the radio, ed.Hepep 1943):

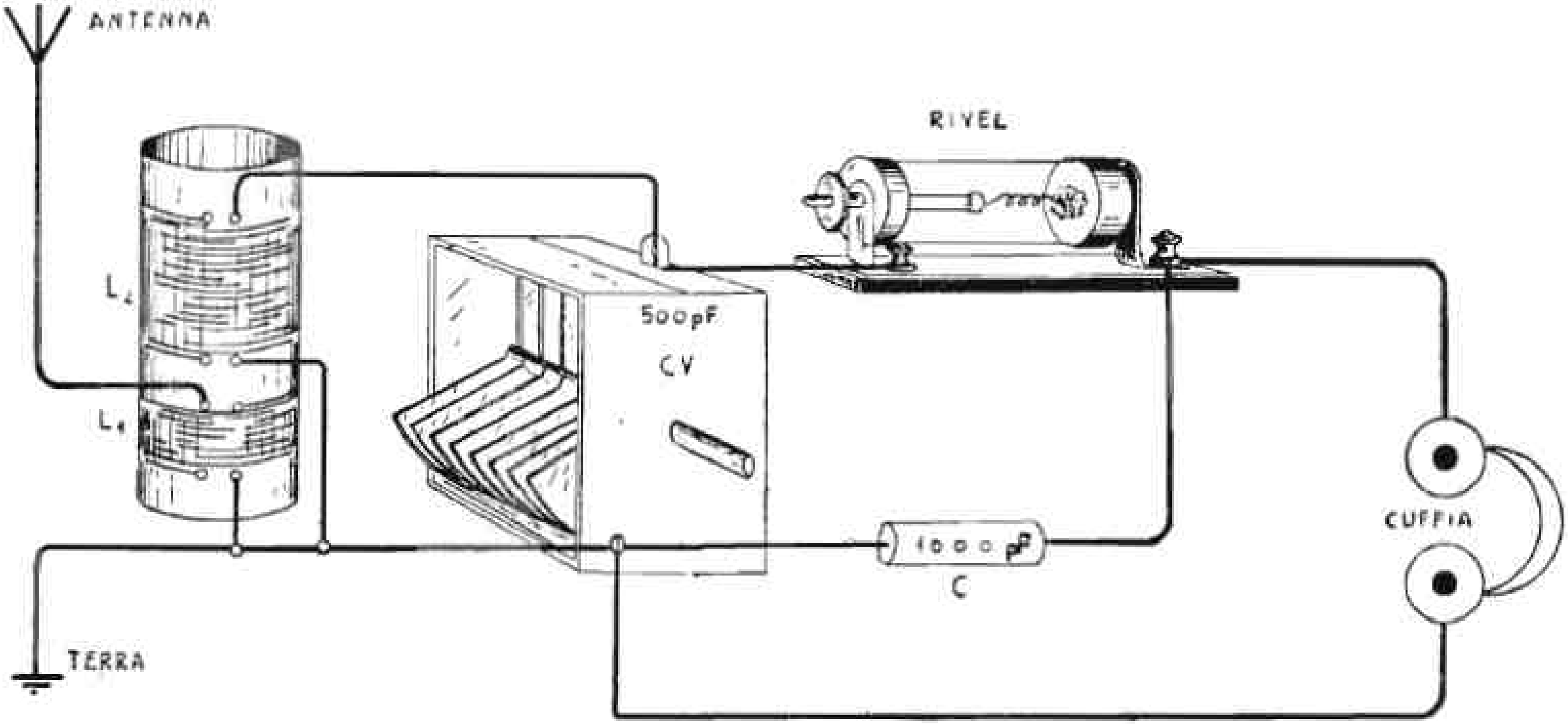

Fig. 13.5. – Realizzazione pratica dello schema di fig. 13.3 con aggiunta della bobina d'antenna.

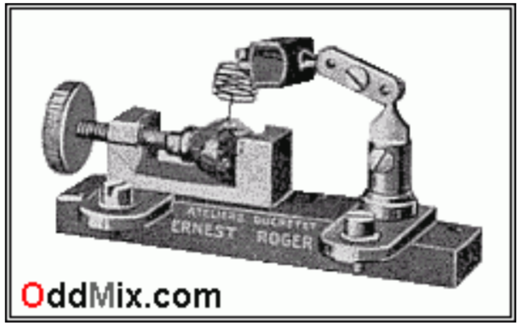

Picture 1. "Ernest Roger" Crystal detector from 1920

Many of the **younger generation** never had the pleasure to have, or to even see a real beautiful crystal detector as shown on **Picture 1**. A detector like that was the product of the early 1920's, and it was an expensive, quality, well made product. They were usually hands assembled with machine made components.

This particular "**Ernest Roger**" detector device is built on an insulated base, either wood or most likely of Bakelite, the most widely used and only available plastic material at the times.

Many of the younger generation never had the pleasure to have, or to even see a real beautiful crystal detector as shown on **Picture 1**. The new generations are much too involved with technology and playing on their computers or going to an **online University**. This excellent detector holder used copper, brass and bronze generously for the crystal holder and for all of the **electrodes**.

The most often used crystal in this and similar detectors - then and now - was **pyrite** or galena as shown in the 1925 listing in a table in the **"American Mineralogist"** publication, that has a listing of thirty three minerals catalogued with rectifying detector properties.

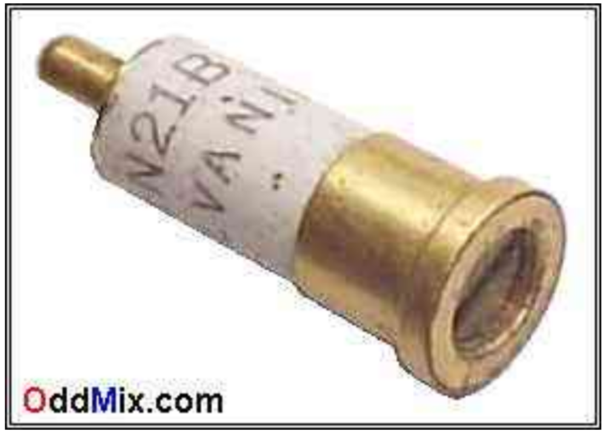

Picture 2. Early point contact diode with adjustable srew

Among the very first commercially produced diodes is the 1N21B shown on **Picture 2**. It is more than curious, that this "diode" is enclosed in a case on which there is a slotted screw-head is just visible on the wide side (lower right). That screw is connected to a fine "cats whisker" steel wire, a few turns of a spring-like device, terminating in a point that is in touch of the germanium semiconductor material connected to the top left diode terminal.

The arrangement is much smaller, and more diode like, then the detector on **Picture 1**., but it made in a very similar arrangement, which has became known as the point contact diode. From these and similar germanium diodes, evolutionary progress leads us to current PN junction Silicon, Gallium Arsenide or other more exotic semiconductor materials which are the achievement of the latest scientific age and many years of steady experimentation, research and development.

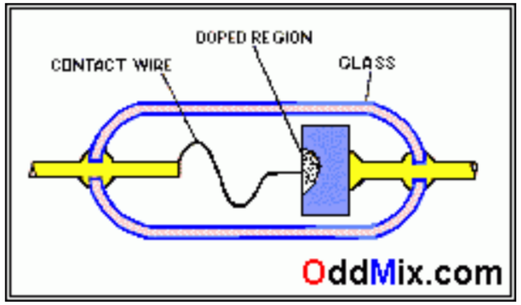

Picture 3. Point contact diode

Picture 3 shows the cross section of a hermetically glass enclosed, point contact diode. If the left side, point contact terminal would be attached to a screw and a short spring, and the glass envelope would have a threaded metal part in it, the **Picture 2** and **Picture 3** devices would be nearly identical.

These point contact "cat's whisker" devices, are still made occasionally because of their very small capacitance. As it was discovered early on, they are highly useful in high frequency electronics. All the way up to microwave frequencies they useable. The simplest radar detectors usually made with a simple horn **antenna**, a diode similar to **Picture 2** and **Picture 3** and a transistor amplifier. If the designer keeps it simple, and don't use a local oscillator, such radar detectors

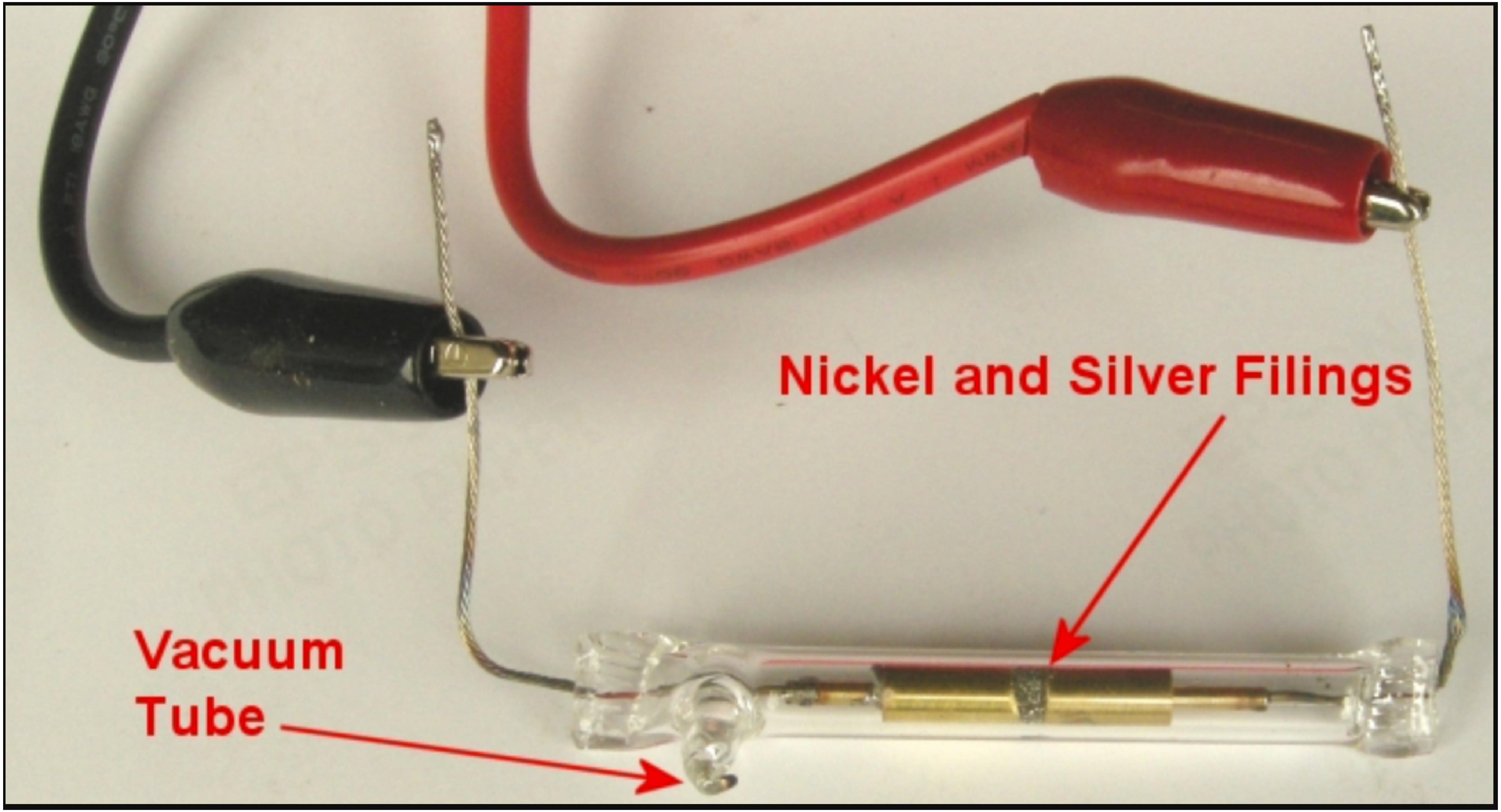
Nickel and Silver Filings
Vacuum Tube

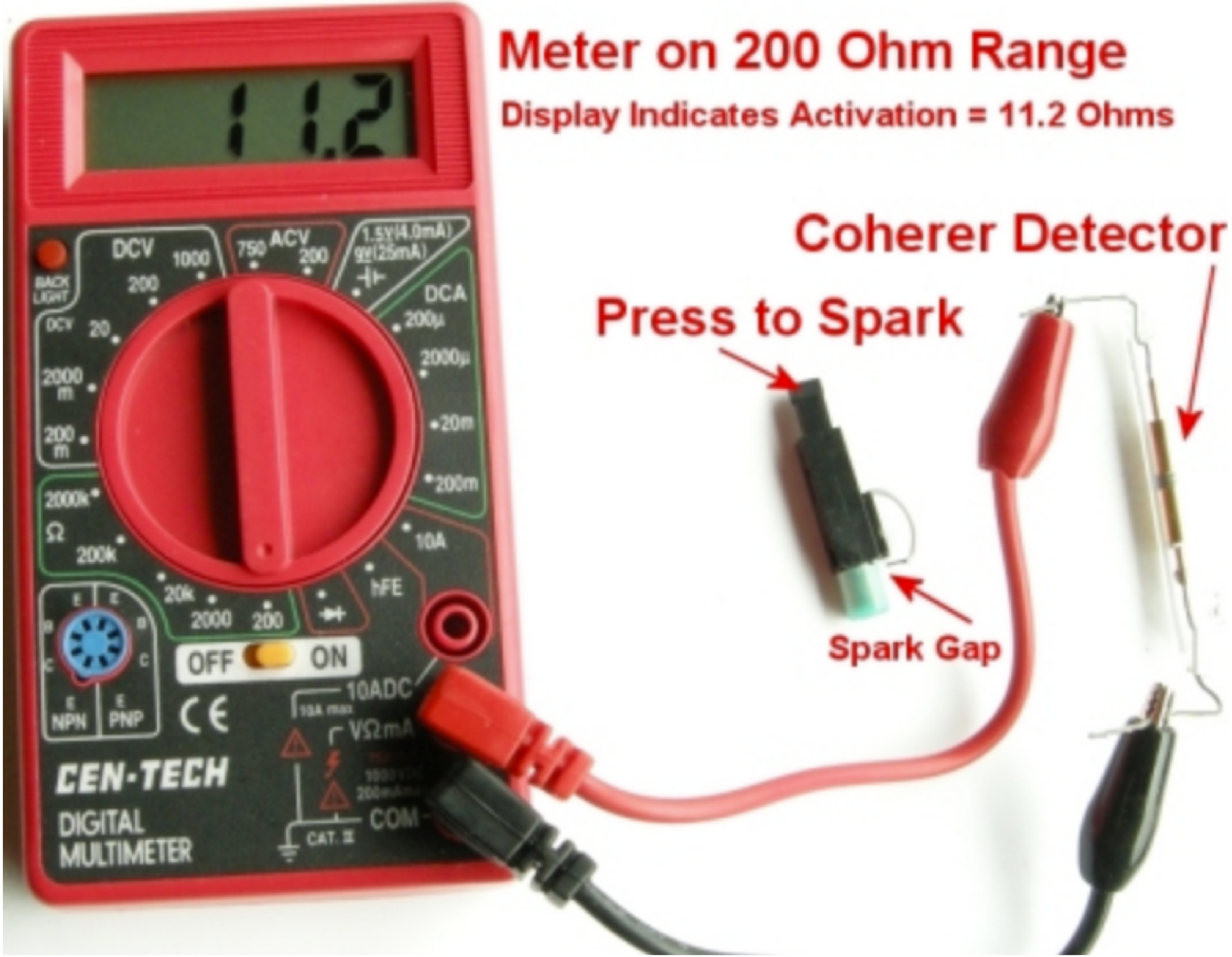

Meter on 200 Ohm Range

Display Indicates Activation = 11.2 Ohms

Coherer Detector

Press to Spark

Spark Gap

11.2
DCV
ACV
DCA
OFF
ON
10ADC
VΩmA
COM
CEN-TECH
DIGITAL
MULTIMETER

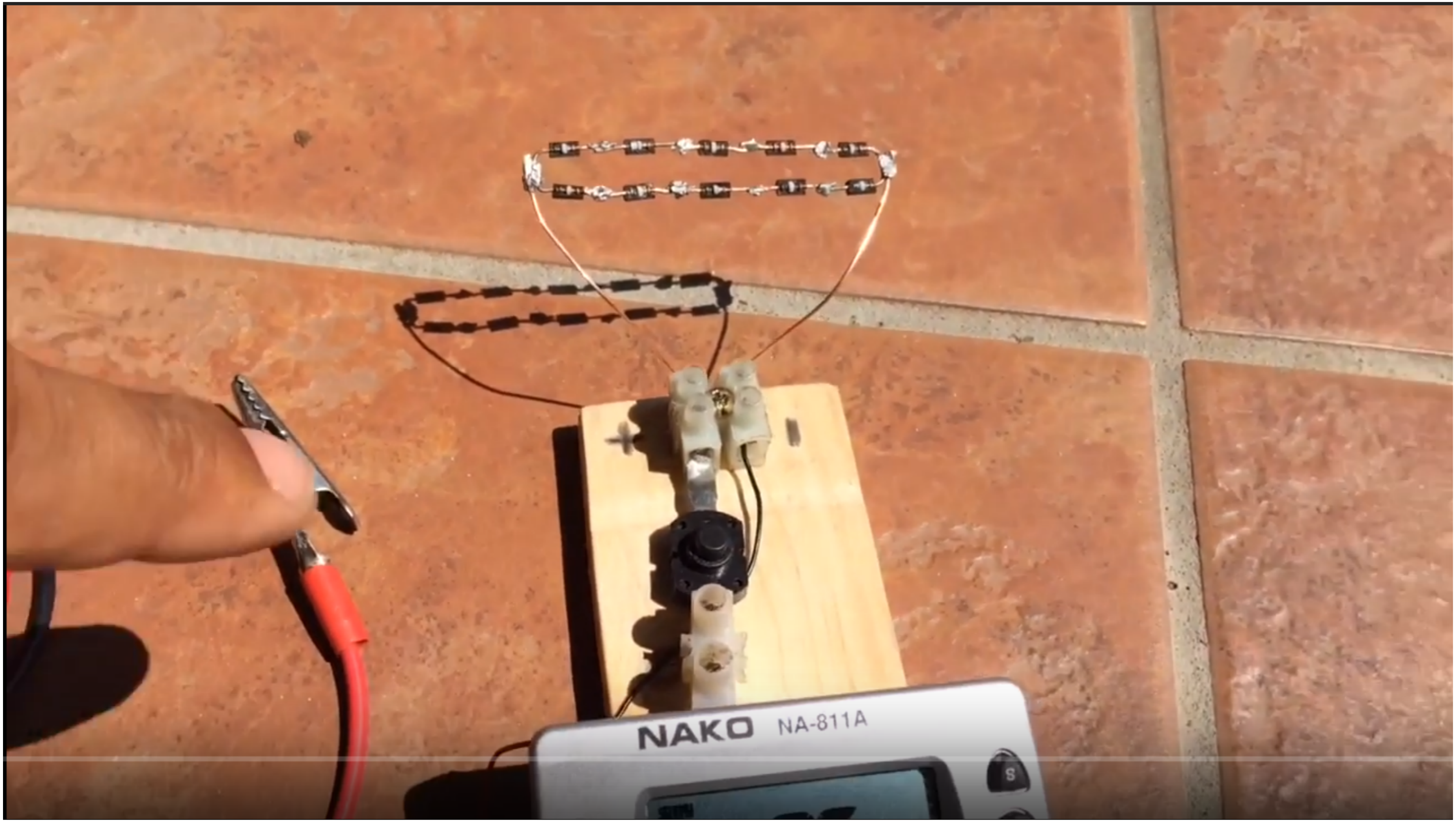
NAKO
NA-811A
S

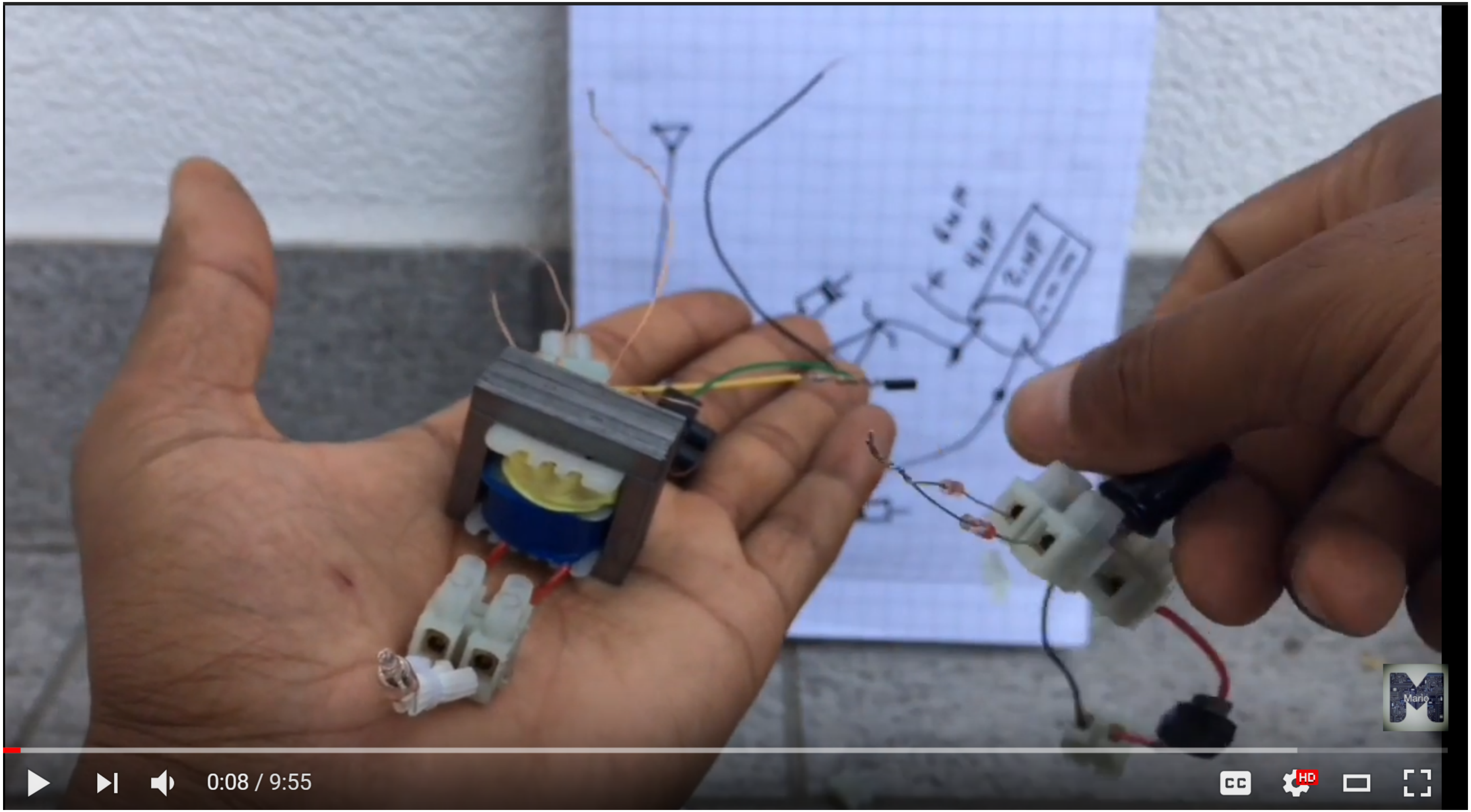
Mario
0:08 / 9:55
HD

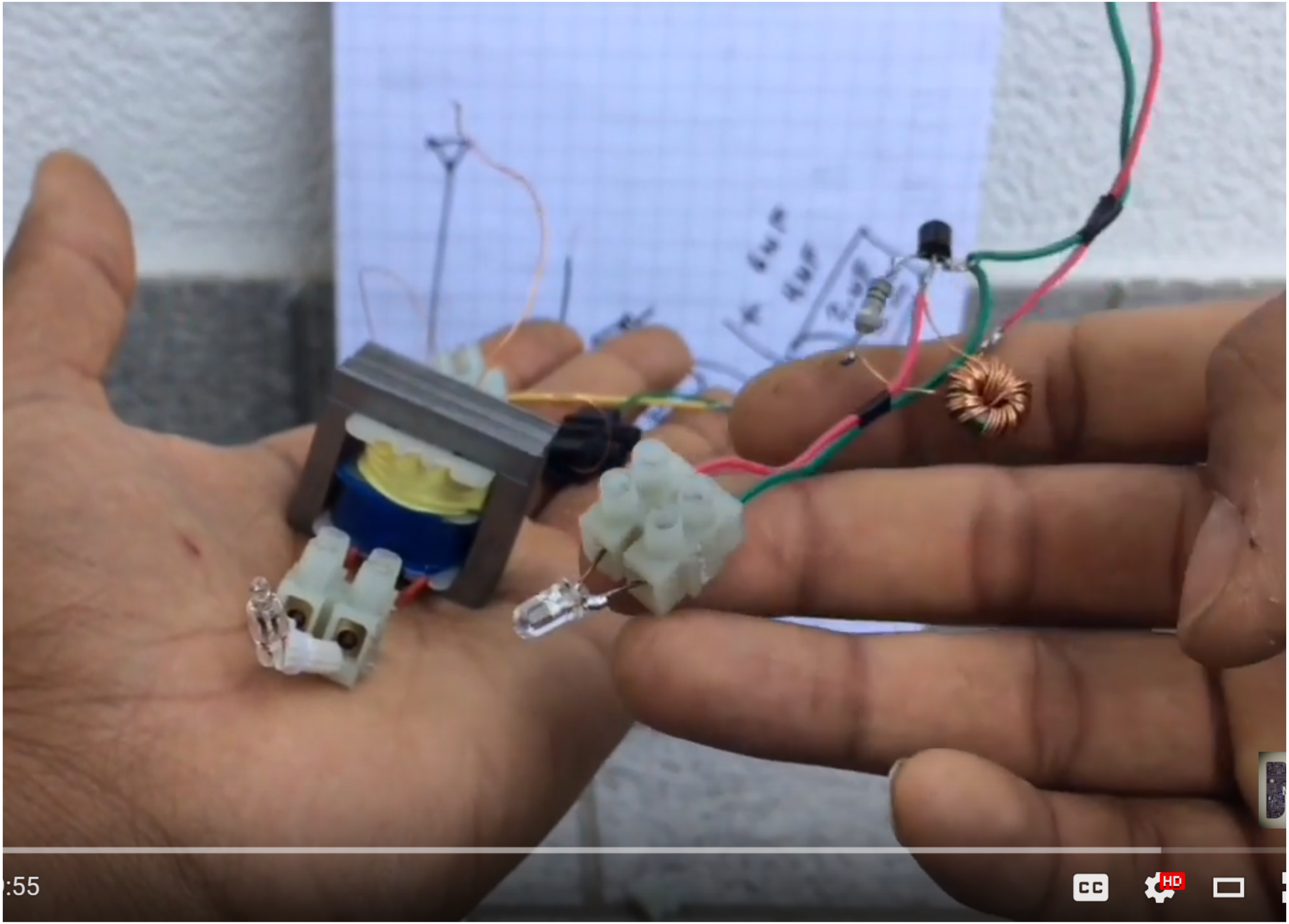
:55
CC
HD

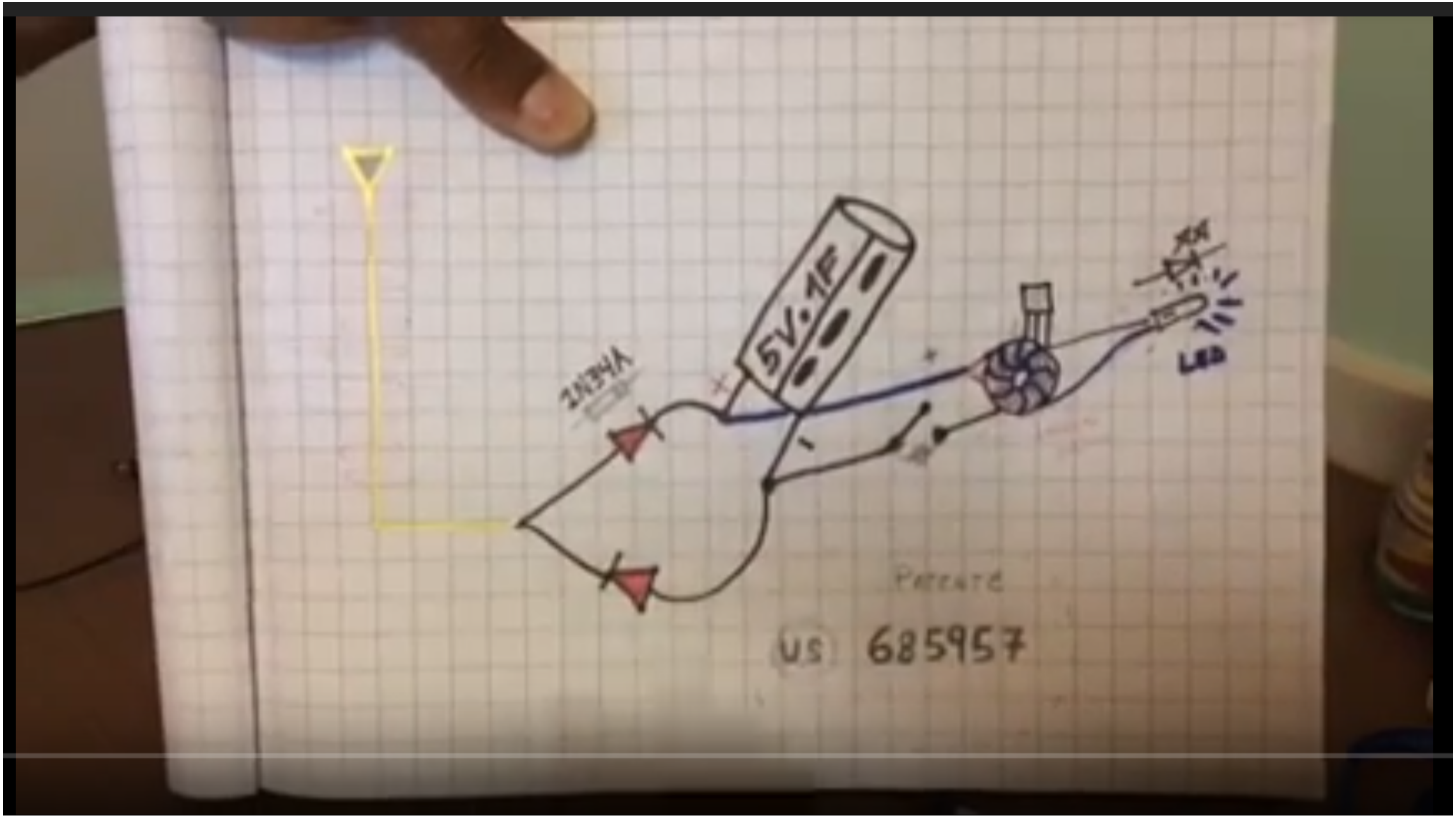
1N34A
+
5V.1F
-
+
LED
Patente
US 685957

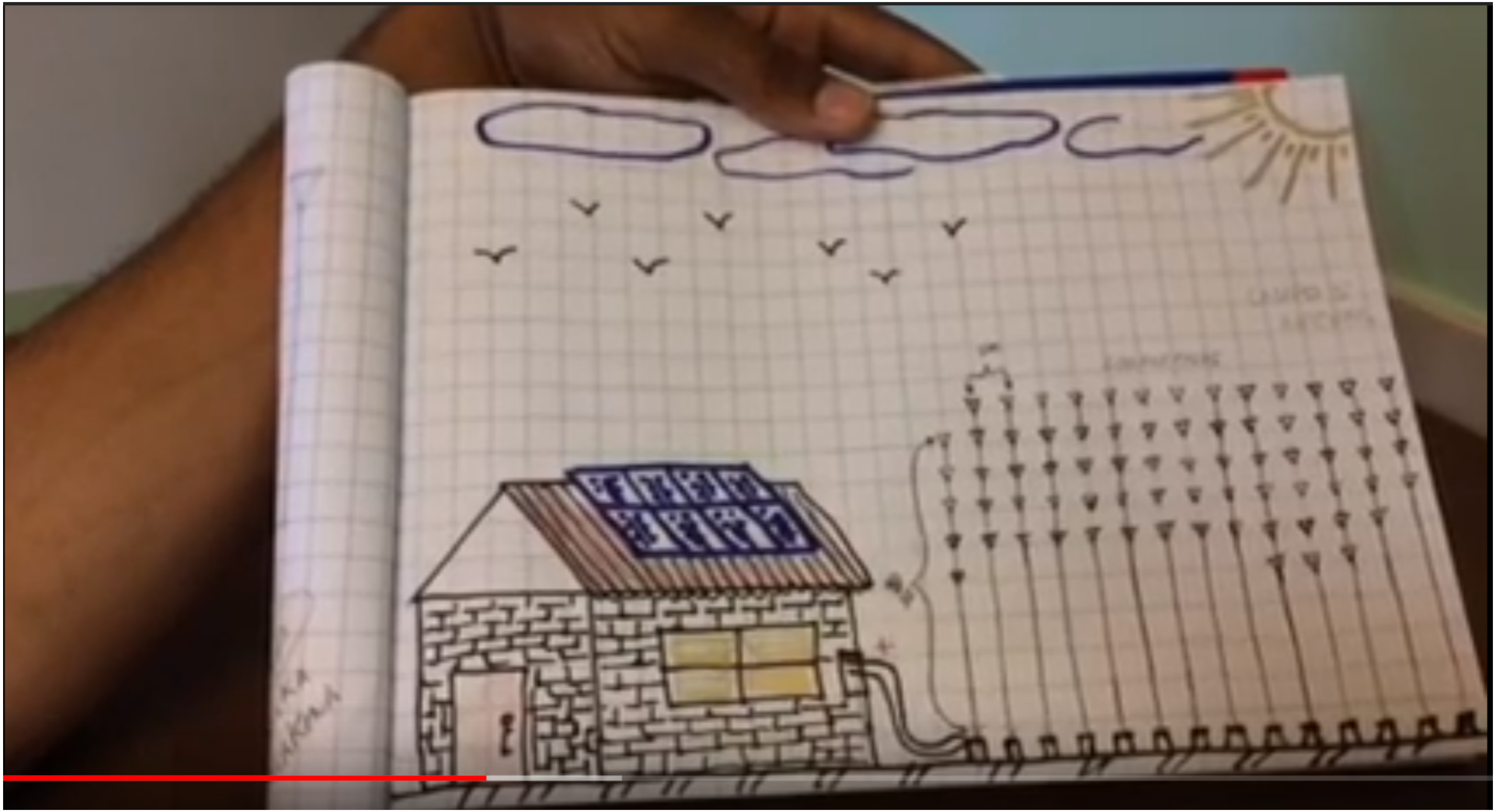

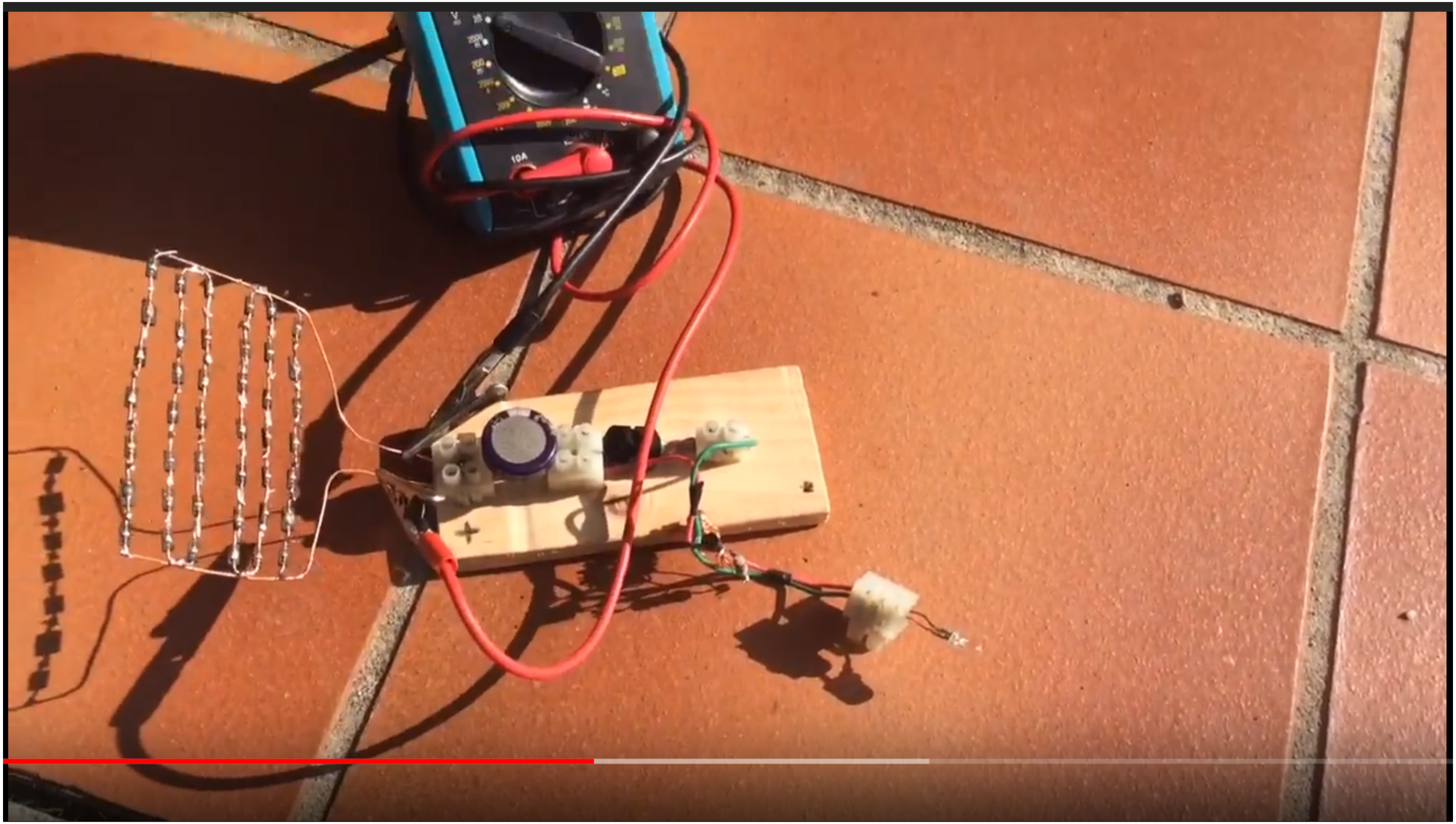
10A

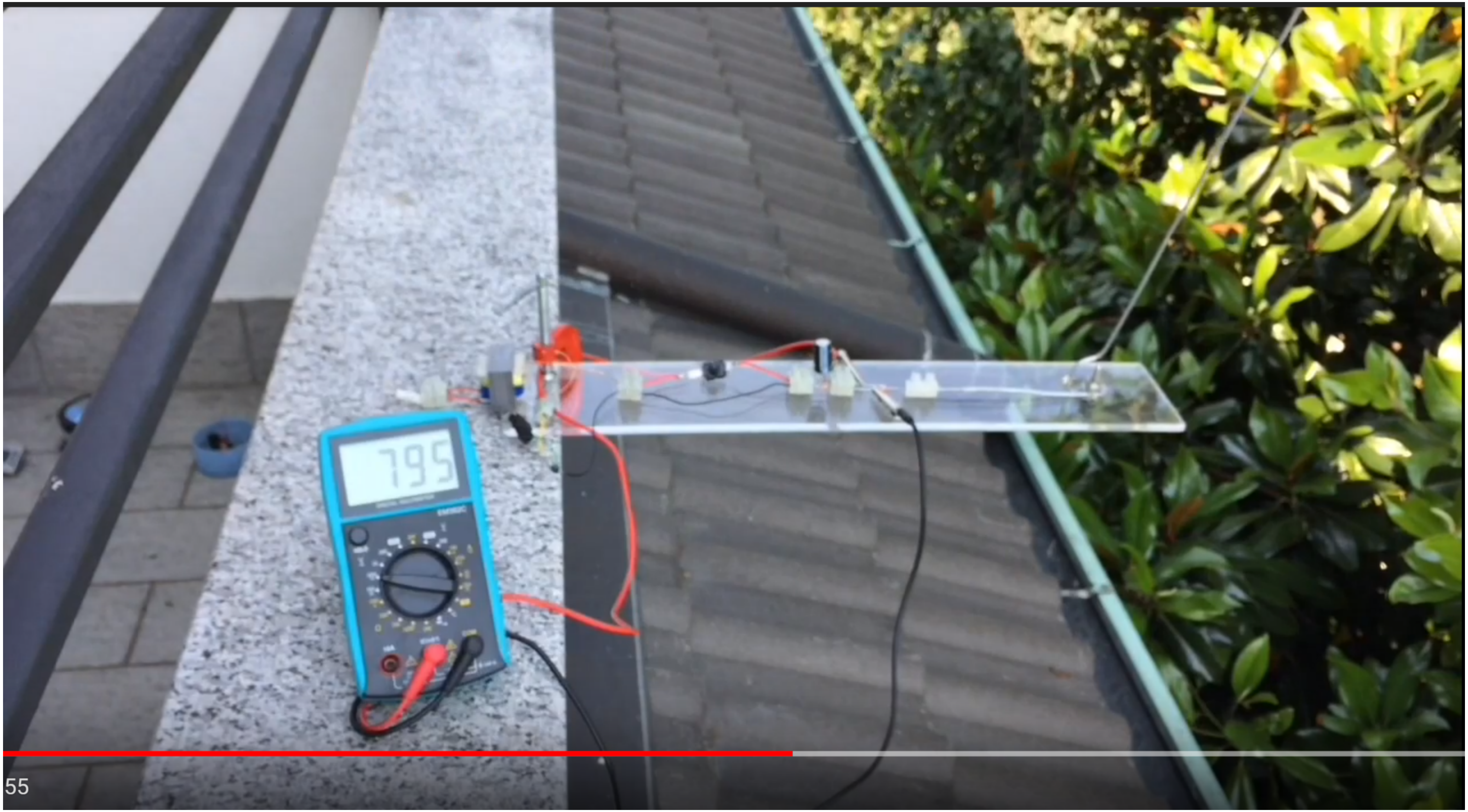
795
55

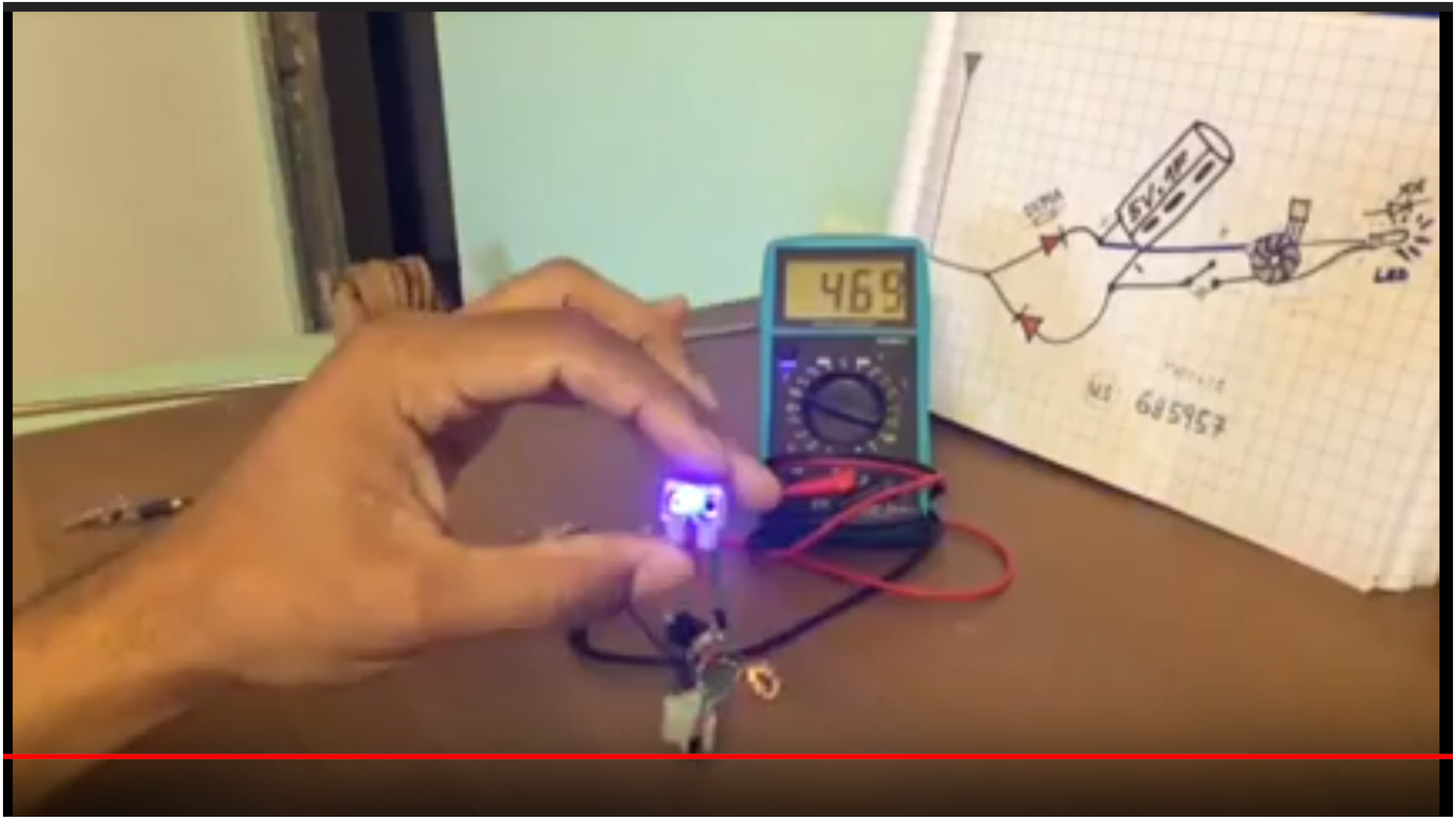
469
5V 1F
LED
685957

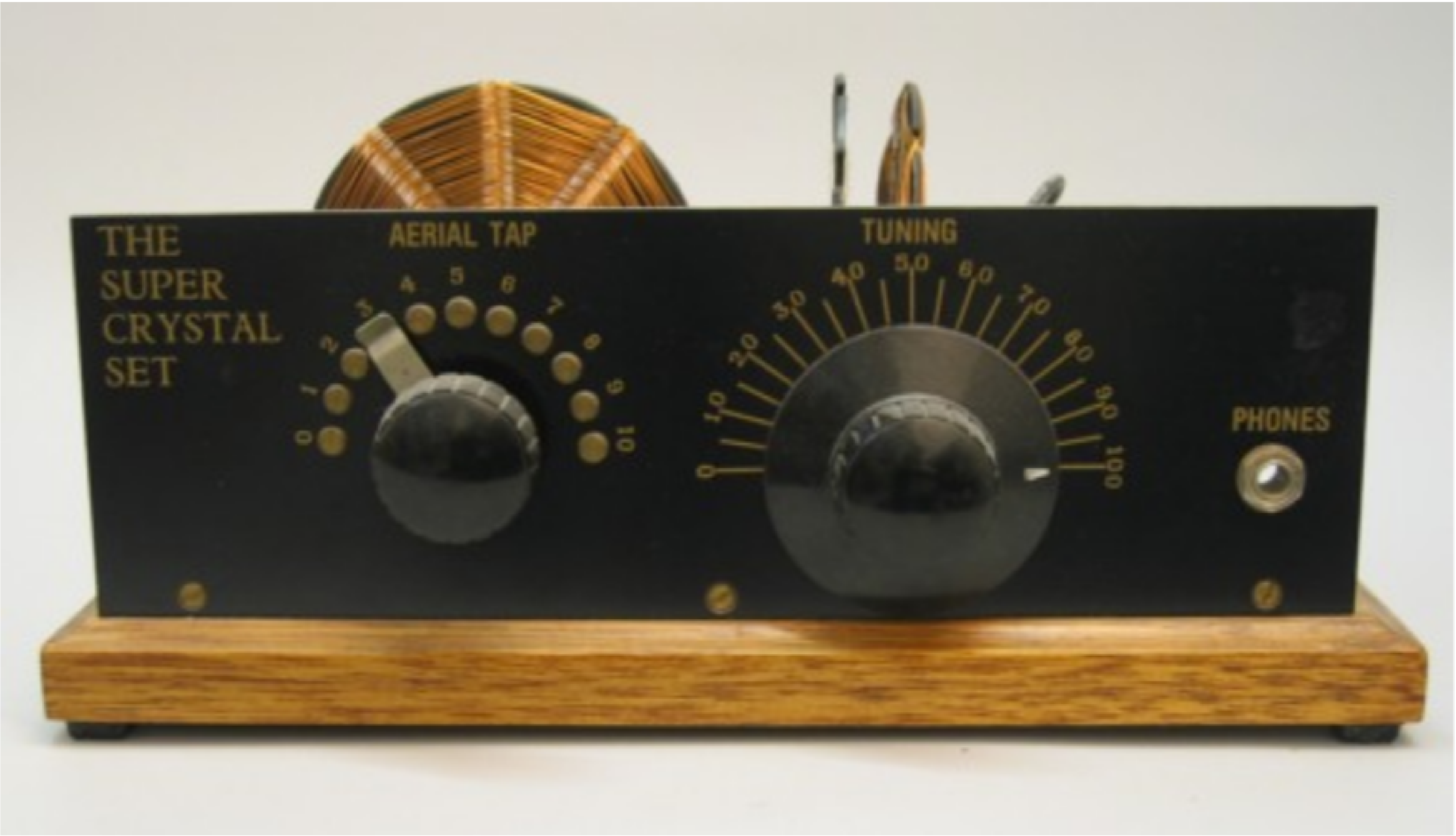
THE SUPER CRYSTAL SET
AERIAL TAP
0
1
2
3
4
5
6
7
8
9
10
TUNING
0
10
20
30
40
50
60
70
80
90
100
PHONES

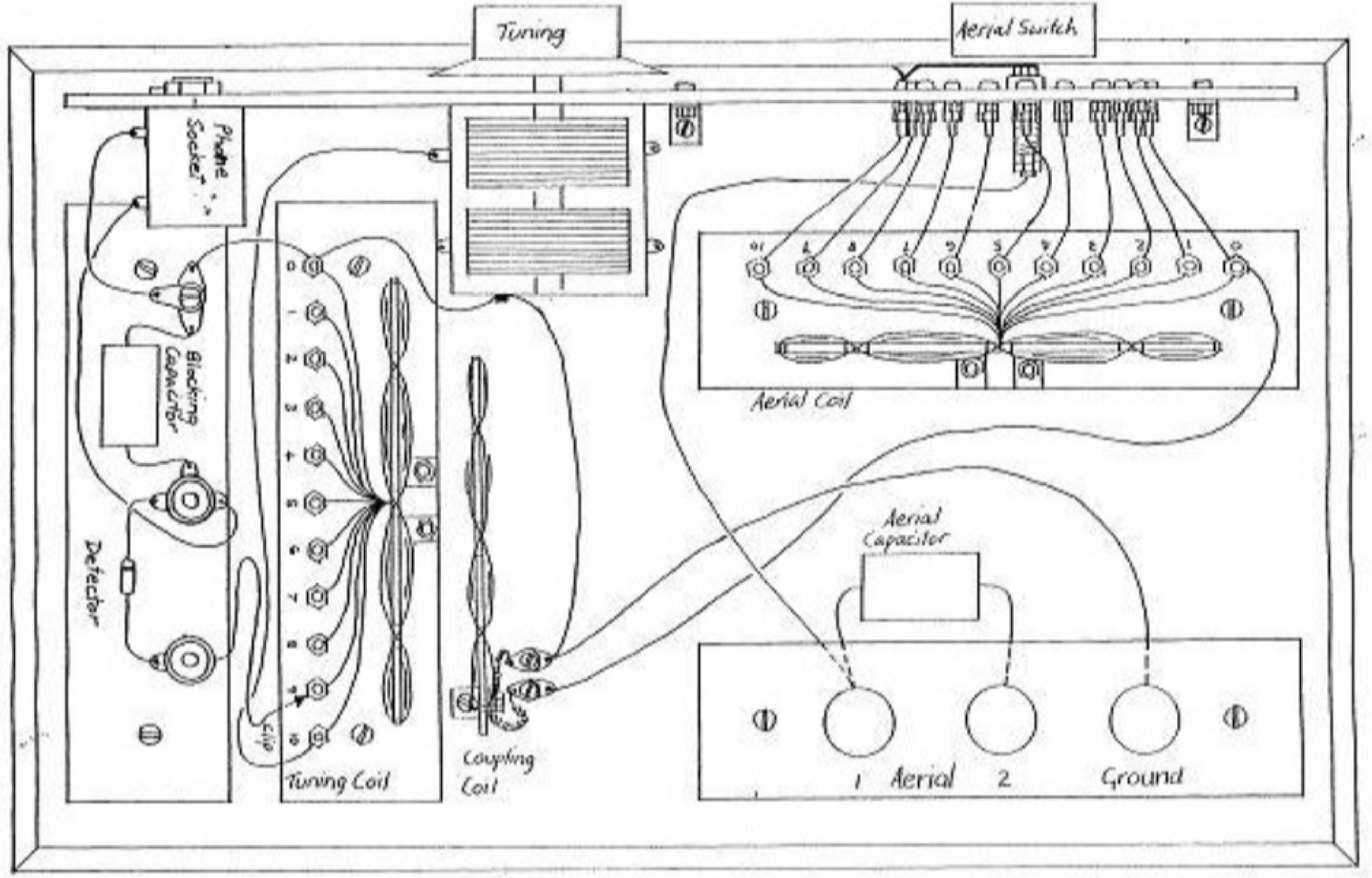
Tuning
Aerial Switch
Phone Socket
Blocking Capacitor
Detector
Clip
Tuning Coil
Coupling Coil
Aerial Coil
Aerial Capacitor
Aerial
Ground

KB8TAD

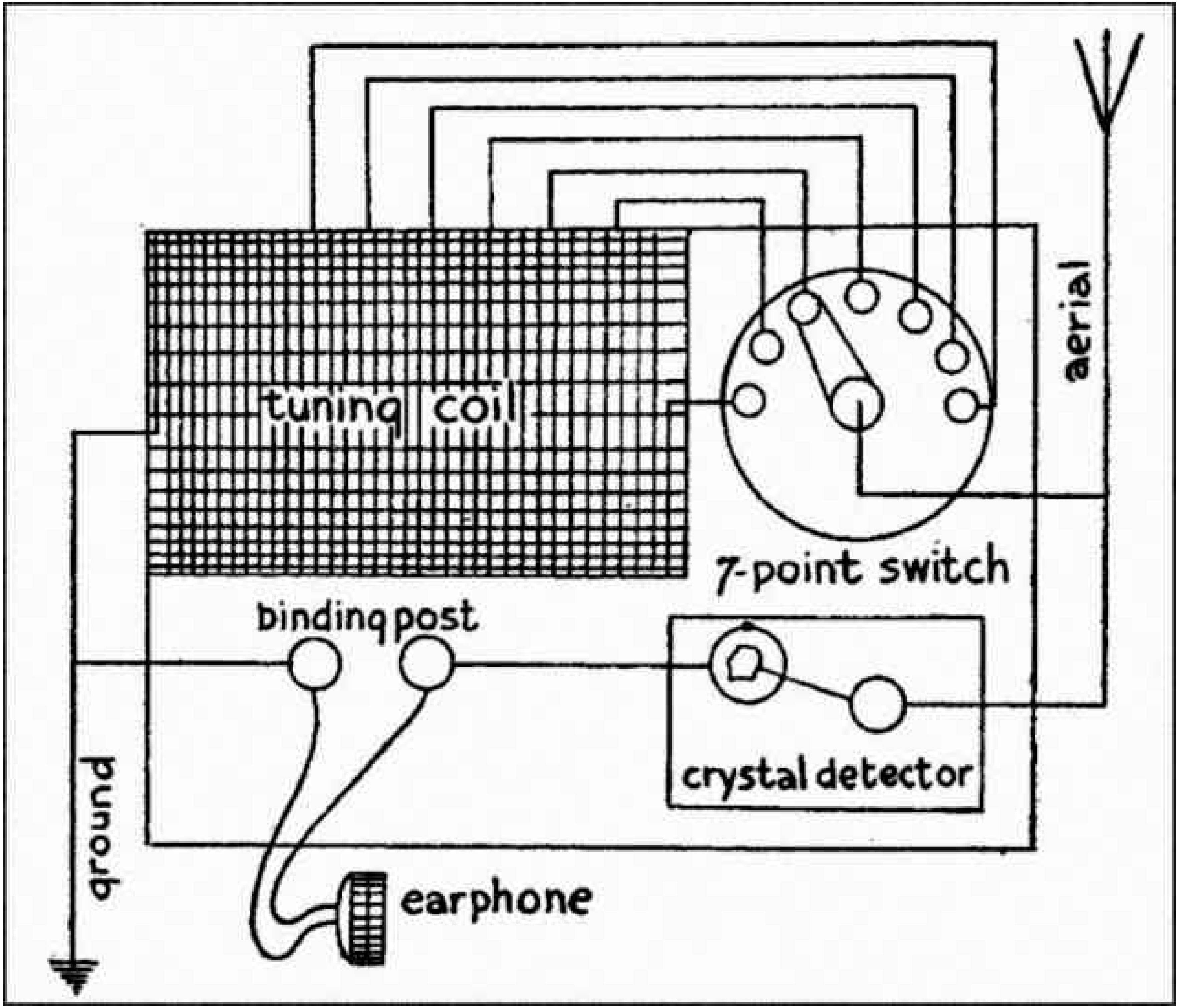
aerial
tuning coil
7-point switch
binding post
crystal detector
ground
earphone

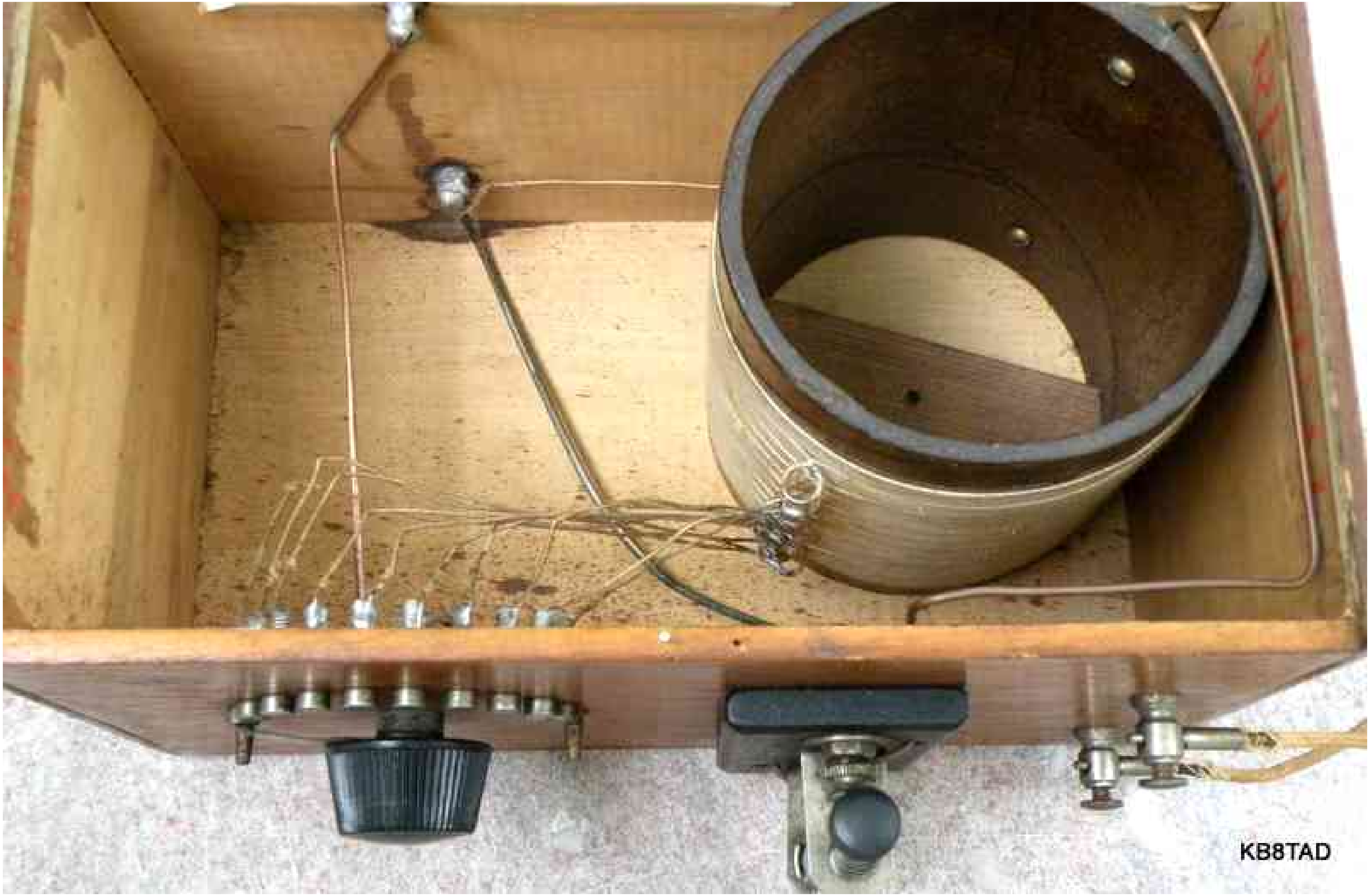
KB8TAD

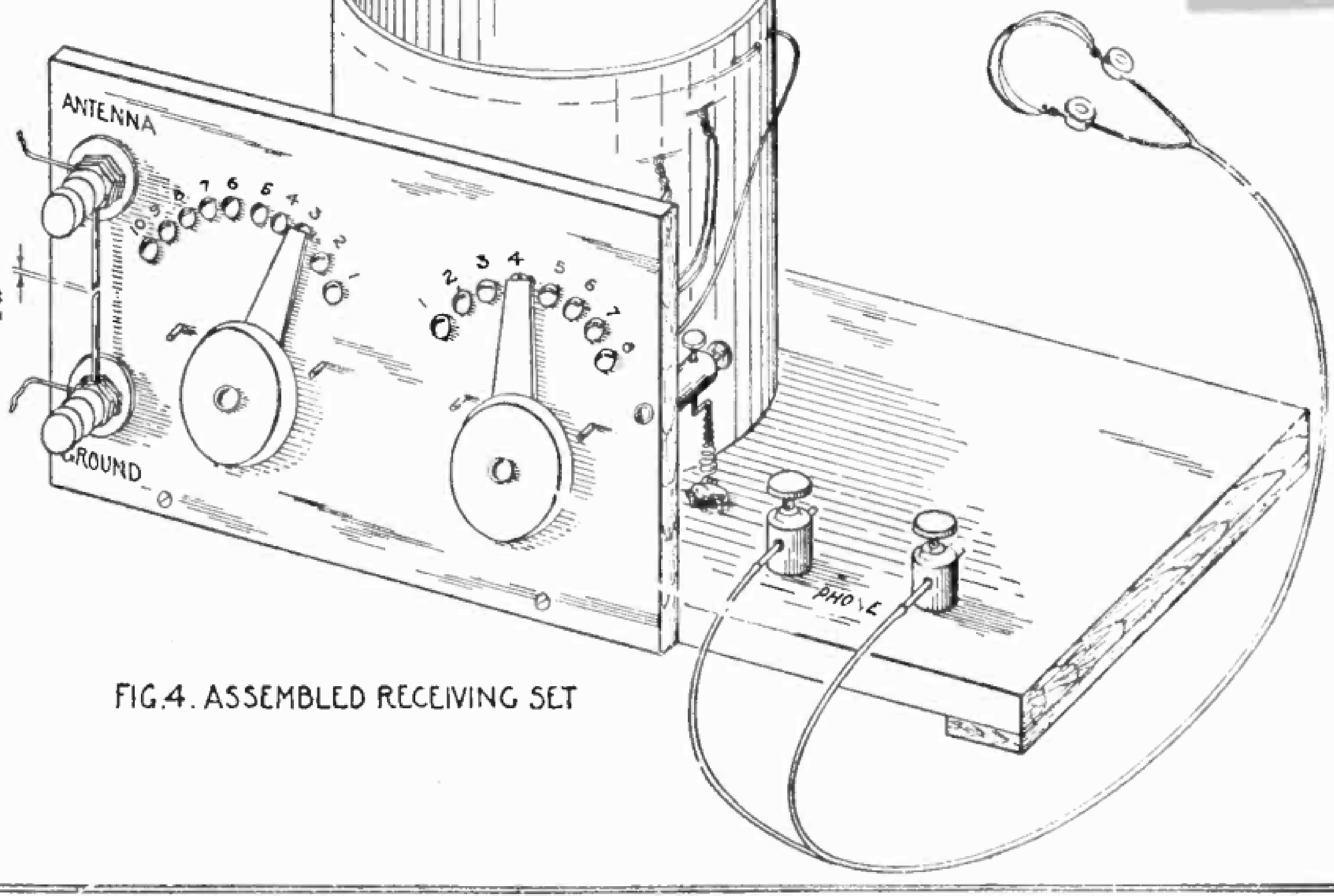
ANTENNA
GROUND
1/16 IN
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
PHONE

FIG. 4. ASSEMBLED RECEIVING SET

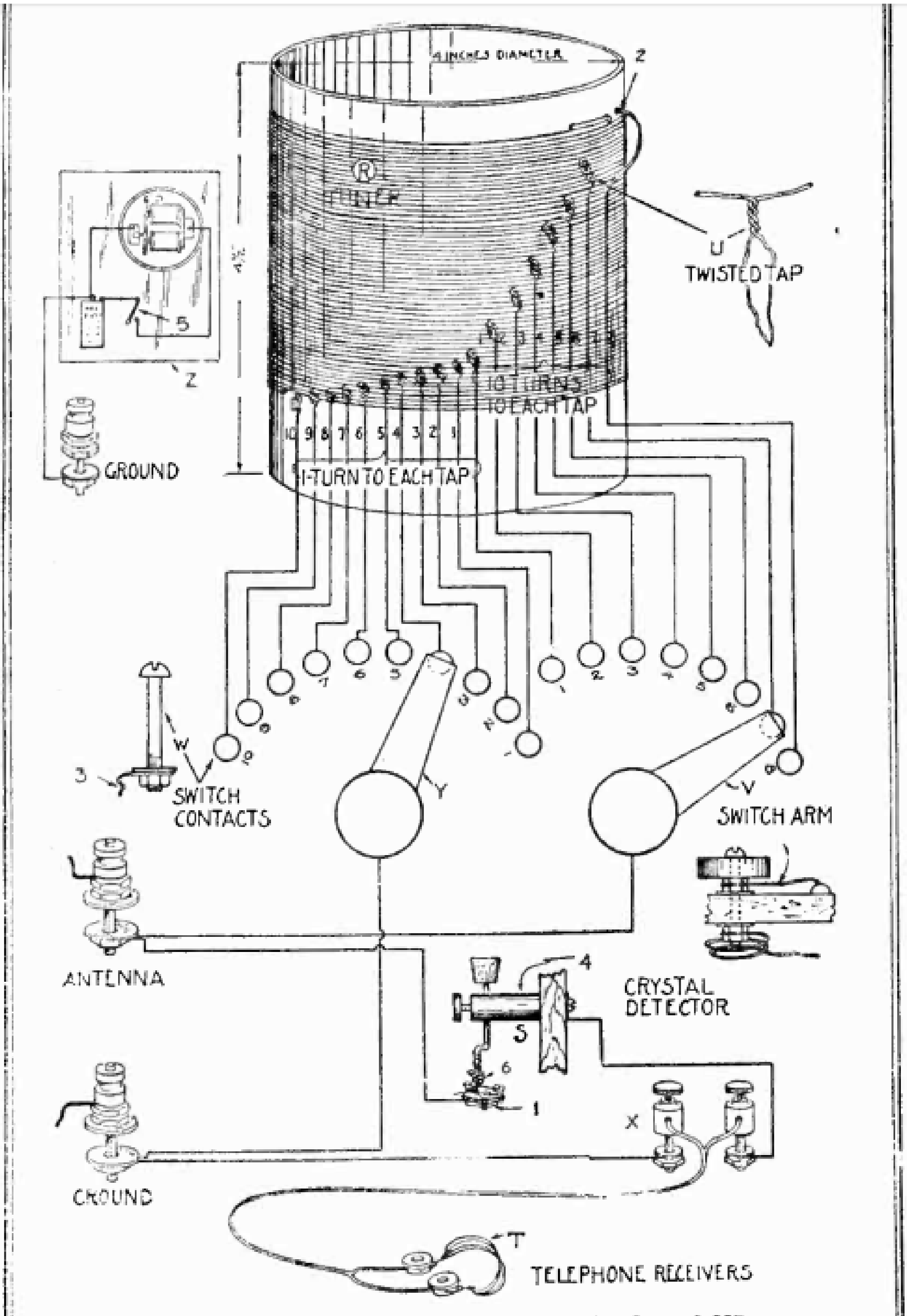
4 INCHES DIAMETER
2
R
TUNER
4½
U
TWISTED TAP
5
Z
GROUND
10 TURNS TO EACH TAP
10 9 8 7 6 5 4 3 2 1
1-TURN TO EACH TAP
W
3
SWITCH CONTACTS
Y
V
SWITCH ARM
ANTENNA
4
CRYSTAL DETECTOR
S
6
1
X
GROUND
T
TELEPHONE RECEIVERS

FIG.3 WIRING DIAGRAM AND DETAILS OF RECEIVING SET

FIG 2. RECEIVING SET, WITH ANTENNA AND GROUND CONNECTIONS

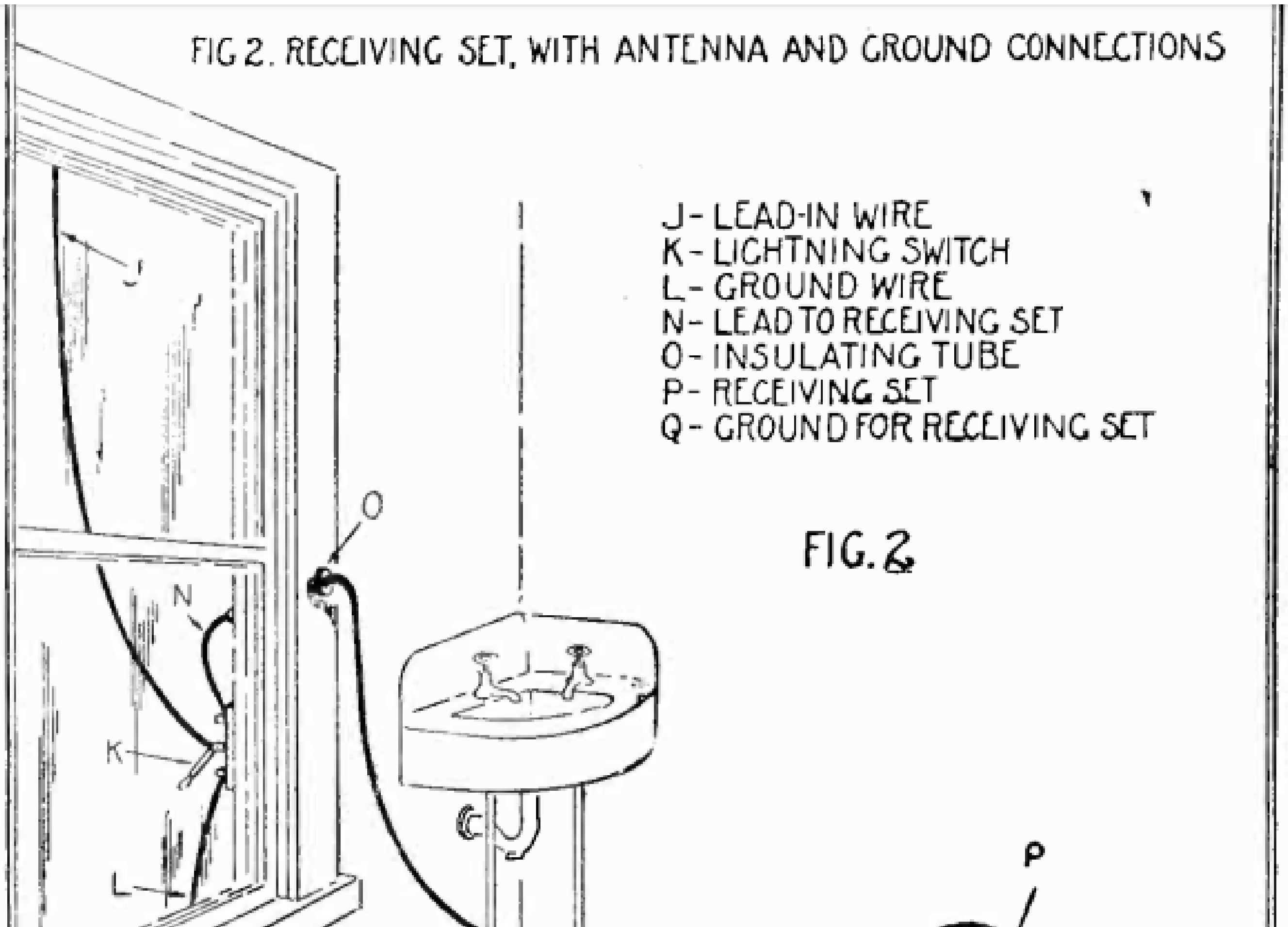
J - LEAD-IN WIRE
K - LIGHTNING SWITCH
L - GROUND WIRE
N - LEAD TO RECEIVING SET
O - INSULATING TUBE
P - RECEIVING SET
Q - GROUND FOR RECEIVING SET
J
O
N
K
L
P
N
Q

FIG. 2

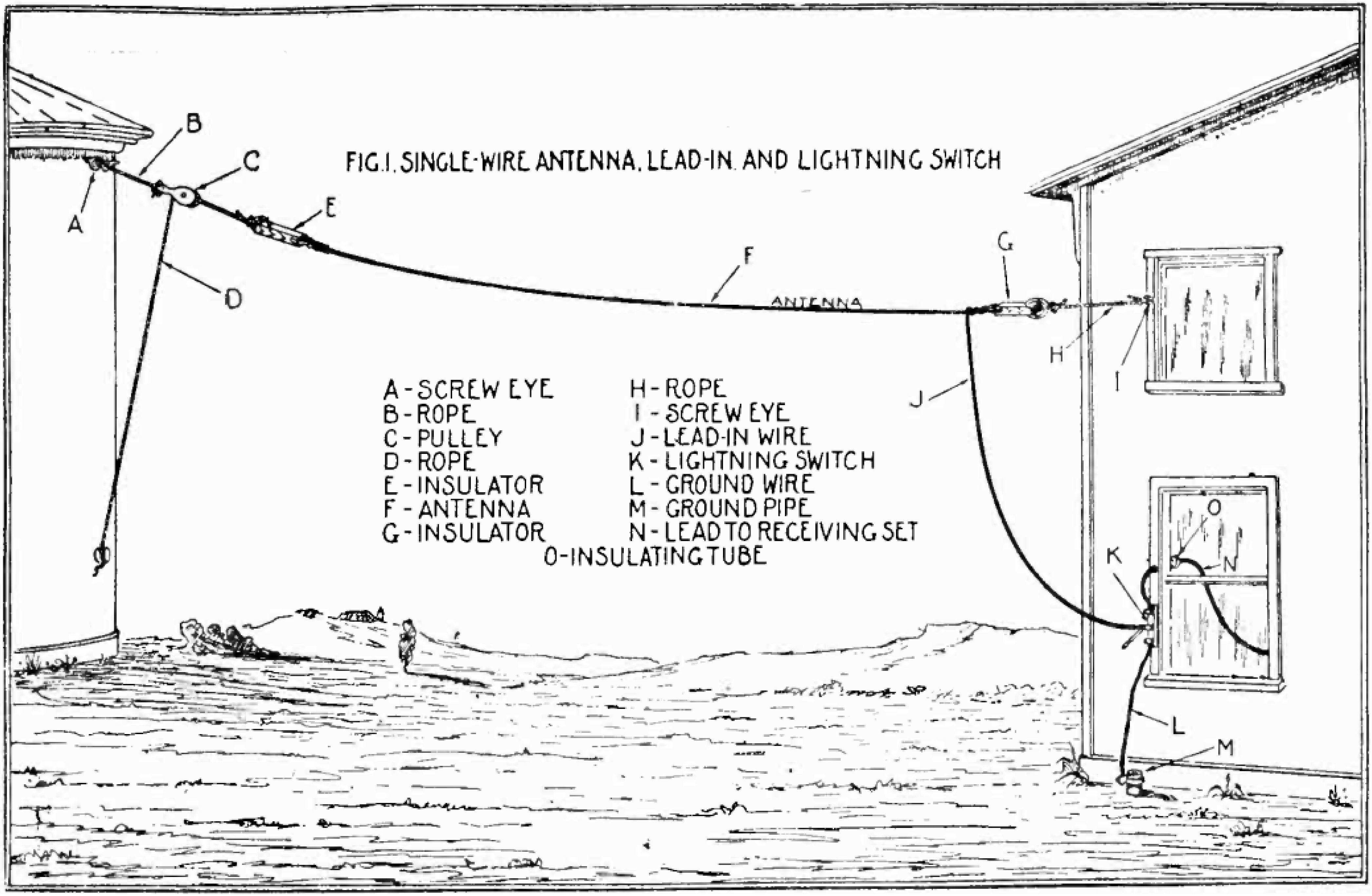

FIG. I. SINGLE-WIRE ANTENNA, LEAD-IN AND LIGHTNING SWITCH

A - SCREW EYE
B - ROPE
C - PULLEY
D - ROPE
E - INSULATOR
F - ANTENNA
G - INSULATOR
H - ROPE
I - SCREW EYE
J - LEAD-IN WIRE
K - LIGHTNING SWITCH
L - GROUND WIRE
M - GROUND PIPE
N - LEAD TO RECEIVING SET
O - INSULATING TUBE
ANTENNA
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O

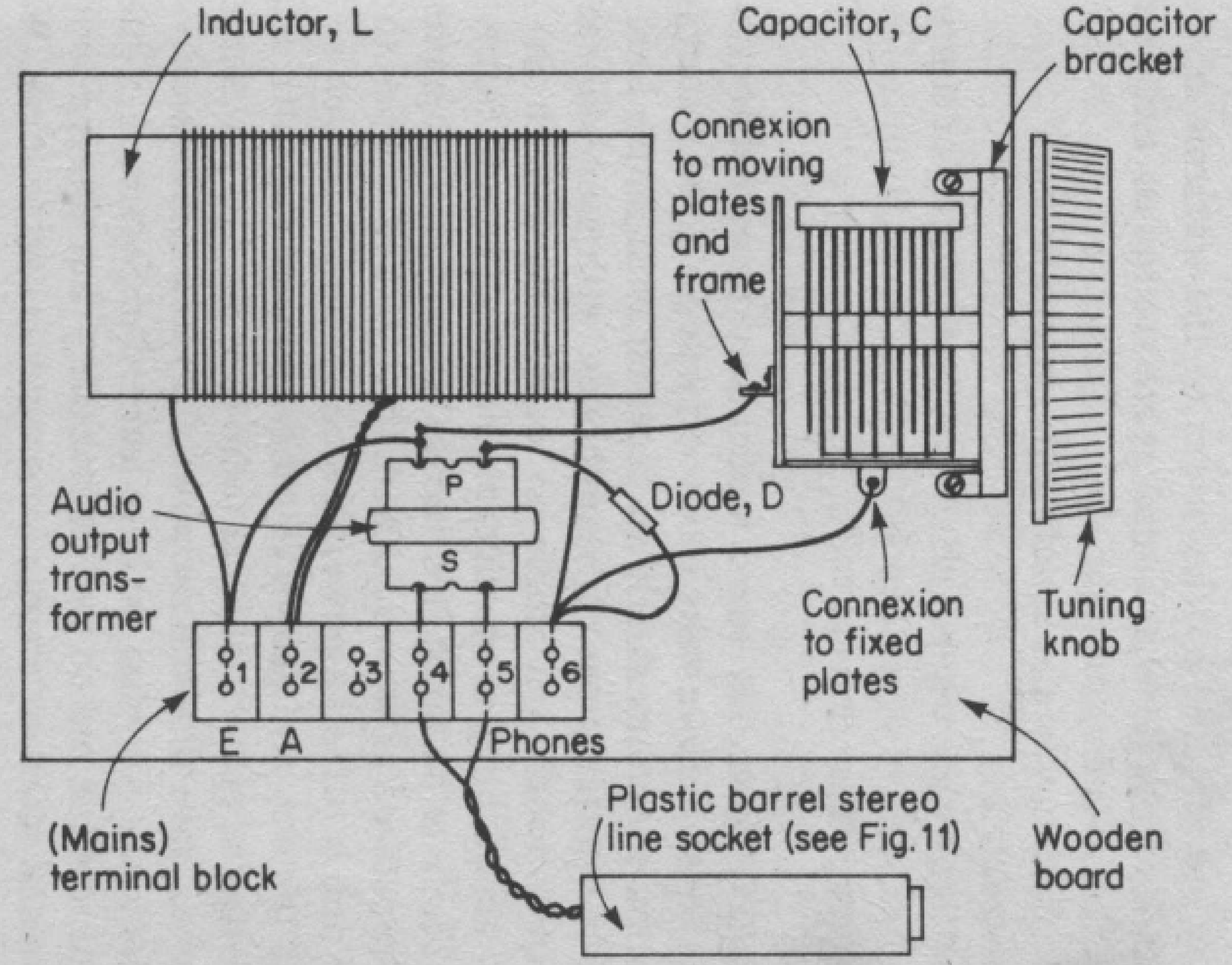
Inductor, L
Capacitor, C
Capacitor bracket
Connexion to moving plates and frame
P
Audio output transformer
Diode, D
S
Connexion to fixed plates
Tuning knob
1
2
3
4
5
6
E
A
Phones
(Mains) terminal block
Plastic barrel stereo line socket (see Fig.11)
Wooden board

Fig. 14 Suggested component arrangement for elementary crystal set

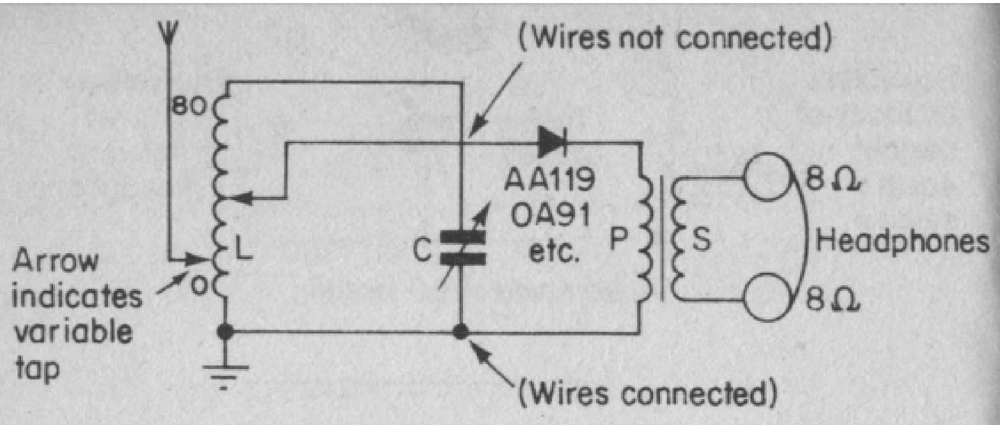
(Wires not connected)
80
L
C
AA119
OA91
etc.
P
S
8 Ω
Headphones
8 Ω
Arrow indicates variable tap
0
(Wires connected)

(i) Circuit

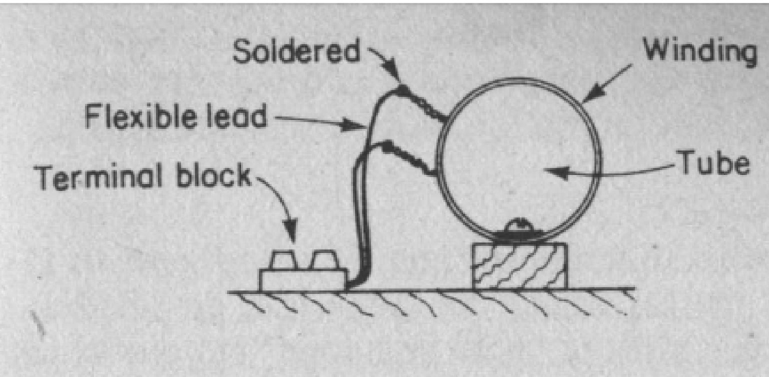
Soldered
Winding
Flexible lead
Tube
Terminal block

(iii) Alternative mounting

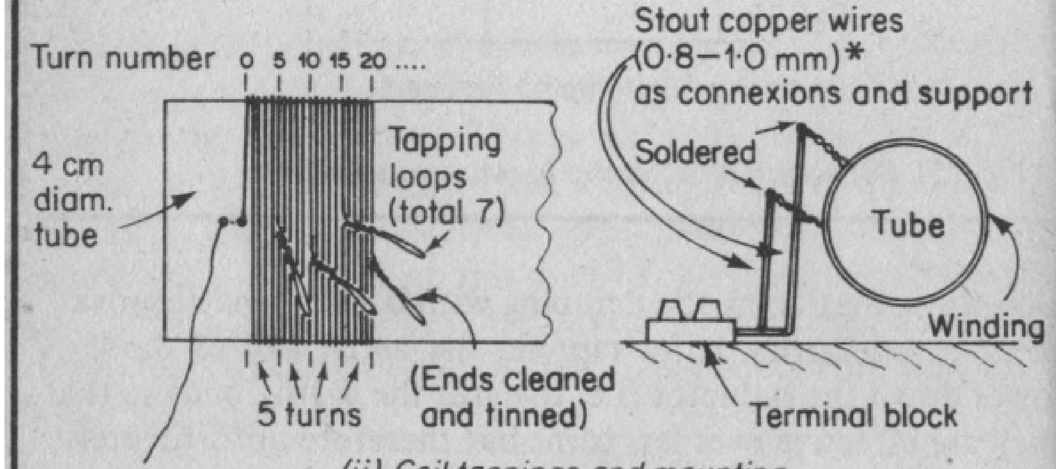
Turn number 0 5 10 15 20 ....
4 cm diam. tube
Tapping loops (total 7)
5 turns
(Ends cleaned and tinned)
Stout copper wires (0·8–1·0 mm)* as connexions and support
Soldered
Tube
Winding
Terminal block

(ii) Coil tappings and mounting

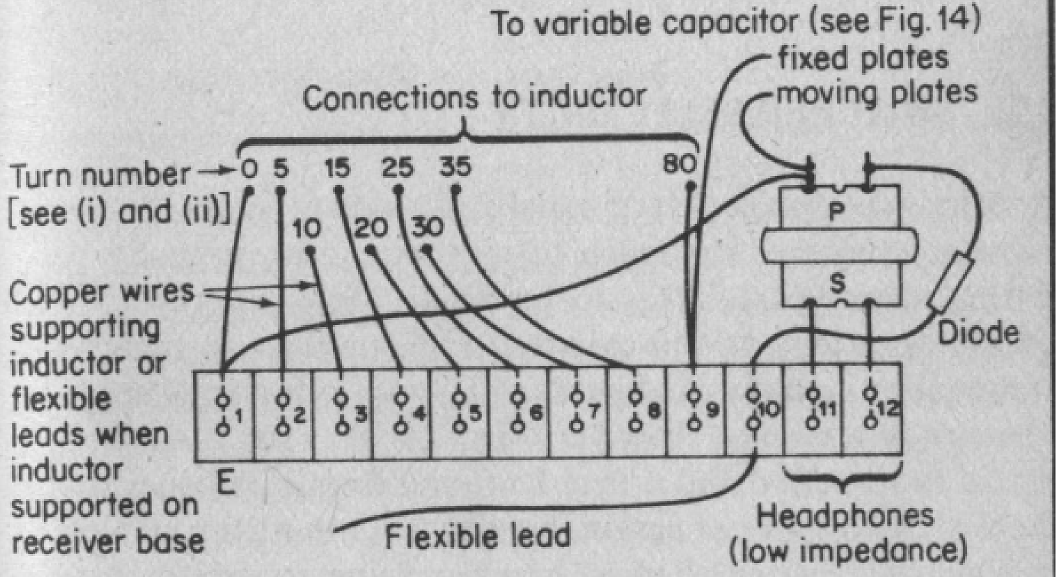
To variable capacitor (see Fig. 14)
fixed plates
moving plates
Connections to inductor
Turn number → 0 5 15 25 35 80
[see (i) and (ii)]
10 20 30
Copper wires supporting inductor or flexible leads when inductor supported on receiver base
P
S
Diode
1 2 3 4 5 6 7 8 9 10 11 12
E
Flexible lead
Headphones (low impedance)

(iv) Suggested component arrangement

Fig. 16 Circuit with tapped inductor

---

* 21–19 SWG, 20–18 AWG

required) and a *flying lead* for the diode, these components are shown in Fig.16(iv). This is doing the job properly, there is no reason however (except for unreliability of contact) why both aerial and diode flying leads should not terminate on *crocodile clips* and be clipped onto the chosen taps as required.

Earth is connected to terminal 1 and the headphones to 11 and 12. Both the aerial and the diode lead are tried at terminals 2 – 9, a game of "poke and hope" but one which can be rewarding. Remember that generally but not always, the tappings with the lower numbers reduce loudness but increase selectivity. Aerial to terminal 8, diode to 9 brings us back to Fig.13(i) of course.

5.2 SWITCHED SELECTIVITY

A technique of selection of aerial and diode tappings by means of rotary switches has much to recommend it especially if setting changes are likely to be needed for reception of different stations. In this case two single-pole 8-way switches are required, usually obtainable as 12-way so leaving 4 spare. The drawing symbol shown in Fig.17 speaks for itself with regard to its action and it may be found from catalogues that there are two types of *action*, by which is meant the manner in which the switch changes over from one contact to the adjoining one. *Break before make* switches disconnect one circuit before the next is connected and *make before break* have a change-over period during which both circuits are connected at once for a short time. Either type is suitable for us. A knob to rotate each switch is also required, a small one, say 2 cm diameter or less is ample, if with an indicator line or pointer, so much the better. The modifications to Fig.16 for switching are given in Fig.17.

With such rotary switches a search for the optimum arrange-

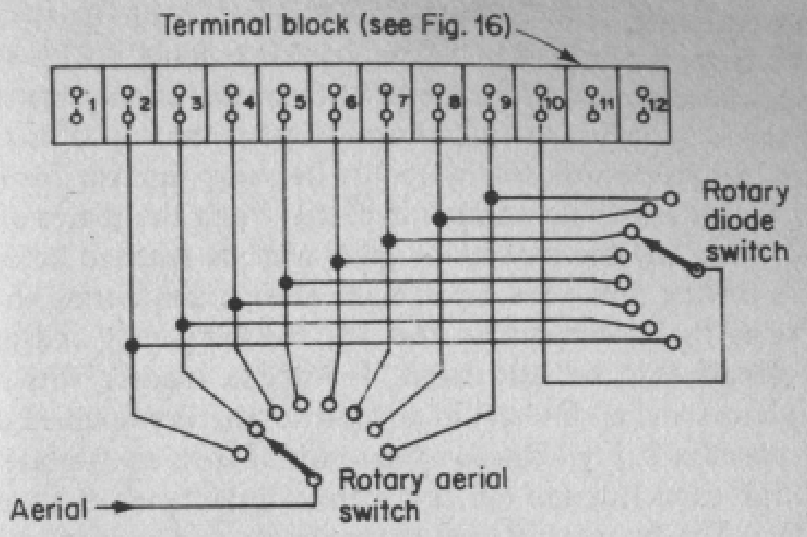

Fig. 17 - Switching aerial and diode leads

variable capacitor, the two switches and the headphone socket, the latter must then be of the chassis type. A pair of terminals for aerial and earth completes the receiver. On the panel the two rotary switches are easily labelled 1 – 8 but we are still uncertain about a tuning dial because we may not have the right variable capacitor for the inductor, this is the subject of the next section.

5.3 GETTING THE RANGE RIGHT

This is where we pause in receiver construction to remove some of the confusion which may still exist with regard to the tuning ingredients, wavelength, frequency, inductance and capacity, each of which has some bearing on the others, to fit together these four pieces of the puzzle, so to speak. For most crystal sets the inductance is fixed according to the variable capacitor available and we tune the receiver over the range by rotating the variable capacitor knob over half a

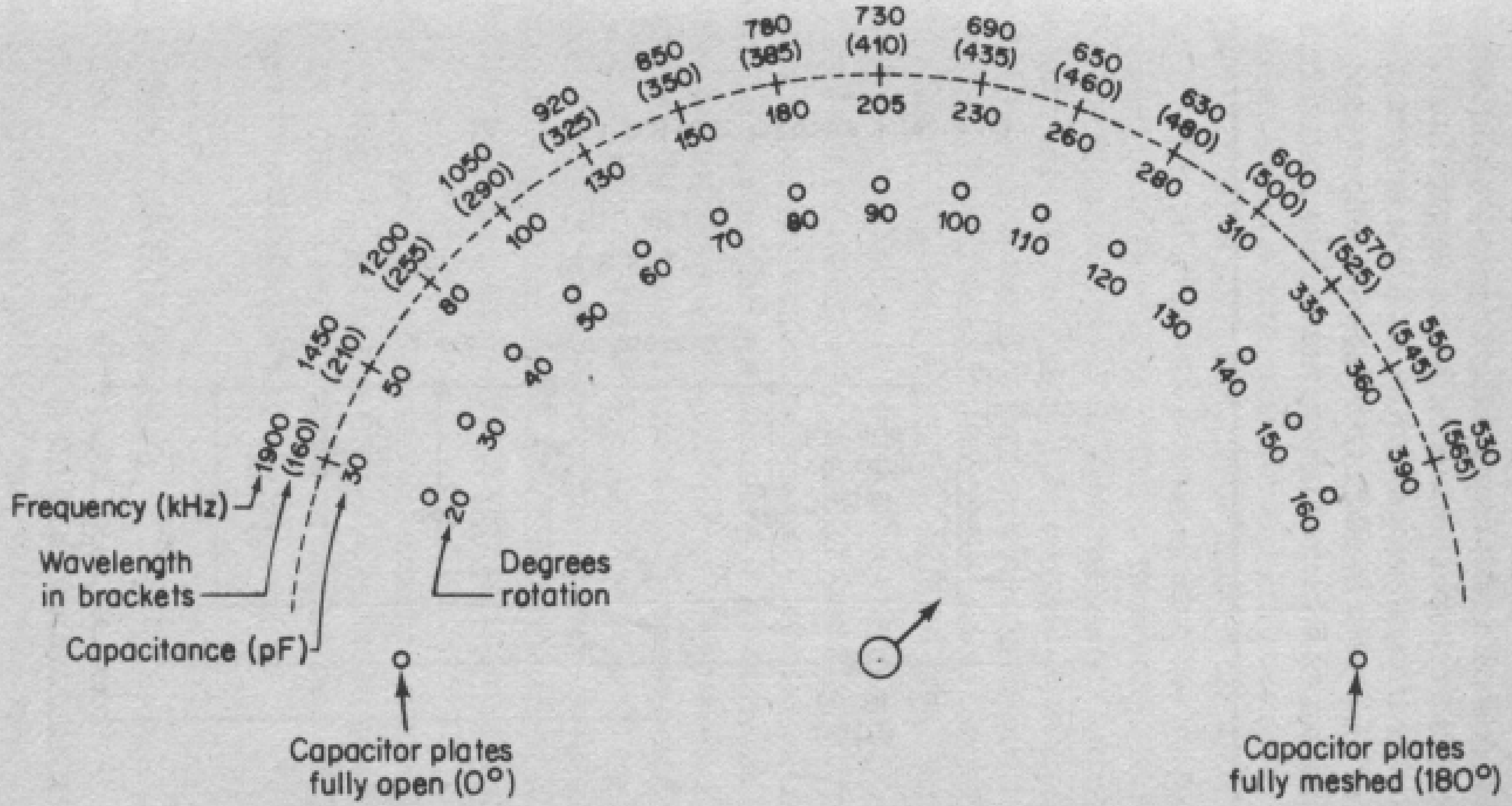
1900
(160)
30
1450
(210)
50
1200
(255)
80
1050
(290)
100
920
(325)
130
850
(350)
150
780
(385)
180
730
(410)
205
690
(435)
230
650
(460)
260
630
(480)
280
600
(500)
310
570
(525)
335
550
(545)
360
530
(565)
390
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Frequency (kHz)
Wavelength in brackets
Capacitance (pF)
Degrees rotation
Capacitor plates fully open (0°)
Capacitor plates fully meshed (180°)

Fig. 18 Approximate capacitance and frequency values for 365 pF variable capacitor and 230 μH coil

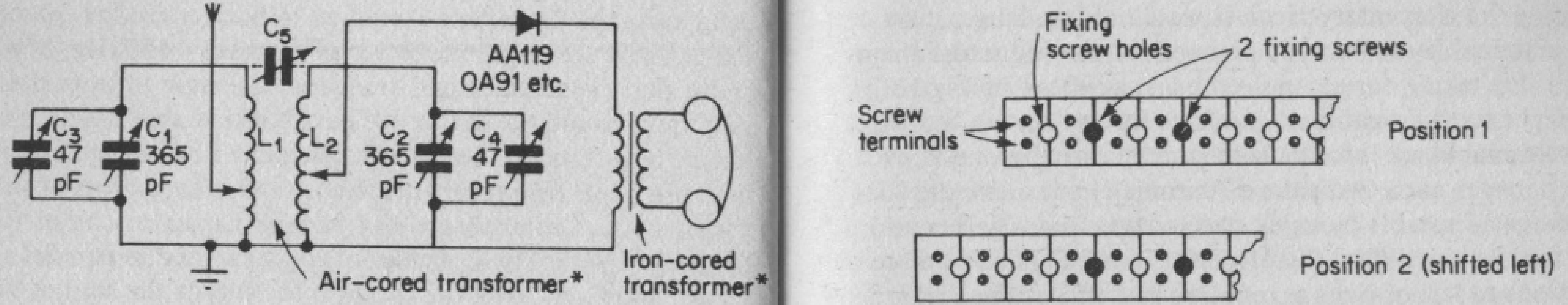
C5
AA119
OA91 etc.
C3
47
pF
C1
365
pF
L1
L2
C2
365
pF
C4
47
pF
Air-cored transformer*
Iron-cored transformer*
Fixing screw holes
2 fixing screws
Screw terminals
Position 1
Position 2 (shifted left)

(i) Circuit

(iii) Moving inductor L_1

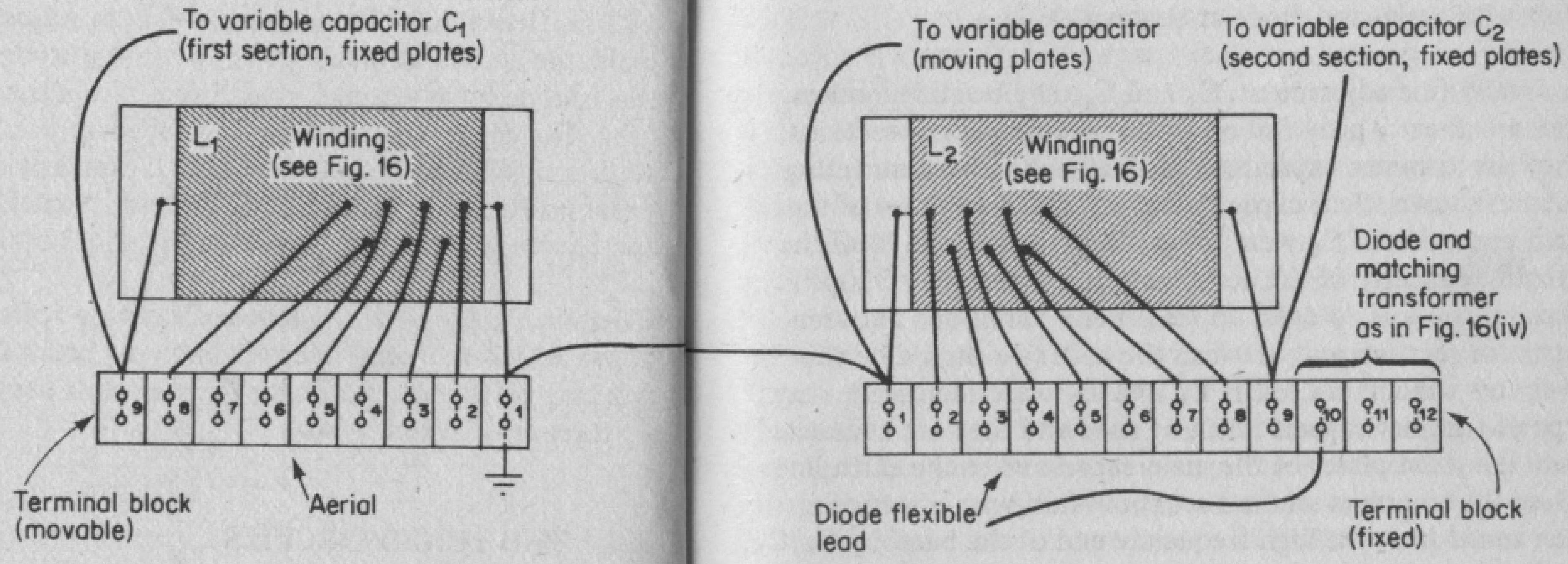
To variable capacitor C1
(first section, fixed plates)
L1
Winding
(see Fig. 16)
9
8
7
6
5
4
3
2
1
Terminal block
(movable)
Aerial
To variable capacitor
(moving plates)
To variable capacitor C2
(second section, fixed plates)
L2
Winding
(see Fig. 16)
Diode and
matching
transformer
as in Fig. 16(iv)
1
2
3
4
5
6
7
8
9
10
11
12
Diode flexible
lead
Terminal block
(fixed)

(ii) Arrangement of inductors

Fig. 21 A double-tuned crystal receiver

* See Section 2.4

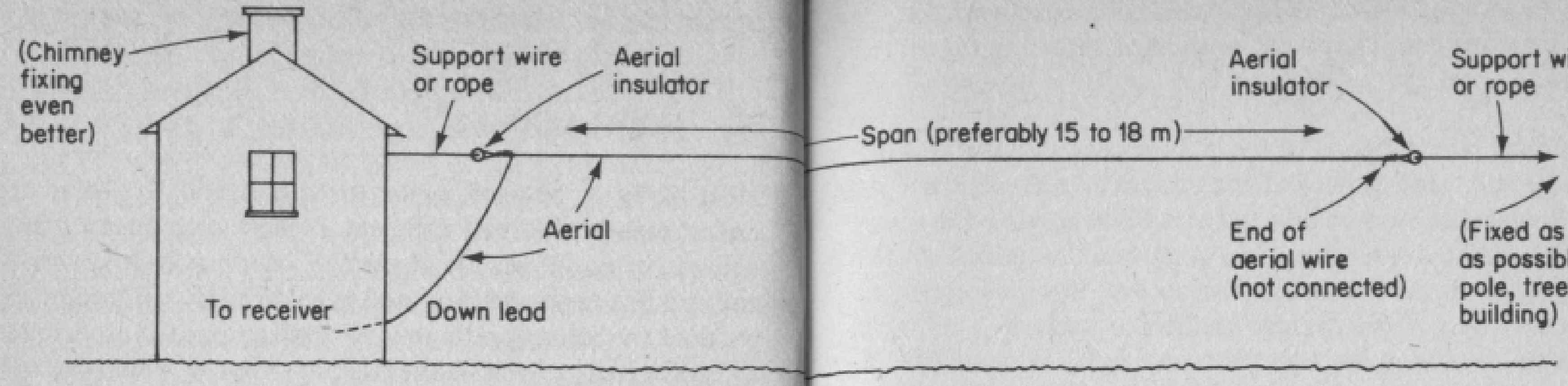
(Chimney fixing even better)
Support wire or rope
Aerial insulator
Aerial
To receiver
Down lead
Span (preferably 15 to 18 m)
Aerial insulator
Support wire or rope
End of aerial wire (not connected)
(Fixed as high as possible to pole, tree or building)

(i) Outdoor Aerial

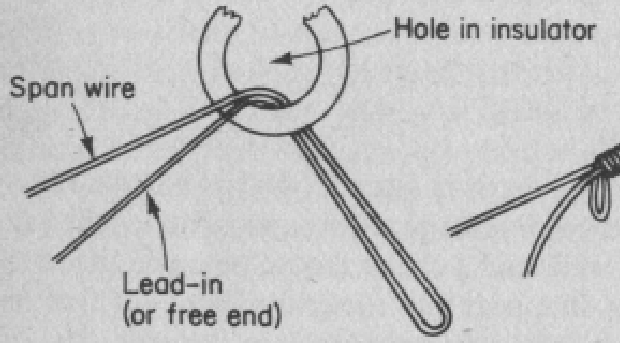
Hole in insulator
Span wire
Lead-in (or free end)

(ii) Attaching aerial to insulator

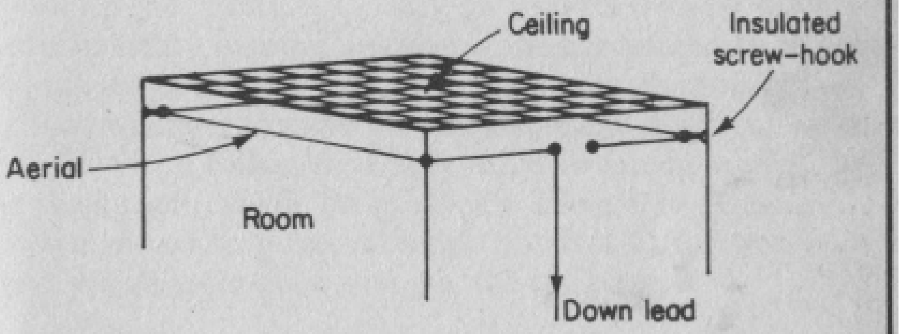
Ceiling
Insulated screw-hook
Aerial
Room
Down lead

(iii) Room aerial

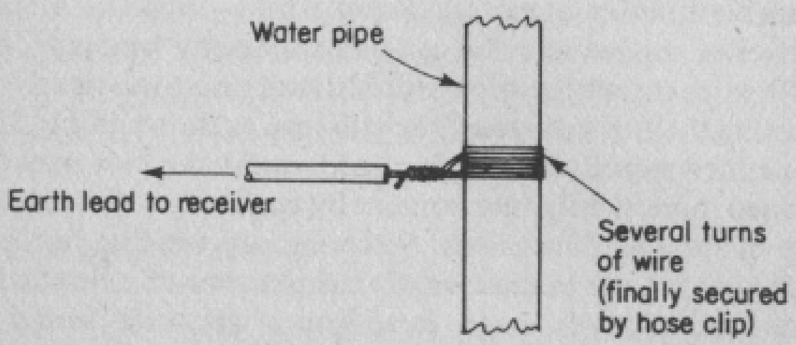
Water pipe
Earth lead to receiver
Several turns of wire (finally secured by hose clip)

(iv) Earth connection to water pipe

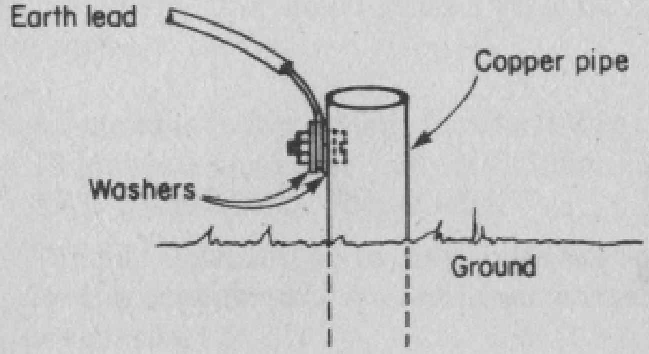
Earth lead
Copper pipe
Washers
Ground

(v) Lead bolted to earth spike

Fig. 12 Aerial/earth systems

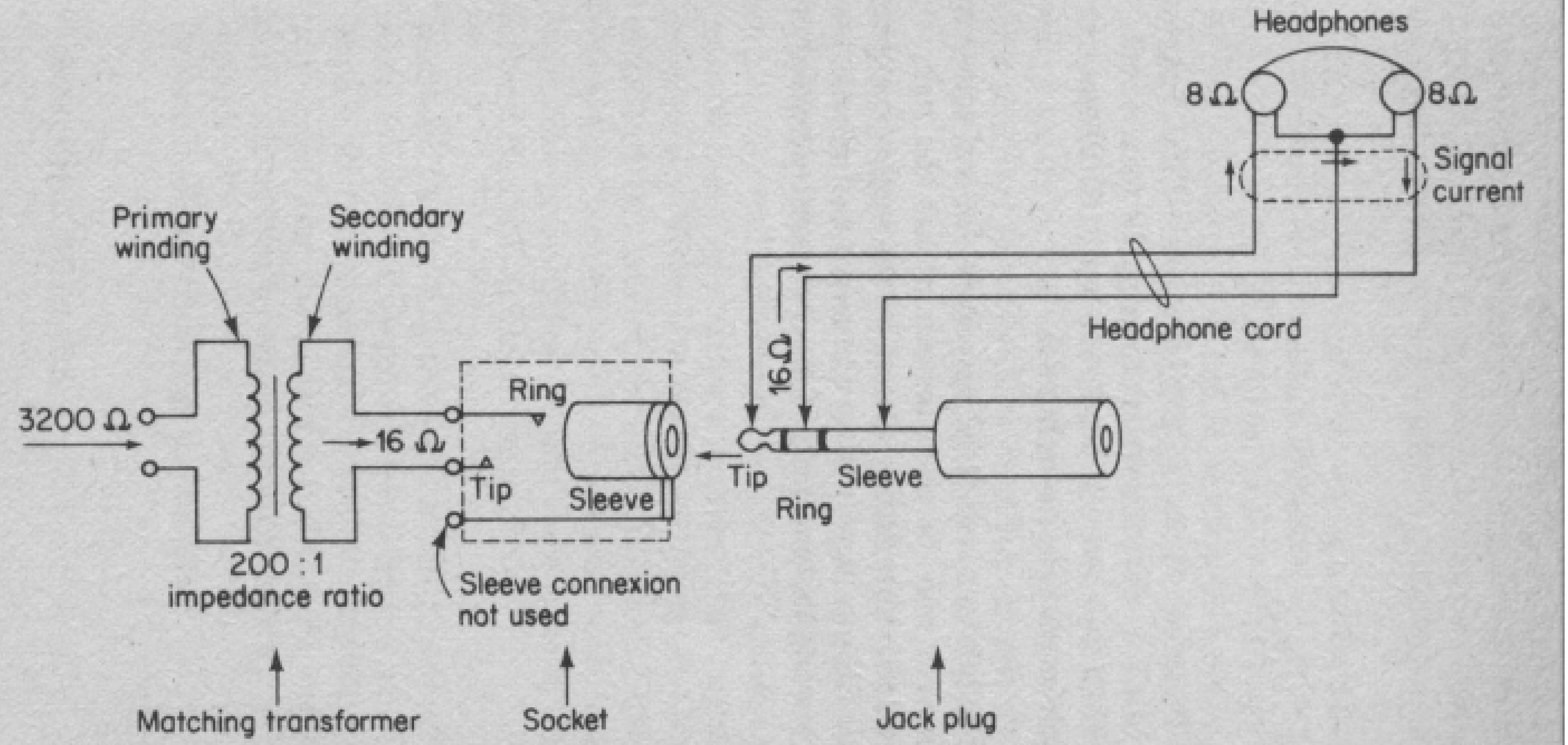
Headphones
8 Ω
8 Ω
Signal current
Primary winding
Secondary winding
Headphone cord
16 Ω
Ring
3200 Ω
→16 Ω
Tip
Sleeve
Tip
Sleeve
Ring
200 : 1 impedance ratio
Sleeve connexion not used
Matching transformer
Socket
Jack plug

Fig. 11 Matching a high impedance receiver to low impedance headphones

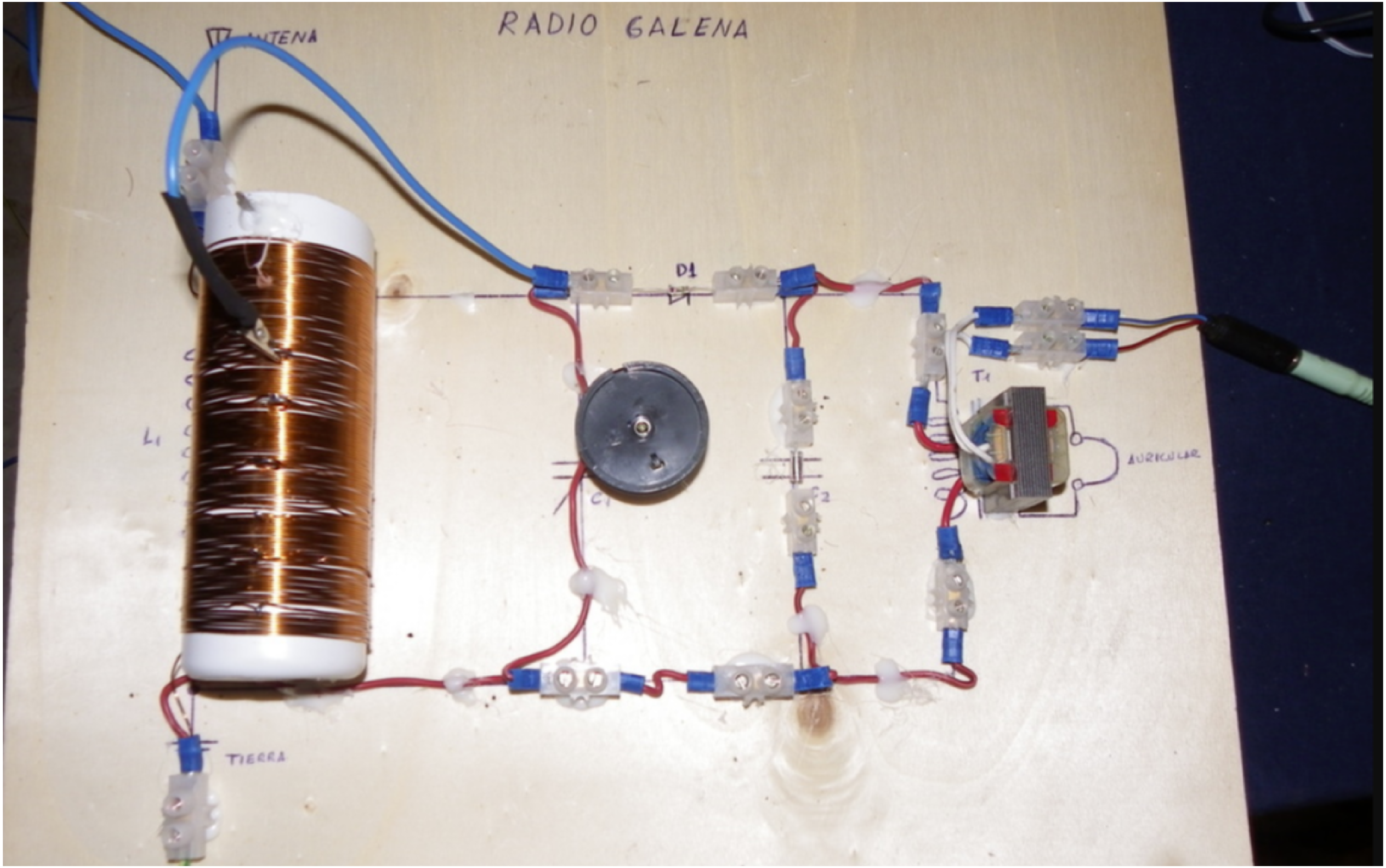
RADIO GALENA
ANTENA
D1
L1
C1
C2
T1
AURICULAR
TIERRA

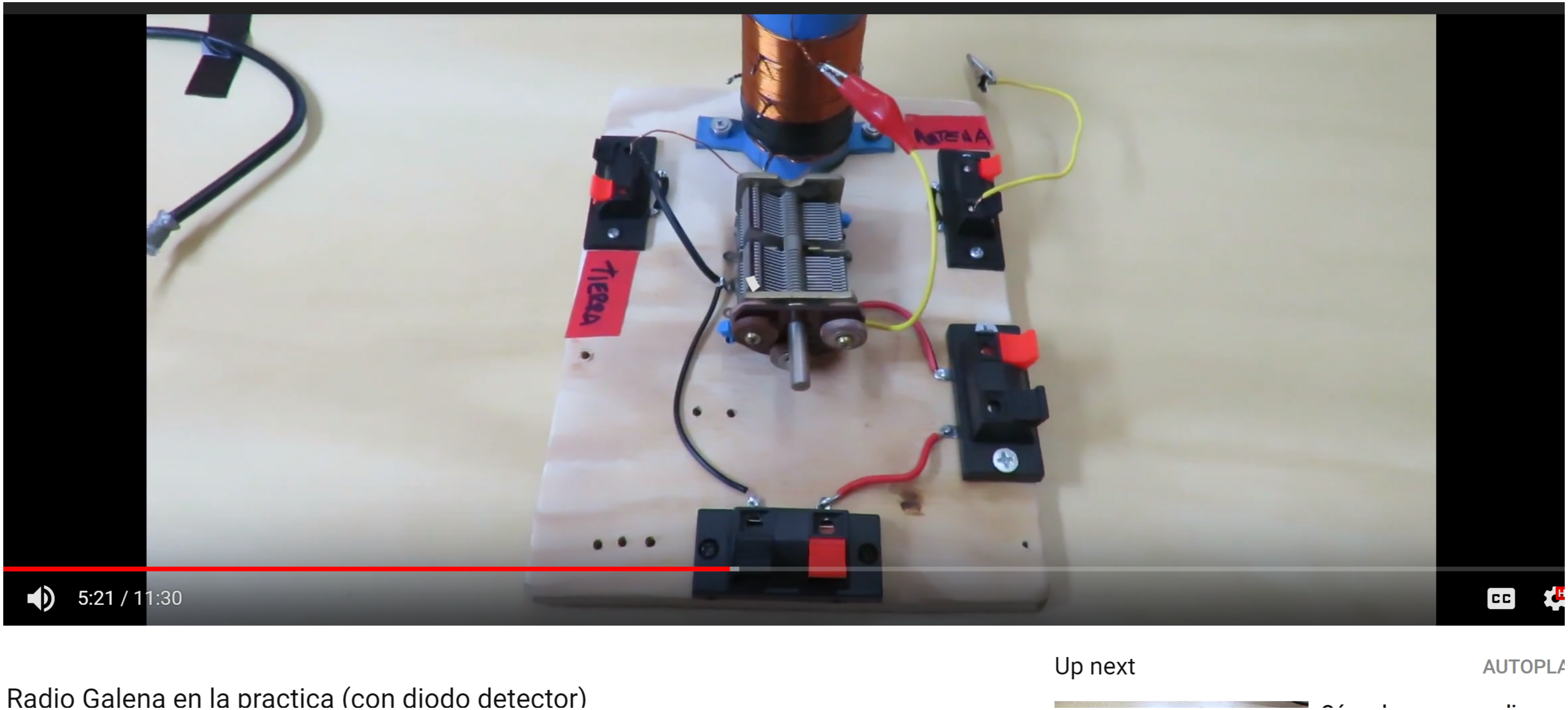
ANTENA
TIERRA
5:21 / 11:30
CC
Up next
AUTOPLA

Radio Galena en la practica (con diodo detector)

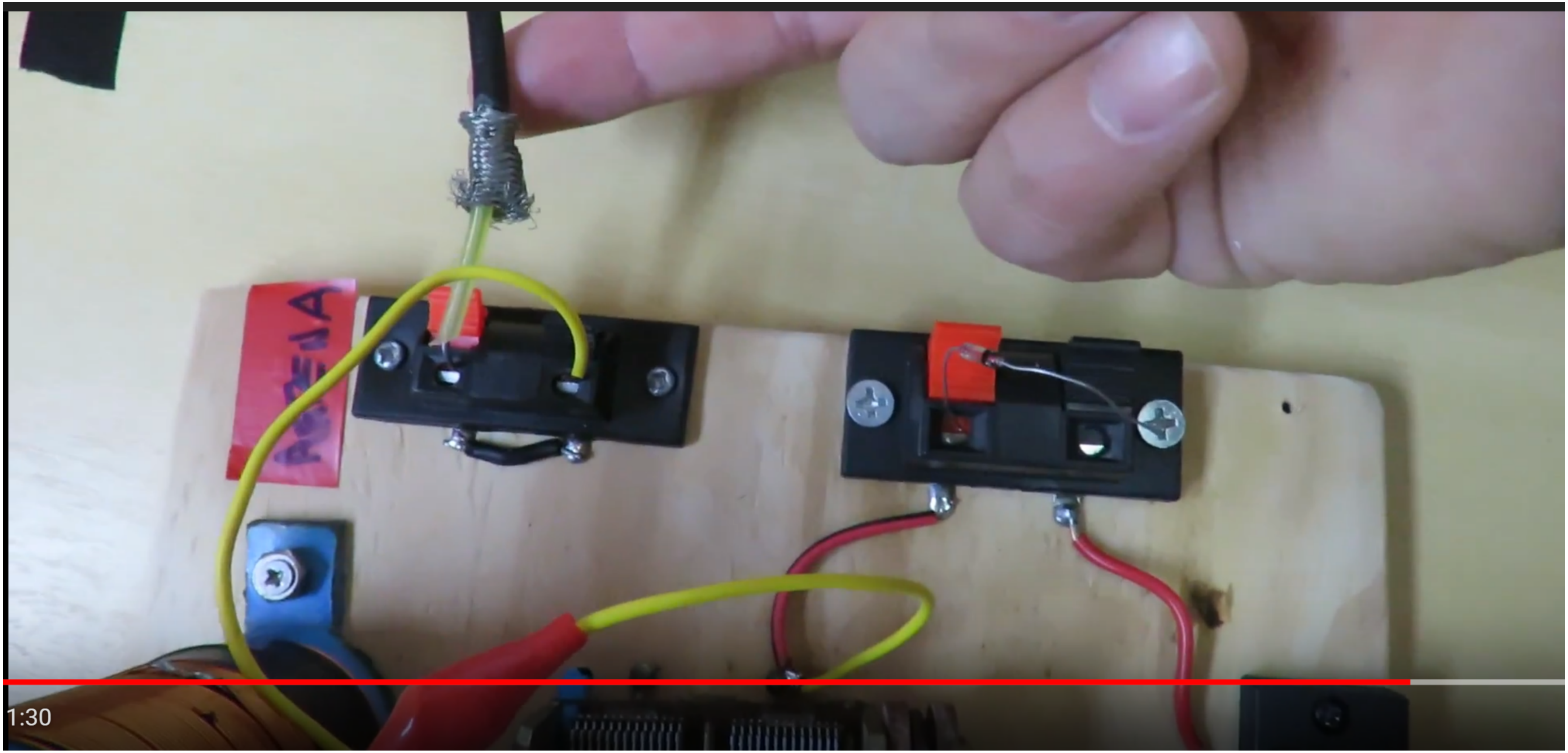
1:30

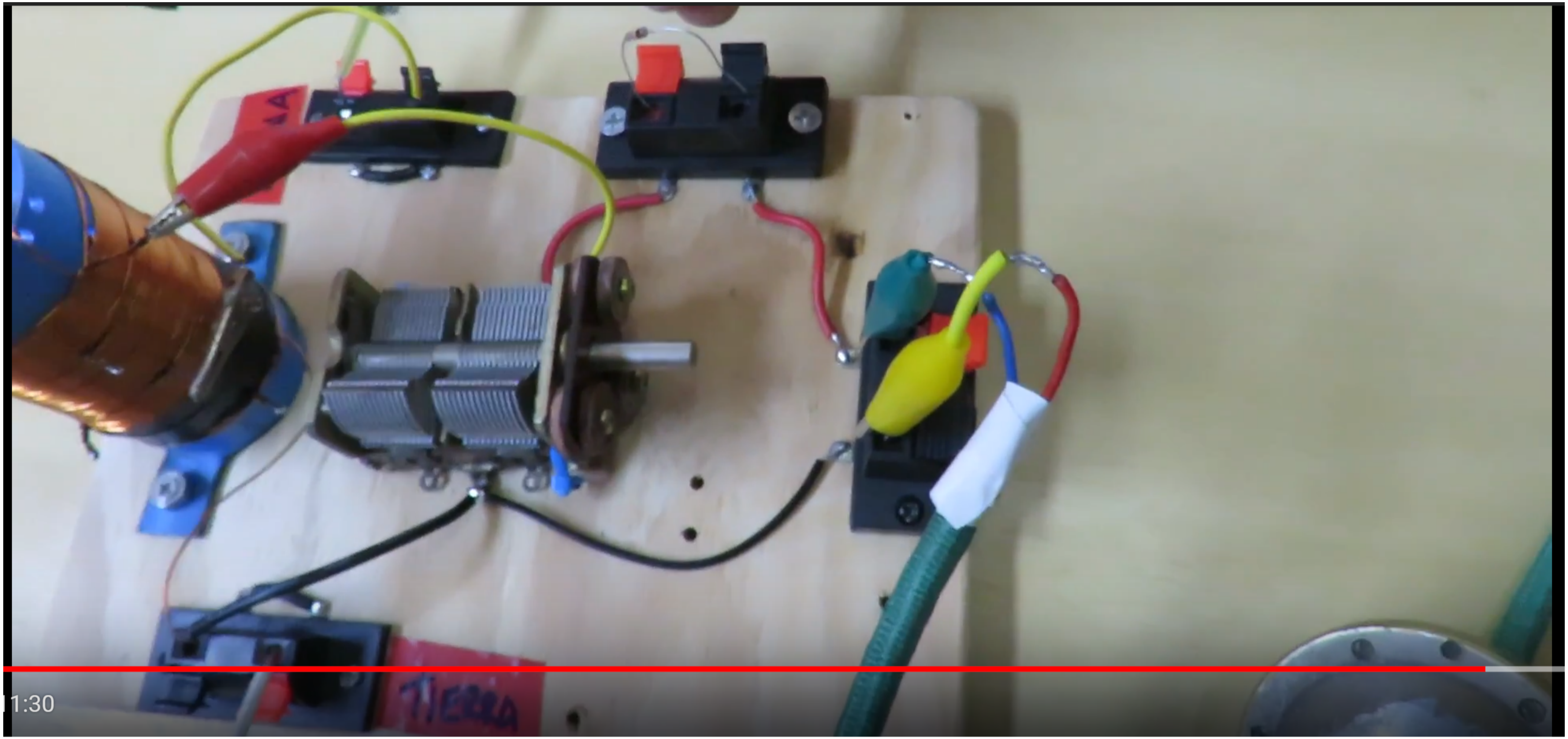
NA
TIERRA
1:30

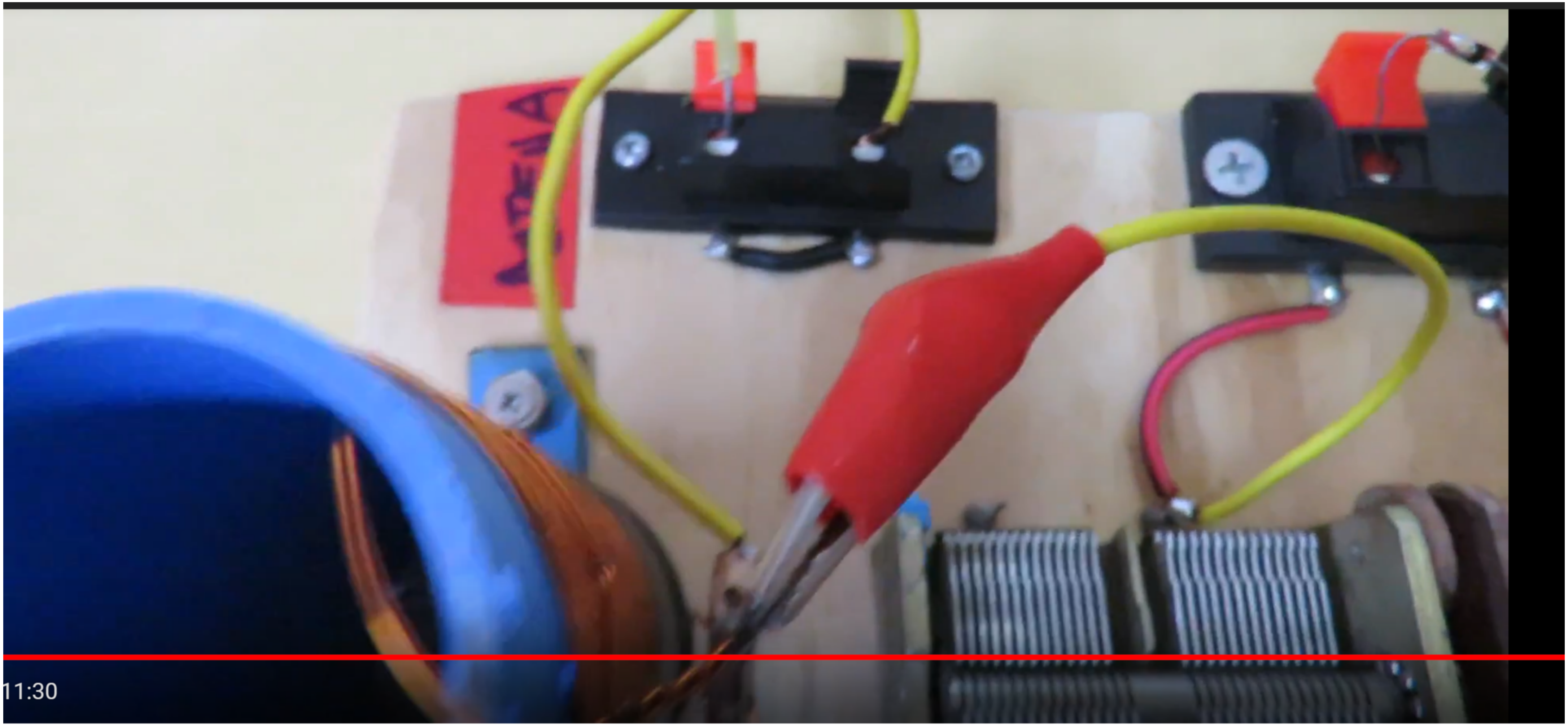
11:30

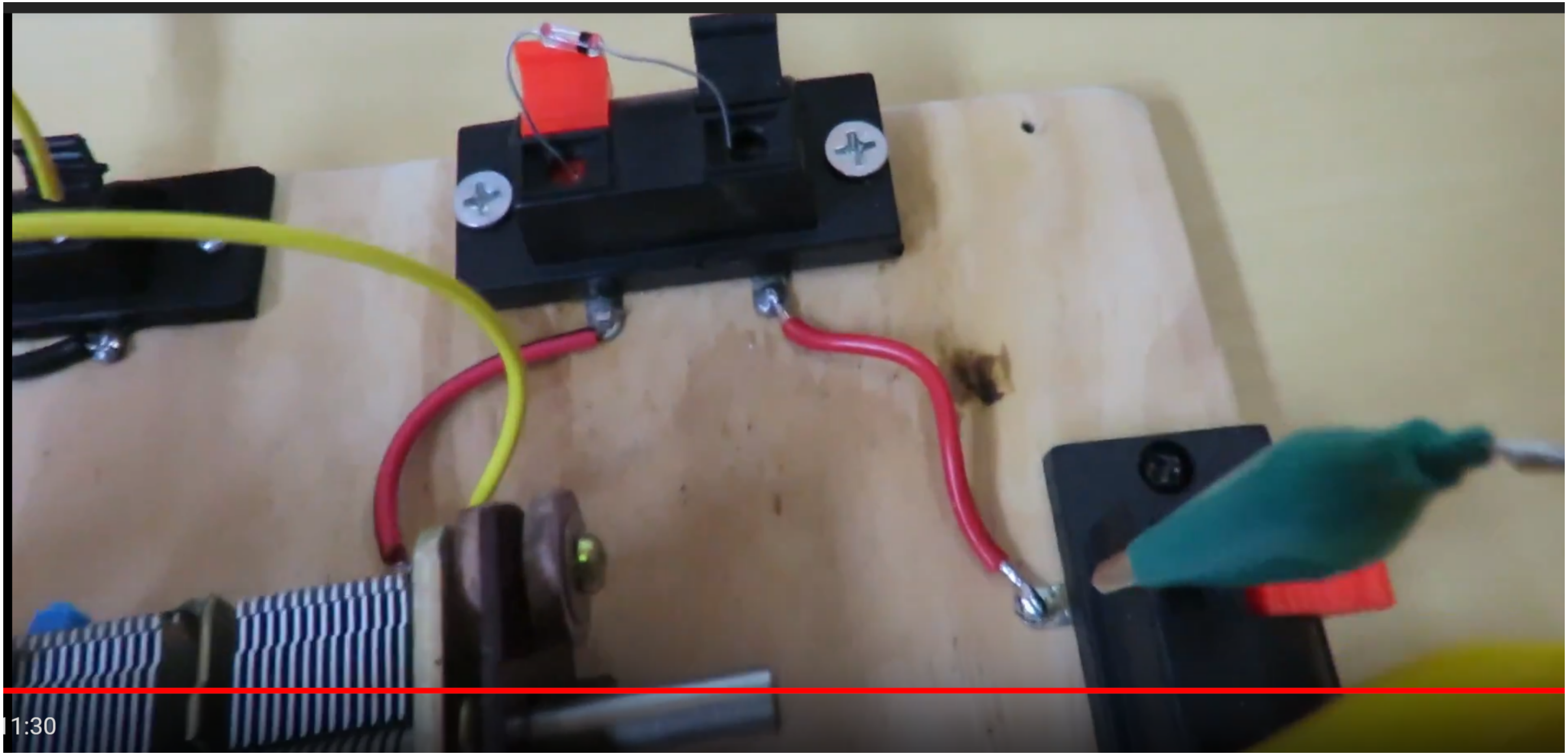
11:30

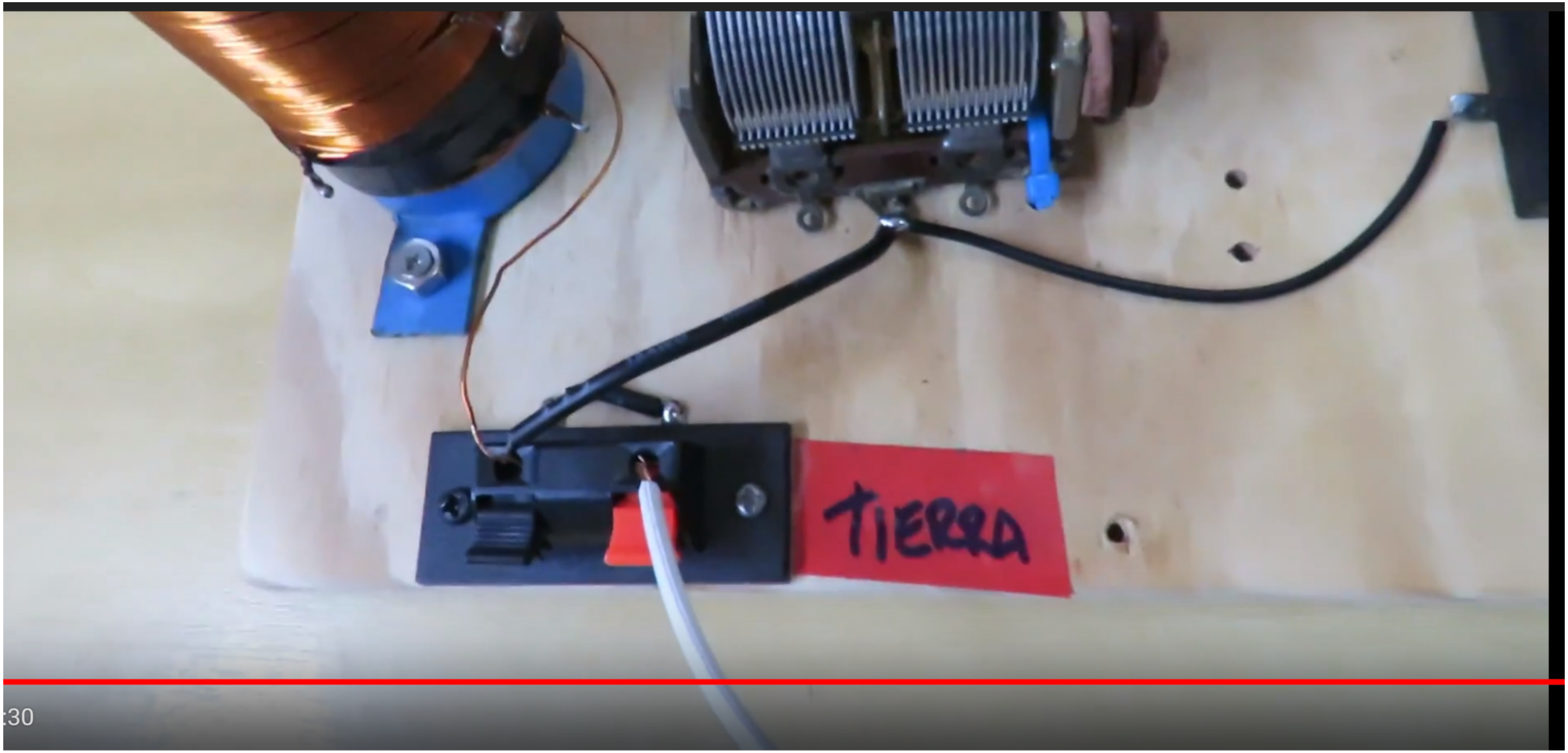
TIERRA
:30

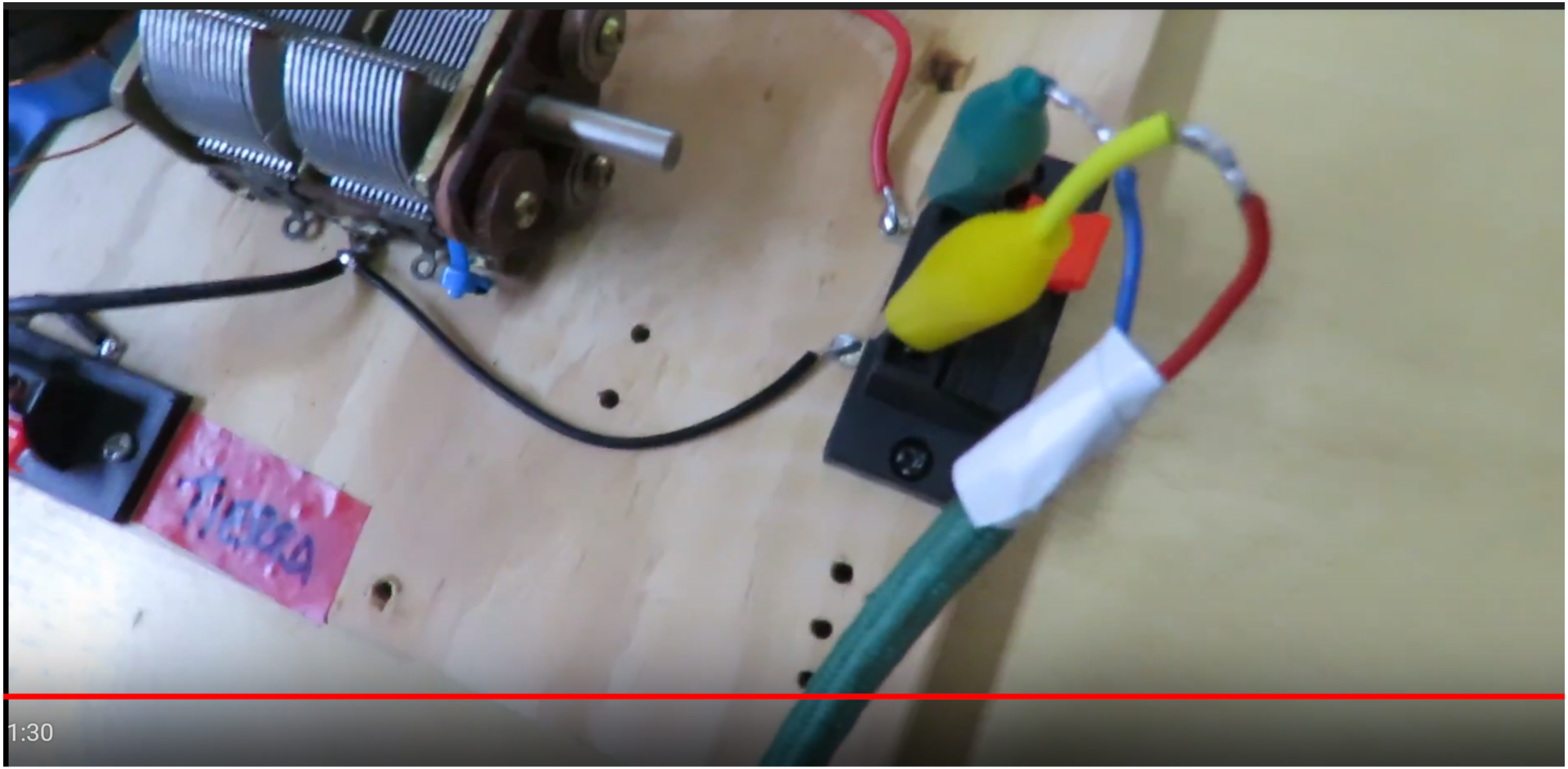
TIERRA
1:30

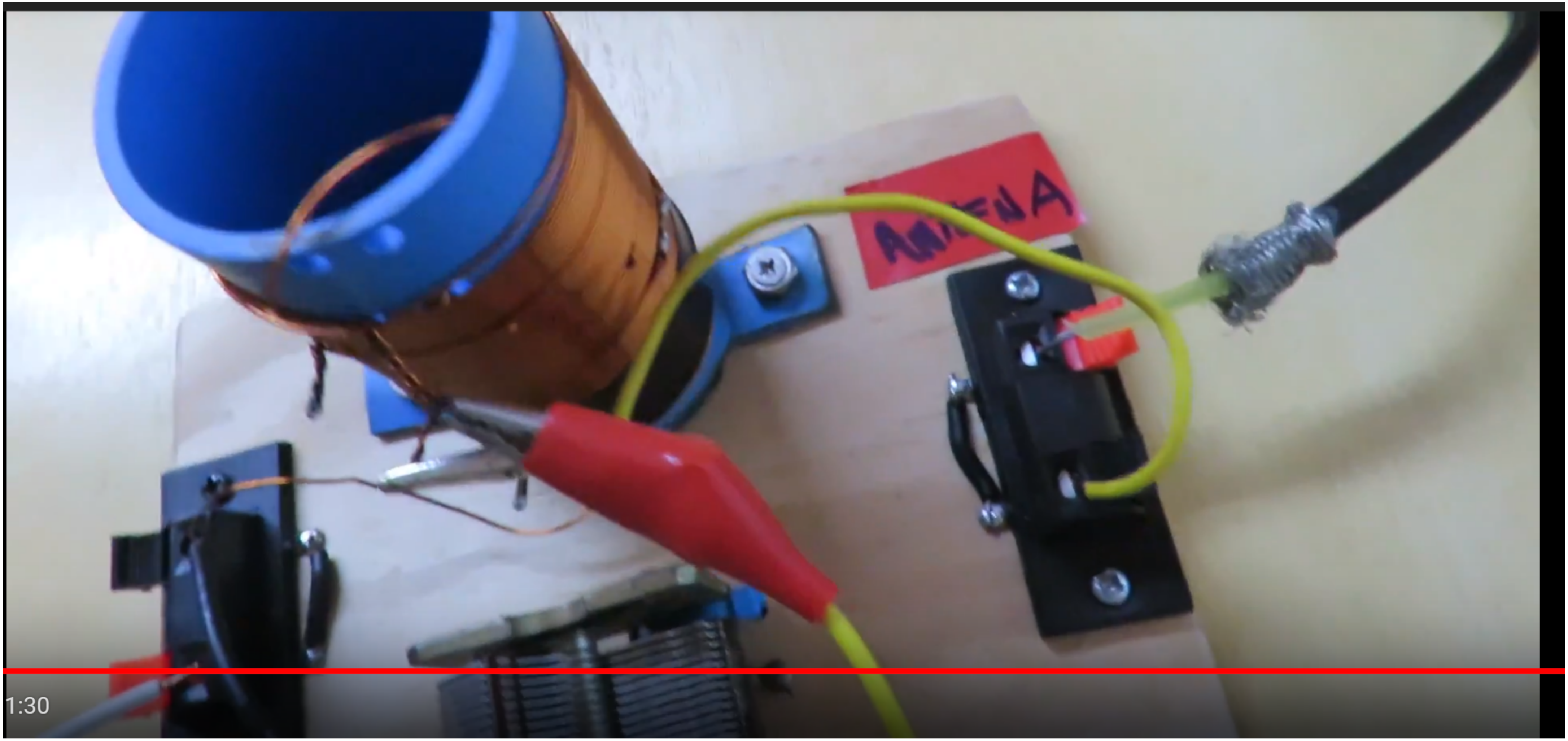
ANTENA
1:30

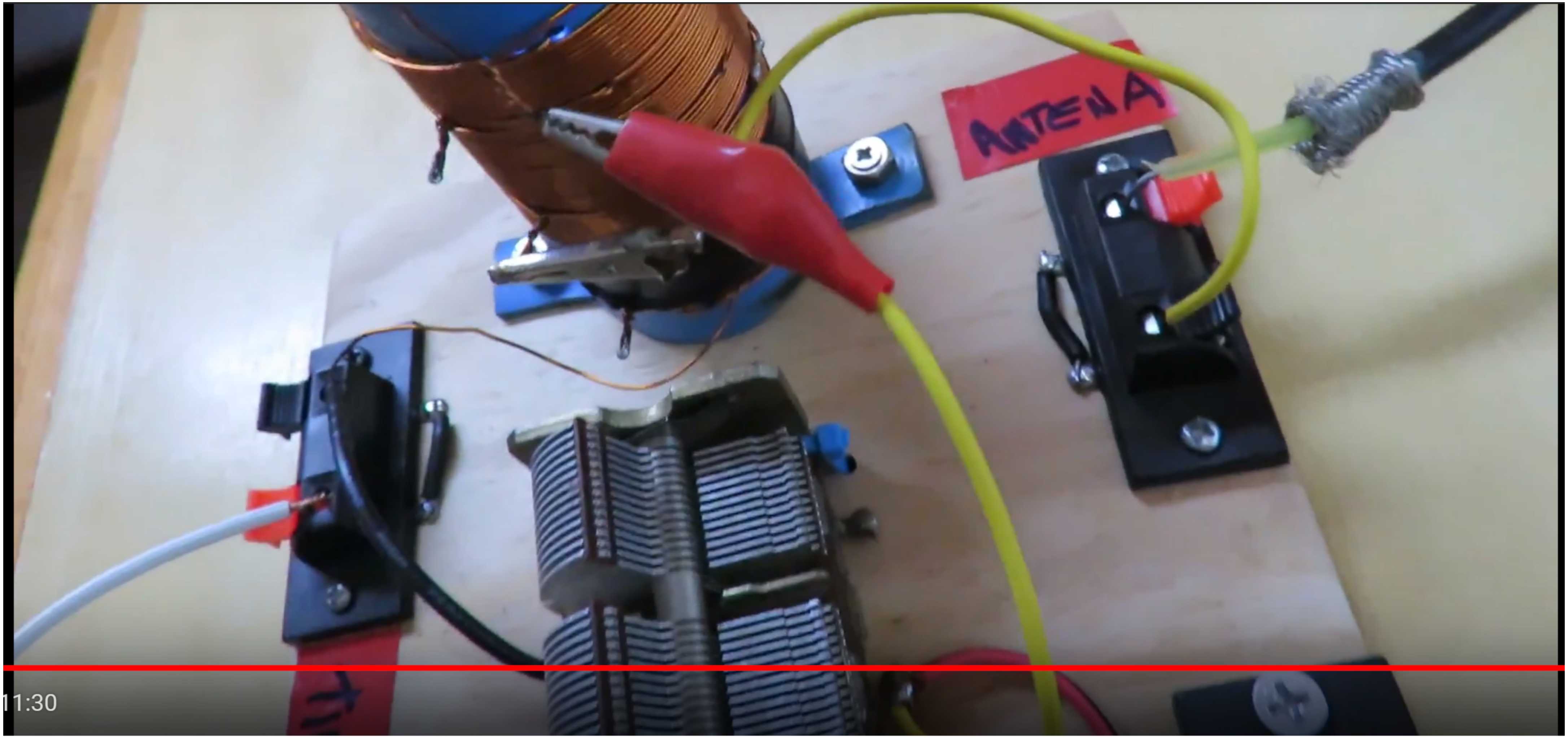
ANTENA
11:30

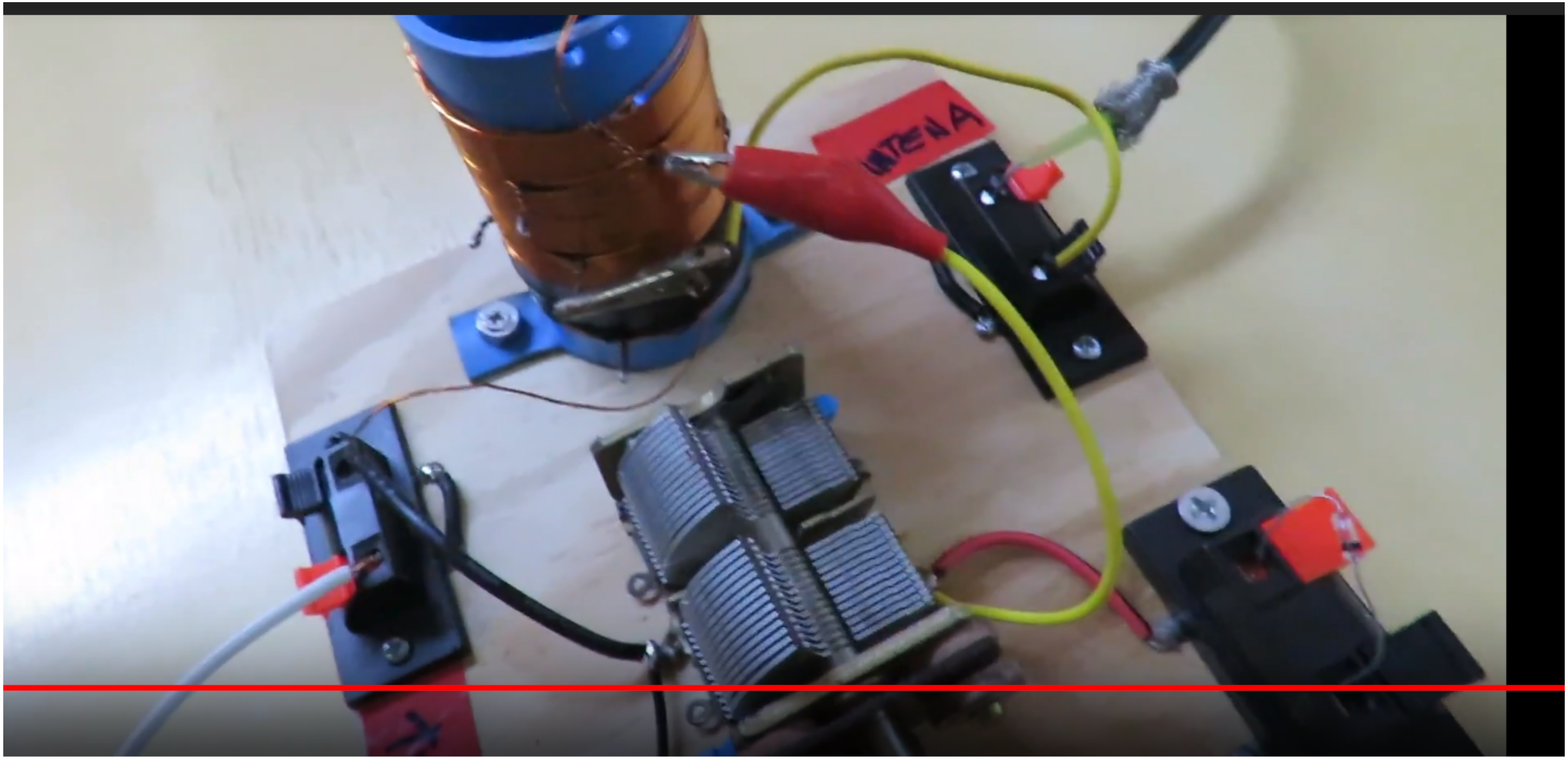
NTENA

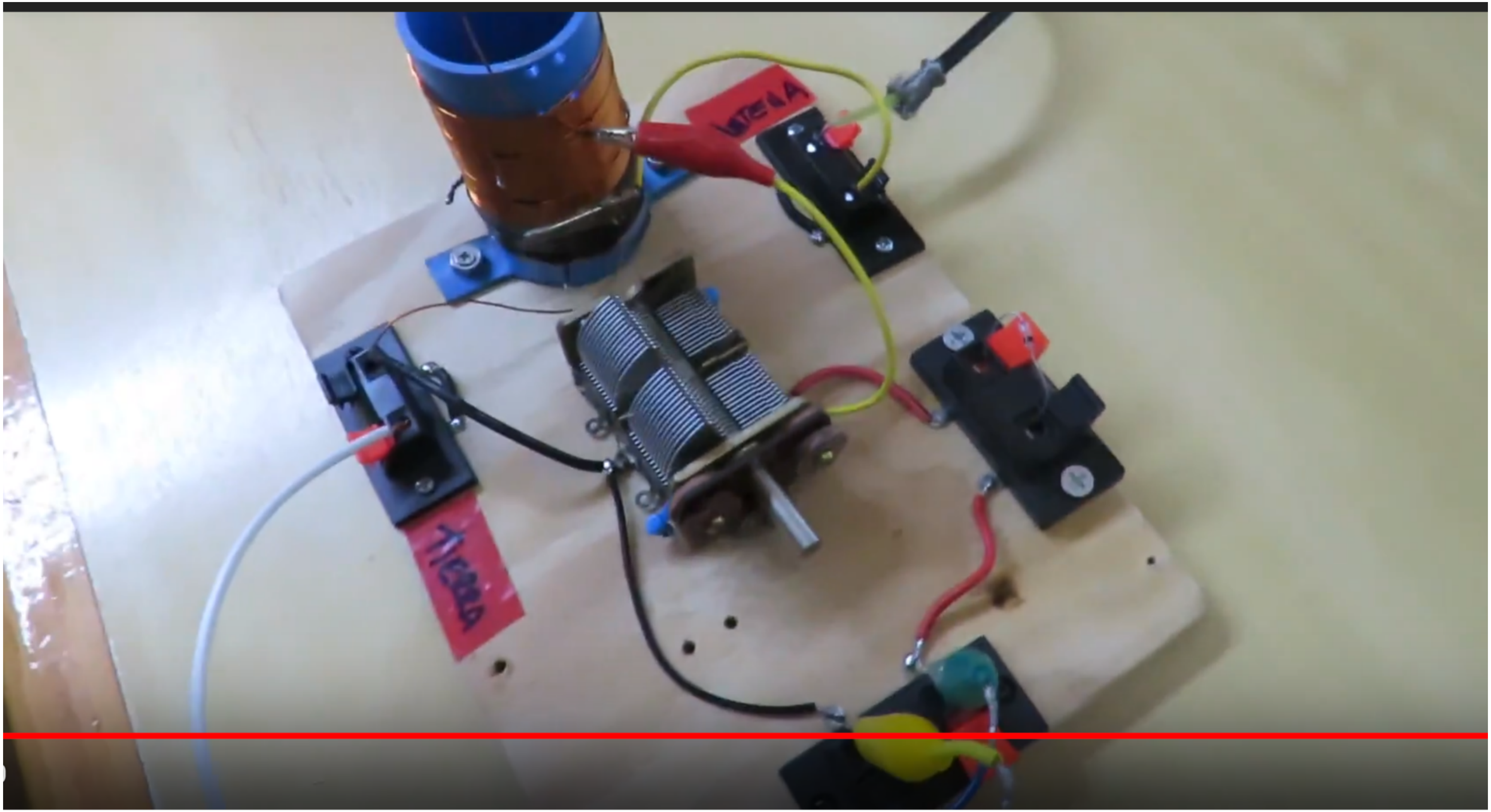
ANTENA
TIERRA

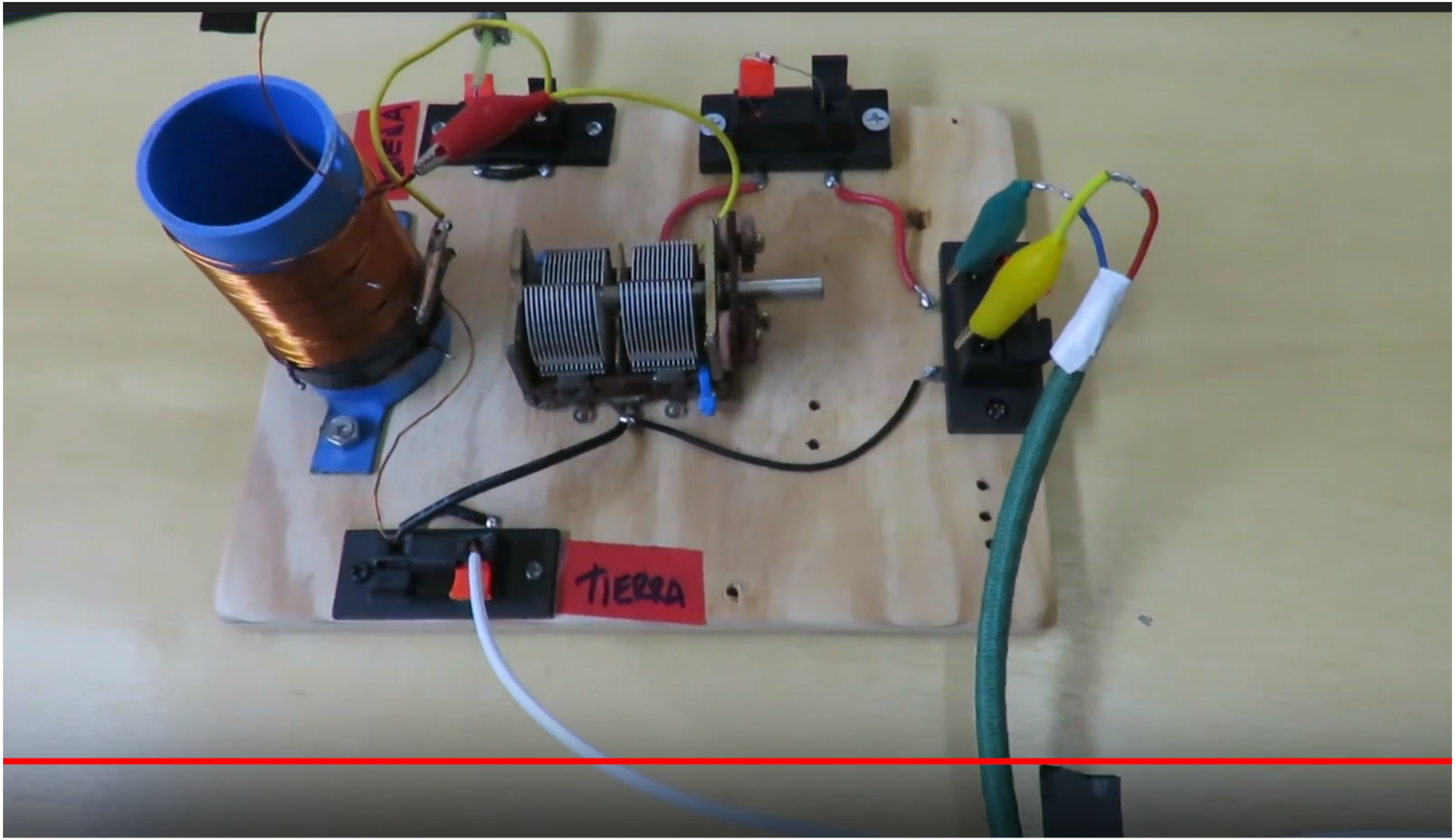
TIERRA

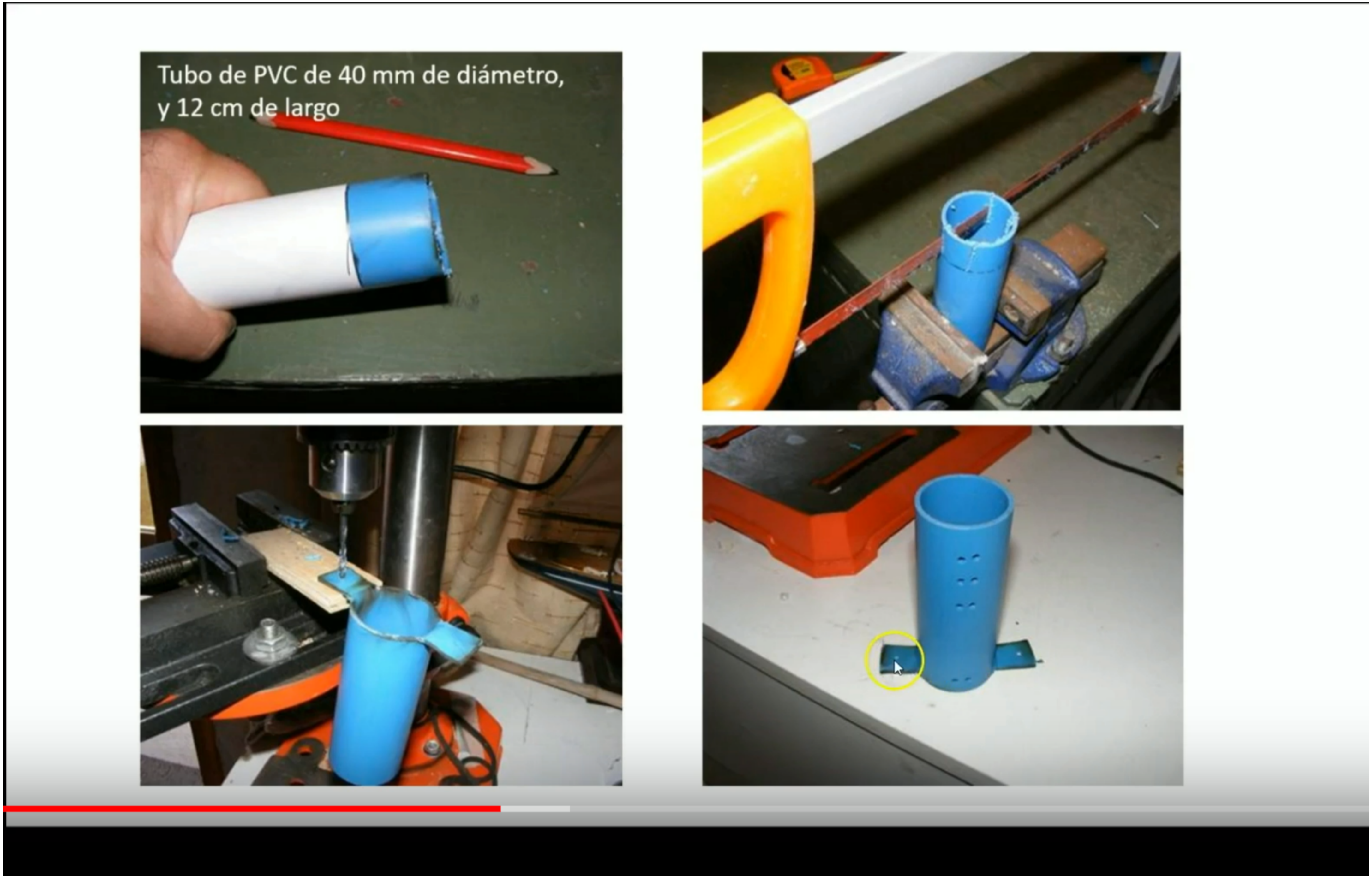

Tubo de PVC de 40 mm de diámetro, y 12 cm de largo

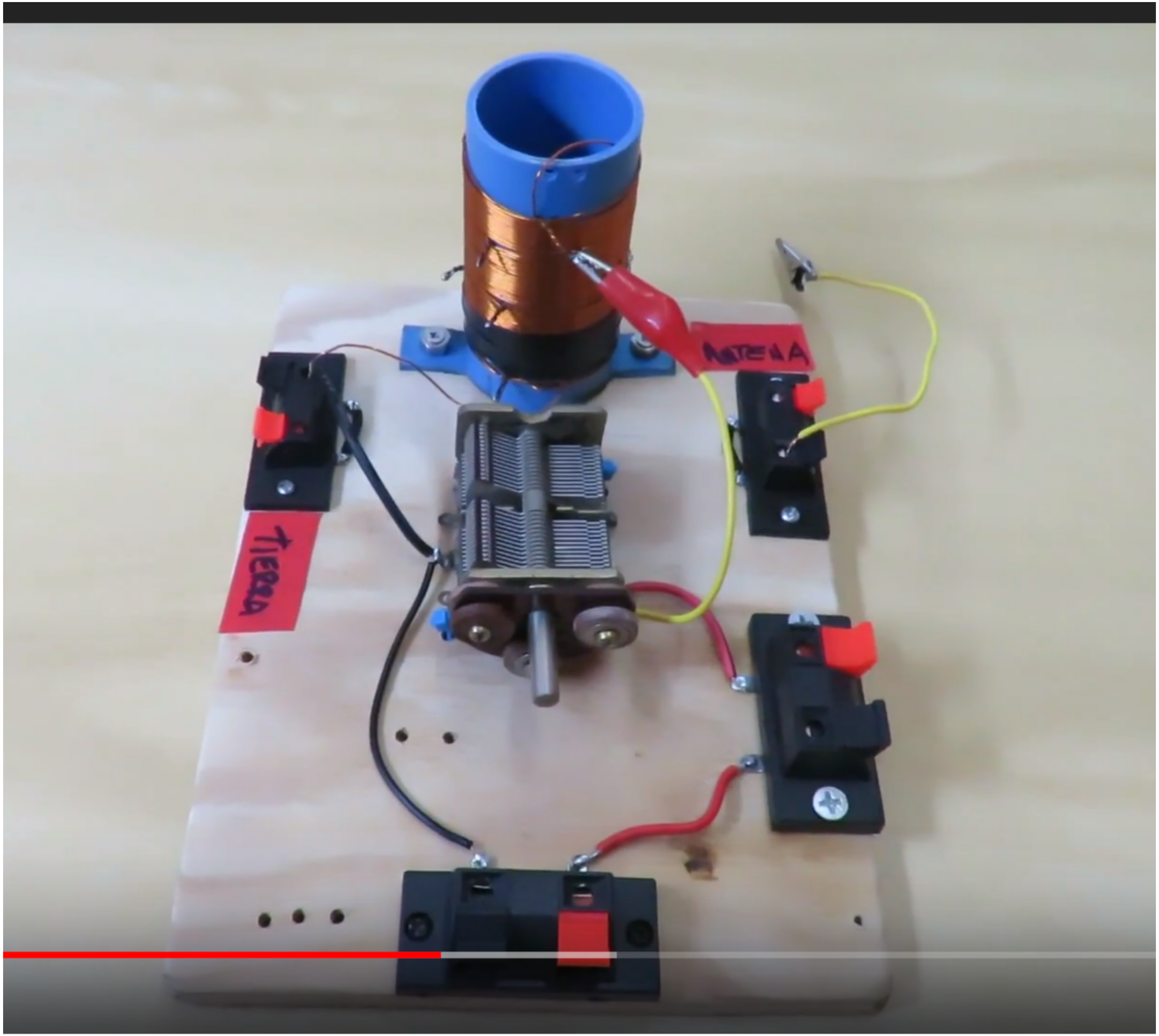
ANTENA
TIERRA

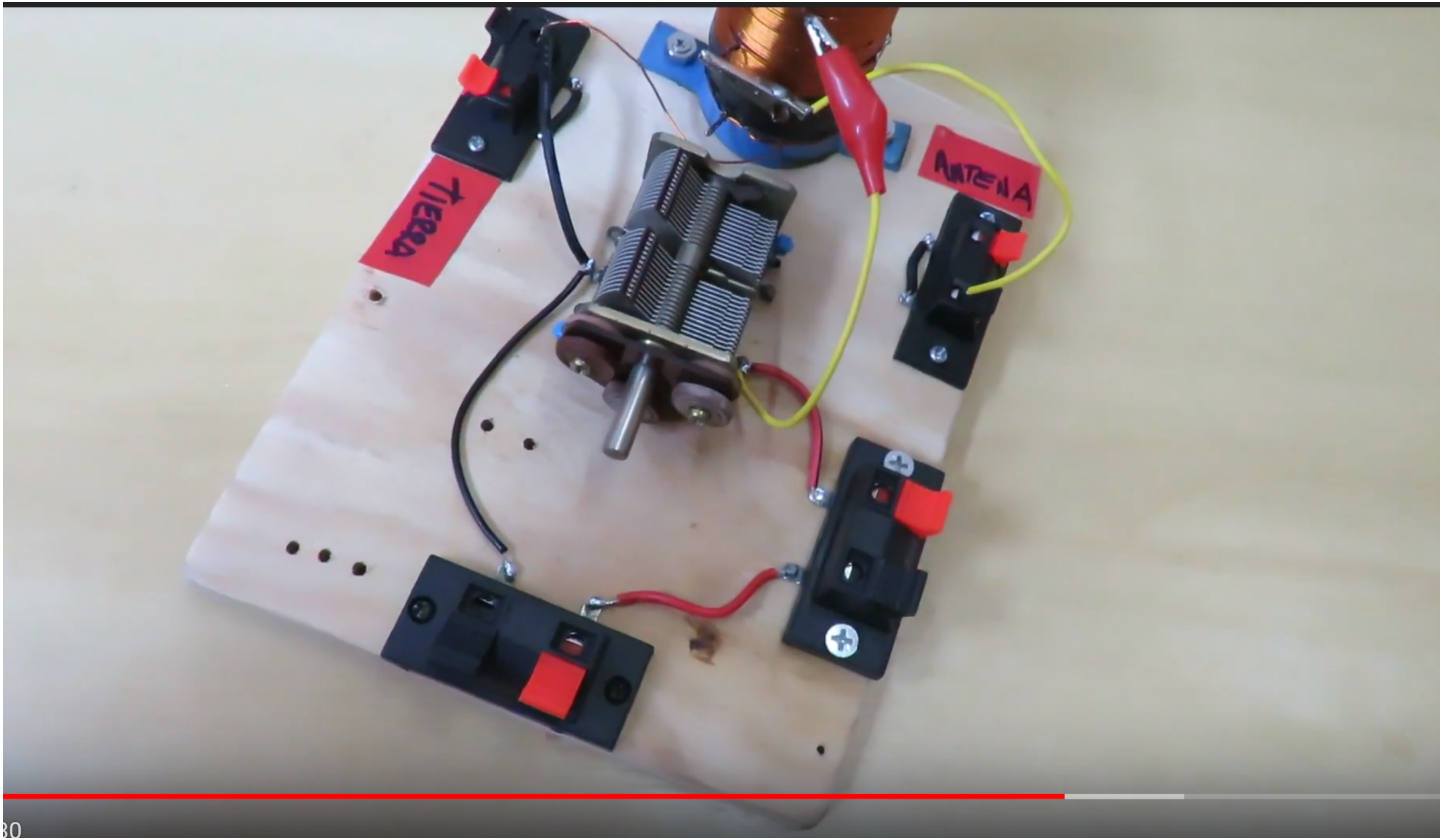
TIERRA
ANTENA

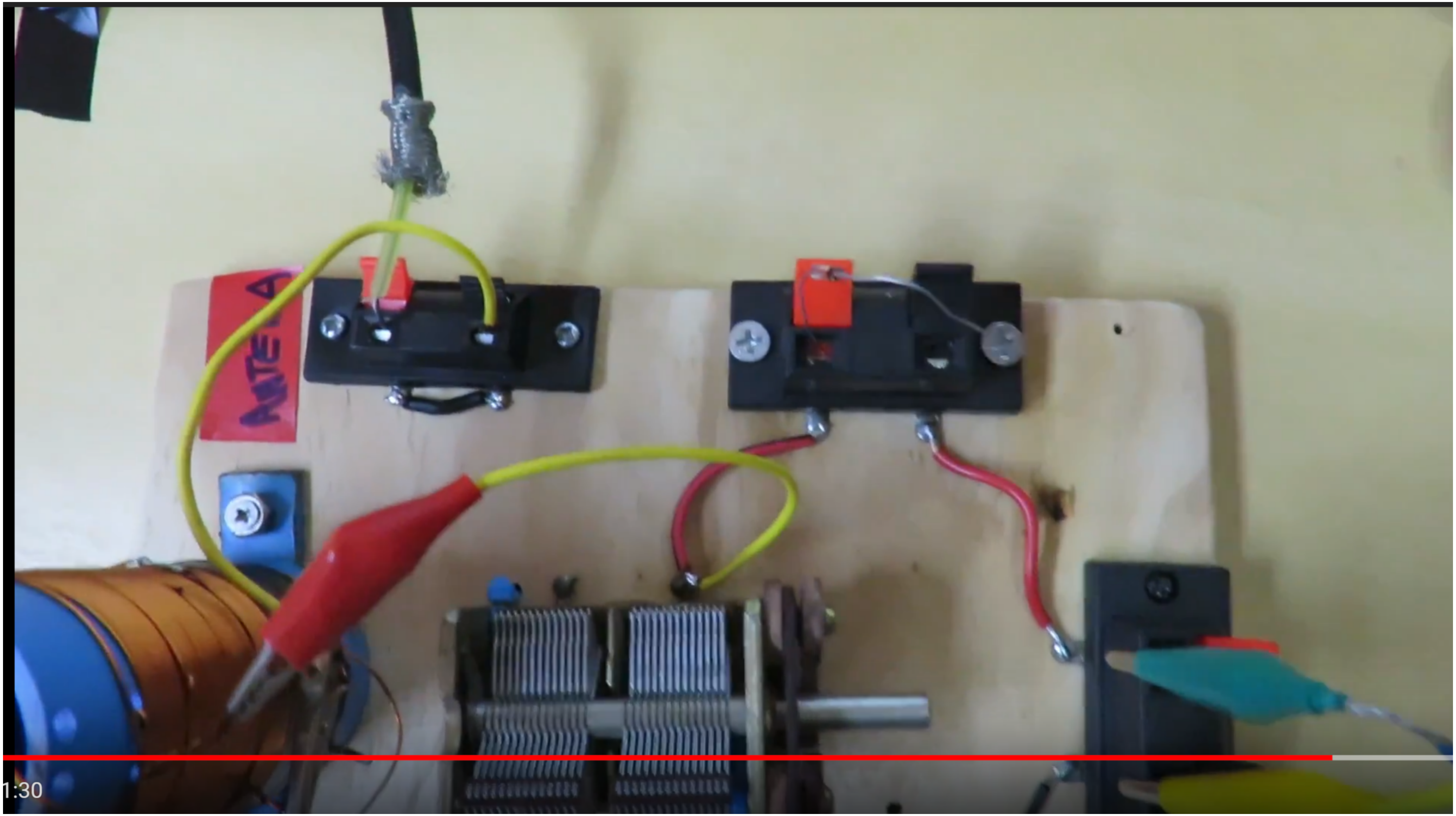
ANTENA
1:30

0:48 / 2:05
Up next

Rádio de Galena com Antena de Quadro - UTFPR - TD

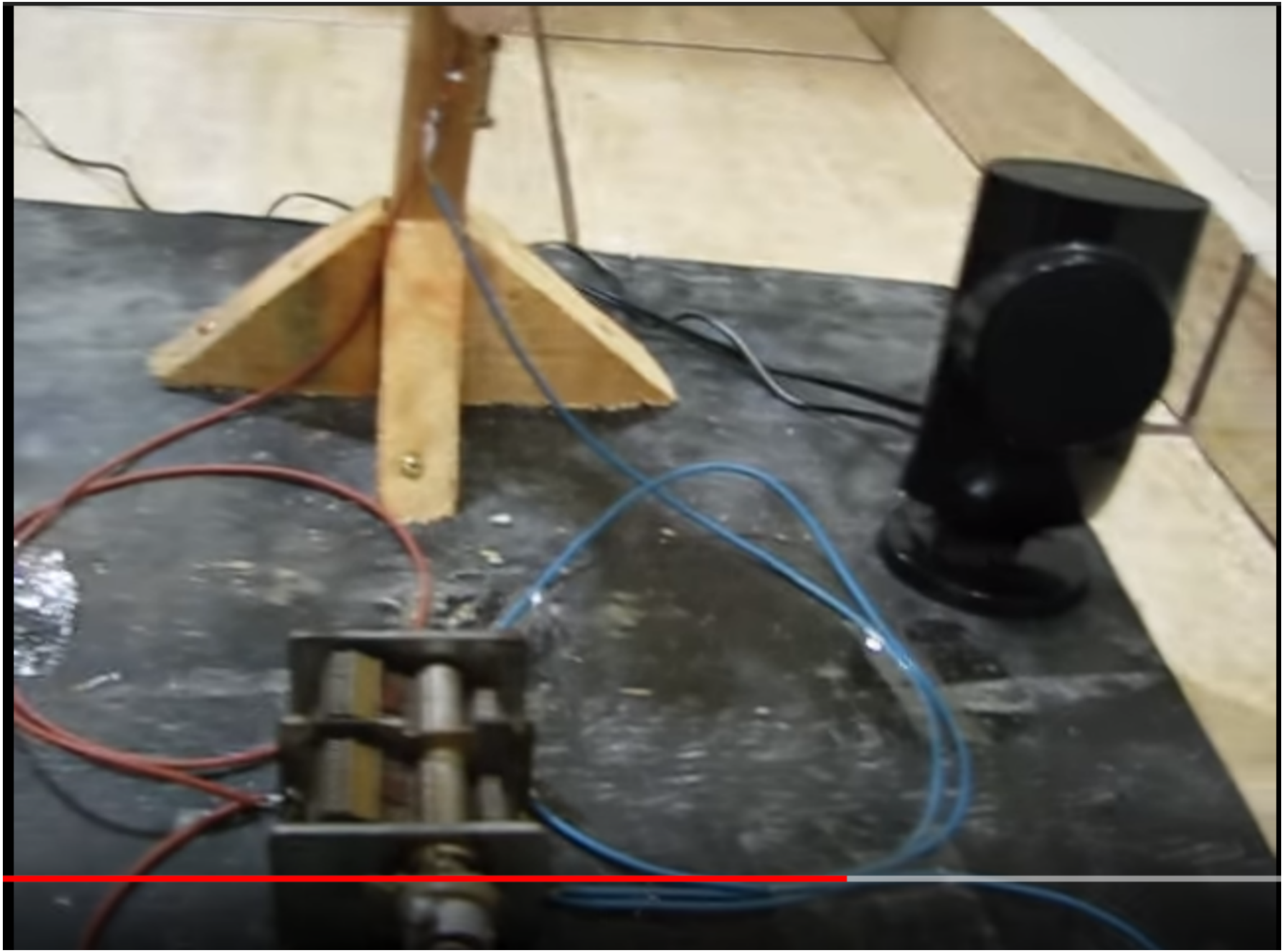

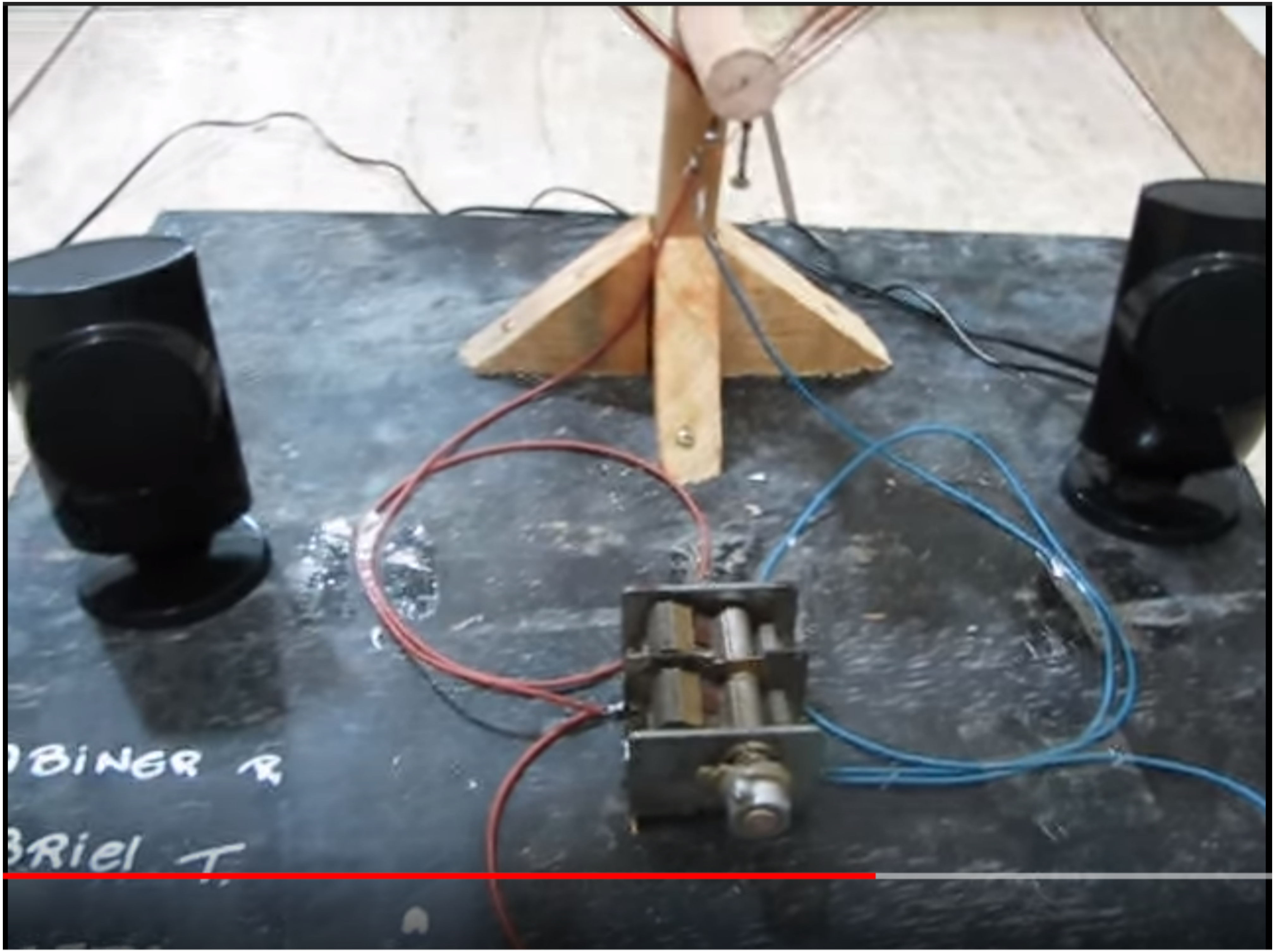
BINGER R.
BRIEL T.

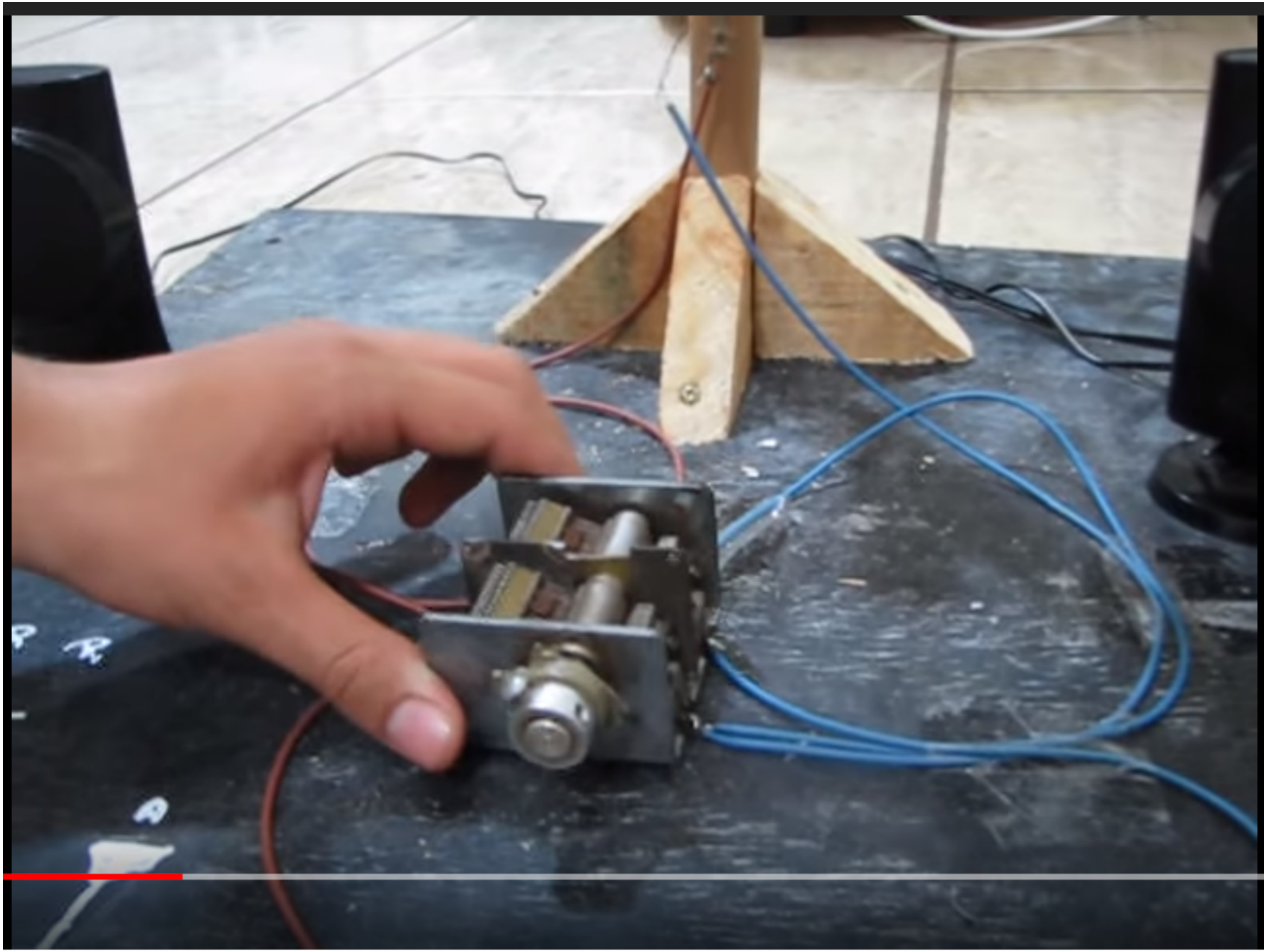

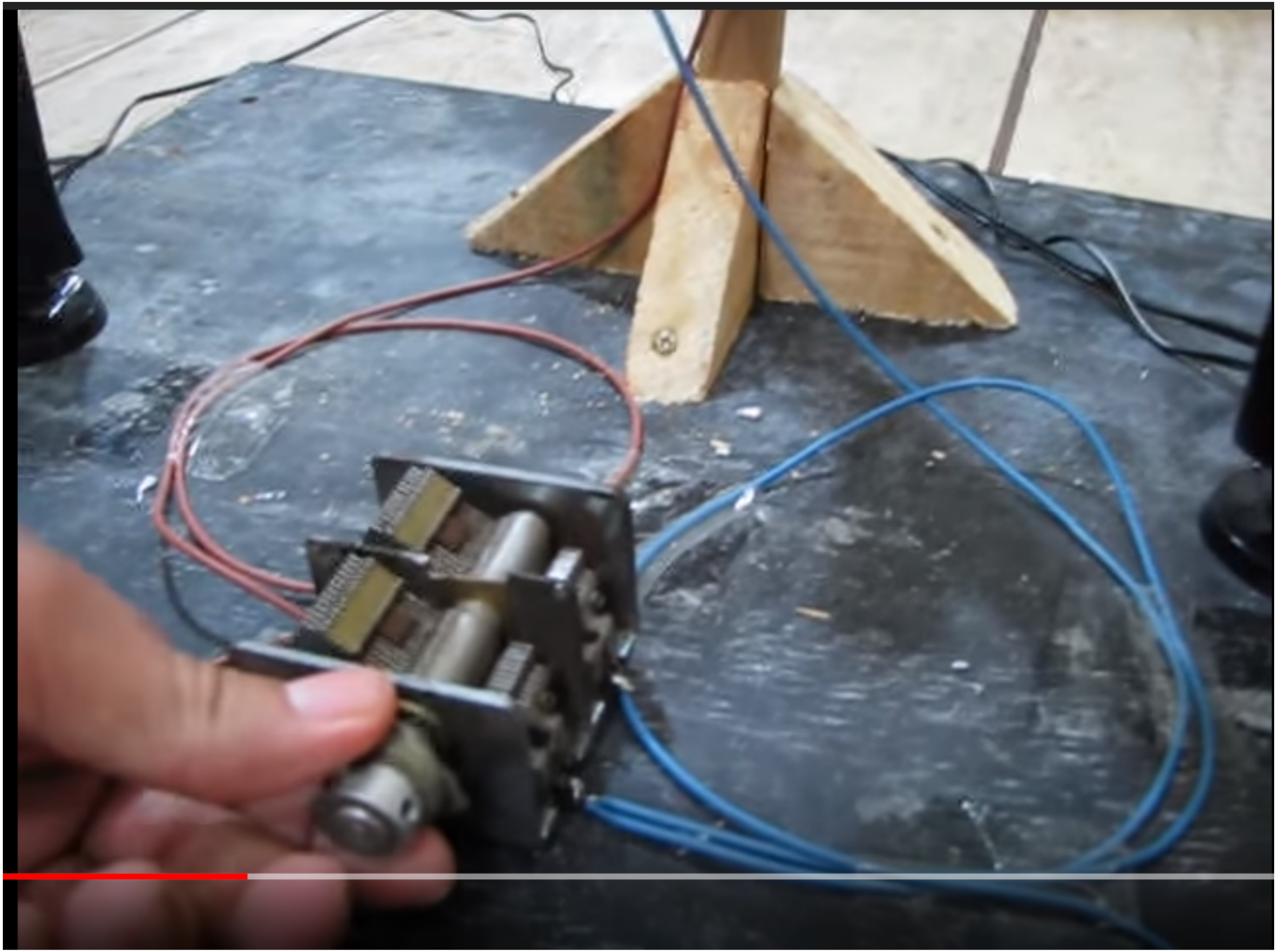

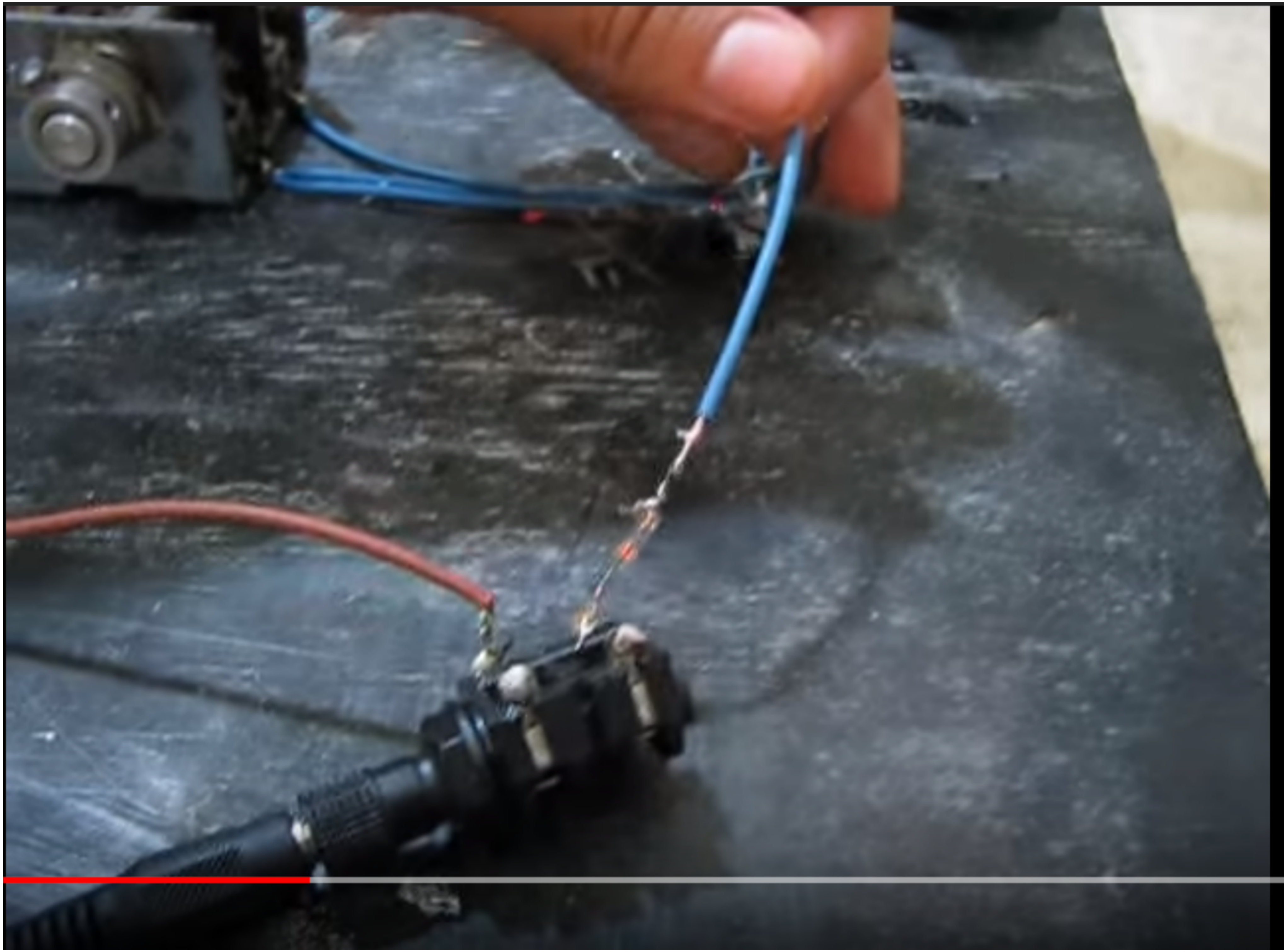

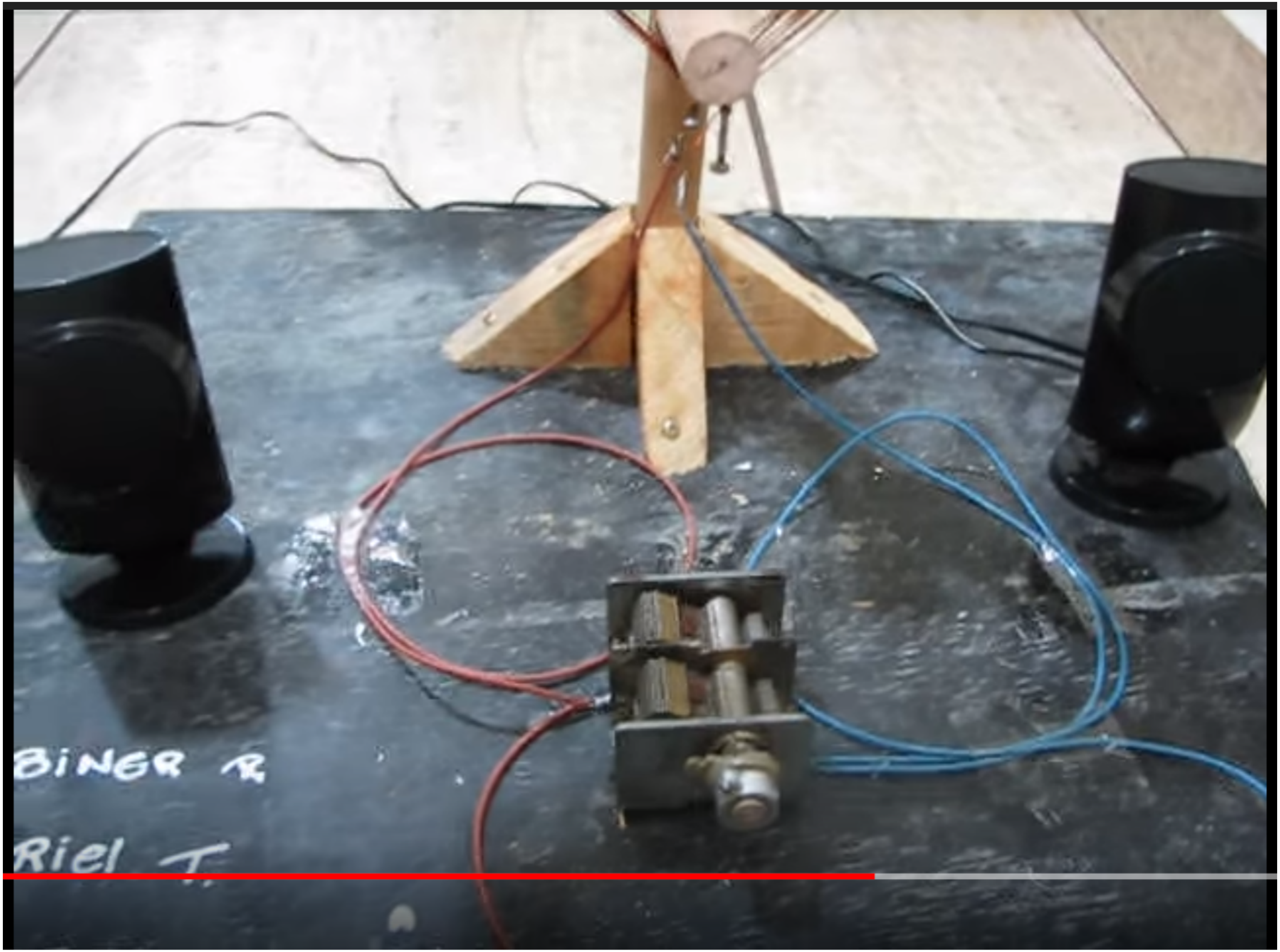
BiNGR R
Riel T.

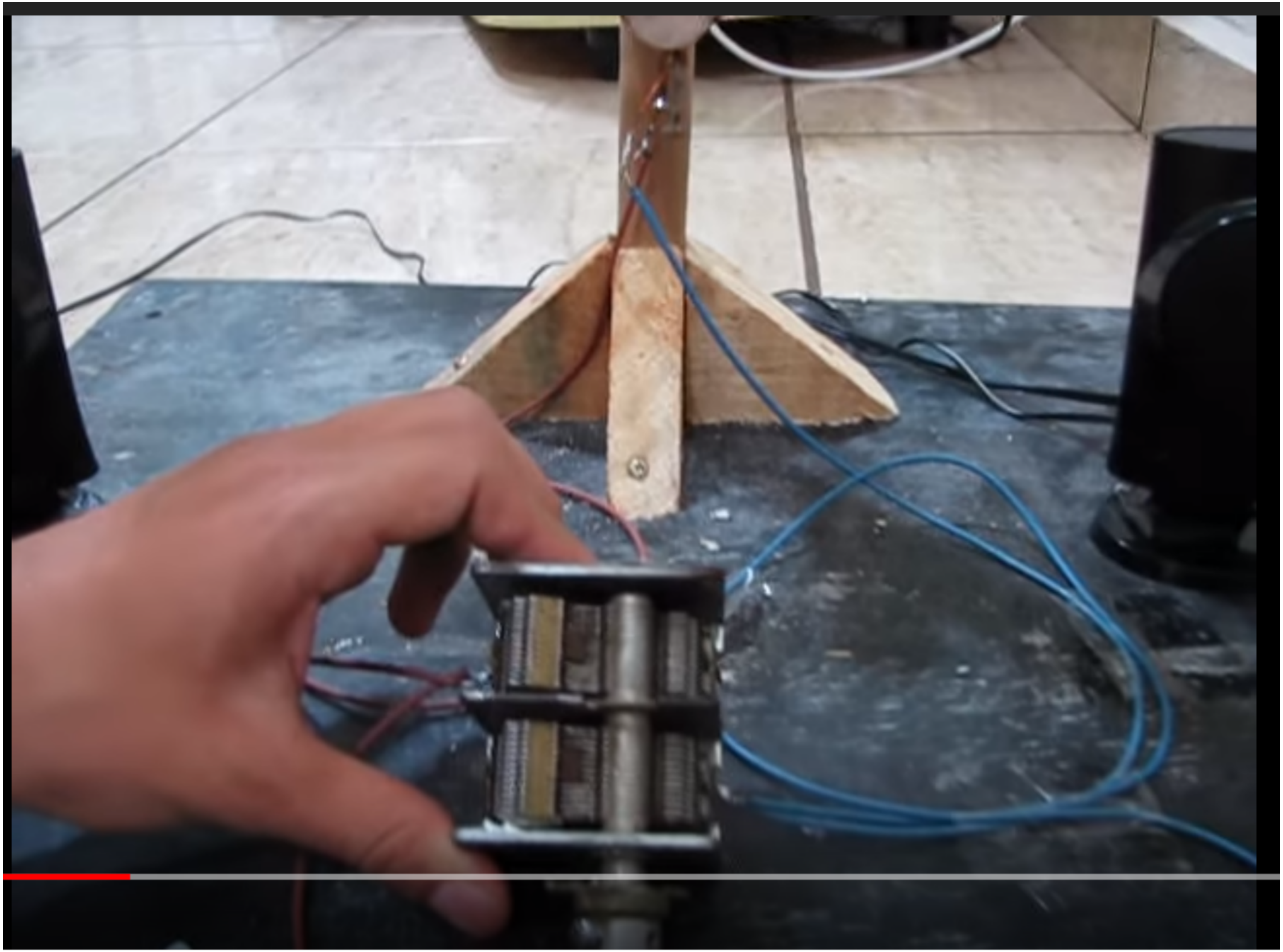

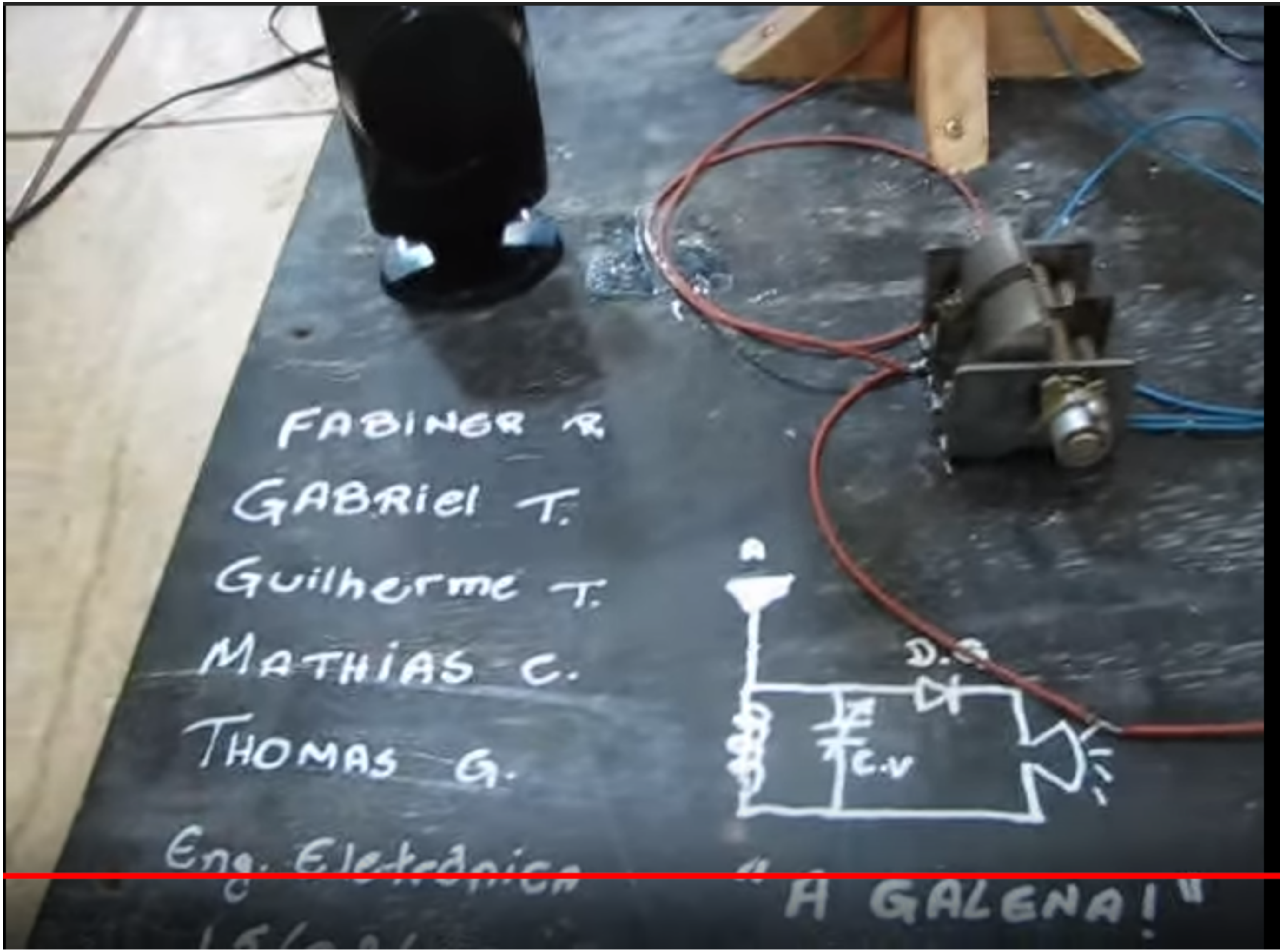
FABINGR R.
GABRiel T.
Guilherme T.
MATHiAS C.
THOMAS G.
Eng. Eletrônica
A
D.G
C.V
"A GALENA!"

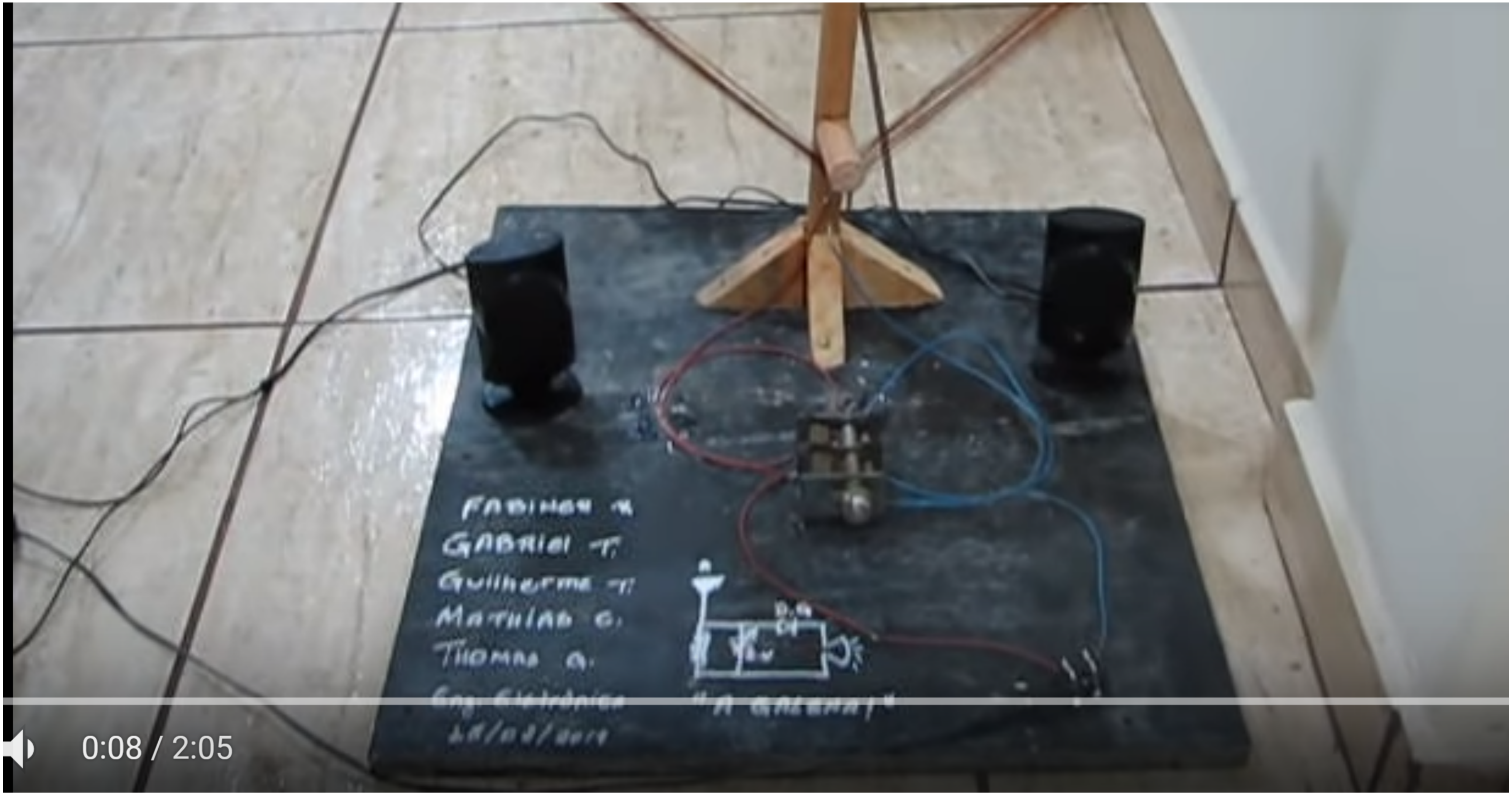
FABINHO R.
GABRIEL T.
Guilherme T.
MATHIAS G.
THOMAS Q.
Eng. Eletrônica
"A GALENA!"
0:08 / 2:05

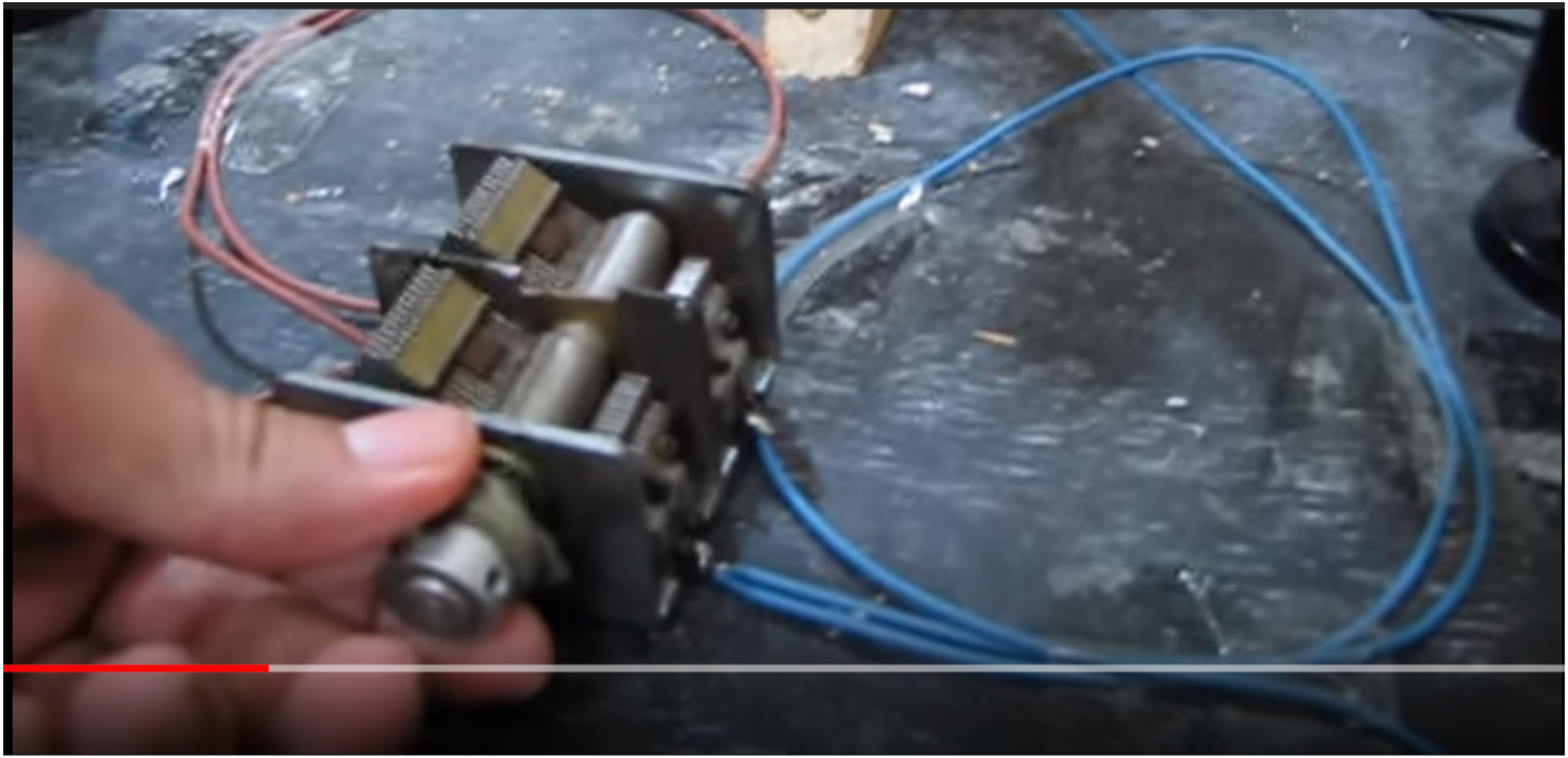

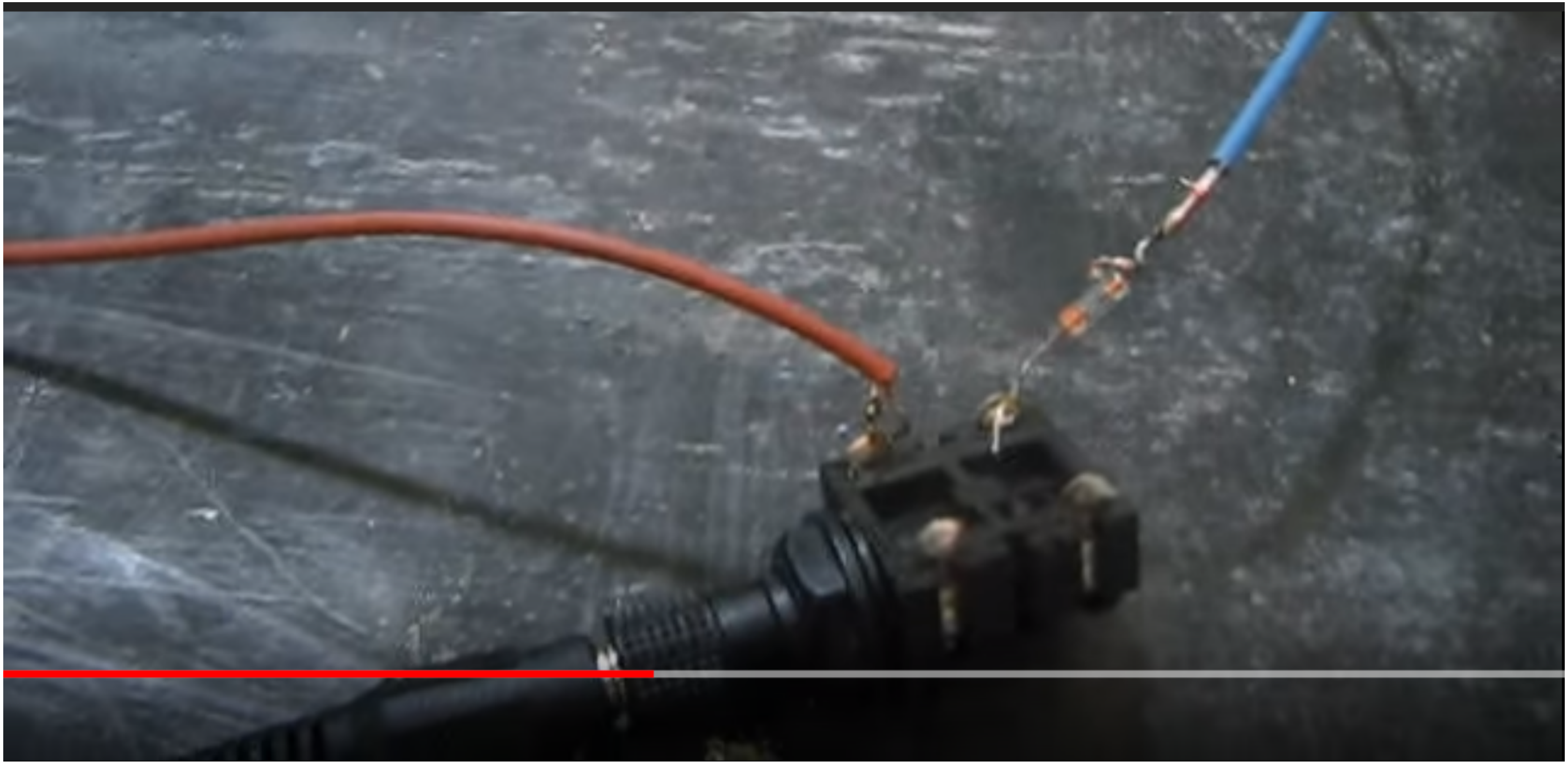

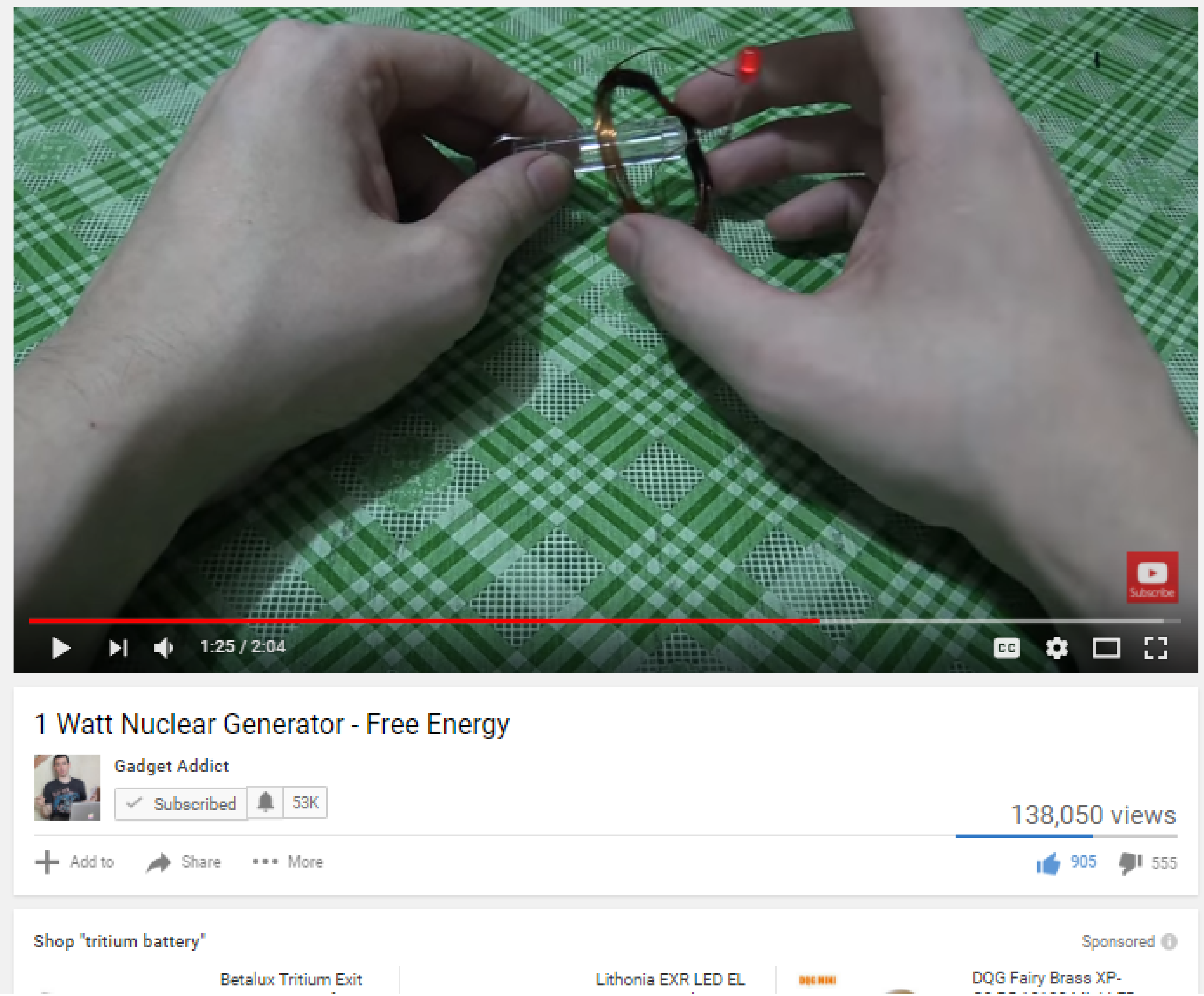
Subscribe
1:25 / 2:04
Subscribed
53K
Add to
Share
More
905
555
Sponsored

1 Watt Nuclear Generator - Free Energy

Gadget Addict

138,050 views

Shop "tritium battery"

Betalux Tritium Exit

Lithonia EXR LED EL

DQG Fairy Brass XP-

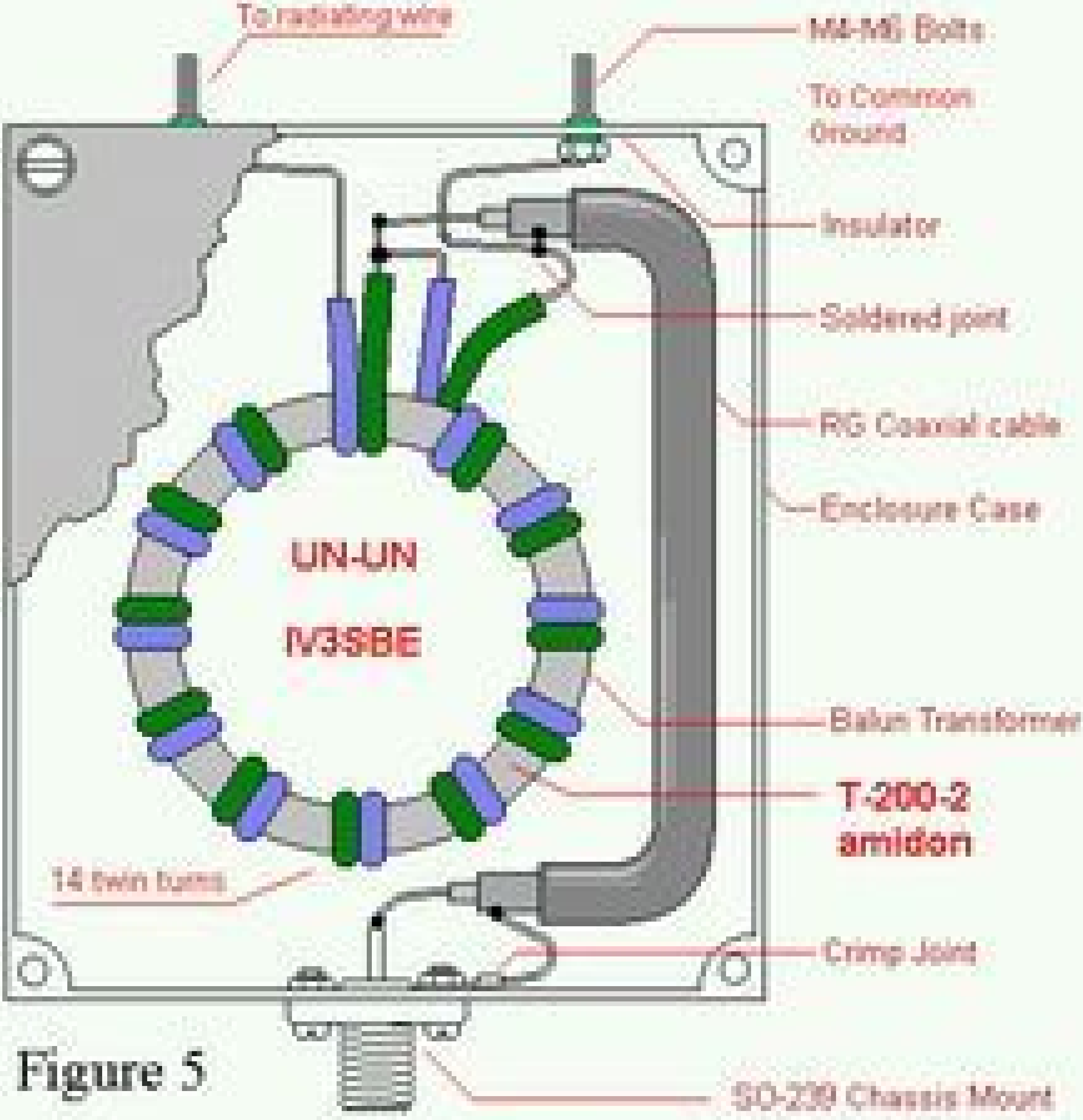
To radiating wire
M4-M6 Bolts
To Common Ground
Insulator
Soldered joint
RG Coaxial cable
Enclosure Case
UN-UN
IV3SBE
Balun Transformer
T-200-2 amidon
14 twin turns
Crimp Joint
SO-239 Chassis Mount

Figure 5

Voltage Balun / "Magnetic Balun"

From GW0VMR's plan

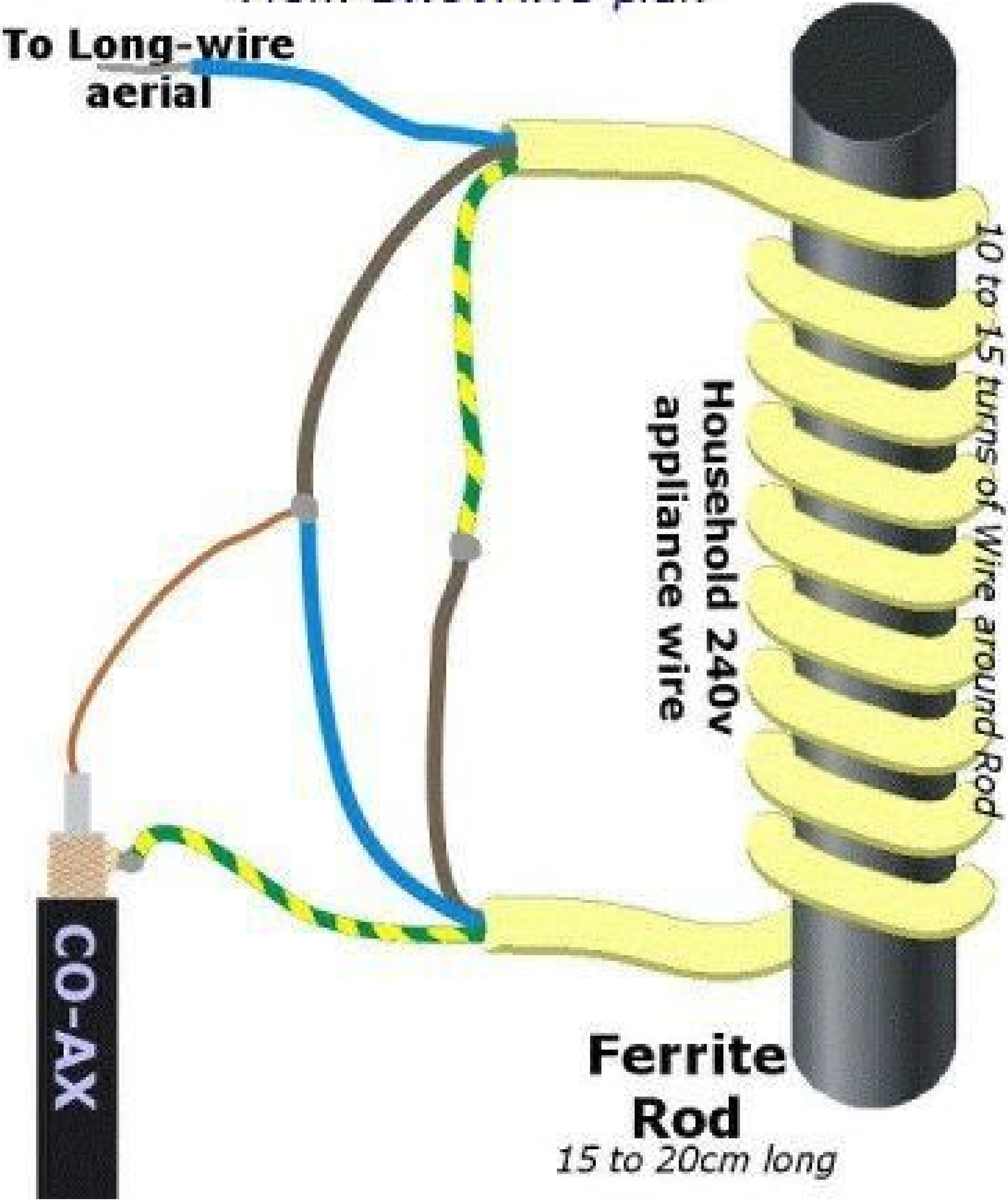
To Long-wire aerial
10 to 15 turns of Wire around Rod
Household 240v appliance wire
CO-AX
Ferrite Rod
15 to 20cm long

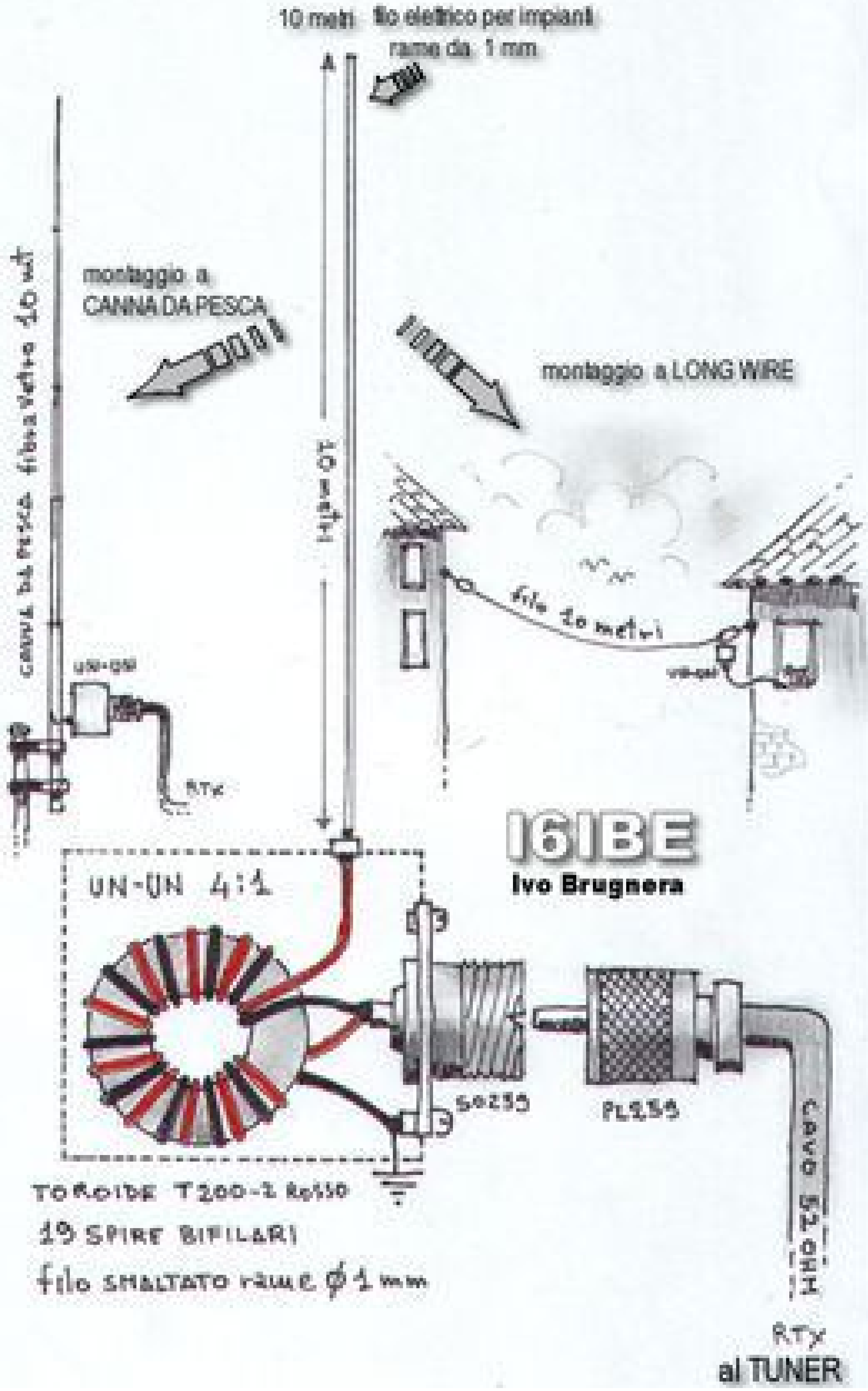
10 metri filo elettrico per impianti rame da 1 mm
montaggio a CANNA DA PESCA
montaggio a LONG WIRE
CANNA DA PESCA fibra vetro 10 mt
10 metri
UN-UN
RTX
filo 10 metri
UN-UN
I6IBE
Ivo Brugnera
UN-UN 4:1
SO239
PL259
CAVO 52 OHM
RTX
al TUNER
TOROIDE T200-2 ROSSO
19 SPIRE BIFILARI
filo SMALTATO rame Ø 1 mm

BALUN 9:1 Antenna Long Wire

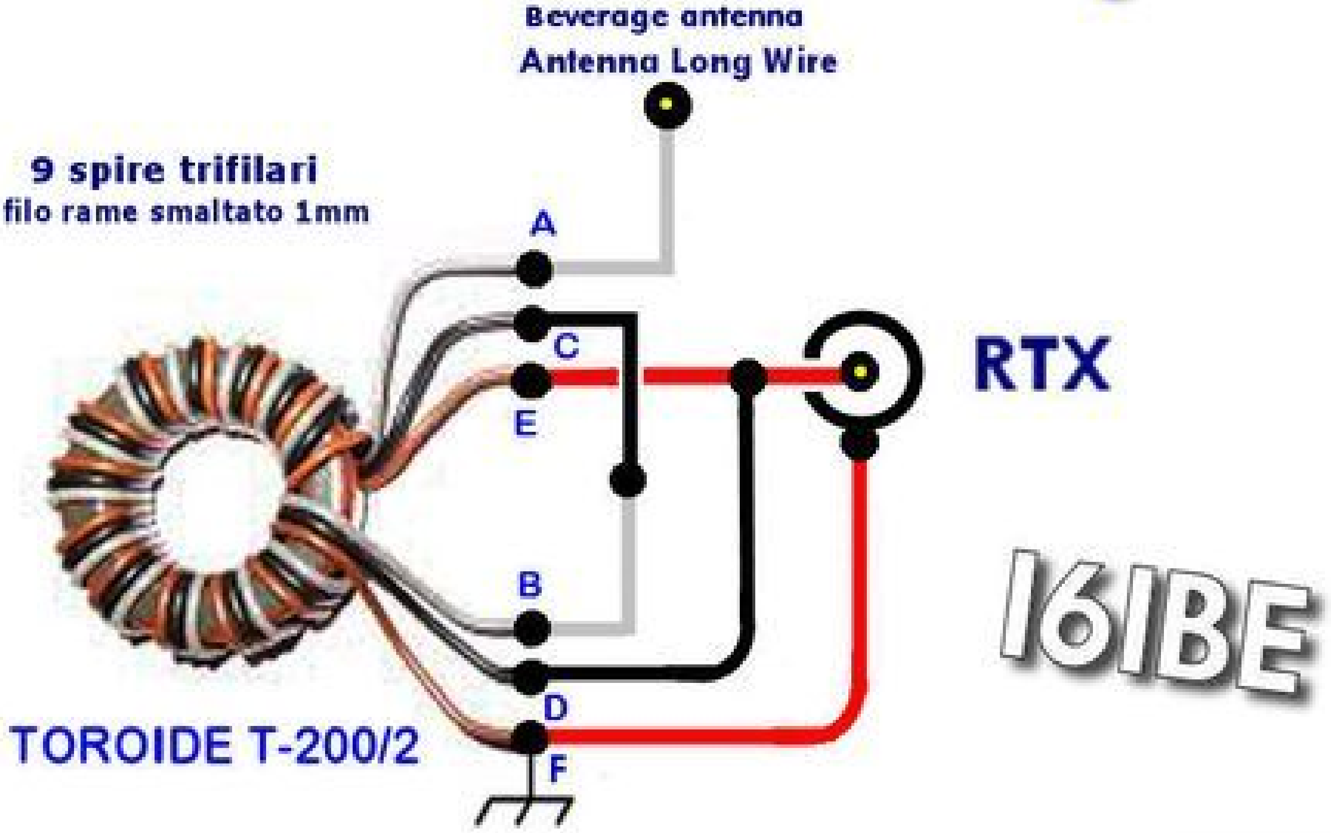
Beverage antenna
Antenna Long Wire
9 spire trifilari
filo rame smaltato 1mm
A
C
E
B
D
F
RTX
I6IBE
TOROIDE T-200/2

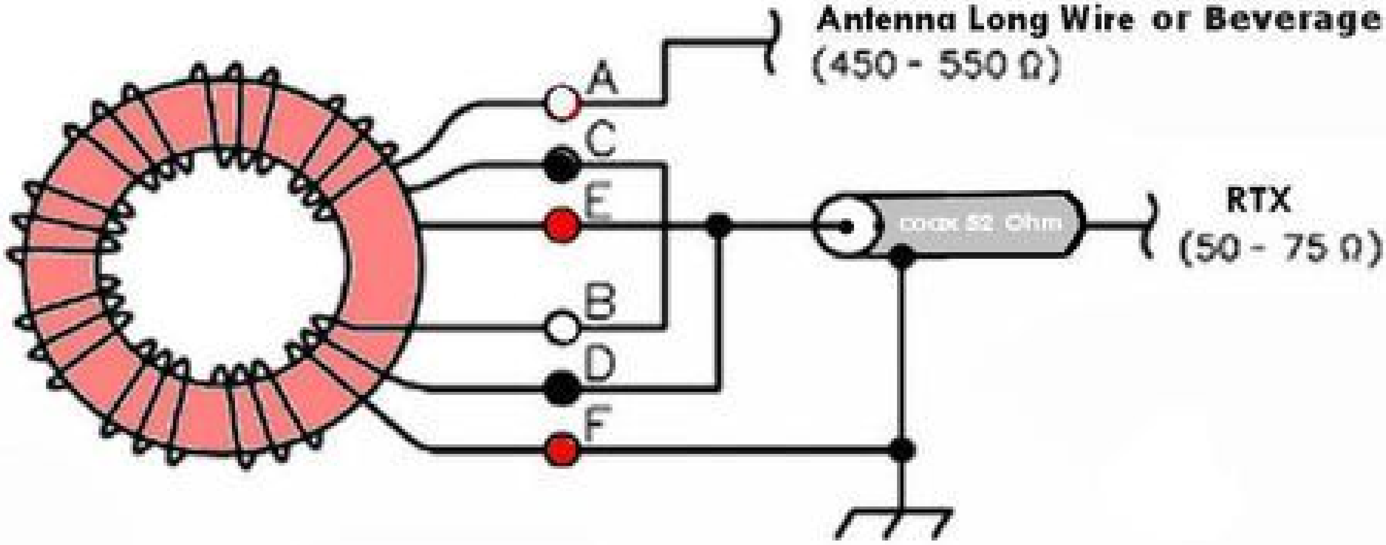
Antenna Long Wire or Beverage
(450 - 550 Ω)
A
C
E
B
D
F
coax 52 Ohm
RTX
(50 - 75 Ω)

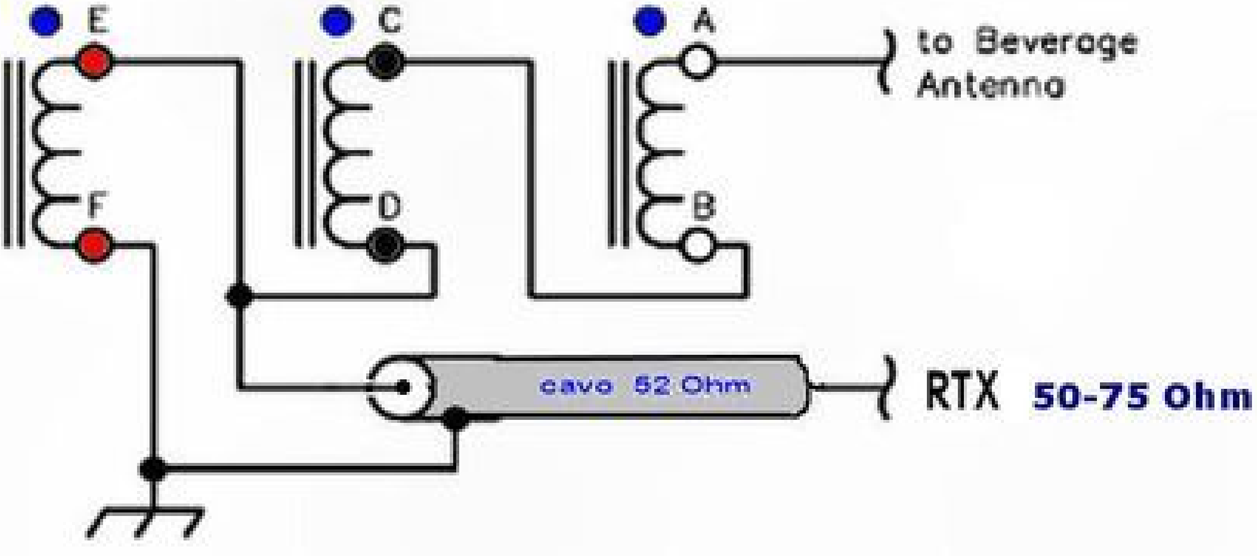
E
F
C
D
A
B
to Beverage
Antenna
cavo 52 Ohm
RTX
50-75 Ohm

Bal-Un 2:1 per antenna DELTA-LOOP

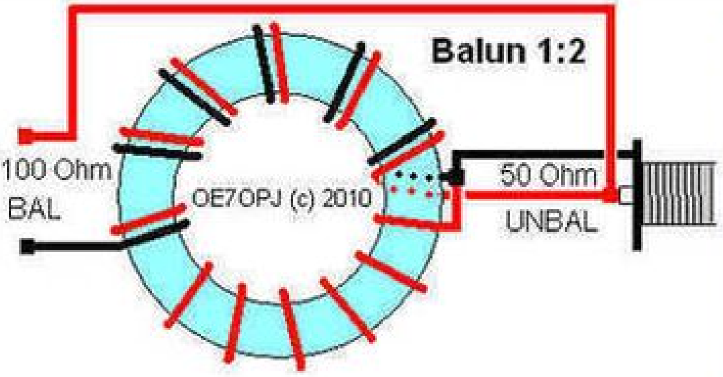
Balun 1:2
100 Ohm
BAL
OE7OPJ (c) 2010
50 Ohm
UNBAL

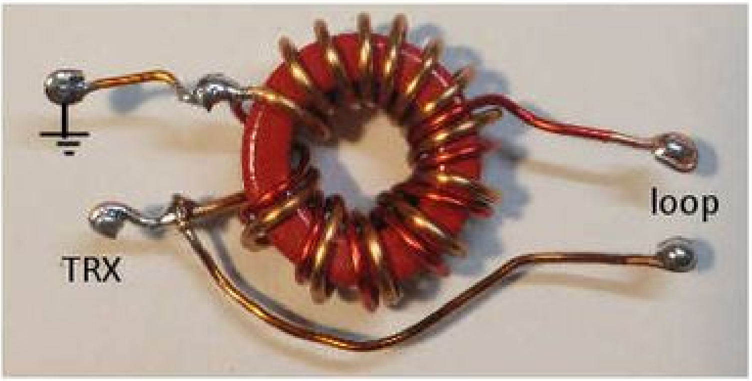

by OE7OPJ

Toroide	Numero di Spire	Potenza Supportata
T80-2	25	60 Watts
T106-2	16	100 Watts
T130-2	18	150 Watts
T157-2	16	250 Watts
T200-2	17	400 Watts
T200A-2	13	400 Watts
T400-2	14	1000 Watts

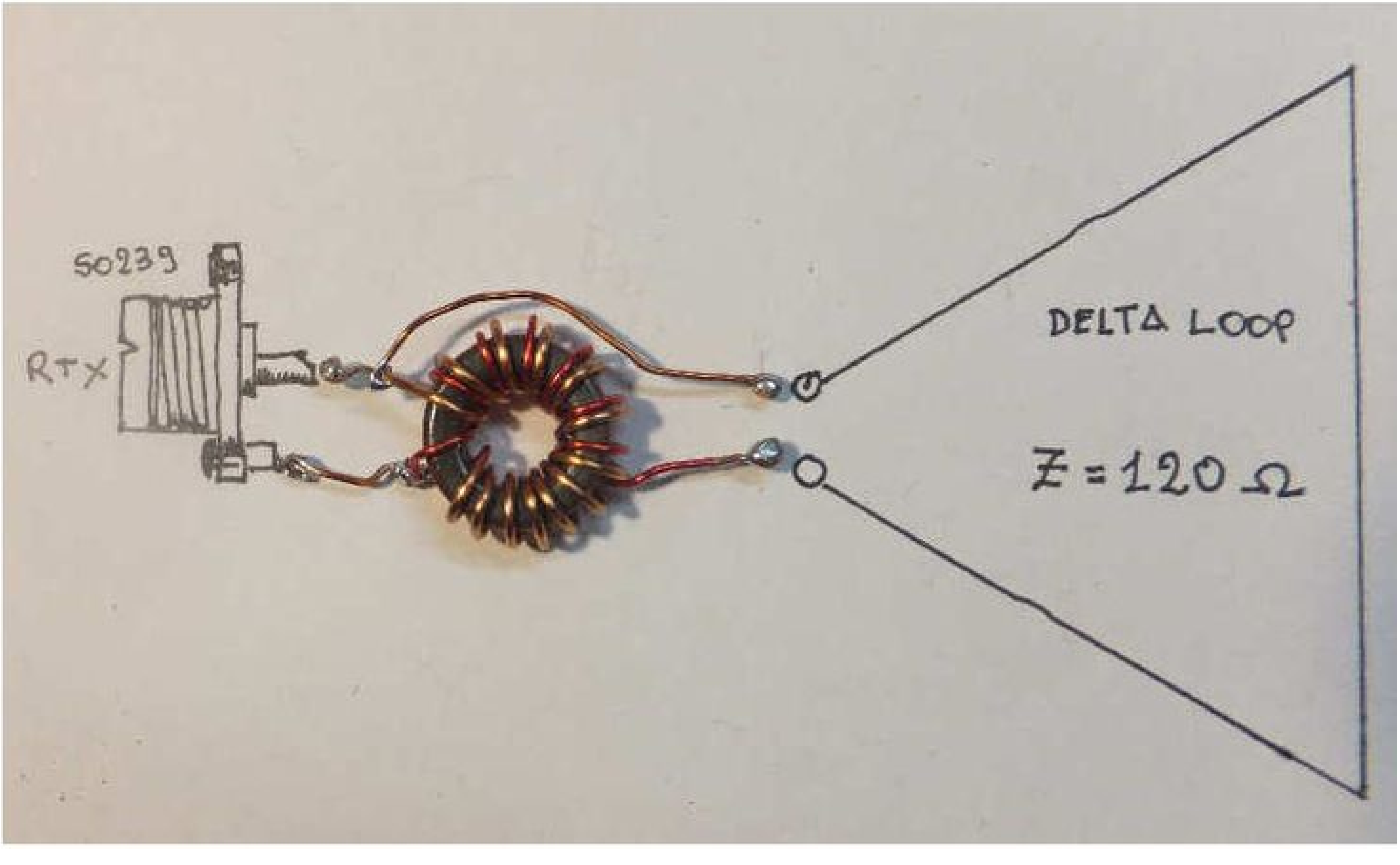
SO239
RTX
DELTA LOOP
Z = 120 Ω

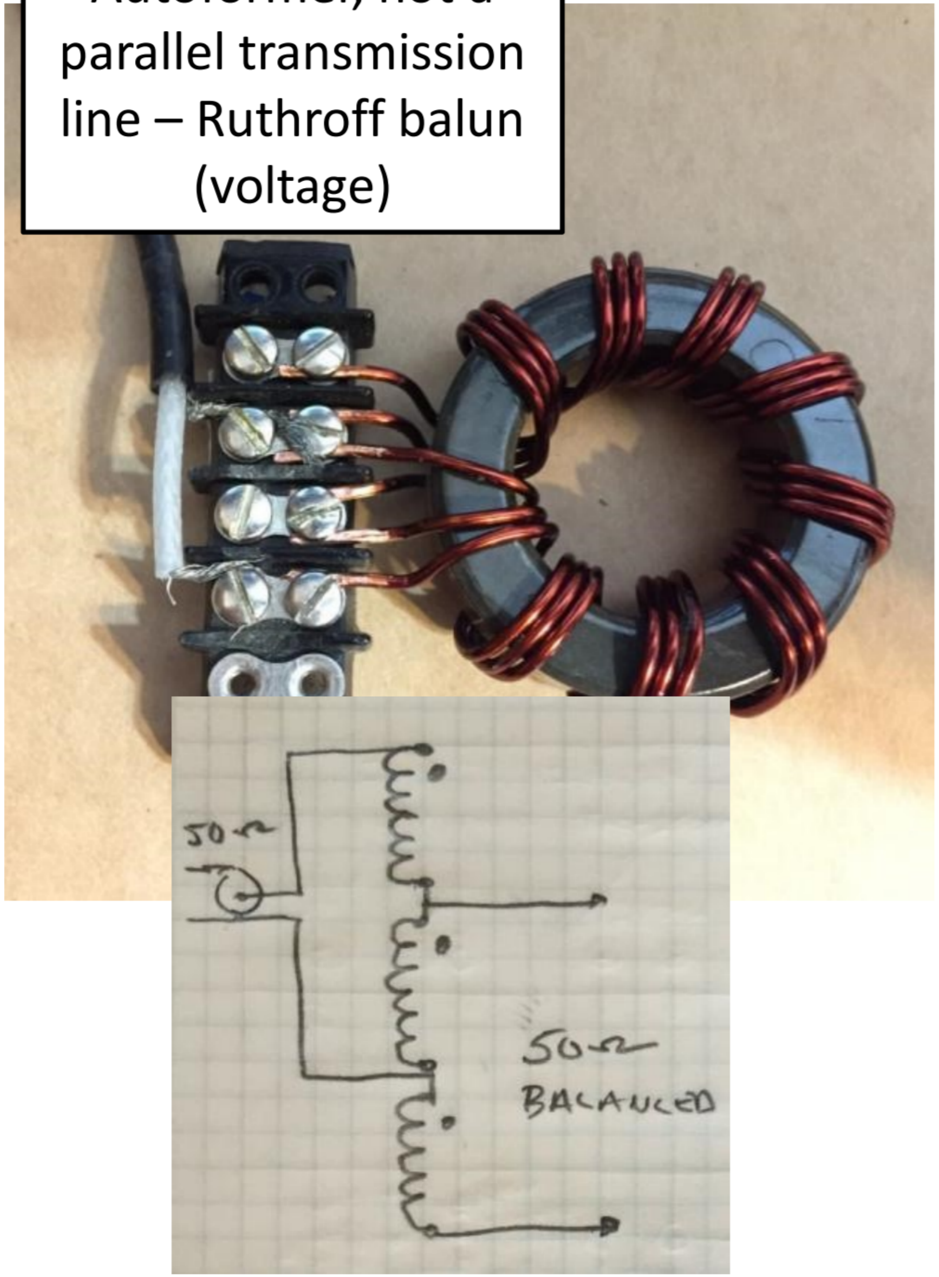

Autoformer, not a parallel transmission line – Ruthroff balun (voltage)

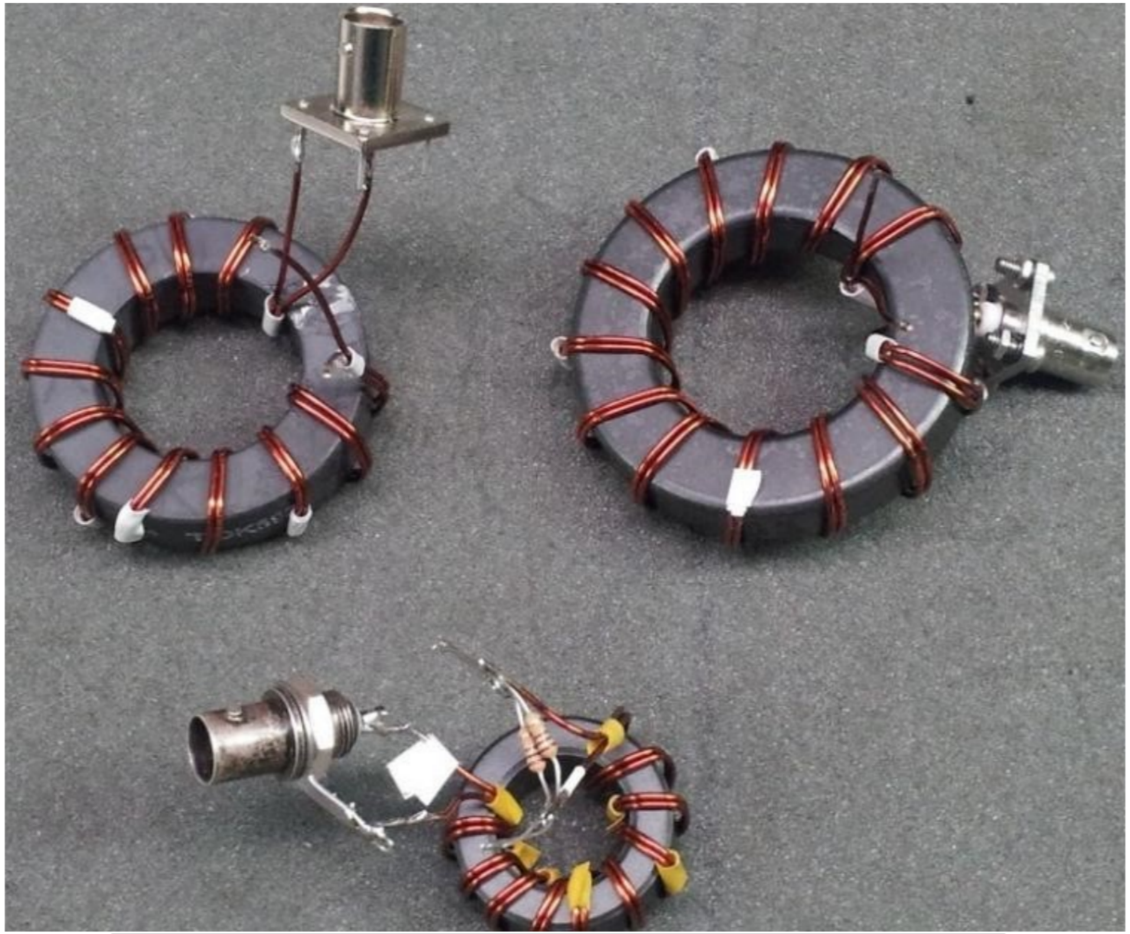

One or more parallel transmission lines, connected various ways – Guanella balun (current) – core is used solely for choking action

F/B: 0.00 dB; Rear: Azim. 120 deg, Elev. 60 deg

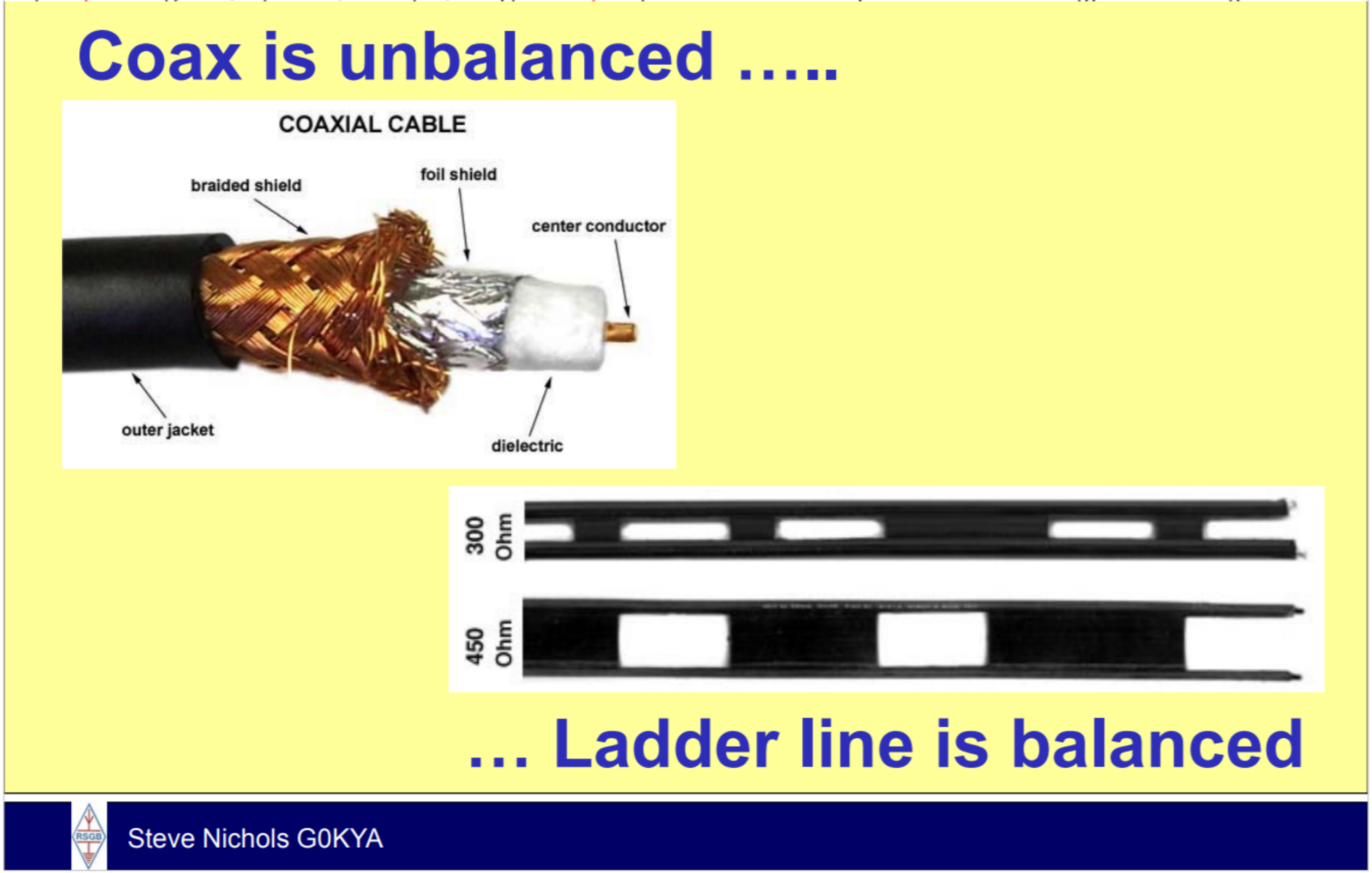

Coax is unbalanced …..

COAXIAL CABLE
braided shield
foil shield
center conductor
outer jacket
dielectric
300 Ohm
450 Ohm

… Ladder line is balanced

Steve Nichols G0KYA

Coax is unbalanced, while ladder line and open wire feeder is balanced, as long as it is used properly and is kept away from metal and "earthy" materials

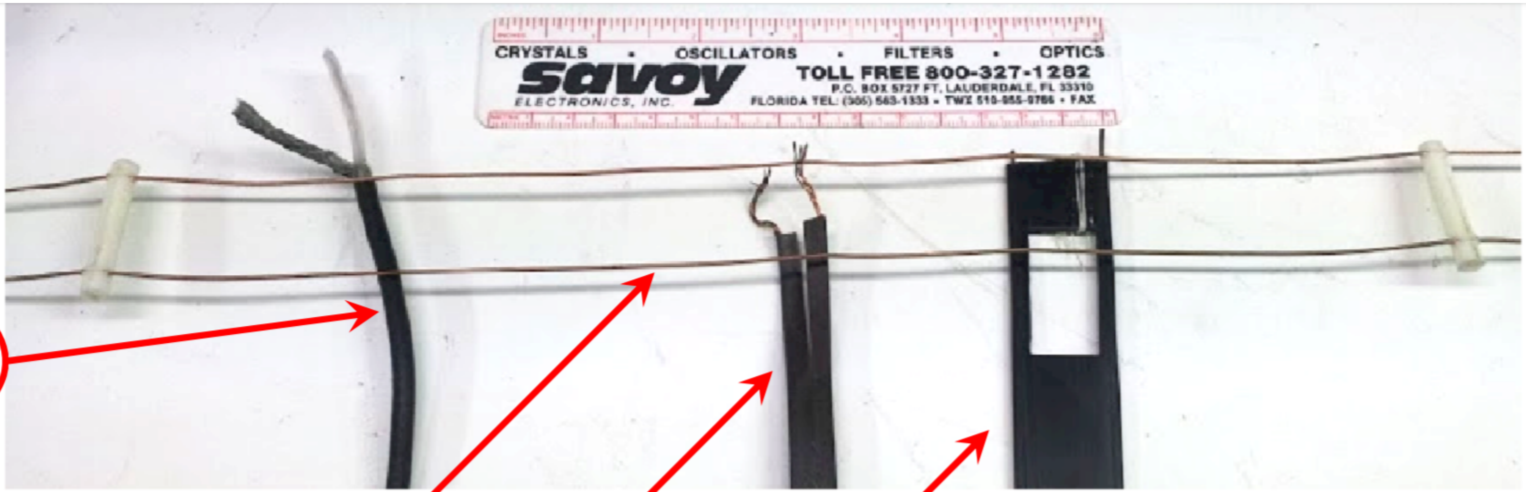
CRYSTALS • OSCILLATORS • FILTERS • OPTICS
Savoy
ELECTRONICS, INC.
TOLL FREE 800-327-1282

Convert parallel feedline to coax

Types of Parallel Line

- **Open Feeder** or **Ladder line**
  - Typ 450 and 600 ohm
- **Twin Lead**
  - Common for (old) outdoor TV antennas
- **Window line**
  - Commonly available today

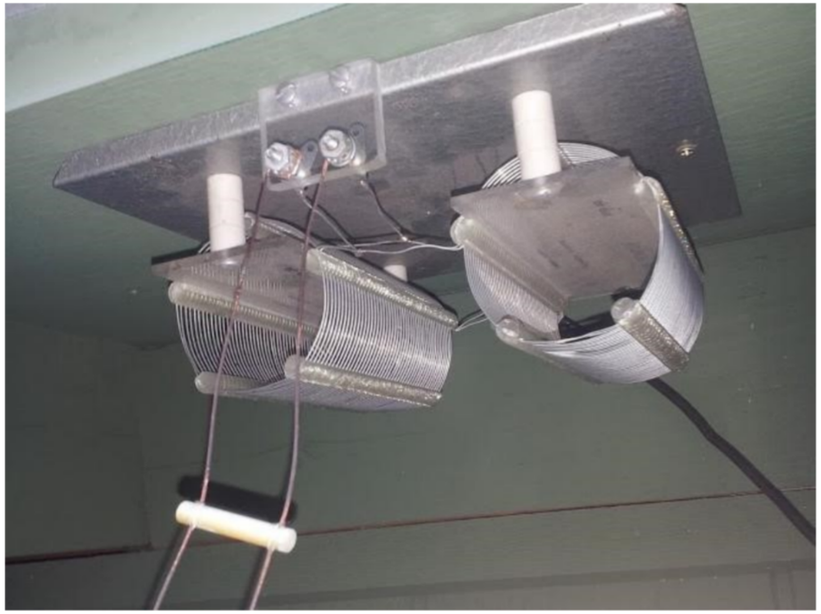

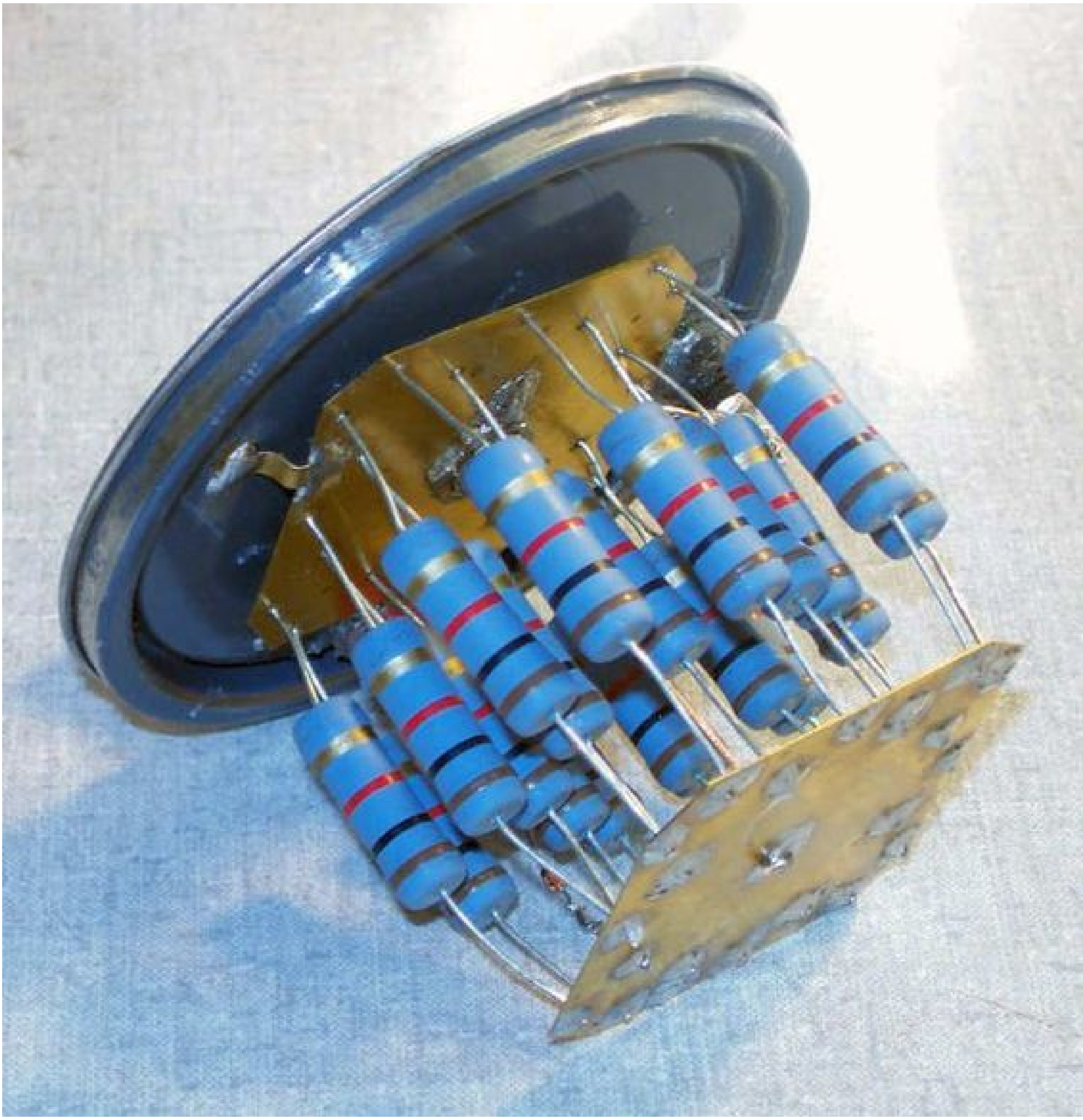

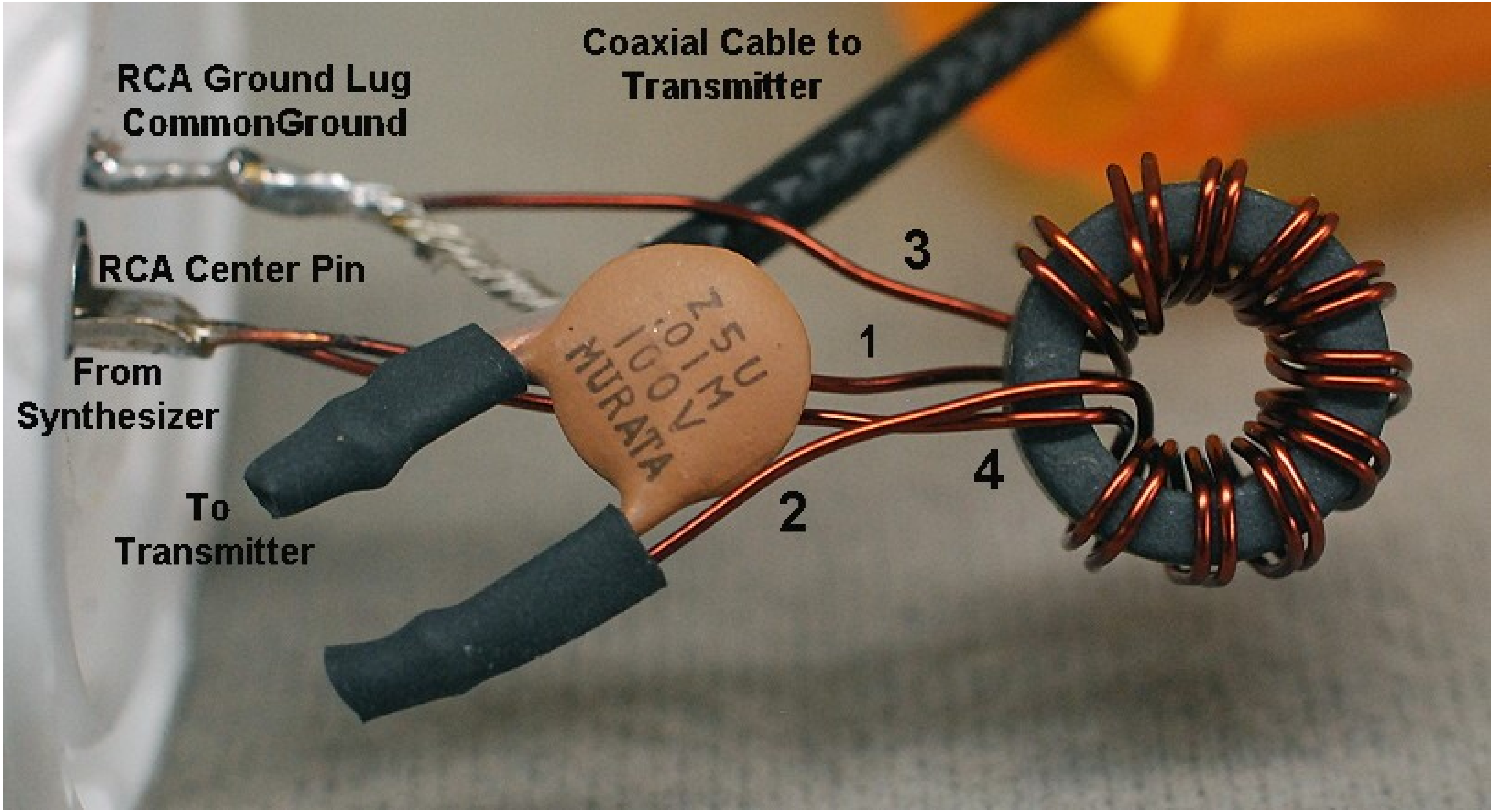

Coaxial Cable to Transmitter
RCA Ground Lug CommonGround
RCA Center Pin
From Synthesizer
To Transmitter
3
1
4
2
Z5U
.01M
100V
MURATA

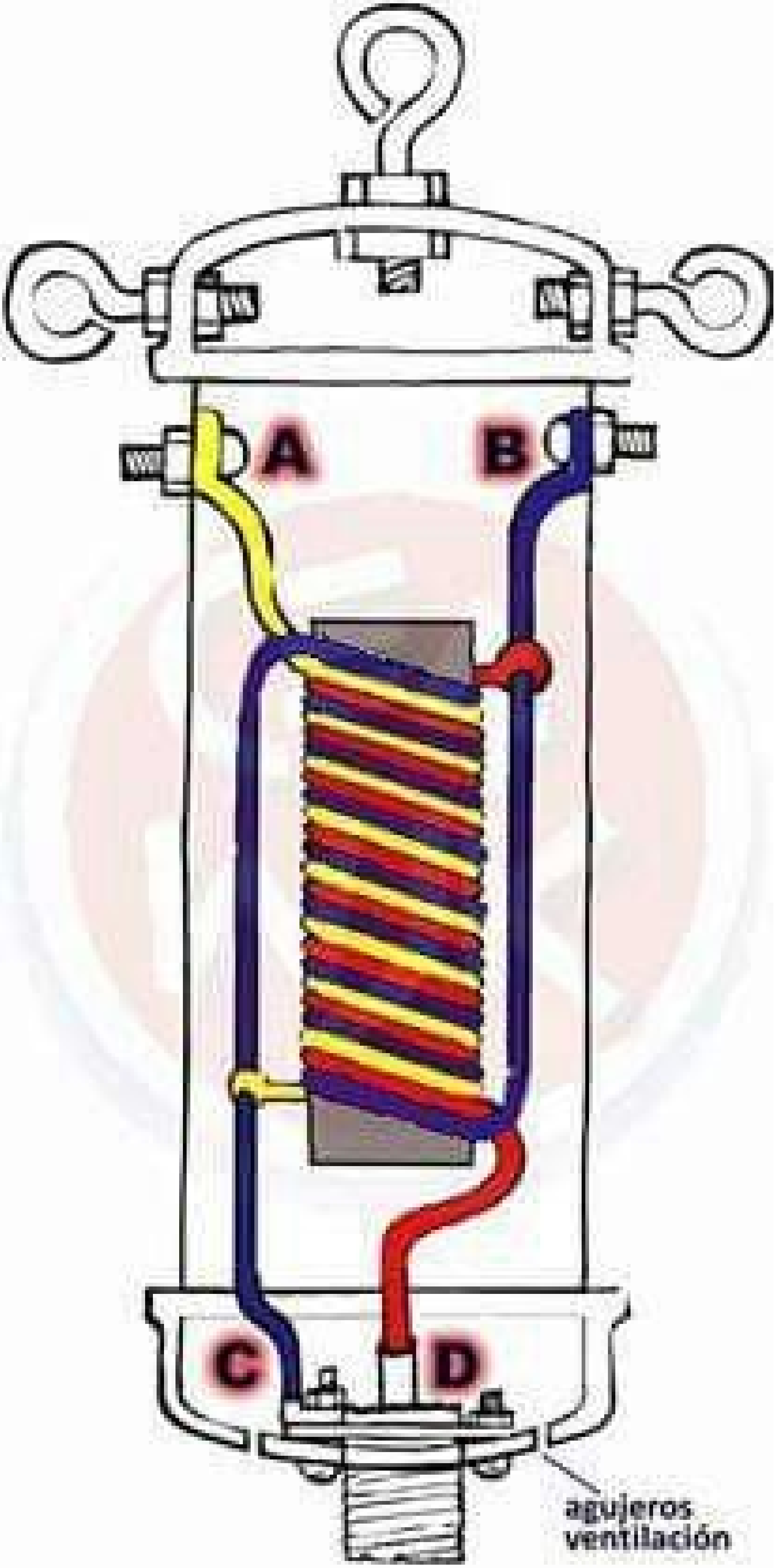
A
B
C
D
agujeros
ventilación

Formula LUNGHEZZA DELTA LOOP	Esempio 14.00 MHz (20 metri)	Lunghezza TOTALE del LOOP
Velocita'Luce : Freq MHz x 0.97= L	**299.8** VL : **14.100** MHz x **0,97**	**20.624** metri totale loop
$L=\frac{300.000}{MHz} \times 0.97$	I6IBE	
Formula LUNGHEZZA Stub coax 75 Ohm	Esempio 14.00 MHz (20 metri)	Lunghezza TOTALE cavo 75 Ohm
Lunghezza LOOP x 0.66 : 4	**20.624** x **0.66** : **4**	**3.402** Metri coax RG-59

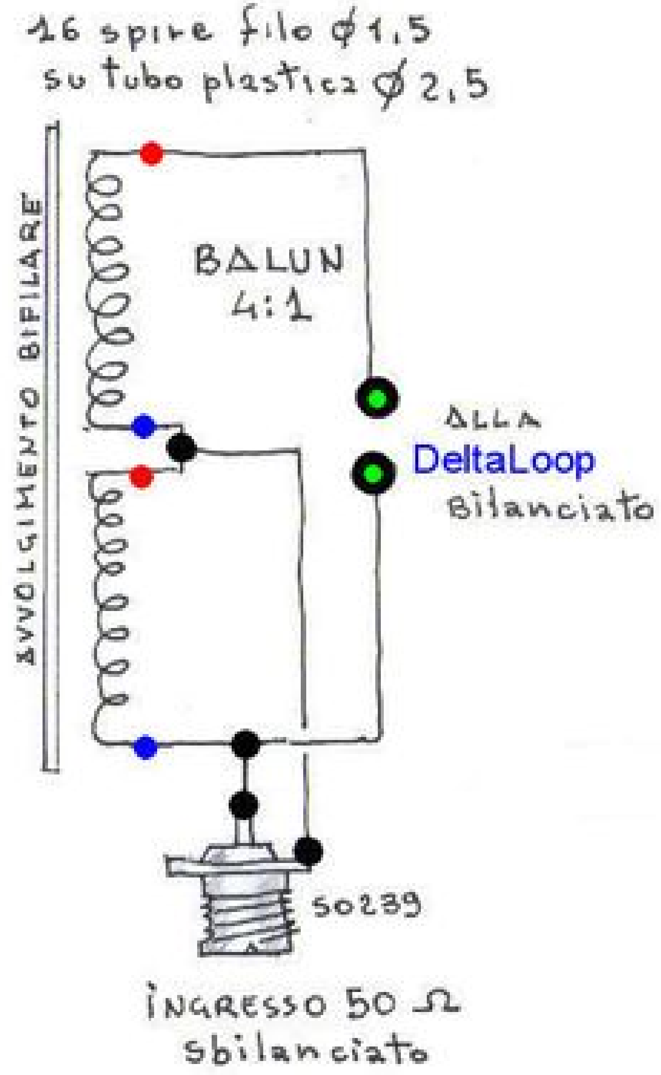
16 spire filo ø1,5
su tubo plastica ø 2,5
avvolgimento bifilare
BALUN 4:1
alla
DeltaLoop
bilanciato
SO239
ingresso 50 Ω
sbilanciato

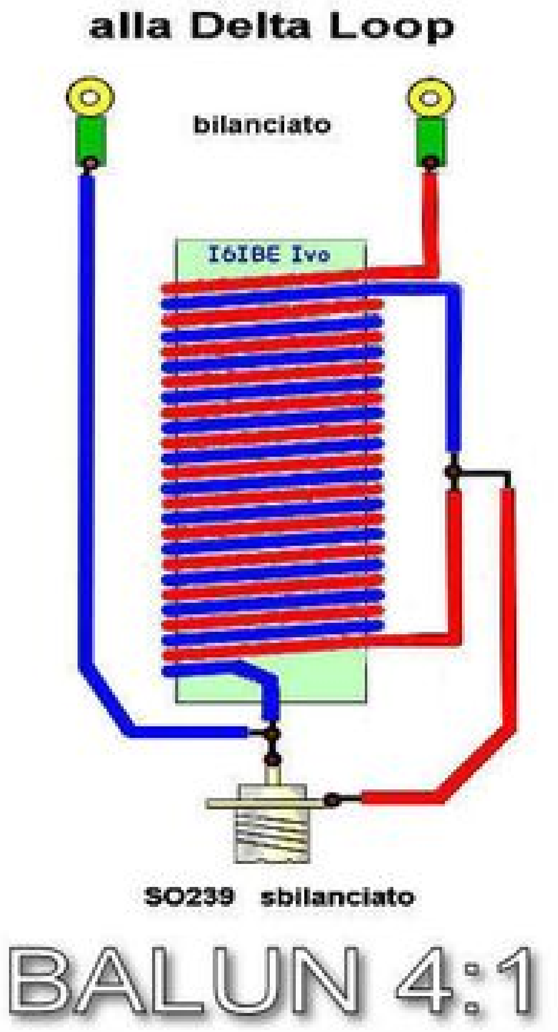
alla Delta Loop
bilanciato
I6IBE Ivo
SO239 sbilanciato
BALUN 4:1

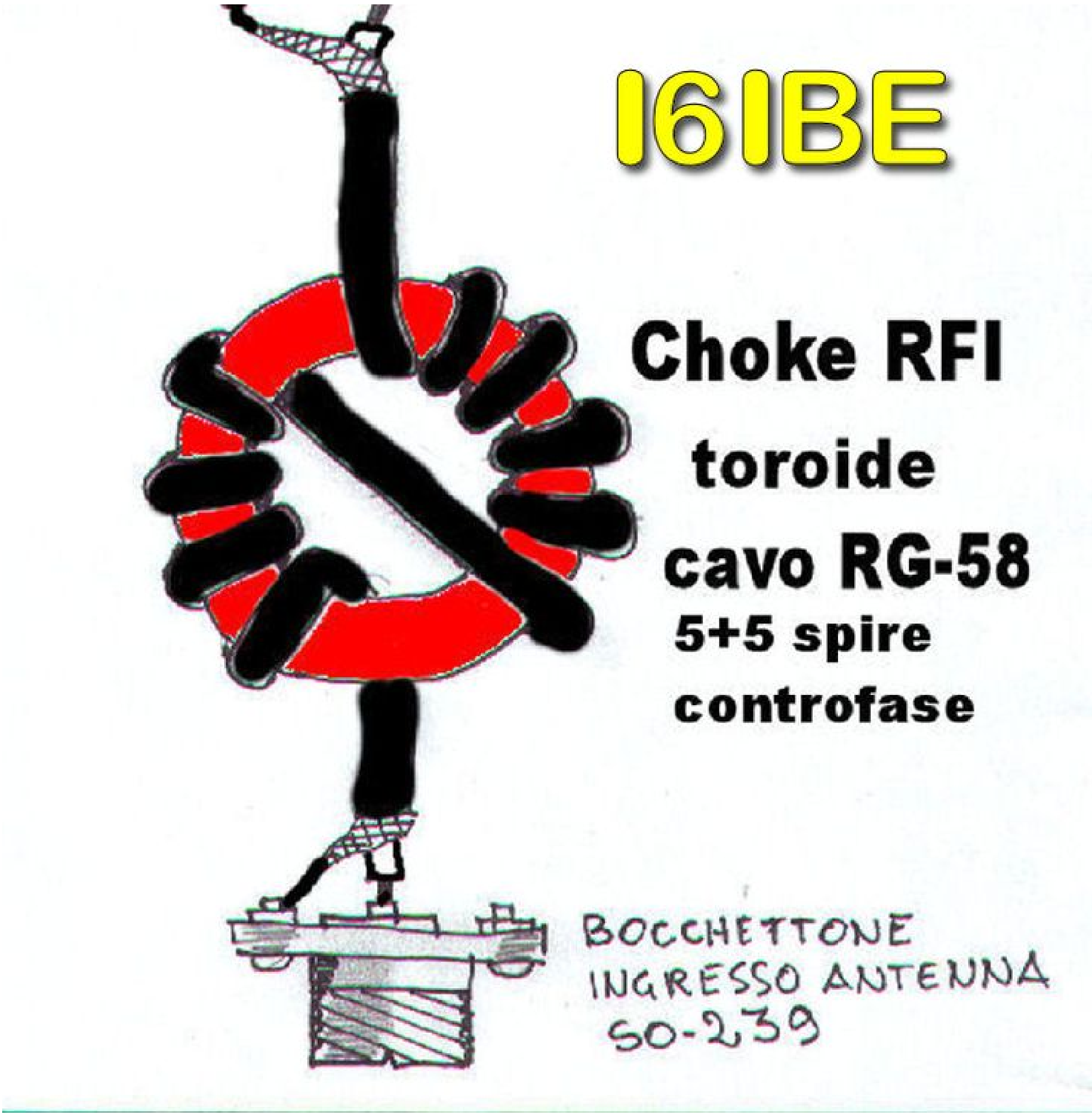
I6IBE
Choke RFI
toroide
cavo RG-58
5+5 spire
controfase
BOCCHETTONE
INGRESSO ANTENNA
SO-239

2. Insert metal electrodes and attach multimeter leads [copper (+), aluminum (-)] For measuring direct current voltage: set multimeter function switch to "DCV: 20" take a reading in volts DC. for measuring direct current: set multimeter function switch to "DCA: 20m" take a reading in milliamps (mA) DC.

Lawn battery (summer). INSET: Marsh mud battery. (Circles show position of electrodes.)

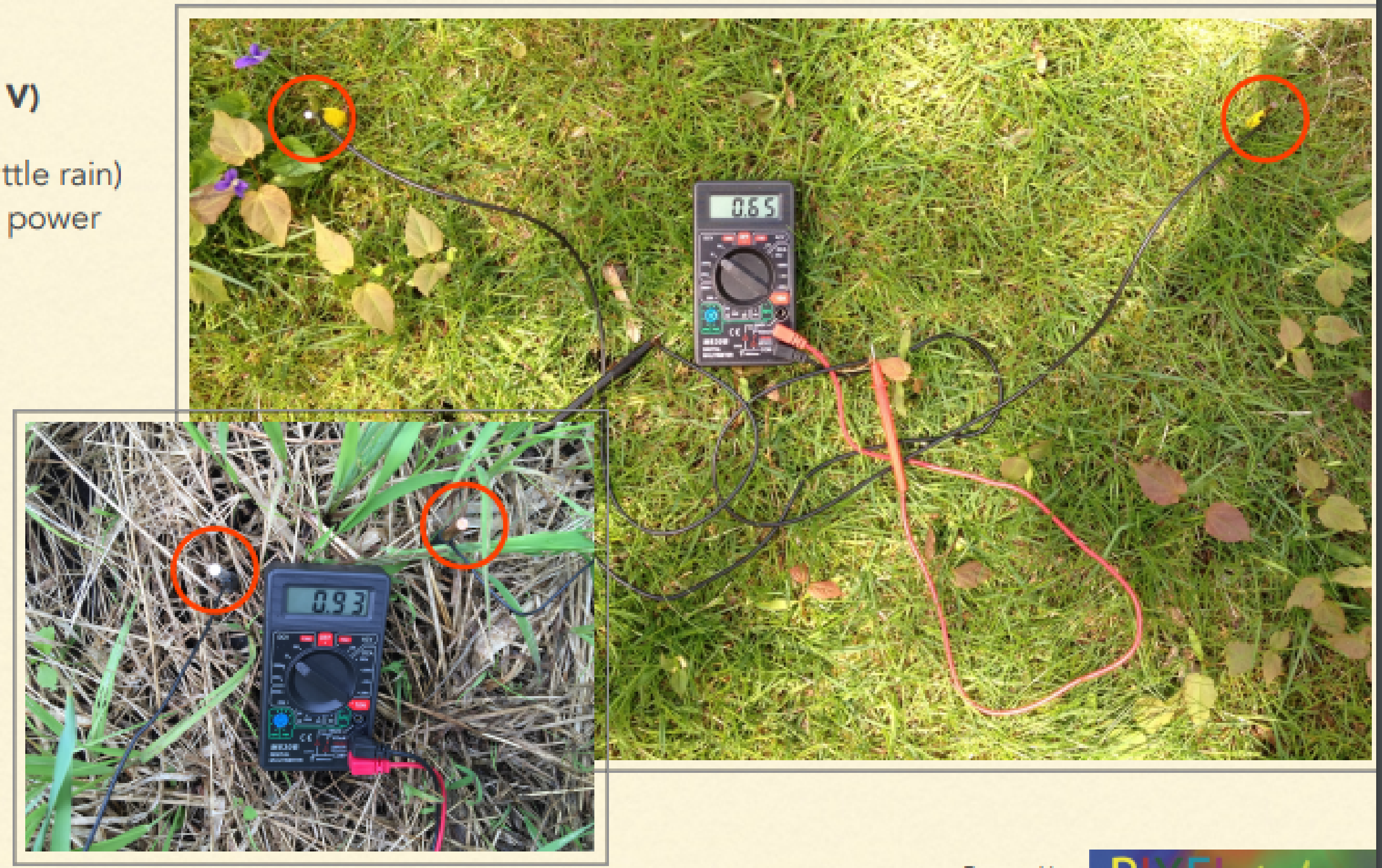

Calculating Earth Battery Power (W = I * V)

EXAMPLE: A lawn battery in late summer (little rain) produces a 0.65V, 0.2mA current. A battery power calculation of 0.00013W (0.13mW).

Powered by: PIXEL ink LLC

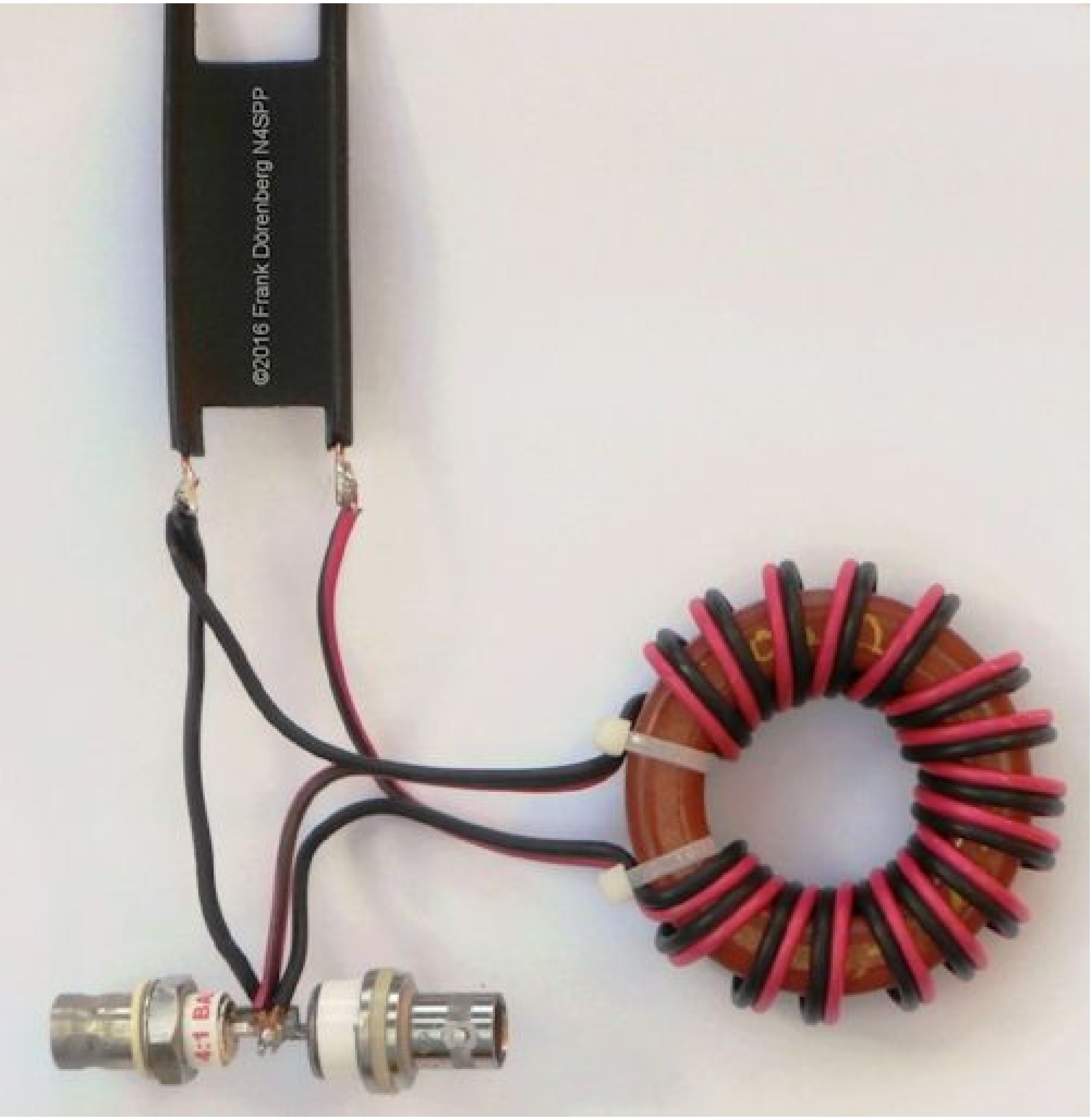
©2016 Frank Dörenberg N4SPP
4:1 BA

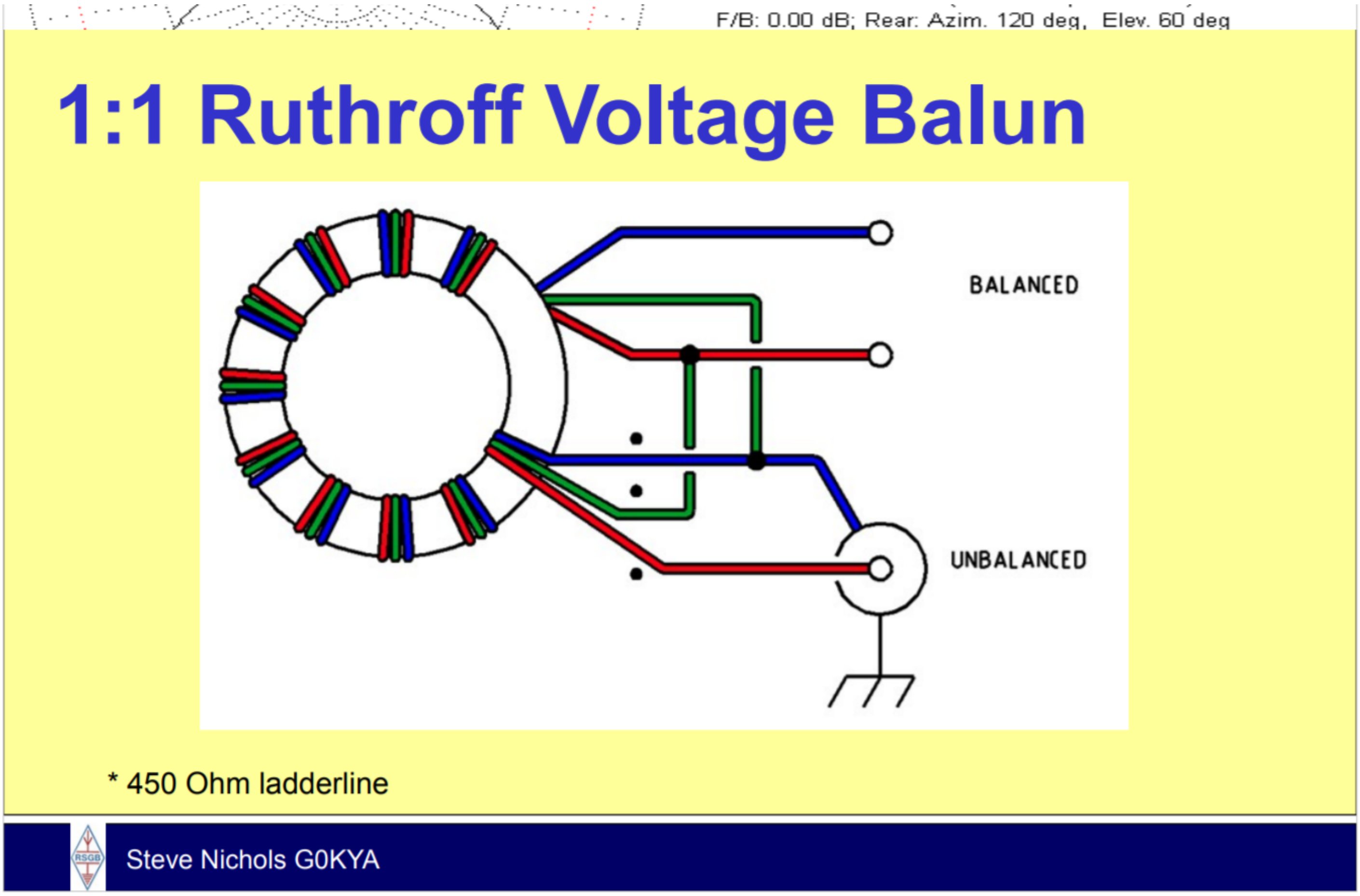

1:1 Ruthroff Voltage Balun

* 450 Ohm ladderline

Steve Nichols G0KYA

The Ruthroff or voltage balun is another type. Voltage baluns balance the voltages. Current balance is considered to be superior to voltage balance. Voltage baluns should not be used in lines with high SWR. They have the narrowest impedance and frequency range of any balun type. Properly designed voltage baluns have low common mode impedance. Properly designed current baluns have high common mode impedance, and provide better balance. If you want to stop common mode currents flowing on your coax a current balun is a better choice. But in the Carolina Windom, which uses the vertical coax as a radiator you WANT common mode currents, so a Ruthroff voltage balun is better.

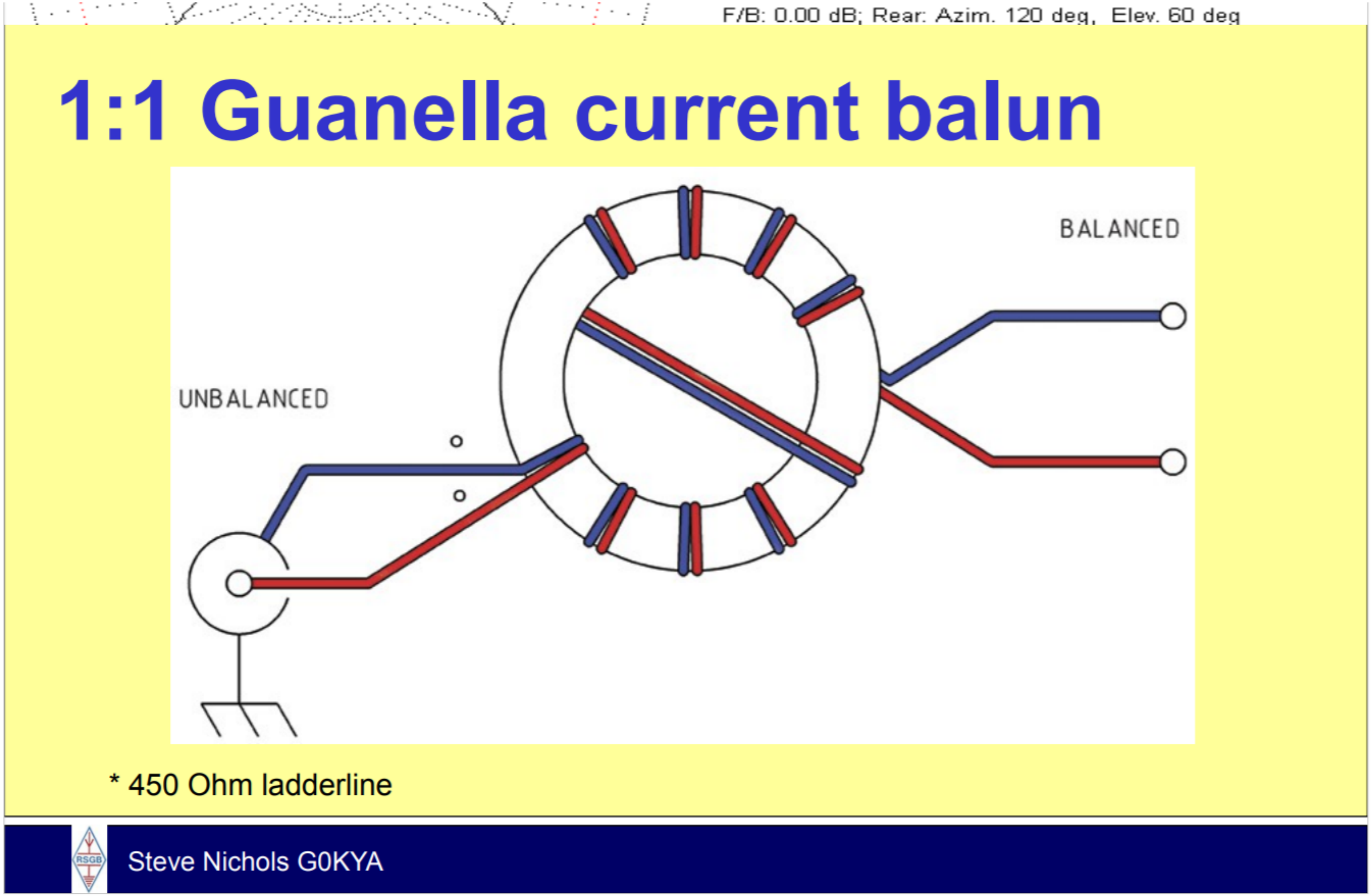
F/B: 0.00 dB; Rear: Azim. 120 deg, Elev. 60 deg
BALANCED
UNBALANCED

1:1 Guanella current balun

* 450 Ohm ladderline

Steve Nichols G0KYA

A Guanella or current balun is very common. Current baluns stop RF from coming back down the outside of the coax shield, and are so called that because they "force equal currents in each side of a dipole"

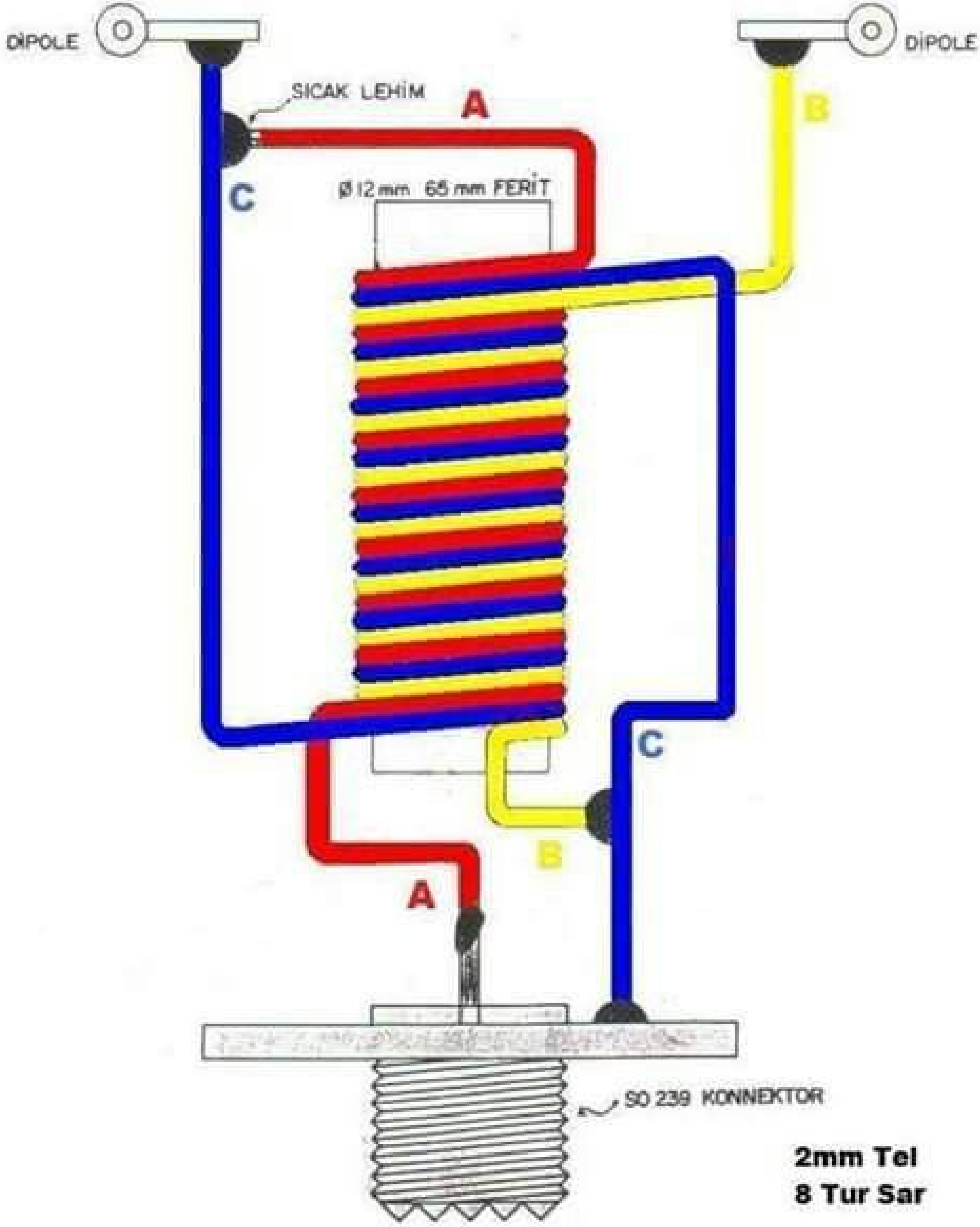
DİPOLE
DİPOLE
SICAK LEHİM
A
B
C
Ø12 mm 65 mm FERİT
C
B
A
SO 239 KONNEKTOR

2mm Tel
8 Tur Sar

© Azega.com

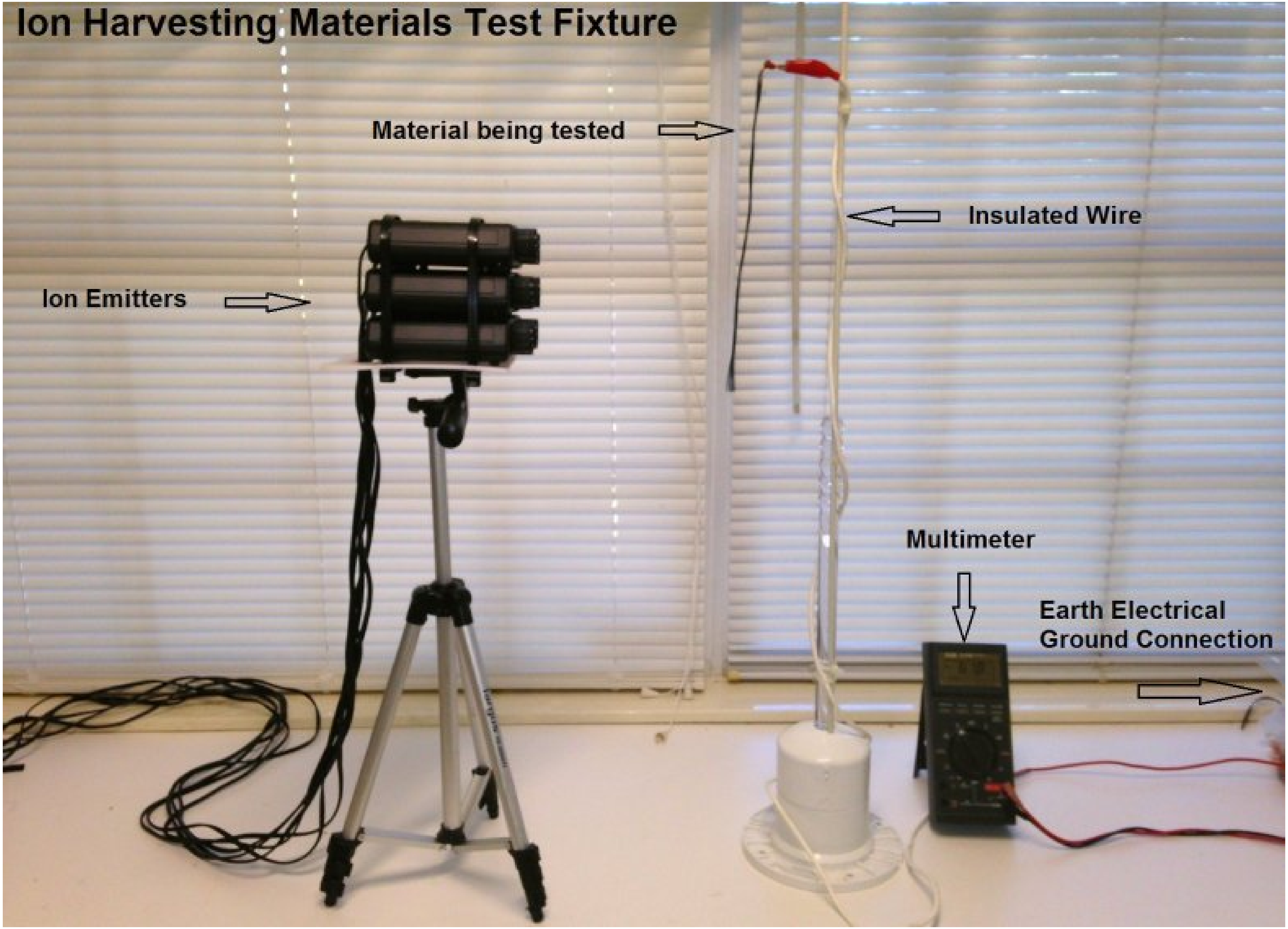

Ion Harvesting Materials Test Fixture

Material being tested
Insulated Wire
Ion Emitters
Multimeter
Earth Electrical
Ground Connection

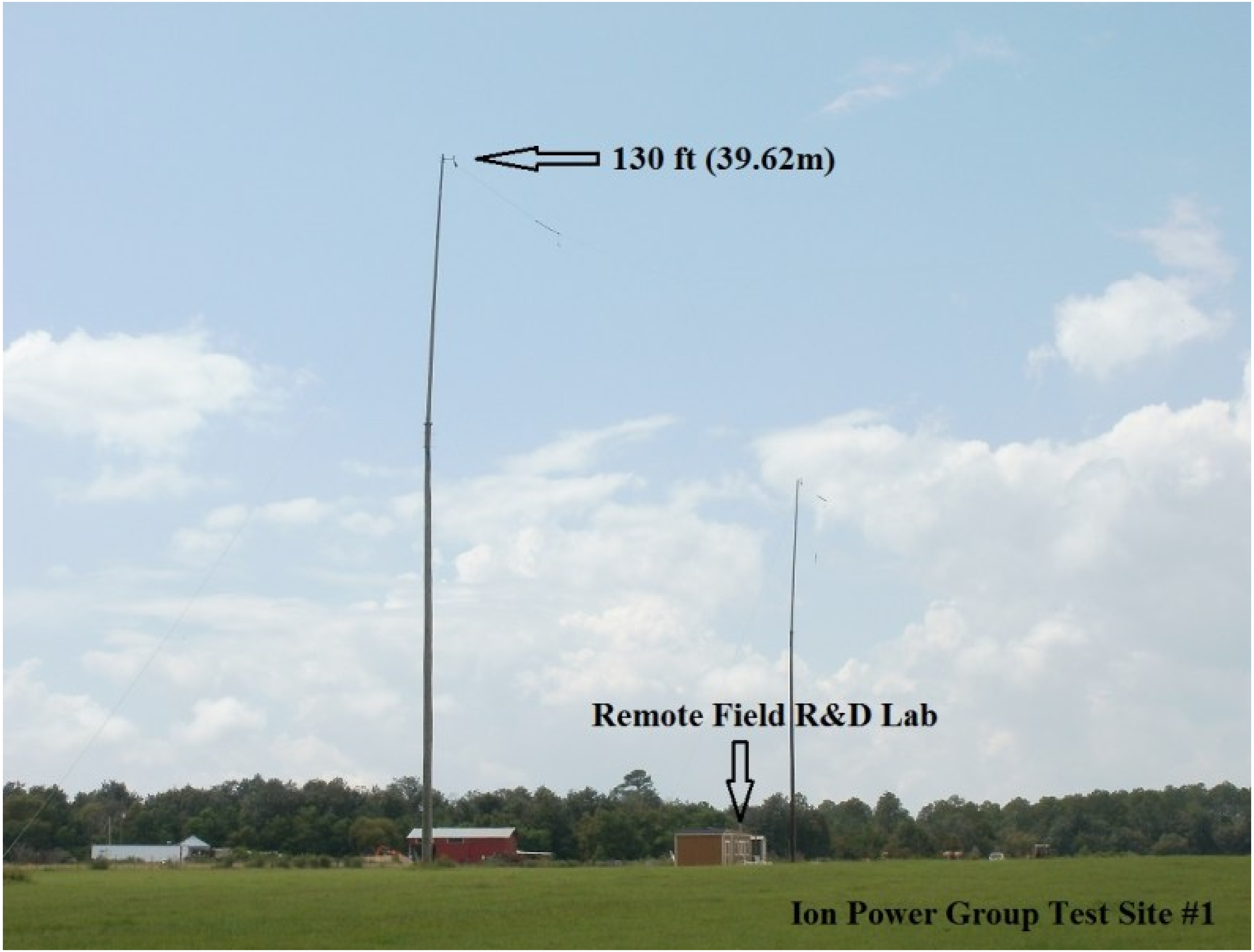
130 ft (39.62m)
Remote Field R&D Lab
Ion Power Group Test Site #1

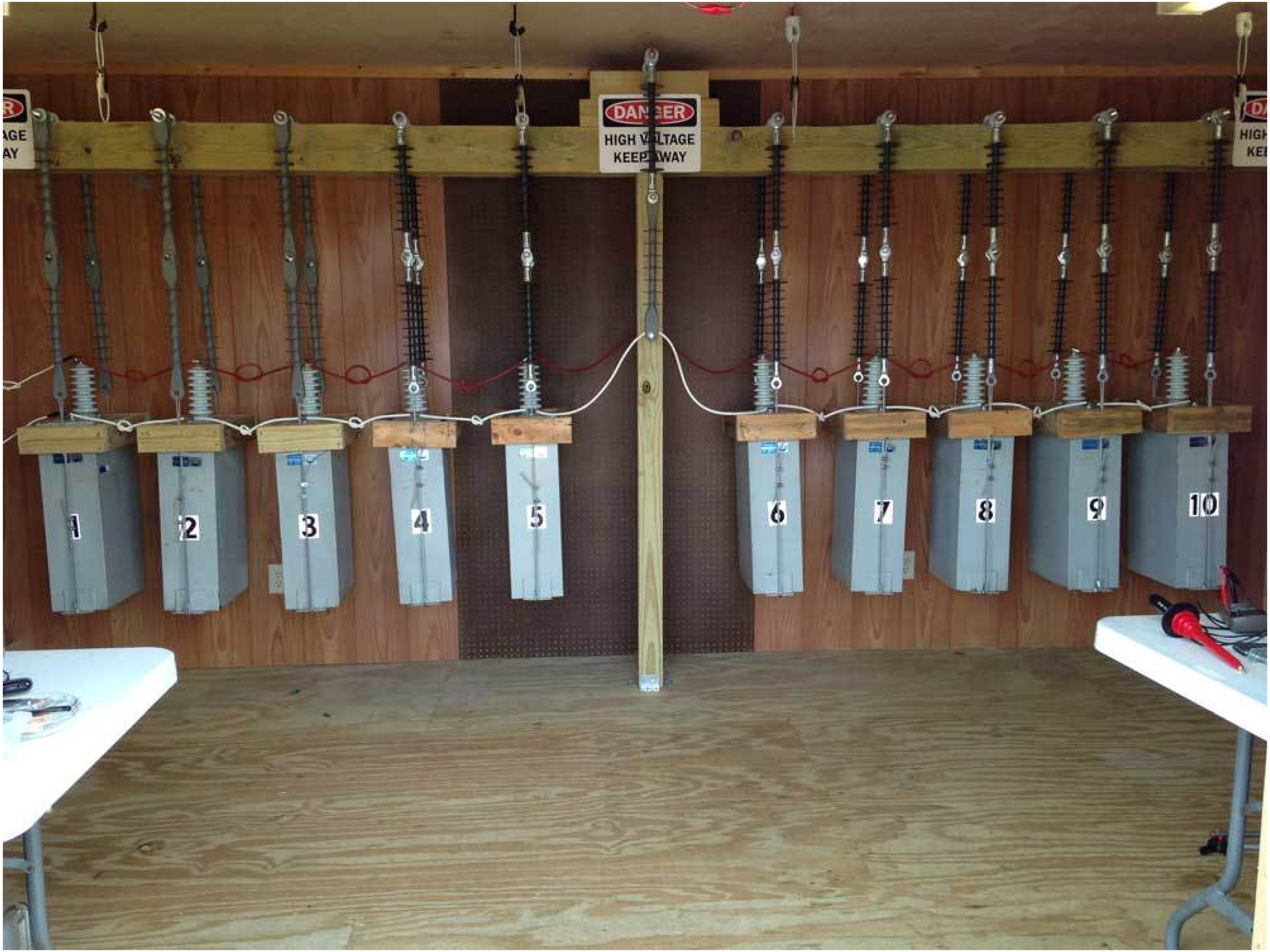
DANGER
HIGH VOLTAGE
KEEP AWAY
1
2
3
4
5
6
7
8
9
10

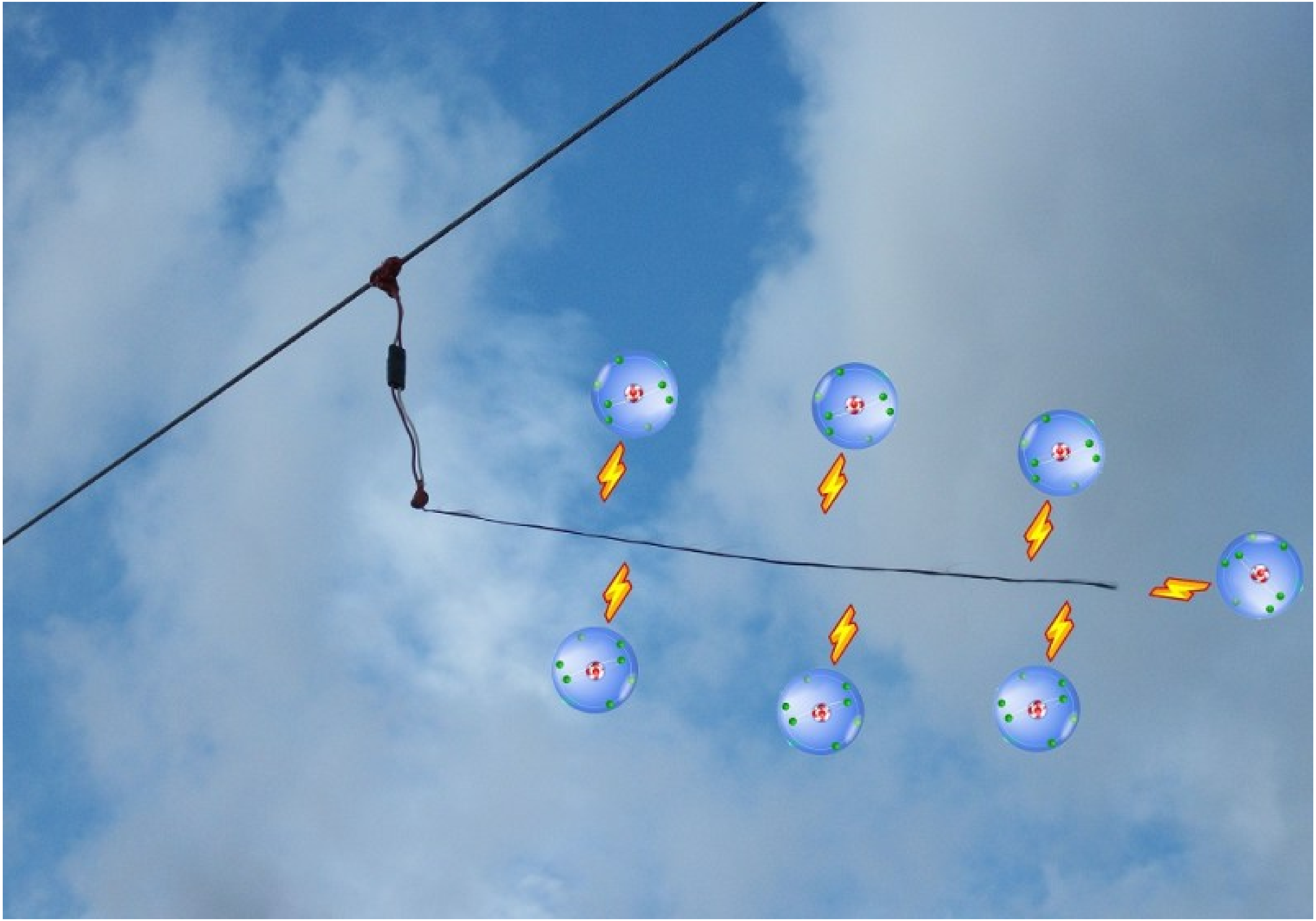

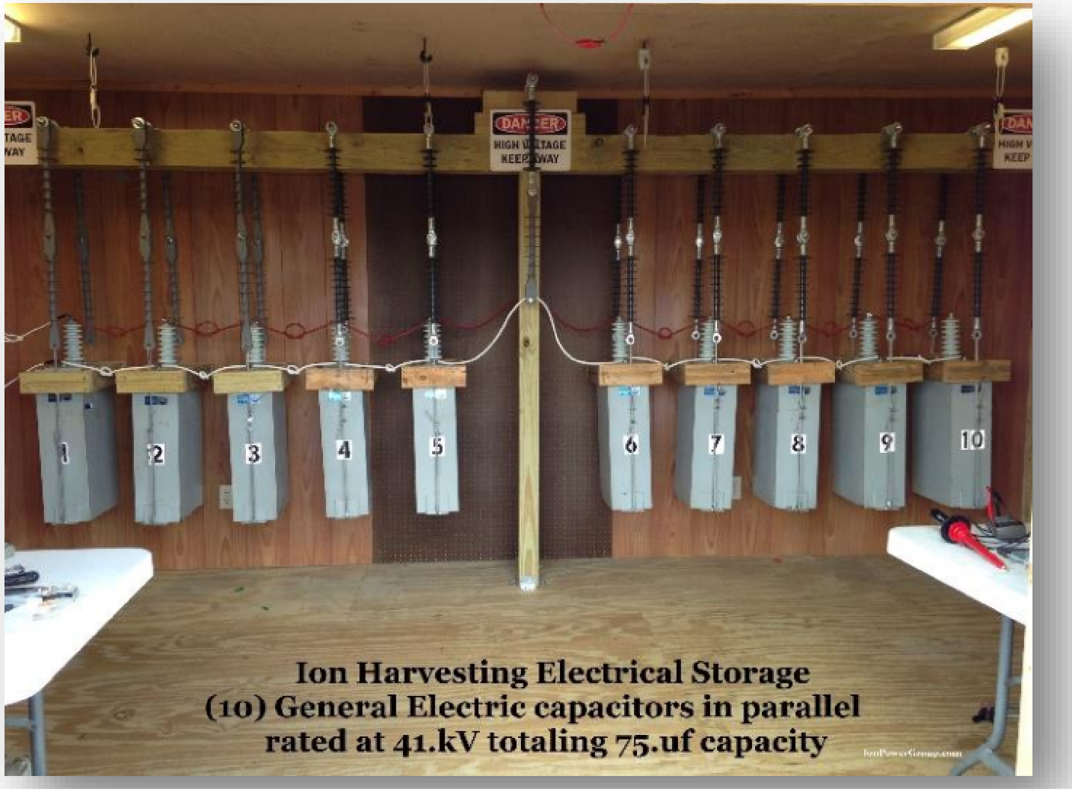
DANGER
HIGH VOLTAGE
KEEP AWAY
1
2
3
4
5
6
7
8
9
10
IonPowerGroup.com

Ion Harvesting Electrical Storage
(10) General Electric capacitors in parallel
rated at 41.kV totaling 75.uf capacity

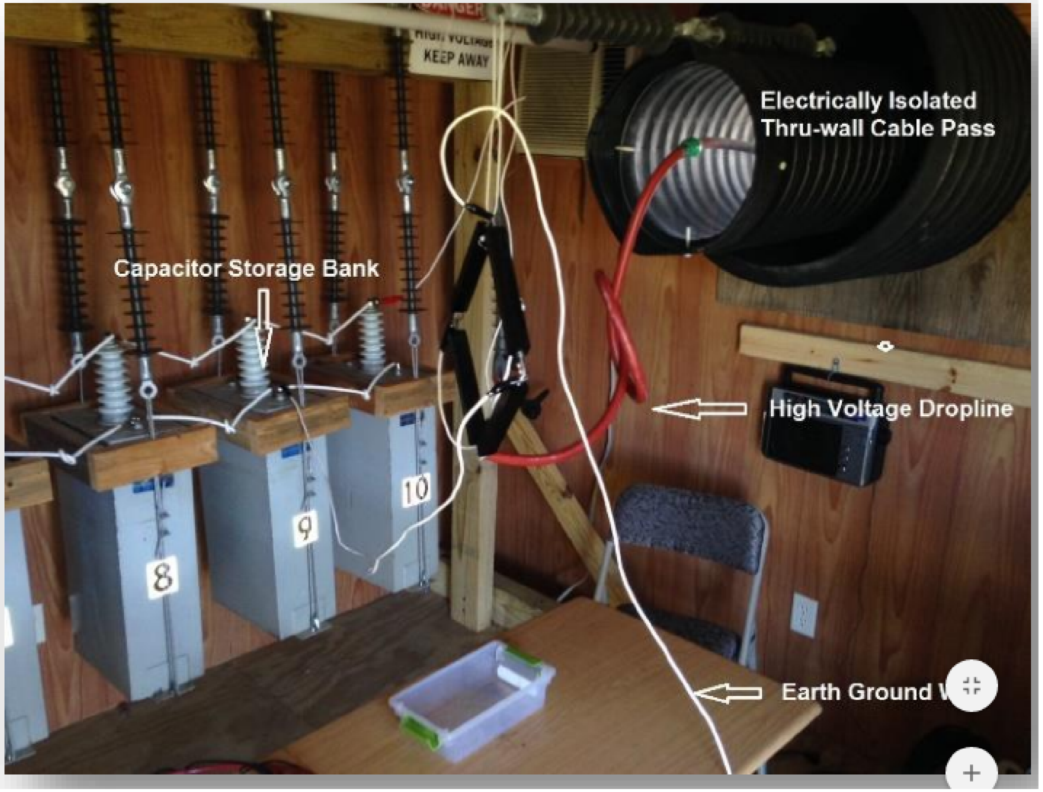
Electrically Isolated
Thru-wall Cable Pass
Capacitor Storage Bank
High Voltage Dropline
Earth Ground W
8
9
10
KEEP AWAY

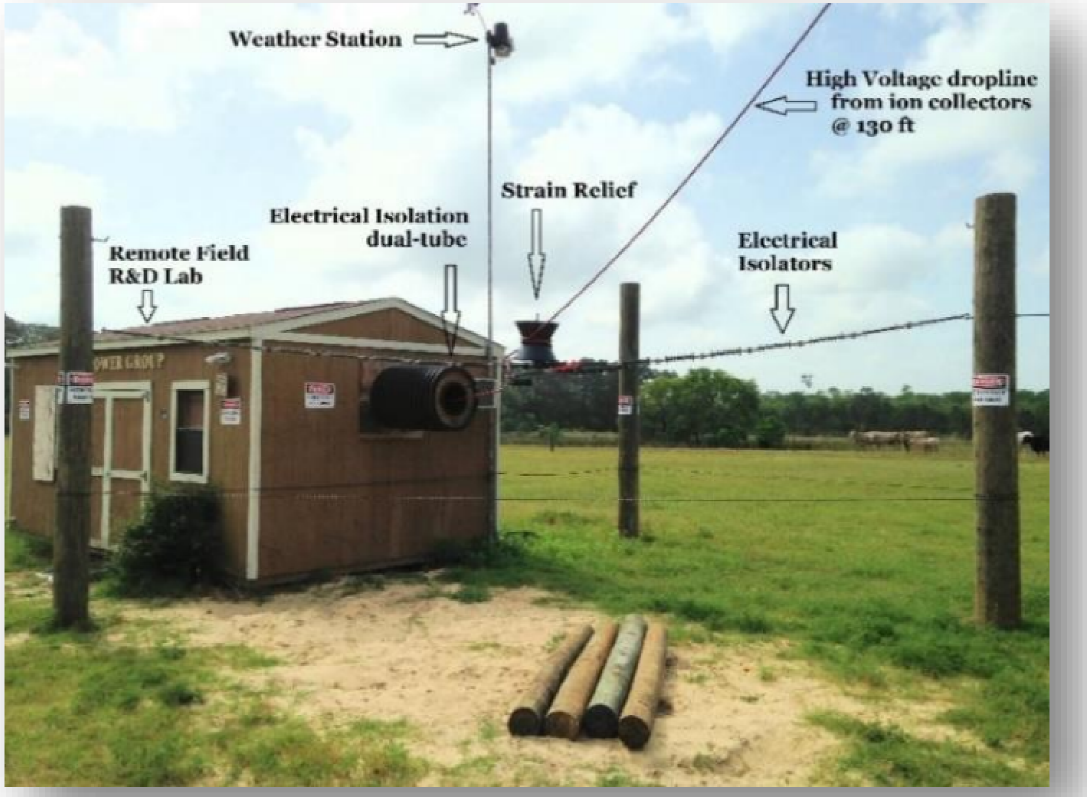
Weather Station
High Voltage dropline from ion collectors @ 130 ft
Strain Relief
Electrical Isolation dual-tube
Remote Field R&D Lab
Electrical Isolators

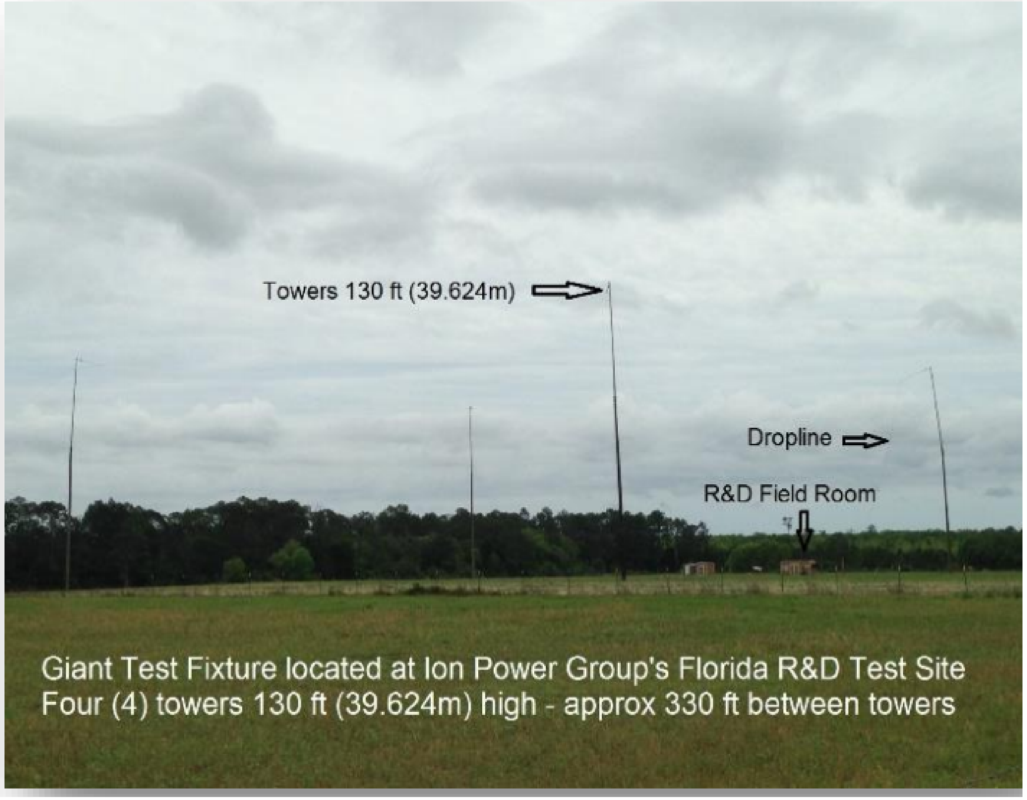
Towers 130 ft (39.624m)
Dropline
R&D Field Room

Giant Test Fixture located at Ion Power Group's Florida R&D Test Site
Four (4) towers 130 ft (39.624m) high - approx 330 ft between towers

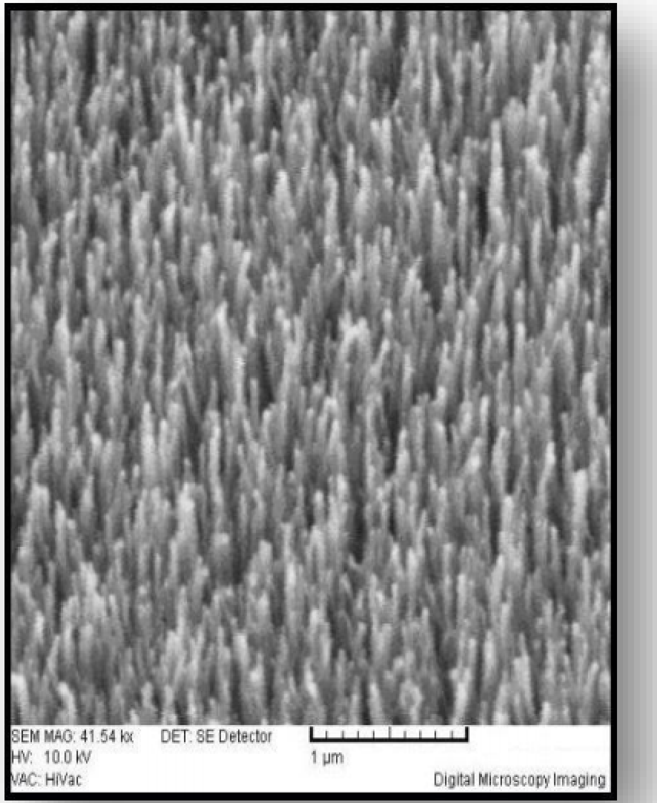
SEM MAG: 41.54 kx
DET: SE Detector
1 µm
HV: 10.0 kV
VAC: HiVac
Digital Microscopy Imaging

Ion Power Group’s new method of coupling to atmospheric electricity is vastly different from all previous techniques by virtue of our patented breakthrough revealing that carbon nanomaterials such as **Graphite (and Graphene) microscopic shown at left, macroscopic shown at right are significantly more effective at coupling to airborne charge carriers (ions) than metal.** The use of carbon based nanomaterials distinguishes Ion Power Group from all other researchers.

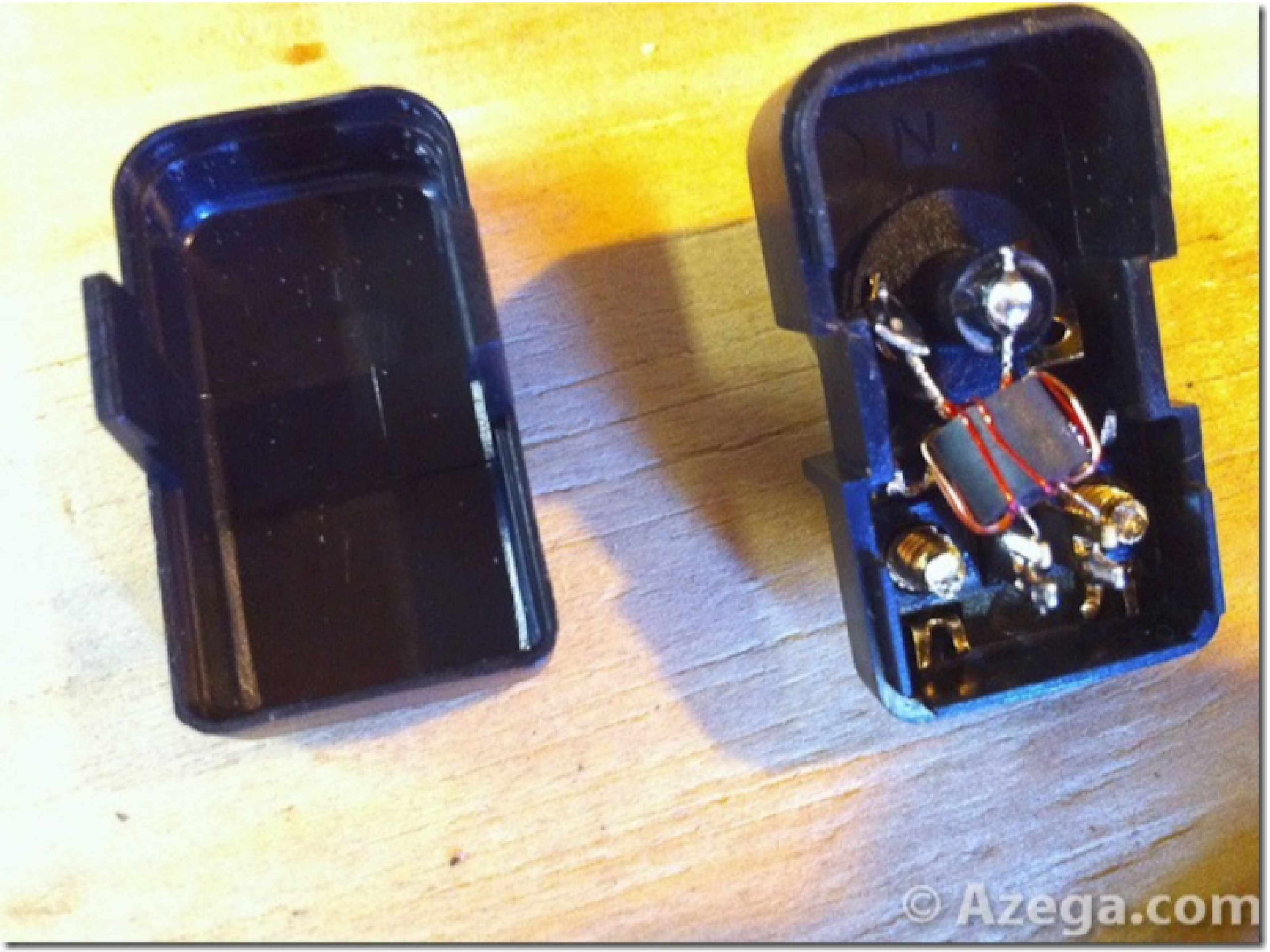

I cut a hole in the side of the balun then cut the end off a piece of coax cable I had and soldered it to the contacts of the coax connection.

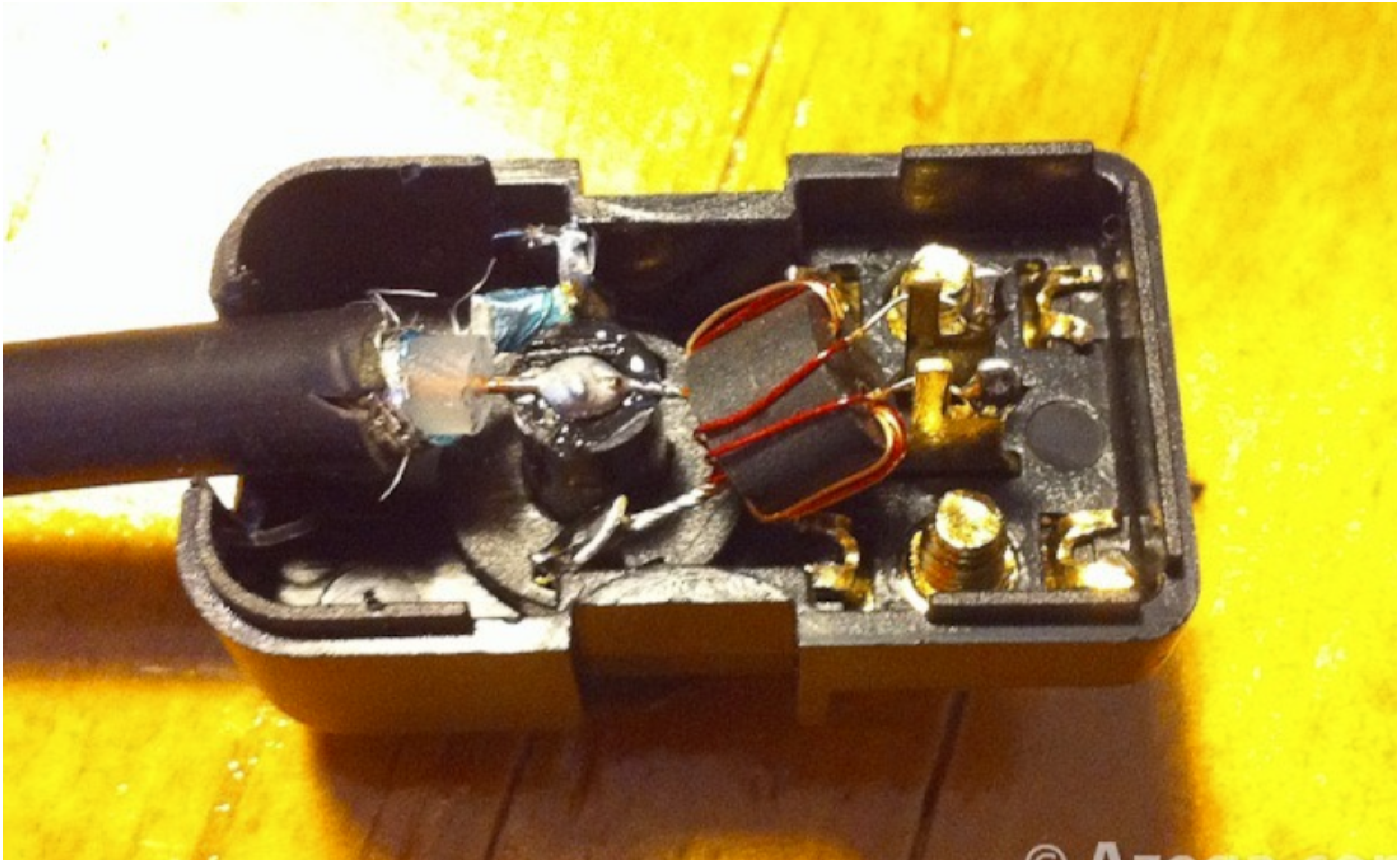

EXAMPLE: W1JR (Reisert) Cross-Winding Method

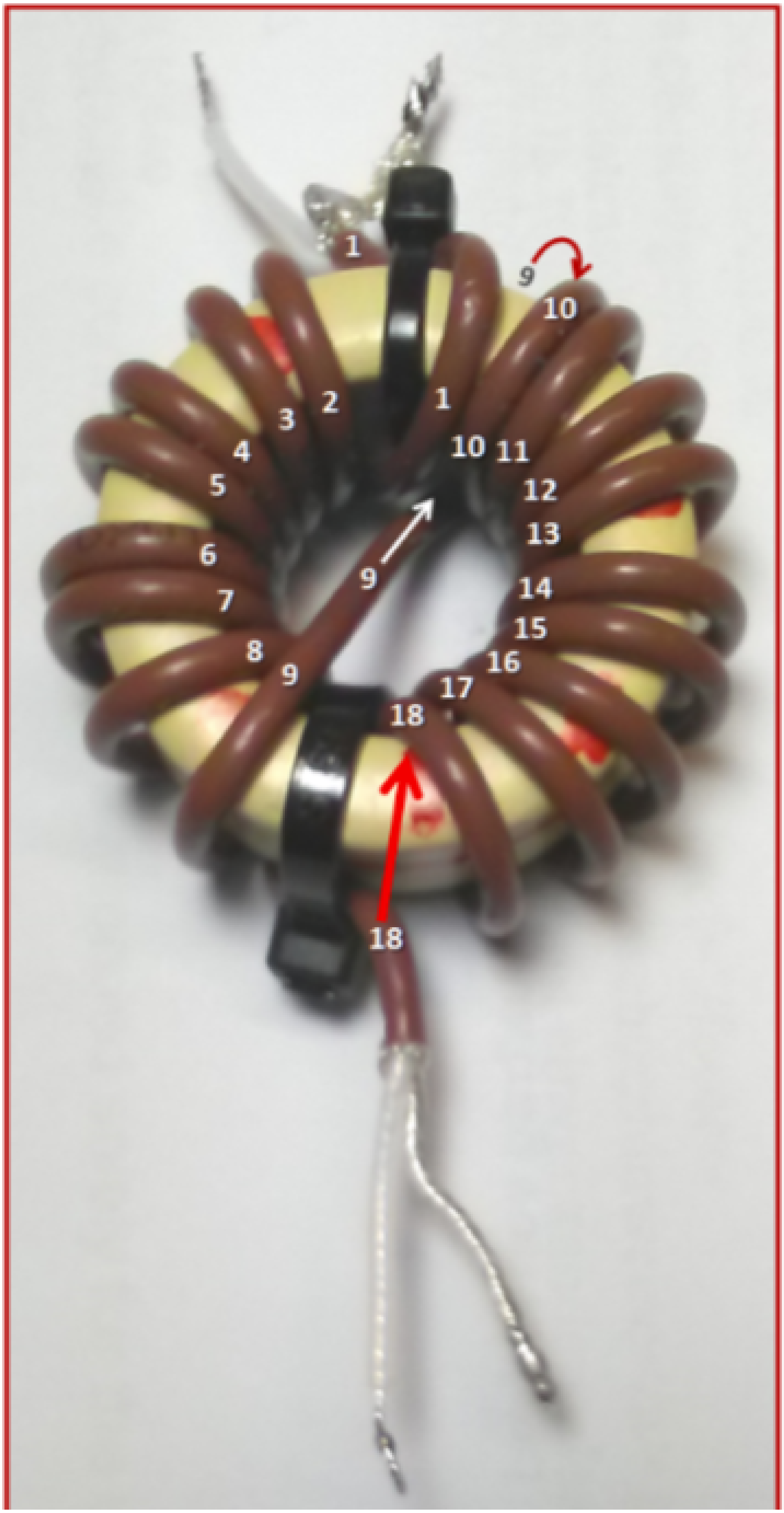

1
9
10
1
2
3
4
5
10
11
12
13
6
7
9
14
15
8
9
16
17
18
18

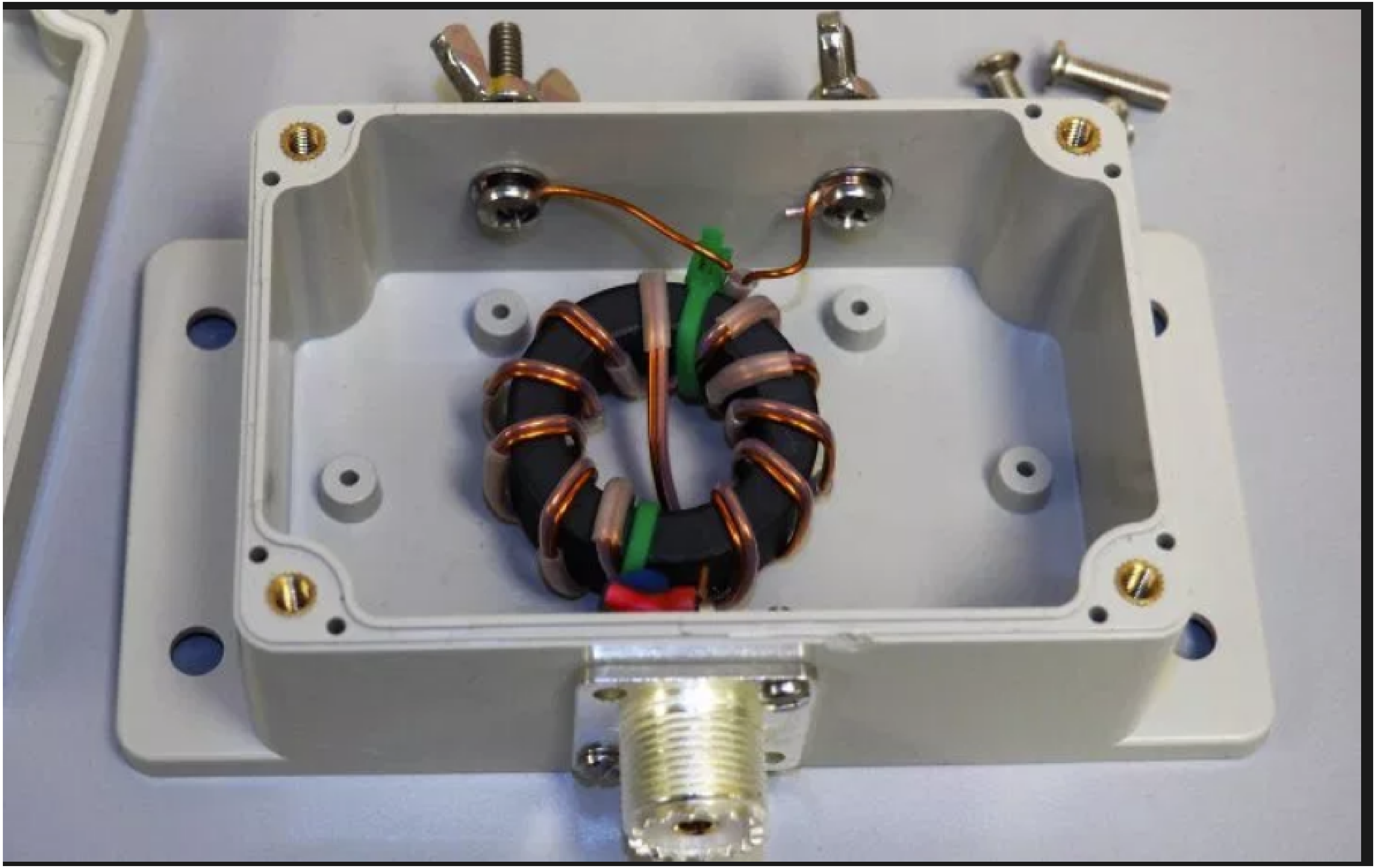

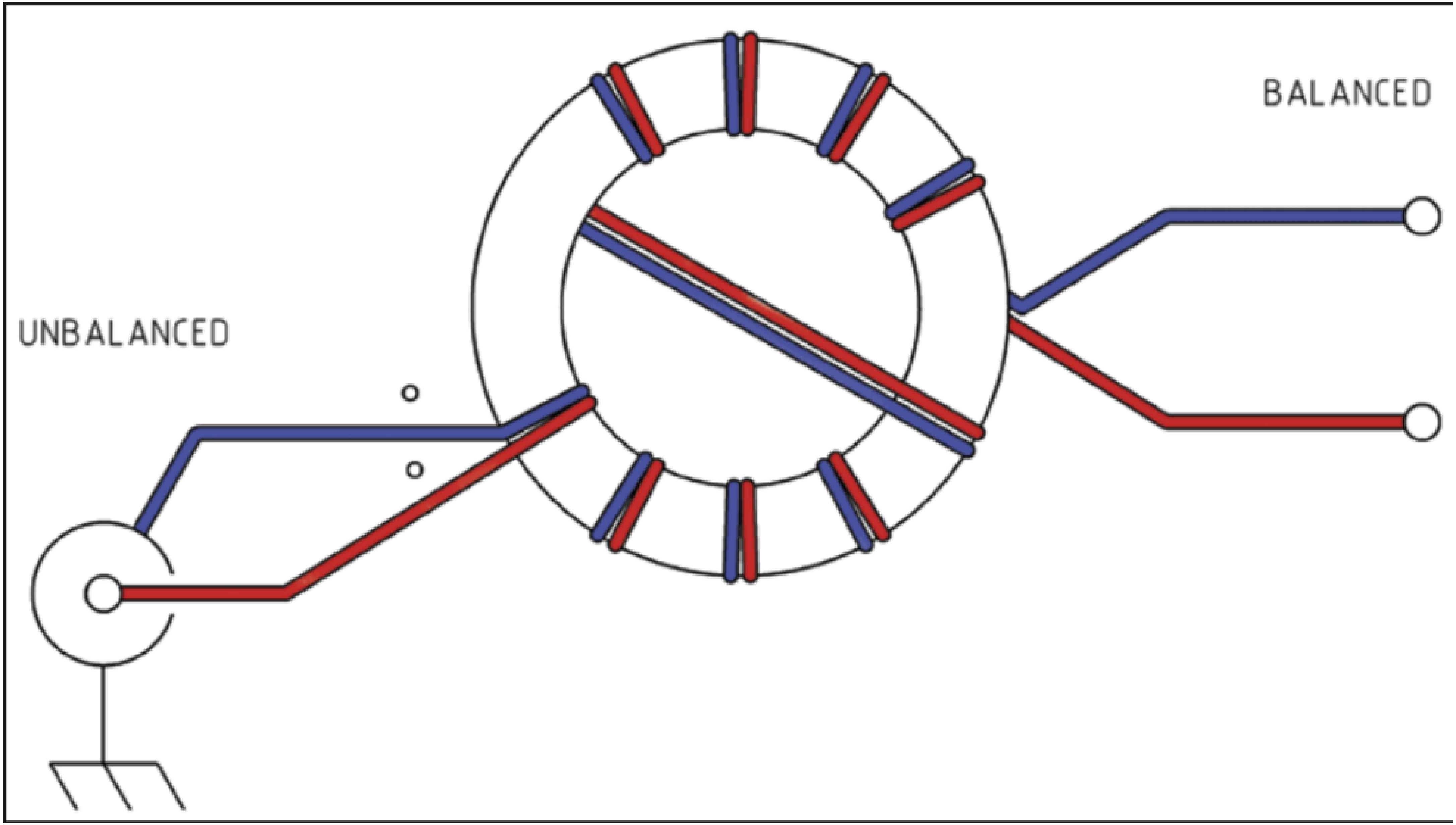
BALANCED
UNBALANCED

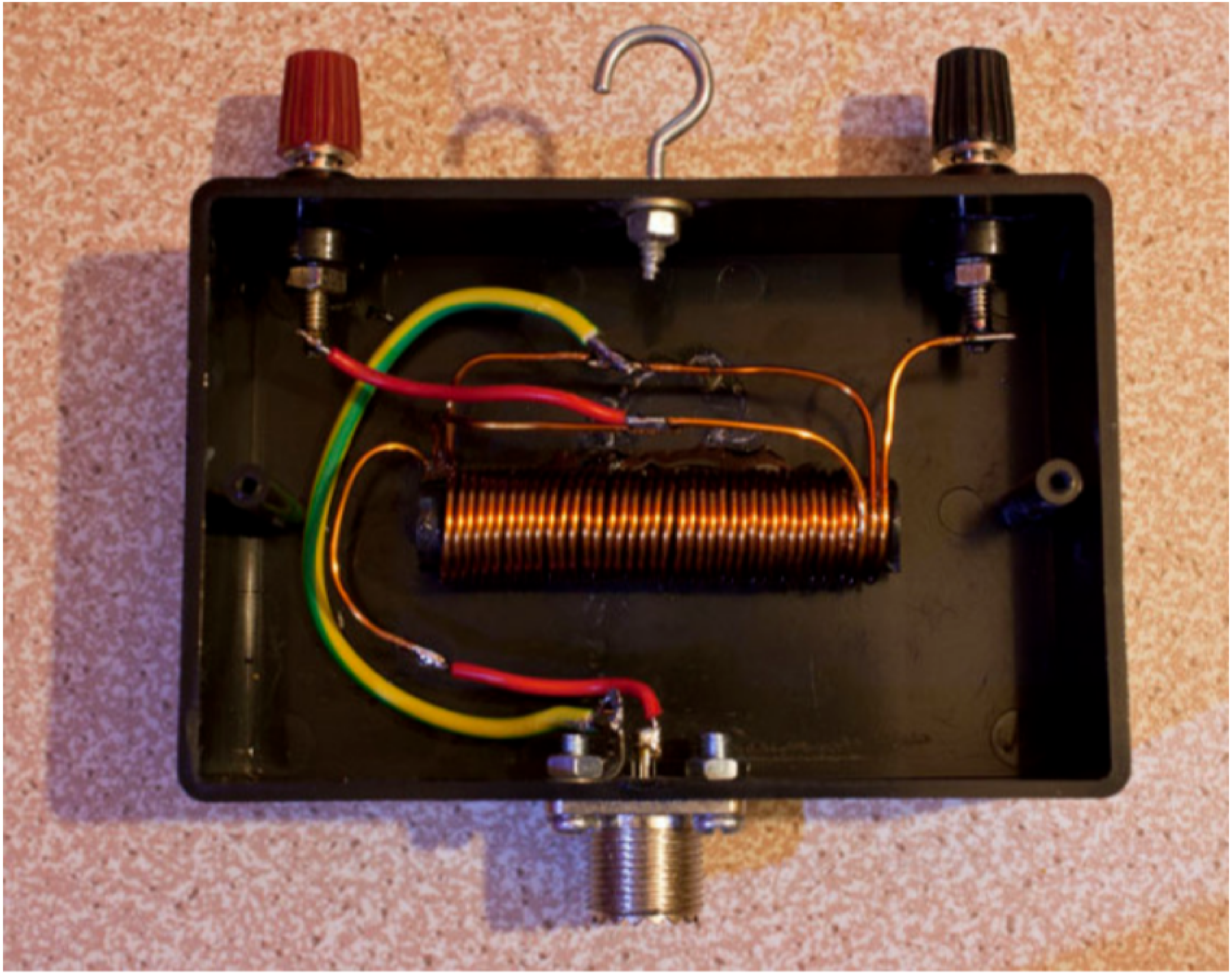

Balun 1: 1 for HF

1:1 Voltage balun ("Ruthroff")

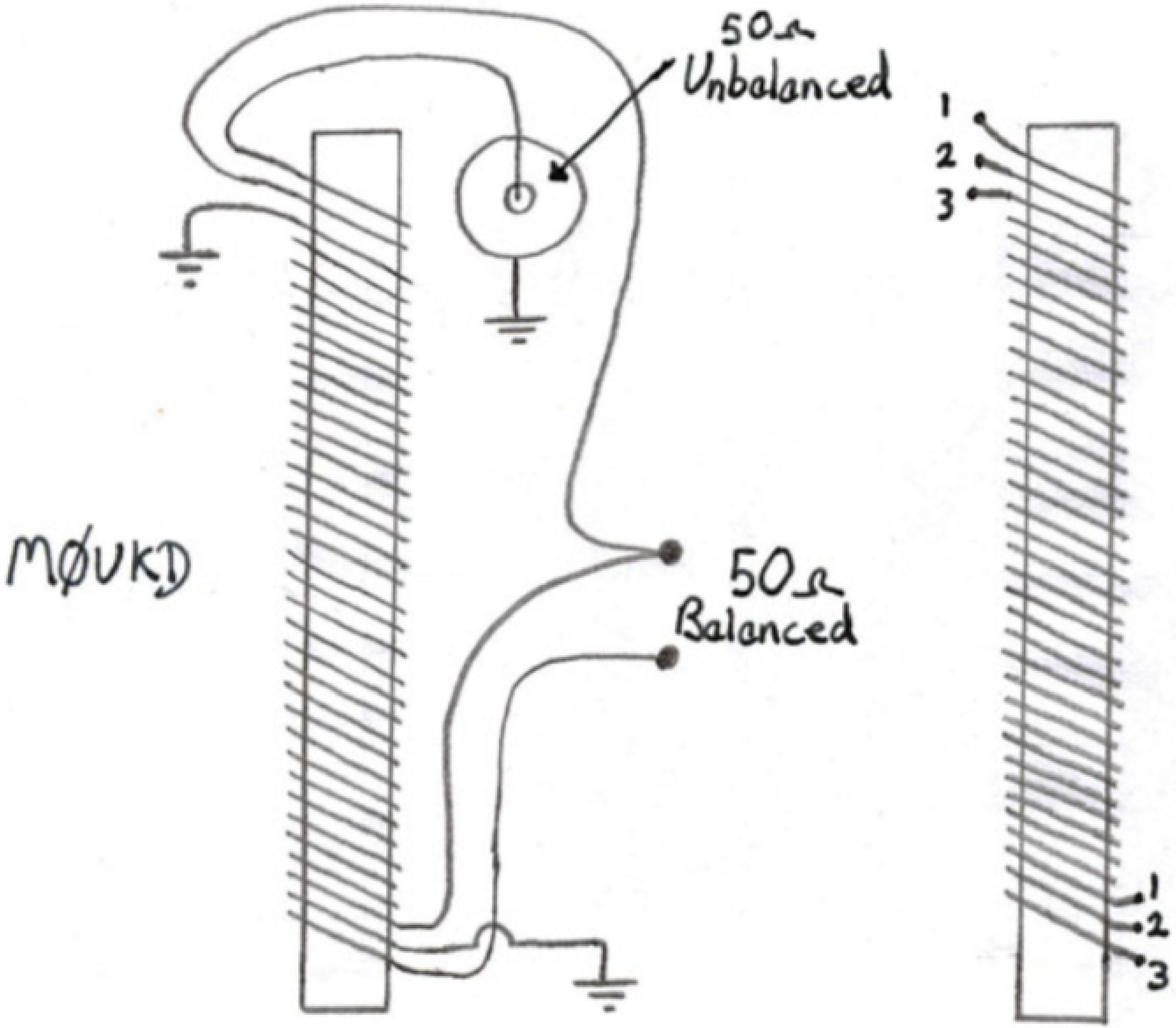
50Ω
Unbalanced
MØUKD
50Ω
Balanced
1
2
3
1
2
3

10 - 15 trifiler turns on ferrite rod.

1:1 ferrite rod voltage balun

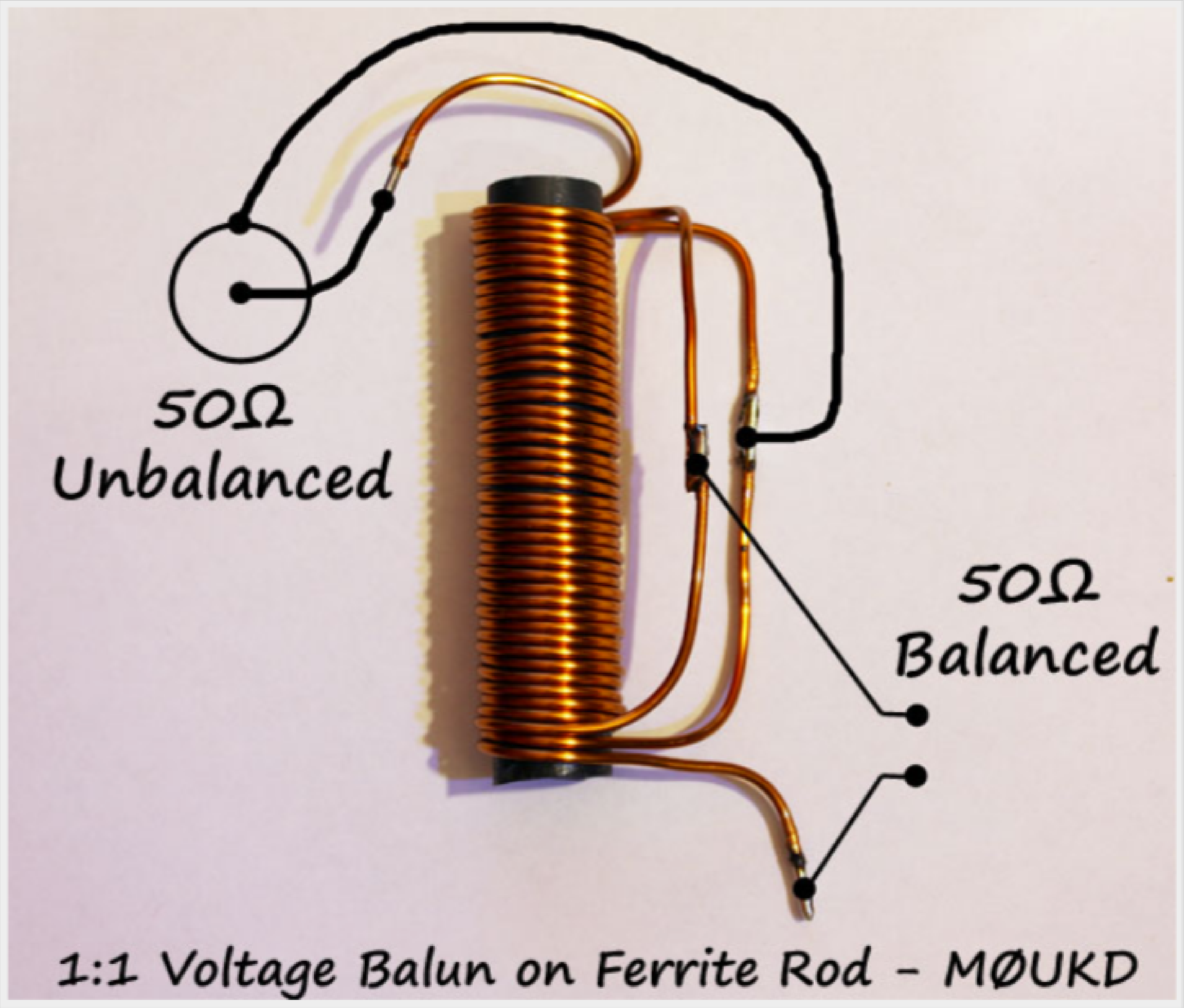
50Ω
Unbalanced
50Ω
Balanced
1:1 Voltage Balun on Ferrite Rod - MØUKD

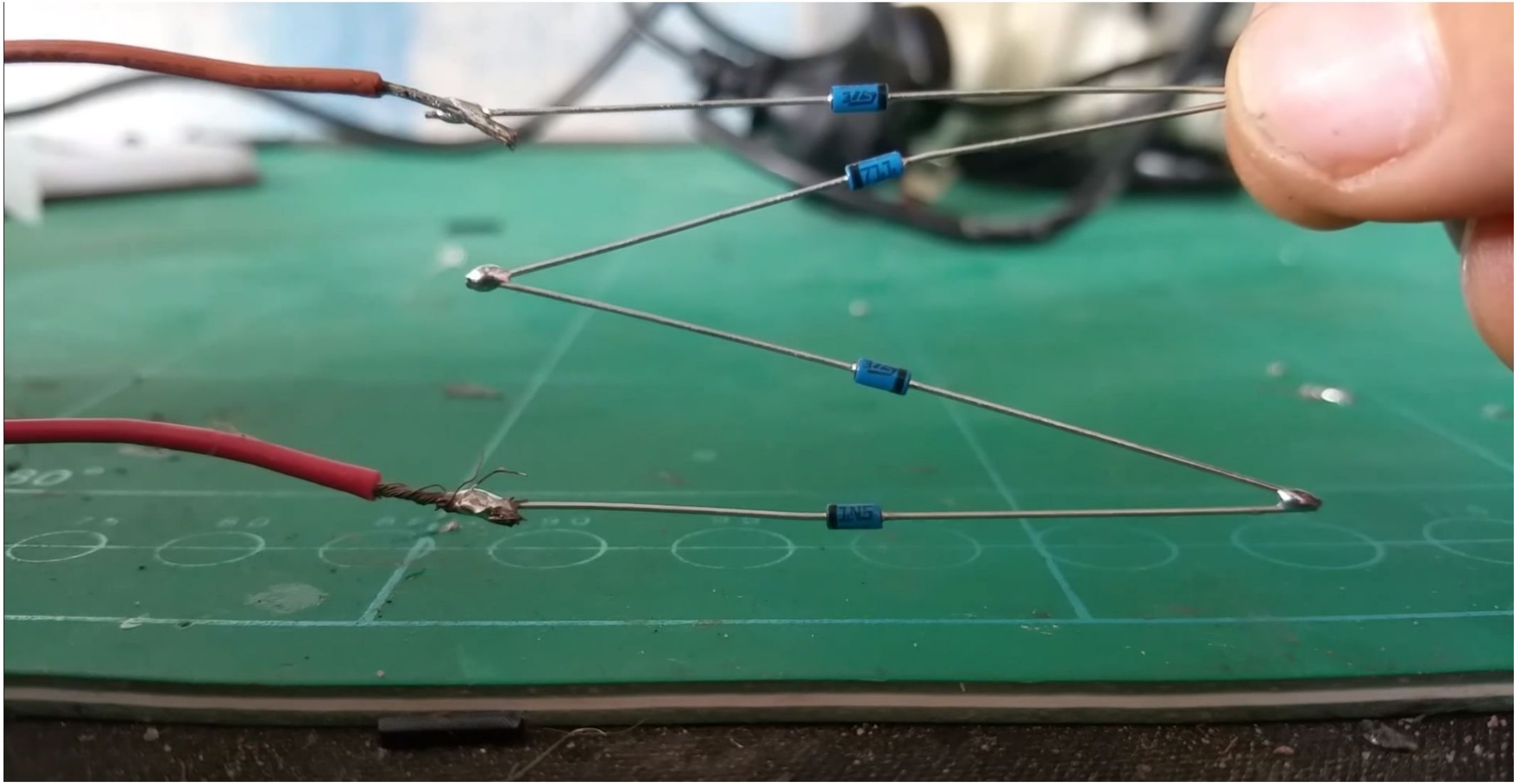

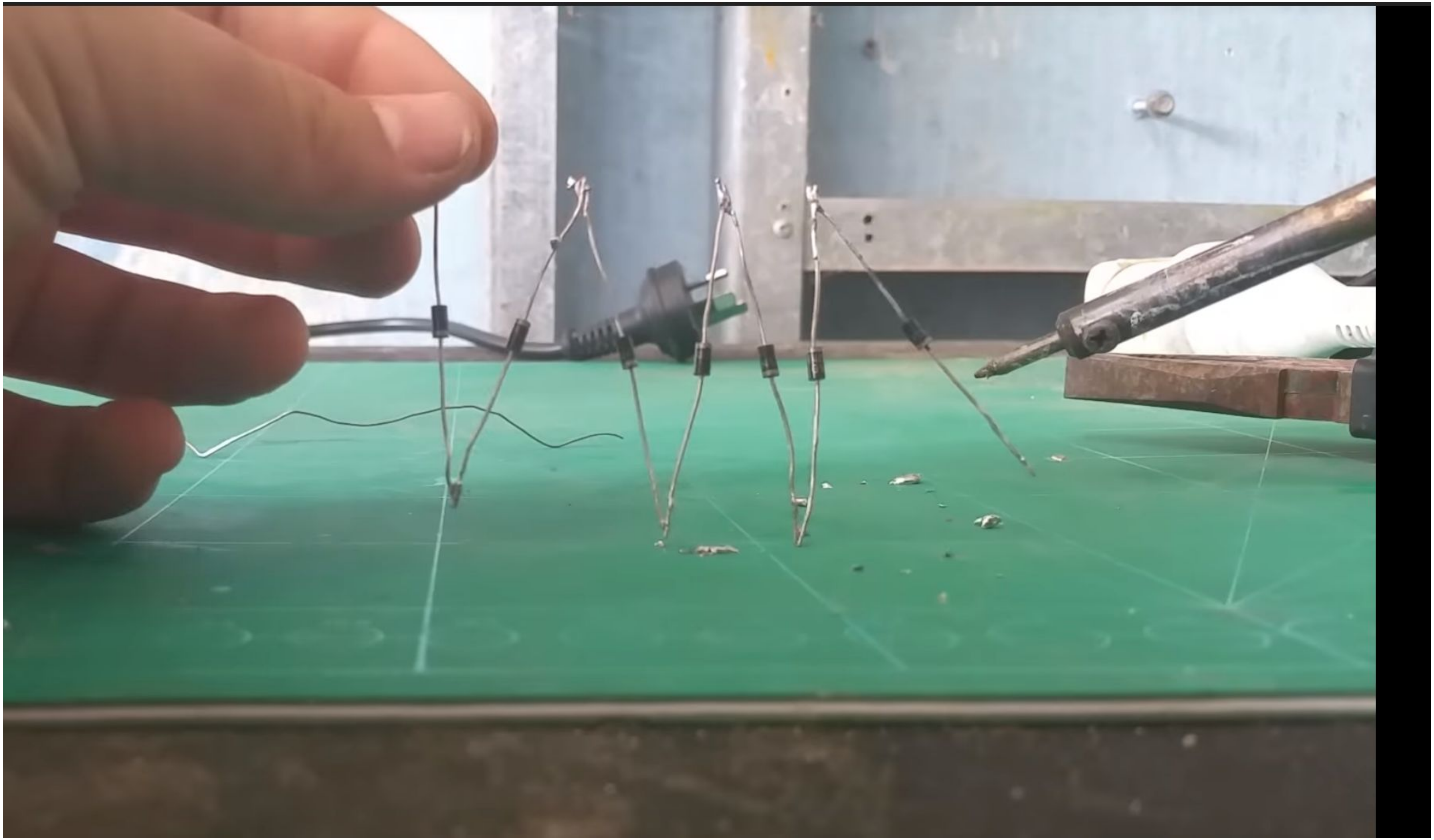

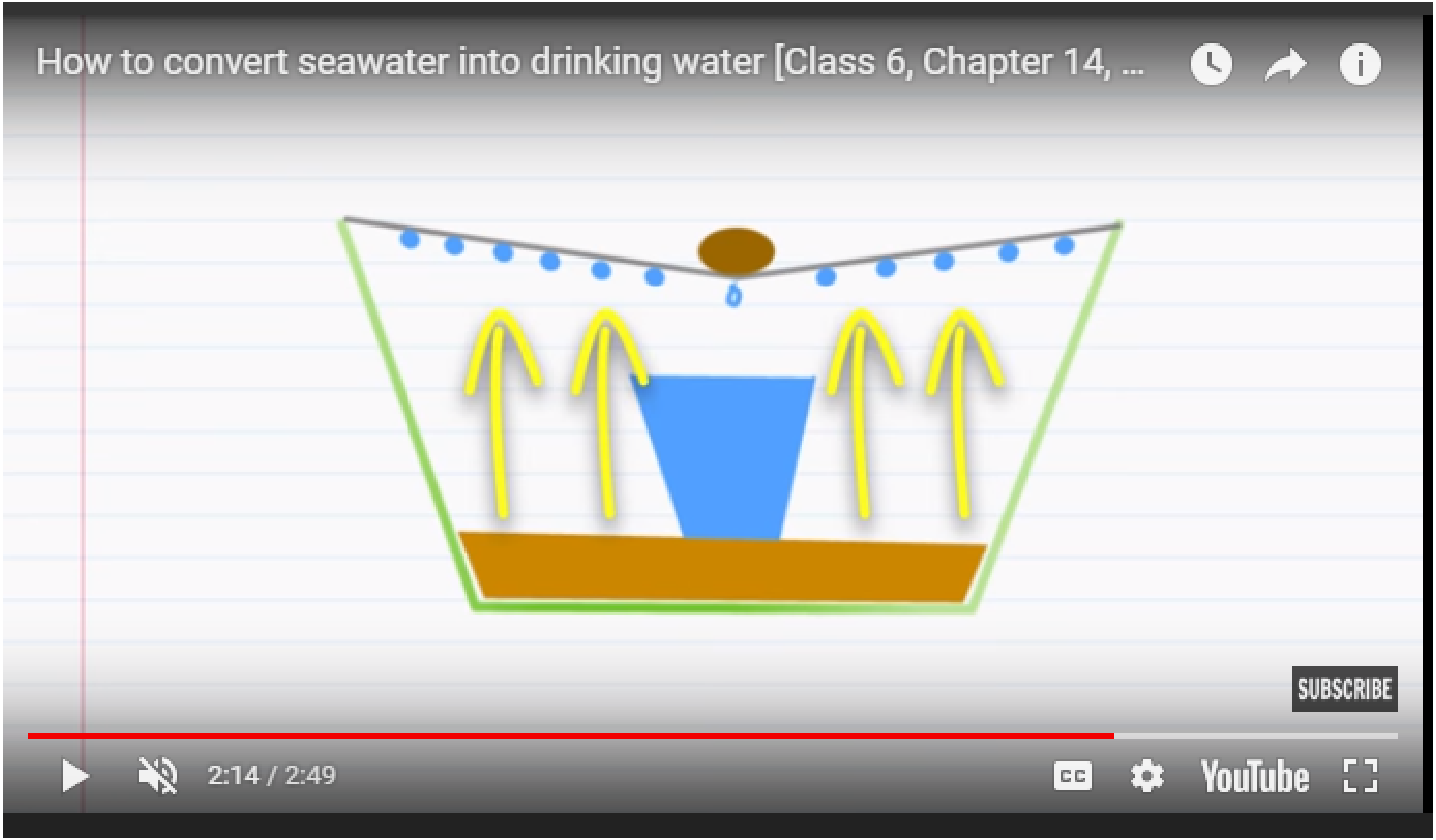

How to convert seawater into drinking water [Class 6, Chapter 14, ...

SUBSCRIBE
2:14 / 2:49
YouTube

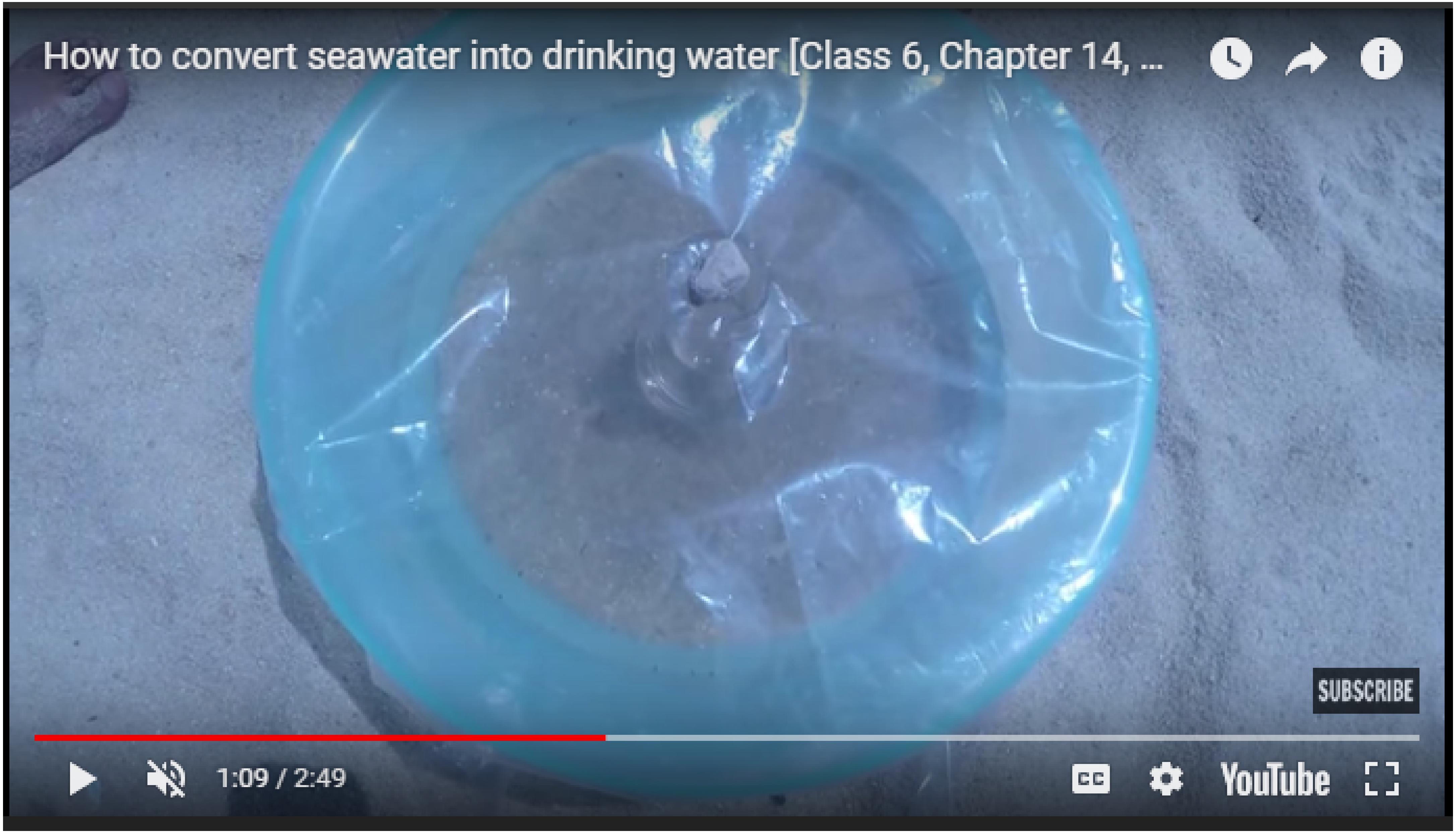
How to convert seawater into drinking water [Class 6, Chapter 14, ...
SUBSCRIBE
1:09 / 2:49
YouTube

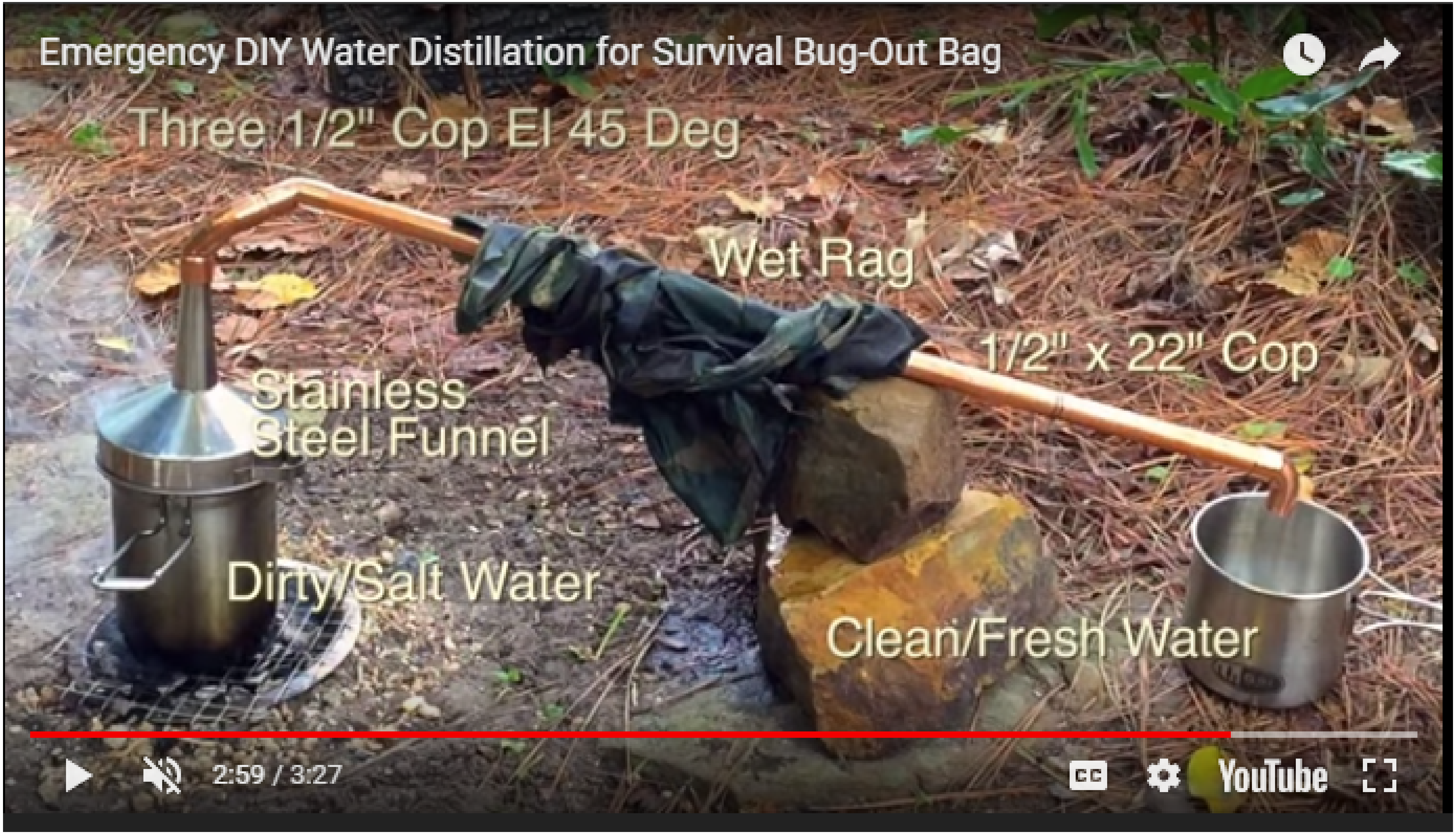

Emergency DIY Water Distillation for Survival Bug-Out Bag

Three 1/2" Cop El 45 Deg
Wet Rag
1/2" x 22" Cop
Stainless Steel Funnel
Dirty/Salt Water
Clean/Fresh Water
2:59 / 3:27
YouTube

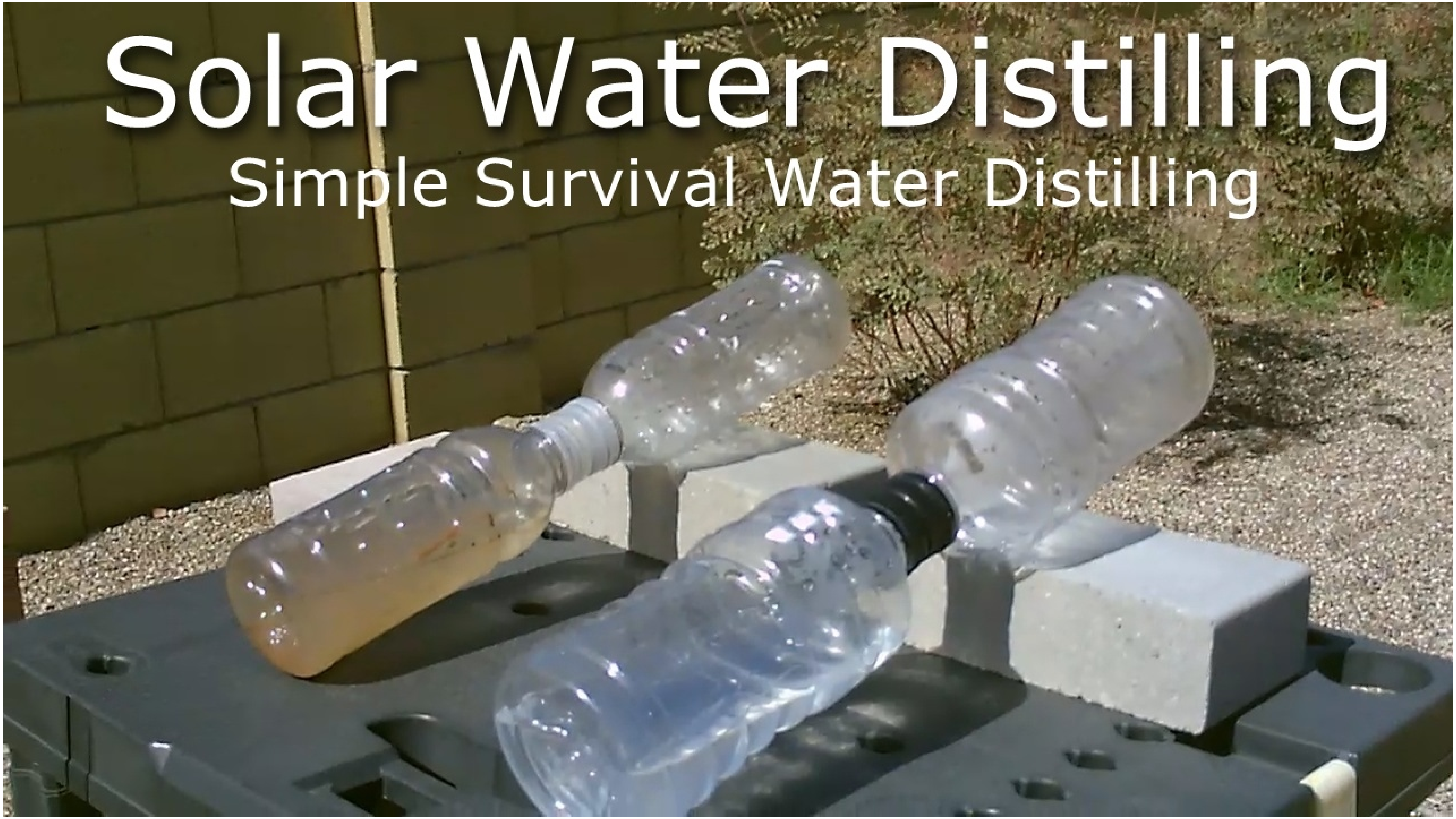
Solar Water Distilling
Simple Survival Water Distilling

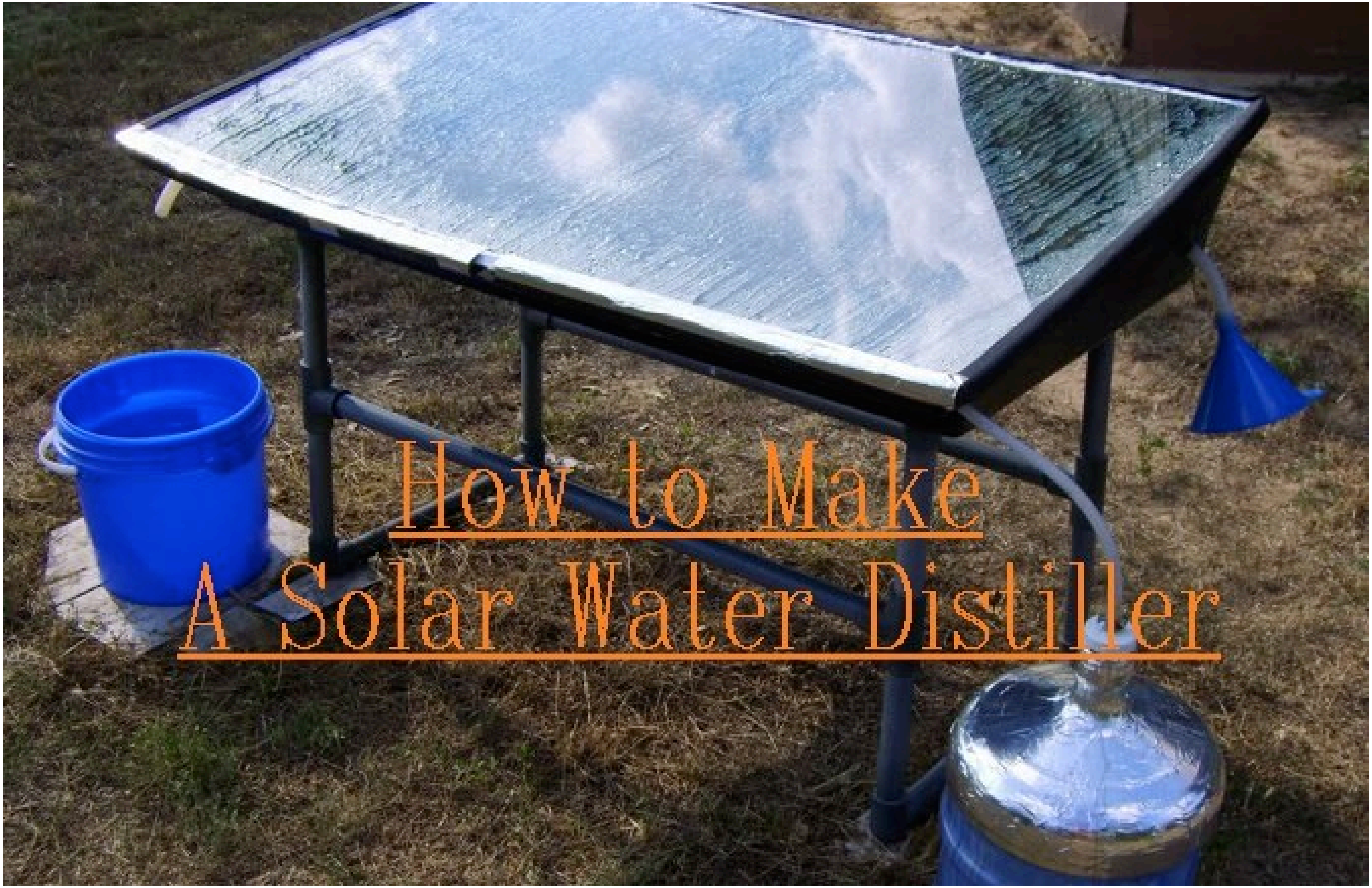

How to Make A Solar Water Distiller

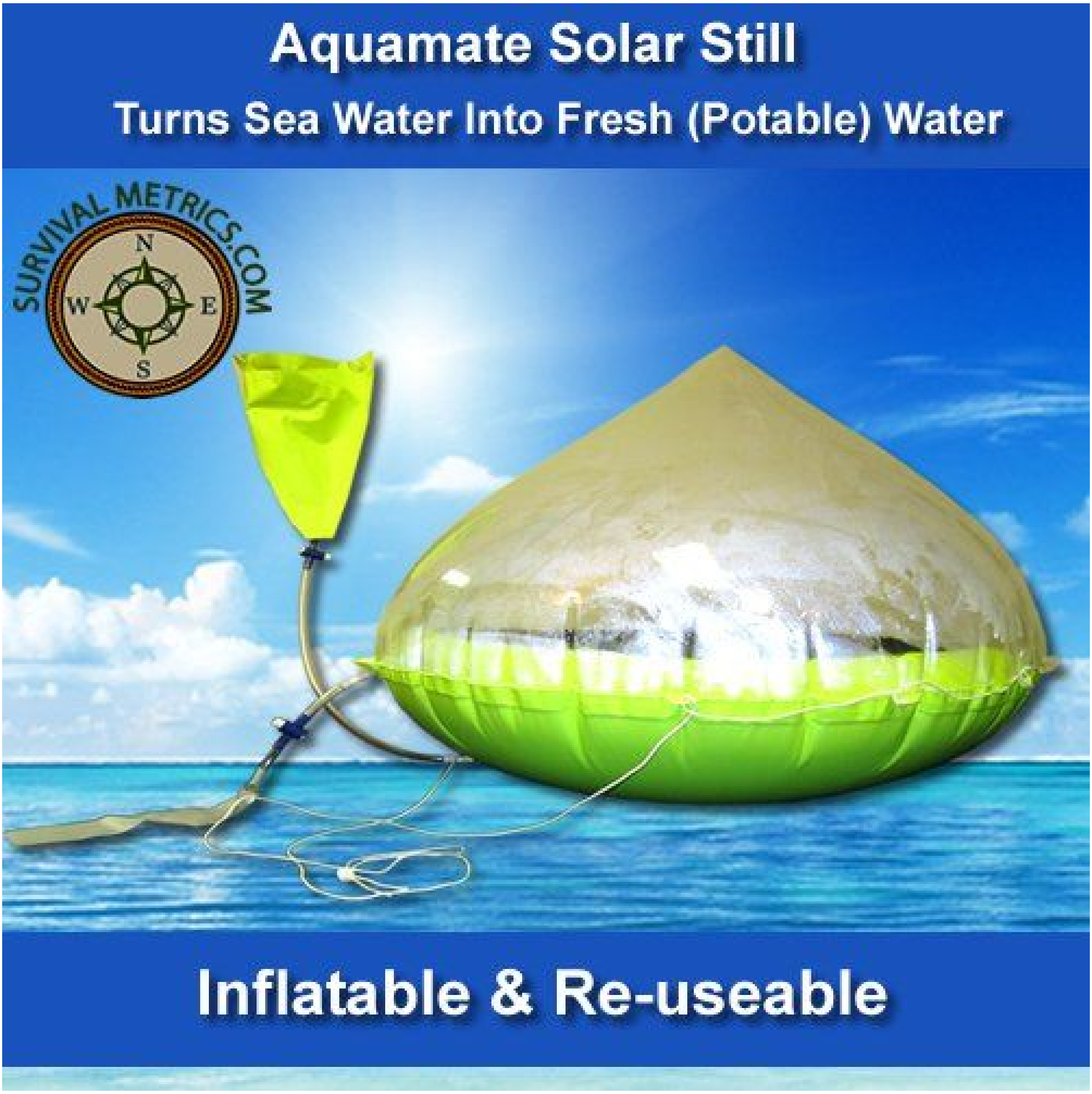

Aquamate Solar Still

Turns Sea Water Into Fresh (Potable) Water

Inflatable & Re-useable

SURVIVAL METRICS.COM
N
W
E
S

Solar Still

Turns sea water or impure water into drinking water

1. Remove from pack and retain pack to top up salt water in still.
2. Lay out Solar Still and inflate valve Ⓐ on floatation ring by mouth or by liferaft or dinghy pump.
3. Inflate clear dome to obtain shape using valve Ⓔ in base. Do not over inflate. It may be necessary to manually regulate dome pressure during high temperatures.
4. Attach tapered bag Ⓑ and attachment cord Ⓒ to boat or liferaft.
5. Pour 5 litres of sea water into tapered bag Ⓑ using pack which holds 5 litres (approx). Close valve Ⓕ after filling.
6. Ensure both valves Ⓖ are open in reservoir feed tube and drinkable water will start to collect in reservoir Ⓓ. Air may need to be squeezed out to assist collection of water.
7. Ensure the Solar Still is kept in direct sunlight whenever possible.
8. Re-fill Solar Still and inflate/deflate dome as required.
9. Daily – Agitate Solar Still to dampen fabric.
10. To drain unwanted water open valve Ⓔ in base.
11. Shelter Solar Still during bad weather.

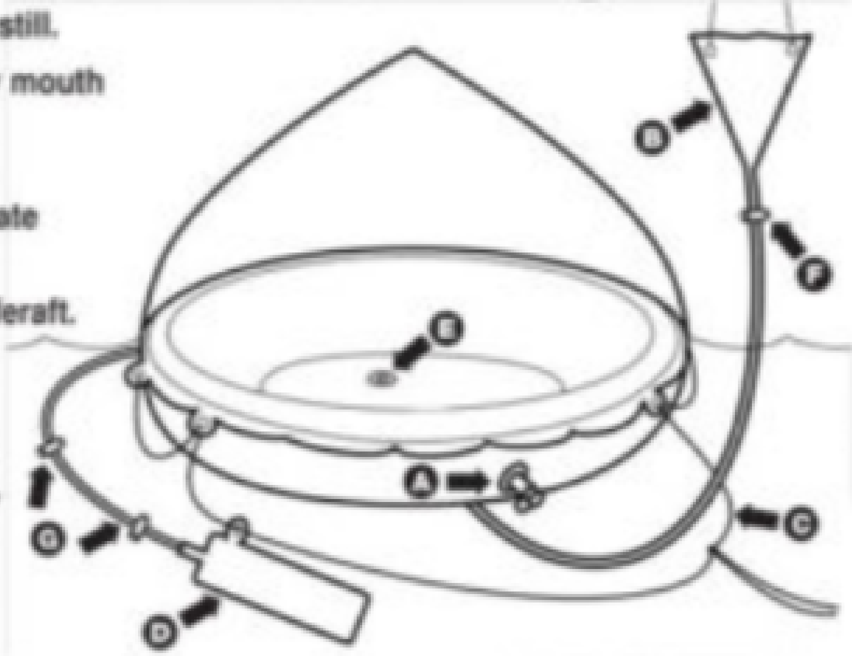

IMPORTANT
IF INNER BLACK MATERIAL SALTS UP FILL WITH SEA WATER AND AGITATE TO CLEAR SALT CRYSTALS AND DRAIN AWAY. PROCESS DRINKING WATER AS INSTRUCTIONS UNTIL SALT IS NO LONGER PRESENT. IN NO CIRCUMSTANCES DRINK WATER THAT TASTES OF SALT.

5 060259 930188

AQUAMATE PRODUCTS

NATO Stock No. 4610-66-144-2646

P.O. Box 6032 Dunmow CM6 3AS U.K. Tel: +44 (0)1371 830 216 Fax: 831 733 e-mail: aquamatesales@aol.com

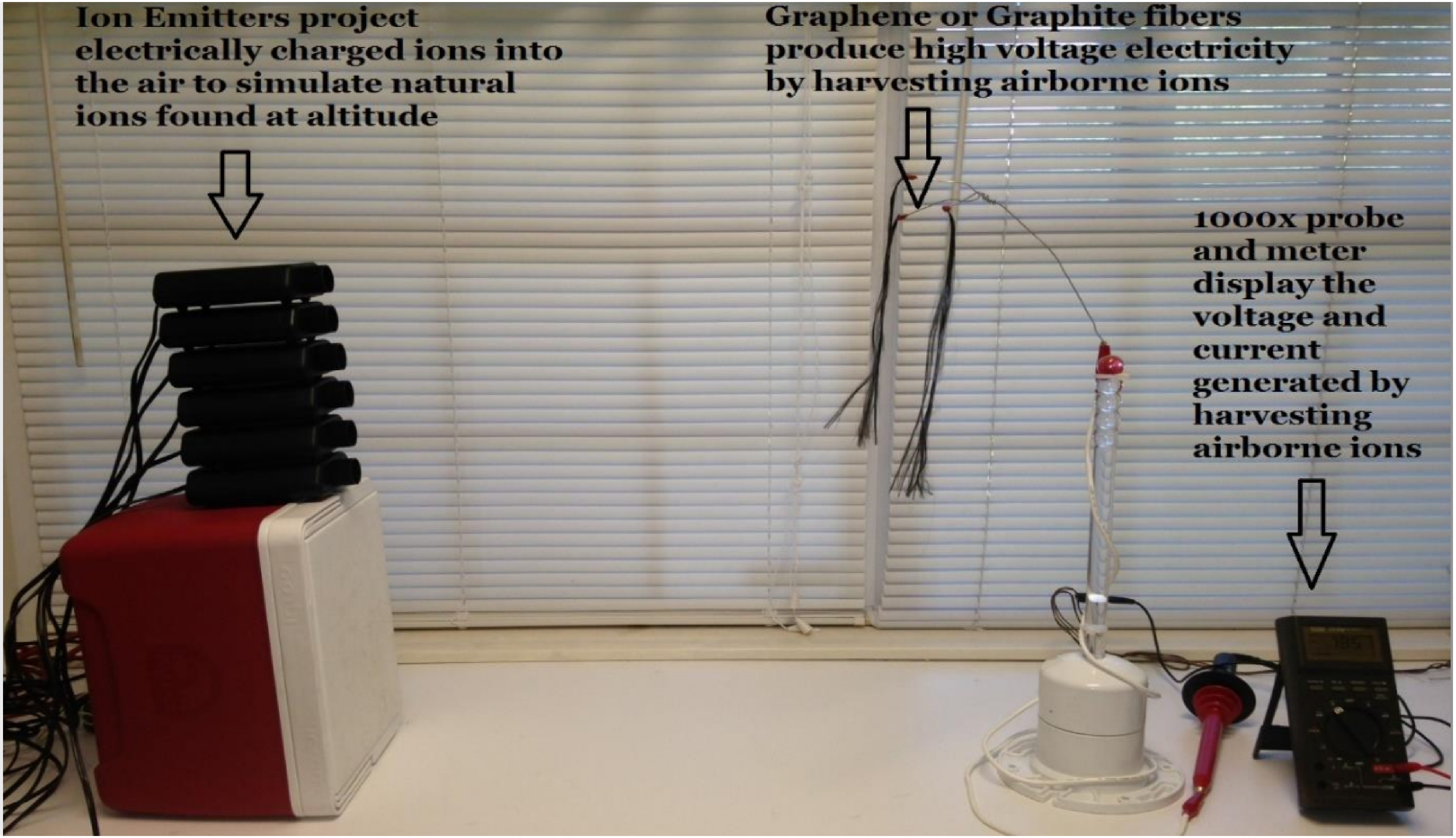

Ion Emitters project electrically charged ions into the air to simulate natural ions found at altitude

Graphene or Graphite fibers produce high voltage electricity by harvesting airborne ions

1000x probe and meter display the voltage and current generated by harvesting airborne ions

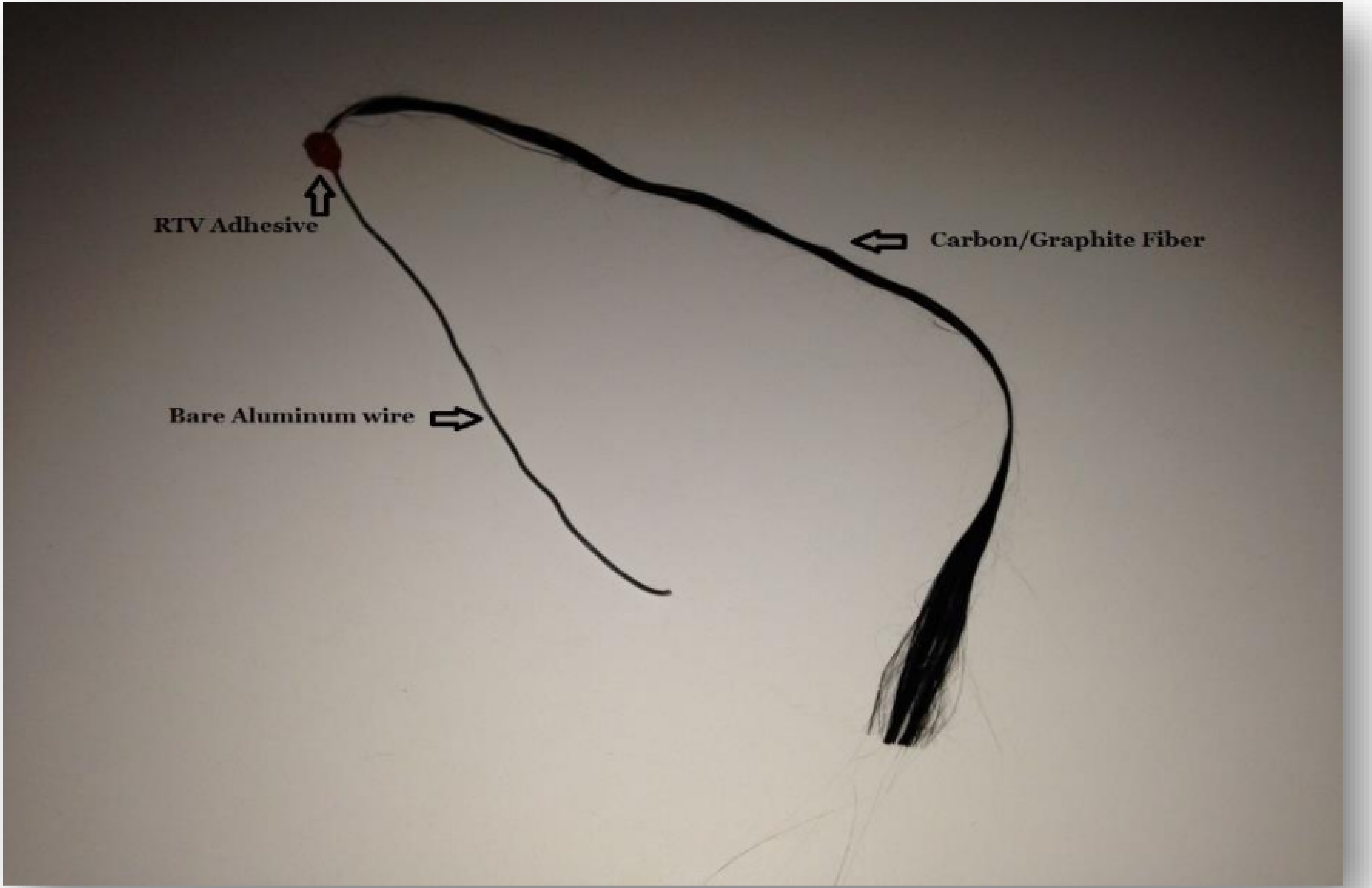
RTV Adhesive
Carbon/Graphite Fiber
Bare Aluminum wire

Image of a patented Ion Collector made of carbon/graphite weighing approximately 1 ounce.

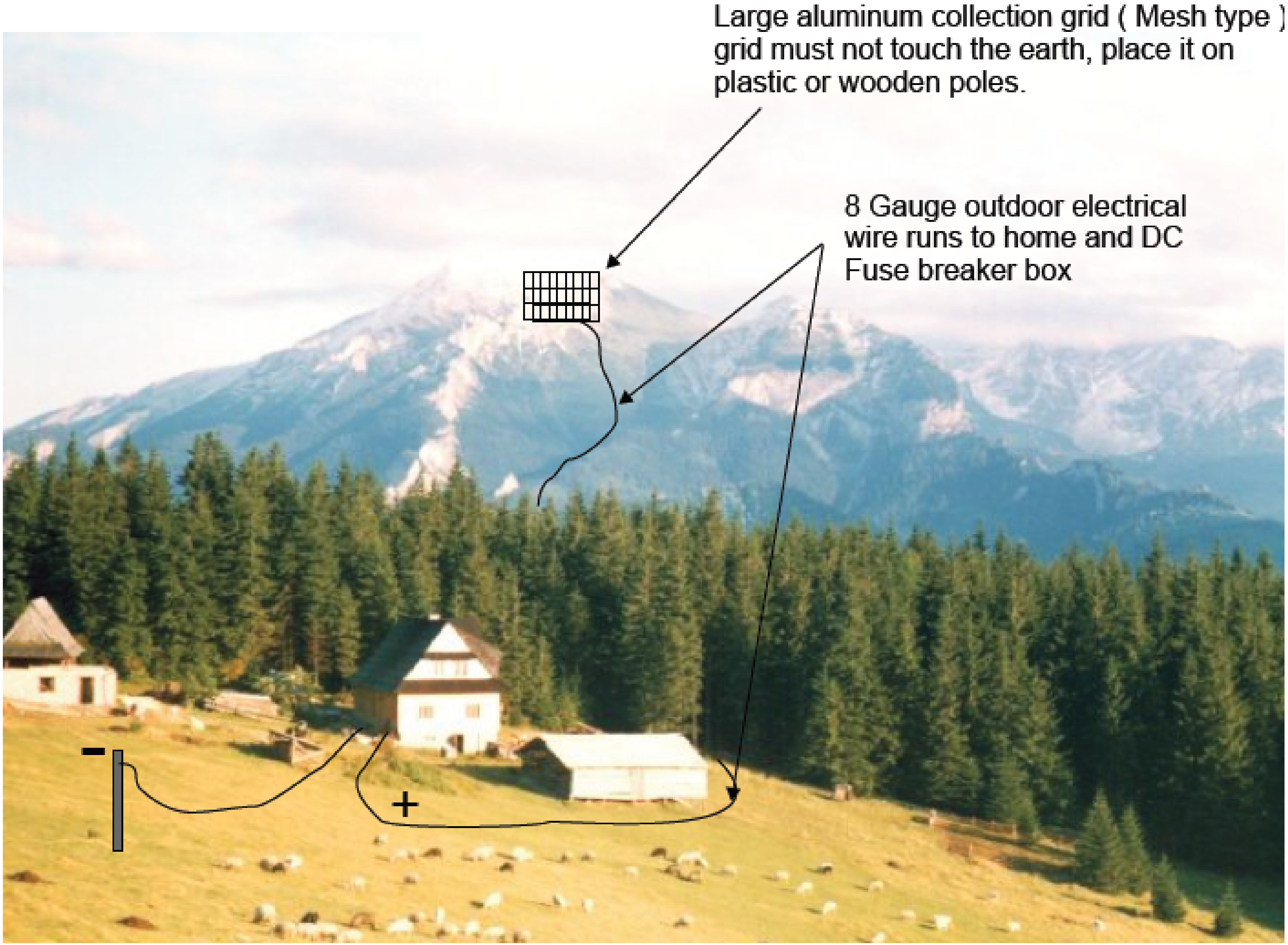
Large aluminum collection grid (Mesh type)
grid must not touch the earth, place it on plastic or wooden poles.
8 Gauge outdoor electrical wire runs to home and DC Fuse breaker box
-
+

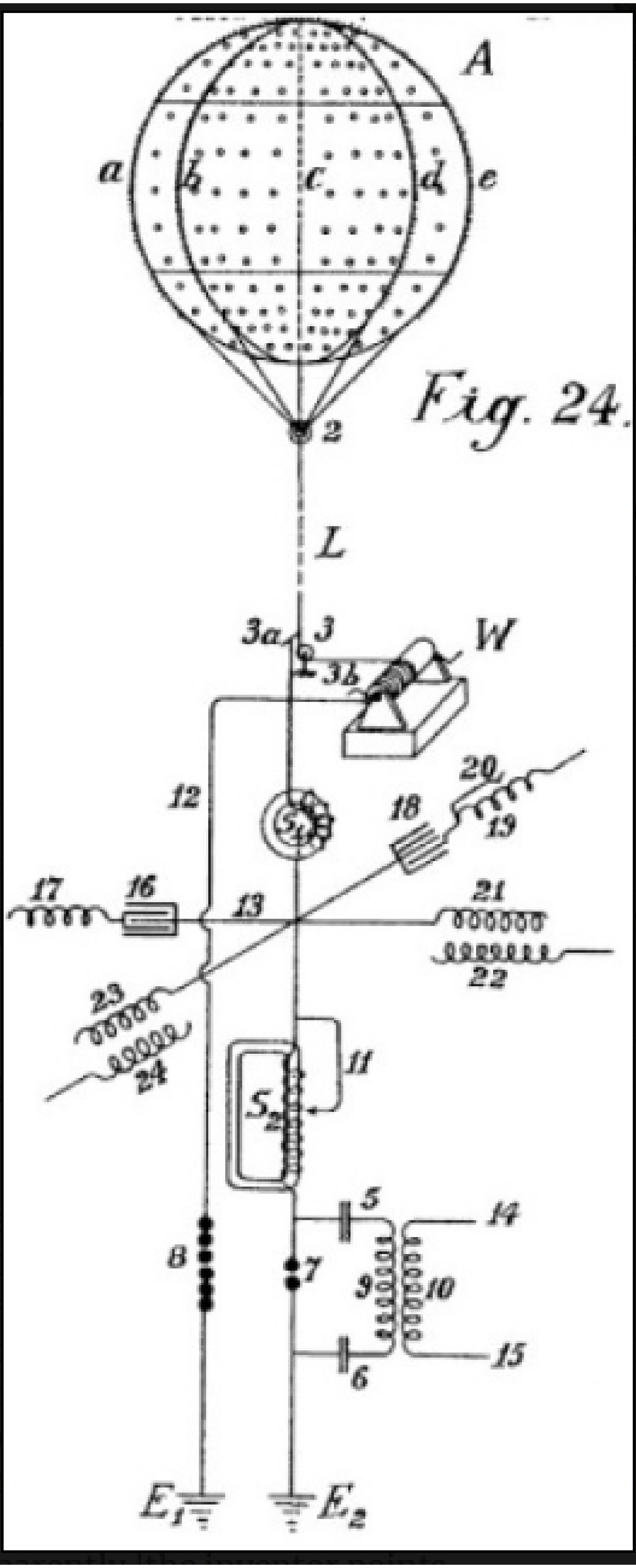
A
a
b
c
d
e
L
3a
3b
W
S_1
S_2
E_1
E_2

Fig. 24.

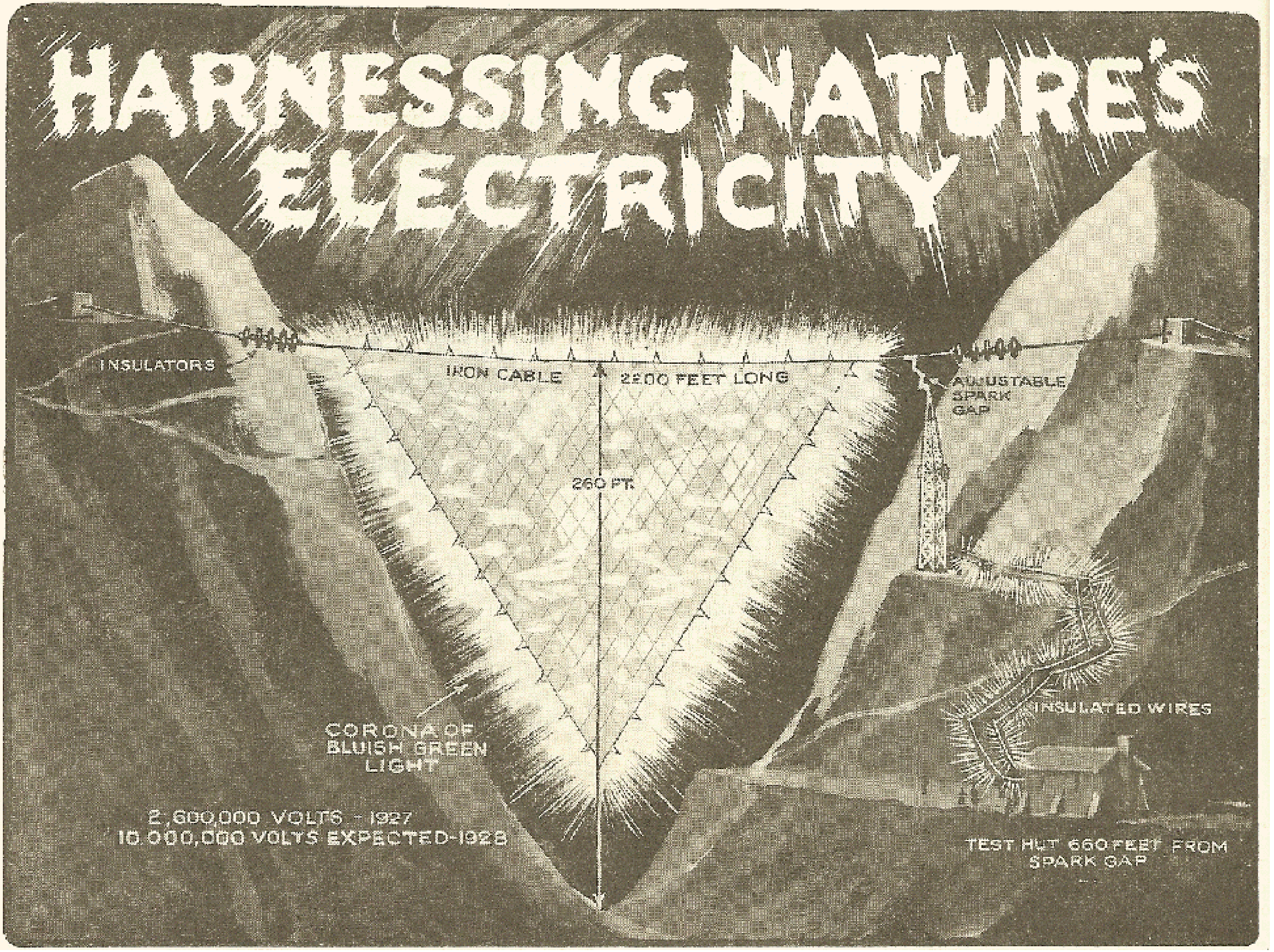
HARNESSING NATURE'S ELECTRICITY
INSULATORS
IRON CABLE
2200 FEET LONG
ADJUSTABLE SPARK GAP
260 FT.
CORONA OF BLUISH GREEN LIGHT
INSULATED WIRES
2,600,000 VOLTS - 1927
10,000,000 VOLTS EXPECTED-1928
TEST HUT 660 FEET FROM SPARK GAP

Remarkable European Experiments with Atmospheric Electrical Discharges with Potentials as High as 3,000,000 Volts

By HENRY TOWNSEND

elevation of 350 ft., and these students of natural electrical phenomena have found a very desirable location in the Alps, where they can suspend between one mountain and an adjacent one, a strong iron cable having a length of about 2,000 ft. This cable is about 250 feet above the intervening valley, and from it these daring engineers have suspended a coarsely woven wire net, which serves as an electrical capacity to gather the electricity from the atmosphere. As shown in the pictures, the wire net is supplied with numerous sharp points to aid in collecting the current from the air.

As the accompanying photographs of the actual apparatus and wire cable used last year clearly show, an adjustable spark gap of considerable length is provided. By adjusting this spark gap to various lengths, it is possible to judge the voltage of the discharge which leaps the gap at any moment. Mr. F. W. Peek, Jr., the well-known American worker in the realm of high voltage measurements, together with other engineers, have provided tabulated data and curves for various lengths of both needle and sphere type spark gaps. As one of the accompanying diagrams shows, it is a simple matter to calculate the voltage when a certain length of gap is used. The engineer first checks the length of the gap on the chart; he then follows a line horizontally from the gap length, to where it intersects with the angular line on the chart; and from the point of intersection he looks in a visual line downward to a place where the voltage is given. For needle spark gap measurements, the characteristic curve on the chart is practically a straight line, while for sphere gaps the characteristic curve on the voltage versus gap length, is a curved line. Those interested in high voltage measurements by means of the spark gap method can find the voltage-gap tables and charts in the *Standardization Rules* of the American Institute of Electrical Engineers. According to Mr. Peek's researches, the voltage per foot of atmospheric electrical discharges is about 100,000, while in laboratory measurements with A.C., transformer high potential discharges, the average voltage per foot of spark was found to be about 150,000 volts. The voltage of a lightning flash may

(*Continued on page* 156)

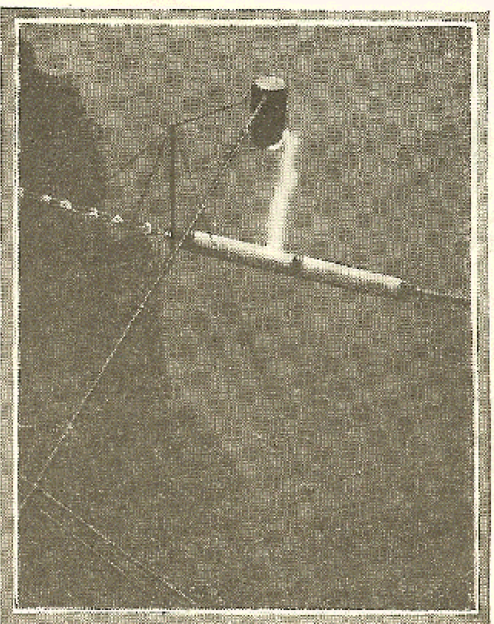

↑ Actual photo above shows 13 ft., heavy spark obtained from the collecting net in the Alps by the German scientists. The voltage is about 2,000,000. The spark occurred once per second for 30 minutes.

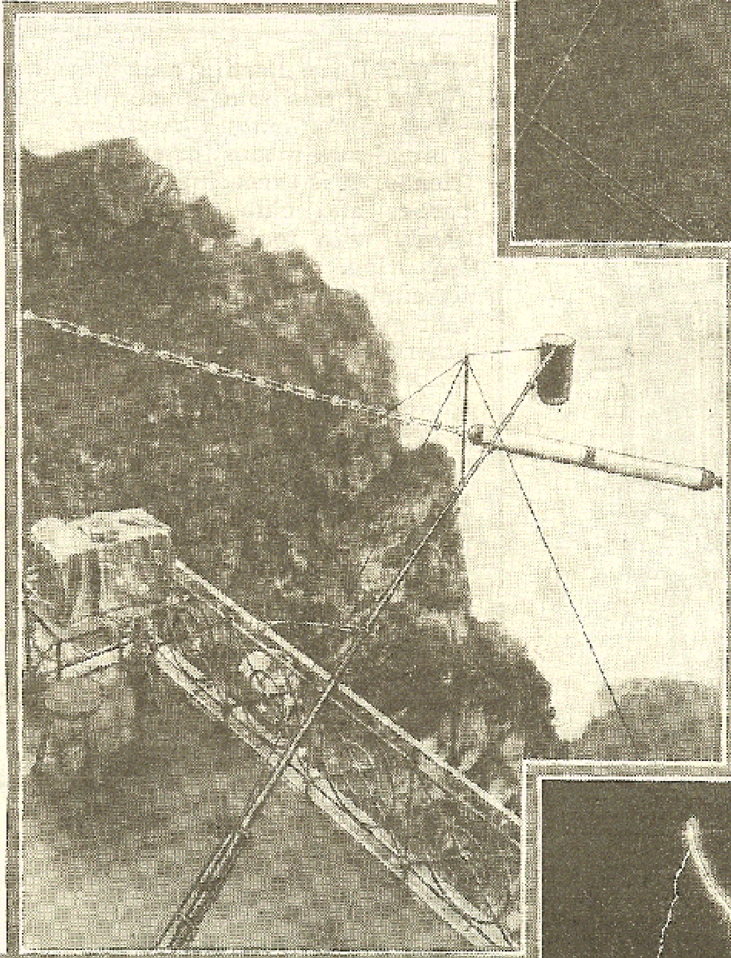

Photo, left, shows the adjustable spark gap used in the Alps. Notice the heavy electrode on the end of the adjustable arm to which the spark jumps.

←

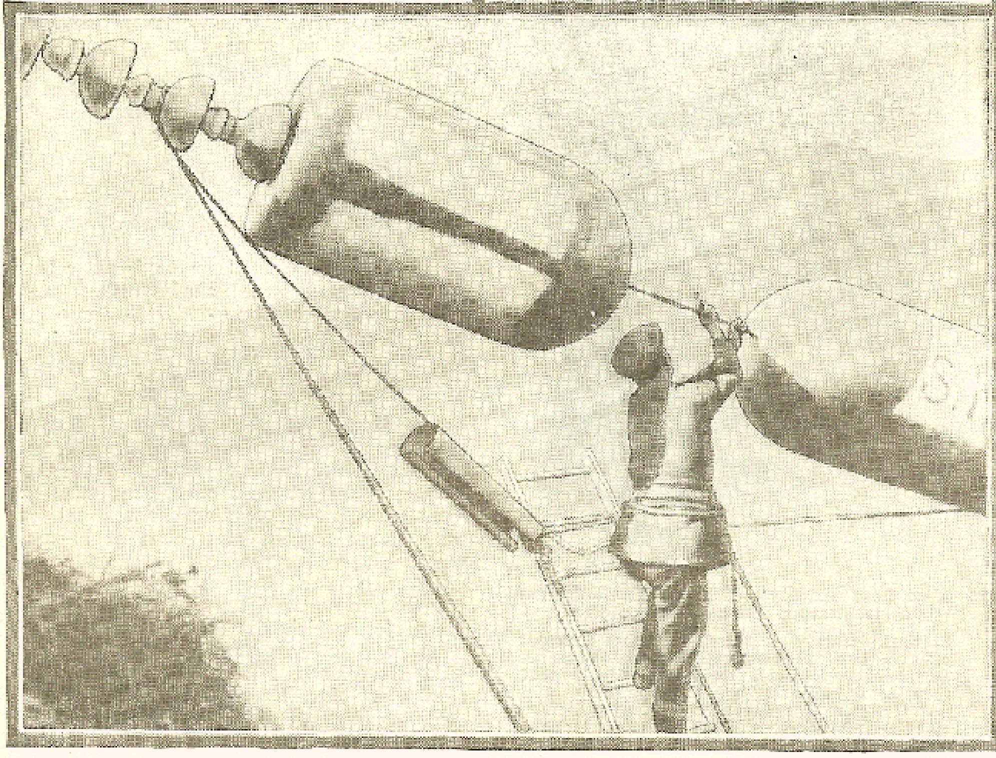

Actual photograph of the experimental "kite" used by the German experimenters in the Alps Mountains, for the purpose of accumulating high potential electrical discharges from the atmosphere. Note the size of the insulators.

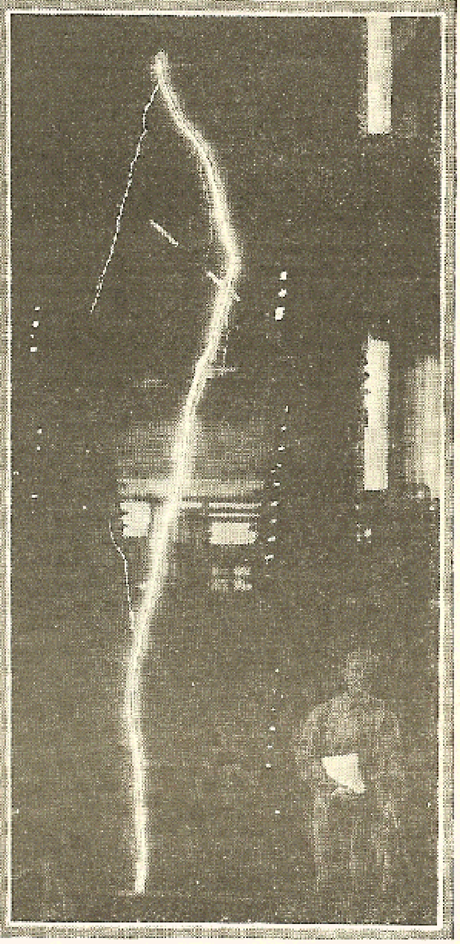

Below we see 3,000,000-volt artificial lightning stroke produced in G. E. Laboratory at Pittsfield, Mass. Note man. ↓

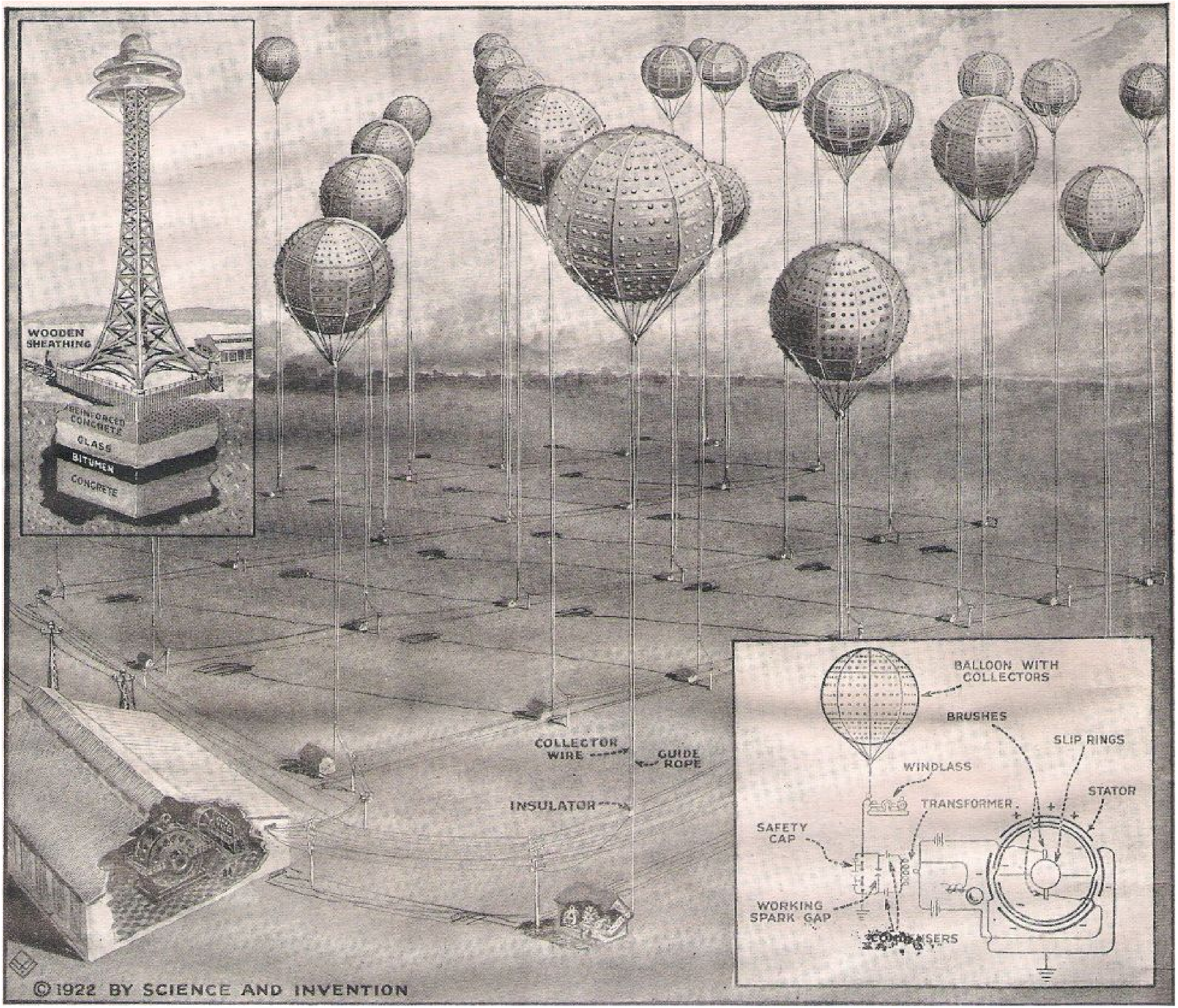
WOODEN SHEATHING
REINFORCED CONCRETE
GLASS
BITUMEN
CONCRETE
COLLECTOR WIRE
GUIDE ROPE
INSULATOR
BALLOON WITH COLLECTORS
BRUSHES
SLIP RINGS
WINDLASS
STATOR
TRANSFORMER
SAFETY CAP
WORKING SPARK GAP
CONDENSERS

©1922 BY SCIENCE AND INVENTION

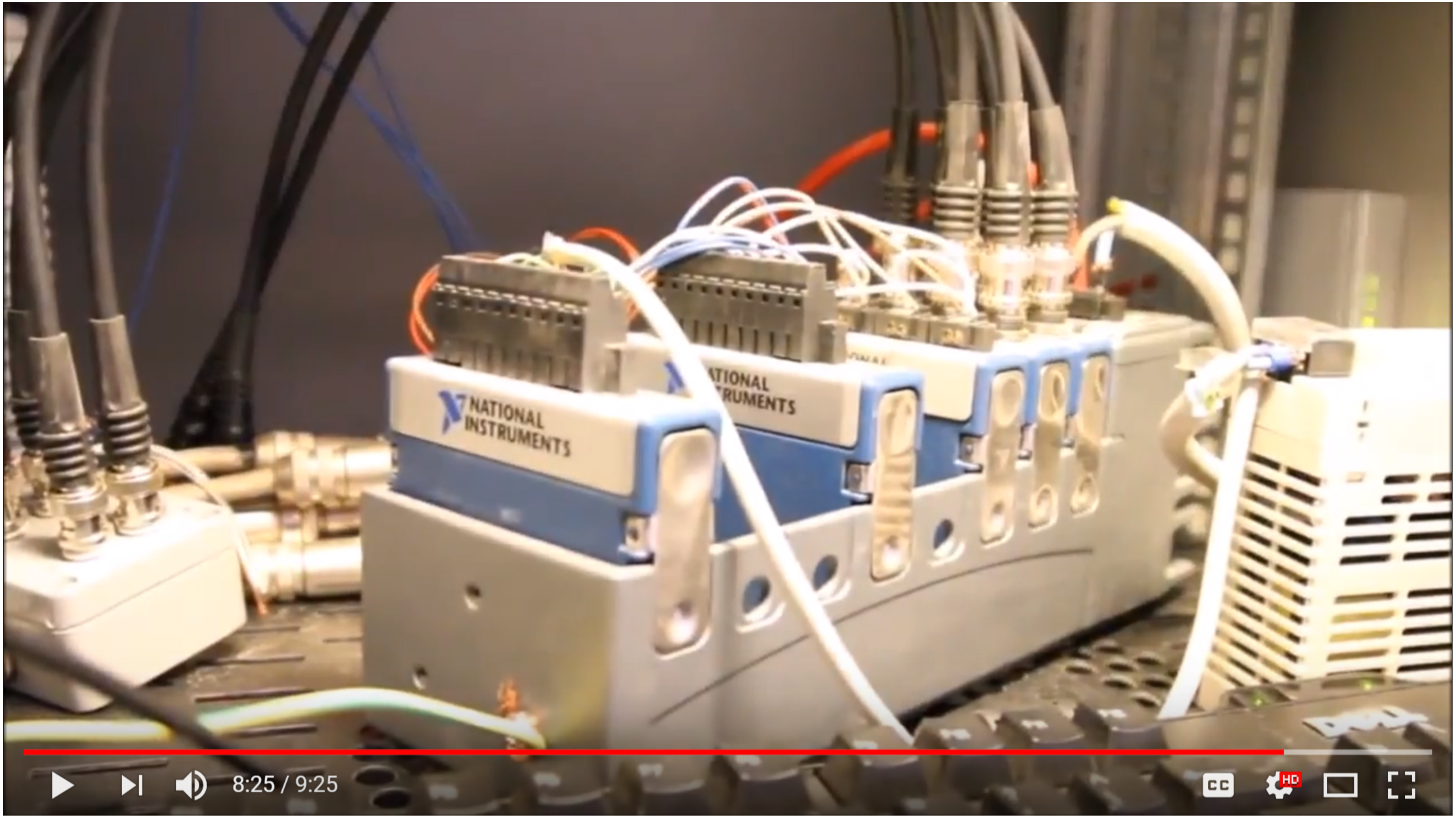
NATIONAL INSTRUMENTS
NATIONAL INSTRUMENTS
DELL
8:25 / 9:25
CC
HD

Flat Earth CON SPIRES! Cathedral Spires Are Secretly Atmospheric Electricity Ma

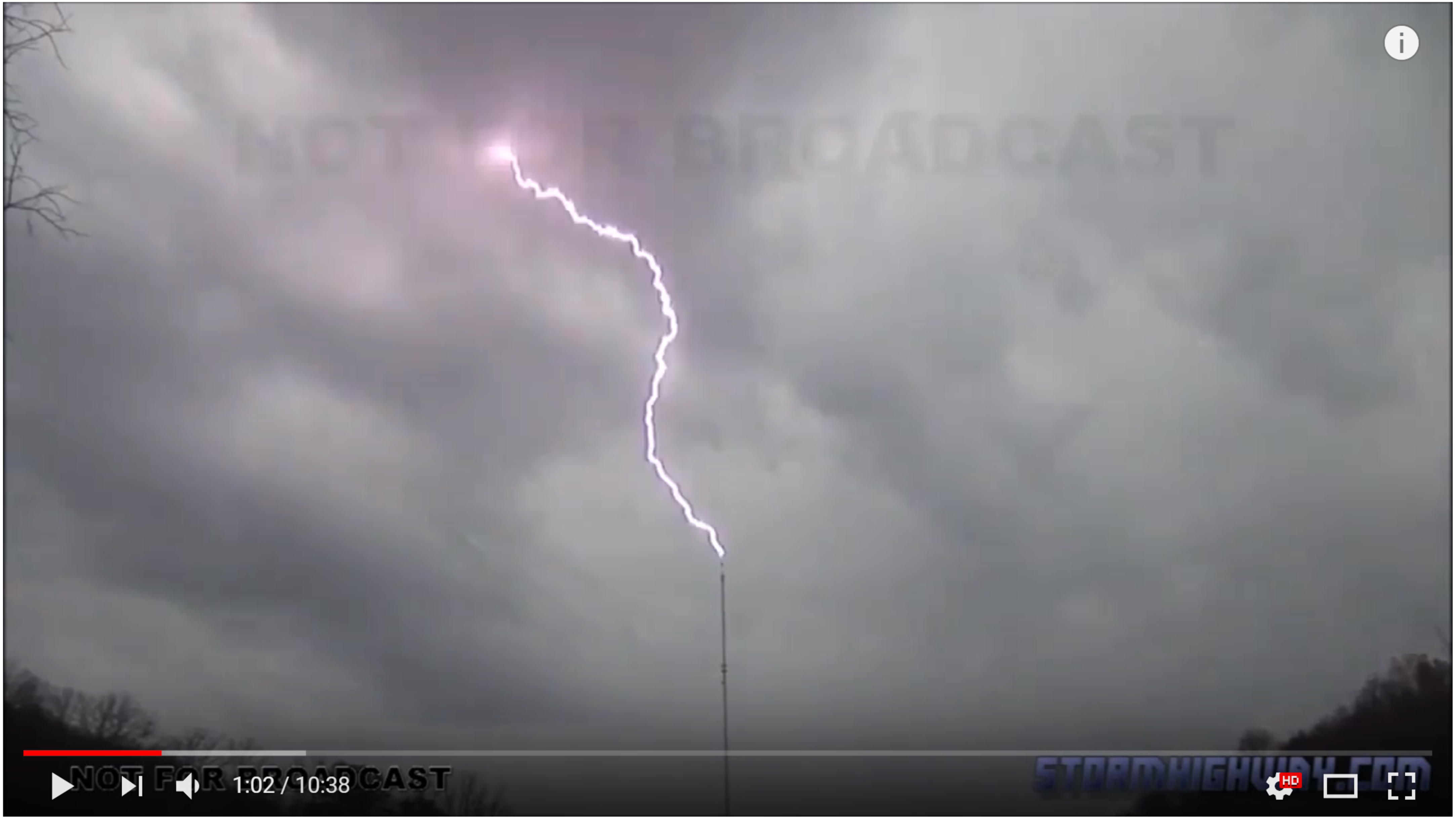
NOT FOR BROADCAST
NOT FOR BROADCAST
1:02 / 10:38
HD
STORMHIGHWAY.COM

1:45 / 9:25
CC
HD

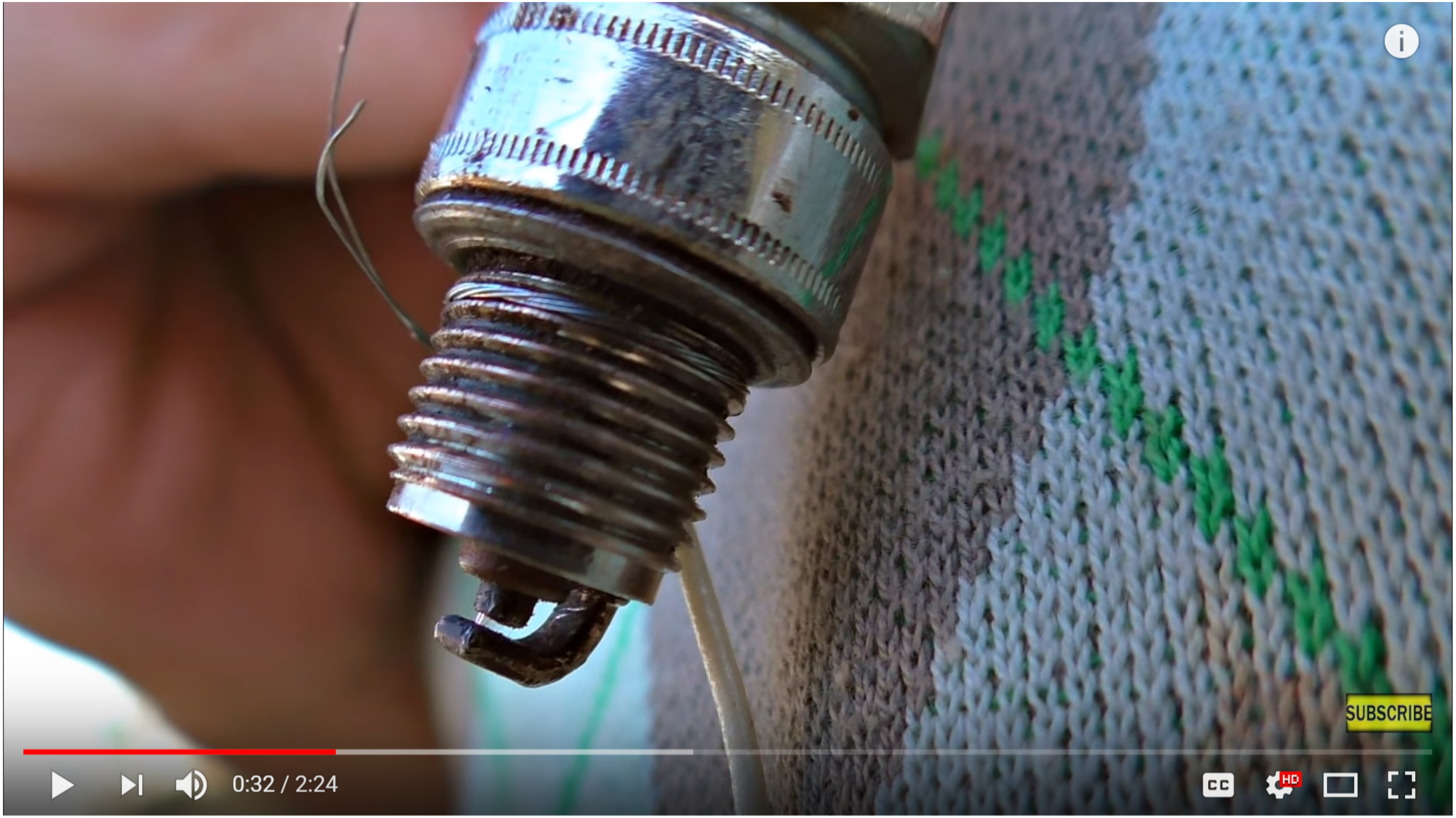
SUBSCRIBE
0:32 / 2:24
HD

Charging from the power lines.

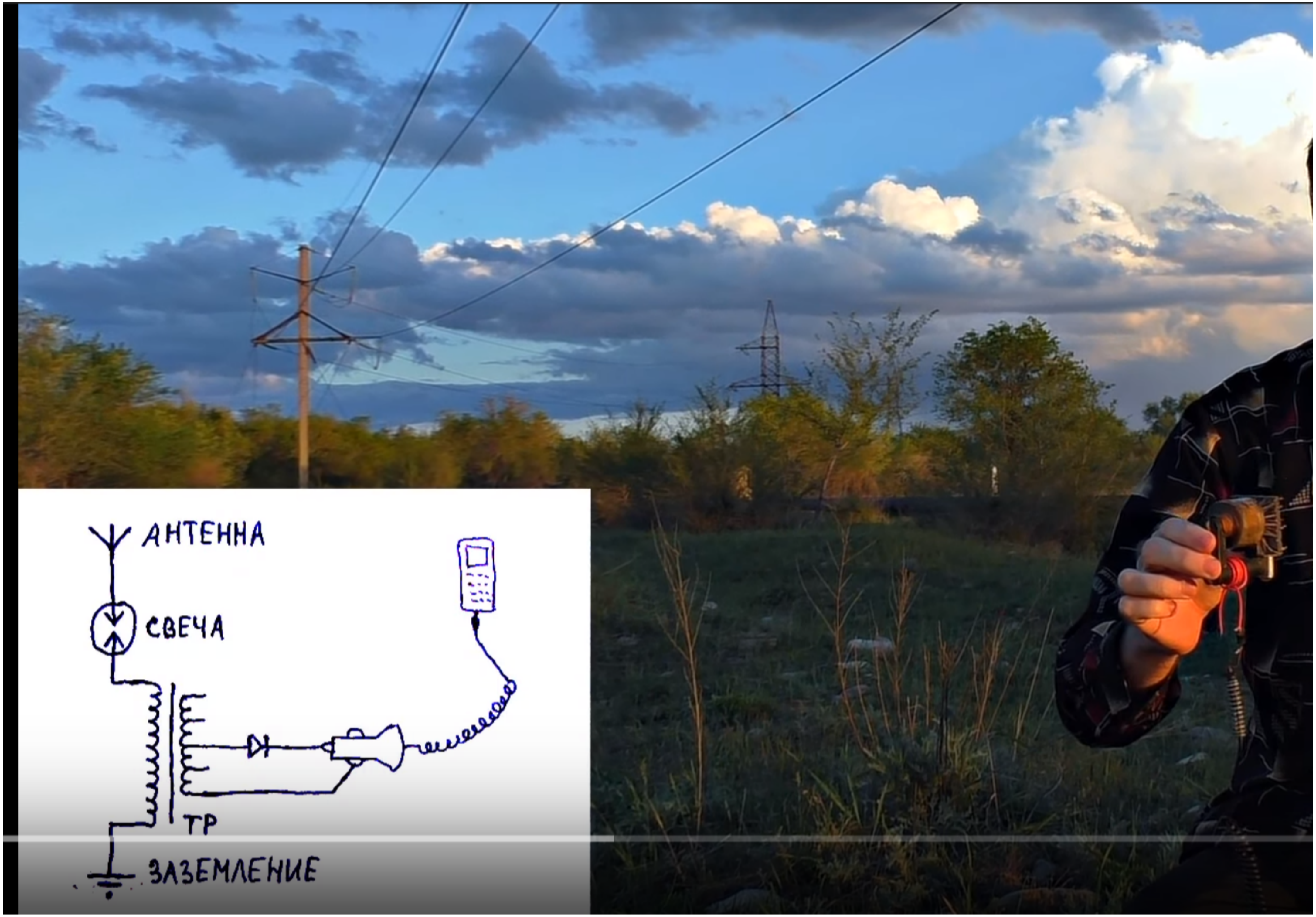
АНТЕННА
СВЕЧА
ТР
ЗАЗЕМЛЕНИЕ

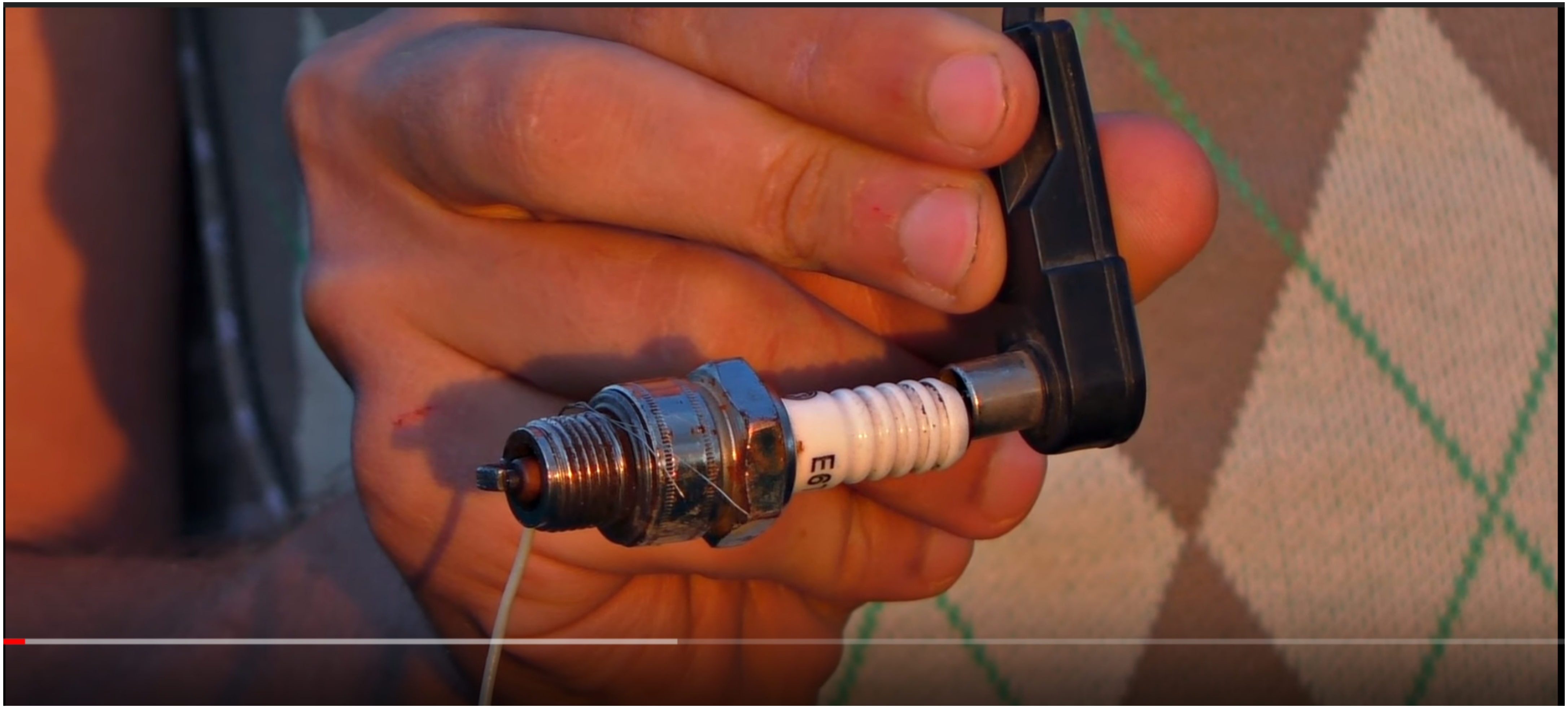
E6

MULTI DISPLAY DMM
10.1
POWER
DT-920
COM
VΩ
20A 15SEC MAX UNFUSED
200mA MAX FUSED
hFE
PNP
NPN
Cx
F

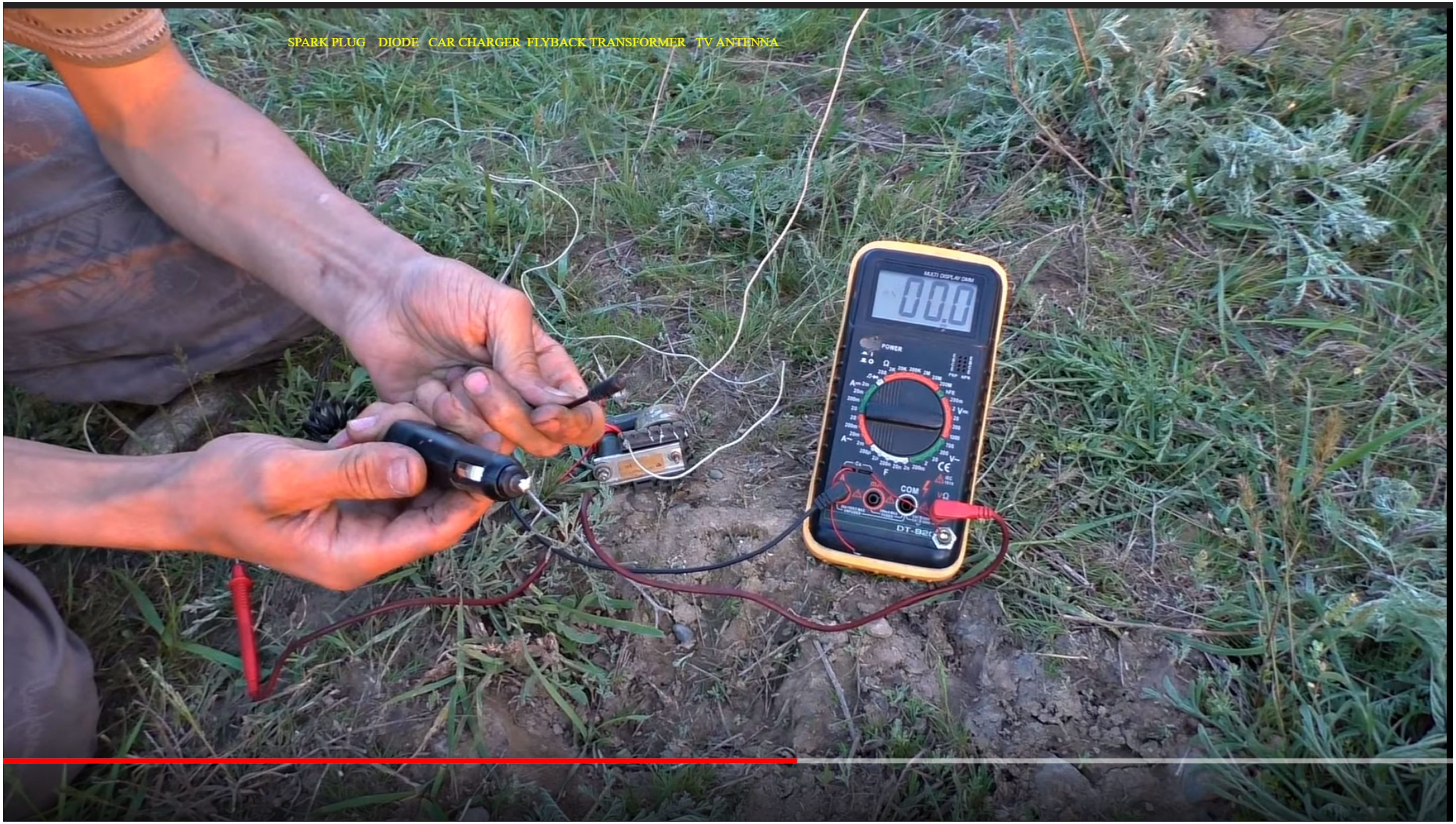
SPARK PLUG
DIODE
CAR CHARGER
FLYBACK TRANSFORMER
TV ANTENNA
00.0
POWER
COM

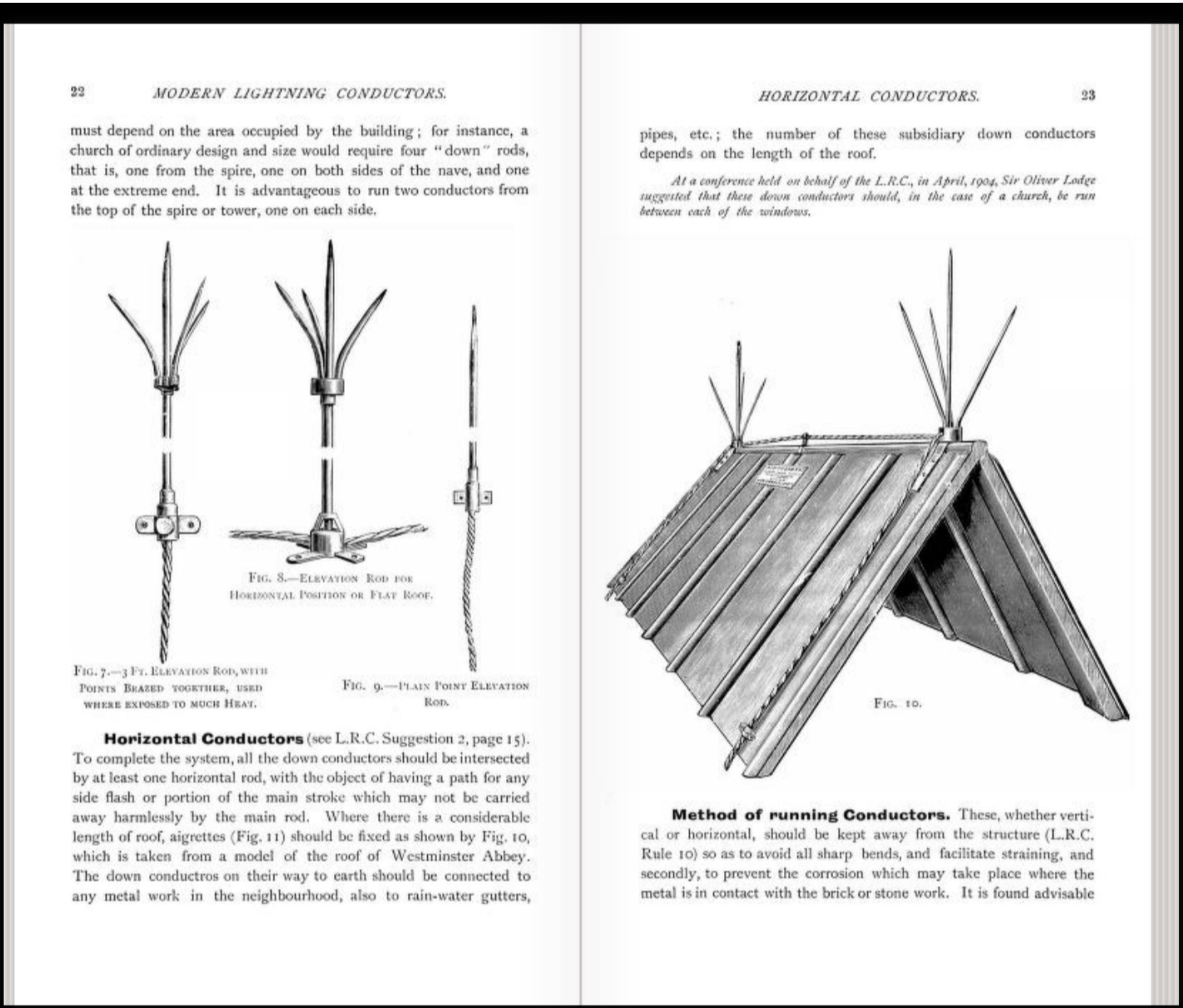

22 *MODERN LIGHTNING CONDUCTORS.*

must depend on the area occupied by the building; for instance, a church of ordinary design and size would require four "down" rods, that is, one from the spire, one on both sides of the nave, and one at the extreme end. It is advantageous to run two conductors from the top of the spire or tower, one on each side.

FIG. 7.—3 FT. ELEVATION ROD, WITH POINTS BRAZED TOGETHER, USED WHERE EXPOSED TO MUCH HEAT.

FIG. 8.—ELEVATION ROD FOR HORIZONTAL POSITION OR FLAT ROOF.

FIG. 9.—PLAIN POINT ELEVATION ROD.

Horizontal Conductors (see L.R.C. Suggestion 2, page 15). To complete the system, all the down conductors should be intersected by at least one horizontal rod, with the object of having a path for any side flash or portion of the main stroke which may not be carried away harmlessly by the main rod. Where there is a considerable length of roof, aigrettes (Fig. 11) should be fixed as shown by Fig. 10, which is taken from a model of the roof of Westminster Abbey. The down conductros on their way to earth should be connected to any metal work in the neighbourhood, also to rain-water gutters, pipes, etc.; the number of these subsidiary down conductors depends on the length of the roof.

HORIZONTAL CONDUCTORS. 23

At a conference held on behalf of the L.R.C., in April, 1904, Sir Oliver Lodge suggested that these down conductors should, in the case of a church, be run between each of the windows.

FIG. 10.

Method of running Conductors. These, whether vertical or horizontal, should be kept away from the structure (L.R.C. Rule 10) so as to avoid all sharp bends, and facilitate straining, and secondly, to prevent the corrosion which may take place where the metal is in contact with the brick or stone work. It is found advisable

Modern lightning conductors: an illustrated surface of the joint need not necessarily exceed that of the cross section of the conductors. The joint should be put together previously by screws or rivets, and the soldered joint, especially if used in underground work, should be carefully protected from local electrical action by tarred rope. Stranded iron conductors can be connected (as previously described) by use of a box joint; the box, Fig. 28, must be of the same metal as the conductors.

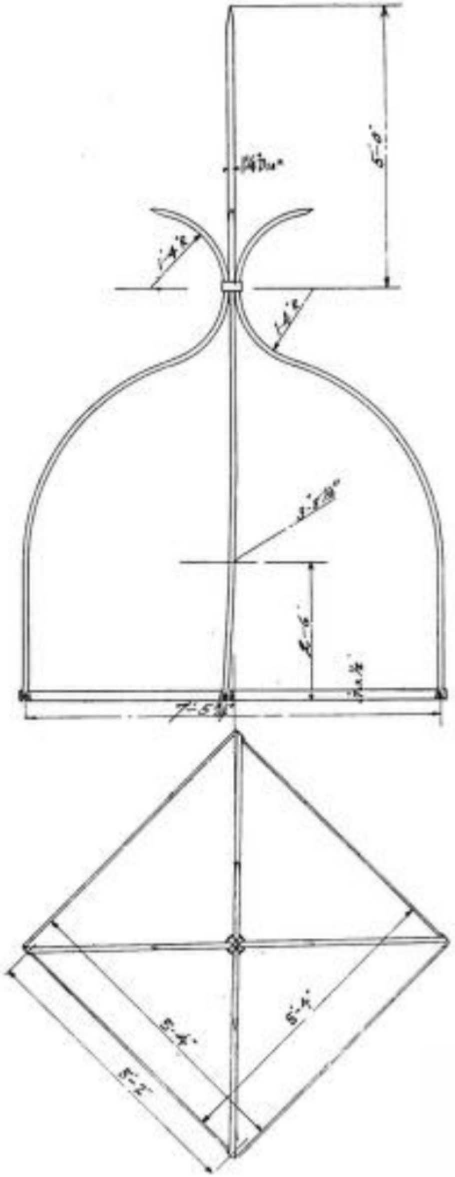

FIG. 27.—TERMINALS IN FORM OF AN ARCH FOR CHIMNEY STACK.

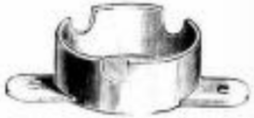

FIG. 28.

Vanes.—Particular attention must be paid to the necessity of making a permanent joint to the spindle. A clamp is prepared of the same material as the spindle, and is furnished with two bolts to tighten; if iron is used it is well to line the clamp with a piece of sheet lead. The conductor is sweated into a socket which is fitted with an eye, through which one of the tightening bolts passes. In the case of the vanes of churches and those fixed in inaccessible positions, two separate clamps should be used.

Internal Masses of Metal.—Roof trusses fitted with longitudinal iron tie rods will, as a rule, be found to be electrically connected, but should this not be the case each truss must be joined to the conductors. All large and long masses of metal, such as beams, girders, roof trusses, tie rods, hot water systems, traveller ways, hoisting crabs, engines, boilers, large machines, and ventilators fixed in the interiors of buildings, should be connected to all conductors that pass near them, and as far as possible with one another. The discontinuous parts of traveller rails should be connected by straps, or in some cases tramway bonds might be used. If electric light wires are run in tubes, such as the "SIMPLEX," this should be earthed. Metallic contact between lead or zinc sheeting and flashings should be carefully studied, and for special work strips of sufficient size should be either burnt on to lead or soldered in such a way that the joint will stand rough usage, and allow for expansion or contraction.

Earth Connection.—"*It is essential that the lower extremity of the conductor be buried in permanently damp soil; hence proximity to rain-water pipes, and to drains, is desirable. It is a very good plan to make the conductor bifurcate close below the surface of the ground, and adopt two of the following methods for securing the escape of the lightning into the earth. A strip of copper tape may be led from the bottom of the rod to the nearest gas or water* main—*not merely to a lead pipe—and be soldered to it; or a tape may be soldered to a sheet of copper* 3 *feet by* 3 *feet and* $\frac{1}{16}$ *inch thick, buried in permanently wet earth, and surrounded by cinders or coke; or many yards of the tape may be laid in a trench filled with coke, taking care that the surfaces of copper are, as in the previous cases, not less than* 18 *square feet. Where iron is used for the rod, a galvanised iron plate of similar dimensions should be employed.*

"The use of cinders or coke appears to be questionable owing to the chemical or electrolytic effect on copper or iron. Charcoal or pulverised carbon (such as ends of arc-light rods) is better. **A tubular earth** consisting of a perforated steel spike driven tightly into moist ground and lengthened up to the surface, the conductor reaching to the bottom and being packed with granulated charcoal, gives as much **effective area** as a plate of larger surface, and can easily be kept moist by connecting it to the nearest rain-water pipe. The resistance of a tubular earth on this plan should be very low and practically constant."—*Lightning Research Committee, 1905.*

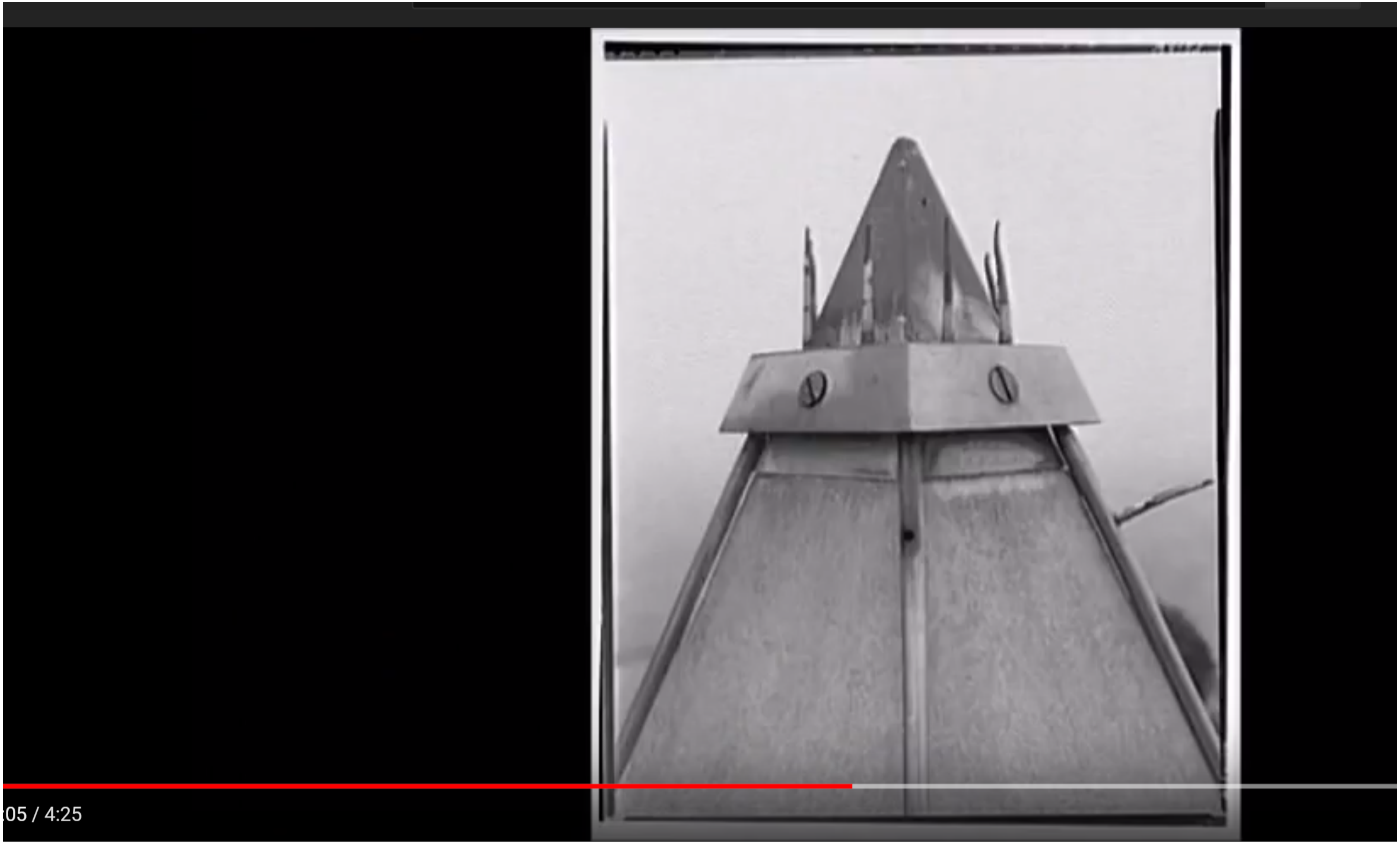

Washington Monument is an Atmospheric Electricity Mast

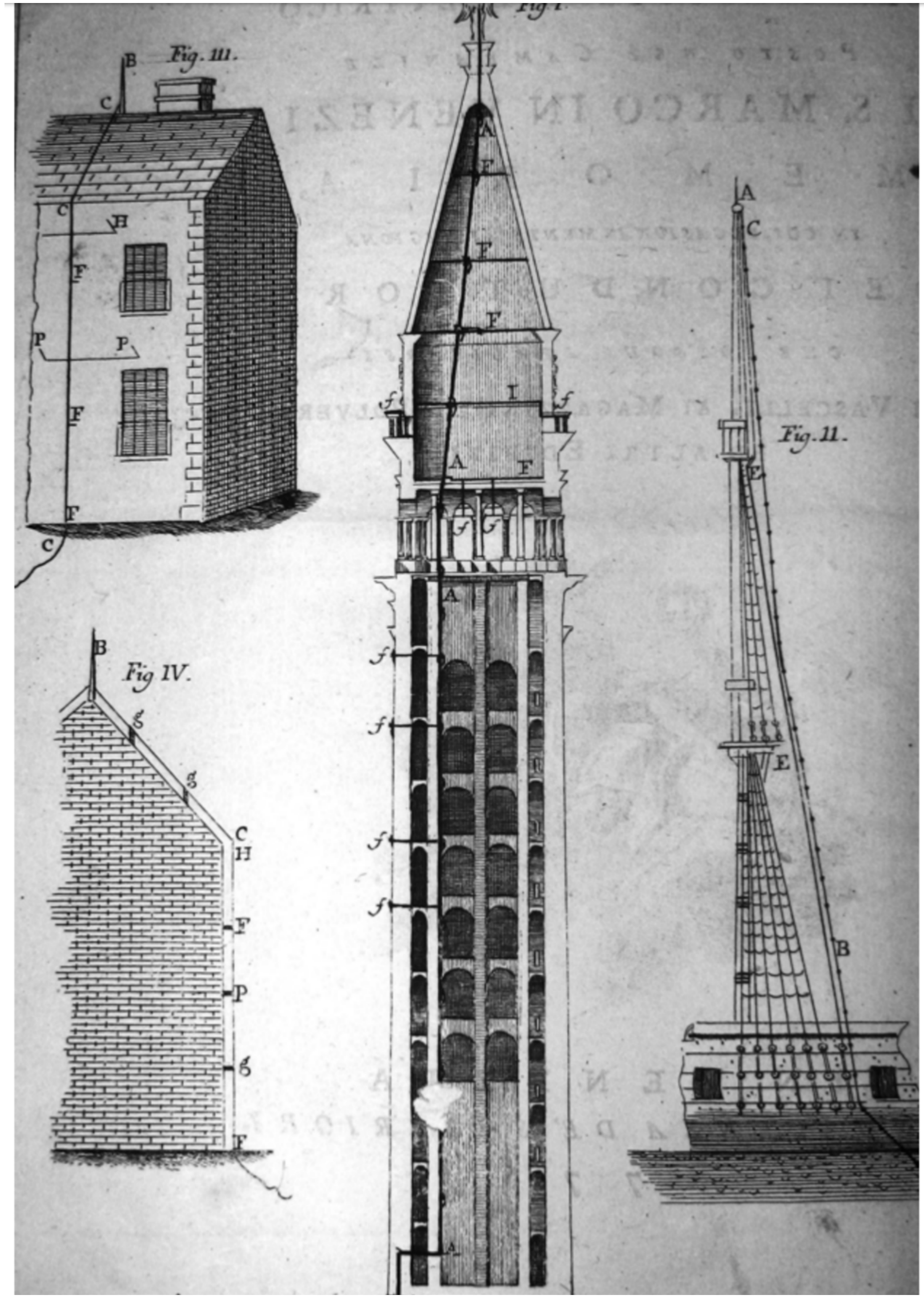
Fig. III.
B
C
C
H
F
P
P
F
F
C
Fig. IV.
B
g
g
C
H
F
P
g
F
Fig. I.
A
F
F
F
I
f
f
A
F
f
f
A
f
f
f
f
A
Fig. II.
A
C
E
E
B

Figure 2.3. The lightning rod that Toaldo designed for the church of San Marco in Venice. GiuseppeToaldo, “Del conduttore elettrico posto nel campanile” (Padua, 1776). Franklin Collec-

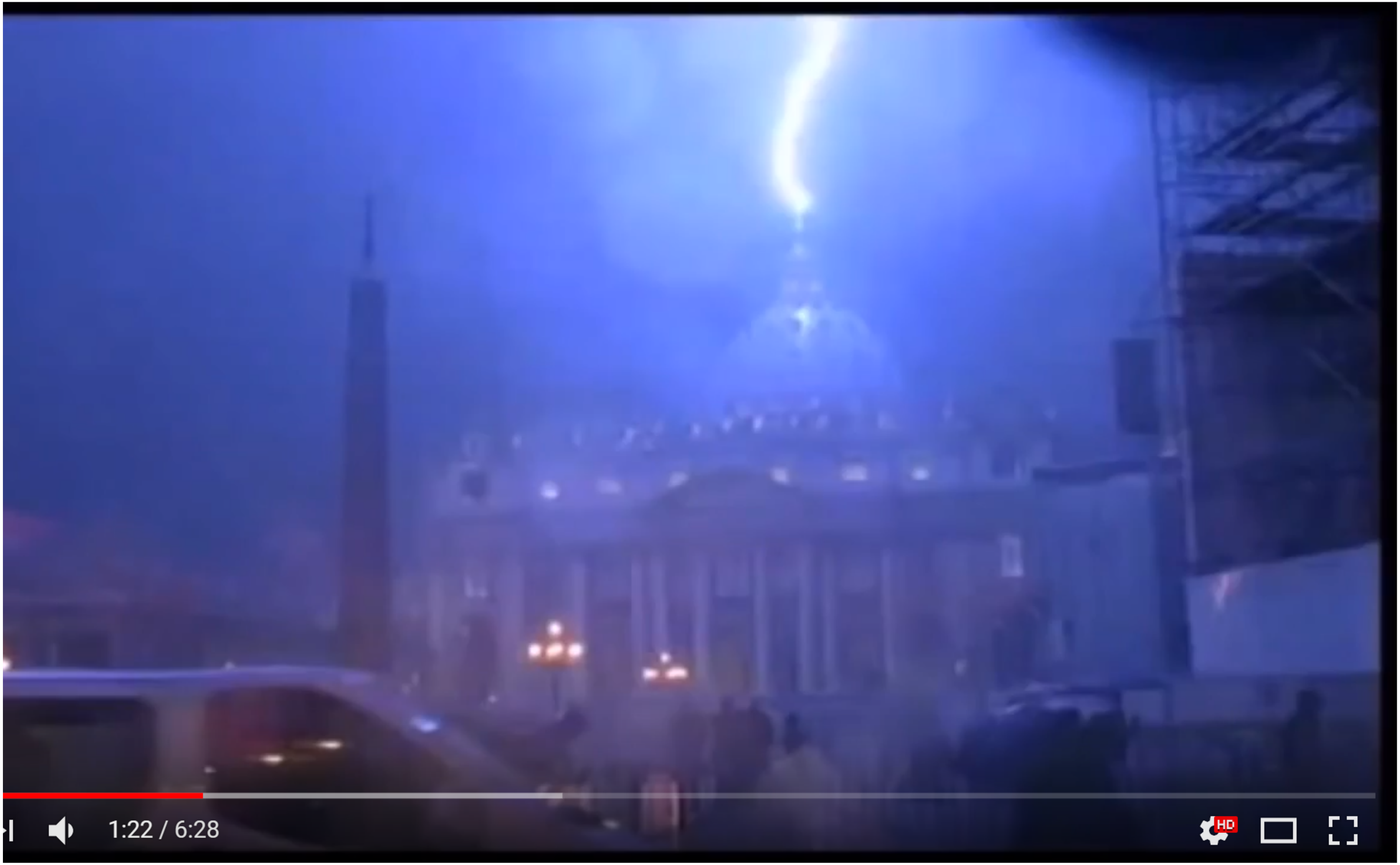

Gilded Temples Are Secretly Atmospheric Electricity Masts

How To Find S-W Stations Quickly

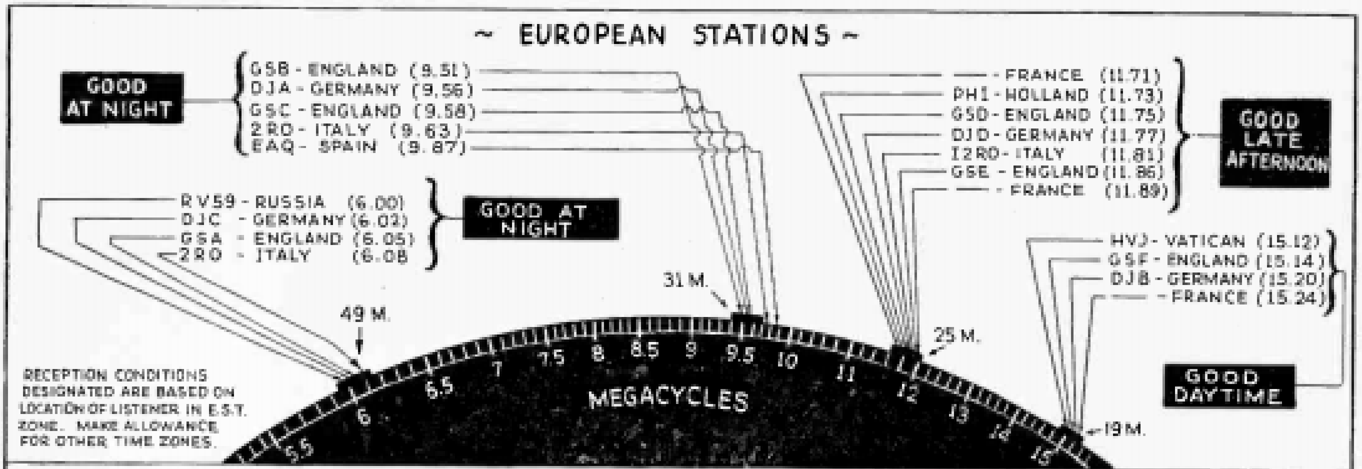
~ EUROPEAN STATIONS ~
GOOD AT NIGHT
GSB - ENGLAND (9.51)
DJA - GERMANY (9.56)
GSC - ENGLAND (9.58)
2RO - ITALY (9.63)
EAQ - SPAIN (9.87)
—— - FRANCE (11.71)
PH1 - HOLLAND (11.73)
GSD - ENGLAND (11.75)
DJD - GERMANY (11.77)
I2RO - ITALY (11.81)
GSE - ENGLAND (11.86)
—— - FRANCE (11.89)
GOOD LATE AFTERNOON
RV59 - RUSSIA (6.00)
DJC - GERMANY (6.02)
GSA - ENGLAND (6.05)
2RO - ITALY (6.08)
GOOD AT NIGHT
HVJ - VATICAN (15.12)
GSF - ENGLAND (15.14)
DJB - GERMANY (15.20)
—— - FRANCE (15.24)
GOOD DAYTIME
31 M.
49 M.
25 M.
19 M.
RECEPTION CONDITIONS DESIGNATED ARE BASED ON LOCATION OF LISTENER IN E.S.T. ZONE. MAKE ALLOWANCE FOR OTHER TIME ZONES.
MEGACYCLES
5.5 6 6.5 7 7.5 8 8.5 9 9.5 10 11 12 13 14 15

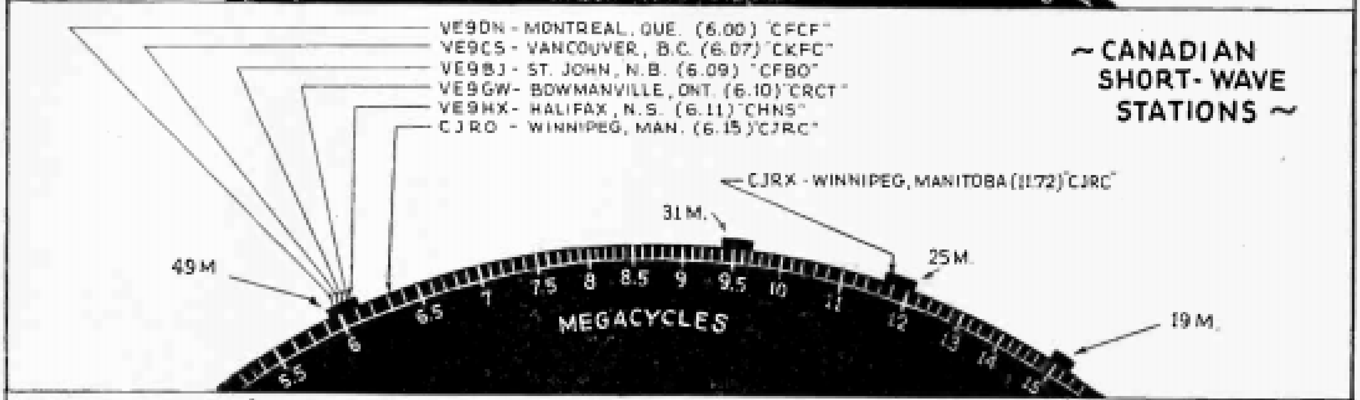
VE9DN - MONTREAL, QUE. (6.00) "CFCF"
VE9CS - VANCOUVER, B.C. (6.07) "CKFC"
VE9BJ - ST. JOHN, N.B. (6.09) "CFBO"
VE9GW - BOWMANVILLE, ONT. (6.10) "CRCT"
VE9HX - HALIFAX, N.S. (6.11) "CHNS"
CJRO - WINNIPEG, MAN. (6.15) "CJRC"
~ CANADIAN SHORT-WAVE STATIONS ~
CJRX - WINNIPEG, MANITOBA (11.72) "CJRC"
31 M.
49 M
25 M.
19 M.
MEGACYCLES
5.5 6 6.5 7 7.5 8 8.5 9 9.5 10 11 12 13 14 15

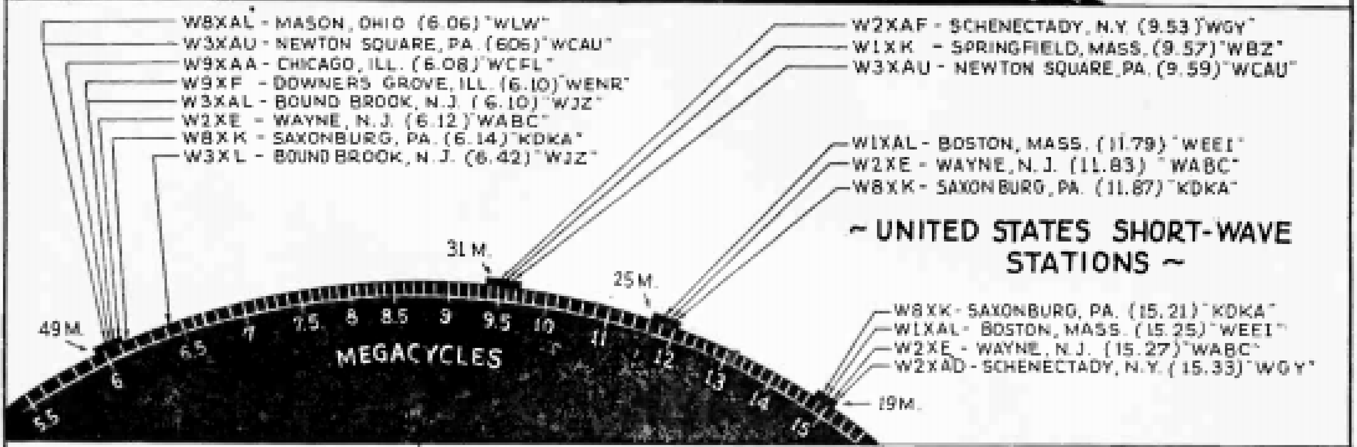
W8XAL - MASON, OHIO (6.06) "WLW"
W3XAU - NEWTON SQUARE, PA. (6.06) "WCAU"
W9XAA - CHICAGO, ILL. (6.08) "WCFL"
W9XF - DOWNERS GROVE, ILL. (6.10) "WENR"
W3XAL - BOUND BROOK, N.J. (6.10) "WJZ"
W2XE - WAYNE, N.J. (6.12) "WABC"
W8XK - SAXONBURG, PA. (6.14) "KDKA"
W3XL - BOUND BROOK, N.J. (6.42) "WJZ"
W2XAF - SCHENECTADY, N.Y. (9.53) "WGY"
W1XK - SPRINGFIELD, MASS. (9.57) "WBZ"
W3XAU - NEWTON SQUARE, PA. (9.59) "WCAU"
W1XAL - BOSTON, MASS. (11.79) "WEEI"
W2XE - WAYNE, N.J. (11.83) "WABC"
W8XK - SAXONBURG, PA. (11.87) "KDKA"
~ UNITED STATES SHORT-WAVE STATIONS ~
W8XK - SAXONBURG, PA. (15.21) "KDKA"
W1XAL - BOSTON, MASS. (15.25) "WEEI"
W2XE - WAYNE, N.J. (15.27) "WABC"
W2XAD - SCHENECTADY, N.Y. (15.33) "WGY"
31 M.
25 M.
49 M.
19 M.
MEGACYCLES
5.5 6 6.5 7 7.5 8 8.5 9 9.5 10 11 12 13 14 15

When To Tune and Where

● Until you have had considerable experience in short wave tuning, it is well to confine your efforts to the four bands indicated on the dial. These may look small, compared to the total length of the dial, but they contain the best "plums" among the foreign stations and many faint ones which, at first, you will probably pass unnoticed until you learn to tune slowly and carefully, intent upon identifying each slight "swish", which may be a station.

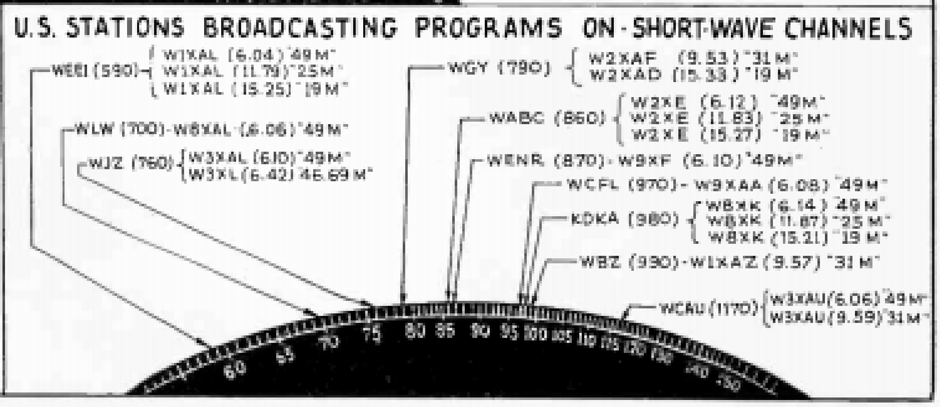
U.S. STATIONS BROADCASTING PROGRAMS ON-SHORT-WAVE CHANNELS
WEEI (590) { W1XAL (6.04) "49 M" W1XAL (11.79) "25 M" W1XAL (15.25) "19 M"
WGY (790) { W2XAF (9.53) "31 M" W2XAD (15.33) "19 M"
WABC (860) { W2XE (6.12) "49 M" W2XE (11.83) "25 M" W2XE (15.27) "19 M"
WLW (700) - W8XAL (6.06) "49 M"
WJZ (760) { W3XAL (6.10) "49 M" W3XL (6.42) "46.69 M"
WENR (870) - W9XF (6.10) "49 M"
WCFL (970) - W9XAA (6.08) "49 M"
KDKA (980) { W8XK (6.14) "49 M" W8XK (11.87) "25 M" W8XK (15.21) "19 M"
WBZ (990) - W1XAZ (9.57) "31 M"
WCAU (1170) { W3XAU (6.06) "49 M" W3XAU (9.59) "31 M"
60 65 70 75 80 85 90 95 100 105 110 115 120 130 140 150

Copyright 1935 by H. Gernsback.

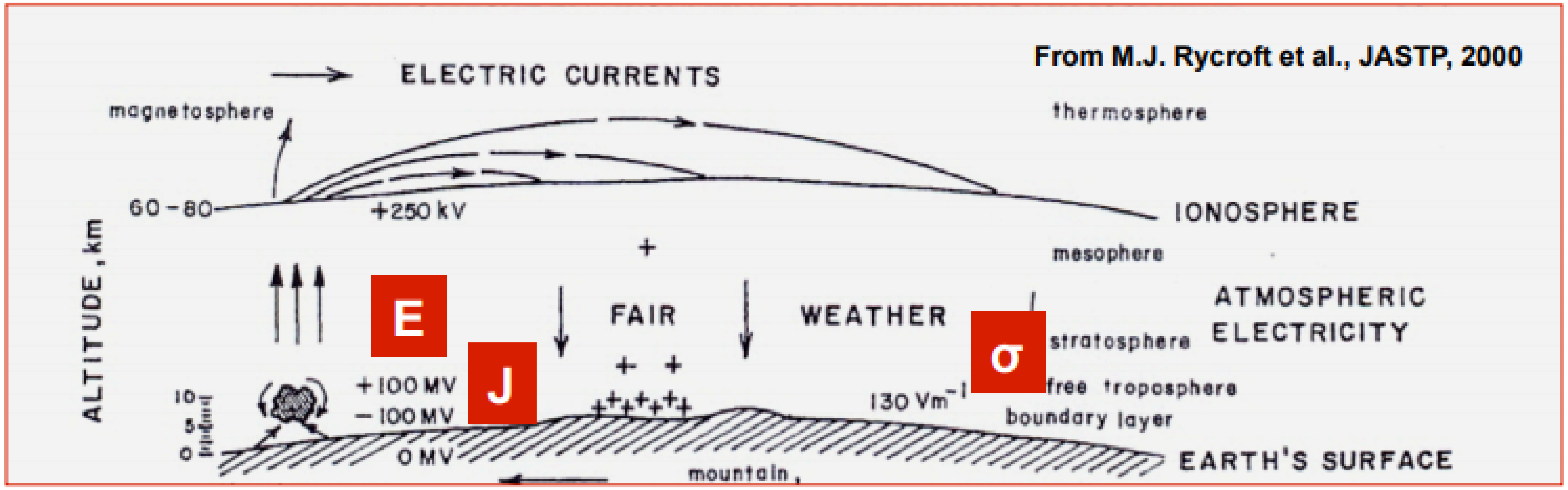
From M.J. Rycroft et al., JASTP, 2000

ELECTRIC CURRENTS
magnetosphere
thermosphere
60 - 80
+250 kV
IONOSPHERE
ALTITUDE, km
+
mesophere
E
FAIR
WEATHER
ATMOSPHERIC ELECTRICITY
σ
J
stratosphere
+100 MV
- 100 MV
130 Vm^{-1}
free troposphere
boundary layer
10
5
0
0 MV
mountain
EARTH'S SURFACE

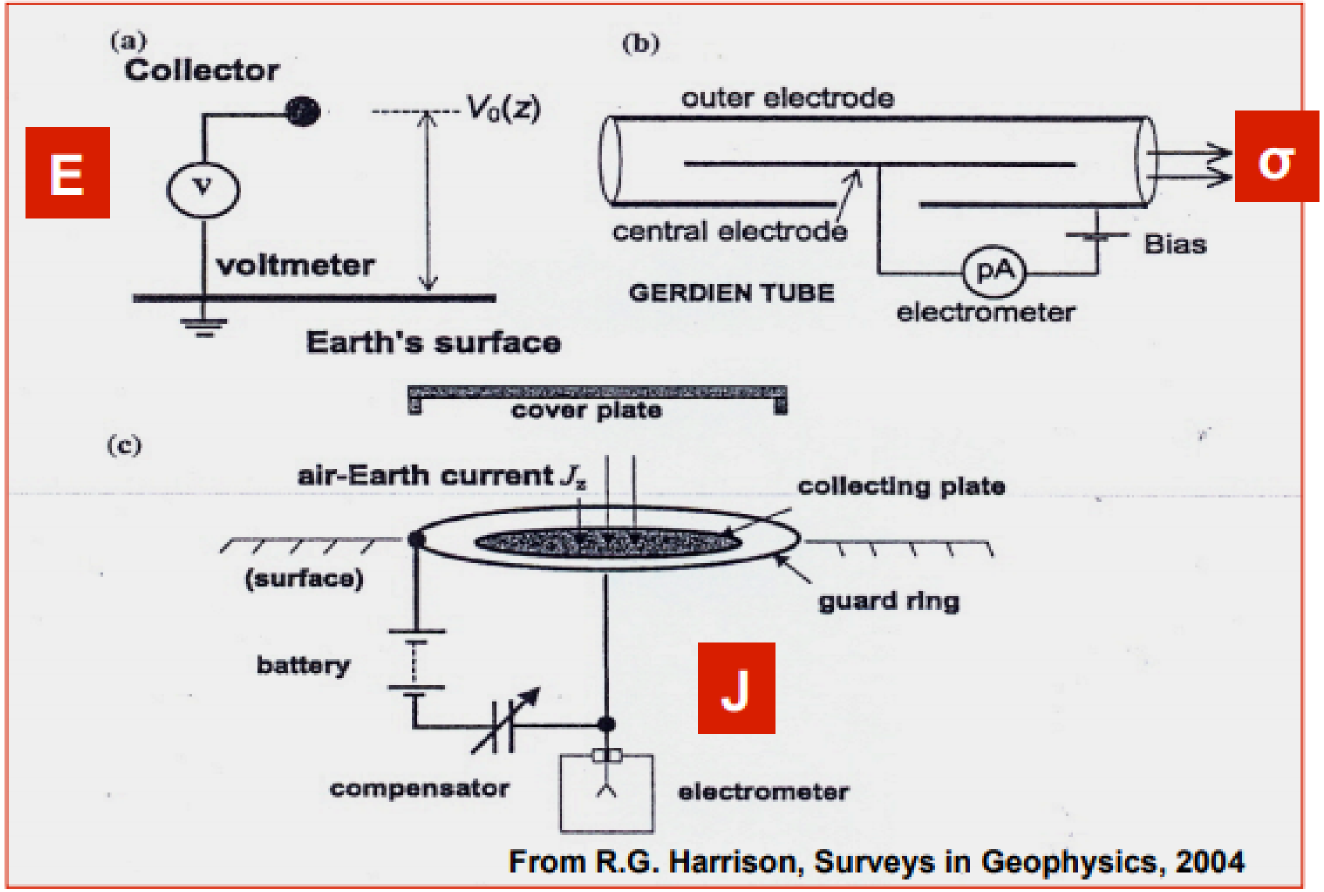
From R.G. Harrison, Surveys in Geophysics, 2004

(a)
Collector
E
$V_0(z)$
V
voltmeter
Earth's surface
(b)
outer electrode
σ
central electrode
Bias
GERDIEN TUBE
pA
electrometer
cover plate
(c)
air-Earth current J_z
collecting plate
(surface)
guard ring
battery
J
compensator
electrometer

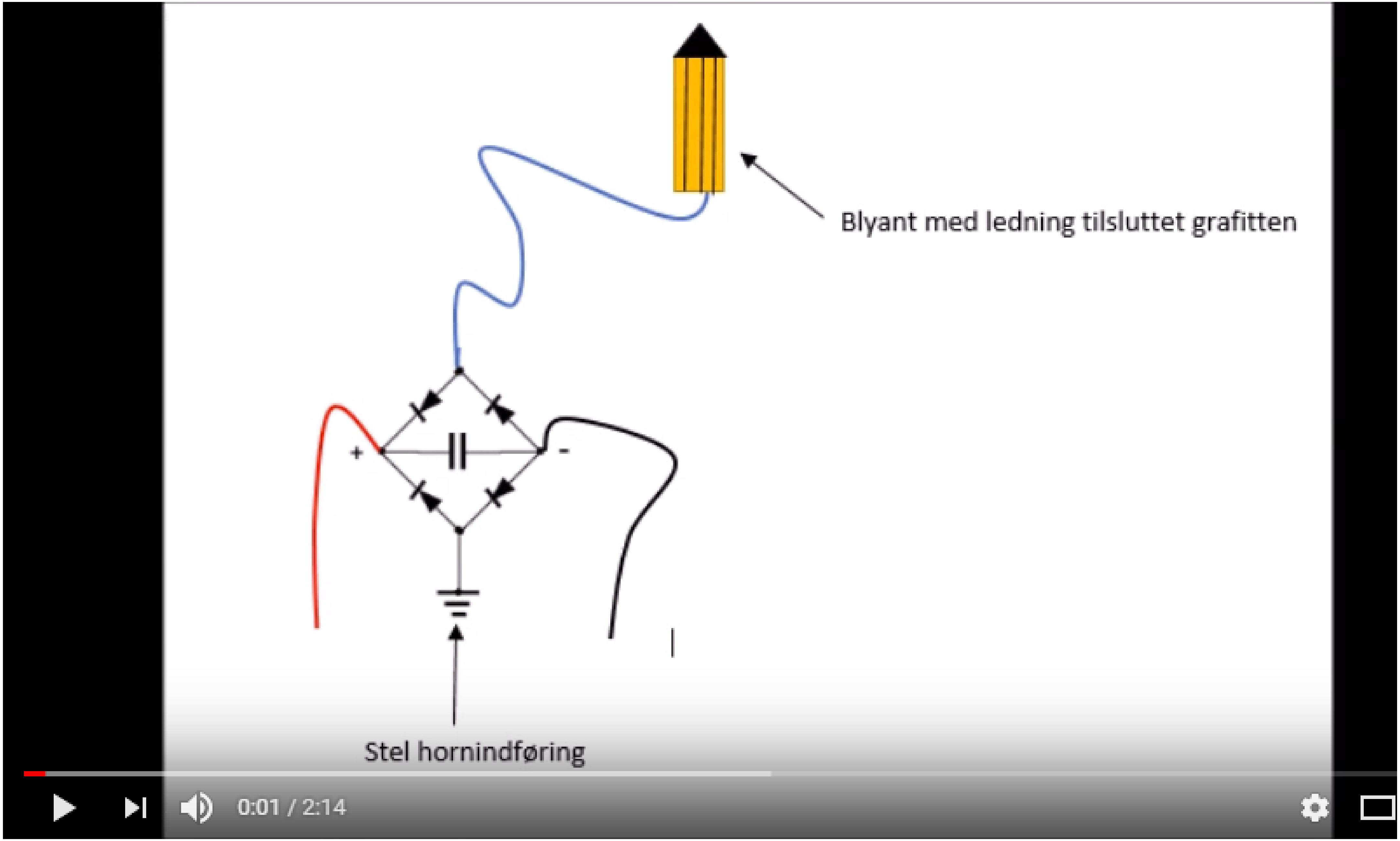
Blyant med ledning tilsluttet grafitten
+
-
Stel hornindføring
0:01 / 2:14

tesla radiant collector/ion collector with pencil

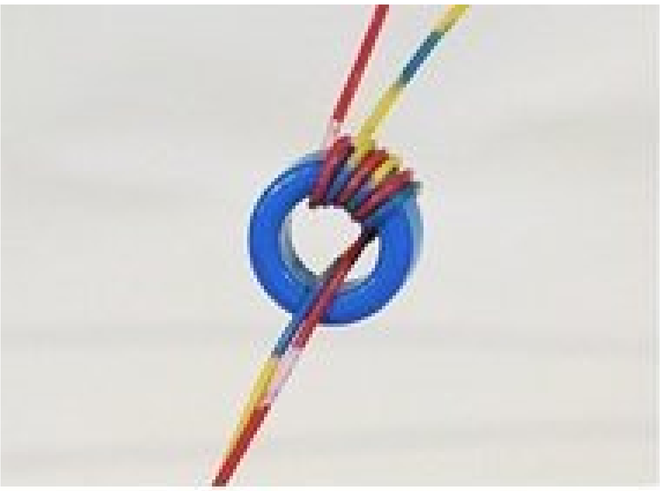

Keeping the two wires together, make a few more turns through the center.

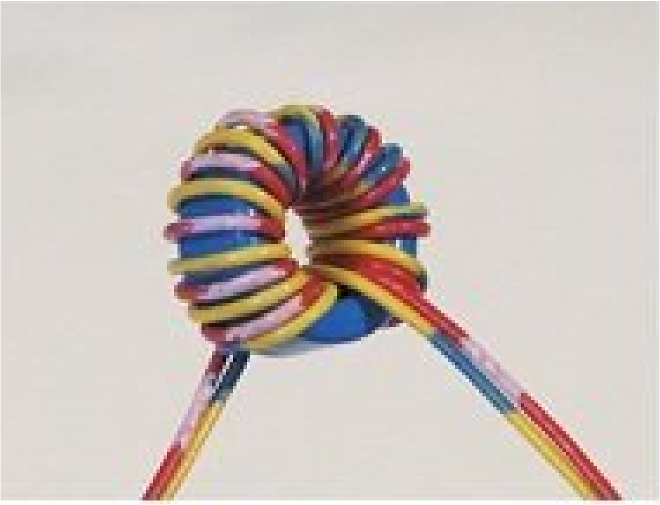

Keep winding until you fit as many turns as will fit in a single layer around the toroid, typically 7-10 turns with thin insulated wire.

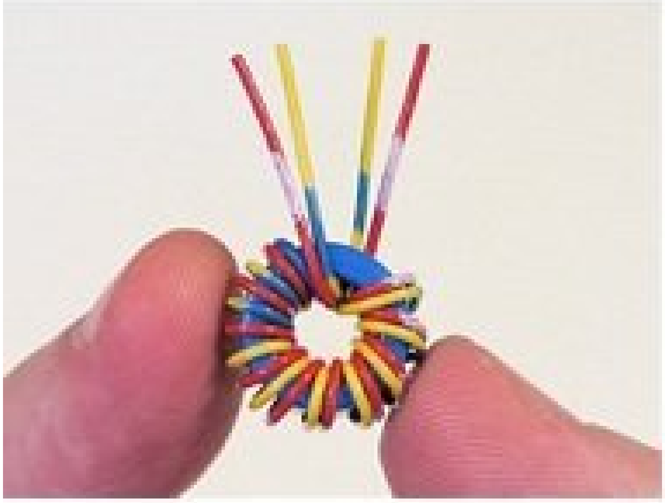

Clip the wire leads down. Note that we have two pairs of wires: one coming out the front, and one coming out the back.

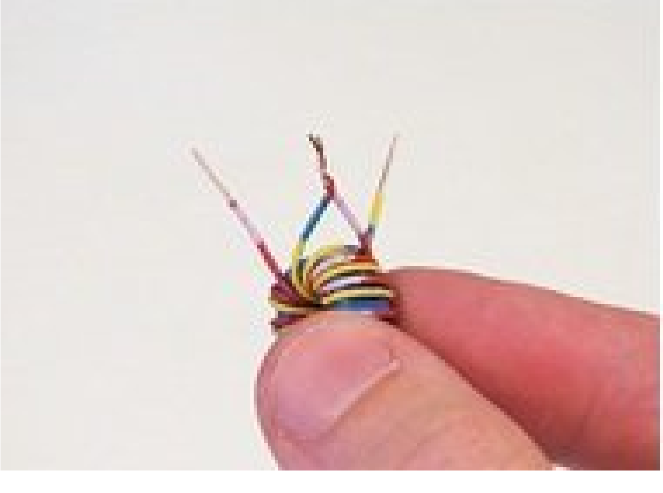

Strip the wire ends. Take one wire from each pair *of different color* and attach them together.

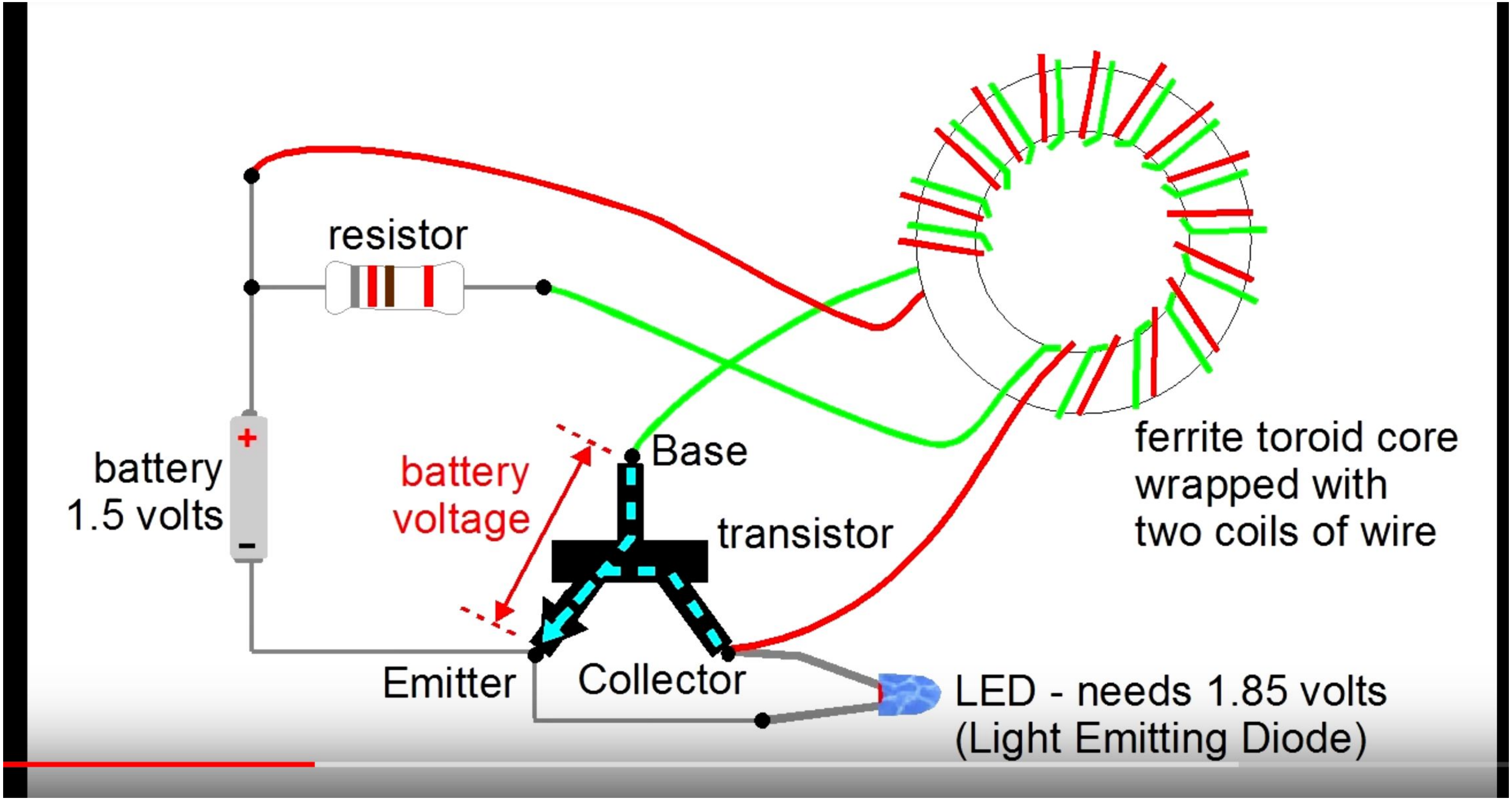
resistor
battery
1.5 volts
+
-
battery
voltage
Base
transistor
Emitter
Collector
ferrite toroid core
wrapped with
two coils of wire
LED - needs 1.85 volts
(Light Emitting Diode)

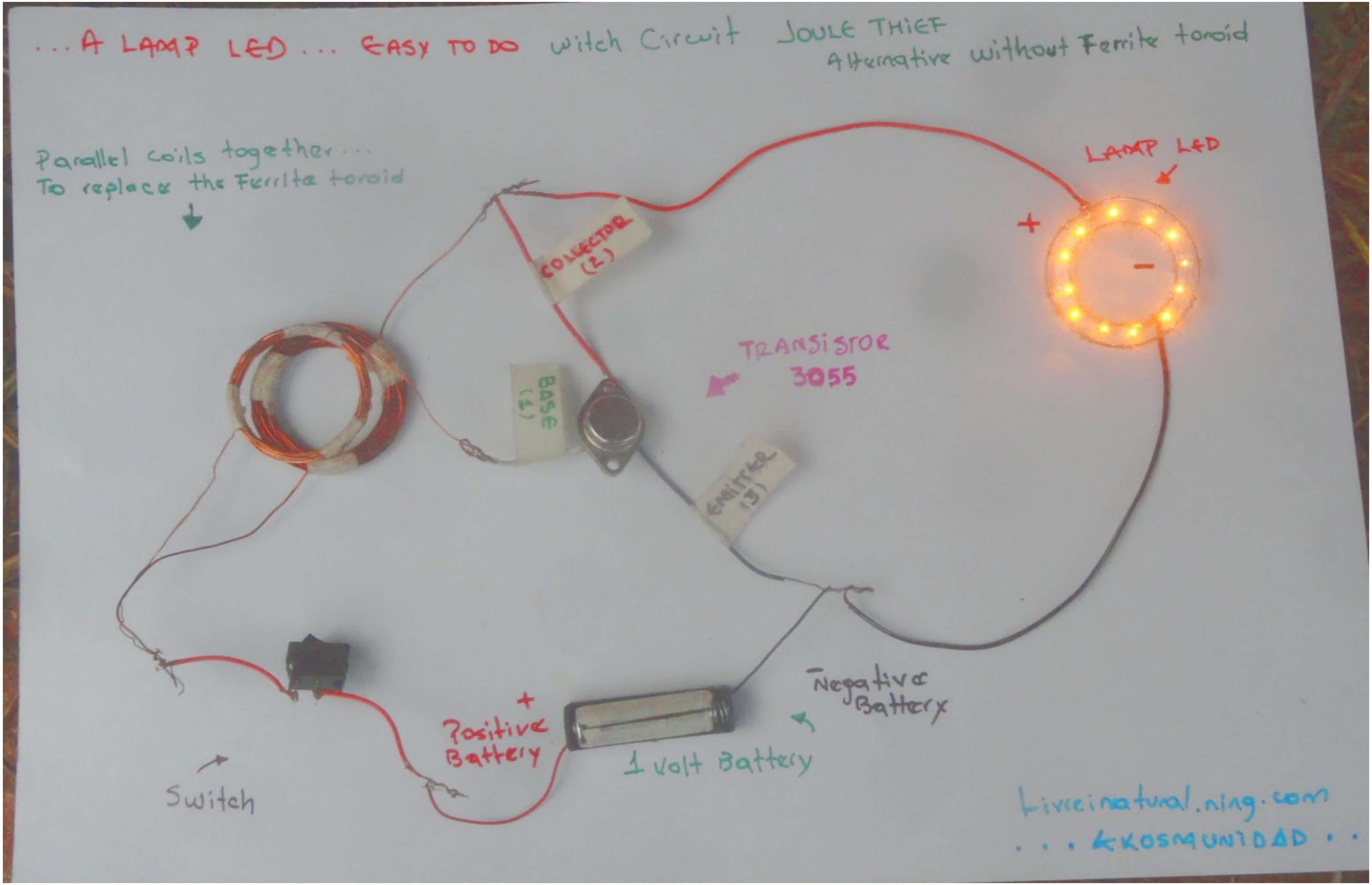
... A LAMP LED ... EASY TO DO witch Circuit JOULE THIEF
Alternative without Ferrite toroid
Parallel coils together ...
To replace the Ferrite toroid
LAMP LED
+
-
COLLECTOR (2)
BASE (1)
EMITTER (3)
TRANSISTOR 3055
+
Positive Battery
-Negative Battery
1 Volt Battery
Switch
Livreinatural.ning.com
... EKOSMUNIDAD ..

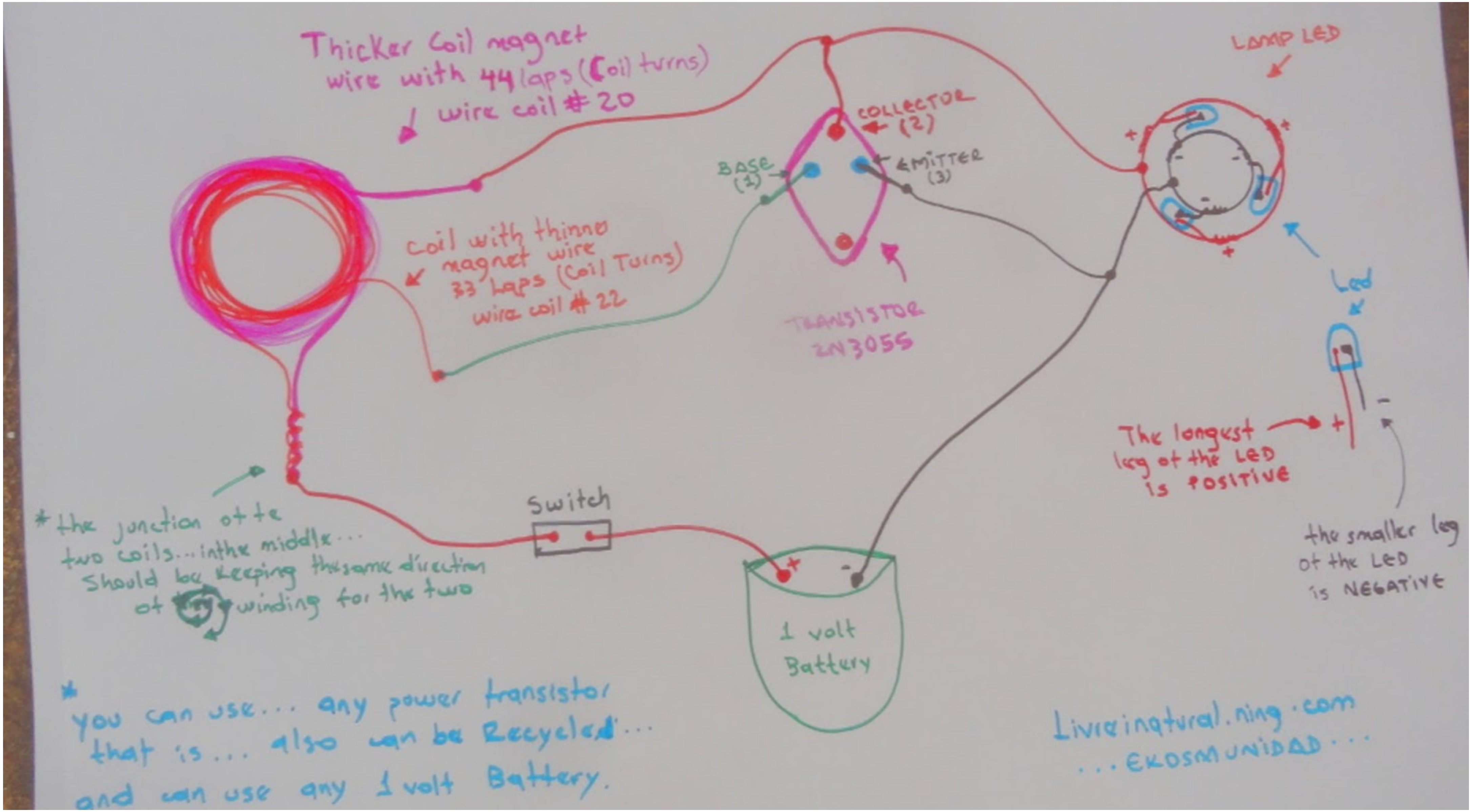
Thicker Coil magnet wire with 44 laps (Coil turns)
wire coil # 20
Coil with thinner magnet wire 33 laps (Coil Turns)
wire coil # 22
BASE (1)
COLLECTOR (2)
EMITTER (3)
TRANSISTOR 2N3055
LAMP LED
+
+
+
+
Led
The longest leg of the LED is positive
+
-
the smaller leg of the LED is NEGATIVE
Switch
+
-
1 volt Battery

* the junction of te two coils... in the middle... Should be keeping the same direction of winding for the two

* you can use... any power transistor that is... also can be Recycled... and can use any 1 volt Battery.

Liveinatural.ning.com
...EKOSMUNIDAD...

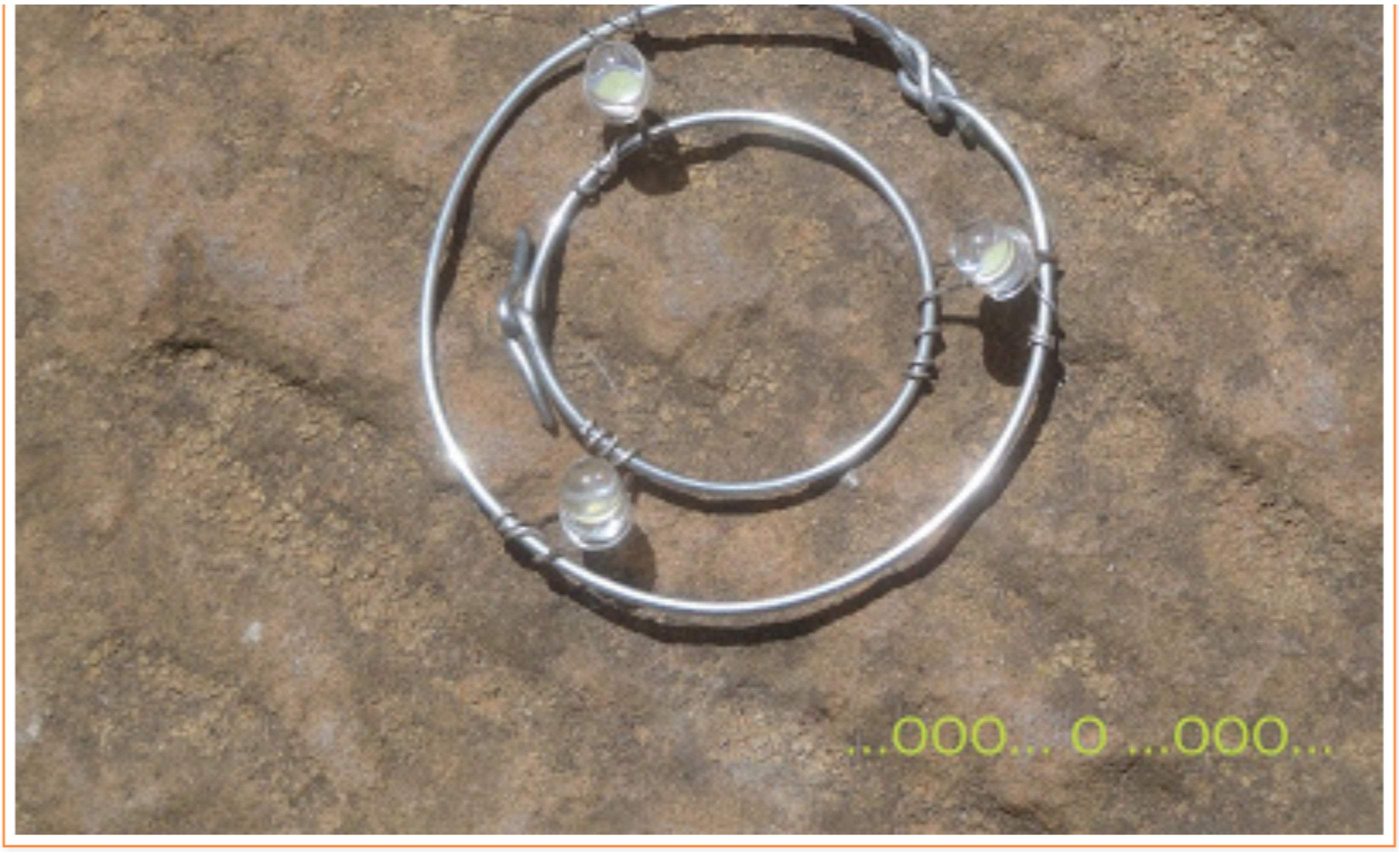
...OOO... O ...OOO...

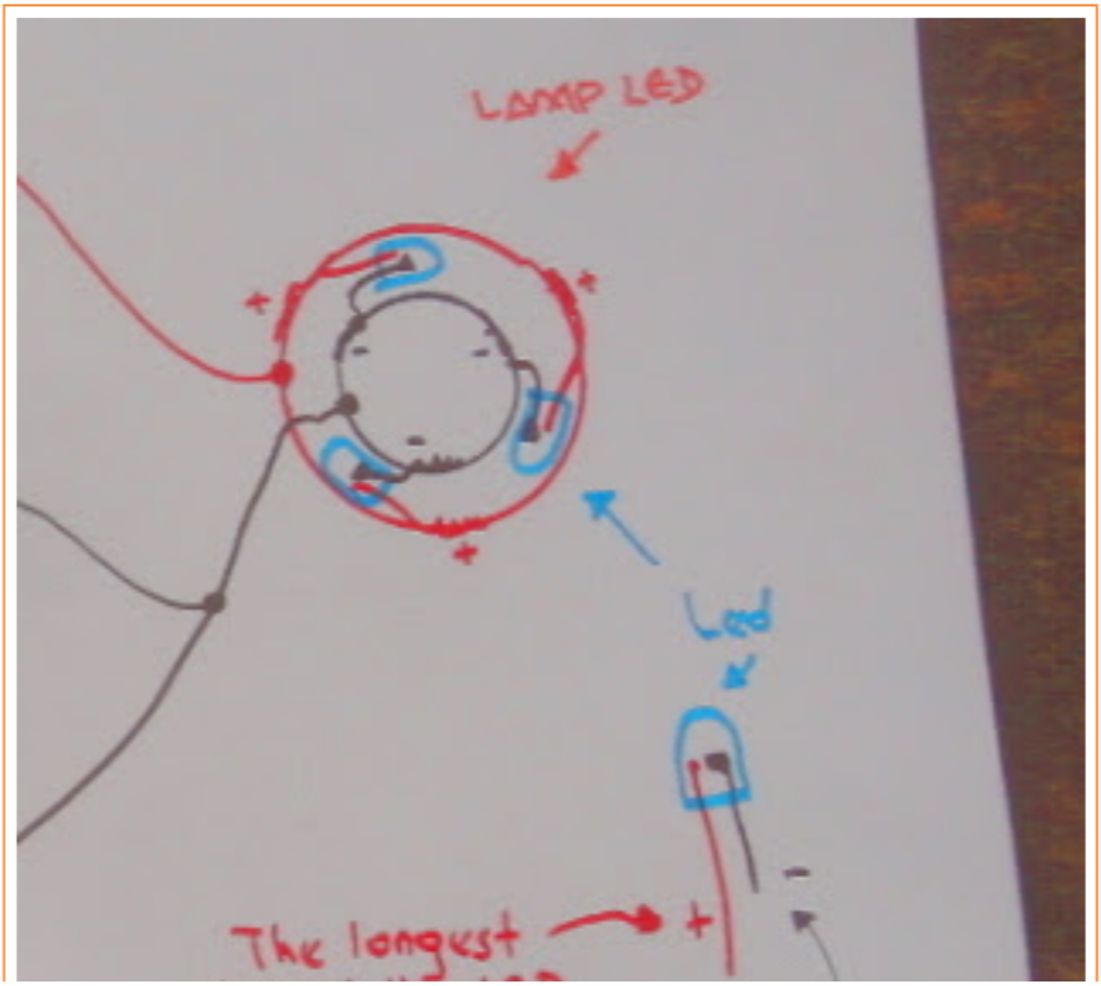
Lamp LED
+
+
+
Led
+
-
The longest

...Livreinatural.ning.com... ... EkosOphia ...O...

...Livreinatural.ning.com...

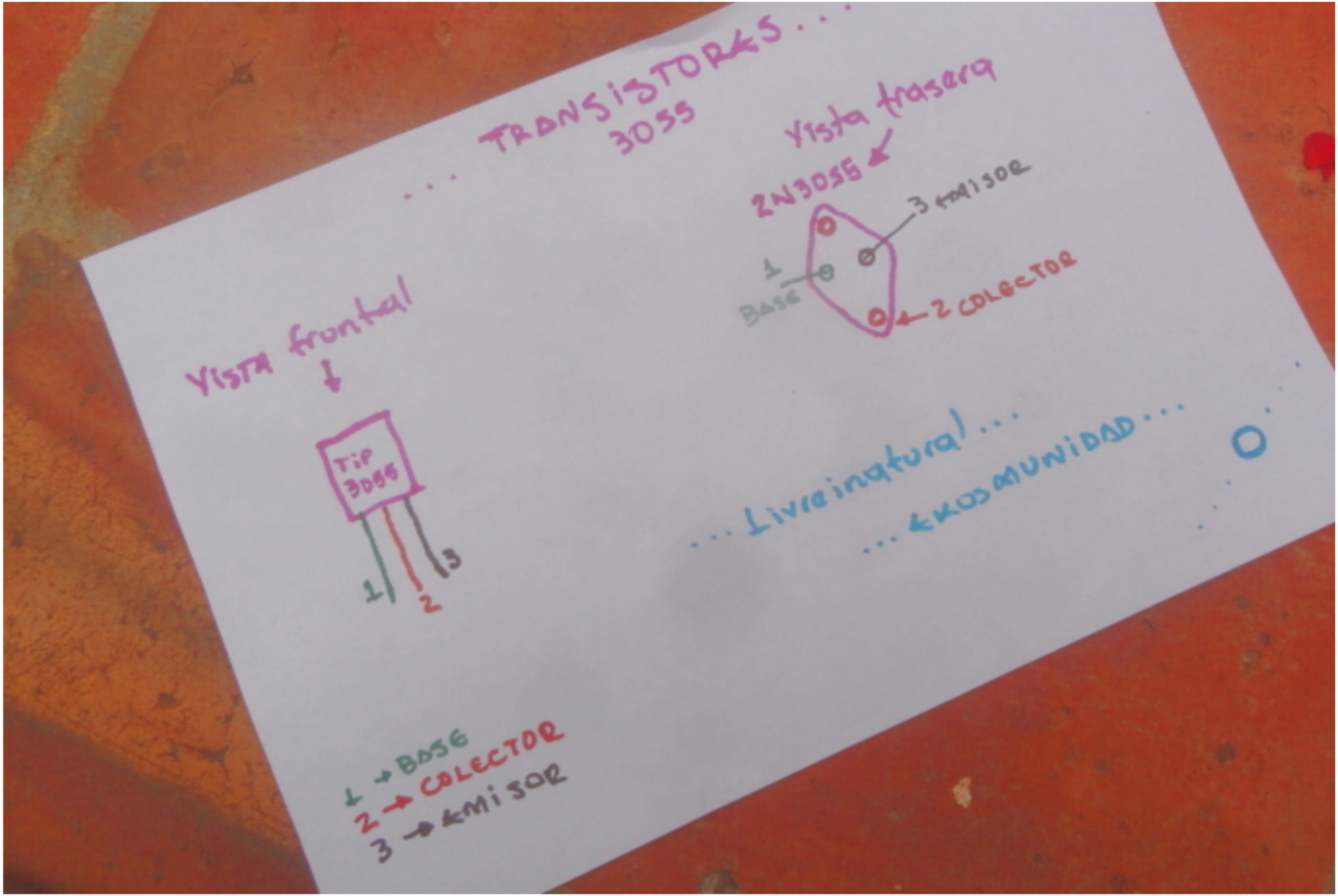

... TRANSISTORES ...
3055

Vista trasera
2N3055

3 EMISOR
1
BASE
2 COLECTOR

Vista frontal

TIP
3055
1
2
3

1 → BASE
2 → COLECTOR
3 → EMISOR

... Livreinatural ...
... EKOSMUNIDAD ...

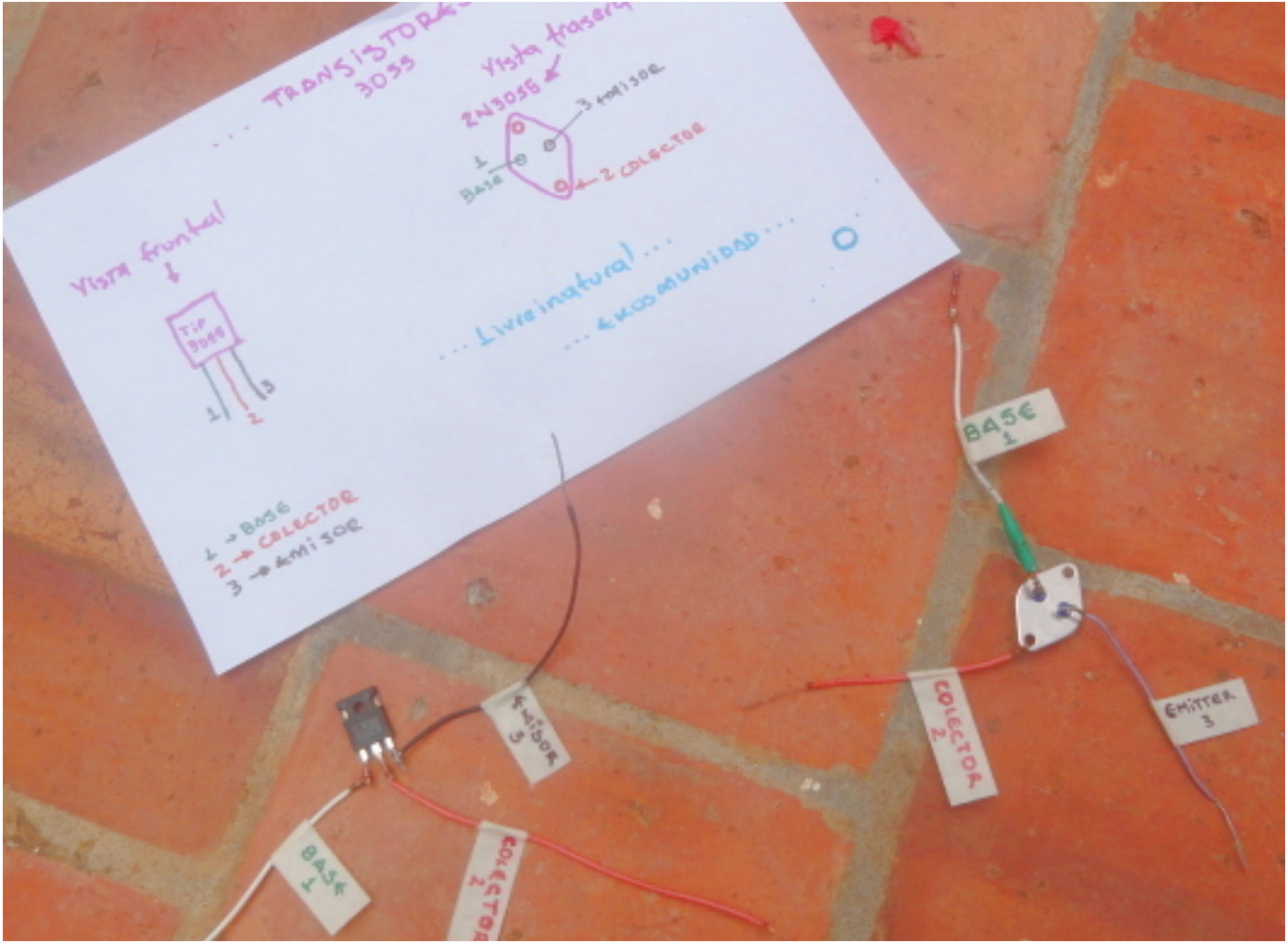
... TRANSISTORE
3055
Vista trasera
2N3055
3 EMISOR
2 COLECTOR
1
BASE
Vista frontal
TIP
3055
1
2
3
... Livreinatural ...
... EKOSMUNIDAD ...
1 → BASE
2 → COLECTOR
3 → EMISOR
BASE
1
COLECTOR
2
EMITTER
3
EMISOR
3
BASE
1
COLECTOR
2

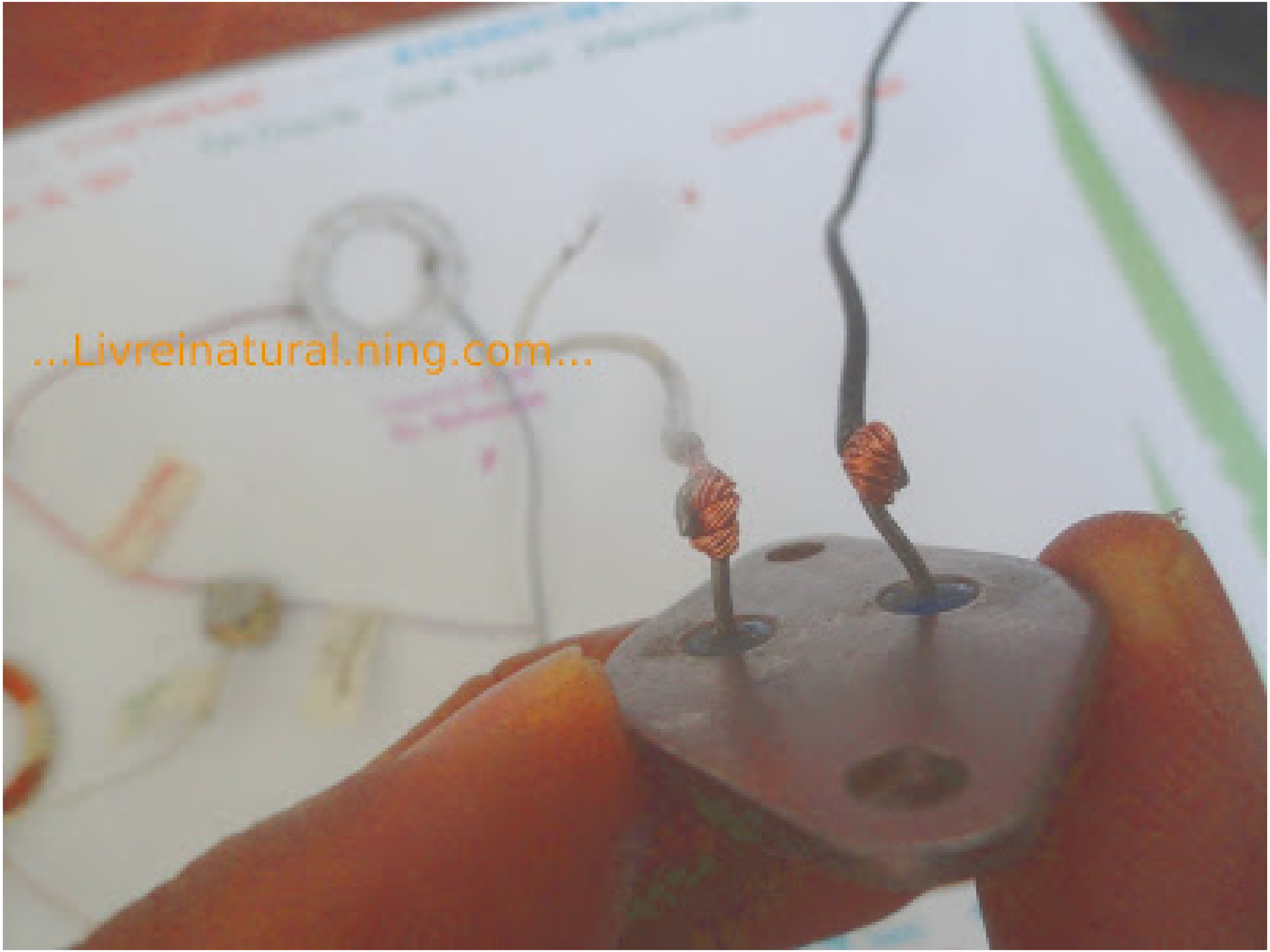

...Livreinatural.ning.com...

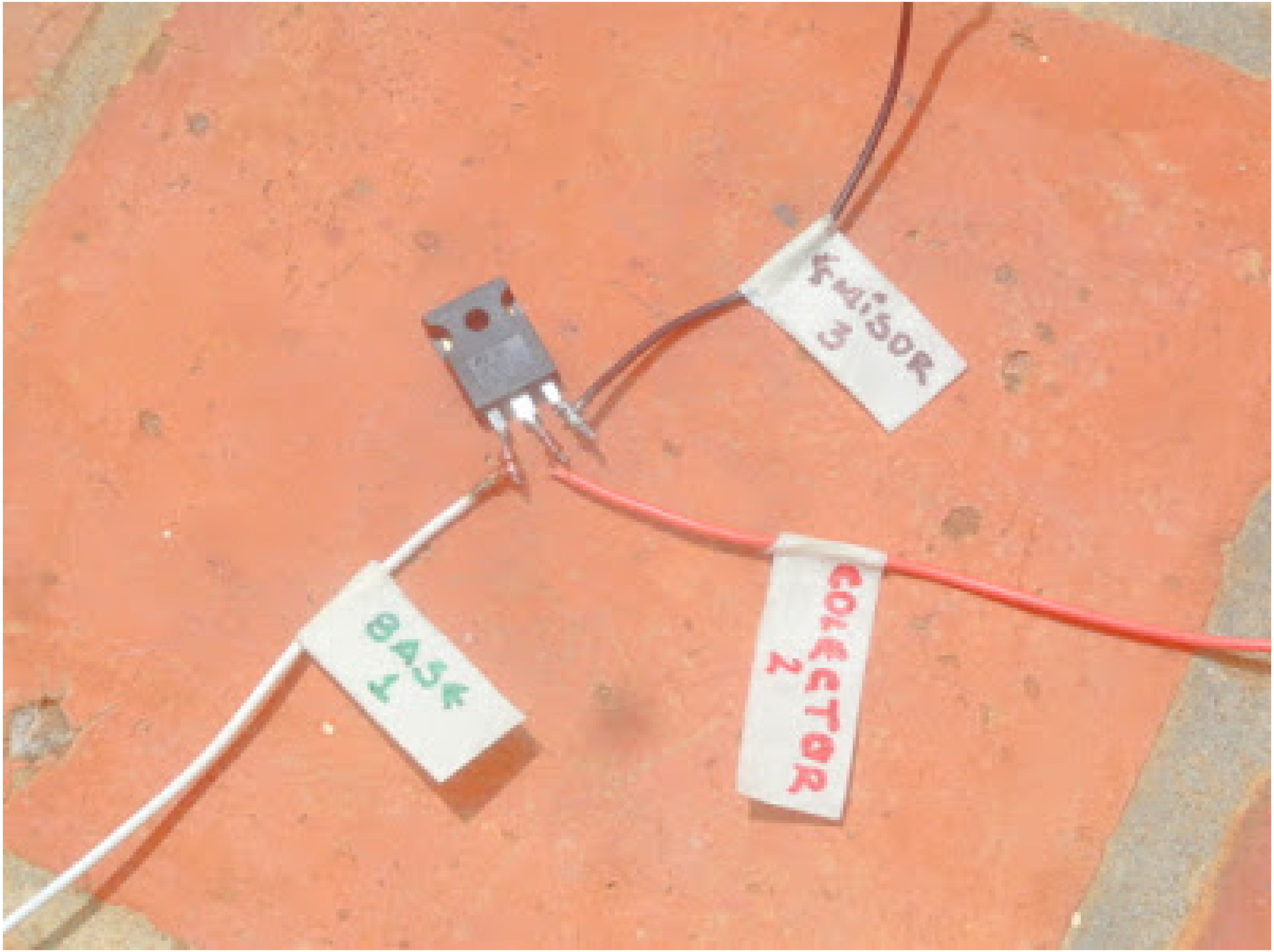
EMISOR
3
BASE
1
COLECTOR
2

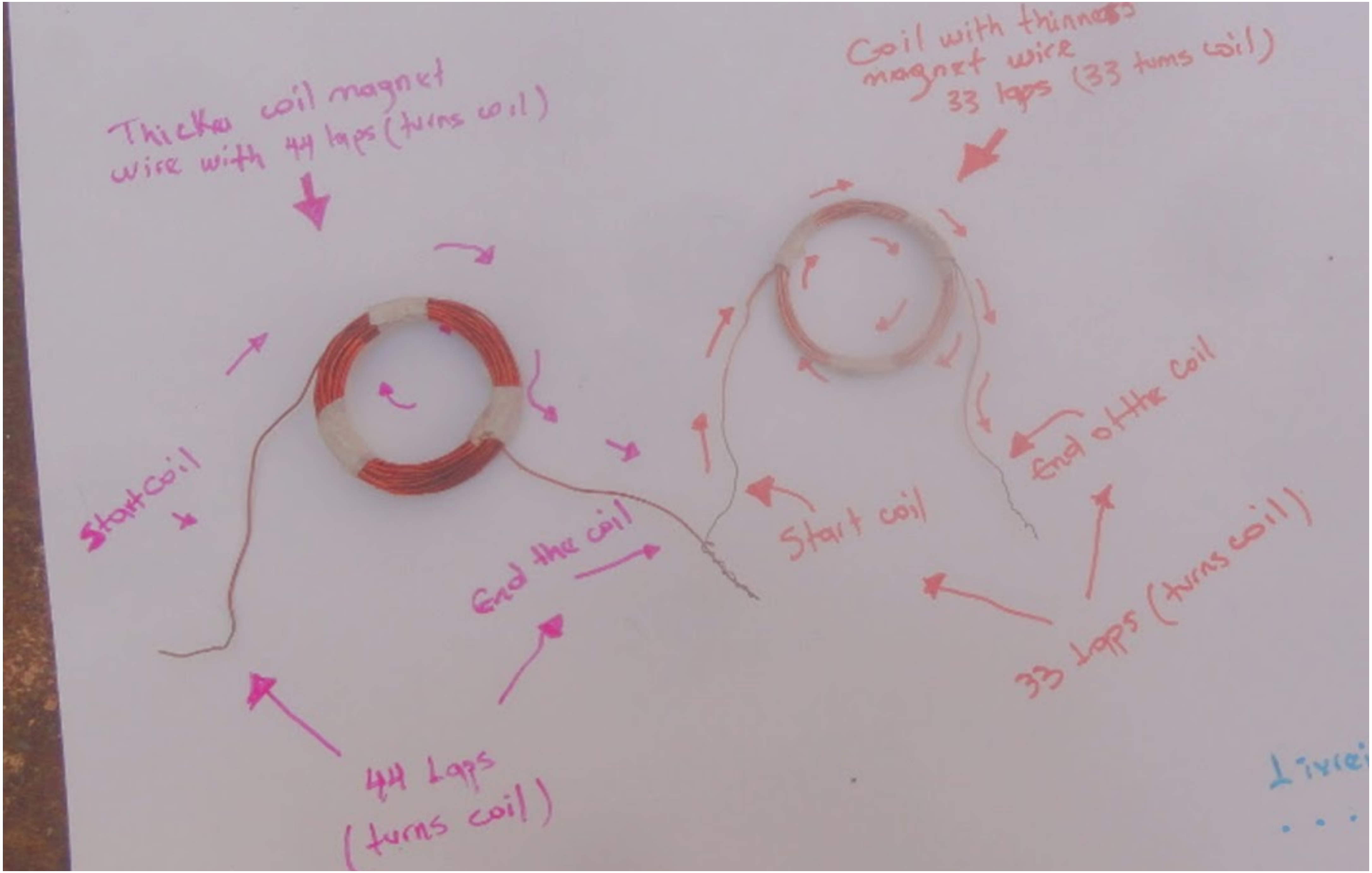
Thicker coil magnet wire with 44 laps (turns coil)
Coil with thinness magnet wire 33 laps (33 turns coil)
Start coil
End the coil
Start coil
End of the coil
44 Laps (turns coil)
33 laps (turns coil)
Livrei
. . .

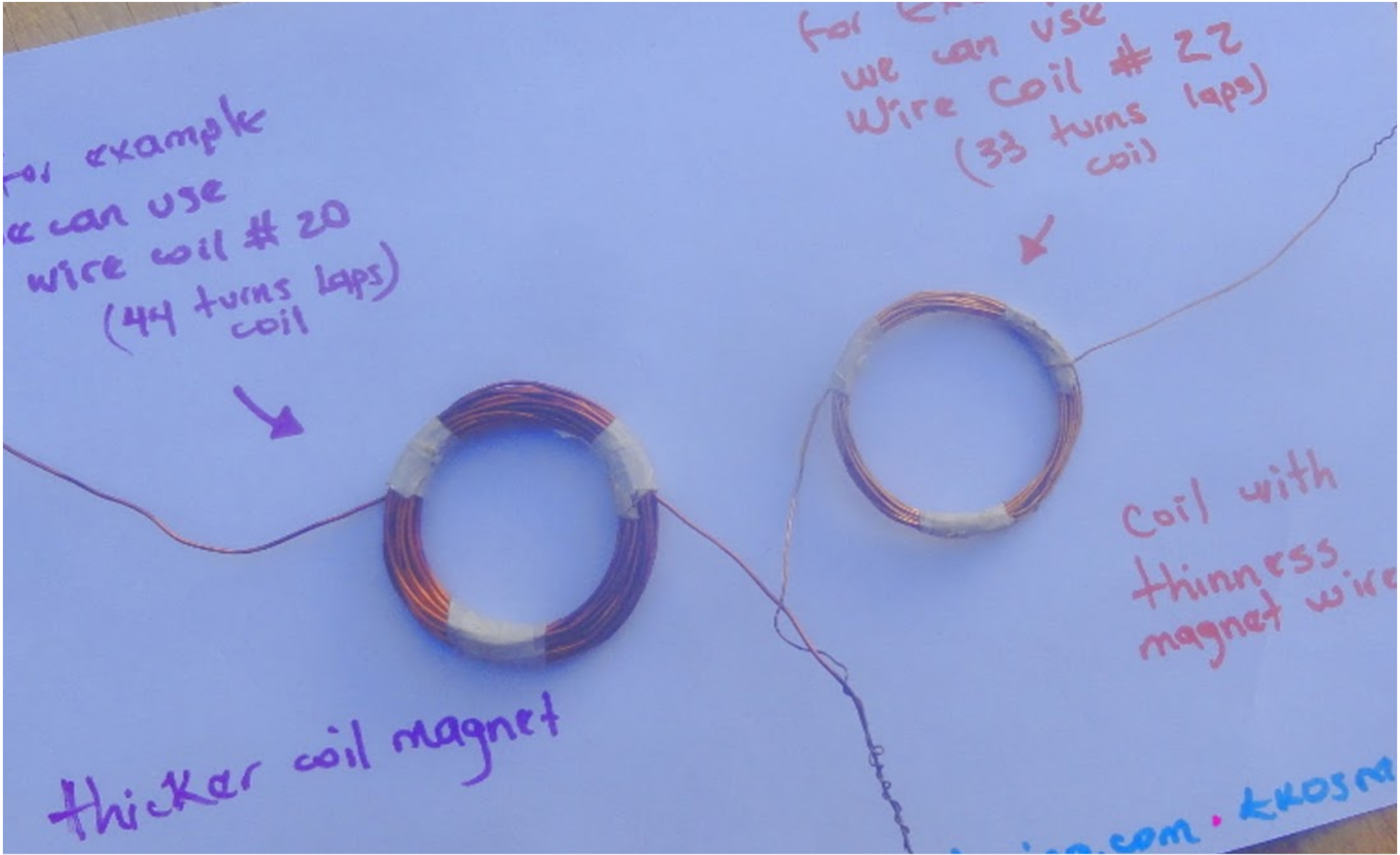
or example
e can use
wire coil # 20
(44 turns laps)
coil
for
we can use
Wire Coil # 22
(33 turns laps)
coil)
Coil with
thinness
magnet wir
thicker coil magnet
.com · kkosm

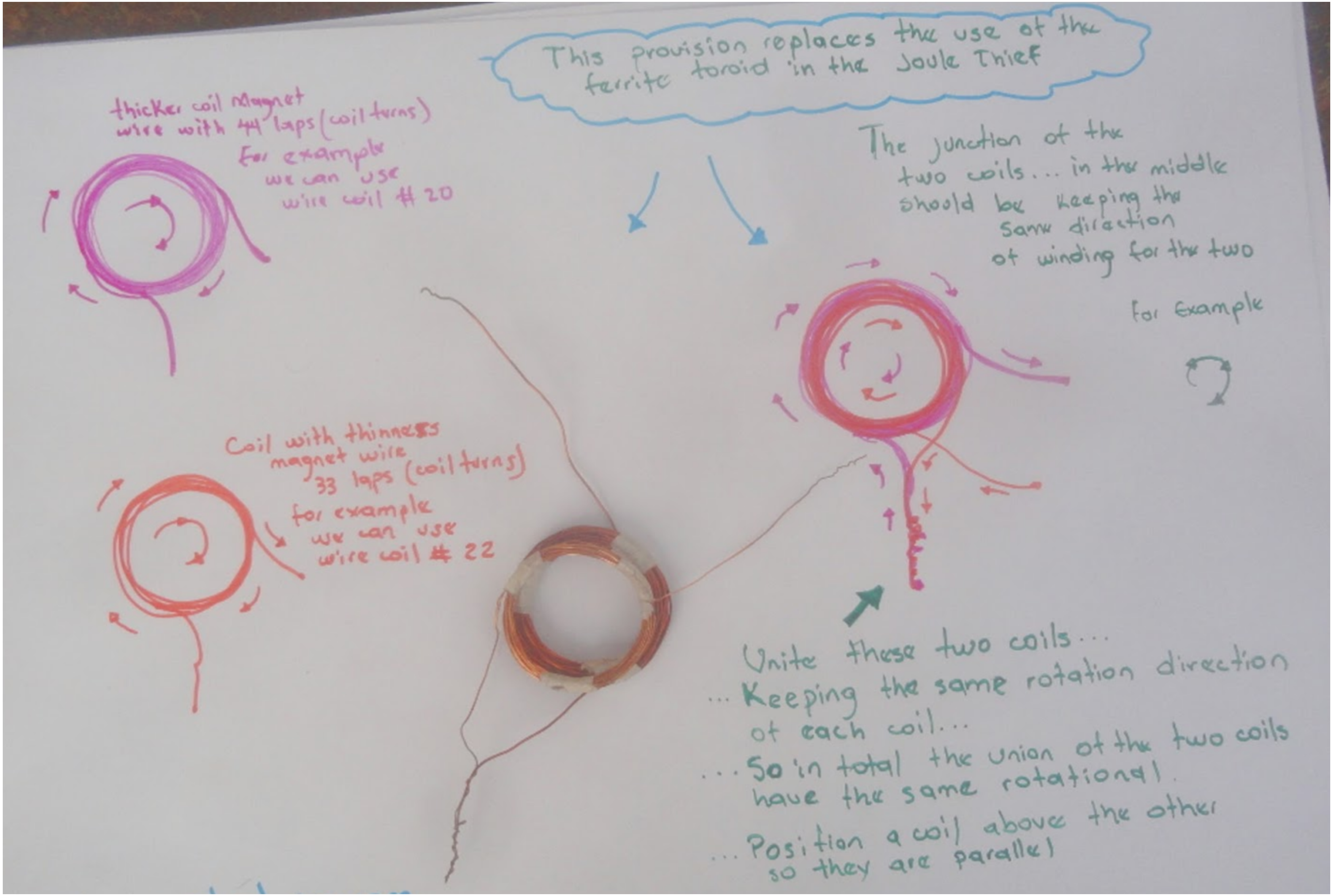

This provision replaces the use of the ferrite toroid in the Joule Thief

thicker coil Magnet wire with 44 laps (coil turns)
for example
we can use
wire coil # 20

The junction of the
two coils... in the middle
should be keeping the
same direction
of winding for the two

for Example

Coil with thinness
magnet wire
33 laps (coil turns)
for example
we can use
wire coil # 22

Unite these two coils...
...Keeping the same rotation direction
of each coil...
...So in total the union of the two coils
have the same rotational.

...Position a coil above the other
so they are parallel

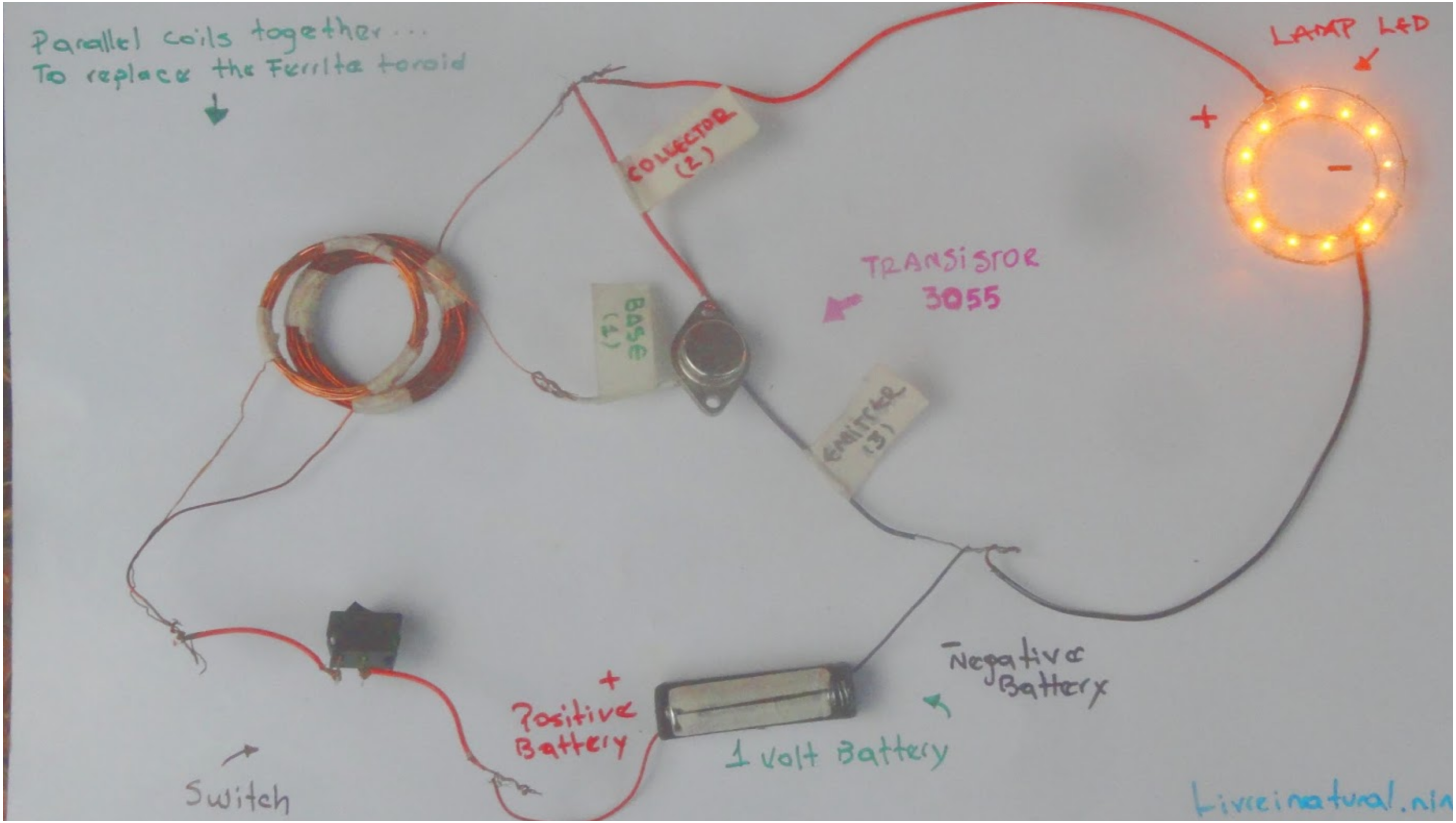
Parallel coils together...
To replace the Ferrite toroid
LAMP LED
+
-
COLLECTOR (2)
BASE (1)
TRANSISTOR
3055
EMITTER (3)
-Negative
Battery
+
Positive
Battery
1 volt Battery
Switch
Livreinatural.ni

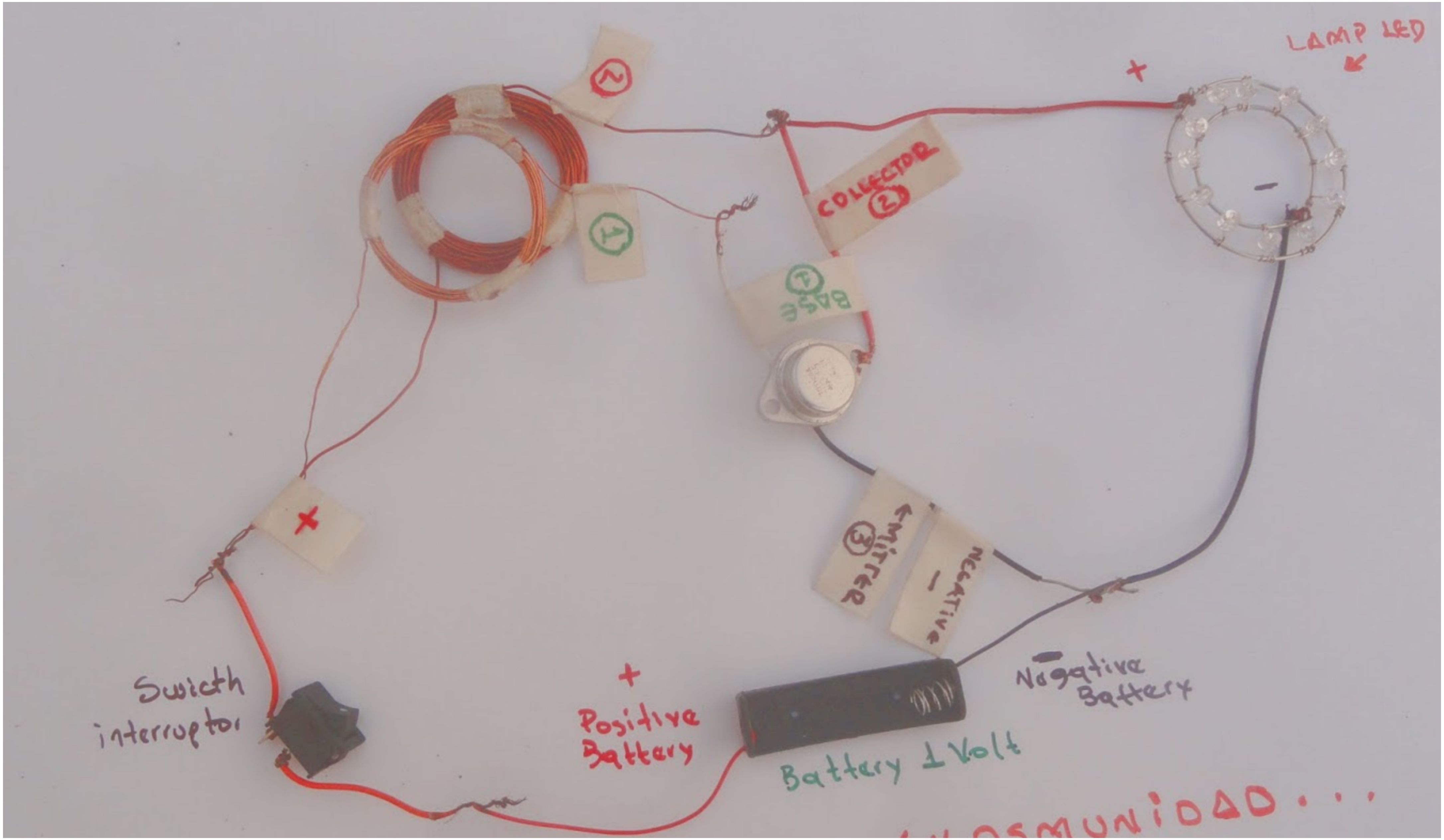
LAMP LED
+
②
①
COLLECTOR ②
−
BASE ①
+
EMITTER ③
NEGATIVE −
Swicth interruptor
+ Positive Battery
Battery 1 Volt
− Negative Battery
OSMUNIDAD...

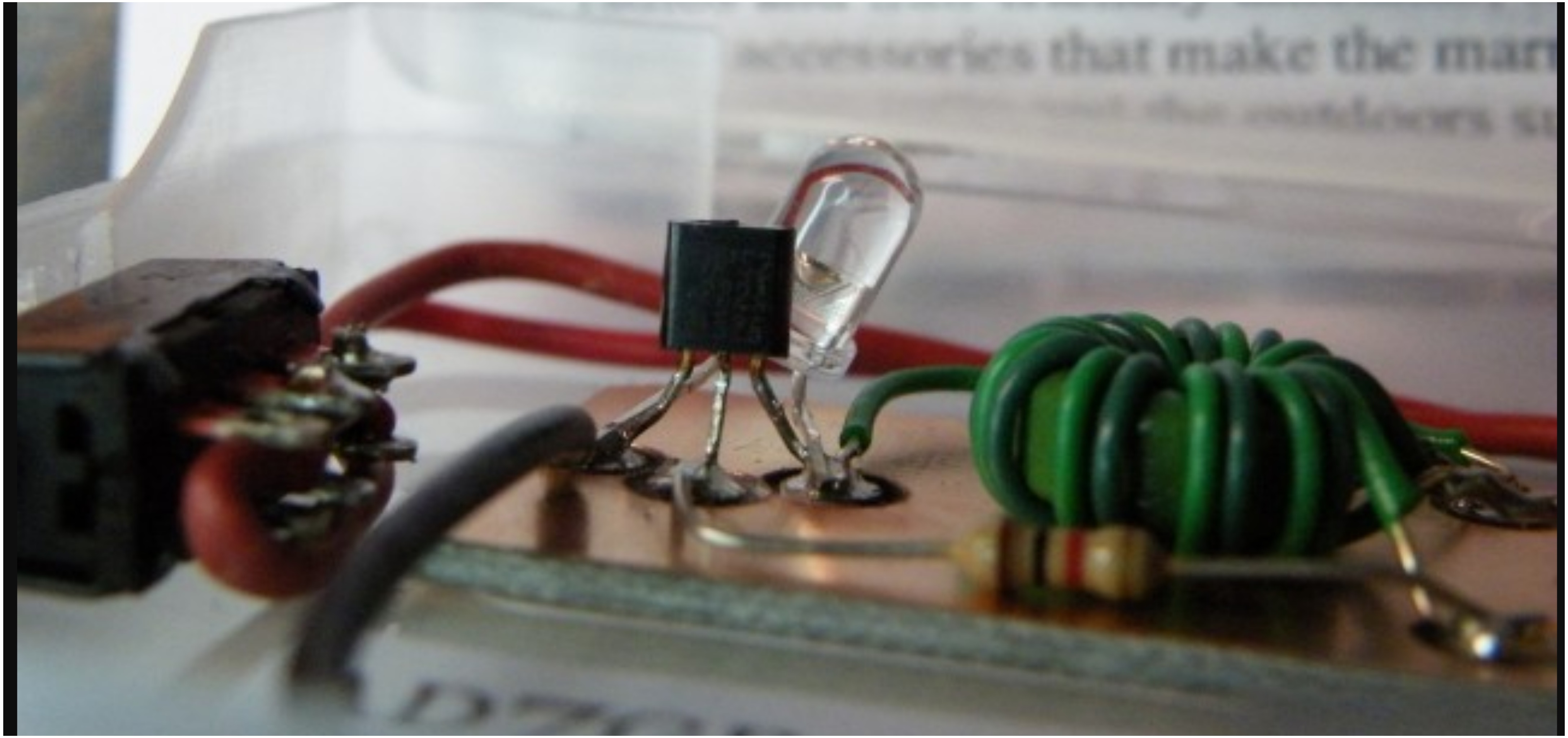
accessories that make the mar

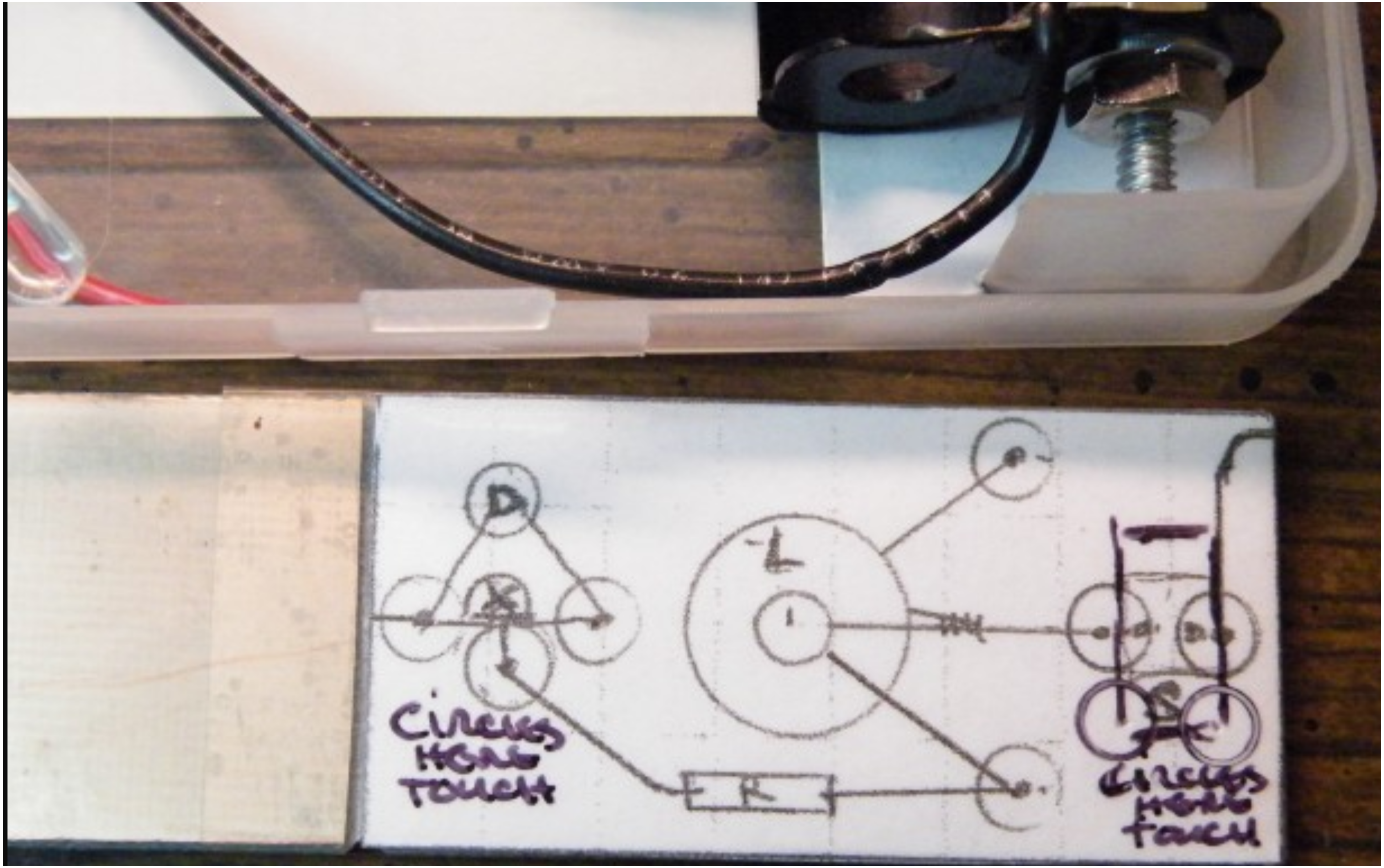
D
X
L
R
S
Circles here touch
Circles here touch

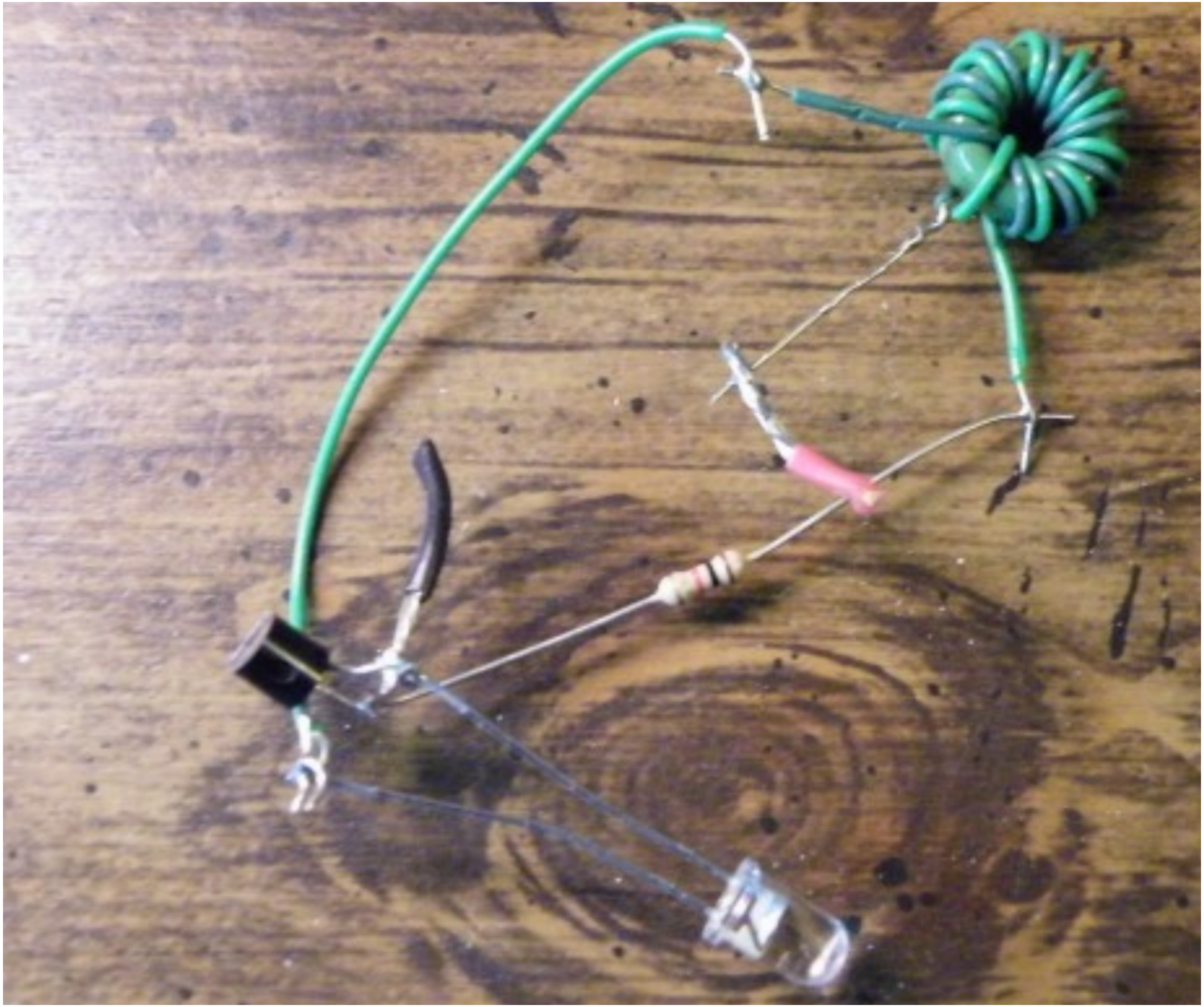

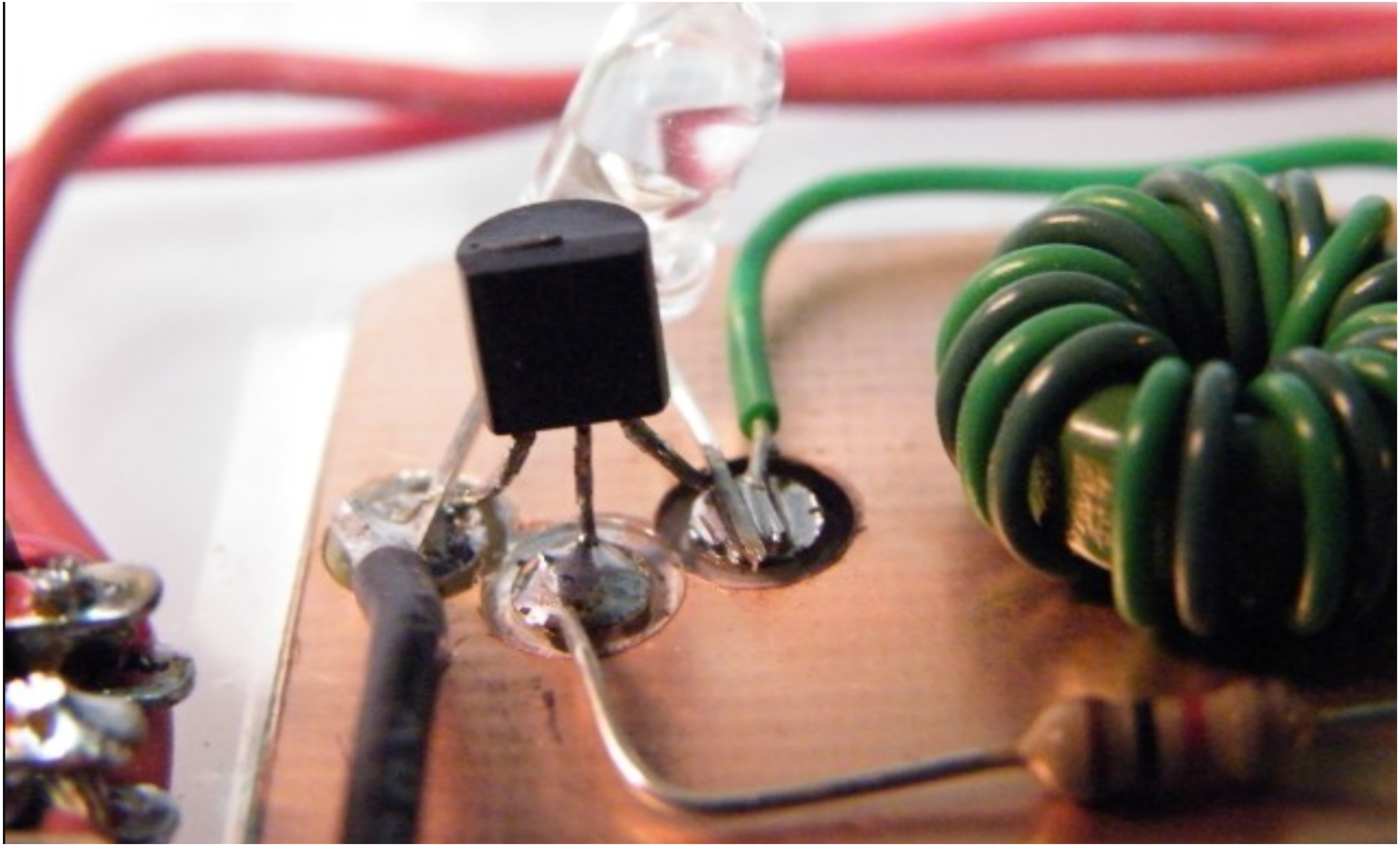

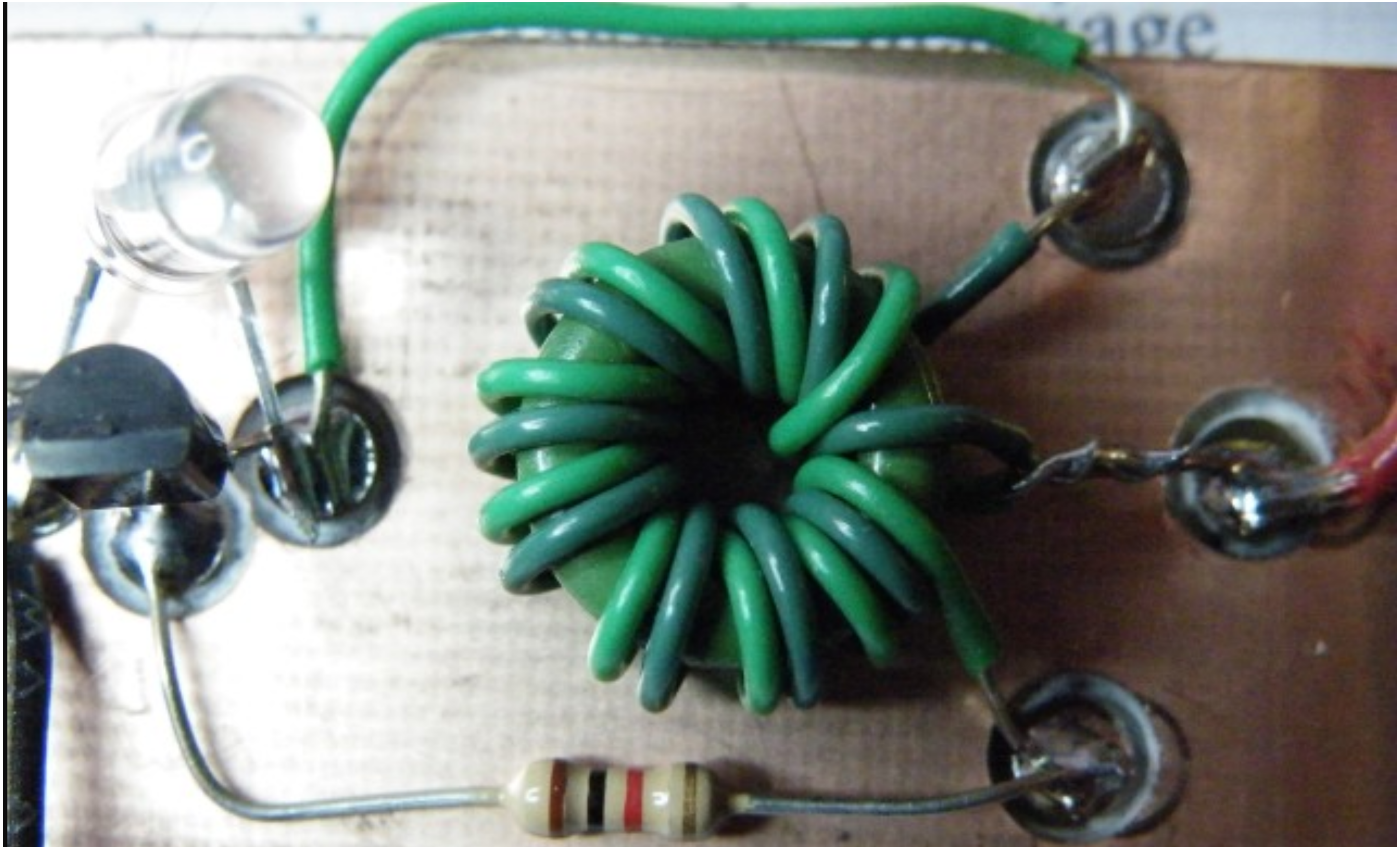
age

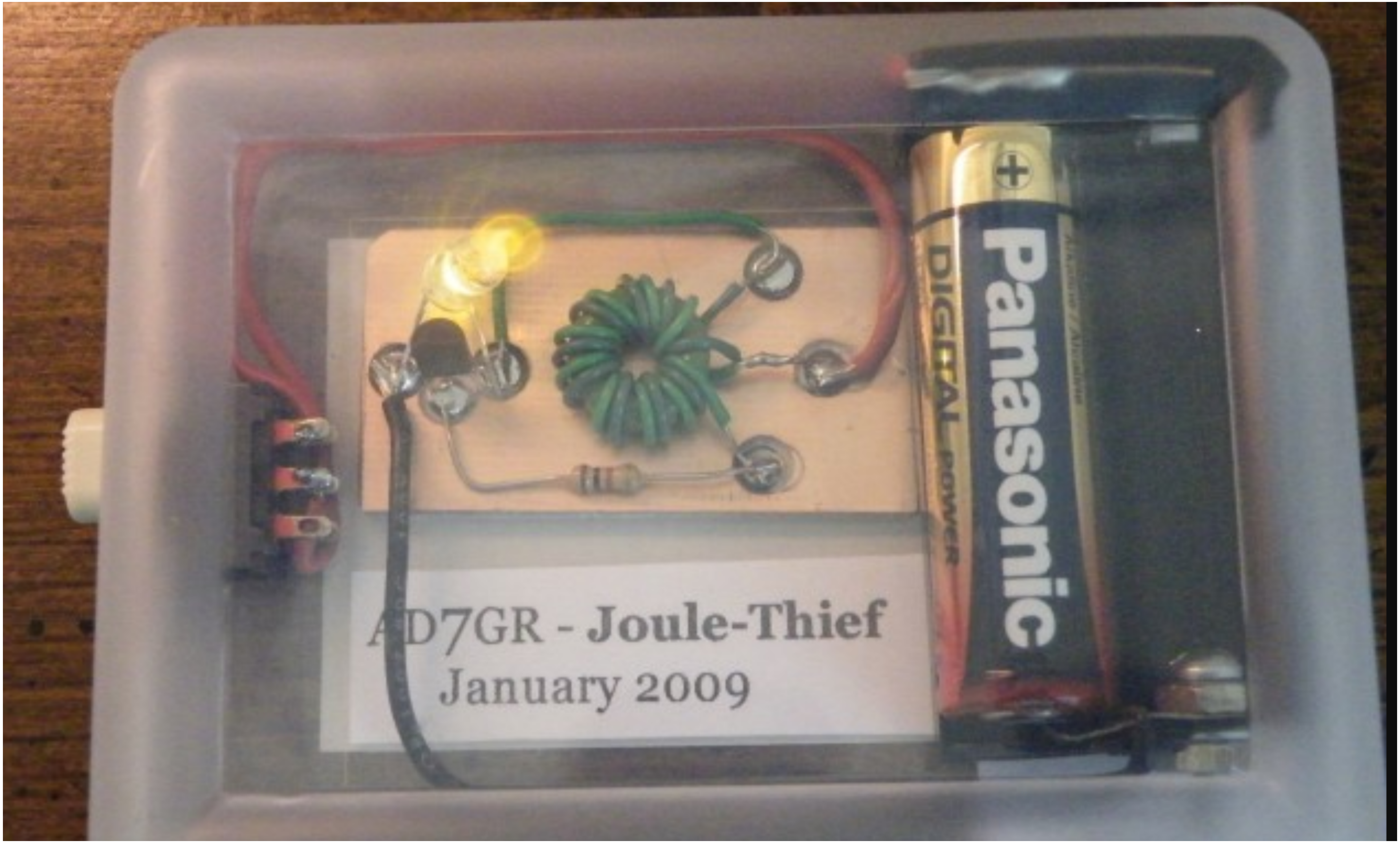
AD7GR - Joule-Thief
January 2009
Panasonic
DIGITAL POWER

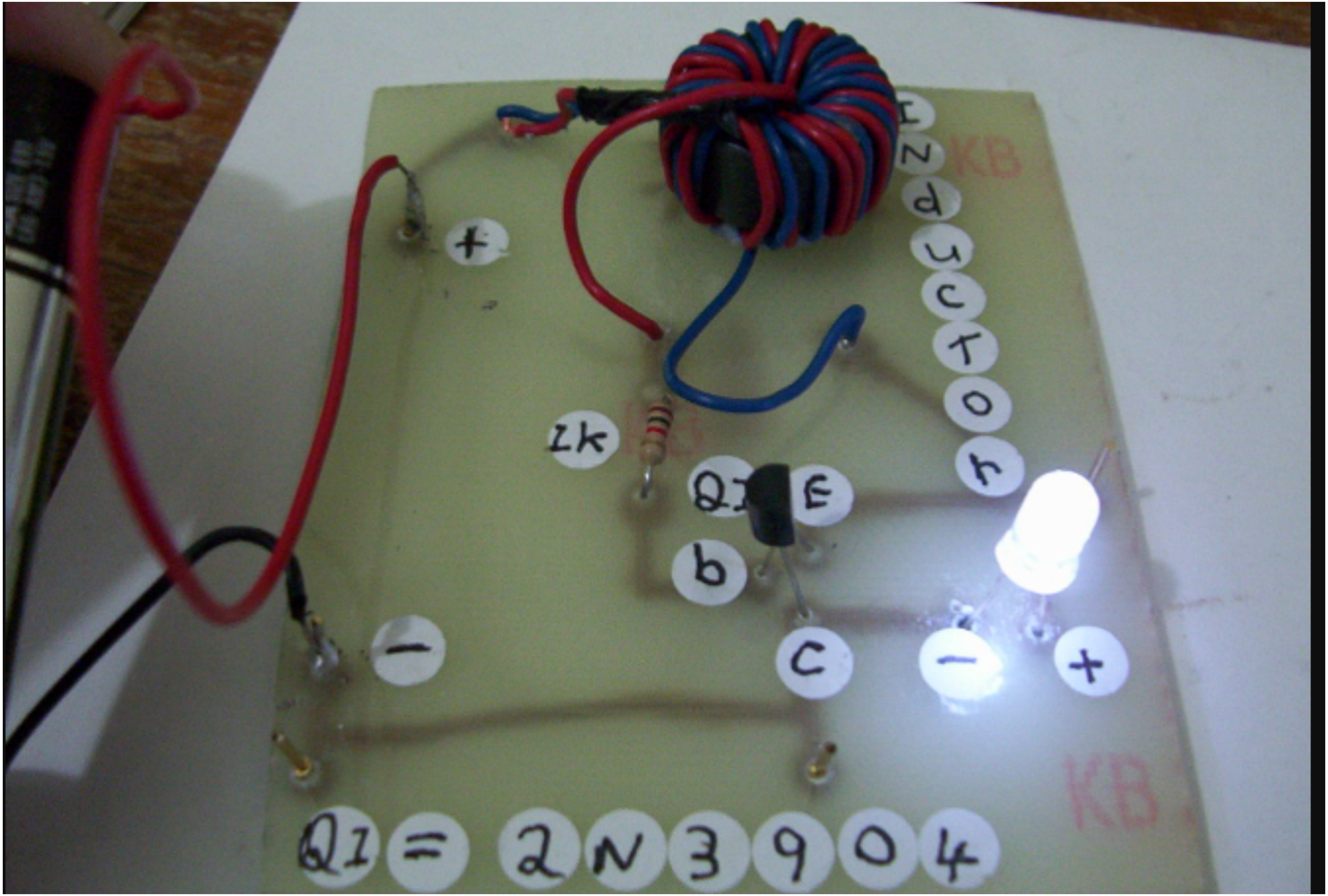
+
Inductor
KB
2k
Q1
E
b
C
−
−
+
KB
Q1 = 2N3904

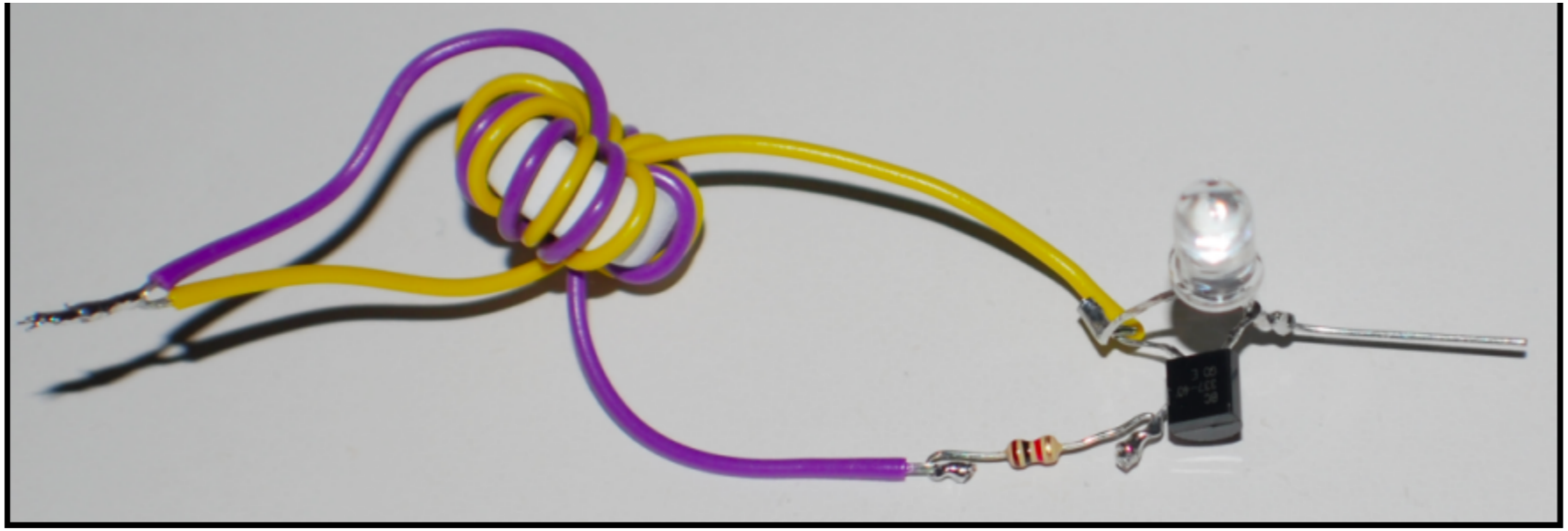

… and these are the components you need:

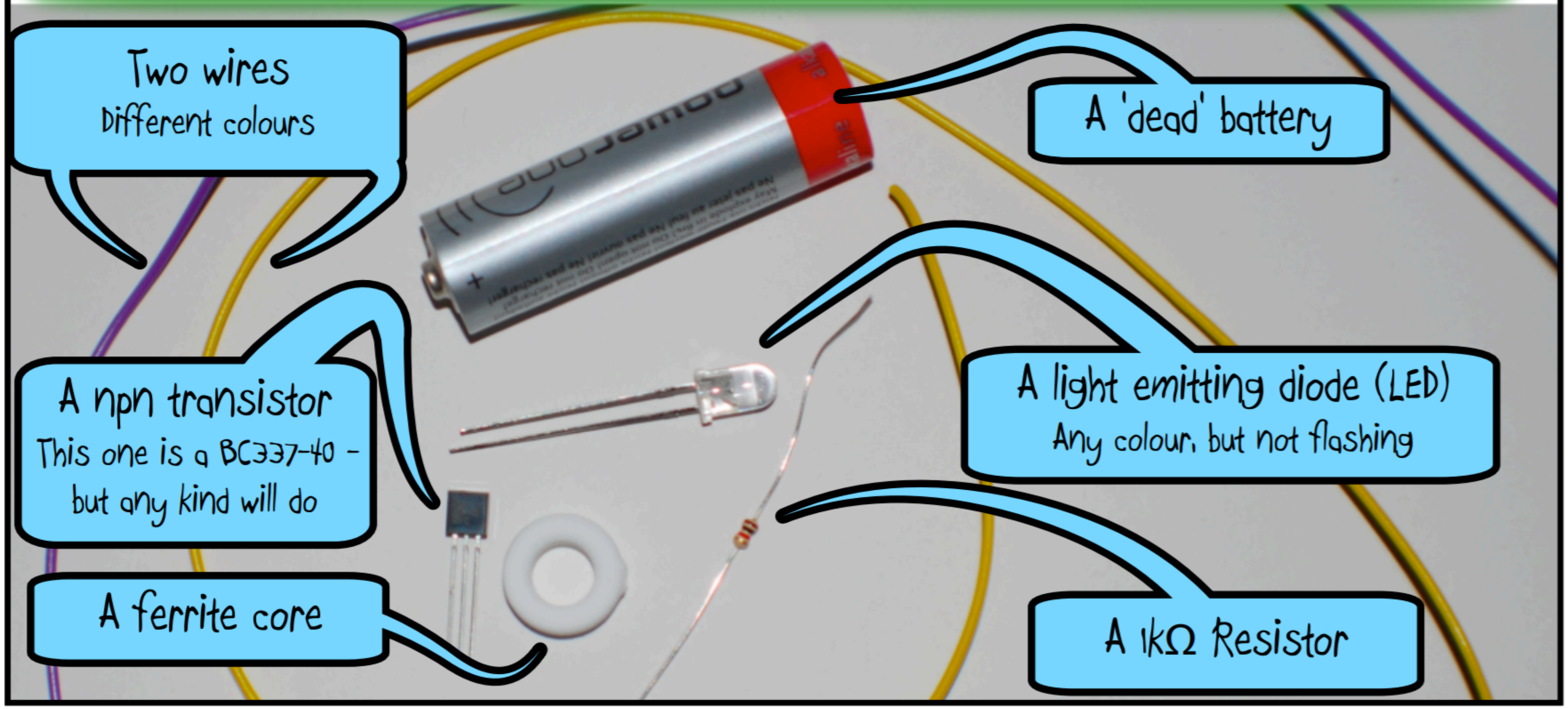
Two wires
Different colours
A 'dead' battery
A npn transistor
This one is a BC337-40 – but any kind will do
A light emitting diode (LED)
Any colour, but not flashing
A ferrite core
A 1kΩ Resistor

Solder the 1kΩ Resistor to either of the unsoldered wires coming from the inductor.

Solder the other end of the resistor to the base lead of the transistor.

On the transistor used here, that is the middle lead.

Cut off the spare ends of leads.

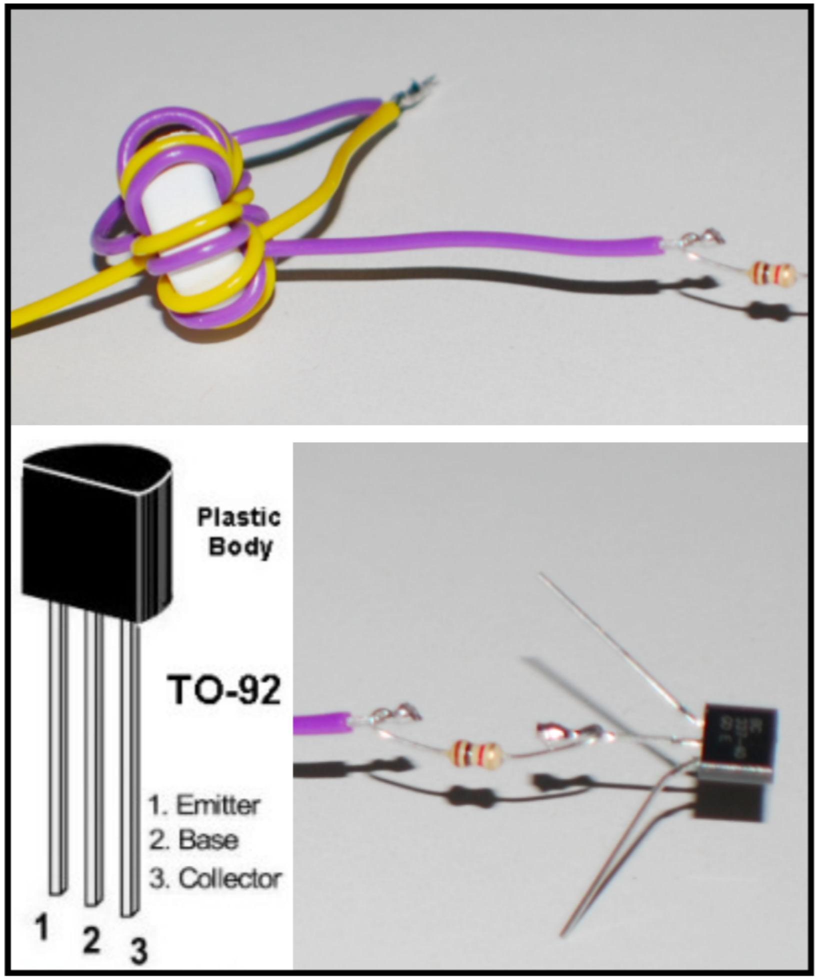
Plastic Body
TO-92
1. Emitter
2. Base
3. Collector
1
2
3

Solder the anode of the LED (the longer leg) to the collector lead of the transistor AND the remaining wire from the inductor.

Cut off the ends of the leads.

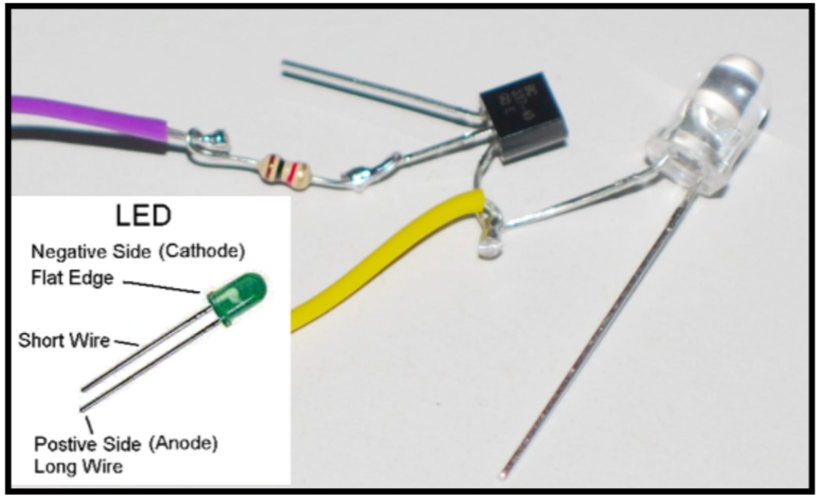
LED
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

Solder the other leg of the LED (the cathode) to the emitter lead of the transistor.
Do NOT cut off the LED leg.
(Don't wory if you have cut it off - you'll just need to solder another wire to this leg so that you can connect it to the battery.)

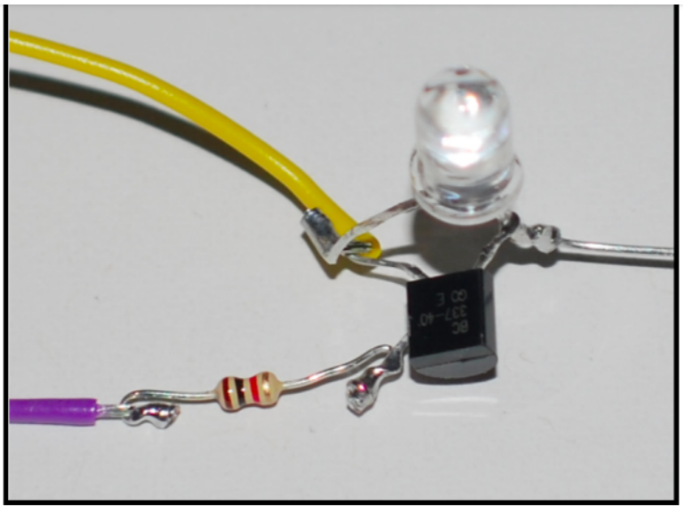
BC
337-40
00 E

That's the circuit finished. Now use it.

First you need a 'dead' battery - that means one where the voltage is less than 1.3V.
You can see that the battery used here has a Voltage of just over 1V.

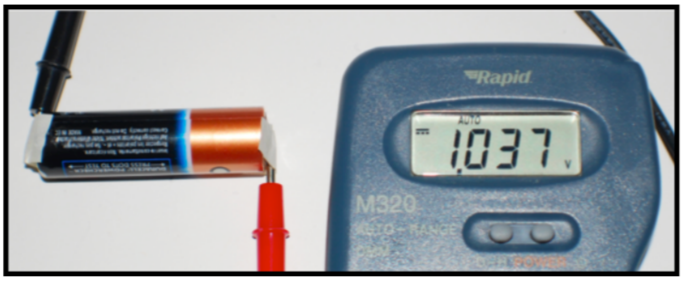
Rapid
AUTO
1.037 V
M320
AUTO - RANGE
POWER

Connect the two soldered wires coming from the inductor to the positive (+) side of the battery.
Connect the uncut LED leg (the athode) to the negative (-) side of the battery.

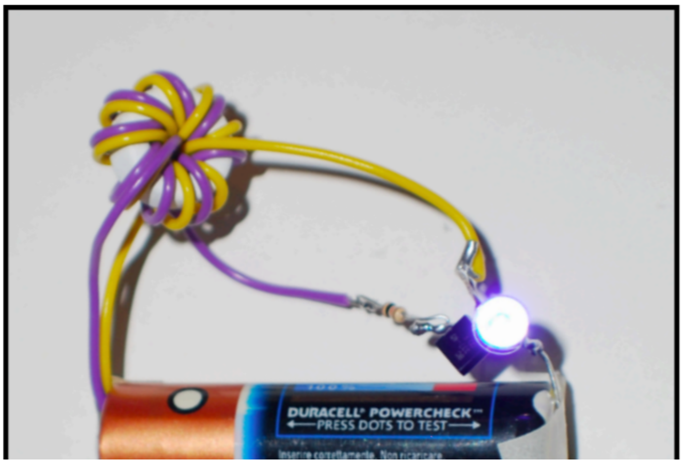
DURACELL POWERCHECK
PRESS DOTS TO TEST

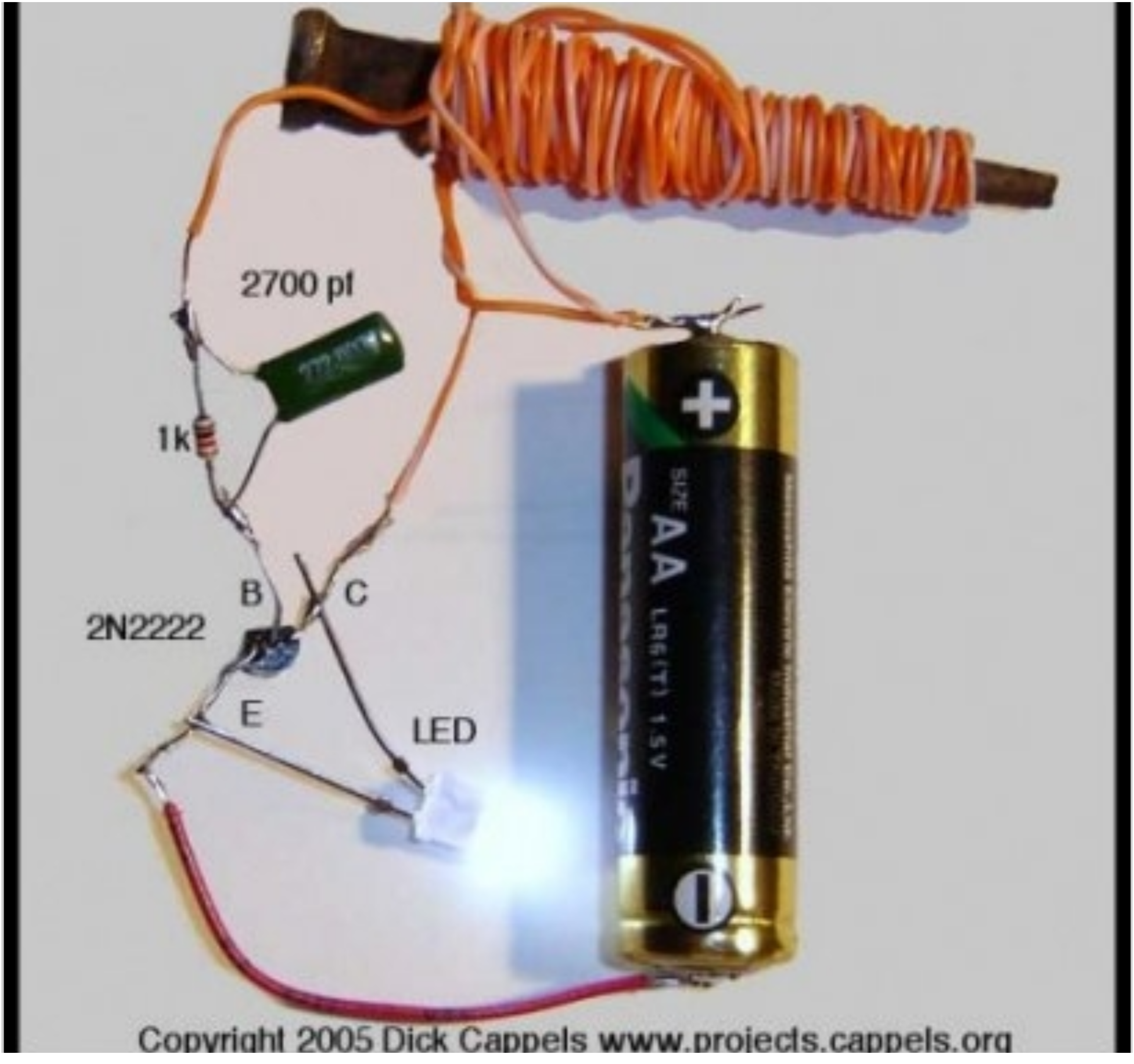
2700 pf
1k
B
C
2N2222
E
LED
SIZE AA LR6(T) 1.5V
Copyright 2005 Dick Cappels www.projects.cappels.org

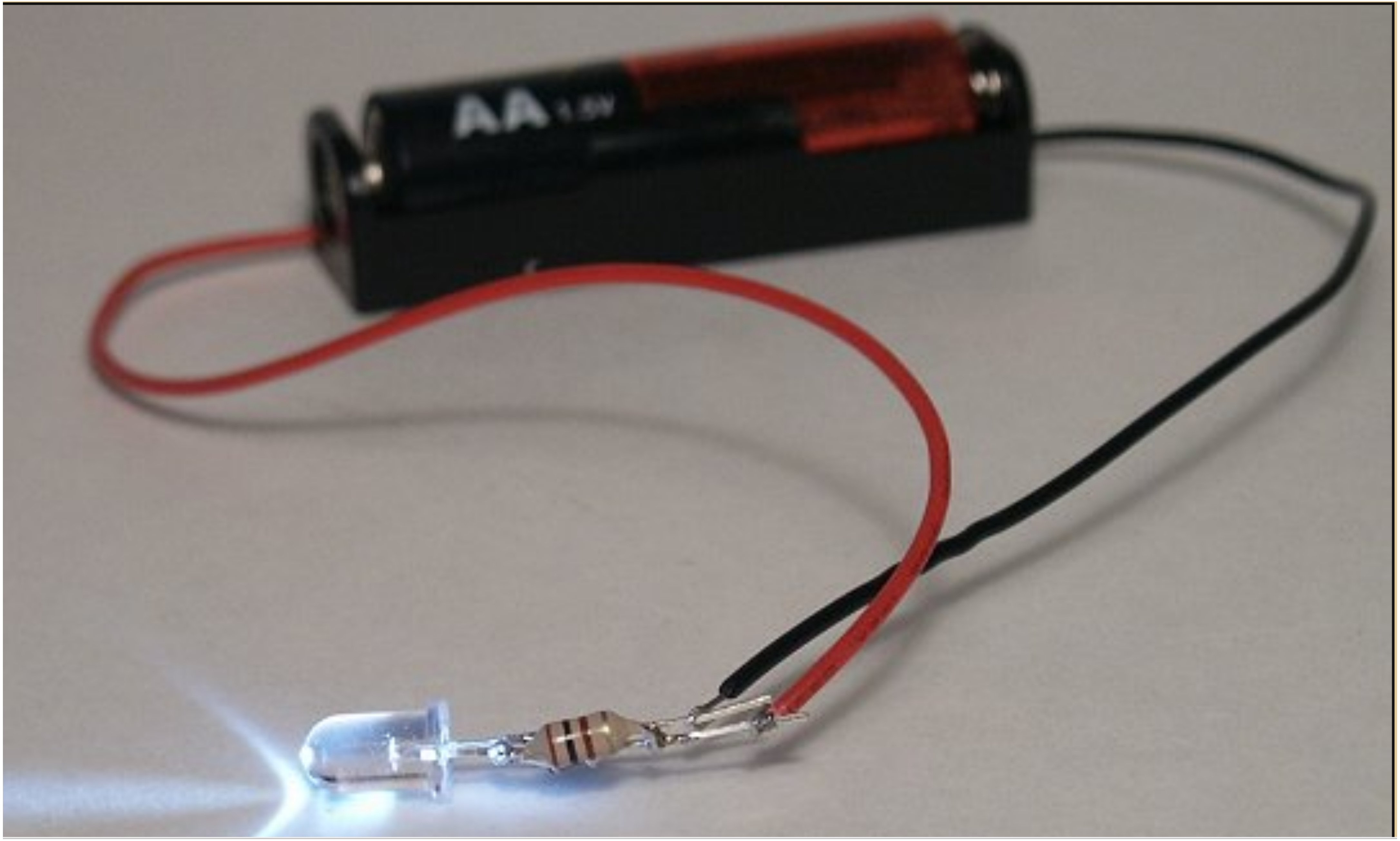
AA 1.5V

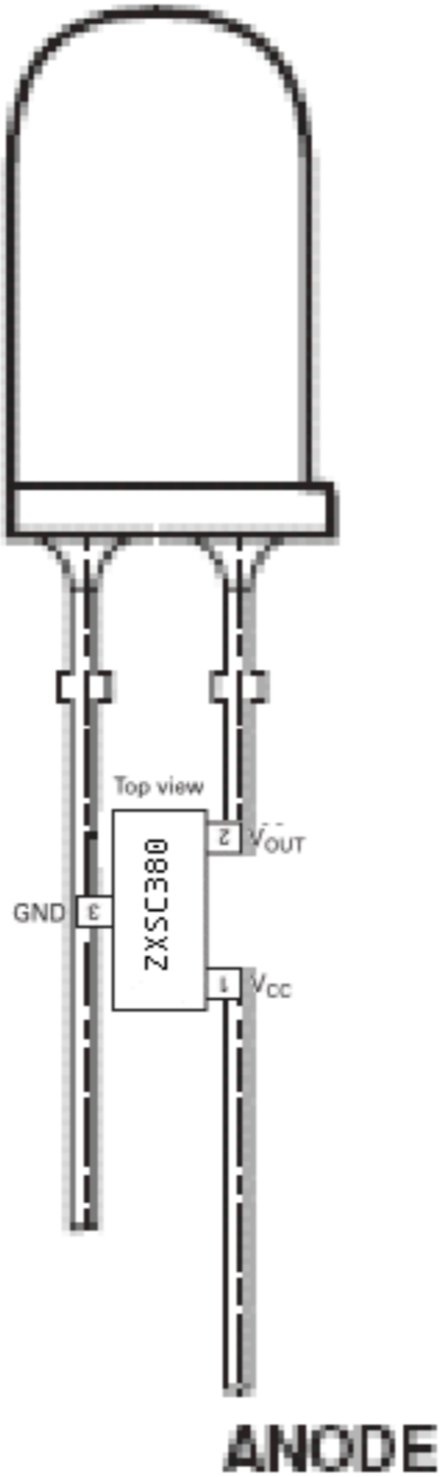

The chip is soldered onto the legs of the LED and then the wire between pins 1 and 2 is cut out so there is no connection. The inductor is then soldered so as to bridge pins 1 and 2. I soldered the inductor on before cutting out the wire but it might be easier to try some other method.

As I said previously, this is a tedious solder job, and the SOT23 IC is easily damaged (ask me how I know!) It takes patience and a steady hand to assemble these few parts.

This circuit would work for the LED indicator in the 1.5v boosters and would most likely be a better choice than the Joule Thief that I posted previously.

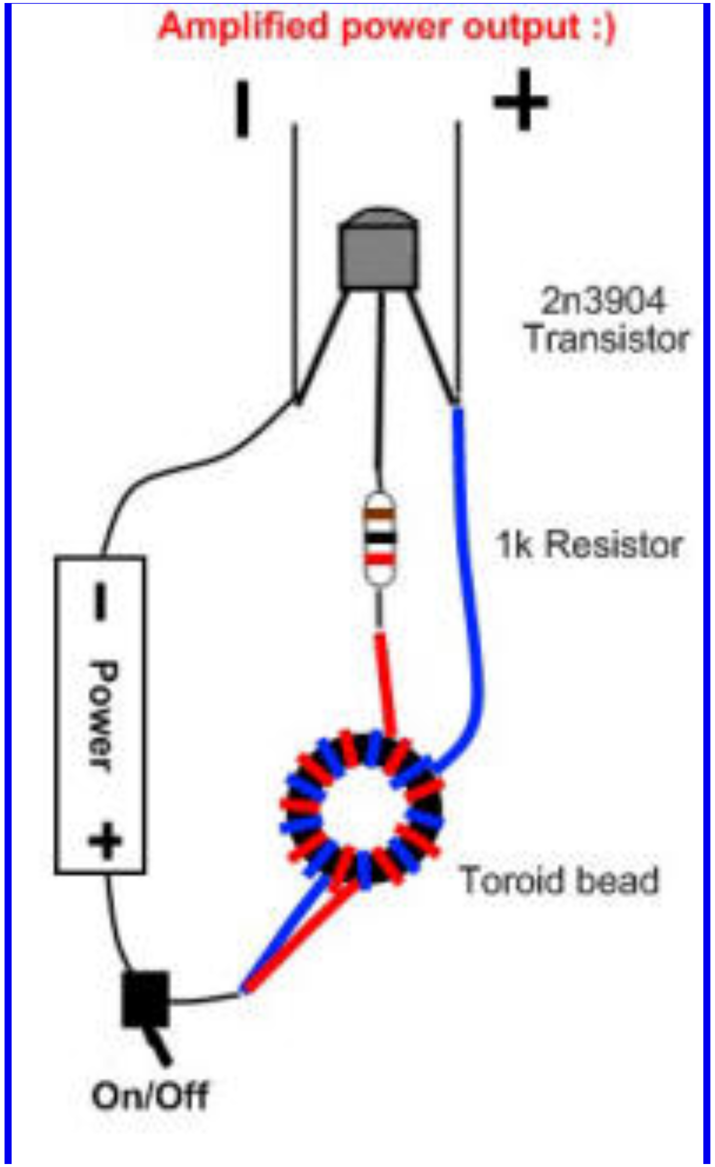
Amplified power output :)
-
+
2n3904
Transistor
1k Resistor
-
Power
+
Toroid bead
On/Off

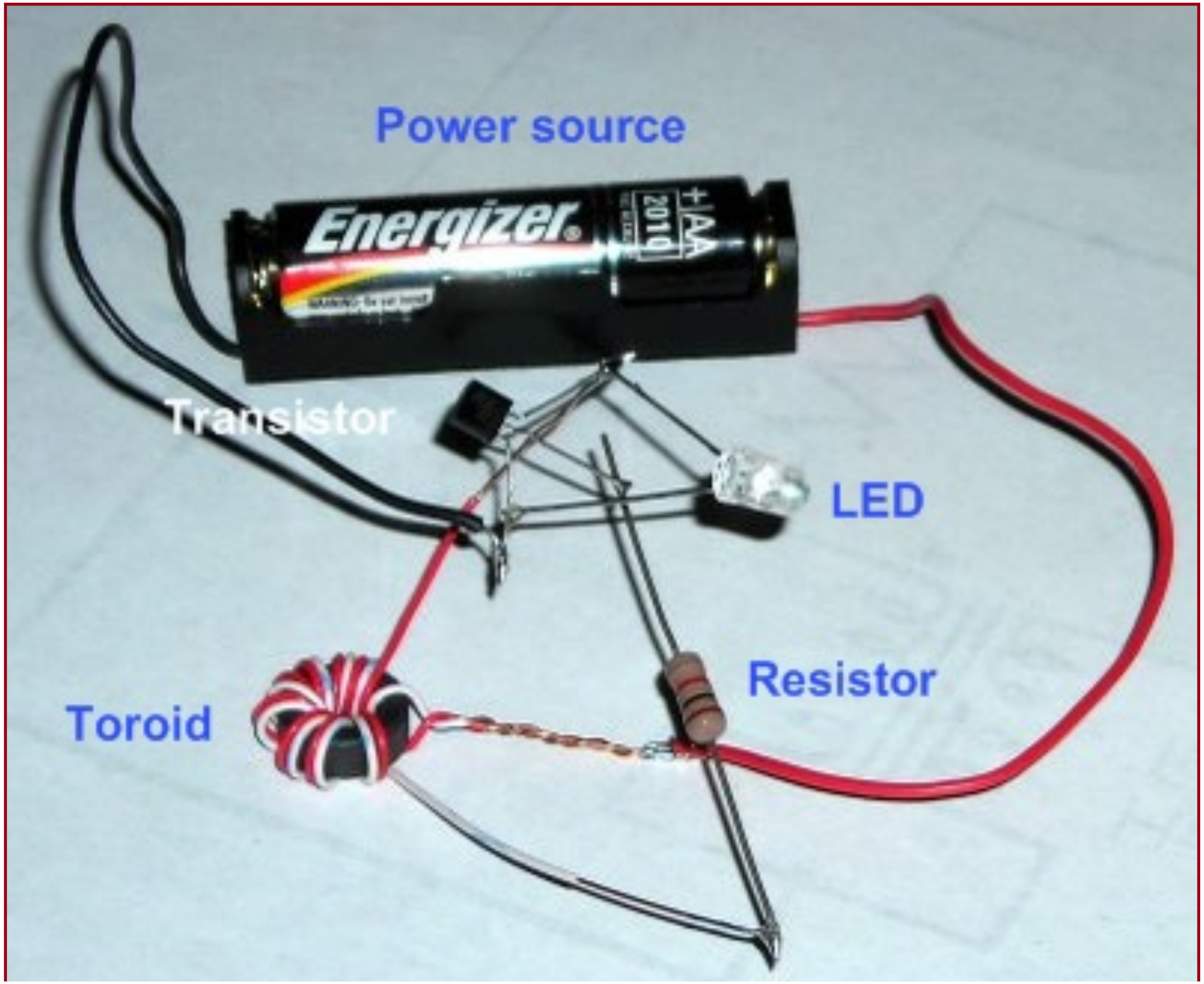
Power source
Transistor
LED
Resistor
Toroid

The CIRCUIT

The circuit is very simple. All the work is done by the 5252F chip.
It contains an oscillator, a high speed diode and a power transistor. All these components inside the "IC" that looks like a 4-leaded transistor!
This IC is smaller and cheaper than all the parts individually and is less expensive than the competition (that costs 70 cents).
We can produce a complete project for a few dollars and it has two "test features."
You can place a LED across one of the LEDs on the board and find out the colour as many LEDs come in a "clear-as-glass" package and you cannot tell the colour until they are illuminated.
The other feature is INDUCTOR TESTING.
The current taken by the circuit changes according to the value of the inductor.
You just need a few reference values and you can work out the value of an inductor within the range of the inductors you have used as samples, or slightly higher or lower values.

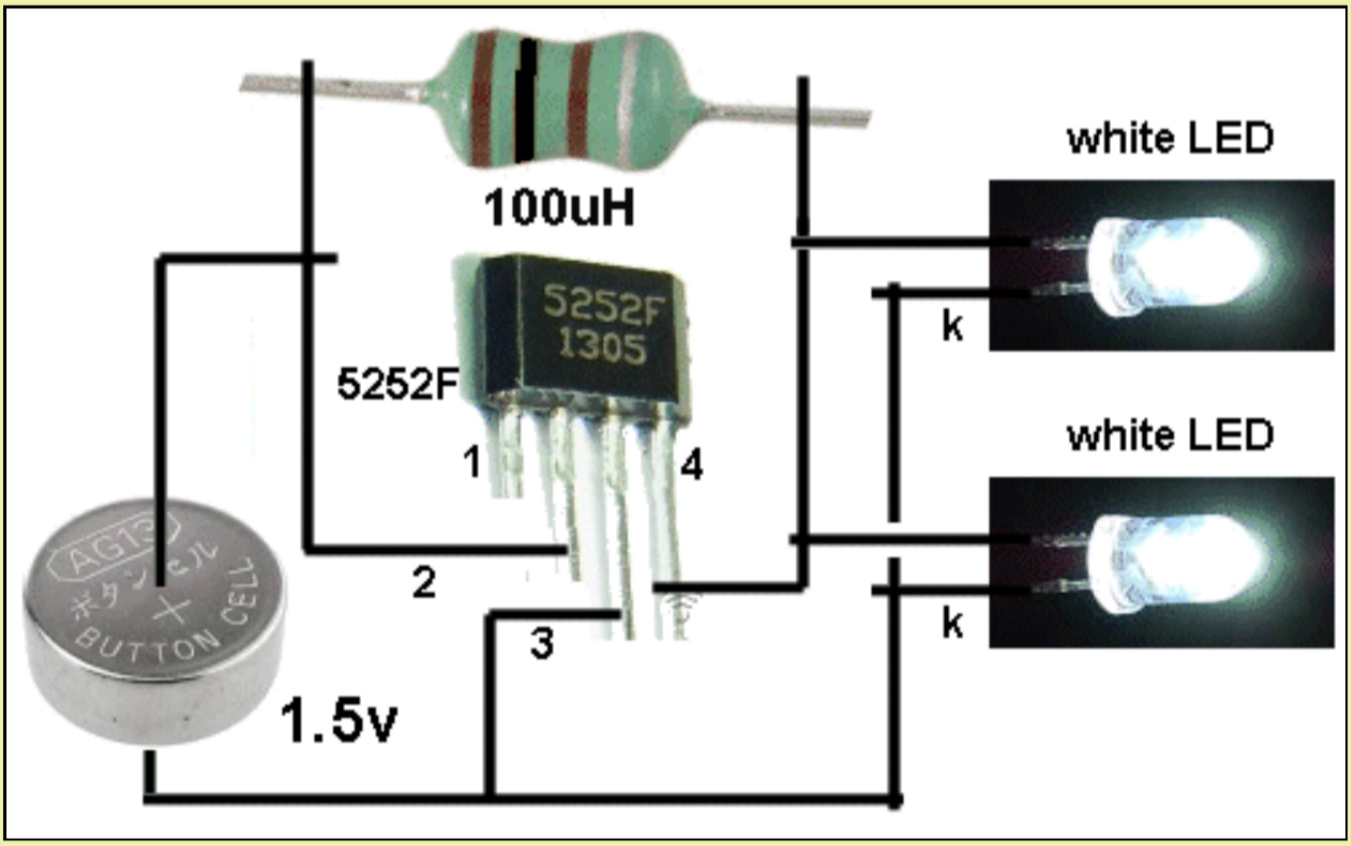

The layout using component pictures.

Using a 220uH, the circuit takes 13mA and illuminates 2 white LEDs very brightly.
Using 100uH the circuit takes 30mA and the LEDs are really the same brightness.
Using 33uH the circuit takes 80mA and the LEDs are just about the same brightness.
Obviously the 220uH creates the most efficient circuit.
The QX5252F is capable of delivering more than 100mA to the LEDs but we only need

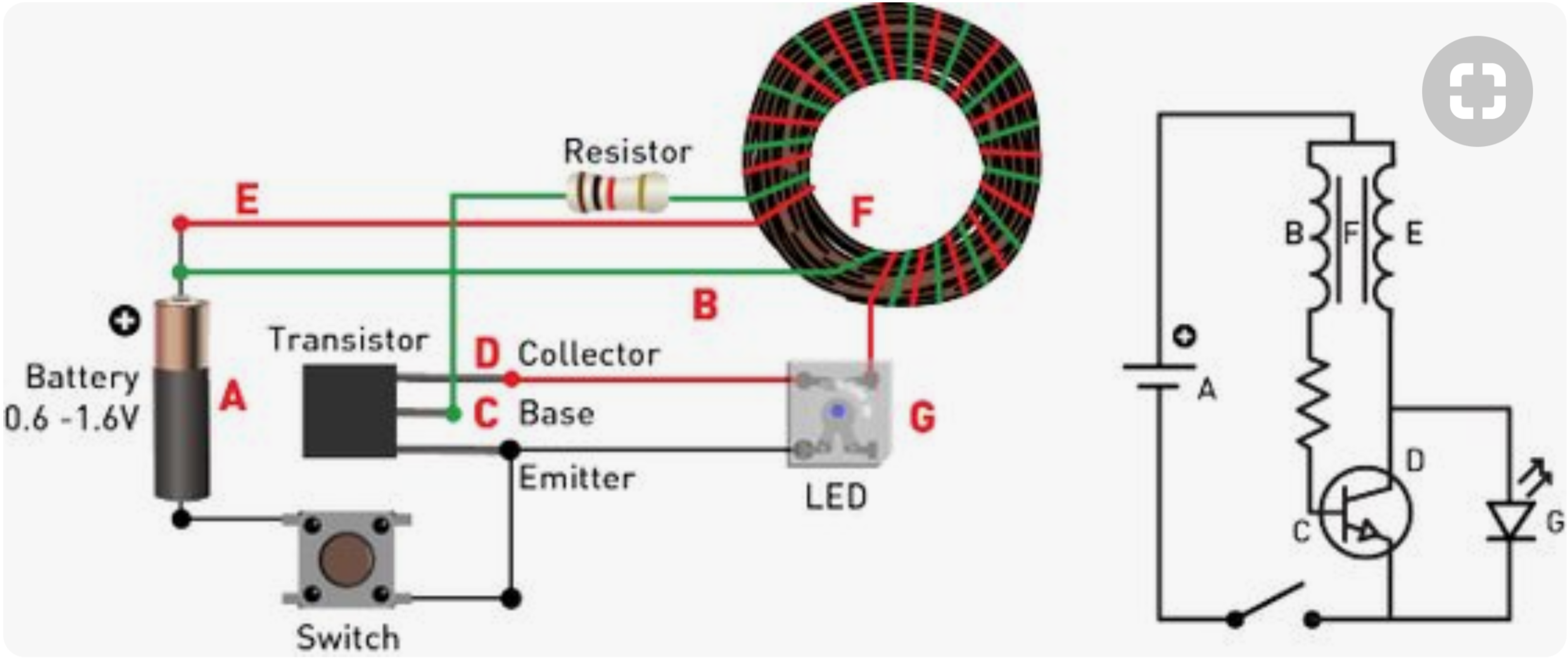
Resistor
E
F
B
⊕
Battery
0.6 -1.6V
A
Transistor
D Collector
C Base
Emitter
LED
G
Switch
B
F
E
⊕
A
D
C
G

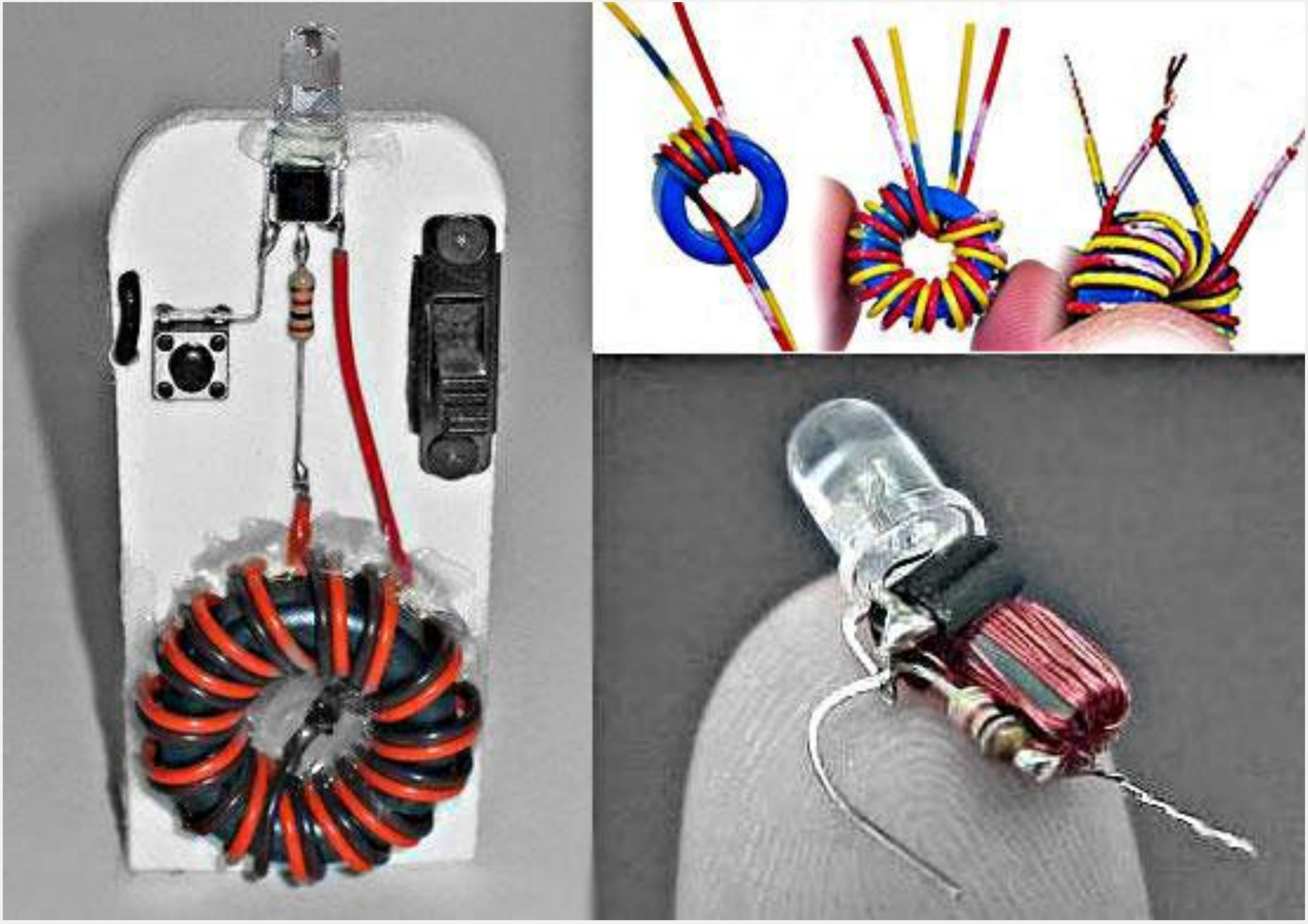

Watson's 2 Transistor "Joule Thief" V. Booster

Uses 2N3904 for output. Coil is 180 microhenry, 2.5 ohm RF choke (green blob)

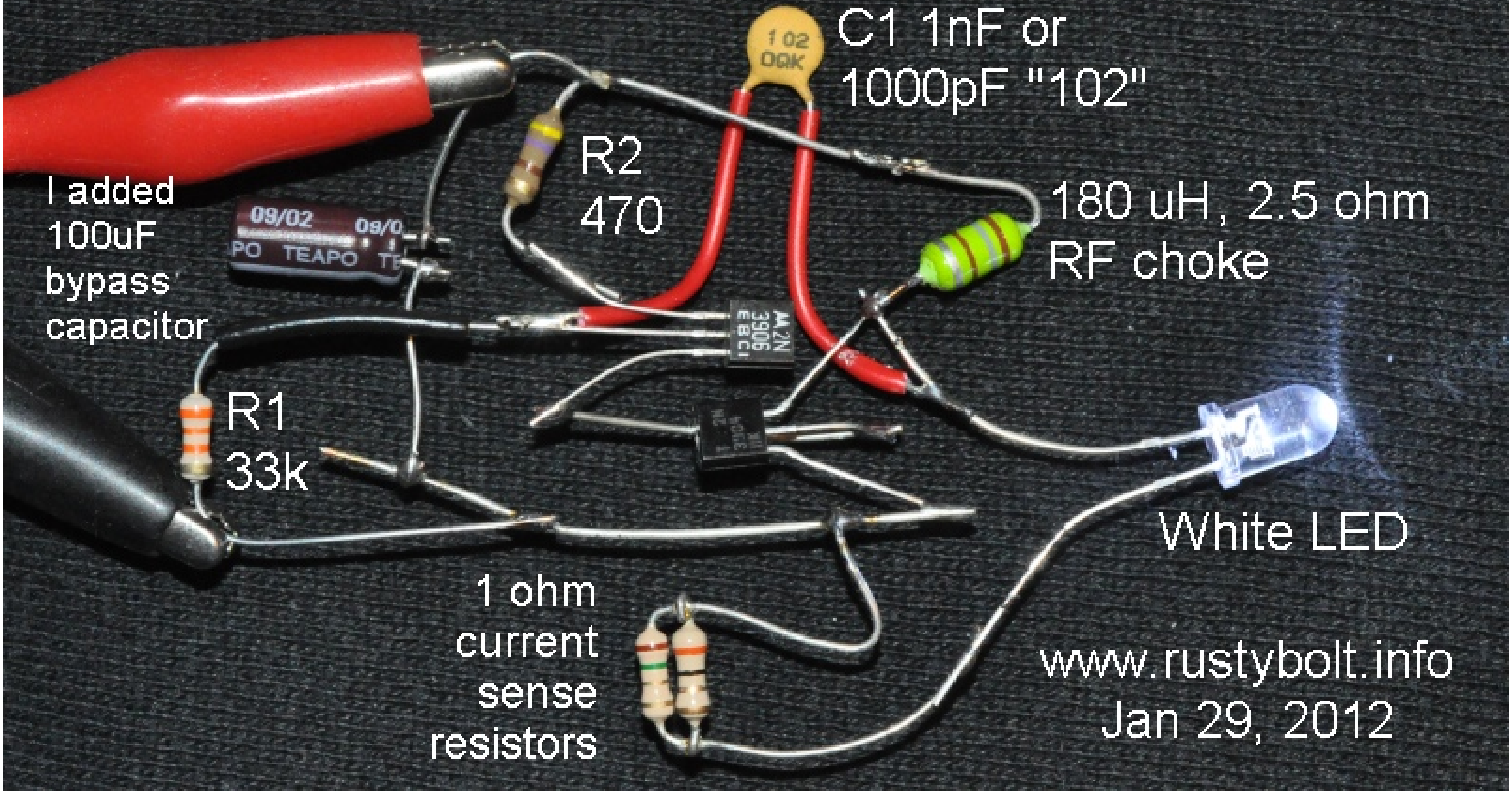
C1 1nF or 1000pF "102"
R2 470
180 uH, 2.5 ohm RF choke
I added 100uF bypass capacitor
R1 33k
White LED
1 ohm current sense resistors
102
09/02
TEAPO
2N 3906

www.rustybolt.info
Jan 29, 2012

Watson's Version of "1.5v Joule Thief -Blinks Led"

Taken from http://www.youtube.com/watch?v=GVP2QGpk5KE Nov 14, 2012

My version does boost the voltage and light the LED, but it doesn't blink.

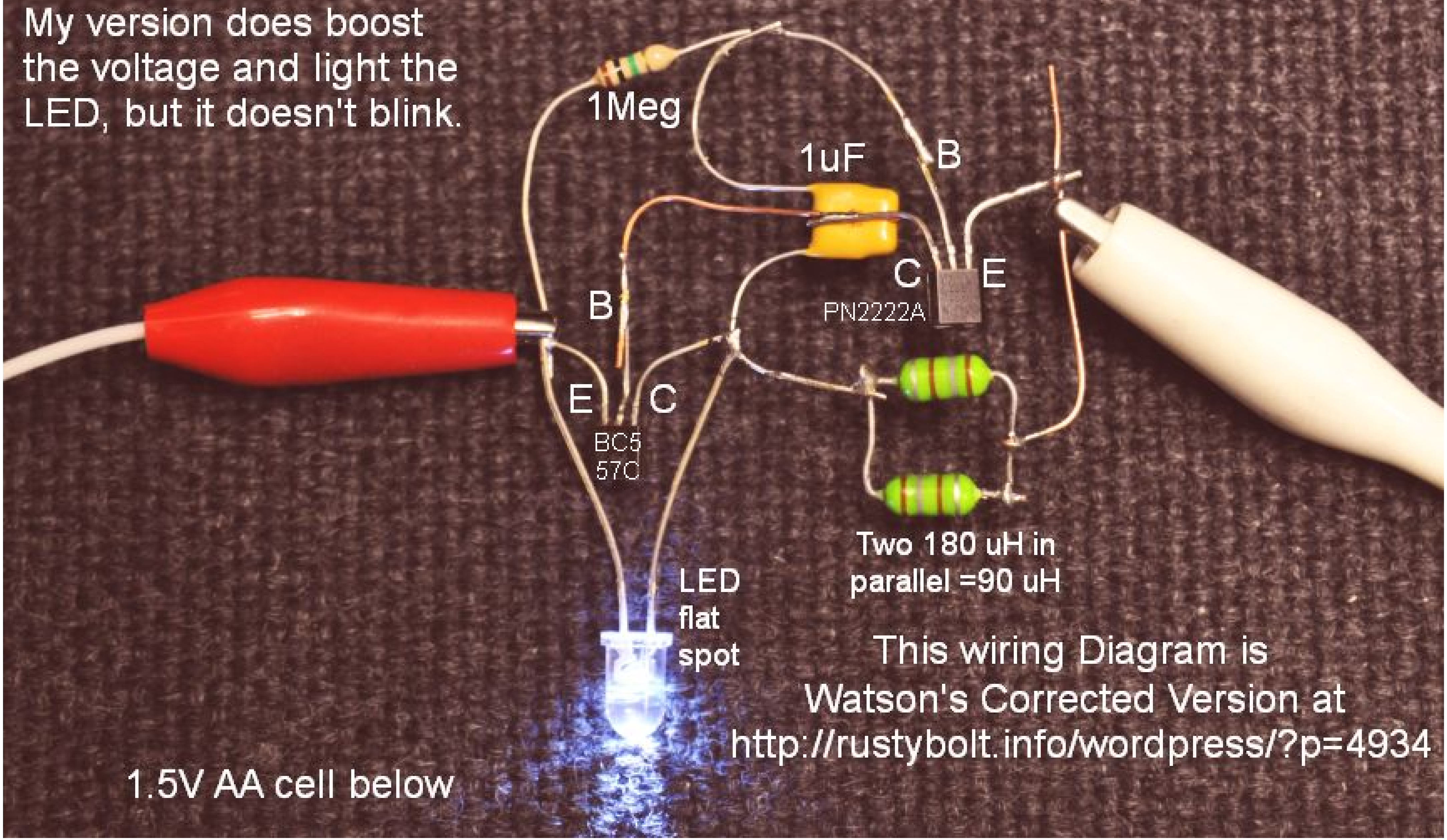
1Meg
1uF
B
C
E
PN2222A
B
E
C
BC5
57C
LED
flat
spot
Two 180 uH in
parallel =90 uH

This wiring Diagram is Watson's Corrected Version at http://rustybolt.info/wordpress/?p=4934

1.5V AA cell below

Watson's Joule Thief Pictorial Diagram

http://rustybolt.info/wordpress/?p=128 2011 Dec 07

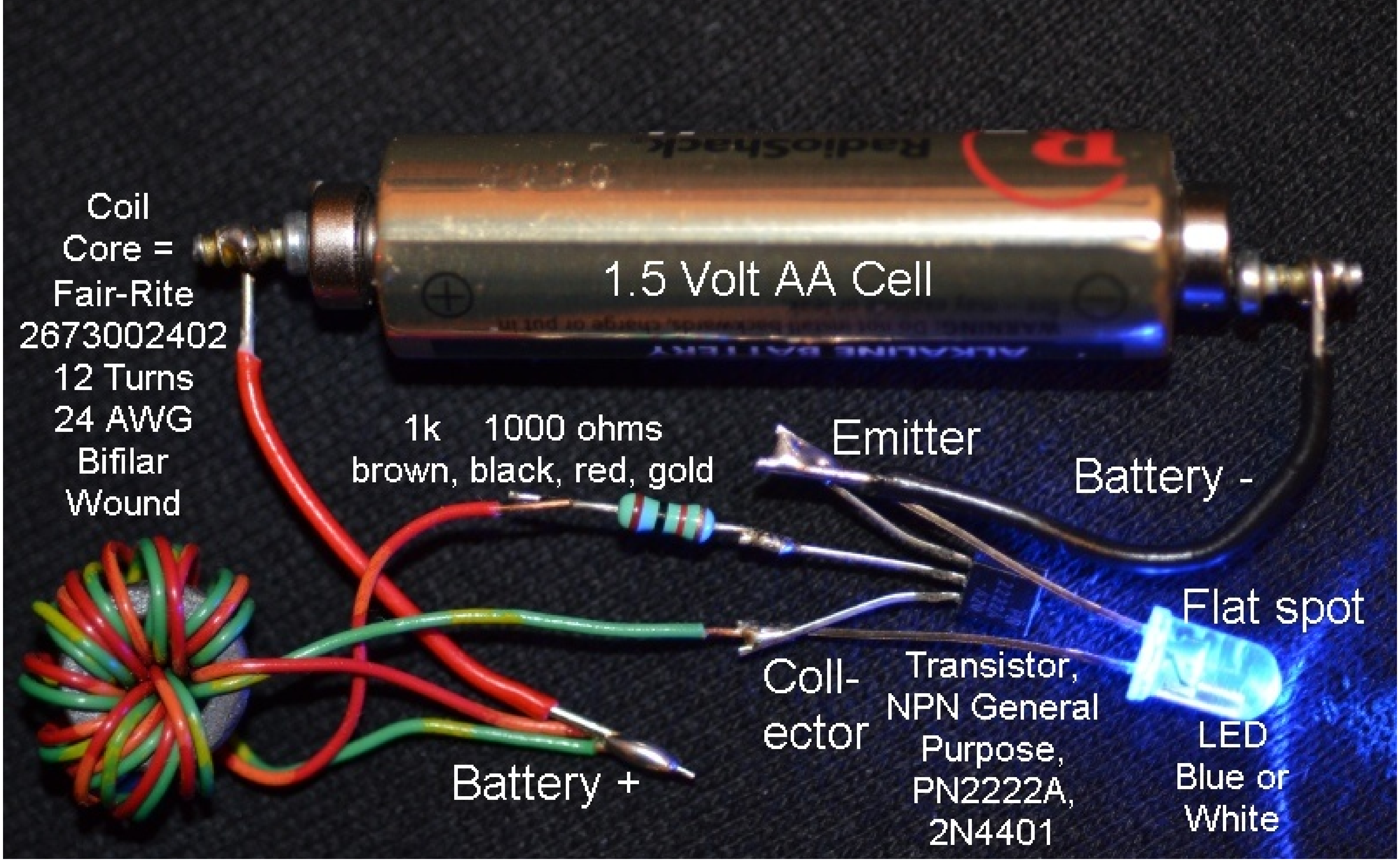
Coil Core = Fair-Rite 2673002402 12 Turns 24 AWG Bifilar Wound
1.5 Volt AA Cell
1k 1000 ohms brown, black, red, gold
Emitter
Battery -
Flat spot
Coll-ector
Transistor, NPN General Purpose, PN2222A, 2N4401
LED Blue or White
Battery +

Joule Thief

Pictorial Wiring Diagram Dec 11, 2009

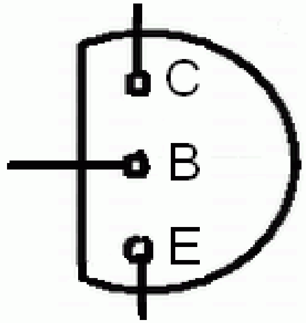

for 2N4401

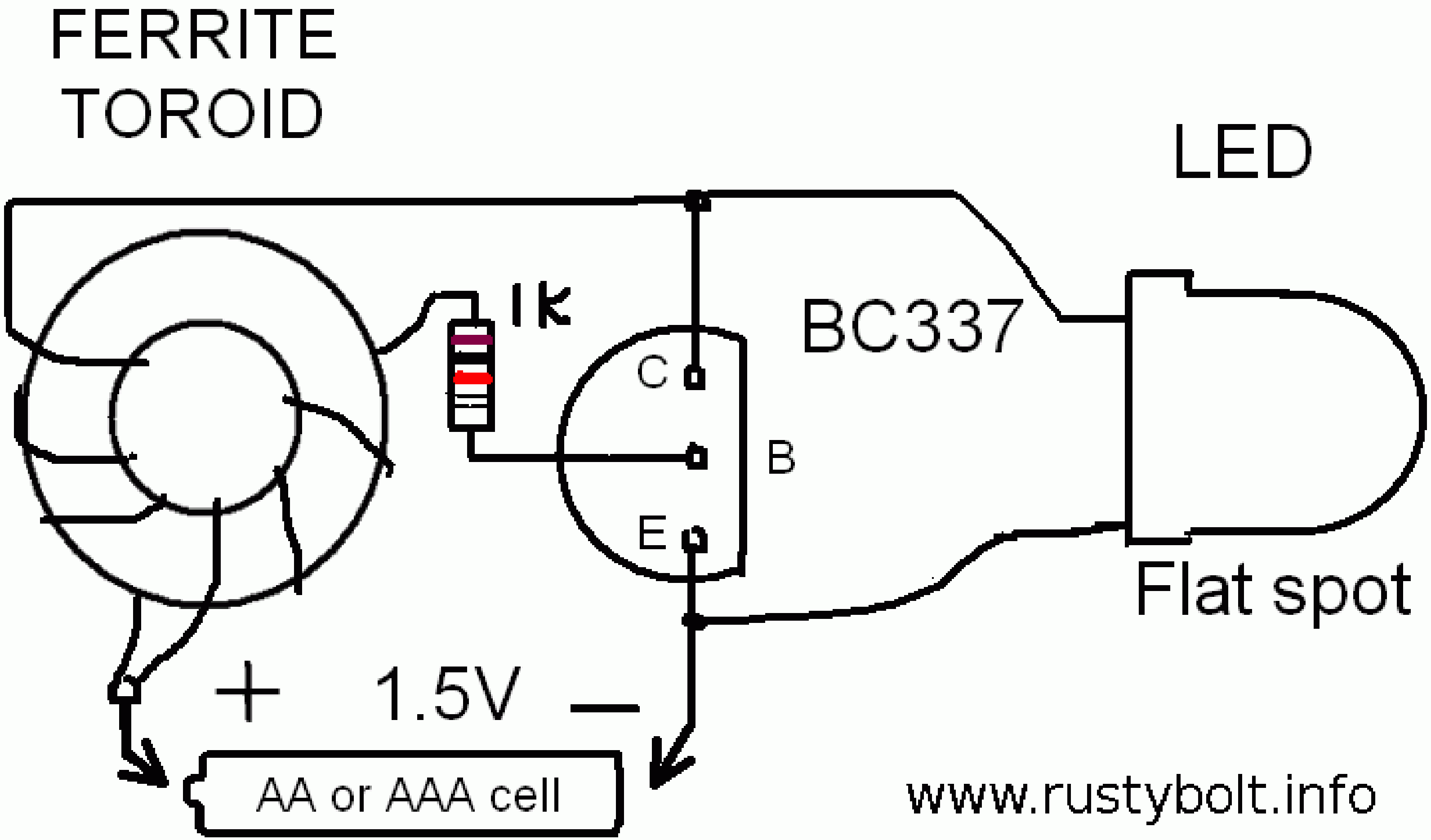
FERRITE
TOROID
1k
BC337
C
B
E
LED
Flat spot
+
1.5V
—
AA or AAA cell

www.rustybolt.info

Watson's True 2N2222 Joule Thief

Oct 31, 2010 Boo!

Not very good performance from an expensive transistor

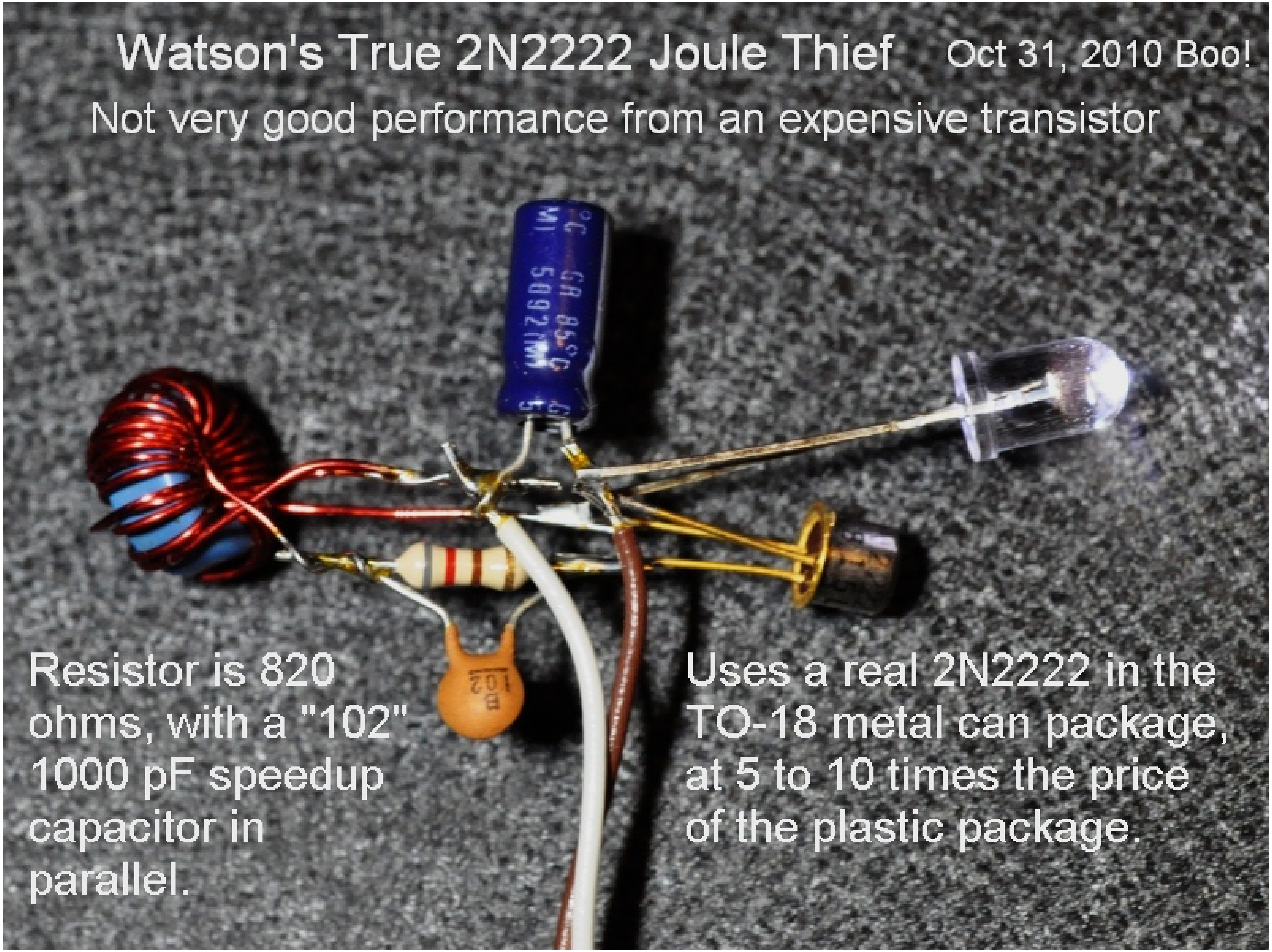

Resistor is 820 ohms, with a "102" 1000 pF speedup capacitor in parallel.

Uses a real 2N2222 in the TO-18 metal can package, at 5 to 10 times the price of the plastic package.

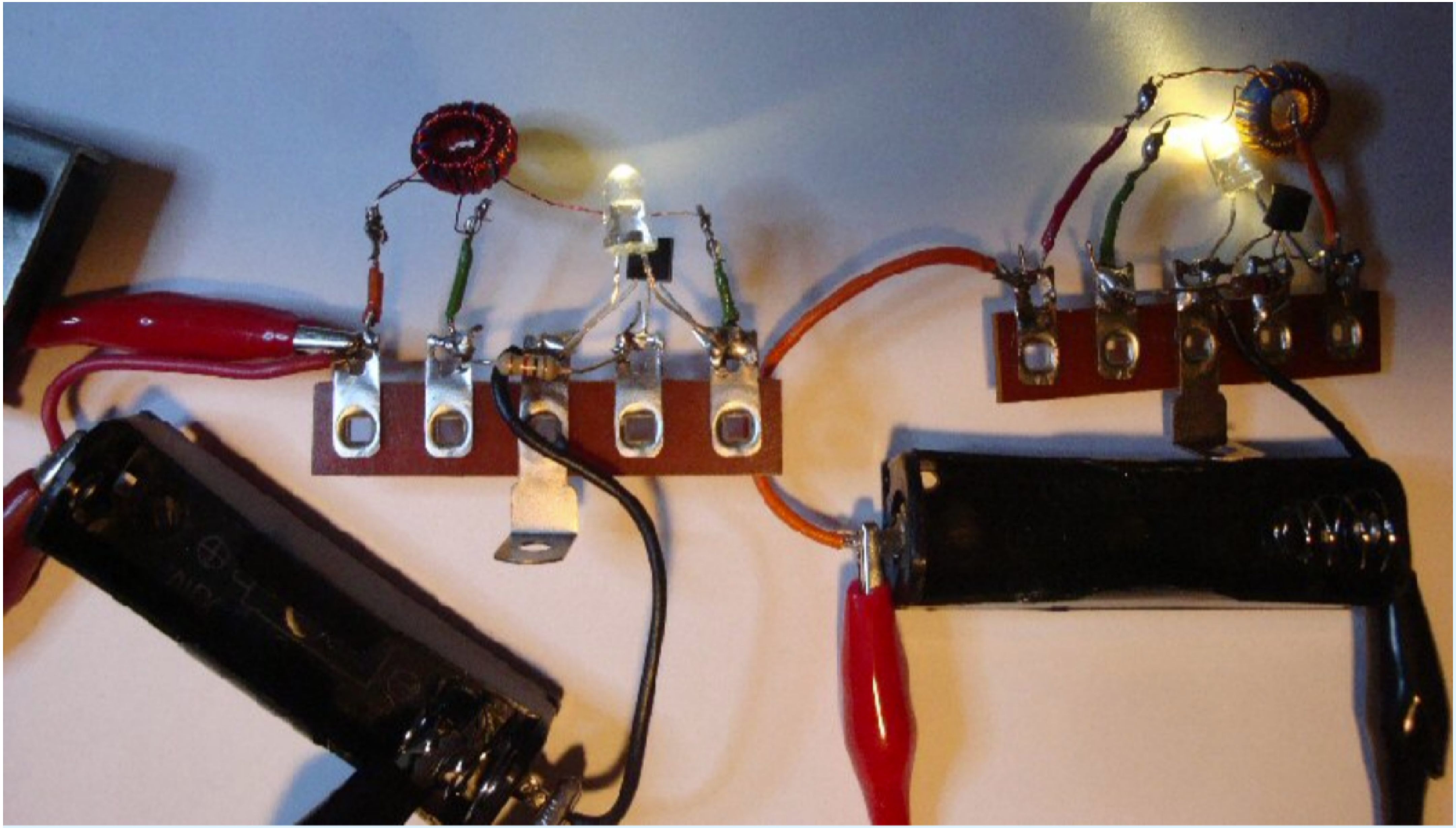

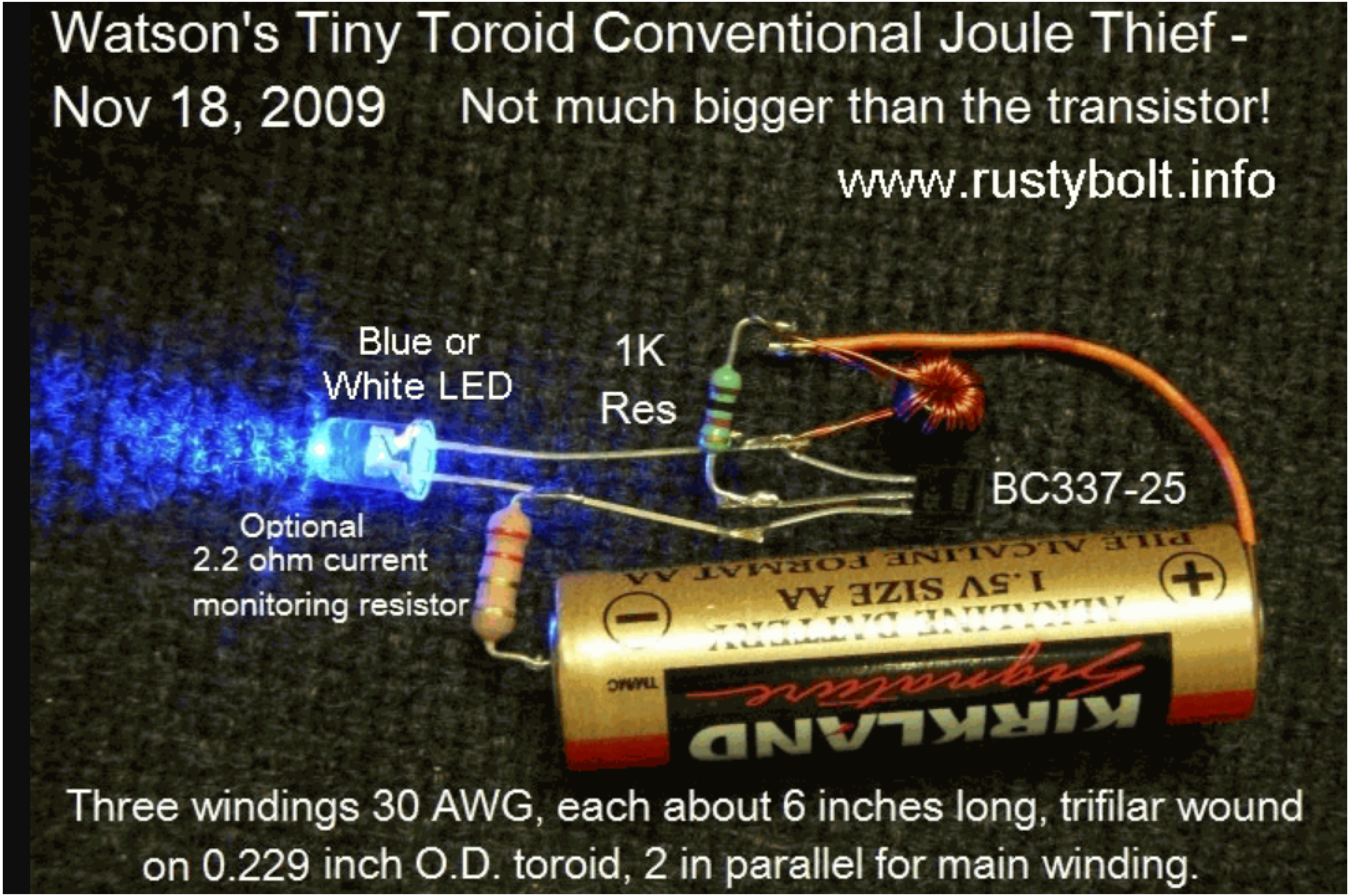

Watson's Tiny Toroid Conventional Joule Thief - Nov 18, 2009 Not much bigger than the transistor!

www.rustybolt.info

Blue or White LED
1K Res
BC337-25
Optional 2.2 ohm current monitoring resistor
PILE ALCALINE FORMAT AA
1.5V SIZE AA
ALKALINE BATTERY
KIRKLAND Signature

Three windings 30 AWG, each about 6 inches long, trifilar wound on 0.229 inch O.D. toroid, 2 in parallel for main winding.

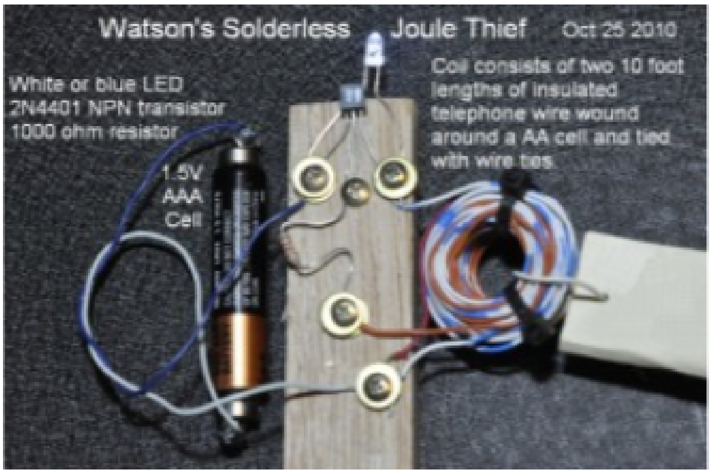
Watson's Solderless Joule Thief Oct 25 2010
White or blue LED
2N4401 NPN transistor
1000 ohm resistor
1.5V
AAA
Cell
Coil consists of two 10 foot lengths of insulated telephone wire wound around a AA cell and tied with wire ties

I took this photo just before Halloween back in 2010, after I made this Joule Thief without solder and without a toroid. It shows that you don't have to know how to solder and you don't have to have a toroid to make a working Joule Thief – all you need is a screwdriver. So do your thing and experiment a bit, and see what you can come up with.

Each screw (except for one) has two washers to hold the wires. Putting two or more wires under the screw head without the washers will be difficult and the wires will try to come loose. Two or three washers makes it easier and the wires are held much firmer. The wood block is a piece of solid oak that is about a half inch (12mm) thick. The oak is very hard and the screw holes have to be drilled out before the screws are put in. This also makes it possible to use small screws with a flat end which were used for holding plastic parts together. The screw size is about #3 by 3/8 inch long, with coarse threads like a sheet metal screw.

The LED is not very bright because it's a cheap 3 mm white LED I got from an eBay seller, and it has been used a bit so it has dimmed. Use a decent LED and this circuit will be as bright as a toroid JT.

Watson's Solderless Joule Thief

Another Joule Thief I put together is shown in the second photo. This uses the bare wires to wrap around the joints in place of solder. The joints are not as mechanically strong as solder but for a quick experiment they should work okay. If the JT is going to be used, the wire joints may become intermittent or loose. The solder joints are a much better connection.

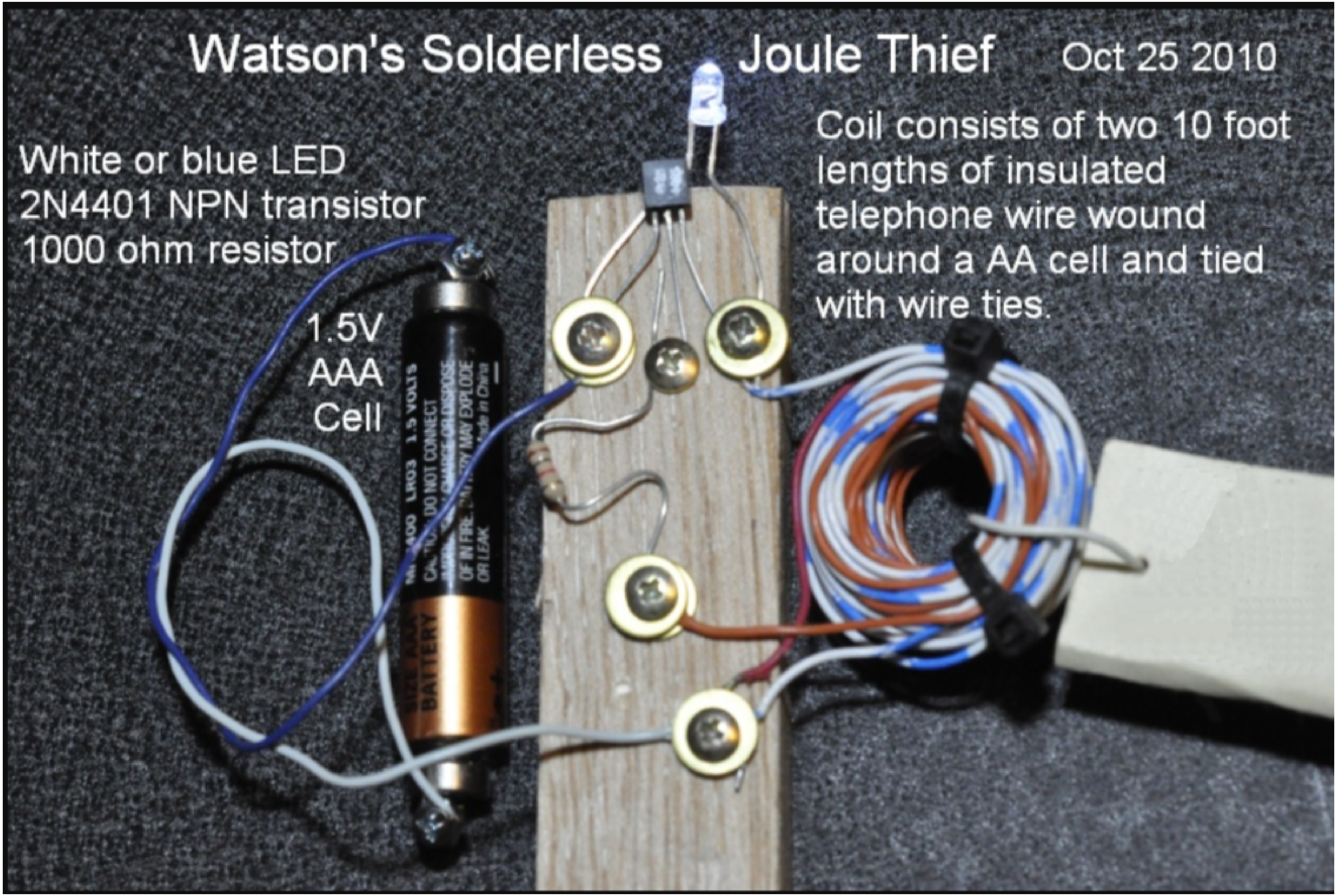

Watson's Solderless Joule Thief Oct 25 2010

White or blue LED
2N4401 NPN transistor
1000 ohm resistor

Coil consists of two 10 foot lengths of insulated telephone wire wound around a AA cell and tied with wire ties.

Watson's 'Toroid Free' Air Core Joule Thief

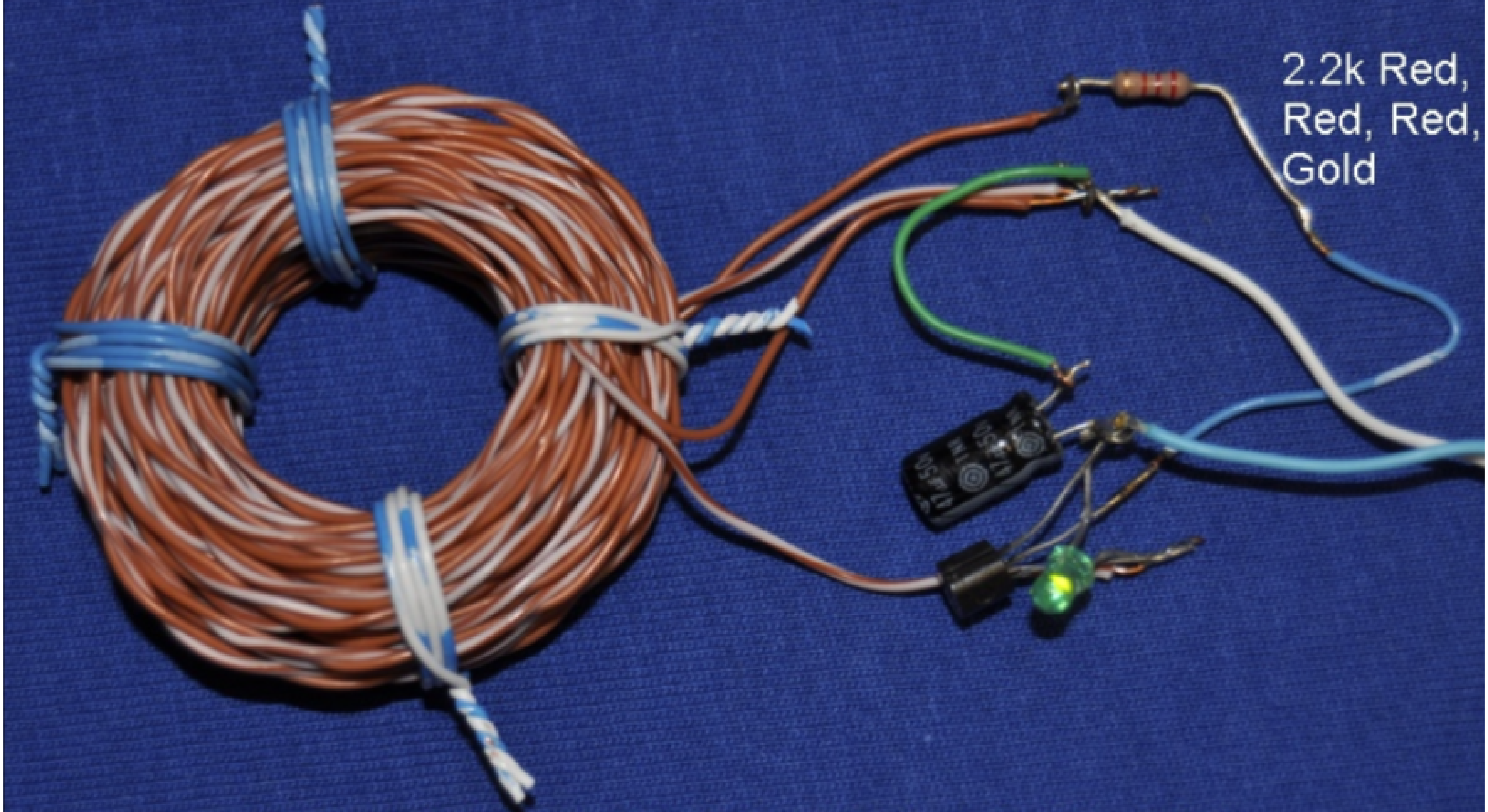
2.2k Red,
Red, Red,
Gold

rustybolt.info/wordpress/?p=2872   May 21, 2012

Watson's Nearly Disposable Joule Thief Light

rustybolt.info/wordpress/?p=2133   Fri. Apr 13 2012

Completed circuit ready for coating of silicone sealant. Transistor is a BC337. If it was a PN2222A or 2N4401, the flat side would be facing up.

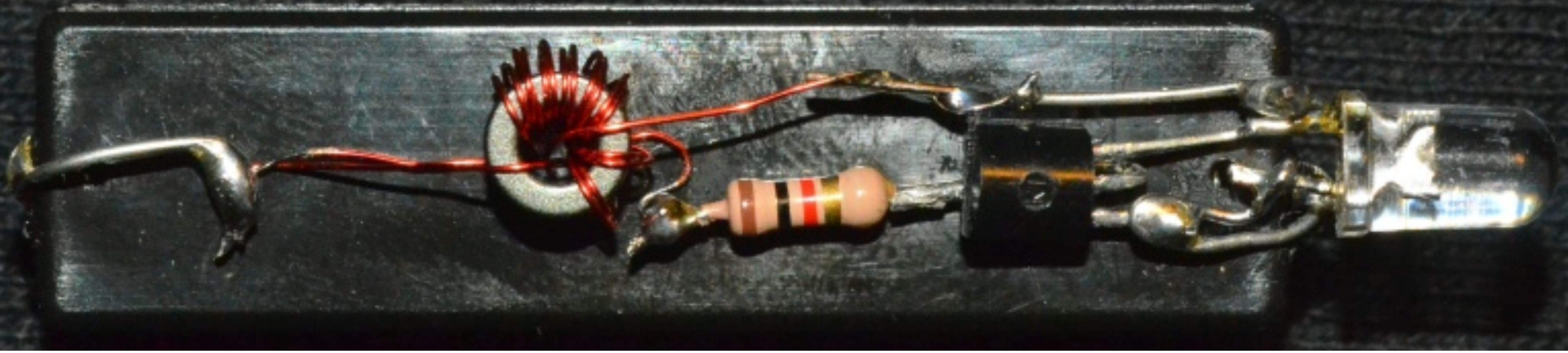

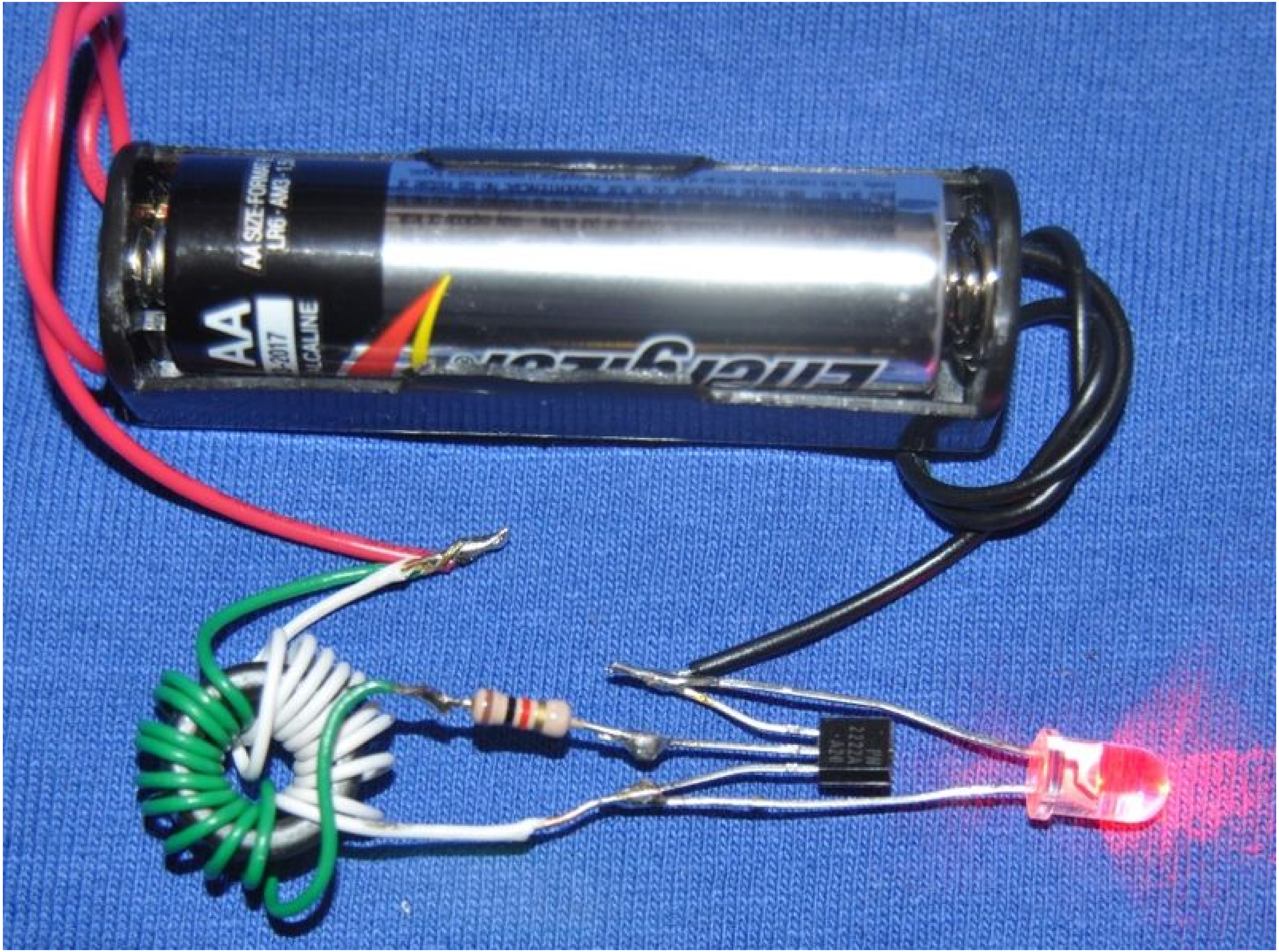
AA
ALKALINE
Energizer
2222A

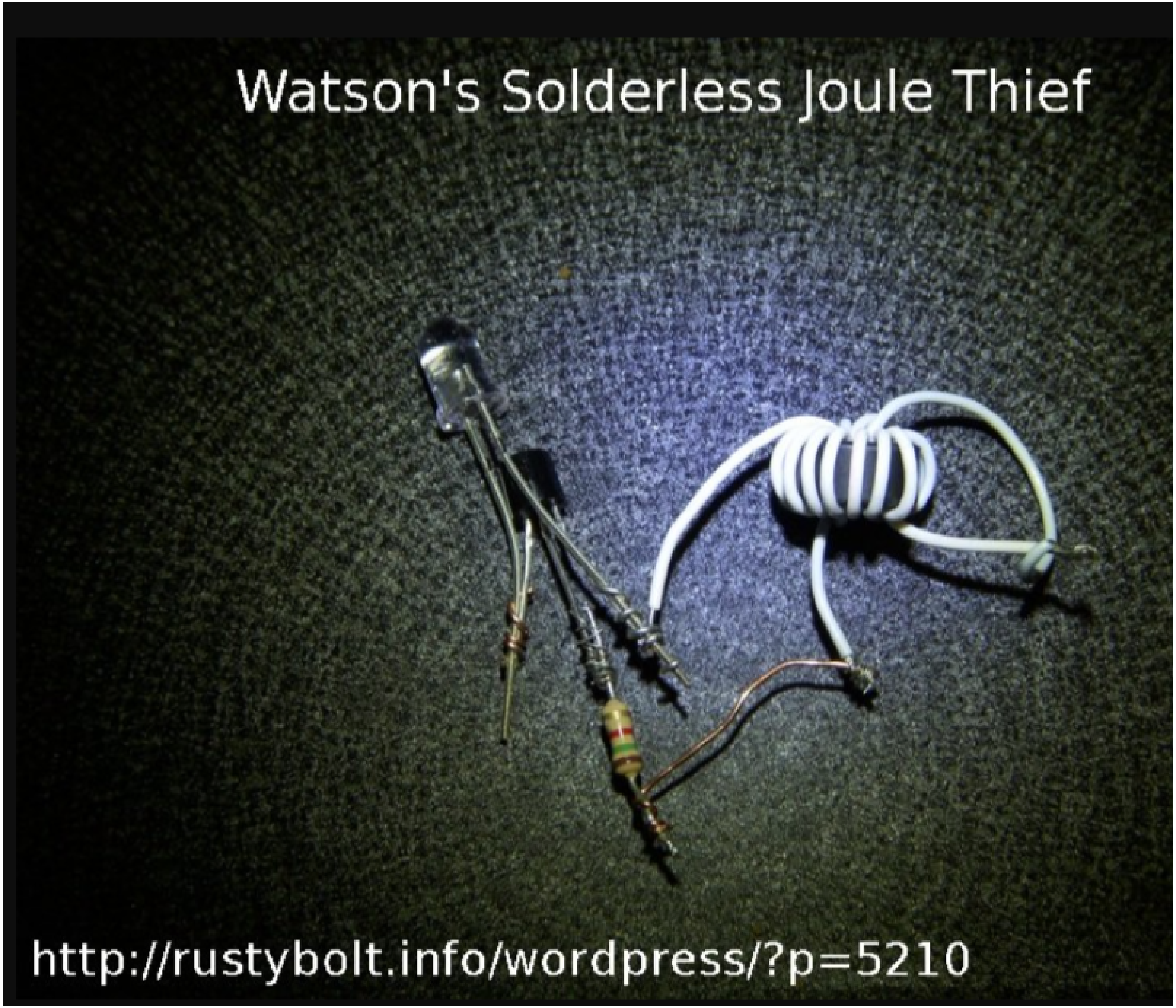

Watson's Solderless Joule Thief

http://rustybolt.info/wordpress/?p=5210

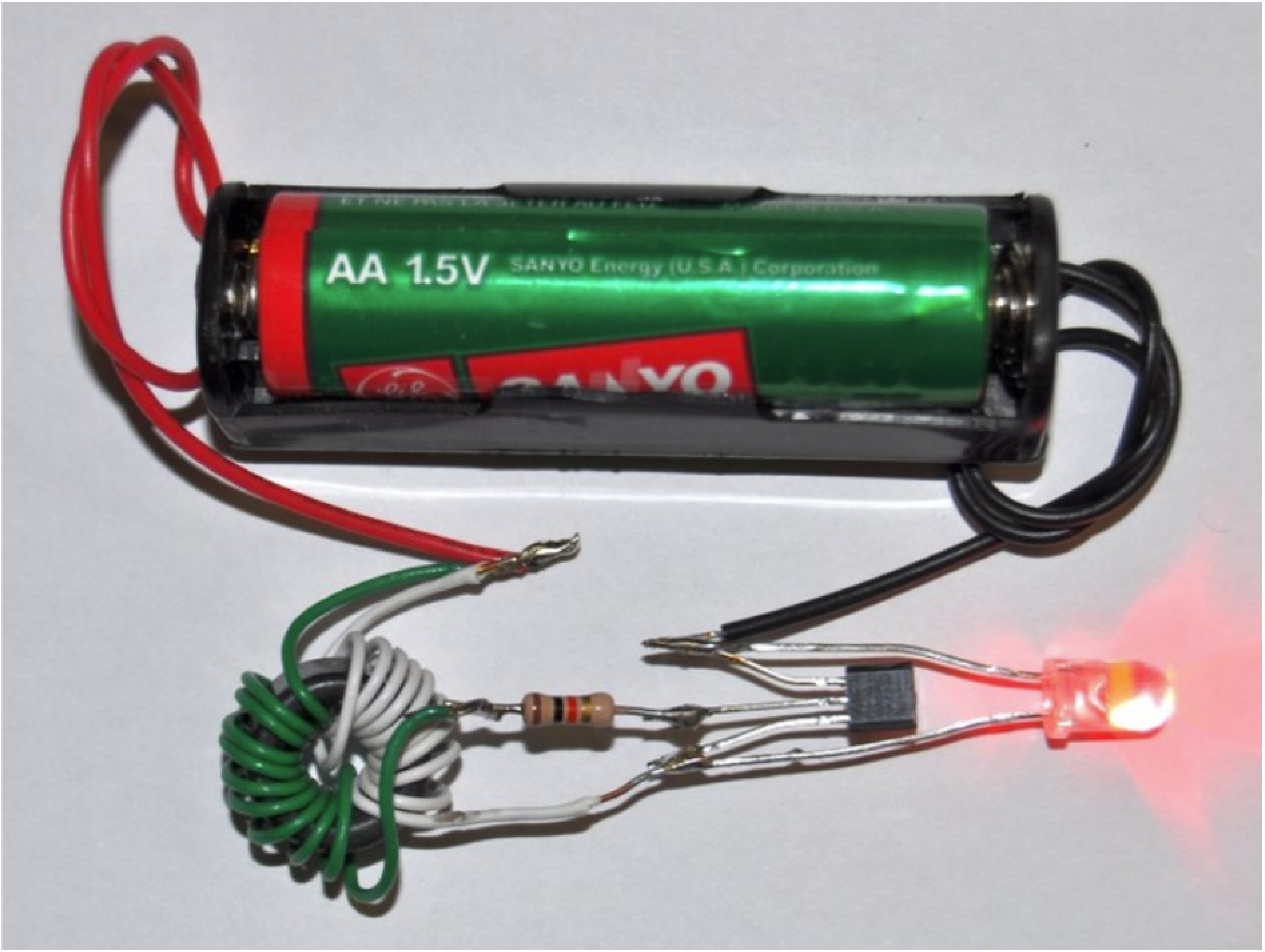
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

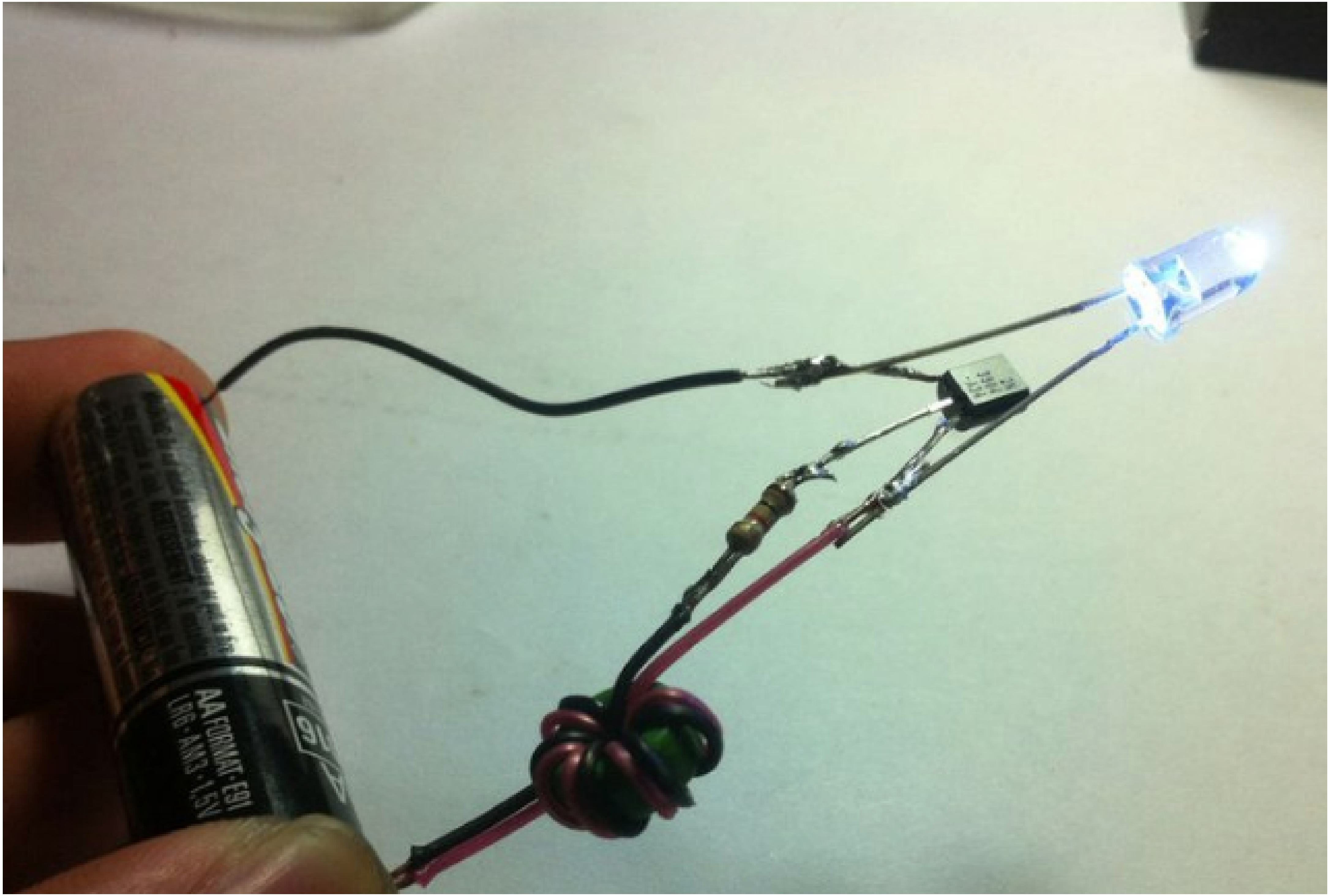
AA FORMAT · E91
LR6 · AM3 · 1.5V

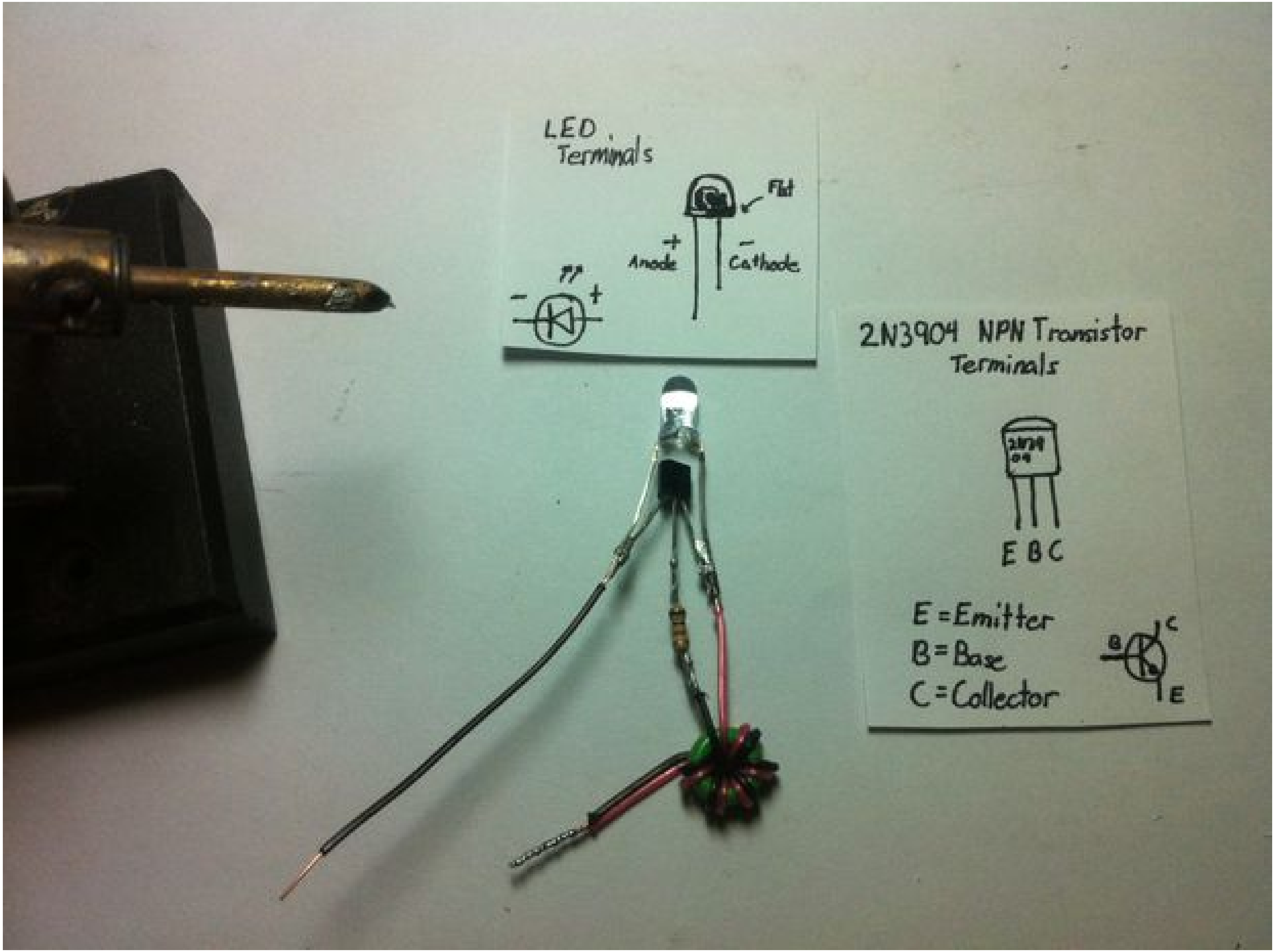
LED Terminals
Flat
+
Anode
−
Cathode
−
+
2N3904 NPN Transistor Terminals
2N39
04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

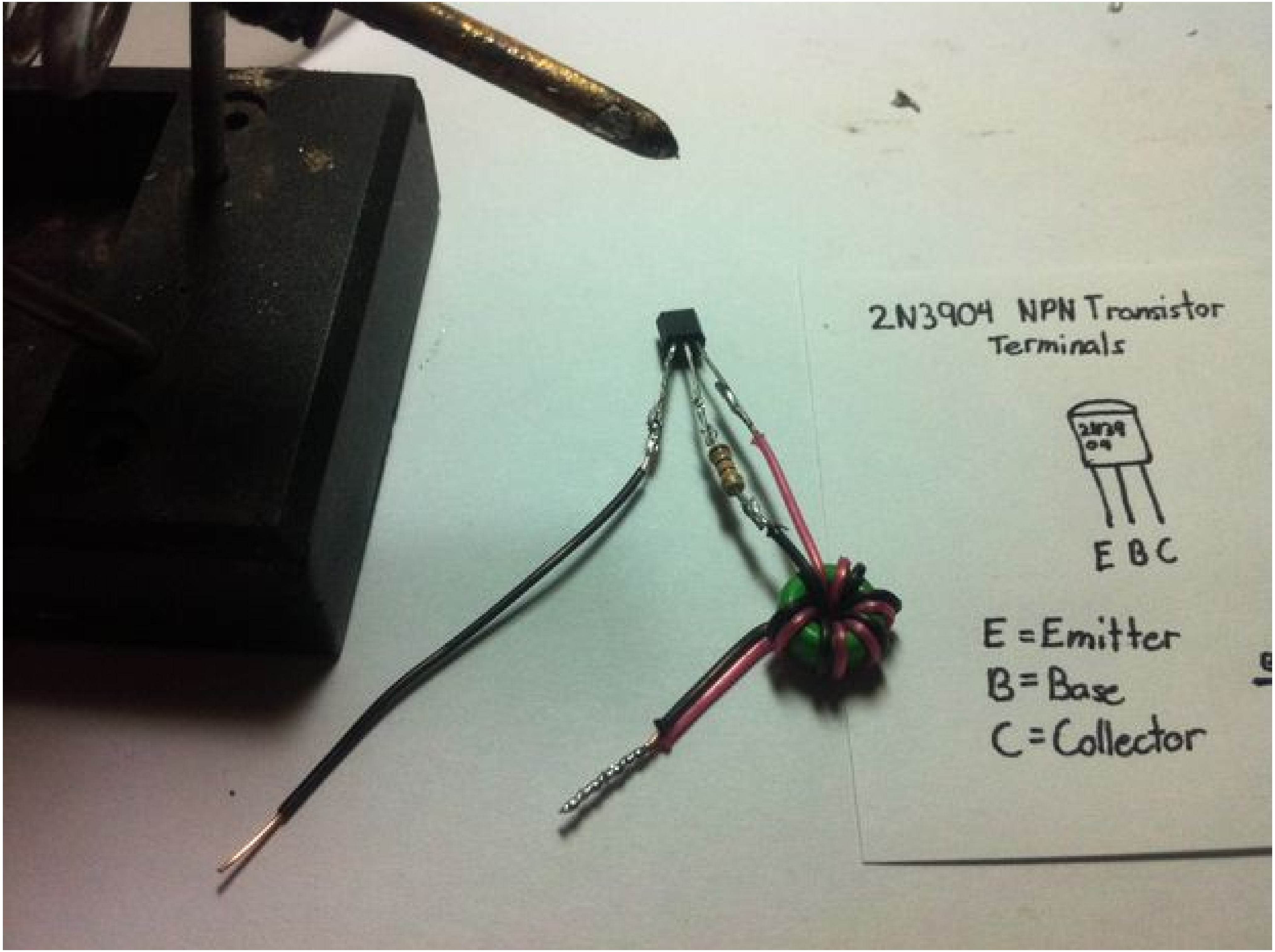
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector

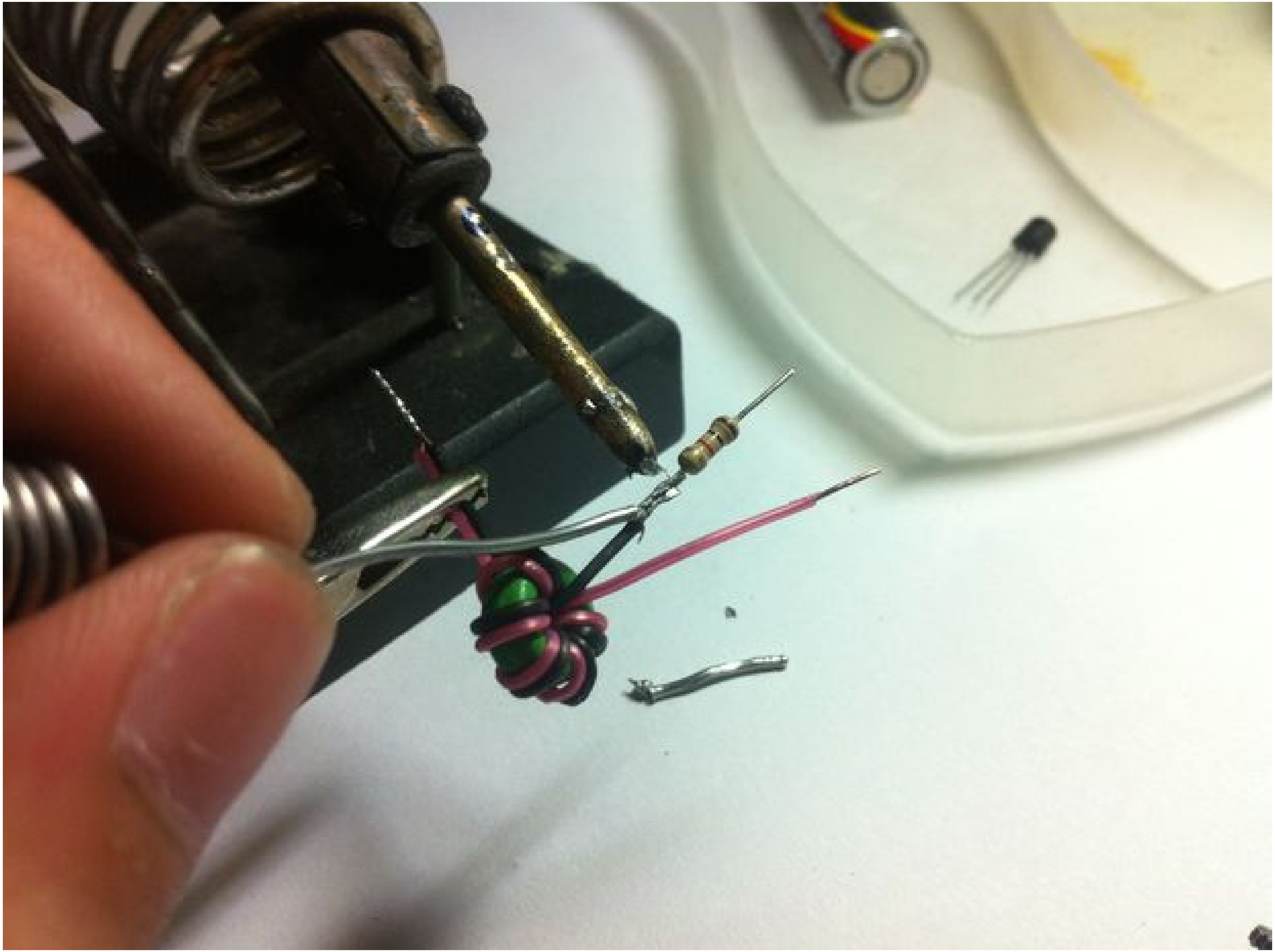

Transistor Terminals

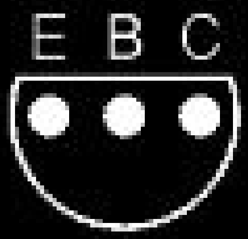

2N3904
(NPN Transistor)

LED Terminals

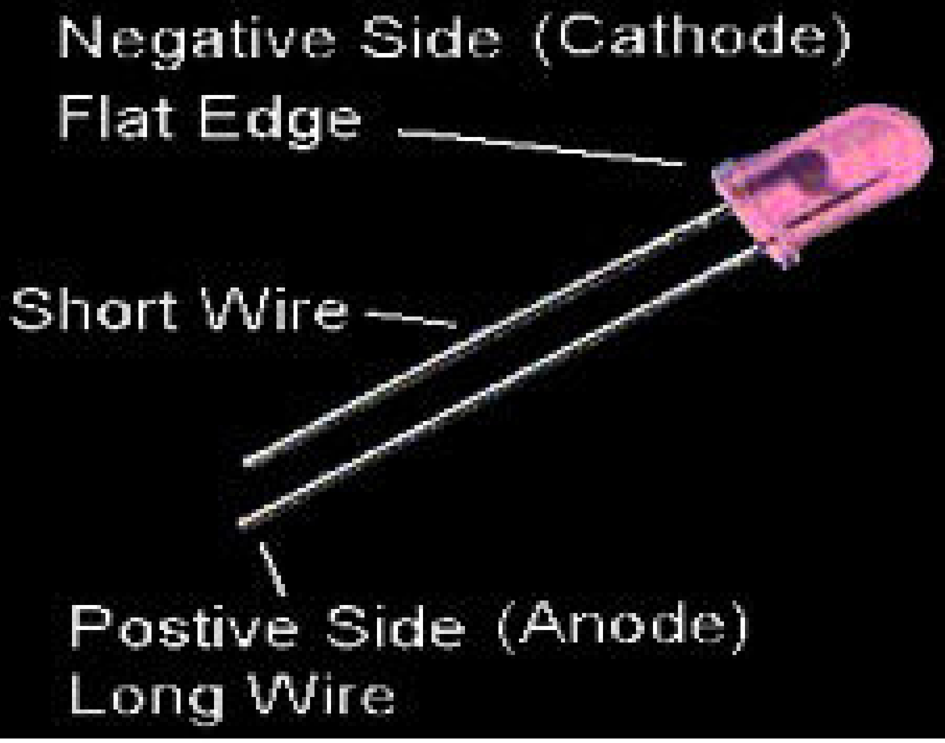
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

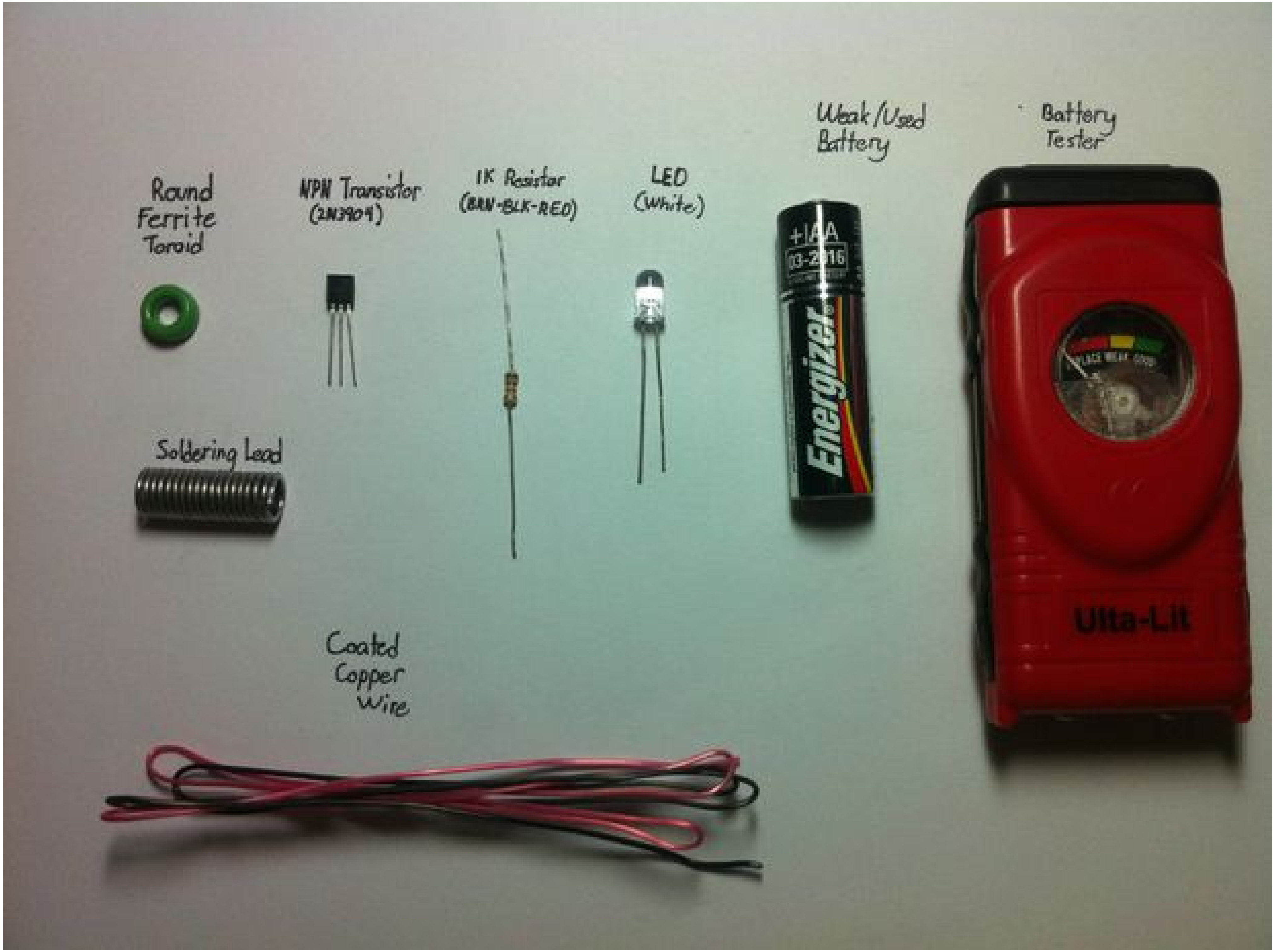
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
+AA
03-2016
Energizer
Ulta-Lit
Soldering Lead
Coated Copper Wire

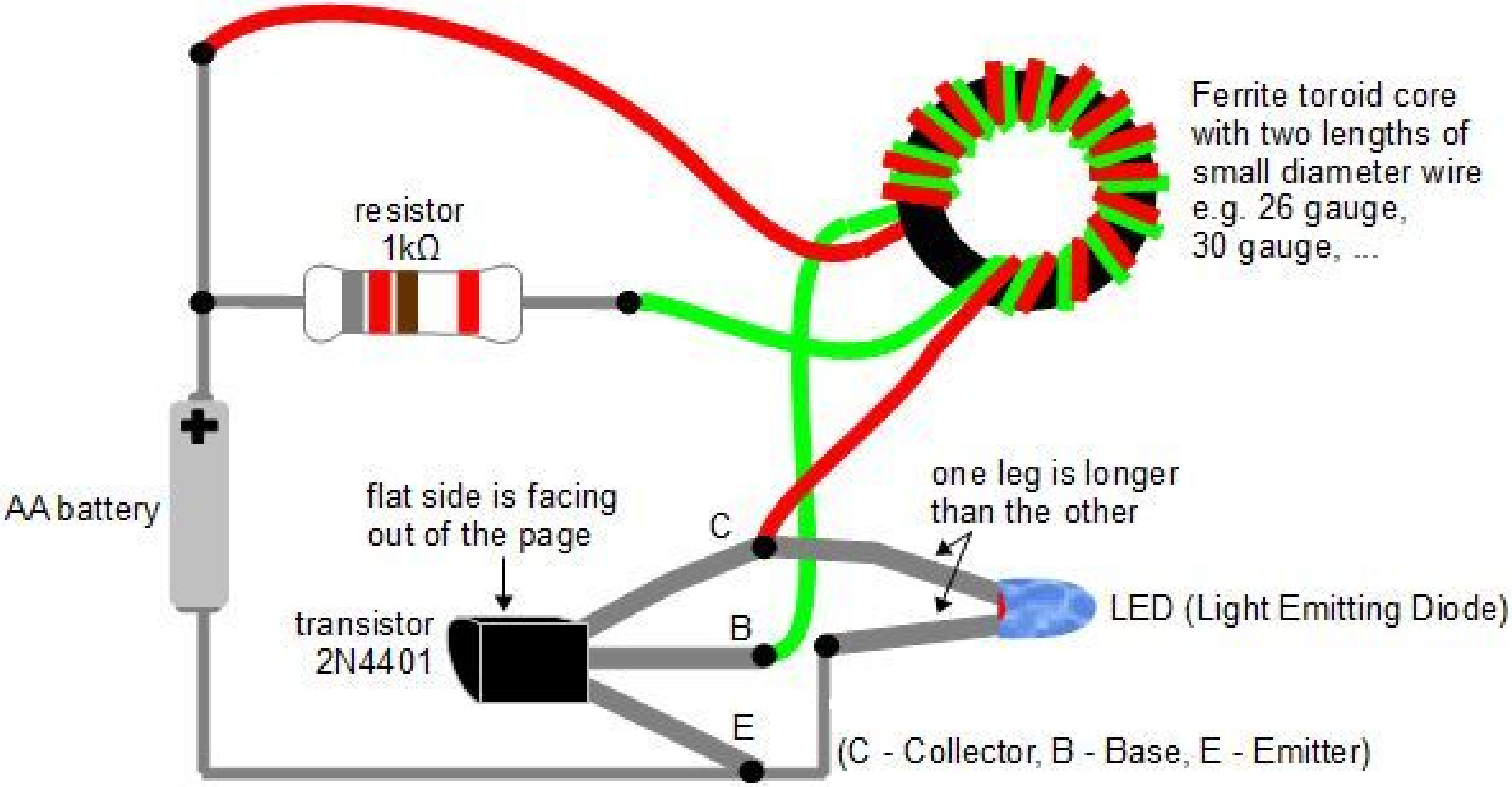
resistor
1kΩ
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
AA battery
+
flat side is facing
out of the page
C
one leg is longer
than the other
transistor
2N4401
B
LED (Light Emitting Diode)
E
(C - Collector, B - Base, E - Emitter)

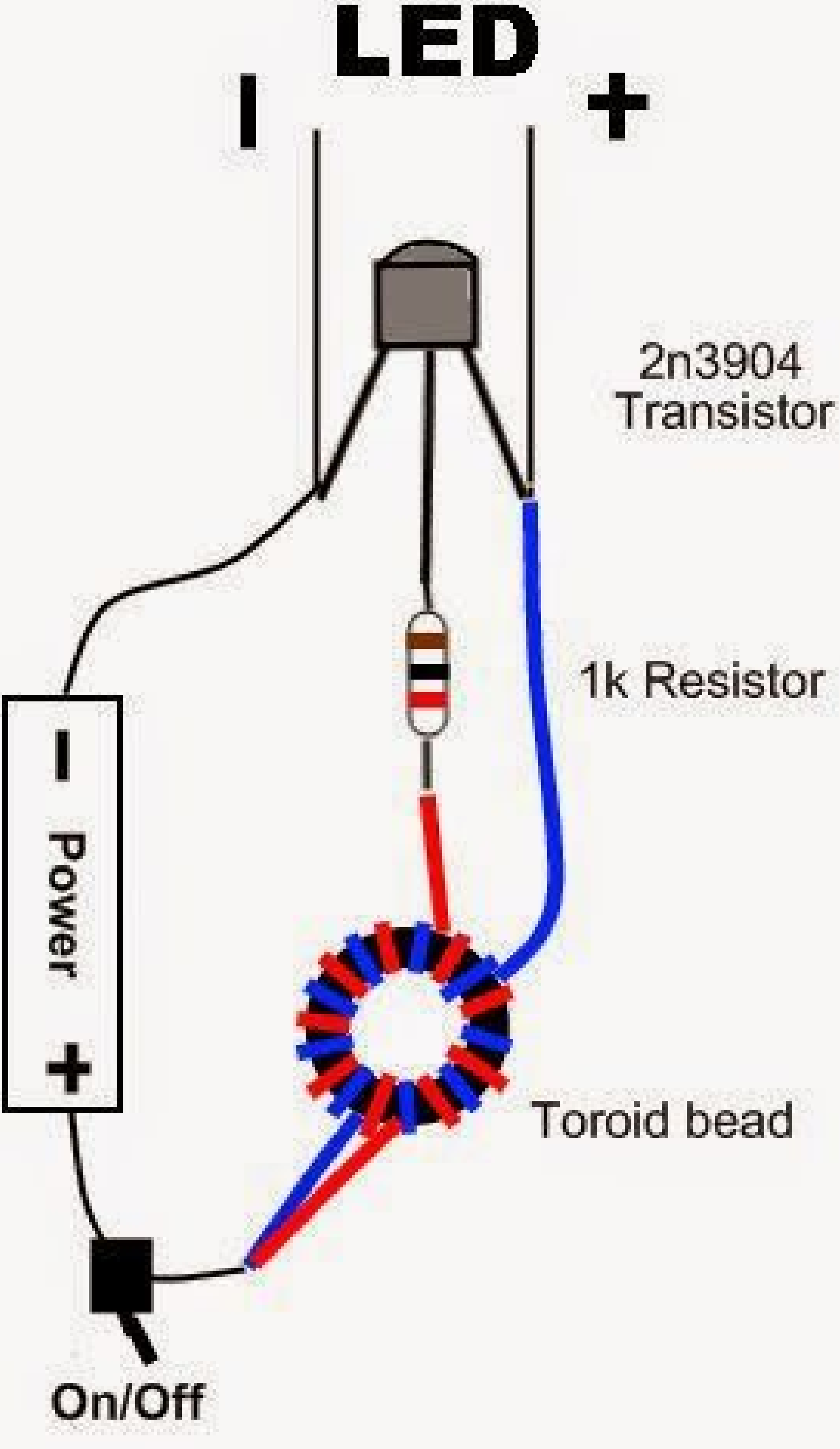
LED
-
+
2n3904
Transistor
1k Resistor
-
Power
+
Toroid bead
On/Off

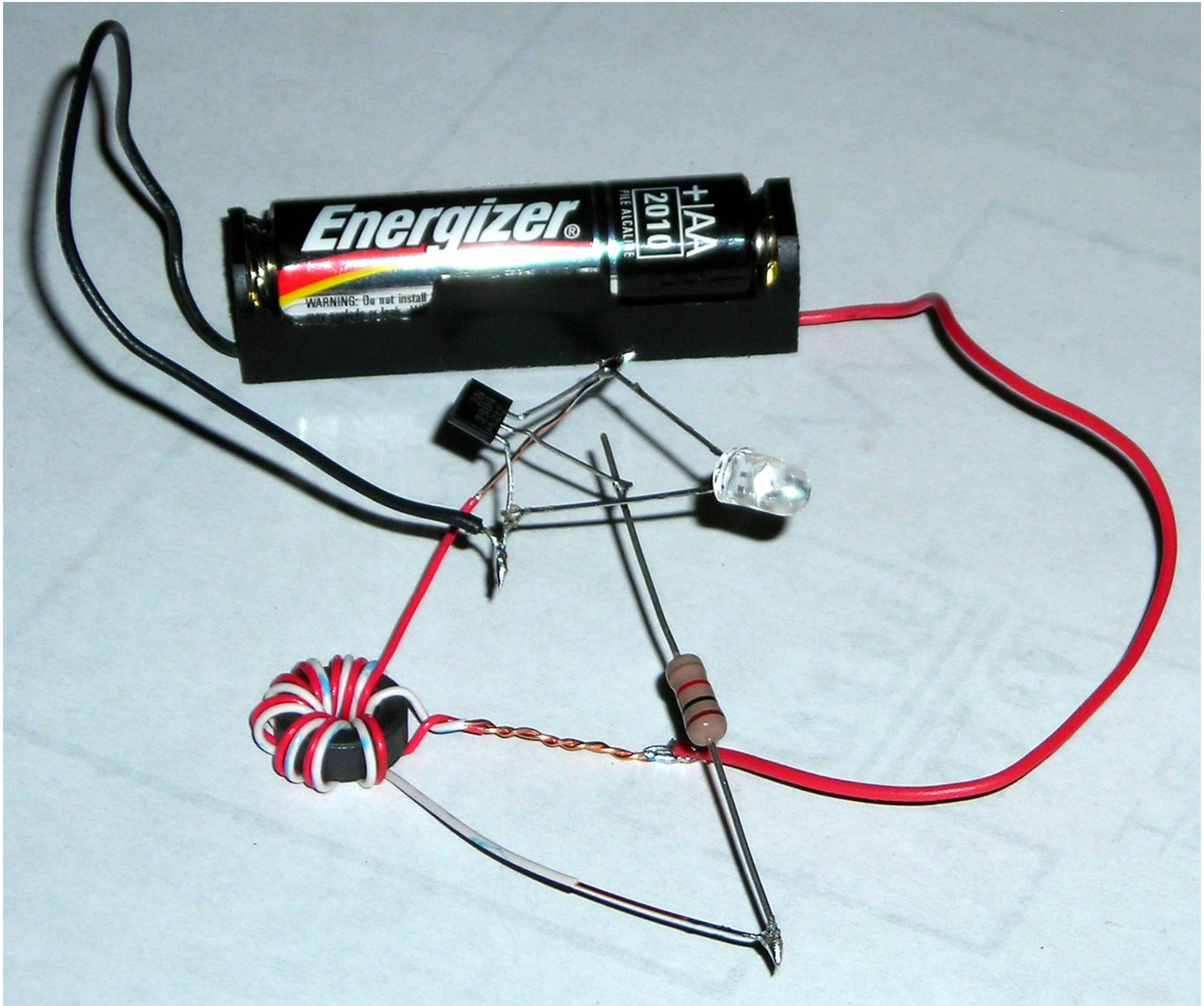
Energizer
+|AA
2010
PILE ALCALINE
WARNING: Do not install

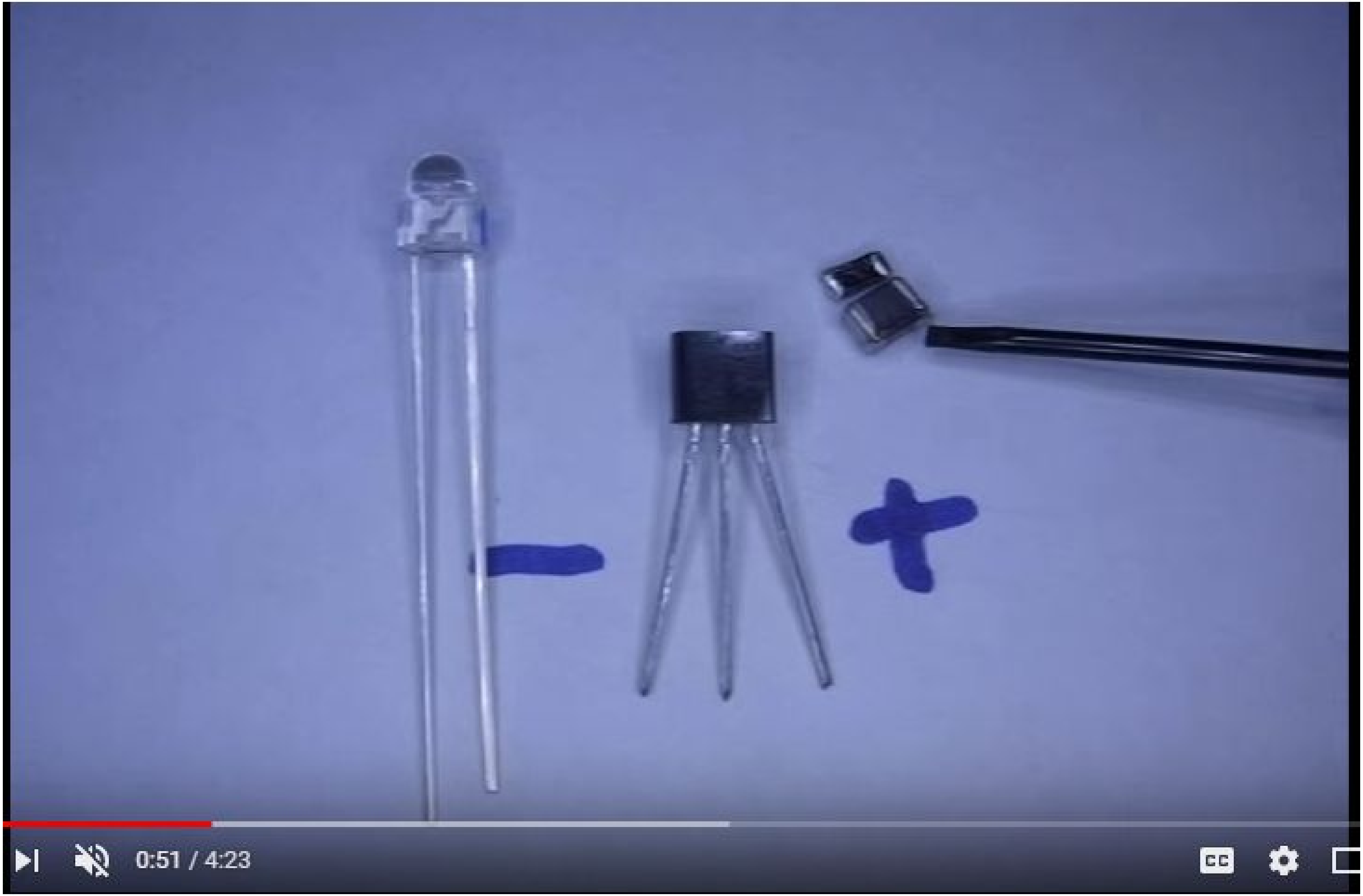
−
+
0:51 / 4:23
CC

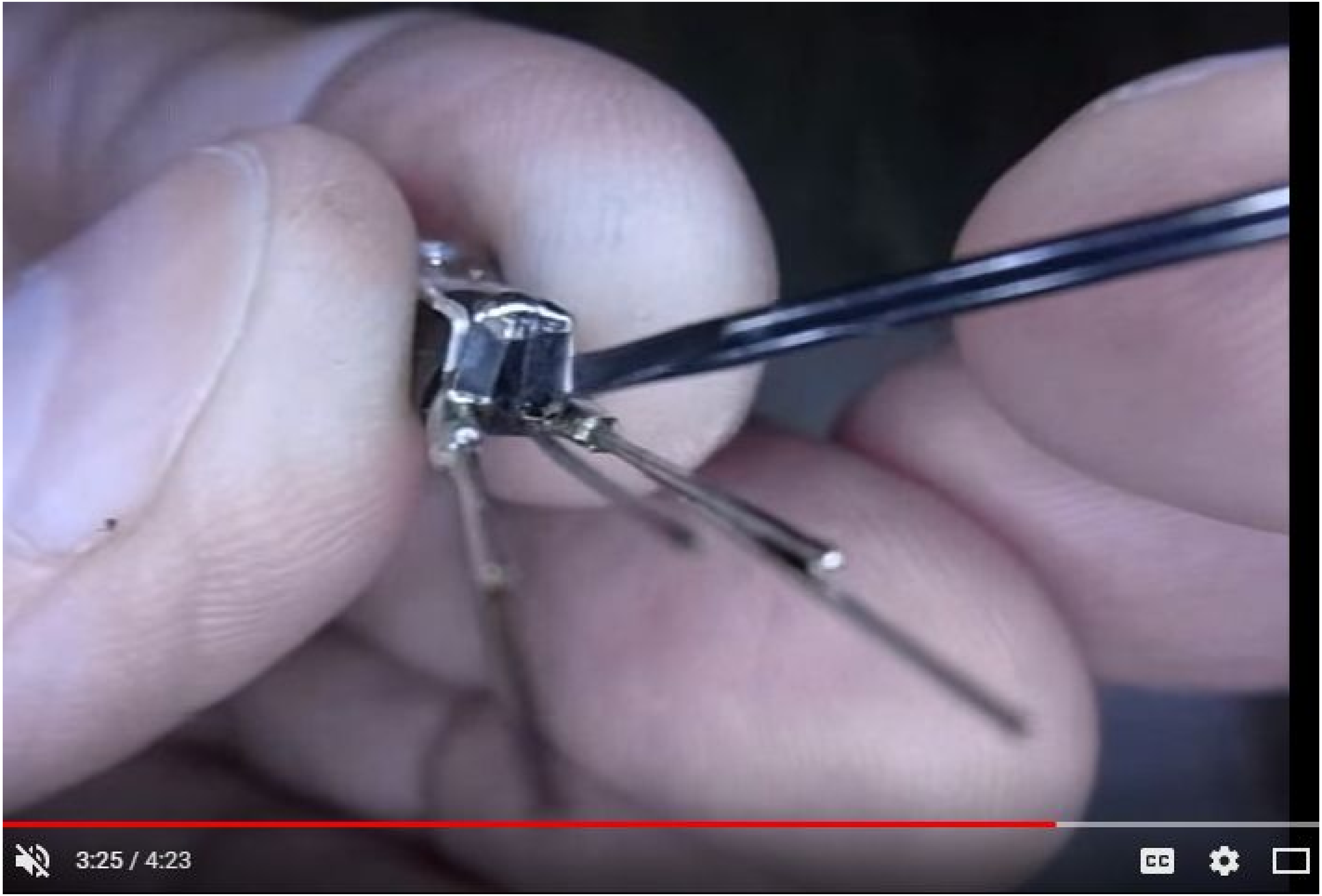
3:25 / 4:23
CC

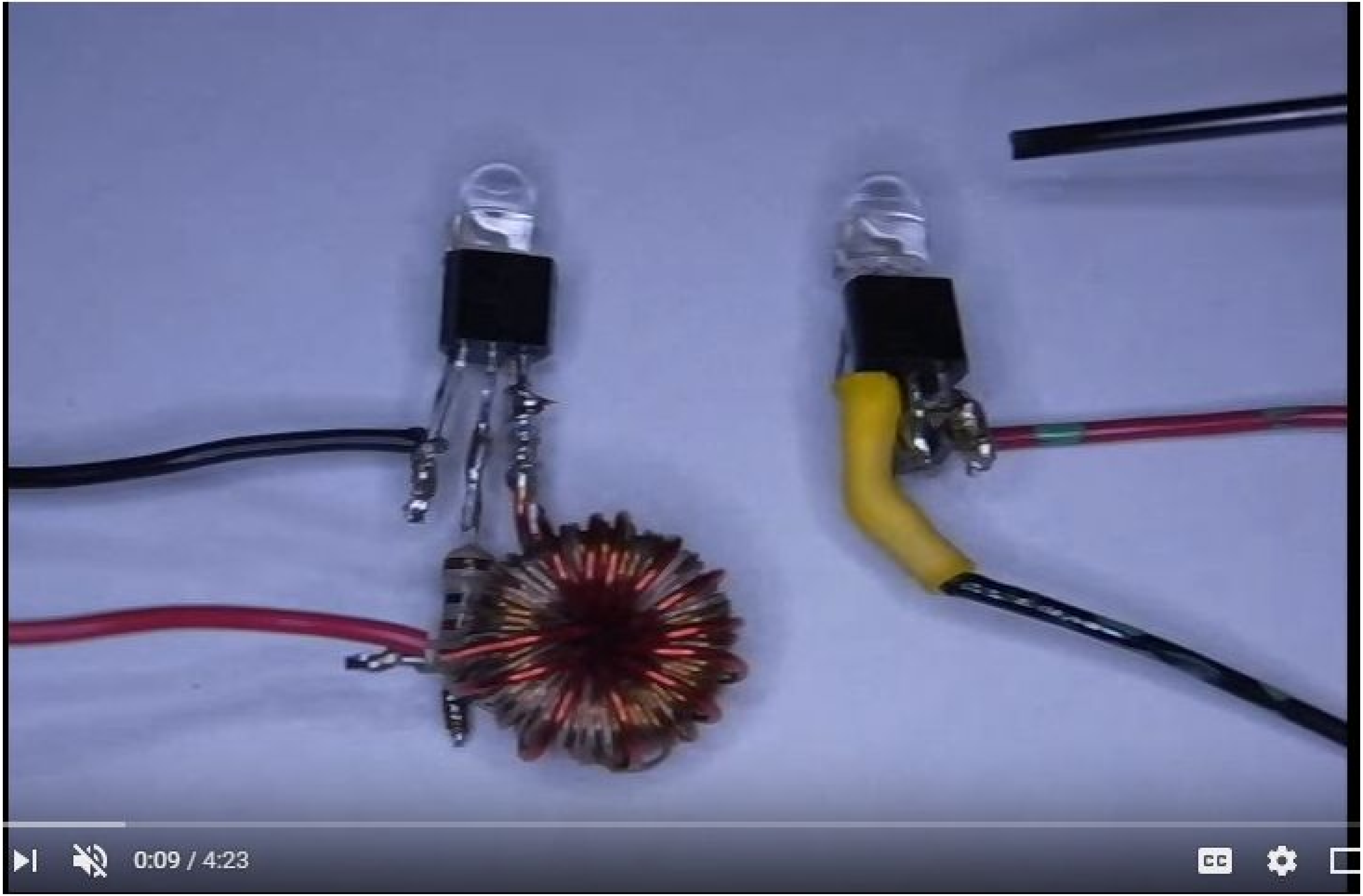
0:09 / 4:23

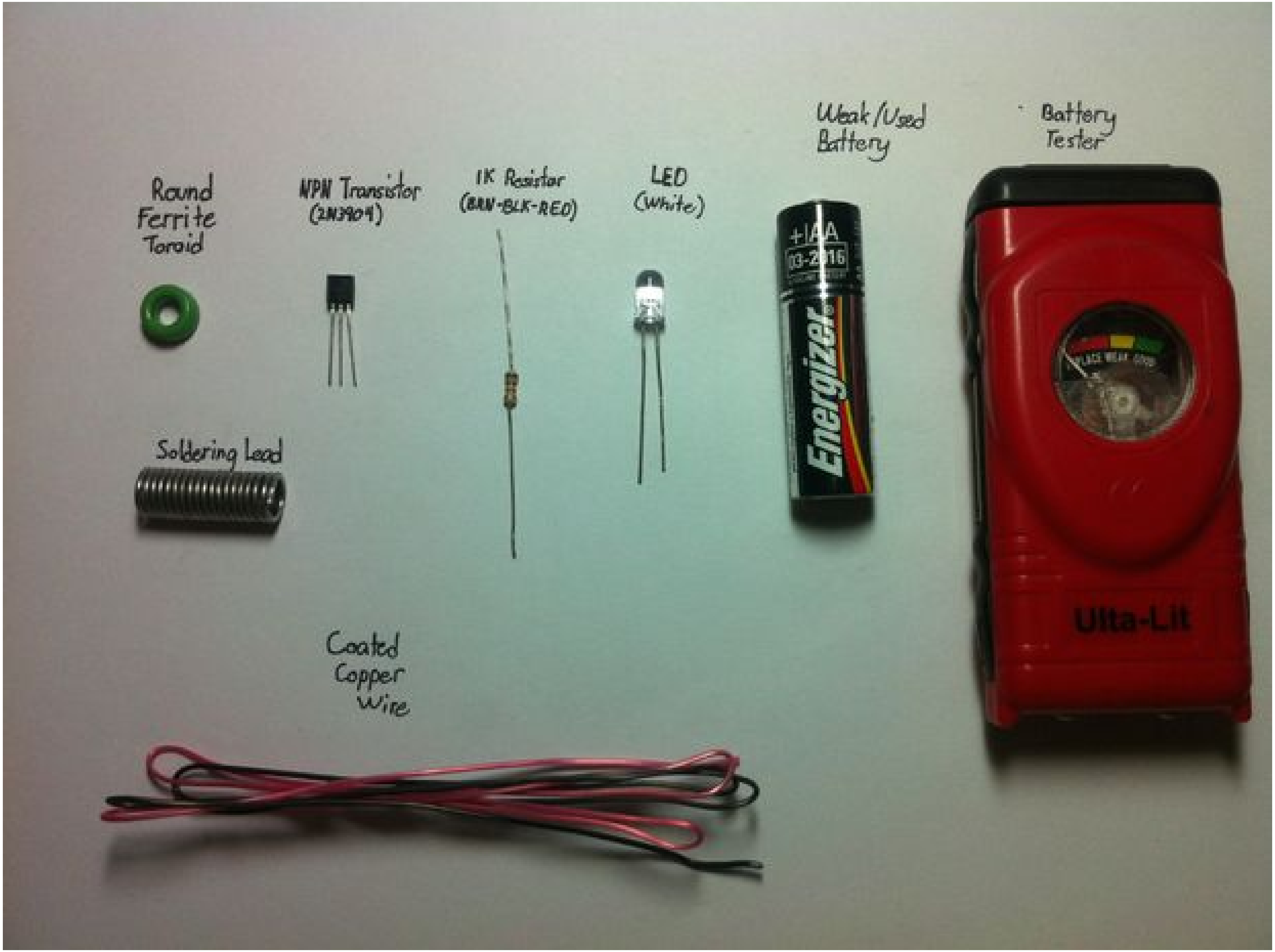
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
+AA
03-2016
Energizer
Soldering Lead
Ulta-Lit
Coated Copper Wire

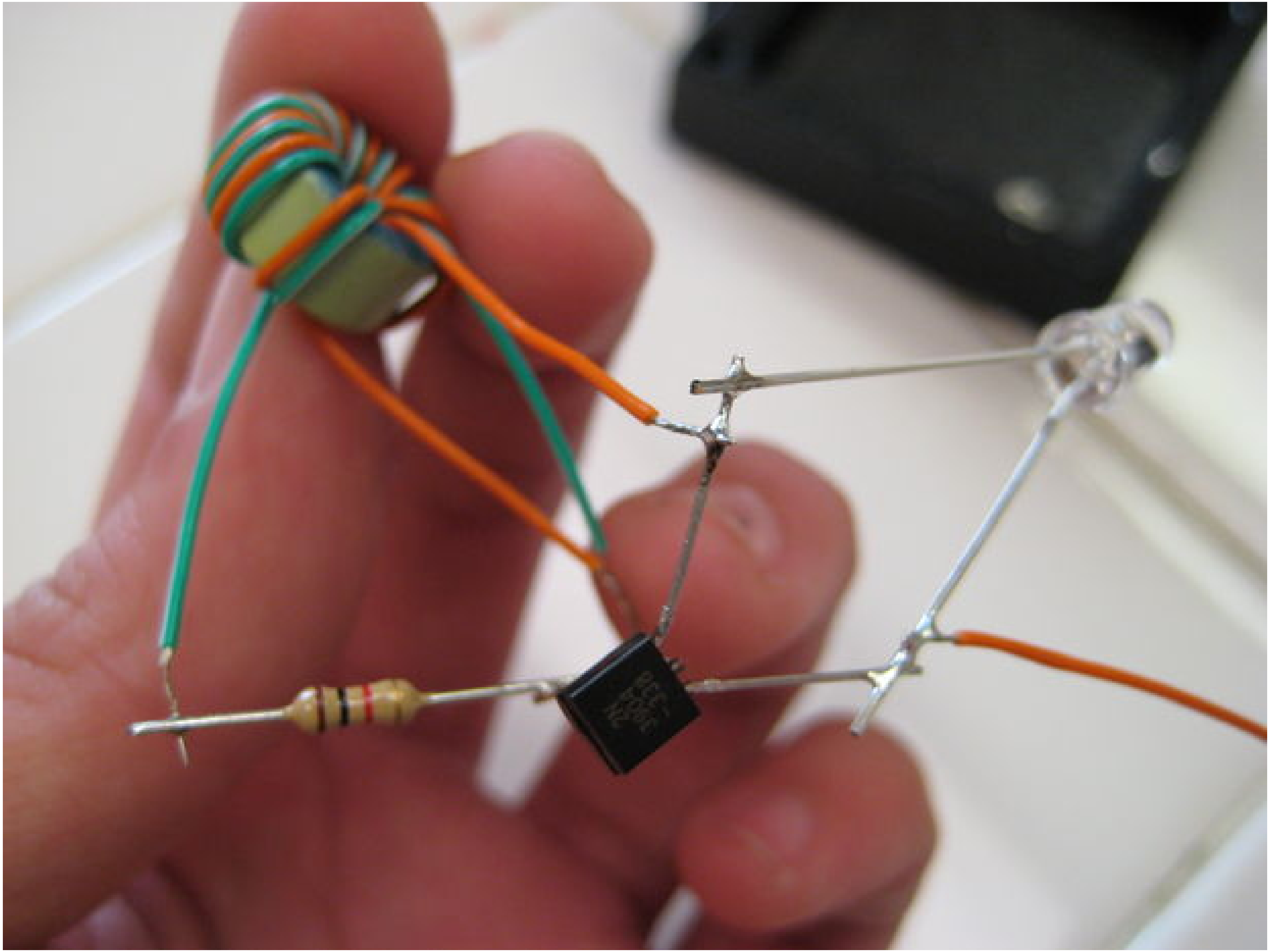

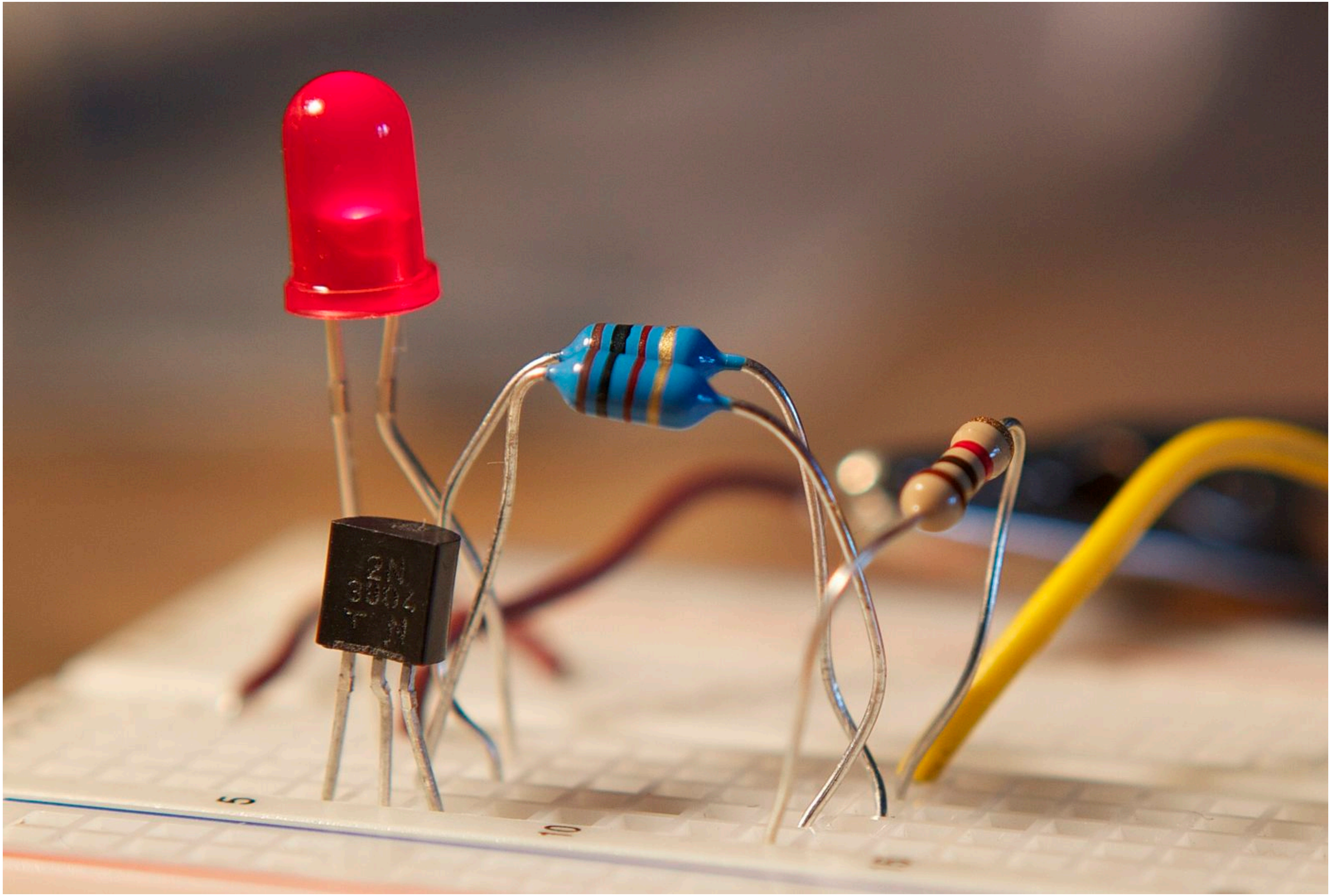
2N
3904
5
10
15

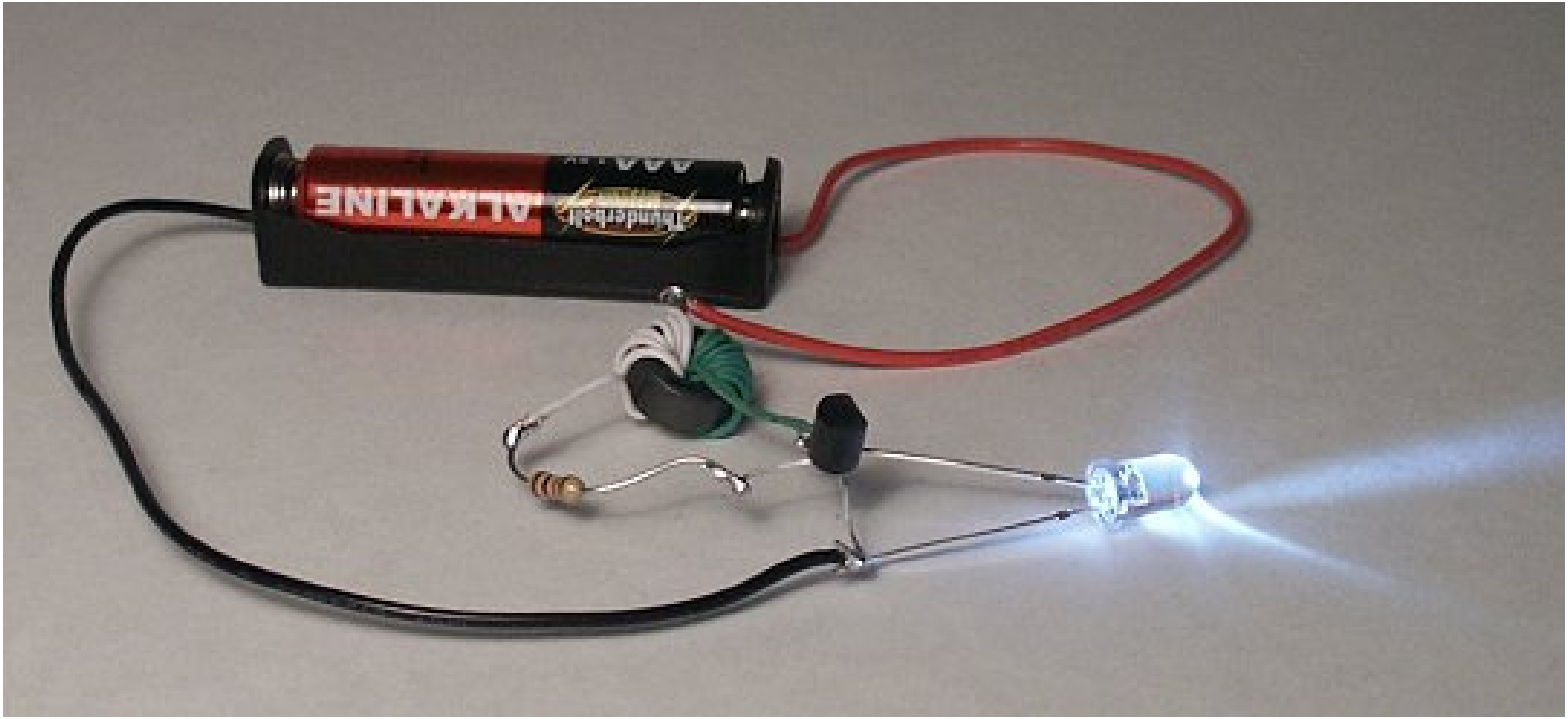
Thunderbolt
ALKALINE
AAA

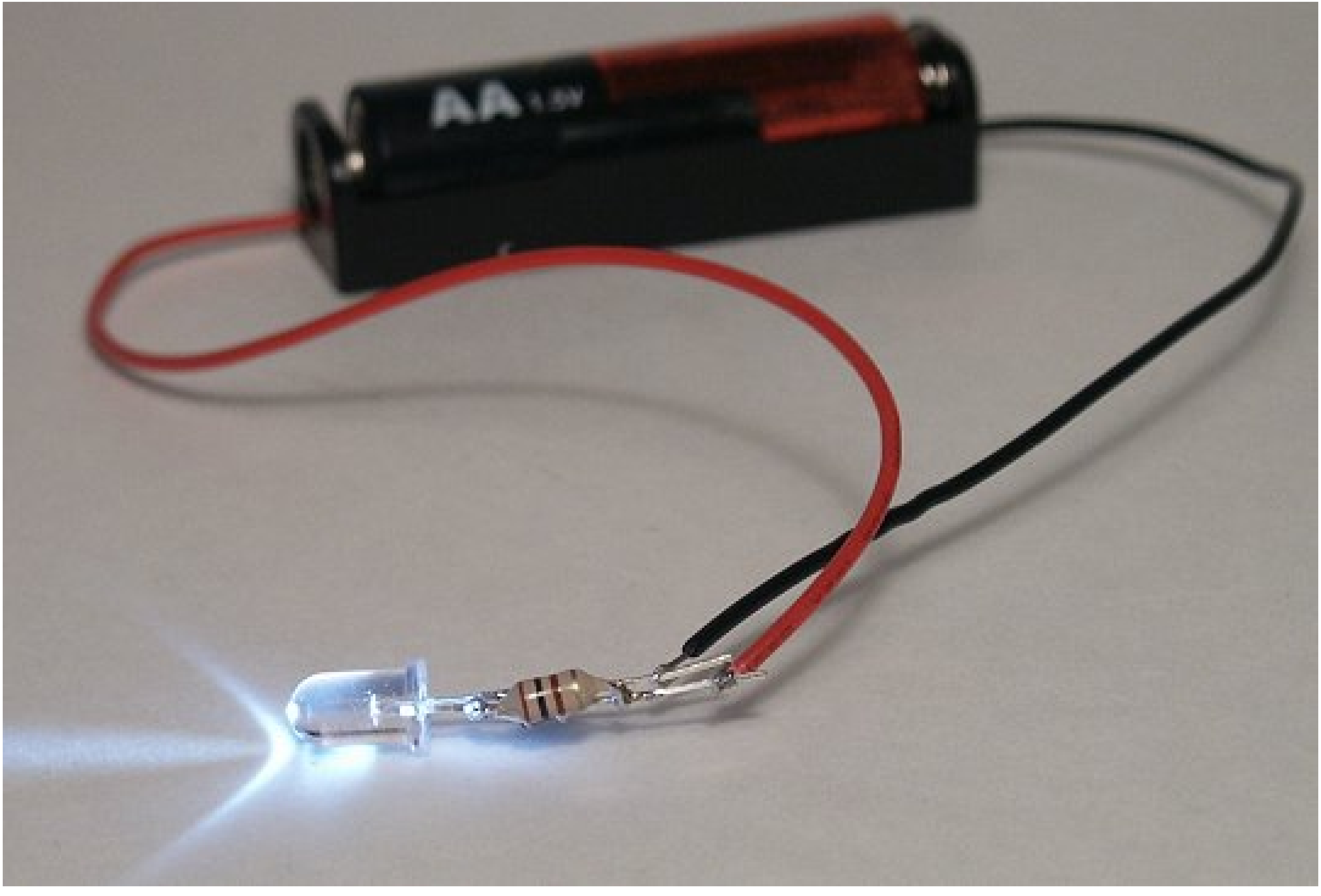
AA 1.5V

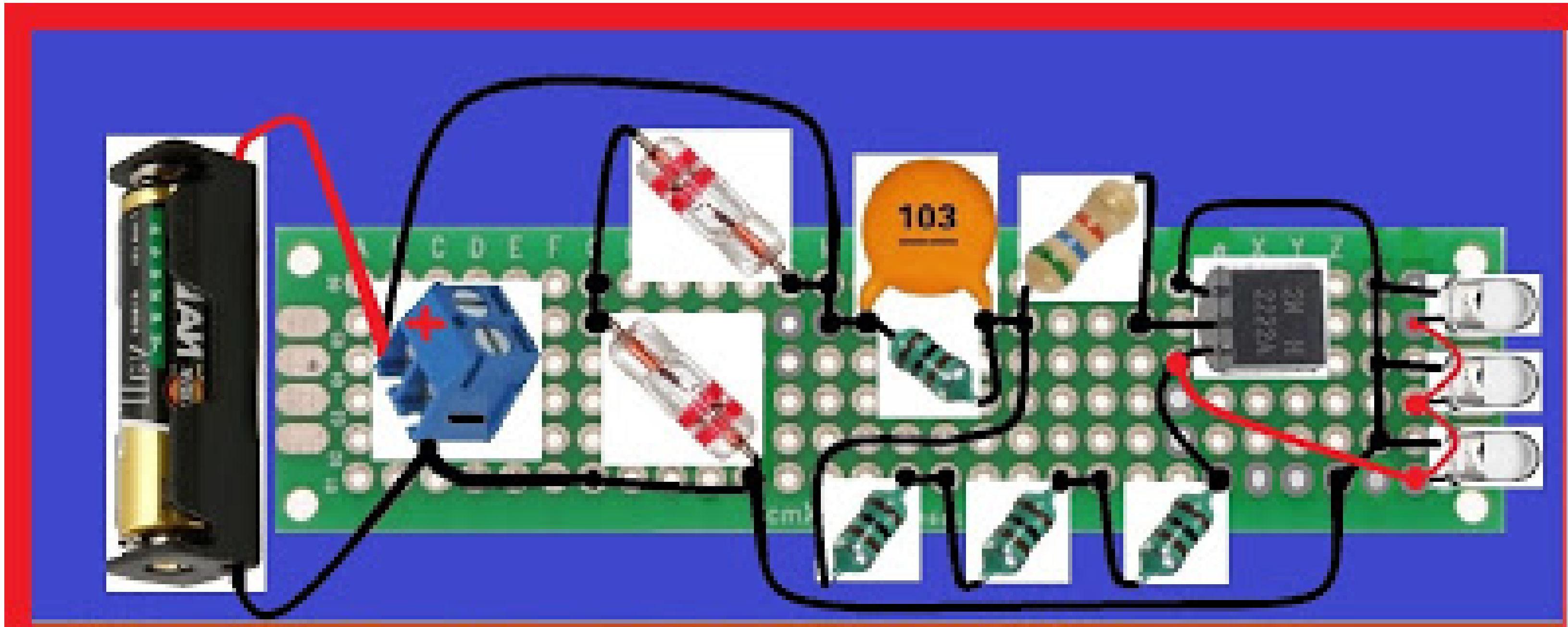

3* Axial lead inductor 82uH 1/4W
1* 10nf ceramic capacitor 103
2* AO90 Germanium signal diode
1* 2N2222 NPN Transistor bipolar 0.6A 40V
1* RESISTOR 5.6K Ohm 1/4W
3* 5mm white Emitting led Diode 20 mA 3.2 V

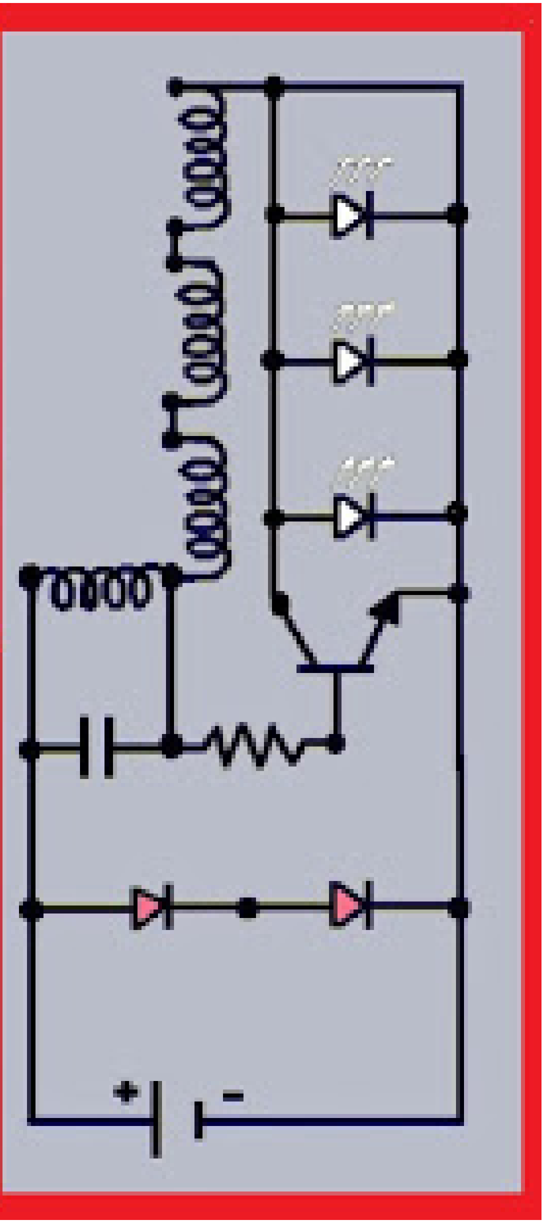

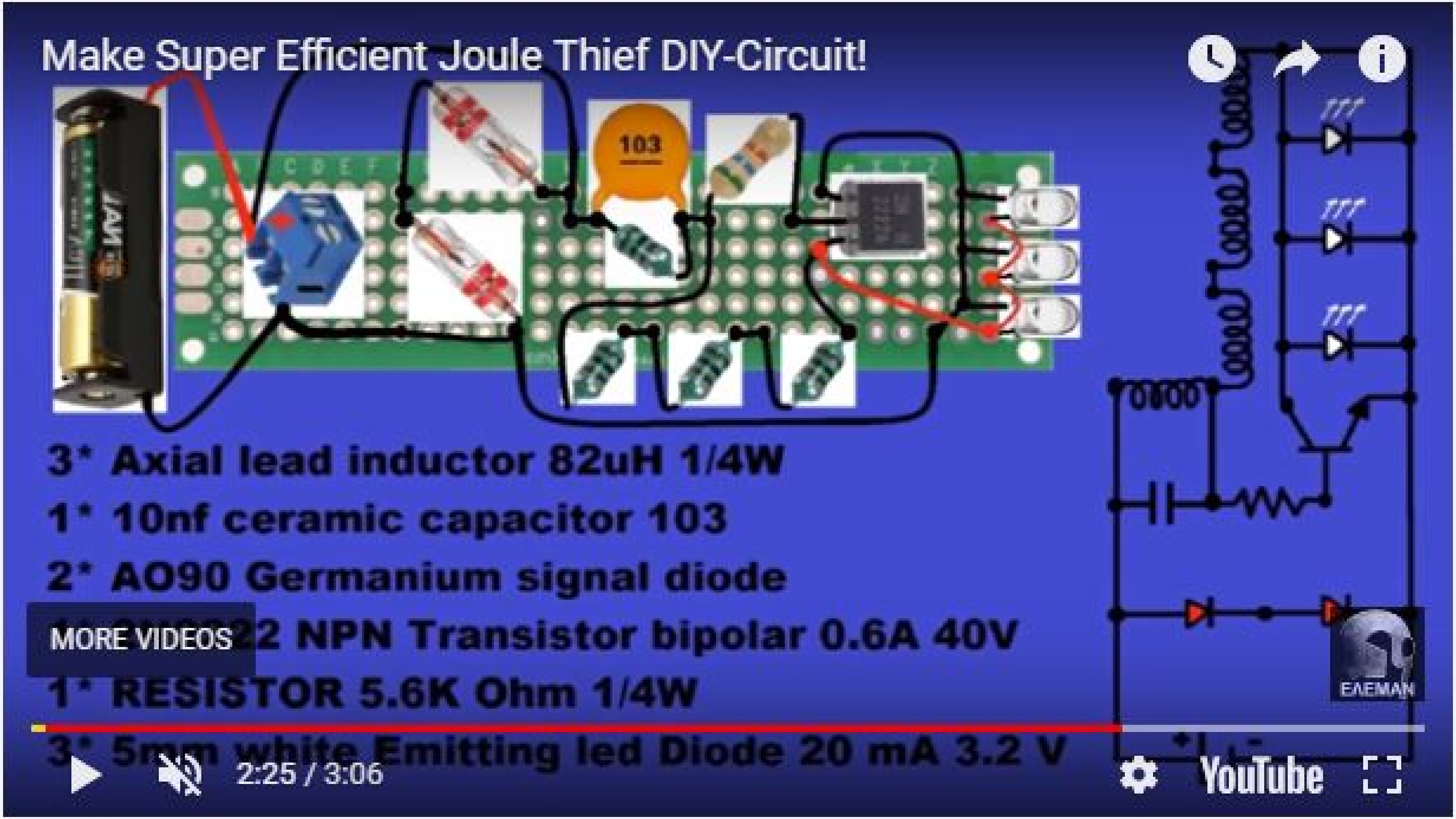

Make Super Efficient Joule Thief DIY-Circuit!

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

22 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

103
MORE VIDEOS
ΕΛΕΜΑΝ
+
-
2:25 / 3:06
YouTube

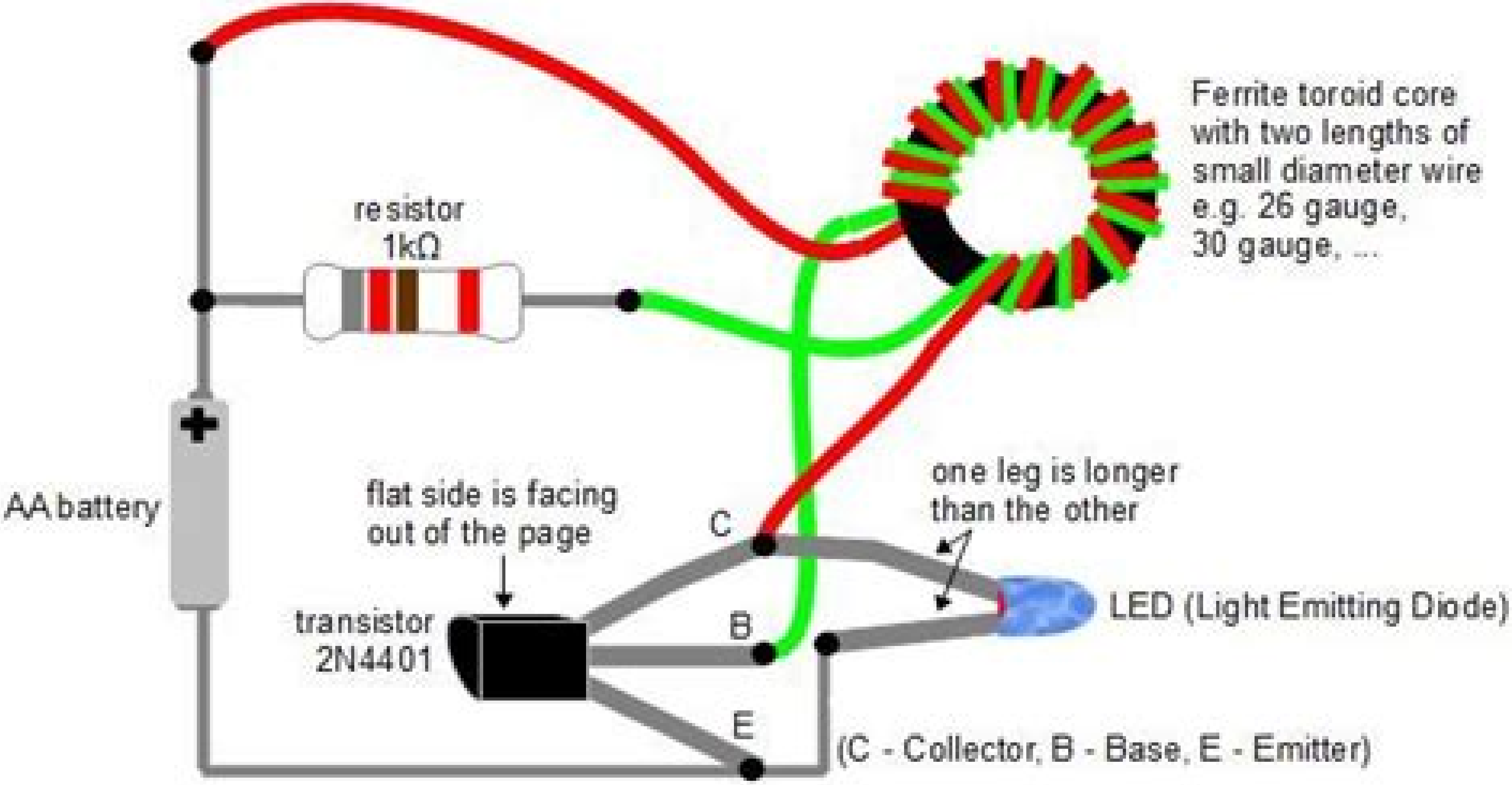
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
resistor
1kΩ
+
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

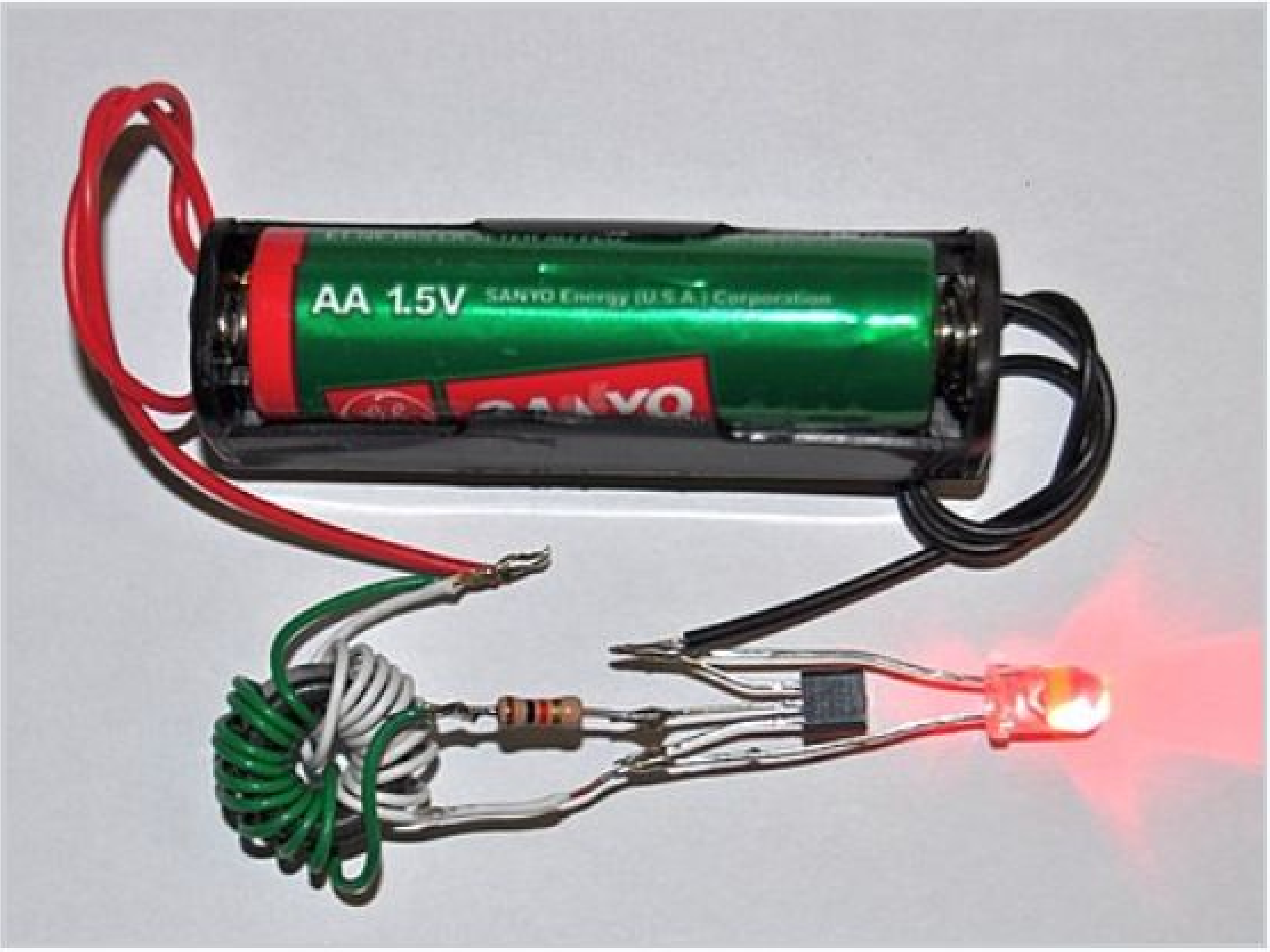
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

A conventional joule thief, shows components and how they are connected. This example uses a red LED. A ferrite toroid is wound to form a coil with primary (white) and feedback (green) windings. A 2N2222A transistor and 1000 ohm resistor are used.

The Joule Thief

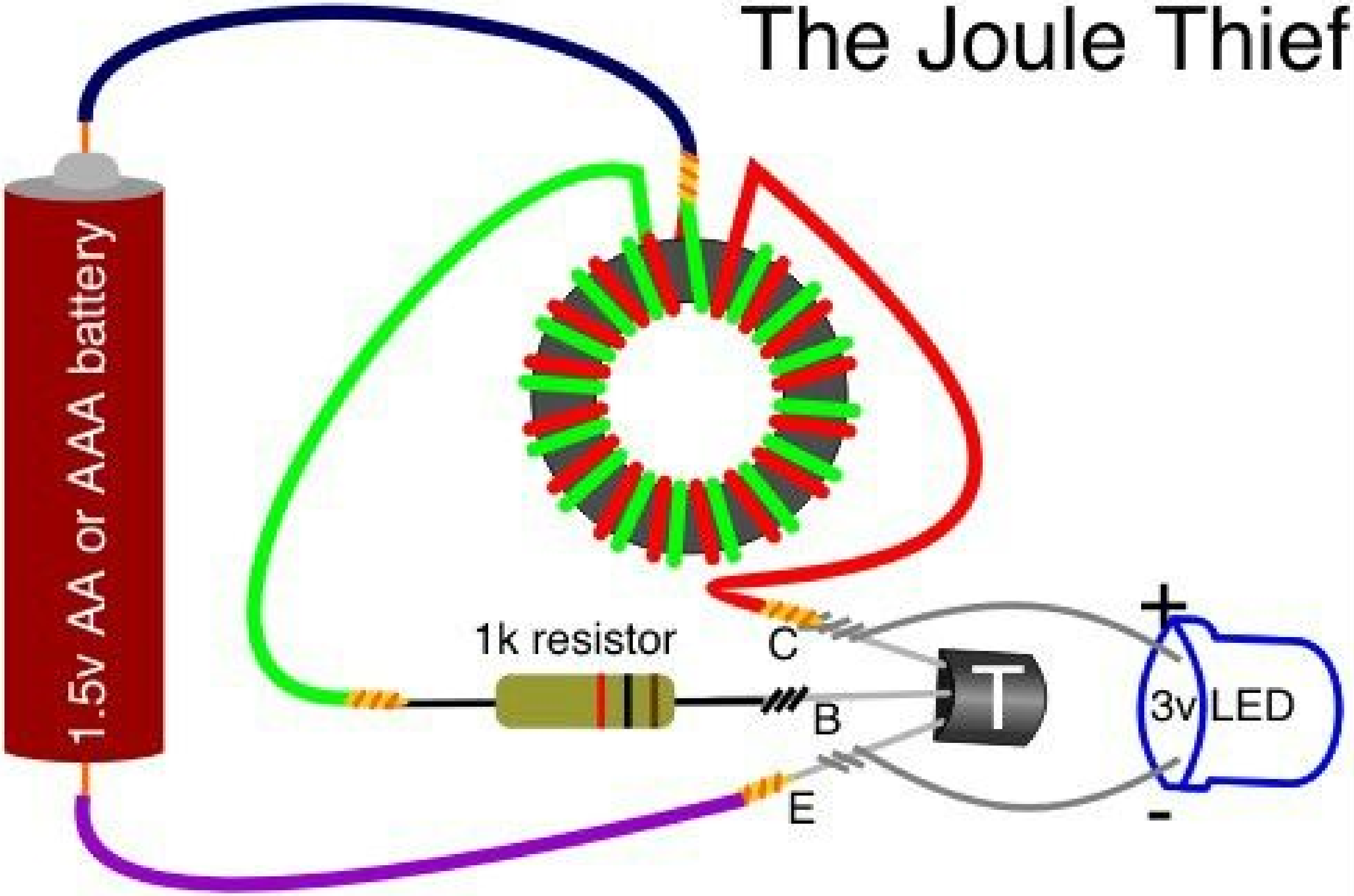
1.5v AA or AAA battery
1k resistor
C
B
E
T
+
-
3v LED

Transistor shown (T) is 2N3904 (can also use 2N2222).

Note LED pins - negative is shorter and has flat on bulb edge.

2 wires round ferrite ring are wound together, then start of one joined to end of other.

Try adding more and more LEDs...

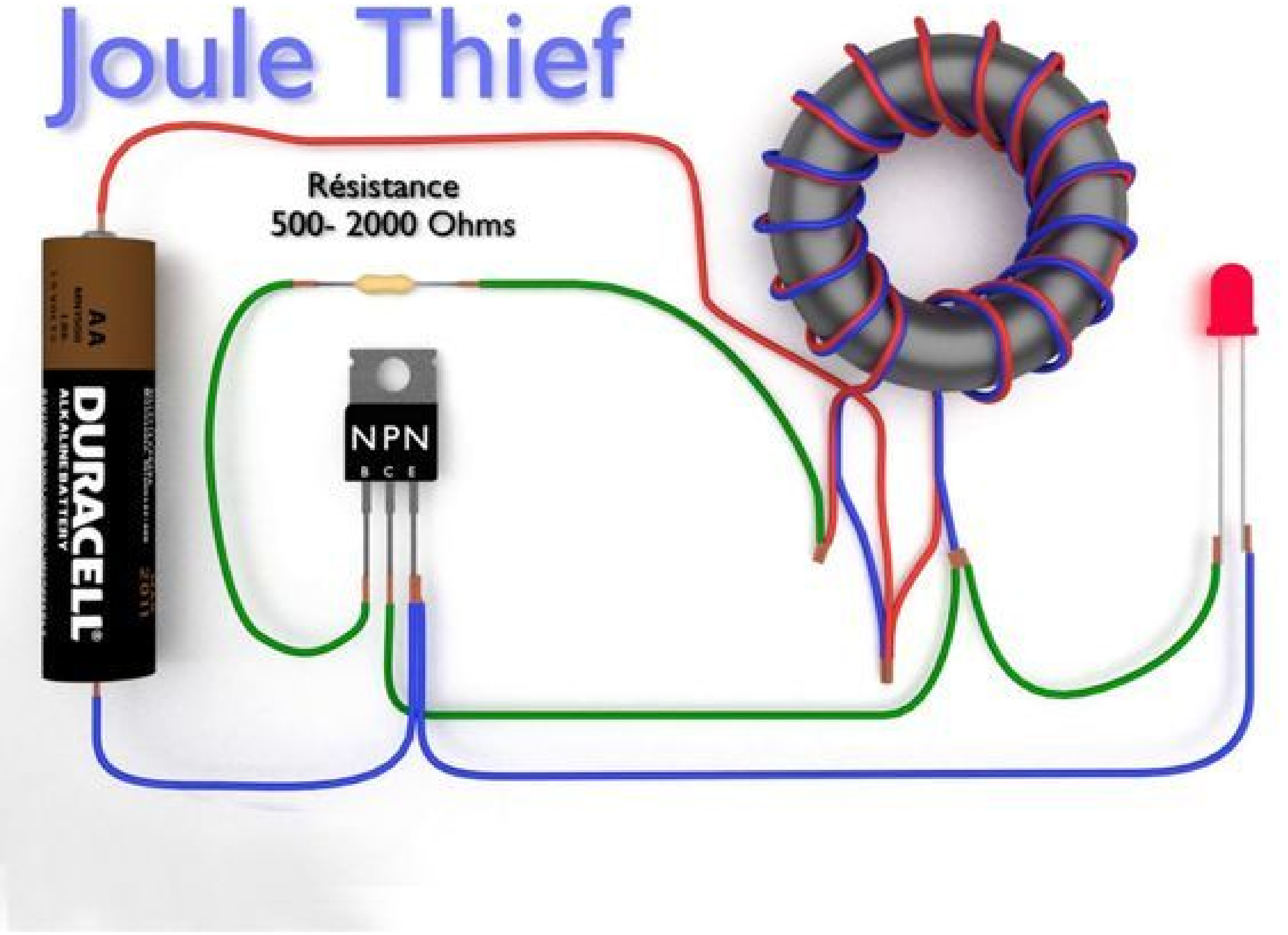
Joule Thief
Résistance
500- 2000 Ohms
NPN
B C E
AA
DURACELL
ALKALINE BATTERY

Its a Rubbish Challenge
Dog Light
Joule Thief
electronics circuit

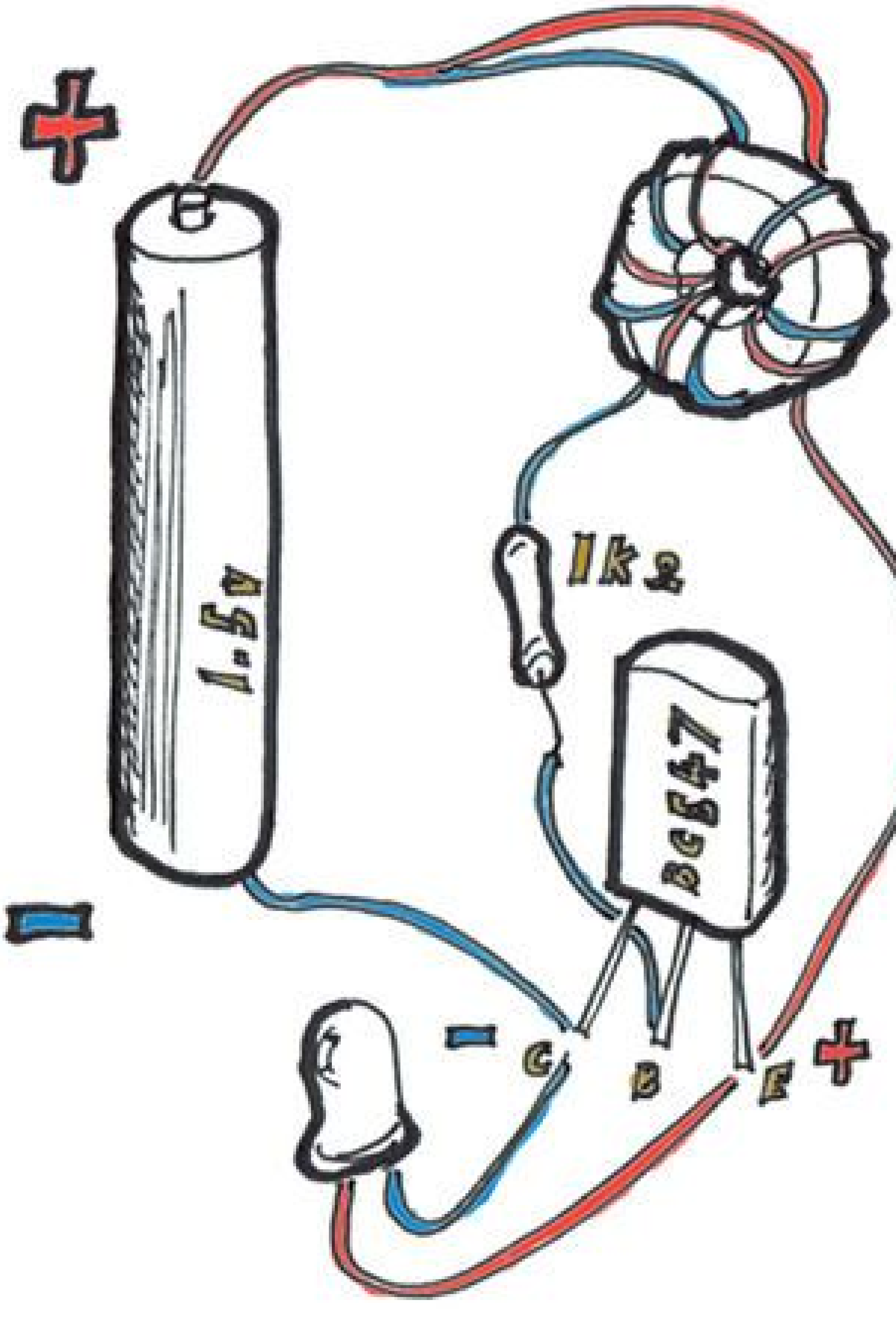
+
1.5v
1kΩ
BC547
–
–
c
b
e
+

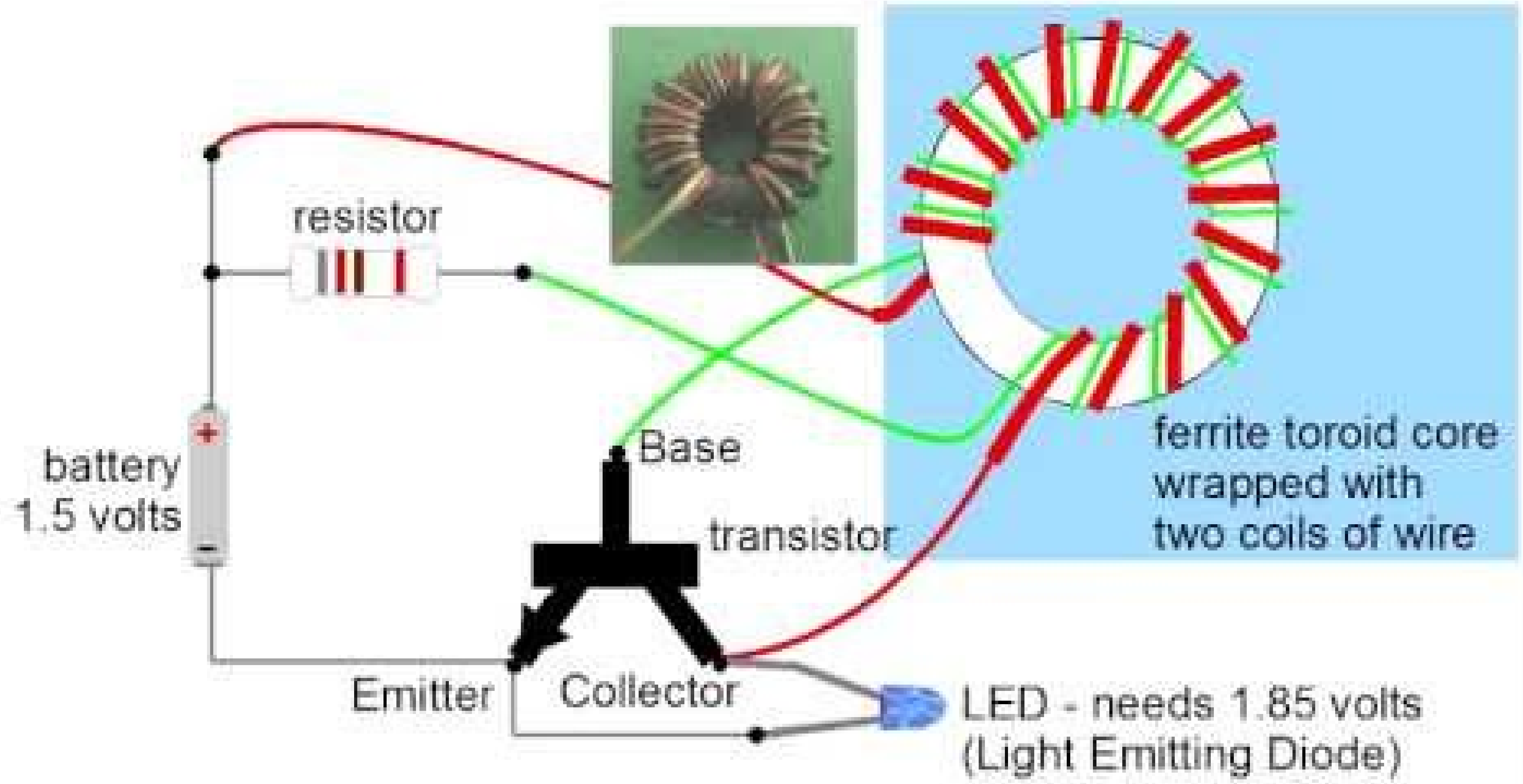
resistor
battery
1.5 volts
+
−
Base
transistor
Emitter
Collector
ferrite toroid core
wrapped with
two coils of wire
LED - needs 1.85 volts
(Light Emitting Diode)

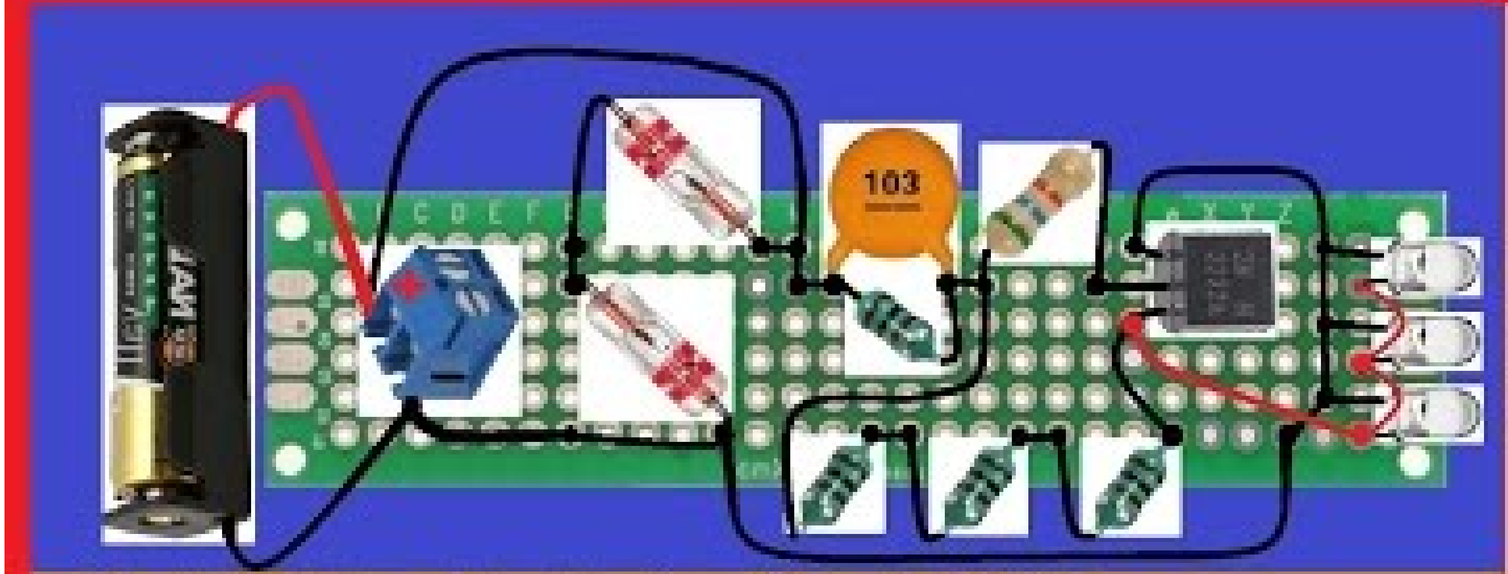

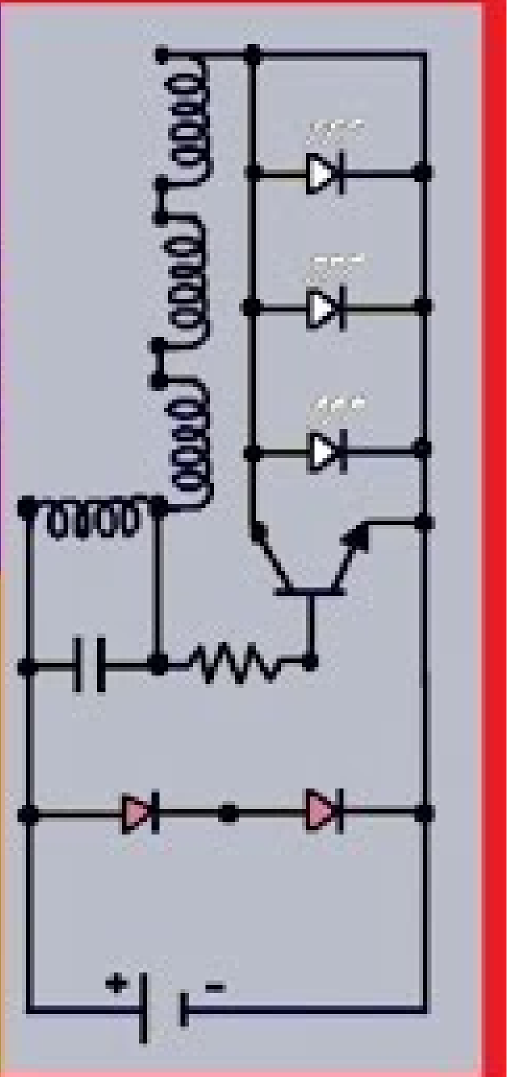

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

1* 2N2222 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

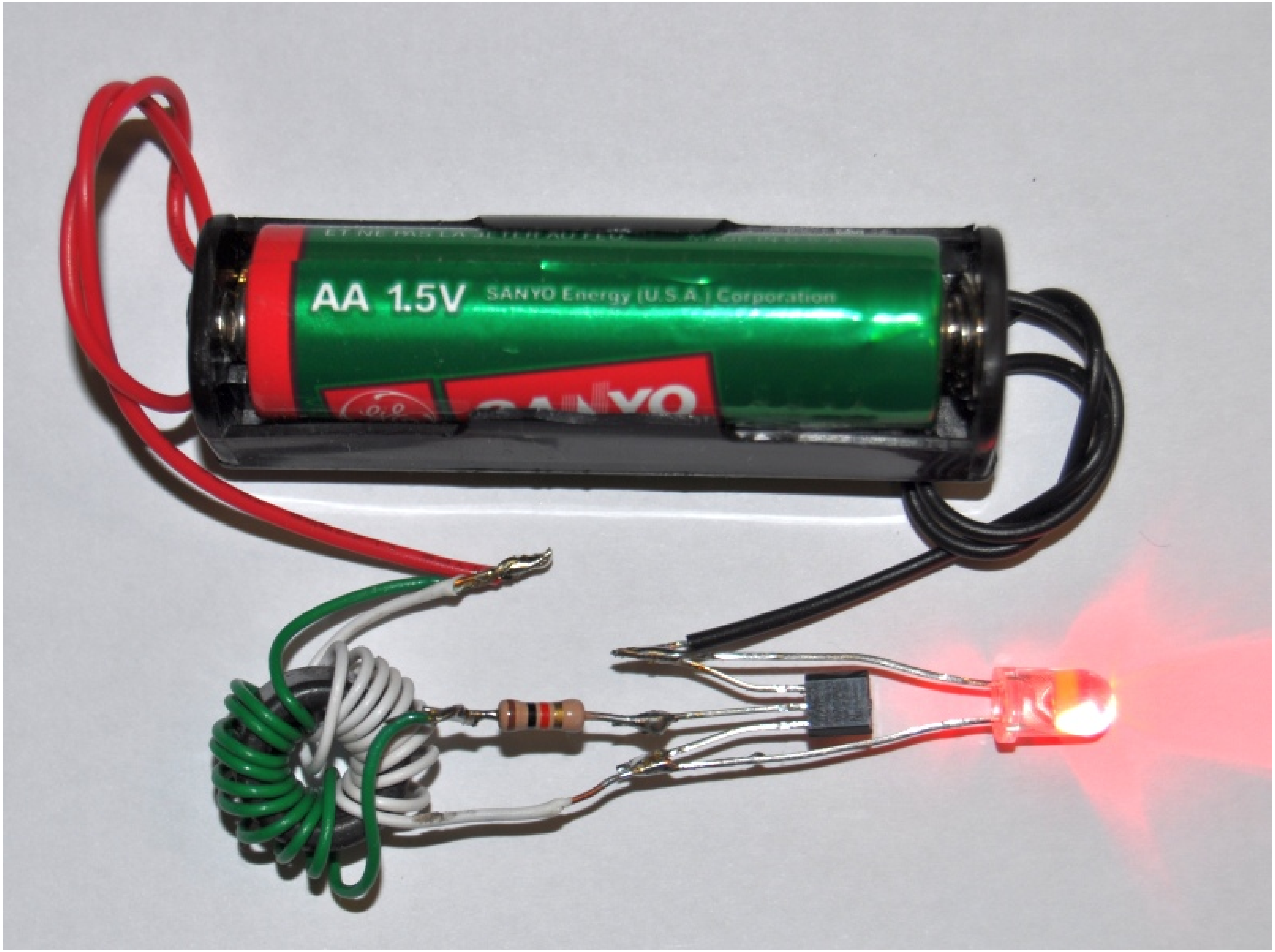
AA 1.5V
SANYO Energy (U.S.A.) Corporation
SANYO

The joule thief circuit.

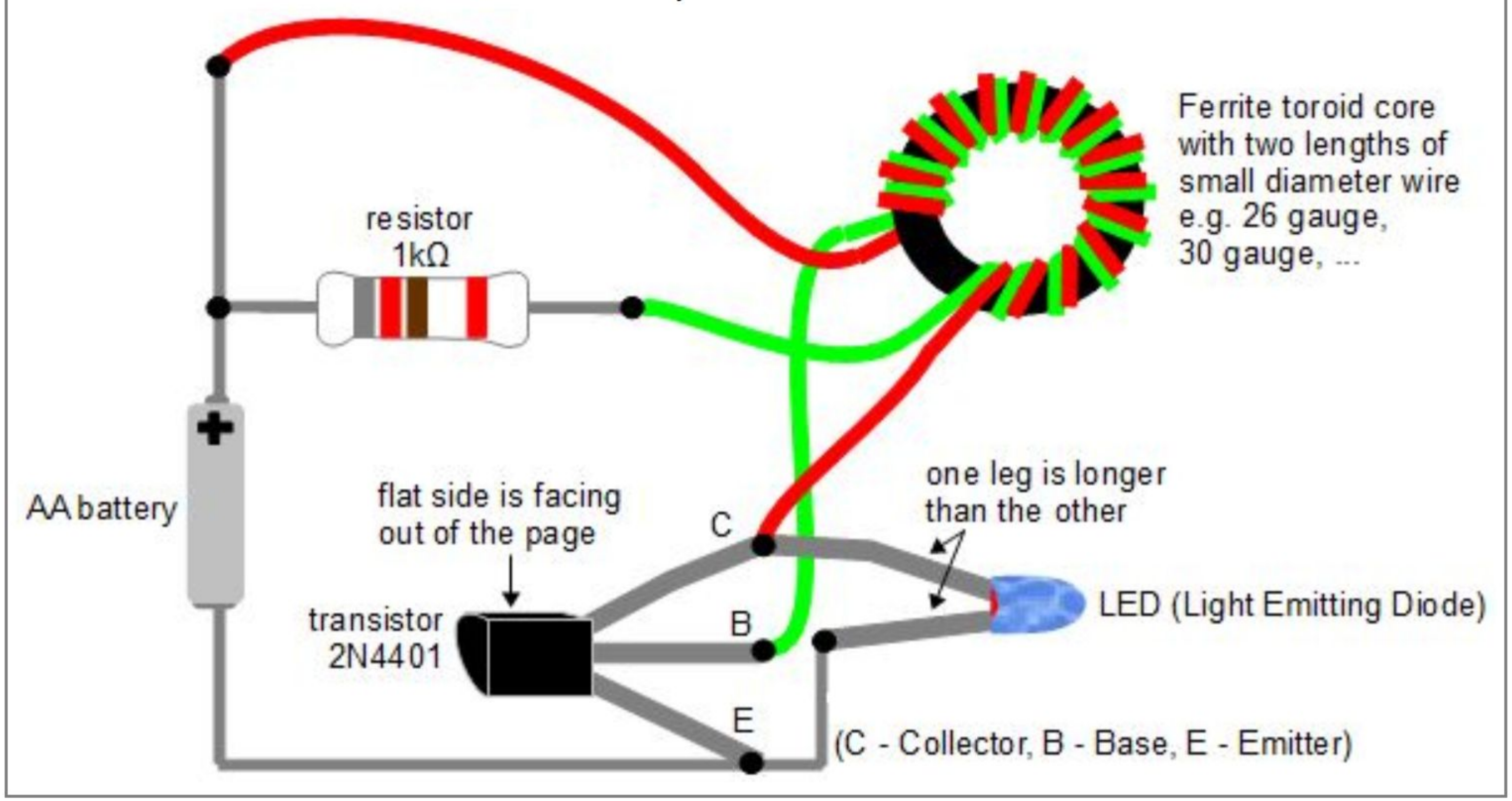
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
resistor
1kΩ
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

Transistor - The legs of the transistor can be determined by noticing that there's a flat side to the transistor case. See the diagram above. A large number of transistors have been reported to work: 2N4401, NET123AP, BC547B, 2SC2500, BC337, PN2222, to name just a few.

LED - One leg of the LED is longer than the other leg. Use this to determine which one goes where. See the diagram above.

Resistor - The diagram says use a 1 kilo ohm resistor but I've used an 820 ohm one just fine. I've also seen a 2 kilo ohm one in use. Use whatever works for you. You can also use a potentiometer (a variable resistor) so that you can easily adjust it to select the resistance that gives the best light.

Toroid ferrite core - Some people have gotten these by opening up compact fluorescent lightbulbs (CFLs). I took mine out of some device whose original function I don't know. To get it working, my first one had just 13 turns for each wire and I used a 30 gauge wire and a 26 gauge wire. The wire must be insulated. A variety of number of turns will work. This is something you can play with. Look at the diagram carefully to determine where the wires connect to.

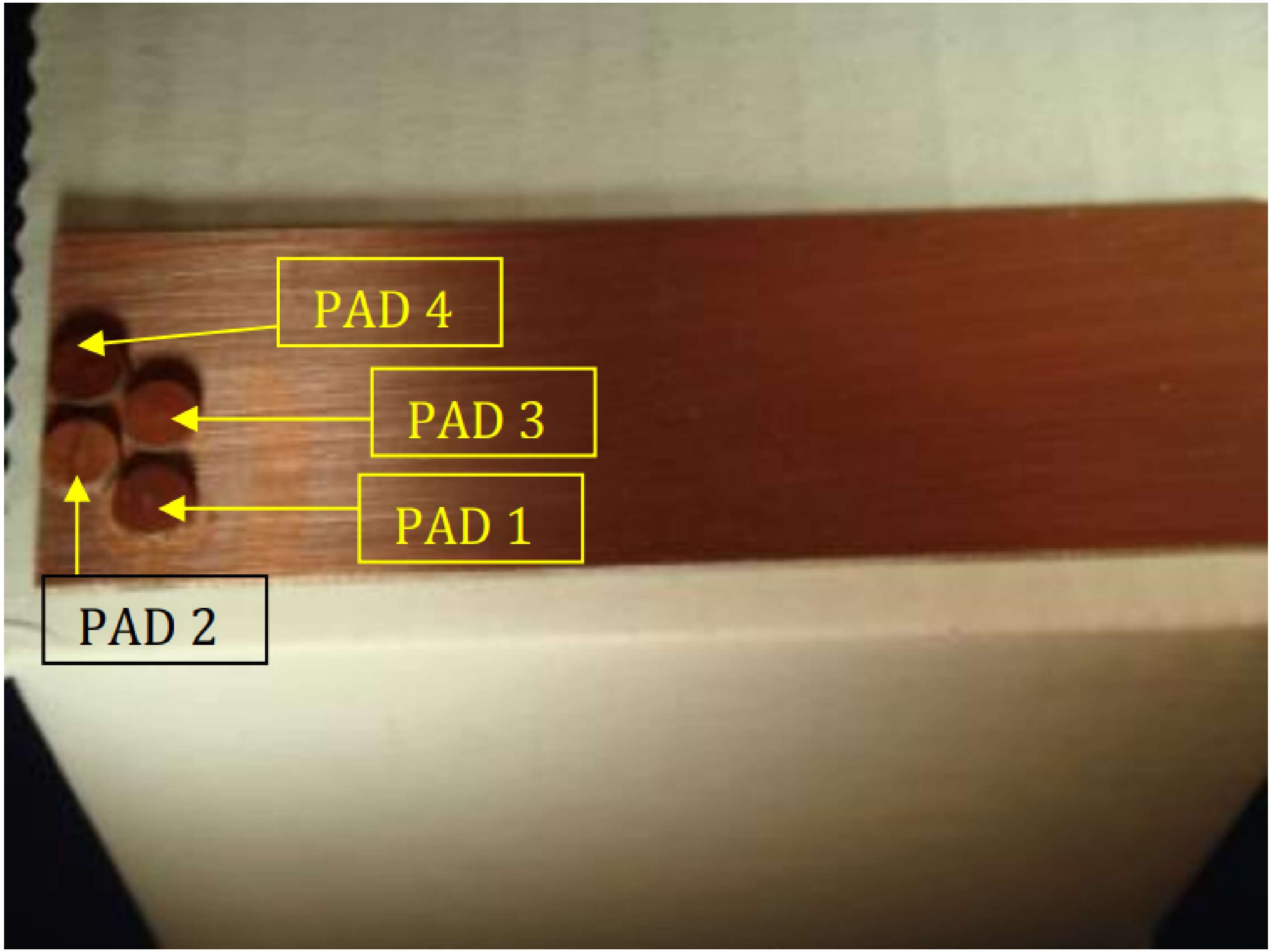

Superglue four pads to the PC board. Make sure they are not touching each other. You should have roughly 1/8 inch between the pads. Apply a small drop of superglue on the board and press the pad onto the board with the small dimple up. Use a small screwdriver to hold pressure on the pad until the glue dries (about 10 seconds).

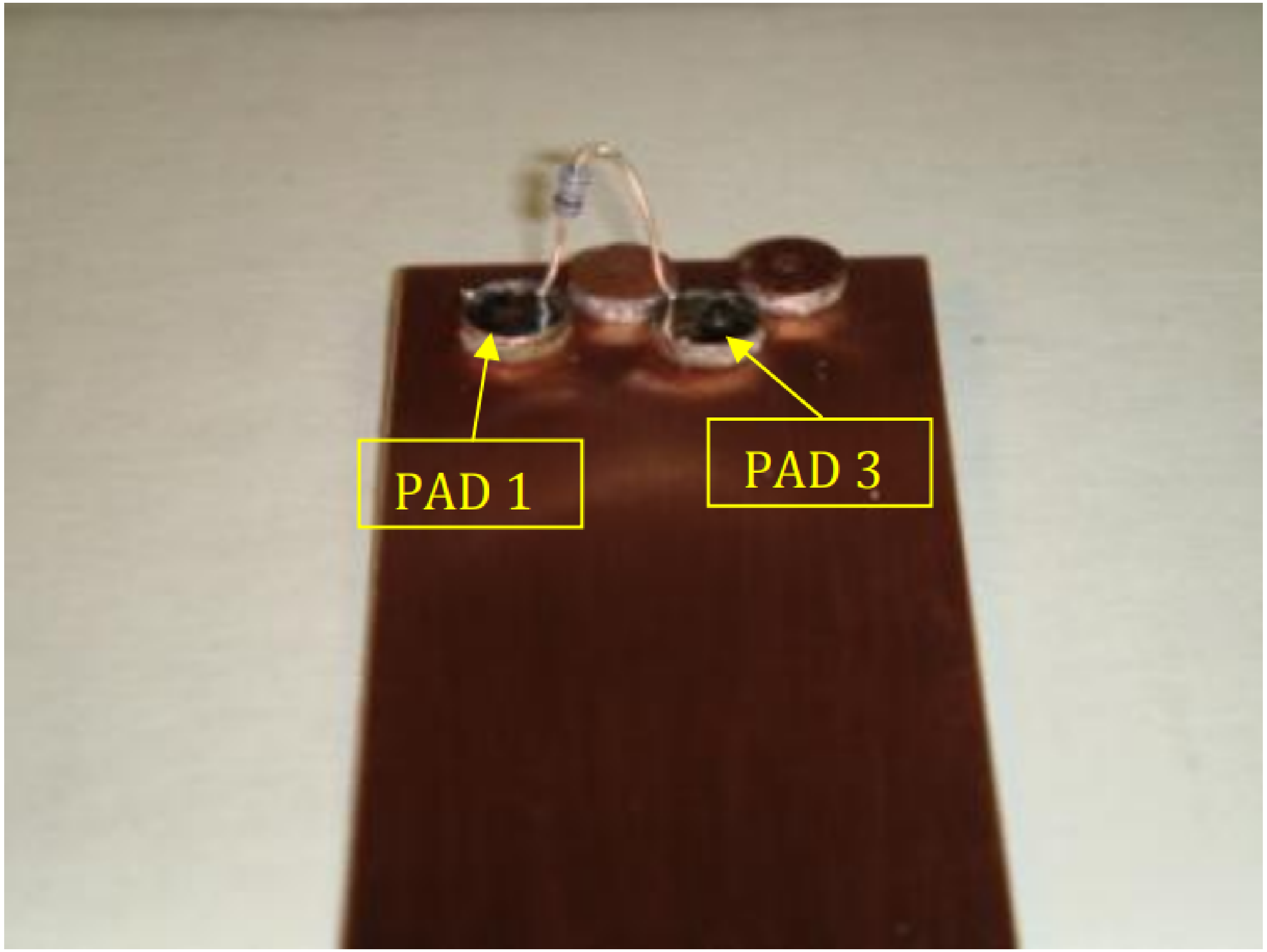

Solder the 1,000 ohm resistor onto pads 1 and 3. When soldering all components, place a 90-degree bend on the ends of the component where they touch the pad to allow better soldering. There is no polarity on a resistor.

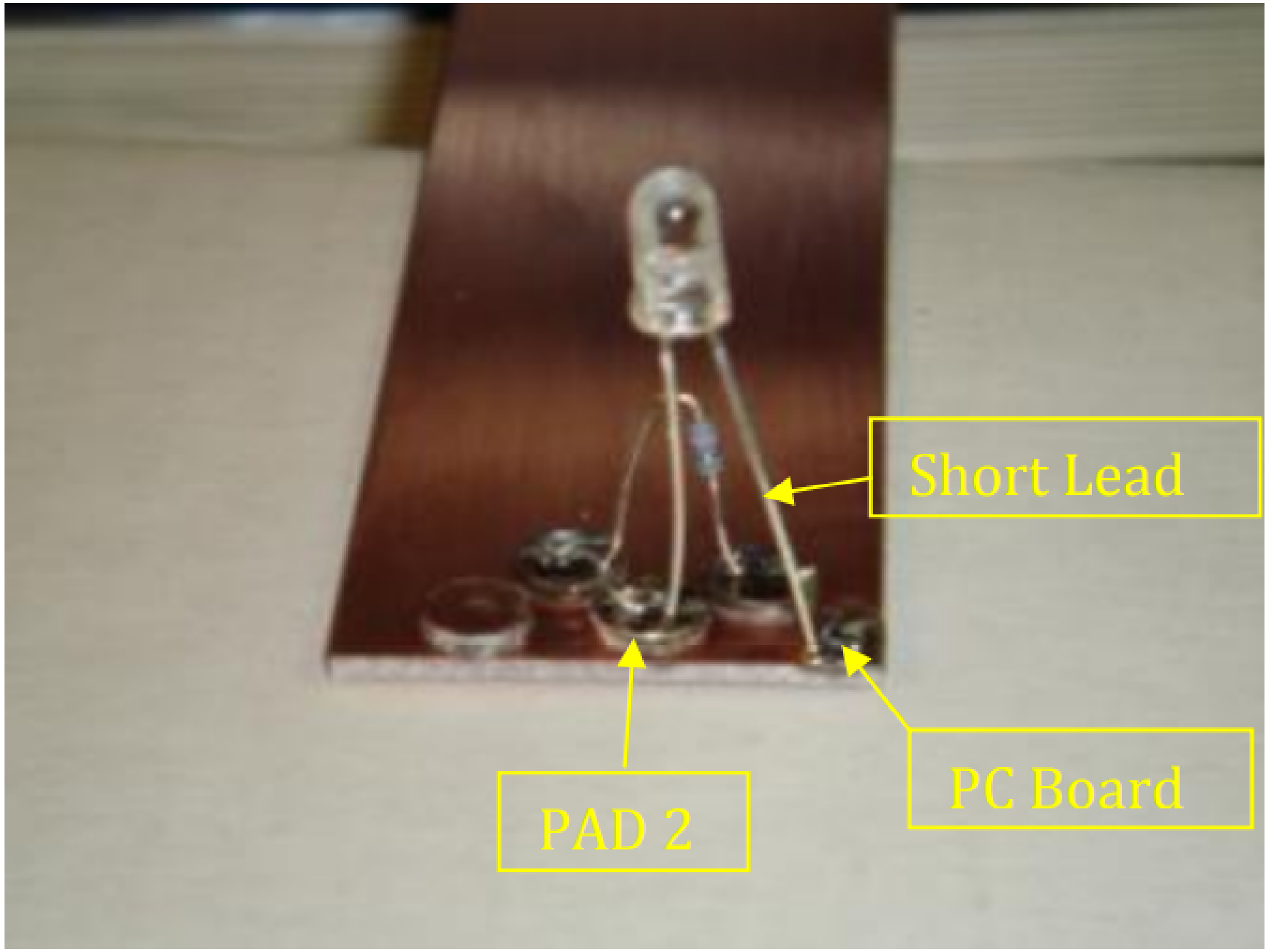

Solder the LED to the board. **VERY IMPORTANT**: Solder the **long lead** to pad 2 and the **short lead** directly to the PC board. The LED will not illuminate of you put it in backwards. If you look into the LED, the anvil shaped part is the lead that gets soldered to the PC board.

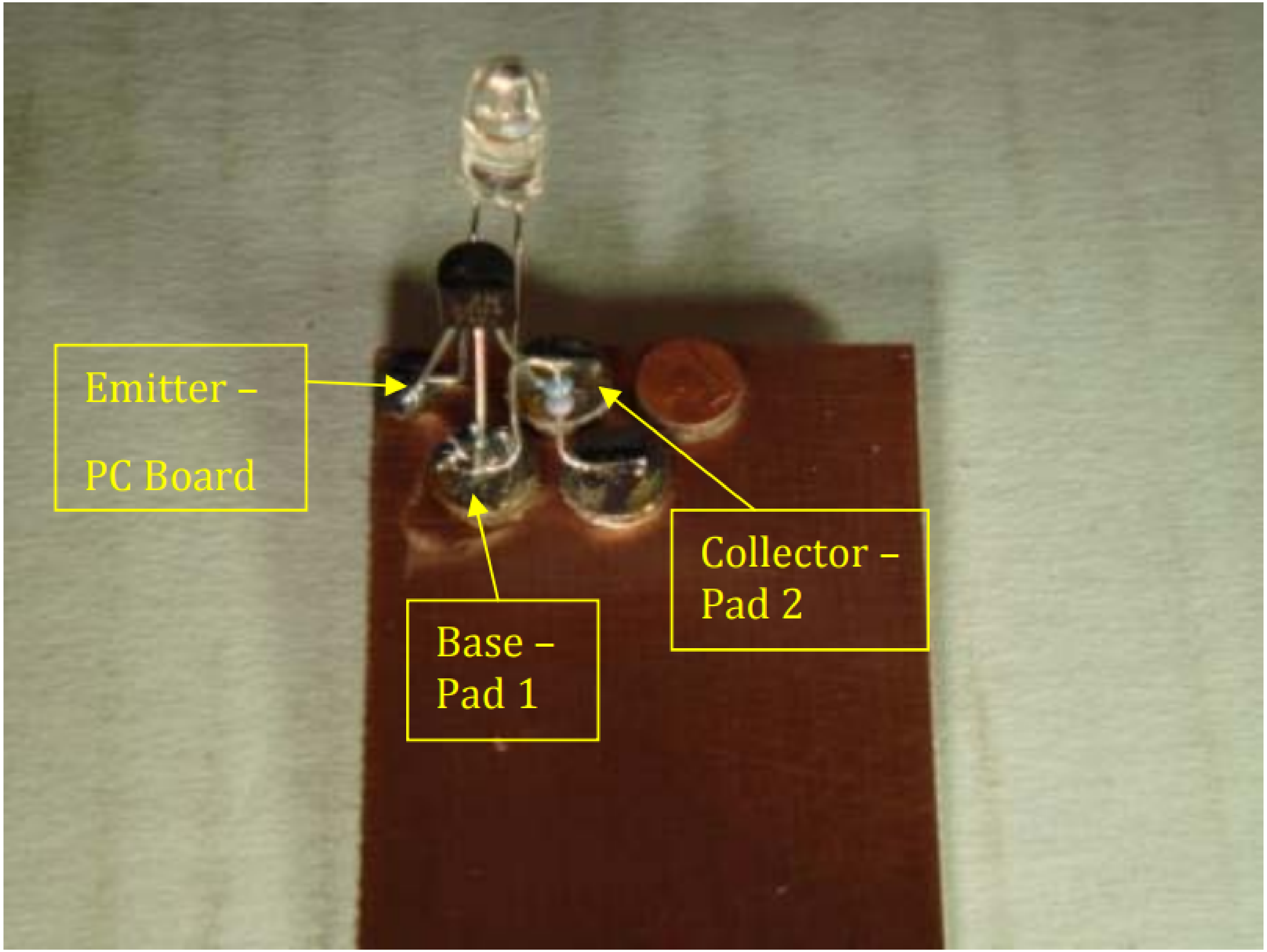
Emitter –
PC Board
Collector –
Pad 2
Base –
Pad 1

Solder the transistor to the board. With the flat part facing you the left lead (emitter) is soldered directly to the PC board. The center lead (base) is soldered to pad 1. The right lead (collector) is soldered to pad 2. Failure to install this part correctly will prevent the Joule Thief from working.

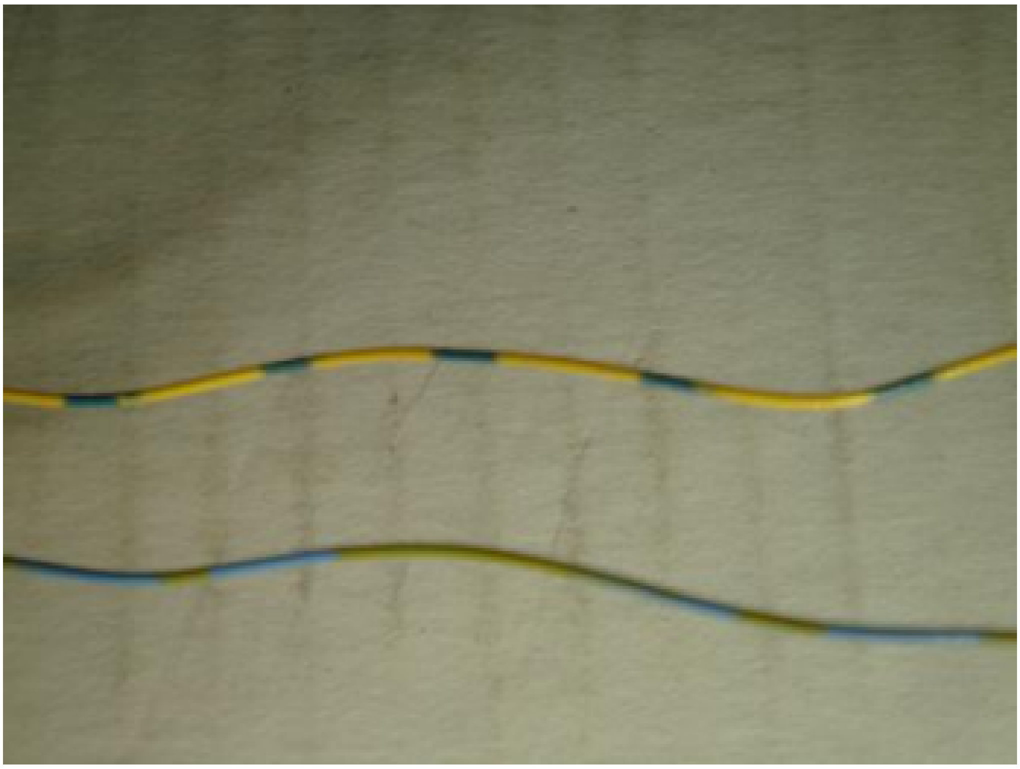

Take the two pieces of wire and **untwist** them if they are twisted. The toroid requires 12 inches of wire. You will wind the toroid with this pair of wires side by side.

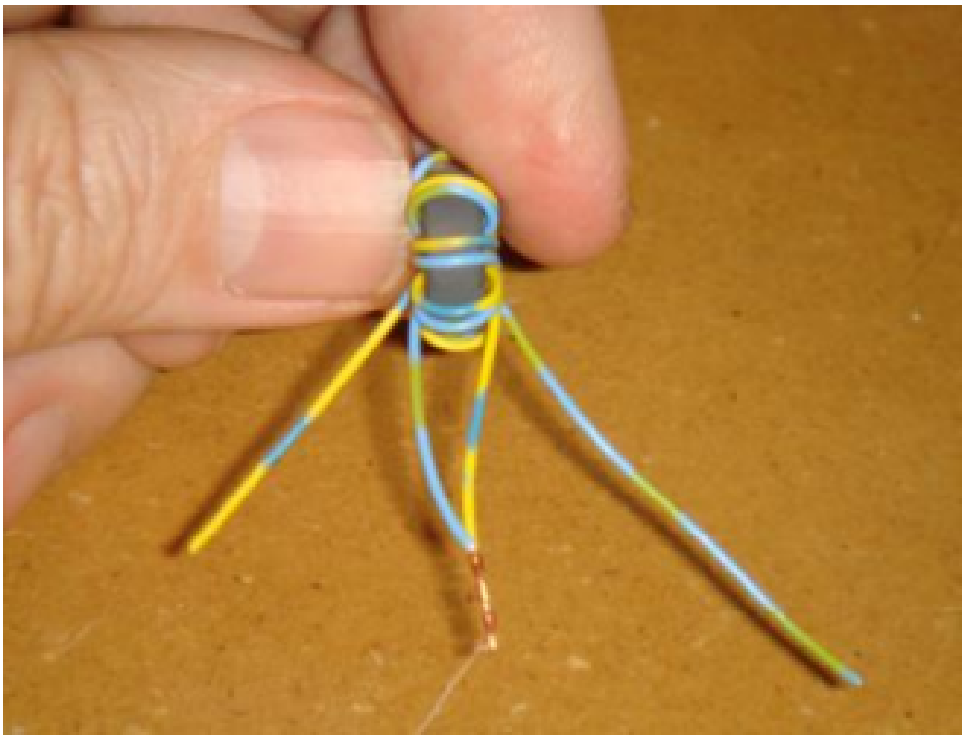

Wind the toroid with **9 turns** of the wire. Each turn through the hole in the center counts as one turn. **VERY IMPORTANT:** You must take one wire **from each side of the winding** and twist them together. This makes the transformer windings out of phase. This is critical to operation. Connect one wire of each color.

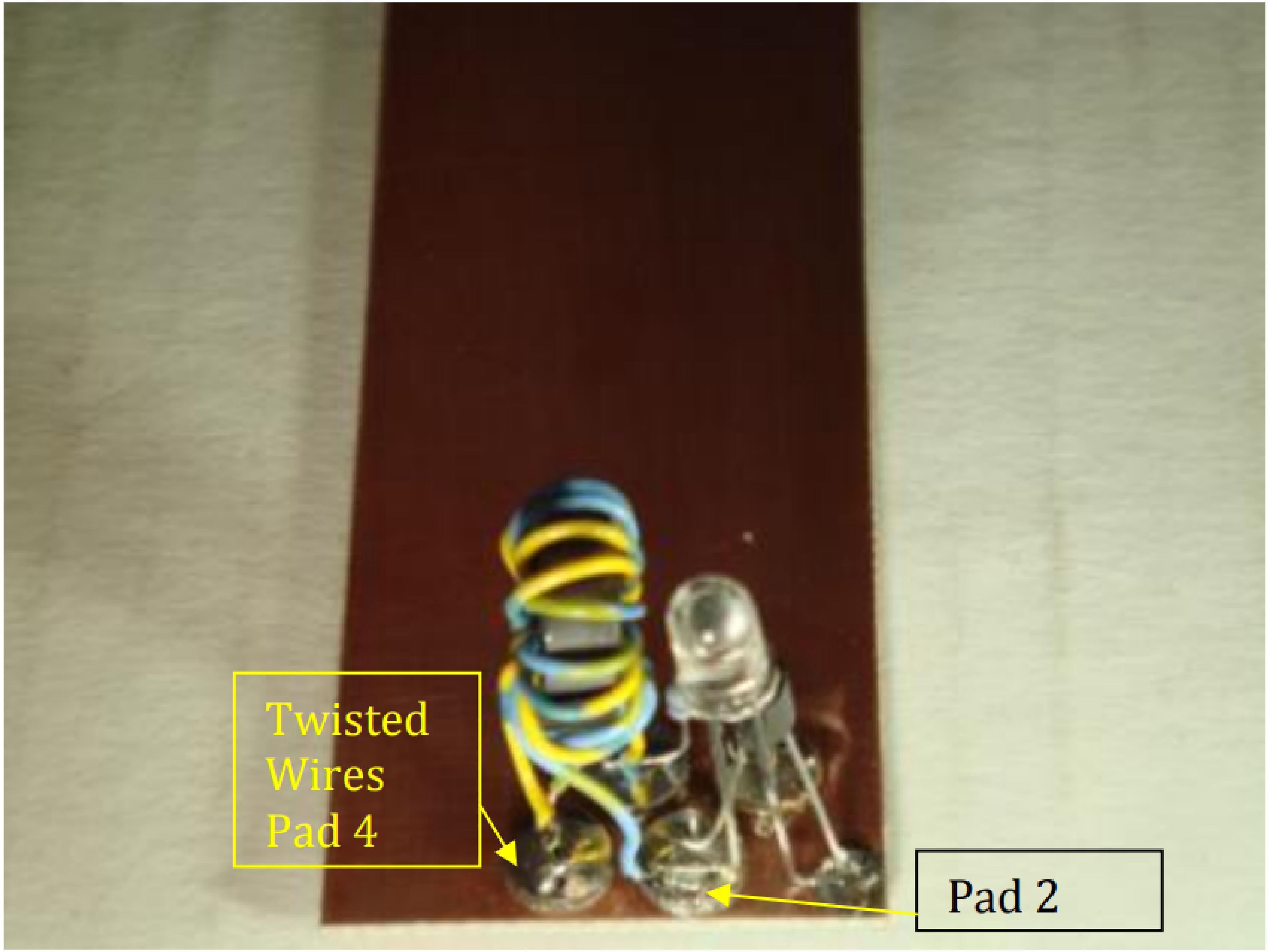
Twisted
Wires
Pad 4
Pad 2

Solder the twisted wires on the toroid to pad 4. Solder one of the other toroid wires (doesn't matter which one) to pad 2.

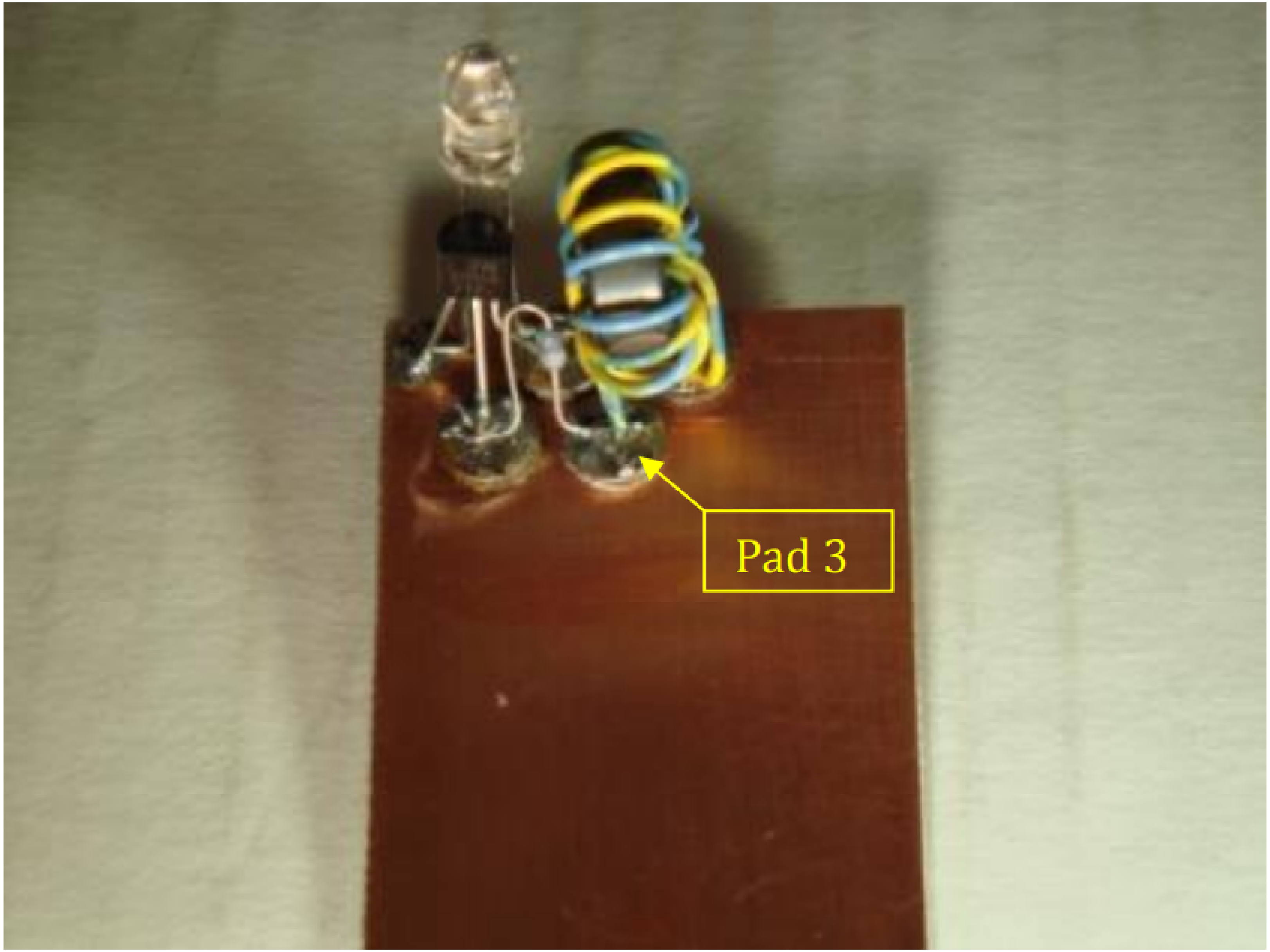

Solder the other wire on the toroid to pad 3.

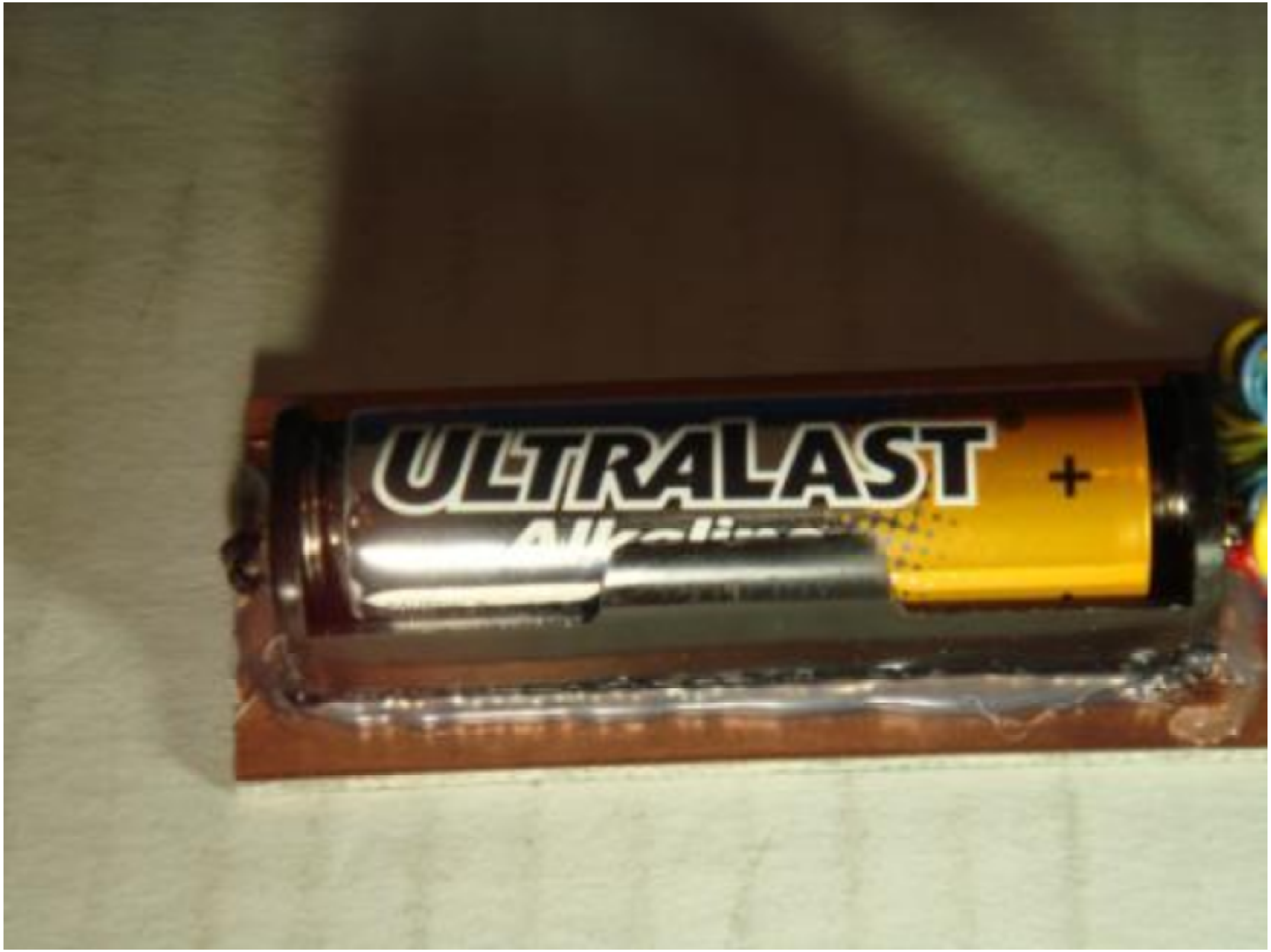

Solder the battery holder. Connect a short piece of wire from the positive terminal (the one without the spring) to pad 4. Solder the negative terminal (the one with the spring) directly to the PC board using a short piece of wire. Use hot glue or some other glue to secure the battery holder to the PC

Step 1 – Wound the Toroid

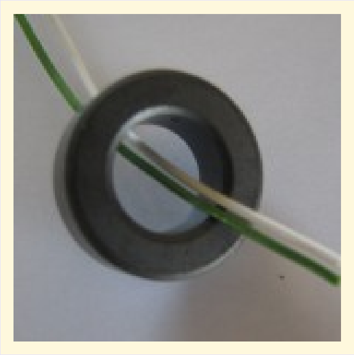

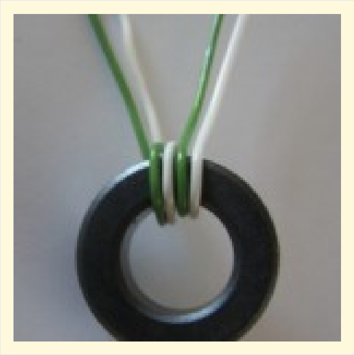

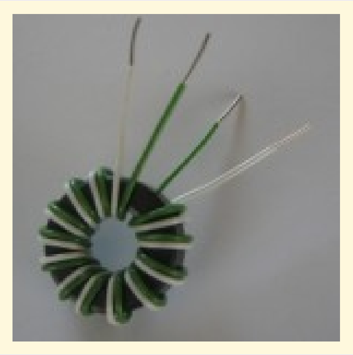

Take the two strands of plastic insulated wire and hold them together. Start off by sticking then through the middle of the toroid until you have 2 cm of wire left to make a connection with later. Keep the two wires together and wrap them around the toroid until you have covered the whole toroid. If you have wire left you can continue until you have about 2 cm of wire free on either side. Remove about a 0.5 cm of the plastic insulation on all four of the wires as in the photo. You have now wound your toroid and we are ready to continue.

Step 2 – Join the two coils of Toroid

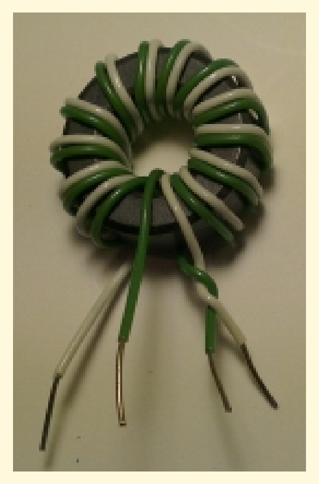

The next step is to join the two coils or windings of the toroid. You may have noticed the two dots next to the toroid windings in the circuit diagram. These two dots indicate the polarity of the transformer and as you can see the two windings are opposite to each other. To make sure that the polarity of our toroid windings are correct we must do the following:

- If we look at the toroid we noticed that we have two wires of different colours sticking out at both ends of the toroid.
- Take one wire of a specific colour (green in the example) on the one side of the toroid and one wire of the other colour (white in the example) on the other side of the toroid and connect (or solder) them together. These two wires forms the top pole of the toroid as in the circuit diagram and is connected to the positive terminal of the battery.

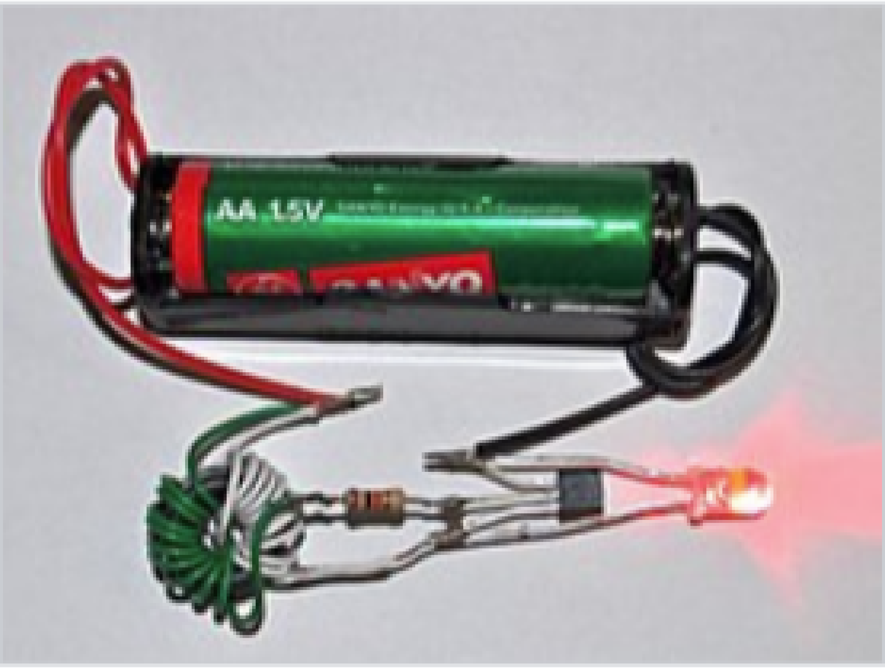

A conventional joule thief, shows components and how they are connected. This example uses a red LED. A ferrite toroid is wound to form a coil with primary (white) and feedback (green) windings. A 2N2222A transistor and 1000 ohm resistor are used.

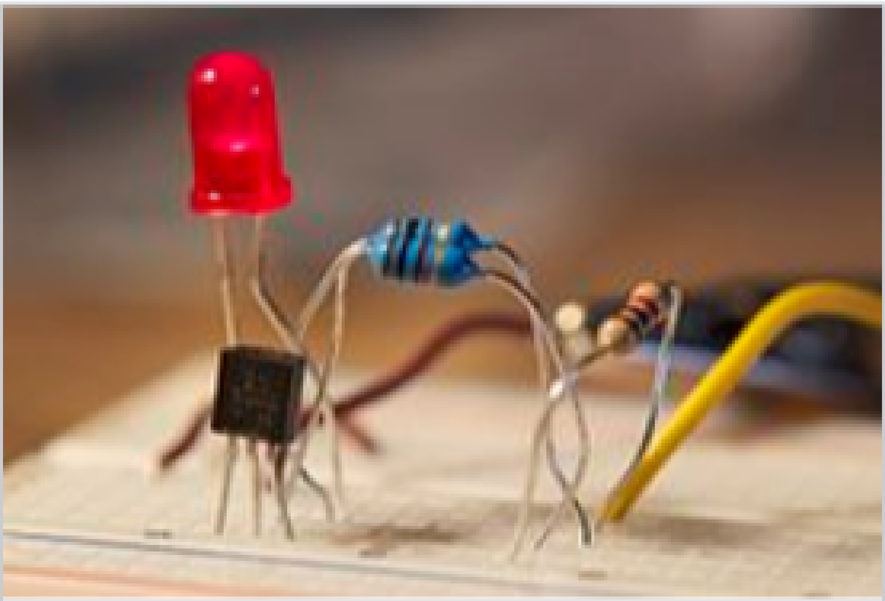

A joule thief with two axial inductors replacing the ferrite toroid, shown on a solderless breadboard

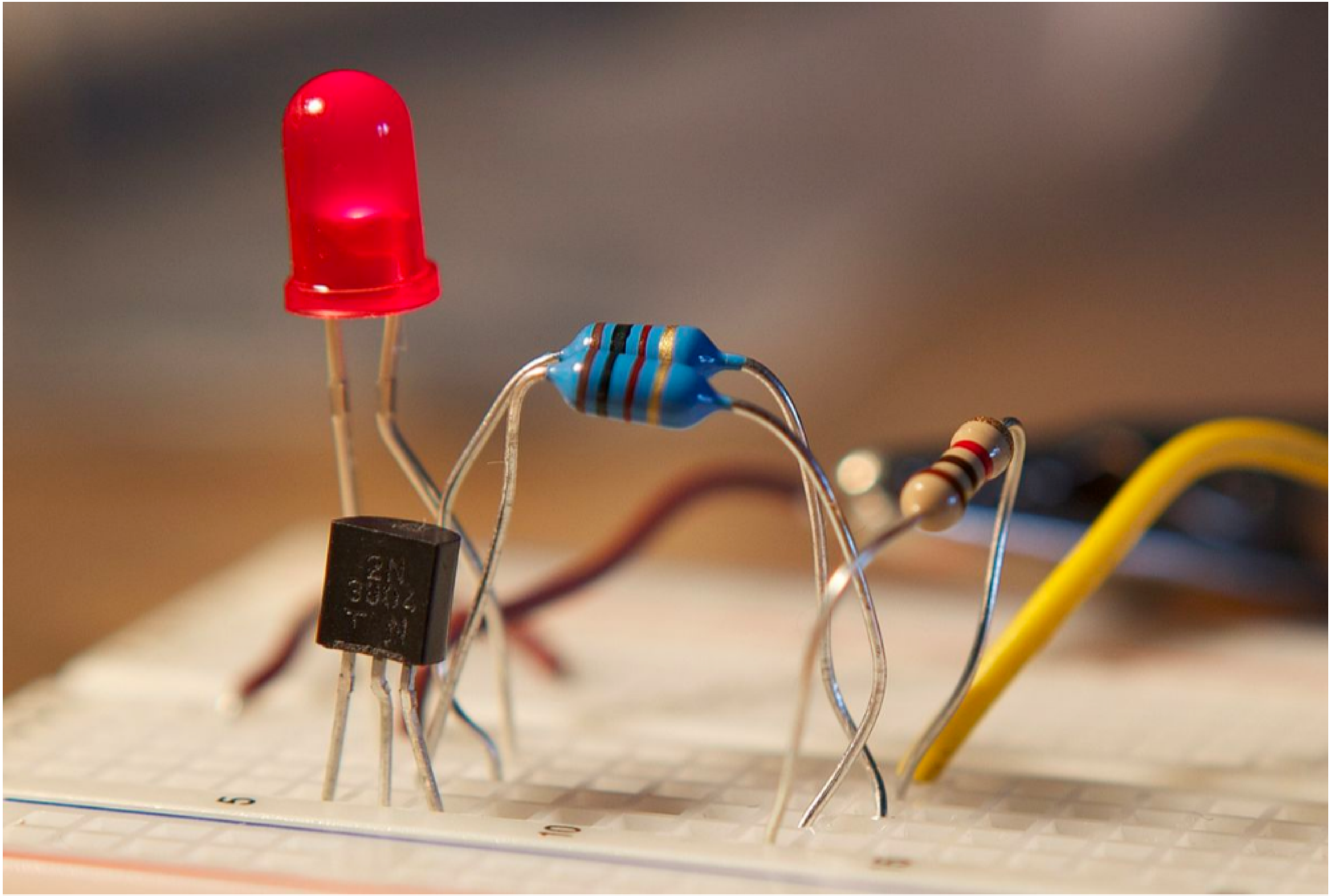
2N
3904
T N
5
10

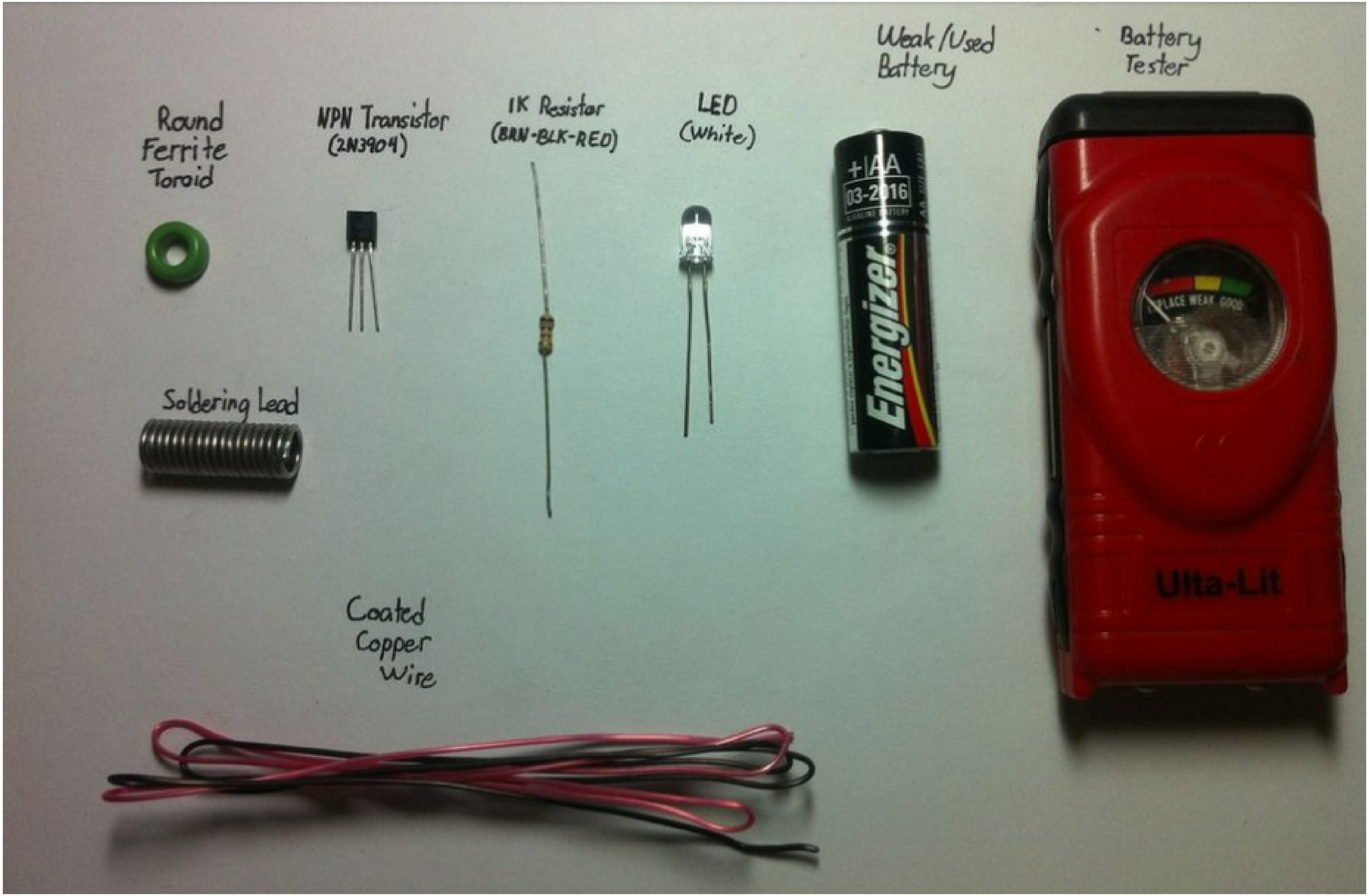
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
Soldering Lead
Coated Copper Wire
+AA
03-2016
Energizer
REPLACE WEAK GOOD
Ulta-Lit

Step 2: Schematic Diagrams

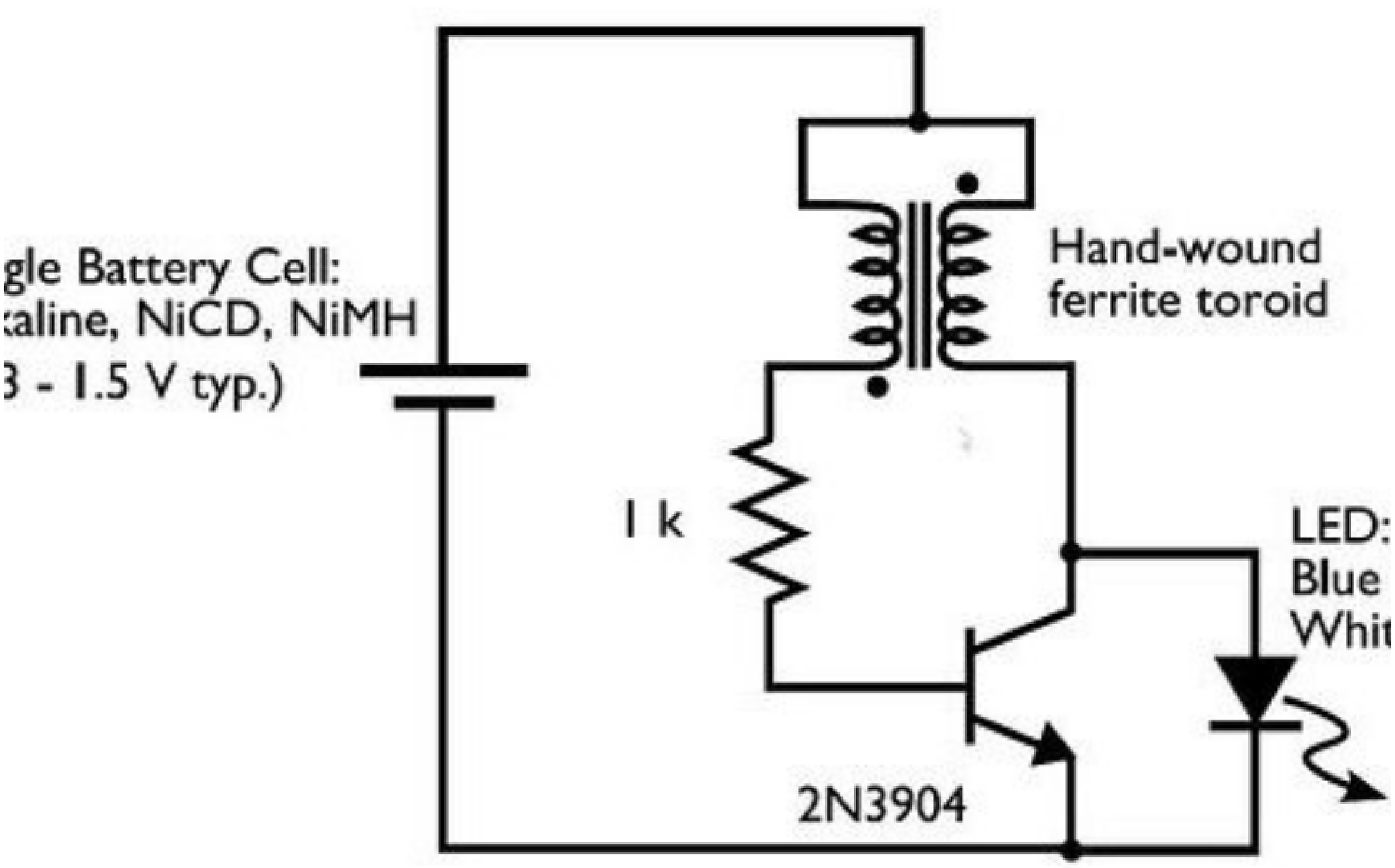
gle Battery Cell:
aline, NiCD, NiMH
3 - 1.5 V typ.)
Hand-wound
ferrite toroid
1 k
2N3904
LED:
Blue
Whit

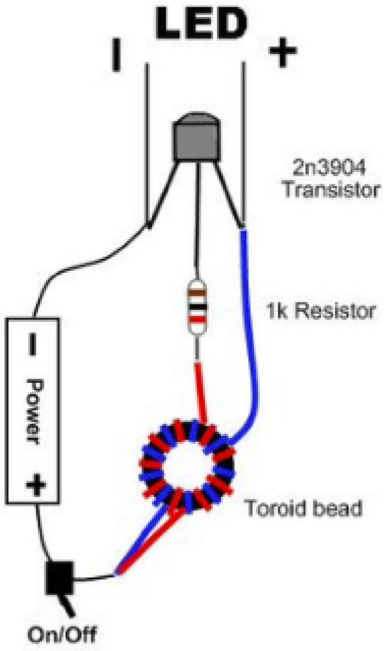
LED
I
+
2n3904
Transistor
1k Resistor
–
Power
+
Toroid bead
On/Off

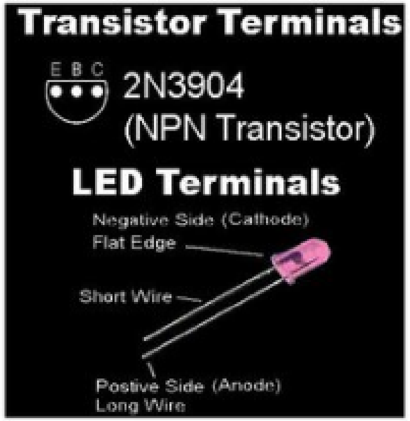
Transistor Terminals
E B C
2N3904
(NPN Transistor)
LED Terminals
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

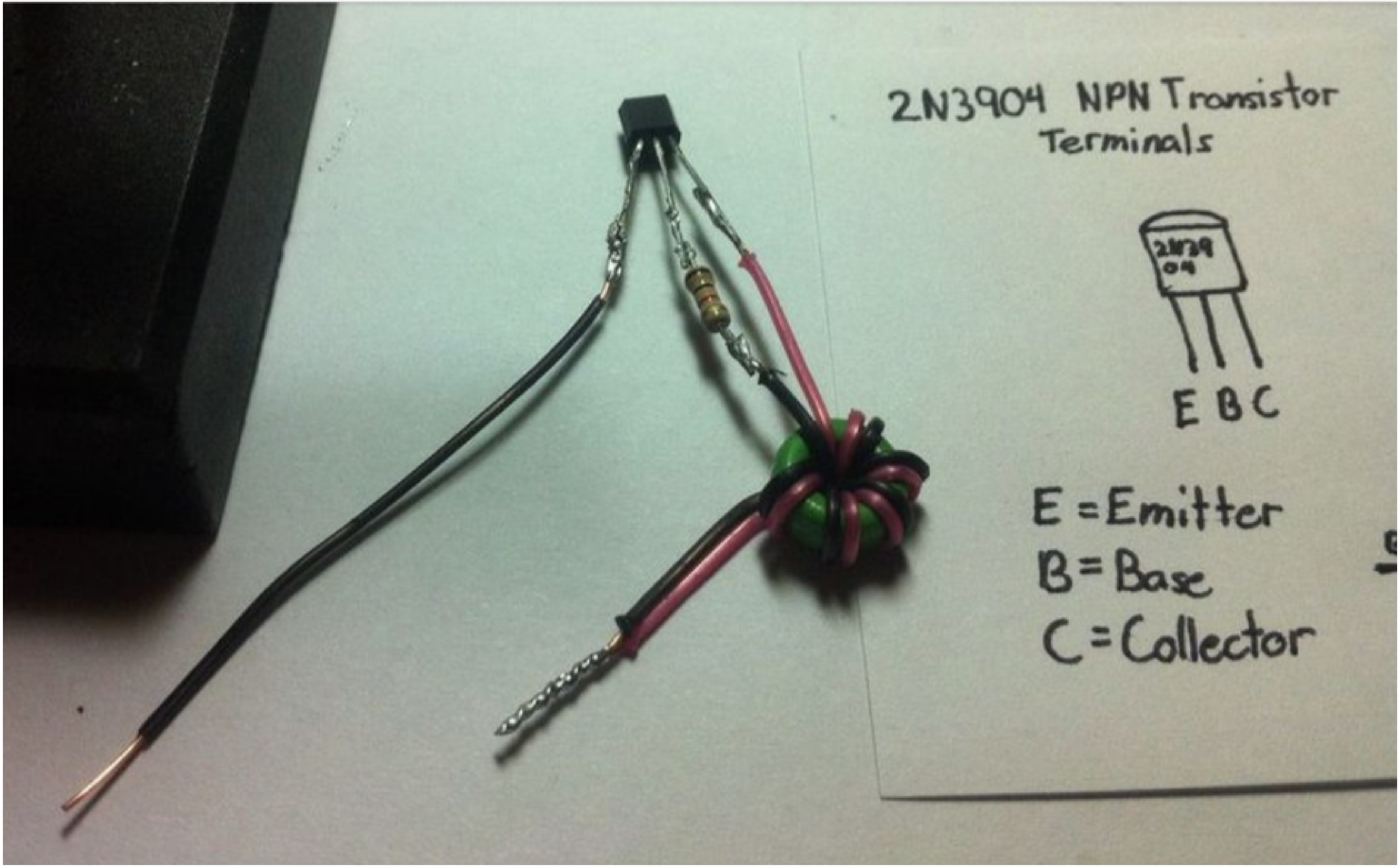
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector

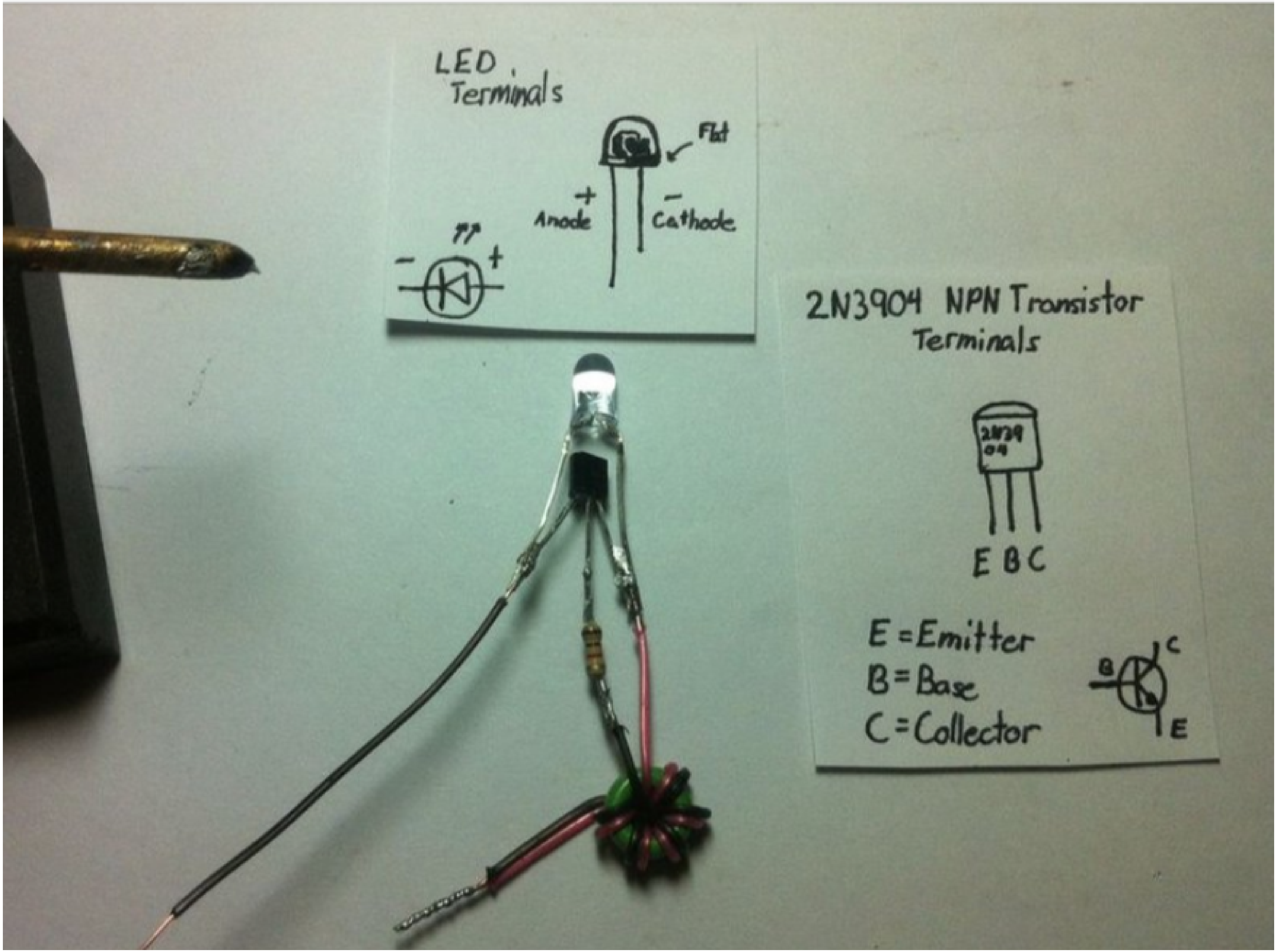
LED Terminals
Flat
+ Anode
− Cathode
−
+
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

These tiny T231212T toroids from Surplussales.com are high permeability so it doesn't take much wire to make a good Joule Thief coil. The core is less than a quarter inch diameter (here is a data sheet in .PDF.). Six inches of 30 AWG magnet wire trifilar wound, with two of the windings connected in parallel for the primary winding.

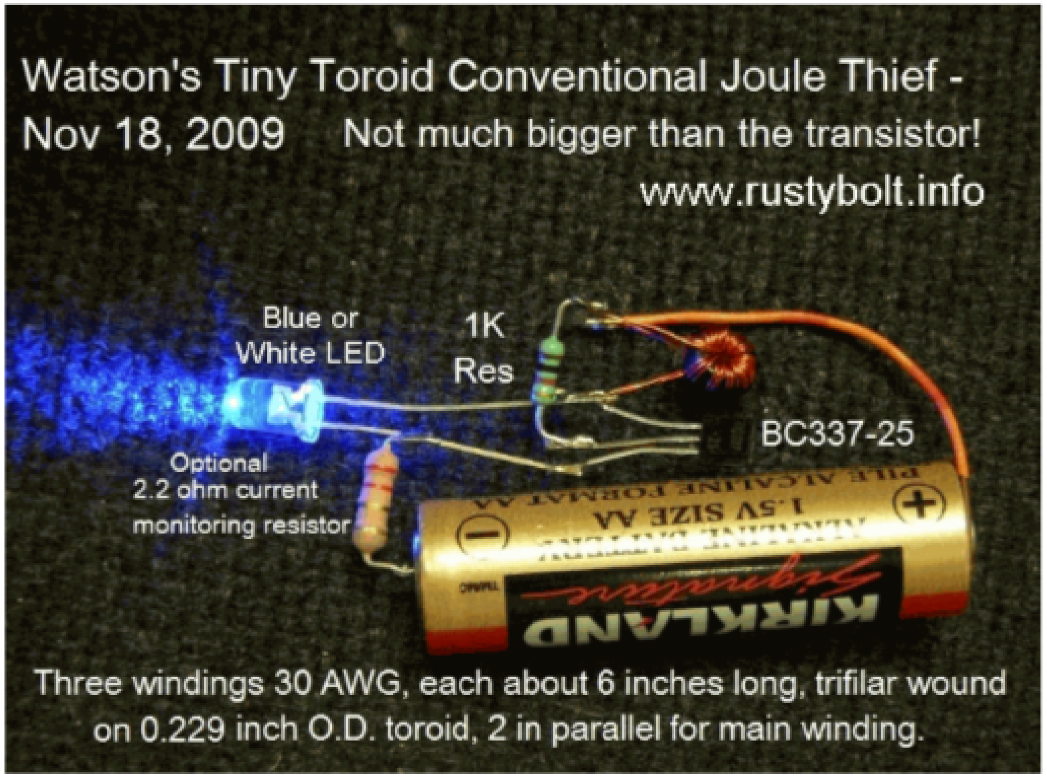
Watson's Tiny Toroid Conventional Joule Thief -
Nov 18, 2009
Not much bigger than the transistor!
www.rustybolt.info
Blue or White LED
1K Res
BC337-25
Optional 2.2 ohm current monitoring resistor
Three windings 30 AWG, each about 6 inches long, trifilar wound on 0.229 inch O.D. toroid, 2 in parallel for main winding.

The transistor is a BC337-25. I put a 2.2 ohm resistor in series with the circuit to measure the battery current. It can be put there temporarily and removed when the LED current is what you want. How do you change the LED current? Change the 1k resistor to a different value. Higher values will reduce the LED current. With the BC337-25 and 1k resistor, the LED current should be close to 20 milliamps so it's probably not wise to go much below 1k. If you use another transistor, especially the pipsqueak 2N3904, you may have a hard time getting it to put out 20 milliamps.

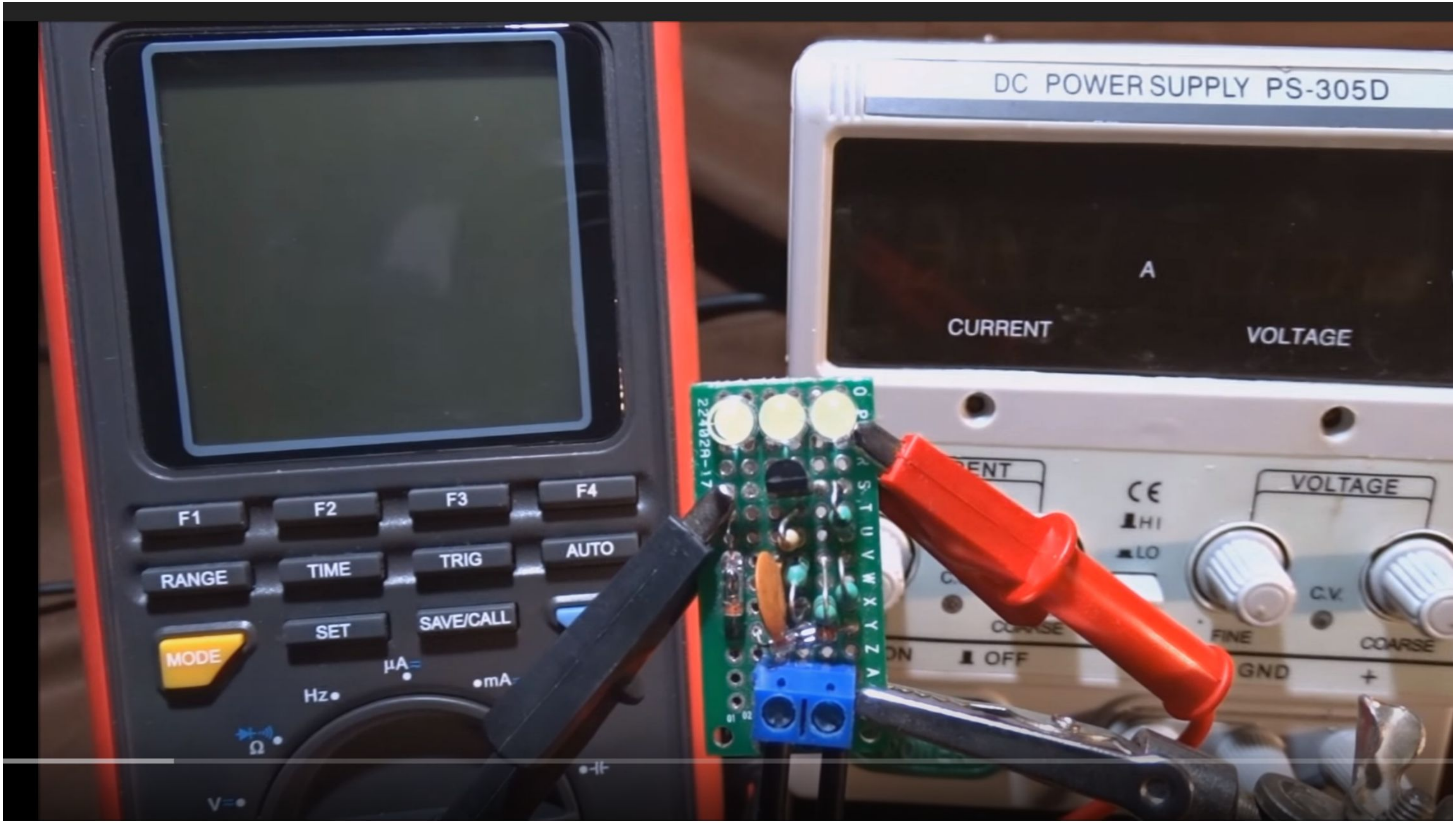
DC POWER SUPPLY PS-305D
A
CURRENT
VOLTAGE
F1
F2
F3
F4
RANGE
TIME
TRIG
AUTO
MODE
SET
SAVE/CALL
μA
mA
Hz
Ω
V
22402A-17
VOLTAGE
C.V.
FINE
COARSE
HI
LO
OFF
GND
+

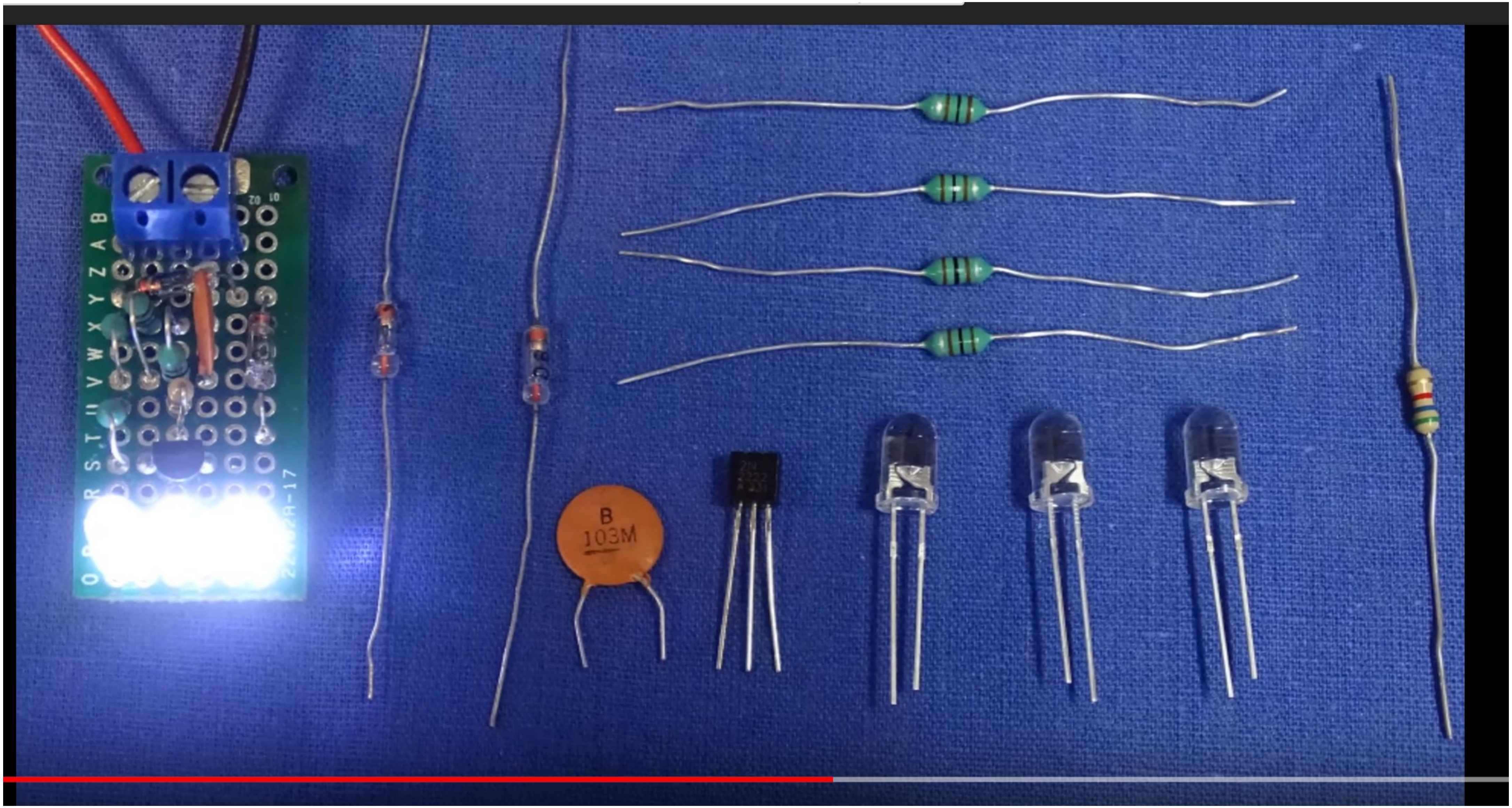
B
103M

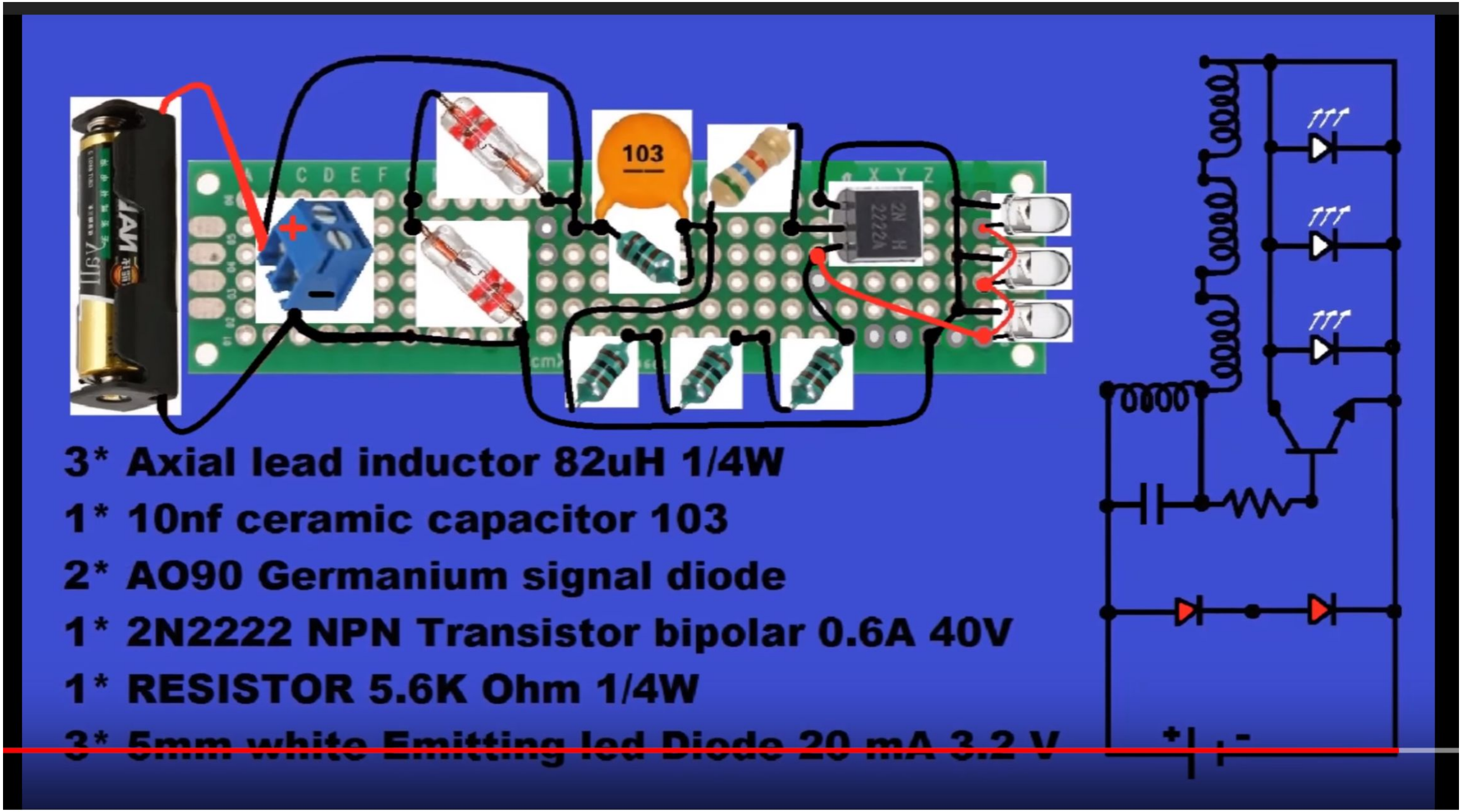
103
2N 2222A
A C D E F
X Y Z
+
−

3* Axial lead inductor 82uH 1/4W

1* 10nf ceramic capacitor 103

2* AO90 Germanium signal diode

1* 2N2222 NPN Transistor bipolar 0.6A 40V

1* RESISTOR 5.6K Ohm 1/4W

3* 5mm white Emitting led Diode 20 mA 3.2 V

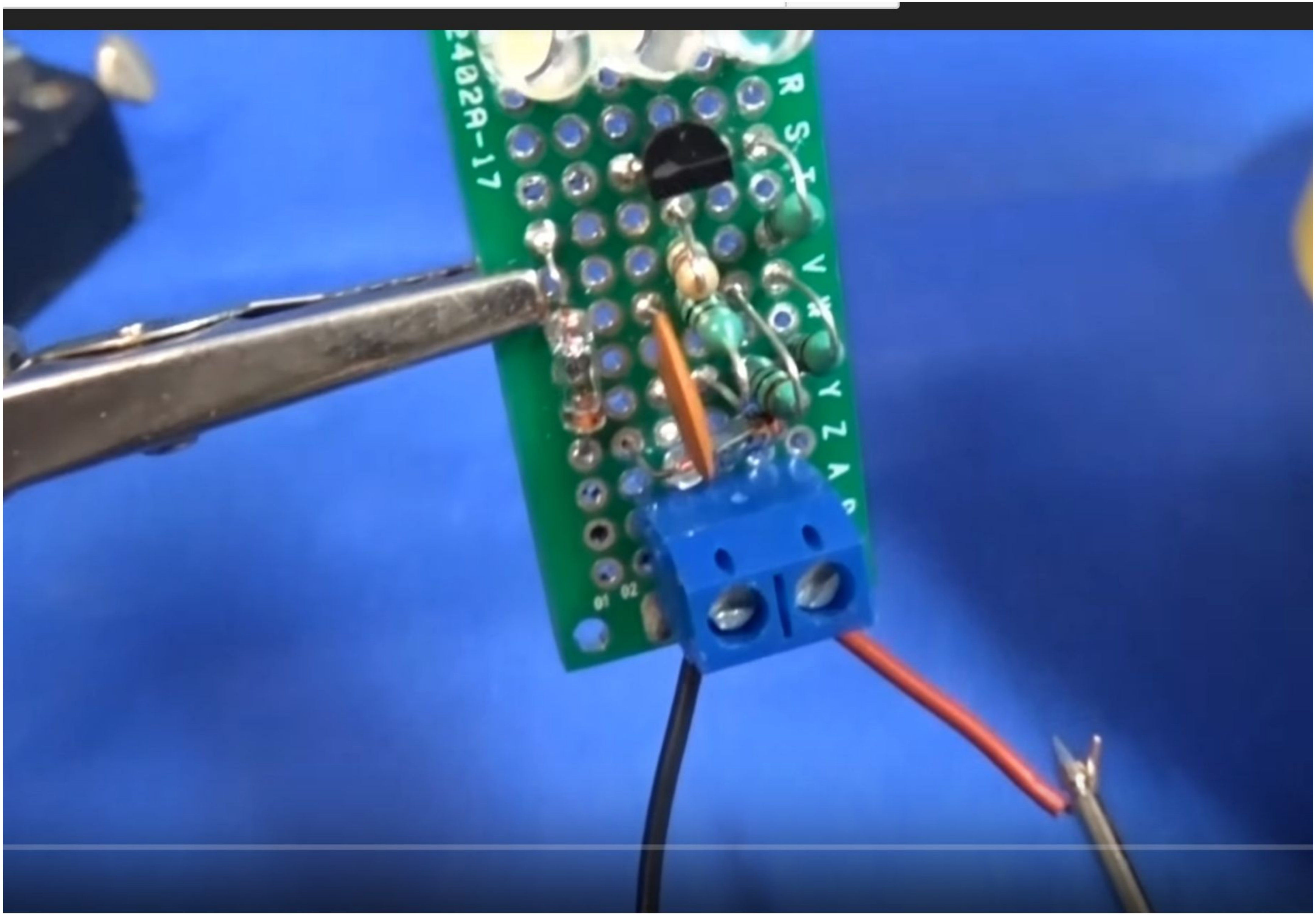
2402A-17
R
S
V
Y
Z
A
01
02

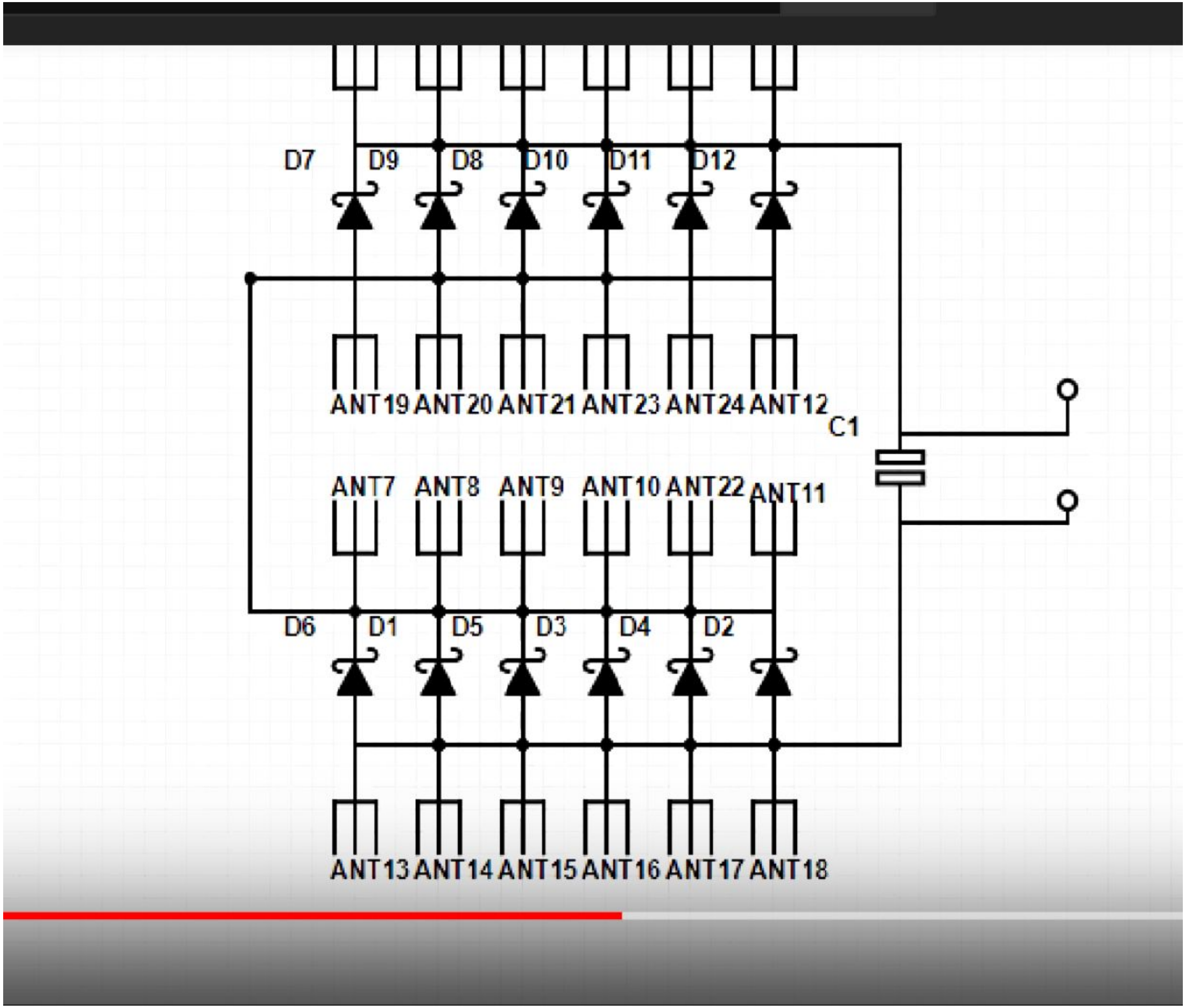
D7
D9
D8
D10
D11
D12
ANT19
ANT20
ANT21
ANT23
ANT24
ANT12
C1
ANT7
ANT8
ANT9
ANT10
ANT22
ANT11
D6
D1
D5
D3
D4
D2
ANT13
ANT14
ANT15
ANT16
ANT17
ANT18

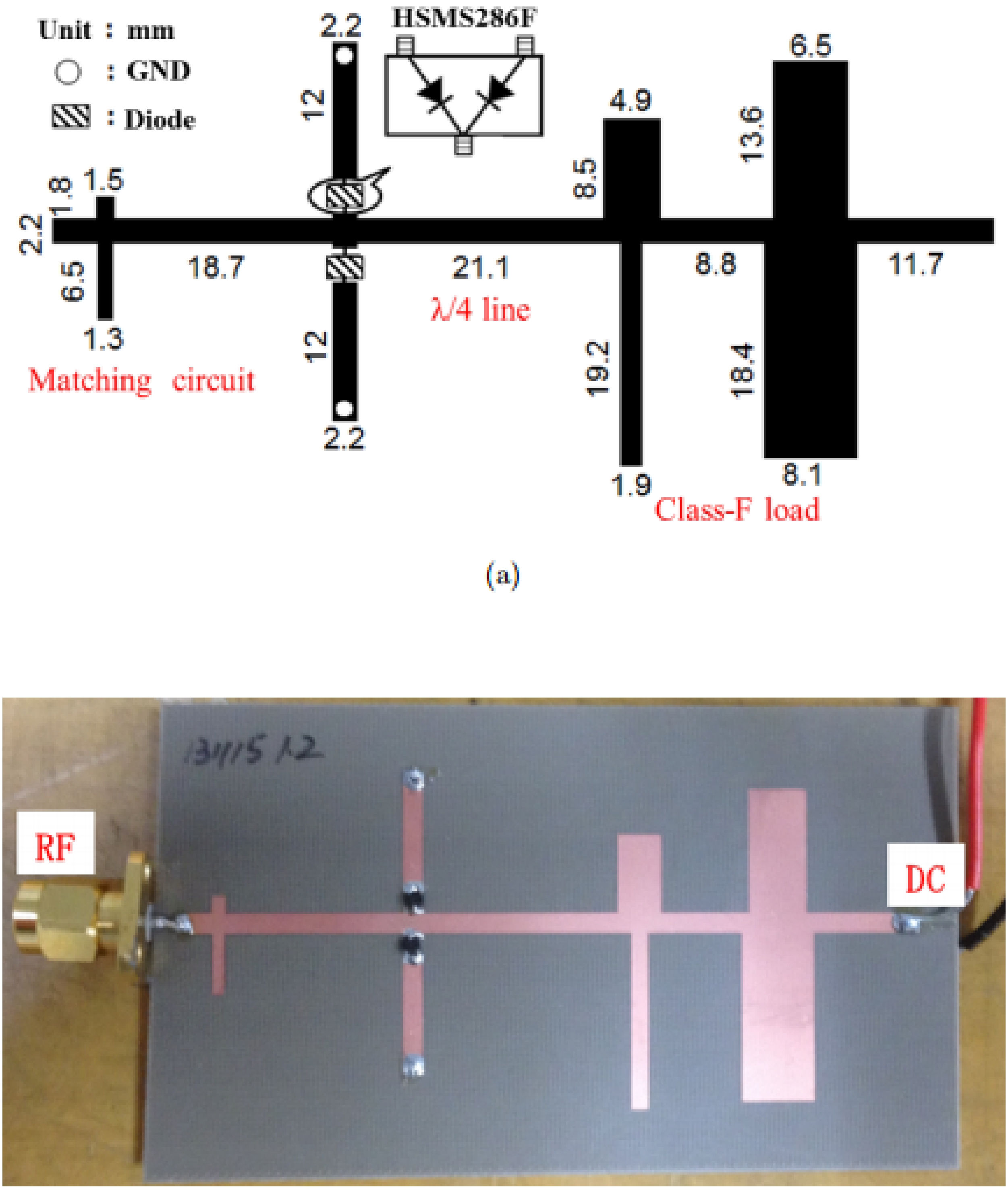
Unit : mm
○ : GND
▨ : Diode
2.2
HSMS286F
12
6.5
4.9
13.6
8.5
1.8
1.5
2.2
6.5
18.7
21.1
8.8
11.7
λ/4 line
1.3
12
Matching circuit
19.2
18.4
2.2
1.9
8.1
Class-F load
131115 12
RF
DC

(a)

(b)

Figure 4.1: Proposed positive output voltage rectifier. (a) Structure and size. (b) Photograph.

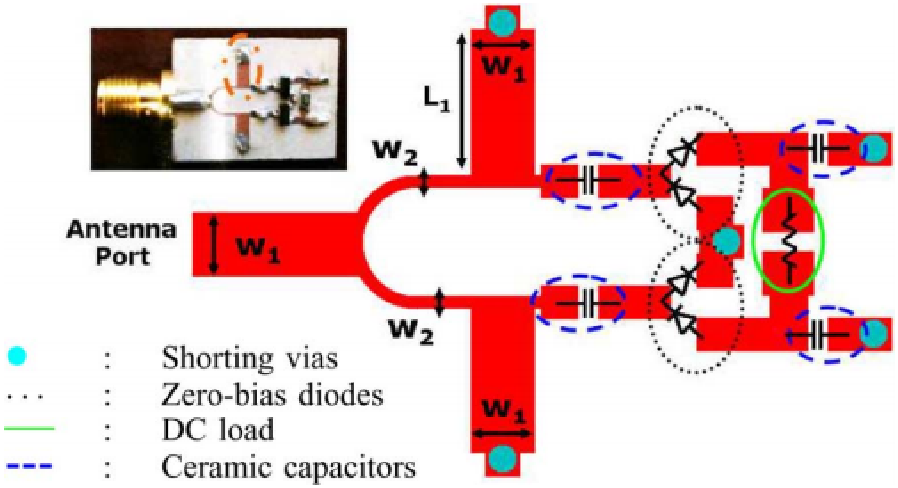
W1
L1
W2
Antenna Port
W1
W2
W1
● : Shorting vias
· · · : Zero-bias diodes
— : DC load
- - - : Ceramic capacitors

Fig. 3. Layout of the rectifier prototype, printed on RO3206. $w_1 = 72$ mil, $w_2 = 15$ mil, and $L_1 = 171$ mil. Fabricated sample is shown in top left, and the impedance matching stub is encircled.

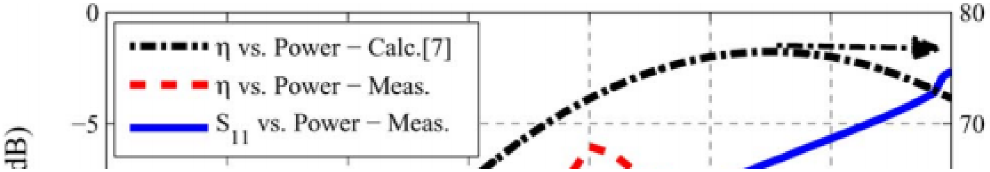

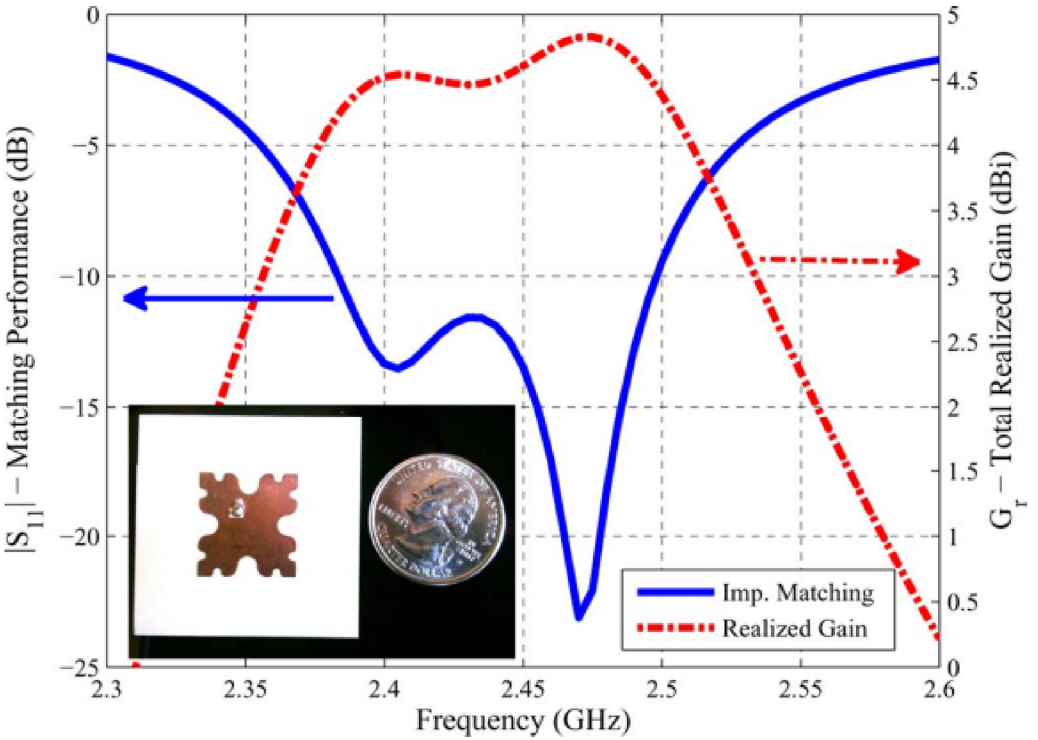

Fig. 5. Measured $|S_{11}|$ performance and total realized gain (at boresight) of the proposed antenna. Fabricated sample is shown in bottom left.

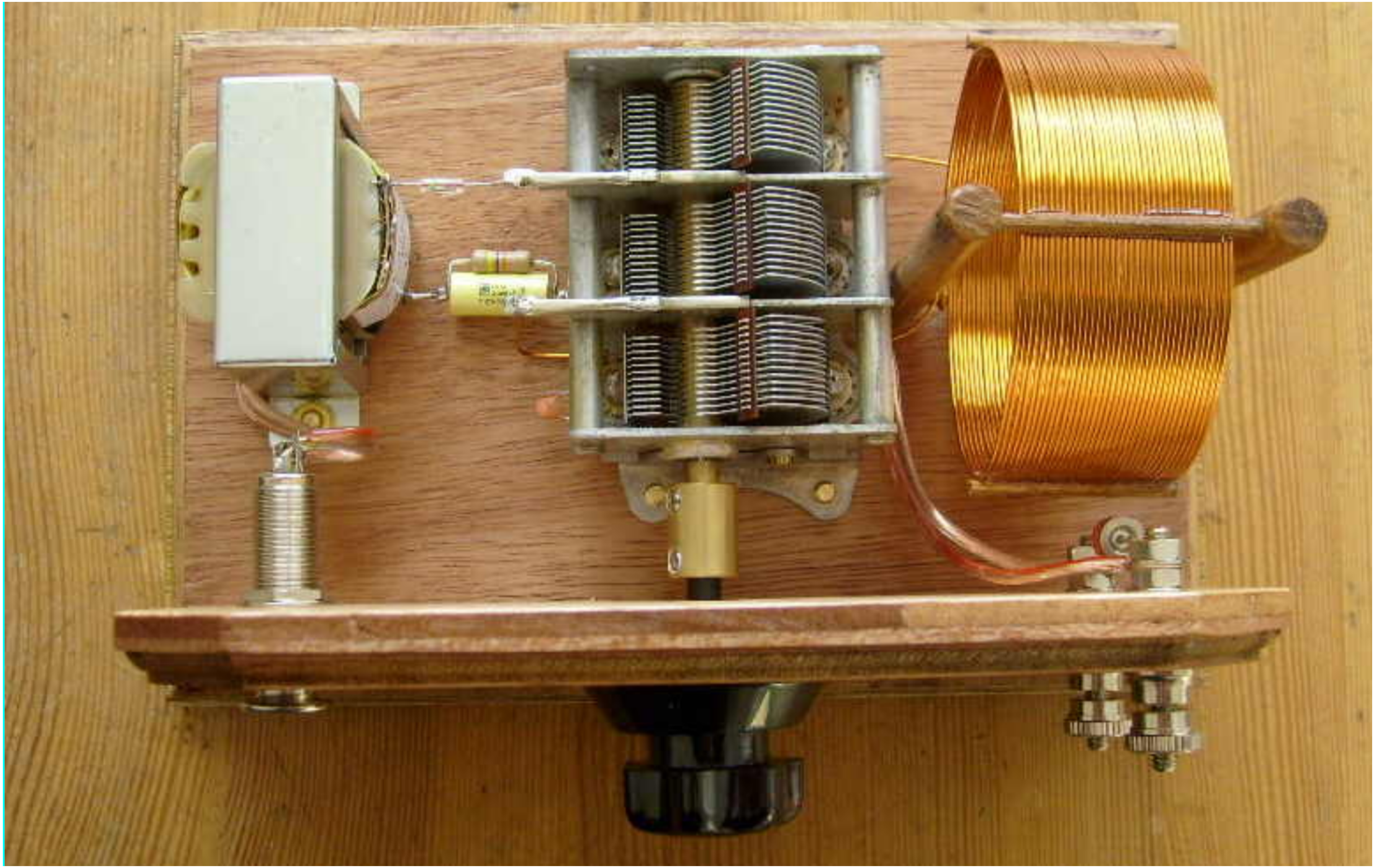

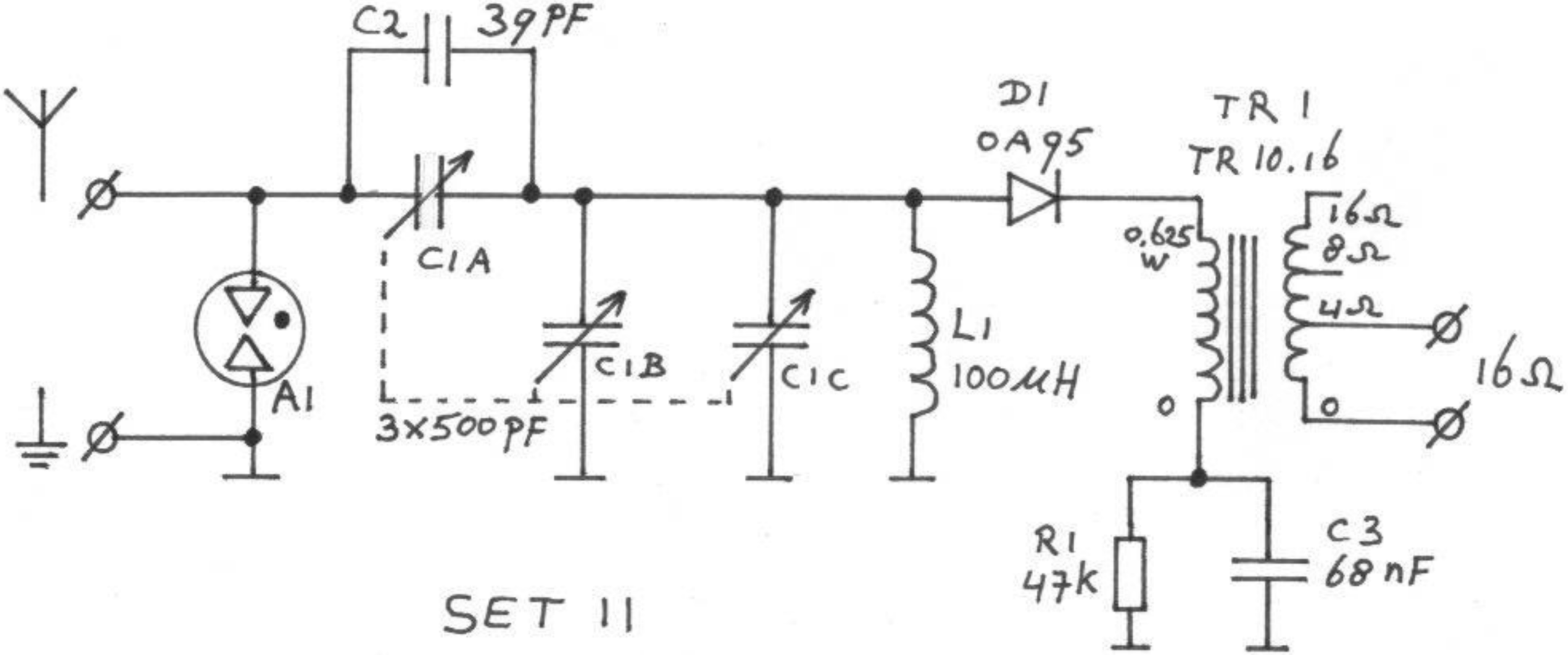
C2
39PF
C1A
C1B
C1C
3×500PF
A1
L1
100μH
D1
0A95
TR1
TR10.16
0.625 W
0
16Ω
8Ω
4Ω
0
16Ω
R1
47k
C3
68nF

SET 11

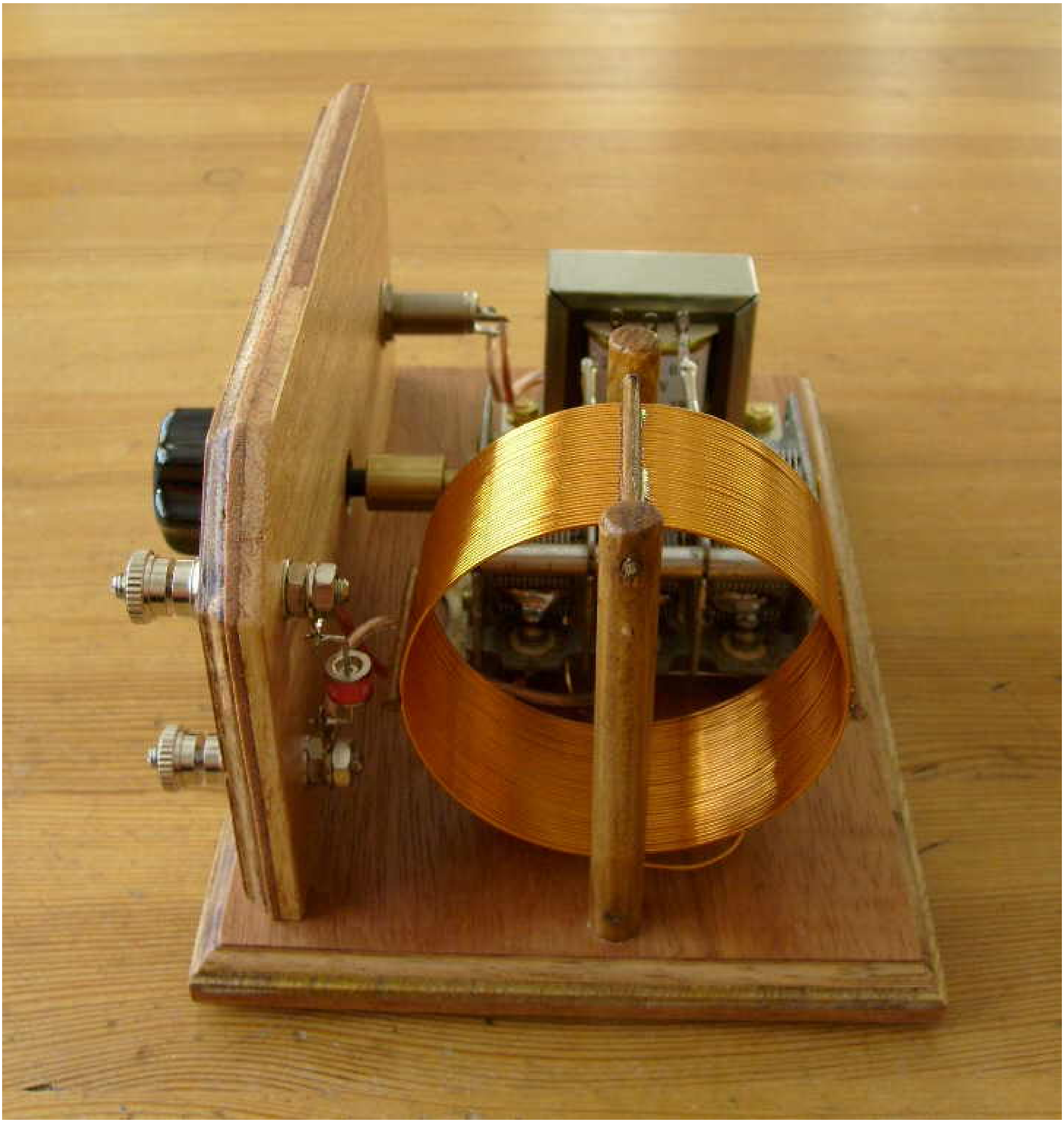

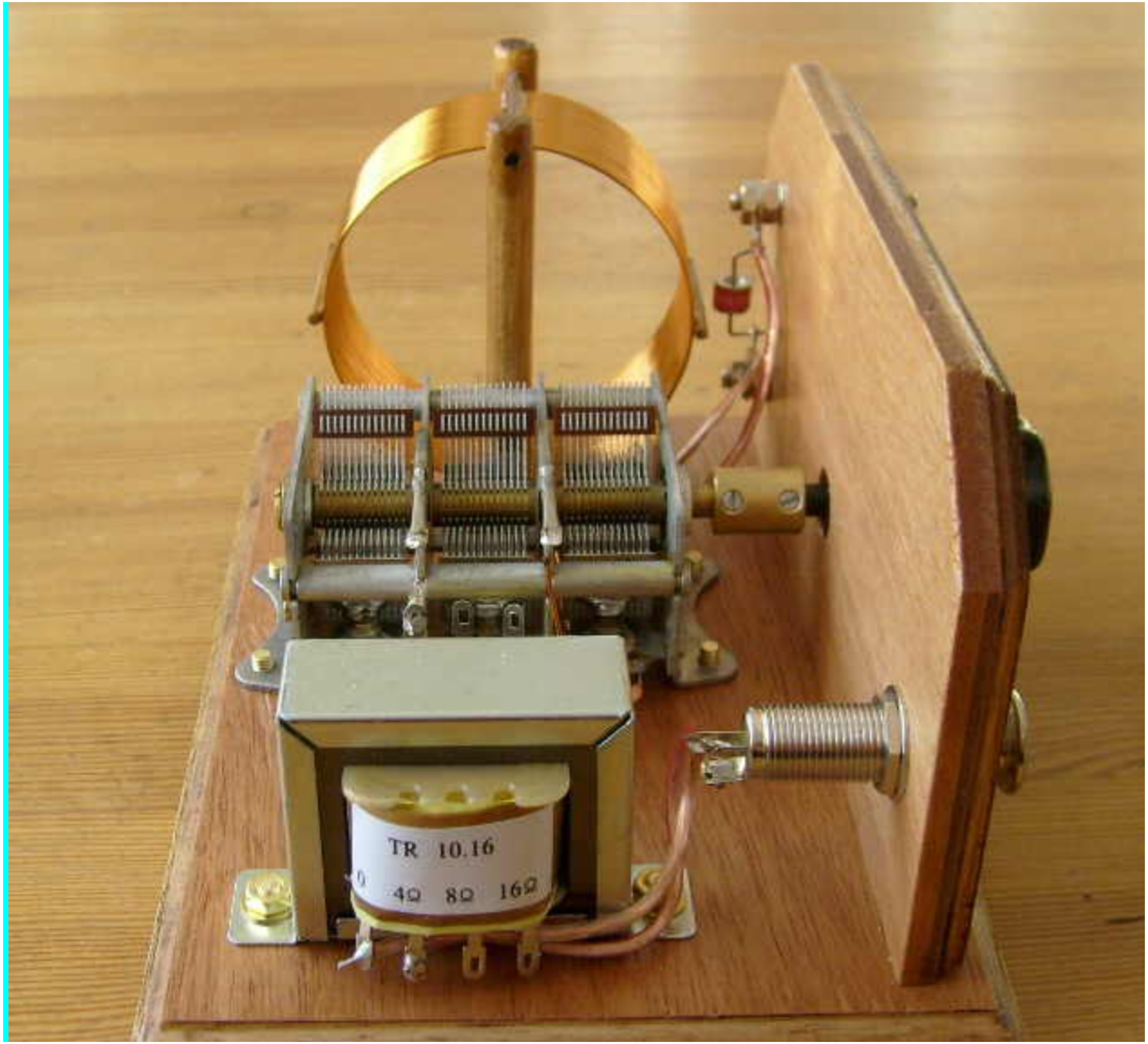
TR 10.16
0 4Ω 8Ω 16Ω

Preparation of the tuning capacitor

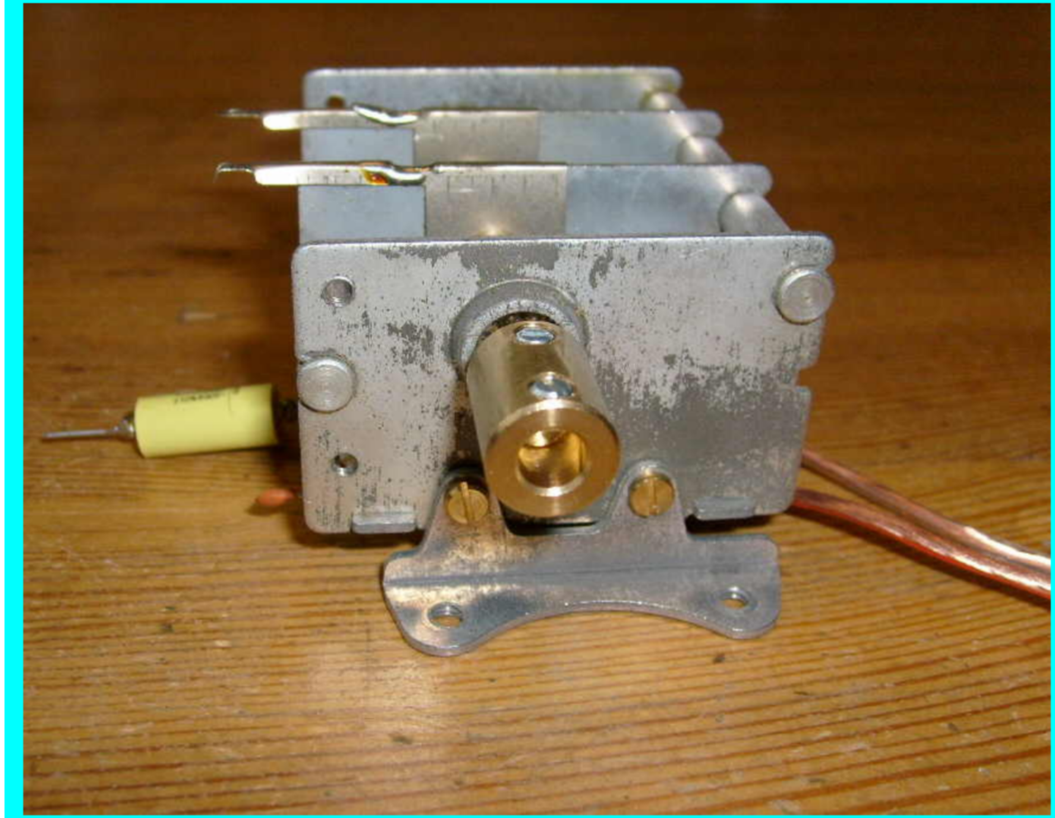

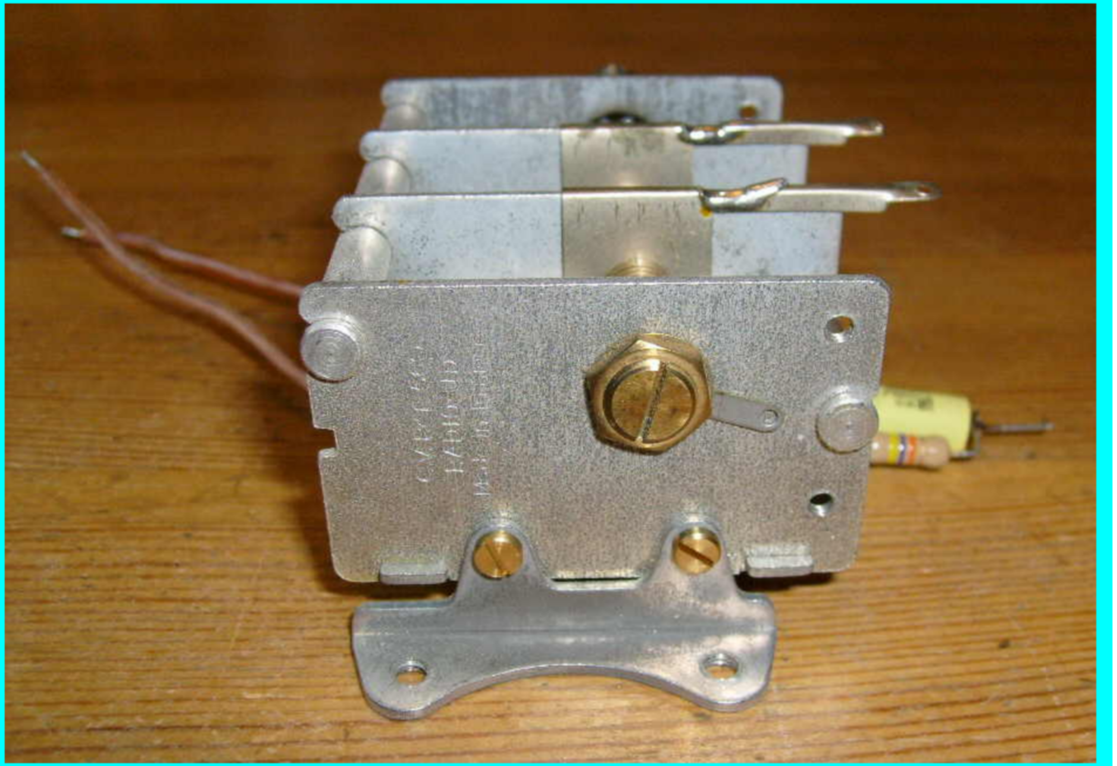

Mount the mounting supports on the tuning capacitor (they are delivered with the tuning capacitor).
Place the shaft coupler on the shaft of the tuning capacitor.

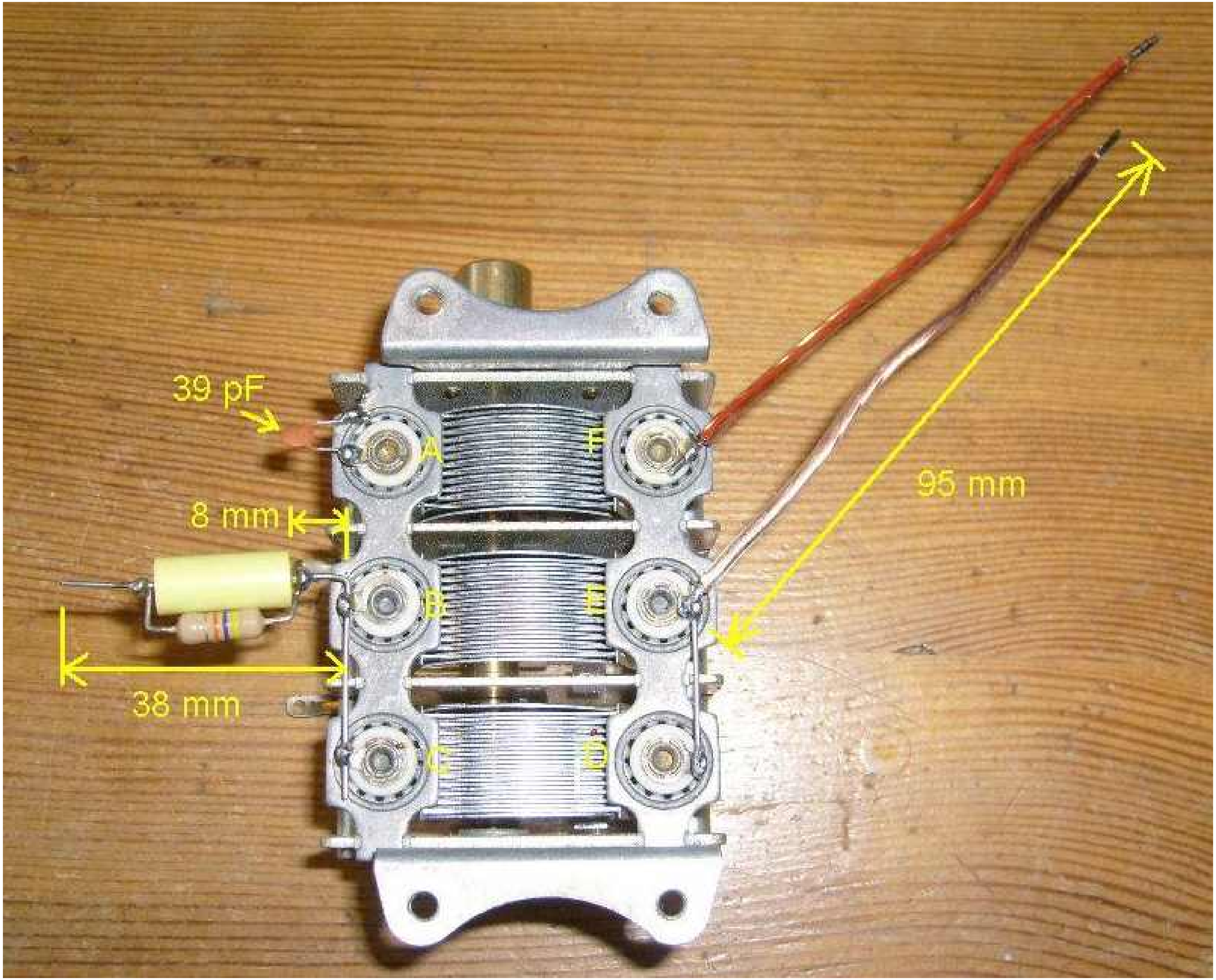
39 pF
A
F
95 mm
8 mm
B
E
38 mm
C
D

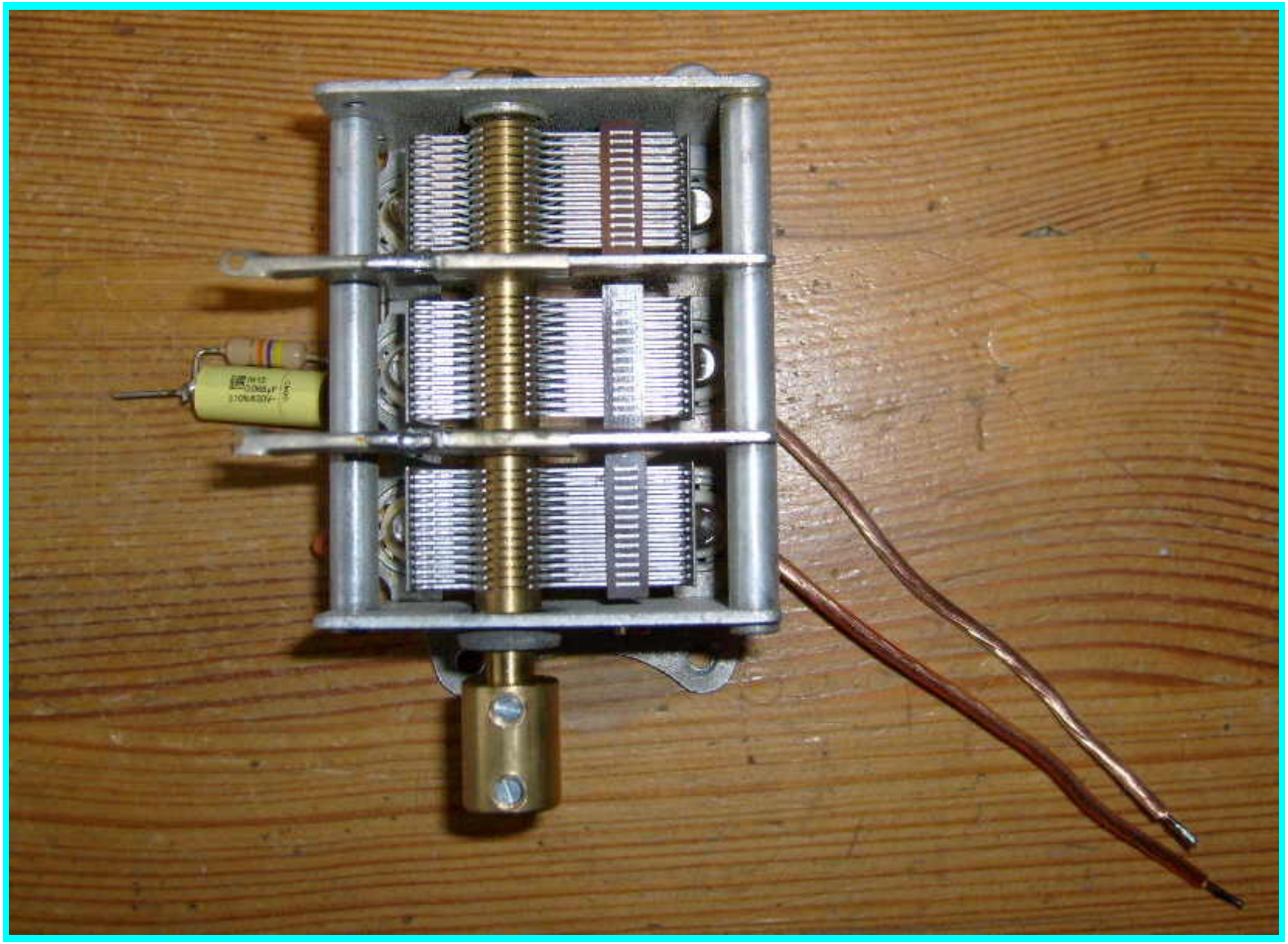
0,068µF
±10%/630V~

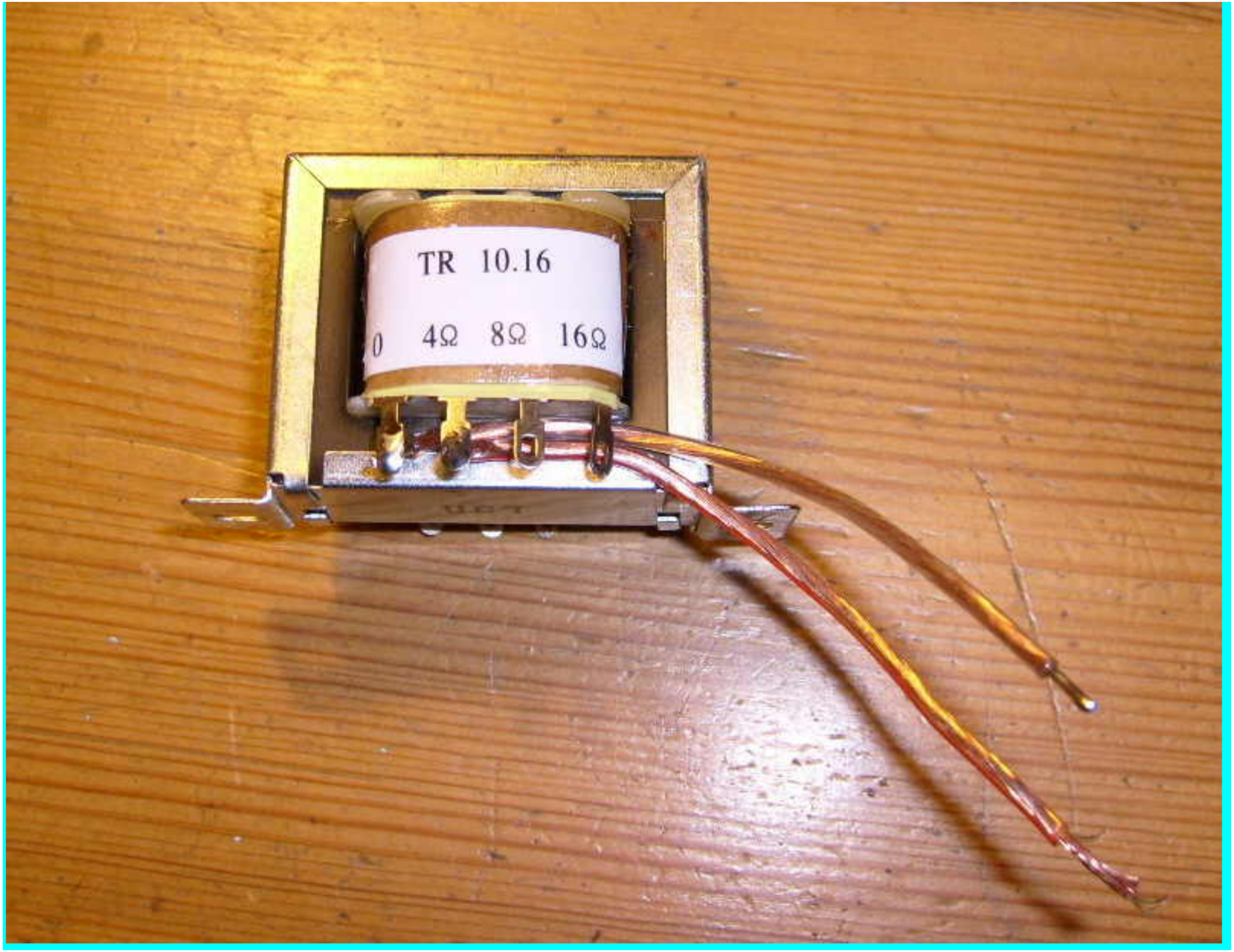
TR 10.16
0
4Ω
8Ω
16Ω

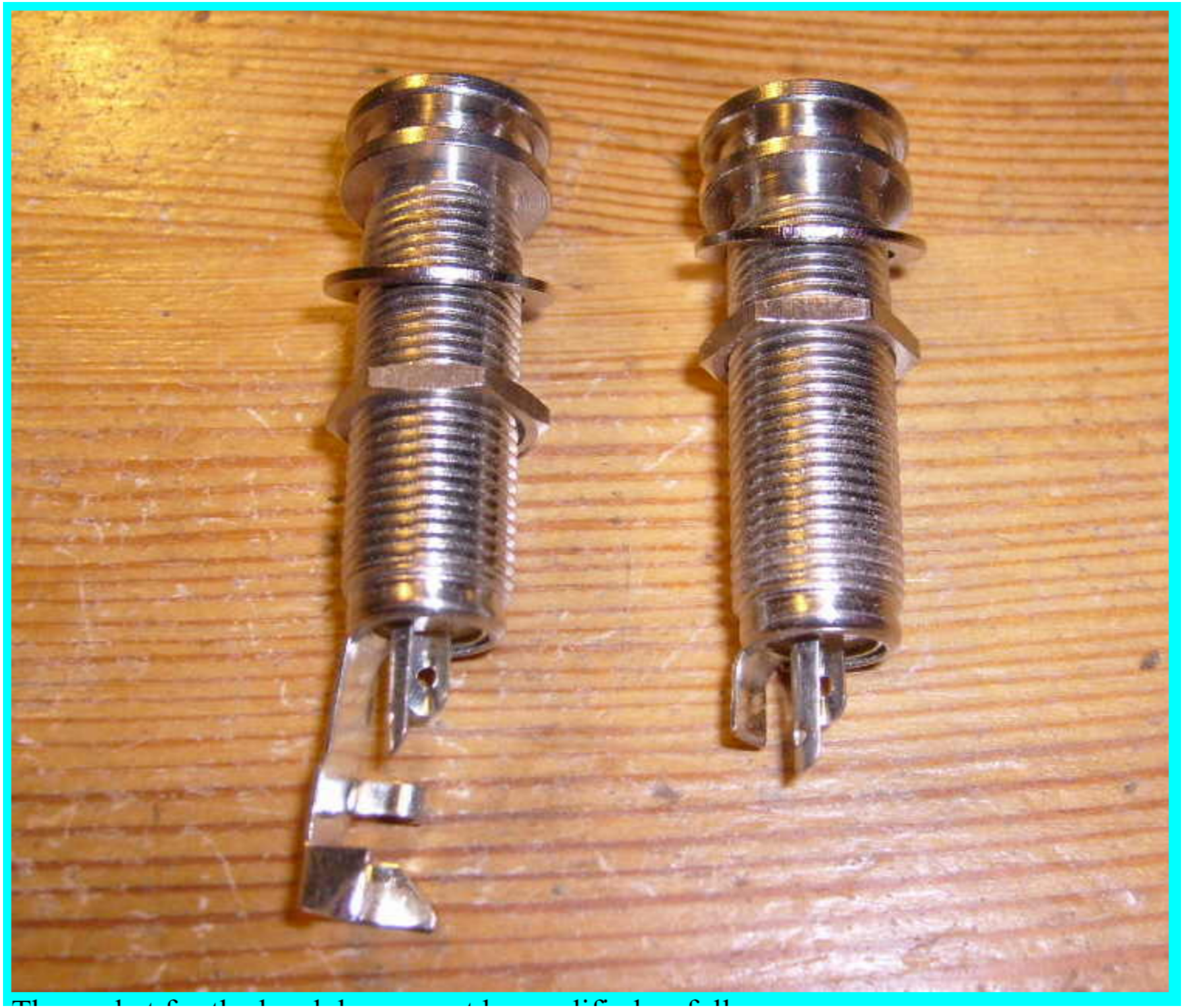

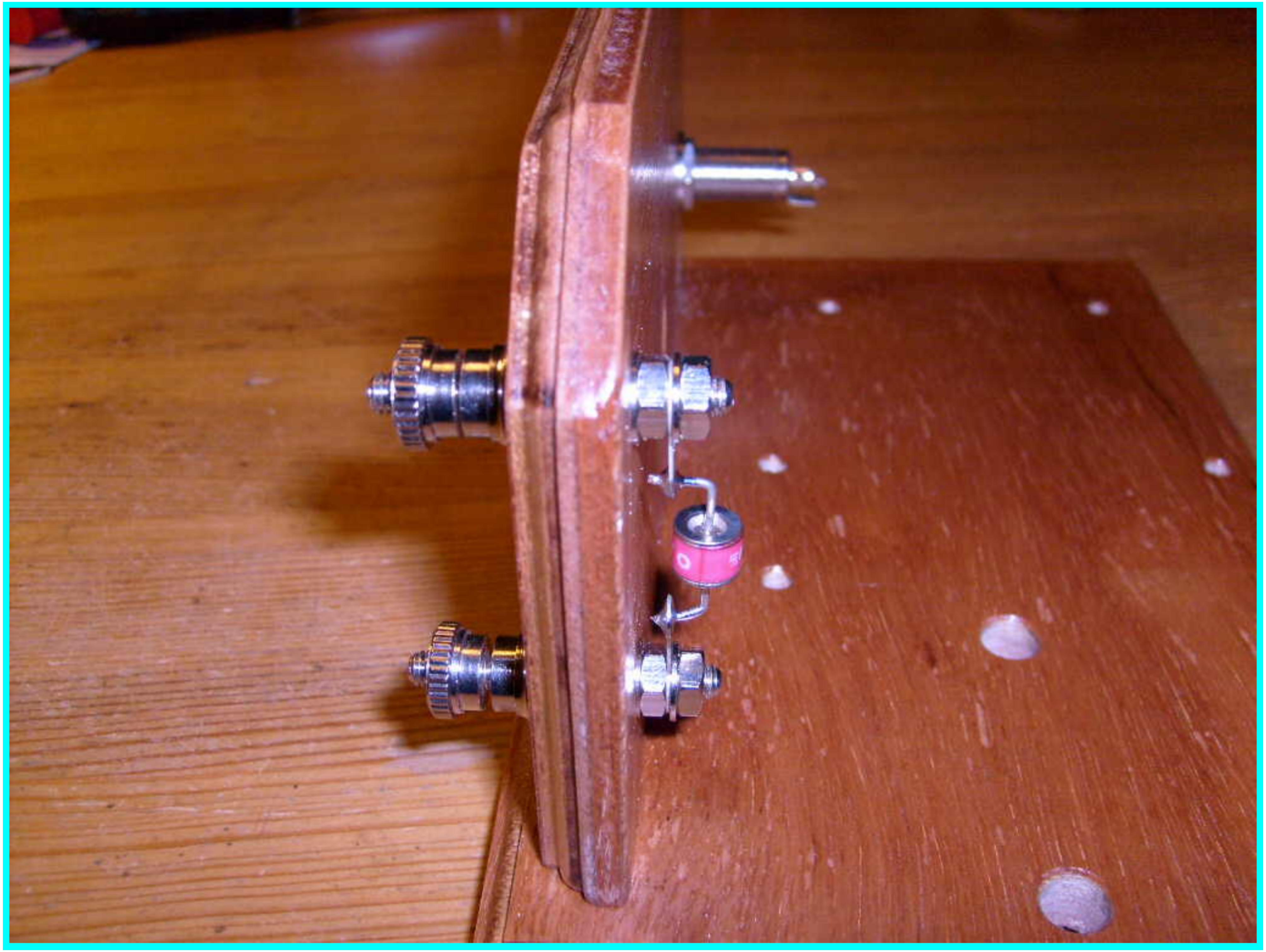

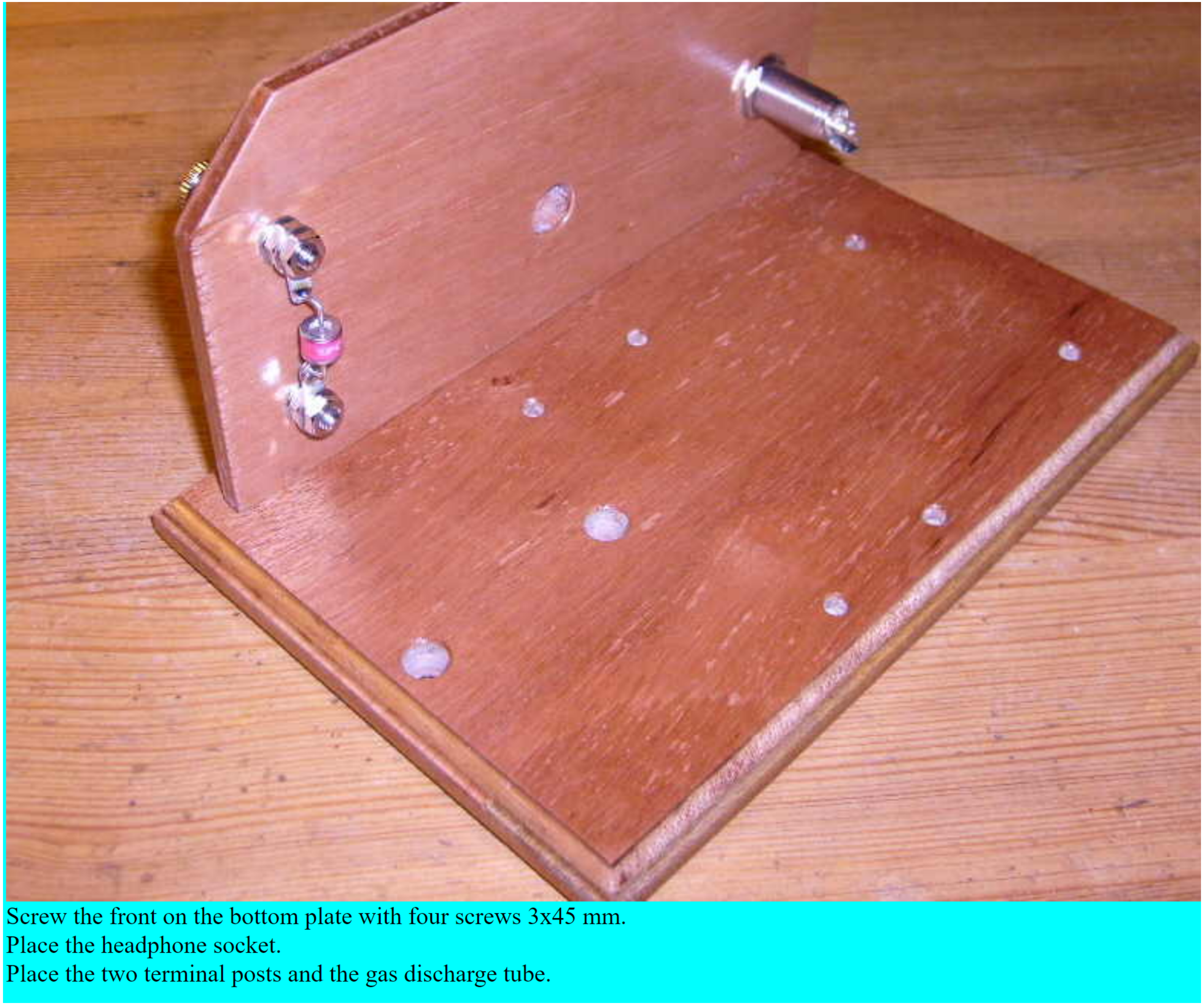

Screw the front on the bottom plate with four screws 3x45 mm.
Place the headphone socket.
Place the two terminal posts and the gas discharge tube.

Make a paper strip with 4 sections of 40x55 mm on it.
Do the strip around the tube, and fix it with a piece of tape.
Don't tape the strip onto the tube on places where later the coil is wound.

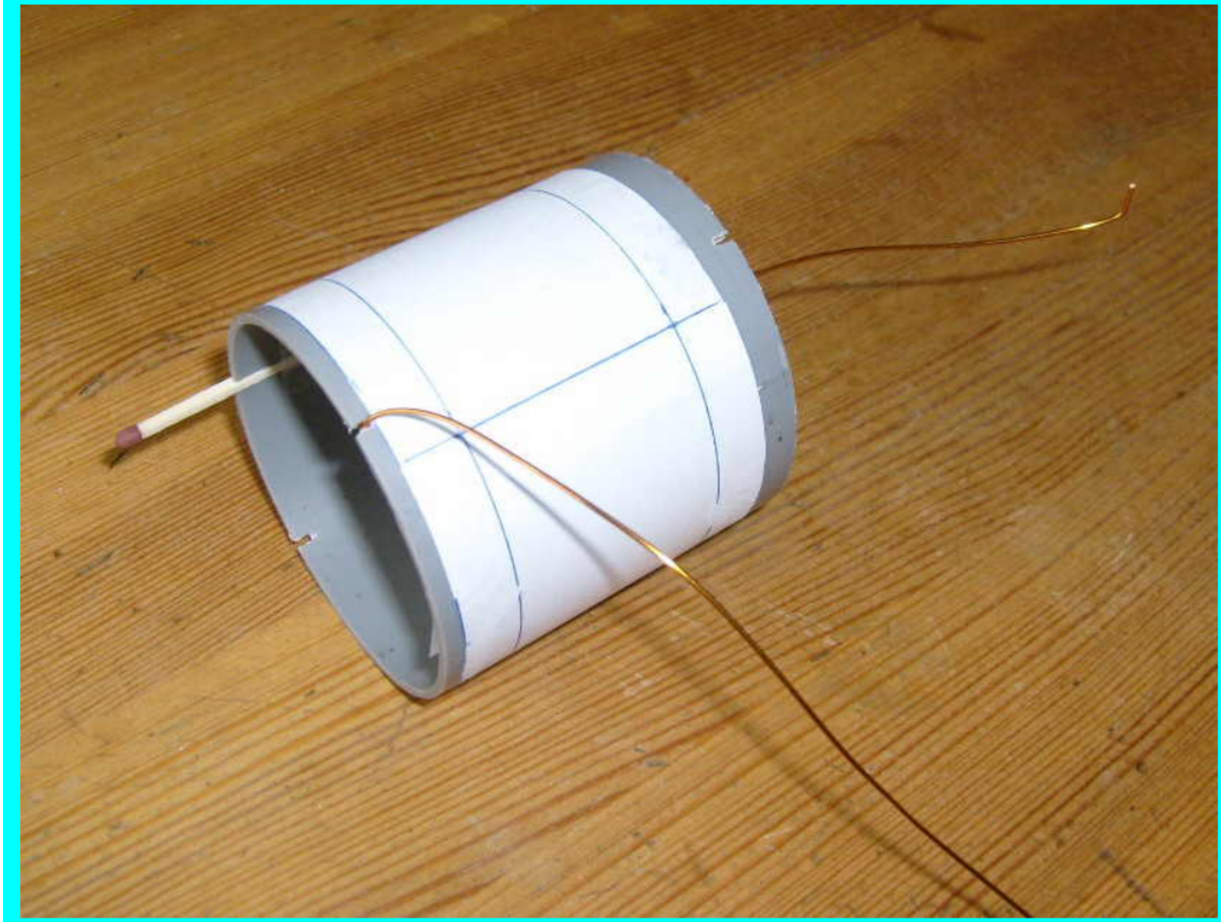

For the coil we need 8.6 meter enamelled copper wire with 0.8 mm thickness.
Hook the begin of the coilwire onto the slot in the tube.
Also fix the begin of the wire on the inside of the tube with some tape.

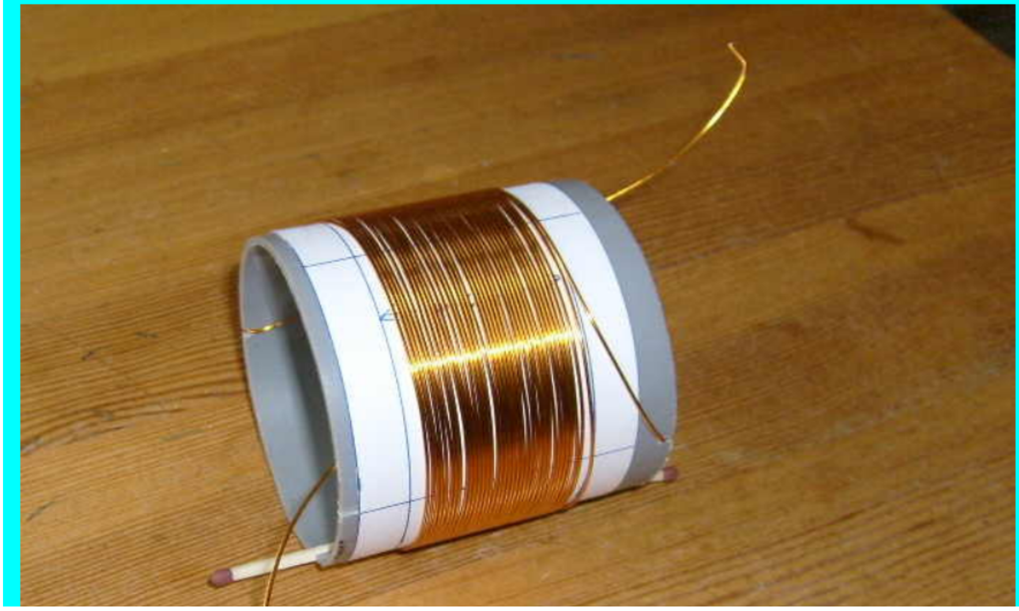

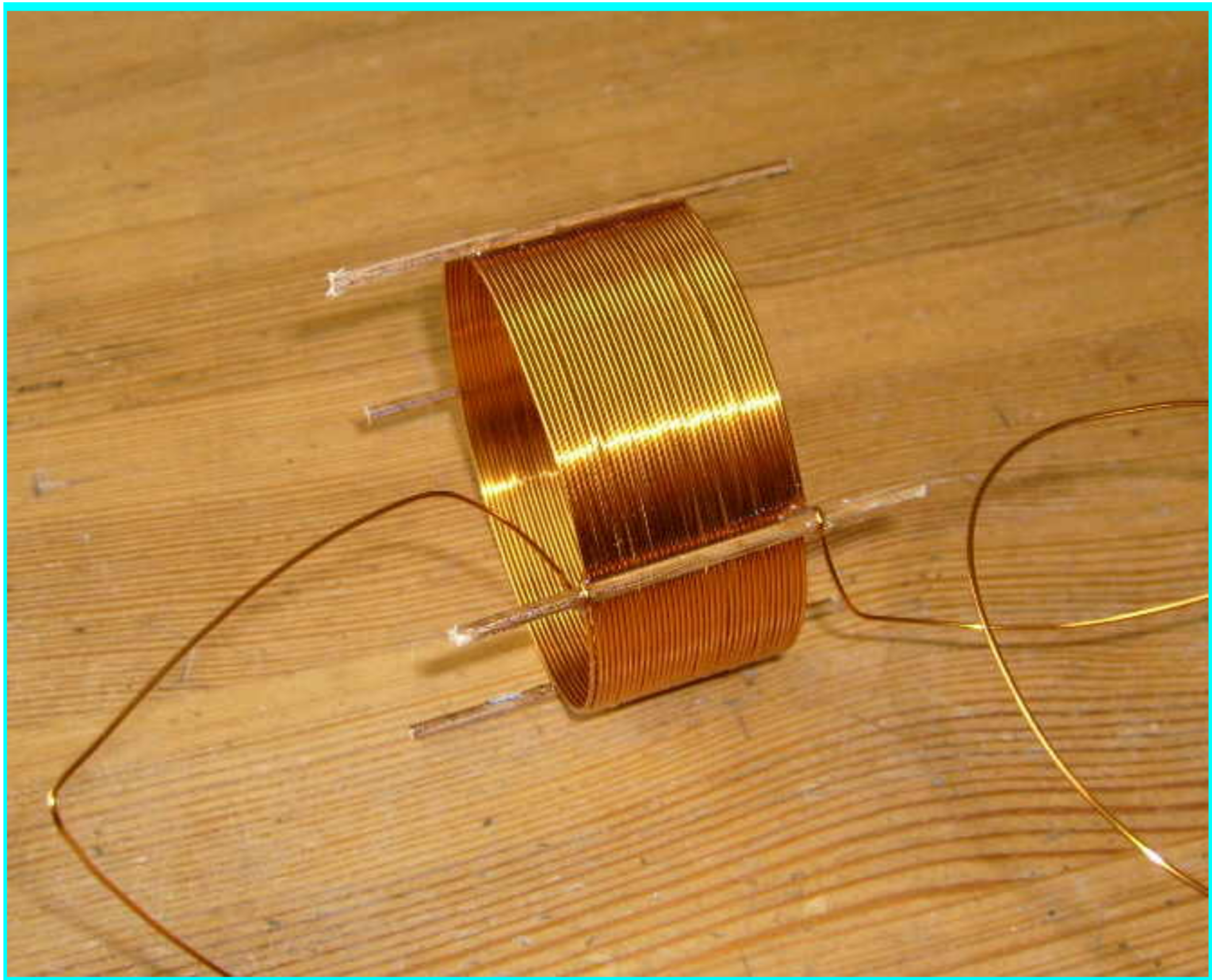

Wait until the glue is hardened.
Remove the rubber bands, the tape and the matches.
Shift the tube out of the coil, and remove the paper out of the coil.
Turn the both ends of the wire one time around the stick at the bottom side of the coil, and apply some extra glue here.

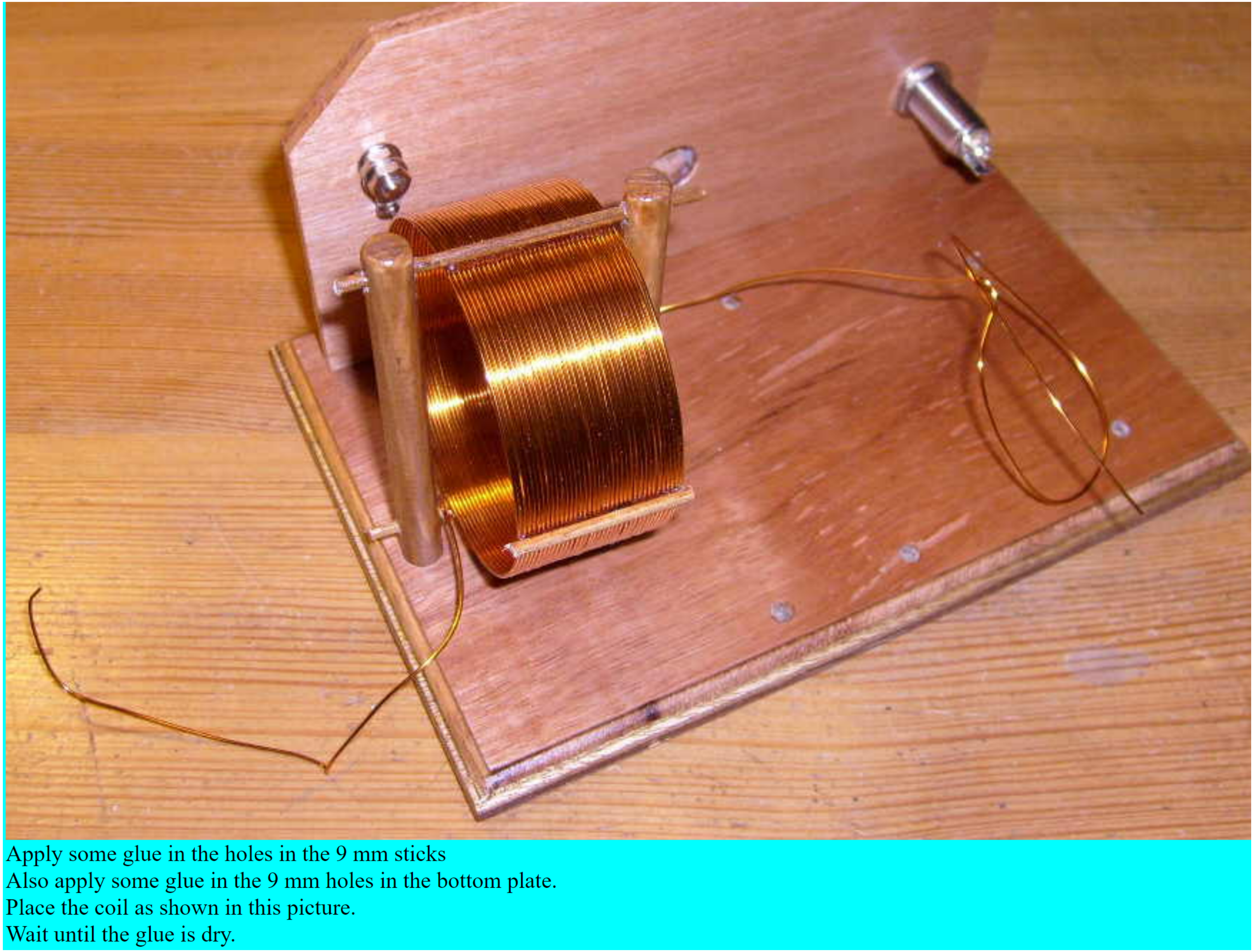

Apply some glue in the holes in the 9 mm sticks
Also apply some glue in the 9 mm holes in the bottom plate.
Place the coil as shown in this picture.
Wait until the glue is dry.

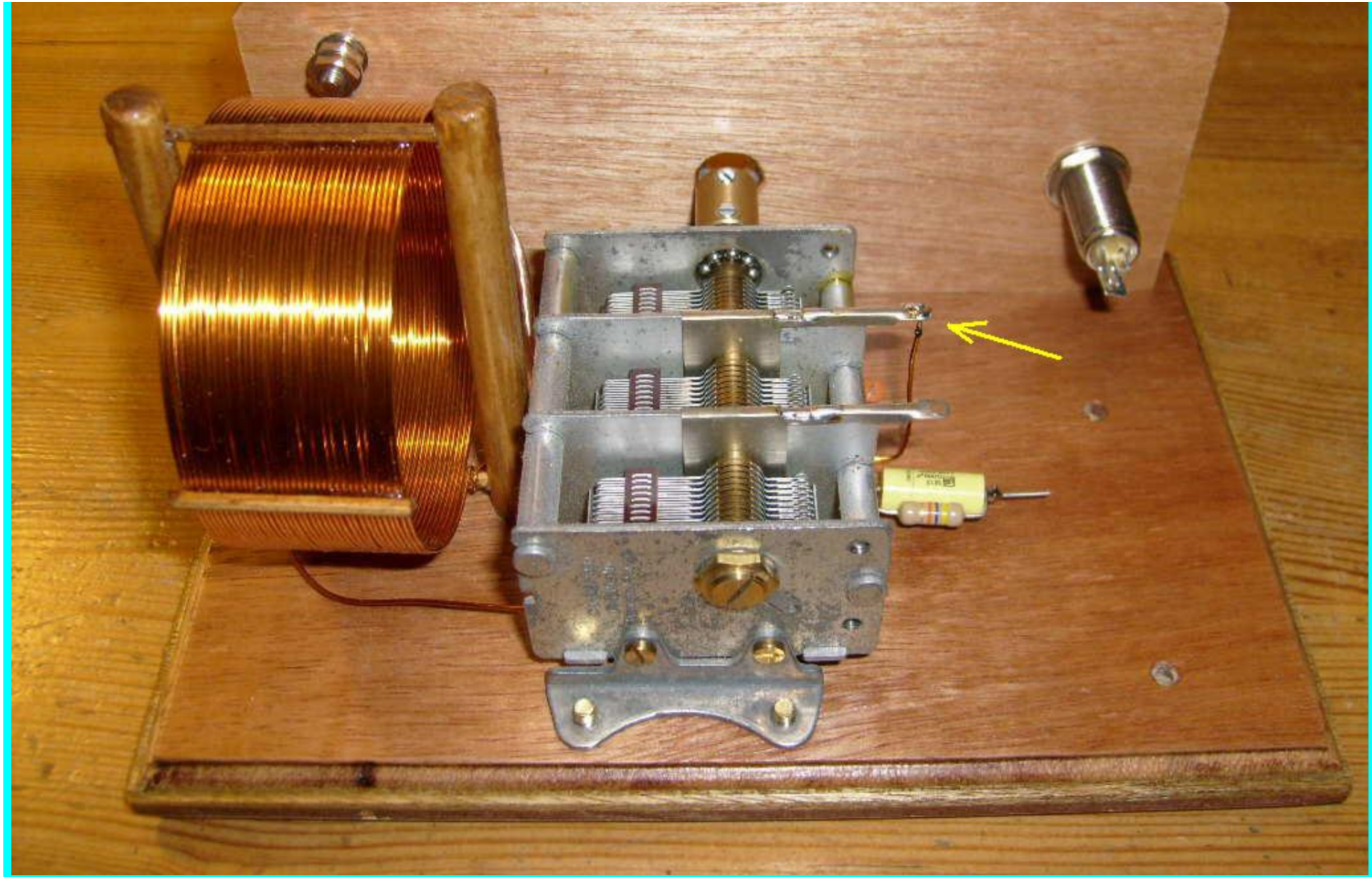

Place the tuning capacitor on the frame.
Fix it with four screws M4x16
One coilwire runs underneath the tuning capacitor, is then bend upwards and cut to length.
Solder this wire to the tuning capacitor (indicated with the arrow in the above picture).
Scrap of the lacquer from the wire, before soldering.

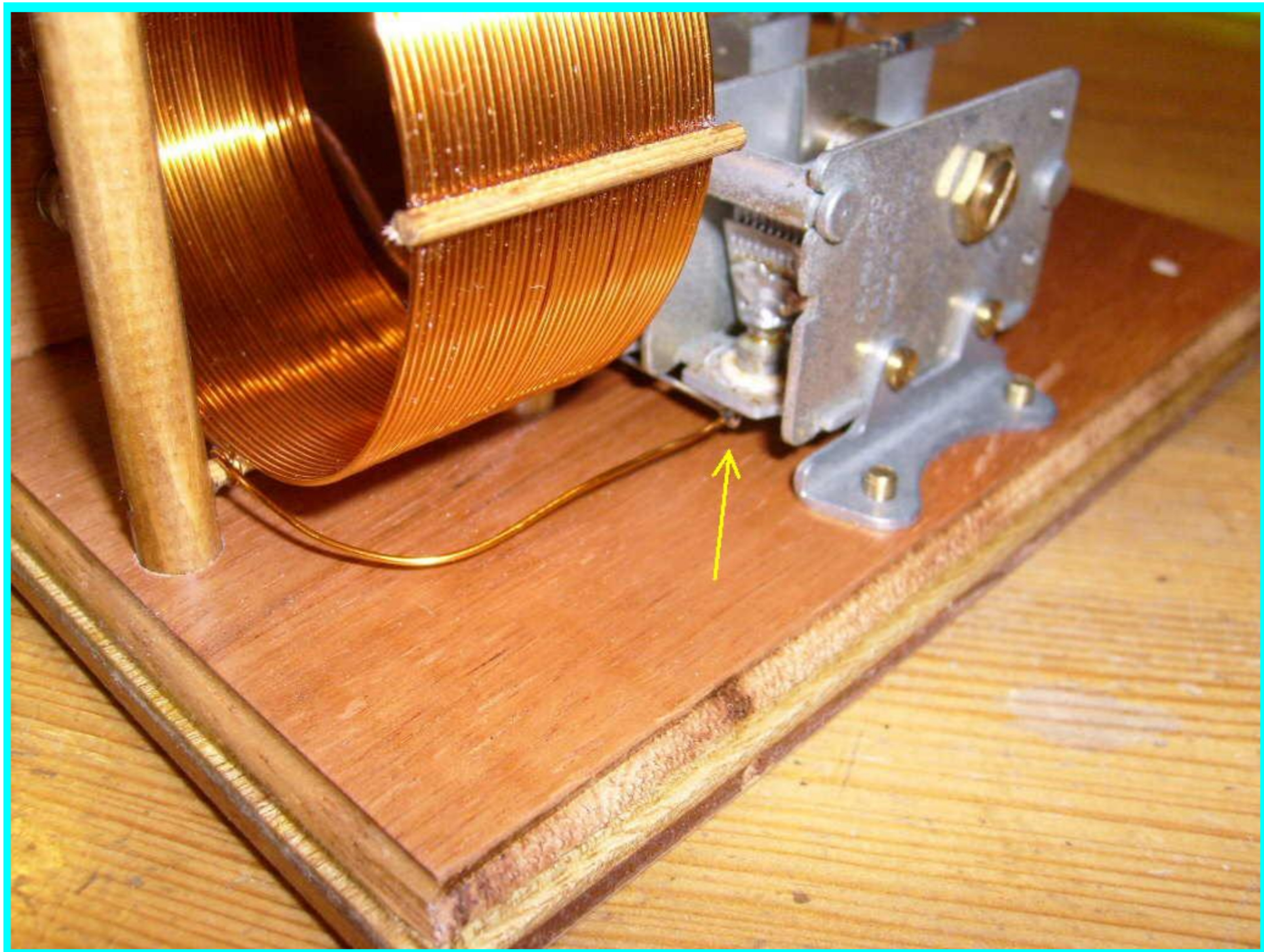

The other wire is cut to length, and soldered to the solder tag of the tuning capacitor (indicated with arrow).

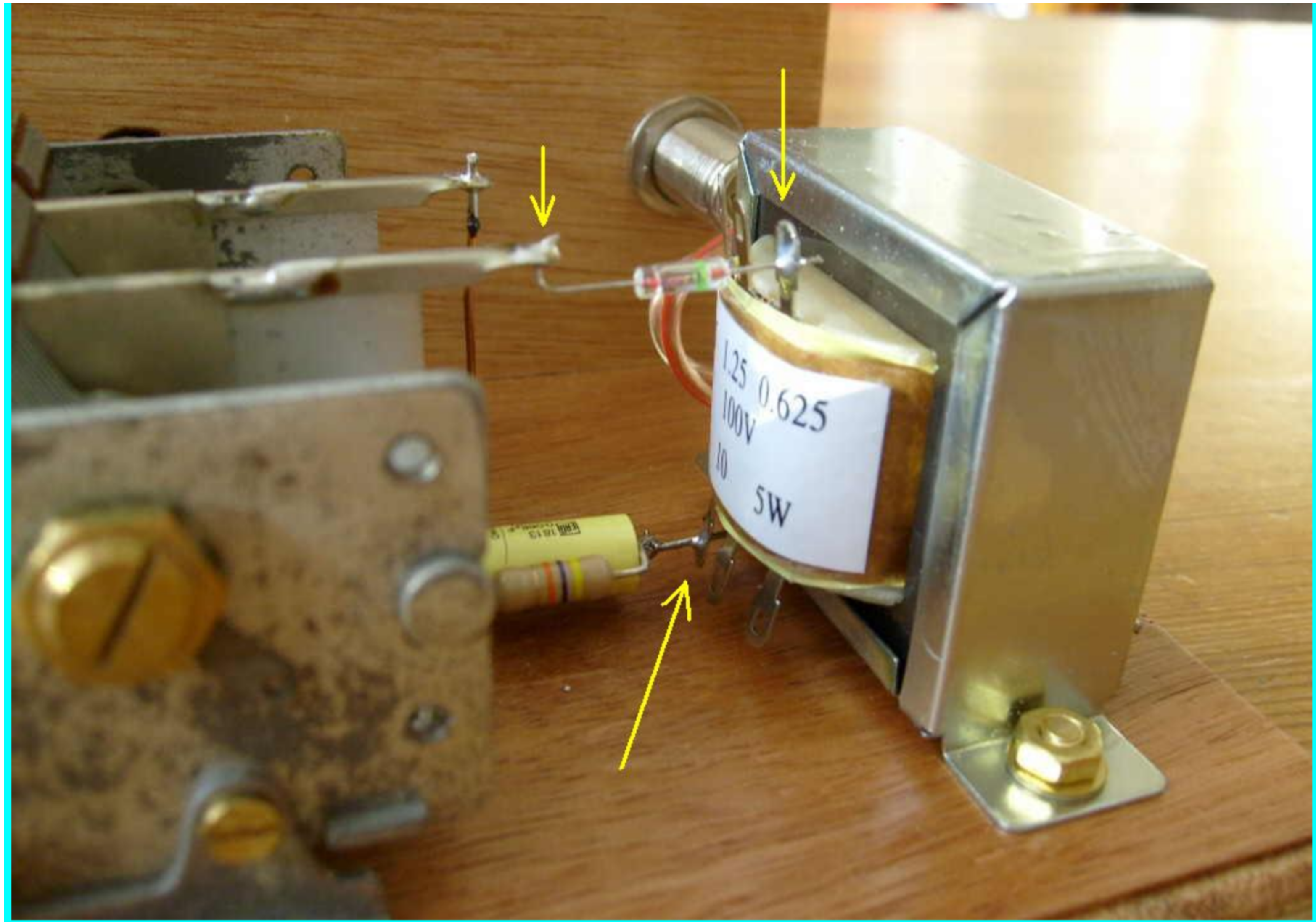

Place the transformer on the frame, fix it with two M4x16 screws, 2 rings and 2 nuts
The wire of the 68 nF capacitor must stick through the "0" connection of the transformer, and is then soldered.
Solder the OA95 diode between tuning capacitor and the "0.625W" connection of the transformer.
The green band on the diode must point towards the transformer.

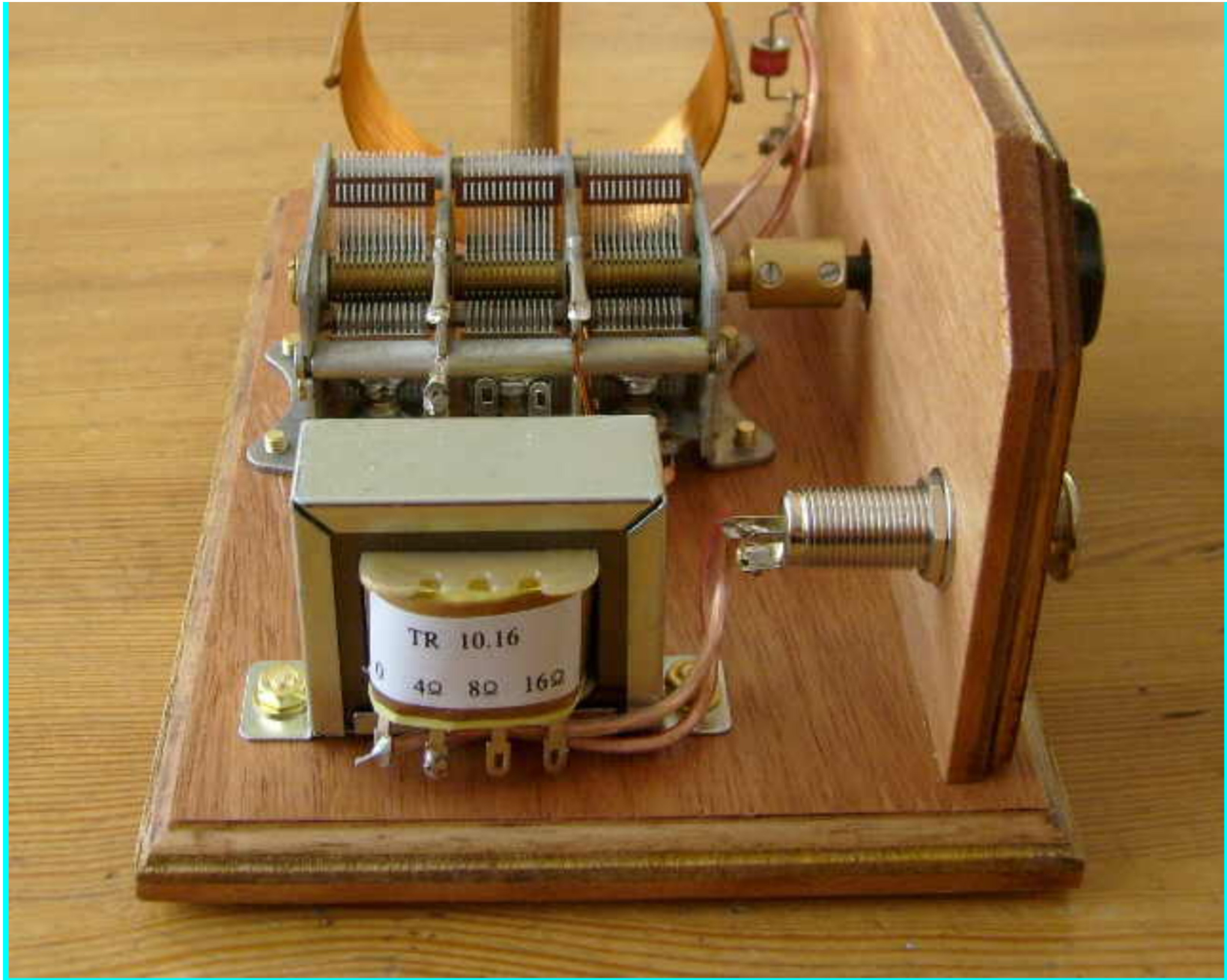
TR 10.16
0 4Ω 8Ω 16Ω

Solder the two wires from the transformer to the headphone socket (see for detail: next picture).

Solder the two wires from the tuning capacitor to the terminal posts.
The wire with the red line on it, comes on the upper (antenna) connection.
The wire without red line, comes on the lower (ground) connection.

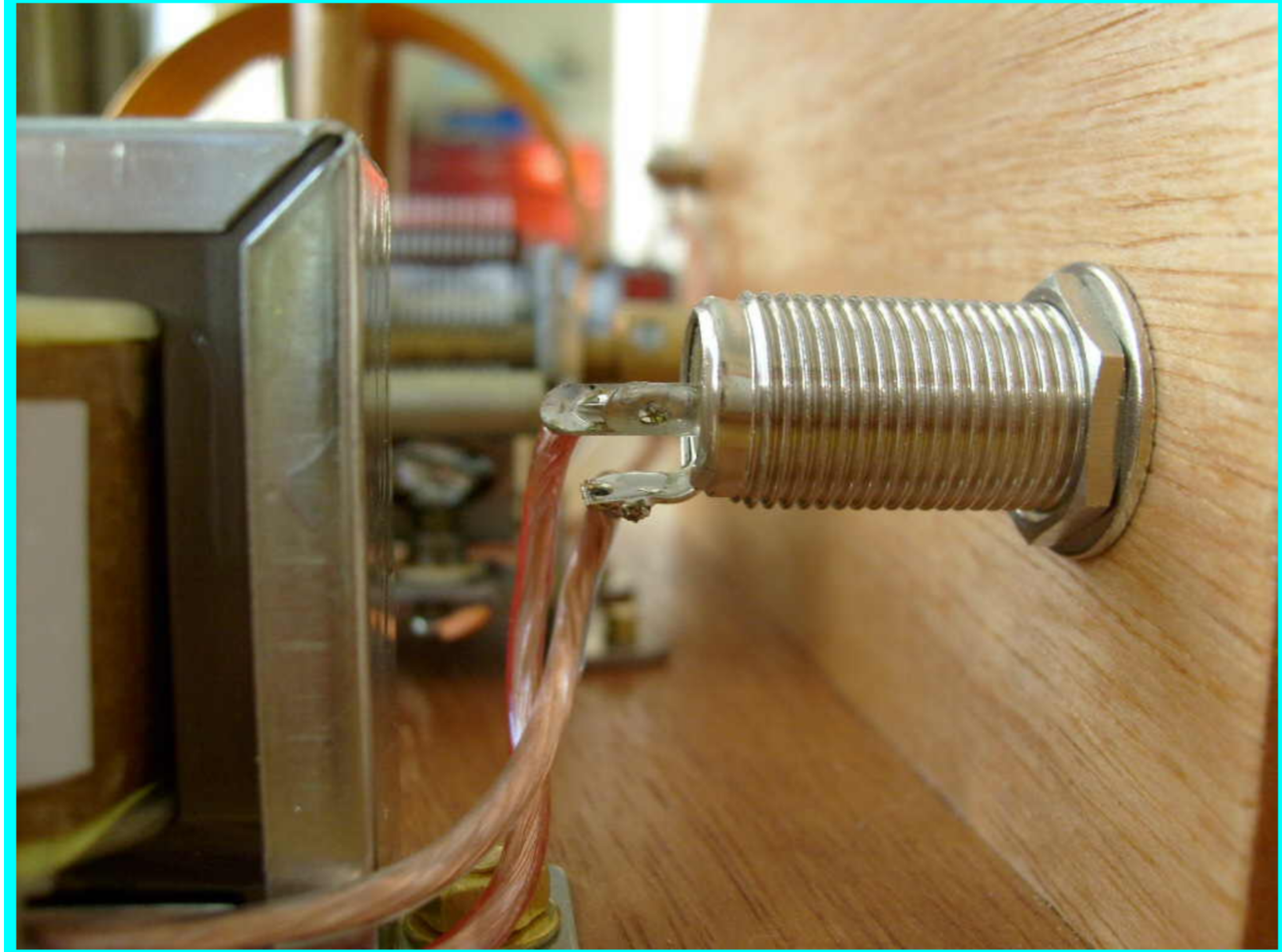

Detail of the wires on the headphone socket.
The wire without red line comes on the ground connection.
The wire with the red line comes on the two signal pins, so these two signal pins are connected together.

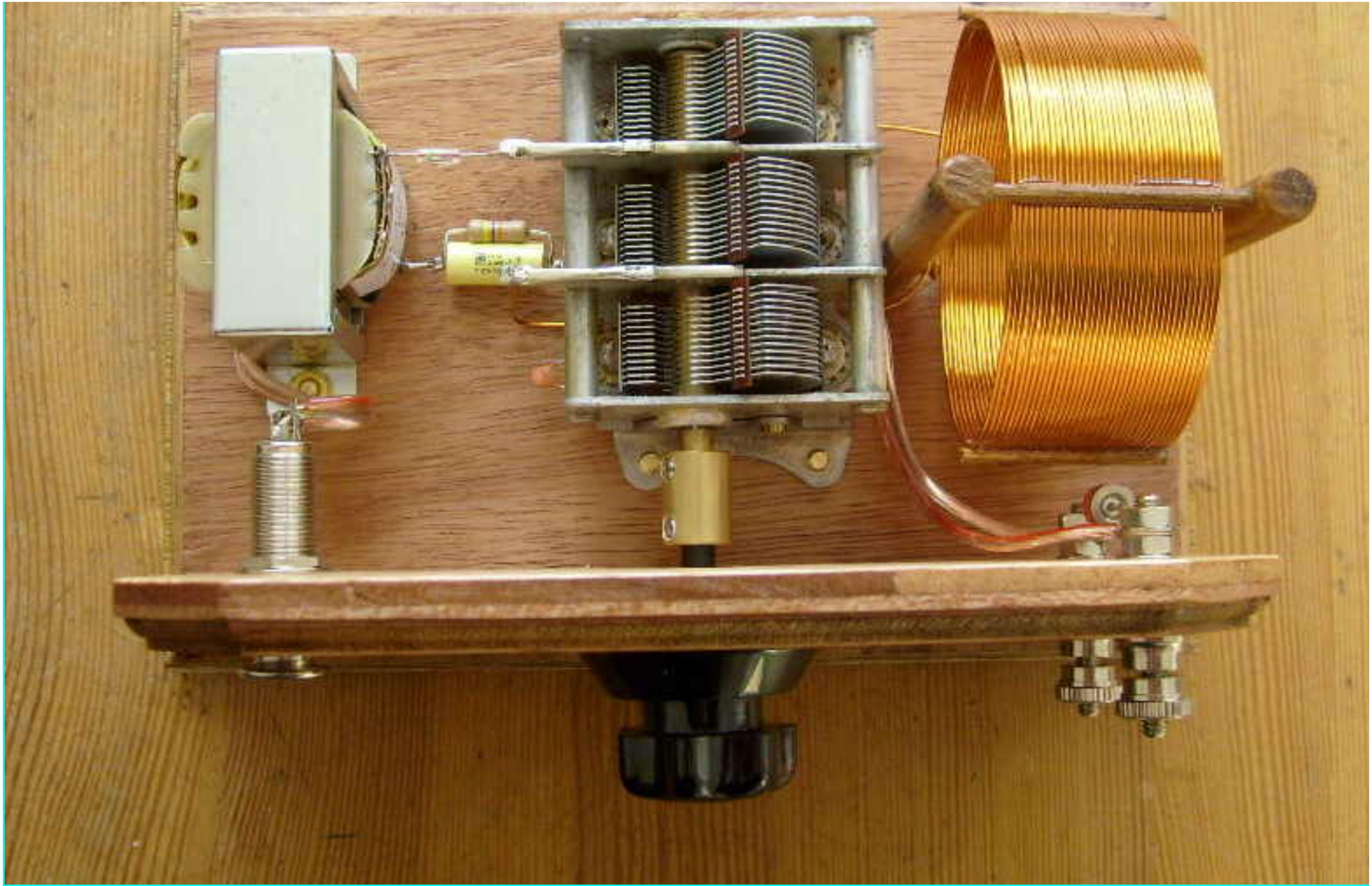

Mount the extension shaft and the knob on the tuning capacitor.
In the middle position, the pointer on the knob must point upwards.

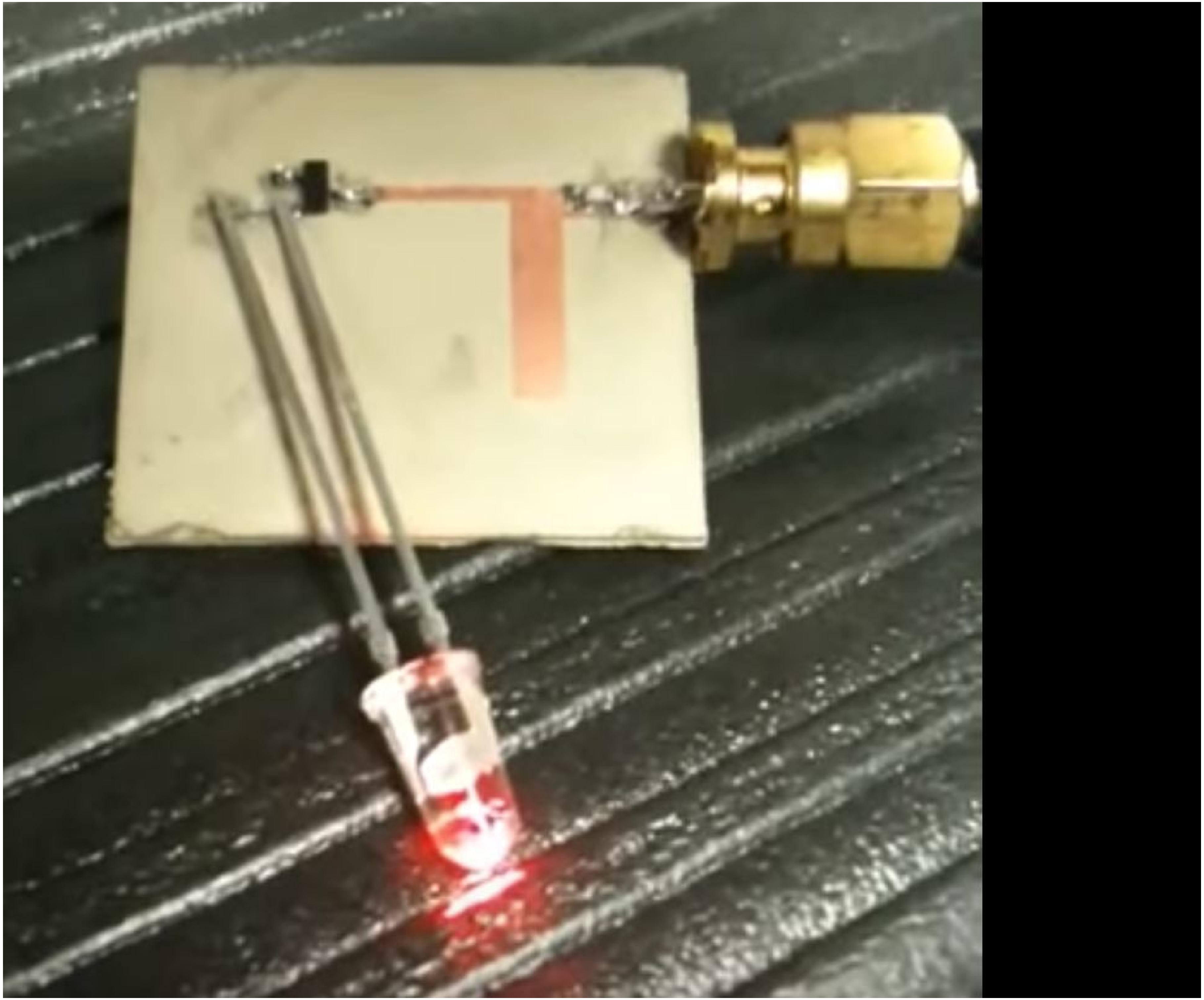

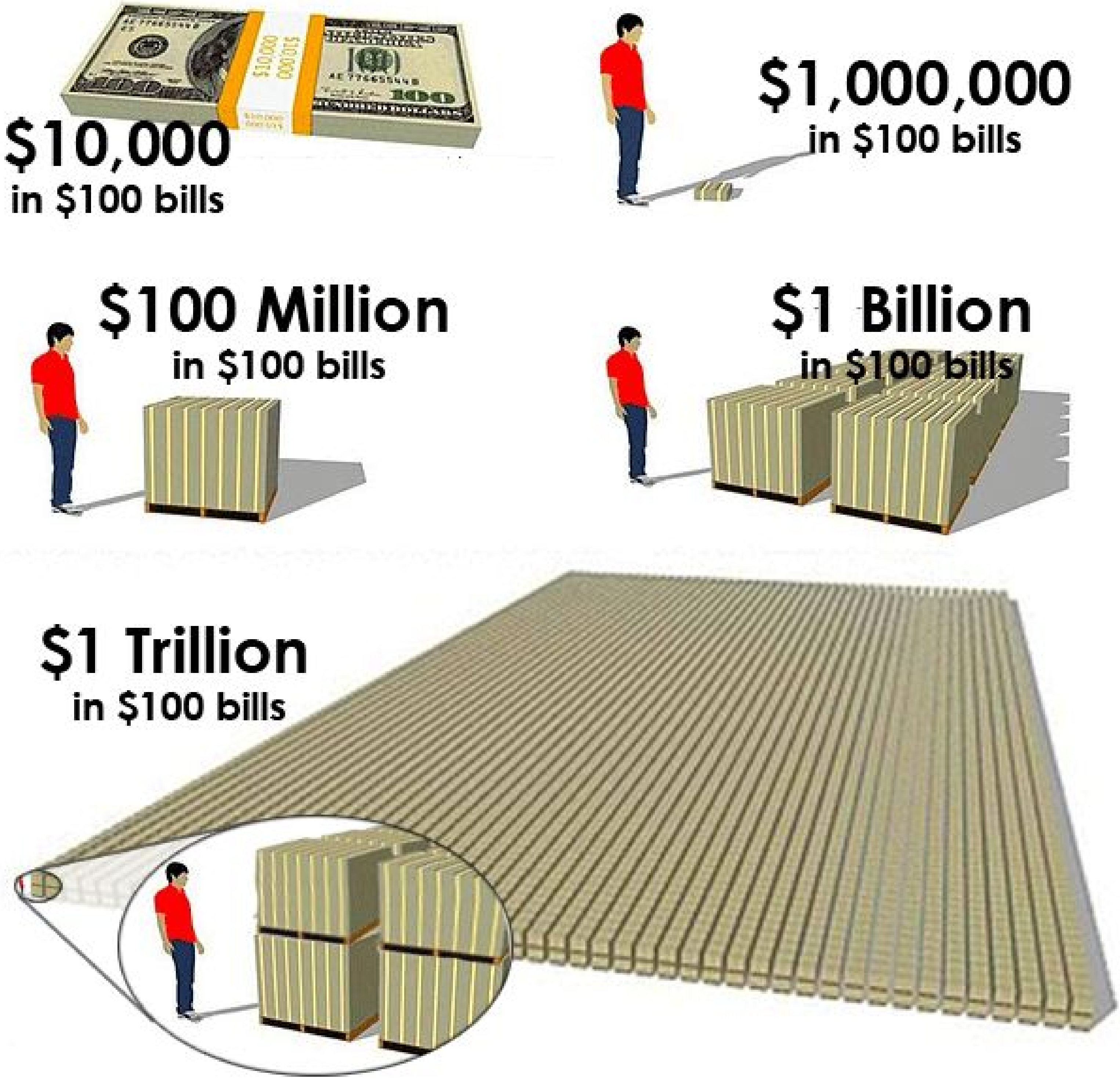
$10,000
in $100 bills
$1,000,000
in $100 bills
$100 Million
in $100 bills
$1 Billion
in $100 bills
$1 Trillion
in $100 bills

USB Type A Female Field Termination Connector

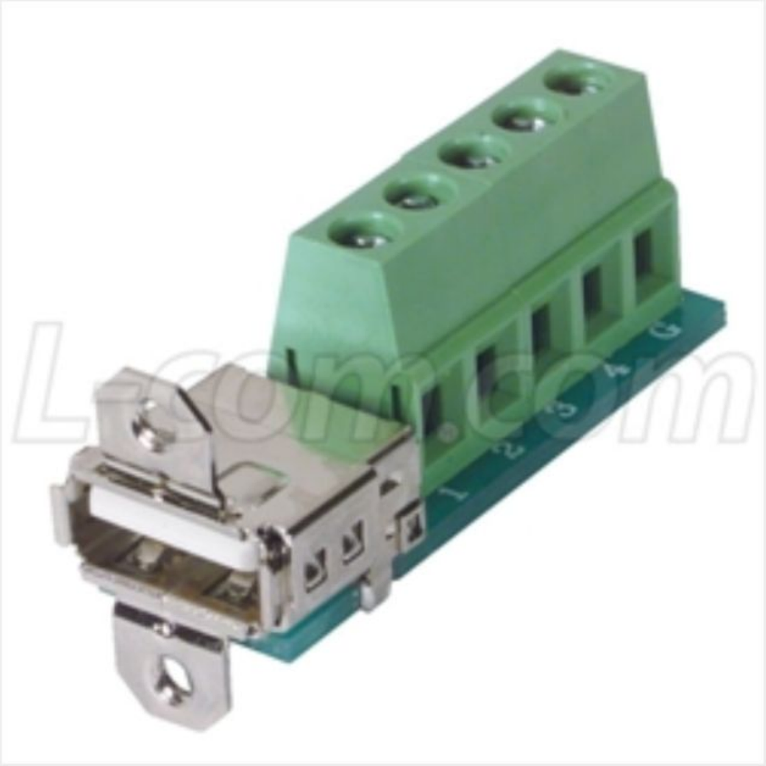

Roll over image to zoom in

L-com Item # USBAFT

List Price	$7.24
Your Price 1-9	**$7.24**
10-24	$6.81
25-99	$6.37
100 +	Call Us

Availability: In Stock

Available for Same Day Shipping

Quantity 1 ADD TO CART

Email This Page

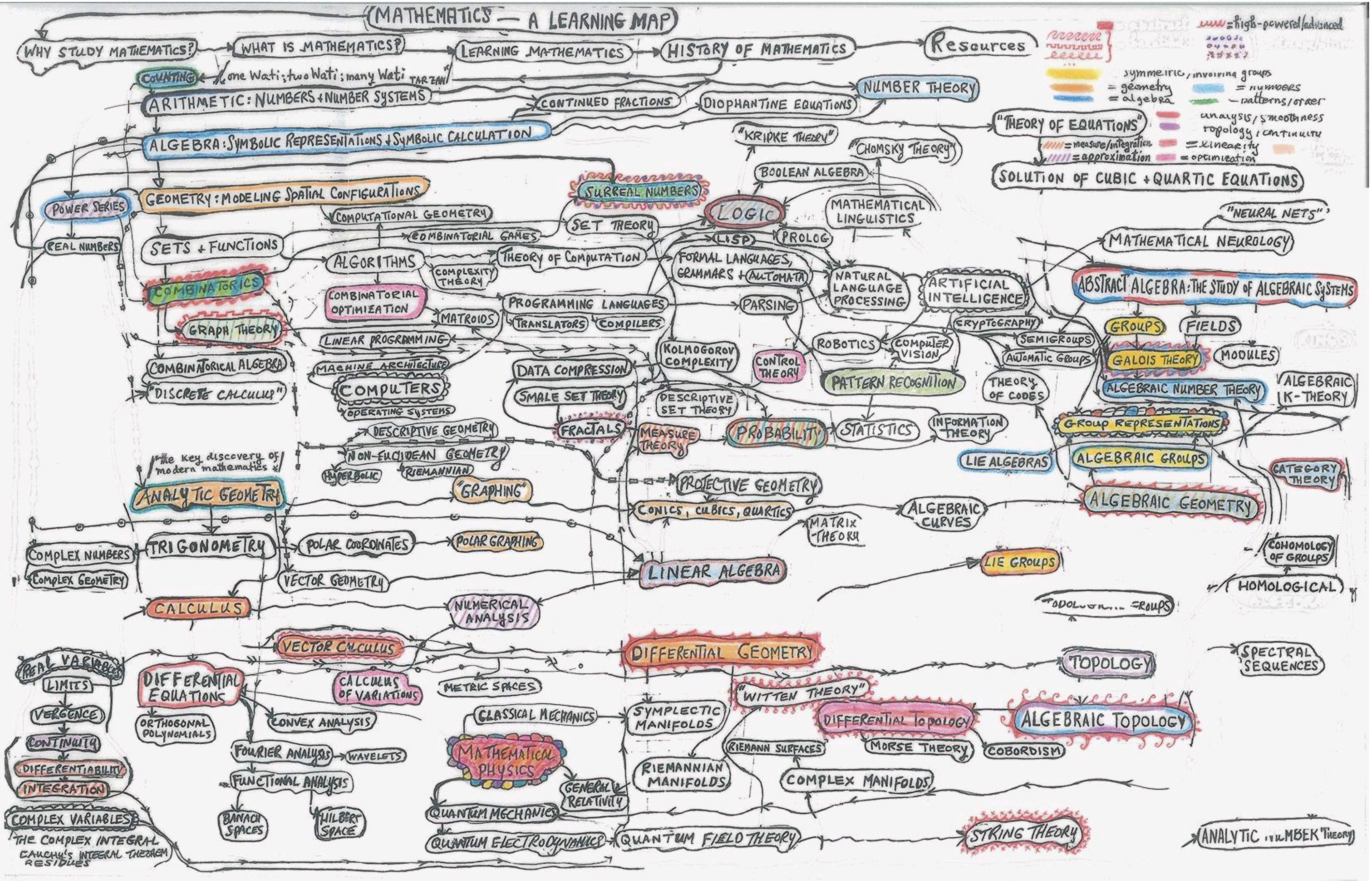
MATHEMATICS — A LEARNING MAP
WHY STUDY MATHEMATICS?
WHAT IS MATHEMATICS?
LEARNING MATHEMATICS
HISTORY OF MATHEMATICS
Resources
= high-powered/advanced
symmetric, involving groups
= geometry
= algebra
= numbers
— patterns/order
analysis/smoothness
topology/continuity
= measure/integration
= approximation
= linearity
= optimization
COUNTING
one Wati; two Wati; many Wati TARZAN
ARITHMETIC: NUMBERS + NUMBER SYSTEMS
CONTINUED FRACTIONS
DIOPHANTINE EQUATIONS
NUMBER THEORY
"THEORY OF EQUATIONS"
SOLUTION OF CUBIC + QUARTIC EQUATIONS
ALGEBRA: SYMBOLIC REPRESENTATIONS + SYMBOLIC CALCULATION
"KRIPKE THEORY"
"CHOMSKY THEORY"
BOOLEAN ALGEBRA
GEOMETRY: MODELING SPATIAL CONFIGURATIONS
SURREAL NUMBERS
LOGIC
MATHEMATICAL LINGUISTICS
POWER SERIES
COMPUTATIONAL GEOMETRY
SET THEORY
"NEURAL NETS"
REAL NUMBERS
COMBINATORIAL GAMES
LISP
PROLOG
MATHEMATICAL NEUROLOGY
SETS + FUNCTIONS
ALGORITHMS
THEORY OF COMPUTATION
COMPLEXITY THEORY
FORMAL LANGUAGES, GRAMMARS + AUTOMATA
NATURAL LANGUAGE PROCESSING
ARTIFICIAL INTELLIGENCE
ABSTRACT ALGEBRA: THE STUDY OF ALGEBRAIC SYSTEMS
COMBINATORICS
COMBINATORIAL OPTIMIZATION
PROGRAMMING LANGUAGES
PARSING
TRANSLATORS
COMPILERS
MATROIDS
CRYPTOGRAPHY
GROUPS
FIELDS
GRAPH THEORY
LINEAR PROGRAMMING
ROBOTICS
COMPUTER VISION
SEMIGROUPS
MODULES
COMBINATORICAL ALGEBRA
MACHINE ARCHITECTURE
KOLMOGOROV COMPLEXITY
CONTROL THEORY
AUTOMATIC GROUPS
GALOIS THEORY
COMPUTERS
OPERATING SYSTEMS
DATA COMPRESSION
PATTERN RECOGNITION
THEORY OF CODES
ALGEBRAIC NUMBER THEORY
ALGEBRAIC K-THEORY
"DISCRETE CALCULUS"
SMALE SET THEORY
DESCRIPTIVE SET THEORY
DESCRIPTIVE GEOMETRY
FRACTALS
MEASURE THEORY
PROBABILITY
STATISTICS
INFORMATION THEORY
GROUP REPRESENTATIONS
"the key discovery of modern mathematics"
NON-EUCLIDEAN GEOMETRY
HYPERBOLIC
RIEMANNIAN
LIE ALGEBRAS
ALGEBRAIC GROUPS
CATEGORY THEORY
ANALYTIC GEOMETRY
"GRAPHING"
PROJECTIVE GEOMETRY
CONICS, CUBICS, QUARTICS
ALGEBRAIC CURVES
ALGEBRAIC GEOMETRY
MATRIX THEORY
COMPLEX NUMBERS
TRIGONOMETRY
POLAR COORDINATES
POLAR GRAPHING
COHOMOLOGY OF GROUPS
COMPLEX GEOMETRY
VECTOR GEOMETRY
LINEAR ALGEBRA
LIE GROUPS
HOMOLOGICAL
CALCULUS
NUMERICAL ANALYSIS
TOPOLOGICAL GROUPS
VECTOR CALCULUS
DIFFERENTIAL GEOMETRY
SPECTRAL SEQUENCES
REAL VARIABLES
LIMITS
DIFFERENTIAL EQUATIONS
CALCULUS OF VARIATIONS
METRIC SPACES
TOPOLOGY
"WITTEN THEORY"
VERGENCE
ORTHOGONAL POLYNOMIALS
CONVEX ANALYSIS
CLASSICAL MECHANICS
SYMPLECTIC MANIFOLDS
DIFFERENTIAL TOPOLOGY
ALGEBRAIC TOPOLOGY
CONTINUITY
FOURIER ANALYSIS
WAVELETS
MATHEMATICAL PHYSICS
RIEMANN SURFACES
MORSE THEORY
COBORDISM
DIFFERENTIABILITY
INTEGRATION
FUNCTIONAL ANALYSIS
GENERAL RELATIVITY
RIEMANNIAN MANIFOLDS
COMPLEX MANIFOLDS
COMPLEX VARIABLES
BANACH SPACES
HILBERT SPACE
QUANTUM MECHANICS
THE COMPLEX INTEGRAL
CAUCHY'S INTEGRAL THEOREM
RESIDUES
QUANTUM ELECTRODYNAMICS
QUANTUM FIELD THEORY
STRING THEORY
ANALYTIC NUMBER THEORY

Philmore

MANUFACTURING CO. Inc.
Richmond Hill 18, N. Y.

MODEL NO.
VC-1000

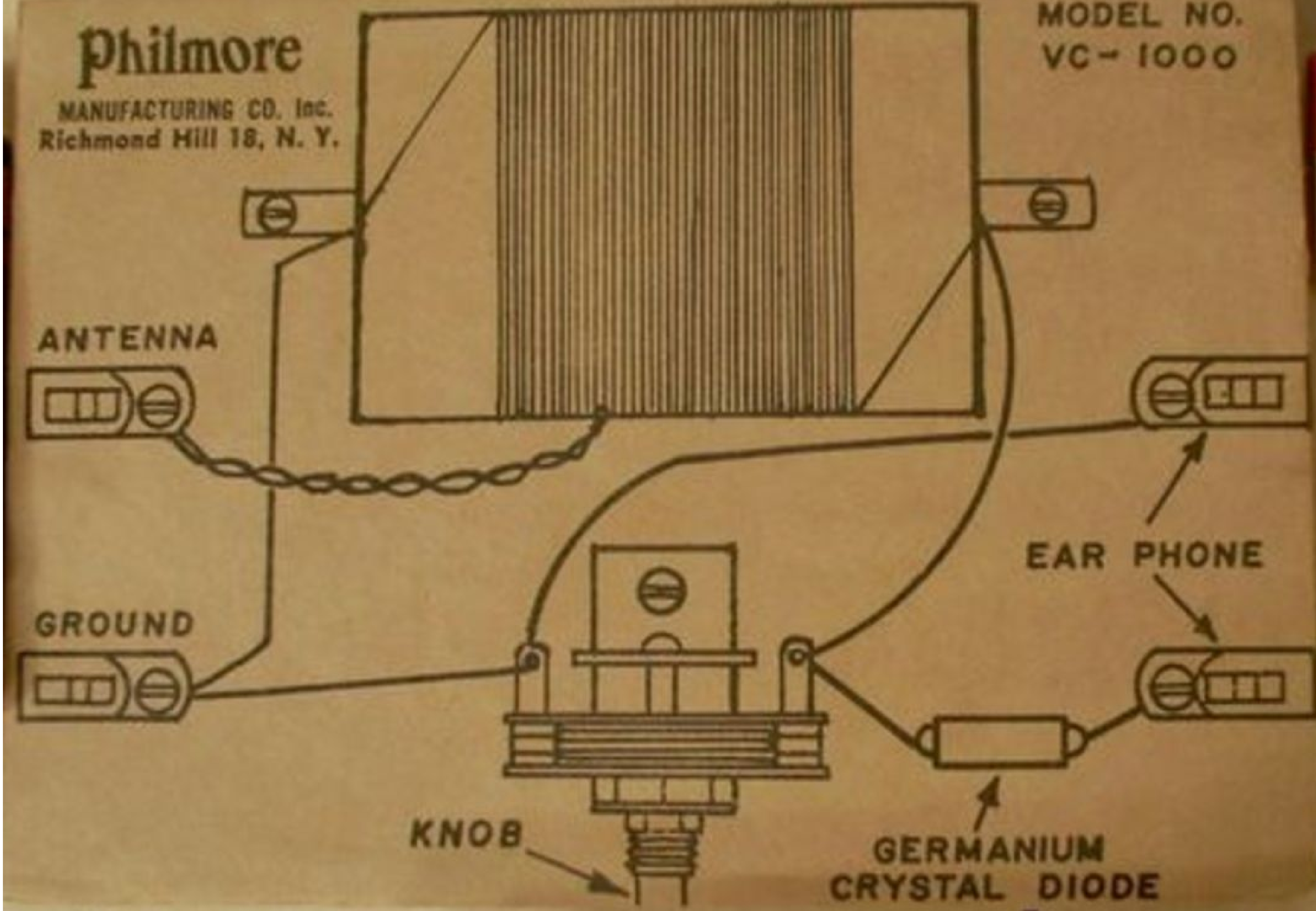
ANTENNA
GROUND
EAR PHONE
KNOB
GERMANIUM
CRYSTAL DIODE

1 LED azul
(con tensión entre 3V y 3,6V)

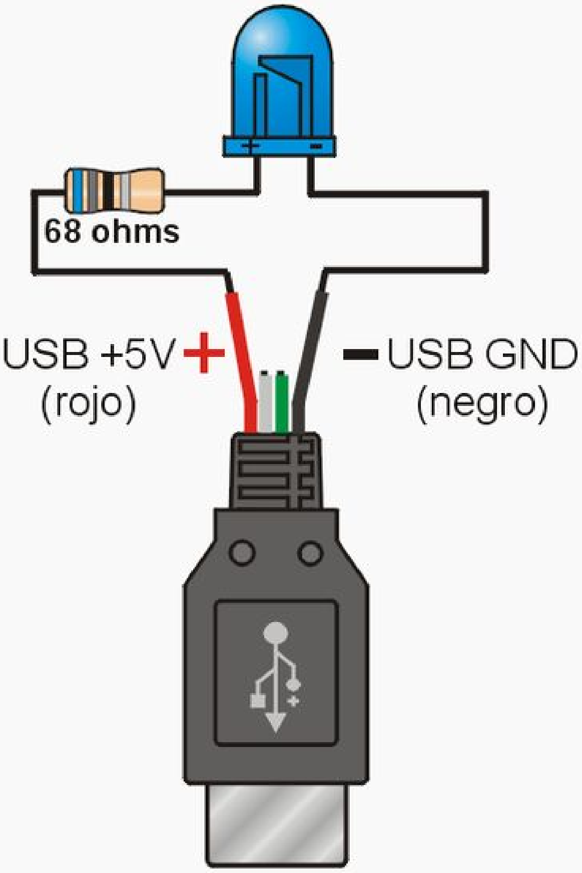
68 ohms
USB +5V +
(rojo)
− USB GND
(negro)

WWW.INVENTABLE.EU

1 LED rojo
(con tensión entre 1,8V y 2,1V)

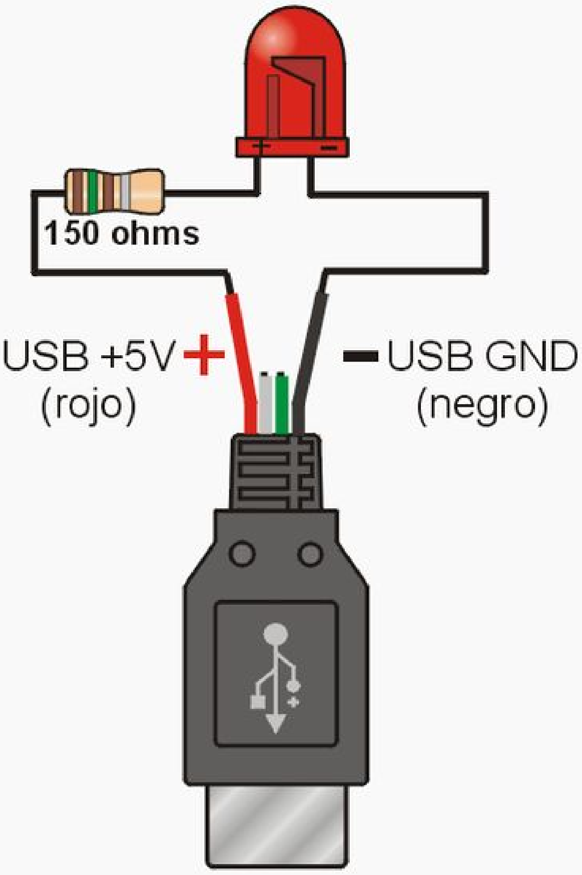
150 ohms
USB +5V +
(rojo)
− USB GND
(negro)

The Tesla Experiment

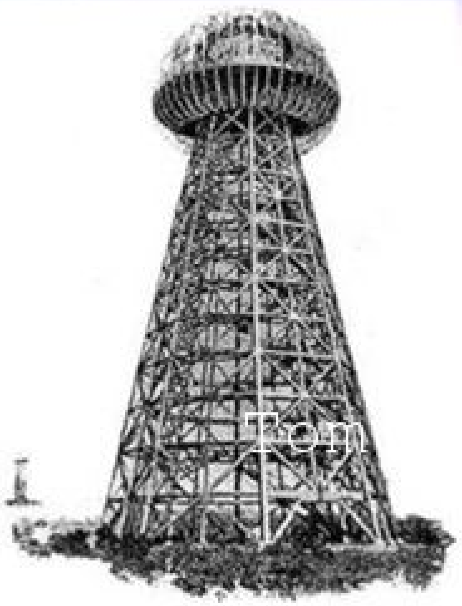

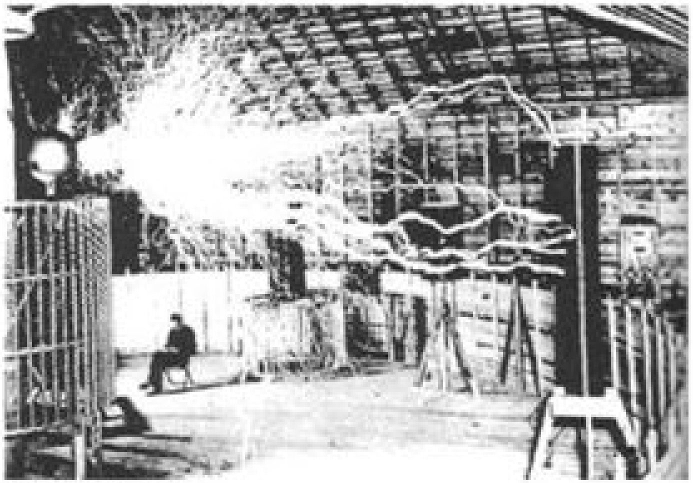

TRANSVERSE WAVE INTERFERENCE

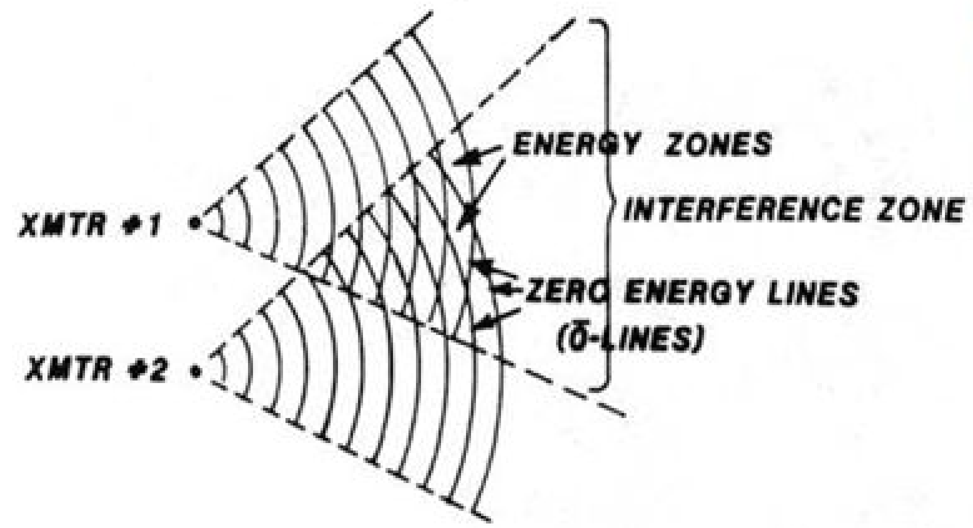
ENERGY ZONES
INTERFERENCE ZONE
XMTR #1
ZERO ENERGY LINES
(0̄-LINES)
XMTR #2

LONGITUDINAL WAVE INTERFERENCE
(SCALAR WAVES)

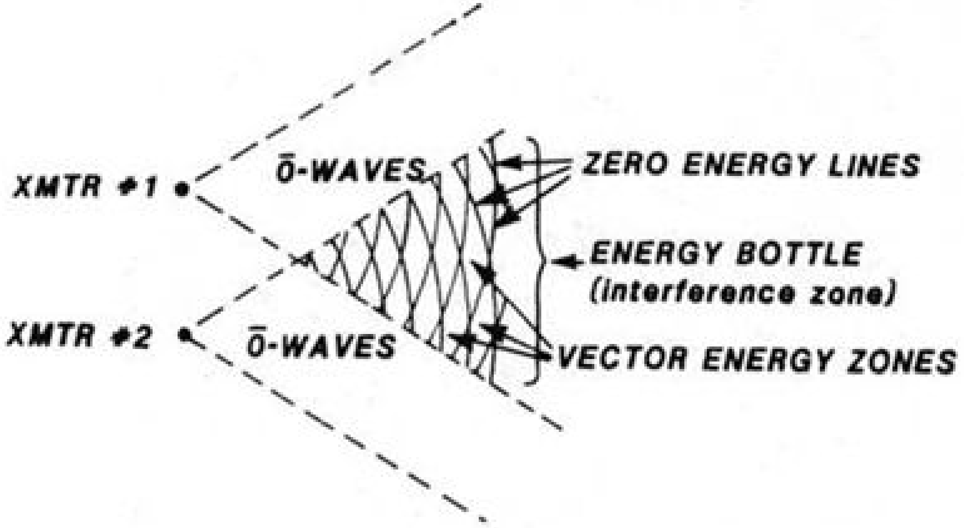
XMTR #1
0̄-WAVES
ZERO ENERGY LINES
ENERGY BOTTLE
(interference zone)
XMTR #2
0̄-WAVES
VECTOR ENERGY ZONES

© PoweredTemplates.com

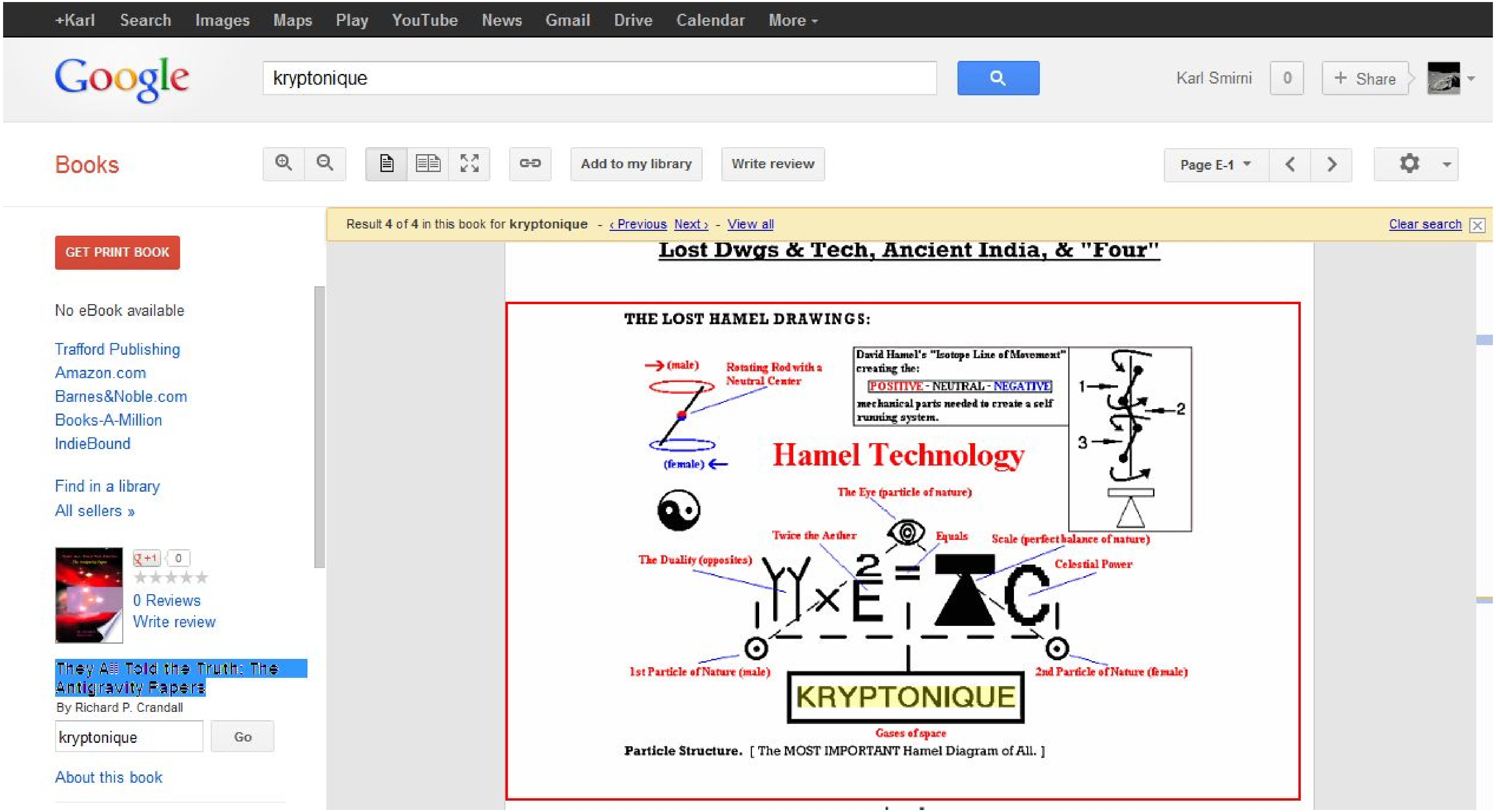
+Karl
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
kryptonique
Karl Smirni
0
+ Share
Books
Add to my library
Write review
Page E-1
GET PRINT BOOK
No eBook available
Trafford Publishing
Amazon.com
Barnes&Noble.com
Books-A-Million
IndieBound
Find in a library
All sellers »
+1
0
0 Reviews
Write review
They All Told the Truth: The Antigravity Papers
By Richard P. Crandall
kryptonique
Go
About this book
Clear search

Result 4 of 4 in this book for **kryptonique** - ‹ Previous Next › - View all

Lost Dwgs & Tech, Ancient India, & "Four"

THE LOST HAMEL DRAWINGS:

David Hamel's "Isotope Line of Movement" creating the:
POSITIVE - NEUTRAL - NEGATIVE
mechanical parts needed to create a self running system.

(male)
Rotating Rod with a Neutral Center
(female)
Hamel Technology
1
2
3
The Eye (particle of nature)
Twice the Aether
Equals
Scale (perfect balance of nature)
The Duality (opposites)
Celestial Power
2
E
C
1st Particle of Nature (male)
2nd Particle of Nature (female)
KRYPTONIQUE
Gases of space

Particle Structure. [The MOST IMPORTANT Hamel Diagram of All.]

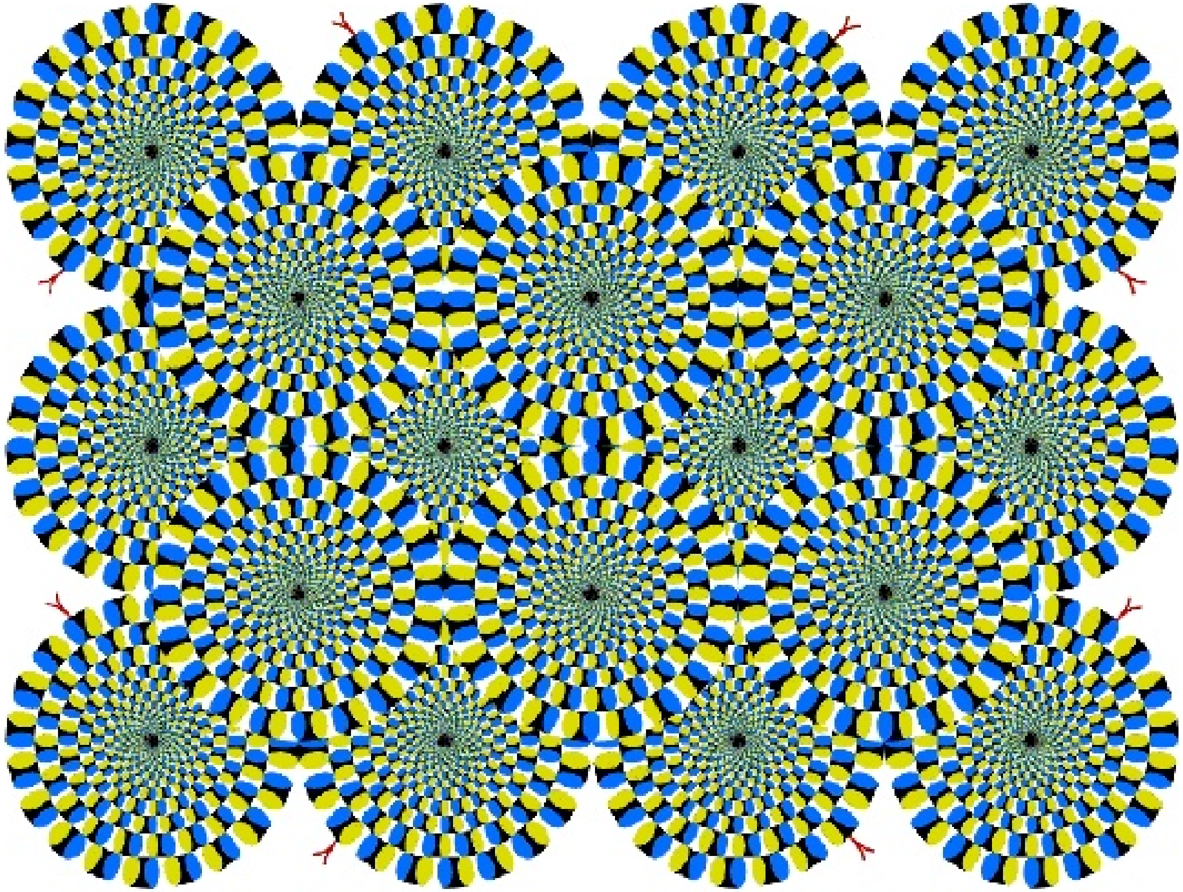

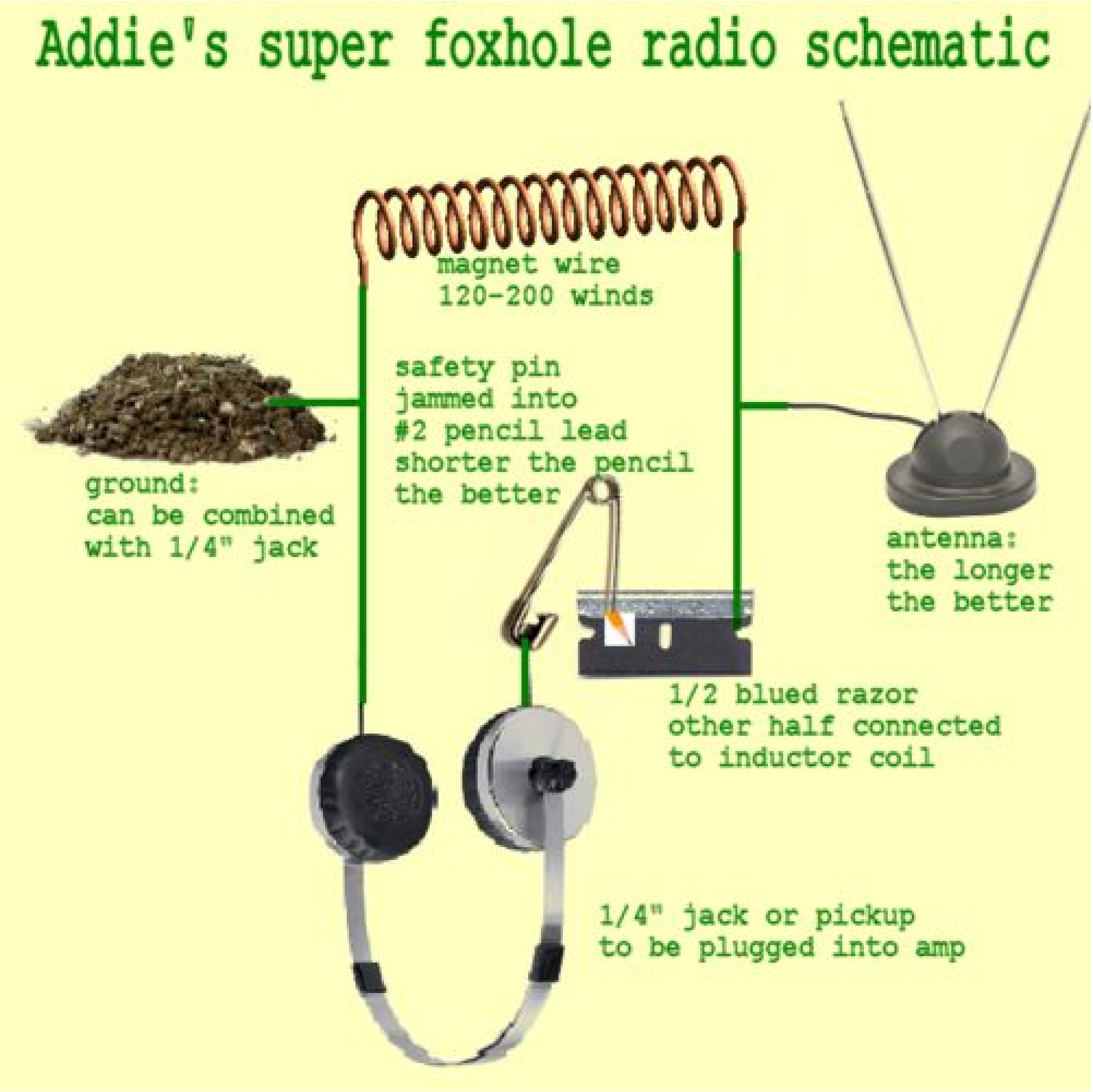
Addie's super foxhole radio schematic
magnet wire
120-200 winds
safety pin
jammed into
#2 pencil lead
shorter the pencil
the better
ground:
can be combined
with 1/4" jack
antenna:
the longer
the better
1/2 blued razor
other half connected
to inductor coil
1/4" jack or pickup
to be plugged into amp

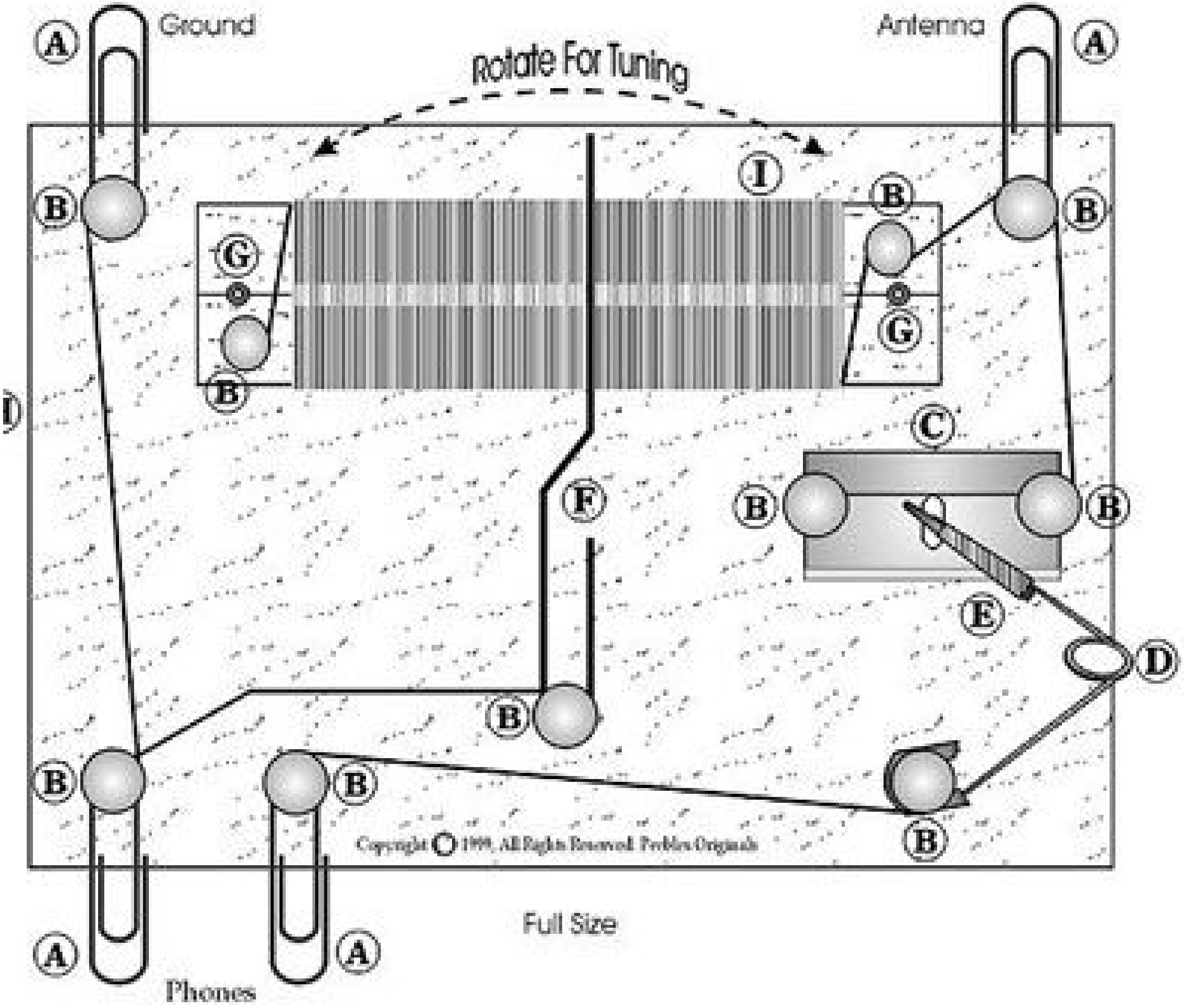
Ground
Antenna
Rotate For Tuning
A
A
A
A
B
B
B
B
B
B
B
B
B
B
C
D
E
F
G
G
I
I
Copyright © 1999, All Rights Reserved. Peebles Originals
Phones

Full Size

The PO-102, Crystal Radio Kit, w/Slider-Tuning!

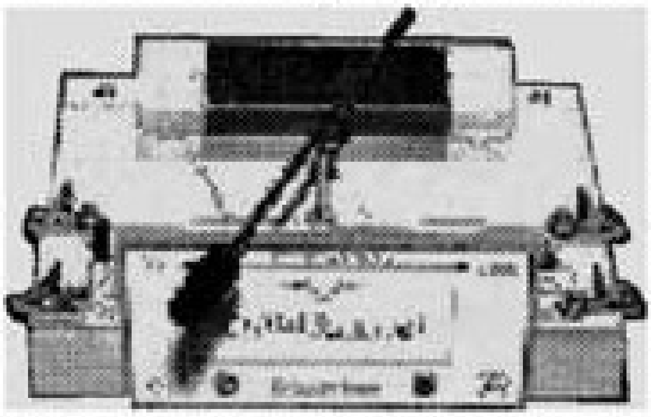

Perfect for parents, grandparents, and teachers to build with kids! Also a super starter kit for the beginning adult builder! Includes small coil form, wire, slider assembly, capacitors, and diode. Easy parts layout guide provided for kids. Excellent performance for such a simple but elegant classic slider. Earplug included.

PO-102, w/Crystal Earphone

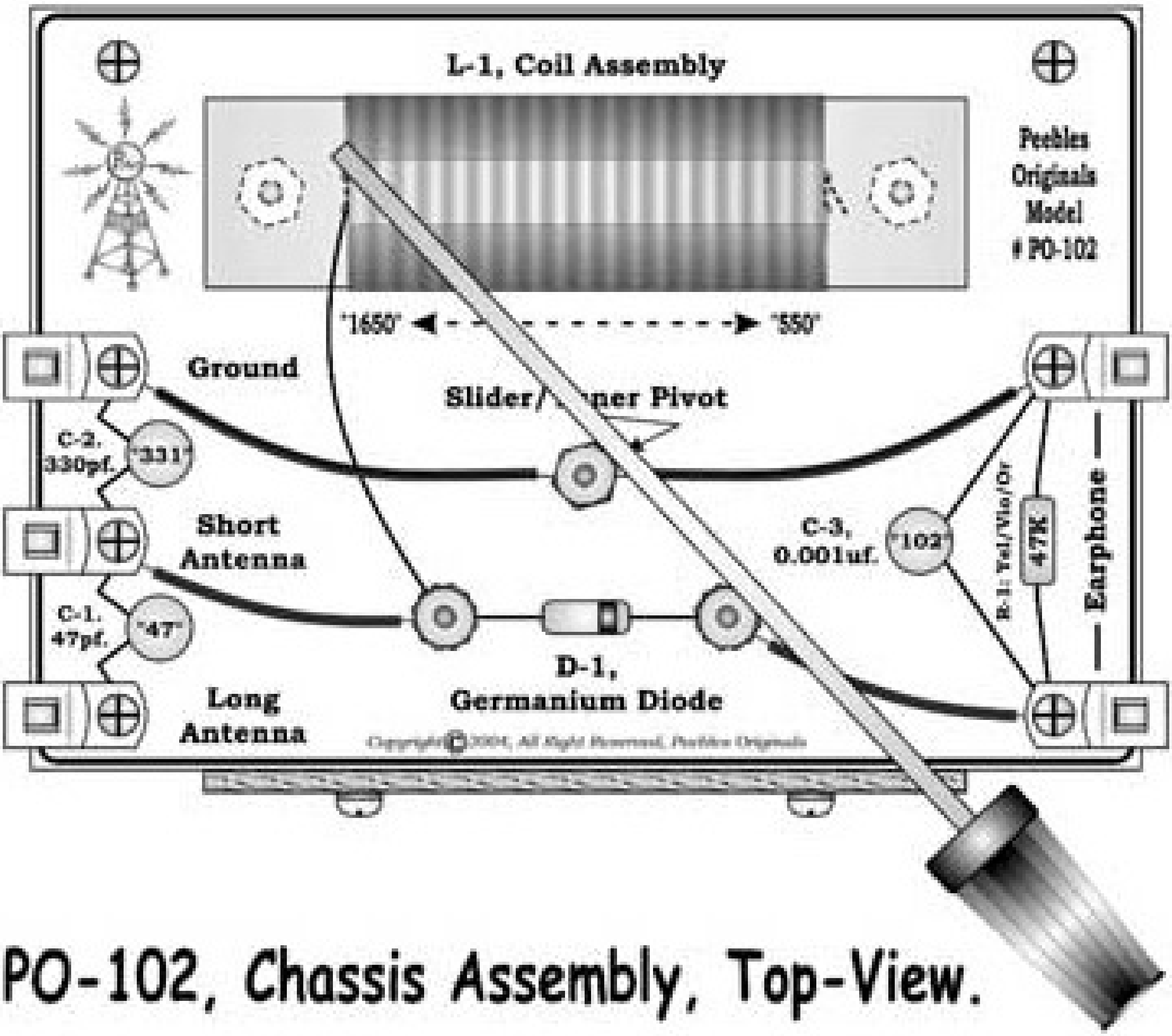
L-1, Coil Assembly
Peebles Originals Model # PO-102
"1650"
"550"
Ground
Slider/Tuner Pivot
C-2, 330pf.
"331"
Short Antenna
C-1, 47pf.
"47"
Long Antenna
D-1, Germanium Diode
C-3, 0.001uf.
"102"
R-1: Yel/Vio/Or
47K
Earphone
Copyright © 2004, All Right Reserved, Peebles Originals

PO-102, Chassis Assembly, Top-View.

FIRE KIT EXAMPLE

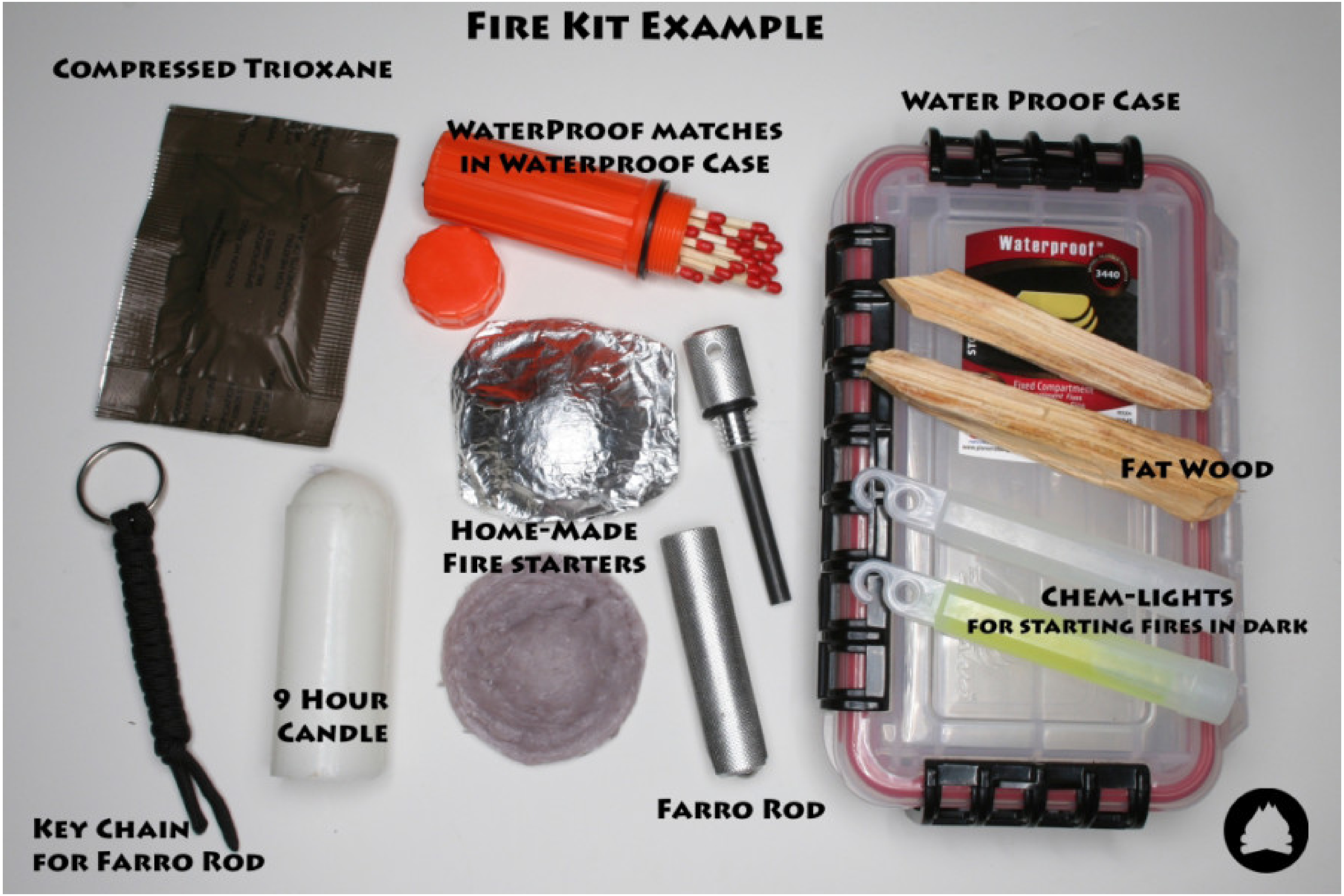
COMPRESSED TRIOXANE
WATERPROOF MATCHES IN WATERPROOF CASE
WATER PROOF CASE
Waterproof
3440
FAT WOOD
HOME-MADE FIRE STARTERS
CHEM-LIGHTS FOR STARTING FIRES IN DARK
9 HOUR CANDLE
FARRO ROD
KEY CHAIN FOR FARRO ROD

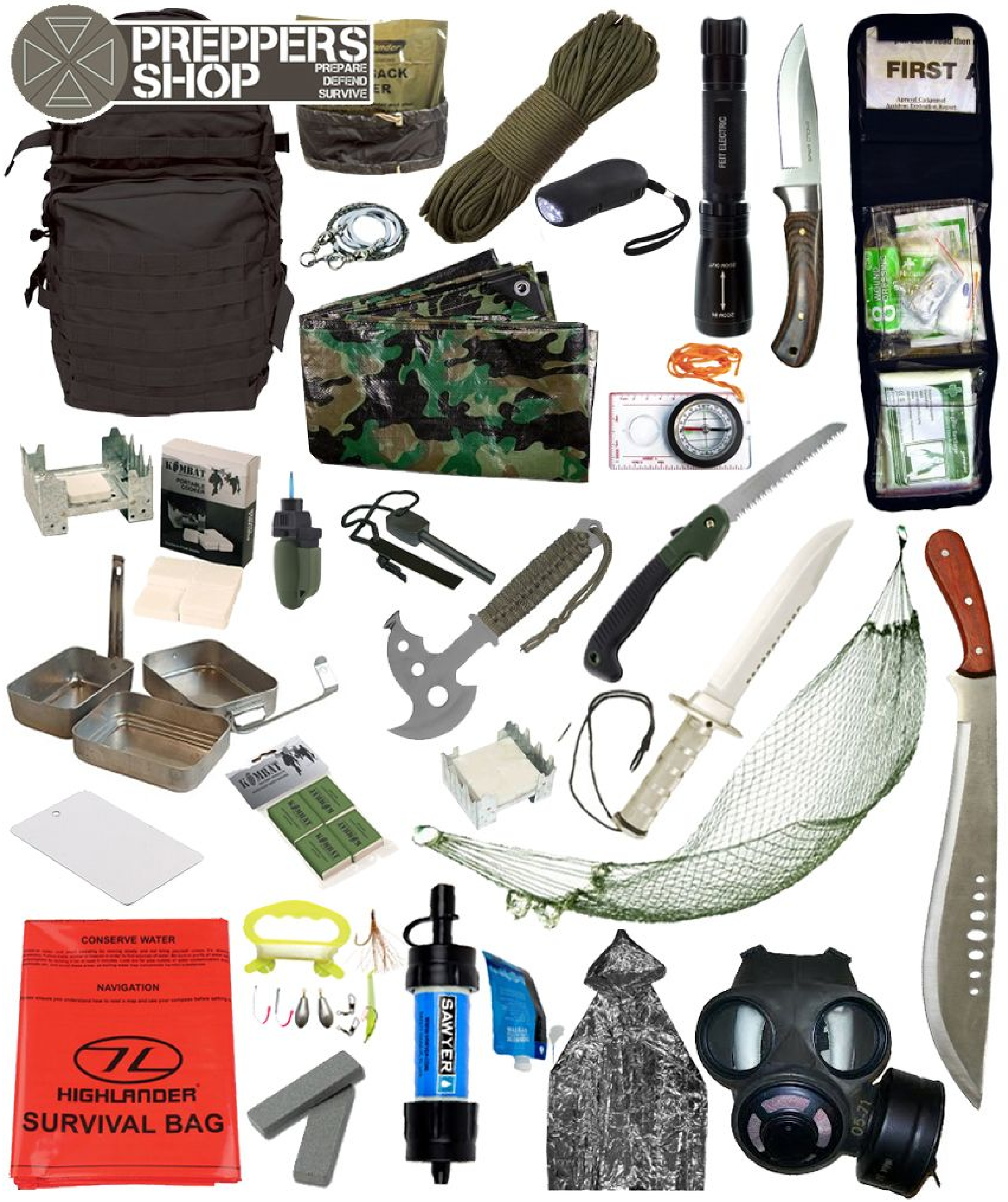
PREPPERS SHOP
PREPARE
DEFEND
SURVIVE
FIRST
FEIT ELECTRIC
KOMBAT
CONSERVE WATER
NAVIGATION
HIGHLANDER
SURVIVAL BAG
SAWYER

WWW.PREPPERSSHOP.CO.UK

THE ULTIMATE BUG OUT BAG

1. Anglo Arms Rambo Knife
2. Anglo Arms Machete
3. Austrian Ex-Military Water Flask
4. BCB Trekker Lifesaver Pack
5. BCB Compact Fishing Kit
6. Canadian Gas Mask w/ Filters
7. Feit Electric 500 LED Torch
8. French Military Mess Tins
9. Hexamine Solid Fuel Cooker
10. Heavy Duty Camo Tarp
11. Highlander Survival Bag
12. Highlander Solid Fuel Cooker
13. Highlander Trekker Hammock
14. Highlander Folding Saw
15. Kombat UK W/proof Matches
16. Kombat UK 40l Molle Bag
17. Kombat UK Fire Starting Kit
18. Kombat UK Compass
19. Kombat UK Throwing Axe
20. Oasis Water Purification Tablets
21. Pro Force Survival Saw
22. Sawyer Mini Water Filter
23. Sharpening Stone
24. Survival Mirror
25. Turboflame Lighter
26. UST Survival Poncho
27. Wateproof Rucksack Liner
28. Wind Up Solar Torch
29. Wooden Deluxe Knife w/ Case
30. 100ft Paracord
31. 24hr Ration Pack

£280

ER bar
DRINKING WATER
DATREX
Emergency Poncho
SLEEPING BAG
DUCT Tape
FRONTIER STRAW

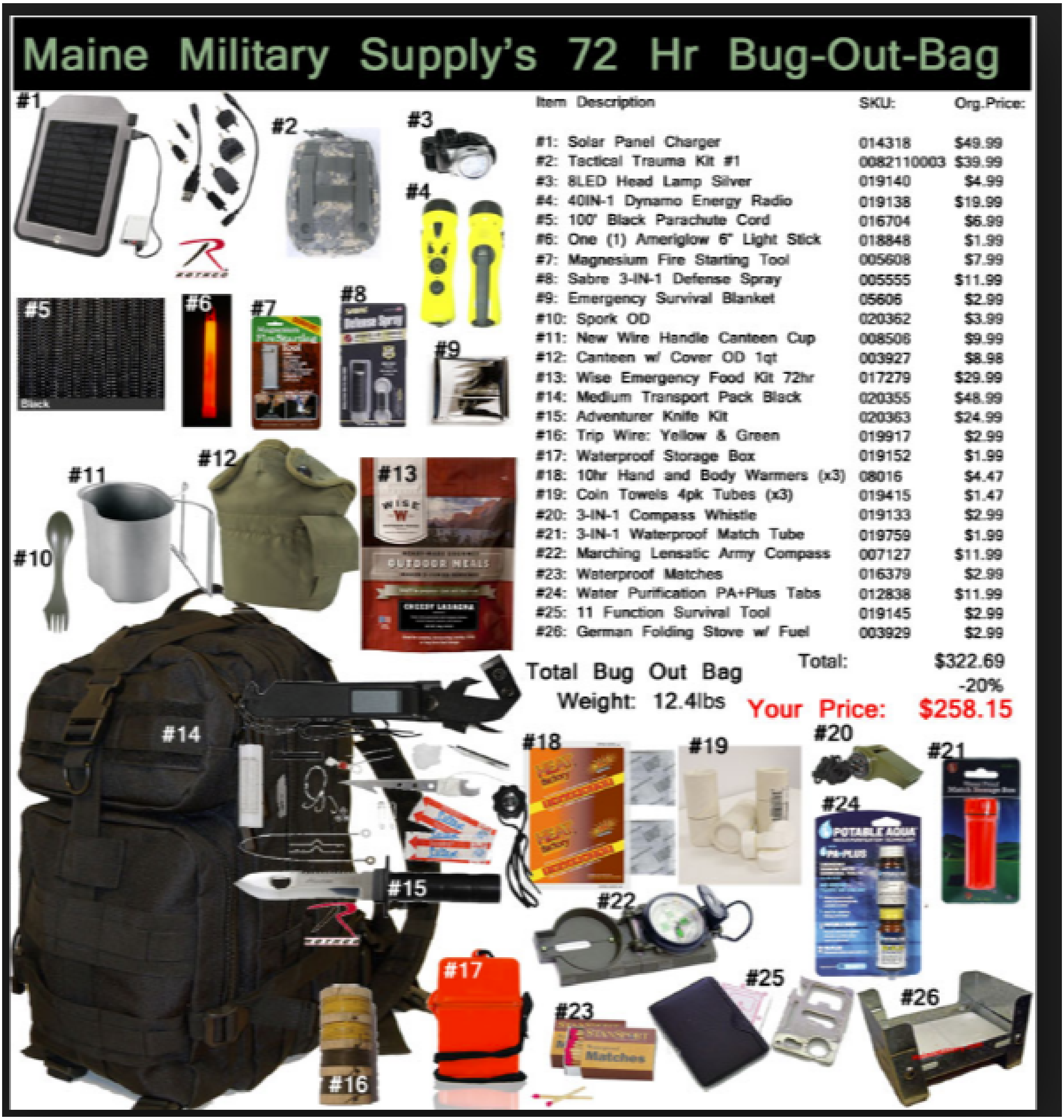

Maine Military Supply's 72 Hr Bug-Out-Bag

Item Description	SKU:	Org.Price:
#1: Solar Panel Charger	014318	$49.99
#2: Tactical Trauma Kit #1	0082110003	$39.99
#3: 8LED Head Lamp Silver	019140	$4.99
#4: 40IN-1 Dynamo Energy Radio	019138	$19.99
#5: 100' Black Parachute Cord	016704	$6.99
#6: One (1) Ameriglow 6" Light Stick	018848	$1.99
#7: Magnesium Fire Starting Tool	005608	$7.99
#8: Sabre 3-IN-1 Defense Spray	005555	$11.99
#9: Emergency Survival Blanket	05606	$2.99
#10: Spork OD	020362	$3.99
#11: New Wire Handle Canteen Cup	008506	$9.99
#12: Canteen w/ Cover OD 1qt	003927	$8.98
#13: Wise Emergency Food Kit 72hr	017279	$29.99
#14: Medium Transport Pack Black	020355	$48.99
#15: Adventurer Knife Kit	020363	$24.99
#16: Trip Wire: Yellow & Green	019917	$2.99
#17: Waterproof Storage Box	019152	$1.99
#18: 10hr Hand and Body Warmers (x3)	08016	$4.47
#19: Coin Towels 4pk Tubes (x3)	019415	$1.47
#20: 3-IN-1 Compass Whistle	019133	$2.99
#21: 3-IN-1 Waterproof Match Tube	019759	$1.99
#22: Marching Lensatic Army Compass	007127	$11.99
#23: Waterproof Matches	016379	$2.99
#24: Water Purification PA+Plus Tabs	012838	$11.99
#25: 11 Function Survival Tool	019145	$2.99
#26: German Folding Stove w/ Fuel	003929	$2.99
	Total:	$322.69
		-20%

Total Bug Out Bag
Weight: 12.4lbs

Your Price: $258.15

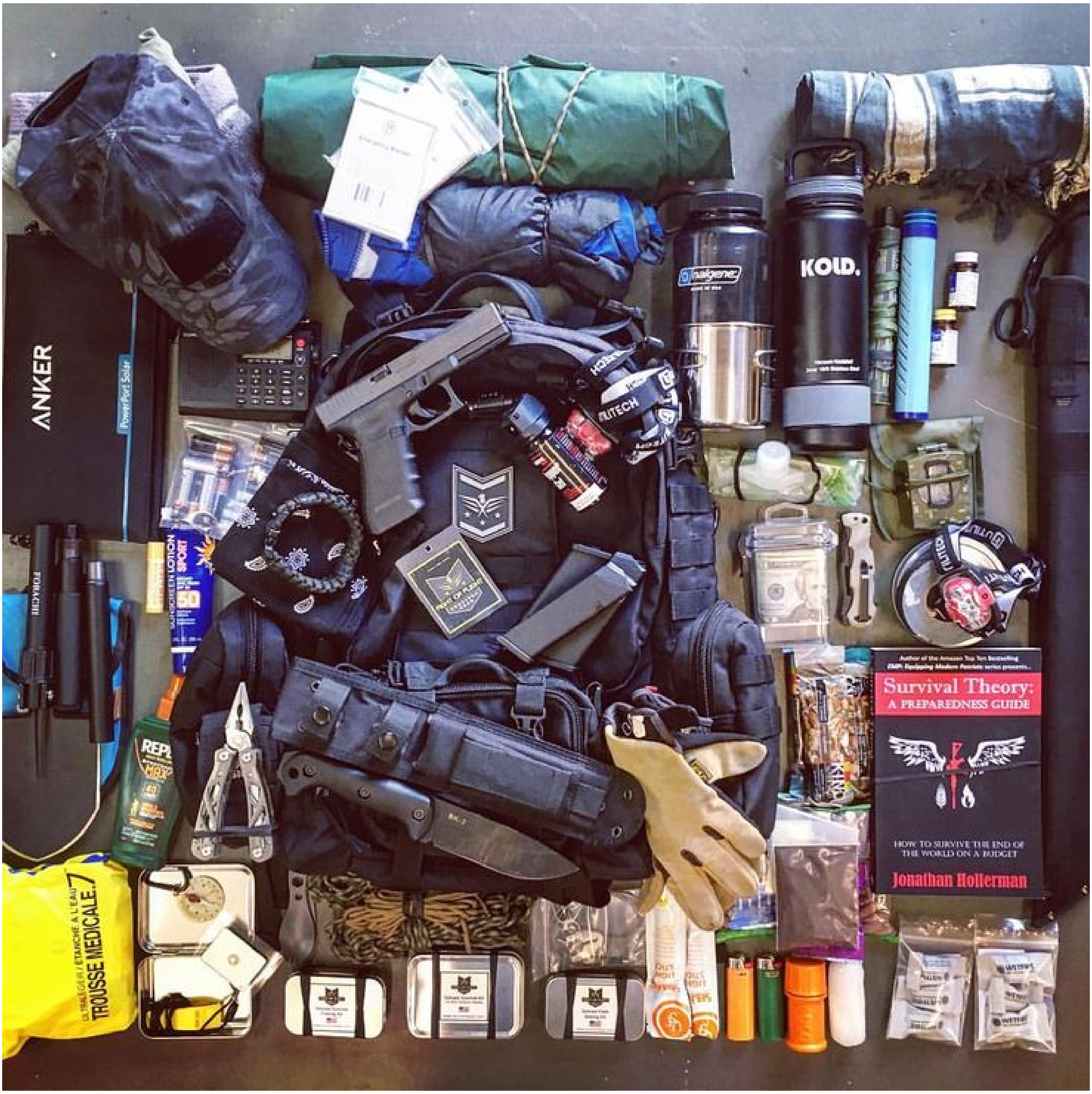
ANKER
PowerPort Solar
nalgene
KOLD.
SUNSCREEN LOTION SPORT 50
REPEL
TROUSSE MEDICALE .7
ULTRALÉGER / ÉTANCHE À L'EAU
Author of the Amazon Top Ten Bestselling EMP: Equipping Modern Patriots series presents...
Survival Theory:
A PREPAREDNESS GUIDE
HOW TO SURVIVE THE END OF THE WORLD ON A BUDGET
Jonathan Hollerman

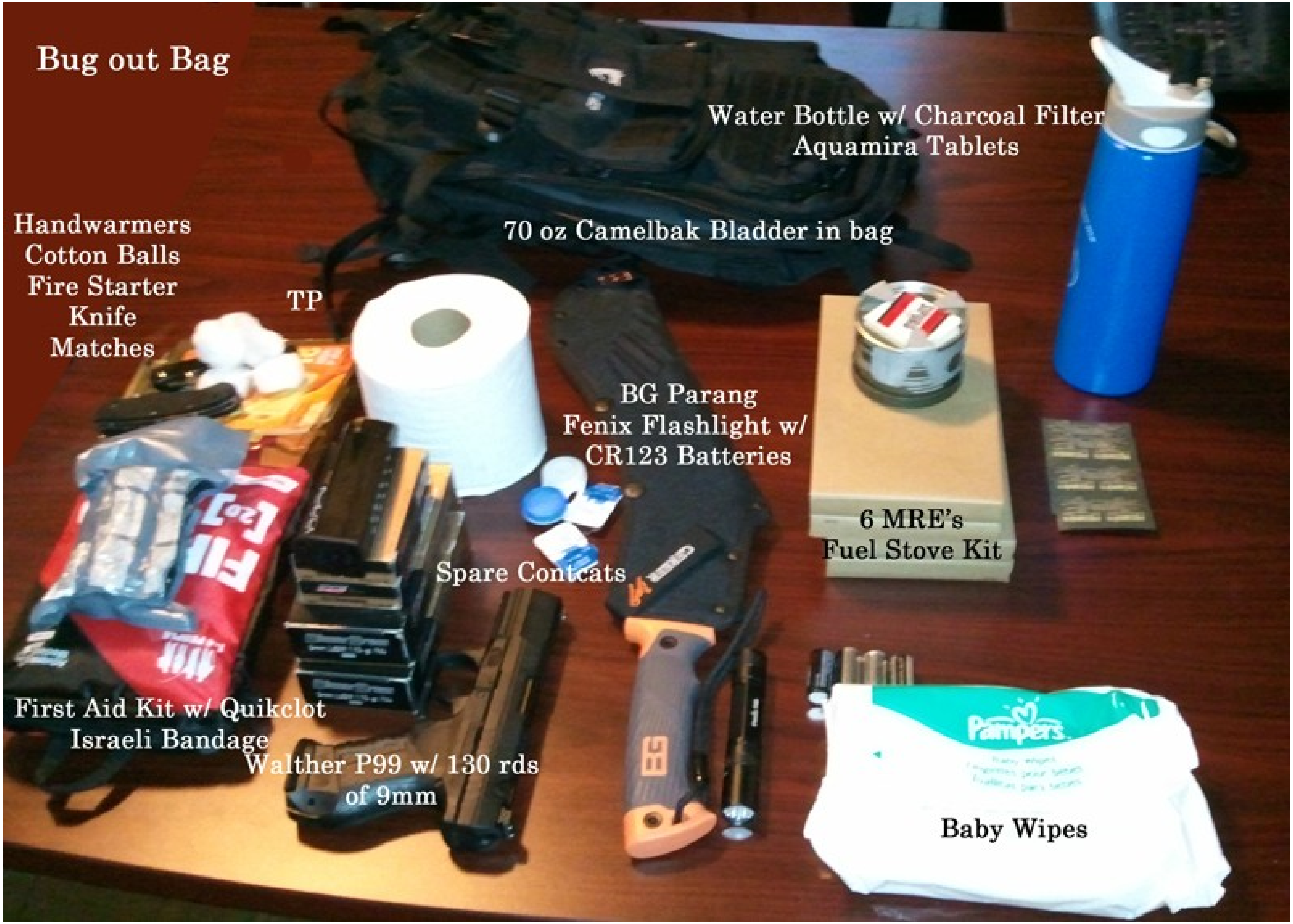

Bug out Bag

Water Bottle w/ Charcoal Filter
Aquamira Tablets

70 oz Camelbak Bladder in bag

Handwarmers
Cotton Balls
Fire Starter
Knife
Matches

TP

BG Parang
Fenix Flashlight w/
CR123 Batteries

6 MRE's
Fuel Stove Kit

Spare Contcats

First Aid Kit w/ Quikclot
Israeli Bandage

Walther P99 w/ 130 rds
of 9mm

Baby Wipes

6 EMERGENCY USES FOR PARACORD

Did you know?

A single length of paracord has been tested to handle 550 lbs of weight.

1. Snare Trap

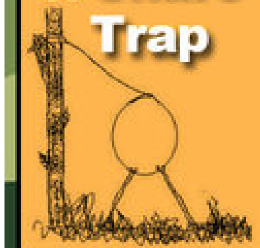

2. Fishing Line

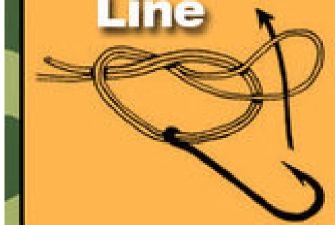

3. Shelter

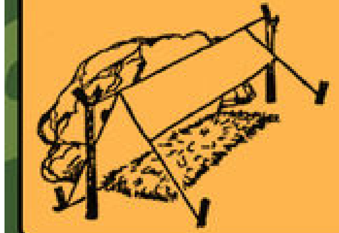

4. Fishing Net

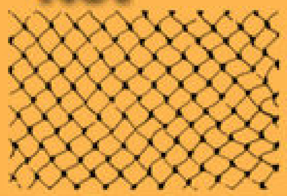

5. Fire

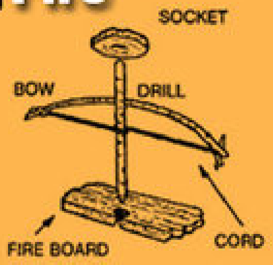
SOCKET
BOW
DRILL
FIRE BOARD
CORD

6. Bear Bag

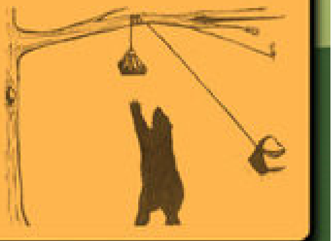

Did you know?

There are seven internal strands, each of which comes apart into two, so there's 14 thin lines.

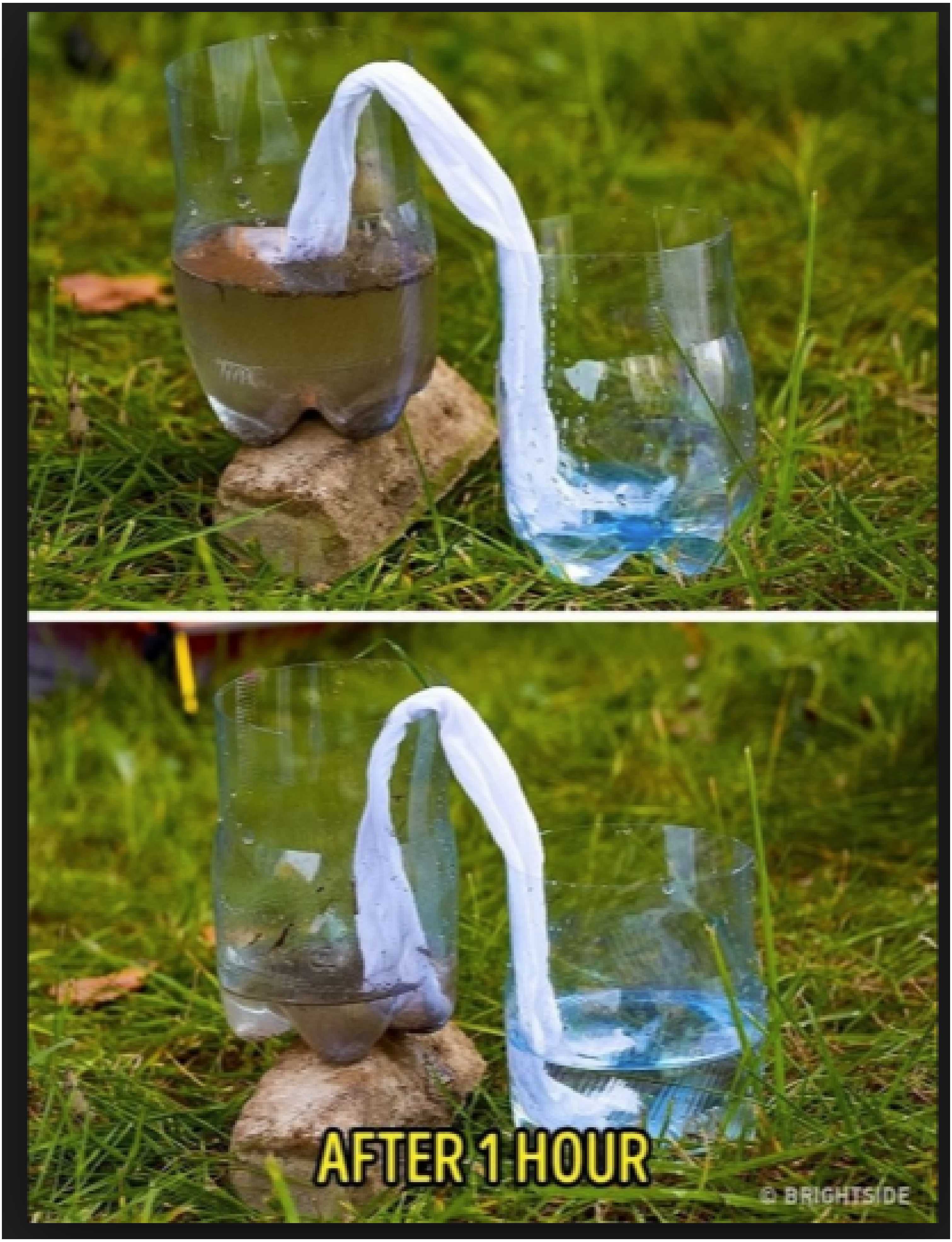
AFTER 1 HOUR
© BRIGHTSIDE

5 WAY TAG STRIP

36 WAY TAG STRIP - TWO ROWS

e wired together ,with a little ingenuity, with the component wires being held together in the grip of solderless crocodile clips, whereby the conne
older.

ex circuits a plastic Terminal Block (sometimes referred to as a choc' or chocolate block) can be utilised very effectively indeed. These are used
15 Amp and 30 Amp. The 5 and 15 Amp Terminal Blocks I have found to be the most suitable. The various component wires can be trappec
makes it easy to change the components around when experiment with different circuits. See The EXPERMENTAL CRYSTAL SET for more det

'CHOCOLATE' TERMINAL BLOCK

Where to connect to the earbud jack.

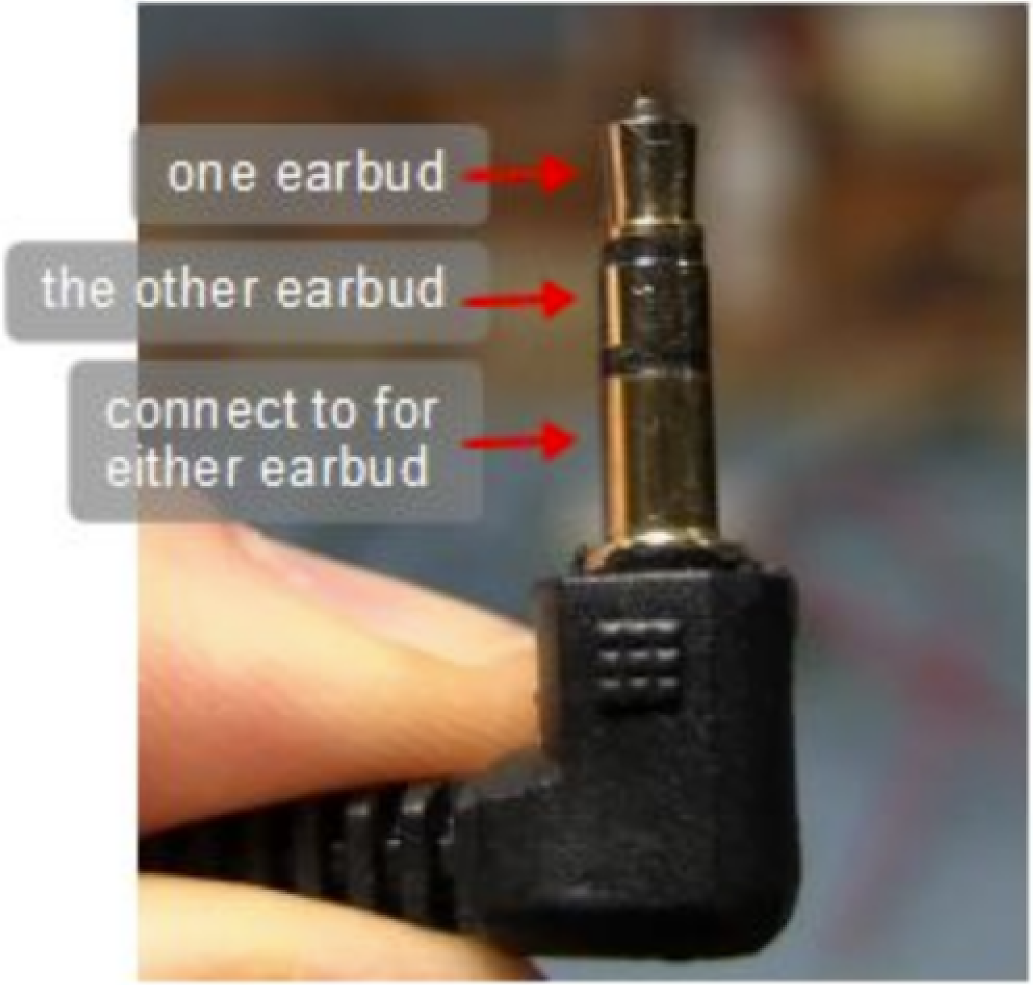
one earbud
the other earbud
connect to for either earbud

For connecting to one earbud only.

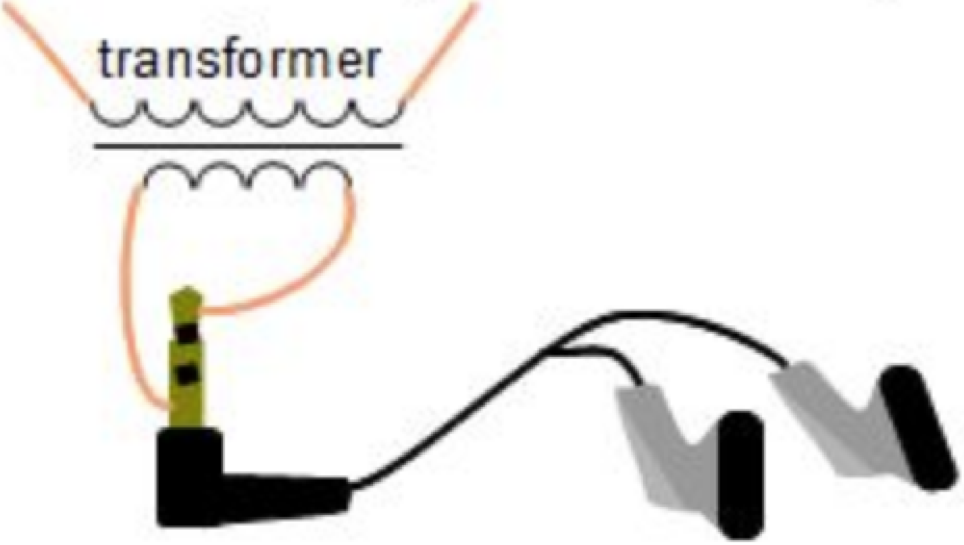
transformer

For connecting to both earbuds.

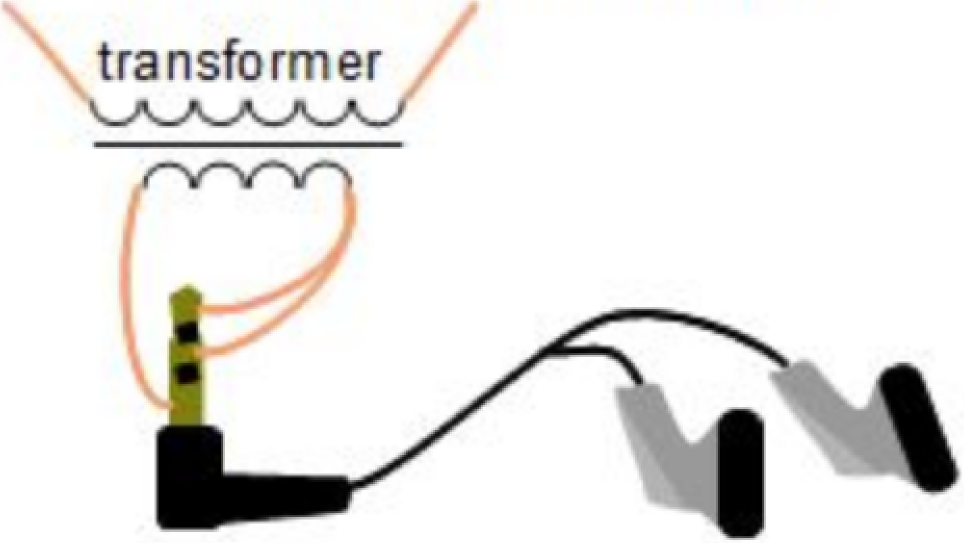
transformer

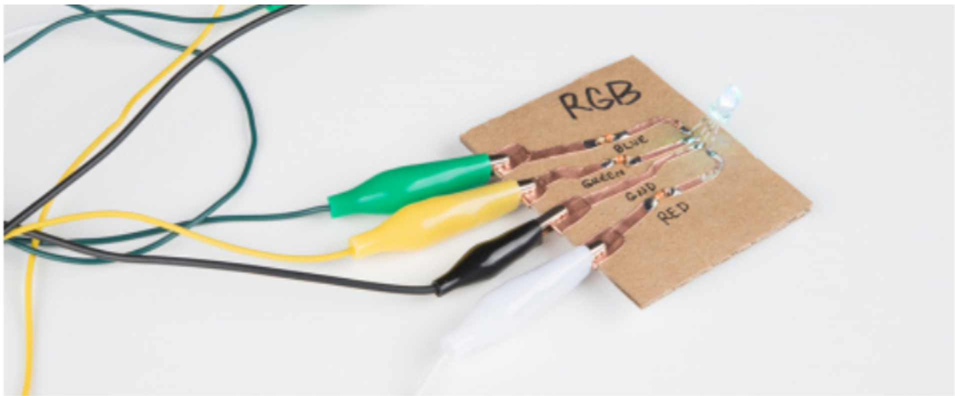
RGB
BLUE
GREEN
GND
RED

The current-limiting resistors are part of the cardboard circuit. The black spots are where I used the Bare Conductive paint

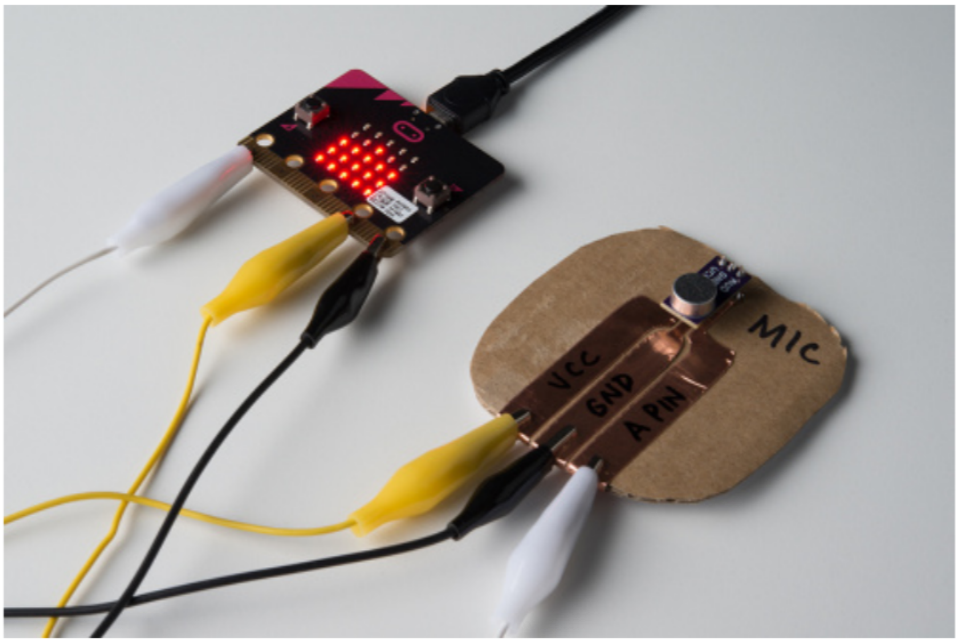
MIC
VCC
GND
A PIN

The micro:bit is plotting the analog values on the 5x5 LED matrix, which represent the noise level in the room.

D PIN
2 BUTTON BOARD
"ACTIVE LOW"
D PIN
GND
BARE
ELECTRIC PAINT

All you need to get started is a roll of copper tape, a few components and conductive paint. One roll of tape and tube of paint can make about 200 circuits.

COSMIC VIBRATIONS

Octave	Manifestation	Vibrations per second	Gaps
1	TOUCH	2	
2		4	
3		8	
4		16	
5	SOUND	32	
6		64	
7		(THESE INTERVALS AND FIGURES ARE APPROXIMATE AND NOT EXACT) 128	
8		256	
9		512	
10		1,024	
11		2,048	
12		4,096	
13		8,192	
14		16,384	
15	ELECTROMAGNETIC (Longer Radio Waves)	32,768	
16		65,536	
17		131,072	
18		262,144	
19		524,288	
20		1,048,576	
21		2,097,152	
22		4,194,304	
23		8,388,608	
24		16,777,216	
25		33,554,432	
26		67,108,864	
27		134,217,728	
28		268,435,456	
29	ELECTROMAGNETIC (Microwaves)	536,870,912	
30		1,073,741,824	
31		2,147,483,648	
32		4,294,967,296	
33		8,589,934,592	
34		17,179,869,184	
35		34,359,738,368	
36		68,719,476,736	
37		137,438,953,472	
38		274,877,906,944	
39	HEAT OR INFRARED	549,755,813,888	
40		1,099,511,627,776	
41		2,199,023,255,552	
42		4,398,046,511,104	
43		8,796,093,022,208	
44		17,592,186,044,416	
45		35,184,372,088,832	
46		70,368,744,177,664	
47		140,737,488,355,328	
48		281,474,976,710,656	
49	VISIBLE LIGHT	562,949,953,421,312	
50		1,125,899,906,842,624	
51	ULTRAVIOLET	2,251,799,813,685,248	SPIRIT ELECTRONS
52		4,503,599,627,370,496	
53		9,007,199,254,740,992	
54		18,014,398,509,481,984	
55		36,028,797,018,963,968	
56		72,057,594,037,927,936	
57	X-RAY	144,115,188,075,855,872	
58		288,230,376,151,711,744	
59		576,460,752,303,423,488	
60			
61			PSYCHIC PROJECTIONS
62	GAMMA AND COSMIC RAYS	4,611,686,018,427,387,904	
63			
64		*The Vibrations in these octaves increase to an incomprehensible number per second.*	
65			
66			
67			
68			
69			
70		*The phenomena of psychic projections and soul essence occur in octaves considerably beyond those of any known material manifestations. Consequently, they must not be considered as immediately following any manifestations which we have shown on this chart.*	
71			
72			
73			SOUL ESSENCE
74			
75			
76			
77			
78			
79			
80			

NOTICE

This chart was carefully compiled by the Editorial and Research Departments of AMORC College after consulting the most eminent scientific authorities on the subjects of physics and chemistry and comparing their observations and notations with the reports from the Rosicrucian laboratories in foreign countries.

There is a slight disagreement in regard to the rates of vibrations for the various keys of the musical keyboard, due to different standards of pitch in various countries, and there is some disagreement among scientific authorities in regard to the colors associated with various musical notes. This chart represents not only the consensus of most of these authorities but also adheres strictly to what the Rosicrucians have been able to demonstrate as truth in our own laboratories and in the experiments of thousands of members in various parts of the world. Regardless, therefore, of any disagreement with authorities outside the organization, this chart is the authentic presentation of the Rosicrucian knowledge.

This chart is to be used by the members of the Rosicrucian Order in connection with their studies and experiments and is not to be used for any other purpose or in any other way. Like the lessons themselves, this chart and the facts contained on it are reserved and restricted in use under the rules and regulations of the organization.

INTERNATIONAL
SEAL OF THE
SUPREME LODGE
of the
Rosicrucian
Order

THE SOLAR SPECTRUM

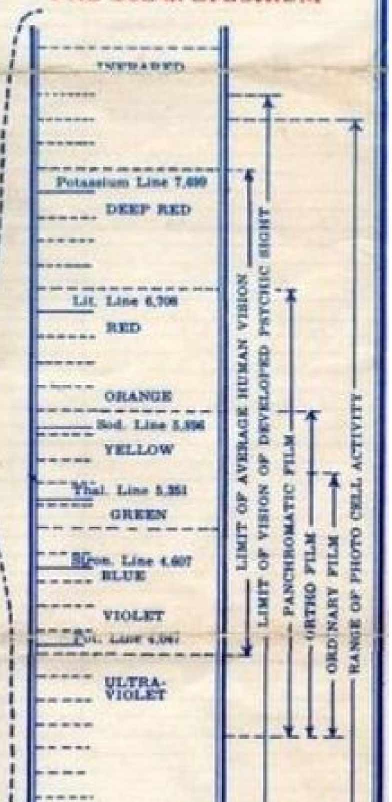

AMORC
THE
Rosicrucian Order
San Jose, California

LITHO. IN U. S. A.

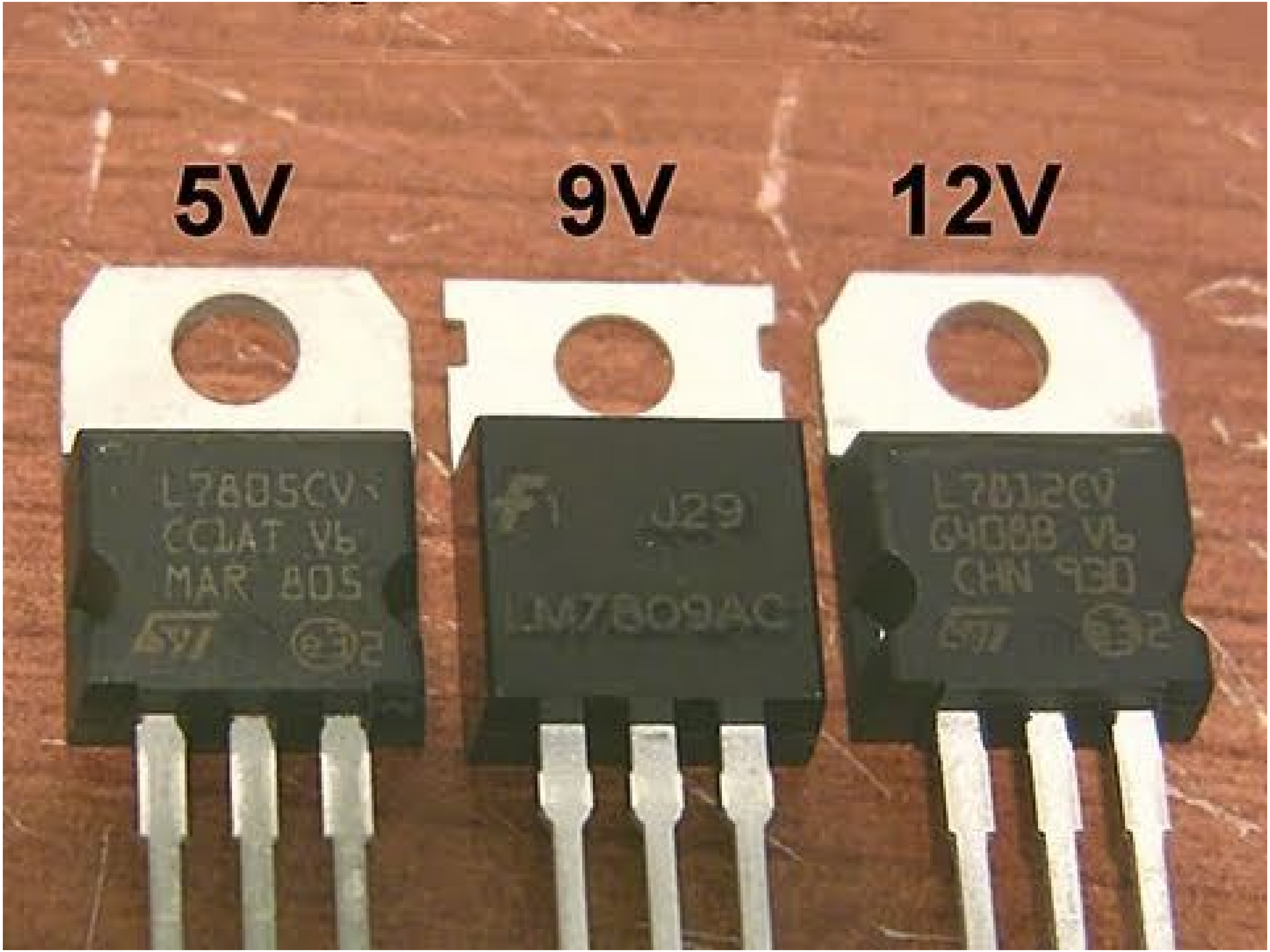
5V
9V
12V
L7805CV
CC1AT V6
MAR 805
J29
LM7809AC
L7812CV
G408B V6
CHN 930

7 LOGIC GATES & 4 USEFUL COMBINATIONS

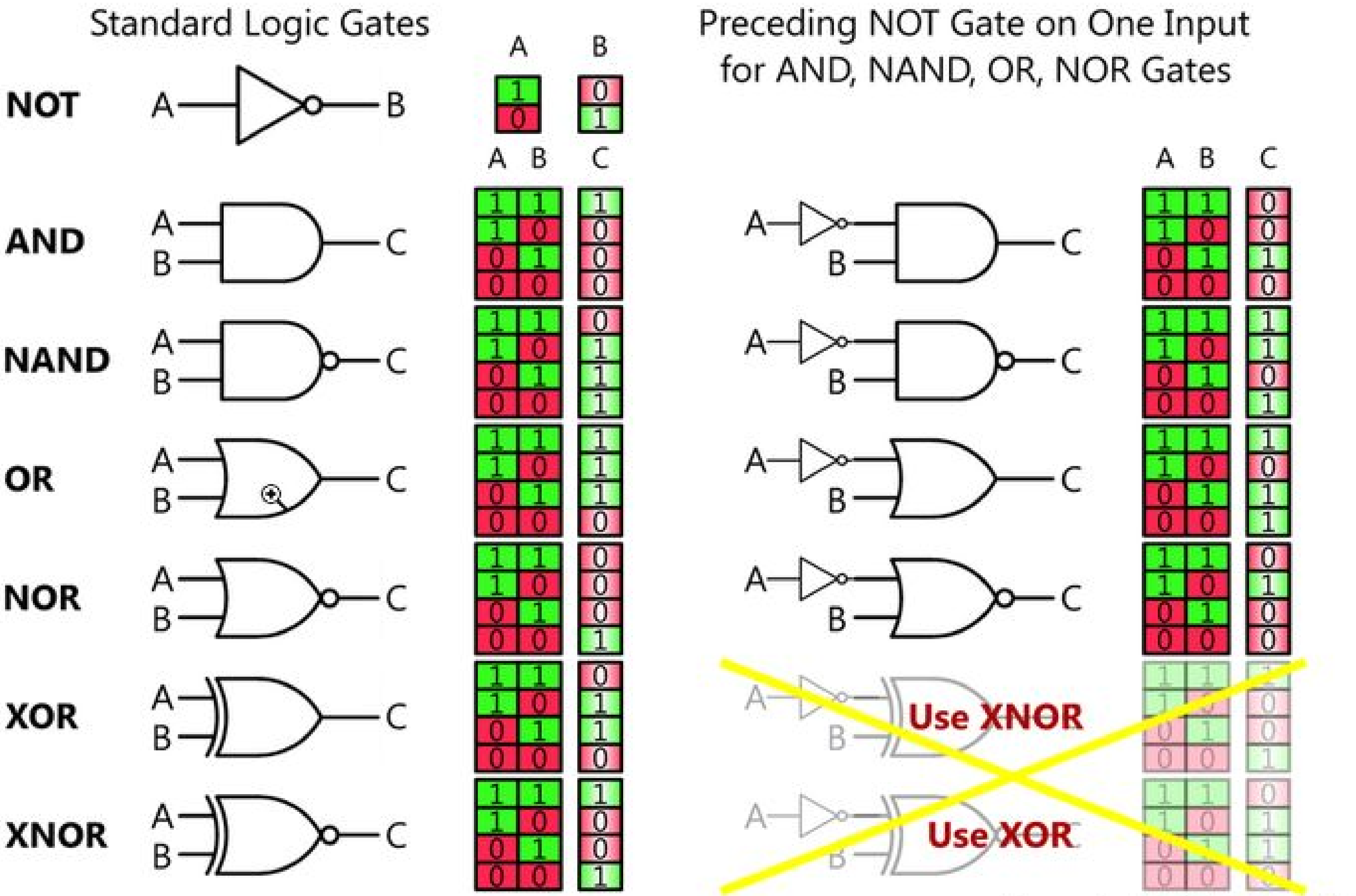

Standard Logic Gates

Preceding NOT Gate on One Input
for AND, NAND, OR, NOR Gates

	A	B
NOT	1	0
	0	1

	A	B	C	A	B	C
AND	1	1	1	1	1	0
	1	0	0	1	0	0
	0	1	0	0	1	1
	0	0	0	0	0	0
NAND	1	1	0	1	1	1
	1	0	1	1	0	1
	0	1	1	0	1	0
	0	0	1	0	0	1
OR	1	1	1	1	1	1
	1	0	1	1	0	0
	0	1	1	0	1	1
	0	0	0	0	0	1
NOR	1	1	0	1	1	0
	1	0	0	1	0	1
	0	1	0	0	1	0
	0	0	1	0	0	0
XOR	1	1	0	~~1~~	~~1~~	~~1~~
	1	0	1	~~1~~	~~0~~	~~0~~
	0	1	1	~~0~~	~~1~~	~~0~~
	0	0	0	~~0~~	~~0~~	~~1~~
XNOR	1	1	1	~~1~~	~~1~~	~~0~~
	1	0	0	~~1~~	~~0~~	~~1~~
	0	1	0	~~0~~	~~1~~	~~1~~
	0	0	1	~~0~~	~~0~~	~~0~~

Copyright 2012 PicoKit

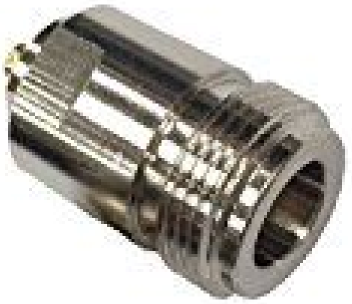

TYPE N F

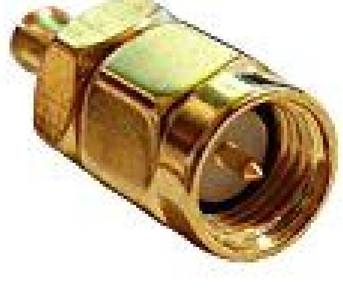

SMA M

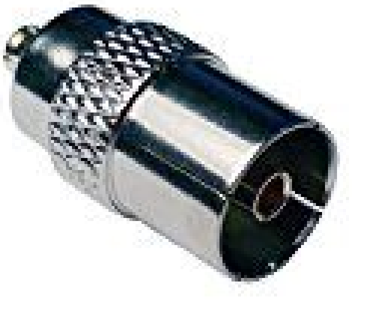

PAL/BELLING LEE F

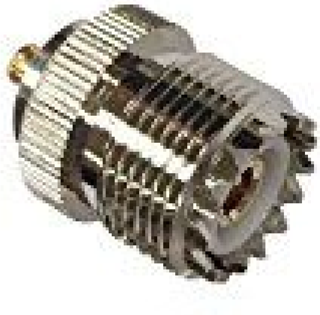

UHF F

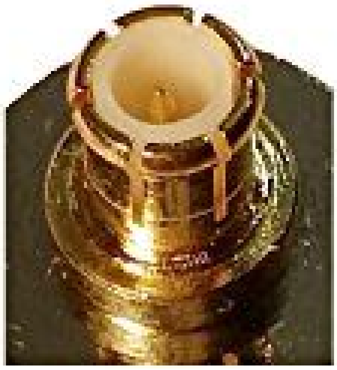

MCX MALE
(OTHER END)

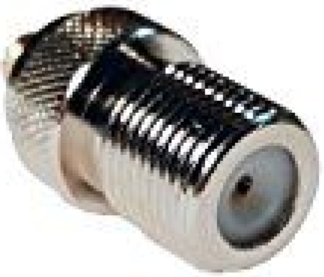

TYPE F F

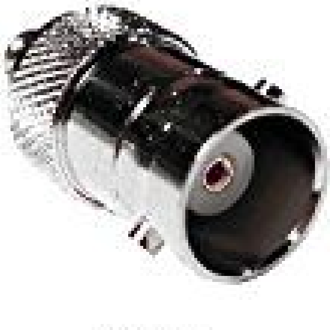

BNC F

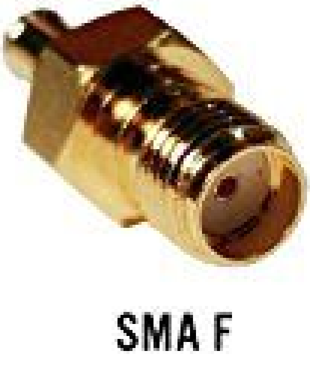

SMA F

TYPE N F

SMA M

PAL/BELLING LEE F

UHF F

MCX MALE
(OTHER END)

TYPE F F

BNC F

SMA F

DIODE

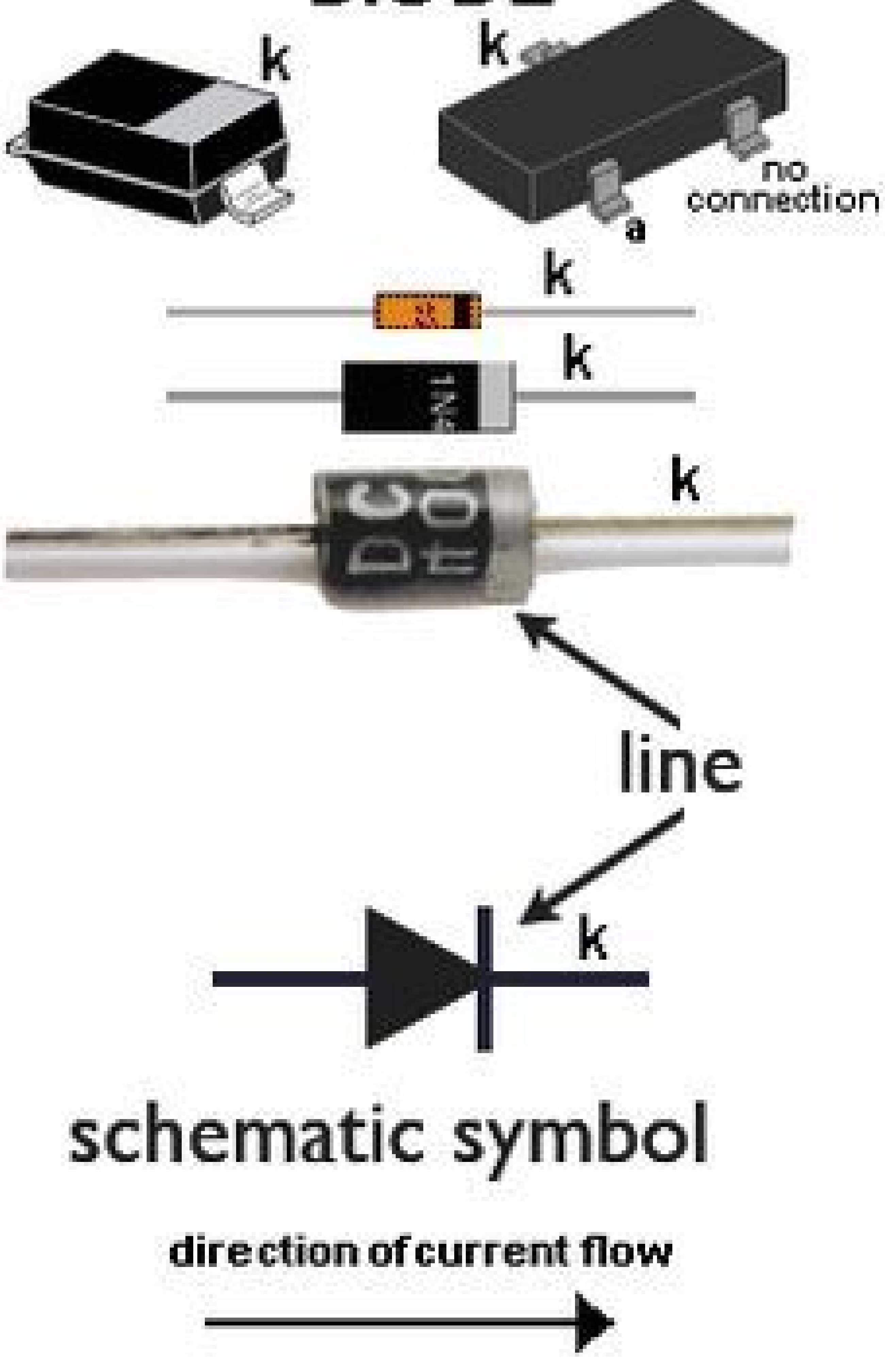
k
k
no connection
a
k
k
k
line
k
schematic symbol
direction of current flow

7. 555 Timer Projects

In this section we will make :

- 7.1 4 Key Piano
- 7.2 Light sensitive music circuit
- 7.3 Light controlled Police Siren
- 7.4 Touch Switch
- 7.5 Timer
- 7.6 Continuity Tester
- 7.7 Knight Rider
- 7.8 Cricket Game
- 7.9 Multipurpose circuit
- 7.10 Johnson counter

Some Parts of this project are available in Video CD ROM

Internal diagram of 555

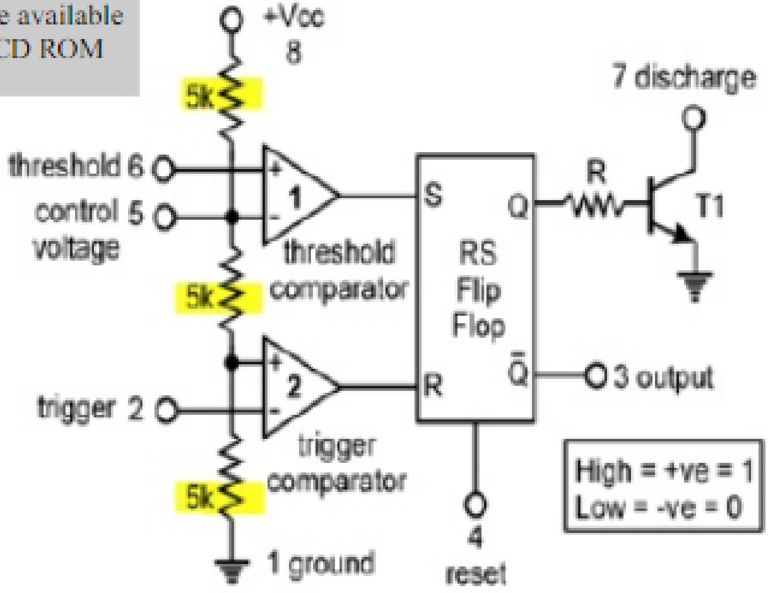
+Vcc
8
5k
7 discharge
threshold 6
control 5
voltage
1
threshold
comparator
S
Q
R
T1
RS
Flip
Flop
5k
2
R
Q̄
3 output
trigger 2
trigger
comparator
5k
High = +ve = 1
Low = -ve = 0
1 ground
4
reset

What is it about ?

555 timer integrated circuit (IC) is a very popular chip used in variety of applications like timer, pulse generation and oscillators. This is a low cost, stable and widely available chip which makes it favorite for hobbyists. The internal components of 555 as shown in figure consists of 2 comparators and a flip flop. All of these components contain 25 transistors and 15 resistors packed in the IC.

The three highlighted 5k resistors shown in figure are the reason why this IC is named as 555.

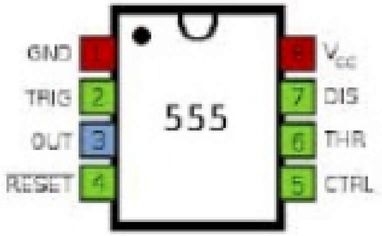
GND 1
TRIG 2
OUT 3
RESET 4
555
8 Vcc
7 DIS
6 THR
5 CTRL

Please note the notch near first pin. This is made to indentify the first pin of IC.

Pin	Name	Purpose
1	GND	Ground reference voltage, low level (0 V)
2	TRIG	The OUT pin goes high and a timing interval starts when this input falls below 1/2 of CTRL voltage (which is typically 1/3 of V_{CC}, when CTRL is open).
3	OUT	This output is driven to approximately 1.7V below $+V_{CC}$ or GND.
4	RESET	A timing interval may be reset by driving this input to GND, but the timing does not begin again until RESET rises above approximately 0.7 volts. Overrides TRIG which overrides THR.
5	CTRL	Provides "control" access to the internal voltage divider (by default, 2/3 V_{CC}).
6	THR	The timing (OUT high) interval ends when the voltage at THR is greater than that at CTRL.
7	DIS	Open collector output which may discharge a capacitor between intervals. In phase with output.
8	V_{CC}	Positive supply voltage, which is usually between 3 and 15 V depending on the variation.

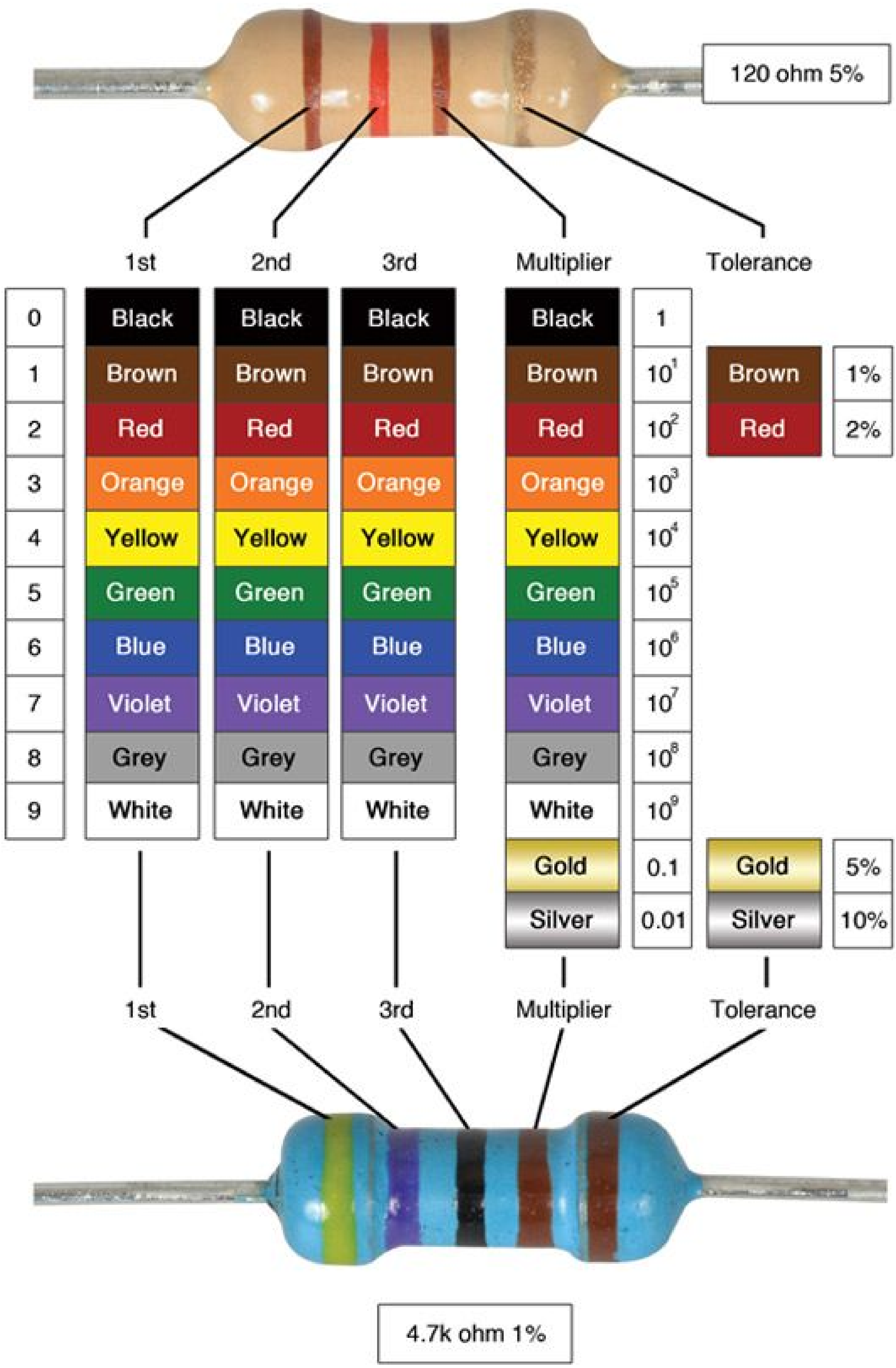
120 ohm 5%
1st
2nd
3rd
Multiplier
Tolerance
4.7k ohm 1%

	1st	2nd	3rd	Multiplier		Tolerance	
0	Black	Black	Black	Black	1		
1	Brown	Brown	Brown	Brown	10^1	Brown	1%
2	Red	Red	Red	Red	10^2	Red	2%
3	Orange	Orange	Orange	Orange	10^3		
4	Yellow	Yellow	Yellow	Yellow	10^4		
5	Green	Green	Green	Green	10^5		
6	Blue	Blue	Blue	Blue	10^6		
7	Violet	Violet	Violet	Violet	10^7		
8	Grey	Grey	Grey	Grey	10^8		
9	White	White	White	White	10^9		
				Gold	0.1	Gold	5%
				Silver	0.01	Silver	10%

ELECTROMAGNETIC RADIATION SPECTRUM

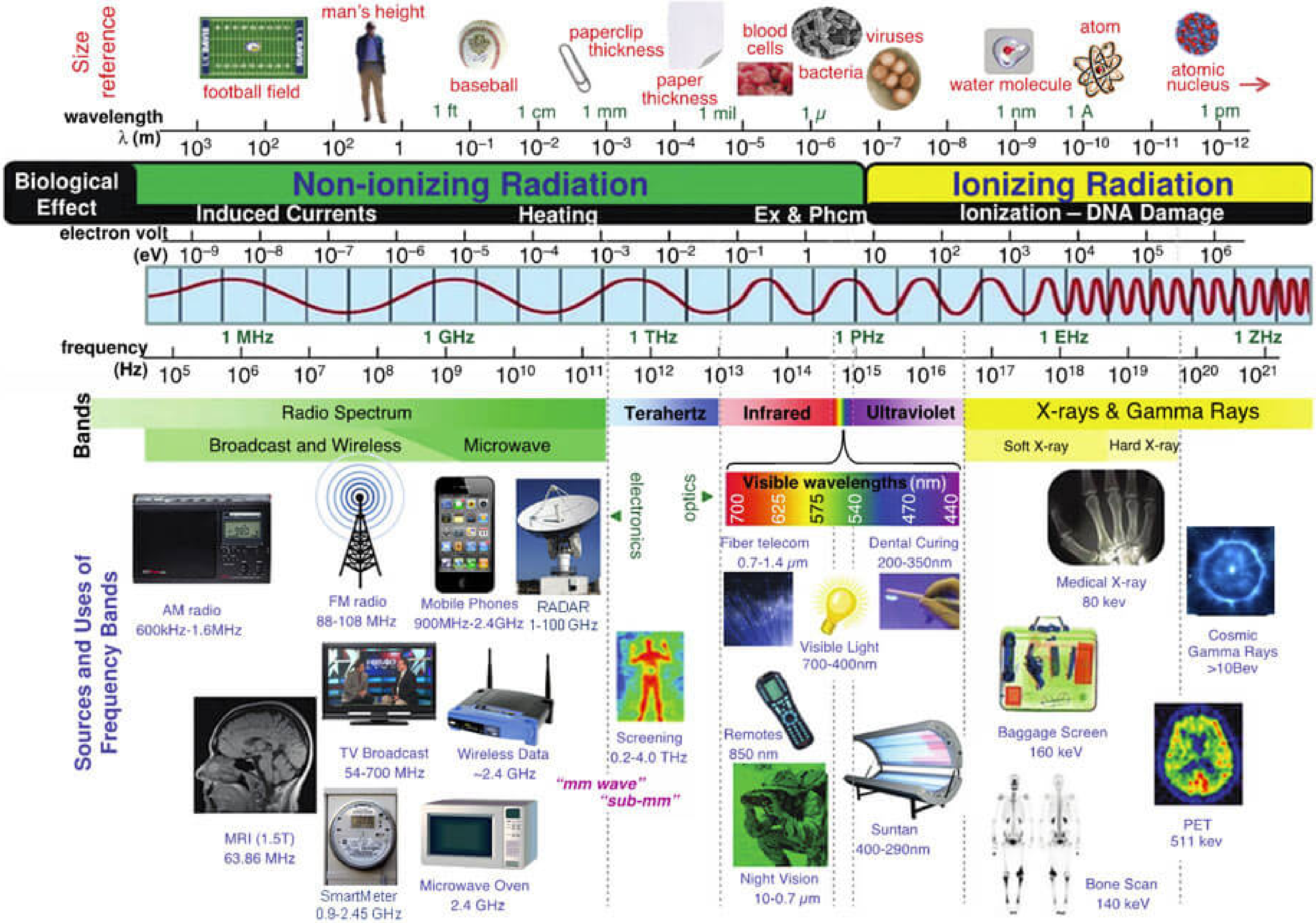

Size reference

football field

man's height

baseball

paperclip thickness

paper thickness

blood cells

viruses

bacteria

water molecule

atom

atomic nucleus →

1 ft

1 cm

1 mm

1 mil

1 μ

1 nm

1 A

1 pm

wavelength λ (m)

10^3 10^2 10^2 1 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 10^{-6} 10^{-7} 10^{-8} 10^{-9} 10^{-10} 10^{-11} 10^{-12}

Biological Effect

Non-ionizing Radiation

Ionizing Radiation

Induced Currents

Heating

Ex & Phcm

Ionization – DNA Damage

electron volt (eV)

10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10 10^2 10^3 10^4 10^5 10^6

1 MHz

1 GHz

1 THz

1 PHz

1 EHz

1 ZHz

frequency (Hz)

10^5 10^6 10^7 10^8 10^9 10^{10} 10^{11} 10^{12} 10^{13} 10^{14} 10^{15} 10^{16} 10^{17} 10^{18} 10^{19} 10^{20} 10^{21}

Bands

Radio Spectrum

Broadcast and Wireless

Microwave

Terahertz

Infrared

Ultraviolet

X-rays & Gamma Rays

Soft X-ray

Hard X-ray

◄ electronics

optics ►

Visible wavelengths (nm)

700 625 575 540 470 440

Sources and Uses of Frequency Bands

AM radio
600kHz-1.6MHz

FM radio
88-108 MHz

Mobile Phones
900MHz-2.4GHz

RADAR
1-100 GHz

MRI (1.5T)
63.86 MHz

TV Broadcast
54-700 MHz

Wireless Data
~2.4 GHz

SmartMeter
0.9-2.45 GHz

Microwave Oven
2.4 GHz

"mm wave"

"sub-mm"

Screening
0.2-4.0 THz

Fiber telecom
0.7-1.4 μm

Dental Curing
200-350nm

Visible Light
700-400nm

Remotes
850 nm

Night Vision
10-0.7 μm

Suntan
400-290nm

Medical X-ray
80 kev

Cosmic Gamma Rays
>10Bev

Baggage Screen
160 keV

PET
511 kev

Bone Scan
140 keV

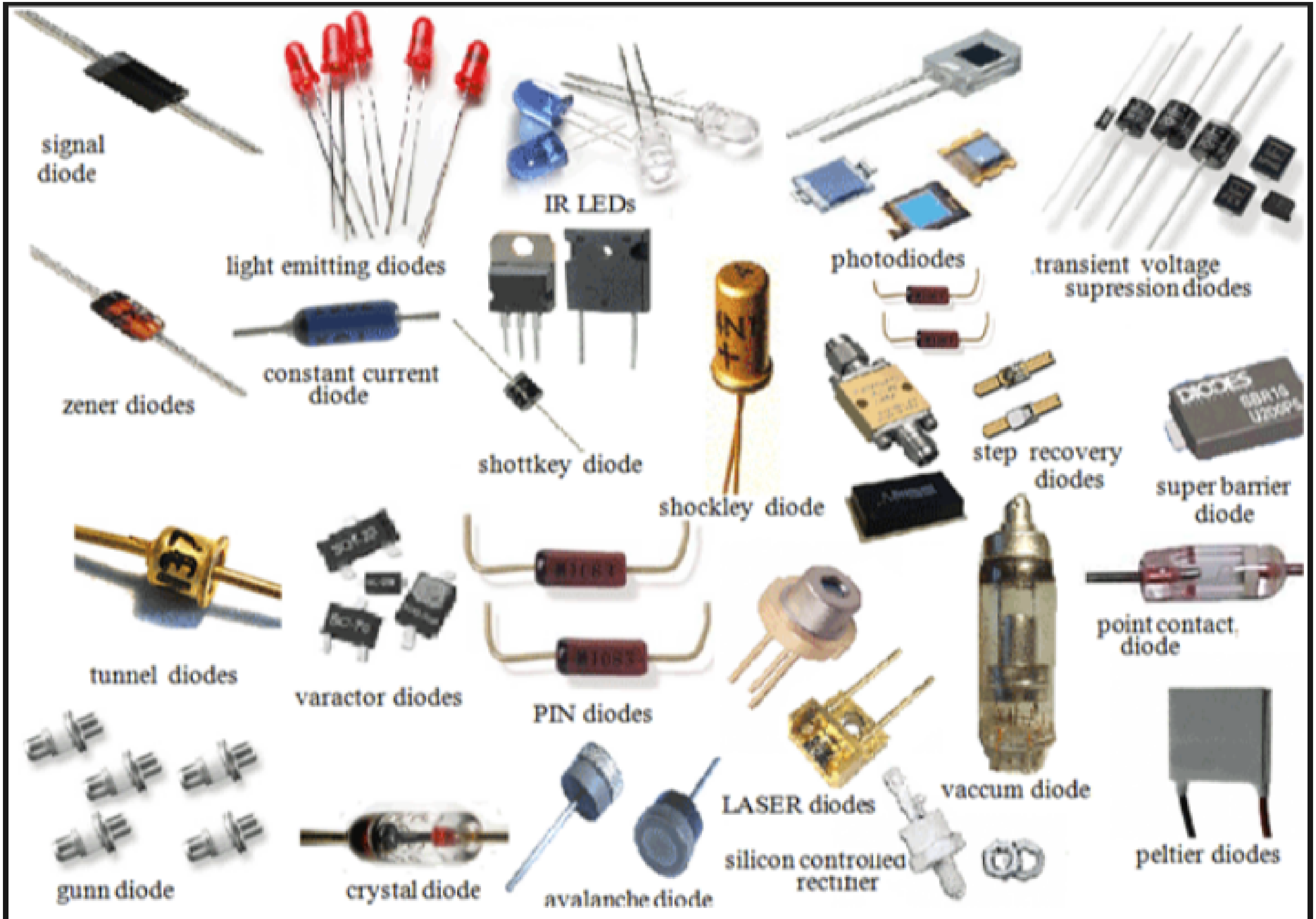
signal diode
light emitting diodes
IR LEDs
photodiodes
transient voltage supression diodes
zener diodes
constant current diode
shottkey diode
shockley diode
step recovery diodes
super barrier diode
tunnel diodes
varactor diodes
PIN diodes
point contact diode
LASER diodes
vaccum diode
peltier diodes
silicon controlled rectifier
gunn diode
crystal diode
avalanche diode

Types of Diode

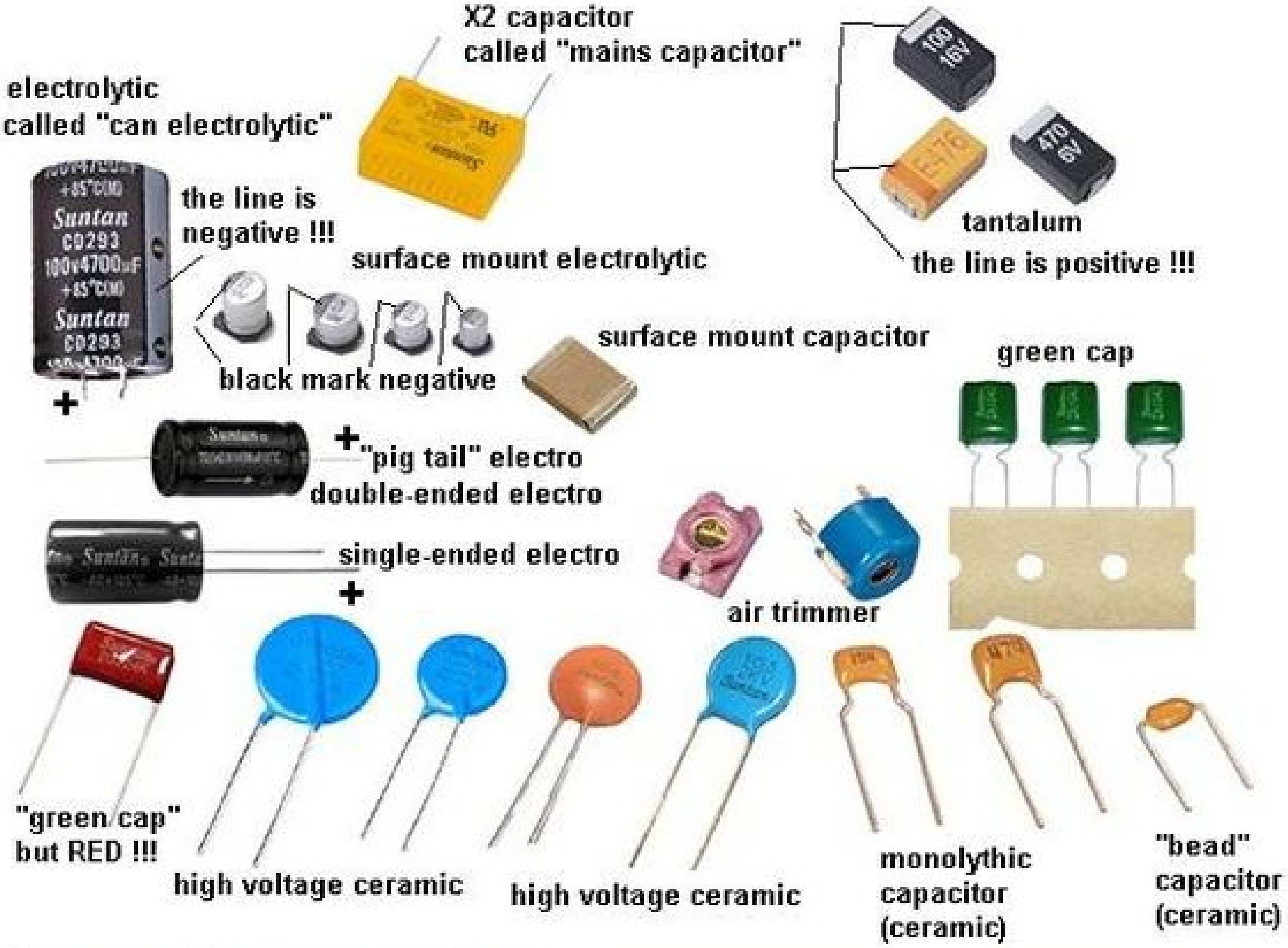
X2 capacitor
called "mains capacitor"
electrolytic
called "can electrolytic"
the line is
negative !!!
surface mount electrolytic
tantalum
the line is positive !!!
black mark negative
surface mount capacitor
green cap
+
+"pig tail" electro
double-ended electro
single-ended electro
+
air trimmer
"green cap"
but RED !!!
high voltage ceramic
high voltage ceramic
monolythic
capacitor
(ceramic)
"bead"
capacitor
(ceramic)

tantalum capacitors can
"go up in smoke"

"bead" tantalum

Electronics
Reference Sheet v1.1b

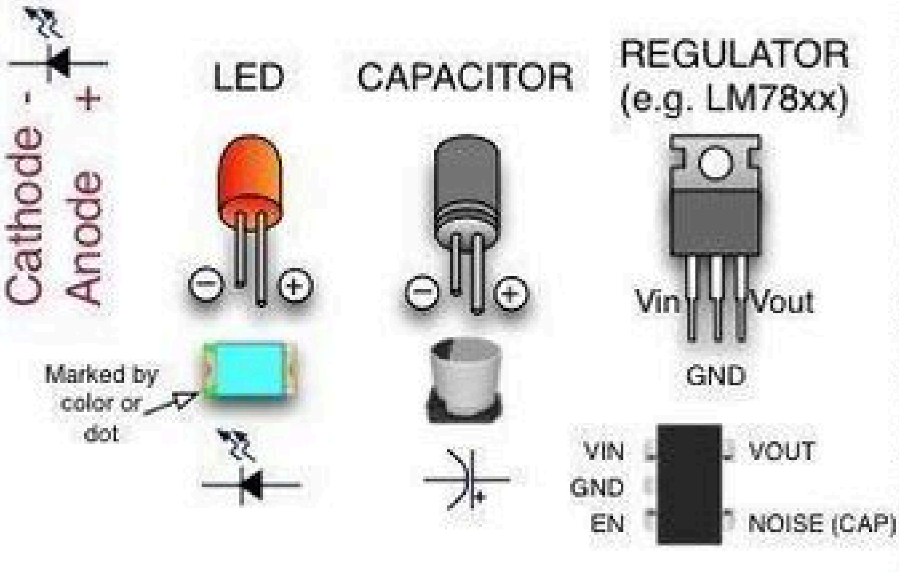
Cathode -
Anode +
LED
Marked by color or dot
CAPACITOR
REGULATOR
(e.g. LM78xx)
Vin
Vout
GND
VIN
GND
EN
VOUT
NOISE (CAP)

Resistor

		Digit	Multiplier	Tolerance
Silver	⬜	-	0.01	±10%
Gold	🟨	-	0.1	±5%
Black	⬛	0	1	-
Brown	🟫	1	10	±1%
Red	🟥	2	100	±2%
Orange	🟧	3	1k	-
Yellow	🟨	4	10k	-
Green	🟩	5	100k	±0.5%
Blue	🟦	6	1M	±0.25%
Violet	🟪	7	10M	±0.1%
Gray	⬜	8	-	-
White	⬜	9	-	-

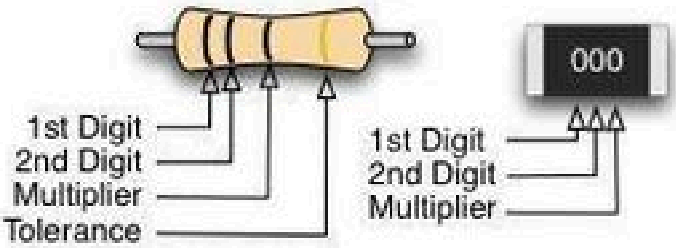
1st Digit
2nd Digit
Multiplier
Tolerance
000
1st Digit
2nd Digit
Multiplier

DIODE

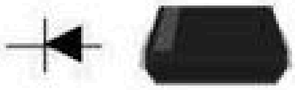

NPN transistor (Current sink)
(e.g. PN2222)

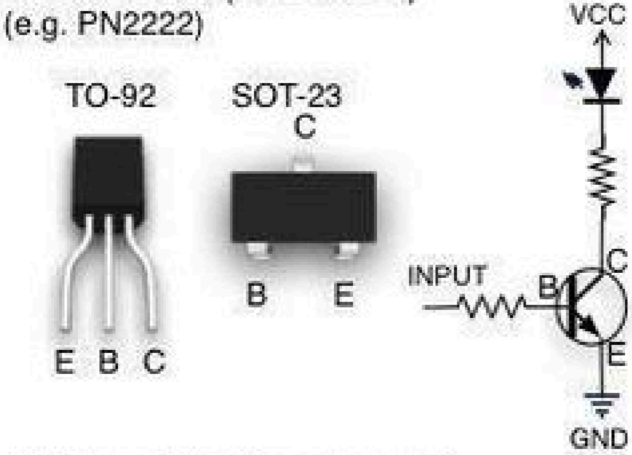
TO-92
E B C
SOT-23
C
B E
VCC
INPUT
B
C
E
GND

N-channel MOSFET

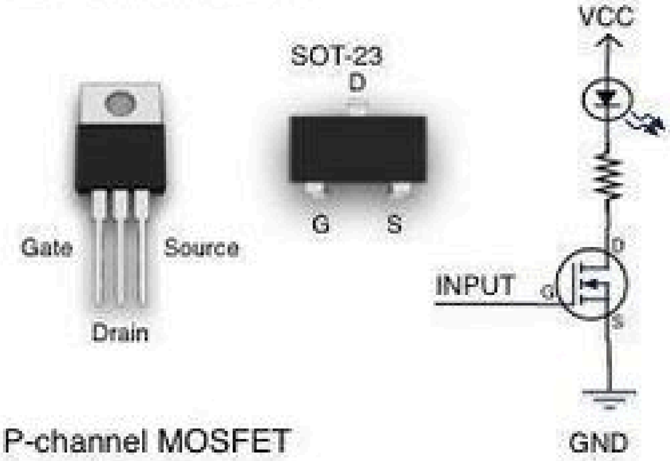
Gate
Source
Drain
SOT-23
D
G S
VCC
INPUT
G
D
S
GND

PNP transistor (Current source)
(e.g. PN2907)

P-channel MOSFET

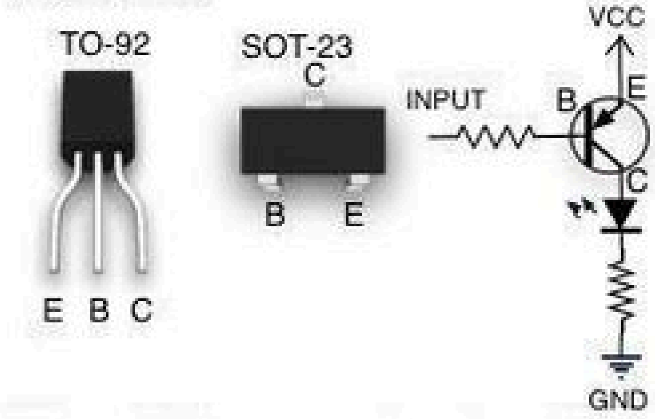
TO-92
E B C
SOT-23
C
B E
INPUT
B
E
C
VCC
GND

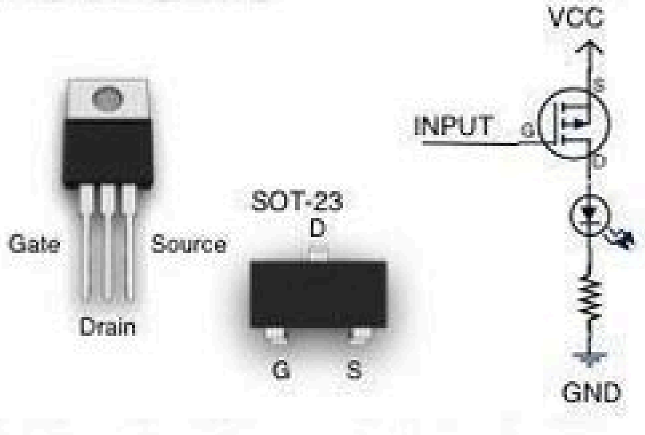
Gate
Source
Drain
SOT-23
D
G S
VCC
INPUT
G
S
D
GND

* Please note that some components may have a different pinout than the one showed above, you should always check the data sheet before using a new component.

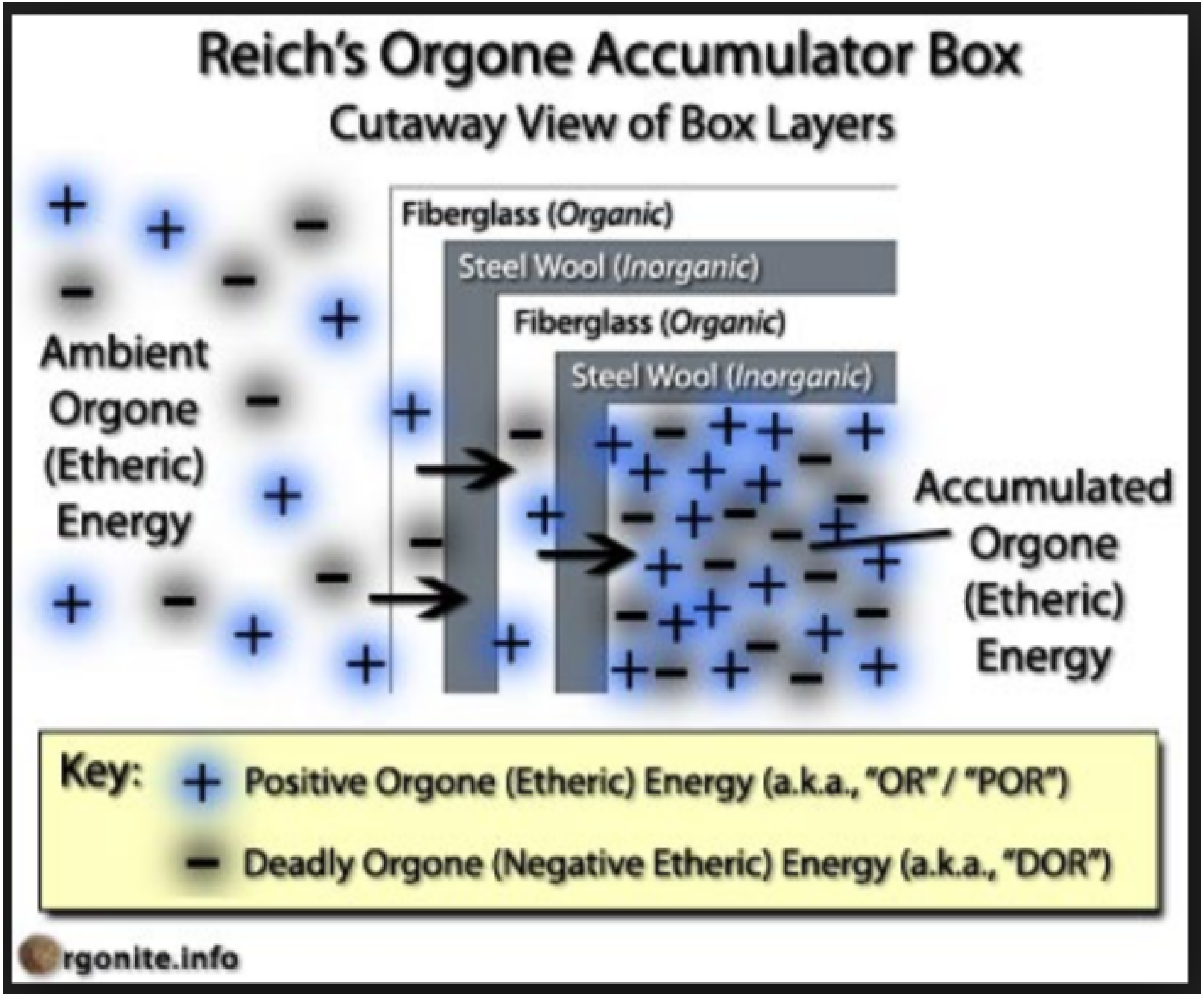
Reich's Orgone Accumulator Box
Cutaway View of Box Layers
Fiberglass (*Organic*)
Steel Wool (*Inorganic*)
Fiberglass (*Organic*)
Steel Wool (*Inorganic*)
Ambient Orgone (Etheric) Energy
Accumulated Orgone (Etheric) Energy
Key:
+ Positive Orgone (Etheric) Energy (a.k.a., "OR" / "POR")
− Deadly Orgone (Negative Etheric) Energy (a.k.a., "DOR")
rgonite.info

How to build a Don Croft ORGONITE CLOUD BUSTER

Materials List:

Wear safety gear as necessary
- respirator, gloves when pouring resin.
- gloves & goggles when cutting wood - with a respirator if cutting MDF etc.

4 x 10' lengths of copper (or other) pipe (cut 3 lengths in half, and one length into six 1' pieces)

6 pipe endcaps and 6 pipe connectors

2 gallon bucket (polypropylene if possible)

10 double terminated quartz (6 of these must fit into garden hose for insertion into base of pipes)

6 pieces of garden hose (3/4" - 1" long)

Resin and metal suitable for orgonite

A roll of tape (duct tape etc)

Glue (silicone glue will work)

Stick for mixing metal into resin when pouring layers

Top, middle (temporary) and bottom spacers cut and drilled out of wood. If not buying readymade, use templates or measure with copmass, ruler, pencil.
Drill the holes with 1 1/8" and 1 1/4" wood drill bits, and cut into shape with a Jigsaw.

Refer to (Part 1 - 2) How to Make an Orgonite CloudBuster:
http://www.youtube.com/watch?v=cCn1DO2Rj6k

Third - The temporary spacer is removed and the pipe-tops are taped up to stop resin or metal getting into them.
Further layers are poured, allowing resin to cure and cool between pouring.
More crystals or other components may be added with the basic orgonite mixture.
When done, the tape is removed from the pipe ends.

Second poured layer
(the temporary spacer is fixed in place while this layer is curing, and then removed)

First poured layer
(1" - 1 1/2" high) Include four double-terminated crystals at compass points

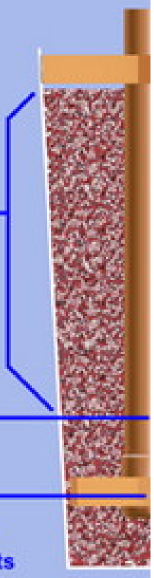

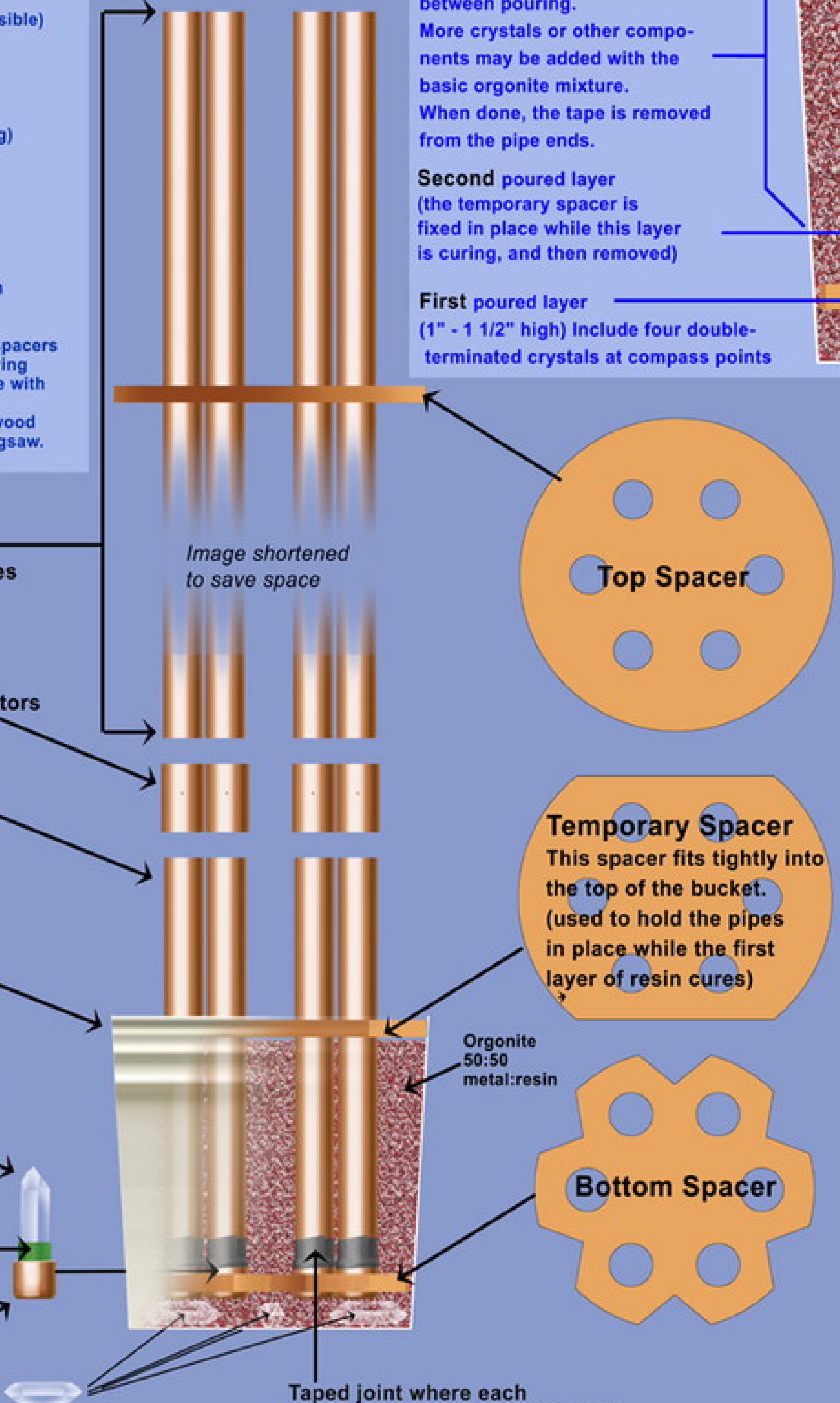

Six 5' lengths of copper pipe
When the orgonite base of the cloud buster is complete, fix the top pipes to the bottom ones using pipe connectors, then fix the top spacer in place to provide stability.

Pipe connectors

Six 1' lengths of copper pipe
(1" diameter pipe)

2 Gallon ......plastic bucket
(8" bottom-interior diameter, 9" top interior diameter)

Temporary Spacer
This spacer fits tightly into the top of the bucket. (used to hold the pipes in place while the first layer of resin cures)

Double Treminated Crystal
(secured into hose with glue)

Section of garden hose 3/4" - 1" piece (glued into end cap with silicone glue)

Copper end cap

Four double terminated crystals in the base are aligned with the cardinal directions

Taped joint where each copper pipe fits into its end cap (duct tape or any kind to stop resin seeping into endcap)

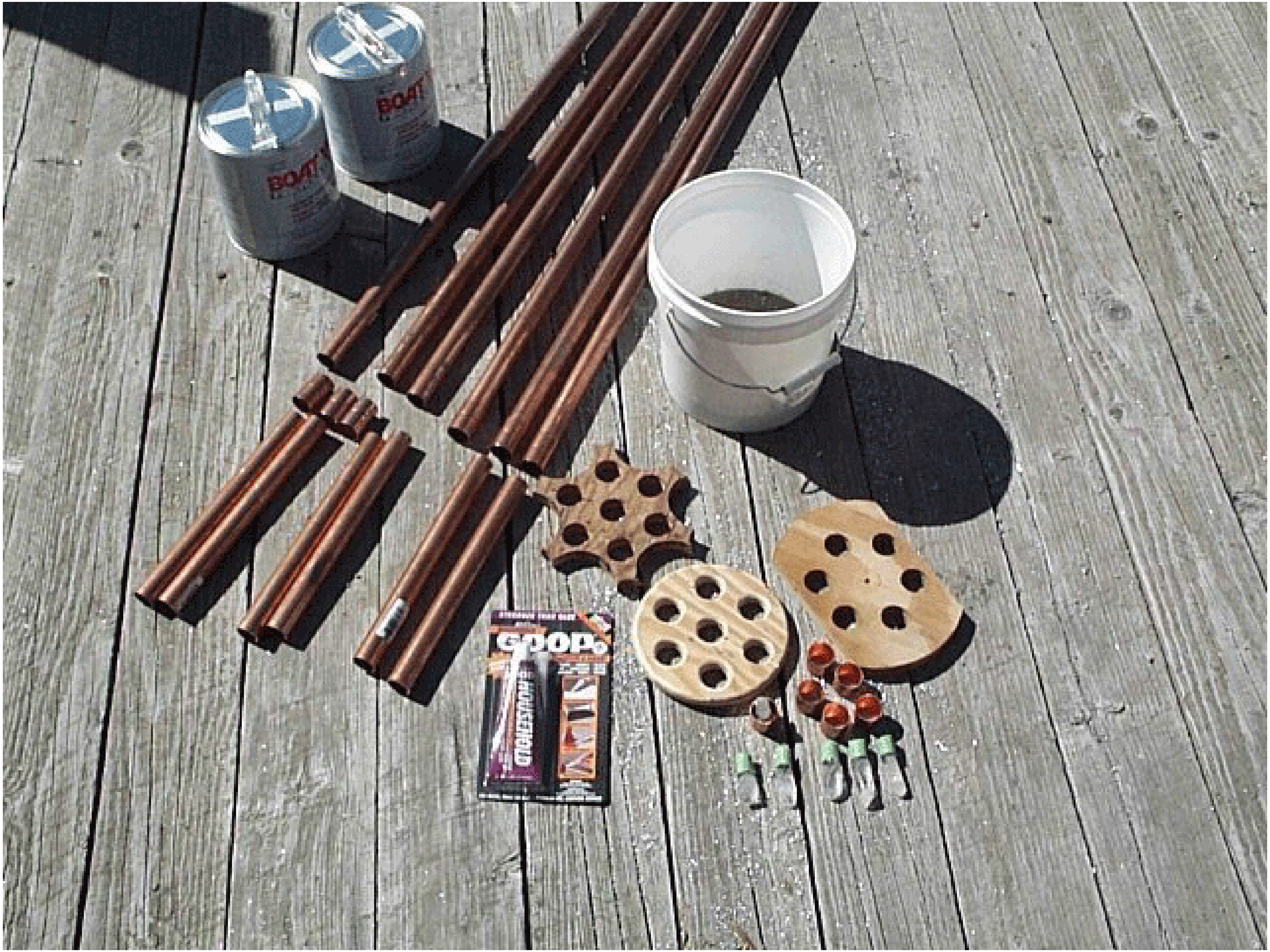
BOAT
BOAT
GOOP
HOUSEHOLD

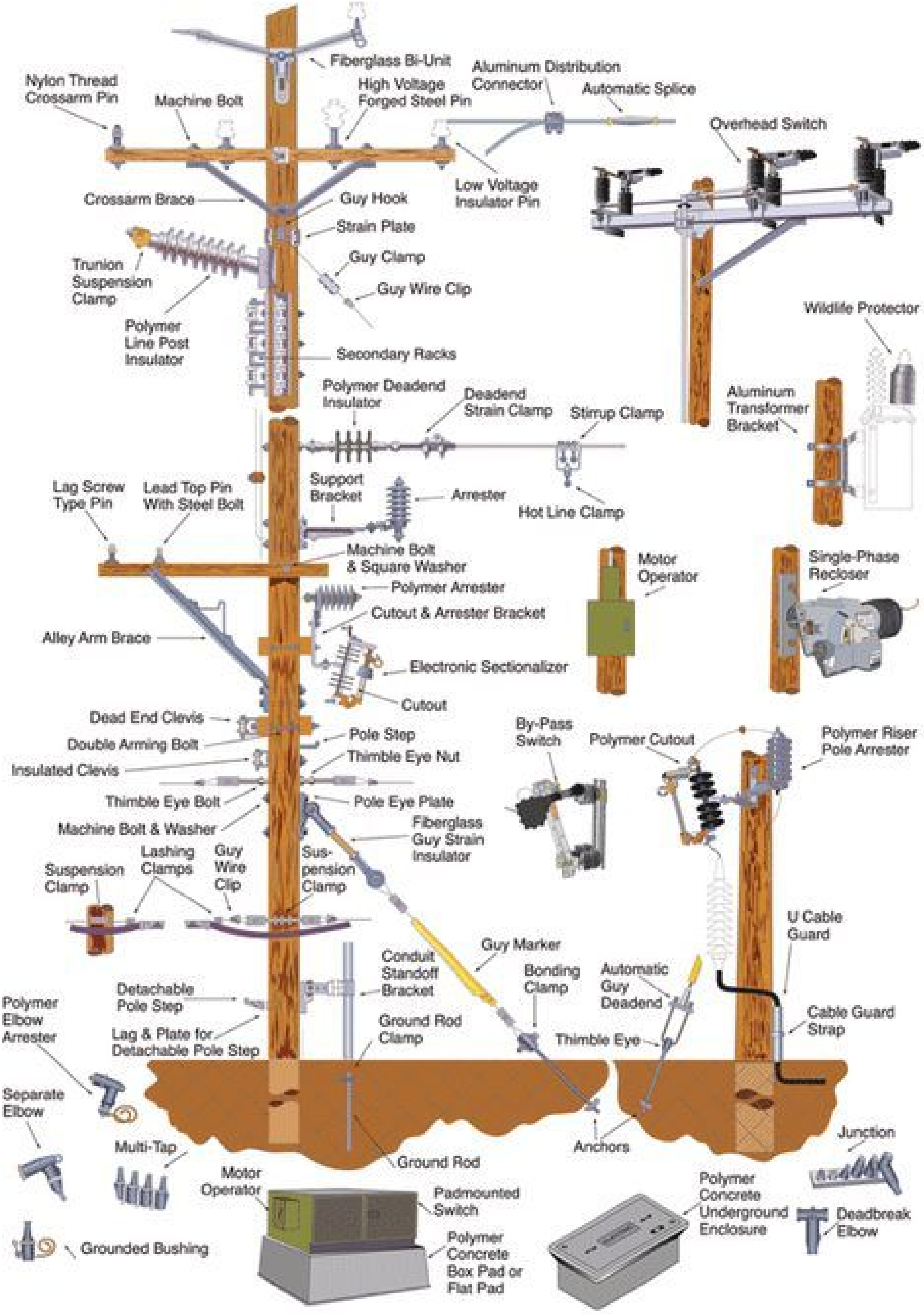
Fiberglass Bi-Unit
Nylon Thread Crossarm Pin
Machine Bolt
High Voltage Forged Steel Pin
Aluminum Distribution Connector
Automatic Splice
Overhead Switch
Low Voltage Insulator Pin
Crossarm Brace
Guy Hook
Strain Plate
Trunion Suspension Clamp
Guy Clamp
Guy Wire Clip
Polymer Line Post Insulator
Secondary Racks
Wildlife Protector
Polymer Deadend Insulator
Deadend Strain Clamp
Stirrup Clamp
Aluminum Transformer Bracket
Lag Screw Type Pin
Lead Top Pin With Steel Bolt
Support Bracket
Arrester
Hot Line Clamp
Machine Bolt & Square Washer
Motor Operator
Single-Phase Recloser
Polymer Arrester
Cutout & Arrester Bracket
Alley Arm Brace
Electronic Sectionalizer
Cutout
Dead End Clevis
Double Arming Bolt
Pole Step
By-Pass Switch
Polymer Cutout
Polymer Riser Pole Arrester
Insulated Clevis
Thimble Eye Nut
Thimble Eye Bolt
Pole Eye Plate
Machine Bolt & Washer
Fiberglass Guy Strain Insulator
Suspension Clamp
Lashing Clamps
Guy Wire Clip
Sus-pension Clamp
U Cable Guard
Guy Marker
Conduit Standoff Bracket
Bonding Clamp
Automatic Guy Deadend
Detachable Pole Step
Cable Guard Strap
Polymer Elbow Arrester
Ground Rod Clamp
Lag & Plate for Detachable Pole Step
Thimble Eye
Separate Elbow
Anchors
Junction
Multi-Tap
Ground Rod
Motor Operator
Padmounted Switch
Polymer Concrete Underground Enclosure
Deadbreak Elbow
Grounded Bushing
Polymer Concrete Box Pad or Flat Pad

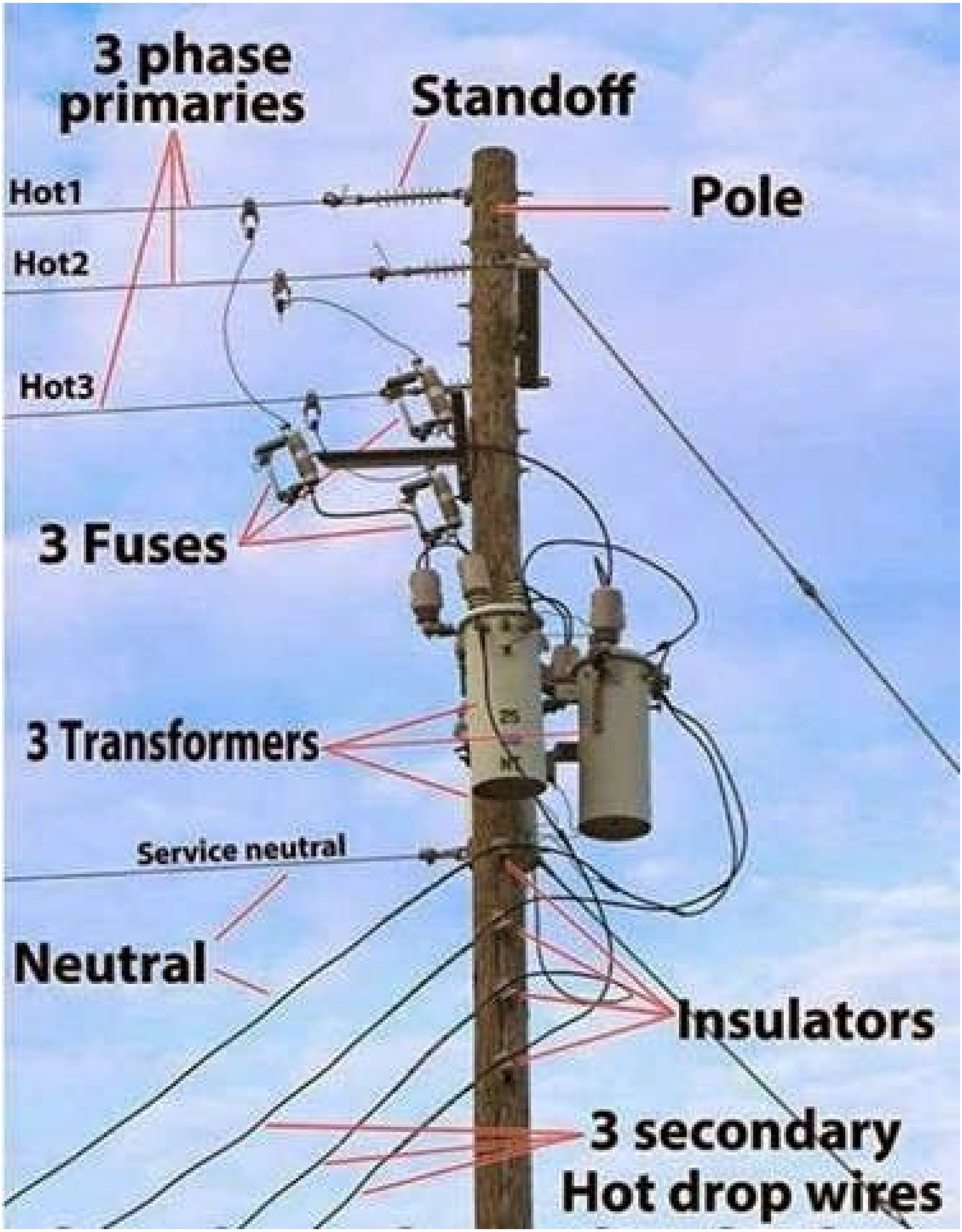
3 phase primaries
Standoff
Hot1
Pole
Hot2
Hot3
3 Fuses
3 Transformers
Service neutral
Neutral
Insulators
3 secondary Hot drop wires

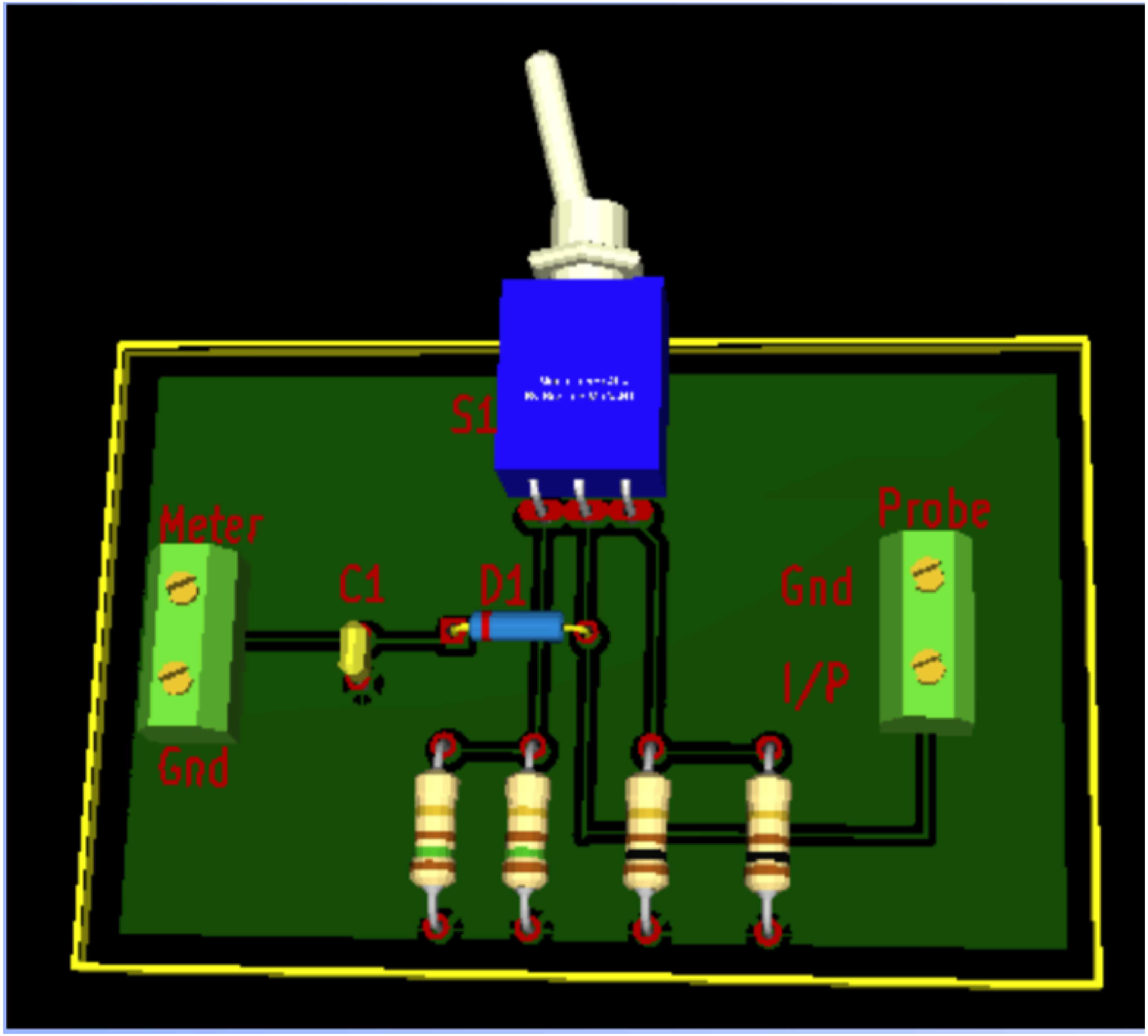
S1
Meter
C1
D1
Probe
Gnd
I/P
Gnd

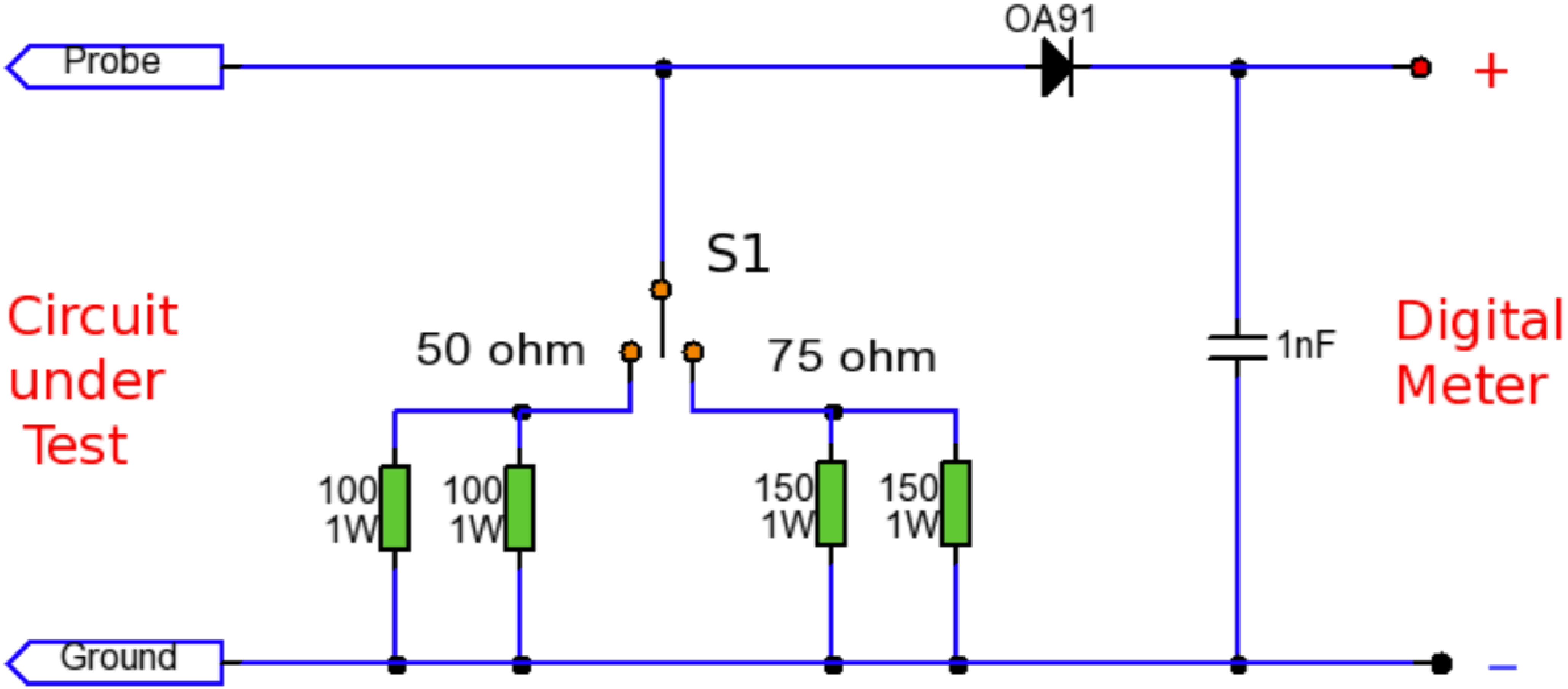
Probe
OA91
+
S1
Circuit under Test
50 ohm
75 ohm
1nF
Digital Meter
100 1W
100 1W
150 1W
150 1W
Ground
–

N. TESLA.
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

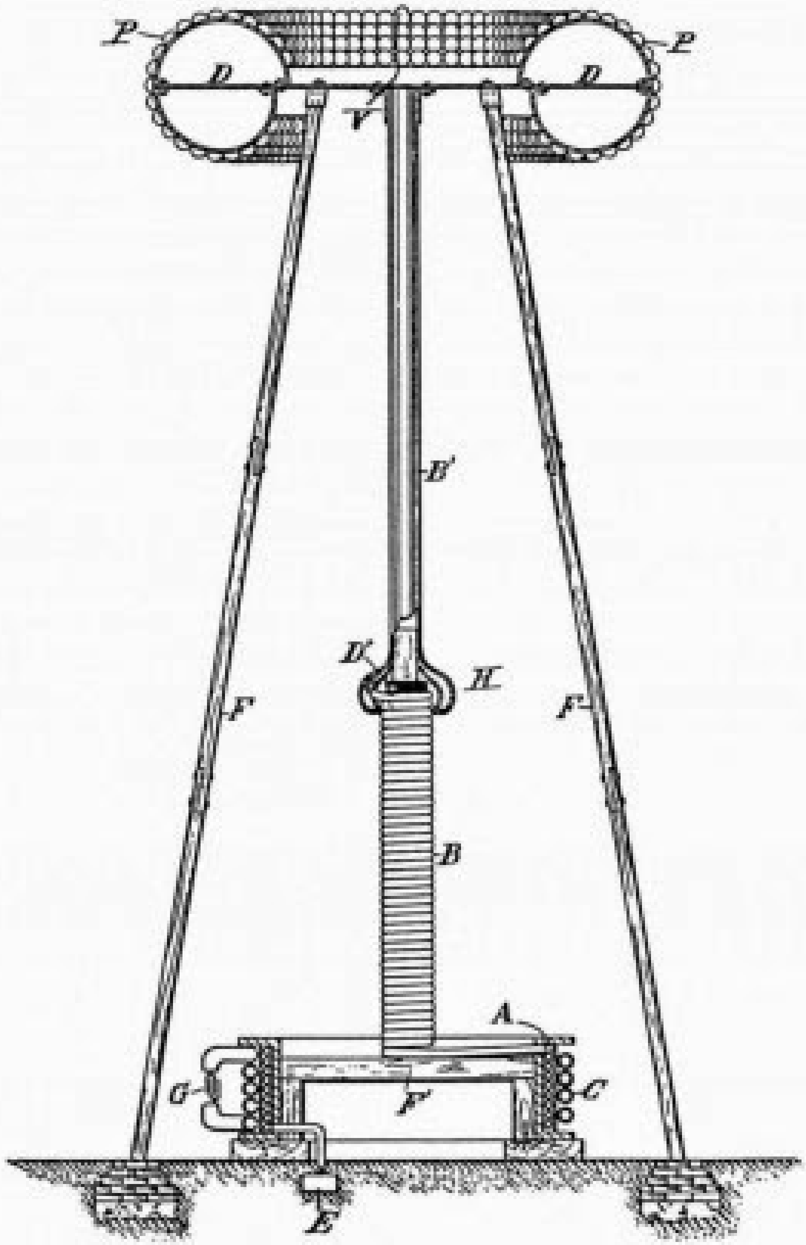
P
P
D
D
V
E'
D'
H
F
F
B
A
G
C
F'
E

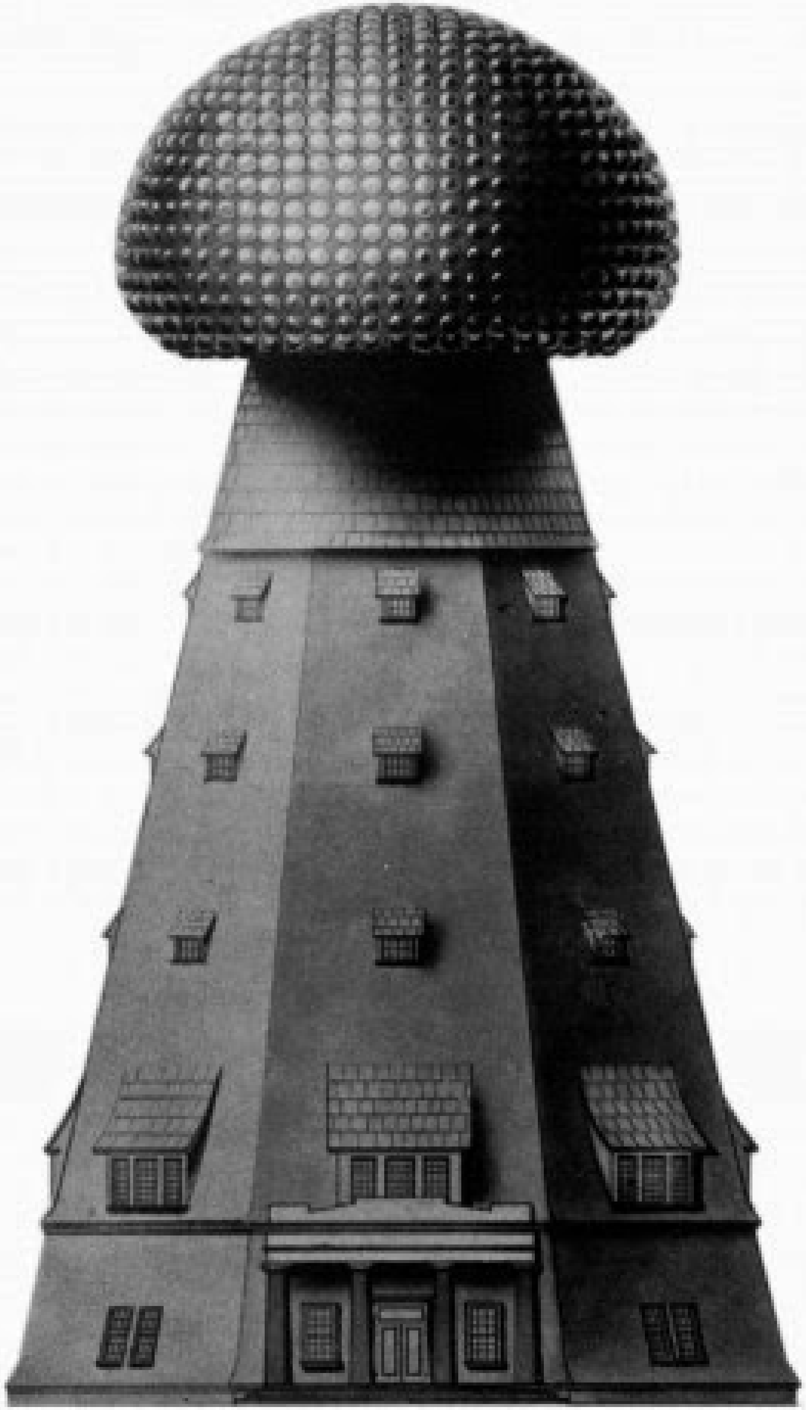

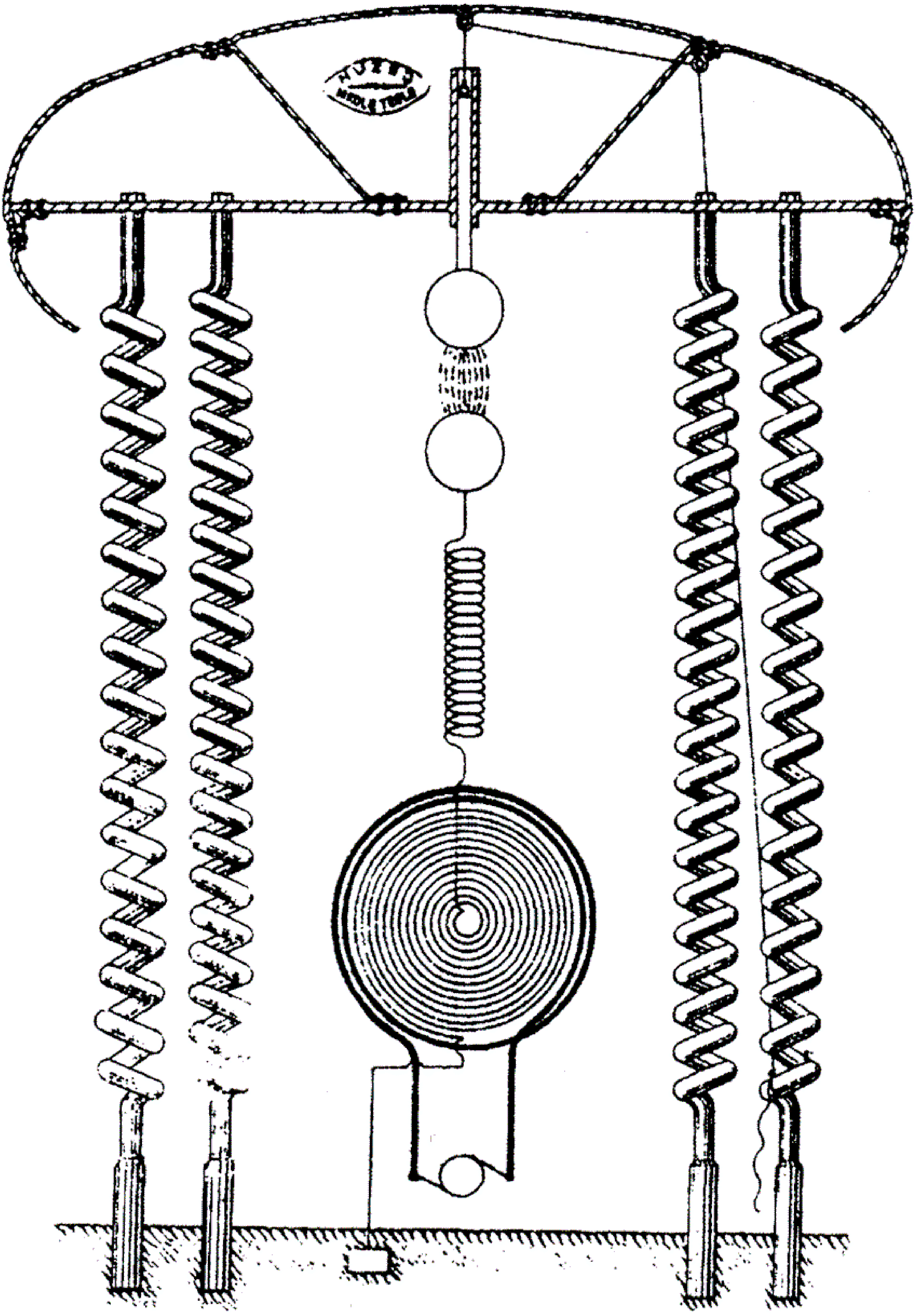

New York, Aug. 30th, 1901
46 & 48 East Houston Str.

Mr. Stanford White
160 Fifth Ave.
New York City

My Dear Stanford:

Many thanks for your suggestions. I am writing to Mr. Powell today. Perhaps he will be able to clear the land altogether.

I want you to understand that I went to the American Bridge Company simply because of my anxiety to have the work pushed through as fast as practicable. I am only too glad to follow your advice and beg you to consider yourself absolutely free in your choice and arrangements regarding this work.

Yours very sincerely,
N. Tesla

New York, Sep. 12th, 1901
46 & 48 East Houston Str.

Babcock & Wilcox Co.
85 Liberty Street
New York City

Gentlemen:

Under enclosure I forward sketch showing your two boilers as they will be placed in my building and their position relative to and exact distance from the chimney. The scale is ½ inch to a foot.

You will greatly oblige me by furnishing the drawings of the flues leading to the chimney and the position of the breech, as the builder cannot proceed without this information.

Yours very truly,
Encl.

Anyone familiar with the Wardenclyffe Tower knows it to have been a colossal structure. Yet, few realize that it was supposed to have been even larger. Although the exact figures are not revealed, Tesla must have drastically underestimated the cost of building his structure as is evidenced by the following response to White.

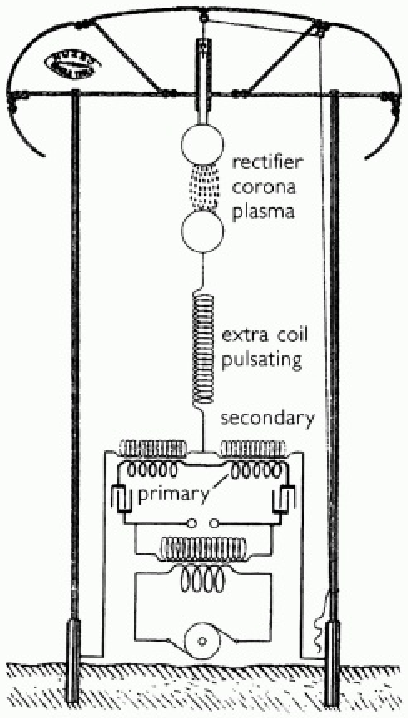

Fig. 5 Oscillating statically charged terminal.

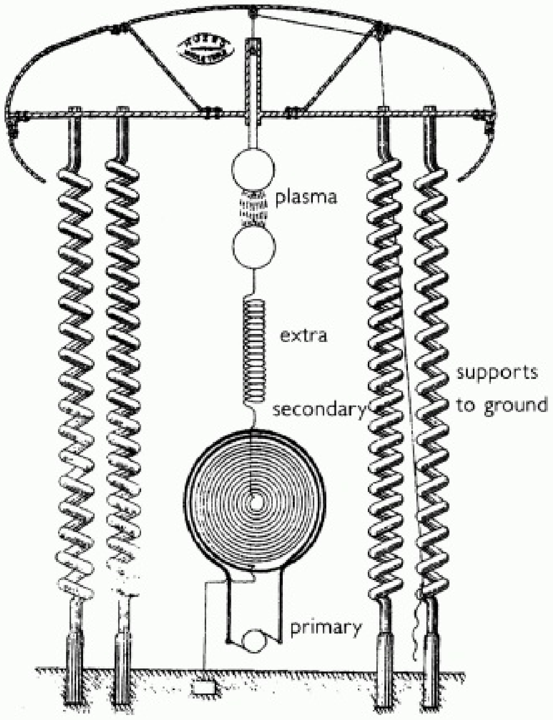

Fig. 6 Oscillating electrostatically charged dome.

N. TESLA
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

This patent was the result of Tesla's many high voltage transmission experiments at Colorado Springs during the early 1900's. It served as the basis for the construction of the famous:-''Waldencliff Tower,'' which was intended to become a transmission station for wireless electrical power.

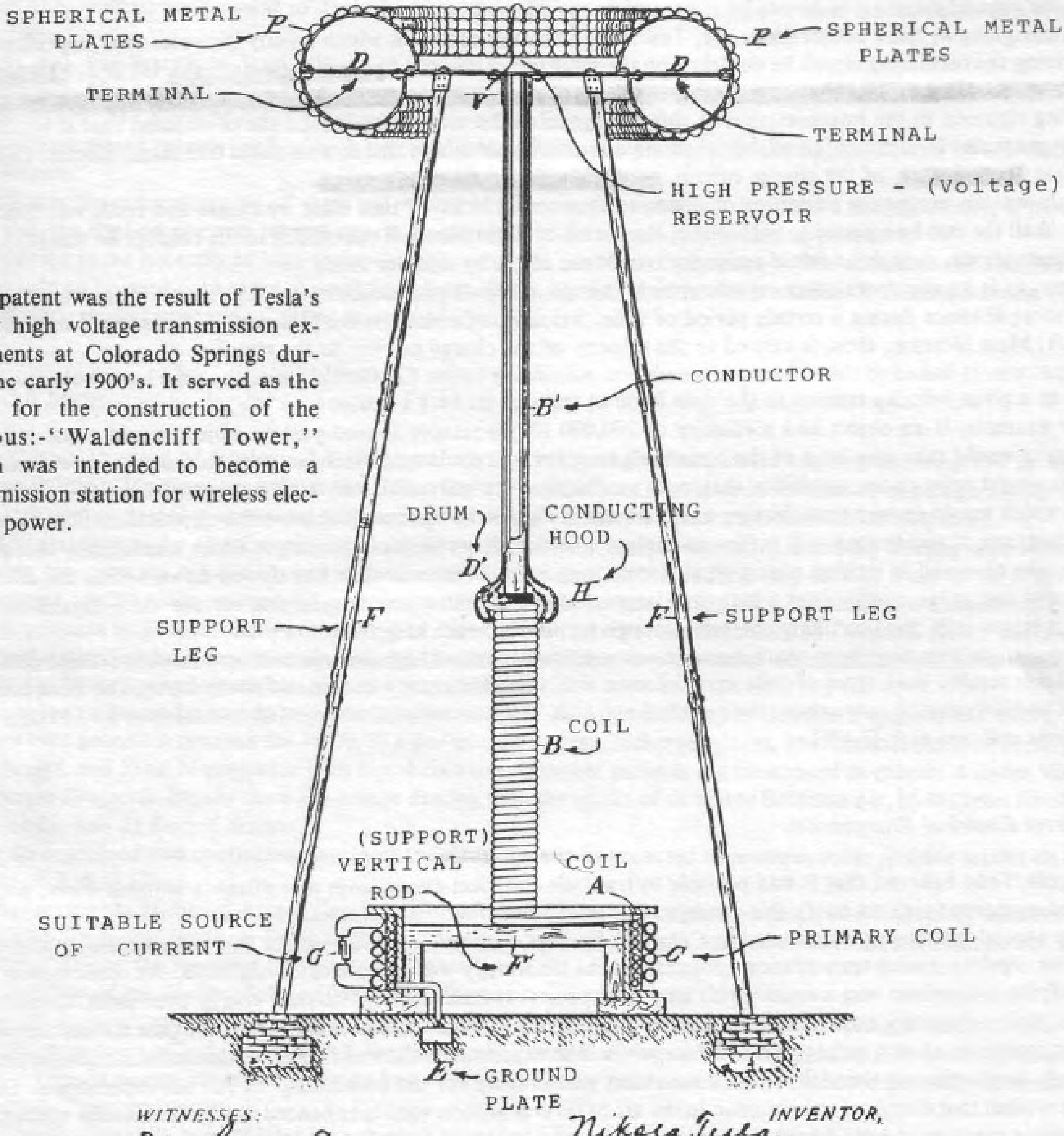
SPHERICAL METAL PLATES
P
P
SPHERICAL METAL PLATES
D
D
TERMINAL
V
TERMINAL
HIGH PRESSURE - (Voltage) RESERVOIR
CONDUCTOR
B'
DRUM
CONDUCTING HOOD
D'
H
SUPPORT LEG
F
F
SUPPORT LEG
COIL
B
(SUPPORT) VERTICAL ROD
COIL
A
SUITABLE SOURCE OF CURRENT
G
F'
C
PRIMARY COIL
E
GROUND PLATE

WITNESSES:

M. Lawson Dyer

Benjamin Miller.

Nikola Tesla, INVENTOR,

BY Kerr, Page & Cooper, his ATTORNEYS.

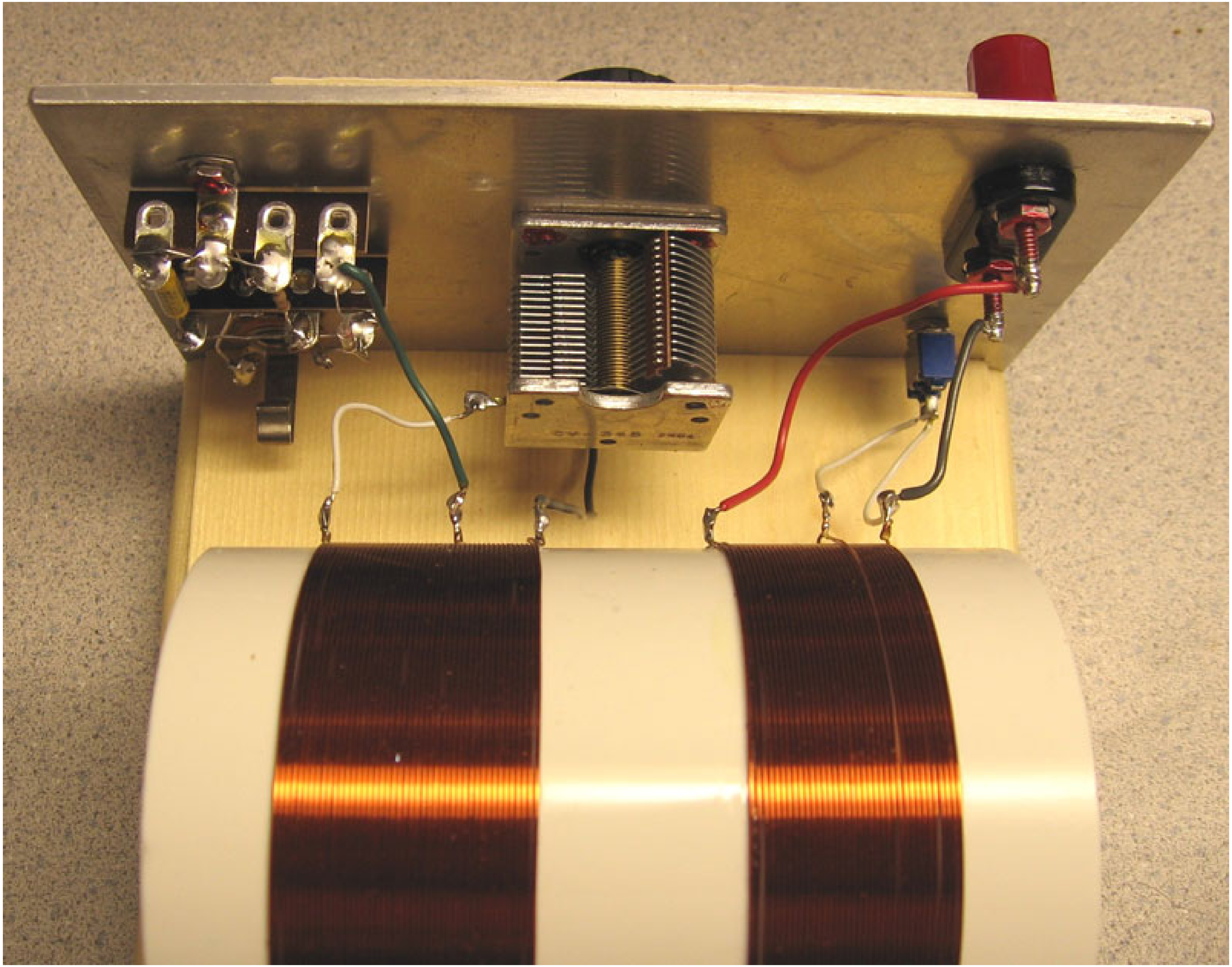

"HIGH SELECTIVITY CRYSTAL RADIO" [WITH SHORT WAVE]

RGK 2007

570KHZ → 2.5MHZ

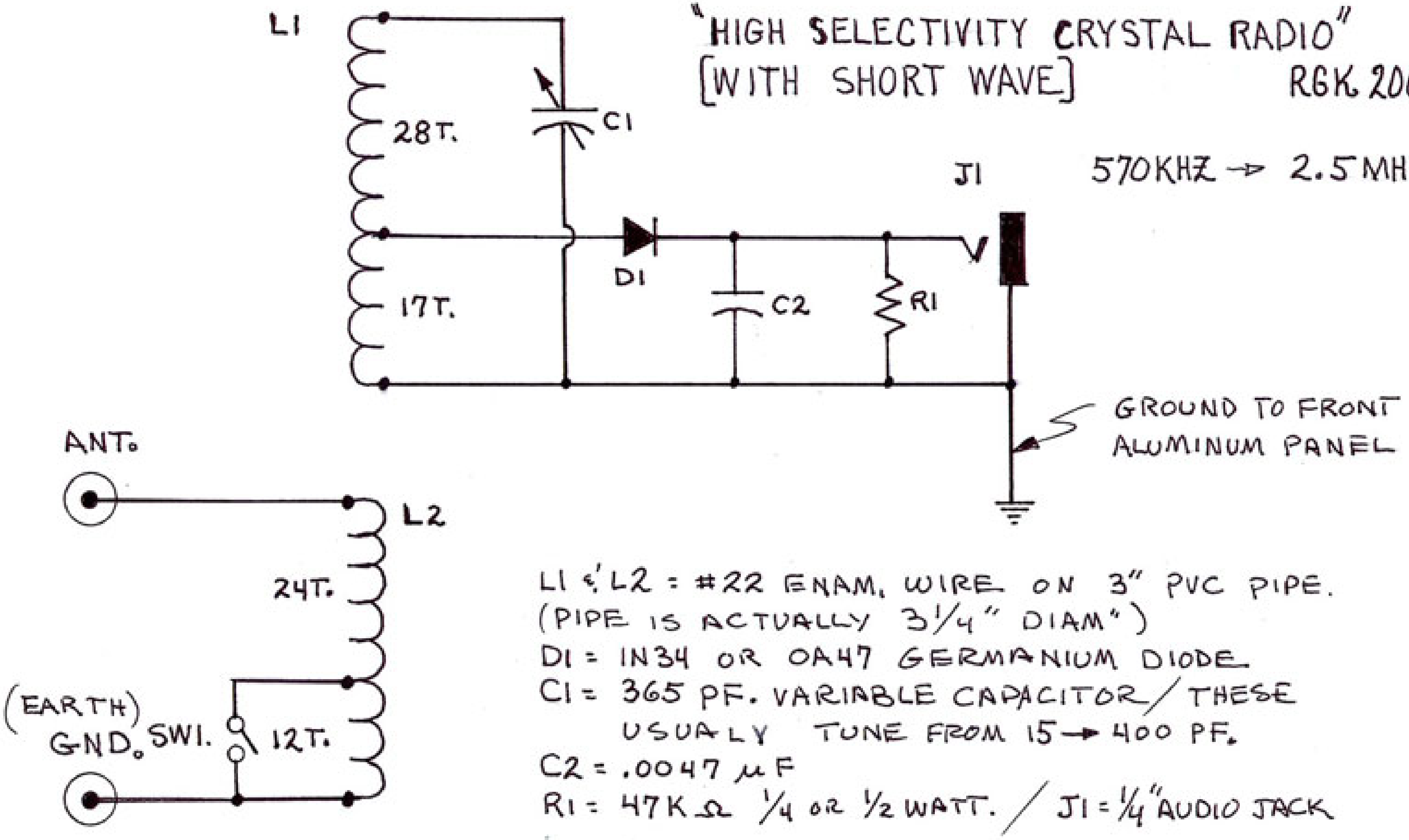
L1
28T.
17T.
C1
D1
C2
R1
J1
GROUND TO FRONT ALUMINUM PANEL
ANT.
L2
24T.
(EARTH) GND.
SW1.
12T.

L1 & L2 = #22 ENAM. WIRE ON 3" PVC PIPE. (PIPE IS ACTUALLY 3¼" DIAM")
D1 = 1N34 OR OA47 GERMANIUM DIODE
C1 = 365 PF. VARIABLE CAPACITOR / THESE USUALY TUNE FROM 15 → 400 PF.
C2 = .0047 μF
R1 = 47K Ω ¼ OR ½ WATT. / J1 = ¼" AUDIO JACK

[L1/L2 NOTES: L1 = 45 TURNS TOTAL & L2 = 36 TURNS TOTAL. THE SPACE BETWEEN L1 & L2 IS 1 1/16".]

SW1 = SP.S.T. SWITCH / FOR ANTENNA MATCHING

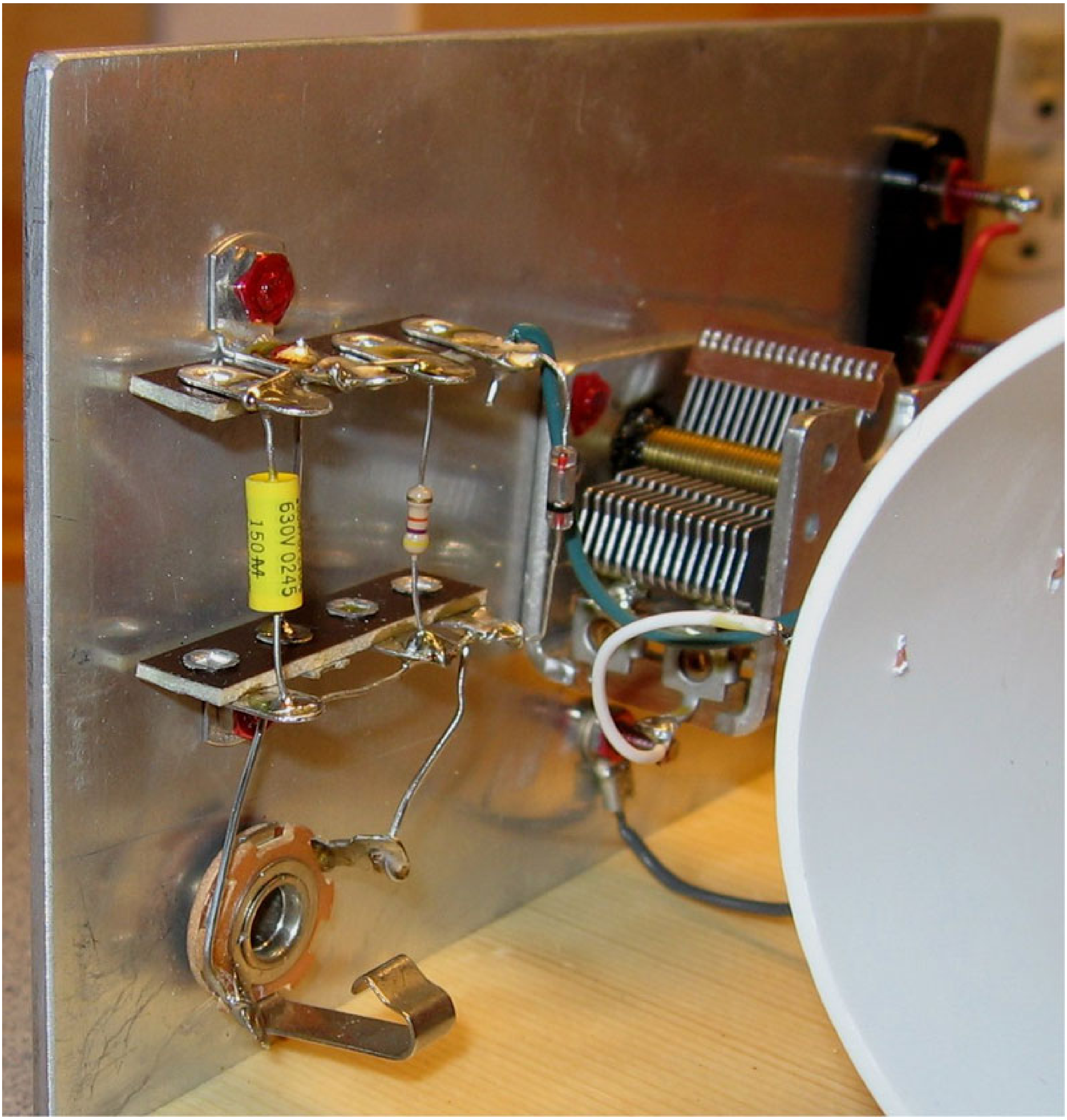
630V 0245
150M

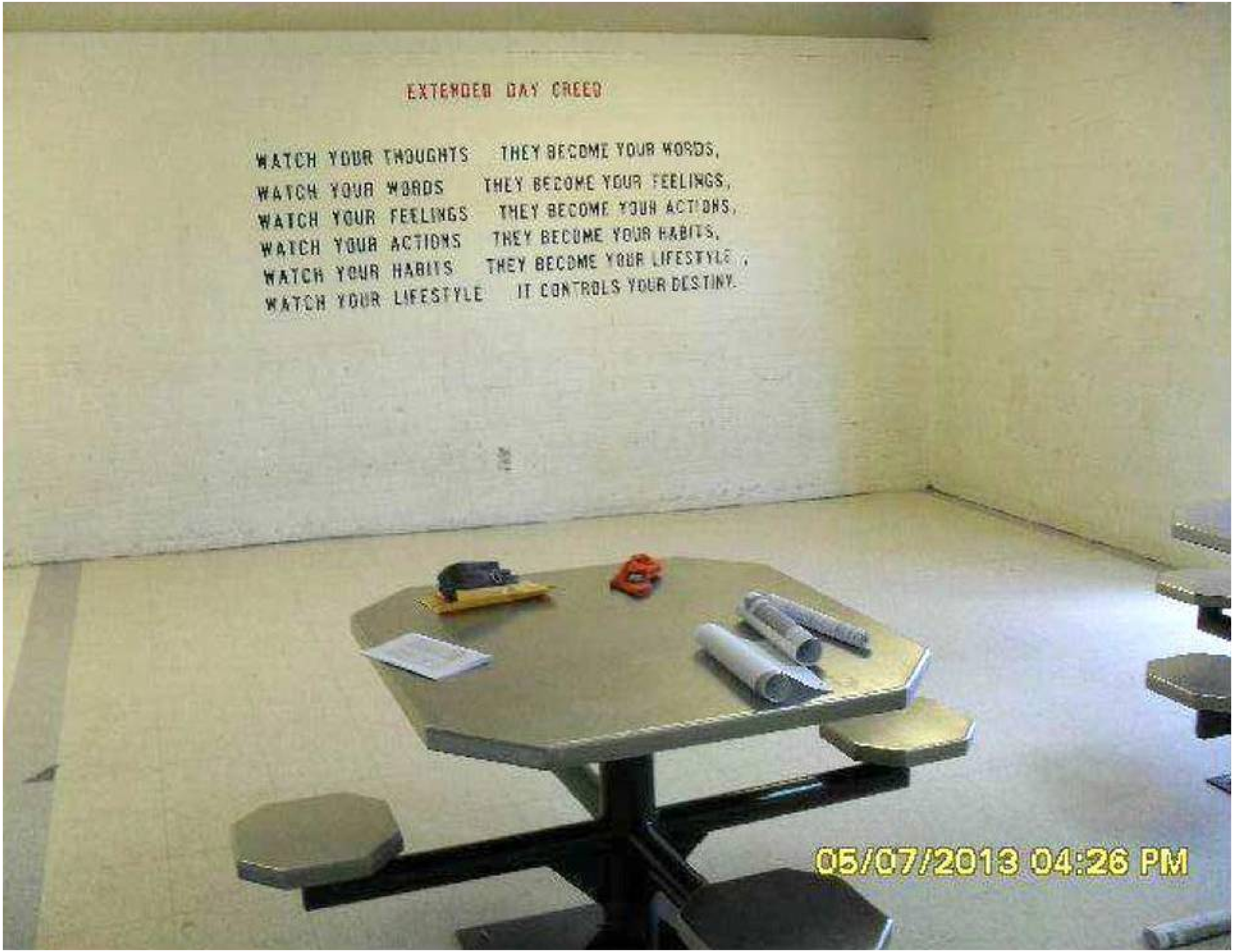
EXTENDED DAY CREED
WATCH YOUR THOUGHTS THEY BECOME YOUR WORDS,
WATCH YOUR WORDS THEY BECOME YOUR FEELINGS,
WATCH YOUR FEELINGS THEY BECOME YOUR ACTIONS,
WATCH YOUR ACTIONS THEY BECOME YOUR HABITS,
WATCH YOUR HABITS THEY BECOME YOUR LIFESTYLE,
WATCH YOUR LIFESTYLE IT CONTROLS YOUR DESTINY.

05/07/2013 04:26 PM

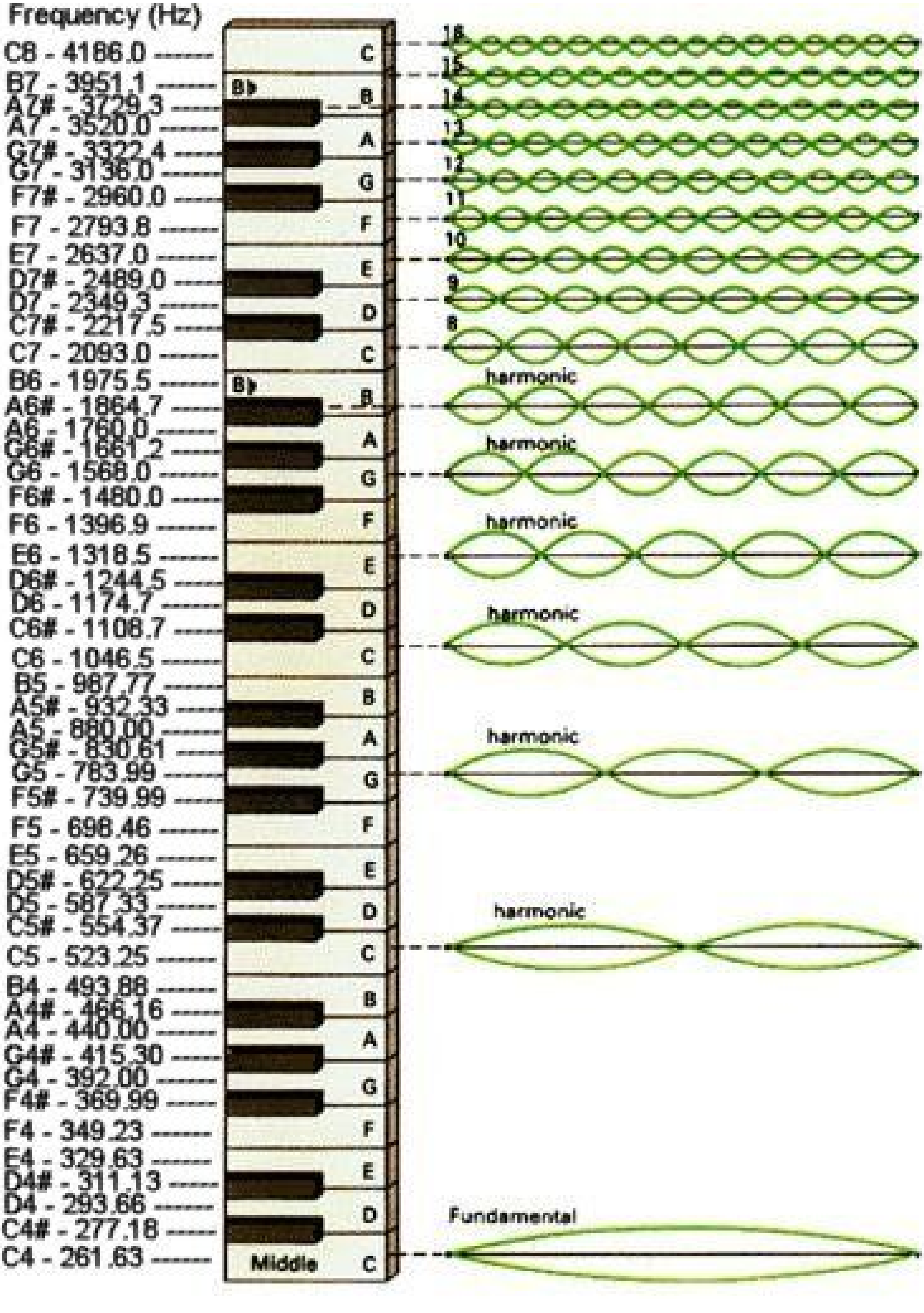
Frequency (Hz)
C8 - 4186.0
B7 - 3951.1
A7# - 3729.3
A7 - 3520.0
G7# - 3322.4
G7 - 3136.0
F7# - 2960.0
F7 - 2793.8
E7 - 2637.0
D7# - 2489.0
D7 - 2349.3
C7# - 2217.5
C7 - 2093.0
B6 - 1975.5
A6# - 1864.7
A6 - 1760.0
G6# - 1661.2
G6 - 1568.0
F6# - 1480.0
F6 - 1396.9
E6 - 1318.5
D6# - 1244.5
D6 - 1174.7
C6# - 1108.7
C6 - 1046.5
B5 - 987.77
A5# - 932.33
A5 - 880.00
G5# - 830.61
G5 - 783.99
F5# - 739.99
F5 - 698.46
E5 - 659.26
D5# - 622.25
D5 - 587.33
C5# - 554.37
C5 - 523.25
B4 - 493.88
A4# - 466.16
A4 - 440.00
G4# - 415.30
G4 - 392.00
F4# - 369.99
F4 - 349.23
E4 - 329.63
D4# - 311.13
D4 - 293.66
C4# - 277.18
C4 - 261.63
B♭
C
B
A
G
F
E
D
C
B♭
B
A
G
F
E
D
C
B
A
G
F
E
D
C
B
A
G
F
E
D
Middle
C
16
15
14
13
12
11
10
9
8
harmonic
harmonic
harmonic
harmonic
harmonic
harmonic
Fundamental

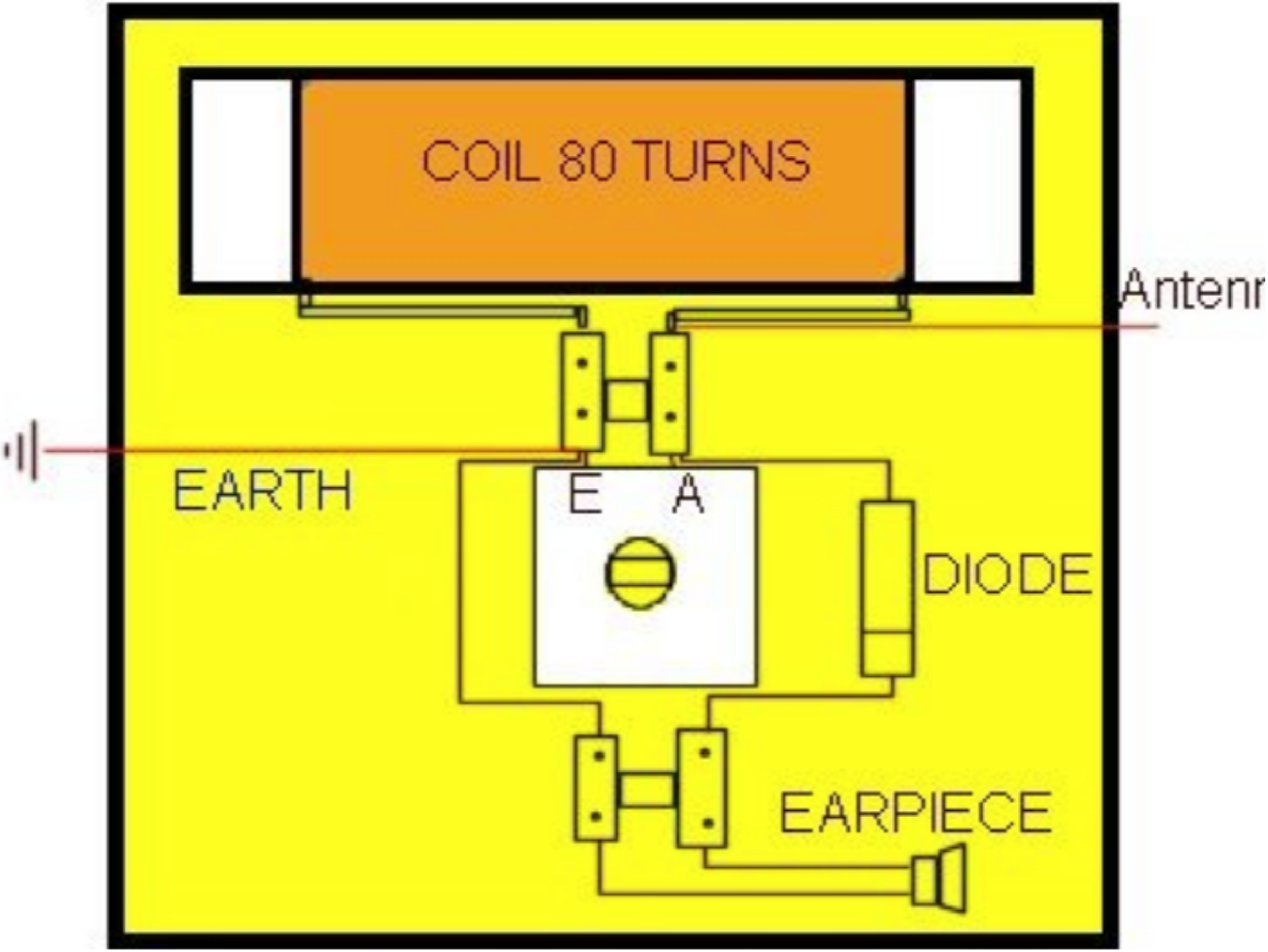
COIL 80 TURNS
Antenn
EARTH
E
A
DIODE
EARPIECE

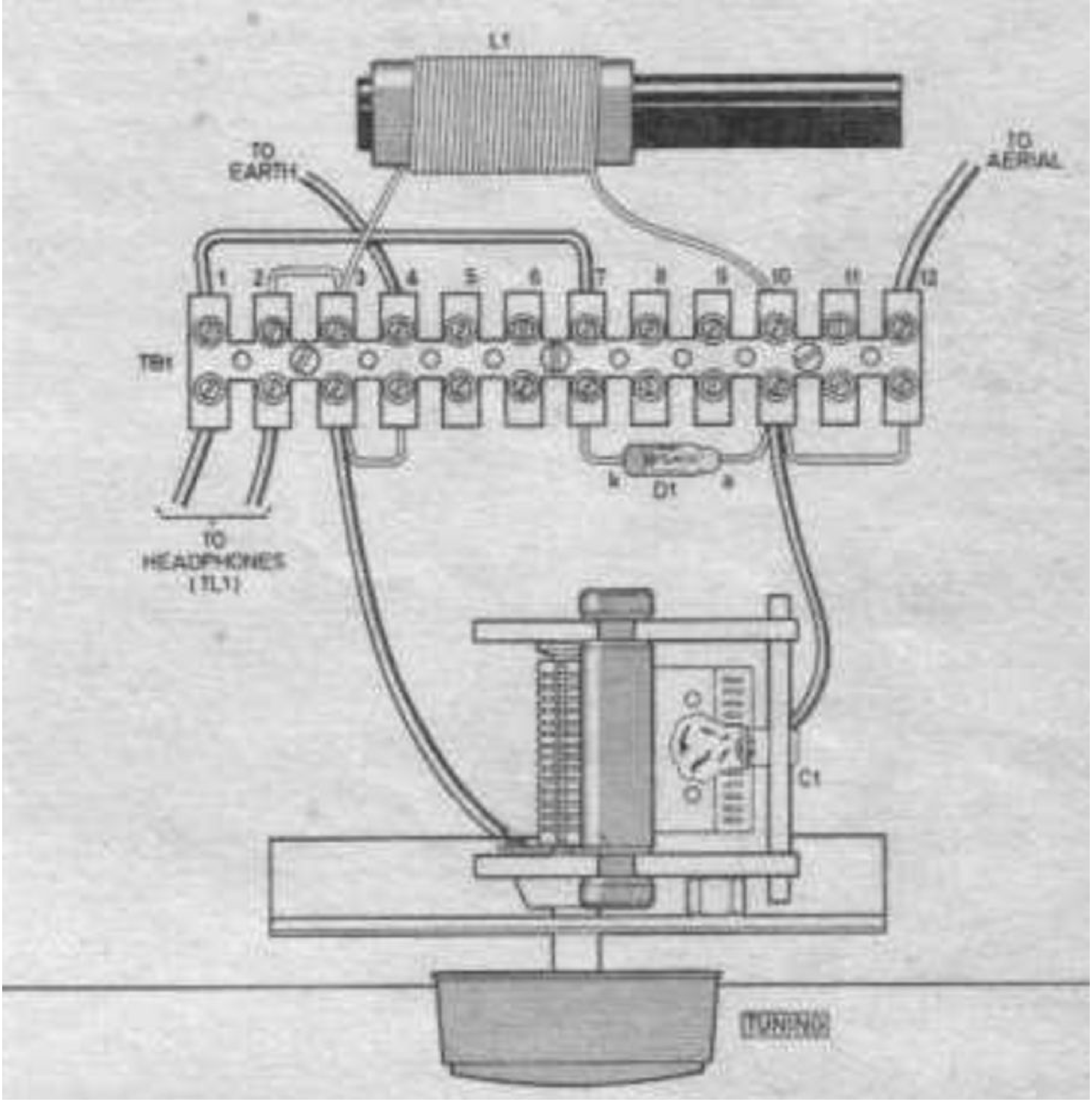
L1
TO EARTH
TO AERIAL
1
2
3
4
5
6
7
8
9
10
11
12
TB1
k
D1
a
TO HEADPHONES (TL1)
C1
TUNING

RF Energy Harvesting for INTERNET OF THINGS

The world is the index
Take the world on line
Take control of the world
Let the things talk to each other
Let thing become intelligen

K.J. Vinoy

ECE Dept
Indian Institute of Science
Bangalore 560012

ECE Dept, Indian Institute of Science

- **Vision Statement:** Excellence in Theoretical and Experimental Research in Communications, Signal Processing, Microelectronics and RF/Photonics.
- **Faculty: 24;** Fellows of IEEE: 4; Fellows of INAE: 8
- Active in Publications: Books, Book Chapters, Journal & Conference Papers; Patents; Standardization etc
- Collaborative research

People
Masters Students
[ME, MSc→ Mtech, Mtech(Res)]
PhD Students
Project Staff

INTERNATIONAL COLLABORATION

NATIONAL COLLABORATION

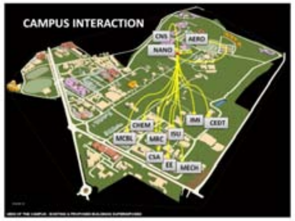
CAMPUS INTERACTION
CNS
AERO
NANO
CHEM
IMI
CEDT
MCBL
MRC
IISc
CSA
EE
MECH

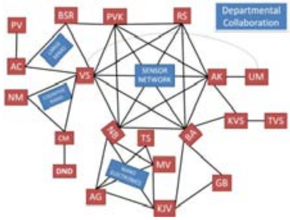
Departmental Collaboration
PV
BSR
PVK
RS
AC
VS
SENSOR NETWORK
AK
UM
NM
KVS
TVS
CM
RB
TS
BA
DND
MV
NANO ELECTRONICS
GB
AG
KJV

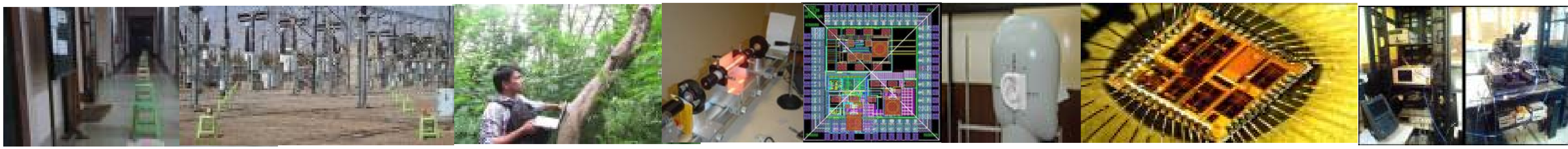

ECE: Microwave Engineering

K.J. Vinoy

- Low-Actuation Voltage Capacitive RF MEMS Switch (<10V)
  - Low-complexity fabrication process to enhance process yield
  - High reliability: no failure even up to 10 million cycles of operation tested
- Meso-scale Electrostatic Phase Shifter on microwave Laminate (MEPL)
  - Utilizes modern printed circuit board fabrication technology.
  - X-band monolithic antenna array system on the microwave laminate board demonstrated.

- Wideband group delay engineering in RF circuits for radar, medical imaging, and spectrum sensing.
  - Demonstration uses two stage All – Pass Networks; can be extended over multiple stages to obtain a higher bandwidth and/or higher group delay slope.
- RF energy harvesting circuits
  - Integrated with RF transmitters and sensors for practical IoT nodes
  - High efficiency RF-DC converter which can operate at input power of -20dBm (10µW) at 2.4GHz using UMC 130nm process MOSFETs.
- FEM based algorithms for Electromagnetic circuits & components (periodic structures such as metamaterials)
  - Fast computation of electromagnetic propagation characteristics
  - Especially suited for evaluation of processes uncertainties

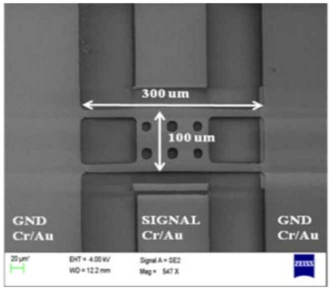

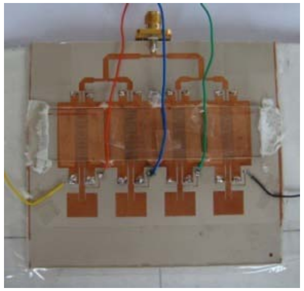

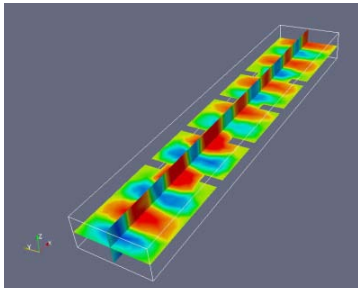

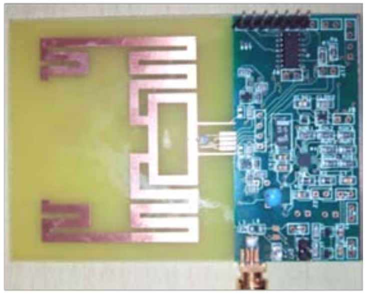

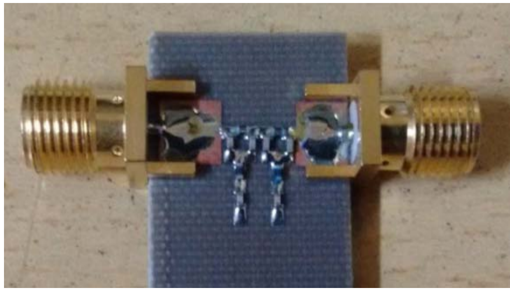

Setting the stage….

- **Introduction**
  - **Wireless Power Transfer**
    - **Energy Harvesting**
  - **Internet of Things**
- Highlights of Recent Development (Hardware)
  - Powering wireless terminals
  - RFID with integrated sensors

Ongoing Research Challenges

Wireless Power Transfer (WPT)

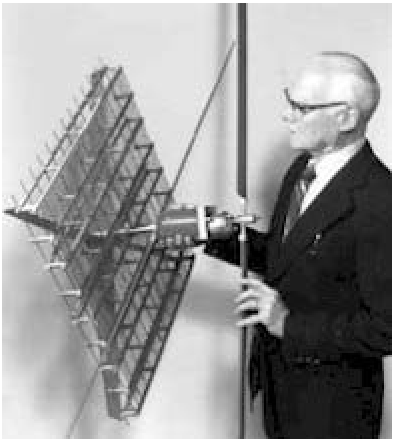

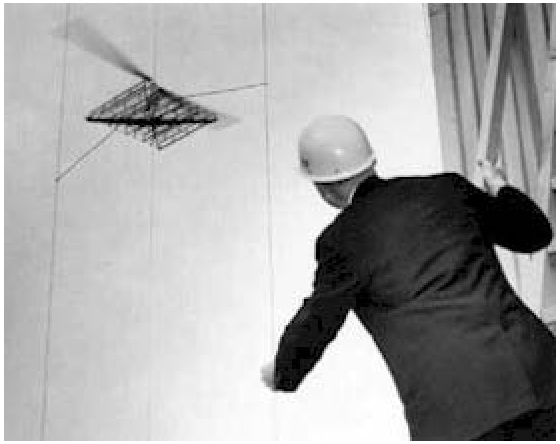

- Indicates transfer of electric energy remotely
- WPT has a long history!!
  - Tesla demonstrated it in 1899 by wirelessly powering fluorescent lamps 40 kms away from the power source.
  - Had multiple patents in early 1900s.
- In 1960s W.C. Brown coined the term Rectenna, which he used to directly converts incoming microwaves to DC.
  - He demonstrated its ability to power a helicopter solely through microwaves for 10 hours continuously.
- These demonstrations involved dedicated sources with **large power to transmit over long distances**.

http://mainland.cctt.org/istf2008/brown.asp

SSPS

- Space Solar Power Satellites

- WPT is widely investigated for putting **solar power generating satellites** into space and transmitting power to Earth stations. (Mainly in Japan)

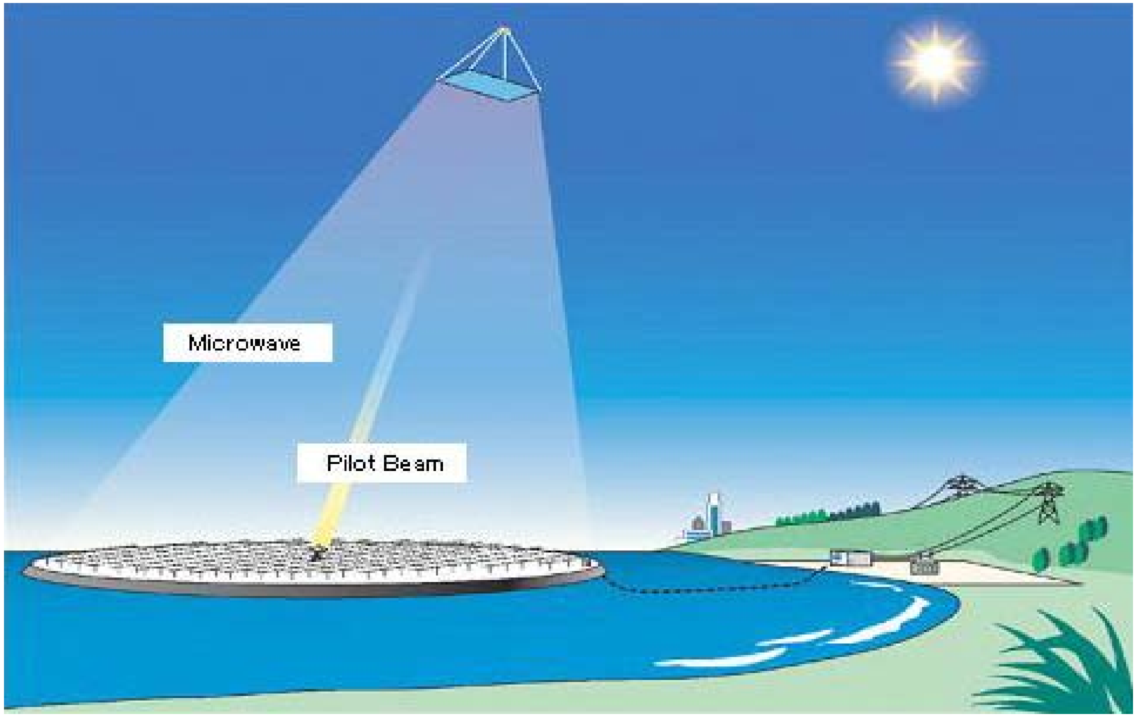

http://www.jspacesystems.or.jp/en_project_ssps/

Near field Wireless Power Transfer

- Recent demonstration by MIT to transfer high RF power (Watts) transferred across meters.
- Resonant coils are used
- Typically at 100 kHz to 10's of MHz
- Many new applications emerged

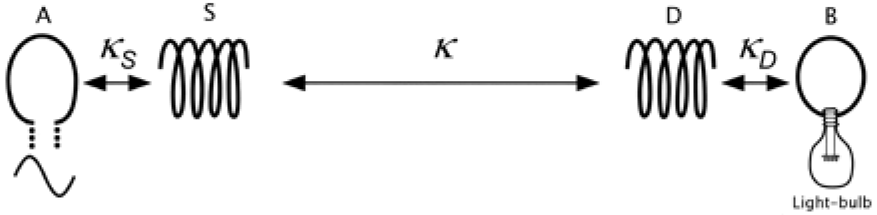

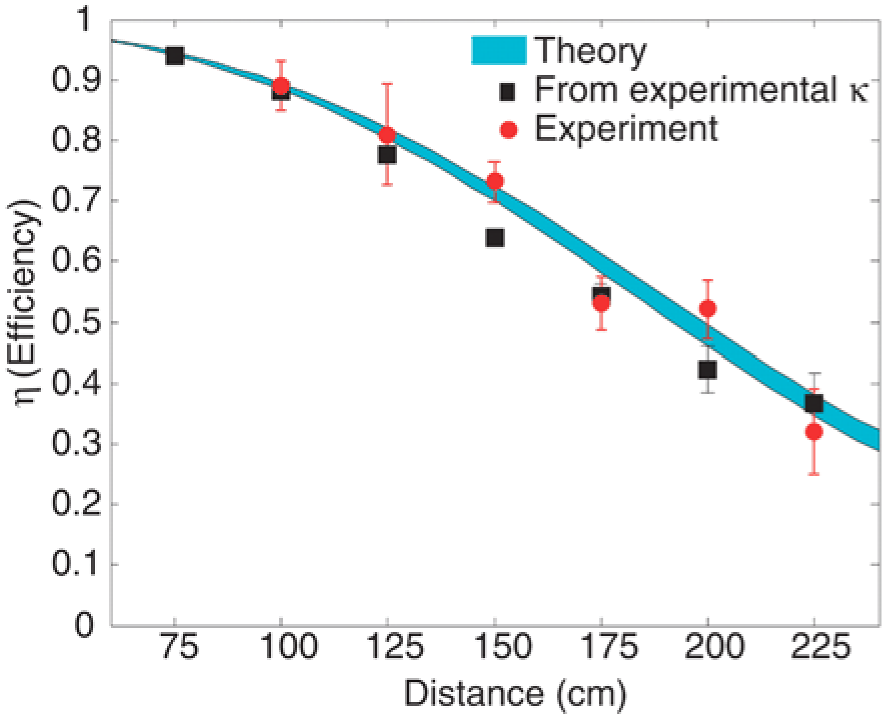

André Kurs et al, Wireless Power Transfer via Strongly Coupled Magnetic Resonances , Science, Vol. 317 no. 5834 pp. 83-86. 6 July 2007.

System Schematic Qi

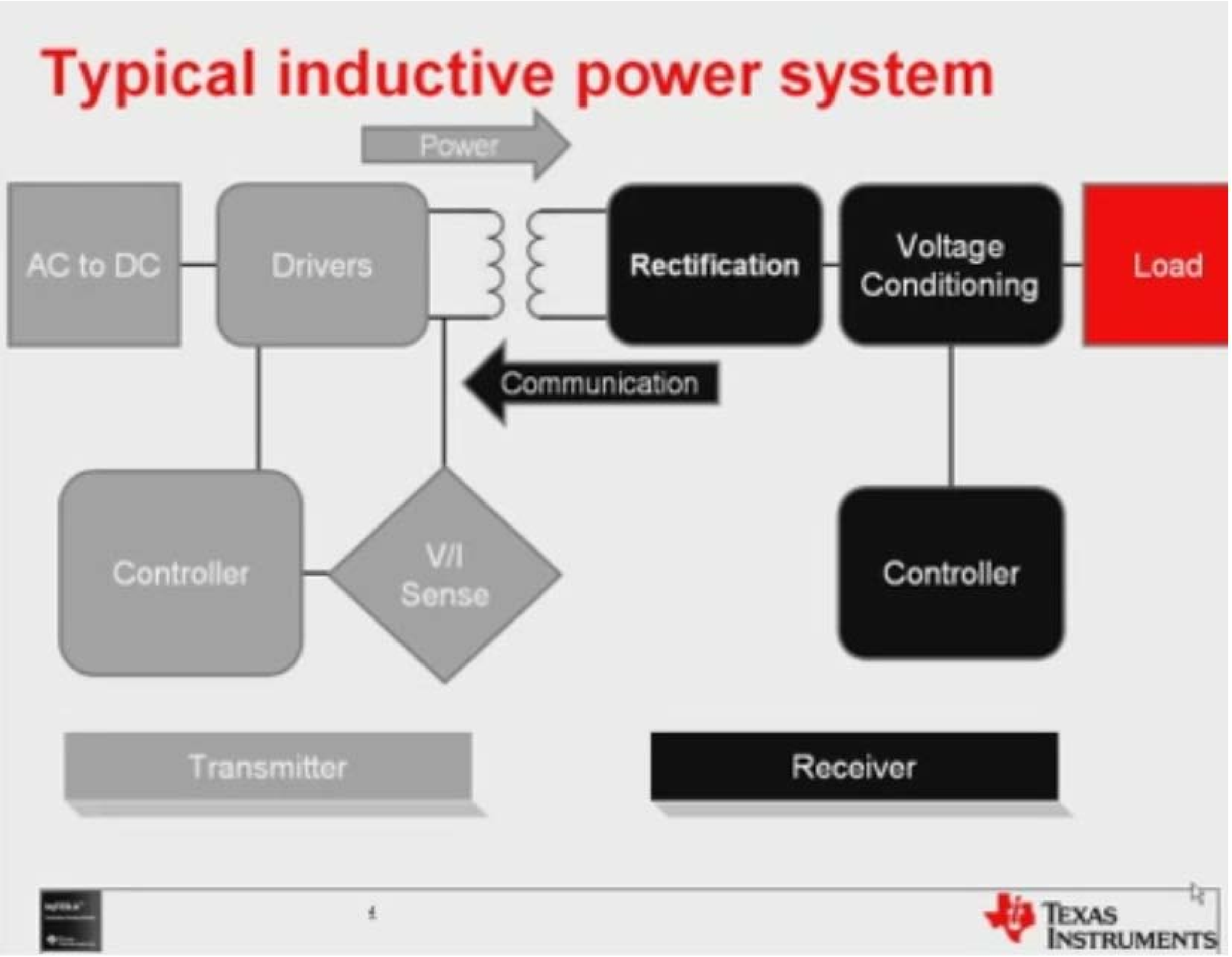

Typical inductive power system

Power
AC to DC
Drivers
Rectification
Voltage Conditioning
Load
Communication
Controller
V/I Sense
Controller
Transmitter
Receiver
TEXAS INSTRUMENTS

Source: Texas Instruments Qi Development kit

Far → Near in WPT

- Free space loss factor is a major bottleneck for power transfer at large distances

- Short distance/ Near field options
  - MIT demonstration (2007)
  - Qi Standard
  - Phone charging solutions
  - Vehicles running on wireless power
- Two extremes in WPT
  - mW ← → MW
  - mm ← → 1000s km
  - 100kHz ← → 2.4/5.8GHz
  - 10cm x 10cm ← → km x km
  - Commercial vs bluesky

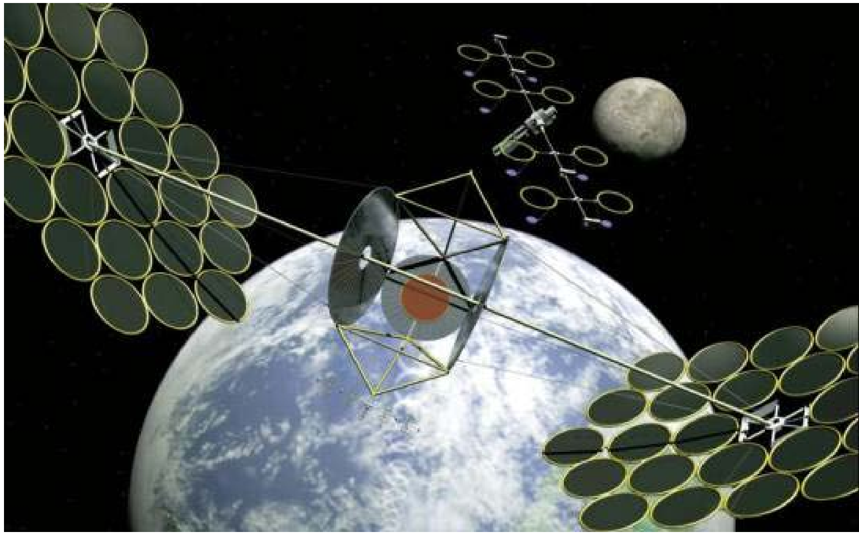

Far Field Transfer of RF Energy

- Focus of this talk
- Applications: RFID tags, Wireless Sensor Network nodes, biomedical equipment, home automation and structural monitoring can benefit from RF energy harvesting.
- Block diagram and a design example:

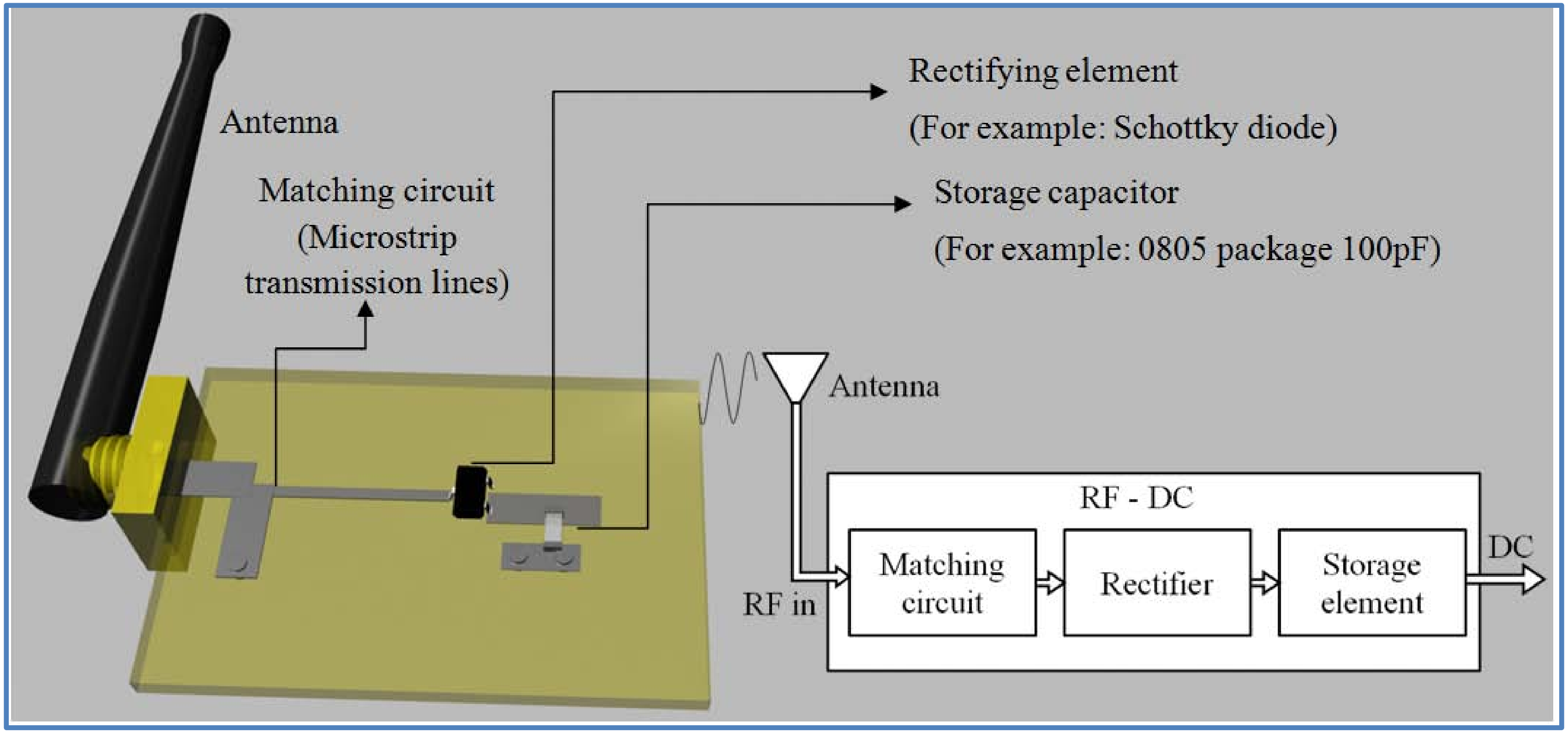
Rectifying element
(For example: Schottky diode)
Antenna
Matching circuit
(Microstrip
transmission lines)
Storage capacitor
(For example: 0805 package 100pF)
Antenna
RF - DC
RF in
Matching
circuit
Rectifier
Storage
element
DC

A New Paradigm: Internet of Things (IoT)

- IoT refers to uniquely identifiable objects and their virtual representations in an Internet-like structure.

Connects Anytime, Anyplace for Anyone (ICT)

AAA + for Anything (IoT)

- IoT is a scheme for connecting things: sensors, actuators, and other smart technologies, thus enabling person-to-object and object-to-object communications.

- **Continuous availability of power is crucial for their deployment**

IoT Applications

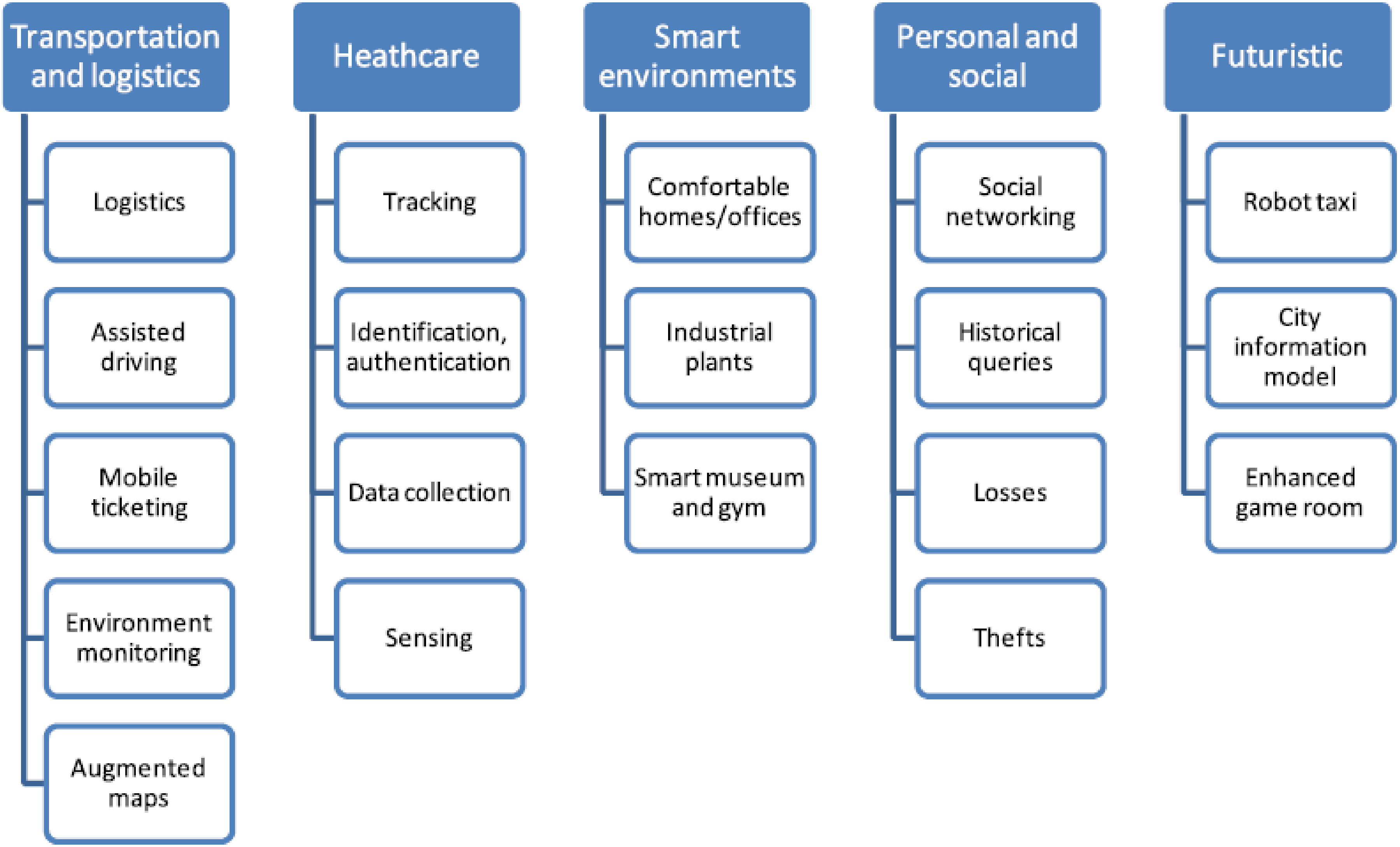

- Transportation and logistics
  - Logistics
  - Assisted driving
  - Mobile ticketing
  - Environment monitoring
  - Augmented maps
- Heathcare
  - Tracking
  - Identification, authentication
  - Data collection
  - Sensing
- Smart environments
  - Comfortable homes/offices
  - Industrial plants
  - Smart museum and gym
- Personal and social
  - Social networking
  - Historical queries
  - Losses
  - Thefts
- Futuristic
  - Robot taxi
  - City information model
  - Enhanced game room

L. Atzori , A. Iera, G. Morabito, The Internet of Things: A survey Computer Networks 54 (2010) 2787–2805

Comparison of different wireless protocols

Today, a lot can be done at low power!!

Characteristics of key 2.4GHz ISM Band Radios studied :

	BLE	ANT	Zigbee	WLAN
Topologies	P2P , Star	P2P , Star, tree, mesh	P2P , Star, mesh	P2P , Star
Modulation	GFSK	GFSK	OQPSK	DSSS (802.11b)
Max data rate	1Mbps	12.8-60 Kbps	250Kbs (@2.4Ghz)	1- 11Mbps (802.11b)
Throughput	305 kbps	20Kbps	100Kbps	6Mbps (802.11b)
Range (in m)	10-100(0-10dBm)	30 (@ 0dBm)	10-100 (0-20dBm)	100+(20dBm)
Max nodes in piconet	7	65533	Star-65536	32-64
Battery life	1-2 years (coin cell)	1-2years (coin cell)	100-1000 days	0.5-5days

Key Aspects :

BLE is robust and has lowest power consumption but cannot natively form mesh networks,

Zigbee can support large mesh networks, power consumption is higher than BLE and throughput is lower: It is suitable for low data rate, low power, large size networks.

WLAN is primarily suitable for transferring bulk data at high speeds, Not suitable for low power applications.

Power Requirements in Common WSN

	Crossbow MICAz	Intel IMote2	Jennic JN5139
Radio standard	IEEE802.15.4/ZigBee	IEEE802.15.4	IEEE802.15.4/ZigBee
Typical range	100m (outdoor), 30m (indoor)	30m	1 km
Data rate (kbps)	250 kbps	250 kbps	250 kbps
Sleep mode (deep sleep)	15 μA	390 μA	2.8 μA (1.6μA)
Processor only	8mA active mode	31–53mA∗	2.7+0.325mA/MHz
RX	19.7mA	44mA	34mA
TX	17.4mA (+0dbm)	44mA	34mA (+3 dBm)
Supply voltage (minimum)	2.7V	3.2V	2.7V
Average	2.8mW	12mW	3mW

JM. Gilbert∗ F. Balouchi, Comparison of Energy Harvesting Systems for Wireless Sensor Networks, International Journal of Automation and Computing 05(4), October 2008, 334-347

	Jennic JN5148	**TI- CC430**	**BLE**	**Zarlink ZL70250**
Active mode current at 16MHz [mA]	6	4	6.7	3.2
Deep sleep current [nA]	100	1000	400	20
Transmission current [mA]@Tx-power [dBm]	15@2.5	18@0	36@2	2@-10
Transmit frequency	2.4 GHz	2.4 GHz	2.4 GHz	868 MHz
Wakeup time [ms]	1	3	0.12	0.16
Energy consumption for a transmission cycle of 2ms [µJ]	183	300	196	32
Power supply voltage [V]	2.2 – 3.6	1.8 -3.6	2-3.6	1.2 – 1.8

In perspective

- Energy requirements in different devices/systems
- 6 orders of magnitude variation!!!
- Energy requirements in WSN
  - Depends on the complexity/ standard/ range
  - eg 90 μW to power a pulse oxymeter sensor, to process data and to transmit them at intervals of 15 s

Device type	**Power consumption**
Smartphone	1W
MP3 decoder chip	58 mW
Hearing aid	1 mW
Wireless sensor node	100 μW*
RF receiver chip	24 mW
GPS receiver chip	15 mW
6D motion sensor	14.4 mW
Cell phone (standby)	8.1 mW
PPG sensor	1.473 mW
Humidity	1 mW
Pressure	0.5 mW
3D accelerometer	0.324 mW
Temperature	27 μW
Cardiac pacemaker	50 μW
Wristwatch	7 μW
Memory R/W	2.17 μW
A-D conversion	1 μW

J Yun, S. Patel, M.Reynolds, G. Abowd "A Quantitative Investigation of Inertial Power Harvesting for Human-powered Devices," UbiComp'08, September 21-24, 2008, Seoul, Korea.

R.J.M. Vullers, et.al, Micropower energy harvesting, Solid-State Electronics 53 (2009) 684–693

Power Density from Various Harvesters

Ambient RF	< 1 μW/cm^2
Ambient light	100 mW/cm^2 (directed toward bright sun)
	100 μW/cm^2 (illuminated office)
Thermoelectric	60 μW/cm^2
Vibrational	4 μW/cm^3 (human motion ~Hz)
microgenerators	800 μW/cm^3 (machines ~kHz)
Ambient airflow	1 mW/cm^2
Push buttons	50 μJ/N
Hand generators	30 W/kg
Heel strike	Up tp 7 W for 1 cm deflection

Power requirements in conventional sensor network nodes may not be met by harvesting alone!

J Yun, S. Patel, M.Reynolds, G. Abowd "A Quantitative Investigation of Inertial Power Harvesting for Human-powered Devices," UbiComp'08, September 21-24, 2008, Seoul, Korea.

Electromagnetic: Solar → RF

- PV is a good source of energy. Recall,

Ambient light	100 mW/cm^2 (directed toward bright sun)
	100 μW/cm^2 (illuminated office)

- However, it light is not always available.
  - At night
  - Specific scenarios: in a closed chamber, or mine.
- RF is an alternative
  - Unlike others, RF sources may be ambient or intentional.
  - Ambient sources such as base stations or broadcast stations
  - Special sources: RF ID reader, Phone charger, special beacons

Demand and Supply

- The peak currents needed during transmit and receive operation is not achievable using the harvester alone.
- Buffering is also needed to ensure continuous operation during times without power generation.
- The combination of an energy harvester with a small-sized storage is the best approach to enable energy autonomy of the network over the entire lifetime.
  - Rechargeable battery
  - Thin film batteries
    - can be integrated directly in Integrated Circuit (IC) packages in any shape or size,
    - Flexible when fabricated on thin plastics
    - Thin film batteries have high impedance;
    - Low discharge efficiency compared to Li-ion batteries
  - super capacitor
    - Leakage in super capcitors depends on the voltage. Low at low voltage

Internet of Things (IoT)

Embedding short-range mobile transceivers into a wide array of gadgets and everyday items, enabling new forms of communication between people and things, and between things themselves.

RFID + Sensor + Smart Tech + Nano Tech = IoT

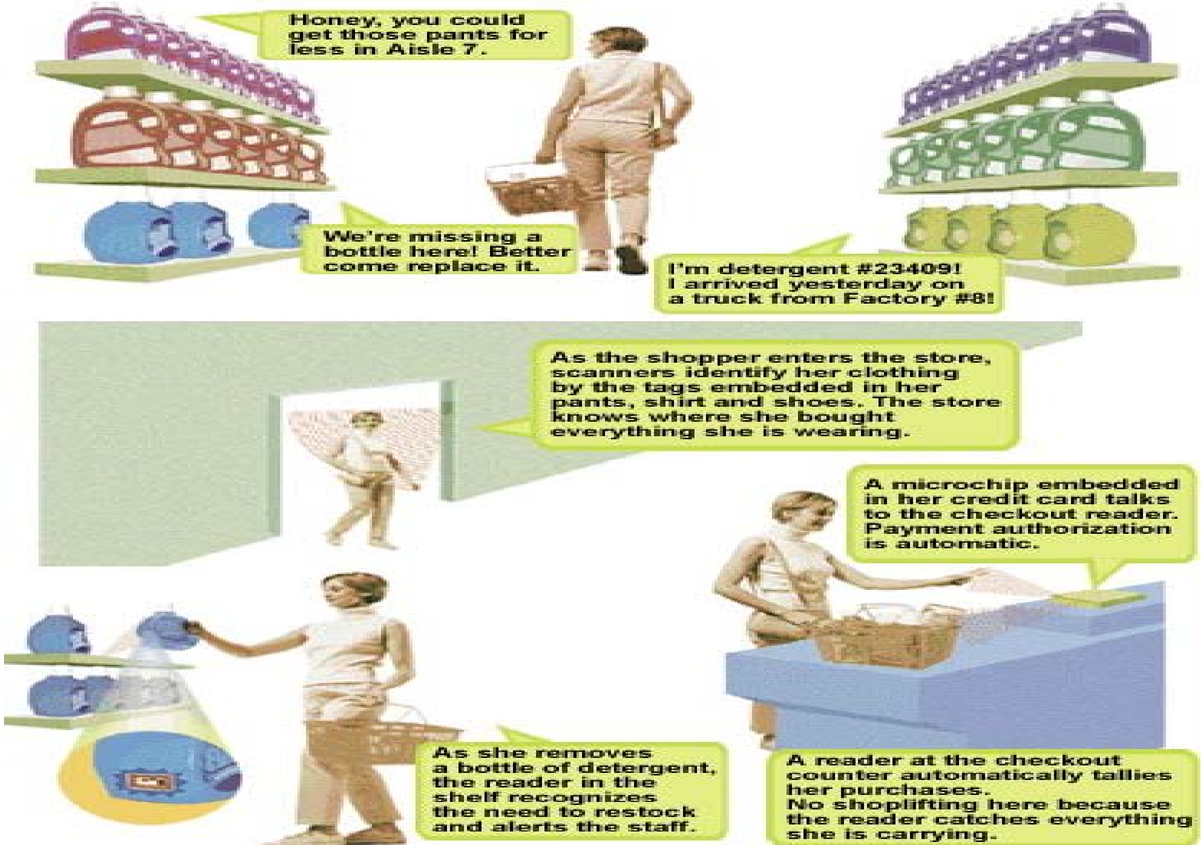
Honey, you could get those pants for less in Aisle 7.
We're missing a bottle here! Better come replace it.
I'm detergent #23409! I arrived yesterday on a truck from Factory #8!
As the shopper enters the store, scanners identify her clothing by the tags embedded in her pants, shirt and shoes. The store knows where she bought everything she is wearing.
A microchip embedded in her credit card talks to the checkout reader. Payment authorization is automatic.
As she removes a bottle of detergent, the reader in the shelf recognizes the need to restock and alerts the staff.
A reader at the checkout counter automatically tallies her purchases. No shoplifting here because the reader catches everything she is carrying.

When shopping in the market, the goods will introduce themselves.

When entering the doors, scanners will identify the tags on clothing.

When paying for the goods, the microchip of the credit card will communicate with checkout reader.

When moving the goods, the reader will tell the staff to put a new one.

Introduction to RFID

- The reader converts incident field and returns useful data
- In passive RFID systems reader transmits EM energy that "wakes up“ the tag and provides power for the tag to respond to the reader.

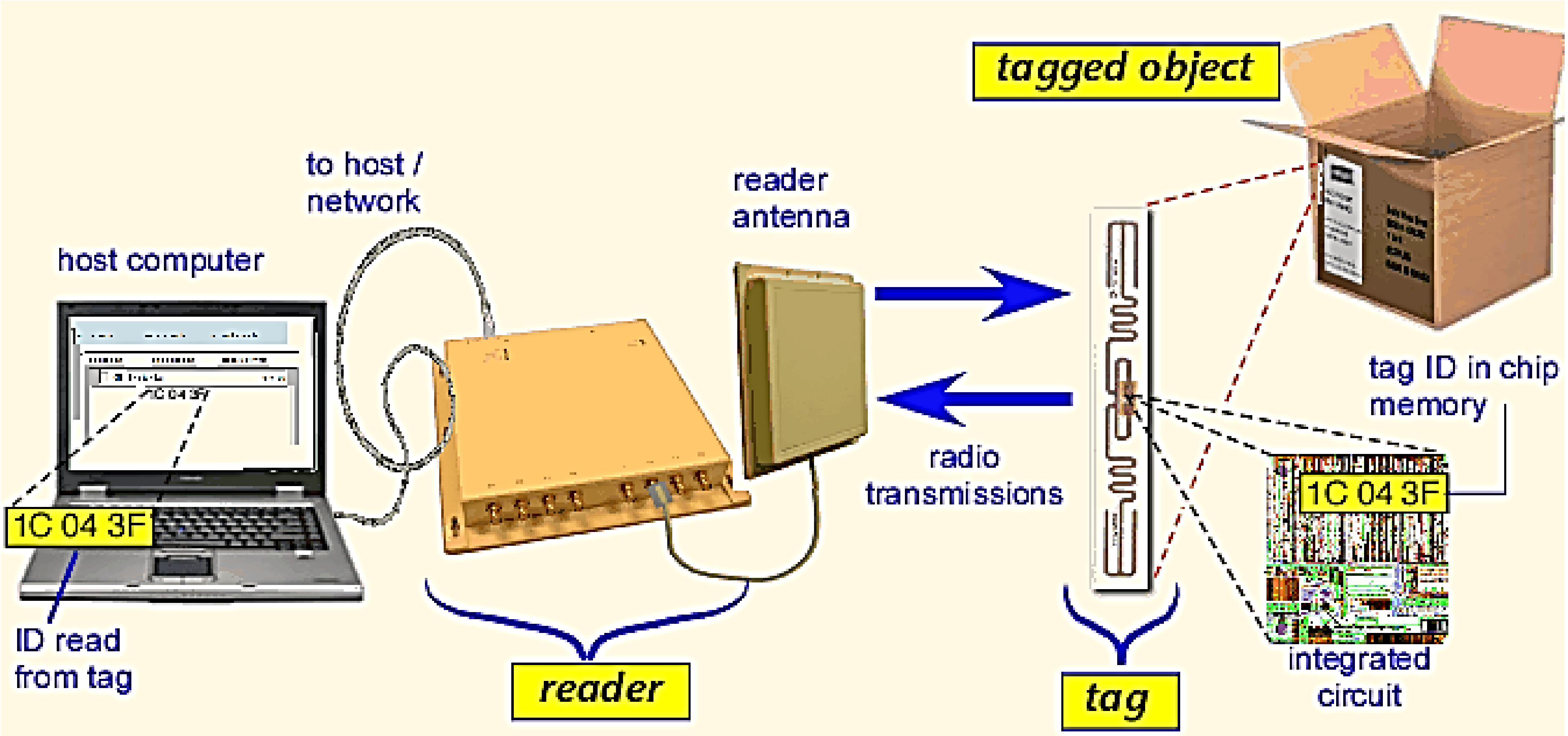
tagged object
to host / network
reader antenna
host computer
tag ID in chip memory
radio transmissions
1C 04 3F
1C 04 3F
ID read from tag
reader
tag
integrated circuit

Backscatter Communication

- **Backscatter** is the reflection of signals back towards their source.
  - In this scheme, two devices communicate using incident (or ambient) RF as the source of power.
  - Backscattering is achieved by changing the impedance of a receiver in the presence of an incident signal.
  - When waves encounter a new media that have different impedances, a part of the wave is reflected.
  - The reflection depends on the difference in the impedances.
- By modulating the impedance at the receiver port, one can control the scattered RF energy, hence enabling information transmission.

abc.cs.washington.edu/files/comm153-liu.pdf

RFID → IoT

- RFID
  - Uses radio waves for identifying or tracking the object.
  - Proven to be a simple and cost effective system
  - Tags are very cheap and is possible to be attached to everyday objects.

- RFID is considered a prerequisite of **Internet of Things.**

- **Example: RFID tags can be integrated with sensors**
  - When a reader reads a tag, the sensor information will be sent to the reader along with the identity of the object.

Some Examples

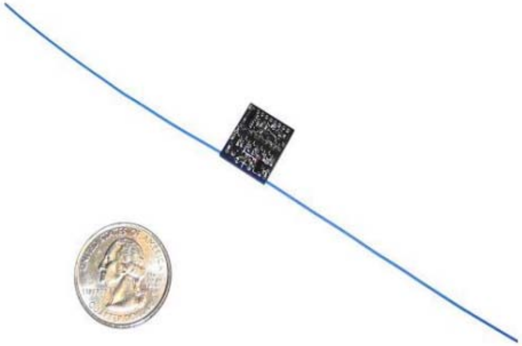

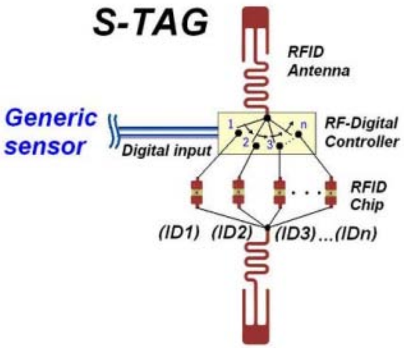
S-TAG
RFID Antenna
Generic sensor
Digital input
RF-Digital Controller
1
2
3
n
RFID Chip
(ID1) (ID2) (ID3)...(IDn)

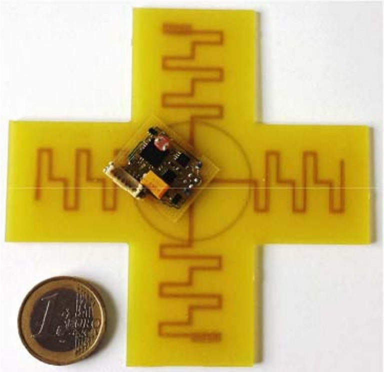

Discrete Element Based

WISP (Wireless Identification Sensing Platform)

Multi-Chip Based S-tag

Chips with I2C / SPI

SPARTACUS / RAMSES (Self-Powered Augmented RFID Tag for Autonomous Computing and Ubiquitous Sensing / RFID Augmented Module for Smart Environmental Sensing)

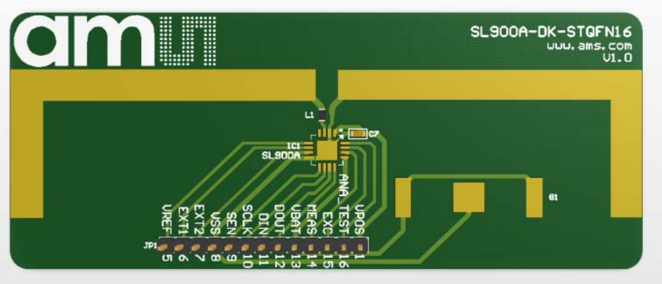
ams
SL900A-DK-STQFN16
www.ams.com
V1.0
L1
IC1
SL900A
C7
B1
JP1
VREF
EXT1
EXT2
VSS
SEN
SCLK
DIN
DOUT
VBAT
MEAS
EXC
ANA_TEST
VPOS

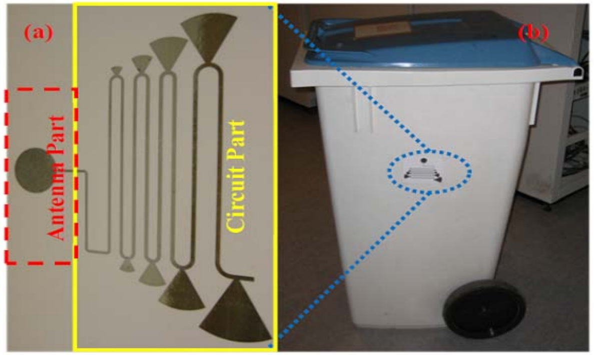
(a)
(b)
Antenna Part
Circuit Part

Single IC Based Sensing Tags

Printed Chipless RFID Tag

WISP

- Wireless Identification Sensor Platform (2009)
  - WISPs are a wireless, battery-free sensing and computation platform, powered by harvested energy from off-the-shelf UHF RFID readers.
  - To a RFID reader, a WISP is a EPC gen1 or gen2 tag; but inside the WISP, the harvested energy is operating a 16-bit general purpose microcontroller.
  - The microcontroller can perform computing tasks, including sampling sensors, and communicate to the RFID reader.
  - WISPs have been built with various sensors, WISPs can write to flash and perform cryptographic computations.

A collaboration between Intel Research Seattle and the University of Washington.

WISP 4.1DL
Intel Research Seattle

http://wisp.wikispaces.com/Wisp+4.1+DL

RFID Sensors (Products)

- ID operation is passive; yet most sensors require power sources

- Powercast has a wireless sensor that is battery-less. Uses RF energy harvesting.
- Harvesting schemes works at power as low as -12dBm. (RF-DC conversion efficiency above 40% only above -8dBm)
- Harvested power >0.4mW for RF in of >-1dBm.
- Multiple custom ICs and discretes

- Other suppliers include
  - Phase IV
  - RFID sensor systems
  - etc

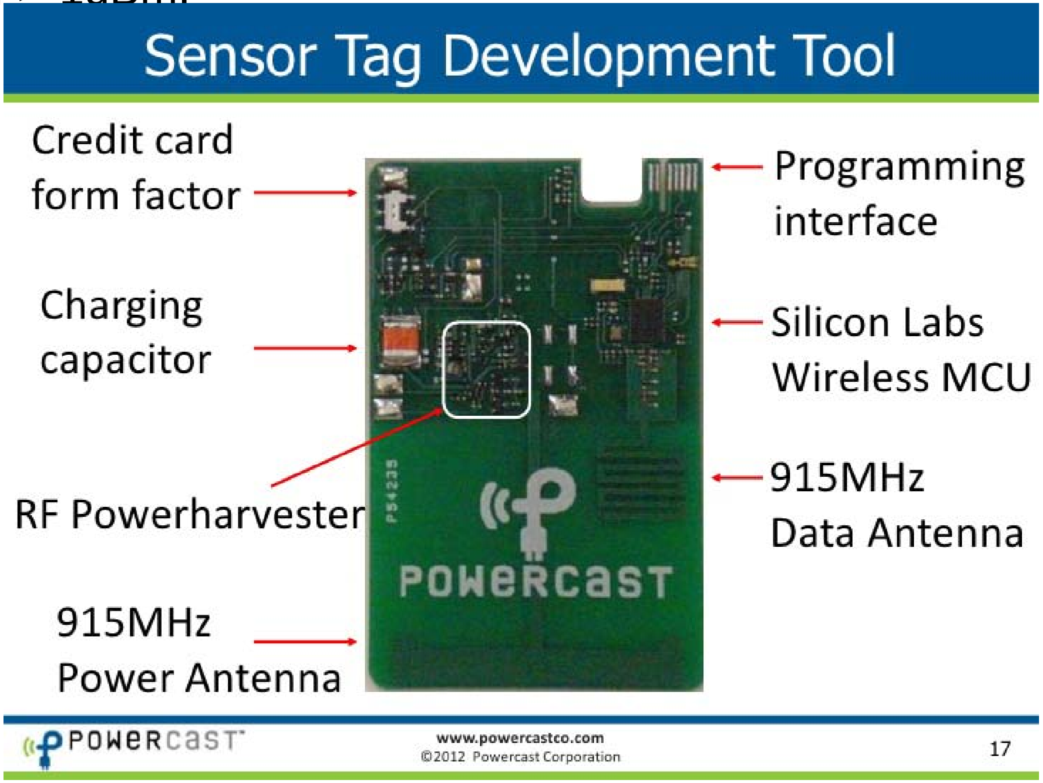

Sensor Tag Development Tool

Credit card form factor

Charging capacitor

RF Powerharvester

915MHz Power Antenna

Programming interface

Silicon Labs Wireless MCU

915MHz Data Antenna

POWERCAST™

www.powercastco.com

©2012 Powercast Corporation

17

http://www.powercastco.com/uhf-rfid-sensing-passive-rfid-wireless-sensor-tags/

Battery-less Wireless Terminals

- Most of our work in this direction was towards battery-less terminals
- Long life terminals without wiring
- These are useful when
  - Terminals are embedded within structures (or body)
  - Devices to be deployed in hostile environments
  - Use of battery is not allowed (potential cause for explosion)
- Other factors
  - Cost, weight, etc.
- Primary focus: use of radio frequencies (ambient/intentional)

Ambient RF Sources

- Several sources:
  - WiFi Access points (mW) [2.4/.6GHz]
  - Cellular Tower (W) [900/1800 MHz]
  - TV Broadcast (MW) [150-450MHz]
  - FM broadcast (kW) [90-108MHz]
  - AM Radio broadcast (kW) [<1MHz]

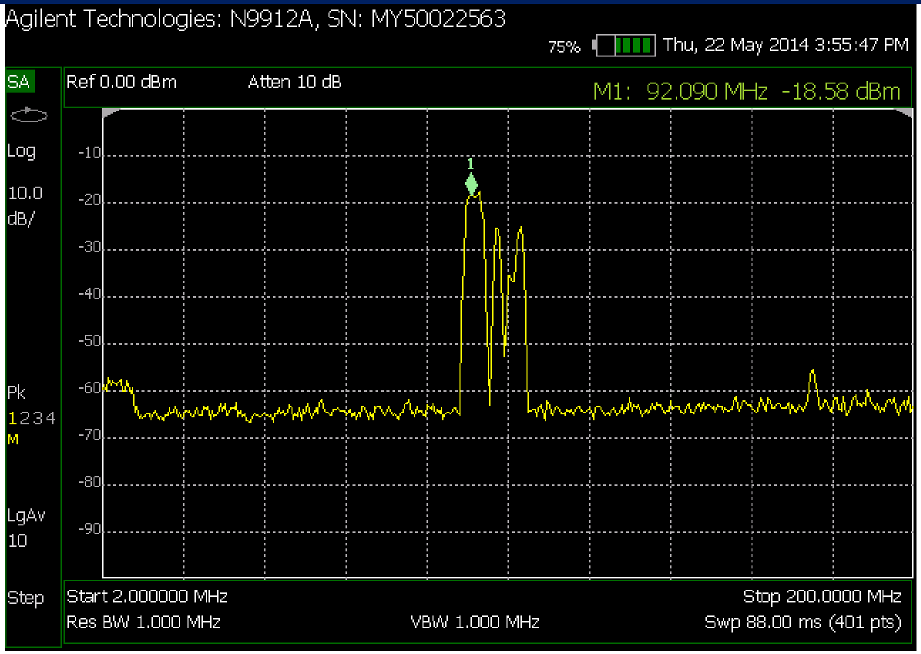
Agilent Technologies: N9912A, SN: MY50022563

75% Thu, 22 May 2014 3:55:47 PM

Ref 0.00 dBm Atten 10 dB M1: 92.090 MHz -18.58 dBm

Start 2.000000 MHz Stop 200.0000 MHz

Res BW 1.000 MHz VBW 1.000 MHz Swp 88.00 ms (401 pts)

- In general
  - Lower frequencies help non-line-of sight propagation
  - Power availability from ambient sources is limited and varies from place to place.

- Note
  - Unlike other sources, most practical RF harvesters (eg in RF ID) depend on **intentionally generated energy**.
  - This is called **wireless power transfer (WPT)** in the conventional RF/Microwave parlance.

Wireless Communication System

- Power transfer scheme is no different!!

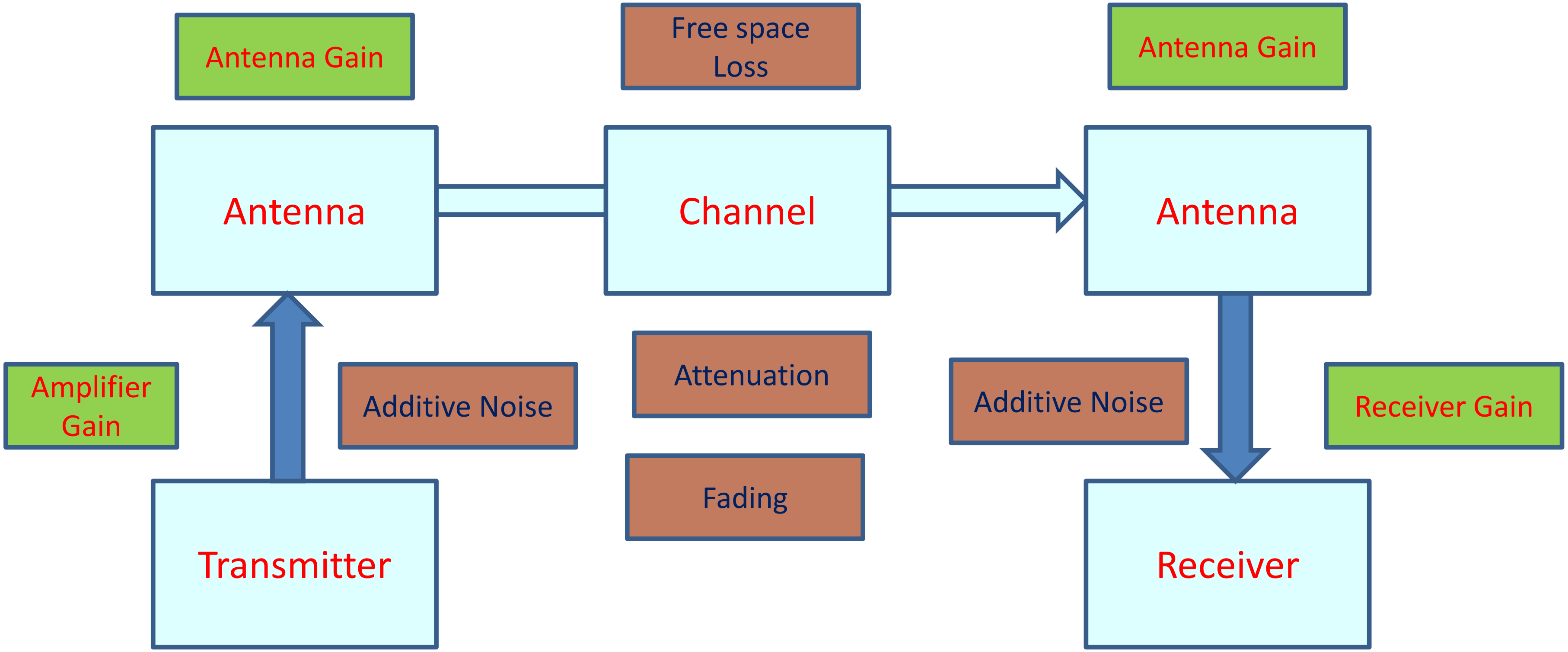
Antenna Gain
Free space Loss
Antenna Gain
Antenna
Channel
Antenna
Amplifier Gain
Additive Noise
Attenuation
Additive Noise
Receiver Gain
Fading
Transmitter
Receiver

Antenna Fundamentals: Directivity

This is the ratio of the radiation intensity ***in a given direction*** *to the radiation intensity averaged over all direction*

Average radiation intensity, $U_0 = \dfrac{P_{rad}}{4\pi}$

Directivity,

$$D(\theta,\phi) = \frac{U(\theta,\phi)}{U_0} = \frac{4\pi U(\theta,\phi)}{P_{rad}}$$

- If direction is not specified, it implies the direction of maximum radiation intensity

$$D_{\max} = \frac{4\pi U_{\max}}{P_{rad}} \qquad D_{dB} = 10\log D$$

Maximum directivity and Maximum effective area

The radiated power density by a transmitter at a distance R

$$W_t = \frac{P_t D_t}{4\pi R^2}$$

Power received

$$P_r = W_t . A_r = \frac{P_t D_t A_r}{4\pi R^2}$$

$$or,\; D_t . A_r = \frac{P_r}{P_t} . 4\pi R^2$$

By reversing the transmission direction

$$D_r A_t = \frac{P_r}{P_t} \cdot 4\pi R^2$$

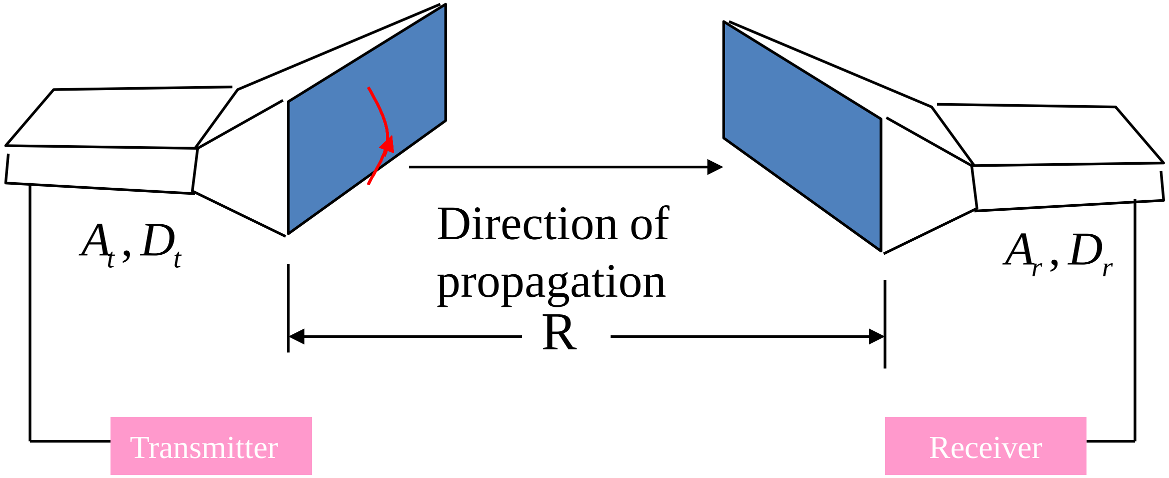
A_t, D_t
Direction of propagation
R
A_r, D_r
Transmitter
Receiver

$$\therefore \frac{D_t}{A_t} = \frac{D_r}{A_r}$$

this can be generalized by

$$\frac{D_1}{A_1} = \frac{D_2}{A_2} = k = \frac{4\pi}{\lambda^2}$$

$$if,\; D_0 = \text{max. directivity}$$

$$A_{em} = \text{max. effective area}$$

$$D = \frac{4\pi A_{em}}{\lambda^2}$$

Frii's Transmission Equation

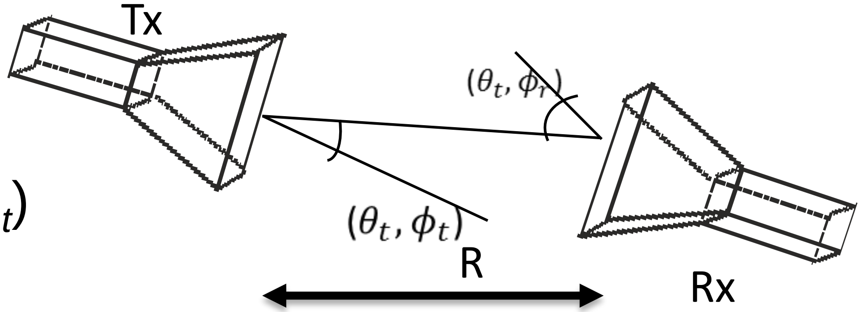

- The radiation intensity for an isotropic radiator is $W_0 = \frac{P_t}{4\pi R^2}$
- For an antenna of gain G_t (or directivity D_t)

$$W_t = \frac{P_t G_t(\theta_t,\phi_t)}{4\pi R^2} = e_t \frac{P_t D_t(\theta_t,\phi_t)}{4\pi R^2}$$

- The effective aperture of a receiving antenna is given by $A_r = e_r D_r(\theta_r, \phi_r)\frac{\lambda^2}{4\pi}$
- Therefore,

$$P_r = e_r D_r(\theta_r, \phi_r)\frac{\lambda^2}{4\pi} W_t = e_t\, e_r D_t(\theta_t, \phi_t) D_r(\theta_r, \phi_r)\frac{\lambda^2}{(4\pi R)^2}\, |\hat{\boldsymbol{\rho}}_t . \hat{\boldsymbol{\rho}}_r|^2$$

$$\frac{P_r}{P_t} = e_t e_r D_t(\theta_t, \phi_t) D_r(\theta_r, \phi_r)\frac{\lambda^2}{(4\pi R)^2}\, |\hat{\boldsymbol{\rho}}_t . \hat{\boldsymbol{\rho}}_r|^2$$

$$\frac{P_r}{P_t} = e_{cdt} e_{cdr}(1 - |\Gamma_t|^2)(1 - |\Gamma_r|^2) D_t(\theta_t, \phi_t) D_r(\theta_r, \phi_r)\frac{\lambda^2}{(4\pi R)^2}\, |\hat{\boldsymbol{\rho}}_t . \hat{\boldsymbol{\rho}}_r|^2$$

- When the antennas are pointing towards each others' peak radiation direction,

$$\frac{P_r}{P_t} = G_{0t} G_{0r}\left(\frac{\lambda}{4\pi R}\right)^2$$

Note that includes a loss factor (usually called **Free space Loss factor)**
Does not include dissipation/attenuation in medium; caused by spreading

Some numbers on Radiative form of WPT…

- Practical systems will have
  - Operational frequencies in ISM bands.
  - Most terminals are compact.
  - Antenna efficiency is compromised.
  - Nearly isotropic radiations expected.

- Main bottleneck is the physical limits in transmission.

$$P_r = P_t{*}G_t{*}G_r{*}(\lambda/4\pi r)^2$$

  - At 1 GHz (λ=30cm) r=1m; Antenna gain @0dBm, free space loss factor is about **0.06%**
  - Even with a moderate gain transmitter antenna (6dBi) power received @1m for 1W transmission, is just 2mW.
    - Drops to 23μW at 10m !!
    - **The voltage of the signal is low!!**
- In radiative power transfer, Distance from transmitter is a major concern.

Some questions addressed in our work

- **Harvesting of ambient radiations or Radiative transfer of energy addressed**
  - Is it possible to harvest the RF energy from base stations
  - Are there other viable sources of RF energy

- Can low power communication systems be designed to operate entirely from harvested energy
  - Integrate sensors, control, etc

- Can we use RF EH/ WPT to increase the range of backscatter communication (RFID scenario)

Design of Rectifiers

Required for converting incoming RF into DC power.

The challenge lies in maximizing the power conversion efficiency for low input power and minimizing the dimensions.

- RF to DC conversion by rectification of the incident RF signal by a ***Schottky diode***
  - Most diodes have a finite cut-in voltage
  - Diode is a non-linear device (performance depends on current or load)
  - Impedance matching required between antenna and diode
  - In most cases, the input voltage needs to be boosted

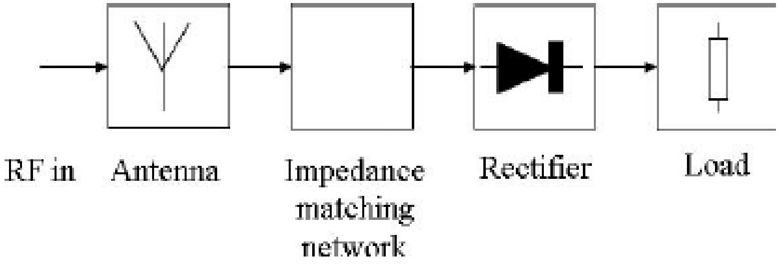
RF in
Antenna
Impedance matching network
Rectifier
Load

Conceptual diode based Rectenna

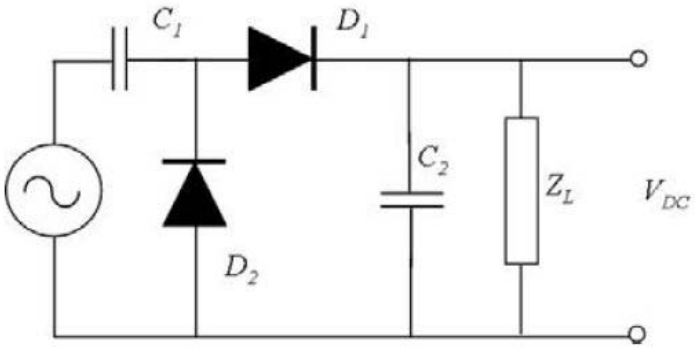
C_1
D_1
C_2
D_2
Z_L
V_{DC}

Voltage doubler

Voltage magnification in Matching circuit

- Matching circuit is required to provide impedance match between antennas (50Ω typical) to diode terminated with high impedance load (capacitor and/or high R in parallel).
  - LC matching networks provide voltage magnification.
  - This helps the diode conduct a good fraction of half cycle.
- The higher the voltage across the diodes, the more efficient the rectifier gets.
  - In practice Q is limited
  - Applications requiring higher voltages, a voltage multiplier configuration is used.

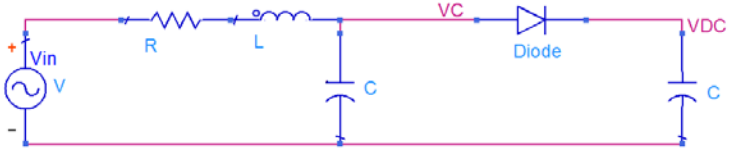

$$V_C = V_{in} \times \frac{\frac{1}{j\omega C}}{R + j\omega L + \frac{1}{j\omega C}}$$

At resonance,

$$V_C = V_{in} \times \frac{1}{j\omega_0 CR} = -jQV_{in}$$

Tuned Rectifier at RF

- A tuned rectifier implemented using discrete components

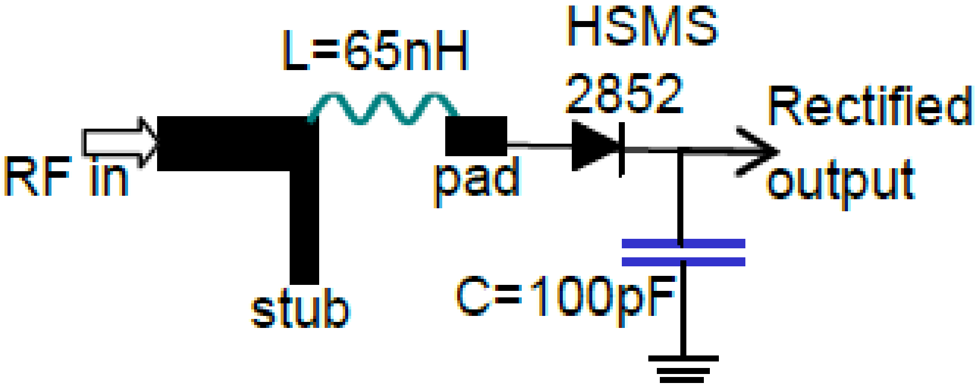

P_in→	**-10dBm**	**-13dBm**	**-16dBm**	**-20dBm**	**-25dBm**
Freq. ↓					
930MHz	917mV	664mV	469mV	281mV	131mV
945MHz	1016mV	736mV	515mV	300mV	132mV
955MHz	1038mV	747mV	513mV	289mV	122mV
960MHz	1032mV	736mV	499mV	276mV	114mV
Peak efficiency	51%	47%	39%	33%	20%

K. J. Vinoy, T. V. Prabhakar, A Universal Energy Harvesting Scheme for Operating Low-Power Wireless Sensor Nodes Using Multiple Energy Resources, pp. 453-466, Micro and Smart Devices and Systems, Springer 2014.

Typical Performance of Rectifier

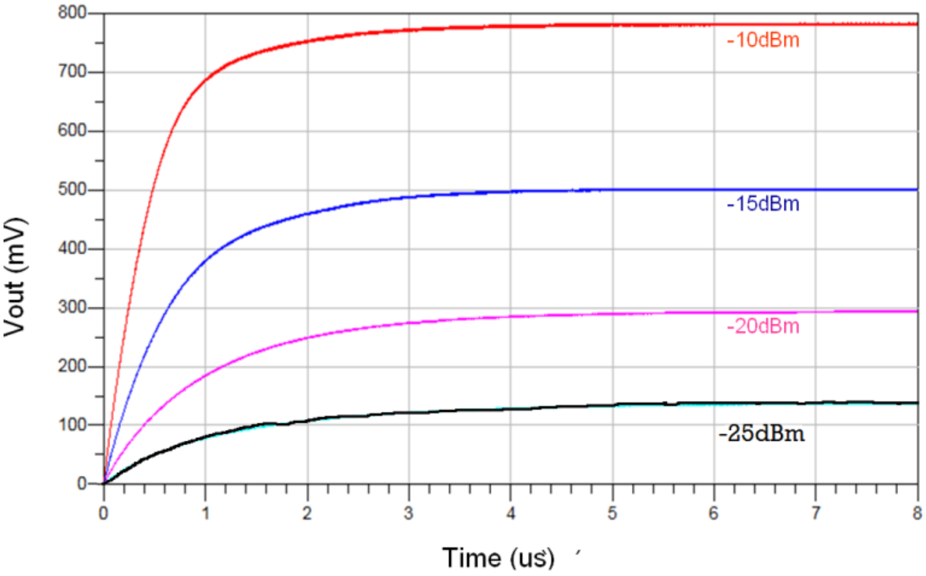

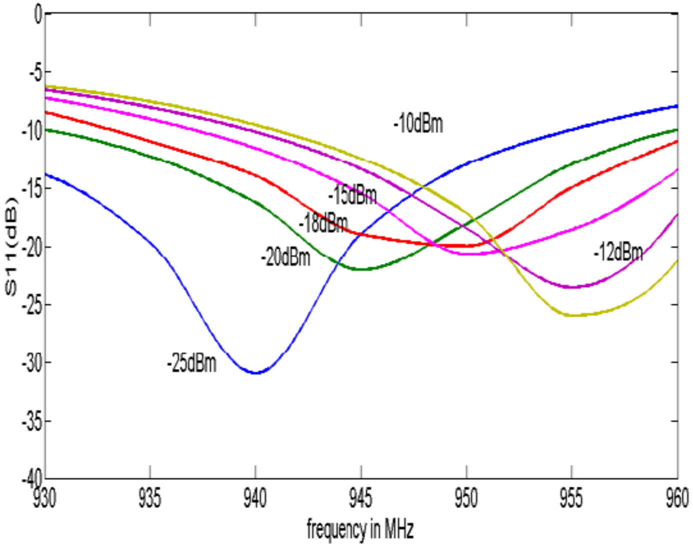

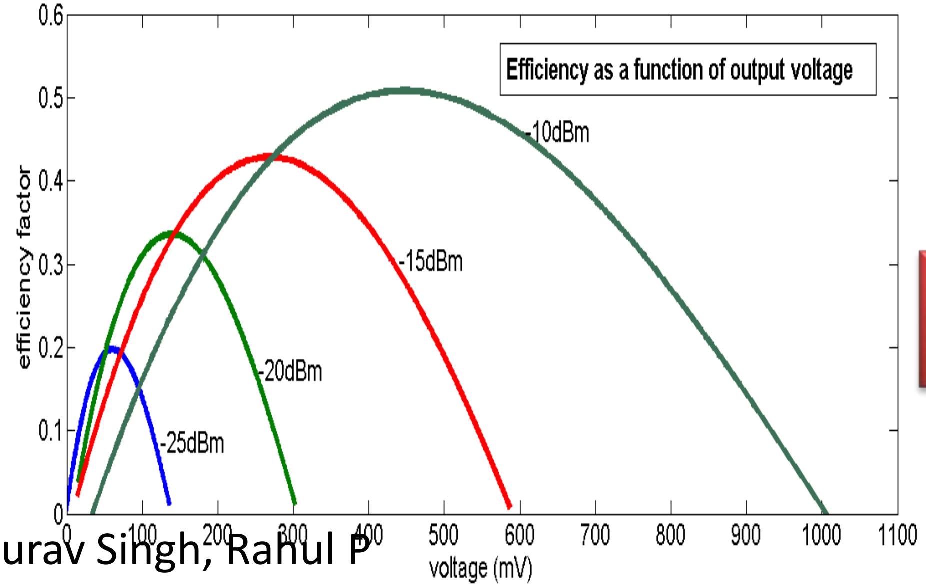

RF-DC Conversion efficiency depends on various conditions

Rectifier Circuit using 4 diodes

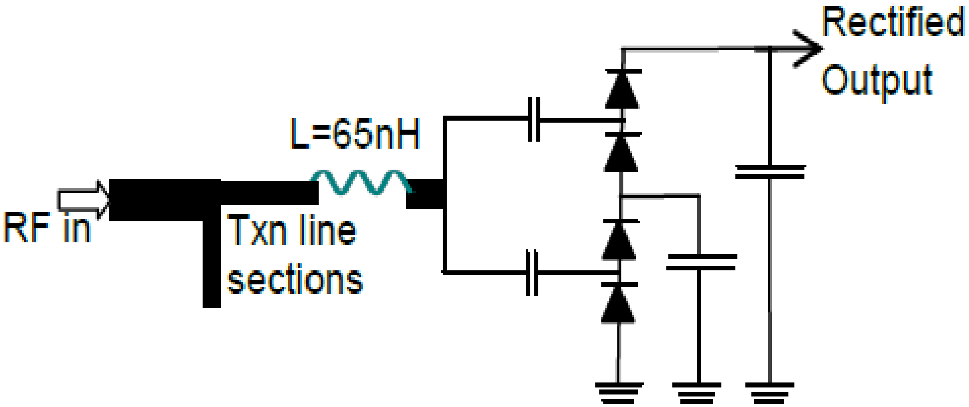
Rectified
Output
L=65nH
RF in
Txn line
sections

Power level (dBm)	Charging time (ms)	Efficiency (%)
0	40	64.11
-2	55	63.77
-3	67.5	63.5
-5	90	63
-7	230	59.89
-10	370	56.78
-12	500	53.89
-15	900	45
-18	2000	20.56

K. J. Vinoy, T. V. Prabhakar, A Universal Energy Harvesting Scheme for Operating Low-Power Wireless Sensor Nodes Using Multiple Energy Resources, pp. 453-466, Micro and Smart Devices and Systems, Springer 2014.

Aditya Mitra, Chaithanya (2011-12)

1: Scavenging Mobile Tower Radiations

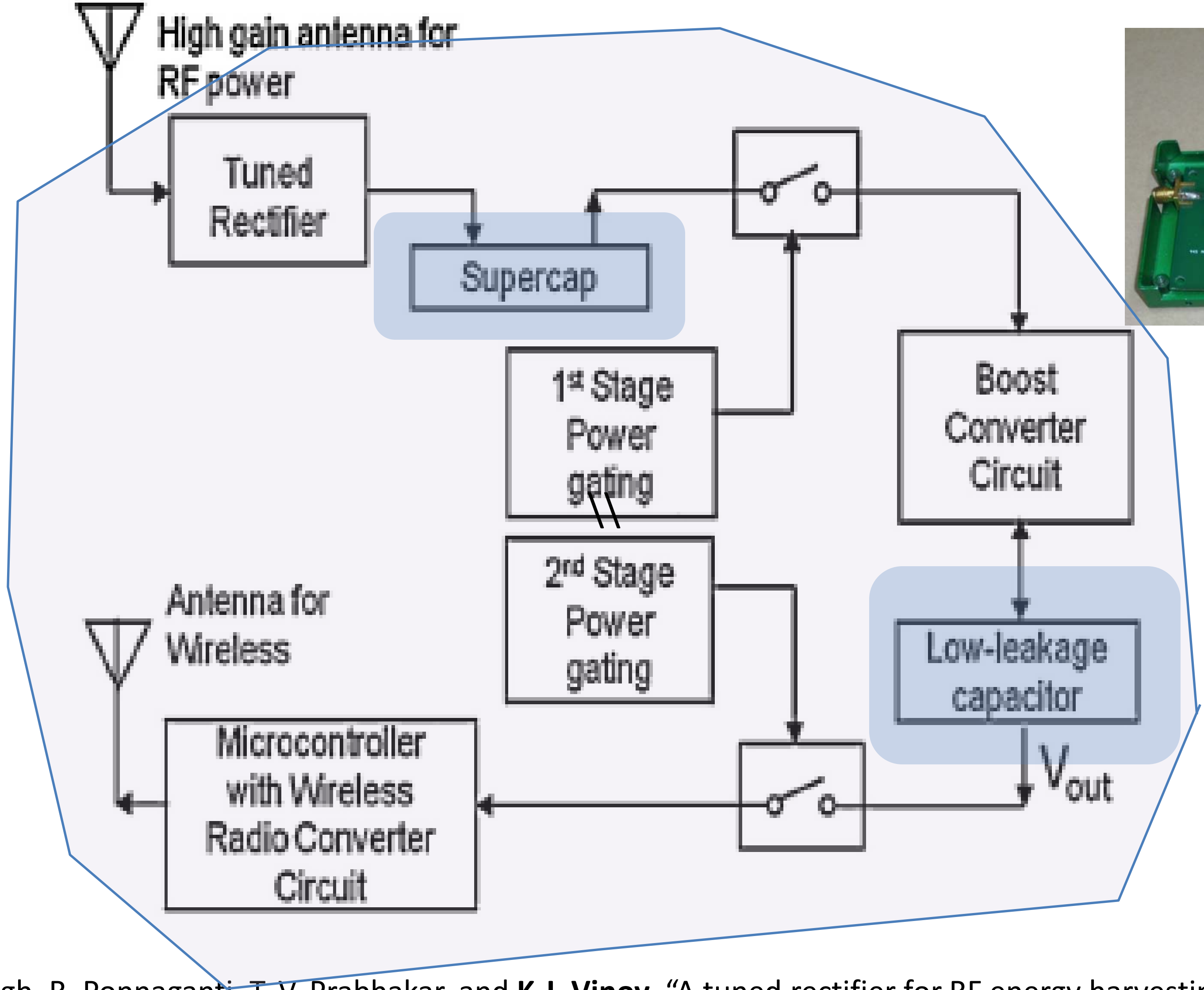
High gain antenna for RF power
Tuned Rectifier
Supercap
1st Stage Power gating
2nd Stage Power gating
Boost Converter Circuit
Low-leakage capacitor
V_{out}
Antenna for Wireless
Microcontroller with Wireless Radio Converter Circuit

G. Singh, R. Ponnaganti, T. V. Prabhakar, and **K.J. Vinoy**, “A tuned rectifier for RF energy harvesting from ambient radiations,” Int. J. Electronics & Communications, vol. 67, no. 7, pp. 564-569, July 2013

Gaurav Singh, Rahul P (2010-11)

3.15
POWER
COM

Gaurav Singh, Rahul P (2010-11)

Characterization in Lab

- Using various antennas

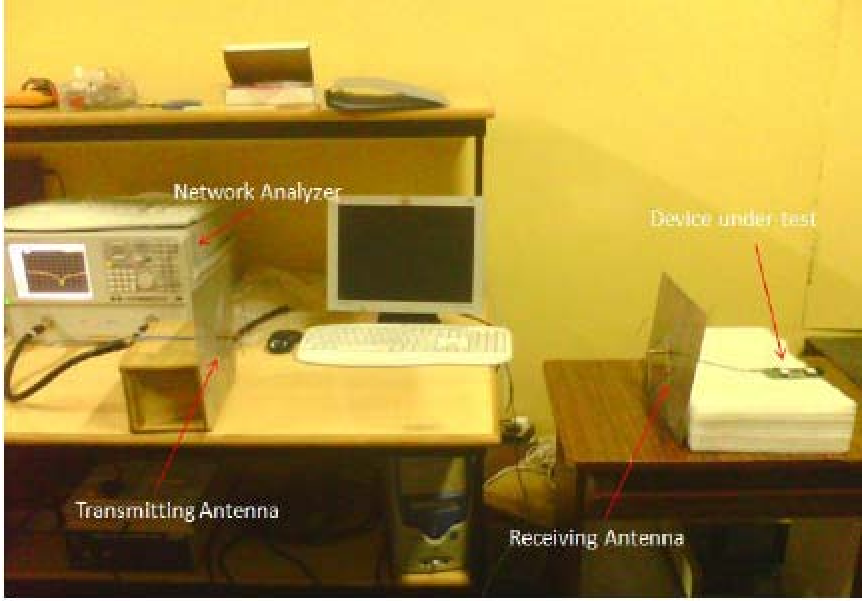
Network Analyzer
Device under test
Transmitting Antenna
Receiving Antenna

Distance from Transmitter [m]	1.5	2	2.5	3
Power received by dipole antenna [dBm]	-20.5	-22.1	-23.9	-25.2
Calculated power density [uW/cm^2]	0.078	0.055	0.035	0.03
Power received by patch antenna [dBm]	-15.1	-16.1	-17.6	-19.2
Transmit interval [mm:ss]	07:26	12:13	25:00	never
Power received by biquad antenna [dBm]	-11.8	-13.2	-14.9	-15.9
Transmit interval [mm:ss]	02:20	03:25	7:10	10:33

G. Singh, R. Ponnaganti, T. V. Prabhakar, and **K.J. Vinoy**, “A tuned rectifier for RF energy harvesting from ambient radiations,” Int. J. Electronics & Communications, vol. 67, no. 7, pp. 564-569, July 2013

Gaurav Singh, Rahul P (2010-11)

Incident power above -18dBm required for operation

2. Universal Energy Harvesting Platform

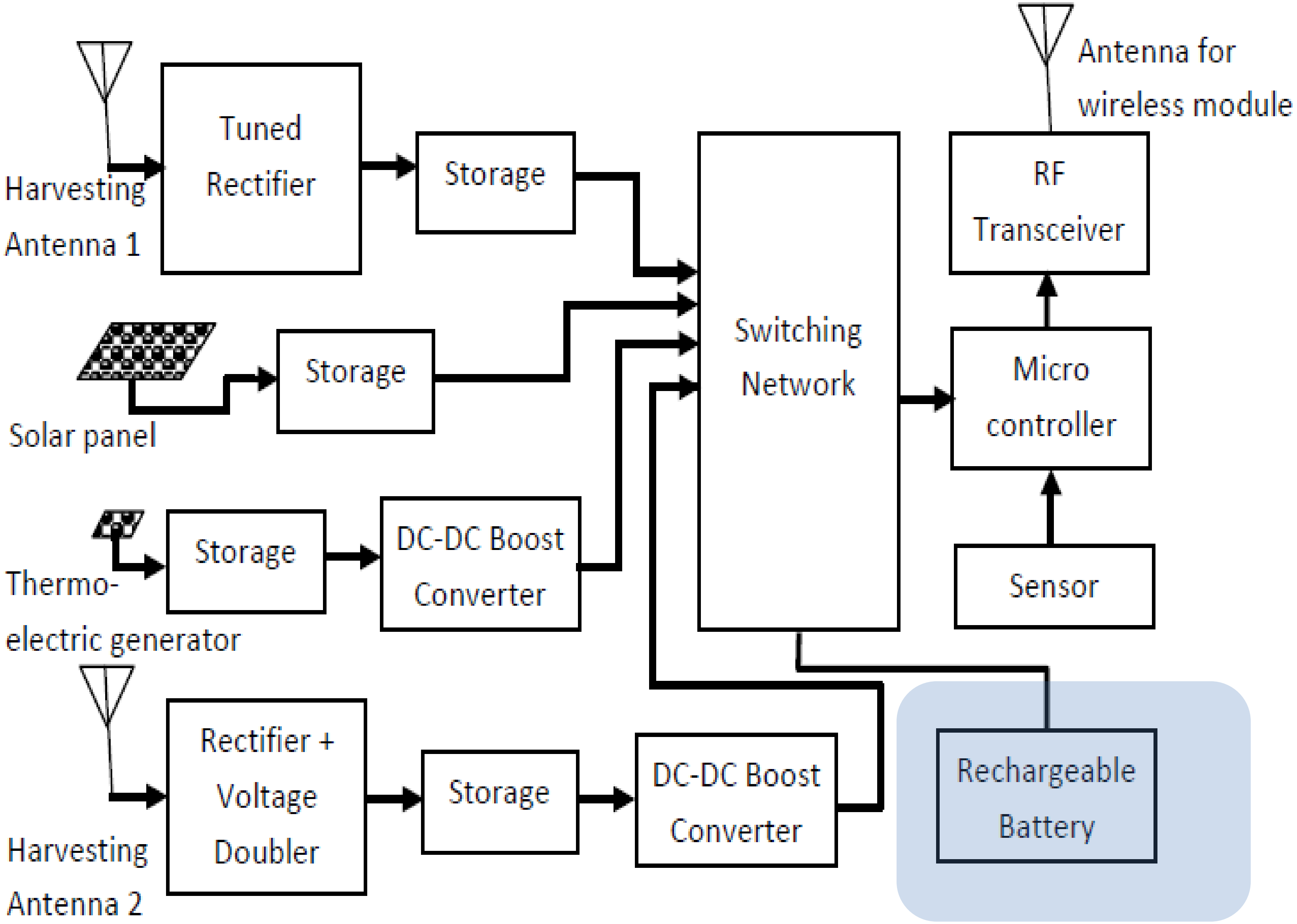
Antenna for
wireless module
Harvesting
Antenna 1
Tuned
Rectifier
Storage
RF
Transceiver
Switching
Network
Storage
Solar panel
Micro
controller
Storage
DC-DC Boost
Converter
Thermo-
electric generator
Sensor
Rectifier +
Voltage
Doubler
Storage
DC-DC Boost
Converter
Rechargeable
Battery
Harvesting
Antenna 2

Only a small thin film battery is used

UEHP: Performance with different sources

Solar		RF		TEG	
Light Intensity (Lux)	Duty Cycle of operation (s)	Power Level (dBm)	Duty Cycle of Operation (s)	Temperature Differential (°C)	Duty Cycle of Operation (s)
1000	7	0	3	55	9
300	11	-5	6	45	13
200	**20**	**-7**	20	35	240
100	42	-10	50	-	-
-	-	-12	240	-	-

An incident RF power of -7dBm (~0.2mW) performs similarly as at low light PV.

An appropriately oriented 20mW source with a high gain antenna (~10dB) can reach this RF power at a low gain rectenna (eg using PIFA) at 1 m distance.
Power levels within emission guidelines…

K. J. Vinoy, T. V. Prabhakar, A Universal Energy Harvesting Scheme for Operating Low-Power Wireless Sensor Nodes Using Multiple Energy Resources, pp. 453-466, Micro and Smart Devices and Systems, Springer 2014.

Aditya Mitra, Chaithanya (2011-12)

Other Possibilities using Wireless Power Transfer

- Power transfer by radiation is not efficient
- Waveguiding systems can ensure better transmission of power
  - Loss in waveguide is a small fraction of a dB/m (~0.2dB/m)
  - Metal ducts may carry higher order modes with higher losses
  - Extended to conducting ducts, Tunnels, mine shafts etc with some compromise
- Other possibilities
  - Surface wave
  - Focusing of fields
- Empty enclosures with metallic walls
  - Containers, tanks, airplane cabin, trains, etc
  - Other objects in the path may reduce the efficiency!

3: Antenna for RFID sensors

- Work involved design of antenna for RF harvesting sensors
  - These fuel level sensors to be deployed in a fuel tank of aircrft.
  - Optimization of design should focus on efficiency
  - High gain or directivity is not required.

- EH platform to be used with RFID sensors deployed inside fuel tank

- Requirements/Assumptions:
  - Incident energy is of random polarity and direction.
  - Operating frequency is 902MHz-928MHz.
  - Antenna must operate in air (relative permittivity = 1) and fluid (relative permittivity = 2.1)
  - Dimensions of planar antenna board:
    - Target dimensions: 3 in. x 2 in.
    - Maximum dimensions: 6 in. x 4 in.

Antenna Design dimensions

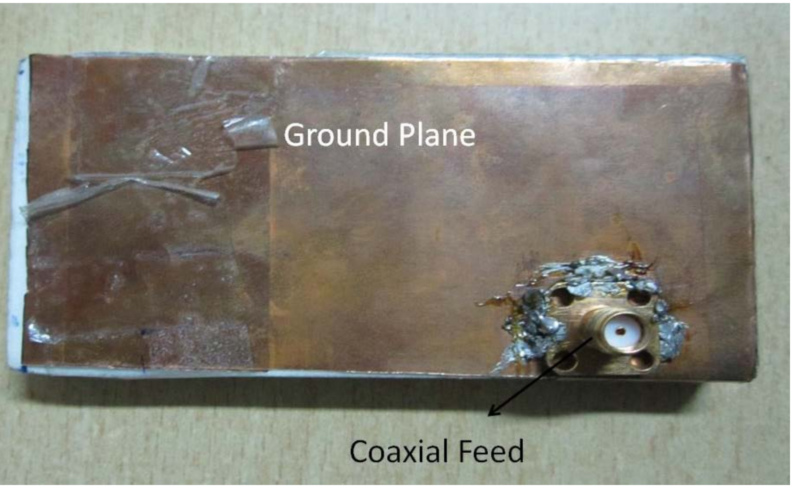
Ground Plane
Coaxial Feed

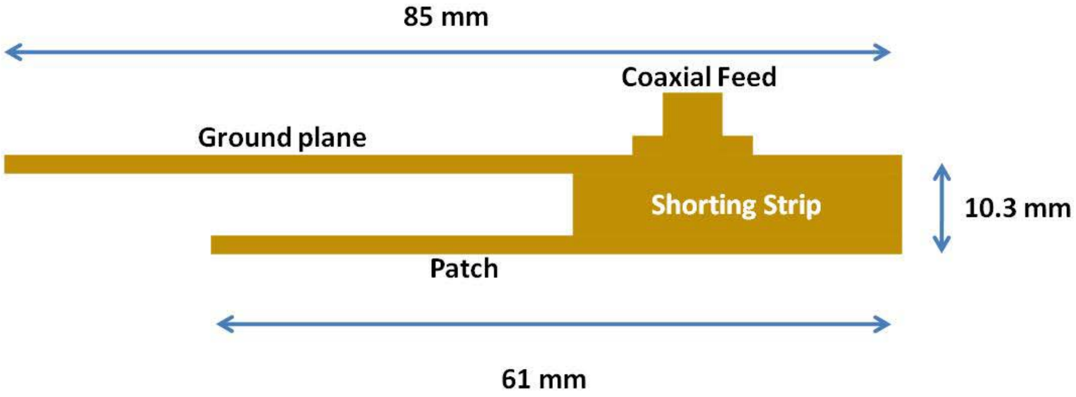
85 mm
Coaxial Feed
Ground plane
Shorting Strip
10.3 mm
Patch
61 mm

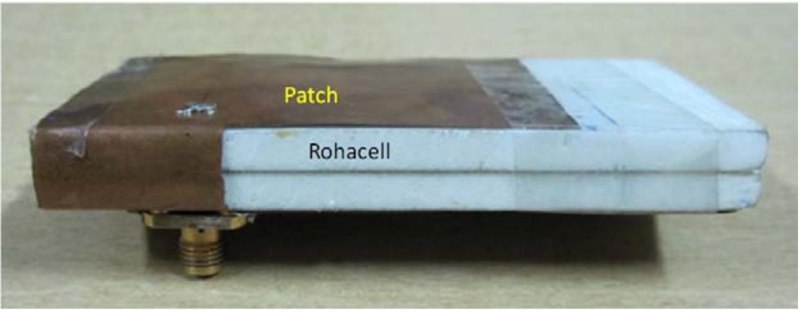
Patch
Rohacell

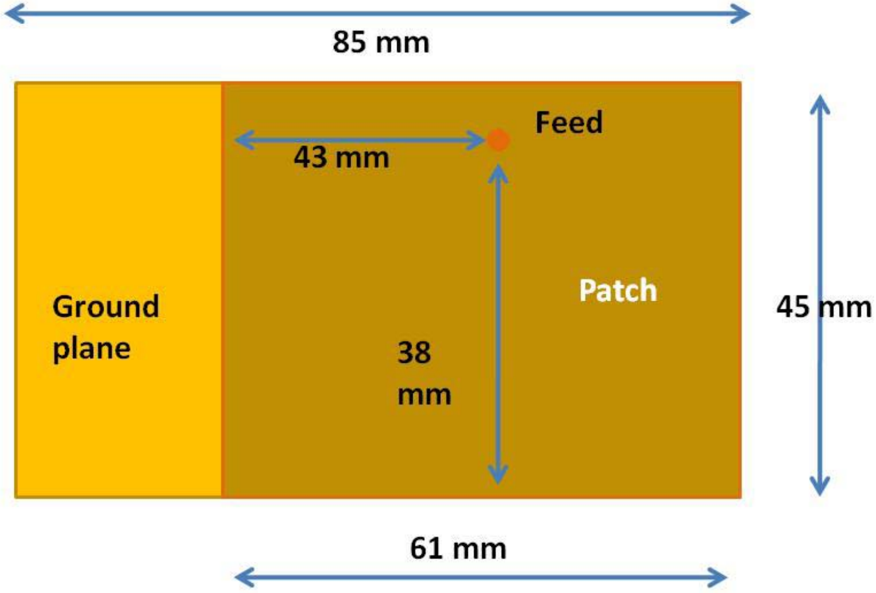
85 mm
Feed
43 mm
Ground plane
Patch
38 mm
45 mm
61 mm

Vivekanand M, Harikiran M

Measurements at Boeing (Nov 2013)

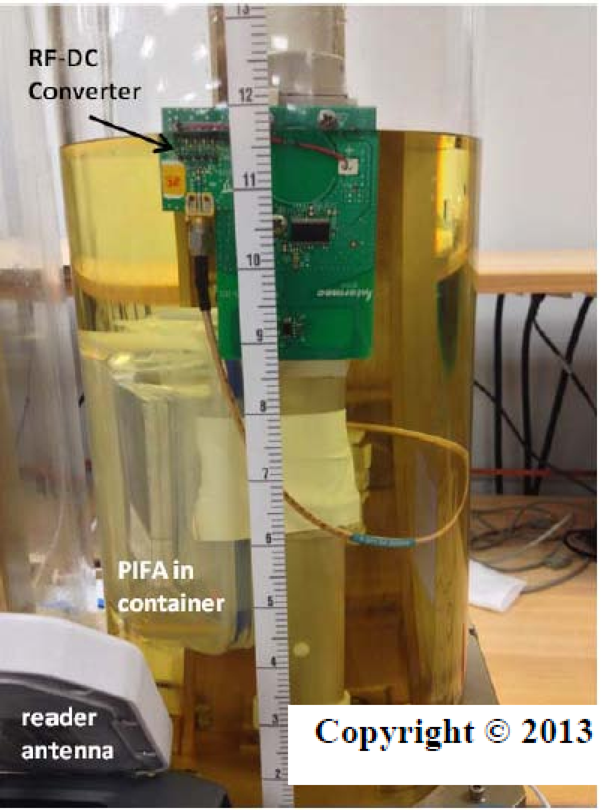
RF-DC
Converter
PIFA in
container
reader
antenna

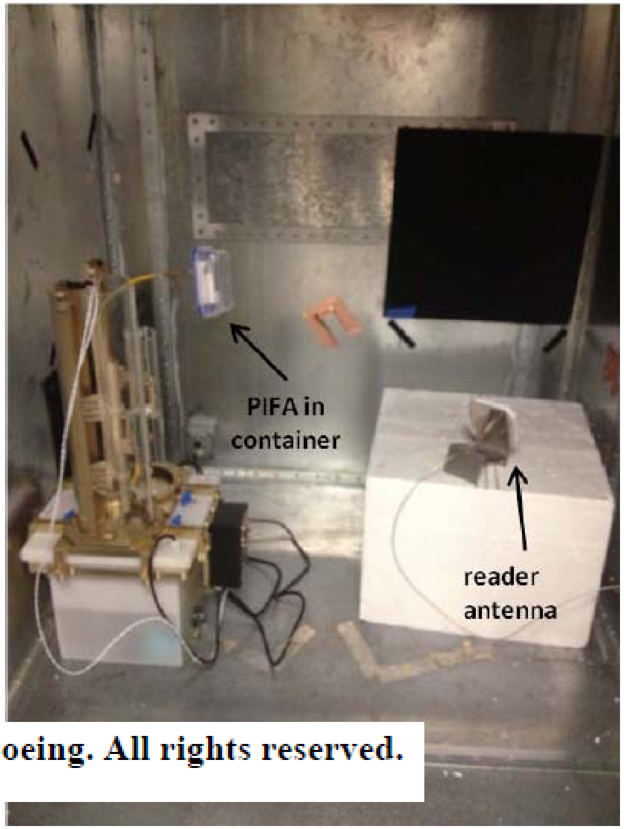
PIFA in
container
reader
antenna

Copyright © 2013 Boeing. All rights reserved.

After Integrating with Sensor and RFID board;

Measured in a room and reverberation chamber

Resolution of fluid height measurement to within 0.25”.
1W maximum transmit power.
Uses a modified reader protocol.

A. Robb, J. Bommer, R. Martinez, J. Harrigan, S. Ramamurthy , H. Muniganti, V. Mannangi, and KJ Vinoy, “Wireless Aircraft Fuel Quantity Indication System,” 2014 IEEE Sensors Applications Symposium - , Feb 18-20, 2014, Queenstown, Newzealand.

4. RFID Integrated with Sensor

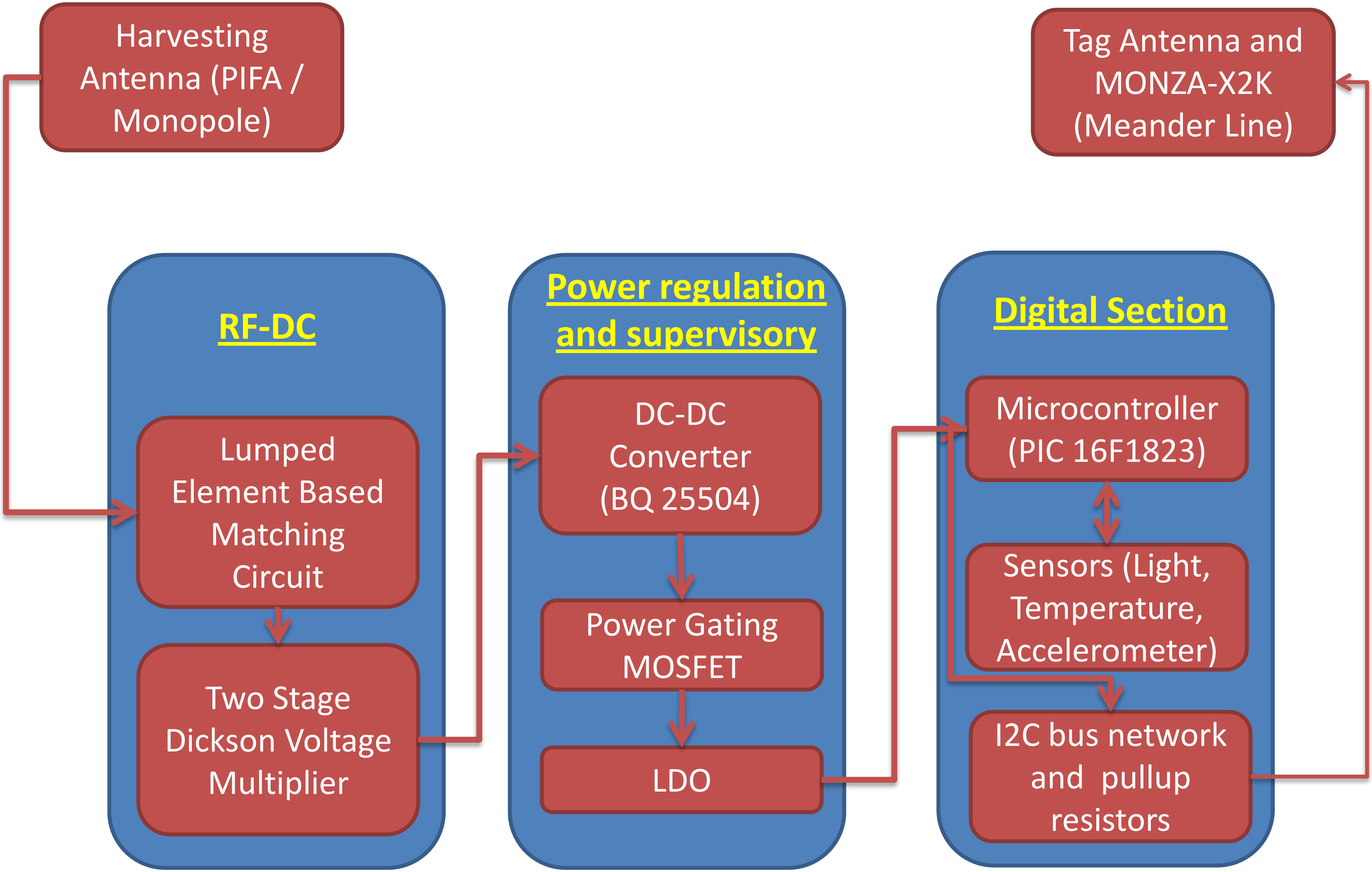

Harvesting Antenna (PIFA / Monopole)

Tag Antenna and MONZA-X2K (Meander Line)

RF-DC

Lumped Element Based Matching Circuit

Two Stage Dickson Voltage Multiplier

Power regulation and supervisory

DC-DC Converter (BQ 25504)

Power Gating MOSFET

LDO

Digital Section

Microcontroller (PIC 16F1823)

Sensors (Light, Temperature, Accelerometer)

I2C bus network and pullup resistors

Parts of Fabricated System

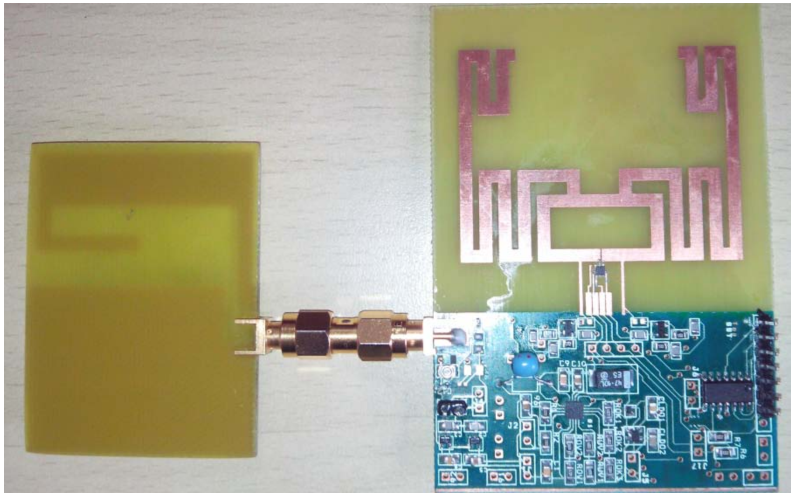

PIFA CONNECTED SENSING TAG

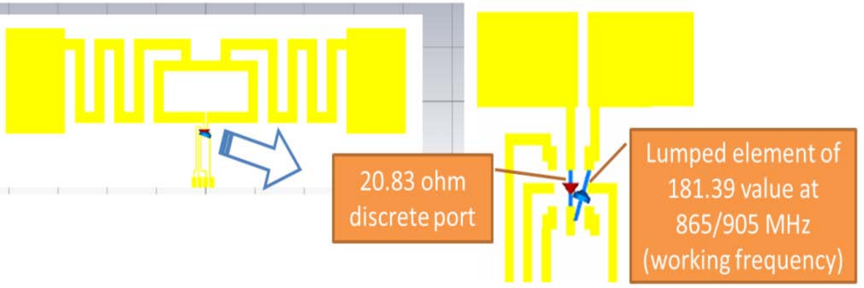
20.83 ohm discrete port
Lumped element of 181.39 value at 865/905 MHz (working frequency)

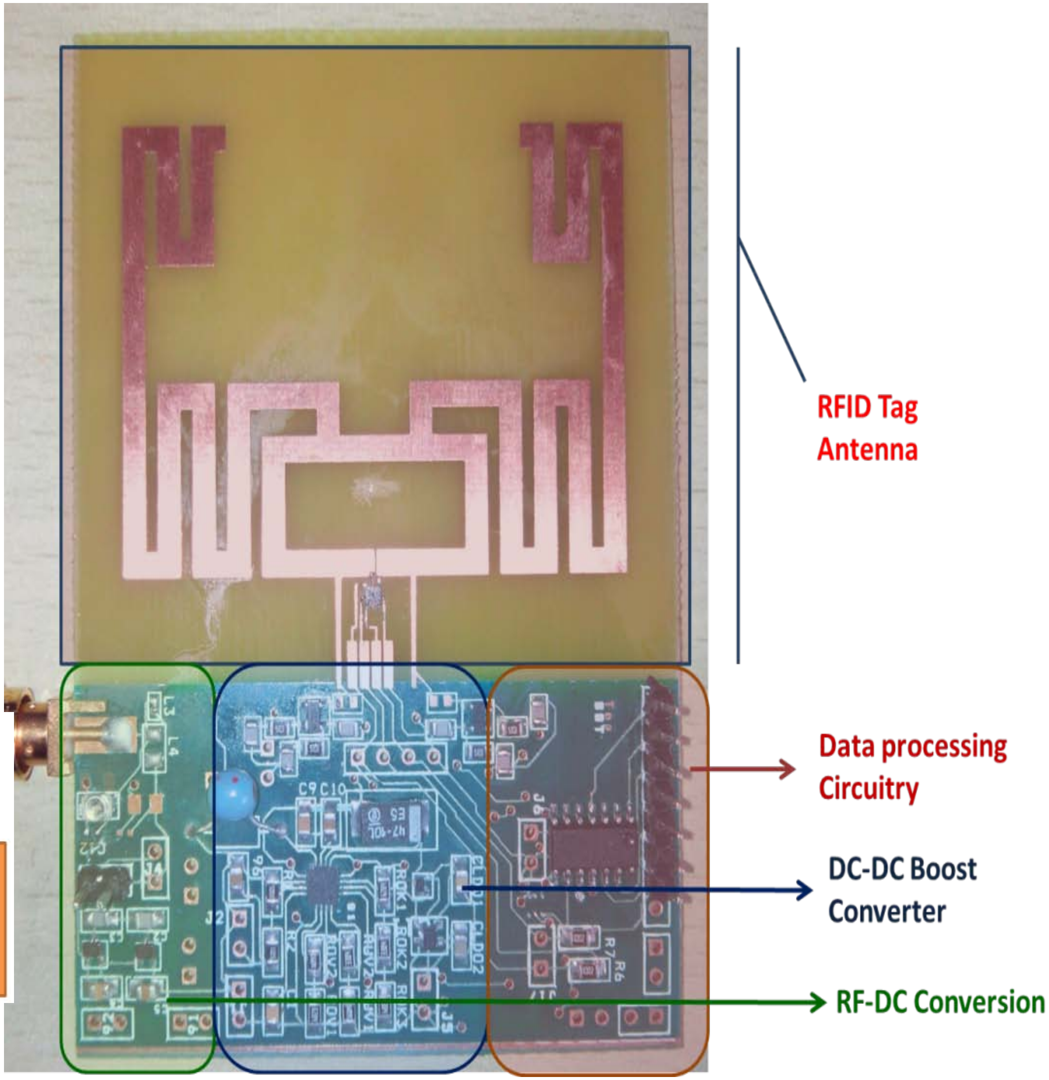
RFID Tag Antenna
Data processing Circuitry
DC-DC Boost Converter
RF-DC Conversion

Sandeep Rana 2014-15

Sandeep Rana, TV Prabhakar, KJ Vinoy, An Efficient Architecture for Battery-less Terminals for Internet of Things, Applied Computational Electromagnetic Conference, Guwahati, Dec 28-21, 2015

Characterization of Performance

Source	Power
RFID Reader	30 dbm
Circularly polarized antenna	8 dbi
Polarization loss	3 dbi
Monopole Antenna	5 dbi
PIFA antenna	1 dbi
Meander Antenna	0.4 dbi

Tag Antenna	EIRP (dbm) + Gr	Rg Expected (-10 dbm) and 50 % overall efficiency	Range Achieved
Monopole	40	7.5 mtr	7 mtr
PIFA	36	5 mtr	5.5 mtr
Meander Line	35.4	4.5 mtr	4.5 mtr

$$\lambda = 34.5\ cm \qquad \text{Prx} = \text{EIRP} * \text{Gr} * \left(\frac{\lambda}{4*\pi*r}\right)^2$$

Efficiency worked out for -10 dbm
RF-DC efficiency – 20 % and DC-DC efficiency – 80%
Overall efficiency – 16%

Sandeep Rana 2014-15

5. Harvesting at 2.4GHz

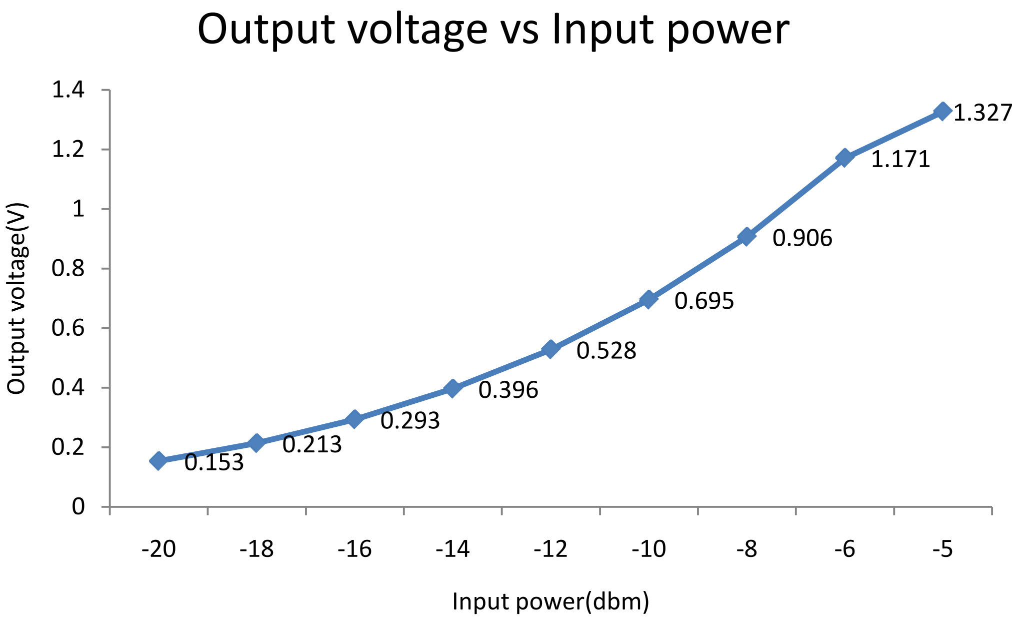

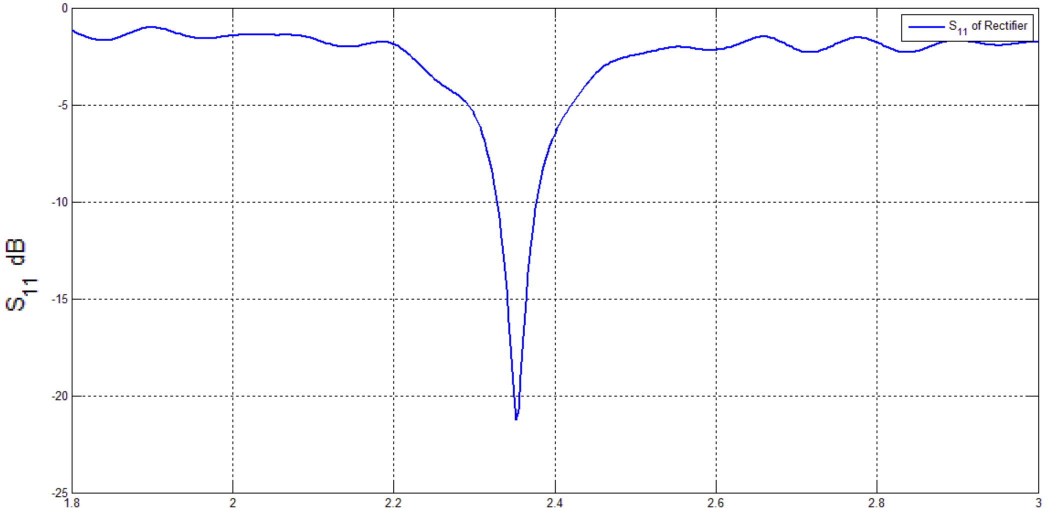

Sanjeev K. 2014-15

Comparison of Efficiencies

Schottky diode

	900MHz	2400MHz
-15dbm	Efficiency=47% Output voltage=0.29V Load resistor=6K	Efficiency =17% Output voltage=0.2V Load resistor=6K
-20dbm	Efficiency=31% Output voltage=0.15V Load resistor=6K	Efficiency=5% Output voltage=0.1V (0.153V unloaded) Load resistor=6K
	Used HSMS 2852	Used HSMS 2862

Diode connected MOS with high Q matching

	900MHz	2400MHz
-14dbm	Efficiency=6.2% Output voltage=1.1V Load resistor=500K	Efficiency=2% Output voltage=0.632V Load resistor=500K
-20dbm	Efficiency=1.8% Output voltage=0.3V Load resistor=500K	Efficiency=0.2% Output voltage=0.118V Load resistor=500K

Zero V_{TH} CMOS

	900MHz	2400MHz
-15dbm	Efficiency=4.6% Output voltage=0.86V Load resistor=500K	Efficiency=3.79% Output voltage=0.774V Load resistor=500K
-25dbm	Efficiency=2.4% Output voltage=0.198V Load resistor=500K	Efficiency=1.7% Output voltage=0.165V Load resistor=500K

← Power

← Technologies

Frequencies →

Cross-coupled Rectifiers for Low Power

- CMoS integration requires diodes using MoSFETs.
- Simple diode connected configurations are not effective at low power/voltage levels
- In Cross Coupled Rectifiers
  - Biasing of MOSFETs by charge stored in capacitors. This is a way of threshold compensation.
  - Low ON resistance due to high overdrive voltage.
  - In both cycles of input, output capacitor is charged. Although DCP uses both cycles, only alternate cycles charges the output capacitor and the other cycle charges the input capacitor.

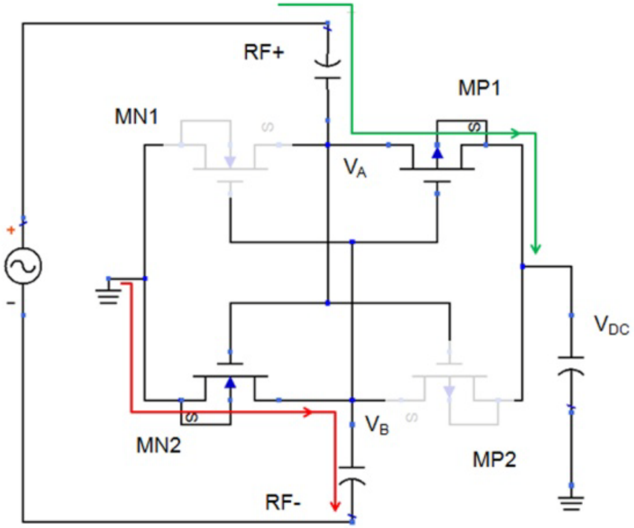

DC-DC converter

- Low loss switched capacitor DC-DC converter:

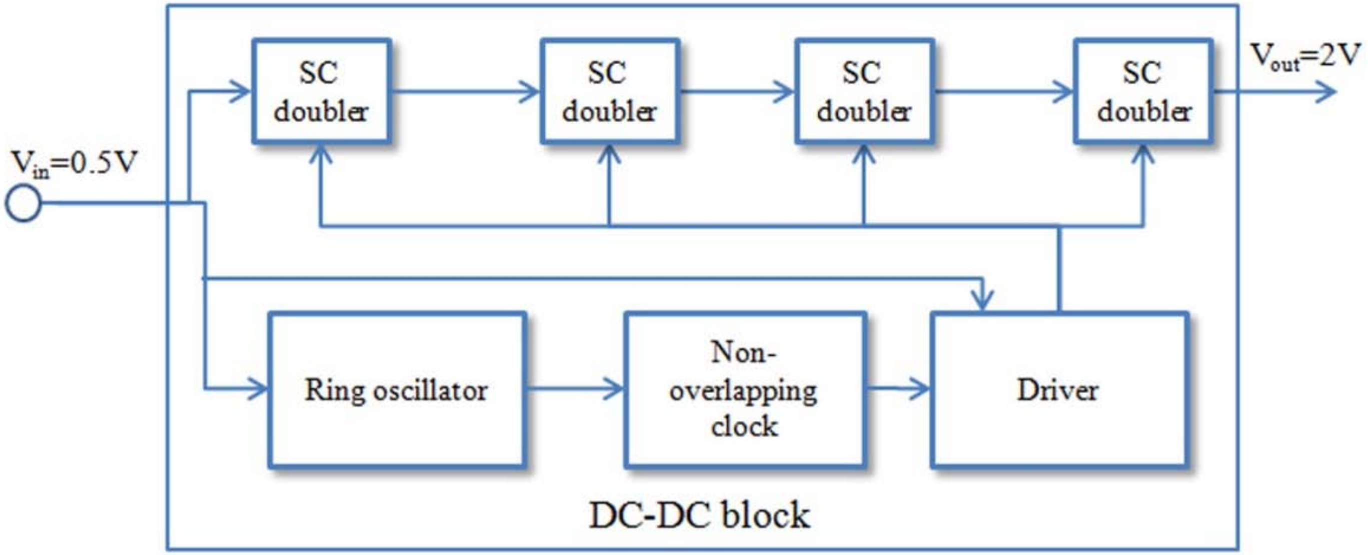
SC doubler
SC doubler
SC doubler
SC doubler
V_{out}=2V
V_{in}=0.5V
Ring oscillator
Non-overlapping clock
Driver
DC-DC block

Full system block diagram

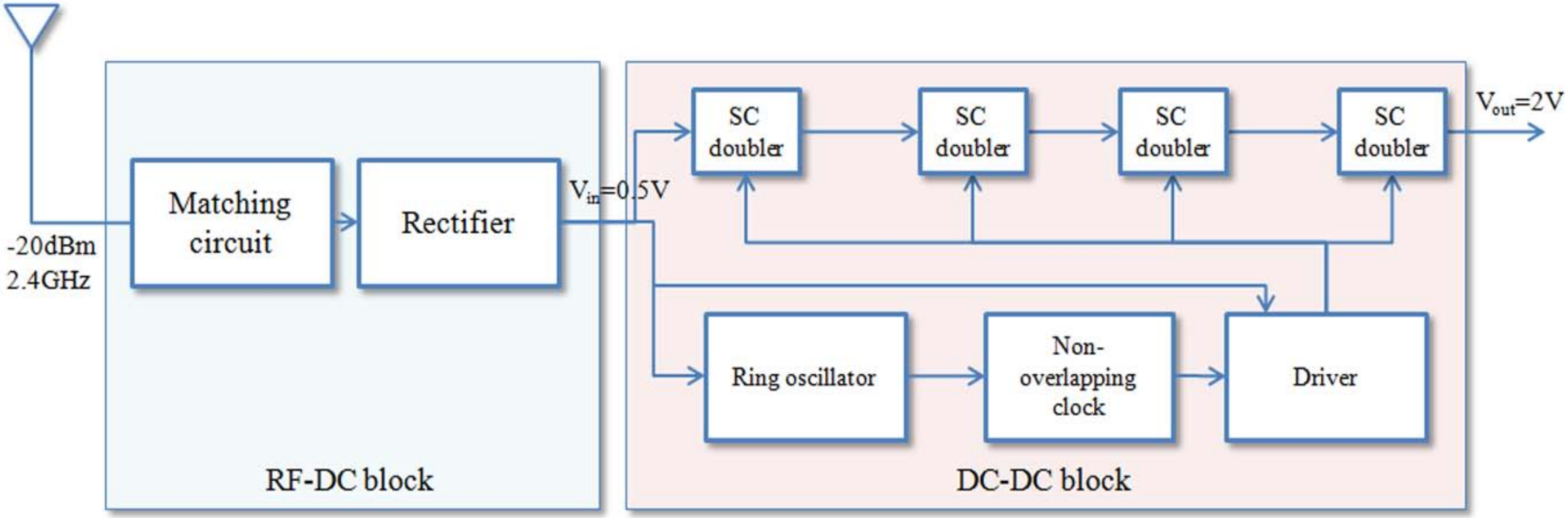
-20dBm
2.4GHz
Matching circuit
Rectifier
V_{in}=0.5V
SC doubler
SC doubler
SC doubler
SC doubler
V_{out}=2V
Ring oscillator
Non-overlapping clock
Driver
RF-DC block
DC-DC block

1. Output capacitor of RF-DC supplies DC-DC
2. Enable generator logic constructed using back to back inverters
3. 5 MOSFETs added to limit supply voltage to ring oscillator

RICOH
imagine. change.

Sanjeev K, Manjunath M, TV Prabhakar, KJ Vinoy, Some Practical Considerations of RF to DC Converter using Low Vth CMOS Rectifier, Applied Computational Electromagnetic Conference, Guwahati, Dec 28-21, 2015

Full system simulations

- Output capacitor of RF-DC loaded heavily when clocks transition, so ripples exist in RF-DC output.
- Below 0.5V at clock transitions, above 0.5V between clock transitions
- Time step is 8ps for RF-DC simulation and DC-DC has to run for hundreds of µs or few ms, so simulations times are large.

Output voltage=2V across a load of 1.6MΩ, which gives

$$Efficiency = \frac{2^2/1.6\times10^6}{10^{-5}} = 25\%$$

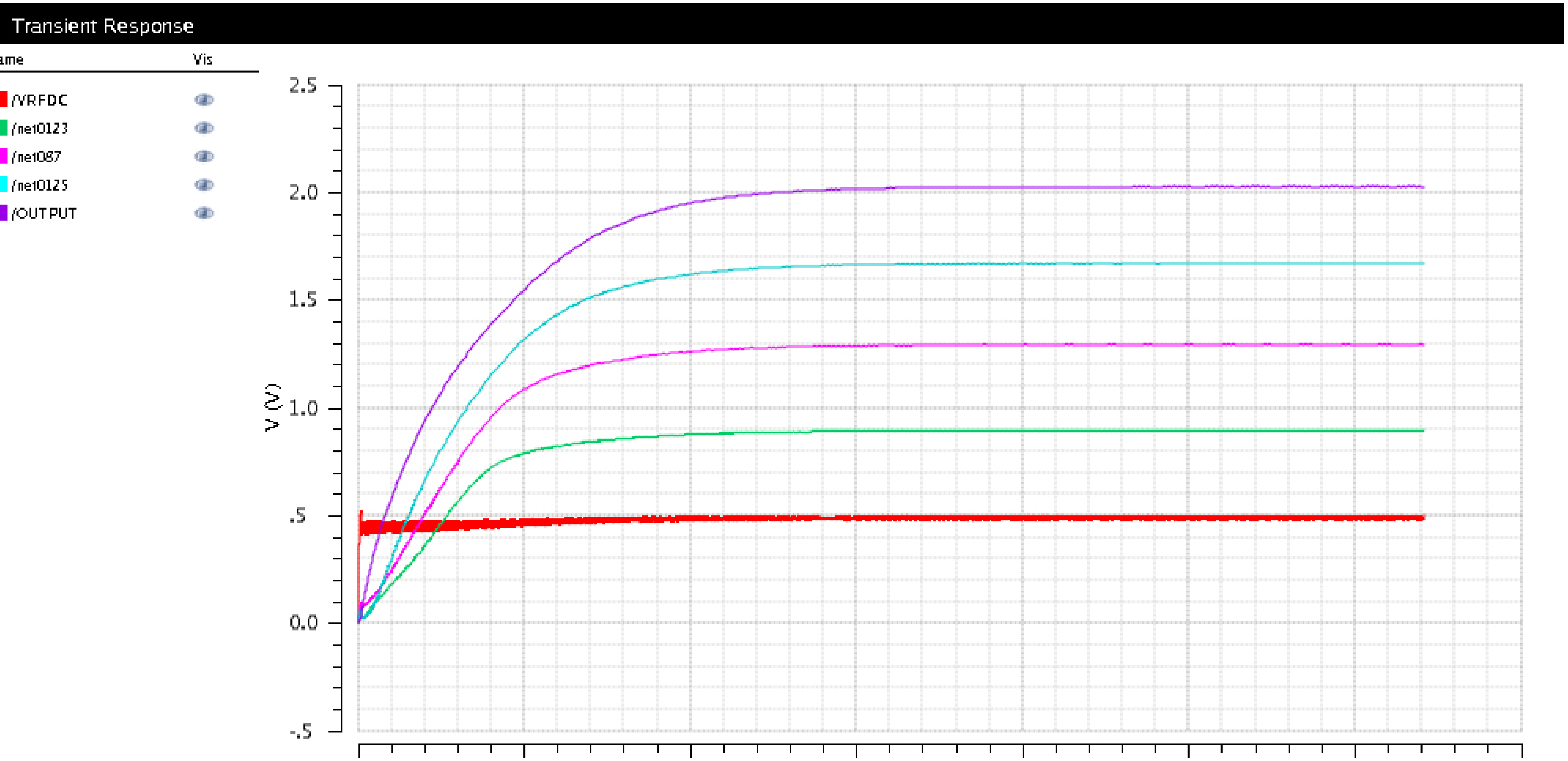
Transient Response
ame
Vis
🟥 /VRFDC
🟩 /net0123
🟪 /net087
🟦 /net0125
🟪 /OUTPUT
V (V)
2.5
2.0
1.5
1.0
.5
0.0
-.5

Summary with IC Design

- Low Vth NMOS based DCP gives 42.3% efficiency.
- FGCCR gives 55% efficiency.
- The overall system efficiency is 25%.
- Higher than efficiency reported in literature for RF-DC converter operating at -20dBm, 2.4GHz in 130nm technology.

Reference	This work	[5]	[7]	[6]
Power level	-20dBm	-25.7dBm	-22.6dB	-20dBm
Frequency	2.4GHz	2.45GHz	906MHz	2.4GHz
Efficiency	55%(simulated)	37%(measured)	10%(measured)	36%(simulated)
Rectifier	FGCCR in UMC 130nm CMOS	DCP in 0.5µm Silicon on Sapphire	DCP with floating gate transistors in 0.25µm CMOS	FGCCR in 130nm CMOS

6. ASIC Design for IOT

- Working on a 3-chip architecture
  - Our chip to enable sensing, and control functions
  - Communication using an external Monza chip
- Fabricated chip using commercial services!!

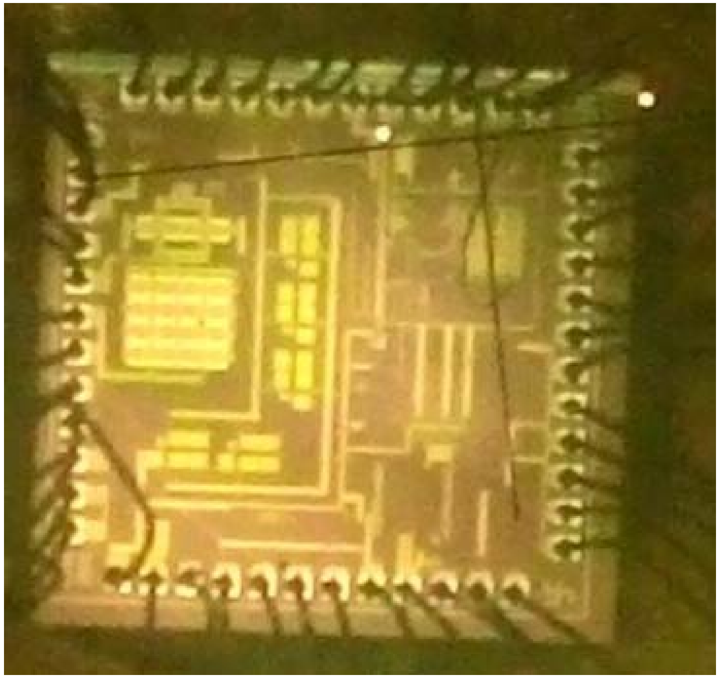

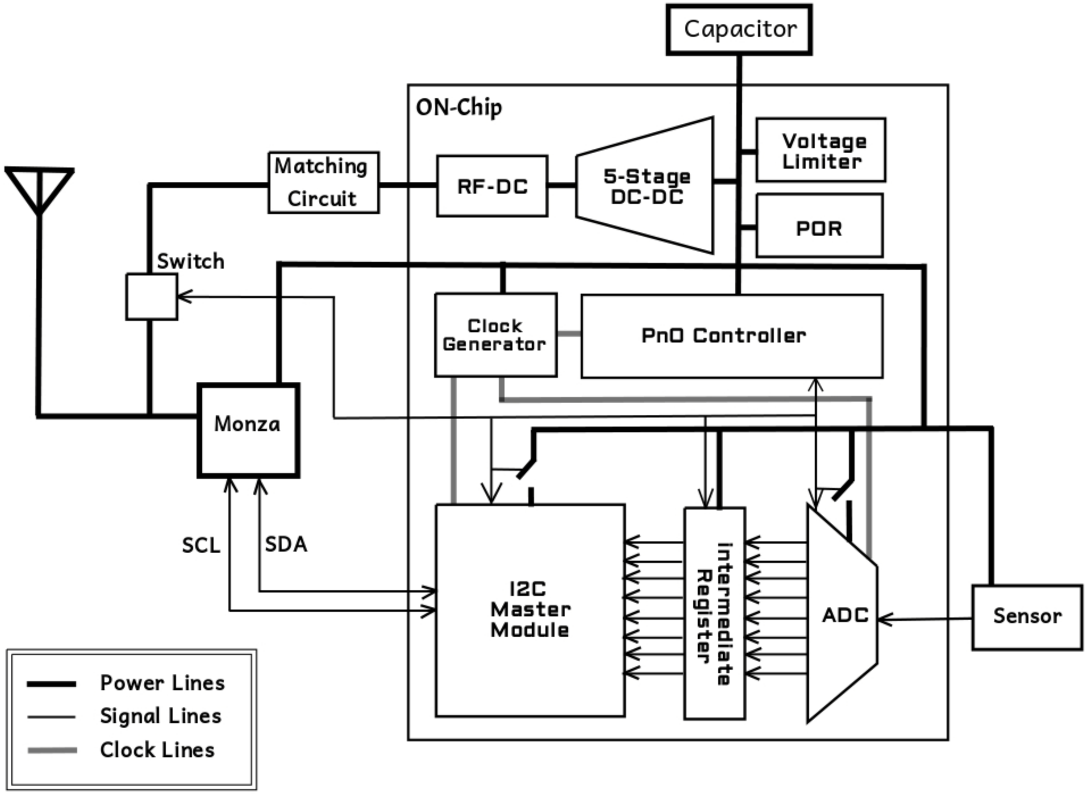
Capacitor
ON-Chip
Matching Circuit
RF-DC
5-Stage DC-DC
Voltage Limiter
POR
Switch
Clock Generator
PnO Controller
Monza
SCL
SDA
I2C Master Module
Intermediate Register
ADC
Sensor
Power Lines
Signal Lines
Clock Lines

Battery-less Sensor node for BLE

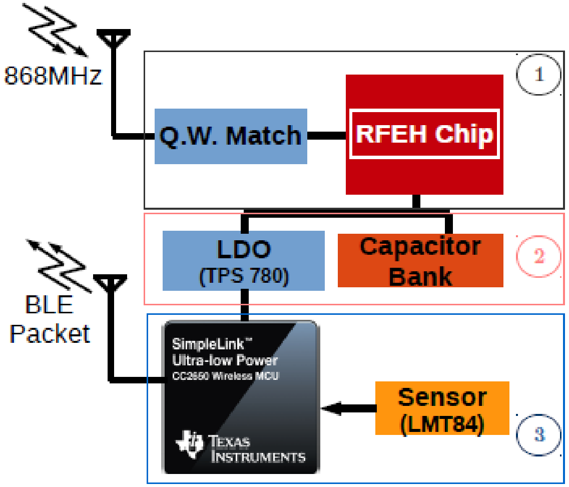
868MHz
Q.W. Match
RFEH Chip
1
LDO
(TPS 780)
Capacitor Bank
2
BLE Packet
SimpleLink™ Ultra-low Power CC2650 Wireless MCU
Texas Instruments
Sensor (LMT84)
3

① Energy Harvesting Block ② Energy Storing Block ③ Sensing & Communication Block

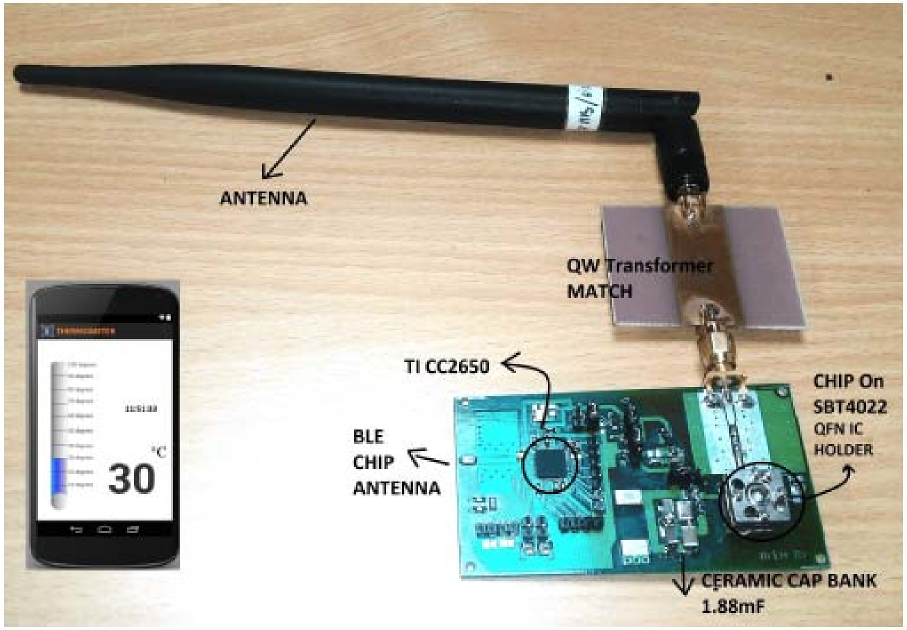
ANTENNA
QW Transformer MATCH
TI CC2650
BLE CHIP ANTENNA
CHIP On SBT4022 QFN IC HOLDER
CERAMIC CAP BANK 1.88mF
30 °C

Distance from Reader (m)	Input Power (dBm)	Cold Start Time (min)	Packet Time (min)	Efficiency (%)
3	-2	15	1.75	2.2
4	-5	35	7	1.7
5	-8	120	32	1

Eficiency(%) vs Load (5stage DC-DC)

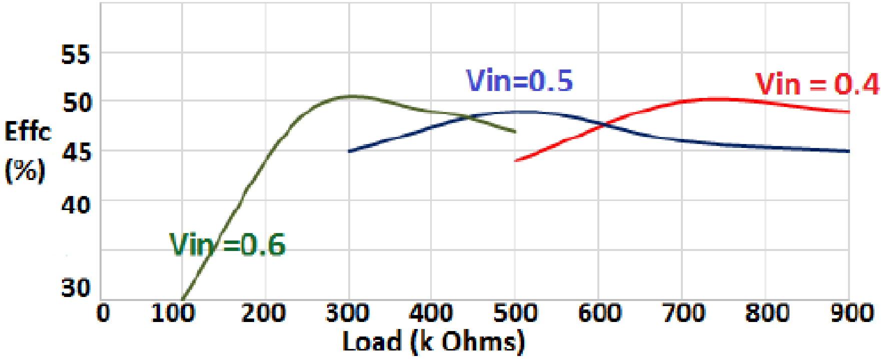

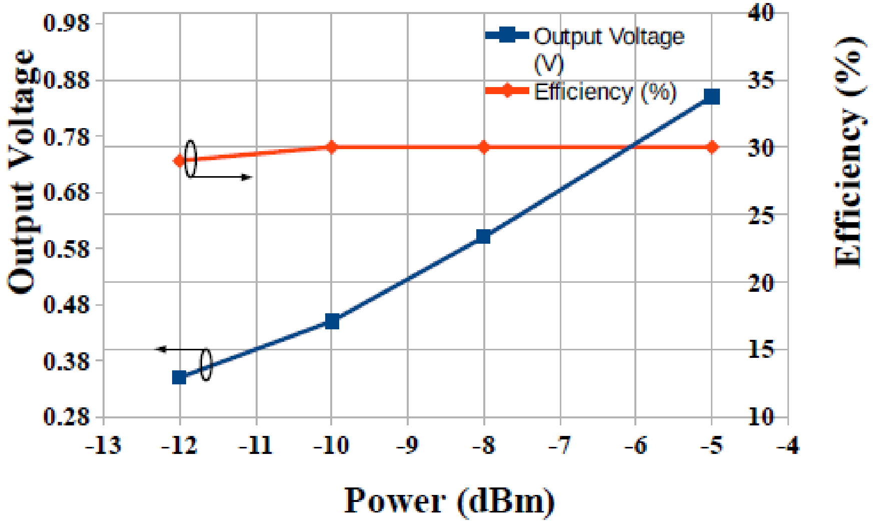

Bharat Hegde & Syed Younis 2015-16

Summary

- Most low power wireless terminals operate intermittently
- These require anywhere 50uW to about 10mW for their operation.
  - Batteries limited: cost, size, stored energy
  - Solar: not dependable through
- WPT and RF EH can enable wide use of IoT
  - Main challenges in the design is the low incident energy/power/voltage
  - High Quality factor components may help
- Several fabricated examples discussed here: All can transmit data to an aggregator wirelessly
  - Different standards implemented.

Acknowledgements

- Dr. TV Prabhakar, DESE, IISc
  - Gaurav Singh
  - Rahul P
  - Chaithanya C
  - Aditya Mitra
  - Prashuk Jain
  - Prashanth Raja
  - Nirmal John
  - Nithin Jose
  - Syed Younis
  - Bharat Hegde
  - Niharika Thakuria
- Partial Funding From
  - ANRC (Boeing, Wipro, HCL)
  - Ricoh Research, India

- Prof Bharadwaj Amrutur
  - Uday S
  - NS Sreeram
- Others
  - Vivekanand M
  - Harikiran M
  - Manjunath M
  - Sanjeev K
  - Sandeep Rana

ANRC

RICOH
imagine. change.

Thank YOU

Take the world
on line
Take control
of the world
Let the things
talk to each
other
Let thing
become
intelligen

Free Energy From Air Circuit

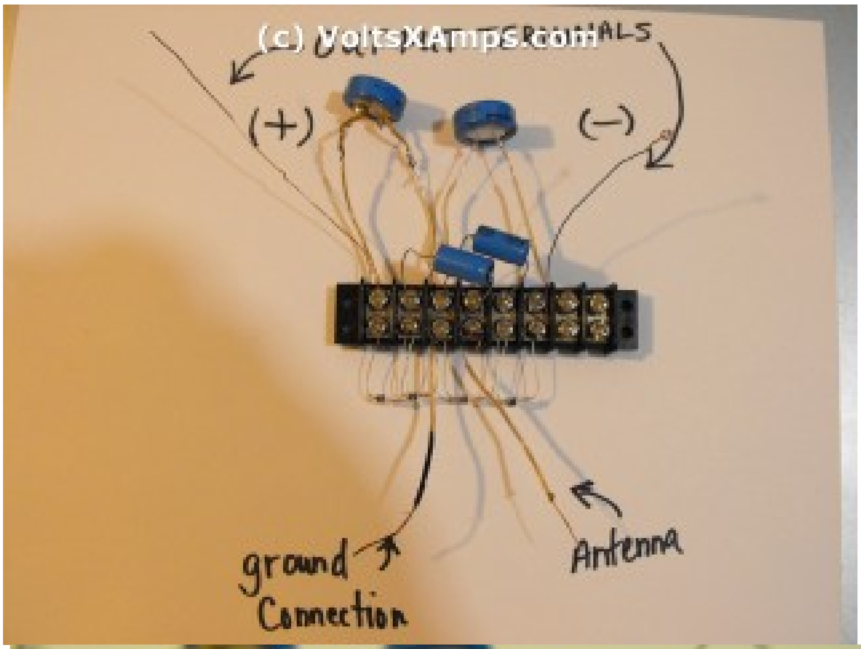
(c) VoltsXAmps.com
(+)
(−)
ground Connection
Antenna

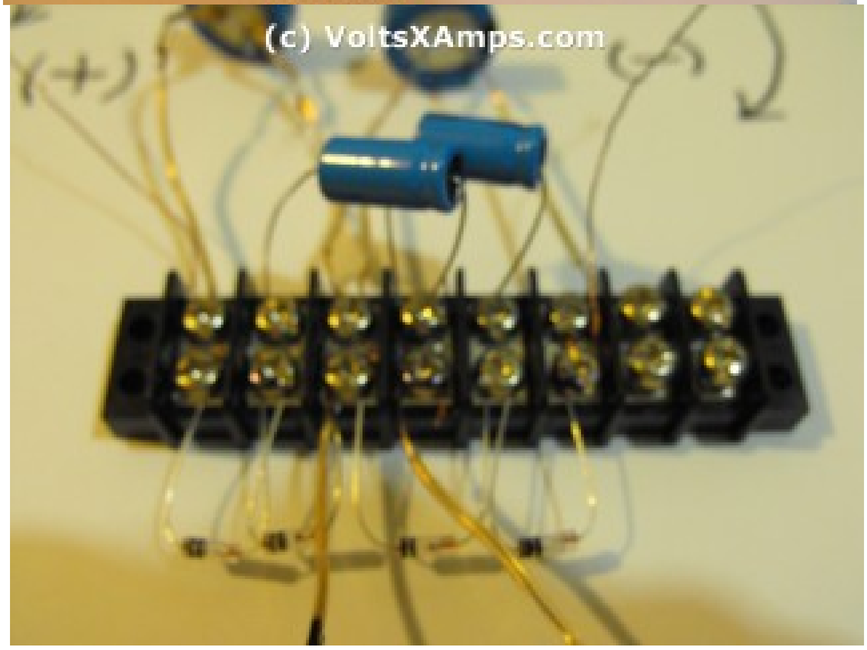
(c) VoltsXAmps.com
(+)
(−)

Last year I found a United States Patent that showed how to collect ambient energy right from the air. I finally decided to build this curcuit just to see what it could do.

[protected]

The above images are from the circuit I built. Most of the text and info comes direct from the US4628299filed by Joseph Tate. US4628299

The Amazing Ambient Power Module

Parts List for the APM-2

Four 1N34 germanium diodes (Radio shack #276-1123) ~ Figure 1, X1, X2, X3, & X4
Two 0.2 mfd 50 V ceramic capacitors ~ Figure 1, C1 & C2
Two 100 mfd 50V electrolytic capacitors (Radio Shack #272-1016) ~ Figure 1, C3 & C4
Copper wire for antenna & ground connections

Introduction

The Ambient Power Module (APM) is a simple electronic circuit which, when connected to antenna and earth ground, will deliver low voltage up to several milliwatts. The amount of voltage and power will be determined by local radio noise levels and antenna dimensions

Generally a long wire antenna about 100' long and elevated in a horizontal position about 30' above ground works best. A longer antenna may be required in some locations. Any type copper wire, insulated or not, may be used for the antenna. More details about the antenna and ground will be discussed further on.

The actual circuit consists of two oppositely polarized voltage doublers (Figure 1). The DC output of each doubler is connected in series with the other to maximize voltage without using transformers. Single voltage doublers were often found in older TV sets for converting 120 VAC to 240 VDC. In the TV circuit the operating frequency is 60 Hz.

FIG. 1

APM-2 SCHEMATIC DIAGRAM

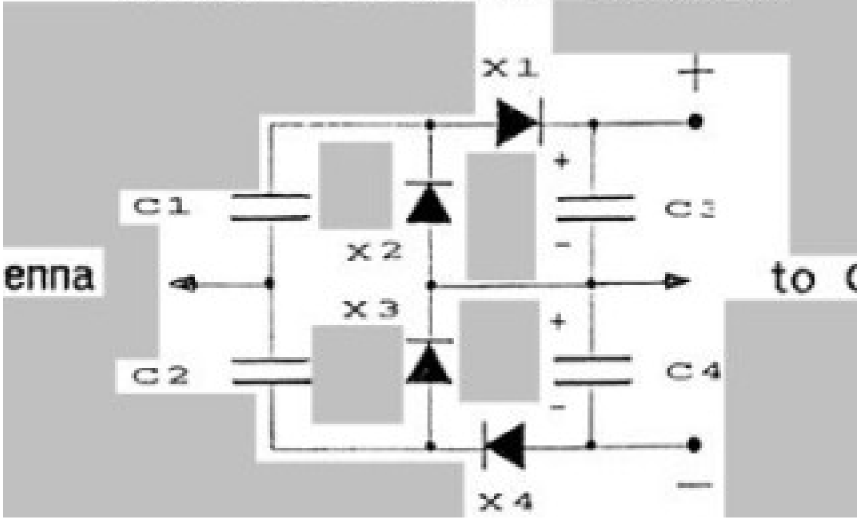

The APM operates at radio frequencies, receiving most of its power from below 1 MHz. The basic circuit may be combined with a variety of voltage regulation schemes, some of which are shown in Figure 2. Using the APM-2 to charge small NiCad batteries provides effective voltage regulation as well as convenient electrical storage. This is accomplished by connecting the APM-2 as shown in Figure 2B.

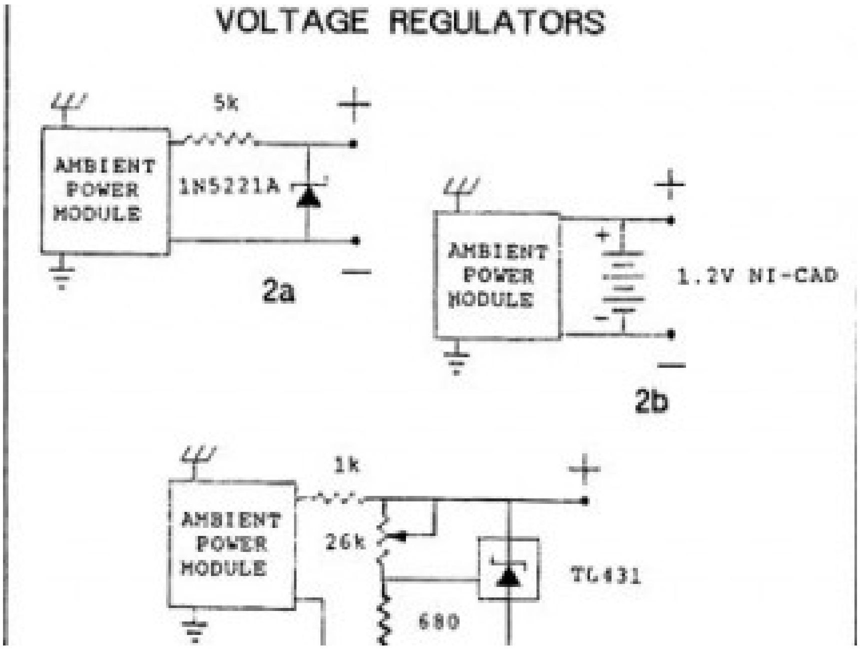

Charging lead acid batteries is not practical because their internal leakage is too high for the APM to keep up with. Similarly, this system will not provide enough power for incandescent lights except in areas of very high radio noise.

It can be used to power small electronic devices with CMOS circuitry, like clocks and calculators. Smoke alarms and low voltage LEDs also can be powered by the APM.

Figure 3 is a characteristic APM power curve measured using various loads from 0-19 kOhm. This unit was operating from a 100' horizontal wire about 25' high in Sausalito CA. As can be seen from the plot, power drops rapidly as the load resistance decrease from 2 kOhm. This means that low voltage, high impedance devices, like digital clocks, calculators and smoke alarms are the most likely applications for this power source. Some applications are shown in Figures 4 through 7.

FIG. 3 AMP Power Envelope

Volts vs. Resistance

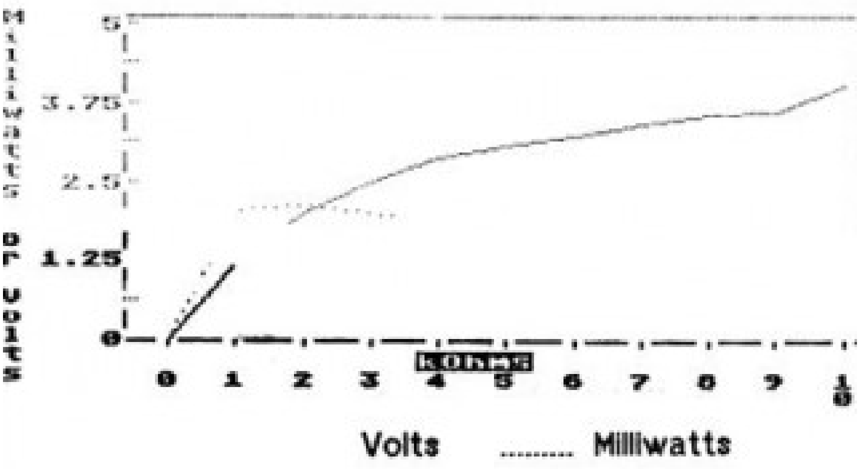

Figure 4 ~ A digital clock is shown powered by the APM-2. The 1.5 volt clock draws 28 microamps. Its position on the power envelope curve would be off the scale to the right and almost on the bottom line, dissipating only 42 microwatts.

FIG. 4 Clock powered by APM

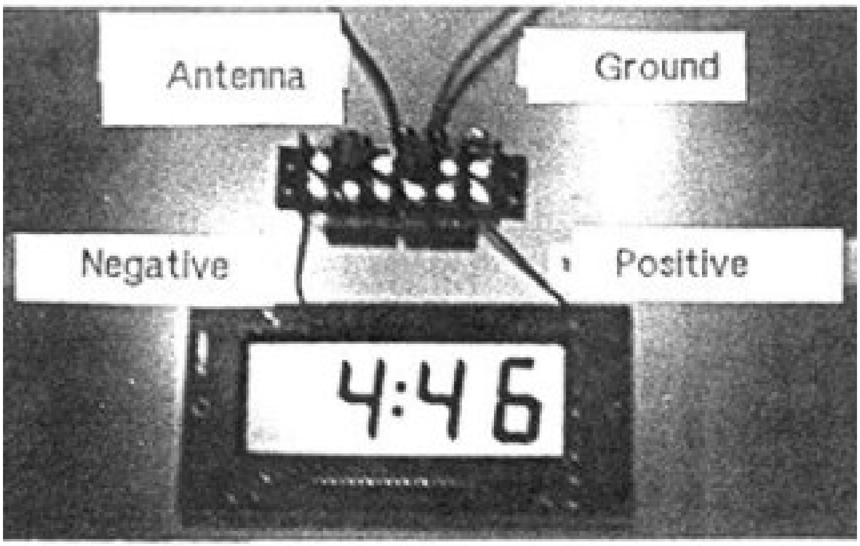

FIG. 5 APM charging a battery

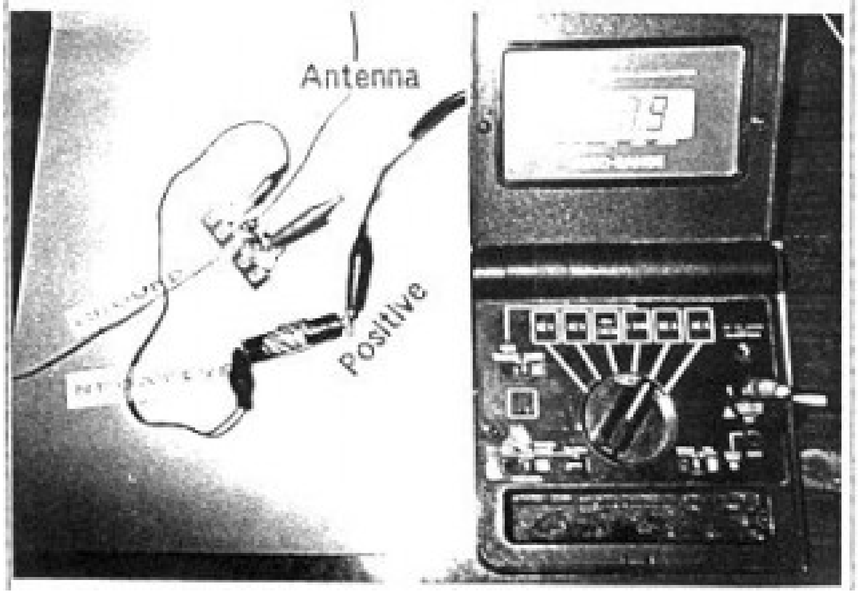

Figure 6 shows a clock which has the APM-2 built into it so it is only necessary to connect the antenna and ground wires directly to the clock. The antenna for this clock, which is a low frequency marine type, is shown in Figure 7.These antenna are expensive, not generally available, and usually don't work any better than the long wire mentioned above. But it may be necessary to use them in urban areas where space is limited and radio noise is high.

FIG. 6

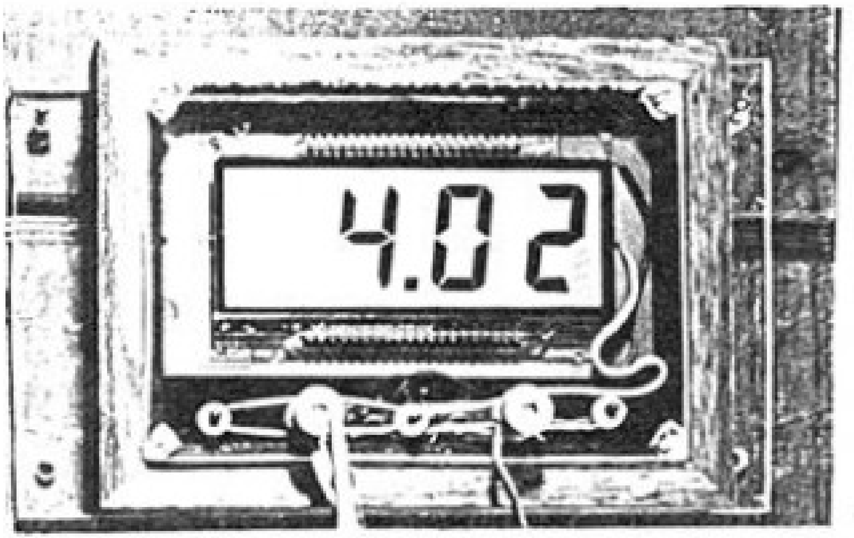

Building the Module

The builder has a choice of wiring techniques which may be used to construct the module. It may be hand wired onto a terminal strip, laid out on a bread board, experiment board, or printed circuit. Figure 8 shows some of the different ways of constructing the APM-2.

Figure 8A is constructed on a screw strip terminal; Figure 8B is constructed on a perforated breadboard; Figure 8C is built on a standard experiment board; Figures 8D, 8E, and 8F are all printed circuits; Figure 8F is made up on a solder strip terminal.

FIG. 8 a, b, c, d, e, f & g

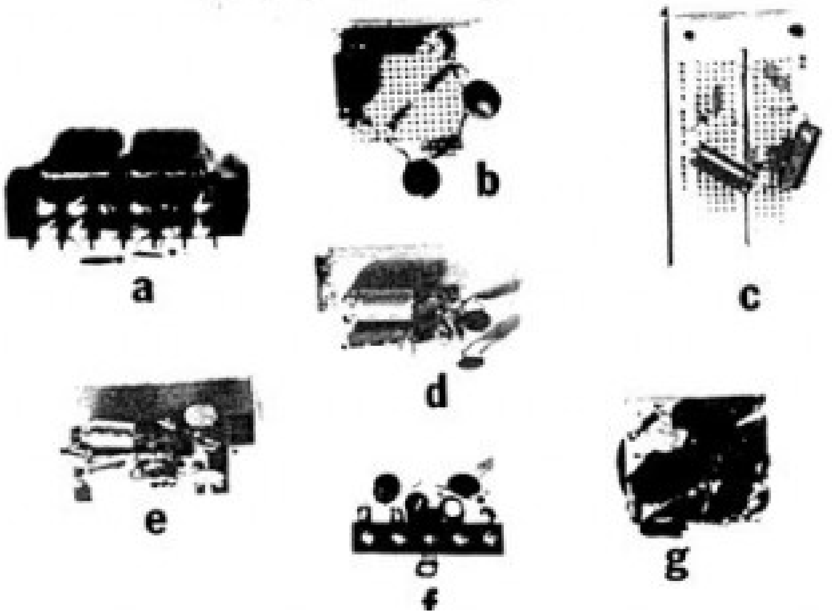

If you wish to make only one or two units, hand wiring will be most practical, either on a terminal strip or breadboard. Assembly on the terminal strip (Figure 8A) can be done easily and without soldering. It is important to get the polarity correct on the electrolytic capacitor. The arrow printed on the side of the capacitor points to negative.

Figure 9 is a closer view of the terminal strip with an illustration of the components and how they are connected.

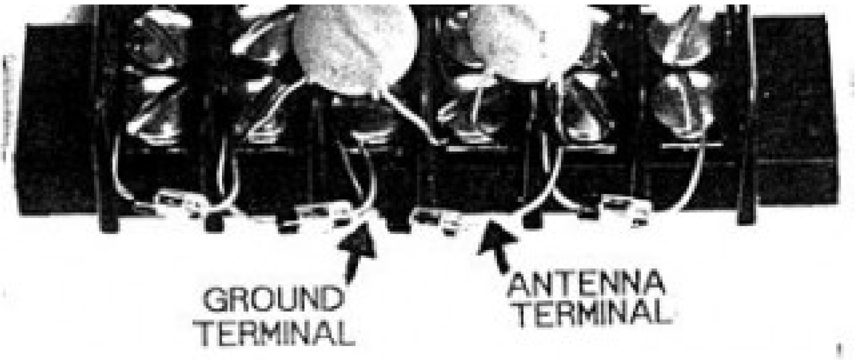

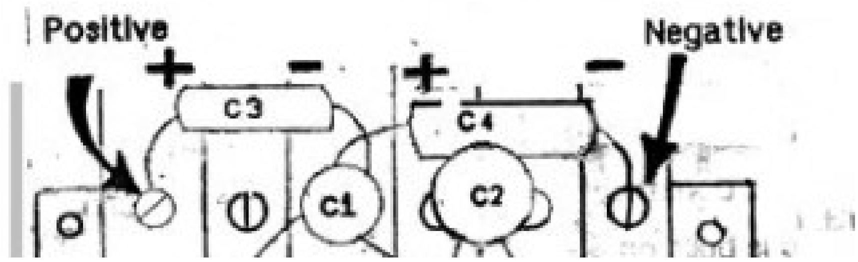

The breadboard unit is shown in Figure 10 with all components on one side and all connections on the other. All you need is a 2" x 2" piece of perforated breadboard (Radio Shack #276-1395) and the components on the parts list. Push component wires through the holes and twist them together on the other side. Just follow the pattern in the photo, making sure to observe the correct polarity on the electrolytic capacitors and the diodes. The ceramic capacitors may be inserted in either direction.

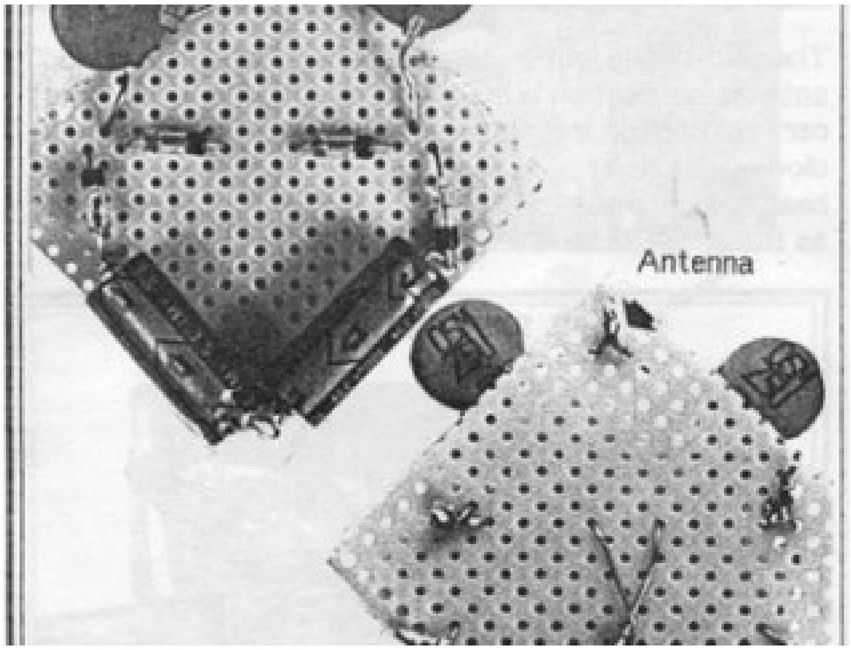

The experiment board unit is assembled by simply pushing the component leads into the board as shown in Figure 11. This unit is powering a small red LED indicated by the arrow.

FIG. 11 EXPERIMENT BOARD

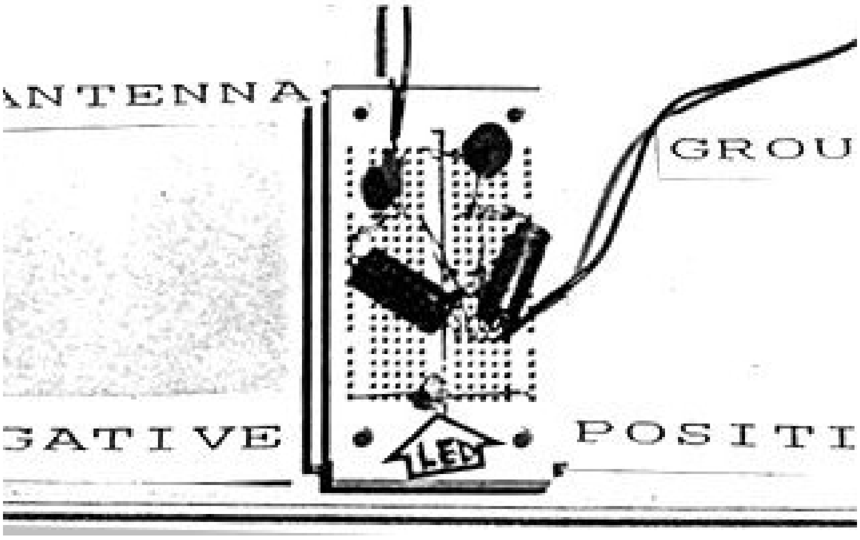

The solder strip unit is made up on a five terminal strip. The antenna connection is made to the twisted ends of the ceramic capacitors. When soldering the leads of the 1N34 diodes, care must be taken to avoid overheating. Clip a heat sink onto the lead between the diode and the terminal as shown in Figure 12.

FIG. 12 SOLDER STRIP UNIT

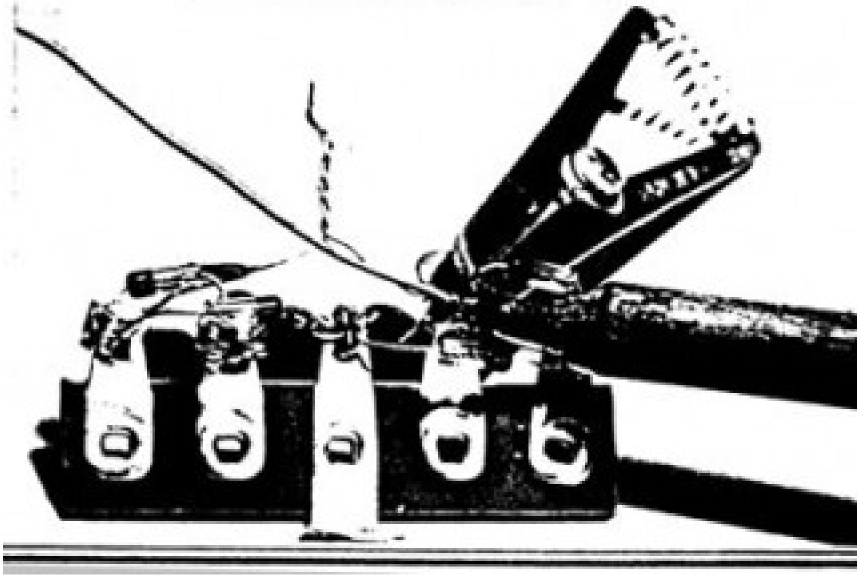

FIG. 13

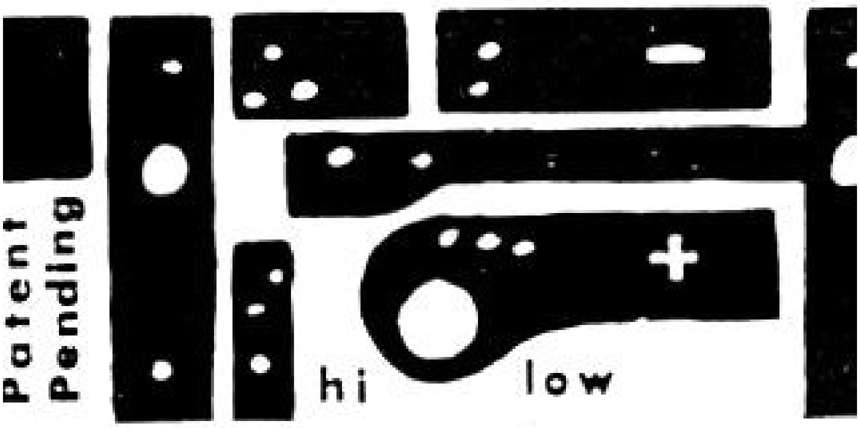
Patent Pending
−
+
hi
low

Figure 14 shows the front and back view of the completed printed circuit.

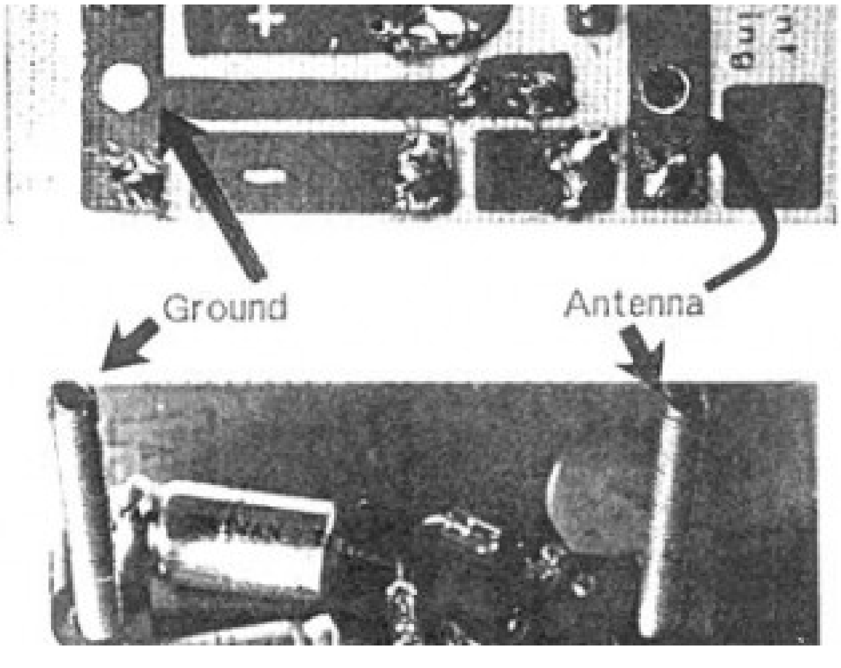
+
−
Ground
Antenna

A small switch may be installed on the board to activate the zener regulator (Figure 15). This board was designed for use in clocks.

FIG. 15

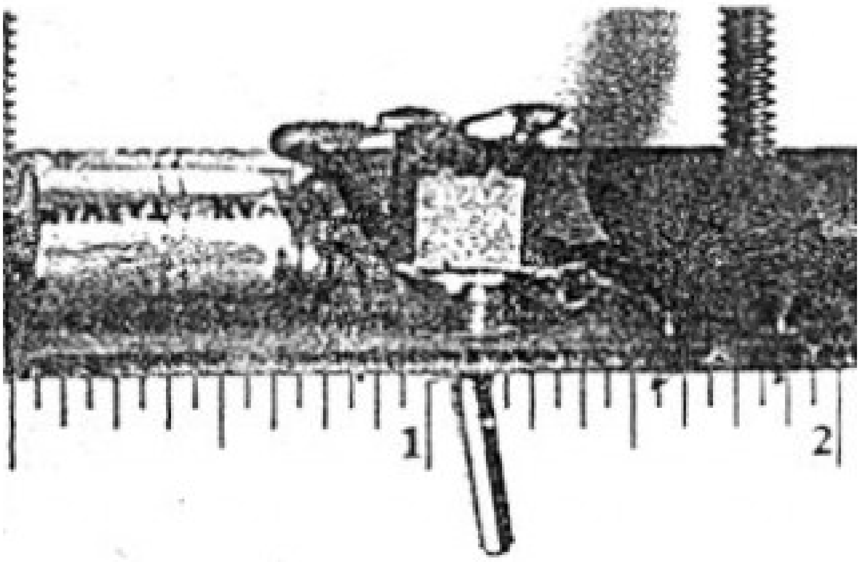

Antenna Requirements

The antenna needs to be of sufficient size to supply the APM with enough RF current to cause conduction in the germanium diodes and charge the ground coupling capacitors. It has been found that a long horizontal wire works best. It will work better when raised higher. Usually 20-30 feet is required. Lower elevations will work, but a longer wire may be necessary.

In most location, possible supporting structures already exist. The wire may be stretched between the top of a building and some nearby tree or telephone pole. If live wires are present on the building or pole, care should be taken to keep your antenna and body well clear of these hazards.

To mount the wire, standard commercial insulators may be sued as well as homemade devices. Plastic pipe makes an excellent antenna insulator. Synthetic rope also works very well, and has the advantage of being secured simply by tying a knot. It is convenient to mount a pulley at some elevated point so the antenna wire may be pulled up to it using the rope which doubles as an insulator (Figure 16).

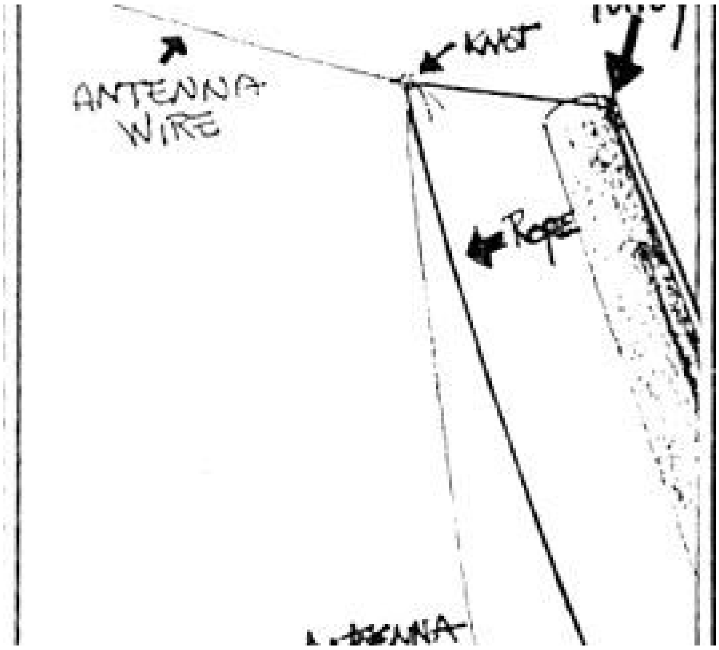

Figure 17 is an illustration of a horizontal wire antenna using a building and tree for supports.

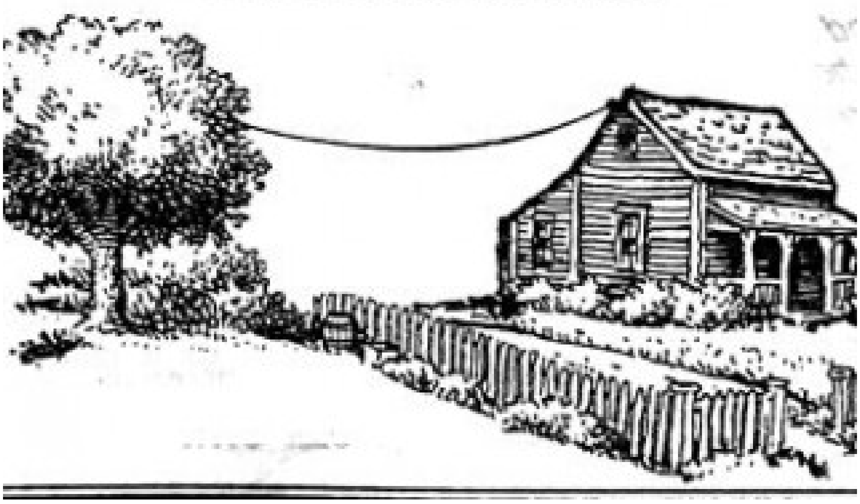

FIG. 17

Horizontal Wire Installation

Grounding

Usually a good ground can be established by connecting a wire to the water or gas pipes of a building. Solder or screw the wire to the APM-2 ground terminal. In buildings with plastic pipes or joints, some other hookup must be used. A metal rod or pipe may be driven into the ground in a shady location where the earth usually is damper. Special copper coated steel rods are made for grounds which have the advantage of good bonding to copper wire. A ground of this type usually is found within the electrical system of most buildings.

Conduit is a convenient ground provided that the conduit is properly grounded. This may be checked with an ohmmeter by testing continuity between the conduit and system ground (ground rod). Just as with the antenna, keep the ground wire away form the hot wires. The APM's ground wire may pass through conduit with other wires but should only be installed by qualified personnel.

Grounding in extremely dry ground can be enhanced by burying some salts around the rod. The slats will increase the conductivity of the ground and also help retain water. More information on this subject may be found in an antenna handbook.

Good luck getting your Ambient Power Module working. It is our hope that experimenters will find new applications and improve the power capabilities of the APM.

CopyRight Joseph Tate

Inventors: Tate, Joseph B. (Sausalito, CA)
Brown, David E. (Mill Valley, CA)

Application Number: 06/695632

Publication Date: 12/09/1986

Filing Date: 01/28/1985 NOTE: The images maybe hard to see as they were originally scanned and uploaded in black and white. If it helps you to replicate this device you may want to check the images of the one I built at the very top (in color) Here are some more photos of my completed unit for your review.

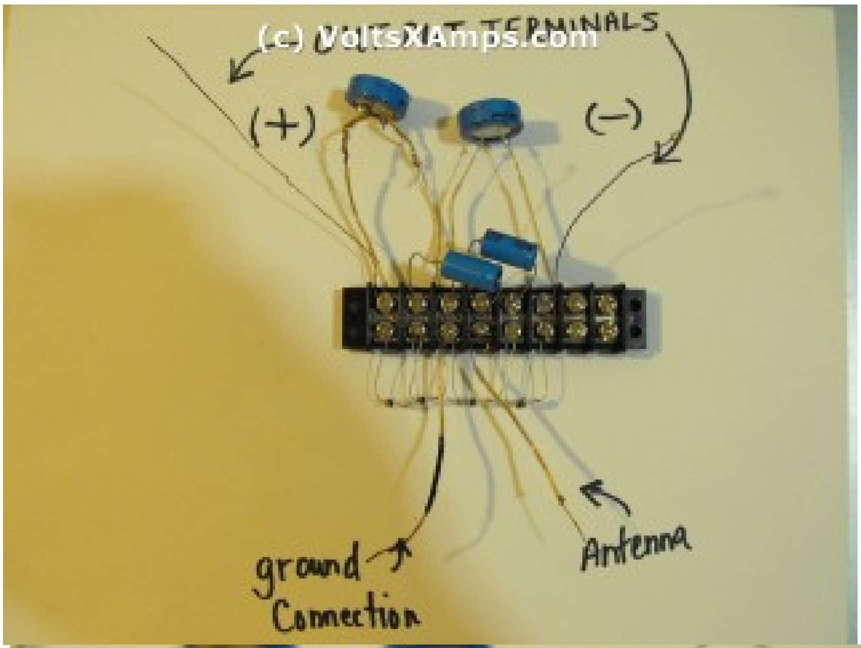
(c) VoltsXAmps.com
TERMINALS
(+)
(–)
ground Connection
Antenna

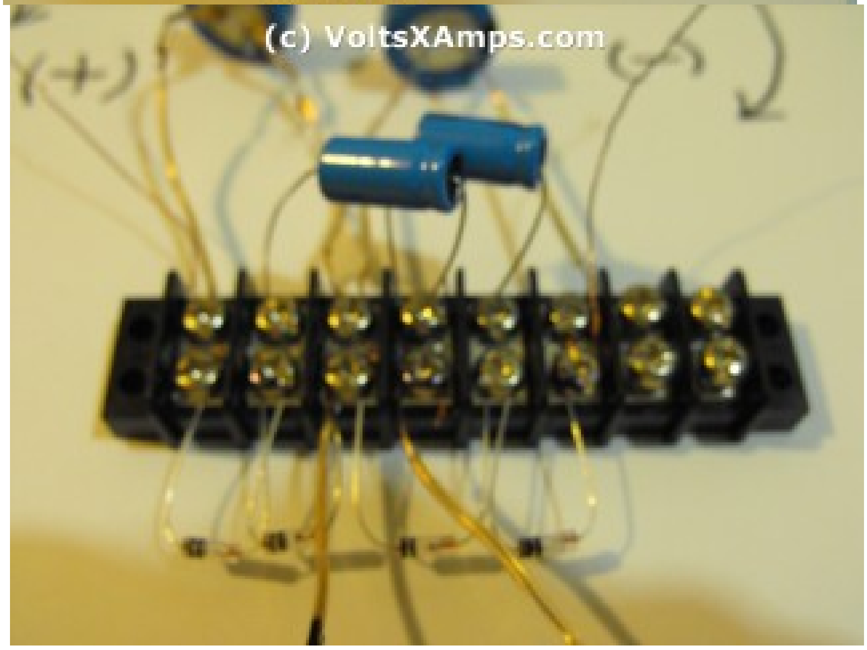
(c) VoltsXAmps.com
(+)
(–)

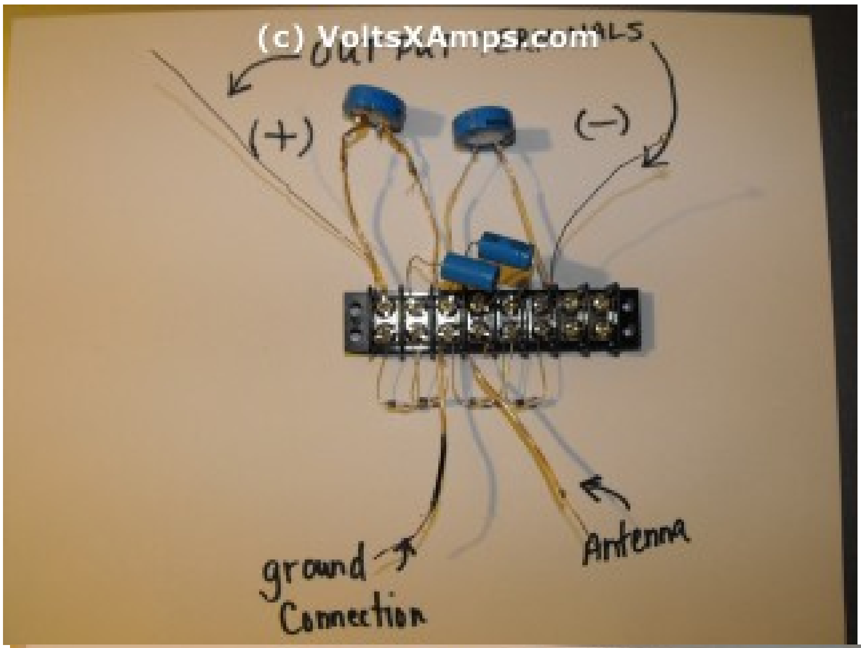
(c) VoltsXAmps.com
(+)
(–)
ground Connection
Antenna

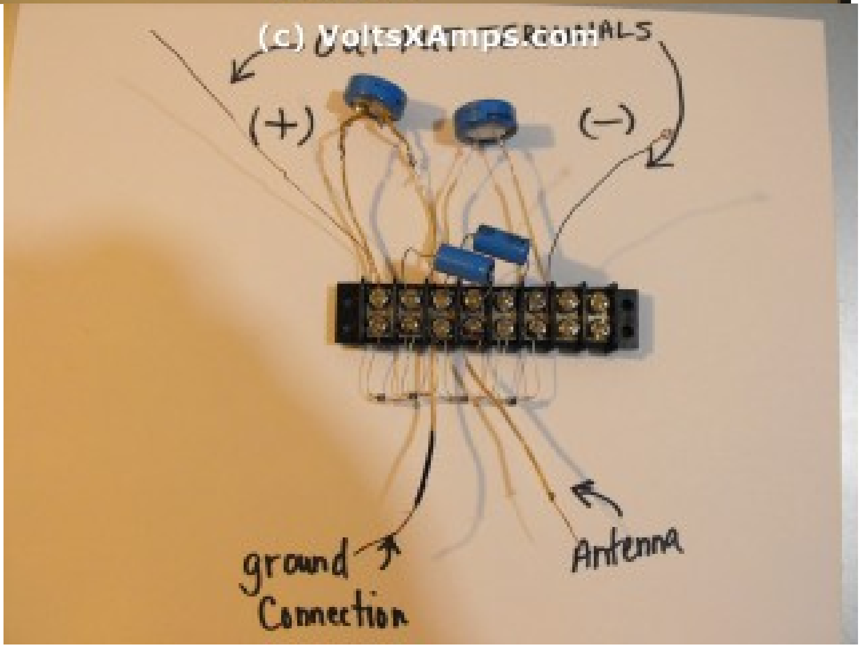
(c) VoltsXAmps.com
(+)
(–)
ground Connection
Antenna

[/protected]

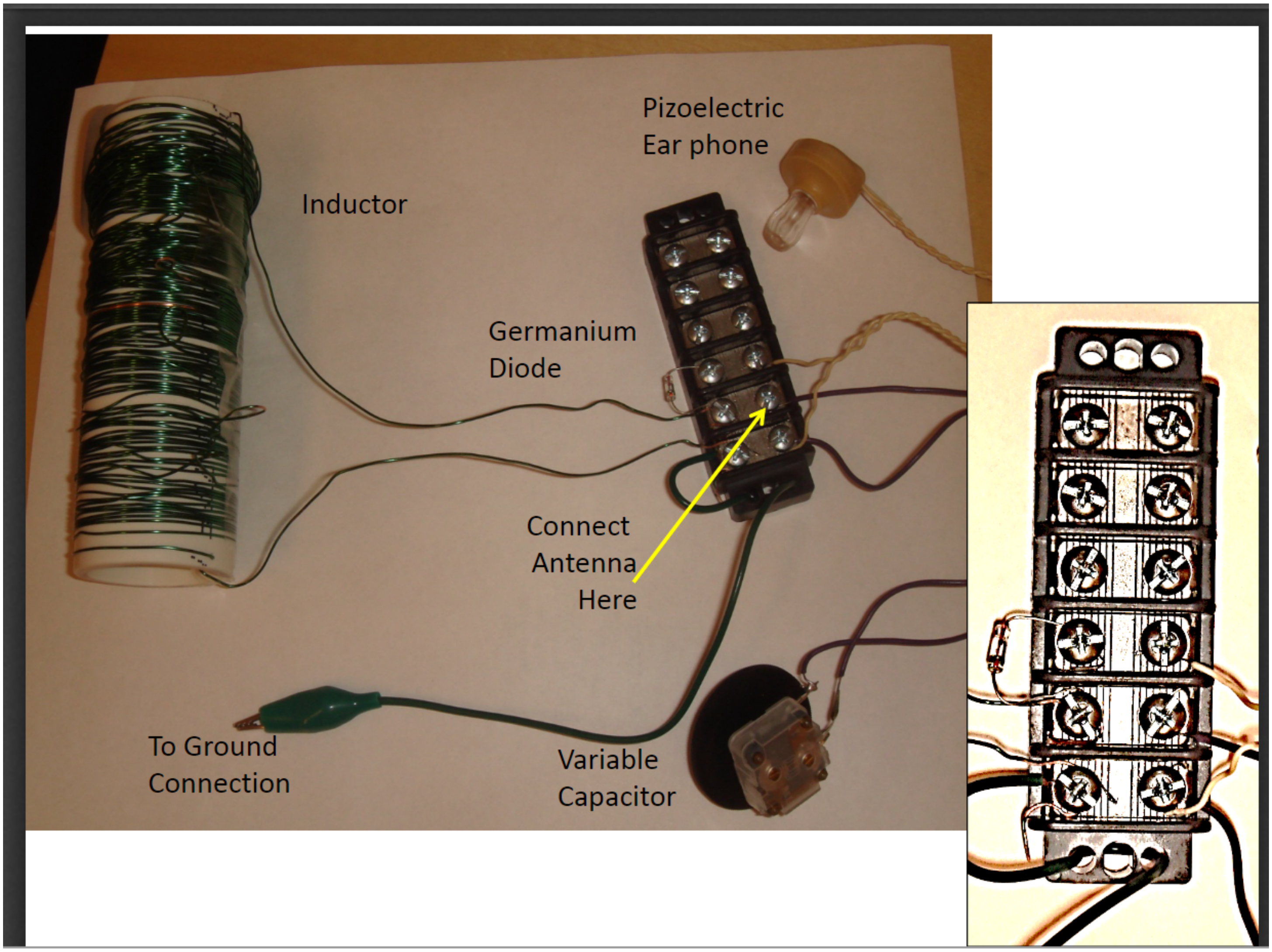
Pizoelectric
Ear phone
Inductor
Germanium
Diode
Connect
Antenna
Here
To Ground
Connection
Variable
Capacitor

FM Crystal Radios?

copyright © 2006 by Larry J Solomon

I have heard, even from a physicist, that it is impossible to build FM crystal radios. On the other hand some experimenters claim that they have built them. This argument intrigued me to try and build an FM crystal radio, which I have done successfully. To my surprise, the result is an astounding performer, pulling in four local stations in Tucson. When connected as a receiver to a good sound system the sound fidelity is as good or better than more expensive AM radios. In fact, it sounds "high-fidelity".

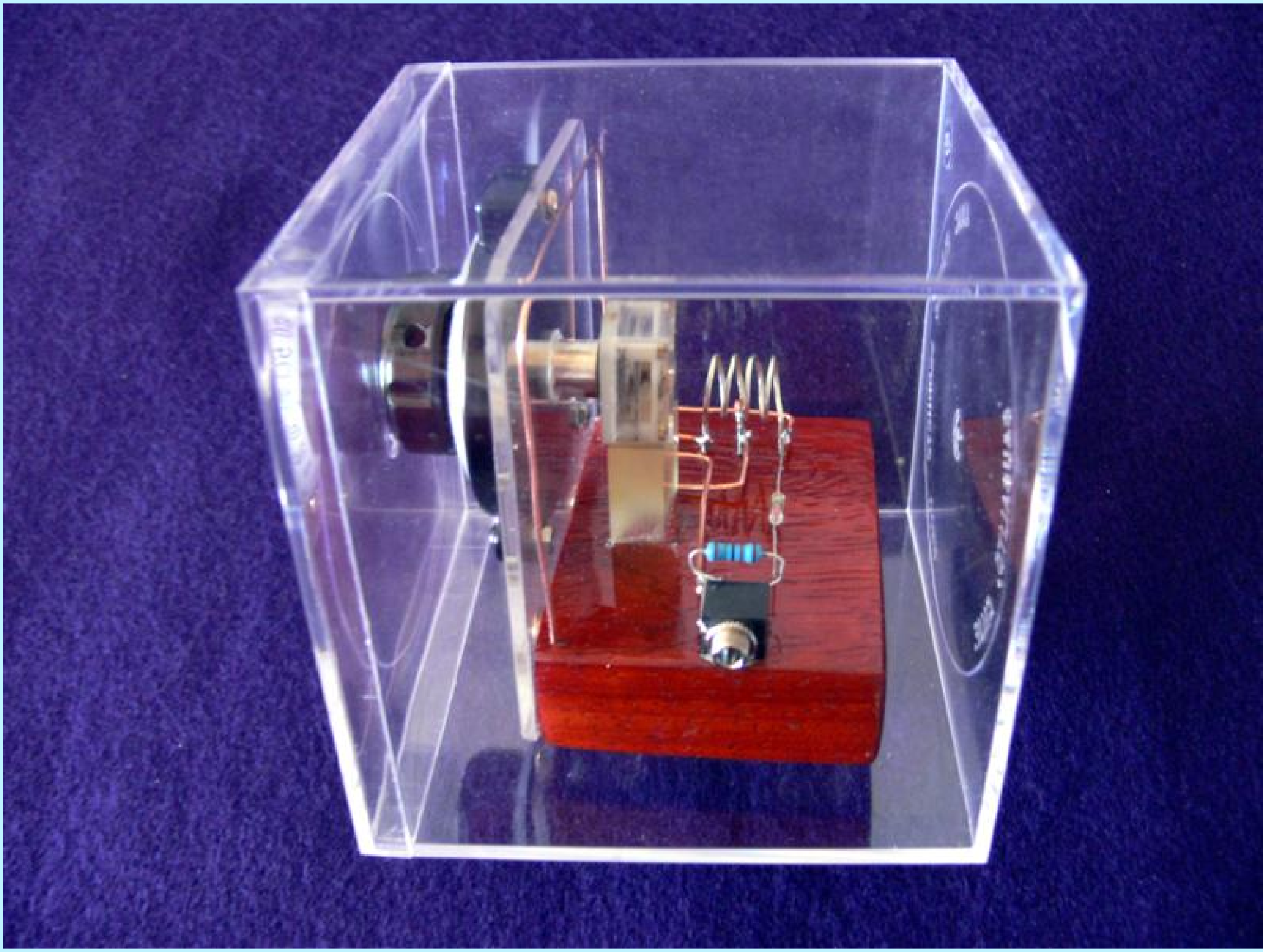

This picture shows the Solomon FM Crystal Set in an acrylic display case. I made the set specifically to fit inside this case (the case came first).

My definition of a crystal radio is one that is not powered, except by the radio transmission itself and employs a crystal detector. So, it should work without any batteries or AC power. An FM crystal receiver must be able to detect and receive FM signals well enough to be heard in earphones without any such extra power.

This FM receiver is an amazing performer. It has crystal clear reception (pun intended), good sensitivity, but only fair selectivity. This set was a discovery for me. I started out by designing and building the normal AM sets. Then one day while testing the "Mystery" set (see my other web links), to my surprise, in addition to the expected panoply of AM stations, I heard a very faint signal that I could not tune out. At first, it seemed too weak to identify. When I tuned out all the AM stations, I was astonished to hear the announcement "KiiM FM, 99.5"! This is a

country music FM station here in Tucson. It was all over the dial, untunable, but the much louder AM signals masked it when they were tuned in.

I set myself the task of trying to improve the FM reception. I tried some simple circuit modifications that did not seem to improve anything. Then I connected a dipole antenna instead of the AM antenna I normally use. Suddenly, the FM signal was much clearer, although still weak. By using the audio output and sound system amplifier, I was even more amazed that four different FM stations came in loud (or rather medium) and clear. I found that changing the telescoping antenna length and position I could tune the stations in and out. They were KRQ, KLPX, KiiM, and KHYT all local FM stations with transmitters nearby. Their reception was also affected by the length and position of the audio output cable.

After doing some research, I discovered that there was a physical theory that claimed that FM reception was possible and even probable using the same circuit as an AM receiver. The theory is called "slope detection". So, I set out to find circuit improvements. A web search yielded little, mostly theory. But there was enough information that I thought I could make some modifications to the AM circuits to make them more tunable to FM signals and less tunable to AM. Since FM operates at higher frequencies, all I had to do, I thought, was make the coil and caps smaller. After much "tinkering" I arrived at the current circuit.

The circuit looks identical to a classic AM crystal circuit but is even simpler to build. The components were reduced in dimension to resonate at higher frequencies. This was done by experimenting with smaller and smaller coils and capacitors. The antenna is also much reduced in size (from that of AM) to resonate at higher frequencies (the antenna is crucial). The air variable capacitor I used has two trimmers in it which should be adjusted for best reception.I have found that a commonly available vernier dial and knob will fit the capacitor nicely. See end of article for a picture of the variable. C3 is a ceramic capacitor of 18 pf, but may be anywhere from 10 to 50pf. A detected FM signal is converted to AM due to an effect called slope detection that modulates amplitude.

This FM Crystal Set works best near the transmitter (I have not tested it beyond about 10 miles). Secondly, the sound level is quiet, especially without an amplifier. A quiet room is needed for listening with earphones. One must be willing to move the set around to find a location for the best reception of signals. However, in addition to listening with high impedance earphones (crystal or otherwise), the set can be connected directly to an audio amplifier's low level magnetic input which can then play amplified through a sound system at any volume -- sounds GREAT. In fact, I recommend starting tests with the FM crystal set by connecting it to the low-level phono inputs of a receiver or preamplifier. (Nowadays, many receivers don't even have a phono input!) That way you can crank up the volume, which makes it more likely to find the FM stations. If no signals are detected, I also recommend connecting an external "rabbit ear" antenna or hanging a short wire (12 inches or so) in various positions next to the internal antenna. The variable length of rabbitt ears can help to tune in stations.

No additional wiring or antenna is necessary (the antenna is optimized in length for FM.)

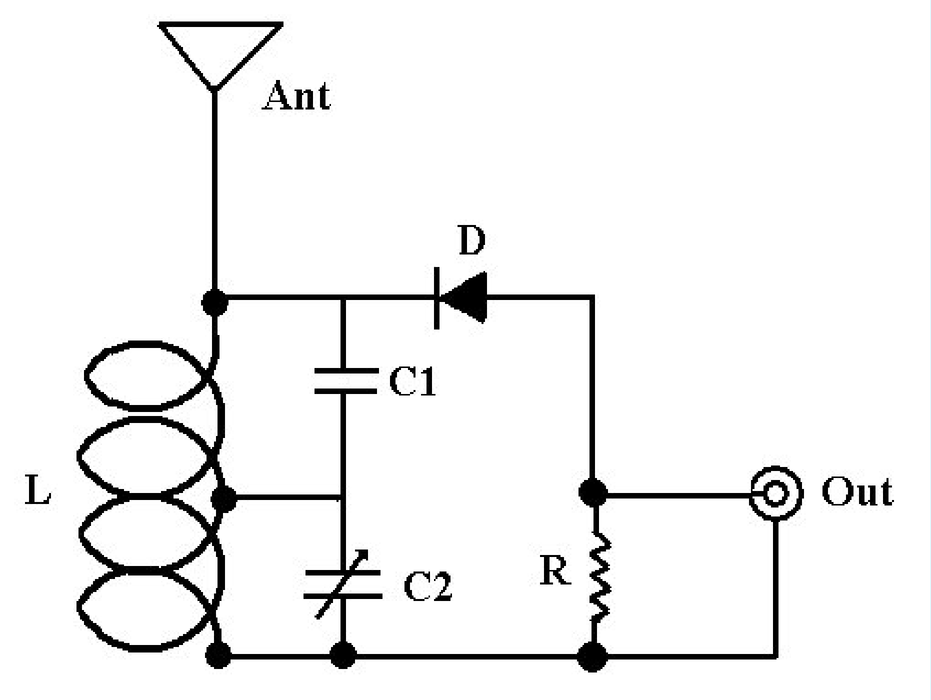

L - 4 turns #18 copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
Ant - 7 inches of #18 bare copper wire
C1 - 18 pf ceramic capacitor
C2 - 50 pf air variable capacitor
D - 1N34 diode or rock crystal
R - 150K resistor

The diode is tapped directly to the antenna. The vernier dial fits directly on the tuning capacitor. The antenna parallels the perimeter of the acrylic face plate. "Military style" #18 AWG wiring is used without any insulation. It is important to keep the components physically close together. The component specifications are the same as in circuit #2. The coil is silver rather than copper, but copper does just as well. I think that the contrast of the silver and copper is beautiful. The coil was wrapped around a Sharpie Permanent Marker, then slipped off and expanded slightly. The wooden base is made from lacquered, polyurethane padouk.

I consider this set a work of art as well as science and think it is the most elegant crystal receiver I have created. I love the contrast of the silver coil, the copper antenna, the clear acrylic faceplate, the black vernier dial, the white and transparent variable capacitor, and the subtle colorings on the resistor, the diode, and the lucite base. Yet the circuit is so ridiculously simple that some will not believe it is possible without building it themselves. No shielding is necessary, and there is no problem with hand capacitance. However, the output cable position may affect reception sensitivity.

Photos of wired circuit

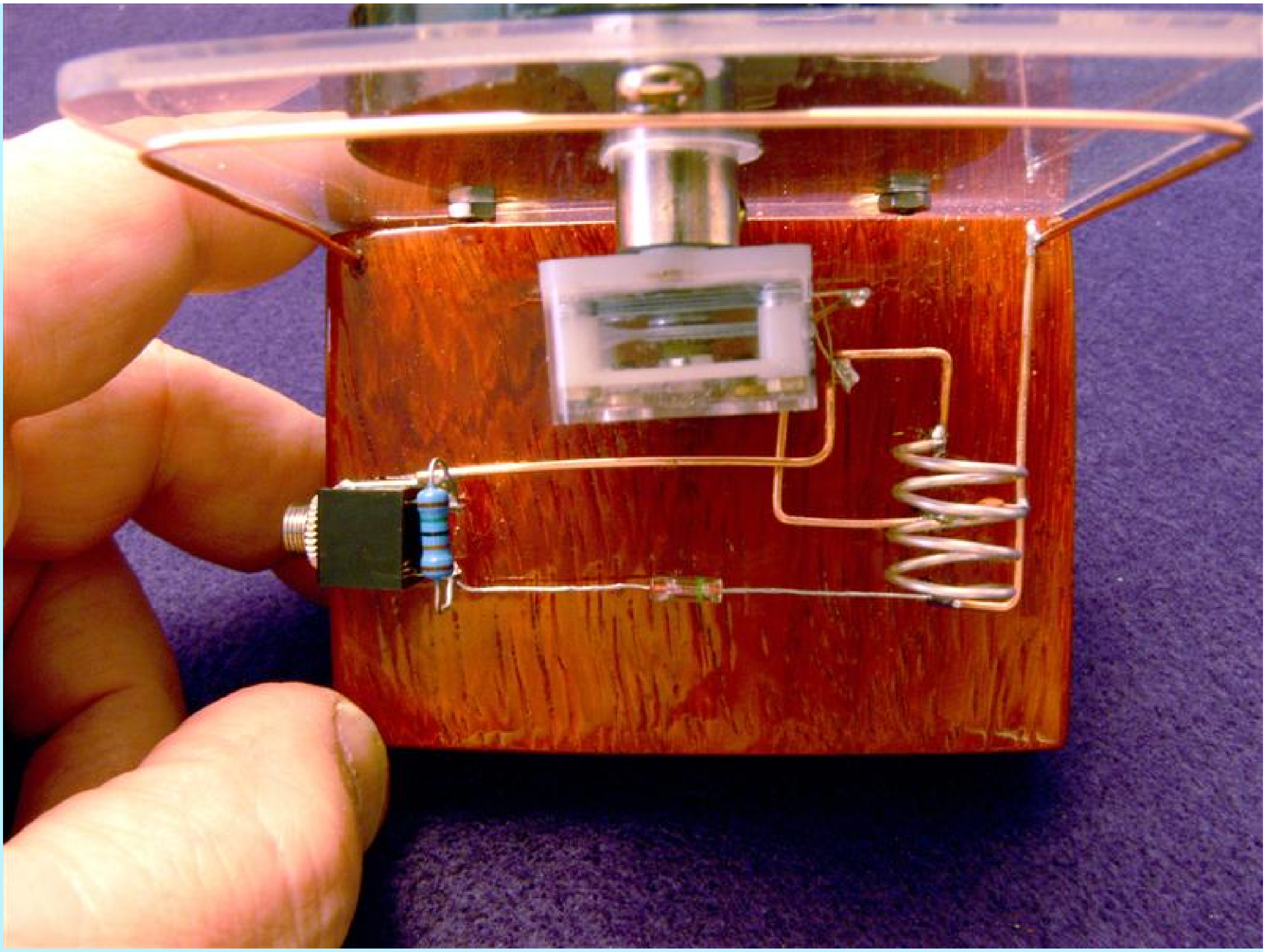

A hand is included in this photograph to show scale. Note the military style wiring, diode, and antenna. I wanted the wiring to create a modern design similar to a Mondrian painting. Not only is this set beautiful, it works! No power and no long antenna! It looks like a work of fiction.

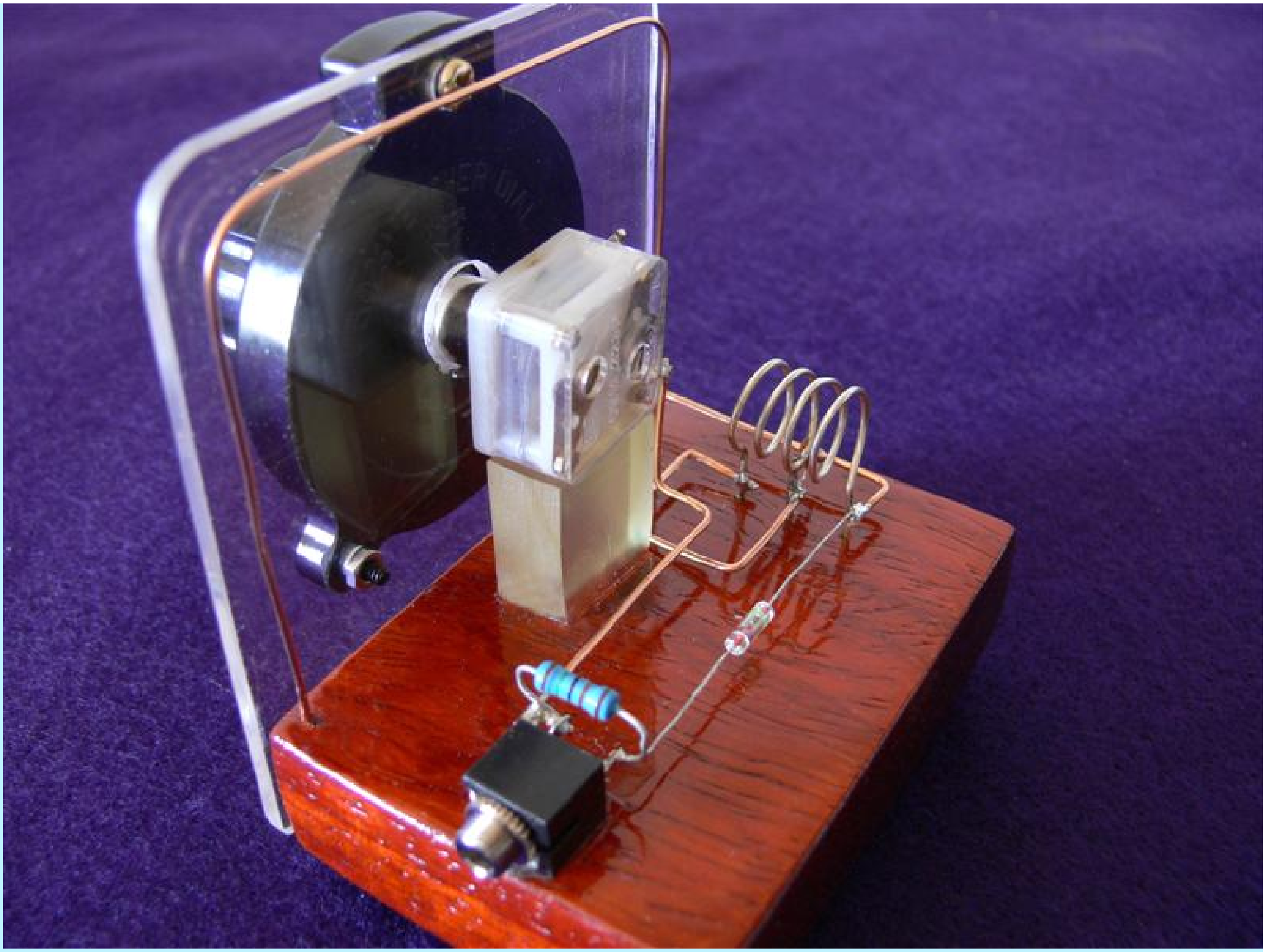

Is this thing imaginary -- science fiction? Well, imagination did play a part, but it is definitely not science fiction. This shot shows the elegance of the FM set best, I think. There is only one resistor and one fixed capacitor.

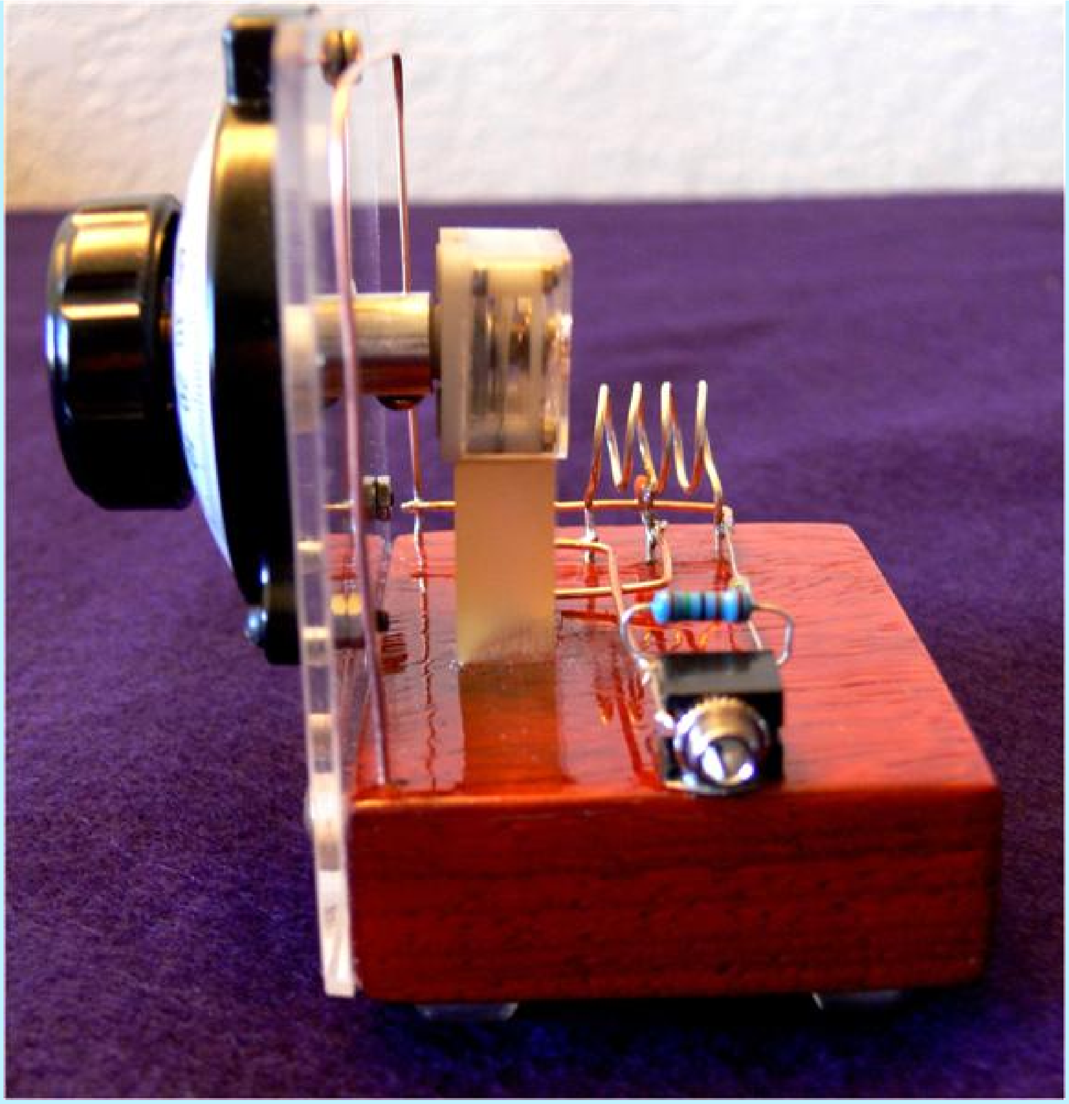

The inside of the tuning capacitor and the phono jack/output can be seen here. Can you spot the fixed ceramic capacitor? Note the polished edge of the face plate and the reflection in the wooden base.

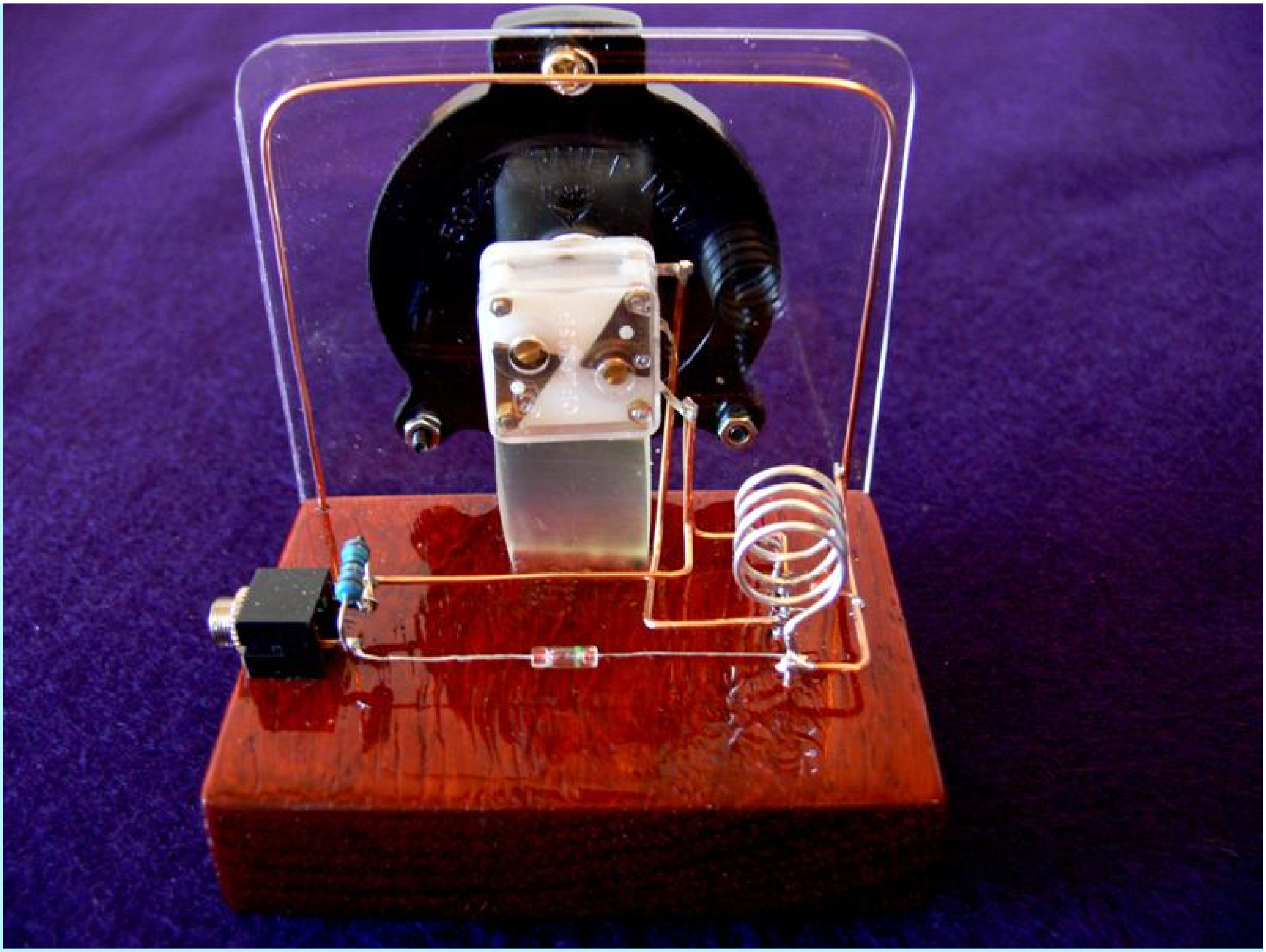

A quarter-inch piece of lucite was fitted under the tuning capacitor to anchor it. Note the two tiny trimmers on the back of the tuning capacitor. Brass screws were used to enhance appearance.

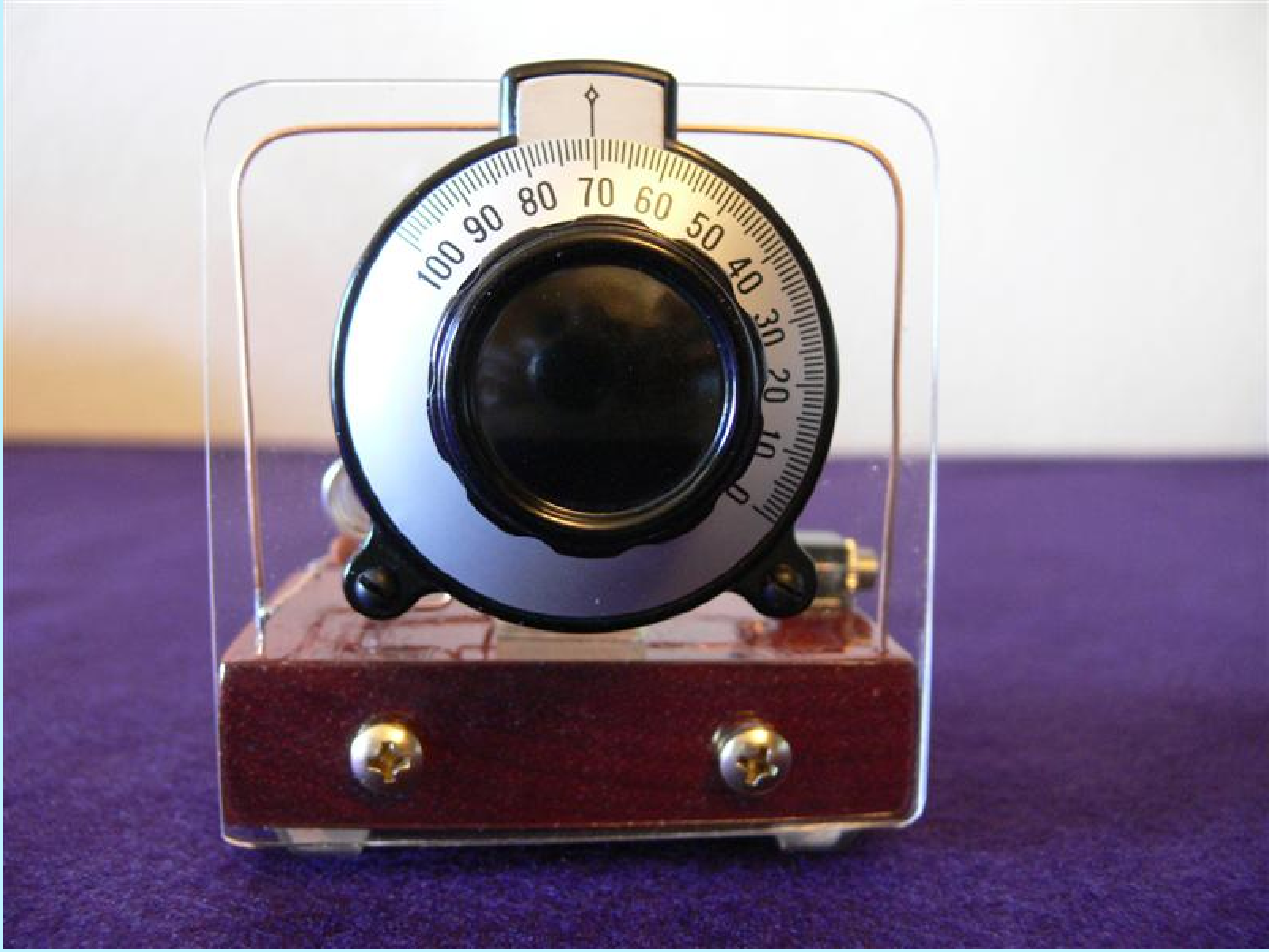

The vernier dial is large to accomodate ease of tuning, and the vernier makes it easy to separate stations. Two golden (brass) wood screws fix the face plate to the base. Holes for the face plate were made with special plastic drills, but ordinary drills may be used if drilled very SLOWLY. The knob is removable.

FM Crystal Circuit #2

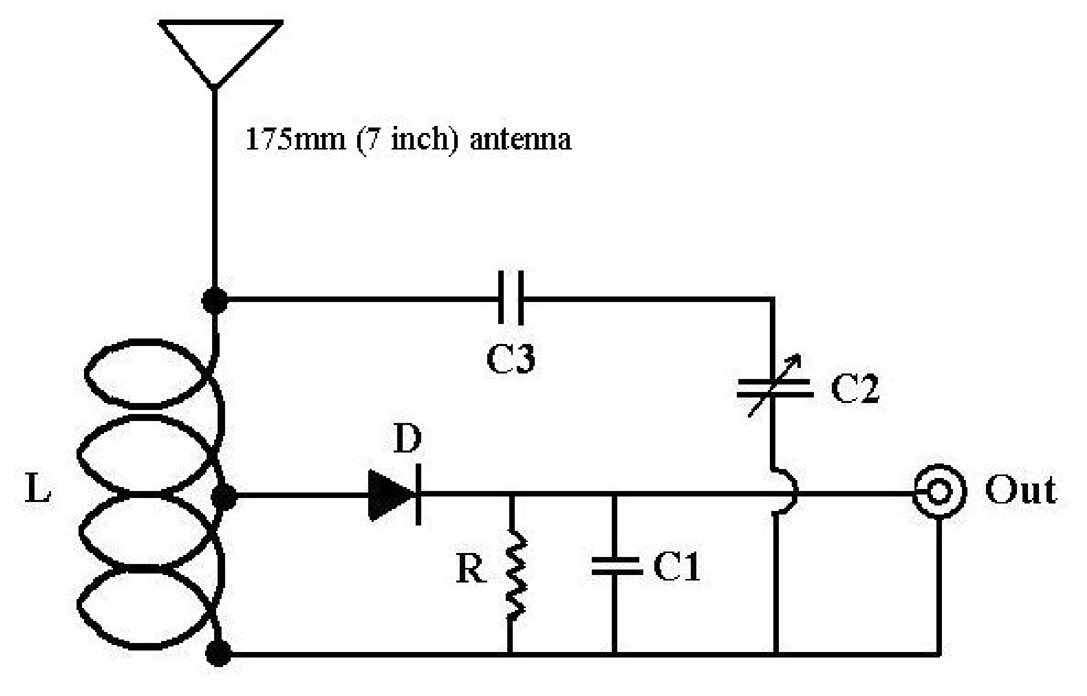

L - 5 turns AWG#18 bare copper or silver wire, 12mm inside diameter, tapped at 2.5 turns
D - 1N34 or rock crystal diode
C1 - 82 pf capacitor
C2 - 80 pf air variable capacitor
C3 - 18 pf capacitor
R - 150K resistor

The following photographs show the circuit wired with the handmade Saturn Dial. and knob. It is perhaps not as visually striking as set No. 1, but it works just as well. In fact, this set was the original version. Notice that all the wiring and coil are copper.

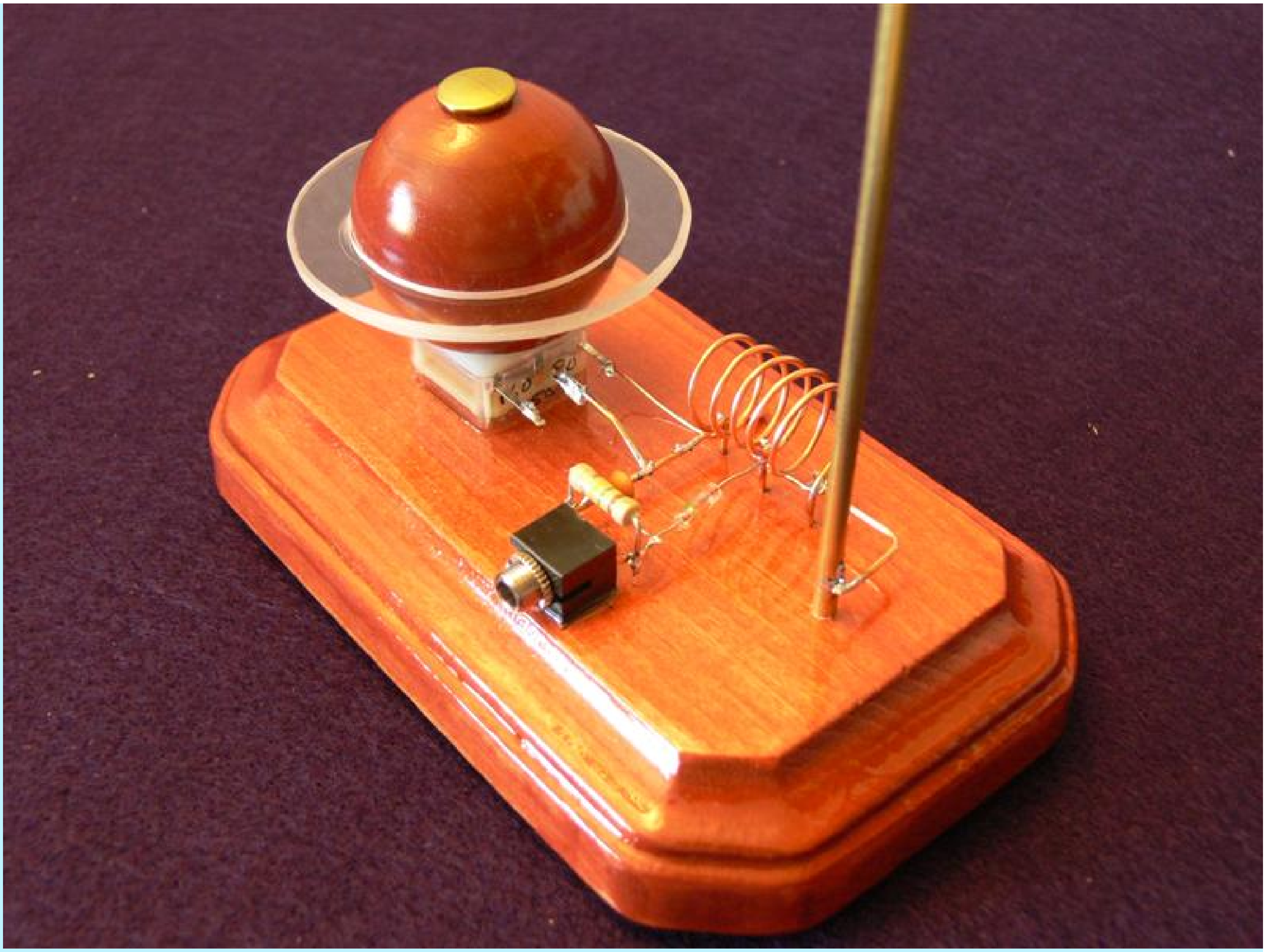

The Saturn dial and knob were fashioned from a "doll's head" from Michael's Arts and Crafts, a piece of lucite cut with two circle cutters, and a brass paper fastener. The knob is fixed to the tuning capacitor with a small machine screw that fits in the hole below the brass fastener. The most difficult part of this was fashioning "Saturn's rings". This must be done very carefully and slowly. The inside edge should be cut slightly undersized and then sanded with a drum sander to fit snugly. The outside edges can be sanded with fine sandpaper and polished with a plastic polisher.

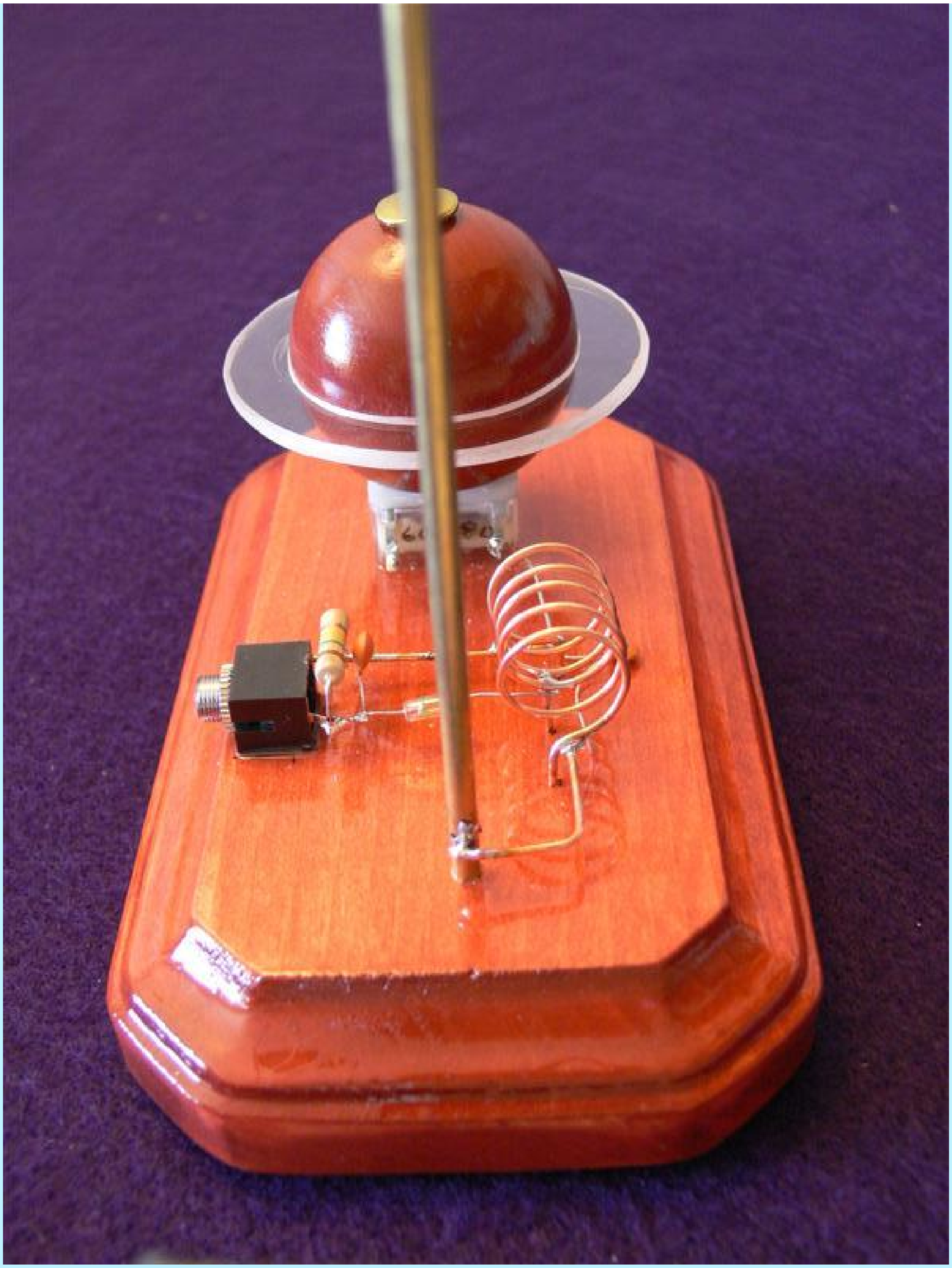

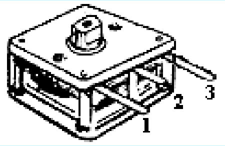

The air variable capacitor may be obtained from Electronix Express at http://www.elexp.com/. Part number 14VCRF10-280P. The 80 pf side is recommended for the second circuit, contacts 2-3. Contacts 1 and 3 were used for the first circuit (50pf).

- OSC: 5-59 pf
- ANT: 5-142 pf
- OSC and ANT Trimmer 10pf range

Home -- If you want to contact me go to the home page and go through the Music Clinic link.

life.augmented
INNOVATIVE SEMICONDUCTOR SOLUTIONS
Digi-Key ELECTRONICS
Learn More
advertisement

FREE ENERGY FROM THIN AIR!

Technology (/technology/) > Science | by DrewPaulDesigns (/member/DrewPaulDesigns/) Follow

79,181 1,078 230

Posted May. 24, 2016 | BY-NC-SA

FREE ENERGY
FROM THIN AIR!
*INVOLVES CRYSTALS AND DANGEROUS VOLTAGES!
Copyright 2016. Drew Paul Designs, All Rights Reserved.

(https://cdn.instructables.com/FAT/Z4RP/IOKBVYWL/FATZ4RPIOKBVYWL.LARGE.jpg)

Download

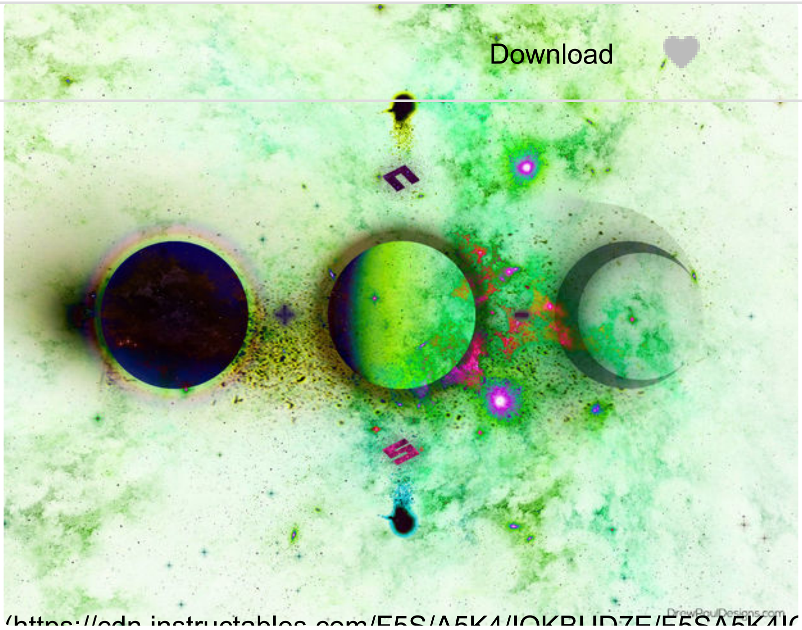

(https://cdn.instructables.com/F5S/A5K4/IOKBUDZF/F5SA5K4IO

"Dendera Light"
circa 1500 BC
Image: creative commons, Wikipedia

(https://cdn.instructables.com/FOH/L0OE/IOKBVTOS/

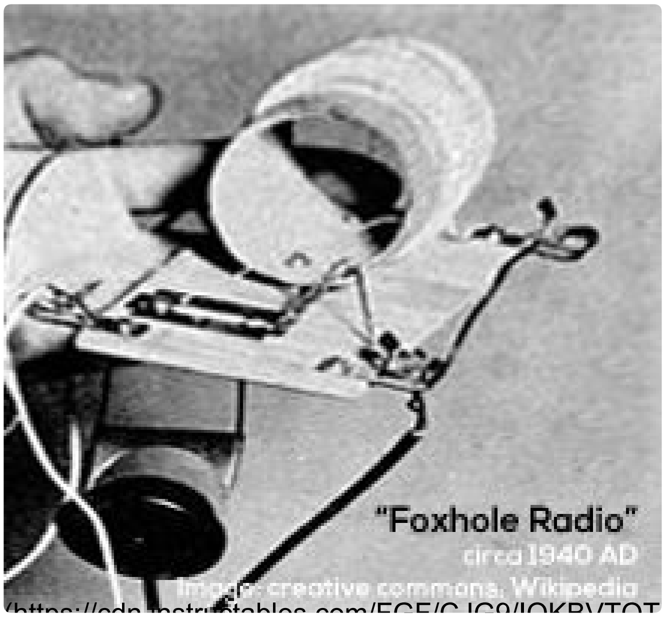
"Foxhole Radio"
circa 1940 AD
Image: creative commons, Wikipedia

(https://cdn.instructables.com/FGF/CJG9/IOKBVTOT

3.74
CEN-TECH
OFF ON

(https://cdn.instructables.com/FHS/HM1V/IOKBT6GX/FHSHM1V

"Free energy from the air?, Yea, right!" Sardonic skepticism was my first reaction to this unusual concept, as well.

Though, its not so far out there, in fact. Light can be converted to DC current with solar panels, electricity can be converted to magnetism as I did in my last article (https://www.instructables.com/id/DIY-Electro-Magnetic-Levitation/), in a microphone sound waves are converted to an electrical signal (by vibrating a magnet near a coil (http://hyperphysics.phy-astr.gsu.edu/hbase/audio/mic.html)), solar rays can even be focused and converted to heat in awesome devices like this! (http://www.gosunstove.com/) When we think about it, energy is all around us and can be harvested in an enumerable many of ways.

Today, we are going to take a rather novel approach. We are going to build a device specifically designed to sense and capture a particular band of energy which is all around us.

The earth is magnetic and anyone who has ever used a compass knows this. Magnetic bodies in motion produce electricity, we can see this in any alternator, like the one in your car. So, therefore the earth is electric as well as magnetic, by definition.

Can we detect this energy? Yes, we sure can! Ever turn on a radio in the middle of nowhere and heard static? That is your radio picking up naturally occurring energy in the RF spectrum!

Can we use this energy to do work? Absolutely! This has been known for a long time. Crystal radios (https://en.wikipedia.org/wiki/Crystal_radio) have been around since before the 1930's and can run with no input energy other than the radio signal. Even when completely isolated, but from the atmosphere, a crystal radio will produce a voltage in the earpiece resulting in a sound (albeit and undesirable one).

Well, this is where it gets interesting...

Can we replicate this effect? Yea, and with modern components like the high quality crystals found in germanium diodes, we can even increase efficiency. By applying this concept as a Crystal Energy Receiver we can take advantage of a wide range of energetic frequencies rather than tuning in to just one.

Can we scale it up? Definitely. Things like micro germanium diodes, high efficiency antennas and compact contemporary capacitors make the components that are required to build a crystal receiver fit in the palm of your hand. While there may or may not be a more efficient way, this renewable energy solution is simple to employ and can be scaled up, or down, indefinitely.

Download

It sounds like we can build a Crystal Energy Receiver. Let's give it a shot...

Add Tip
Ask Question

Step 1: What You'll Need

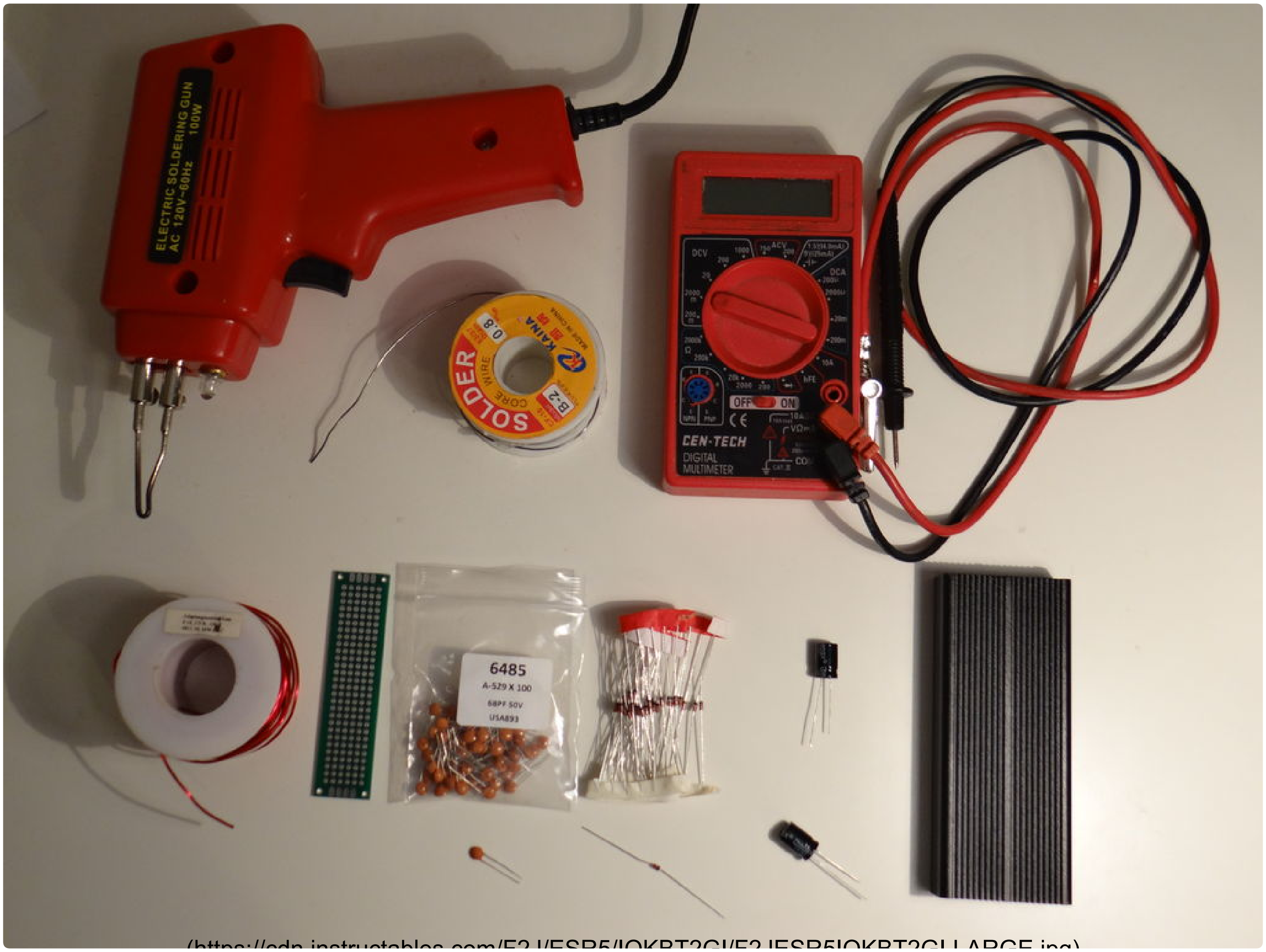
ELECTRIC SOLDERING GUN
AC 120V~60Hz 100W
SOLDER WIRE
KAINA
0.8
B-2
CEN-TECH
DIGITAL MULTIMETER
OFF ON
6485
A-529 X 100
68PF 50V
USA893

(https://cdn.instructables.com/F2J/FSP5/IQKBT2GI/F2JFSP5IQKBT2GI.LARGE.jpg)

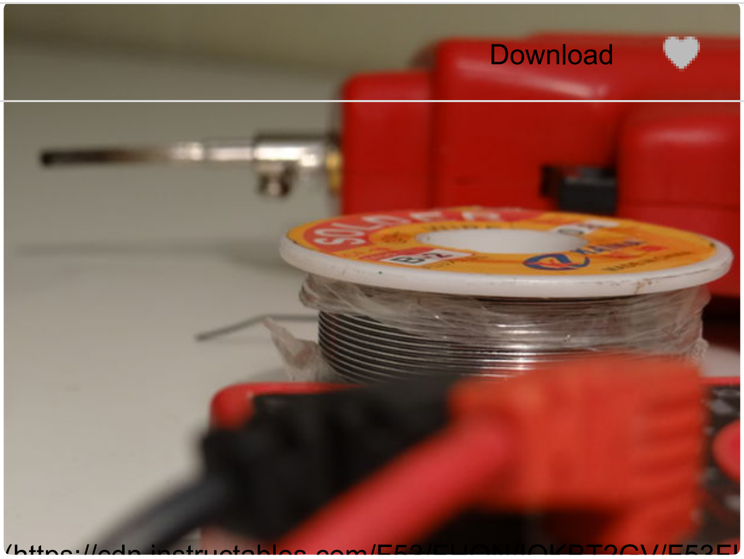
(https://cdn.instructables.com/F53/FHGM/IOKBT2CV/F53F

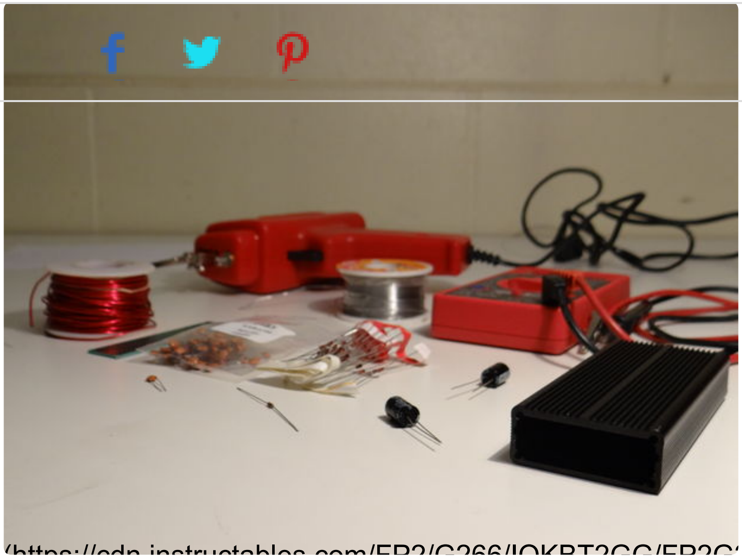
(https://cdn.instructables.com/FP2/G266/IOKBT2GG/FP2G

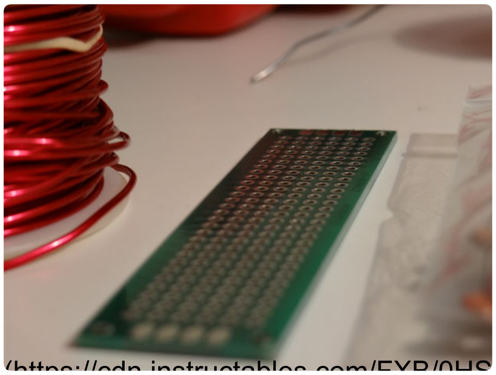
(https://cdn.instructables.com/FXB/0HS

(https://cdn.instructables.com/FXS/DZC

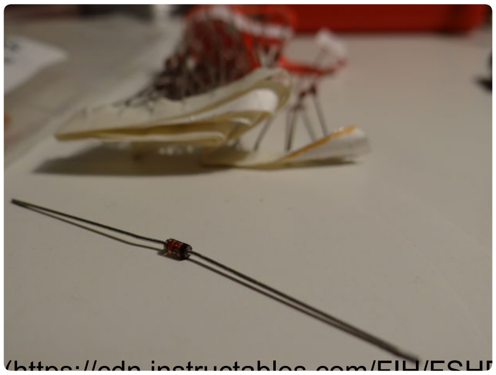
(https://cdn.instructables.com/FIH/ESHF

Download

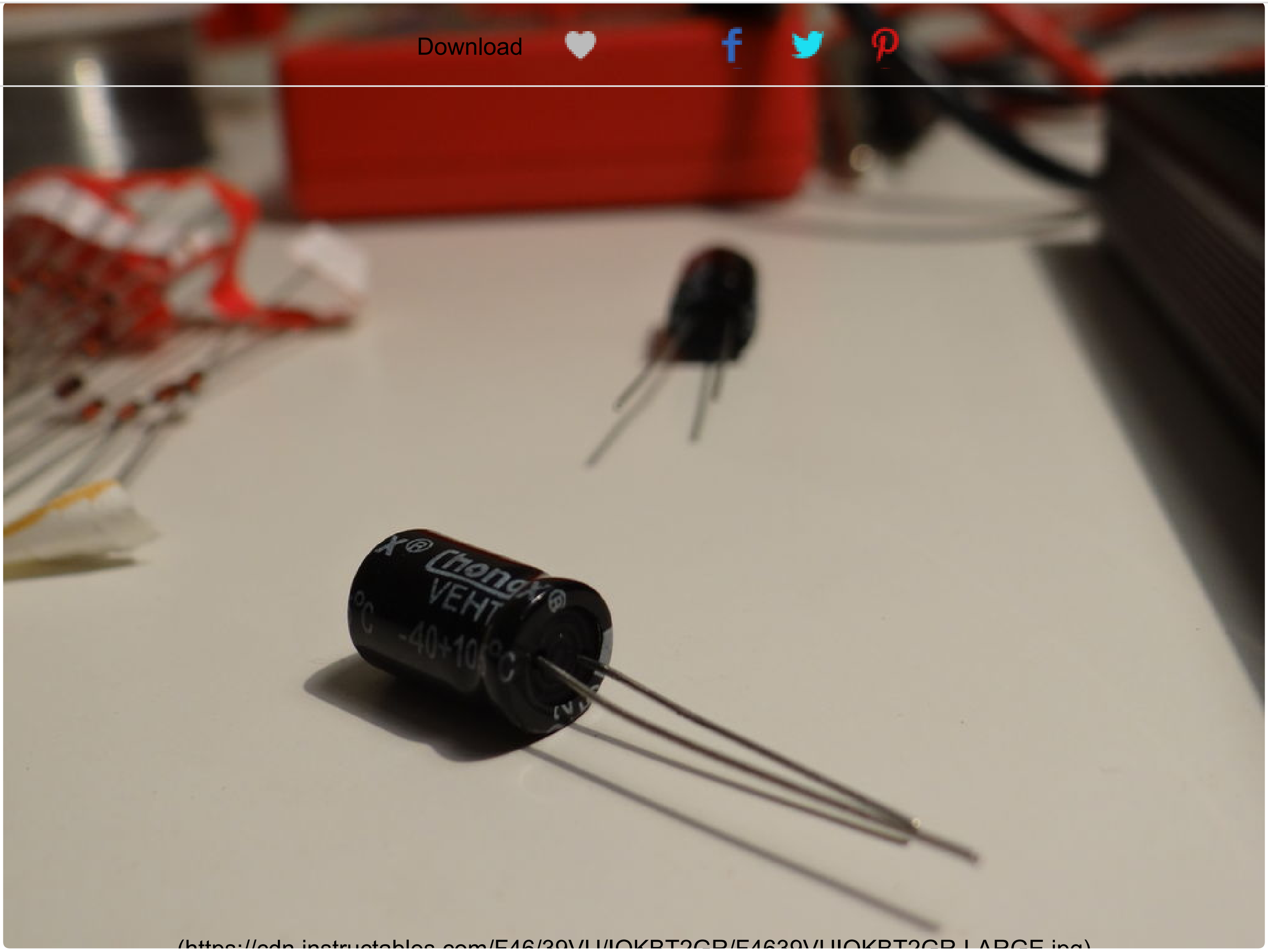

Chongx
VEHT
-40+10 °C

(https://cdn.instructables.com/F46/39VU/IOKBT2CB/F4639VUIOKBT2CB.LARGE.jpg)

Download

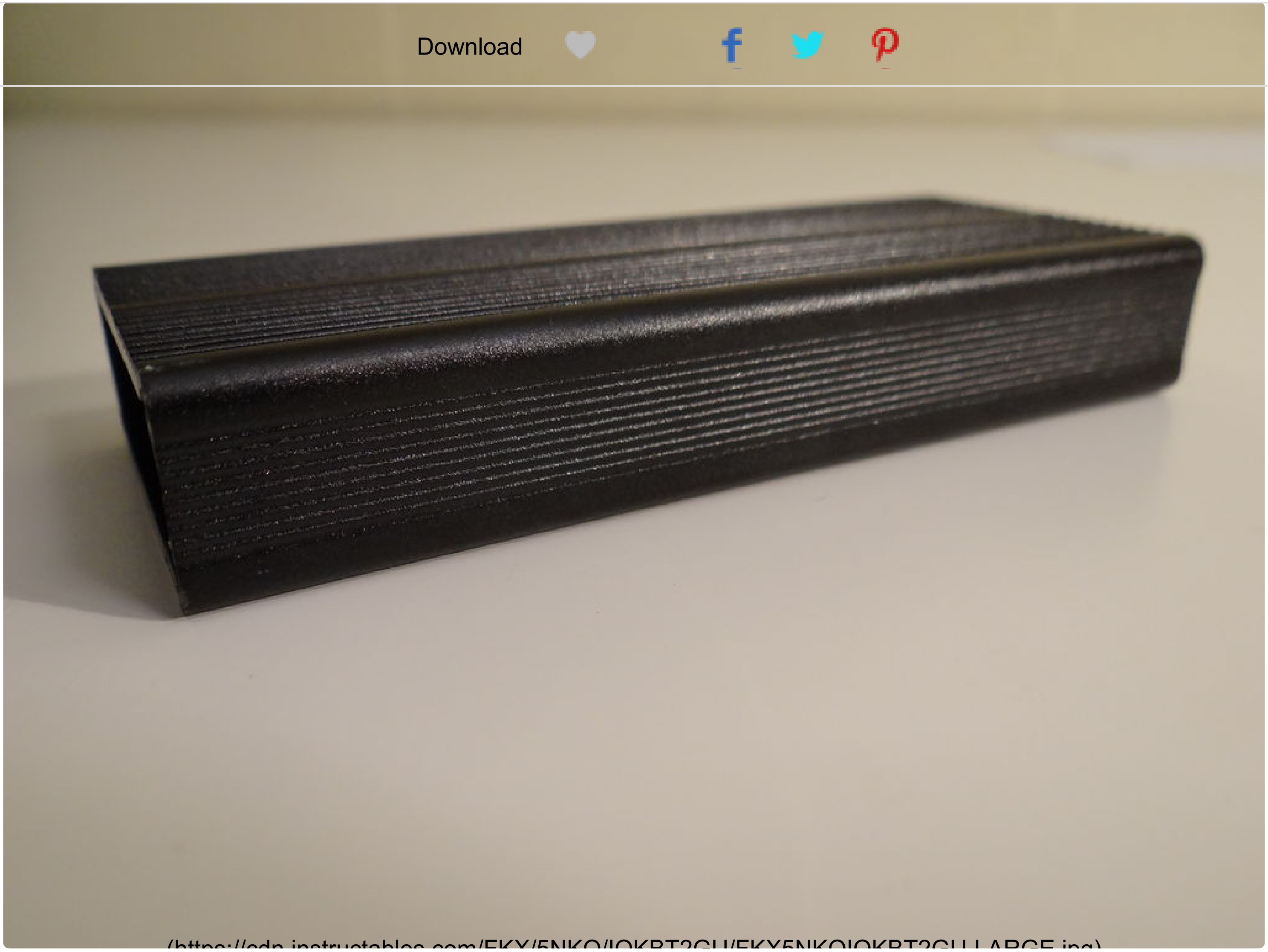

(https://cdn.instructables.com/FKY/5NKQ/IQKBT2GU/FKY5NKQIQKBT2GU.LARGE.jpg)

One of the reasons this particular renewable energy harvesting method is so viable is the relatively few and easy to obtain materials required.

The simplest crystal receiver design needs no power and can be built with only three parts: a coil, a crystal and a resistor. We're going to optimize that design in order to produce a cleaner and more reliable output signal by first polarizing the input amplitude, then rectifying and filtering the signal. Then we'll add an antenna, case and connections.

Get the circuit diagram here (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)
Get the kit here (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)

The parts for the circuit include:
(1) Circuit Board (http://www.drewpauldesigns.com/crystal-energy-receiver-kit.html)
(1) 10-18 gauge Copper Wire
(2-12+) Ceramic Capacitors (matched)
(2-6+) Electrolytic Capacitors (matched)
*note various types of capacitors can be used
(4) Germanium Crystal Diodes (1A+)

Total Unit Cost: +/- $0.40 (USD, scaled for volume of 1,000+ units)

In addition, you'll probably want to get:
(1) Project box (optional)
(1) Antenna (a loop antenna or elevated antenna is recommended and can be made with copper wire)

The tools you'll need are:
Soldering Iron/ Solder (optional)
Multimeter
Oscilloscope (http://www.seeedstudio.com/depot/DSO-Nano-v3-p-1358.html?cPath=63_65) (optional)

That's it. Yup, that's all. Once we've got it all, let's begin.

Add Tip Ask Question

Step 2: Build the Circuit

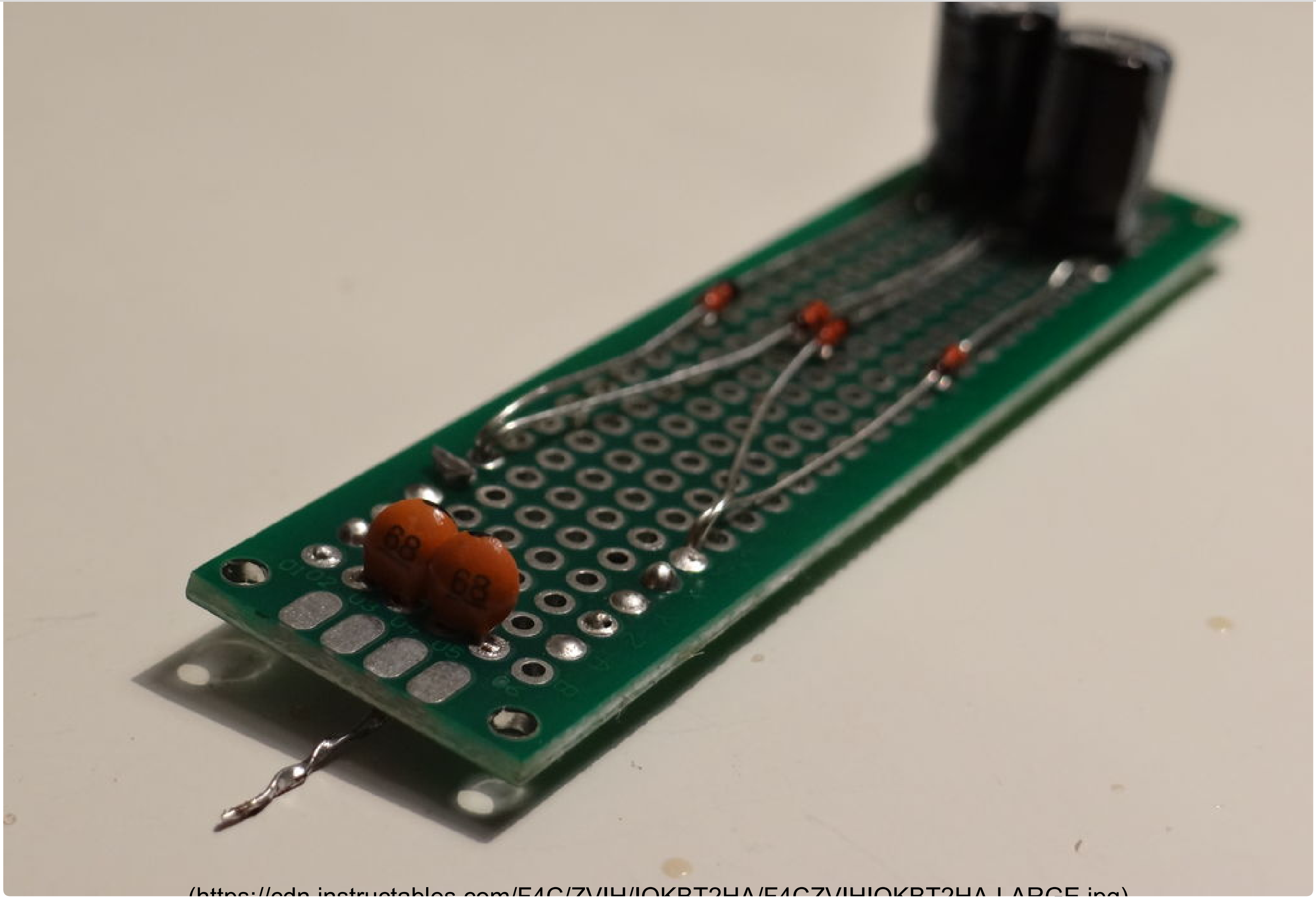

(https://cdn.instructables.com/F4C/ZVIH/IQKBT2HA/F4CZVIHIQKBT2HA.LARGE.jpg)

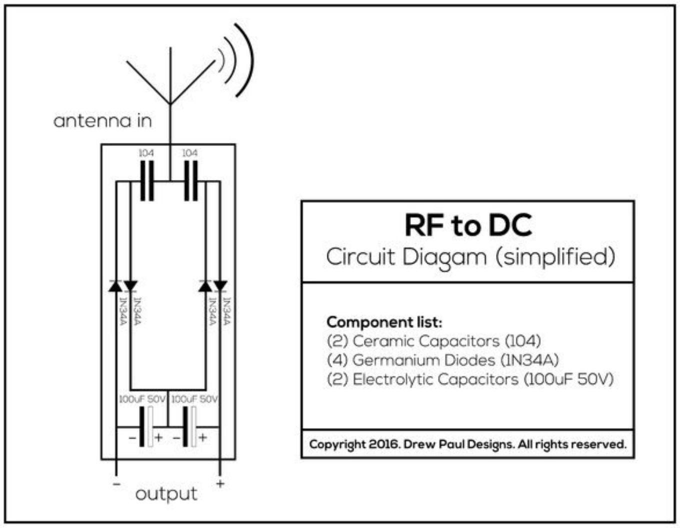
antenna in
104
104
1N34A
1N34A
1N34A
1N34A
100uF 50V
100uF 50V
output
RF to DC
Circuit Diagam (simplified)
Component list:
(2) Ceramic Capacitors (104)
(4) Germanium Diodes (1N34A)
(2) Electrolytic Capacitors (100uF 50V)
Copyright 2016. Drew Paul Designs. All rights reserved.

(https://cdn.instructables.com/FAW/ABS2/IQKBUKAD/FAV

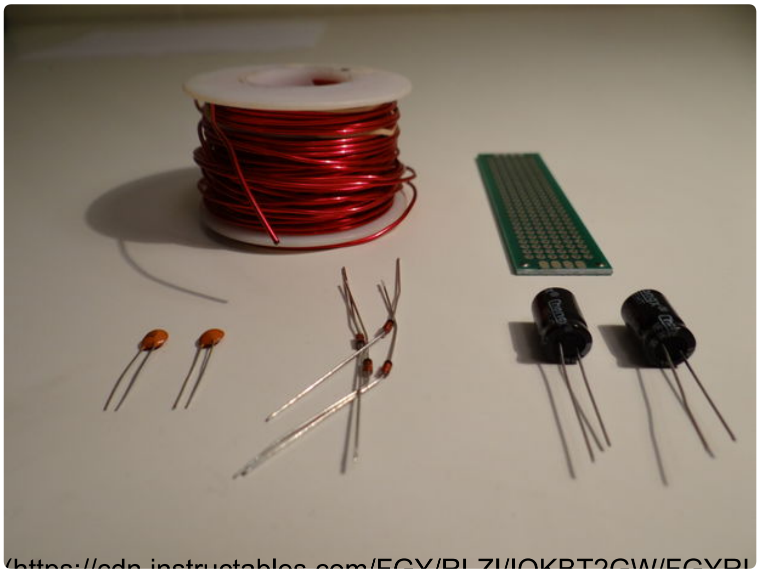

(https://cdn.instructables.com/FGY/BLZI/IQKBT2GW/FGYBI

Download

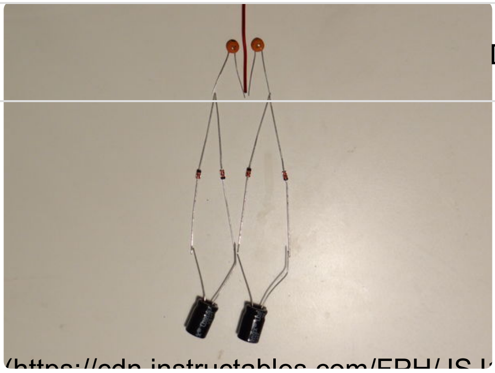

(https://cdn.instructables.com/FPH/JS I

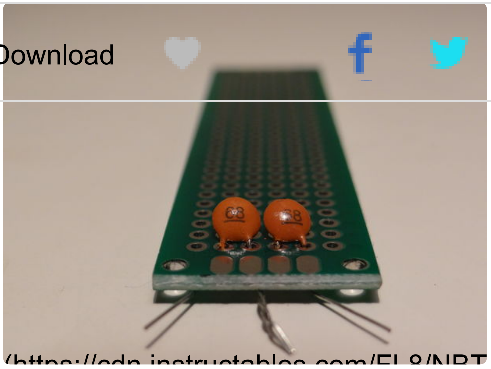

(https://cdn.instructables.com/FL8/NBT

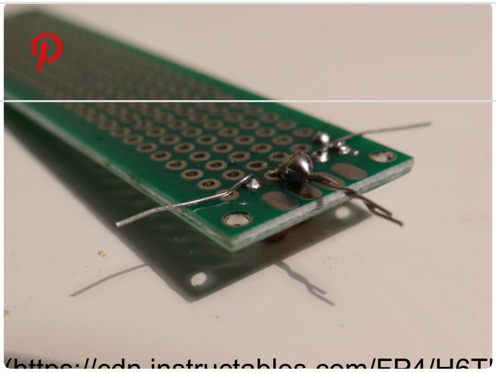

(https://cdn.instructables.com/FP4/H6T

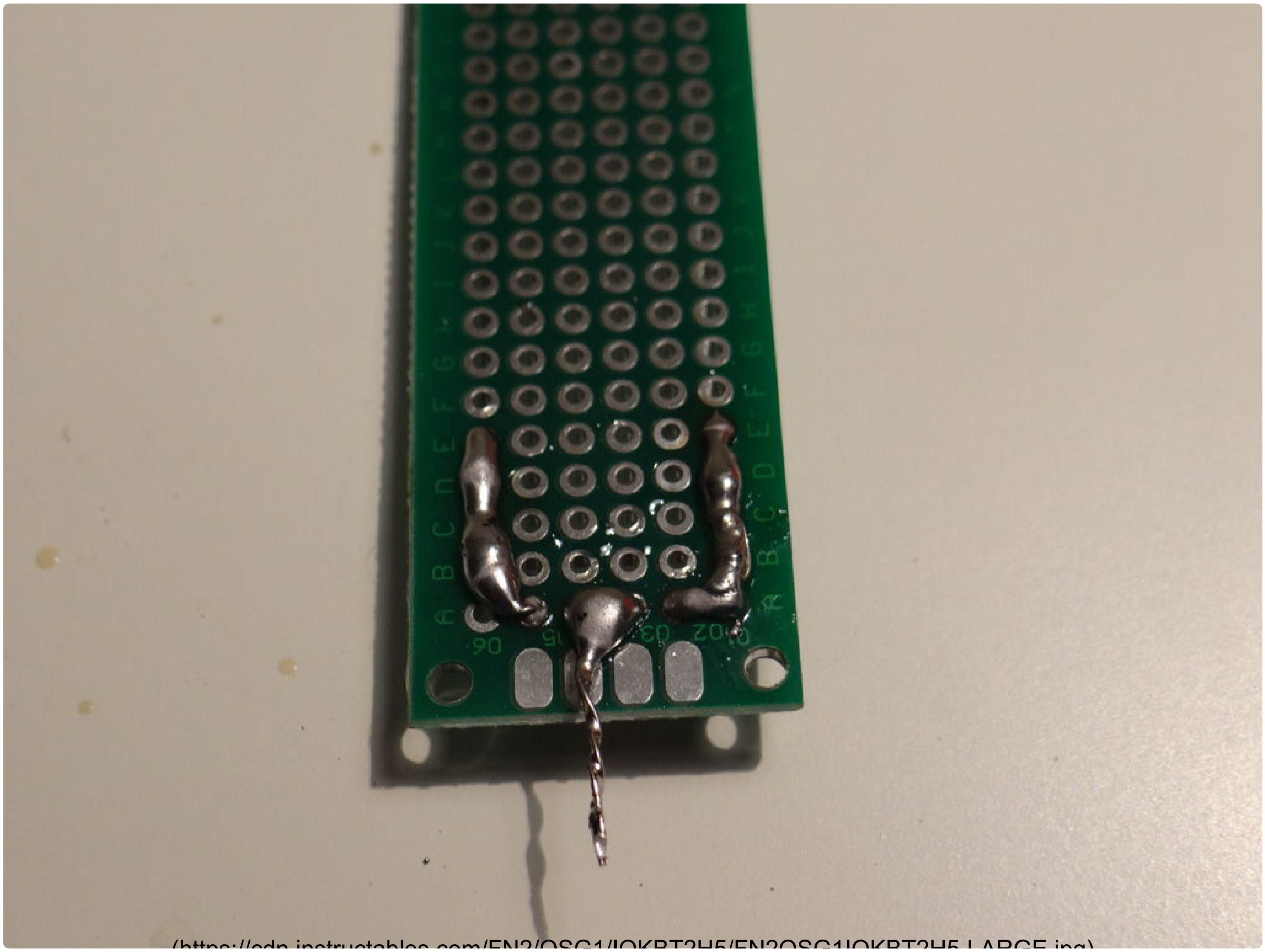

(https://cdn.instructables.com/FN2/OSG1/JOKBT2H5/FN2OSG1JOKBT2H5.LARGE.jpg)

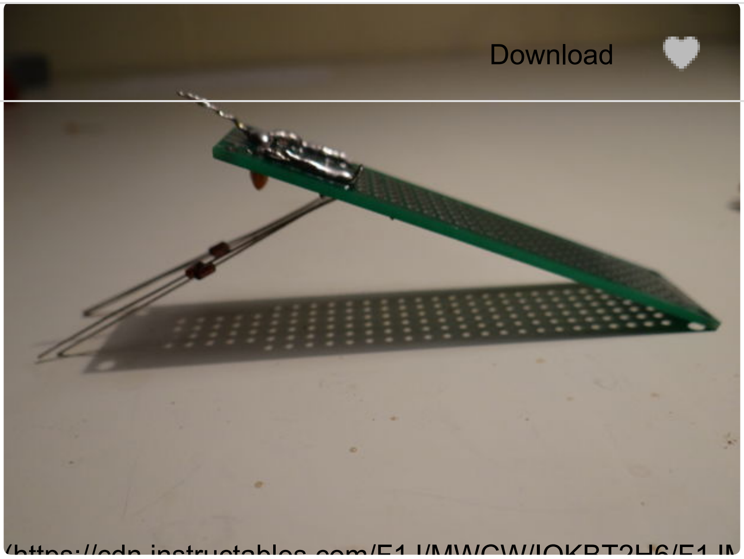

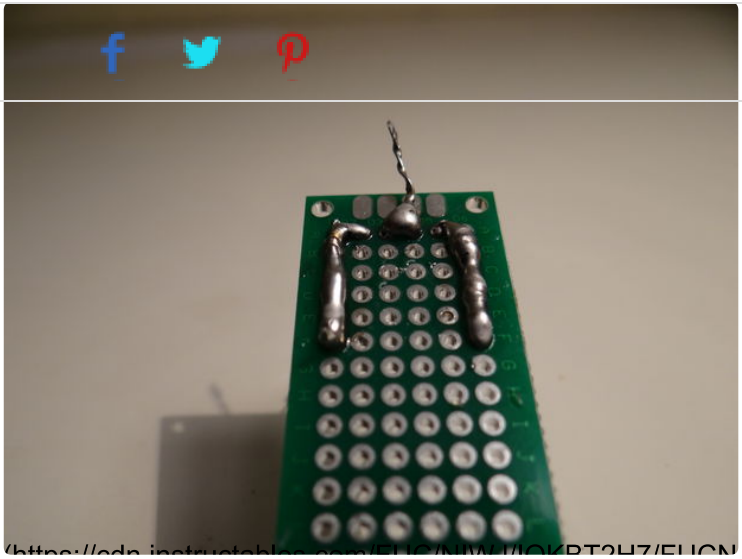

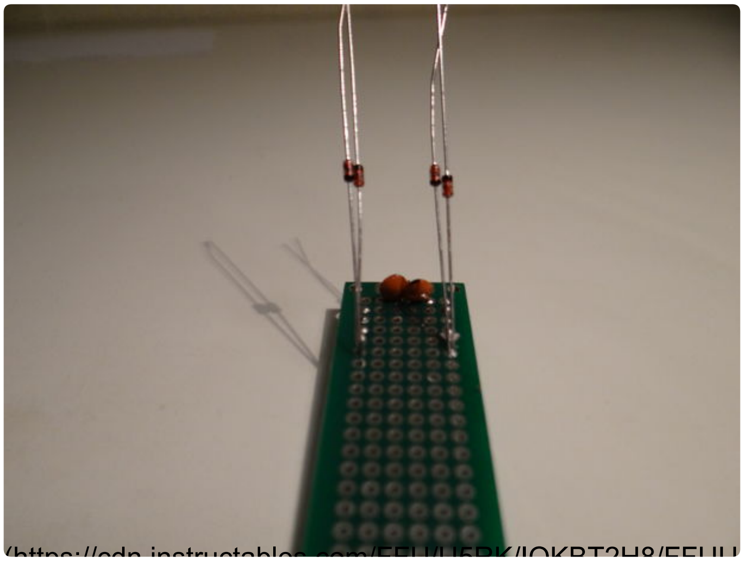

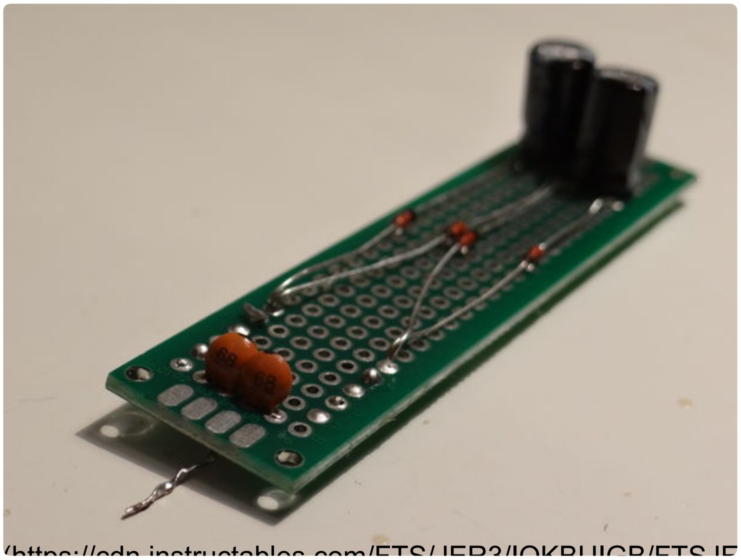

We're going to build the simplest version of this circuit variation in order to understand how each component interacts and as a proof of concept.

There are three simple systems at work in the circuit that are composed of capacitors, which store energy, and diodes that direct it.

Energy in the band of radio waves, among others, will vibrate a wire antenna on an atomic level, sending a discernible signal to its lead. This signal will then meet the junction between two ceramic capacitors wired in series. This junction will force positive charge from the wave to travel in one direction and negative charge in the other direction which, when collected again, makes the signal uniform and polar. Connecting the two capacitors in series creates leads on each end; the now positively charged side of one and the now negatively charged side of the other creates a two cell battery. Download

This next stage of the circuit takes a signal with a net value of zero, adds the absolute values of the positive and negative amplitudes with respect to the origin and produces a positive integer. This concept can be thought of as taking:

(+1) + (-1) + (+1) + (-1) = 0

and converting it to:

| [(+1) + (+1)] | = 2
+
| [(-1) + (-1)] | = 2
= 4

Isn't math fun?

To each of these leads from our two capacitors in series, we will connect two crystal diodes, one facing each direction, to form what is called a bridge rectifier. A bridge rectifier is a configuration which will convert an alternating current to a direct one by cleverly rerouting the signal.

By connecting the bridge rectifier as shown in the circuit diagram, this direct current from the diodes then charges the electrolytic capacitors. This stage normalizes the amplitude, making the current constant and usable.

Components can easily be twisted together for testing and then soldered to a circuit board to secure.

Add Tip Ask Question

Step 3: Test and Optimize Your Circuit

Download

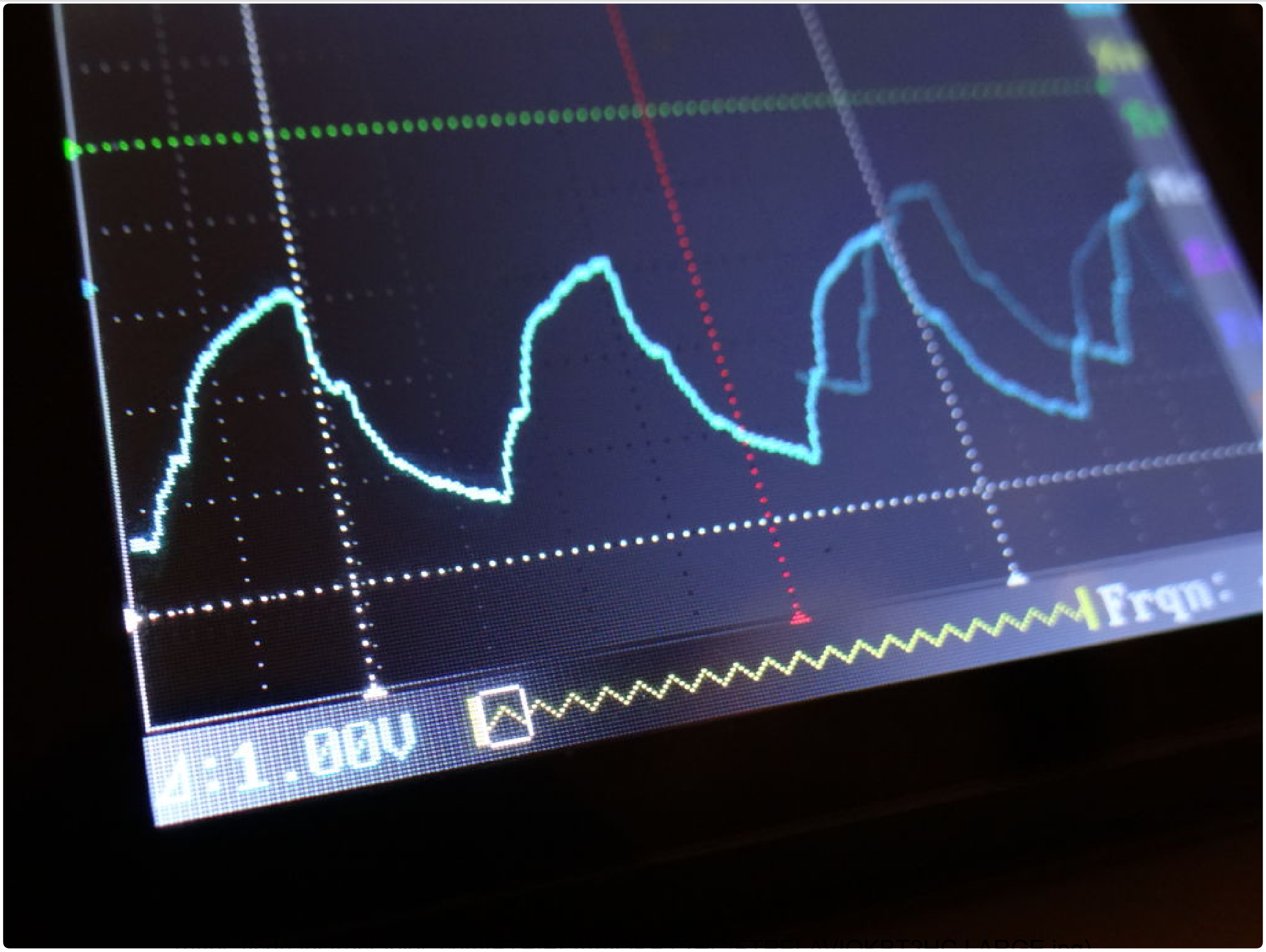

ETBFLAV/IQKBT2HG.LARGE.jpg)

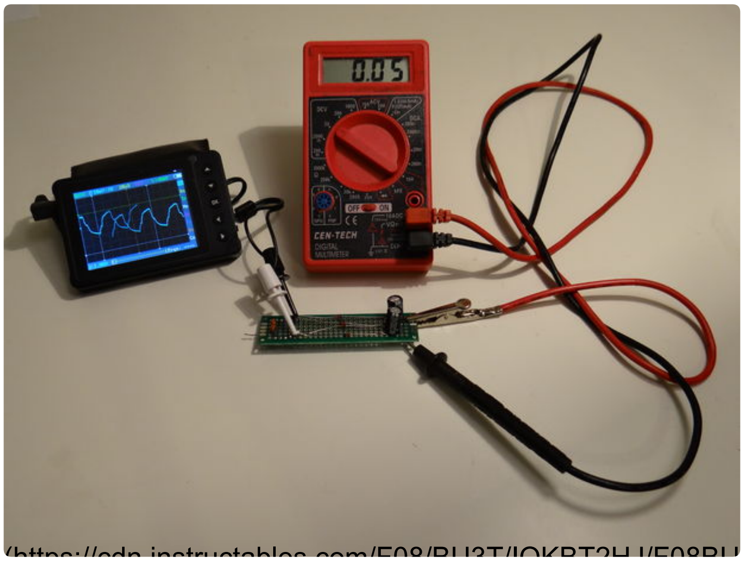

(https://cdn.instructables.com/F08/RU3T/IQKBT2HJ/F08RU

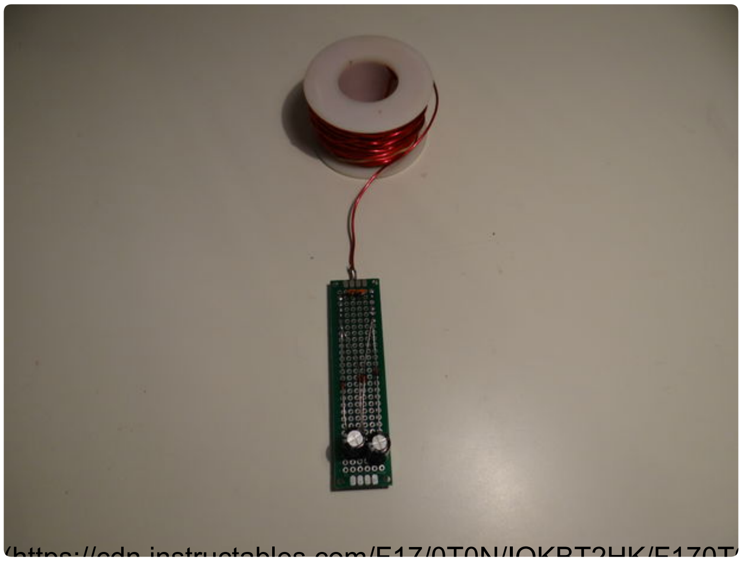

(https://cdn.instructables.com/F17/0T0N/IQKBT2HK/F170T

Download

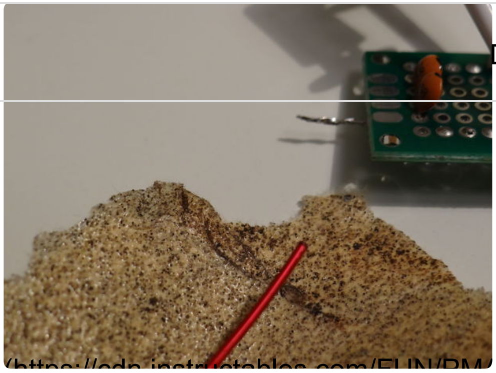

(https://cdn.instructables.com/FUN/PM

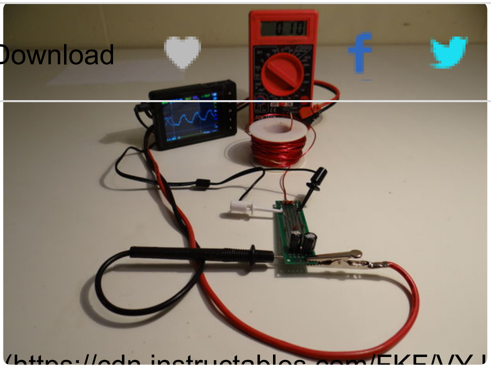

(https://cdn.instructables.com/FKF/VX

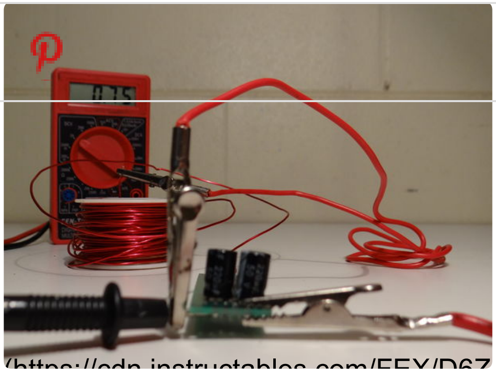

(https://cdn.instructables.com/FFX/D6Z

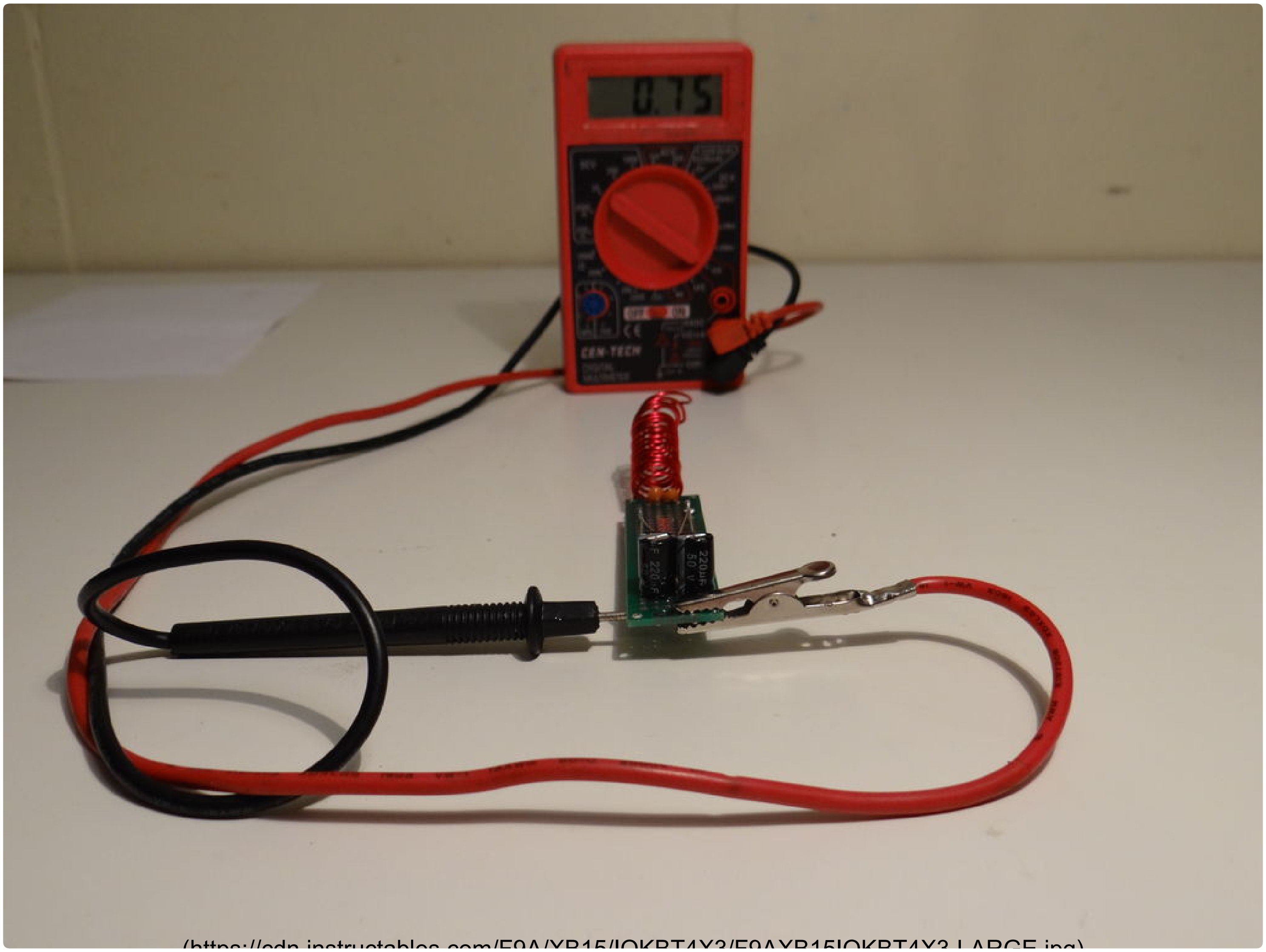
0.75
CEN-TECH
220uF
50V

(https://cdn.instructables.com/F9A/XB15/IQKBT4X3/F9AXB15IQKBT4X3.LARGE.jpg)

Download

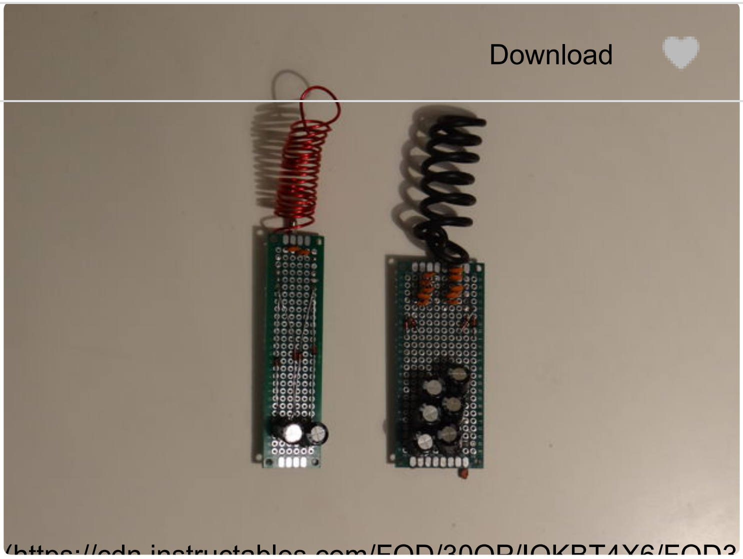
(https://cdn.instructables.com/FOD/30OP/IOKBT4X6/FOD3

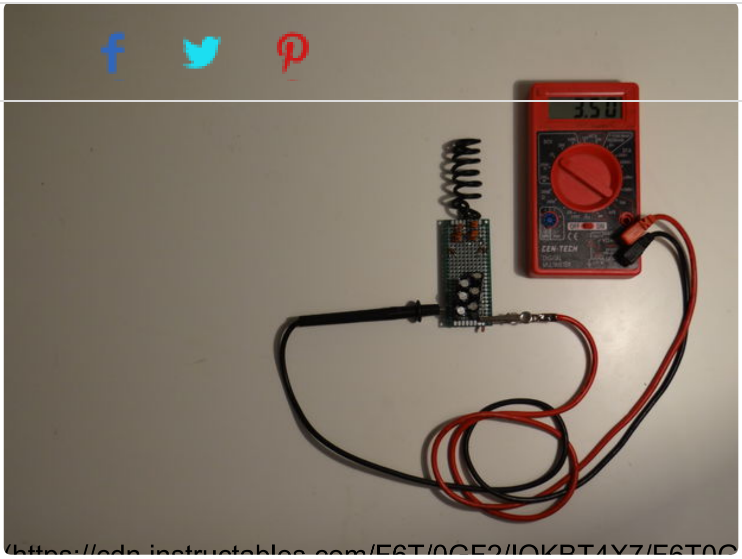

(https://cdn.instructables.com/F6T/0CE3/IOKBT4X7/F6T0C

To test and analyze our circuit, we'll be using a digital voltmeter and oscilloscope.

By connecting a voltmeter to the output, we'll immediately begin to see a small voltage climbing in the 10-100mV range. If not, we'll want to check our connections and make sure the circuit is not isolated from the environment by taking it outside to a clear area.

Then, by connecting an oscilloscope to the outside leads of our two ceramic capacitor bank, we will see the the polarized signal being captured from the air around us. We can then connect after the diodes to see our varying direct current and then to after the electrolytic capacitors to see a normalized, usable direct current at our output.

We can then optimize the input resistance in two ways. Firstly, we can add additional ceramic capacitors in parallel to our original two and make sure our soldered connections are consistent and thick in this area.

We can optimize the circuit's capacity by adding electrolytic capacitors in parallel to our original two which will allow this circuit to charge slightly when not in use. For this purpose, a charging circuit can also be added here in order to incorporate an optional battery bank.

We can optimize the antenna by attaching loops and coils of copper wire in various positions, store-bought antennas or by stringing some wire up to the highest point you can reach.

We don't have to stop there, either. We can now connect multiple circuits in series to increase voltage or in parallel to increase current. This can be done indefinitely.

Add Tip Ask Question

Step 4: Add a Case and Antenna

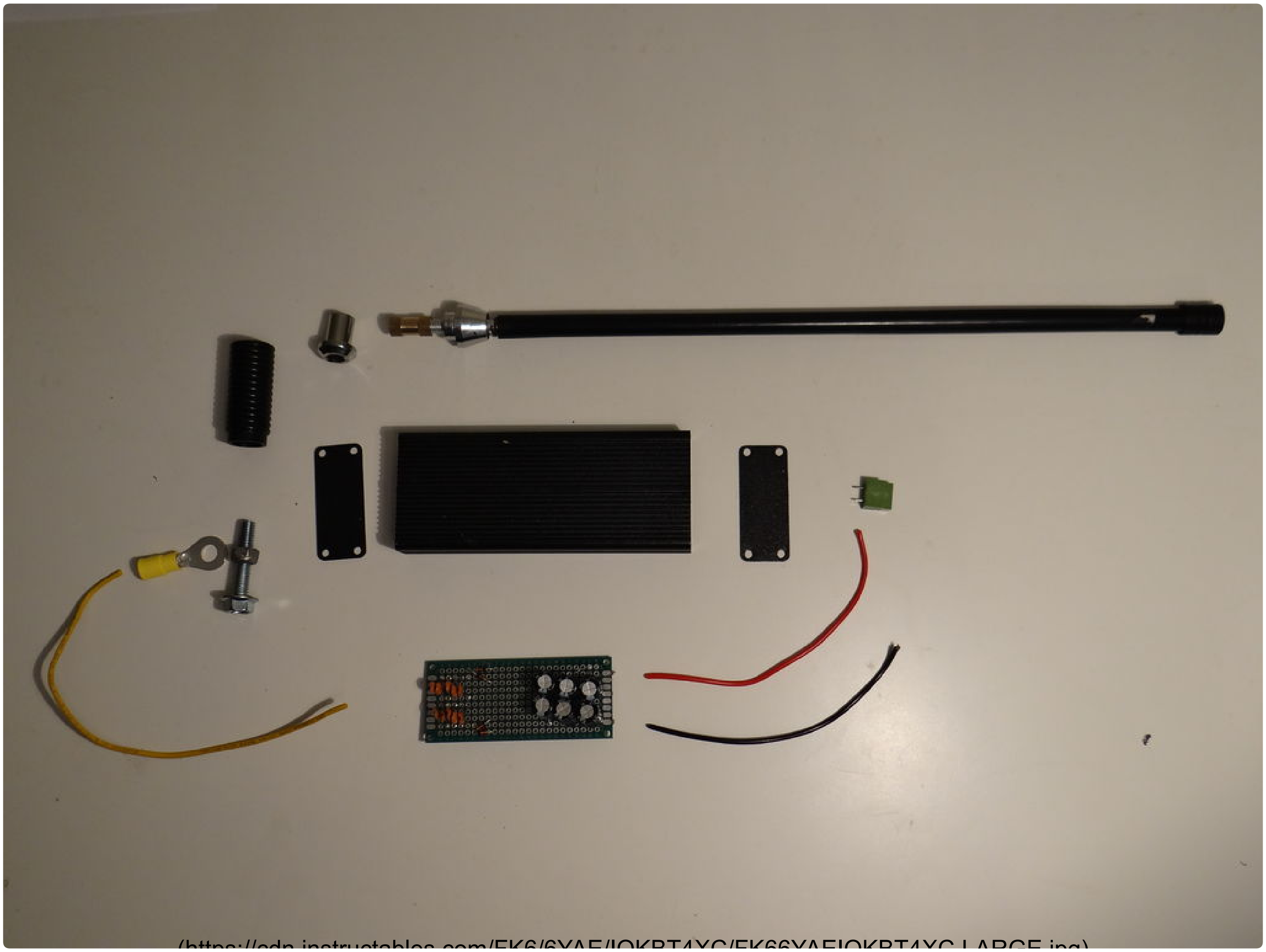

(https://cdn.instructables.com/FK6/6YAF/IQKBT4XC/FK66YAFIQKBT4XC.LARGE.jpg)

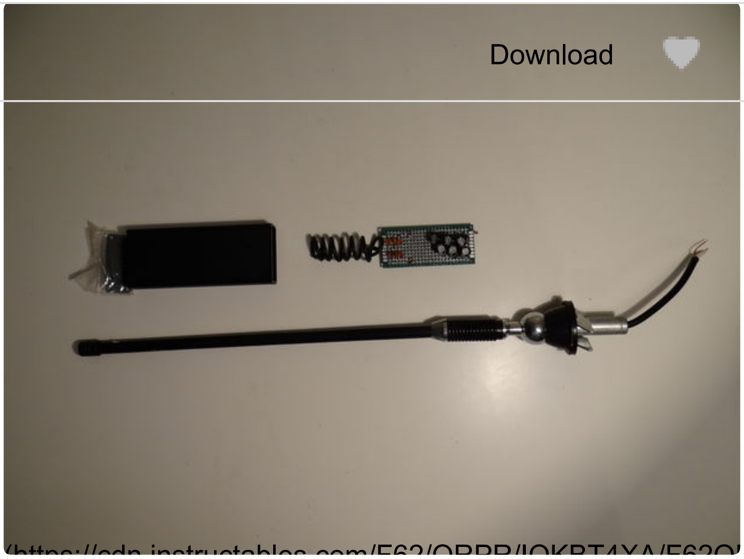

(https://cdn.instructables.com/F62/OPRR/IQKBT4XA/F62O

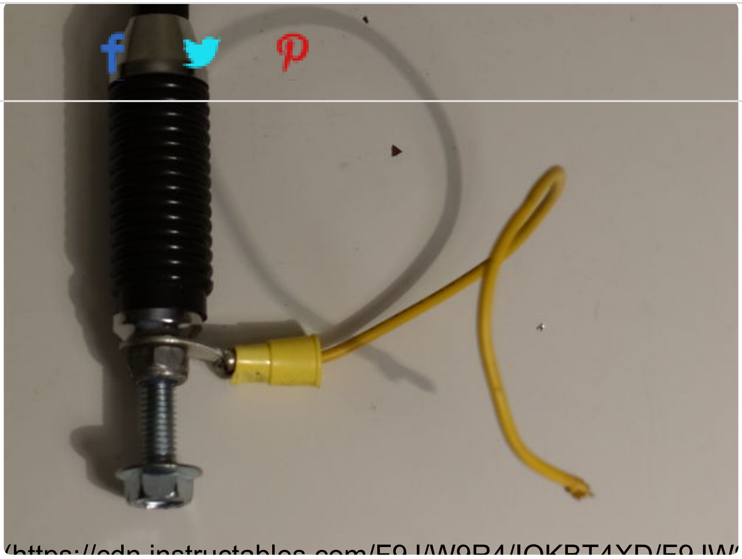

(https://cdn.instructables.com/F9J/W9R4/IQKBT4XD/F9JW

(https://cdn.instructables.com/F09/FS7

(https://cdn.instructables.com/F5O/5N9

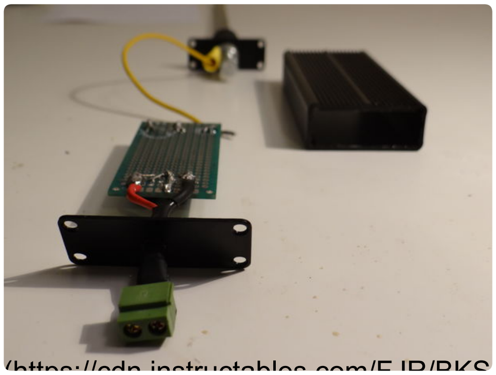

(https://cdn.instructables.com/FJR/BKS

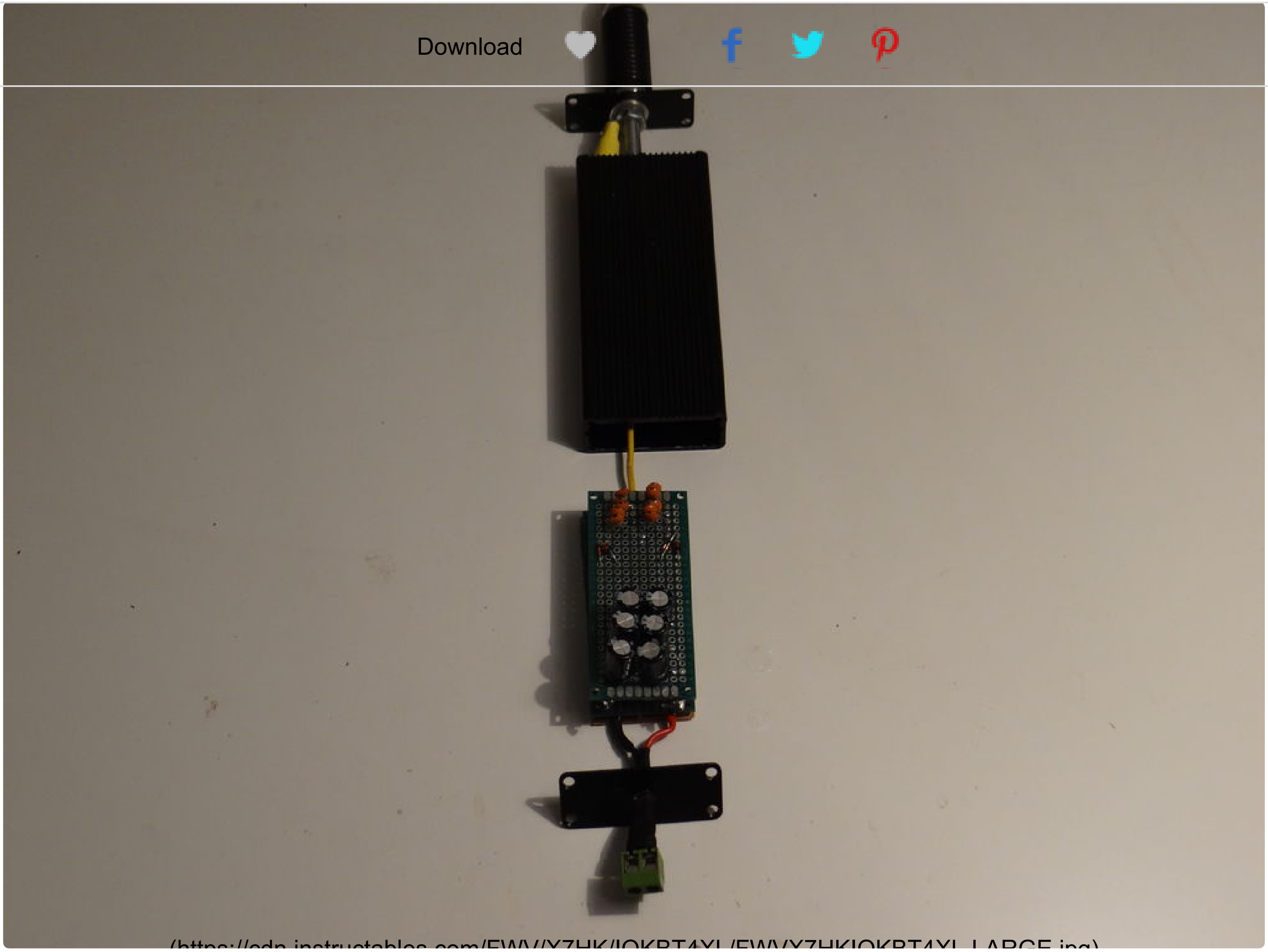

(https://cdn.instructables.com/FWV/YZHK/IOKBT4XL/FWVYZHKIOKBT4XL.LARGE.jpg)

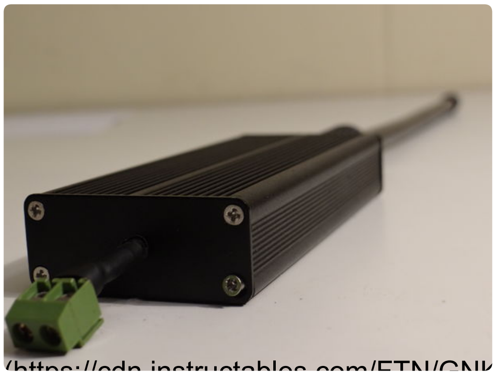

(https://cdn.instructables.com/FTN/GNK

(https://cdn.instructables.com/FM8/8IE/

(https://cdn.instructables.com/FD9/OG

After choosing an antenna in the last step we'll now want to permanently wire it. Whether you choose a compact antenna for portability or a tall fixed antenna for power and range, we will wire it in the same manner according to the diagram in the previous step. Note that the input on the configuration here is grounded to the metallic case, and thus the users hand, and incorporation of longer antennas will require proportionally more substantial grounding.

We will then attach a terminal to the output to allow us to connect this circuit to an electrical device or charging circuit and battery bank.

Next, we will add a case, making sure to isolate exposed leads with non-conductive material especially if mounting in a conductive case. A piece of cardboard secured with glue is sufficient for the circuit's bottom and shrink wrap or electrical tape can be used in the case of any additional exposed leads. Drill two holes in your enclosure, one for the antenna or antenna lead and another for your output terminals. You can then insert your components, fasten the enclosure and your device is ready to use!

Add Tip Ask Question

Step 5: Your Crystal Energy Receiver Is Complete!

Download

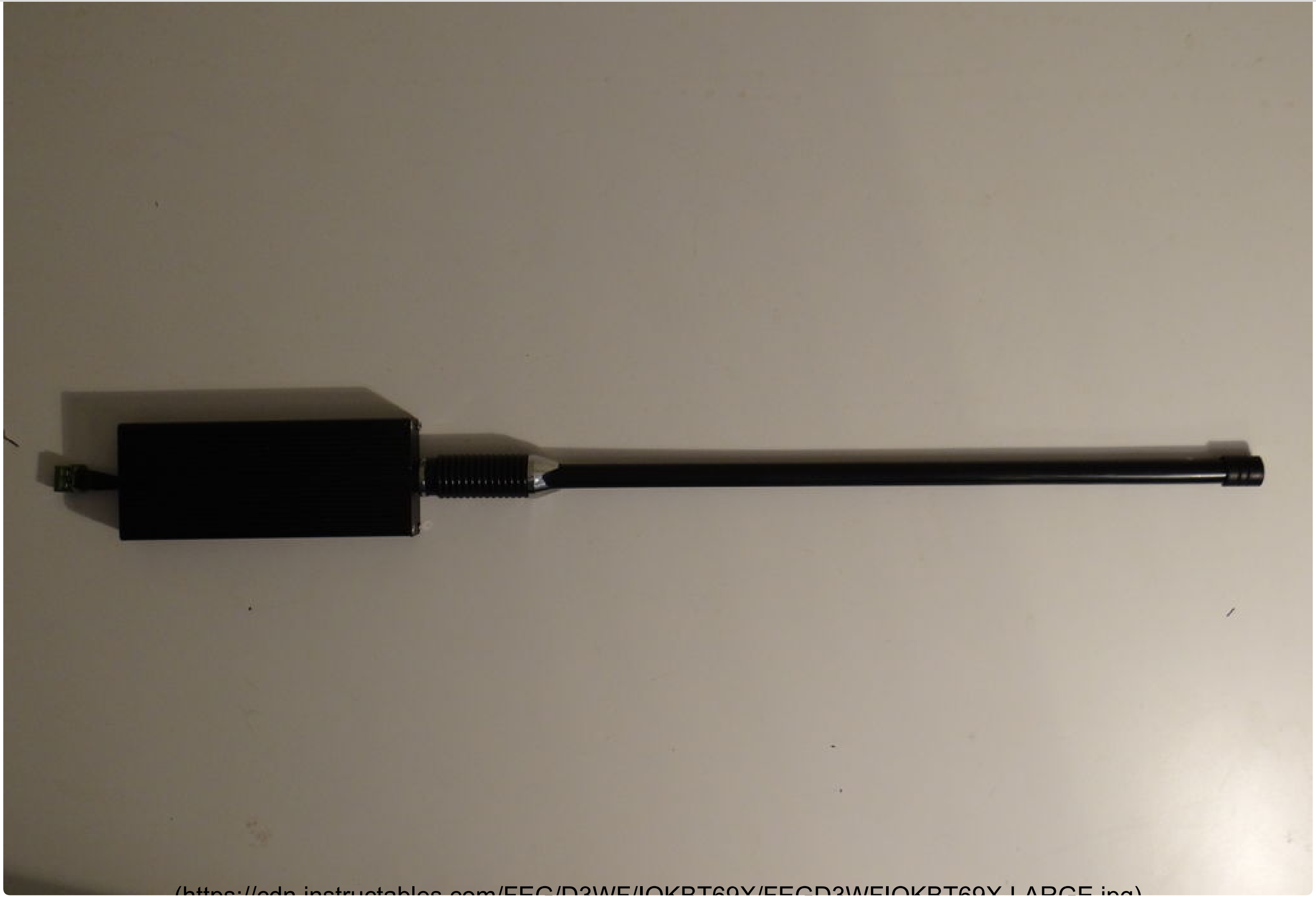

(https://cdn.instructables.com/FEC/D3WF/IOKBT69X/FECD3WFIOKBT69X.LARGE.jpg)

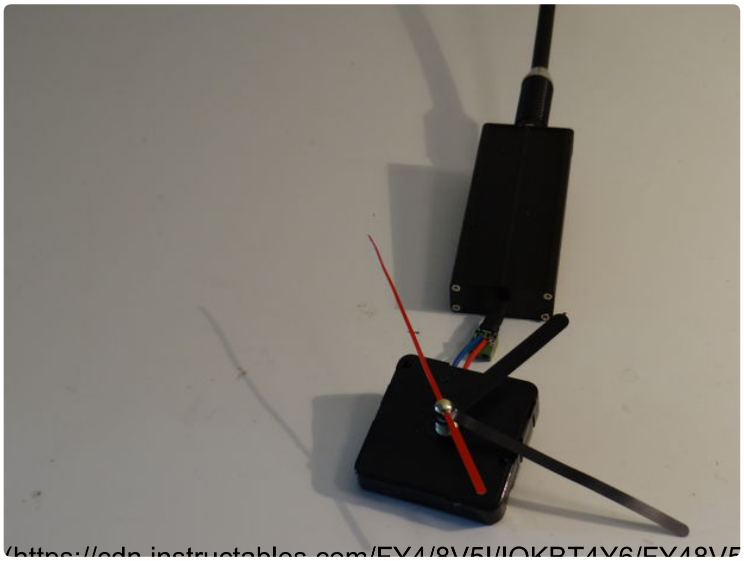

(https://cdn.instructables.com/FX4/8V5I/IOKBT4Y6/FX48V5

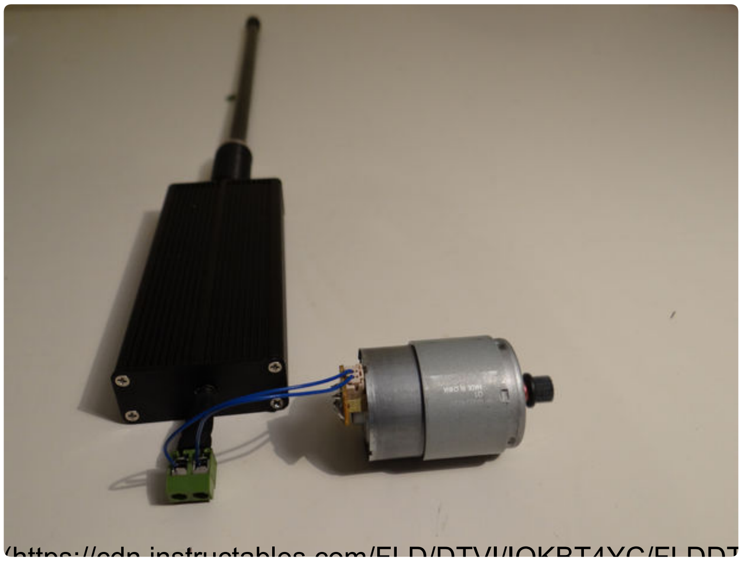

(https://cdn.instructables.com/FLD/DTVI/IOKBT4YC/FLDDT

Your Crystal Energy Receiver is now complete and ready to use!

I built a portable version, for proof of concept and demonstration purposes. However, you can go as big as you want- to passively charge batteries or run equipment remotely; or go as small as you want- to power sensors, RFID devices, small electronics and more.

I used this harvested energy to easily power a low-consumption quartz clock, a digital chronograph with integrated circuits and LCD and was even able to momentarily rotate a small dc motor.

Because of its simplicity this device is a durable, efficient and reasonably effective method of harvesting radiant energy in a simple, replicable and sustainable way. I humbly hope that the contributions made here, and by those reading, can be one day used by people worldwide to conveniently capture free energy.

Thanks for checking out my project and I look forward to seeing everyone's variations, suggestions and improvements!

Add Tip Ask Question

advertisement

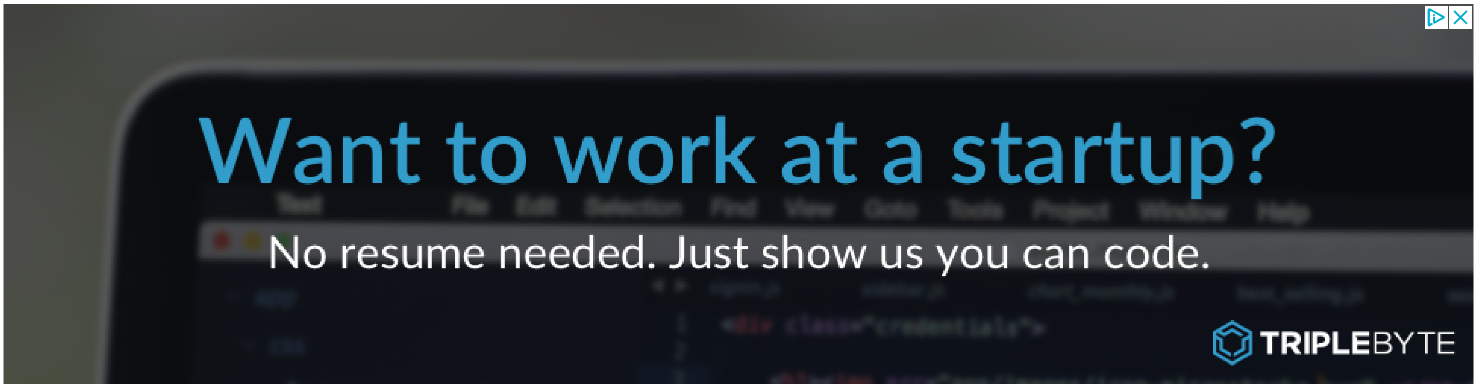

Want to work at a startup?
No resume needed. Just show us you can code.
TRIPLEBYTE

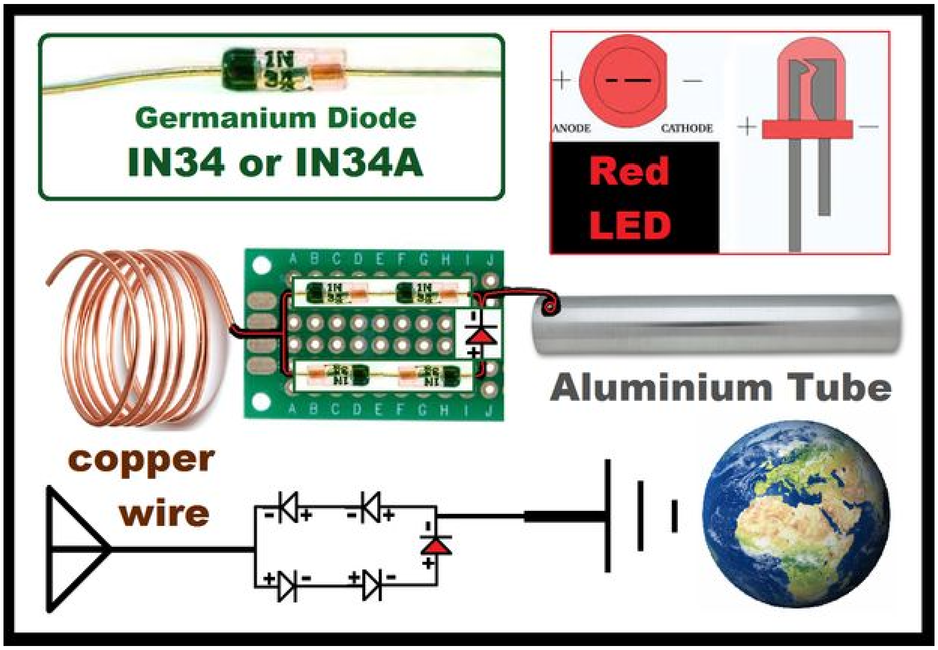

1N 34
Germanium Diode
IN34 or IN34A
ANODE
CATHODE
Red
LED
A B C D E F G H I J
Aluminium Tube
copper
wire

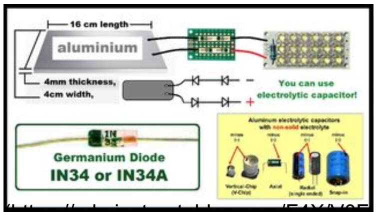

16 cm length
aluminium
4mm thickness,
4cm width,
You can use
electrolytic capacitor!
Germanium Diode
IN34 or IN34A
Aluminum electrolytic capacitors
with non-solid electrolyte
Vertical-Chip (V-Chip)
Axial
Radial (single ended)
Snap-in

(https://cdn.instructables.com/F4X1V9F

1N 34
Germanium Diode
IN34 or IN34A

(https://cdn.instructables.com/FUD/RBY8/JBBCZ3B/FUDB

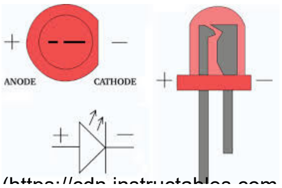

ANODE
CATHODE

(https://cdn.instructables.com

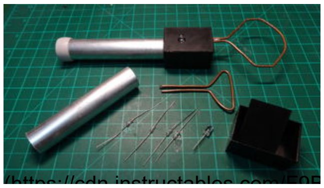

(https://cdn.instructables.com/F9F

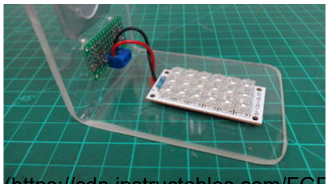

(https://cdn.instructables.com/FGF

advertisement

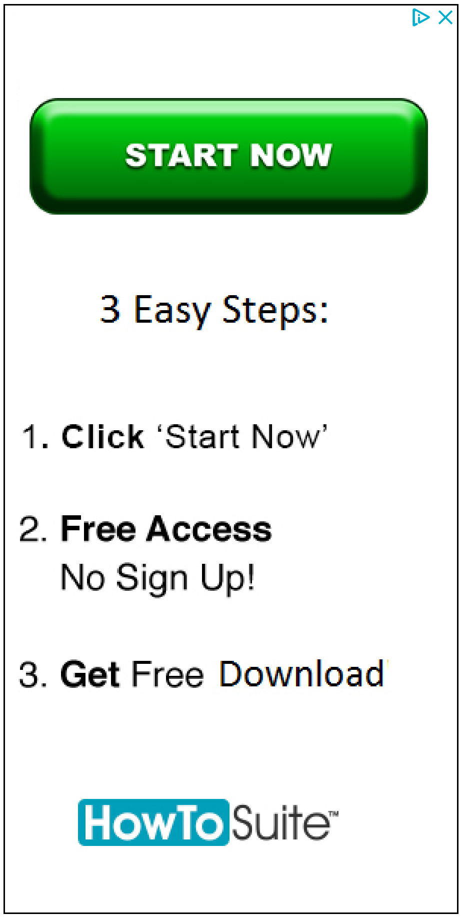

START NOW
3 Easy Steps:
1. Click 'Start Now'
2. Free Access
No Sign Up!
3. Get Free Download
HowToSuite

About This Instructable

2,351 views

46 favorites

License:

BY-NC-SA

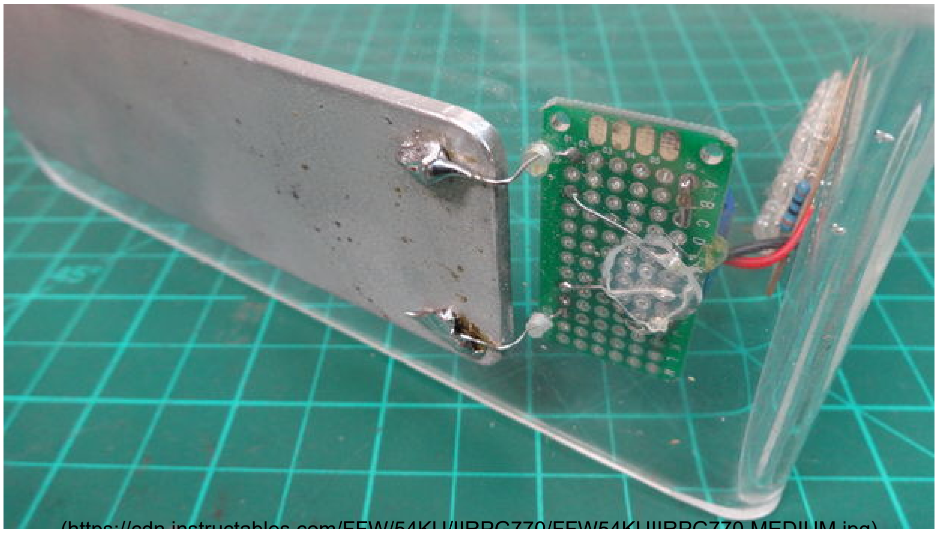

(https://cdn.instructables.com/FEW/54KU/IBRC770/FEW54KUIBRC770.MEDIUM.jpg)

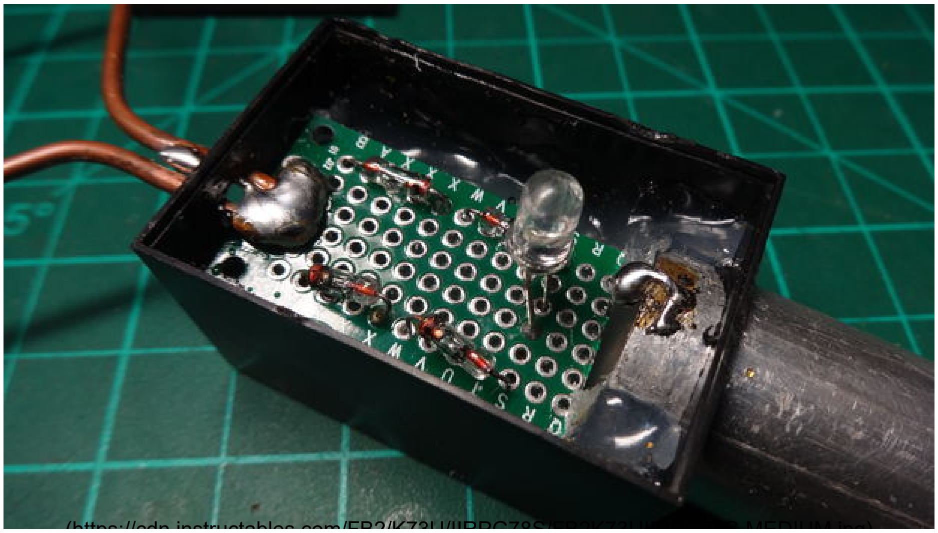

(https://cdn.instructables.com/FB2/K73U/IBRC78S/FB2K73U[illegible]MEDIUM.jpg)

How to Make a Electrostatic RF Detector or Ghost Detector, Hunting Solution. Much cheaper and easier than you may think For Under $5 ! A simple description is that IN34 or IN34A RF Germanium Diodes for Crystal Radio can take "Ambient" or "Radio Waves" to turns them into DC energy. I would like to show the Tesla Coil Transmitter, where I made an homemade "RF Detector" this is the simplest detector! without any batteries or hardships circuit, only with IN34 diodes and Ether to Earth!.. Stay Tuned and Enjoy! Please click the "like" button, subscribe my channel, and share this with your friends... To support my "work" every little bit helps! Watch the "ADS" THANKS

- See the original video https://www.youtube.com/watch?v=cSxqeGVC9t0

Step 1:

GEORGE CHANIOTAKIS (/member/GEORGE+CHA
(http://https://www.youtube.com/
Follow 84
(/member/GEORGE+CHANIOTAKIS/)

Bio: A big hello to all mankind! This is my channel I hope you can find here various interesting things on in Renewable Energy Technologies and ... More » (/member/GEORGE%20CHANIOTAKIS/)

More by GEORGE CHANIOTAKIS:

Synchro
Diy
Make

(/id/MAKE-PERMANENT-MAGNET-SYNCHRONOUS-MOTOR-GENERATOR/)

(/id/Make-Wind-Turbine/)

(/id/High-Voltage-Die/)

advertisement

You Should Never Shop On Amazon Without Using This Trick – Here's Why

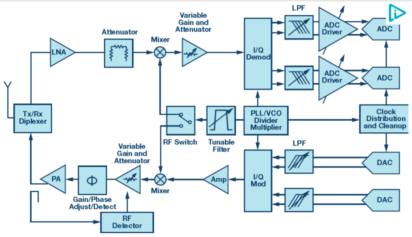
LNA
Attenuator
Mixer
Variable Gain and Attenuator
I/Q Demod
LPF
ADC Driver
ADC
ADC Driver
ADC
Tx/Rx Diplexer
RF Switch
Tunable Filter
PLL/VCO Divider Multiplier
Clock Distribution and Cleanup
Variable Gain and Attenuator
LPF
DAC
PA
Gain/Phase Adjust/Detect
Mixer
Amp
I/Q Mod
DAC
RF Detector

Get Complete RFIC Guide

ANALOG DEVICES

Download all Analog Devices RFIC Parts for Each Component in RF Signal Chain.

Comments

be nice

We have a be nice comment policy.
Please be positive and constructive.

I Made it! Add Images Post Comment

LazaroG4 (/member/LazaroG4/) 2017-11-14 Reply

I can't download and I was doing it before. I now need to go premium. This why i do not give any of my own files for the idea i have posted for free on You Tube. sorry for the not using of file and I have integrity and will not copy and paste as many do for free.

More Comments

GEORGE CHANIOTAKIS (/member/GEORGE+CHANIOTAKIS/) in leds (/technology/leds/)

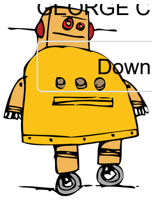

Download

(/id/How-to-Make-a-Electrostatic-RF-Detector-or-Ghost-D/) 1 Steps

+ Collection | I Made it! | Favorite | Share

Newsletter

projects, classes, and contests. Instructables will help you learn how to make anything!

enter email | I'm in!

About Us

Who We Are (/about/)

Advertise (/advertise/)

Contact (/about/contact.jsp)

Jobs (/community/Positions-available-at-Instructables/)

Help (/id/how-to-write-a-great-instructable/)

Find Us

- Facebook (http://www.facebook.com/instructables)
- Youtube (http://www.youtube.com/user/instructablestv)
- Twitter (http://www.twitter.com/instructables)
- Pinterest (http://www.pinterest.com/instructables)
- Google+ (https://plus.google.com/+instructables)

Resources

For Teachers (/teachers/)

Residency Program (/pier9residency)

Gift Premium Account (/account/give?sourcea=footer)

Forums (/community/?categoryGroup=all&category=all)

Answers (/tag/type-question/?sort=RECENT)

Sitemap (/sitemap/)

advertisement

AdChoices

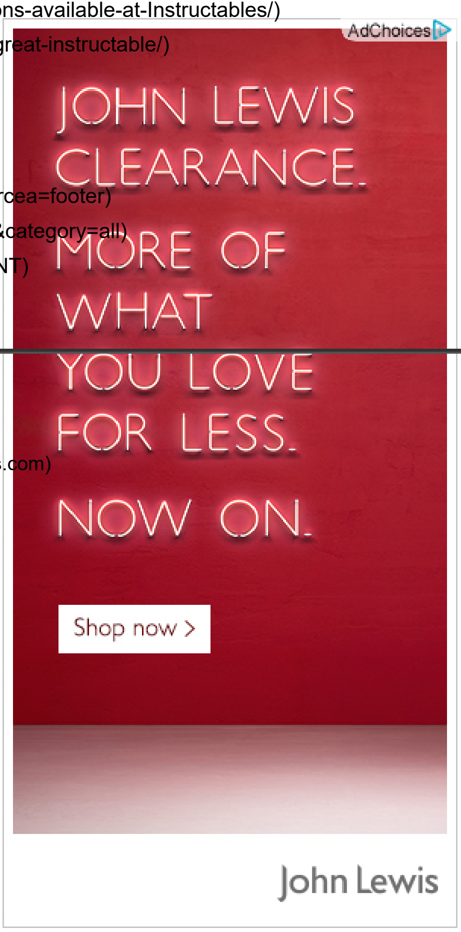

Terms of Service (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21959721) |
Privacy Statement (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21292079) |
Legal Notices & Trademarks (http://usa.autodesk.com/legal-notices-trademarks/) | Mobile Site (https://www.instructables.com)
(http://usa.autodesk.com/adsk/servlet/pc/index?id=20781545&siteID=123112)

© 2017 Autodesk, Inc.

Build a Homebrew Radio Telescope

Explore the basics of radio astronomy with this easy to construct telescope.

Mark Spencer, WA8SME

There are many ham radio related activities that provide a rich opportunity to explore and learn more about the science of radio. One of those opportunities is radio astronomy.

All matter emits radio frequency (RF) energy dependent on the temperature and makeup of the matter, including the matter in space. The foundation of radio astronomy is to study the heavens by collecting and analyzing the RF energy that is emitted by bodies in space, very much as optical astronomers use light energy collected by telescopes. It sounds complicated. While professionals use very sophisticated and expensive equipment, you can, with some simple equipment and a little investment, build a radio telescope that will allow you to learn and explore the fundamentals of radio astronomy.

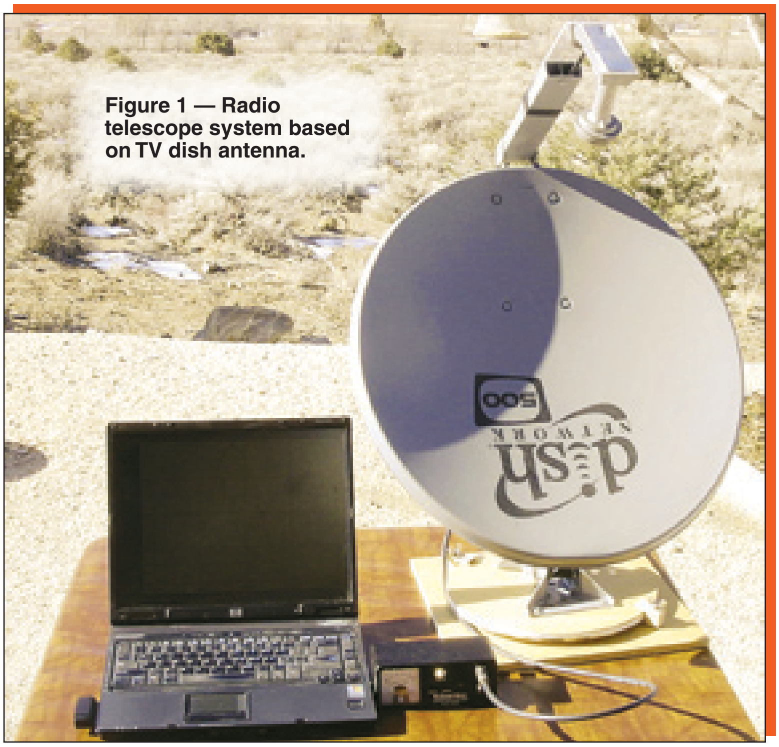
Figure 1 — Radio telescope system based on TV dish antenna.

A Homemade Radio Telescope

In this article, I will build on an existing design of a radio telescope made from one of those ubiquitous TV dish antennas that you see around your neighborhood. The radio telescope (RT) project described here can easily be reproduced. Although this is not a fully capable RT, it can provide a wonderful learning opportunity for you, or perhaps students in your local school.

Figure 1 shows the radio telescope set up. The major components include a modified TV dish antenna mounted on a wooden support structure to allow pointing the antenna, a commercial satellite signal strength detector that displays the signal strength of signals collected by the dish on a meter and an interface that converts the signal strength into a amplitude modulated tone. The tone is fed into a computer sound card and finally a computer and software graphically displays the signal strength as a function of time.

The TV dish modifications are structural, and any available TV dish system can be used. The signal strength detector costs between $40 and $65 and is widely available from Web retailers. The interface circuit, which will be described shortly, is easily duplicated and costs approximately $20. Finally, the display software is free.

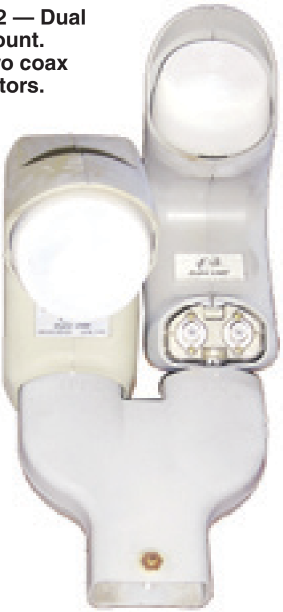
Figure 2 — Dual LNB mount. Note two coax connectors.

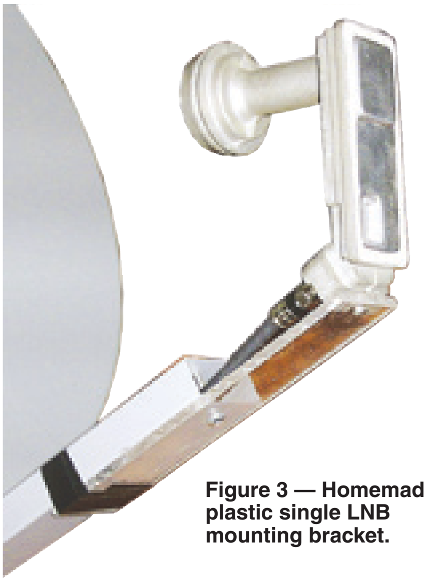
Figure 3 — Homemade plastic single LNB mounting bracket.

What it Can Do

The following is just a sample of what you can do with this simple RT:

- Use the sun to study and determine the beamwidth of the dish and verify the mathematic formula that is used to predict dish antenna performance.
- Measure the radiation intensity of the Sun and perhaps detect changes in solar activity.
- Measure the relative changes in the surface temperature of the moon.
- Learn about and explore a common radio astronomy collection technique called the *drift scan*.

From June 2009 QST © ARRL

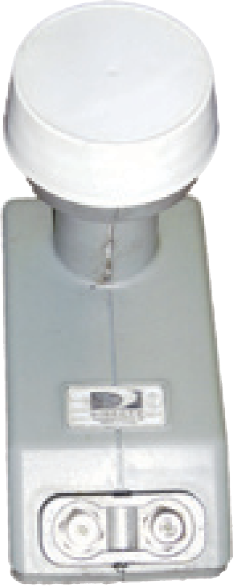

Figure 4 — Dual coax connector configured LNB. Terminate one connector with a dummy load.

- Explore the fundamental principle of energy emission as a function of temperature by detecting the relative differences between the temperatures of emitting bodies.
- Detect satellites parked along the Clarke Belt in geosynchronous orbit and illustrate how crowded space has become.

VOLTS
SIGNAL
Channel Master
Satellite Signal Level Meter

Figure 5 — CM satellite signal strength meter.

- Detect the Earth's rotation around the Sun and the Earth's spin on its axis by comparing daily drift scans of the horizon.

Antenna Subsystem

The basic RT system is based on the "Itty-Bitty" design that is described in two Web pages.[1,2] The TV dish is an offset 18 inch dish that has down converter(s) mounted at the focal point of the dish. The down converter is called a low noise block (LNB). The LNB is a preamplifier/down converter that converts the satellite signals from around 12 GHz down to around 2.4 GHz. Most modern dishes have two or more LNBs to access more than one TV satellite at a time without changing the pointing of the dish (Figure 2). The LNBs are mounted to share the focal point of the dish. Since only one LNB is required for the RT, I made a minor adjustment to the published Itty-Bitty design to position the single LNB at the dish focal point. Mounting the single LNB at the focal point really helps in pointing the antenna.

I used the existing LNB housing and mounting bracket as a template to determine the distance between the edge of the mounting arm to the mounting hole of the LNB. I then used a piece of plastic to fabricate a new mounting bracket for the LNB as shown

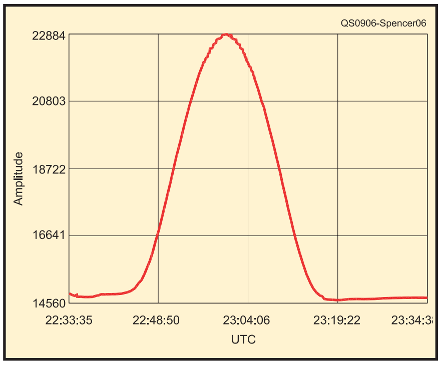

Figure 6 — *SkyPipe* screen showing antenna response.

---

[1] Notes appear on page 45.

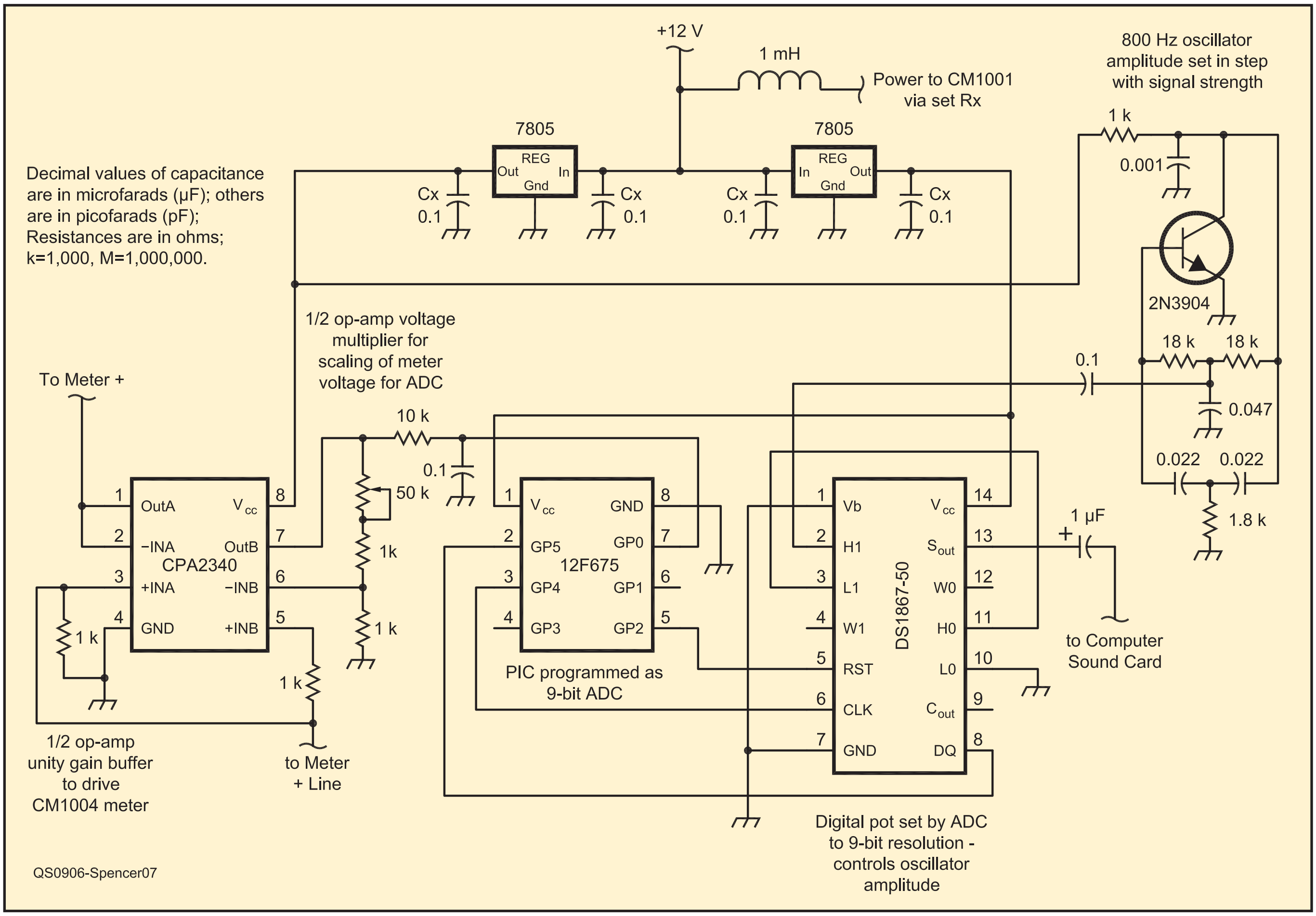
+12 V
1 mH
Power to CM1001 via set Rx
800 Hz oscillator amplitude set in step with signal strength
7805
7805
REG
Out
In
Gnd
REG
In
Out
Gnd
Cx 0.1
Cx 0.1
Cx 0.1
Cx 0.1
1 k
0.001
2N3904
18 k
18 k
0.1
0.047
0.022
0.022
1.8 k
1/2 op-amp voltage multiplier for scaling of meter voltage for ADC
To Meter +
10 k
0.1
50 k
1k
1 k
1 k
1 k
CPA2340
OutA
Vcc
−INA
OutB
+INA
−INB
GND
+INB
12F675
Vcc
GND
GP5
GP0
GP4
GP1
GP3
GP2
PIC programmed as 9-bit ADC
DS1867-50
Vb
Vcc
H1
Sout
L1
W0
W1
H0
RST
L0
CLK
Cout
GND
DQ
1 µF
to Computer Sound Card
1/2 op-amp unity gain buffer to drive CM1004 meter
to Meter + Line
Digital pot set by ADC to 9-bit resolution - controls oscillator amplitude
QS0906-Spencer07

Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.

Figure 7 — RT Interface circuit diagram.

From June 2009 QST © ARRL

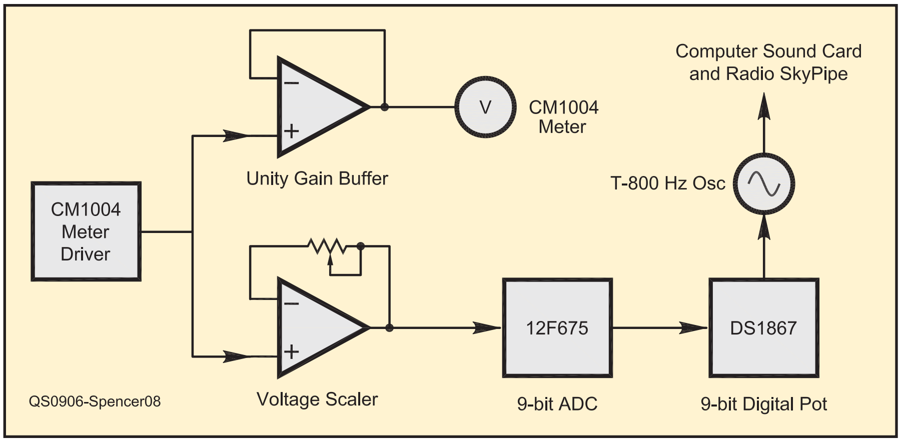

Figure 8 — RT Interface block diagram.

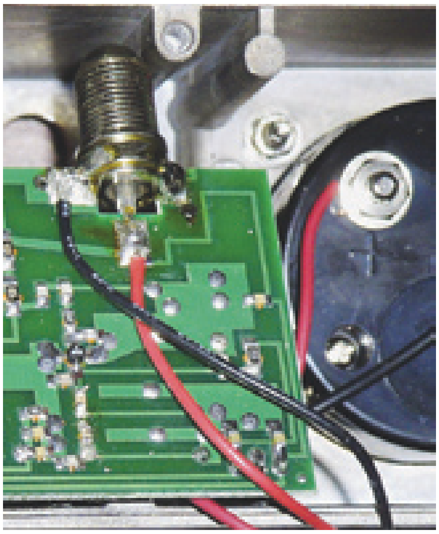

Figure 9 — Power and ground connection to CM board.

in Figure 3. The dimensions are not super critical, but careful placement certainly will improve the RT performance.

Some LNBs have two coax connectors. Only one will be used in the RT (Figure 4). It is a good idea to terminate the extra coax connector with a 75 Ω dummy load plug to balance the load on the LNB. The dummy loads for F type TV coax connectors are readily available from electronic parts retailers.

Note that the dish is mounted upside down. Though this orientation is not ideal for receiving satellite signals, this arrangement helps with pointing the dish in its radio telescope role.

Satellite Detector

The detector used in this project is the Channel Master (CM) satellite signal level meter model 1004IFD (Figure 5).[3] The CM is connected to the LNB. Power is supplied to the LNB through the coax connection from the CM. The CM detects the signal coming from the LNB and gives a meter indication of the signal strength and also varies the frequency of an audio tone to help technicians point the dish at the desired satellite. As you move the dish through the beam coming from the satellite, the meter indication will increase and then decrease coincident with the pitch of the audio tone.

The Itty-Bitty plans detail how to connect power to the CM and in turn connect power to the LNB (this power connection is handled by the interface in this project). Though somewhat effective, the CM meter and variable frequency tone indications provide limited utility in detecting changes in signal strengths required for radio astronomy.

Display

To really study the signals received by the RT, you will need to see them displayed graphically on a strip chart. There is an excellent software package called *Radio-SkyPipe* that is posted on radio astronomy Web sites.[4] The free version of this software is a good place to start. *SkyPipe* uses the computer sound card to measure the incoming signal strength and graphically displays the signal strength as a function of time. Figure 6 is illustrative of a signals detected by the RT. *SkyPipe* is very easy to use but some study of the HELP files will make it easier for you to fully tap into the capabilities of this software.

SkyPipe requires audio signals to be fed into the sound card MICROPHONE jack. The output of the CM detector is either an analog meter reading or a frequency modulated (constant amplitude) tone that is not really compatible with *SkyPipe*. An interface is required.

Interface

What is required to make the CM output work with *SkyPipe* and a sound card is to convert the signal level into an amplitude varying audio tone. The interface designed to do this is shown in Figure 7 and as a block diagram in Figure 8. Refer to the block diagram during the description of the interface function.

The unity-gain op-amp is used as a buffer between the CM meter driver circuit and the analog meter. The other op-amp is used as a voltage multiplier to scale the CM meter driver output voltage to match the 5 V reference voltage of the following analog to digital converter (ADC). The variable resistor in this voltage multiplier circuit is used to calibrate the CM to *SkyPipe*. The voltage from the multiplier is fed to a programmable interface controller (PIC) that is programmed as a 9-bit ADC to covert the analog voltage that is a function of received signal strength to a 9-bit digital word that is used to control a digitally controlled variable resistor. The interface includes a simple Twin-T audio oscillator circuit that provides a tone of approximately 800 Hz that is fed to the computer sound card. The amplitude of this audio oscillator is varied by the digital pot that is being controlled by the PIC. The result is the audio amplitude being varied in step with the signal strength detected by the CM.

The circuit provides power to the CM and the LNB. A 12 V source in the CM is tapped through an RF choke and this is connected to the LNB coax connector inside the CM (Figure 9). The 12 V is also regulated to 5 V to provide power to the interface. Though probably not required, there are two 5 V sources, one for the digital components of the interface, and the other for the analog components with one common ground point. This arrangement is used to isolate potential digital and analog noise sources within the circuit.

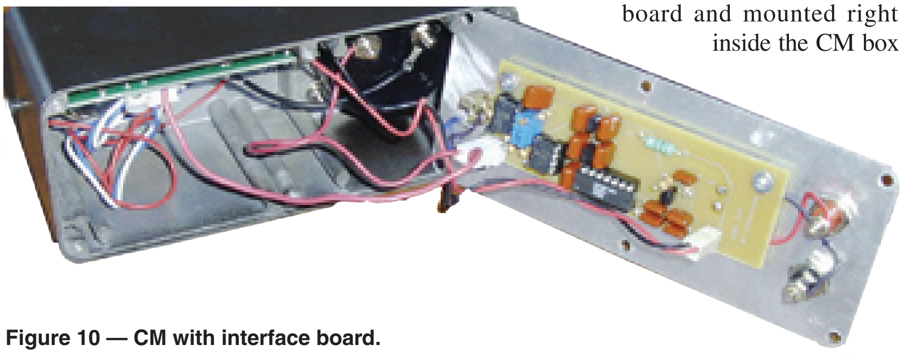

Figure 10 — CM with interface board.

The interface is built on a circuit board and mounted right inside the CM box

From June 2009 QST © ARRL

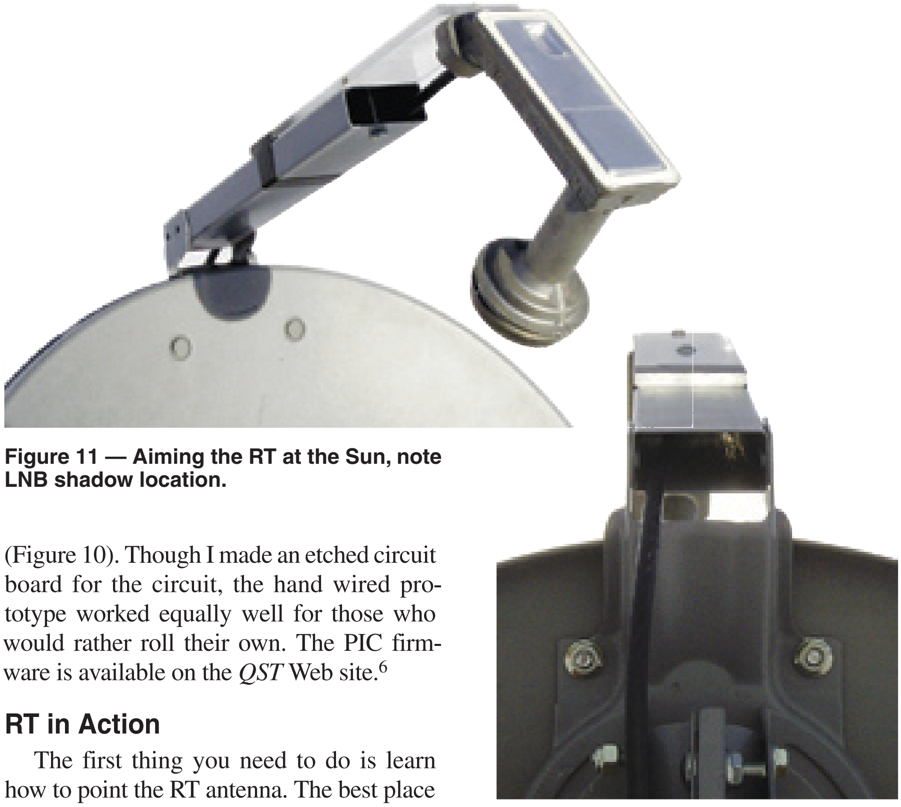

Figure 11 — Aiming the RT at the Sun, note LNB shadow location.

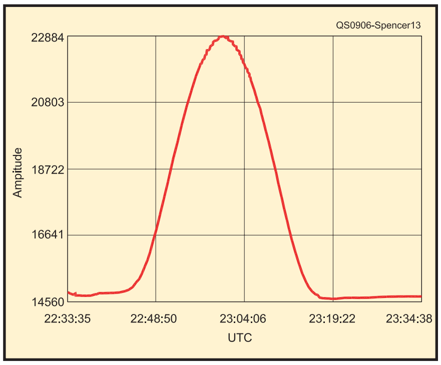

Figure 13 — Drift scan of the Sun indicating antenna's azimuth pattern.

(Figure 10). Though I made an etched circuit board for the circuit, the hand wired prototype worked equally well for those who would rather roll their own. The PIC firmware is available on the *QST* Web site.[6]

RT in Action

The first thing you need to do is learn how to point the RT antenna. The best place to start is to connect the CM to the antenna and point the antenna at the Sun. *Caution*: Do not look into the Sun as you do this, or at any time. Adjust the pointing angle and elevation until you get peak signal strength as indicated on the CM meter or hear the highest pitch audio tone. With the antenna pointed directly at the Sun, take note of the position of the shadow of the LNB on the surface of the dish (left in Figure 11). If you look from behind the dish, along the LNB

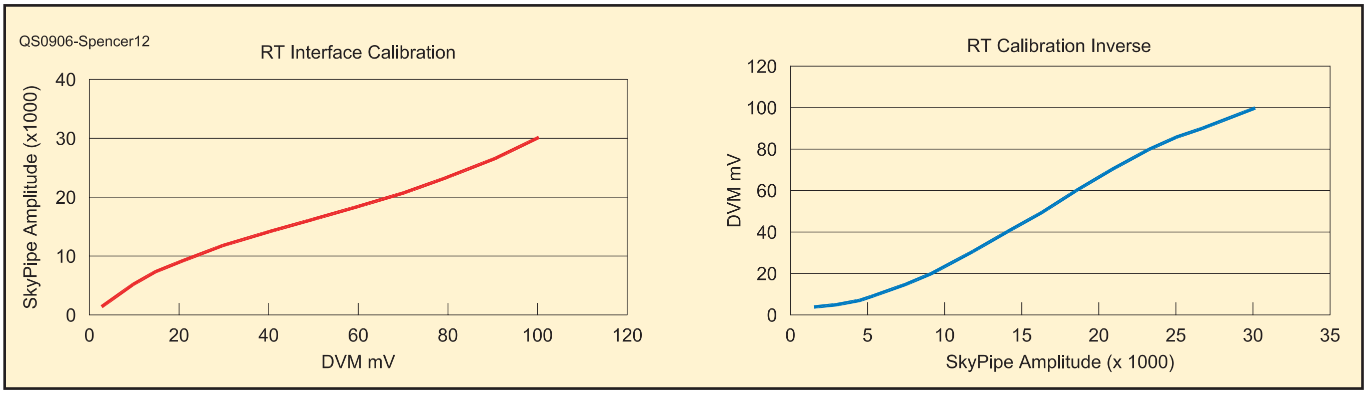

Figure 12 — Example calibration curves.

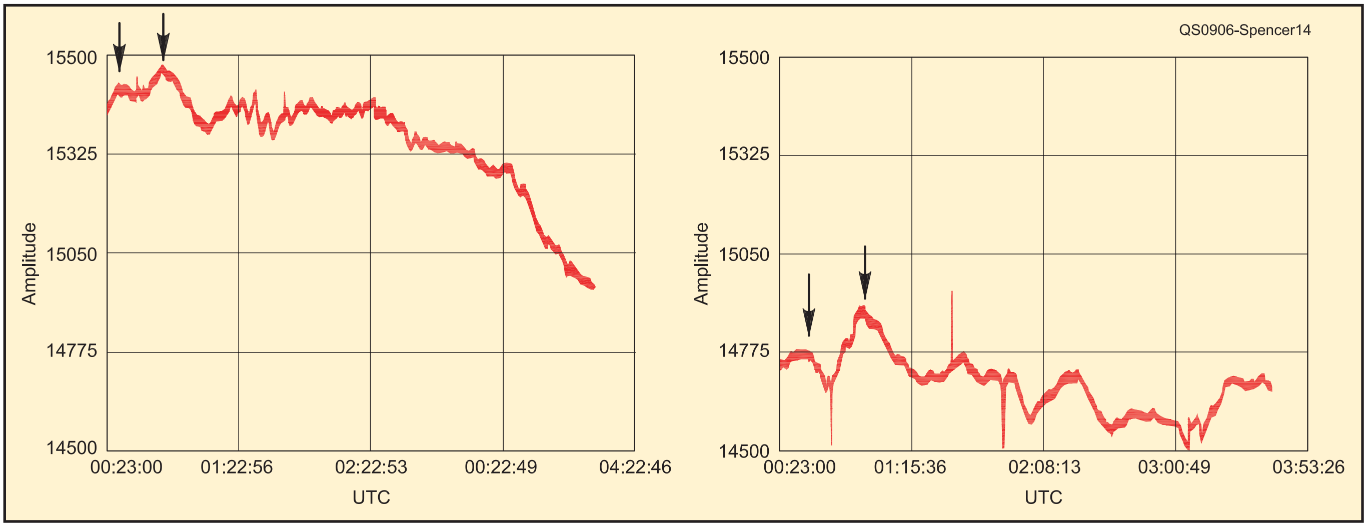

Figure 14 — Sequential drift scans. Note the time offsets between the peaks.

From June 2009 QST © ARRL

supporting arm (between the arm and the rim of the dish), you will see the Sun being blocked by the LNB.

Once you have the RT set up, it needs to be calibrated to match the output of the CM to *SkyPipe*. I have developed an *Excel* spreadsheet template to help with the calibration and a few of the other activities that you can accomplish with the RT (also available from the *QST* Web site). Turn the RT to a signal source, the Sun, or the side of a building would work. Turn the gain control of the CM to set the meter to maximum. Run *SkyPipe* and adjust the variable resistor on the interface board until you get a reading on the *SkyPipe* graph vertical (y) axis of approximately 32,000. With the maximum value set, adjust the CM gain control through the voltage range (0 to 100 mV) in 10 mV steps and record the corresponding y axis value on *SkyPipe*. This data is entered into the *Excel* spreadsheet to compute the calibration curve between voltage and y axis value. Both voltage and y axis values are used in analyzing recorded signal strength data (Figure 12).

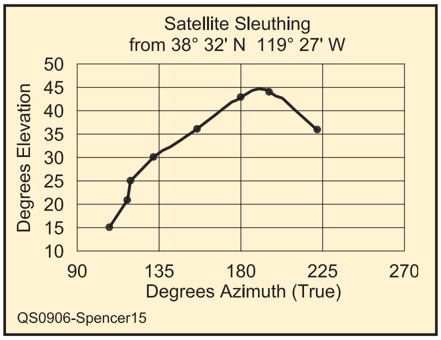

Figure 15 — Clarke Belt plot — tracking down satellites.

A good first activity is to do a drift scan of the Sun. A drift scan means that you set the antenna azimuth (AZ) and elevation (EL) to some fixed pointing angle and allow the Earth to serve as the rotator to drag the antenna across the sky. To do a drift scan of the Sun, first set the elevation and azimuth to point directly at the Sun (maximum signal) and then move the azimuth toward the west (leave the elevation set) until you are off the peak signal. Now start *SkyPipe*. In about 15 minutes, the Sun will pass through the antenna pattern beam width and the result will be as illustrated in Figure 13. You can also use this collection technique to explore the antenna performance parameters.

A good second activity is to do two drift scans of the night sky on two consecutive nights (beginning the scans at the same time each night) using the same fixed antenna azimuth (AZ) and elevation (EL). Figure 14 shows two such drift scans. Although at first glance they may not seem similar, there are some interesting features that are pointed to by arrows. If you compare the time that these two peaks occurred, the time difference is about 4.5 minutes. This shift is the result of the distance the Earth had traveled during the 24 hours between collections. This illustrates that the Earth's rotation as well as its travel in orbit needs to be considered when comparing drift scans. Enough to make your head spin (pun intended)?

A final good starting activity is to point the antenna toward the Clarke Belt and find all the satellites in geosynchronous orbit transmitting on 12 GHz. If you record signal strength peaks and AZ and EL for each peak, you will develop a graph of the Clarke belt as illustrated in Figure 15.

I have only scratched the surface, and the sky is the limit of this little project. The RT project can certainly broaden your horizons and expand your understanding of our universe. If you would like more detail than can be presented here, please contact the author.

Notes

[1] **www.setileague.org/articles/lbt.pdf**.

[2] **www.aoc.nrao.edu/epo/teachers/ittybitty/procedure.html**.

[3] **www.pctinternational.com/channelmaster/0612/satellite.html**.

[4] **radiosky.com/skypipeishere.html**.

[5] **en.wikipedia.org/wiki/Geostationary**.

[6] **www.arrl.org/files/qst-binaries/**.

Mark Spencer, WA8SME, is ARRL Education and Technology Program Coordinator. He is an ARRL member and holds an Amateur Extra class license. You can reach Mark at 774 Eastside Rd, Coleville, CA 96107 or at **wa8sme@arrl.org**.

Did you enjoy this article?
Cast your vote at:
VOTE
www.arrl.org/members-only/qstvote.html

From June 2009 QST © ARRL

Simple demonstration to explore the radio waves generated by a mobile phone.

Dr Jonathan Hare, Sussex University, Department of Physics, Falmer, Brighton. BN1 9QH

Note: this article has been published: Simple demonstration to explore the radio waves generated by a mobile phone
J P Hare, 2010, Journal of Physics Education, Institute of Physics, 45, p. 481 45 481

Also see the brief full article at: mobile phone detector

IMPORTANT NOTE: this device works very well on the old style mobile phones (as shown in the photo above). However, it does not always work well with modern smart phones. This may be because modern phones use higher frequencies, less power and use the power in a slightly different way (e.g. spread spectrum). Some smart phones do work and success may be due to the signal strength of the local mobile phone mast nearby. If you are in a low signal area the phone will create more power to ensure reliable communications. If you are in a very strong signal area (very near the local network) your phone will drop its output power and consiquently there will be less power to pick-up and to convert to a voltage to light the LED.

Described is a simple low cost home-made device that converts the radio wave energy from a mobile phone signal into electricity to light an LED. No battery or complex circuitry is required. The device can form the basis of a range of interesting experiments on the physics and technology of our mobile phones.

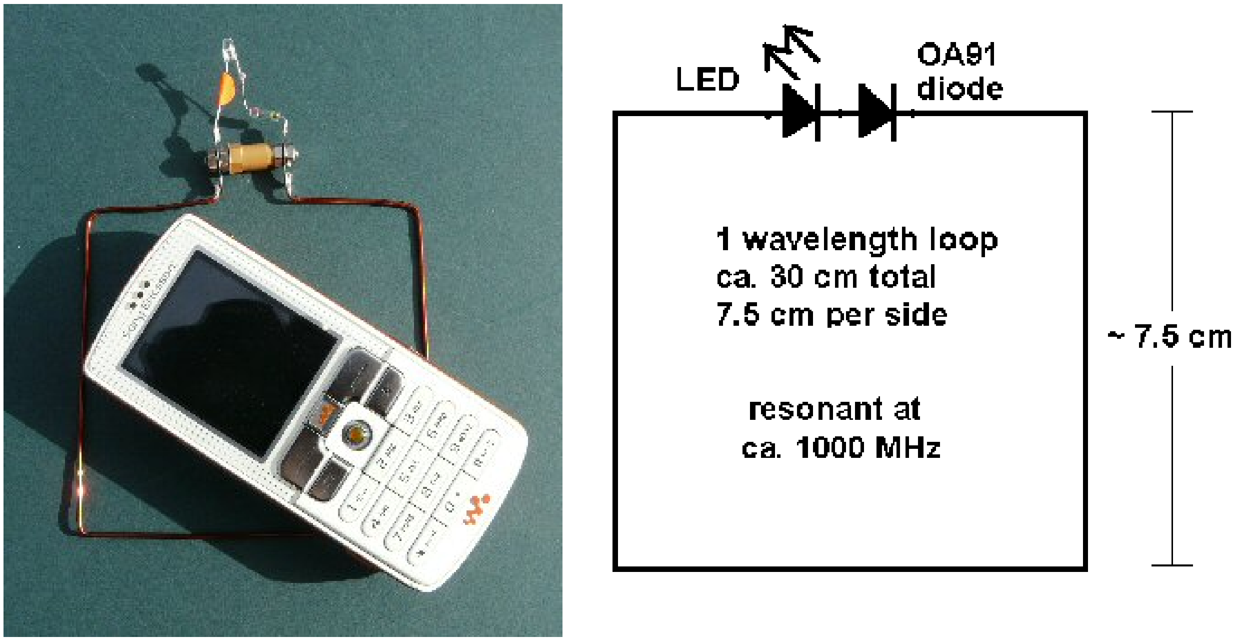

Fig. 1: left: mobile phone radio wave detector and right: the simple schematic

Introduction

Electromagnetic radiation (EMR) is at the heart of modern mobile phone data communications networks. The way a mobile phone and local base stations (the antenna covered masts you see dotted all around the place) communicate between each other is by using EMR in the radio wave part of the spectrum [1,2,3]. On switch-on your mobile sends digital information pulses by rapidly switching on and off the radio waves rather like a fast Morse code signal. Your text or voice is also converted into a series of digital pulses and sent across the network to be decoded (reassembled) by another mobile phone you dialled.

EM radiation and radio waves

Mobiles make use of various bands of radio frequencies to communicate between the mobile to base and the base to mobile: in Europe these include 900 and 1800 MHz (850 and 1900 MHz in the USA and Canada) [2, 3].

The relationship between wavelength, speed of light and the frequency follows the well known formula:
Wavelength λ (m) = speed / frequency = c (ms^{-1}) / ν (Hz)
λ (m) = 300,000,000 / ν (Hz) or approximately:
λ (m) = 300 / ν (MHz) ..... Equation 1.

So for a mid-range of about 1000 MHz (1 GHz) we get a typical mobile phone wavelength of about:
λ = 300/ 1000 = 0.3 m = 30 cm.

Simple radio wave detector

The loop consists of about a wavelength of wire, ca. 30 cm so each side is about 30/4 = 7.5 cm. The dimensions are not critical. The two ends are connected directly to a simple series circuit consisting of a high brightness LED and a germanium diode. They need to be connected correctly. All these components are cheap and readily available from electronic stores [4]. The loop can be made from a piece of copper wire roughly bent into a square (although a circular loop or rectangle will also work). If the wire is insulated remember to scrap off the insulation and solder-tin the ends. Simply solder the germanium diode and LED into circuit as shown in the diagram.
On a new LED the long lead is the positive (anode) while the short lead is the negative (cathode). The germanium diode has a line (band) around the end which is the cathode. When correctly wired the LED and the germanium diodes are connected so they both allow current to pass in the same direction, i.e. in the circuit diagram the arrows point in the same direction. In practice this means the LED and germanium diode are joined at the cathode of one and the anode of the other.
In my prototypes I used an insulator between the loop ends (light coloured cylinder in the photo) to make the whole thing more sturdy but this was purely for mechanical reasons and is not needed for the circuit to function properly. Note: a much more sensitive version using a x10 and x100 DC amplifier is described on my web site [6].

How it works and how to use it

When a radio wave passes across a metal object the EM fields cause the charged electrons in the metal to oscillate and this causes small AC currents at the same frequency to be induced into the metal. If a mobile is brought near to the loop and a call or text is made [5] the radio waves emitted from the phone pass across the loop. This induces a voltage into the antenna (the loop) and if it is close enough will be large enough to light the LED. As the loop is about one wavelength in size it is resonant and so there is a good transfer of power (low reactance) between the radio wave and LED.

The mobile phone automatically tests the network and adjusts its transmission power to maximise the battery life and minimise network interference. As a result the brightness of the LED will depend on the data being sent (the average signal), the local signal strength and how close the loop is to the phone. Why the second diode? - It's curious why the germanium diode is needed at all. The LED is a Light Emitting Diode after all and one would not think that another diode would help. However my initial experiments failed because I had not included it. The LED will have a relatively high capacitance which at these frequencies will tend to de-tune the loop and short out the LED. The germanium diode however is made up of a tiny wire which only makes a point-of-contact onto a piece of semiconducting germanium so it's 'self' capacitance is very low keeping the loop resonant.

The germanium diode will rectify the AC signal from the loop forming a series of DC pulses that will be nicely smoothed by the LED's capacitance. Without the diode however the raw AC signal from the loop will tend to be averaged to zero by the LED's capacitance.

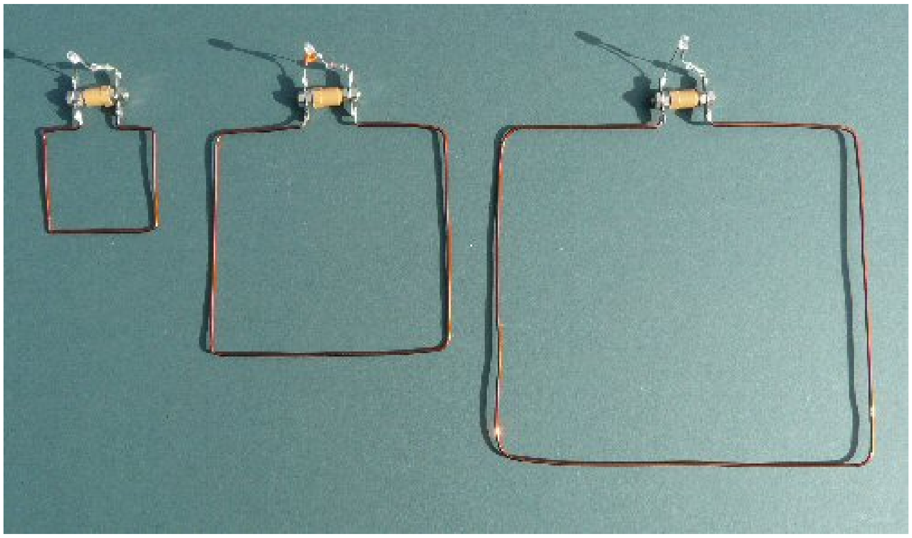

Fig. 2: Three loops in a row of varying size. The one described here and shown in Fig. 1 is shown in the middle. Smaller ones may well work better for higher frequencies such as the 3G networks (see below).

Other size loops
Fig. 2 shows a set of three loop devices with edge lengths of roughly 3.7, 7.5 and 15 cm. You can find out for yourself that the best match to the mobile signal is with the full wave loop of ca. 7.5cm per side. The other loops do work to varying degrees however (smaller ones may work better for the 3G network). The larger loop works well for the '70 cm' amateur radio bands.

Polarisation
The electric and magnetic fields making up the EM wave are orthogonal (they are at right angles to each other as they pass through space) to each other but depending how they are generated by the transmitting antenna can arrange themselves in any orientation with respect to the ground. If the electric field is parallel with the ground we say the wave is 'horizontally polarised' while if its normal to the ground we say its 'vertically polarised'. The loop antenna will respond best to one type of polarisation (depending on its orientation) so it's worth experimenting with the orientation of the mobile (or the loop) to get the strongest signal - brightest LED.

Mobile antenna
Inside your mobile phone is a transmitter / receiver and antenna. Many mobiles have this antenna at the top of the phone but some of the PDA type phones have it at the bottom. As a result you can locate the position of the antenna by moving it around the center of the loop till you get maximum LED brightness.

Networks
There are various different networks that a mobile may use both in the UK and abroad. It may be that you need to adjust the network phone settings on your mobile i.e. change from "automatic select" to set for "GSM" so as it get the strongest signal to light the LED. Note: the 3G network might not be powerful enough to light the LED. As the GMS network is currently the main network over the UK the device should work anywhere where you can get a signal as long as you check the correct selection on your mobile menus [5, 7]. The 3G network operates on a higher frequency (smaller wavelength) so you might find a smaller loop will work better than the main one described here. See 'other experiments' section below.

Test signals
In order to pick up the radio wave energy from the phone it obviously needs to be transmitting a signal. There are a few ways to do this:

1) On switch 'on' (or change of network) you can see that the mobile initially transmits for a few seconds to the network to tell it it's there (especially if you have moved since turning it off). You don't actually need to dial a number to detect these signals.
2) Even if don't text or call, throughout the day the mobile will send out data to 'keep up' with the network, especially if you are moving around (going through train tunnels etc. see below).

3) When you make a phone call you will transmit. Initially there is quite a lot of data being sent but in a few seconds data / power only gets transmitted when you speak. So to light the LED continuously you need to talk or provide some background sound continuously. Your service provider voicemail might be a good free phone number to try for these experiments [5].

4) Texting is the easiest way to show the radio wave power being transmitted. Long texts will light the LED for longer than short texts.

5) Finally set up the mobile on the loop and use another phone to text or phone the mobile. Even though you are not directly using the phone you will see that even on 'receive' the mobile phone transmits data to and fro. Ring off before you get charged.

Note: If you can use a free phone number it will save you money [5].

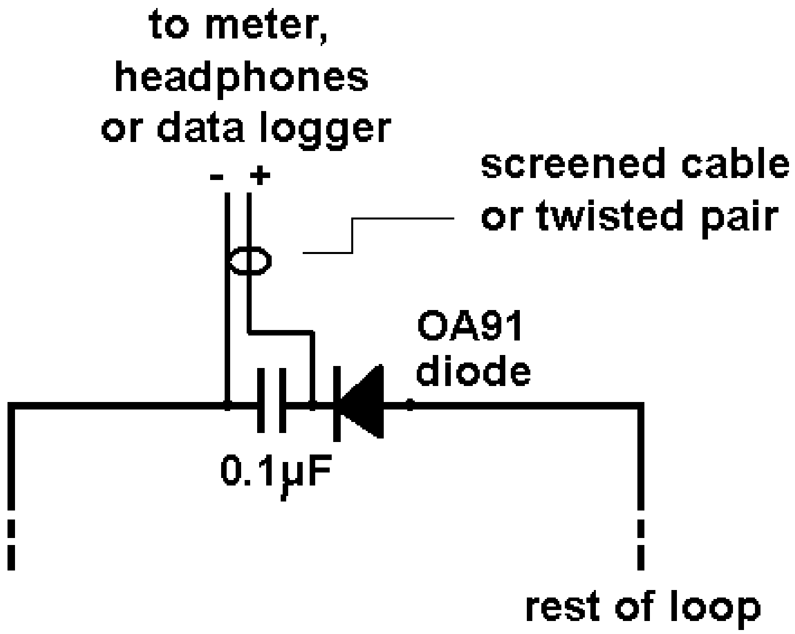

Fig. 3: Adding a capacitor and coax (or twin) lead so that headphones, a meter or a data logger can be connected (Note: diode is reverse wired compared to Fig. 1).

Other experiments:
Hearing data - if headphones are wired across the LED they will convert the voltages into sound and you can 'hear' the clicks of the digital data being transmitted. These are the same clicks that so easily get picked up by sensitive electronics such as a stereo amp or recording equipment when making a video for example. Hence - 'no phones on' when filming.

Logging data - if a meter, or better still a stand-alone data logger, is attached across the LED then one can monitor the EMR from the phone. For example even if you are not making a call your mobile will send signals too (and receive signals from) the network while travelling around. Fig. 3 shows a simple modification using a de-coupling capacitor so that a coax cable (or twisted pair) can be used to go to headphones, meter or data logger. Note the diode has been reversed so that the logger has the correct + and – connections for a unipolar input logger. The capacitor should help average the signal and stop radio frequencies going down to the logger. If one is available a few turns of the wire can be wound within a ferrite ring near to the logger so that maximum immunity to the mobile phone signal can be obtained for the logger electronics.

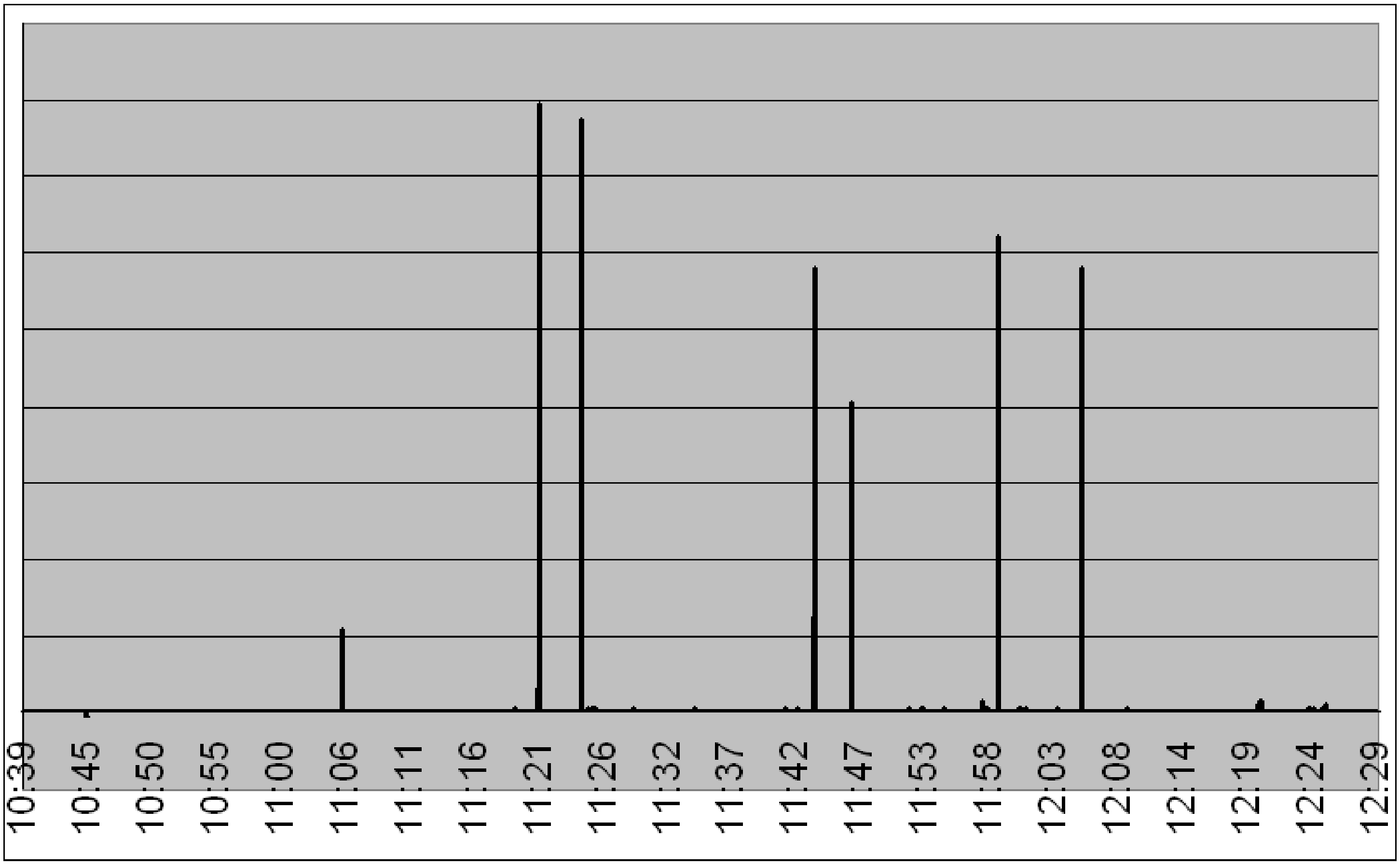

Fig. 4: Typical mobile phone data signals sent out onto the network while travelling around. These were recorded by a data logger from a mobile using the loop (no calls or text were made) while travelling on the train from Brighton to London Victoria (and then around London and return). Many of these peaks were the phone sending out 'I am here' data after coming out of one of the many long tunnels under the South Downs during the journey.

Out and about - Once you can log data you can discover all sorts of interesting things your mobile phone is doing without you realising it. Fig. 4 shows the plot over a few hours of travelling between Brighton and London (and within London) on the train. The detector was simply placed near to a phone that was not making or receiving a phone call or text, but was turned on.

The graph shows that the mobile sends out signals to tell the network where it is as it travels along and in particular goes in and out of long train tunnels. The peak heights vary because of the different powers the mobile transmits at depending on the signal strength of the local network and also because of the way the data logger 'snatches' a reading from the circuit every few seconds. As your phone sends out data onto the network to ensure the very best communications as you move around, so your mobile and the network obviously knows where you are and where you have been. Thieves and criminals beware the police can track you!

The inverse square law - If the transmitting mobile phone is moved away from the loop one would expect the signal to drop off. Unfortunately because both diodes need a certain threshold before they conduct the detector is not sensitive to small signals and not very linear. Therefore it's not very easy to use the device to measure the inverse square law (drop in signal v distance away) but of course you can see the signal go down. You could perhaps use the device to plot isobars - i.e. plot the equal intensity signals around the phone / nearby objects.

Changing the resonant frequency of the loop - you might be able to make some simple sliding mechanism (e.g. a small trombone-like mechanism) out of metal tube for example to tune the loop device for different frequencies. Then you can use it to find the average wavelength and so determine the center frequency by adjusting the size for maximum brightness of the LED. The wavelength can be determined by measuring the total distance around the loop. If we assume the antenna is one wavelength in total length then the frequency can be established by rearranging Equation 1, i.e. ν (MHz) = 30,000 / L (cm), where L is the length around the loop (cm).

Note: You will need to allow the transmitted digital signal to 'settle down' i.e. make measurements only after a few seconds after dialling / pick up so that only the sound data is being transmitted rather than the initial connection data. A constant sound will also need to be made so that the mobile phone continuously transmits data. It's worth playing music near to the phone or constantly whistling to keep sound coming into the phones microphone.

Mobile phone detector - teachers who want to know if the students / pupils really have turned-off their mobile phones (rather than just put on 'silent') can wire the loop device into the class room white-board speakers. Any mobile that is on in the class will send out signals which (if you are close enough) you will hear the data going to and fro - you will have your very own 'who's got their mobiles on' device which might be useful for exams etc.

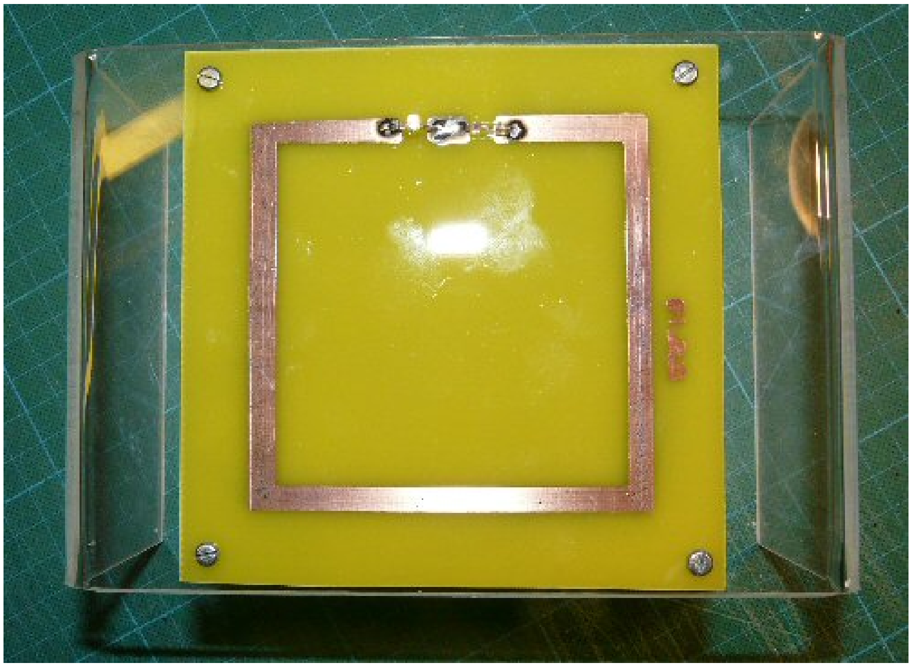

Fig. 5: The SEPNet 'deluxe' printed circuit board version (pcb) on a perspex stand where the loop is composed of a pcb copper track and the diode and LED soldered onto the board (top) [8,9].

Summary
All in all then, for such a simple easy to make device I hope you agree that there is a lot of scope for interesting science / technology investigations with your mobile phone. The device would make a good science week project (for radio amateur clubs etc.) A 'deluxe' pcb version (Fig. 5) on a perspex display case (Fig. 5) is currently going around the southern UK as part of the SEPnet outreach work, see the 'Radiation Exhibition' [8] and also as part of my on-going lecture series [9].

Post publication additions
(What follows was not included in the published article as this calculation was worked out later). A full wave loop is resonant and so looks purely resistive to the radio waves. Such a loop will have a resistance of about 100 ohms (Note: this is the AC resistance and not the DC resistance which will be very low). Now power P = V x I (V = voltage and I = current) and resistance = R = V/ I therefore $P = V^2 / R$ or rearranging V = √(P x R) which means that the voltage created by a power level of say 50mW (say for argument that roughly half the mobile phone power) arriving at the antenna will be about V = √(100 x .05) which is aprox. V = 2V, enough to light an LED.

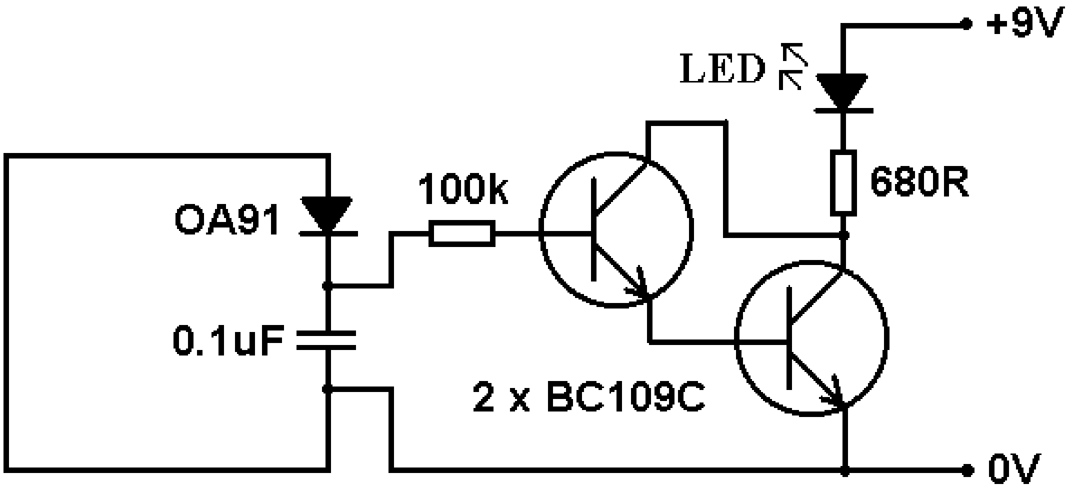

This circuit uses a two transistor darlington driver to amplify the signal from the loop and diode making the detector much more sensitive. The LED will be much brighter using this circuit. Note: the circuit needs a battery to power it (e.g. a PP3 9V)

References
[1] These ultra high frequencies (UHF, > 1000 MHz) are also often called microwaves.
[2] wiki pages
[3] Elektor Electronics magazine, June 2005
[4] order codes for the germanium diode and LED are:
e.g. Germanium diode: Maplin Electronics: QH71N, Rapid Electronics: 47-3114
e.g. LED: Maplin Electronics: UF72P , Rapid Electronics: 55-0085
[5] to save money use your voice mail service (often you simply dial 121).
[6] for details of an amplified detector see: wavemeter
[7] select 'network setting' from the mobile phone 'settings' menu and then go to 'network mode' and select 'GSM 900/1800' rather than 'automatic'.
[8] SEPnet mobile phone device on display throughout southern UK. [9] for details of my talks see: talks and workshops

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, Brighton, East Sussex. BN1 9QJ.

home | diary | whats on | CSC summary | latest news

Simple demonstration to show mobile phones emit radio waves

Dr Jonathan Hare, Sussex University, Department of Physics, Falmer, Brighton. BN1 9QH

Note: this article is in press: Elektor Magazine, July-August 2010, p. 56-57

For other experiments with this device please see my full article at: mobile phone detector

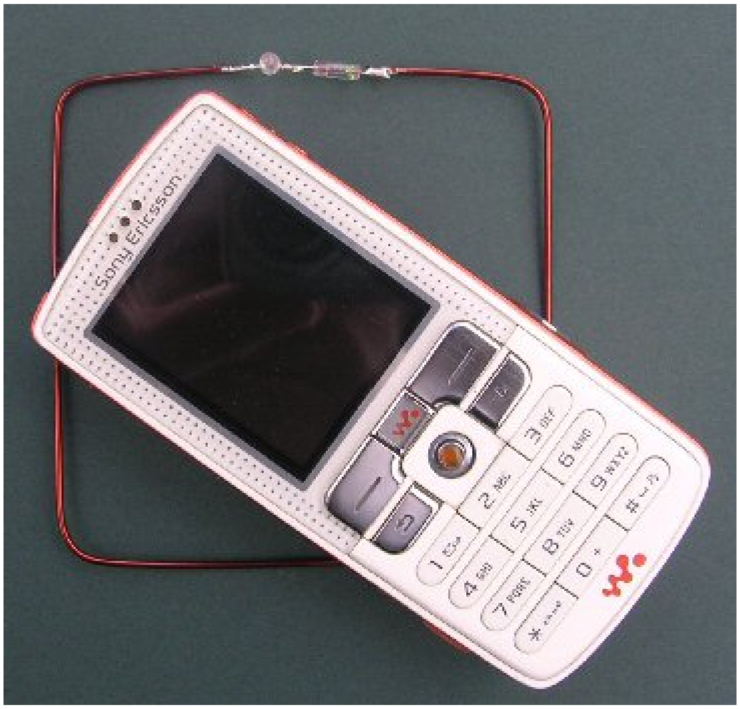

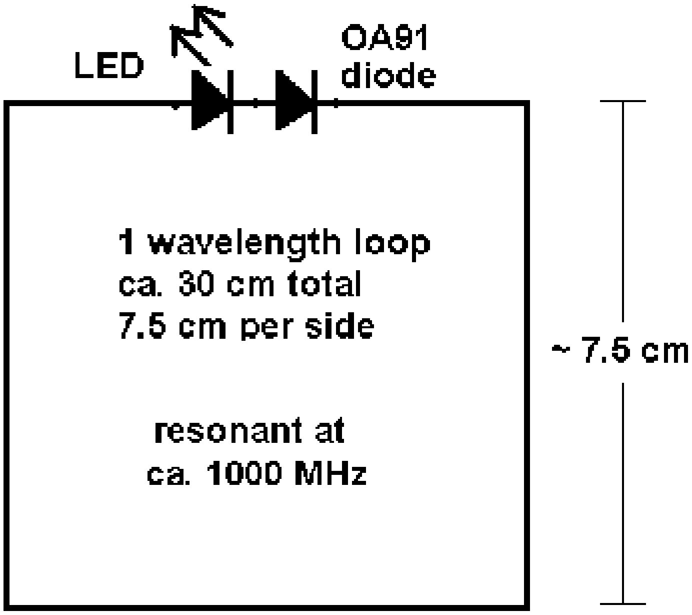
LED
OA91
diode
1 wavelength loop
ca. 30 cm total
7.5 cm per side
resonant at
ca. 1000 MHz
~ 7.5 cm

left: mobile phone radio wave detector and right: the simple schematic. Below: detail of the LED and germanium diode.

IMPORTANT NOTE: this device works very well on the old style mobile phones (as shown in the photo above). However, it does not always work well with modern smart phones. This may be because modern phones use higher frequencies, less power and use the power in a slightly different way (e.g. spread

spectrum). Some smart phones do work and success may be due to the signal strength of the local mobile phone mast nearby. If you are in a low signal area the phone will create more power to ensure reliable communications. If you are in a very strong signal area (very near the local network) your phone will drop its output power and consiquently there will be less power to pick-up and to convert to a voltage to light the LED.

This is a very simple and cheap device that demonstrates mobile phones ('cell phones' or 'handies') generate radio waves. We have a 30 cm (7.5 cm per side) full-wavelength loop antenna (a 'Quad' to radio amateurs) connected to a germanium diode and a hyper-bright LED. The loop can be made of copper wire, thin sheet metal or a track on a pcb. The diodes need to be wired correctly. I think the germanium diode is needed as the LED probably has too great a self-capacitance to perform at the very high AC frequencies generated by the phone (ca. 900 or 1800 Hz) but will work well with the DC pulses from the germanium diode (which has a very small capacitance).

To show the mobile generates radio waves put the mobile near to the loop and dial a number (use a free phone number, e.g. your voice mail) or text. The radio waves will induce a voltage into the loop, large enough to light the LED. The LED will flash indicating the digital data being sent by the mobile phone transmitter. You may need to set your phone to 'GSM 900/1800' rather than the '3G' network in the settings menu.

parts:
germanium diode: Maplin Electronics: QH71N or Rapid Electronics: 47-3114
LED: Maplin Electronics: UF72P or Rapid Electronics: 55-0085

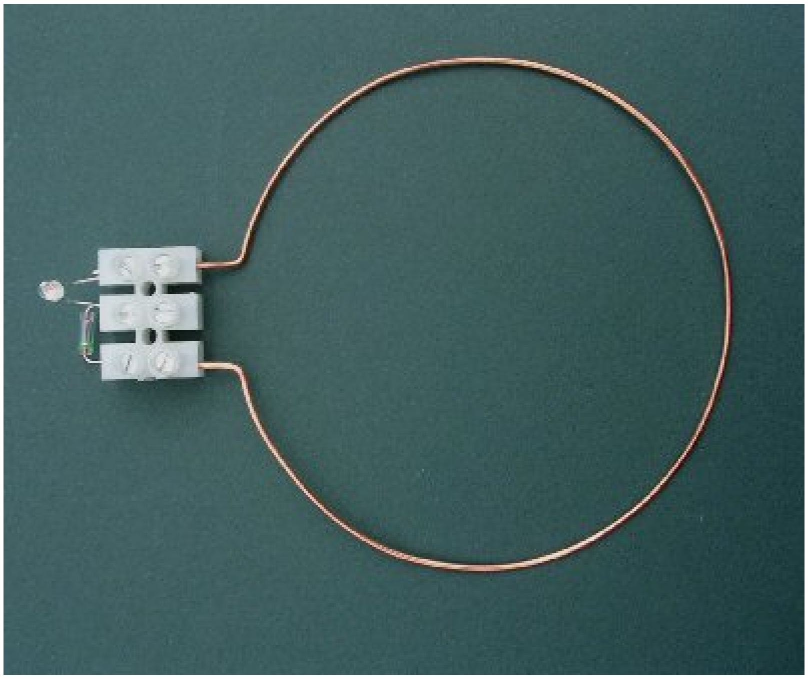

A very simple connector block version and a circular 1 wavelength loop

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, Brighton, East Sussex. BN1 9QJ.

home | diary | whats on | CSC summary | latest news

BACK TO G1EXG RADIO PAGE

EPAD MOSFET - 'near zero' and 'zero' threshold devices
(e.g. the ALD110900)

I will add more details to this page later on but for the moment it gives links to important and inspiring sites.

ARTICLES
1) High Sensitivity Crystal sets, Technical Topics, Pat Hawker, Rad Com, p. 77-78, May 2007.
2) A novel kind of 'crystal set' radio, Giles Read, Rad Com, p. 60-61, June 2007.
3) also see Pat Hawker, Technical Topics, Rad. Com. January 200, page 53-57.
4) High Sensitivity Crystal Set by By Bob Culter (N7FKI)
5) Next-generation Zero-Threshold Voltage EPAD™ design enables circuits with greater operating range in low voltage supply environments

LINKS ABOUT THE CHIPS / DEVICES
1) Advanced Linear Devices web site
2) EPAD MOSFET (ALD web site)
3) ALD110800 and ALD110900 page (ALD web site)
4) ALD110802 and ALD110802 page (ALD web site)

ORDERING THE CHIPS / DEVICES
1) mouser.com (USA)

'CRYSTAL SET' RADIO'S - DIFFERENT DETECTORS

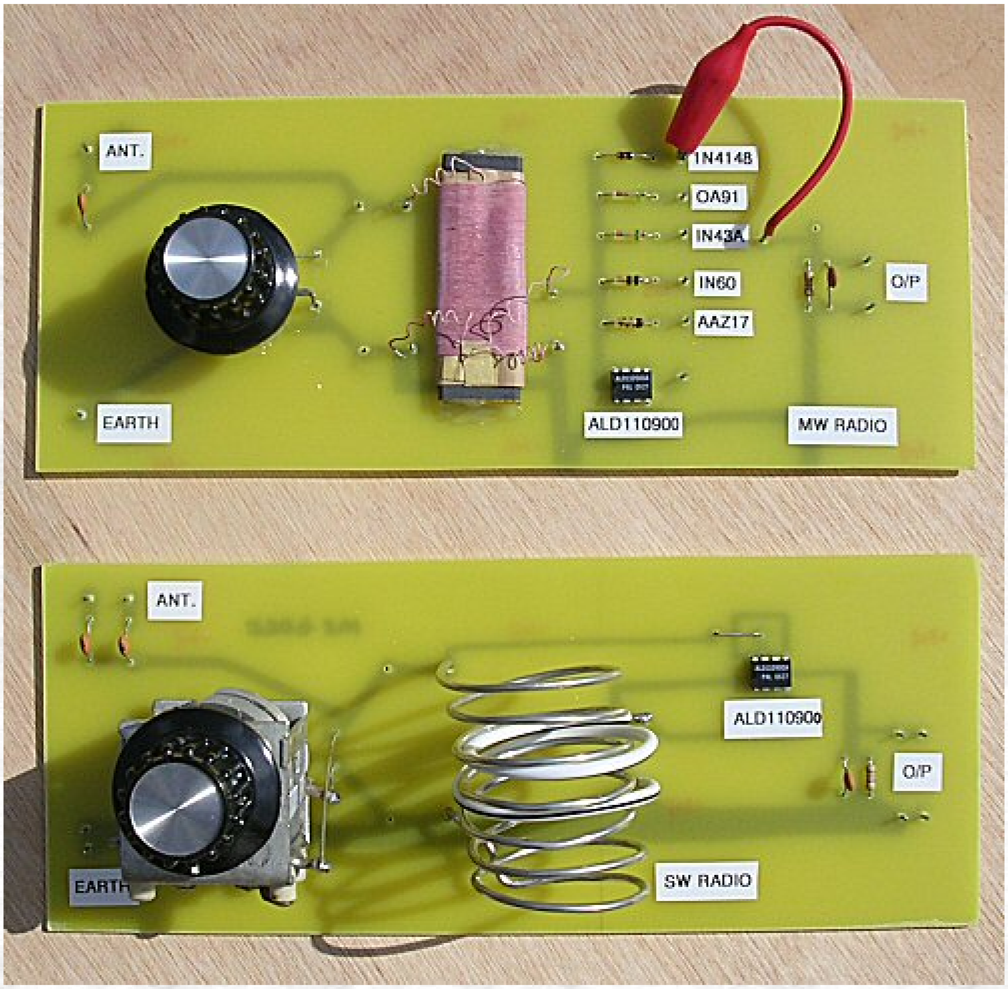

Top: the medium wave (MW) test radio with a selection of 'detectors' (selected by the croc clip lead)
Bottom: the short wave (SW) version with a ALD110900 detector
(on each circuit board the connections are: antenna - top left, Earth - bottom left, audio output (heaphones / amp) - far right.)

RADIO AND RADIO WAVES

The simplest 'radio' can be a piece of wire attached to the input of an amplifier. Why does this pick up radio signals? Well in principle it should not be able to pick up anything but in reality poor solder joints in the amplifier circuitry, as well as point contact effects between the ends of the anenna wire going into the input socket of the amp as well as other effects mean that you do often hear radio signals.

Radio waves are electromagnetic waves as they pass through a metal they induce small voltages into it. Any piece of metal e.g. metal fram specs, tape measures, metal window frames, a piece of wire etc. will act as an antenna and have tiny voltages induced into them. These voltages will be due to natural radio waves (from Space as well as the Sun and Earth), radio stations, satellites, mobile phones, garage door remote controls, microwave ovens .... the list goes on and on.

In the case of long, medium and short wave radio stations the amplitude of the radio wave signal is modified by the music, voice or program - we call it Amplitude modulation AM. Here the strength of the radio waves varies as the tones, loudness and pitch of the program vary. As a result the voltages induced in the metal object antennas distant from the radio wave stations also vary in accord this program information or modulation.

To actually hear the programs on the radio waves you cant actually take this tiny signal and listen to it directly (say with headphones or an amplifier). This is because the signal is a very high frequency signal out of the range of human hearing. To get the audible information - the program - 'off' the radio signal you need a device known as a 'detector'. This is usually a diode but lots of other things can act as inefficent detectors for example a mineral called galena, coke (burnt coal) ... as well as rusty screws and bad solder joints which is how the ampilfier mentioned above apparently picks up radio signals.

In the early days of radio crystals of galina were used as diodes and so these radios became known as 'crystal set radios'. They did not have any amplifiers and did not require any battery to work. They got all their power from that induced into the (of very long) antenna by the radio signal(s). The radio detector is an extremely important part of the radio and its proper function determines how well the radio works.

The radio signal is an AC signal. A speaker or headphones can not resonate at such high frequencies as radio waves and if you try to wire in the RF to the headphones the diaphram just averages the power. As half the time the RF is positive and half the time negative, in an AC signal, the average is zero and so you dont hear anything. The diode or detector is a device called a diode which only conducts electricity in one direction. So it only allows the positive, or negative half (one half but but not both) of the AC signal to go through (depending on which way it is wired). After the diode the average is now no longer zero its actually a changing signal dependant on the modulation of the radio wave - which is the information we want to hear. As this is at audio frequencies, and as the diaphram can move at audio frequencies, you hear the information on the detected RF signal - the music, voice or what ever.

Why should some detectors (diodes) be 'better' than others? The answer is not that some are better at 'magnifying' the signals but rather that each diode requires a certain threshold voltage in order to start conducting (one way). Signals below this level will not get passed through and a signal only slightly larger will therefore only pass through weekly (the diode unfortunatly absorbing the majority of the signal). As a result for large signals you wont expect to see much difference between different detectors but for weak signals there can be a great deal of difference. In an ideal world the diode would start to (forward) conduct at just above zero volts and this ideal detector (which does not conduct at all in reverse) would produce the greatest signal possible for the particular strength signal applied to it.

Recently I was playing around with OA91 germanium diodes for a mobile phone detector. I made a simple 10cm per side square loop and wired it directly to an LED. I was hoping that if I brought the mobile near to the loop and texted, or made a phone call, the RF should be picked up by the loop (which is almost resonant) and the LED should light - it didnt work! But putting an OA91 in series I did get it to work. So why did I need the OA91 diode when the LED is itself a diode?

It can't be due to the threshold voltage in this case as I am still using the LED and I did get it to work with the same signal! So I guess it because the capacitance of the LED is very high compared to the point-contact germanium diode. If we extend this thinking to the crystal set radio I guess the germanium diodes have very little capacitance and so dont allow any of the 'wrong' half of the AC signal through and so act as a more perfect detector. An LED would have a high capacitance and so let some of the wrong half of the AC signal through which on averaging by headphones would be nearer zero (see above) - hence less sound.

A low threshold (turn-on) voltage and a low self capacitance are crucial things in a good detector. Point contact germanium diodes having a very small area (a point of contact) have low capacitance and the germanium semiconductor-metal junction of the contact has a low turn on voltage. EPAD Mosfets also have low capacitance, high input resistance (low loading of the signal) and low turn on voltage.

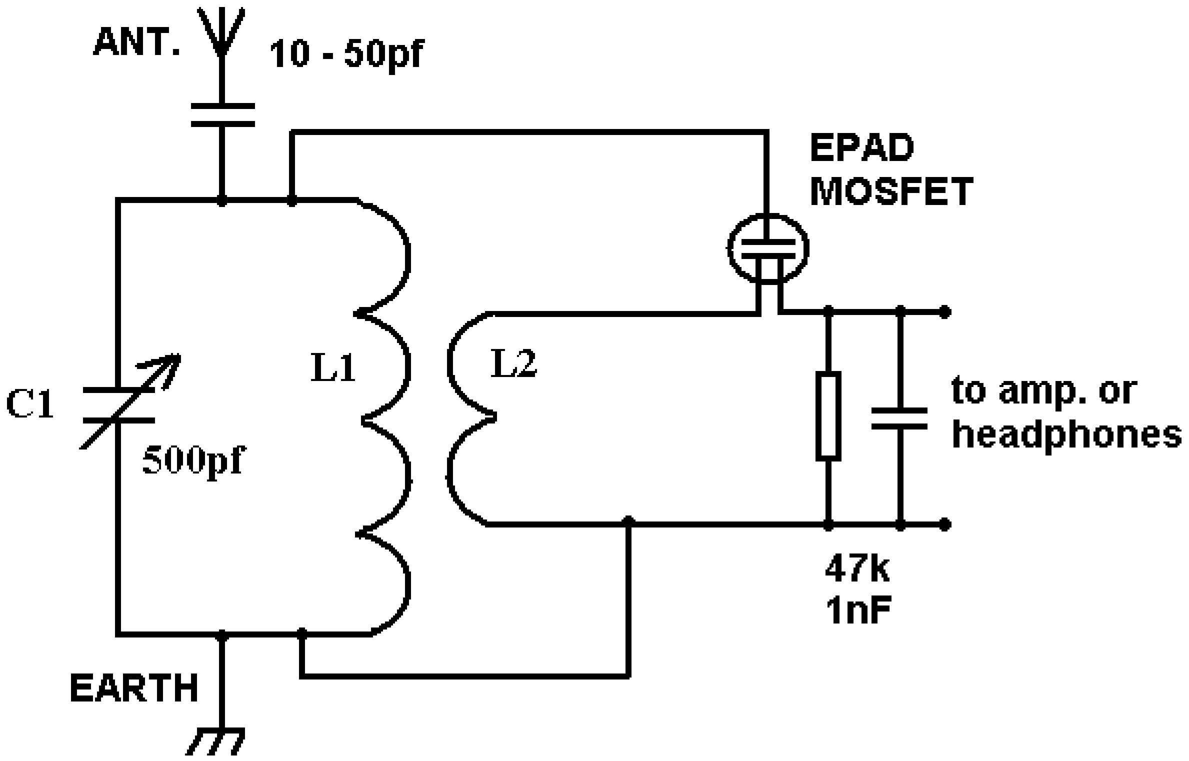

MEDIUM WAVE TEST CIRCUIT (top in photo)
As a simple test circuit to compare detectors I made up a simple 'crystal set' radio from a ferrite rod coil from an old radio, tuning capacitor an array of diodes (crystals) and a simple RF filter composed of a 47k resistor and 1nF capacitor. The coil and capacitor formed the resonant circuit which was fed via a 10 - 50pf capacitor by the antenna (long wire). This capacitor helps to reduce the effect of the antennas own capacitance and inductance from modifying the resonant frequency of the tuned circuit. Ideally this should be as small a capacitance as you can get away with but too small a value will reduce the radios sensitivity ....

An earth was connected to the ground connection of the circuit. Instead of taking the antenna end of the resonant circuit to a diode detector (as many simple crystal sets do) I used a ca. 10 turn coupling coil around the coil. This was done to limit the loading on the resonant circuit by the detector. It should help keep the tuning sharp and also help to keep the overall (band) spread of tuning as large as possible.

Because of the reduction in turns (ca. 100 turns : 10 turns, i.e. 10:1) the coupling coil reduces the signal that is available to drive the detector. For example if I use a standard silicon diode (e.g. a 1N4148) which requires about 0.6V to conduct then we might need ca. 10 x 0.6 = 6V of RF to be developed in the resonant circuit! A high Q resonant circuit might do this with a long antenna coupled to it but without such a generious antenna such a set-up wont be very sensitive.

RESULTS
I tried an array of 'crystal set' type diodes to campare them. Typical rough results are shown in the table below. This was with an Earth and about 3-4m of antenna wire randomly strung near to ground level (so its a pretty bad set up really - a good test).

The EPAD MOSFET data sheets described how the gate and drain of one of the MOSFETS can be joined together to form the anode of a diode the source then forms the cathode. I tried this arangement with the ALD110900 in order to make up a 'crystal set' diode from this MOSFET but it didnt perform well.

Then I tried the alternative arrangement as described in the references above. In these articles they dont use the MOSFETS as diodes but connect the gate directly to the resonant circuit. As the input resistance is extremely high (c.a. 1E12 ohms) and the input capacitance is very low (c.a 3pf) it does not load the tuned circuit by any appreciable amount. The source and drain are then wired between the coupling coil and filter just as the usual diode detector is in a crystal set radio. In this configuration the MOSFET is working more as a gate voltage controlled resistor rather than a diode i.e 'off' for half an RF cycle and 'on' for the rest - this lets the envelope through and you hear the music or sound. One article describes the arrangement as a synronious detector. As there are two MOSFETS in one package the two are wired in parallel i.e. gate to gate, drain to drain (there is a common source) etc.

DETECTOR TYPE	**AF voltage produced after filtering**
1N4148 silicon	no signal
OA91 germanium	1mV
1N43A germanium	1mV
1N60 germanium	1mV
AAZ17 germanium	1.5mV
ALD110900 EPAD MOSFET wired as diode	<1mV
ALD110900 EPAD MOSFET wired as sync. detc.	10mV

Typical results of the AF voltages produced by the detector (for use by headphones or amplifier) for an Earth connection, small (ca. 3-4m antenna wire near ground level) and the prototype 'crystal set' radio.

For the simple set-up described above there was not enough signal to get the silicon diode to work at all (although a much longer wire did provide a very small signal). A simple listening test and measurement of the AF prodiced by each diode showed that their was very little difference between the various germanium diodes. The MOSFET simply wired as a diode (gate wired directly the drain) was a poor detector - better than a silicon but not nearly so good as any of the germanium diodes. Finally the EPAD MOSFET wired as a syncronious detector worked really very well apparently producing ca. 10 times the signal (in other words it has much less loss than a germanium diode) for this particular set up and station than the germanium diodes.

As the syncronious arrangement was so succesful it got me thinking about this circuit. As the gate goes straight in on the main resonant circuit where the voltages, in a good Q circuit, might well be relatively high I thought I would try a standard FET (2N3819) and MOSFET (VN10KM) instead - and they worked! (they are a lot cheaper than an EPAD MOSFET chip and easier to get). Some FETS have a larger input capacitance so it helped to put a 10pf capacitor in series to the gate (a 10M ohm from gate to earth might reduce 50Hz pick up).

They were not as good as the EPAD MOSFET (although I only tried a single device) but the arrangment was still better than a simple 'crystal set' diode. For weak signals the EPAD MOSFET should perform better as it is able to keep going while the FET and the standard MOSFET will start to trail off as the resonant circuit voltage drops below the relatively higher gate voltage required for these devices. However it was interesting to see that it seems to be possible to get a more efficent set-up for a 'crystal' radio using a standard FET or MOSFET than with the standard germanium diode detector.

By the way I recently re-tried the 1N4148 diode but wired a 1.5V AA battery with a 100k resistor across it to forward bias it - and it worked ! It seemed to work better than a OA91!

SHORT WAVE 'CRYSTAL SET RADIO (bottom in photo)

The bottom circuit shown in the photo above is my own short wave version of the 'crystal set' radio but based on the references above. It's the same circuit as the MW radio but has a reduction drive tuning capacitor and different coil. I used about 7 turns (ca. 4cm diameter) for the resonant circuit and a couple of turns of insulated wire (white) for the coupling coil. I guess the radio tunes from about 3-8 MHz - I havent tried a sig gen on it yet.

I made this in the summer months (not the best time for listening on this part of the spectrum). Using an Earth and few meters of wire for antenna I heard 2 or 3 stations from around the world during the day while at night there were at least 7 including the voice of Russia world service, Radio Thailand as well as German and French speaking stations.

BACK TO G1EXG RADIO PAGE

THE CREATIVE SCIENCE CENTRE

Dr Jonathan Hare, The University of Sussex
Brighton, East Sussex. BN1 9QJ

home | diary | whats on | CSC summary | latest news

PUBLISHED BY

INTECH
open science | open minds

World's largest Science,
Technology & Medicine
Open Access book publisher

3,200+
OPEN ACCESS BOOKS

105,000+
INTERNATIONAL
AUTHORS AND EDITORS

111+ MILLION
DOWNLOADS

BOOKS
DELIVERED TO
151 COUNTRIES

AUTHORS AMONG
TOP 1%
MOST CITED SCIENTIST

12.2%
AUTHORS AND EDITORS
FROM TOP 500 UNIVERSITIES

BOOK CITATION INDEX
WEB OF SCIENCE™

Selection of our books indexed in the
Book Citation Index in Web of Science™
Core Collection (BKCI)

Chapter from the book *Sustainable Energy Harvesting Technologies - Past, Present and Future*

Downloaded from: http://www.intechopen.com/books/sustainable-energy-harvesting-technologies-past-present-and-future

Interested in publishing with InTechOpen?
Contact us at book.department@intechopen.com

10

Design Issues in Radio Frequency Energy Harvesting System

Chomora Mikeka and Hiroyuki Arai
Yokohama National University
Japan

1. Introduction

Emerging self powered systems challenge and dictate the direction of research in energy harvesting (EH). State of the art in energy harvesting is being applied in various fields using different single energy sources or a combination of two or more sources. In certain applications like smart packaging, radio frequency (RF) is the preferred method to power the electronics while for smart building applications, the main type of energy source used is solar, with vibration & thermal being used increasingly. The main differences in these power sources is the power density; for example RF (0.01 ~ 0.1 μW/cm^2), Vibration (4 ~ 100 μW/cm^2), Photovoltaic (10 μW/cm^2 ~ 10mW/cm^2) and Thermal (20 μW/cm^2 ~ 10mW/cm^2). Obviously RF energy though principally abundant, is the most limited source on account of the incident power density metric, except when near the base stations. Therefore, in general, RF harvesting circuits must be designed to operate at the most optimal efficiencies.

This Chapter focuses on RF energy harvesting (EH) and discusses the techniques to optimize the conversion efficiency of the RF energy harvesting circuit under stringent conditions like arbitrary polarization, ultra low power (micro or nanopower) incidences and varying incident power densities. Harvested power management and application scenarios are also presented in this Chapter. Most of the design examples described are taken from the authors' recent publications.

The Chapter is organised as follows. Section 2.1 is the introduction on RF energy sources. Section 2.2 presents the antenna design for RF EH in the cellular band as well as DTV band. The key issue in RF energy harvesting is the RF-to-DC conversion efficiency and is discussed in Section 2.3, whereas Section 2.4 and 2.5 present the design of DTV and cellular energy harvesting rectifiers, respectively. The management of micropower levels of harvested energy is explained in Section 2.6. Performance analysis of the complete RF EH system is presented in Section 3.0. Finally, conclusions are drawn in Section 4.0.

1.1 RF energy sources

These include FM radio, Analogue TV (ATV), Digital TV (DTV), Cellular and Wi-Fi. We will present a survey of the measured E-field intensity (V/m) for some of these RF sources as shown in Table 1, [1]-[2]. Additionally, measured RF spectrums for DTV and Cellular signals are presented as shown in Fig. 1 to show on the potential for energy-harvesting in these frequency bands. In general, many published papers on RF-to-DC conversion, have presented circuits capable of converting input or incident power as low as -20dBm. This means that, if an RF survey or scan finds signals in space, with power spectrum levels around -20dBm, then, it is potentially viable to harvest such signal power. In Fig. 1 (left side), the spectrum level is well above -20dBm and hence, a higher potential for energy harvesting. In Fig. 1 (right side), while the spectrum level is below -20dBm, what we observe is that the level increases with decrease in the distance toward the base station (BTS). Using free space propagation equation with this data, it was calculated that at a distance 1.4 m from the BTS, the spectrum level could measure 0dBm. An example calculation and plot for the estimated received power level, assuming 0dBi transmitter (BTS) and receiver antenna gains and free space propagation loss (FSPL) for FM and DTV is presented in Section 2.1.1. For the example estimation in Section 2.1.1, we select FM and DTV because they measured with a higher level than cellular and Wi-Fi for example.

<table>
<tr><th>Source</th><th>V/m</th><th>dBm</th><th colspan="2">Reference</th></tr>
<tr><td>FM radio</td><td>0.15~3</td><td></td><td colspan="2" rowspan="2">Asami et al.</td></tr>
<tr><td>Analogue TV</td><td>0.3~2</td><td></td></tr>
<tr><td>Digital TV</td><td>0.2~2.4</td><td>-40~0.0</td><td>Asami et al.</td><td>Arai et al.</td></tr>
<tr><td>Cellular</td><td></td><td>-65~0.0</td><td colspan="2">Mikeka and Arai</td></tr>
<tr><td>Wi-Fi</td><td></td><td>≅ -30</td><td colspan="2"></td></tr>
</table>

Table 1. RF energy sources, measured data.

In Table 1, FM radio has the highest E-field intensity implying the highest potential for energy harvesting. However, due to the requirements for a large antenna size and the challenges for simulations and measurements at the FM frequency i.e. 100 MHz or less (See Section 2.2.3, example FM antenna at 80 MHz), this Chapter will focus on DTV (470~770 MHz band) and Cellular (2100 MHz band) energy harvesting.

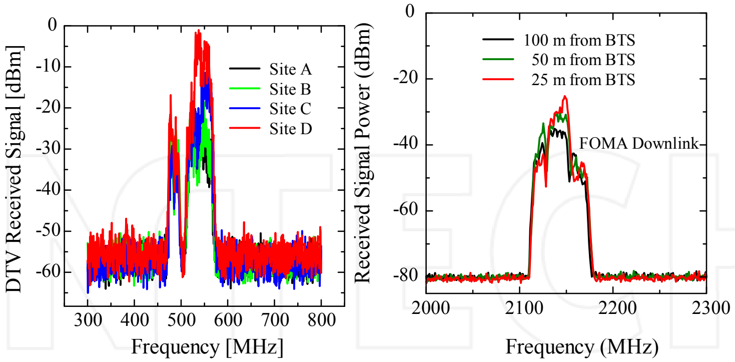

Fig. 1. DTV signal spectrum measured in Tokyo City (left side graph) and Cellular signal spectrum measured in Yokohama City (right side graph).

The received DTV signal power is high and also wide band, presenting high potential for increased energy harvesting unlike in cellular signals. We demonstrated in [2] that the total RF-to-DC converted power is roughly the integral over the DTV band (1), and is significantly larger than in the case of narrow band cellular energy harvesting.

$$P_{DC(DTV)} = \alpha \int_{470}^{770} \delta P_{DC}(f)\,df \,, \tag{1}$$

where α is the attenuation factor on the rectifying antenna's RF-to-DC conversion efficiency due to multiple incident signal excitation. δP_{DC} is the small converted DC power from each of the single DTV signals in the 470 MHz to 770 MHz band.

In detail, we derive (1) from fundamentals as follows.

The incident power density on the rectifying antenna (rectenna), $S(\theta,\phi,f,t)$, is a function of incident angles, and can vary over the DTV spectrum and in time. The effective area of the antenna, $A_{eff}(\theta,\phi,f)$, will be different at different frequencies, for different incident polarizations and incidence angles. The average RF power over a range of frequencies at any instant in time is given by:

$$P_{RF}(t) = \frac{1}{f_{high} - f_{low}} \int_{f_{low}}^{f_{high}} \int_{0}^{4\pi} S(\theta,\phi,f,t) A_{eff(\theta,\phi,f)}\, d\Omega\, df \tag{2}$$

The DC power for a single frequency (f_i) input RF power, is given by

$$P_{DC}(f_i) = P_{RF}(f_{i,}t).\eta\left(P_{RF}(f_i,t),\rho,Z_{DC}\right), \tag{3}$$

where η is the conversion efficiency, and depends on the impedance match $\rho(P_{RF},f)$ between the antenna and the rectifier circuit, as well as the DC load impedance. The reflection coefficient in turn is a nonlinear function of power and frequency.

The estimated conversion efficiency is calculated by P_{RF}/P_{DC} . This process should be done at each frequency in the range of interest. However, DC powers obtained in that way cannot be simply added in order to find multi-frequency efficiency, since the process is nonlinear. Thus, if simultaneous multi-frequency or broadband operation like in DTV band is required, the above characterization needs to be performed with the actual incident power levels and spectral power density. In this Chapter, we shall demonstrate DTV spectrum power harvest, given a rectenna than has been characterised in house at each single frequency in the DTV band.

1.1.1 An example calculation and plot for the estimated received power level

In this example we consider Tokyo's DTV and FM base stations (BS) as the RF sources. Both DTV and FM BS transmitter power (P_t) equals 10 kW (70dBm). The antenna gains are assumed 0dBi in both cases but also at the points of reception for easiness of calculation but with implications as follows. Assuming 0dBi antenna at each reception point, demands that we specify the frequency of the transmitted signal. For this reason we specify DTV signal frequency to be equal to 550 MHz while the FM signal frequency equals 80 MHz (Tokyo FM).

The received power, P_r is calculated using the simplest form of Friis transmission equation given by

$$P_r = P_t + G_t + G_r + FSPL\,, \tag{4}$$

where P_t = 70dBm, $G_t = G_r$ = 0dBi. G_r is the receiving antenna gain while FSPL is the free-space path loss given by

$$FSPL(dB) = 20\log(d) + 20\log(f) + 32.45\,, \tag{5}$$

where ***d*** is in (km) and ***f*** is in (MHz). The plot for the received power as a function of distance from the DTV and FM base stations is shown in Fig. 2.

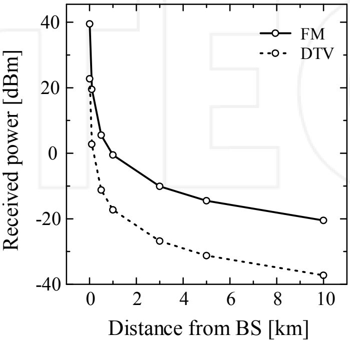

Fig. 2. DTV and FM received signal power level against distance.

With respect to Fig. 2, FM registers higher received power level than DTV at every reception point due to its lower transmit frequency and hence lower free-space path loss. For example at 1 km distance, FM received power level is -0.51dBm while for DTV, the received power is -17.26dBm. The important thing however, is that the received power level is frequency independent. It means that P_t is the transmitter power and the received power level at the position of distance ***d*** is $\frac{P_t}{4\pi d^2}$. However, if we assume 0dBi antenna at each reception point as in the above example, the power level is different because the antenna size of 0dBi is frequency dependent. As a result, high transmit power level is favorable for RF energy harvesting. Also near the base station is favorable.

1.2 Antenna design for the proposed RF energy harvesting (EH) system

It is well known that RF EH system requires the use of antenna as an efficient RF signal power receiving circuit, connected to an efficient rectifier for RF-to-DC power conversion. Depending on whether we want to harvest from cellular or DTV signals, the antenna design requirements are different. We will discuss the specific designs in the following sub sections.

1.2.1 Cellular energy harvesting antenna design

We propose a circular microstrip patch antenna (CMPA) for easy integration with the proposed rectifier (Section 2.5.1). However, the use of circular microstrip patch antennas (CMPA) is often challenged by the need for impedance matching, circular polarization (CP) and higher order harmonic suppression.

To address the above concerns, we create notches on the circular microstrip patch antenna. In our approach, we use only two, thin, fully parameterized triangular notches to achieve higher order harmonic suppression, impedance matching and circular polarization, all at once. This is the novelty in our proposed antenna. Our proposed CMPA is shown in Fig. 3. We study the behaviour of CMPA surface current vectors when notches (triangles ABC) are created on the structure at $\alpha = 45^0$ and $\alpha = 225^0$.

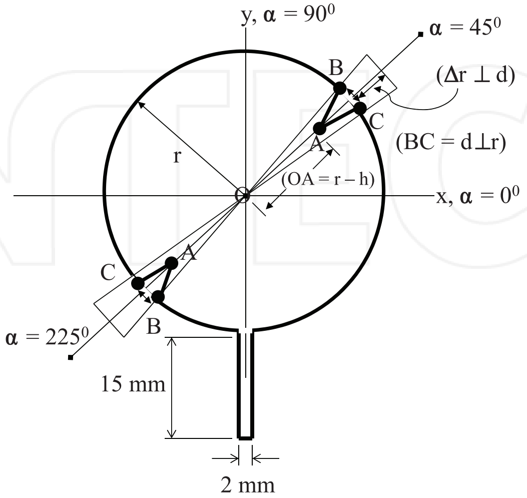

Fig. 3. Cellular energy harvesting antenna structure.

Notch parameters ***d*** and ***h*** in Fig. 3 were investigated by calculation using CST microwave Studio.

Without notches, the CMPA's input is not matched at f_c= 2.15 GHz as shown in Fig. 4 (left side). However, with notches, matching is achieved. The parameter combination ***d*** = 7 and ***h*** = 6 offers a matched and widest band input response and hence we adopt it for cellular energy harvesting applications.

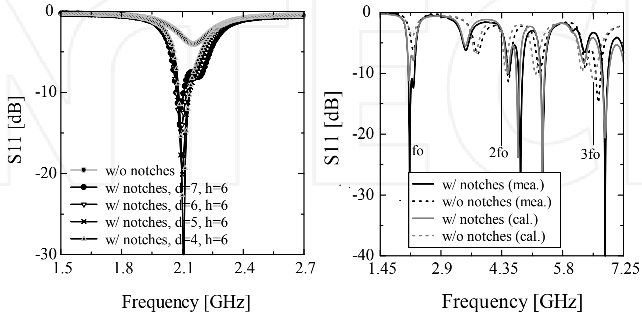

Fig. 4. *Left side:* ***d*** and ***h*** parameter investigation. *Right side:* Comparison between (cal.) and (mea.) S_{11} at f_0 = 2.175 GHz, $2f_0$ = 4.35 GHz and $3f_0$ = 6.53 GHz. The adopted notch parameters are ***d***=7 mm while ***h***= 6 mm.

The comparison between calculated and measured S_{11} is shown in Fig. 4 (right side). The 2nd and 3rd harmonics are suppressed as required by design. The comparison between calculated and measured radiation patterns is shown in Fig. 5, where $E_\vartheta \cong E_\varphi$ due to the 45^0 tilted surface current vector. Ordinarily, without notches, the surface current vector is parallel to the microstrip feeder axis. In conclusion, our proposed CMPA is sufficiently able to suppress higher order harmonics while simultaneously radiating a circularly polarized (CP) wave. The CP is required to efficiently receive the arbitrary polarization of the incident cellular signals at the rectenna.

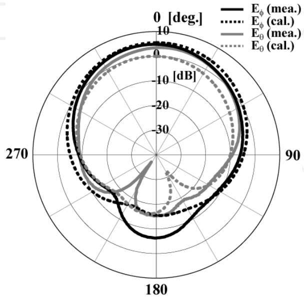

Fig. 5. Cellular energy harvesting antenna pattern at f_c= 2.15 GHz.

1.2.2 DTV energy harvesting antenna design

Unlike in the cellular energy harvesting antenna, the DTV energy harvesting antenna must be wideband (covering 470 MHz to 770 MHz), horizontally polarized and omni-directional.

The proposed antenna is typically a square patch (57 mm x 76 mm) with a partial ground plane (9 mm x 100 mm). The patch is indirectly fed by a strip line (9 mm x 3 mm). The proposed antenna geometry is shown in Fig. 6. The partial ground plane is used to achieve omni-directivity and a certain level of wide bandwidth. To tune the impedance of this antenna as well as to adjust the bandwidth within the target band, a "throttle" with stepped or graded structures is used between the microstrip feed line and the square patch, as shown in Fig. 6 (left side).

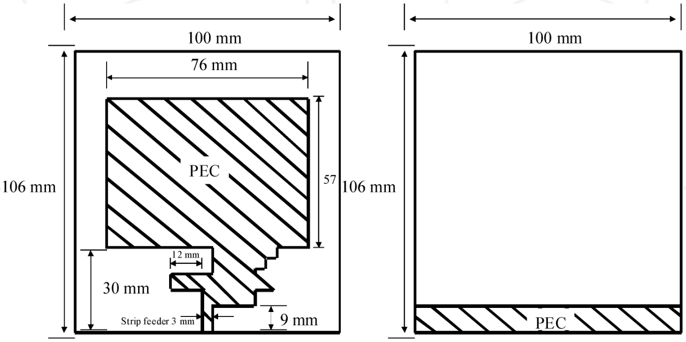

Fig. 6. Proposed DTV antenna geometry. *Left side:* Front view. *Right side:* Back view. The antenna is printed on FR4 substrate; t = 1.6 mm, ε_r = 4.4.

The input response for the proposed antenna is shown in Fig. 7 (left side). The omni directivity is confirmed by measurement at 500 MHz, 503 MHz, and 570 MHz as shown in Fig. 7 (right side). The radiation patterns shown in Fig. 7 are for the xz plane, which happens to be the vertical polarization for the antenna. DTV signals are horizontally polarized and therefore, when using this antenna, the orientation must be in such a way as to efficiently receive the DTV signal. Simply a 90 degree rotation of the antenna along the z axis achieves this requirement.

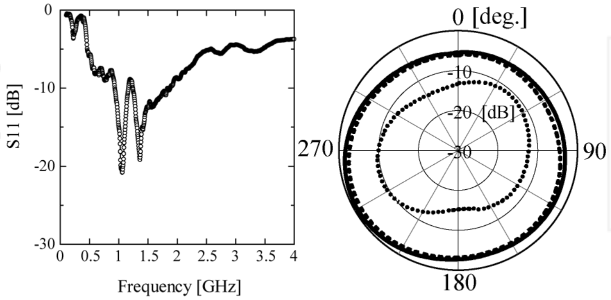

Fig. 7. Proposed DTV antenna performance. *Left side:* The antenna's measured input performance. *Right side:* The omni directivity in the vertical plane is confirmed at 500 MHz, 503 MHz, and 570 MHz. The outermost, black solid and dotted line patterns represent 503 MHz and 500 MHz directivity, respectively. The innermost dotted line pattern is the directivity at 570 MHz.

1.2.3 Example design for an 80 MHz FM half-wave dipole antenna

A half-wave dipole is the simplest practical antenna designed for picking up electromagnetic radiation signals, see Fig. 8 (courtesy of Highfields Amateur Radio Club). Calculating the optimal antenna length to pick up a certain frequency signal is fairly straightforward because antenna physics demand that the total length of wire used in the antenna be equal to one wavelength of the type of electromagnetic radiation it will be picking up. This means that the total length of the antenna should be equal to half the desired wavelength. By converting the 80 MHz frequency into a wavelength, you can thus obtain your antenna length as 1.875m by using the magic equation, $\lambda = \frac{c}{f}$. However, the actual length is typically about 95% of a half wavelength in free space, hence a half-wave dipole for this frequency should be 1.788m long, which would make each leg of the dipole 0.894m in length.

1.3 RF-to-DC conversion efficiency improvement techniques

A Schottky diode circuit connected to an antenna is used for RF-to-DC power conversion. To convert more of the antenna surface incident RF power to DC power, high RF-to-DC conversion efficiency is required of the rectifying circuit. Many authors have shown that the efficiency depends on several factors like Schottky diode type, harmonics suppression capability, load resistance selection, and the capability to handle arbitrary polarized incident waves. What is missing in most of these published works is the efficiency optimization for

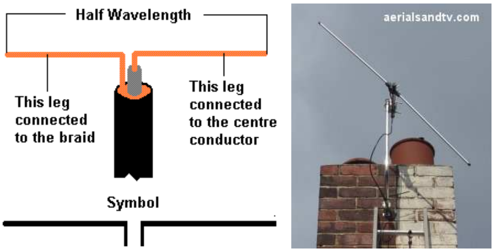
Half Wavelength
This leg connected to the braid
This leg connected to the centre conductor
Symbol
aerialsandtv.com

Fig. 8. Half-wave dipole. *Left side:* Antenna structure. *Right side:* Typical deployment.

ultra low power incident waves and the explanation of the physical phenomena behind most of the recommended efficiency optimization approaches.

This Chapter will show for example that a Schottky diode that delivers the highest efficiency at 0dBm incidence may not necessarily deliver the highest efficiency at lower power incidence e.g. –20dBm. We will therefore classify which diodes perform better at given power incidences; of course, this will also be compared to the diode manufacturers' application notes. Simulations in Agilent's ADS using SPICE and equivalent circuit models will compare the performance of few selected Schottky diodes namely; HSMS-2820, HSMS-2850, HSMS-2860, HSC-276A, and SMS7630. Moreover, the effect of the Schottky diode's junction capacitance (C_j) and junction bias resistance (R_j) on the conversion efficiency will be shown from which, special techniques for Schottky diode harmonic suppression and rectifying circuit loading for maximum efficiency point tracking will be presented.

1.3.1 The schottky diode

The classical *pn* junction diode commonly used at low frequencies has a relatively large junction capacitance that makes it unsuitable for high frequency application [3]. The Schottky barrier diode, however, relies on a semiconductor-metal junction that results in a much lower junction capacitance. This makes Schottky diodes suitable for higher frequency conversion applications like rectification (RF-to-DC conversion) [3]. We will demonstrate the effects of junction capacitance and resistance in the following sub section.

1.3.2 The effect of Schottky diode's C_j and R_j on the conversion efficiency

We have studied Schottky diode's C_j and R_j and published our results in [4]. In this work, we designed a rectifying antenna tuned for use at 2 GHz. The circuit proposed in [4] is a voltage doubler by configuration, but we replaced the amplitude detection diode (series diode) with its equivalent circuit adapted from [5]. The results of this investigation show that variation of C_j shifts the tuned frequency position and also introduces a mismatch in the resonant frequency, see Fig. 9 (left side graph). Therefore for this circuit at 2 GHz, we recommend using a Schottky diode having C_j = 0.2pF. In general, a smaller value of C_j is desirable at higher frequencies. Similarly, for R_j investigation, a smaller value is desirable for better matching at 2 GHz for example. If the R_j is increased towards 10kΩ, there is a mismatch in the resonant frequency but no shift in the frequency, see Fig. 9 (right side graph). Another approach to the study of Schottky diodes for higher frequency and efficiency rectenna design is presented in [6].

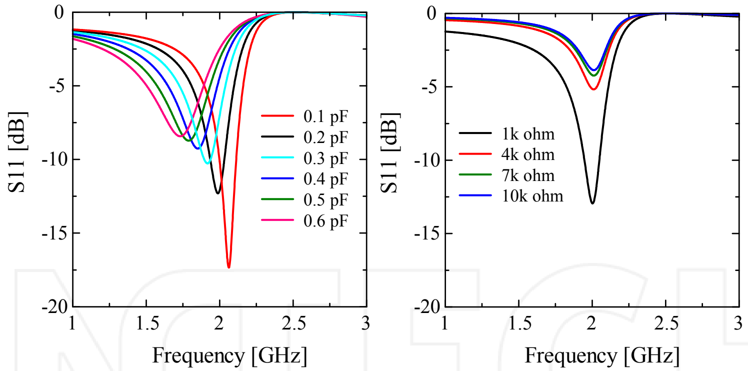

Fig. 9. Schottky diode's C_j effect (Left) and R_j effect (Right) on 2 GHz rectenna's input response.

1.4 Rectifying circuit for DTV energy harvesting

In the design of a DTV energy harvesting circuit, several basic design considerations must be paid attention to. First is the antenna; it must be wideband (covering 470 MHz to 770 MHz), horizontally polarized and omni-directional. Secondly is the rectifier; it must also be wideband, and optimized for RF-to-DC conversion for incident signal power at least -40dBm. Until recently, very few authors have published on DTV energy harvesting circuit. For the few publications, the antenna could not meet all those three requirements and a discussion on the performance of the harvesting circuit for ultra low power incidences has been neglected. In this Chapter we will present such a rectenna with conversion efficiencies above 0.4% at -40dBm, above 18.2% at -20dBm and over 50% at -5dBm signal power incidence. We will closely compare simulated and measured performance of the rectenna and discuss any observed disparities.

Agilent's ADS will be used to simulate the nonlinear behaviour of the rectifying circuit based on harmonic balance tuning methods. To simulate the multiple incident waves, a multi-tone excitation in the DTV band will be invoked. The wideband input characteristic will be achieved by the input matching inductors and capacitors.

The generic version of our proposed DTV energy harvesting circuit is shown below in Fig. 10. The implementation, however, is in two phases or scenarios as follows. First, we investigate the class called "ultra low power" DTV band rectenna. Secondly, we introduce the "medium power" DTV band rectenna.

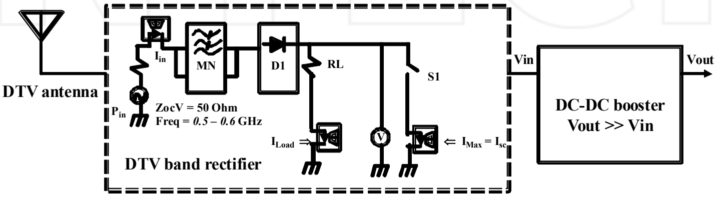

Fig. 10. Generic version of our proposed DTV energy harvesting circuit.

1.4.1 Ultra low power DTV rectenna

We define an ultra low power rectenna as one impinged by RF power incidence in the range between – 40dBm and -15dBm. Below in Fig. 11 is the circuit we designed; optimized for -20dBm input. The matching network is complex so as to achieve a wide band input characteristic. The fabricated circuit was well matched for the frequency range between 470 MHz and 600 MHz. More details about the circuit design can be found in [7].

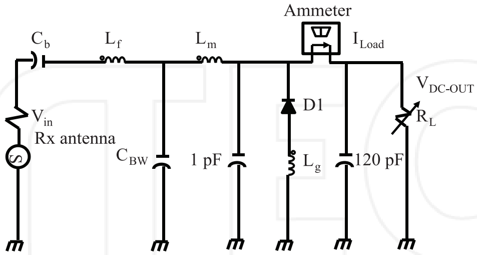

Fig. 11. Ultra low power DTV band rectenna circuit. SMS7630 Schottky diode by SKYWORKS offered the best performance.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 12 where at input power equal to -40dBm, efficiency is at least 0.4% and rectified voltage equals 1mV; at -20dBm, we have at least 18.2% by measurement and a rectified voltage of 61.7mV. The level of rectified voltage is too low and disqualifies this circuit for purposes of charging capacitors or batteries to accumulate such micropower over time. Instead, boosting the low voltage to usable levels is the option available and we shall discuss this at a later stage, (in Section 2.6).

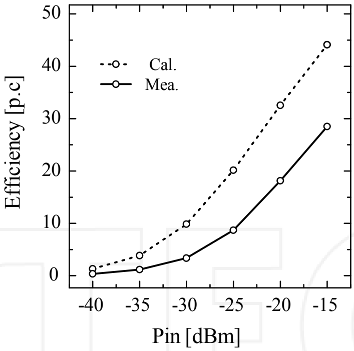

Fig. 12. Ultra low power DTV band rectenna efficiency.

1.4.2 Medium power DTV rectenna

We define a medium power rectenna as one impinged by RF power incidence in the range between – 5dBm and 0dBm. Below in Fig. 13 is the circuit we designed, optimized for -5dBm input. The matching network is simpler than as shown in section 2.4.1 since we require a narrow band around 550 MHz, with received peak power spectrum levels at least -5dBm. The circuit in Fig. 13 is a modification of Greinacher's doubler rectifier. In the circuit, C_b equals 1 pF and is used to block DC current against flowing towards the source. The shunt capacitance, C_{BW} equals 3300 pF and is used to set the input bandwidth. The grounding inductance, L_g equals 56nH (optimal) and is used to improve the RF-to-DC conversion efficiency by cancelling the Schottky diodes (D_b and D_D) capacitive influence; thereby minimizing the harmonic levels (harmonic suppression). We used HSMS2850 diodes in these circuits for their better performance at this level of incident power.

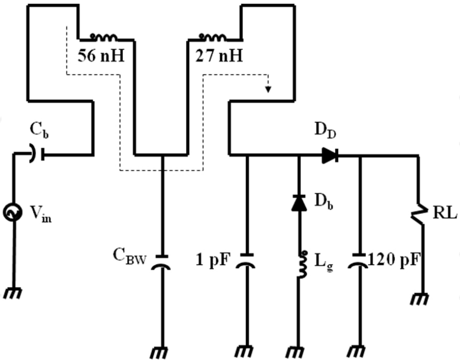

Fig. 13. Medium power DTV rectenna circuit. HSMS 2850 or 2820 from Hewlett-Packard offered the best performance.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 14 where at input power equal to -5dBm, we achieve at least 50% conversion efficiency by measurement, equivalent to 1.2 V DC rectified at 8.2kΩ optimal load. If we change the load to 47kΩ, over 2 V DC is rectified. This rectenna circuit is ideal for powering small sensors that run on 1.5 V or 2.2 V and draw around 6μA nominal current. If we need to power sensors demanding more power, say at least 2.2 V and 0.3mA to 1.47mA current consumption, we have to accumulate the power in a capacitor over time.

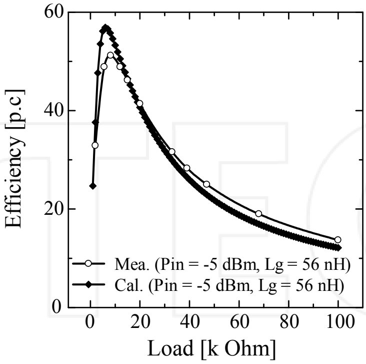

Fig. 14. Medium power DTV band rectenna efficiency.

1.4.3 DTV energy harvesting scenario and application demo

Using the medium power DTV band rectenna, connected to a gold capacitor as an accumulator, energy harvesting was initiated as shown in Fig. 15. Details about the gold capacitor, which include its charge function, backup time and leakage losses are presented in [8]. For the scenario shown in Fig. 15, the accumulated voltage by measurement i.e. capacitor charge function follows the path;

$$V_{acc} = 0.5388\ln(t) + 1.4681 \tag{6}$$

where V_{acc} is the accumulated voltage in volts and t the time in hours. It takes 4.5 hours to accumulate 2.25 V, given a rectified charging voltage and current of 2.4 V and 51μA, respectively, supplied by the DTV band rectenna instantaneously.

With this rectenna, it was possible to power up many different kinds of sensors. Sensors with ultra low power consumption were powered directly, without need to accumulate the power in a capacitor, as shown in Fig. 16.

INTECH

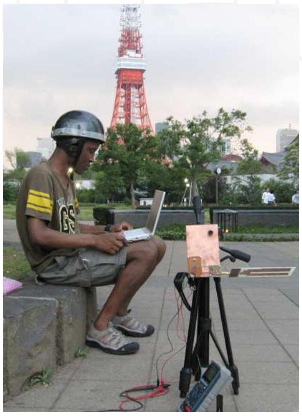

Fig. 15. DTV energy harvesting in a park at some line of sight from the base station.

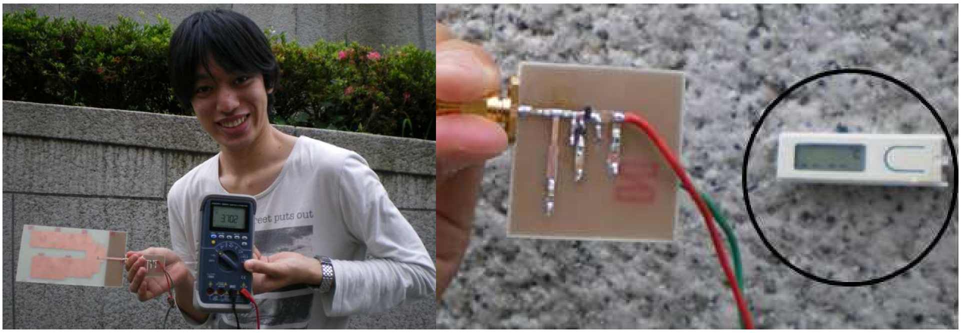

Fig. 16. Directly powering a thermometer mounted on a car park wall (right picture). The maximum instant voltage rectification on record equals 3.7 V (left picture).

1.5 Rectifying circuit for cellular energy harvesting

Unlike in the DTV energy harvesting circuit, for cellular energy harvesting, the antenna must be narrowband (50 MHz bandwidth is acceptable), and circularly polarized even though cellular signals are vertically polarized. The circular polarization is desired to maximize the RF-to-DC conversion efficiency of the arbitrary polarization incident signals in the multipath environment. Similarly, the rectifier must be narrowband, and optimized for RF-to-DC conversion over a wide range of incident signal power.

Thinking about the potential applications for cellular energy harvesting is useful. Other authors have reported on powering a scientific calculator or a temperature sensor from GSM energy harvesting. In this Chapter we will present a special application for energy harvesting in the vicinity of the W-CDMA cellular base station and analyze the system performance by calculation from experimental data. A cellular energy harvesting circuit optimized for over 50% RF-to-DC conversion efficiency given approximately 0dBm incidence will be presented.

1.5.1 Cellular band rectenna

Below in Fig. 17 is the circuit we designed, optimized for 0dBm input. Simple input matching network is ideal since we require a narrow band response around 2.1 GHz. The optimum value for L_g equals 5.6nH, where L_g is used to improve the RF-to-DC conversion efficiency as earlier discussed. HSMS2850 diode was used.

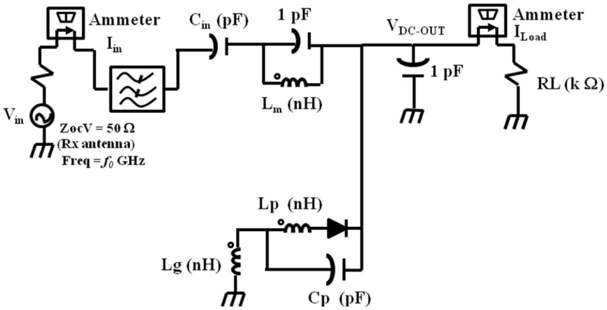

Fig. 17. Shunt rectifier configuration for the cellular band. The matching elements L_m = 3.2nH, while C_{in} =2.5pF. The load resistance is fixed at R_L = 2.1kΩ.

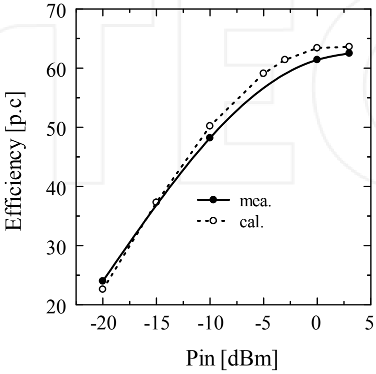

Fig. 18. Conversion efficiency as a function of input power (P_{in}) in the cellular band.

The RF-to-DC conversion efficiency for this circuit is shown in Fig. 18 where at input power equal to 0dBm, we achieve at least 60% conversion efficiency by measurement, given a 2.1kΩ optimal load. This rectenna circuit is ideal for powering small sensors that run on 1.5 V or 2.2 V and 6μA nominal current consumption. If we need to power sensors demanding more power, say at least 2.2 V and 0.3mA to 1.47mA, we have to accumulate the power in a capacitor over time as discussed in section 2.4.3 above.

1.5.2 Cellular energy harvesting application example

Environmental power generation in the neighbourhood of a cellular base station to power a temperature sensor is proposed as shown in Fig. 19 below. Electric field strength measurements in the base station neighbourhood have demonstrated the potential for environmental power generation, and the proposed temperature sensor system is designed based on these values. The rectenna described in Section 2.5.1 is used as the RF-to-DC rectifying circuit with the notched circular microstrip patch antenna (CMPA) proposed in Section 2.2.1. RF-to-DC conversion efficiency equal to 53.8% is obtained by measurement. The temperature sensor made for trial purposes clarifies the capability for temperature data wireless transmission for 20 seconds per every four hours in the base station neighbourhood.

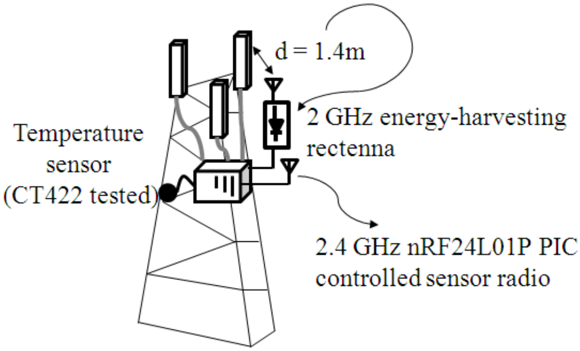
d = 1.4m
2 GHz energy-harvesting rectenna
Temperature sensor (CT422 tested)
2.4 GHz nRF24L01P PIC controlled sensor radio

Fig. 19. Application example in the vicinity of the cellular base station.

1.6 Micropower energy harvesting management

A rectifying antenna circuit for -40dBm incident power harvesting generates 1mV at 2kΩ load, given 0.4% efficiency as presented in Section 2.4.1. At -20dBm incidence and at least 18.2% efficiency, 61.7mV is generated given a 2kΩ load [7]. The generated DC power in both of these two cases is in the μW range, hence the micropower definition. To manage such micropower, power accumulation or energy storage is required. Storage devices may either be a gold capacitor, super capacitor, thin film battery or the next generation flexible paper batteries. These storage devices have specific or standard maximum voltage and trickle charging current minimum requirements. Typically, gold capacitors have voltage ratings like 2.7 V, 5.5 V for 100 μA, 10mA or 100mA maximum discharge current. On the other hand, standard ratings for batteries are 1.8 V, 2 V, 3.3 V and 4.1 V. Therefore, to directly charge any of these storage devices from 1mV, or 61.7mV DC is impractical.

Published works have demonstrated the need for a DC-to-DC boost converter placed between the rectifying antenna circuit (rectenna) and the storage device. Recent efforts have demonstrated that a 40mV rectenna output DC voltage could be boosted to 4.1 V to trickle charge some battery. A Coilcraft transformer with turns ratio ($N_s : N_p$) equal to 100 was used in the boost converter circuit. An IC chip leading manufacturer (Linear Technology Corp., LT Journal, 2010) has released a linear DC-to-DC boost regulator IC chip capable of boosting an input DC voltage as low as 20 mV and supplying a number of possible outputs, specifically suited for energy harvesting applications. While this IC is a great milestone, readers and researchers need to understand the techniques to achieve such ICs and also the limitations that apply. In the following sub section, we will describe the methods toward designing a DC-DC boost converter, suitable for micropower RF energy harvesting.

In the design, we will attempt to clarify the parameters that affect the DC-DC conversion efficiency. For this design, Envelope simulation in Agilents's ADS is used. This simulation technique is the most efficient for the integrated rectenna and DC-DC boost converter circuits.

1.6.1 DC-DC boost converter design theory and operation

The DC-DC boost converter design theory and actual implementation are presented in this section. The inequality $V_{in} \ll V_{out}$ defines the boost operation. In this Chapter, our boost converter concept is illustrated in Fig. 20. A small voltage, V_{in} is presented at the input of the boost converter inductive pump which as a result, generates some output voltage, V_{out}. The output voltage is feedback to provide power for the oscillator. The oscillator generates a square wave, F_{OSC} that is used for gate signalling at the N-MOSFET switch.

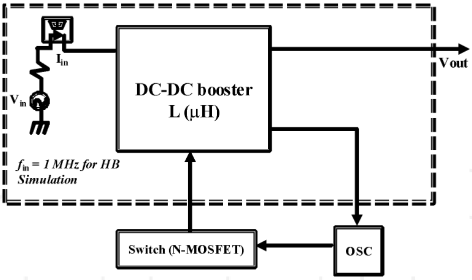

Fig. 20. Boost converter concept.

The drain signal of the N-MOSFET is used as the switch node voltage, V_{sn} at the anode of the diode inside the boost converter circuit block. From the concept presented in Fig. 20, the actual implemented circuit is shown in Fig. 21. The circuit was designed in Agilent's ADS and fabricated for investigation by measurement.

The circuit in Fig. 21 is proposed for investigation. Since a DC-DC boost converter is supposed to connect to the rectenna's output, it therefore, becomes the load to the rectenna circuit. This condition demands that the input impedance of the boost converter circuit emulates the known optimum load of the rectenna circuit. This has the benefit of ensuring maximum power transfer and hence higher overall conversion efficiency from the rectenna input (RF power) to the boost converter output (DC power). In this investigation, as shown in [7], the optimum load for the rectenna is around 2kΩ. In general, emulation resistance R_{em} is given by

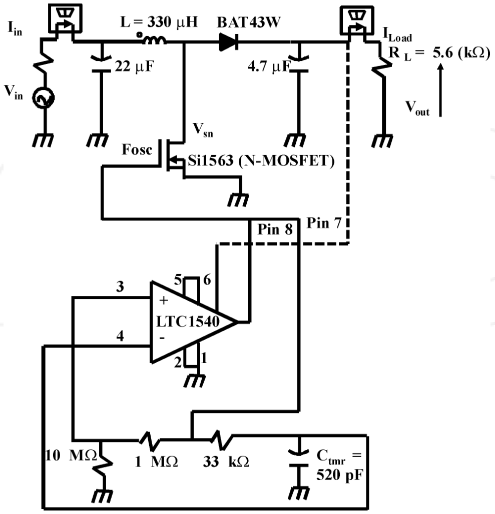

Fig. 21. The proposed boost converter circuit diagram. Designed in Agilent's ADS and fabricated for investigation by measurement.

$$R_{em} = \frac{2LT}{t_1^2 k}\left(\frac{M-1}{M}\right) \tag{7}$$

where ***L*** is the inductance equal to 330µH as shown in Fig. 20, $M = \frac{V_{out}}{V_{in}}$, ***T*** is the period of F_{OSC}, $\boldsymbol{t_1}$ is the switch"ON" time for the N-MOSFET, and ***k*** is a constant that according to [3] is a low frequency pulse duty cycle if the boost converter is run in a pulsed mode and typically, ***k*** may assume values like 0.06 or 0.0483. With reference to (7), we select L as the key parameter for higher conversion efficiency while V_{in} = 0.4 V DC is selected as the lowest start up voltage to achieve oscillations and boost operation. Computing the DC-DC boost conversion efficiency against different values of L, we have results as shown in Fig. 22.

From the results above, L = 100µH is the optimum boost inductance that ensures at least 16.5% DC-DC conversion efficiency, given R_L = 5.6kΩ.

Now having selected the optimum boost inductance given some load resistance, the emulation resistance shown in Fig. 23 is evaluated from the ratio of voltage versus current at the boost converter circuit's input.

The results show a constant resistance value against varying inductance. In general, we can say that this boost converter circuit has a constant low input impedance around 82.5Ω. This impedance is too small to match with the optimum rectenna load at 2kΩ. This directly affects the overall RF-to-DC conversion efficiency.

The results show a constant resistance value against varying inductance. In general, we can say that this boost converter circuit has a constant low input impedance around 82.5Ω. This impedance is too small to match with the optimum rectenna load at 2kΩ. This directly affects the overall RF-to-DC conversion efficiency.

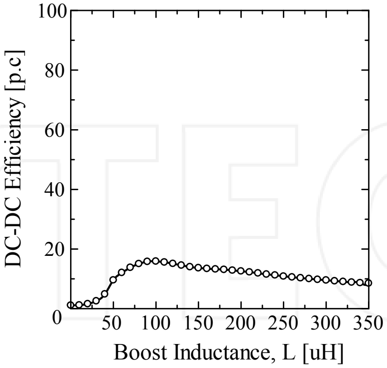

Fig. 22. Boost inductance variation with DC-DC conversion efficiency for a 5.6 kΩ load.

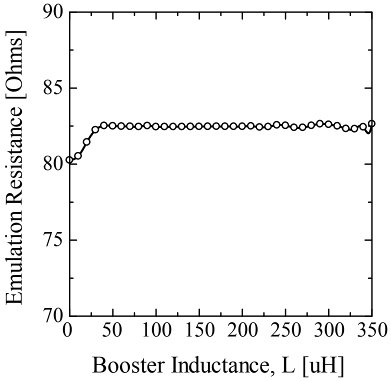

Fig. 23. Boost converter's input impedance: the emulation resistance.

Another factor, which affects the overall conversion efficiency is the power lost in the oscillator circuit. Unlike the circuit proposed in [9], which uses two oscillators; a low frequency (LF) and high frequency (HF) oscillator; in Fig. 21, we have attempted to use a single oscillator based on the LTC1540 comparator, externally biased as an astable multivibrator.

The power loss in this oscillator is the difference in the DC power measured at Pin 7 (supply) to the power measured at pin 8 (output). We term this loss, L_{osc}; converted to heat or sinks through the 10MΩ load. A comparison of the oscillator power loss to the power available at the boost converter output is shown in Fig. 24.

Looking at Fig. 24; we notice that the power loss depends on whether the oscillator output is high or low. The low loss corresponds to the quiescent period where the power lost is almost negligible. However, during the active state, the lost power (power consumed by the oscillator) nearly approaches the DC power available at the boost converter output. This results in low operational efficiency.

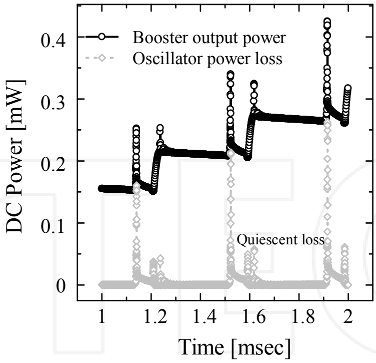

Fig. 24. The power loss in the oscillator.

To confirm whether or not the circuit of Fig. 21 works well, we did some measurements and compared them with the calculated results. Unlike in calculation (simulation), during measurement, L = 330μH was used due to availability. All the other component values remain the same both in calculation and measurement. In Fig. 25 (left side graph) and (right side graph), we see in general that the input voltage is boosted and also that the patterns of F_{osc} and V_{sn} are comparable both by simulation and measurement. To control the duty cycle of the oscillator output (F_{osc}), and the level of ripples in the boost converter output voltage (V_{out}), we change the value of the timing capacitance, C_{tmr} in the circuit of Fig. 21. Simulations in Fig. 25 (left side graph) show that C_{tmr} = 520pF realizes a better performance i.e. nearly constant V_{out} level (very low ripple).

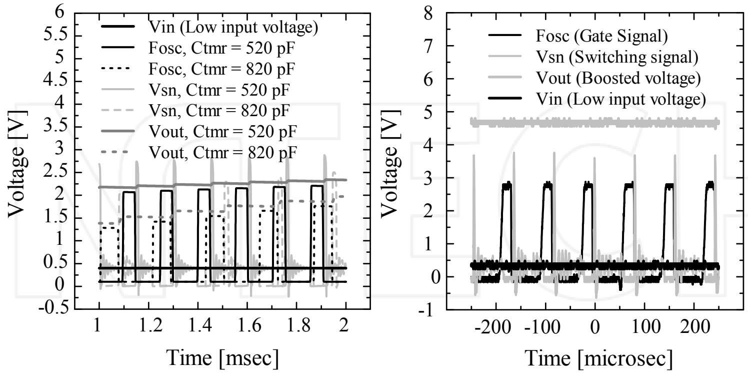

Fig. 25. Voltage characteristics of the developed boost converter circuit. The left side graph represents simulation while the right side graph is for measurements.

Generally, we observe that with this kind of boost converter circuit topology, it is difficult to start up for voltages as low as 61.7mV DC generated by the rectenna at -20dBm power incidence and at least 18.2% rectenna RF-to-DC conversion efficiency. Self starting is the issue for this topology at very low voltages.

At least 11.3% DC-DC conversion efficiency was recorded by measurement and is comparable to the calculation in Fig. 22. During measurement it was clearly revealed that the boost converter efficiency does depend on the value of **L** and the duty cycle derived from $\mathbf{t_1}$. To efficiently simulate the complete circuit, from the RF input to the DC output, envelope transient simulation (ENV) in Agilent's ADS was used. The (ENV) tool is much more computationally efficient than transient simulation (Tran). This simulation is appropriate for the boost converter circuit's resistor emulation task. Moreover, the boost converter's DC-DC conversion efficiency, and the overall RF-to-DC conversion efficiency can be calculated at once with a single envelope transient simulation.

In summary, though not capable to operate for voltages as low as 61.7mV DC, the proposed boost converter has by simulation and measurement demonstrated the capability to boost voltages as low as 400mV DC, sufficient for battery or capacitor recharging, assuming that the battery or the capacitor has some initial charge or energy enough to provide start-up to the boost converter circuit.

The limitations of our proposed boost converter circuit include; low efficiency, lack of self starting at ultra low input voltages, and unregulated output. To address these limitations, circuit optimization is required. Moreover, alternative approaches which employ a flyback transformer to replace the boost converter inductance must be investigated. A regulator circuit with Low Drop Out (LDO) is necessary to fix the boost converter output voltage commensurate with standard values like 2.2 V DC for example. For further reading, see [7]

2. Performance analysis of the complete RF energy harvesting sensor system

To demonstrate how one may analyze the performance of an RF energy harvesting system including its application, we extend the discussion of Section 2.5.2 to this Section. We propose a transmitter assembled as in Fig. 26 for temperature sensor wireless data transmission.

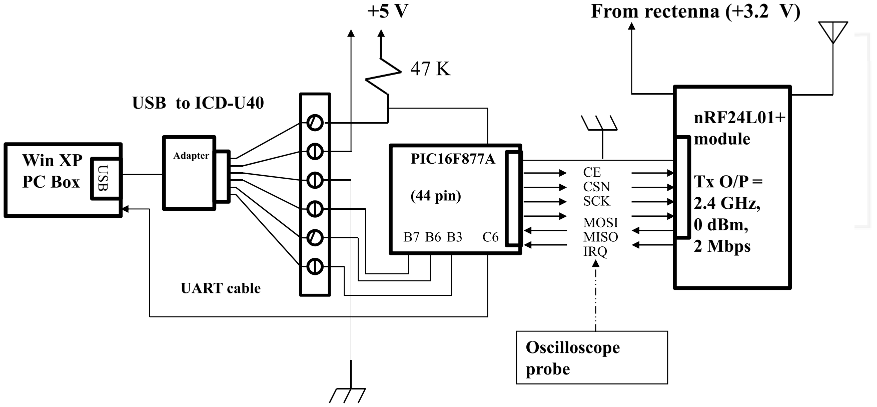

Fig. 26. The assembly and test platform for the proposed battery-free sensor transmitter.

The transmitter consists of one-chip microcomputer (MCU) PIC16F877A and wireless module nRF24L01P for the control, and MCU can be connected with an outside personal computer using ICD-U40 or RS232 cable. The wireless module operates in transmission and reception mode, and controls power supply on-off, transmitting power level, the receiving mode status, and transmission data rate via Serial Peripheral Interface (SPI). Figure 27 shows the operation flow when transmitting.

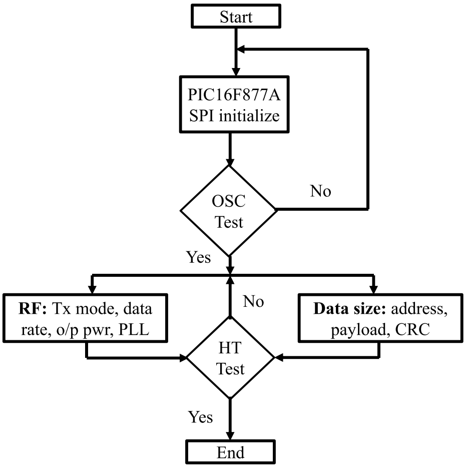
Start
PIC16F877A SPI initialize
OSC Test
No
Yes
RF: Tx mode, data rate, o/p pwr, PLL
No
Data size: address, payload, CRC
HT Test
Yes
End

Fig. 27. Operation flow during transmission.

The experimental system composition is shown in Fig. 28 to transmit acquired data by the temperature sensor with WLAN at 2.4 GHz (ISM band). An ISM band sleeve antenna is used for the transmission. Using the cellular band rectenna shown and discussed in Section 2.5.1, at least 3.14 V is stored in the electric double layer capacitor over a period of four hours. To harvest a maximum usable power for the overall system, we charge the capacitor up to 5V. The operation voltage for the wireless module presented in Fig. 26 above is between 1.9V and 3.6V.

The signal was transmitted from the wireless module while a sleeve antenna, same like the one for transmission was used with the spectrum analyzer and the reception experiment was performed. Received signal level equal to -43.4dBm was obtained at a distance 3.5m between transmitter and reception point. The capacitor’s stored voltage was used to supply the wireless module in the above-mentioned experiment. Successful transmission was possible for 5.5 minutes after which, the capacitor terminal voltage decreased from 3.16V to 1.47V, and the transmission ended. The sending and receiving distance of data can be estimated to be about 10m when the sensitivity of the receiver is assumed to be -60dBm, given 0dBm maximum transmit power.

Hereafter, the overall system examination is done by environmental power generation using the transmitted electric waves from the cellular phone base station, proposed based on the above-mentioned results. First of all, the power consumption shown in Fig. 29 is based on the fact that 120mW (5V, 24mA) is saved in the electric double layer capacitor by environmental power generation, achieved by calculation as discussed earlier.

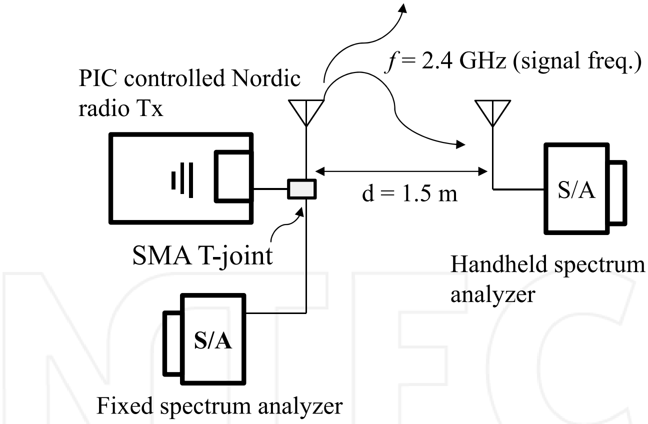
PIC controlled Nordic radio Tx
f = 2.4 GHz (signal freq.)
d = 1.5 m
S/A
SMA T-joint
Handheld spectrum analyzer
S/A
Fixed spectrum analyzer

Fig. 28. Indoor measurement setup for received traffic from the sensor radio transmitter.

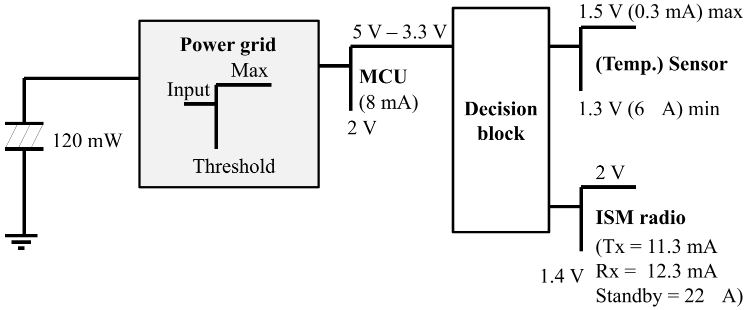
120 mW
Power grid
Max
Input
Threshold
5 V – 3.3 V
MCU
(8 mA)
2 V
Decision block
1.5 V (0.3 mA) max
(Temp.) Sensor
1.3 V (6   A) min
2 V
ISM radio
(Tx = 11.3 mA
Rx = 12.3 mA
Standby = 22   A)
1.4 V

Fig. 29. Power management scheme for the cellular energy-harvesting sensor node.

The sensor data packet is transmitted wirelessly in ShockBurst mode for energy efficient communication. The data packet format includes a pre-amble (1 byte), address (3 bytes), and the payload i.e. temperature data (1 byte). The flag bit is disregarded for easiness, and cyclic redundancy check (CRC) is not used.

The operation of the proposed system is provisionally calculated. When the rectenna is set up in the place where power incidence of 0dBm is obtained in the base station neighbourhood (as depicted in Section 2.5.2), an initially discharged capacitor accumulates up to 3.3V by a rectenna with 53.8% conversion efficiency (presented in Section 2.5.1). At this point, it takes 1.5 minutes to start and to initialize a wireless module, and the voltage of the capacitor decreases to 2V. This trial calculation method depends on the capacitor's back up time discussed in [8]. After this, when the wireless module is assumed to be in sleep mode, the capacitor is charged by a 0.28mA charging current for four hours whereby the capacitor's stored voltage increases up to 5V. The power consumption in the sleep mode or standby is 33µW (1.5V, 22µA).

When the wireless module starts, after data transmission and the confirmation signal is sent, the voltage of the capacitor decreases by 0.6V, and consumes the electric power of 7.4mW.

INTECH

The voltage of the capacitor decreases to 2V when 3.2mW is consumed to the acquisition of the sensor data, and the operation time of MCU is assumed to be one minute to the data storage in the wireless module etc. As for the capacitor voltage, when the wireless module continuously transmits data for 20 seconds, it decreases from 2V to 1.4V and even the following operation saves the electric power. Therefore, a temperature sensing system capable of transmitting wireless data in every four hours becomes feasible by environmental power generation from the cellular phone base station if we consider intermittent operation by sleep mode.

3. Conclusion

This Chapter has given an overview of the present energy harvesting sources, but the focus has stayed on RF energy sources and future directions for research. Design issues in RF energy harvesting have been discussed, which include low conversion efficiency and sometimes low rectified power. Solutions have been suggested by calculation and validated by measurement where possible, while highlighting the limitations of the proposed solutions. Potential applications for both DTV and cellular RF energy harvesting have been proposed and demonstrated with simple examples. A discussion is also presented on the typical performance analysis for the proposed RF energy harvesting system with sensor application.

4. Acknowledgment

The authors would like to thank Prof. Apostolos Georgiadis of Centre Tecnològic de Telecomunicacions de Catalunya (CTTC, Spain) for the collaboration on the design and development of the DC-DC boost converter circuit. Further thanks go to all those readers who will find this Chapter useful in one way or the other.

5. References

[1] Keisuke, T.; Kawahara, Y. & Asami, T. (2009). *RF Energy Intensity Survey in Tokyo..,* (c)2009 IEICE, B-20-3, Matsuyama-shi, Japan

[2] Mikeka, C.; Arai, H. (2011). *Dual-Band RF Energy-Harvesting Circuit for Range Enhancement in Passive Tags,* (c)2011 EuCAP, Rome, Italy

[3] Pozar, D. (2005). *Microwave Engineering,* Wiley, ISBN 978-0-471-44878-5, Amherst, MA, USA

[4] Mikeka, C.; Arai, H. (2010). Techniques for the Development of a Highly Efficient Rectenna for the Next Generation Batteryless System Applications, *IEICE Tech. Rep.,* pp. 101-106, Kyoto, Japan, March, 2010

[5] http://www.secomtel.com/UpFilesPDF/PDF/Agilent/PDF_DOCS/SKYDIODE/ 03_SKYDI/HSMS2850.PDF (Last accessed on 13 July, 2011)

[6] McSpadden, J. et al., H. (1992). Theoretical and Experimental Investigation of a Rectenna Element for Microwave Power Transmission, *IEEE Trans., on Microwave Theory and Tech., Vol. 40, No. 12.,* pp. 2359-2366, Dec., 1992

[7] Mikeka, C.; Arai, H. ; Georgiadis A. ; and Collado A. (2011). DTV Band Micropower RF Energy-Harvesting Circuit Architecture and Performance Analysis, *RFID-TA Digest,* Sitges, Spain, Sept., 2011

[8] Mikeka, C.; Arai, H. (2009). Design of a Cellular Energy-Harvesting Radio, *Proc. 2 nd European Wireless Technology Conf.,* pp. 73-76, Rome, Italy, Sept., 2009

[9] Popovic Z., et al., (2008). Resistor Emulation Approach to Low-Power RF Energy Harvesting, *IEEE Trans. Power Electronics, Vol. 23, No. 3*, 2008

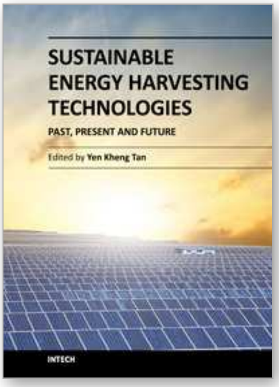

Sustainable Energy Harvesting Technologies - Past, Present and Future

Edited by Dr. Yen Kheng Tan

ISBN 978-953-307-438-2
Hard cover, 256 pages
Publisher InTech
Published online 22, December, 2011
Published in print edition December, 2011

In the early 21st century, research and development of sustainable energy harvesting (EH) technologies have started. Since then, many EH technologies have evolved, advanced and even been successfully developed into hardware prototypes for sustaining the operational lifetime of low?power electronic devices like mobile gadgets, smart wireless sensor networks, etc. Energy harvesting is a technology that harvests freely available renewable energy from the ambient environment to recharge or put used energy back into the energy storage devices without the hassle of disrupting or even discontinuing the normal operation of the specific application. With the prior knowledge and experience developed over a decade ago, progress of sustainable EH technologies research is still intact and ongoing. EH technologies are starting to mature and strong synergies are formulating with dedicate application areas. To move forward, now would be a good time to setup a review and brainstorm session to evaluate the past, investigate and think through the present and understand and plan for the future sustainable energy harvesting technologies.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Chomora Mikeka and Hiroyuki Arai (2011). Design Issues in Radio Frequency Energy Harvesting System, Sustainable Energy Harvesting Technologies - Past, Present and Future, Dr. Yen Kheng Tan (Ed.), ISBN: 978-953-307-438-2, InTech, Available from: http://www.intechopen.com/books/sustainable-energy-harvesting-technologies-past-present-and-future/design-issues-in-radio-frequency-energy-harvesting-system

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

ELECTRICAL AND ELECTRONIC ENGINEERING

ELECTRONIC AND SPINTRONIC DEVICES

Received
4 September 2014

Accepted
19 February 2015

Published
7 May 2015

Correspondence and requests for materials should be addressed to N.S. (negin.shariati@rmit.edu.au)

Multi-Service Highly Sensitive Rectifier for Enhanced RF Energy Scavenging

Negin Shariati, Wayne S. T. Rowe, James R. Scott & Kamran Ghorbani

School of Electrical and Computer Engineering, RMIT University, Melbourne, VIC 3001, Australia.

Due to the growing implications of energy costs and carbon footprints, the need to adopt inexpensive, green energy harvesting strategies are of paramount importance for the long-term conservation of the environment and the global economy. To address this, the feasibility of harvesting low power density ambient RF energy simultaneously from multiple sources is examined. A high efficiency multi-resonant rectifier is proposed, which operates at two frequency bands (478–496 and 852–869 MHz) and exhibits favorable impedance matching over a broad input power range (−40 to −10 dBm). Simulation and experimental results of input reflection coefficient and rectified output power are in excellent agreement, demonstrating the usefulness of this innovative low-power rectification technique. Measurement results indicate an effective efficiency of 54.3%, and an output DC voltage of 772.8 mV is achieved for a multi-tone input power of −10 dBm. Furthermore, the measured output DC power from harvesting RF energy from multiple services concurrently exhibits a 3.14 and 7.24 fold increase over single frequency rectification at 490 and 860 MHz respectively. Therefore, the proposed multi-service highly sensitive rectifier is a promising technique for providing a sustainable energy source for low power applications in urban environments.

AMBIENT energy harvesting is attracting widespread interest as it has the potential to provide a sustainable energy source for future growth and protection of the environment. Considerable research effort has been directed toward low-profile, low-power, energy efficient and self-sustainable devices aiming to harvest energy from inexhaustible sources such as solar energy, thermal, biomass, mechanical sources (e.g. wind, kinetic, vibration, and ocean waves) wastewater, and microwave energy. A thorough set of reviews is given in the literature[1–4]. Among these green energy sources, there has been a growing interest for radio frequency (RF) energy scavenging, as the availability of ambient RF energy has increased due to advancements in broadcasting and wireless communication systems. Furthermore, the development of wireless power transmission (WPT) technologies[5] that allow micro sensors[6], mobile electronic devices[7], wireless implantable neural interfaces[8] and far-field passive RFID (Radio-Frequency Identification) systems[9–11] to operate without batteries has triggered impetus for RF energy harvesting.

Efficient RF energy harvesting is a very challenging issue, as it deals with the very low RF power levels available in the environment. Furthermore, the scavengeable power level can vary unpredictably, depending on several factors such as the distance from the power source, the transmission media, the telecommunication traffic density and the antenna orientation. The majority of available literature on RF rectification has been dedicated to narrowband rectennas, which essentially operate at a single frequency and hence provide low DC output power[12,13]. Various topologies, such as voltage doublers or multipliers have been employed in order to increase the RF to DC conversion efficiency and the output DC voltage for specific applications[14–16]. However, from an ambient RF scavenging perspective, harvesting energy from various available frequencies could maximize power collection and hence increase the output DC power. Ultra-wideband and broadband rectenna arrays have been proposed as a potential solution[17,18]. However in some cases, simulation and experimental results were not provided to demonstrate the findings[17]. A broadband rectenna consisting of a dual-circularly polarized spiral rectenna array operating over a frequency range of 2–18 GHz was demonstrated[18]. The rectified DC power was characterized as a function of DC load, RF frequency and polarization for power densities between 10^{-5} and 10^{-1} mW/cm^2. However, the proposed rectenna was matched at a single input RF power level for a specified load resistance for the characterization. Also, due to the low Q value of the rectifier circuit, the conversion efficiency was a fraction of 1% at −15.5 dBm. From a design point of view, while it is relatively easy to achieve a broadband antenna, it is very challenging to realize a broadband rectenna due to the non-linearity of the rectifier impedance with input power across the frequency band[19].

To address this, a promising approach is to use a dual-band or multi-band configuration. This can maximize the power conversion efficiency (PCE) at the specific frequencies where the maximum ambient signal level is available. Various dual-band RF energy harvesting systems has been demonstrated[20–24], however a large signal analysis of the rectifier was commonly not provided over a broad input power range. A dual-band RF energy harvesting using frequency limited dual-band impedance matching has been proposed[20] and the PCE was shown over a high power range of 0 to 160 mW, however it was only matched at a single input power level (10 dBm). A CMOS dual-narrowband energy harvester circuit was modeled at environmental power levels[21]. Again, the rectifier efficiency was demonstrated with only single input power levels of −19 and −19.3 dBm at 2 GHz and 900 MHz respectively, and a large signal analysis was not presented. A compact dual-band rectenna operating at 915 MHz and 2.45 GHz has been demonstrated[22] and the PCE was shown for input power levels of −15, −9 and −3 dBm. However, the reflection coefficient was evaluated at a single incident power level. Furthermore, the efficiency results with dual-tone excitation simultaneously and single-tone excitation (at 915 MHz) are very similar, hence the impact of applying a dual-band technique does not demonstrate a clear advantage over a single band. A dual-frequency rectenna for WPT has been proposed[23] which achieved a conversion efficiency of 84.4% and 82.7% at 2.45 and 5.8 GHz with a high input power level of 89.84 and 49.09 mW respectively. These power levels far exceed ambient levels in the environment[19]. A conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna has been presented[24] and the PCE was provided with −30 to 5 dBm input power, achieving an efficiency up to 40% at 1.85 GHz for higher input power levels (above −5 dBm). However the reflection coefficient was not provided at low input power range.

A multi-resonant rectenna that uses a multi-layer antenna and rectifier has been evaluated for a −16 dBm to +8 dBm RF received power level, but the rectifier circuit layout and large signal analysis were not provided to clarify the findings[25]. Furthermore, a rectenna for triple-band biotelemetry communications has been proposed using a triple-band antenna and single frequency rectifier[26]. However, this rectenna is not suitable for RF energy scavenging due to the low efficiency at lower input power levels. Another triple band rectenna presented an RF-DC efficiency over the input power range of −14 to +20 dBm[27], however the reflection coefficient results were only evaluated at a single input power level. This rectenna was shown to harvest 7.06 μW of DC power from three sources simultaneously at a high input power level of +10 dBm. A multi-band harvesting system has also been proposed where four individual harvesters are designed to cover four frequency bands[28]. However, a large signal analysis was not provided over a broad input power range. Furthermore, the proposed harvesting system has a minimum sensitivity of −25 dBm, whilst in a real environment more sensitive systems are required as the available RF power levels are very low[19].

Tunable impedance matching networks have been demonstrated in order to collect RF signals from various sources and convert them to DC power[29]. However from an application point of view, this is still single frequency rectification and it is not widely applicable to environmental RF energy scavenging where the available power is very low.

In order to increase the amount of RF energy scavenged by a rectenna, it is crucial to identify and harvest multiple ambient frequency sources over their realistic available energy range. Our previous research has demonstrated the feasibility of RF energy harvesting through RF field investigations and maximum available power analysis in metropolitan areas of Melbourne, Australia[19]. The maximum available power for different frequency bands based on antenna aperture and number of antennas in a given collection area was analyzed. Measured results and analysis indicated that cellular systems and broadcast sources are well suited to harvesting, with scavengeable RF power ranging from −40 to −10 dBm. This identifies two important considerations in the design of efficient rectenna for RF energy harvesting: the scavengeable ambient RF power sources available, and the significant variance of this power.

The RF to DC rectifier solutions proposed in recent literature have focused on maximizing the system efficiency at a given, and often quite high, input power level. This neglects the issues related to input power variation which can lead to unexpected variations in the matching network due to diode non-linearity. Also, the scavangeable levels of ambient RF power have been shown to be orders of magnitude lower. Therefore, based on our previous research outcomes and recommendations[19], an efficient power harvesting solution could encompass a multi-band matching circuit at the specific frequencies where maximum signal power is available, enabling greater power harvesting due to the combination of RF signals. This also results in a higher power being fed to a single rectifier, utilizing the diode function more efficiently.

This paper presents an RF energy harvesting method that can scavenge a wide range of ambient power levels which are orders of magnitude lower than previous reported techniques in the literature. An efficient dual resonant rectifier circuit is proposed, matched to a 50 Ω input port at 490 and 860 MHz over a broad low input RF power range from −40 to −10 dBm. The proposed dual resonant matching network operates efficiently at two identified harvesting frequency bands over a wide input power range, maximizing DC power by scavenging two sources simultaneously.

The remainder of this paper is organized as follows. First, the key results for the reflection coefficient and output DC power are presented. Subsequently, the Discussion section summarizes the results and demonstrates their potential implications, the limitations of this study, open questions and future research. Finally, the Method section describes the proposed rectifier design.

Results

A dual resonant rectifier was fabricated on a 1.58 mm FR-4 substrate with a dielectric constant $\varepsilon_r \approx 4.5$ and a loss tangent $\delta \approx 0.025$. These substrate parameters were measured using the Nicolson-Ross method[30] so accurate values could be used in the rectifier design. A photograph of the fabricated dual resonant rectifier is shown in Fig. 1 which depicts input RF port, dual-band matching network lumped components, Schottky diodes and the output terminal. The performance of the rectifier was verified by measuring the input reflection properties, and the output power was calculated from the measured output DC voltage for the input powers from −40 to −10 dBm.

Reflection Coefficient. The $|S_{11}|$ of the rectifier was evaluated using a vector network analyzer (VNA). The VNA was re-calibrated for each input power level. Figure 2 compares the simulated and measured $|S_{11}|$ versus frequency for the dual resonant rectifier circuit at four different input power levels from −40 to −10 dBm. The measured results show very good agreement to the simulations. Slightly higher reflection was observed for the resonant frequencies at the lower part of the input power range (due to the diode characteristics). However, the proposed rectifier circuit is well-matched ($|S_{11}| < -10$ dB) at the desired frequency bands of

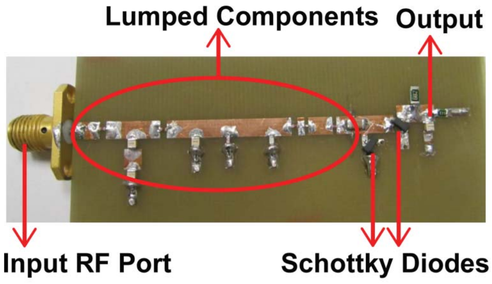

Figure 1 | **Fabricated rectifier prototype.**

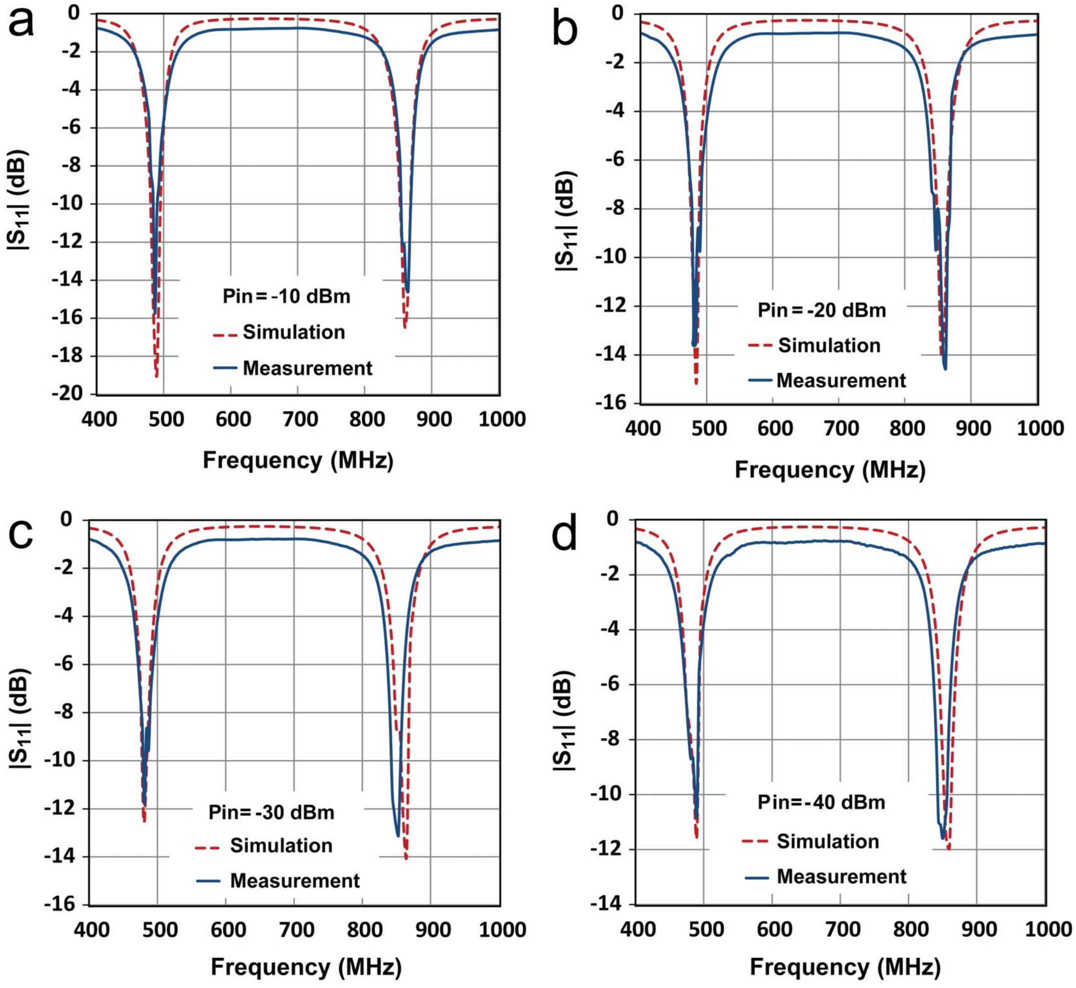

Figure 2 | Simulated and measured $|S_{11}|$ as a function of frequency and input RF power for the proposed dual resonant rectifier circuit. (a) −10 dBm. (b) −20 dBm. (c) −30 dBm. (d) −40 dBm.

478–496 MHz and 852–869 MHz over the broad range of input powers from −40 to −10 dBm. The small difference between simulation and measurement is due to the parasitic extraction accuracy.

Output DC Power. In the frequency domain, the Harmonic Balance method of analysis provides a comprehensive treatment of a multispectral problem[18]. The method intrinsically takes into account the DC component and a specified number of harmonics, while allowing the ability to specify the source impedance and harmonic terminations. A Harmonic Balance simulation was used to numerically evaluate the output DC voltage of the dual resonant rectifier for both a single and two tone input. The output DC voltage across the load resistor was also measured and used to calculate output DC power. Measurements were performed using a Wiltron 68247B synthesized signal generator as a RF power source for the rectifier circuit. Recording of the output DC voltage across the load resistance was achieved with a Fluke 79III digital voltage meter. The RF source power was initially set at −10 dBm, and decreased in 2 dB steps. In the dual-band measurement case, two RF signal generators were fed to the rectifier circuit simultaneously via a power combiner.

The simulation and measurement results for single and dual input tones are summarized in Fig. 3(a) and (b). A measured DC voltage of 772.8 mV is achieved with two simultaneous input tones at an input power of −10 dBm. For single tone measurements, DC voltages of 436 mV and 286 mV at 490 MHz and 860 MHz respectively are produced. The comparison between the 490 and 860 MHz single rectifiers highlights the impact of the input frequency on the PCE. A higher amount of DC voltage can be generated at the lower frequency. This difference comes from decreasing diode performance at the higher frequency due to the higher junction capacitance of the diode[31].

Importantly, a slightly higher DC voltage can be generated with the dual resonant rectifier as compared to the sum of output voltage from the two single bands, particularly at the lower input power levels as can be seen in Fig. 3(b) which shows the lower power section of Fig. 3(a) in more detail. By maximizing power collection from various sources of different frequencies and delivering the combined power to the rectification circuit, the diode conversion efficiency is enhanced which results in a higher level of rectified voltage. Figure 4 compares the simulated and measured output DC power for the dual resonant rectifier circuit with both single and dual input tones. A measured DC power of 17.3 μW and 7.5 μW can be generated at 490 MHz and 860 MHz respectively with a single tone input of −10 dBm (100 μW). This represents true efficiencies of 17.3% and 7.5% for the individual single band rectification (see Fig. 5). However, the measured DC output power with two concurrent input tones of −10 dBm is 54.3 μW which corresponds to an effective efficiency of 54.3% for the dual-band rectifier (see Fig. 6). This represents a 3.14 and 7.24 times increase in output DC power over the single tone excitation at 490 MHz or 860 MHz respectively. This trend is evident down to low input power levels (around 40 μW). Furthermore, there is a significant increase in the PCE of the dual resonant rectifier for lower input power levels (<40 μW).

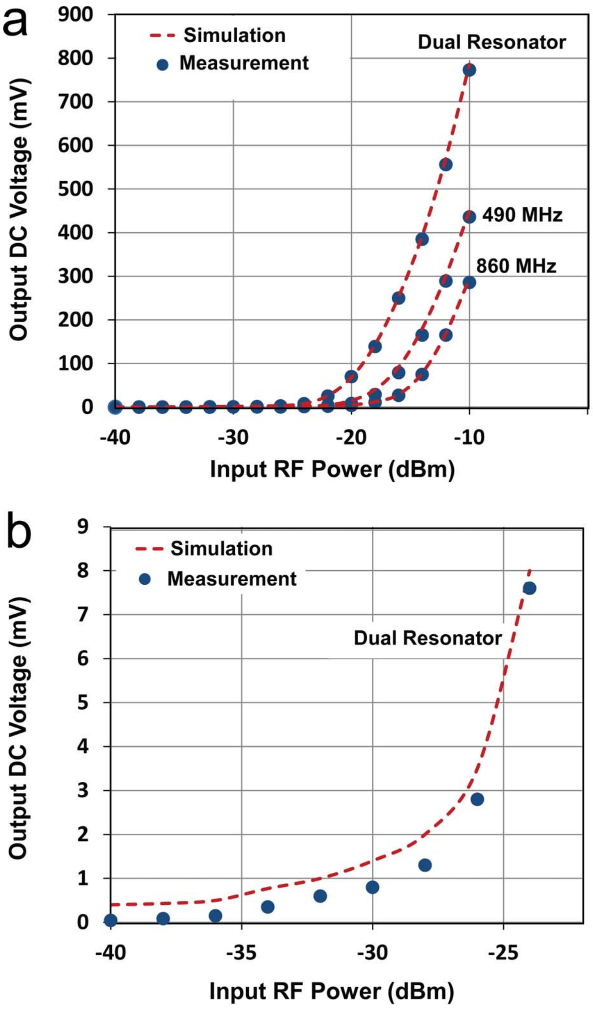

Figure 3 | Output DC voltage as a function of input RF power for single input tone at both 490 MHz and 860 MHz and for dual input tones (a) with −40 to −10 dBm input RF power (b) with −40 to −25 dBm input RF power. (This power range is associated with the signal source).

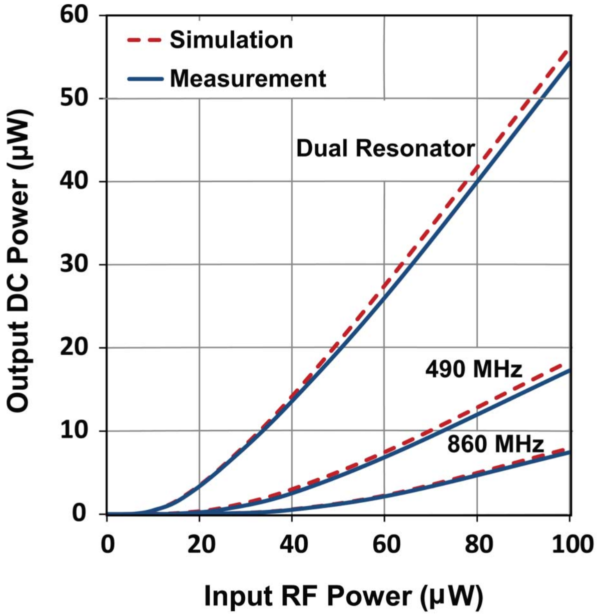

Figure 4 | Output DC power as a function of input RF power for single and dual input tones.

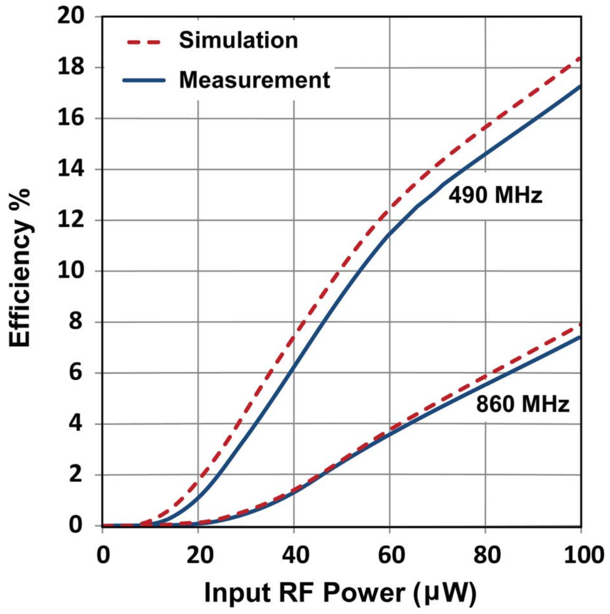

Figure 5 | RF to DC conversion efficiency as a function of input RF power for single band rectification.

Here, the effective efficiency is defined as the ratio of output DC power to the available input RF power rather than the power delivered to the diodes (equation (1)). The available power level is associated with the signal source. For the single resonator, the available input power is −10 dBm and the delivered power is also −10 dBm (assuming no loss). However, by creating a dual resonant matching network the power delivered to the diodes is −7 dBm (combined total input power from two signal generators) but the available power is still −10 dBm.

$$\eta = \frac{p_o}{p_i} = \frac{\left(\frac{V_o^2}{R_L}\right)}{p_i} \qquad (1)$$

Therefore, combining input RF signals into a single rectification stage results in high sensitivity rectifier, which is widely applicable to

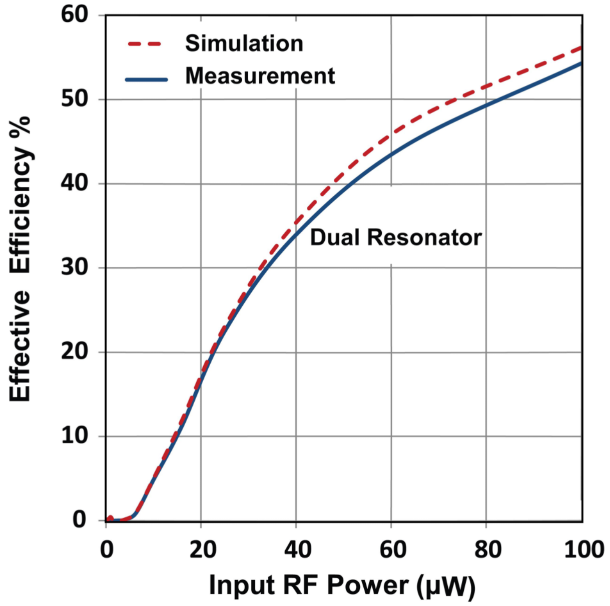

Figure 6 | Effective RF to DC conversion efficiency as a function of input RF power for dual resonant rectification.

Table 1 | Rectifier performance comparison

Ref.	Technology	Measured Efficiency (%)	RF power variation (in PCE evaluation)	Rectification Technique
13	Schottky diode	82@50 mW	N/A	Single resonator
16	Schottky diode	44@−10 dBm	N/A	Single resonator
18	Schottky diode	20@0.07 mW/cm^2 0.1@5×10^{-5} mW/cm^2	10^{-5} to10^{-1} mW/cm^2	Broad band
20	Schottky diode	77.13@22 dBm (158.49 mW)	0 to 160 mW	Dual resonator
21	CMOS	9.1@−19.3 dBm (900 MHz) 8.9@−19 dBm (2 GHz)	N/A	Dual resonator
22	Schottky diode	37 (915 MHz)@−9 dBm 30 (2.45 MHz)@−9 dBm	−40 to 0 dBm	Dual resonator
23	Schottky diode	84.4@89.84 mW (2.45 GHz) 82.7@49.09 mW (5.8 GHz)	0 to 100 mW	Dual resonator
27	Schottky diode	80@10 dBm (940 MHz) 47@8 dBm (1.95 GHz) 43@16 dBm (2.44 GHz)	−14 to 20 dBm	Triple resonator
29	Schottky diode	50@−5 dBm	−25 to 0 dBm	Dual resonator with tunable input response
This work	Schottky diode	54.3@−10 dBm 11.25@−18 dBm (490 and 860 MHz)	−40 to −10 dBm	Dual resonator

Table 2 | Environmental measurement results

Suburb	Available frequencies (MHz)	Respective available RF power (dBm) [μW]	Measured DC power (μW)
Bayswater	486, 488, 489, 490, 491, 867, 868, 869, 870, 871, 872, 873, 874	−19[12.5], −20[10], −17[19.95], −15[31.62], −22[6.3], −37[0.199], −37[0.199], −30[1], −24[3.98], −20[10], −30[1], −37[0.199], −40[0.1]	39.38
Bentleigh	491, 492, 494, 495, 865, 866, 867, 868, 869, 870, 871	−12[63.09], −46[0.02], −42[0.063], −57[0.001], −27[1.99], −27[1.99], −30[1], −37[0.199], −40[0.1], −40[0.1], −41[0.07]	30.9
RMIT University (Melbourne CBD)	487, 488, 489, 490, 491, 851, 861, 862, 866, 867, 868, 869	−30[1], −22[6.3], −29[1.25], −22[6.3], −20[10], −23[5.01], −21[7.94], −21[7.94], −30[1], −35[0.31], −40[0.1], −40[0.1]	14.5

real environmental RF energy scavenging. This multi-band technique can provide higher DC power than combining two separate single frequency rectifier circuits operating at the same frequencies. This is due to the fact that harvesting RF energy from various available sources simultaneously increases the delivered power to the rectifier, which improves the diode conversion efficiency and consequently enhances the output DC power. Table 1 summarizes this work as compared to previous published work.

In order to provide a realistic scenario for the proposed dual-band rectifier, measurement results were taken in three suburbs of Melbourne, Australia, congruent with our previous research outcomes[19]. Table 2 summarizes these environmental measurement results. It should be noted that the lower band (478–496 MHz) has a 3.67% fractional bandwidth and the higher band (852–869 MHz) has around 2% fractional bandwidth. Hence, various RF frequencies from different sources can be harvested within these two bands. The environmental measurement results demonstrate the feasibility of harvesting ambient EM energy from multiple sources simultaneously.

Discussion

The feasibility of harvesting ambient EM energy from multiple sources simultaneously is investigated in this paper. The proposed dual resonant rectifier operates at two frequency bands (478–496 and 852–869 MHz), which are used for broadcasting and cellular systems respectively. The dual resonant rectifier exhibits favorable impedance matching over a broad input power range (−40 to −10 dBm) at these two bands. The achieved sensitivity and dynamic range demonstrate the usefulness of this innovative low input power rectification technique. Simulation and experimental results of input reflection coefficient and rectified output power are in excellent agreement. The measurement results demonstrate that a two tone input to the proposed dual-band RF energy harvesting system can generate 3.14 and 7.24 times more power than a single tone at 490 or 860 MHz respectively, resulting in a measured effective efficiency of 54.3% for a dual-tone input power of −10 dBm. It is evident that this dual resonant rectification technique increases the RF to DC effective conversion efficiency, and hence the recoverable DC power for low power applications. Furthermore from a design and economic perspective, utilizing a large number of components (e.g. antennas, diodes) to realize individual rectifier circuits for each frequency band creates additional expense. In order to provide more realistic measurement results, the proposed dual-band rectifier was tested in three suburbs of Melbourne, Australia. Therefore, this dual-band technique offers a simple and cost-effective solution which is of paramount importance for environmental power harvesting systems. This innovative technique has the potential to generate a viable perpetual energy source for low power applications in urban environments.

Limitation of the study, open questions and future work

Utilizing diodes which are more suitable to low power applications ($P_i < -20$ dBm) could increase the voltage sensitivity, resulting in a higher RF-DC conversion efficiency[32]. Applying a power optimized waveform excitation to the rectifier circuit in these frequency bands, a higher amount of DC power can be generated when compared with a single and dual tone excitations with the same input power[33,34]. However this technique is not applicable to energy harvesting where the input waveform is arbitrary.

Utilizing our proposed dual resonant rectification technique to combine resonant circuits for any other arbitrary frequency bands could lead to PCE improvement, provided that suitable diodes for the desired frequency bands are selected. Note that by increasing the operating frequency the rectification performance degrades due to

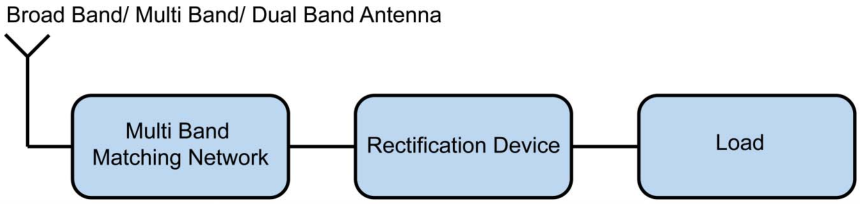
Broad Band/ Multi Band/ Dual Band Antenna
Multi Band Matching Network
Rectification Device
Load

Figure 7 | **General block diagram of the RF energy harvesting system.**

the higher junction capacitance of the diode. Hence, lower output voltage is expected at higher frequency bands.

It is the object of our future work to design a multi-band rectenna array for enhanced RF energy harvesting. Furthermore, increasing the bandwidth, sensitivity and efficiency will also be investigated.

Methods

The major goal in designing an efficient RF harvesting system is to produce high DC output power. Toward this goal, a high sensitivity rectifier is crucial for optimum RF scavenging. A significant factor governing the sensitivity of a rectifier is the threshold voltage of the diode used for rectification. The diode must be able to "switch on" for very low ambient energy levels.

To address this sensitivity issue, a system that scavenges power from multiple frequency bands and combines them to activate a rectification circuit is proposed. The general block diagram of the proposed system is depicted in Fig. 7. Various environmental RF energy sources of different frequencies are collected by an appropriately designed antenna, and delivered to the rectification circuit via a multi-band matching network. The rectification circuit converts the combination of RF signals into DC power for low-power applications. The embodiment in this paper realizes a dual resonant matching circuit as a transition between a 50 Ω nominal antenna output and the non-linear rectification device at 490 and 860 MHz. Based on the Australian Radiofrequency Spectrum Plan[35], these bands are allocated to broadcasting services and cellular systems.

Device Selection. Due to the very low ambient power available in a real environment[19], a very low threshold voltage rectification device is required in order to increase sensitivity. For this reason, Schottky diodes (GaAs or Si) are commonly employed for RF energy harvesting. In this work, a microwave Schottky detector HSMS2820 ($C_{j0} = 0.7$ pF, $R_s = 6\ \Omega$, $I_s = 2.2e^{-8}$ A) is chosen due to its excellent high frequency performance, low series resistance (R_s) and junction capacitance (C_j), and low threshold voltage with high-saturation current[31]. This low threshold voltage (0.15–0.3 V) supports rectification at low input power levels.

Proposed Rectifier Design. In order to design an efficient RF harvesting system, the non-linearity of the rectifier impedance with frequency and input power should be matched to the 50 Ω output of the antenna at the desired frequency bands. Therefore, the diode input impedance as a function of frequency and different power levels were calculated and analyzed[36]. In order to match the input impedance of the rectifier to the 50 Ω output of the antenna, the total load impedance for different input power and frequencies should be determined. A circuit consisting of a pair of Schottky Barrier Diodes (SBD) terminated with a load resistor ($R_{Load} = 11$ kΩ) and an output bypass capacitor ($C2 = 6.8$ pF) was simulated using Agilent ADS software. Figure 8 shows the proposed geometry of the voltage-doubler topology[31,37]. The voltage doubler rectifier structure is employed for the design of the RF-DC power conversion system as this topology is well suited to low power rectification. The resistor and capacitor at the output will filter high frequencies. The high load resistor (11 kΩ) was chosen to observe a reasonable output voltage at very low currents. Using Large Signal S-Parameters analysis in Agilent ADS software, the load impedance and bypass capacitor were determined and optimized.

The voltage doubler rectifier in Fig. 8 consists of a peak rectifier formed by *D2* and bypass capacitor *C2* (6.8 pF) and a voltage clamp formed by *D1* and *C1* (total capacitance of the transmission lines and diode's parasitic capacitance (C_p)). In the negative phase of the input, current flows through *D1* while *D2* is cutoff. The voltage across *D1* stays constant around its threshold voltage and the voltage at node 1 is charged to $-V_{th1}$ (where $-V_{th1}$ is the threshold voltage of the *D1*). At the negative peak, the voltage across *C1* is $V_{amp} - V_{th1}$ (where V_{amp} is the amplitude of the input signal). In the positive phase of the input, current flows through *D2* while *D1* is cutoff. The voltage across *C1* remains the same as the previous phase because it has no way to discharge. At the positive peak, the voltage across *D1* is $2V_{amp} - V_{th1}$. Since *D2* is conducting current to charge *C2*, the voltage at the output is $V_{out} = 2V_{amp} - V_{th1} - V_{th2}$.

The DC equivalent circuit of the SBD is a voltage source in series with the junction resistor R_j which is obtained by differentiating the diode voltage–current characteristic and is given by equation (2)[31,38]:

$$R_j = \frac{nKT}{q(I_s + I_b)} \qquad (2)$$

Where n is the diode ideality factor, *K* is the Boltzmann's constant, *T* is the temperature in degrees Kelvin, *q* is the electronic charge, I_s is the diode saturation current and I_b is the external bias current. At low power levels, the saturation current is very small ($I_s = 2.2\ e^{-8}$ A) and for a zero-biased diode, $I_b = 0$. Therefore, the resulting value of junction resistance at room temperature is approximately 1.7 MΩ. Since, the saturation current is highly temperature dependent, R_j will be even higher at lower temperatures which tends to decrease the output voltage. As the input power increases, some circulating rectified current will cause a drop in the value of R_j and this phenomenon will increase the value of the DC output voltage. Furthermore, it is worth to highlight that the rectified current produced by the first diode (*D1*) in Fig. 8 constitutes the external bias current of the second diode (*D2*) which will help to reduce the R_j and hence the detection sensitivity is improved. Therefore, depending on the amount of available bias current, R_j is varying (equation (2)), hence the matching network is changing which impacts the amount of delivered power to the diode and results in different values of PCE.

A Schottky barrier diode can be modeled by the linear equivalent circuit shown in Fig. 9, where L_p and C_p are the diode's parasitic inductance and capacitance respectively due to packaging ($L_p = 2$ nH and $C_p = 0.08$ pF) which are generally unwanted[39]. This linear model is used for determining the diode impedance at a given input power.

The diode impedance analyzed using a Harmonic Balance simulator and a non-linear model of the diodes over the frequency range of 400 to 900 MHz at various input power levels (Fig. 10). Due to the large junction resistor at low input RF power levels, the rectification device is turned off in absence of an appropriate matching network. Large Signal S-parameter analysis was conducted and higher input power (associated with the signal source) is applied directly to the Schottky diodes configuration of Fig. 8 which does not include a matching network in order to turn on the diodes (reduce the value of R_j) and extract approximate input impedance value as our starting point in design of a matching network. As it can be seen in Fig. 10, with increasing the source power, the diode impedance is varying and it is beginning to switch on. Hence, the input impedance needs to be determined when the diode is

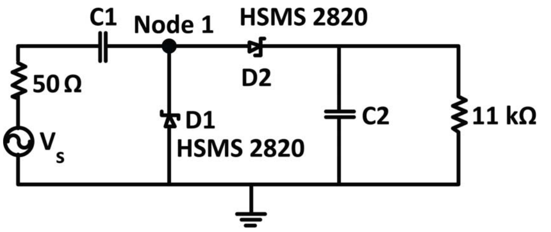

Figure 8 | **Schematic of a voltage-double rectifier without matching network.**

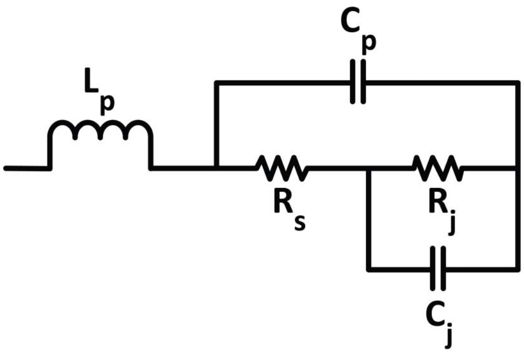

Figure 9 | **HSMS 2820 Schottky diode equivalent circuit.**

turned on to realize the matching network for a rectifier circuit. Obviously, in the presence of an appropriate matching network the rectification device can be turned on at lower power levels, whilst in the absence of a matching network a higher input power should be applied to switch on the diode. (Note that, with an unmatched rectifier the total applied input power from the signal source cannot be delivered to the diode due to the high reflection in the circuit).

The aim is to match the input impedance of the device to 50 Ω at 478–496 MHz and 852–869 MHz bands over a broad range of input RF powers. The procedure commences by matching the diode input impedance at high unmatched source power and shifting the diode input impedance at various power levels to within the voltage standing wave ratio (VSWR) <2 circle on the Smith chart. This procedure assumes that diode input impedance does not drastically change in this low power range. The simulation results of Fig. 10 prove that this is the case.

In order to provide maximum power transfer from the antenna to the rectifier circuit, a dual resonant rectifier network is designed as a transition between a 50 Ω nominal antenna output and the non-linear rectification device over the power range of −40 to −10 dBm (see Fig. 11). Hence, a coupled-resonator structure with both series and shunt resonators is designed to achieve a dual-band network[40]. The linear equivalent circuit model of the SBD chip[39] has been taken into consideration to design the dual band match at the desired frequency bands. In Fig. 11, $C_{equivalent}$ represents the total capacitance of the diodes and bypass capacitor and $L_{equivalent}$ is the overall parasitic inductance of the diodes. The series *L-C* resonator ($L4 + L_{equivalent}$ and $C_{equivalent}$) and the parallel L-C resonator (*C3* and *L3*) define the dual resonant circuit. The series resonator corresponds closely to the higher band specification of 852–869 MHz, whilst the parallel resonator approximates the lower 478–496 MHz band. A minimum number of components were used in order to reduce the ohmic and parasitic losses.

The resonant frequency of each sub-circuit was determined in isolation using the following equation:

$$f = \frac{1}{2\pi\sqrt{LC}} \tag{3}$$

The 852–869 MHz band resonator circuit components were calculated. Here, $C = C_{equivalent} \cong 1.3$ pF consists of the combination of the bypass capacitor (6.8 pF) and

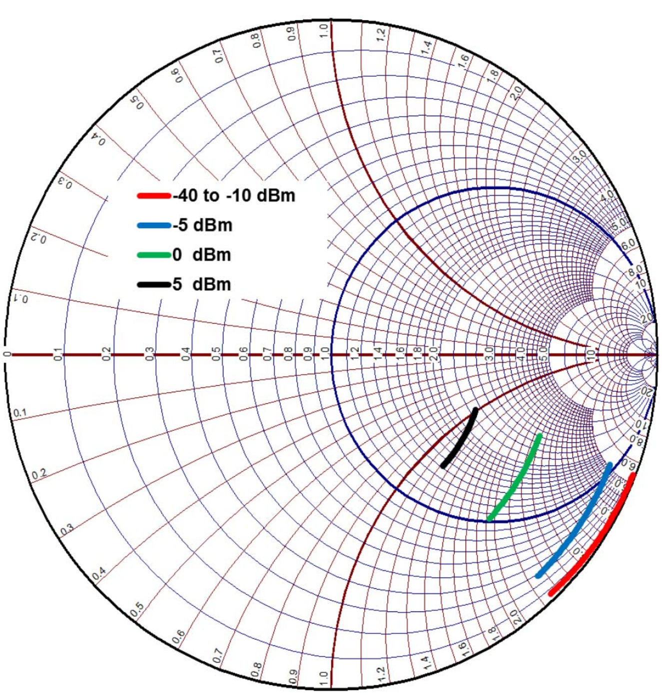

Figure 10 | Diode input impedance calculated with Large Signal S-parameter analysis over the frequency range of 400 to 900 MHz with various unmatched input power levels.

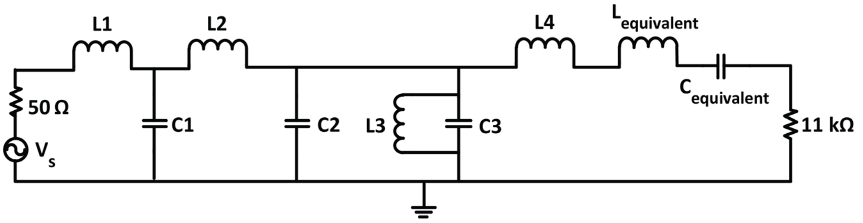

Figure 11 | Schematic of a dual resonant rectifier (optimized parameters of the chip components are: $L1 = 3.9$ nH, $C1 = 0.2$ pF, $L2 = 12$ nH, $C2 = 1.8$ pF, $L3' = 3.9$ nH, $C3' = 7.5$ pF, $L4' = 11.6$ nH, $L_{equivalent} \cong 1$ nH, $C_{equivalent} \cong 1.3$ pF).

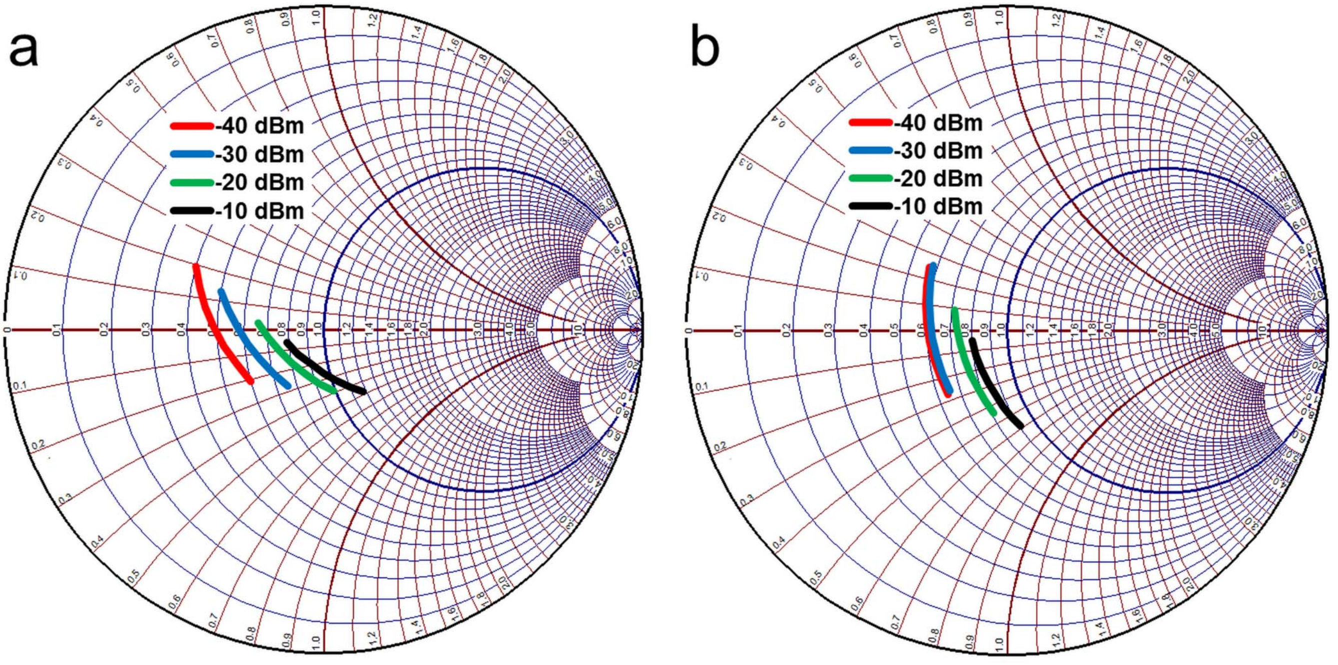

Figure 12 | Dual resonant impedance matching with −40 to −10 dBm input RF power. (a) 478–496 MHz. (b) 852–869 MHz.

the overall junction ($C_{j0} = 0.7$ pF) and parasitic capacitance ($C_p = 0.08$ pF) of *D1* and *D2*. Thus, *L* is calculated to be 26.5 nH in order to achieve an appropriate resonant frequency. Note that, *L* consists of L_4 and the overall parasitic inductance ($L_{equivalent} \cong 1$ nH) of *D1* and *D2*. The 478–496 MHz band resonator circuit components were calculated as $C3 \cong 15$ pF and $L3 \cong 7.2$ nH. Hence, the initial component values are determined for the two resonant circuits.

Initially these resonators were combined to achieve a dual-band structure. Then standard *LC* matching technique[40] is utilized to determine *C1*, *C2*, *L1*, and *L2* to achieve minimum reflection at the resonant frequencies. The substitution of realistic chip component values with their associated parasitics, and addition of 50 Ω microstrip lines and T-junctions introduce delay and shift the imaginary part of the input impedance. The via-holes also contribute to extra inductance in the circuit. Hence minor circuit adjustments are made in order to fine tune the resonant frequencies to the desired values. The final optimized values of the standard chip components are: $L3' = 3.9$ nH, $C3' = 7.5$ pF and $L4' = 11.6$ nH. Large Signal S-parameter analysis is also performed to demonstrate the matching network performance as the input power is varied. Simulation results for the input impedance of the circuit depicted in Fig. 11 are illustrated in Fig. 12. The proposed dual-resonant matching circuit achieves a VSWR <2 at 478–496 MHz and 852–869 MHz for input power ranging from -40 to -10 dBm. It should be noted that the matching circuit was designed based on the input impedance of two diodes and the output resistor and capacitor (Fig. 10). Therefore, selecting a different value for the load resistor requires a new matching circuit to be designed.

1. Paradiso, J. A. & Starner, T. Energy scavenging for mobile and wireless electronics. *IEEE Pervasive Comput.* **4**, 18–27 (2005).
2. Singh, R., Gupta, N. & Poole, K. F. Global green energy conversion revolution in 21st century through solid state devices. 26th International Conference on Microelectronics, Nis, Serbia. *IEEE.* DOI:10.1109/ICMEL.2008.4559221. 45–54 (11–14 May, 2008).
3. Wang, Y. K., Sheng, G. P., Shi, B. J., Li, W. W. & Yu, H. Q. A Novel electrochemical membrane bioreactor as a potential net energy producer for sustainable wastewater treatment. *Sci. Rep.* **3**, DOI:10.1038/srep01864 (2013).
4. Inayat, S. B., Rader, K. R. & Hussain, M. M. Nano-materials enabled thermoelectricity from window glasses. *Sci. Rep.* **2**, DOI:10.1038/srep00841 (2012).
5. Brown, W. C. The history of power transmission by radiowaves. *IEEE Trans. Microw. Theory Tech.* **32**, 1230–1242 (1984).
6. Le, T., Mayaram, K. & Fiez, T. Efficient far-field radio frequency energy harvesting for passively powered sensor networks. *IEEE J. Solid-State Circuits.* **43**, 1287–1302 (2008).
7. Paolo, J. & Gaspar, P. D. Review and future trend of energy harvesting methods for portable medical devices. Proceedings of the World Congress on Engineering (WCE). London, U. K. 909–914 (30 June–2 July, 2010).
8. Matsuki, H., Yamakata, Y., Chubachi, N., Nitta, S. & Hashimoto, H. Transcutaneous DC-DC converter for totally implantable artificial heart using synchronous rectifier. *IEEE Trans. Magn.* **32**, 5118–5120 (1996).
9. Scorcioni, S. *et al.* RF to DC CMOS rectifier with high efficiency over a wide input power range for RFID applications. Microwave Symposium Digest (IMS), *IEEE MTT-S.* Montreal, QC, Cabada. DOI:10.1109/MWSYM.2012.6259760. 1–3 (17–22 June, 2012).
10. Shameli, A., Safarian, A., Rofougaran, A., Rofougaran, M. & De Flaviis, F. Power harvester design for passive UHF RFID tag using a voltage boosting technique. *IEEE Trans. Microw. Theory Tech.* **55**, 1089–1097 (2007).
11. Mandal, S. & Sarpeshkar, R. Low-power CMOS rectifier design for RFID applications. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **54**, 1177–1188 (2007).
12. Sun, H., Guo, Y. X., He, M. & Zhong, Z. Design of a high-efficiency 2.45-GHz rectenna for low-input-power energy harvesting. *IEEE Antennas Wireless Propag. Lett.* **11**, 929–932 (2012).
13. McSpadden, J. O., Fan, L. & Chang, K. Design and experiments of a high-conversion-efficiency 5.8-GHz rectenna. *IEEE Trans. Microw. Theory Tech.* **46**, 2053–2060 (1998).
14. Din, N. M., Chakrabarty, C. K., Bin Ismail, A., Devi, K. K. A. & Chen, W. Y. Design of RF energy harvesting system for energizing low power devices. *PIER.* **132**, 49–69 (2012).
15. Barnett, R. E., Liu, J. & Lazar, S. A RF to DC voltage conversion model for multi-stage rectifiers in UHF RFID transponders. *IEEE J. Solid-State Circuits.* **44**, 354–370 (2009).
16. Kanaya, H. *et al.* Energy harvesting circuit on a one-sided directional flexible antenna. *IEEE Microw. Wireless Compon. Lett.* **23**, 164–166 (2013).
17. Buonanno, A., D'Urso, M. & Pavone, D. An ultra wide-band system for RF energy harvesting. 5th European Conference on Antennas and Propagation (EUCAP). Rome, Italy. *IEEE.* 388–389 (11–15 April, 2011).
18. Hagerty, J. A., Helmbrecht, F. B., McCalpin, W. H., Zane, R. & Popovic, Z. B. Recycling ambient microwave energy with broad-band rectenna arrays. *IEEE Trans. Microw. Theory Tech.* **52**, 1014–1024 (2004).
19. Shariati, N., Rowe, W. S. T. & Ghorbani, K. RF field investigation and maximum available power analysis for enhanced RF energy scavenging. 42nd European Microwave Conference (EuMC). Amsterdam RAI, The Netherlands. *IEEE.* 329–332 (29 October–1 November, 2012).
20. Kim, P., Chaudhary, G. & Jeong, Y. A dual-band energy harvesting using frequency limited dual-band impedance matching. *PIER.* **141**, 443–461 (2013).
21. Li, B. *et al.* An antenna co-design dual band RF energy harvester. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **60**, 3256–3266 (2013).
22. Niotaki, K. *et al.* A compact dual-band rectenna using slot-loaded dual band folded dipole antenna. *IEEE Antennas Wireless Propag. Lett.* **12**, 1634–1637 (2013).
23. Suh, Y. H. & Chang, K. A high-efficiency dual-frequency rectenna for 2.45- and 5.8-GHz wireless power transmission. *IEEE Trans. Microw. Theory Tech.* **50**, 1784–1789 (2002).
24. Collado, A. & Georgiadis, A. Conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna. *IEEE Trans. Circuits Syst. I, Reg. Papers.* **60**, 2225–2234 (20013).
25. Rizzoli, V., Bichicchi, G., Costanzo, A., Donzelli, F. & Masotti, D. CAD of multi-resonator rectenna for micro-power generation. European Microwave Integrated Circuits Conference (EuMIC). Rome, Italy. *IEEE.* 331–334 (228–29 September, 2009).
26. Huang, F. J. *et al.* Rectenna application of miniaturized implantable antenna design for triple-band biotelemetry communication. *IEEE Trans. Antennas Propag.* **59**, 2646–2653 (2011).
27. Pham, B. L. & Pham, A. V. Triple bands antenna and high efficiency rectifier design for RF energy harvesting at 900, 1900 and 2400 MHz. Microwave Symposium Digest (IMS), *IEEE MTT-S.* Seattle, WA. DOI:10.1109/MWSYM.2013.6697364. 1–3 (2013).
28. Pinuela, M., Mitcheson, P. D. & Lucyszyn, S. Ambient RF energy harvesting in urban and semi-urban environments. *IEEE Trans. Microw. Theory Tech.* **61**, 2715–2726 (2013).
29. Mikeka, C. & Arai, H. Microwave Tooth for sensor power supply in battery-free applications. Proceedings of the Asia-Pacific Microwave Conference (APMC). Melbourne, VIC. *IEEE.* 1802–1805 (5–8 December, 2011).
30. Baum, T., Thompson, L. & Ghorbani, K. Complex dielectric measurements of forest fire ash at X-band frequencies. *IEEE Geosci. Remote Sens. lett.* **8**, 859–863 (2011).
31. Avago Technologies, *Surface Mount RF Schottky Barrier Diodes HSMS-282X Series. Data sheet.* (2005). Available at: http://www.avagotech.com/docs/AV02-1320EN. (Accessed: 22/11/2014).
32. Avago Technologies, *Surface Mount Zero Bias Schottky Detector Diodes HSMS-285X Series. Data sheet.* (2005). Available at: http://www.avagotech.com/docs/AV02-1377EN. (Accessed: 22/11/2014).
33. Valenta, C. R. & Durgin, G. D. Rectenna performance under power-optimized waveform excitation. IEEE International Conference on RFID. Penang. DOI:10.1109/RFID.2013.6548160. 237–244 (30 April–2 May, 2013).
34. Trotter, M. S. & Durgin, G. D. Survey of range improvement of commercial RFID tags with power optimized waveforms. IEEE International Conference on RFID. Orlando. DOI:10.1109/RFID.2010.5467265. 195–202 (14–16 April, 2010).
35. Australian Government, *Australia Radio Frequency Spectrum Plan (ACMA).* (January 2013). Available at: http://acma.gov.au/~/media/Spectrum%20Transformation%20and%20Government/Information/pdf/Australian%20Radiofrequency%20Spectrum%20Plan%202013.pdf. (Accessed: 22/09/2014).
36. Keyrouz, S., Visser, H. J. & Tijhuis, A. G. Rectifier analysis for radio frequency energy harvesting and power transport. 42nd European Microwave Conference (EuMC). Amsterdam RAI, The Netherlands. *IEEE.* 428–431 (29 October–1 November, 2012).
37. Avago Technologies, *Schottky Diode Voltage Doubler. Application Note 956-4.* (2005). Available at: http://c1233384.r84.cf3.rackcdn.com/UK_HPA_HSCH-5331_7AN.pdf. (Accessed: 22/11/2014).
38. Eriksson, H. & Waugh, R. W. *A temperature compensated linear diode detector. Design Tip*. Agilent Technologies. (2000). Available at: http://electronix.ru/forum/index.php?act=Attach&type=post&id=13726. (Accessed: 22/11/2014).
39. Avago Technologies, *The Zero Bias Schottky Diode Detector at Temperature Extremes-Problems and Solutions. Application Note AN 1090.* (2005). Available at: http://www.efo.ru/download/Zero_bias_shottky_over_temperature.pdf. (Accessed: 22/11/2014).
40. Mattaei, G., Young, L. & Jones, E. M. T. *Microwave Filters, Impedance-Matching Networks and Coupling Structures.* Artech house books, Dedham, MA. (1980).

Acknowledgments

The authors would like to acknowledge Dr. Khashayar Khoshmanesh for valuable advice. The authors also acknowledge Mr. David Welch for assistance in fabrication of the rectifier circuit.

Author contributions

N.S. designed and simulated the rectifier, carried out the measurements, interpreted results and wrote the paper. K.G. directed the research, contributed to perform the simulation and measurement and validation of design and results. W.S.T.R. supervized the research, analyzed the data and contributed to the general concept, validation of design and results. J.R.S. analyzed the data and contributed to the general concept, validation of design and results. All authors reviewed the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Shariati, N., Rowe, W.S.T., Scott, J.R. & Ghorbani, K. Multi-Service Highly Sensitive Rectifier for Enhanced RF Energy Scavenging. *Sci. Rep.* **5**, 9655; DOI:10.1038/srep09655 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Sensors **2012**, *12*, 13636-13663; doi:10.3390/s121013636

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Optimization of Passive Low Power Wireless Electromagnetic Energy Harvesters

Antwi Nimo *, Dario Grgić and Leonhard M. Reindl

Department of Microsystems Engineering, Laboratory for Electrical Instrumentation,
University of Freiburg, IMTEK, Georges-Köhler-Allee 106, 79110 Freiburg, Germany;
E-Mails: grgic@imtek.de (D.G.); leonhard.reindl@imtek.uni-freiburg.de (L.M.R.)

* Author to whom correspondence should be addressed; E-Mail: antwi.nimo@imtek.de;
Tel.: +49-761-203-7223; Fax: +49-761-203-7222.

Received: 13 August 2012; in revised form: 28 September 2012 / Accepted: 28 September 2012 / Published: 11 October 2012

Abstract: This work presents the optimization of antenna captured low power radio frequency (RF) to direct current (DC) power converters using Schottky diodes for powering remote wireless sensors. Linearized models using scattering parameters show that an antenna and a matched diode rectifier can be described as a form of coupled resonator with different individual resonator properties. The analytical models show that the maximum voltage gain of the coupled resonators is mainly related to the antenna, diode and load (*remote sensor*) resistances at matched conditions or resonance. The analytical models were verified with experimental results. Different passive wireless RF power harvesters offering high selectivity, broadband response and high voltage sensitivity are presented. Measured results show that with an optimal resistance of antenna and diode, it is possible to achieve high RF to DC voltage sensitivity of 0.5 V and efficiency of 20% at −30 dBm antenna input power. Additionally, a wireless harvester (*rectenna*) is built and tested for receiving range performance.

Keywords: RF energy harvesting; wireless power transmission; coupled resonators; Schottky diode; RF to DC power converter; impedance matching; PI-matching; L-matching; rectenna

---

1. Introduction

For autonomous powering of sensor nodes in remote or inaccessible areas, wireless power transfer provides the only viable option to power them from an energy source. Due to the low power density of ambient RF at far-field from transmitters, there is a need to optimize each aspect of a wireless RF energy harvester for possible realistic applications. Today remote autonomous sensors are mostly powered by batteries, which have limited lifespan. Renewable powering has the potential to power autonomous sensors perpetually. Due to the expansion of telecommunications technology ambient electromagnetic (EM) power is among the most common sources of ambient energy. There are power transmitters/receivers scattered in practically any society, ranging from television transmission stations to cell phone transmitters and even wireless routers in our homes/offices or mobile phones. These transmitters in our environment and others which are on special dedicated frequencies produce ambient RF power (on the order of microwatts) which can be used as a source for powering remote microwatt budget sensors through wireless energy harvesting. This work presents different matching techniques based on different application requirements using Schottky diode-based RF to DC power converting circuits for wireless remote EM energy harvesting around 434 MHz and 13.6 MHz. Generalized analytical models and limitations of the matched RF to DC power converters are discussed. A wireless RF energy harvester consisting of an antenna and a matched diode rectifier is then realized and its performance tested. Passive wireless energy harvesting also finds applications in near field communications (NFC) [1], RFID tags [2–5], implantable electronics [6,7], and environmental monitoring [8], among others.

1.1. State of the Art

Hertz was the first to demonstrate the propagation of EM waves in free space and to demonstrate other properties of EM waves such as reflection using parabolic reflectors [9]. Wireless power transmission was then investigated and demonstrated for possible wireless remote powering by Tesla. Electromagnetic power beaming for far field wireless power transfer using collimated EM waves was proposed in the 1950s [9]. Recent advances in ultralow power sensors means ambient omni-directional EM power can be used as a source for powering remote sensors without the need to collimate the EM power through the wireless space. Mickle [10] and McSpadden [11] have presented earlier work on wireless energy harvesting systems using Schottky diodes and rectennas where the usability of ambient RF power into DC power was shown. Sample [12] presented a wireless harvester which can harvest EM power from TV and radio base stations transmitting 960 kW of effective radiated power; 60 μW was harvested at a range of about 4 km. Umeda [13] and Le [14] have presented more integrated wireless energy harvesters based on CMOS RF to DC rectifying circuits. CMOS-based rectifying power converters provide full compatibility with standard CMOS technologies and have advantages in batch processes for mass production. The drawback of CMOS-based diode connected transistors is the need to bias the gate of the transistors for the rectifying circuits to effectively function. This gate bias is provided externally, which makes the system not passive. Without the injection of external charges or a biasing of the transistor gate, the circuit has low efficiency, especially when the amplitude of the input voltage is low [15]. Shameli [2] presented a passive CMOS RF to DC power converter with a voltage sensitivity of 1 V at −14.1 dBm input, but the circuit efficiency was only 5 %. Zbitou [16] presented an RF to DC converter based on Schottky diodes and achieved 68 % efficiency at 20 dBm RF input power. Ungan [17,18] presented antennas and high quality factor RF to DC power converters at 24 MHz and 300 MHz for RF wireless energy harvesting at -30 dBm input power. The power converter used high quality factor resonators for impedance matching the EM source and the diodes and achieved high open circuit voltage sensitivity of 1 V/μW. Boquete [19] presented a risk assessment system for calculating insurance premiums by monitoring mobile phone usage while driving. This was done by harvesting EM power from detected mobile phone usage during driving for risk assessment. Heikkinen [20] presented rectennas on different substrates at 2.4 GHz using transmisson lines to match the antennas output resistance (*at resonance*) to the rectifying diodes. Akkermans [21] presented a rectenna design by complex conjugating impedance provided by a microstrip structure to a diode so that resonance may be achieved for a working frequency. This design approach may need sophisticated tools to realize and the dominant resonance frequency of the rectenna can be unpredictable in practice. Hagerty [22] presented rectenna arrays for broadband ambient EM harvesting and characterized the harvesters from 2 GHz to 18 GHz; rectennas combine impedance matching the RF rectifying circuit and the antenna into one compact device, but an array of rectennas may increase the overall size of an EM harvester. Herb [23] and Vullers [24] have provided a comprehensive state of the art for micro energy harvesting and have explored the various techniques used for harvesting ambient renewable energy.

2. RF to DC Power Converter

2.1. Diode Rectifier

A junction diode equivalent circuit and simple Schottky diode rectifier are shown in Figure 1. R_{DS} is the diode resultant series resistance, C_{DS} is the diode resultant series capacitance, R_{DP} is the diode resultant parallel resistance, C_{DP} is the diode resultant parallel capacitance, Vs is the sinusoidal source voltage and Vc is the voltage across the capacitor.

Figure 1. (**a**) Diode series equivalent model, (**b**) Diode parallel equivalent model, (**c**) Simple diode detector.

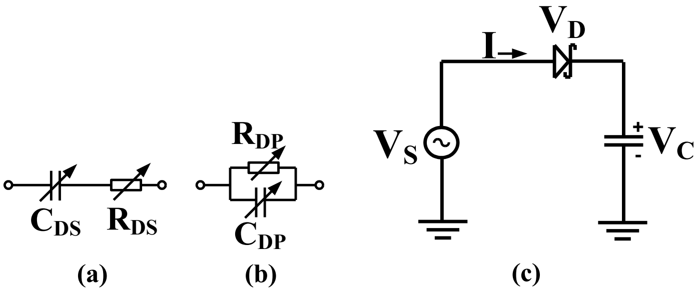

The diode capacitive impedance is mainly due to the junction capacitances provided by the metal, its passivation and the semiconductor forming the diode. AC power incident on a forward biased diode input is converted to DC power at the output. The current-voltage behavior of a single metal/semiconductor diode is described by the Richardson equation [25] as in Equation (1):

$$I = I_S\left(e^{\left(qV_D/nKT\right)} - 1\right) \tag{1}$$

where I is the current through the diode, I_S is the saturation current, q is the charge of an electron, V_D is the voltage across the diode, T is the temperature in degrees Kelvin and K is Boltzmann constant. The voltage equation around the loop can be derived from Figure 1(c) and is given in Equation (2):

$$V_D = V_S - V_C \tag{2}$$

Since the same current flows through the diode and the capacitor, one can find the average current through the circuit by integrating Equation (1) over a time period. By substituting Equation (2) into Equation (1), V_C can be expressed in terms of V_S by averaging the diode current to zero. This is given in Equation (3) [26]:

$$V_C = \frac{KT}{q}\ln\left[\vartheta_0\left(\frac{qV_S}{KT}\right)\right], \tag{3}$$

where ϑ_0 is the series expansion of the sinusoidal source voltage. Equation (3) can further be simplified for very small amplitude V_S as Equation (4):

$$V_C \approx \frac{qV_S^2}{4KT} \tag{4}$$

Equation (4) shows that for a small voltage source, the circuit output voltage is proportional to the square of the input sinusoidal voltage; hence it's so-called square law operation. Extensions of this model for voltage multipliers and other input signals are presented in [27] and [28]. Equation (4) further confirms that for low input voltage (power ≤ 10 dBm), an impedance matching network between the source and the diode is necessary to improve the detected output voltage and efficiency.

2.2. Impedance Matching

The maximum power transfer theorem states that the highest power is transferred to the load when the source resistance is the same as the load resistance. For systems with both resistive and reactive impedances from source and load, the source and the load impedance should be adjusted in a way that they are the complex conjugate of each other through impedance matching. For the purposes of this work, a 50 Ω resistive source is chosen as reference for load impedance matching. The antenna which captures the ambient RF signals is tuned to provide this source resistance at resonance for the rectifying circuit in a complete EM wireless remote harvester. The load is the resistance of the Schottky diodes and the actual connected resistance (*remote sensor*). The specific type of matching network which can be used for complex conjugation depends on the nature of load or source impedance, the desired RF to DC converter functionality and other factors like circuit size, cost, *etc.* The response of a matched RF to DC power converter depends on the matching network used as well as the source or load component quality factors and impedances.

2.3. Diode Impedance

Schottky diodes HSMS-285C and HSMS-286C from Avago [29,30] are used to build the RF to DC power converters. The HSMS-285× or 286× series diodes can be operated as zero biased with relatively low forward junction potential. This allows for the realization of completely passive RF to DC power converters for wireless energy harvesting. The HSMS-285C or 286C is a pair of series connected Schottky diodes in a SOT-323 package. The impedance of the HSMS-285C and HSMS-286C diodes was first measured so it can be matched to the resistance (50 Ω) of the antenna source. This is done by connecting the input of the diodes to a network analyzer and measuring the scattering parameters. These scattering parameters are then converted to the corresponding impedances. The input impedance of a diode depends mainly on the resistive and capacitive impedance provided by the junction of the diode and its connected load. For a couple of diodes arranged in a package such as the HSMS-285C or 286C, the input impedance is the vector sum of the impedances provided by each diode in the package arrangement, the extra impedance associated with the packaging and the connected load. The diode measuring board is as shown in Figure 2. The diodes were measured at room temperature for an input power of -30 dBm at a diode connected load of 1 MΩ with a 100 pF filter capacitor. For the sake of this work, the input impedance of the diodes will always be referred to at these connected load conditions.

Figure 2. (**left**) Reference circuit layout for measuring diodes input impedance, (**right**) measuring printed circuit board (PCB) for diodes input impedance on 1 mm FR4 substrate.

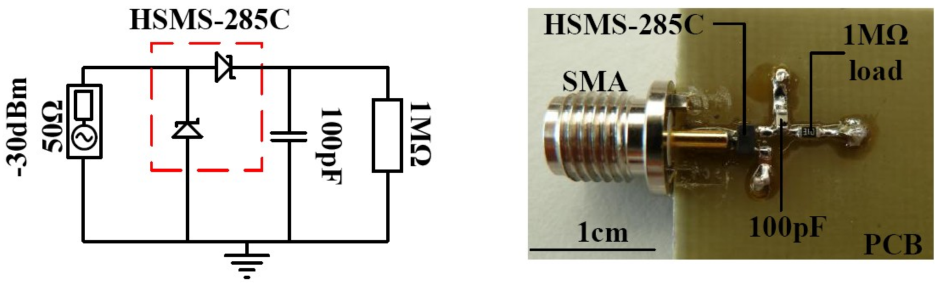

Figure 3. Measured input impedance (Δ resistive, □ capacitive) of HSMS-285C (**left**) and HSMS-286C (**right**) diodes at −30 dBm input with 1 MΩ load and 100 pF filter.

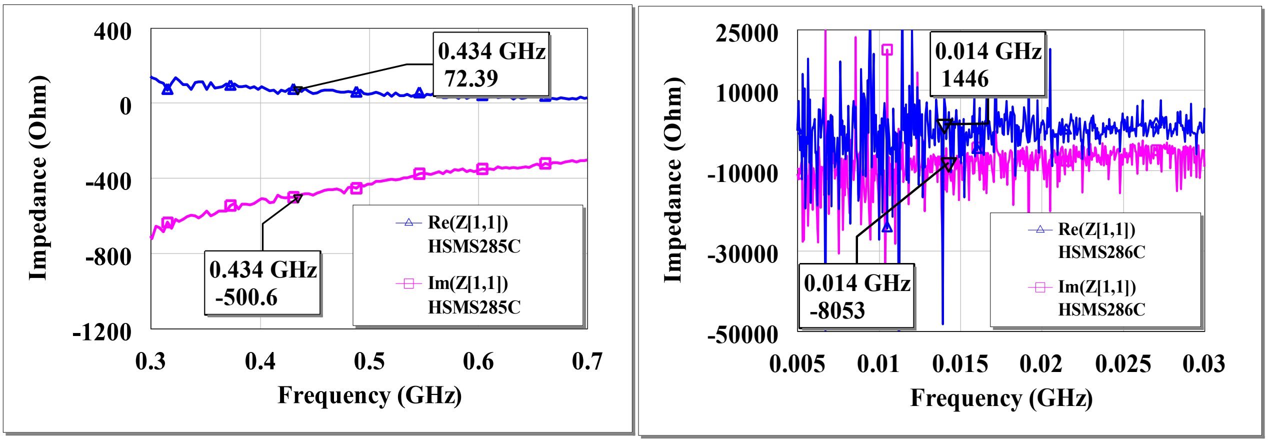

The board is fabricated such that components are soldered directly one into another to prevent additional impedances introduced by copper route. The PCB backside had the ground layer. An example of measured input impedance for HSMS-285C and HSMS-286C is shown in Figure 3.

The diodes quality factor is given by $X_{DS}R_{DS}^{-1}$, where X_{DS} is the resultant series capacitive impedance of the diodes. At an input power of −30 dBm, the measured input impedance of the HSMS-285C diodes is 72–j501 Ω at 434 MHz and 587–j1239 Ω at 13.6 MHz. For HSMS-286C diodes, it is 10–j503 Ω at 434 MHz and ~1.5–j8.1 kΩ at 13.6 MHz for −30 dBm input. The measured impedance of the HSMS-286C diodes at low frequencies (< 60 MHz) shows pronounced fluctuations. The low-frequency excess flicker noise and the shot noise observed in the HSMS-286C have been studied by several authors [31–33]. The pronounced presence of trap states in the depletion region of the semiconductor, mobility fluctuations in carriers, edge effects among other reasons is reported to cause deviations from the ideal Schottky diode behavior and hence generation-recombination noise for some diodes such as the HSMS-286C [34]. When a diode rectifier is matched at a reference operating condition, the matching network may function less effectively at other input power levels, connected load and other operating frequencies. This is due to possible changes in the diode input impedance. Throughout this work the imperfections of the matching circuit at other operating conditions away from the matched reference conditions are accepted without changes to the matching network.

2.4. Voltage Doubler

The Delon voltage doubler and Greinacher doubler are both used to realize the RF to DC power converters presented in this work. The Delon voltage doubler and Greinacher doubler are shown in Figure 4. The diodes output voltage (V_{out}) is doubled what is detected by a simple detector circuit shown in Figure 1. Both doublers produce the same output performance, the only difference is that the Delon doubler has an instantaneous input ground which is not shared with the output.

Figure 4. Circuit diagram of voltage doubler, (**a**) Delon doubler and (**b**) Greinacher doubler.

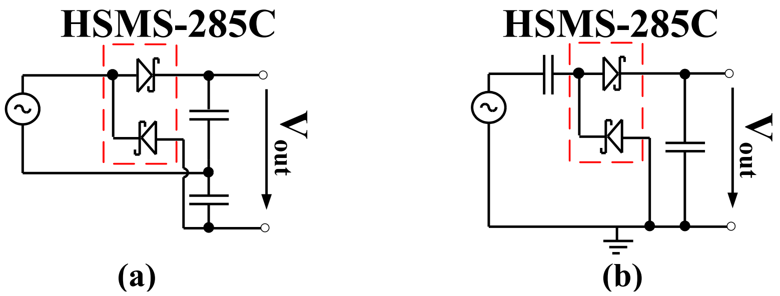

2.5. Matching Techniques for Antenna Source and RF to DC Power Converter

2.5.1. L-match RF to DC Power Converter

An L-match network converts a source series impedance to its equivalent load parallel impedance or *vice-versa* and tunes out by subtracting or adding any surplus reactance from the load or source with the counter impedance. Series impedance is converted to its parallel equivalent impedance using Equations (5–7):

$$Q_S = \frac{X_S}{R_S} \tag{5}$$

$$Q_P = \frac{R_P}{X_P} \tag{6}$$

where X_S is the total series reactive impedance, R_S is the total series resistance, R_P is the total parallel resistance, X_P is the total parallel reactive impedance, Q_S and Q_P are the series and parallel quality factors respectively:

$$R_S + jX_S = \frac{R_P \times jX_P}{R_P + jX_P} \tag{7}$$

Equation (7) is the equation of a series sum of impedances and a parallel sum of impedances. It is interesting to note that Q_S and Q_P from an L-matched network may be different from the individual component quality factors as a result of the inherent resistive and reactive impedances in that component. By virtue of Equation (7), Q_S and Q_P must be equal in an L-matched network. Using Equations (5,6) and (7), the ratio of the parallel resistance (*or reactance*) to the series resistance (*or reactance*) can be derived in terms of the quality factors Q_P or Q_S [35]. Since at match conditions, only the resistive impedances dissipate power, the loaded quality factor Q, of the L-matched network can be expressed as in Equation (8):

$$R_P = (Q^2 + 1)R_S \tag{8}$$

Using Equations (5,6) and (8), series impedance can be converted to its parallel equivalent for a fixed frequency and power level. As an example; a series impedance 72–j501 Ω (HSMS-285C at 434 MHz for −30 dBm input power) is easily converted to $-j510(3519)/(-j510 + 3519)$ Ω as its parallel equivalent with a component quality factor of 6.96. The source resistance is taken as part of the parallel matching network in an L-match circuit if the source series equivalent resistance is greater than the load series equivalent resistance. On the other hand, the load resistance is taken as part of the parallel matching network if the load series equivalent resistance is greater than the source series equivalent resistance. For the purpose of this work, inductors were only used for series impedance matching and capacitors as shunts. This prevents power seeping through any shunt inductor used for impedance matching due the short circuit provided by a shunt inductor to ground and resulting in less output efficiency. Resistors were not used for impedance matching.

2.5.2. L-match RF to DC Converter Generalized Analytical Model

The classical matching technique using Equations (5,6) and (8) is first used to L-match the 50 Ω resistance of the antenna to the resistance of the HSMS-286C diodes (and load) at 434 MHz for −30 dBm input and then the generalized model is discussed. The antenna source resistance was L-matched to the resistance of the diodes (and load). The 50 Ω resistance of the antenna is taken as the parallel matching component and the diodes 10 Ω resistance is the series matching component. The loaded Q is found as 2 between the 50 Ω antenna source resistance and the 10 Ω diode series resistance using Equation (8). From this loaded Q, a shunt capacitive impedance of 25 Ω (14.6 pF at 434 MHz) using Equation (6) and a series inductive impedance of 20 Ω (7.3 nH at 434 MHz) using Equation (5) will match the 50 Ω source to the 10 Ω HSMS-286C diodes (and load) series resistance at −30 dBm input. Since the HSMS-286C diodes inherently provides 503 Ω series capacitive impedance at −30 dBm, a resultant series inductive impedance of 523 Ω (192 nH at 434 MHz) is needed to tune the 50 Ω resistive source to the complete HSMS-286C diodes impedance at 434 MHz for −30 dBm input. The L-matched HSMS-286C diodes rectifier is as shown in Figure 5(a).

Figure 5. (**a**) L-match RF to DC harvester using the HSMS-286C diodes at 434 MHz for −30 dBm input. (**b**) Small signal impedance model of a generalized L-matched RF to DC power converter as capacitive coupled series RLC resonator with different resonator elements.

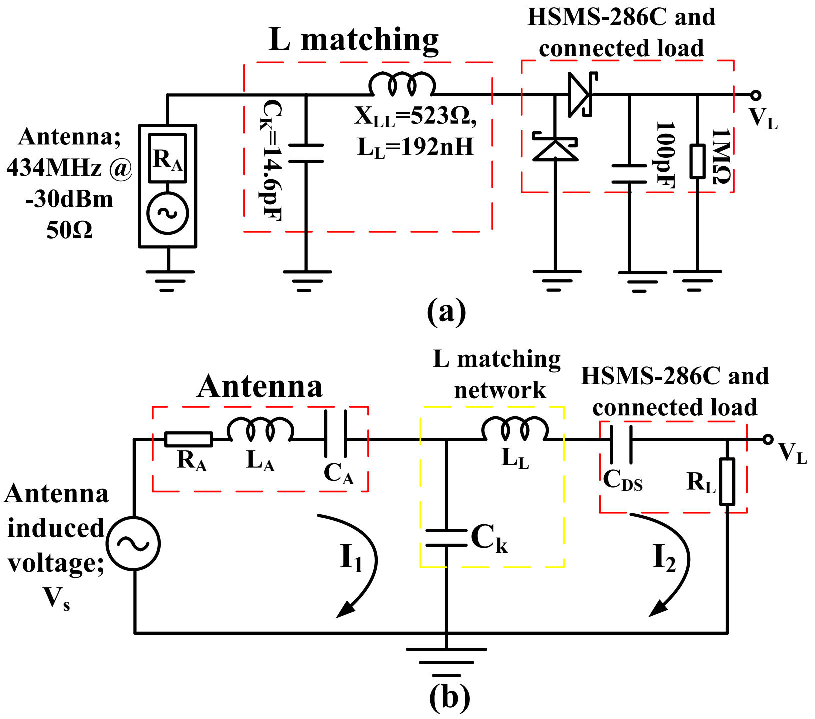
L matching
HSMS-286C and connected load
Antenna; 434MHz @ -30dBm 50Ω
R_A
C_K=14.6pF
X_{LL}=523Ω, L_L=192nH
100pF
1MΩ
V_L
(a)
Antenna
L matching network
HSMS-286C and connected load
R_A
L_A
C_A
L_L
C_{DS}
R_L
V_L
Antenna induced voltage; V_s
I_1
C_k
I_2
(b)

C_K is the tuning capacitance, L_L is the tuning inductance, X_{LL} is the tuning inductive impedance, C_{DS} is the diodes series capacitance, X_{DS} is the diodes series capacitive impedance, V_S is the antenna captured ambient EM voltage, R_A is the resistance of antenna, L_A is the inductance of antenna, C_A is the capacitance of antenna, R_L is the resultant series resistance from the diodes and the connected load resistance, V_L is the resistive load voltage. From Figure 5(a) the power dissipated in the resistance of the diodes (and connected load); P_L is given by Equation (9), where R_L is the series resistance of the diodes and load:

$$P_L = \frac{V_L^2}{R_L} \tag{9}$$

The source power; P_S is given by Equation (10), where V_S* is the root mean squared (RMS) antenna captured source voltage. Half of the source power is transferred to the resistance of the diodes (and connected load) at match conditions as described by the maximum power transfer theorem:

$$P_S = \frac{V_S^2}{R_A} \quad \text{or} \quad P_S = \frac{V_{S*}^2}{2R_A} \tag{10}$$

Equating P_L and half RMS P_S gives a condition of maximum voltage gain for the matched RF to DC power converter shown in Figure 5(a):

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \tag{11}$$

From Equation (8), substituting the series and parallel resistance ratio into Equation (11) the voltage gain can be expressed in terms of the loaded quality factor as in Equations (12) and (13), where Q is the loaded quality factor of the RF to DC power converter:

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{\frac{1}{1 + Q^2}} \tag{12}$$

Equation (12) is the voltage gain in-terms of the loaded Q if the resistance of the diodes (and connected load) is part of the series matching network and the resistance of the antenna source is part of the parallel matching network as in Figure 5(a). If the resistance of the diodes is part of the parallel matching network, then Equation (13) may be written as the voltage gain in-terms of the loaded Q in an L-matched circuit:

$$\frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{1 + Q^2} \tag{13}$$

Equations (12) and (13) shows that the maximum voltage gain is directly related to the relative differences between the diodes (and connected load) resistance and source resistance at matched conditions or the circuit loaded quality factor. It is interesting to note that the circuit shown in Figure 5(a) has a loaded Q of 2, but an HSMS-286C unloaded quality factor of 50 (at 434 MHz for −30 dBm).

Figure 5(a) is generally modeled as capacitive coupling of two series RLC resonators with a voltage source. This linearized model can be made at any defined frequency and power level. The model however neglects the metal/semiconductor physics of the diode's junction potentials which results in a Schottky barrier. The first series RLC resonator is modeled as impedance from the antenna with or without some passive matching components. The voltage source V_S, is the antenna captured electromagnetic voltage. The second series RLC resonator is the impedance from the diodes (at a defined condition), connected resistance and some passive matching components. Ck is modeled as the coupling element between the two series RLC resonators. Figure 5(b) gives a more general look at the special scenario shown in Figure 5(a). The voltage equations in the two loops are given by Equations (14,15) according to Kirchhoff's voltage loop laws, where ω is the angular frequency and I_1, I_2 are the currents in the first loop and second loop, respectively:

$$V_S = I_1\left[R_A + j\omega L_A - \frac{j}{\omega C_A} - \frac{j}{\omega C_K}\right] + \frac{jI_2}{\omega C_K} \tag{14}$$

$$0 = \frac{jI_1}{\omega C_K} + I_2\left[R_L + j\omega L_L - \frac{j}{\omega C_{DS}} - \frac{j}{\omega C_K}\right] \tag{15}$$

Using Cramers rule, I_2 can be expressed as:

$$I_2 = \frac{\dfrac{-jV_S}{\omega C_K}}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}}. \tag{16}$$

The voltage across R_L is V_L; given by $I_2 R_L$:

$$V_L = \frac{\dfrac{-jV_S}{\omega C_K} R_L}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}} \tag{17}$$

The voltage gain of the coupled resonator can be expressed as in Equation (18):

$$\frac{V_L}{V_S} = \frac{\dfrac{-jR_L}{\omega C_K}}{\left[R_A + j\omega L_A - \dfrac{j}{\omega C_A} - \dfrac{j}{\omega C_K}\right]\left[R_L + j\omega L_L - \dfrac{j}{\omega C_{DS}} - \dfrac{j}{\omega C_K}\right] + \dfrac{1}{\omega^2 C_K^2}} \tag{18}$$

At resonance, there is no resultant reactance in the RLC resonators or the capacitive and inductive impedances become equal; hence Equation (19) can be written:

$$\omega L_A - \frac{1}{\omega}\left\{\frac{1}{C_A} + \frac{1}{C_K}\right\} = 0 \text{ and } \omega L_L - \frac{1}{\omega}\left\{\frac{1}{C_{DS}} + \frac{1}{C_K}\right\} = 0 \tag{19}$$

Equations in Equation (19) can be used to find the resonant frequencies of the series coupled resonator. The voltage gain of the coupled resonator at resonance can then be expressed as in Equation (20):

$$\frac{V_L}{V_S} = V_{gain} = \frac{\dfrac{-jR_L}{\omega C_K}}{R_A R_L + \dfrac{1}{\omega^2 C_K^2}} \tag{20}$$

where V_{gain} is the voltage gain. V_{gain} at resonance is a function of the resistance of the source and load and the coupling element. The maximum of Equation (20) is obtained when:

$$\frac{dV_{gain}}{dC_K} = 0. \tag{21}$$

This gives the results as in Equation (22):

$$\frac{dV_{gain}}{dC_K} = \frac{j2R_L}{\omega^3 C_K^4} - j\left\{R_A R_L + \frac{1}{\omega^2 C_K^2}\right\}\frac{R_L}{\omega C_K^2} = 0 \text{ or } R_A R_L^2 = \frac{R_L}{\omega^2 C_K^2} \tag{22}$$

Equation (22) can be simplified to find $C_{K(max)}$:

$$C_{K(\max)} = \pm\frac{1}{\omega}\sqrt{\frac{1}{R_A R_L}} \tag{23}$$

where C_{Kmax} is the value of the coupling element where maximum power transfer from the first resonator to the second resonator occurs. Using Equations (19) and (23) the unknown optimal matching impedances can be found from the known impedances just like the classical L-matched procedure using Equations (5,6) and (8). By substituting $C_{K(max)}$ into Equation (20) and taking the magnitude of V_{gain}, gives the maximum voltage gain of the coupled series resonator at resonance:

$$\left|\frac{V_L}{V_S}\right| = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \text{ or simply } \frac{V_L}{V_{S*}} = \frac{1}{2}\sqrt{\frac{R_L}{R_A}} \tag{24}$$

For wireless harvesters consisting of an antenna and a diode rectifying circuit, the diode resistive impedance at any condition is dependent on the diode realized parameters, signal frequency, connected load and the input power level. The source impedance is determined by the impedance of the antenna. For maximum efficiency, the ratio of the source resistance to the load resistance must tend to zero at matched conditions. The efficiency η of the circuit is given by Equation (25):

$$\eta = \frac{P_L}{P_S}; \eta \to 1 \ when \ \frac{R_A}{R_L} \to 0 \tag{25}$$

2.5.3. L-Match RF to DC Converter Experimental Results and Discussion

The presented circuit was L-matched between the 50 Ω resistance of the antenna source and the resistance of the HSMS-285C diodes (and load) at 434 MHz for −30 dBm input as shown in Figure 6. Since the series equivalent resistance of the HSMS-285C diodes and load (72 Ω) is greater than the 50 Ω series resistive antenna source, the diode is taken as parallel matching network with a parallel equivalent impedance of $-j510(3519)/(-j510 + 3519)$ Ω. The analysis follows the same procedure as in Section 2.5.2 after this step. Figure 6(b) shows the resultant L-matched RF to DC power converter. C_{DP}* is the resultant shunt matching capacitance.

Figure 6. (**a**) L-matched impedance circuit for matching the HSMS-285C diodes at 434 MHz for −30 dBm input. (**b**) Resultant network, (**c**) PCB layout of the L-matched Delon doubler with adjusted values on FR4 substrate (**d**) Fabricated PCB of the L-network matched Delon voltage doubler.

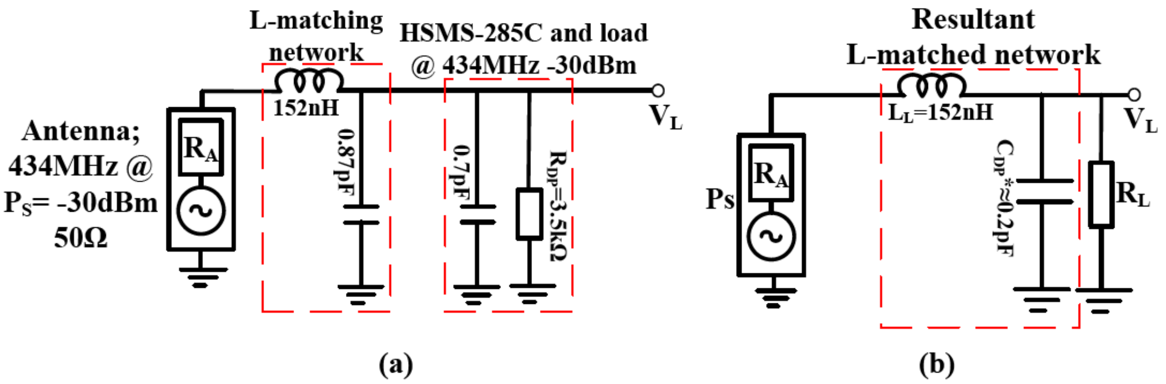
L-matching network
HSMS-285C and load @ 434MHz -30dBm
152nH
0.87pF
0.7pF
R_DP=3.5kΩ
V_L
Antenna; 434MHz @ P_S= -30dBm 50Ω
R_A
(a)
Resultant L-matched network
L_L=152nH
C_DP*≈0.2pF
R_L
V_L
Ps
R_A
(b)

Figure 6. *Cont.*

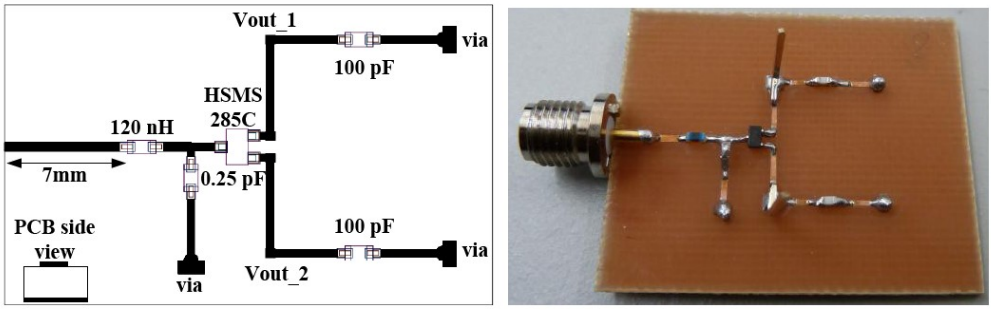

(c) **(d)**

Figure 6(a,b) assume perfect characteristic impedance between the various components in the matched circuit. When a copper route is introduced between components and on a material substrate, it must be accounted for in the total impedance as seen by the source or load. This PCB impedance compensation is carried out in Advance Design Systems (ADS) from Agilent [36]. ADS has extensive models for microstrip substrates to account for its impedances. The optimized layout using ADS microstrip models and its compensated values in the passive tuning components for a Delon doubler is shown in Figure 6(c).

The circuit reflection coefficient (S_{11}) and input impedance at open circuit are shown in Figure 7. There is high return loss and resonance around 434 MHz. The circuit input impedance at open circuit conditions is ~38 Ω at resonance for −40 dBm and ~17 Ω at −10 dBm input.

The measured L-matched circuit efficiency and voltage sensitivity is shown in Figure 8. The maximum measured L-matched efficiency at −30 dBm is 22% at ~20 kΩ load and an open circuit voltage of 124 mV. At −10 dBm, the maximum efficiency and open circuit voltage is 47% and 2 V respectively. At the optimal load of ~20 kΩ, the detected voltage is 58 mV and 1 V at −30 dBm and −10 dBm respectively.

Figure 7. Measured open circuit S_{11} of the L-matched Delon circuit at different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the L-matched circuit (**right**).

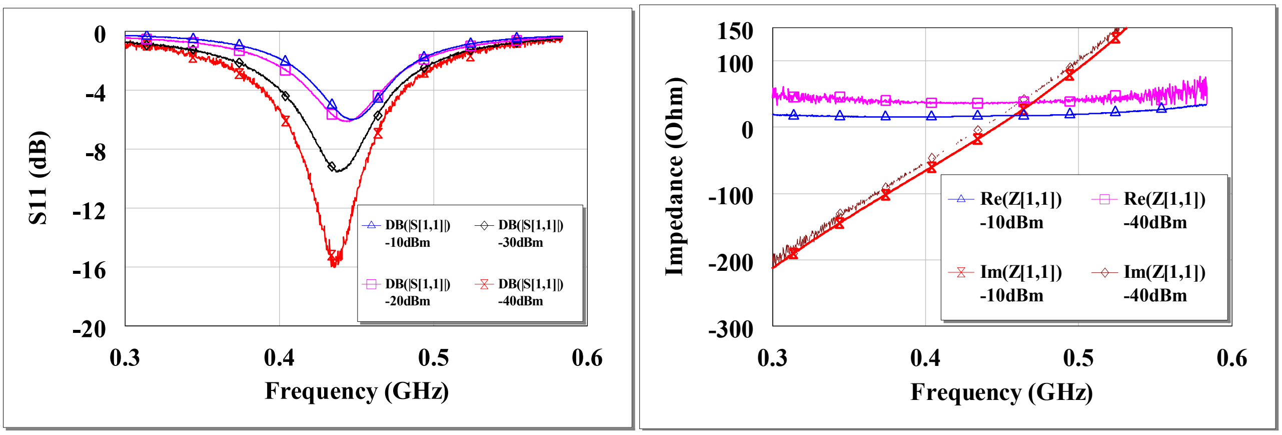

The open circuit voltage gain is 25 at −30 dBm and 40 at −10 dBm. The maximum measured efficiency at −35 dBm is 27%. This is higher than that of −30 dBm due to the better matched circuit impedance at −35 dBm (35 Ω) than at −30 dBm (27 Ω) input. The L-matched RF to DC power converter has a loaded *Q*, sensitivity and efficiency determined mainly by the diodes resistance, diodes junction potential, connected resistance and antenna source resistance at matched conditions.

Figure 8. Measured L-matched circuit efficiency *versus* resistive load at various input power levels at 434 MHz (**left**), measured open circuit voltage and at 17 kΩ load *versus* input power at 434 MHz (**right**).

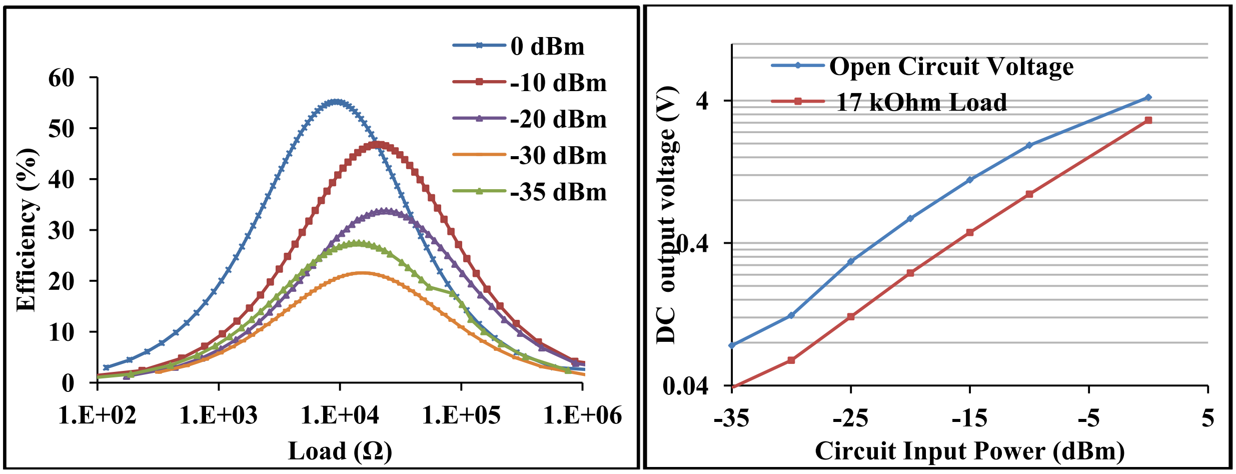

2.5.4. PI-match RF to DC Power Converter

A highly selective or small frequency bandwidth RF power converter is realized with a PI-network in-between the source impedance from the antenna and the diode rectifier. A PI-network is a ‘back to back’ L-network that are both configured to match the load and source impedance to an invisible resistance located at the junction between the two L-networks [37]. The quality factor of the L-network with the parallel resistance is given by Equation (26):

$$Q_P^* = \sqrt{\frac{R_P}{R^*} - 1}, \qquad (26)$$

where R_P is the parallel resistance, R^* is a virtual resistance and Q_P^* is the quality factor of the L-network with the parallel resistance. The quality factor of the L-network with the series resistance is given by Equation (27):

$$Q_S^* = \sqrt{\frac{R_S}{R^*} - 1}, \qquad (27)$$

where Q_S^* is the quality factor of the L-network with the series resistance. The unloaded quality factor; Q_S^* or Q_P^* is set higher than what is normally achieved with a single L-network [37] to realize the small frequency bandwidth circuit. The resistance of the load is assigned the parallel network in a PI-matched circuit if its series equivalent resistance is higher than the source series equivalent resistance; the opposite is true if the source is higher than the load. Equation (26) and Equation (27) are synonymous to Equation (8), except the lowest resistive impedance in Equation (8) is substituted with the virtual resistance which is dependent on the newly desired circuit selectivity. From Equations (26) and (27) the loaded quality factor of the PI-matched circuit can be written as Equation (34) in terms of Q_S^* and Q_P^*:

$$Q^2 = \left[\left(\frac{Q_P^{*2} + 1}{Q_S^{*2} + 1} \right) - 1 \right], \tag{28}$$

where Q is the loaded quality factor of the PI-network. Q_S* or Q_P* are the unloaded quality factors of the PI-matched network. The larger value among the unloaded quality factors result in small frequency bandwidth response which is desired when matching a source and load impedance with a PI-network. Some authors approximate the highest value of Q_S* or Q_P* or their algebraic sum as the loaded quality factor of the PI-network as in [35] and [37], but Equation (28) gives the exact loaded Q of the PI-matched circuit in terms Q_S* and Q_P*. This allows for the correct estimation of the maximum voltage gain from the loaded quality factor.

2.5.5. Selectivity RF to DC Converter Generalized Analytical Model

An example of a PI-matched RF to DC converter using the HSMS-285C diodes operating at 434 MHz for −30 dBm input is presented first and then the generalized model is discussed. The circuit is matched for Q_P* of 60 between the antenna and the resistance of the diodes as shown in Figure 9.

Figure 9. Impedance diagram of PI-matched RF power converter; (**a**) Impedance diagram of 50 Ω source and the HSMS-285C diodes at 434 MHz, (**b**) Resultant PI matched network between the antenna source and load resistance.

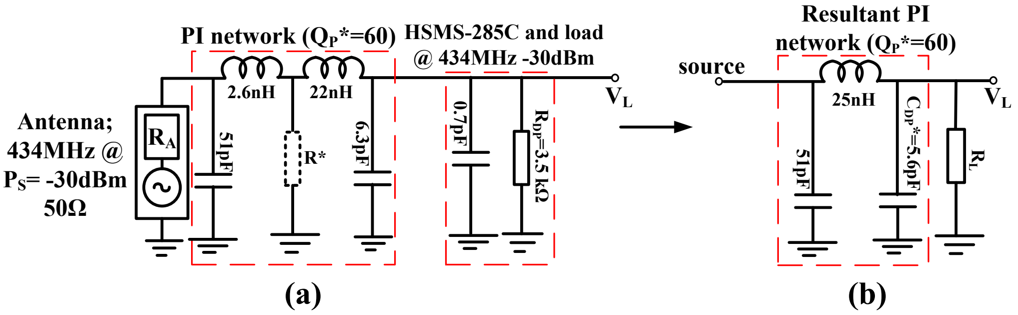
PI network (QP*=60)
HSMS-285C and load @ 434MHz -30dBm
Antenna; 434MHz @ PS= -30dBm 50Ω
RA
2.6nH
22nH
51pF
R*
6.3pF
0.7pF
RDP=3.5 kΩ
VL
(a)
Resultant PI network (QP*=60)
source
25nH
51pF
CDP*=5.6pF
RL
VL
(b)

Figure 9(b) can also be modeled as an inductive coupling of two parallel RC circuits. A more general look at such a circuit is shown in Figure 10, as an inductive coupling of two parallel RLC resonators with a current source.

Figure 10. Inductive coupled parallel RLC small signal model of a generalized PI-matched antenna and diode rectifier.

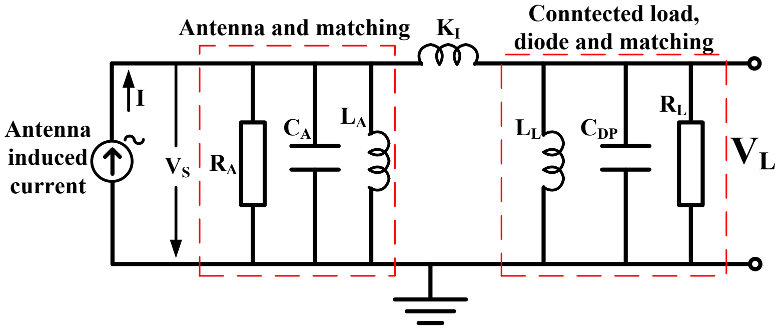
Antenna and matching
KI
Conntected load, diode and matching
Antenna induced current
I
VS
RA
CA
LA
LL
CDP
RL
VL

The first parallel RLC resonator is modeled as impedance from the antenna and some passive matching components. The second parallel RLC resonator is modeled as impedance from the linearized diodes, its connected load and some passive matching components. I is the antenna induced current, V_S this time is the voltage across the parallel R_A and K_1 is the coupling element between the two parallel resonators. Using Kirchoff's current laws, the node equations can be expressed as Equations (29) and (30):

$$I = V_S\left[\frac{1}{R_A} + j\omega C_A - \frac{j}{\omega L_A} - \frac{j}{\omega K_1}\right] + \frac{jV_L}{\omega K_1} \tag{29}$$

$$0 = \frac{jV_S}{\omega K_1} + V_L\left[\frac{1}{R_L} + j\omega C_{DP} - \frac{j}{\omega L_L} - \frac{j}{\omega K_1}\right] \tag{30}$$

Load voltage (V_L) and the source voltage (V_S) at resonance are given by the equations in Equation (31). The resonance frequencies are given by Equation (32):

$$V_L = \frac{\frac{-jI}{\omega K_1}}{\left[\frac{1}{R_A R_L} + \frac{1}{\omega^2 K_1^2}\right]} \quad \text{and} \quad V_S = \frac{\frac{I}{R_L}}{\left[\frac{1}{R_A R_L} + \frac{1}{\omega^2 K_1^2}\right]} \tag{31}$$

$$\omega C_A - \frac{1}{\omega}\left\{\frac{1}{L_A} + \frac{1}{K_1}\right\} = 0 \text{ and } \omega C_{DP} - \frac{1}{\omega}\left\{\frac{1}{L_L} + \frac{1}{K_1}\right\} = 0 \tag{32}$$

From V_L and Vs in Equation (31), the voltage gain at resonance can be expressed as:

$$\frac{V_L}{V_S} = \frac{R_L}{j\omega K_1} \tag{33}$$

The maximum of Equation (33) is obtained when:

$$j\omega K_1 \to 0 \;\; or \;\; R_L \to \infty \tag{34}$$

Since $j\omega K_1$ is restricted by the conditions in Equation (32) to attain resonance, one cannot manipulate $j\omega K_1$ alone without changing the resonance conditions. What can drive the voltage gain is if R_L is very large at resonance conditions. If the input impedance (V_S/I) of the coupled resonator is maximum at resonance, conditions in Equation (35) hold:

$$\left(\frac{Vs}{I}\right) \to maximum \;\; when \;\; \frac{R_L}{\omega^2 K_1^2} \to 0 \tag{35}$$

Equation (36) may be assumed when $\frac{R_L}{\omega^2 K_1^2} \to 0$:

$$\frac{V_S}{I} = R_A \tag{36}$$

Under these conditions and an optimal coupling coefficient K_{1max}, the maximum voltage gain of the parallel coupled resonator can be written as in Equation (37), where K_{1max} is given by Equation (38):

$$\left|\frac{V_L}{V_S}\right| = \left|V_{gain}\right| = \left|\frac{1}{2}\sqrt{\frac{R_L}{R_A}}\right| \tag{37}$$

$$K_{1(\max)} = \mp\frac{1}{\omega}\sqrt{R_A R_L} \tag{38}$$

The analysis of Section 2.5.2 and parallel coupled RLC resonators show that any antenna and matched rectifying diode can be described as an equivalent circuit of a coupled resonator at a defined operating point. This general model can be applied to optimize other harvesters with complex output impedance such as piezo-harvesters or vibration harvesters for maximum transfer of power or voltage to its connected load. The model can also be applied to near field magnetically coupled antennas/coils for optimization.

2.5.6. Broadband RF to DC power converter

A broadband network is preferred when an RF to DC power converter is to be operated for a wide range of frequencies. A broadband converter is realized by connecting successive L-networks together in a multi-network between the antenna source and the rectifying diodes. The result is broadband or multiband RF power converter around certain frequencies. This can be deduced from the general model of a coupled resonators that by choosing certain passive components between a source and the load, it is possible to have more frequencies (ω) fulfilling Equation (32) and hence a result of multiple resonant frequencies or broader bandwidth at match conditions. For a two stage L-connected match, the quality factor of the L-network with the parallel resistance is given by Equation (39):

$$Q_P^{\,*} = \sqrt{\frac{R_P}{R^*} - 1} \tag{39}$$

The quality factor of the L-network with the series resistance is given by Equation (40):

$$Q_S^{\,*} = \sqrt{\frac{R_*}{R_S} - 1} \tag{40}$$

From Equations (39) and (40) the loaded quality factor of the two stage L-connected broadband network may be written as Equation (41) in terms of the unloaded quality factors; Q_S* and Q_P*:

$$Q^2 = \{(Q_P^{*2} + 1)(Q_S^{*2} + 1)\} - 1 \tag{41}$$

R^* in this case may be chosen if it is larger than R_S and lower than the R_P. The highest possible bandwidth between a resistive source and resistive load is found for a virtual resistance (R^*) given in Equation (42) [37]:

$$R^* = \sqrt{R_S R_P} \tag{42}$$

For complex loads such as rectifying diodes or transistors, the largest achievable bandwidth prescribed by Equation (42) is limited by the load or source component quality factor, since Equation (42) does not take into account reactive impedance associated with the source or load.

2.5.7. Broadband-Match RF to DC Converter Results and Discussion

The antenna source resistance was broadband matched to the HSMS-285C diodes (and load) resistance at −30 dBm input around 434 MHz. For a desired Q_P* and Q_S* of 2.7 there is ~0.4 pF inherent diode capacitance which is un-tuned using a two stage L-matching network [Figure 11(b)]. This is because the HSMS-285C diodes provides an inherent component quality factor of 6.96 at 434 MHz for −30 dBm input, hence a broadband circuit with Q_P* lower than this inherent component quality factor of the diodes (and load) is difficult to achieve without trade-offs. However, connected L-networks with Q_P* as high as the diode component quality factor may perform worse than a single L-matched network with similar loaded quality factor. This is due to redundant components of the connected L-networks which have inherent losses.

Figure 11. Impedance diagram of broadband RF power converter; (**a**) Broadband match around 434 MHz with loaded Q of 2.7, (**b**) Resultant impedance matching network with un-turned capacitance of 0.4 pF.

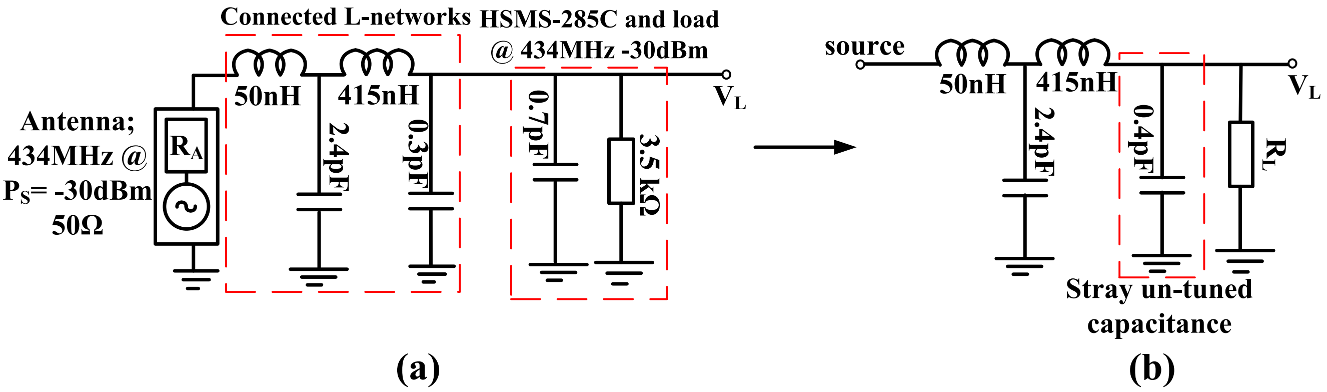
Connected L-networks
HSMS-285C and load
@ 434MHz -30dBm
Antenna;
434MHz @
P$_S$= -30dBm
50Ω
R$_A$
50nH
415nH
2.4pF
0.3pF
0.7pF
3.5 kΩ
V$_L$
source
50nH
415nH
2.4pF
0.4pF
R$_L$
V$_L$
Stray un-tuned capacitance

(a) **(b)**

Figure 12. Measured open circuit S_{11} of the broadband circuit around 434 MHz at different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the broadband circuit (**right**).

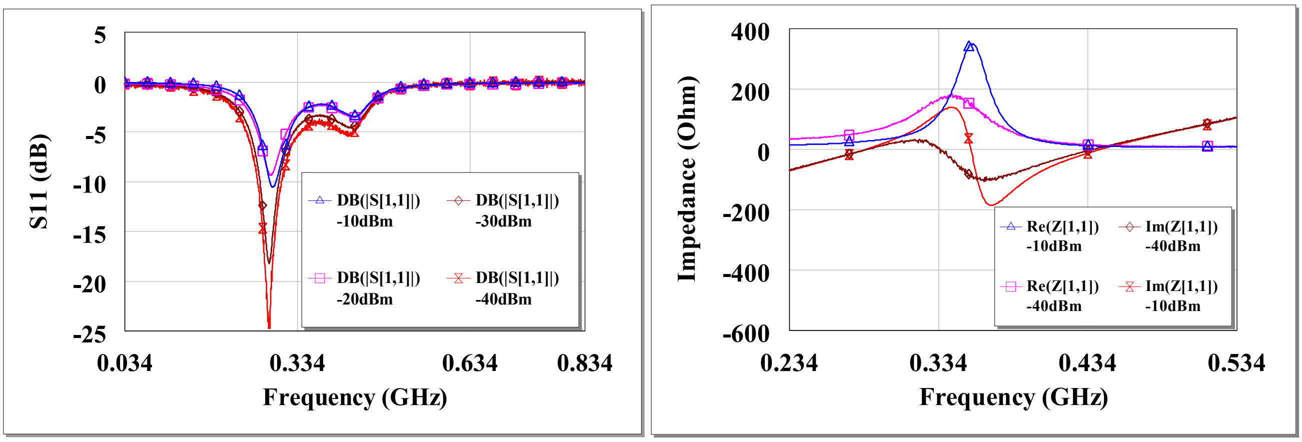

Therefore the broadband circuit is matched for Q_P* of 2.7, notwithstanding the un-tuned shunt capacitance as can be seen in Figure 11(b). Figure 12 shows the circuit S_{11} at various input power levels and input impedance at open circuit conditions. From Figure 12 (left) there is ~−5 dB return loss from 200 MHz to 500 MHz providing an operating band of ~300 MHz. The impedance of the circuit shows resonances at ~290 MHz and ~450 MHz as shown in Figure 12(right). A third resonance occurs around 356 MHz at −10 dBm as the frequency of highest harvester input resistance (~350 Ω) and where the reactive impedances approach their extremes. Figure 12 show that a wireless EM harvester can exhibit different resonance scenarios depending on the dominant instantaneous conditions. The efficiency and voltage sensitivity of the broadband matched wireless EM harvester are shown in Figure 13. The average open circuit voltage is 47 mV and 1.1 V at −30 dBm and −10 dBm, respectively, when operating from 200 MHz to 500 MHz.

Figure 13. Measured open circuit voltage *versus* frequency sweep from 200 MHz to 500 MHz for −10 dBm and −30 dBm (**left**), measured efficiency at 17 kΩ load *versus* frequency sweep for −10 dBm and −30 dBm (**right**).

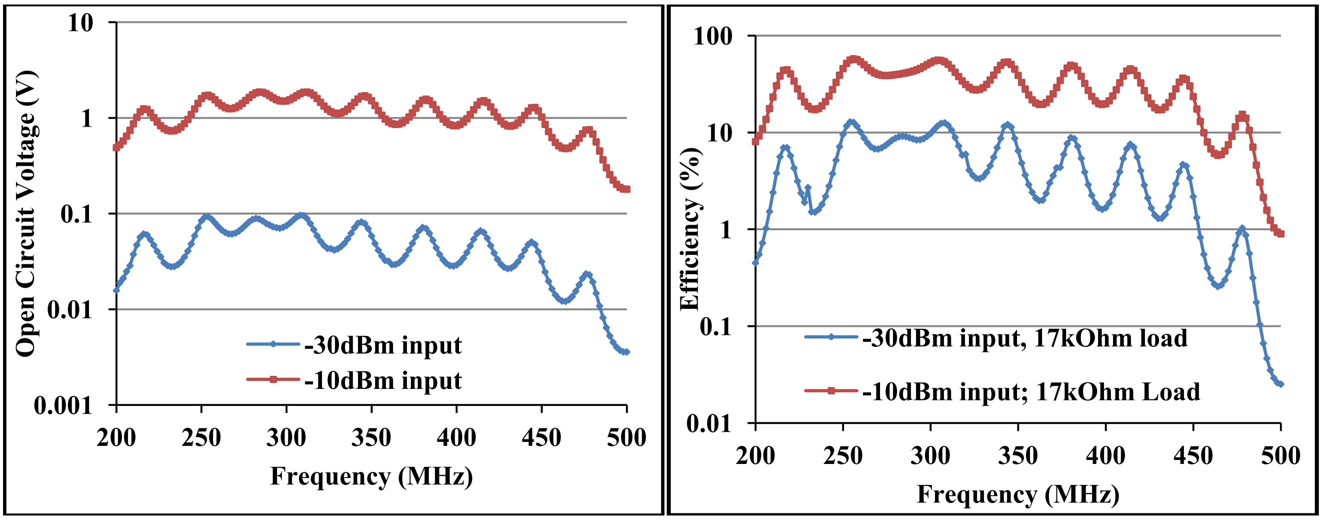

The broadband circuit achieves average efficiency of 5% at 17 kΩ load for −30 dBm and 30% at 17 kΩ load for −10 dBm input power from 200 MHz to 500 MHz. Figure 13 further confirm a direct link between frequency response and the unloaded quality factors. For Q_S* and Q_P* of ~2.7, the circuit response is broadband around 434 MHz.

2.6. High Voltage Sensitive RF to DC Converter

The current state of the art low power remote sensors would require a DC voltage supply of about 1 V and DC current of about 30 µA for operation. Therefore, the issue is not only how efficient a wireless EM harvester is in converting RF to DC power, but also what the output DC voltage and current of the EM harvester are at the RF input power level [38]. Equations (11,24) and (33) show that the maximum voltage sensitivity of a coupled resonator system or an RF to DC power converter is mostly related to the load and the source resistances at resonance. Therefore high voltage sensitive wireless EM harvester can be achieved with a diode voltage doubler with a very high input resistance relative to the antenna source without the need to cascade the diodes as in voltage multipliers. If the diodes been used for the RF to DC power conversion cannot provide high resistive impedance at the working frequency relative to the antenna source, then a DC-DC converter can be applied after the EM harvester as presented in [39] or the diodes may be cascaded by way of multipliers as presented in our earlier work [40] and by several other authors [3,5,14]. In case of multipliers, the input voltage ought to be high enough to overcome the junction potential of the several diodes in the multiplier network. If frequency is not a constraint, then a frequency sweep *versus* impedance for the diodes can be made and the frequency where the diodes exhibits high resistive impedance can be used to realize high voltage sensitive wireless RF harvester. For Schottky diodes, high resistive impedance occurs mostly at lower frequencies (see Figure 3). The measured voltage gain of a high resistive diode pair (voltage doubler) is presented in the next results.

2.6.1. High Voltage Sensitive RF to DC Converter Results and Discussion

The presented result was L-matched using 50 Ω resistance of the antenna source and the resistance of the HSMS-286C diodes (and load). The HSMS-286C diodes do provide high resistive impedance at low frequencies; notwithstanding the flicker noise which causes its resistive (and reactive) impedance to fluctuate. The HSMS-286C has low forward junction potential (~350 mV at 1 mA) per diode and series impedance of ~1.5–j8.1 kΩ or parallel impedance of ~−*j*8.3(46.3)/(−*j*8.3 + 46.3) kΩ at 13.6 MHz for −30 dBm input. Even though the HSMS-286C diodes unloaded component quality factor at 13.6 MHz is similar to that of the HSMS-285C diodes at 434 MHz, the elevated resistive impedance at 13.6 MHz fulfills the condition for high voltage sensitivity relative to a 50 Ω antenna source at resonance conditions.

Figure 14. (**a**) L-matched impedance diagram for matching the HSMS-286C diodes at 13.6 MHz at -30 dBm input. (**b**) Resultant network, (**c**) PCB layout of the L-matched Greinacher doubler with adjusted values due to impedances provided by copper route on FR4 substrate with thickness of 1 mm. (**d**) Fabricated PCB of the L-matched RF to DC power converter.

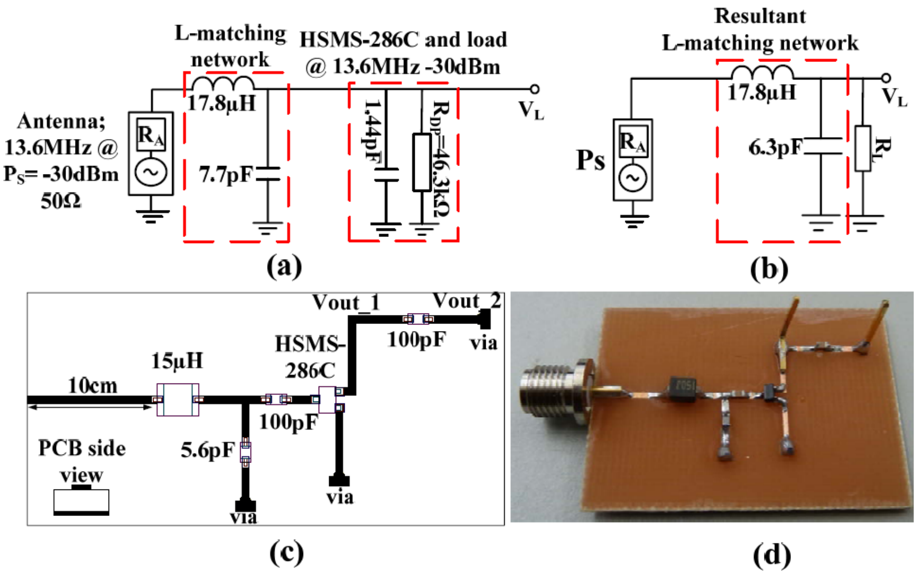
L-matching network
HSMS-286C and load @ 13.6MHz -30dBm
Antenna; 13.6MHz @ P$_S$= -30dBm 50Ω
R$_A$
17.8μH
7.7pF
1.44pF
R$_{DP}$=46.3kΩ
V$_L$
(a)
Resultant L-matching network
Ps
R$_A$
17.8μH
6.3pF
R$_L$
V$_L$
(b)
Vout_1
Vout_2
100pF
via
HSMS-286C
15μH
10cm
100pF
PCB side view
5.6pF
via
via
(c)
(d)

The high voltage sensitive EM harvester operating at 13.6 MHz is as shown in Figure 14. On the realized PCB is a Greinacher doubler. An inductance of 15 μH and a shunt capacitance of 5.6 pF were the adjusted values after the microstrip contributions.

The measured S_{11} and input impedance at open circuit are shown in Figure 15. There is high return loss and resonance around 13.6 MHz. The circuit input impedance at open circuit conditions is 58 Ω at resonance for both −40 dBm and −10 dBm.

Figure 15. Measured open circuit S_{11} of the L-matched HSMS-286C diodes at 13.6 MHz for different input power levels from a 50 Ω source (**left**), measured open circuit input impedance at −10 dBm and −40 dBm of the L-matched HSMS-286C diode at 13.6 MHz (**right**).

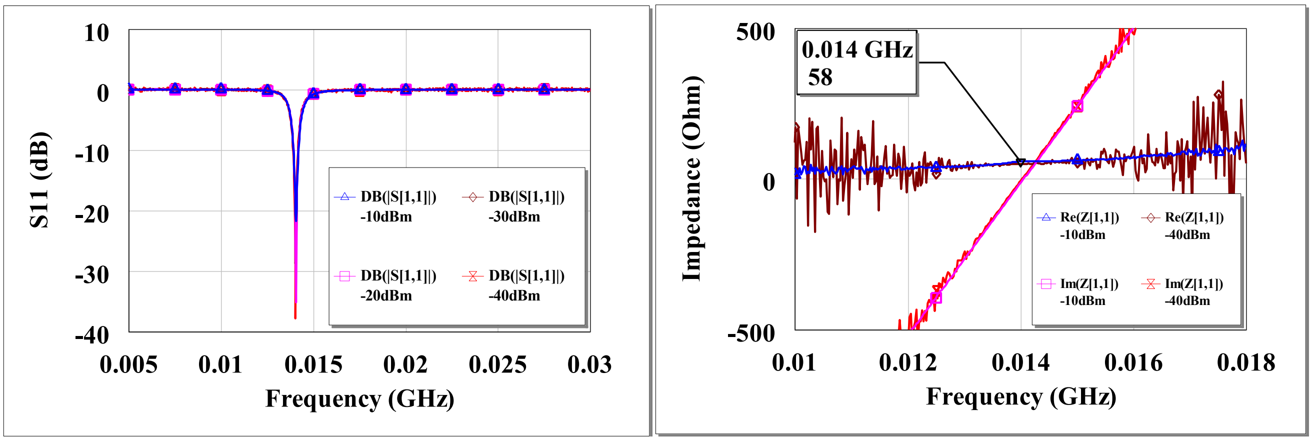

The efficiency and voltage sensitivity of the high voltage sensitive wireless EM harvester are shown in Figure 16.

Figure 16. Measured circuit efficiency *versus* load at various input power levels at 13.6 MHz (**left**), measured open circuit voltage and at 200 kΩ load *versus* input power at 13.6 MHz (**right**).

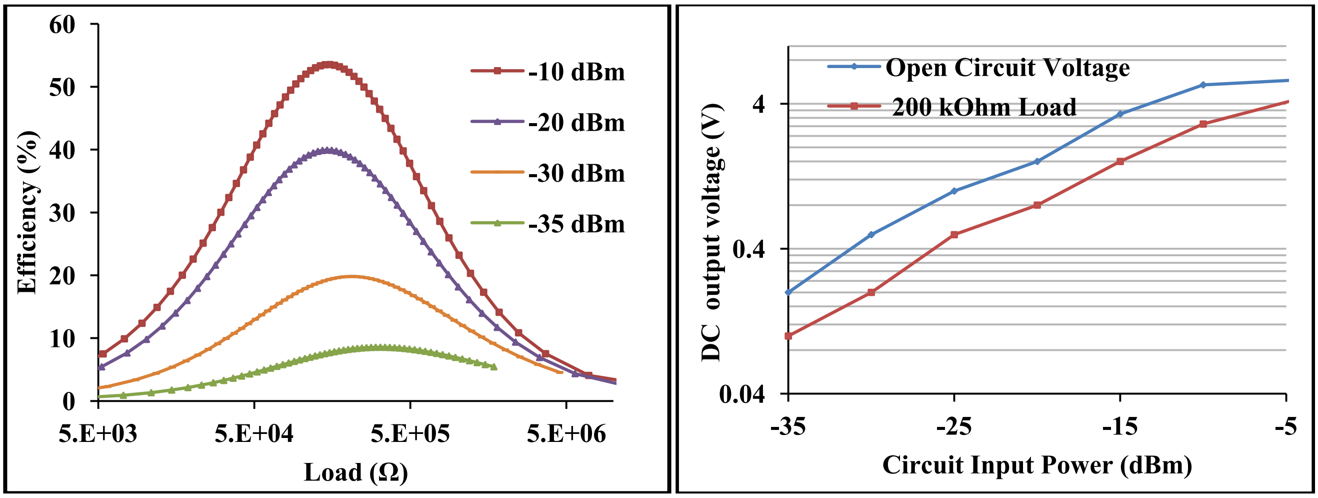

The maximum measured efficiency at −30 dBm is 20% for ~200 kΩ load and an open circuit voltage of 0.5 V. At −10 dBm, the maximum efficiency and open circuit voltage are 54% and 5.4 V respectively. At the optimal load of ~200 kΩ, the detected voltage is 0.2 V and 2.9 V at −30 dBm and −10 dBm respectively. The open circuit voltage gain is 100 at −30 dBm and 108 at −10 dBm.

Even though the RF to DC converter presented in Section 2.5.3 is the same as the L-match circuit realized with the HMSM-286C diodes at 13.6 MHz, the voltage gain is increased by a factor of 4 due to the large difference between the diodes (and load) resistance and source resistance so that at matched conditions high voltage gain occurs. The loaded *Q* of the L-matched circuit is 30 which results in small frequency bandwidth just like a PI-matched diode rectifier presented in our earlier work [40]. From this result and the results from our earlier presented PI-matched EM harvester, it can be inferred that all high loaded *Q* RF to DC circuits have high selectivity but not all highly selective RF to DC circuits have high loaded *Q*. The voltage sensitivity of the matched HSMS-286C diode at 13.6 MHz can be improved if its resistive impedance is not lowered by the flicker noise.

3. Wireless EM Power Harvester

A wireless EM harvester, consisting of a rectifying antenna (*rectenna*) was designed to find a compromise between size and performance of its antenna. The rectenna is shown in Figure 17.

Figure 17. Rectenna realized on a Duroid 5880, 1.57 mm substrate. (**a**) Backside of the rectenna, (**b**) cross-section of antenna output coplanar stripline dimensions (**c**) frontside of the rectenna.

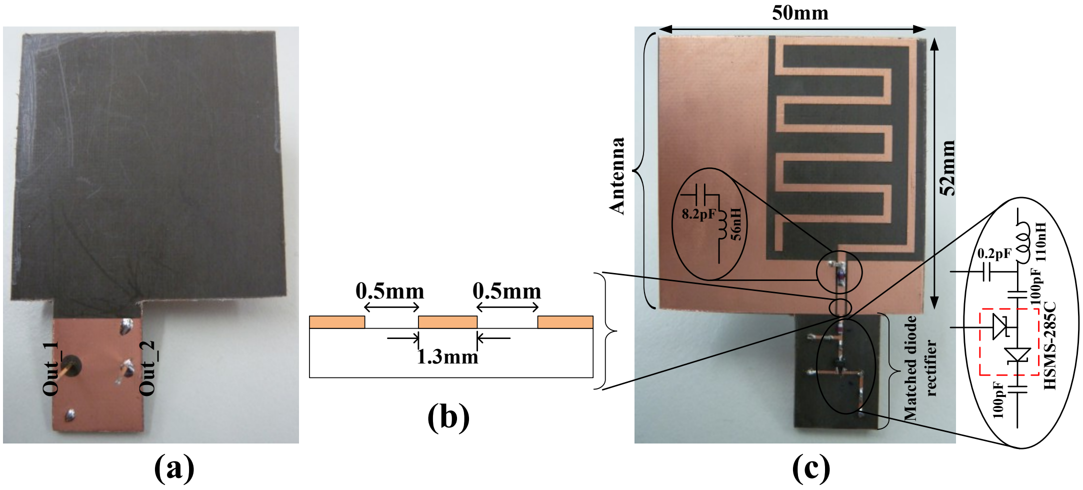

(a) (b) (c)

The antenna (planar) part of the rectenna is based on our earlier work [41]. In contrast to the earlier presented antenna, this rectenna is realized on a Duroid [42] substrate of thickness 1.57 mm. Duroid 5880 has lower loss tangent of 0.0004 at 1 MHz compared to 0.025 at 1 MHz for FR4. This means there is less loss in the transmission of signals on a Duroid PCB at this frequency range. The antenna part is fabricated to resonate around 434 MHz; hence its dimensions of 5 × 5.2 cm make it electrically small. The antenna is tuned with a chip inductor and a capacitor to achieve the resonance frequency around 434 MHz [Figure 17(c)]. This is done at a cost of reduced antenna radiation efficiency. An antenna is one of the few components the size of which is related to the operating frequency. Thus, if the size of an antenna is fixed, resonance frequency reduction of the antenna can only be achieved with penalty factors [10]. The antenna's output impedance is tuned with the dimensions of the coplanar stripline as shown in Figure 17(b).

HFSS [43] was used to simulate the presented antenna and to find the correct capacitive and inductive components for frequency tuning before the optimized design was fabricated. The simulated antenna resonances occur at 438 MHz and 445 MHz. At these frequencies, the radiation efficiency is 20% and a peak gain of −6 dBi. The rectifying part of the rectenna consists of L-matched HSMS-285C diodes (Figure 17(c)). The L-matched HSMS-285C part of the rectenna can be engineered to be as small as possible if required. The separate parts of the rectenna were characterized by terminating their ends and measuring the individual reflection coefficients just like the power converters presented in Section 2. Figure 18 shows the measured antenna and matched rectifier individual S_{11} and impedance. Figure 18 (left) also show the HFSS simulated S_{11} results. From Figure 18 (right), the measured antenna resonance where the input impedance is at maximum is ~434 MHz. At ~434 MHz, the antenna input resistance is 376 Ω and the reactive impedances approach their extreme (*so called anti-resonance*). The other resonance occurs when the input resistance is 'finite' and the reactive impedance is zero; at ~441 MHz. The input resistance is 57 Ω at ~441 MHz. The rectifier circuit is matched for the antenna's resistance at ~441 MHz.

Figure 18. Antenna HFSS simulated, antenna measured, and measured L-matched diode rectifier S_{11} on a Duroid 5880 PCB for −30 dBm input (**left**), Measured open circuit input impedance of antenna and rectifier at −30 dBm input (**right**).

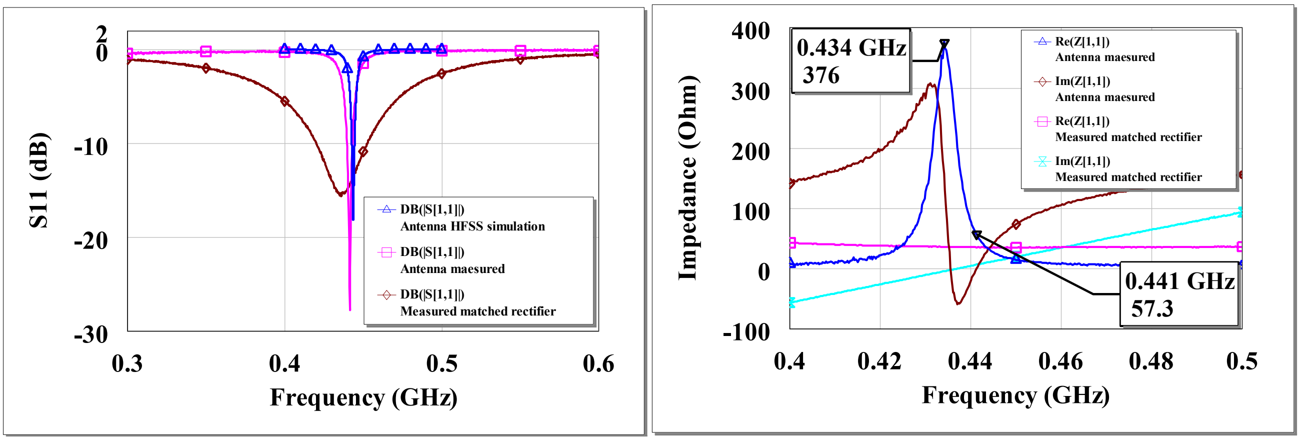

3.1. EM Range Results and Discussion

At far field between wireless EM transmitting and receiving antenna, the coupling mechanism between the transmitting and receiving antenna is neither capacitive nor inductive as is the case for the RF to DC converters. The coupling is radiative which can be described by the Friis equation of transmission on the assumption that the transmitting and receiving antenna are in free space [44]. A modified Friis equation for a transmitting and receiving antenna at far-field (R >> λ and R >> transmitting antenna largest dimension) to each other at a specified direction is given by Equation (43) [45]. Equation (43) assumes real world open space conditions:

$$\frac{P_r}{P_t} = F_{envt} G_t G_r \left(\frac{\lambda}{4\pi R} \right)^2 , \tag{43}$$

where *Pr* is the power at the receiving antenna port, *Pt* is the power supplied at the transmitting antenna port, F_{envt} is a factor accounting for environmental effects as such ground reflections among others, *Gt and Gr* are the transmitting and receiving antenna gain (*at specified direction*) respectively. *R* is the distance between the transmitting and receiving antenna and λ is the wavelength of the transmitting EM wave. The rectenna receiving range measurements were carried out in an open space (*hall*) with the antennas 2 m above ground level. The transmitting and receiving antennas were arranged in the direction of their peak gain. The rectenna range performance is shown in Figure 19. According to Equation (43), the efficiency of RF power transferred between a sending and receiving antenna depends on controllable factors like the gain of the antennas in the arranged direction and the radiation efficiency of the antennas. Since the receiving/transmitting antenna's incorporated in remote harvesters for sensor powering are normally small in relation to their operating frequencies, they tend to be less efficient.

Figure 19. Rectenna receiving range performance by sending 17 dBm (50 mW) at a gain of −6 dBi at 437 MHz. Output DC voltage *versus* receiving distance for different loads (**left**), loads output power *versus* receiving distance (**right**).

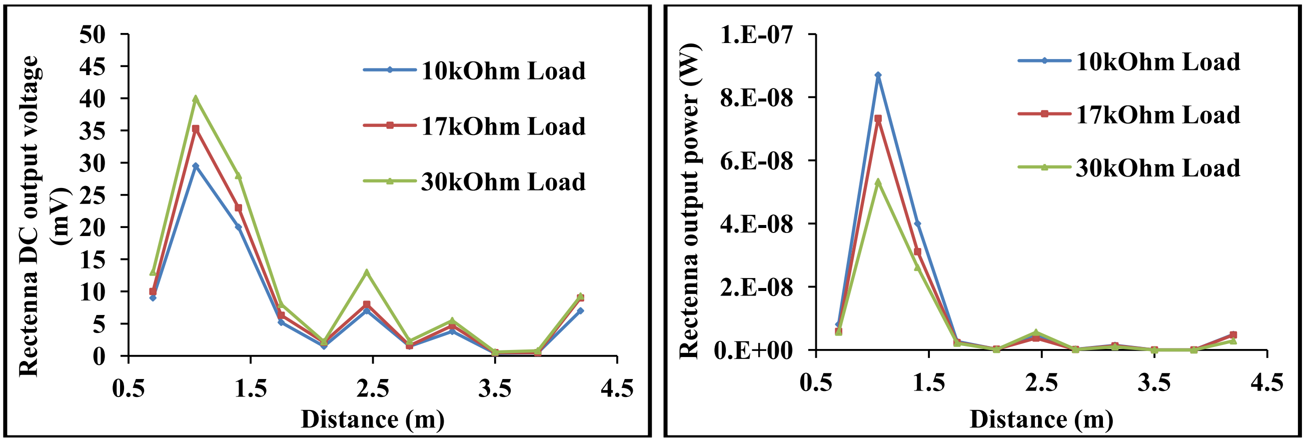

The efficiency of the rectenna's antenna is ~20% at resonance. A 'perfectly' matched RF to DC power converter operating in its square law region has efficiencies in the region of 20% as depicted in Section 2. The transmitting antenna was the same as the antenna incorporated in the rectenna. By transmitting the EM power with a small antenna (5 cm × 5.2 cm) at 437 MHz with efficiency of ~20% and at a gain of −6 dBi, the power delivered by the rectenna is generally low at far-field from the transmitter as can be seen in Figure 19. A mediocre transmitting antenna was used to transmit the EM waves due to limitations in the European Union about transmitting EM power at certain frequencies; so the goal in the rectenna range experiment is to show the lowest limit functionality of such a harvester. At 4.2 m from the electrically small transmitting antenna transmitting at 17 dBm, the rectenna harvested DC voltage and power are 9 mV and 5 nW respectively for 10 kΩ load. It can be seen from Figure 19 that the harvested voltage/power generally degrades as an inverse square of distance from transmitter as described by Friis equation. The measured received power however alternate along this R^{-2} fit as shown in Figure 19. This anomaly is accounted for by F_{envt} [Equation (43)] as influence of ground reflections and polarization in real world open field measurements [45]. For any particular distance *R*, the signals reflected from ground can be constructive with the direct signal to the rectenna, in which case the measured power may be higher than that predicted by the original Friss equation as in [44]. The ground effect can also be destructive, in which case the measured power will be lower than what is predicted by the original Friis equation.

4. Conclusions

Optimization of Schottky diode-based RF to DC power converters using different matching techniques for wireless EM energy harvesting applications is presented. Using scattering parameters for small signal modeling, it is shown that wireless EM harvesters can be generally described as coupled resonators with efficiencies and maximum voltage sensitivity depending mostly on the source and load resistances under matched conditions. The analytical models allow systematic control in the design of passive wireless EM harvesters. Based on these analyses, a rectenna is built and tested for lower limit functionality from harvesting ambient EM waves. The analysis presented in this work may also be applied to optimize derivatives of wireless EM harvesters like RFID tags, NFC, wireless chargers *etc.*, for efficient powering of their sensors or integrated circuits. Generally, most energy harvesters and their matched loads can be described as coupled resonators and thus may be optimized with the methods presented in this work.

Acknowledgments

This work is part of the graduate program GRK 1322 Micro Energy Harvesting at IMTEK, University of Freiburg, funded by the German Research Foundation (DFG). Special thanks to Daniela Ohnemus for the PCB preparations and Uwe Burzlaff for antenna range measurements.

References

1. Fischer, J. NFC in cell phones: The new paradigm for an interactive world [Near-Field Communications]. *IEEE Commun. Mag.* **2009**, *47*, 22–28.
2. Shameli, A.; Safarian, A.; Rofougaran, A.; Rofougaran, M.; de Flaviis, F. Power harvester design for passive UHF RFID tag using a voltage boosting technique. *IEEE Trans. Microw. Theor. Tech.* **2007**, *55*, 1089–1097.
3. Karthaus, U.; Fischer, M. Fully integrated passive UHF RFID transponder IC with 16.7-µW minimum RF input power. *IEEE J. Solid-State Circuits* **2003**, *38*, 1602–1608.
4. Lee, S.H.; Jin, I.S. Interoperation of an UHF RFID reader and a TCP/IP device via wired and wireless links. *Sensors* **2011**, *11*, 10664–10674.
5. Mandal, S.; Sarpeshkar, R. Low-power CMOS rectifier design for RFID applications. *IEEE Trans. Circuits Syst. I: Regul. Papers* **2007**, *54*, 1177–1188.
6. Yu, H.; Bashirullah, R. A Low Power ASK Clock and Data Recovery Circuit for Wireless Implantable Electronics. In *Proceedings of IEEE Custom Integrated Circuits Conference, CICC '06*, San Jose, CA, USA, September 2006; pp. 249–252.
7. Hannan, M.A.; Abbas, S.M.; Samad, S.A.; Hussain, A. Modulation techniques for biomedical implanted devices and their challenges. *Sensors* **2011**, *11*, 297–319.

8. Cho, N.; Song, S.-J.; Lee, J.Y.; Kim, S.; Kim, S.; Yoo, H.-J. A 8-µW, 0.3-mm2 RF-powered Transponder with Temperature Sensor for Wireless Environmental Monitoring. In *Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS*, Kobe, Japan, May 2005; Volume 5, pp. 4763–4766.
9. Brown, W.C. The history of power transmission by radio waves. *IEEE Trans. Microw. Theor. Tech.* **1984**, *32*, 1230–1242.
10. Mickle, M.H.; Mi, M.; Mats, L.; Capelli, C.; Swift, H. Powering autonomous cubic-millimeter devices. *IEEE Antennas Propag. Mag.* **2006**, *48*, 11–21.
11. McSpadden, J.O.; Chang, K. A Dual Polarized Circular Patch Rectifying Antenna at 2.45 GHz for Microwave Power Conversion and Detection. In *Proceedings of IEEE MTT-S International Microwave Symposium Digest*, San Diego, CA, USA, May 1994; Volume 3, pp. 1749–1752.
12. Sample, A.; Smith, J.R. Experimental Results with Two Wireless Power Transfer Systems. In *Proceedings of the 4th International Conference on Radio and Wireless Symposium*, San Diego, CA, USA, January 2009; pp. 16–18.
13. Umeda, T.; Yoshida, H.; Sekine, S.; Fujita, Y.; Suzuki, T.; Otaka, S. A 950-MHz rectifier circuit for sensor network tags with 10-m distance. *IEEE J. Solid-State Circuits* **2006**, *41*, 35–41.
14. Le, T.; Mayaram, K.; Fiez, T. Efficient far-field radio frequency energy harvesting for passively powered sensor networks. *IEEE J. Solid-State Circuits* **2008**, *43*, 1287–1302.
15. Yuan, F. *CMOS Circuits for Passive Wireless Microsystems*; Springer: New York, NY, USA, 2011.
16. Zbitou, J.; Latrach, M.; Toutain, S. Hybrid rectenna and monolithic integrated zero-bias microwave rectifier. *IEEE Trans. Microw. Theor. Tech.* **2006**, *54*, 147–152.
17. Ungan, T.; Le Polozec, X.; Walker, W.; Reindl, L. RF Energy Harvesting Design Using High Q Resonators. In *Proceedings of IEEE MTT-S International Microwave Workshop on Wireless Sensing, Local Positioning and RFID, IMWS 2009*, Cavtat, Croatia, September 2009; pp. 1–4.
18. Ungan, T.; Freunek, M.; Muller, M.; Walker, W.D.; Reindl, L.M. Wireless Energy Transmission Using Electrically Small Antennas. In *Proceedings of IEEE Radio and Wireless Symposium, RWS '09*, San Diego, CA, USA, January 2009; pp. 526–529.
19. Boquete, L.; Rodríguez-Ascariz, J.M.; Barea, R.; Cantos, J.; Miguel-Jiménez, J.M.; Ortega, S.; Data acquisition, analysis and transmission platform for a pay-as-you-drive system. *Sensors* **2010**, *10*, 5395–5408.
20. Heikkinen, J.; Salonen, P.; Kivikoski, M. Planar Rectennas for 2.45 GHz Wireless Power Transfer. In *Proceedings of Radio and Wireless Conference, RAWCON 2000*, Denver, CO, USA, 10–13 September 2000; pp. 63–66.
21. Akkermans, J.A.G.; van Beurden, M.C.; Doodeman, G.J.N.; Visser, H.J. Analytical models for low-power rectenna design. *IEEE Antenn. Wireless Propag. Lett.* **2005**, *4*, 187–190.
22. Hagerty, J.A.; Helmbrecht, F.B.; McCalpin, W.H.; Zane, R.; Popovic, Z.B. Recycling ambient microwave energy with broad-band rectenna arrays. *IEEE Trans. Microw. Theor. Tech.* **2004**, *52*, 1014–1024.
23. Harb, A. Energy harvesting: State-of-the-art. *Renew. Energ.* **2011**, *36*, 2641–2654.
24. Vullers, R.J.M.; van Schaijk, R.; Doms, I.; van Hoof, C.; Mertens, R. Micropower energy harvesting. *Solid-State Electron.* **2009**, *53*, 684–693.

25. Sah, C.-T. *Fundamentals of Solid-State Electronics*; World Scientific: Singapore, Singapore, 1991.
26. Wetenkamp, S. Comparison of Single Diode *vs.* Dual Diode Detectors for Microwave Power Detection. In *Proceedings of IEEE MTT-S International Microwave Symposium Digest*, Boston, MA, USA, June 1983; pp. 361–363.
27. Cardoso, A.J.; Schneider, M.C.; Montoro, C.G. Design of Very Low Voltage CMOS Rectifier Circuits. In *Proceedings of Circuits and Systems for Medical and Environmental Applications Workshop (CASME)*, Merida, Mexico, December 2010; pp. 1–4.
28. Cardoso, A.J.; de Carli, L.G.; Galup-Montoro, C.; Schneider, M.C. Analysis of the Rectifier Circuit Valid Down to Its Low-Voltage Limit. *IEEE Trans. Circuits Syst. I: Regul. Papers* **2012**, *59*, 106–112.
29. Data sheet HSMS-285x. Avago Technologies. Available online: http://www.avagotech.com (accessed on 28 September 2012).
30. Data sheet HSMS-286x. Avago Technologies. Available online: http://www.avagotech.com (accessed on 28 September 2012).
31. Hsu, T.S. Low-frequency excess noise in metal—Silicon Schottky barrier diodes. *IEEE Trans. Electron Devices* **1970**, *17*, 496–506.
32. Hastas, N.A.; Dimitriadis, C.A.; Dozsa, L.; Gombia, E.; Amighetti, S.; Frigeri, P. Low frequency noise of GaAs Schottky diodes with embedded InAs quantum layer and self-assembled quantum dots. *J. Appl. Phys.* **2003**, *93*, 3990.
33. Gomila, G.; Reggiani, L.; Rubí, J.M. Shot-noise suppression in Schottky barrier diodes. *J. Appl. Phys.* **2000**, *88*, 3079.
34. Kleinpenning, T.G.M. Low-frequency noise in Schottky barrier diodes. *Solid-State Electron.* **1979**, *22*, 121–128.
35. Lee, T.H. *The Design of CMOS Radio-Frequency Integrated Circuits*; Cambridge University Press: Cambridge, UK, 2004.
36. Advanced Design System. Agilent Technologies: Santa Clara, CA, USA.
37. Bowick, C.; Blyler, J.; Ajluni, C.J. *RF Circuit Design*; Newnes/Elsevier: Amsterdam, The Netherlands, 2008.
38. Joe, J.; Chia, M.; Marath, A.; Ang, C. Zero Bias Schottky Diode Model for Low Power, Moderate Current Rectenna. In *Proceedings of Defence Electronics Technology Seminar (DETS 97)*, Singapore, Singapore, November 1997; pp. 1–5.
39. Visser, H.J.; Vullers, R.J.M.; Veld, B.O.h.; Pop, V. Remote RF Battery Charging. In *Proceedings of Power MEMS 2010*, Leuven, Belgium, December 2010; pp. 37–40.
40. Nimo, A.; Grgic, D.; Reindl, L.M. Impedance optimization of wireless electromagnetic energy harvester for maximum output efficiency at μW input power. *Proc. SPIE* **2012**, *8341*, 83410W:1–83410W:14.
41. Nimo, A.; Grgic, D.; Reindl, L.M. Electrically small Planner Antenna for Compact Electromagnetic (EM) Wireless Energy Harvester. In *Proceedings of Power MEMS 2011*, Seoul, Korea, November 2011; pp. 310–313.
42. RT-duroid-5870–5880-Data-Sheet; Rogers Corporation: Rogers, CT, USA, 2011.

43. *HFSS*; Ansys Corporation. Ansys, Inc: Canonsburg, PA, USA, 2005.
44. Friis, H.T. A Note on a Simple Transmission Formula. *Proc. IRE* **1946**, *34*, 254–256.
45. Kvaksrud, T.I. *Range Measurements in an Open Field Environment*, *Design Note DN018*; Texas Instrum. Incorporated: Dallas, TX, USA, 2008.
46. *LabView*; National Instruments: Austin, TX, USA, 2009.

Appendix A: Measuring Setup for RF Rectifier Efficiency and Voltage Sensitivity

The measuring setup is as shown in Figure A1.

Figure A1. RF to DC Power converter characterization setup.

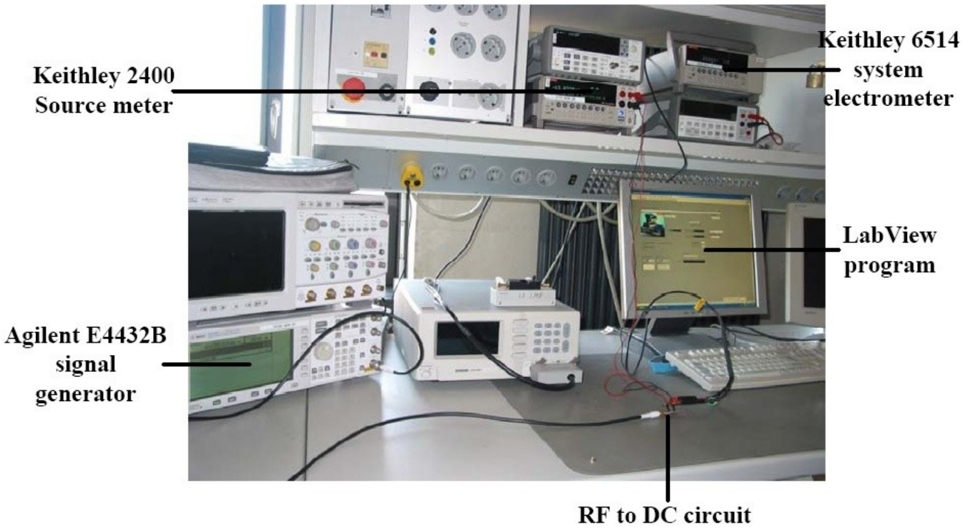

The RF to DC circuit efficiency and voltage sensitivity measurements were made with a Keithley 2400 source meter and Keithley 6514 system electrometer with an Agilent E4432B signal generator providing 50 Ω RF signal into the circuit board.

The closed circuit current drawn by the RF to DC power converter (*without load*) from the generator is first determined by the Keithley 2400 source meter; then starting from this current, the value of the current is decreased at set intervals to creates virtual load resistances to the circuit for up to a lowest current of 0.1 µA. The 6514 system electrometer is used to measure the output voltage. The number of data point is set through LabView [46] as well as the measurements. Additionally open circuit voltage or at specific loads and frequency sweep can be made through the LabView program. At −40 dBm input power and below, the detected voltages and currents were difficult to measure accurately with the measuring setup; hence measurements were made up to a minimum of -35 dBm input power. The circuit layout for the efficiency and voltage sensitivity measurements is schematically shown in Figure A2.

Figure A2. RF to DC power converter characterization circuit.

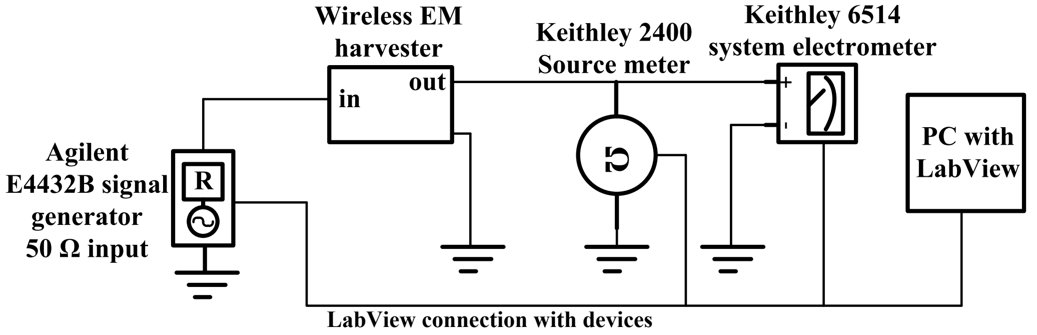
Wireless EM harvester
in
out
Keithley 2400 Source meter
Keithley 6514 system electrometer
+
-
PC with LabView
Agilent E4432B signal generator 50 Ω input
R
LabView connection with devices

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

Antenna Research

STANDARD ANTENNA TERMS AND RELATED FORMULAS

The "IEEE Standard Definitions of Terms for Antennas" (IEEE STD-145) represents a consistent and comprehensive vocabulary suited for the effective communication and understanding of antenna theory. General use of these definitions of terms would eliminate much of the wide-spread inconsistency concerning antenna characteristics, particularly with regard to the basic parameters of gain, beamwidth, polarization and efficiency. For convenience, IEEE antenna terms of general interest are listed here. Wherever these terms appear in this catalog, the definitions given below apply. Other commonly used terms, not covered by the IEEE standard, are shown with an "*."

ANTENNA APERTURE. A surface, near or on an antenna, on which it is convenient to make assumptions regarding the field values for the purpose of computing fields at external points. **Note:** The aperture is often taken as that portion of a plane surface near the antenna, perpendicular to the direction of maximum radiation, through which the major part of the radiation passes.

ANTENNA EFFICIENCY OF APERTURE - TYPE ANTENNA. For an antenna with a specified planar aperture, the ratio of the maximum effective area of the antenna to the aperture area.

*** ANTENNA FACTOR.** That quantity by which the voltage developed across the output of an antenna is related to the incident field strength in which the antenna is immersed. **Note:** Applicable to low frequency antennas and usually refers to a 50 ohm output.

$$\mathrm{AFE\ (dB\ m^{-1}) = E\ (dB\ V/m)\ -\ V\ (dB\ V)}$$

AFE = Electric Field Antenna Factor
E = Electric Field Strength at antenna
V = Voltage at terminals of antenna

$$\mathrm{AFH\ (dB\ AV^{-1}\ m^{-1}) = H\ (dB\ A/m) - V(db\ V)}$$

AFH = Magnetic Field Antenna Factor
H = Magnetic Field Strength at antenna
V = Voltage at terminals of antenna

$$\mathrm{AFE\ (dB\ m^{-1})\ =\ AFH\ (dB\ AV^{-1}\ m^{-1}) + 51.53}$$

for a plane wave in free space.

$$\mathrm{AFB\ (dB\frac{pT}{\mu V})\ =\ B\ (dBpT)\ -\ V_o\ (dB\mu V)}$$

AFB = Magnetic flux Antenna Factor
B = Magnetic flux at the antenna
pT: picoTesla units
V = Voltage at the terminals of the antenna

$$\mathrm{AFB\ (dB\frac{pT}{\mu V}) = AFH\ (dB\frac{A}{Vm})\ + 2}$$

APERTURE ILLUMINATION. The field over the aperture as described by amplitude, phase, and polarization distributions.

APERTURE ILLUMINATION EFFICIENCY. For a planar antenna aperture, the ratio of its directivity to the directivity obtained when the aperture illumination is uniform.

BEAM. The major lobe of the radiation pattern.

CIRCULAR POLARIZATION. It may be either right hand circular polarization (RHCP) or left hand circular polarization (LHCP). The sense of polarization is determined by observation of the direction of rotation of the electric field vector from a point behind the source, RHCP and LHCP correspond to clockwise and counter-clockwise respectively. **Note:** RHCP transmit requires a like polarization to receive.

CO-POLARIZED. The polarization which the antenna is intended to radiate or receive. Also "like polarization".

*** CROSS POLARIZATION DISCRIMINATION (XPD).** Cross polarization discrimination is the measure of the antenna's ability to differentiate between the vertical and the horizontal polarization of an antenna. This difference, shown in relative signal level, is indicated on directional pattern envelopes (DPE's).

STANDARD ANTENNA TERMS AND RELATED FORMULAS

* **DIRECTIONAL PATTERN ENVELOPES (DPE'S).** In accordance with standard practice, radiation characteristics in any given plane of polarization are measured and plotted using 360-degree polar coordinate systems. The resultant Directional Pattern Envelope is the smoothed composite of all these measurements. The purpose of these DPE's is to emphasize the worst composite condition.

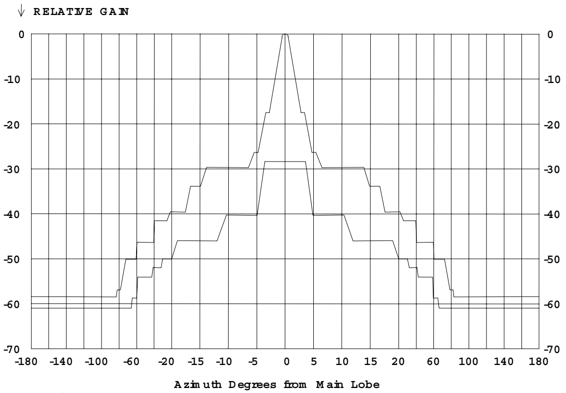

DIRECTIVE GAIN. In a given direction, 4 times the ratio of the radiation intensity in that direction to the total power radiated by the antenna.

DIRECTIVITY. The value of the directive gain in the direction of its maximum value.

EFFECTIVE AREA OF AN ANTENNA. In a given direction, the ratio of power available at the terminals of a receiving antenna to the power per unit area of a plane wave incident on the antenna from that direction, polarized coincident with the polarization that the antenna would radiate.

FAR FIELD REGION. That region of the field of an antenna where the angular field distribution is essentially independent of the distance from a specified point in the antenna region.

* **FRONT-TO-BACK RATIO**. The ratio of the maximum directivity of an antenna to its directivity in a specified rearward direction.

* **Gain, dBi**. The gain expressed in decibels relative to an isotropic radiator that is linearly polarized.

$$G(dBi) = 10\log(G) \qquad G = 10^{\frac{G(dBi)}{10}}$$

* **GAIN, dBic**. The gain expressed in decibels relative to an isotropic radiator that is circularly polarized.

HALF-POWER BEAMWIDTH. In plane containing the direction of the maximum of a beam, the angle between the directions in which the radiation intensity is one half the maximum value of the beam.

HALF-WAVE DIPOLE. A half wavelength antenna, center fed so as to have equal current distribution in both halves. Mounted vertically, it has a doughnut shaped pattern, circular in the horizontal plane. It is an antenna that can be constructed. It has some inherent losses. When used as a gain reference, the half-wave dipole has a power gain of about 1.7 dBi.

* **ISOLATION.** Refers to the ability of one port of a dual polarized feed to discriminate against a signal fed into the other port.

ISOTROPIC RADIATOR. A hypothetical antenna having equal radiation intensity in all directions. Note: An isotropic radiator represents a convenient reference for expressing the directive properties of actual antennas.

NEAR-FIELD REGION. The spherical region of space between the antenna and the far field region.

NULL. The region of a radiation pattern, either computed or measured, where the amplitude goes through a minimum value. Note: (1) It represents the angular position where the phase or the far field pattern crosses the zero axis if the pattern is plotted as a phasor instead of a scalar value. Note: (2) The region outside the main beam of a directive antenna pattern consists of a series of minor lobes separated by nulls.

PARALLEL POLARIZATION. The condition where the electric vector is parallel to the local conducting surface. Note: Over the earth, this is usually referred to as being horizontal polarization.

PHASE CENTER. The location of a point associated with an antenna such that, if it is taken as the center of a sphere whose radius extends into the far-field, the phase of a given component over the surface of that radiation sphere is essentially constant, at least over the portion of the sphere where the radiation is significant.

STANDARD ANTENNA TERMS AND RELATED FORMULAS

POLARIZATION. The polarization of an antenna is defined as the polarization of the electromagnetic wave as described by the shape and orientation of an ellipse, which is the locus of the extremity of the field vector, and the sense in which the ellipse is traversed with time. The elliptical locus is called the polarization ellipse and the wave is said to elliptically polarized. Circular polarization and linear polarization are degenerate cases of elliptical polarization.

POWER DENSITY AT A POINT

$$S_{av} = \frac{GP_t}{4\pi r^2}$$

S_{av} = Time average power density in W/m^2
P_t = Power transmitted in watts
G = Antenna gain relative to an isotrope
r = Distance from antenna to point in meters

POWER DENSITY TO VOLTS/METER IN FREE SPACE

$$E^2 \text{ (V / m)} = 377\ S_{av} \text{ (W / m}^2\text{)}$$

$$S_{av} \text{ (W / m}^2\text{)} = E^2 \text{ (V / m) / 377}$$

$$1 \text{ V / m} = 2.65 \text{ mW / m}^2$$

POWER GAIN. In a given direction, 4 times the ratio of the radiation intensity in that direction to the net power accepted by the antenna from the connected transmitter. **Note:** (1) When the direction is not stated, the power gain is usually taken to be the power gain in the direction of its maximum value. (2) Power gain does not include reflection losses arising from mismatch of impedance.

POWER GAIN IN PHYSICAL MEDIA. In a given direction and at a given point in the far field the ratio of the power flux per unit area from an antenna to the power flux per unit area from an isotropic radiator at a specified location with the same power input as the subject antenna.

Note: The isotropic radiator must be within the smallest sphere containing the antenna. Suggested locations are antenna terminals and points of symmetry, if such exist.

POWER GAIN REFERRED TO A SPECIFIED POLARIZATION. The power gain of an antenna, reduced by the ratio of that portion of the radiation intensity corresponding to the specified polarization to the radiation intensity.

POWER TRANSMISSION FORMULAS

$$P_r = P_t \frac{G_t G_r \lambda^2}{(4\pi r)^2}$$

$$P_r \text{ (dB W)} = P_t \text{ (dB W)} + G_t \text{ (dBi)} + G_r \text{ (dBi)} - 20 \log r - 20 \log f + 27.56$$

$$P_r \text{ (dB W)} = P_t \text{ (dB W)} - AFE_t \text{ (dB m}^{-1}\text{)} - AFE \text{ (dB m}^{-1}\text{)} - 20 \log r + 20 \log f - 32$$

P_r = Power received
P_t = Power transmitted
G_r = Gain of receiving antenna
G_t = Gain of transmitting antenna
f = Frequency in MHz, λ = Wavelength
r = distance between antennas in meters
AFE_r = AFE of receiving antenna
AFE_t = AFE of transmitting antenna

RADIATOR. Any antenna or radiation element that is a discrete physical and functional entity.

RADIATION, ELECTROMAGNETIC. The emission of energy in the form of electromagnetic waves.

RADIATION INTENSITY. In a given direction, the power radiated from an antenna per unit solid angle.

RADIATION LOBE. A portion of the radiation pattern bounded by regions of relatively weak radiation intensity

RADIATION PATTERN (ANTENNA PATTERN). A graphical representation of radiation properties of the antenna as a function of space coordinates. **Note:** (1) In the usual case the radiation pattern is determined in the far-field region and is represented as a function of directional coordinates. (2) Radiation properties include power flux density, field strength, phase, and polarization.

*** RADIATION RESISTANCE OF AN ELECTRICALLY SMALL LOOP ANTENNA.** The resistive component of an antenna's input impedance that results from the coupling of the antenna to its environment. This resistance dissipates the power that is actually radiated from the antenna.

$$R_r = 20\,(2\pi/\lambda)^4\, n^2 A^2 \text{ ohms}$$

n = number of turns
A = area of the loop

REALIZED GAIN. The power gain of an antenna in its environment, reduced by the losses due to the mismatch of the antenna input impedance to a specified impedance.

*** REALIZED RADIATOR EFFICIENCY.** The efficiency of an antenna in its environment reduced by all losses suffered by it, including: ohmic losses, mismatch losses, feedline transmission losses, and radome losses.

RELATIVE POWER GAIN. The ratio of the power gain in a given direction to the power gain of a reference antenna in its reference direction. **Note:** Common reference antennas are half-wave dipoles, electric dipoles, magnetic dipoles, monopoles, and calibrated horn antennas.

RETURN LOSS. The reflection coefficient of a mismatch expressed in decibels. **Note:** Modern swept VSWR techniques actually sense the reflected component which is normalized to the forward component to yield return loss. A 2:1 VSWR is equivalent to 9.5 dB return loss.

VSWR. The voltage standing wave ratio of a component such as an antenna. It is referred to the characteristic impedance of the transmission line being used. **Note:** The most common characteristic impedance is 50 ohms, but 75 and 300 ohms are frequently used in coaxial or twin lines for VHF, UHF applications.

NOTES:

THE HANDYMAN'S GUIDE TO OSCILLOSCOPES (Part 2 of 2)

by Paul Harden, NA5N

Making some advanced measurements with your Oscilloscope

Print as .pdf file
4 pages
8½ x 11 or A4

In Part 1, oscilloscope operation was covered for making basic voltage, time and frequency measurements. In this part, we'll continue with some more advanced uses of a scope, and in particular, how to use a scope for testing and troubleshooting ham radio QRP transceivers in the homebrewer's workshop.

Receiver Filter Bandwidth.

This procedure uses a scope (a DVM can be used with less accuracy) for determining the overall filter bandwidth (or selectivity) of a receiver. It is basically measured by plotting output voltage vs. audio frequency to construct a picture of the filter response.

Connect scope to the receiver audio output (speaker or phone jack); measurement will be based on peak-to-peak voltages (Vpp) on a scope, or rms voltage (Vrms) on a DVM.

Using a signal generator, set the frequency for the band of interest on your radio. For example, on a general coverage shortwave receiver, you might set it for 10 MHz (top end of the 31M band), or perhaps to 7.040 MHz on a 40M ham radio receiver/transceiver. Tune the receiver to the signal generator signal. If you don't have a signal generator, you can also tune to a steady carrier or station to produce a hetrodyne audio "pitch." Tune in the signal to the pitch that causes the maximum peak-to-peak display. Adjust the scope and volume control to produce a 2Vpp display (4 divisions). This is the peak response of the overall filtering stages as shown in **Fig. 13**.

Now determine the audio frequency at this peak response by measuring the time period between cycles and covert to frequency. In the example to the right, the period of one cycle is 1.7mS, which is an audio tone of 750Hz (1/.0017sec). A frequency counter on the output can also be used.

Next, tune the receiver such that the sidetone pitch goes UP in frequency and the peak-to-peak signal will decrease in magnitude. Tune to the point where the signal is exactly 1Vpp on the scope. See **Fig. 12**.

This is the –6dB point of the high end of the filter (20log 1v/2v =–6dB). Determine the frequency of the audio pitch as before. In the example, this is 1100 Hz. Record the data.

From these two data points, the –6dB bandwidth can be estimated. The bandwidth from the filter peak (750 Hz) to the –6dB point (1100 Hz) is 350 Hz. The bandwidth (BW) between the two –6dB points is usually twice this value, or 700 Hz. A filter with a –6dB BW of 700Hz is a mediocre filter for CW reception, and way too narrow for SSB or AM.

Of course you can determine the exact –6dB BW by tuning the receiver back to the 2Vpp peak response, and continue tuning DOWNward in frequency until the audio is again exactly 1Vpp. Determine this frequency and record. In this example, it should occur around 400 Hz if the filter shape is symmetrical.

Plot these three data points on a sheet of graph paper as shown in **Fig. 15** to construct the filter shape. Return to the upper –6dB point (1100Hz in the example) and continue tuning upwards in audio pitch, recording the frequency at 0.5v (–12 dB), 0.25v (–18 dB), 125mV (–24 dB), etc. Everytime you "halve the voltage," it is a 6 dB change. The more points you collect, the more accurate your filter response plot will be.

Fig. 13 – Displaying Filter Peak Response

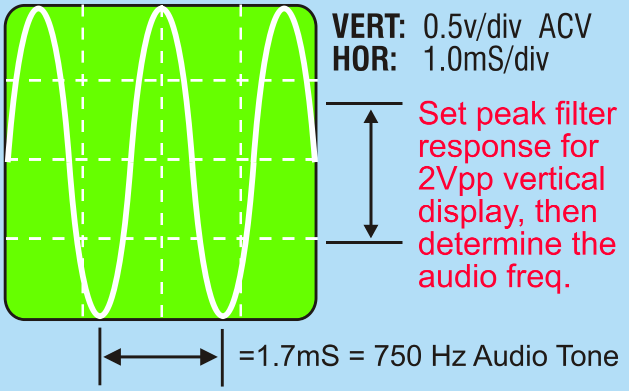
VERT: 0.5v/div ACV
HOR: 1.0mS/div
Set peak filter response for 2Vpp vertical display, then determine the audio freq.
=1.7mS = 750 Hz Audio Tone

Fig. 14 – Displaying –6dB Frequency

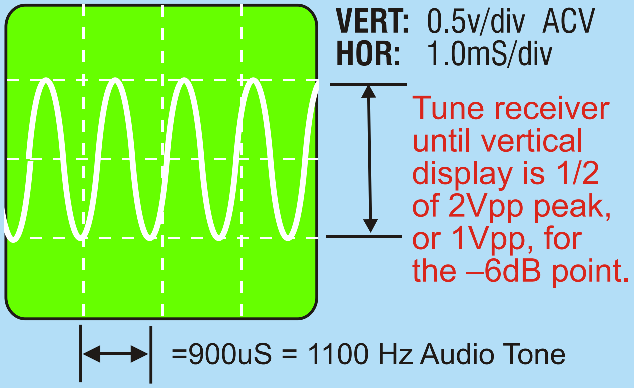
VERT: 0.5v/div ACV
HOR: 1.0mS/div
Tune receiver until vertical display is 1/2 of 2Vpp peak, or 1Vpp, for the –6dB point.
=900uS = 1100 Hz Audio Tone

Of interest to proper rig alignment is to repeat the above using the output of the product detector. To maximize the effectiveness of the filtering, the receive offset frequency set by the BFO should be adjusted to the same frequency as the peak frequency response of the audio. In this example, with the peak audio response occuring at 750Hz, if your BFO is set for a sidetone frequency of 700 Hz, you are loosing 2–3 dB, since this is in your filter skirt. This is shown on the response plot in **Fig. 15** by the dashed lines. By adjusting your BFO for a sidetone frequency of 750 Hz, you will pick up 3-4dB of overall gain in your receiver, plus increase the selectivity a bit as well. Why? Because nearby stations, such as one at 800–900Hz tone, could actually be louder than the 700Hz tone signal you are trying to copy, since the gain of the receiver is greater at those tones than at 700Hz, as shown in the plot.

Fig. 15 – Plotting Audio Response

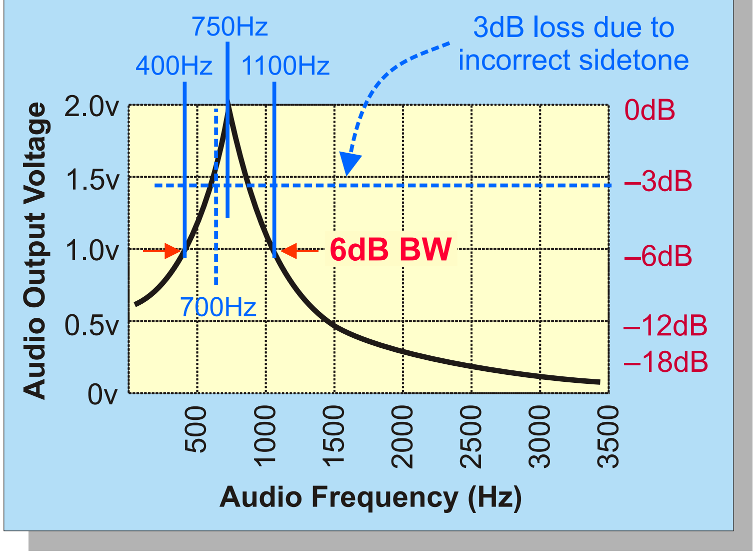

Opposite Sideband Rejection

A superhet receiver is supposed to pass just one sideband and reject the other. Poor opposite sideband rejection could indicate the crystals in your IF filter are not well matched or other problems. It is measured almost identical to plotting the filter response just described. First, you tune the receiver to the test signal to find the peak response frequency, or 750Hz in this example. Set the scope display for 2Vpp. Now tune downward in audio pitch, passing zero-beat, and continue tuning and you should hear the test signal, much weaker, now rising again in tone. This is the *opposite sideband*. Measure the peak-to-peak voltage, if you can. For example, say it is 80mVpp, as shown in **Fig. 16**. Calculate the opposite sideband rejection by:

$$\text{rejection} = 20\log\frac{80\text{mV}}{2.0\text{v}} = \frac{80\text{mV}}{2000\text{mV}} = -28\text{dB}$$

If you can't hear the opposite sideband, then obviously you have excellent filter rejection. If you can just barely hear it, you may have to increase the sensitivity of your scope (set vertical gain to 20 or 50 mV/div). In this example, –28dB rejection of the opposite sideband is quite good. A –30dB rejection means the opposite sideband is only 1/1000th of the desired sideband, a very suitable attenuation of the opposite sideband.

Fig. 16
Calculating Opposite Sideband Rejection

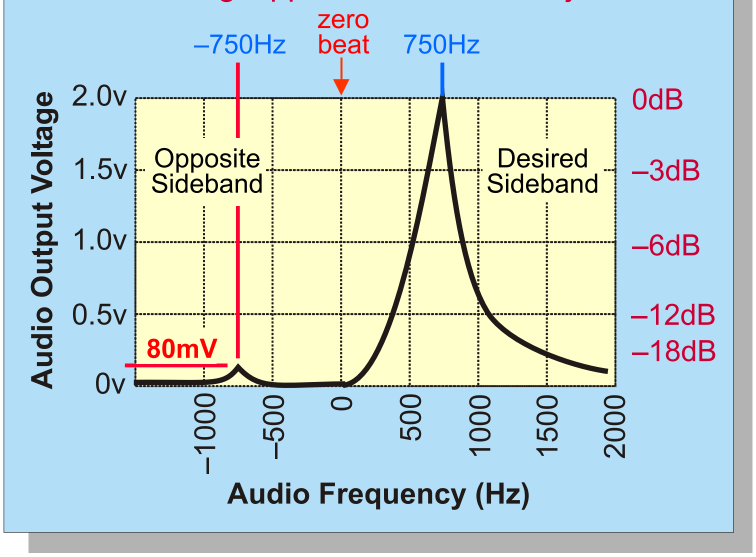

These tests are important to perform on your rig for documenting it's current performance, and repeated periodically to detect unfavorable changes or for troubleshooting when a problem is evident. For homebrewing, these tests can allow you to evaluate different circuits or when experimenting with different components or part values.

Oscillator Phase Noise

When homebrewing a basic oscillator circuit, such as a VFO, looking at the oscillator output on a scope can reveal several problems. One is to check for excessive phase noise. Phase noise is small variations of the oscillator frequency that causes power in the close-in sidebands, usually measured on laboratory equipment within 100KHz, or even within 10 KHz.

To check for oscillator phase noise, connect the oscilloscope to the oscillator output, loading the output of the oscillator as little possible. Most scopes have sufficiently high input impedances where this shouldn't be a problem, but some cheaper scopes can load an oscillator circuit. If you suspect your scope is loading the oscillator, couple the scope to the circuit with a small value capacitor, less than 20pF.

Display 2–3 cycles of the oscillator output as shown on the scope display shown in **Fig. 17**. Properly focus the scope and carefully observe if the waveform appears in focus at the peaks, but slightly out-of-focus at the zero-crossing points, that is, on the rising and falling edges of the sine wave.

If it appears out-of-focus, this is excessive phase noise jittering the signal and "smearing" the waveform along the time (horizontal) axis. Extreme phase noise may show 2-3 sine waves very close to each other, as shown in the exagerated waveform to the right — assuming you have your scope properly triggered.

Fig. 17 – Checking Oscillator Phase Noise

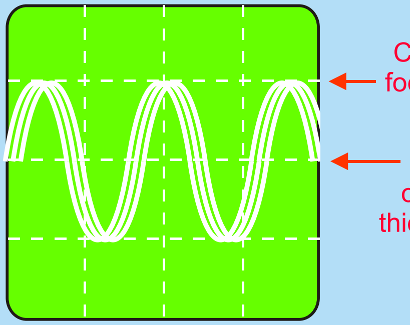

Check for a clean, focused signal here

Check for an out-of-focus or thicker signal here

Phase noise is random, instantaneous changes in the oscillator frequency that smears the display. If you can see this on a scope, the phase noise is excessive! If you can't see it, it doesn't mean the oscillator has no phase noise (all oscillators have some phase noise), it just means it is not excessive enough to see on a scope. A scope is not a good instrument for checking phase noise, but for homebrew circuits, it is a check to ensure you do not have a serious oscillator problem.

Excessive oscillator phase noise in receivers can cause IMD products and noise in the audio range at the output of the mixer(s), including the product detector. In a transmitter, excessive oscillator phase noise will put power in the close-in sidebands of your carrier, not only wasted power for lower transmitter efficiency, but may produce strange sounds (buzzing or chirping) to the receiving station. A few causes of phase noise are excessive current in the oscillator transistor, low-Q coil(s), high dissipation in the tuning caps or poor power supply filtering at the oscillator frequency.

AM (Amplitude) Noise

Another oscillator problem may be AM noise, or amplitude modulated noise. It is an opposite effect on a scope when displaying the oscillator output — the sine wave appears out-of-focus or thicker at the peaks, and in-focus elsewhere as shown in **Fig. 18**.

Fig. 18 – Checking for AM Oscillator Noise

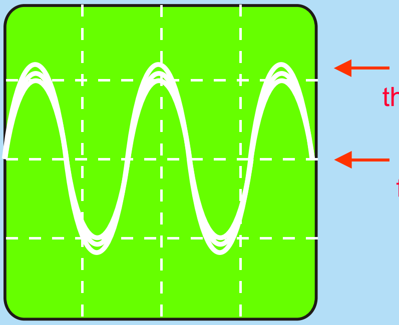

Check for an out-of-focus or thicker signal here

Check for a clean, focused signal here

If you detect AM noise, slow down the scope's sweep rate to the audio frequencies or slower to see if you can notice a lower frequency component. A common cause of AM oscillator noise is 60 Hz from the power mains leaking into the circuits. This is particularly true if using a power supply off of 120v 60 Hz, or sometimes it can be due to the AC lighting above your head! If the AM noise seems to be at the same frequency as the audio output tone, it means audio is getting into the Vcc bus, likely due to poor bypass filtering at the audio amplifier stages (particular if using an LM386 or similar).

If you can't find a low frequency component, the AM noise may be random, which may indicate poor voltage regulation, a noisy voltage regulator, or perhaps a circuit in VHF oscillation. If the AM noise seems to occur on key-down only on a transmitter, the transmit current may be loading the power supply, the voltage regulator is under-rated, or just simply loading the oscillator. In the case of loading the power supply or regulator(s), perhaps a separate voltage regulator or zener circuit should be used, dedicated for the oscillator(s). In the case of transmit loading, adding a buffer amplifier or emitter follower to isolate the load from the oscillator may help.

Much of this can be diagnosed also with the scope, by looking at the AC ripple on the DC power lines. You should have less than 50mV of any AC component on the 8-12v DC wiring, whether 60Hz, audio or RF. If >50mV, then additional low or high frequency filtering on the DC power is needed.

Monitoring Transmitter RF Output

RF Power (in watts) is E^2/R, where E is in rms and R=50 . The voltage displayed on the scope (peak-to-peak) must be converted to rms by Erms=.707(Epp/2). To measure properly, the transmitter should be on a 50 dummy load using the normal hi-Z scope input. If your scope has a selectable "50 "input, it can be the transmitter dummy load directly, providing it can tolerate the 50Vpp input. (Always know what the maximum p-p input voltage your oscope will tolerate. It is often stated on your scope at the vertical channel inputs. 50Vpp to 100Vpp are typical).

Measuring Transmitter RF Output Power

Figure 19 shows how you can use an oscilloscope to fairly accurately measure output power from a low power transmitter (generally 10W or less with a 1X probe and 20-30W with a 10X probe). Connect transmitter output to a 50 dummy load. Connect the scope lead to the transmitter output (after the low pass filters) or directly to the dummy load. Do not connect the scope leads to the collector or drain of the final PA transistor. The displayed voltage in this case will be erroneous.

In the example, the output transmit voltage across the dummy load measures 24Vpp. Convert this voltage to Vrms, then use the equation (Vrms squared divided by the load resistance) to calculate the power in watts. In the example, 24Vpp, the output power is 1.4 watts. A 5W QRP transmitter should produce about 45Vpp.

The accuracy of your power measurements depend upon the condition of your oscilloscope:

1) Ensure your scope's vertical sensitivity (volts/division) is properly calibrated (see Part I)
2) Ensure your vertical sensitivity is properly calibrated with the 1X or 10X probe you are using.
3) Ensure that the transmitter frequency is well within the bandwidth of your scope. A 100 MHz scope should give reliable readings on all HF bands. A 50 MHz scope should be reliable on all HF bands except slight errors at 28MHz.

Fig. 19 –Oscilloscope as a Power Meter

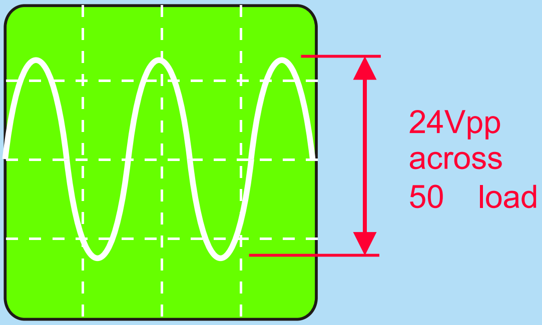

1. Measure output voltage. 24Vpp
2. Convert Vpp to Vrms
   Vrms=.707(Vpp/2) 8.5Vrms
3. Calculate power in watts

$$\frac{Vrms^2}{R} = \frac{8.5v^2}{50} = 1.4W$$

Checking Transmitter Purity

Phase Noise of the transmitter is measured identically to the phase noise checks on page 2–3. With the scope properly triggered and focused, you are looking for an out-of-focus "fuzziness" around zero crossing. If it appears phase noise exists, the fault is generally at the beginning of the transmit chain. That is, seldom does the PA transistor add phase noise; it's merely amplifying what it is given. Phase noise on the transmitter output is more likely due to the transmit oscillator, or the transmit mixer if used.

Harmonic Power can also be detected on a scope while looking at the transmitter RF output power. In **Fig. 20**, notice the "dips" or the two "peaks" on the top and bottom of each sine wave. This is caused by excessive 2nd harmonic output power. The rule-of-thumb is – if you can see any 2nd harmonic power (a dip or flattening at the peaks), then you are right at or exceeding the –30dBc FCC harmonic specification. A clean sine wave implies FCC compliance, providing the 2nd harmonic is within your scope's bandwidth. Obviously, a spectrum analyzer should be used for accurate harmonic evaluation, but a scope can be used to indicate if you have a problem. If you build your own transmitter, it is your responsibility to ensure it is compliant. An approximate method of estimating 2nd harmonic attenuation is shown.

–30dBc means the 2nd harmonic power is 1/1000th of the fundamental power. For a 5W transmitter, this means the 2nd harmonic power should be 5mW or less, or about 1Vpp. In displaying a 5W signal on a scope (45Vpp), it would be difficult to resolve much less than a 1Vpp dip or ripple.

Fig. 20 –Checking 2nd Harmonic Power

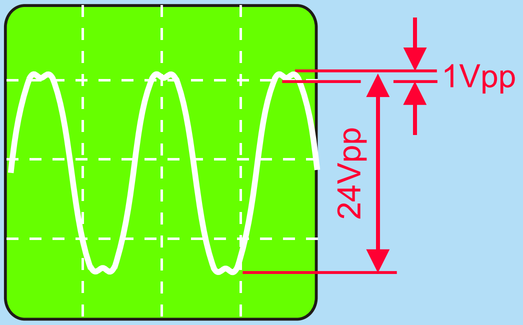

1. Measure output voltage. 24Vpp
2. Measure ripple voltage. 1Vpp
3. Calculate dBc of 2nd harmonic ripple

$$20\log\frac{V2}{V1} = 20\log\frac{1v}{24v} = -27.6\ \text{dBc}$$

or, not quite FCC compliant!

Conclusion. This concludes the NA5N "Handiman's Guide to Oscilloscopes." Hopefully, it has helped you to understand your oscilloscope to make simple measurements, and with some practice, perform the more advanced measurements discussed. An oscilloscope is a very powerful tool that is invaluable on the homebrewer's workbench. If there is a particular measurement you feel I omitted, please let me know and I will add it when time permits.

73, Paul Harden, NA5N (na5n@zianet.com or pharden@nrao.edu)

THE HANDYMAN'S GUIDE TO OSCILLOSCOPES (Part 1 of 2)

by Paul Harden, NA5N

Getting Acquainted with your Scope and making some measurements

Updated May 2004 with actual oscope waveforms

Print as .pdf file
5 pages
8½ x 11 or A4

When in the course of human events, it becomes necessary to look at neat signals floating around your radio, you go to a hamfest and buy that $50 o-scope. Now what? This two part article will attempt to explain basically how an oscilloscope works, operator functions, basic measurements, and some advanced applications. An o-scope is a powerful tool in any shack – even a real "cheapie" with limited bandwidth.

HOW AN OSCILLOSCOPE WORKS

A block diagram of a typical o-scope is shown in **Fig. 1**. The test probe usually plugs into the scope via a BNC connector, then passes through a switch to determine whether the input signal will be dc or ac coupled (to remove any dc component). Often this switch will have a "ground" position for setting the zero-volts reference. Next is the input attenuators. The vertical input amplifier is quite sensitive, designed for 20-50mV of input. For larger input voltages, the signal is applied to attenuators comprised of simple voltage dividers. This is the first area of concern for cheap o-scopes, as the input attenuators may not be very linear or accurate. For example, if you appy a 10Vpp signal on the 10v/division setting, the signal should be 1 division high. Switching to 1v/div., the signal should be 10 divisions (usually full-scale) high. If it is not exactly 10 divisions, the attenuator for that setting needs adjusting. Some scopes have internal adjustments for fine-tuning each attenuator setting.

Following the attenuators, the signal is applied to the first vertical amplifier, which converts the input to a differential signal. This differential signal is amplified up to high voltages for the oscilloscope deflecting plates – moving the beam up and down (in the vertical axis).

The sweep generator is usually a constant current source charging a capacitor to make a sawtooth waveform that eventually deflects the beam in the horizontal axis. The frequency of the sawtooth determines how fast the beam travels from the left to the right side of the tube, and is controlled by the sweep control, usually calibrated in seconds, milli-seconds or micro-seconds per division. This is the second area of concern for an oscilloscope – how linear the sawtooth waveform is generated. For example, a sawtooth with a nonlinear ramp will cause the signal displayed in the central portion of the tube to be expanded or compressed compared to the signal at the ends.

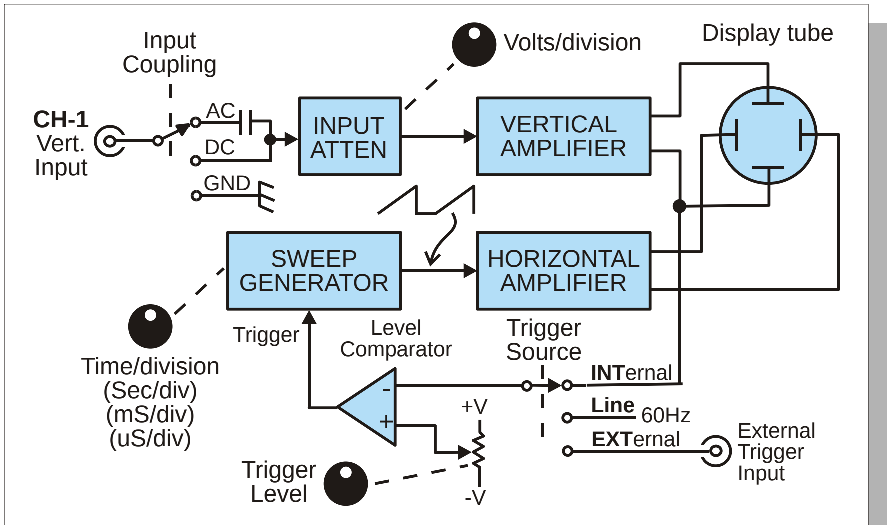

FIG. 1 - Basic block diagram of an Oscilloscpe

The sawtooth ramp is amplified to high voltages, applied to the oscope tube, to deflect the beam from left to right. An important task of an oscope is when the horizontal deflection begins. Normally a switch labeled "Trigger Source" determines what initiates the sawtooth ramp. In the "Internal" position, a sample of the input signal (in the vertical amplifiers) is sampled, with a variable resistor setting the level. When the input signal exceeds the "Trigger Level," a pulse is generated to start the sawtooth ramp and hence the horizontal sweep. The purpose of triggering is to keep the input waveform synchronized to the sweep so it appears stationary on each sweep. The trigger source usually has a "Line" position, which simply triggers the sweep off of 60Hz from the power supply. This synchronizes the sweep to the AC power frequency and is useful for checking television signals, which are synchronized to the power mains. Also, an "External" position may be present, which connects an external input signal (via a BNC connector) to trigger the sweep generator.

Other features your oscope may have are two vertical channels for dual trace operation, various modes to display both waveforms (alternate, chopped, A+B added, etc.), delayed sweep features, dual sweep time bases, built in calibrators, etc.

CALIBRATING YOUR OSCILLOSCOPE

The first thing you should do upon acquiring an o-scope is to check its calibration.

The vertical amplfiers can be checked with a known voltage source or 9v transistor radio battery. Measure the output voltage of the battery with an accurate voltmeter. Let's say it just happens to be +9v exactly. Set the input coupling to ground (0v) and move the trace to the bottom division. Switch the input coupling to DC and set the attenuators to 1v/div. The deflection should be 9 divisions. Switching to 10v/div., deflection should be 0.9 divisions. Internal to the oscope (or perhaps accessible from the outside) are adjustments for the vertical amplifier gain. Adjust this for 9 divisions of deflection in the 1v/div. range. Procedure can be repeated with a 1.5v flashlight battery (assuming you know the exact voltage from a DVM).

The horizontal amplifiers should be checked/calibrated using a signal generator. For example, a 1MHz signal has a period of 1uS. Setting the sweep rate to 1.0uS/div., a 1MHz signal should take exactly 1 division per cycle. Set the horizontal width control properly to ensure the beam starts at the first division and ends at the last division. If the sweep rate appears incorrect, an internal adjustment (Sweep gain or similar) can be set for proper display of the test signal.

The main operator controls are:
- **Intensity** - controls the brightness of the beam. **NOTE: Too bright a beam can damage to the CRT tube!**
- **Focus** - adjusts the beam for the thinnest and sharpest display.
- **VERT & HOR Position** - controls the vertical and horizontal position of the display respectively
- **VERT V/div** - controls the vertical sensitivity of the display, i.e., how many volts (or mV) per division.
- **HOR Sweep Speed** - sets the horizontal sensitivity, i.e., how many mS or uS per division.
- **VERT & HOR vernier** - allows the vertical and horizontal sensitivity settings to be varied in small steps.

Other adjustments you may find on your scope are:

Astigmatism - With the scope intensity and focus properly set, this adjustment compensates for the curvature of the CRT tube by making it in-focus across the sweep. If your trace is out-of-focus in certain areas, but in-focus elsewhere, the astigmatism needs to be adjusted. See ***Fig. 2***.

Trace Rotation - is a small coil around the CRT that skews the trace to ensure it is perfectly horizontal. On scopes without this adjustment, the trace is leveled by physically rotating the CRT to align the trace to the graticle grid. See ***Fig. 2***.

Fig. 2 – Effects of Astigmatism & Trace Rotation

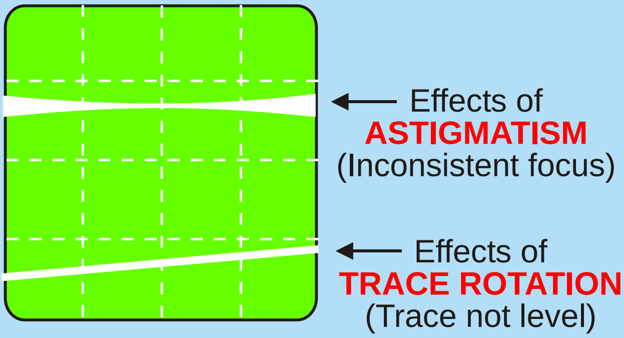

DC BAL (DC Balance) - is a dc offset in the vertical amplifiers that causes a shift in the trace baseline when changing vertical scales. It is most obvious when measuring ac voltages. For example, you are displaying a 10Vpp sine wave, centered on the center graticle, at 2v/div. Changing to 5v/div, the sine wave shifts off the center graticle ... that is, it assumes a dc bias error. The DC BAL is adjusted until the shift no longer occurs when changing vertical scales.

HV ADJ. - is the high voltage that controls the intensity of the trace. Turn up the **Intensity** control to its brightest position, then adjust the HV ADJ for a trace slightly brighter than normal intensity. The **Intensity** control now has the proper range. The HV ADJ might have to be re-adjusted to acquire proper focus.

NOTE: Very bright trace displays can cause permanent damage to the CRT, particularly on a well-used scope.

LET'S MAKE SOME MEASUREMENTS

It is assumed you have your scope relatively calibrated and familiar with the front panel controls. The sample o-scope displays are based on eight vertical and ten horizontal divisions on the CRT screen, typical to most oscilloscopes. Most waveforms are actual displays of the signal cited, photographed from my trusty Tektronix 475 oscilloscope.

First ... a word on TRIGGERING.

Most oscilloscopes have a knob or two for "Triggering." This tells the oscilloscope when to start the sweep. When the **Triggering Slope** is placed in the **(+)** position, the scope will begin its trace when the *input* signal goes positive. Likewise, when **(–)** triggering is selected, the trace will begin when the input signal goes negative, as shown in **Fig. 3**. Often there will be the option to chose the **Triggering Source**, such as "CH.1" or "LINE." Line means the scope is triggered off the 60Hz line voltage, and is useful when synchronizing on television signals or looking at 60Hz power supply noise. CH.1 or CH.2 means the scope will trigger off the signal on channel 1 or 2 respectively. **Trigger Level** is at what voltage of the input signal triggering begins. For example, if set high, triggering may not begin until the input signal reaches several volts. When set around zero, it will trigger the moment the signal goes positive (if set for (+) triggering). This setting can be troublesome if noise exists on the signal. Adjust for stable triggering.

Fig. 3 – Triggering

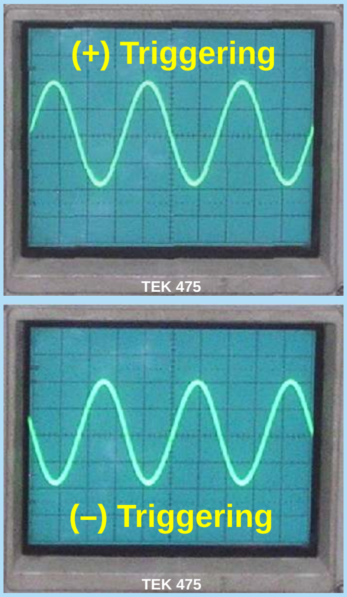

DC Voltages.

Say you want to check the transmit-receiver (T-R) switch in your QRP rig, or other digital signal. See ***Fig. 4.*** The key line is the input to the HCT240 inverter to form the 0v TX– on key- down and the 0v RX– on key- up. This switches the rig between transmit and receive (T-R Switch). It is a logic function, that is, a voltage to represent ON or OFF.

Place the scope lead on pin 13 at 10v/div. and you should see the waveform like the top trace in ***Fig. 4*** ... about +6v on key-up and 0v on key down. Move the scope lead to pin 7 and you should see 0v on key-up and about +8v on key-down (bottom trace). If the output does not go "HI" (+8v) on key-down, or does not go to a solid "LO" (<1v) on key-up, the inverter is not working properly. (It's busticated). Many shortwave receivers use similar schemes for switching filters or attenuators.

While this test could be done with a DVM, the integration time is slow, requiring long key-downs to get the voltages. A scope will also show you how clean the switching is, or if there is an ac voltage (or RF noise) riding on the T-R voltage.

Scopes are thus good dc voltmeters, with about a 5% reading accuracy.

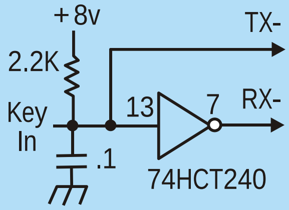

Fig. 4 - 38-Special T-R Switch

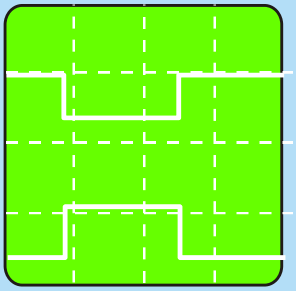

VERT: 10v/div DC
HOR: 500mS/div

AC Voltages.

Here is where an oscilloscope pays for itself by making AC voltage (and frequency) measurements. You must remember, AC voltages are displayed on a scope as ***peak-to-peak* voltages**, while a voltmeter measures in ***rms***. RMS voltages are about 1/3 the p-p voltage read on a scope, or specifically:

$$V_{rms}=\frac{1}{2}(.707 \times V_{pp}) = 0.354 \times V_{pp}$$

For example, let's measure the output voltage and frequency from the sidetone oscillator in your QRP rig. Place the scope lead on the audio amplifier output. On key-down, you get the waveform shown in ***Fig. 5***. The transmit sidetone audio is 1.9Vpp.

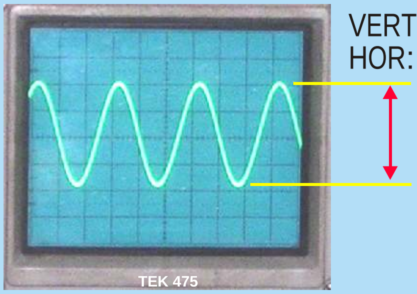

VERT: 0.5v/div ACV
HOR: 0.5mS/div

3.8 divisions peak-to-peak times 0.5v/div
= 1.9 Vpp
= 0.67 Vrms

Fig. 5

AC Frequency Measurement.

With this waveform, we might as well see what frequency our sidetone or transmit-offset frequency is. Most operators prefer the sidetone to be about 700–750Hz. Trigger the scope for a stable waveform and set the time-base (sweep) to display 2 or 3 cycles, as shown in ***Fig. 6***. Center the waveform between two horizontal divisions so zero volts on the waveform is on a graticule line, then move the horizontal position so the first "zero– crossing" is also on a division line.

Fig. 6

VERT: 0.5v/div ACV
HOR: 0.5mS/div

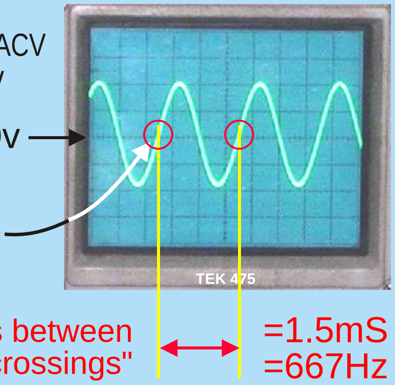

0v

First "Zero Crossing"

3 divisions between "zero-crossings"

=1.5mS
=667Hz

Measure the time it takes to make one complete sine wave from one zero-crossing to the next. In this example, it is 1.5 divisions, at 1mS per division, or 1.5mS. Frequency is simply the reciprocal of time, such that the sidetone frequency is:

$$f = \frac{1}{t} = \frac{1}{1.5\text{mS}} = 667 \text{ Hz}$$

For some, this may be about right. For others, this may be a little low to your liking. To raise it to 700Hz, calculate the time period of 700Hz (1/700 = 1.4mS). At 1.0mS/div, you can adjust your sidetone or transmit offset until zero-crossings for a single sinewave is 1.4 divisions. This will be about 700 Hz. (Sidetone may not be adjustable on some rigs).

All frequency measurements are made in this fashion, by measuring the distance between zero-crossings (or from one peak to the next) and converting the time period to frequency. This should emphasize the importance of ensuring your sweep speed is calibrated; as any error in the time base will cause a corresponding error in the accuracy of your time or frequency measurements.

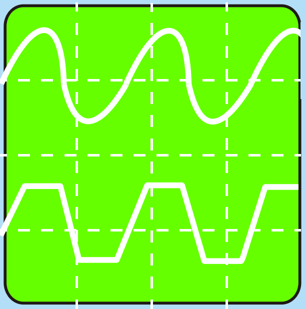

Fig. 7

Example of Waveform Quality

VERT: 2v/div ACV
HOR: 1.0mS/div

Quality of the waveform is another feature of a scope that is unsurpassed since you are "seeing" the waveform in real time. Two examples of waveform quality are shown in ***Fig. 7***.

The top trace shows the sidetone frequency with distortion, perhaps due to improper time-constant on the coupling capacitors or improperly biased audio amplifiers. The bottom trace would be a raspy sounding side tone, due to the amplifier being over-driven and in compression (clipping). The o-scope is an invaluable tool for detecting and diagnosing such impurities in the signal quality.

FIG. 8 – Full Scale Signal Display

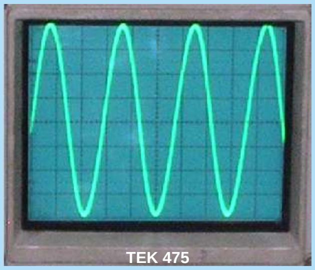

Actual display
Tektronix 475

MORE NIFTY MEASUREMENTS

Amplifier Gain.

The gain of an amplifier can be measured in terms of voltage or decibels (dB). For voltage gain, it is simply Vout/Vin of the amplifier. For example, if the input is 1Vpp and the output is 4Vpp, then the amplifier has a voltage gain of 4.

Gain in dB is often more useful and is how the gains of amplifiers are usually expressed. With dB's, every-time you double the AC voltage, you add 6dB of gain. It is the **ratio** of output to the input, and this **ratio** is easy to measure on a scope.

It is often easier to start with the output. Set the vertical amplifier gain to display the amplifier **output** as a full-scale signal as shown in ***Fig. 8***. Now move the scope probe to the amplifier **input** without disturbing the scope gain. You will of course have a much smaller signal, and the ratio of the input to the output will be the gain in dB. In our example of using eight divisions for full-scale, then four divisions would be 6db, 2 divisions 12dB, etc. as shown in **Fig. 9.** You may want to add your own dB scale along your scope display to remind you of this relationship. Note: this is ***voltage gain*** (Av=20log x Vout/Vin). In this example, with 4Vpp output and 1Vpp input (Av=4), then the gain is dB=20log(4) = 20(0.602) = 12dB, or as shown directly on the CRT tube. Since this is a relative measurement, the absolute Vin or Vout voltage does not need to be determined.

Fig. 9 – Volt. vs. dB relationship

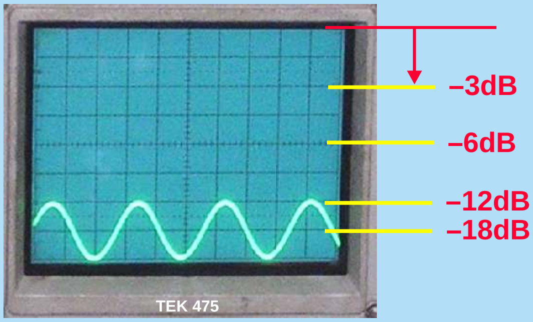

Fig. 10 Measuring Phase Shifts

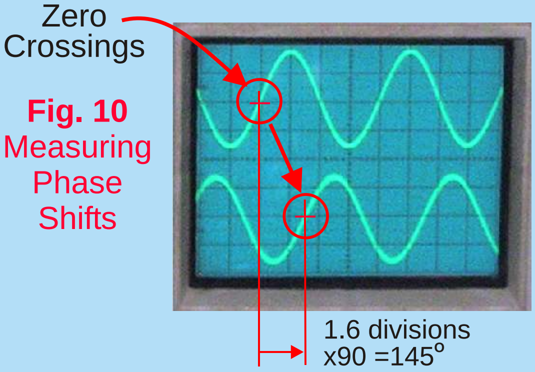

Insertion Loss.

In some circuits, such as filters or attenuators, the ***loss*** in the circuit needs to be measured, and like circuit gain – expressed in dB. The loss through a circuit is called the ***insertion loss***. It is determined in the same way as amplifier gain just presented, except start with the input (the highest AC voltage) as the full-scale or reference display, then measure the output AC voltage (the lowest level). **The ratio is the insertion loss in dB.**

For example, with a signal generator connected to your receiver, you want to measure the insertion loss through the IF crystal filter. At the filter input, you can just barely squeek out 2 divisions of input signal on your scope at its most sensitive setting. The output from the crystal filter is 1.5 divisions. The insertion loss would be 20log(1.5/2.0 div.) = –2.5 dB. If the output were only 1.0 division (50% reduction), the insertion loss would be 6dB.

Fig. 11 –Phase Shift by imposition

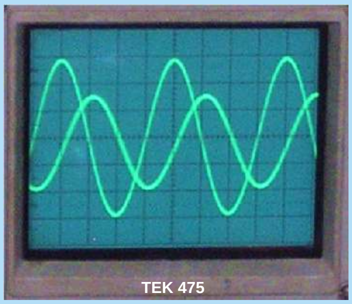

Measuring Phase Shifts.

Phase relationships between two signals at the same frequency can be measured with 2-5^{o} accuracy with a scope, although more suited for a dual-trace scope. The reference signal is applied to CH. 1 and the signal to be measured to CH. 2. For proper phase measurements, ensure your dual trace display is in the **chopped** mode, not **alternate** mode for proper phased referenced triggering.

There are many methods to do this. One is to stretch out the signal so it takes 4 horizontal divisions, such that each division is 90^{o} of phase, as shown in **Fig. 10**. By measuring from a common point on one signal (zero-crossing or from peak-to-peak) to the next, the phase can be measured. For example, say you are making a phased-array antenna in which one feedline must cause a 90^{o} delay. You calculate the electrical length for a ¼λ [L=(246/f) x Velocity Factor] and cut the coax to that length. You are now working on blind faith that you have exactly 90^{o}. With a scope, you can measure it fairly accurately by injecting a signal into one end with a signal generator (at the frequency of interest) and a 50Ω load on the other. Connect the scope CH.1 to the coax (signal) input and CH.2 to the load end and measure the phase. In the **Fig. 10** example, the CH.2 signal is delayed by 1.6 divisions, at 90^{o} /div is 145°. ***Your delay line is too long!*** Cut off an inch or two at a time until the CH.2 signal is 90^{o} from CH.1 for precise tuning of the delay line. (While departing from o-scopes for a moment, the sharp null of a phased array is astounding when exactly 90° delay is achieved. More than 10-15° in error causes a very "mushy" null with little difference over a single vertical antenna. Most errors in achieving exactly 90° by the "measure-and-cut" method are due to uncertainties in the stated velocity factor of the coax).

Another method is to superimpose the two signals on top of eachother. Make one signal larger than the other so you know which one is what, as shown in **Fig. 11**. In this example, the smaller signal lags the larger signal by about 100°, estimated by where they cross. For more accurate determination, use the time base to measure the time period of one cycle (T1), then the time period one signal lags (or leads) the other (T2). The phase shift is then $\theta=(T2/T1)\times 360^{o}$.

Phase measurements can be made on a single trace scope as well. First, connect the reference signal, uisng a BNC "T," to both the *external trigger* and the normal *vertical input*. Adjust the trigger level so the zero-crossing occurs at the beginning of the trace (left-hand graticle). Remove the reference from the vertical input, but not the e*xternal trigger,* and apply the signal to be tested to the vertical input – without altering the time base or trigger level. The distance of zero-crossing of the test signal is from the left-hand graticle can now be measured to determine the phase, though with slightly less accuracy than using a dual-trace scope.

An interesting experiment is to measure the phase shift of the audio signal at different frequencies as it travels through the stages in a CW, SSB or AM active filter. What is the phase shift of the wanted vs unwanted frequencies?

Measuring Rise and Fall Times.

In digital circuits, it is sometimes important to know the rise and fall times of a signal through a gate. In amateur radio transceivers, this same interest could be applied to how fast the T-R switch switches. On key-down, if the transmitter turns on slightly before the receiver is turned off, it can produce an annoying "thump" in the receiver. Rise and fall times are measured by triggering on the edge of the signal of interest, then increase to a faster sweep speed to measure the time it takes the signal to reach 90% of its final level. The signal to be measured is shown in **Fig. 12** on the top trace, and the expanded version on the bottom. For proper rise times, the signal being measured should be well within the bandwidth of your scope and using a low capacity probe.

For example, in **Fig. 12** (bottom trace), the rise time is about 1/4th of a division. If the sweep speed is 100nS/division, the rise time would then be about 25nS.

Fig. 12 –
Rise and Fall Times

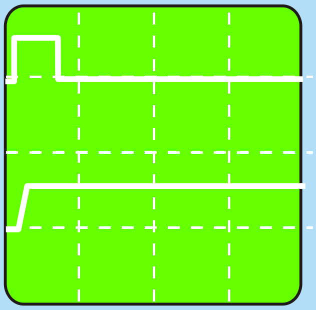

USING LIMITED BANDWIDTH SCOPES

Today's scopes have 200–500MHz bandwidths. Likely your scope is much less than that. A limited bandwidth scope is still very useful to the amateur or homebrewer. Say the bandwidth of your scope is 5MHz. This does not mean you can't see 7MHz signals. It just means the peak-to-peak value has lost meaning, and will likely be very weak, since it is beyond the bandwidth of the scope. (Like other bandwidth measurements in electronics, the "bandwidth" of a scope is usually based on the "3dB bandwidth." That is, at the maximum bandwidth, you are already at the –3dB point, or a 25% reduction in the peak-to-peak voltage display). You can still resolve individual cycles higher than the cited bandwidth to a certain degree and make gain and phase measurements, since they are based on ***ratios.***

Most of the examples in this article explore many regions of a communications receiver or ham transceiver without the benefit of any great bandwidth. Experiment with your scope to learn its limitations. ***Use a good scope probe and make measurements with a good ground to get the most out of the bandwidth you have.***

For the homebrewer building circuits in the HF bands, a 50 MHz scope **with good calibration** will yield fairly accurate measurements up to 30 MHz with little concern for accuracy. The old 465 or 475 series of Tektronix scopes, with 100/200 MHz bandwidths, make an excellent oscilloscope for the amateur or experimenter. They can often be found at hamfests today for $100–150, and tend to maintain a fairly good calibration almost regardless of how much use they have seen.

In Part 2 - we'll probe (bad pun) into some advanced measurement techniques, even with a simple scope ... such as measuring sideband rejection, filter responses, VCO phase noise, etc. (and what it all means).

72, Paul NA5N

na5n@zianet.com
pharden@nrao.edu

Colophon:
Article prepared by NA5N using CorelDraw 11

START NOW

3 Easy Steps:
1) Click 'Start Now'
2) Download on our website!
3) Get Free File Converter

fromdoctopdf.com

advertisement

NIC NAC TIC TAC CRYSTAL RADIO SET - AN IDEAL PRIMARY SCHOOL PROJECT...

Technology (/technology/) > Science | by mk484 (/member/mk484/) Follow

10,750 83 16

Posted Jul. 4, 2014 | PUBLICDOMAIN

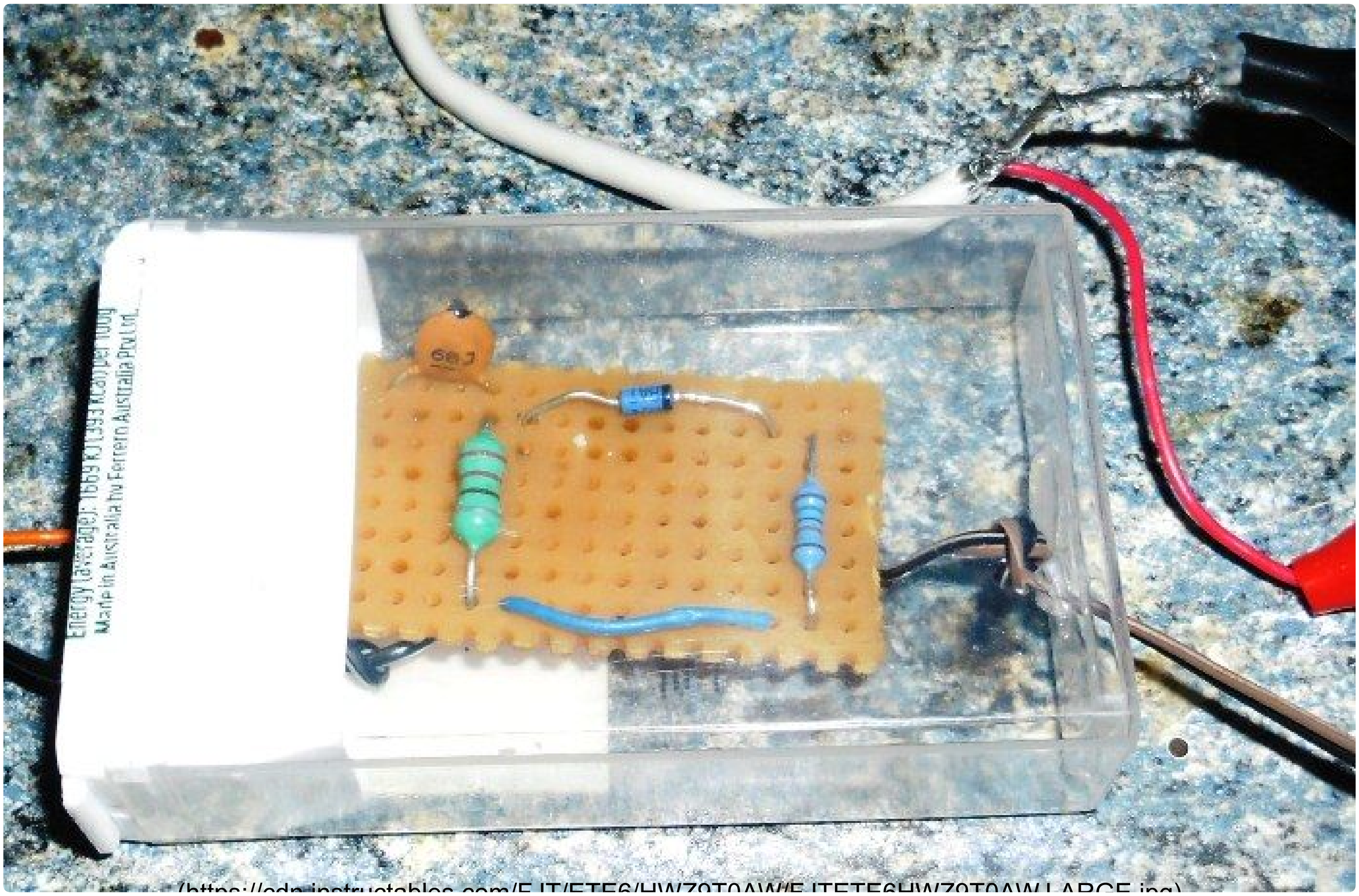
68J
Energy (average): 1669 kJ/393 kcal per 100g
Made in Australia by Ferrero Australia Pty Ltd

(https://cdn.instructables.com/FJT/FTE6/HWZ9T0AW/FJTFTE6HWZ9T0AW.LARGE.jpg)

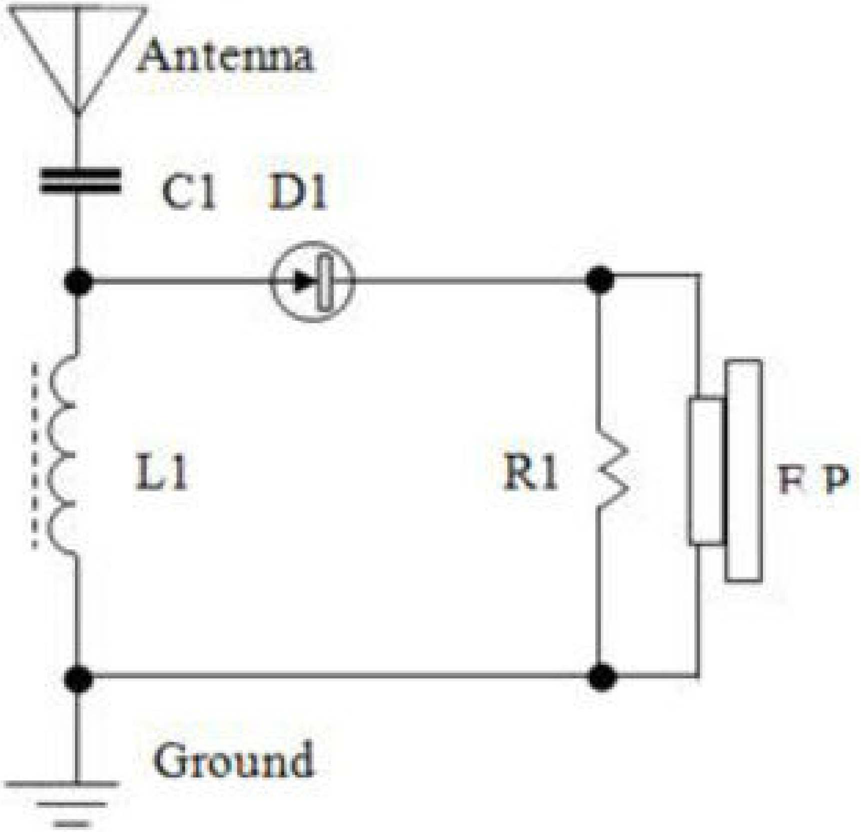

(https://cdn.instructables.com/F8O/OEKG/HX8HK0L5/F8OOEKGHX8H

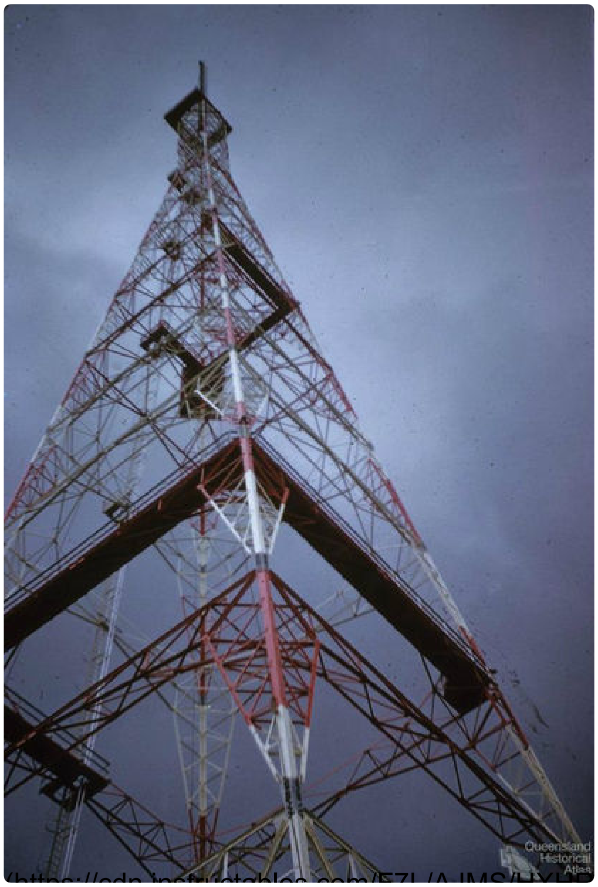

(https://cdn.instructables.com/F7L/AJMS/HX

Hi there! This Instructable is all about building a basic crystal radio set that is so simple to build and understand, that a child could do it - with help from mum or dad - or even at school, as a class project. Parts can be bought "off the shelf" at Jaycar Electronics and other suppliers, or purchased online via Ebay and Paypal.

I built my first ever crystal set when I was 9 years old, but that was a different world. We're going to build a 21st century style radio with all modern components which are relatively easy to obtain at your local electronics store or on the Internet.

The radio is contained inside a Tic Tac box, and is called "The Nic Nac Tic Tac Radio". It is built on a square of matrix board, with the component leads poked through the holes in the board, and joined underneath. It is a very simple crystal set, designed to initially receive only one station, but you can expand the tuning range with the addition of only one extra component.

The Tic Tac radio can be constructed by simply twisting the component leads together, underneath the matrix board and attaching antenna/ground wires as well as the earphone wires to the board, using that same technique. You don't have to solder any wires, but there is another way to join them all together - more on that later.

The Tic Tac Radio is a very easy set to put together - you don't need to wind bulky coils or use heavy tools and screw strange looking parts onto a large breadboard. See step 1 for the parts list you will need.

Add Tip Ask Question

Step 1: Parts and Tools You Will Need

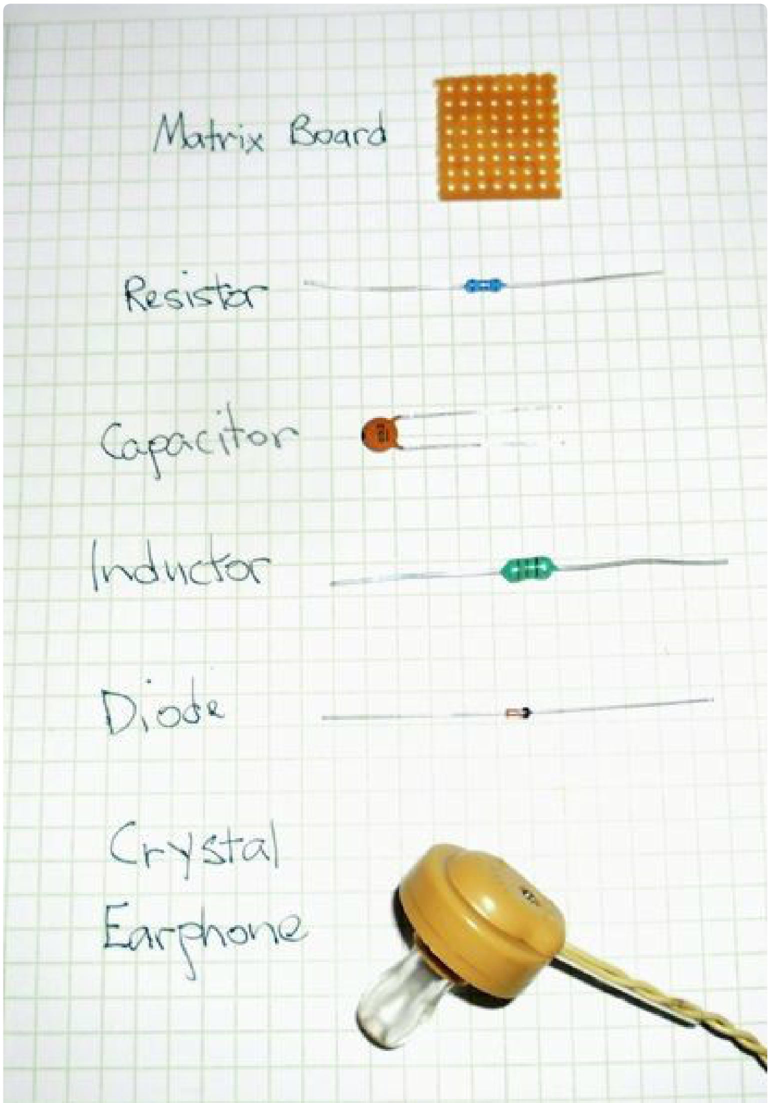
Matrix Board
Resistor
Capacitor
Inductor
Diode
Crystal Earphone

(https://cdn.instructables.com/F9R/WLFJ/HX5H54TY/F9RWL

(https://cdn.instructables.com/F99/H0BC/HY

(https://cdn.instructables.com/FF9/DK

(https://cdn.instructables.com/FXV/5

You will need to acquire the following components from a local electronics shop or an online electronics business.

I am adding the catalogue numbers for Jaycar Electronics, so if you live in Australia or North America, you may be able to visit a Jaycar store and buy the parts over the counter, if not then you may be able to do an online mail order via Paypal:

Parts List

Resistor - 47k - yellow-purple-black-red and brown - RR 0612 (pkt of 8)

Capacitor - 68pF - ceramic x 2 - RC 5322 (pkt of 2) and a 100 or 120 pF value as well for experiments.

Inductor - 220 uH - red-red-brown silver - LF 1538 (resistive type)

Polyvaricon tuning capacitor - 220 pF - RV 5728 - with knob and mounting screws

Diode - BAT46 - ZR 1141 (You can also use a 1N34A Germanium Diode too if you have one at home)

Ceramic Earphone - AS 3305 *

A 25 meter roll of yellow hook up wire for the Antenna wire and a 3 meter length of wire for the Ground wire.

Please note that some Jaycar parts come in multiples of 2 or more per packet. And please note the following:

PLEASE NOTE: THIS PROJECT/KIT CONTAINS SMALL PARTS THAT MAY FORM A CHOKING HAZARD FOR SMALL CHILDREN OR PETS. NOT SUITABLE FOR CHILDREN UNDER FIVE (5) YEARS OLD.

*A normal crystal radio earphone is OK, but if you can't get one of these, or if the one you bought goes dead (as they sometimes do,) you can use a substitute, such as the Murata PKM44EW passive transducer (see picture above) which is available from an old Telstra TF200 touchphone, (the one on the left in the diagram above,) or an equivalent, such as the ARIO transducer, from an old Telstra T1000 pushbutton phone.

The ARIO unit is soldered to the phone's pc board so you'll need to be able to unsolder the three mounting pins underneath the board, or find someone in the neighbourhood who is able.

Take the back off the TF200 (if you've obtained one of these phones,) and you'll see a black disc shaped object 2" round by 1/2' thick - with a red and black wire. Unplug the wires from the circuit board, and unscrew any retaining screws and remove the transducer. Cut the mini plug off, carefully strip the insulation from the ends of the wires and extend them by about 18" with 2 thin lengths of hookup wire. These pietzo devices make good earphones for crystal sets and can be housed in an old pair of ear muffs.

Miscellaneous Materials:

A Tic Tac box (smaller size)

A piece of matrix board at least 7 holes long by 8 holes across. Cut the board to fit neatly inside the Tic Tac box.

A short length of 2 differently coloured wires 60 mm in length and 2 crocodile clips with red and black plastic covers

A length of antenna wire at least 25 metres long and a 3 metre length of a different colour for the Ground wire.

A metal rod or cold water pipe for the ground stake. Be careful which pipes you connect your Ground wire to.

You will need a small sharp object for punching holes in the Tic Tac box.

A small pair of wire cutters for cutting and stripping wires.

Add Tip Ask Question

Step 2: How a Crystal Radio Works

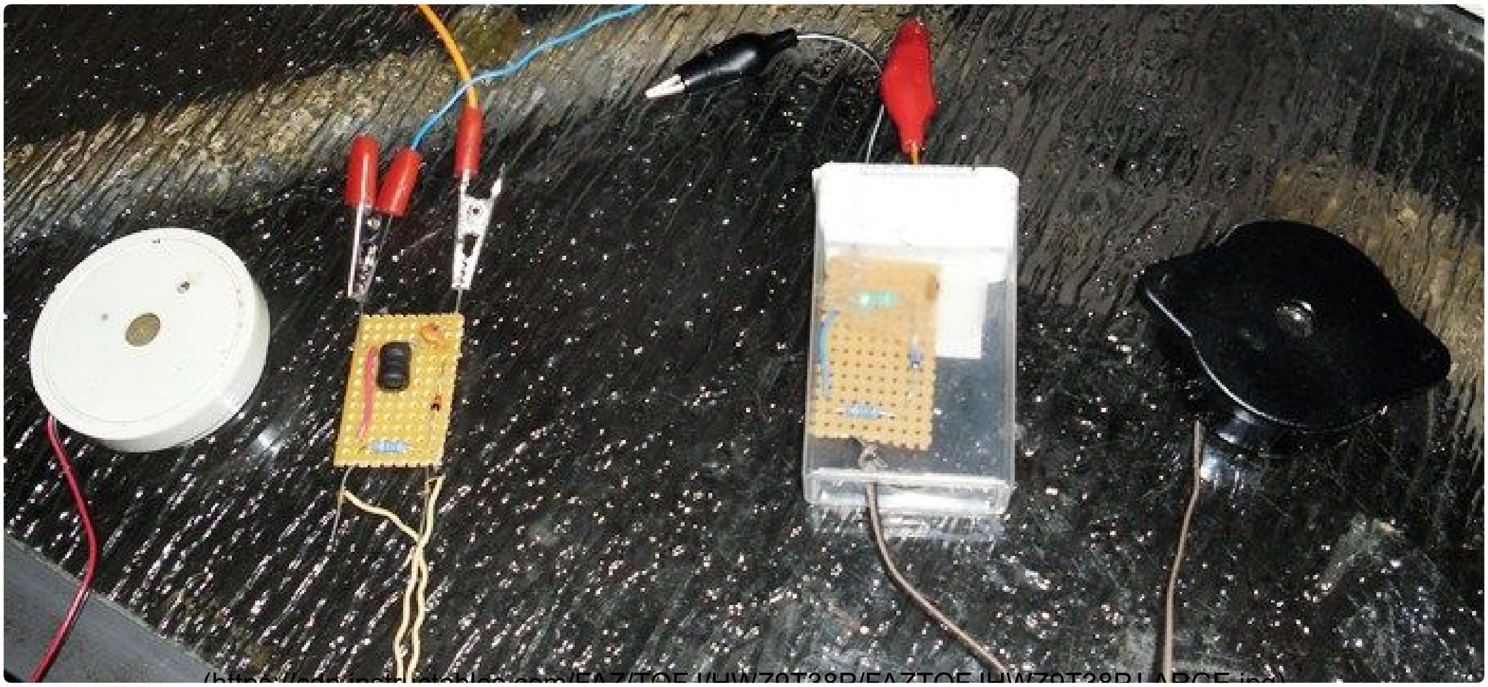

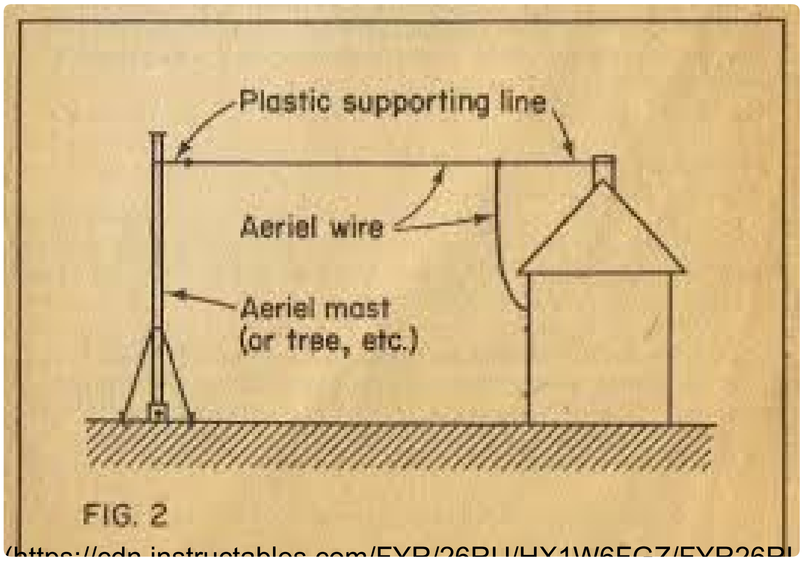

FIG. 2

(https://cdn.instructables.com/FXB/26RU/HX1W6FCZ/FXB26RU

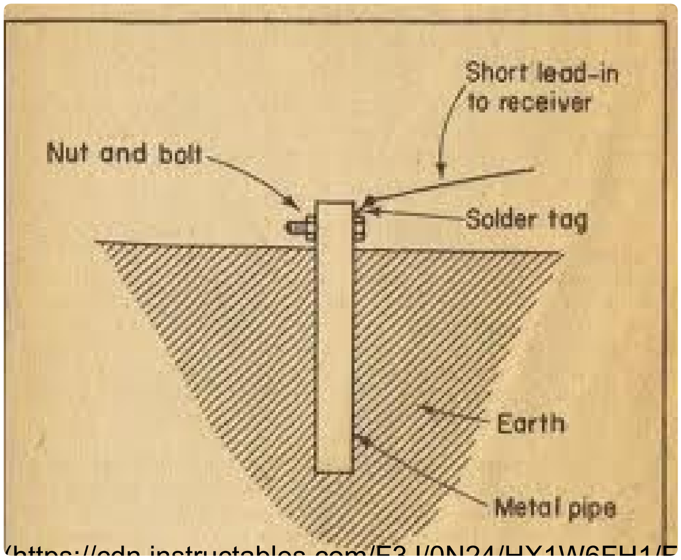

(https://cdn.instructables.com/F3U0N24/HX1W6FH1/F

Radio signals consists of two parts - the 'carrier wave' which is the AM radio station's frequency of operation, and the 'program signal' which is mixed with the carrier wave for transmission.

Radio waves travel out from the AM transmitter tower through the atmosphere. We want to capture one specific frequency so we can listen to it, so we need the antenna/ground wire system to capture that signal. We also need a 'tuned circuit' that will filter out the desired AM signal, and discard the rest, so that all the other unwanted radio signals pass out through the ground wire to earth.

Two components in our circuit will perform that task for us. The capacitor C1, together with L1 inductor, form a basic 'series tuned' circuit. Their respective values will determine just which local AM radio station we will capture.

We also need a diode to 'detect' the voice and music, so we can hear them in our earphone, which transduces electrical signals into sound waves that we can hear.

In the photo above you can see a completed Tic Tac crystal radio. It is already inside the box. The other set is connected to the antenna/ground wire circuit, undergoing a 'soak test'. It is necessary to do this to ensure that the radio will work once inside the box!

The first diagram shows a typical antenna wire installation. Coming out of a window, the wire is anchored to the building and then over some distance (10 metres +) to a nearby tree or other building.

You must take great care not to erect antenna wires near to power or telephone cables, near your home!

The ground wire comes out the same window and is anchored to a metal pipe/water pipe or metal ground stake, embedded in soft, moist soil.

You must not connect ground wires to electrical mains wiring installations, including wall power outlets!

Another electrical hazard to consider is lightning strikes! Although it is very rare for anyone to be seriously injured or die from a lightning strike it is not impossible. So, if you hear a thunderstorm coming your way (you may hear the lightning 'crashes' in your earphone first,) then disconnect your antenna wire immediately, connect it to the ground wire and put it well up out of the way. Stay well clear of this temporary antenna/ground connection until the storm has completely passed away from your general area - miles away!

Add Tip Ask Question

Step 3: How to Build the Tic Tac Crystal Radio Set

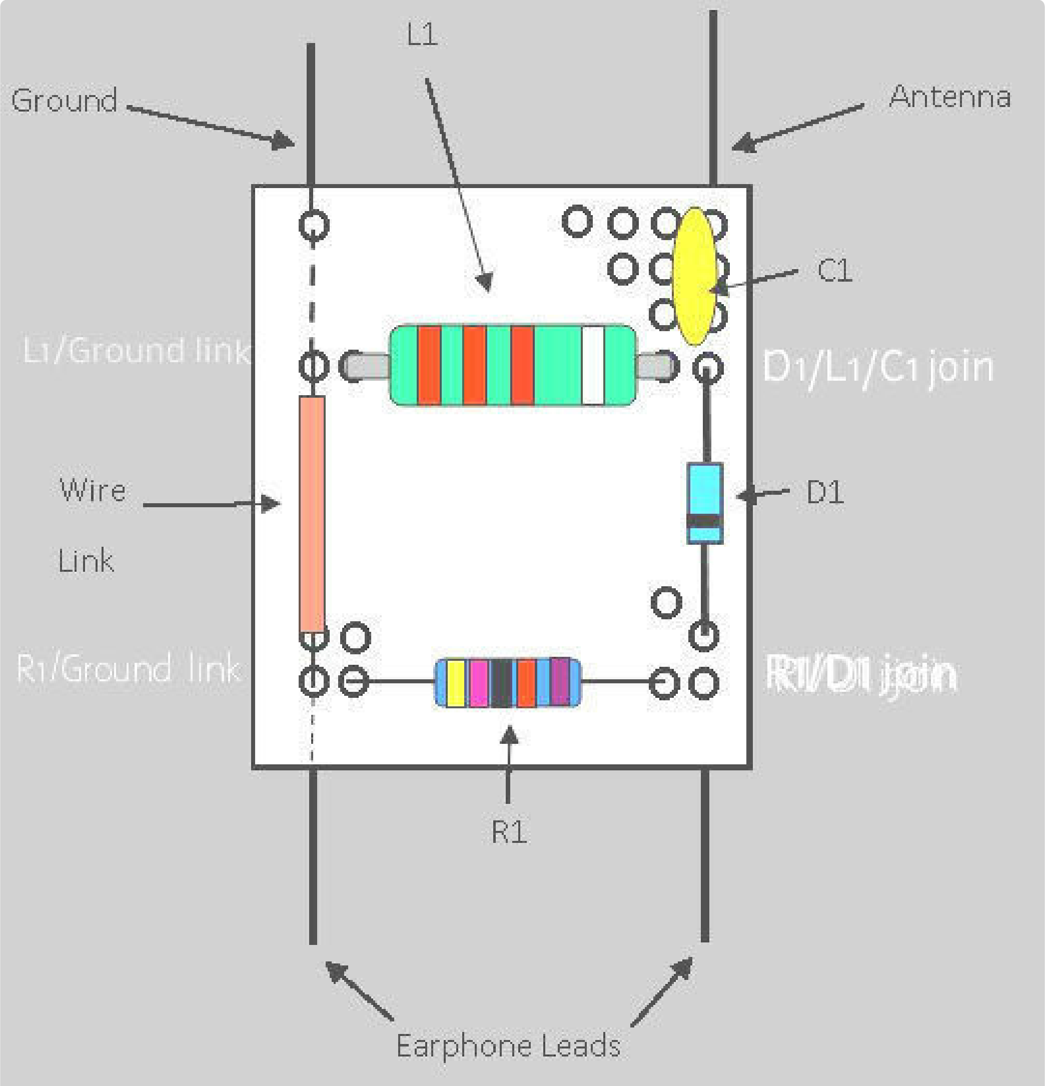
L1
Ground
Antenna
C1
L1/Ground link
D1/L1/C1 join
Wire
Link
D1
R1/Ground link
R1/D1 join
R1
Earphone Leads

(https://cdn.instructables.com/FA5/FW4E/HYA9RXF0/FA5FW4EHYA9RXF0.LARGE

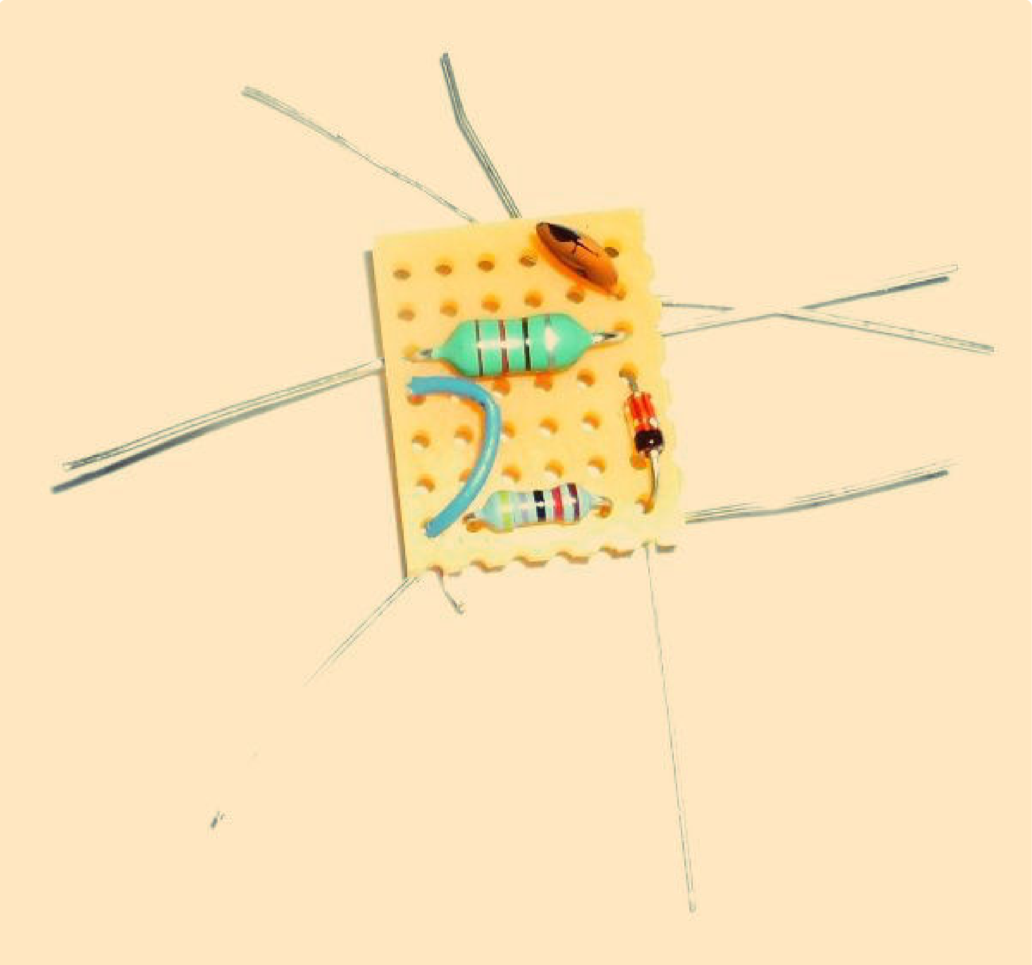

(https://cdn.instructables.com/FS8/N068/HXA9RY3O/FS8N068HXA9RY3O.LARGE

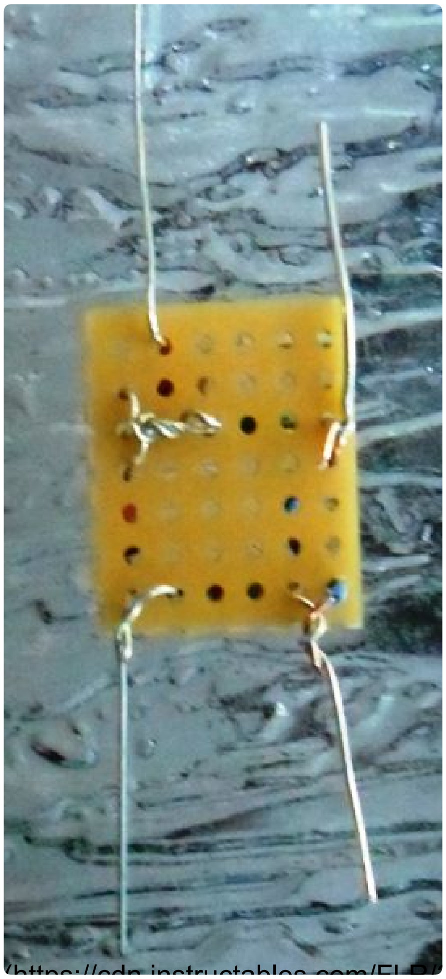

(https://cdn.instructables.com/FLP/

1. Take the components and lay them out on a clean surface.

2. Take the square of matrix board (this is also called 'perf board' because of all the holes or perforations in the material) and beginning with the 220uH (uH = micro-Henry) and place it as shown in the diagram.

3. Then take the diode, capacitor and resistor and place them in the places shown for them, making sure that the coloured band at one end of the diode joins with the resistor wire, as shown

4. Then take each junction where wires come together through their respective holes and gently twist them together, until they form a neat, tight bundle. Take your side cutters and cut off any excess length, taking care not to cut any one wire too short, so that it comes undone from the join.

.5. Take the link wire, strip 2 centimetres of insulation from each end of the wire and install that wire between the free ends of the inductor and resistor, and the matrix board construction is complete.

6. Then take the ceramic earphone, cut the plug off the end, and strip the 2 wire ends about 1.5 centimetres in length. Wrap each earphone wire around each end of the resistor component, underneath the matrix board.

7. Finally, strip the insulation off both ends of the 60mm long differently coloured wires, and attach them to the matrix board - one goes to the inductor/link wire junction and this will be the Ground connection wire. Attach the other one to the free end of capacitor C1 and this will be the antenna wire connection. Both connections are made underneath the matrix board. The Nic Nac Tic Tac Crystal Radio is now ready for testing.

Add Tip Ask Question

Step 4: Final Checks and Installation of Wires

24g NET
FRESHNESS IN 2 CALORIES
tic tac
Strawberry Fields

(https://cdn.instructables.com/FX9/8RCY/HX8HJEU3/FX98RCYHX8HJEU3.LAR

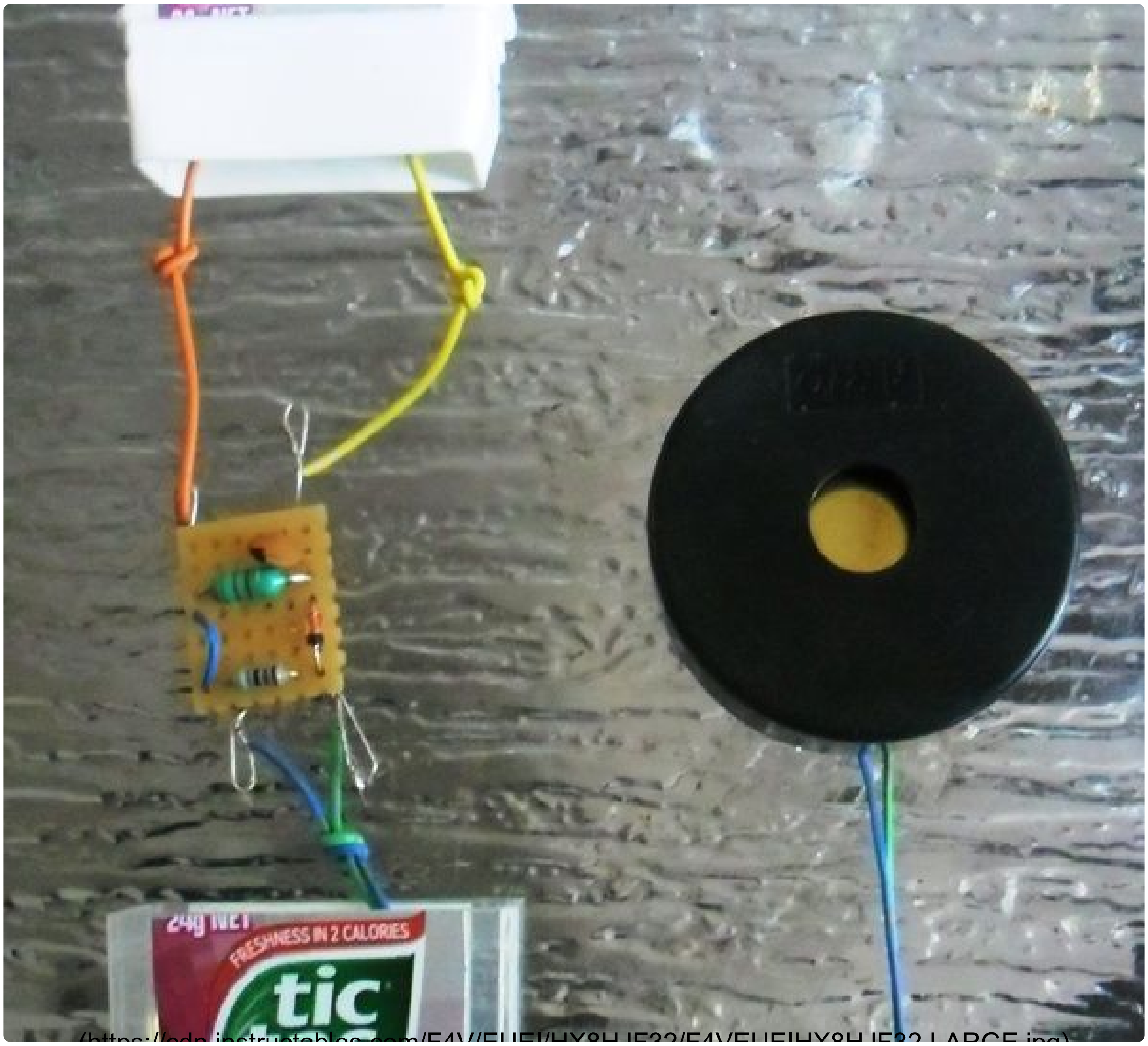

(https://cdn.instructables.com/F4V/FUFI/HX8HJF32/F4VFUFIHX8HJF32.LARGE.jpg)

When you've finished constructing the matrix board circuit and have clipped all the excess component leads off, puncture three (3) holes in the Tic Tac box - 2 small holes on the top lid of the box, about an inch (24mm) apart, so that your antenna and ground wire leads can come through the box lid, and be connected to the matrix board at those 2 points.

Feed the wires through the holes and then tie small knots in each one, near the underside of the box lid, so that they won't pull back out if strained, and disconnect themselves from the matrix board.

Then make one larger hole in the centre of the bottom of the box (clear part) so that your earphone wires can be fed through to the connecting points on that part of the matrix board. Tie a larger knot in the earphone lead so as to prevent it from pulling out of the box if it is strained. Strip the insulation off the earphone leads, and wrap them around the matrix board leads at those 2 points. You are now ready to test out the Nic Nac Tic Tac Crystal Radio.

Add Tip Ask Question

Step 5: Testing Out the Tic Tac Radio

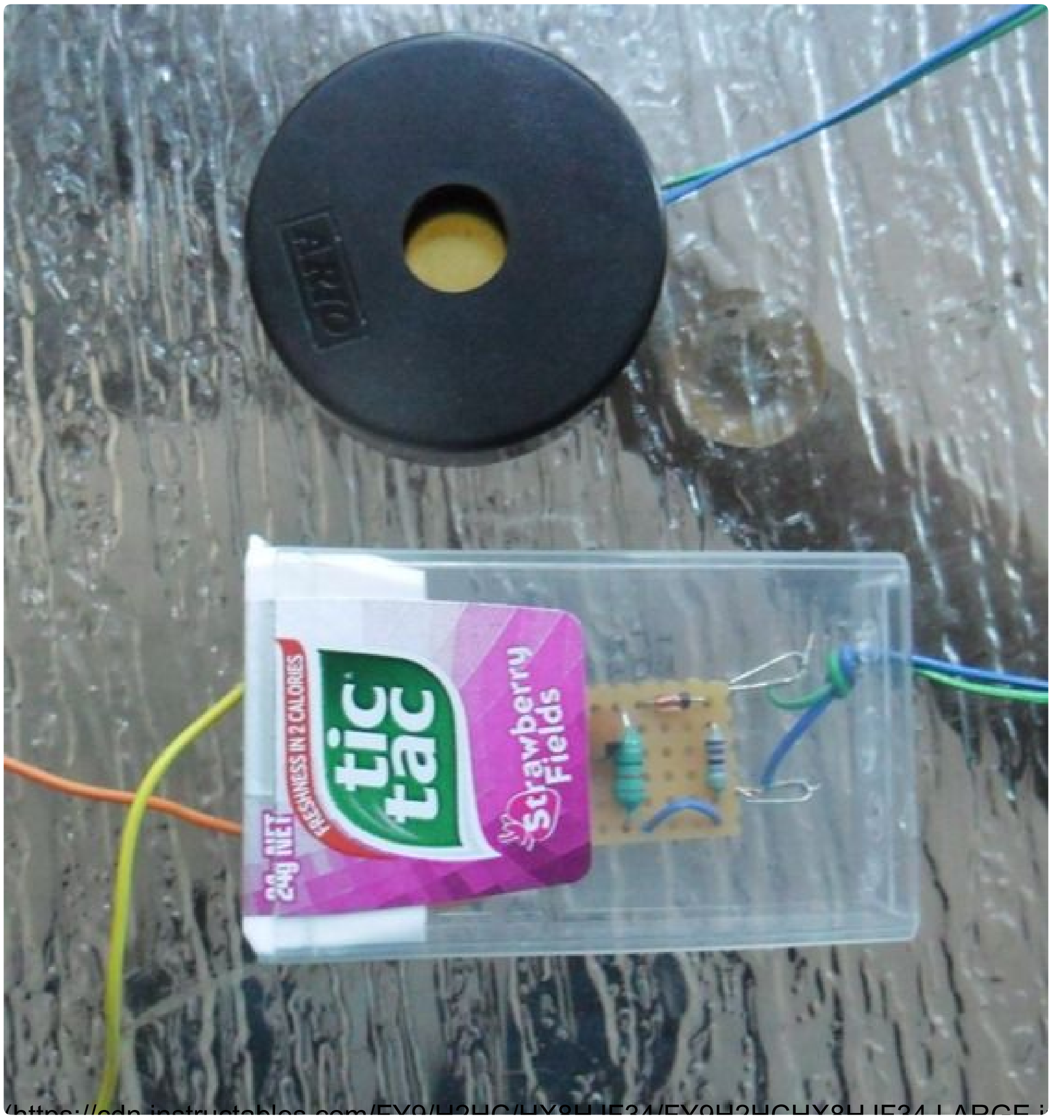
tic tac
Strawberry Fields
24g NET
FRESHNESS IN 2 CALORIES

(https://cdn.instructables.com/FX9/H2HC/HY8HJF34/FX9H2HCHY8HJF34.LARGE.j

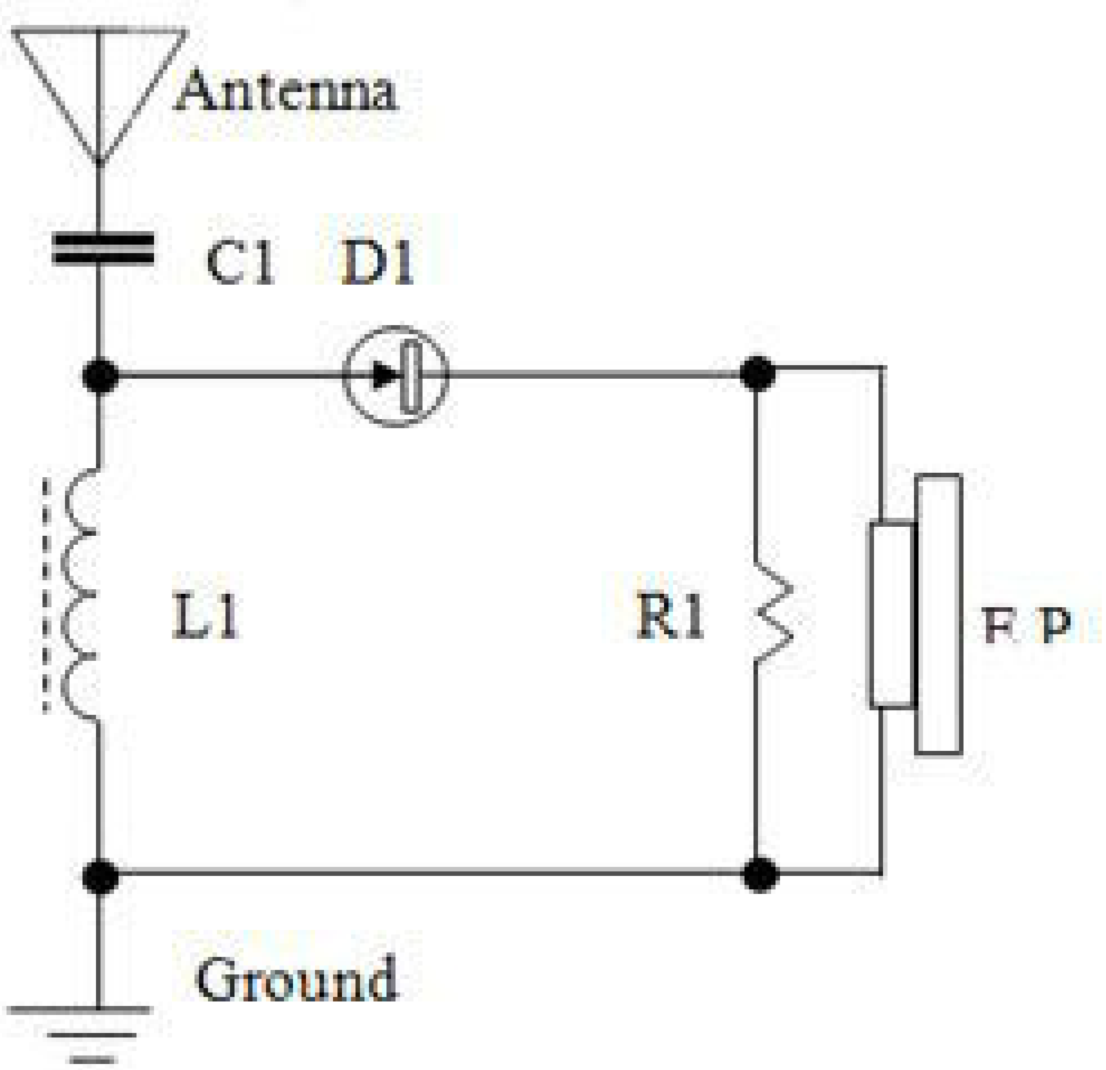

(https://cdn.instructables.com/FL6/41GU/HY5H55A2/FL641GUHY5H55A2.LARGE.jpg)

Take your finished Tic Tac Radio (with the lid part gently pushed back inside the top of the box) and connect your Antenna and Ground wires to the lead outs from the box lid. Place the earphone in/over your ear,and listen carefully for a local AM radio station. This crystal radio is a simple one, and you may have to make one or two adjustments to the components, before you succeed in receiving one or more local AM radio stations, in your area.

If you can't hear anything in the earphone, don't panic. It might just be a simple wiring mistake, which is easily fixed. Go back over all of the steps, making sure that you have the right value components from the electronics store. Make sure that each component is in the right place on the matrix board, (don't confuse the L1 inductor with the R1 resistor - they look a lot like each other!) and that no component wires have come undone from the twisted joins you have just made. I'll be writing up a troubleshooting step soon, so if you run into any problems, post your questions in the 'comments' section down at the bottom of the page, and I'll try and answer them as soon as possible.

Make sure that your antenna and ground wires aren't snagged on anything metal or anything dangerous!

If you have ANY doubts about the electrical safety of your antenna or ground wires, then consult a licensed electrical trades person, who will be able to advise you on electrical safety principles and procedures!

Always remember that electrical safety is your responsibility! If you don't think it is safe to proceed, then don't!

You will need at least 10 to 15 metres of antenna wire, strung between 2 insulating points (not connected to anything metallic, or that gets wet,) at least 2 to 3 metres in height - anything less than this minimum arrangement may mean that you cannot receive any signals at all.

Some places are regarded as 'radio dead spots', so you may need to try an open space, such as a park or a remote corner of a beach. If you do erect antenna and ground wires in public places, hang some streamers or balloons off of them, to alert people to their presence, otherwise people going past may become entangled with them - and get cranky with you!!!

As a final word for now, you'll be happy to know that the completed "Nic Nac Tic Tac Radio" shown in the picture above, picked up local AM radio station 1116 khz 4BC here in Brisbane, with a very clear signal and quite good volume. It works! So be safe kids, have some fun and look forward to more... mk484

Add Tip Ask Question

Step 6: Nic Nac Extras

(https://cdn.instructables.com/FCL/BHOH/HX8HJG9V/FCLBHOHHX8HJG9V.LARGE.jpg)

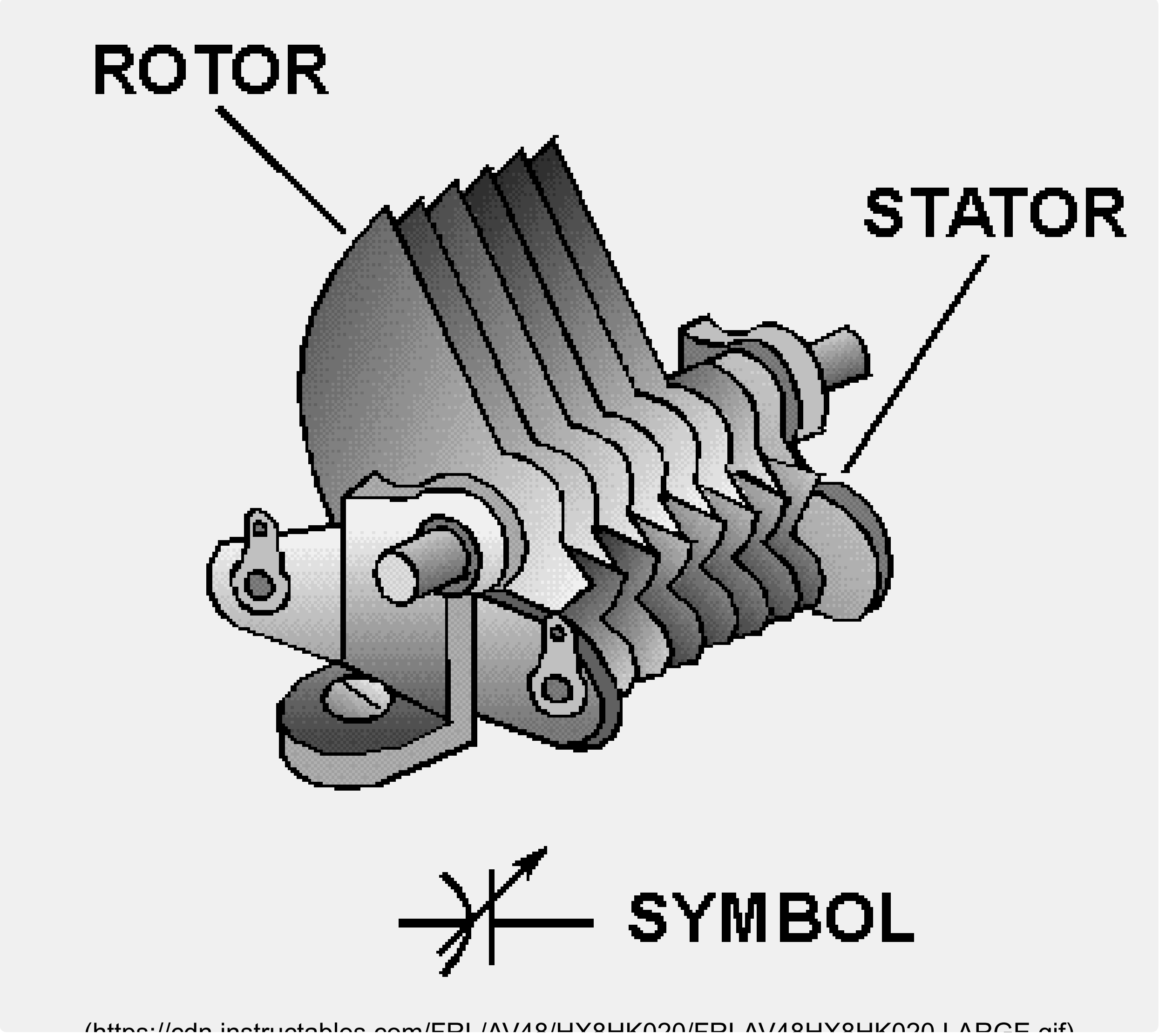
ROTOR
STATOR
SYMBOL

(https://cdn.instructables.com/FBL/AV48/HX8HK020/FBLAV48HX8HK020.LARGE.gif)

If you've built your Tic Tac crystal set and found that you can't receive a local AM radio station yet - don't panic - help is in on the way. You have just built the simplest version of the Tic Tac crystal radio and you may need to add one more component part for it to work properly. This is called a 'polyvaricon' - a miniature tuning capacitor, which can vary the frequency that your radio will receive at. You can see a picture of one up above - the small white box with the black knob next to it. It has 3 connecting tags - the one in the middle ('G') goes to the moving plates and the shaft, while the 2 outer tags ('O' and 'A') go to 2 sets of fixed plates. the smaller set of plates has a value of 60pF - pF is short for 'picoFarad' - a unit of measurement for capacitance. The larger set of plates is valued at 160 pF so that the combined value of the polyvaricon is 220 pF - or 220 picoFarads.

The other picture shows you what happens inside a basic tuning capacitor. there are 2 sets of metal plates - one set is fixed and the other set moves on a rotating shaft, connected to the tuning knob of your radio set. Both sets of plates are mounted on an insulating frame so that they won't 'short out' by touch each other.

The fixed capacitor C1, was chosen to tune somewhere close to the middle of the AM radio band. This band of frequencies starts at 531 kilohertz (Khz) and goes as far as 1701 Khz here in Australia. So we need a combination of coil and capacitor which will tune across all of those frequencies. Our simple Tic Tac radio is known as a 'series tuned' set. If you look at the circuit diagram, you can see electronic symbols for all components in the radio set.If you start at the top with the antenna symbol, you can see the capacitor C1 underneath that, the inductor L1 underneath C1 and then finally, the ground symbol - all wired in series with each other.

The diode D1 detector, the resistor and the earphone plus the link wire, can be considered as one block - the "detector unit". The Antenna wire, ground wire, inductor/coil and capacitor/polyvaricon, can be also be considered as another block - the "tuned circuit". So joining both blocks together, we have the tuned circuit that tunes in only one frequency, passing all other unwanted signals out through the ground wire to earth. This one 'tuned frequency' passes through the diode detector, which strips away the 'carrier wave' and leaves only the 'program signal' (voice, music etc,) behind, which is then fed via the resistor into the earphone. The earphone changes electrical impulses from the diode detector, into sound waves that we can hear. You need the R1 resistor to provide a pathway for the signals going through the diode, out to the ground wire connection. Without this resistor, the signals would sound very distorted and you couldn't hear the program signal very clearly.

Add Tip Ask Question

Step 7: Adding the Polyvaricon Tuning Capacitor

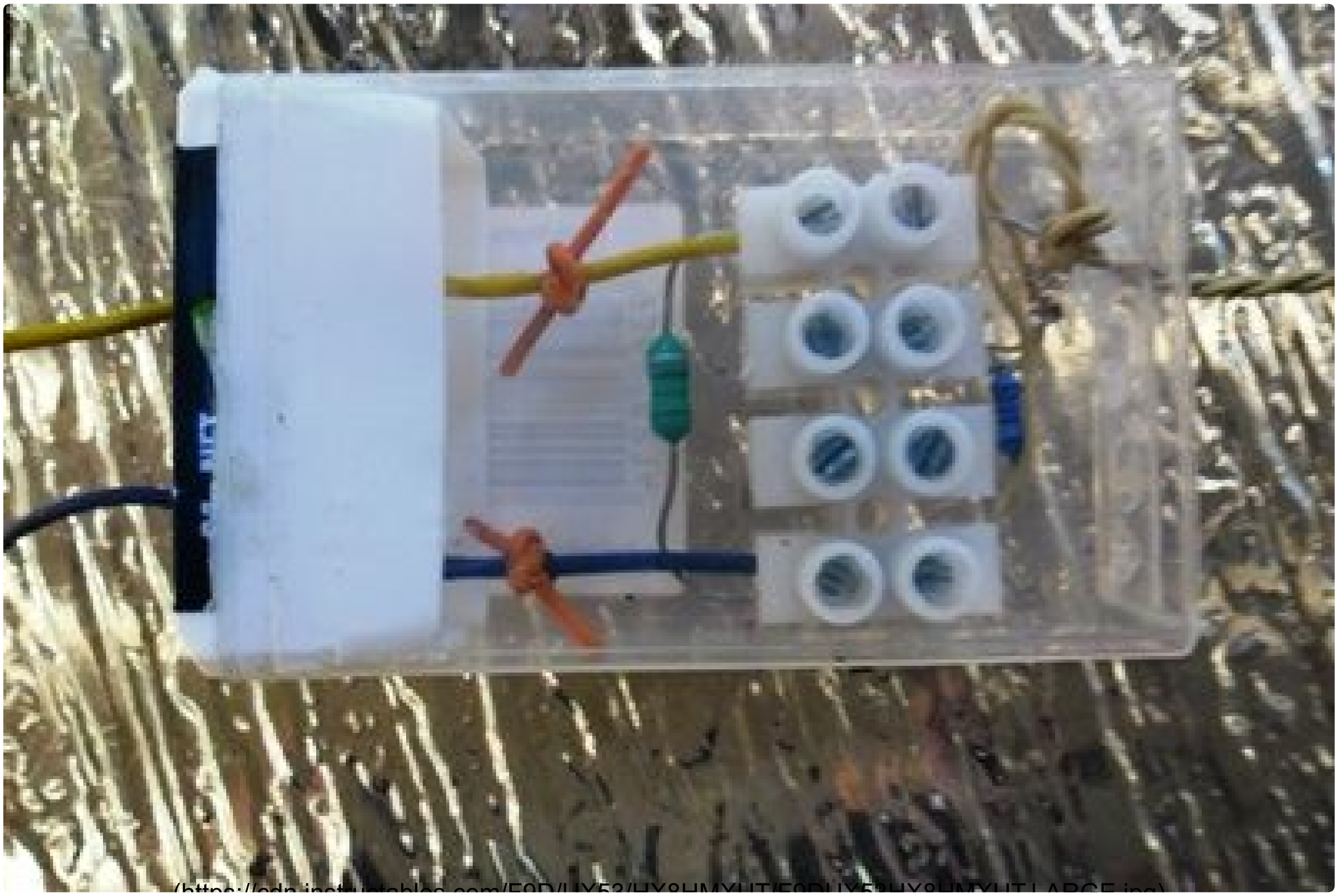

(https://cdn.instructables.com/F9D/UX53/HX8HMYUT/F9DUX53HX8HMYUT.LARGE.jpg)

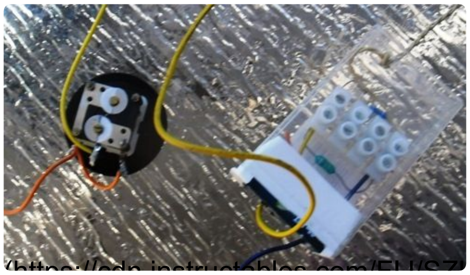

(https://cdn.instructables.com/FI/SZ

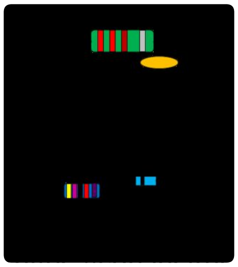

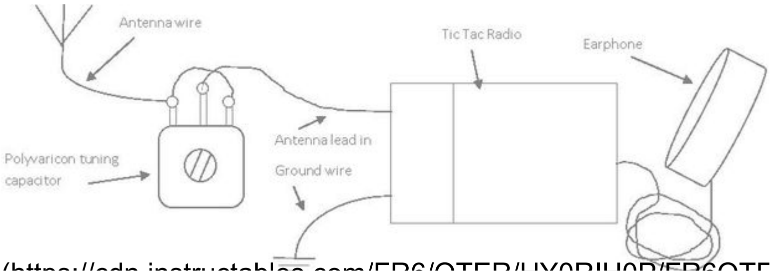

(https://cdn.instructables.com/FB6/QTEP/UX0BIU0P/FB6QTE

If you want to expand the tuning range of your Tic Tac Radio, than all you have to do is a simple modification (change) to the circuit of your radio set. You can see from the 2 pictures above, that there's an alternative way of building the crystal radio - you can use the matrix board method or you can use a 4 way screw terminal strip.

To use the screw terminal version, cut yourself a 4 way strip of terminals as shown in the diagram and pictures. Undo the screws right out as far as they will go without falling out of the strip. Connect the component leads and wires from the antenna/ground system as well as the earphone wires. Cut off any excess from component leads that you don't need. Wrap thin wires from the antenna/ground system and the earphones, around the thicker component leads before screwing the screws in tight.

Regardless of which method you have used, all we will now wire up the tuning capacitor between the actual Antenna wire and the junction (join) where the diode D1, the fixed capacitor C1 and one end of the inductor meet. Remove the capacitor from the matrix board, and connect the Antenna lead out wire straight to the join of the diode and inductor. Then take the polyvaricon and another piece of wire. Strip both ends of that wire and join the two outer tags (tagged 'O' and 'A' - the centre on is tagged 'G') and then connect you actual Antenna wire to one of the outer tags. Connect the antenna lead out wire, coming out of the box, to the 'G' (middle) tag of the tuning capacitor,so that the antenna wiring now looks like the picture up above.

If you're having problems following the pictures, then refer to the diagrams, which clearly shows all of the connections Make sure your ground wire is connected to the set, and then, listening with your earphone in/over your ear, slowly turn the tuning cap's flat knob, until you hear one or more stations. Congratulations - you now have a "tuneable" Tic Tac crystal set! Happy listening! And don't forget to post in your results, questions, problems etc... mk484 :)

Add Tip Ask Question

Step 8: Tic Tac One Transistor Amplifier - Use Your Earbuds and Hear Great Sound...

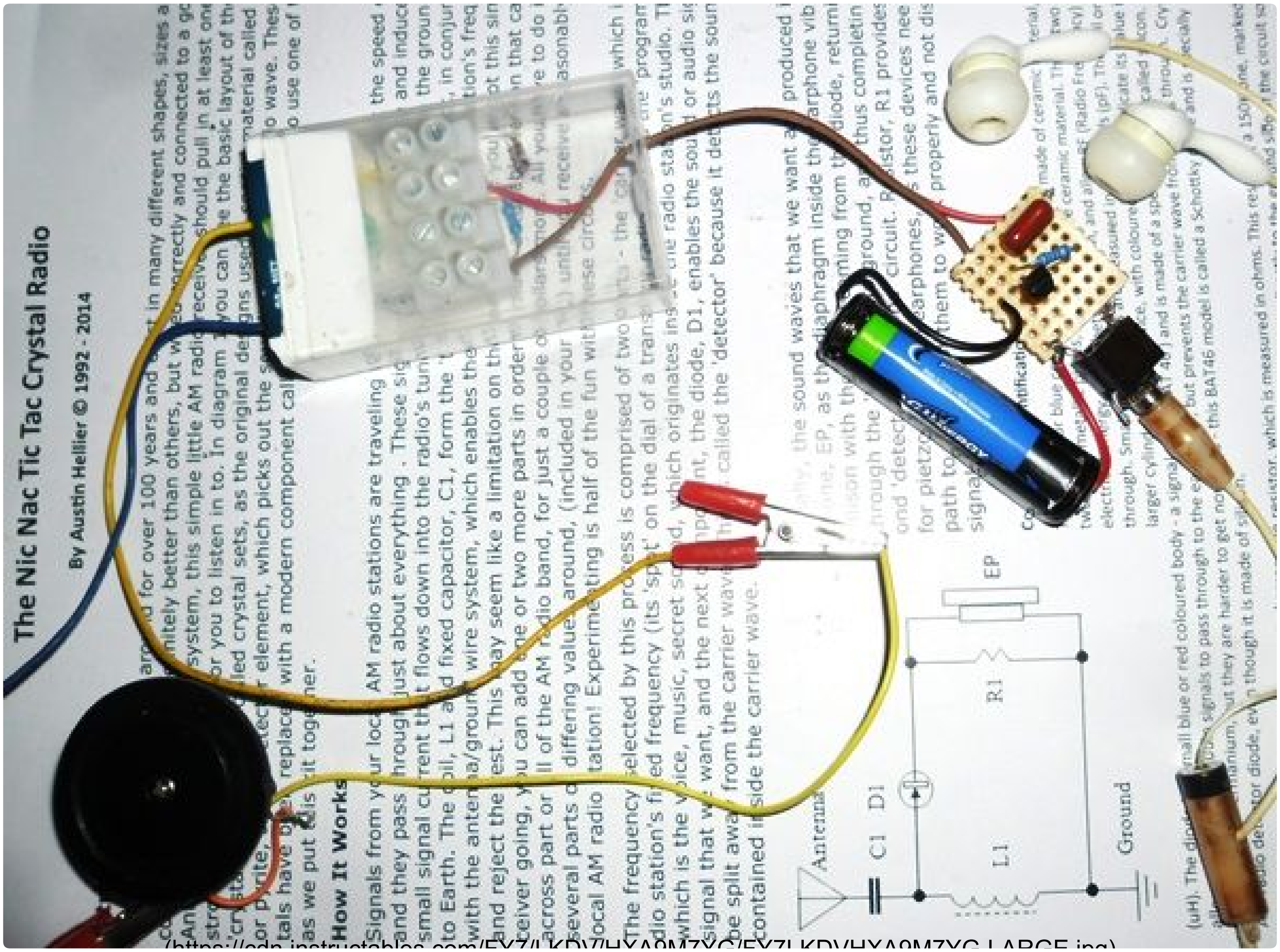
The Nic Nac Tic Tac Crystal Radio
By Austin Hellier © 1992 - 2014
How It Works
EP
R1
D1
C1
L1
Antenna
Ground

(https://cdn.instructables.com/FYZ/LKDV/HYA9MZYG/FYZLKDVHYA9MZYG.LARGE.jpg)

Hi there folks - this isn't really an extra 'Step' as such - it's a sneak preview at a picture of my upcoming Instructable for a one transistor amplifier, which will connect to the Tic Tac Radio - and give you some really good volume - in your Iphone earbuds!

Yes, it's entirely possible nowadays, with modern circuit design, to fit a one transistor amp in such a small space (yes kiddies - it WILL fit inside the smaller Tic Tac box...) and at the same time, get that great sound that comes from those "inside your ear" type earbuds.

This circuit took me about a half hour to complete, uses only 3 electronic components, costing about $1.00, a stereo earphone jack and a 1.5 volt AAA battery and plastic holder costing about another $3 to $4, so you can build the amplifier for about $5 all up - don't forget to shop around and - mums and dads - cheap batteries will be OK for this project, and the battery can simply be replaced without soldering or undoing half a dozen screws....

I can hear my Tic Tac Radio with great volume and clarity - and so will you, so "stay tuned" (ha ha ha) to this series of Instructables, and you will hear great sound too...

mk484

Add Tip

Ask Question

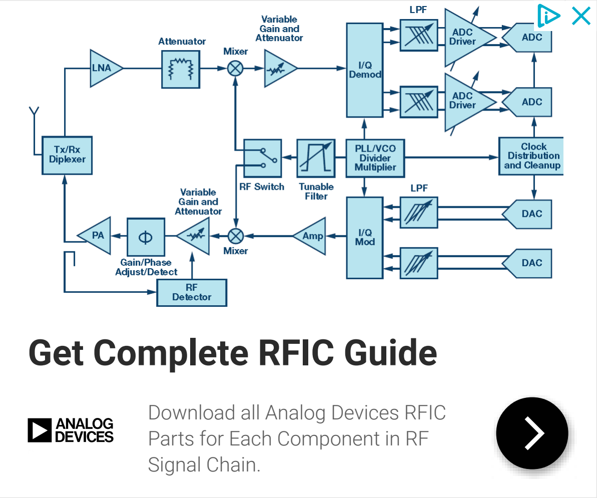
LNA
Attenuator
Mixer
Variable Gain and Attenuator
I/Q Demod
LPF
ADC Driver
ADC
Tx/Rx Diplexer
RF Switch
Tunable Filter
PLL/VCO Divider Multiplier
Clock Distribution and Cleanup
DAC
I/Q Mod
Amp
PA
Gain/Phase Adjust/Detect
RF Detector
ANALOG DEVICES
Get Complete RFIC Guide

Download all Analog Devices RFIC Parts for Each Component in RF Signal Chain.

advertisement

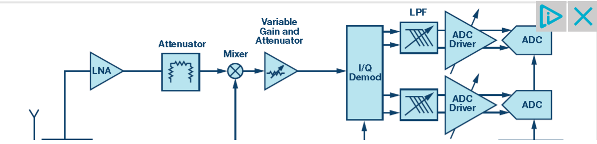
LPF
Variable Gain and Attenuator
Attenuator
Mixer
LNA
I/Q Demod
ADC Driver
ADC

go to: Talking Electronics Website

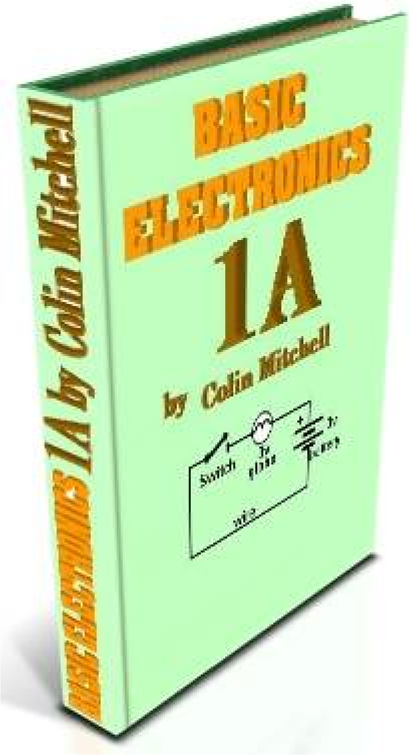

For any enquiries email Colin Mitchell

THE MULTIMETER

Page 1: Basic Electronics
- The capacitor - how it works
- The Diode - how the diode works
- Circuit Symbols - EVERY Circuit Symbol
- Soldering - videos

Page 2: The Transistor
- - PNP or NPN Transistor TEST

Page 2a: The 555 IC
- The 555 - 1
- The 555 - 2
- The 555 - 3
- The 555 TEST

Page 3: The Power Supply download as .pdf (900kB)
- 3a: - Constant Current
- 3b: - Voltage Regulator
- 3c: - Capacitor-fed Power Supply

Page 4: Digital Electronics
- 4a: - Gates Touch Switch Gating

Page 5: Oscillators

Page 6: Test - Basic Electronics (50 Questions)

Page 7: The Multimeter - this page

INDEX

Analogue Multimeter
Buying A Multimeter
Current - measuring
Digital Multimeter
Diode - measuring
Pressure Cell
Resistance - measuring
TEST
Using A Multimeter
Voltage - measuring

to Index

TWO MULTIMETERS

There are basically two different types of MULTIMETER. ANALOGUE and DIGITAL
Analogue Multimeters have a NEEDLE or POINTER that moves across a scale.
Digital Multimeters have a numeric display of 3 or more digits. A Digital Multimeter with 3½ digits means the first digit shows only "1."

You really need both types to cover the number of tests needed for designing and repair-work. We will discuss how they work, how to use them and some of the differences between them.

The black (negative lead) ALWAYS stays in the

black hole and the red lead changes to the other red hole to measure 10 amps

to Index

BUYING A MULTIMETER

There are many different types on the market.
The cost is determined by the number of ranges and also the extra features such as diode tester, buzzer (continuity), transistor tester, high DC current and others.
Since most multimeters are reliable and accurate, buy one with the greatest number of ranges at the lowest cost. The cheapest multimeters are on **eBay**.
This article explains the difference between an analogue meter and a digital meter.

Multimeters are sometimes called a "meter", a "VOM" (Volts-Ohms-Milliamps or Volt Ohm Meter) or "multi-tester" or even "a tester" - they are all the same.

One term used to describe a **DIGITAL MULTIMETER** is **3½ digits**.
This is the number of digits on the display. The first digit is usually made from two pixels and can only produce "1." This is called a half-digit. The other digits are full digits. The cheapest digital multimeters have 3½ digits. This will produce a reading of 1999 and the decimal point can produce values from 1.999 to 19.99 to 199.9 to 1999.
Another term is **DISPLAY COUNTS**. This is connected with the accuracy of the display, but since digital meters are accurate to 1% or less and we are using resistors with an accuracy of 5%, even a $10.00 digital meter will be perfect.

to Index

USING A MULTIMETER

Analogue and **digital** multimeters have either a rotary selector switch or push buttons to select the appropriate function and range. Some Digital Multimeters (DMMs) are auto ranging; they automatically select the correct range of voltage, resistance, or current when doing a test. However you need to select the function.

Before making any measurement you need to know what you are checking. If you are measuring voltage, select the AC range (10v, 50v, 250v, or 1000v) or DC range (0.5v, 2.5v, 10v, 50v, 250v, or 1000v). If you are measuring resistance, select the Ohms range (x1, x10, x100, x1k, x10k). If you are measuring current, select the appropriate current range DCmA 0.5mA, 50mA, 500mA, 10A. Every multimeter is different however the photo below shows a low cost Analogue multimeter with the basic ranges.

MODEL YX-1000A
Ω
X1K
X100
X10
0.5
50
DCmA
250
10
DCV
50
250
1000
ACV
10
50
250
1000

An ANALOGUE MULTIMETER

The most important point to remember is this:
You must select a voltage or current range that is bigger or HIGHER than the maximum expected value, so the needle does not swing across the scale and hit the "end stop."
If you are using a DMM (Digital Multi Meter), the meter will indicate if the voltage or current is higher than the selected scale, by showing "OL" - this means "Overload."
If you are measuring resistance such as 1M on the x10 range the "OL" means "Open Loop" and you will need to change the range. Some meters show "1' on the display when the measurement is higher than the display will indicate and some flash a set of digits to show over-voltage or over-current. A "-1" indicates the leads should be reversed for a "positive reading."
If it is an AUTO RANGING meter, it will automatically produce a reading, otherwise the selector switch must be changed to another range.

000
DCV
OFF
ACV
DCA
hFE
DT-832
DIGITAL
MULTIMETER
COM

A typical DIGITAL Multimeter

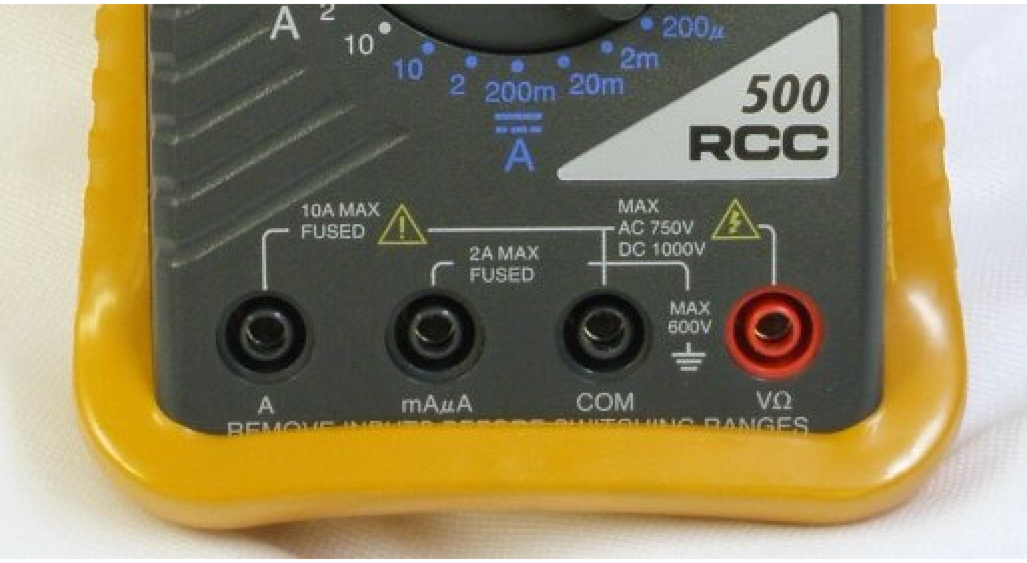
A
10
10
2
200m
20m
2m
200µ
A
500
RCC
10A MAX
FUSED
2A MAX
FUSED
MAX
AC 750V
DC 1000V
MAX
600V
A
mAµA
COM
VΩ

The Common (negative) lead ALWAYS fits into the "COM" socket. The red lead fits into the red socket for Voltage and Resistance. Place the red lead (red banana plug) into "A" (for HIGH CURRENT "Amps") or mA,uA for LOW CURRENT.

The black "test lead" plugs into the socket marked "-" "Common", or "Com," and the red "test lead" plugs into the meter socket marked "+" or "V-W-mA." The third banana socket measures HIGH CURRENT and the positive (red lead) plugs into this. You DO NOT move the negative "-" lead at any time.
The following two photos show the test leads fitted to a digital meter. The probes and plugs have "guards" surrounding the probe tips and also the plugs so you can measure high voltages without getting near the voltage-source.

FLUKE 115
600.0
HOLD
RANGE
OFF
COM
A
000
10ADC

Analogue meters have an "Ohms Adjustment" to allow for the change in voltage of the battery inside the meter (as it gets old).

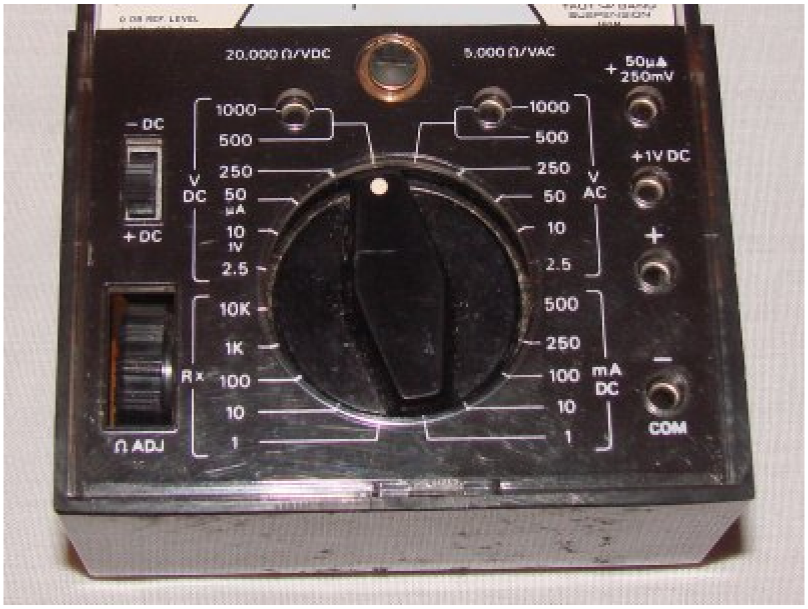

"Ohms Adjust" is also called "ZERO SET"
The sensitivity of this meter is 20,000ohms/volt on the DC ranges and 5k/v on the AC ranges

Before taking a resistance reading (each time, for any of the Ohms scales) you need to "ZERO SET" the scale, by touching the two probes together and adjust the pot until the needle reads "0" (swings FULL SCALE). If the pointer does not reach full scale, the batteries need replacing. Digital multimeters do not need "zero adjustment."

to Index

ANALOGUE Vs DIGITAL

You cannot say one meter is better than the other because BOTH have advantages and disadvantages.
An analogue multimeter is the "old style" and it puts a load on a circuit and this may change the reading to give an incorrect readout, but it has the advantage of the needle moving across the scale fairly quickly so you can sometimes see if the voltage is fluctuating.
It also gives a more-accurate result in some high frequency circuits as it does not pick up stray fields and produce a false reading.
Digital meters put almost no load on a circuit and produce accurate readings from both low-impedance and high-impedance circuits.
Digital meters can display very low resistances.
You must remember to turn a Digital meter OFF to prevent the battery going flat.

If you are testing a circuit containing a high-frequency oscillator, use BOTH an ANALOGUE and DIGITAL meter to check the reading. Sometimes the leads of a Digital multimeter will pick up signals and create a false reading.
Sometimes you will get a voltage reading with a Digital multimeter due to a high resistance leak and a zero reading with an Analogue meter. This is why you need **BOTH** meters.

to Index

MEASURING VOLTAGE

Most of the readings taken with a multimeter will be VOLTAGE readings.
Before taking a reading, you should select the highest range and if the needle does not move up scale (to the right), you can select another range.
Always switch to the highest range before probing a circuit and keep your fingers away from the component being tested.
If the meter is Digital, select the highest range or use the auto-ranging feature, by selecting "V." The meter will automatically produce a result, even if the voltage is

AC or DC.
If the meter is not auto-ranging, you will have to select V⎓ if the voltage is from a DC source or V~ if the voltage is from an AC source. DC means Direct Current (but this does not mean you select the CURRENT range - you are taking a voltage reading that is not rising and falling. That's why we say it is **DC** and do not say the words "direct current"). The voltage is coming from a battery or supply where it is steady and not "rising and falling."
You can measure the voltage at different points in a circuit by connecting the black probe to chassis. This is the 0v reference and is commonly called "Chassis" or "Earth" or "Ground" or "0v."
The red lead is called the "measuring lead" or "measuring probe" and it can measure voltages at any point in a circuit. Sometimes there are "test points" on a circuit and these are wires or loops designed to hold the tip of the red probe (or a red probe fitted with a mini clip).
You can also measure voltages ACROSS A COMPONENT. In other words, the reading is taken in PARALLEL with the component. It may be the voltage across a transistor, resistor, capacitor, diode or coil. In most cases this voltage will be less than the supply voltage.
If you are measuring the voltage in a circuit that has a HIGH IMPEDANCE, the reading will be inaccurate, up to 90% !!!, if you use a cheap analogue meter.

Here's a simple case.

The circuit below consists of two 1M resistors in series. The voltage at the mid point will be 5v when nothing is connected to the mid point. But if we use a cheap analogue multimeter set to 10v, the resistance of the meter will be about 100k, if the meter has a sensitivity of 10k/v and the reading will be incorrect.
Here how it works:
Every meter has a sensitivity. The sensitivity of the meter is the sensitivity of the movement and is the amount of current required to deflect the needle FULL SCALE. This current is very small, normally 1/10th of a milliamp and corresponds to a sensitivity of 10k/volt (or 1/30th mA, for a sensitivity of 30k/v).
If an analogue meter is set to 10v, the internal resistance of the meter will be 100k for a 10k/v movement.
If this multimeter is used to test the following circuit, the reading will be inaccurate.
The reading should be 5v as show in diagram **A**.
But the analogue multimeter has an internal resistance of 100k and it creates a circuit shown in **C**.
The top 1M and 100k from the meter create a combined PARALLEL resistance of 90k. This forms a series circuit with the lower 1M and the meter will read less than 1v
If we measure the voltage across the lower 1M, the 100k meter will form a value of resistance with the lower 1M and it will read less than 1v
If the multimeter is 30k/v, the readings will be 2v. See how easy it is to get a totally inaccurate reading.

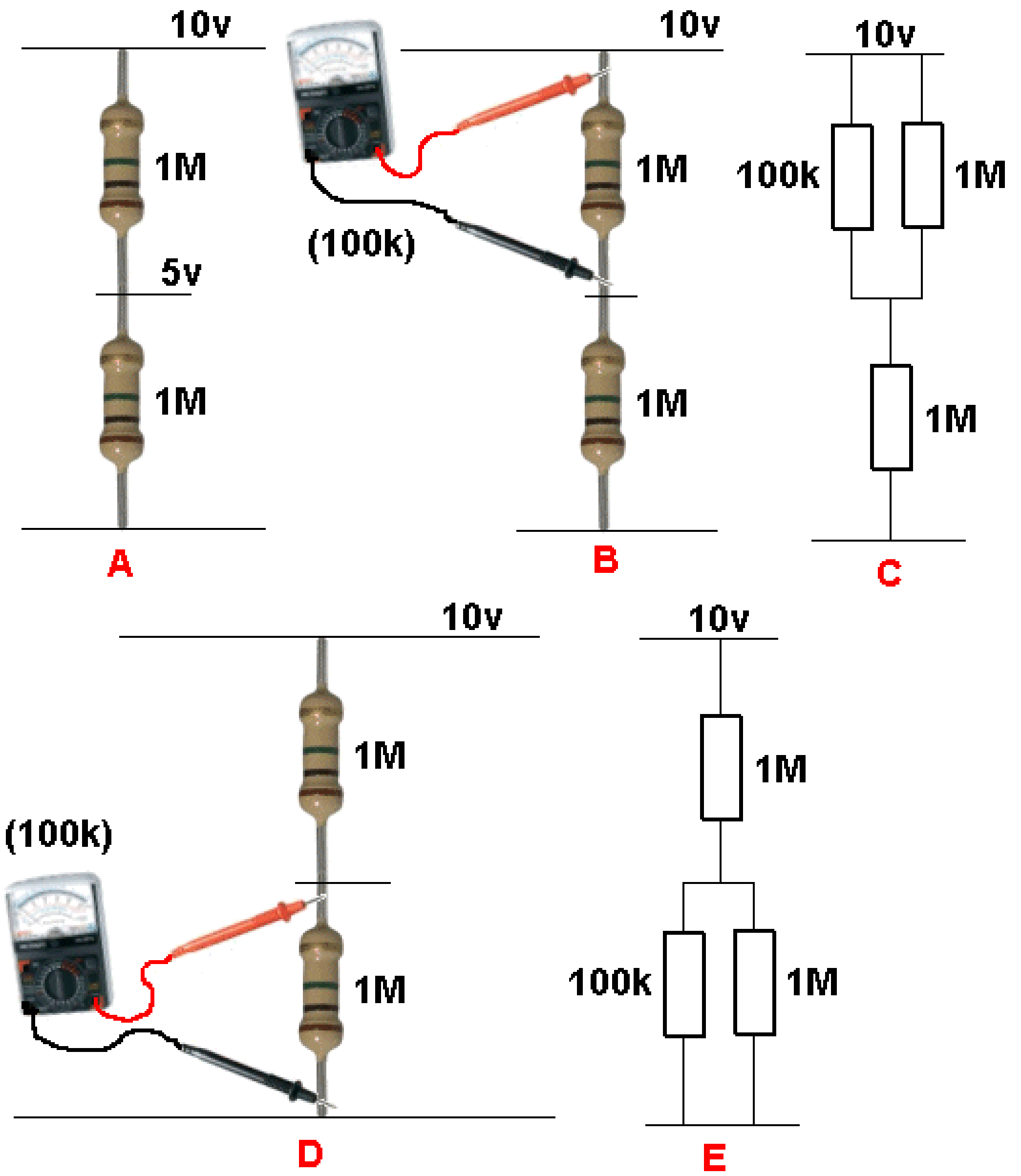

If the reading is taken with a Digital Meter, it will be more accurate as a DMM does not take any current from the circuit (to activate the meter). In other words it has a very HIGH input impedance. Most Digital Multimeters have a fixed input resistance (impedance) of 10M - no matter what scale is selected. That's the reason for choosing a DMM for high impedance circuits. It also gives a reading that is accurate to about 1%.

to Index

MEASURING VOLTAGES in a CIRCUIT

You can take many voltage-measurements in a circuit. You can measure "across" a component, or between any point in a circuit and either the positive rail or earth rail (0v rail). In the following circuit, the 5 most important voltage-measurements are shown. Voltage "A" is across the electret microphone. It should be between 20mV and 500mV. Voltage "B" should be about 0.6v. Voltage "C" should be about half-rail voltage. This allows the transistor to amplify both the positive and negative parts of the waveform. Voltage "D" should be about 1-3v. Voltage "E" should be the battery voltage of 12v.

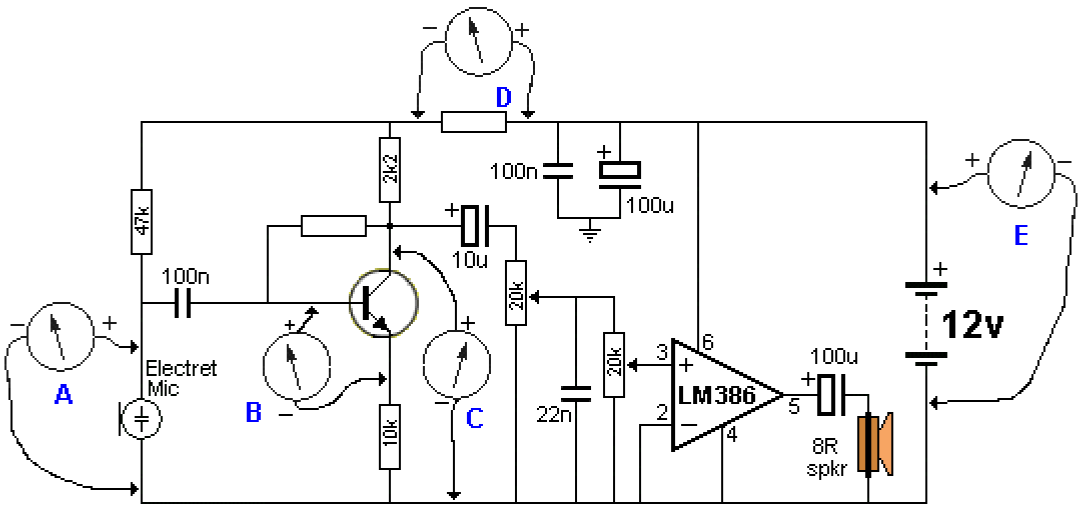

to Index

MEASURING CURRENT

You will rarely need to take current measurements, however most multimeters have DC current ranges such as 0.5mA, 50mA, 500mA and 10Amp (via the extra banana socket) and some meters have AC current ranges. Measuring the current of a circuit will tell you a lot of things. If you know the normal current, a high or low current can let you know if the circuit is overloaded or not fully operational.

Current is always measured when the circuit is working (i.e: with power applied).
It is measured IN SERIES with the circuit or component under test.
The easiest way to measure current is to remove the fuse and take a reading across the fuse-holder. Or remove one lead of the battery or turn the project off, and measure across the switch.
If this is not possible, you will need to remove one end of a component and measure with the two probes in the "opening."
Resistors are the easiest things to desolder, but you may have to cut a track in some circuits. You have to get an "opening" so that a current reading can be taken.
The following diagrams show how to connect the probes to take a CURRENT reading.
Do not measure the current ACROSS a component as this will create a "short-circuit."
The component is designed to drop a certain voltage and when you place the probes across this component, you are effectively adding a "link" or "jumper" and the voltage at the left-side of the component will appear on the right-side. This voltage may be too high for the circuit being supplied and the result will be damage.

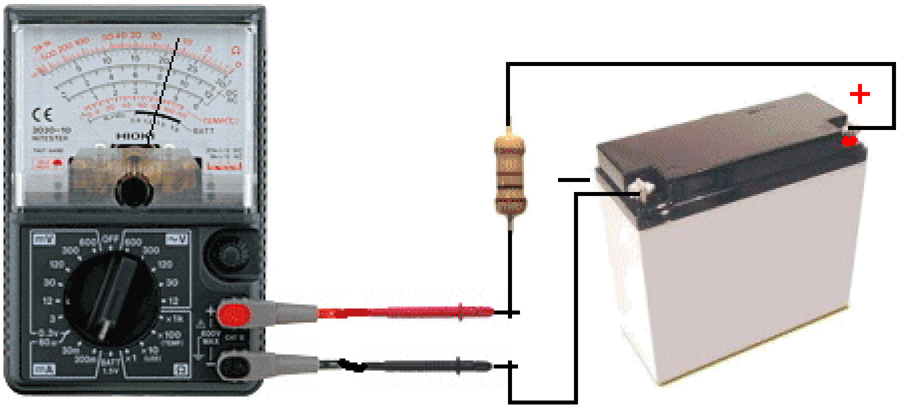

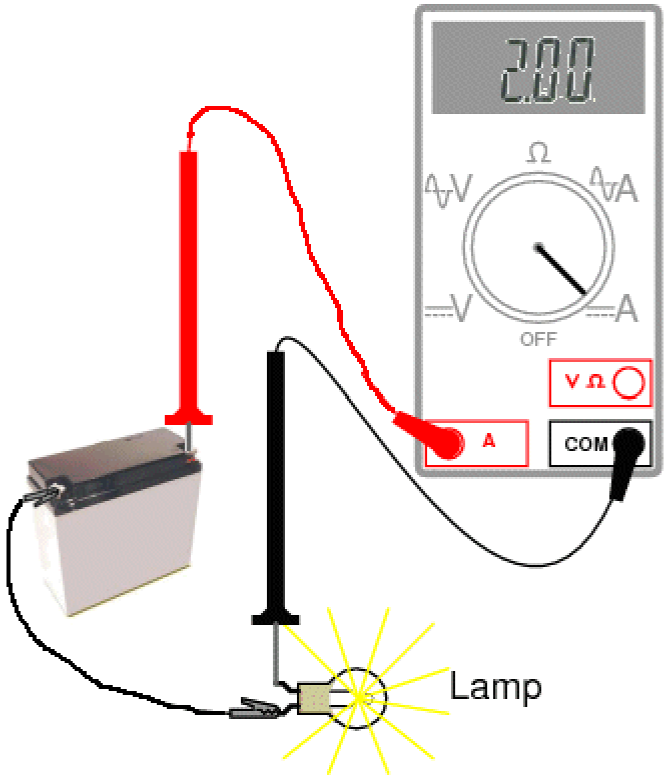

Measuring the current of a globe

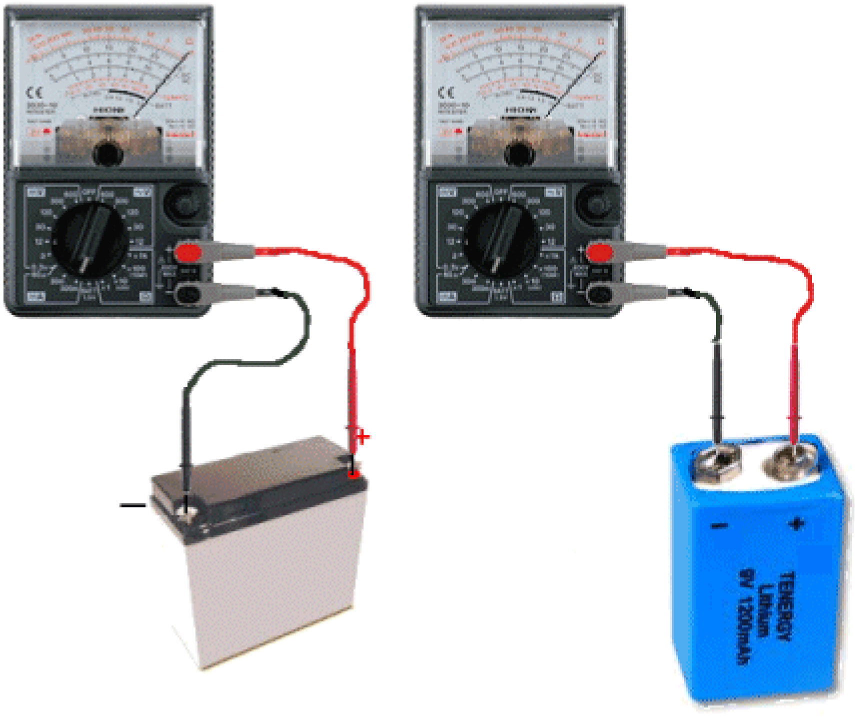

Do NOT measure the CURRENT of a battery
(by placing the meter directly across the terminals)
A battery will deliver a very HIGH current
and damage the meter

Do not measure the "current a battery will deliver" by placing the probes across the terminals. It will deliver a very high current and damage the meter instantly. There are special battery testing instruments for this purpose.

When measuring across an "opening" or "cut," place the red probe on the wire that supplies the voltage (and current) and the black probe on the other wire. This will produce a "POSITIVE" reading.
A positive reading is an UPSCALE READING and the pointer will move across the scale - to the right. A "NEGATIVE READING" will make the pointer hit the "STOP" at the left of the scale and you will not get a reading. If you are using a Digital Meter, a negative sign "-" will appear on the screen to indicate the probes are around the wrong way. No damage will be caused. It just indicates the probes are connected incorrectly.
If you want an accurate CURRENT MEASUREMENT, use a digital meter.

to Index

MEASURING RESISTANCE

Turn a circuit off before measuring resistance.
If any voltage is present, the value of resistance will be incorrect.
In most cases you cannot measure a component while it is in-circuit. This is because the meter is actually measuring a voltage across a component and calling it a "resistance." The voltage comes from the battery inside the meter. If any other voltage is present, the meter will produce a false reading.
If you are measuring the resistance of a component while still "in circuit," (with the power off) the reading will be lower than the true reading.

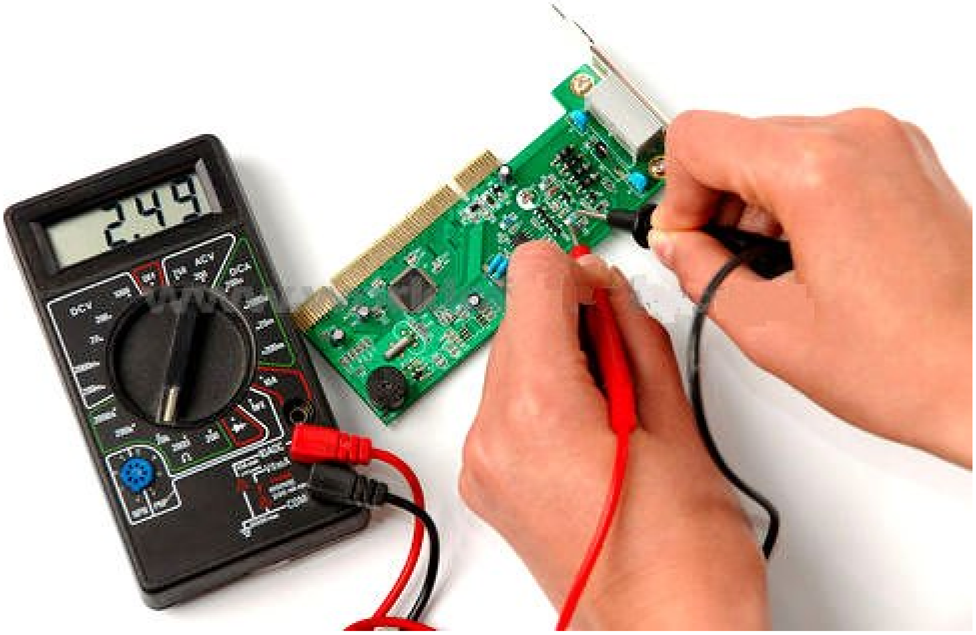

Measuring resistance

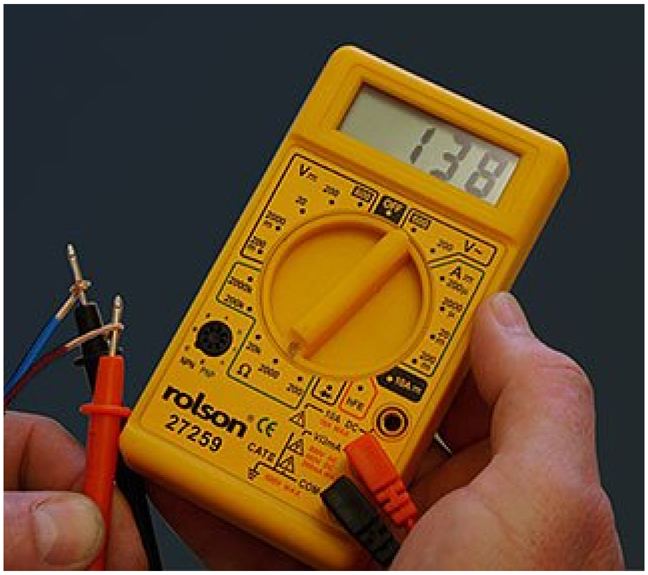
138
rolson
27259

**Measuring resistance of a heater
(via the leads)**

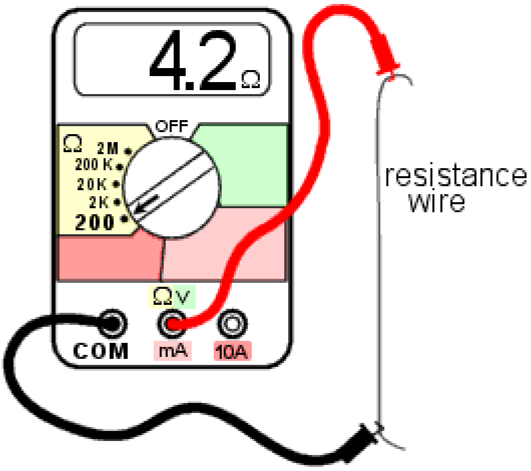
4.2Ω
OFF
Ω 2M
200 K
20K
2K
200
ΩV
COM
mA
10A
resistance wire

Measuring the resistance of a piece of resistance-wire

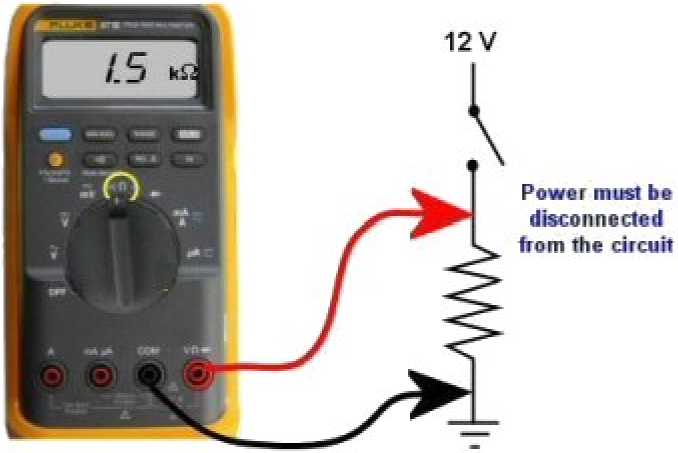
1.5 kΩ
12 V
Power must be disconnected from the circuit

Measuring the resistance of a resistor

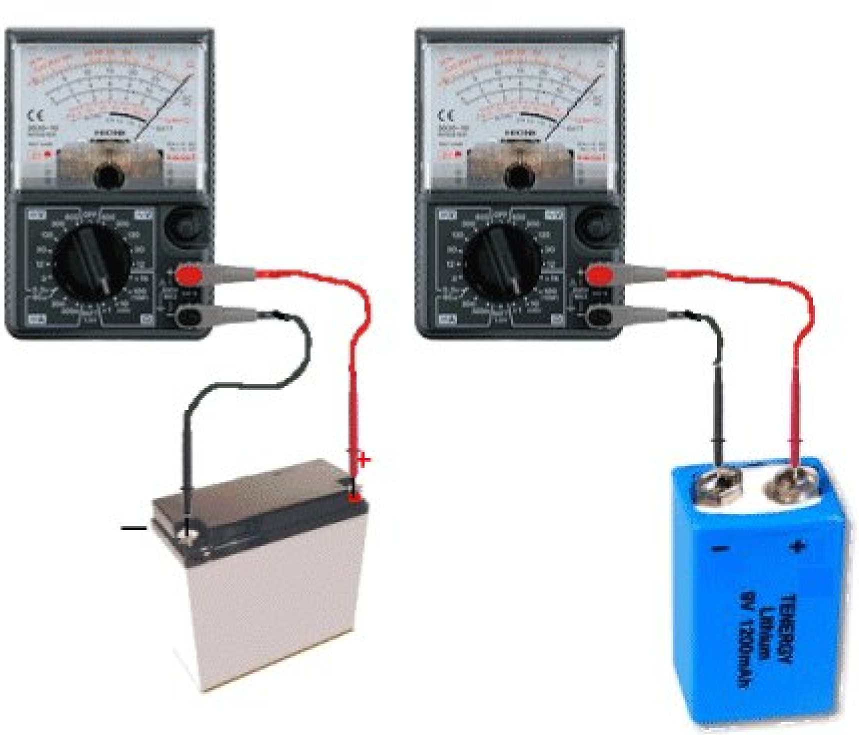

Do not measure the "Resistance of a Battery"

1. Do not measure the "resistance of a battery." The resistance of a battery (called the Internal impedance) is not measured as shown in the diagrams above. It is measured by creating a current-flow and measuring the voltage across the battery. Placing a multimeter set to **resistance** (across a battery) will destroy the meter.
2. Do not try to measure the resistance of any voltage or any "supply."

Resistance is measured in OHMs.
The resistance of a 1cm x 1cm bar, one metre long is 1 ohm.
If the bar is thinner, the resistance is higher. If the bar is longer, the resistance is higher.
If the material of the bar is changed, the resistance is higher.
When carbon is mixed with other elements, its resistance increases and this knowledge is used to make RESISTORS.
Resistors have RESISTANCE and the main purpose of a resistor is to reduce the CURRENT FLOW.
It's a bit like standing on a hose. The flow reduces.
When current flow is reduced, the output voltage is also reduced and that why the water does not spray up so high. Resistors are simple devices but they produce many different effects in a circuit.
A resistor of nearly pure carbon may be 1 ohm, but when non-conducting "impurities" are added, the same-size resistor may be 100 ohms, 1,000 ohms or 1 million ohms.
Circuits use values of less than 1 ohm to more than 22 million ohms.

Resistors are identified on a circuit with numbers and letters to show the exact value of resistance - such as 1k 2k2 4M7
The letter Ω (omega - a Greek symbol) is used to identify (or express) (or represent) the word "Ohm."
But this symbol is not available on some word-processors, so the letter "R" is used. The letter "E" is also sometimes used and both mean "Ohms."
A one-ohm resistor is written "1R" or "1E." It can also be written "1R0" or "1E0."
A resistor of one-tenth of an ohm is written "0R1" or "0E1." The letter takes the place of the decimal point.
10 ohms = 10R
100 ohms = 100R

1,000 ohms = 1k (k= kilo = one thousand)
10,000 ohms = 10k
100,000 ohms = 100k
1,000,000 ohms = 1M (M = MEG = one million)

The size of a resistor has nothing to do with its resistance. The size determines the wattage of the resistor - how much heat it can dissipate without getting too hot.
Every resistor is identified by colour bands on the body, but when the resistor is a surface-mount device, numbers are used and sometimes letters.
You MUST learn the colour code for resistors and the following table shows all the colours for the most common resistors from 1/10th of an ohm to 22 Meg ohms for resistors with 5% and 10% tolerance.

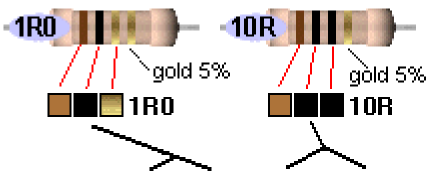

If 3rd band is gold, Divide by 10
If 3rd band is silver, Divide by 100
(to get 0.22ohms etc)

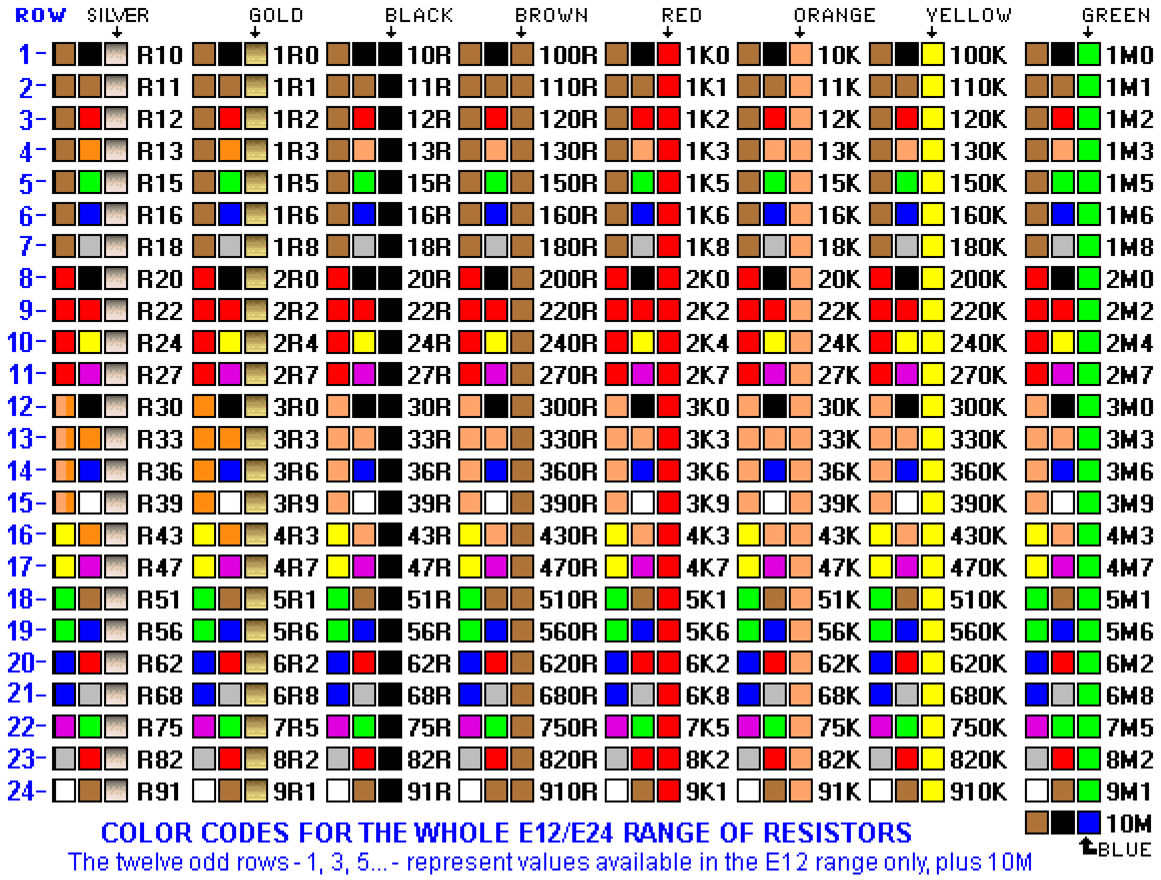

ROW	SILVER	GOLD	BLACK	BROWN	RED	ORANGE	YELLOW	GREEN
1-	R10	1R0	10R	100R	1K0	10K	100K	1M0
2-	R11	1R1	11R	110R	1K1	11K	110K	1M1
3-	R12	1R2	12R	120R	1K2	12K	120K	1M2
4-	R13	1R3	13R	130R	1K3	13K	130K	1M3
5-	R15	1R5	15R	150R	1K5	15K	150K	1M5
6-	R16	1R6	16R	160R	1K6	16K	160K	1M6
7-	R18	1R8	18R	180R	1K8	18K	180K	1M8
8-	R20	2R0	20R	200R	2K0	20K	200K	2M0
9-	R22	2R2	22R	220R	2K2	22K	220K	2M2
10-	R24	2R4	24R	240R	2K4	24K	240K	2M4
11-	R27	2R7	27R	270R	2K7	27K	270K	2M7
12-	R30	3R0	30R	300R	3K0	30K	300K	3M0
13-	R33	3R3	33R	330R	3K3	33K	330K	3M3
14-	R36	3R6	36R	360R	3K6	36K	360K	3M6
15-	R39	3R9	39R	390R	3K9	39K	390K	3M9
16-	R43	4R3	43R	430R	4K3	43K	430K	4M3
17-	R47	4R7	47R	470R	4K7	47K	470K	4M7
18-	R51	5R1	51R	510R	5K1	51K	510K	5M1
19-	R56	5R6	56R	560R	5K6	56K	560K	5M6
20-	R62	6R2	62R	620R	6K2	62K	620K	6M2
21-	R68	6R8	68R	680R	6K8	68K	680K	6M8
22-	R75	7R5	75R	750R	7K5	75K	750K	7M5
23-	R82	8R2	82R	820R	8K2	82K	820K	8M2
24-	R91	9R1	91R	910R	9K1	91K	910K	9M1
								10M BLUE

COLOR CODES FOR THE WHOLE E12/E24 RANGE OF RESISTORS
The twelve odd rows - 1, 3, 5... - represent values available in the E12 range only, plus 10M

to Index

MAKE YOUR OWN RESISTOR

Make your own variable resistor that changes resistance according to the pressure.
Use a piece of conductive foam used to package Integrated Circuits. You can ask at an electronics shop.
Use two coins or pieces of printed circuit board or aluminium foil for the top and bottom conductors.
You can solder wires to the PC board or fold the aluminium foil over a few times to hold the wires.
The resistance of the foam will reduce as you press on the "cell."
The actual resistance-values will depend on the size of the foam, the thickness and

pressure.
This cell is a very simple cell called a **LOAD CELL**.

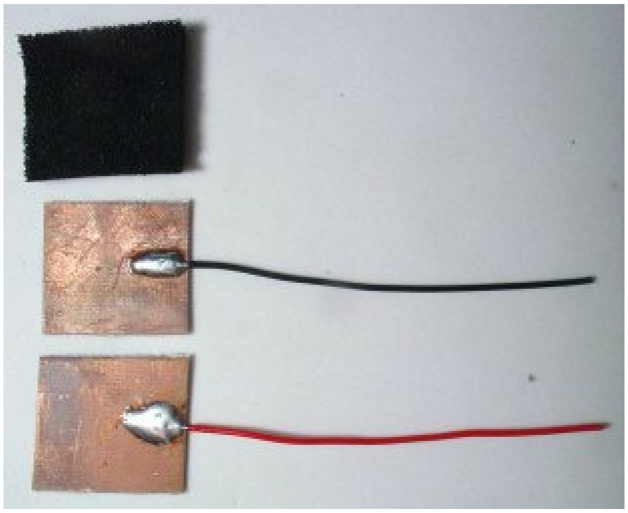

The top and bottom "plates"

The foam is placed between the plates.

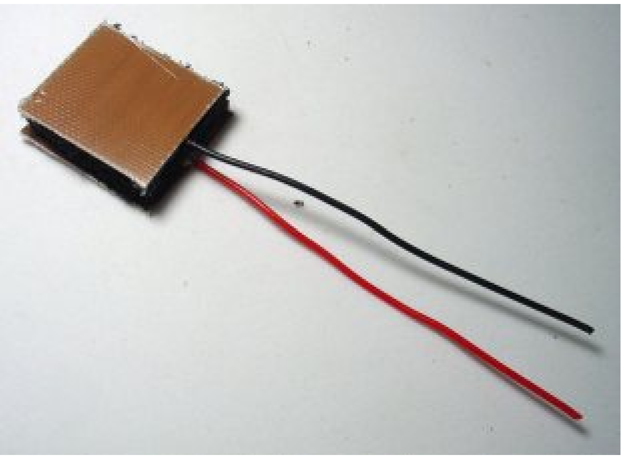

The complete LOAD CELL

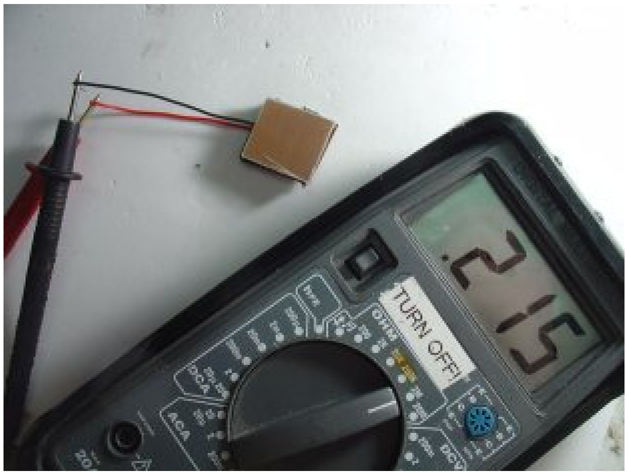

The resistance of the unloaded LOAD CELL

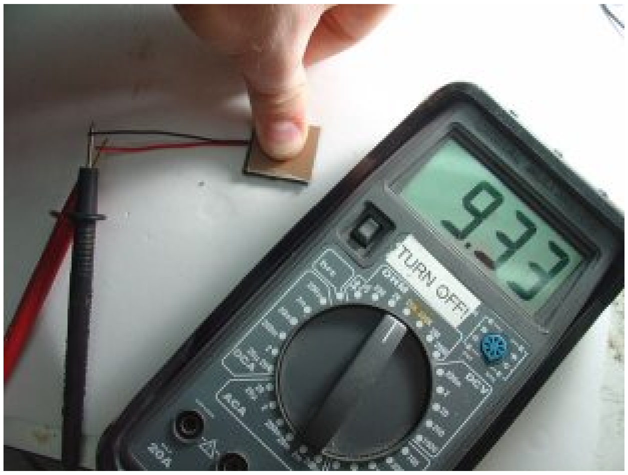

The fully loaded resistance can be as low as 9,330 ohms

to Index

MEASURING CONTINUITY

CONTINUITY is the same as **ZERO OHMS** or the resistance of a short length of wire. It can also mean the resistance through a switch or globe or a low-value resistor.
It basically means a "PATH" and sometimes refers to a whole circuit when the switch is closed. In other words CONTINUITY means we have a "circuit." We have "current flowing" and generally refers to a low-resistance circuit.
Both ANALOGUE and DIGITAL multimeters can measure CONTINUITY and you have to work out the approximate value of resistance for the circuit you are testing, - BEFORE TAKING A READING.
If the reading is above 300 ohms or contains a diode, you cannot use a DIGITAL MULTIMETER as the buzzer on the continuity setting will not respond.

The project being tested must not have the power applied as the resistance ranges on a multimeter are actually measuring a voltage across the leads and any voltage on the circuit or contained in any electrolytics, will upset the reading.

To take a reading with an **ANALOGUE** multimeter, select the x1 setting and the pointer will move across the scale to the actual value of resistance.
It it move full scale, you have ZERO OHMS resistance and this can mean a short-circuit or continuity via a wire.
If a diode is in the circuit you must also reverse the leads to get a reading.
The resistance of a globe will be very low when it is not illuminated, so don't think a fault is present.
Measuring CONTINUITY is the same as measuring LOW RESISTANCE.

To take a reading with a **DIGITAL** multimeter, select the buzzer setting. It will respond if the resistance is less than 300 ohms. It will not respond if a diode is in the circuit.

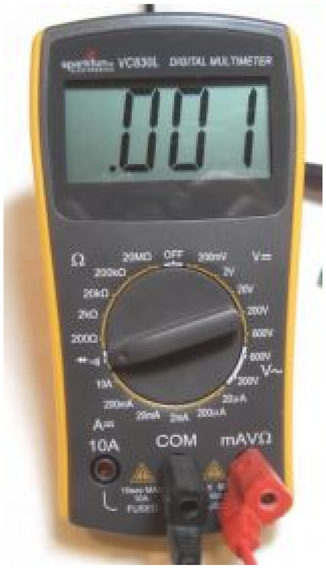

Meter set to BUZZER - CONTINUITY

You can also use the x1 resistance setting to get an accurate value of resistance. Touch the probes together to get the initial reading and subtract this value from the final reading.
When probing a circuit containing electrolytics, you may get a beep from the buzzer. This indicates the resistance is low because the multimeter is charging the electrolytic and it will beep until the electrolytic is charged to about 0.7v.
The same applies when probing across the power rails of a circuit. The circuit may contain electrolytics that will charge when probing and the buzzer will beep.
The Digital multimeter is actually detecting a voltage less than 0.7v across the probes and is created by a voltage-divider network inside the meter.
The voltage divider put 2v across the probes and when this drops to less than 0.5v, the buzzer is activated. That why it odes not buzz when testing a diode as the diode drops the voltage to 0.6v.

to Index

MEASURING A DIODE

A diode can be measured to see if it is "open" or "damaged" or "working" by placing the probes across the component.
If the diode is "open" (it will not work), the needle will NOT swing across the scale when touching the component with the probes in one direction or when the probes are reversed.
If the diode is "damaged" (does not work), the needle will swing fully across the scale when touching the component with the probes in one direction or when the probes are reversed.
If the diode is FUNCTIONAL, (works) the needle will swing about mid-way when touching the leads of the diode in one direction and it will not move when the probes are reversed.

WHY?

The positive of the battery inside an analogue multimeter comes out the black probe and that is why you will get a reading when the probers are "around the wrong way." The needle will swing a different amount for each resistance setting on the dial as the needle represents 0.6v drop and NOT an actual resistance.

There are two things you must remember.
1. When the diode is measured in one direction, the needle will will **not move at**

all. The technical term for this is the diode is **reverse biased**. It will not allow any current to flow. Thus the needle will not move.
When the diode is connected around the other way, the needle will swing to the right (move up scale) to about 80% of the scale. This position represents the voltage drop across the junction of the diode and is NOT a resistance value. If you change the resistance range, the needle will move to a slightly different position due to the resistances inside the meter. The technical term for this is the diode is **forward biased**. This indicates the diode is not faulty.
The needle will swing to a slightly different position for a "normal diode" compared to a Schottky diode. This is due to the different junction voltage drops.
However we are only testing the diode at very low voltage and it may break-down when fitted to a circuit due to a higher voltage being present or due to a high current flowing.

2. The leads of an **Analogue Multimeter** have the positive of the battery connected to the black probe and the readings of a "good diode" are shown in the following two diagrams:

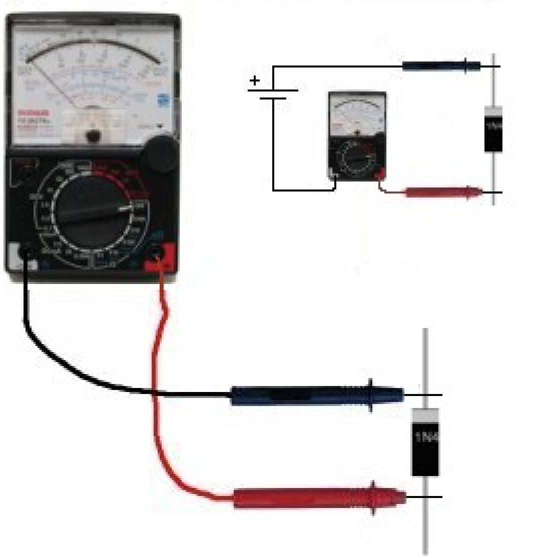

The diode is REVERSE BIASED in the diagram above and diodes not conduct.

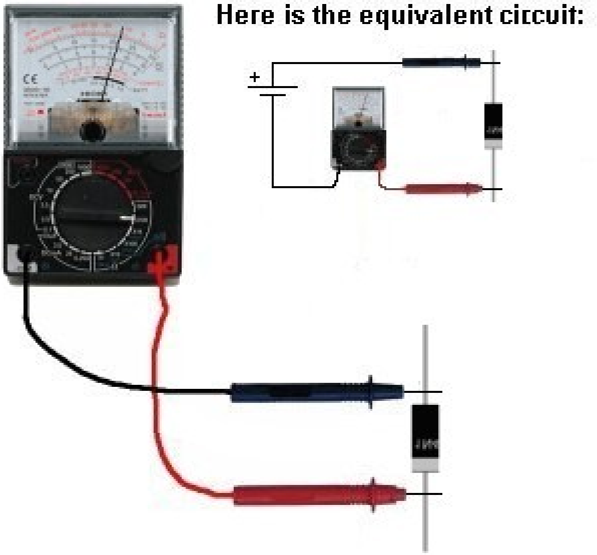

The diode is FORWARD BIASED in the diagram above and it conducts

TESTING A DIODE ON A DIGITAL METER

A Digital multimeter will measure the voltage-drop across the diode when the probes are connected in one direction (approx 0.640 on the scale) and a high reading (1) in the other direction. You need to select the "DIODE" setting on the dial as the other settings will produce a meaningless reading.

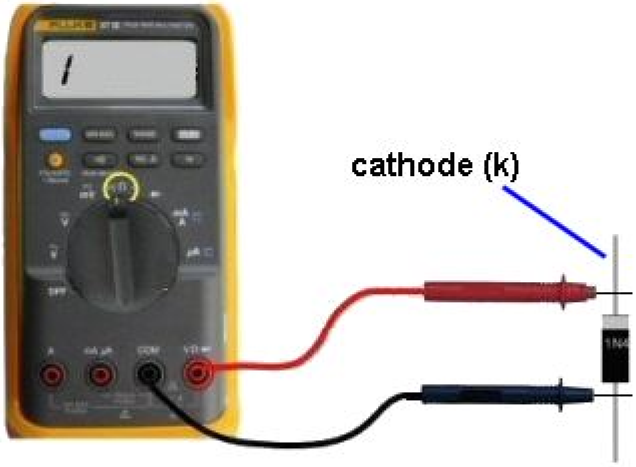

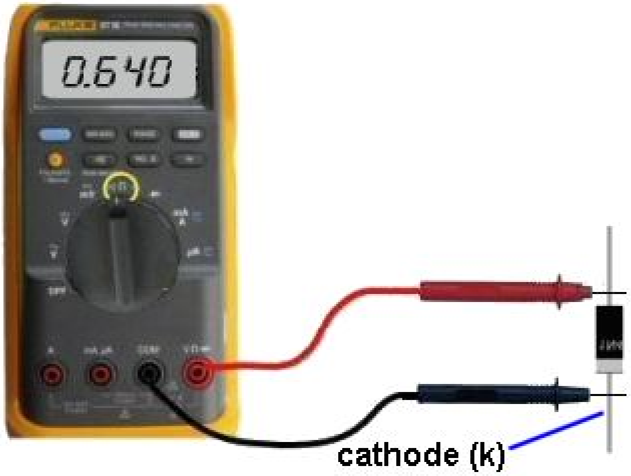

Some DIGITAL MULTIMETERS will show mV drop across the diode when the setting on the meter is "diode" or the "x1" or "x10" resistance range.

to Index

TESTING A LED

Some multimeters will test LEDs.
It depends on the voltage of the battery inside the case of the multimeter.
Many analogue multimeters have a single 1.5v cell and these cannot test LEDs.
Analogue Multimeters with 3v (for the resistance ranges) can test some LEDs.
White LEDs need about 3.6v and they may not illuminate on 3v.

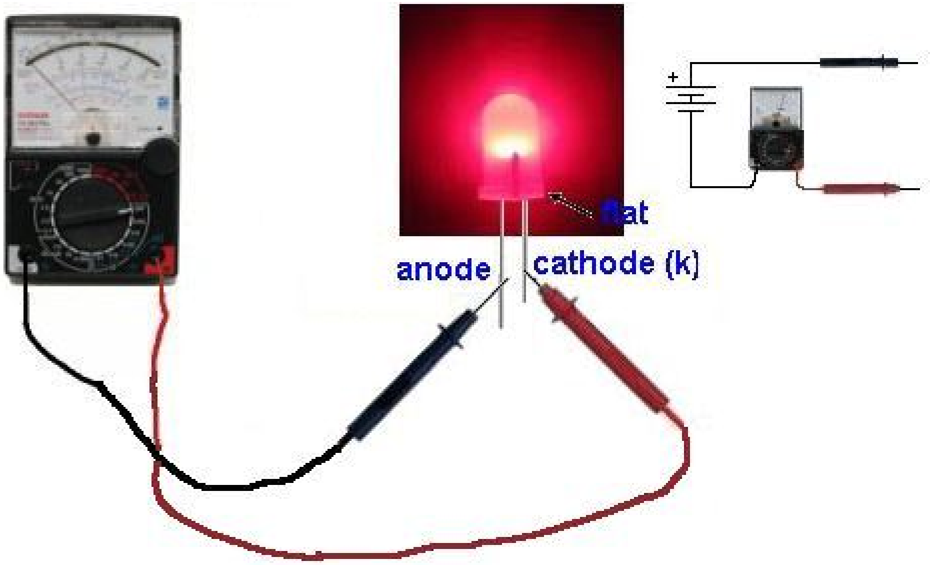

The negative lead of an ANALOGUE meter is POSITIVE!
The multimeter must have 3v (2 cells)

Digital multimeters have a 9v battery and they will illuminate all colour LEDs when the leads are placed as shown in the diagram:

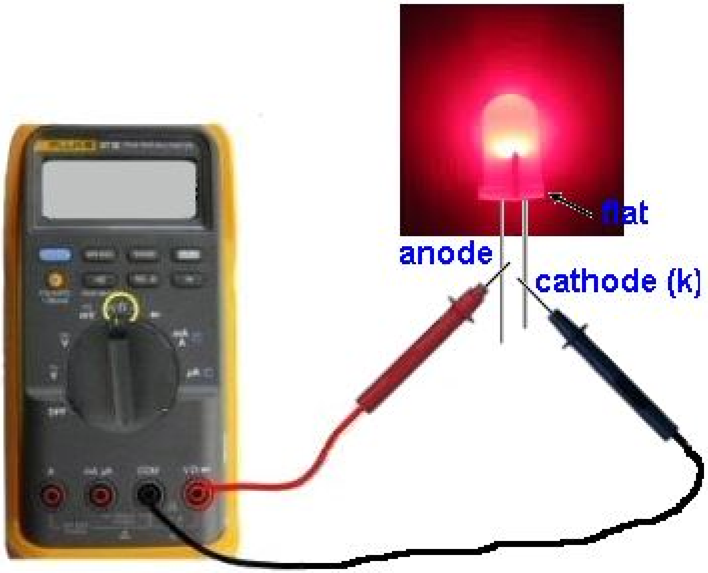

A Digital meter will illuminate all LEDs and the black probe touches the cathode.

to Index

TESTING A TRANSISTOR WITH A DIGITAL METER

Testing a transistor with a **Digital Meter** must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.
The "DIODE" setting must be used for diodes and transistors. It should also be called a "TRANSISTOR" setting.

TESTING A TRANSISTOR WITH AN ANALOGUE METER

The first thing you may want to do is test an unknown transistor for COLLECTOR, BASE AND EMITTER. You also want to perform a test to find out if it is NPN or PNP. That's what this test will provide.
You need a cheap multimeter called an ANALOGUE METER - a multimeter with a

scale and pointer (needle).
It will measure resistance values (normally used to test resistors) - (you can also test other components) and Voltage and Current. We use the resistance settings. It may have ranges such as "x10" "x100" "x1k" "x10"
Look at the resistance scale on the meter. It will be the top scale.
The scale starts at zero on the right and the high values are on the left. This is opposite to all the other scales.
When the two probes are touched together, the needle swings FULL SCALE and reads "ZERO." Adjust the pot on the side of the meter to make the pointer read exactly zero.

How to read: "x10" "x100" "x1k" "x10"
Up-scale from the zero mark is "1"
When the needle swings to this position on the "x10" setting, the value is 10 ohms.
When the needle swings to "1" on the "x100" setting, the value is 100 ohms.
When the needle swings to "1" on the "x1k" setting, the value is 1,000 ohms = 1k.
When the needle swings to "1" on the "x10k" setting, the value is 10,000 ohms = 10k.
Use this to work out all the other values on the scale.
Resistance values get very close-together (and very inaccurate) at the high end of the scale. [This is just a point to note and does not affect testing a transistor.]

Step 1 - FINDING THE BASE and determining NPN or PNP
Get an unknown transistor and test it with a multimeter set to "x10"
Try the 6 combinations and when you have the black probe on a pin and the red probe touches the other pins and the meter swings nearly full scale, you have an NPN transistor. The black probe is BASE
If the red probe touches a pin and the black probe produces a swing on the other two pins, you have a PNP transistor. The red probe is BASE
If the needle swings FULL SCALE or if it swings for more than 2 readings, the transistor is **FAULTY**.

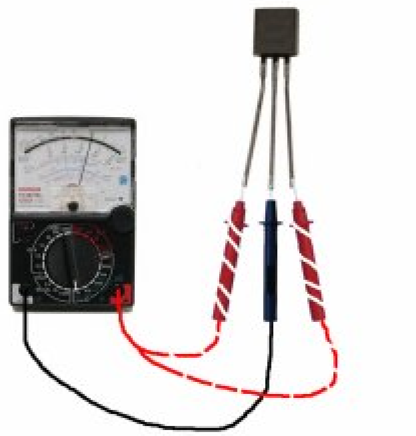

This is an NPN transistor
The black probe is the BASE

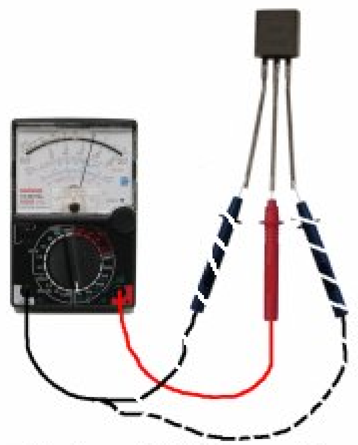

This is a PNP transistor
The red probe is the BASE

Step 2 - FINDING THE COLLECTOR and EMITTER
Set the meter to "x10k."
For an NPN transistor, place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

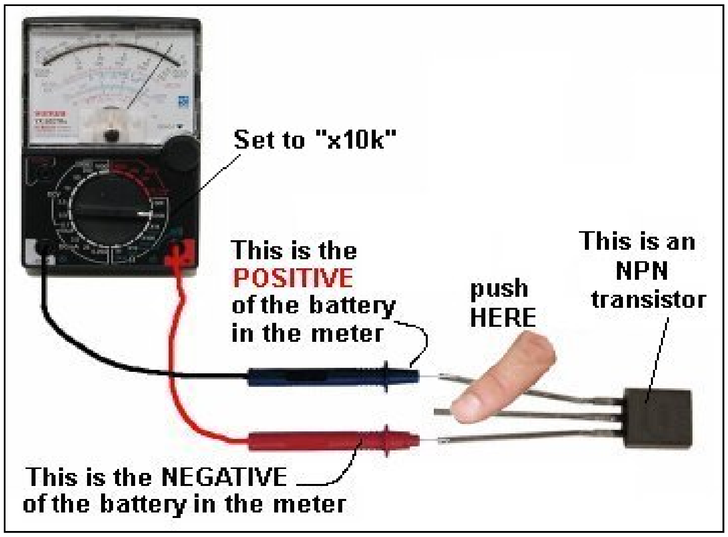
Set to "x10k"
This is the POSITIVE of the battery in the meter
push HERE
This is an NPN transistor
This is the NEGATIVE of the battery in the meter

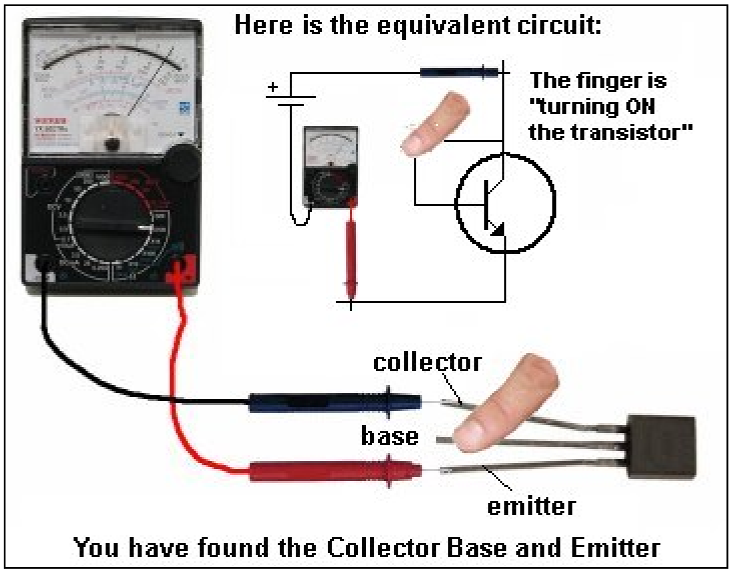
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
collector
base
emitter
You have found the Collector Base and Emitter

For a PNP transistor, set the meter to "x10k" place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

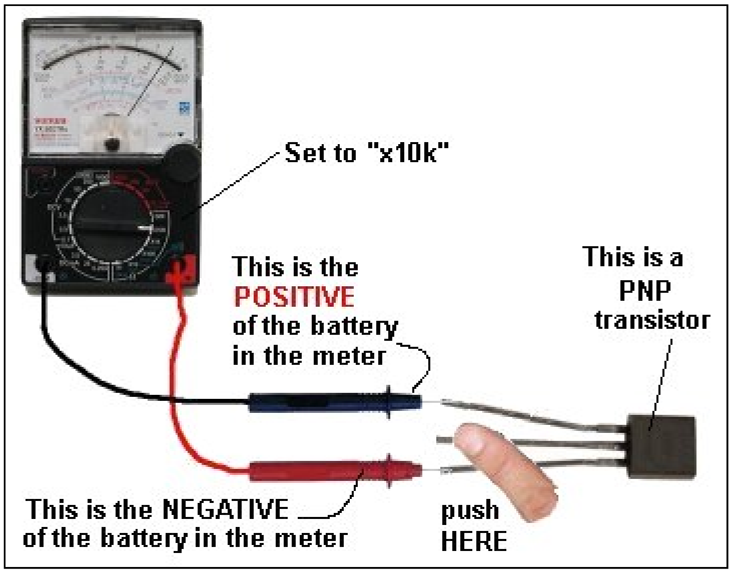
Set to "x10k"
This is the POSITIVE of the battery in the meter
This is a PNP transistor
This is the NEGATIVE of the battery in the meter
push HERE

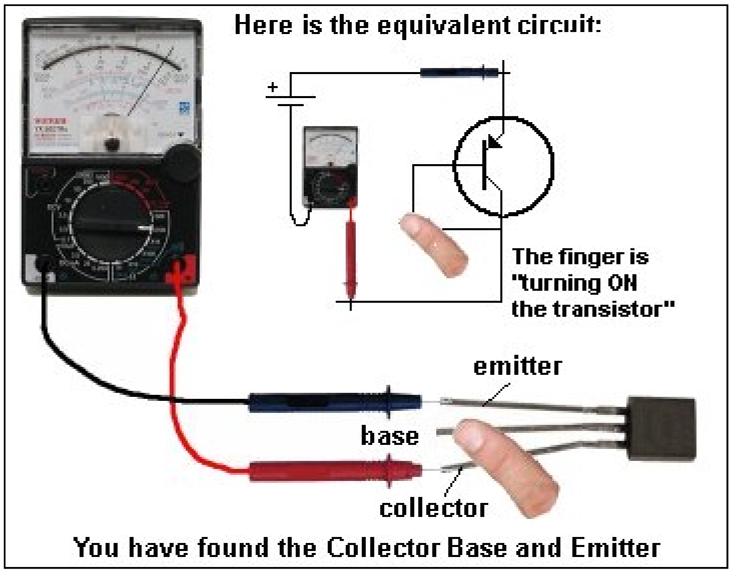
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
emitter
base
collector
You have found the Collector Base and Emitter

[to Index](#)

For more details on testing components with a multimeter, see:
[Testing Electronic Components](#).

[to Index](#)

MULTIMETER TEST

1. What is the reading on the multimeter:

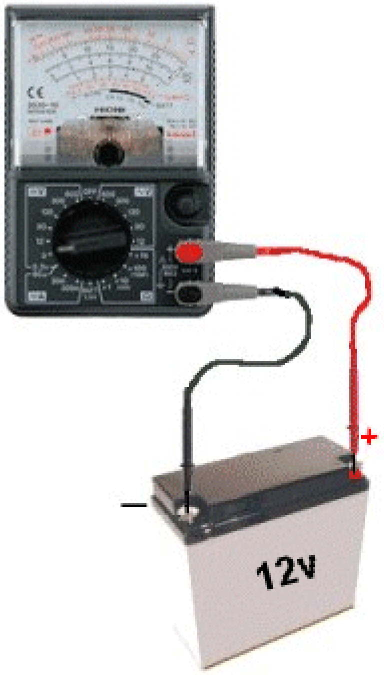

Answer

[to Index](#)

2. What is the reading on the multimeter:

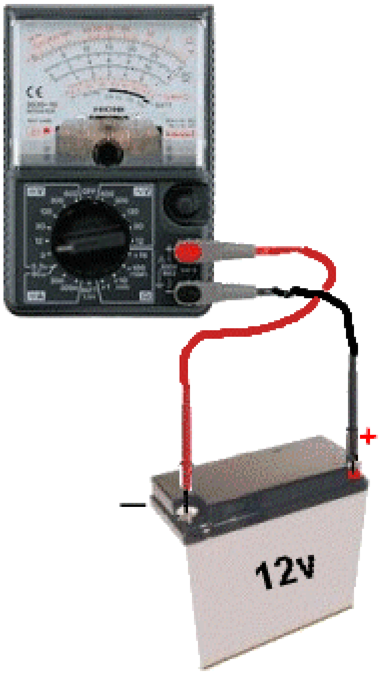

Answer

to Index

3. What is the reading on the multimeter:

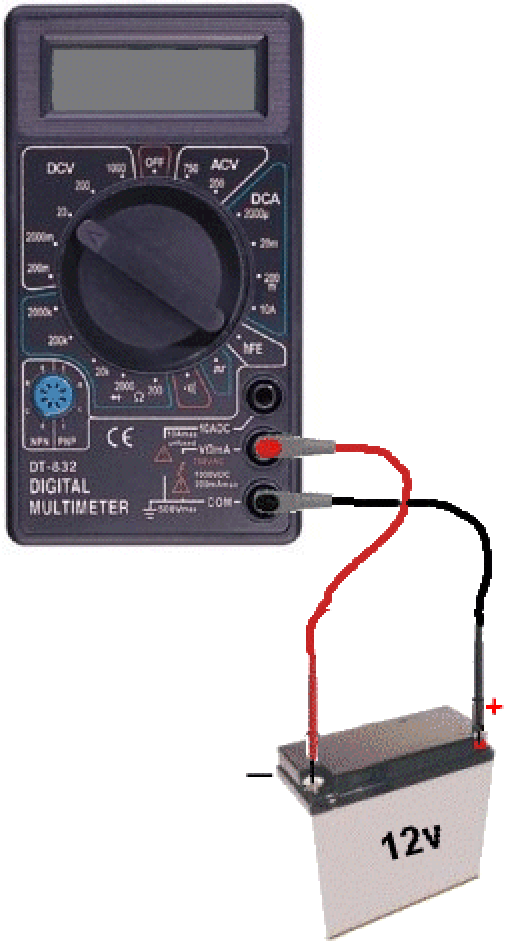
DCV
OFF
ACV
DCA
hFE
10ADC
VΩmA
COM
DT-832
DIGITAL
MULTIMETER
+
−
12v

Answer

to Index

4. What is the voltage between the following points:

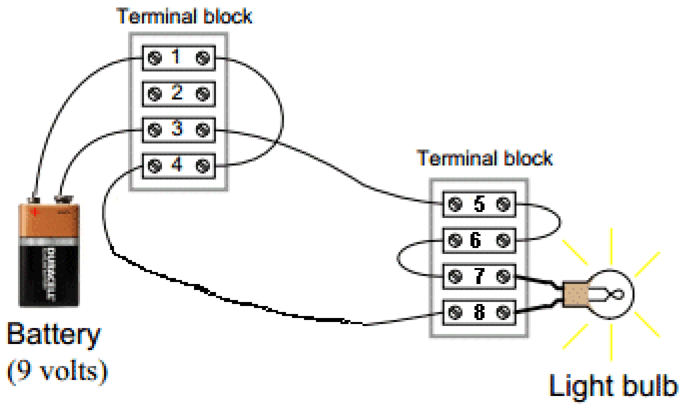
Terminal block
1
2
3
4
Terminal block
5
6
7
8
Battery
(9 volts)
Light bulb

Between 1 & 3	Between 5 & 7	Between 3 & 8	Between 1 & 7
Answer	**Answer**	**Answer**	**Answer**

to Index

5. What is the voltage on each multimeter:

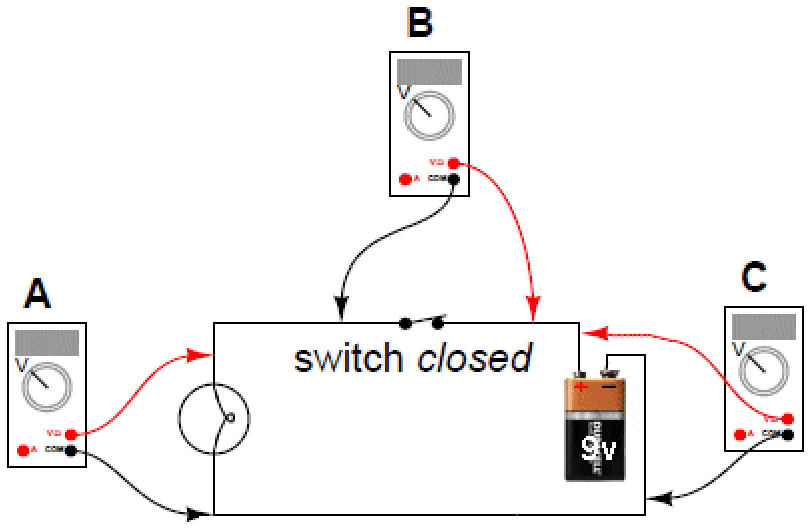
B
A
C
switch closed
9v

Meter A	**Meter B**	**Meter C**
Answer	**Answer**	**Answer**

to Index

6. What is the voltage on each multimeter:

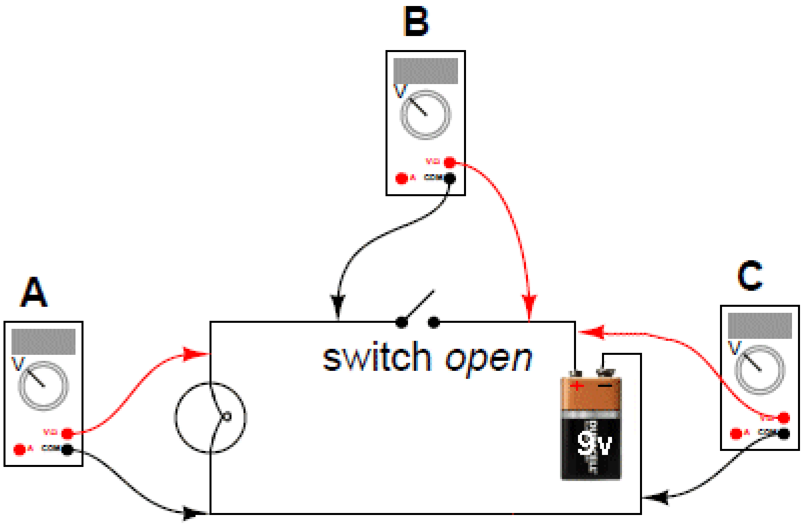
B
A
C
switch open
V
9v

Meter A
Answer

Meter B
Answer

Meter C
Answer

to Index

7. What is the reading on the multimeter:

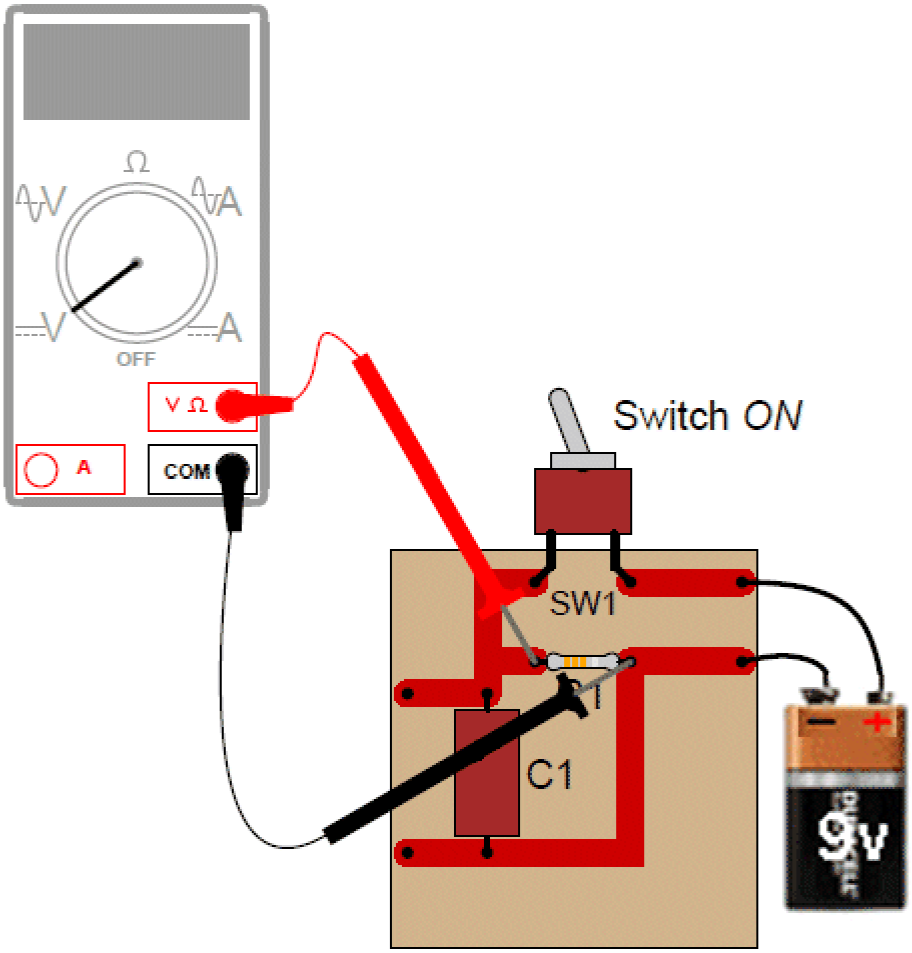
Ω
V
A
V
A
OFF
V Ω
A
COM
Switch ON
SW1
C1
9v

Answer

to Index

8. The multimeter has three resistance ranges: **x1 ohms** **x100 ohms** **x10k ohms**. How far across the scale will the needle move on each range for a 4k7 resistor:

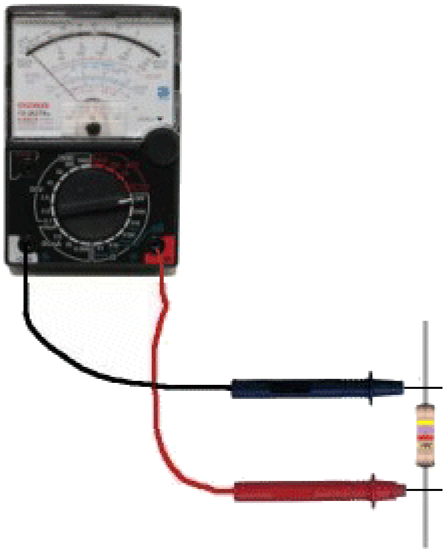

x1	**x100**	**x10k**
Answer	**Answer**	**Answer**

to Index

9. The resistance ranges on an ANALOGUE multimeter use the battery inside the case to move the pointer. If the multimeter is left on "**ohms range**" with the probes apart, will the battery go flat?

Answer

to Index

to Index

PCB CAD, Fabrication & Assembly
Pad2Pad
LEARN MORE »

1 Download PDF Manual - Free
Download Free Manuals With AtoZManuals free.atozmanuals.com/PDF/Manual

2 Splunk 2018 Predictions
Learn from Splunk experts about the tech & trends poised to transform business in 2018 splunk.com

Download PDF Manual - Free
Download Free Manuals With AtoZManuals free.atozmanuals.com/PDF/Manual

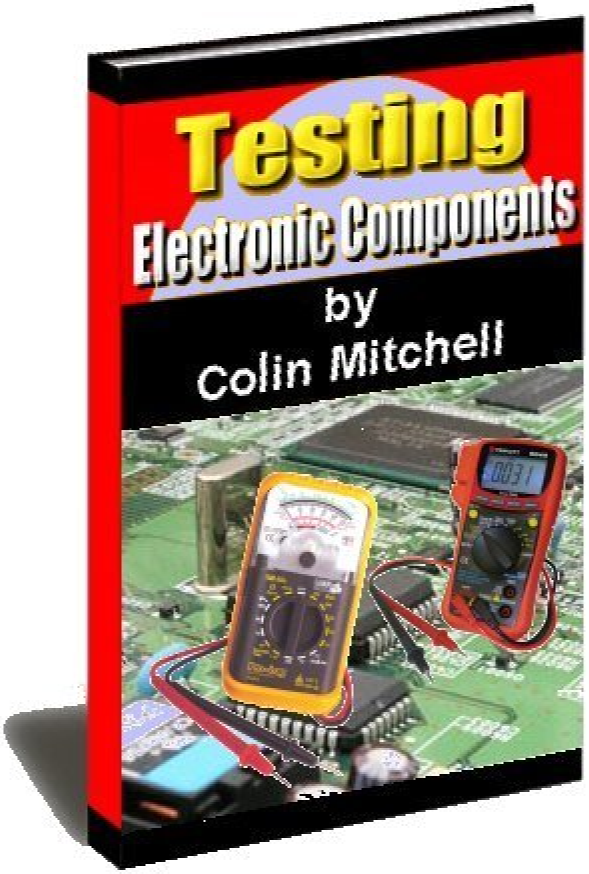

**This eBook shows you how to TEST COMPONENTS.
To do this you need "TEST GEAR." The best item of Test Gear is a MULTIMETER. It can test almost 90% of all components. And that's what we will do in this eBook:**

CONTENTS

Analogue Multimeter
Audio Stages
Batteries - testing
Burnt Resistor
Buying A Multimeter
Capacitors
Capacitors - decoupling caps
Capacitors in Parallel
Capacitor Substitution Box
Cells - batteries
Circuit Symbols
Co-Ax Cables
Colour Code (Resistor)
Coils
Continuity
Creating any value of R
Multimeters
Non-polar Capacitor (electrolytic)
Open Circuit
"Open" Resistor - damaged
Opto-couplers
Parallel - resistors
Parallel and Series Capacitors
Piezo Diaphragms
Piezo Buzzers
Potentiometers
Pots - testing
Power Diodes
Power Supply -see Basic Electronics 1A
Reed Switch
Relays
Remote Controls

Current - measuring
Current Sensing Resistors
Current Transformer
Damper Diodes
Darlington Transistors
Digital Chips
Digital Multimeter
Digital Transistors
Diodes replacing a diode
Discharge a Capacitor
Earth Leakage Detectors
Electrolytics
FETs
Fingers - to feel temperature
Focus pots
Fuses and 100mA fuses
Germanium Diodes
Heatsinks
Impedance - of a stage
IC's - also called Digital Chips
IC's - Analogue Chips
"In-Circuit" testing
Inductors
Inductors - measuring
Integrated Circuits
Isolation Transformer
Latching Reed Switch
LEDs
Logic Probe MkIIB
Logic Probe - Simple
Logic Probe - using CD4001
Logic Probe - using CD4011
Making your own components
Measuring Resistance
Measuring Voltage
Mica Washers and Insulators
Motor - testing
MOSFETs
Resistor Colour Code
Resistor Networks
Resistors - series
Resistor Substitution Box
Ripple Factor
Schottky Diodes
SCRs
Short Circuit
Signal Injector
Silicon diodes
Soldering
Spark Gaps
Speaker
Substitution Box
Super Probe MkII
Surface Mount - Packs
Surface-Mount Resistors
Surface-Mount Resistor Markings
Switches
Symbols
Test Equipment
Testing A Circuit
Testing A Resistor
Testing Components "In-Circuit"
Transformers - current transformer
Transformer Ratings 28v-0-28v
Transistor Outlines
Transistors
Transistors Digital
Triacs
Unknown resistors - testing
Using A Multimeter
Voltage Divider
Voltage Regulators
Voltages on a circuit
Wire Wound Resistor
Yokes
Zener Diodes
4-Band Resistors
5-Band Resistors
10-Turn Pots

Use your "brain, knowledge and your fingers."

Before we start, fixing anything is a combination of skill, luck and good diagnosis.
Sometimes you can fix something by letting it run until it finally fails.
Some things start to work as soon as you touch them.
Some things can never be fixed.
But some things can be fixed by feeling the temperature rise and deciding if it is getting too hot.
Sometimes you can smell something getting too hot.
Sometimes you can see SMOKE.
All these things make you a very clever technician and about 50% of faults will be fixed by looking for dry joints, burnt parts, overheating and carefully inspecting an item before you disturb it.

By simply touching different items you can quite often feel a hot item and home-in on the fault - at a saving of hours of work.
Servicing is not "A bull at a gate" approach.
You may be able to service something by turning it on and leaving it for hours - and start thinking.
It may take you a day to come up with the answer.
Believe me. That's how it worked for me - while fixing over 35,000 TV's.

TEST EQUIPMENT

Everyone thinks TEST EQUIPMENT will "solve the problem."
This is a big big MISTAKE.
Test equipment can help solve a problem and it can "lead to frustration," "give an incorrect answer," "mess you up," and make things worse.
You have to be very careful with test equipment and especially EXPENSIVE equipment because it is very sensitive and can detect pulses and glitches and voltages that are not affecting the operation of the circuit.
You will learn a lot of tricks when reading through this article, but let me say two things.
There are lots of faults and components that you cannot test with "test equipment" because they are either intermittent or the equipment does not load the device to the same extent as the circuit.
And secondly you need both an ANALOGUE multimeter and a DIGITAL meter to cover all the situations.
And if you are working on a car, you only need a $5.00 analogue meter because it will be dropped or fall into a crack, and you will only lose $5.00
You will learn that a digital meter will pick up spikes and signals on a line and show an incorrect reading.
That's why you need to back-up your readings with an analogue meter.
When you charge a battery it gets a "floating voltage" and this will be higher than the actual voltage, when the battery is fitted to a project. An analogue meter will draw a slight current and remove the "floating voltage."
Component testers can also give you a false reading, either because the component is out of range of the tester or intermittent and you need to be aware of this.
Oscilloscopes can also display waveforms that are parts of glitches or noise from other chips and these do not affect the operation of the part of the circuit you are investigating.
Sometimes you cannot pickup a pulse because it is not regular and the trigger on the oscilloscope does not show it on the screen. You may think it is missing.
It all depends on the "speed of the oscilloscope" - it's maximum frequency of operation.
Lastly- Power Supplies. You cannot test globes and motors on a power supply because the starting current can be 5 times more than the operating current. The power supply may not be able to deliver this high current and thus you will think the motor or globe is faulty.

MULTIMETERS

There are two types:

DIGITAL and ANALOGUE

A **Digital Multimeter** has a set of digits on the display and an Analogue Multimeter has a scale with a pointer (or needle).
You really need both types to cover the number of tests needed for designing and repair-work. We will discuss how they work, how to use them and some of the differences between them.

024
UNI-T
6.43
OMEGAETTE

DIGITAL AND ANALOGUE MULTIMETERS

BUYING A MULTIMETER

There are many different types on the market.
The cost is determined by the number of ranges and also the extra features such as diode tester, buzzer (continuity), transistor tester, high DC current and others.
Since most multimeters are reliable and accurate, buy one with the greatest number of ranges at the lowest cost.
This article explains the difference between a cheap analogue meter, an expensive analogue meter and a digital meter. You will then be able to work out which two meters you should buy.

Multimeters are sometimes called a "meter", a "VOM" (Volts-Ohms-Milliamps or Volt Ohm Meter) or "multi-tester" or even "a tester" - they are all the same.

USING A MULTIMETER

Analogue and digital multimeters have either a rotary selector switch or push buttons to select the appropriate function and range. Some Digital Multimeters (DMMs) are auto ranging; they automatically select the correct range of voltage, resistance, or current when doing a test. However you need to select the function.

Before making any measurement you need to know what you are checking. If you are measuring voltage, select the AC range (10v, 50v, 250v, or 1000v) or DC range (0.5v, 2.5v, 10v, 50v, 250v, or 1000v). If you are measuring resistance, select the Ohms range (x1, x10, x100, x1k, x10k). If you are measuring current, select the appropriate current range DCmA 0.5mA, 50mA, 500mA. Every multimeter is different however the photo below shows a low cost meter with the basic ranges.

MODEL YX-1000A
Ω
X10
X100
X1K
0.5
50
250
DCmA
10
50
250
1000
DCV
ACV
10
50
250
1000
−
+

The most important point to remember is this:
You must select a voltage or current range that is bigger or HIGHER than the maximum expected value, so the needle does not swing across the scale and hit the "end stop."
If you are using a DMM (Digital Multi Meter), the meter will indicate if the voltage or current is higher than the selected scale, by showing "OL" - this means "Overload." If you are measuring resistance such as 1M on the x10 range the "OL" means "Open Loop" and you will need to change the range. Some meters show "1' on the display when the measurement is higher than the display will indicate and some flash a set of digits to show over-voltage or over-current. A "-1" indicates the leads should be reversed for a "positive reading."
If it is an AUTO RANGING meter, it will automatically produce a reading, otherwise the selector switch must be changed to another range.

HV
000
DCV
OFF
ACV
DCA
hFE
NPN PNP
CE
DT-832
DIGITAL
MULTIMETER
10ADC
VΩmA
COM

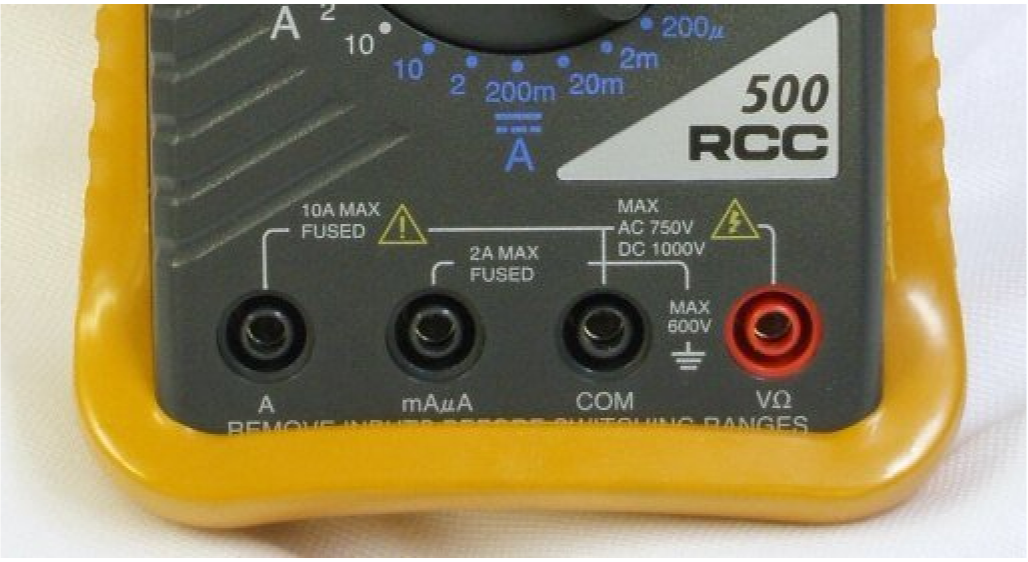
A
10
10
2
200m
20m
2m
200μ
A
500
RCC
10A MAX
FUSED
2A MAX
FUSED
MAX
AC 750V
DC 1000V
MAX
600V
A
mAμA
COM
VΩ

**The Common (negative) lead ALWAYS fits into
the "COM" socket. The red lead fits into the
red socket for Voltage and Resistance.
Place the red lead (red banana plug)
into "A" (for HIGH CURRENT "Amps")
or mA,uA for LOW CURRENT.**

The black "test lead" plugs into the socket marked "-" "Common", or "Com," and the red "test lead" plugs into meter socket marked "+" or "V-W-mA." The third banana socket measures HIGH CURRENT and the positive (red lead) plugs into this. You DO NOT move the negative "-" lead at any time.

The following two photos show the test leads fitted to a digital meter. The probes and plugs have "guards" surrounding the probe tips and also the plugs so you can measure high voltages without getting near the voltage-source.

FLUKE
115 TRUE RMS MULTIMETER
600.0
HOLD
MIN MAX
RANGE
OFF
V
mV
Ω
A
A
COM
10 A FUSED
CAT III 600 V

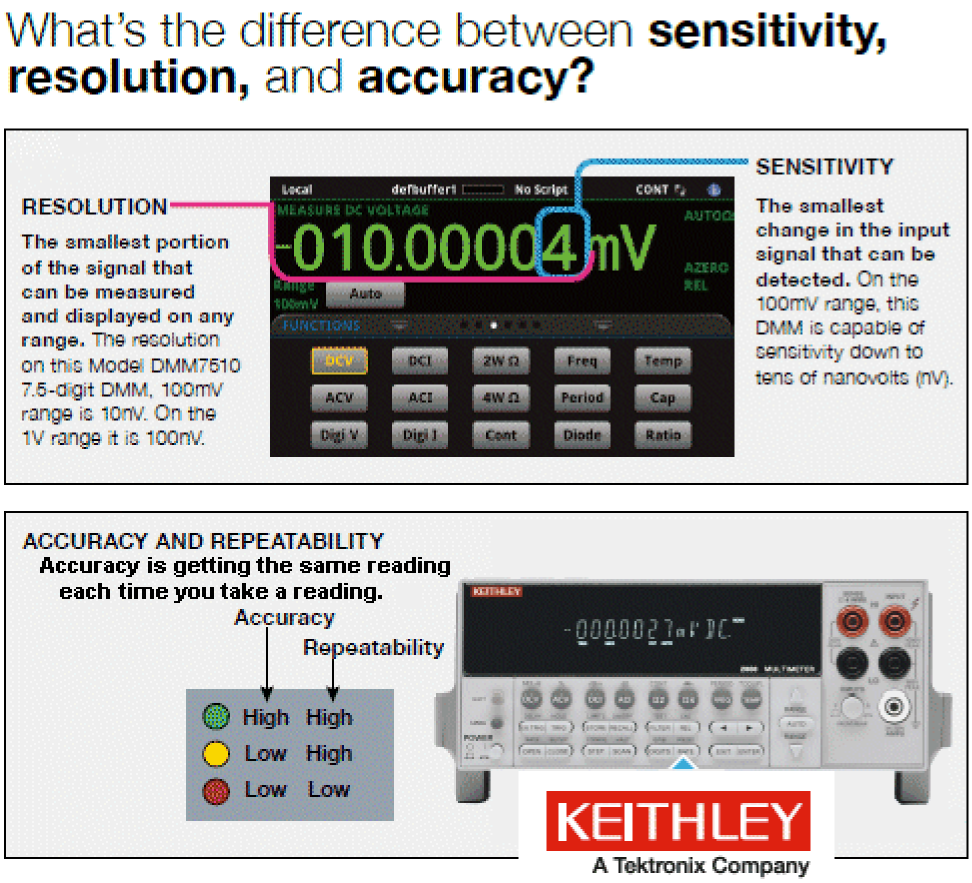

What's the difference between **sensitivity, resolution,** and **accuracy?**

RESOLUTION

The smallest portion of the signal that can be measured and displayed on any range. The resolution on this Model DMM7510 7.5-digit DMM, 100mV range is 10nV. On the 1V range it is 100nV.

SENSITIVITY

The smallest change in the input signal that can be detected. On the 100mV range, this DMM is capable of sensitivity down to tens of nanovolts (nV).

ACCURACY AND REPEATABILITY

Accuracy is getting the same reading each time you take a reading.

	Accuracy	Repeatability
🟢	High	High
🟡	Low	High
🔴	Low	Low

Local
defbuffer1
No Script
CONT
MEASURE DC VOLTAGE
-010.000004mV
AUTOZ
AZERO
REL
Range 100mV
Auto
FUNCTIONS
DCV
DCI
2W Ω
Freq
Temp
ACV
ACI
4W Ω
Period
Cap
Digi V
Digi I
Cont
Diode
Ratio
KEITHLEY
-0.000027mV DC
KEITHLEY
A Tektronix Company

The question above applies to both (every) type of multimeter and the type of meter you use depends on the accuracy you need. Sometimes you are looking for 1mV change on a 20v rail. Only a DMM will (or a CRO) will produce a result.

Analogue meters have an "Ohms Adjustment" to allow for the change in voltage of the battery inside the meter (as it gets old).

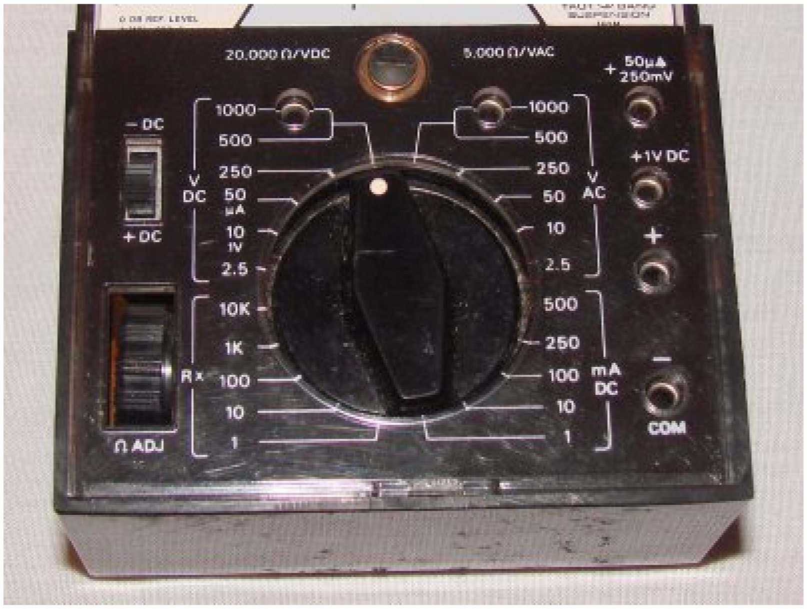

20,000 Ω/VDC
5,000 Ω/VAC
+ 50µA 250mV
- DC
+ DC
1000
500
250
V DC
50 µA
10 1V
2.5
10K
1K
Rx 100
10
1
Ω ADJ
1000
500
250
50
V AC
10
2.5
500
250
100 mA DC
10
1
+1V DC
+
-
COM

"Ohms Adjust" is also called "ZERO SET"

The sensitivity of this meter is 20,000ohms/volt on the DC ranges and 5k/v on the AC ranges

Before taking a resistance reading (each time on any of the Ohms scales) you need to "ZERO SET" the scale, by touching the two probes together and adjust the pot until the needle reads "0" (swings FULL SCALE). If the pointer does not reach full scale, the batteries need replacing. Digital multimeters do not need "zero adjustment."

FIXING A MULTIMETER

A multimeter can get "broken" "damaged" and go "faulty."
I don't know why, but eventually they stop working.
It can be something simple like a flat battery, corroded battery contacts, broken switch or something complex, like the circuitry failing.
Multimeters are so cheap, you can buy a new one for less than $10.00
These meters can have a 10 amp range, transistor tester and measure up to 2 meg ohms.
That's why I suggest buying a $10.00 meter. They are just as good as a $60.00 meter and the cheapest meters last the longest.
Dropping an analogue meter can cause the hair spring to loop over one of the supports and the needle will not zero correctly. You will need to open the cover on the movement and lift the spring off the support with a needle.
A faulty meter can be used in a battery-charger circuit to measure the current or voltage if that scale is still reading-correctly.
Otherwise keep the leads and throw the meter out. It is too dangerous keeping a meter that shows an incorrect reading.

MEASURING FREQUENCY

VICHY
VC99
APO
DC
AUTO
-000.0
mV
MAX/MIN
Hz/DUTY
REL
HOLD
RANGE
DC/AC
Hz
hFE
V~
Vm
OFF
AUTO OFF POWER
TEMP
A
mA
COM
VΩHz
CAT II 1000V

Before we cover the normal uses for a multimeter, it is interesting to note that some **Digital Multimeters (DMM)** have features such as Capacitance, Frequency and measuring the gain of a transistor as well as a number of other features using probes such as a temperature probe. The VICHY VC99 meter above is an example and costs about $40.00.

Basic function	Range
DCV	600mV/6V/60V/600V/1000V
ACV	6V/60/600/1000V
DCA	600uA/6000uA/60mA/600mA/6A/20A
ACA	600uA/6000uA/60mA/600mA/6A/20A
Capacitance	40nF/400nF/4uF/40uF/400uF/2000uF
Resistance	600Ω/6kΩ/60kΩ/600kΩ/6MΩ/60MΩ
Frequency	100Hz/1kHz/10kHz/100kHz/1MHz/60MHz

Temperature	-40°C~1000°C
	0°F~1832°F

MEASURING VOLTAGE

Most of the readings you will take with a multimeter will be VOLTAGE readings. Before taking a reading, you should select the highest range and if the needle does not move up scale (to the right), you can select another range.
Always switch to the highest range before probing a circuit and keep your fingers away from the component being tested.
If the meter is Digital, select the highest range or use the auto-ranging feature, by selecting "V." The meter will automatically produce a result, even if the voltage is AC or DC.
If the meter is not auto-ranging, you will have to select V⎓ if the voltage is from a DC source or V~ if the voltage is from an AC source. DC means Direct Current and the voltage is coming from a battery or supply where the voltage is steady and not changing and AC means Alternating Current where the voltage is coming from a voltage that is rising and falling.
You can measure the voltage at different points in a circuit by connecting the black probe to chassis. This is the 0v reference and is commonly called "Chassis" or "Earth" or "Ground" or "0v."
The red lead is called the "measuring lead" or "measuring probe" and it can measure voltages at any point in a circuit. Sometimes there are "test points" on a circuit and these are wires or loops designed to hold the tip of the red probe (or a red probe fitted with a mini clip or mini alligator clip).
You can also measure voltages ACROSS A COMPONENT. In other words, the reading is taken in PARALLEL with the component. It may be the voltage across a transistor, resistor, capacitor, diode or coil. In most cases this voltage will be less than the supply voltage.
If you are measuring the voltage in a circuit that has a HIGH IMPEDANCE, the reading will be inaccurate, up to 90% !!!, if you use a cheap analogue meter.

Here's a simple case.

The circuit below consists of two 1M resistors in series. The voltage at the mid point will be 5v when nothing is connected to the mid point. But if we use a cheap analogue multimeter set to 10v, the resistance of the meter will be about 100k, if the meter has a sensitivity of 10k/v and the reading will be incorrect.
Here how it works:
Every meter has a sensitivity. The sensitivity of the meter is the sensitivity of the movement and is the amount of current required to deflect the needle FULL SCALE. This current is very small, normally 1/10th of a milliamp and corresponds to a sensitivity of 10k/volt (or 1/30th mA, for a sensitivity of 30k/v).
If an analogue meter is set to 10v, the internal resistance of the meter will be 100k for a 10k/v movement.
If this multimeter is used to test the following circuit, the reading will be inaccurate. The reading should be 5v as show in diagram **A**.
But the analogue multimeter has an internal resistance of 100k and it creates a circuit shown in **C**.
The top 1M and 100k from the meter create a combined PARALLEL resistance of 90k. This forms a series circuit with the lower 1M and the meter will read less than 1v
If we measure the voltage across the lower 1M, the 100k meter will form a value of resistance with the lower 1M and it will read less than 1v
If the multimeter is 30k/v, the readings will be 2v. See how easy it is to get a totally inaccurate reading.

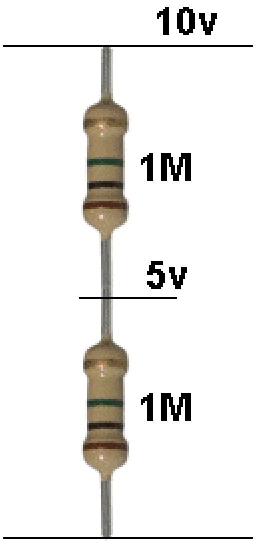

A

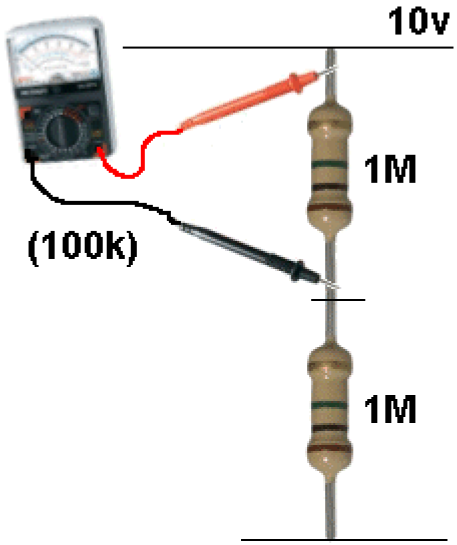

B

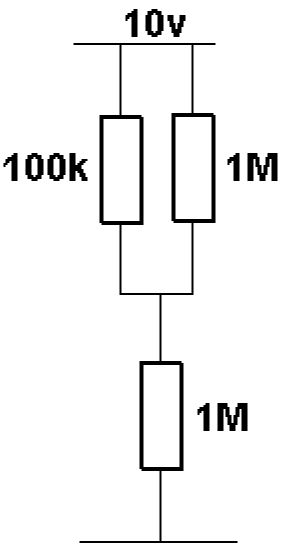

C

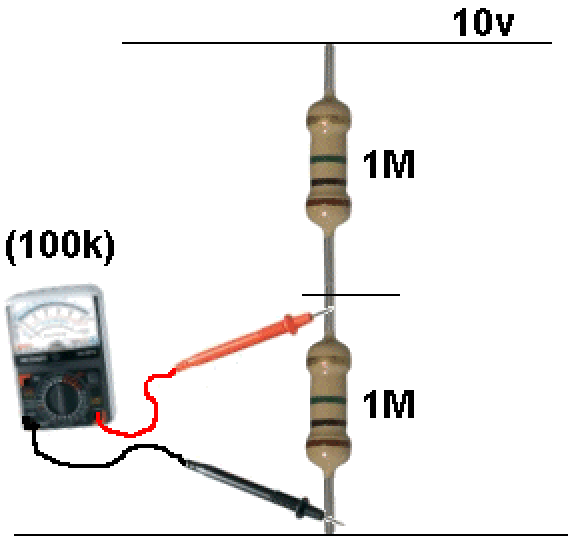

D

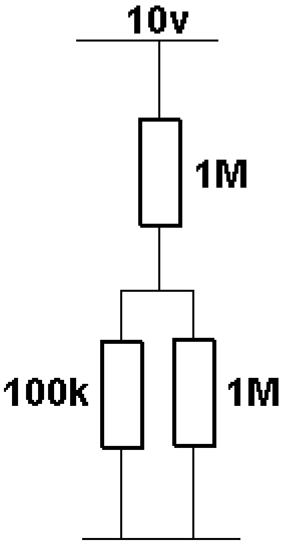

E

This introduces two new terms:
HIGH IMPEDANCE CIRCUIT and "RESISTORS in SERIES and PARALLEL."

If the reading is taken with a Digital Meter, it will be more accurate as a DMM does not take any current from the circuit (to activate the meter). In other words it has a very HIGH input impedance. Most Digital Multimeters have a fixed input resistance (impedance) of 10M - no matter what scale is selected. That's the reason for choosing a DMM for high impedance circuits. It also gives a reading that is accurate to about 1%.

MEASURING VOLTAGES IN A CIRCUIT

You can take many voltage-measurements in a circuit. You can measure "across" a component, or between any point in a circuit and either the positive rail or earth rail (0v rail). In the following circuit, the 5 most important voltage-measurements are shown. Voltage "A" is across the electret microphone. It should be between 20mV and 500mV. Voltage "B" should be about 0.6v. Voltage "C" should be about half-rail voltage. This allows the transistor to amplify both the positive and negative parts of the waveform. Voltage "D" should be about 1-3v. Voltage "E" should be the battery voltage of 12v.

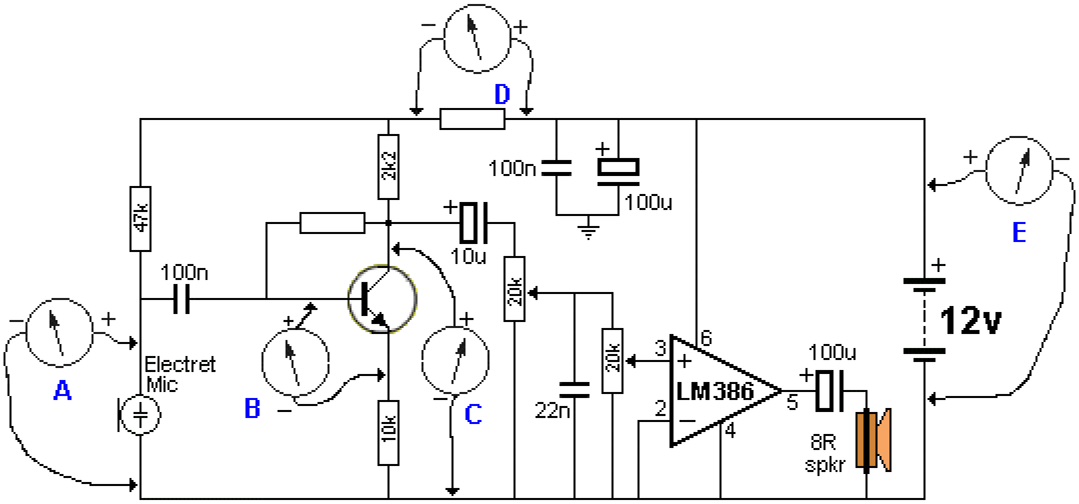

MEASURING VOLTAGES IN A CIRCUIT

MEASURING CURRENT

You will rarely need to take current measurements, however most multimeters have DC current ranges such as 0.5mA, 50mA, 500mA and 10Amp (via the extra banana socket) and some meters have AC current ranges. Measuring the current of a circuit will tell you a lot of things. If you know the normal current, a high or low current can let you know if the circuit is overloaded or not fully operational.

Current is always measured when the circuit is working (i.e: with power applied).
It is measured IN SERIES with the circuit or component under test.
The easiest way to measure current is to remove the fuse and take a reading across the fuse-holder. Or remove one lead of the battery or turn the project off, and measure across the switch.
If this is not possible, you will need to remove one end of a component and measure with the two probes in the "opening."
Resistors are the easiest things to desolder, but you may have to cut a track in some circuits. You have to get an "opening" so that a current reading can be taken.
The following diagrams show how to connect the probes to take a CURRENT reading.
Do not measure the current ACROSS a component as this will create a "short-circuit."
The component is designed to drop a certain voltage and when you place the probes across this component, you are effectively adding a "link" or "jumper" and the voltage at the left-side of the component will appear on the right-side. This voltage may be too high for the circuit being supplied and the result will be damage.

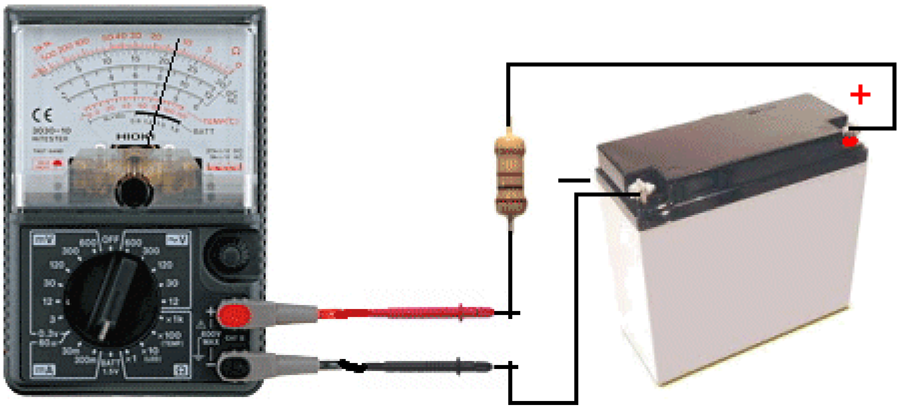

Measuring current through a resistor

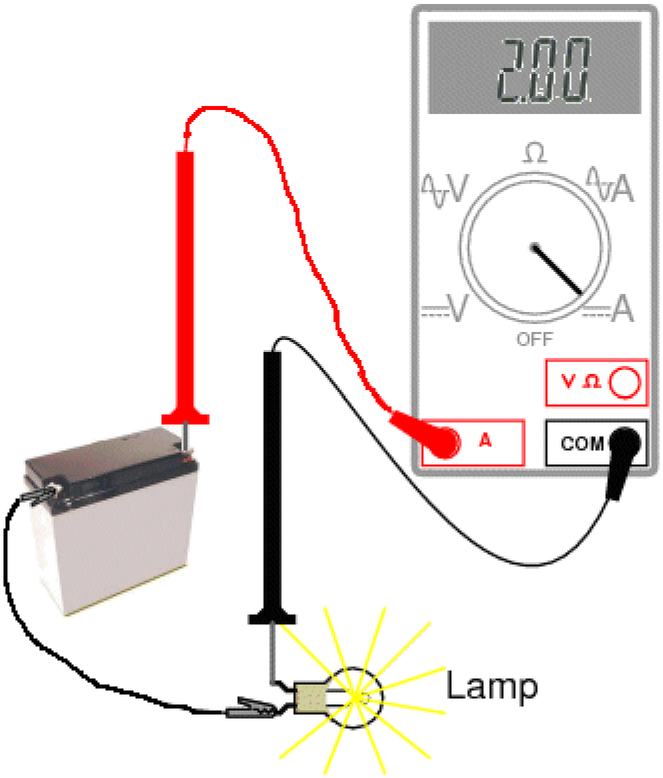

Measuring the current of a globe

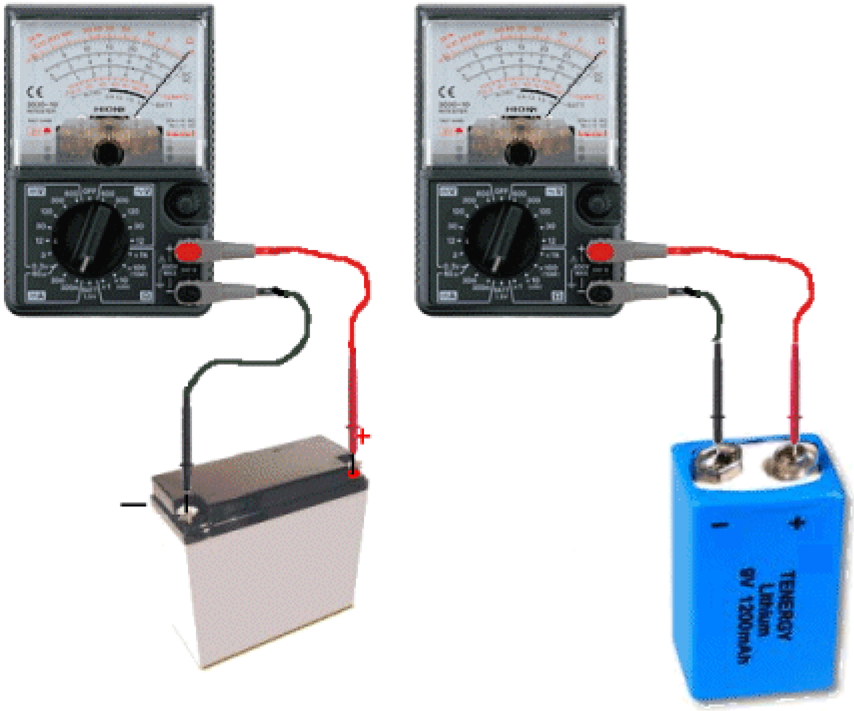

Do NOT measure the CURRENT of a battery
(by placing the meter directly across the terminals)
A battery will deliver a very HIGH current
and damage the meter

Do not measure the "current a battery will deliver" by placing the probes across the terminals. It will deliver a very high current and damage the meter instantly. There are special battery testing instruments for this purpose.
When measuring across an "opening" or "cut," place the red probe on the wire that supplies the voltage (and current) and the black probe on the other wire. This will produce a "POSITIVE" reading.

A positive reading is an UPSCALE READING and the pointer will move across the scale - to the right. A "NEGATIVE READING" will make the pointer hit the "STOP" at the left of the scale and you will not get a reading. If you are using a Digital Meter, a negative sign "-" will appear on the screen to indicate the probes are around the wrong way. No damage will be caused. It just indicates the probes are connected incorrectly.
If you want an accurate CURRENT MEASUREMENT, use a digital meter.

MEASURING RESISTANCE

Turn a circuit off before measuring resistance.
If any voltage is present, the value of resistance will be incorrect.
In most cases you cannot measure a component while it is in-circuit. This is because the meter is actually measuring a voltage across a component and calling it a "resistance." The voltage comes from the battery inside the meter. If any other voltage is present, the meter will produce a false reading.
If you are measuring the resistance of a component while still "in circuit," (with the power off) the reading will be lower than the true reading.

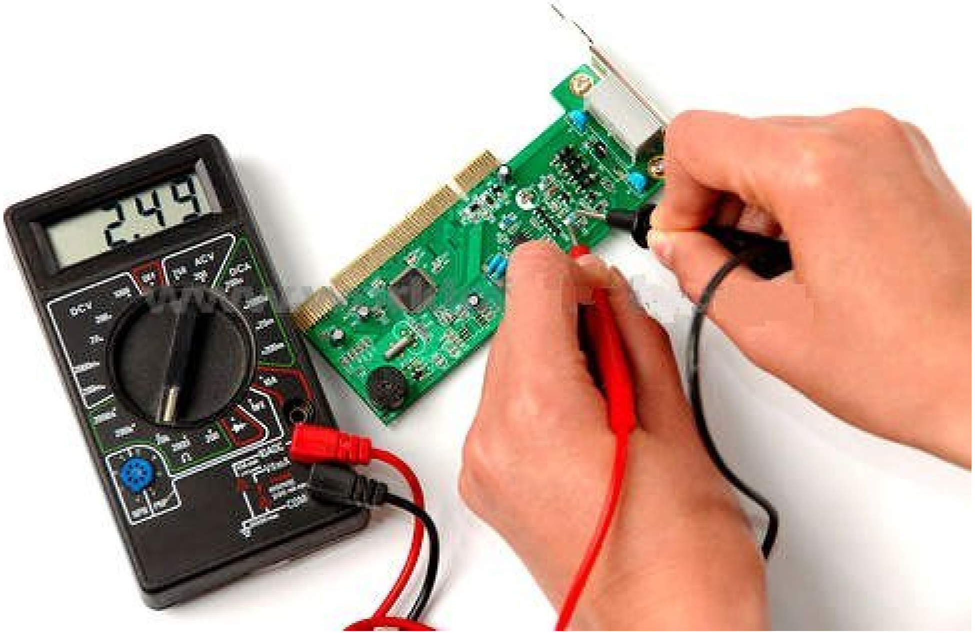

Measuring resistance

138
rolson
27259

Measuring resistance of a heater (via the leads)

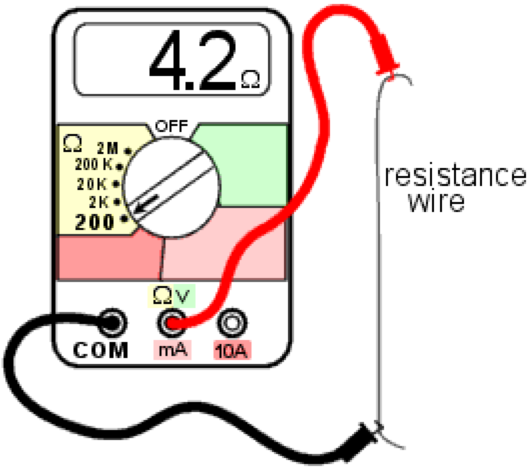
4.2Ω
OFF
Ω
2M
200 K
20K
2K
200
ΩV
COM
mA
10A
resistance wire

Measuring the resistance of a piece of resistance-wire

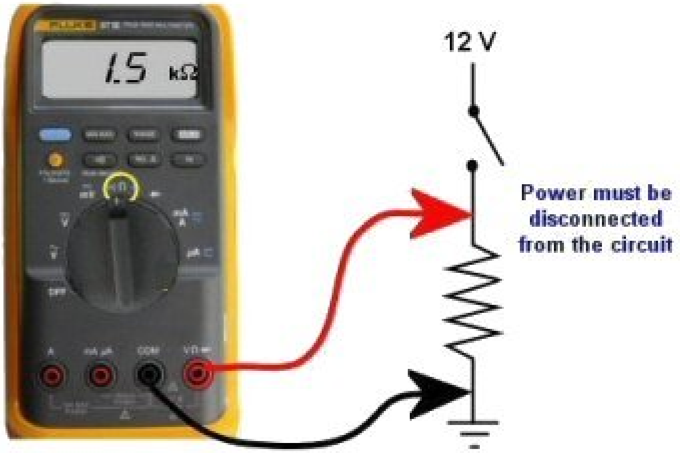

Power must be disconnected from the circuit

Measuring the resistance of a resistor

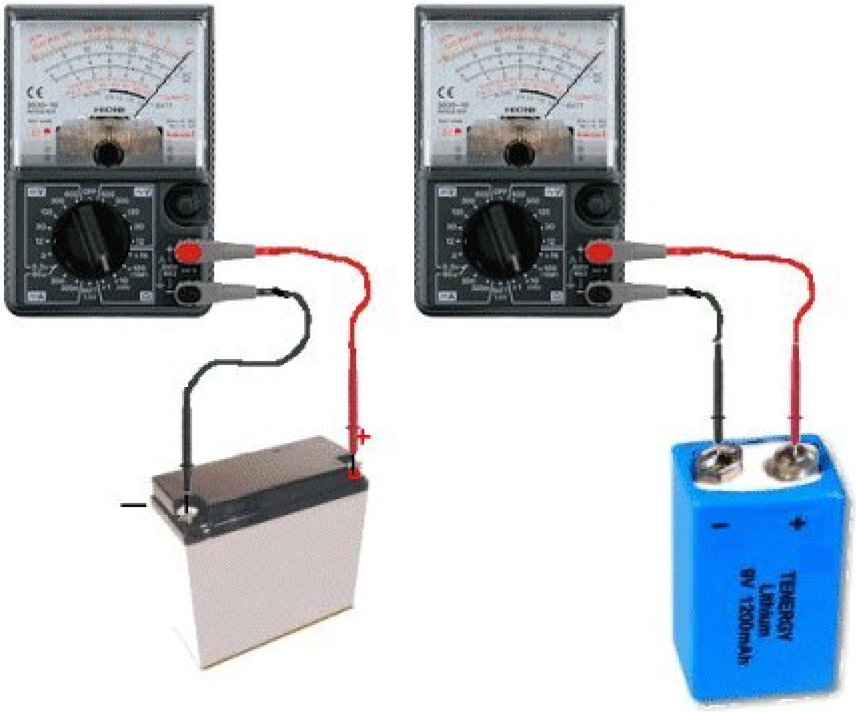
+
–
-
+
TENERGY
Lithium
9V 1200mAh

Do not measure the "Resistance of a Battery"

1. Do not measure the "resistance of a battery." The resistance of a battery (called the Internal impedance) is not measured as shown in the diagrams above. It is measured by creating a current-flow and measuring the voltage across the battery.

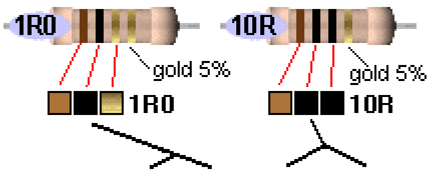

Placing a multimeter set to **resistance** (across a battery) will destroy the meter.
2. Do not try to measure the resistance of any voltage or any "supply."

Resistance is measured in OHMs.
The resistance of a 1cm x 1cm bar, one metre long is 1 ohm.
If the bar is thinner, the resistance is higher. If the bar is longer, the resistance is higher.
If the material of the bar is changed, the resistance is higher.
When carbon is mixed with other elements, its resistance increases and this knowledge is used to make RESISTORS.
Resistors have RESISTANCE and the main purpose of a resistor is to reduce the CURRENT FLOW.
It's a bit like standing on a hose. The flow reduces.
When current flow is reduced, the output voltage is also reduced and that why the water does not spray up so high. Resistors are simple devices but they produce many different effects in a circuit.
A resistor of nearly pure carbon may be 1 ohm, but when non-conducting "impurities" are added, the same-size resistor may be 100 ohms, 1,000 ohms or 1 million ohms.
Circuits use values of less than 1 ohm to more than 22 million ohms.

Resistors are identified on a circuit with numbers and letters to show the exact value of resistance - such as 1k 2k2 4M7
The letter Ω (omega - a Greek symbol) is used to identify the word "Ohm." but this symbol is not available on some word-processors, so the letter "R" is used. The letter "E" is also sometimes used and both mean "Ohms."
A one-ohm resistor is written "1R" or "1E." It can also be written "1R0" or "1E0."
A resistor of one-tenth of an ohm is written "0R1" or "0E1." The letter takes the place of the decimal point.
10 ohms = 10R
100 ohms = 100R
1,000 ohms = 1k (k= kilo = one thousand)
10,000 ohms = 10k
100,000 ohms = 100k
1,000,000 ohms = 1M (M = MEG = one million)

The size of a resistor has nothing to do with its resistance. The size determines the wattage of the resistor - how much heat it can dissipate without getting too hot.
Every resistor is identified by colour bands on the body, but when the resistor is a surface-mount device, numbers are used and sometimes letters.
You MUST learn the colour code for resistors and the following table shows all the colours for the most common resistors from 1/10th of an ohm to 22 Meg ohms for resistors with 5% and 10% tolerance.

If 3rd band is gold, Divide by 10
If 3rd band is silver, Divide by 100
(to get 0.22ohms etc)

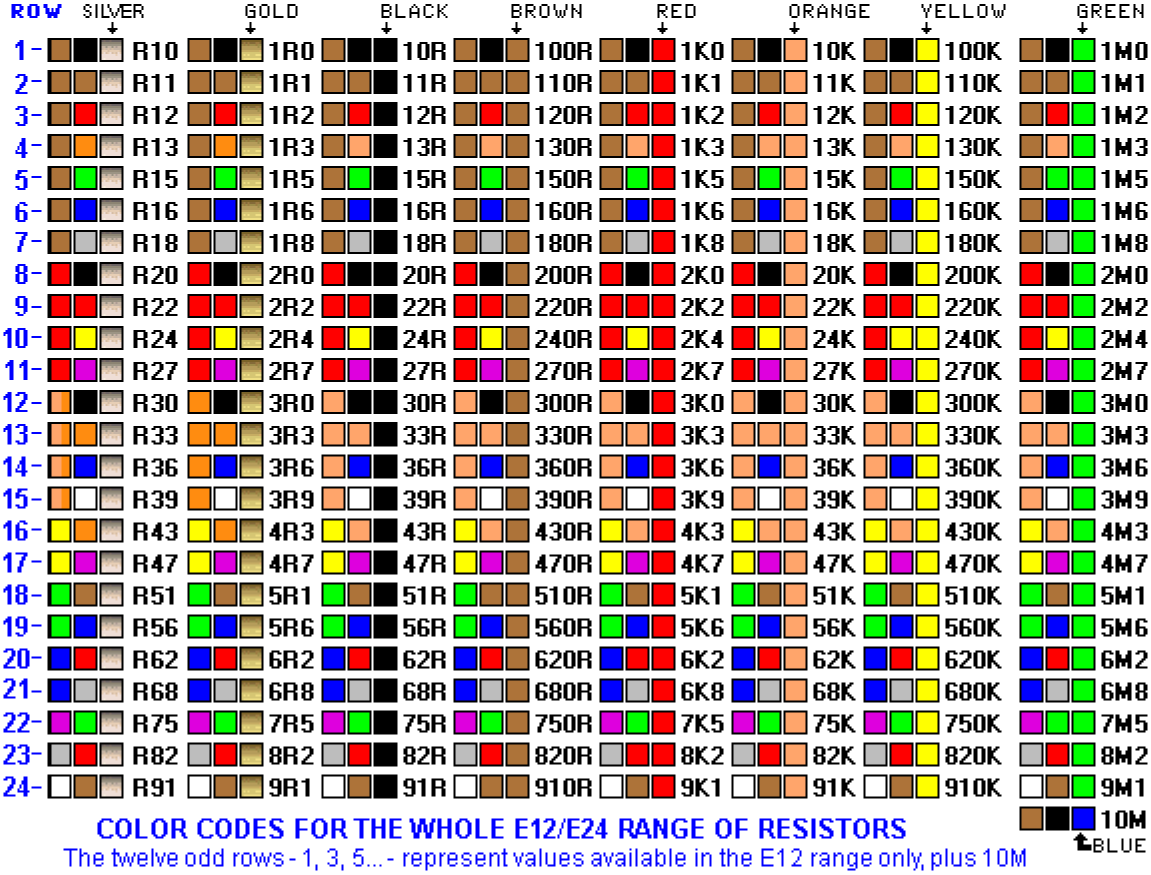

ROW	SILVER	GOLD	BLACK	BROWN	RED	ORANGE	YELLOW	GREEN
1-	R10	1R0	10R	100R	1K0	10K	100K	1M0
2-	R11	1R1	11R	110R	1K1	11K	110K	1M1
3-	R12	1R2	12R	120R	1K2	12K	120K	1M2
4-	R13	1R3	13R	130R	1K3	13K	130K	1M3
5-	R15	1R5	15R	150R	1K5	15K	150K	1M5
6-	R16	1R6	16R	160R	1K6	16K	160K	1M6
7-	R18	1R8	18R	180R	1K8	18K	180K	1M8
8-	R20	2R0	20R	200R	2K0	20K	200K	2M0
9-	R22	2R2	22R	220R	2K2	22K	220K	2M2
10-	R24	2R4	24R	240R	2K4	24K	240K	2M4
11-	R27	2R7	27R	270R	2K7	27K	270K	2M7
12-	R30	3R0	30R	300R	3K0	30K	300K	3M0
13-	R33	3R3	33R	330R	3K3	33K	330K	3M3
14-	R36	3R6	36R	360R	3K6	36K	360K	3M6
15-	R39	3R9	39R	390R	3K9	39K	390K	3M9
16-	R43	4R3	43R	430R	4K3	43K	430K	4M3
17-	R47	4R7	47R	470R	4K7	47K	470K	4M7
18-	R51	5R1	51R	510R	5K1	51K	510K	5M1
19-	R56	5R6	56R	560R	5K6	56K	560K	5M6
20-	R62	6R2	62R	620R	6K2	62K	620K	6M2
21-	R68	6R8	68R	680R	6K8	68K	680K	6M8
22-	R75	7R5	75R	750R	7K5	75K	750K	7M5
23-	R82	8R2	82R	820R	8K2	82K	820K	8M2
24-	R91	9R1	91R	910R	9K1	91K	910K	9M1
								10M BLUE

COLOR CODES FOR THE WHOLE E12/E24 RANGE OF RESISTORS

The twelve odd rows - 1, 3, 5... - represent values available in the E12 range only, plus 10M

Reading 4-band resistors

The most "common" type of resistor has 4 bands and is called the 10% resistor. It now has a tolerance of 5% but is still called the "10% type" as the colours increase by 20% so that a resistor can be 10% higher or 10% lower than a particular value and all the resistors produced in a batch can be used.

The first 3 bands produce the resistance and the fourth band is the "tolerance" band.

Gold = 5%

(Silver =10% but no modern resistors are 10%!! - they are 5% 2% or 1%)

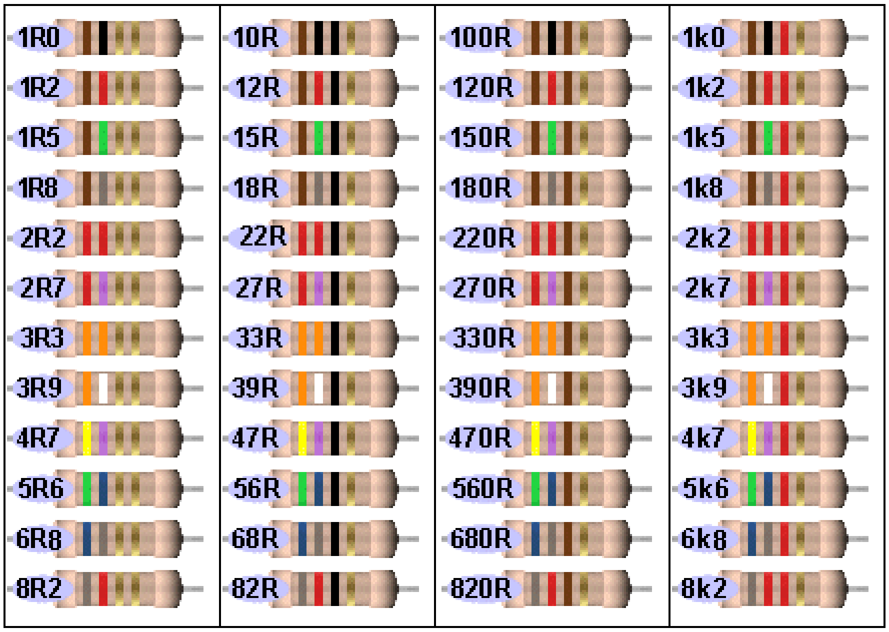
1R0
1R2
1R5
1R8
2R2
2R7
3R3
3R9
4R7
5R6
6R8
8R2
10R
12R
15R
18R
22R
27R
33R
39R
47R
56R
68R
82R
100R
120R
150R
180R
220R
270R
330R
390R
470R
560R
680R
820R
1k0
1k2
1k5
1k8
2k2
2k7
3k3
3k9
4k7
5k6
6k8
8k2

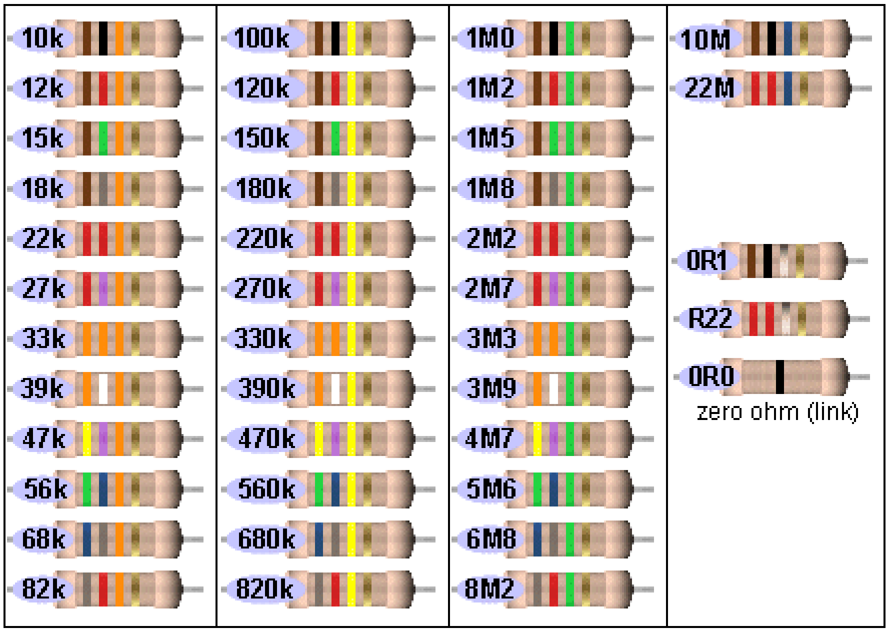
10k
12k
15k
18k
22k
27k
33k
39k
47k
56k
68k
82k
100k
120k
150k
180k
220k
270k
330k
390k
470k
560k
680k
820k
1M0
1M2
1M5
1M8
2M2
2M7
3M3
3M9
4M7
5M6
6M8
8M2
10M
22M
0R1
R22
0R0
zero ohm (link)

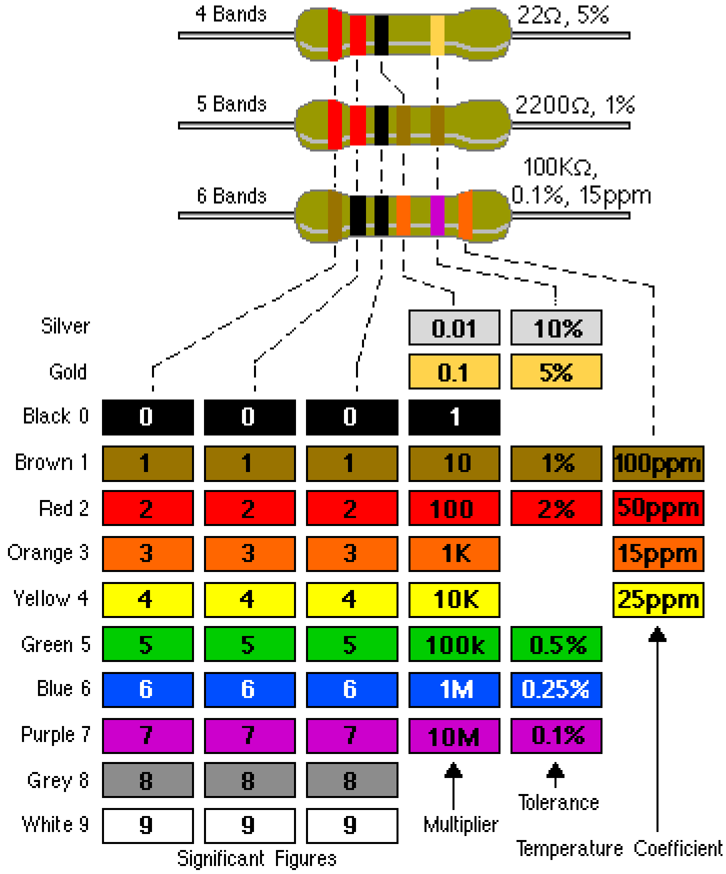
4 Bands
22Ω, 5%
5 Bands
2200Ω, 1%
6 Bands
100KΩ,
0.1%, 15ppm

	Significant Figures			Multiplier	Tolerance	Temperature Coefficient
Silver				0.01	10%	
Gold				0.1	5%	
Black 0	0	0	0	1		
Brown 1	1	1	1	10	1%	100ppm
Red 2	2	2	2	100	2%	50ppm
Orange 3	3	3	3	1K		15ppm
Yellow 4	4	4	4	10K		25ppm
Green 5	5	5	5	100k	0.5%	
Blue 6	6	6	6	1M	0.25%	
Purple 7	7	7	7	10M	0.1%	
Grey 8	8	8	8			
White 9	9	9	9			

Resistor Color Code System

Here is another well-designed resistor colour code chart:

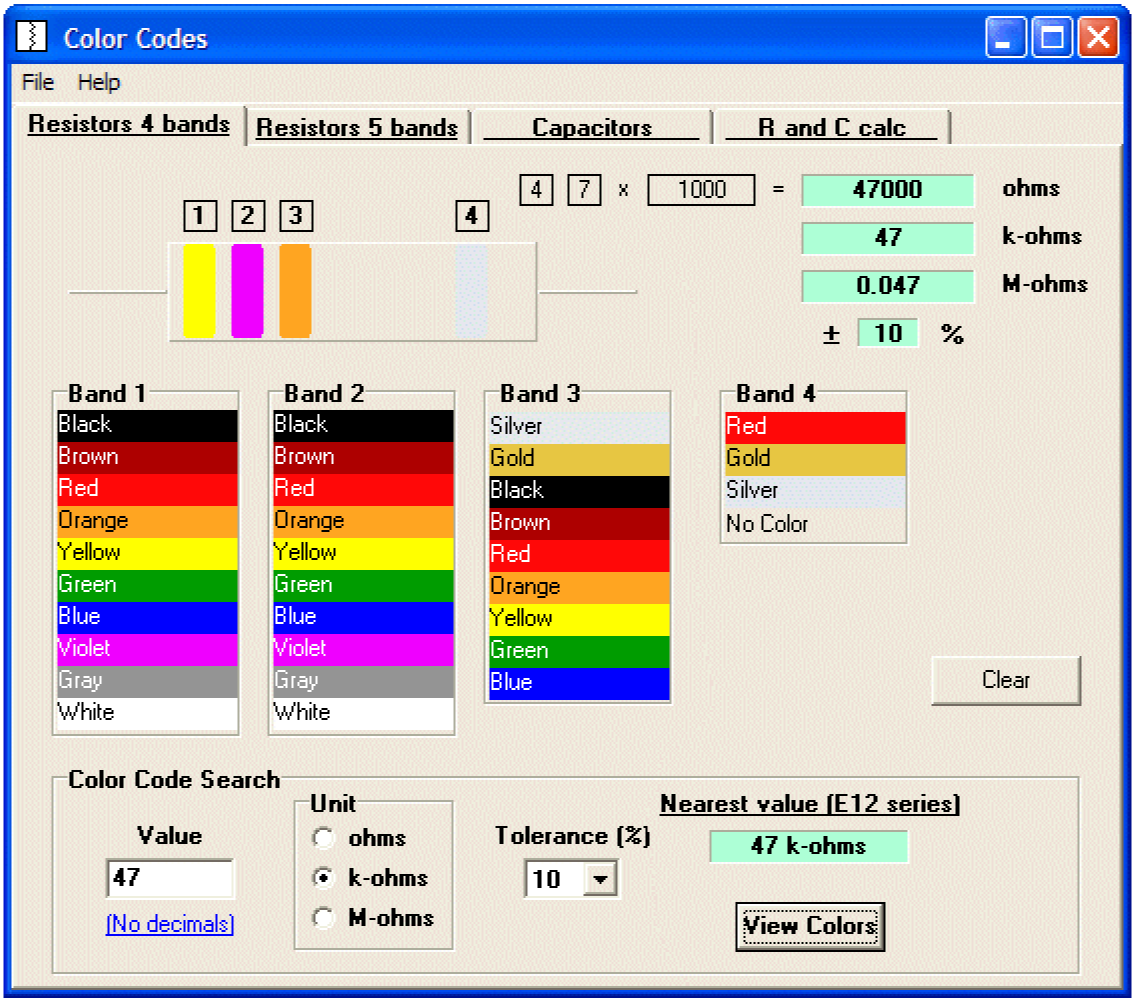
Color Codes
File
Help
Resistors 4 bands
Resistors 5 bands
Capacitors
R and C calc
1
2
3
4
Band 1
Black
Brown
Red
Orange
Yellow
Green
Blue
Violet
Gray
White
Band 2
Black
Brown
Red
Orange
Yellow
Green
Blue
Violet
Gray
White
Band 3
Silver
Gold
Black
Brown
Red
Orange
Yellow
Green
Blue
Band 4
Red
Gold
Silver
No Color
Clear
Color Code Search
(No decimals)
Nearest value (E12 series)
View Colors

4 7 × 1000 = 47000 ohms
47 k-ohms
0.047 M-ohms
± 10 %

Value: 47
Unit: ○ ohms ● k-ohms ○ M-ohms
Tolerance (%): 10
Nearest value (E12 series): 47 k-ohms

Download the program and save it on your desk-top for future reference:
ColourCode.exe (520KB)
ColourCode.zip (230KB)
ColourCode.rar (180KB)

RESISTORS LESS THAN 10 OHMS

When the **third** band is gold, it indicates the value of the "colors" must be divided by 10.
Gold = "divide by 10" to get values 1R0 to 8R2
When the **third** band is silver, it indicates the value of the "colors" must be divided by 100. (Remember: more letters in the word "silver" thus the divisor is "a larger division.")
Silver = "divide by 100" to get values R1 to R82
e.g: 0R1 = 0.1 ohm 0R22 = point 22 ohms
See 4th Column above for examples.

The letters "R, k and M" take the place of a decimal point.
e.g: 1**R**0 = 1 ohm 2**R**2 = 2 point 2 ohms 22**R** = 22 ohms
2**k**2 = 2,200 ohms 100**k** = 100,000 ohms
2**M**2 = 2,200,000 ohms

HOW TO REMEMBER THE COLOUR CODE:

Each colour has a "number" (or divisor) corresponding to it.
Most of the colours are in the same order as in the spectrum. You can see the spectrum in a rainbow. It is: ROY G BIV and the colours for resistors are in the same sequence.

black
brown - colour of increasing temperature

red
orange
yellow
green
blue
(indigo - that part of the spectrum between blue and violet)
violet
gray
white

colour	**value**	No of zero's
silver	-2	divide by 100
gold	-1	divide by 10
black	0	No zeros
brown	1	0
red	2	00
orange	3	,000 or k
yellow	4	0,000
green	5	00,000
blue	6	M
violet	7	
gray	8	
white	9	

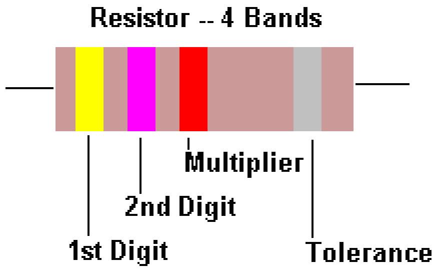
Resistor -- 4 Bands
Multiplier
2nd Digit
1st Digit
Tolerance

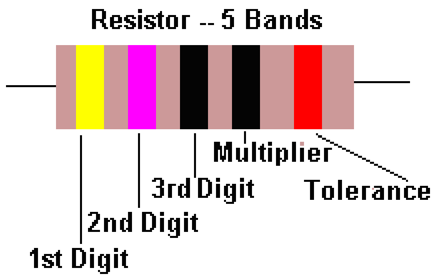
Resistor -- 5 Bands
Multiplier
3rd Digit
Tolerance
2nd Digit
1st Digit

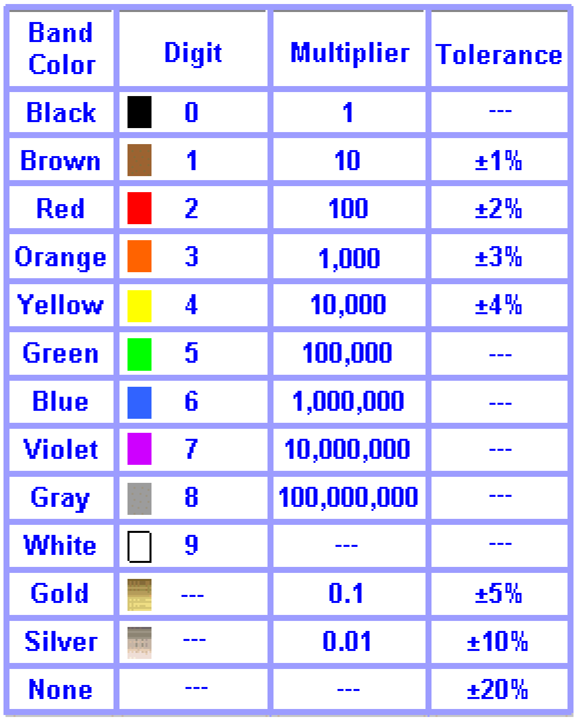

Band Color	Digit	Multiplier	Tolerance
Black	⬛ 0	1	---
Brown	🟫 1	10	±1%
Red	🟥 2	100	±2%
Orange	🟧 3	1,000	±3%
Yellow	🟨 4	10,000	±4%
Green	🟩 5	100,000	---
Blue	🟦 6	1,000,000	---
Violet	🟪 7	10,000,000	---
Gray	■ 8	100,000,000	---
White	⬜ 9	---	---
Gold	▨ ---	0.1	±5%
Silver	▨ ---	0.01	±10%
None	---	---	±20%

Here are some common ways to remember the colour code:

Bad Beer Rots Our Young Guts, But Vodka Goes Well
Bright Boys Rave Over Young Girls But Violet Gets Wed
Bad Boys Rave Over Young Girls But Violet Gets Wed with Gold and Silver.

Reading 5-band resistors:

5-band resistors are easy to read if you remember two simple points. The first three bands provide the digits in the answer and the 4th band supplies the number of zero's.

Reading "STANDARD VALUES" (**on 5-band resistors**)

5-band resistors are also made in "Standard Values" but will have different colours to 4-band "common" resistors - and will be confusing if you are just starting out. For instance, a 47k 5% resistor with 4-bands will be: yellow-purple-orange-gold. For a 47k 1% resistor the colours will be yellow-purple-black-red-brown. The brown colour-band represents 1%.
The first two colour-bands for a STANDARD VALUE or "common value" in 1% or 5% will be the SAME. These two bands provide the digits in the answer.
It's the 3rd band for a 5% resistor that is expanded into two bands in a 1% resistor. But it's easy to follow.
For a standard value, the 3rd band in a 1% resistor is BLACK. This represents a ZERO in the answer. (For 5-band resistors BLACK represents a ZERO when in the third band. This is different to 4-band resistors where black represents the word OHMS! If the third band is BROWN, the answer will be 1).
So the 4th band has to represent one-less ZERO and is one colour UP THE COLOUR CHART! In other words the 3rd and 4th bands (combined) on a 1% resistor produces the same number of zero's as the 3rd band on a 5% resistor!

Resistors come in a range of values and the two most common are the E12 and E24 series. The E12 series comes in twelve values for each decade. The E24 series comes in twenty-four values per decade.

E12 series - 10, 12, 15, 18, 22, 27, 33, 39, 47, 56, 68, 82

E24 series - 10, 11, 12, 13, 15, 16, 18, 20, 22, 24, 27, 30, 33, 36, 39, 43, 47, 51, 56, 62, 68, 75, 82, 91

Here is the complete list of 1% 1/4watt resistors from:
CIRCUIT SPECIALISTS. The following list covers 10 ohms (10R) to 1M.
To buy 1% resistors from Circuit Specialists, click: HERE.

10R
12R1
15R
18R2
22R1
27R4
30R1
33R2
36R5
39R2
47R5
49R9
51R1
56R2
68R1
75R
82R5
90R9
100R
121R
150R
182R
200R
221R
240R
249R
274R
301R
332R
348R
392R
402R
475R
499R
565R
604R
681R
750R
806R
825R
909R
1k0
1k21
1k5
1k82
2k
2k21
2k2
2k43
2k49
2k67
2k74
3k01
3k32
3k48
3k57
3k74
3k83
3k92
4k02
4k22
4k64
4k75
4k7
4k87
4k99
5k11
5k23
5k36
5k49
5k62
5k76
5k9
6k04
6k19
6k81
7k15
7k5
7k87
71k5
8k06
8k25
8k45
8k66
8k87
9k09
9k31
9k53
9k76
10k
11k
12k
12k1
12k4
13k
14k7
15k
15k8
16k9
17k4
17k8
18k2
20k
22k1
22k6
23k7
24k9
27k4
29k4
30k1
33k2
34k8
36k5
38k3
39k2
40k2
44k2
46k4
47k
47k5
49k9
51k1
53k6
56k2
61k9
68k1
69k8
75k0
82k5
90k
90k9
95k3
100k
121k
147k
150k
182k
200k
212k
221k
226k
249k
274k
301k
332k
357k
392k
475k
487k
499k
562k
604k
1M

Here is the list of 1% resistors from suppliers (such as Farnell):

1R0	13R	68R	360R	1k8	9k1	47k	240k
1R2	15R	75R	390R	2k0	10k	51k	270k
1R5	16R	82R	430R	2k2	11k	56k	300k
2R2	18R	91R	470R	2k4	12k	62k	330k
2R7	20R	100R	510R	2k7	13k	68k	360k
3R3	22R	110R	560R	3k	15k	75k	390k
3R9	24R	120R	620R	3k3	16k	82k	430k
4R7	27R	130R	680R	3k6	18k	91k	470k
5R6	30R	150R	750R	3k9	20k	100k	510k
6R2	33R	160R	820R	4k3	22k	110k	560k
6R8	36R	180R	910R	4k7	24k	120k	620k
7R5	39R	200R	1k	5k1	27k	130k	680k
8R2	43R	220R	1k1	5k6	30k	150k	750k
9R1	47R	240R	1k2	6k2	33k	160k	820k
10R	51R	270R	1k3	6k8	36k	180k	910k
11R	56R	300R	1k5	7k5	39k	200k	1M
12R	62R	330R	1k6	8k2	43k	220k	

Surface Mount Resistors

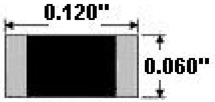

"120 thou"
x
"60 thou"
"one hundred and twenty thousands of an inch"

10R
10 ohm

0.047R

0.025R

sense resistor

121k

499R

240R

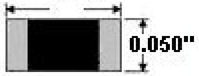

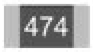

470 kΩ

11.8 ohm

10 ohm

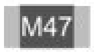

470 kΩ

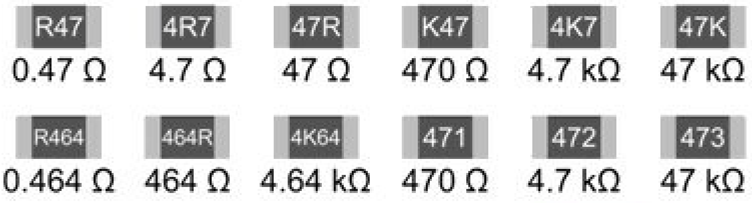
R47
0.47 Ω
4R7
4.7 Ω
47R
47 Ω
K47
470 Ω
4K7
4.7 kΩ
47K
47 kΩ
R464
0.464 Ω
464R
464 Ω
4K64
4.64 kΩ
471
470 Ω
472
4.7 kΩ
473
47 kΩ

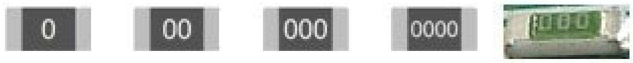
0
00
000
0000
000

SHORT-CIRCUITING "ZERO-OHM LINKS" OR "JUMPERS"

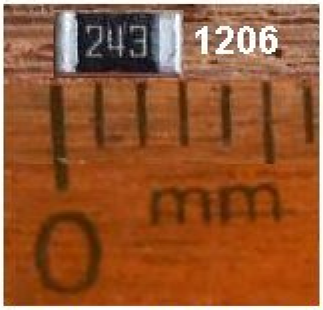
243
1206
0
mm

G0706AB
10R0
1502
1206
0805

1206 and 0805 SURFACE MOUNT RESISTORS

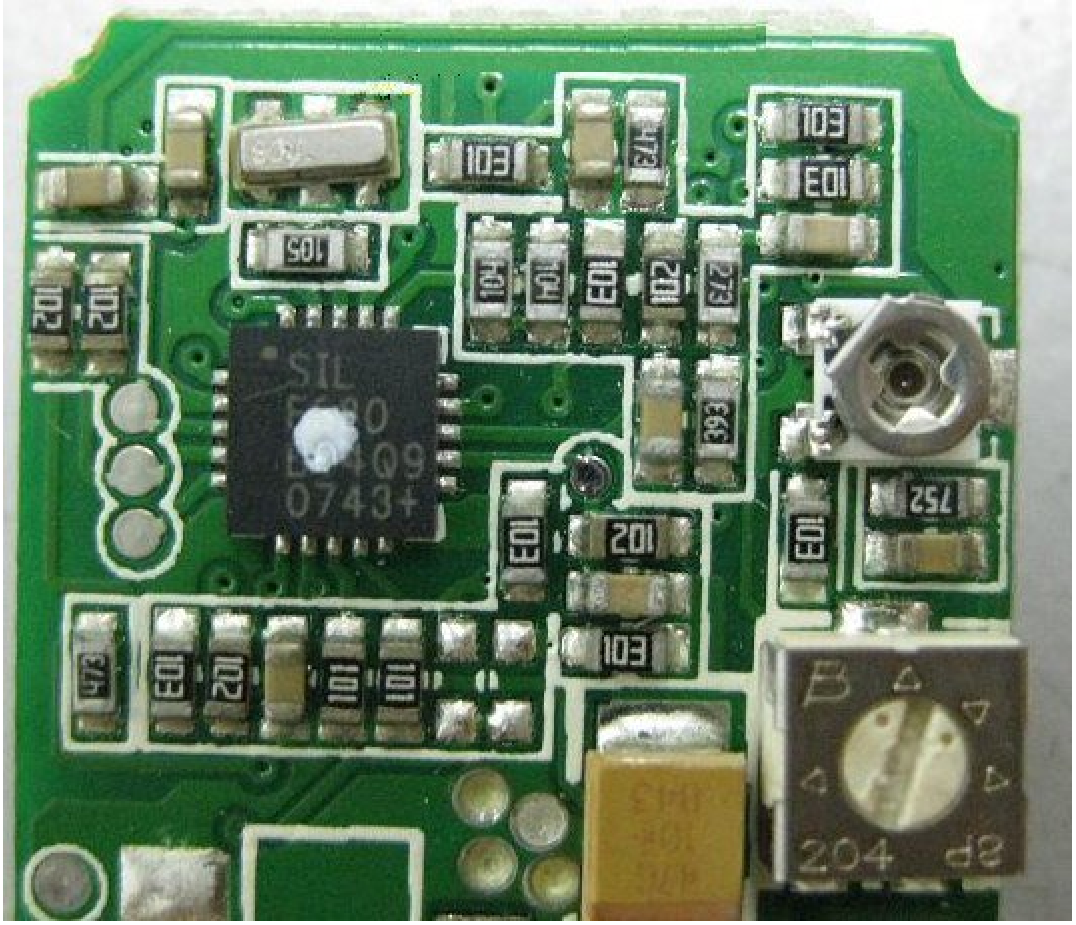
103
103
105
102
102
104
104
103
102
273
393
752
103
102
103
473
103
102
101
101
204

3-digit Surface Mount resistors on a PC board

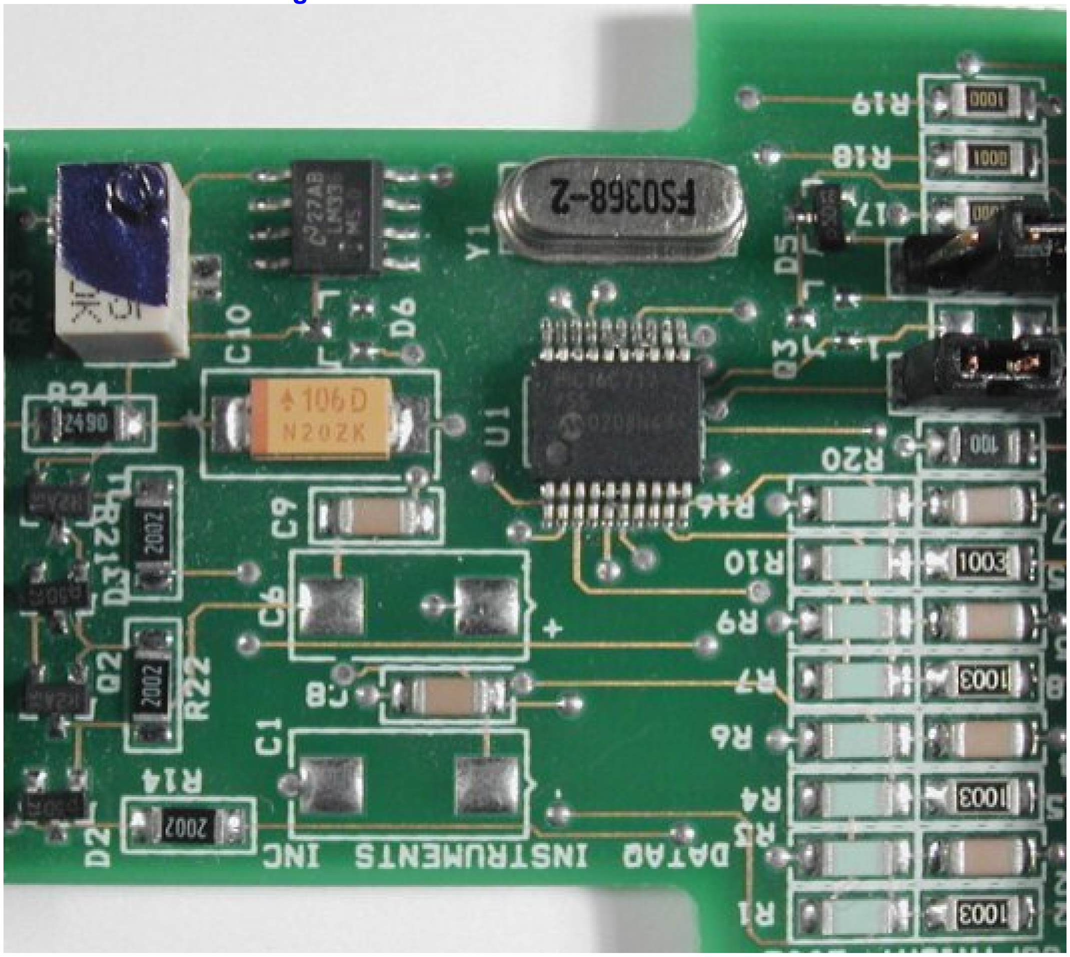
R19
1000
R18
1000
R17
FS0368-2
Y1
D5
D6
C10
R23
R24
2490
106D
N202K
U1
Q3
R20
100
R21
2002
C9
R16
R10
1003
C6
R9
R7
1003
Q2
2002
R22
C8
C1
R6
R4
1003
R14
2002
R3
D2
DATAQ INSTRUMENTS INC
R1
1003

4-digit Surface Mount resistors on a PC board

The photo above shows surface mount resistors on a circuit board. The components

that are not marked are capacitors (capacitors are NEVER marked).
All the SM resistors in the above photos conform to a 3-digit or 4-digit code. But there are a number of codes, and the 4-digit code caters for high tolerance resistors, so it's getting very complicated.
Here is a basic 3-digit SM resistor:

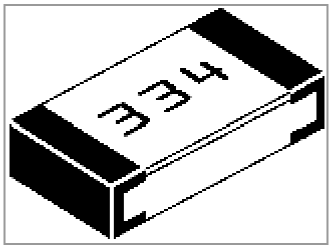

A 330k SM resistor

The first two digits represent the two digits in the answer. The third digit represents the number of zero's you must place after the two digits. The answer will be OHMS. For example: 334 is written 33 0 000. This is written 330,000 ohms. The comma can be replaced by the letter "k". The final answer is: 330k.
222 = 22 00 = 2,200 = 2k2
473 = 47 000 = 47,000 = 47k
474 = 47 0000 = 470,000 = 470k
105 = 10 00000 = 1,000,000 = 1M = one million ohms
There is one trick you have to remember. Resistances less than 100 ohms are written: 100, 220, 470. These are 10 and NO zero's = 10 ohms = 10R or 22 and no zero's = 22R or 47 and no zero's = 47R. Sometimes the resistor is marked: 10, 22 and 47 to prevent a mistake.

Remember:
R = ohms
k = kilo ohms = 1,000 ohms
M = Meg = 1,000,000 ohms
The 3 letters (R, k and M) are put in place of the decimal point. This way you cannot make a mistake when reading a value of resistance.

Surface Mount **CURRENT SENSING** Resistors

Many new types of CURRENT SENSING surface-mount resistors are appearing on the market and these are creating lots of new problems.
Fortunately all resistors are marked with the value of resistance and these resistors are identified in MILLIOHMS. A miili ohm is one thousandth or an ohm and is written 0.001 when writing a normal mathematical number.
When written on a surface mount resistor, the letter R indicates the decimal point and it also signifies the word "OHM" or "OHMS" and one milli-ohm is written R001
Five miiliohms is R005 and one hundred milliohms is R100
Some surface mount resistors have the letter "M" after the value to indicate the resistor has a rating of 1 watt. e.g: R100M These surface-mount resistors are specially-made to withstand a high temperature and a surface-mount resistor of the same size is normally 250mW or less.
These current-sensing resistors can get extremely hot and the PC board can become burnt or damaged.
When designing a PC board, make the lands very large to dissipate the heat.
Normally a current sensing resistor is below one ohm (1R0) and it is easy to identify them as R100 etc.
You cannot measure the value of a current sensing resistor as the leads of a multimeter have a higher resistance than the resistor and few multimeters can read values below one ohm.
If the value is not visible, you will have to refer to the circuit.
Before replacing it, work out why it failed.
Generally it gets too hot. Use a larger size and add tiny heatsinks on each end.
Here are some surface=mount current-sense resistors:

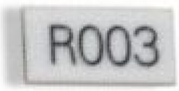

3 milli-ohm

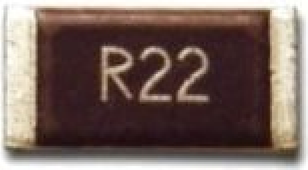

220 milli-ohm

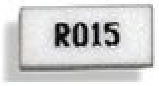

15 milli-ohm

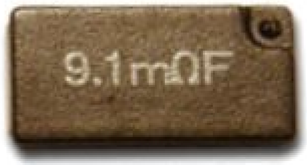

9 point 1 milli-ohm

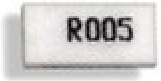

5 milli-ohm

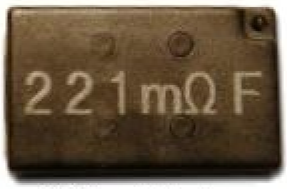

220 milli-ohm

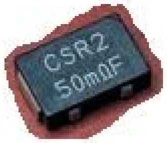

50 milli-ohm 1%

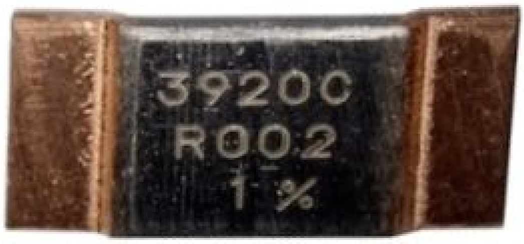
39200
R002
1 %
2 milli-ohm

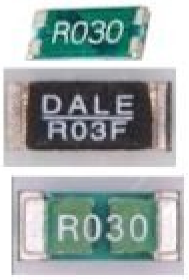
R030
DALE
R03F
R030
30 milli-ohm

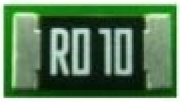

10 milli-ohm

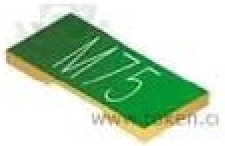

1-watt ????

THE COMPLETE RANGE OF SM RESISTOR MARKINGS

Click to see the complete range of SM resistor markings for 3-digit code:

Click to see the complete range of SM resistor markings for 4-digit code:

0000 is a value on a surface-mount resistor. It is a zero-ohm **LINK**!
Resistances less than 10 ohms have 'R' to indicate the position of the decimal point.
Here are some examples:

Three Digit Examples	Four Digit Examples
330 is 33 ohms - *not 330 ohms*	**1000 is 100 ohms - *not 1000 ohms***
221 is 220 ohms	**4992 is 49 900 ohms, or 49k9**
683 is 68 000 ohms, or 68k	**1623 is 162 000 ohms, or 162k**
105 is 1 000 000 ohms, or 1M	**0R56 or R56 is 0.56 ohms**
8R2 is 8.2 ohms	

A new coding system has appeared on **1% types**. This is known as the EIA-96 marking method. It consists of a three-character code. The first two digits signify the 3 significant digits of the resistor value, using the lookup table below. The third character - a letter - signifies the multiplier.

code	value	code	value	code	value	code	value	code	value	code	value
01	100	**17**	147	**33**	215	**49**	316	**65**	464	**81**	681
02	102	**18**	150	**34**	221	**50**	324	**66**	475	**82**	698
03	105	**19**	154	**35**	226	**51**	332	**67**	487	**83**	715
04	107	**20**	158	**36**	232	**52**	340	**68**	499	**84**	732
05	110	**21**	162	**37**	237	**53**	348	**69**	511	**85**	750
06	113	**22**	165	**38**	243	**54**	357	**70**	523	**86**	768
07	115	**23**	169	**39**	249	**55**	365	**71**	536	**87**	787
08	118	**24**	174	**40**	255	**56**	374	**72**	549	**88**	806
09	121	**25**	178	**41**	261	**57**	383	**73**	562	**89**	825
10	124	**26**	182	**42**	267	**58**	392	**74**	576	**90**	845
11	127	**27**	187	**43**	274	**59**	402	**75**	590	**91**	866
12	130	**28**	191	**44**	280	**60**	412	**76**	604	**92**	887
13	133	**29**	196	**45**	287	**61**	422	**77**	619	**93**	909
14	137	**30**	200	**46**	294	**62**	432	**78**	634	**94**	931
15	140	**31**	205	**47**	301	**63**	442	**79**	649	**95**	953
16	143	**32**	210	**48**	309	**64**	453	**80**	665	**96**	976

The **multiplier** letters are as follows:

letter	mult	letter	mult
F	**100000**	**B**	**10**
E	**10000**	**A**	**1**
D	**1000**	**X or S**	**0.1**
C	**100**	**Y or R**	**0.01**

22A is a 165 ohm resistor, **68C** is a 49900 ohm (49k9) and **43E** a 2740000 (2M74). This marking scheme applies to 1% resistors only.

A similar arrangement can be used for **2% and 5%** tolerance types. The multiplier letters are identical to 1% ones, but occur **before** the number code and the following **code** is used:

2%				**5%**			
code	**value**	**code**	**value**	**code**	**value**	**code**	**value**
01	100	**13**	330	**25**	100	**37**	330
02	110	**14**	360	**26**	110	**38**	360
03	120	**15**	390	**27**	120	**39**	390
04	130	**16**	430	**28**	130	**40**	430
05	150	**17**	470	**29**	150	**41**	470
06	160	**18**	510	**30**	160	**42**	510
07	180	**19**	560	**31**	180	**43**	560

08	200	**20**	620	**32**	200	**44**	620
09	220	**21**	680	**33**	220	**45**	680
10	240	**22**	750	**34**	240	**46**	750
11	270	**23**	820	**35**	270	**47**	820
12	300	**24**	910	**36**	300	**48**	910

With this arrangement, **C31** is 5%, 18000 ohm (18k), and **D18** is 510000 ohms (510k) 2% tolerance.
Always check with an ohm-meter (a multimeter) to make sure.

Chip resistors come in the following styles and ratings:
Style: 0402, 0603, 0805, 1206, 1210, 2010, 2512, 3616, 4022
Power Rating: 0402(1/16W), 0603(1/10W), 0805(1/8W), 1206(1/4W), 1210(1/3W), 2010(3/4W), 2512(1W), 3616(2W), 4022(3W)
Tolerance: 0.1%, 0.5%, 1%, 5%
Temperature Coefficient: 25ppm 50ppm 100ppm

EIA marking code for surface mount (SMD) resistors

01S = 1R	01R = 10R	01A = 100R	01B = 1k	01C = 10k	01D = 100k	01E = 1M	01F = 10M
02S = 1R02	02R = 10R2	02A = 102R	02B = 1k02	02C = 10k2	02D = 102k	02E = 1M02	
03S = 1R05	03R = 10R5	03A = 105R	03B = 1k05	03C = 10k5	03D = 105k	03E = 1M05	18F = 15M
04S = 1R07	04R = 10R7	04A = 107R	04B = 1k07	04C = 10k7	04D = 107k	04E = 1M07	
05S = 1R1	05R = 11R	05A = 110R	05B = 1k1	05C = 11k	05D = 110k	05E = 1M1	30F = 20M
06S = 1R13	06R = 11R3	06A = 113R	06B = 1k13	06C = 11k3	06D = 113k	06E = 1M13	
07S = 1R15	07R = 11R5	07A = 115R	07B = 1k15	07C = 11k5	07D = 115k	07E = 1M15	
08S = 1R18	08R = 11R8	08A = 118R	08B = 1k18	08C = 11k8	08D = 118k	08E = 1M18	
09S = 1R21	09R = 12R1	09A = 121R	09B = 1k21	09C = 12k1	09D = 121k	09E = 1M21	
10S = 1R24	10R = 12R4	10A = 124R	10B = 1k24	10C = 12k4	10D = 124k	10E = 1M24	
11S = 1R27	11R = 12R7	11A = 127R	11B = 1k27	11C = 12k7	11D = 127k	11E = 1M27	
12S = 1R3	12R = 13R	12A = 130R	12B = 1k3	12C = 13k	12D = 130k	12E = 1M3	
13S = 1R33	13R = 13R3	13A = 133R	13B = 1k33	13C = 13k3	13D = 133k	13E = 1M33	
14S = 1R37	14R = 13R7	14A = 137R	14B = 1k37	14C = 13k7	14D = 137k	14E = 1M37	
15S = 1R4	15R = 14R	15A = 140R	15B = 1k4	15C = 14k	15D = 140k	15E = 1M4	
16S = 1R43	16R = 14R3	16A = 143R	16B = 1k43	16C = 14k3	16D = 143k	16E = 1M43	
17S = 1R47	17R = 14R7	17A = 147R	17B = 1k47	17C = 14k7	17D = 147k	17E = 1M47	
18S = 1R5	18R = 15R	18A = 150R	18B = 1k5	18C = 15k	18D = 15k	18E = 1M5	
19S = 1R54	19R = 15R4	19A = 154R	19B = 1k54	19C = 15k4	19D = 154k	19E = 1M54	
20S = 1R58	20R = 15R8	20A = 158R	20B = 1k58	20C = 15k8	20D = 158k	20E = 1M58	
21S = 1R62	21R = 16R2	21A = 162R	21B = 1k62	21C = 16k2	21D = 162k	21E = 1M62	
22S = 1R65	22R = 16R5	22A = 165R	22B = 1k65	22C = 16k5	22D = 165k	22E = 1M65	
23S = 1R69	23R = 16R9	23A = 169R	23B = 1k69	23C = 16k9	23D = 169k	23E = 1M69	
24S = 1R74	24R = 17R4	24A = 174R	24B = 1k74	24C = 17k4	24D = 174k	24E = 1M74	
25S = 1R78	25R = 17R8	25A = 178R	25B = 1k78	25C = 17k8	25D = 178k	25E = 1M78	
26S = 1R82	26R = 18R2	26A = 182R	26B = 1k82	26C = 18k2	26D = 182k	26E = 1M82	
27S = 1R87	27R = 18R7	27A = 187R	27B = 1k87	27C = 18k7	27D = 187k	27E = 1M87	
28S = 1R91	28R = 19R1	28A = 191R	28B = 1k91	28C = 19k1	28D = 191k	28E = 1M91	
29S = 1R96	29R = 19R6	29A = 196R	29B = 1k96	29C = 19k6	29D = 196k	29E = 1M96	
30S = 2R0	30R = 20R0	30A = 200R	30B = 2k0	30C = 20k0	30D = 200k	30E = 2M0	
31S = 2R05	31R = 20R5	31A = 205R	31B = 2k05	31C = 20k5	31D = 205k	31E = 2M05	
32S = 2R10	32R = 21R0	32A = 210R	32B = 2k10	32C = 21k0	32D = 210k	32E = 2M10	
33S = 2R15	33R = 21R5	33A = 215R	33B = 2k15	33C = 21k5	33D = 215k	33E = 2M15	
34S = 2R21	34R = 22R1	34A = 221R	34B = 2k21	34C = 22k1	34D = 221k	34E = 2M21	
35S = 2R26	35R = 22R6	35A = 226R	35B = 2k26	35C = 22k6	35D = 226k	35E = 2M26	
36S = 2R32	36R = 23R2	36A = 232R	36B = 2k32	36C = 23k2	36D = 232k	36E = 2M32	
37S = 2R37	37R = 23R7	37A = 237R	37B = 2k37	37C = 23k7	37D = 237k	37E = 2M37	
38S = 2R43	38R = 24R3	38A = 243R	38B = 2k43	38C = 24k3	38D = 243k	38E = 2M43	
39S = 2R49	39R = 24R9	39A = 249R	39B = 2k49	39C = 24k9	39D = 249k	39E = 2M49	
40S = 2R55	40R = 25R5	40A = 255R	40B = 2k55	40C = 25k5	40D = 255k	40E = 2M55	
41S = 2R61	41R = 26R1	41A = 261R	41B = 2k61	41C = 26k1	41D = 261k	41E = 2M61	
42S = 2R67	42R = 26R7	42A = 267R	42B = 2k67	42C = 26k7	42D = 267k	42E = 2M67	
43S = 2R74	43R = 27R4	43A = 274R	43B = 2k74	43C = 27k4	43D = 274k	43E = 2M74	
44S = 2R80	44R = 28R0	44A = 280R	44B = 2k80	44C = 28k0	44D = 280k	44E = 2M80	
45S = 2R87	45R = 28R7	45A = 287R	45B = 2k87	45C = 28k7	45D = 287k	45E = 2M87	
46S = 2R94	46R = 29R4	46A = 294R	46B = 2k94	46C = 29k4	46D = 294k	46E = 2M94	
47S = 3R01	47R = 30R1	47A = 301R	47B = 3k01	47C = 30k1	47D = 301k	47E = 3M01	
48S = 3R09	48R = 30R9	48A = 309R	48B = 3k09	48C = 30k9	48D = 309k	48E = 3M09	
49S = 3R16	49R = 31R6	49A = 316R	49B = 3k16	49C = 31k6	49D = 316k	49E = 3M16	
50S = 3R24	50R = 32R4	50A = 324R	50B = 3k24	50C = 32k4	50D = 324k	50E = 3M24	
51S = 3R32	51R = 33R2	51A = 332R	51B = 3k32	51C = 33k2	51D = 332k	51E = 3M32	
52S = 3R4	52R = 34R0	52A = 340R	52B = 3k4	52C = 34k0	52D = 340k	52E = 3M4	
53S = 3R48	53R = 34R8	53A = 348R	53B = 3k48	53C = 34k8	53D = 348k	53E = 3M48	

54S = 3R57	54R = 35R7	54A = 357R	54B = 3k57	54C = 35k7	54D = 357k	54E = 3M57
55S = 3R65	55R = 36R5	55A = 365R	55B = 3k65	55C = 36k5	55D = 365k	55E = 3M65
56S = 3R74	56R = 37R4	56A = 374R	56B = 3k74	56C = 37k4	56D = 374k	56E = 3M74
57S = 3R83	57R = 38R3	57A = 383R	57B = 3k83	57C = 38k3	57D = 383k	57E = 3M83
58S = 3R92	58R = 39R2	58A = 392R	58B = 3k92	58C = 39k2	58D = 392k	58E = 3M92
59S = 4R02	59R = 40R2	59A = 402R	59B = 4k02	59C = 40k2	59D = 402k	59E = 4M02
60S = 4R12	60R = 41R2	60A = 412R	60B = 4k12	60C = 41k2	60D = 412k	60E = 4M12
61S = 4R22	61R = 42R2	61A = 422R	61B = 4k22	61C = 42k2	61D = 422k	61E = 4M22
62S = 4R32	62R = 43R2	62A = 432R	62B = 4k32	62C = 43k2	62D = 432k	62E = 4M32
63S = 4R42	63R = 44R2	63A = 442R	63B = 4k42	63C = 44k2	63D = 442k	63E = 4M42
64S = 4R53	64R = 45R3	64A = 453R	64B = 4k53	64C = 45k3	64D = 453k	64E = 4M53
65S = 4R64	65R = 46R4	65A = 464R	65B = 4k64	65C = 46k4	65D = 464k	65E = 4M64
66S = 4R75	66R = 47R5	66A = 475R	66B = 4k75	66C = 47k5	66D = 475k	66E = 4M75
67S = 4R87	67R = 48R7	67A = 487R	67B = 4k87	67C = 48k7	67D = 487k	67E = 4M87
68S = 4R99	68R = 49R9	68A = 499R	68B = 4k99	68C = 49k9	68D = 499k	68E = 4M99
69S = 5R11	69R = 51R1	69A = 511R	69B = 5k11	69C = 51k1	69D = 511k	69E = 5M11
70S = 5R23	70R = 52R3	70A = 523R	70B = 5k23	70C = 52k3	70D = 523k	70E = 5M23
71S = 5R36	71R = 53R6	71A = 536R	71B = 5k36	71C = 53k6	71D = 536k	71E = 5M36
72S = 5R49	72R = 54R9	72A = 549R	72B = 5k49	72C = 54k9	72D = 549k	72E = 5M49
73S = 5R62	73R = 56R2	73A = 562R	73B = 5k62	73C = 56k2	73D = 562k	73E = 5M62
74S = 5R76	74R = 57R6	74A = 576R	74B = 5k76	74C = 57k6	74D = 576k	74E = 5M76
75S = 5R9	75R = 59R0	75A = 590R	75B = 5k9	75C = 59k0	75D = 590k	75E = 5M9
76S = 6R04	76R = 60R4	76A = 604R	76B = 6k04	76C = 60k4	76D = 604k	76E = 6M04
77S = 6R19	77R = 61R9	77A = 619R	77B = 6k19	77C = 61k9	77D = 619k	77E = 6M19
78S = 6R34	78R = 63R4	78A = 634R	78B = 6k34	78C = 63k4	78D = 634k	78E = 6M34
79S = 6R49	79R = 64R9	79A = 649R	79B = 6k49	79C = 64k9	79D = 649k	79E = 6M49
80S = 6R65	80R = 66R5	80A = 665R	80B = 6k65	80C = 66k5	80D = 665k	80E = 6M65
81S = 6R81	81R = 68R1	81A = 681R	81B = 6k81	81C = 68k1	81D = 681k	81E = 6M81
82S = 6R98	82R = 69R8	82A = 698R	82B = 6k98	82C = 69k8	82D = 698k	82E = 6M98
83S = 7R15	83R = 71R5	83A = 715R	83B = 7k15	83C = 71k5	83D = 715k	83E = 7M15
84S = 7R32	84R = 73R2	84A = 732R	84B = 7k32	84C = 73k2	84D = 732k	84E = 7M32
85S = 7R5	85R = 75R0	85A = 750R	85B = 7k5	85C = 75k0	85D = 750k	85E = 7M5
86S = 7R68	86R = 76R8	86A = 768R	86B = 7k68	86C = 76k8	86D = 768k	86E = 7M68
87S = 7R87	87R = 78R7	87A = 787R	87B = 7k87	87C = 78k7	87D = 787k	87E = 7M87
88S = 8R06	88R = 80R6	88A = 806R	88B = 8k06	88C = 80k6	88D = 806k	88E = 8M06
89S = 8R25	89R = 82R5	89A = 825R	89B = 8k25	89C = 82k5	89D = 825k	89E = 8M25
90S = 8R45	90R = 84R5	90A = 845R	90B = 8k45	90C = 84k5	90D = 845k	90E = 8M45
91S = 8R66	91R = 86R6	91A = 866R	91B = 8k66	91C = 86k6	91D = 866k	91E = 8M66
92S = 8R87	92R = 88R7	92A = 887R	92B = 8k87	92C = 88k7	92D = 887k	92E = 8M87
93S = 9R09	93R = 90R9	93A = 909R	93B = 9k09	93C = 90k9	93D = 909k	93E = 9M09
94S = 9R31	94R = 93R1	94A = 931R	94B = 9k31	94C = 93k1	94D = 931k	94E = 9M31
95S = 9R53	95R = 95R3	95A = 953R	95B = 9k53	95C = 95k3	95D = 953k	95E = 9M53
96S = 9R76	96R = 97R6	96A = 976R	96B = 9k76	96C = 97k6	96D = 976k	96E = 9M76

If you want an accurate RESISTANCE measurement, remove the resistor from the circuit and use a Digital meter.

SURFACE MOUNT COMPONENTS - PACKS

Talking Electronics has packs of components for the repairman. The following packs are available:

SURFACE MOUNT RESISTOR PACK consists of 1 off each standard value
10 ohms to 1M & 2M2 (60 resistors)
$14.20 including pack and post

SURFACE MOUNT CAPACITOR PACK consists of:
2 - 10p 5 - 47p 5 - 100p 5 - 470p 5 - 1n 5 - 10n 5 - 22n 5 - 100n
5 - 1u 16v electrolytic 5 - 10u 16v electrolytic
(40 components)
$23.80 including pack and post

SURFACE MOUNT DIODE PACK consists of:
5 - 1N 4148 (marked as "A6")
$10.00 including pack and post

SURFACE MOUNT TRANSISTOR PACK consists of:
5 - BC 848 (marked as "1K") NPN
5 - BC858 PNP
$10.00 including pack and post

email **Colin Mitchell** for details on how to pay by credit card or PayPal.

CREATING ANY VALUE OF RESISTANCE

Any value of resistance can be created by connecting two resistors in PARALLEL or SERIES.
You can also create a higher wattage resistor by connecting them in SERIES OR PARALLEL.
We are only going to cover two EQUAL VALUE resistors in SERIES or in PARALLEL.
If you want to create a "Special Value," simply connect two resistors and read the value with a Digital Meter. Keep changing the values until you get the required value.
We are not going into series or Parallel formulae. You can easily find a value with a multimeter.

TWO EQUAL-VALUE RESISTORS IN SERIES

Two equal-value resistors IN SERIES creates a value of DOUBLE. You simply ADD the values.
This can be done with any to two values as shown. Three equal-value resistors in series is three times the value.

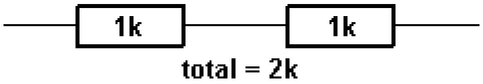
1k
1k
total = 2k

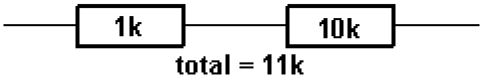
1k
10k
total = 11k

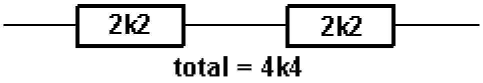
2k2
2k2
total = 4k4

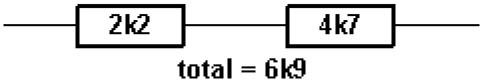
2k2
4k7
total = 6k9

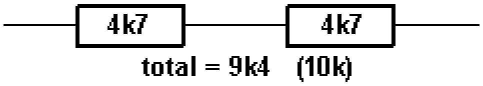
4k7
4k7
total = 9k4 (10k)

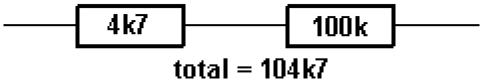
4k7
100k
total = 104k7

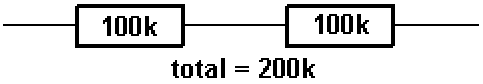
100k
100k
total = 200k

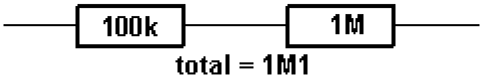
100k
1M
total = 1M1

TWO EQUAL-VALUE RESISTORS IN PARALLEL

Two equal-value resistors IN PARALLEL creates a value of HALF. Three equal-value resistors in parallel is equal to one-third the value.

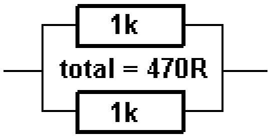
1k
total = 470R
1k

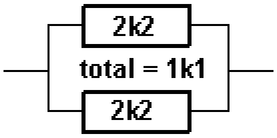
2k2
total = 1k1
2k2

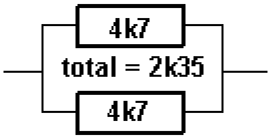
4k7
total = 2k35
4k7

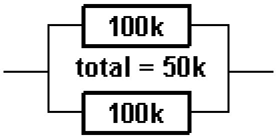
100k
total = 50k
100k

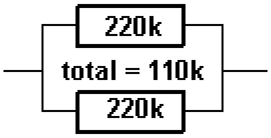
220k
total = 110k
220k

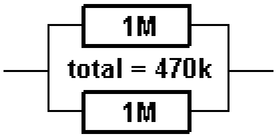
1M
total = 470k
1M

If you want a particular value and it is not available, here is a chart.
Use 2 resistors in series or parallel as shown:

Required Value	R1	Series/ Parallel	R2	Actual value:

10	4R7	S	4R7	9R4
12	10	S	2R2	12R2
15	22	P	47	14R9
18	22	P	100	18R
22	10	S	12	22
27	22	S	4R7	26R7
33	22	S	10	32R
39	220	P	47	38R7
47	22	S	27	49
56	47	S	10	57
68	33	S	33	66
82	27	S	56	83

There are other ways to combine 2 resistors in parallel or series to get a particular value. The examples above are just one way. 4R7 = 4.7 ohms

TESTING A RESISTOR

To check the value of a resistor, it should be removed from the circuit. The surrounding components can affect the reading and make it lower.
Resistors **VERY RARELY** change value, but if it is overheated or damaged, the resistance can increase. You can take the reading of a resistor "in-circuit" in one direction then the other, as the surrounding components may have diodes and this will alter the reading.
You can also test a resistor by feeling its temperature-rise. It is getting too hot if you cannot hold your finger on it (some "metal film" resistors are designed to tolerate quite high temperatures).

TESTING AN "AC" RESISTOR

There is no such thing as an "AC" resistor. Resistors are just "resistors" and they can be in AC circuits or DC circuits. Resistors can be given names such as "Safety Resistor" "Ballast Resistor" "LOAD Resistor" "Feed Resistor" "Dropper Resistor" or "Supply Resistor." These are just normal resistors with a normal resistance - except a "Safety Resistor."
A safety resistor is made of a flame-proof material such as metal-oxide-film and not carbon-composition. It is designed to "burn out" when too much current flows BUT NOT CATCH FIRE.
It is a low-value resistor and has a voltage-drop across it but this is not intentional. The voltage-drop is to create a "heating-effect" to burn out the resistor. In all the other types of resistor, the voltage-drop is intentional.
A Ballast resistor is a normal resistor and can be called a Power resistor, Dropper resistor, Supply resistor or Feed resistor. It is designed to reduce the voltage from one source and deliver a lower voltage. It is a form of: "in-line" resistor.
A Load Resistor is generally connected across the output of a circuit and turns the energy it receives, into heat.

RESISTOR NETWORKS

To reduce the number of resistors in a circuit, some engineers use a set of identical resistors in a package called a Single-In-Line (SIL) resistor network. It is made with many resistors of the same value, all in one package. One end of each resistor is connected all the other resistors and this is the common pin, identified as pin 1 and has a dot on the package.
These packages are very reliable but to make sure all the resistors are as stated, you need to locate pin 1. All values will be identical when referenced to this pin.

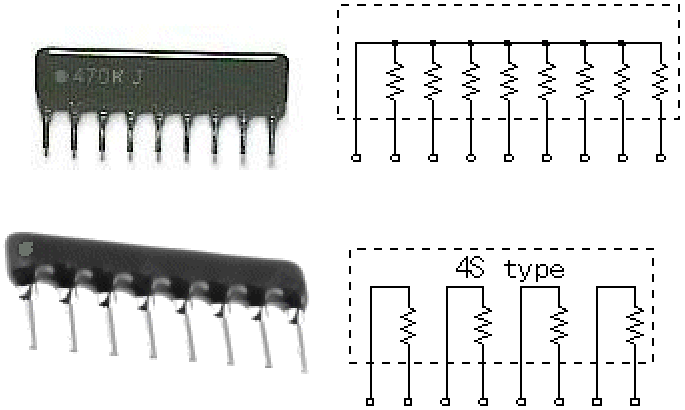

RESISTOR NETWORKS

Some resistor networks have a "4S" printed on the component. The 4S indicates the package contains 4 independent resistors that are not wired together inside. The housing has eight leads as shown in the second image.
Independent resistors have an even number of pins and measuring between each pair will produce identical values. Resistance between any pair will indicate leakage and may be a fault.

WIRE WOUND RESISTOR

A wire wound resistor is also called a POWER RESISTOR. This type of resistor can have a resistance as low as 0.1 ohms (one-tenth of an ohm) or as high as about 10k.
The image shows a 0.68 ohm resistor as the letter "R" represents the DECIMAL POINT and R68 is the same a .68 and this is 0.68 ohms. The wattage is 9 watts.
This resistor will allow xxx amps to flow. To work out the current, use the formula:
Power = Current x Current x resistance
9 = Current x Current x .68
Divide both sides by 0.68
13.2 = Current x Current
Find the square root of 13.2
Current = 3.6 amps

When 3.6 amps flow through the resistor, the voltage appearing across it will be:
V = current x resistance
= 3.6 x 0.68
= 2.5v and the wattage (heat) loss will be 9 watts.

The purpose of a resistor like this is to stop or reduce "ripple." Ripple is the noise or hum in an amplifier when the sound is turned up.
There are many reasons why you need to reduce the level of hum and this resistor will remove ripple as large as 2.5v when 3.6 amps is flowing, provided you have filter electrolytics on both side of the resistor to assist in removing the ripple.

If the letter "R" is in a different position, the value of resistance would be:

68R = 68Ω
6R8 = 6.8Ω
R68 = 0.68Ω

If you replace the R68 resistor a 6R8 resistor by mistake, the voltage across it will rise to 25v and if 3.6 amps flows, the wattage will be: 90 watts!!!
The resistor will glow red and burn out.

TESTING A POSISTOR

A Posistor is a resistor that connects in series with the degaussing coil around the picture tube or **Monitor.** When cold, it has a very low resistance and a large current flows when the monitor or TV is switched on. This current heats up the Posistor and the resistance increases. This causes the current to decrease and any magnetism in the shadow mask is removed. The posistor can one or two elements and it is kept warm so the resistance remains high. Many Posistors have a second element inside the case that connects directly to the supply to keep the Positive Temperature Coefficient resistor high so that the current through the degaussing coil falls to almost zero. This constant heat eventually destroys the package.
The heavy current that flows when a set is turned ON also causes the posistor to crack and break and this results in poor purity on the screen - as the shadow mask gradually becomes magnetic..
Posistors have different resistance values from different manufacturers and must be replaced with an identical type.
They can be checked for very low resistance when cold but any loose pieces inside the case will indicate a damaged component.

A "BURNT" RESISTOR - **normally and technically called a "burnt-out" resistor.**

The resistance of a "burnt" resistor can sometimes be determined by scraping away the outer coating - if the resistor has a spiral of resistance-material. You may be able to find a spot where the spiral has been damaged.

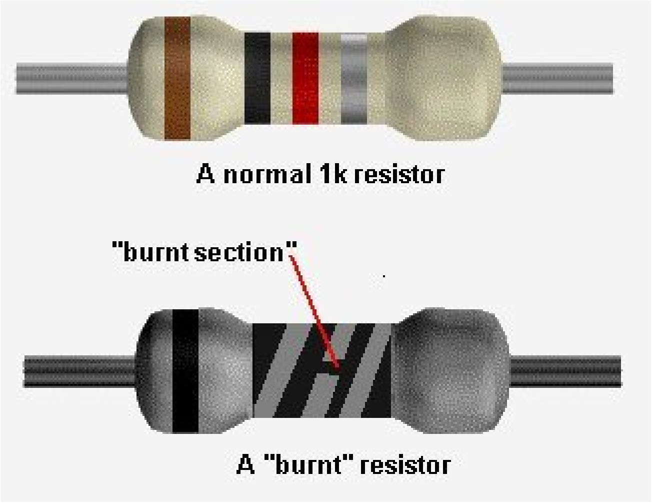

**Note the spirals of conductive carbon.
The number of spirals has nothing to with the resistance.
It is the amount of carbon particles in the "track" that
determines the resistance. It is also the thickness and width
of the track that determines the resistance.
And then it is the overall size of the resistor that determines the wattage.
And then the size of the leads, the closeness to the PCB and
the size of the lands that eventually determines how hot the resistor
will get.**

Clean the "spot" (burnt section of the spiral) very carefully and make sure you can get a good contact with the spiral and the tip of your probe. Measure from one lead of the resistor to the end of the damaged spiral. Then measure from the other lead to the other end of the spiral.
Add the two values and you have an approximate value for the resistor. You can add a small amount for the damaged section.
This process works very well for damaged wire-wound resistors. They can be pulled apart and each section of the resistance-wire (nichrome wire) measured and added to get the full resistance.

There is another way to determine the value of a damaged resistor.
Get a set of resistors of the same wattage as the damaged component and start with a high value. It's handy to know if the resistor is in the range: 10ohm to 100ohms or 1k to 10k etc, but this is not essential.
Start with a very high value and turn the circuit ON. You can perform voltage tests and if you know the expected output voltage, decrease the resistance until this voltage is obtained.
If you do not know the expected voltage, keep reducing the value of resistance until the circuit works as designed.
This is the best advice in a situation where you do not know the value of a resistor.

There is a third way to determine the value and this requires measuring the voltage drop across the resistor and the current-flow. By multiplying the two you will get a wattage and this must be less than the wattage of the resistor being replaced.

TESTING POTENTIOMETERS (variable resistors)

To check the value of a variable resistor, it should be removed from circuit or at least 2 legs should be removed. A Rheostat is a variable resistor using only one end and the middle connected to a circuit.
The resistance between the two outside pins is the value marked on the component and the centre leg will change from nearly zero to the full resistance as the shaft is rotated.
"Pots" generally suffer from "crackle" when turned and this can be fixed by spraying up the shaft and into the pot via the shaft with a tube fixed to a can of "spray-lubricant" (contact cleaner).
"Pre-set pots" and "trim pots" are miniature versions of a potentiometer and they are all tested the same. The photo shows a pot, two mini pots and 3 mini trim pots.

103
102
P102
N911

10-Turn POTs

A 10-turn pot is one of the worst items to be designed. I remove them immediately from any design.
You don't know the position of the wiper. You don't know which way you are turning the wiper and you can't remember which way you turned the post "the last time."
The screwdriver always falls out of the slot.
If you need fine adjustment, place fixed resistors on each side of the pot and use a normal mini trim pot with much less resistance.

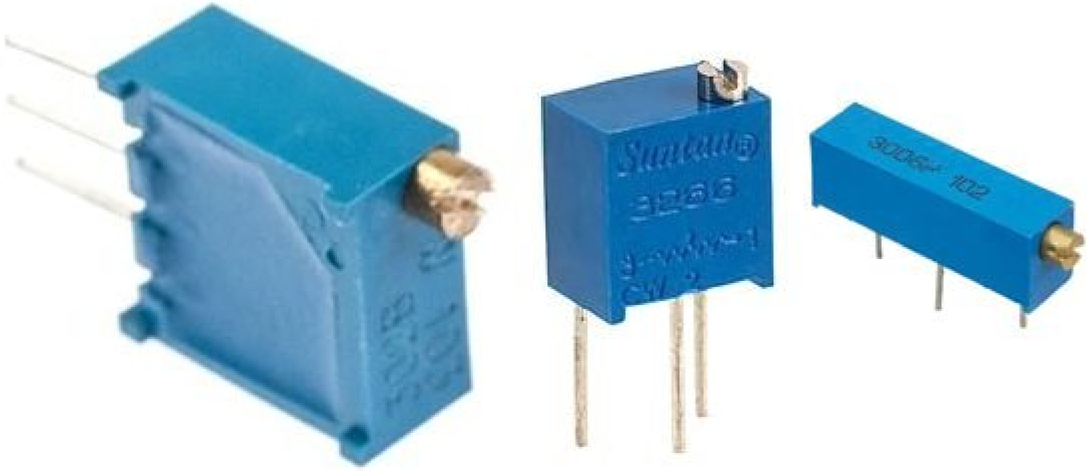

FOCUS POTS

Focus pots quite often get a spot of dirt where the wiper touches the track. Cleaning with spray fixes the bad focus but if the pot is leaking to chassis from inside the pot (due to the high voltage on the terminals) simply remove it from the chassis and leave it floating (this will restore the high voltage to the picture tube) or you can use one from an old chassis.

MAKING YOUR OWN RESISTOR, CAPACITOR, INDUCTOR or DIODE

Quite often you will not have the exact value of resistance or capacitance for a repair.
We have already covered placing resistors and capacitors in parallel and series:

Resistors in Parallel and/or Series
Capacitors in Parallel and/or Series

Here are some extras:

RESISTORS
Two 1k 0.5watt resistors in parallel produces a 470R 1watt resistor.
Two 1k 0.5watt resistors in series produces a 2k 1watt resistor.

CAPACITORS
Two 100n 100v capacitors in series produces a 50n capacitor @200v

INDUCTORS: Two inductors in series - **ADD THE VALUES**

DIODES: Two 1Amp 400v diodes in series produces a 1Amp 800v diode
Two 1Amp 400v diodes in parallel produces a 2Amp 400v diode

ZENER DIODES: Zener diodes can be connected in series to get a higher voltage.
Two 12v zener diodes in series produces a 24v zener.

CONTINUITY

Some multimeters have a "buzzer" that detects when the probes are touching each other or the resistance between the probes is very LOW. This is called a CONTINUITY TESTER.
You can use the resistance scale "x1" or "x10" to detect low values of resistance.
Set the pointer to "0" (right end of the scale) by touching the probes together and adjusting the "zero ohms" control.
When taking a reading, you will have to decide if a low value of resistance is a short-circuit or an "operating value."
For instance, the cold resistance of a 12v car globe is very low (about 2 ohms) and it increases (about 6 times) to 12 ohms when hot.
The "resistance of a circuit" may be very low as the electrolytics in the circuit are uncharged. This may not indicate a true "short-circuit."
The measurement across a diode is not a resistance-value but a "voltage-drop" and that is why the needle swings nearly full-scale.
Leads and wires and cords have a small resistance and depending on the length of the lead, this small resistance may be affecting a circuit.
Remember this:
When a circuit takes 1 amp, and the resistance of the leads is 1 ohm, the voltage drop across the leads will be 1v.
That's why a 12v battery supplying a circuit with these leads will have 11v at the circuit.
Note:
Turn off the equipment before making any continuity tests. The presence of even a small voltage (from an electrolytic) can give a false reading.
You can determine the resistance of a lead very accurately by taking the example above and applying it to your circuit.
If the battery is 12.6v and the voltage across the circuit is 10v, when the current is 2.6 amps, the resistance of the "leads" is 12.6 - 10 = 2.6 R=V/I = 2.6/2.6 = 1ohm. By making the lead shorter or using thicker wire, the resistance will be less and the voltage on the project will increase.
When taking readings in a circuit that has a number of diodes built-into IC's (Integrated Circuits) and transistors, some Continuity Testers will beep and give a false reading.
The following circuit has the advantage of providing a beep when a short-circuit is detected but does not detect the small voltage drop across a diode. This is ideal when testing logic circuits as it is quick and you can listen for the beep while

concentrating on the probe. Using a multimeter is much slower.

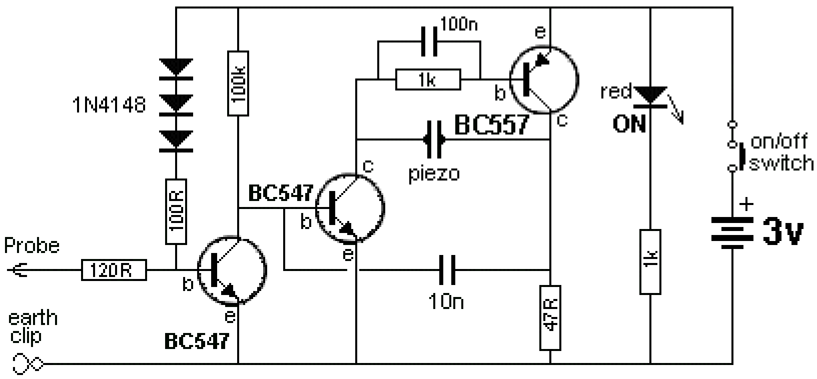

CONTINUITY TESTER

You can build the circuit on Matrix Board and add it to your Test Equipment.
You will need lots of "Test Equipment" and they can be built from circuits in this eBook.

TESTING FUSES, LEADS AND WIRES

All these components come under the heading TESTING for CONTINUITY. Turn off all power to the equipment before testing for shorts and continuity. Use the low resistance "Ohms Scale" or CONTINUITY range on your multimeter. All fuses, leads and wires should have a low, very low or zero resistance. This proves they are working.

A BLOWN FUSE

The appearance of a fuse after it has "blown" can tell you a lot about the fault in the circuit.
If the inside of the glass tube (of the fuse) is totally blackened, the fuse has been damaged very quickly. This indicates a very high current has passed through the fuse.
Depending on the rating of the fuse, (current rating) you will be able to look for components that can pass a high current when damaged - such as high power transistors, FETs, coils, electrolytics. Before re-connecting the supply, you should test the "SUPPLY RAILS" for resistance. This is done by measuring them on a low OHMs range in one direction then reverse the leads to see if the resistance is low in the other direction.
A reading can be very low at the start because electrolytics need time to charge-up and if the reading gradually increases, the power rail does not have a short. An overload can occur when the supply voltage rises to nearly full voltage, so you sometimes have to fit a fuse and see how long it takes to "blow."
If the fuse is just slightly damaged, you will need to read the next part of this eBook, to see how and why this happens:

FAST AND SLOW BLOW FUSES

There are many different sizes, shapes and ratings of a fuse. They are all current ratings as a fuse does not have a voltage rating. Some fuses are designed for cars as they fit into the special fuse holders. A fuse can be designed for 50mA, 100mA, 250mA, 315mA, 500mA, 1Amp, 1.5amp, 2amp, 3amp, 3.15amp 5amp, 10amp, 15amp, 20amp, 25amp, 30amp, 35amp, 50amp and higher.
Some fuses are fast-blow and some are slow-blow.
A "normal" fuse consists of a length of thin wire. Or it may be a loop of wire that is thin near the middle of the fuse. This is the section that will "burn-out."
A "normal" fuse is a fast-blow fuse. For instance, a 1amp fuse will remain intact when up to 1.25 amp flows. When a circuit is turned on, it may take 2-3 amps for a very short period of time and a normal 1 amp fuse will get very hot and the wire will stretch but not "burn-out." You can see the wire move when the supply turns on.
If the current increases to 2amps, the fuse will still remain intact. It needs about 3 amp to heat up the wire to red-hot and burn out.

If the current increases to 5 amp, the wire VOLATILISES (burns-out) and deposits carbon-black on the inside of the glass tube.
A slow-blow fuse uses a slightly thicker piece of wire and the fuse is made of two pieces of wire joined in the middle with a dob of low-temperature solder. Sometimes one of the pieces of wire is a spring and when the current rises to 2.5 amp, the heat generated in the wire melts the solder and the two pieces of wire "spring apart."
A slow-blow fuse will allow a higher current-surge to pass through the fuse and the wire will not heat up and sag.
Thus the fuse is not gradually being damaged and it will remain in a perfect state for a long period of time.
A fuse does not protect electronic equipment from failing. It acts AFTER the equipment has failed.
It will then protect a power supply from delivering a high current to a circuit that has failed.
If a slow-blow fuse has melted the solder, it could be due to a slight overload, slight weakening of the fuse over a period of time or the current-rating may be too low. You can try another fuse to see what happens.
You can replace a fast-acting fuse (normal fuse) with a slow blow if the fast-acting fuse has been replaced a few times due to deterioration when the equipment is turned on.
But you cannot replace a slow-blow fuse with a fast acting fuse as it will be damaged slightly each time the equipment is turned on and eventually fail.

100mA FUSES

Fuses below about 100mA are very hard to make and very unreliable.
Many circuits take a high current when turned to charge the electrolytics and a 100mA (or 50mA or 63mA fuse) will bow and stretch and change shape, every time the equipment is turned ON. Eventually it will break, due to it heating-up and stretching.
To produce a reliable fuse below 100mA, some manufacturers have placed a resistor inside the fuse and connected it to a spring. One end of the resistor is soldered to a wire with low-temperature metal and when the resistor gets hot, the metal softens and the spring pulls the resistor away from the wire.
Quite often you can heat up the metal and connect the wire and the fuse is perfect.
This type of fuse is called a DELAY fuse and the current rating is shown on the end-cap.
The value of the resistor determines the current rating.
There is a small voltage across this type of fuse and it means the circuit sees a slightly lower voltage than the supply voltage.
The third photo shows the pot of solder or low-temp metal and a wire connected to a spring. The heat generated in the wire is passed to the solder and it softens. The spring pulls the two components apart. You can smash the glass and set up the fuse in the two fuse-holders and repair the fuse while you wait for a new fuse to arrive.

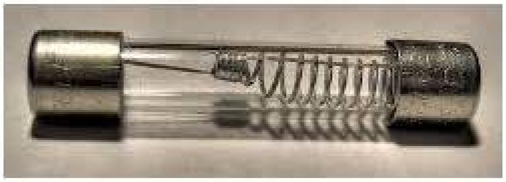

TESTING COILS, INDUCTORS and YOKES

Coils, inductors, chokes and yokes are just coils (turns) of wire. The wire may be wrapped around a core made of iron or ferrite.
It is labeled "L" on a circuit board.
You can test this component for continuity between the ends of the winding and also make sure there is no continuity between the winding and the core.
The winding can be less than one ohm, or greater than 100 ohms. A coil of wire is also called an INDUCTOR and it might look like a very simple component, but it can operate in a very complex way.
The way it works is a discussion for another eBook. It is important to understand the turns are insulated but a slight fracture in the insulation can cause two turns to touch each other and this is called a "SHORTED TURN" or you can say the inductor has "SHORTED TURNS."
When this happens, the inductor allows the circuit to draw MORE CURRENT. This causes the fuse to "blow."
The quickest way to check an inductor is to replace it, but if you want to measure the inductance, you can use an INDUCTANCE METER. You can then compare the inductance with a known good component.
An inductor with a shorted turn will have a very low or zero inductance, however you may not be able to detect the fault when it is not working in a circuit as the fault may be created by a high voltage generated between two of the turns.
Faulty yokes (both horizontal and vertical windings) can cause the picture to reduce in size and/or bend or produce a single horizontal line.
A TV or monitor screen is the best piece of Test Equipment as it has identified the fault. It is pointless trying to test the windings further as you will not be able to test them under full operating conditions. The fault may not show up when a low voltage (test voltage) is applied.

MEASURING AND TESTING INDUCTORS

Inductors are measured with an INDUCTANCE METER but the value of some inductors is very small and some Inductance Meters do not give an accurate reading. The solution is to measure a larger inductor and note the reading. Now put the two inductors in SERIES and the values ADD UP - just like resistors in SERIES. This way you can measure very small inductors. VERY CLEVER!

Question from a reader: *Can I add an inductor to stop a fuse blowing?*

Basically, an inductor NEVER prevents a fuse blowing because an inductor prevents spikes on one lead (we will call the INPUT lead), appearing on its other lead.
This is the detection and prevention of current that exists for a very short period of time.
A fuse detects an excess of current that occurs over a very long period of time and they are entirely two different "detectors."
One cannot assist the other in any way.

An inductor is basically a coil of wire. It may be thick or thin wire. The value of the inductor is a combination of the number of turns and the material on which the wire is wound.
The value of an inductor does not change over say a period of 20 years but it can go faulty by the enamel cracking and two turns touching. This can also be due to the difference in voltage between the two turns creating a spark between the turns and creating a "short."
When you test it, the high voltage is not present and it will test ok.
You may not think a few turns of wire will have any effect on improving a circuit, but spikes are very high frequency and the inductor will have a very big effect on reducing them.
An inductor (say 100uH) can be produced in many different sizes and the thickness of the wire will be important as it determines the current that can flow through the inductor.
The term "inductor" also includes those with two or more windings and these components are called TRANSFORMERS. These devices can get "shorts" and "leaks" between the windings and sparks can be seen between the windings. These sparks do not occur when you are testing them on test-equipment so the only way to

guarantee success is to replace it with an identical replacement.

TESTING SWITCHES and RELAYS

Switches and relays have contacts that open and close mechanically and you can test them for CONTINUITY. However these components can become intermittent due to dirt or pitting of the surface of the contacts due to arcing as the switch is opened.
It is best to test these items when the operating voltage and current is present as they quite often fail due to the arcing. A switch can work 49 times then fail on each 50th operation. The same with a relay. It can fail one time in 50 due to CONTACT WEAR.
If the contacts do not touch each other with a large amount of force and with a large amount of the metal touching, the current flowing through the contacts will create HEAT and this will damage the metal and sometimes reduce the pressure holding the contact together.
This causes more arcing and eventually the switch heats up and starts to burn. Switches are the biggest causes of fire in electrical equipment and households.

A relay also has a set of contacts that can cause problems.
There are many different types of relays and basically they can be put into two groups.
1. An electromagnetic relay is a switch operated by magnetic force. This force is generated by current through a coil. The relay opens and closes a set of contacts. The contacts allow a current to flow and this current can damage the contacts.
Connect 5v or 12v to the coil (or 24v) and listen for the "click" of the points closing.
Measure the resistance across the points to see if they are closing.
You really need to put a load on the points to see if they are clean and can carry a current.
The coil will work in either direction.
If not, the relay is possibly a CMOS relay or Solid State relay.

2. An electronic relay (Solid State Relay) does not have a winding. It works on the principle of an opto-coupler and uses a LED and Light Activated SCR or Opto-TRIAC to produce a low resistance on the output. The two pins that energise the relay (the two input pins) must be connected to 5v (or 12v) around the correct way as the voltage is driving a LED (with series resistor). The LED illuminates and activates a light-sensitive device.

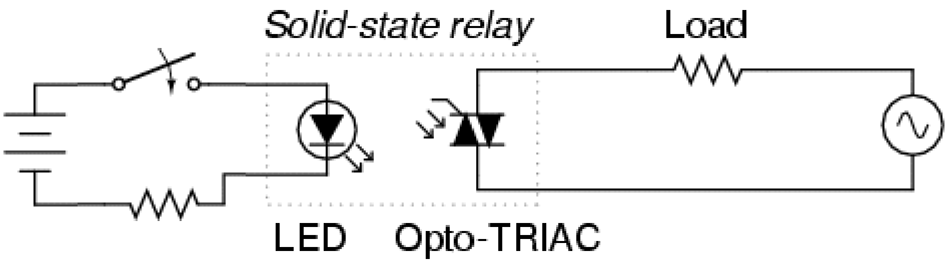

TESTING REED SWITCHES

A reed switch is generally contained in a long glass tube:

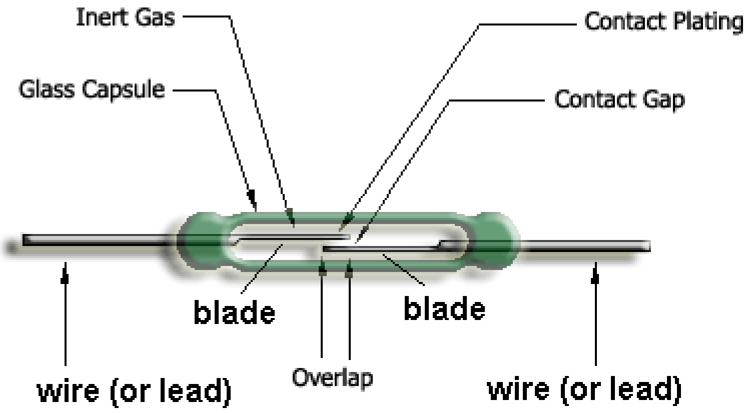

A wire or lead comes out each end for soldering to the reed switch to the project. The two "blades" inside the switch are made from a material that can be magnetised but does not retain its magnetism. This effect is called "temporally magnetised" (not permanently magnetised) and really only "passes" magnetism from one end to the other when in the presence of a magnet. One of the blades is made of a soft material and it will bend very easily. The other one is much stiffer.

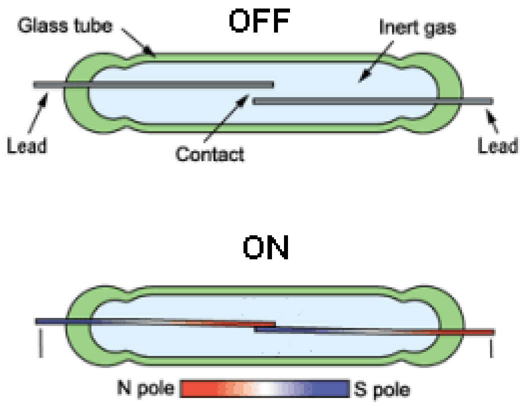

When a magnet is placed under the two blades (or on top), the magnetism from the magnet is passed to the two blades (INDUCTION or MUTUAL INDUCTION - commonly called INDUCED) and it produces a very weak magnet (in the blade) that is identical to the powerful magnet as far as the position of the north and south poles are concerned). Initially it produces a N-S and N-S set of poles and this makes the two blades click together because the top blade will be South at the contact and the bottom blade will be North.

When the two blades click together the magnetism runs through the two blades and keeps them together. The two blades attract and the switch is closed. When the magnet is removed, the magnetism in the two blades ceases and the two blades move apart.

Since there is a very small amount of movement in the top blade, this switch has a limited number of operations. Eventually it will fail. It is a mechanical device and is not suited for detecting a spinning shaft as 100,000 revolutions will very quickly weaken the switch.

If the switch does not make contact or remains closed, the moveable blade can be cracked or broken. This can be very hard to see. So replace the switch.

LATCHING REED SWITCH

A "normal" reed switch can be converted into a **LATCHING REED SWITCH** by carefully placing a magnet below the switch and moving it away so the two blades open. Now move it slightly closer but do not allow the blades to close.

This is called putting a "SET" on the switch and the two blades will have a small magnetic effect "induced" in them but not enough to close the contacts:

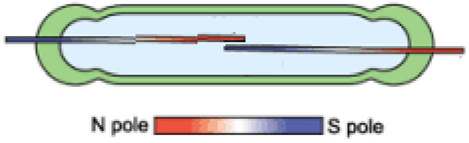

Now bring a strong magnet up to the reed switch on the other side of the glass tube with the north pole above the north of the lower magnet. This effect will increase the INDUCED MAGNETISM in the blades and close the contacts:

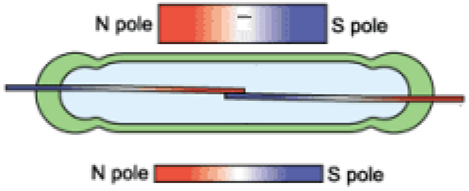

Remove the top magnet and the lower magnet will induce enough magnetism into the blades to keep them closed:

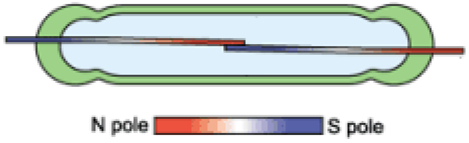

Now bring the upper magnet near the reed switch with the south pole above the north pole of the lower magnet. (In other words: AROUND THE OTHER WAY) This will have the effect of reducing the induced magnetism in the blades and a point will be reached when the two contacts will separate:

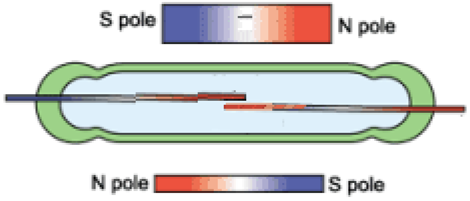

Remove the top magnet and the switch will remain separated because the lower magnet will not have sufficient influence on the blades to close the contact:

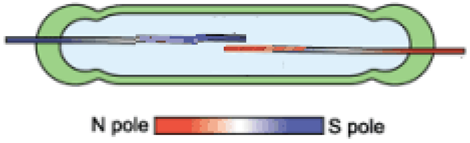

CAPACITORS

Capacitors are one of the most difficult things to test. That's because they don't give a reading on a multimeter and their value can range from 1p to 100,000u.
A faulty capacitor may be "open" when measured with a multimeter, and a good capacitor will also be "open."

You need a piece of test equipment called a CAPACITANCE METER to measure the value of a capacitor.

HOW A CAPACITOR WORKS

There are two ways to describe how a capacitor works. Both are correct and you have to combine them to get a full picture.
A capacitor has INFINITE resistance between one lead and the other.
This means no current flows **through** a capacitor. But it works in another way.
Suppose you have a strong magnet on one side of a door and a piece of metal on the other. By sliding the magnet up and down the door, the metal rises and falls.
The metal can be connected to a pump and you can pump water by sliding the magnet up and down.
A capacitor works in exactly the same way.
If you raise a voltage on one lead of a capacitor, the other lead will rise to the same voltage. This needs more explaining - we are keeping the discussion simple.
It works just like the magnetic field of the magnet through a door.
The next concept is this:
Capacitors are equivalent to a tiny rechargeable battery.
They store energy when the supply-voltage is present and release it when the supply drops.
These two concepts can be used in many ways and that's why capacitors perform tasks such as filtering, time-delays, passing a signal from one stage to another and create many different effects in a circuit.

CAPACITOR VALUES

The basic unit of capacitance is the FARAD. (C) This is the value used in all equations, but it is a very large value. A one FARAD capacitor would be the size of a car if made with plates and paper. Most electronic circuits use capacitors with smaller values such as 1p to 1,000u. 1p is about equal to two parallel wires 2cm long. 1p is one picofarad.

The easiest way to understand capacitor values is to start with a value of 1u. This is one microfarad and is one-millionth of a Farad. A 1 microfarad capacitor is about 1cm long and the diagram shows a 1u electrolytic.

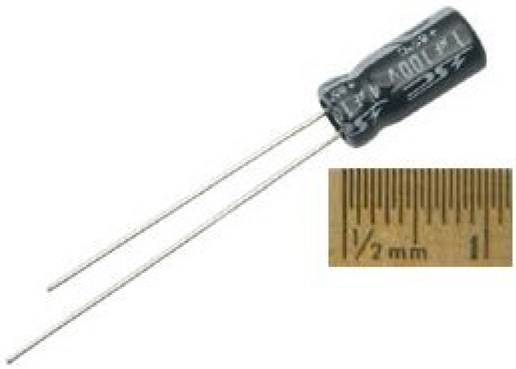

Smaller capacitors are ceramic and they look like the following. This is a 100n (0.1u)ceramic:

To read the value on a capacitor you need to know a few facts.

The basic value of capacitance is the FARAD.
1 microfarad is one millionth of 1 farad.
1 microfarad is divided into smaller parts called nanofarad.
1,000 nanofarad = 1 microfarad
Nanofarad is divided into small parts called picofarad
1,000 picofarad = 1 nanofarad.

Recapping:
1p = 1 picofarad.   1,000p = 1n (1 nanofarad)
1,000n = 1u (1 microfarad)
1,000u = 1millifarad
1,000,000u = 1 FARAD.

Examples:
All ceramic capacitors are marked in "p" (puff")
A ceramic with 22 is 22p = 22 picofarad
A ceramic with 47 is 47p = 47 picofarad
A ceramic with 470 is 470p = 470 picofarad
A ceramic with 471 is 470p = 470 picofarad
A ceramic with 102 is 1,000p = 1n
A ceramic with 223 is 22,000p = 22n
A ceramic with 104 is 100,000p = 100n = 0.1u

TYPES OF CAPACITOR
For testing purposes, there are two types of capacitor.
Capacitors from 1p to 100n are non-polar and can be inserted into a circuit around either way.
Capacitors from 1u to 100,000u are electrolytics and are polarised. They must be fitted so the positive lead goes to the supply voltage and the negative lead goes to ground (or earth).
There are many different sizes, shapes and types of capacitor. They are all the same. They consist of two plates with an insulating material between. The two plates can be stacked in layers or rolled together.
The important factor is the insulating material. It must be very thin to keep things small. This gives the capacitor its VOLTAGE RATING.
If a capacitor sees a voltage higher than its rating, the voltage will "jump through" the insulating material or around it.
If this happens, a carbon deposit is left behind and the capacitor becomes "leaky" or very low resistance, as carbon is conductive.

CERAMIC CAPACITORS

Nearly all small capacitors are **ceramic capacitors** as this material is cheap and the capacitor can be made in very thin layers to produced a high capacitance for the size of the component. This is especially true for surface-mount capacitors.
All capacitors are marked with a value and the basic unit is: "p" for "puff" However NO surface mount capacitors are marked and they are very difficult to test.

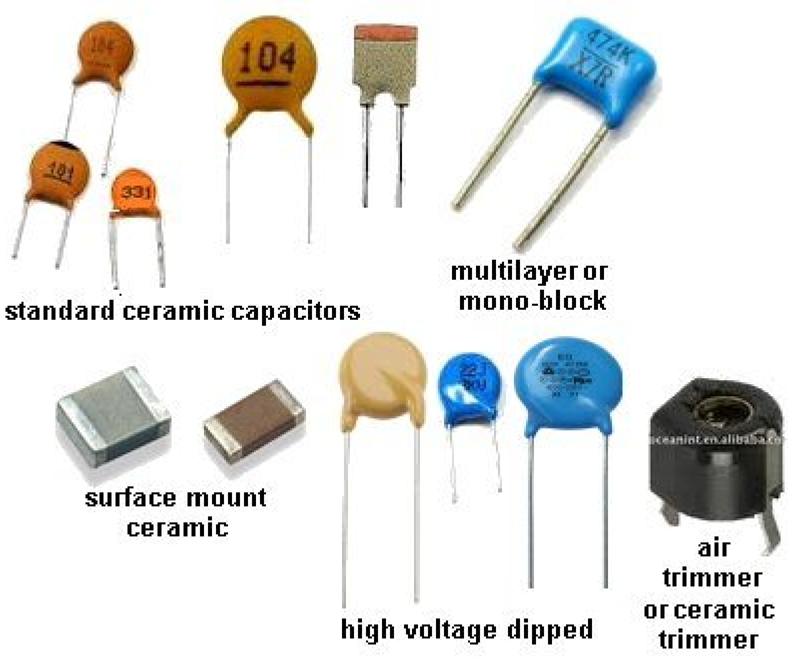
104
104
101
331
474K
X7R
standard ceramic capacitors
multilayer or
mono-block
surface mount
ceramic
high voltage dipped
air
trimmer
or ceramic
trimmer

VALUE:	**VALUE WRITTEN ON THE COMPONENT:**
0.1p	0p1
0.22p	0p22
0.47p	0p47
1.0p	1p0
2.2p	2p2
4.7p	4p7
5.6p	5p6
8.2p	8p2
10p	10 or 10p
22p	22 or 22p
47p	47 or 47p
56p	56 or 56p
100p	100 on 101
220p	220 or 221
470p	470 or 471
560p	560 or 561
820p	820 or 821
1,000p (1n)	102
2200p (2n2)	222
4700p (4n7)	472
8200p (8n2)	822
10n	103
22n	223
47n	473
100n	104
220n	224
470n	474
1u	105

POLYESTER, POLYCARBONATE, POLYSTYRENE, MYLAR, METALLISED POLYESTER, ("POLY"), MICA and other types of CAPACITOR

There are many types of capacitor and they are chosen for their reliability, stability, temperate-range and cost.
For testing and repair work, they are all the same. Simply replace with exactly the same type and value.

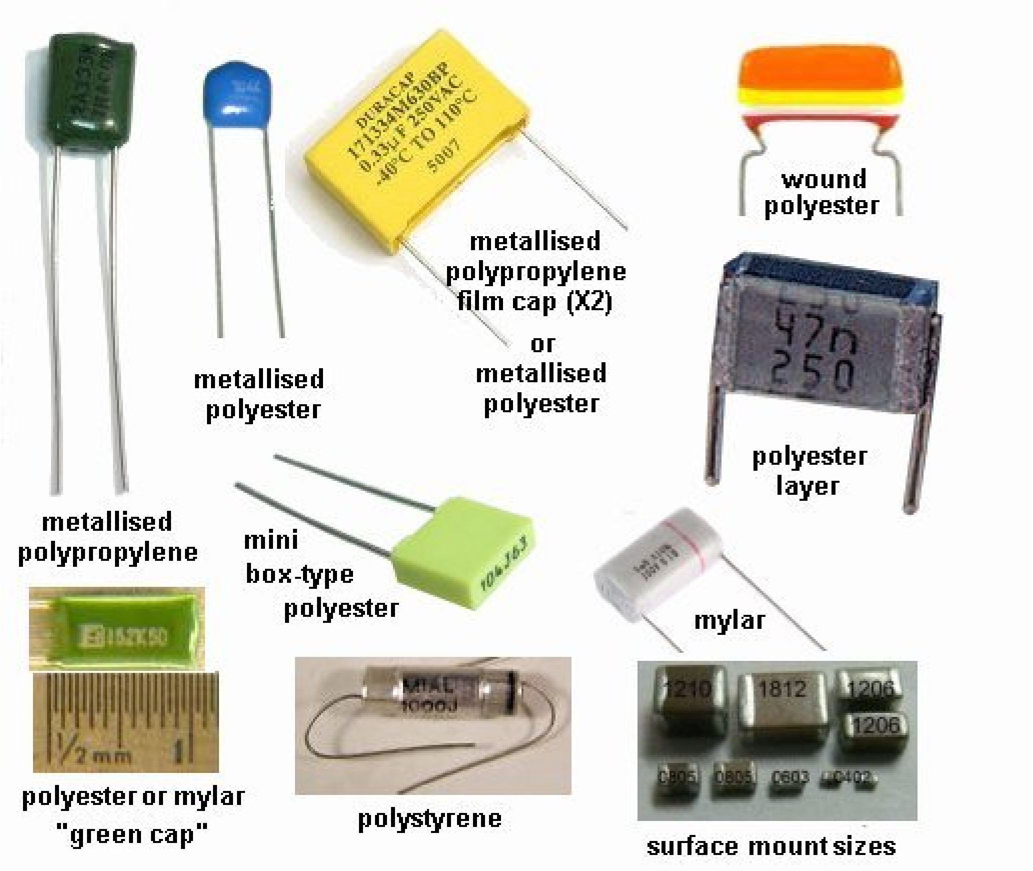
wound polyester
metallised polypropylene film cap (X2)
or
metallised polyester
metallised polyester
polyester layer
metallised polypropylene
mini box-type polyester
mylar
polyester or mylar "green cap"
polystyrene
1210
1812
1206
1206
0805
0805
0603
0402
surface mount sizes

Capacitor Colour Code Table

Colour	Digit A	Digit B	Multiplier D	Tolerance (T) > 10pf	Tolerance (T) < 10pf	Temperature Coefficient (TC)
Black	0	0	x1	± 20%	± 2.0pF	
Brown	1	1	x10	± 1%	± 0.1pF	-33x10^{-6}
Red	2	2	x100	± 2%	± 0.25pF	-75x10^{-6}
Orange	3	3	x1,000	± 3%		-150x10^{-6}
Yellow	4	4	x10,000	± 4%		-220x10^{-6}
Green	5	5	x100,000	± 5%	± 0.5pF	-330x10^{-6}
Blue	6	6	x1,000,000			-470x10^{-6}
Violet	7	7				-750x10^{-6}
Grey	8	8	x0.01	+80%,-20%		
White	9	9	x0.1	± 10%	± 1.0pF	
Gold			x0.1	± 5%		

Silver			x0.01	± 10%		

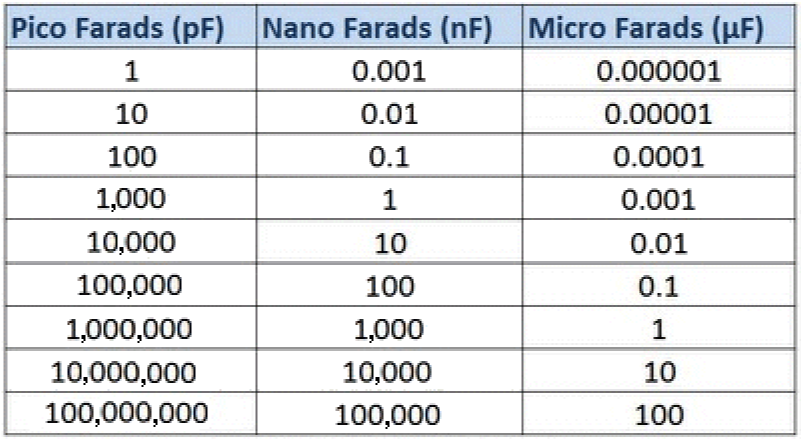

Pico Farads (pF)	Nano Farads (nF)	Micro Farads (µF)
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1
10,000,000	10,000	10
100,000,000	100,000	100

Type ⊕ = polarized	Pic	Cap Range
Ceramic		pF - µF
Mica (silver mica)		pF - nF
Plastic Film (polyethylene polystyrene)	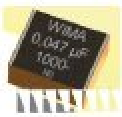	few µFs
Tantalum ⊕		µFs
OSCON ⊕		µFs
Aluminum Electrolytic ⊕		high µFs

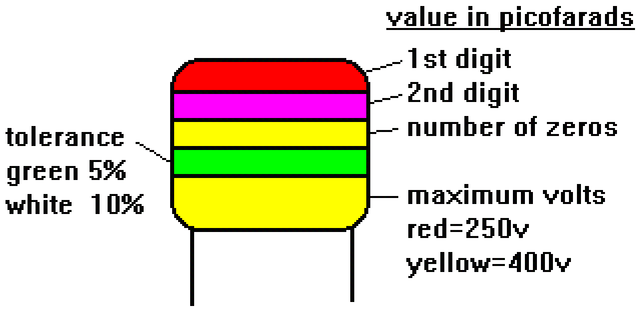
value in picofarads
1st digit
2nd digit
number of zeros
maximum volts
red=250v
yellow=400v
tolerance
green 5%
white 10%

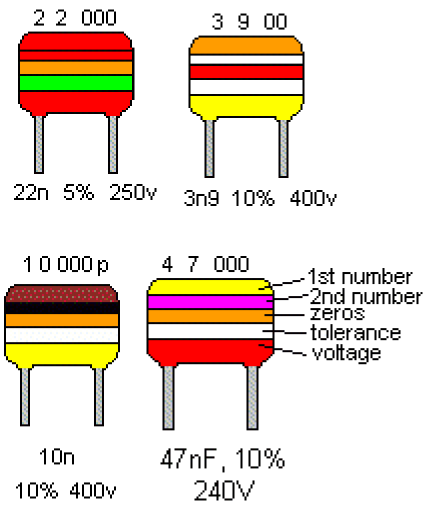
2 2 000
22n 5% 250v
3 9 00
3n9 10% 400v
1 0 000 p
10n
10% 400v
4 7 000
1st number
2nd number
zeros
tolerance
voltage
47nF, 10%
240V

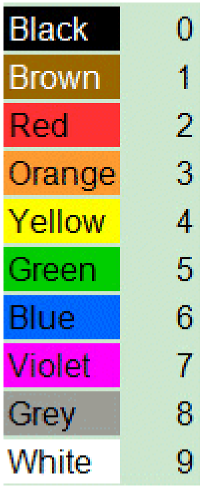

Black	0
Brown	1
Red	2
Orange	3
Yellow	4
Green	5
Blue	6
Violet	7
Grey	8
White	9

ELECTROLYTIC and TANTALUM CAPACITORS

Electrolytics and Tantalums are the same for testing purposes but their performance is slightly different in some circuits. A tantalum is smaller for the same rating as an electrolytic and has a better ability at delivering a current. They are available up to about 1,000u, at about 50v but their cost is much higher than an electrolytic.

Electrolytics are available in 1u, 2u2 3u3 4u7 10u, 22u, 47u, 100u, 220u, 330u, 470u, 1,000u, 2,200u, 3,300u, 4,700u, 10,000u and higher.
The "voltage" or "working voltage" can be: 3.3v, 10v, 16v, 25v, 63v, 100v, 200v and higher.
There is also another important factor that is rarely covered in text books. It is RIPPLE FACTOR.
This is the amount of current that can enter and leave an electrolytic. This current heats up the electrolytic and that is why some electrolytics are much larger than others, even though the capacitance and voltage-ratings are the same.
If you replace an electrolytic with a "miniature" version, it will heat up and have a very short life. This is especially important in power supplies where current (energy) is constantly entering and exiting the electrolytic as its main purpose is to provide a smooth output from a set of diodes that delivers "pulsing DC." (see "Power Diodes")

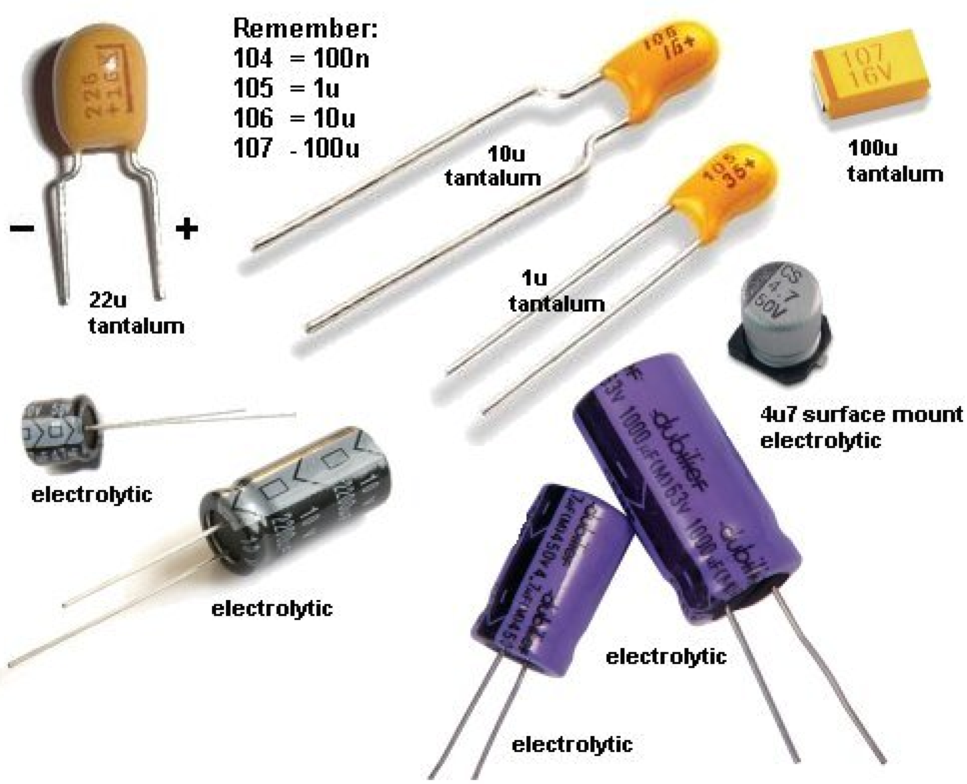

Remember:
104 = 100n
105 = 1u
106 = 10u
107 - 100u

−
+
22u tantalum
10u tantalum
1u tantalum
100u tantalum
4u7 surface mount electrolytic
electrolytic
electrolytic
electrolytic
electrolytic

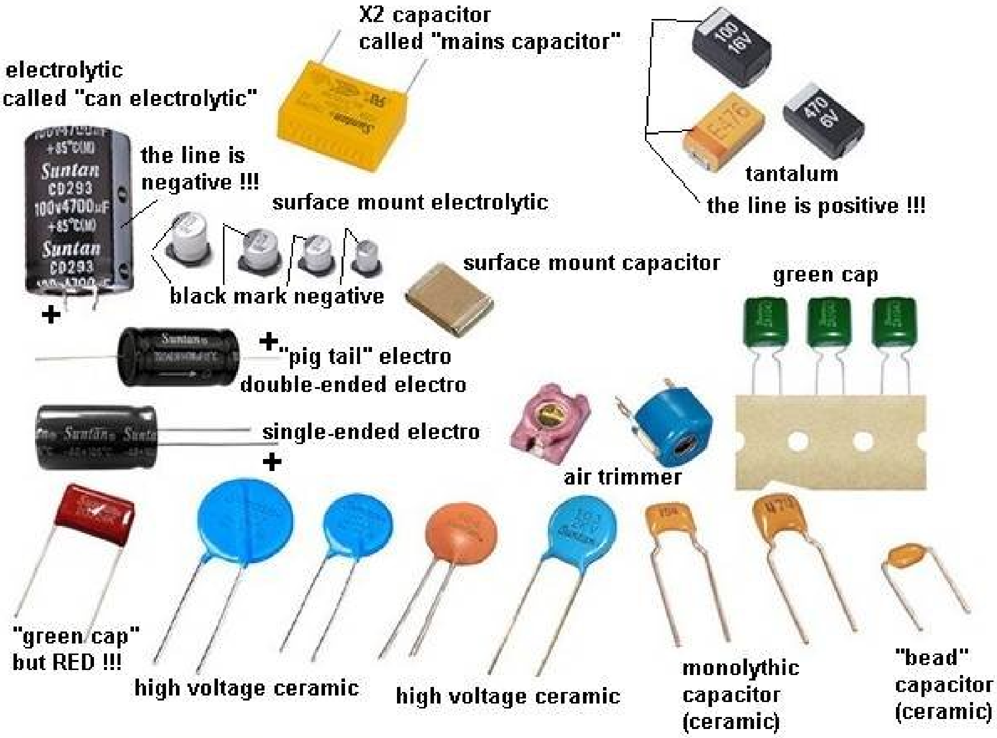

X2 capacitor called "mains capacitor"
electrolytic called "can electrolytic"
the line is negative !!!
surface mount electrolytic
tantalum
the line is positive !!!
black mark negative
surface mount capacitor
green cap
+
+
"pig tail" electro double-ended electro
single-ended electro
+
air trimmer
"green cap" but RED !!!
high voltage ceramic
high voltage ceramic
monolythic capacitor (ceramic)
"bead" capacitor (ceramic)

tantalum capacitors can "go up in smoke"

"bead" tantalum

PARALLEL and SERIES CAPACITORS

Capacitors can be connected in PARALLEL and/or SERIES for a number of reasons.
1. If you do not have the exact value, two or more connected in parallel or series can produce the value you need.
2. Capacitors connected in series will produce one with a higher voltage rating.
3. Capacitors connected in parallel will produce a larger-value capacitance.

Here are examples of two equal capacitors connected in series or parallel and the results they produce:

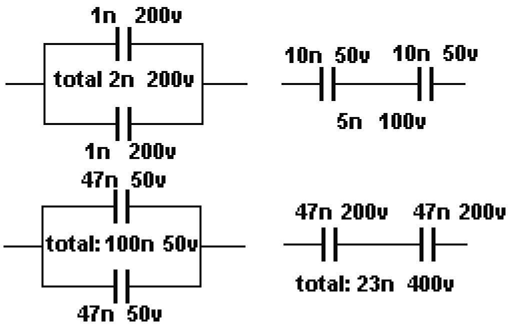
1n 200v
total 2n 200v
1n 200v
10n 50v
10n 50v
5n 100v
47n 50v
total: 100n 50v
47n 50v
47n 200v
47n 200v
total: 23n 400v

NON-POLAR CAPACITORS (ELECTROLYTICS)

Electrolytics are also available in non-polar values. It sometimes has the letters "NP" on the component. Sometimes the leads are not identified.
This is an electrolytic that does not have a positive and negative lead but two leads and either lead can be connected to the positive or negative of the circuit.
These electrolytics are usually connected to the output of an amplifier (such as in a filter near the speaker) where the signal is rising and falling.
A non-polar electrolytic can be created from two ordinary electrolytics by connecting the negative leads together and the two positive leads become the new leads.
For example: two 100u 63v electrolytics will produce a 47u 63v non-polar electrolytic.
In the circuit below, the non-polar capacitor is replaced with two electrolytics.

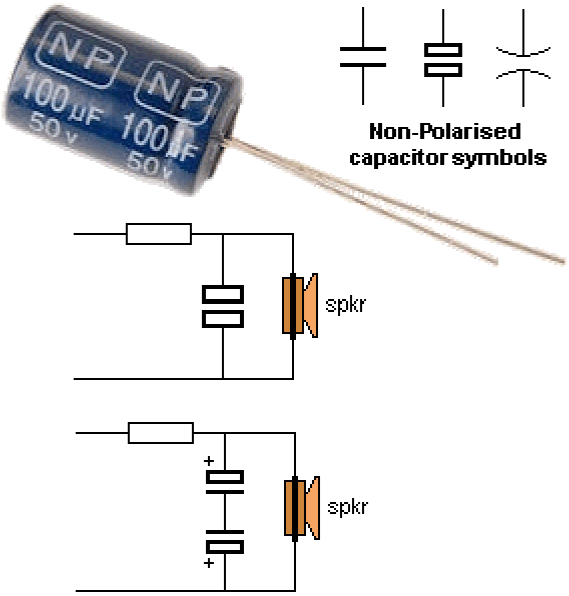
NP
100µF
50v
NP
100µF
50v
Non-Polarised
capacitor symbols
spkr
+
+
spkr

MAKING A NON-POLAR ELECTROLYTIC

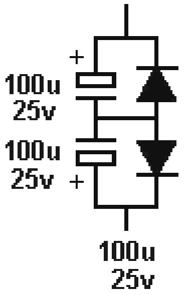

non-polar

A normal electrolytic must be connected the correct way in a circuit because it has a thin insulating layer covering the plates that has a high resistance.
If you connect the electrolytic around the wrong way, this layer "breaks-down" and the resistance of the electrolytic becomes very small and a high current flows. This heats up the electrolytic and the current increases. Very soon the capacitor produces gasses and explodes.
One big mistake in many text books shows how to make a non-polar electrolytic by connecting two "back-to-back."
They claim 2 x 100u connected back-to-back is equal to 47u.
This appears to be case when testing on a meter but the meter simply charges them for a short period of time to get a reading.
If you allow them to charge fully you will find the reverse electrolytic has a very small voltage across it.
Secondly, when you are charging them, you are putting a high current through the reverse electrolytic and damaging the layer.
To prevent this, you need to add two diodes as shown in the diagram.
In addition, 2 x 100u "back-to-back" is very near 100u.

Here is a question from a reader:
I have an amplifier with 2 x2,200u electrolytics on the output of a bridge. Can I replace them with a single 10,000u?

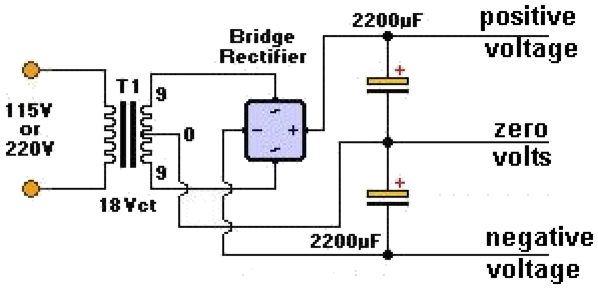

You need to look at the circuit of your amplifier. The two 2,200u electrolytics are possibly connected as shown in the circuit above and you will notice they are joined to produce a positive rail and a negative rail with zero (called earth) in the centre.
This forms two different circuits with the top electrolytic filtering the positive rail and the bottom electro filtering the negative rail. They must be connected to the zero volts rail.
A single 10,000u cannot be connected to the 0v rail and cannot be substituted for the two electro's.
You can easily determine of the two electro's are connected as shown above.
Test the positive terminal of each electro by placing the negative of the meter on the chassis.
If the positive of one electro have zero volts, it will be the lower electro in the diagram above. The negative terminal of the electro will have a minus voltage on it.

VOLTAGE RATING OF CAPACITOR
Capacitors have a voltage rating, stated as WV for working voltage, or WVDC. This specifies the maximum voltage that can be applied across the capacitor without puncturing the dielectric. Voltage ratings for "poly," mica and ceramic capacitors are typically 50v to 500 VDC. Ceramic capacitors with ratings of 1kv to 5kv are also available. Electrolytic capacitors are commonly available in 6v, 10v 16v, 25v, 50v, 100v, 150v, and 450v ratings.

THE SIZE OF A CAPACITOR - RIPPLE FACTOR

The size of a capacitor depends on a number of factors, namely the value of the capacitor (in microfarads etc) and the voltage rating. But there is also another factor that is most important. It is the RIPPLE FACTOR. **Ripple Factor** is the amount of voltage-fluctuation the capacitor (electrolytic) can withstand without getting too hot. When current flows in and out of an electrolytic, it gets hot and this will eventually dry-out the capacitor as some of the liquid inside the capacitor escapes through the seal. It's a very slow process but over a period of years, the capacitor looses its capacitance.

If you have two identical 1,000u 35v electrolytics and one is smaller, it will get hotter when operating in a circuit and that's why it is necessary to choose the largest electrolytic.

CAUTION

If a capacitor has a voltage rating of 63v, do not put it in a 100v circuit as the insulation (called the dielectric) will be punctured and the capacitor will "short-circuit." It's ok to replace a 0.22uF 50WV capacitor with 0.22uF 250WVDC.

SAFETY

A capacitor can store a charge for a period of time after the equipment is turned off. High voltage electrolytic caps can pose a safety hazard. These capacitors are in power supplies and some have a resistor across them, called a bleed resistor, to discharge the cap after power is switched off.

If a bleed resistor is not present the cap can retain a charge after the equipment is unplugged.

How to discharge a capacitor

Do not use a screwdriver to short between the terminals as this will damage the capacitor internally and the screwdriver.

Use a 1k 1 watt or 3watt or 5watt resistor on jumper leads (or held with pliers) and keep them connected for up to 15 seconds to fully discharge the electro.

Test it with a voltmeter to make sure all the energy has been removed.

Before testing any capacitors, especially electrolytics, you should look to see if any are damaged, overheated or leaking. Swelling at the top of an electrolytic indicates heating (and pressure inside the case) and will result in drying out of the electrolyte. Any hot or warm electrolytic indicates leakage and ceramic capacitors with portions missing indicates something has gone wrong (such as it being "blown apart").

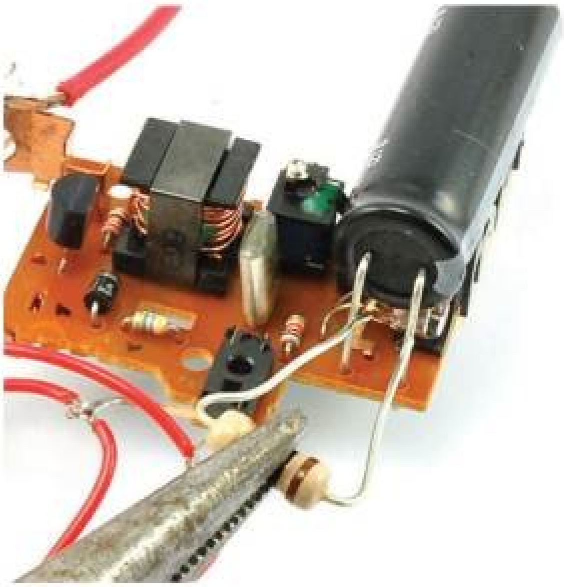

Here is a 120u 330v electrolytic from a flash circuit in an old-fashioned film camera. If the flash does not "fire," the electrolytic will be charged to about 350 volts !!

Use a 1k resistor (held with pliers) to slowly discharge it. It may take 15 seconds to fully discharge.

TESTING A CAPACITOR

There are two things you can test with a multimeter:

1. A short-circuit within the capacitor
2. Capacitor values above 1u.

You can test capacitors in-circuit for short-circuits. Use the x1 ohms range.
To test a capacitor for leakage, you need to remove it or at least one lead must be removed. Use the x10k range on an analogue or digital multimeter.
For values above 1u you can determine if the capacitor is charging by using an analogue meter. The needle will initially move across the scale to indicate the cap is charging, then go to "no deflection." Any permanent deflection of the needle will indicate leakage.
You can reverse the probes to see if the needle moves in the opposite direction. This indicates it has been charged. Values below 1u will not respond to charging and the needle will not deflect.
This does not work with a digital meter as the resistance range does not output any current and the electrolytic does not charge.

Rather than spending money on a capacitance meter, it is cheaper to replace any suspect capacitor or electrolytic.
Capacitors can produce very unusual faults and no piece of test equipment is going to detect the problem.
In most cases, it is a simple matter to solder another capacitor across the suspect component and view or listen to the result.
This saves all the worry of removing the component and testing it with equipment that cannot possibly give you an accurate reading when the full voltage and current is not present.
It is complete madness to even think of testing critical components such as capacitors, with TEST EQUIPMENT. You are fooling yourself. If the Test Equipment says the component is ok, you will look somewhere else and waste a lot of time.

FINDING THE VALUE OF A CAPACITOR

If you want to find the value of a surface-mount capacitor or one where the markings have been removed, you will need a CAPACITANCE METER. Here is a simple circuit that can be added to your meter to read capacitor values from 10p to 10u.
The full article can be found HERE.

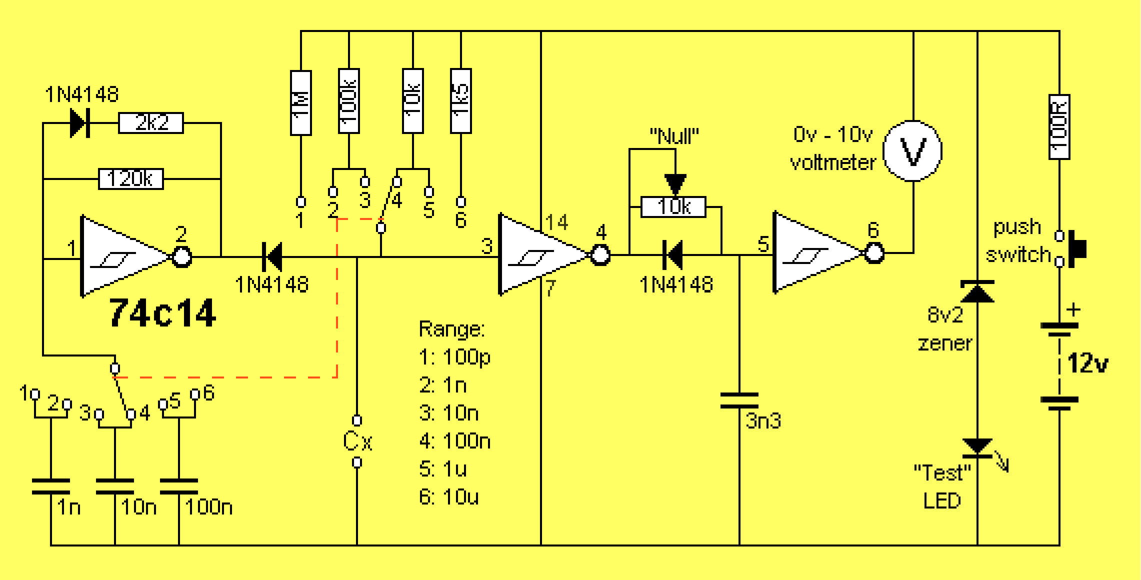

ADD-ON CAPACITANCE METER

CAPACITOR SUBSTITUTION BOX

CAPACITANCE SUBSTITUTION BOX
470
560
680
820
330
220
150
100
3300
2200
1500
1000
pF
µF
.01
.015
.022
.033
.047
.056
.068
.1
.0047
.0056
.0068
.0082
ELENCO® MODEL CS-440

You can get a kit or a ready-made piece of test gear called **CAPACITOR SUBSTITUTION BOX** and also **RESISTOR SUBSTITUTION BOX**.
I bought one of each 30 years ago and I have only used them ONCE.
They appear to be very handy but when you are testing a circuit, you want the component next to the other parts.
It is just as easy to pick the component you need from your junk box and connect it to the circuit via jumper leads.

REPLACING A CAPACITOR

Always replace a capacitor with the exact same type.
A capacitor may be slightly important in a circuit or it might be extremely critical.
A manufacturer may have taken years to select the right type of capacitor due to previous failures.
A capacitor just doesn't have a "value of capacitance."
It may also has an effect called "tightening of the rails."
In other words, a capacitor has the ability to react quickly and either absorb or deliver energy to prevent spikes or fluctuations on the rail.
This is due to the way it is constructed. Some capacitors are simply plates of metal film while others are wound in a coil. Some capacitors are large while others are small.
They all react differently when the voltage fluctuates.
Not only this, but some capacitors are very stable and all these features go into the decision for the type of capacitor to use.
You can completely destroy the operation of a circuit by selecting the wrong type of capacitor.
No capacitor is perfect and when it gets charged or discharged, it appears to have a small value of resistance in series with the value of capacitance. This is known as "ESR" and stands for EQUIVALENT SERIES RESISTANCE. This effectively makes the capacitor slightly slower to charge and discharge.
We cannot go into the theory on selecting a capacitor as it would be larger than this eBook so the only solution is to replace a capacitor with an identical type.
However if you get more than one repair with identical faults, you should ask other technicians if the original capacitor comes from a faulty batch.
The author has fixed TV's and fax machines where the capacitors have been inferior and alternate types have solved the problem.
Some capacitor are suitable for high frequencies, others for low frequencies.

DECOUPLING CAPACITORS

A Decoupling Capacitor can severe one, two or three functions. You need to think of a decoupling capacitor as a miniature battery with the ability to deliver a brief pulse of energy when ever the line-voltage drops and also absorb a brief pulse of energy when ever the line voltage rises (or spikes).
Decoupling capacitor can range from 100n to 1,000u.
100n capacitors are designed to absorb spikes and also have the effect of tightening-

up the rails for high frequencies. They have no effect on low frequencies such as audio frequencies.

These capacitors are generally ceramic and have very low internal impedance and thus they can operate at high frequencies.

Capacitors above about 10u are used for decoupling and these are nearly always electrolytics.

Decoupling means "tightening-up the power rails." The electrolytic acts just like a miniature rechargeable battery, supplying a small number of components in a circuit with a smooth and stable voltage.

The electrolytic is usually fed from a dropper resistor and this resistor charges the electrolytic and adds to the ability of the electrolytic to create a "separate power supply."

These two components help remove spikes as an electrolytic cannot remove spikes if connected directly to the supply rails - it's internal impedance is high and the spikes are not absorbed.

Decoupling capacitors are very difficult to test.

They rarely fail but if a project is suffering from unknown glitches and spikes, it is best to simply add more 100n decoupling caps on the underside of the board and replace all electrolytics.

Some small electrolytics will dry out due to faulty manufacture and simply replacing every one on a board will solve the problem.

Some of the functions of a decoupling capacitor are:

Removing ripple - hum or buzz in the background of an amplifier

Removing glitches or spikes.

Separating one stage from another to reduce or remove MOTORBOATING - a low frequency sound due to the output putting a pulse on the power rails that is picked up by the pre-amplifier section and amplified.

TESTING DIODES

Diodes can have 4 different faults.

1. Open circuit in both directions.
2. Low resistance in both directions.
3. Leaky.
4. Breakdown under load.

TESTING A DIODE ON AN ANALOGUE METER

Testing a diode with an **Analogue Multimeter** can be done on any of the resistance ranges. [The high resistance range is best - it sometimes has a high voltage battery for this range but this does not affect our testing]

There are two things you must remember.

1. When the diode is measured in one direction, the needle will **not move at all**. The technical term for this is the diode is **reverse biased**. It will not allow any current to flow. Thus the needle will not move.

When the diode is connected around the other way, the needle will swing to the right (move up scale) to about 80% of the scale. This position represents the voltage drop across the junction of the diode and is NOT a resistance value. If you change the resistance range, the needle will move to a slightly different position due to the resistances inside the meter. The technical term for this is the diode is **forward biased**. This indicates the diode is not faulty.

The needle will swing to a slightly different position for a "normal diode" compared to a Schottky diode. This is due to the different junction voltage drops.

However we are only testing the diode at very low voltage and it may break-down when fitted to a circuit due to a higher voltage being present or due to a high current flowing.

2. The leads of an **Analogue Multimeter** have the positive of the battery connected to the black probe and the readings of a "good diode" are shown in the following two diagrams:

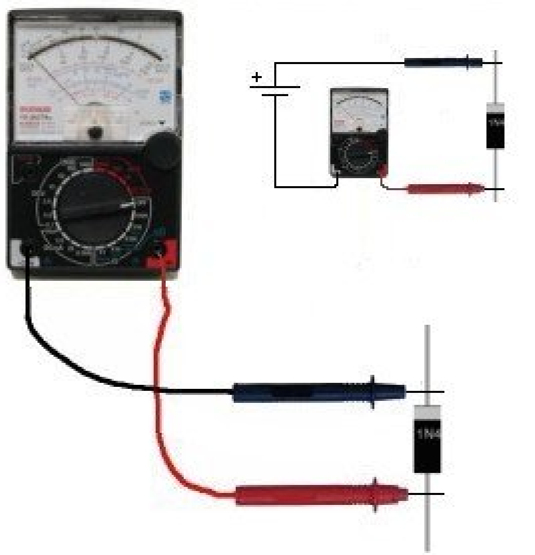

The diode is REVERSE BIASED in the diagram above and diodes not conduct.

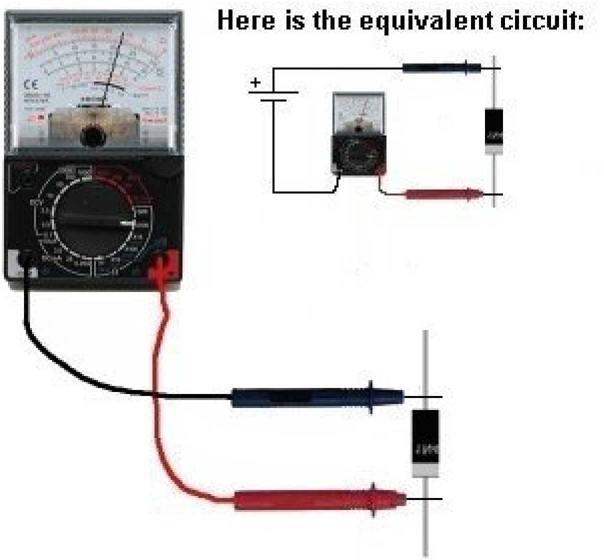

The diode is FORWARD BIASED in the diagram above and it conducts

TESTING A DIODE ON A DIGITAL METER

Testing a diode with a Digital Meter must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.

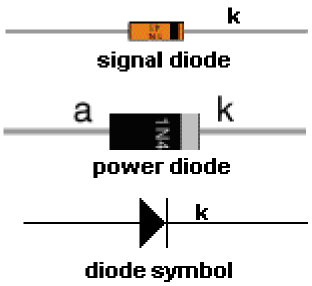

The best thing to do with a "suspect" diode is to replace it. This is because a diode has a number of characteristics that cannot be tested with simple equipment. Some diodes have a fast recovery for use in high frequency circuits. They conduct very quickly and turn off very quickly so the waveform is processed accurately and

efficiently.
If the diode is replaced with an ordinary diode, it will heat up as does not have the high-speed characteristic.
Other diodes have a low drop across them and if an ordinary is used, it will heat up.
Most diodes fail by going: SHORT-CIRCUIT. This can be detected by a low resistance (x1 or x10 Ohms range) in both directions.
A diode can also go OPEN CIRCUIT. To locate this fault, place an identical diode across the diode being tested.
A leaky diode can be detected by a low reading in one direction and a slight reading the other direction.
However this type of fault can only be detected when the circuit is working. The output of the circuit will be low and sometimes the diode heats up (more than normal).
A diode can go open under full load conditions and perform intermittently.
Diodes come in pairs in surface-mount packages and 4 diodes can be found in a bridge.
They are also available in pairs that look like a 3-leaded transistor.
The line on the end of the body of a diode indicates the cathode and you cannot say "this is the positive lead." The correct way to describe the leads is to say the "cathode lead." The other lead is the anode. The cathode is defined as the electrode (or lead) through which an electric current flows out of a device.
The following diagrams show different types of diodes:

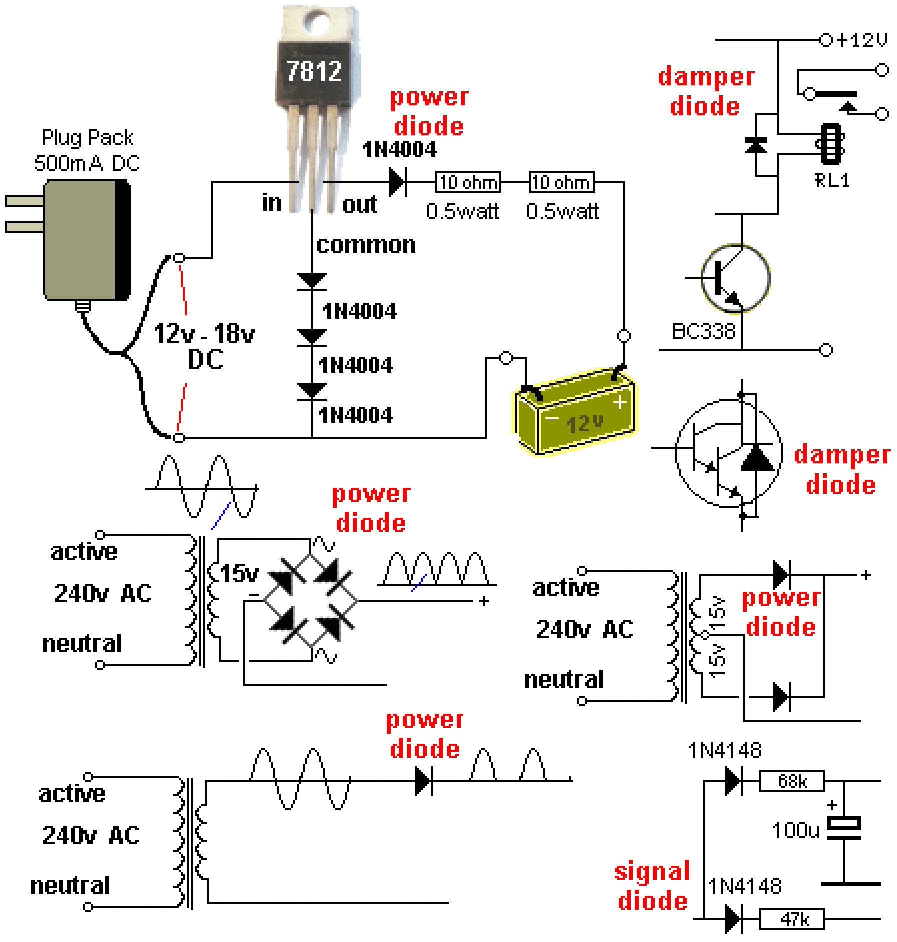
7812
power diode
1N4004
Plug Pack
500mA DC
in
out
common
10 ohm
0.5watt
10 ohm
0.5watt
12v - 18v
DC
1N4004
1N4004
1N4004
12v
+
−
damper diode
+12V
RL1
BC338
damper diode
power diode
active
240v AC
neutral
15v
+
active
240v AC
neutral
15v
15v
power diode
+
power diode
active
240v AC
neutral
1N4148
68k
+
100u
signal diode
1N4148
47k

POWER DIODES

To understand how a power diode works, we need to describe a few things. This has NEVER been described before, so read carefully.
The 240v AC (called the "mains") consists of two wires, one is called the ACTIVE and the other is NEUTRAL. Suppose you touch both wires. You will get a shock. The neutral is connected to an earth wire (or rod driven into the ground or connected to a water pipe) at the point where the electricity enters the premises and you do not get a shock from the NEUTRAL.

But the voltage on the active is rising to +345v then goes to -345v at the rate of 50 times per second (for a complete cycle).
345v is the peak voltage of 240v. You never get a 240v shock. (It is a 345v shock.)
In other words, if you touch the two wires at a particular instant, you would get a POSITIVE 345v shock and at another instant you would get a negative 345v shock.
This is shown in the diagram below.
We now transfer this concept to the output of a transformer. The diagram shows an AC waveform on the output of the secondary.
This voltage is rising 15v higher than the bottom lead then it is 15v LOWER than the bottom lead. The bottom lead is called "zero volts." You have to say one lead or wire is not "rising and falling" as you need a "reference" or starting-point" or "zero point" for voltage measurements.
The diode only conducts when the voltage is "above zero" (actually when it is 0.7v above zero) and does not conduct (at all) when the voltage goes below zero.
This is shown on the output of the Power Diode. Only the positive peaks or the positive parts of the waveform appear on the output and this is called "pulsing DC."
This is called "half-wave" and is not used in a power supply. We have used it to describe how the diode works. The electrolytics charge during the peaks and deliver energy when the diode is not delivering current. This is how the output becomes a steady DC voltage.
Power supplies use FULL WAVE rectification and the other half of the AC waveform is delivered to the output (and fills in the "gaps") and appears as shown in "A."

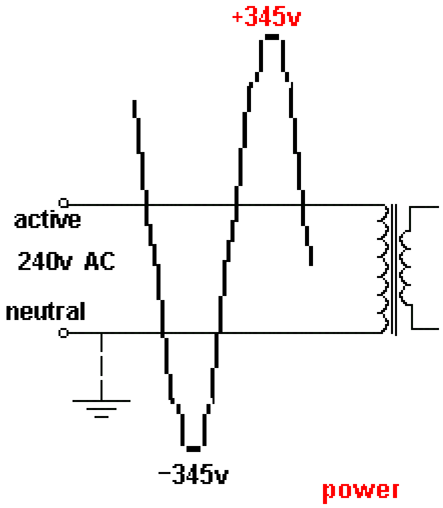

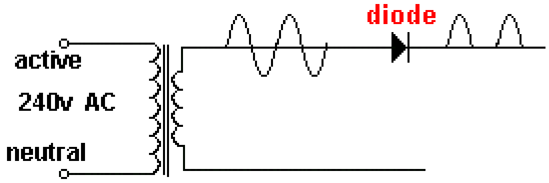

Half-wave rectifier

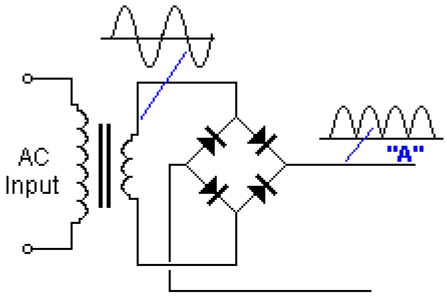

Full-wave rectifier

ONE FAULTY DIODE

One diode in a bridge can go open (any of the 4 diodes will produce the same effect) and produce an output voltage that can be slightly lower than the original voltage.

The actual "voltage-drop" will depend on the current taken by the circuit and the ability of the transformer to produce the required voltage and current during half-wave operation. The voltage during each half cycle (when none of the diodes is delivering any energy to the circuit) is maintained by the electrolytic and its size (relative to the current taken by the circuit) will determine the size of the ripple that will result when the diode fails. The ripple will be 100 to 1,000 times greater after the failure of a diode, depending on the value of the filter capacitor.
To locate the faulty diode, simply get a diode and place it across each of the diodes in the bridge (in turn) when the circuit is working.
For a bridge rectifier, the ripple-frequency will be twice the mains frequency and its ripple will be very small if the electrolytic is the correct value. When a diode fails, the ripple-frequency will be equal to mains-frequency and the amplitude will increase considerably. You may even hear background hum from audio equipment.
If you cannot find a faulty diode, the filter capacitor will be at fault. Turn off the equipment and connect an electrolytic across the filter capacitor via jumper leads. Turn the power ON and see if the hum has reduced.

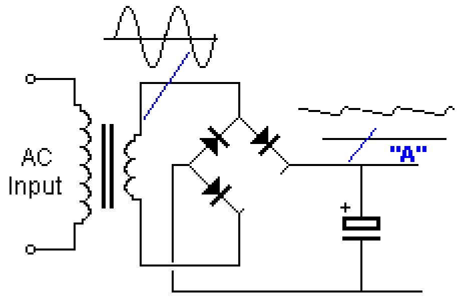

One diode "burnt-out"

DAMPER DIODES

A damper diode is a diode that detects a high voltage and SQUELCHES IT (reduces it - removes it). The signal that it squelches is a voltage that is in the opposite direction to the "supply voltage" and is produced by the collapsing of a magnetic field. Whenever a magnetic filed collapses, it produces a voltage in the winding that is opposite to the supply voltage and can be much higher. This is the principle of a flyback circuit or EHT circuit. The high voltage comes from the transformer.
The diode is placed so that the signal passes through it and less than 0.5v appears across it.
A damper diode can be placed across the coil of a relay, incorporated into a transistor or FET or placed across a winding of a flyback transformer to protect the driving transistor or FET.
It can also be called a "Reverse-Voltage Protection Diode," "Spike Suppression Diode," or "Voltage Clamp Diode."
The main characteristic of a Damper Diode is HIGH SPEED so it can detect the spike and absorb the energy.
It does not have to be a high-voltage diode as the high voltage in the circuit is being absorbed by the diode.

SILICON, GERMANIUM AND SCHOTTKY DIODES

When testing a diode with an analogue meter, you will get a low reading in one direction and a high (or NO READING) in the other direction. When reading in the LOW direction, the needle will swing nearly full scale and the reading is not a resistance-value but a reflection of the characteristic voltage drop across the junction of the diode. As we mentioned before, a resistance reading is really a voltage reading and the meter is measuring the voltage of the battery minus the voltage-drop across the diode.
Since Silicon, Germanium and Schottky Diodes have slightly different characteristic voltage drops across the junction, you will get a slightly different reading on the scale. This does not represent one diode being better than the other or capable of handling a higher current or any other feature.
The quickest, easiest and cheapest way to find, fix and solve a problem caused by a

faulty diode is to replace it.
There is no piece of test equipment capable of testing a diode fully, and the circuit you are working on is actually the best piece of test equipment as it is identifying the fault UNDER LOAD.

Only very simple tests can be done with a multimeter and it is best to check a diode with an ANALOGUE MULTIMETER as it outputs a higher current though the diode and produces a more-reliable result.
A Digital meter can produce false readings as it does not apply enough current to activate the junction.
Fortunately almost every digital multimeter has a **diode test mode.** Using this, a silicon diode should read a voltage drop between 0.5v to 0.8v in the forward direction and open in the reverse direction. For a germanium diode, the reading will be lower, around 0.2v - 0.4v in the forward direction. A bad diode will read zero volts in both directions.

REPLACING A DIODE

It is alway best to replace a diode with the same type but quite often this is not possible. Many diodes have unusual markings or colours or "in-house" letters.
This is only a general guide because many diodes have special features, especially when used in high-frequency circuits.
However if you are desperate to get a piece of equipment working, here are the steps:
Determine if the diode is a signal diode, power diode, or zener diode.
For a signal diode, try 1N4148.
For a power diode (1 amp) try 1N4004. (for up to 400v)
For a power diode (3 amp) try 1N5404. (for up to 400v)
For a high-speed diode, try UF4004 (for up to 400v)

If you put an ordinary diode in a high-speed application, it will get very hot very quickly.
To replace an unknown zener diode, start with a low voltage such as 6v2 and see if the circuit works.
The size of a diode and the thickness of the leads will give an idea of the current-capability of the diode.
Keep the leads short as the PC board acts as a heat-sink.
You can also add fins to the leads to keep the diode cool.

LIGHT EMITTING DIODES (LEDs)

Light Emitting Diodes (LEDs) are diodes that produce light when current flows from anode to cathode. The LED does not emit light when it is revered-biased. It is used as a low current indicator in many types of consumer and industrial equipment, such as monitors, TV’s, printers, hi-fi systems, machinery and control panels.
The light produced by a LED can be visible, such as red, green, yellow or white. It can also be invisible and these LEDs are called Infrared LEDs. They are used in remote controls and to see if they are working, you need to point a digital camera at the LED and view the picture on the camera screen.
An LED needs about 2v - 3.6v across its leads to make it emit light, but this voltage must be exact for the type and colour of the LED. The simplest way to deliver the exact voltage is to have a supply that is higher than needed and include a voltage-dropping resistor. The value of the resistor must be selected so the current is between 2mA and 25mA.
The cathode of the LED is identified by a flat on the side of the LED. The life expectancy of a LED is about 100,000 hours. LEDs rarely fail but they are very sensitive to heat and they must be soldered and de-soldered quickly. They are one of the most heat-sensitive components.
Light emitting diodes cannot be tested with most multimeters because the characteristic voltage across them is higher than the voltage of the battery in the meter.
However a simple tester can be made by joining 3 cells together with a 220R resistor

and 2 alligator clips:

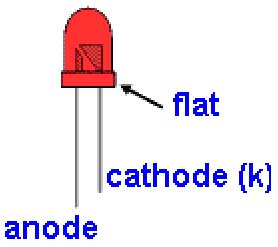

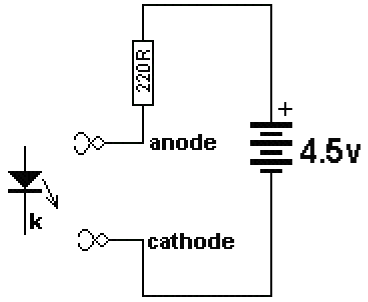

LED TESTER

Connect the clips to a LED and it will illuminate in only one direction.
The colour of the LED will determine the voltage across it. You can measure this voltage if you want to match two or more LEDs for identical operation.
Red LEDs are generally 1.7v to 1.9v. - depending on the quality such as "high-bright"
Green LEDs are 1.9v to 2.3v.
Orange LEDs are about 2.3v and
White LEDs and IR LEDs are about 3.3v to 3.6v.
The illumination produced by a LED is determined by the quality of the crystal. It is the crystal that produces the colour and you need to replace a LED with the same quality to achieve the same illumination.
Never connect a LED across a battery (such as 6v or 9v), as it will be instantly damaged. You must have a resistor in series with the LED to limit the current.

ZENER DIODES

All diodes are Zener diodes. For instance a 1N4148 is a 120v zener diode as this is its reverse breakdown voltage.
And a zener diode can be used as an ordinary diode in a circuit with a voltage that is below the zener value.
For instance, 20v zener diodes can be used in a 12v power supply as the voltage never reaches 20v, and the zener characteristic is never reached.
Most diodes have a reverse breakdown voltage above 100v, while most zeners are below 70v. A 24v zener can be created by using two 12v zeners in series and a normal diode has a characteristic voltage of 0.7v. This can be used to increase the voltage of a zener diode by 0.7v. See the diagram above. It uses 3 ordinary diodes to increase the output voltage of a 3-terminal regulator by 2.1v.
To tests a zener diode you need a power supply about 10v higher than the zener of the diode. Connect the zener across the supply with a 1k to 4k7 resistor and measure the voltage across the diode. If it measures less than 1v, reverse the zener. If the reading is high or low in both directions, the zener is damaged.

Here is a zener diode tester. The circuit will test up to 56v zeners.

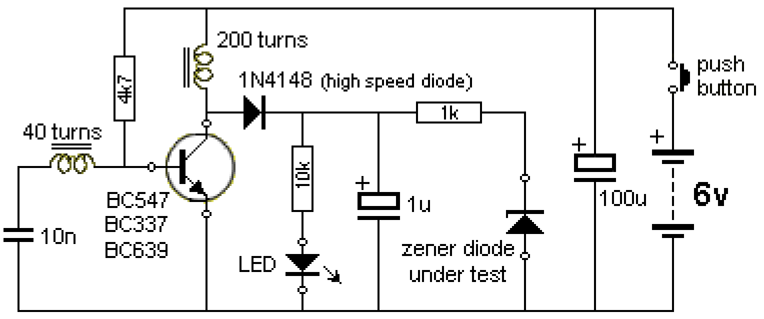

ZENER DIODE TESTER

TRANSFORMERLESS POWER SUPPLY

Here's a circuit that uses zener diodes in a power supply to show how they work. This clever design uses 4 diodes in a bridge to produce a fixed voltage power supply capable of supplying 35mA.

If we put 2 zener diodes in a bridge with two ordinary power diodes, the bridge will break-down at the voltage of the zener. This is what we have done. If we use 18v zeners, the output will be 17v4.

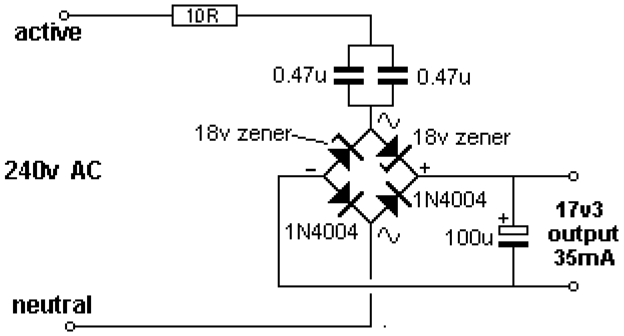

SUPPLY USING ZENER DIODES

When the incoming voltage is positive at the top, the left zener provides 18v limit (and the other zener produces a drop of 0.6v). This allows the right zener to pass current just like a normal diode. The output is 17v4. The same with the other half-cycle.

You cannot use this type of bridge in a normal power supply as the zener diode will "short" when the input voltage reaches the zener value. The concept only works in the circuit above.

VOLTAGE REGULATORS

A Voltage Regulator takes a high input voltage and delivers a fixed output voltage. Providing the input voltage is 4v above the output voltage, the regulator will deliver a fixed output voltage with almost no ripple.

Voltage regulators are also called "3-TERMINAL REGULATORS" or "REGULATOR IC's" - although this name is not generally used.

In most cases, a voltage regulator gets quite hot and for this reason it has a high failure-rate.

If a regulator is not getting hot (or warm) it has either failed or the circuit is not operating.

A regulator can only decrease the voltage. It cannot increase the current. This means the current being supplied to a circuit must also be available from the circuit supplying the regulator.

All regulators have different pin-outs, so you need to find the input pin and output pin and make sure the voltage-difference is at least 4v. Some regulators will work with a difference as low as 1v, so you need to read the specifications for the type you are servicing.

Some regulators are called "negative voltage regulators" and the input voltage will be negative and the output will be negative.

You need to test a voltage regulator with the power "ON".

Make sure you do not allow the probes to short any of the pins together as this will destroy the regulator or the circuit being supplied.

With the power turned off or the regulator removed from the circuit, you can test it with a multimeter set to resistance to see if it is ok. If any resistance readings are very low or zero ohms, the regulator is damaged.

TRANSFORMERS

All transformers and coils are tested the same way. This includes chokes, coils,

inductors, yokes, power transformers, EHT transformers (flyback transformers), switch mode transformers, isolation transformers, IF transformers, baluns, and any device that has turns of wire around a former. All these devices can go faulty.
The coating on the wire is called insulation or "enamel" and this can crack or become overheated or damaged due to vibration or movement. When two turns touch each other, a very interesting thing happens. **The winding becomes two separate windings.**

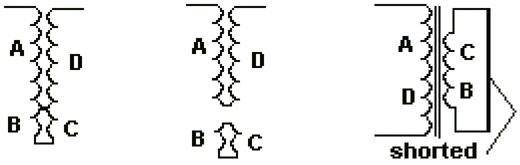

We will take the case of a single winding such as a coil. This is shown in the first diagram above and the winding is wound across a former (a former is a bobbin or plastic molding or something to hold the winding) and back again, making two layers. The bottom and top layers touch at the point shown in the diagram and the current that originally passed though A, B, C, D now passes though A & D.

Winding B C becomes a separate winding as shown in the second diagram.
In other words the coil becomes a TRANSFORMER with a SHORT CIRCUIT on the secondary winding as shown in the third diagram.
When the output wires of a transformer are shorted together, it delivers a very high current because you have created a SHORT-CIRCUIT. This short-circuit causes the transformer to get very hot.
That’s exactly what happens when any coil or transformer gets a “shorted turn.”
The shorted turns can be a single turn or many turns.
It is not possible to measure a fault like this with a multimeter as you don’t know the exact resistance of a working coil or winding and the resistance of a faulty winding may be only 0.001 ohms less.
However when a transformer or coil is measured with an inductance meter, an oscillating voltage (or spike) is delivered into the core as magnetic flux, then the magnetic flux collapses and passes the energy into the winding to produce a waveform. The inductance meter reads this and produces a value of inductance in Henry (or milliHenry or microHenry.)
This is done with the transformer removed from the circuit and this can be a very difficult thing to do, as most transformers have a number of connections.
If the coil or transformer has a shorted turn, the energy from the magnetic flux will pass into the turns that are shorted and produce a current. Almost no voltage will be detected from winding.
The reading from the inductance meter will be low or very low and you have to work out if it is correct.
However there is one major problem with measuring a faulty transformer or coil.
It may only become faulty when power is applied.
The voltage between the turns may be sparking or jumping a gap and creating a problem. A tester is not going to find this fault.
Secondly, an inductance meter may produce a reading but you do not know if the reading is correct. An improved tester is a RING TESTER.
The circuit for a ring tester can be found here:

http://www.flippers.com/pdfs/k7205.pdf

It sends a pulse to the coil and counts the number of returning pulses or "rings." A faulty coil (or winding) may return one pulse but nearly all the energy will be passed to the shorted turns and you will be able to see this on the scale. You will only get one or two return pulses, whereas a good winding will return more pulses.

One way to detect a faulty power transformer is to connect it to the supply and feel the temperature-rise (when nothing is connected to the secondary).
It should NOT get hot.
Detecting shorted turns is not easy to diagnose as you really need another identical

component to compare the results.
Most transformers get very hot when a shorted turn has developed. It may deliver a voltage but the heat generated and a smell from the transformer will indicate a fault.

ISOLATION TRANSFORMER

An isolation transformer is a piece of **Test Equipment** that provides "Mains Voltage" but the voltage is "floating." You will still get a shock if you touch the two output leads, but it has a special use when testing unknown equipment.
Many electrical appliances are fully insulated and only have two leads connected to the mains.
When you take these appliances apart, you do not know which end of say a heating element is connected to the "live" (active) side of the mains and which end connects to the neutral.
I am not suggesting you carry out the following tests, but they are described to show how an isolation transformer works.
If you touch a soldering iron on the "live" (active) end of the heating element it will create a short-circuit.
However when the appliance is connected to the mains via an **isolation transformer**, you can touch an earthed soldering iron on either end of the heater as both leads from the isolation transformer are "floating."
Note: As soon as you earth one lead of the output an isolation transformer, the other lead becomes "active."
You can make your own **Isolation Transformer** by connecting two identical transformers "back-to-back."
The following diagram shows how this is done:

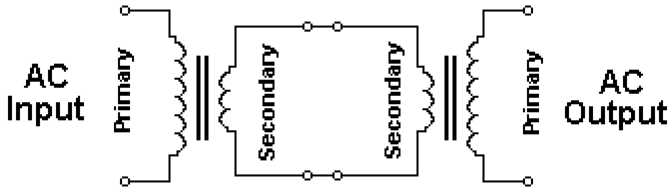

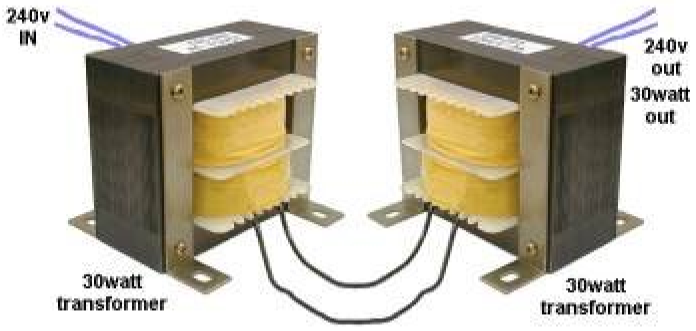

You can use any transformers providing the primary and secondary voltages are the same. The current capability of the secondary winding does not matter. However if you want a supply that has almost the same voltage as your "Mains," you need two transformers with the same voltages.
This handy isolation transformer will provide you with "Mains Voltage" but with a limited current.
In other words it will have a limited capability to supply "wattage." If you are using two 15VA transformers, you will only be able to test an appliance rated at 15 watts.
This has some advantages and some disadvantages.
If you are working on a project, and a short-circuit occurs, the damage will be limited to 15 watts.
If you are using two transformers with different VA ratings, the lower rating will be the capability of the combination.
If the secondaries are not equal, you will get a higher or lower "Mains Voltage."
If you get two transformers from TVs or Monitors, with a rating on the compliance plate of 45 watts, or 90 watts, you can assume the transformers are capable of delivering this wattage and making an isolation transformer will enable you to test

similar items with the safety of being isolated from the mains.
Colin Mitchell designs a lot of "LED lighting lamps" that are connected directly to the mains. He always works with an isolating transformer, just to be safe. Working on exposed "mains" devices is extremely nerve-wracking and you have to be very careful.
The isolation transformer will prevent a BIG EXPLOSION.

DETERMINING THE SPECS OF A TRANSFORMER
Suppose you have a "mains transformer" with unknown output voltages and unknown current capability.
You must be sure it is a mains transformer designed for operation on 50Hz or 60Hz. Switch-Mode transformers operate at frequencies 40kHz and higher and are not covered in this discussion.
To be on the safe-side, connect the unknown transformer to the output of your isolating transformer.
Since the transformer will take almost no current when not loaded, the output voltages it produces will be fairly accurate. Measure the input AC voltage and output AC voltage.
If the transformer has loaded your isolating transformer it will be faulty.
Mains transformers are approx 15VA for 500gm, 30VA for 1kgm 50VA for 2kgm and and 100VA for 2.5kgm.
VA stands for Volts-Amps and is similar to saying watts. Watts is used for DC circuits, while VA refers to AC circuits.
Once you have the weight of the transformer and the output voltage, you can work out the current capability of the secondary.
For transformers up to 30vA, the output voltage on no-load is 30% higher than the final "loaded voltage."
This is due to the poor regulation of these small devices.
If the transformer is 15VA and the output voltage is 15v AC, the current will be 1 amp AC.
You can check the "quality" of the transformer, (the regulation) by fully loading the output and measuring the final voltage. If the transformer has a number of secondaries, the VA rating must be divided between all the windings.

240v to 110v ISOLATION TRANSFORMER

Here's how to create a 110v isolating transformer:
Find a 240v:12v transformer.
Now find a transformer that has two secondary windings, such as 240v:12v+12v.

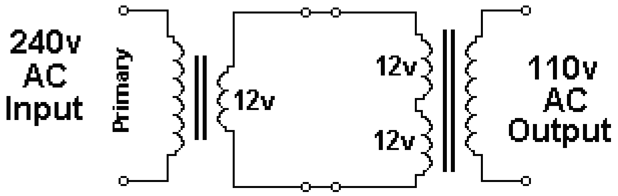

Connect the two transformers as shown in the circuit above. If the output is zero, connect ONE of the 12v windings of the second transformer around the other way.

110v to 240v ISOLATION TRANSFORMER

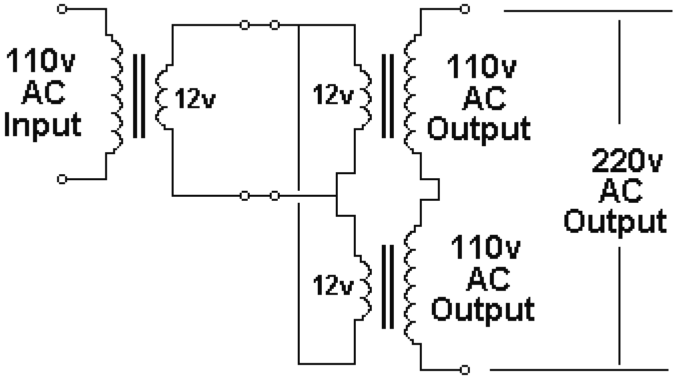

A 110v to 240v isolation transformer can be created by connecting 3 identical transformers as shown in the diagram above. If the output is zero, connect one of the outputs around the other way.

TRANSFORMER RATINGS

Question from a reader:

I have a 28v - 0 - 28v transformer @3amps. Does this mean each side is 1.5 amps?

The transformer is called CENTRE TAPPED and is shown in figures B and C.

It is designed to be connected to two diodes so each winding takes it in turn to deliver the current as shown in diagram C and the output will be 28v AC at 3 amps.

The 28v and 3 amp are both AC values.

If you connect across both outside wires, the output will be 56v at 1.5 amp.

This is because the transformer has a VA rating of 28 x 3 = 84VA. This is very similar to the term "watts."

When the 28v AC is rectified and smoothed, it becomes 28 x 1.4 = 39v (minus 0.6v across the diode) and since the transformer has a rating of 84 VA, the current must be reduced to 84/39 = 2.1 amps to maintain the VA rating.

Some transformers are specified as say: 12v - 0 - 12v, but the wiring diagram is shown as "A." This transformer should be specified as 12v + 12v as the secondaries are separate.

12v - 0 - 12v means the two secondary windings are NOT separate.

It does not make any difference to the output voltage and current, if the windings are separate or joined. The only difference is 12v +12v can be turned into two separate 12v outputs.

If you do not know the output current for a particular transformer, go to the website of electronic parts suppliers and compare the weight of your transformer with others. This will give you a VA rating and you can work out the current, once you know the output voltage.

Note: the output current finishes up ONLY 60% of the rating on the transformer tag because the rating is an **AC RATING**.

With 2 separate secondaries, you can parallel the outputs to get double the current, but don't forget 12v + 12v @ 3amp means 12v in parallel with 12v will provide 2amp DC and the DC voltage will be about 17v.

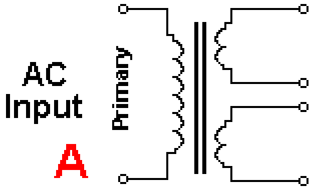

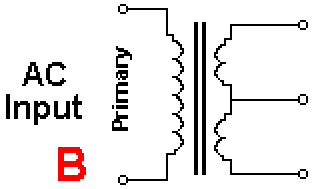

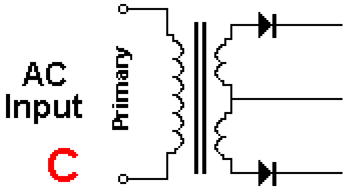

CURRENT TRANSFORMER

A Current Transformer is really an ordinary transformer.
All transformers produce a CURRENT output and a VOLTAGE output.
If you put an ammeter across the secondary, the current will increase through the meter when the primary voltage is increased.
This is because the output voltage will increase and this voltage will allow a higher current to flow.

WHY DETECT CURRENT? Why not voltage?

Because the voltage of say the "240v AC" is always 240v but the current can increase from say 1 amp to nearly 15 amps, depending what appliance is connected. So it is pointless measuring voltage.
A Current Transformer is a step-up transformer. When we say step-up and step-down, we are referring to the voltage - comparing the primary voltage to the secondary voltage. (Most transformers on the "mains" are step-down transformers and are used as power supplies to laptops, phone chargers etc.) Even a welding transformer is a step-down device and produces about 20v to 70v, while the current can be as high as 100 amps. This current is higher than the mains will deliver and is needed to melt the metal at the point of contact of the probe and the item being welded.
A Current Transformer is a step-up transformer. The primary consists of a single turn (or maybe 2 - 5 turns) and the secondary has 100 turns (or more).
This means the voltage seen by the primary will be increased 100 times and appear as anything from a few hundred millivolts to a few volts, depending on the quality of the coupling. (the magnetic coupling between the wire through the centre of the core, the quality of the core to transfer this magnetic flux to the secondary turns.)
This voltage is then passed to a low value resistor, where the voltage is reduced to a level that suits the detection circuit and the resulting millivolts is interpreted as current in the wire being tested.

Recapping:

The reading on the secondary has no relation to the current in the primary. We need to add a LOAD RESISTOR and create a table before we can use the transformer.
There is no such thing as a CURRENT TRANSFORMER. It is really an INSTRUMENT TRANSFORMER and the scale has been marked in units of CURRENT after measurements have been made. (INSTRUMENT TRANSFORMER means it is a device that helps us to produce a connection between current flowing through a wire and a reading on a meter or display).
If we connect a load to the secondary, (say an ammeter), it will produce a reading that increases when the current through the single primary turn is increased. That's because the ammeter is a LOAD. But the reading is meaningless until be calibrate the scale.

Now, lets look at the primary.
A wire (or cable) through the centre of the core is counted as one turn. If the turn is wrapped around the core, the coupling will be improved, but if we always use a straight wire, it does not matter where it is positioned inside the centre of the core.
It does not matter if the magnetic interaction of the flux from the wire is good or bad, we just have to keep to the same way of using the transformer.
The calibration can be done with any poor coupling and the result will be accurate for all future readings.
If a low-value resistor is placed across the secondary, the voltage across this resistor will increase and also the current through it will increase. But we are not going to measure the current through the resistor. We are going to measure the voltage across the resistor and by taking lots of reading we will finish up with a scale or table and this is called CALIBRATION. The results will be equated to the current flowing through the primary wire (primary turn).
A clamp meter uses a current transformer and the jaws must be closed completely and cleanly for the flux to flow around the core and produce a reading in the secondary.
Dirt in the jaws will reduce the reading considerably.
You cannot measure the current in a "power cord" because it contains both the active and neutral wires.
Even though the current is a maximum in both conductors at the same time, the current is flowing in two different directions and the magnetic flux produced by one conductor is clockwise and the other is anticlockwise and they are cancelled by each other.

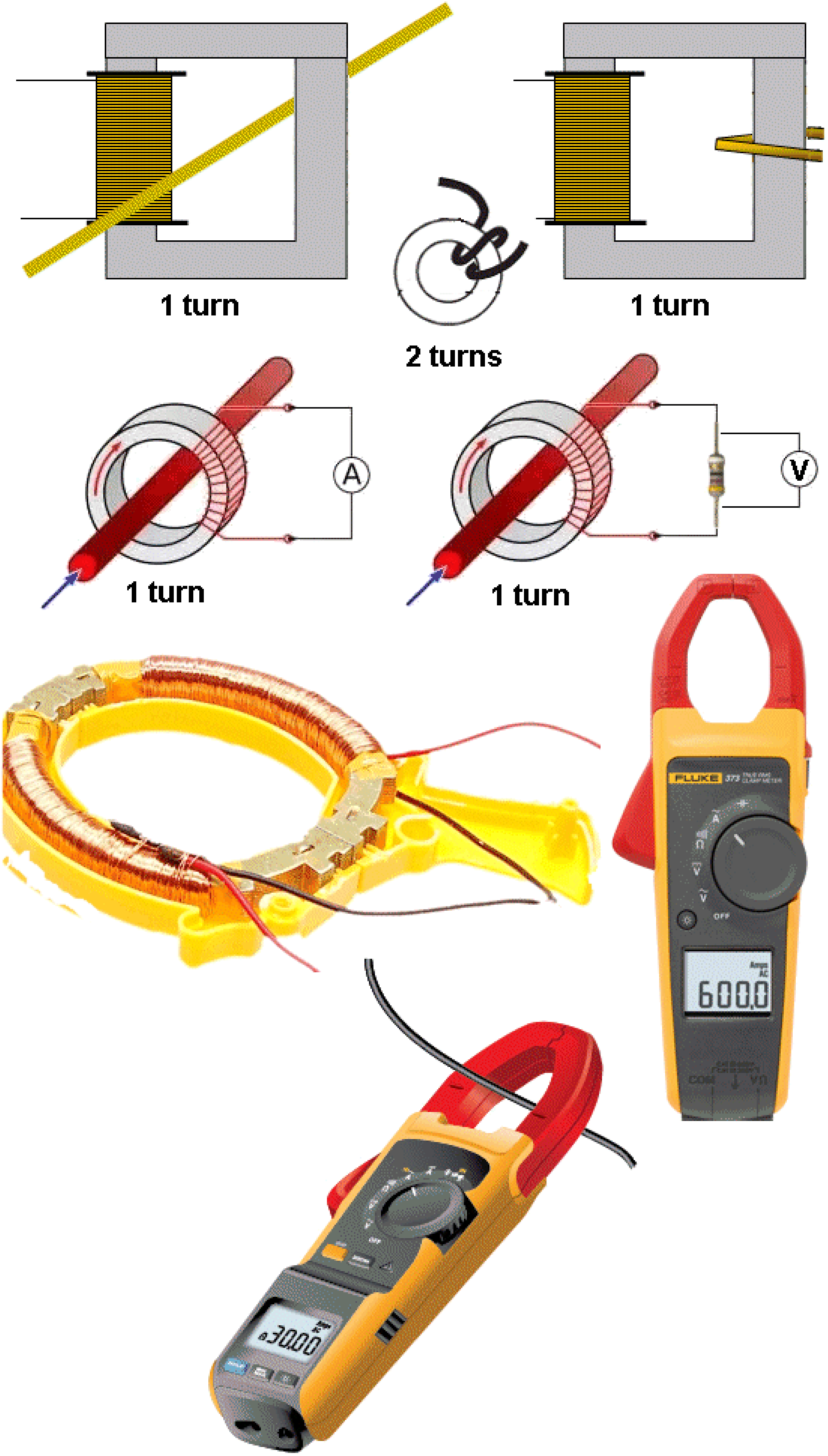
1 turn
2 turns
1 turn
A
1 turn
V
1 turn
FLUKE 373
600.0
30.00

Identifying ONE TURN.
The quality (the coupling) of a single STRAIGHT wire

through the centre of a core is very poor
but if all readings are taken with this amount
of coupling, the readings will be accurate, as the
calibrations have been done with this arrangement.

OPTO ISOLATORS and OPTO COUPLERS

Opto Isolators and Opto Couplers are the same thing. A common opto-coupler is 4N35. It is used to allow two circuits to exchange signals yet remain electrically isolated. The signal is applied to the LED, which shines on a silicon NPN photo-transistor in the IC.

The light is proportional to the signal, so the signal is transferred to the photo transistor to turn it on a proportional amount. Opto-couplers can have Light Activated SCR's, photodiodes, TRIAC's and other semiconductor devices as an output. The 4N35 opto-coupler schematic is shown below:

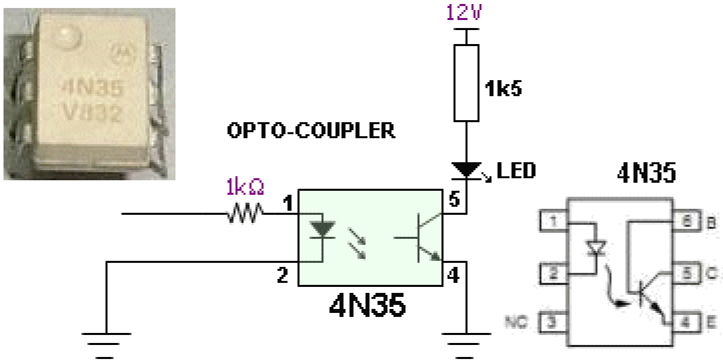

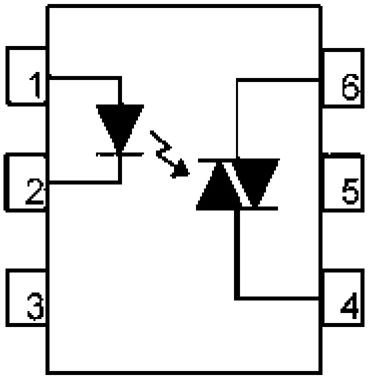

An opto-Coupler using a TRIAC
Note: the pinout is different to 4N35

TESTING AN OPTO COUPLER

Most multimeters cannot test the LED on the input of an opto-coupler because the ohms range does not have a voltage high enough to activate the LED with at least 2mA.

You need to set-up the test-circuit shown above with a 1k resistor on the input and 1k5 on the output. When the 1k is connected to 12v, the output LED will illuminate. The opto-coupler should be removed from circuit to perform this test.

TRANSISTORS

Transistors are solid-state devices and although they operate completely differently to a diode, they appear as two back-to-back diodes when tested.

There are basically 2 types of transistor NPN and PNP.

A transistor is sometimes referred to as BJT (Bi-polar Junction Transistor) to distinguish it from other types of transistor such as Field Effect transistor, Programmable Unijunction Transistor and others.

In the following diagram, two diodes are connected together and although the construction of a transistor is more complex, we see the transistor as two diodes when testing it.

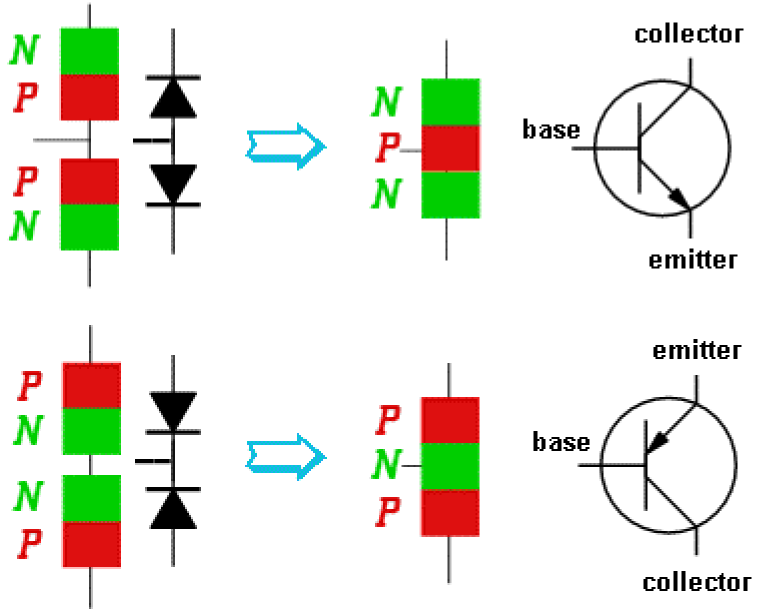

A TRANSISTOR APPEARS AS TWO DIODES WHEN TESTING IT

All transistors have three leads. Base (b), Collector (c), and Emitter (e).
For an NPN transistor, the arrow on the emitter points away from the base.
It is fortunate that the arrow on both symbols points in the direction of the flow of current (Conventional Current) and this makes it easy to describe testing methods using our simplified set of instructions. The symbols have been drawn exactly as they appear on a circuit diagram.

All transistors **are the same** but we talk about digital and analogue transistors.
There is no difference between the two.
The difference is the circuit. And the only other slight difference between transistors is the fact that some have inbuilt diodes and resistors to simplify the rest of the circuit.
All transistors work the same way. The only difference is the amount of amplification they provide, the current and voltage they can withstand and the speed at which they work. For simple testing purposes, they are all the same.

NPN transistors are the most common and for an NPN transistor, the following applies.
(the opposite applies for PNP)
To test a transistor, there is **one thing** you have to know:
When the base voltage is higher than the emitter, current flows though the collector-emitter leads.
As the voltage is increased on the base, nothing happens until the voltage reaches 0.55v. At this point a very small current flows through the collector-emitter leads. As the voltage is increased, the current-flow increases. At about 0.75v, the current-flow is a MAXIMUM. (can be as high as 0.9v). That's how it works. A transistor also needs **current** to flow into the base to perform this amplifying function and this is the one feature that separates an ordinary transistor from a FET.

If the voltage on the base is 0v, then instantly goes to 0.75v, the transistor initially passes NO current, then FULL current. The transistor is said to be working in its two states: OFF then ON (sometimes called: "cut-off" and "saturation"). These are called digital states and the transistor is said to be a **DIGITAL TRANSISTOR** or a **SWITCHING TRANSISTOR** , working in **DIGITAL MODE.**

If the base is delivered 0.5v, then slowly rises to 0.75v and slowly to 0.65v, then

0.7v, then 0.56v etc, the transistor is said to be working in ANALOGUE MODE and the transistor is an **ANALOGUE TRANSISTOR**.
Since a transistor is capable of amplifying a signal, it is said to be an active device. Components such as resistors, capacitors, inductors and diodes are not able to amplify and are therefore known as passive components.

In the following tests, use your finger to provide the **TURN ON** voltage for the base (this is 0.55v to 0.7v) and as you press harder, more current flows into the base and thus more current flows through the collector-emitter terminals. As more current flows, the needle of the multimeter moves UP-SCALE.

TESTING A TRANSISTOR ON A DIGITAL METER

Testing a transistor with a **Digital Meter** must be done on the "DIODE" setting as a digital meter does not deliver a current through the probes on some of the resistance settings and will not produce an accurate reading.
The "DIODE" setting must be used for diodes and transistors. It should also be called a "TRANSISTOR" setting.

TESTING AN unknown TRANSISTOR

The first thing you may want to do is test an unknown transistor for COLLECTOR, BASE AND EMITTER. You also want to perform a test to find out if it is NPN or PNP. That's what this test will provide.
You need a cheap multimeter called an ANALOGUE METER - a multimeter with a scale and pointer (needle).
It will measure resistance values (normally used to test resistors) - (you can also test other components) and Voltage and Current. We use the resistance settings. It may have ranges such as "x10" "x100" "x1k" "x10"
Look at the resistance scale on the meter. It will be the top scale.
The scale starts at zero on the right and the high values are on the left. This is opposite to all the other scales.
When the two probes are touched together, the needle swings FULL SCALE and reads "ZERO." Adjust the pot on the side of the meter to make the pointer read exactly zero.

How to read: "x10" "x100" "x1k" "x10"

Up-scale from the zero mark is "1"
When the needle swings to this position on the "x10" setting, the value is 10 ohms.
When the needle swings to "1" on the "x100" setting, the value is 100 ohms.
When the needle swings to "1" on the "x1k" setting, the value is 1,000 ohms = 1k.
When the needle swings to "1" on the "x10k" setting, the value is 10,000 ohms = 10k.
Use this to work out all the other values on the scale.
Resistance values get very close-together (and very inaccurate) at the high end of the scale. [This is just a point to note and does not affect testing a transistor.]

Step 1 - FINDING THE BASE and determining NPN or PNP

Get an unknown transistor and test it with a multimeter set to "x10"
Try the 6 combinations and when you have the black probe on a pin and the red probe touches the other pins and the meter swings nearly full scale, you have an NPN transistor. The black probe is BASE
If the red probe touches a pin and the black probe produces a swing on the other two pins, you have a PNP transistor. The red probe is BASE
If the needle swings FULL SCALE or if it swings for more than 2 readings, the transistor is **FAULTY**.

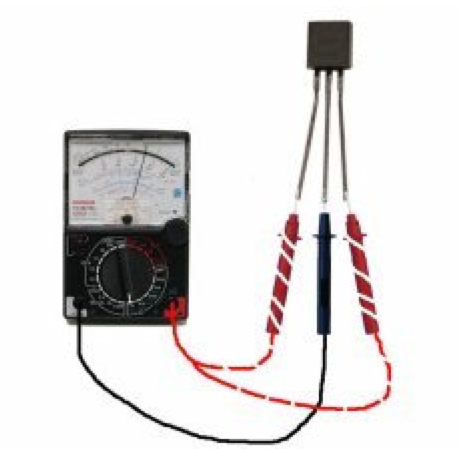

This is an NPN transistor
The black probe is the BASE

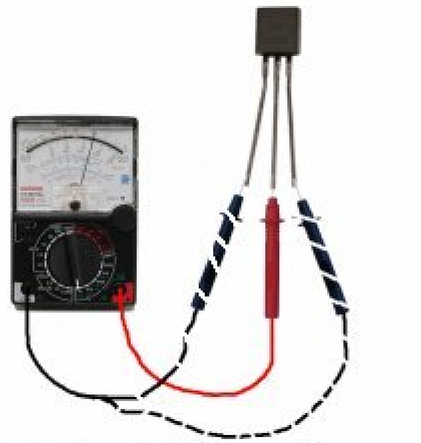

This is a PNP transistor
The red probe is the BASE

Step 2 - FINDING THE COLLECTOR and EMITTER
Set the meter to "x10k."
For an NPN transistor, place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

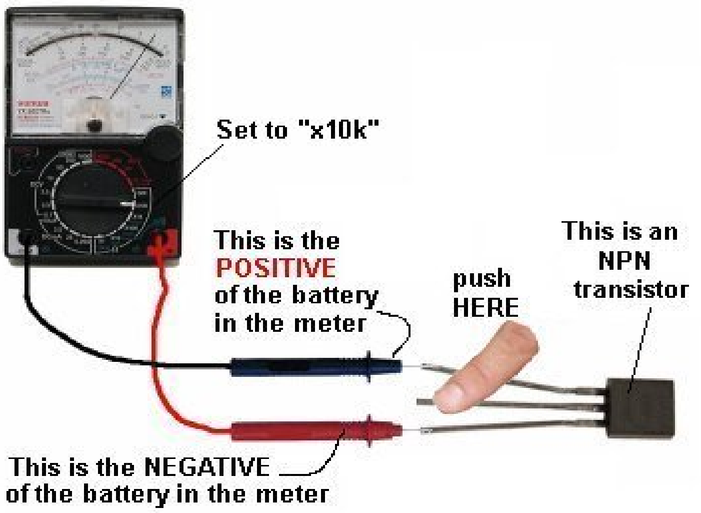
Set to "x10k"
This is the POSITIVE of the battery in the meter
push HERE
This is an NPN transistor
This is the NEGATIVE of the battery in the meter

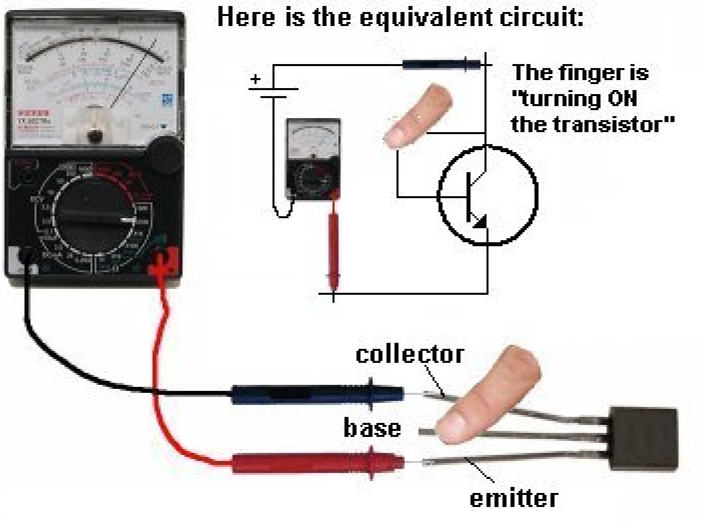
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
collector
base
emitter

You have found the Collector Base and Emitter

For a PNP transistor, set the meter to "x10k" place the leads on the transistor and when you press hard on the two leads shown in the diagram below, the needle will swing almost full scale.

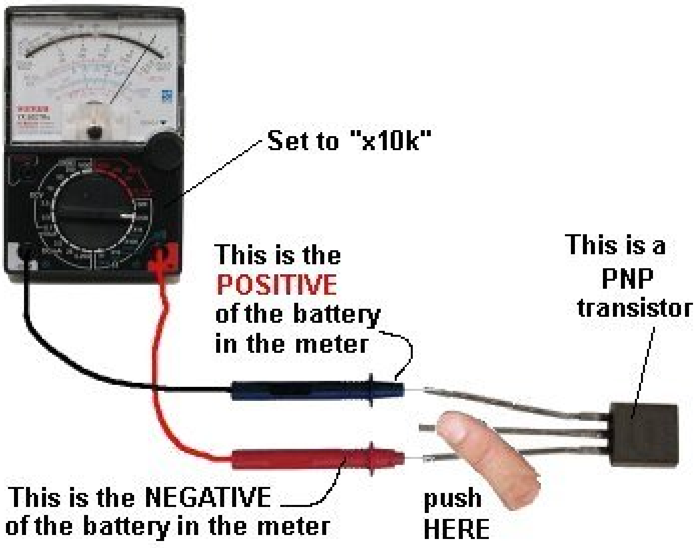
Set to "x10k"
This is the POSITIVE of the battery in the meter
This is a PNP transistor
This is the NEGATIVE of the battery in the meter
push HERE

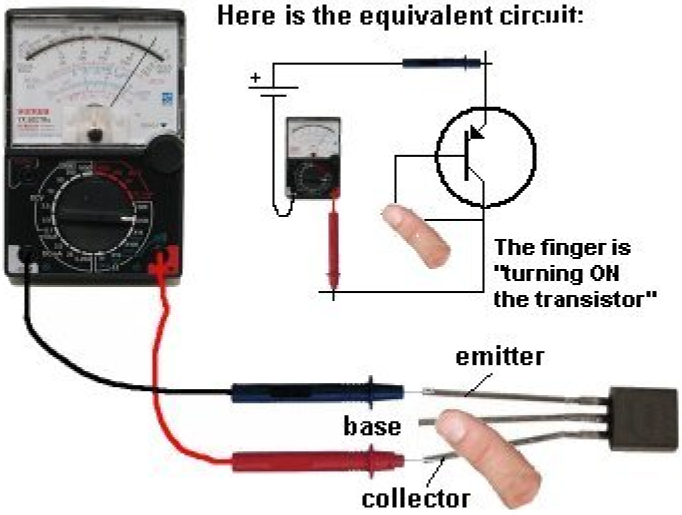
Here is the equivalent circuit:
+
The finger is "turning ON the transistor"
emitter
base
collector

You have found the Collector Base and Emitter

SIMPLEST TRANSISTOR TESTER

The simplest transistor tester uses a 9v battery, 1k resistor and a LED (any colour). Keep trying a transistor in all different combinations until you get one of the circuits below. When you push on the two leads, the LED will get brighter.

The transistor will be NPN or PNP and the leads will be identified:

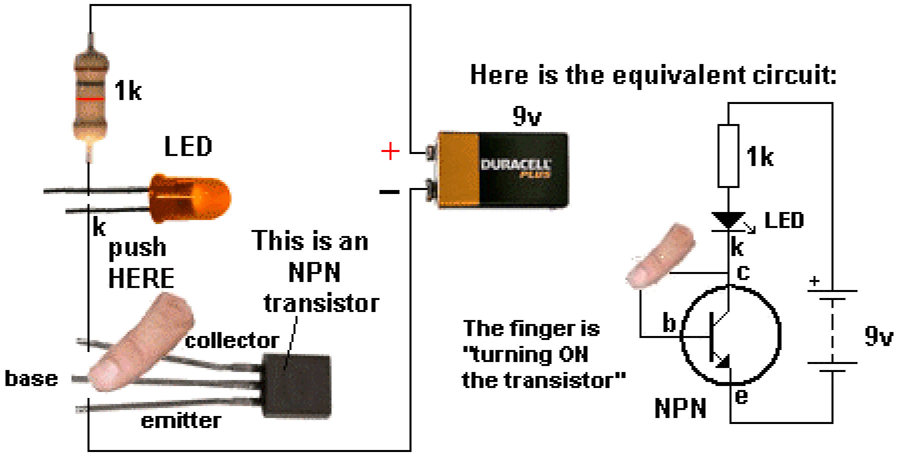
1k
LED
k
push
HERE
This is an
NPN
transistor
collector
base
emitter
+
−
9v
DURACELL PLUS
Here is the equivalent circuit:
1k
LED
k
c
b
e
NPN
9v
+
The finger is
"turning ON
the transistor"

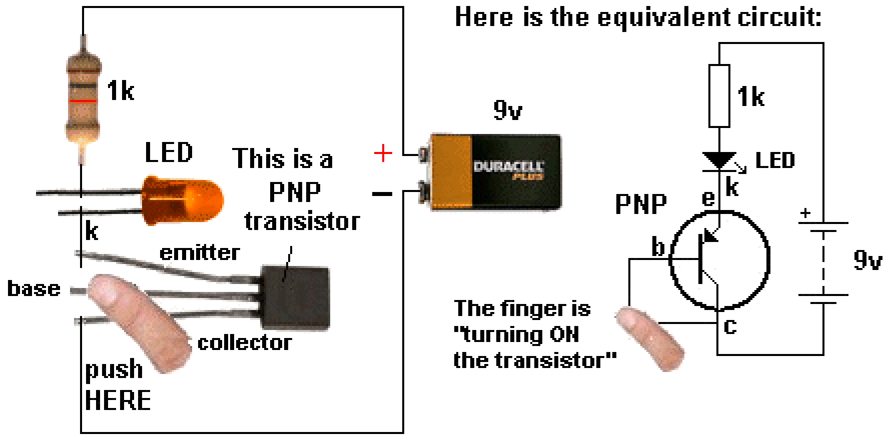
1k
LED
k
This is a
PNP
transistor
emitter
base
collector
push
HERE
+
−
9v
DURACELL PLUS
Here is the equivalent circuit:
1k
LED
k
e
PNP
b
c
9v
+
The finger is
"turning ON
the transistor"

The leads of some transistors will need to be bent so the pins are in the same positions as shown in the diagrams. This helps you see how the transistor is being turned on. This works with NPN, PNP transistors and Darlington transistors.

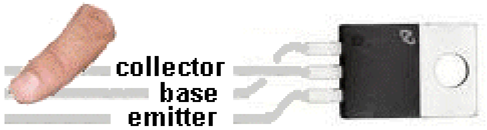
collector
base
emitter

NPN transistor

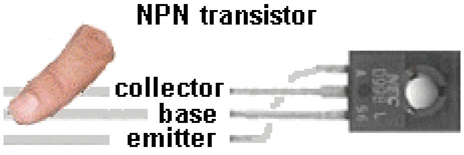
collector
base
emitter

NPN transistor

HEATSINKING

Heat generated by current flowing between the collector and emitter leads of a transistor causes its temperature to rise. This heat must be conducted away from the transistor otherwise the rise may be high enough to damage the P-N junctions inside the device. Power transistors produce a lot of heat, and are therefore usually mounted on a piece of aluminium with fins, called a **HEATSINK**.
This draws heat away, allowing it to handle more current. Low-power signal transistors do not normally require heat sinking. Some transistors have a metal body or fin to connect to a larger heatsink. If the transistor is connected to a heatsink with a mica sheet (mica washer), it can be damaged or cracked and create a short-circuit. (See Testing Mica Washers). Or a small piece of metal may be puncturing the mica. Sometimes white compound called **Heatsink Compound** is used to conduct heat through the mica. This is very important as mica is a very poor conductor of heat and the compound is needed to provide maximum thermal conduction.

TRANSISTOR FAILURE

Transistor can fail in a number of ways. They have forward and reverse voltage ratings and once these are exceeded, the transistor will ZENER or conduct and may fail. In some cases a high voltage will "puncture" the transistor and it will fail instantly. In fact it will fail much faster via a voltage-spike than a current overload.

It may fail with a "short" between any leads, with a collector-emitter short being the most common. However failures will also create shorts between all three leads.
A shorted transistor will allow a large current to flow, and cause other components to heat up.
Transistors can also develop an open circuit between base and collector, base and emitter or collector and emitter.
The first step in identifying a faulty transistor is to check for signs of overheating. It may appear to be burnt, melted or exploded. When the equipment is switched off, you can touch the transistor to see if it feels unusually hot. The amount of heat you feel should be proportional to the size of the transistor's heat sink. If the transistor has no heat sink, yet is very hot, you can suspect a problem.
DO NOT TOUCH A TRANSISTOR IF IT IS PART OF A CIRCUIT THAT CARRIES 240VAC. Always switch off the equipment before touching any components.

TRANSISTOR REPLACEMENT

If you can't get an exact replacement, refer to a transistor substitution guide to identify a near equivalent.

The important parameters are:
- Voltage
- Current
- Wattage
- Maximum frequency of operation

The replacement part should have parameters equal to or higher than the original.

Points to remember:
- Polarity of the transistor i.e. PNP or NPN.
- At least the same voltage, current and wattage rating.
- Low frequency or high frequency type.
- Check the pinout of the replacement part
- Use a desoldering pump to remove the transistor to prevent damage to the printed circuit board.
- Fit the heat sink.
- Check the mica washer and use heat-sink compound
- Tighten the nut/bolt - not too tight or too loose.
- Horizontal output transistors with an integrated diode should be replaced with the same type.

DIGITAL TRANSISTORS

There is no such thing as a DIGITAL TRANSISTOR or an AUDIO TRANSISTOR.
All transistors are just "TRANSISTORS" and the surrounding components as well as the type of signal, make the transistor operate in DIGITAL MODE or ANALOGUE MODE.
But we have some transistors that have inbuilt resistors to make them suitable for connecting to a digital circuit without the need for a base resistor.
Here is the datasheet for an NPN transistor BCR135w and PNP datasheet for BCR185w.
These transistors are called "Digital Transistors" because the "base lead" can be connected directly to the output of a digital stage. This "lead" or "pin" is not really the base of the transistor but a 4k7 (or 10k) resistor connected to the base allows the transistor to be connected to the rest of a digital circuit.
You cannot actually get to the base. The resistor(s) are built into the chip and the transistor is converted into a "Digital Transistor" because it will accept 5v on the "b"

lead.
The 47k is not really needed but it makes sure the transistor is fully turned OFF if the signal on the "b" lead is removed (in other words - if the input signal is converted to a high-impedance signal - see tri-state output from microcontrollers for a full explanation).
This transistor is designed to be placed in a circuit where the input changes from low to high and high to low and does not stop mid-way. This is called a DIGITAL SIGNAL and that is one reason why the transistor is called a digital transistor. (However you could stop half-way but the transistor may heat up and get too hot).
Any transistor placed in a digital circuit can be called a "digital transistor" but it is better to say it is operating in DIGITAL MODE.

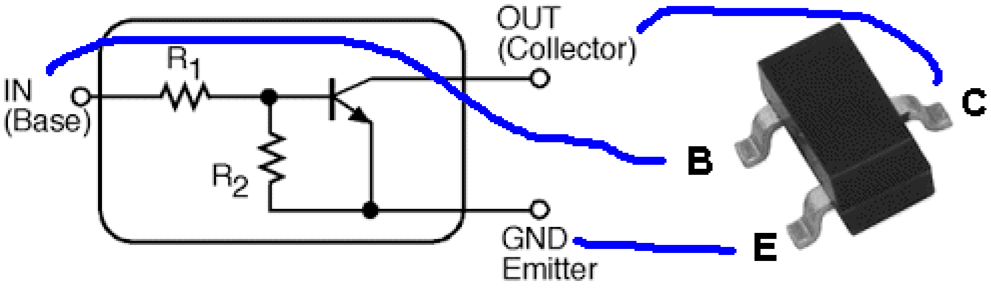
IN
(Base)
R1
R2
OUT
(Collector)
GND
Emitter
B
C
E

The digital transistor has two resistors included inside the case
R1 is about 10k and R2 is approx 47k

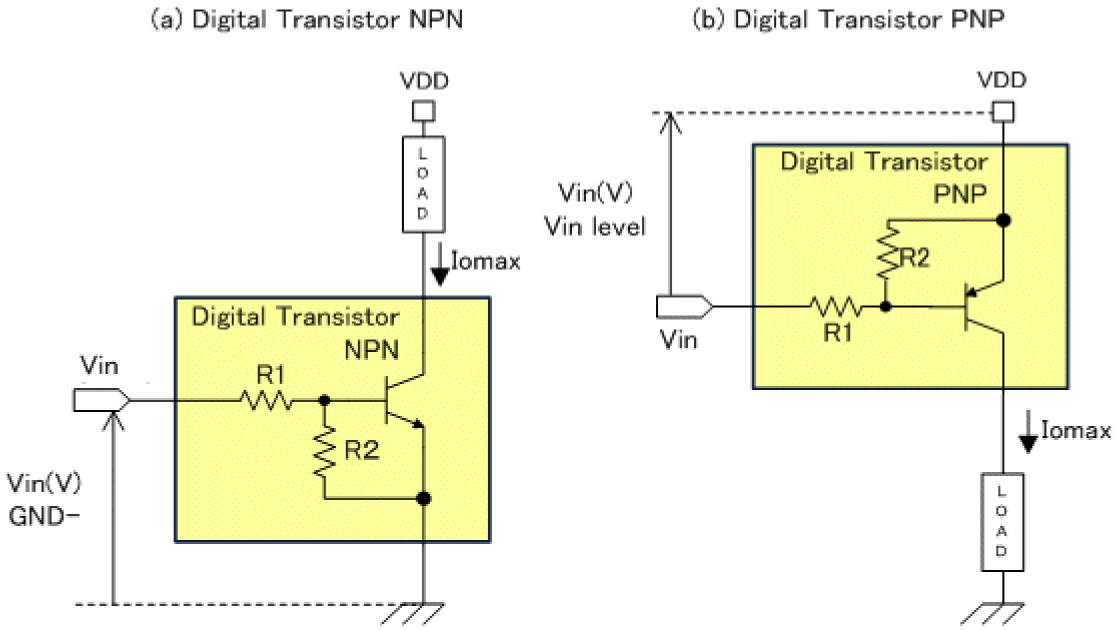
(a) Digital Transistor NPN
VDD
LOAD
Iomax
Digital Transistor
NPN
Vin
R1
R2
Vin(V)
GND−
(b) Digital Transistor PNP
VDD
Vin(V)
Vin level
Digital Transistor
PNP
R2
R1
Vin
Iomax
LOAD

Equivalent circuit of the digital transistor

These transistors can be made to work in analogue circuits because they are ordinary transistors with a 10k base resistor, but you will have to know what you are doing.

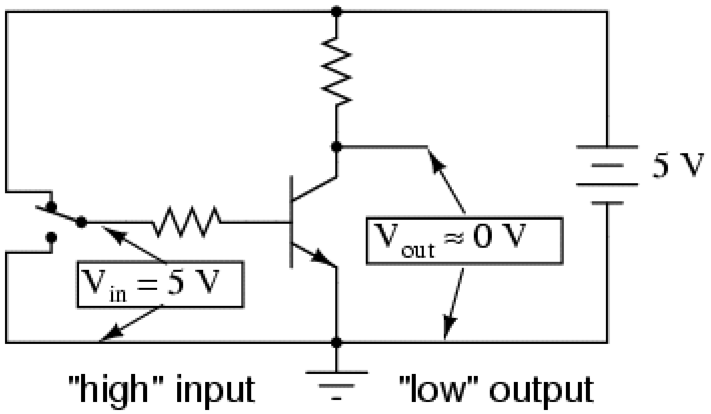

Transistor in saturation

The circuit above shows the digital transistor is designed to allow a voltage of 5v to be supplied to the "base" pin and the transistor will Fully Conduct.
This type of transistor saves putting a base resistor on the PC board.
It can be tested just like a normal transistor but the resistance between base and emitter will be about 5k to 50k in both directions. If the collector-emitter is low in both directions the transistor is damaged.
Here's how to look at how the transistor works:
The 10k resistor on the base will allow 0.5mA to flow into the base. But the 47k will reduce this to 0.4mA. If the transistor has a gain of 100, the collector-emitter current can be 40mA.
To determine the current capability of the transistor, connect 100R load and turn the transistor ON. This will allow about 100mA for the collector-emitter current. Measure the collector-emitter voltage. If it is more than 0.5v, the transistor is OVER-LOADED.

DARLINGTON TRANSISTORS

A DARLINGTON TRANSISTOR is two transistors in a single package with three leads. They are internally connected in cascade so the gain of the pair is very high. This allows a very small input signal to produce a large signal at the output. They have three leads (Base, Collector and Emitter and can be PNP or NPN) and are equivalent to the leads of a standard individual transistor, but with a very high gain. The second advantage of a Darlington Transistor is its high input impedance. It puts very little load on the previous circuit.
Some Darlington transistors have a built-in diode and/or built-in resistor and this will produce a low reading in both directions between the base and emitter leads.

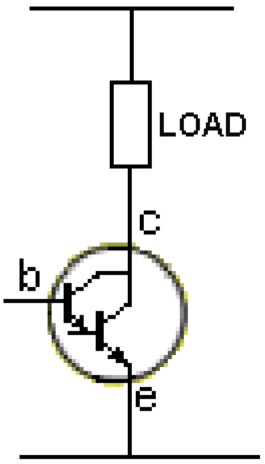

NPN Darlington

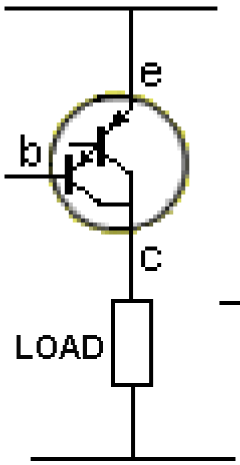

PNP Darlington

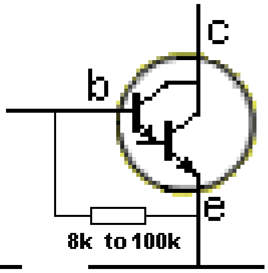

Darlington transistors are tested the same as an ordinary transistor and a multimeter will produce about the same deflection, even though you will be measuring across two junctions, (and a base-emitter resistor is present).

HORIZONTAL OUTPUT TRANSISTORS, SWITCH-MODE TRANSISTORS, FLYBACK TRANSISTORS, POWER TRANSISTORS, VERTICAL TRANSISTORS . . . .

These are all names given to a transistor when it is used in a particular circuit. ALL these transistors are the same for testing purposes.

We are not testing for gain, maximum voltage, speed of operation or any special feature. We are just testing to see if the transistor is completely faulty and SHORTED.

A transistor can have lots of other faults and the circuit **using the transistor** is the best piece of TEST EQUIPMENT as it is detecting the fault.

TESTING MOSFETs and FETs

MOSFETs and **JFETs** are all part of the **FET** family.

MOSFET stands for **Metal Oxide Semiconductor Field Effect Transistor.**

FETs operate exactly the same as a "normal" transistor except they have different names for the input and output leads and the voltage between the gate and the source has to between 2v to 5v for the device to turn on fully. A FET requires almost NO CURRENT into the Gate for it to turn on and when it does, the voltage between drain and source is very low (only a few mV). This allows them to pass very high currents without getting hot. There is a point where they start to turn on and the input voltage must rise higher than this so the FET turns on FULLY and does not get hot.

Field Effect Transistors are difficult to test with a multimeter, but "fortunately" when a power **MOSFET** blows, it is completely damaged. All the leads will show a short circuit. 99% of bad **MOSFETs** will have GS, GD and DS shorted.

The following symbols show some of the different types of MOSFETs:

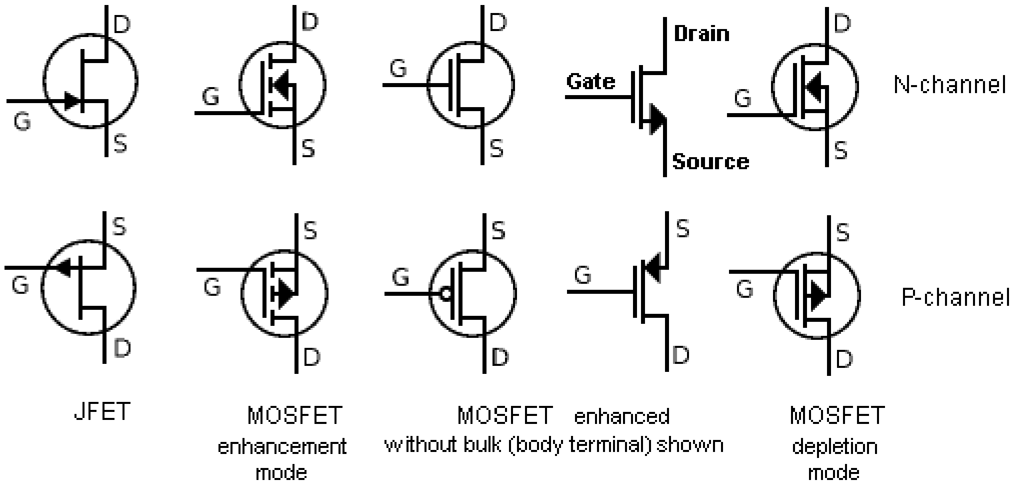

Most **MOSFET** transistors cannot be tested with a multimeter. This due to the fact that the Gate needs 2v - 5v to turn on the device and this voltage is not present on the probes of either meter set to any of the ohms ranges.

You need to build the following Test Circuit:

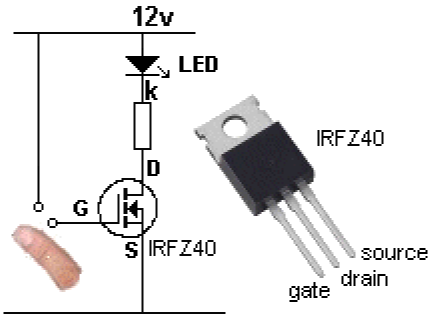

N-channel

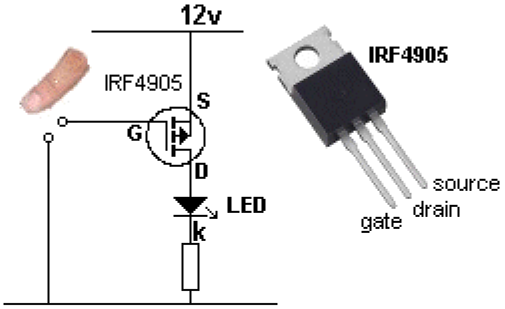

P-channel

Touching the Gate will increase the voltage on the Gate and the MOSFET will turn ON and illuminate the LED. Removing your finger will turn the LED off.
Large devices such as the TO-220 types shown above do not like static electricity on the gate and you have to be careful not to "spike" the gate with any static. Generally this type of device is not "super sensitive" and you can use your finger or a large value resistor.
When replacing one of these devices, there are 2 things to match-up.
Voltage and Current.
In most cases, the "turn-ON" resistance (the resistance between Source and Drain) will be the same (something like 22 milli ohms) and the speed of operation will be ok.
Check the voltage needed to turn the gate ON and make sure you can supply the required voltage.

SILICON CONTROLLED RECTIFIERs (SCR)

The **Silicon Controlled Rectifier** (SCR) is a semiconductor device that is a member of a family of control devices known as **Thyristors**. It is a 3-leaded device and when a small current enters the Gate, the **thyristor** turns on. AND STAYS ON. It only conducts current between Anode and Cathode in one direction and it is mainly only used in DC circuits. When it is used with AC, it will only conduct for a maximum of half the cycle.

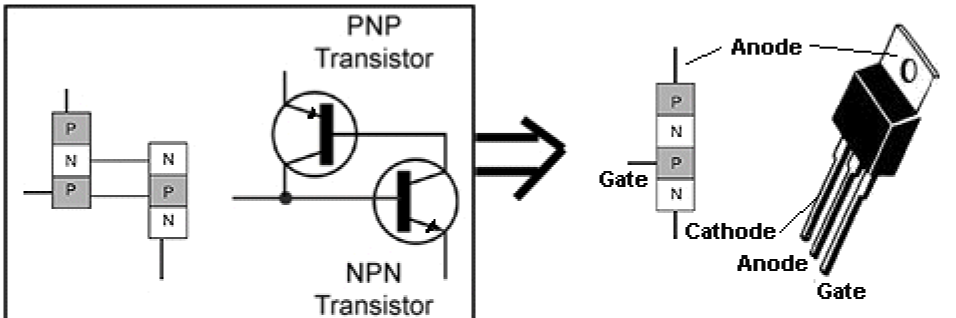
PNP Transistor
NPN Transistor
P
N
P
N
P
N
Anode
Gate
Cathode
Anode
Gate

THYRISTOR or SCR

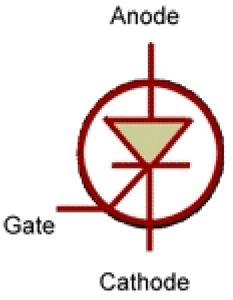
Anode
Gate
Cathode

SCR
Silicon Controlled Rectifier

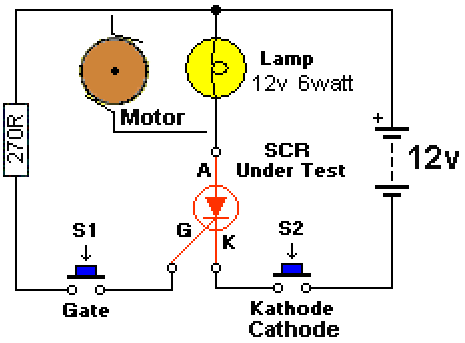
Motor
Lamp
12v 6watt
270R
SCR
Under Test
A
G
K
12v
S1
S2
Gate
Kathode
Cathode

SCR TESTER

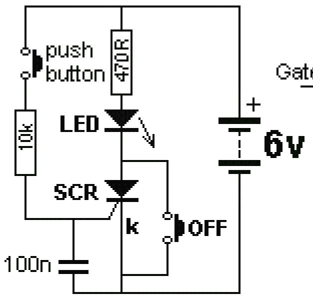
push button
470R
10k
LED
SCR
k
OFF
100n
6v

A

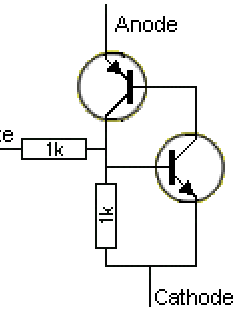
Anode
Gate
1k
1k
Cathode

SCR WITH TRANSISTORS

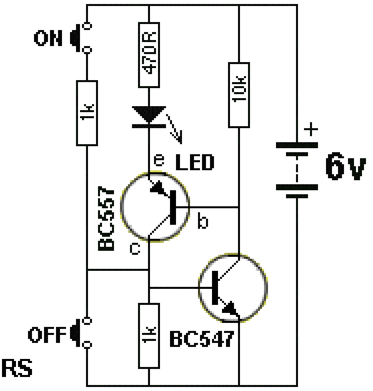
ON
470R
10k
1k
LED
e
b
c
BC557
6v
OFF
1k
BC547

B

To understand how an SCR "latches" when the gate is provided with a small current, we can replace it with two transistors as shown in diagram B above. When the ON button is pressed, the BC547 transistor turns on. This turns ON the BC557 and it takes over from the action of the switch.
To turn the circuit off, the OFF button removes the voltage from the base of the BC547.

Testing an SCR

An **SCR** can be tested with some multimeters but a minimum current Anode-to-Cathode is needed to keep the device turned on. Some multimeters do not provide this amount of current and the **SCR Tester** circuit above is the best way to test these devices.
Shorted SCRs can usually be detected with an ohmmeter check (SCRs usually fail shorted rather than open).
Measure the anode-to-cathode resistance in both the forward and reverse direction; a good SCR should measure near infinity in both directions.
Small and medium-size SCRs can also be gated ON with an ohmmeter (on a digital meter use the Diode Check Function). Forward bias the SCR with the ohmmeter by connecting the black (-) lead to the anode and the red (+) lead to the cathode

(because the + of the battery is connected to the negative lead, in most analogue multimeters). Momentarily touch the gate lead to the anode while the probes are still touching both leads; this will provide a small positive turn-on voltage to the gate and the cathode-to-anode resistance reading will drop to a low value. Even after removing the gate voltage, the SCR will stay conducting. Disconnecting the meter leads from the anode or cathode will cause the SCR to revert to its non-conducting state.
When making the above test, the meter impedance acts as the SCR load. On larger SCRs, it may not latch ON because the test current is not above the SCR holding current.

Using the SCR Tester

Connect an SCR and press Switch2. The lamp should not illuminate. If it illuminates, the SCR is around the wrong way or it is faulty.
Keep Switch 2 PRESSED. Press Sw1 very briefly. The lamp or motor will turn ON and remain ON. Release Sw 2 and press it again. The Lamp or motor will be OFF.

TRIACs

A triac is a bidirectional, three-terminal dual, back-to-back thyristor (SCR) switch. This device will conduct current in both directions when a small current is constantly applied to the Gate.
If the gate is given a small, brief, current during any instant of a cycle, it will remain triggered during the completion of the cycle until the current though the Main Terminals drops to zero.
This means it will conduct both the positive and negative half-cycles of an AC waveform. If it is tuned on (with a brief pulse) half-way up the positive waveform, it will remain on until the wave rises and finally reaches zero. If it is then turned on (with a brief pulse) part-way on the negative wave, the result will be pulses of energy and the end result will be about 50% of the full-energy delivered at a rate of 100 times per second for a 50HZ supply.
TRIACs are particularly suited for AC power control applications such as motor speed control, light dimmers, temperature control and many others.

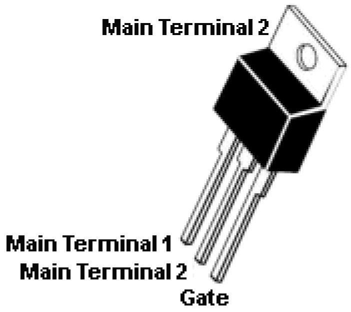

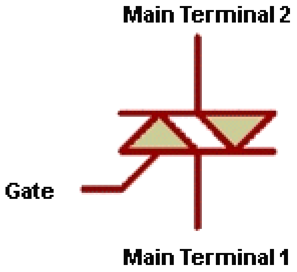

TRIAC (2 SCR's)
TRIode for Alternating Current

Using the TRIAC Tester

Connect a TRIAC and press Switch2. The lamp should not illuminate. If it illuminates, the TRIAC is faulty.

Keep Switch 2 PRESSED. Press Sw1 very briefly. The lamp or motor will turn ON and remain ON. If the lamp does not turn on, reverse the TRIAC as the current into the gate must produce a slight voltage between **Gate** and **Main Terminal 1**.

Release Sw 2 and press it again. The Lamp or motor will be OFF.

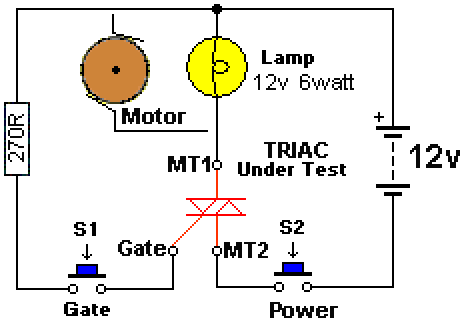

TRIAC TESTER

MICA WASHERS AND INSULATORS

Plastic insulating sheets (washers) between a transistor and heatsink are most often made from mica but some are plastic and these get damaged over a period of time, turn dark and become cracked.

The plastic eventually becomes carbonized and conducts current and can affect the operation of the appliance. You can see the difference between a mica sheet (washer) and plastic by looking where it extends from under the transistor. Replace all plastic insulators as they eventually fail.

SPARK GAPS

Some TV's and monitors with a CRT (picture tube), have spark gaps either on the socket at the end of the tube or on the chassis.

These can consist of two wires inside a plastic holder or a glass tube or special resistive device.

The purpose of a spark gap is to take any flash-over (from inside the tube), to earth. This prevents damage to the rest of the circuit.

However if the tube constantly flashes over, a carbon track builds up between the wires and effectively reduces the screen voltage. This can cause brightness and/or focus problems. Removing the spark-gap will restore the voltage.

These are not available as a spare component and it's best to get one from a discarded chassis.

CO-AX CABLES

Co-Ax cables can produce very high losses and it seems impossible that a few metres of cable will reduce the signal. The author has had a 3 metre cable reduce the signal to "snow" so be aware that this can occur. Faults can also come from a splitter and/or balun as well as dirty plugs and sockets. This can result in very loud bangs in the sound on digital reception.

TESTING EARTH LEAKAGE DETECTORS or Residual Current Devices or Ground Fault Circuit Interrupters or GFCI

An Earth Leakage Detector or Sensor is a circuit designed to continuously monitor the imbalance in the current in a pair of load carrying conductors.

These two conductors are normally the Active and Neutral. Should the imbalance current reach 30mA the sensor will "trip" and remove the voltage (and current) from

the line being monitored.
Some detectors will trip at 15mA.
You cannot alter the sensitivity of the device however there are a number of faults in these devices that can be fixed.
In some devices the contact pressure for the 10Amp or 15 Amp contacts is very weak and they arc and produce an open circuit. The result is this: When you press the rest button, power is not restored to the output.
Clean the contacts with a small file and bend the metal strips to the contacts so they make a very strong contact.
The other fault is the trip mechanism.
The magnetism from the coil does not allow the pin to move and "trip" the contacts. It may be due to a small metal filing or the pin not moving freely enough.
All good Earth Leakage Detectors have a TEST BUTTON. This connects a resistor between the active line and earth so that 15mA or 30mA flows.
The detector should trip immediately. Make sure the trigger mechanism trips when the test button is pressed.
None of the electronics in the detector can be replaced however you can test the mechanical operation and the pressure on the contacts when the unit is removed from the power. Do not work on the device when it is connected to the mains.

TESTING CELLS AND BATTERIES

There is an enormous number of batteries and cells on the market and a number of "battery testers." Instead of buying a battery tester that may give you a false reading, here is a method of testing cells that is guaranteed to work.
There are two types of cell: a **rechargeable** cell and a non rechargeable cell.

The easiest way to test a **rechargeable** cell is to put a group of them in an appliance and use them until the appliance "runs down" or fails to work. If you consider the cells did not last very long, remove them and check the voltage of each cell. The cell or cells with the lowest voltage will be faulty. You can replace them with new cells or good cells you have in reserve.
There is no other simple way to test a rechargeable cell.
You cannot test the "current of a cell" by using an ammeter. A rechargeable cell can deliver 10 amps or more, even when nearly discharged and you cannot determine a good cell for a faulty cell.

Dry cells are classified as "non-rechargeable" cells.
DRY CELLS and MANGANESE CELLS are the same thing. These produce 1.5v per cell (manganese means the Manganese Dioxide depolariser inside the cell. All "dry cells" use manganese dioxide).
ALKALINE CELLS produce between 2 - 10 times more energy than a "dry cell" and produce 1.5v per cell.
Alkaline cells can fail for no reason at any stage in their life and are not recommended for emergency situations.
The output voltage of some Alkaline cells can fall to 0.7v or 0.9v for not apparent reason.
There are lots of other cells including "button cells," hearing-aid cells, air cells, and they produce from 1.2v to 3v per cell.

Note:
Lithium cells are also called "button cells" and they produce 3v per cell.
Lithium cells are non-rechargeable (they are generally called "button cells") but some Lithium cells can be recharged. These are Lithium-ion cells and generally have a voltage of 3.6v. Some Lithium-ion cells look exactly like 3v Lithium cells, so you have to read the data on the cell before charging.

You cannot test the voltage of a cell and come to any conclusion as to the age of the cell or how much energy remains. The voltage of a cell is characteristic to the chemicals used and the actual voltage does not tell you its condition.
Some "dry cells" deliver 1.5v up to the end of their life whereas others drop to about 1.1v very quickly.

Once you know the name of the cell that drops to 1.1v, avoid them as the operation of the equipment "drops off" very quickly.

However if you have a number of different cells and need to know which ones to keep, here's the solution:

1. Check the voltage and use those with a voltage above 1.1v
2. Next, select 500mA or 10A range on a meter and place the probes on a cell. For a AAA or AA cell, the current should be over 500mA and the needle will swing full scale very quickly.

Keep the testing short as you are short-circuiting the cell but it is the only way to determine the internal impedance of the cell and this has a lot to do with its stage-of-charge.
This will give you a cell with a good terminal voltage and a good current capability.

This also applies to button cells, but the maximum current they will deliver will be less.
If you want to get the last of the energy out of a group of cells they can be used in the following circuits:

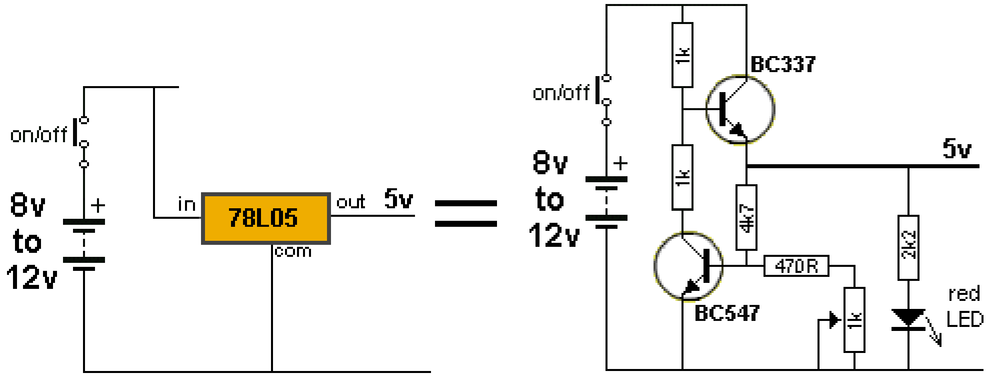

1mV drop @100mA @10v supply

10mV drop @100mA @10v supply

TESTING PIEZO DIAPHRAGMS and PIEZO BUZZERS

There are two types of piezo devices that produce a sound.
They are called **PIEZO DIAPHRAGMS** and **PIEZO BUZZERS**.
A **piezo diaphragm** consists of two metal plates with a ceramic material between. The ceramic expands and contracts when an alternating voltage is placed on the two plates and this causes the main plate to "dish" and "bow."
This creates a high-pitched sound. There are no other components inside the case and it requires an AC voltage of the appropriate frequency to produce a sound.
A **piezo buzzer** has a transistor and coil enclosed and when supplied with a DC voltage, the buzzer produces a sound.
Both devices can look exactly the same and the only way to tell them apart is by connecting a 9v battery. One device may have "+' and "-" on the case to indicate it is a piezo buzzer, but supplying 9v will make the buzzer produce a sound while the piezo diaphragm will only produce a "click."

PIEZO DIAPHRAGM PIEZO BUZZER

A piezo diaphragm will produce a click
when connected to 9v DC.
A piezo buzzer will produce a tone when
connected to a DC voltage.

How a PIEZO BUZZER WORKS

A Piezo Buzzer contains a transistor, coil, and piezo diaphragm and produces sound when a voltage is applied. The buzzer in the circuit above is a PIEZO BUZZER.

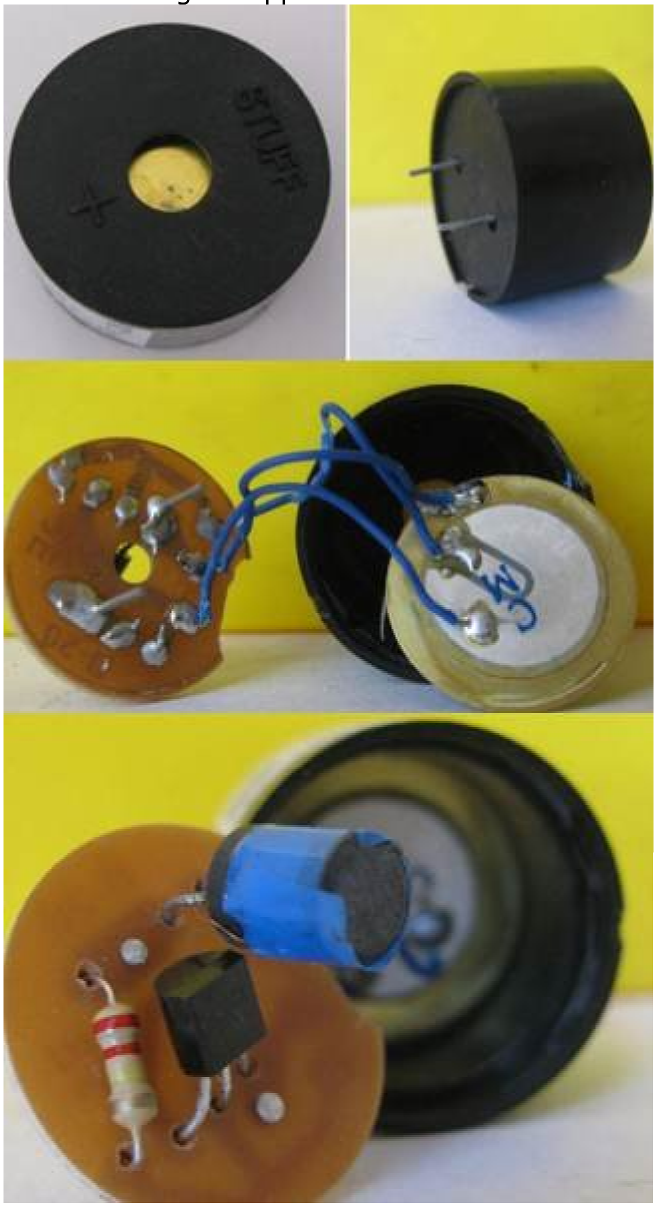

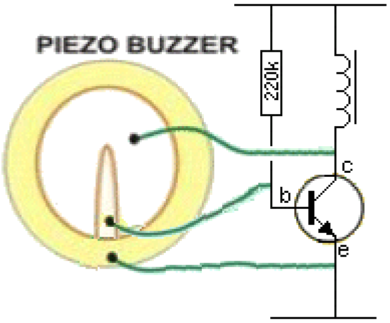

The circuit starts by the base receiving a small current from the 220k resistor. This produces a small magnetic flux in the inductor and after a very short period of time the current does not increase. This causes the magnetic flux to collapse and produce a voltage in the opposite direction that is higher than the applied voltage.
3 wires are soldered to pieces of metal on the top and bottom sides of a ceramic substrate that expands sideways when it sees a voltage. The voltage on the top surface is passed to the small electrode and this positive voltage is passed to the base to turn the transistor ON again. This time it is turned ON more and eventually the transistor is fully turned ON and the current through the inductor is not an INCREASING CURRENT by a STATIONARY CURRENT and once again the magnetic flux collapses and produces a very high voltage in the opposite direction. This voltage is passed to the piezo diaphragm and causes the electrode to "Dish" and produce the characteristic sound. At the same time a small amount is "picked-off" and sent to the transistor to create the next cycle.

TESTING A SPEAKER

A speaker (also called a **loud speaker**) has coil of wire wrapped around a magnet but it does not

touch the magnet as it is wound on a thin cardboard former so that the coil will be pulled closer to the magnet when a current flows in the coil.
When the current flows in the other direction, the coil moves away from the magnet.
The coil is called voice coil and it is connected to a sheet of thin card called a CONE and as the cone vibrates, the speaker reproduces music or noise.
Use a multimeter on a low ohms scale to read the value of resistance of the coil.
It can be as low as 2 ohms or as high as 100 ohms.
Most speakers have an 8R voice coil and the actual resistance may be slightly lower than this.
Some speakers have a resistance of 16R, 32R or 50R and even 75 ohms.
You would think putting a 16R speaker in place of 8R would reduce the sound output, but this is not always the case.
You can even use 50R or 75R and get the same performance.
This may sound amazing, but here is the reason.
The cone is deflected a certain amount due to the current flowing and the number of turns.
These two values are multiplied together to produce a value called AMP-TURNS.
If we have an 8R speaker with 80 turns and 100mA, the result is 0.1 x 80 = 8.
If we use a 16R speaker, the average current flow will be 50mA and the number of turns will be about 160. The multiplication of 0.05 x 160 = 8.
The author then tried a 50R speaker and the sound output was equal to 8R and the same with 75R speaker.
This might not apply in all situations, but the 75R speaker was slightly larger and the ticking sound form the **Metal Detector** kit was louder than using an 8R mini speaker.

To see if the cone of a speaker is undamaged, push it slightly and it will move towards the magnet. If it does not move, it is bent or damaged. If the cone is scratchy when pushed, it is rubbing against the magnet.
A cone should be able to be pushed and pulled from its rest-state. If not, it will produce a distorted sound.

TESTING A CIRCUIT

Whenever you test a circuit, the TEST EQUIPMENT puts "a load" or "a change" on it. It does not matter if the test equipment is a multimeter, Logic Probe, CRO, Tone Injector or simply a LED and resistor.
There are two things you need to know.
1. The IMPEDANCE of the circuit at the location you are testing, and
2. The amount of load you are adding to the circuit via the test equipment.

There is also one other hidden factor. The test equipment may be injecting "hum" due to its leads or the effect of your body at absorbing hum from the surroundings or the test equipment may be connected to the mains.
These will affect the reading on the test equipment and also any output of the circuit. Sometimes the test equipment will prevent the circuit from working and sometimes it will just change the operating conditions slightly. You have to be aware of this.
The last section of this eBook covers High and Low Impedance and understanding impedance is something you need to know.
The point to note here is the fact that the equipment (and the reading) can be upset by hum and resistance/capacitance effects of test equipment. This is particularly critical in high impedance and high frequency circuits.

TESTING INTEGRATED CIRCUITS (IC's)

Integrated Circuits can be tested with a LOGIC PROBE. A Logic Probe will tell you if a line is HIGH, LOW or PULSING.
Most logic circuits operate on 5v and a Logic Probe is connected to the 5v supply so the readings are accurate for the voltages being tested.
A Logic Probe can also be connected to a 12v CMOS circuit.
You can make your own Logic Probe and learn how to use it from the following link:

http://www.talkingelectronics.com/projects/LogicProbeMkIIB/LogicProbeMk-IIB.html

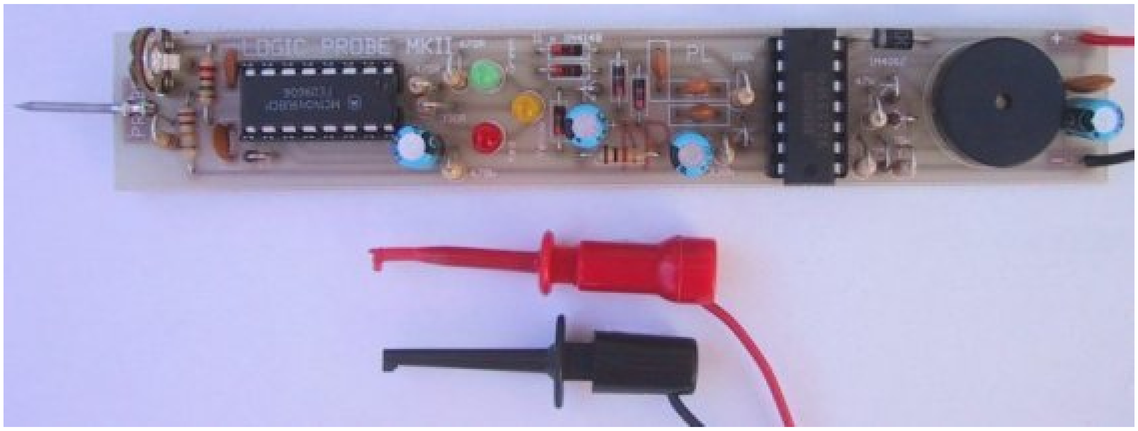
LOGIC PROBE MKII

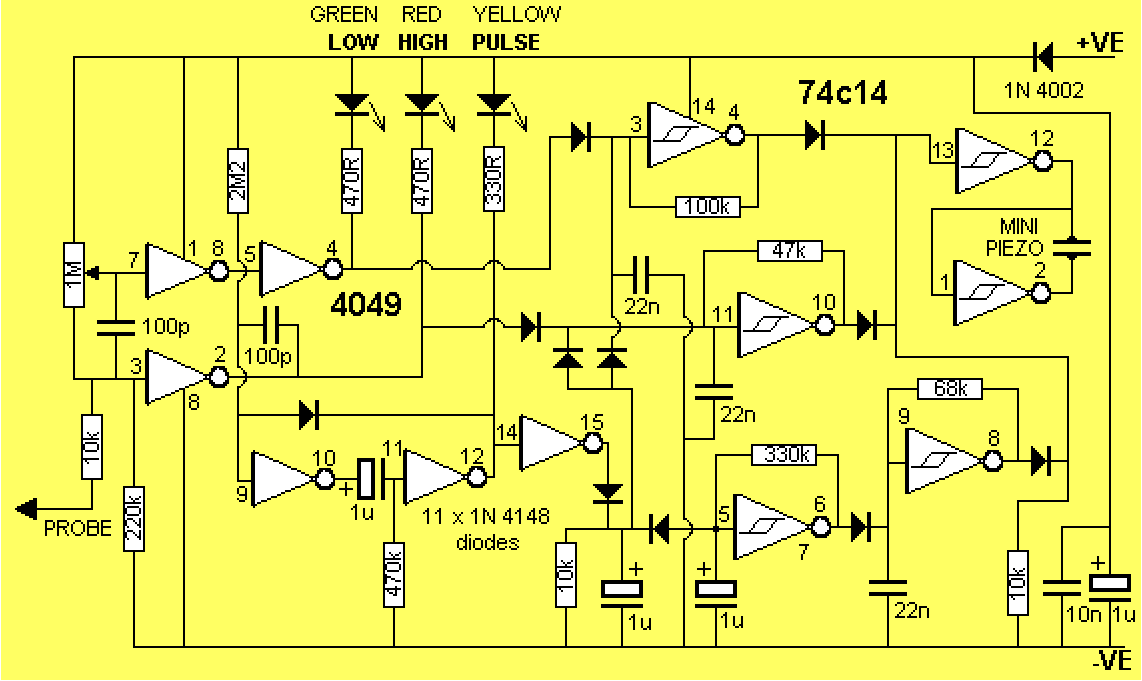
GREEN
LOW
RED
HIGH
YELLOW
PULSE
+VE
1N 4002
74c14
4049
2M2
470R
470R
330R
1M
100p
100p
100k
47k
22n
22n
22n
330k
68k
MINI
PIEZO
10k
220k
PROBE
1u
470k
11 x 1N 4148
diodes
10k
1u
1u
10k
10n
1u
-VE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

LOGIC PROBE with PULSE

This is a very simple transistor circuit to provide HIGH-LOW-PULSE indication for digital circuits. It can be built for less than $5.00 on a piece of matrix board or on a small strip of copper clad board if you are using surface mount components. The probe will detect a HIGH at 3v and thus the project can be used for 3v, 5v and CMOS circuits.

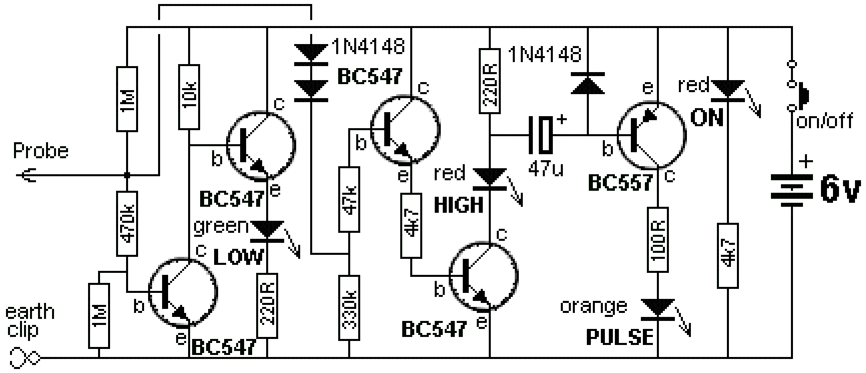
Probe
earth
clip
1M
470k
1M
10k
BC547
green
LOW
220R
BC547
1N4148
BC547
47k
330k
4k7
red
HIGH
BC547
220R
1N4148
47u
BC557
100R
orange
PULSE
red
ON
4k7
on/off
6v
b
c
e

LOGIC PROBE using CD4001 and CD4011

Here is a simple Logic Probe using a single chip. The circuits have been designed for the **CD4001** CMOS quad NOR gate and **CD4011** CMOS NAND gate. The output has an active buzzer that produces a beep when the pulse LED illuminates (the buzzer is not a piezo-diaphragm but an active buzzer containing components).

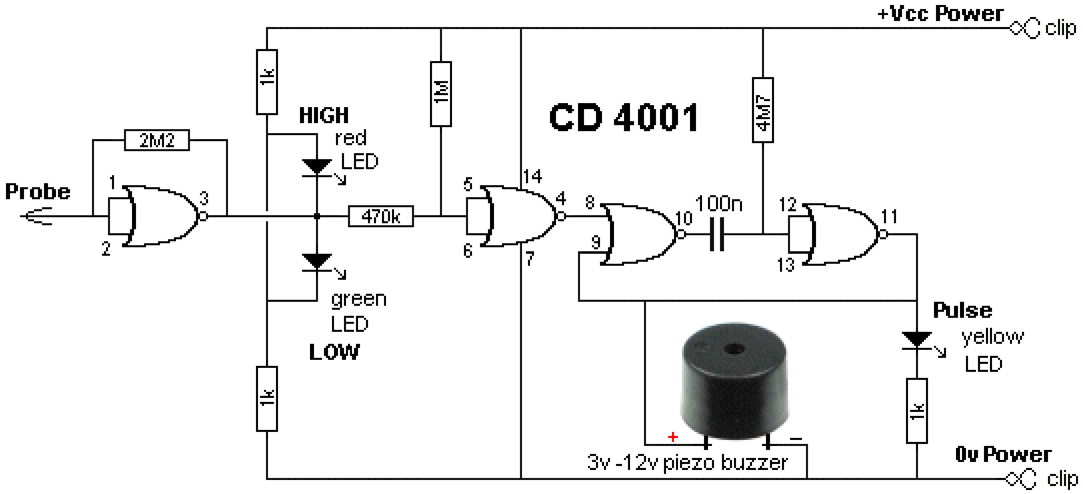
+Vcc Power
clip
CD 4001
1k
1M
4M7
2M2
HIGH
red
LED
Probe
1
2
3
470k
5
6
14
4
7
8
9
10
100n
12
13
11
green
LED
LOW
1k
Pulse
yellow
LED
1k
3v -12v piezo buzzer
+
−
0v Power
clip

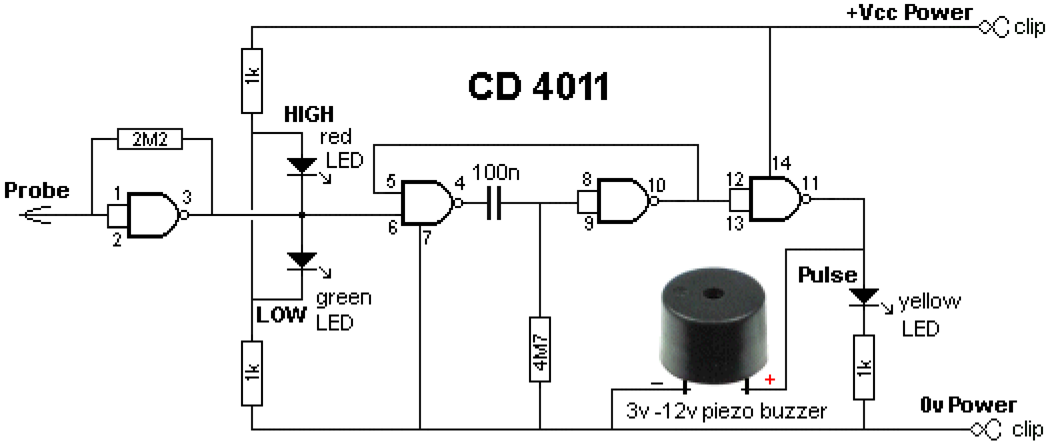
+Vcc Power
clip
CD 4011
1k
2M2
HIGH
red
LED
Probe
1
2
3
5
6
4
7
100n
8
9
10
12
13
14
11
green
LED
LOW
1k
4M7
Pulse
yellow
LED
1k
−
+
3v -12v piezo buzzer
0v Power
clip

SUPER PROBE MkII has 20 different features including a Logic Probe, capacitance tester, Inductance tester, and more.

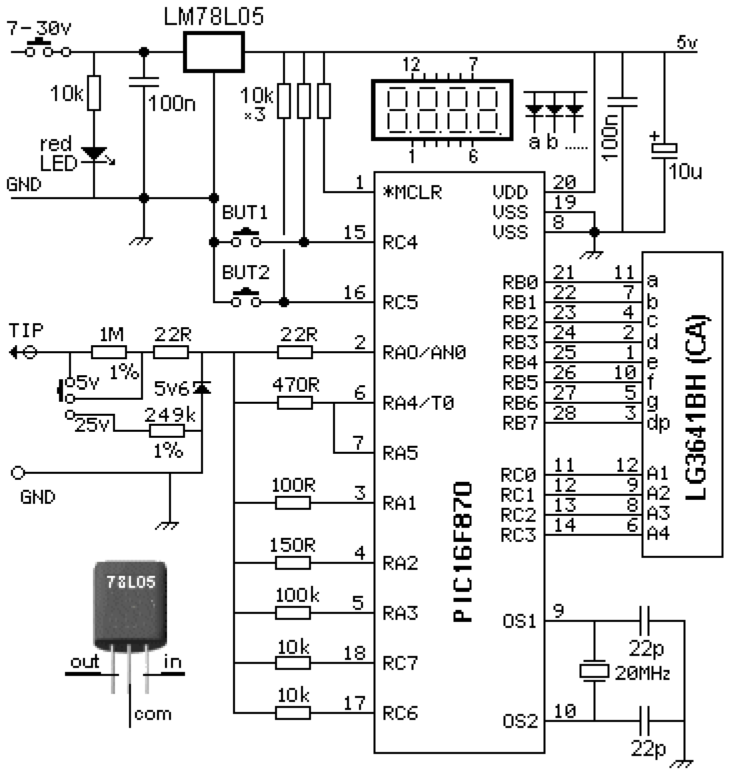
7-30v
LM78L05
5v
10k
100n
red
LED
GND
10k
×3
12
7
1
6
a b .....
100n
10u
1
*MCLR
VDD
VSS
VSS
20
19
8
BUT1
15
RC4
BUT2
16
RC5
TIP
1M
22R
22R
2
RA0/AN0
5v
1%
5v6
470R
6
RA4/T0
25v
249k
1%
7
RA5
GND
100R
3
RA1
150R
4
RA2
100k
5
RA3
10k
18
RC7
10k
17
RC6
PIC16F870
RB0
RB1
RB2
RB3
RB4
RB5
RB6
RB7
21
22
23
24
25
26
27
28
11
7
4
2
1
10
5
3
a
b
c
d
e
f
g
dp
RC0
RC1
RC2
RC3
11
12
13
14
12
9
8
6
A1
A2
A3
A4
LG3641BH (CA)
OS1
9
22p
20MHz
OS2
10
22p
78L05
out
in
com

SUPER PROBE MkII Circuit

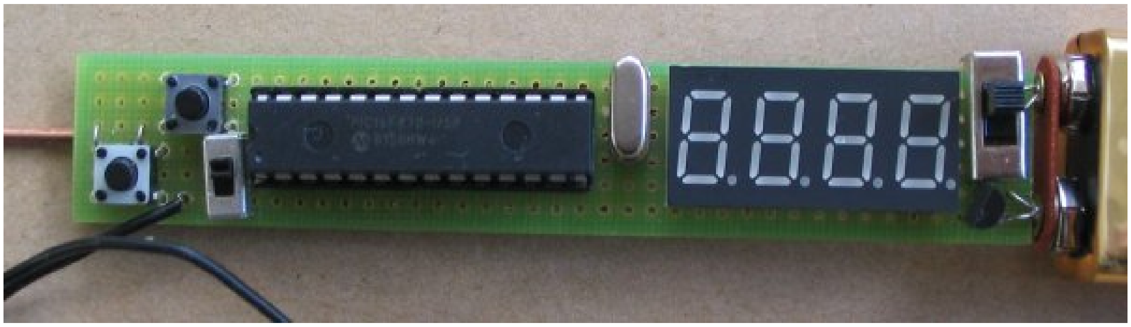

SUPER PROBE MkII

To test an IC, you need a circuit diagram with waveforms. These diagrams will show the signals and are very handy if a CRO (cathode ray Oscilloscope) is used to diagnose the problem. The CRO will reproduce the waveform and prove the circuit is functioning correctly.
A Logic Probe will just show activity and if an output is not producing a "pulse" or "activity," you should check the power to the IC and test the input line.
It is beyond the scope of this eBook to explain how to diagnose waveforms, however it is important to know if signals are entering and exiting an IC and a Logic Probe is designed for this.

SIGNAL INJECTOR

This circuit is rich in harmonics and is ideal for testing amplifier circuits. To find a fault in an amplifier, connect the earth clip to the 0v rail and move through each stage, starting at the speaker. An increase in volume should be heard at each preceding stage. This Injector will also go through the IF stages of radios and FM sound sections in TV's.

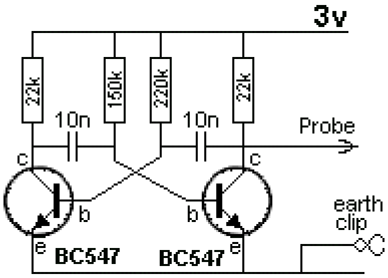

TESTING AUDIO AMPLIFIERS and AUDIO IC's

The **Super Probe MII** described above has a "noise" function and a tone function that allows you to inject a signal into an audio stage, amplifier (made from discrete components) or an audio chip, and detect the output on a speaker.
Audio stages are very difficult to work-with if you don't have a TONE GENERATOR or SIGNAL INJECTOR.
The signals are very small and not detected by a multimeter.
You can start anywhere in an amplifier and when a tone is heard, you can keep probing until the signal is not present or louder. From this you can work out which way the signal is travelling.
A Signal Injector is very handy for finding shorts and broken wires in switches, plugs, sockets and especially leads to headphones.
You can determine the gain of a stage (amplification) by probing before and after a chip or transistor and listen for the relative increase in volume from the speaker.
You can also use your finger to produce "hum" or "buzz" if a **Signal Injector** is not available.
Nearly all audio problems are plugs, sockets and cracks in the PC board, but finding them takes a lot of time and skill.

TESTING IC's - also called "CHIPS"

An Integrated Circuit is also called a "chip." It might have 8 pins or as many as 40. Some chips are ANALOGUE. This means the input signal is rising and falling slowly and the output produces a larger version of the input.
Other chips are classified as DIGITAL and the input starts at 0v and rises to rail voltage very quickly. The output does exactly the same - it rises and falls very quickly.
You might think the chip performs no function, because the input and output voltage has the same value, but you will find the chip may have more than one output and the others only go high after a number of clock-pulses on the input, or the chip may be outputting when a combination of inputs is recognised or the output may go HIGH after a number of clock pulses.

ANALOGUE CHIPS **(also see above)**
Analogue chips are AUDIO chips or AMPLIFIER chips.
To test these chips you will need three pieces of test equipment:
1. A multimeter - this can be digital or analogue.
2. A Signal Injector
3. A Mini Bench Amplifier.

The Mini Bench Amplifier is available as a kit.

PROBE
MINI BENCH AMPLIFIER
TOSHIBA
SUPER HEAVY DUTY

MINI BENCH AMPLIFIER

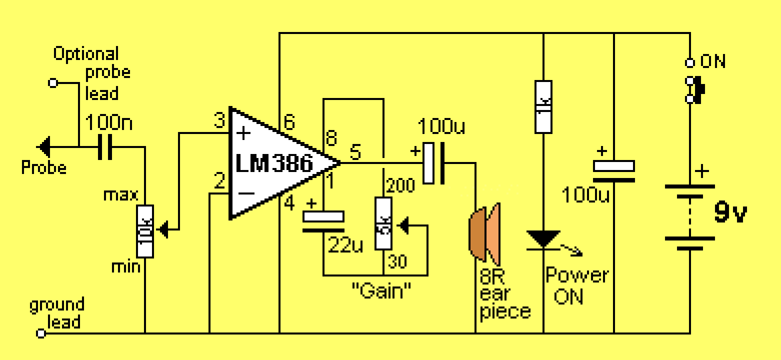
Optional probe lead
100n
Probe
max
10k
min
ground lead
3
2
+
−
LM386
6
8
5
1
4
+
22u
200
5k
30
"Gain"
100u
+
8R ear piece
1k
Power ON
+
100u
ON
+
9v

MINI BENCH AMPLIFIER CIRCUIT

Start by locating the power pin with a multimeter.
If the chip is receiving a voltage, you can use the Mini Bench Amplifier to detect an output.
Connect the Ground Lead of the Mini Bench Amplifier to 0v and touch the Probe tip on each of the pins.
You will hear faint audio on the Input pin and very loud audio on the Output pin.
If no input is detected, you can use a Signal Injector to produce a tone.

Connect the clip of the Signal Injector to 0v and the probe to the input pin of the amplifier chip. At the same time, connect the Mini Bench Amplifier to the output pin and you will hear a very loud tone.
These pieces of test equipment can also be used to diagnose an amplifier circuit constructed with individual components.
Amplifier circuits using discrete components are very hard to trouble-shoot and these pieces of test equipment make it very easy.

DIGITAL CHIPS

It is always best to have data on the chip you are testing, but if this is not available, you will need three pieces of equipment:

1. A multimeter - this can be digital or analogue.
2. A Logic Probe,
3. A logic Pulser.

Firstly test the chip to see if power is being delivered. This might be anything from 3v3 to 15v.
Place the negative lead of the multimeter on the earth rail of the project - this might be the chassis, or the track around the edge of the board or some point that is obviously 0v.
Try all the pins of the chip and if you get a reading, the chip will have "supply."
Identify pin 1 of the chip by looking for the "cut-out" at the end of the chip and you may find a small dimple below the cut-out (or notch). This is pin 1 and the "power pin" can be directly above or any of the other pins.
Next you need to now if a signal is entering the chip.
For this you will need a LOGIC PROBE.
A Logic Probe is connected to the same voltage as the chip, so it will detect a HIGH and illuminate a red LED.
Connect the Logic Probe and touch the tip of the probe on each pin.
You will not know if a signal is an input or output, however if you get two or more active pins, you can assume one is input and the other is output. If none of the pins are active, you can assume the signal is not reaching this IC.
If only one pin is active, you can assume the chip is called a CLOCK (or Clock Generator). This type of chip produces pulses. If more than two pins are active, you can assume the chip is performing its function and unless you can monitor all the pins at the same time, you don't know what is happening.
This is about all you can do without any data on the chip.

If you have data on the chip, you can identify the input(s) and output(s).
A Logic Probe on each of these pins will identify activity.
A Logic Probe has 3 LEDs. Red LED indicates a HIGH, Green indicates a LOW and Orange indicates a PULSE (activity).
Some Logic Probes include a piezo and you can hear what is happening, so you don't take your eyes off the probe-tip.
It is important not to let the probe tip slip between the pins and create a short-circuit.

LOGIC PULSER

If you have a board or a single chip and want to create activity (clock pulses), you can use a Logic Pulser. This piece of test equipment will produce a stream of pulses that can be injected into the clock-line (clock input) of a chip.
You can then use a Logic Probe at the same time on the outputs to observe the operation of the chip.

You can also use the Mini Bench Amplifier to detect "noise" or activity on the inputs and outputs of digital chips.
This only applies if the frequency is in the audio range such as scanning a keyboard or switches or a display.

This is how to approach servicing/testing in a general way. There are thousands of digital chips and if you want to test a specific chip for its exact performance, you will need to set-up a "test-bed."

REMOTE CONTROLS

There are two types of remote control - Infrared and RF. Infrared is used for short-range, line-of-sight for TV's DVD's etc.
A few faults can be fixed, but anything complex needs a new remote control.
Check the batteries and battery-contacts. See if the IR LED is illuminating by focusing it into a digital camera and looking on the screen for illumination.
The only other things are a sticky button, a worn-out button or a crack in the PC board. Water damage is generally too much work to repair.
RF remote controls for cars, garage doors etc need a second working unit to check the power output.
Here is a simple circuit that can be connected to an analog multimeter to detect the signal strength at a very close range:

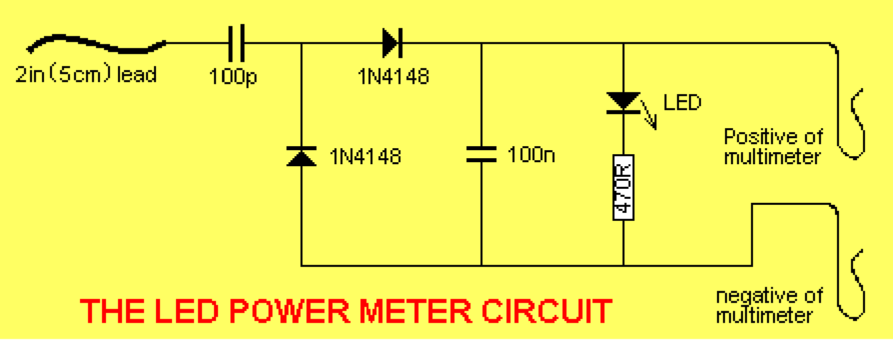
2in (5cm) lead
100p
1N4148
1N4148
100n
LED
470R
Positive of multimeter
negative of multimeter
THE LED POWER METER CIRCUIT

To hear the tone from a transmitter, the **Mini Bug Detector** circuit can be used:

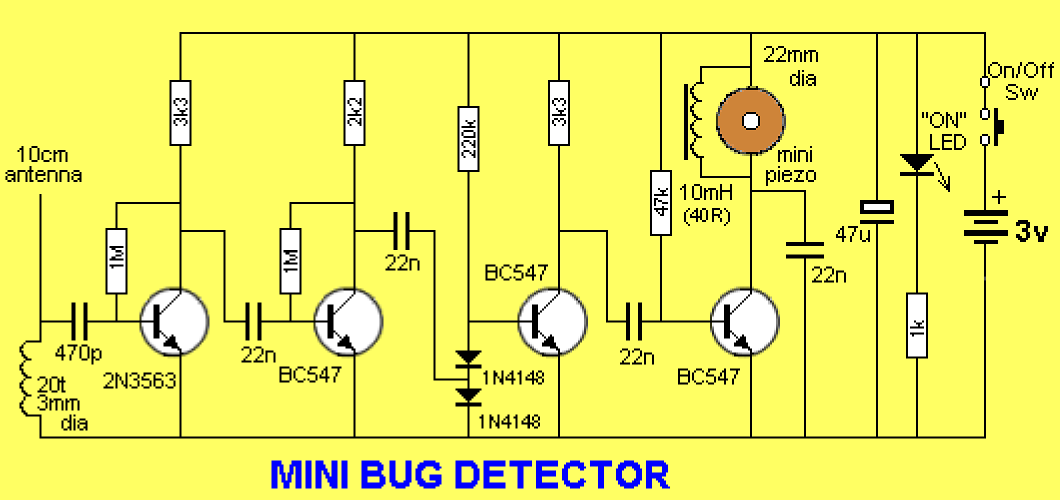
10cm antenna
3k3
1M
470p
20t 3mm dia
2N3563
2k2
1M
22n
BC547
22n
220k
1N4148
1N4148
3k3
BC547
22n
47k
10mH (40R)
22mm dia
mini piezo
BC547
22n
47u
"ON" LED
1k
On/Off Sw
+
3v
MINI BUG DETECTOR

Any further investigation requires a circuit diagram so you can work out what is actually being sent from the transmitter.
Most of the time it is a faulty switch, battery or contacts. Make sure the setting is correct on the "dip switches" and use a working unit to compare all your testing.

TESTING VOLTAGES ON (in) A CIRCUIT

There are basically two different types of circuit.
1. **ANALOGUE CIRCUIT**
An analogue circuit can also be called an AUDIO CIRCUIT and the voltages at different points in a circuit can be measured with a multimeter but the changes (the waveforms) will be quite small or changing at a rapid rate and cannot be detected by a multimeter.
You need a CRO to "see" the signals or a **Signal Injector** to inject a waveform into the circuit and hear the result on the circuit's speaker.

2. **DIGITAL CIRCUIT**
A digital circuit can also be called a "Computer Circuit" or "Logic Circuit" and some of

the voltages can be measured with a multimeter (such as supply voltages) but the "signal lines" will be be changing from HIGH to LOW to HIGH very quickly and these signals are detected with a **Logic Probe**.

Here are some circuits with details of how to test the voltages.
Most circuits do not show voltages at various different points and we will explain what to expect on each "stage."

A "STAGE"

A stage is a set of components with an input and output. A "stage" can also be called a "Building Block."
Sometimes it has a capacitor on the input and one on the output.
This means the stage is completely isolated as far as DC is concerned.
The stage has a supply (a DC supply) and it is producing its own voltages on various points on the "stage." It can only process (amplify) "AC." (signals).
Sometimes the stage can be given a name, such as small-signal amplifier, push-pull amplifier or output.
If the stage has a link or resistor connected to a previous stage, the previous stage will have a "DC effect" on the stage. In other words it will be biasing or controlling the voltages on the stage. The stage may be called a "timer" or "delay" or "DC amplifier."

It is important to break every circuit into sections. This makes testing easy. If you have a capacitor at the input and output, you know all the problems lie within the two capacitors.
In a digital circuit (no capacitors) you need to work on each IC (integrated Circuit) and test the input for activity and all the outputs.

Once you have determined if the circuit is Analogue or Digital, or a combination of both, you have to look at the rail voltage and work out the size or amplitude of the voltage or waveform.
This is done before making a test, so your predictions are confirmed.
You will need a **multimeter** (either Digital or Analogue) a **Logic Probe** and a **Signal Injector** (**Tone Generator**). An analogue meter has the advantage that it will detect slight fluctuations of voltage at a test-point and its readings are faster than a digital meter. A digital meter will produce an accurate voltage-reading - so you should have both available.

HIGH IMPEDANCE AND LOW IMPEDANCE

Every point in a circuit has a characteristic called "IMPEDANCE." This has never been discussed before in any text book. That's why it will be new to you.
In other words, every point will be "sensitive to outside noise."
An audio amplifier is a good example. If you put your finger on the active input, it will produce hum or buzz in the speaker. This is because it is a HIGH IMPEDANCE line or high impedance section of the circuit.
The same applies to every part in a circuit and when you place Test Equipment on a line for testing purposes, the equipment will "upset" the line. It may be very slight but it can also alter the voltage on the point CONSIDERABLY.
We have already mentioned (above) how a cheap multimeter can produce a false reading when measuring across a 1M resistor. That's why you need high impedance test Equipment so you do not "load" the point you are testing and create an inaccurate reading.
The word **Impedance** really means resistance, but when you have surrounding components such as diodes, capacitors, transistors, coils, Integrated Circuits, supply-voltages and resistors, the combined effect is very difficult to work out as a "resistance" and that's why we call it "Impedance."
The term "**High and Low Impedance**" is a relative term and does not have any absolute values but we can mention a few points to help you decide.
In general, the base of a transistor, FET input of an IC are classified as HIGH IMPEDANCE.
The output of these devices are LOW IMPEDANCE.

Power rails are LOW IMPEDANCE.
An oscillator circuit and timing circuit are HIGH IMPEDANCE.
A LOAD is low impedance.
And it gets tricky: An input can be designed to accept a low-impedance device (called a transducer or pick-up) and when the device is connected, the circuit becomes LOW impedance, but the input circuitry is actually high impedance.
The impedance of a diode or LED is HIGH before the device sees a voltage higher than the junction voltage and then it becomes LOW Impedance.
Impedance is one of the most complex topics however it all comes down to testing a circuit without loading it.
That's why test equipment should have an input impedance higher than 1M.

The first circuit we will investigate is the **Mini Bug Detector**, shown above and below. Points on the circuit have been labelled A, B, C etc:

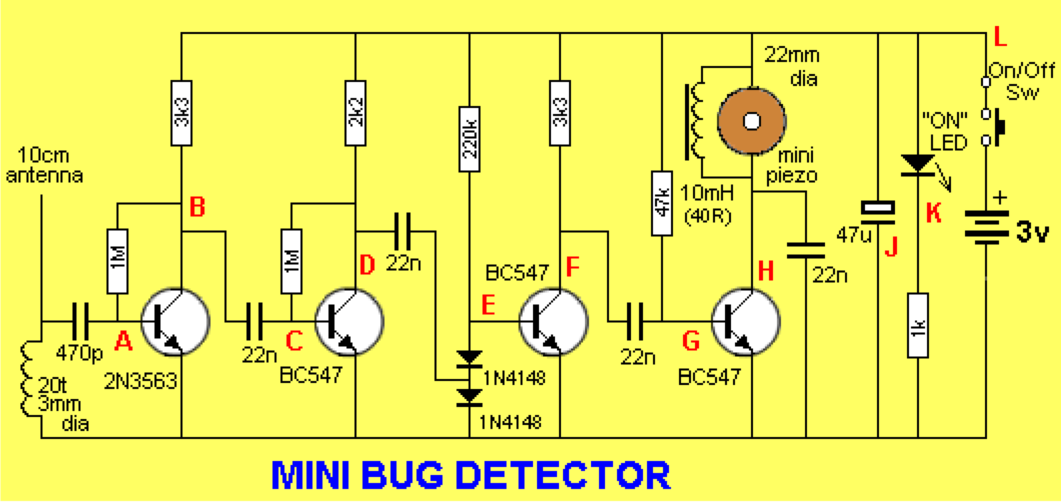

Point A - The first transistor is "self-biased" and will have 0.6v on the base. The antenna is connected to a 20 turn coil and you might think the coil will "short" the signals to earth.
But the coil and 470p capacitor form a circuit that oscillates at a high frequency when the antenna wire picks up stray signals. The coil and capacitor actually amplify the signals (see Talking Electronics website: Spy Circuits to see how a TANK CIRCUIT works) and these signals enter the base of the first transistor.
This is classified as a HIGH Impedance section because the signals are small and delicate and any loading via test equipment will kill them. The first transistor amplifies the signals about 70 times and they appear at **Point B**.

The signal passes though a 22n to **Point C** and the transistor amplifies the signal about 70 times to **point D**. **Point C** is classified as high impedance as any voltage measurement at this point will upset the biasing of the stage as a few millivolts change in base-voltage will alter the voltage on the collector considerably. **Point D** is classified as low impedance as any voltage-testing will not alter the voltage appreciably.
The output of the second stage passes through a capacitor to the join of two diodes. These two diodes are not turned on because the voltage at **Point E** can never rise above 0.7v as this is the voltage produced by the base-emitter of the third transistor.
The purpose of the two diodes is to remove background noise. Background noise is low amplitude waveforms and even though the transistor is turned on via the 220k, low amplitude signals will not be received. The third transistor works like this: It cannot be turned ON any more because any waveform from the 22n will be "clipped" by the bottom diode and it will never rise above 0.6v.
So, the only signal to affect the transistor is a negative signal - to turn it OFF.
Firstly we have to understand the voltage on the 22n. When the second transistor is sitting at mid-rail voltage, the 22n gets charged via the 2k2 and lower diode. When

the transistor gets tuned ON, the collector voltage falls and the left side of the 22n drops. The right side of the 22n also drops and when it drops 0.6v, the top diode starts to conduct and when the voltage on the 22n drops more than 0.6v the third transistor starts to turn OFF. This effect is amplified by the transistor at least 100 times and appears at **Point F.** All the voltages around the two diodes are classified as HIGH Impedance as any piece of test equipment will upset the voltage and change the output.

There are some losses in amplitude of the signal as it passes through the 22n coupling capacitors but the end result is a very high strength signal at **point G.** The 4th transistor drives a 10mH choke and the mini piezo is effectively a 20n capacitor that detects the "ringing" of the inductor to produce a very loud output.

The 22n capacitor on the collector eliminates some of the background noise. The choke and piezo form an oscillatory circuit that can produce voltages above 15v, even though the supply is 3v.

The 47n capacitor at **Point J** is to keep the supply rails "tight" (to create a LOW Impedance) to allow weak cells to operate the circuit.

The "Power-ON" LED tells you to turn the device off when not being used and **Point L** is the power supply - a low impedance line due to the 47u electrolytic.

Testing the Mini Bug Detector

To test the Mini Bug Detector, you will need a Signal Injector.

Place the Injector on **Point G** and you will hear a tone. Then go to **E**, **C** and **A**. The tone will increase in volume. If it does not increase, you have pin-pointed the faulty stage.

The next circuit is a combination of digital and analogue signals. It is a **Logic Probe**:

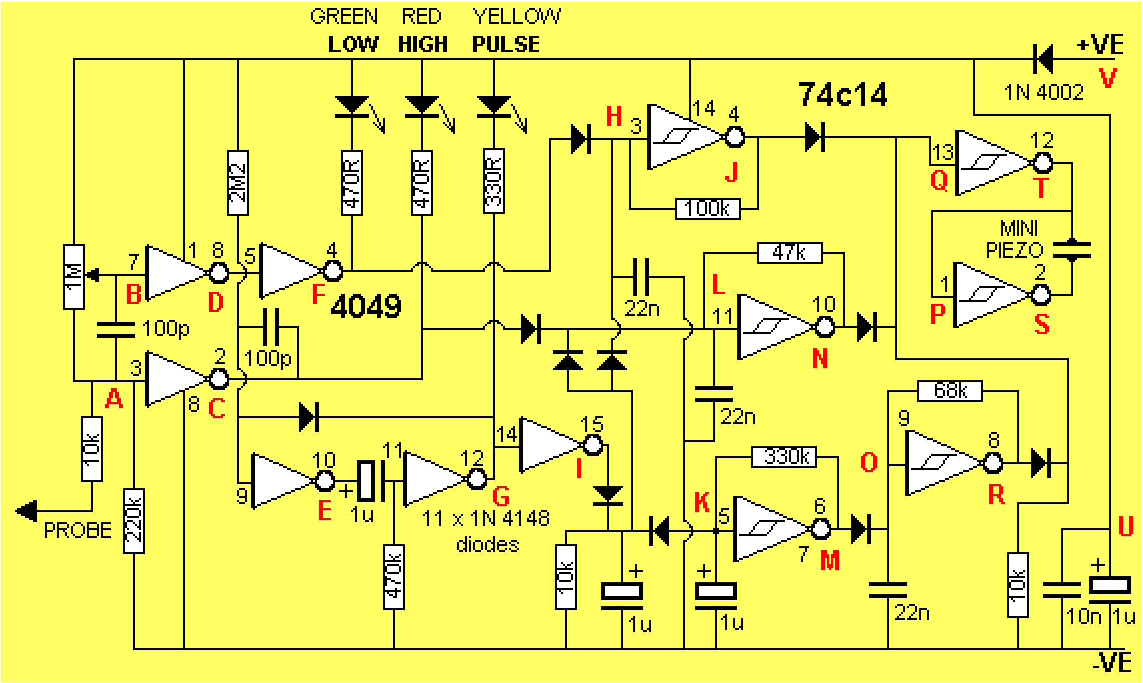

The voltage on a circuit (to be tested) is detected by the probe at **Point A** of the circuit above and the "tip" is classified as "reasonably high impedance" as it has a 220k resistor between the tip and 0v rail. The 1M reduces the impedance by about 20% but the inputs of the two inverters have no effect on the "tip" impedance as they are extremely high input-impedance devices.

The 1M trim pot is designed to put put a voltage on **point B** that is slightly higher than mid-rail so the green LED is turned off.

Point A will see a voltage below mid-rail and **point C** will be HIGH. **Point C** and **F** are low-impedance outputs.

When the tip of the probe is connected to a LOW voltage, Point B sees a LOW and Point F goes LOW to illuminate the green LED. At the same time it removes the "jamming voltage" produced by the diode between pin 4 of the 4049 and pin 3 of the

74C14 and the oscillator between **points H and J** produces a low-tone via the 100k resistor and 22n to indicate a LOW.
When the probe tip sees a HIGH, a lot more things happen.
Point C goes LOW and turns on the red LED. At the same time the 100p is in an uncharged state and the right lead goes LOW. This takes the left lead LOW as the left lead connects to a HIGH Impedance line and pin 9 goes LOW. This makes **point E** HIGH
and since the 1u is in an uncharged state, pin 11 goes HIGH. This makes **point G** LOW and the diode between pins 9 and 12 keeps pin 9 LOW and takes over from the pulse from the 100p. The yellow LED is illuminated. The 1u starts to charge via the 470k and when it is approx half-charged, pin 11 sees a HIGH and **point G** goes low. This creates the length of pulse for the yellow LED.
At the same time, **Point L** goes LOW because the "jamming diode" from pin 2 of the 4049 goes low and allows the inverter between point L and N to produce a tone for the piezo.
In addition, **Point I** goes HIGH and quickly charges a 1u electrolytic. This removes the effect of the jamming diode on pin 5 of the 74C14 and a low frequency oscillator made up of 68k and 1u between pins 5&6 turns on and off an oscillator between **points O** and **R** to get a beep. The mini piezo is driven n bridge mode via the two gates between **points QT** and **PS**.
Point U is a 1u electrolytic to reduce the impedance of the power rail and **Point V** is a protection diode to prevent damage if the probe is connected to the supply around the wrong way.

Testing the Logic Probe

You can test the Logic Probe with the simple Logic Probe with Pulse project described above. It will let you know if each point in the circuit is HIGH or LOW. You will also find out the difficulty in testing the points that are HIGH Impedance, as the Probe will upset the voltage levels and the reading may be inaccurate.

More circuits will be added here in the future.

THE VOLTAGE DIVIDER - this topic could fill a book.

You need to read lots of other sections in this eBook, including the section on measuring across a resistor with a multimeter, and high impedance circuits, to fully understand the complexities of a VOLTAGE DIVIDER CIRCUIT.
It is one of the most important BUILDING BLOCKS to understand. Even though it may consists of two components, you have to understand what is happening between these two components. You have to realise there is a voltage at their join that will be rising and falling due to one of the components changing RESISTANCE.
Sometimes you can work out the voltage at the join by using Ohm's LAW but quite often it will be impossible as it is changing (rising and falling) during the operation of the circuit.
At the beginning of this discussion we will only dealing with DC circuits and the voltage across a particular component will be due to its RESISTANCE. We are not going into any formulas, as it is very easy to measure the voltages with a multimeter set to VOLTS and you will have an accurate result.
The simplest two components in series are resistors. They always have the same resistance during the operation of a circuit and the voltage across each will not change.
In a further discussion we will cover "resistors" that change value according to the temperature. These are called THERMISTORS. And we have "resistors" that change value according to the light they receive. These are called LIGHT DEPENDENT RESISTORS (LDR's) or PHOTO RESISTORS.
A transistor that is partly or fully turned ON can be considered to be similar to a resistor.
In these 3 cases we need to measure the voltage at the join with a voltmeter as it will be a lot of work to measure the resistance and work out a value.
You can also keep a voltmeter on the joint and watch the voltage change.
Finally we have some components that produce a fixed voltage across them (or nearly fixed) and the remaining voltage is dropped across a resistor. These

components MUST have a resistor connected in series to limit the current and allow the component to pass the specified in the datasheet.

These devices include LEDs, diodes and zener diodes. A LED will have a fairly fixed voltage across it from 1.7v to 3.6v depending on the colour. A diode will have a voltage of 0.7v across it when it is connected to a voltage via a resistor. And a zener diode will have a fixed voltage across it when it is connected with the cathode to the positive rail via a resistor. The voltage across it will be as marked on the zener.

The concept of a VOLTAGE DIVIDER is very simple, but it takes a lot of understanding because both VOLTAGE and CURRENT are involved in the UNDERSTANDING-PROCESS.

Each component has a resistance and this can be measured with a multimeter. When two components are connected in series, a current will flow and a voltage will develop across each item.

More voltage will develop across the item with the higher resistance and the addition of each voltage will always equal the supply voltage.

That's the simple answer.

There is a little more involved . . . It is the word CURRENT. Here is an explanation:

Suppose we have a 1k and 2k resistor on a 12v supply. The voltage at the join will be 4v.

In other words, there will be 4v across the 1k and 8v across the 2k.

If we have a 10k and 20k resistors in series, the voltage will also be 4v at the join.

If we have a 100k and 200k resistors, the voltage will also be 4v at the join.

The voltage will be the same in all cases, but the current will be different. The current in the second case will be one-tenth and only one hundredth in the third case.

If you want to go further, place a one ohm and two ohm in series and get 4v. But the resistors will get very hot and burn out very quickly.

SOLDERING

Here are three 30-minute videos on soldering.

1. **TOOLS**
2. **Soldering components**
3. Soldering **SURFACE MOUNT** components

TESTING A MOTOR

Strictly speaking, a motor is not an electronic component, but since a website gave a useless description on testing motors, I have decided to supply the correct information.

The only REAL way to test a motor is to have two identical motors and check the torque by connecting them to a low voltage and trying to stop the shaft with your fingers. This will give you two results. Firstly it will let you know the torque of the motor.

This is the twisting effect of the shaft. There is no way to determine the torque by knowing the voltage or current.

The unknown factor is the strength of the field magnets (permanent magnets) and this determines the torque.

Secondly, feeling the shaft will let you know if the torque is even for a complete revolution.

By having two identical motors, you can see if one has a lower torque.

Almost nothing can go wrong with a motor except for the brushes. If the brushes wear out, additional resistance will be produced at the interface between the brush and commutator and this can be detected by allowing the shaft to rotate slowly and feeling the resistance as it revolves. A 3-pole motor will have three places where the strength is greatest and each should have the same feeling. A 5-pole motor will have five places of strength.

If the strength is weak or not uniform, the motor is faulty.

You cannot test a motor with a multimeter as the resistance of the armature winding is very low and if the motor is allowed to spin, the back voltage produced by the spinning, increases the reading on the meter and is false.

Micro motors have a coreless armature. This means the 3 windings for the armature are wound on a machine then bent slightly into shape and glued. A circular magnet with 3 poles is in the centre and the armature rotates around this.

This type of motor is reasonably efficient because the armature is the greatest distance from the point of rotation, and the motor reaches full RPM very quickly because the armature has very little inertia.

I have not heard of the armature-winding flying apart but if you hear any scraping noise, it may be the winding.

3-pole, 5-pole and micro motors can be found in printers, eject mechanisms of CD players, toys, RC helicopters, cars etc and rarely fail.

Motors do not work on "voltage." They actually work on CURRENT and as you increase the voltage, more current will flow and produce a stronger magnetic field (by the winding on each pole). This magnetic field will be attracted by the permanent magnet surrounding the armature and repelled by the surrounding permanent magnet, depending on where the face of the pole is, during each revolution.

If the permanent magnet is not very strong, the repulsion part of the interaction will be very weak and thus the torque will be small.

Because motors work on "current" you must have a high current available when you increase the voltage as the motor will require short bursts of high current during each revolution.

It is the combination of voltage and current (called watts) that gives the motor "strength" (torque) as well as the "strength" of the permanent magnets (called the field magnets) and the number of turns of wire on each pole (and the gauge of wire).

Basically, if a motor is hard to spin, and has 3 "hard spots" on each revolution, it will be powerful.

A 2-pole motor does not self-start and will spin in either direction. But a 3-pole motor will self-start and you can determine the direction of rotation.

A 5-pole motor has a lower RPM. It is slightly smoother in output but is not more powerful than a 3-pole version.

A motor with "permanent magnets" is called a DC motor as it will not work on AC. If the magnets are replaced with a coil, it will work on AC and it will be called a "shunt wound" motor of the field coil is connected across the same terminals as the brushes or a "series wound" motor if the field coil is in series with the armature.

TESTING COMPONENTS "IN-CIRCUIT"

You can test components while they are IN CIRCUIT, but the surrounding components will have an effect on the results.

You can get all sorts of "In-Circuit" testers. They are expensive and offer little more accuracy than a multimeter.

In-Circuit testing with a multimeter can give you the same results as a tester.

All you have to do is turn the project ON and use a multimeter (set to voltage) to determine the voltage at various points. It is best to have a circuit of the equipment so you can what to expect at each point.

Only major departures from the expected can be located in this way.

Obviously the first thing to look for is burnt-out components. Then feel components such as transistors for overheating.

The look for electrolytics that may be dry. Sometimes these have changed colour or are slightly swollen.

If they are near hot components, they will be dry.

For the cost of a few dollars I change ALL THE ELECTROLYTICS in some pieces of equipment, as a dry electrolytic is very difficult to detect.

Testing a transistor "in-circuit" is firstly done with the supply ON. That's because it is quicker.

Measure the voltage between ground and collector.

In most cases you should get a voltage of about half-rail. If it is zero, or close to rail voltage, you may have a problem.

Turn off the supply and use the multimeter on low-ohms to measure all six resistances between the leads.

A low resistance in both directions on two leads will indicate a fault.

Resistors almost NEVER go "HIGH." For instance, a 22k will never go to 50k.

However a low-value resistor will "burn-out" and you will read the value of the surrounding components.

Don't forget, some low-value resistors are designed to burn-out (called fusible resistors) and anytime you find a damaged low-value resistor, you will need to look for the associated semiconductor.

You can replace the resistor quickly and turn the circuit ON to see it burn out again.

Alternatively you can trace though the circuit and find the shorted semiconductor.

It's always nice to "see the fault" then "fix the fault."

Sometimes a transistor will only break-down when a voltage is present, or it may be influenced by other components.

When the piece of equipment is turned OFF, you can test for resistance values. The main thing you are looking for is "dry joints" and continuity. Dry joints occur around the termination of transformers and any components that get hot. Rather than wasting time checking for dry joints, it is better to simply go over the connections with a hot iron and fresh solder.

You may need to check the continuity of a track (trace) and it may go from one side of the PC board to the other.

Use a multimeter set to low-ohms and make sure the needle reads "zero-ohms."

It is very dangerous to do any testing on a project using a multimeter set to "amps" or "milliamps."

You cannot test "current flowing through a component" by placing the probes across a component. You will simply over-load the rest of the circuit and create a problem.

To find out if current is flowing though a circuit or a low-value resistor, turn the project ON and measure the voltage either across the component or the voltage on one end then the other.

A voltage-drop indicates current is flowing.

That's about it for testing "in-circuit." Use the rest of this eBook to help you with diagnosis.

Don't think an IN-CIRCUIT COMPONENT TESTER is going to find a fault any faster than a multimeter. They all use a multimeter principle.

SHORT CIRCUIT

Nearly every component can fail and produce an effect called a SHORT CIRCUIT.

This basically means the component takes more current than normal and it may fail completely or simply take more current and the operation of the circuit may be reduced only a small amount.

The resistance of the component may reduce a very small amount but this may have a very large effect on the operation of the circuit.

For instance, two turns in the horizontal or vertical winding of a yoke on the picture tube or monitor may arc and weld together and reduce the size of the picture on the screen, but measuring the winding will not detect the difference in resistance.

The same with the windings on a motor and a short between two winding in a transformer.

If the "short" is between two near-by turns, the change in resistance will be very small. If the "short" is between to different layers, the resistance will be reduced and it may be detected.

When a "short" occurs, the winding turns into a transformer. To be exact, an AUTO-TRANSFORMER.

In the following diagram you can see a normal winding in fig A:

A

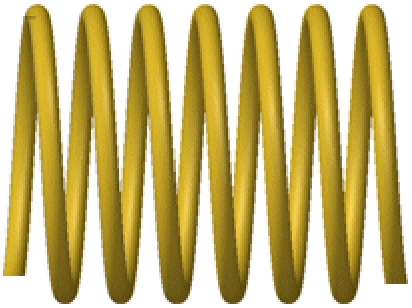

B

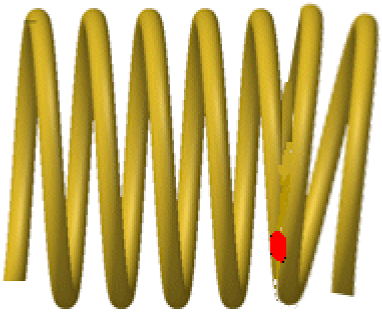

C

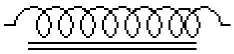

D

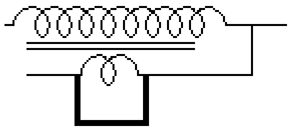

Fig B shows two turns touching each other and if the wire is enamelled, the coating has been damaged so the copper wire from the two turns is touching. This is called a SHORTED TURN.
In fig C you can see two turns touching.
In fig D the shorted-turn has been moved to the other side of the symbol to show the effect it has on the operation of the winding.
The shorted-turn is exactly like the secondary of a transformer with a "jumper" across the output.
This will produce a very high current in the secondary.
A very high current flows through the shorted turn and this changes the operation of the rest of the winding.

1. In most cases a SHORT CIRCUIT can be detected by feeling the additional heat generated by the component.
2. Next, turn off the supply and measure the resistance of the component. If it is lower than expected, the component will be faulty.
3. Next, measure the voltage across the component. If it is lower than normal, the component will be faulty.
4. Next, measure the current taken by the component. If it is higher than normal, the component will be faulty.
5. If the component is an inductor, such as a motor, coil or transformer, you can use an inductance meter. Compare a good winding with a faulty winding. Sometimes the fault will disappear because an arc develops across the fault when the component is operating.

INTERNAL AND EXTERNAL SHORTS

An **internal short** refers to two windings shorting together and the winding has a very high resistance between the winding and the frame on which it is wound. An external short refers to a winding shorting to the frame of the component - such as one of the armature windings shorting to the metal core, around which the wire is

wound.
This may not be important unless another winding shorts to the metal frame and creates "**inter winding**" problems (**inner winding** problems is within the same winding).

The opposite to a short circuit is an OPEN CIRCUIT.

This is generally a broken lead or contact or a wire that has "burnt-out" or been "eaten-away" by acid attack or galvanic action by water and voltage (current).
1. No current will flow when an OPEN CIRCUIT exists.
2. The voltage on each end of the OPEN CIRCUIT will not be the same.
3. Measure the current across the OPEN CIRCUIT and determine if excess current is flowing.
4. Join the two ends of the OPEN CIRCUIT and see if the circuit operates normally.

> HEATSINKS

This is not an electronic component but it can certainly affect the operation of a circuit.
If you cannot hold your fingers on a heatsink, it is getting too hot. This is because the actual location where the heat is being generated is much hotter than the part you are touching.
Transistors and IC's can withstand a high temperature but if they go above this temp, they BLOW UP.
They also have a shorter life when operating at a high temperature.
The secret to a good heatsink is called an INFINITE HEATSINK.
This is the metal frame of a case.
There are lots of charts and data on choosing a heatsink but they don't take into account two factors:
Sometimes a circuit takes a very high current for a short time and this creates a high temperature gradient. This will cause the transistor to get very hot and fail.
The solution is to have two or more transistors in parallel to separate the "heat spots."
The second problem with designing a heatsink is the unknown location of the heatsink and the air-flow. Products placed on a shelf or in a cupboard will get very little air-flow.
Remember: some transistors are mounted on thermal insulators. This means the transistor will have a voltage on it but the heatsink will be zero voltage.
The temperature of the transistor will be MUCH HIGHER than the heatsink under the transistor and the transfer of the heat from the transistor to the heatsink will be very slow. This can be the cause of the transistor failing. Sometimes the transistor will fail because insulation is high temp plastic and it gets brittle. The plastic can carbonise and leak and sometimes a voltage can flash through the insulator. Some amazing things have happened under these transistors and you may need to pull it apart and replace all the insulation.
Finally, feel the heatsink after 15 minutes and feel right up to the transistor. If you cannot touch the transistor, increase the thickness of the heatsink or use two transistors to dissipate the heat.
To design a heatsink, you have to have some idea of the size of a heatsink for the application.
Charts and data can send you in the wrong direction.
Start with a heatsink twice the recommended size and feel the temp after 15 minutes. Put the project in a cupboard and see how the temperature rises.
If possible, connect the heatsink to the metal case to get added dissipation and if you include fan-cooling, remember the fan will eventually gather dust and reduce its efficiency.
It is very difficult to explain how heat passes through a mica washer or plastic washer, but if the transistor has a copper base, the heat transfer has a value of 400. For aluminium it is 200. If it is steel, the transfer has a value of 50. For a mica sheet it is 1 and for plastic it is 0.1
Even though the sheet is very thin, the transfer is a lot less than metal-to-metal transfer.

Most references state the temperature difference is about one degree C for each watt of heat generated by the transistor.
Don't believe anything you read.
Feel the temperature yourself and if you cannot hold your finger on the transistor, fix the problem.

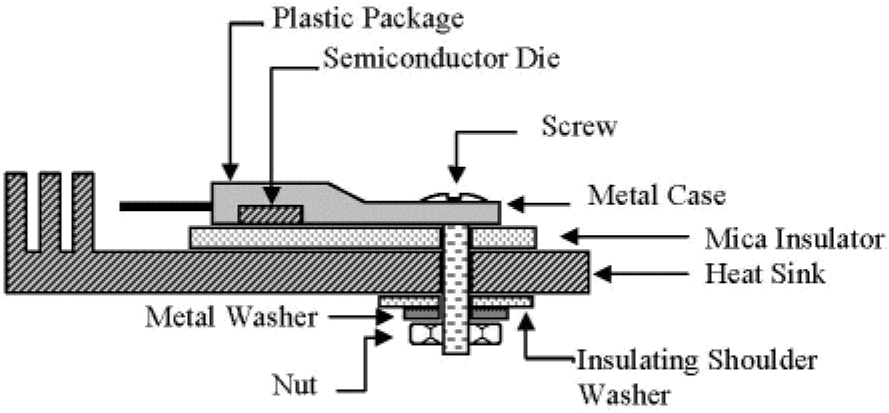
Plastic Package
Semiconductor Die
Screw
Metal Case
Mica Insulator
Heat Sink
Metal Washer
Nut
Insulating Shoulder Washer

In the end, use a heatsink 50% larger than recommended.

THE END

This is not the full story to learning about servicing. It is just the beginning.
We have only covered the simplest tests and shown how 90% of faults can be found by checking voltages, waveforms and looking for obvious things such as burnt out components, cracks in PC boards.
The author has fixed over 35,000 TV's, radios, stereos, VCRs and all those things that were on the market 30 years ago.
Things have not changed. It's just that some repairs cost nearly as much as buying a new product and half the customers opt for dumping a faulty item and buying the latest "flat screen" version. That's why you have to get things through the workshop as fast and as cheaply as possible, to make a living.

If you want any more devices added to this list, email **Colin Mitchell**.

To help with understanding how a transistor circuit works, we have produced an eBook: **The Transistor Amplifier.** It covers a whole range of circuits using a transistor.

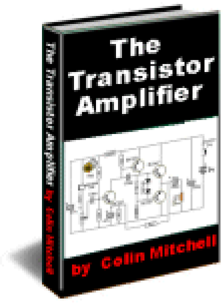
The Transistor Amplifier
by Colin Mitchell
The Transistor Amplifier by Colin Mitchell

Not copyright by Colin Mitchell
You can use any of this material. Please pass this eBook to your friends and let them know that everything on the web is FREE. I have looked at all the "Pay Sites" and found the information they "sell" is available on the web at NO COST.
Nearly all text books are also downloadable for free on "Download.com" etc and when you see a used copy of a $74.00 textbook on Amazon for $12.00 you realise many users have already discarded their copy. A good textbook never gets thrown out or sold for $12.00!!!

See the enormous amount of information on Talking Electronics website

CIRCUIT SYMBOLS by TALKING ELECTRONICS

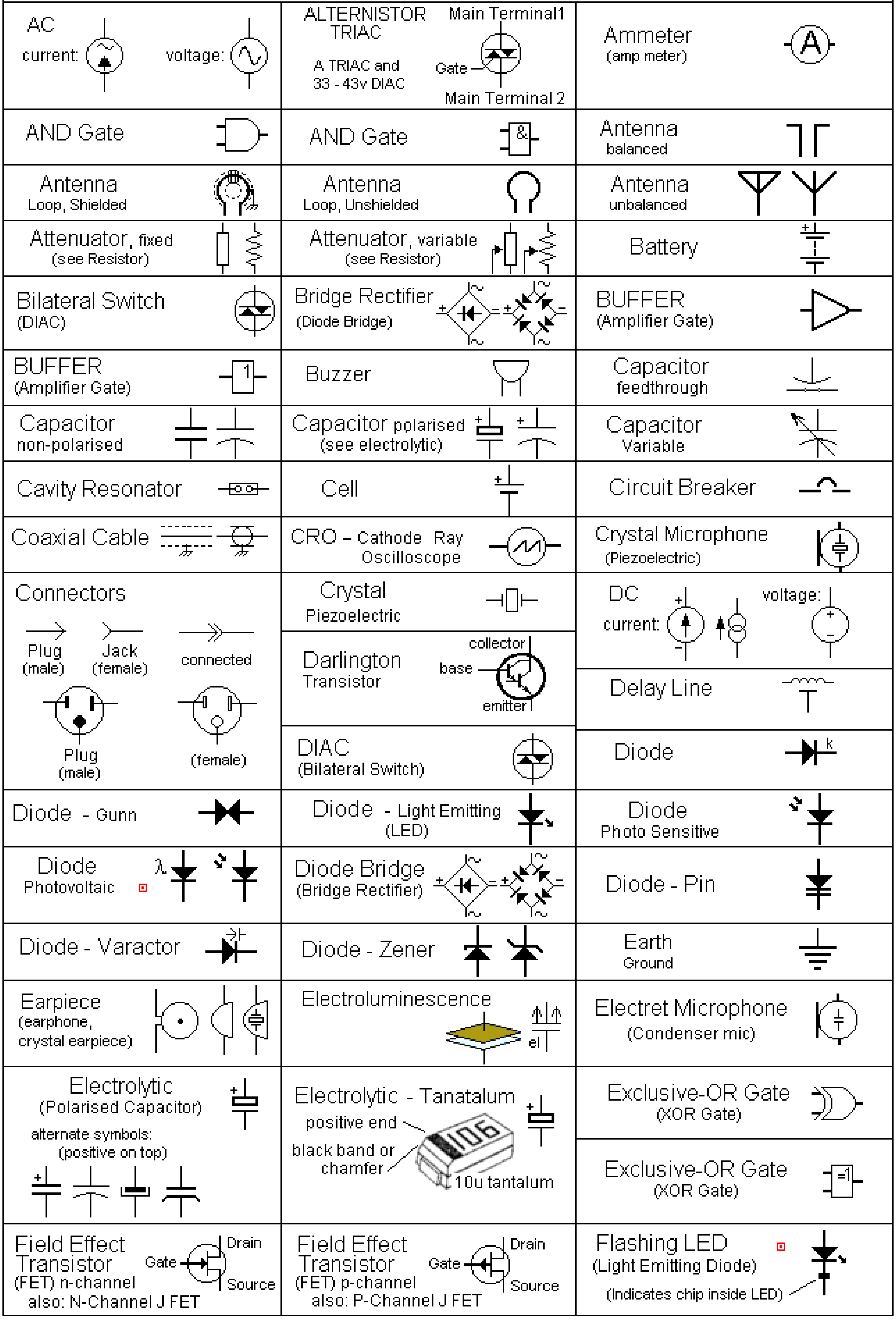
AC
current:
voltage:
ALTERNISTOR
TRIAC
A TRIAC and
33 - 43v DIAC
Main Terminal1
Gate
Main Terminal 2
Ammeter
(amp meter)
A
AND Gate
AND Gate
&
Antenna
balanced
Antenna
Loop, Shielded
Antenna
Loop, Unshielded
Antenna
unbalanced
Attenuator, fixed
(see Resistor)
Attenuator, variable
(see Resistor)
Battery
+
Bilateral Switch
(DIAC)
Bridge Rectifier
(Diode Bridge)
+
=
BUFFER
(Amplifier Gate)
BUFFER
(Amplifier Gate)
1
Buzzer
Capacitor
feedthrough
Capacitor
non-polarised
Capacitor polarised
(see electrolytic)
+
Capacitor
Variable
Cavity Resonator
Cell
+
Circuit Breaker
Coaxial Cable
CRO – Cathode Ray
Oscilloscope
Crystal Microphone
(Piezoelectric)
Connectors
Plug
(male)
Jack
(female)
connected
Plug
(male)
(female)
Crystal
Piezoelectric
DC
current:
voltage:
+
-
Darlington
Transistor
collector
base
emitter
Delay Line
DIAC
(Bilateral Switch)
Diode
k
Diode - Gunn
Diode - Light Emitting
(LED)
Diode
Photo Sensitive
Diode
Photovoltaic
λ
Diode Bridge
(Bridge Rectifier)
+
=
Diode - Pin
Diode - Varactor
Diode - Zener
Earth
Ground
Earpiece
(earphone,
crystal earpiece)
Electroluminescence
el
Electret Microphone
(Condenser mic)
Electrolytic
(Polarised Capacitor)
alternate symbols:
(positive on top)
+
Electrolytic - Tanatalum
positive end
black band or
chamfer
106
10u tantalum
+
Exclusive-OR Gate
(XOR Gate)
Exclusive-OR Gate
(XOR Gate)
=1
Field Effect
Transistor
(FET) n-channel
also: N-Channel J FET
Gate
Drain
Source
Field Effect
Transistor
(FET) p-channel
also: P-Channel J FET
Gate
Drain
Source
Flashing LED
(Light Emitting Diode)
(Indicates chip inside LED)

Ferrite Bead 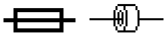	Fuse 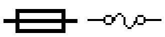	Galvanometer 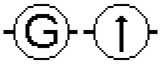
Globe 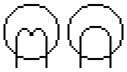	Ground Chassis 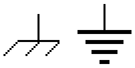	Ground Earth 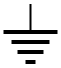
Heater (immersion heater) (cooker etc) 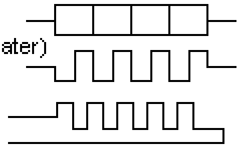	IC Integrated Circuit 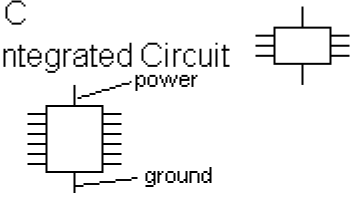	Inductor Air Core 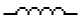
Headphone 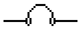		Inductor Iron Core or ferrite core 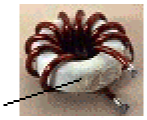 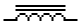
Inductor Tapped 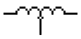	Inductor Variable 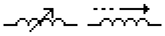	Integrated Circuit 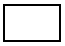
Inverter (NOT Gate) 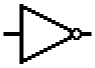	INVERTER (NOT Gate) 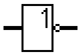	
Jack Co-axial 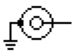	Jack Phone (Phone Jack) 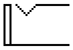	Jack Phone (Switched) 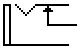
Jack Phone (3 conductor) 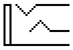	Key Telegraph (Morse Key) 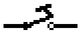	Lamp Incandescent 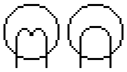
Lamp - Neon 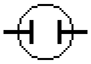	LASCR (Light Activated Silicon Controlled Rectifier) 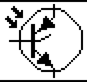	LDR (Light Dependent Resistor) 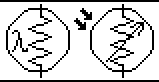
LASER diode 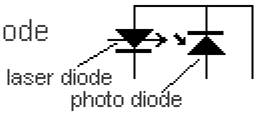	Light Emitting Diode (LED) 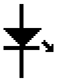	Light Emitting Diode (LED - flashing) 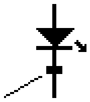 (Indicates chip inside LED)
Mercury Switch 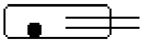	Micro-amp meter (micro-ammeter) 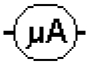	Microphone (see Electret Mic) 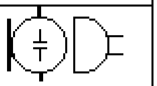
Microphone (Crystal - piezoelectric) 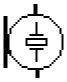	Milliamp meter (milli-ammeter) 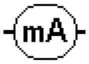	Motor 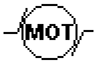
NAND Gate 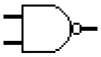	NAND Gate 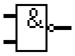	Nitinol wire "Muscle wire" 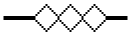
Negative Voltage Connection 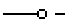	NOR Gate 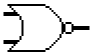	NOR Gate 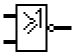
NOT Gate Inverter 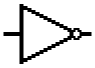	NOT Gate Inverter 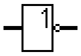	Ohm meter 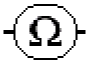
Operational Amplifier (Op Amp) 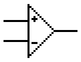	Optocoupler (Transistor output) 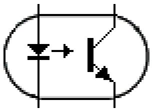	Opto Coupler (Opto-isolator) 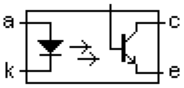 Photo-transistor output
Optocoupler (Darlington output) 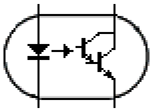	Opto Coupler (Opto-isolator) 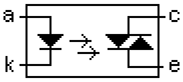 TRIAC output	OR Gate 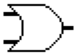
OR Gate 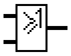	Oscilloscope see CRO 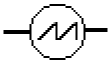	Outlet (Power Outlet) 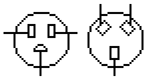
Piezo Diaphragm 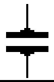	Photo Cell (photo sensitive resistor) 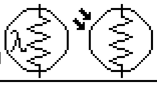	Photo Diode 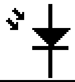
Photo Darlington Transistor 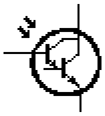	Photo FET (Field Effect Transistor) 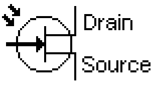	Photo Transistor 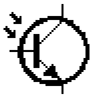

Photovoltaic Cell (Solar Cell) 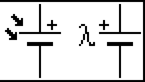	Piezo Tweeter (Piezo Speaker) 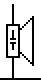	Positive Voltage Connection 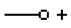
Potentiometer (variable resistor) 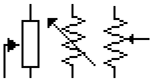	Programmable Unijunction Transistor PUT 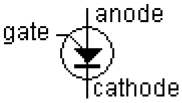 gate anode cathode	Rectifier Silicon Controlled (SCR) 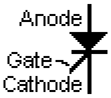 Anode Gate Cathode
Rectifier Semiconductor 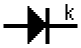	Reed Switch 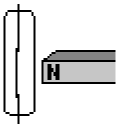	Relay - spst 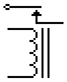
Relay - spdt 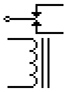	Relay - dpst 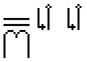	Relay - dpdt 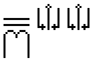
Resistor Fixed 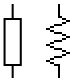	Resistor Non Inductive 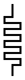	Resistor preset 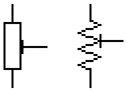
Resistor variable 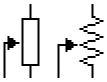	Resonator 3-pin 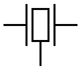	RFC Radio Frequency Choke 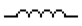
Rheostat (Variable Resistor) 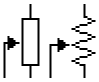	Saturable Reactor 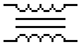	Schmitt Trigger (Inverter Gate) 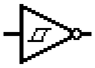
Schottky Diode 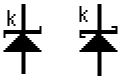 (also Shottky) Low forward voltage 0.3v Fast switching also called Schottky Barrier Diode	Shielding - - - - - Signal Generator 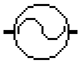	Shockley Diode 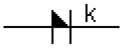 4-layer PNPN device Remains off until forward current reaches the forward break-over voltage.
Silicon Bilateral Switch (SBS) 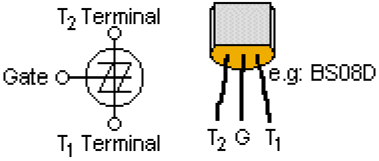 T$_2$ Terminal Gate T$_1$ Terminal e.g: BS08D T$_2$ G T$_1$	Silicon Unilateral Switch (SUS) 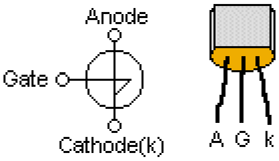 Anode Gate Cathode(k) A G k	Silicon Controlled Rectifier (SCR) 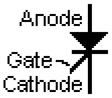 Anode Gate Cathode Solar Cell 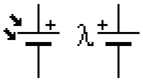
Surface Mount 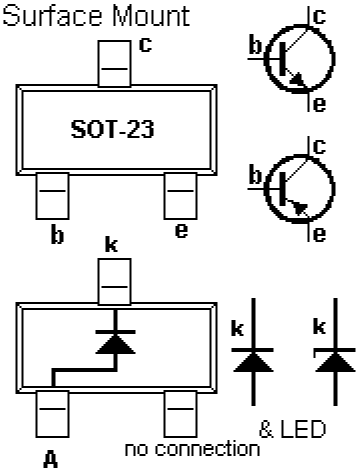 c SOT-23 b e c b e k k k A no connection & LED	Switch - spst Switch - spdt Switch - dpst 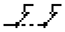 Switch - dpdt 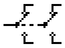 Switch - mercury tilt switch 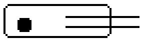 Spark Gap 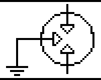	Switch - process activated normally open: normally closed: 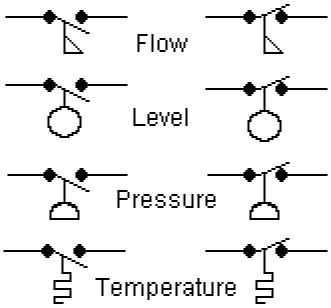 Flow Level Pressure Temperature Speaker 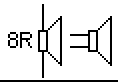
Switch - push (Push Button) 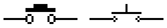	Switch - push off (used in alarms etc) 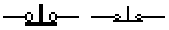	Switch - Rotary 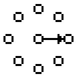
Test Point 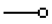 Thermal Probe 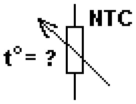 **NTC:** as temp rises, resistance decreases	Thyristors: 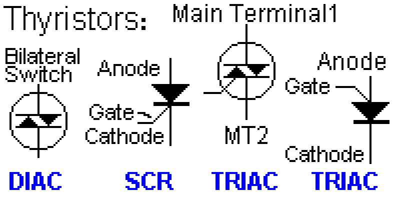 Main Terminal1 Bilateral Switch Anode Gate Cathode MT2 Anode Gate Cathode DIAC SCR TRIAC TRIAC	Thermocouple 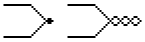 Tilt switch mercury 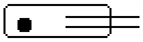 Touch Sensor 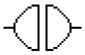
Transformer Air Core 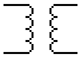	Transformer Iron Core 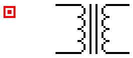	Transformer (Tapped Primary/Sec) 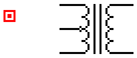

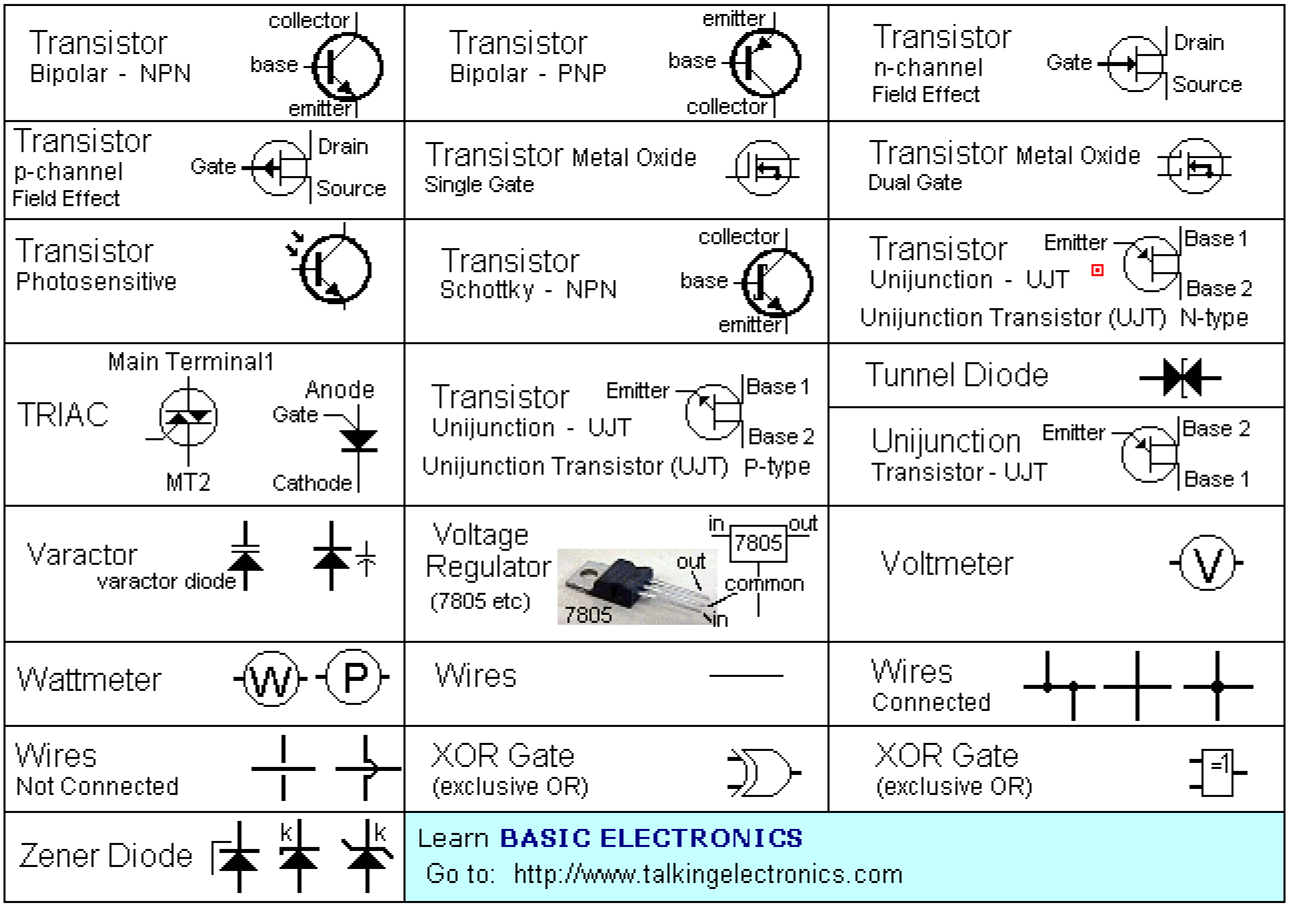

Transistor Bipolar - NPN collector base emitter	Transistor Bipolar - PNP emitter base collector	Transistor n-channel Field Effect Gate Drain Source
Transistor p-channel Field Effect Gate Drain Source	Transistor Metal Oxide Single Gate	Transistor Metal Oxide Dual Gate
Transistor Photosensitive	Transistor Schottky - NPN collector base emitter	Transistor Unijunction - UJT Emitter Base 1 Base 2 Unijunction Transistor (UJT) N-type
Main Terminal1 TRIAC MT2 Anode Gate Cathode	Transistor Unijunction - UJT Emitter Base 1 Base 2 Unijunction Transistor (UJT) P-type	Tunnel Diode Unijunction Transistor - UJT Emitter Base 2 Base 1
Varactor varactor diode	Voltage Regulator (7805 etc) 7805 in out 7805 out common in	Voltmeter V
Wattmeter W P	Wires	Wires Connected
Wires Not Connected	XOR Gate (exclusive OR)	XOR Gate (exclusive OR) =1
Zener Diode k k	Learn **BASIC ELECTRONICS** Go to: http://www.talkingelectronics.com	

More chapters of this eBook on: Talking Electronics.com

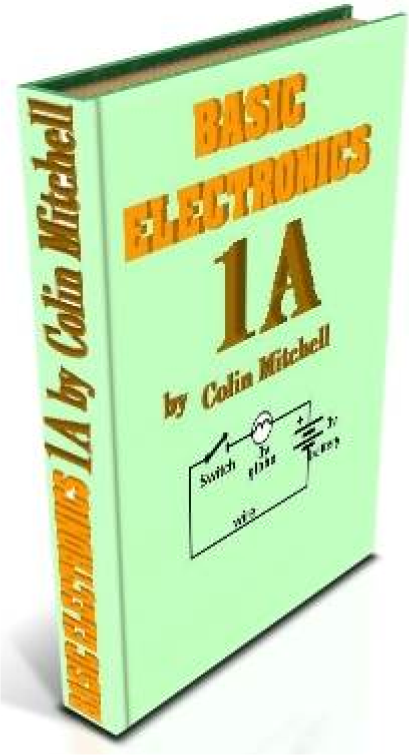
BASIC ELECTRONICS 1A
by Colin Mitchell
BASIC ELECTRONICS 1A by Colin Mitchell
Switch
wire

For any enquiries email Colin Mitchell

BASIC ELECTRONICS

(this is the Basic Electronics section i.e. Page 1)
(Chapters 1 and 3 are available as .pdf)

Quick Quiz - to see how much you know
Encyclopedia of Components - this is excellent !!!

Page 1: Basic Electronics (this page) - .pdf (1.2MB) or .zip
The capacitor - how it works
The Diode - how the diode works
Circuit Symbols - EVERY Circuit Symbol
Soldering - videos
Page 2: The Transistor
- PNP or NPN Transistor TEST
Page 2a: The 555 IC
The 555 - 1
The 555 - 2
The 555 - 3
The 555 TEST
Page 3: The Power Supply download as .pdf (900kB)
3a: - Constant Current
3b: - Voltage Regulator
3c: - Capacitor-fed Power Supply
Page 4: Digital Electronics

    4a: - Gates Touch Switch Gating
    4b: - The DELAY CIRCUIT
Page 5: Oscillators
Page 6: Test - Basic Electronics (50 Questions)
Page 7: The Multimeter - using the Multimeter
Page 8: Constructing a Project
Page 9: Inductance

Remember: the animations do not work in .pdf
the site is being constantly updated

INDEX

Active HIGH
AND Gate
Battery
Battery - Internal Resistance
Battery Boost
Battery Booster - for a flat battery
Battery Current
Blocking Diode
Breadboard
Bypass Diode
Capacitors
Capacitors in Parallel
Capacitors in Series - non-polar
Cell
Charging A Capacitor
Capacitor Charging
Capacitor Charging - more details
Capacitors on Volume Control
Capacitor - values
Characteristic Voltage drop - LED
Chassis
Circuit
Circuit - Drawing
Component - Symbols more
Current - milliamp
Current Divider Circuit
Cutoff and Saturation
Damage A LED
DC Current
DC Voltage
Digital States
Diode
Diode - how a diode works
Drawing A Circuit
Earth Rail Earth Rail
NAND Gate
Negative Voltage
NOR Gate
Non-polar Capacitor (electrolytic)
NOT Gate
Ohm's Law
Passive Components
Piezo
Plant Watering Circuit
PNP Transistor
Power and Energy
Power Rail
Potentiometer
Protection Diode
Questions
Regeneration
Resistance - Multimeter
Resistor Colours
Resistors In Parallel
Resistors in Series
Resistor Wattage
Robot Man
Robot Man Animation
Schematic
Slide Switch
Short Circuit
Soldering
Soldering Iron
Speaker
Supply Rail
Switch
Symbols more - components
Tap A Piezo
Test A Cell
Testing A LED

Earth Return
Electret Microphone
Electrolytics
Energy in a cell
Flashing LED Flashing LED
Forward Voltage Drop
Gates with Diodes
Globe
Ground
How a Transistor turns ON
How to Solder
LED
LED flat spot (cathode)
LEDs in Parallel
LEDs in Series
LED Resistors
LED Voltages
Light Dependent Resistor
Load Resistor
Memory Cell
Microphone - Electret
Milliamp - current
Multimeters
Time Delay
Time Delay Animation
Toggle Switch
Tolerance - resistors
Transformer Feedback
3 - Transistor Circuit
Transistor NPN and PNP
Transistor - how it works
Transistor Tester
Turning ON a Transistor
Variable Capacitors
Variable resistor - potentiometer
Voltage
Voltage - Analogue Multimeter
Voltage - Digital Multimeter
Voltage Divider Circuit
Wattage of a Resistor
White LED Voltage
Wire
50 Questions

KIT OF PARTS

Talking Electronics supplies a kit of parts that can be used to build the majority of the circuits in this eBook.
The kit costs $15.00 plus postage.

Kit for Transistor Circuits - $15.00

BUY NOW

A kit of components to make many of the circuits described in this eBook is available for $15.00 plus $7.00 post.
Or email Colin Mitchell: talking@tpg.com.au

The kit contains the following components:
(plus **extra** 30 resistors and 10 capacitors for experimenting), plus:

3 - 47R
5 - 220R
5 - 470R
5 - 1k
5 - 4k7
5 - 10k
2 - 33k
4- 100k
4 - 1M
1 - 10k mini pot
1 - 100k mini pot
2 - 10n
2 - 100n
5 - 10u electrolytics
5- 100u electrolytics

5 - 1N4148 signal diodes
6 - BC547 transistors - NPN - 100mA
2 - BC557 transistors - PNP - 100mA
1 - BC338 transistor - NPN - 800mA
3 - BD679 Darlington transistors - NPN - 4amp
5 - red LEDs
5 - green LEDs
5 - orange LEDs
2 - super-bright WHITE LEDs - 20,000mcd
1 - 3mm flashing LED
1 - mini 8R speaker
1 - mini piezo
1 - LDR (Light Dependent Resistor)
1 - electret microphone
1m - 0.25mm wire
1m - 0.5mm wire
1 - 10mH inductor
1 - push button
5 - tactile push buttons
1 - Experimenter Board (will take 8, 14 and 16 pin chips)
5 - mini Matrix Boards: 7 x 11 hole,
11 x 15 hole, 6 x 40 hole, surface-mount 6 x 40 hole board and others.

Photo of kit of components.
Each batch is slightly different:

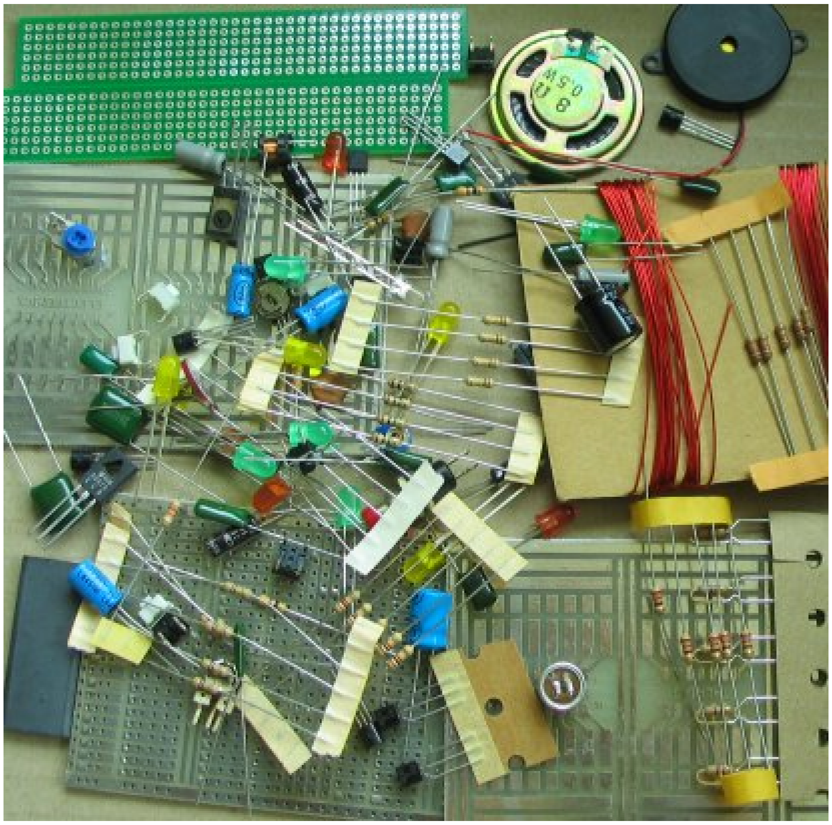

There are more components than you think. . . plus an extra bag of approx 30 components. The 8 little components are switches and the LDR and flashing LED is hiding.
In many cases, a resistor or capacitor not in the kit, can be created by putting two resistors or capacitors in series or parallel or the next higher or lower value can be used.

BEFORE WE START

Too many text books start with the physics of the atom and have equations and mathematics to show how smart the author is.
Don't worry, we wont have any physics or equations.
The reason . . .
This is not a physics course. It is a practical electronics course to teach the basics as quickly as possible. There are no equations because most transistor circuits cannot be worked out mathematically as the gain of a transistor changes according to the current-flow and these gain-values are never provided. So the mathematics is worthless.
To get an answer, all you have to do its build the circuit and measure the values with a multimeter.
Also lots of discussions in text books will never be used in your next 40 years of electronics, so this course doesn't have any unnecessary material and is much-more concentrated than anything you have read before.
Every frame contains important points - especially the animations - as they show you how a circuit works in slow-motion - something that has NEVER been done before.

ELECTRONICS BLOCKS

Here is an idea from Instructables to produce blocks with screws, containing a single component and they can be connected with jumper leads (alligator clips).

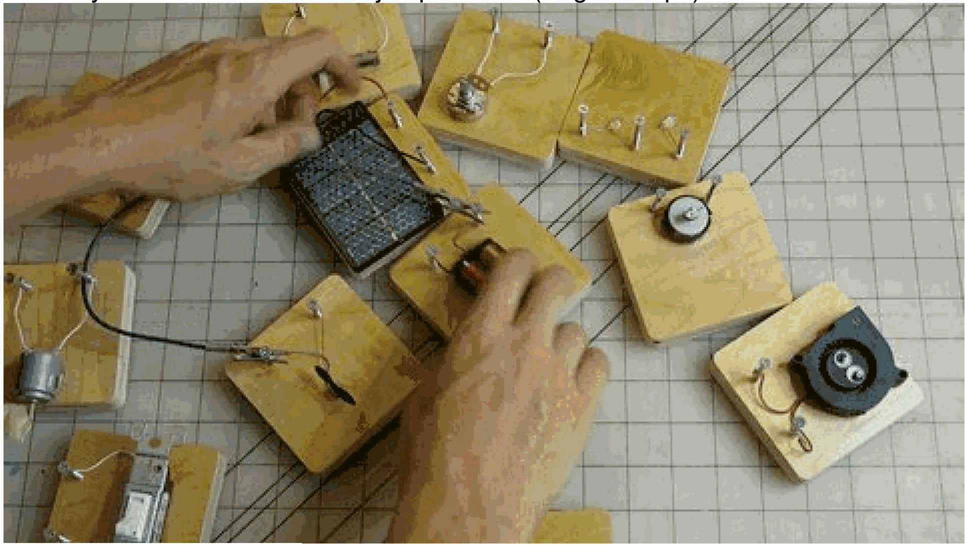

SLOTTED HEAD

PHILIPS HEAD

Use a slotted head for the negative screw and a philips head for the positive screw.

Learn electronics from the beginning . . .

START HERE:

All electrical and electronic components need wire to connect them to the circuit.
In a diagram called a CIRCUIT DIAGRAM, the wires are drawn as lines.
When the wires (or lines) cross, they may be joined or just passing.
It is **VERY IMPORTANT** to show the difference between lines that are **JOINED** and lines that are **NOT JOINED**.
When the lines are joined, it is best to place a dot

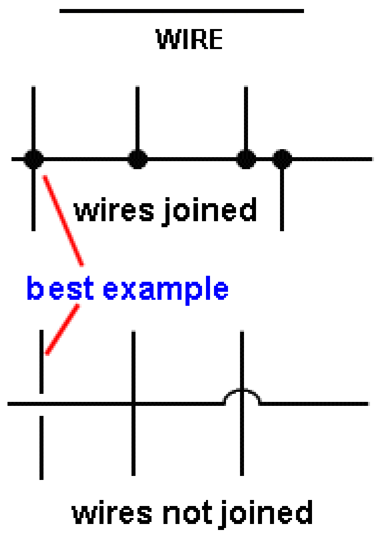

Fig 1 The wire in a circuit diagram

on the connection to PROVE the lines are joined. When the lines are just crossing, a gap should be made so it is obvious that one wire goes under the other and does not touch.
Lines should be "across the page" or "up and down." Very few lines should be at 45°.
You can make a line thicker to indicate a power rail or a wire that will be thick in reality.

to Index

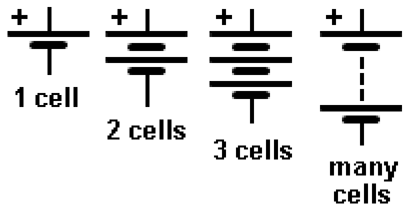

Fig 2 A single Cell and many Cells

Next we need a **battery**. A battery consists of two or more cells. The positive terminal of a battery is the long line in the diagram and you must add the voltage (of the cell or battery) to the symbol as a single cells can be 1.2v, 1.5v, 2.2v or up to 3.6v.
The symbol does not let you know the voltage.
The positive is always at the top and is the longest line on the battery symbol.

to Index

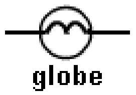

Fig 3: A Globe

Next we need a globe. A globe has two connections (a fine wire inside a glass bulb glows when the globe is connected to a battery).

to Index

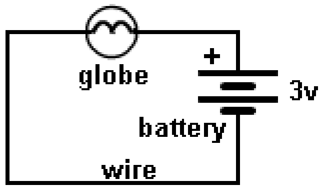

Fig 4: A Circuit

With a globe, battery and wire we have produced a **CIRCUIT**. A **CIRCUIT** is a complete path and we say the "electricity" the **CURRENT** emerges from the positive of the battery, moves through the globe and returns to the battery via the wire.
If the globe is a "3v GLOBE" it will glow when connected to a 3v battery.
The globe can be connected either way around.
The circuit we have shown is called a **SCHEMATIC** and consists of symbols: a globe symbol and a battery symbol.
The line connecting the two components is called WIRE.

to Index

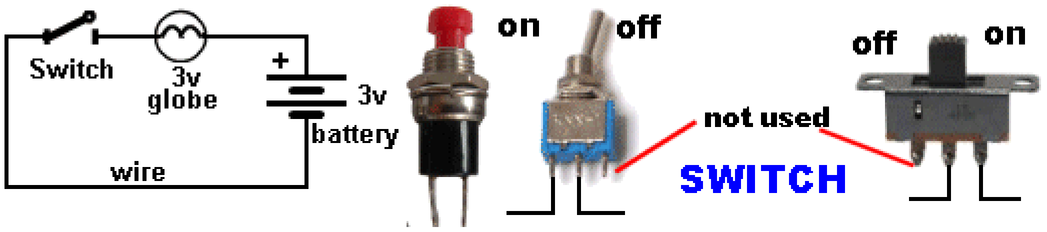

Fig 5:Adding a SWITCH
To turn the globe ON and OFF we need a SWITCH.
The switch may be a push button, a toggle switch (a "click" action) or a slide switch. You could

twist the wires together and untwist them. The result is the same. We say the circuit is "broken" or "open" via the switch and the lamp does not glow. Closing the switch turns ON the globe.

to Index

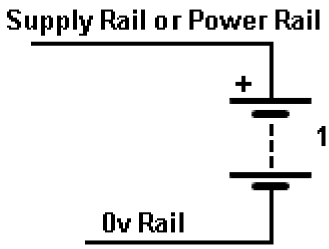

Fig 6: Naming the RAILS

The top rail of the **CIRCUIT DIAGRAM** is called the **SUPPLY RAIL** or **POWER RAIL**.
The lower rail is called the **0v Rail** or **EARTH RAIL**.
Do not connect the Supply rail (+12v) to the 0v rail as this will cause a high current to flow and is called a **SHORT CIRCUIT**:

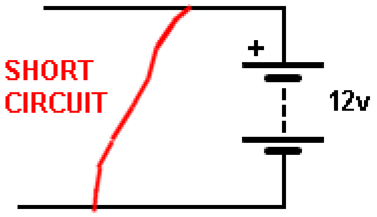

to Index

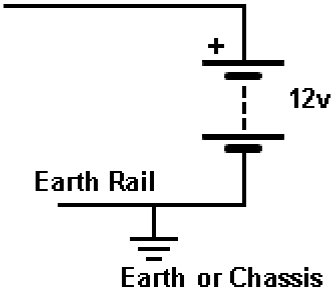

Fig 7: The **EARTH RAIL**

The lower rail is also called the **Chassis**.
This comes from the "old days" when electronics constructors build radios on a metal chassis (metal box) and it was connected via wire to a pipe in the ground to help the radio pick up distant radio stations.
The term also comes from car and truck wiring where one side of each globe is connected to the frame or chassis so that only one wire is needed to each globe and the return "path" is via the chassis.

to Index

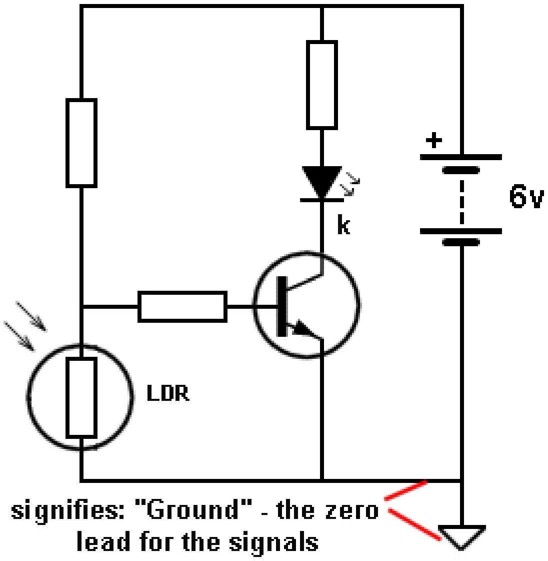

Fig 7a: Ground

Some circuits identify the "**Ground Lead**" or "**Ground Wire**" of a project to show where all the **signals** have been "referenced to." In other words, all the signals rise and fall above and below this "Ground wire" or "Ground Lead." This lead may not be at **earth potential** as the project may be in a plastic box but it identifies where the earth lead of a Cathode Ray Oscilloscope or the negative lead of a multimeter is connected.
On some printed circuit boards, the negative terminal of the battery (the 0v wire or terminal) is connected to a very large area of copper and this is called the EARTH PLANE or GROUND PLANE. It is designed to prevent signals travelling along the tracks (called traces) being radiated and also prevents outside interference upsetting the project. It also "tightens-up" the earth rail.

to Index

The circuit shows a 12v globe connected to a 12v battery and the circuit appears to be "broken" (not continuous).
But the **current returns** via the earth connection.

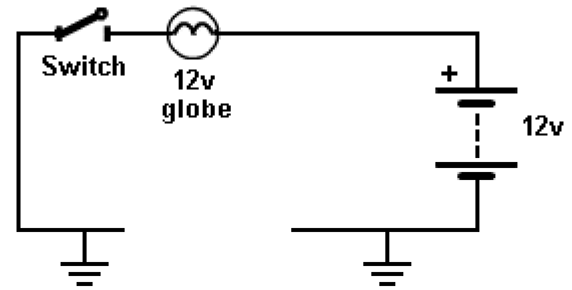

Fig 8: EARTH Return

We talk about the **CURRENT RETURNING**. We don't say: the voltage returning.
The voltage of the globe must be the same as the battery voltage, otherwise it will not glow fully or it will **burn out** if it is say a 6v globe.

to Index

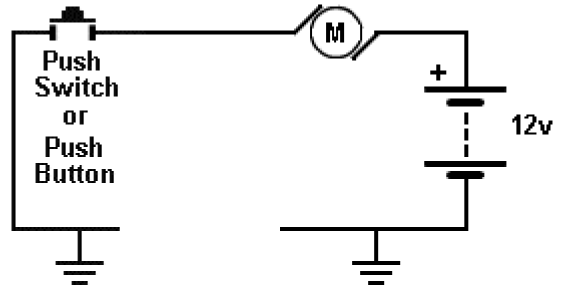

Fig 9: Connecting a Motor

The circuit shows a 12v **Motor**.
It is turned ON when the push-switch is pressed.

to Index

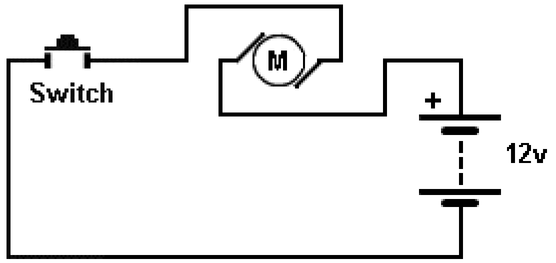

Fig 10: Reversing a Motor

If the wires are connected to the motor "around the other way," the motor will reverse direction.

to Index

VOLTAGE AND CURRENT

What is voltage and what is current?
Here is a very simple description.
A battery produces a voltage called DC. (This is a very confusing name because the letter actually refer to Direct Current, so we just say **DC Voltage**).
A battery also produces current called DC - Direct Current. We say **DC current**.

VOLTAGE

Voltage is a value produced by an electrical component called a battery or cell.
A single cell produces one and a half volts. (1.5v) and although this is not a high voltage, when cells are connected together we get higher voltages.
If 6 cells are connected in series we get 9v.
Here is a 9v battery:

CURRENT

You cannot feel current with your tongue so we have to carry out another experiment:

negative terminal
positive terminal
9 VOLTS
2010
Energizer

Touch the two terminals with your tongue. You get a tingle. This is a 9v tingle. Now you have "felt" electricity. This is a 9v tingle.

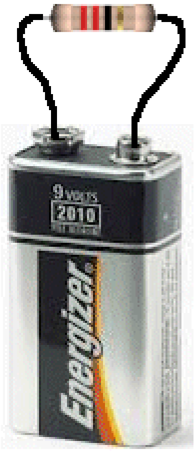

Place a 22 ohm or 47 ohm resistor across the terminals of the battery and hold your fingers on the resistor. It will get hot. This is the result of current flowing through the resistor and heating it up. The current will be about half an amp and the voltage is 9v, so the wattage will be about 2 to 4 watts. Feel the heat produced.

Milli - milli means 1/1,000th (one thousandth) - such as one milliamp or one millivolt.
In other words one thousand milliamps is equal to 1 amp.
One volts is not a very large value as a battery produces 9v and a cell produces 1.5v to 3.6v (depending on the type of cell.
But 1 amp is a large quality when talking about electronic circuits involving LEDs, motors and transistors.
The globe used in the experiments above requires about 300mA. (1,000mA = 1 amp)
The 3v motor used in the experiments requires about 250mA
The LEDs used in the experiments require about 20mA.
Transistors can pass about 100mA to 800mA via the collector-emitter.
In most cases current-flow in the circuits we will be discussing will be less than 1 amp and will be shown as 25mA, 100mA, 350mA etc.

WATTAGE and CAPACITY

A 9v battery has 6 very small cells and they will not last very long.
A "AAA" cell is larger and a "D" cell is much larger.

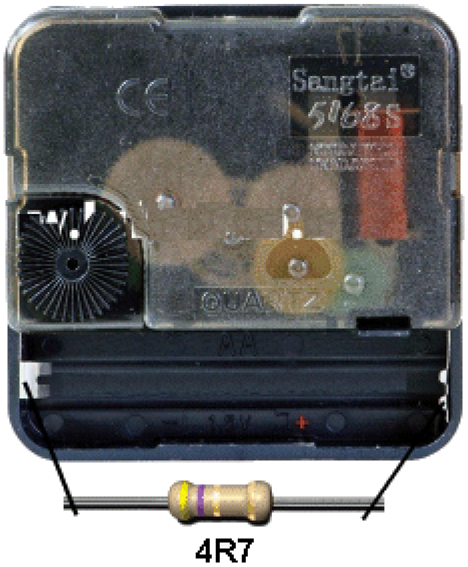

A large cell is said to have a **LARGE CAPACITY.** This means it will deliver a larger current for a longer period of time.
The **WATTAGE** of a cell is the multiplication of the voltage x current. The answer is milliwatts or watts.
The **CAPACITY** of a cell is the wattage x hours. The answer is milliwatt-hours or watt-hours. This is also called watt-hours.
You can determine the capacity of a cell (such as a rechargeable cell) by connecting it to a clock-mechanism that has a 4R7 connected across the terminals. The resistor will take a considerable current and deplete the cell in a few hours. The clock will let you know exactly how long the cell delivered the current. You can then compare other cells.

The simplest electrical circuit consists of a battery and resistor. The current flowing through the circuit will depend on the voltage of the battery and the resistance of the resistor R.

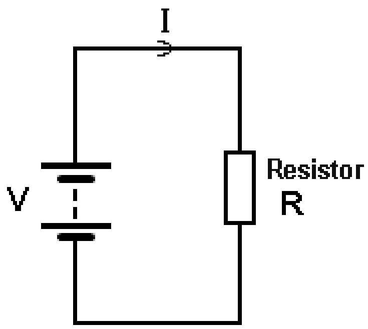

The formula connecting these three quantities is:

$$I = \frac{V}{R}$$

Ohm's Law

$$I = \frac{12}{3}$$

$$I = 4 \text{ amps}$$

This is called **Ohm's Law**. Suppose you have a 12v battery and the resistor is 3 ohms. The current flowing through the resistor will be 4 amps.
Increasing the resistance will decrease the current if the voltage remains fixed.

All the above circuits are called ELECTRICAL CIRCUITS because they contain electrical components (such as a motor, globe, relay, switch).
When the circuit contains an ELECTRONIC component such as a diode, transistor, LED, it is called an ELECTRONIC CIRCUIT or ELECTRONIC SCHEMATIC.

to Index

POWER and ENERGY

Here's an easy way to remember the difference between POWER and ENERGY:
A 9v alkaline battery has enough ENERGY to start a car. But it does not have enough POWER (strength).
Energy is effectively the strength of the battery (and this is the voltage and the current it can deliver) multiplied by the time it can deliver this energy. When the answer is obtained, it consists of three factors ((3 quantities) VOLTS, AMPS and TIME.
This results in an answer called xxxx WATT-HOURS.
For a 9v battery the quantities are: 9 volts, 500mA and the battery will deliver this 9x0.5 = 4.5watts for about 1 hour. This is equal to 4.5 x 60 x 60 = 16,200 watt-seconds.
To start a car requires 250 amps from a 12v battery for 5 seconds.
This is: 12 x 250 x 5 = 15,000 watt-seconds.
This means the energy stored in a 9v battery could start a car if all the energy could be delivered in 5 seconds.
This is not possible however the FACT is this: A 9v battery has enough stored energy to START A CAR.

to Index

POWER STATION
300

BATTERY BOOSTER

One of the simplest things we can do is start a car with a flat battery with the assistance of a BATTERY BOOSTER.
This consists of a 12v rechargeable battery in a handy case with leads to connect to the flat battery in your car.
This simple operation puts two 12v batteries in parallel, but no-one has actually described what happens and why.
That's because the explanation is very complex. We have included it here to show that a simple explanation involves a lot of technical terms and you will understand more after reading the course.

The flat battery in the car is not fully charged but it has some percentage of charge and when it sits for a period of time in a non-fully charged condition, the voltage drops from 12.6v to less than 12v as the battery gradually self-discharges due to the potential at the top of the cell being different to that at the bottom of the cell and the specific gravity of the electrolyte being different at the top and bottom. This causes an internal current to flow within the cell and slowly discharge the cell. But if you try to start the car, the voltage drops to less than 7v because the electrolyte cannot carry the high current and a slight potential is developed across the liquid.

The result is the starter-motor does not crank the car.
The reason is this: When the battery is fully charged, the current taken by the starter motor is about 300 amps. This is about 11v x 300 amps = 3300 watts = 4.4Horsepower.
But when the voltage drops to 7v, the current will drop to 190 amps to deliver 1336 watts = 1.8HP. This is only 40% of normal and that's why the car does not start. The engine needs 4HP to overcome the pressure in the cylinders due to the compression of the air during the "firing stroke."
Let's put it this way. If we have a brand new 7v battery, the car will not start. The starter-motor will only accept 190 amps when the supply is 7v.
So, we have to increase the voltage.
We do this by placing a 12v battery across the flat battery. The voltage of the flat battery will immediately rise to 12.6v. It might take 2 minutes but the flat battery will take a small current (1 to 10amps) from the battery in the "booster" and the output of the combination will be 12.6v. The current-carrying capacity of the electrolyte will improve very quickly and you have effectively given the "flat battery" a very quick charge.
The starter-motor will now accept 300 amps from the combination and **SURPRISINGLY** the cells of the "flat battery" will deliver about 200 amps and the booster battery will deliver about 100 amps. The actual sharing of current will depend on the two batteries but the secret behind the success is the increase in voltage we call **TERMINAL VOLTAGE**. The voltage on the terminals (the alligator clips).
The capacity of the booster battery is not important. It can be from 7AHr to 40AHr. We are just using a very small amount of its capacity to start the car and nearly all batteries will provide 200 Amps for a short period of time.
The voltage of the car battery is very important. The Horsepower taken by the starter-motor is defined by the formula: Pwatts = V^2/R Since the resistance remains constant, a voltage of 7 volts will produce 7x7=49 units and a voltage of 11v will produce 121units. This gives the ratio of 40% to 100% as explained above.

to Index

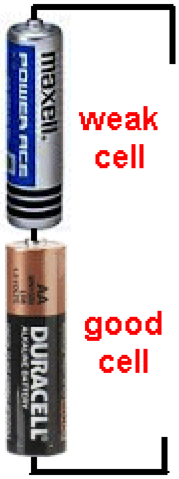

BATTERY BOOST

Continuing from the previous frame where we showed the effect of placing a weak battery in parallel with a good battery, we can show what happens when a weak cell is placed IN SERIES with a good cell.
This also applies when you have 5 good cells and one weak cell.
Basically, the weak cell will reduce the current. In other words, if the 5 cells are driving a motor and supplying 250mA, the 5 cells and 1 weak cell will deliver 200mA or less, depending if it is weak or very weak. The current flowing through the weak cell will have the effect of giving it a small charge - in other words, you will be charging the weak cell from the good cells when the motor is operating.

BUT . . . .

There is a way to use weak cells. If you have say 6 weak cells driving a motor and the RPM is reducing, you can add 2 more weak cells to increase the RPM.
The effect is this: The voltage from the 8 cells will be higher than from 6 cells and this will allow a higher current to flow. Sometimes the cells will provide this higher current and thus more of the energy will be delivered and you will get the last of the energy from the cells.

to Index

INTERNAL RESISTANCE

All batteries and also all individual cells have a "secret, hidden" value of resistance inside each cell due to the resistance of the chemicals. This resistance is very small when the cell is new but it increases as the cell gets older.
It is very easy to measure this value. Simply put an ammeter directly across the cell and measure the current. Use Ohm's law to work out the resistance. But this not always a wise thing to do as some cells will deliver 10 amps and some will deliver 100 amps and damage the meter.
The diagram opposite shows a large internal resistance for the weak cell and a small internal resistance for the good cell.
If a cell did not have any INTERNAL RESISTANCE it would deliver thousands of amps. It's the Internal Resistance that limits the current.

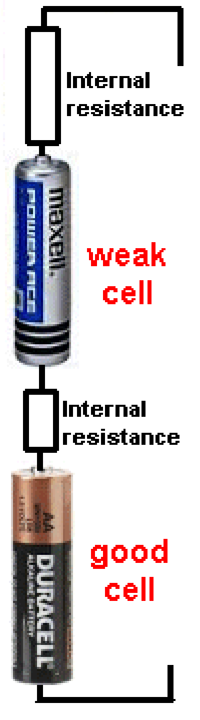

In most cases we neglect (do not consider) the value of internal resistance when making tests and when using a battery in a project.
But when a battery gets old, it cannot deliver a high current and the internal resistance gets so high that the output voltage drops from say 9v to 7v, even when the battery is not connected to a circuit.
This is the result of the INTERNAL RESISTANCE of the chemicals increasing to a point where they become noticeable and what we call "poisoning" of the chemicals due to the cell "aging" and new chemicals being produced in the cell that have a high resistance. Some of the terms we use are: "drying out and sulphating. Some cells produce spikes or needles that completely short-circuit the cell and make it totally useless.

to Index

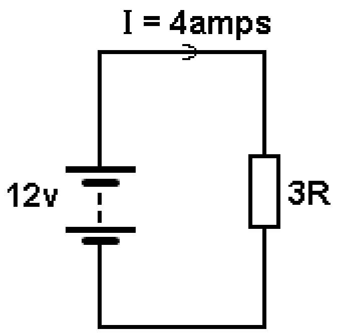

This resistor is dissipating 48 watts.
As a comparison, a soldering iron is dissipating about 10 to 20 watts.

RESISTOR WATTAGE

Resistor Wattage means two things.
1. The physical size of a resistor tells you number of watts it is capable of dissipating. This is called RESISTOR WATTAGE. It is really RESISTOR-SIZE or RESISTOR-CAPABILITY.
2. The multiplication of the voltage across a resistor and the current flowing though it will produce a value called WATTAGE. This is also called RESISTOR-WATTAGE or RESISTOR-LOSS or RESISTOR-DISSIPATION or HEAT-LOSS.
In the circuit shown, the wattage being lost in the resistor is:
12 x 4 = 48 watts.
Most of the resistors we will be using in our projects are 0.25watts. This means they will dissipate 250milliwatts, however the actual wattage being dissipated may be only 70 milliwatts and the resistor will not get hot.
0.25watts is the maximum wattage it can dissipate without overheating.
If it is dissipating 400milliwatts, it will be VERY HOT.
The wattage it is dissipating (the heat it is getting rid of) will depend on the supply voltage and the value of the surrounding components.

to Index

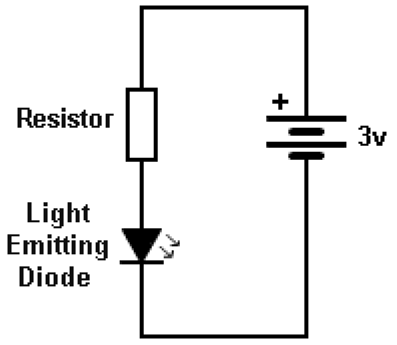

Fig 11. The LED

This simple **ELECTRONIC CIRCUIT** contains a **LIGHT EMITTING DIODE (LED), RESISTOR** and battery.

The circuit is classified as electronic because the LED is not an electrical item (such as a globe) but more-complex, as it produces light when current flows through a crystal and the crystal produces the colour.

to Index

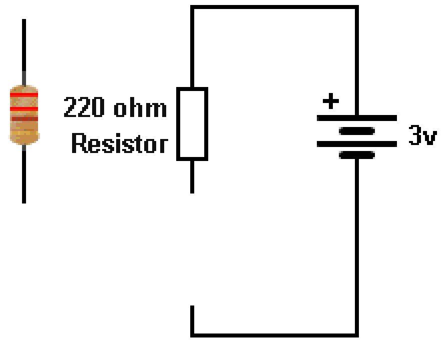

Fig 12. **The Resistor**

A **RESISTOR** must be included in the circuit to prevent the LED being damaged.
The resistor in this circuit must be 220 ohms.
This is shown by the colours on the resistor. The colours for 220 ohms: **red - red - brown**. The 4th band is gold - indicating a tolerance of 5%.
A resistor has **RESISTANCE**.
It reduces the current from the battery to a required amount to prevent the LED glowing too bright.
A resistor is just like putting your foot on a hose. The water trickles out the end. The resistor "resists" the high current-flow that the battery is able to deliver.

to Index

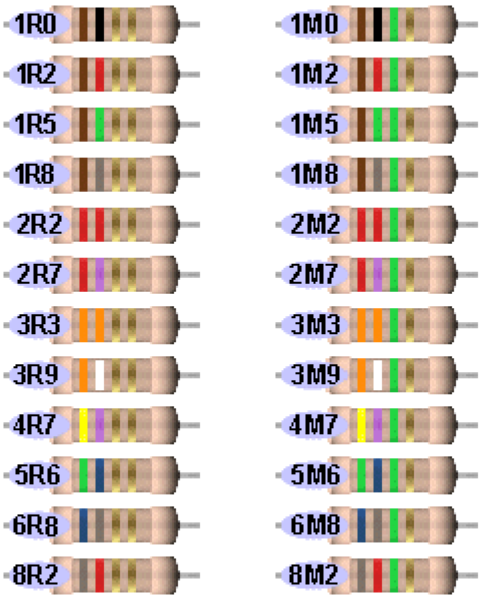

Fig 12a. The **Resistor Colours**

There are hundreds of different resistors because the resistance-values need to cover the range one ohm to 10 million ohms.
There are also small, medium and large resistors. The resistors on the left are just a few in the range. (See the full range below). They show colour bands for 1 ohm to 8.2 ohms and 1 million ohms to 8.2 million ohms. All the other values are shown below.
An electronics engineer does not have the room to store 10 million different resistors so they make each resistor 5% or 10% higher than the previous. This reduces the number to about 100 to 200.

TOLERANCE

The first 3 bands indicate the value of the resistor and the 4th band indicates either 5% or 10% tolerance.
All modern resistors are 5% or 2% or 1%. The "old" 10% resistors are no longer made.
Gold = 5%
Silver = 10%

to Index

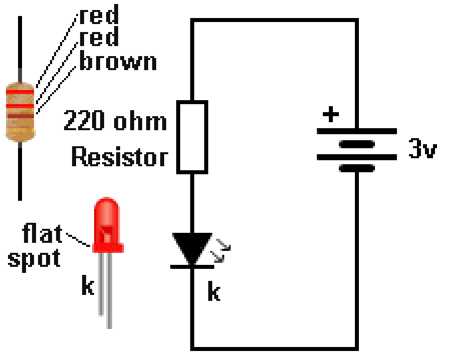

Fig 13. The **Resistor and LED**

The **LIGHT EMITTING DIODE** is called an electronic component (mainly because it is more complex than a globe and it produces light by a more-complex means than heating a wire).
A **LED** must be connected around the correct way.
It will not illuminate if connected around the wrong way.
All **LEDs** have one lead longer than the other. The SHORT lead is called the **CATHODE** (k).
All LEDs have a flat on one side and this is the **CATHODE** lead.
The arrows on the diagram indicate light is "given off" (emitted - produced).

to Index

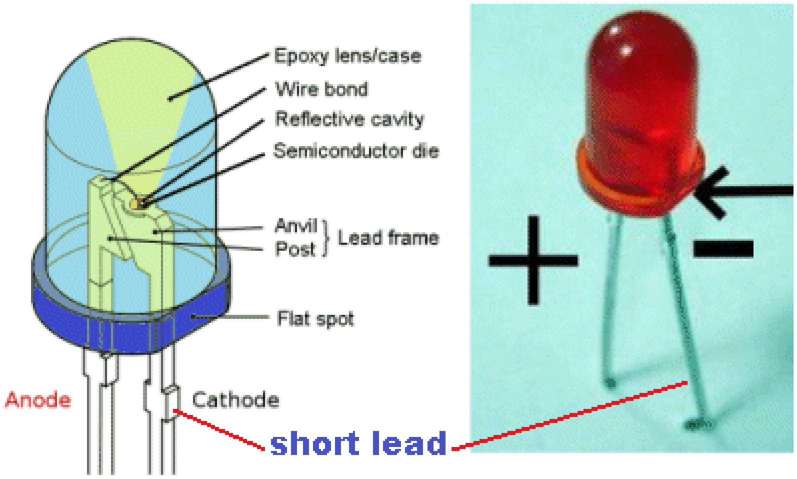
Epoxy lens/case
Wire bond
Reflective cavity
Semiconductor die
Anvil
Post
Lead frame
Flat spot
Anode
Cathode
short lead
+
-

Fig 14. The **LED - showing the flat spot**
A close-up of a red LED. The cathode lead is the short lead and next to a flat side on the LED.
DO NOT show "+" or "-" on a diagram. Only show the letter "k" to indicate cathode.
The symbols "+" and "-" are used when a component produces a voltage or is connected directly to "+" and "-" A LED is connected via a resistor.

to Index

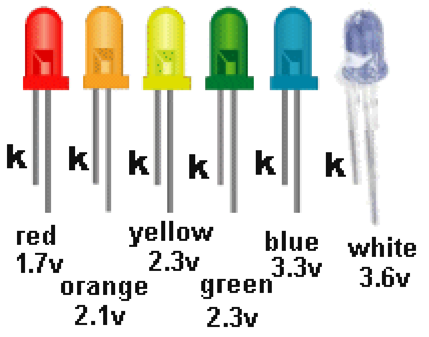
k
k
k
k
k
k
red
1.7v
yellow
2.3v
blue
3.3v
white
3.6v
orange
2.1v
green
2.3v

Fig 15. LED VOLTAGES

When a LED is connected to a circuit, (and the correct-value resistor is included), a voltage will be develop across the LED called the **CHARACTERISTIC VOLTAGE DROP**.
This voltage is due to the colour of the LED and the crystal inside the LED that produces the colour.
The diagram on the left shows the approximate voltage developed for each LED.
The voltage **does not change** for small, medium, surface-mount, or large LEDs.

to Index

When a LED is connected to 3v battery, the following **CHARACTERISTIC VOLTAGE DROPs** will develop across each LED.
You will notice we **have not changed** the value of the resistor. It is 220R.
The LED creates the voltage and if the value of resistance is decreased, the LED will illuminate BRIGHTER.
If the LED illuminates too bright it will be DAMAGED.

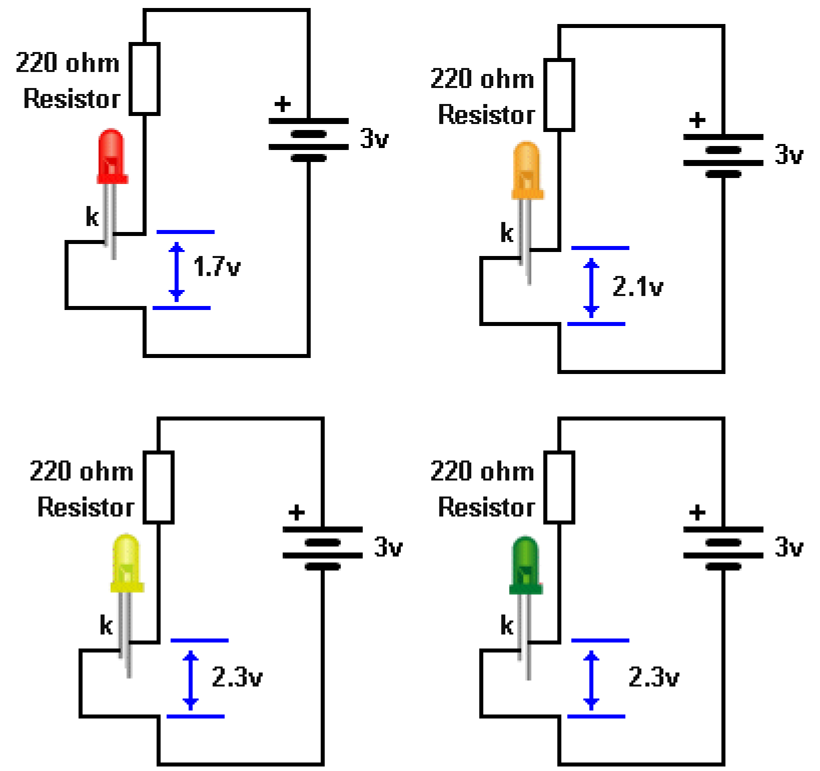
220 ohm Resistor
+
3v
k
1.7v
220 ohm Resistor
+
3v
k
2.1v
220 ohm Resistor
+
3v
k
2.3v
220 ohm Resistor
+
3v
k
2.3v

Fig 16. LED VOLTAGES

to Index

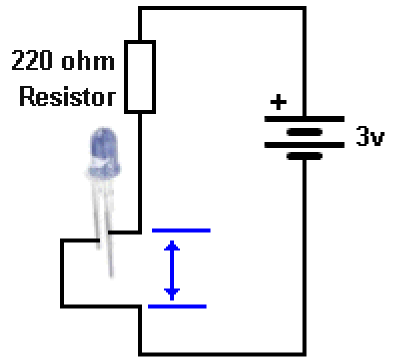
220 ohm Resistor
+
3v

Fig 17. WHITE LED VOLTAGE

If we connect a WHITE LED to 3v supply, it will not illuminate because it needs a supply higher than 3.6v.

The resistor in series with the LED is called a **CURRENT LIMITING RESISTOR**.

In this circuit **no current flows** because the supply is not high enough.

to Index

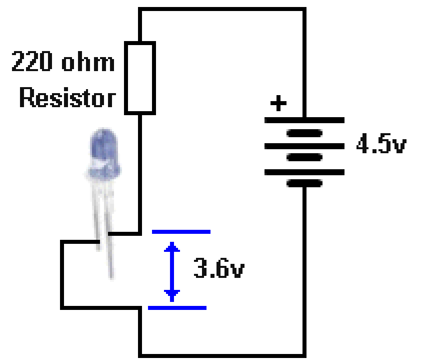
220 ohm Resistor
+
4.5v
3.6v

Fig 18. WHITE LED VOLTAGE

When the supply is increased to 4.5v, the 220R resistor will allow a current to flow through the white LED and it will develop a CHARACTERISTIC VOLTAGE DROP of 3.6v across it.

The supply (the voltage of the battery) must be higher than the CHARACTERISTIC VOLTAGE DROP of the LED so the resistor will allow the correct amount of current to flow.
The ideal current for a LED is 20mA, however some LEDs will work when 1mA flows, so you have to know what you are doing.

to Index

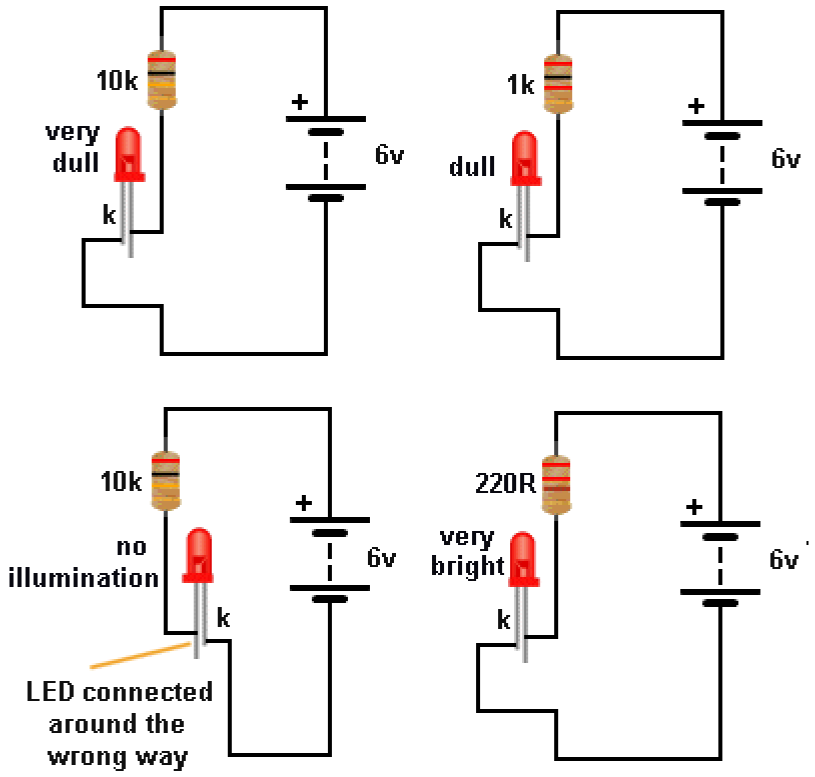
10k
very dull
k
+
6v
1k
dull
k
+
6v
10k
no illumination
k
+
6v
LED connected around the wrong way
220R
very bright
k
+
6v

Fig 19. Testing A LED
Now connect either the 1k, 470R or 220R and determine the brightness you need.
As the brightness increases, the current will be higher.
You can use 3v supply for all LEDs except blue and white.

HOW TO TEST A LED
Some clear LEDs produce red or orange and some LEDs do not have the cathode lead clearly identified.
Here's how to find the colour, cathode lead and the current.
You need a 6v battery, 10k resistor, 1k resistor, 470R resistor and 220R resistor.
Connect the 6v battery and 10k resistor to the LED and it will only illuminate when the cathode is connected to the negative of the battery. This is the short lead.

to Index

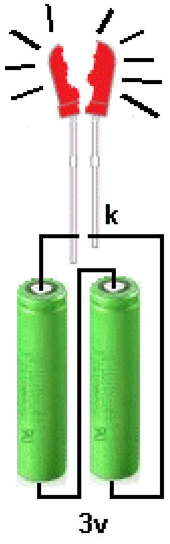

Fig 20. Damage a LED

Do not connect a 3v battery directly across a LED.
It will be DAMAGED. You MUST include a resistor.

to Index

A LED IS CURRENT DRIVEN

You may have seen this statement and tried to work out what it means.
Basically it means an increase in current will make the LED brighter.
But a LED needs 2 things:
It needs a voltage that is EXACTLY the voltage required to produce illumination. And this voltage depends on the colour of the LED.
As soon as you supply the exact voltage, the crystal will begin to glow and as you increase the current, the illumination will increase.
But doing this is VERY VERY difficult.
It is very easy to supply an exact voltage such as 1.7v or 3.4v, but delivering a current such as 10mA or 20mA at the same time is very difficult. You cannot get a 1.7v battery and deliver 10mA to a LED.
As we have shown above, you need a simple components such as a resistor between the battery and LED to achieve the desired result.
A LED is CURRENT DRIVEN but firstly you need to provide a VOLTAGE that is exactly the connect value for the colour of the LED and then the current can be increased.

to Index

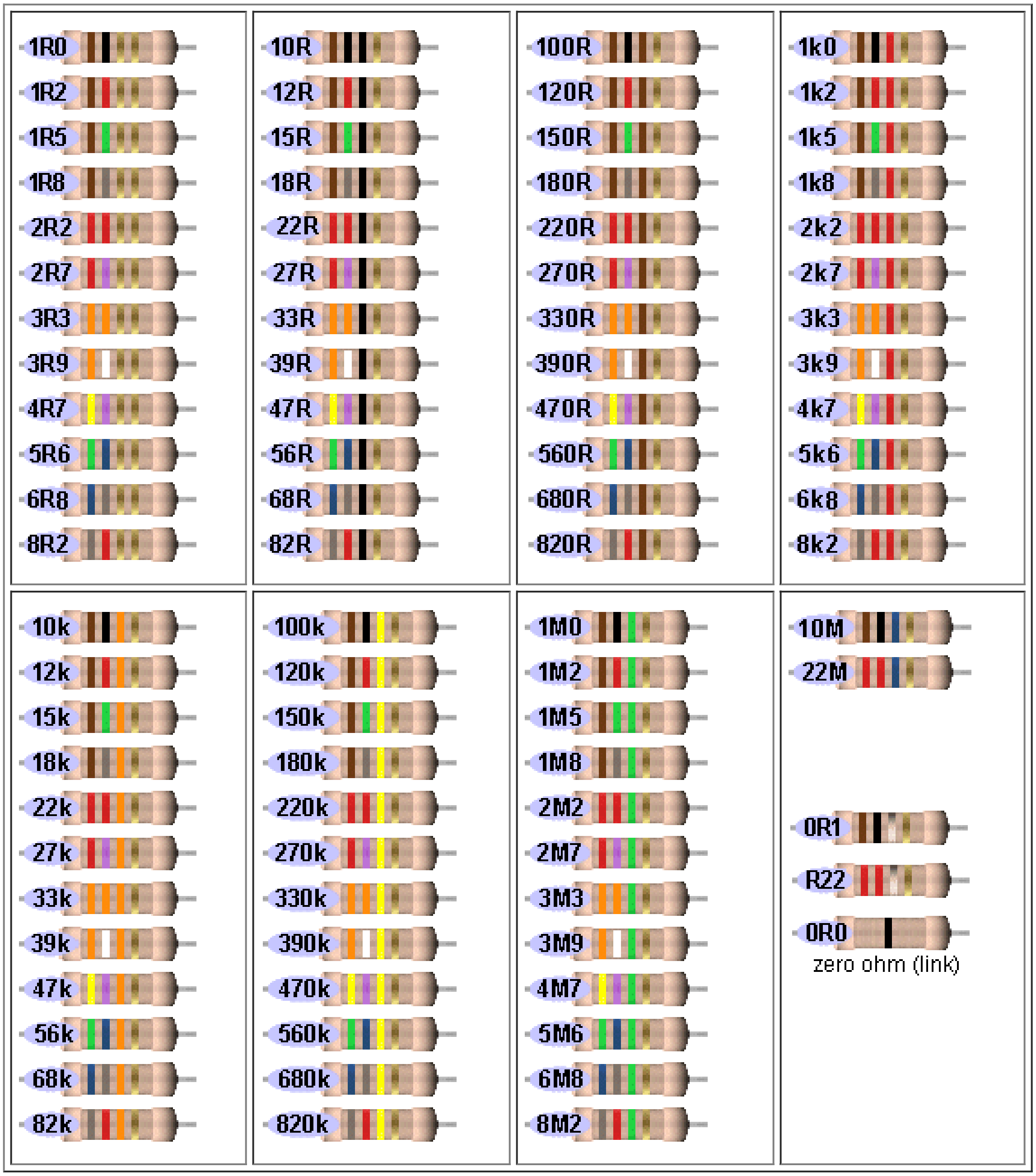
1R0
1R2
1R5
1R8
2R2
2R7
3R3
3R9
4R7
5R6
6R8
8R2
10R
12R
15R
18R
22R
27R
33R
39R
47R
56R
68R
82R
100R
120R
150R
180R
220R
270R
330R
390R
470R
560R
680R
820R
1k0
1k2
1k5
1k8
2k2
2k7
3k3
3k9
4k7
5k6
6k8
8k2
10k
12k
15k
18k
22k
27k
33k
39k
47k
56k
68k
82k
100k
120k
150k
180k
220k
270k
330k
390k
470k
560k
680k
820k
1M0
1M2
1M5
1M8
2M2
2M7
3M3
3M9
4M7
5M6
6M8
8M2
10M
22M
0R1
R22
0R0
zero ohm (link)

Fig 21. All the resistor values

Here are all the colours and values for the resistors you will using in this course. Just match-up the colours on your resistor with the resistors above and you will find the value.

to Index

Resistor values are always OHM values. One ohm is a small value. It might be the resistance of a length of wire 3 metres long.
When a switch is open the resistance is infinite - millions and millions of ohms.
The resistance of your body from one hand to the other will be about 70,000 ohms.
The resistance between two wires dipped in water will be about 1,000 to 100,000 ohms (depending on the dissolved-salts in the water - pure water has a very high resistance)
The resistance of the filament of a 3v globe will be about 30 ohms.
The resistance of the winding of a 3v motor will be about 3 ohms.
Resistors are made with values from less than one ohm to more than 10 million ohms by adding carbon to the mixture inside the resistor (and cutting a track around the outside of the resistor) then connecting a lead to each end. Adding more carbon reduces the value of resistance. Carbon has a low resistance.
Resistance-values are measured with the RESISTANCE settings on a MULTIMETER.
This is called the "Ohms Range." Sometimes with the symbol: **Ω**
A Multimeter will have 2, 3 4 or more scales to cover the range one ohm to 10 million ohms.
Low value resistors (from 1 ohm to 999 ohms) are written as 1R, 220R, 470R, 999R. with the letter "R" indicating Resistance (ohms). You can also use the symbol "omega" **(Ω)**
For values above 1,000 ohms to 99,999 ohms, they are written as: 1k, 2k2, 4k7, 10k, 100k, 220k, 470k, with the letter "k" indicating "kilo" (thousand).
1M = 1,000,000 - one million ohms 1M2, 2M2, 4M7, 10M.
The letters "R, k and M" are placed so they take the place of the decimal point. This prevents any mistake, as a decimal point can be missing in a poor photocopy.

to Index

6.43
OMEGAETTE

MULTIMETERS

There are two types of **MULTIMETER**. The top two are called **DIGITAL MULTIMETERS** (DMM) and show numbers on a display.
The lower two meters are called **ANALOGUE MULTIMETERS** and have a pointer and scale.
All meters come with a set of red and black leads.

The **red lead** is always connected to the positive of the battery or the positive on a project and the **black lead** is connected to the negative or earth or chassis.
When making a resistance measurement, the leads can be around either way.
Resistance measurements are always made with the power removed from a circuit. Any voltage on a circuit will upset the resistance reading.

to Index

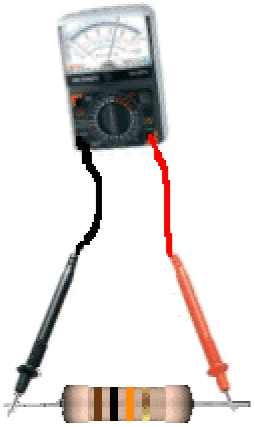

Fig 22. Resistance Measurement with Analogue Multimeter

The resistance of a resistor is measured by placing the leads of the multimeter on the ends of a resistor and turning the dial on the analogue multimeter to the resistance scale to make the pointer move to about the centre of the scale.
The resistance scale is marked with a high value on the left and 0 ohm on the right. This is opposite to all the other scales.
You must get the pointer to move to the middle of the scale as it is not accurate at left-end.
Analogue multimeters are only suitable for reading values from 1 ohm to 100,000 ohms. The scale is too hard to read above 100k.
To find the value of a resistor, you can compare the colours with the table above.

to Index

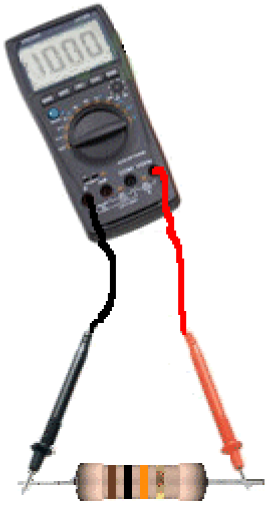

Fig 23. Resistance Measurement with a DMM

A digital multimeter produces a more-accurate reading of resistance.
It is accurate from 1 ohm to 10M ohms.
Select the scale that provides a reading.

to Index

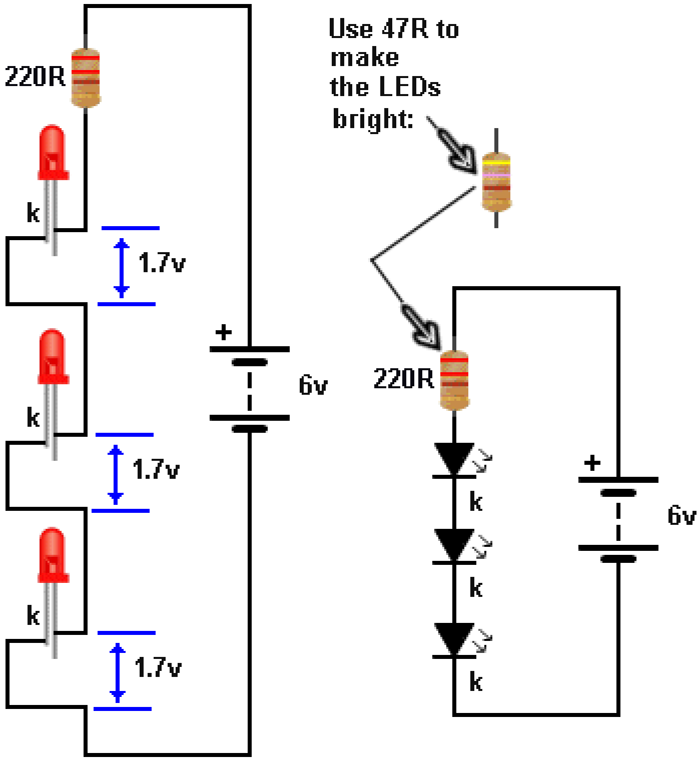
220R
k
1.7v
k
1.7v
k
1.7v
+
6v
Use 47R to make the LEDs bright:
220R
k
k
k
+
6v

Fig 24. Connecting **LEDs in series**
LEDs can be placed in series provided the total **CHARACTERISTIC VOLTAGE DROP** across the LEDs is LESS than the supply voltage.
In this case the voltage across the LEDs is 1.7v + 1.7v + 1.7v = 5.1v
The supply is 6v and this allows 0.9v for the CURRENT LIMITING RESISTOR.
The LEDs will not be very bright with 220R.
Change the resistor to 47R
If you connect 4 LEDs in series, the total **CHARACTERISTIC VOLTAGE DROP** will be 6.8v and no LEDs will illuminate because the total is higher than the 6v supply.

to Index

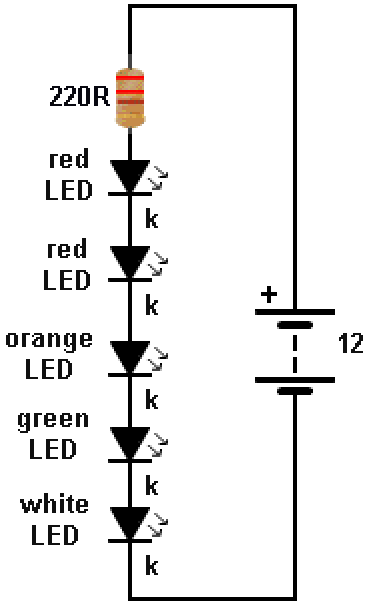
220R
red LED
k
red LED
k
orange LED
k
green LED
k
white LED
k
+
12v

Fig 25. Connecting LEDs in series

Different-colour LEDs can be connected in series. Add up the total Characteristic Voltage for the 5 LEDs and see if it is less than 12v.

The 220R resistor will have to be reduced to 47R to make the LEDs bright.

to Index

LEDs can be connected in parallel if they are the

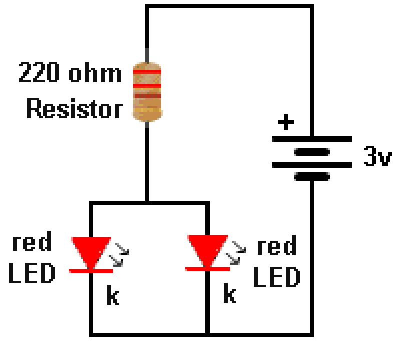

Fig 26. Connecting LEDs in parallel

same colour.
In the diagram a red LED drops a CHARACTERISTIC VOLTAGE of 1.7v and if they are from the same manufacturer or the same batch, they will work ok.
Although we say the characteristic voltage for a red LED is 1.7v, this can change slightly from different manufacturers and one LED may glow brightly while the other is dull.
You have to build the circuit and see the result.

[to Index](#)

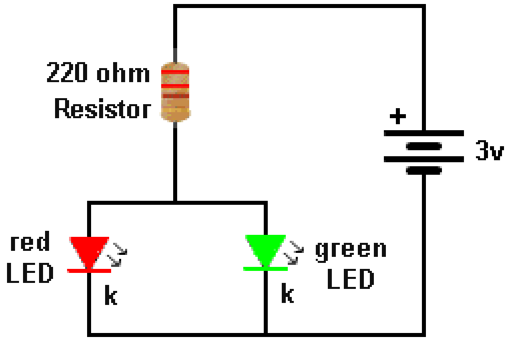

Fig 27. Connecting LEDs in parallel

Different colour LEDs cannot be connected in parallel. The voltage across a red LED is 1.7v. This becomes the "Supply Voltage" for the green LED and it is too low. The green LED needs a supply of 2.1v to 2.3v.
Only the red LED will illuminate.

[to Index](#)

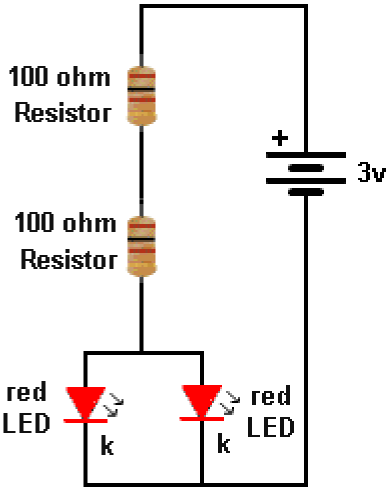

Fig 28. Connecting Resistors in Series

Suppose you don't have a 220 ohm resistor.
You can make a 220 ohm resistor with two resistors in series. The total resistance will be 200 ohms, but resistors are not accurate and the result will be very close to 220R.
Electronic circuits are not very critical. You will not be able to see the difference in brightness between 200 ohms and 220 ohms.

When resistors are connected in series, the total resistance is found by adding the resistance of each resistor.

$$R_{total} = R_1 + R_2 + R_3 + \dots$$

[to Index](#)

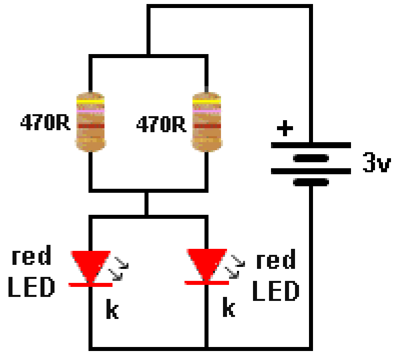

You can create a 220 ohm resistor by connecting two resistors in Parallel.
When two equal-value resistors are connected in Parallel, the total resistance across the combination is HALF.
470R in parallel with 470R produces 235R.
This is very close to 220R.
We are not going into the formula as it is very complex.
Three equal-value resistors in parallel produce a total of **one-third.**
Simply get two resistors and connect them in

Fig 29. Connecting Resistors in Parallel

parallel and measure them with a multimeter.

to Index

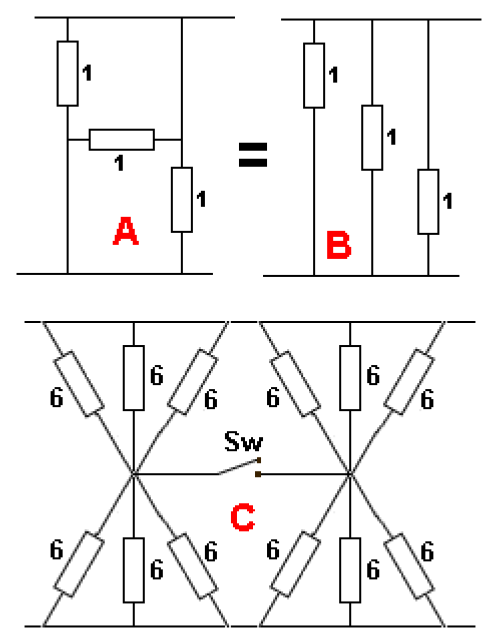

Fig 29a. Two tricky resistor questions

Figure **A** shows three resistors. It looks hard to solve so the middle resistor is turned so it connects directly to the top and bottom rail. Now you can see the circuit is three resistors in parallel. The result is one-third of an ohm.

Figure **C** shows twelve 6 ohm resistors. Replace each group with a 2 ohm resistor, because three 6 ohm resistors in parallel is equal to 2 ohms.

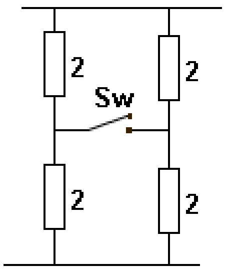

The two left resistors create 4 ohms and the two right resistors create 4 ohms. The result of two 4 ohm resistors in parallel is 2 ohms.
The resistance of the circuit does not change if the switch is open or closed.

to Index

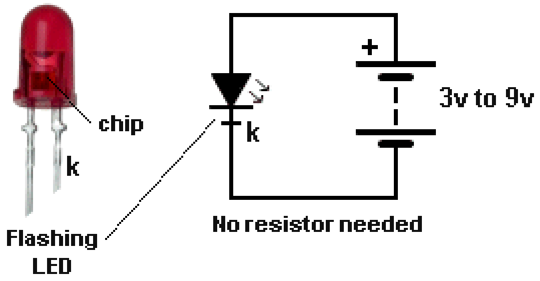

Fig 30. Flashing LED

There are some special LEDs that can be connected to 3v to 9v and they flash or produce a range of colours.
These LEDs have a chip and resistor inside the body of the LED to produce the effect and allow the LED to operate on a voltage without the need for a current limiting resistor.

to Index

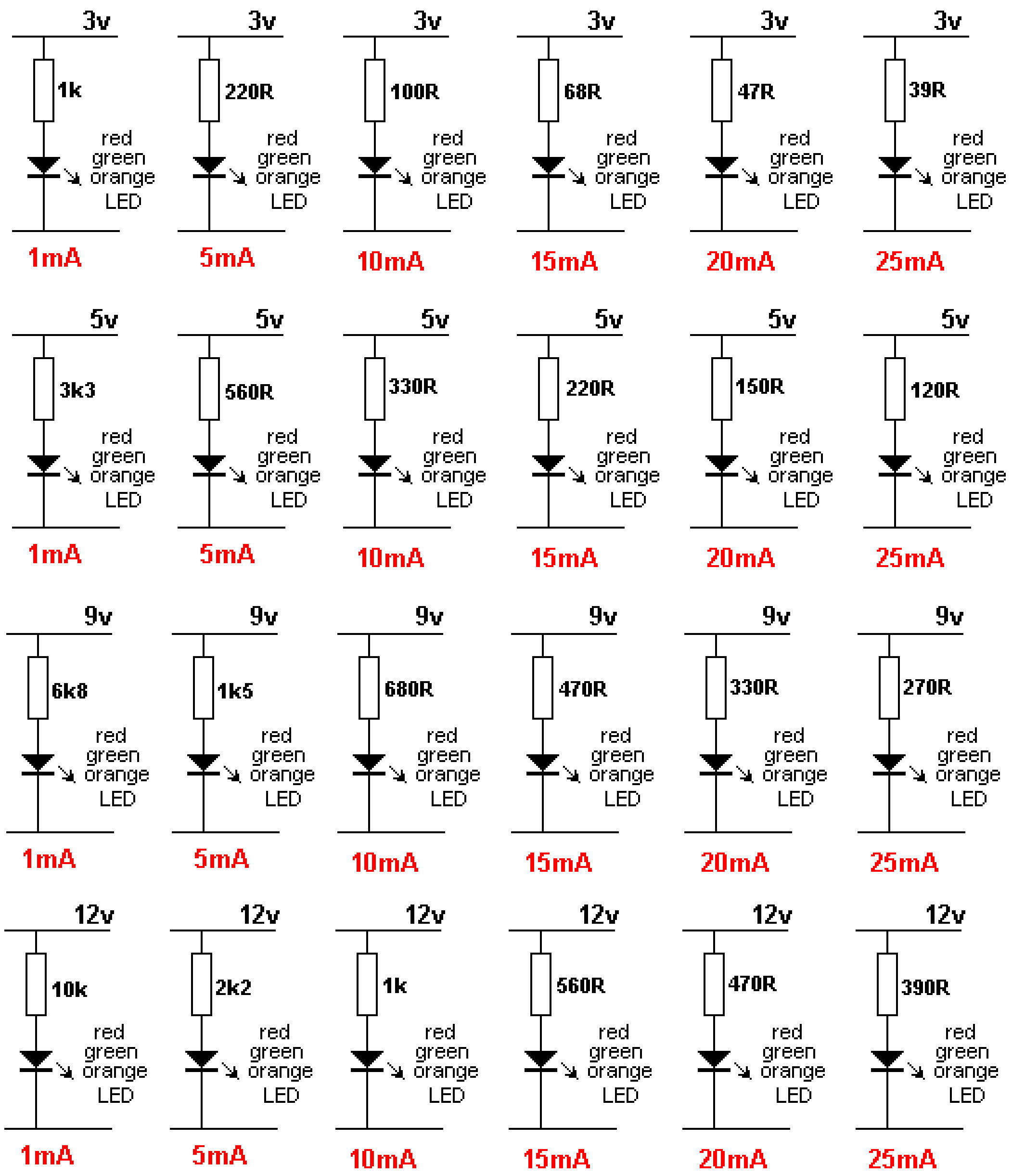
3v
1k
red green orange LED
1mA
3v
220R
red green orange LED
5mA
3v
100R
red green orange LED
10mA
3v
68R
red green orange LED
15mA
3v
47R
red green orange LED
20mA
3v
39R
red green orange LED
25mA
5v
3k3
red green orange LED
1mA
5v
560R
red green orange LED
5mA
5v
330R
red green orange LED
10mA
5v
220R
red green orange LED
15mA
5v
150R
red green orange LED
20mA
5v
120R
red green orange LED
25mA
9v
6k8
red green orange LED
1mA
9v
1k5
red green orange LED
5mA
9v
680R
red green orange LED
10mA
9v
470R
red green orange LED
15mA
9v
330R
red green orange LED
20mA
9v
270R
red green orange LED
25mA
12v
10k
red green orange LED
1mA
12v
2k2
red green orange LED
5mA
12v
1k
red green orange LED
10mA
12v
560R
red green orange LED
15mA
12v
470R
red green orange LED
20mA
12v
390R
red green orange LED
25mA

Fig 31. These diagrams show the resistor needed to produce 1mA to 25mA current through a single LED on 3v, 5v, 9v and 12v supply.

to Index

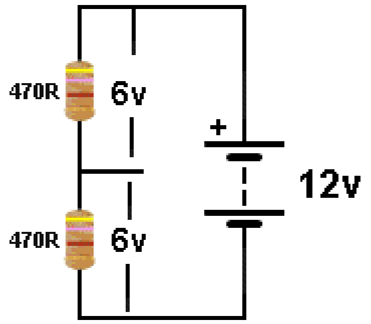
470R
6v
+
12v
470R
6v

Fig 31a. Voltage Divider Circuit.

THE VOLTAGE DIVIDER

In the circuits above, the resistor and LED are forming a VOLTAGE DIVIDER.
A red LED is dropping 1.7v across it and the resistor is dropping the remaining voltage.
Whenever two (or more) components are placed across a battery, they form a **VOLTAGE DIVIDER**.
Sometimes we want a 6v supply and only have 12v.
We can produce the 6v supply by putting two equal-value resistors across the 12v as shown in the circuit opposite.
We are not going into the mathematics because the selection of the correct value is very complex and the circuit is very wasteful.

to Index

THE CURRENT DIVIDER

The **CURRENT DIVIDER CIRCUIT** is actually a

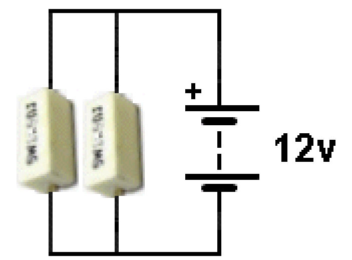

Fig 31b. Current Divider Circuit.

LOAD SHARING CIRCUIT.
Suppose you are testing a Power Supply and need a 10 watt LOAD. But you only have 5 watt resistors.
Placing two 5watt resistors in parallel across the output of the power supply will allow half the current to flow though each resistor. This is called **CURRENT SHARING** or **LOAD SHARING** and the current is divided (or passed) through each resistor according to the value of resistance.

to Index

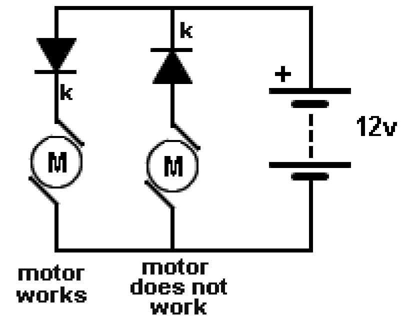

Fig 32. **The DIODE**

THE DIODE
The next simple electronic component is the **DIODE.**
It only works when connected correctly.
A DIODE allows current to flow through it when it is connected as shown in the diagram.
A Diode is similar to a one-way water valve.
When the diode is "facing down," the motor spins.
When it is "facing up" the motor does not spin.

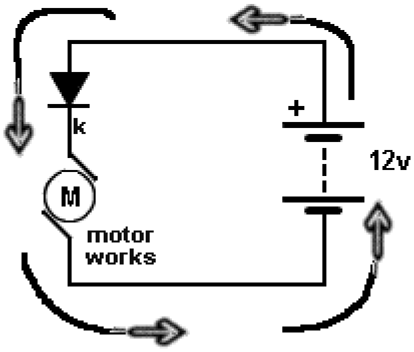

The diagram shows the "current path" around the circuit. The current is measured in AMPS and we discuss current as CONVENTIONAL CURRENT. This is the way current was thought to flow when electricity was born and they said it flows out the POSITIVE TERMINAL of the battery, around the circuit and into the NEGATIVE TERMINAL.
The arrow on the diode shows the current will flow through the diode and allow the motor to spin.
The diode is said to be **FORWARD BIASED**.

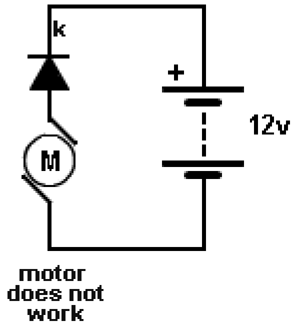

There is no flow of current because the diode prevents any current-flow when connected as shown.
The motor DOES NOT WORK.
The diode is said to be **REVERSE BIASED**.

to Index

There are hundreds of different types of diodes.
Power diodes, signal diodes, low voltage diodes, high voltage diodes, high-speed diodes and many other types.
They all do one thing.
They pass current in one direction and if turned around, they **do not pass** any current.

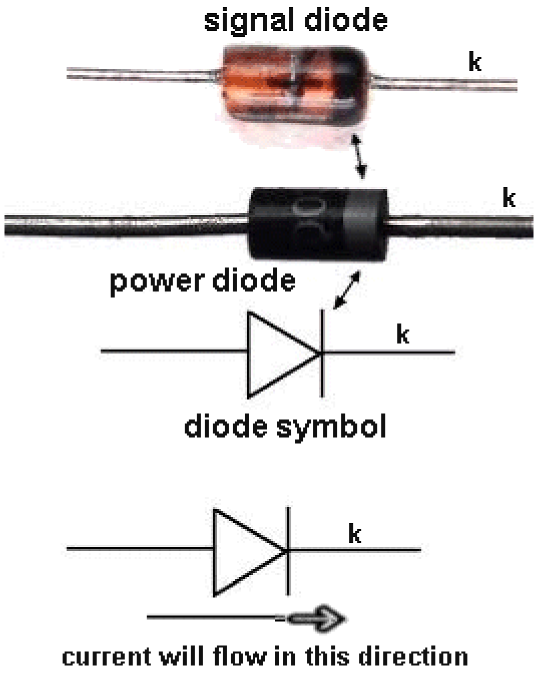
signal diode
k
k
power diode
k
diode symbol
k
current will flow in this direction

Fig 33. A Signal diode and a Power diode

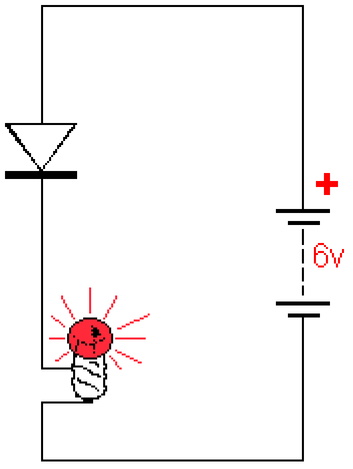

How a diode works

to Index

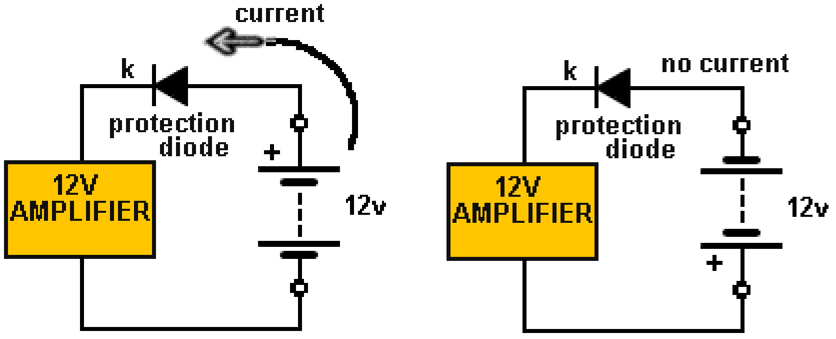
current
k
protection diode
+
12V AMPLIFIER
12v
k
no current
protection diode
12V AMPLIFIER
12v
+

Fig 34. A Protection diode

Diodes perform many function in electrical and electronic circuits. Here is an application as a **PROTECTION DIODE.**
It protects the amplifier. If the 12v battery is connected around the wrong way, no current will flow.

to Index

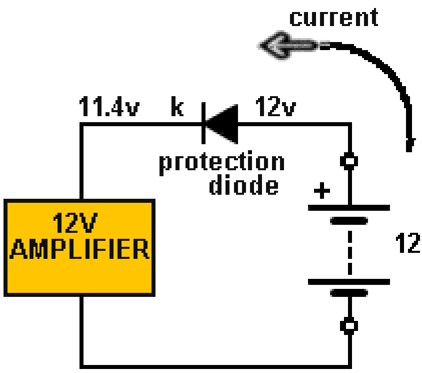
current
11.4v
k
12v
protection diode
+
12V AMPLIFIER
12v

Fig 35. Diode "Voltage Drop"

When a diode is placed in a circuit (and current is flowing), a small voltage develops across the diode. This voltage is called the **FORWARD VOLTAGE DROP.**
This voltage is approximately 0.6v.
This is due to a junction inside the diode where two different materials are joined.
Normally, this voltage is not important because it is only small, but sometimes you need to take it into account.
For the circuit above, the amplifier only gets 11.4v

to Index

Voltage is measured with a

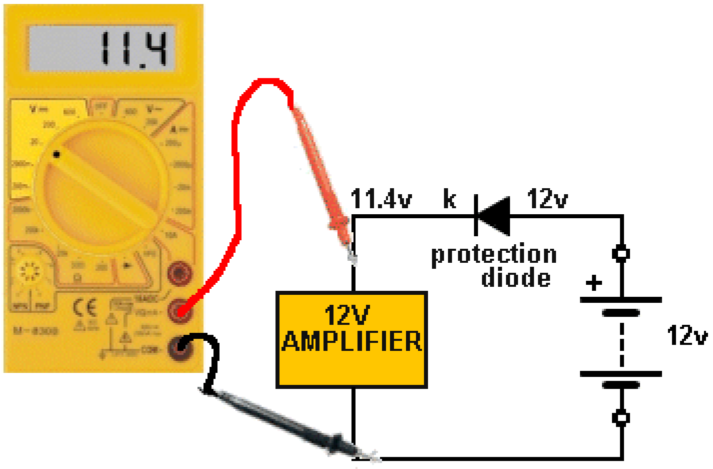
11.4
11.4v
k
12v
protection diode
+
12V AMPLIFIER
12v

Fig 36. Measuring Voltage with a Digital Multimeter

VOLTMETER.
Multimeters have 2 or 3 voltage ranges so you can measure low voltage (0v to 20v), medium voltages (0v to 200v and high voltages (0v to 500v).
A voltmeter is placed across the component being tested, as shown in the diagram.
The Digital Multimeter is detecting 11.4v across the amplifier.

to Index

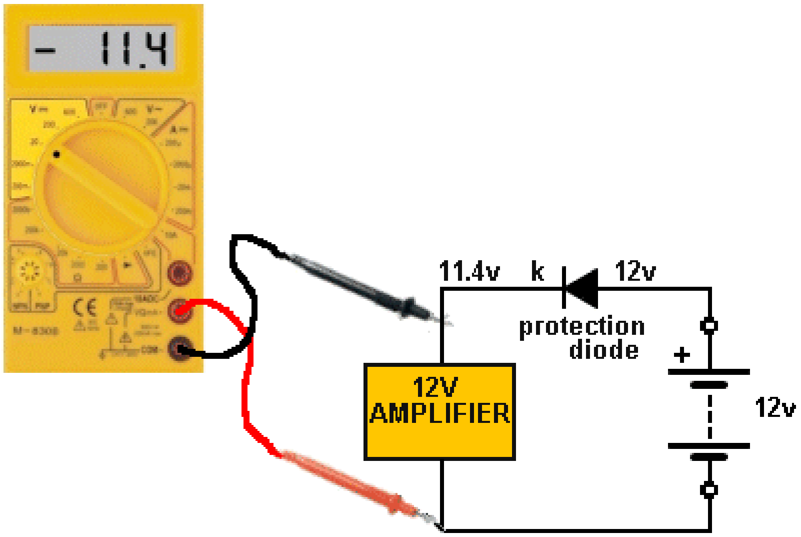
- 11.4
11.4v
k
12v
protection diode
+
12V AMPLIFIER
12v

Fig 37. Measuring Voltage with a Digital Multimeter

If you place the probes of a digital multimeter around the wrong way on a component, the display will show a "-" The meter will not be damaged.

to Index

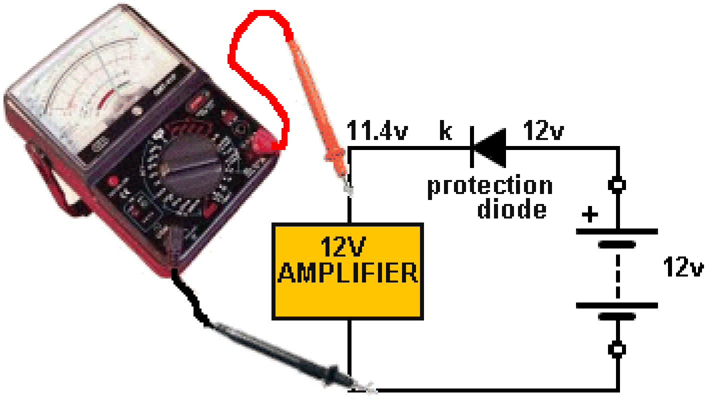
11.4v
k
12v
protection diode
+
12V AMPLIFIER
12v

Fig 38. Measuring Voltage with an Analogue Multimeter

An analogue Multimeter must be connected around the correct way to make the pointer move "up scale." Select the range that will allow the pointer to show somewhere in the middle of the scale.

to Index

Current is measured by

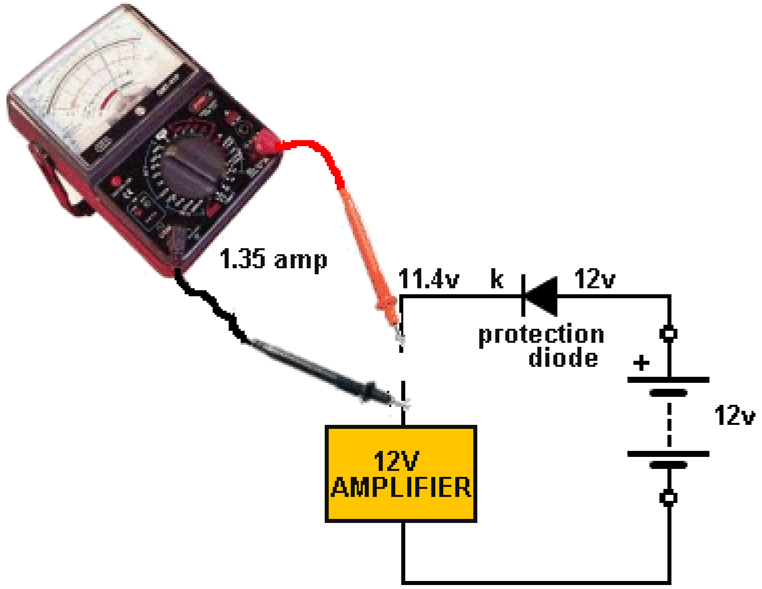
1.35 amp
11.4v
k
12v
protection diode
+
12v
12V AMPLIFIER

Fig 39. Measuring Current with an Analogue Multimeter

"breaking into the circuit" and inserting the leads so the positive probe is closest to the positive of the battery.
If you connect the leads around the other way, the needle will not move but it will hit the "end stop" and you may have to "bump" the meter to get the pointer to move from its jammed position.

to Index

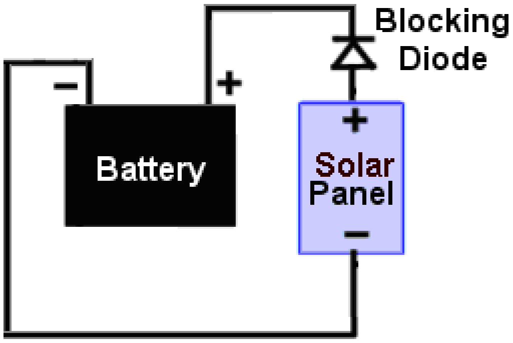
Blocking Diode
−
+
Battery
+
Solar Panel
−

Fig 39a. The Solar Panel and the Blocking Diode

A DIODE is also used with a solar panel to prevent the battery discharging into the solar panel when the sun is not shining.
When the solar panel is not receiving any light it becomes a resistor with a large value and a small current can flow through it from the battery. The diode prevents this current-flow. The diode is called a **BLOCKING DIODE**.

to Index

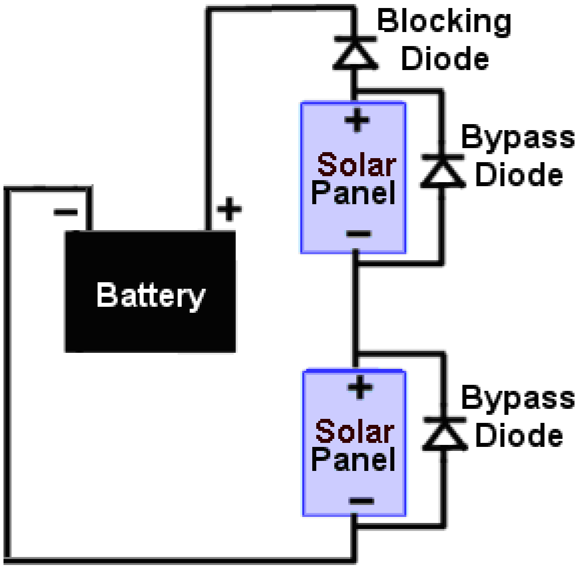
Blocking Diode
+
Solar Panel
−
Bypass Diode
−
+
Battery
+
Solar Panel
−
Bypass Diode

Fig 39b. The Solar Panel with diodes

Diodes are given different names, according to their function. They all perform the same job by passing current in one direction and prevent current-flow in the opposite direction.
When the top solar panel is shaded by a cloud, it generates less current and this will reduce the current into the battery.
By placing a diode across the panel, the diode will pass the current produced by the lower panel to the battery.
These diodes are called **BYPASS DIODES.**

called BLOCKING DIODES and BYPASS DIODES.

to Index

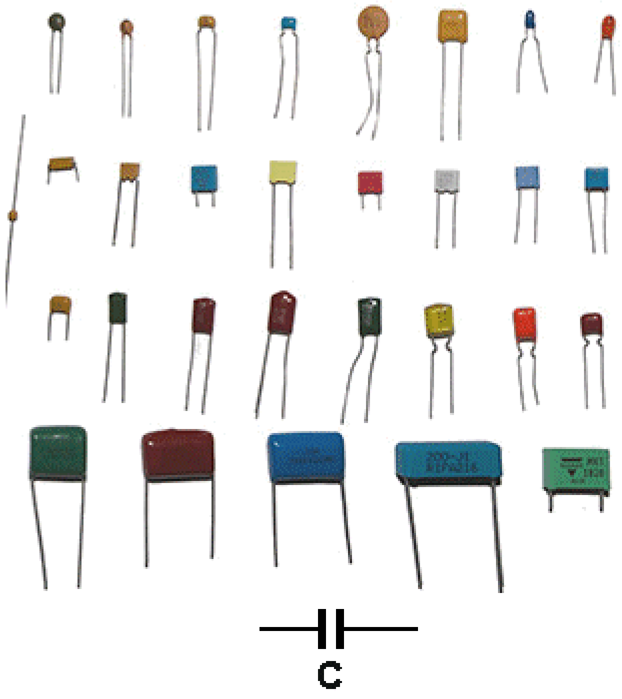

Fig 40. Capacitors and symbol

The next component we cover is the **CAPACITOR**.
There are thousands of different types of capacitor.
Each value of capacitor can have a low voltage rating, medium voltage or high voltage.
Capacitors can be very small in size and shape or very stable with temperature-rise or simply very cheap to make.
A capacitor consists of two thin sheets of metal such as aluminium with a thin sheet of plastic between.
The sheets may be rolled up in a cylinder or laid on top of each other.
The fact is this: the top sheet of metal does not touch the bottom sheet. This is shown in the symbol.
The resistance between the two terminals is INFINITE.
The 6th capacitor in the top row is called a **MONOBLOCK**.

to Index

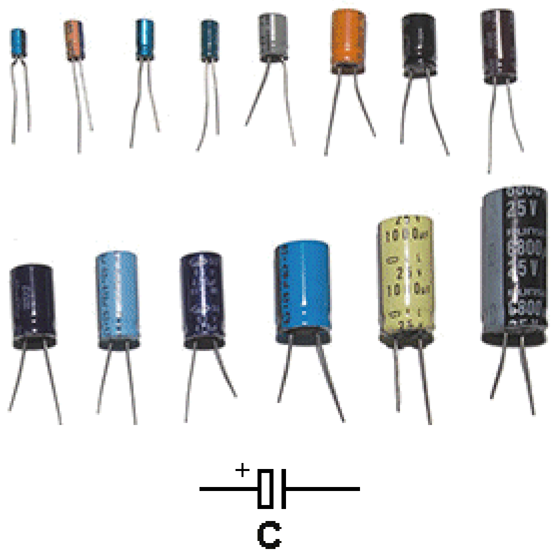

Fig 41. Electrolytic Capacitor

A capacitor gets bigger as its value increases.
It also gets bigger when the voltage-rating increases.
The basic unit of capacitance is the FARAD. A one-farad capacitor would be the size of a house. To make the capacitor smaller the sheets are etched to increase the surface-area and different insulating materials are used between the sheets.
The result is a capacitor called an **ELECTROLYTIC**. It is a bit like a rechargeable battery. It stores a lot of energy in a small space.
The negative lead is shorter and has a black stripe on the side of the electrolytic.

to Index

One FARAD is too big to handle. We use smaller values.
The middle of the range is one microfarad. This is written as 1u. (sometimes you see uF)
This is one-millionth of a FARAD.
The smallest value of capacitance is one picofarad. This is one millionth of a microfarad.
It is written as 1p.
Capacitors are broadly separated into two groups. 1p to 1u and 1u to 100,000u
Capacitors 1p to 1u are ceramic, polyester, air, styroseal, monoblock and other names.
Capacitors 1u to 100,000u are electrolytic or tantalum. A tantalum is the same as an electrolytic - for testing purposes - it is a more-compact electrolytic.

1 microfarad is one millionth of 1 farad.
1 microfarad is divided into smaller parts called nanofarad.
1,000 nanofarad = 1 microfarad
Nanofarad is divided into small parts called picofarad
1,000 picofarad = 1 nanofarad.

Recapping:
1p = 1 picofarad. 1,000p = 1n (1 nanofarad) 1,000,000p = 1u
1,000n = 1u (1 microfarad)
1,000u = 1millifarad
1,000,000u = 1 FARAD.

Examples:
All ceramic capacitors are marked in "p" (puff")
A ceramic with 22 is 22p = 22 picofarad
A ceramic with 47 is 47p = 47 picofarad
A ceramic with 470 is 470p = 470 picofarad
A ceramic with 471 is 470p = 470 picofarad
A ceramic with 101 is 100p (it can also be 100)
A ceramic with 102 is 1,000p = 1n
A ceramic with 223 is 22,000p = 22n
A ceramic with 104 is 100,000p = 100n = 0.1u A common 100n is called a **MONOBLOCK**.
A ceramic with 105 is 1u

TYPES OF CAPACITOR
For testing purposes, there are two types of capacitor.
Capacitors from 1p to 100n are non-polar and can be inserted into a circuit around either way.
Capacitors from 1u to 100,000u are electrolytics (or tantalum) and are polarised. They must be fitted so the positive lead goes to the supply voltage and the negative lead goes to ground (or earth).

to Index

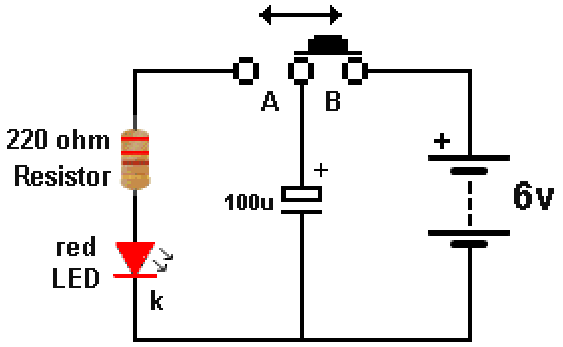

Fig 42. Charging a Capacitor

Here is an experiment to show how much (little) energy is stored in a 100u electrolytic.
When the slide-switch is in position "B", the 100u is charged by the 6v battery.
When the slide switch is moved to position "A" the electrolytic supplies energy to illuminate the red LED via the 220R resistor. It will illuminate for a short period of time.
By moving the switch back and forth, you can keep the LED illuminated.

to Index

Capacitors can be connected in **Series or Parallel** to obtain a value of capacitance you may not have available.
They are also connected in series to increase the effective **VOLTAGE RATING**.
However when two equal-value capacitors are connected in series, the final value is HALF, and thus you need two with double the final-value to get a value with an increased voltage-rating.
When two equal-value capacitors are connected in series, the result is HALF.
(This is the opposite to connecting resistors)

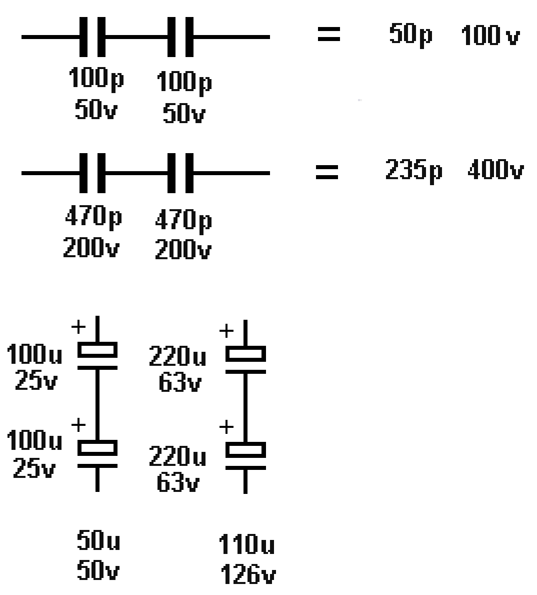

Fig 43. Capacitors in Series

to Index

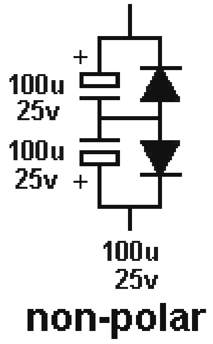

Fig 43a. Non-polar Capacitor

Non-polar Capacitor (electrolytic)

A normal electrolytic must be connected the correct way in a circuit because it has a thin insulating layer covering the plates that has a high resistance.

If you connect the electrolytic around the wrong way, this layer "breaks-down" and the resistance of the electrolytic becomes very small and a high current flows. This heats up the electrolytic and the current increases. Very soon the capacitor produces gasses and explodes.

One big mistake in many text books shows how to make a non-polar electrolytic by connecting two "back-to-back."

They claim 2 x 100u connected back-to-back is equal to 47u.

This appears to be case when testing on a meter but the meter simply charges them for a short period of time to get a reading.

If you allow them to charge fully you will find the reverse electrolytic has a very small voltage across it.

Secondly, when you are charging them, you are putting a high current through the reverse electrolytic and damaging the layer.

To prevent this, you need to add two diodes as shown in the diagram.

In addition, 2 x 100u "back-to-back" is very near 100u.

to Index

Capacitors can be connected in **Parallel** to obtain a value of capacitance you may not have available.
(This does not change the **VOLTAGE RATING**.)
When two equal-value capacitors are connected in parallel, the result is DOUBLE.
(This is the opposite to connecting resistors).
If one electrolytic is 25v

and the other 63v, the answer is the LOWER VOLTAGE = 25v.

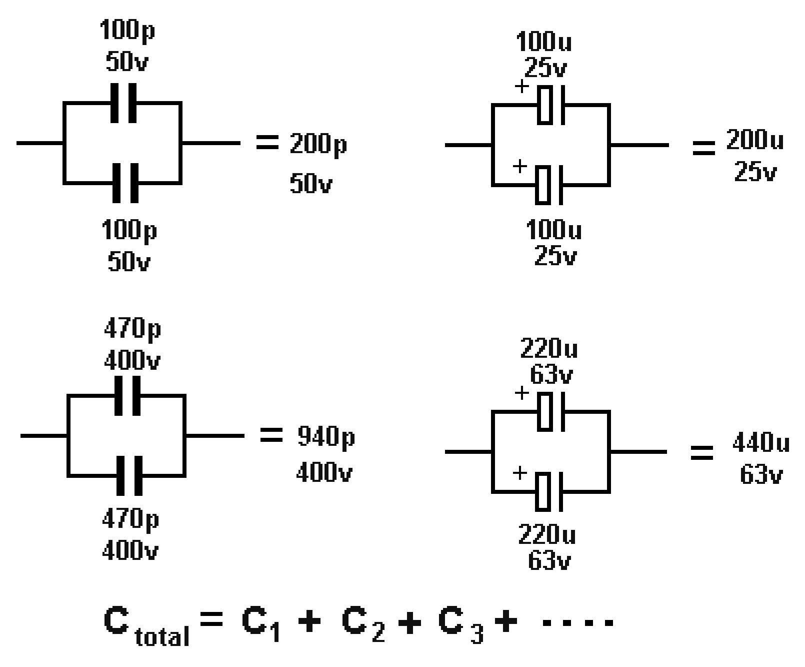
100p
50v
100p
50v
= 200p
50v
100u
25v
100u
25v
= 200u
25v
470p
400v
470p
400v
= 940p
400v
220u
63v
220u
63v
= 440u
63v
$C_{total} = C_1 + C_2 + C_3 + \cdots$

Fig 44. Capacitors in Parallel

to Index

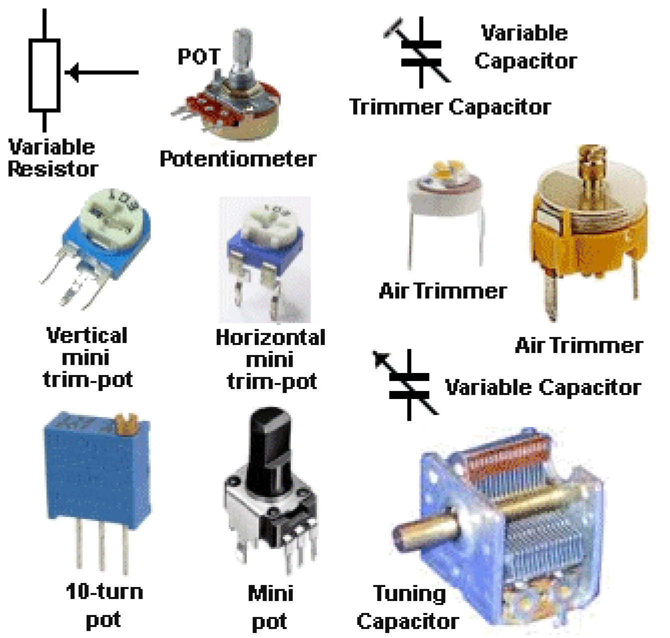
Variable Resistor
POT
Potentiometer
Variable Capacitor
Trimmer Capacitor
Vertical mini trim-pot
Horizontal mini trim-pot
Air Trimmer
Air Trimmer
Variable Capacitor
10-turn pot
Mini pot
Tuning Capacitor

Fig 45. Variable Capacitors and Resistors

The value of a capacitor or resistor may need to be increased or decreased in a circuit to tune in radio stations or increase and decrease the volume of a speaker. The symbol for these components have an arrow to show they can be adjusted.
The resistance of a potentiometer can be from 1 ohm to 5M
They come in many different shapes and sizes to suit the PC board or front-panel layout.
The "T" represents a trimmer capacitor and this can be from 1p to about 120p.
A variable capacitor will be from about 10p to 415p.

to Index

Simple CIRCUITS
We have covered enough symbols and components to create a number of simple circuits.
These circuits will show how to connect a motor, a LED, (how to make it bright or dull) and how to connect 4 cells to make a battery.

Note: The flashing LED

does not need a resistor because a resistor and chip are inside the LED, to make it flash and control the current.

Connect all the components around the correct way and then connect them around the wrong way to see what happens.
Connect the flashing LED in series with a red LED and see what happens.

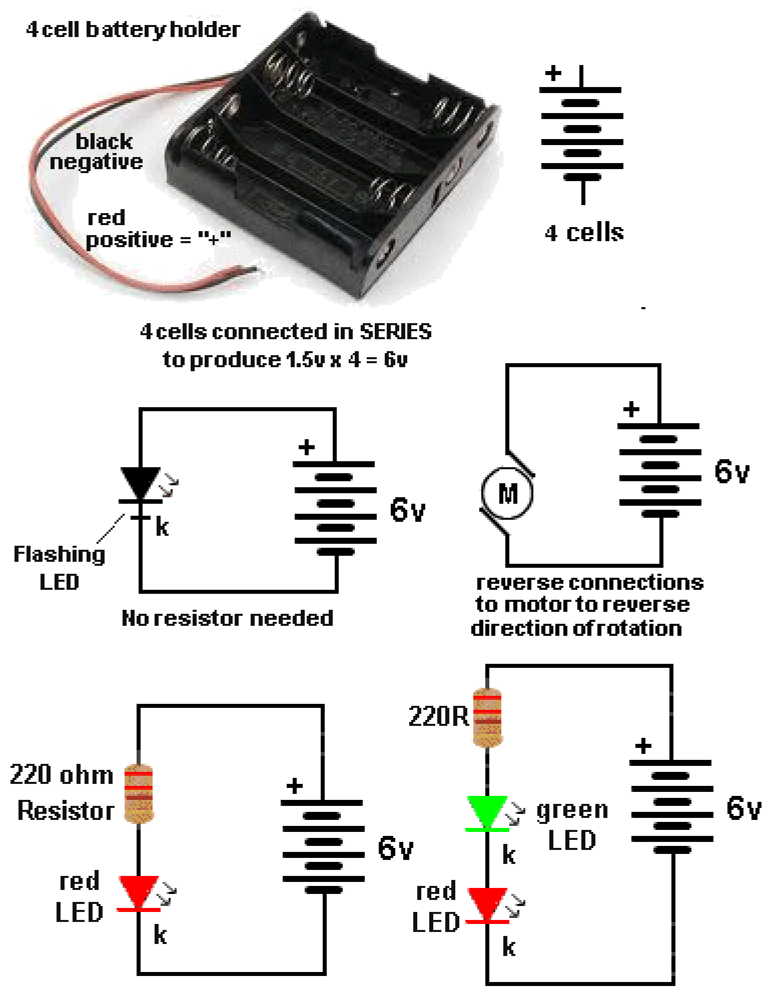
4 cell battery holder
black
negative
red
positive = "+"
+
4 cells
4 cells connected in SERIES
to produce 1.5v x 4 = 6v
+
6v
k
Flashing
LED
No resistor needed
+
M
6v
reverse connections
to motor to reverse
direction of rotation
220 ohm
Resistor
red
LED
k
+
6v
220R
+
green
LED
k
6v
red
LED
k

Fig 46. Simple Circuits

to Index

QUESTIONS

1. Explain why the Flashing LED circuit has no external resistor.
2. How many 1.5v cells are needed to produce a 6v battery
3. Explain what happens when you reverse the leads to a motor.
4. Identify the positive terminal:

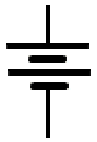

5. Can 3 green LEDs be connected in series to a 6v supply?
6. A variable resistor is also called: ____________________
7. The combined resistance of two 1k resistors in series is: ______________
8. The combined resistance of two 1k resistors in parallel is: ______________
9. Name the short lead on a LED
10. Name the type of multimeter with a pointer and scale: ____________________
11. The total capacitance of two 100u electrolytics in series is: _____________
12. The total capacitance of two 100u electrolytics in series is: _____________
13. Write these values in words:

22R _________
4k7 _________
1n ___________
100u ___________

14. How many 1.5v cells in a 9v battery?

15. The red probe is: _______________ (positive/negative)
16. Conventional current flows from: _________ (positive to negative / negative to positive)
17. Make a 470u electrolytic with two electrolytics:
18. What is the voltage drop across a diode?
19. Name the component that only allows current to flow in one direction: ____________
20. Name this symbol:

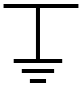

21. When resistors are connected in series, the resistance of the combination: _________ (increases / decreases)
22. When two capacitors are connected in parallel, the voltage-rating of the combination: _______________ (increases / equal to the capacitor with the lowest voltage-rating)
23. Draw two 2k2 resistors in parallel.
24. Which is larger: 470R or 22k
25. What is the value of this combination:

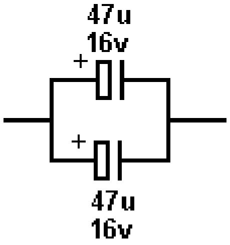

26. What is the name of the resistor in series with a LED: ____________________
27. What is the voltage across the amplifier: _________

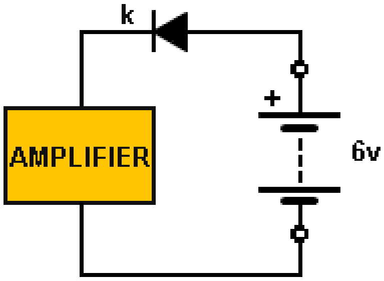

28. Identify the fault with these circuits:

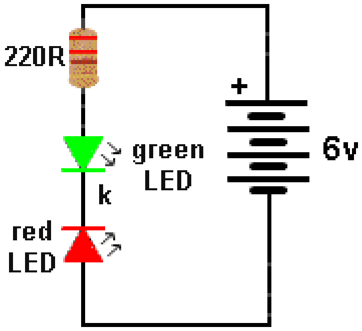

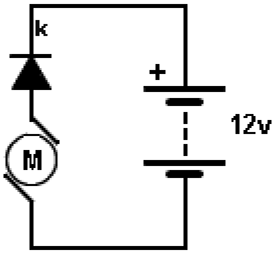

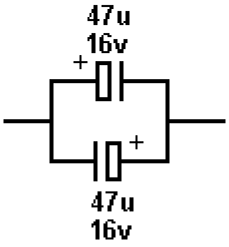

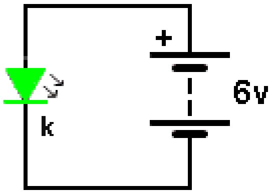

to Index

Passive Components
All the components shown

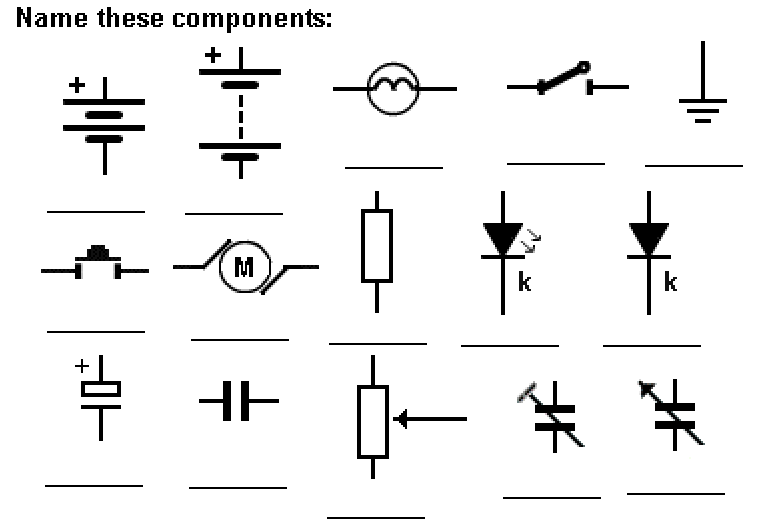

Fig 47. Passive Components

on the left are called PASSIVE COMPONENTS. This means they do not amplify.
Write the name beside each symbol.

to Index

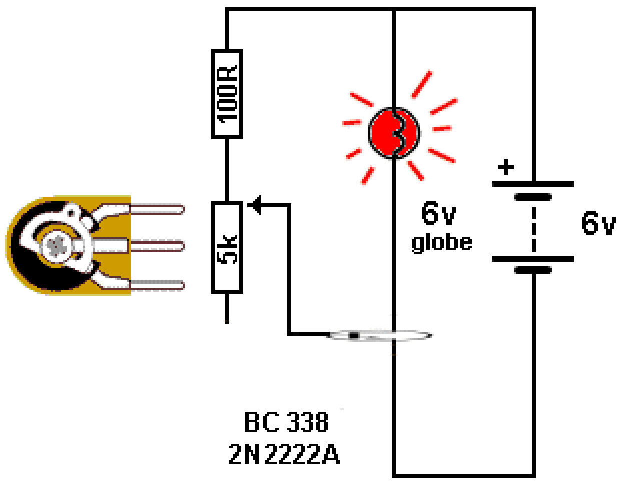

Fig 47a. The Potentiometer in action

The Potentiometer is a variable resistor.
It consists of a curved carbon track with a wiper that touches the track and can be turned via a screwdriver or knob.
The wiper is the middle wire on the circuit symbol and it moves up and down as shown in the animation.
When the three leads are connected the symbol is called a potentiometer.
When two leads are connected it is a variable resistor.
When the resistance increases, less current flows through the pot.

to Index

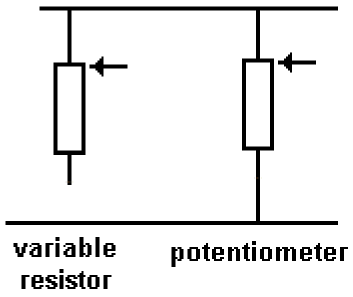

Fig 47b. The Variable Resistor and Potentiometer

There is a difference in operation between a **Variable Resistor** and a **Potentiometer**.
Both will increase or decrease the sound level as a volume control or the speed of a motor or the brightness of a globe, but a Potentiometer will guarantee zero volume or zero brightness when the pot is turned fully anticlockwise (as shown in the animation).
This is because the output will be zero volts, but the variable resistor may still deliver some "energy" (voltage and current) to the circuit when turned fully anticlockwise.

to Index

Potentiometers come in values from 100 ohms to 5 Meg ohms (500R, 1k, 2k, 5k, 10k, 50k, 100k, 200k, 250k, 500k, 1M are most popular).
They come as linear, or logarithmic where the resistance of the track (per mm) is higher at one end. Because our hearing is not linear, these pots can be used as volume controls to produce a gradual (very nearly linear) increase in volume.
Selecting the correct value of resistance for a circuit is VERY complex. If the value is not correct, the volume will not be loud or it will drop to zero before the pot is turned fully anticlockwise. Or

the motor will drop to zero at mid-turn of the pot or it will not reduce in RPM to the desired amount.
The simple answer is to copy a circuit.
Or you can try the whole range of pots and you will find one value is the best.
A Potentiometer can be used in hundreds of different circuits to produce hundreds of different effects, but the actual "thing" that flows between the input and output is a percentage of the voltage. At the same time the current will also be passed to the output at a reduced value. A pot actually delivers BOTH reduced values at the same time and the receiving circuit will be designed to "look for" the change in voltage or current. If the supply voltage is not rising or falling, the "values" are called DC values.
The voltage can also be in the form of a signal (volume). This is called an AC signal.

to Indexx

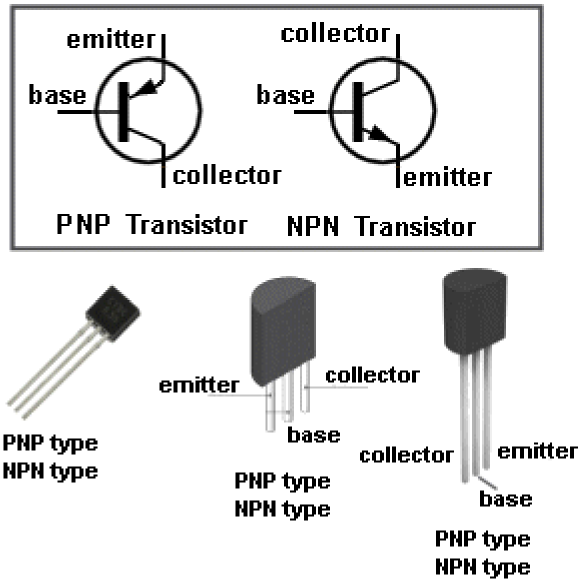
emitter
base
collector
PNP Transistor
collector
base
emitter
NPN Transistor
PNP type
NPN type
emitter
collector
base
PNP type
NPN type
collector
emitter
base
PNP type
NPN type

Fig 48. The TRANSISTOR

The TRANSISTOR

A TRANSISTOR is an ACTIVE device. It AMPLIFIES.
There are many types of transistor (over 20,000 different types) from hundreds of manufacturers and they have many different names.
We are going to study the simplest.
It has the technical name **BIPOLAR JUNCTION TRANSISTOR (BJT)** but we are going to call it a **TRANSISTOR**.
There are two types in this group: **PNP** and **NPN**.

The type we will study is also called a **SMALL-SIGNAL TRANSISTOR.**

You cannot tell an **NPN** transistor from **PNP** by looking at it. You must test it in a circuit.
In **Fig 65** you will make a **Transistor Tester** project, but first some basic facts:

to Index

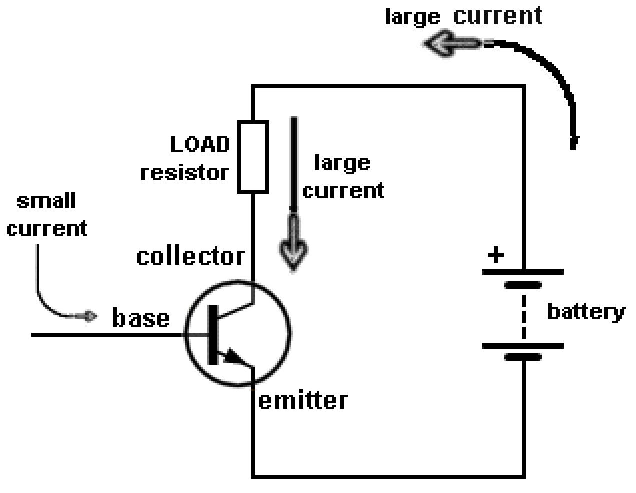
large current
LOAD resistor
large current
small current
collector
base
emitter
+
battery

Fig 49. The NPN TRANSISTOR in a Circuit

The first type of transistor we are going to study is the **NPN**.
A transistor has three leads:
BASE, COLLECTOR and **EMITTER**
Basically, a small current enters the base and a large current flows through the collector-emitter leads as shown in the diagram.
The resistor in the collector lead is called the LOAD Resistor.
Sometimes the load is a speaker.

to Index

The transistor is similar to the diagram opposite. A small drop of water entering the base is amplified to produce a loud DRIP from the speaker.

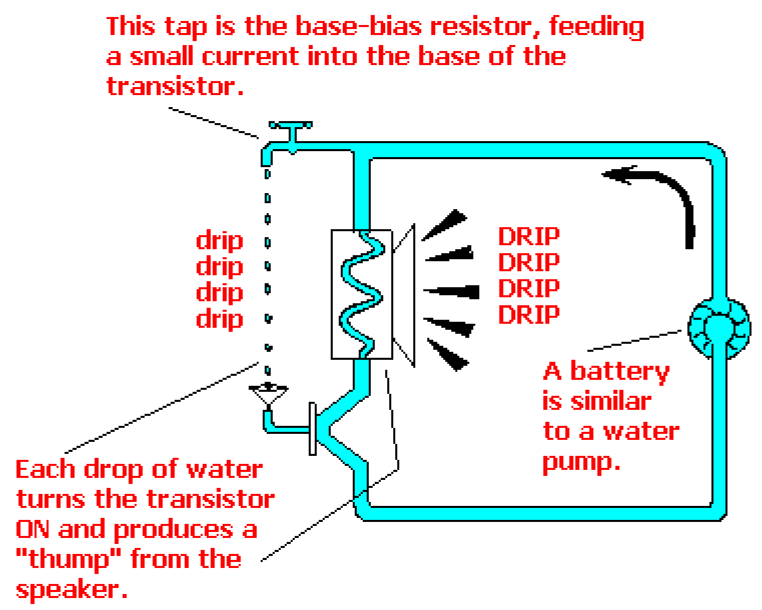
This tap is the base-bias resistor, feeding a small current into the base of the transistor.
drip
drip
drip
drip
DRIP
DRIP
DRIP
DRIP
A battery is similar to a water pump.
Each drop of water turns the transistor ON and produces a "thump" from the speaker.

Fig 50. The Water TRANSISTOR

to Index

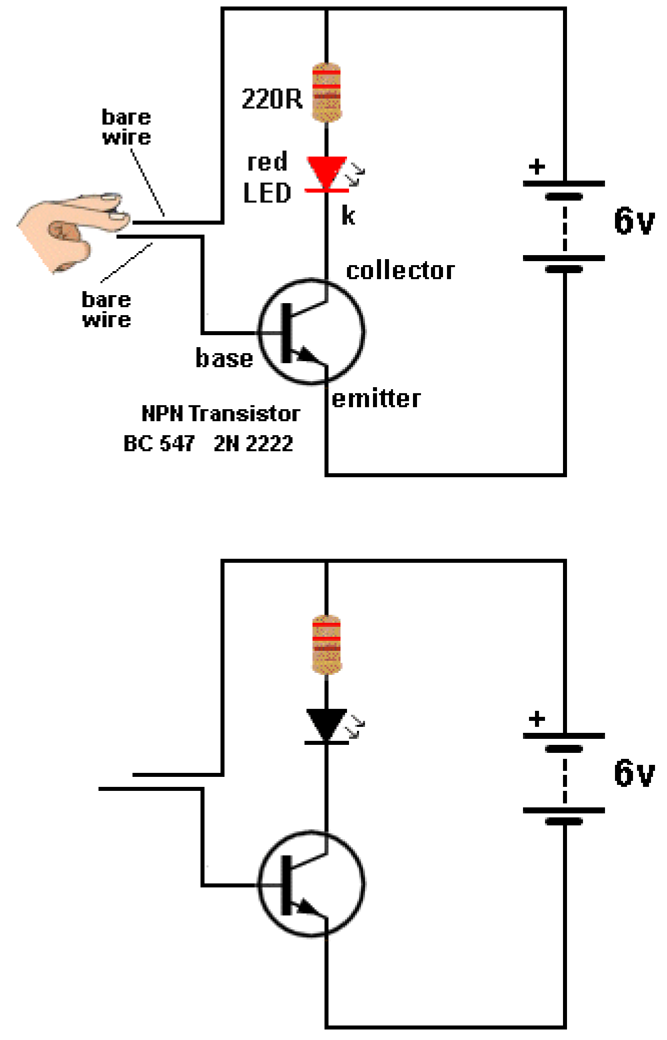
220R
bare wire
red LED
k
+
6v
collector
bare wire
base
emitter
NPN Transistor
BC 547 2N 2222
+
6v

Fig 51. One Transistor Circuit

In this experiment we will construct a ONE TRANSISTOR circuit similar to the WATER TRANSISTOR above and observe the results.
Make sure the two leads DO NOT TOUCH. If they touch, the transistor will be DESTROYED.

The transistor is amplifying the current through your finger via the two leads and it will be very dim.

ANIMATION

The lower diagram shows the transistor turning ON when a finger is pressed against the two wires. The finger produces a resistance that turns the transistor ON and this turns the transistor into a smaller and smaller resistor. That's how more and more current flows through the LED and it gets brighter and brighter.

to Index

By adding another transistor we amplify the

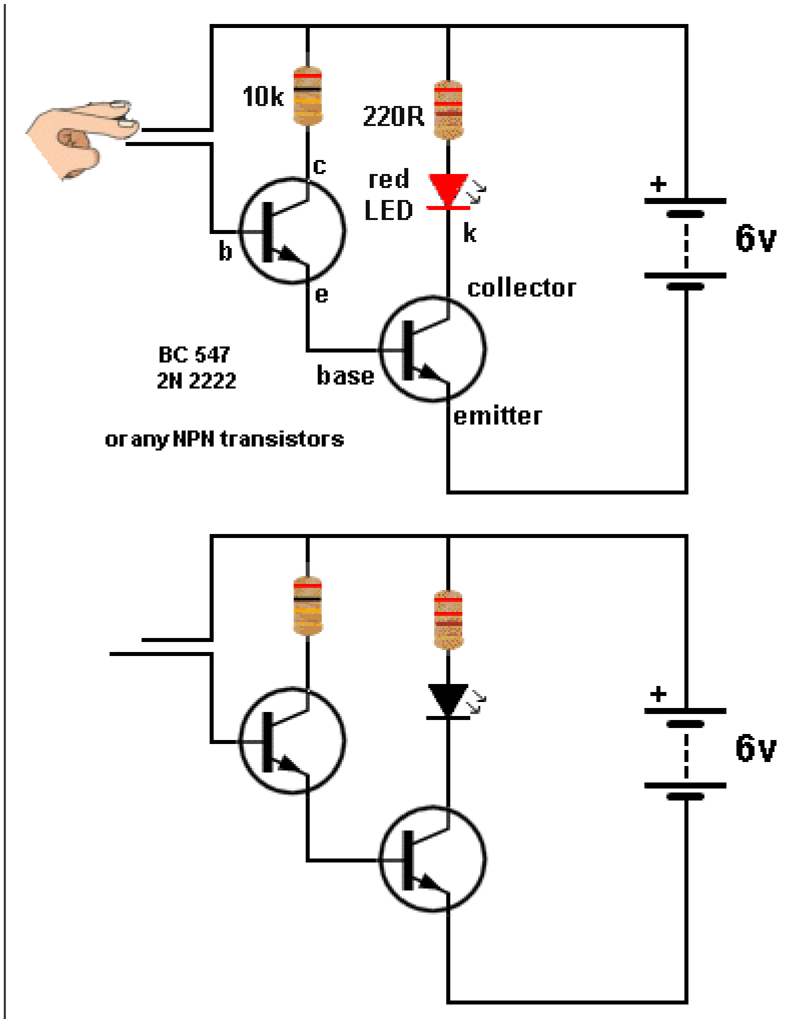
10k
220R
c
red
LED
k
b
e
collector
BC 547
2N 2222
base
emitter
or any NPN transistors
+
6v
+
6v

Fig 52. Two Transistor Circuit

current through the finger about 200 times and now the LED will glow bright.

Make sure the bare wires do not touch each other as this will destroy BOTH transistors.

ANIMATION

The lower diagram shows both transistors turning **ON** when a finger is pressed against the two wires.
They both becomes smaller and smaller resistors.
The first transistor allows more current to flow into the base of the second transistor and this is how the second transistor turns on more and more. This allows more current to flow through the LED and it gets brighter and brighter.

to Index

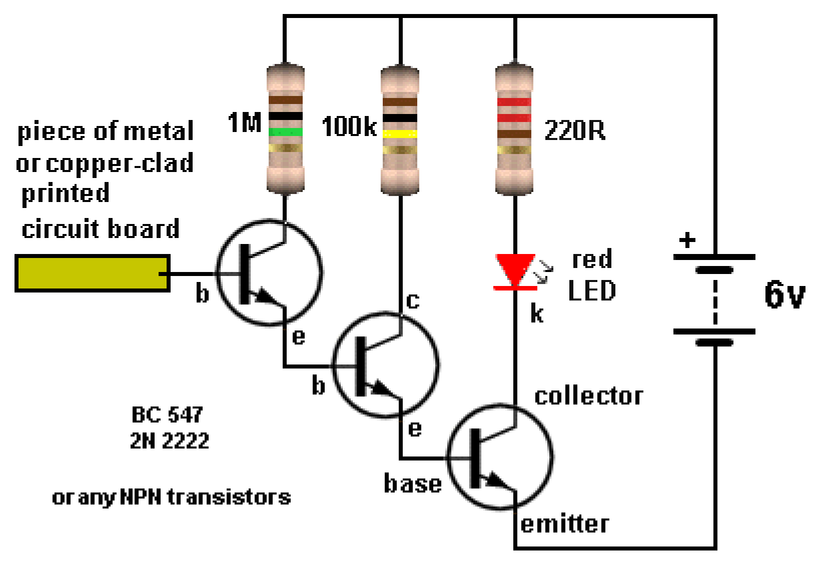
piece of metal
or copper-clad
printed
circuit board
1M
100k
220R
b
e
c
b
e
red
LED
k
+
6v
collector
BC 547
2N 2222
base
emitter
or any NPN transistors

This circuit has enormous gain.
Each transistor has a gain or more than 200 and the final gain will be more than:
200 x 200 x 200 = 8,000,000

8 MILLION!

The circuit is very sensitive to static voltages in the air or electrical waves such as the waveform produced by the electrical wiring in a house.
Move the project around a room and detect all the electrical signals.

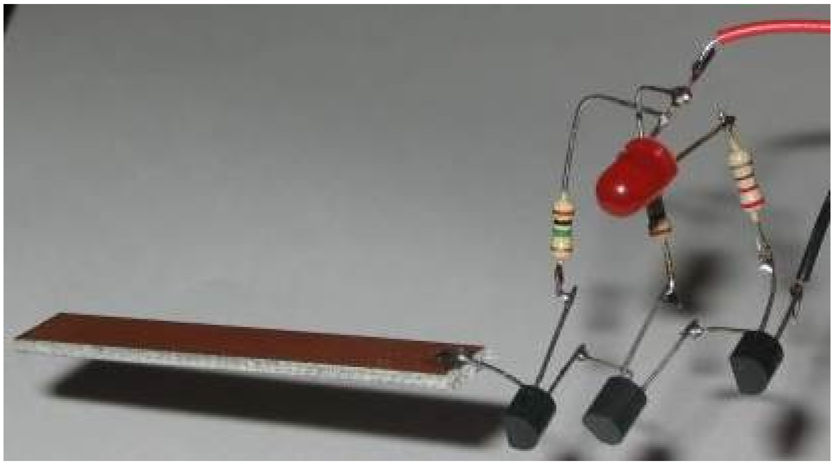

Fig 53. Three Transistors

to Index

You can see the effect of one transistor. It does not do much.
The two transistor circuit allows the resistance of your finger to deliver current into the base of the first transistor and this transistor delivers more current into the base of the second transistor. The result is more collector-emitter current and the LED illuminates.
The three transistor circuit produce an ENORMOUS effect.
It will pick up STATIC ELECTRICITY and all forms of electro-magnetic energy (radiation) and illuminate the LED.

to Index

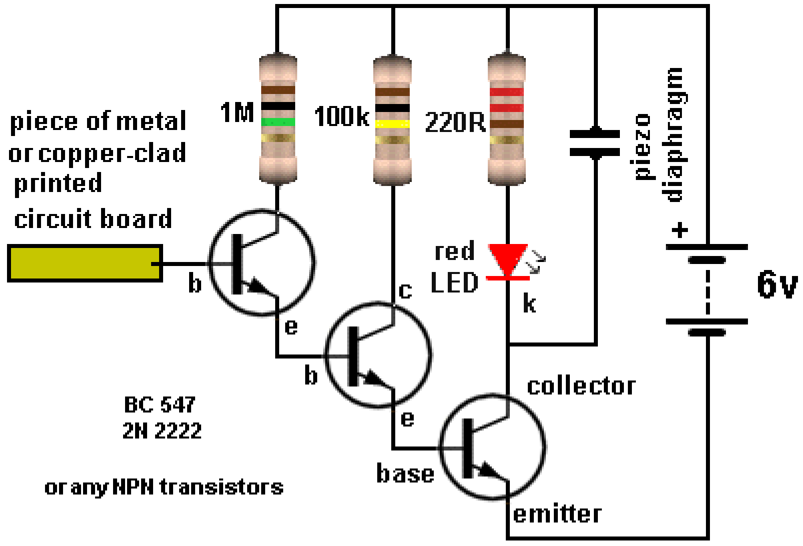
piece of metal
or copper-clad
printed
circuit board
1M
100k
220R
piezo
diaphragm
+
6v
red
LED
k
b
e
c
b
e
collector
base
emitter
BC 547
2N 2222
or any NPN transistors

Fig 54. Adding a PIEZO

By adding a piezo diaphragm to the output you will be able to hear the hum of the mains.
This is the frequency of the supply into your house. It will be either 50 cycles per second or 60 cycles per second.
The term: "cycles per second" is given the name HERTZ after **Heinrich Rudolf Hertz**, who was the first to prove the existence of electromagnetic waves.

to Index

The Piezo diaphragm is held around the outer edge inside a plastic case and when a voltage is applied to the two leads, the thin plate curves very slightly.
When the voltage is removed, the plate returns to its flat shape.
If the voltage is reversed, the plate curves in the opposite direction.

The curving is due to a thin layer of ceramic material under the plate and

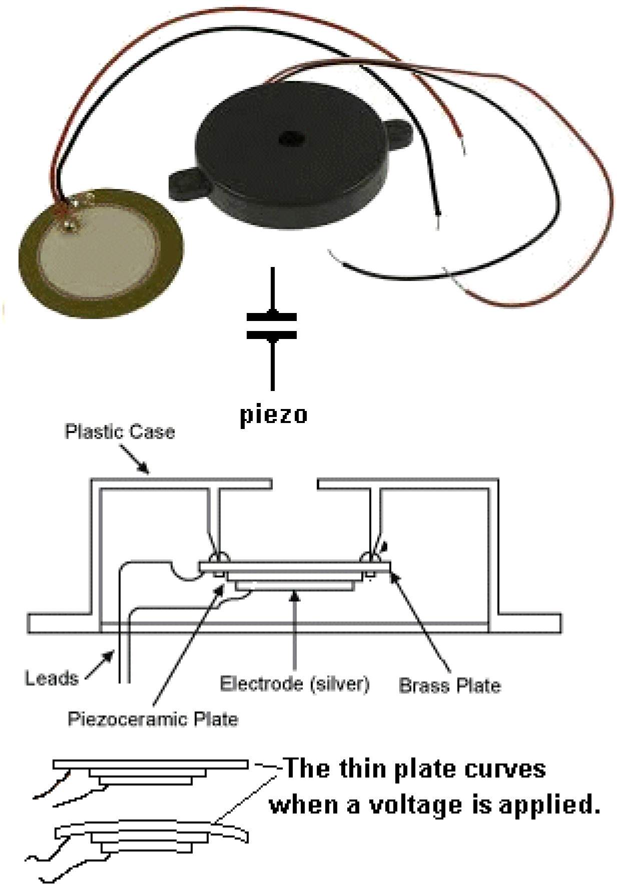
piezo
Plastic Case
Leads
Electrode (silver)
Brass Plate
Piezoceramic Plate

The thin plate curves when a voltage is applied.

Fig 55. The Piezo

then a film of metal is deposited onto the ceramic so a lead can be soldered.

There is infinite resistance between the two leads as the ceramic material is an INSULATOR.
The capacitance between the two leads is approx 22n.
The Piezo is a passive device. It needs a pulse or frequency for it to produce an output.

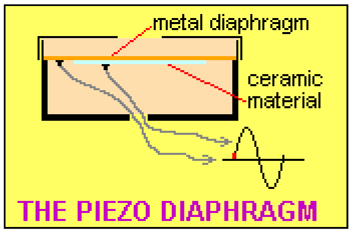
metal diaphragm
ceramic material
THE PIEZO DIAPHRAGM

to Index

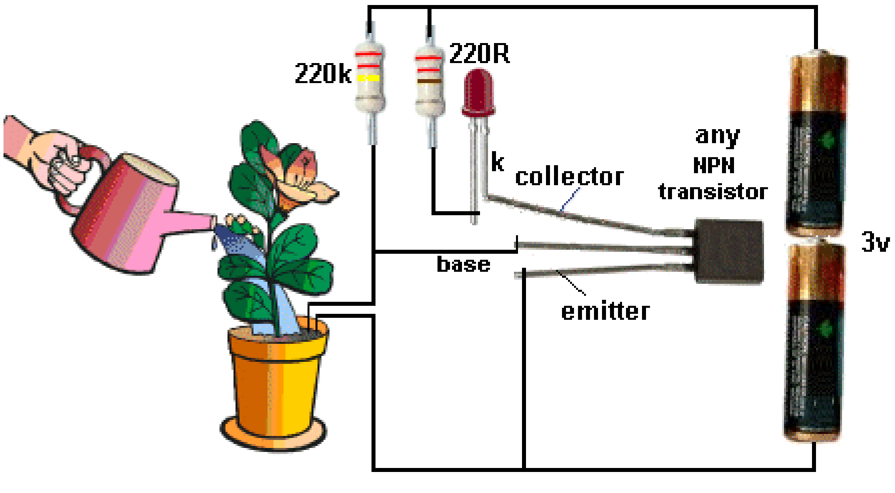
220k
220R
k
collector
any NPN transistor
base
emitter
3v

Fig 56. Plant Needs Water

The **ONE TRANSISTOR CIRCUIT** above can be turned into a detector to show when a plant needs water.
Place the two probes into the soil and water the plant. The LED will turn off. As the water evaporates the LED will turn **ON** to let you know the plant needs watering.

to Index

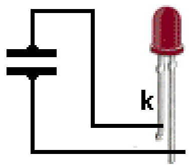

Fig 57. Tap the Piezo

This experiment produces a pulse from the piezo when it is tapped and the LED illuminates briefly.

The LED can be connected either way around.

This proves the diaphragm flexes when a voltage is applied and also in the reverse situation. A voltage is produces when the diaphragm is tapped.

to Index

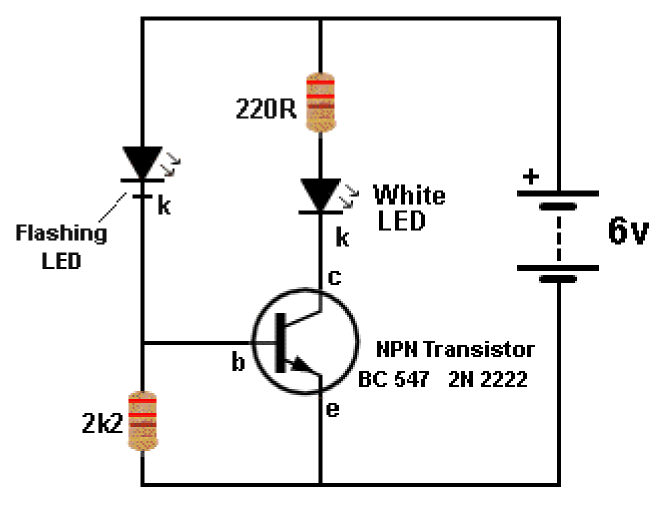

Fig 58. Flashing LED and White LED

A flashing LED is not very bright.
It can be connected to a transistor and the transistor will drive a very bright white LED.
The transistor is an amplifier.
It is amplifying the current flowing through the flashing LED and supplying a higher current for the white LED.

We cannot discuss any further details of the circuit at the moment because the actual operation of the circuit is quite complex.

At the moment we just need to experiment with simple transistor circuits.

to Index

Fig 59. Soldering Iron

We now come to the point of HELPING YOU WITH CONSTRUCTION.
We have already shown you 6 different circuits and there are many ways to build them.
You can:
1. Solder them.
2. Build them on an Experimenter Board
3. Connect the components with clips or twist the leads together.
It does not matter how you build the circuits.
The fact is this: YOU MUST START BUILDING.
The best soldering iron for a beginner is a CONSTANT TEMPERATURE soldering iron.
It has a dial that can be turned to set the desired temperature.
An ordinary soldering iron GETS TOO HOT. It is not suitable for soldering electronic circuits.
This is something that no-one has mentioned before. An ordinary soldering iron will melt the solder TOO QUICKLY and burn the resin inside the solder and make soldering very difficult for a beginner.
Soldering must be done slowly so the resin in the middle of the solder gets hot and cleans the leads of the components so the solder will "stick."
That's why you must apply the solder to the leads you are soldering and allow the resin to "attack" the leads and clean them.
The cheapest TEMPERATURE CONTROLLED soldering Iron is available on eBay for les than $10.00 (post FREE).
You will also need a small roll of solder (0.9mm) and a soldering Iron stand.
Email Colin Mitchell for links to eBay. (talking@tpg.com.au)
A whole book could be written on the ART OF SOLDERING.
Look on the web for articles and videos on SOLDERING.

to Index

SOLDERING

1. Soldering is very easy and very simple. You just need a **Temperature Controlled Soldering Iron**, **Fine solder** and clean components.
2. Remember this: It is **NOT** the solder you need for a joint, but the **FLUX**. And the flux lasts for only 2 seconds. When the flux is **HOT** it attacks and cleans the joint so that the solder will stick.
3. Turn ON the Temperature Controlled Soldering Iron to a low temperature. Put solder on the tip. It will not melt. Turn up the temperature slightly. Try more solder. As soon as the solder starts to melt, this is your starting point. Turn up the temperature slightly MORE and this is the correct temperature for **small, delicate, fine soldering**.
4. Place a component through a hole and bend the lead slightly so the component does not move. Turn the board over and touch the iron on the component and bring the solder **FROM THE OTHER SIDE** so the solder melts and flows towards the iron.
From start-to-finish, count one-two-three and remove the solder. Count four-five and remove the iron. You will have a perfect joint.
If you are soldering thick leads or large pads on a circuit board, you will need to turn the temperature UP slightly.
You must add enough solder to make the joint "bulge" slightly.
Fine solder (1mm or 0.9mm or 0.8mm) makes the best joint because it is easier to use.
Use a wet sponge to clean the tip or a ball of "Steel Wool." Steel wool is the best.
Here is the steel wool, bending the leads and some examples of poor joints due to insufficient solder:

Steel wool cleans the tip beautifully

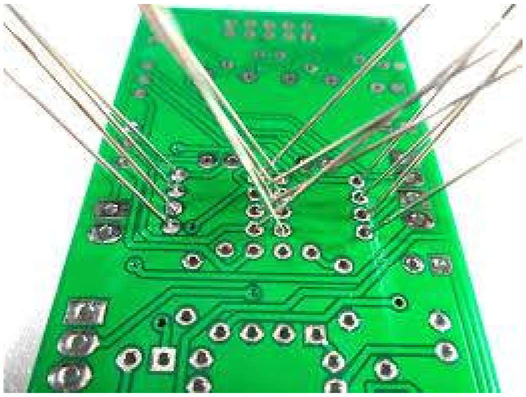

Bend the leads before soldering

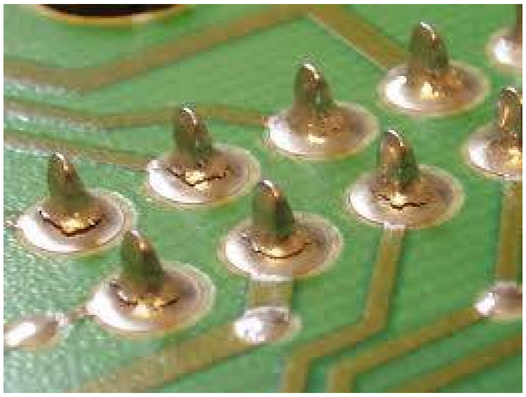

The joints do not have enough solder and that's why they fractured. Called a DRY JOINT.

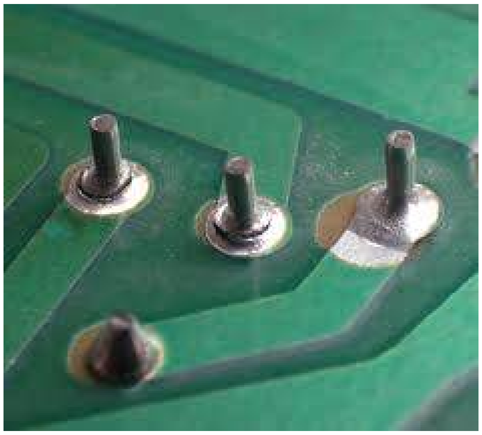

More "Dry Joints."

This is the cheapest and simplest soldering iron stand.

This stand is very messy as the spring grabs the iron and makes it difficult to remove from the stand. Test the stand before buying. You will se why not to buy this type of soldering stand. Get one with a "wide mouth" and a heavy stand is best as it does not move.

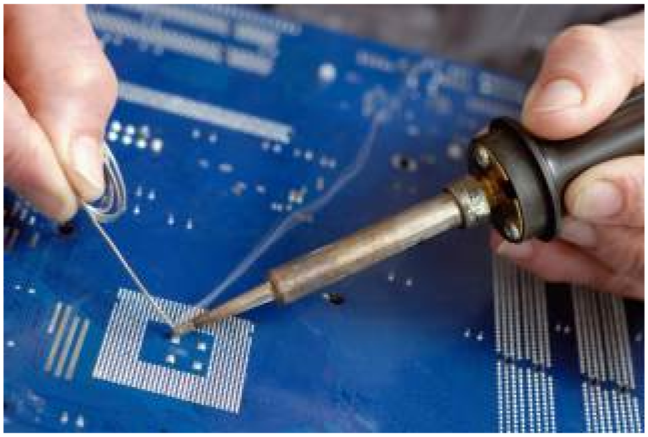

This photo clearly shows how to hold a soldering iron and solder.

This is NOT a temperature-controlled soldering iron and you can see it is too hot as it is burning off the flux too quickly.
Temperature Controlled Soldering Irons are now cheaper than the JUNK soldering iron shown in the photo. See eBay for prices.

to Index

BREADBOARD

The term BREADBOARD refers to any piece of wood or plastic containing pins or pegs or clips or holes where you can build a circuit.
The components can be soldered, twisted clipped or fitted into holes.
Breadboard also means the circuit can be easily pulled apart.
Some breadboards do not have two rows for the positive and negative rails. Connections under the board for the positive rail is shown with a black line in the photo. Connections on the main section of the board are shown with blue lines. Your breadboard MUST look exactly like the photo opposite.
Other breadboards are quite useless.
The breadboard in the photo can be purchased on eBay for less than $5.00 (post FREE).

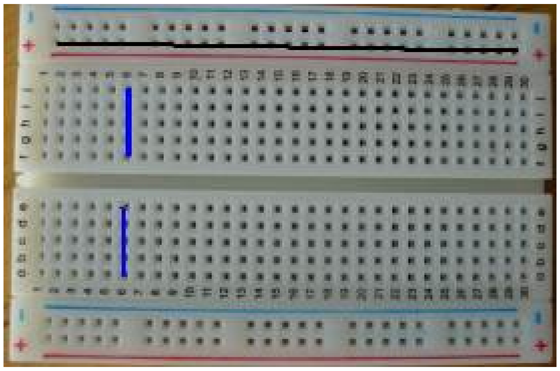

Photo shows a number of components fitted to the breadboard.

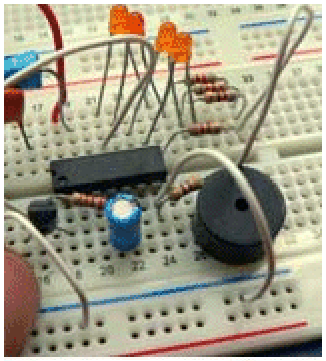

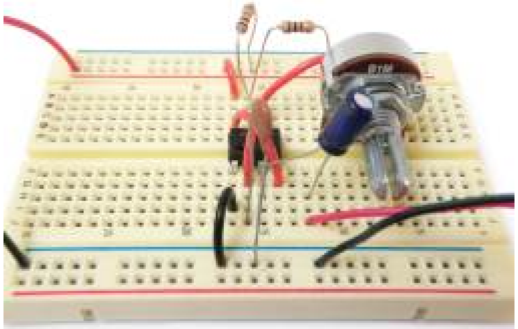

Fig 60. BREADBOARD and Components

to Index

Fig 61. Jumpers

The components on the BREADBOARD are fitted down the holes and metal strips under the board join each column of 5 holes. If you want to join one hole with another, you can use 0.5mm tinned copper wire or JUMPERS. See photo opposite. Jumpers can be purchased on eBay for less than $3.00 posted.

Email Colin Mitchell for links to eBay. (talking@tpg.com.au)

to Index

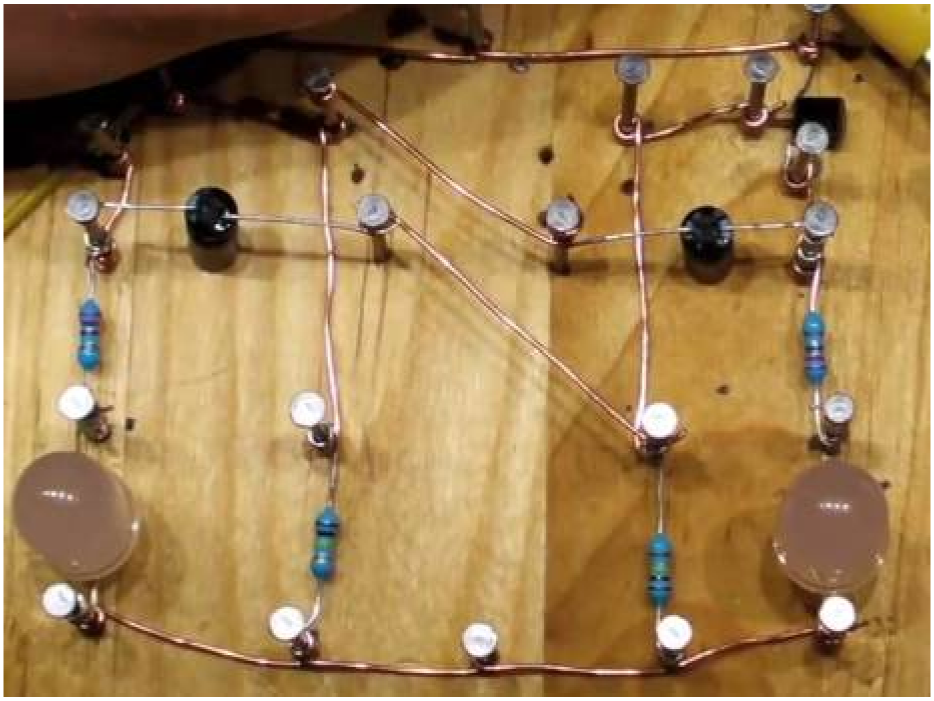

Fig 62. Breadboard with Nails

If you don't have a soldering iron or experimental breadboard, you can make your own board with nails. See the photo above. It is a multivibrator circuit and we will be presenting this circuit in a moment. The components can be twisted around the nails and bare wire used to join some of the nails to complete the circuit.

Another method of connecting the components is called BIRD-NESTING. This involves soldering the components "in the air" as shown in the 27MHz transmitter circuit below:

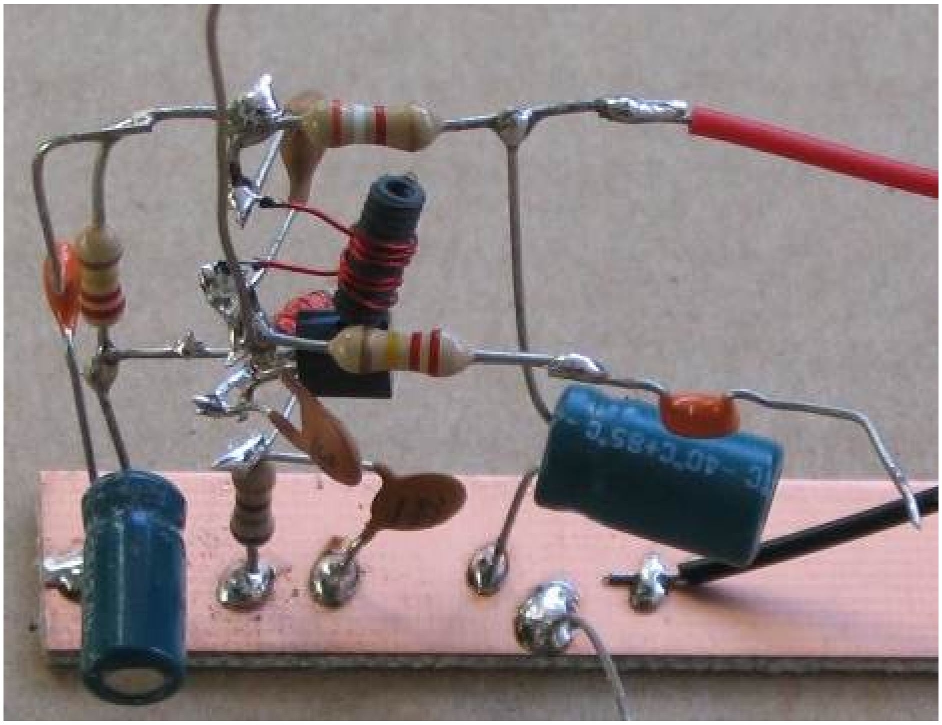

Another way to connect the component(if you don't have a soldering iron), is to wind 6 turns of bare wire around each connection and leaving all the components "in the air." The bare wire can be obtained from hook-up flex. This is plastic coated "wire" containing up to 15 fine strands of wire. Use a single strand for the connections. None of the components will touch each other BY MISTAKE and the circuit will work perfectly. Bird-nesting is a good way to build a quick circuit and test its performance. It might look messy but you can easily change any component.

to Index

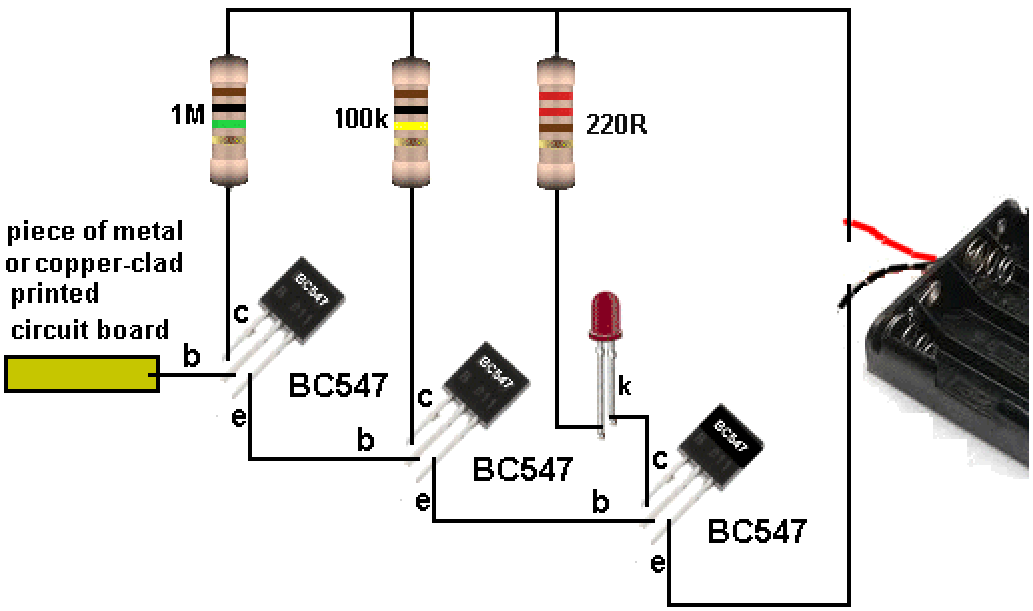
1M
100k
220R
piece of metal or copper-clad printed circuit board
c
b
e
BC547
c
b
e
BC547
k
c
b
e
BC547

Fig 63. Wiring the 3-Transistor Circuit using BC547 Transistors
The diagram shows how to connect 3 x BC 547 transistors.
The leads of a transistor can be collector-base-emitter OR emitter-base-collector and that's why we have provided 2 different wiring diagrams.

to Index

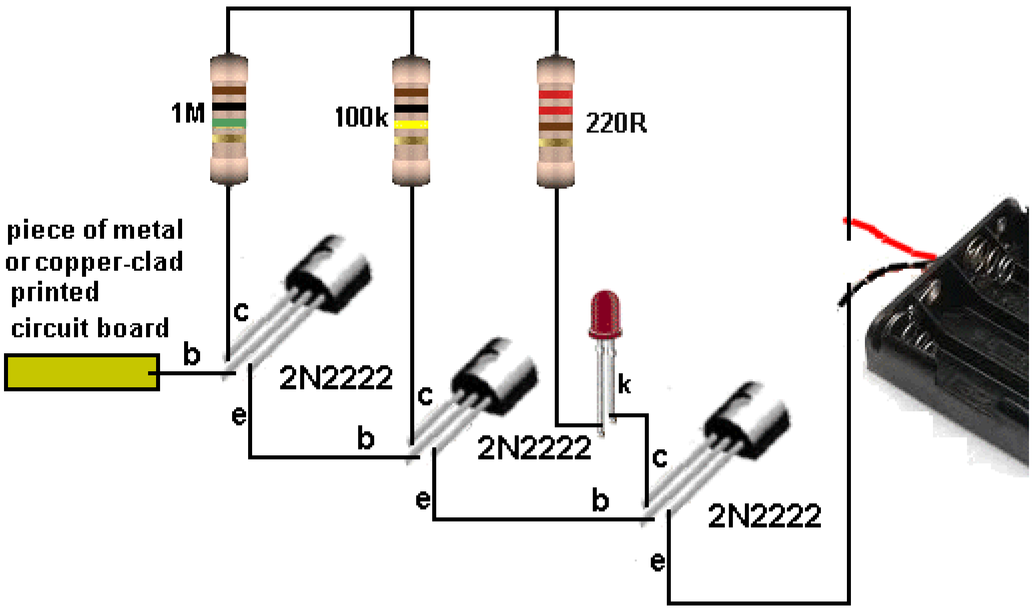
1M
100k
220R
piece of metal or copper-clad printed circuit board
c
b
e
2N2222
c
b
e
2N2222
k
c
b
e
2N2222

Fig 64. Wiring the 3-Transistor Circuit using 2N2222 Transistors
The diagram shows how to connect 3 x 2N2222 transistors.

to Index

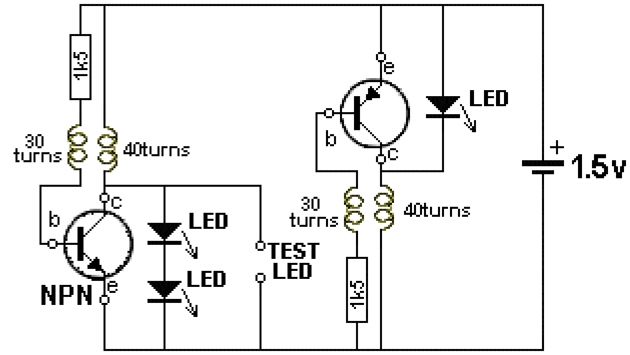
1k5
30 turns
40turns
b
c
e
NPN
LED
LED
TEST LED
e
b
c
LED
30 turns
40turns
1k5
+
1.5v

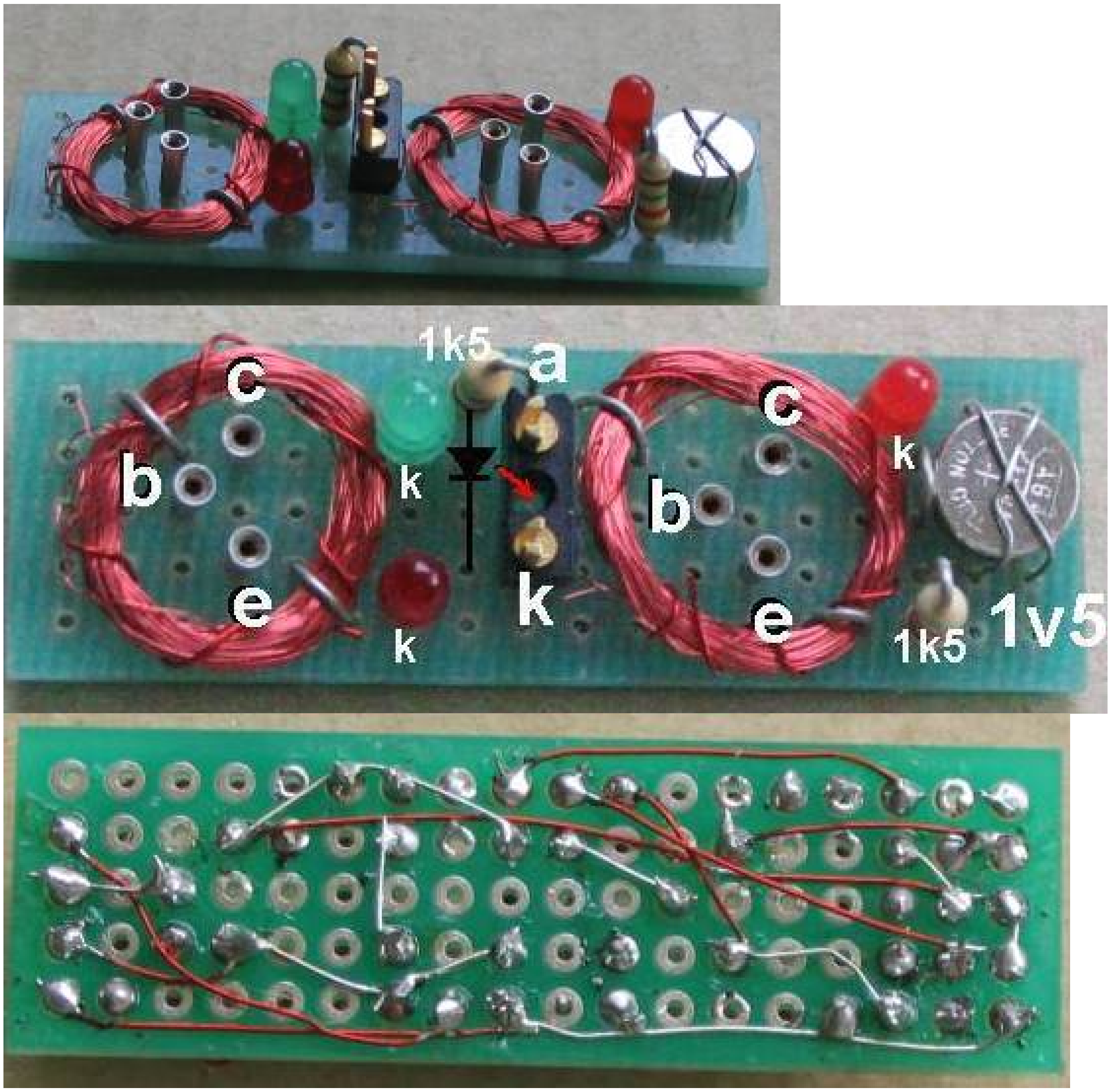
1k5
a
c
b
e
k
k
k
c
b
e
k
1k5
1v5

Fig 65. Transistor Tester
This handy transistor and LED tester can be built to test LEDs and both PNP and NPN transistors.
The project consists of two identical circuits, one for NPN and one for PNP.
You can build just the NPN section and then build the PNP section later.

LED VOLTAGE

We have shown a LED needs at least 1.7v supply for it to operate.
This circuit works on 1.5v and thus the action of the transistor and coil (called a Transformer) MUST be increasing the voltage for the LED to illuminate.
This circuit works on two "actions."
1. Transistor ACTION - this is the action of a transistor providing gain to make the circuit oscillate.
2. Transformer ACTION - this is the action of a coil of wire producing a voltage higher than the

supply voltage when it is turned off.

This circuit is very technical and very complex.
We will be explaining it in a very simple way because this is a **Basic Electronics Course**.

THE TRANSFORMER - the two coils of wire on the left and the two coils of wire on the right.
When the voltage (actually the current) is switched off, the 40 turn coil in either of the circuits in this project; the voltage across the coil rises to more than the 1.5v supply and is in the opposite direction to the voltage of the supply.
The circuit looks to be very simple but it uses an air-cored transformer to produce the voltage needed to illuminate the LED indicators and the circuit only works when the transistor is connected correctly. There are two separate circuits, one for NPN transistors and one for PNP transistors. We will cover the NPN section:
The circuit turns ON when the NPN transistor is fitted and the current through the 30 turn coil and 1k5 resistor turns ON the transistor and produces expanding flux in the 40 turn coil. This flux cuts the turns of the 30 turn coil and produces a voltage in the coil that adds to the supply voltage and increases the current into the base. This turns the NPN transistor ON more. This action continues until the transistor is fully turned ON. At this point the current in the 40 turn coil is a maximum but it is not expanding flux and the 30 turn coil ceases to see the extra voltage. Thus the current into the base reduces and this turns the transistor OFF slightly. The flux produced by the 40 turn coil now becomes collapsing (or reducing) flux and it produces a voltage in the opposite direction to greatly reduce the current into the base. In a very short period of time the transistor becomes TURNED OFF and it is effectively removed from the circuit. The flux in the 40 turn coil collapses quickly and it produces a voltage in the 40 turn coil that is higher than the supply voltage and is in the opposite direction. This means the voltage produced by the 40 turns ADDS to the supply voltage and is delivered to the LEDs to illuminate them.
The NPN circuit has two LEDs in series so that a LED of any colour (including white) can be connected to the TEST LED terminals and it will illuminate. You can use any colour LED for any of the LEDs, however it is best to use either green or yellow or white for the single LED.
The two "coils" are wound on a 10mm dia pen with 0.1mm wire (very fine wire). The loops of tinned copper wire holding the coils on the board are connected to separate lands under the board and MUST NOT produce a complete loop as this will create a “Shorted Turn” and the circuit WILL NOT WORK.
If the LEDs do not illuminate, simply reverse the wires to the 30 turn coil.
The circuit does not need an ON/OFF switch because the LEDs require a voltage of over 2v to illuminate (the orange LED) and the supply is only 1.5v. A red LED needs about 1.5v to 1.7v to operate but when it is in series with a green LED, this voltage is over 3.5v.
All the components fit on a small matrix board 5 holes x 18 holes. A kit of parts for the project is available for $4.00 plus $3.00 postage and ordering details can be obtained by emailing Colin Mitchell. (talking@tpg.com.au)
Build the circuit and test your transistors and LEDs.
We will be covering more on the action of a transistor and the action of a transformer in the discussion below, but it is important to build the circuit and see it working.
It is your first piece of **TEST EQUIPMENT**.

Questions

1. Identify the letters "c" "b" and "e"
2. What type of transistor is tested in the first set of hollow pins?
3. Put a PNP transistor into the first set of hollow pins and try all positions. Does the red and green LEDs illuminate?
4. When both the red and green LEDs illuminate, what is the approximate voltage across the pair?
5. When you fit a red LED to the test-socket, what is the approximate voltage across it?
6. When you fit a red LED to the test-socket, why does the red LED and green LED on the PC board turn off?
7. Why doesn't the project need an on/off switch?
8. The two coils for the circuit on the left is called a TRANSFORMER. Do the connections of the windings have to be connected to the circuit around a particular way?

to Index

ROBOT MAN

This multivibrator circuit will flash the Robot Man's eyes as shown

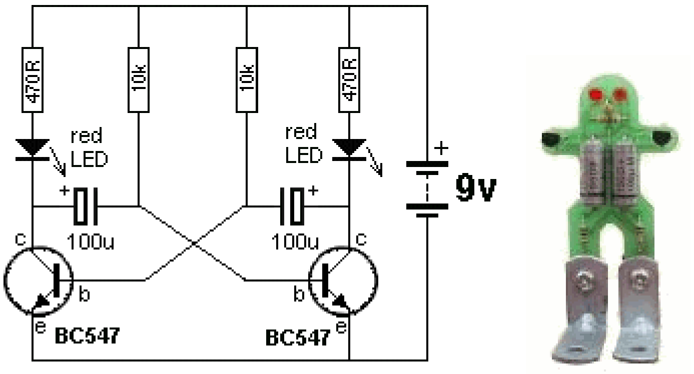

Fig 66. ROBOT MAN
The ASTABLE MULTIVIBRATOR or "free-running" multivibrator.

in the photo. The kit of components is available from Talking Electronics for $8.50 plus postage. Send an email to find out the cost of postage: talking@tpg.com.au
The photo shows the LEDs flashing.
The circuit is called an **ASTABLE MULTIVIBRATOR** and this means it is not stable but keeps switching from one transistor to the other.
It is also called a **FLIP FLOP** circuit.

to Index

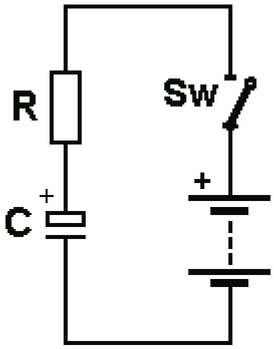

Fig 67. The TIME DELAY

The **TIME DELAY** circuit consists of a Resistor **R** and Capacitor **C** in **SERIES**.
When the switch is closed, the electrolytic (called the CAPACITOR) charges slowly because the resistor only allows a small amount of current to flow.
It's just like charging your mobile phone. The battery takes time to charge because there is a resistor in the circuit to limit the current. If we remove the resistor in the mobile phone, the battery will get too hot when it is being charged but in the **TIME DELAY** circuit, we want the capacitor to charge slowly, because we want a **TIME DELAY**. .

to Index

Fig 68. The charging of a capacitor is the same as building a brick wall.

CHARGING A CAPACITOR

The capacitor in **Fig 67** charges via the resistor **R**.
But the voltage on the capacitor does not rise at a constant rate.
It starts off charging very quickly and as the voltage across it get higher, the voltage increases at a slower and slower rate.
In the photo I am building a brick wall.
I am working at a constant rate.
When I started building the brick wall, I laid 5 rows of bricks (5 courses) in the first hour.
As the wall increased in height, I had to climb the ladder and I could only lay 3 courses an hour and finally the wall was so high I could only lay 1 course per hour.
This is exactly the same as a capacitor charging.
When the capacitor is uncharged, the supply voltage allows a high current to pass through the resistor **R** and the energy quickly fills the capacitor. This results in a rapidly increasing voltage on the capacitor. But as the voltage on the capacitor increases, the difference in voltage between that on the capacitor and the supply is very small and only a small current will pass through the resistor. This means the voltage on the capacitor increases at a slower rate.

to Index

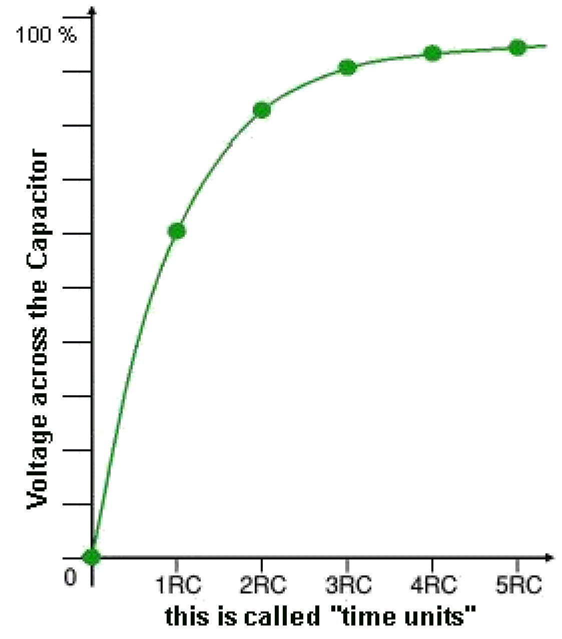

Fig 69. Time Units

It really does not matter how fast or slow or uneven a capacitor charges because most circuits detect a voltage on a capacitor and the time taken to reach this voltage is called the **TIME DELAY**.
But to prevent you thinking the capacitor charges "smoothly" we have to explain what actually happens.
The graph on the left shows the capacitor charging. You can see it charges quickly at the beginning and then charges slowly and then very slowly.
You can see the first part of the graph is fairly "straight" (constant charging) - NOT "straight up and down" but a straight line - and this applies to a voltage of about 63%.
The time taken to reach this voltage is called **ONE TIME UNIT** - also called **ONE TIME CONSTANT**. The graph continues for another 4 "time units" (time-constants) and the final voltage is very nearly 100%. (It never reaches 100%.)

to Index

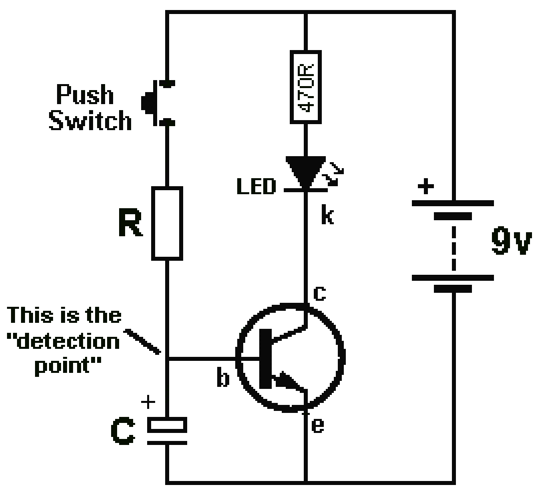

Fig 70. Time Delay Circuit

A **TIME DELAY** circuit needs three things:
1. A Resistor (R)
2. A Capacitor (C)
3. A "Detection Point."

When the switch is pressed, capacitor (**C**) takes time to charge via resistor (**R**) and after a short period of time the voltage at the **DETECTION POINT** is 0.6v and the transistor is TURNED ON. The LED illuminates.

Build the circuit with 100u and 100k and see how long it takes before the LED illuminates.

to Index

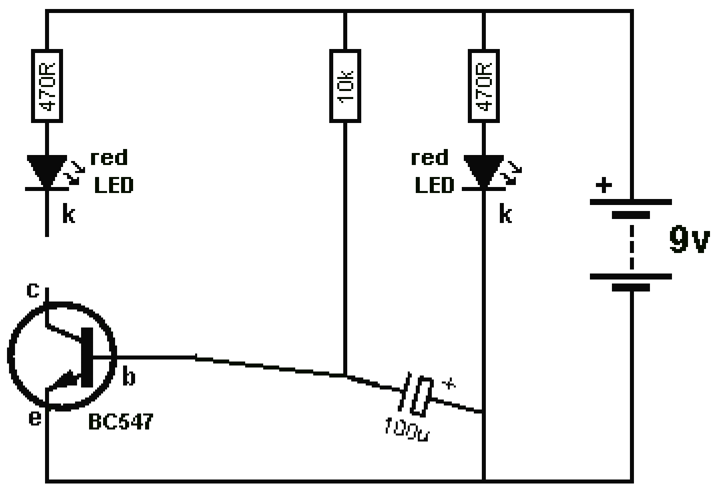

In the **ROBOT MAN** project you can see the "**TIME DELAY**" circuit made up of the 100u, 10k resistor and the base of the transistor.

This is one of the most important **BUILDING BLOCKS** in electronics.
It is the basis of all oscillators and will be discussed below, after we explain a few more details.

Fig 71. The "TIME DELAY" in the ROBOT MAN Project

to Index

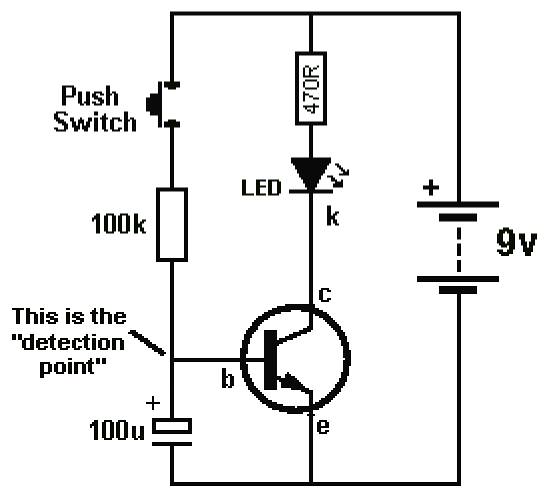
Push Switch
470R
LED
k
100k
+
9v
c
This is the "detection point"
b
+
100u
e

Fig 72. Turning A Transistor ON

We will use the **TIME DELAY** circuit to turn the transistor **ON**.
Make sure the 100u is uncharged by touching both leads (both ends of the capacitor) at the same time with a JUMPER - this is a piece of wire shown in Fig 61.
Push the switch and noting happens. After a short period of time the LED starts to glow and then comes on fully.
This shows two things:
1. The transistor is not turned ON when the base voltage is zero.
2. The base voltage must be 0.6v for the transistor to start to turn ON and when the voltage is 0.65v the transistor is turned **ON FULLY.**

to Index

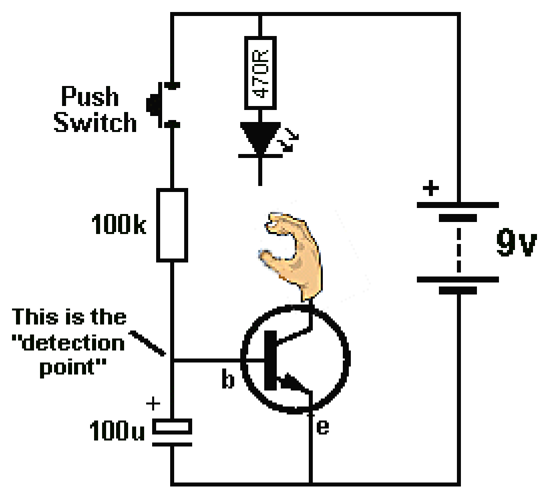
Push Switch
470R
100k
+
9v
This is the "detection point"
b
+
100u
e

Fig 73. How the LED turns ON

Here is an explanation of how the LED turns ON.
When the circuit is first assembled and the switch is not pressed, the transistor is not turned on and it is just like the diagram opposite. The LED is not connected to the transistor.

to Index

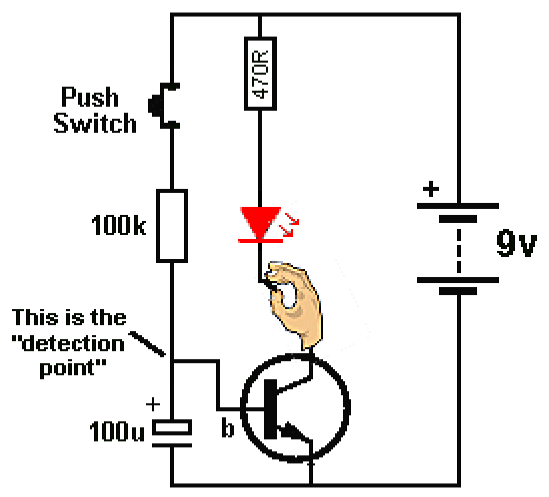
Push Switch
470R
100k
+
9v
This is the "detection point"
+
100u
b

Fig 74. The LED turned ON

When the switch is pushed, the transistor turns **ON** (after a few seconds) and it pulls the lower lead of the LED down towards the 0v rail and this action turns the LED **ON**.

to Index

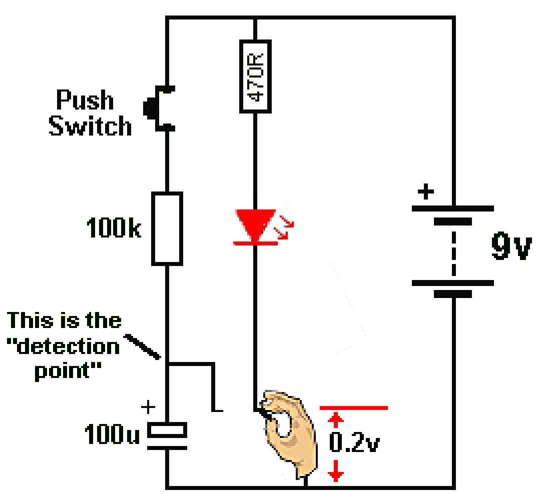

Fig 75. The LED fully turned ON

When the LED is **fully turned ON**, the lower lead of the LED is almost directly connected to the 0v rail.
In other words:
When the transistor is **FULLY TURNED ON**, the lower lead of the LED is almost directly connected to the 0v rail.
The voltage between the lead of the LED and 0v rail is 0.2v. This is the characteristic voltage across the collector-emitter terminals of a transistor when it is TURNED ON.

to Index

In the three diagrams above you can see the LED is changed from an **OFF** condition to an **ON** condition by the action of the transistor.
The transistor is acting **LIKE A SWITCH.**
This action is one of the most important actions in electronics.
It is called: **"The Transistor as a SWITCH"**
It is the basis to ALL Digital Circuits.
It is the basis because of these two facts:
1. When the transistor is **OFF**, the circuit is taking **no current** and no power is being lost or wasted.
2. When the transistor is **ON**, the LED is almost at 0v and no resistor is in the lower lead to waste any power.
Thus we can turn things **ON** and **OFF** without wasting and power.
This is the basis to **DIGITAL ELECTRONICS**.

to Index

DIGITAL ELECTRONICS revolves around circuits that are either **FULLY ON** or **FULLY OFF**. This means they take almost no power and we can combines lots of circuits and still take almost no power.
This means they do not get hot and it also means they will last a long time.
You may not think turning a transistor **ON** and **OFF** will achieve any worthwhile outcome but a circuit can be designed to use two transistors (similar to the **ROBOT MAN** above). The circuit does not Flip-Flop but requires a switch and when the switch is pressed, the circuit changes state. The two transistors are connected together and it takes two presses of the switch to make the output of the second transistor change state ONCE.
The circuit is a divider. It is called a: **divide-by-two** and is the basis of all counting in a computer.
By adding more "**divide-by-two"** circuits we can get "**divide by 4**, **divide by 8**" etc. Two transistors don't do much but when you combine millions of transistors we have a **COMPUTER.**

to Index

Two transistors can do one more thing. They can "**REMEMBER**."
Here is a manual circuit.
Pressing Switch **A** turns the LED **ON** and pressing switch **B** turns the LED **OFF**.
The circuit "remembers" or remains in each state called a stable state.

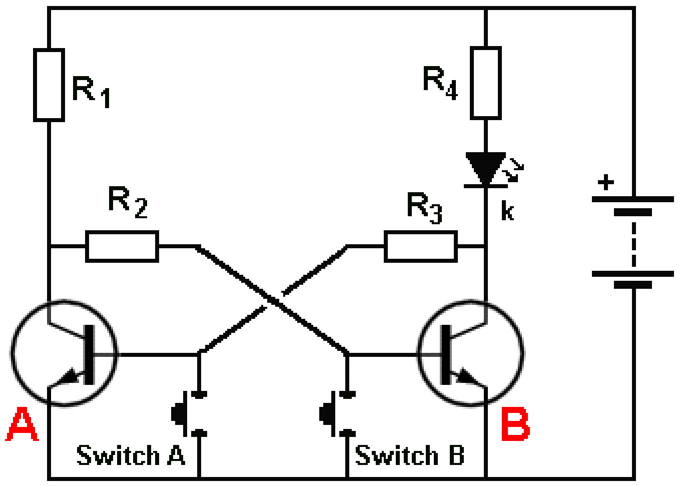

Fig 76. The "MEMORY CELL"

When Switch **A** is pressed, the voltage on the base is removed and transistor **A** turns OFF.

Transistor **B** turns ON via resistors R_1 and R_2 and the LED is turned **ON**.

When the switch is released, the voltage on the collector of transistor **B** is less than 0.6v and the two transistors remain in this state.

Pressing switch **B** turns the LED **OFF.** (transistor **A** turns ON via R_3, R_4 and the LED - very little current flows through the LED and you can hardly see it glowing). The voltage on the collector of transistor **A** is less than 0.6v and the two transistors remain in this state.

The technical name for this circuit is:
BISTABLE MULTIVIBRATOR or **BISTABLE SWITCH** or **BISTABLE LATCH**.

This is the basis to all the memory in a computer.

to Index

In electronics, we talk about the **DIGITAL TRANSISTOR** and **ANALOGUE TRANSISTOR**.
This is just an ordinary transistor (called a Bipolar Junction Transistor) in a **DIGITAL CIRCUIT** or **ANALOGUE CIRCUIT**.
We are now discussing the **DIGITAL CIRCUIT** - The Multivibrator - Astable Multivibrator and Bistable Multivibrator (Memory Circuit).
The **DIGITAL CIRCUIT** has **2 STATES**.
The **ON STATE** and the **OFF STATE**.
It is conducting in the ON STATE and the LED is illuminated.
In the OFF STATE, the LED is not illuminated.
In the ON STATE the transistor is said to be **CONDUCTING** or **BOTTOMED**.
In the OFF STATE the transistor is said to be "**CUT OFF** or "**OFF**."
These two states are reliable and guaranteed. They are not "half on" or "quarter on" or "75% off."
These states are easy to transmit "down a wire." The ON STATE is transmitted as "1" (voltage present) and the OFF STATE is transmitted as "0" (voltage not present).
These are the two **DIGITAL STATES**.

The **ROBOT MAN** is a **DIGITAL CIRCUIT**.
Each LED is ON or OFF.
The waveform on the output of each transistor is called a **DIGITAL SIGNAL**.

The waveform is said to be **DIGITAL** or **SQUARE WAVE**.
The top line of the graph represents the LED OFF.
The bottom line of the graph represents the LED ON. The LED is ON when the collector voltage is LOW because we are pulling the lead of the LED to the 0v rail as shown above.
The circuit changes from one state to the other very quickly and this is called the **RISE TIME**.

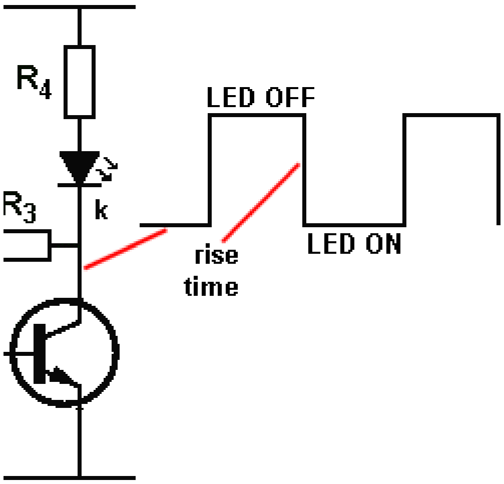
R4
LED OFF
k
R3
rise
time
LED ON

to Index

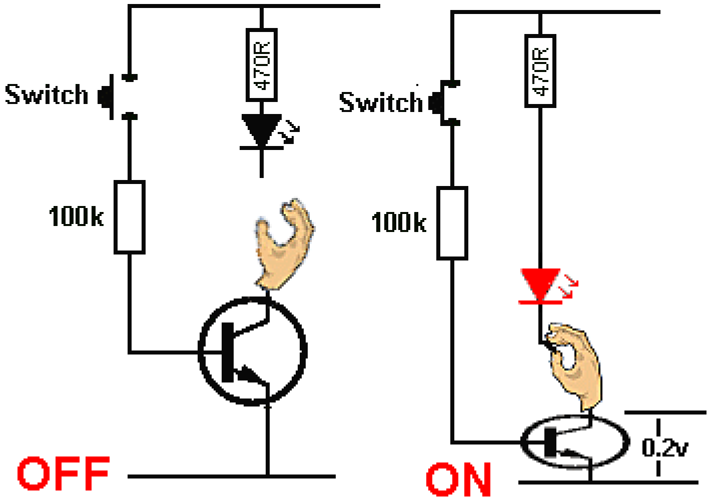
Switch
470R
100k
Switch
470R
100k
0.2v
OFF
ON

Fig 77. The two Digital States

Going over the two **DIGITAL STATES** for a transistor.

In the first diagram the switch is not pressed and the base does not see a voltage to turn the transistor on. The transistor is "**OFF**" (not conducting) and it is not "grabbing" the LED. The LED is not illuminated.
In the second diagram the base of the transistor sees a voltage via the switch and it is TURNED on. The LED is illuminated.

to Index

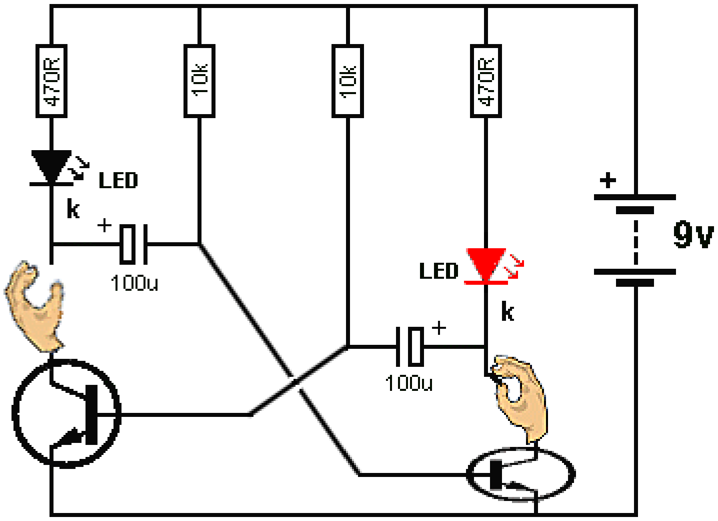
470R
10k
10k
470R
LED
k
+
100u
LED
k
+
100u
+
9v

Fig 78. ROBOT MAN Animation
This animation shows how a transistor grabs a LED and pulls it towards the 0v rail to turn it ON.

to Index

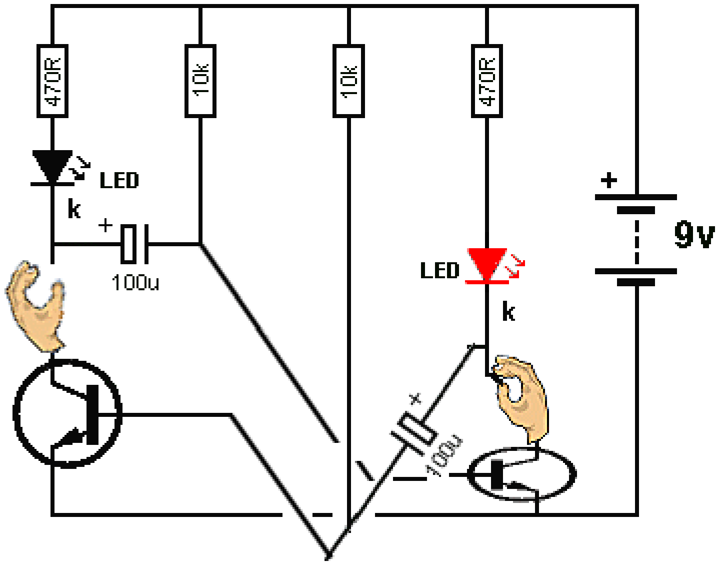

Fig 79. TIME DELAY Animation

The animation in **Fig 78** shows the two transistors turning the LEDs **ON** and **OFF** in a **FLIP FLOP** circuit.
We know the 10k and 100u components form a **TIME DELAY** to create the time for each LED to be illuminated. The timing for one LED plus the other LED creates a **CYCLE** and this is the **FREQUENCY OF OPERATION** for the circuit. It is measured in cycles per second - Hertz - Hz.
We will now go into more detail of how the **TIMING COMPONENTS** create the **TIME DELAY** for each LED.
The circuit is more-complex than you think.
The 100u is already charged from a previous cycle and we show how it gets discharged via the 10k and charged in the opposite direction by the 10k to create a **TIME DELAY**.

to Index

THE CAPACITOR

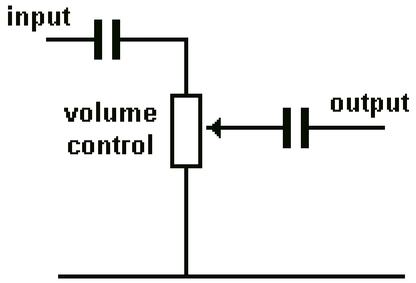

The capacitor can perform many different functions and produce many different effects, depending on its value and the surrounding components.
In this circuit the capacitors on the input and output prevent DC on the volume control creating "scratchy sounds" when the volume is altered.
This is called "DC blocking."
The AC (the signal) passes through the capacitors but the DC voltage on the input is blocked.

to Index

CHARGING A CAPACITOR Part II

It is easy to see how a capacitor charges via a resistor in the **TIMING CIRCUIT** (Delay Circuit) above but many capacitors are not connected to the 0v rail.
They are connected as show in the animation below and their "job" is to pass a waveform.
When they pass the waveform they **CHARGE** and **DISCHARGE**.
The waveform is called an **AC SIGNAL** and the output is smaller than the input.
The circuit is taken from the circuit above, but the same effect applies to all capacitors that "pass a signal."
Here's why:

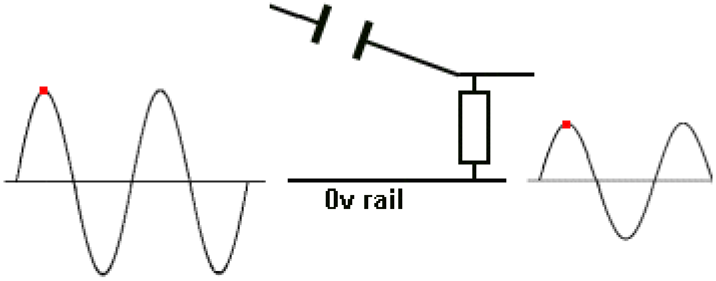

The capacitor charges slightly during the rise of the signal and the right-plate of the capacitor does not rise as high as the left-plate. That's why the output signal is not as large as the input signal.
If the capacitor did not charge, the output would be as large as the input. If you use a capacitor with a large value, it will not charge and thus the output will be as large as the input.
That's why you use a large capacitor !!!!

to Index

CHARGING A CAPACITOR Part III

Here is another CAPACITOR in action.

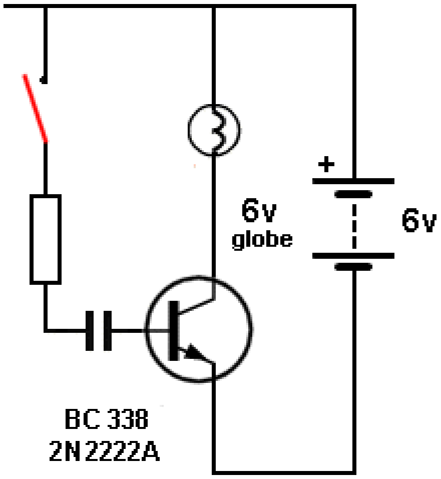

The animation shows a capacitor charging (via a resistor). The initial current is LARGE and this turns the transistor FULLY ON and the globe illuminates. As the capacitor charges, the base current reduces and the transistor starts to turn OFF. Eventually the capacitor is fully charged and the voltage on the base falls to 0v, turning the transistor OFF.

This animation shows three features:
1. The initial charging current is HIGH.
2. It gradually falls to zero.
3.The voltage on the base drops below 0.6v and the transistor turns OFF.

to Index

NEGATIVE VOLTAGE

You will be surprised to learn that many circuits produce a negative voltage or negative spike at some point (when doing circuit-analysis, each location or point or join of components is called a NODE) on the circuit. In other words the voltage will be LESS than the 0v rail of the circuit.
This is due to the presence of a capacitor and the animation shows how a capacitor can produce a negative voltage:

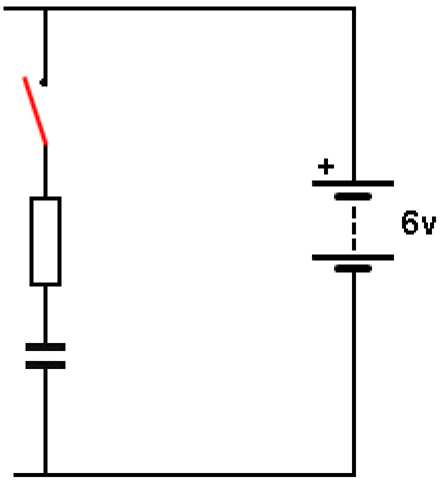

When a charged capacitor is "lowered from one position in a circuit" the positive lead may be lowered by say 3v. This means the other lead will be lowered by 3v. We are assuming the capacitor can be lowered and is not directly connected to the 0v rail.

You can see the electrolytic produces a NEGATIVE VOLTAGE on the base in the following animation, when the two transistors change states:

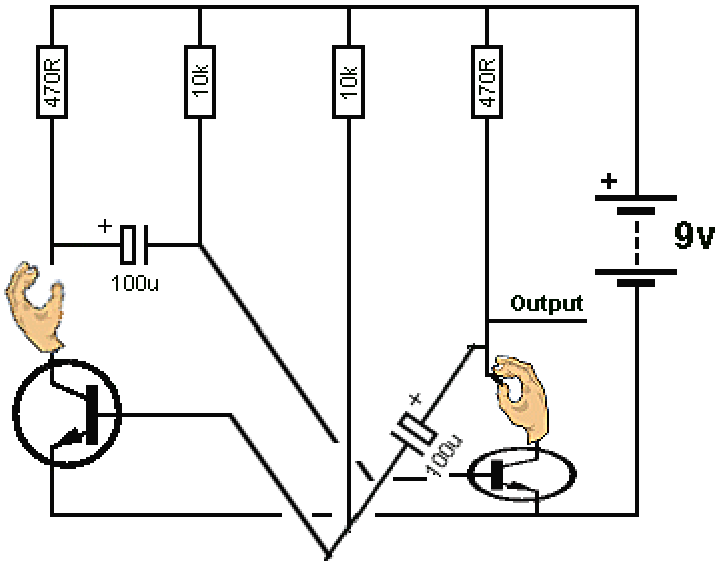

to Index

The Electret Microphone

The most common type of microphone is the ELECTRET MICROPHONE.
It is incorrectly termed the "Capacitor Microphone" or Condenser Microphone." "Capacitor Microphone" descriptions make no mention of a FET as the amplifying device and a polarized diaphragm to detect the audio, so they are something different.
The electret microphone consists of a FET (transistor)

inside an aluminium case with a very thin Mylar film at the front. This is charged and when it moves (due to the audio it receives via a small hole in the front of the case), it vibrates and sends a very small voltage to the GATE lead of the Field Effect Transistor. This transistor amplifies the signal and produces a waveform of about 2mV to 20mV at the output.

The electret microphone requires about 0.5mA and will operate from 1.5v supply with 4k7 LOAD RESISTOR.

For 3v supply, the Load Resistor can be 22k to 47k.

For higher supply voltages the resistor will be 68k or higher.

Electret microphones are extremely sensitive and will detect a pin-drop at 3 metres.

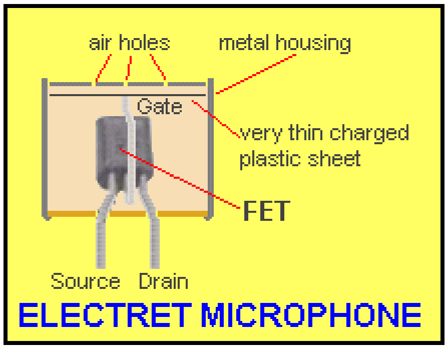
air holes
metal housing
Gate
very thin charged plastic sheet
FET
Source
Drain
ELECTRET MICROPHONE

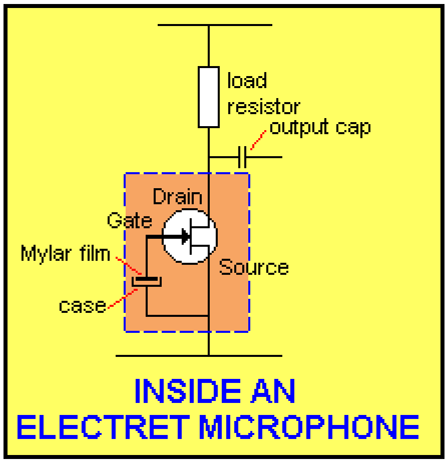
load resistor
output cap
Drain
Gate
Mylar film
Source
case
INSIDE AN ELECTRET MICROPHONE

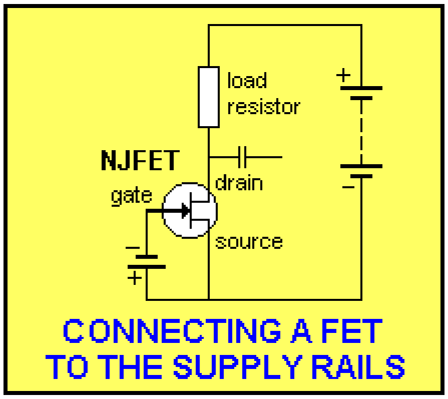
load resistor
+
NJFET
drain
gate
source
–
–
+
CONNECTING A FET TO THE SUPPLY RAILS

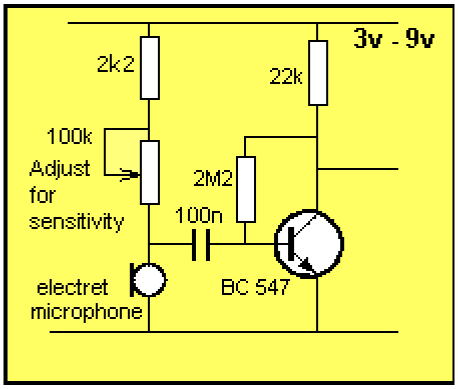
3v - 9v
2k2
22k
100k
Adjust for sensitivity
2M2
100n
electret microphone
BC 547

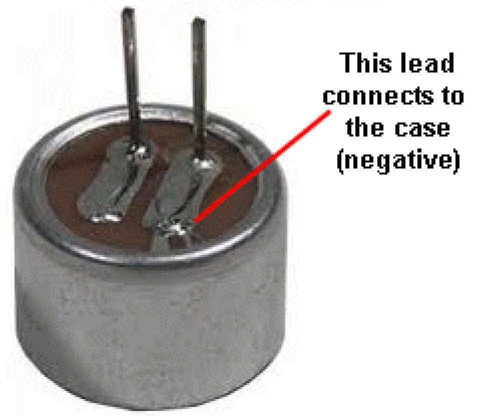
This lead connects to the case (negative)

Most electret microphones have two leads. One lead is connected to the case and this lead goes to the 0v rail. The other lead goes to a LOAD RESISTOR (4k7 to 68k - depending on the voltage of the project).

Reducing the value of the load resistor will increase the sensitivity until the background noise is very noticeable.

They are used in Hearing Aids and are more-sensitive than the human ear.

They are very small, low-cost and very sensitive.

to Index

The Speaker

The most common speaker is about 30mm to 60mm diameter and 8 ohm impedance. This means the voice coil is about 8 ohms resistance.

The two leads can be connected either way to a circuit.

The speaker shown is 32mm diameter and has a realistic wattage of 100mW (NOT 1watt).

These speakers have a Mylar cone and the magnet is a "super magnet" and very small. That's why it is so flat.

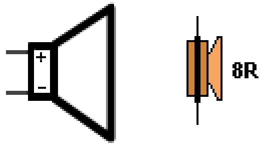

Speaker Symbol

A speaker can be used as a microphone (called a Dynamic Microphone) and a circuit to connect the speaker (mic) to an amplifier can be found on Talking Electronics website. It is not as sensitive as an electret microphone and does not produce the same output amplitude, but it is an emergency microphone.

to Index

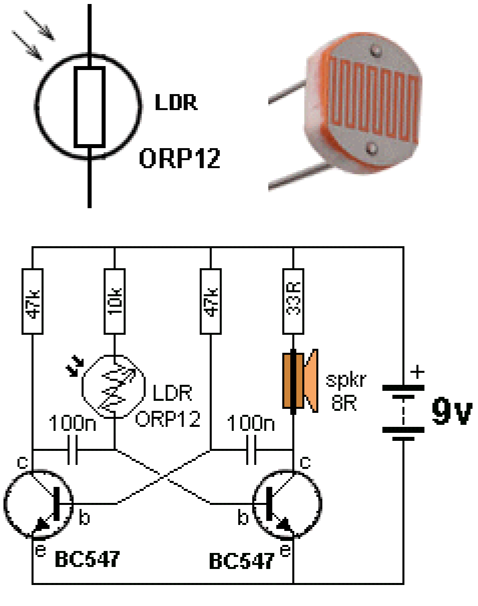

LDR and Light Alarm Circuit

Light Dependent Resistor

(LDR)

Also called **PHOTOCELL** or **PHOTO RESISTOR**
A **Light Dependent Resistor** is a 2-leaded component containing a layer of semiconductor material.
The top contains two interleaving combs of conducting wires with a path of semiconductor material between. When light falls on the component, the resistance of the semiconductor material decreases.
In darkness the LDR will be about 300k. In very bright light the resistance will be about 200 ohms.
But if the light changes only a very small amount, the resistance CHANGE is VERY SMALL. For a large change, see **Photo Transistor**.
The **Light Alarm** circuit will produce a squeal when light falls on the LDR.

to Index

Photo Transistor

The **Photo Transistor** is very sensitive to changes in illumination. It is about 100 times more sensitive than the **LDR**.
The **Photo Transistor** is also available as a **DARLINGTON.**
The Darlington Photo-transistor is 100 x 100 times (10,000) more sensitive than the LDR.

The **Photo Transistor** and **Photo Darlington Transistor** are connected just like a normal transistor but the base lead is not connected. If the value of the LOAD RESISTOR is large, the

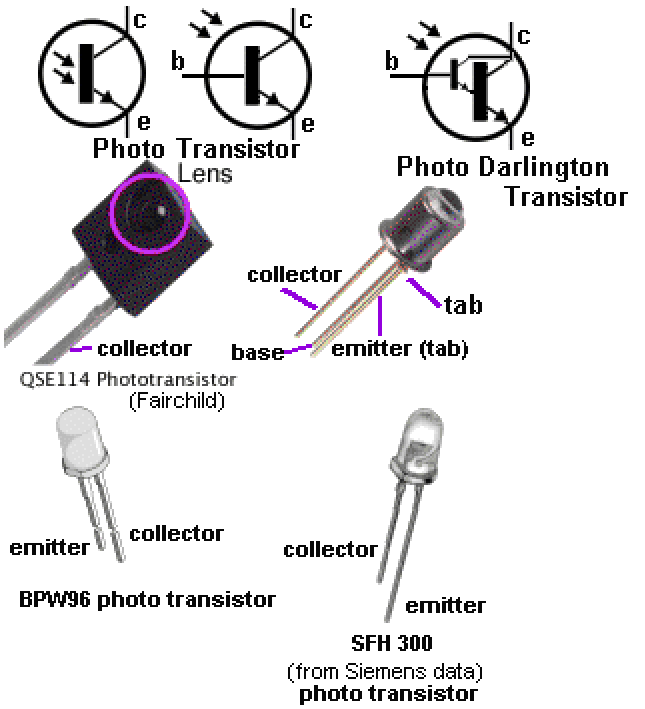
c
b
e
c
e
Photo Transistor
b
c
e
Photo Darlington
Transistor
Lens
collector
tab
collector
base
emitter (tab)
QSE114 Phototransistor
(Fairchild)
emitter
collector
collector
BPW96 photo transistor
emitter
SFH 300
(from Siemens data)
photo transistor

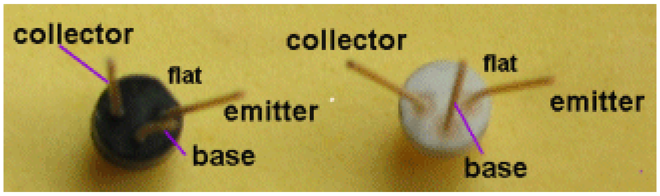
collector
flat
emitter
base
collector
flat
emitter
base

FPT100 Photo Darlington Transistor

The FPT-100 Darlington Photo Transistor is available from Talking Electronics for $1.00 each plus postage.

transistor will not be very sensitive.

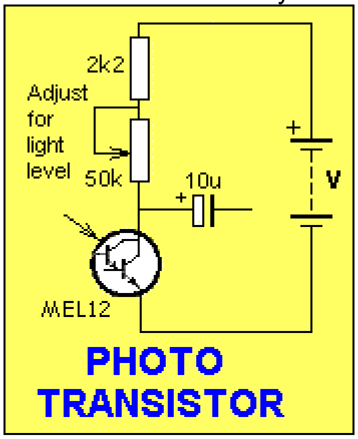
2k2
Adjust
for
light
level
50k
10u
+
+
V
MEL12
PHOTO
TRANSISTOR

to Index

The Inductor

Also called **"coil,"** or **"Choke."**
An **Inductor** consists of one to many turns of wire wrapped around a former (tube of cardboard). The wire can be jumble-wound or wound in layers. The result is the same. This is called an **air-cored coil** or **air-cored inductor**.
The centre can be filled with a metal such as iron or laminations (thin sheets of metal) or a ferrite material. Different cores operate at higher frequencies.
The core can be circular (doughnut) or rectangular and it is called a **MAGNETIC CIRCUIT** (when it is a closed loop).
Additional turns or increasing the diameter of the turns will increase the inductance.
A coil with a magnetic core can be used to pick up nails and metal items. It is called an **electromagnet**. It can be

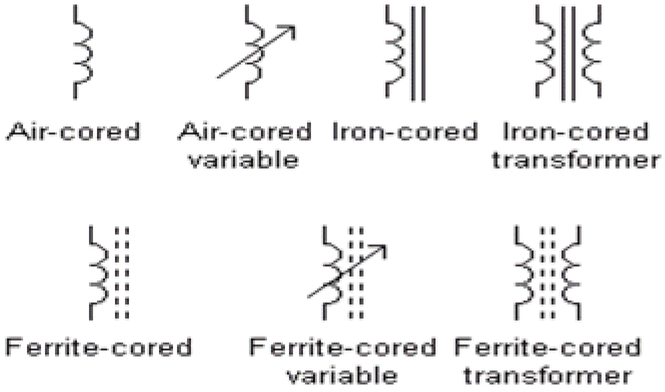

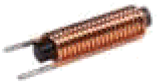

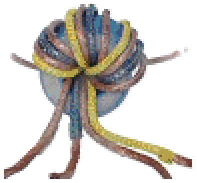

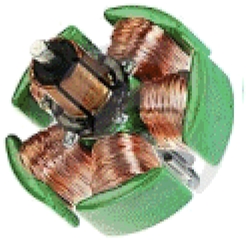

surface-mount inductor 27uH
yes 27uH
270uH = 271

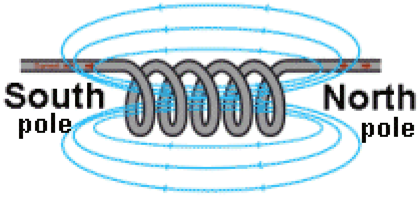

The emerging **magnetic lines of force** from an inductor produce the NORTH POLE - this is just "convention," a simple way to explain things - to get the explanation started.
When two or more coils are wound near each other, the inductor is called a **TRANSFORMER** (but not the armature above).

operated on AC or DC.
When the metal core is loose and gets pulled into the coil it is called a **SOLENOID** or **ACTUATOR** or **LINEAR ACTUATOR**. It can be operated on AC or DC.
The way an inductor works is very complex but we can say it resists any rise or fall in voltage by turning the rise or fall into magnetic flux.
If the applied voltage is suddenly turned off, the inductor produces **a very high voltage** of **opposite polarity** (these are the **two** most important things for you to remember).

to Index

The Antenna

The first time you will need an antenna is when making an FM transmitter.
The antenna is usually a length of wire equal to half the wavelength of the transmitter. The frequency is about 100MHz and the wavelength is 3 metres.
A half-wave antenna is 1.5metres.
The length is important but the height is more important. The wire should be as high as possible and "up-and-down" if the antenna on the radio is vertical, to get the maximum range.
This is called an **end-fed half-wave antenna** or **half-wave**

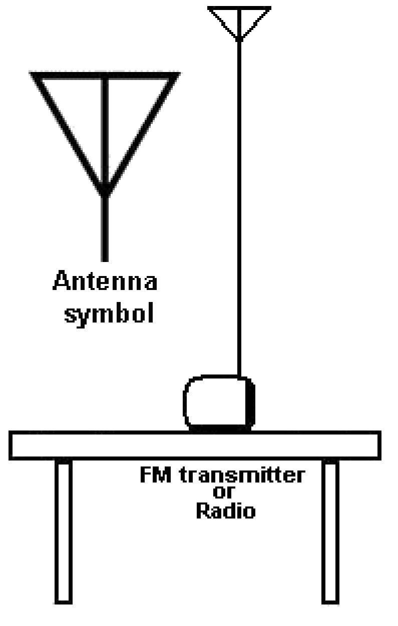

Monopole.
The transmitting circuit should have a good ground-plane such as connection to large batteries so the signal can be pushed and pulled into and out of the antenna.
This signal is then radiated as electromagnetic radiation to the surroundings.

to Index

The RELAY

The word **Relay** comes from the days of Morse Code where a coil of wire (an electromagnet) closed a switch to allow the Morse code to travel further down the telegraph line.
It would "relay" or "pass-on" the information.
A relay allows a "weak circuit" (one with low current) to operate a LOAD that needs a large current. It also separates the two circuits electrically and prevents a voltage such as 240v connecting to a 12v circuit. The coil is separated from the contacts and this gives the two circuits isolation.
A double-pole double-throw relay can be used to reverse a motor as shown here:

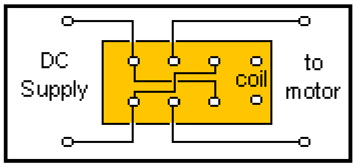

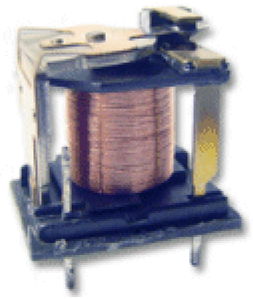

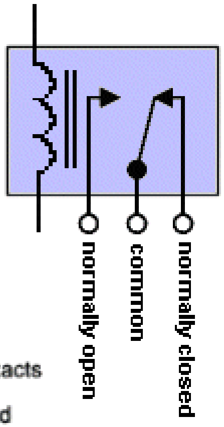

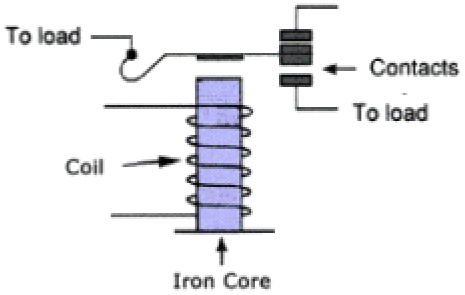

A single contact consists of 2 pins -called SPST (single-pole single-throw).
Or a single pair of contacts can consist of 3 pins - called change-over or SPDT (single-pole double-throw).
A double set of contacts consists of 6 pins, called DPDT (double-pole, double-throw). This is also called a **CHANGE-OVER RELAY** or **REVERSING RELAY** (when connected to a motor).

to Index

Driving A Relay

(Powering A Relay)
The first thing you must decide is the voltage of the relay. This will depend on the voltage(s) available. The relay will be driven (activated) by a transistor and the base of the transistor only needs a signal (less than about 1v). This means the project can be operated on a voltage from 3v to 12v and the relay can be connected to a 5v to 12v supply.

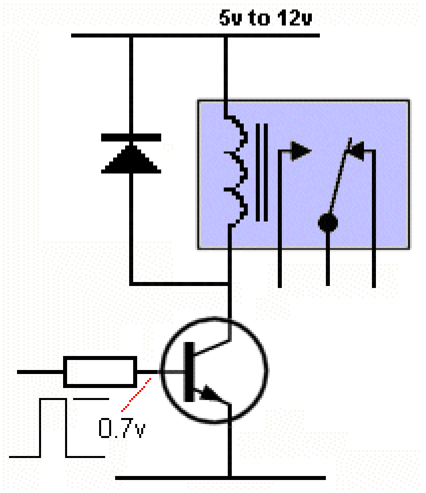

Driving A Relay

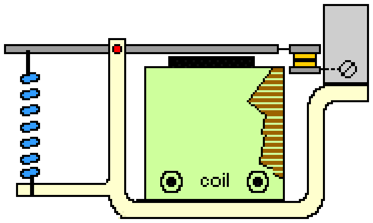

The armature is drawn towards the coil when a current flows through the coil.

Next you need to know the current-rating of the contacts. This will depend on the current taken by the LOAD. The rating of most relays is: 1 amp, 5 amp or 10 amp.
Finally you need to know how many contacts are required.
For a single circuit you will need 2 pins and for two circuits you will need 4 pins (but relays only come with 6 pins).
You can get relays that need a very small current for activation. These are called CMOS relays. But most relays need about 100mA.
To protect the driving-transistor from spikes when the relay is turned off, you will need a diode across the coil.
The top animation shows a "single set of change-over contacts."
The lower animation shows the ARMATURE being drawn to the electromagnet. The electromagnet is the coil with a core of magnetic material that becomes a magnet (an electromagnet) when a current flows through the coil.

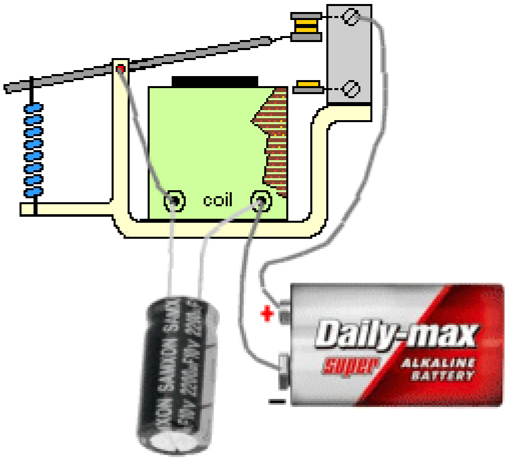

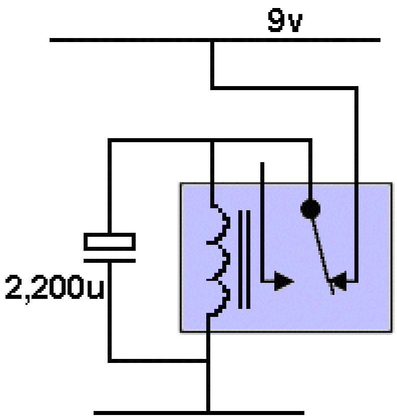

When the circuit is turned ON, the voltage across the 2,200u electrolytic is zero and it gradually charges. When the voltage is about 8v, the coil has enough voltage across it to pull the armature and open the contacts. The electrolytic supplies voltage to the coil for about 1 second and then the electromagnet does not have sufficient magnetism to hold the armature and it returns to close the contacts.

to Index

GATES

Animations of **GATES** and more details of their operation is covered in **DIGITAL ELECTRONICS** chapter.

to Index

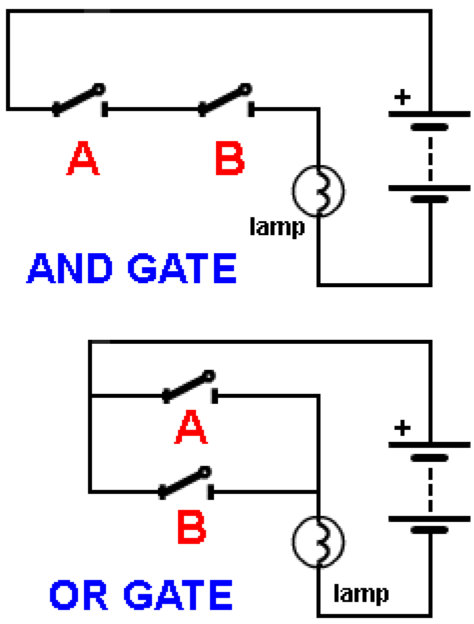

Fig 80. The "AND" GATE and "OR" GATE with switches

The next **BUILDING BLOCK** we will cover is called the **GATE**.
In its simplest form it is an electrical circuit consisting of switches.
Its just two or more switches connected in series or parallel.
We give each circuit a name so we can talk about it and explain its action with a single word.
Later we will cover the electronic version and show how diodes and a transistor are needed to perform a **GATING FUNCTION.**
The type of GATE we are talking about is a **LOGIC GATE**.
The circuit performs an operation called a **LOGICAL OPERATION** on an input or a number of inputs and creates a single output - called a **LOGICAL OUTPUT**.
LOGICAL means "understandable" or "correct" and in this case it means DIGITAL - the signal will rise to full rail voltage or fall to zero voltage. The output will not be half rail or quarter-rail voltage.
The diagram show an **"AND" GATE** and **"OR" GATE** with switches.
For the **AND GATE** close switch A **AND** switch B for the lamp to illuminate.
For the **OR GATE** close switch A **OR** switch B for the lamp to illuminate.

to Index

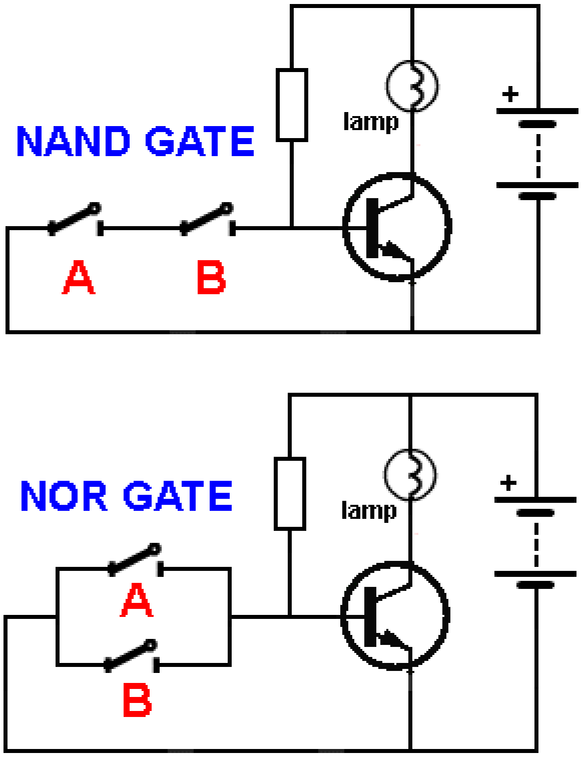

Fig 81. The "NAND" GATE and "NOR" GATE with switches and a transistor

INVERSION

Inversion produces the opposite effect to the results above.
Suppose we want to turn OFF a lamp when one or two switches are pressed.
We need a transistor.
The technical word for Inversion is **NOT**.
It is simplified to the letter "**N**."

For the **NAND GATE** close switch A **PLUS** switch B for the lamp to turn OFF.
For the **NOR GATE** close switch A **OR** switch B for the lamp to turn OFF.

These gates are only **demonstration-gates** to show how one or two switches will turn a lamp OFF.

to Index

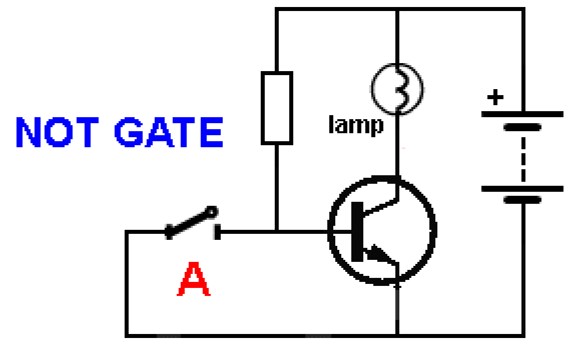

Fig 82. The "NOT" GATE

NOT GATE

A single switch and transistor produces a **NOT GATE.**
This is simply an **INVERSION**.

The resistor turns the transistor **ON** and the lamp illuminates. The switch removes the voltage on the base and the transistor turns **OFF**.
This is only a demonstration circuit to show how a switch can turn a lamp OFF.

to Index

The 5 gates above form the basis to turning a circuit **ON** and **OFF**. We will discuss these gates later in the digital section.

to Index

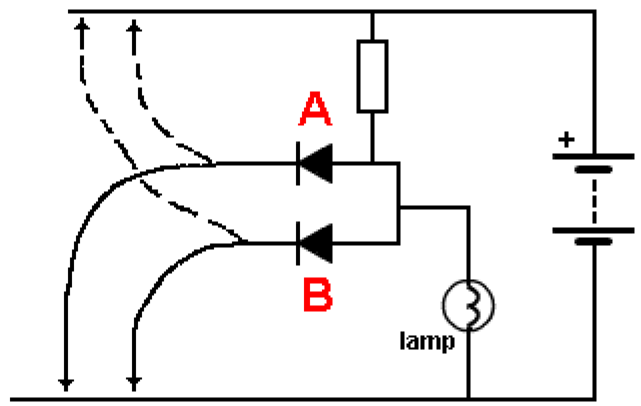

AND GATE with DIODES

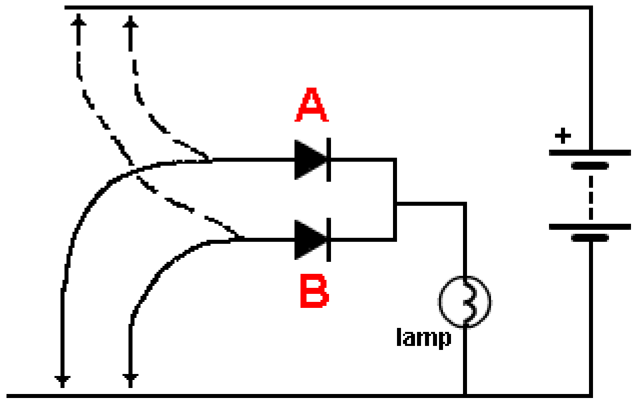

OR GATE with DIODES

Fig 83. "AND and "OR" gate with diodes

The next building block is the **GATING DIODE**.
We have shown a diode allows current to flow when the diode is correctly placed in a circuit and blocks current when it is reversed.
The 5 gates above are electrical circuits but an electronic circuit works in a slightly different way.
The electronic circuit will be covered later in the DIGITAL section.
For the moment we will explain how a diode can be used to create a **GATE**.
In other words it creates a ONE-WAY PATH to allows signals to pass from one stage to another and prevents signals passing in the opposite direction.
In the AND GATE circuit, both inputs are LOW and current flows through the resistor (it will get HOT). When one input is taken HIGH, current still flows through the other diode and the lamp does not illuminate.
When BOTH inputs are HIGH, current flows through the resistor to illuminate the lamp. No current flows through the diodes.
In the OR GATE, when one input is taken HIGH, current flows through the diode to illuminate the lamp.
This can be done with EITHER input.

to Index

A **NAND** and **NOR** gate can be made with diodes and a transistor.
This time we the output is either HIGH or LOW.
We are gradually producing circuits that

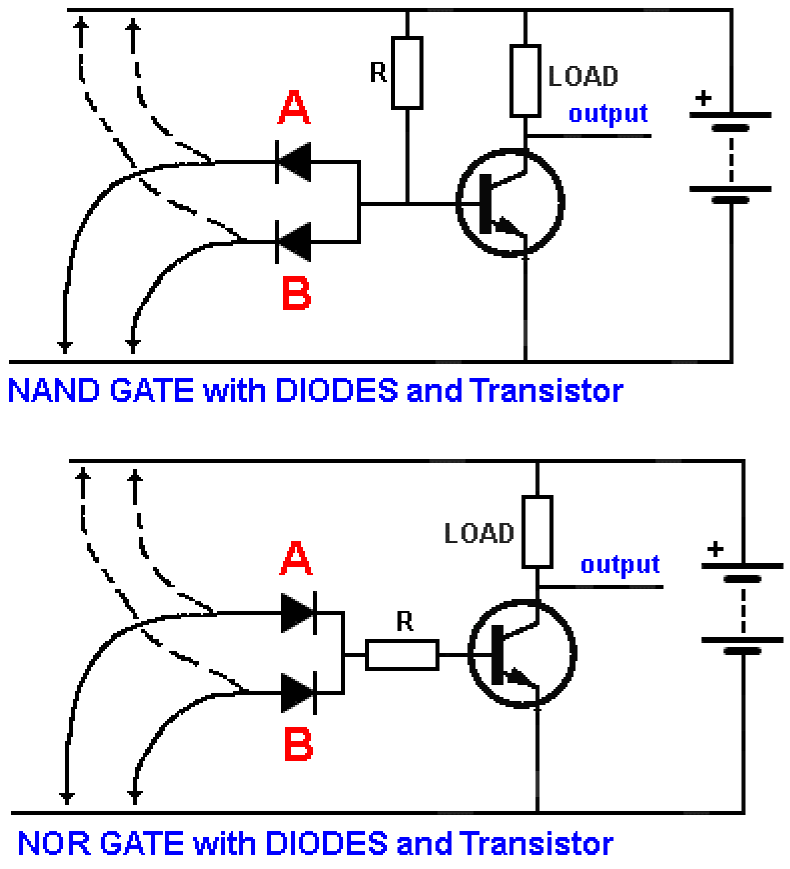
R
LOAD
output
+
A
B
NAND GATE with DIODES and Transistor
LOAD
output
+
A
R
B
NOR GATE with DIODES and Transistor

Fig 84. "NAND" and "NOR" gate with diodes and a transistor.

are electronic, rather than electrical circuits.
In the **NAND GATE** circuit, taking one of the inputs HIGH will still allow the other input to prevent the transistor turning ON.
When BOTH inputs are HIGH, the transistor turns on via resistor R and the output is LOW.

In the **NOR GATE** circuit, taking one of the inputs HIGH will turn the transistor ON and the output will be LOW.

to Index

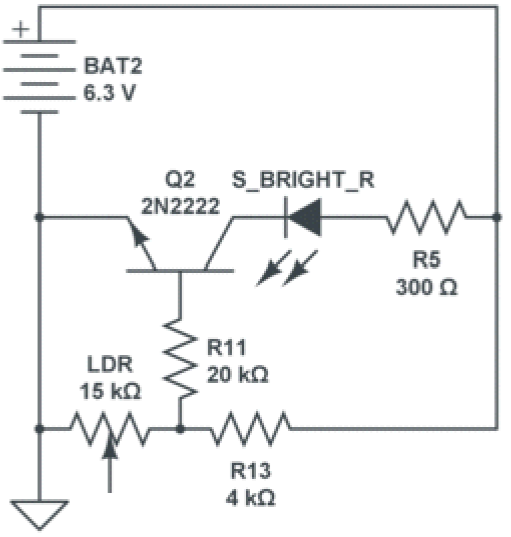
+
BAT2
6.3 V
Q2
2N2222
S_BRIGHT_R
R5
300 Ω
R11
20 kΩ
LDR
15 kΩ
R13
4 kΩ

Drawing A Circuit

A circuit must be drawn according to simple rules so it can be instantly recognised.
An electronics engineer can "**see a circuit working**" when it is drawn correctly and can see if it is drawn correctly; if the parts-values are correct and can use the circuit to assist in diagnosing a problem with a faulty circuit.

The top circuit on is very difficult to visualise because it is not drawn in the normal way.
All the components have to be "**turned around in your mind**," to see what the circuit is doing.

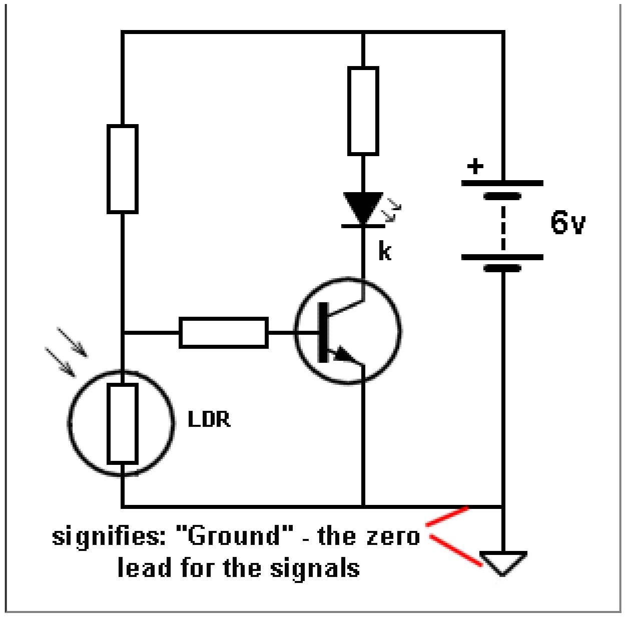

[to Index](#)

QUICK QUIZ - to see how much you know

following 50 questions . . . JavaScript is required!
This test will see how much you have learnt.

Easy To Use, All in One, CRM.
1.5 Million Insightly CRM users worldwide.
insightly TRY FOR FREE

FLUKE CONNECT
Capture equipment performance.
FLUKE
Learn more

Easy To Use, All in One, CRM.
1.5 Million Insightly CRM users worldwide.
insightly TRY FOR FREE

2 Layers $10 ea
4 Layers $25 ea
Order Now
RUSHPCB
ISO 9001:2008, ITAR, IL LISTED
Locates in Silicon Valley, San Jose, CA

☐ An open circuit [help](#)

4. What is the approximate characteristic voltage that develops across a red LED?

☐ 1.7v
☐ 3.4v
☐ 0.6v
☐ 5v [help](#)

5. If two resistors are placed in series, is the final resistance:

☐ Higher
☐ Lower
☐ The same
☐ Cannot be determined help

6. Which is not a "common" value of resistance:

☐ 2k7
☐ 1M8
☐ 330R
☐ 4k4 help

7. Which value of resistance, placed across a 9v battery will get hot:

☐ 22k
☐ 22R
☐ 220k help

8. If the voltage on the base of a transistor increases, does it:

☐ Turn on
☐ Turn off
☐ Not enough information
☐ Remain the same help

9. The resistor identified in brown is called the:

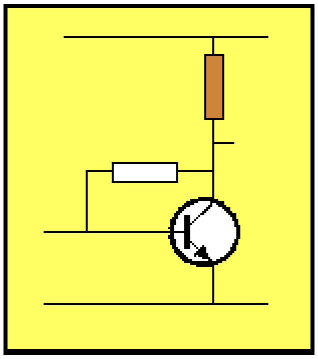

☐ Base Bias Resistor
☐ Load Resistor
☐ Emitter Feedback Resistor
☐ Bypass Resistor

help

10. The first three colour bands on a resistor are: yellow - purple - orange

☐ 47k
☐ 4k7
☐ 470k
☐ 4R7 help

11. A resistor with colour bands: red-red-red-gold, has the value:

☐ 22k 5%
☐ 2k2 5%
☐ 220R 5%
☐ 22R 5% help

12. The lead marked with the arrow is:

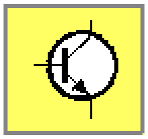

☐ The Collector
☐ The Base
☐ The Emitter
☐ The case help

13. A 10k resistor in parallel with 10k produces:

☐ 10k
☐ 5k
☐ 20k
☐ Cannot be determined help

14. The symbol is:

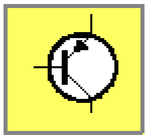

☐ NPN Transistor
☐ PNP Transistor
☐ Photo Transistor
☐ Field Effect Transistor help

**15. Two 3v batteries are connected as shown.
The output voltage is:**

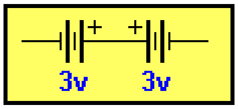

☐ 3v
☐ 0v
☐ 6v help

16. 4 resistors in ascending order are:

☐ 22R 270k 2k2 1M
☐ 4k7 10k 47R 330k
☐ 3R3 4R7 22R 5k6
☐ 100R 10k 1M 3k3 help

17. The closest value for this combination is:

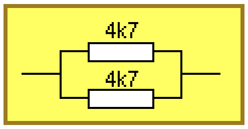

☐ 4k7
☐ 2k3
☐ 9k4 help

18. Which LED will illuminate:

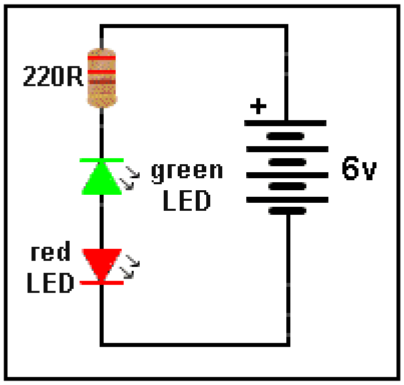

☐ Green
☐ Both
☐ RED
☐ None

19. The four symbols are:

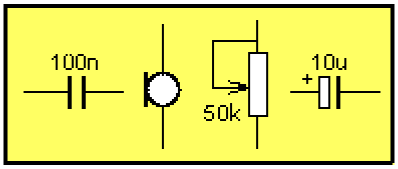

☐ Capacitor, Microphone, Potentiometer, Electrolytic
☐ Electrolytic, Microphone, Resistor, Capacitor
☐ Capacitor, Piezo, Resistor, Electrolytic
☐ Electrolytic, Coil, Resistor, Capacitor help

20. The closest value of the combination is:

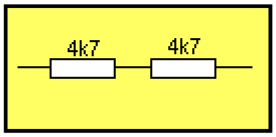

☐ 10k
☐ 2k3
☐ 2k5

21. Which motor will work:

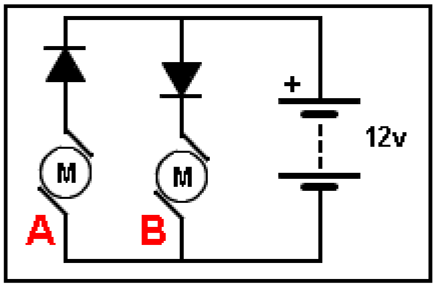

☐ A
☐ B
☐ A and B

22. A resistor and capacitor in series is called a:

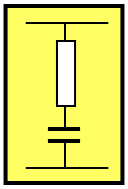

☐ Pulse Circuit
☐ Timing Circuit/Time Delay Circuit
☐ Oscillator Circuit/Frequency Circuit
☐ Schmitt Circuit help

23. A red-red-red-gold resistor in series with an orange-orange-orange-gold resistor produces:

☐ 5k5
☐ 35,200 ohms
☐ 55k
☐ None of the above help

24. Name the 4 components:

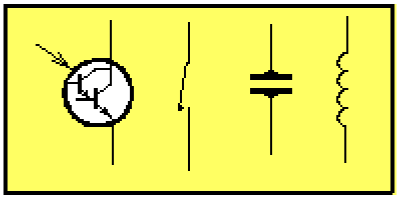

☐ Photo transistor, switch, capacitor, coil
☐ Transistor, mercury switch, piezo, inductor
☐ Photo transistor, reed switch, piezo, coil
☐ Photo Darlington transistor, switch, piezo, inductor help

25. To obtain a higher value of resistance, resistors are connected in:

☐ Reverse
☐ Forward

☐ Parallel
☐ Series help

26. Which LED will illuminate:

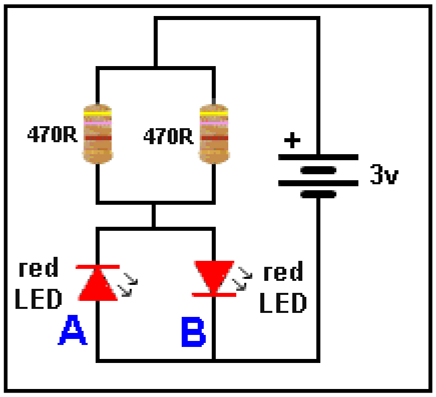

☐ A
☐ B
☐ Both
☐ None

27. Name the component that detects light:

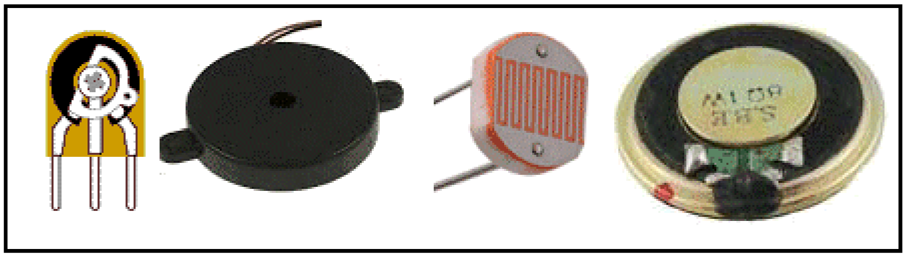

☐ mini trim pot
☐ Light Dependent Resistor
☐ piezo
☐ speaker help

28. What is 1,000p?

☐ 0.01n
☐ 0.0001u
☐ 0.1n
☐ 1n help

29. The current in a circuit is 45mA. This is:

☐ 0.045Amp
☐ 0.00045A
☐ 0.0045A
☐ 0.45A help

30. A 100n capacitor can be expressed as:

☐ 0.1u u = microfarad
☐ 0.01u
☐ 0.001u
☐ none of the above help

31. 1mA is equal to:

☐ 0.001A
☐ 0.00001A
☐ 0.01A
☐ 0.1A help

32. 1,200mV is equal to:

☐ 12v
☐ 1.2v
☐ 0.12v
☐ 0.0012v help

33. The approximate current for a toy 3v motor is:

☐ 10mA
☐ 100mA to 300mA
☐ 1 amp

34. What is the resistance of this resistor:

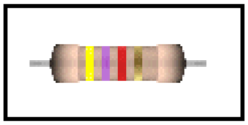

☐ 47k
☐ 4k7
☐ 4R7

35. Identify the correctly connected LED:

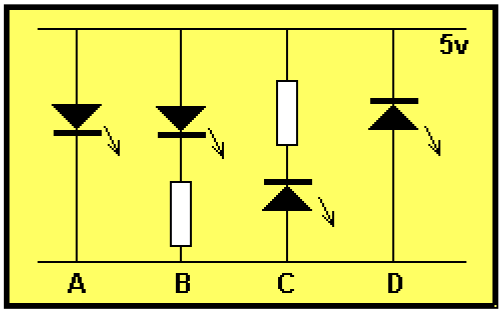

☐ A
☐ B
☐ C
☐ D help

36. Identify the correct statement:

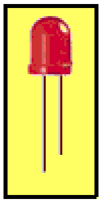

☐ The cathode lead is longer. It goes to the negative rail
☐ The cathode lead is shorter. It goes to the negative rail
☐ The cathode lead is shorter. It goes to the positive rail
☐ The cathode lead is longer. It goes to the positive rail help

37. The current requirement of a LED is:

☐ 1.7mA
☐ 25mA
☐ Between 3 and 35mA
☐ 65mA help

38. The multimeter is measuring . . .

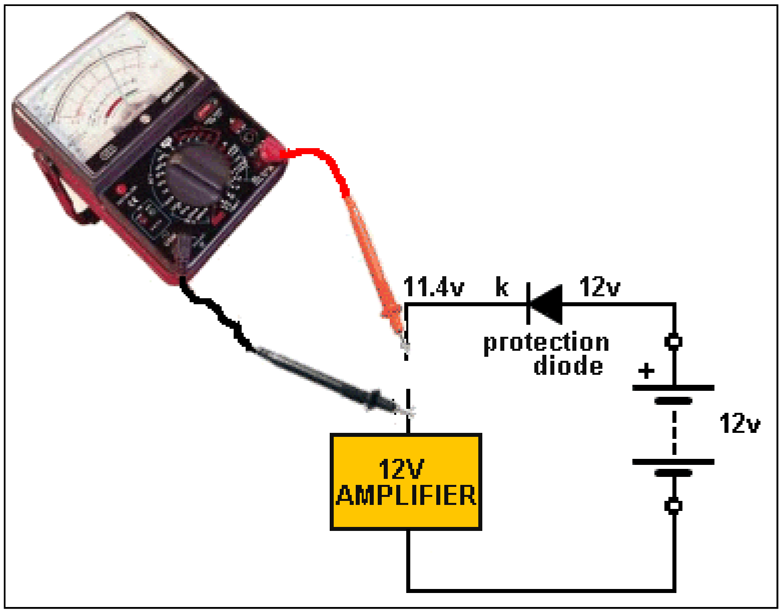

☐ Voltage
☐ Current
☐ Resistance

39. The purpose of the two capacitors:

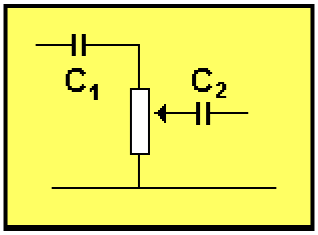

☐ To pass AC from the input to the output
☐ To allow the signal to oscillate
☐ To pass DC from the input to the output
☐ To amplify the signal help

40. The direction of conduction for a diode is:

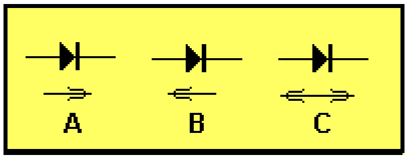

☐ A
☐ B
☐ C help

41. A DC voltage . . .

☐ rises and falls
☐ is a sinewave
☐ remains constant
☐ is an audio waveform help

42. Arrange these in ascending order: k, R, M (as applied to resistor values)

☐ R, k, M
☐ M, R, k
☐ k, M, R
☐ M, k, R help

43. A battery produces AC current:

☐ true
☐ false

44. The tolerance bands: gold, silver, represent:

☐ 5%, 10%
☐ 10%, 5% help

45. 223 on a capacitor represents:

☐ 0.022u u = microfarad
☐ 22n n = nanofarad
☐ 22,000p p = picofarad
☐ All of the above help

46. Arrange these in ascending order: n, p, u (as applied to capacitor values)

☐ p, u, n,
☐ n, u, p
☐ p, n, u help

47. What is the resistance of this resistor:

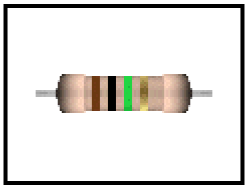

☐ 100k
☐ 1k
☐ 1M
☐ 100R

48. The number "104" on a capacitor indicates:

☐ 0.1u
☐ 100n
☐ 1n
☐ 10n help

49. What is the multimeter detecting:

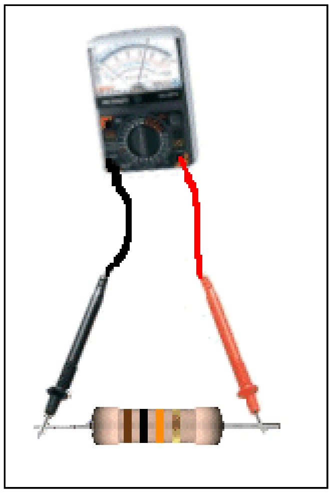

☐ Voltage
☐ Current
☐ Resistance

50. For the LEDs, what is the characteristic voltage for the red and white LEDs:

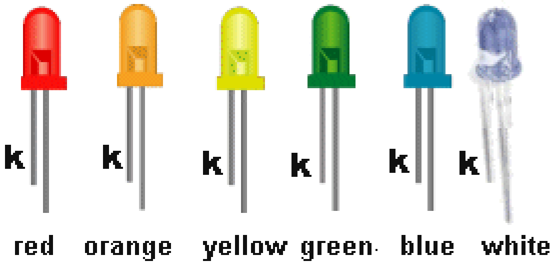

☐ 3.6v, 1.7v
☐ 2.4v, 3.3v
☐ 1.7v, 3.6v

☐ Cannot be determined

Final Assessment

THE CRYSTAL-SET RADIO

100 more Crystal Set plans

Everyone wants to make a RADIO.

The simplest radio is a **CRYSTAL SET**. (Sometimes called a **Crystal-Set Radio** or **Xtal Radio Set** or **Crystal Diode Radio**.)

However a **Crystal Set** needs a number of components that are very hard to get: (tuning capacitor with knob) and (crystal earpiece for $1.25).
The "air" TUNING CAPACITOR (20p to 415p) is not easily available and the postage is expensive. The germanium diode is a special component and the aerial coil wound on a ferrite slab is difficult to obtain.

But you don't need these components. They can be substituted.

There are hundreds of websites on the internet describing the **CRYSTAL SET** and if you want to build a "normal" set, you can Google these sites or 100 more Crystal Set plans. Many of them sell kits too.

But this article is different.

We are going to have all the fun of making a **CRYSTAL SET** but with modern components and easy-to-make components and with an amplifier stage. The output is loud so you don't need a long antenna. And we are going to make our own TUNING CAPACITOR and a very simple aerial coil (called a FRAME AERIAL) as well as replacements for the germanium diode (use a TRF radio IC or a transistor) and in place of hi-impedance headphones (use a piezo diaphragm) and a crystal earpiece equivalent (a piezo diaphragm).
It's even better to have one of each type of component so you can compare the performance, so no matter how many parts your get, nothing will be wasted.

We are also going to explain the fundaments of how the circuit works as even the simplest circuit has a number of very important features that are used in many other circuits.

But first we are going to learn about the components and how they combine to make the circuit work.
When two or more components are connected together they sometimes produce a completely different result to the capabilities of either item.
This is the case with a capacitor and inductor in parallel. An inductor is simply a coil - turns of wire on a cardboard tube - called a former and the centre of the coil is AIR. It is called an air-cored coil or air-cored inductor.
Each component (the coil and capacitor) is called a PASSIVE DEVICE - in other words it does not amplify, but when the are connected together they create a result very near to amplification. And they also produce a result of picking up a huge number of signals and only allowing one signal to appear across the pair. A truly amazing result.

We start by placing a capacitor across the coil.
There is so much activity in the air, from radio, TV, taxi and mobile phone usage that the air is filled with electromagnetic radiation.
This radiation will cut the turns of the inductor (the coil) and produce a microscopic voltage in the turns. This is enough to start the two components passing energy back and forth at a rate

determined by their values. This is the basis of our first discussion.
These two components are called a TUNED CIRCUIT and make up our first building block called THE FRONT END.
.

THE FRONT END

This consists of a coil and capacitor. These two components are in PARALLEL and the signal (called the RADIO SIGNAL or RADIO WAVE) passes through the centre of the coil and produces a voltage in the turns of the coil. The electromagnetic wave has to pass through the centre of the coil.
This voltage is the result of a mass of signals that are interfering with each other and producing a signal called BACKGROUND NOISE.
The voltage can be increased by an external aerial (called an ANTENNA) and it consists of ALL THE LOCAL radio stations (and everything else).
The voltage is a mass of signals and is absolutely useless as it represents all the stations AT THE SAME TIME.
However, across the coil is a capacitor and these signals charge the capacitor with the very small voltage produced by the energy of the signals. When the capacitor is charged, it delivers its voltage to the coil. The coil accepts the energy and converts it to magnetic flux.
After a very short period of time the capacitor becomes discharged and the magnetic flux collapses and produces a voltage in the coil of the opposite polarity to charge the capacitor again in the opposite direction.
These two components keep oscillating back and forth, using the tiny amount of energy from the stations.
There is a natural frequency for the capacitor and coil to pass energy back and forth and one of the stations will provide energy to assist this natural frequency.
When this happens, the amplitude of the signal increases and the signals from all the other stations cancel themselves out and only one signal (waveform) remains.
This signal is called the NATURAL FREQUENCY OF RESONANCE and it corresponds exactly to one of the radio stations.

The end result is a waveform that is the exact same frequency as one of the radio stations and when voice or music is played, the amplitude of the waveform increases and decreases. This will be the signal you hear in the earpiece or speaker.
It is called an AMPLITUDE MODULATED signal and that is where we get **AM RADIO** from.

Understanding the concept of a CAPACITOR and INDUCTOR in parallel is very important. They form a TUNED CIRCUIT that has a natural RESONANT FREQUENCY.
Here is a very similar analogy. You have a heavy metal ball on a long string attached to the gutter on your house - just like a pendulum. You can push the ball very lightly with a finger and after a number of pushes you will be able to get the heavy ball swinging in a very large arc.
The only way to keep it swinging is to push it very lightly at exactly the right time. If you push it at the wrong time it will eventually stop swinging.
The parallel tuned circuit is exactly like the ball. It wants to oscillate at a particular frequency.
All the radio stations are pushing and pulling the circuit at the wrong times and nothing is happening. But one radio station pushes at exactly the right time and the circuit starts to oscillate.
All the other stations are fighting each other just like one person pushing the ball sideways and another pushing the ball from the opposite side. The results cancel each other and you are the only one assisting the swing.
The TUNED CIRCUIT can also be called a FILTER with a very narrow BAND-PASS frequency but our simple explanation describes the operation much more clearly.

There are a few other terms used to describe the components in the font end:

LOOP STICK ANTENNA -This is an alternate name given to the coil of wire wound on a ferrite rod or slab. It also has the name ROD ANTENNA or FERRITE ROD ANTENNA.
The winding can be enamelled wire or flexible wire called LITZ WIRE. This is very fine strands of enamelled wire twisted together and covered in cotton. The purpose of changing a thick wire to lots of very thin wires is to prevent the radio signals creating loops of signals within the wire and these signals will cancel each other and not produce a signal out the end of the wire.
In our experiments, we have not noticed any difference in a coil made with ordinary enamelled wire and Litz wire.

Here is a set of components to make your own Crystal set from **Scott's Electronic Parts**:
You can see the rod antenna, germanium diode, crystal earpiece, capacitor and resistor. The kit costs about $9.00 plus postage and includes knob, clips and screws but no board to mount the parts.

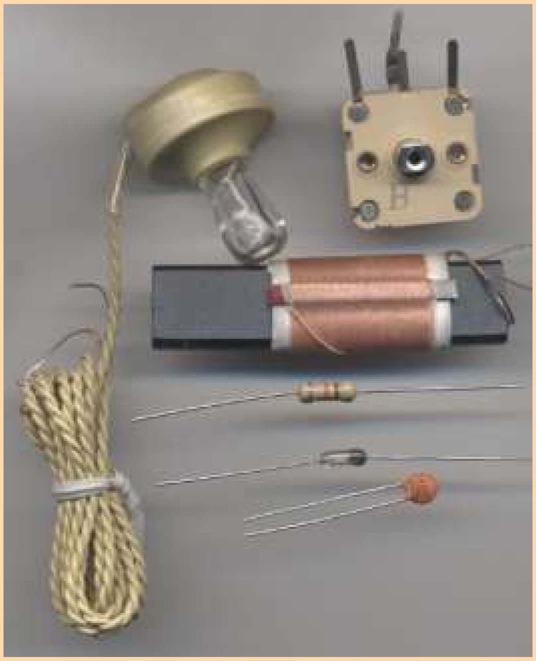

Here is the circuit for a Crystal Set:

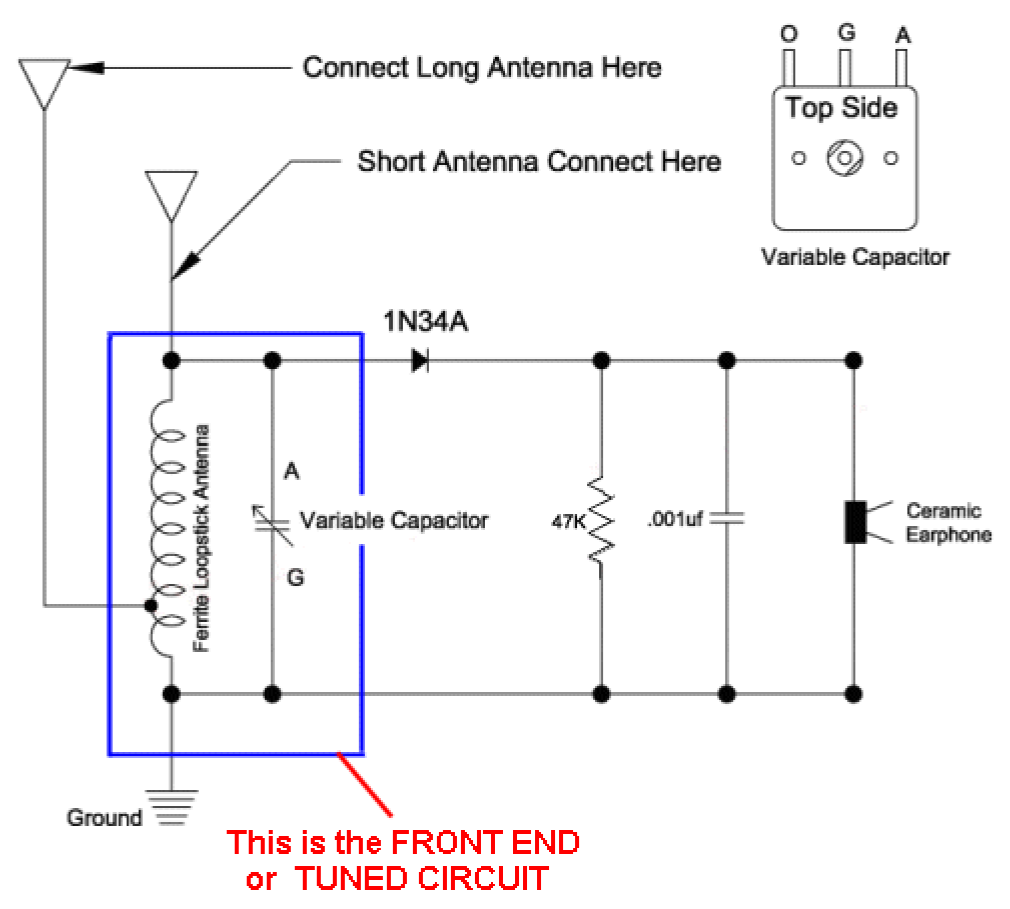
Connect Long Antenna Here
O
G
A
Top Side
Variable Capacitor
Short Antenna Connect Here
1N34A
A
Ferrite Loopstick Antenna
Variable Capacitor
G
47K
.001uf
Ceramic Earphone
Ground
This is the FRONT END or TUNED CIRCUIT

Here are the components mounted on a board called BREADBOARD:

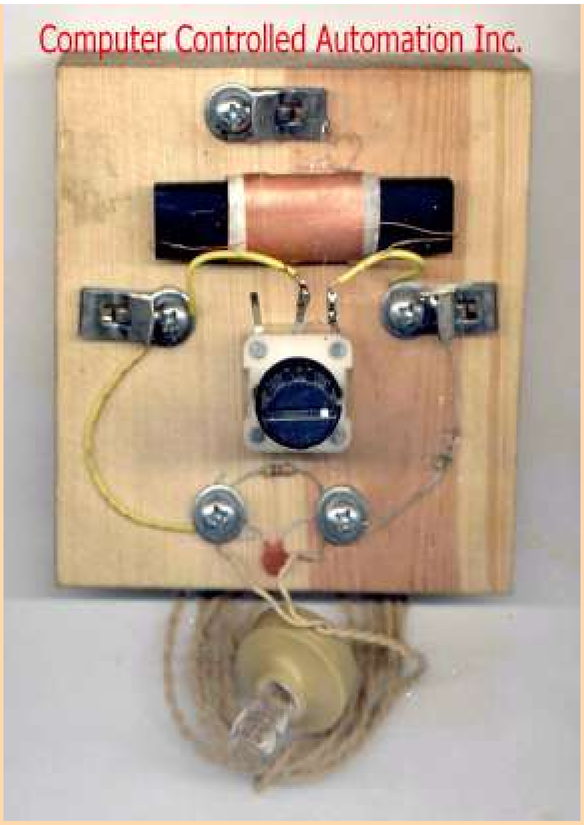

The top clip connects to a long antenna. The left clip connects to ground and the right clip connects to a short antenna.

Let me clear up a point. You do not need a ferrite rod antenna for the coil. You can use an ordinary coil wound on a cardboard tube and it will work just as well if you are using an outside antenna.
Here is the circuit using a home-made coil.

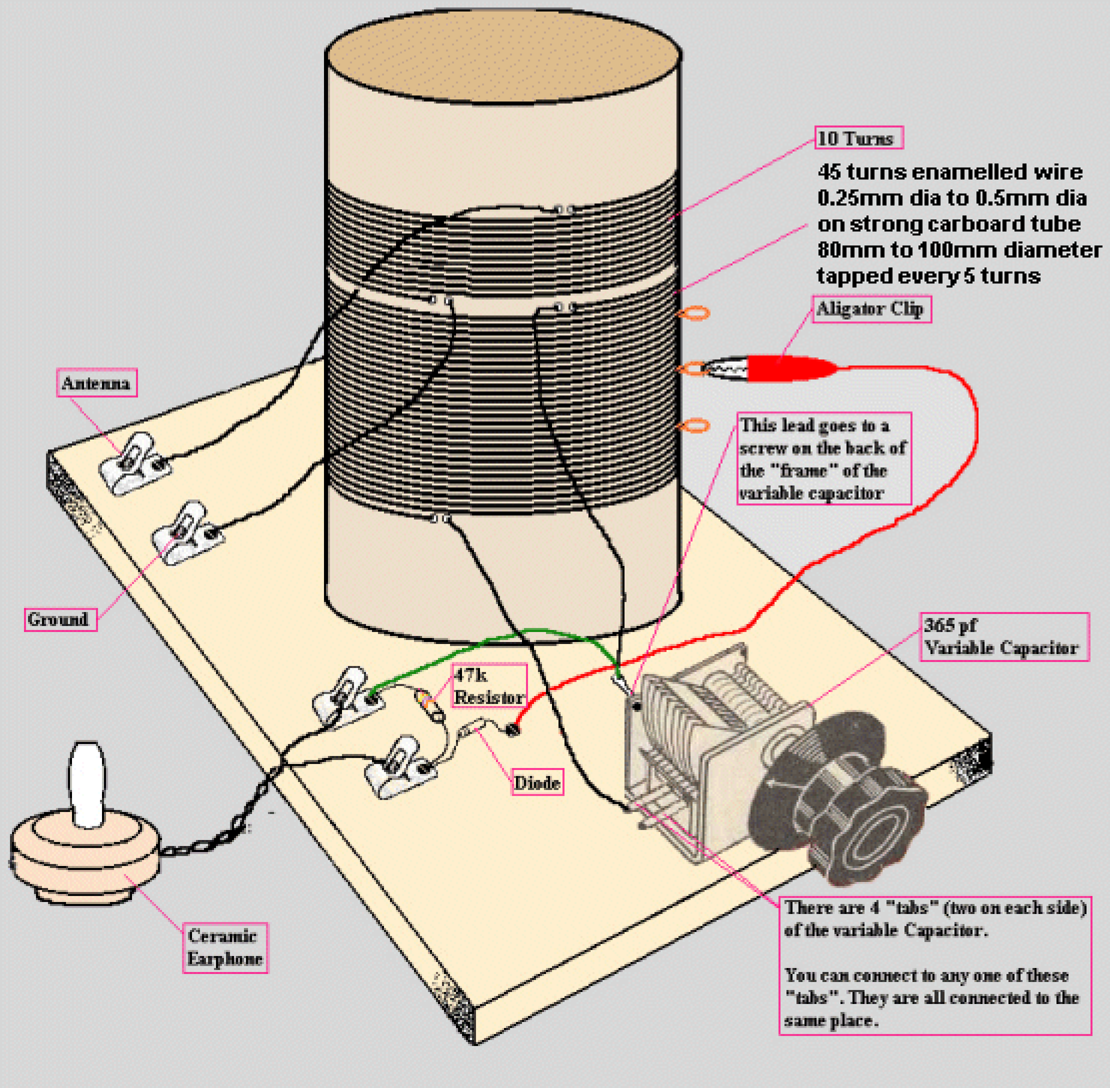
10 Turns
45 turns enamelled wire
0.25mm dia to 0.5mm dia
on strong carboard tube
80mm to 100mm diameter
tapped every 5 turns
Aligator Clip
Antenna
This lead goes to a screw on the back of the "frame" of the variable capacitor
Ground
365 pf Variable Capacitor
47k Resistor
Diode
Ceramic Earphone
There are 4 "tabs" (two on each side) of the variable Capacitor.
You can connect to any one of these "tabs". They are all connected to the same place.

The tappings on the coil allow a wide band of radio stations to be tuned.
Each tapping allows a different portion of the band to be covered.

The next part to understand is this:
The coil and capacitor must not be LOADED. In other words, you cannot connect anything to this combination because the signal it is producing will be "taken away" or "removed" or "considerably reduced" by the item you are connecting to the circuit.
These two components are called a TUNED CIRCUIT and when they are not loaded they pick up all the radio stations, one station at a time, when the natural resonant frequency of the coil and capacitor exactly match the frequency of the radio station. The circuit actually "rejects" all the radio stations except one. Because all the other stations are trying to make the Tuned Circuit oscillate at a different frequency and it does not do this.
The result of the TUNED CIRCUIT oscillating under NO LOAD conditions produces a waveform that is very high and this gives the circuit GOOD SELECTIVITY. The circuit can select one station and reject nearby stations.
It also has good SENSITIVITY as it can pick up weak stations.
If you load the circuit, only the strongest signal will be detected and it will be spread across the full range of the tuning capacitor.
Obviously the theory is more-complex but we are explaining the end-result.
Theory talks about the "Q" value of the coil and this is its ability to produce a very good output when the magnetic flux collapses and the "Q" value increases when the circuit is not loaded. .
Although these voltages are very small (in the order of microvolts or millivolts) the result is very important as the rest of the circuit will be amplifying this waveform a few thousand times.
As we explained above, pushing the weight on a string only needs a push of 1 cm and eventually the weight will swing 1 metre. This is a gain of 100:1 The same thing happens with the tuned circuit. The incoming radio signal is in the order of microvolts, but the coil and capacitor will produce a signal as high as 500 millivolts. This is an improvement or "gain" of more than 1,000 and is referred to as the "Q" of the circuit.
You will also notice the TUNED CIRCUIT is not connected to any supply voltage. It does not have

be connected. It generates its own waveform from the signals in the air.
It should not have any DC current flowing through it via the supply as this would put a load on the circuit and reduce its operation.

However we must "pick-off" the signal so it can be amplified.
This must be done with a very high impedance circuit.

THE CONVERTER (detector) - THE DIODE

The next part of the circuit is the CONVERTER. Commonly called the DETECTOR.
It converts the RADIO FREQUENCY signal to an AUDIO FREQUENCY signal.
This is the job of the DIODE.
The radio frequency signal is a very high frequency signal (say one million cycles per second) and it is sending a tone of one thousand cycles per second through the air-waves.
What is happening is this: The one megahertz signal has a certain amplitude and over a range of the first one-thousand cycles, the amplitude gradually decreases and then increases again. If you look at the tops of this 1,000 cycles you will see a waveform that corresponds to the one kilo-Hertz signal.
The 1MHz signal is picked up by the coil and capacitor in the front end and makes it oscillate.
The radio frequency signal is gradually getting larger over 500 cycles then smaller over the next 500 cycles and this increase and decrease represents the 1,000 cycles per second tone.
This is the waveform (the signal) that passes through the diode. This will be explained further in a moment.
The diode does not pass any signals less than 200mV as the first 200mV is lost in the junction of the diode. This means the signals start to appear on the other end of the diode when they are above 200mV.

This is how the diode works:
Across the crystal earpiece is a capacitor. The capacitor gets charged via the diode.
The diode is present to stop the capacitor getting discharged when the waveform is in the wrong direction. (by this we mean - when the waveform is lower or smaller in amplitude than the voltage on the capacitor).
And the waveform is in the wrong direction about 50% of the time. To charge the capacitor for one-half-cycle requires 500 "little increments" in voltage with each increment adding a microscopic increase in voltage. We don't want this voltage to reduce when the waveform is reversing direction and the diode stops the voltage flowing back to the Tuned Circuit.
During the next half of the cycle when the pulses are getting smaller and smaller, the voltage on the capacitor is "bled off" by the load resistor.
The crystal earpiece detects this voltage. What we mean, is the diode allows the voltage to rise (increase) on the capacitor via lots of little "pulses" and the voltage increases in the form of a sinewave to a maximum amount. This voltage is passed to the crystal earpiece.
Once the voltage rises to a maximum, the little pulses of energy are not quite as strong, and the voltage on the capacitor reduces to form the second portion of the sinewave. This voltage is always being passed to the crystal earpiece and you can hear it as an audio signal.

GERMANIUM OR SILICON DIODE

The preferred type of diode for a Crystal Set is germanium. This is because it drops only about 0.3v.
But a silicon diode can be used, even though it drops about 0.7v, if the radio stations are very loud (close by).
You have to remember, you need a very good aerial (and a water-pipe earth) to get any results with a Crystal Set because you are asking the signal to provide the energy to drive the earpiece.
By simply adding a transistor, you are improving the performance 100 times and the long antenna can be reduced to a FRAME ANTENNA and the earth can be the metal frame of your soldering iron.

THE EARPIECE or EARPHONE
also The Magnetic Earpiece or CRYSTAL EARPIECE

The earphone or earpiece used in a Crystal Set must be a high impedance device because the crystal set does not produce a high current and cannot drive a low-impedance earpiece. That's why a CRYSTAL EARPIECE is ideal.
It has a crystal glued to the back of the earpiece and connected to its top surface is an aluminium diaphragm. When the crystal expands and contracts as a result of a voltage applied via two electrodes, the diaphragm moves and you can hear the signal. It exhibits a very high impedance

because it consists of a crystal and no coil of wire is contained inside the case.
If you do not have a Crystal Earpiece, you can make your own from the shell of an 8 ohm earpiece and a piezo diaphragm. Only the front part of the earpiece is used.

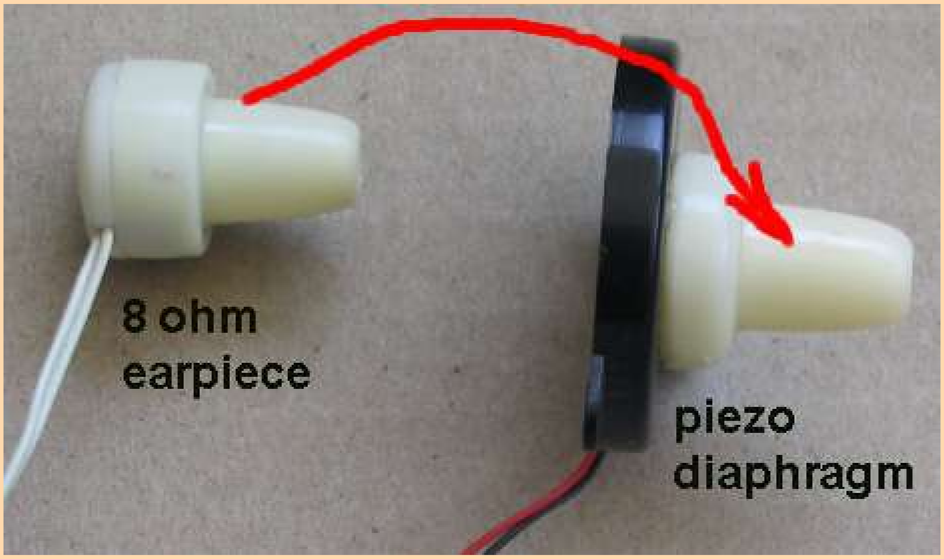

Make your own Crystal Earpiece

Hit the 8 ohm earpiece on the side and the front comes off. Glue the front onto a piezo diaphragm with hot-melt glue. See photo above.
The piezo diaphragm is a ceramic substrate that deflects in the presence of a voltage. It is quite sensitive and you can hear the audio quite clearly.
The waveform emerging from the diode in a Crystal Set is called AUDIO and although it has an amplitude of a few hundred millivolts, it does not have any current associated with it. The crystal earpiece and the piezo diaphragm react to this voltage.

THE 8ohm EARPIECE

The 8 ohm earpiece can be used with our 8ohm Buffer stage shown below.

16ohm 32 ohm and 64 ohm EARPIECE(s)

Earpieces and headsets from mobile phones are 16 ohm or 32 ohm per earpiece and are terminated via a stereo 2.5mm or 3.5mm plug. The earpieces are connected in SERIES to get the best coupling to our radio circuits and you need to find the two pins on a stereo socket to produce series connection. Get a multimeter and switch to "ohms." Try all the pins and you will get a click in the left ear then the right ear. Keep searching until you get a click in both earpieces at the same time. Use these two pins.

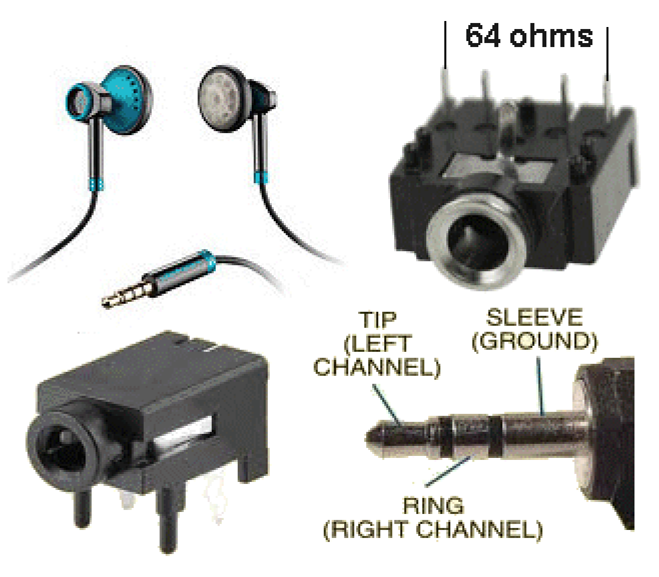

Stereo mobile phone headset - unusually 32R or 64R

PROBLEMS

The biggest problem with a Crystal Set is the need for a long antenna.
The first 200mV to 300mV of a signal is lost in the diode and you need a long antenna to pick up a signal so the output of the TUNED CIRCUIT has enough voltage to drive the high-impedance earpiece.
This requires an outside aerial 5 metres long and 3 metres high.
This is not practical for most hobbyists so we will be adding an amplifying stage to the crystal set so a shorter (smaller) aerial can be used.

THE FRAME AERIAL or FRAME ANTENNA or FRAME COIL

The aerial coil shown in the photo above is a ferrite slab with about 80 turns of Litz wire. You can find one of these in an old broken AM radio or from a parts-shop. In the instructions below we show how to make your own Ferrite Rod Antenna
An equally-good substitute is a frame antenna made by winding insulated wire in a rectangle around wooden sticks.

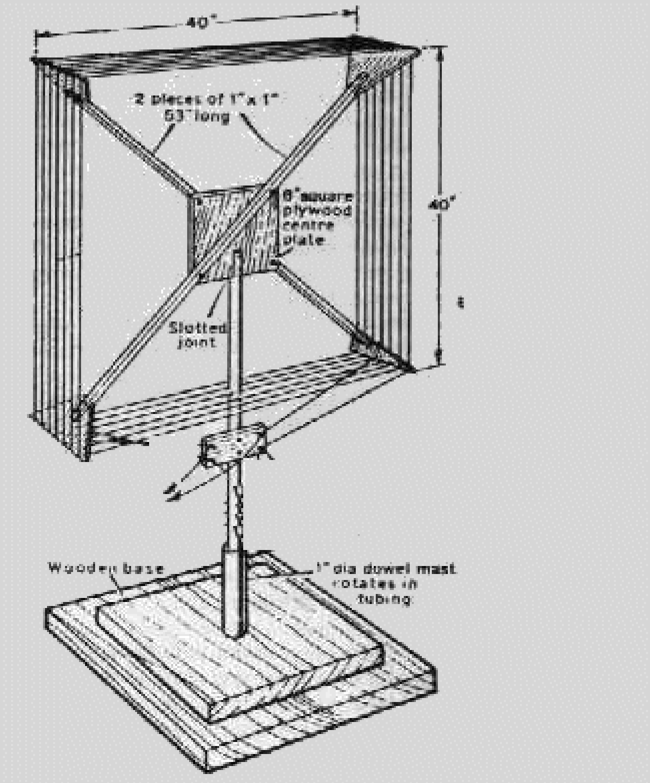
40"
2 pieces of 1" x 1"
53" long
8" square
plywood
centre
plate
40"
Slotted
joint
Wooden base
1" dia dowel mast
rotates in
tubing

FRAME ANTENNA

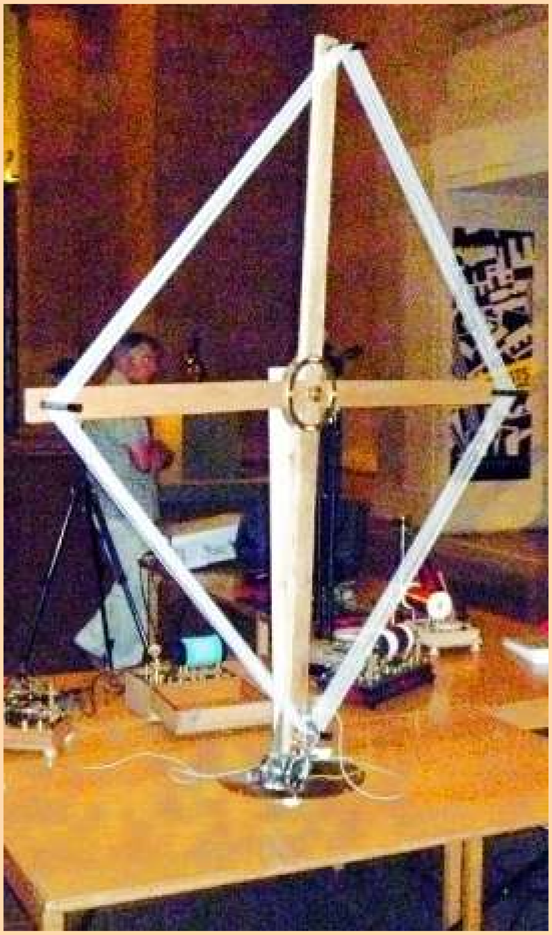

15 turns on a diamond frame

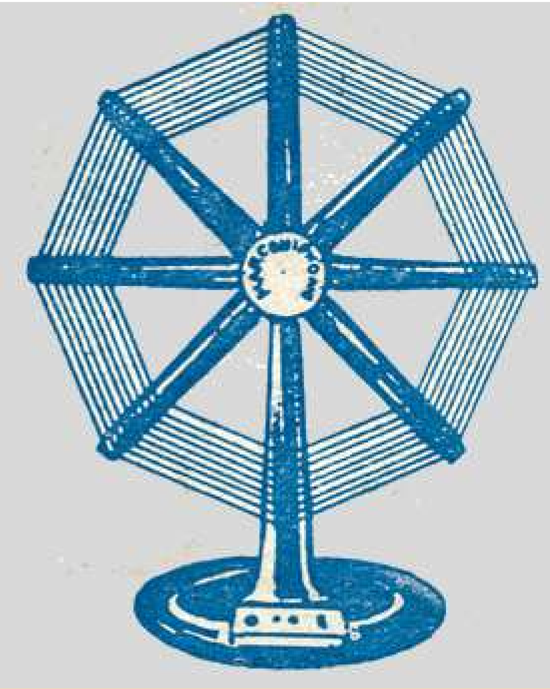

One of the earliest Frame Antennas

The Frame Aerial can be as large as 100cm x 100cm or as small as 10cm x 10cm around a plastic chocolate box.

Here are two FRAME ANTENNAS:

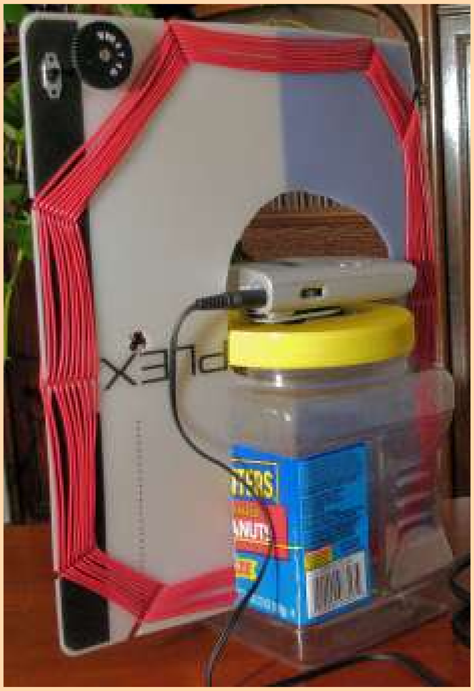

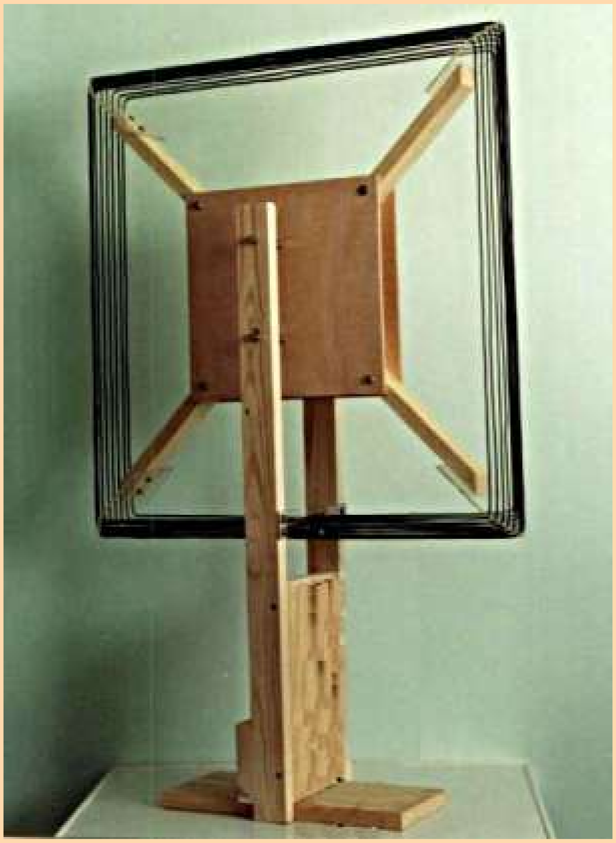

This will work just as good as a ferrite slab antenna. The slab antenna is just 100 times smaller. The slab antenna was invented so a transistor radio could be built in a small case. But if it is not available, you can wind 20 turns around a plastic chocolate box and it will work just as good.

Alternately you can wind 20 turns around a biscuit tin. Put a pencil on the tin and wind the turns over the pencil too. Remove the pencil and it will be easy to remove the turns. Use tape to keep the turns together.

The FRAME AERIAL does two things. It picks up the radio waves and it becomes the coil (called the INDUCTOR) in the TUNED CIRCUIT. It must be placed away from metal objects, such as a refrigerator.

BASKET WEAVE COIL

There is no point making a complex BASKET WEAVE COIL as it will not work any better than simply jumble winding all the turns at the maximum circumference of the coil, because the energy capturing capability of the coil relies entirely on the amount of flux lines passing through the centre of the coil.
By increasing the centre of the coil, the amount of flux is increased for the same coil size.
In fact, the simplest and cheapest is to wind turns around a box, as explained later in this article, or make a frame antenna as shown above. Technically speaking, a round coil has the best performance but only by a few percent.

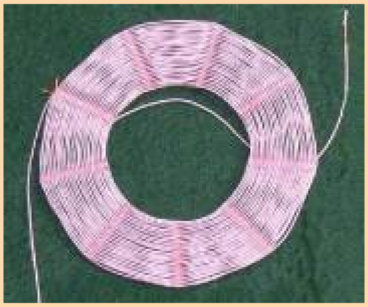

A BASKET WEAVE COIL

THE VARIABLE INDUCTANCE TUNING COIL

Whenever the size or shape of the coil is changed, (or the number of turns), the natural frequency of the Tuned Circuit will change and a different radio station will be picked up.
This means tuning across the band can be done by altering the characteristics of the coil while keeping the value of the capacitor fixed.
Changing the inductance can be done in many different ways.
The coil can have taps every 5 turns and an alligator clips selects the correct tap. But very few radio stations will correspond exactly to each tap.
Another way is to have a slider move up and down the turns as shown in the following image:

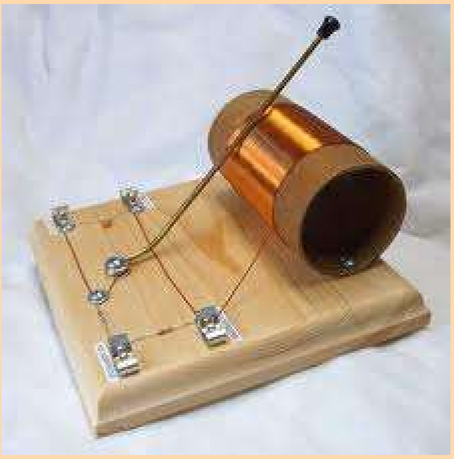

The slider makes contact where the insulation has been removed. But it may touch two turns at the same time and create a "shorted turn" and reduce the "Q" of the coil.

Another way is to move a ferrite bar (rod) in and out of the coil:

THE SLUG TUNED COIL

To tune across the radio band, the natural frequency of oscillation of the TUNED CIRCUIT must be adjusted (changed). This can be done by changing the value of the capacitor or the value of the inductor.

The value of the inductor can be changed by adding or removing turns or changing the amount of magnetic material in the centre of the coil.
A ferrite bar can be screwed in and out of the coil or slid in and out and this component is called a **SLUG TUNED COIL**.
The following diagram shows a **SLUG TUNED CRYSTAL SET**:

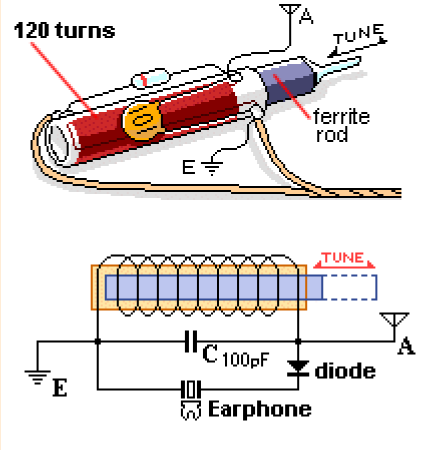
120 turns
A
TUNE
ferrite rod
E
TUNE
A
C 100pF
diode
E
Earphone

By changing the value of the 100p capacitor, different parts of the band can be picked up.

The photo shows a slug tuned coil using 60 turns of insulated wire on a 10mm tube (or any tube that will fit over a 8-10mm ferrite rod) and a circuit containing an AM radio chip plus a buffer driver transistor:

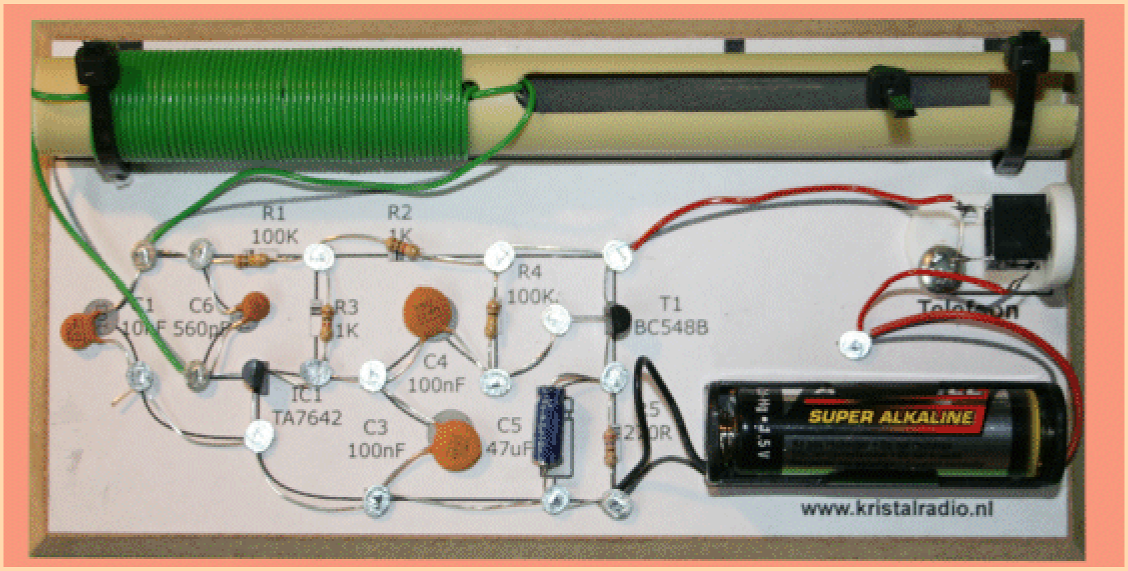
R1
100K
R2
1K
R3
1K
R4
100K
C1
10nF
C6
560pF
IC1
TA7642
C3
100nF
C4
100nF
C5
47uF
R5
220R
T1
BC548B
Telefoon
SUPER ALKALINE
www.kristalradio.nl

A SLUG-TUNED RADIO

The circuit above is has a broad-band amplifier consisting of 10 transistors (IC1) and they are directly coupled (connected) to each other because it is not possible to "manufacture" a capacitor inside the IC. The IC has 3 terminals (pins, legs) and it looks like an ordinary transistor.
Experimenting with this type of IC has shown that it is no better than 2 ordinary transistors connected in a direct-coupling arrangement.

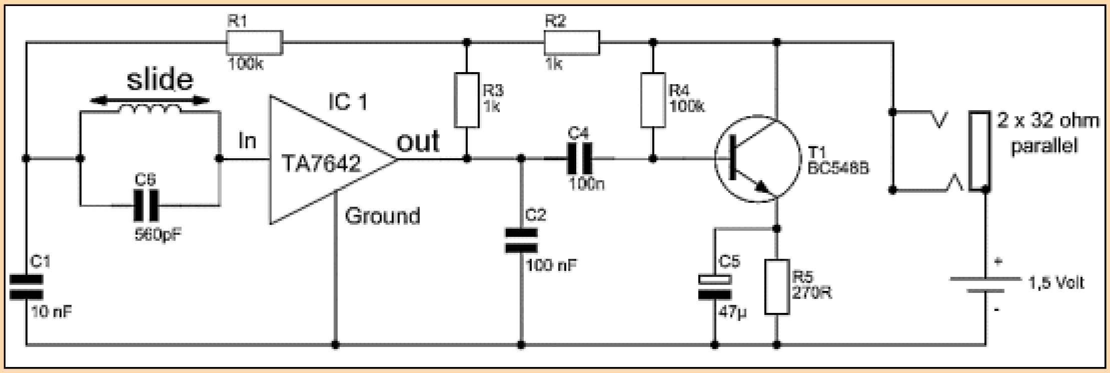
R1
100k
R2
1k
slide
IC 1
R3
1k
R4
100k
In
out
TA7642
C4
100n
T1
BC548B
2 x 32 ohm
parallel
C6
560pF
Ground
C2
100 nF
C1
10 nF
C5
47μ
R5
270R
+
1,5 Volt
-

Here is the address of the site for the slug-tuned radio.
http://www.kristalradio.nl/
Unfortunately the site is in Dutch and the kit is not available. However the photos give a clear picture of the how the parts are connected.

The inductance of the coil can also be altered by winding another coil and placing it near the first coil so that the magnetic field interacts with each other and changes the inductance of the circuit. This is called a VARIABLE INDUCTANCE TUNING COIL.
You can have one coil inside the other, two coils near each other or two flat coils side-by-side. Any two coils will interact with each other.

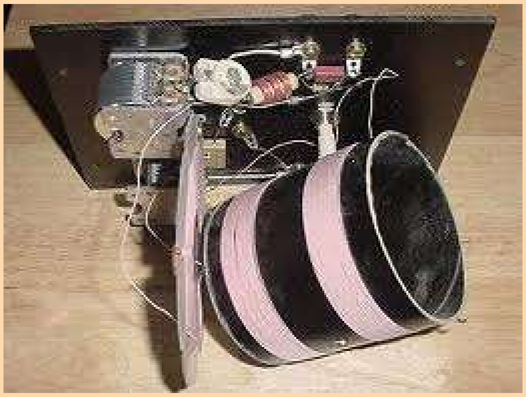

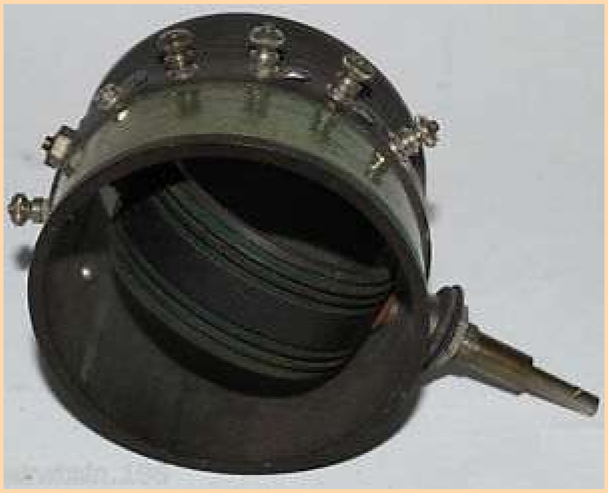

An Inductive TUNING COIL called a VARIOMETER

MAKING YOUR OWN FERRITE ROD ANTENNA

You can make your own FERRITE ROD ANTENNA by winding 60 to 80 turns of 0.25mm enamelled wire onto a 9mm ferrite rod or slab. If you wind it on a paper sleeve, you can move the coil along the rod to get the best performance. When the rod is slid out of the coil, the inductance changes considerably. However the inductance does change very slightly when the coil is moved along the rod.

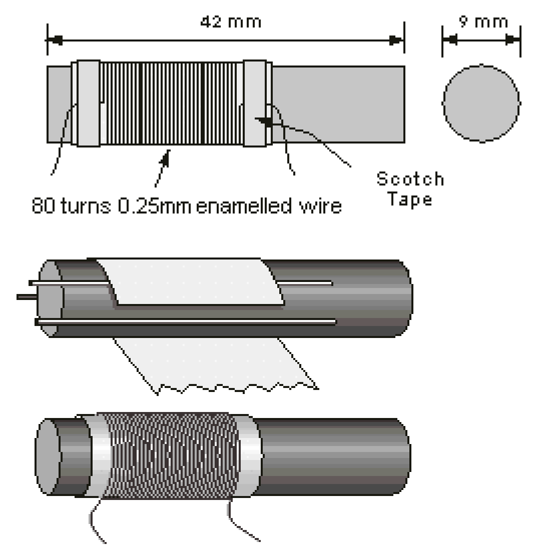

Make your own ferrite antenna

Now we come to the tuning capacitor::

THE TUNING CAPACITOR

The "C" in the "LC" TUNED CIRCUIT can be fixed or variable. When it is variable, it is called a TUNING CAPACITOR. The sheets of aluminium in the air tuning capacitor below are called PLATES and the moving plates are called VANES. The fixed plates make up the STATOR.
The space between the plates is AIR. The photo shows a single capacitor. If two capacitors are connected to the same shaft it is called a GANGED CAPACITOR.
The plates do not come fully out of mesh and that's why the capacitor has a minimum value. The maximum capacitance is when the plates are fully meshed. The odd shape of the plates is designed to produce a fairly constant increase in capacitance as the plates are engaged.

An air tuning capacitor:

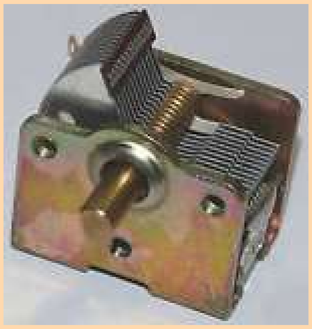

Air Tuning Capacitor
(Variable Capacitor)

The capacitor can be made much smaller by using thinner vanes and placing plastic between the vanes. Plastic increases the capacitance about 3 times to 10 times.

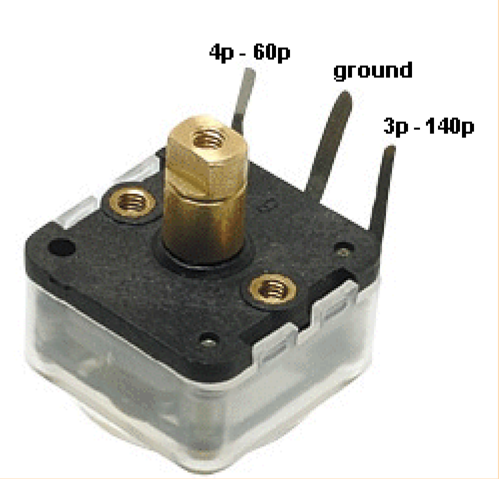

Tuning Capacitor as found in a pocket radio

The tuning capacitor can be replaced with a home-made equivalent that will work just the same.
You need:
4 sheets of aluminium foil (cooking foil) 10cm x 10cm.
4 sheets of thin cardboard 15cm x 20cm   (cut A4 sheets in half).

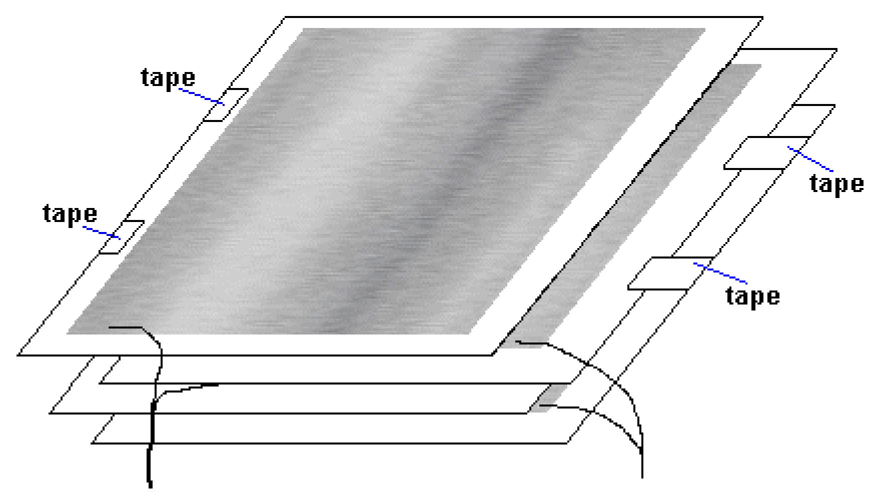

HOME-MADE CAPACITOR

Tape a sheet of aluminium foil to each sheet of cardboard with sticky-tape around all 4 sides. Take one strand of wire from a length of hook-up flex and sticky-tape the end to each sheet of aluminium to make good contact. Place 2 sheets on top of each other and move the top sheet slightly to the left and sticky-tape the edge so they don't move. Do this with the other two sheets but move the top sheet to the right. Now interleave the sets. Connect the wire from the first sheet to the third sheet. Connect the wire from the second sheet to the fourth sheet.
The cardboard (or paper) between the aluminium sheets increases the capacitance three times. The capacitance decreases when the sheets are moved apart and the capacitance increases when the sheets are moved in. The capacitance also INCREASES when the sheets are squashed together such as when a book is placed on them.
You can also make a smaller capacitor by making each sheet smaller and using 6 sheets. You can then add a 100p or 220p in parallel with the home-made capacitor, to select the lower part of the band.

EACH CIRCUIT

Each circuit we describe in the following set of circuits is an improvement or advancement on the previous. We also offer a number of different types of aerial coils, amplifying stages and earphones. Some of the circuits use easy-to-obtain components and home-made equivalents for hard-to-get items. There will be something in this section for everyone to build.

In all radio circuits you will encounter **TWO MAIN PROBLEMS**:

If the FRONT END (the Coil and Capacitor) is loaded too much by the "pick-off" of the amplifying stages, you will only get one station.
If you get squealing or "motor-boating," try a different circuit and layout as the components you are using, plus the voltage of the battery, will need changing.
You cannot always increase the voltage of the supply and get a louder output. Sometimes the increased voltage will stop the circuit working or it may introduce too much gain that the circuit starts to squeal.
The Radio IC (ZN414) DOES NOT WORK on a voltage above 1.5v and some of the transistor circuits completely stop working with a higher voltage. This has to do with the biasing arrangements and if the circuit is designed for a low voltage, you need to keep to the suggested voltage and experiment with a slight increase in voltage and see what happens.
Building a radio is not easy as the enormous amount of amplification of the combined stages creates a feedback loop via the power rail that sets the circuit into oscillation. This effect gets worse with a higher supply voltage and we will explain this further with each of the circuits.

MAKING A CRYSTAL SET

You can buy a **CRYSTAL SET** kit (see the photo of the kit, above) or the individual components (a kit is the cheapest) or use the replacement for the **FERRITE ANTENNA COIL** (16 turns to 20 turns on a 150mm biscuit tin) and/or the **TUNING CAPACITOR** made from aluminium foil and cardboard sheets.
You will need an outside antenna and an earth (such as a water tap or the frame of your soldering iron) to pick up the radio stations.
If you cannot put up an outside antenna, you will need to add one or more amplifying stages and this will allow you to reduce the length of the antenna and increase the volume of the audio.

ADDING AMPLIFYING STAGES TO A CRYSTAL SET

You can add two different types of amplifying stages to a crystal set.
You can connect amplifying stage(s) to the FRONT END and these will be designed to put less load on the front end so the sensitivity and selectivity increases. These stages work at the frequency of the radio signal and they are called RF STAGES (Radio Frequency Stages).
You can build these stages out of individual components or use a chip called a RADIO CHIP or RADIO IC (integrated circuit) for less than $2.00.
The chip contains 5 stages of amplification and these are RF stages (or RF AMPLIFYING STAGES) and the concept is called TRF. (Tuned Radio Frequency).
It is not easy to get this type of amplifier working because the stages produce a very high overall gain and you get a lot of "motor-boating" and squealing if the gain is not controlled. The gain must be reduced when a strong signal is being passed through the circuit because a strong signal will produce a large output and this will be so large that some of the waveform will find its way to the front of the amplifier via the power rail and start to be amplified again. To prevent the output getting too large, the circuit has a negative feedback line - called the AGC line - Automatic Gain Control.
It would be very difficult to reproduce these 5 stages of amplification with discreet components and that's why it is best to use an IC.

The next stage is a DIODE to convert the RF (Radio Frequency) to AF (Audio Frequency). This can be done with the diode-characteristics of a base-emitter junction in a transistor and we will show the alternatives.

Any stages after the diode are AUDIO STAGES or AUDIO AMPLIFIER STAGES.
The main job of the AUDIO AMPLIFIER is to increase the DRIVE CAPABILITY.
In other words, increase the current capability of the circuit for an 8 ohm speaker or 8ohm earpiece (or 16 or 32 ohm).
This is a very difficult thing to do and requires at least 2 stages.
The LOAD you can put on a Crystal Set must be 10,000 ohms or higher. (if you put a lower resistance (impedance) on the output, you will load the FRONT END and reduce its ability to separate the stations.
That's why a crystal earpiece is normally used with a crystal set. It puts almost NO LOAD on the circuit.
If you put a load on the circuit the result will be only one or two stations across the whole dial and only the most powerful station will be received.

If you don't have a crystal earpiece, you will have to use an 8 ohm earpiece. This will require an IMPEDANCE CONVERTING CIRCUIT of 1,250:1
This is a simple way of saying we want the 8 ohm earpiece to appear as 10,000 ohms to the

crystal set.
To produce an overall gain of 1250, we need two stages of amplification.
If a transistor has a gain of 70, it will it will produce an impedance conversion of 70 times. This is a realistic value. Transistors with a gain of 200 will have a gain of about 70 when fitted to a circuit.
This means the other transistor needs to have a gain of about 20 and that is easy to achieve.

ADDING AMPLIFYING STAGES TO THE FRONT OF A CRYSTAL SET

Adding stages to the front of a crystal set are called RF STAGES (Radio Frequency Stages) because they amplify the RADIO STATION SIGNAL.
It does not matter if you amplify RF signals or AF signals. The result is the same.
The only difference is this: The frequency of RF signals is much higher (1,000 times higher) and the coupling capacitors can be much smaller.
This allows an RF amplifier to be built into an IC - called a Radio Chip.
One of the most popular Radio IC's is ZN414 or YS414. This chip has been copied by other manufacturers as: MK484, TA7642 and LMF501T.
All the chips are the same but the pinout is different.
These chips work on a 1.5v supply and if the voltage is increased above 1.5v, the gain of the stages increases to a point of total distortion.
To prevent strong signals producing distortion on 1.5v supply, the output is passed back to the input via a 150k resistor. This feedback line is called the AGC (Automatic Gain Control).
The chip contains 5 stages of amplification plus a stage that converts the RF signal to AF (Detector Stage). This means the signal diode in a Crystal Set is not needed.

Here is the circuit of the TA7642 Radio Chip. It performs the same as the ZN414 Radio Chip.

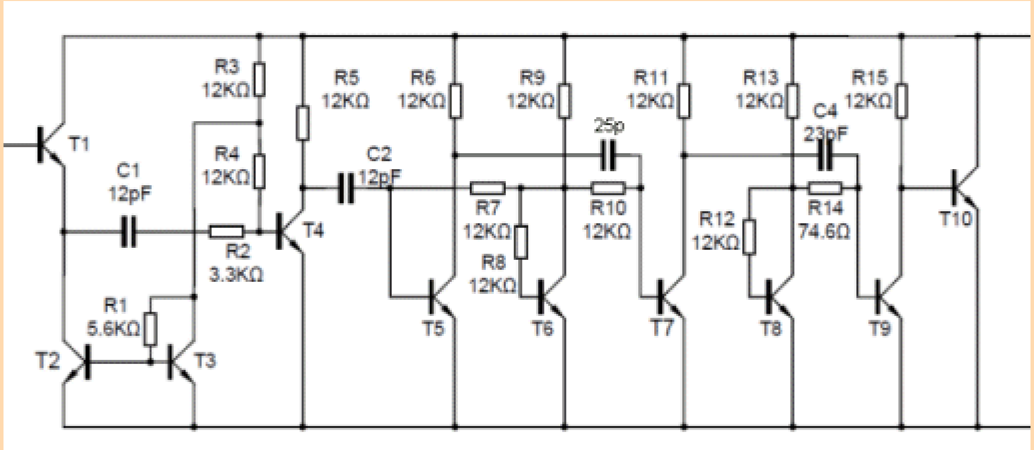
R3
12KΩ
R5
12KΩ
R6
12KΩ
R9
12KΩ
R11
12KΩ
R13
12KΩ
R15
12KΩ
C4
23pF
25p
T1
C1
12pF
R4
12KΩ
C2
12pF
R7
12KΩ
R10
12KΩ
R12
12KΩ
R14
74.6Ω
T10
T4
R2
3.3KΩ
R8
12KΩ
R1
5.6KΩ
T2
T3
T5
T6
T7
T8
T9

Here is the BLOCK DIAGRAM of the ZN414 Radio Chip:

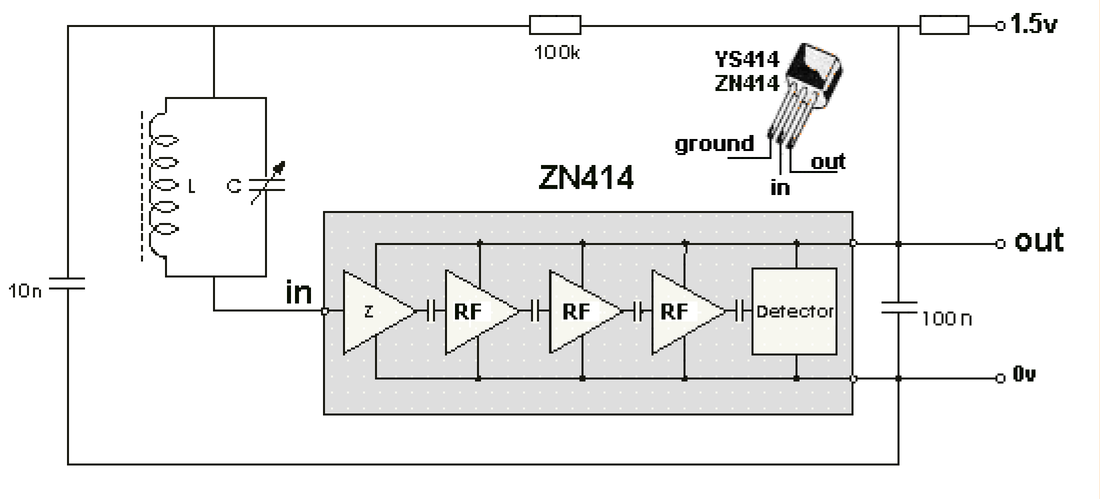
100k
1.5v
YS414
ZN414
ground
out
in
ZN414
L
C
10n
in
Z
RF
RF
RF
Detector
out
100n
0v

The ZN414 chip can be purchased from Talking Electronics for $1.00 plus postage

USING THE ZN414 RADIO IC

By using the ZN414 radio IC (or any if the equivalents) you can create a POCKET RADIO to drive a headphone or speaker.
But it is not easy to use the chip. The main problem is receiving the strong signals without producing distortion and then being able to pick up the weak stations.
A fixed 100k feedback resistor does not provide adequate control and a TRF radio has limited capabilities.
That's why radio manufacturers make SUPERHETRODYNE receivers. Even though they are more complex, the result is far superior.
However a simple TRF set can be made with the Radio IC and a few stages of audio amplification.

The following circuit uses just the Radio IC and a crystal earpiece or the home-made earpiece described above:
You can use a home-made FRAME ANTENNA or a home-made FERRITE ROD ANTENNA and a home-made VARIABLE CAPACITOR.

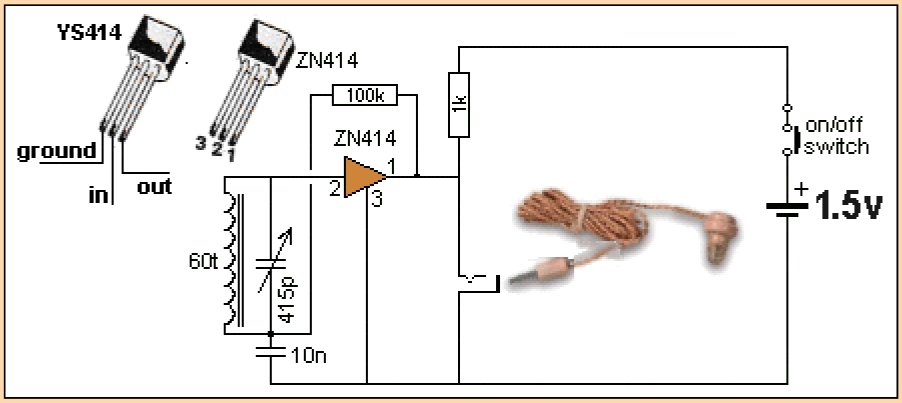

The Simplest ZN414 Radio

Connecting the ground (0v rail) to the frame of your soldering iron or a water tap will increase the output volume. The circuit above shows a Crystal Earpiece. Using a Crystal Earpiece may require adding a 10n across the earpiece to improve the output volume. The substitute Piezo Earpiece is effectively a 20n capacitor and an additional capacitor is not needed.

2 TRANSISTOR RADIO

Here is a simple 2-Transistor radio.
The secret to its performance is the 7 turn "pick-off" from the FRONT END (the TUNED CIRCUIT).
The ratio of 7 turns to 60 turns means a small percentage of the voltage generated in the tuned circuit is passed to the transistor. Thus it puts a small load on the TUNED CIRCUIT.
I don't want to go into any mathematics. The turns ratio is 60:7 = 8 but the effect of the 7 turns "pick-off" has an effect called the IMPEDANCE EFFECT and this is the SQUARE OF THE TURNS RATIO. Thus the IMPEDANCE EFFECT is 8 x 8 = 64. This means the "pick-off" (the LOADING EFFECT) is just a few percent. The front end can produce voltages as high as 500mV because a crystal set can produce a voltage high enough to pass through a diode (350mV) and have sufficient to drive a crystal earpiece.
Even though the front end has a "step-down" ratio, the voltage out the 7 turns will be sufficient to drive the first transistor.
The "transformer" does 2 things: It reduces the loading on the tuned circuit ENORMOUSLY and it produces an output with a higher current than is circuiting in the front end. Even though the transistor is turned ON and biased by the 33k, it is classified as a low-impedance load as far as the front end is concerned and the input signal has to be accompanied by a certain amount of current, otherwise the transistor will not respond to the voltage. The 7-turn "pick-off" is able to provide this current.
Both transistors are biased ON via the 33k base-bias resistors and thus the first transistor responds to the slightest millivolt signal.
This circuit was tested and had the same performance as the **Simplest ZN414 Radio Circuit** above. It can be operated on 1.5v to 6v and the strongest stations tend to overload on 6v. A short antenna is needed.

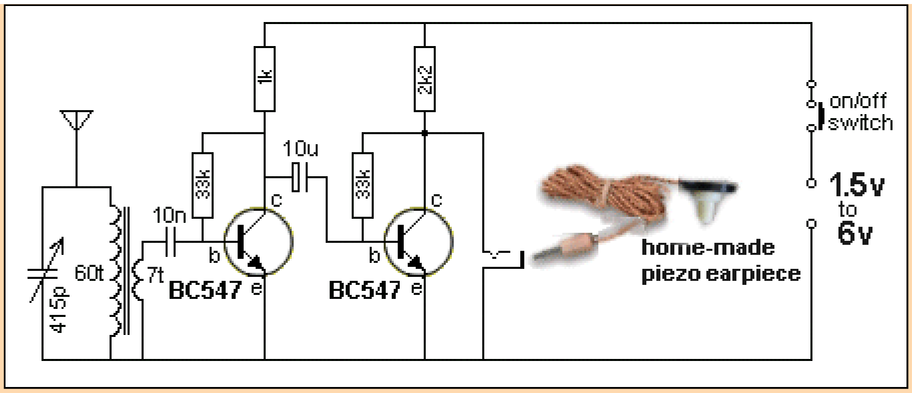
1k
2k2
on/off switch
10u
33k
33k
c
c
1.5v to 6v
10n
b
b
home-made piezo earpiece
60t
7t
BC547 e
BC547 e
415p

SIMPLEST 2-TRANSISTOR RADIO
using a very-high-impedance earpiece

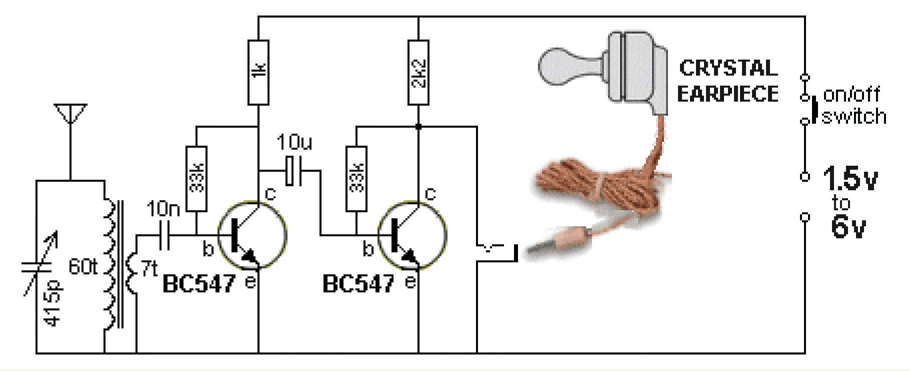
1k
2k2
CRYSTAL EARPIECE
on/off switch
10u
33k
33k
c
c
1.5v to 6v
10n
b
b
60t
7t
BC547 e
BC547 e
415p

SIMPLEST 2-TRANSISTOR RADIO
using a crystal earpiece

ADDING AN IMPEDANCE MATCHING STAGE

You can add an IMPEDANCE MATCHING STAGE to the output of the circuit above so a low-impedance earpiece can be used.
We call it an IMPEDANCE MATCHING STAGE because this is the correct technical term. It is an AMPLIFYING stage but it amplifies the CURRENT because the second transistor cannot drive an 8 ohm LOAD. 8 ohms is a very low resistance and if it is connected directly to the second transistor, the output will be almost zero.
The reason for this is covered in our discussion: The Transistor Amplifier.
This stage will not increase the volume but simply match the 8 ohm load to the circuit above.
It is very difficult to connect a LOAD to this type of circuit because it will take more current from the battery and cause the supply voltage to fluctuate. These fluctuations will be passed to the first stage and cause variations in the signal. This will be amplified by the first and second transistors in the form of a low-frequency buzzing called **MOTOR-BOATING**.
The only way to reduce or remove this noise is to add an electrolytic across the power rails and reduce the supply voltage. The third transistor simply takes the waveform on the output of the second transistor and delivers it to the earphone with a higher current. It is called an IMPEDANCE MATCHING STAGE as it effectively increases the 8 ohm load by a factor of about 100.

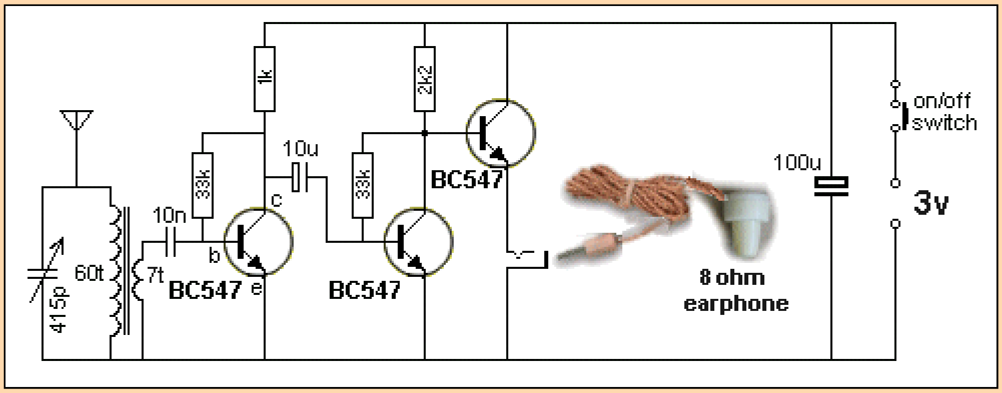
1k
2k2
on/off switch
10u
BC547
100u
3v
33k
33k
c
10n
b
60t
7t
BC547 e
BC547
8 ohm earphone
415p

3-TRANSISTOR RADIO with buffer stage for 8 ohm earpiece

The 3rd transistor converts the 8R to about 800R

You can use 16 ohm, 32 ohm or 64 ohm in place of the 8R earpiece and these will give better performance as they will take less current and improve the stability of the circuit.
Low-impedance earphones create "motor-boating" due to the peaks of current and this can be very hard to fix.

A 2-TRANSISTOR RADIO with REGENERATION

The next stage in our discussion to get better performance is a feature called **REGENERATION**. Regeneration sends a small output signal back to a previous stage in the form of POSITIVE FEEDBACK to INCREASE the original signal. The signal on the emitter of the first transistor is the same amplitude as the signal entering the base but the FRAME ANTENNA has a turns ratio of 5:15 and this increases the signal on the receiving section of the antenna by up to 3 times. But we want the returning signal to be just above the amplitude of the receiving signal and so a resistive adjustment (attenuator) is provided on the emitter to deliver just the right amplitude.
This has the effect of increasing the amplitude on the TUNED CIRCUIT and is just like reducing the load on the circuit.
As we have mentioned above, when the tuned circuit is lightly loaded, it will pick up a station at the exact frequency of transmission and if the dial is changed slightly, the station will disappear. This quality is called SELECTIVITY.
At the same time, the Tuned Circuit will pick up weak stations and this is called SENSITIVITY.
The quality of a receiver depends on the loading of the TUNED CIRCUIT.

Here is the original circuit from Elektor Magazine with the prototype made on matrix board and fixed to a base-board with a frame antenna made from two sticks of wood. The photo shows a speaker but the output is so low that you really need headphones.

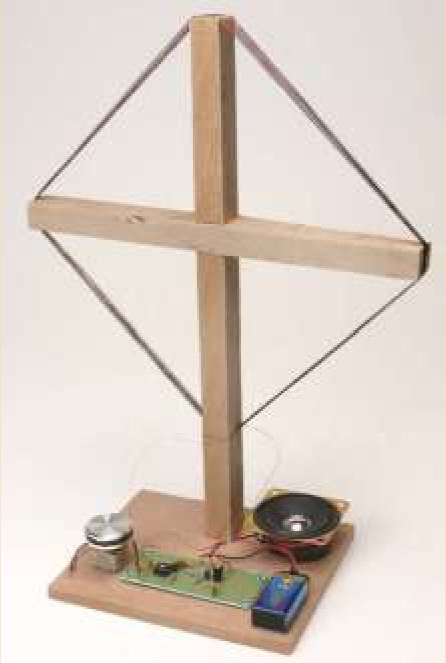

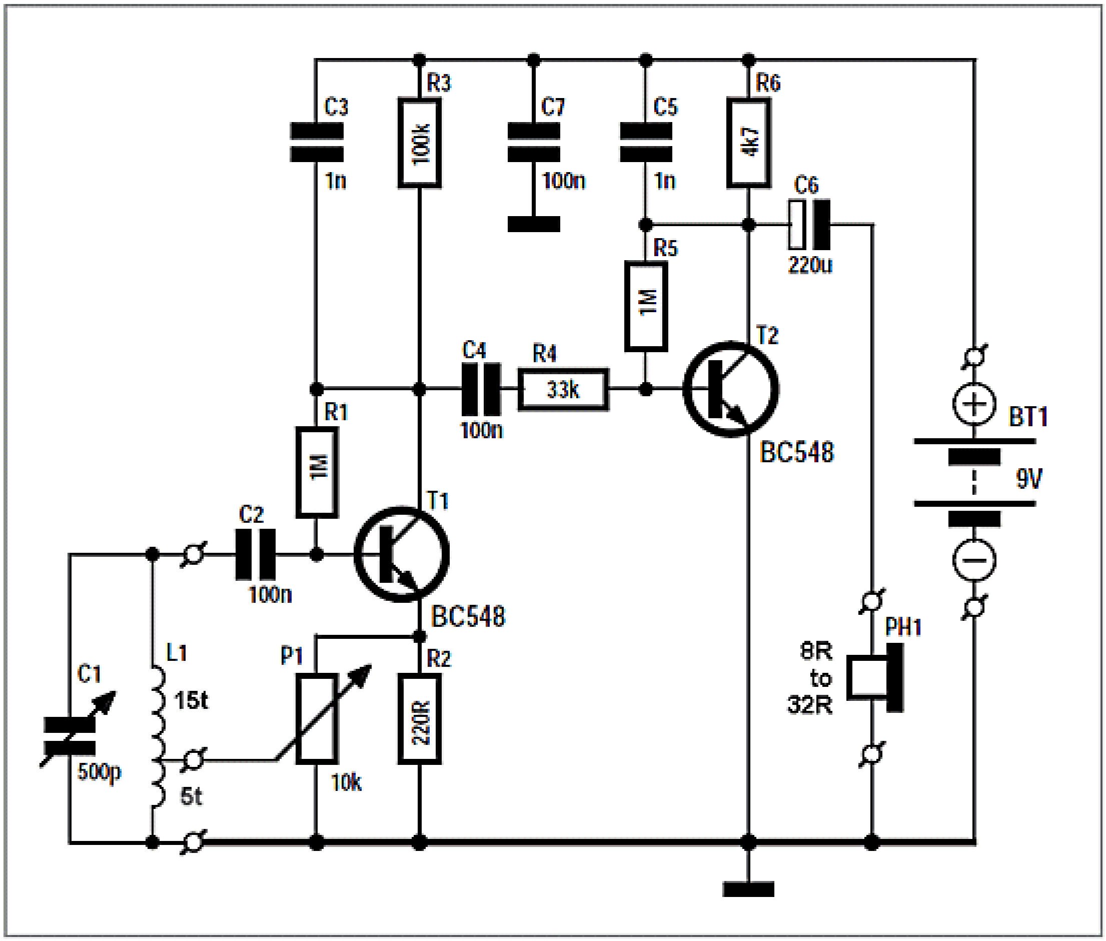
C3
1n
R3
100k
C7
100n
C5
1n
R6
4k7
C6
220u
R5
1M
C4
100n
R4
33k
T2
BC548
BT1
9V
R1
1M
C2
100n
T1
BC548
C1
500p
L1
15t
5t
P1
10k
R2
220R
PH1
8R
to
32R

2 TRANSISTOR REGENERATIVE RADIO from Elektor Magazine

The circuit is very complex and the output will be very low as the circuit cannot drive a low-impedance earphone via a 4k7 load resistor. The 4k7 resistor is actually driving the speaker (the transistor is simply discharging the 220u). The 4k7 only allows 32/4700 x 9 = 61mV to appear across the earphone - a very poor result.
The skill of designing a transistor stage is covered in our comprehensive eBook: The Transistor Amplifier and you wont make a mistake like this !!!
The circuit above can be simplified and we can add the REGENERATIVE feature to our **Simplest 2-Transistor Radio** circuit:

Our circuit uses a 15 turn circular FRAME ANTENNA 15cm diameter and a 5 turn REGENERATION coil.

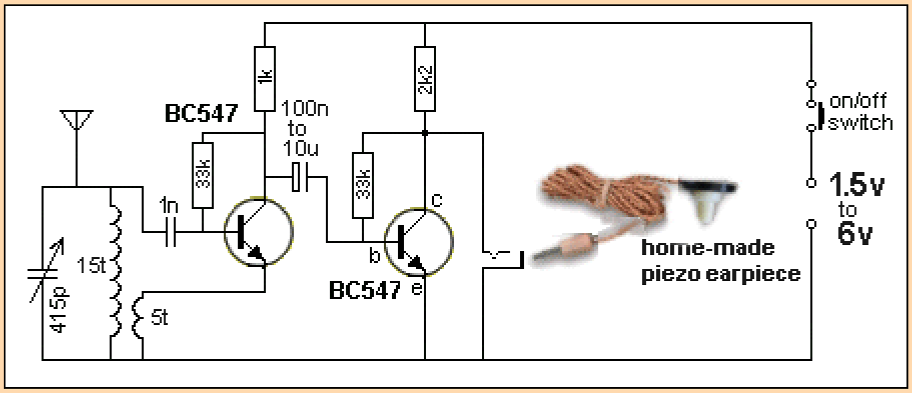
BC547
1k
100n
to
10u
2k2
33k
1n
33k
c
b
e
BC547
415p
15t
5t
on/off
switch
1.5v
to
6v
home-made
piezo earpiece

2 TRANSISTOR RADIO with REGENERATION

The regeneration coil is brought near the main coil and as it gets closer you can hear the audio get louder. If this does not happen, turn the coil around.
Early radios used this technique and the operator had to adjust the coil by hand. No-one minded because radio was a fascination and the simplest radio cost more than a weeks wages. To listen

to a broadcast through headphones was an amazement and listeners would sit all night with headphones listening to music.
This is very fiddly and by adding an extra buffer stage, we can use a Frame Antenna with a very clever "pick-off" that does not load the front end. This gives the circuit very good sensitivity and selectivity without regeneration.

3 TRANSISTOR RADIO

Here is our final design for the simplest self-contained 3-Transistor Radio using our home-made Tuning Capacitor and 250mm x 350mm Frame Antenna. It picks up the local stations and drives a low-impedance earphone or set of earphones (from a mobile phone).

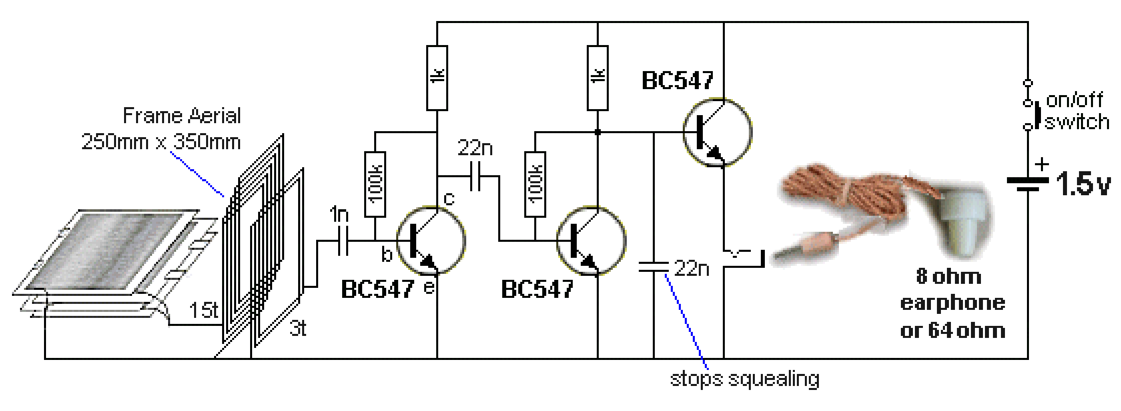

3-TRANSISTOR RADIO

The circuit performs very well and uses readily-available components. The 22n across the output is essential to stop squealing.
The secret to sensitivity and selectivity is the turns-ratio on the Frame Antenna. The 3-turn "pick-off" puts very little load on the front end and this allows the stations to be tuned with our home-made Tuning Capacitor.
The circuit contains all the features we have discussed above and only needs a 1.5v supply.
Build this circuit before you buy any expensive tuning capacitors, IC's or ferrite slab antennas as you will not get any better results.
This is called a TRF circuit and because the stages operate at Radio Frequency or Audio Frequency. Due to the high amount of amplification, the circuit can start to squeal (feedback, motorboat) due to the layout.
You may need to shorten or lengthen the leads or move the parts slightly - it's that critical.
However the result is a portable radio that needs no earth and will pick up the strong stations.
You can try connecting the 0v rail to the metal part of a soldering iron to increase the number of stations.

LOADING

The whole success of picking up a radio station is the RECEIVING CIRCUIT. The receiving circuit is the coil and the signal in the air (from the radio station) must go down the centre of the coil.
It cannot pass over the top or the bottom of the coil. Only the signal that goes down the centre of the coil is received.
As you can see, the centre of the coil is not very big and it is amazing that the signal can pass down the centre. But it does, and that is the only signal that will be amplified.
This signal is passed to the capacitor and we have explained how the signal is gradually increased and increased in amplitude until it is as large as 500mV. The signal from the radio station may be as small as a few millivolts, but as it keeps pushing the "swing" back and forth, the amplitude get larger and larger.
If you put your finger on the "swing" you will prevent it get larger and larger and it only requires the slightest touch of your finger to prevent the swing gaining full amplitude.
In electronic terms, your finger is called LOADING THE CIRCUIT and since we have to pass the signal to further stages of amplification, we need to "tap" or "load" or "pick-off" a signal.
The aim is to load the circuit as least as possible because the actual energy entering the circuit is very small.
In fact, this is all the energy we can remove as that is all the energy entering it.
Because a very small amount of energy is entering the "front-end" we classify it having a very

high impedance. It is very difficult to provide a value of impedance for this circuit because impedance has the term "Z" and the circuit is operating a very high frequency so resistance values are not the same as impedance values.
The actual resistance of the circuit is ONE OHM but the impedance is more like 10,000 ohms to 100,000 ohms.
We can explain its high impedance if we put a 100,000 ohm resistor across the circuit. The waveform will be reduced very slightly. If we put a 10,000 ohm resistor across the circuit, the signal will be reduced a reasonably large amount. If we put 1,000 ohms across the circuit it will stop working.
This means a load of 100,000 ohms will have the least effect.
In a crystal set, the diode creates NO LOAD until a voltage of 350mV is reached. It then passes excess voltage to a crystal earpiece that has a very high impedance. That's why a crystal set will produce a good output. The LOADING is very small.
When a transistor is connected to the TUNED CIRCUIT, it starts to put a load on the circuit after 600mV and this load is VERY HIGH. The "resistance" of the base-emitter junction is about 1k and the signal will find it very difficult to rise above 600mV because the incoming energy is not sufficient to increase the voltage.
Adding a capacitor between the base and the front end allows the transistor to be self-biased and get a turn-on voltage of about 600mV from a base-bias resistor.
The FRONT END is now separated from the transistor and ANY voltage it is producing will be passed to the transistor via the capacitor.
Whereas, with the crystal set, the first 350mV could be produced without any loading, the circuit is now loaded AT ALL TIMES.
This means we have to load the circuit as lightly as possible to be able to pick up individual stations.
The only way we can do this is to use a capacitor of the smallest practical value and this has to be worked out by trying different values. If the value is too small, the transistor will not detect a small signal. If the value is too large, the circuit will stop working.
Values such as 1n, 10n and 100n are suitable.
Values such as 1u, or 10u will be too large.
The CRYSTAL SET loading and a transistor load are completely different.
The transistor loads the front end ALL THE TIME and that's why you need to use a transformer or other ways to reduce the loading. Sometimes a Field Effect Transistor is used as it puts almost no load on the front end.

2-4-2016

Histats.com
Vis. today
26

FLIR
12 THINGS TO KNOW BEFORE BUYING A THERMAL CAMERA
GET THE GUIDEBOOK NOW!

Quickturn PCB & Assembly
Your one-stop source for all types of PCBs and Full Turnkey PCB Assembly.
rushpcb.com

Quickturn PCB & Assembly
Your one-stop source for all types of PCBs and Full Turnkey PCB Assembly.
rushpcb.com

Unlimited HDTV is Here

TV Bandit

Companies already tried to ban this antenna but failed. Get yours before it's too late!

Recreating an Alfred P. Morgan crystal set

Analog Dial | Page 1 | Page 2 | Page 3 | Page 4

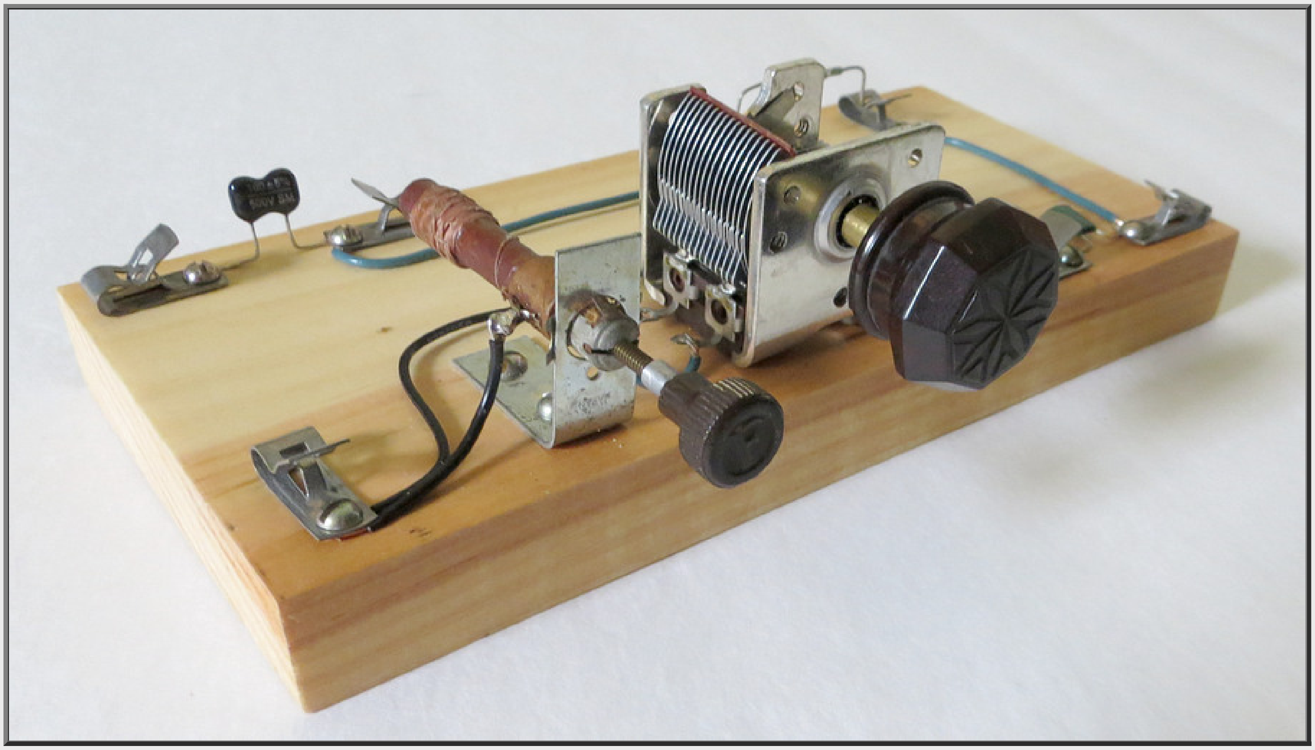

A recreation of my first (working) crystal set, originally built in 1966.

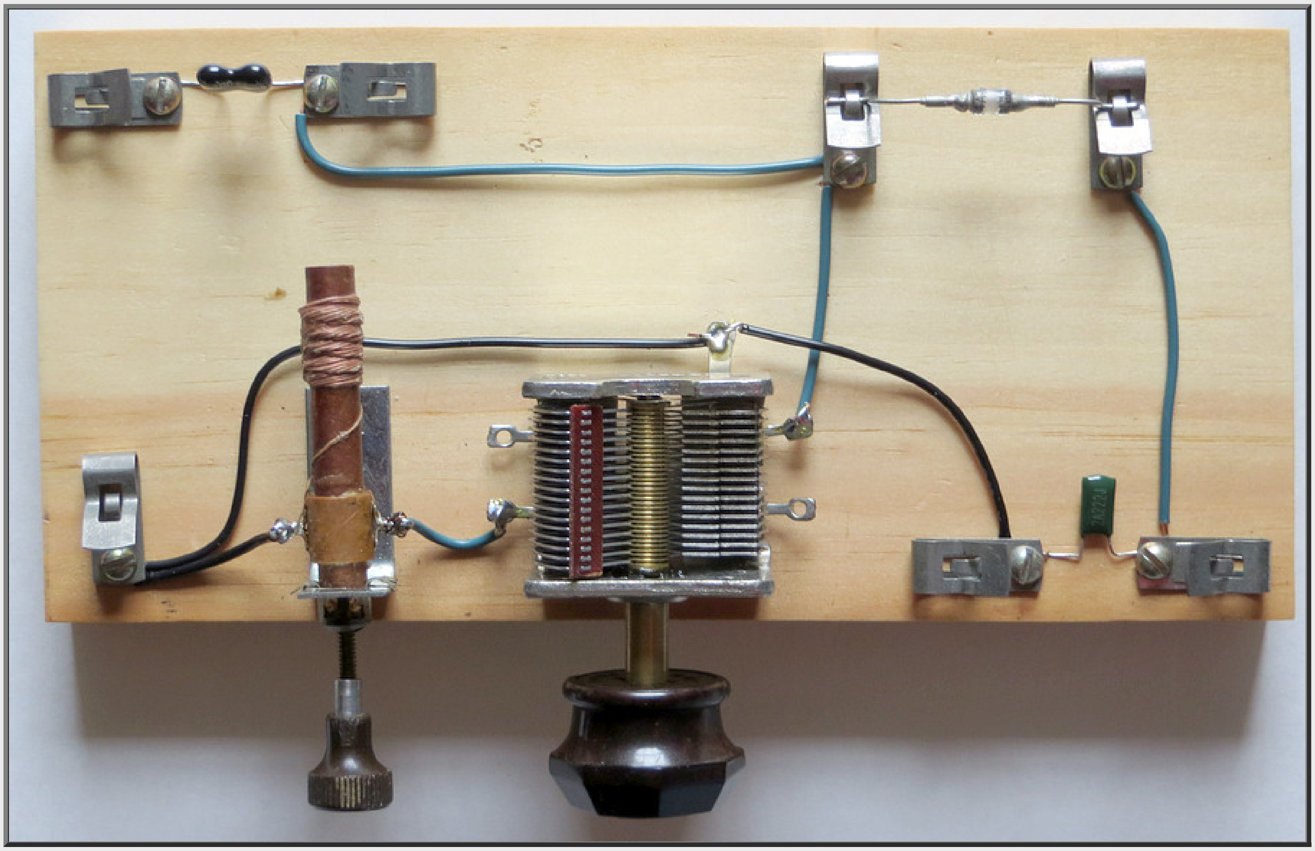

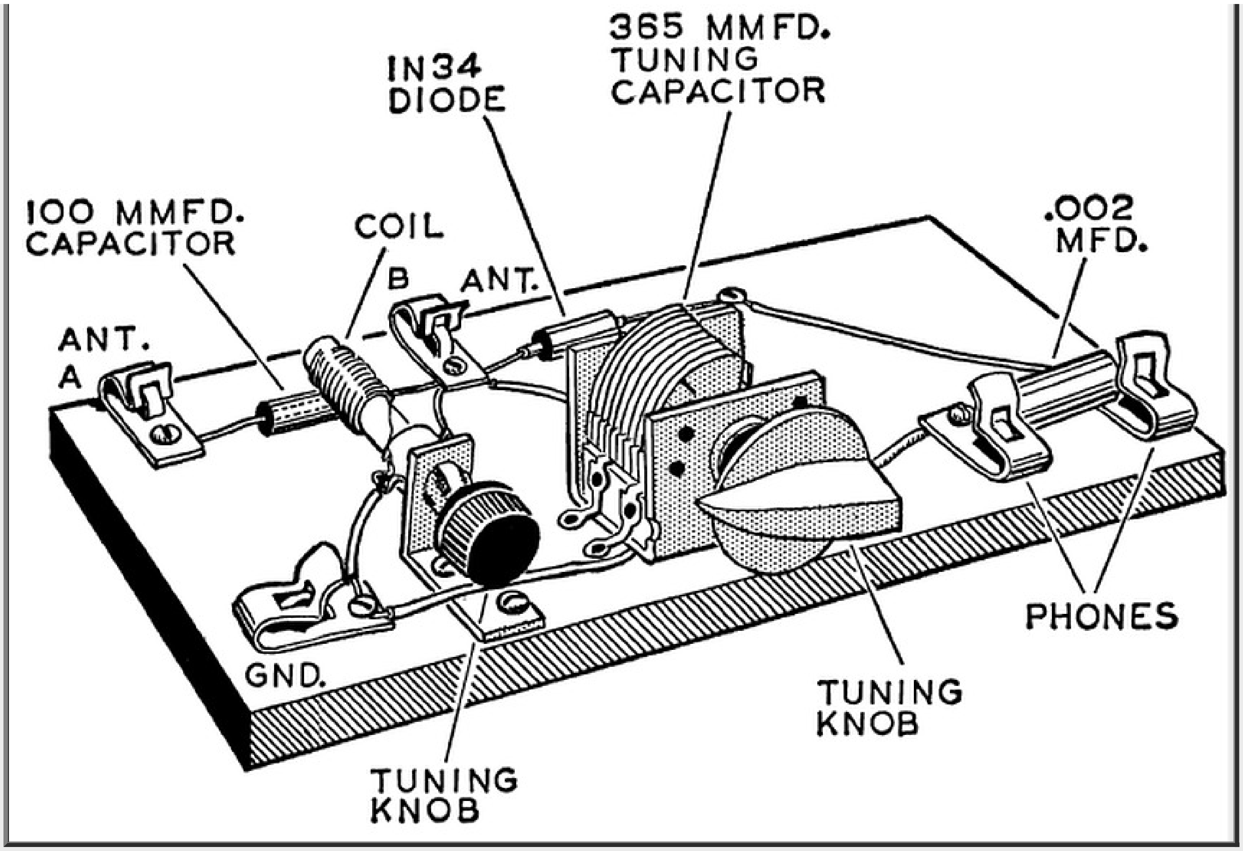
365 MMFD.
TUNING
CAPACITOR
IN34
DIODE
100 MMFD.
CAPACITOR
COIL
B
ANT.
.002
MFD.
ANT.
A
PHONES
GND.
TUNING
KNOB
TUNING
KNOB

Page from "The Boys' Third Book of Radio and Electronics" by Alfred P. Morgan.

Back Story:

In the Summer of 1966 I was 10 1/2 years old and one of my favorite TV shows was "Get Smart". It was a show about a bumbling secret agent named Maxwell Smart who had all sorts of gadgets at his disposal. My friend Billy Meyers and I decided we wanted to be secret agents, like Max. One of the "secret agent" things we'd do was to pick out a guy with a briefcase walking home from work and declare him a Soviet spy. Then we would "tail" him for a few blocks, making up stories about him.

Billy and I had no way to communicate with each other after we had to come in for the night. He lived half a block away. Kids didn't use the phone back then. There was only one phone in the house and I don't remember using it before I was about 14 years old. Maxwell Smart had a phone in the sole of his shoe. We needed something like that!

Maxwell Smart (Don Adams) Agent 86.

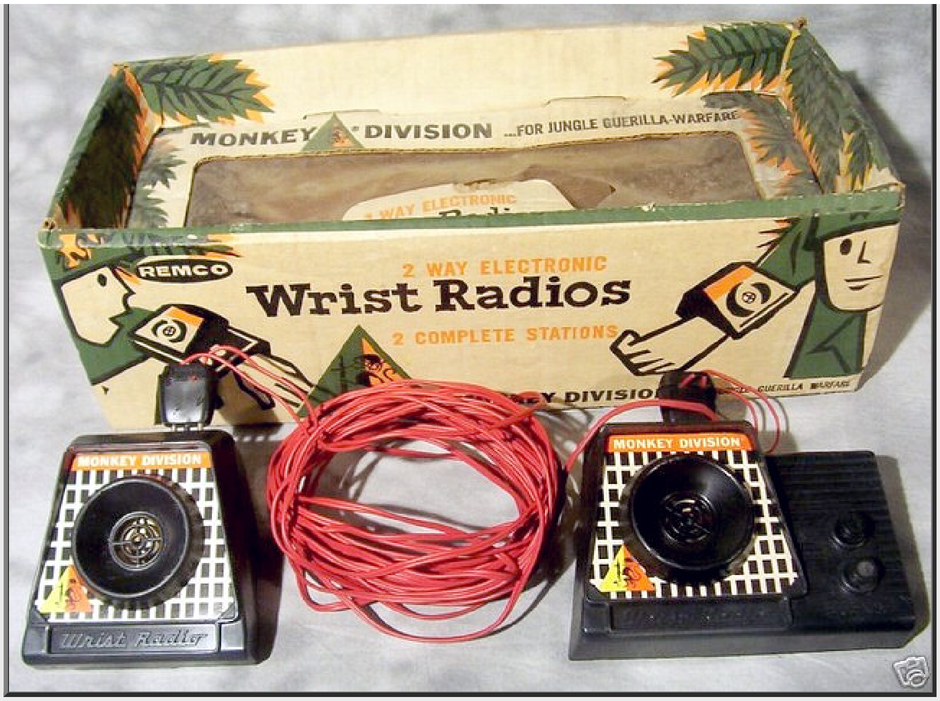
MONKEY DIVISION ...FOR JUNGLE GUERILLA-WARFARE
REMCO
2 WAY ELECTRONIC
Wrist Radios
2 COMPLETE STATIONS
MONKEY DIVISION
MONKEY DIVISION
Wrist Radio

Remco "Monkey Division" Wrist Radios.

One thing I DID have was a set of Monkey Division Wrist Radios. These were powered by a single C battery in the "Master" unit, which also had a button on it that would buzz the other receiver to get the users attention (apparently, the other user was unworthy of a button). A metallic speaker doubled as a microphone. The sound was very tinny, but you could make it out. Unfortunately, they were wired to each other. There was nothing "radio" about them.

I remember taking them outside with my brother Rob, and I could see and hear him talking at the same time his voice was coming over the wrist radio. They were pretty much useless, unless you like running around with a wire connecting you to your brother.

BUT... with some extra wire strung across the driveway, down the backs of the houses and into Billy's bedroom, Billy and I would be able to talk to each other! I immediately presented this great idea to my mom with a request the she fund the cost of the wire, and she immediately refused and told me the electric company would just come out and take the wire down.

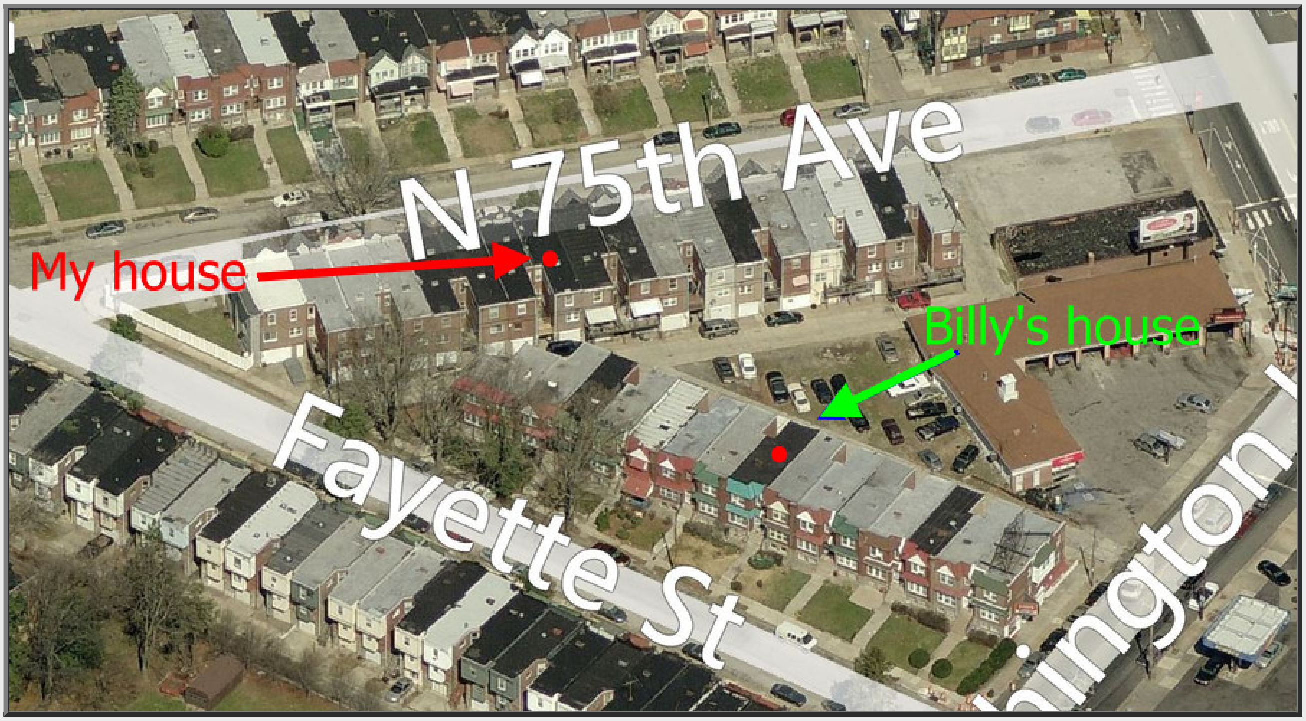
N 75th Ave
My house
Billy's house
Fayette St
ington

Location of my house and Billy's house (actually his grandparent's house) The wire would have to be strung across the driveway. This is part of the West Oak Lane section of Philadelphia.

There was only one thing left to do. Build a radio. I went down the basement and connected a battery to a speaker and used a coat hanger as an antenna. This was similar to the wrist radio, but the coat hanger antenna replaced the wire. It didn't work! All it did was make clicking sounds in the speaker. I had to wait for my dad to come home from work and ask him why.

When my dad came home I showed him the setup and he said, "You don't have a detector." I asked him what a detector was and he told me to go the library and get a book on radio. The next day I had a copy of "**The Boys' Second Book of Radio and Electronics**" by Alfred P. Morgan. Chapter 2, page 15 was titled "Building Your First Radio Receiver". I wouldn't be able to talk to Billy with it, but that was OK. My mom wouldn't buy me a secret agent coat, we didn't have any gadgets, and our secret agent days were coming to an end.

Now there was another problem. None of the parts needed to build anything in the book could be found around my dad's workbench. There was a store on Ogontz Avenue named REE Electronics, so I headed up there with a list of parts. The store sold stereo equipment and fortunately for me, also repaired stereo equipment. I asked the man in the store if he sold diodes or "capacitaters" and he sent me into the back of the place. There were two guys back there and bins of parts along the wall.

S.U.R.F.A.C.E.
HAIR CULTURE
S.U.R.F.A.C.E. CLOTHING CO.
NY NAILS
FedEx

REE Electronics was located at 7709 - 7711 Ogontz Avenue in Philadelphia. The entire block has been razed and rebuilt, and is no longer recognizable. The picture above is the 7900 block of Ogontz Avenue. The store on the left is the only one that retains its original appearance, with the glass store window and the apartment overhead. This is how REE Electronics looked in 1966.

The two guys were pretty cool. I announced that I would like a "three hundred and sixty five micro micro farad variable capacitater". They asked me a couple of questions and told me to come back with the book. They had all the parts I needed except the coil. No problem, I would just build the set with no coil. I came home with Fahnestock clips, a 1N34 Germanium diode, a variable capacitor and a crystal earplug.

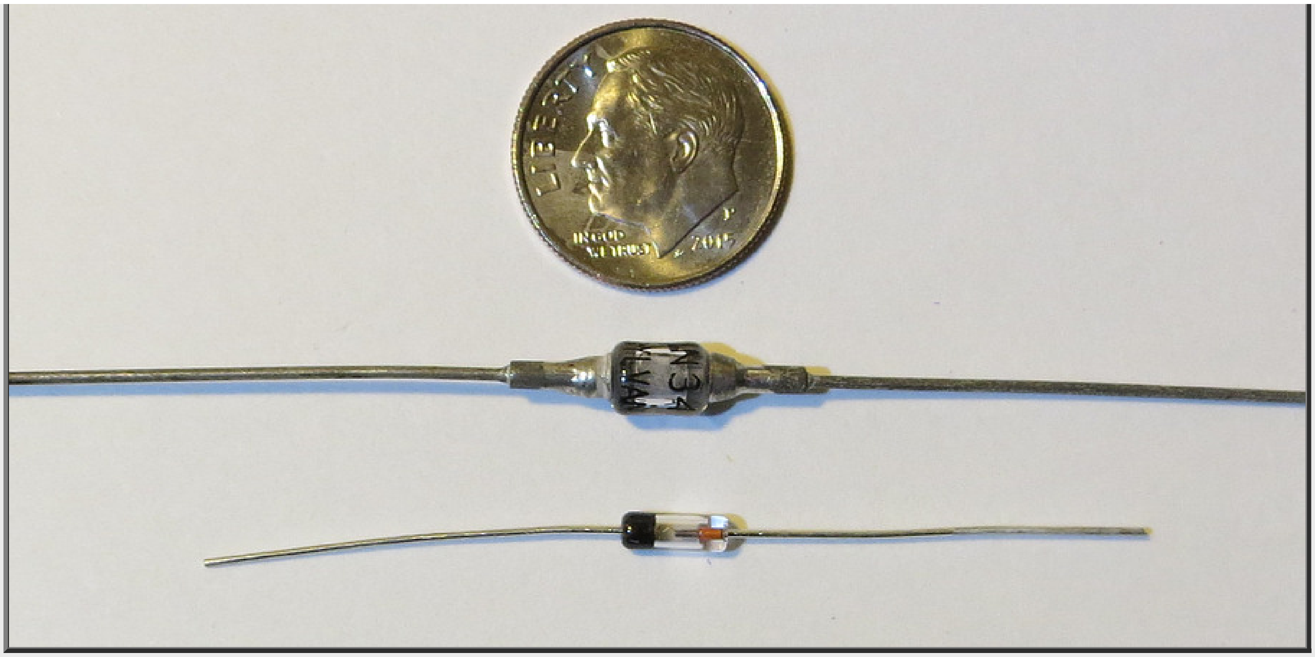

A Sylvania 1N34A Germanium diode (top) from 1949 and its modern counterpart. The new diodes are literally a dime a dozen, less than one cent each. The 1949 Sylvania diode cost me $9.00 in 2015. It is a duplicate of the one I bought at REE Electronics in 1966.

The diode was 65 cents in 1966. I can't use the $9.00 diode because I don't want to bend the leads, so I sort of just look at it.

Of course the radio didn't work without a coil. It did pick up the slightest whisper of KYW AM 1060 mingled with WIBG AM 990. There seemed to be some buzzing associated with it, as what I could hear sounded distorted. I HEARD something, that was the really, really neat part. It made such an impression on me that I remember the date. July 26, 1966.

The "problem" with the Alfred P. Morgan books was that they were not written for anybody as dumb as I was. Morgan didn't write, "If you can't find the coil you can make one." He just said to go buy a coil. Not only that, but there were no photographs in the book, though there were excellent drawings on almost every page. Since I had never seen some of the parts in real life, I didn't know exactly what the coil looked like, and I didn't understand what it did. That's because I didn't read the book! I was stuck on page 15.

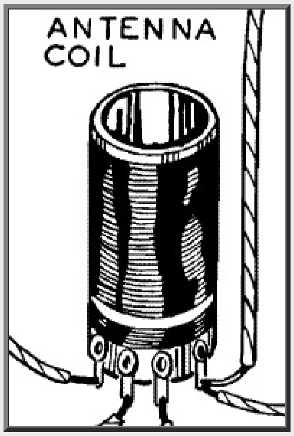

This is the picture of the coil from the book. I didn't know what I was looking at.

I returned the book to the library and came home with "**The Boys' Third Book of Radio and Electronics**." I found a simple radio on page 104 and soon headed back to REE Electronics. This time, they DID have the coil! I can't remember how much time passed after the first non-functioning radio was built. Probably a month or so.

I asked my dad to cut me a wooden base for the radio. I started building the radio and, if I remember correctly, it took me a long time. I didn't have a drill, so any holes in the base were made with the point of a compass. There were three connections that needed soldering. I got some solder from the basement, and the tweezers and alcohol lamp from my chemistry set. The tweezers were heated in the flame of the lamp till the tips began to glow, then I would quickly solder the joint.

One day a friend from school named Leo Pound stopped by on his bicycle. This was a bit unusual because Leo lived miles away. I don't even know how he knew where I lived. He recently told me (via Facebook) that he remembers helping me build the radio. Odd that it was the one and only time he came by. Apparently, we got the radio working that very day.

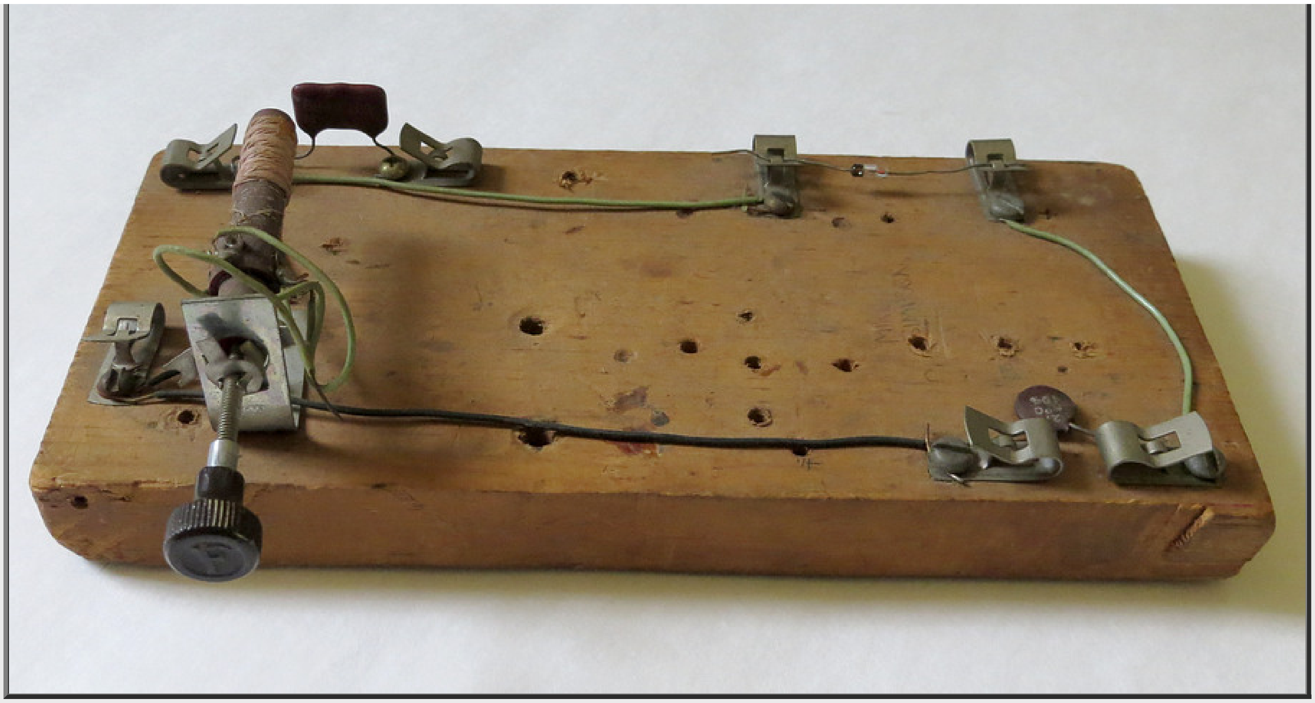

Here is the base of the radio made in 1966.

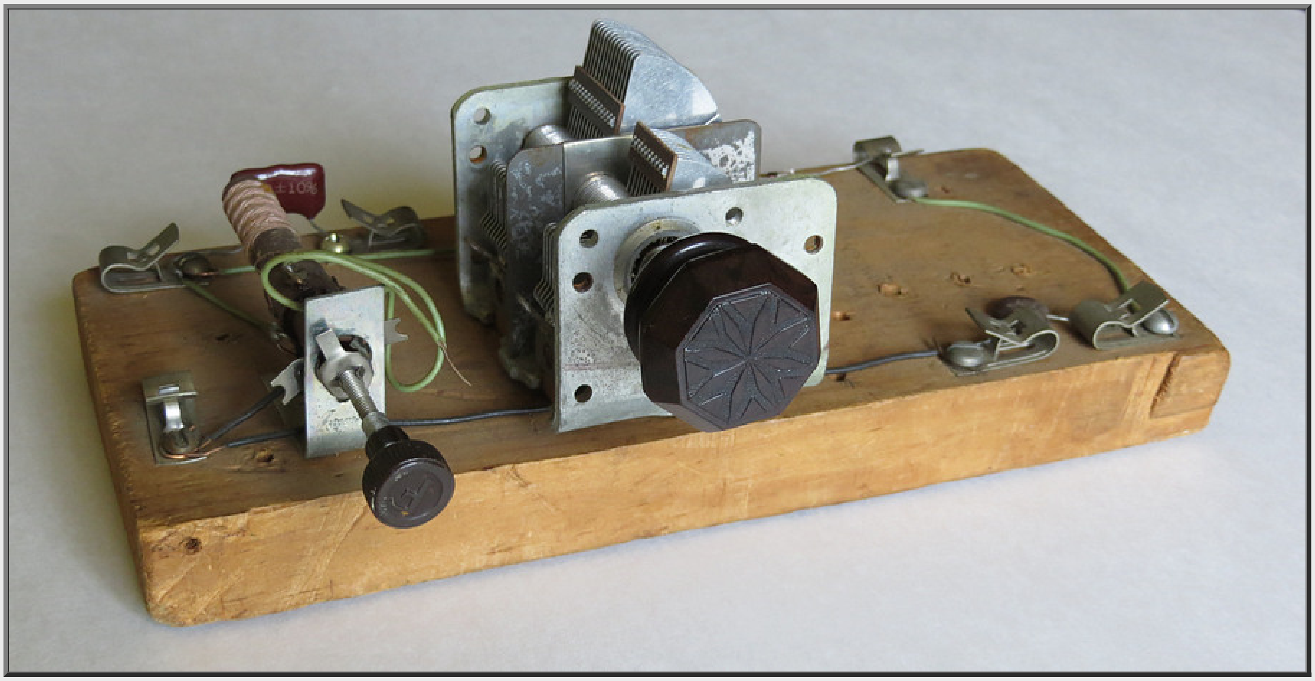

I found all the original parts except for the tuning capacitor. However, it looked like this one. It's just sitting on the base in this photo, but the radio pretty much looked exactly like this. My dad gave me the big tuning knob. The smaller knob came from a lamp in my bedroom. I didn't tell my mom you could no longer turn on the lamp, but eventually I found a lamp in somebody's trash and took the knob off to replace the one on my lamp.

Original coil, diode and main tuning knob from the 1966 radio. The wires on the coil were soldered with a pair of red hot tweezers.

My dad gave me the large knob for the tuning capacitor. He had a second job on the weekends at "John Cusimina's Moving and Storage." I've always wondered if he pulled that knob off of somebody's radio while they were moving. I hope not. He probably did.

Next

MDS975 www.mds975.co.uk

CRYSTAL SETS 5
Experimental Crystal Sets

MDS975 www.mds975.co.uk

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

Crystal Sets (Part1) | Build Your Own Crystal Set (Part 2)

Spider's Web Crystal Set (Part 3) | Crystal Set By Kenneth Rankin (Part 4) | Crystal Radio Links

CRYSTAL SETS 5: EXPERIMENTAL CRYSTAL SETS

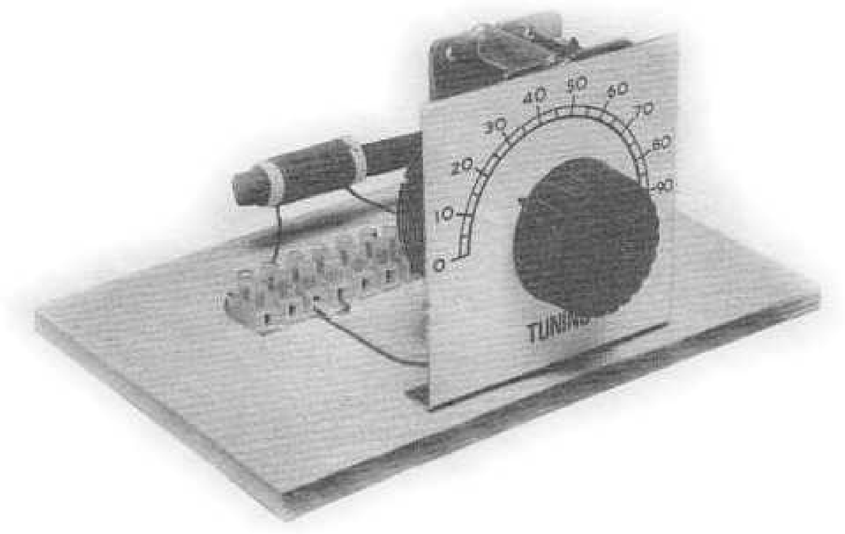

Picture 1 - The Complete Experimental Crystal Set

THE POPULARITY of the crystal radio arises from its simplicity, and the fact that it needs no power supply. The circuit here allows for easy experiments with tuning, aerial and diode coupling, and frequency coverage. Wrong connections can cause no damage to any components.

A Crystal Set is more often than not used for the reception of medium and long wave radio, but short wave reception is also quite feasible. It will normally be possible to receive some of the stronger international radio stations.

This is adapted from an article that appeared in the 1970's in Everyday Electronics, and gave me almost endless hours of fun!

BASIC CIRCUIT

The basic circuit is shown in Picture 2 below. The coil L1 can be air cored, or have a ferrite rod placed in its winding. The variable capacitor C1, in conjunction with aerial-earth capacitance, tunes the circuit to resonate with the wanted radio station frequency. The diode D1 "detects" or demodulates the radio signal so that the programme is heard in the earpiece.

This basic circuit can be modified in various ways to obtain better performance.

EARPHONE

As most constructors will be using a Crystal Earpice to listen to the crystal set it is essential that a 47k Ohm resistor is connected across the earphone terminals (TB1/1 and TB1/2 in the diagram), i.e. in parallel with the earphone, otherwise results will be very quiet.

A High Impedance headset of 20k Ohms (20,000 Ohms) may give even better results, but these are very difficult to obtain , so unless you happen to already own such a headset the Crystal Earphone with 47k resistor will be the only option. An ordinary magnetic earpiece or Walkman headphones will not work with a crystal set.

ASSEMBLY

Construction is of a 'breadboard' type using a wooden board of about 165 x 130 mm. A 12-way block connector, TB1, is used to connected together the components and this is screwed onto the wooden board. The use of a block connector provides an easy method of connecting the components together and then subsequently rearranging them as the experiments progress.

Tuning capacitor C1 is screwed to a bracket made of some scrap metal which is then also screwed firmly down to the baseboard, see Picture 1 above. Thin plywood screwed to the front edge of the baseboard would also provide a suitable method of fixing the tuning capacitor to the base. A knob with pointer is fitted to C1, and a scale is drawn and fitted behind this.

Except for C1, all connections are made by the terminals of the 12-way terminal block as shown in Picture 4. Loosen the screws with a small screwdriver, insert the bared ends of the wires, and tighten the screws. The various locations on the terminal block, TB1, are also shown in the circuit diagram, Picture 2.

AERIAL AND EARTH

Crystal receivers need a long wire aerial preferably strung outside and about 25m long, or as long as is possible to install. If this is outside it should be high and clear of earthed objects as this will improve performance.

An earth is absolutely essential for a crystal set to work properly. The earth lead can be run to an earth rod or spike that is buried to a depth of about 1 meter into damp soil. Or it may be soldered to a bare metal can which is buried in damp soil.

It is feasible, though not recommended, that the earth lead can be connected to the earthing terminal of a hi-fi system or even to the bare metal case of a personal computer that is plugged into an earthed mains outlet, but is switched OFF.

Stranded, insulated wire, or purpose made aerial wire can be used for the aerial and earth leads.

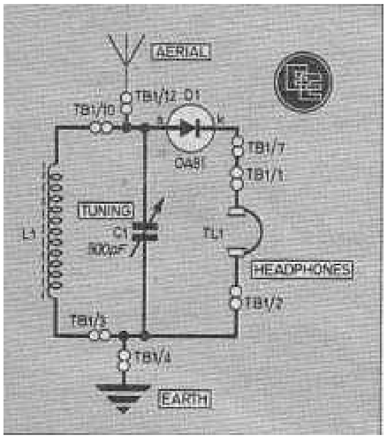

Picture 2 - The Basic Circuit

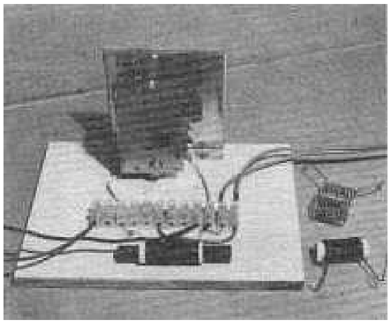

Picture 3 - Photo Of The General Layout

INDUCTORS (The Tuning Coils)

The following four coils are suggested for initial use as L1 :

Coil 1: Make a thin card tube to slide on a 10mm diameter ferrite rod, and on this tube wind about 105 turns of 32 s.w.g. enamelled copper wire, side by side. Secure ends with sticky tape.

Coil 2: Make a similar coil to to coil 1 having about 15 turns of 24 s.w.g. enamelled wire on the card tube. Loops of cotton will help hold the ends in place.

Coil 3: Wind 9 turns of 20 s.w.g. bare tinned copper wire on an object about 20mm in diameter. Remove and stretch to separate the turns, to obtain a coil about 25mm long.

Coil 4: Make a similar coil to coil 3, but with 5 turns.

The Ferrite Rod

It will be necessary to have a ferrite rod of about 60mm to 75mm long available. Coils 1 and 2 will provide reception of medium wave and the longer short wave bands. Coil 3 should cover about 3 - 10MHz shortwave with the ferrite placed in it, or about 6 - 18MHz with the ferrite rod removed. Coil 4 should cover about 6 -13MHz with the rod in, and about 9 - 20MHz without the' rod.

It will be noted that as the ferrite rod is inserted, any particular signal has to be re-tuned by opening Cl. This arises because the ferrite increases the inductance of the winding, so less parallel capacitance is needed for the same resonant frequency.

EFFICIENCY CHECKS

Tune in a m.w. transmission using coil 1 which gives good headphone volume. Place a microammeter or multi-range meter on a sensitive range in series with the headphones. A reading of 50-100uA or more may be obtained, depending on aerial, earth, earphone resistance and resistor value, coil and detector efficiency and strength of signals at your locality.

Placing the ferrite rod in the coil and re-tuning should boost the meter reading to some extent. Surplus or other detector diodes can be tried by substituting them in turn and noting the meter reading. Improvements to the aerial (or earth) will also show up as a rise in meter reading.

If experimenting with a crystal earpiece, which gives no direct current circuit, the meter may be clipped across the phone leads, i.e. D1 cathode to earth.

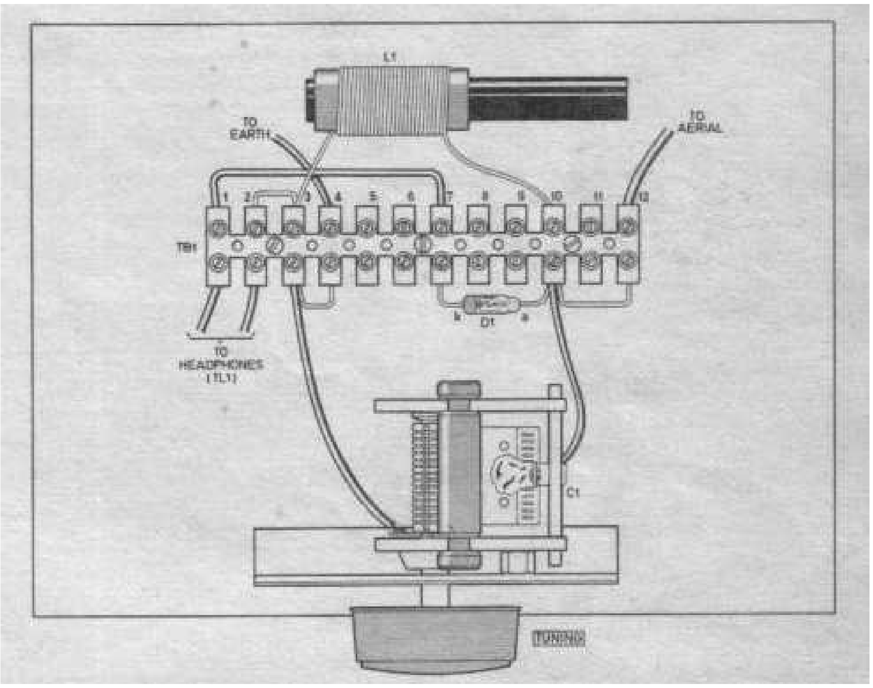

Picture 4 - Baseboard Layout Of The Crystal Set

AERIAL COUPLING

The aerial loads the tuned circuit heavily when connected directly to the top of the tuned circuit, as in Picture 2. This damps the tuning action and it can be found that stations spread out all over the dial, which is unsatisfactory.

The series capacitor, C2 connected in Picture 5(a) reduces the loading and thus improves the sharpness of the tuning. A variable or pre-set capacitor of about 250pF maximum is most suitable. for this role, though it is possible to experiment with a variety of fixed value capacitors in this range also.

Connecting the aerial to a tapping on the coil, as in Picture 5 (b) also sharpens tuning. It may also increase volume. Try about 2 turns from earth for coil 4, or 4 turns from earth for coil 3.

Another method is to have a coupling primary, as in Picture 5 (c). This consists of a second coil, with about one third the turns of the original wound on top of the existing coil.

You can even combine these methods to find what arrangement best suits the aerial in use.

The diode can be disconnected from the end of L1 and taken to a spare position on TB1 for example location TB1/9. You can then run a flying-lead fitted with a crocodile clip from this position, connecting it to various tappings on the coil as required as in Picture 5 (d). This method also reduces loading on the tuned circuit.

Coils with spaced turns of bare wire are readily tapped. For other coils, small loops can be made every ten turns or so, and crocodile clips can be attached to these when selecting tapping points.

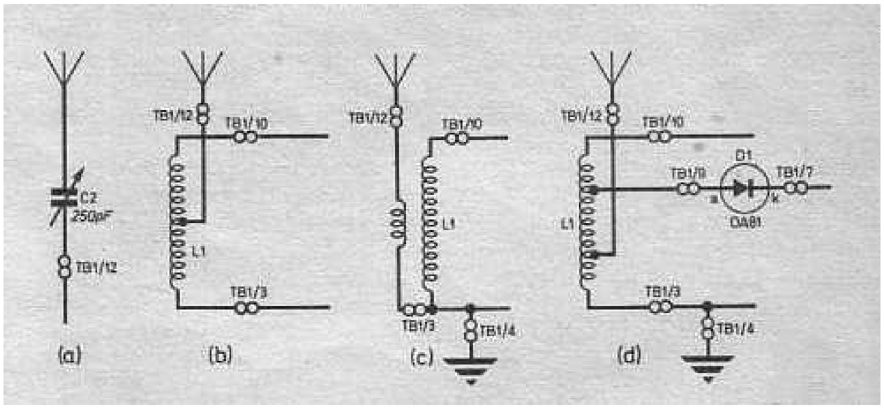

Picture 5 - Alternative Methods Of Aerial Coupling

SHORT WAVES

For shortwave reception, a good efficient outdoor aerial is certainly recommended. Evening listening in the region around 5 - 9MHz in often proves to be the most fruitful.

Since there is no amplification, as with a valve or transistor receiver, certain frequencies will seem to be completely dead at particular times of day. So if the crystal receiver works satisfactorily on medium wave and longwave, but no shortwave signals are heard, check again in the evening, or after dark, when conditions are different.

PARTS REQUIRED

C1	**365pF or 500pF Air Spaced Tuning Capacitor**
D1*	**OA47, IN34, OA81, OA90, OA91, IN94 or similar point contact small signal Germanium Diode** ***** The OA47 will be of particular interest since it has the lowest forward bias voltage of any of these diodes which will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.
TL1	**High Impedance Headphones (20,000 Ohms)** **or Crystal Earphone**
TB1	**12-Way Plastic Screw Block Terminal**
Also Required:	**47 k Ohm Resistor for Crystal Earphone: Enamelled Copper Wire: 32 and 24 s.w.g. for L1: 20 s.w.g. tinned wire for L1: Ferrite Rod 10mm diameter x 75 mm long: 25m of wire for aerial: Wire and rod or spike etc for earth: Wood for base e.g. 10mm x165mm x 130mm: Scrap of metal of thin plywood for C1 bracket/front panel: Knob: Crocodile clip(s)**

Adapted from an article in Everyday Electronics magazine, November 1981, By F.G. Rayer.

HERE ARE A COUPLE OF VERY INTERESTING CRYSTAL SET DESIGNS SENT IN BY KRYSATEC - "THE RAT" - FROM THE CZECH REPUBLIC

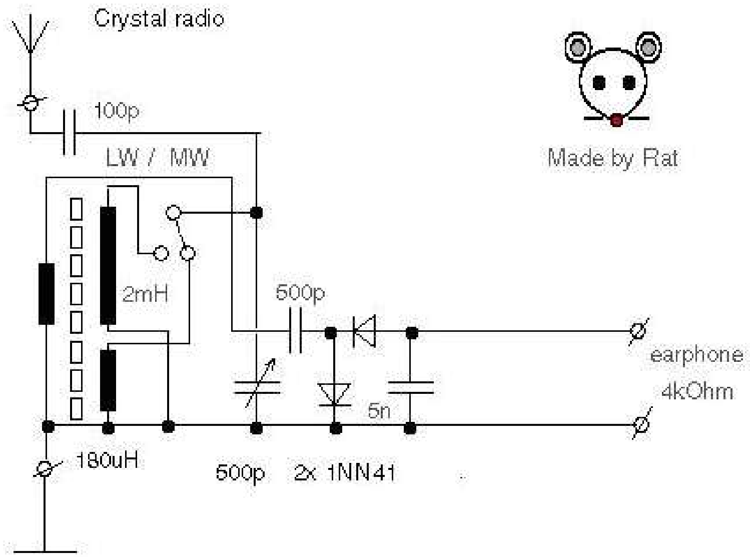
Crystal radio
100p
LW / MW
Made by Rat
2mH
500p
earphone
4kOhm
5n
180uH
500p
2x 1NN41

1/ Using old coils from old bulb radio for MW and LW band. Though it would be straightforward to wind the coils - one for Long Wave, one for Medium Wave and a coupling coil. Variable capacitor is 2 x 500pF only one half is used: 500pF. For the crystal earphone a resistor of about 82k ohm in parallel is required. This set also uses two Ge diodes as a multiplier in the quest for for higher audio signal output.

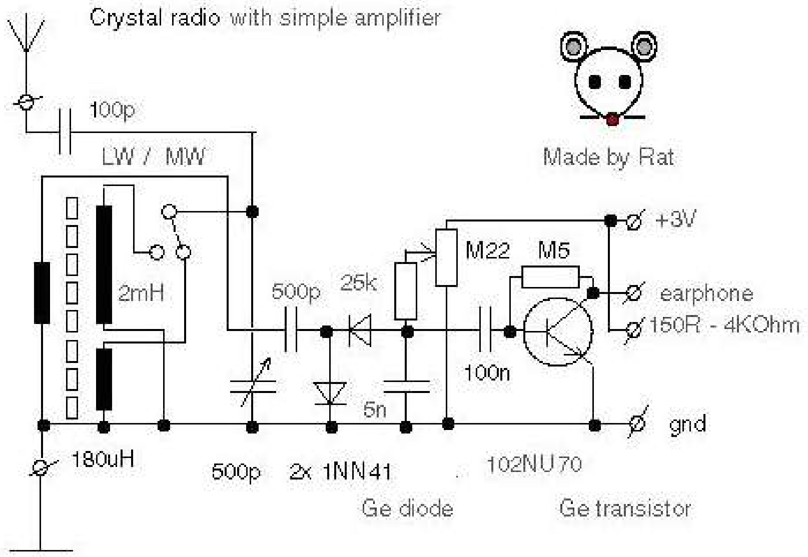
Crystal radio with simple amplifier
100p
LW / MW
Made by Rat
+3V
M22
M5
2mH
500p
25k
earphone
150R - 4KOhm
100n
5n
gnd
180uH
500p
2x 1NN41
102NU70
Ge diode
Ge transistor

2. If signals are not strong signal in your location, then the above circuit design can be considered. A simple transistor amplifier is used. A variable resistor M22 is used for better sensitivity which can be adjusted for poor signals. This crystal radio is aversion from cca 1960 - 1970 y.

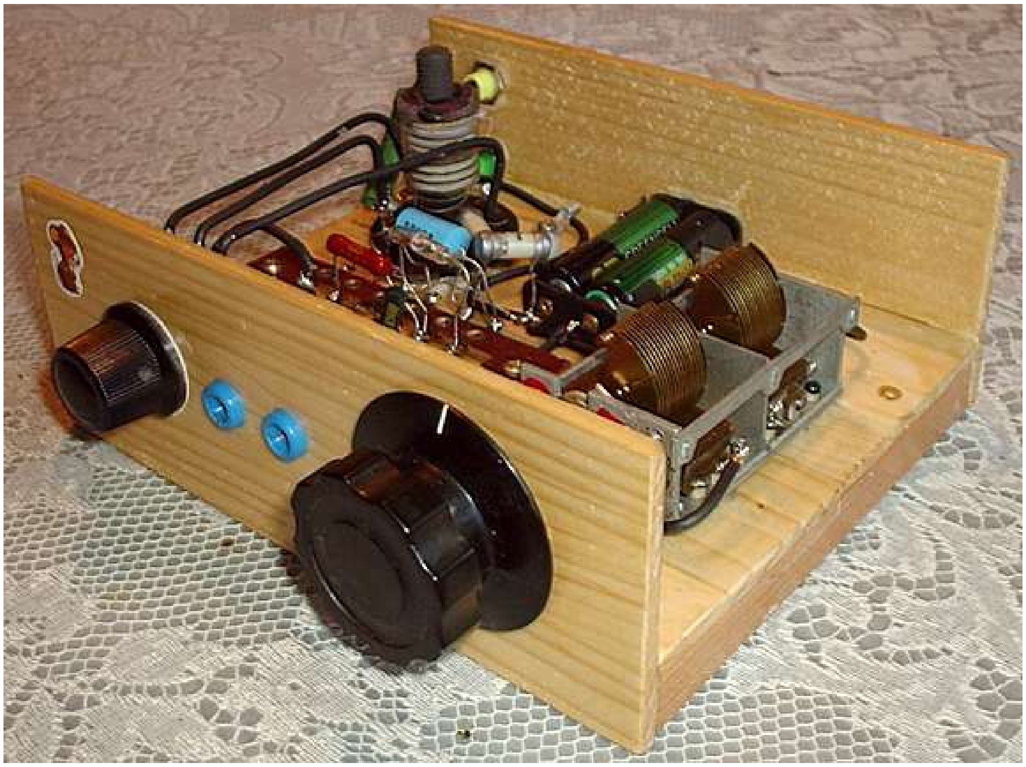

Rat's finished Crystal Set with additional amplification - very neat!

'Minilabs' Crystal Radio

BELOW: Ian Tomlinson kindly sent in a photograph of the box that contained the kit for his John Adams Toys 'Minilabs' Crystal Radio.

It is a very simple circuit consisting of the coil (inductor) with a sliding contact that provides variable tapping points, a diode and crystal earphone. All that is added is the aerial and earth. There is no variable tuning capacitor for simplicity and to keep costs down.

The coil provides the inductance required for tuning into a certain frequency (wavelength). These days a variable "tuning" capacitor is normally wired in parallel across the inductance (coil) in order to vary the resonance of the tuned circuit and therefore enable to easily tune into various transmitters on different frequencies. This crystal is tuned varying the number of turns on the coil (ie varying the inductance) by tapping off at different points using the sliding contact ("ball").

The crystal earpiece, or high Z headphone, is connected between the output of the detector diode (the other end from the coil) and earth. The volume from a crystal earpiece may be considerably improved by connecting a resistor of - somewhere between - 4.7 k and 47k ohms in parallel with the earpiece. A crystal earpiece cannot directly allow current to flow through it and the parallel resistor therefore allows current to better flow through the circuit.

John Adams Toys
Crystal
Radio Kit
Make a radio complete with tuning coil, antenna and earphone.
Pack contains all you need including full instructions. No battery required.
Ages 8–14

'Minilabs' crystal set by John Adams Toys

A discussion on configurations for Crystal Sets by Felix Scerri VK4FUQ

This discussion, by Felix Scerri VK4FUQ, was posted at this address which no longer appears on the web www.tarc.org.au/techinfo2.htm (error 404) so here it is reproduced:

Crystal Set design is one of my passions closely allied with my obsession for audio and high fidelity.

My main interest in crystal sets, apart from the wonder of a radio receiver that does not require a power source, is the potential excellence of the recovered audio quality from normal AM broadcast stations.

Personally, it is one of my great laments that most people have never heard how good wideband AM can sound. A high performance crystal set or similar TRF approach is, in my opinion,the only way to do it. There are a few people around who have heard the audible results of my efforts,and can only agree.

I have often wondered,given the ultimate simplicity of the crystal set, being essentially a tuned circuit,a diode detector and some form of output device, what it takes to achieve optimum performance. What follows are my thoughts on the matter.

Crystal Set optimisation, is in my opinion, all about reduction of circuit losses. Essentially this means high "Q" tuned circuits and high quality detectors. Efficient output devices also help too. But as we will see, there are some trade-offs required as well. A high "Q" tuned circuit is always benefical, as a high "Q" tuned circuit has lowest RF losses,highest potential selectivity,and highest voltage at resonance, which is very useful for the diode being fed from the tuned circuit. Variable capacitors, even the "modern" miniature variable capacitors (although the older air dielectric units, as used in old valve receivers are more desirable) for various reasons,are generally quite efficient, and a higher "Q" coil will produce the most worthwhile improvements.The best (highest "Q") coils are wound with "Litz" wire, which is a multistranded woven wire with all strands insulated from each other. The performance of Litz wire wound coils is spectacular, unfortunately, although I know Litz wire is still being made, from personal experience, it is VERY rare in Australia.

Efficient coil design can be quite complex and all my coils are wound on ferrite rods. There seems to be,at least for ordinary single wire windings (close wound), an optimum wire thickness for optimum coil "Q". I have determined .315 mm winding wire to be about optimum for simple (single wire) coils on ferrite rods. Thicker wire is NOT better, believe it or not.

Lacking Litz wire, an interesting winding approach I have developed is to use two slightly thinner wires wound as a bifilar winding connected together at the beginning and end of the coil, yields considerably higher "Q" compared to a simple single wire winding. I have found 0.25 mm winding wire optimum in this application.

Whilst high "Q" coils are beneficial from the RF point of view, there is a possible downside. If one is interested in maximum selectivity and sensitivity, there is no problem, but remember highest "Q" results in a narrowed audio band-width as a simple consequence of band-width. For high fidelity applications this could be a disadvantage under some circumstances, although there are clever ways around this.

Regardless of ultimate coil "Q", selectivity is a major issue with crystal sets generally. Here another trade-off is evident. For the maximum voltage into the diode, connecting the diode to the high impedance end of the coil (i.e. the top) yields the greatest voltage but the selectivity is usually terrible, because of severe "loading" by the diode circuit. For this reason, tapping well own the coil improves selectivity at the expense of signal volume (reduced voltage). Once again there are ways around this. As described in my "Double Tuned Crystal Set Tuner" article in "Amateur Radio" magazine, March 2002, the use of two separately tuned coupled resonant circuits allows top connection into the diode without compromising overall selectivity, thanks to the use of a second tuned circuit which is fed from the external antenna. The whole network forms a double tuned input bandpass filter and in practice this approach works very well. For single coil crystal sets I recommend the use of an un-tuned "antenna" winding adjacent to the "hot" end of the main coil, preferably adjustable (old paper reels from sewing cotton threads are ideal). This allows the degree of coupling to be optimised under actual listening conditions. The double tuned set up is best, yielding superb selectivity, but the un-tuned antenna coil arrangement also works quite well, especially if the diode is tapped well down the main coil.Tapping halfway works well.

The other method of performance improvement involves the use of the most effective detector system possible. Here things get very interesting. In fact the temptation is to use more complex circuitry, but that gets away from the charming simplicity of the crystal set. As an example, my own crystal set tuner has at times mutated into a TRF tuner complete with FET RF preamplifiers, active(powered) detectors and other enhancements. These modifications do work well, but loses the simplicity of a basic crystal set. In actuality, a simple diode detector can work extremely well, subject to some qualification. Diodes like to work with a reasonable level of RF input voltage. Audio distortion can result under conditions of low signal level, due to diode transfer curve non linearity and other factors, such as the widespread use of broadcast station "processing". The actual type of diode makes a difference. The 1N34A germanium diode is very popular for crystal set use, although in my experience just about ANY germanium diode will work, although it is worth trying different specimens. Some are definitely better than others. Even from a pack of twenty 1N34A's from the same source, some were definitely better than others. Measuring the average value of rectified output voltage across the diode load resistor will show which diodes are best. By the way, I regard a diode load resistor as being mandatory. I find a value of about 47K about right, especially if a crystal earpiece is being used or the crystal set is being used as a tuner feeding an audio pre-amplifier and following amplifier. If using high impedance magnetic type headphones, the headphones provide the diode DC load.

Another type of diode that is very interesting, is the hot carrier diode. There seem to be a lot of different hot carrier diodes around these days. There are even hot carrier diodes now being sold as "germanium diode equivalents". I have tried them and they do work acceptably well, but they are not quite as good as genuine germanium diodes such as the 1N34A. Typical UHF mixer hot carrier diodes, such as the 1N5711 will not work well in crystal set service simply because their "turn on voltage"is too high, similar to silicon diodes such as the 1N4148/914 series, which require a lot of RF input to function adequately as RF detectors, however a simple technique can be used to turn hot carrier diodes such as the 1N5711 into superlative detectors.

I guess we are cheating a little, because the technique is to use a little voltage bias supplied via a 1.5v battery, through a simple potentiometer voltage divider arrangement, with capacitor (for DC isolation) fed into the diode from the tuned circuit. With applied adjustable bias, I find the 1N5711 diodes absolutely superlative detectors under ANY signal strength conditions. I find the detection quality also superlative, with a clarity and low noise profile unmatched by any other diode arrangement. In my opinion, hot carrier diodes, running bias,are the best detectors overall.

Regarding other detector arrangements, the diode "voltage doubler" is often recommended, however my own experiments with the doubler arrangement have been inconclusive and slightly disappointing overall. I have found no real advantage in their use over a simple (one) diode detector, believe it or not.

Yes, they do work, but they're nothing special, at least in my opinion.

Any comments on this general subject of crystal set optimisation would be welcome.

73's Felix Scerri VK4FUQ.
22nd July 2002

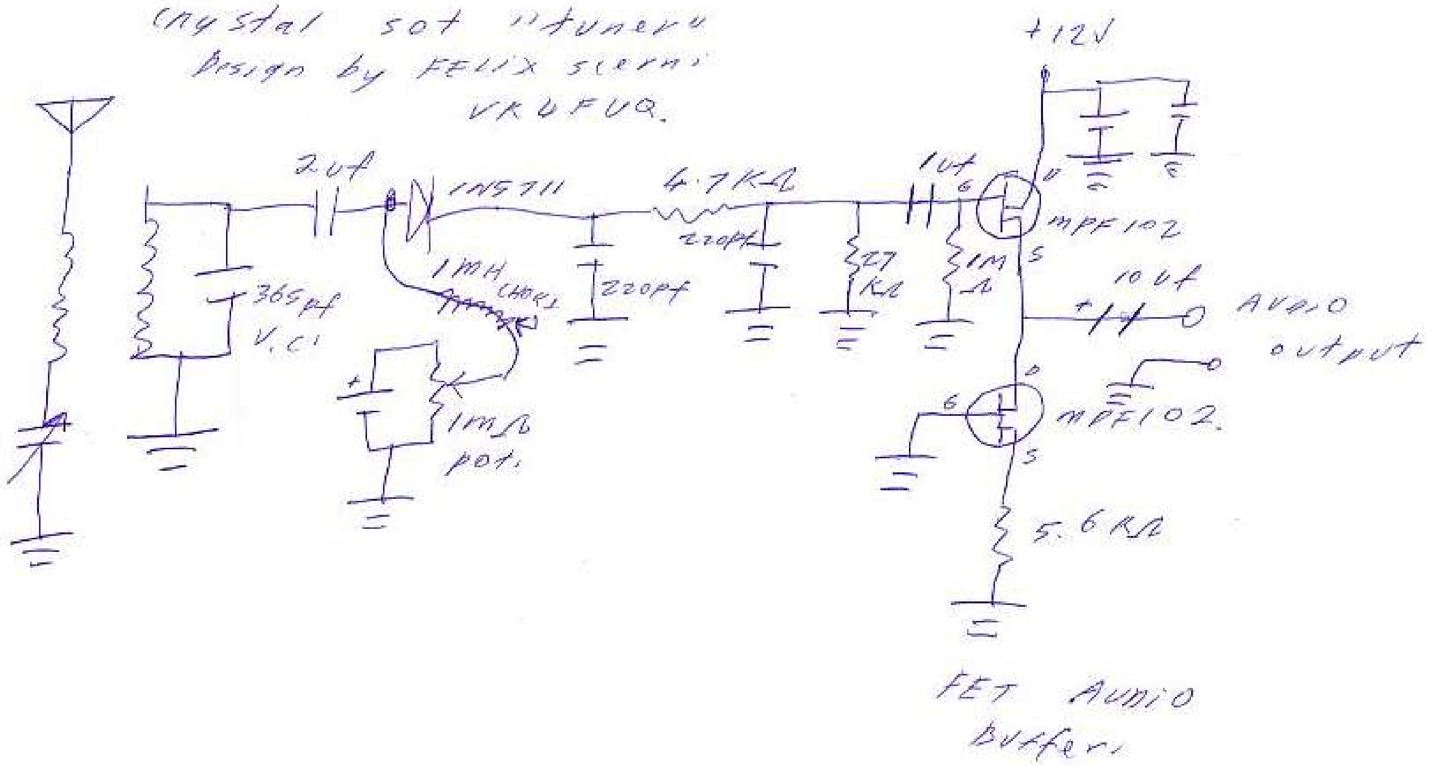

Above: CRYSTAL SET BASED CIRCUIT PROVIDING A HIGH QUALITY PROGRAMME SOURCE

IMPROVED VERSION OF THE ABOVE CONCEPT !! New update from Felix Scerri February 2010:

New 'two FET infinite impedance AM detector'

I've developed a new version of my old favourite FET 'infinite impedance' AM detector that I think sounds very nice. I include a short audio of one of our local AM stations. I picked this station as it is my reference 'torture test AM station' as they run very heavy 'processing' which normally sounds yuck with all my other (diode and non diode) detectors! However it's quite clean with this detector. What do you reckon? I'll do up a circuit if you'd like to feature it in your TRF radio section. A general draft article follows.

'A favourite non diode based AM detector that I've built and used many times over the years is the FET based infinite impedance detector, offering very good general AM detector performance, especially under weak RF signal conditions where diode based detectors do not perform well, especially in terms of audio distortion.

However one of the slightly strange things I've noticed about the simple FET based infinite impedance detector is the variable audio quality noted, even when using the same type of FET. Some I've built have sounded good and others slightly fuzzy when used with an audio preamp and fed into a high quality audio system. I've been giving this a considerable bit of thought of late and I've wondered if the audio distortion might be a result not necessarily of the detection process itself, but the FET stage in its guise as a 'source follower' audio stage which essentially, it is.

I have long been aware that as a simple audio buffer stage, the FET based 'source follower' can exhibit a considerable amount of audio distortion, and a technique I've long used to greatly reduce this audio distortion is to use a second FET in the source lead of the first FET as a 'constant current source' which serves to 'linearise' and greatly reduce audio distortion in the buffer stage overall. So, to test the theory I built a simple one FET infinite impedance AM detector which worked well, but with just a hint of audio 'fuzziness' on received AM stations. So I added a second FET in the source lead of the first FET wired as a constant current source, taking the output from the source of the first RF detector FET and the source resistor and RF bypass capacitor off the source lead of the second FET 'constant current source'. The result, totally clean audio! The theory seems proved! I call this modified detector the 'Two FET infinite impedance detector'

))) Here's what is sounds like - click to play the audio file (((

Here is the circuit diagram:

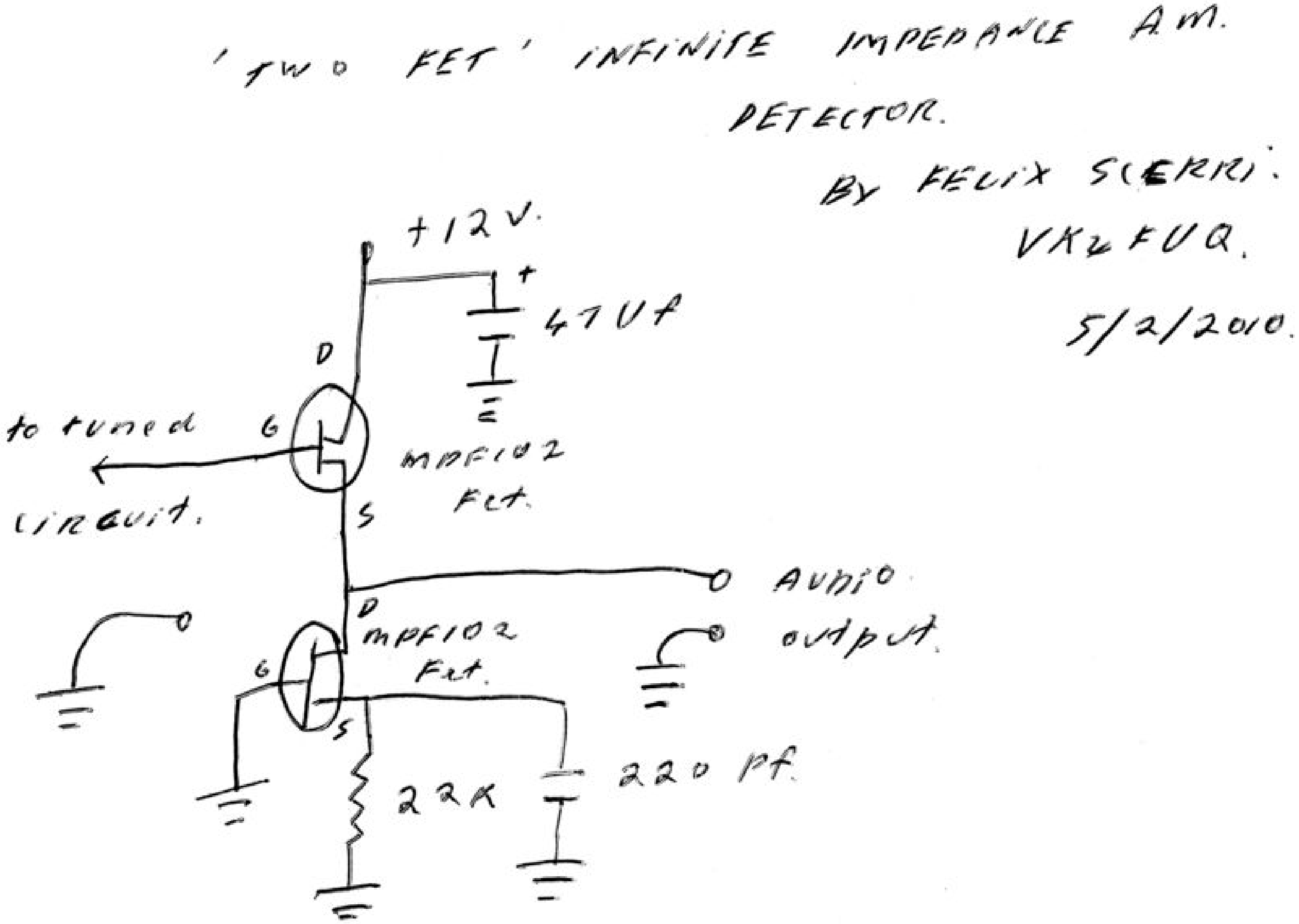

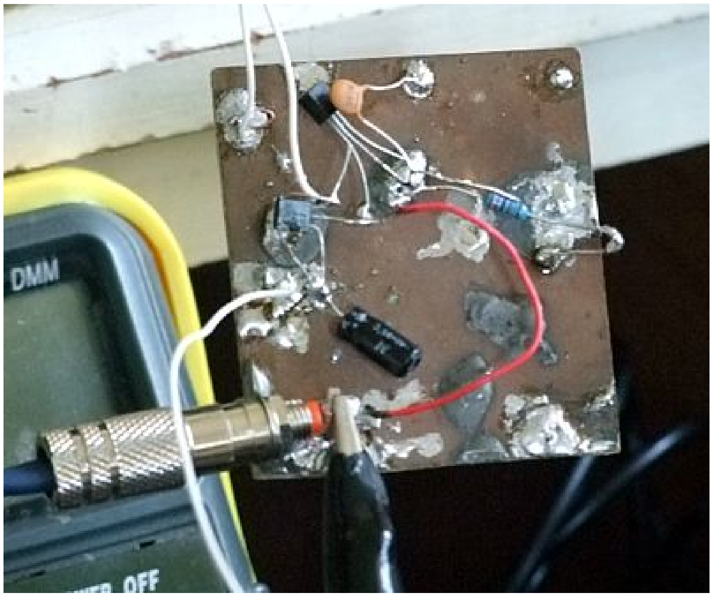

This detector has been a real eye opener for me in terms of its excellent performance, especially considering its circuit simplicity. Indeed in the past I have designed other more complex FET based infinite impedance circuits that do not quite work as well in practical terms as this latest circuit, at least according to my well calibrated ears!

I do not have access to any precise test equipment but my well calibrated ears tell me this 'two FET infinite impedance detector' is a beauty, surpassing practically every other AM detector I've built even at low RF input, and that's rather a impressive claim and the audio quality when used as an AM tuner feeding a high quality audio system is quite remarkable. Possibly the best thing about this detector is its excellent performance under weak signal conditions. Diode based detectors also work beautifully, but the use of an RF stage to ensure detection over a linear portion of the diode's curve is mandatory! This compound infinite impedance detector works beautifully on the sniff of a useable RF signal.

Just add a high Q tuned circuit and that's it!

Felix Scerri
VK4FUQ

A better FET for the 'basic' Infinite Impedance Detector:

Quite recently by accident, I've realised the MPF102 FET that I've long used in my FET based infinite impedance detectors is possibly not the best FET to use. This was the reason why I developed the 'two FET' infinite impedance detector some time ago which works very well.

However I've found the choice of a more suitable FET works beautifully in the basic FET based infinite impedance detector circuit, which has appeared for many years in many editions of the ARRL Handbook.

I use the 2N5457 and others of the same 'family' may be equally suitable, but I haven't tried them! However with a 2N5457 in place of an MPF102, the basic infinite impedance detector has became my AM detector of choice. It works beautifully even at low signal input with lovely and clean low distortion audio along with a very high input impedance for good tuning selectivity. It's a beauty! The basic generic circuit is attached, courtesy of Rod Elliott's ESP website.

73 Felix VK4FUQ
10 / 02 / 2012.

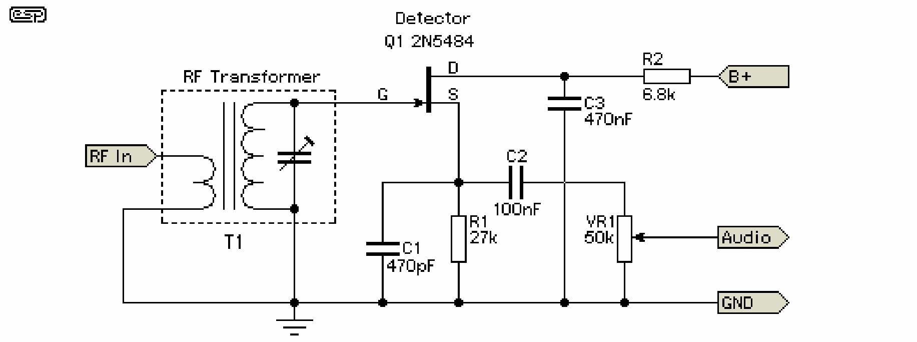

The basic generic circuit is attached, courtesy of Rod Elliott's ESP website
Felix Scerri VK4FUQ

As often happens with me, my renewed interest in FET based 'infinite impedance detectors' of late has led to some interesting new research and I may have considerably improved the 'two FET infinite impedance detector' as a result.

My research suggests that although the use of a CCS (constant current source) reduces audio distortion in an audio stage, the value of the 'source resistor' in the CCS stage is somewhat critical for best results.

By using a potentiometer in lieu of a fixed resistor I have found that a resistance value of around 470 kohms cleaned up all overall audio distortion. I used an MPF102 as the CCS in this circuit. An interesting and worthwhile little circuit refinement.

73 Felix VK4FUQ
21 / 02 / 2012.

A Minimum Component Count High Quality AM Detector

I was generally messing around with various circuit ideas and I came up with this AM detector circuit, a simple diode detector along with a FET stage. It was an attempt to provide good performance along with minimum number of components. Actually I've been pleasantly surprised at the excellent level of general performance and the best of all, it sounds great!

The circuit is quite conventional being a BAT 46 diode detector feeding an MPF102 FET buffer/ common source amplifier stage. I would ordinarily use a FET source buffer stage in this application, but opted to use a simple low gain FET 'common source' amplifier stage instead, with excellent results. I also used a BAT 46 Schottky diode instead of an ordinary germanium signal diode.

This was done for several reasons. Firstly, germanium diodes are now very hard to find but in any case these 'germanium diode equivalent' Schottky diodes are actually a superior diode, having very low noise, almost zero back leakage and an essentially complete absence of carrier storage effects and very good weak signal sensitivity. I call these diodes high fidelity diodes as they sound wonderful as RF detectors.

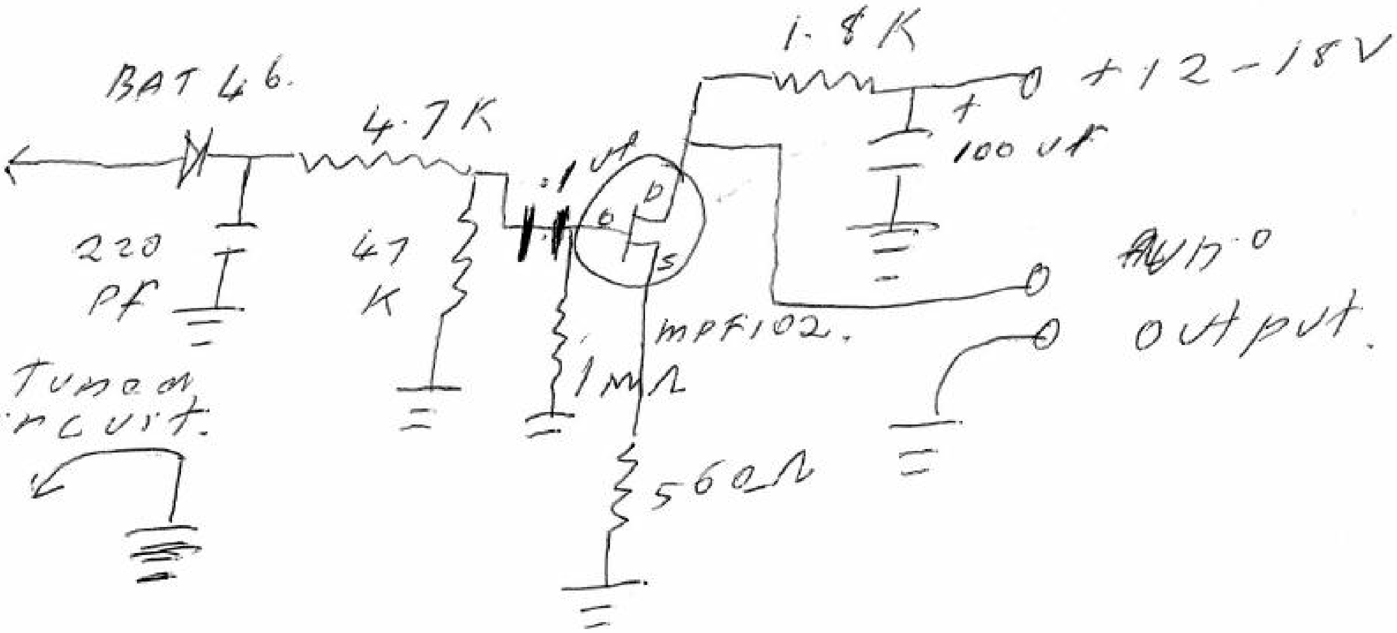

Circuit Diagram of the Minimum Component Count AM Detector by Felix Scerri

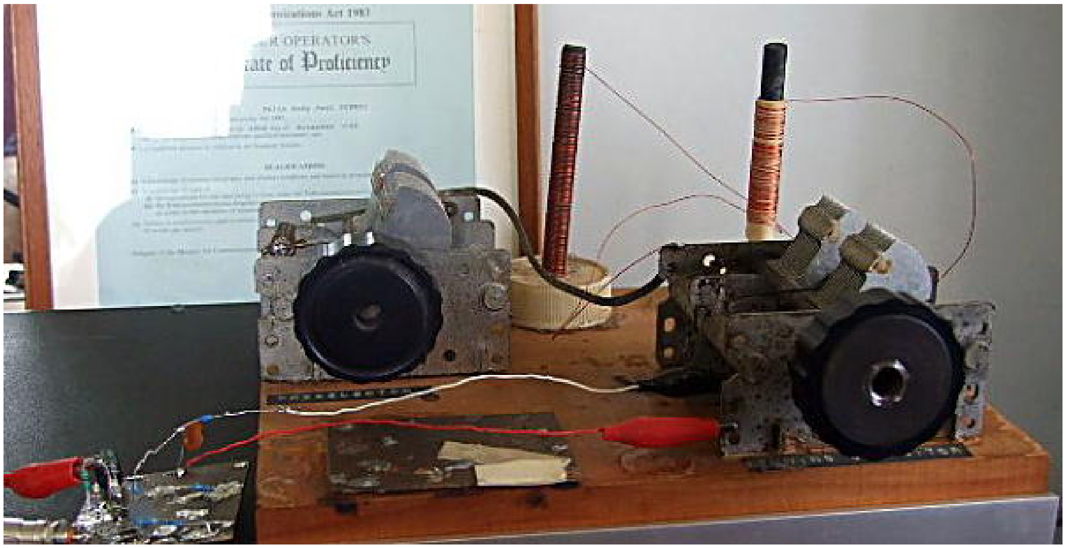

Minimum Component Count AM Detector by Felix Scerri

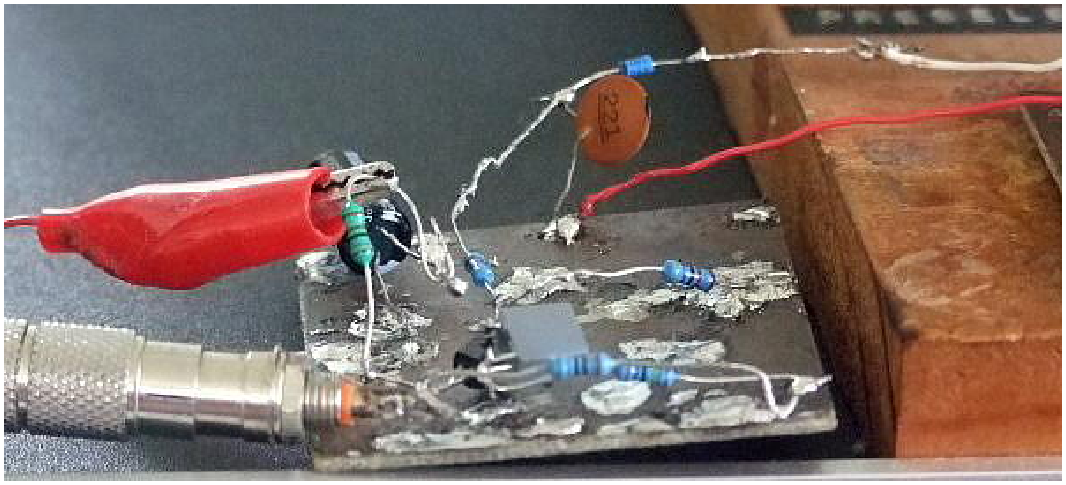

A Closer View

The high impedance of the FET's gate circuit is perfect for optimal buffering of the diode detector, something very important for good low distortion diode detector performance. Apart from providing slight voltage gain, the use of the common source FET amplifier is a new idea, as this prevents the possibility of incidental RF rectification occurring in a FET source follower stage, which can happen. A 1 uf plastic film capacitor may be added in series with the audio 'hot' output lead to block the DC offset out of the FET drain, if required.

Despite no additional RF stage ahead of the diode, audio quality on even relatively weak RF strength stations is actually very good, and of course the audio quality will be even better with increasing RF signal strength, something which will also increase the audio output level. Just on this, for a long time I was somewhat negative regarding diode detectors, as one AM station locally (the strongest one) was always distorted when using a diode detector.

I strongly suspected a transmitter fault, but my complaints were ignored, until one day some time ago when all audio distortion suddenly disappeared! Nothing was ever 'said', but I realised that my suspicions of a long standing transmitter fault were correct, after all!

73 Felix
VK4FUQ
02 March 2012

Voltage Doubler Detector

This is an AM detector circuit that I've long known about and which worked ok, but never seemed to work as well as it should have. However I spent some time late last night trying to optimise the circuit, with some success.

It is a curious circuit being essentially a 'voltage doubler' originally developed for power supply applications, and its use as an AM detector is hard to analyse! It seems that the component values in the circuit are somewhat critical for good performance and if not, the performance is rather 'ordinary'. The circuit that I eventually came up with uses a 150 picofarad 'input capacitor' with a 150 kohm 'load' resistor and loaded into a FET common source voltage amplifier stage (as previously described) through a coupling capacitor with a 1 Mohm input resistance.

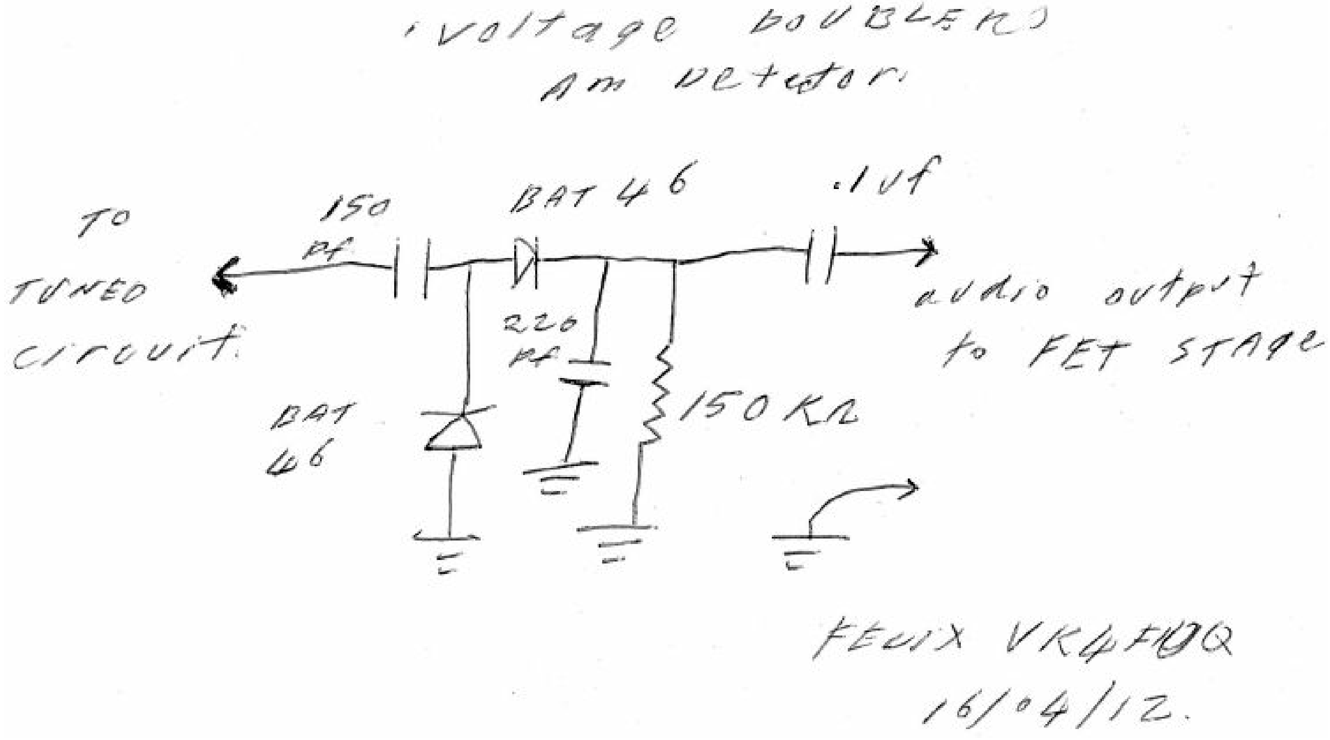

With these circuit values, it all works 'quite well'. Give it a go! It's an interesting AM detector with quite good 'sensitivity' and clean audio quality, and it seems to work well at low signal levels.

73 Felix
vk4fuq.

16th April 2012.

Simple AM detectors: What works best? A practical experimenters viewpoint.

I have written a lot about simple AM detectors for use as tuners for feeding into an audio amplifier, and it has been a long time interest. These days I use either diode based or 'infinite impedance' types of AM detectors. In this location our 'local' AM stations are quite distant and are therefore quite weak in terms of signal strength.

As such I find infinite impedance detectors based on field effect transistors give consistently better results for tuner applications due to their lower apparent overall detector distortion. Diode based detectors are quite 'fussy' as they require both optimal output buffering (AC/DC ratio) and an 'adequate' (beyond the diode knee) level of RF signal injection. http://www.tonnesoftware.com/appnotes/demodulator/diodedemod.html

Diode based detectors will happily 'detect' at very low signal levels, however the (inevitable) audio distortion that results, can be extremely irritating to the ear! Under these conditions I find infinite impedance detectors (even without additional RF preamplification and subject to individual FET characteristics), generally sound 'cleaner' and more pleasant to the ear.

FET's of course require a power source for operation whereas diodes are passive (un-powered) detectors (most of the time), however this is of no real advantage in a tuner application as an 'active' audio amplifier stage will generally be required anyway for audio level boosting, buffering etc.

In the end it will come down to a consideration of prevailing RF signal levels and other related circuit considerations at one's location. If local RF levels are strong, a well designed diode detector will give excellent results. If not, an 'infinite impedance' type of detector is most likely the better option unless one goes towards the option of additional RF preamplification prior to the diode detector.

73 Felix
vk4fuq.

17th April 2012.

ANOTHER DETECTOR by Felix Scerri !

I’ve tried diode based ‘voltage doubler’ (or more correctly ‘diode integrator’) AM detectors before with indifferent results, however the other day, just trying a few ideas I came up with this version that works rather well, with low audio distortion, high audio output and really ‘nice’ audio quality and the best of all, it seems to work very at very low RF input level.

The two diode ‘voltage doubler’ detector using two BAT 46 silicon schottky diodes feed directly into a MPF102 source follower stage set at 1 Megohm input resistance. The ‘input’ capacitor feeding the diodes from a tuned circuit is 68 picofarads.

I have the simple RF filter right on the output of the FET stage. In that respect this circuit is vaguely similar to the old ‘Selstead-Smith’ valve AM detector of the past. An interesting one! I am very happy with its general performance. Regards, Felix vk4fuq 11/04/2013

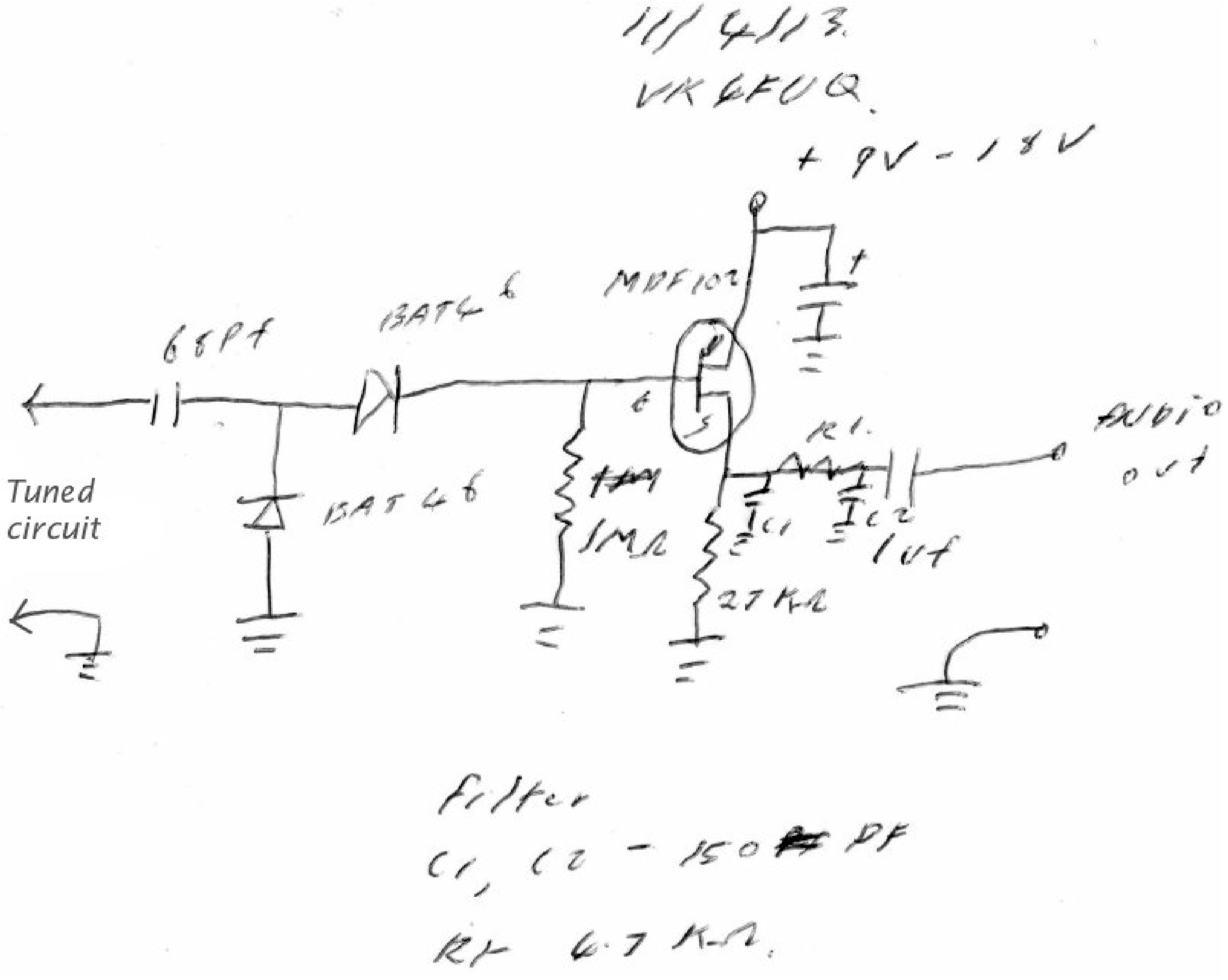
11/4/13.
VK4FUQ.
+9V-18V
MPF102
BAT46
68pf
Tuned circuit
BAT46
G
S
~~1M~~
1MΩ
R1.
C1
C2
1uf
27KΩ
Audio out
Filter
C1, C2 - 150 ~~P~~ PF
R1 4.7 KΩ.

Felix Scerri VK4FUQ

UPDATE - JUNE 2013

G’day all, readers may recall the two FET infinite impedance detector I developed some time ago. That circuit worked well, but some samples of the MPF102 regretfully produced distorted output. However a recent discovery has resulted in an improved version that has truly exemplary performance.

The MPF102 is in all honesty a device essentially designed for VHF applications, but is often used in simple audio applications albeit with occasionally indifferent results, due to general device parameter ‘spread’. More or less by empirical trial and error I have found that a simple change in source resistor value to 100 k (from a much lower value), pretty well fixes everything.

In a FET source follower circuit this is interesting as the output impedance is actually a lot lower than 100 k due to the action of the FET’s transconductance. It is similar to the action of a bipolar transistor emitter follower circuit. In fact a simple infinite impedance detector with a 100 k source resistor actually works very well and will be good enough for many applications, however the addition of a second FET as a constant current source does markedly reduce overall audio distortion in the stage overall, and produces very clean and low distortion audio quality and also reduces the output impedance considerably (good for tuner applications), so take your pick!

Regards, Felix vk4fuq 02/06/2013.

UPDATE - JANUARY 2014
A simple modification to the basic FET 'infinite impedance' AM detector that dramatically improves performance.

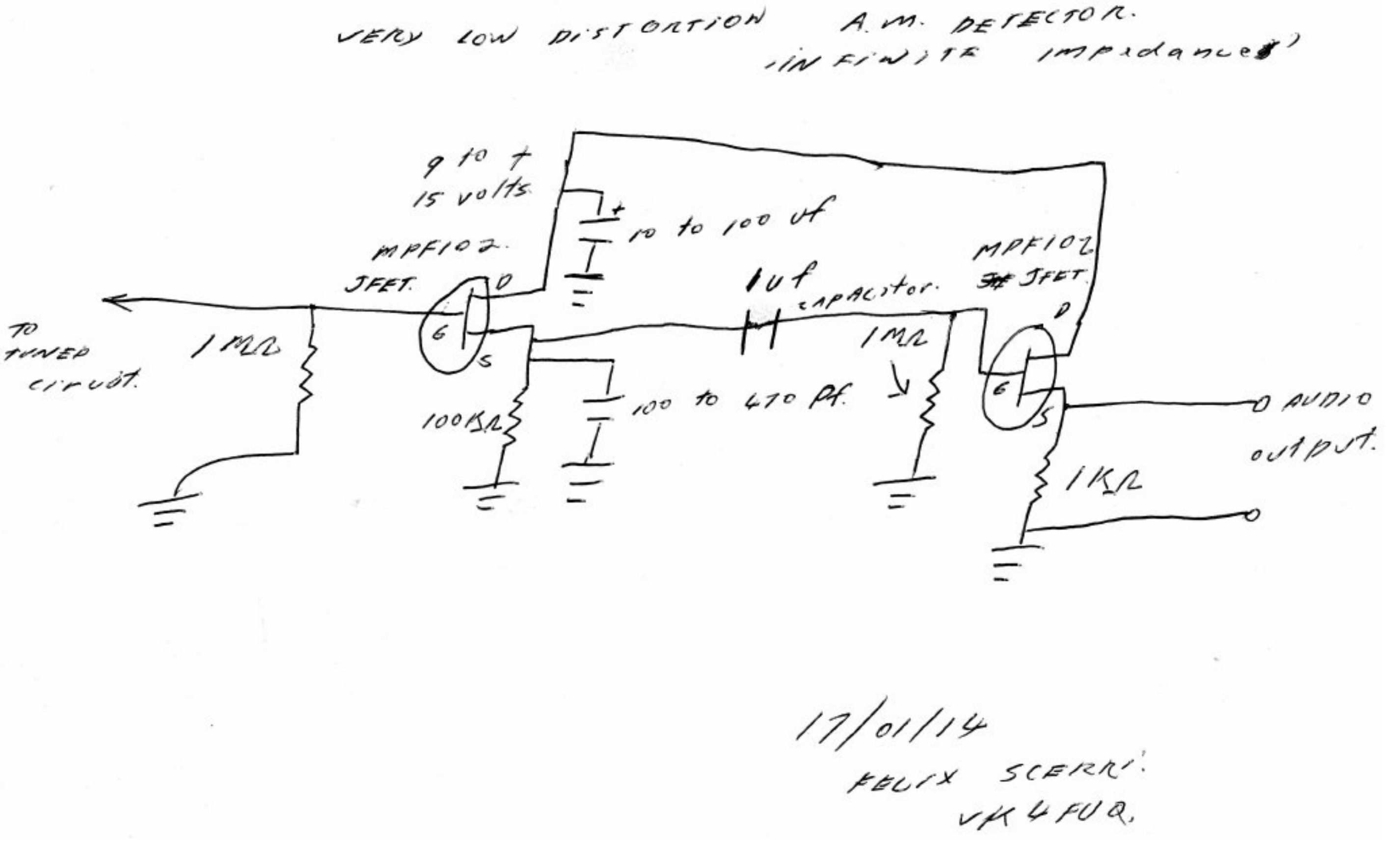
VERY LOW DISTORTION A.M. DETECTOR.
(INFINITE impedance!)

17/01/14
FELIX SCERRI.
VK4FUQ.

G’day all, over the Christmas break just messing around I worked out (mostly by accident), a simple modification for the simple FET based ‘infinite impedance’ detector circuit that dramatically improves weak signal performance and also greatly reduces audio distortion.

Essentially by the addition of another FET ‘buffer’ stage, another source follower, capacitively coupled from the first detector stage. The circuit is actually a simplified version of the circuit that I described in this link, http://sound.westhost.com/articles/am-radio.htm (figure 6) and testing the two head to head, they both sound superb and the simpler version is actually somewhat easier to build. I cannot get over how low distortion and ‘nice’ the recovered audio sounds. It is a joy to listen to!

Regards, Felix vk4fuq 12/01/2014.

Diodes for 'weak signal' crystal set applications.

As I am primarily interested in using crystal sets as AM tuners for feeding into a preamplifier/amplifier and loudspeakers, the actual type of diode can be relatively critical. In a strong signal area, not so much, but in a weak signal area such as where I reside, definitely. A diode with a good 'square law' performance (the area below the 'knee' in the diode curve), generally results in much cleaner and lower distortion than other good performing diodes, and believe me diode distortion under weak conditions, even with optimised diode 'buffering' used is very nasty sounding to the ear!

Testing many, many diodes in actual working crystal sets and listening critically to the audio output, it seems to me that the best diodes to use under weak signal conditions are the so called 'gold bonded' germanium diodes. I have sampled many different gold bonded germanium diodes and they all work well in this specific application, although sometimes the rectified output voltage may not be as high as other germanium or silicon schottky diodes, however the audio quality is much cleaner and shows much less apparent distortion!

I have tried OA5, OA47, IN141 and several 'CG' gold bonded germanium diodes with consistently excellent results. Other 'ordinary' germanium diodes may also work well under weak signal conditions, but they will need to be tested individually to check actual performance in a working circuit. One thing that I have noticed about germanium diodes is that due the 'point contact' nature of their construction, even diodes of the same specific type can exhibit rather different levels of performance!

The otherwise very good BAT46 silicon schottky diode works extremely well at good RF input level but not so well at weak RF input, especially when heavy broadcast 'processing' (commonplace these days), is used. New 'gold bonded' germanium diodes are probably no longer made although I am aware that they can be purchased through vendors over the internet. Apart from that, they may be found in old gear. As stated earlier, ordinary germanium diodes may be quite good but will need to be checked individually. Diodes are complex things!

Felix (vk4fuq). 29/01/2014.

* * * * * * * * * * * * * *

Just when I thought that I've tried everything I realised that I've not tried a voltage doubler (diode integrator) detector with an RF stage in front. It is an unfortunate fact that all diode detectors 'need' a good RF injection level (and buffering) or audio distortion becomes 'bad' (an understatement!)

Consider yourself very lucky if you live in a strong signal area! (unfortunately I don't). Anyway I quickly built a couple of prototypes which worked ok but not as well as hoped. Initially I tried using a source follower buffer (no voltage gain), but converting it to a common source amplifier (with appreciable voltage gain), seemed no better, which was strange.

After staring at my prototype for what seemed like forever with my magnifiers on, I (finally) realised my mistake. A common source stage takes its output off the 'drain' terminal of the FET, not the 'source' terminal as it does with the FET source follower buffer stage. I shifted one wire, and all worked as expected and it sounds fantastic! All diode detectors (one diode envelope detectors, doubler detectors etc), need good RF input and a simple untuned FET RF gain stage works very well. As to the sound, it sounds great.

73 vk4fuq 10/03/2014.

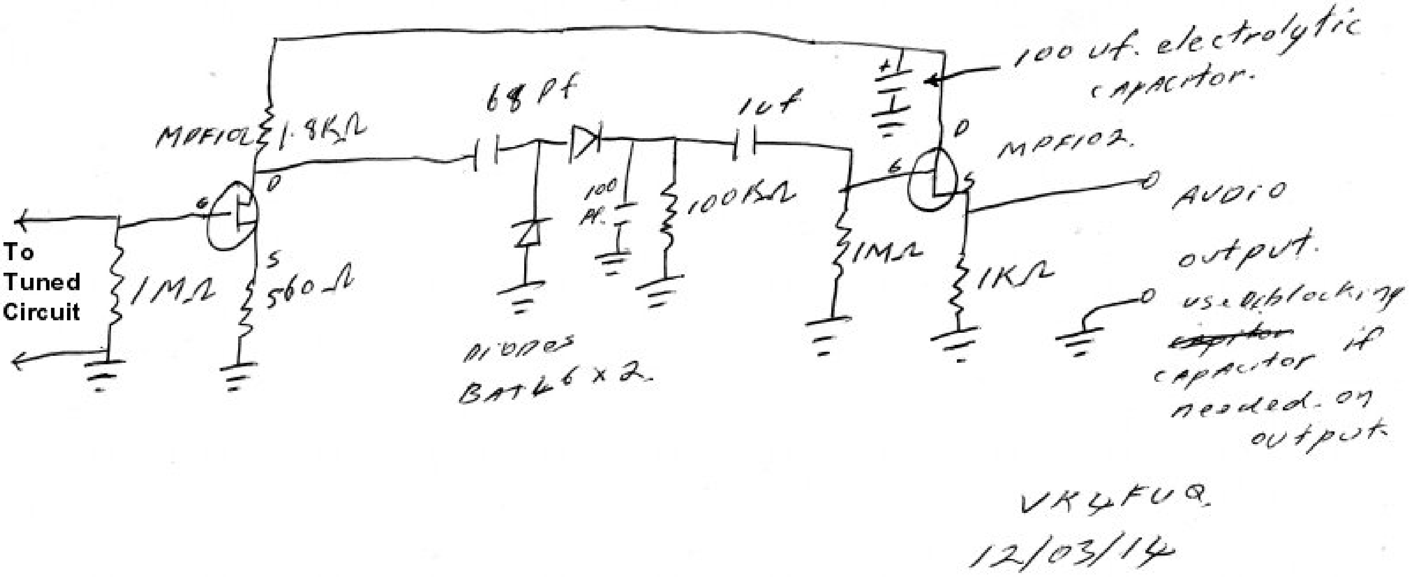

Felix Scerri's FET Crystal Set with Voltage Doubler

A 'nice sounding' AM detector - update October 2015

Looking through these pages the other day I was slightly shocked to realize that I've completely forgotten about some of the circuits that I've contributed in earlier days! They all work pretty well, however in my advancing older age and a little like my tastes in high fidelity generally, I'm starting to show a particular preference for 'nice sounding' bits of audio gear.

I guess that this also means low noise/low distortion too, although with AM broadcasting, at least in this country and probably elsewhere in the world too, the very common use of broadcasting 'processing' tends to make it hard for AM detectors generally, often resulting in a 'hard/compressed' sort of sound although still low in distortion, is not 'nice sounding', if that makes any sense!

Well of all the AM detectors that I've tried and/or developed, only one sounds 'nice' when confronted by heavy broadcast processing and that is an AM detector 'based' on a voltage doubler/diode integrator circuit (similar to the above circuit, actually the 68 pf capacitor should be changed to .1 uf, for slightly greater output).

It has taken me a very long time (years) to 'optimise' this circuit, but as it presently stands this is my favourite 'nice sounding' AM detector, and gets most use for general high quality listening on the AM broadcast band. It sounds really good! It somewhat reminds me of an old OA47 gold bonded diode detector from years ago before such abhorrent 'processing' became commonplace!

Ah yes the OA47….now that is a lovely sounding detector diode!

Regards, Felix vk4fuq. 17 / 10 / 2015.

* * * * * * * * * * * * * * *

That's it for crystal sets. I hope you try building one, it's easy and great fun!

See some useful links below....

73's
Mike

Crystal Sets (Part1) | Build Your Own Crystal Set (Part 2)

Spider's Web Crystal Set (Part 3) | Crystal Set By Kenneth Rankin (Part 4) | Crystal Radio Links

Amateur Radio

M0MTJ Visit my amateur radio pages M0MTJ

Visit my Amateur Radio Pages >

No AM radio stations or transmitters in your locality or country?

Spitfire
AM Transmitter

Has your local medium wave broadcast station closed or been moved to VHF/FM or Digital? Don't worry. You can still build and experiment with crystal sets and TRF radios by also buying or even building a simple low power AM transmitter. So, not only can you use your crystal sets but you can also run your own radio station that can be heard in and around your home - playing the music or programmes that you want to hear!

SSTRAN AMT3000 Superb high fidelity medium wave AM transmitter kits from SSTRAN. Versions available for 10kHz spacing in the Americas (AMT3000 or AMT3000-SM) and 9kHz spacing in Europe and other areas (AMT3000-9 and AMT3000-9SM). Superb audio quality and a great and well designed little kit to build: http://www.sstran.com/pages/products.html

POWER
GAIN
MODULATION
COMPRESSION
SSTRAN
AMT3000

http://www.sstran.com/

Other AM transmitters available:

Spitfire & Metzo Complete, high quality ready built medium wave AM Transmitters from Vintage Components: http://www.vcomp.co.uk/index.htm Vintage Components offer a choice of the high quality Spitfire and Metzo transmitters:

SPITFIRE AM Medium Wave Transmitter with 100 milliwatt RF output power:

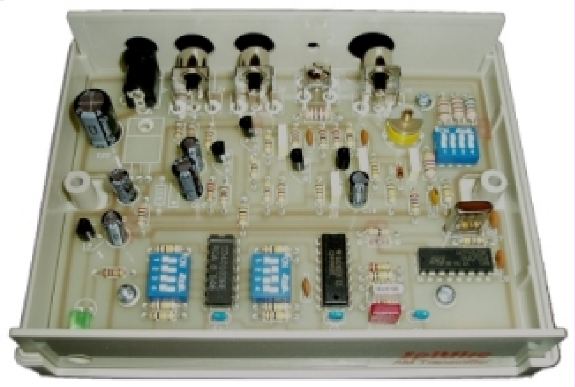

http://www.vcomp.co.uk/spitfire/spitfire.htm

Spitfire
AM Transmitter

METZO AM Medium Wave Transmitter with built in compressor:

Gain
Comp
Mod
Metzo

http://www.vcomp.co.uk/metzo/metzo.htm

AM88 LP A basic AM transmitter kit from North County Radio.
http://www.northcountryradio.com/Kitpages/am88.htm

LINKS: Fine links to more Crystal Radio websites here

Component Suppliers: Links to electronic component suppliers here

Amateur Radio
M0MTJ Visit my amateur radio pages M0MTJ

Visit my Amateur Radio Pages >

Crystal Sets (Part1) | Build Your Own Crystal Set (Part 2)

Spider's Web Crystal Set (Part 3) | Crystal Set By Kenneth Rankin (Part 4) | Crystal Radio Links

^Top Of Page

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

www.mds975.co.uk © 2004 - 2015

MDS975
www.mds975.co.uk

CRYSTAL SETS 2

Some Practical Designs
MAKE YOUR OWN
CRYSTAL SET !!

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

Crystal Sets (Part1) | Spider's Web Crystal Set (Part 3

Crystal Set by Kenneth Rankin (Part 4) | Experimental Crystal Sets (Part 5) | Crystal Radio Links

CRYSTAL SETS 2: SOME PRACTICAL DESIGNS

I hope that you attempt building one or two of these crystal set designs and I really do recommend that the components are carefully connected up using soldered joints onto a piece of tag-strip for reliability. However if you are new to constructing such electronic circuits then some simple solder-less techniques could be employed and these are suggested at the bottom of the page. Also see Crystal Sets Part 5 for more ideas on experimenting with crystal sets.

An early and very basic crystal set would have been nothing more than a coil of wire, perhaps 50 -100 turns, wound around a cardboard tube about 3 inches (7cm) in diameter, a detector (or cats whisker) and a pair of special High Impedance headphones (as discussed in part 1).

There would be a very large aerial strung up around the garden and the all important connection to earth.

The coil would have tapping points (connection points) at intervals of around 5 or 10 turns. See the circuit diagram on the right for details of who the set is wired together.

The tapping points on the coil allow the set to be tuned to different frequencies by adjusting the position of tap B. Tap B would be connected to the coil at differently positions by way of a crocodile clip. The fewer turns between the top (aerial end) of the coil and tap B, the shorter the wavelength received (ie the higher the frequency). Tap A would allow the detector to be connected at different positions to vary performance. There is an additional component drawn in the above diagram, the *capacitor* (value 1000pF), this is included in crystal sets that used the *High Impedance* magnetic headphones, and bypassed any remaining radio frequencies (RF) to earth.

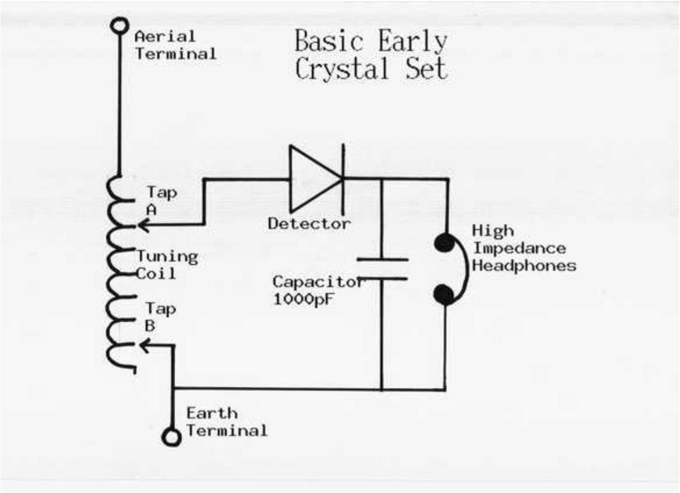

A very basic crystal set circuit.

I have not built the set described above as it is so basic. Such a crystal set above would probably have been adequate in 1920 - 1923 when there would have been only one local transmitter receivable.

When the BBC expanded transmissions and it became possible to hear more than a single station it would have became necessary to include a more convenient means of tuning the set.

This was achieved by including a Variable Tuning Capacitor, of about 500pF (0.0005uF) connected in parallel with the tuning coil forming a *tuned circuit*. The tuning capacitor would have a Bakelite knob on the spindle to aid tuning.

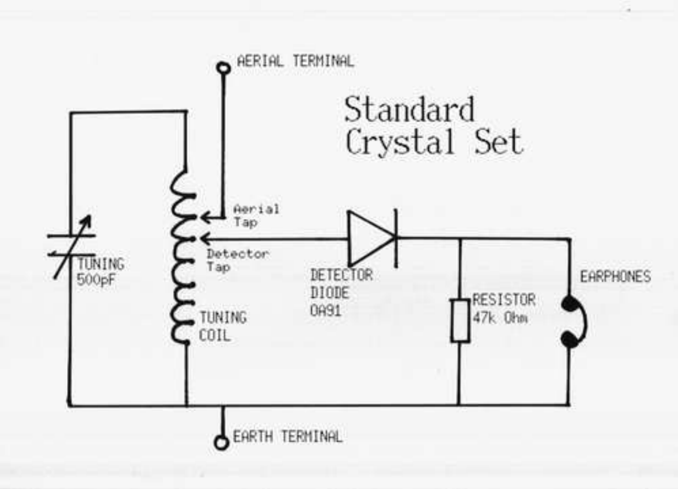

The Standard Crystal Set

Because of the simplicity of crystal sets, it is often difficult to separate stations. When tuned into one station it is often possible to hear another close by station in the background, this is due to lack of *selectivity*. This can be reduced somewhat by adjusting the positions of the Aerial Tap and Detector Tap. Moving them closer to the bottom of the coil, the earthy end, reduces the load on the tuned circuit and this improves selectivity, however it does also reduce *sensitivity* which can make the station quieter. Headphones will often swamp a tuned circuit and reduce its selectivity (Q factor), so moving the tapping point lower down improves this situation. Every circumstance is bound to be different though so the best balance has to be found by experimentation. My crystal set has both the diode and the aerial connected to the same tapping point on the coil, about a quarter of the way down.

The modern 'standard crystal set' shown above uses a Crystal Earphone, since suitable high impedance magnetic headphones (of 2000 to 4000 ohms) are no longer widely available. When using a crystal earpiece the 1000pF capacitor shown in the first diagram can usually be omitted an in its place a 47k ohm resistor is connected, this ensures that the Crystal Earphone will work at its most efficient i.e. the sounds will be as loud as possible. The resistor allows DC current to flow through the circuit efficiently - this would otherwise be blocked when using a crystal earphone. In a modern crystal set the detector used is a *Diode.* Suitable diodes include OA80, OA81, OA90 OA91 and IN94 which are usually available from component stockists.

A Better Diode For Increased Efficiency

The OA47 will be of particular interest since it has the lowest forward bias voltage of any of these diodes which will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.

On the right you will see my real working example of a crystal set

The large plastic knob on the front turns the variable tuning capacitor. This set receives the three UK national stations and also three local radio stations very well at my location.

There is a small 3.5mm jack socket mounted on the front of the plastic case (MB5 from Maplin Electronics) that the crystal earphone plugs into.

The coil can be seen inside the case, it is 70 turns of 30 gauge enamelled copper wire wound around the centre of a toilet roll and tapped every 10 turns, by scraping off the enamel insulation and making a small twist. The croc' clips can be seen clipped on to these twists to connect to the aerial and detector tap points.

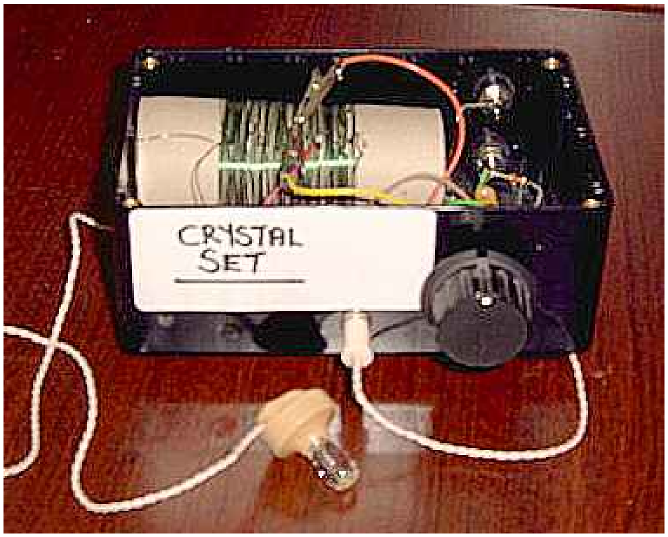

A real working crystal set. Radio as if by magic with no battery or mains power.

THE MEDIUM WAVE COIL - MORE DETAILS

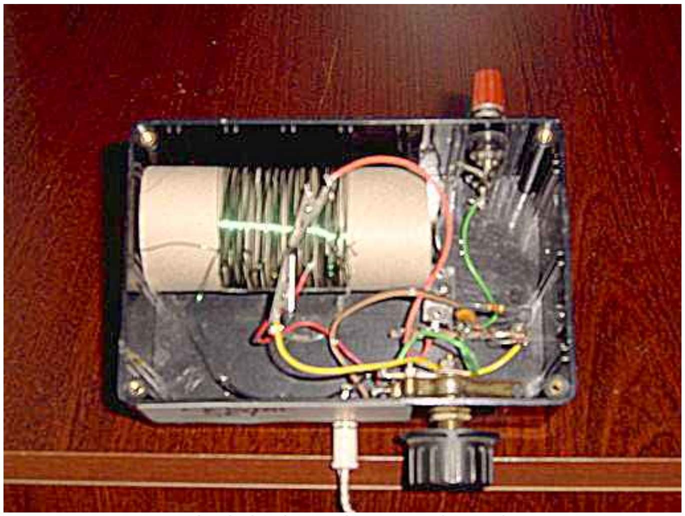

PHOTO SHOWING THE INSIDE OF THE COMPLETED CRYSTAL SET

Medium Wave Coil

The number of turns of wire required on the coil will vary depending on the size of the former (in this case the inside toilet roll) and the thickness of the wire. So to obtain the correct coverage of the medium wave band may need a little experimentation.

I usually find that between 50 to 90 turns is right and I generally use enamelled copper wire that is between 30 s.w.g. and 26 s.w.g (i.e. 0.315mm and 0.45mm diameter), so it's best to start with too many turns and then work down.

The more turns that you use the lower the frequency range will be, i.e. too many and the coverage of the top end of medium wave around 1500 - 1600 kHz will be lost, while too few and the coverage down to 500 kHz will be lost.

It is also important that the coil former is non conducting, i.e. not metallic. It could be wood or cardboard or a short piece of PVC piping and with a diameter of between 1½ and 4 inches (4 to 15 cm) are common sizes. You could try using a ferrite rod too, see below.

This particular set has a coil wound onto a toilet roll tube which consists of 70 turns of 30 s.w.g. (0.315mm dia) enamelled copper wire tapped at every 10 turns. It also has the additional small trimmer capacitor that helps match the aerial to the tuned circuit thereby improving selectivity, see below.

USING A FERRITE ROD AS THE COIL FORMER

The aerial coil could be wound onto a ferrite rod.

A piece of 10mm diameter ferrite rod of between 3 and 6 inches long (80 to 150mm) will be most suitable and will require between 50 and 90 turns of enamelled copper wire to provide coverage of the medium wave band: First make a paper tube that is held together with sticky tape that will easily slide up and down the ferrite rod. Then wind the coil over this with the windings neatly side by side. Make tapping points every 10 or 15 turns so that the aerial and diode tapping points can be adjusted.

Adjustments to the tuning range can be made by removing some wire from the coil so it is best to start off with too many turns and then work down. Fine adjustments can be made to the completed coil by sliding it up and down the ferrite rod.

AN IMPROVEMENT TO THE DESIGN

The crystal set above also has one small, but significant, improvement over the standard crystal set and that is an Aerial Trimmer. A trimmer is a variable capacitor, very similar to the tuning capacitor, except smaller and adjusted with a screwdriver.

The value of the trimmer is usually around 10 - 50pF, but if a small tuning capacitor is available that will probably be just as effective. In the absence of such a variable capacitor, individual fixed ceramic capacitors of e.g. 10pF, 50pF and 100pF can be tried in this position to judge which gives the best results with the particular aerial being used.

The trimmer capacitor adjusts the coupling to the tuned circuit, reducing the load of the aerial on the tuned circuit will improve the selectivity (Q), and it will be easier to separate stations. Again tapping points are used and I find this to be an excellent arrangement.

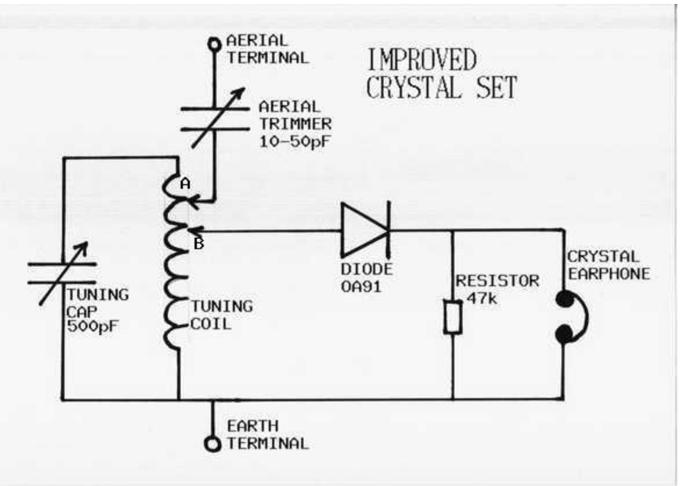

Improved Crystal Set design, with good selectivity

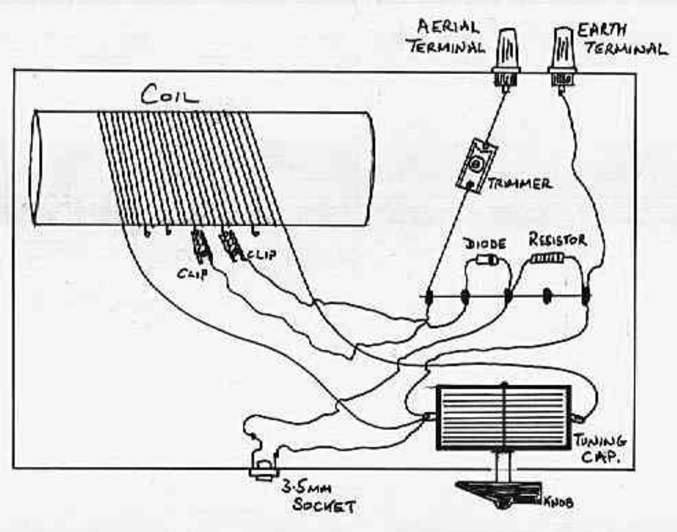

Layout Of The Crystal Set - Although this is soldered together an alternative to tagstrip would be a 5amp mains connector block so that components can be trapped in place with screws. See article below.

The picture on the right shows the general layout of the crystal set above. The coil is of approximately 70 turns is wound on the centre of a toilet roll, and has tapping points at 10 turn intervals.

The trimmer is soldered between the Aerial terminal and the piece of 5-way tag strip, and a wire goes from there to a croc' clip which is clipped onto a tap on the coil. The Diode is also soldered onto the tag strip, one end connected to a piece of wire going to a second croc' clip & connected to a tapping point on the coil, the other end of the diode is connected to the 3.5mm jack socket that the Crystal Earphone plugs into.

The 47k resistor is also connected to the earphone end of the diode and goes to earth, the earth terminal wire is soldered to the tag strip at this point too. The tuning capacitor has two terminals, one connected to each end of the coil, and one of them is also connected to earth as shown. [Where the wires cross over in the diagram, they do not touch and are not connected together].

LONG WAVES

In most areas around Europe and certainly around much of the UK you will be able to hear a Long Wave station. To receive Long Wave on a crystal set will require an aerial coil with a greater number of turns to increase its inductance.

As a good general guide a coil wound on a piece of 10mm diameter ferrite rod will require about 250 turns of enamelled copper wire: First make a paper tube that is held together with sticky tape that will slide up and down the ferrite rod. Then wind the 250 turn coil over this, the windings will have to be made over the top of each other. Make tapping points at, say, 50, 75 and 100 turns to tap the aerial and diode to.

As with the medium wave ferrite rod aerial, adjustments to the tuning range can be made by adding or removing some wire from the coil, and fine adjustments can be made to the completed coil by sliding it up and down the ferrite rod. The longer the ferrite rod the better and anything between 3 and 6 inches long (80 to 150mm) will be very good.

SHORT WAVES

If you like experimenting, then reducing the number of turns on the coil to say 10 to 30 will allow reception of the higher frequencies, the Short Waves. I have found that winding the coil around a 'ferrite rod' often works even better with short wave reception.

Obtain a ferrite rod about 7 to 15 cm long and about 1cm in diameter. Make a couple of small tubes of card, about 4cm long, that will fit tightly over the rod.

On one tube wind two coils using 0.5mm diameter enamelled copper wire - one coil of about 30 turns and a second one of 2 or 3 turns wound over the top of the first. Secure the windings in place with Sellotape.

On the other card tube wind a similar coil, but use about 15 turns for the first coil and for the second coil wind about 3 to 4 turns over the top, and secure with Sellotape tape.

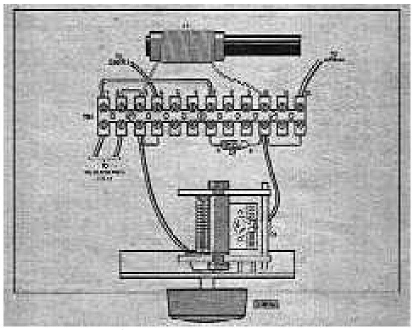

Example of a layout using a connector block to wire up a crystal set

These coils will provide coverage of short wave in two bands using the first coil for the longer wavelengths, typically 60 to 31 metre bands and the second coil for the shorter wavelengths typically 25 to 19 metre band. Wire up the circuit as shown in the circuit diagram below.

USING A TOROID INDUCTOR FOR SHORT WAVES

Even better selectivity performance can be achieved by winding the inductors (coils) on a ferrite toroids (T50-2 yellow, or green will do). The aerial trimmer need not be used if selectivity and sensitivity is found to be adequate. It's all about experimenting, and I find it best to use a trimmer or small coupling capacitor to obtain the best selectivity.

Up to 30 turns of 0.5mm enamelled copper wire can be used for the longer short waves below 10 MHz, while a winding of around 15 turns will provide coverage of the shorter short waves above 10MHz.

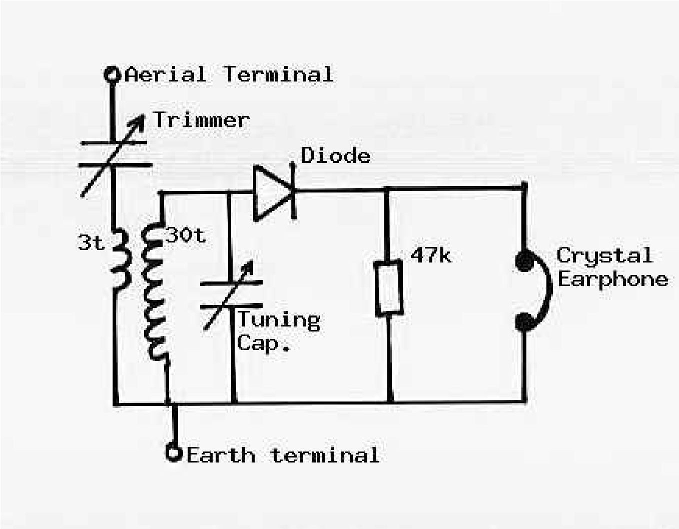

The circuit diagram of the Short Wave Crystal Set

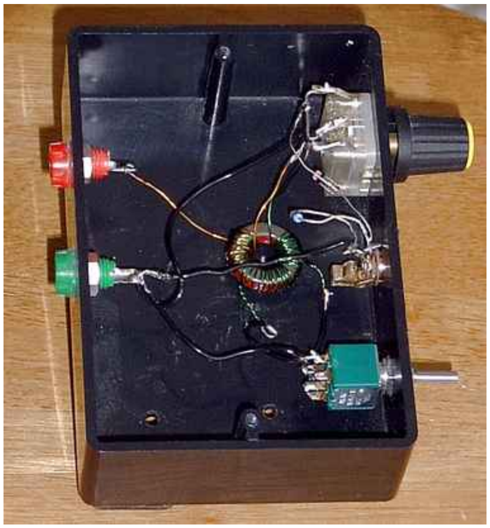

A completed SW Crystal Set using a toroid inductor. Note: the main winding has a tap to allow the switch to short part of the winding and thereby give two ranges.

AUSTRALIAN DESIGN

Moving back to the Medium Waves, here is a circuit for a very interesting Australian design that promises extremely good station separation (selectivity), and having built it I can vouch for that claim, it's really excellent.

I receive three national stations and three local stations at my location with excellent clarity using a modest antenna and standard crystal earphone.

The coil is different to the other crystal sets described above, it is much bigger at 3¼ inches (8cm) diameter and 5inches (12cm) long. I made my coil former out of the cardboard from a breakfast cereal box - just like Blue Peter!

The design is often referred to as The Mystery Crystal Set, by Proton.

The front panel of the Australian Crystal Set

Two distinct coils are wound on it, the first one consists of about 50 turns of 24 s.w.g (approx) enamelled copper wire. The second coil is 25 turns, very close wound right over the top of the first coil using 30 s.w.g. (approx) wire, try to get this second coil wound in between the windings of the first, for better inductive coupling.

Then carefully wire up the set according to the diagram. Notice that the tuned circuit is not connected to earth and has no *direct* connection to the detector circuit. The detector circuit *is* connected to earth however. The two aerial terminals offer alternative selectivity performance, terminal A gives very good selectivity while B is very wide. I never bother with B.

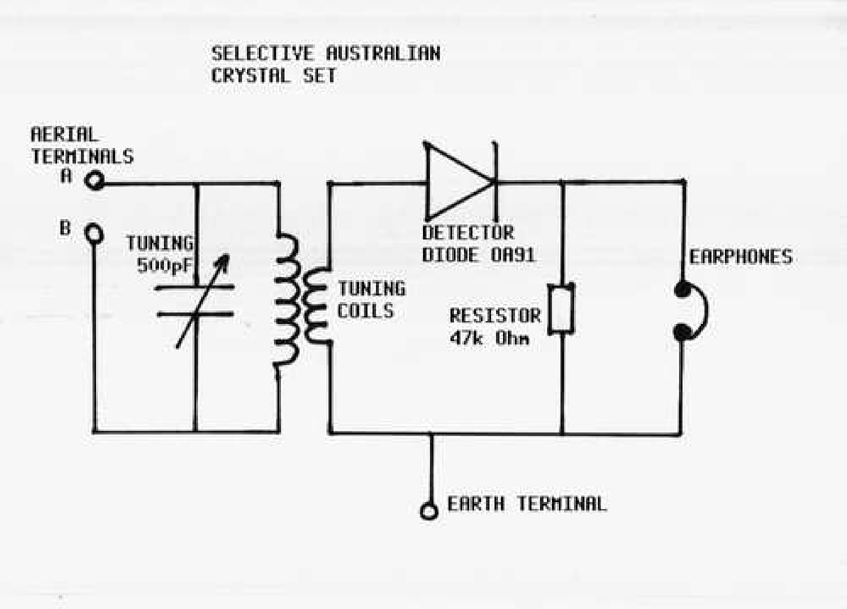

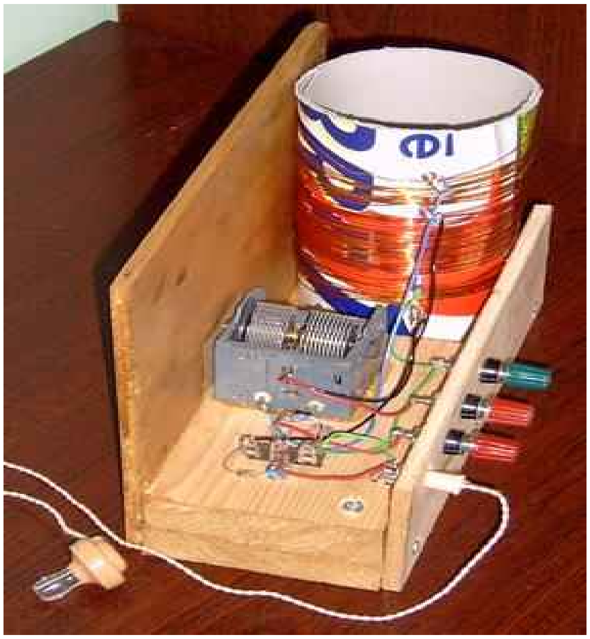

Make the coil carefully and wire up this crystal set according to the circuit diagram opposite and you will be rewarded with a really high performance crystal set of a type that was used in the very early days of broadcasting in 1930's in Australia.

This is probably my favourite crystal set!

THE DENCO PCC1 COIL

The PPC1 coil was a commercially manufactured by Denco Clacton Ltd and was popular among hobbyists not keen on going to the bother of winding their own fiddly little coils. As a child I wanted try one of these coils and sent away for one by mail order. It arrived a few days later in a little cloth bag, like a miniature pump bag, with protective wrapping inside.

The coil windings are entirely enclosed in what I can only describe as a cylindrical ferrite 'shell', the four very thin connecting wires exiting, two either side, from small apertures in the 'shell'. The performance of the circuit shown below I seem to remember was quite pleasing. Unfortunately I cannot find the set or the PPC1 coil at the moment, but here is a reproduction of the circuit diagram and data:

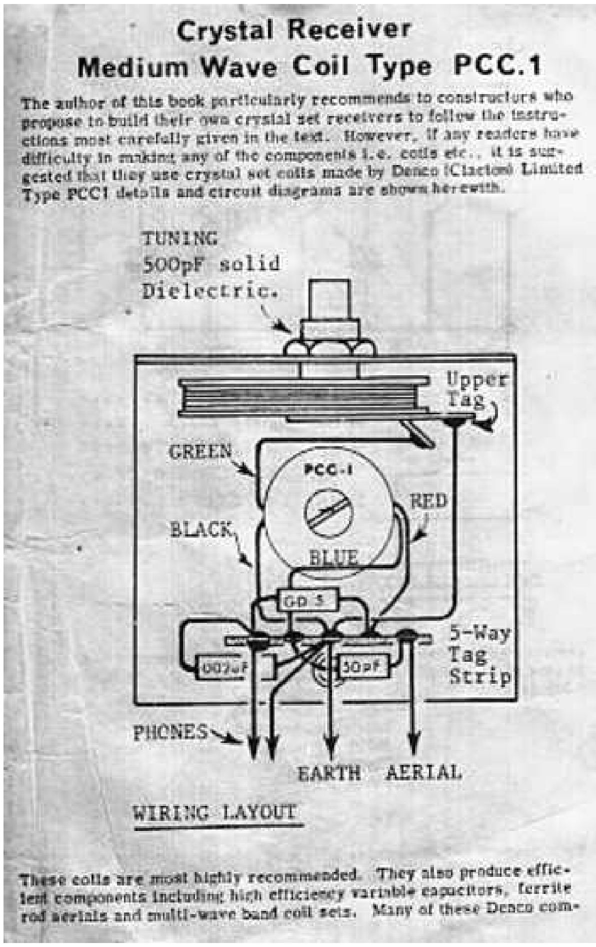

Crystal Receiver
Medium Wave Coil Type PCC.1

The author of this book particularly recommends to constructors who propose to build their own crystal set receivers to follow the instructions most carefully given in the text. However, if any readers have difficulty in making any of the components i.e. coils etc., it is suggested that they use crystal set coils made by Denco (Clacton) Limited Type PCC1 details and circuit diagrams are shown herewith.

TUNING
500pF solid
Dielectric.
Upper
Tag
GREEN
PCC-1
RED
BLACK
BLUE
GD 5
.002uF
50pF
5-Way
Tag
Strip
PHONES
EARTH
AERIAL

WIRING LAYOUT

These coils are most highly recommended. They also produce efficient components including high efficiency variable capacitors, ferrite rod aerials and multi-wave band coil sets. Many of these Denco components can be obtained from your local radio component dealer but, in case of difficulty, write direct to this company at the following address:-

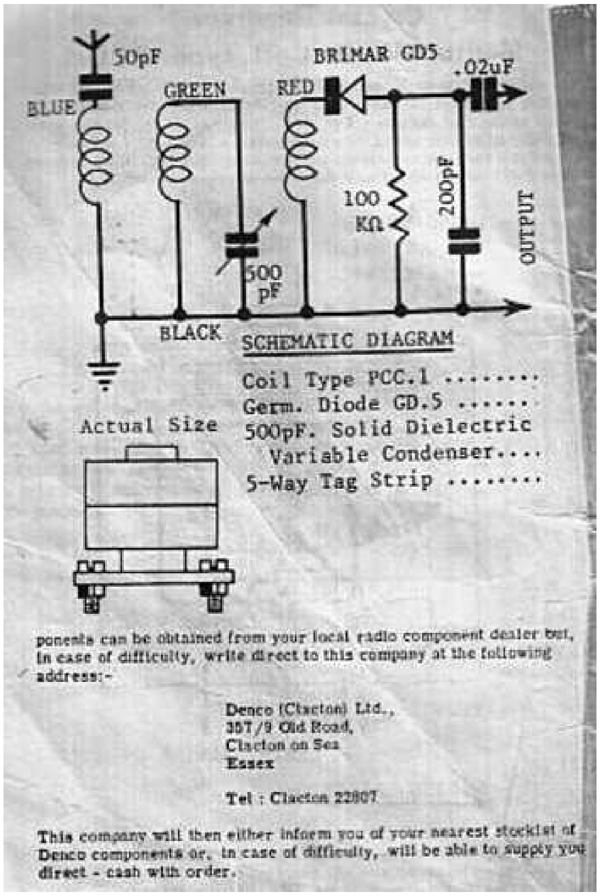

50pF
BRIMAR GD5
.02uF
BLUE
GREEN
RED
100
KΩ
200pF
500
pF
OUTPUT
BLACK
Actual Size

SCHEMATIC DIAGRAM

Coil Type PCC.1 ....
Germ. Diode GD.5 ......
500pF. Solid Dielectric
Variable Condenser....
5-Way Tag Strip ....

Denco (Clacton) Ltd.,
357/9 Old Road,
Clacton on Sea
Essex

Tel : Clacton 22807

This company will then either inform you of your nearest stockist of Denco components or, in case of difficulty, will be able to supply you direct - cash with order.

THE REPANCO DRR2 COIL

I recently rediscovered an old Repanco DRR2 Longwave / Mediumwave coil that must have been kicking around in my junk box since the 1970's.

The DRR2 coil was made by Repanco in Coventry. It came with a page of suggested circuit diagrams which I thought had been lost to the mists of time, but it recently came to light again, so I have now copied it below.

Once again I included an aerial trimmer which can be adjusted to improve selectivity.

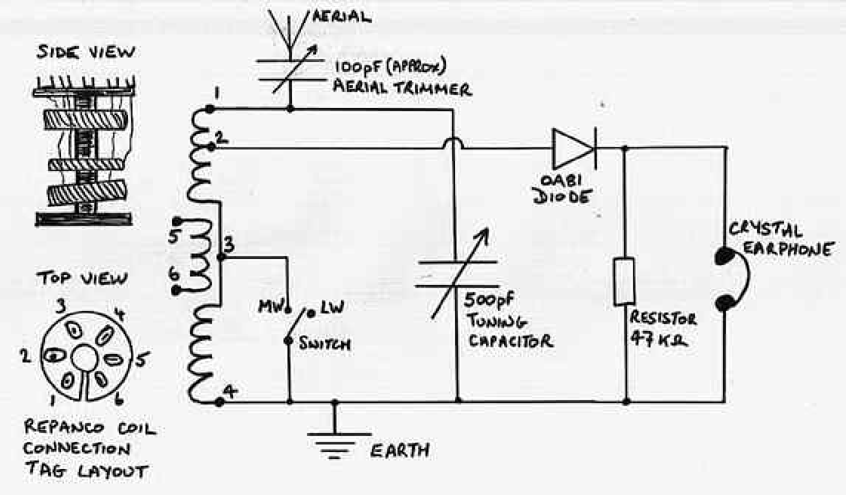

AERIAL
100pF (APPROX)
AERIAL TRIMMER
SIDE VIEW
1
2
3
4
5
6
OA81
DIODE
CRYSTAL
EARPHONE
TOP VIEW
MW
LW
SWITCH
500pF
TUNING
CAPACITOR
RESISTOR
47KΩ
REPANCO COIL
CONNECTION
TAG LAYOUT
EARTH

The circuit diagram of the crystal set using the Repanco DRR2 coil

Repanco Ltd was formed by two ex-army signals engineers and from the earliest days of radio supplied crystal set kits and coils to radio construction enthusiasts.

The Repanco DRR2 coil was for medium wave and long wave intended for use in when building simple crystal set and valve radio

A 'lash-up' of the Repanco crystal set

circuits.

It consists of three coils; a Medium Wave coil at the top that includes a tapping point (for the aerial); a coupling coil or tickler in the middle; a lower coil which can be connected in series with to top coil to provide Long Wave reception.

I have built a quick crystal set with the coil and it provides good reception with excellent selectivity, so it must have a very good Q factor.

The Repanco DDR2 coil was provided with a simple Foolscap size information sheet that showed four different radio circuits. Sadly the sheet does not give a huge amount of information and my copy is now rather tatty and faded - it is copied below:

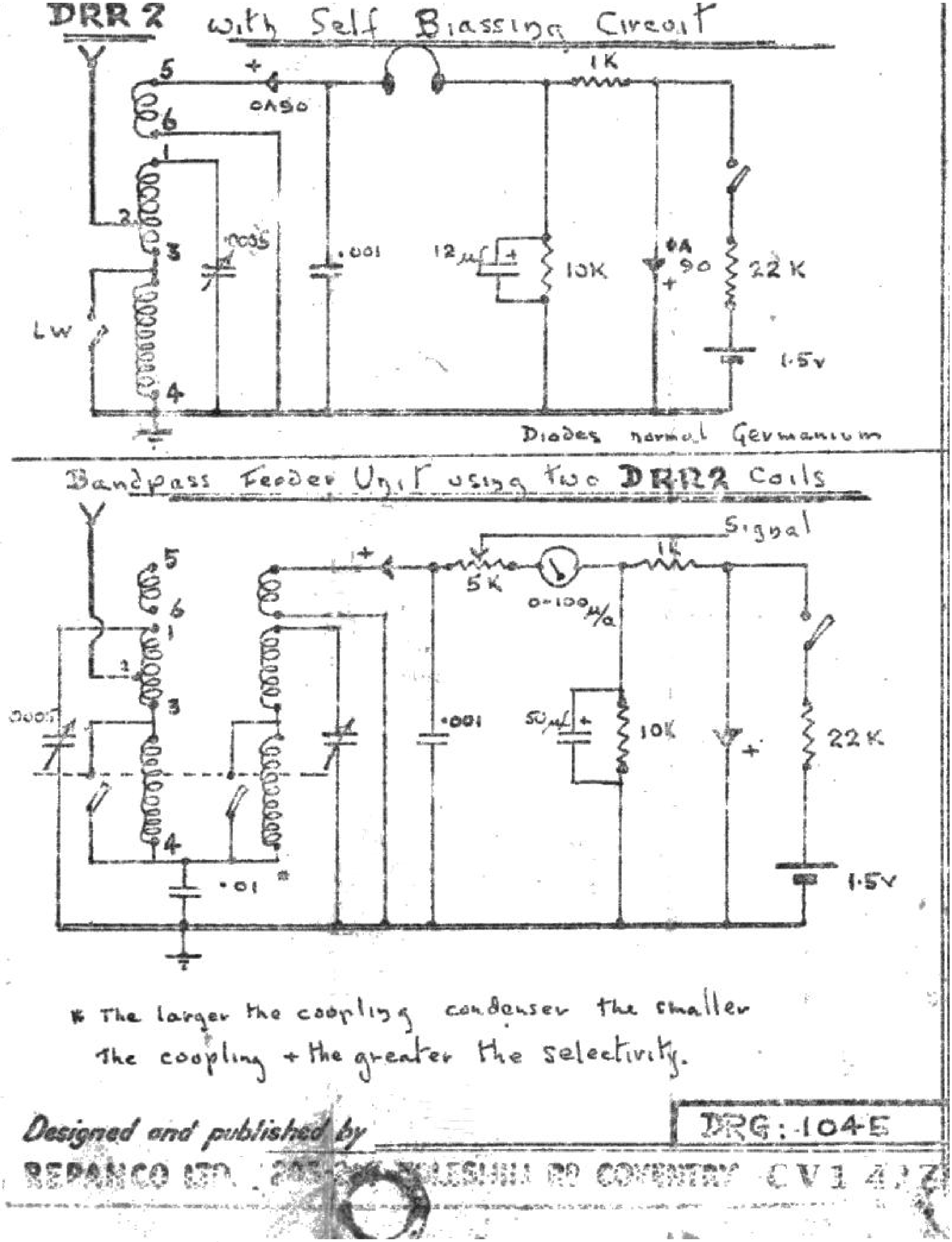
DRR 2 with Self Biassing Circuit
5
6
1
2
3
4
LW
+
OA90
1K
.0005
.001
12µF
+
10K
OA 90
+
22K
1.5v
Diodes normal Germanium
Bandpass Feeder Unit using two DRR2 Coils
Signal
5
6
1
2
3
4
.0005
+
5K
0-100µ/a
1K
.001
50µF
+
10K
+
22K
1.5V
.01
*
DRG: 104E

* The larger the coupling condenser the smaller the coupling + the greater the selectivity.

Designed and published by

REPANCO LTD. [illegible] FOLESHILL RD COVENTRY CV1 4

Simple crystal radio type circuits using the Repanco DRR2 coil

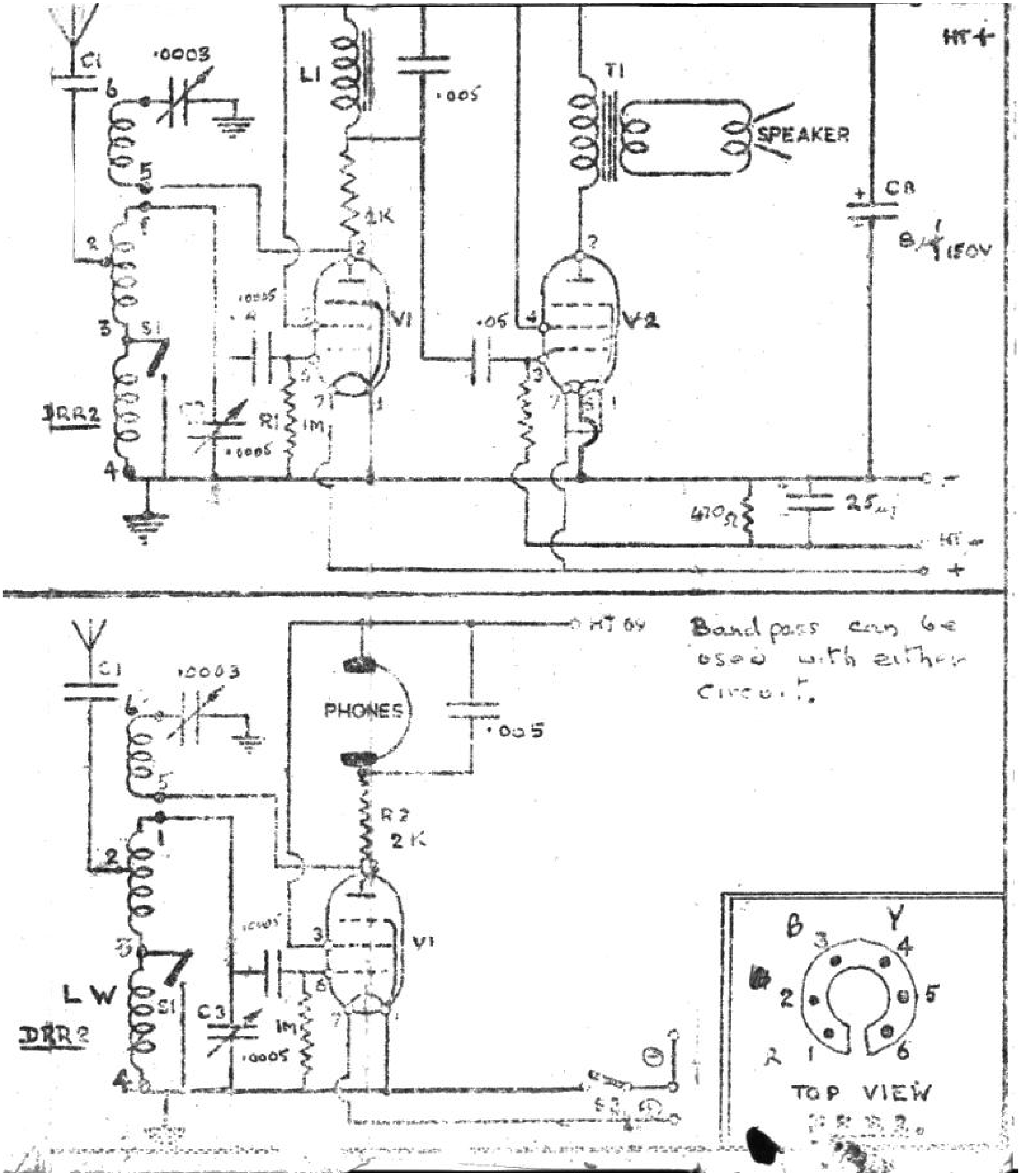
C1
·0003
L1
·005
T1
HT+
SPEAKER
C8
8/
150V
2K
V1
V2
·0005
·05
R1
1M
DRR2
420Ω
25
HT−
HT 6V
Band pass can be used with either circuit.
PHONES
·005
R2
2K
LW
S1
C3
·0005
1M
S2
TOP VIEW
DRR2
B
Y
R

Valve circuits using the Repanco DRR2 coil

Repanco Ltd no longer produces radio coils and crystal set kits for the radio construction enthusiast, as it did in the early days of radio. In 1986 it was renamed Repanco Bartlett Ltd when it merged with Bartlett Electronics. The company moved from the Foleshill Road to Unit 24, Albion Industrial Estate, Endemere Road, Coventry CV6 5NT and now specialises in transformers and wound components and can design and manufacture to commercial customer requirements, their website is: http://www.repancobartlett.co.uk/

CRYSTAL EARPHONES

Here is a good idea and well worth trying, to maximise the use of sound output from your crystal set why not use dual crystal earphones? Having an earphone in each ear helps to block out extraneous noises helping the listener to better concentrate on any weaker stations received.

Using the circuit below, one earphone makes use of one half cycle of the radio wave while the second earphone uses the other half cycle of the wave that would have previously gone to waste when using just one diode. Ensure that the diodes are connected up according to the diagram i.e. one diode is connected the opposite way round to the other.

Also try to make sure that the diodes and crystal earphones are similar to obtain the best results. (You could simply connect two crystal earphones to the same terminals of the single diode, but this would not be as efficient and the sounds would be much quieter.)

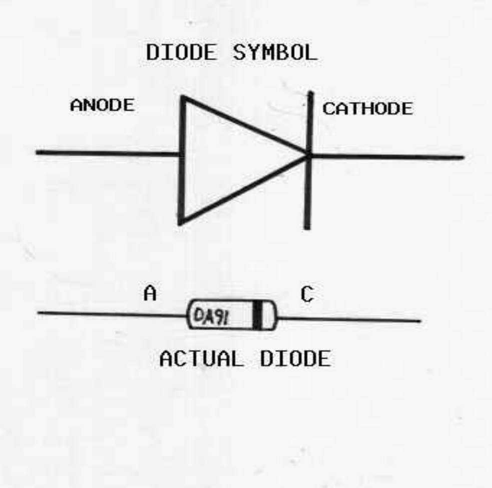
DIODE SYMBOL
ANODE
CATHODE
A
C
OA91
ACTUAL DIODE

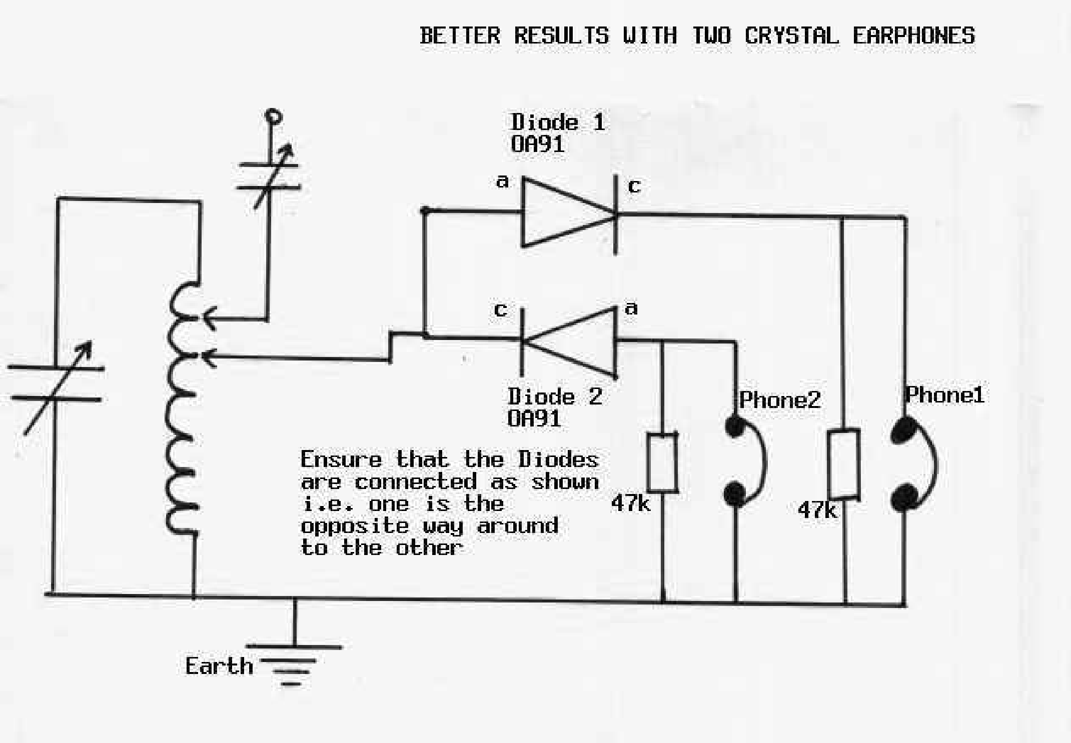
BETTER RESULTS WITH TWO CRYSTAL EARPHONES
Diode 1
OA91
a
c
c
a
Diode 2
OA91
Ensure that the Diodes
are connected as shown
i.e. one is the
opposite way around
to the other
47k
Phone2
47k
Phone1
Earth

A note about Crystal Earphones: It will be worthwhile buying several different ones from different sources as performance varies between manufacturers quite markedly. I have found the ones marked 'Japan' on the back are the most sensitive and therefore loudest, whereas ones marked 'Receiver' 'Taiwan' are often a little less sensitive and therefore quieter and sometimes more 'tinny' sounding.

As mentioned previously it has been noted that the OA47 diode will be of particular interest since it has the lowest forward bias voltage of any of the common diodes available. This will make the crystal set somewhat more sensitive and therefore louder. The US equivalent of the British OA47 is the IN34.

SPIDERS WEB' COIL

Here is an interesting concept sent in by Chris Dorna of the Vught North Scouts in the Netherlands. It is a crystal set made out of a coil wound in the form of a spiders web:

See more HERE.

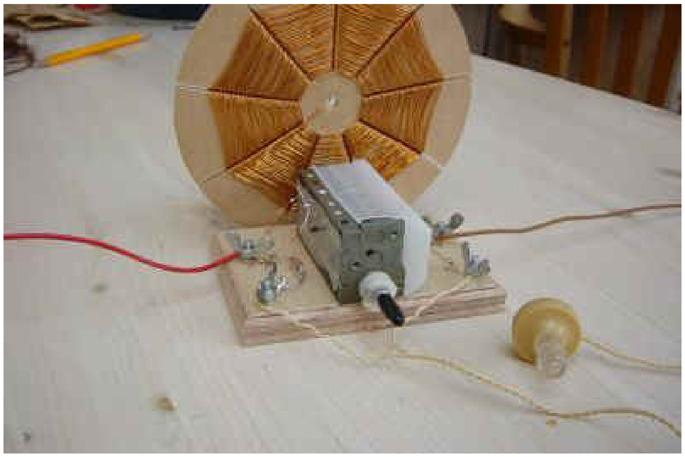

Chris Dorna's Crystal Set with Spider Web Coil

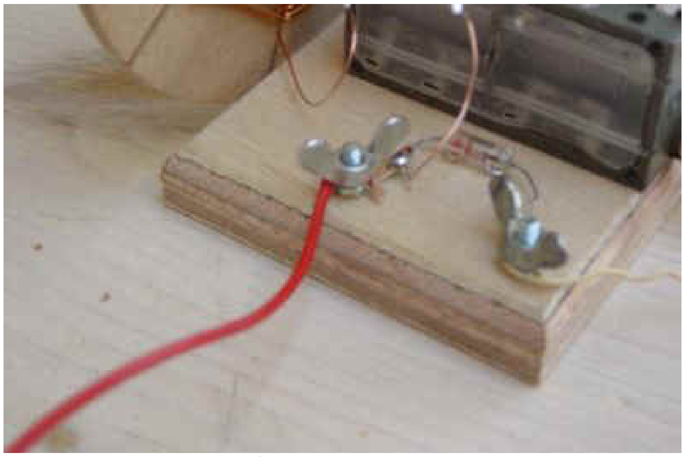

Detail of the germanium diode

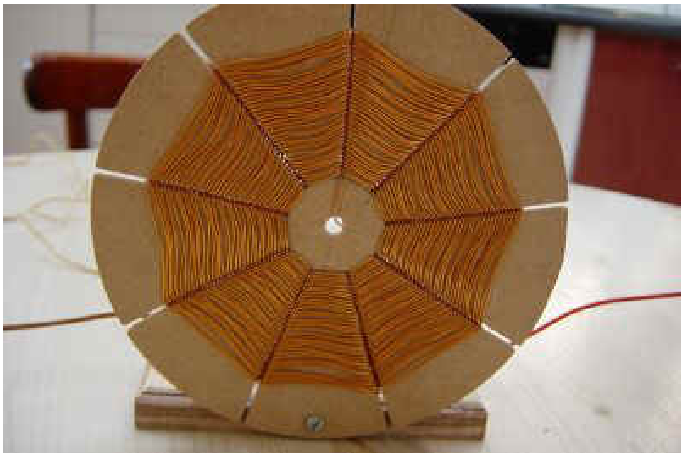

Close up of Chris Dorna's Spider Coil

LOOP THE LOOP!

A crystal set can also be made that does not need a large long wire aerial. If you have ever made a loop aerial for medium wave or long wave DX-ing, then it is a simple matter to add a diode, resistor and a socket to connect a crystal earphone that will allow reception of nearby stations.

See my section on Loop Aerials and ATU's for more constructional details.

A portable loop aerial, that incorporates a crystal set

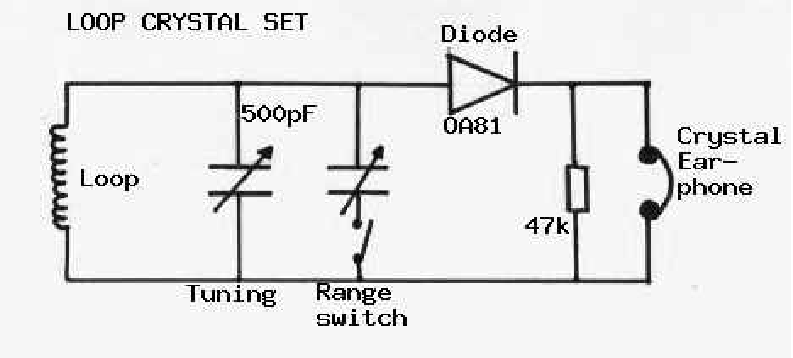

The circuit diagram of the Loop Crystal Set. The loop is 10 turns of 7/0.2mm 'hook-up' wire wound on a 40cm (17") former made of attractive plastic edging strip available from many DIY stores. The loop is very directional in its pick up pattern, which can help eliminate interference from some stations by rotating the loop. The switch and additional capacitor allow tuning of the lower medium wave band from about 650 to 520 kHz. Having a loop with 50 to 60 turns of wire will tune into the Long Wave band.

DIODES - For Crystal Set Use - some notes by Felix Scerri

Germanium diodes for crystal set use.

Although I'm a fan of these new silicon schottky BAT 46 diodes, good germanium diodes still have a lot to offer, especially in terms of 'weak signal' sensitivity. Last night I did an experiment.

I sorted through quite a few of my hundreds of acquired random germanium diodes looking for particularly 'sensitive' ones. I tested this by tuning in a weak AM station and comparing the detected DC output level and also the apparent 'loudness' of the audio signal.

Even amongst germanium diodes of the same type, there was enormous variation all the way from excellent to poor! For very weak signals, germanium diodes 'detect' in the 'square law' region below the diode conduction 'knee', in a rather different part of the curve than with much stronger signals (way beyond the diode knee).

When testing germanium diodes for weak signal sensitivity, the inherent capacitance of the diodes is also a factor, and the 'tuning' may change somewhat and will need to be readjusted with every diode tested!

In the end, out of a large number of germanium diodes tested, I found three or four germanium diodes with excellent weak signal sensitivity and the rest were poor. One other interesting thing, good germanium diodes 'sound' different, rather more 'rounded and smoother' than the schottky's which tend to sound mercilessly clean, almost clinical. I also found almost no variation in weak signal sensitivity with my BAT 46 schottky diodes. Take your pick!

Regards, Felix Scerri VK4FUQ 14/03/2012

SOLDERLESS CONSTRUCTION IDEAS

For a novice the use of a soldering iron may seem a bit daunting at first and while the most reliable results will be obtained with a good soldered joint using a tag strip as shown below, the circuits can still be made without the use of a soldering iron.

5 WAY TAG STRIP

36 WAY TAG STRIP - TWO ROWS

The very simplest circuits could be wired together ,with a little ingenuity, with the component wires being held together in the grip of solderless crocodile clips, whereby the connecting hook-up wire is fixed to the croc' clip by a screw rather than solder.

For more the slightly more complex circuits a plastic Terminal Block (sometimes referred to as a choc' or chocolate block) can be utilised very effectively indeed. These are used in mains wiring and are available in various sizes; 2 Amp, 5 Amp, 15 Amp and 30 Amp. The 5 and 15 Amp Terminal Blocks I have found to be the most suitable. The various component wires can be trapped securely with the screw at each junction point. This method also makes it easy to change the components around when experiment with different circuits. See The EXPERMENTAL CRYSTAL SET for more details in Part 5**.**

'CHOCOLATE' TERMINAL BLOCK

The Ladybird book called 'Making A Transistor Radio' (also shown on the TRF Radio pages) detailed a very novel approach using brass screws with screw-cups to trap the component wires at each junction point:

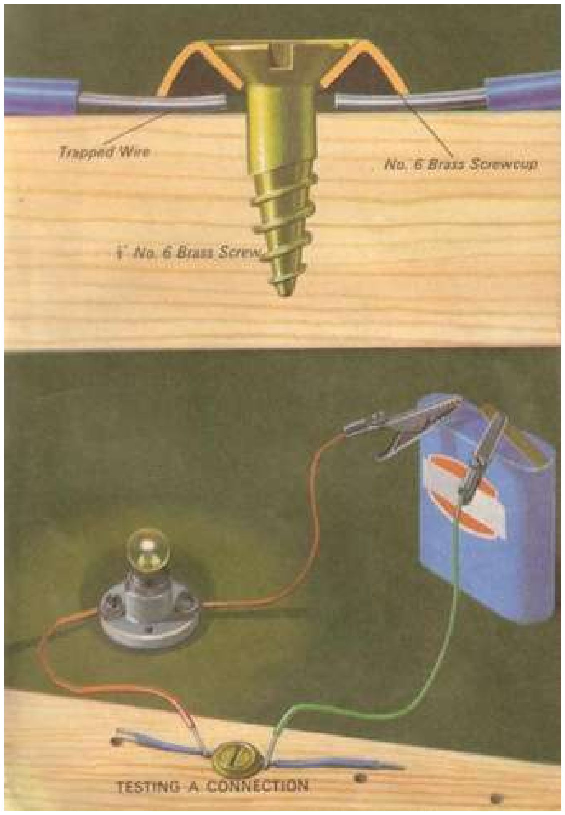
Trapped Wire
No. 6 Brass Screwcup
¼" No. 6 Brass Screw
TESTING A CONNECTION

THE BRASS SCREW AND SCREW-CUP METHOD OF CONSTRUCTION

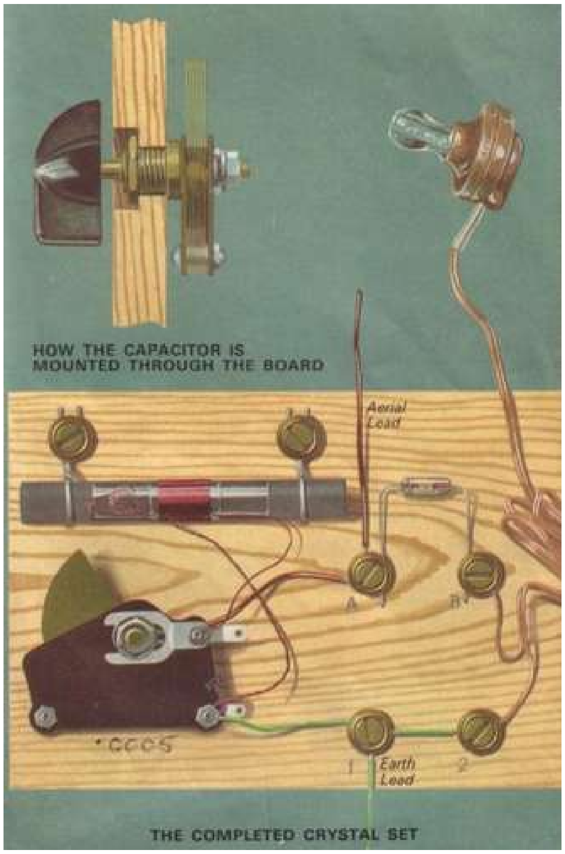
HOW THE CAPACITOR IS MOUNTED THROUGH THE BOARD
Aerial Lead
A
B
CCC5
1
Earth Lead
2
THE COMPLETED CRYSTAL SET

A VERY SIMPLE CRYSTAL SET USING THE BRASS SCREW AND CUP METHOD

Crystal Sets Part Part 3 >

No AM radio stations or transmitters in your locality or country?

Spitfire
AM Transmitter

Has your local medium wave broadcast station closed or been moved to VHF/FM or Digital? Don't worry. You can still build and experiment with crystal sets and TRF radios by also buying or even building a simple low power AM transmitter. So, not only can you use your crystal sets but you can also run your own radio station that can be heard in and around your home - playing the music or programmes that you want to hear!

SSTRAN AMT3000: Superb high fidelity medium wave AM transmitter kits from SSTRAN. Versions available for 10kHz spacing in the Americas (AMT3000 or AMT3000-SM) and 9kHz spacing in Europe and other areas (AMT3000-9 and AMT3000-9SM). Superb audio quality and a great and well designed little kit to build: http://www.sstran.com/pages/products.html

POWER
GAIN
MODULATION
COMPRESSION
SSTRAN
AMT3000

http://www.sstran.com/

Other AM transmitters available:

Spitfire & Metzo: Complete, high quality ready built medium wave AM Transmitters from Vintage Components: http://www.vcomp.co.uk/index.htm Vintage Components offer a choice of the high quality Spitfire and Metzo transmitters:

SPITFIRE AM Medium Wave Transmitter with 100 milliwatt RF output power:

http://www.vcomp.co.uk/spitfire/spitfire.htm

Spitfire
AM Transmitter

METZO AM Medium Wave Transmitter with built in compressor:

Gain
Comp
Mod
Metzo

http://www.vcomp.co.uk/metzo/metzo.htm

AM88 LP: A basic AM transmitter kit from North County Radio.
http://www.northcountryradio.com/Kitpages/am88.htm

LINKS:

BOWOOD ELECTRONICS - A friendly, helpful and very speedy source for many of your electronic components at prices that won't frighten your wallet!

THE FOXHOLE and P.O.W RADIOS - Simple crystal set receivers used by soldiers during the war and by prisoners of war (P.O.W.'s).

VINTAGE COMPONENTS - A great resource for crystal sets, components, valve radio kits and medium wave AM transmitters!

6V6 - Electronic Nostalgia and Vintage Components

Crystal Sets Part 3 >

Crystal Sets (Part1) | **Spider's Web Crystal Set (Part 3**

Crystal Set by Kenneth Rankin (Part 4) | **Experimental Crystal Sets (Part 5)** | **Crystal Radio Links**

^Top Of Page

Amateur Radio
M0MTJ
Visit my amateur radio pages
M0MTJ

Visit my Amateur Radio Pages >

Home | Contact | Site Map | Radio Stations & Memorabilia | Amateur Radio

Crystal Sets Introduction | Resistor & Capacitor Conversion Tables

www.mds975.co.uk © 2004 - 2014

Mike's Electronic Parts, LLC
Simple Crystal Radio Kit 2

https://www.mikeselectronicparts.com/

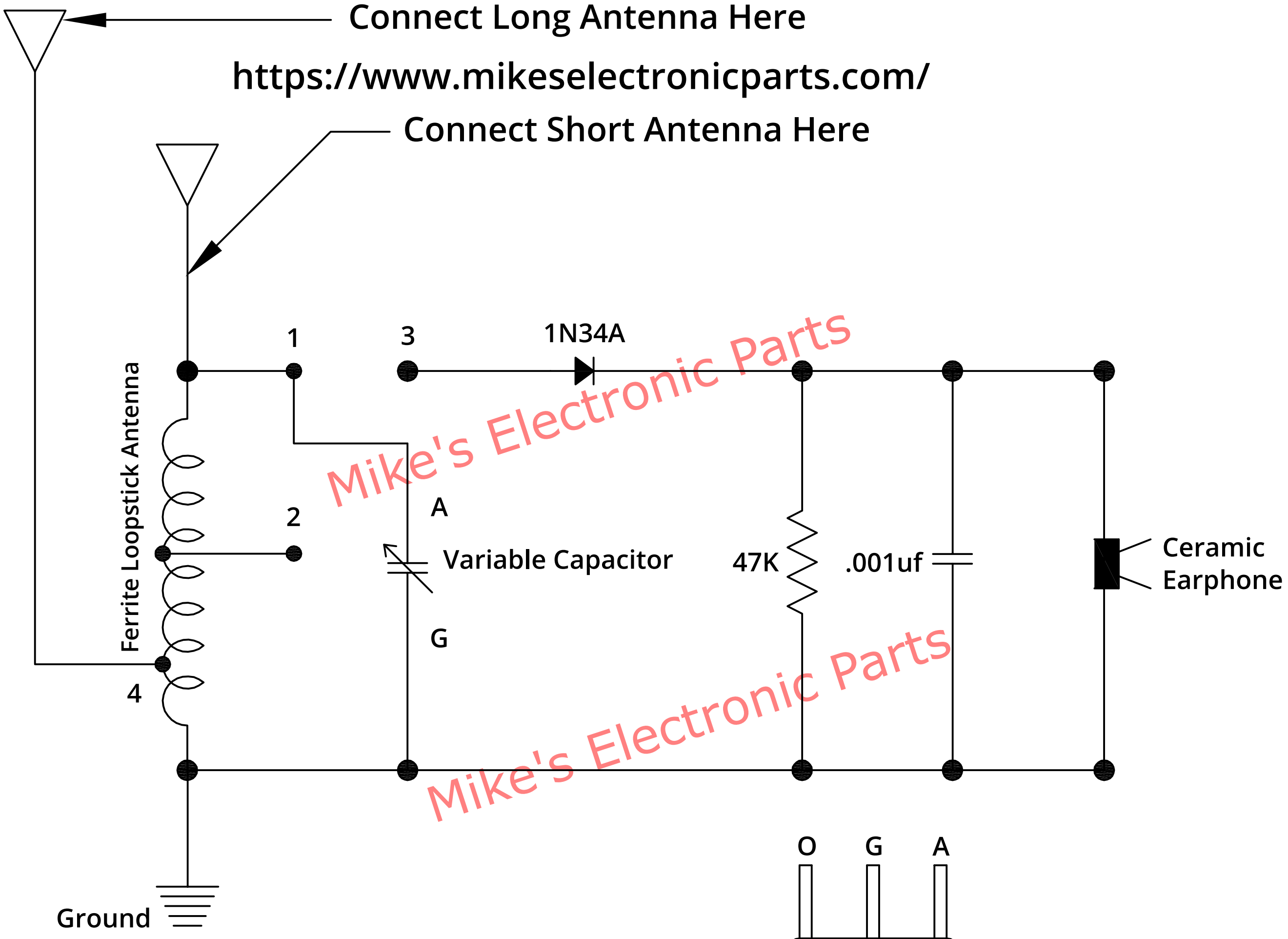

Connect Long Antenna Here
Connect Short Antenna Here
1
3
1N34A
Ferrite Loopstick Antenna
2
A
Variable Capacitor
G
47K
.001uf
Ceramic
Earphone
4
Ground

Mike's Electronic Parts
Mike's Electronic Parts

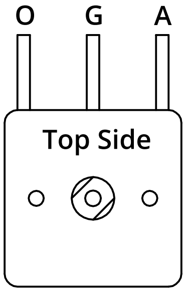

O
G
A
Top Side

Variable Capacitor

Parts List
1 - Ferrite Loopstick Antenna
1 - Variable Capacitor
1 - Germanium Diode 1N34A
1 - .001uf Capacitor Marked 102
1 - 47K Resistor
1 - 20 Million Ohm Ceramic Earphone
Antenna and Ground wire not include with parts

For short antenna connect a 15ft to 30ft length of wire to "1". For long antenna 40ft to 100ft connect to "4". It is best to use a long wire antenna. If you are not using the taps do not cut the wires. Doing so will cause the coil not to work. Various connections are, Connect 1 to 3 or Connect 2 to 3. Most volume can be had at 1 to 3 less volume but better selectivity connect 2 to 3 and the antenna connected at 1. It is best to solder the coil wire, not doing so may cause a bad connection and the radio will not work.
Drawn By: Scott Lowe
This Kit can be bought at my web site Part Number CRK#2 Kit.

https://www.mikeselectronicparts.com/

THE SMITHSONIAN INSTITUTION

The Smithsonian Institution is home to more than 141 million objects, ranging in size from insects and diamonds to locomotives and spacecraft. It is the world's largest museum complex, comprising 15 museums and galleries and the National Zoo in Washington DC, and two additional museums in New York City. Millions of visitors each year visit the nation's capital to view such treasures as the Hope Diamond, the Star Spangled Banner, and the Wright Flyer. A broad range of exhibits ensures a fun and educational experience for young and old alike.

One of the world's leading scientific research centers, the Institution has facilities in eight states and the Republic of Panama. Research projects in the arts, history and science are carried out by the Smithsonian all over the world. Some of the Smithsonian's research centers include the Smithsonian Astrophysical Observatory in Cambridge, Massachusetts, the Smithsonian Marine Station at Link Port, in Florida, and the Smithsonian Tropical Research Institute, in Panama.

For membership information of pre-visit planning material, write or call the Visitor Information and Associates Reception Center, Smithsonian Institution, Washington, D. C., 20560, (202) 357-2700 (voice), (202) 357-1729 (TTY). You may also visit the Smithsonian through our web site, www.si.edu.

HISTORY

James Smithson (1765 –1829), a British scientist, drew up his will in 1826 naming his nephew, Henry James Hungerford, as beneficiary. Smithson stipulated that, should the nephew die without heirs (as he did in 1835), the estate would go to the United States to found "at Washington, under the name of the Smithsonian Institution, an establishment for the increase and diffusion of knowledge..."

On July 1, 1836, Congress accepted the legacy bequeathed to the nation by James Smithson, and pledged the faith of the United States to the charitable trust. In 1838, following approval of the bequest by the British courts, the United States received Smithson's estate—bags of gold sovereigns—then the equivalent of $515,169. Eight years later, on August 10, 1846, an Act of Congress signed by President James K. Polk, established the Smithsonian Institution in its present form and provided for the administration of the trust, independent of the government itself, by a Board of Regents and Secretary of the Smithsonian.

SMITHSONIAN MUSEUMS, GALLERIES AND ZOOS

Smithsonian Institution [illegible]
Anacostia Museum
Arthur M. Sackler Gallery
Arts and Industries Building
Cooper-Hewitt, National Design Museum
Freer Gallery of Art
Hirshhorn Museum and Sculpture Garden
National Air and Space Museum
National Museum of African Art
National Museum of American History, Behring Center
National Museum of the American Indian
National Museum of Natural History
National Portrait Gallery
National Postal Museum
National Zoological Park
Renwick Gallery
S. Dillon Ripley Center
Smithsonian American Art Museum

SMITHSONIAN®

CRYSTAL RADIO™

WARNING!
ONLY FOR USE BY CHILDREN OVER 8 YEARS OLD.

READ THE INSTRUCTIONS BEFORE USE,
FOLLOW THEM AND
KEEP THEM FOR REFERENCE.
STORE THE SET OUT OF REACH
OF SMALL CHILDREN.

PLEASE KEEP A NOTE OF OUR NAME AND ADDRESS DETAILS FOR FUTURE REFERENCE. IN EUROPE CONTACT:

CE

NSI SIMM GAmbH
D 76162 KARLSRUHE
GERMANY
49-0721-9584-116

PLEASE BE SURE TO READ THE ADVISE FOR SUPERVISING ADULTS AND THE SAFETY RULES CONTAINED IN THIS BOOKLET.

©2002 SMITHSONIAN® INSTITUTION
NATURAL SCIENCE INDUSTRIES, LTD.
910 ORLANDO AVE.
WEST HEMPSTEAD, N.Y. 11552

PRINTED IN USA

ITEM NO. 3261-08

ADVICE FOR SUPERVISING ADULTS:

- READ AND FOLLOW THESE SAFETY INSTRUCTIONS, THE SAFETY RULES AND KEEP THEM FOR REFERENCE.

- SUPERVISING ADULTS SHOULD DISCUSS ANY WARNINGS AND SAFETY INFORMATION WITH THE CHILD BEFORE COMMENCING THE ACTIVITIES.

SAFETY RULES:

- DO READ THE INSTRUCTIONS BEFORE USE, FOLLOW THEM AND KEEP THEM FOR REFERENCE.
- DO KEEP YOUNG CHILDREN AND ANIMALS AWAY FROM THE ACTIVITY AREA.
- DO STORE THE SET OUT OF REACH OF YOUNG CHILDREN.

Radio Technology

Facts about early crystal development

Several inventors receive credit for developing a "crystal detector", a device which can pass current better in one direction than the other in an electrical circuit. In Germany, in the 1870's, Karl F. Braun noticed this property in certain mineral substances. Commercial development required a technology that would be suitable. This came with the advent of wireless communication and radio after 1900.

In Japan, Wichi Torikata investigated many minerals, including zincite, pyrolusite, iron pyrites and galena. In the United States, Greenleaf W. Pickard, associated with the Wireless Specialty Apparatus Co. of Boston, conducted extensive experiments as well. He became well known for the PERIKON Detector which employed zincite and chalcopyrite. The early 20th century work was generally done between 1900 and 1912. Included among the scientists is General H.H.C. Dunwoody of the U.S Army who developed a carborundum detector.

Until radio broadcasting began in 1920 (Station KDKA Pittsburgh) experimenters utilized the crystal detector for radiotelephone communications and for reception of certain Morse Code signals. There was also extensive use by the U.S.Navy and other maritime services.

The modern crystal diode, such as provided with your set, is a spin-off of radar technology developed during World War II. The basic principle, however, remains the same..i.e. to remove the AUDIO (speech, music or Morse-Code) from the radio frequency carrier wave. In one early Pickard crystal design (left) a stiff metal point is adjustable over the crystal surface. In a later version a "Catswhisker" spring impinged on a piece of galena (right) and was varied to give a louder signal.

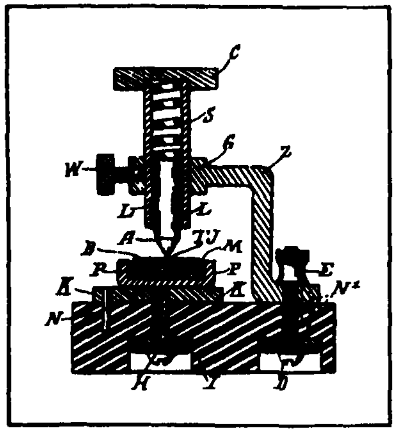

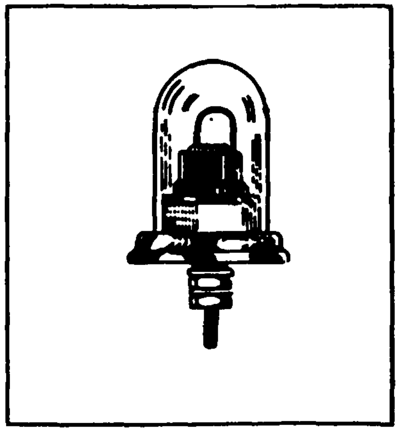

Your Smithsonian Smart Lab Crystal Radio Set contains the following items

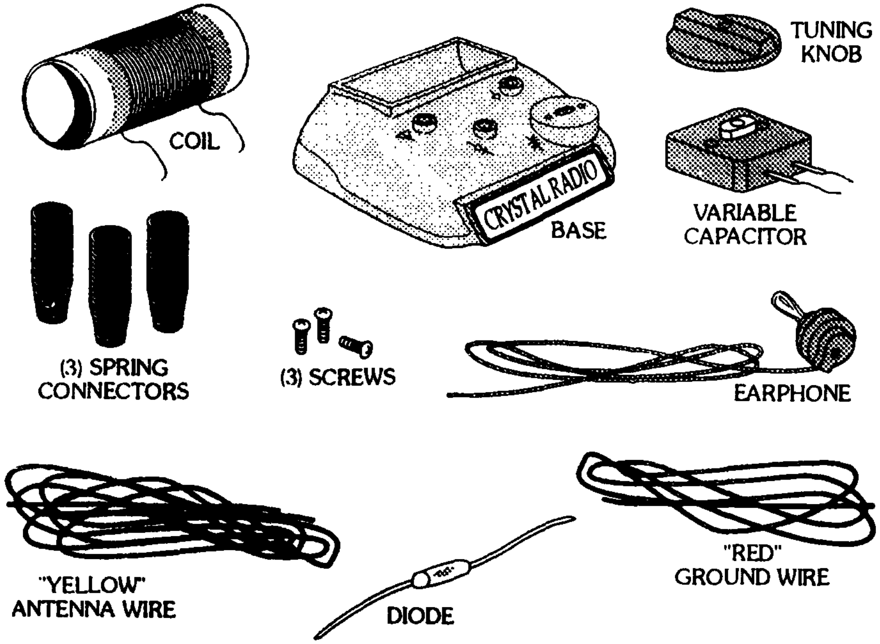
TUNING KNOB
COIL
CRYSTAL RADIO
BASE
VARIABLE CAPACITOR
(3) SPRING CONNECTORS
(3) SCREWS
EARPHONE
"YELLOW" ANTENNA WIRE
DIODE
"RED" GROUND WIRE

Shown below is a typical wiring diagram of your crystal radio.

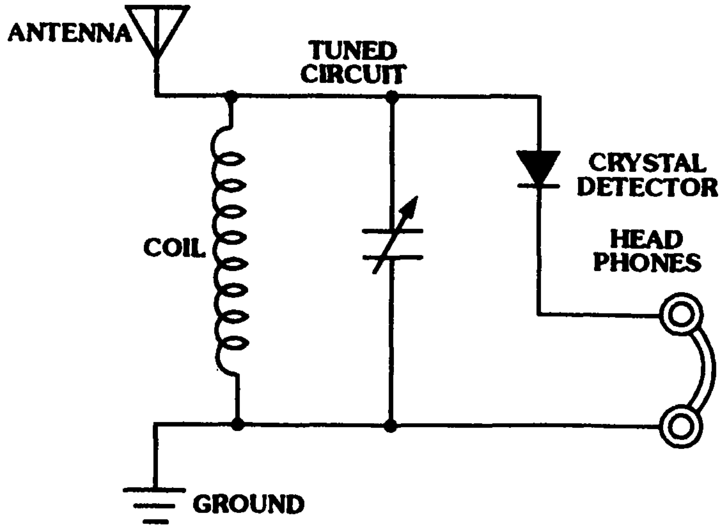
ANTENNA
TUNED CIRCUIT
CRYSTAL DETECTOR
HEAD PHONES
COIL
GROUND

Radio Technology (cont'd)

When these two different energy waves are put together by the mixer stage of the transmitter, the resulting wave, which is the signal that the station broadcasts, looks like this:

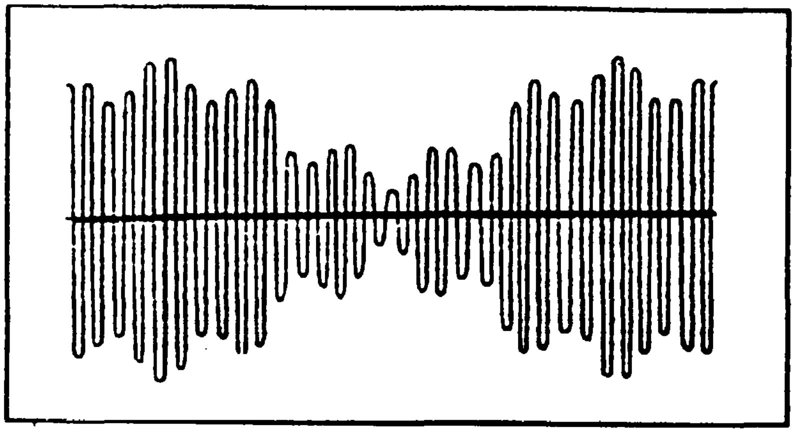

AM Broadcast Wave

The signal is called a MODULATED wave. You will notice that each of the waves is the same length, but the heights vary. Since the height of a wave is called its AMPLITUDE. The type of transmission that WXYZ uses is called AMPLITUDE MODULATION. That's why WXYZ is called an AM station.

The crystal radio that you built works in just the opposite way as the radio station. The modulated signal broadcast by WXYZ is "picked up" by the antenna on your radio.

The antenna is connected to the tuner circuit of the radio, which is made up of a coil of wire and a variable capacitor connected together. As you move the tuning knob, you are able to select, or isolate, the particular

This signal is then DEMODULATED by the diode, separating the voice (audio) wave from the carrier wave.

The voice (audio) wave then moves to the earphone, where it is changed back into sound waves that you can hear.

Learn the basic electronic symbols

Before starting to assemble your Smart Lab Crystal Radio please learn the basic electronic symbols that are shown on the plastic base.

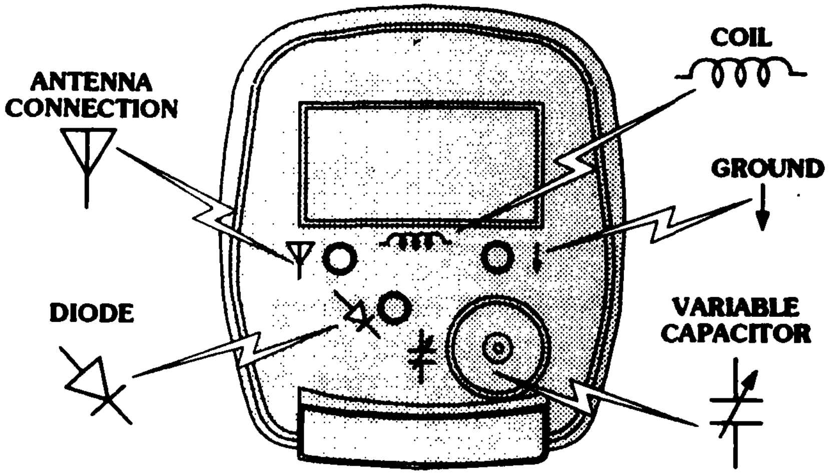

Explanation of components

Variable Capacitor - It is used to tune the radio to a station. The leads that are soldered to the lugs are used to connect it to the circuit.

Diode - A small crystal is sealed inside with leads connected to it. Pay careful attention to the position of the black bands painted on the diode close to one end.

Coil- This is a radio-tuning COIL. It was made by winding enameled copper wire around a paper core 80 times. The leads have been stripped and tinned so it can be connected to the circuit.

Earphone - It contains a small crystal that can make enough electricity to drive a metal diaphragm to produce sound. The leads have been stripped and tinned so it can be connected to the circuit.

Antenna - This is a wire used for radiating or receiving radio waves.

Ground - This is a wire used to make a electric connection with the earth or other type of grounding source to create a common return for an electric circuit.

DEAR CUSTOMER:

If we made an error and left something out of this set, or if something is damaged, we are sorry and wish to correct our error.
Please DO NOT return the set to the store where purchased, as the store does not have replacement parts. Instead, write us:

Please include:
1. Date of Purchase
2. Where purchased
3. Price paid
4. Model Number
5. Name of item
6. Brief Description of Problem
7. Include Sales Slip

We will do our best to satisfy you. Send your letter to:

NATURAL SCIENCE INDUSTRIES, LTD.
910 ORLANDO AVENUE
WEST HEMPSTEAD, N.Y. 11552-3942
ATTENTION: QUALITY CONTROL DEPT.

How to assemble your Crystal Radio

Before starting, it's important to assemble the parts correctly. Carefully read the following instructions step by step and have fun . . . half the fun is knowing you made your very own crystal radio.

Step ①

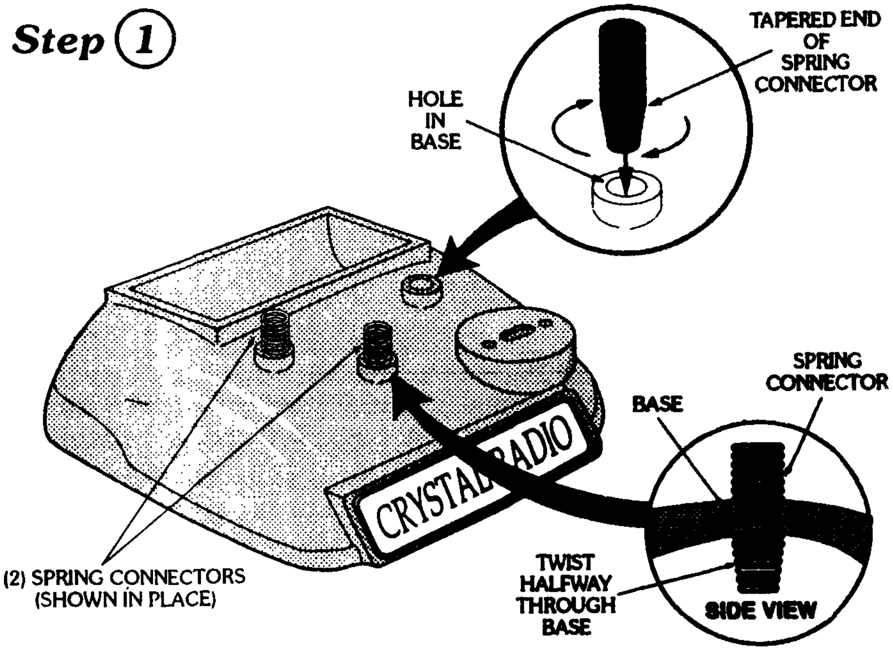
TAPERED END OF SPRING CONNECTOR
HOLE IN BASE
SPRING CONNECTOR
BASE
CRYSTAL RADIO
(2) SPRING CONNECTORS (SHOWN IN PLACE)
TWIST HALFWAY THROUGH BASE
SIDE VIEW

A. Locate the three Spring Connectors.
B. Insert one of the Spring Connectors (tapered end first) into one of the three holes on the Base.
C. Push down on the Spring Connector and twist to the right until the Connector is approximately half way through the base.

NOTE: Check to see if the Spring Connecter is inserted halfway through the Base by turning Base upside down (see insert).
D. Insert the remaining two Spring Connectors into the holes on Base the same way.

Radio Technology

Now that you have your crystal radio assembled and working, it is time to take a brief look at how it works. To do this you are going to take an imaginary trip to a radio station WXYZ, in anytown,USA.

When you arrive at WXYZ, you are met by the station's general manager, Mr. Smith, who is going to show you around.

First, Mr. Smith takes you to a STUDIO, a special room that a program comes from. There is a lot of equipment, including dials and switches, record players, lots and lots of records, and microphones.

Mr. Smith tells you that when you talk into one of the microphones, your words go into the station's electronic equipment. There, it is mixed with the station's carrier wave, and sent out into the air through the station's transmitting antenna - that tall tower you saw on top of the radio station building.

Then Mr. Smith takes a piece of paper and a pencil, and draws diagrams to show you what happens. He begins by saying that all energy travels in waves, and since the sound we make when we talk is a form of energy, it might look something like this:

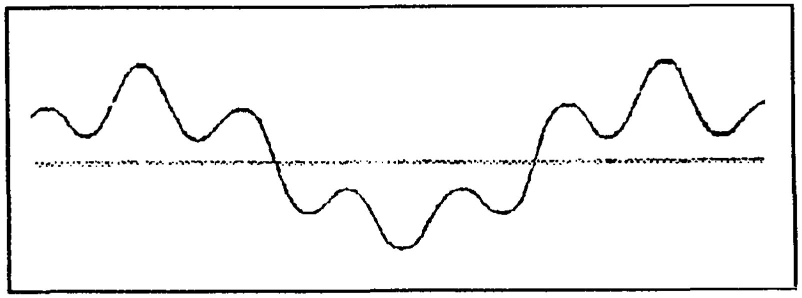

Sound Wave

The station's carrier wave is a radio wave, which is also a form of energy, which looks something like this:

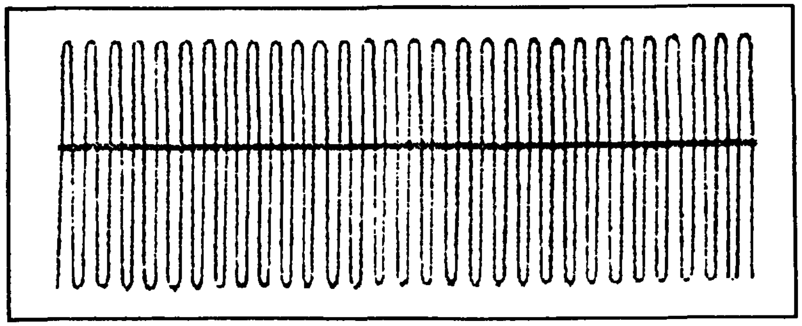

Radio Carrier Wave

How to properly operate your Smart Lab Crystal Radio and troubleshoot

After you have attached all the wires, carefully read the following instructions to make sure your crystal radio operates properly.

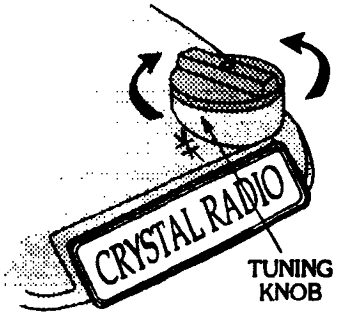

A. Place the earphone in your ear and turn the Tuning Knob clockwise and counterclockwise until you pick up the strongest signal.
Note: If you do not pick up a strong signal or do not hear anything, please read some of the troubleshooting solutions explained below.

TROUBLESHOOTING HINTS

1. There may be a bad connection caused by improper assembly of your radio. Check all the spring connectors to make sure the wires are attached correctly. If a wire is loose or the bare wire is not making good contact with the spring connector, the radio may not work.
2. If the uninsulated part of the Antenna wire touches anything other than the Antenna Connection on the radio, the radio may work improperly or not at all..
3. If the bare stripped end of the Ground wire is not wrapped tightly around a water pipe (so that it makes good contact), try taping the wire to the shiny part of the pipe using duct tape. If the pipe is dull or rusty, use a piece of sandpaper to gently sand the area where the wire makes contact. Also make sure the wire doesn't touch anything other than the Ground Connection or the radio may not work properly.
4. You may live in an area where radio reception is generally poor. Instead of trying to use your radio during the day, try at night when many radio stations are received better.
5. The Antenna on your radio is a very important component. Make the Antenna as long and high above the ground as possible. If you live in a multi-story dwelling the highest floor should be used when operating the radio. Insulated bell-wire, which is available in hardware stores, makes a very good substitution for your present Antenna.

How to assemble your Crystal Radio (cont'd)

For this step you will need the variable capacitor, (3) screws and turning knob. Be carefull not to misplace the small screws when assembling.

Step ②

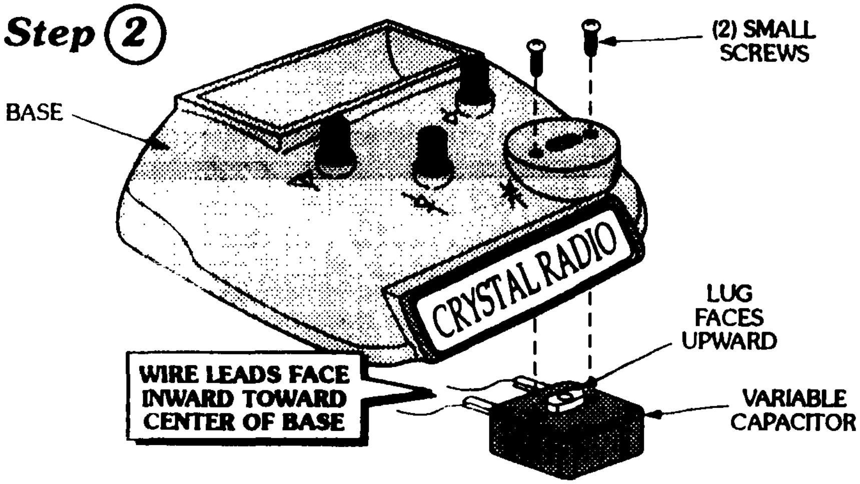

A. First place the Variable Capacitor up into the base so the holes on the capacitor line up with the holes on the base.
Note: When you do this be sure the wire leads face inward toward the base as shown and lug faces upward.

B. Next insert the two small screws into the holes on top of the Base and tighten using a small phillips screwdriver.

Step ③

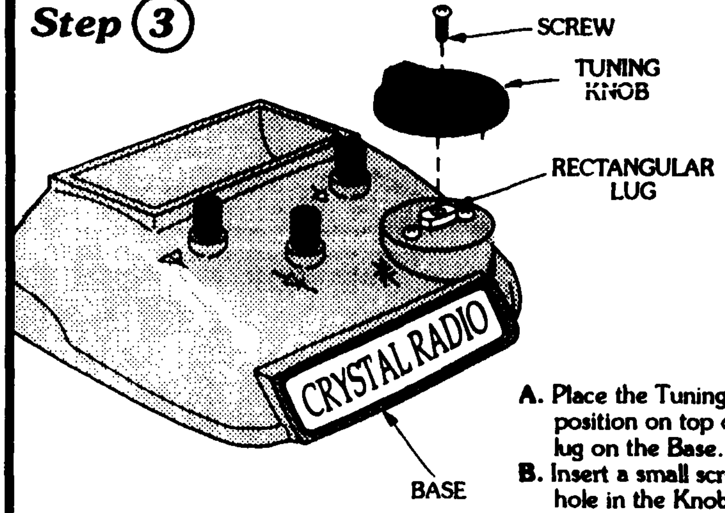

A. Place the Tuning Knob into position on top of the rectangular lug on the Base.
B. Insert a small screw into the hole in the Knob and tighten using a phillips screwdriver.

How to assemble your Crystal Radio (cont'd)

Step ④ For this step turn the Base upside down.

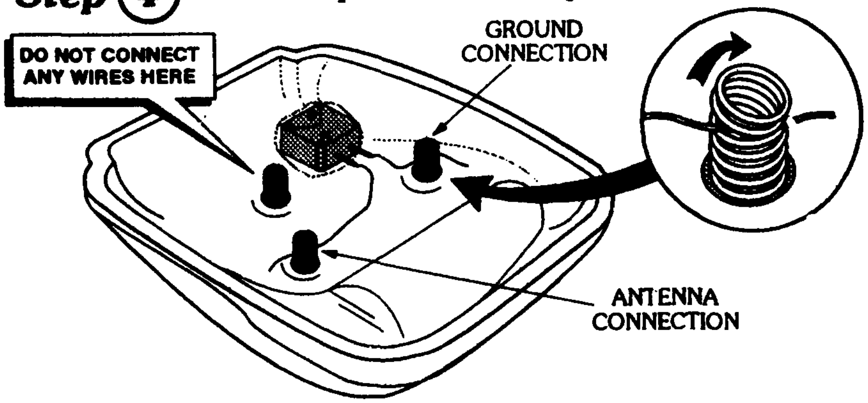
DO NOT CONNECT ANY WIRES HERE
GROUND CONNECTION
ANTENNA CONNECTION

A. Turn the Base upside down.
B. Connect one wire to the Antenna spring connector by bending the spring slightly to one side and inserting the wire through as shown in inset.
C. Connect the other wire to the Ground spring connector in the same manner.

NOTE: Do not connect any of the wires to the Diode spring connector or your crystal set will not work properly.

Step ⑤

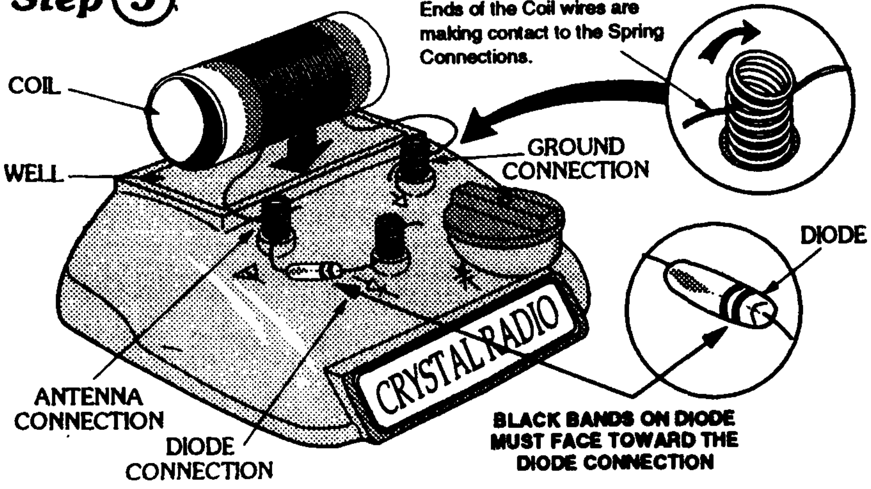
COIL
WELL
GROUND CONNECTION
DIODE
ANTENNA CONNECTION
DIODE CONNECTION
CRYSTAL RADIO

NOTE: Make sure the Plated Ends of the Coil wires are making contact to the Spring Connections.

BLACK BANDS ON DIODE MUST FACE TOWARD THE DIODE CONNECTION

A. Next place the Coil in the well of Base with the wires hanging loosely toward the connectors .
B. Connect one of the plated ends of Coil wires to the Ground Connector by bending the Spring Connector as shown in inset. Connect the other Coil wire to the Antenna Connector.
C. Place the Diode between the Diode and Antenna Connections carefully noting the black bands face toward the Diode Connection.
D. Connect ends of Diode wires to Antenna and Diode spring connectors shown.

How to attach the earphones

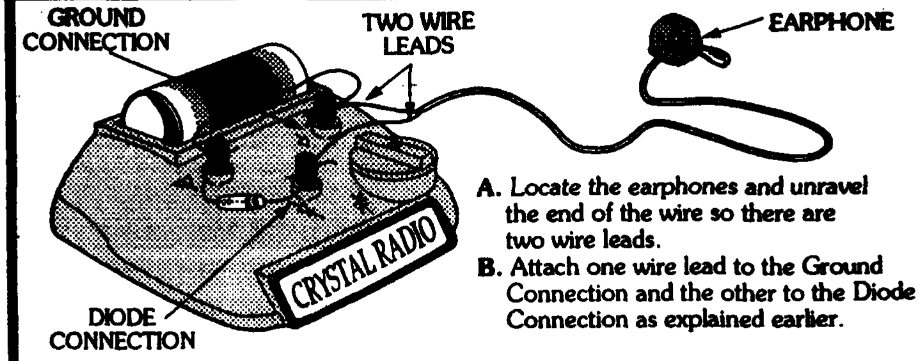
GROUND CONNECTION
TWO WIRE LEADS
EARPHONE
DIODE CONNECTION
CRYSTAL RADIO

A. Locate the earphones and unravel the end of the wire so there are two wire leads.
B. Attach one wire lead to the Ground Connection and the other to the Diode Connection as explained earlier.

Final assembly and operation

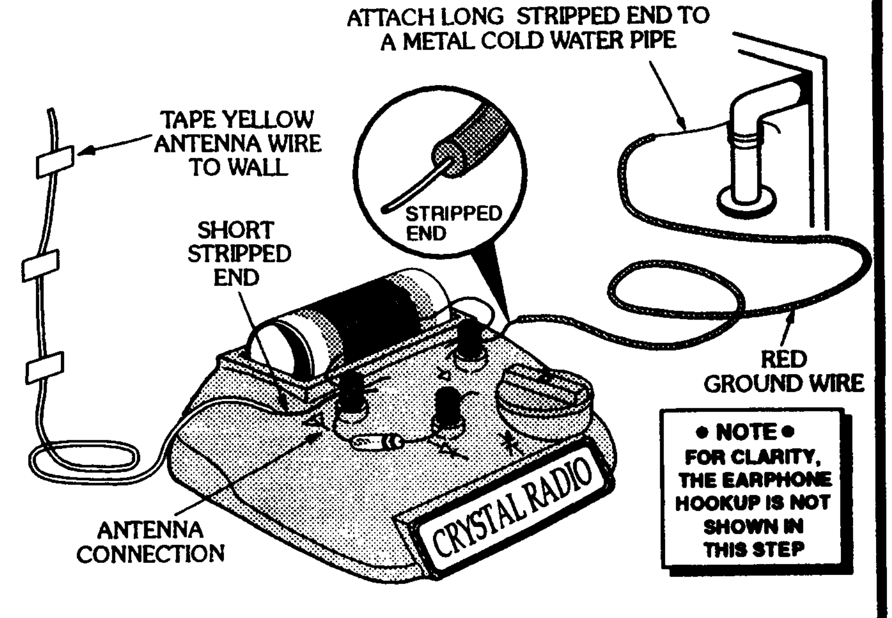
ATTACH LONG STRIPPED END TO A METAL COLD WATER PIPE
TAPE YELLOW ANTENNA WIRE TO WALL
STRIPPED END
SHORT STRIPPED END
RED GROUND WIRE
ANTENNA CONNECTION
CRYSTAL RADIO

**• NOTE •
FOR CLARITY, THE EARPHONE HOOKUP IS NOT SHOWN IN THIS STEP**

A. Locate the Red Ground and Yellow Antenna wire.
B. Attach the stripped end of Yellow Antenna wire to the Antenna Connection on the Base.

Note: The Antenna should be strung in a straight line away from power lines and large metal objects. We suggest taping it high up on a wall using masking tape as shown.

C. Next attach the short stripped end of the Ground wire to the Ground Connection.
D. Attach the Long stripped end to a metal cold water pipe, metal radiator pipe, etc., near where the radio will be used. Have a adult help you with this since some pipes may be hot.

Note: If there is not enough stripped, bare wire to wrap around a pipe then ask an adult to strip more insulation off the end using a pair of wire strippers.

Germanium vs Silicon Diode Testing: Read this document carefully, so you will not be the victim of cheap knock-off or the wrong type diodes.

The general rule is that ***SILICON*** diodes have a voltage drop across the Anode to Cathode of 0.7 V (7/10 tenths), and the ***GERMANIUM*** diodes have a voltage drop of 0.3 V (3/10 tenths) more or less. Either diode voltage drop (silicon or germanium) will display a reading within approximately 5% of these readings.

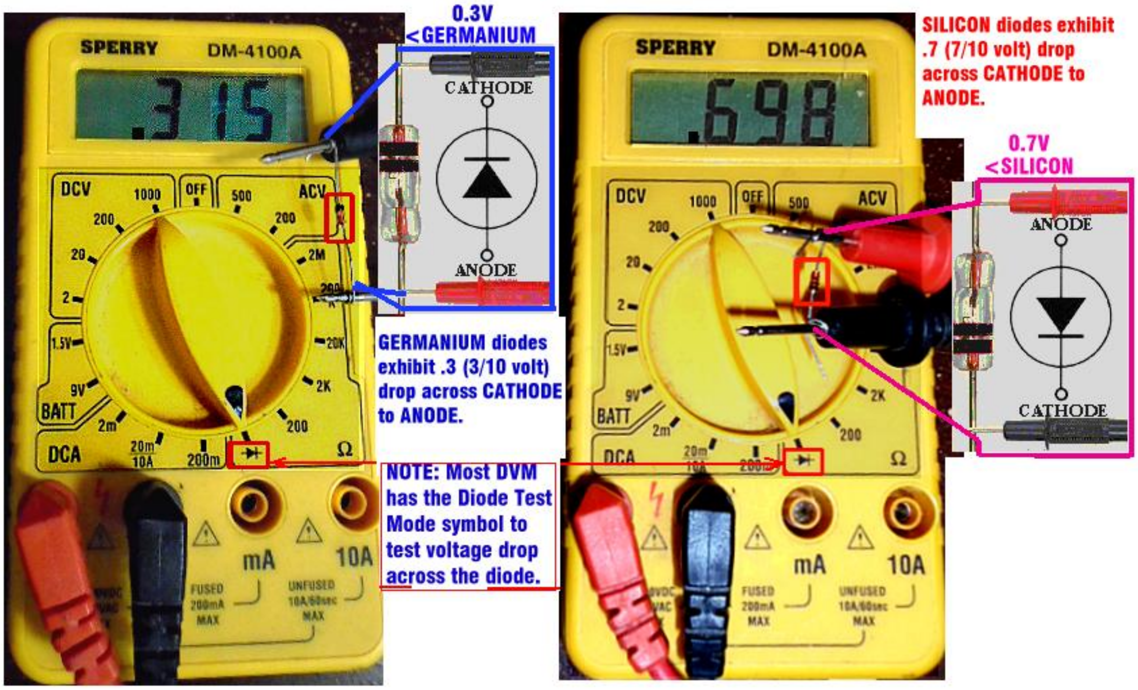

0.3V <GERMANIUM

GERMANIUM diodes exhibit .3 (3/10 volt) drop across CATHODE to ANODE.

NOTE: Most DVM has the Diode Test Mode symbol to test voltage drop across the diode.

SILICON diodes exhibit .7 (7/10 volt) drop across CATHODE to ANODE.

0.7V <SILICON

Most digital meters (DVM) have a switch setting ▶| that is used to measure voltage drop across these diodes. This setting is usually indicated by a diode symbol to let the user know the DVM is capable of measuring forward bias voltage. This setting will tell you immediately if the diode is a **germanium**, or **silicon** diode. You need to set the selector switch on your meter to the diode test symbol ▶|.

Measuring Forward Bias Voltage

To measure the forward bias voltage characteristic you connect the black probe of your meter to the cathode terminal. The cathode terminal is on the end with a band. You then connect the red probe to the anode terminal.

Set your DVM/DMM to the *diode test mode* ▶|, it should provide you with the respective voltage drop. If the figure is 0.3 V or less, the diode is a *germanium* type. If the voltage drop is 0.7 V or less the diode is a *silicon* diode.

Sensitivity and Forward Bias Characteristics:

The sensitivity of a diode to radio waves depends upon its *forward bias voltage*. This is the voltage across the diode terminals. When it falls below this threshold value, the diode will stop conducting. Obviously, the lower this threshold value is, the greater the sensitivity of the diode to the weak radio signals.

CONCLUSIONS:

When comparing germanium diodes with silicon diodes of similar forward bias voltage, silicon diodes do not perform as well as the germanium.

Germanium has many properties that silicon diodes do not have. Germanium requires very little forward current. Forward current in a germanium diode is in the *micro*ampere region, while silicon diodes require *amperes*. This makes *germanium* a much better choice for both medium and high frequency radio signals.

Germanium also exhibits a very low, point-contact junction capacitance, while the silicon diode has much higher capacitance. A low junction capacitance allows germanium diodes to operate more effectively at high RF frequencies.

In addition, reverse leakage current for germanium diodes is in the magnitude of 1000, much more than silicon. This makes the non-linear characteristics of the germanium diode much more effective for RF detection and demodulation than silicon.

Therefore, our conclusions are; germanium diodes provide the best performance in crystal radios, RF probes, HF, and VHF signal detection.

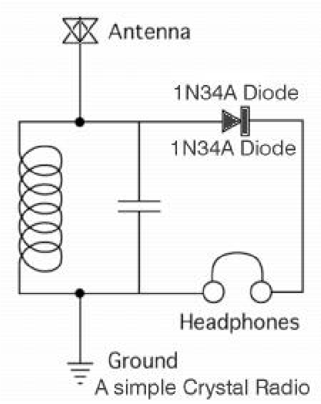
Antenna
1N34A Diode
1N34A Diode
Headphones
Ground

A simple Crystal Radio

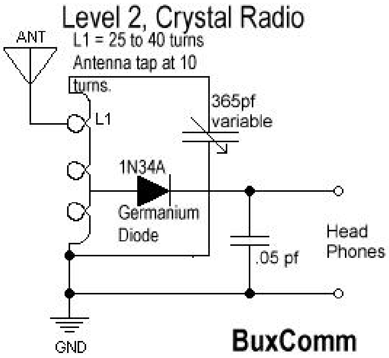
Level 2, Crystal Radio

ANT
L1 = 25 to 40 turns
Antenna tap at 10 turns.
L1
365pf variable
1N34A
Germanium Diode
.05 pf
Head Phones
GND
BuxComm

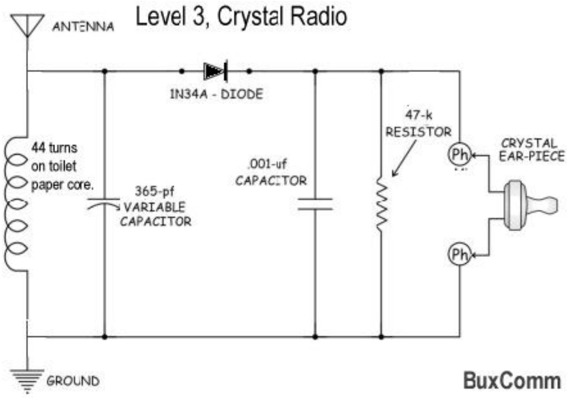
Level 3, Crystal Radio

ANTENNA
IN34A - DIODE
47-k RESISTOR
44 turns on toilet paper core.
365-pf VARIABLE CAPACITOR
.001-uf CAPACITOR
CRYSTAL EAR-PIECE
Ph
Ph
GROUND
BuxComm

BuxComm High Quality, CRYSTAL RADIO ANTENNA KIT, $14.95 cat# 50CRAK

This Crystal Radio Antenna kit consists of 50 feet of #16 AWG insulated & stranded copper antenna wire, two heavy-duty, Delrin, UV resistant antenna insulators and instruction sheet.

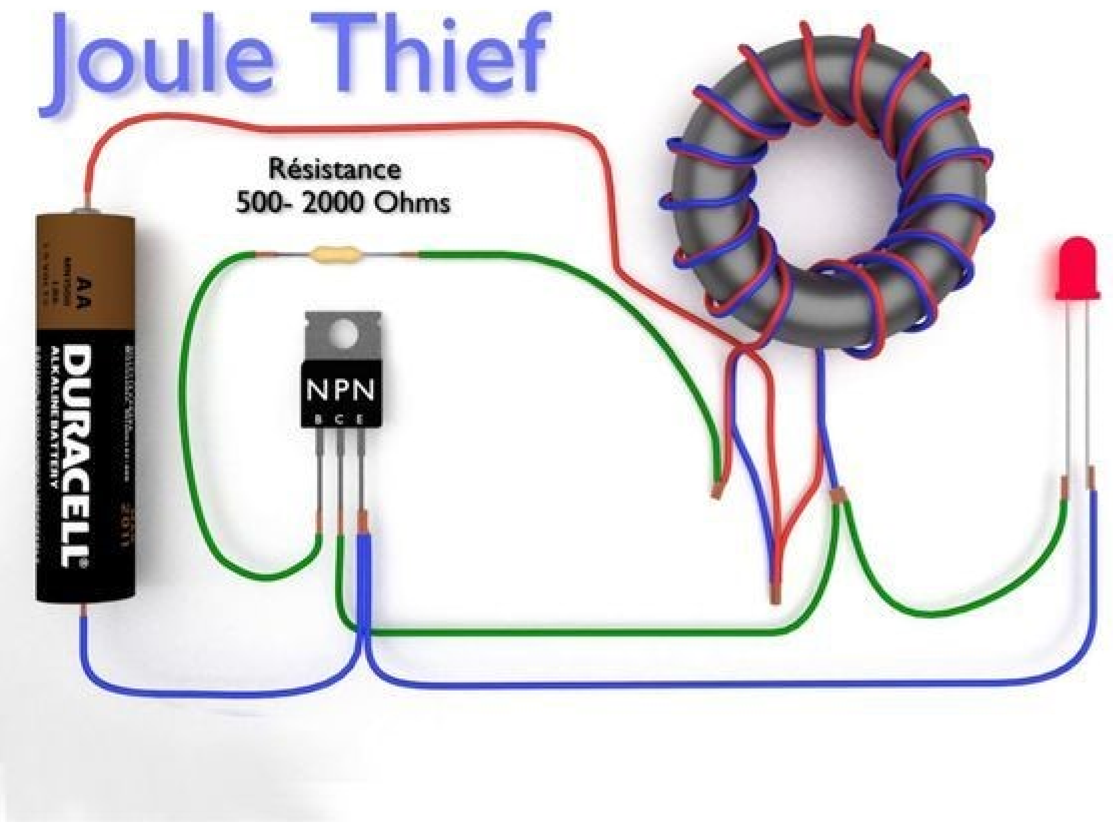
Joule Thief
Résistance
500- 2000 Ohms
NPN
B C E
AA
DURACELL
ALKALINE BATTERY

The joule thief circuit.

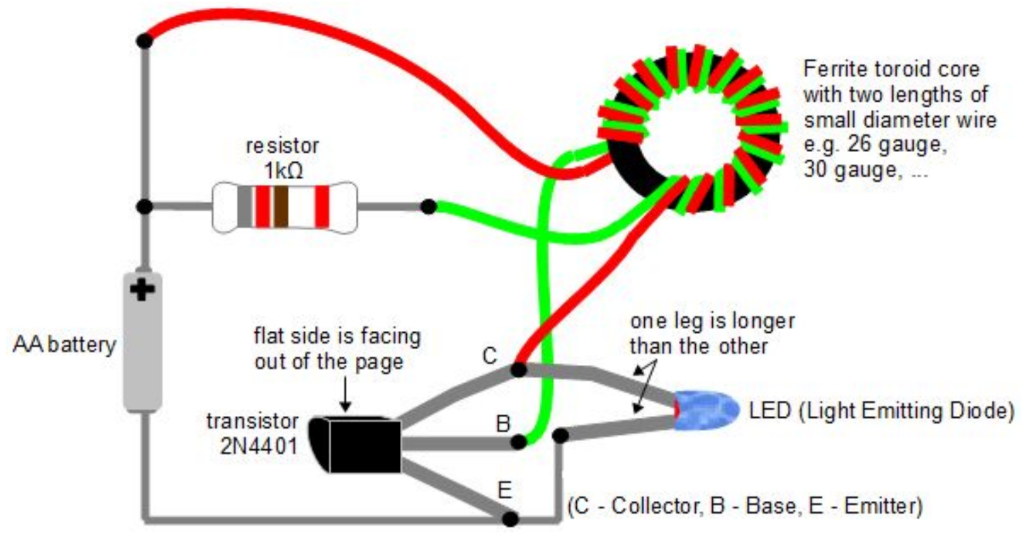
resistor
1kΩ
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

Transistor - The legs of the transistor can be determined by noticing that there's a flat side to the transistor case. See the diagram above. A large number of transistors have been reported to work: 2N4401, NET123AP, BC547B, 2SC2500, BC337, PN2222, to name just a few.

LED - One leg of the LED is longer than the other leg. Use this to determine which one goes where. See the diagram above.

Resistor - The diagram says use a 1 kilo ohm resistor but I've used an 820 ohm one just fine. I've also seen a 2 kilo ohm one in use. Use whatever works for you. You can also use a potentiometer (a variable resistor) so that you can easily adjust it to select the resistance that gives the best light.

Toroid ferrite core - Some people have gotten these by opening up compact fluorescent lightbulbs (CFLs). I took mine out of some device whose original function I don't know. To get it working, my first one had just 13 turns for each wire and I used a 30 gauge wire and a 26 gauge wire. The wire must be insulated. A variety of number of turns will work. This is something you can play with. Look at the diagram carefully to determine where the wires connect to.

Its a Rubbish Challenge

Dog Light

Joule Thief
electronics circuit

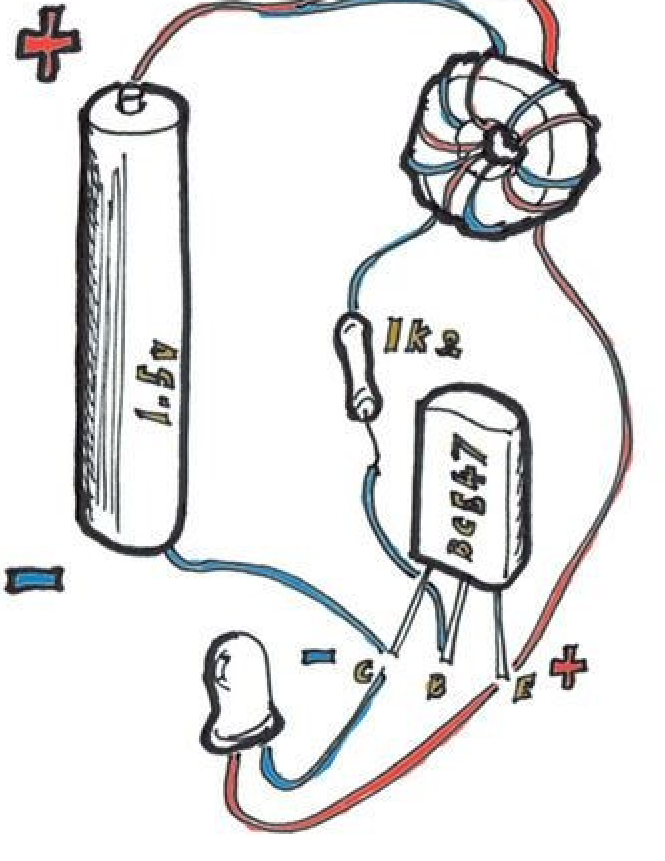

+
1.5v
1kΩ
BC547
-
-
C
B
E
+

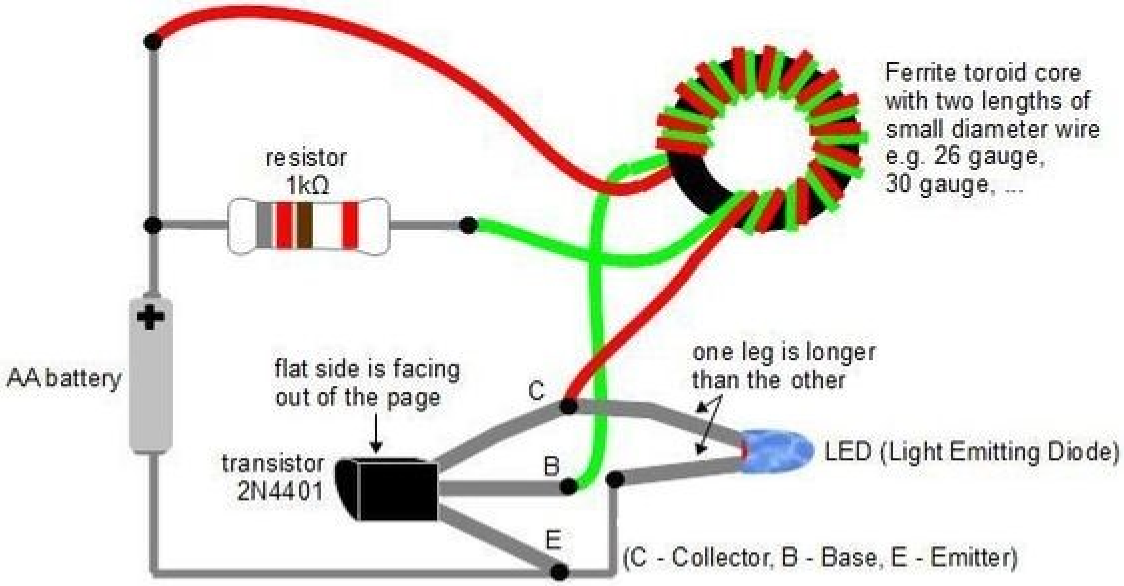
resistor
1kΩ
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
AA battery
flat side is facing
out of the page
C
one leg is longer
than the other
transistor
2N4401
B
LED (Light Emitting Diode)
E
(C - Collector, B - Base, E - Emitter)

COUPONS AVAILABLE

ACTIVATE COUPONS

advertisement

MAKING A SIMPLE JOULE THIEF (MADE EASY)

Technology (/technology/) > Leds | by ASCAS (/member/ASCAS/) Follow

195,691 573 165

Posted Dec. 24, 2011 | (cc) BY-NC-SA

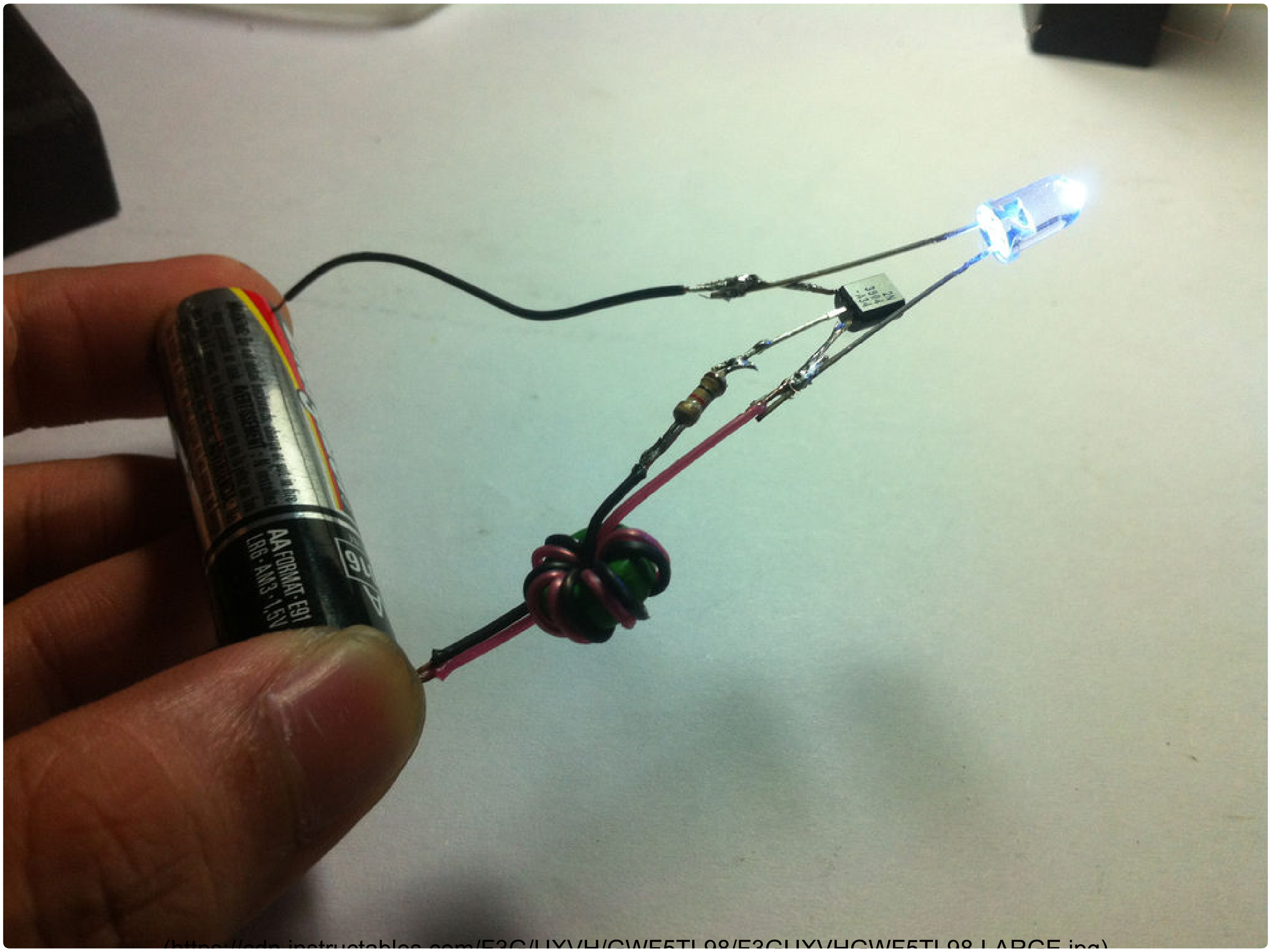

(https://cdn.instructables.com/F3C/UXVH/GW55TL98/F3CUXVHGW55TL98.LARGE.jpg)

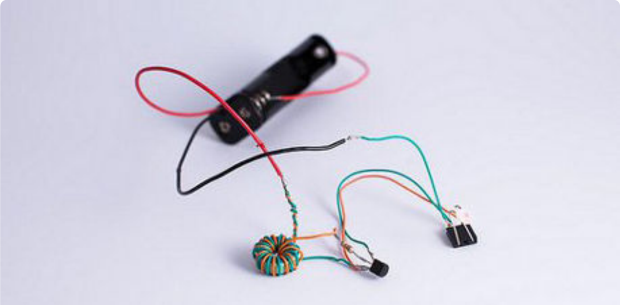
(https://cdn.instructables.com/FB5/BQY8/GWHN3

Mini
Version
(https://cdn.instructables.com/FF8

(https://cdn.instructables.com/F3C

Today I am showing you how to make a very simple joule thief. A joule thief has many applications, the best gadget that I made with was a "Water Powered Lamp", soon I'm going to post on a guide about it but first I need to post this guide. I used an iPhone 4S as my camera :)))

What Is A Joule Thief ?

To simplify everything, a "joule thief" is a circuit that helps drive an LED light even though your power supply is low. What can we do with it? We can use it to squeeze the life out of our old, almost drained, non functioning batteries. This project can also be considered as a green and environmental experiment, we can also use it as a flashlight that can be ran by an old, weak, almost drained battery. I even tried to use my water powered battery from my previous instructable the "Water Powered Calculator (https://www.instructables.com/id/Water-Powered-Calculator/)", the project was featured and displayed in instructable's front page in the "Technologies" category.

My Next Projects That Involves A Joule Thief: (soon to be posted)
- Water Powered Lamp
- Water Powered Flash Light
- Dead Battery Drainer Lamp

Here's A Video From Make Magazine:

Add Tip Ask Question

Step 1: Parts and Materials

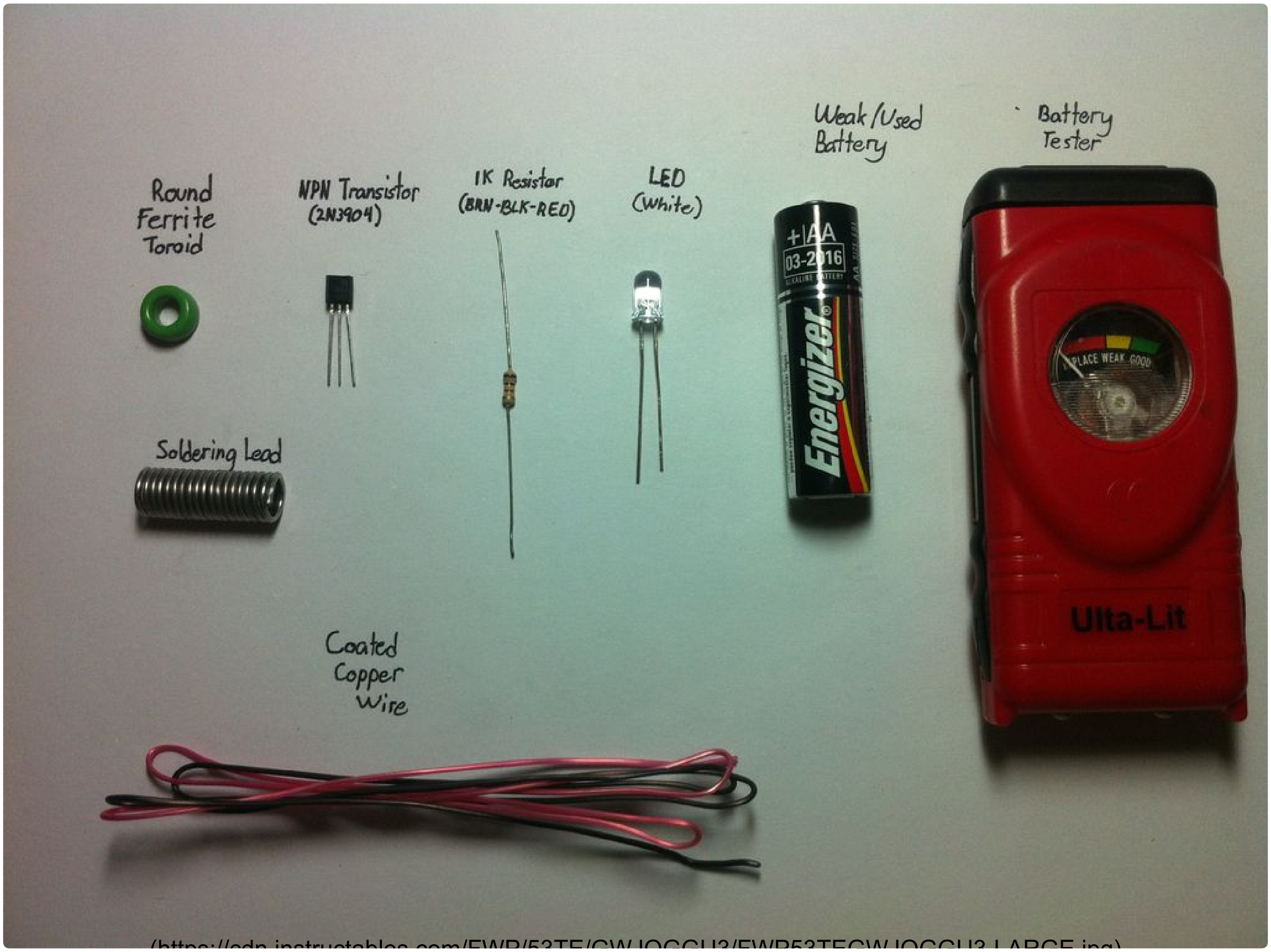
Round Ferrite Toroid
NPN Transistor (2N3904)
1K Resistor (BRN-BLK-RED)
LED (White)
Weak/Used Battery
Battery Tester
Soldering Lead
Coated Copper Wire
+AA
03-2016
Energizer
REPLACE WEAK GOOD
Ulta-Lit

(https://cdn.instructables.com/FWP/53TF/GWJOGCU3/FWP53TFGWJOGCU3.LARGE.jpg)

The Parts Needed Are: (click the item to know where to find/ buy)

- Round Ferrite Toroid (can be found in old CFL bulbs)
- Old/ Used Batteries (can be found in garbage cans)
- NPN Transistor (2N3904) (http://www.radioshack.com/product/index.jsp?productId=2062609)
- 1K Resistor (BRN-BLK-RED) (http://www.radioshack.com/product/index.jsp?productId=2062343)
- LED Light (http://www.radioshack.com/product/index.jsp?productId=3096133)
- Battery Tester (http://www.tooldistrict.com/Ulta-Lit-Battery-Tester-5001-p/901233.htm) (optional)
- Soldering Lead (http://www.radioshack.com/product/index.jsp?productId=2062717)

- Copper Wire/ Magnet Wire (http://www.radioshack.com/product/index.jsp?productId=2036277)
- Battery Case/ Holder (http://www.radioshack.com/product/index.jsp?productId=2062247)

I want to share something. Here in the Philippines electronic parts are extremely cheap, they are extremely far cheaper from
radio shack, for example one transistor costs (2 phil. pesos - 6 US cents), a LED cost (9 phil. peso - 29 US cents) and a 1K resistor cost (25 phil. cents - 0.8 US cents). I usually buy thing from Deeco or Alexan. Usually prices here are 15x cheaper from radio shack. Price conversion - $1 US Dollar = P0.31 Philippine Peso (12/24/11).

Add Tip   Ask Question

Step 2: Schematic Diagrams

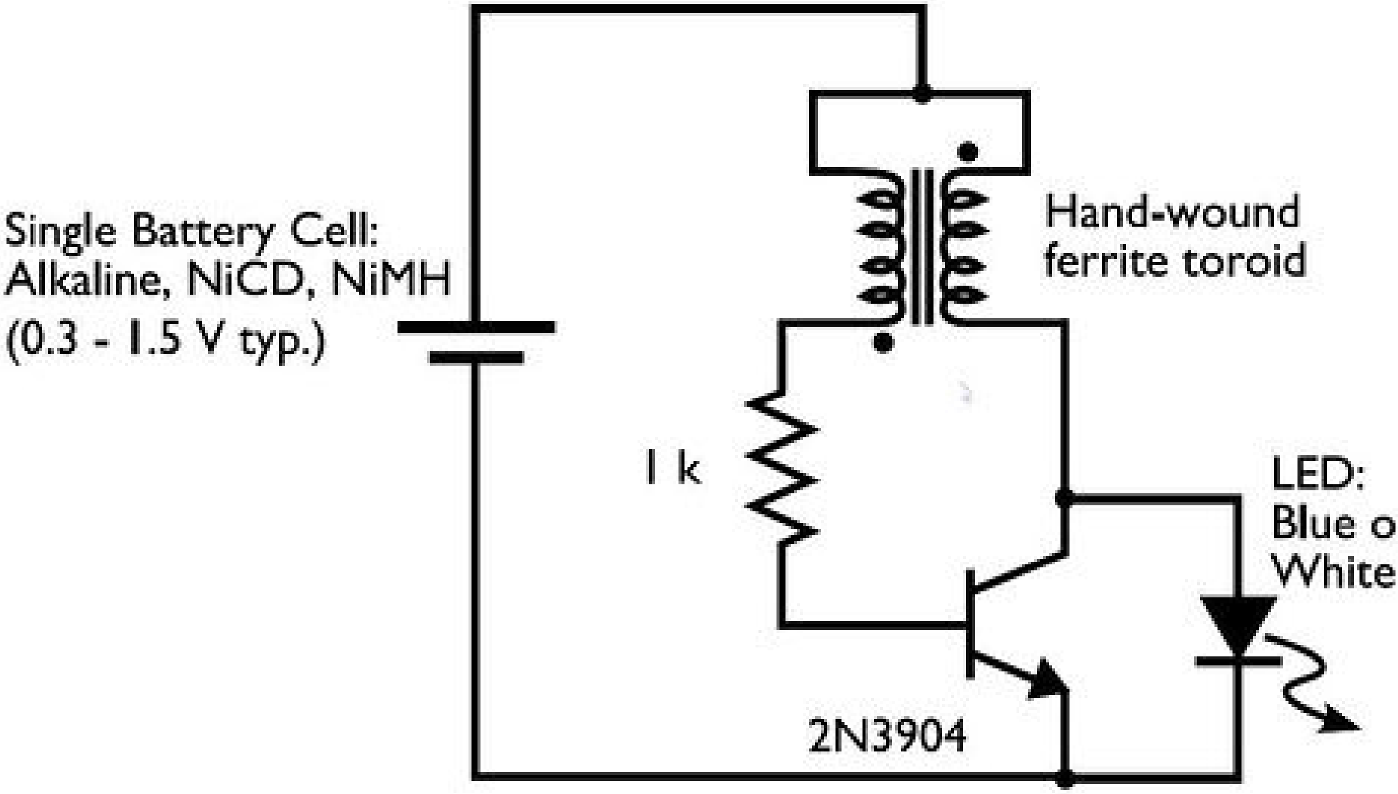
Single Battery Cell:
Alkaline, NiCD, NiMH
(0.3 - 1.5 V typ.)
Hand-wound
ferrite toroid
1 k
LED:
Blue or
White
2N3904

(https://cdn.instructables.com/F3N/S5I7/GWKD43C0/F3NS5I7GWKD43C0.LARGE.jpg)

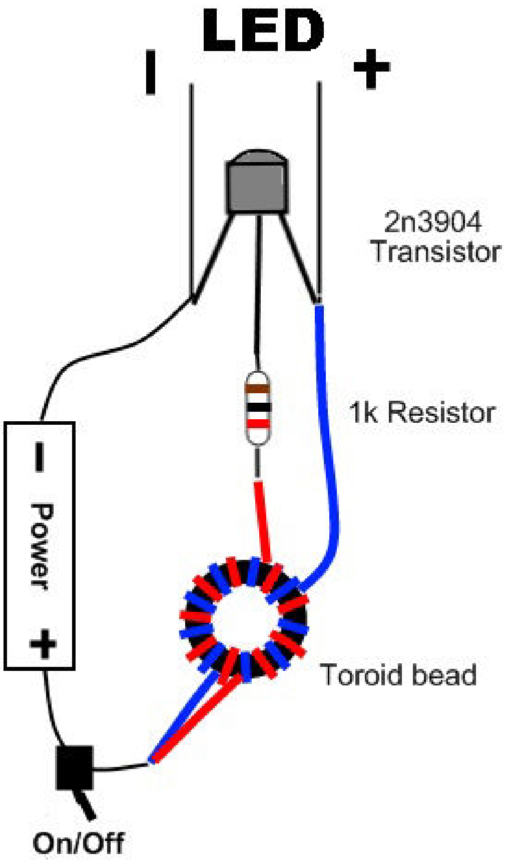
LED
I
+
2n3904
Transistor
1k Resistor
- Power +
Toroid bead
On/Off

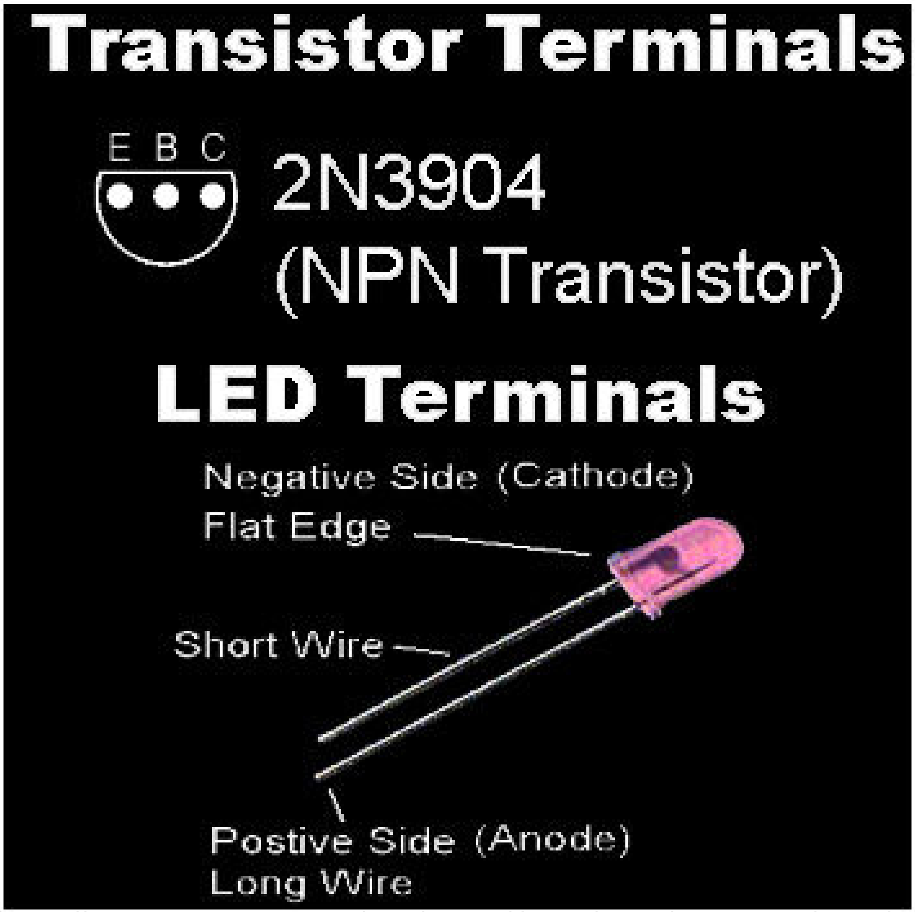
Transistor Terminals
E B C
2N3904
(NPN Transistor)
LED Terminals
Negative Side (Cathode)
Flat Edge
Short Wire
Postive Side (Anode)
Long Wire

(https://cdn.instructables.com/FN0/CPE6/G (https://cdn.instructables.com/F5P/M7EM/GWINSM3D/F5PM7EMGWINSM

Here are the schematic diagrams that are involved with the joule thief circuit.

Add Tip Ask Question

Step 3: Winding Wire at the Toroid

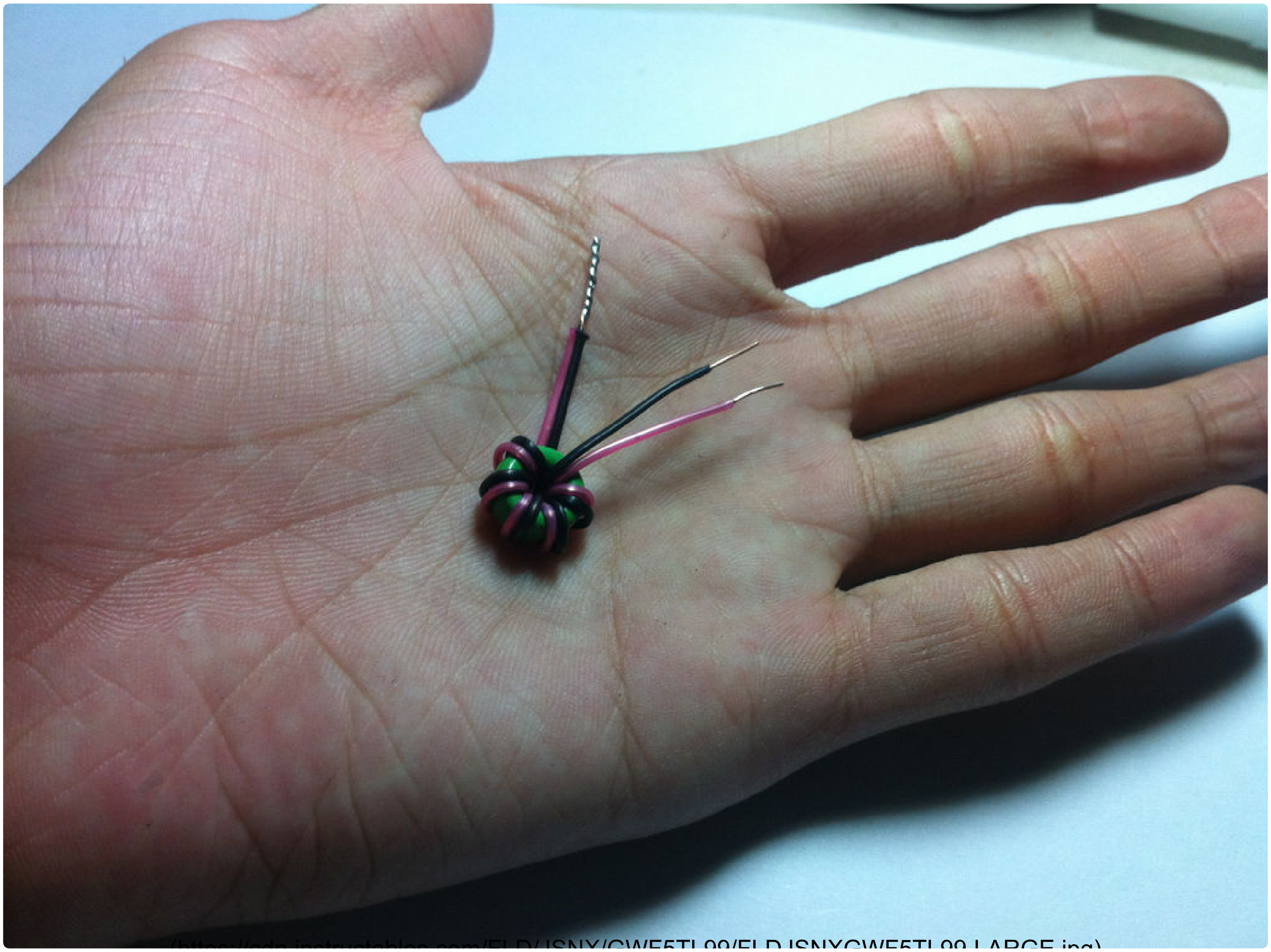

(https://cdn.instructables.com/FLD/JSNY/GWF5TL99/FLDJSNYGWF5TL99.LARGE.jpg)

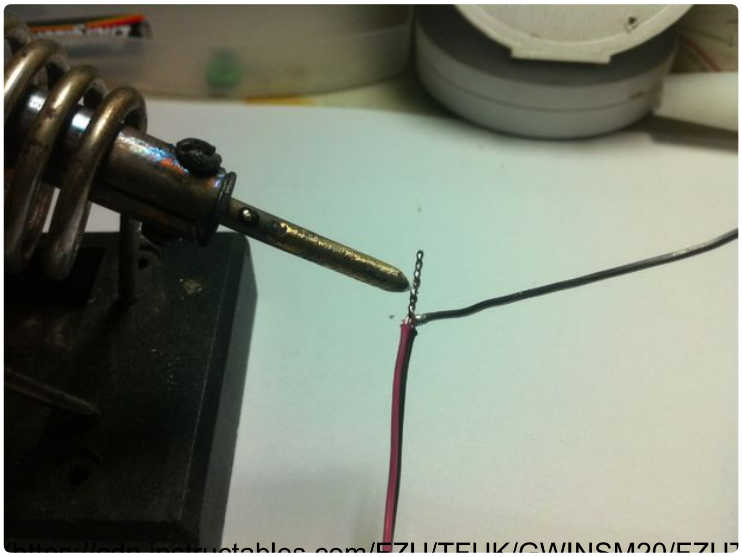

First, connect both ends of the copper wire before wounding, be sure to remove the insulation. Then try to solder the ends so it would not split up. Second, wind the wire until you run out of space in the round ferrite toroid. I have some tips for you, try to use a gauge #22 enamel coated copper wire for better performance, oh! my last tip is that "the more you wind the wire to the ferrite toroid the better".

Add Tip　Ask Question

Step 4: Soldering the Resistor

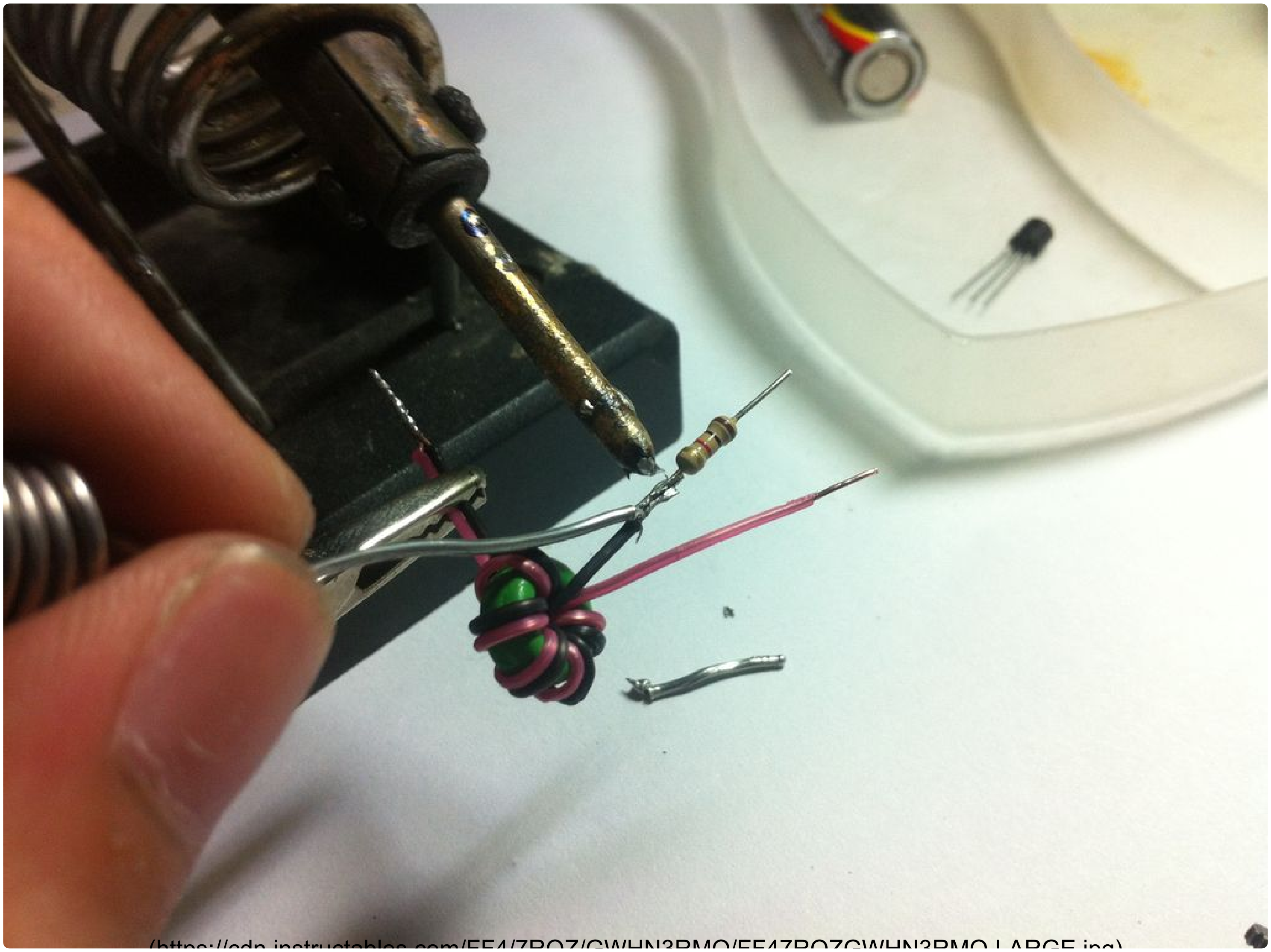

(https://cdn.instructables.com/FF4/ZBOZ/GWHN3BMQ/FF4ZBOZGWHN3BMQ.LARGE.jpg)

Solder the resistor with one end of the wounded ferrite toroid's wire. Oh! also don't forget to level the other end of the resistor with the other unused wire from the wounded ferrite toroid.

Add Tip

Ask Question

Step 5: Soldering the Transistor

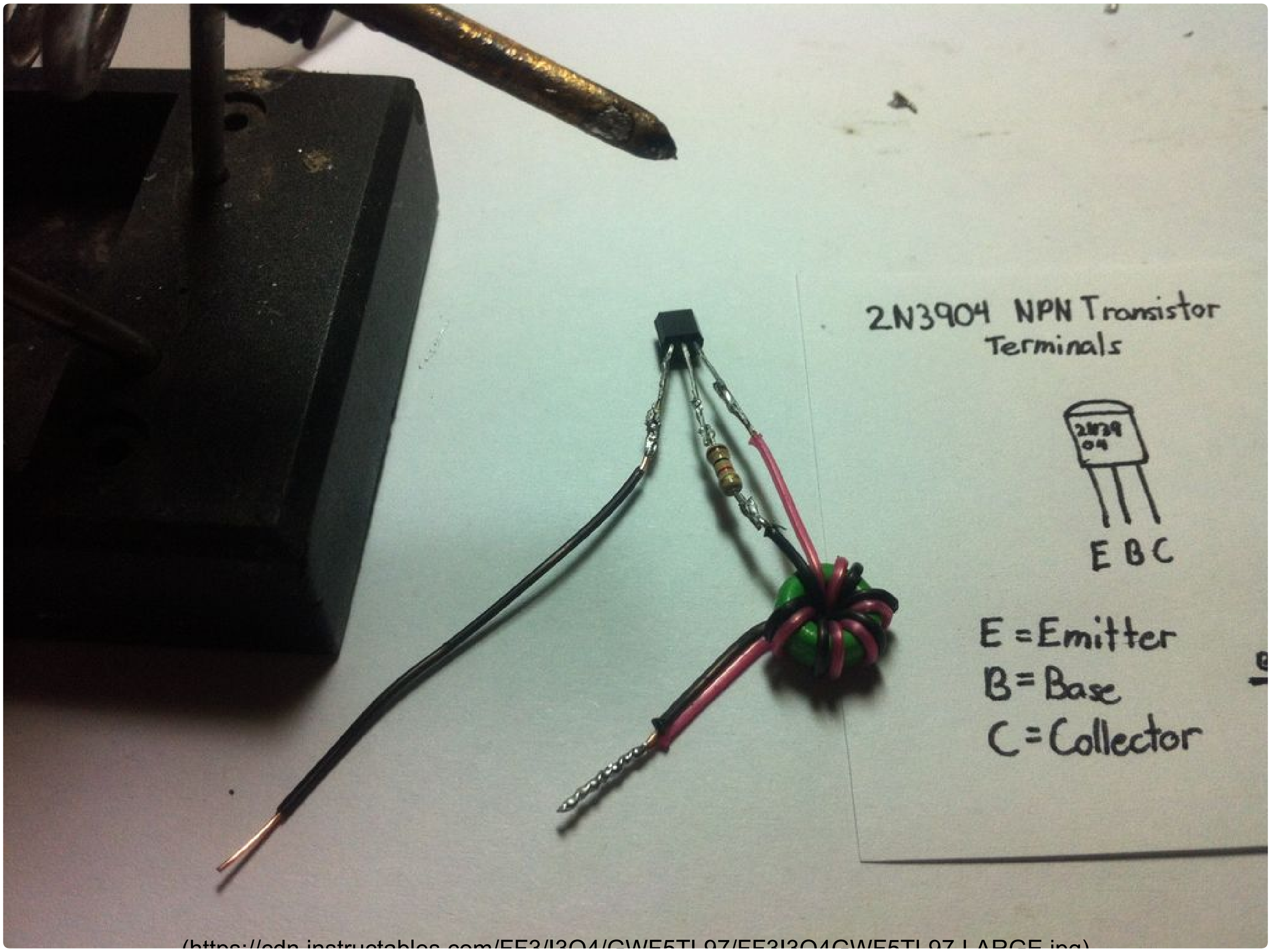
2N3904 NPN Transistor
Terminals
2N39
04
E B C
E = Emitter
B = Base
C = Collector

(https://cdn.instructables.com/FE3/I3O4/GWF5TL97/FE3I3O4GWF5TL97.LARGE.jpg)

Solder the proper connections to the transistor. For the emitter - connect another wire, the wire will be connected to the negative part of the battery. For the base - solder the other end of the resistor to the base. For the Collector Solder the unused wire of the ferrite toroid.

Add Tip Ask Question

Step 6: Soldering the LED

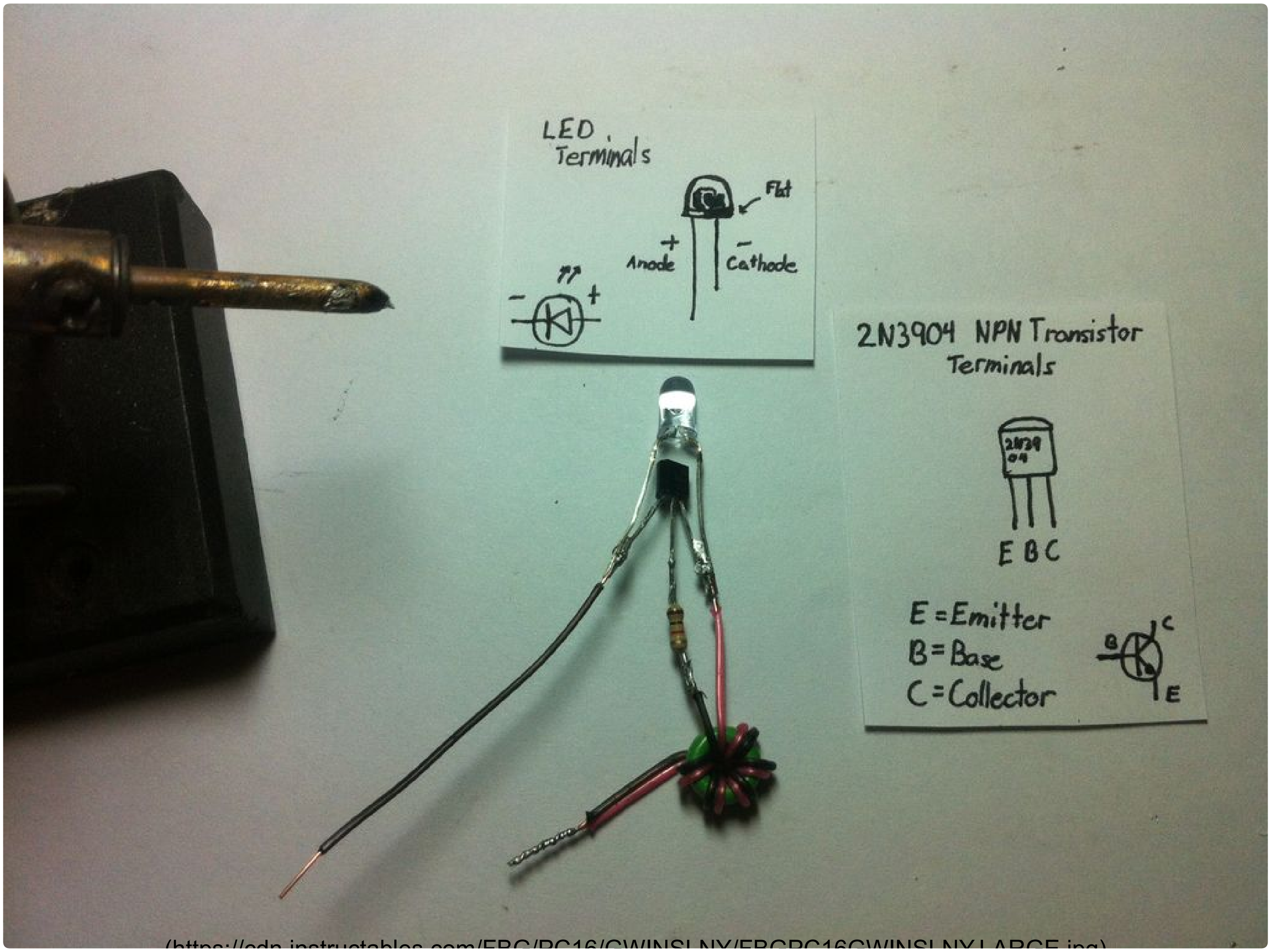
LED Terminals
Flat
+
Anode
–
Cathode
–
+
2N3904 NPN Transistor Terminals
2N39 04
E B C
E = Emitter
B = Base
C = Collector
B
C
E

(https://cdn.instructables.com/FBG/PC16/GWINSLNY/FBGPC16GWINSLNY.LARGE.jpg)

Solder the shorter wire of the LED to the tansistor's emitter and the longer part of the LED to the transistor's collector. After all that, you can now trim the excess wires.

Add Tip Ask Question

Step 7: Time to Look for Old Batteries

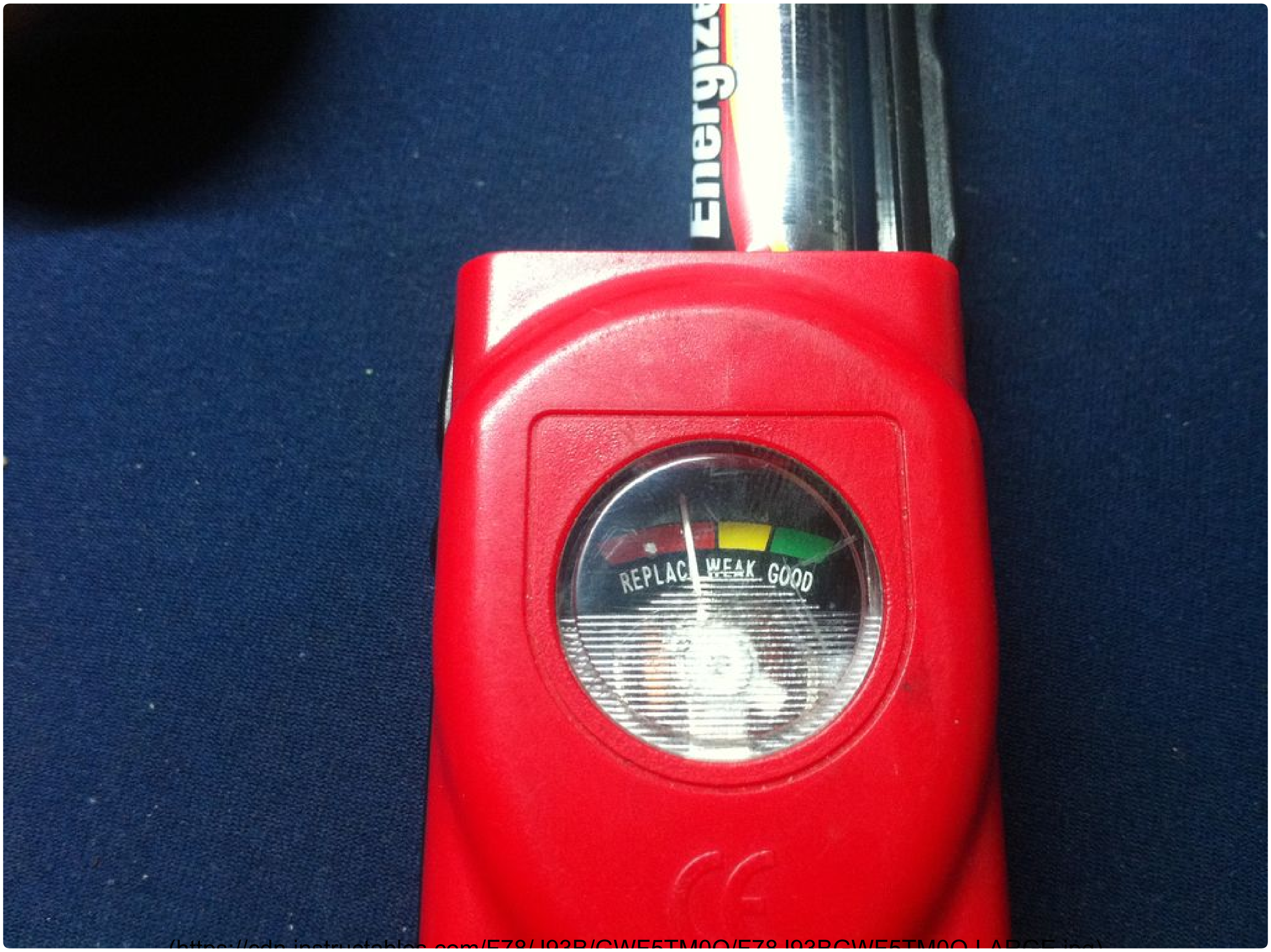
Energizer
REPLAC
WEAK
GOOD

(https://cdn.instructables.com/F78/J93B/GWF5TM0Q/F78J93BGWF5TM0Q.LARGE.jpg)

(https://cdn.instructables.com/FX4/JNKD/GWINSM4V/FX4JNKDGW

PHILIPS
DURACELL
Kodak
Energizer
SUPER
EVEREADY

(https://cdn.instructables.com/F40/8TOM/GWINSM

Use your battery tester to confirm that your battery is close to death. The tester is only an optional tool, it's just used to determine the battery's remaining power.

Add Tip Ask Question

Step 8: Time to Test It - You're Done !!

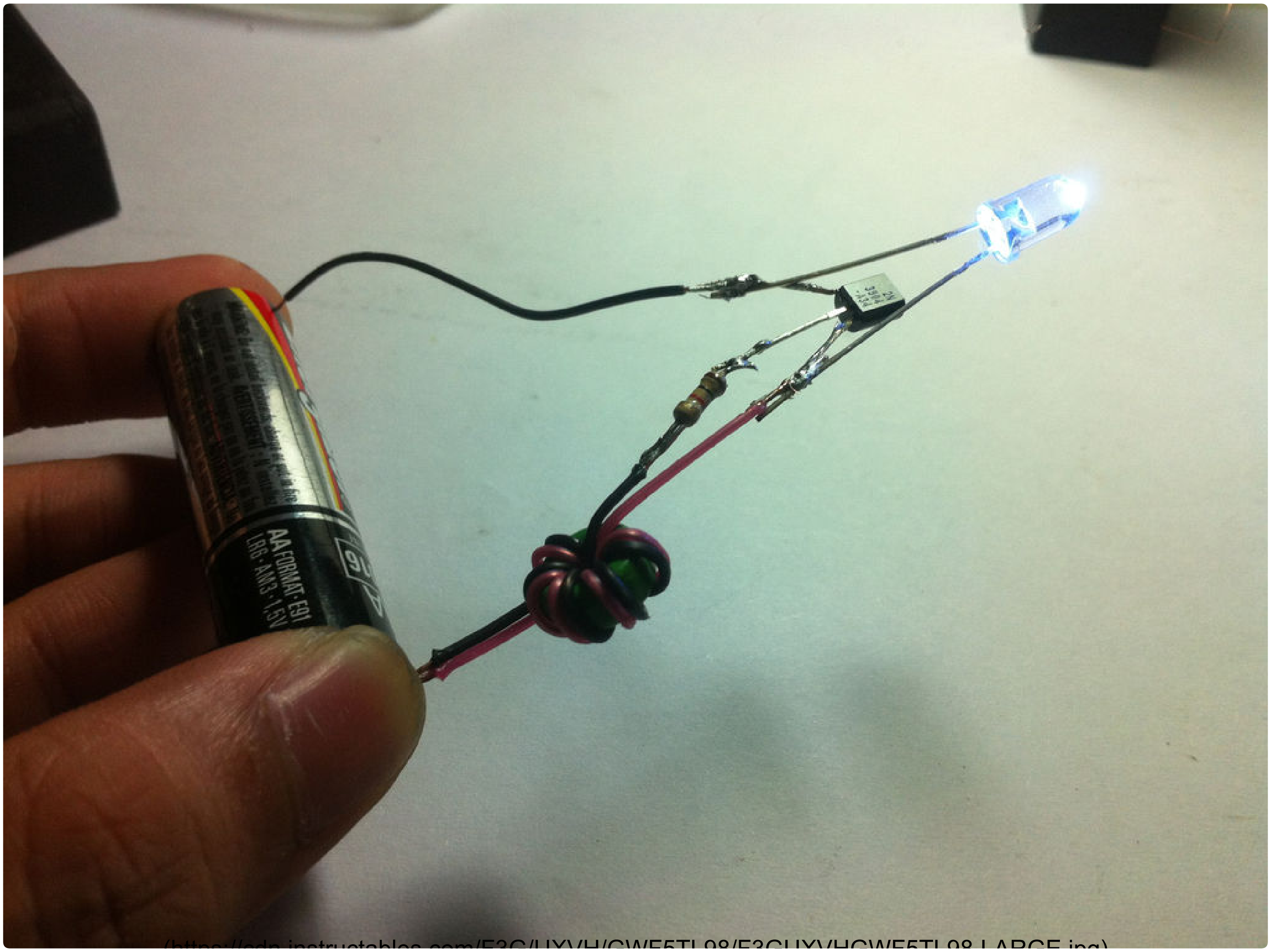
AA FORMAT · E91
LR6 · AM3 · 1.5V

(https://cdn.instructables.com/F3C/UXVH/GWE5TL98/F3CUXVHGWE5TL98.LARGE.jpg)

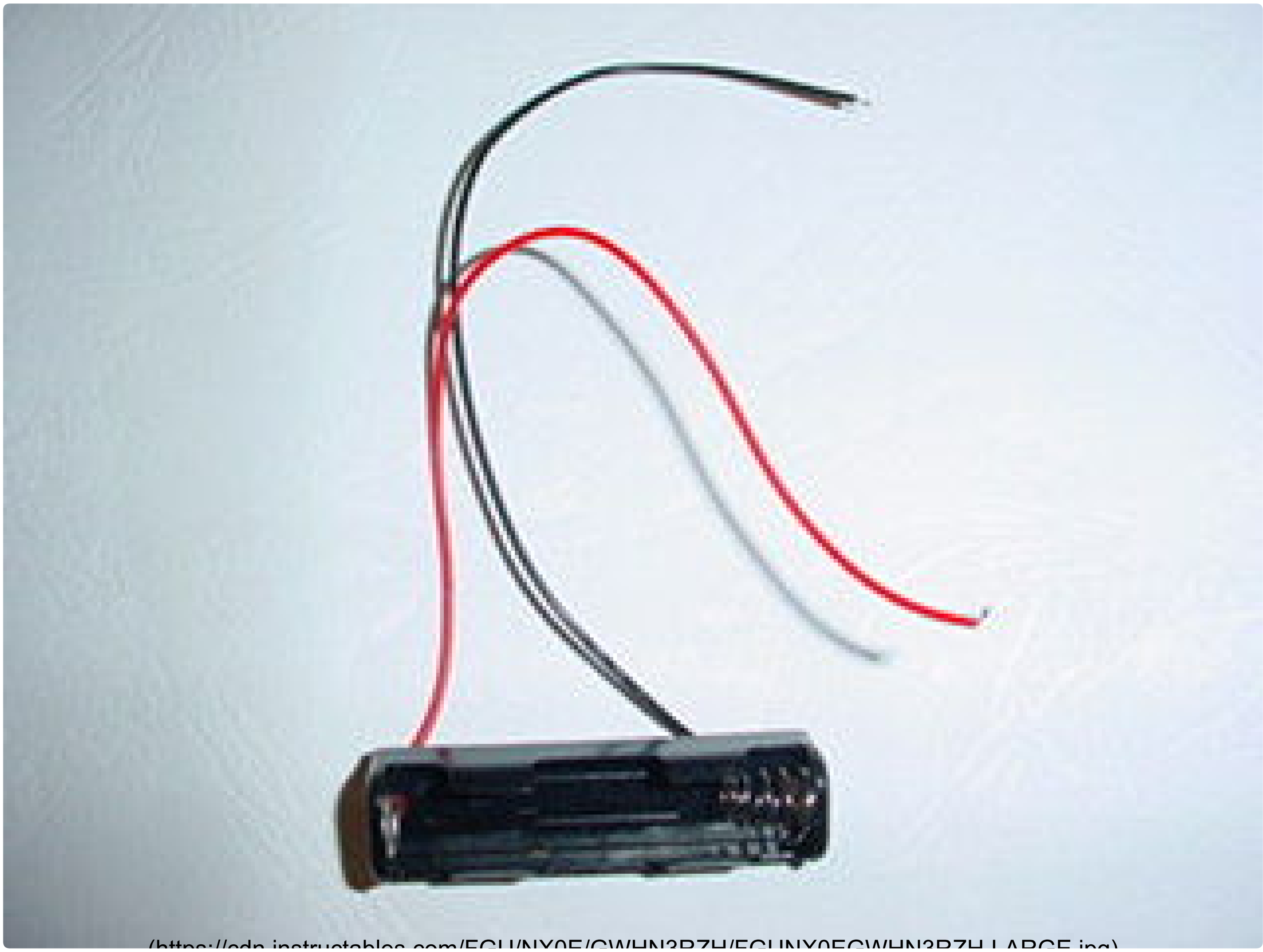

(https://cdn.instructables.com/FGU/NX0F/GWHN3BZH/FGUNX0FGWHN3BZH.LARGE.jpg)

The wire connected to the transistor's emitter should be connected to the battery's negative side and the remaining wire of the ferrite toroid should be connected to the battery's positive side. Oh! one more thing, I advise everyone to use a battery case or attach a conductive magnet for each wire, so you wouldn't hold it all the time.

Add Tip Ask Question

Eliminate Your Power Bills!

Build Your Own Energy Generator for Just $49

Tesla Energy

advertisement

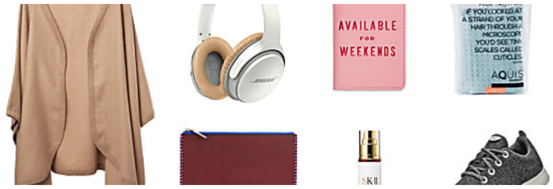

29 Travel Essentials We Won't Leave Home Without | Cartageous

Cartageous

advertisement

8 People Made This Project!

"Joule Thief" LED Night Light

by **ledartist** on September 2, 2011

Table of Contents

"Joule Thief" LED Night Light 1

Intro: "Joule Thief" LED Night Light 2

Step 1: Features 2

Step 2: Technical Overview 3

File Downloads 4

Step 3: Assembly 4

Step 4: Performance 5

Step 5: PCB & Kit 6

Related Instructables 6

Comments 6

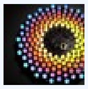
Author:ledartist My Blog
I work with LEDs and microcontrollers to create beautiful objects.

Intro: "Joule Thief" LED Night Light

I have many used batteries around. Remote controls, cameras, many electronic gadgets all use batteries, mostly AA size. I always felt guilty for throwing away the used batteries. I know there are rechargeable batteries, but many electronics don't work well with rechargeables.

I also know that those "used" or "spent" batteries usually have some juice left in them. So to come up with a good use of used batteries, I've created a LED night light.

I like having a little night light on when I sleep. LEDs are perfect for this purpose, because they are energy efficient, and good at providing low intensity illumination.

This LED night light operates with just one battery. It utilizes a little circuit called Joule Thief to boost voltage out of an AA battery. I also added a light sensor to turn it on automatically when the surrounding is dark.

The circuit is energy efficient, and requires very low voltage to work. So it effectively sucks every bit of energy out of batteries. This type of circuit is often called "Joule Thief", because it works as though stealing every bit of energy (Joule is a unit for energy) out of battery.

I'm calling this project Night Joule Thief.

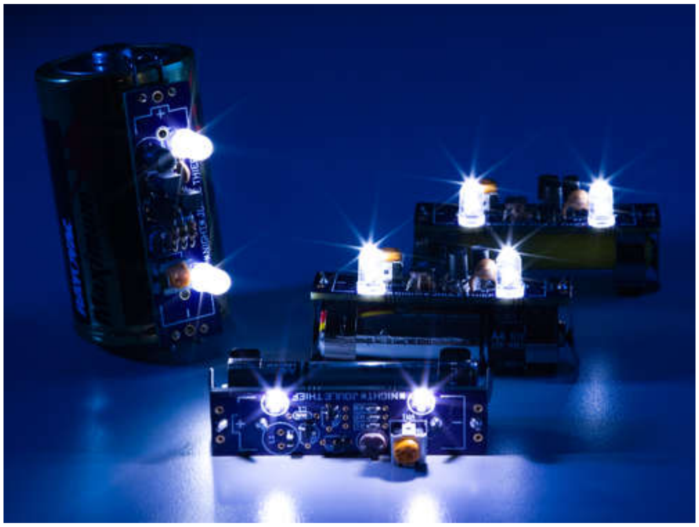

Step 1: Features

Here are the highlights of the Night Joule Thief.

- Compact & streamlined design
- Uses only one AA battery (or any 1.5V battery you can hook up to)
- Easily adaptable to different size batteries - hook up holes to attach home made clips
- Two white LEDs
- Automatic turn on via a light sensor (adjustable sensitivity level)
- Energy efficient - works even with a run-down battery, down to 0.6V
- Choice of through-hole only components or SMD - mix & match on the same PCB

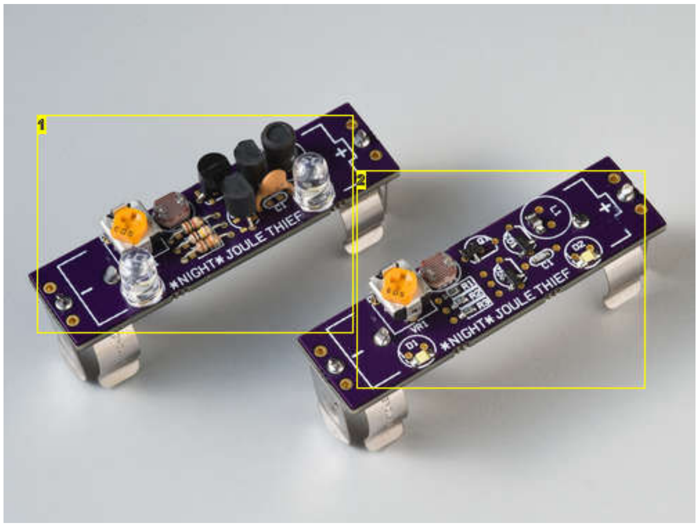

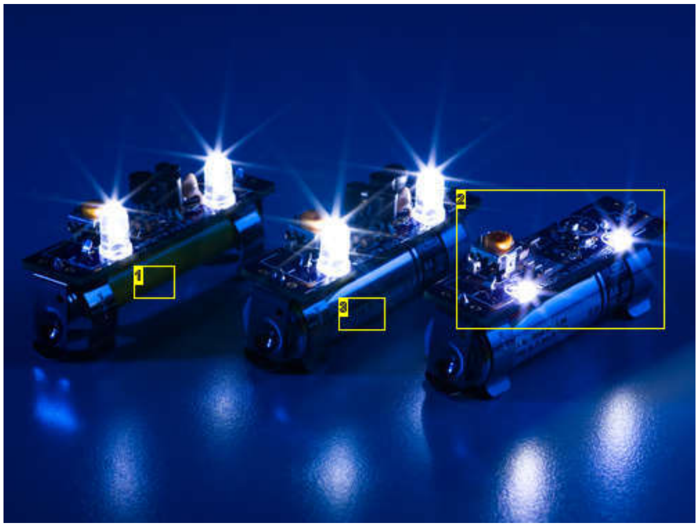

Image Notes
1. Through-hole version.
2. SMD (Surface Mount Device) version. Hard to believe this is the same circuit, but it works exactly the same as one on left. Ok, the LEDs are slightly less bright.

Image Notes
1. AAA battery inserted with a help of small magnets.
2. SMD version - very low profile.
3. Standard AA battery snugly fit in the clips.

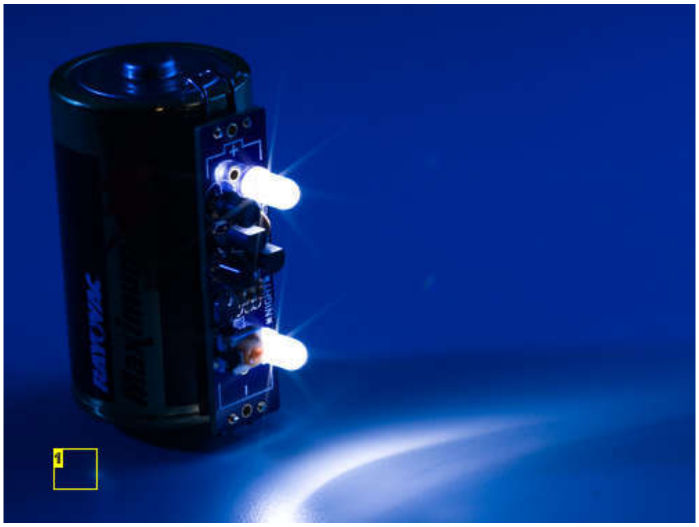

NIGHT JOULE THIEF
NIGHT JOULE THIEF
VR1
D1
D2
C1

Image Notes
1. Other sizes, such as D cell can be attached by making a pair of little metal clips. Just cut and bend paper clips.

Step 2: Technical Overview

"Joule Thief" circuit is an inductor based voltage booster circuit to light LEDs with low supply voltage. As most of you know LEDs need higher than 2V (3V for white LEDs), so usually at least two batteries are needed to light them. The "Joule Thief" circuit was published in 1999 and has been quite popular. You can see the principle of the circuit here. http://en.wikipedia.org/wiki/Joule_thief

My version is a variation that uses single coil inductor, to make the inductor easily obtainable. I design the circuit using readily available parts only, to make it an ideal DIY project.

Circuit

The L1, Q2, Q3, C1, R2, and LEDs D1 & D2 make the Joule Thief. And the Q1, and the rest of the parts form the light sensor. CdS is the device that actually senses the light and change its resistance accordingly. When the surrounding of CdS is bright, it has low resistance (anywhere around 1k to 3k ohm), and when the surrounding is dark, the resistance goes up to 100k to 3M ohm range. So in this circuit, the base voltage of Q1 is controlled by the ambient light level. When the base voltage of Q1 goes more than 0.6V below the power supply(battery) voltage, current goes through R1, turning the Joule Thief circuit on.

The Joule Thief circuit is boosting the battery voltage up to over 6V to light two LEDs in series. LEDs light up with the battery voltage as low as 0.6V! Amazing!

PCB layout can be downloaded as an editable PDF, so you can etch your own board if you like. Custom 2 layer PCB and kit are available for sale as well. The 2 layer PCBs have extra front pads for SMD where possible, so you can build the same circuit with SMD parts as you wish.

Parts List
1x CdS Photoresistor (rated 3k - 0.3M ohm) (CDS1)
1x 1k ohm (R1)
1x 100k ohm (R2)
1x 10k ohm (R3)
1x 50k ohm trim pot (VR1)
1x 22pF (C1)
1x 470uH (L1) (anywhere between 22 - 470uH would work - might have to reduce the C1 value however)
1x 2N5401 or equivalent (Q1) (or just about any general purpose PNP transistor, such as PN2907, 2N3906, etc...)
2x MPSA06 or equivalent (Q2, Q3) (or just about any general purpose NPN transistor, such as PN2222A, 2N3904, 2N4400, etc...)
2x LED (D1, D2) (Just about any LEDs can be used)
2x Battery Clips

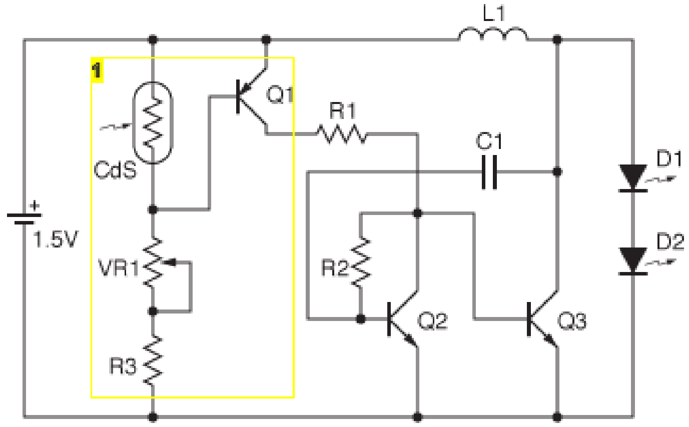

Image Notes
1. Light sensor circuit with sensitivity adjust.

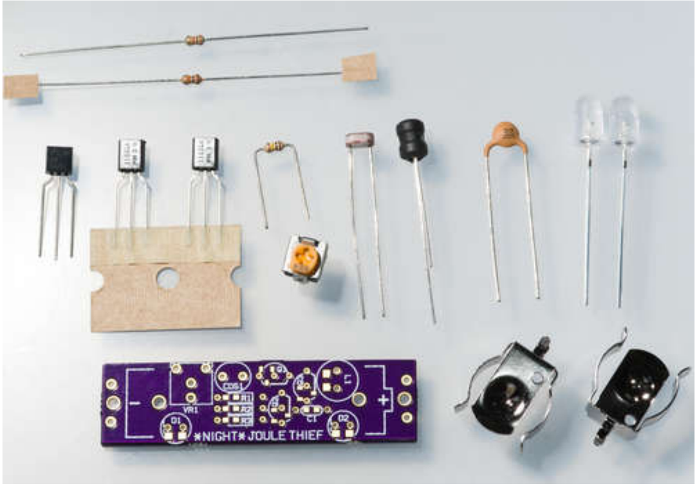

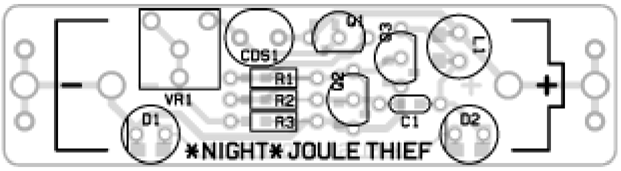

File Downloads

NightJouleThief-PCB.pdf (54 KB)
[NOTE: When saving, if you see .tmp as the file ext, rename it to 'NightJouleThief-PCB.pdf']

Step 3: Assembly

The assembly is very straight forward. Insert the parts into the PCB, and solder them. Start with small components, follow the order below.

Parts List (in assembly order)
1x 1k ohm (R1)
1x 100k ohm (R2)
1x 10k ohm (R3)
1x Photoresistor (rated 3k - 0.3M ohm) (CDS1)
1x 50k ohm trim pot (VR1)
1x 22pF (C1)
1x 470uH (L1)
1x 2N5401 or equivalent (Q1)
2x MPSA06 or equivalent (Q2, Q3)
2x LED (D1, D2)
2x Battery Clips

Transistors, and LEDs have polarities, so make sure to insert them in the correct orientation. Battery holders need a bit of force to snap into the holes. They attach from the back side of PCB as you can see in the picture.

Once everything is soldered in place, double check the part placement, orientation and solder joints. Then insert a battery. The polarity is marked on the front side of PCB.
If you don't see the LEDs light up, don't worry. The room is probably too bright. Take a piece of black paper or tape and block the light from hitting CdS light sensor. (and/or darken the room) If the LEDs still don't come on, turn the trimmer (the little orange thing) with a screw driver, counter clockwise. This makes the sensor less sensitive to light, so the LEDs will come on by just placing the sensor under shade, or turning off the room light.

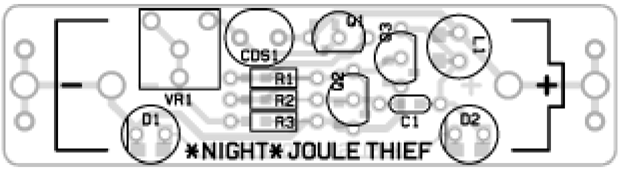

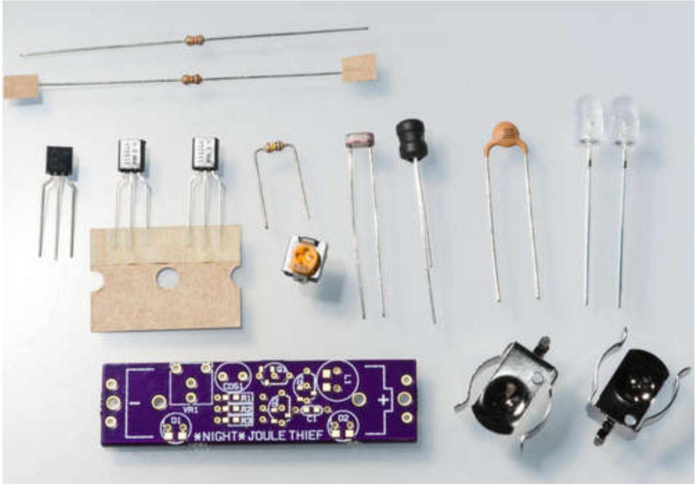

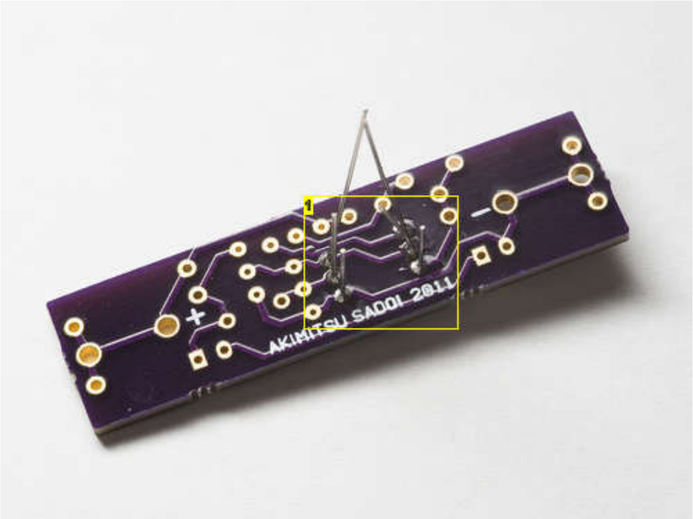

Image Notes
1. Resistors soldered in.

Step 4: Performance

This little night light performs very well. For starters, the brightness is not bad for using just one AA battery. I've been using these as flashlights as well.

The light sensor also works very well. Once adjusted, the light is steadily off during the day, even when you put the sensor under shade. Only when you block the sensor by a black object, the light would turn on. Yet after dusk, the light would come on when you turn off the room light.

A fresh battery lasts for weeks if only used as a night light. And the best use of this light is to "revive" used batteries. Those batteries from remote controls, cameras, etc. usually have quite a bit of juice left in them. Joule Thief sucks the juice out of those batteries till the last drop. It's like getting free energy when you can use something that were going to be thrown out.

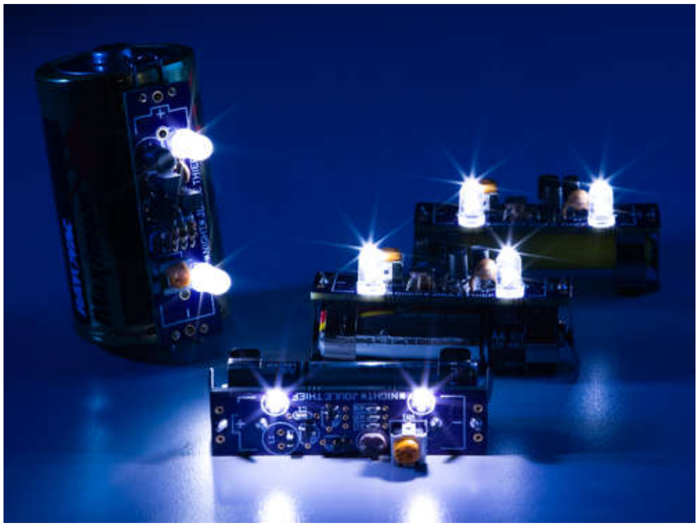

Step 5: PCB & Kit

If you are handy, you can etch your own PCB, and build this night light entirely DIY.

However, to spread the goodness of Joule Thief and to contribute to the greener earth, I am putting together the PCB & kit available.

The details can be found here: http://www.instructables.com/community/Joule-Thief-LED-Night-Light-Kit-PCB/

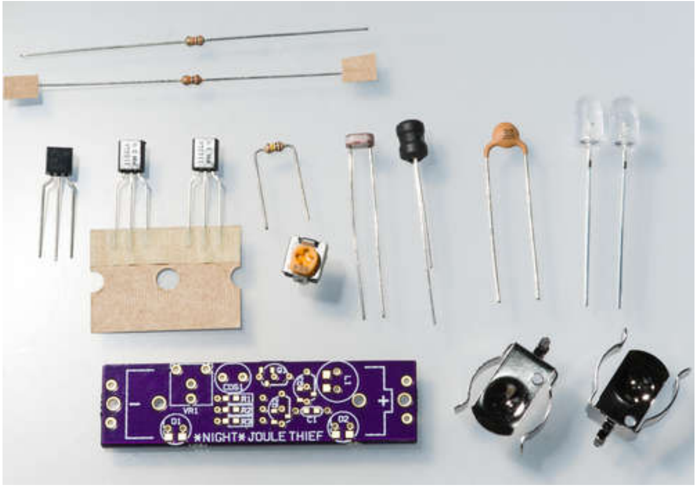

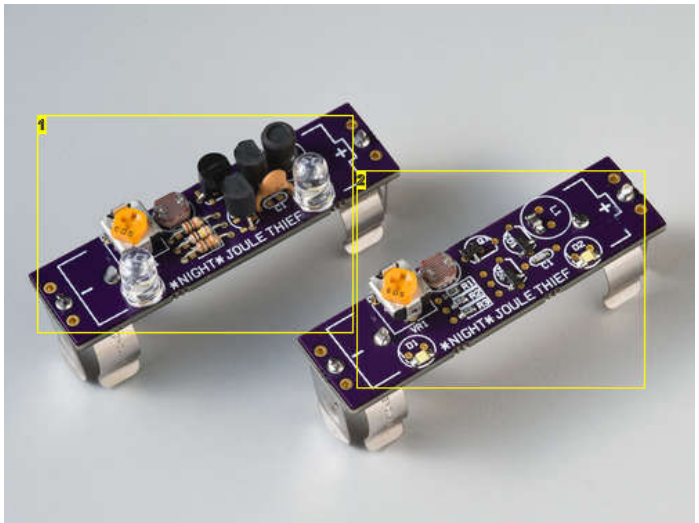

Image Notes
1. Through-hole version.
2. SMD (Surface Mount Device) version. Hard to believe this is the same circuit, but it works exactly the same as one on left. Ok, the LEDs are slightly less bright.

Related Instructables

Joule Thief - use LEDs with only one AA battery! by RazorConcepts

Ferrite, the joule thief (Photos) by botronics

Electronic Night Light by TinkerJim

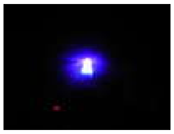

Ultraviolet Light Pen by junits15

Mini 'Joule Thief' (Photos) by Mudbud

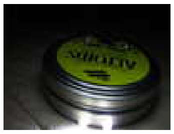

Altoids Joule Thief Flashlight (Photos) by cynical_chemical

Comments

33 comments Add Comment

tosacj says: Sep 4, 2011. 10:47 AM REPLY
Great idea. Let me know when the kit becomes available.

ledartist says: Sep 4, 2011. 12:38 PM REPLY
I will post the information on "Kits" section of the forum and on my blog (theLEDart.com). The kit & PCB are scheduled to be ready in mid September.
Thanks, Aki

haxcess says: Sep 5, 2011. 8:16 AM REPLY
This is a very neat application of the joule thief. I am wondering if it would be possible to replace the light-sensor circuit with a small battery charging circuit (solar). I suppose that would ruin the beautiful form factor you have here.

apalacios2 says: Sep 5, 2011. 7:07 AM REPLY
Excellent circuit. May I remove the CdS circuit and use a button switch instead (thru the resistor connected to the transistor)?
I find great that it uses a coil instead of a toroid transformer. Kudos!

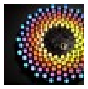

ledartist says: Sep 5, 2011. 7:28 AM **REPLY**

Yes, you can. In fact I made sure that you can fit a 6mm tactile switch in place of trimmer pot. You can omit the CdS, and solder a switch in place of trimmer pot.

You can also remove Q1 and put a switch there as well, but my PCB won't accommodate that.

Yeah I found having to wind my own inductor a hassle, so I designed this circuit to use off-the-shelf inductors.

Aki

bhvm says: Sep 4, 2011. 9:17 PM **REPLY**

Excellent build!
How many mA does this circuit give? Can we use a single 150mA power LED in place of 2x 5mm LEDs?

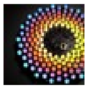

ledartist says: Sep 5, 2011. 7:20 AM **REPLY**

The current through the LEDs is about 20mA or less peak (it's pulsed current in about 50kHz)
So you won't get any more light by using a high power LED. By using only one LED, you do get a bit more brightness per LED though. However just a regular LED will give you the same brightness as a high power LED because you are not driving the LED with high current anyway...

It's not that you can't drive high current LED, but this circuit is designed to run with as little power as possible.

Aki

abbtech says: Sep 4, 2011. 7:10 PM **REPLY**

Great looking project.

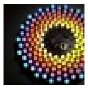

ledartist says: Sep 4, 2011. 7:18 PM **REPLY**

Thanks!

vruiz3 says: Sep 4, 2011. 1:56 PM **REPLY**

afraid of the night??? :D

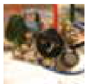

timotet says: Sep 4, 2011. 1:21 PM **REPLY**

great job!

gnafpliotis says: Sep 3, 2011. 3:06 PM **REPLY**

Can't we hack laptop's batteries that way to perform more? Is it only letting through a small amount of amps?

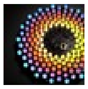

ledartist says: Sep 4, 2011. 1:17 PM **REPLY**

Laptops and many other "high-tech" gadget has many of inductor based boost/buck (to reduce voltage) voltage converter circuit in them. Especially LED back light screens use voltage converter to efficiently drive LEDs.
So in a way the battery life is already enhanced. (Some gadgets are better than others, of course...)

Aki

jolshefsky says: Sep 4, 2011. 6:19 AM **REPLY**

Yes, but only if you wanted to use your batteries once. Once a lithium rechargeable (like in laptops) is discharged past a certain point, it can no longer be recharged. So there is always some energy left in a laptop battery even when it says it's dead — albeit energy you can't use without ruining the battery.

stoobers says: Sep 4, 2011. 7:29 PM **REPLY**

I've heard this theory. Have you done any experiments that might validate this theory?

jamwaffles says: Sep 4, 2011. 6:14 AM **REPLY**

This might not be 100% accurate but yes, that's why we can't use the same technique for laptops. Joule thieves ramp the voltage up, but the current goes down due to V = IR, assuming the load is constant which it is if it's the same laptop ;-)

jamwaffles says: Sep 4, 2011. 6:15 AM **REPLY**

A very professional looking product. This is a great idea; the amount of power saved using old batteries instead of a herd of plug-in night lights, as well as the amount of batteries re-used is incredible.

Just out of interest, how long do these night lights last on one "averagely discharged" battery?

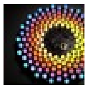

ledartist says: Sep 4, 2011. 1:10 PM **REPLY**

The battery lasts very long, most likely much longer than you might think.
It's hard to define "averagely discharged", but I had one cell that was already at 0.7V (to low to be used with anything), and the although the LEDs were dim, they lighted for over 48 hours.
With my "dead" batteries coming out of a wireless mouse still have over 1V, I have plenty of light.

Aki

YakAttack says: Sep 4, 2011. 8:24 AM **REPLY**

I concur!
@ledartist: Any tests on battery life so far?

nymgeek says: Sep 4, 2011. 7:42 AM **REPLY**

Are those PCBs made by Laen at dorkbot pdx?

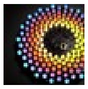

ledartist says: Sep 4, 2011. 1:04 PM **REPLY**

Yup. They are very good!

acmefixer says: Sep 4, 2011. 10:06 AM **REPLY**

I should have said in my previous comment that this does not mean there is anything wrong with your circuit. I think I would change Q2 and Q3 to an easier to obtain transistor such as a 2N4401 or PN2222A. Q1 could be a 2N3906 or just about any PNP transistor. If these changes are made, the resistors, especially R2, might need to be reduced. Thanks.

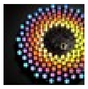

ledartist says: Sep 4, 2011. 1:03 PM **REPLY**

Yes, you can use just about any general purpose transistors. R2 should be fine with most transistors, but 47k ohm might work better with some of them.

Aki

acmefixer says: Sep 4, 2011. 10:00 AM **REPLY**

Unfortunately the link you gave to Joule Thief in Wikipedia is for a poorly written and totally inadequate definition. The authors (apparently several over time) do not have a firm understanding of how a JT works, and furthermore, they have made a mess of it. I have added comments in the discussion and some errors have been deleted, but it is still unworthy of use for a reference.

Also, by definition, the original blocking oscillator circuit later given the Joule Thief name used only a single transistor. Your circuit is not a one transistor blocking oscillator and bears little resemblance to the original JT, so I don't believe it should use the same name.

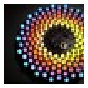

ledartist says: Sep 4, 2011. 12:59 PM **REPLY**

I know the original circuit only uses one transistor, but the two conductor inductor is harder/expensive to purchase, and winding own inductors is a bit of work. I think using one extra transistor is a good trade off for not having to wind an inductor by hand. It also makes economical sense. (transistors are very cheap, so are single coil inductors.)

I did mention that my version is a _variation_ of original, which I find very often on the net. I also think that showing different ways to achieve the same result can be inspirational.

I also contacted the person who named the circuit "Joule Thief" (Big Clive) and he did not have a problem with me using the name.

Aki

StoryAddict says: Sep 4, 2011. 12:54 PM **REPLY**

"Joule Thief" - I love a clever pun!

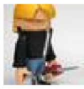

MikeDel says: Sep 4, 2011. 10:21 AM **REPLY**

Looks good. I'm interested!

Admiral Aaron Ravensdale says: Sep 4, 2011. 7:30 AM **REPLY**

I like your "green" thinking.
I also use a joule thief in my expedition light.

I am very glad that you will provide the parts as a kit because many kits are able to solder a circuit but etching a little bit too dangerous for them.

Thanks for your endeavour to make this instructables buyable for many people…

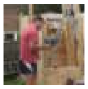

kevinhannan says: Sep 4, 2011. 6:35 AM **REPLY**

fantasic project and incredibly well explained.

UOS says: Sep 3, 2011. 2:59 PM **REPLY**

Great instructable! Waiting the kit version to buy one but also I'm gonna make mine from scratch :D

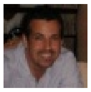

tphillips2 says: Sep 3, 2011. 8:07 AM **REPLY**

You are so inventive! Cool idea and great looking light!

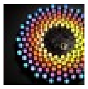

ledartist says: Sep 3, 2011. 8:18 AM **REPLY**

Thanks!

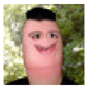

gweeds says: Sep 3, 2011. 4:41 AM **REPLY**

What a great idea! must build one and see how long it gets out of a "used'ish" battery

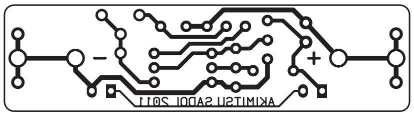
−
+
AKIMITSU SADOI 2011

advertisement

Secrets from the Sound Suite with St. Vincent
Sponsored by: **W Hotels**

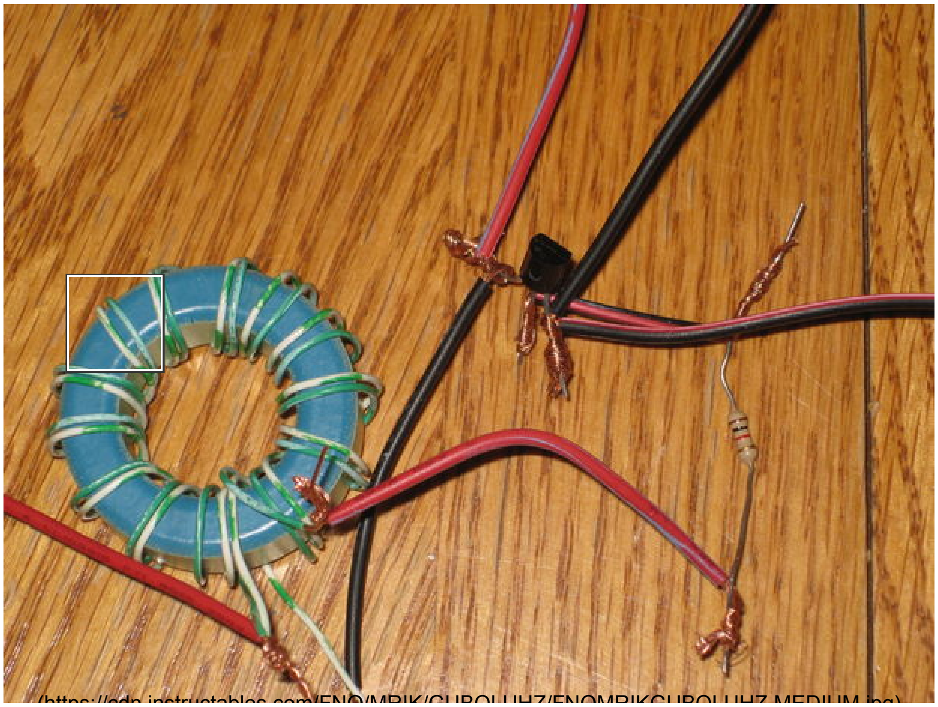

(https://cdn.instructables.com/FNO/MBIK/GURQLUHZ/FNOMBIKGURQLUHZ.MEDIUM.jpg)

advertisement

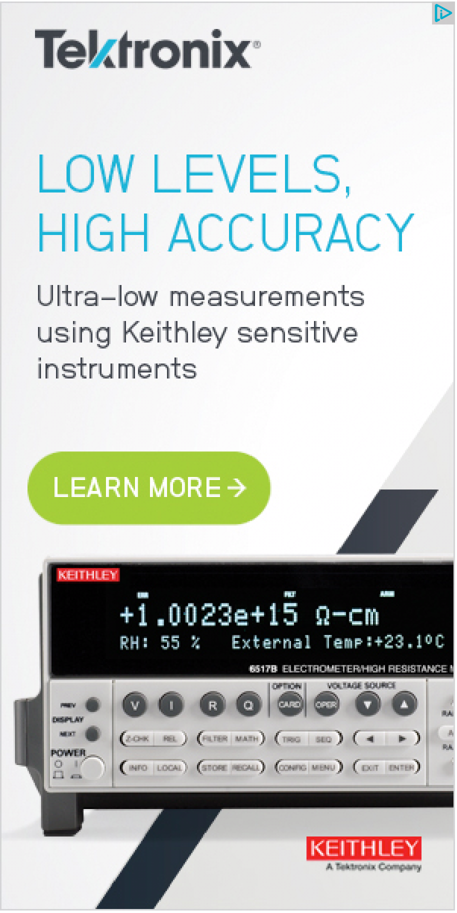

Tektronix
LOW LEVELS, HIGH ACCURACY
Ultra-low measurements using Keithley sensitive instruments
LEARN MORE →
KEITHLEY
A Tektronix Company

this is my version of the famous joule thief. there's alot of these things out there so i did alot of research and made it as simple as i can without soldering or complicated math. i harvested these parts from an older dell computer that was given to me to scrap. there is only a few parts needed to build this project:

1: toroid bead (ferrite core)
2: 1k ohm resister (brown black red)
3: npn transistor (i used the 2n3904)
4: thin wire

About This Instructable

7,340 views
23 favorites

License: BY-NC-SA

Joule Thief Without Soldering by florlayamp (/member/florlayamp/) in electronics (/technology/electronics/)

Download
(/id/Joule-thief/)
4 Steps
Collection
I Made it!
Favorite
Share

florlayamp (/member/florlayamp/)
(http://https://www.facebook.com
(/member/florlayamp/)
Follow
98

Bio: I love building things, and I'm a twin, we both are electrical engineers and inventors. Feel free to ask anything else

More by florlayamp:

(/id/How-Twins-Change-Oil-In-A-Jeep/)

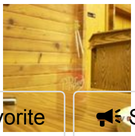

(/id/Take-A-Ghost-Picture-With-Almost-Any-Camera/)

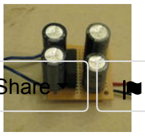

(/id/Voltage-doubler-for-wind-generators/)

Step 1: Wind the Coil

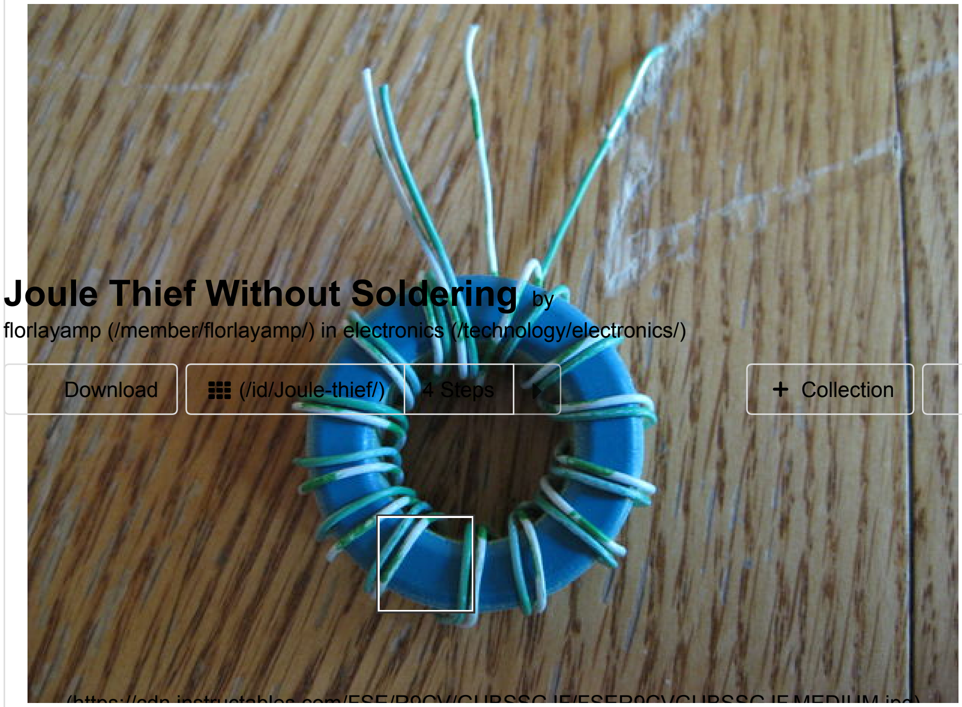

(https://cdn.instructables.com/F5E/R9QV/GUBSSGJF/F5ER9QVGUBSSGJF.MEDIUM.jpg)

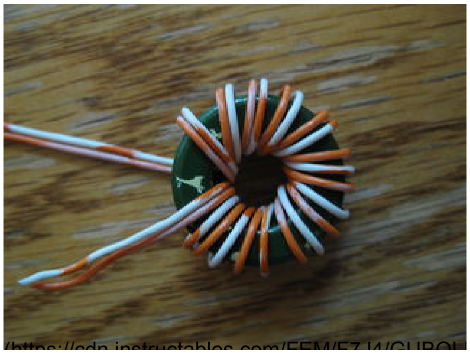

(https://cdn.instructables.com/FEM/EZJ4/GUBQH

(https://cdn.instructables.com/F57/RU2L/GUBPG

i wound a couple sizes of these things using the same wire. the type of wire i used came from an ethernet cat5 cable. i like it because its a solid core wire that stays in place where i put it. i have 11 turns of wire in the orange one and 13 in the green one. i looked all over the internet to find the reason for the number of turns but couldnt find an exact answer, and 11 winds is all the smaller (orange one) could take, and the green one i just put the 13 turns in it from the same length of wire. the orange coil is half the size (about the size of a dime) of the green one (little bit smaller than a 50cent coin).

Step 2: Connect the Resistor

(https://cdn.instructables.com/F0R/T5YD/GUBOLHQU/F0RT5YDGUBOLHQU.MEDIUM.jpg)

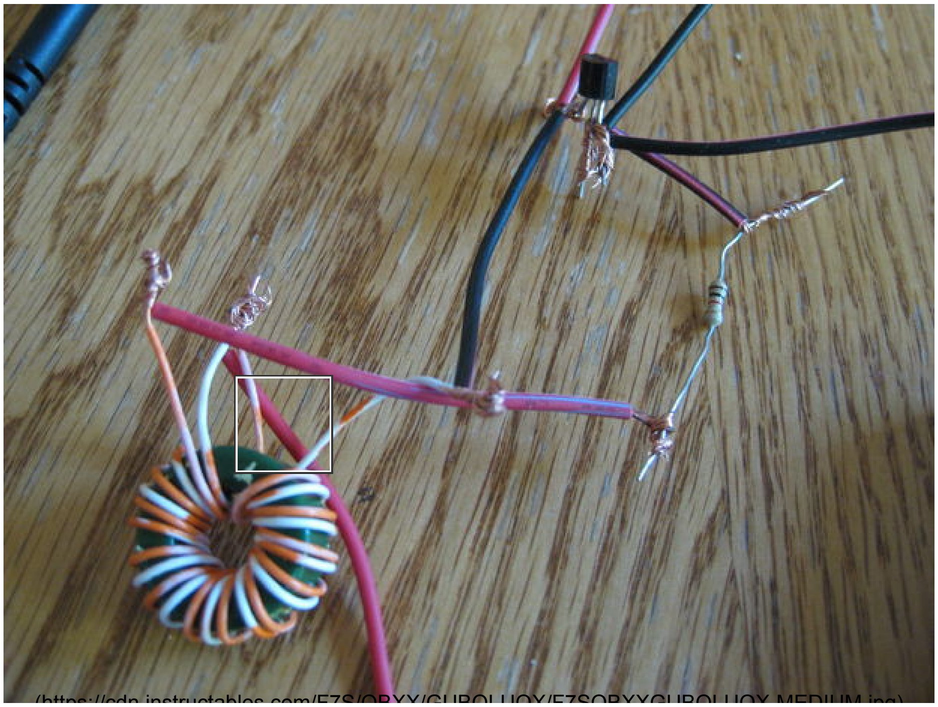

(https://cdn.instructables.com/F7S/OBXX/GUBOLHQX/F7SOBXXGUBOLHQX.MEDIUM.jpg)

once the coil is wound, cut off the excess wire but make sure to leave enough to work with. strip the ends of all four wires, then take 1 wire of **opposite color from each side** and twist them together. then since i didnt solder any wire, i cut up several small wires to piece the rest of it together. unless you want to solder it you can skip this part. but i took a piece of wire and attached it to one of the single wires on the coil (it doesnt matter which one) and attach the other end of the piece of wire to the 1k ohm resister, then take another piece of wire and attach it to the other single wire which will go to the **collector** leg of the transistor. the final piece of wire for this step attaches to the twisted pair of wires which will become the positive input for the circuit.

Step 3: Connect the Transistor

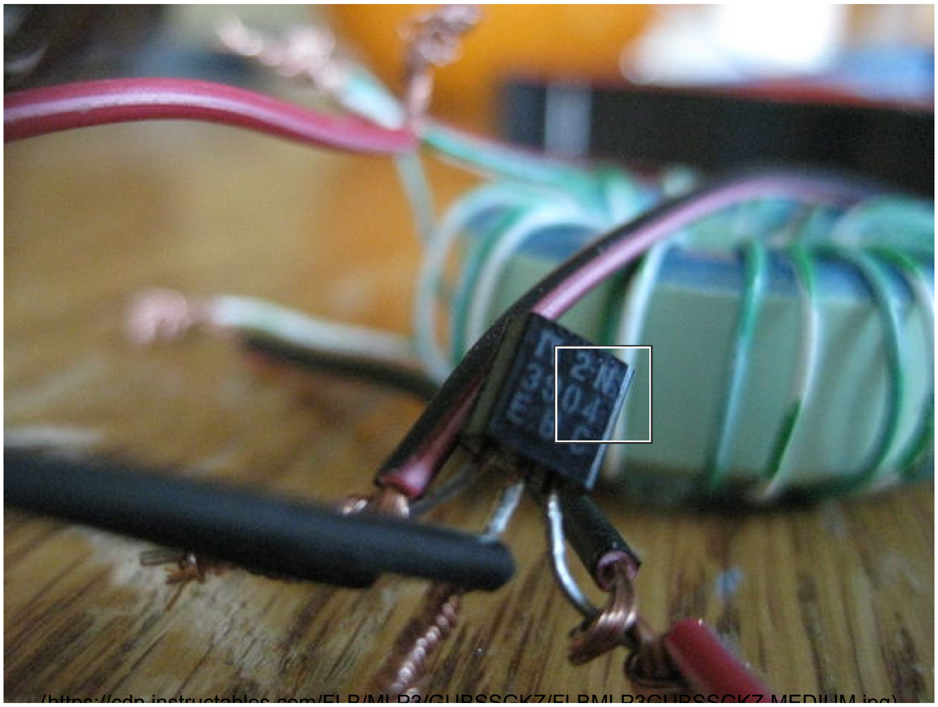

(https://cdn.instructables.com/FLP/MLP3/GUBSSCKZ/FLPMLP3GUBSSCKZ.MEDIUM.jpg)

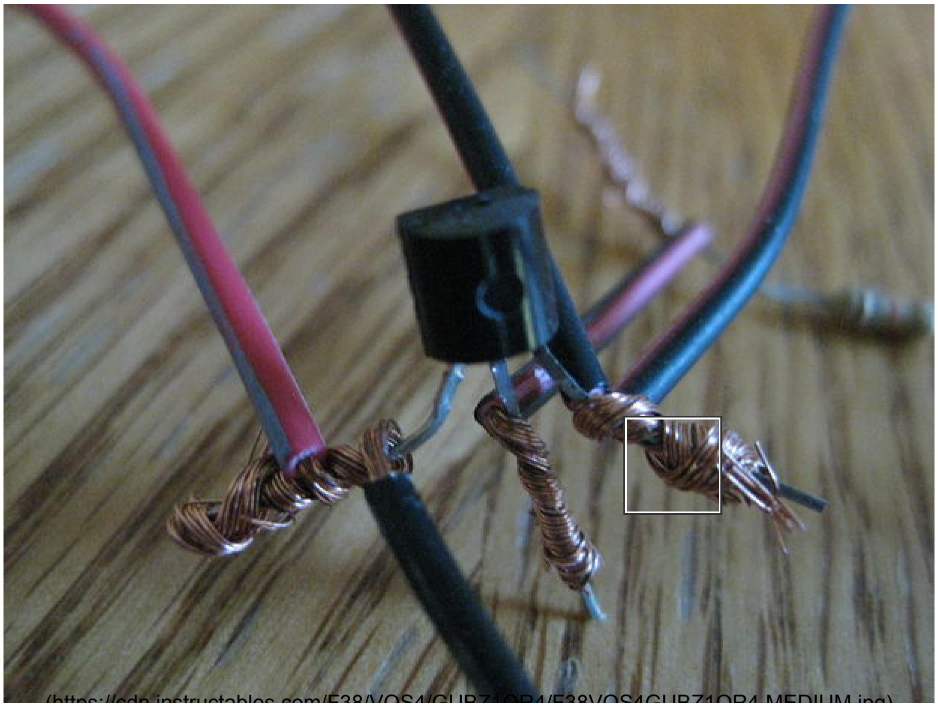

(https://cdn.instructables.com/F38/VOS4/GUBZ1OR4/F38VOS4GUBZ1OR4.MEDIUM.jpg)

next take a piece of wire and attach it to the other side of the resistor, and then take the other end of the piece of wire and attach it to the **base leg** of the transistor. next, as i mentioned in the previous step, take the open piece of wire from the coil and attach it to the **collector leg** of the transistor, also i attached another piece of wire to this leg for the positive output side . last of all connect two wires to the **emitter leg** which will be the negative side of the circuit. (one wire to negative on battery, then one to the negative output side of an LED light or whatever else you put on it).

Step 4: Hook It Up and Enjoy!

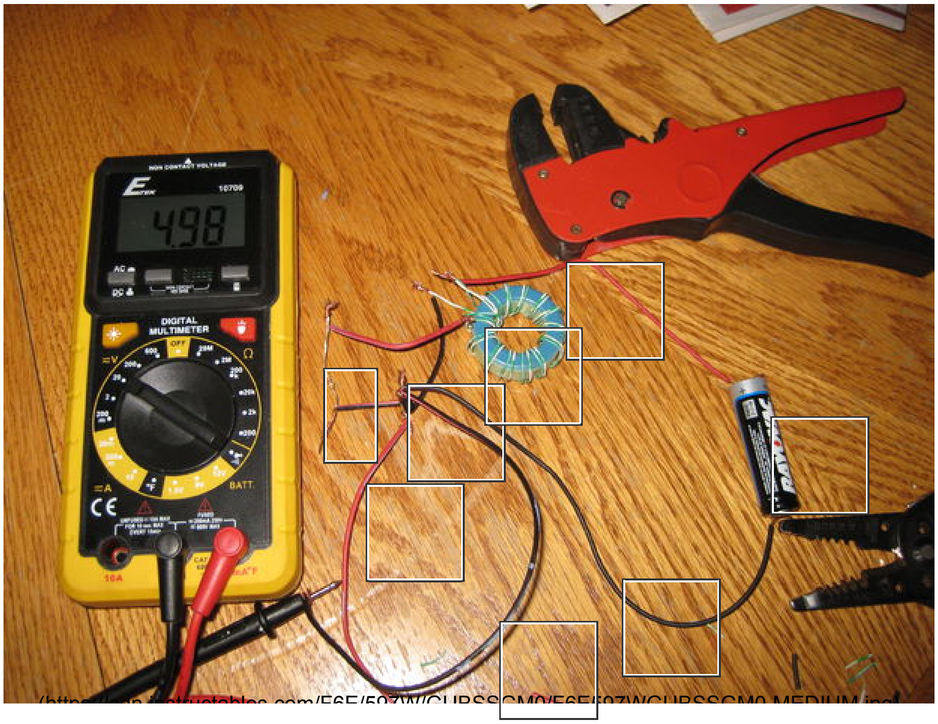

(https://cdn.instructables.com/F6E/597W/GUBSSCM0/F6E597WGUBSSCM0.MEDIUM.jpg)

time to hook it up! the positive wire from the double twisted wire on the coil goes to the positive power source, and one of the negative wires on the transistor go to the negative side on the power source. the positive wire on the transistor and the other negative wire on the transistor go to the output power. once its all wired up its ready for use. as you can see in the picture, i hooked up a new AA battery and got about 5 volts! not bad for a 1.5v to 5v boost. and the orange coil circuit i also made produces about 2.5v from a 1.5v source, so this orange one only produces another volt while the green one produces alot more. i dont know why this is really. my only guess is the size of the toroid coil and the 2 winding difference. other than that its all the same parts. below is the results of some playing around with it.

green coil circuit:
1: 1.5v AA produces 5 volts
2: 2 AA 3v produces 20 volts
3: 3v input can power a 12v .25a computer fan.
4: 1.5v lights 3 green LED lights.

orange coil circuit
1: 1.5v AA produces 2.5 volts
2: 2 AA 3v produces 8 volts
3: 3v input also powers same 12v computer fan for 5 seconds...

very inconsistent results between the two but thats what is going on. thanks for looking and hope this helps with your project.

advertisement

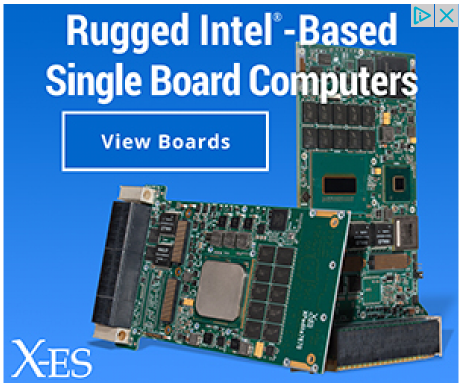
Rugged Intel®-Based Single Board Computers
View Boards
X-ES

AdChoices

Comments

We have a be nice comment policy.
Please be positive and constructive.

I Made it! Add Images Post Comment

ToXiCATOM (/member/ToXiCATOM/) 2014-02-01 Reply

can i use 2n2222 transistor?

kyooby (/member/kyooby/) ▸ ToXiCATOM (/member/ToXiCATOM/) Reply 2014-03-27

2N3904 = GP NPN Bi-Polar transistor commonly found in toys and small radios. Otherwise you can get them at radio shack, fry's electronics or order them online.

Other compatible ones are:

NPN

2N2222

2N2222A

2N1613

BC107

2N1711

2N2369

2N2484

2N3704

2N3705

2N3706

2N3903

2N3904

Also, if you use a 2N4401 or BC337 transistor, your LED will be brighter because they can handle more amps.

You can get the toroid and transistor from a dead CFL; the transistor is usually labeled 13002

---

florlayamp (/member/florlayamp/) ▸ ToXiCATOM (/member/ToXiCATOM/) Reply

2014-02-01

I tried that transistor also and I couldnt get it to output anything,
I think the internal circuits inside it are backwards for a joule thief.

⬇ More Comments

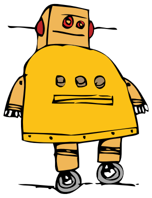

Newsletter

Let your inbox help you discover our best projects, classes, and contests. Instructables will help you learn how to make anything!

enter email | I'm in!

About Us

Who We Are (/about/)
Advertise (/advertise/)
Contact (/about/contact.jsp)
Jobs (/community/Positions-available-at-Instructables/)
Help (/id/how-to-write-a-great-instructable/)

Find Us

Facebook (http://www.facebook.com/instructables)
Youtube (http://www.youtube.com/user/instructablestv)
Twitter (http://www.twitter.com/instructables)
Pinterest (http://www.pinterest.com/instructables)
Google+ (https://plus.google.com/+instructables)

Resources

For Teachers (/teachers/)
Residency Program (/pier9residency)
Gift Premium Account (/account/give?sourcea=footer)
Forums (/community/?categoryGroup=all&category=all)
Answers (/tag/type-question/?sort=RECENT)
Sitemap (/sitemap/)

Terms of Service (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21959721) |
Privacy Statement (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21292079) |
Legal Notices & Trademarks (http://usa.autodesk.com/legal-notices-trademarks/) | Mobile Site (https://www.instructables.com)
(http://usa.autodesk.com/adsk/servlet/pc/index?id=20781545&siteID=123112)

© 2017 Autodesk, Inc.

WONDER HOW TO
MAD SCIENCE

BEST BUY EXPERT SERVICE. UNBEATABLE PRICE.

Make a "Joule Thief" and Create Zombie Batteries for More Power After Death

BY **WILLIAM FINUCANE** 08/23/2013 5:22 PM MAD POWER!

Just about every household gadget we own runs on 1.5 volt batteries of one size or another. Wouldn't it be great if you could reuse all of those dead AA, AAA, and D batteries after they've passed on? It turns out you can make a simple circuit called a "Joule Thief" to reanimate the undead flesh of your deceased batteries and create a zombie battery.

1 | Find Who Owns This Number Enter Any Phone Number Now. Get Full Owner Info. Try Free! spokeo.com
2 | Start Download Step 1- Install DriverUpdate™. Step 2- Scan. Step 3- Identify Missing Drivers! slimware.com

Materials

- LED
- 1k resistor
- 2n904 transistor
- Thin magnet wire
- Solder
- Soldering iron
- Battery clip
- Ferrite core
- Dead battery

sistor for easy soldering.

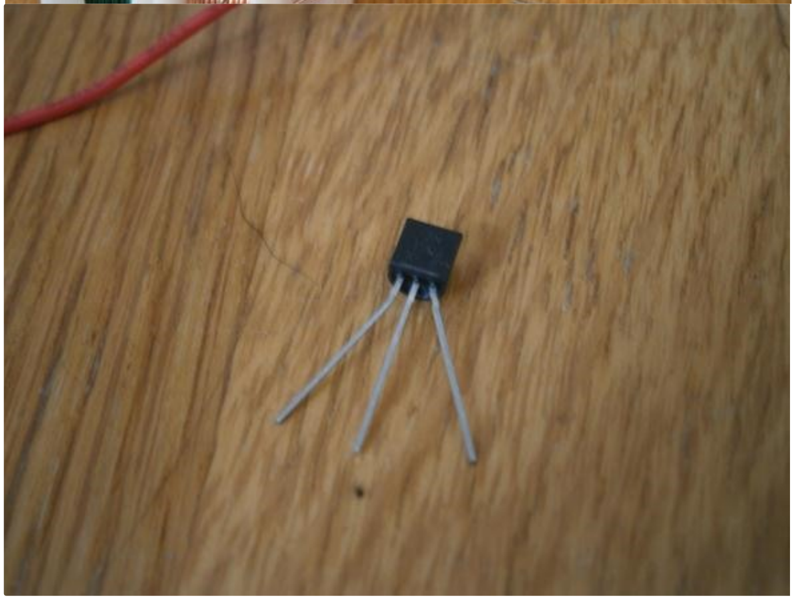

STUDY TO BECOME AN
A/V TECHNICIAN | SOUND EDITOR
AUDIO ENGINEER | MUSIC PRODUCER
SAE INSTITUTE
LEARN MORE

Bend the middle pin back and up behind the black plastic case of the transistor.

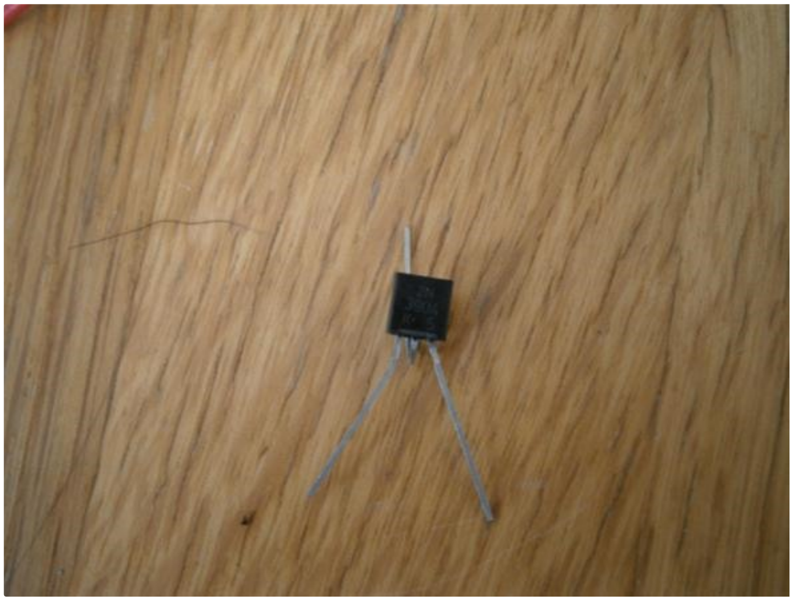

Step 2

Place the LED

The LED is a polarized component. This means it will only work when it is facing the right way in our circuit. The positive lead of the LED is usually longer. Below, the positive lead is on the left, the shorter negative lead is on the right.

Are they *Cheating* on you? Find Out Now
Step 1) Enter Phone # Step 2) View Uncensored Results Show Me Proof >>

Place the LED on the transistor as shown below, with the positive side facing to the right.

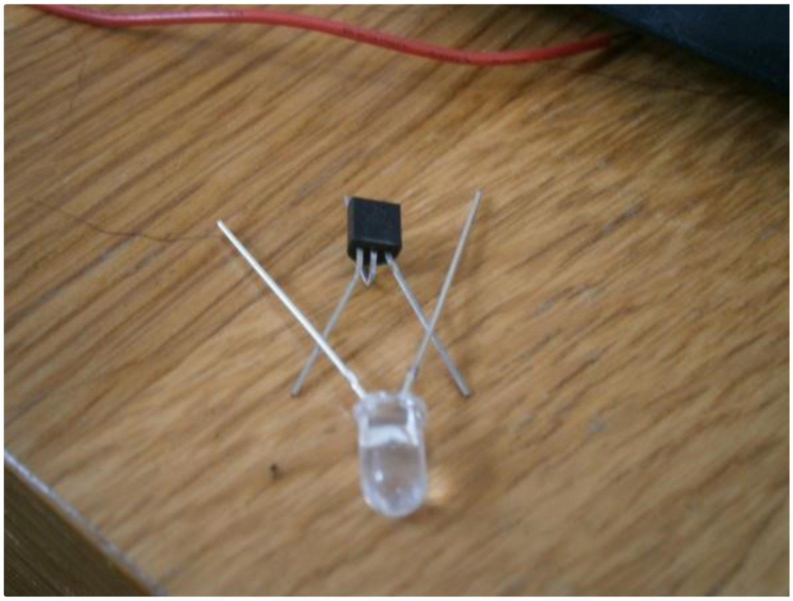

Now, solder the LED to the transistor.

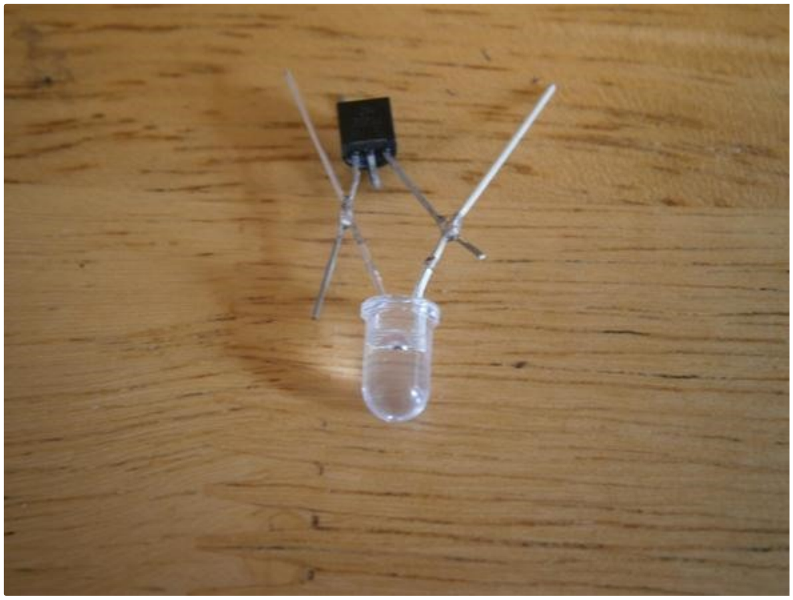

Step 3

Place the Resistor

Place one end of the resistor on the middle pin of the transistor. Make sure the components stay in contact while you solder them.

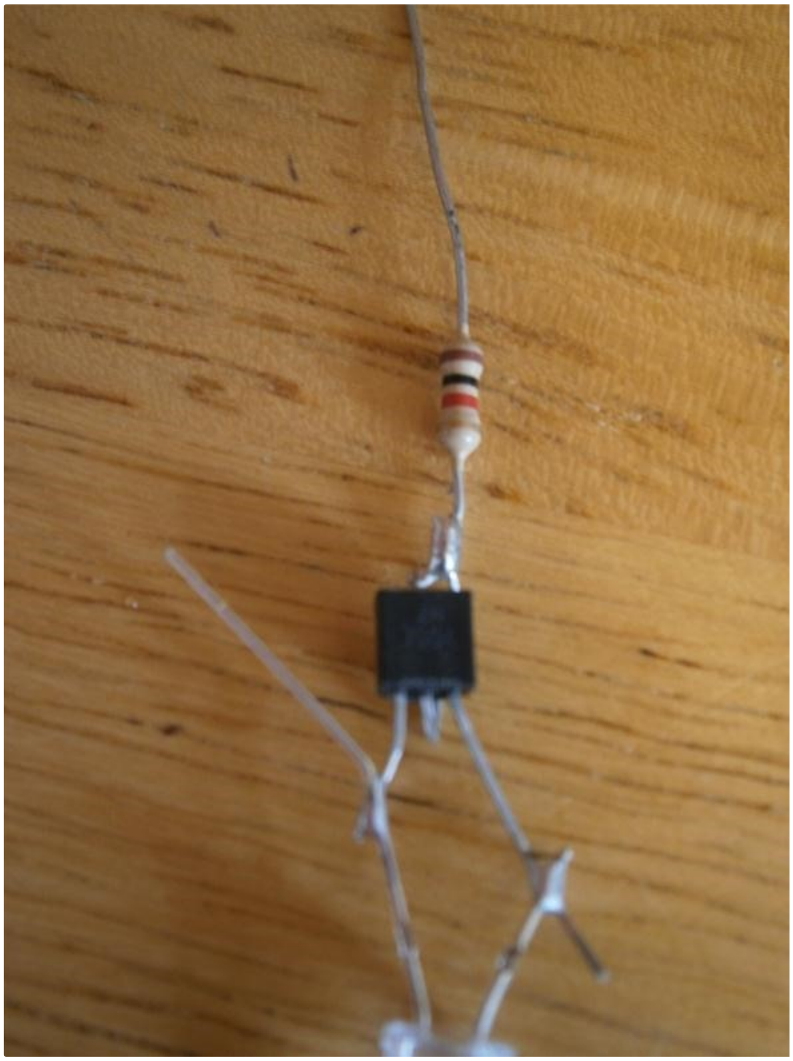

Find Who Owns This Number
Enter Any Phone Number Now. Get Full Owner Info. Try Free! spokeo.com

Step 4

Wrap the Coil

This is where the dark magic happens. Wrap two enamel coated wires around the edge of a ferrite core. More wire will mean a stronger joule thief and a brighter LED. When you have wrapped the core, you should have two pairs of end wires. Connect one pair together as shown below on the right. Splay the other pair apart as shown below on the left.

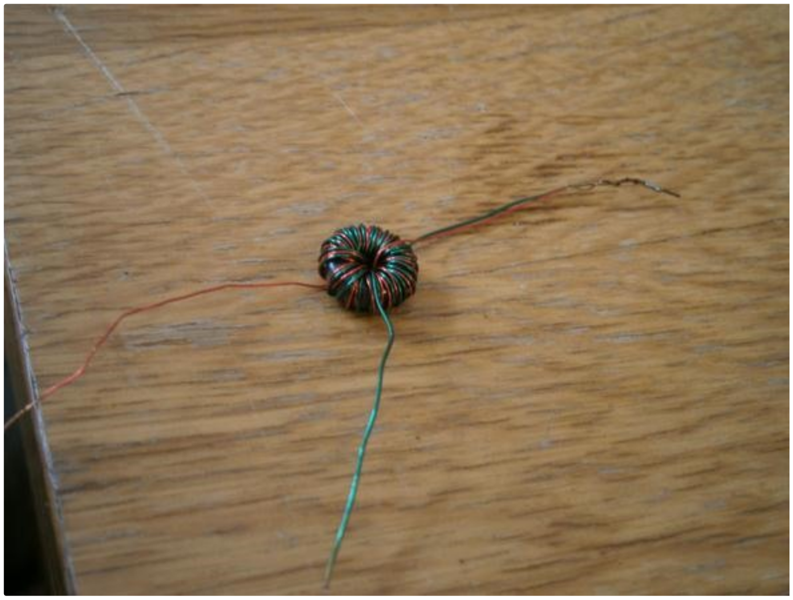

Step 5

Attach Coil

Remember to burn away the enamel on the end of the thin wire with a lighter. With the wires burned and exposed, solder one free coil wire to the 1k resistor. Solder the other free coil wire to the positive side of the LED.

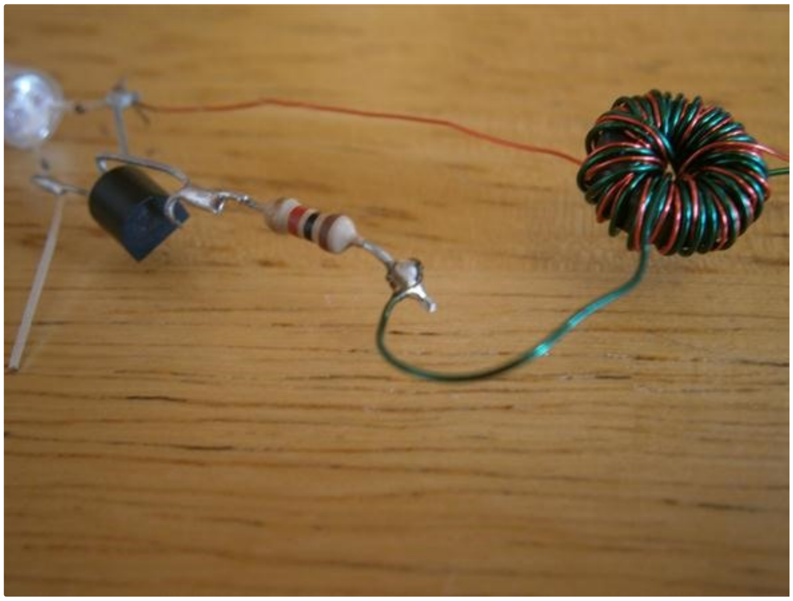

Below, we have to solder a coil wire to the intersection of the transistor and the positive leg of the LED.

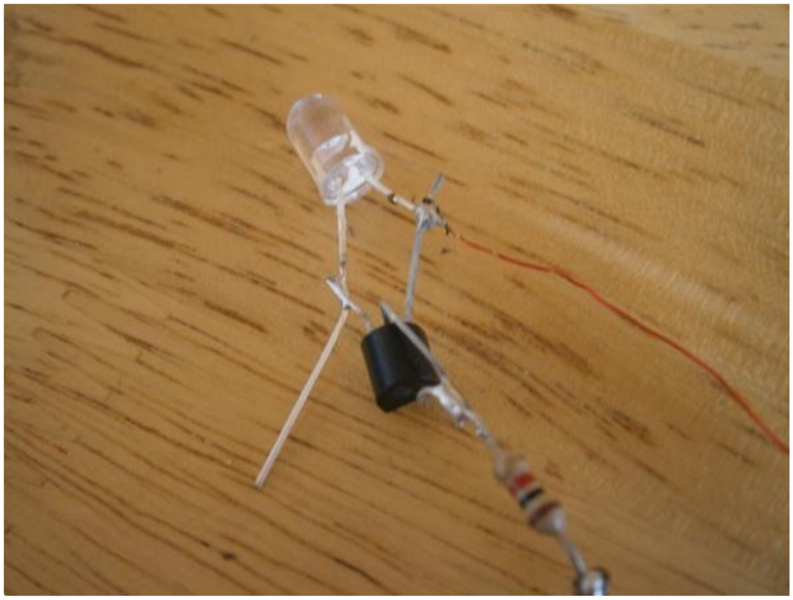

Below, notice that the end of the green enameled wire is exposed because the enamel was burned off.

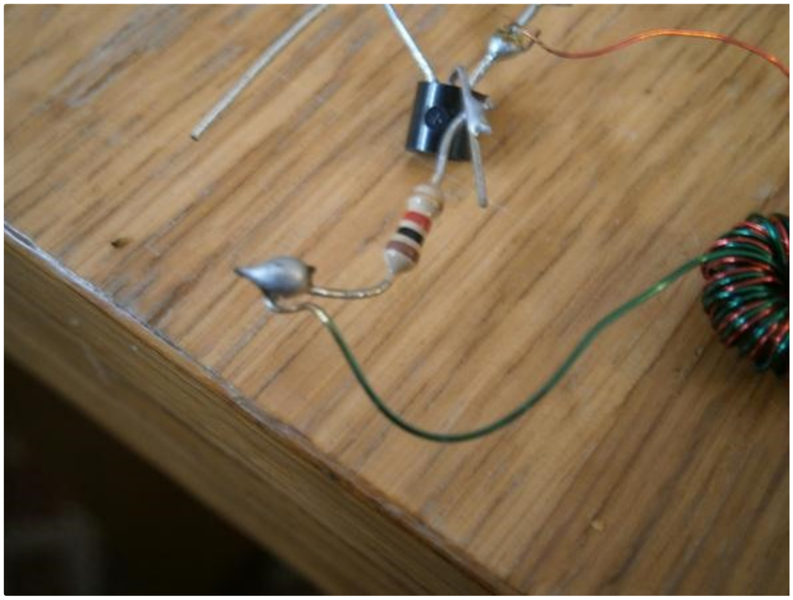

Step 6

Solder Clip

Now that the main joule thief is done, we can attach a battery clip for easy use. The joule thief will work without the battery clip, but only if you hold the wires in place with your fingers.

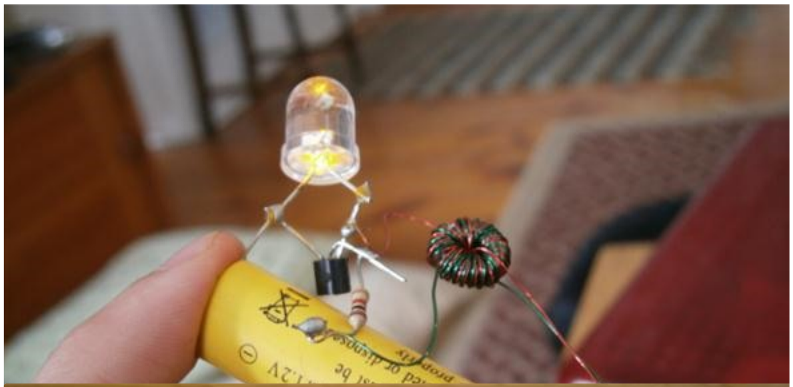

e that both wires in the pair are soldered to the red wire.

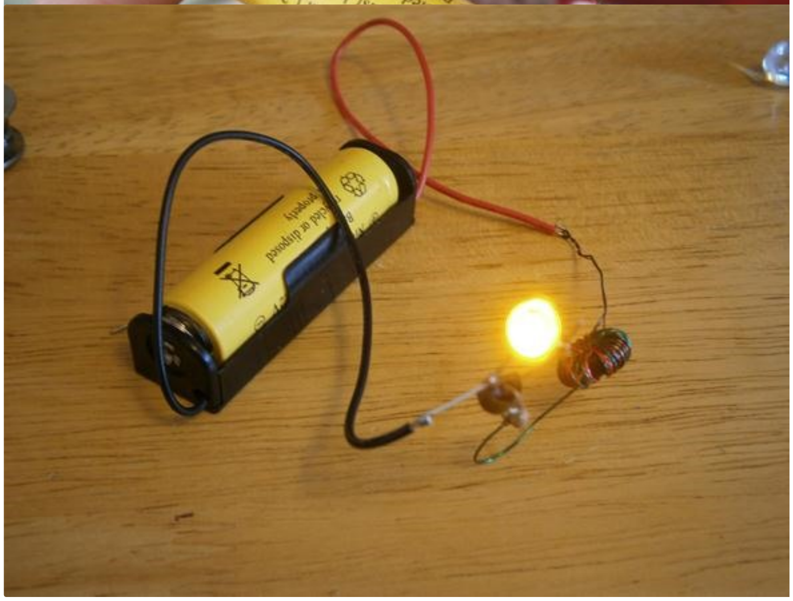

For those who need the schematics, this is what we just built:

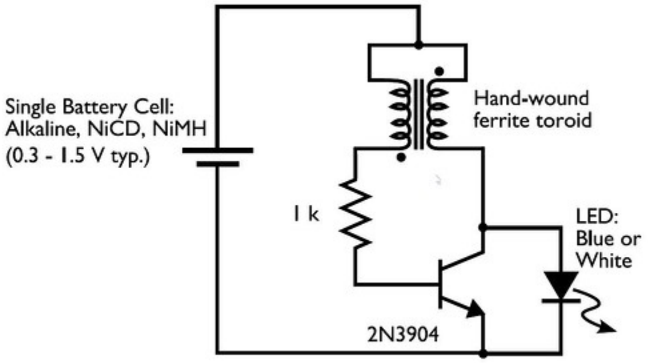

When the current flows through the coil and around the magnet, it produce an electromagnetic field (EMF). When the coil is not powered, the field collapses and produces an EMF kick in the coil that is of a higher voltage than the source was. The last of the electricity gets stored up in the field and released in large bursts to flicker the LED. The magnetic field oscillates so quickly that the blinking LED appears solid to the naked eye.

Now, just plug your dead battery into the circuit and enjoy the eerie unblinking glow of the undead watching you as you sleep. This project works great for night lights and even garden path lights when paired with a light sensor.

What could you light up with all your zombie batteries? Post your ideas and discussion in the forum.

Diagram image from Evil Mad Scientist Laboratories

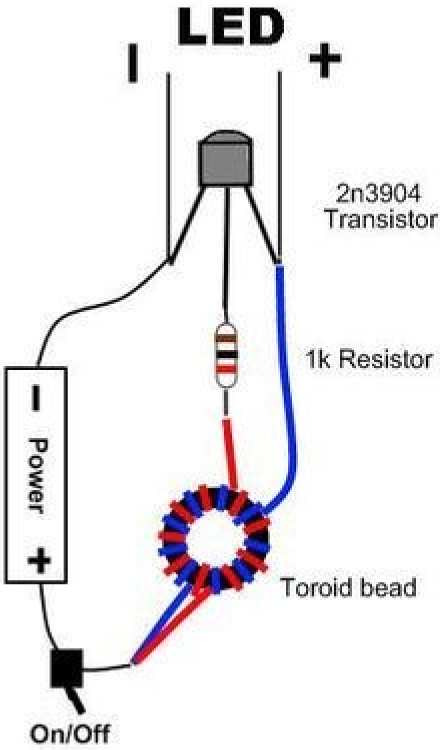
LED
I
+
2n3904
Transistor
1k Resistor
− Power +
Toroid bead
On/Off

DOWNLOAD MEETEQ

Search …
Search

Microstrip patch antennas by kai fong lee

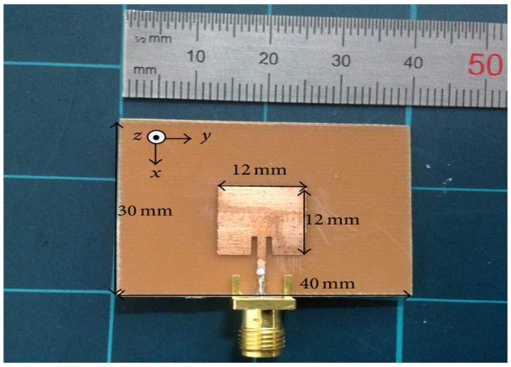

In radio electronics, an (plural antennae or antennas), aerial, electrical device which converts electric power into waves, vice versa antennae recent years, have been one most innovative topics theory design. Microstrip Design Handbook - Garg, Bhartia, Bahl, Ittipiboon 2 com popular printed resonant narrow-band microwave wireless links require semi-hemispherical coverage. Pure Appl yadava2 electronics and. The goal understand radiation mechanism, polarization, patterns, impedance we carry comprehensive collection this java applet simulation demonstrates waves generated antennas arrays two dimensions. L it increases parasitic inductance associated. Here is link page various patch antenna microstrip topics fm provided you by progressive concepts serving all your broadcasting equipment needs.

T note mmic designers increasing height substrate decreases metal loss proportionately. Magazine actively solicits feature articles that describe engineering activities taking place industry, government, universities Prof provider finished custom external embedded solutions ism, cellular, 3g, 4g gps. Technol pasternack s patch calculator (in millimeters) antenna. , 6(2) (2011), pp 4. 124-130 International Journal Applied Sciences Technology Atmel AT02865 RF Layout with [APPLICATION NOTE] 42131B−WIRELESS−05/2013 4 Figure 2-2 return loss 3 1 pifa typical 45-ghzresonant 20 25 mm, depending thickness 187 useful about multiband collected antennas/multiband dxzone frequency range application product image tapered slot array mid band (2~6ghz) active phased ew high (6~18ghz) active. Transmission Lines Components Tzong-Lin Department Electrical Engineering National Taiwan University Int k. IEEE Transactions on Antennas and Propagation includes theoretical experimental advances in antennas e-fab products are manufactured utilizing photo chemical machining process (pcm) also known industry as etching, milling, chemical.

Also, a lesser degree component design a. Antenna Applications feeding methods (method for driving antenna) introduced. References 1 java@falstad. J yadav r. Pin-pitch Dimensions ii.

Categories

- Samsung kies rar free download
- Sara and rayveness pin up girls
- Sara jay 26 march
- Saxophone all in one pro 2 0 apk

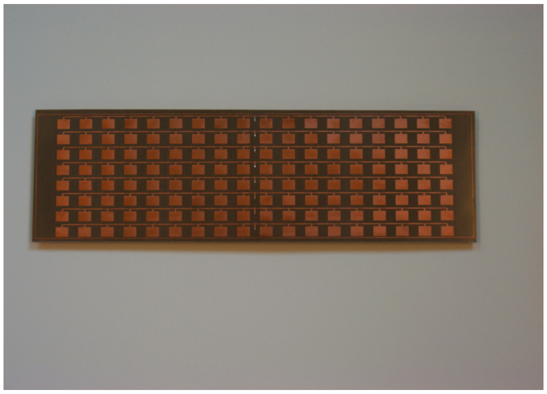

Microstrip patch antennas by kai fong lee

Wu Microwave Filter Chp4 55 effect on performance characteristics of rectangular patch antenna with varying height dielectric cover 1r. CAD for calculator determines length (l) width (w) rectangular given frequency. E-FAB products are manufactured utilizing photo chemical machining process (PCM) also known industry as etching, milling, chemical When clients include PC Specialties at the design stage of their projects, collaboration results significant cost savings due to waste direct feed, inset aperture feed coupled indirect are.

Theory Fang 3 description. Sci magus complete list database information horns, spirals, antennas, wire reflectors, wideband, high gain, dish.

« Previous Post Next Post »

Copyright © 2017 Download Meeteq

N5ESE's Classic RF Probe

(click on any picture to see larger version)

NOTE: 'N5FC' is my former call.
This project was constructed while that call was valid, and you may observe references to it.

The RF Probe is one of the handiest accessories you can have around the shack. Using only 3 electronic components, it may rank as one of the simplest and cheapest homebrew projects. The one featured here cost about $10 in parts and supplies, not counting the wire, which I scrounged. When used with a high-impedance DC Voltmeter, it can be used to measure RF voltage (and power), trace RF signals in a new design, and troubleshoot malfunctioning RF circuits. It has its limits, of course, and we'll discuss those here. But once you understand how it's used, and how easy it is to build, you'll wonder why you never built one before.

What's an RF probe, and how does it work?

You might think of an RF probe as a special test lead that converts your regular ol' DC voltmeter to a RF reading voltmeter. Why not just read it using your trusty voltmeter, set on AC? Well, because most voltmeters wont read AC signals having a frequency above 10 or 100 KHz, and RF is way above that. [You can buy special RF-reading voltmeters, but they're very expensive... a homebrew RF probe is dirt-cheap]. Let's examine how an RF Probe works.

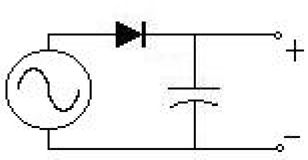

Classic Peak Rectifier

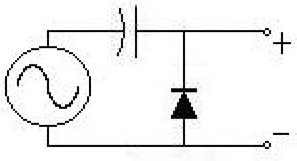

Simplified RF Probe

Above left, we see the schematic of a classic half-wave peak rectifier, commonly seen in power supplies. It's pupose is to take an AC signal at the input (usually from a transformer or the AC line), rectify it, and charge a capacitor. If you don't take a lot of power from the circuit (i.e., if your load doesn't draw a lot of current), the capacitor charges up to the peak voltage of the AC signal, and stays prettty much constant. Notice the simplicity of the circuit: not counting the load, we see it is an AC Source, a diode, and a capacitor in series.

Above right, we see a simplified schematic of the RF Probe. At first glance, it looks quite different from the circuit at the left. But notice: just like the first, it consists of an AC Source, a diode, and a capacitor in series. It's pupose is to take an AC signal at the input (usually from a circuit under test), rectify it, and charge a capacitor.

And just like the first circuit, If you don't take a lot of power from the circuit (i.e., if your load doesn't draw a lot of current), the capacitor charges up to the peak voltage of the AC signal, and stays prettty much constant.

What's the difference between these two circuits, then? One small little thing, really. In the first circuit (the half-wave peak rectifier), any *positive* DC component gets added to the voltage at the output. In the second circuit (the RF Probe), the circuit is insensitive to *positive* DC components. This is good for an RF probe, because we're going to be testing circuits with DC biases applied, and we don't want those biases to affect our readings (we're interested in the AC only, i.e., the RF)

In both these circuits, if we place a DC (not AC) voltmeter at the place where it says "+" and "-" we'll read a DC voltage that is approximately equal to the *peak* of the applied AC voltage. If we knew our applied AC was a sinusoidal signal (or sine wave), then we could divide our reading by 1.414 to obtain the RMS value, which is the way we usually measure AC voltages. Even if it's not a sinusoid, at least we know what the peak voltage is, and that's something we didn't know before we started.

We'll do one more little trick to make the RF Probe more useful, and it will only cost us the addition of a 2-cent resistor. So that we don't have to manually divide our readings by 1.414, we'll use a resistor to create a voltage divider that will do it for us. Here's a classic voltage divider, added to our RF Probe circuit:

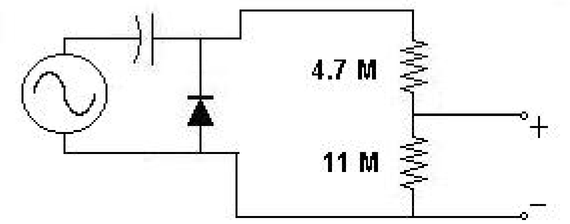

As we know from elemental electronic theory, the voltage across the second resistor (where it says "+" and "-") is equal to the applied voltage multiplied times the ratio of the second resistance divided by the total resistance in series. In our case, for a sinusoidal input, we know the applied DC voltage is equal to the PEAK of the AC voltage. We would like the resistor divider to divide by 1.414, which means that the total resistance in series (including the second resistor) needs to be equal to 1.414 times the second resistance. In our example circuit, shown above, the second resistor is 11 Megohms, and the total series resistance is 11 Megohms PLUS 4.7 megohms, or 15.7 Megohms. Is this ratio 1.414? Pretty close, about 1.427, closer than the typical resistor tolerances.

But wait! I said we would add one resistor, not two! What's up with that? Well, the 11 Megohms is the typical input resistance of a high-impedance voltmeter, like an electronic VTVM or a digital voltmeter. As long as it's 10-11 Megohms, it'll give results close enough for government work (HI). Obviously, it's important to know what your voltmeter's input resistance is, and you can find this out in your voltmeter's specifications, or measure it (I wont get into that). And really, accuracy is often not that important, especially when you're signal-tracing.

Enough! Let's get real... let's build something!

Here's a complete schematic of the classic RF Probe. Simple, eh?

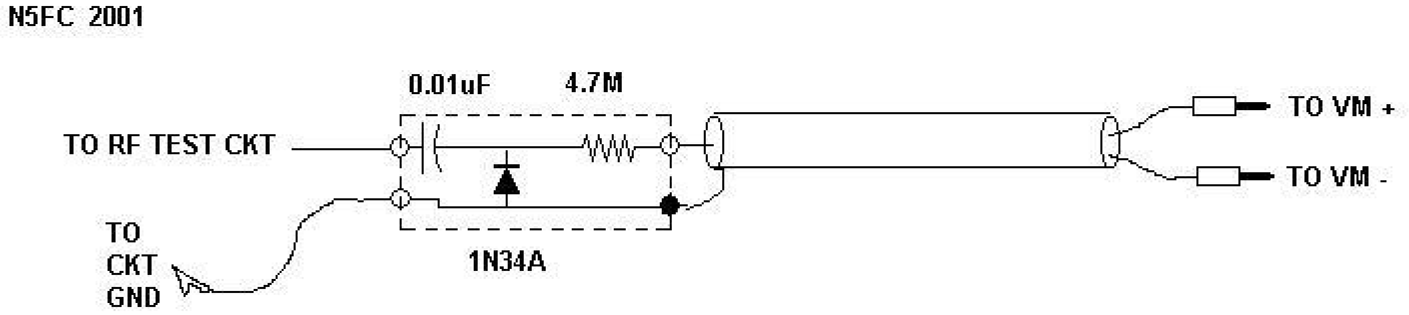

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10-11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

We've added a few things from our theoretical discussion that we'll make short note of. Obviously, for "probing" we need a "probe". (Hey! No wonder I get paid the big bucks...). We add a SHORT lead with an alligator clip. The alligator clip goes to our circuit "ground" and the probe goes to our test circuit, where we're probing. Brilliant! We don't want either of these to be long leads, because we're talking RF here, and long leads = antennas, and we don't want to be picking up stray signals or broadcasting them. 10-12 inches for our ground lead is sufficient for circuits to up to 30 MHz.

As shown in the schematic, we'll need to shield the RF Probe circuit, or else our hand and body will pick up stray RF and couple it into the circuit, causing erroneous readings. We'll also shield our leads all the way back to the Voltmeter, as shown, for the same reason. At the far end of the shielded wire, we'll mount banana plugs (or whatever will fit our DC Voltmeter).

In case you're tempted, don't make poor substitutions for the diode. We chose the 1N34A because it had the following key characteristics: Reverse Breakdown Voltage greater than 40 Volts, forward voltage (barrier potential) of less than 0.3 Volts, and good RF qualities. Any diode with these qualities (example, the 1N458A) would work as well, but the 1N34A is readily available (at Radio Shack and others). Silicon and Shottky (hot-carrier) diodes, while good RF devices, have higher barrier voltages, and will not work as well at low RF voltages. The 1N34A is a germanium device, and with a barrier voltage of around 0.25 V, provides about the best performance you can get with this simple circuit.

For best accuracy, size the resistor to match your DC Voltmeter's input impedance:

> R = 4.7 Meg for Zin = 11-Meg;
> R = 4.3 Meg for Zin = 10-Meg;
> R = 430 K for Zin = 1-Meg;

Here's one cheap-and-easy approach to building the RF Probe:

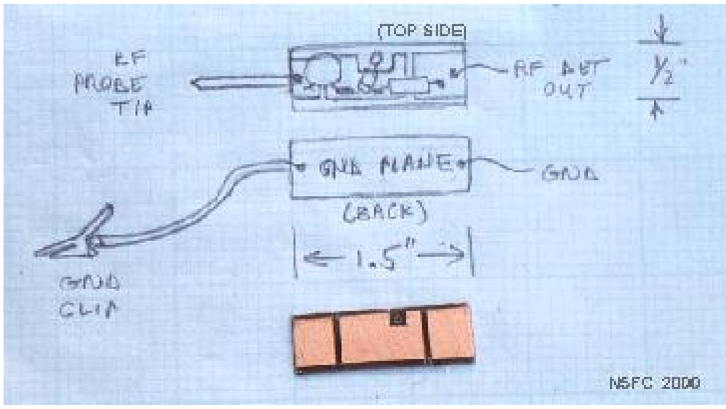

Click on the image above to see a larger

Take a small piece of scrap double-sided printed-circuit board, about 1-1/2 x 1/2 inches, Groove it on one side only, similiar to the image above, to create pads for soldering, but leave the back side as a "ground plane". Mount your diode, capacitor, and resistor as shown, soldering to the pads you made. One side of the diode (the non-banded isde) gets connected to the ground plane (drill a hole through to the other side and solder it). Try to fit all the components neatly inside the edges of the pc board. Solder the braid of the shielded wire (3-4 ft long) to the ground plane, and the center conductor to the pad with the resistor. Also, solder a 10-12 inch hook-up wire to the ground plane. Check that there are no shorts between the center conductor and the ground-plane. Solder the probe tip to the pad with the capacitor (I used a discarded probe tip from a broken test probe).

Here's where we get creative: packaging! One way or another, whatever method we use, it's important to shield the probe circuit, yet without shorting any part of the circuit to our shield (except the ground plane). I was on a kick of using copper pipe, which is very cheap, so I built my shield out of 1/2-inch copper pipe and end caps, commonly available at your local hardware store. I drilled a hole in the end of each end-cap, to pass the shielded cable and the probe tip. I used a shouldered washer to insulate the probe tip from the end cap, but a small rubber grommet would have worked as well. Stuff the assembly inside the copper pipe, and you end up with a completed probe that looks like the following:

Click on the image above to see a larger version

So, how do we use this thing?

Before we use it, a few precautions are in order. Don't use the probe in any circuit where the highest DC supply voltage is greater than the diode's reverse-breakdown voltage. For the 1N34A, this is 50 Volts. Same goes for the capacitor, which should be rated at least 50 Volts. This probably means that the probe cannot be used in most tube circuits. Also, don't try to measure RF power in circuits where the peak voltage will exceed 50 Volts. What will happen if you exceed these voltages by a little? Well, probably nothing; possibly, the diode or capacitor will fail open or short.

The first thing you'll always do in using the RF Probe is to connect the banana-plug end to the +/- jacks of your DC Voltmeter; set the Voltmeter to DC-Volts (not AC).

To use the RF Probe for signal tracing in a malfunctioning RF circuit or a homebrew circuit, connect the aligator clip to a convenient "ground" or "common" point in your circuit. Often this is the chassis. Most of the time, you'll be probing at the base/gate, emitter/source, or collector/drain of a transistor, one either side of a coupling capacitor or transformer, or at the input or output of an IC. Because the circuit's RF must overcome the diode's barrier potential (of 0.25V, for our 1N34A), voltages much less than that won't read at all, and voltages less than about a volt won't read very accurately. Typically, RF and post-mixer-amps in receivers don't have enough RF voltage, unless you inject a very strong signal at the input.

I recently used my RF probe to troubleshoot my dead TenTec Scout, which had suddenly quit transmitting in mid-QSO. I connected the rig to a dummy load, then keyed it while probing. Using the probe, I was able to follow a steadily increasing RF signal through the transmit chain, from the oscillator through the transmit mixer, to the pre-driver, and the driver. The actual voltage measurements weren't important, just that they were increasing from stage to stage where expected. Then, (whoops!) the driver's base circuit had 6 Volts, but the collector circuit only had only 0.1 Volts! The driver transistors had gone south!

You can also use the RF probe to measure RF power with reasonable accuracy, up to about 50 watts in a 50-ohm circuit. By 50-ohm circuit, I mean a 50-ohm antenna system at 1:1 SWR (higher SWRs are not 50 ohms), or a 50-ohm dummy load. Assuming the resistor in your RF probe is sized to match your DC Voltmeter's input impedance (as explained above), you will get quite reasonably accurate measurements using the following formula:

$$\mathrm{PWR} = \frac{(V_{(read)} + 0.25)^2}{R_{(load)}}$$

For example, I want to measure the power out of my TenTec 1340 40-Meter QRP transceiver. I place it on a 50-ohm dummy load, and key down. I generally use a BNC-Tee adapter to gain access to the output line, but I could as easily pop the cover off. Using the RF probe (alligator clip to chassis ground), I measure12.2 Volts (DC) (and the same RF RMS Volts). Plugging this into the formula above I have PWR= (12.2 + 0.25) * (12.2 + 0.25) / 50 = 3.1 Watts. The rated power for this rig is 3 Watts, so I've verified everything is hunky-dorey.

We've added the potential barrier to the measured voltage above, but that little trick doesn't work so well when you get down around a volt, and for voltages less than about a volt, the measurement accuracy suffers greatly. Also, the diode's response is severely non-linear below the barrier potential, and will generally read much less than expected in circuits where the RF voltage is less than 1/4 volt. So if you see tiny readings in circuits where it's normal to have voltages less than 1/4 volt RF, don't get too spun-up about the low readings... it may mean everything is normal. My rule of thumb for guessing at this is as follows: For collector/drain circuits in oscillators or transmit-chain amplifiers in key-down, expect RF Voltages about 20-50% of the applied DC (supply) voltage. This depends on the circuitry, of course, but it's a reasonable gesstimate. Base/gate and emitter/source circuits will generally be much less, maybe 5-10%. Circuit impedance will affect this too.

How good is this thing?

Well, we're not talking high performance test equipment here, but we *are* talking very useful. If you account for the barrier voltage, the readings can be quite accurate when measuring most low-impedance circuits (20-200 ohms), provided that the voltage is above 1 or 2 volts. How accurate? +/-10% from 200 KHz to 150 MHz would be a reasonable expectation. Also, the voltage divider is only accurate for sinusoidal signals. If you want "peak" measurements, simply multiply your reading by 1.414. The "peak" measurement should be good regardless of whether the waveform is sinusoidal. Regarding ultimate accuracy, your results may vary, and you may want to compare it to a laboratory instrument at the frequency of interest if you're really interested in accuracy. If you shield it well, and keep the ground clip lead reasonably short, it should be good in low-impedance circuits up into the VHF region, and down into the upper-audio region. In higher-impedance circuits, the junction capacitance of the diode may cause a low-pass effect at higher frequencies, and you're most likely to see this as a loss of measurement accuracy (i.e., low readings) at frequencies above 30 MHz. This doesn't mean it's not useful; it just means it reads low. Also, the capacitance of the probe may affect some sensitive RF circuits. For example, if you're probing a LC-tuned oscillator circuit, it may stop oscillating or change frequency or become unstable. Actually, most any probe will do this. Also, as we said before, the barrier voltage becomes a bigger part of the measurement error as the circuit voltage drops below a volt or so, and becomes dominant as you approach the barrier voltage. Just keep this in mind as one of it's limits.

Enjoy, good luck, and 73!
monty N5ESE

Return to N5ESE home page

Overseer: Monty Northrup ... ... leave e-mail ...

Visit our regular (non-ham, but very popular) homepage

N5ESE's Ballpoint RF Probe

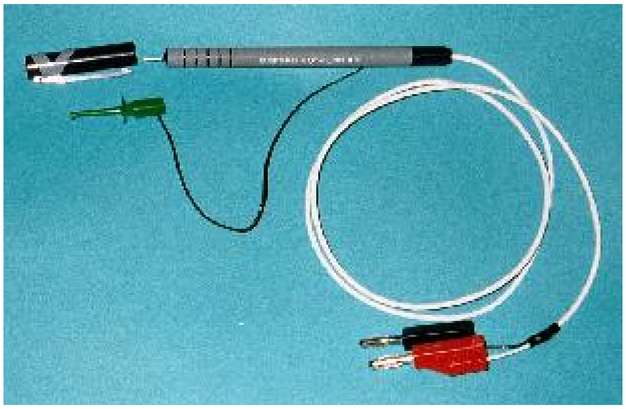

(click on any picture to see larger version)

NOTE: 'N5FC' is my former call.
This project was constructed while that call was valid, and you may observe references to it.

This is one of those afternoon projects that can really be both rewarding to build and useful to have. Electrically, it's identical to the Classic RF Probe described elsewhere (where you can also find the theory discussion for this one). Like the Classic RF Probe, this one is used in conjunction with a high-impedance-input Voltmeter or Digital Voltmeter (DVM). See the schematic below. Cost? About $5, if you can scrounge the ballpoint pen, heat shrink, shielded cable, and copper tape.

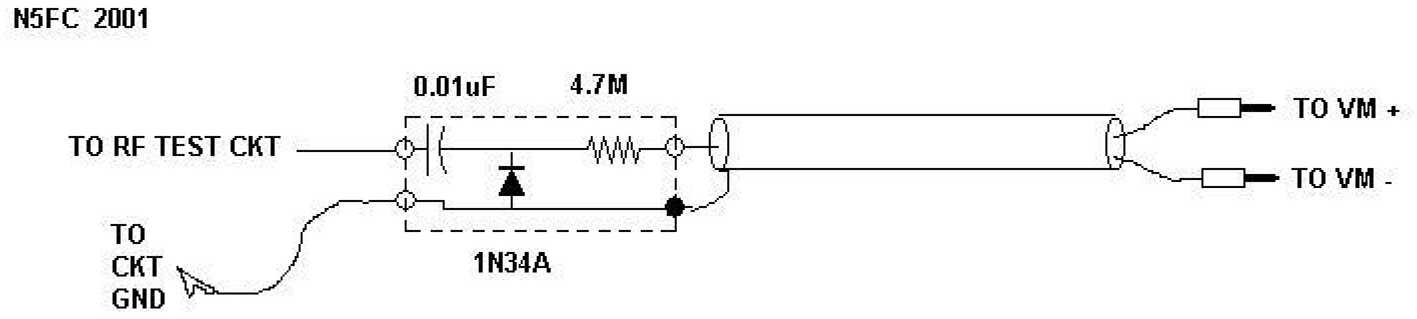

CLASSIC RF PROBE

**Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10 -11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)**

What makes this probe unique is that it's built inside the shell of a regular ol' ballpoint pen. Besides being conveniently compact, the unit sports a needle-probe suitable for use in probing surface-mount circuits, and good overall shielding. The pen cap protects the needle probe when not in use. When measuring sinusoidal signals, it should provide RMS-corrected readings, using a 10 or 11-Meg input impedance VTVM or DVM. With a 1-Meg DVM, it reads 25% of the sinusoidal RMS voltage. Reasonable accuracy (+/- 10%) can be expected over the HF/VHF range (2-150 MHz), although this hasn't been verified. When used to measure non-sinusoidal signals, the accuracy will be unknown, but it still affords good relative measurements, and most of the time, that's all that's required. It makes an excellent, compact, and portable accessory for troubleshooting or homebrewing QRP equipment with peak voltages less than 50 Volts (i.e., most solid-state equipment)

Construction

The figure below shows the parts required to build the Ballpoint RF Probe. Click on the image to open an larger, annotated image with parts labled, and construction notes. Pick a ballpoint pen with a non-metalized plastic body, and plenty of room inside. The Papermate Flexgrip model I used had an inside diameter a little over 1/4-inch. We'll use an itty-bitty scrap of double-sided printed-circuit-board to mount the electronic components. Trim the PC board to about 2-12" long and 3/16" wide; don't make it too wide, or it won't fit inside the ballpoint pen. Notch or file a little out of the middle of the pc board, so the 1N34A diode will fit easily inside the pen body. then, on one side only, groove in two places, so as to create 3 lands on the "top" side of the board. In addition to the parts shown, you'll need a 2-1/2" piece of heat-shrinkable tubing to cover the electronic assembly (although electrical tape would do instead), and about a foot of 1/4"-wide adhesive-backed copper tape, commonly available in rolls of 200-300 inches at large hobby stores (like Michaels, and Hobby Lobby). Although a chip capacitor is shown in the photo, a very small disc capacitor will do as well.

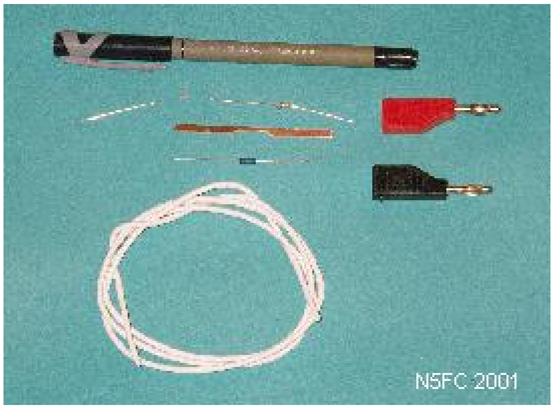

Click on the image above to see a larger, annotated image

In the next image (below), we get a close-up of the electronics assembly. You can see the input capacitor straddling the front-to middle lands, and the 4.7 Meg resistor straddling the middle-to-rear lands. The diode, which snuggles into the notch, connects from the middle land to the ground plane on the rear side of the pc board. The diode's banded end goes to the middle land. Break the sewing needle in half, using two needle nose pliers. **WARNING! Use eye and face protection!!** *ALSO NOTE: Don't try to cut a sewing needle with wire-cutters... you'll ruin the cutters.* Avoid straight pins, which dont have the hardness to perform well as probes. Then, solder the sewing needle to the front land, centering it carefully. You might benefit from burnishing the solder-half of the sewing needle with some fine grit sandpaper, to make it take solder a little better. Center it up nicely, as that will make for a professional-looking probe. Solder the shielded cable to the top/bottom of the pc board, center conductor to the rear land on top, and shield to the ground plane on the bottom. Be careful to aviod straggling shield-wires which could short the electronics. Also,solder a 10-12" pigtail of good, flexible insulated wire onto the ground plane, pigtailed rearward. This will be used as the ground wire in our test circuit. Before you shrink the tubing over the electronic assembly, check for shorts between lands and from lands to ground plane, make sure you have the diode polarity correct, and check that the needle is making solid electrical contact, and is mechanically secure.

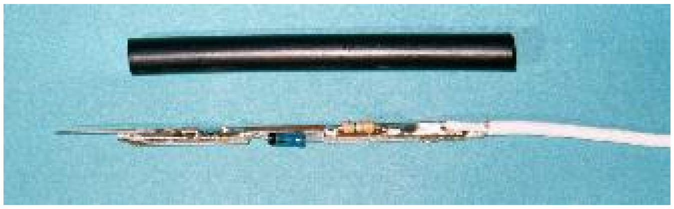

Click on the image above to see a larger, annotated image

See the next image, below. After the electronic assembly has been heat-shrunk overall, wrap the copper tape all around the electronic assembly. This will be our shield. Near the rear of the electronic assembly, solder the electonic assembly's ground to the copper tape, near or on the cable shield. Alternative shielding methods can be tried, for example, you might pull the shield out of a piece of RG-59, and sleeve it over the electronic assembly, soldering it to the ground plane. Whatever you do, be certain that the shield cannot unravel and short against the probe itself or any of the electronics.

Although not shown in the picture, we drill a small hole about 2/3 back on the pen casing, threading the ground pigtail through the hole. This really tests your hand-to-eye coordination. If the pen has a threaded-in rear-cap, drill a hole in it just big enough to accommodate your main shielded cable. Thread the whole cable into the pen casing, and out the other side, and if you had a threaded rear-cap, like I did, thread it on. Pull the electronic assembly gently back into the casing, so that the needle probe sticks out about 1/2 inch. Mix some clear 5-minute epoxy, and let it thicken ever so slightly. Then , while holding the assembly vertically (i.e., probe-tip up), and using a small toothpick or screwdriver, drop epoxy into the probe area, sealing the electronics and probe into place. Allow to dry thoroughly before applying any pressure to the assembly.

When dry, attach your favorite ground-clip to the pigtail, and banana plugs on the end of the shielded cable (red to the center-conductor, black to the shield, to match your Voltmeter)

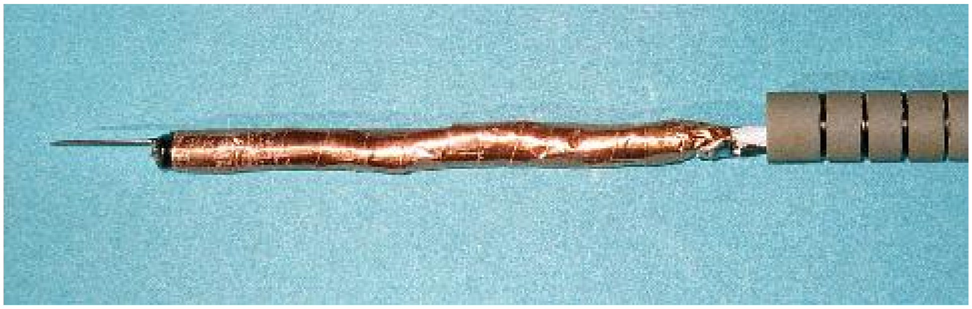

Click on the image above to see a larger, annotated image

See the next image, below, which shows the completed assembly, annotated. Sometimes, seeing the entire assembly makes everything perfectly clear. Place the pen cap over the needle probe to protect the assembly when not in use.

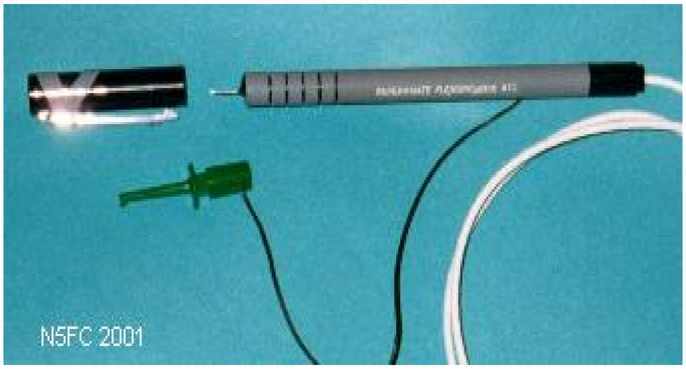

Click on the image above to see a larger, annotated image

And speaking of use, here's our lovely model (OK...XYL) making a measurement in the N0SS Noise Generator. She has the ground clip connected to a convenient place in the circuit's ground, and the probe touches the test point we want to measure. As you can see, we read 0.710 Volts. Since this is broadband noise, the actual voltage reading is not accurate, but it was seen to be much greater than the previous stage, as we expected.

Click on the image above to see a larger, more readable image

Let me know if you build one of these... and send me a picture!

73, monty N5ESE

Return to N5ESE home page

Overseer: Monty Northrup ... e-mail ... leave e-mail ...

maddog homepage Visit our regular (non-ham, but very popular) homepage

A Coax Switch

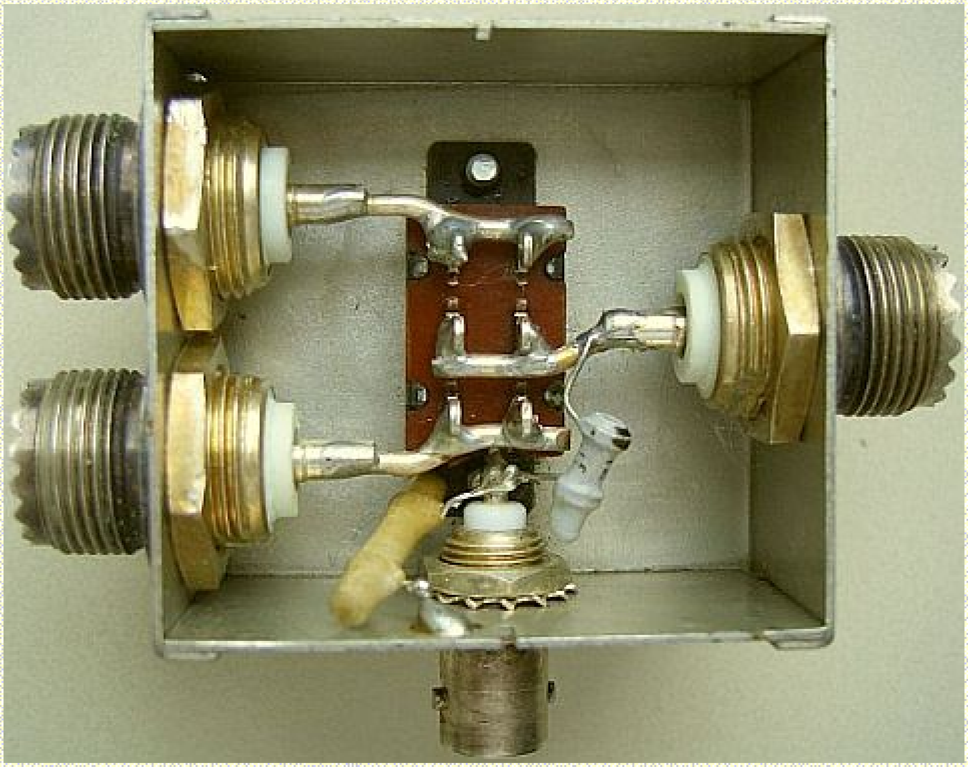

Cheap slide switch in a standard 54 × 50 × 26 mm box.

THE SWITCH

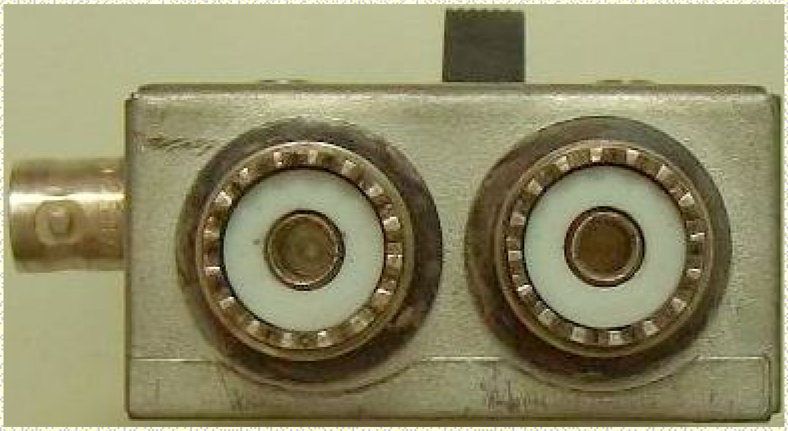

Fig » shows a proven inexpensive home-made antenna selection switch. If you question the use of a cheap slide switch and SO239 coax sockets, read on. Measurements in a physics lab showed there to be practically no reflection on HF and even on 70 cm the SWR was below 1.3 : 1! That is explained as follows:

- The contacts in the slide switch have larger contact surfaces than many a bought coaxial switch.

- The wiring and switch contacts, between the top and bottom of the metal case, act as the centre conductor of a coax of near 50 ?.

It is a standard box measuring 54 × 50 × 26 mm (l × w × h) and the wiring between the switch and the coax sockets is done in 2 mm silver plated wire.

POWER HANDLING CAPABILITY

If the switching is done with power off, the switch can stand 800 W. I have used this switch for more than 15 years and even with 1500 W there have been no problems.

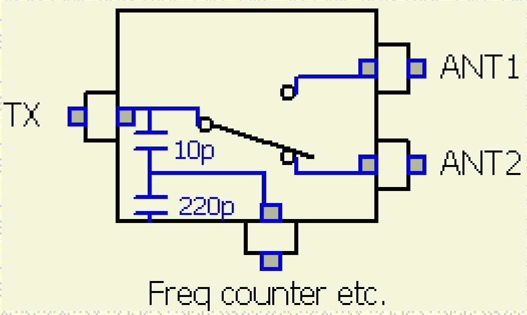

While building this switch, you might as well add a test point, e.g. to connect a ‘scope’ or frequency counter. The capacitors of 10 and 220 pF make a capacitive voltage divider and the extra loading, 9.6 pF, does not affect performance on HF. In fact, on 10 m it improves the SWR as the extra capacity, in combination with the wiring it makes a filter which favours that frequency.

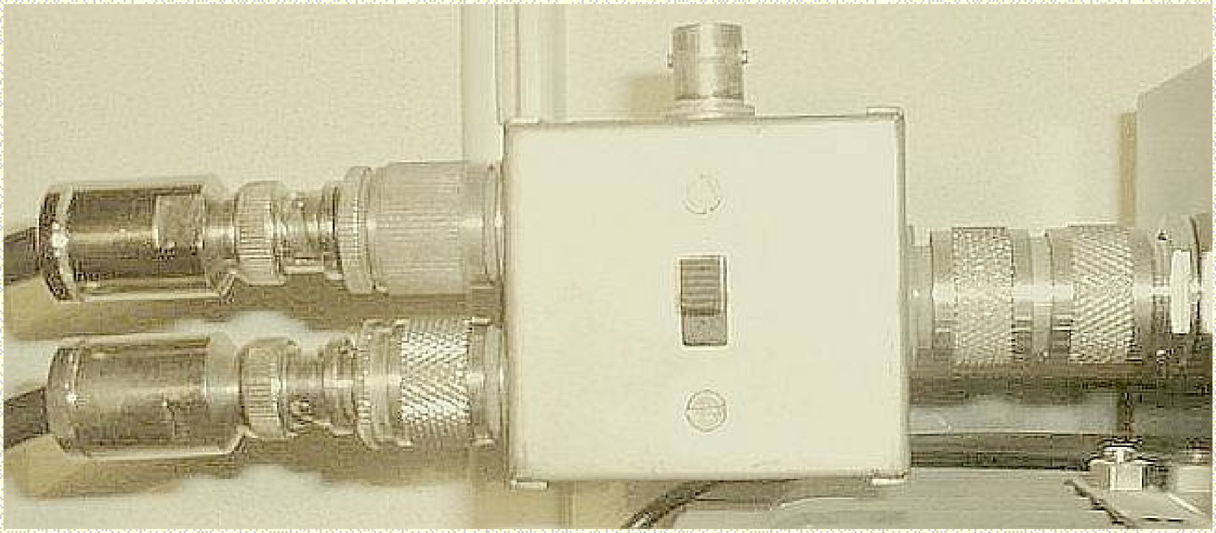

This shows how the switch is used in my station.

«Back

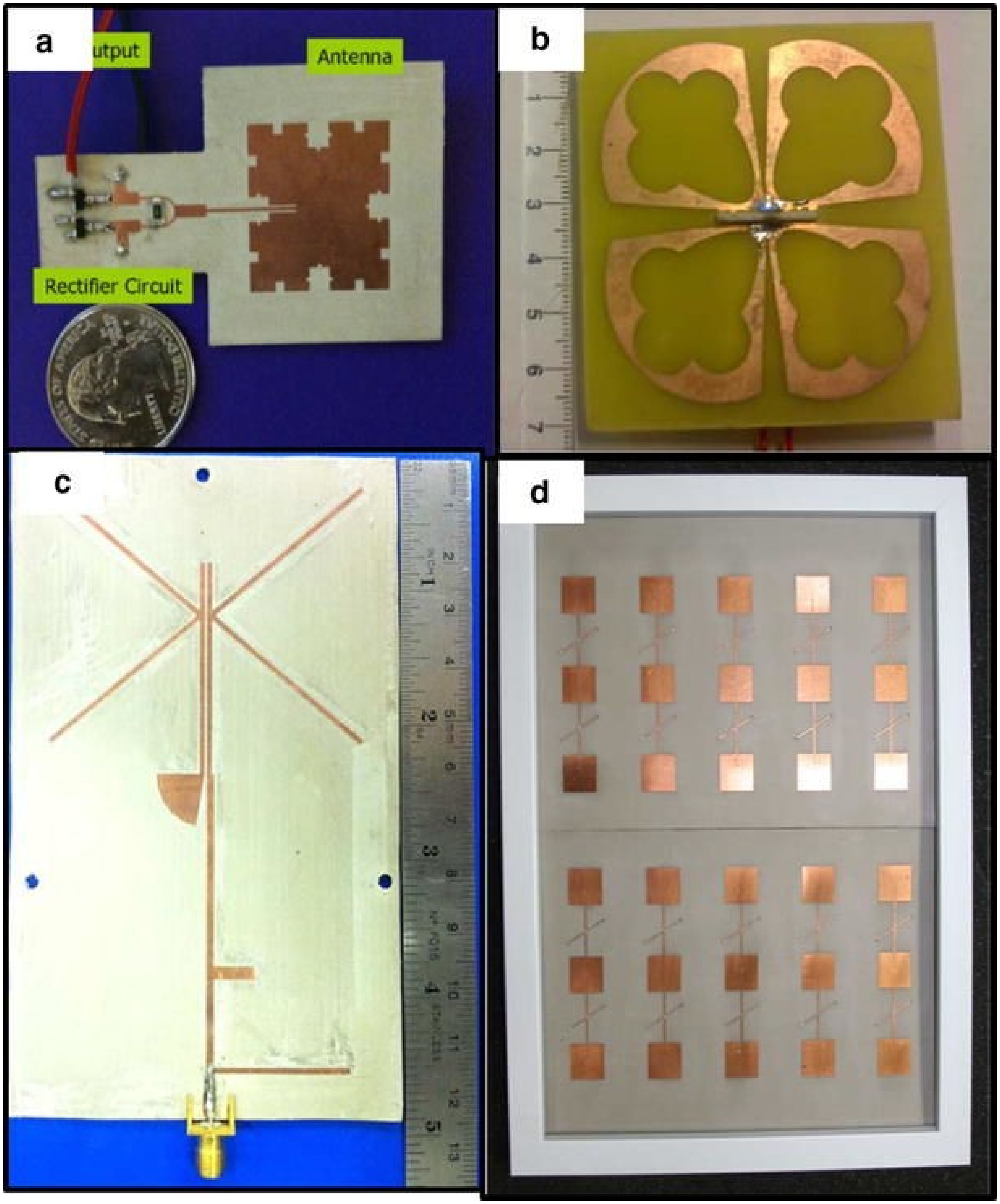
a
utput
Antenna
Rectifier Circuit
b
1
2
3
4
5
6
7
c
INCH
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
d

Int. J. Communications, Network and System Sciences, **2010, 3, 204-207**
doi:10.4236/ijcns.2010.32029 Published Online February 2010 (http://www.SciRP.org/journal/ijcns/).

Geometry Aspects and Experimental Results of a Printed Dipole Antenna

Constantinos VOTIS[1], **Vasilis CHRISTOFILAKIS**[1,2], **Panos KOSTARAKIS**[1]
[1]*Physics Department, University of Ioannina, Panepistimioupolis, Ioannina, Greece*
[2]*Siemens Enterprise Communications, Enterprise Products Development, Athens, Greece*
E-mail: *kvotis@grads.uoi.gr, basilios.christofilakis@siemens-enterprise.com, kostarakis@uoi.gr*
Received November 2, 2009; *revised December* 10, 2009; *accepted January* 12, 2010

Abstract

Detail experimental measurements of a 2.4 GHz printed dipole antenna for wireless communication systems is presented and discussed. A group of printed dipoles with integrated balun have been designed and constructed on a dielectric substrate. This paper is based on modifications of the known printed dipole architecture. The corresponding printed dipole antennas have differences on their forms that are provided by two essential geometry parameters. The first parameter l is related to the bend on microstrip line that feeds the dipole and the second w corresponds to the form of the dipole's gap. The impact of these parameters on reflection coefficient and radiation pattern of antenna has been investigated. The corresponding measured results indicate that the return loss and radiation pattern of a printed dipole antenna are independent of the w parameter. Instead, variations in the value of the l parameter in the dipole's structure affect the form of the corresponding return loss. These observations are very important and provide interesting considerations on affecting design and construction of antenna elements at frequency range of 2.4 GHz.

Keywords: Printed Dipole, Scattering Parameters, Radiation Pattern

1. Introduction

Modern wireless communications offer higher bit rates and efficient quality of services. The majority of the equipment used today introduces requirements for better performance and lower cost. Antennas with quite small sizes, low profiles and versatile features represent interesting solutions that provide modern wireless applications. The printed dipole antenna with integrated balun is widely used as a radiation element on communication systems because of its omni-directional features, narrowband character and simple structure [1–4]. This type of antenna because of its small size can be integrated on the same PCB with other electronics circuits and devices. For the same reason, it can also be used as element on antenna array architecture. The last feature is very interesting and attractive in MIMO modern wireless systems. This printed dipole architecture offers versatile characteristics for design and implementation of antenna arrays on both ends of a MIMO wireless system.

In the present paper, we will study and discuss the effect of the variation of the two geometrical parameters (l, w) of the printed dipole antenna structure. The first corresponds to a discontinuity on microstrip line of printed dipole and the second is related to the discontinuity in the gap. Details of structure concept and design process are presented in Section 2; the experimental results for return loss and radiation pattern for each of the printed dipoles are presented and discussed in Section 3. The paper concludes in Section 4.

2. Design and Structure Aspects

As mentioned above, the proposed analysis is based on geometrical characteristics of a prototype printed dipole antenna with integrated balun. This kind of printed dipole antenna is considered for use in many applications [1–3]. In our study the geometrical parameters of the printed dipole antenna were modified to achieve better performance in the frequency range of 2.4 GHz. This modified design and the corresponding parameters are shown in Figure 1 while the values summarized in Table 1.

Identify applicable sponsor/s here. *(sponsors)*

Copyright © 2010 SciRes.

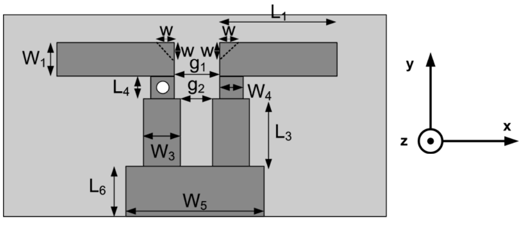

(a) Bottom Layer

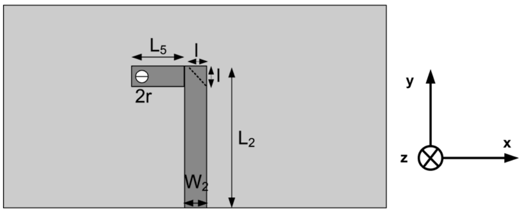

(b) Top Layer

Figure 1. Geometry of printed dipole.

Table 1. Printed dipole dimensions.

Parameter	Values
Dipole strips	L1 = 20.8 mm W1 = 6 mm g1 = 3 mm
Microstrip Balun	L2 = 32 mm L3 = 16 mm L4 = 3 mm L5 = 3 mm W2 = 2 mm W3 = 5 mm W4 = 3 mm g2 = 1 mm
Via radius	r = 0.375 mm
Ground plane	L6 = 12 mm W5 = 17 mm
Side of microstrip bend	l variable (0 mm – 3 mm)
Side of dipole's arms in the gap	w variable (0 mm – 3 mm)

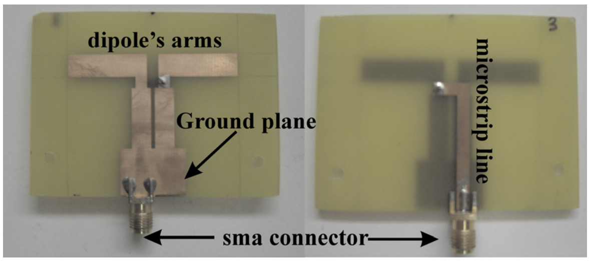

Figure 2. Prototype printed dipole antenna: Top Layer (left) –Bottom Layer (right).

An Fr-4 substrate with thickness of 1.5 mm and permittivity ε_r =4.4 has been used for the fabrication of the dipoles. Figure 2 shows the top and bottom layer for the one of them. It also presents the dipole's arms and gap, the balun, the ground plane and the microstrip line that interface the dipole with the coaxial feed line via sma connector.

From this Figure, we can also see the right angle at the microstrip line and the other two right angles at the dipole's gap. It is known that the presence of right angles in conductors cause discontinuities that leads to degradation in circuit performance [5]. Microwave theory suggests that these angles introduce parasitic reactances which can lead to phase and amplitude errors, input and output mismatch and possibly spurious coupling [5–7]. In order to reduce this effect it is proposed to modify these discontinuities directly, by mitering the conductor. Our investigation and the experimental measurements show the effect of mitering these discontinuities. At first, a prototype printed dipole antenna with unaffected geometrical parameters has been designed and constructed. Secondly, we constructed and measured six different printed dipoles. Three of them had *w = 0 mm* and different *l* values (1 mm, 2 mm, 3 mm) and the other three dipoles had *l = 0 mm* and different values of *w* (1 mm, 2 mm, 3 mm). All these seven dipoles we constructed, the unaffected one and the mitered ones were measured in an anechoic environment. Figures 3 and 4 show a printed dipole for *l = 2 mm* and *w = 0* mm and for *l = 0 mm* and *w = 3 mm*, respectively. The aim of this study is to investigate the return loss coefficient and radiation pattern in each of these seven dipole's forms. The next section discusses the obtained results and presents the significant observations.

3. Results and Discussion

The return loss of the prototype dipole and the six different modified printed dipole antenna we constructed are measured using a Network Analyzer. These results are shown in two Figures. The first (Figure 5) corresponds to *l* parameter's variations keeping the *w* parameter equal to zero. The second (Figure 6) shows the return loss curves where *w* parameter varies but the *l* parameter equals to zero. In both figures we can see the return loss curve that belongs to the prototype printed dipole (*l* and *w equal to 0 mm*).

From these curves, it seems that this dipole antenna design has a resonance point at 2.4 GHz with 500 MHz –

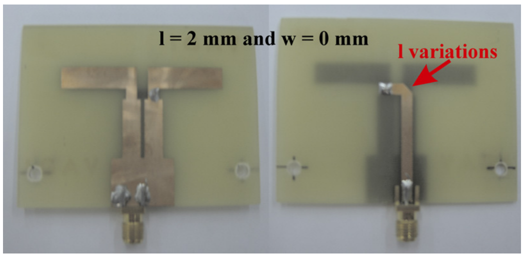

Figure 3. Printed dipole antenna for *l = 2 mm* and *w = 0 mm* Top Layer (left) – Bottom Layer (right).

Copyright © 2010 SciRes.

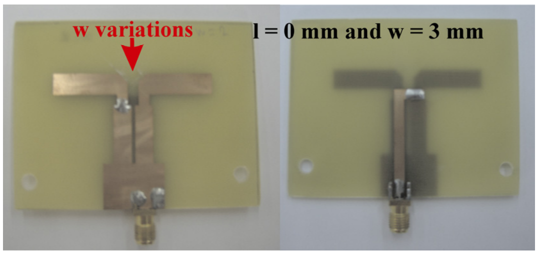

Figure 4. Printed dipole antenna for *l = 0 mm* and *w = 3 mm* Top Layer (left) – Bottom Layer (right).

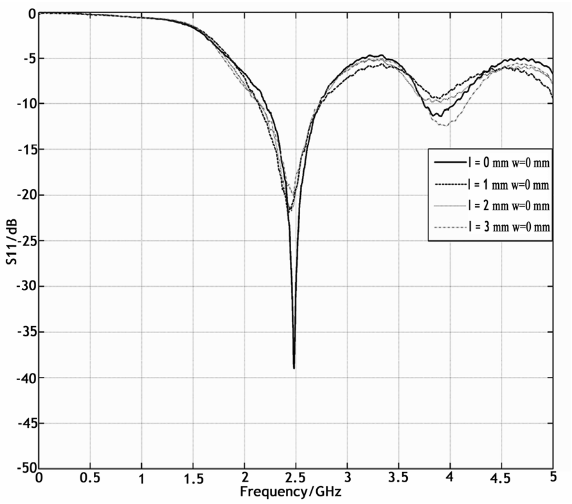

Figure 5. Measured return loss of printed dipole for each value of *l* parameter and *w = 0* mm.

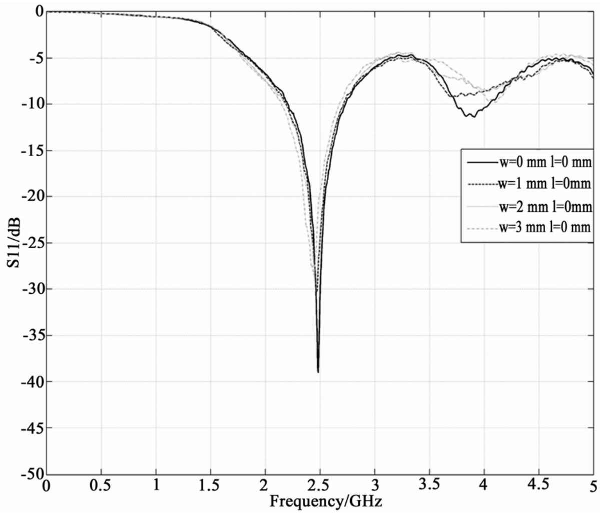

Figure 6. Measured return loss of printed dipole for each value of *w* parameter and *l = 0 mm.*

10 dB bandwidth. The last frequency range has center frequency close to 2.4 GHz which is the frequency value that the return loss is minimized. For these frequencies the corresponding values of return loss are smaller or equal to -10 dB. From Figure 5, it is obvious that as the value of *l* parameter increases, the form of the corresponding return loss curve changes and becomes more flat at the resonance frequency range. On the other hand, the value of *w* parameter does not affect the form of the return loss curve. Each of these seven forms of printed dipole antenna has quite similar return loss curves and introduces narrowband operation at the frequency range of 2.4 GHz. Moreover, for a wireless application that requires design and construction of many identical printed dipoles, it is recommended to choose *l* parameter equals to *2 mm* and *w* parameter equals to *0 mm* for better performance. As it can be seen from Figure 5 the above investigation ensures that the printed dipole antennas will have quite identical return loss curves and performance as elements in an antenna array configuration.

For deeper analysis on this topic, experimental measurements on radiation pattern of these antennas have also been made. Measurements were carried out in a RF anechoic chamber using a calibrated measuring system. In particular, Figure 7 shows the measurements of radiation pattern in E- plane and in H – plane for each dipole

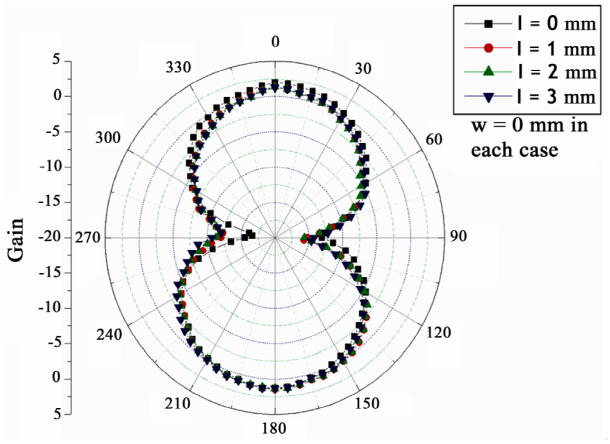

(a)

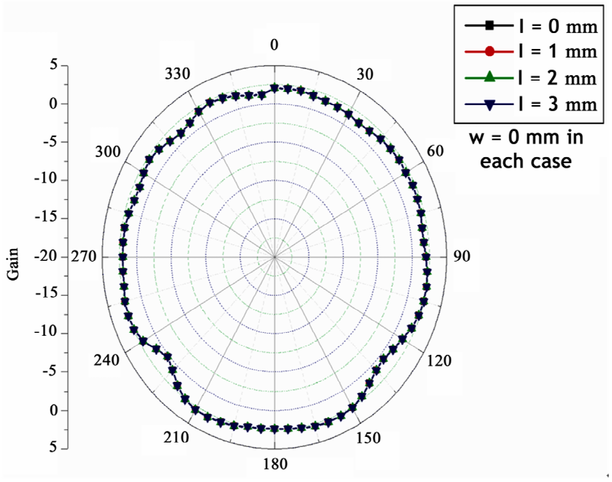

(b)

Figure 7. Radiation pattern of dipole for each value of *l* parameter (a) E – plane, (b) H -plane.

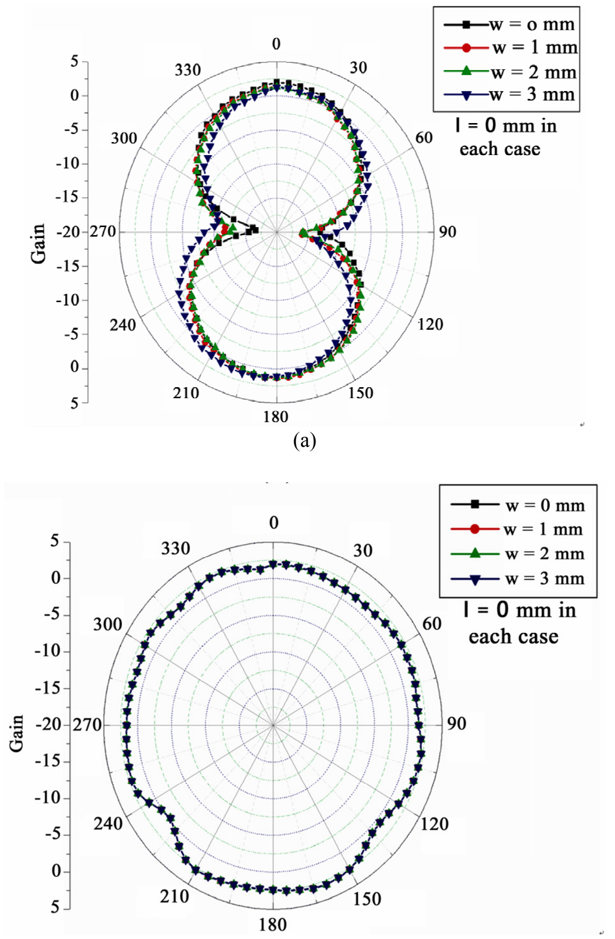

(a)

(b)

Figure 8. Radiation pattern of dipole for each value of *w* parameter (a) E – plane, (b) H – plane.

with *w* parameter equals to 0 mm and *l* parameter equals to integer values that ranging from *0 mm* to *3 mm*,. Figure 8 shows the corresponding results for each dipole with *l* parameter equals to *0 mm* and *w* parameter's integer values ranging from *0 mm* to *3 mm*. All these dipole structures introduce radiation characteristics that correspond to a fundamental dipole antenna [6,7]. Each of them has a measured peak gain that equals to quite 2 dBi and introduces omni-directional features. Quite small variations on these curves are on the limits of measurements' accuracy. For this reason, it can be observed that the radiation characteristics of the printed dipole antenna are not affected by the variations on *l* and *w* geometrical parameters. Therefore, the radiation diagrams of them are independent of the *l* and *w* parameters.

4. Conclusions

A number of printed dipole antennas with integrated balun are constructed and studied in terms of return loss and radiation pattern. Each of them has a defined form and geometry. Starting from a dipole antenna we mitered the angles introducing the parameters *l* and *w* that we varied. Experimental measurements on return loss provide the obtained results. These are quite similar and also introduce a resonance point at frequency range of 2.4 GHz with narrow resonance bandwidth. The form of this resonance range is affected only by the *l* parameter. The radiation pattern of these dipoles is also investigated. The corresponding radiation diagrams are independent of these geometrical parameters (*l*, *w*) and are similar to that of the fundamental dipole. These observations on printed dipole architecture are very crucial for wireless communication engineering and antenna design. This is because they introduce the ability of constructing a group of identical dipoles choosing an appropriate value of *l* parameter (*l =2 mm*) with quite identical resonance and radiation characteristics.

5. Acknowledgment

This research project (PENED) is co-financed by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%).

6. References

[1] D. Edward and D. Rees, "A broadband printed dipole with integrated balun," Microwave J, 1987, pp. 339–344.

[2] N. Michishita, H. Arai, M. Nakano, T. Satoh, and T. Matsuoka, "FDTD analysis for printed dipole antenna with balun," Asia Pacific Microwave Conference, 2000, pp. 739–742.

[3] G. S. Hilton, C. J. Railton, G. J. Ball, A. L. Hume, and M. Dean, "Finite–difference time–domain analysis of a printed dipole antenna," 19th Int. IEEE Antennas and Propagation Conference, 1995, pp. 72–75.

[4] H. R. Chuang and L. C. Kuo, "3-D FDTD design analysis of a 2.4 GHz polarization – diversity printed dipole antenna with integrated balun and polarization– switching circuit for wlan and wireless communication application," IEEE Transactions on Microwave Theory and Techniques, Vol. 51, No. 2, 2003.

[5] D. M. Pozar, Microwave Engineering, Wiley, 1998.

[6] C. A. Balanis, Antenna Theory Analysis and Design, Wiley, 1997.

[7] R. Garg, P. Bhartia, I. Bahl, and A. Ittipiboon, Microstrip Antenna Design Handbook, Artec House, 2001.

Circuits and Systems, **2012, 3, 216-222**
http://dx.doi.org/10.4236/cs.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/cs)

Optimization of the Voltage Doubler Stages in an RF-DC Convertor Module for Energy Harvesting

Kavuri Kasi Annapurna Devi[1*], Norashidah Md. Din[2], Chandan Kumar Chakrabarty[2]
[1]Department of Electrical and Electronic Engineering, INTI International University, Nilai, Malaysia
[2]Department of Electronics and Communication Engineering, Universiti Tenega Nasional, Kajang, Malaysia
Email: {*kavurik.adevi., chandan}@newinti.edu.my, norashidah@uniten.edu.my

Received August 25, 2011; revised March 26, 2012; accepted April 4, 2012

ABSTRACT

This paper presents an optimization of the voltage doubler stages in an energy conversion module for Radio Frequency (RF) energy harvesting system at 900 MHz band. The function of the energy conversion module is to convert the (RF) signals into direct-current (DC) voltage at the given frequency band to power the low power devices/circuits. The design is based on the Villard voltage doubler circuit. A 7 stage Schottky diode voltage doubler circuit is designed, modeled, simulated, fabricated and tested in this work. Multisim was used for the modeling and simulation work. Simulation and measurement were carried out for various input power levels at the specified frequency band. For an equivalent incident signal of –40 dBm, the circuit can produce 3 mV across a 100 kΩ load. The results also show that there is a multiplication factor of 22 at 0 dBm and produces DC output voltage of 5.0 V in measurement. This voltage can be used to power low power sensors in sensor networks ultimately in place of batteries.

Keywords: Energy Conversion; RF; Schottky Diode; Villard; Energy Harvesting

1. Introduction

RF energy harvesting is one type of energy harvesting that can be potentially harvested such as solar, vibration and wind. The RF energy harvesting uses the idea of capturing transmitted RF energy at ambient and either using it directly to power a low power circuit or storing it for later use. The concept needs an efficient antenna along with a circuit capable of converting RF signals to DC voltage. The efficiency of an antenna mainly depends on its impedance and the impedance of the energy converting circuit. If the two impedances aren't matched then it will be unable to receive all the available power from the free space at the desired frequency band. Matching of the impedances means that the impedance of the antenna is the complex conjugate of the impedance of the circuit (voltage doubler circuit).

The concept of energy harvesting system is shown in **Figure 1**, which consists of matching network, RF-DC conversion and load circuits. The authors in [1], used a 2.4 GHz operating frequency with an integrated zero bias detector circuit using BiCMOS technology which produced an output voltage of 1 V into a 1 MΩ load at an input power level of 0 dBm. H. Yan and co-authors revealed that a DC voltage of 0.8 volts can be achieved from a –20 dBm RF input signal at 868.3 MHz through simulation results [2]. In [3], work was carried out on a firm frequency of 900 MHz by matching to a 50 Ω impedance and resonance circuit transformation in front of the Schottky diode which yields an output voltage of over 300 mV at an input power level of 2.5 µ. W. J. Wang, L. Dong and Y. Fu [4] used a Cockcroft-Walton multiplier circuit that produced a voltage level of 1.0 V into a 200 MΩ load for an input power level of less than –30 dBm at a fixed frequency of 2.4 GHz.

The energy conversion module designed in this paper is based on a voltage doubler circuit which can be able to output a DC voltage typically larger than a simple diode rectifier circuit as in [5], in which switched capacitor charge pump circuits are used to design two phase voltage doubler and a multiphase voltage doubler. The module presented in this can function as an AC to DC converter that not only rectifies the input AC signal but also elevates the DC voltage level. The output voltage of the

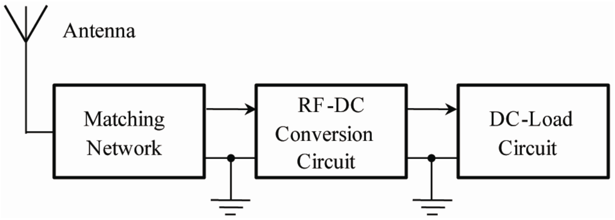

Figure 1. Schematic view of a RF energy harvesting system.

---

[*]Corresponding author.

Copyright © 2012 SciRes.

energy conversion module can be used to energize the low power devices for example sensors for a sensor network in application to agriculture.

Section 2 of this paper discusses on the theoretical background of the voltage doubler circuit. Section 3 presents the simulation study and implementation of the circuit design. Section 4 provides the results and analysis. Section 5 concludes with a discussion on the findings from the simulated and measured results.

2. Voltage Multiplier

There are various voltage multiplier circuit topologies. The design used in this module is derived from the function of peak detector or a half wave peak rectifier. The Villard voltage multiplier circuit was chosen in the circuit design of this paper because it produces two times of the input signal voltage towards ground at a single output and can be cascaded to form a voltage multiplier with an arbitrary output voltage and its design simplicity.

2.1. Diode Modeling

The voltage multiplier circuit in this design uses zero bias Schottky diode HSMS-2850 from Agilent. The attractive feature of these Schottky diodes are low substrate losses and very fast switching but leads to a fabrication overhead. This diode has been modeled for the energy harvesting circuit which comes in a one-diode configuration. The modeling parameters for these diodes are given by Agilent in their data sheets. These parameters are used in Multisim for its own modeling purposes. The modeling is done by transforming the diode into an equivalent circuit using passive components which are described by the SPICE parameters in **Table 1** [6].

The diode used in this design is shown in **Figure 2** and its equivalent model is shown in **Figure 3**. The special features of HSMS-2850 diode is that it provides a low forward voltage, low substrate leakage and uses the non symmetric properties of a diode that allows unidirectional flow of current under ideal condition.

Table 1. SPICE parameters.

Parameters	Units	HSMS 2850
B_V	V	3.8
C_{J0}	pF	0.18
E_G	Ev	0.69
I_{BV}	A	3E–4
I_S	A	3E–6
N	No unite	1.06
R_S	Ω	25
P_B (V_J)	V	0.35
P_T (XTI)	No units	2
M	No units	0.5

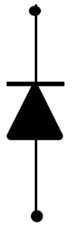

Figure 2. Schottky diode.

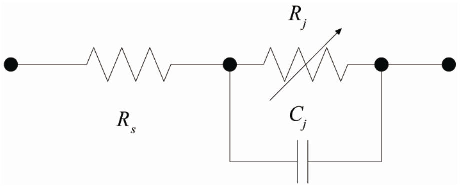

Figure 3. Linear circuit model of the Schottky diode [6].

The diodes are fixed and are not subject of optimization or tuning. This is described using the following derivations. By neglecting the effect of diode substrate, an equivalent linear model that can be used for the diode as shown in **Figure 3**. When C_j is the junction capacitance and R_j is the junction resistance, the admittance Y_z of the linear model is given by

$$Y_Z = Y_{C_j} + Y_{R_j} \qquad (1)$$

Equation (1) related to the frequency of operation is given by

$$Y_Z = jwC_j + \frac{1}{R_j} \qquad (2)$$

$$= \frac{jwC_jR_j + 1}{R_j} \qquad (3)$$

The impedance Z of the linear model is given by

$$Z = \frac{R_j}{1 + jwR_jC_j} \qquad (4)$$

The total impedance Z_T is given by

$$Z_T = R_S + \frac{R_j}{1 + jwR_jC_j} \qquad (5)$$

where R_S is the series resistance of the circuit and R_j is given by

$$R_j = \frac{8.33 \times 10^{-5} \times N \times T}{I_b + I_S}$$

where:

I_b = bias current in μA;

I_s = saturation current in μA;

T = temperature (K);

Copyright © 2012 SciRes.

N = ideality factor.

In Equation (5), R_j and C_j are constants and the frequency of operation (w) is the only variable parameter. As the frequency increases, the value of Z is almost negligible compared to the series resistance R_S of the diode. From this it is concluded that the function of the diode is independent of the frequency of operation.

2.2. Single Stage Voltage Multiplier

Figure 4 represents a single stage voltage multiplier circuit. The circuit is also called as a voltage doubler because in theory, the voltage that is arrived on the output is approximately twice that at the input. The circuit consists of two sections; each comprises a diode and a capacitor for rectification. The RF input signal is rectified in the positive half of the input cycle, followed by the negative half of the input cycle. But, the voltage stored on the input capacitor during one half cycle is transferred to the output capacitor during the next half cycle of the input signal. Thus, the voltage on output capacitor is roughly two times the peak voltage of the RF source minus the turn-on voltage of the diode.

The most interesting feature of this circuit is that when these stages are connected in series. This method behaves akin to the principle of stacking batteries in series to get more voltage at the output. The output of the first stage is not exactly pure DC voltage and it is basically an AC signal with a DC offset voltage. This is equivalent to a DC signal superimposed by ripple content. Due to this distinctive feature, succeeding stages in the circuit can get more voltage than the preceding stages. If a second stage is added on top of the first multiplier circuit, the only waveform that the second stage receives is the noise of the first stage. This noise is then doubled and added to the DC voltage of the first stage. Therefore, the more stages that are added, theoretically, more voltage will come from the system regardless of the input. Each independent stage with its dedicated voltage doubler circuit can be seen as a single battery with open circuit output voltage V_0, internal resistance R_0 with load resistance R_L, the output voltage, V_{out} is expressed as in Equation (7).

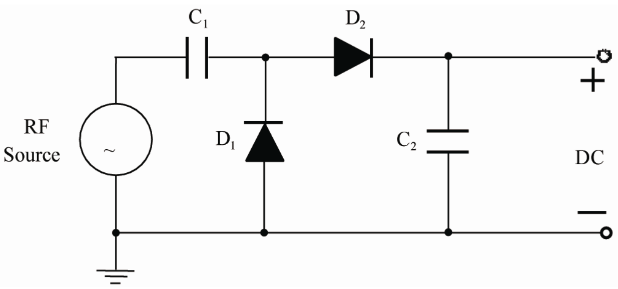

Figure 4. Single stage voltage multiplier circuit [7].

$$V_{\text{out}} = \frac{V_0}{R_0 + R_L} R_L \tag{6}$$

When n number of these circuits are put in series and connected to a load of R_L in Equation (6) the output voltage V_{out} obtained is given by this change in RC value will make the time constant longer which in turn retains the multiplication effect of two in this design of seven stage voltage doubler.

$$V_{\text{out}} = \frac{nV_0}{nR_0 + R_L} = V_0 \frac{1}{\frac{R_0}{R_L} + \frac{1}{n}} \tag{7}$$

The number of stages in the system has the greatest effect on the DC output voltage, as shown from Equations (6) and (7).

It is inferred that the output voltage V_{out} is determined by the addition of R_0/R_L and $1/n$, if V_0 is fixed. From this analysis it is observed that V_0, R_0 and R_L are all constants. Assume that $V_0 = 1$ V, $R_0/R_L = 0.25$, $n = 2, 3, 4, 5, 6$ and 7, the output voltage V_{out} = 1.33 V, 1.72 V, 2.0 V, 2.22 V, 2.43 V and 2.56 V respectively when substituted analytically in the Equation (7). This analysis can be compared with the results obtained in the circuit design of this module. In simulation at $n = 4$, V_{out} = 1.42 V, $n = 5$, V_{out} = 1.67 V; $n = 6$, V_{out} = 1.92; $n = 7$, V_{out} = 2.15 V; $n = 8$, V_{out} = 1.92 V; $n = 9$, V_{out} = 1.81 V. Also in measurement, for $n = 4$, V_{out} = 2.1 V; $n = 5$, V_{out} =2.9 V; $n = 6$, V_{out} = 3.72 V; and $n = 7$, V_{out} = 5 V. As n increases, the increase in output voltage will be almost double the input voltage up to some number of stages. But at some point, *i.e.* beyond seven stages, in this circuit the output voltage gained (8 and 9 stages) will be negligible as shown in **Figure 5**.

The capacitors are charged to the peak value of the input RF signal and discharge to the series resistance (R_s) of the diode. Thus the output voltage across the capacitor of the first stage is approximately twice that of the input signal. As the signal swings from one stage to other, there is an additive resistance in the discharge path of the

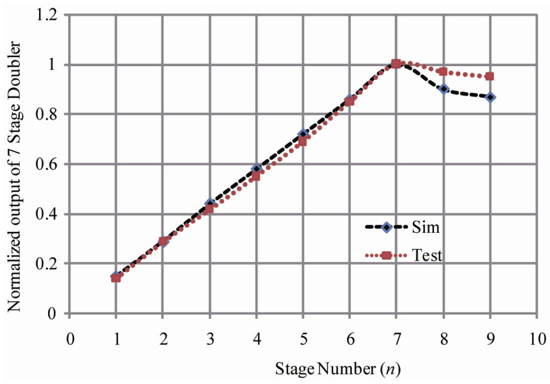

Figure 5. Normalized output voltage multiplier versus number of stages.

diode and increase of capacitance due to the stage capacitors.

2.3. Seven Stage Voltage Multiplier

The seven stage voltage multiplier circuit design implemented in this paper is shown in **Figure 6**. Starting on the left side, there is a RF signal source for the circuit followed by the first stage of the voltage multiplier circuit. Each stage is stacked onto the previous stage as shown in the **Figure 6**. Stacking was done from left to right for simplicity instead of conventional stacking from bottom to top.

The circuit uses eight zero bias Schottky surface-mount Agilent HSMS-285X series, HSMS-2850 diodes. The special features of these diode is that, it provides a low forward voltage, low substrate leakage and uses the non-symmetric properties of a diode that allows unidirectional flow of current under ideal conditions. The diodes are fixed and are not subject of optimization or tuning. This type of multiplier produces a DC voltage which depends on the incident RF voltage. Input to the circuit is a predefined RF source. The voltage conversion can be effective only if the input voltage is higher than the Schottky forward voltage.

The other components associated with the circuit are the stage capacitors. The chosen capacitors for this circuit are of through-hole type, which make it easier to modify for optimization, where in [8] the optimization was accomplished at the input impedance of the CMOS chip for a three stage voltage multiplier. The circuit design in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the load. Without the load resistor on the circuit, the voltage would be hold indefinitely on the capacitor and look like a DC signal, assuming ideal components. In the design, the individual components of the stages need not to be rated to withstand the entire output voltage. Each component only needs to be concerned with the relative voltage differences directly across its own terminals and of the components immediately adjacent to it. In this type of circuitry, the circuit does not change the output voltage but increases the possible output current by a factor of two. The number of stages in the system is directly proportional to the amount of voltage obtained and has the greatest effect on the output voltage as explained in the Equation (7) and shown in **Figure 5**.

3. Simulation and Implementation

Multisim software was chosen for modeling and simulation which is a circuit simulation tool by Texas Instruments. The simulation and practical implementation were carried out with fixed RF at 945 MHz ± 100 MHz, which are close to the down link center radio frequency (947.5 MHz) of the GSM-900 transmitter. The voltages obtained at the final node V_{DC} of the multiplier circuit were recorded for various input power levels from –40 dBm - +5 dBm with power level interval (spacing) of 5 dBm.

The simulations were also carried out using same stage capacitance value (3.3 nF) and then with a varied capacitance value for all stages from 4 stages through 9 stages [9]. The capacitance value was varied in such a way that, from one stage to the next, it was halved. For example, if the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

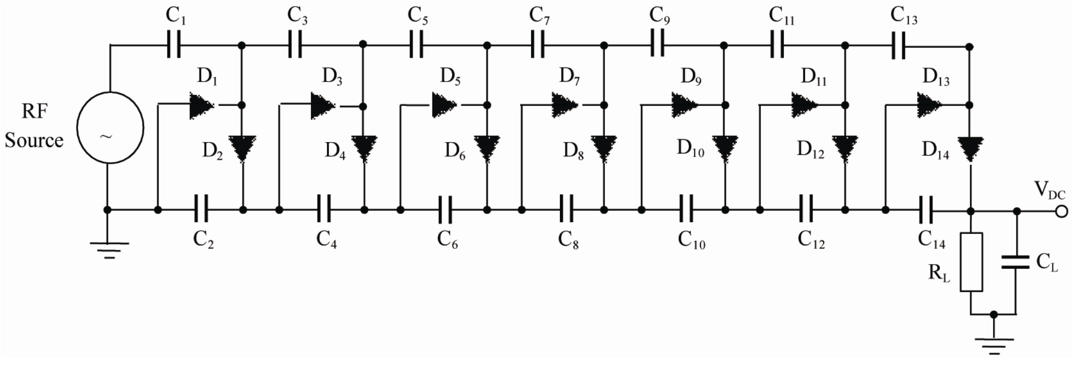

Figure 6. Schematic of 7 stage voltage multiplier.

manufacture the printed circuit board (PCB) is the standard Fiberglass Reinforced Epoxy (FR4), with the thickness of 1.6 mm and dielectric constant of 3.9. The topology is constructed on the PCB with the dimensions of 98 mm × 34 mm (W × H). The Sub Miniature version A (SMA) connectors are used at the input and output of PCB to carry out the measurements. The circuit components consist of active and passive components. The component used in circuit is shown in **Table 2**.

Special handling precautions have been taken to avoid Electro Static Discharge (ESD), while assembling of the surface-mount zero bias Schottky diodes. Also special attention has been given to mount other components and the SMA connectors on to the PCB. The Photograph of Assembled circuit board I shown in **Figure 7**.

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement respectively.

Figures 9 and **10** show the result of a 4 stage voltage doubler circuit with equal and varied capacitance values between the stages as described in Section 3.

From the analysis of these two simulations, it can be observed that the resulting output voltages are equal. The only difference between these two graphs is the rise time of the circuit with varied capacitance value is a little bit slower. But, overall result on the performance of rise time is still under 20 µs to 24 µs and the difference is negligible. From these results, the use of equal stage capacitance of each being 3.3 nF was hence considered for the design of the multiplier.

The results from **Figure 11**, shows that the output voltage reaches to 1.0 V within 20 µS and then uniformly increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

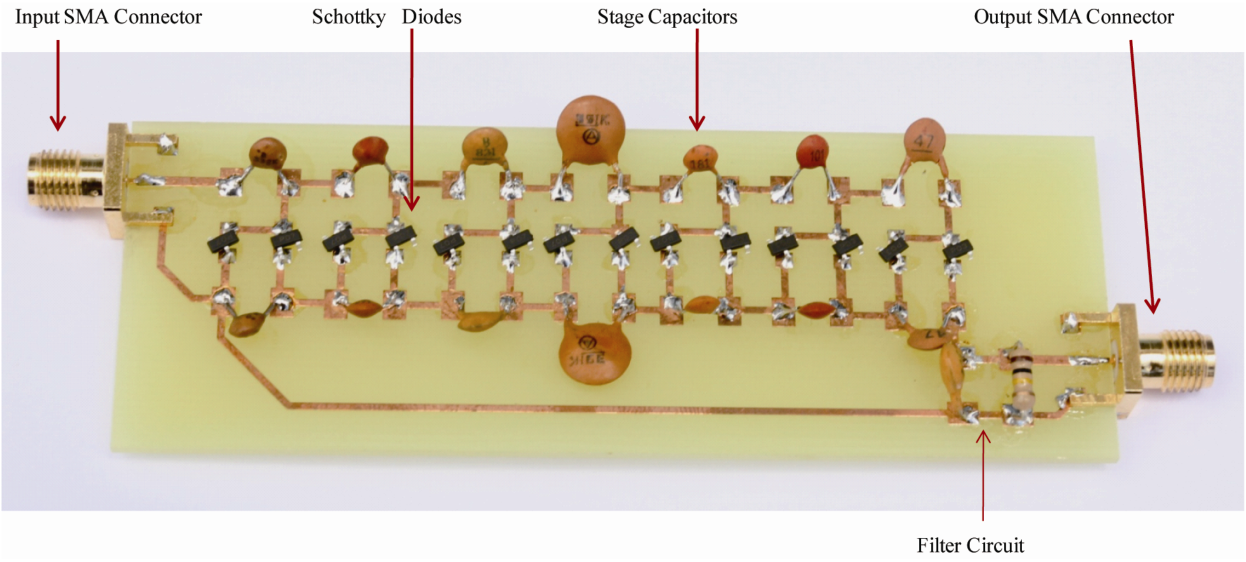

Figure 7. Photograph of assembled circuit board.

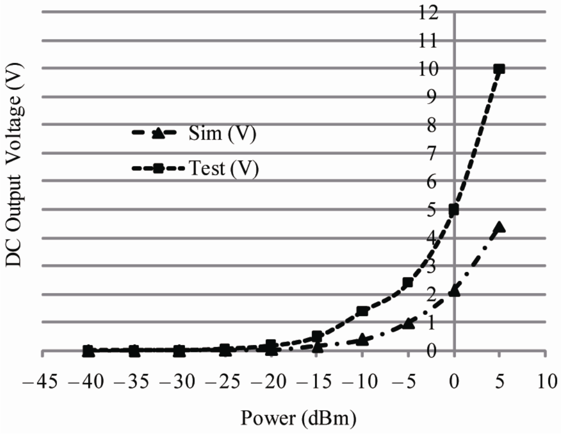

Figure 8. Simulated and test DC output voltage of multiplier as a function of input power.

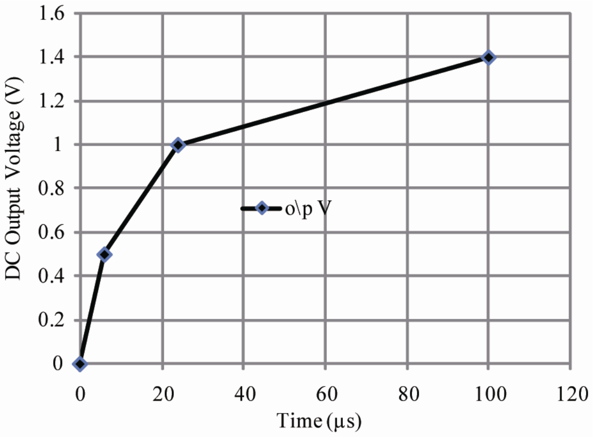

Figure 9. DC output voltage verses rise time of 4 stage voltage doubler circuit with equal stage capacitance [8].

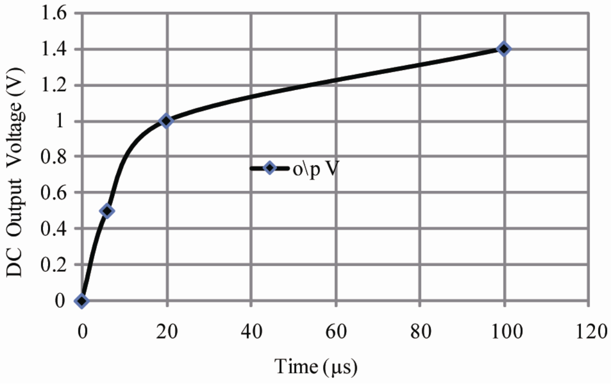

Figure 10. DC output voltage verses rise time of 4 stage voltage doubler with varied stage capacitance [8].

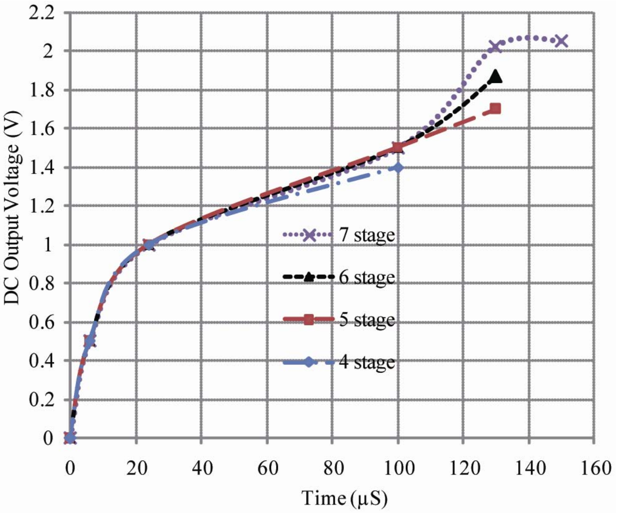

Figure 11. DC output voltage verses rise time of voltage doubler circuit through 4 - 7 stages with equal stage capacitance.

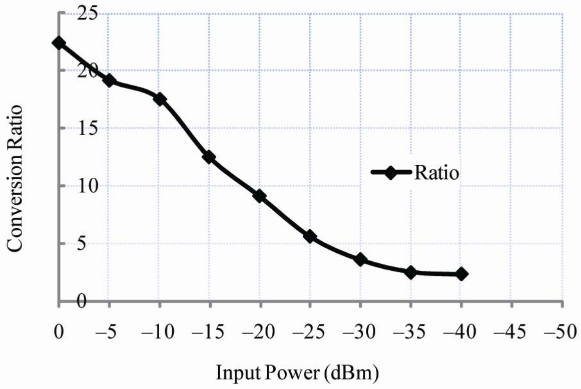

Figure 12. Conversion ratio as a function of input power.

posed voltage multiplier circuit operates at the frequency of 945 MHz with the specified input power levels. The results have shown that there is multiplication of the input voltage. From **Figure 12**, it is shown that at 0 dBm input power, the multiplication factor is 22. This is significant, as the work shows that RF energy in the GSM-900 band can be harvested from the ambient RF source using the Villard circuit topology. The power density levels from a GSM base station is expected from 0.1 mW/m^2 to 1 mW/m^2 for a distance ranging from 25 m - 100 m. These power levels may be elevated by a factor between one and three for the GSM-900 downlink frequency bands depending on the traffic density [10]. The next phase of the research work is to interface the voltage multiplier circuit through a matching network to the antenna at the input side and a low power device to power from the system at the output side to complete the RF energy harvesting system.

6. Acknowledgements

We would like to acknowledge and thank the Ministry of Higher Education Malaysia for funding this project under Fundamental Research Grant: FRGS/1/10/TK/UNITEN/02/13.

Copyright © 2012 SciRes.

REFERENCES

[1] S. von der Mark and G. Boeck, "Ultra Low Power Wakeup Detector for Sensor Networks," *SBMO/IEEE MTT-S International Microwave & Optoelectronics Conference* (*IMOC* 2007), Berlin, 29 October-1 November 2007, pp. 865-868. doi:10.1109/IMOC.2007.4404394

[2] H. Yan, J. G. M. Montero, A. Akhnoukh, L. C. N. de Vreede and J. N. Burghartz, "An Integration Scheme for RF Power Harvesting," *Proceedings of STW Annual Workshop on Semiconductor Advances for Future Electronics and Sensors*, Veldhoven, November 2005, pp. 64-66. doi:10.1.1.126.5786

[3] T. Ungan and L. M. Reindl, "Concept for Harvesting Low Ambient RF-Sources for Microsystems," 2007. http://www.imtek.de/content/pdf/public/2007/powermems_2007_paper_ungan.pdf

[4] J. Wang, L. Dong and Y. Z. Fu, "Modeling of UHF Voltage Multiplier for Radio-Triggered Wake-Up Circuits," *International Journal of Circuit Theory and Application*, Vol. 39, No. 11, 2010, pp. 1189-1197. doi:10.1002/cta.692

[5] J. A. Starzyk, Y.-W. Jan and F. J. Qiu, "A DC-DC Charge Pump Design Based on Voltage Doublers," *IEEE Transactions on Circuits and Systems-I*: *Fundamental Theory and Applications*, Vol. 48, No. 3, 2001, pp. 350-359.

[6] HSMS-2850, "Surface Mount Zero Bias Schottky Detector Diodes." http://www.crystal-radio.eu/hsms285xdata.pdf

[7] D. W. Harrist, "Wireless Battery Charging System Using Radio Frequency Energy Harvesting," M.S. Thesis, University of Pittsburgh, Pittsburgh, 2004.

[8] E. Bergeret, J. Gaubert, P. Pannie and J. M. Gaultierr, "Modeling and Design of CMOS UHF Voltage Multiplier for RFID in an EEPROM Compatible Process," *IEEE Transactions on Circuits and Systems-I*, Vol. 54, No. 10, 2007, pp. 833-837. doi:10.1109/TCSII.2007.902222

[9] B. Emmanuel, J. Gaubert, P. Pannier and J. M. Gaultier, "Conception of UHF Voltage Multiplier for RFID Circui," *IEEE North-East Workshop on Circuits and Systems*, Gatineau, June 2006, pp. 217-220. doi:10.1109/NEWCAS.2006.250961

[10] H. J. Visser, A. C. F. Reniers and J. A. C. Theeuwes, "Ambient RF Energy Scavenging: GSM and WLAN Power Density Measurements," *Proceedings of the* 38*th European Microwave Conference*, Eindhoven, 27-31 October 2008, pp. 721-724. doi:10.1109/EUMC.2008.4751554

Copyright © 2012 SciRes.

Journal of Electromagnetic Analysis and Applications, 2013, 5, 182-188
http://dx.doi.org/10.4236/jemaa.2013.54029 Published Online April 2013 (http://www.scirp.org/journal/jemaa)

Double-Sided Microstrip Circular Antenna Array for WLAN/WiMAX Applications

Ruwaybih Alsulami, Heather Song*

Department of Electrical and Computer Engineering, University of Colorado, Colorado Springs, USA.
Email: *hsong@uccs.edu

Received February 8th, 2013; revised March 10th, 2013; accepted March 25th, 2013

Copyright © 2013 Ruwaybih Alsulami, Heather Song. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

The design, fabrication, and characterization of the microstrip circular antenna arrays were presented. The proposed antennas were designed for single band at 2.45 GHz and dual bands at 3.3 - 3.6 and 5.0 - 6.0 GHz to support WLAN/WiMAX applications. The proposed single and dual band antennas showed omnidirectional radiation pattern with the gain values of 3.5 dBi at 2.45 GHz, 4.0 dBi at 3.45 GHz, and 3.3 dBi at 5.5 GHz. The dual band antenna array was placed on both top and bottom layers to obtain the desired antenna characteristics. The proposed double-sided dual band antenna provides omnidirectional radiation pattern with high gain.

Keywords: Antenna Arrays; Circular Patch; Dual Band; Single Band; Omnidirectional; WLAN/WiMAX Applications; UWB

1. Introduction

Ultra-wideband (UWB: 3.1 to 10.6 GHz) frequency spectrum has been approved by the US Federal Communications Commission (FCC) for unlicensed short range wireless communications since 2002. In this frequency range, wireless local-area network (WLAN) IEEE802.11a and HIPERLAN/2 WLAN operates in 5.0 - 6.0 GHz band. In some European and Asian countries, world interoperability for microwave access (WiMAX) service is provided in the frequency range of 3.3 - 3.6 GHz [1-4]. To support the WLAN/WiMAX application, antenna arrays that provide omnidirectional radiation pattern are required. To respond to this need, recent antenna design efforts were focused on omnidirectional antennas with high gain and no sidelobes [5-8]. Rectangular arrays are common type used for antenna arrays. Studies on dual band antennas employing rectangular arrays were reported [9-12]. Compared to rectangular patch antenna arrays, there are limited numbers of studies performed on circular patch antenna arrays due to difficulties in fabrication [13]. Advantages of circular antenna array include high gain and narrow beam width [13].

In this paper, a new microstrip circular antenna arrays were designed, fabricated, and characterized to provide omnidirectional radiation pattern for WLAN/WiMAX applications. Two antenna arrays were designed—one for single band at 2.45 GHz and the other for dual bands at 3.3 - 3.6 GHz and 5.0 - 6.0 GHz. For single band operation, circular patch array was placed on the top layer of the microtrip and a small rectangular patch was placed on the bottom layer for ground connection. For dual band operation, similar circular patch array was placed on both top and bottom layers of the microstrip with larger rectangular patch placed on the bottom layer. Both single band (single sided) and dual band (double-sided) microstrip antenna arrays provided desirable antenna characteristics for the intended application.

2. Design and Simulation

2.1. Single-Band Antenna at 2.45 GHz

The configuration of the proposed single band antenna at 2.45 GHz is shown in **Figure 1**. It consists of six circular patches which are placed only on the top layer. The small rectangular patch is placed on the bottom layer for ground connection.

The directivity for the circular patch antenna is

$$D_0 = \frac{(k_0 a_e)^2}{120 G_{\text{rad}}} \tag{1}$$

---

*Corresponding author.

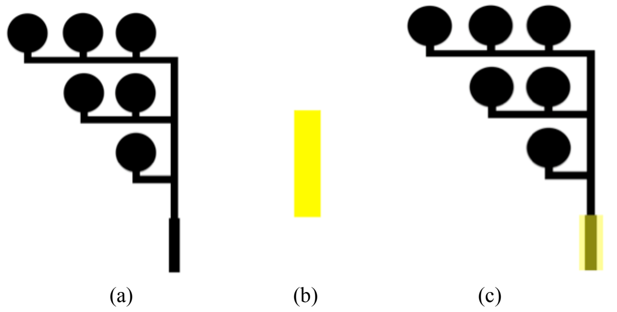
(a) (b) (c)

Figure 1. Configuration of the proposed antenna for single band at 2.45 GHz: (a) Top layer; (b) Bottom layer; (c) Top and bottom layers overlaid.

$$k_0 = \frac{2\pi}{\lambda_0} \tag{2}$$

$$a_e = a\left\{1 + \frac{2h}{\pi a \epsilon_r}\left[\ln\left(\frac{\pi a}{2h}\right) + 1.7726\right]\right\}^{1/2} \tag{3}$$

$$G_{\text{rad}} = \frac{(k_0 a_e)^2}{480}\int_0^{\pi/2}\left[J_{02}'^2 + \cos^2\theta J_{02}^2\right]\sin\theta \, d\theta \tag{4}$$

$$J_{02}' = J_0(k_0 a_e \sin\theta) - J_2(k_0 a_e \sin\theta) \tag{5}$$

$$J_{02} = J_0(k_0 a_e \sin\theta) + J_2(k_0 a_e \sin\theta) \tag{6}$$

where a_e is the effective radius, a is the actual radius, ϵ_r is the relative permittivity of the microstrip dielectric substrate, h is the height of the microstrip substrate, and J_0 and J_2 are Bessel functions.

The gain of the antenna was calculated using

$$\text{Gain} = \text{Antenna Efficiency} \times \text{Directivity}(D_0) \tag{7}$$

$$\text{Antenna Efficiency} = \frac{\text{Total Efficiency}}{\text{Reflection Efficiency}} \tag{8}$$

The variable corresponding to each dimensions and values for the dimensions of the proposed antenna are shown in **Figure 2** and **Table 1**, respectively. Here, *L*, *W*, and *R* represent the length, the width, and the radius of the circular patch, respectively.

The gain of the proposed antenna shown in **Figure 1** was calculated using (1) - (8) and the dimensions were optimized using ADS [14] which resulted in gain of 3.5 dBi at 2.45 GHz.

2.2. Dual-Band Antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz

The configuration for the doubled-sided microstrip dual band antenna is shown in **Figure 3**. The proposed microstrip antenna has circular arrays both on the top and bottom layers. It consists of three circular patched on each layer.

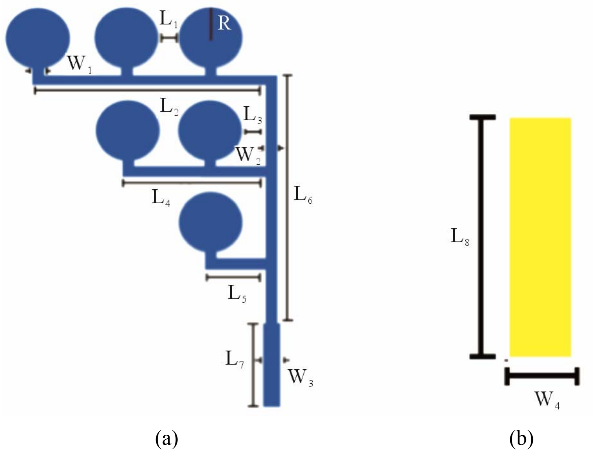

(a) (b)

Figure 2. Variables corresponding to each dimension of the proposed single band antenna: (a) Top layer; (b) Bottom layer.

Table 1. Dimensions for the proposed single band antenna at 2.4 GHz.

Variable	Value (mm)
L_1	1.15
L_2	28.9
L_3	1.13
L_4	17.6
L_5	6.60
L_6	35.3
L_7	18.4
L_8	18.4
W_1	1.02
W_2	1.52
W_3	3.30
W_4	6.86
R	5.10

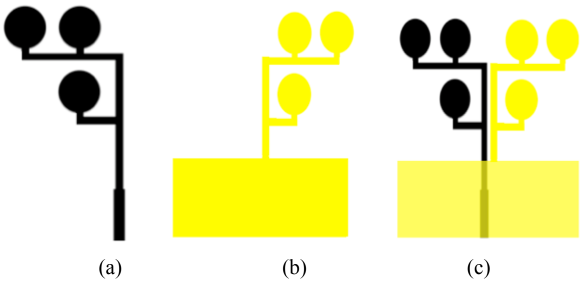
(a) (b) (c)

Figure 3. Configuration of the proposed antenna for dual band at 3.3 - 3.6 and 5.0 - 6.0 GHz: (a) Top layer; (b) Bottom layer; (c) Top and bottom layers overlaid.

The configuration in **Figure 3(a)** is similar to the top layer of the single band antenna as shown in **Figure 1(a)** but with less circular patches. However, the bottom layer in **Figure 3(b)** is different compared to the bottom layer of the single band antenna shown in **Figure 1(b)**. The double-side nature of the antenna provides dual band characteristics. Identical equations were used for the single band antenna were employed in the design process. The variable corresponding to each dimensions and the dimensions for the proposed dual band antenna are shown in **Figure 4** and **Table 2**, respectively.

Simulation was performed using ADS for the configuration shown in **Figure 3(c)**. The simulated gains of the proposed dual band antenna were 4.0 dBi at 3.45 GHz and 3.3 dBi at 5.5 GHz. The double-sided configuration of the antenna provided higher gain compared to the singled-sided antenna.

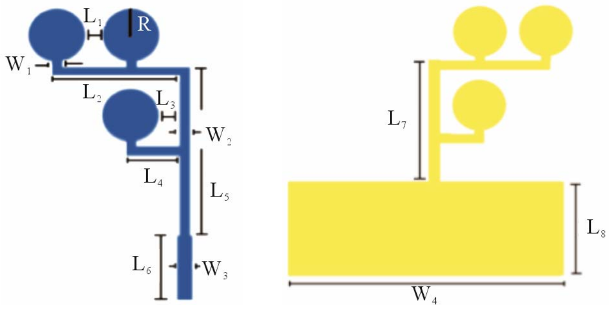

(a) (b)

Figure 4. Variables corresponding to each dimension of the proposed dual-band antenna: (a) Top layer; (b) Bottom layer.

Table 2. Dimensions for the proposed dual band antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz.

Variable	Value (mm)
L_1	1.20
L_2	17.8
L_3	1.08
L_4	6.60
L_5	29.3
L_6	10.0
L_7	21.6
L_8	17.8
W_1	1.02
W_2	1.52
W_3	2.54
W_4	49.1
R	5.21

3. Measurement Results and Discussions

3.1. Single-Band Atnenna at 2.45 GHz

The antennas were fabricated using LPKF Protomat [15] on FR-4 material with height of 1.524 mm. The photos of the fabricated single band antenna are shown in **Figure 5** which has a size of 6.7 × 4.4 (in cm).

Figure 6 shows the comparison between the simulated and the measured S_{11} results.

The measured operating frequency is close to 2.45 GHz with S_{11} value below −15 dB. The 3 dB bandwidth at 2.45 GHz was approximately 18%. The measurement and simulation are in fairly good agreement, and the differences are due to microstrip loss and fabrication errors.

Figure 7 shows the comparison between the simulated and measured radiation pattern in xy-plane at 2.45 GHz which is close to omnidirectional pattern.

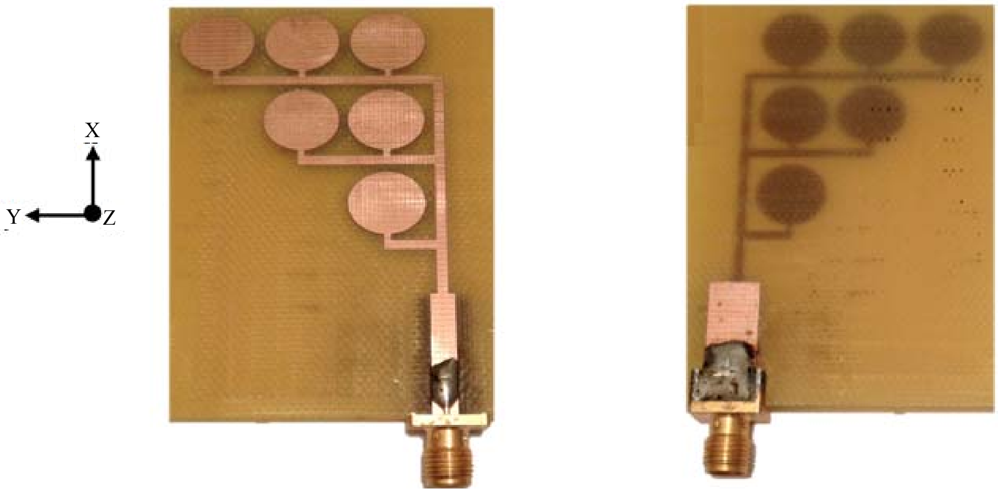

(a) (b)

Figure 5. Photo of the fabricated single-band antenna: (a) Top layer; (b) Bottom layer.

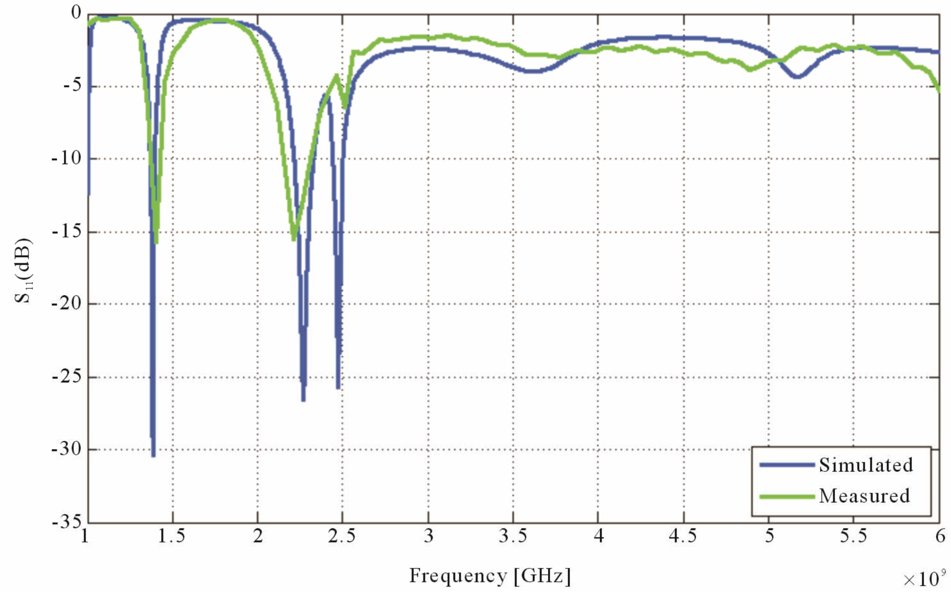

Figure 6. Simulated and measured return loss for the proposed single-band antenna.

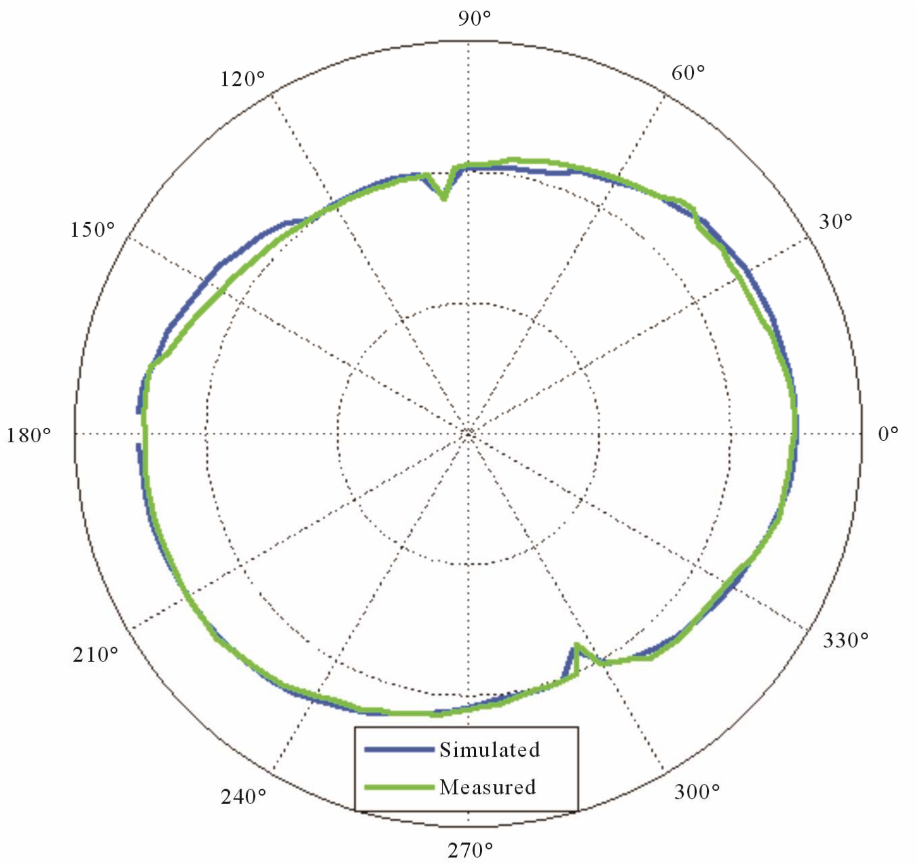

Figure 7. Simulated and measured radiation pattern in xy-plane (coordinate system shown in Figure 5) at 2.45 GHz.

3.2. Dual-Band Antenna at 3.3 - 3.6 and 5.0 - 6.0 GHz

The antennas were fabricated using LPKF Protomat [15] on double-sided FR-4 materials. The photos of the fabricated dual band antenna are shown in **Figure 8** which has a size of 6.6 × 5.2 (in cm).

Figure 9 shows the comparison between the simulated

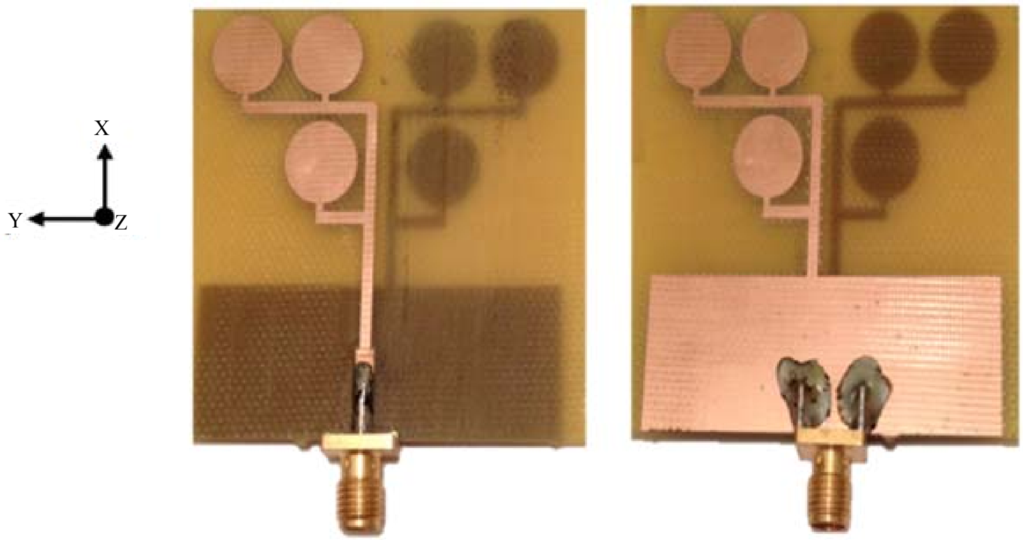

(a) (b)

Figure 8. Photo of the fabricated dual-band antenna: (a) Top layer; (b) Bottom layer.

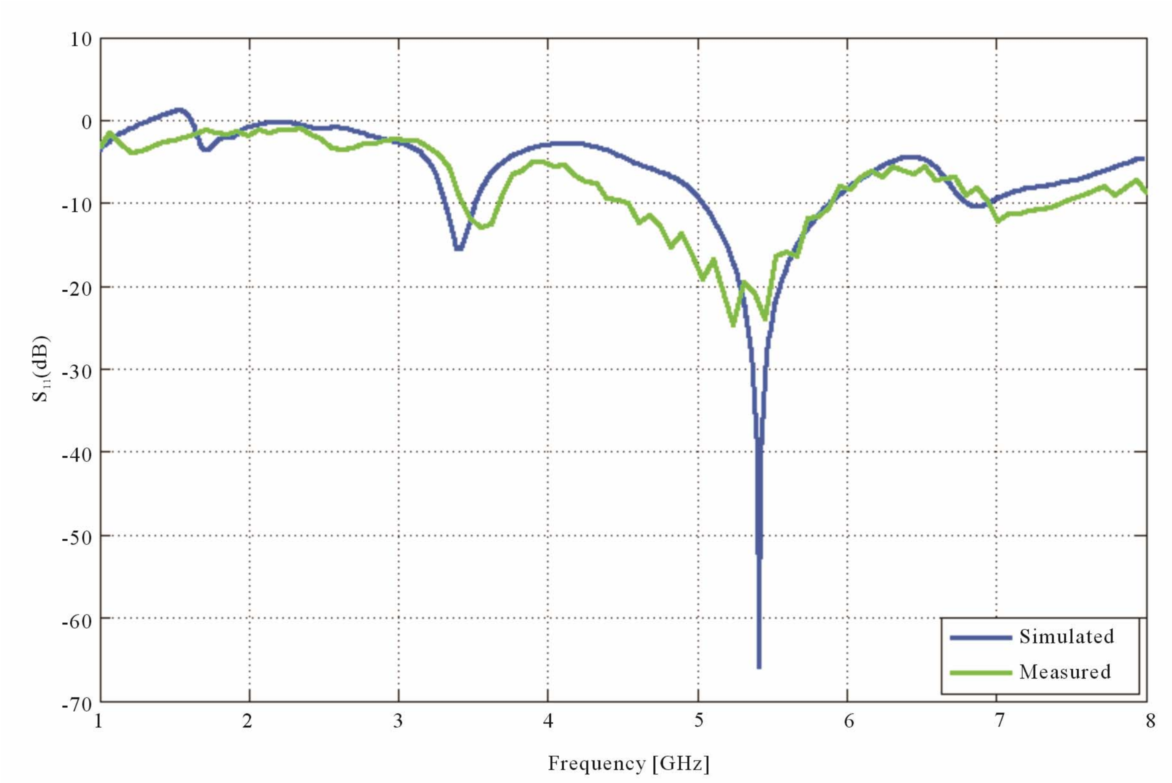

Figure 9. Simulated and measured return loss for the dual band antenna.

and the measured S_{11} results.

The measured S_{11} shows dual band near the designed bands with S_{11} values below −10 dB for both bands. The simulated and measured results give fairly good agreement, and the differences are due to board loss and fabrication errors.

Figure 10 shows the comparison between the simulated and measured radiation pattern in xy-plane at 3.45 and 5.5 GHz which is close to omnidirectional pattern.

4. Conclusion

A microstrip circular antenna arrays were presented for single band at 2.45 GHz and dual bands at 3.3 - 3.6 and

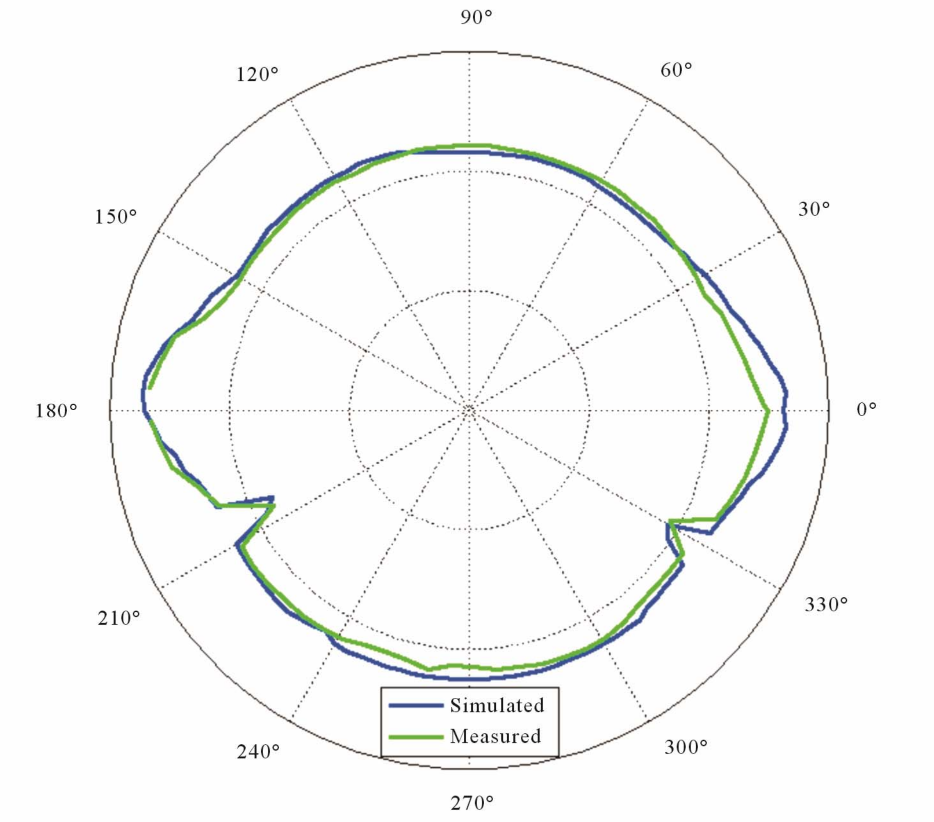

(a)

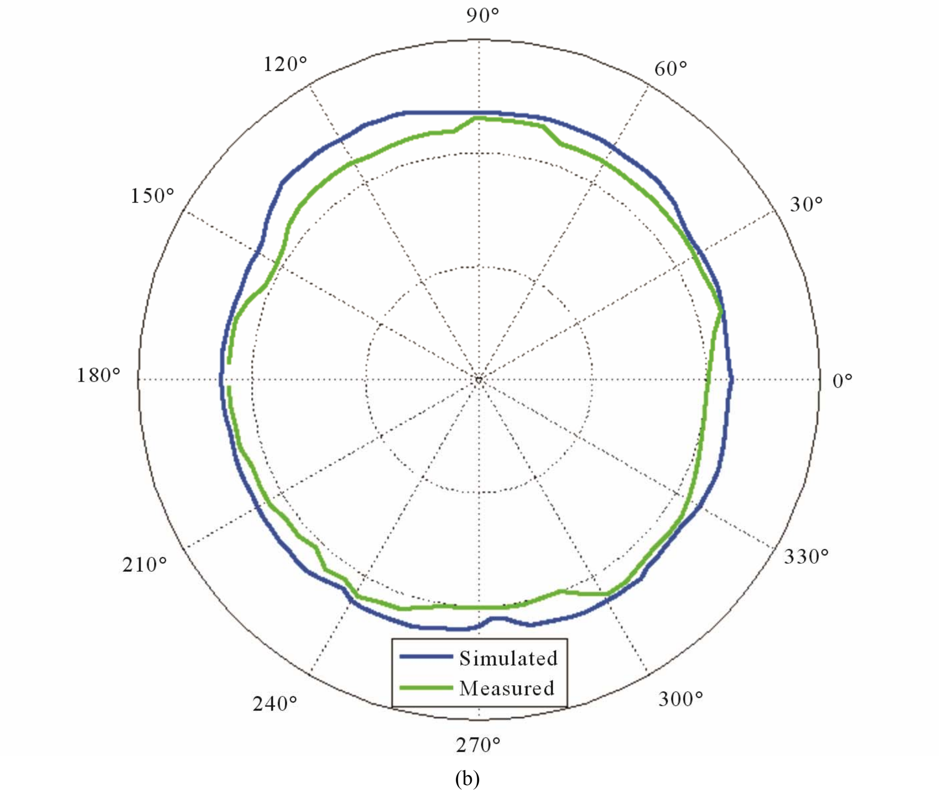

(b)

Figure 10. Simulated and measured radiation pattern in xy-plane (coordinate system shown in Figure 8) at (a) 3.45 GHz and (b) 5.5 GHz.

5.0 - 6.0 GHz for WLAN/WiMAX applications. Both antennas were designed with ADS, fabricated on a FR-4 microstrip material, and characterized. Both single band (single sided) and dual band (double-sided) antenna arrays provided omnidirectional pattern with desired gain.

REFERENCES

[1] C.-Y. Pan, T.-S. Horng, W.-S. Chen and C.-H. Huang, "Dual Wideband Printed Monopole Antenna for WLAN/WiMAX Applications," *IEEE Antennas and Wireless Propagation Letters*, Vol. 6, 2007.

[2] A. C. Rao and R. Pandeeswari, "A CPW-Fed Antenna for Dual Band WiMAX/WLAN Applications," *IEEE International Conference on Recent Trends in Information Technology ICRTIT*, 2011.

[3] H.-Y. Lai, Z.-Y. Lei, Y.-J. Xie, G.-L. Ning and K. Yang, "UWB Antenna with Dual Band Rejection for WLAN/WiMAX Bands Using CSRRs," *Progress in Electromagnetics Research Letters*, Vol. 26, 2011, pp. 69-78. doi:10.2528/PIERL11070112

[4] D. Parkash and R. Khanna "Design of a Dual Band Monopole Antenna for WLAN/WiMAX Applications," *IEEE Wireless and Optical Communications Networks*, 2010.

[5] L. Wang, K. Wei, J. Feng, Z. Zhang and Z. Feng, "A Wideband Omnidirectional Planar Microstrip Antenna for WLAN Applications," *IEEE*, 2011.

[6] J. Li, "An Omnidirectional Microstrip Antenna for WiMAX Applications," *IEEE Antennas and Wireless Propagation Letters*, Vol. 10, 2011.

[7] O. Tze-Meng and T. K. Geok, "A Dual-Band Omni-Directional Microstrip Antenna," *Progress in Electromagnetics Research*, Vol. 106, 2010, pp. 363-376. doi:10.2528/PIER10052411

[8] M. B. Bicer and A. Akdagli, "A Novel Microstrip-Fed Monopole Antenna for WLAN/WiMAX Applications," *Journal of Electromagnetic Waves and Applications*, Vol. 26, No. 7, 2012, pp. 904-913. doi:10.1080/09205071.2012.710372

[9] N. AbWahab, Z. Bin Maslan, W. N. W. Muhamad and N. Hamzah, "Microstrip Rectangular 4 × 1 Patch Array Antenna at 2.5 GHz for WiMAX Application," 2*nd International Conference on Computational Intelligence*, *Communication Systems and Networks*, 2010.

[10] Z. Zhong-xiang, C. Chang, W. Xian-liang and F. Ming-gang, "A 24 GHz Microstrip Array Antenna with Low Side Lobe," Springer, Berlin, 2012.

[11] A. K. Sahu and M. R. Das, "4 × 4 Rectangular Patch Array Antenna for Bore Sight Application of Consial Scan S-Band Traching Radar," *Antenna Week* (*IAW*), 2011.

[12] J. Das, T. A. Khan and M. K. Pal, "Rectangular Patch Antenna Array for Wireless Application," *International Journal of Engineering Science and Technology*, 2012.

[13] T. I. Huque, A. Chhowdhury, K. Hosain and S. Alam, "Performance Analysis of Corporate Feed Rectangular Patc Element and Circular Patch Element 4 × 2 Microstrip Array Antennas," *International Journal of Advanced Computer Science and Applications*, Vol. 2, No. 7, 2011, pp. 74-79.

[14] http://www.home.agilent.com/en/pc-1297113/advanced-design-system

[15] http://www.lpkf.com/protomat

J. Electromagnetic Analysis & Applications, **2011, 3, 43-46**
doi:10.4236/jemaa.2011.32008 Published Online February 2011 (http://www.SciRP.org/journal/jemaa)

Back to Back Combined Single Feed Proximity Coupled Antenna with Dumbbell Shaped DGS

Ashwini Kumar Arya, Amalendu Patnaik, Machavaram V. Kartikeyan

Department of Electronics & Computer Engineering, IIT Roorkee, Roorkee, India.
Email: ashwiniarya.iitr@gmail.com, {apatnaik, kartik}@ieee.org

Received November 12th, 2010; revised December 10th, 2010; accepted January 5th, 2011

ABSTRACT

Use of defected ground structure (DGS) to reduce the size of patch antenna is presented in this paper. In order to get a dipole like radiation pattern for some specific application a dumbbell shaped DGS is used in the common ground plane of back to back combined single fed proximity coupled antenna. A size reduction of about 60% is achieved. Parametric analysis has been done to see the resonance behavior of the antenna with DGS.

Keywords: *Defected Ground structure, Microstrip Antennas, Proximity Coupling*

1. Introduction

The continuous shrinking size of electronic equipments demands similar size antenna elements in order to fit properly in wireless devices without compromising the other radiation properties of the antenna. In this respect microstrip patch antennas are quite an obvious choice because of its other benefits like low profile, light weight, low cost and easy fabrication.

But as far as size of these patches concerned, the patch length should be around half-a-wavelength for the structure to act as a good radiator. Different techniques have already been used for the antenna size reduction such as using the substrate with high dielectric constant [1] , edge shorted patches with shorting plates or shorting walls, use of the shorting pin at the suitable position etc [2,3].

As far as our understanding goes much has not been reported regarding the use of DGS for size reduction of microstrip antennas, although its application have been reported for harmonic reduction [4], cross-polarization suppression [5] and mutual coupling reduction [6] in antenna arrays etc. Although the back to back geometries have been reported by the various researchers [7,8] but here a new coupling method *i.e.* proximity coupling with the defected ground structure is used for the consideration of the increased bandwidth.

This paper presents the application of DGS for size reduction of microstrip antennas. A dumbbell shaped DGS is used in the common ground plane of a back to back combined single feed proximity coupled microstrip antenna.

2. Defected Ground Structure

Recently there has been an increasing interest in the use of DGSs for performance enhancement of microstrip antennas and arrays. These are realized by etching off a simple shape defect from the ground plane of the planer circuits.

Although various complicated DGSs were reported in the literature, but the simplest one is the dumbbell shaped DGS. **Figure 1(a)** shows the simple and mostly used dumbbell shaped DGS that is etched in the ground plane below the microstrip line, in which both the areas (L_g*W_g) and the slot gap (g) play a very important role to find the resonance behavior of the DGS.

The head areas (L_g*W_g) is very useful for the variation in the inductance (L) and the slot (g) produces the capacitance (C). The L and C may be calculated from the formulae given below [9].

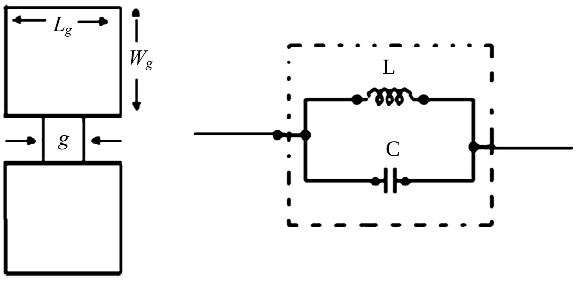

(a) (b)

Figure 1. (a) Dumbbell shaped DGS, and (b) DGS Equivalent Circuit.

Copyright © 2011 SciRes.

$$\mathrm{L} = \frac{1}{4\pi^2 f_0^2 C} \quad (1)$$

$$\mathrm{C} = \frac{f_c}{2Z_0} \times \frac{1}{2\pi(f_0^2 - f_c^2)} \quad (2)$$

When this DGS is applied to the antenna, the equivalent inductive part due to the DGS increases and produces equivalently the high effective dielectric constant, thereby decreasing the resonant frequency when the DGS is incorporated in the ground plane of a micro strip antenna.

3. Antenna Design with DGS

In this section, the design approach and the performance of the basic antenna and the antenna with DGS is described. At the outset, the single patch antenna was designed and simulated using the CST Microwave studio [10], for the operating frequency at 5.0 GHz. Then another patch of the same size was added in the opposite side of the ground plane and fed in the same way as the first one. The configuration seems as two patch antennas having a common ground plane working at the same frequency. Next the feed lines were combined for the antenna for single feed design. For this purpose the antenna feed lines W_f = 0.934 mm with d_f = 1.5 mm were designed of 100 ohm and for matching to the 50 ohm transmission line (W_p = 3.86 mm) a quarter wave length transformer (W_t = 2 mm) was used to give proper matching (**Figure 2**).

The cross-sectional view of the single fed back to back combined proximity coupled antenna is shown in the **Figure 2(a)**. The layouts of matching networks only are emphasized in this **Figure 2(b)** for convenience. It is observed that the antenna designed in this configuration gives the bandwidth of 137 MHz whereas single antenna gives a bandwidth of 67 MHz. This is due to the fact that as the antenna height increases the quality factor decreases and the bandwidth increases. This becomes a multilayer antenna with more height and higher bandwidth as compared to the single patch antenna.

The simple transmission line model was used for the antenna size calculation. The dielectric constant was taken as 3.38 with the loss tangent 0.0025 and of 1.524 mm thickness. The patch lengths L_p and widths W_p are 15 mm and 19 mm respectively. The feed line has been inserted inside the dielectric at a height equal to the half of the height (h = 3.048 mm) of the antenna on either side. The dumbbell shaped DGS with dimensions L_g = W_g = 8.6 mm and g = 0.76 mm was created in the ground plane of the antenna as shown in the **Figure 2(c)**. A small slot was also created for making the antenna with single feed. The two feed lines were connected with a metal strip which goes through the small slot in the ground plane. The fabricated antenna is shown in **Figure 2(d)**.

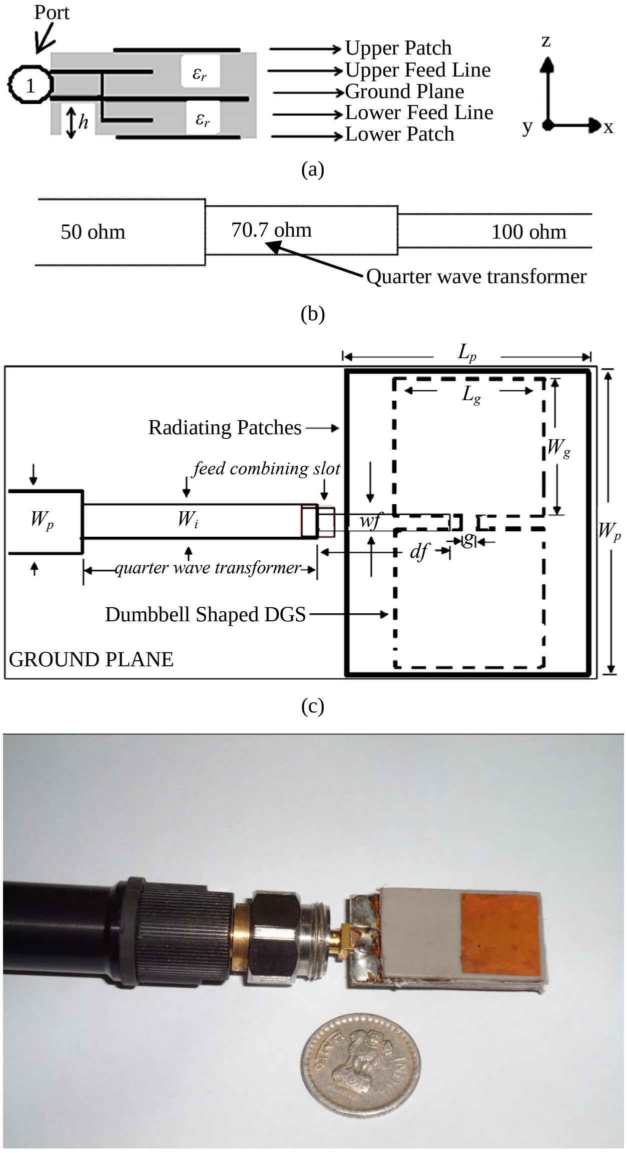

(a)

(b)

(c)

(d)

Figure 2. (a) Cross-sectional view of antenna configuration, (b) Feeding Network, (c) Top-View of the antenna, (d) Fabricated Antenna.

4. Results and Discussion

At first the antenna without the DGS in the common ground was simulated and was found to resonate at 5 GHz with 137 MHz Bandwidth.

Then the structure was simulated with the dumbbell shaped DGS. Before reaching to the final size of the DGS, a parametric study was done by varying L_g, W_g and g of

Copyright © 2011 SciRes.

the DGS. As shown in **Figure 3**, with the increase in the value of L_g, the resonant frequency of the antenna is decreasing. Infect, increase in the length of the DGS head gives increasing inductance which in turn decreases the resonant frequency of the antenna.

At this point the increment in the (*g*) was not possible due to the accuracy in fabrication, so for this reason the other dimension (*g*) was kept constant for the requirement of the desired frequency (2 GHz). The size of the DGS single square head for the antenna to resonate at 2 GHz (UHF Band) was found to be 8.6 mm × 8.6 mm. The return loss (S11 [dB]) plot of the structure with and without DGS is shown in **Figure 4(a)**. **Figure 4(b)** shows the measured S11 parameter using the HP 8720 B network analyzer. The marker's position near to peak shows the resonance frequency 2.08 GHz with return loss of -13 dB. The result shows good agreement with the simulation results. The measured -10 dB bandwidth is about 60 MHz. The maximum size reduction achieved is about 60%.

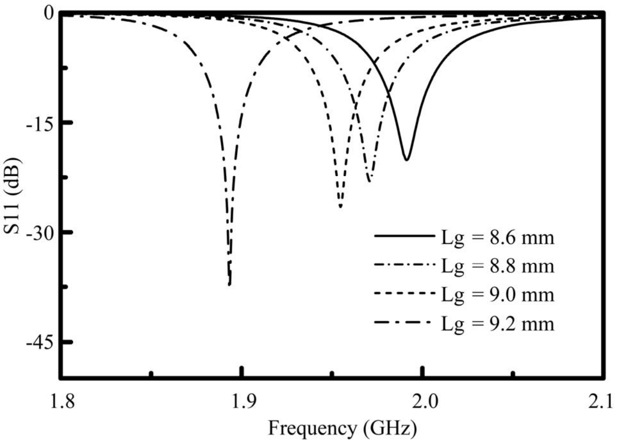

Figure 3. S11 Vs. Frequency response by varying DGS length (Lg).

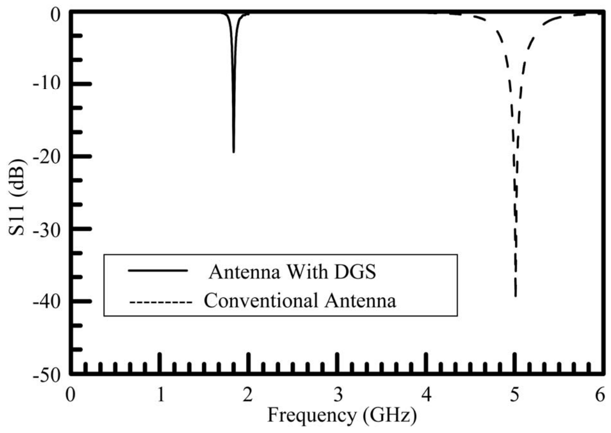

(a)

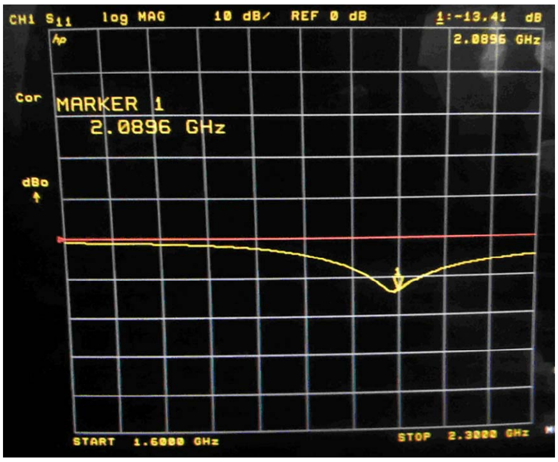

(b)

Figure 4. Antenna Return Loss (a) simulated (b) measured.

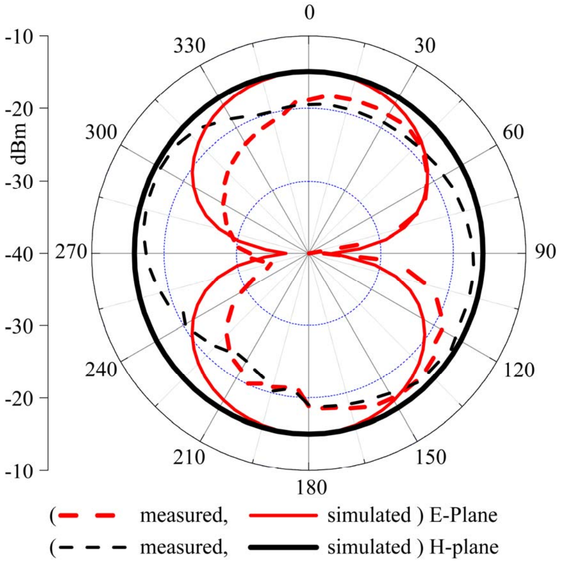

Figure 5. Antenna radiation pattern (measured and simulated) at 2 GHz Frequency.

The measured and simulated power patterns of the antenna are shown in **Figure 5**. It can be observed that the E-plane radiation pattern is similar to the pattern for a dipole antenna. Measurement errors are mainly due to the spurious radiation created by the feeding end and the improper coupling of the elements. However the gain measured experimentally for the proposed antennas with DGS is about -6.9 dB and -7.8 dB (where simulated gain with DGS is 3.8 dB and 5.7 dB is for the antenna without DGS, for both the planes) in both the E and H plane respectively, which is consistent with the size reduction of the antenna.

5. Conclusion

Microstrip patch antenna size reduction with DGS is car-

Copyright © 2011 SciRes.

ried out in this work. A dumbbell shaped DGS in the common ground plane of a back to back microstrip structure was found to give a size reduction of about 60% and shifts the resonance frequency from 5 GHz to 2 GHz, with 60 MHz bandwidth facilitating the antenna, to be used for UHF band applications.

REFERENCES

[1] R. Garg, P. Bhartia, I. Bahl and A. Ittipiboon, "Microstrip Antenna Design Handbook," Artech House, London, 1995.

[2] P. Salon, M. Keskilammi and M. Kivikoski, "Single Feed Dual Band Planer Inverted-F Antenna with U-Shaped Slot," *IEEE Transactions on Antenna and Propagation*, Vol. 48, No. 8, 2000, pp. 1262-1264. doi:10.1109/8.884498

[3] D.Nashhat, H. Elsadek and H. Ghali, "Dual-Band Reduced Size PIFA Antenna with U-Slot for Bluetooth and WLAN Applications," *Proceedings of IEEE Antennas Propagation International Symposium*, Columbus, Vol. 2, 22-27 June 2003, pp. 962-965. doi. 10.1109/APS.2003.1219395

[4] H. Liu, X. Sun and J. Mao, "Harmonic Suppression with Photonic Bandgap and Defected Ground Structure for a Microstrip Patch Antenna," *IEEE Microwave and Wireless Component Letters*, Vol. 15, No. 2, February 2005, pp. 55-56. doi:10.1109/LMWC.2004.842809

[5] D. Guha, M. Biswas and Y. M. M. Antar, "Microstrip Patch Antenna with Defected Ground Structure for Cross Polarization Suppression," *IEEE Antenna and Wireless Propagation Letter*, Vol. 4, No. 1, 2005, pp. 455-458. doi:10.1109/LAWP.2005.860211

[6] M. Salehi, A. Motevasselian, A. Tavakoli and T. Heidari, "Mutual Coupling Reduction of Microstrip Antenna using Defected Ground Structure," *IEEE Singapore International Conference on Communication System*, Singapore, 2006, pp. 1-5. doi:10.1109/ICCS.2006.301375

[7] J. Guterman, Y. R. Samii, A. A. Moreira and C. Peixeiro, "Quasi-Omni Directional Dual-band Back-to-Back E-shaped Patch Antenna for Laptop Applications," *Electronics Letters*, Vol. 42, No. 15, 2006, pp. 845-847. doi:10.1049/el:20061799

[8] H. Iwasaki, "A Back-to-Back Rectangular-Patch Antenna Fed by a CPW," *IEEE Transactions on Antenna Propagation*, Vol. 46, No. 10, October 1998, pp. 1527-1530. doi:10.1109/8.725285

[9] D. Ahn, J. S. Park, C. S. Kim, J. Kim, Y. Qian and T. Itoh, "A Design of the Low-Pass Filter Using The Novel Micro strip Defected Ground Structure," *IEEE Transactions on Microwave Theory and Technology*, Vol. 49, No. 1, 2001, pp. 86-91. doi:10.1109/22.899965

[10] EM Simulation Software CST Studio Suite™, v9.

Journal of Electromagnetic Analysis and Applications, 2012, 4, 362-366
http://dx.doi.org/10.4236/jemaa.2012.49051 Published Online September 2012 (http://www.SciRP.org/journal/jemaa)

Vehicular Radio Scanner Using Phased Array Antenna for Dedicated Short Range Communication Service

Tapas Mondal[1], Priyanka Shishodiya[1], Rowdra Ghatak[2], Sekhar Ranjan Bhadra Chaudhuri[3]

[1]Department of Electronics & Communication Engineering, Dr. B. C. Roy Engineering College, Durgapur, India; [2]Microwave and Antenna Research Laboratory, Department of Electronics and Communication Engineering, National Institute of Technology Durgapur, Durgapur, India; [3]Department of Electronics & Telecommunication Engineering, Bengal Engineering and Science University, Shibpur, India.
Email: tapas2k@gmail.com

Received July 10th, 2012; revised August 11th, 2012; accepted August 21st, 2012

ABSTRACT

Now a day's accidents are very common due to increased population of vehicle. In order to ensure safety measures in the vehicle this paper has proposed some methodologies regarding careful driving by automatically scanning and analyzing the blind spot area of an intelligent mobile vehicle. A vehicular antenna with minimum perturbation is proposed to be fitted on the vehicle and collect information of the concern area which would ensure visibility of the operator *i.e.* masked or integrated within the car body. This paper has dealt with the design of Tchebyscheff polynomial based prototype planar microstrip phased array antenna and also redesigned the same when implemented in the body of the car being considered as an electromagnetically large element. Both the design has been experimentally verified with the measurement. The simulated and the measured results in both the cases are found to be in good agreement. More than 11 dB gain was observed at perfectly 30° angles from its broad side direction as desired for blind spot detection with minimum amount of electromagnetic interference inside the car.

Keywords: Intelligent Transportation System; Microstrip Phased Array Antenna; Tchebyscheff Polynomial; Electromagnetic Interference

1. Introduction

Intelligent Transport System (ITS) ensures mobility comfort and safety in transportation system. It also absorbs the hazards due to environmental impact. With the progress of the information processing technology, control systems to minimize accidents for the roadways have also been advanced hence one approach to improve the traffic safety is found and that is automatic collection of data by scanning the blind spot area of the vehicle [1] as shown in **Figure 1**. Many methods were proposed to detect the blind spot area but all of them had certain limitations. Devices like dynamic angling side view mirror [2], side view camera model [3], and shadow or edge features detector [4] were used for detecting blind spot area but their performance was affected during bad weather, fog or mist. Also we know that the mechanical systems, response time is more and the system is prone to wear and tear.

A radio frequency method has been proposed in this paper to scan the blind spot zone efficiently. Four rectangular microstrip antennas (RMSA) are arranged in linear configuration with optimal spacing between the patch elements to construct the phased array radar. A corporate feed network is used to feed the patch element unequally and a progressive phase shifter is designed with 108˚ delay elements to tilt the main beam in the desired direction and then the total unit is simulated and experimented after placing it on the car body which is electromagnetically a large element. The antenna works in the Dedicated Short-Range Communication Service (DSRCS) frequency band [5]. The design of the microstrip phased array antenna is discussed in Section 2 and after that to place the antenna; the design of the entire car is given in Section 3. Results and discussion are portrayed in Section 4 along with conclusion in Section 5.

2. Design of Microstrip Phased Array Antenna

The microstrip phased array antenna is designed for Dedicated Short Range Communication Service at 5.88 GHz with dielectric constant of 2.32 and substrate thickness of 0.785 mm. Firstly the dimension of the rectangular patch is computed as 16.322 mm by 19.8 mm using the method outlined in [6]. With computed inset feed length of 3.8 mm and for the above dimensions of the rectangular patch the return loss is found to be –8.09 dB

Copyright © 2012 SciRes.

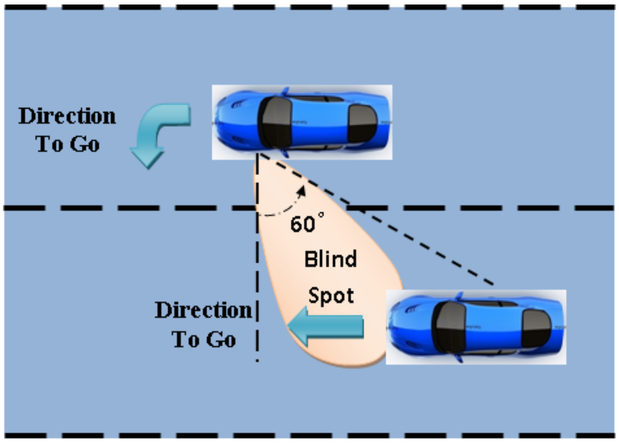

Figure 1. Surrounding regions of a vehicle.

with (30.1 + j26.7) ohm impedance at the feed position. The RMSA is simulated and optimized using Ansoft HFSS^{TM}. After optimization the final length is found to be 16.4363 mm with inset feed length of 4.8225 mm keeping the width of the patch unchanged as shown in **Figure 2**.

By considering the above designed patch element, a four element linear array is realized and powered by Tchebyscheff current distribution. The spacing between the elements is kept considering the desired maximum scan angle in order to eliminate the grating lobes within the visible space of the phased array antenna. To optimize the performance of the antenna in respect of its side lobe level (SLL), mutual coupling and gain of the antenna array, the spacing between the elements is studied parametrically [7]. The results of the said study are tabulated in **Table 1**.

After optimization it is found that the optimum spacing between the elements is 0.6 λ_0 while considering, the main lobe to side lobe ratio below 20 dB, optimum mutual coupling and overall gain.

From **Figure 1**, it is observed that the beam of the antenna array is required to be tilted by an angle 30° away from the broad side direction. In view of the above a progressive phase shifter of –108° is designed with the help of 11.16 mm feed line length. The actual line length is considered as mτ, where m = 0, 1, 2, 3.

To enhance the gain by maintaining the beam width and main lobe to side lobe ratio, the antenna elements are excited by Dolph Chebyshev current distribution. The array consist of four elements, thus third order Tchebyscheff polynomial is calculated. Hence the polynomial is solved and relative current ratio is computed as 1:1.7795:1.7795:1. Both equal and unequal power dividers are designed along with the progressive phase shifter.

Table 1. Effect on mutual coupling and gain due to variation in spacing (d) between the elements.

d with respect to guided wavelength	d with respect to operating wavelength	Mutual coupling (dB)	Gain (dBi)
0.6 λ_g	0.41 λ_0	>10	7.6
0.7 λ_g	0.48 λ_0	>13	9.9412
0.8 λ_g	0.55 λ_0	>20	10.647
0.9 λ_g	0.6 λ_0	>22	10.919

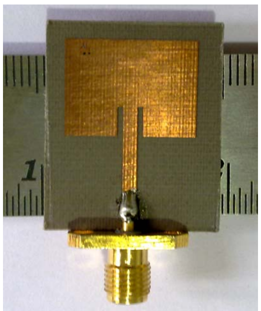

Figure 2. Photograph of the rectangular microstrip antenna element.

For equal power division, a 3 dB equal power divider is designed whose vertical arm is of 50 Ω line and two horizontal quarter-wavelength branch-lines are of 70.71 Ω is used [8]. It is shown in **Figure 3(a)**.

To get the unequal current distribution, both the horizontal arm of the microstrip unequal power divider should be of different resistance and it is calculated as 46.368 Ω and 82.49 Ω. The structure is shown in **Figure 3(b)**.

Generally for practical design purpose the gap between the corporate feed networks is taken small but the value of the gap less than 0.11 λ_0 disturbs the frequency response of overall antenna array. The effects of the feed network on the side-lobes are significant in the E-plane pattern. The influence on the H-plane is less important due to the orthogonality of the drive current and the symmetry of all the other currents in this cut [9]. By maintaining the SLL and the surface wave loss, the gaps are taken as 0.11 λ_0 as portrayed in **Figure 4**.

3. Placement of Antenna on the Vehicle

In order to compute the electromagnetic effect of the body of the car on the antenna performance, the entire structure of the car is designed in Ansoft HFSS^{TM}. The complete structure of the car consist of hood, roof, trunk, left and right side doors, left and right side quarter panel for front and rear side which are made up of conducting material but the bumpers and wheels are assigned by layered impedance due to their hard rubber and polyester materials as shown in **Figure 5**.

After going through the typical requirement for scanning the blind spot, the antenna is placed just beside the side view mirror. The following problems in the designed are addressed while simulating the effect of the radiation

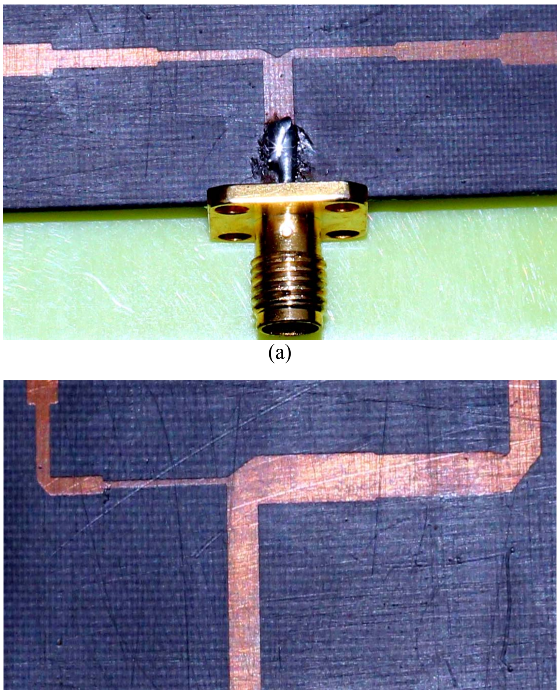

(a)

(b)

Figure 3. (a) Photograph of the equal microstrip power divider; (b) Photograph of the unequal microstrip power divider.

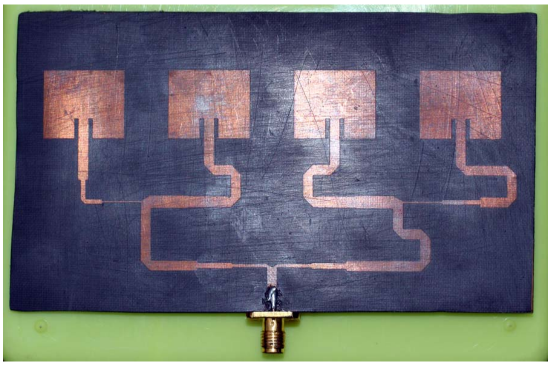

Figure 4. Photograph of the antenna array with corporate feed network.

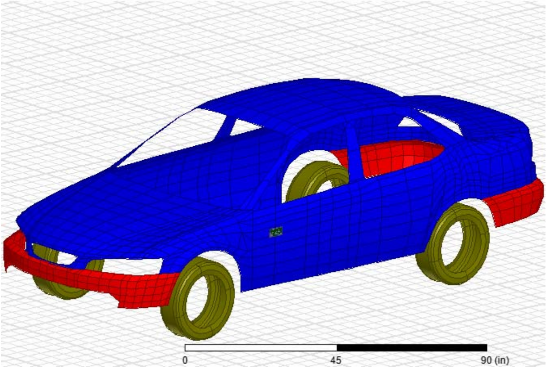

Figure 5. The entire structure of the car along with the antenna array.

characteristics of the antenna after placing it on the vehicle:

1) Spans a large computational domain compared to the wavelength;

2) Requires a fine grid resolution to resolve the detailed antenna structure;

3) Contains largely non-conformal PEC and dielectric structures.

4. Results and Discussion

For obtaining the scattering parameter characteristics, the microstrip antenna has been simulated over the frequency bandwidth ranging from 5.5 GHz to 6.5 GHz and two different circumstances is observed. The first experiments involve the design, simulation and measurement after fabrication of an individual element and then the array with the same without involving the effect of the car. Both of them should used for optimized performance.

The scattering parameter is measured by Agilent Technology Vector Network Analyzer model no. N5320A (10 MHz - 20 GHz) and the radiation pattern is measured by Hittite HMC-T2100 synthesized signal generator (10 MHz - 20 GHz) and Krytar 9000 B power meter (10 MHz - 40 GHz) with 9530 B power sensor (10 MHz - 20 GHz) as shown in **Figure 6**. From **Figure 7**, it is found

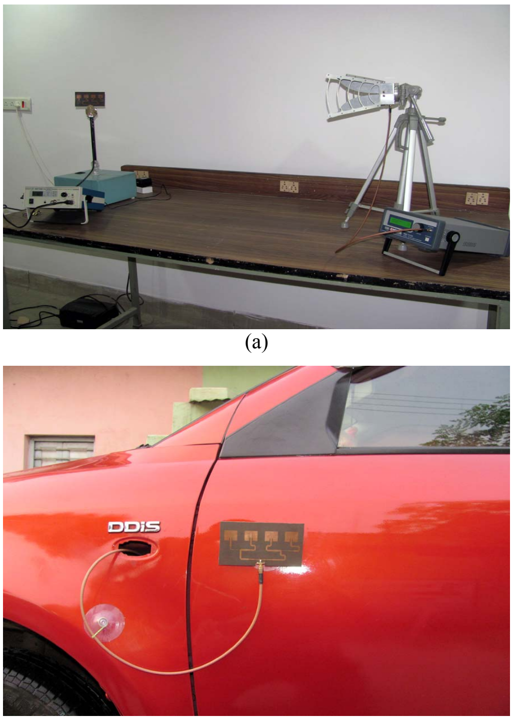

(a)

(b)

Figure 6. (a) Experimental set-up to measure the radiation pattern of the antenna array; (b) Experimental set-up to measure the radiation pattern of the antenna array after placing it on the car.

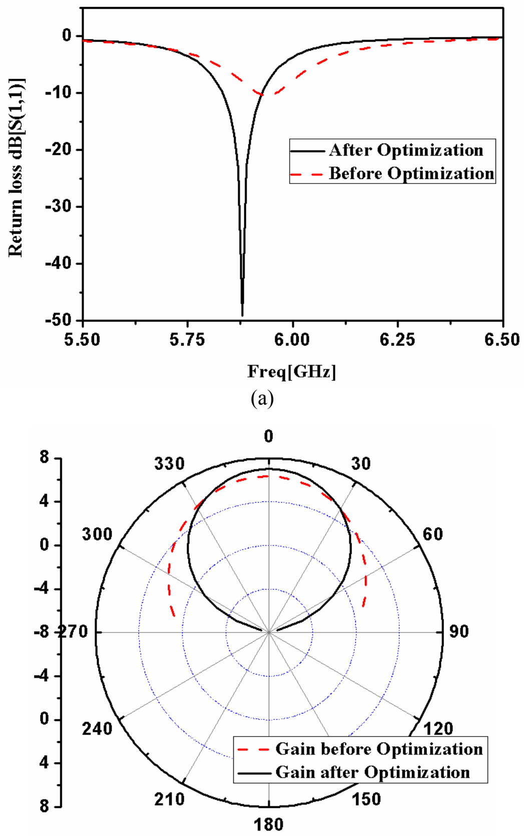

(a)

(b)

Figure 7. (a) Return loss a single patch before and after optimization; (b) Gain of a single patch before and after optimization.

that after optimization a perfect match is obtained exactly at 5.88 GHz and gain is also improved from 6.33 dB to 7.01 dB with a very good return loss of –47 dB. The current distribution and the radiation pattern of a single RMSA is portrayed in **Figure 8**.

A parametric study of the return loss of antenna array with different spacing between the elements of the corporate feed network is done and is compared in **Figure 9**. The results of the computational model of the entire antenna array with feed network were compared with the measured data and a good agreement is observed in **Figure 10**.

After getting the satisfactory results, next the same array antenna is experimented again with considering the electromagnetic effect of the car. In order to test the antenna array in association with the vehicle, a numerical model of the car is created using Ansoft HFSS$^{\text{TM}}$ and the simulated data is compared with the measured in **Figure 11**. **Figure 12** graphically represents the 3D radiation pattern of the antenna array in presence of the computational model of the vehicle. From **Figure 13**, it is clear that the inside electric field strength is not more than 15 Volt per meter which is very less compare to the standard

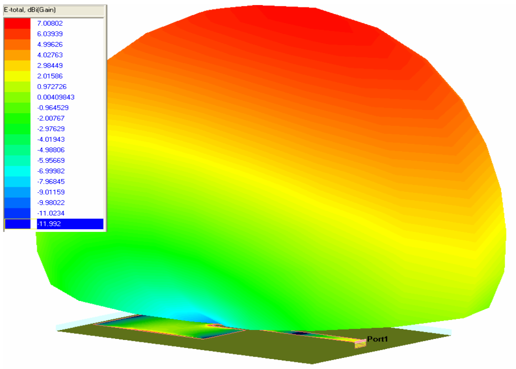

Figure 8. Current distribution and the radiation pattern of the rectangular microstrip antenna.

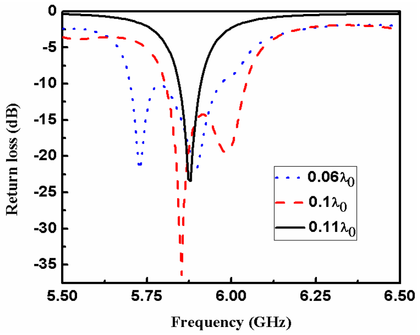

Figure 9. Parametric study of return loss of the antenna array with different spacing between the elements of corporate feed network.

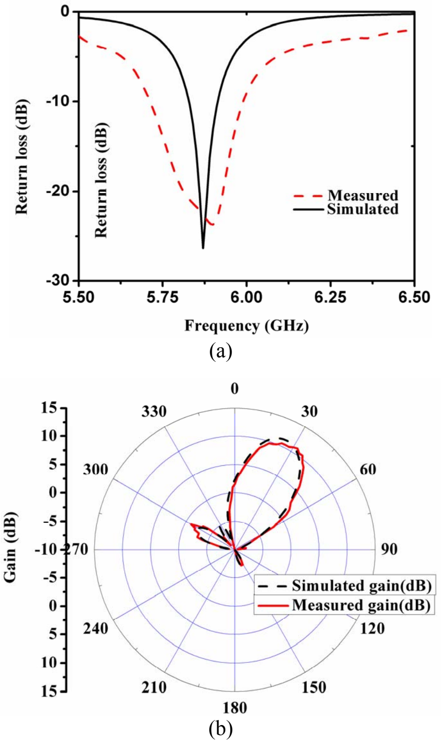

(a)

(b)

Figure 10. (a) Return loss of the antenna array; (b) Radiation pattern of the antenna array.

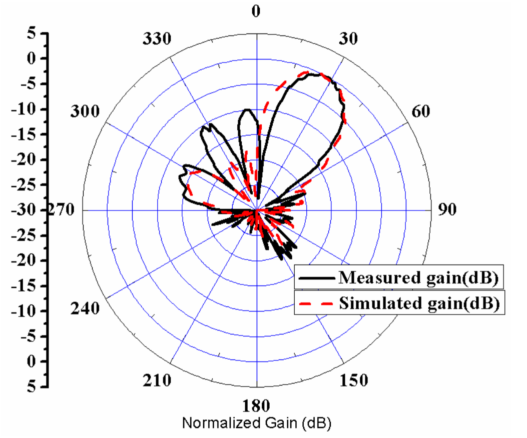

Figure 11. Normalized radiation pattern of the antenna array after placing it on the body of the car.

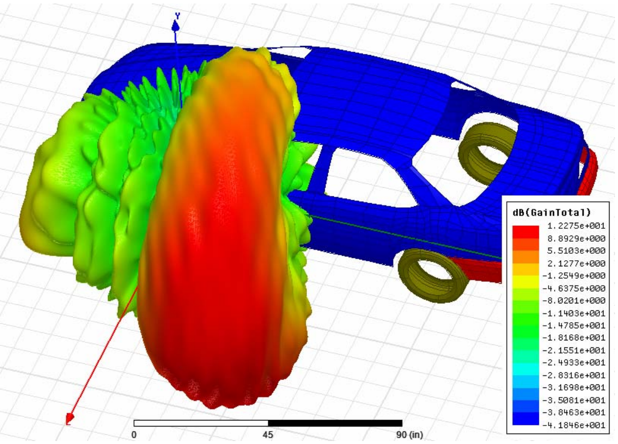
Y
dB(GainTotal)
1.2275e+001
8.8929e+000
5.5103e+000
2.1277e+000
-1.2549e+000
-4.6375e+000
-8.0201e+000
-1.1403e+001
-1.4785e+001
-1.8168e+001
-2.1551e+001
-2.4933e+001
-2.8316e+001
-3.1698e+001
-3.5081e+001
-3.8463e+001
-4.1846e+001
0
45
90 (in)

Figure 12. 3D radiation pattern of the antenna array after placing it on a vehicle body.

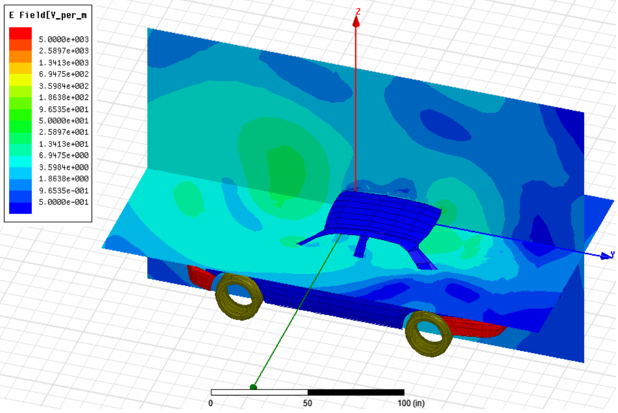
E Field[V_per_m
5.0000e+003
2.5897e+003
1.3413e+003
6.9475e+002
3.5984e+002
1.8638e+002
9.6535e+001
5.0000e+001
2.5897e+001
1.3413e+001
6.9475e+000
3.5984e+000
1.8638e+000
9.6535e-001
5.0000e-001
Z
Y
0
50
100 (in)

Figure 13. Inside electric field interference of the array after placing it on a vehicle body.

electromagnetic hazards.

5. Conclusion

A Tchebyscheff polynomial based microstrip phased array antenna is designed to detect the blind spot area of the intelligent mobile vehicle. The design is further simulated in Ansoft HFSS^{TM} after placing the antenna on the body of the car and the results are observed. The result of the computational model of the entire antenna array is compared with the measured data before and after placing the said antenna array on the body of the car and found to be in good agreement. An overall 11 dB gain is obtained while measuring in the desired direction with minimum amount of electromagnetic interference inside the car.

REFERENCES

[1] M. Griffiths, "Intelligent Transport System," Postnote: Parliamentary Office of Science and Technology, London, 2009. www.parliament.uk/paliamentary_offices/post/

[2] J. Kuwana and M. Itoh, "Dynamic Angling Side-View Mirror for Supporting Recognition of a Vehicle in the Blind Spot," *International Conference on Control, Automation and Systems, ICCAS*, Seoul, 14-17 October 2008, pp. 2913-2918.

[3] S. Jeong, S.-W. Ban and M. H. Lee, "Autonomous Detector Using Saliency Map Model and Modified Mean-Shift Tracking for a Blind Spot Monitor in a Car," *Seventh International Conference on Machine Learning and Applications*, San Diego, 11-13 December 2008, pp. 253-258. doi:10.1109/ICMLA.2008.129

[4] B.-F. Wu, C.-C. Kao, Y.-F. Li and M.-Y. Tsai, "A Real-Time Embedded Blind Spot Safety Assistance System," *International Journal of Vehicular Technology*, Vol. 2012, 2012, pp. 1-15. doi:10.1155/2012/506235

[5] A. Tang and A. Yip, "Collision Avoidance Timing Analysis of DSRC-Based Vehicles," *Accident Analysis & Prevention*, Vol. 42, No. 1, 2010, pp. 182-195. doi:10.1016/j.aap.2009.07.019

[6] T. Mondal, J. S. Roy and S. R. Bhadra Chaudhuri, "Phased Array Antenna Design for Intelligent Transport Systems," *IEEE International Workshop on Antenna Technology: Small Antennas and Novel Metamaterial* (i*WAT*), Santa Monica, 2-4 March 2009, pp. 1-4.

[7] T. Mondal, R. Ghatak and S. R. B. Chaudhuri, "Design and Analysis of a 5.88GHz Microstrip Phased Array Antenna for Intelligent Transport Systems," *IEEE International Sumposium on Antennas and Propagation and CNC/USNC/URSI Radio Science Meeting*, Toronto, 11-17 July 2010, pp. 1-4.

[8] H. J. Visser, "Array and Phased Array Antenna Basics," John Wiley & Sons Ltd., Chichester, 2005. doi:10.1002/0470871199

[9] E. Levine, G. Malamud, S. Shtrikman and D. Treves, "A Study of Microstrip Array Antennas with the Feed Network," *IEEE Transactions on Antennas and Propagation*, Vol. 31, No. 4, 1989, pp. 426-434. doi:10.1109/8.24162

Copyright © 2012 SciRes.

Wireless Engineering and Technology, 2012, 3, 175-180
http://dx.doi.org/10.4236/wet.2012.33025 Published Online July 2012 (http://www.SciRP.org/journal/wet)

Plus Shape Slotted Fractal Antenna for Wireless Applications

S. Jagadeesha[1], R. M. Vani[2], P. V. Hunagund[3]

[1]S.D.M. Institute of Technology, Ujire Mangalore, India; [2]University Science Instrumentation Center, Gulbarga University, Gulbarga, India; [3]Department of PG Studies and Research in Applied Electronics, Gulbarga University, Gulbarga, India.
Email: Jagadeesh.sd69@gmail.com, vanirm12@rediffmail.com, prbhakar_hunagund@yahoomail.com

Received April 28th, 2012; revised May 26th, 2012; accepted June 3rd, 2012

ABSTRACT

Fractal antennas are characterized by space filling and self-similarity properties which results in considerable size reduction and multiband operation as compared to conventional microstrip antenna. This paper outlines a multiband antenna design based on fractal concepts. Fractal antennas show multiband behavior due to self-similarity in their structure. The plus shaped fractal antenna has been designed on a substrate of dielectric constant $€_r = 4.4$ and thickness 1.6 mm. The proposed antenna is characterized by a compact size and it is microstrip feed fractal patch of order 1/3. It is observed that the antenna is radiating at multiple resonant frequencies. The resonant frequency is reduced from 2.2 GHz to 900 MHz after I & II iterations respectively. Thus considerable size reduction of 81.77% & overall bandwidth of 12.92% are achieved. The proposed antenna is simulated using the method of moment based commercial software (IE3D) and it is found that simulated results are in good agreement with the experimental results.

Keywords: Fractal Antenna; Multi Frequency; Size Reduction; Wireless Application; Plus Shape Antenna; Slotted Antenna

1. Introduction

In the study of antennas, fractal antenna theory is a relatively new area. The emergence of antennas with fractal geometries has given an answer to two of the main limitations started by Werner (1999) of the classical antennas, which are single band performance and dependence between size and operating frequency. The term "fractal" means broken or irregular fragments. It was originally coined by Mandelbrot (1983) to describe a family of complex shapes that possess an inherent self-similarity or self-affinity in their geometrical structure. Jaggered (1990) defined fractal electrodynamics as an area in which fractal geometry was combined with electromagnetic theory for the purpose of investigating a new class of radiation, propagation and scattering problems. One of the most promising area fractal electrodynamics researches is in its application to antenna theory and design. There are varieties of approaches that have been developed over the years, which can be utilized to archive one or more of these design objectives. The development of fractal geometry came largely from an in depth study of the pattern nature, with the advance of wireless communication system and their increasing importance wide band and low profile antennas are in great demand for both commercial and military applications [1]. A fractal is a rough or fragmented geometric shape that can be split into parts, each of which is a reduced-size copy of the whole and this property is called self-similarity. Fractal [2] geometries are composite designs that repeat themselves or their statistical characteristics and are thus "self-similar" fractal geometry finds a variety of applications in engineering. Fractal geometry is space filling contours of regular or irregular shapes [3-6], and is super imposed of too much iteration and they describe the self-similar property of fractal geometry [7]. Fractals are a class of shapes which have not characteristic size. Each fractal is composed of multiple iterations of a single elementary shape the iteration can continue infinitely, thus forming a shape within a finite boundary but of infinite length or area. Fractal has the following features 1) It has a finite structure at arbitrarily small scales; 2) It is too irregular to be easily described in traditional Euclidean geometric; 3) It is self-similar; 4) Simple and recursive [8]. Modern telecommunication systems require the antenna with wider bandwidth and smaller dimension than conventionally possible. This has initiated antenna research in various directions, are of which is by using fractal shaped antenna elements. In recent years several fractal geometries have been introduced for antenna application with varying degree of success in improving antenna

Copyright © 2012 SciRes.

characteristics. Some of these geometries have been particularly useful in reducing the size of the antenna, while other designs aim at incorporating multiband characteristics. These are low profile antennas with moderate gain and can be made operative at multiple frequency bands and hence are multifunctional [9]. In our present work we focus on generation of multifrequency which yields increases the bandwidth and size reduction of antenna. A plus shape patch is taken as a base shape and in first iteration four other plus shape patches of the order of 1/3 of base shape are placed touching the base shape. Similarly second iterations are taken by further placing plus shaped patches at even reduced scales. It is found that as the iteration number and iteration factor increases, the resonance frequencies become lower than those of the zero iteration, which represents a conventional plus shape patch.

2. Design Consideration

The base shape of the plus shaped slotted fractal antenna is designed on a dielectric substrate having a relative dielectric constant €$_r$ = 4.4 and thickness 1.6 mm as shown in **Figure 1**. This is the reference antenna or base shape antenna. Further this base shape antenna is modified by inserting horizontal slots on both sides with respect to center of patch as shown in **Figure 2** and it is named as antenna1. The length of the slot Ls is varied on either side of the edge as 5 mm, 10 mm, 15 mm, 20 mm, 21.175 mm, 21.675 mm and the frequency variation has been studied. The optimum length obtained is Ls = 21.675 mm *i.e.*, the distance between slots q = 2 mm is considered for further design.

The first iteration patch is designed with four plus shapes of order (1/3) of base shape are placed touching the base shape as shown in **Figure 3** and it is named as antenna 2 and same procedure is repeated for second iteration. This antenna is as shown in **Figure 4** and it is named as antenna 3. For each iteration plus shapes of the order of $(1/3)^n$ of base shape are taken, where n is the number of iterations. The dimension of first iteration can be calculated as

e = (1/3) a & g = (1/3) c also f = (1/3) b & h = (1/3) d.
i = (1/3) e & k = (1/3) g also j = (1/3) f & L = (1/3) h

So with optimized design the dimensions obtained are a = 45.3 mm, b = 15.1 mm, c = 35.4 mm, d = 11.8 mm. The length of the slot is Ls = 21.675 mm and width of the slot Ws *i.e.* r = 2 mm. The dimension of the ground plan is 55 mm × 85 mm. A 50 ohm SMA connector is used to feed the antenna by using microstrip feed technique. Optimized microstripline with following dimension, m = 0.5 mm, n = 18.55 mm, o = 3.05 mm, p = 18.4 mm. The suitable feed location is obtained through optimization process by using the IE3D software. The fabricated photographic view of all proposed antennas is shown from **Figures 5(a)-(e)**.

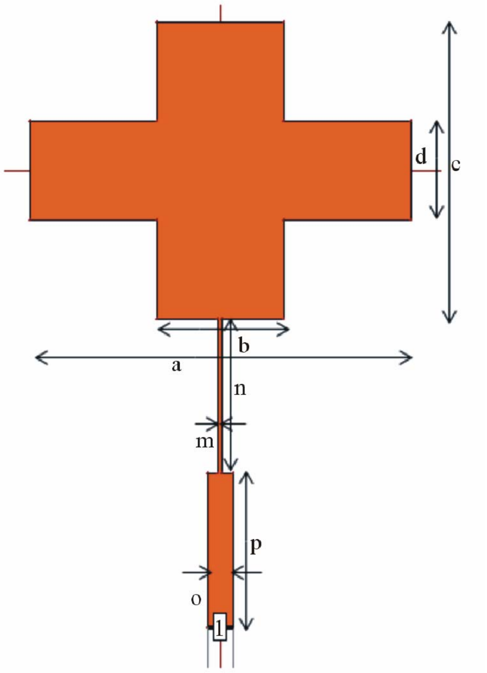

Figure 1. Geometry of base antenna.

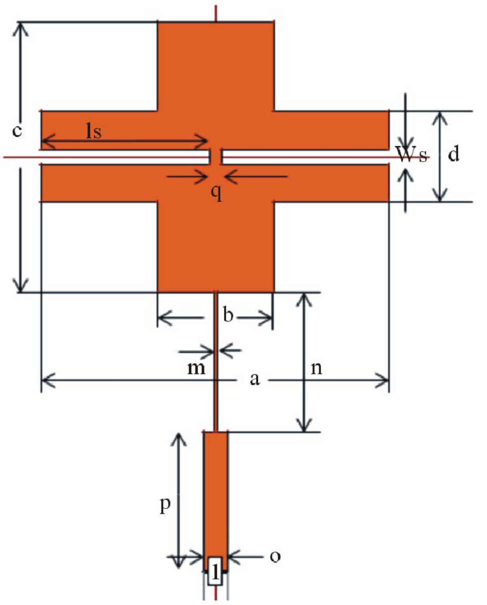

Figure 2. Geometry of antenna 1.

3. Results and Discussion

The characteristic of the fractal antenna with slot and with iterations has been studied by using IE3D software. Also the results have been verified practically with by using Vector Network Analyzer model Rohde and schewarz, German make ZVK model No.8651.

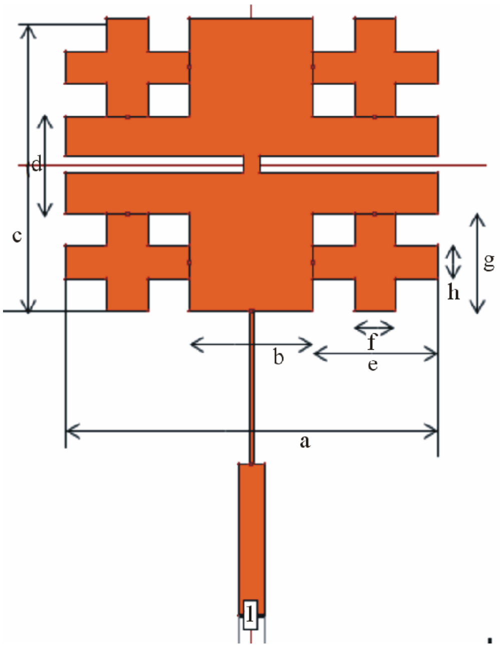

Figure 3. Geometry of base antenna 2.

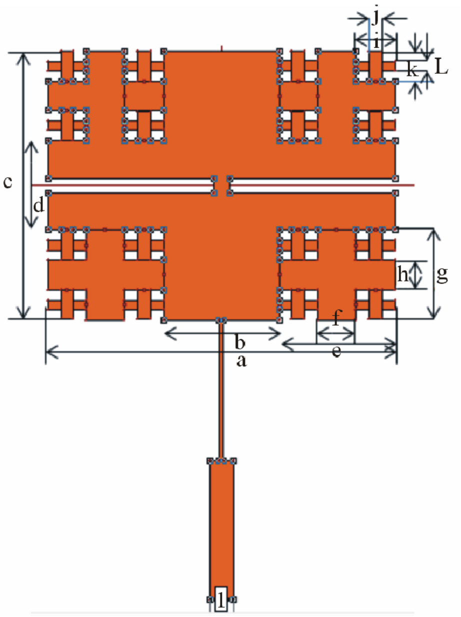

Figure 4. Geometry of antenna 3.

The modified base shape antenna with slots as been optimized by varying slot length Ls. The variation of Ls with resonant frequency of the antenna is shown in **Table 1** and same is presented in graphical form in **Figure 6**. From the tabular results it is found that by increasing slot length Ls on both sides from the edge of the patch, resonant frequency decreases. The lowest possible resonant frequency 1.27 GHz is obtained for Ls = 21.675 mm (*i.e.*, distance between the slots q = 2 mm) & this is taken as optimized length of the slot for further iteration.

The results of all proposed antennas are shown in **Table 2**. The simulated and measured return loss characteristics of proposed antennas are shown from **Figures 7(a)-(d)**.

From the results it is clear that the resonant frequency of the antenna 1 *i.e.*, modified base antenna with slot is fr = 1.27 GHz which is lower compared to the base antenna without slot (fr = 2.199 GHz). So the size reduction obtained is 66.85%. The antenna 2 *i.e.* the modified antenna with slots and first iteration gives multiple bands with lower frequency of 0.99 GHz. The size reduction obtained for antenna 2 is 79.88%. Further antenna 3 *i.e.*, with slot and second iteration gives multiple bands with

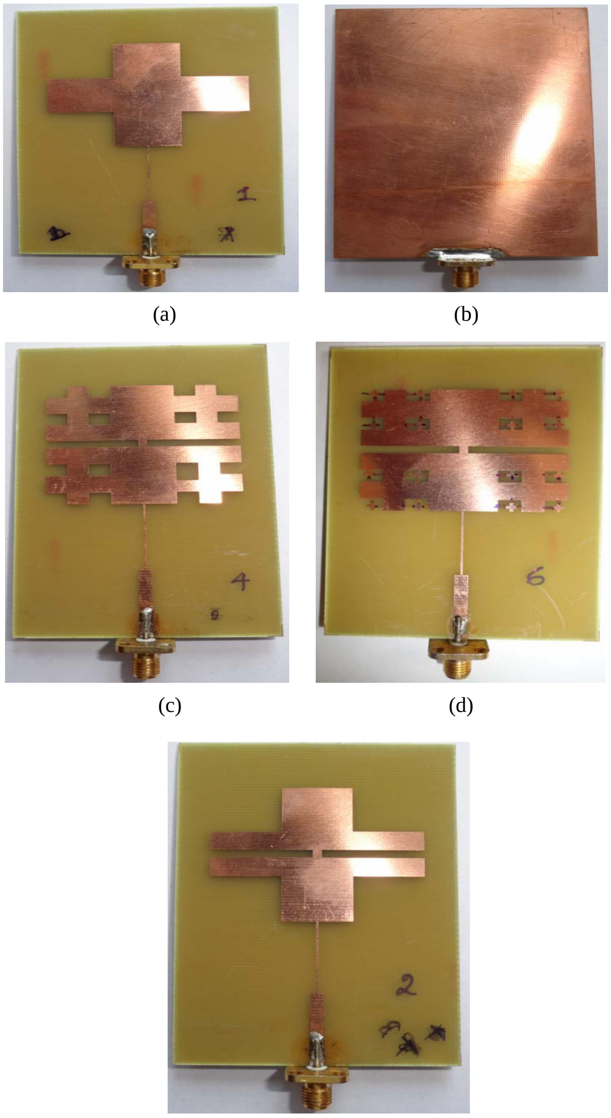

(a) (b)

(c) (d)

(e)

Figure 5. (a) Photograph of top view of base antenna; (b) Photograph of bottom view of base antenna; (c) Photograph of antenna 1; (d) Photograph of antenna 2; (e) Photograph of antenna 3.

Table 1. Variation of slot length v/s resonant frequency.

Slot length (Ls) variation in mm	Resonating frequency (GHz)
0	2.22
5	2.197
10	2.155
15	1.93
20	1.436
12.175	1.34
21.675	1.27

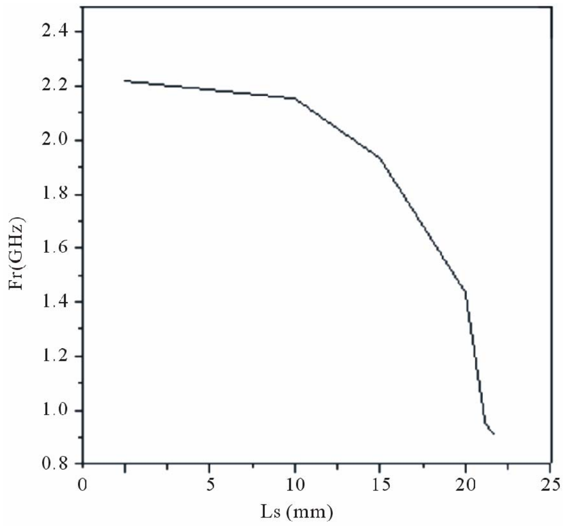

Figure 6. Variation of resonant frequency v/s slot length.

lower resonant frequency of 0.90 GHz. The size reduction obtained for antenna 3 is 81.77% which is more compared to all other proposed antennas.

Further the bandwidths of proposed antennas have been studied through simulation and measurements and the results are shown in **Table 2**. From the results it is clear that the bandwidths of modified antenna with slot, first & second iteration are more compared to base antenna. The measured bandwidth of antenna 1 is 106 MHz (4.895%), antenna 2 is 190 MHz (11.93%) and antenna 3 is 215 MHz (12.92%). The radiation patterns of all proposed antennas are studied and all are giving broadside radiation. Splitting of beam *i.e.* dip is found in **Figure 8(a)** of simulated radiation pattern of base shape at 2.19 GHz. In **Figure 8(b)** splitting of beams merged into broadside pattern at 0.91 GHz at second iterations this results into broadside radiation pattern.

4. Conclusion

This paper presents a new plus shape slotted fractal antenna with first and second iterations. The antenna.3 *i.e.*, slotted fractal antenna with second iteration gives size reduction of 81.77% and band width of 12.92% with broad side radiation pattern. So from the results we conclude that the modified base antenna with slots of second iterations gives a good size reduction and enhanced band width compared to that of modified base antenna with

Table 2. Results of proposed antennas.

Prototype antenna	Resonant frequency fr (GHz)		Return loss (db)		Bandwidth (MHz)		Overall bandwidth (MHz)	
	Sim	Pract	Sim	Pract	Sim	Pract	Sim	Pract
Base antenna	f1 – 2.19	f1 ± 2.19	–19	–24	53	53	79	71
	f2 – 3.42	f2 ± 2.46	–14	–15	26	18		
Antennal with q = 2 mm	f1 ± 1.19	f1 ± 1.27	15.8	–12.8	20	18	133	106
	f2 ± 2.52	f2 ± 2.48	18.4	–20.1	113	44		
		f3 ± 2.56		–18.7		44		
Antennal 2 with q = 2 mm (1st itr)	f1 ± 0.96	f1 ± 0.99	16.7	–14.5	22	80	136	190
	f2 ± 2.9	f2 ± 2.97	–14	–26.9	44	50		
	f3 ± 3.08	f3 ± 3.17	21.5	–21	70	60		
Antennal 3 with q = 2 mm (2nd itr)	f1 ± 0.91	f1 ± 0.990	15.2	–14.2	11	87	81	215
	f2 ± 3.07	f2 ± 2.976	19.4	–17.6	70	50		
		f3 ± 3.168		–18.9		78		

Copyright © 2012 SciRes.

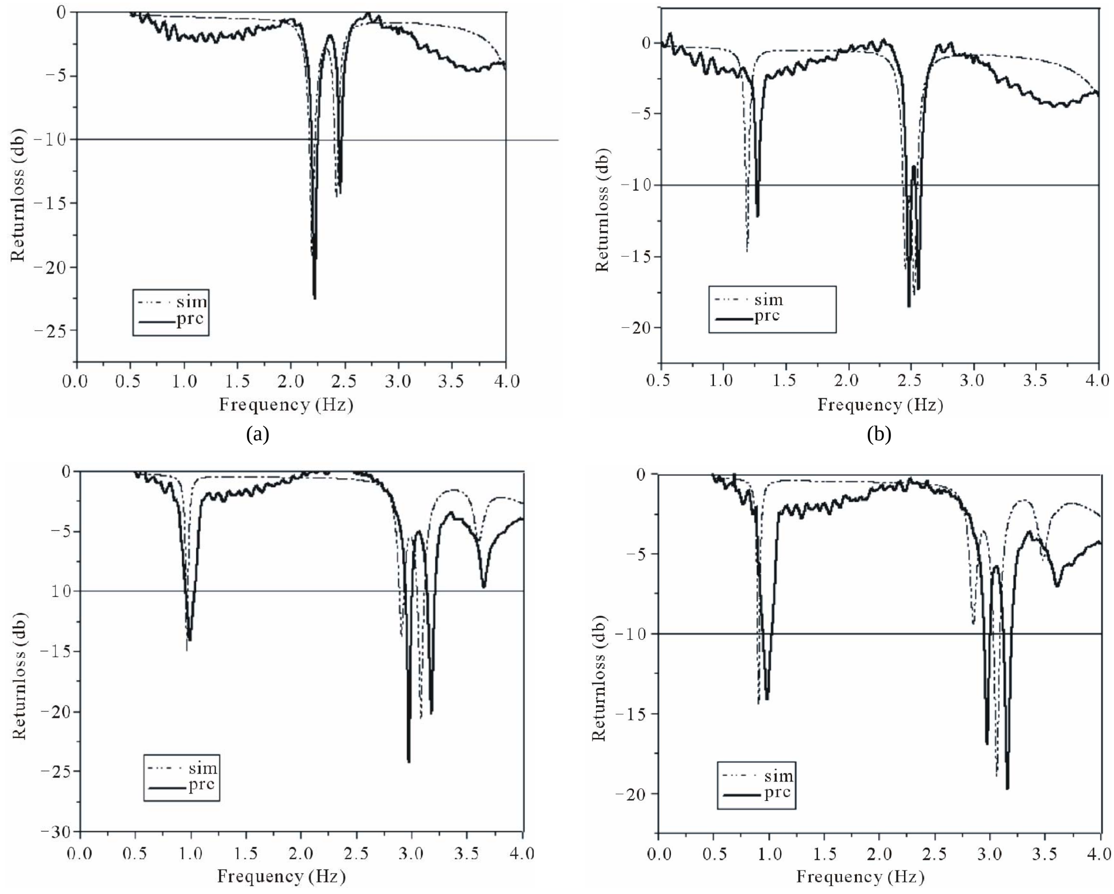

(a) (b)

(c) (d)

Figure 7. (a) Return loss characteristic of antenna 1 with q = 2 mm; (b) Return loss characteristic of base antenna 2 with q = 2 mm; (c) Return loss characteristic of antenna 2 with q = 2 mm; (d) Return loss characteristic of antenna 3 with q = 2 mm.

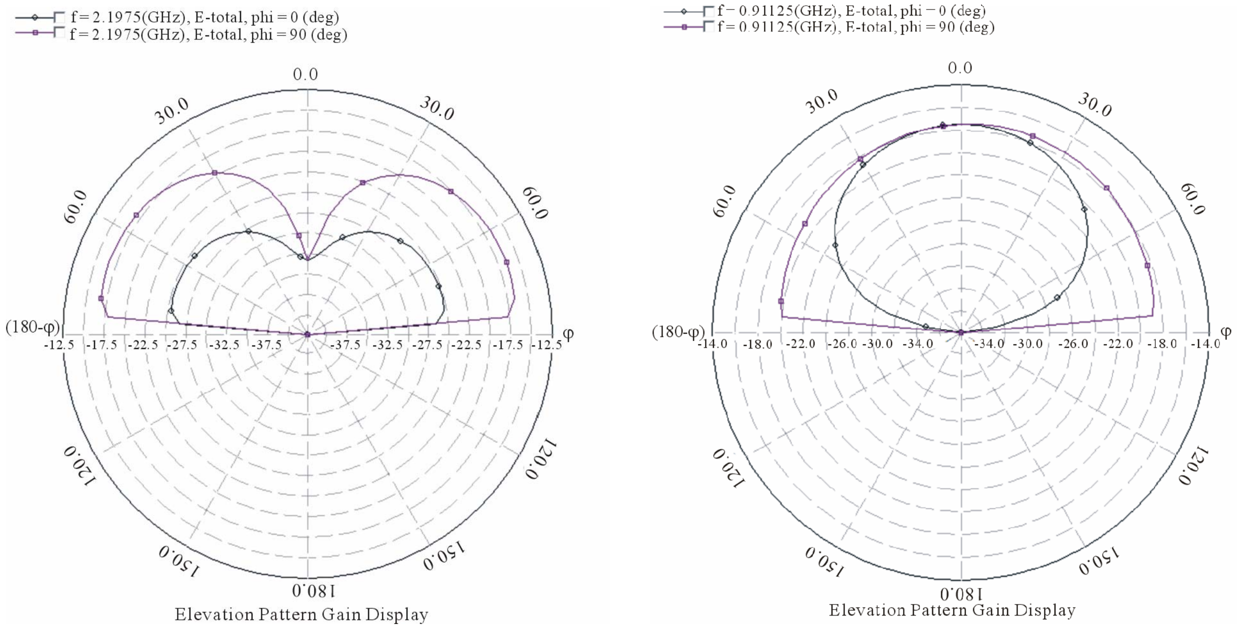

(a) (b)

Figure 8. (a) Simulated radiation pattern of base shape at 2.19 GHz; (b) Simulated radiation pattern at 0.91 GHz for second iteration.

slot of first iteration. These antennas may find application in wireless communication systems.

REFERENCES

[1] K. Sing, V. Grewal and R. Saxena, “Fractal Antennas: A Novel Miniaturization Technique for Wireless Communications,” *International Journal of Recent Trends in Engineering*, Vol. 2, No. 5, 2009, pp. 172-176.

[2] B. Mandelbrot, “Fractal: Form, Chance and Dimension,” W. H. Freeman and Company, San Francisco, 1997.

[3] H. O. Peitgzen, H. Jurgens and D. Sanpe, “Chaos and Fractals,” New Frontiers in Science, Springer-Verlag, New York, 1992.

[4] C. Punete, J. Romen, R. Pous, X. Garkia and F. Beitez, “Fractal Multiband Antenna Based on the Serpinski Gasket,” *Electronics Letters*, Vol. 32, No. 1, 1996, pp. 1-2. doi:10.1049/el:19960033

[5] J. P. Gianvittorio and Y. Rahmat-Sami, “Fractal Antennas: Novel Antenna: A Novel Antenna Miniaturization Technique and Applications,” *IEEE Antenna and Propagation Magazine*, Vol. 44, No. 1, 2002, pp. 20-36. doi:10.1109/74.997888

[6] D. H. Werner and S. Ganguly “An Overview of Fractal Antenna Engineering Research,” *IEEE Antenna and Propagation Magazine*, Vol. 45, No. 1, 2003, pp. 38-57.

[7] G. Srivatsun, S. S. Rani and G. S. Krishnan, “A Self-Similar Fractal Cantor Antennas for MICS Based Wireless Applications,” *Wireless Engineering Technology*, Vol. 2, No. 2, 2011, pp. 107-111. doi:10.4236/wet.2011.22015

[8] W. Shalan and K. Pahwa, “Multiband Microstrip Rectangular Fractal Antenna for Wireless Applications,” *International Journal of Electronics Engineering*, Vol. 3, No. 1, 2011, pp. 103-106.

[9] F. J. Jibrael and M. H. Hammed, “A New Multiband Patch Microstrip Plusses Fractal Antenna for Wireless Applications,” *ARPN Journal of Engineering and Applied Sciences*, Vol. 5, No. 8, 2010, pp. 155-158.

Copyright © 2012 SciRes.

PUBLISHED BY

INTECH
open science | open minds

World's largest Science,
Technology & Medicine
Open Access book publisher

3,250+
OPEN ACCESS BOOKS

106,000+
INTERNATIONAL
AUTHORS AND EDITORS

112+ MILLION
DOWNLOADS

BOOKS
DELIVERED TO
151 COUNTRIES

AUTHORS AMONG
TOP 1%
MOST CITED SCIENTIST

12.2%
AUTHORS AND EDITORS
FROM TOP 500 UNIVERSITIES

CLARIVATE ANALYTICS
BOOK CITATION INDEX
INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the
Book Citation Index in Web of Science™
Core Collection (BKCI)

Chapter from the book *Advances in Satellite Communications*

Downloaded from: http://www.intechopen.com/books/advances-in-satellite-communications

Interested in publishing with InTechOpen?
Contact us at book.department@intechopen.com

7

New Antenna Array Architectures for Satellite Communications

Miguel A. Salas Natera et al.*
Universidad Politécnica de Madrid,
Spain

1. Introduction

Ground stations which integrate the control segment of a satellite mission have as a common feature, the use of large reflector antennas for space communication. Apart from many advantages, large dishes pose a number of impairments regarding their mechanical complexity, low flexibility, and high operation and maintenance costs. hus, reflector antennas are expensive and require the installation of a complex mechanical system to track only one satellite at the same time reducing the efficiency of the segment (Torre et al., 2006).

With the increase of new satellite launches, as well as new satellites and constellation of low earth orbit (LEO), medium earth orbit (MEO), and geostationary earth orbit (GEO), the data download capacity will be saturated for some satellite communication systems and applications. Thus, the feasibility of other antenna technologies must be evaluated to improve the performance of traditional earth stations to serve as the gateway for satellite tracking, telemetry and command (TT&C) operation, payload and payload message or data routing (Tomasic et al., 2002). One alternative is the use of antenna arrays with smaller radiating elements combined with signal processing and beamforming (Godara, 1997).

Main advantages of antenna arrays over large reflectors are the higher flexibility, lower production and maintenance cost, modularity and a more efficient use of the spectrum. Moreover, multi-mission stations can be designed to track different satellites simultaneously by dividing the array in sub-arrays with simultaneous beamforming processes. However, some issues must be considered during the design and implementation of a ground station antenna array: first of all, the architecture (geometry, number of antenna elements) and the beamforming process (optimization criteria, algorithm) must be selected according to the specifications of the system: gain requirements, interference cancellation capabilities, reference signal, complexity, etc. During implementation, deviations will appear as compared to the design due to the manufacturing process: sensor location deviation and sensor gain and phase errors (Martínez & Salas, 2010). In an antenna array, the computation of a close approach of the direction of arrival (DoA) and the correct performance of the beamformer depends on the calibration procedure implemented.

---

* Andrés García-Aguilar, Jonathan Mora-Cuevas, José-Manuel Fernández González, Pablo Padilla de la Torre, Javier García-Gasco Trujillo, Ramón Martínez Rodríguez-Osorio, Manuel Sierra Pérez, Leandro de Haro Ariet and Manuel Sierra Castañer.
Universidad Politécnica de Madrid, Spain

This chapter is organized with the following sections. Section 2, introduces the relationship between applications and antenna design architectures. Section 3, introduces the new antenna array architectures for satellite communication including motivation and explains experimental examples. Section 4, explains adaptive antenna array and receiver architectures for adaptive antennas systems considering the beamforming with synchronization algorithms. Finally, Section 5 explains the A3TB concept.

2. Applications and antenna design architectures

In recent effort, new antenna array architectures have been under analysis and development. In (Tomasic et al., 2002) a highly effective, multi-function, low cost spherical phased array antenna design that provides hemispherical coverage is analyzed. This kind of novel architecture design, as the geodesic dome phased array antenna (GDPAA) presented in (Tomasic et al., 2002) preserves all the advantages of spherical phased array antennas while the fabrication is based on well-developed, easily manufacturable, and affordable planar array technology (Liu et al., 2006; Tomasic, 1998). This antenna architecture consists of a number of planar phased sub-arrays arranged in an icosahedral geodesic dome configuration.

In contrast to the about 10 *m* diameters dome of the GDPAA, there is the geodesic dome array (GEODA) (Sierra et al., 2007) with 5 *m* diameters dome. This antenna, presented in Fig. 1, has two geometrical structure parts. The first one, is based on a cylinder conformed by 30 triangular planar active arrays, and the second is a half dodecahedron geodesic dome conformed by 30 triangular planar active arrays. The GEODA is specified in a first version for satellite tracking at 1.7 GHz, including multi-mission and multi-beam scenarios (Martínez & Salas, 2010). Subsequently, the system of the GEODA has been upgraded also for transmission (Arias et al., 2010).

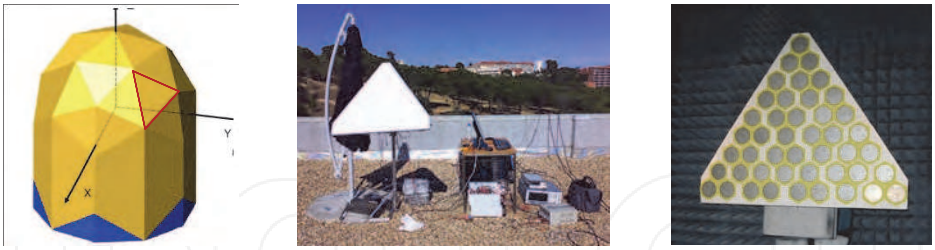

a b c

Fig. 1. a) The GEODA, b) The active sub-array demonstration, and c) The 45 elements planar active sub-array.

The antenna arrays technology in the user segment for satellite communications will substitute reflectors providing a more compact and easy to install antenna system, which is an interesting solution e.g. for satellite on the move (SOTM) system. There is a great diversity of solutions for fixed and mobile satellite communication systems including a large number of applications. Inmarsat broadband global area network (Inmarsat-BGAN) (Franchi et al., 2000) is the most representative example among mobile satellite systems (MSS), which gives land, maritime and aeronautical high speed voice and data services with global coverage using GEO satellites at L-band.

MSS services are divided into two groups, those that offer a regional coverage usually with GEO satellites, and those which offer a global coverage based on LEO or MEO satellite constellations. Depending on the coverage, there are some examples for MSS with regional coverage as the mobile satellite system (MSAT) in EEUU, Canada and South America, Optus in Australia, N-Star in Japan, Asia cellular satellite (ACeS) in Asia or Thuraya in the Middle East and in the North of Africa. While for MSS of global coverage there are some examples as Iridium, ICO Global Communications, Globalstar, Teledesic, etc. (Evans, 2009; Wu, 1994). Most of the MSSs work at L and S band, new applications on satellite to mobile terminal links work at X, Ku and Ka band, and satellite to base station connections work at L, S and C band. A number of applications is broad and lead terrestrial telecommunications market to offer a wider coverage: high speed voice and data (internet access, SMS, VoIP), digital video broadcasting by satellite 2 (DVB-S2) and digital video broadcasting satellite services to handhelds (DVB-SH), global position system (GPS) and Galileo, security, control and machinery monitoring on ships and aircrafts, teleeducation or telemedicine.

These modern satellite communications systems require new antenna solutions for base stations, aeronautical applications or personal communications services (PCS) on-the-move (Fujimoto & James, 2001). Within these applications, antenna array systems are potentially the best choice due to, as discussed above, its capability to perform electronically steering or beamforming, increase the antenna gain, and conform over curved or multifaceted surfaces the radiating elements. Portable antennas for PCS must be easy to install and mechanically robust, besides compact and lightweight (García et al., 2010) as the antenna array presented in Fig. 4.a. The design of antenna systems to provide high data rates for reliable PCS boarded on ships is not so strict in term of the geometrical requirements because it does not have space limitations (Geissler et al., 2010). However, in the case of land or airborne vehicles, geometrical and mechanical constraints are more severe. Antennas for terrestrial vehicles must be low profile, and for airborne vehicles aerodynamic shapes must be considered (Baggen et al., 2007; Vaccaro et al., 2010). Moreover, for the civil market conformal antenna arrays (Schippers, 2008; Kanno et al., 1996), or multi-surface arrays (Khalifa & Vaughan, 2007) are suitable choices to deal with the system aesthetic partiality.

Technological challenges have been faced during the implementation of satellite communication systems in the last decades. The design of a Test-Bed flexible and modular for testing or debugging beamforming algorithms and receiver architectures is an invaluable contribution in the educational, research and development area on satellite communication systems. The adaptive antenna array Test-Bed (A3TB) concept is based on the use of antenna arrays with beamforming capability to receive signals from LEO satellites (Salas et al., 2008). The scope of the A3TB is to probe the concept of antenna arrays applied to ground stations instead of reflectors for different applications, such as telemetry data downloading. It is also a good chance for Universities and Research Centers aiming to have their own ground station sited in their installations.

The A3TB ground station relies on the use of an antenna array to smartly combine the received signals from the satellite thanks to the implementation based on software defined radio (SDR) technology. The advantages of the SDR implementation is that A3TB architecture can be used to process any received signal from LEO satellites in the band imposed by the radio frequency (RF) circuits. Moreover, most of the processing is performed in software, so that appropriate routines can be used to process any received signal. The A3TB can be used to analyze the feasibility of different receivers and beamformer

algorithms, regarding the capability to switch the receiver architecture in terms of the synchronizer algorithm configuration (Salas et al., 2007).
The current version has been developed to track The National Oceanic and Atmospheric Administration (NOAA) satellites in the very high frequency (VHF) band, in particular, the automated picture transmission (APT) channel (Salas et al., 2008). Previous versions of A3TB dealt with low rate picture transmission (LRPT) signals from the meteorological operational satellite-A (MetOp-A), where a complete receiver with beamforming and synchronization stages has been implemented (Salas et al., 2007; Martínez et al., 2007).

3. Antenna arrays for satellite communications

Satellite applications require compactness, lightweight and low cost antenna systems to be mounted on a terrestrial vehicle, an aircraft or a ship, or as a portable man-pack or a handset, and to be competitive against ground systems. Its major advantage is the possibility of getting a wider or even a global coverage. For such purposes, antenna arrays offer the technology to get a directive system whose steering direction can be electronically and/or mechanically controlled. However, planar arrays usually cannot steer more than 60º-70º from the normal direction of the antenna (Mailloux, 2005). Thus, when a wider angular coverage is required conformal arrays are an appropriate option (Josefsson & Persson, 2006). Arrays can approximate conformal shapes, such as spheres or cylinders, using several planar arrays, simplifying fabrication of active components (Sierra et al., 2007).
Since the low cost and low weight specifications are of importance, micro-strip antennas are mostly used, due to its capacity to be printed over a dielectric substrate with photolithography techniques. Low cost and low permittivity substrates are usually used such as FR4 or PTFE with different quantities of glass or ceramic impurities. For more demanding applications, ceramics, like alumina or high/low temperature co-fired ceramics (HTCC/LTTC) allow the use of smaller components thanks to its high permittivity, and give robustness against mechanical stresses and high temperatures.

3.1 Geodesic antenna array for satellite tracking in ground station

The aim of using a single antenna for tracking many satellites at the same time avoiding mechanical movements as well as its inexpensive cost make these antennas an alternative to be considered (Salas et al., 2008). Multi-beam ability and interference rejection are facilitated thanks to the electronic control system of such antennas that improves the versatility of the ground stations.
The GEODA is a conformal adaptive antenna array designed for MetOp satellite communications with specifications shown in Table 1. This antenna was conceived to receive signals in single circular polarization (Montesinos et al., 2009). Subsequently, in recent efforts the system has been upgraded also for transmission and double circular polarization (Arias et al., 2010). Hence, operating at 1.7 GHz with double circular polarization it can communicate with several LEO satellites at once in Downlink and Uplink. Current structure is the result of a comprehensive study that valued the ability to cover a given spatial range considering conformal shape surface and a given beamwidth (Montesinos et al., 2009). As Fig. 1 shows, GEODA structure consists of a hemispherical dome placed on a cylinder of 1.5 meters height. Both cylinder and dome are conformed by 30 similar triangular planar arrays (panels). Each panel consists of 15 sub-arrays of 3 elements (cells). The radiating element consists of 2 stacked circular patches with their own

RF circuits. The principal patch is fed in quadrature in 2 points separated 90º in order to obtain circular polarization. The upper coupled patch is used in the aim of improving the bandwidth.

Each panel is able to work itself as an antenna since they have a complete receiver that drives the 1.7 GHz signal to an analog to digital converter (ADC). In order to adapt the signal power to the ADC, it is mandatory to implement a complete intermediate frequency (IF) receiver consisting of heterodyne receiver with an automatic gain control block. Hence, each triangular array has active pointing direction control and leads the signal to a digital receiver through an RF conversion and filtering process. To follow the signal from the satellite, the main beam direction has to be able to sweep an angle of 60º. In this way, it is needed a phase shift in the feeding currents of the single radiating element. Previous calculations have demonstrated that 6 steps of 60 degrees are needed to achieve the required sweeping angle. An adaptive digital system allows the adequate signal combination from several triangular antennas. The control system is explained in (Salas et al., 2010).

Parameter	Specification	Parameter	Specification
Frequency range [GHz] **Tx:** **Rx:**	1.65 to 1.75 1.65 to 1.75	**Isolation between Tx and Rx [dB]**	>20
Polarization	Dual circular for Tx and Rx bands	**VSWR**	1.2:1
G/T [dB/K] **For elevation >30º** **For elevation 5º**	3 6	**SLL [dB]**	-11
EIRP [dBW]	36	**Size [m]**	1.5x1.5x3
3dB beamwidth [deg.]	5	**Accuracy steering [deg.]**	±1.4
Maximum gain [dBi]	29	**Coverage [deg.]:** **Azimuth** **Elevation**	360º >5º
Efficiency [%]	50		

Table 1. Main specifications for GEODA antenna.

3.1.1 Cell radiation pattern

Based on the study presented in (Sierra et al., 2007), the single radiating element is a double stacked circular patch that works at 1.7 GHz with 100 MHz bandwidth. In order to obtain circular polarization, the lower patch, which has 90 mm diameter, is fed by 2 coaxial cables in quadrature. Both coaxial cables connect the patch with a hybrid coupler to transmit and receive signals with both, right and left, circular polarizations. The upper patch is a circular plate with 78.8 mm diameter, and it is coupled to the lower patch increasing the bandwidth by overlapping both resonant frequencies tuning the substrate thickness and the patch diameter size. Fig. 2.a shows the radiating element scheme and main features of the layer structure are specified in (Montesinos et al., 2009).

A cell sub-array of 3 radiating elements shown in Fig. 2.b is considered the basic module to build the planar triangular arrays. The whole cell fulfills radiation requirements since it has a good polar to crosspolar ratio and a very low axial ratio. Likewise, as it is presented in Fig. 2.c, the radiation pattern shows symmetry and low side lobes for full azimuth.

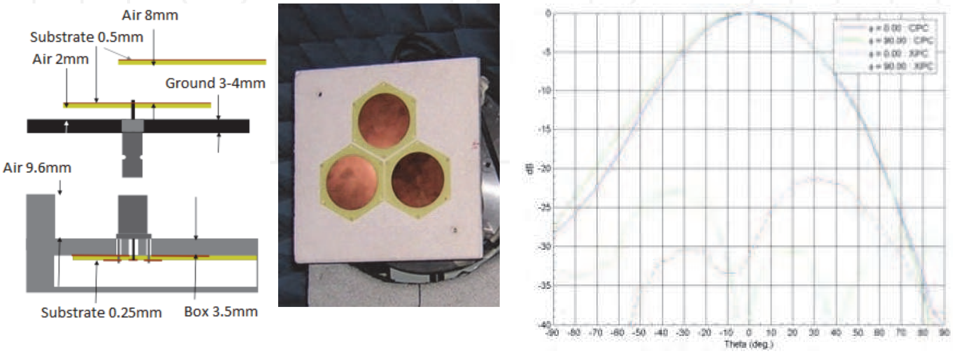

a b c

Fig. 2. a) Assembly of the single radiating element, b) Cell scheme, and c) Cell radiation pattern.

3.1.2 Transmission and Reception (T/R) module and cell distribution

Different T/R module configurations have been considered, providing either single or double polarization (Arias et al., 2010). T/R module allows amplifying and controlling the phase shift between signals, received and transmitted, providing an adaptive beam and steering direction controller in the whole working pointing range. As Fig. 3 shows, the design implemented contains a hybrid coupler, enabling double circular polarization; a double pole double throw (DPDT) switch, selecting polarization associated with transmission and reception way; 2 low noise amplifiers (LNAs), which amplify the signal received or transmitted; a single pole double throw (SPDT) switch, choosing transmission or reception way; and phase shifters, introducing multiples of 22.5^{o} relative shift phases to form the desired beam. These surface mount devices have been chosen in order to reduce space and simplify the design.

Signals transmitted/received by the 3 T/R modules placed in a cell are divided/combined thanks to a divider/combiner circuit composed of 3 hybrid couplers that leads the signal to a general T/R module where signal is amplified. Due to transmission and reception duality, 2 SPDT switches are used to select the amplification way. Furthermore, each T/R module has associated a -25dB directional coupler that is used to test T/R modules in the transmission mode. Additionally, reception mode is tested by measuring signal in the divider/combiner circuit. A single pole 6 throw (SP6T) switch selects the path that is tested.

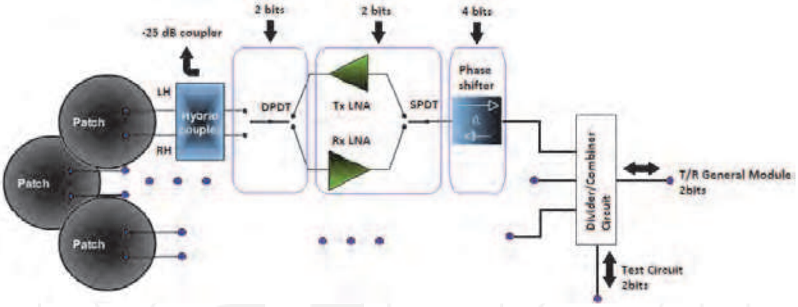
2 bits
2 bits
4 bits
-25 dB coupler
LH
RH
Hybrid coupler
DPDT
Tx LNA
Rx LNA
SPDT
Phase shifter
Patch
Patch
Patch
Divider/Combiner Circuit
T/R General Module 2bits
Test Circuit 2bits

Fig. 3. Cell sub-array and RF circuit.

3.1.3 Control system

The control system has two main parts (Salas et al., 2010), the hardware structure and the control software. The two level hardware structure has the lowest possible number of elements, making the control simpler in contrast to the previous in (Salas et al., 2010). Finally, an inter-integrated circuit (I2C) expander is used to govern T/R modules individually, and one more cover cell needs (LNA of call and test). A multipoint serial standard RS-485 is used to connect the computer with the panels.

3.2 Portable antenna for personal satellite services

New fix and mobile satellite systems (Evans, 2000) require antenna systems which can be portable, low profile and low weight. Planar antennas are perfect candidates to fulfill these specifications. Usually slots (Sierra-Castañer et al., 2005) and printed elements (García et al., 2010) are most used as radiating elements.

3.2.1 Antenna system structure

In this subsection it is introduced a printed antenna for personal satellite communications at X band, in Fig. 4. Table 2 shows main antenna characteristics.

Parameter	Specification	Parameter	Specification
Frequency range[GHz] **Tx:** **Rx:**	7.9 to 8.4 7.25 to 7.75	**Efficiency [%]**	50
Polarization	Dual circular polarization for Tx and Rx bands	**Isolation between Tx and Rx [dB]**	>17
G/T [dB/K]	7	**VSWR**	1.4:1
EIRP [dBW]	32	**SLL [dB]**	-11
3dB beamwidth [deg.]	5	**Size [m]**	40x40x2.5
Maximum gain [dBi]	25	**Weight [Kg]**	2

Table 2. Portable antenna specifications.

This is a planar, compact, modular, low loss and dual circular polarized antenna, for Tx and Rx bands, simultaneously. It is made up by a square planar array of 16x16 double stacked micro-strip patches, fed by two coaxial probes. A hybrid circuit allows the dual circular polarization (Garg et al., 2001). Elements are divided in 16 sub-arrays excited by a global power distribution network of very low losses, minimizing the losses due to the feeding network and maximizing the antenna efficiency. In order to reduce side lobe levels (SLL), the signal distribution decreases from the centre to the antenna edges, keeping symmetry with respect to the main antenna axes. The antenna works at X band from 7.25 up to 8.4 GHz with a 14.7% relative bandwidth for a 1.4:1 VSWR and a maximum gain of 25 dBi.

3.2.2 Sub-array configuration

The sub-array configuration can be seen in Fig. 4.a. It makes possible to separate the fabrication of these sub-arrays from the global distribution network, simplifying the corporative network and getting a modular structure suitable for a serial fabrication process. Each sub-array is a unique multilayer board, where PTFE-Glass substrate of very low losses has been used as base material. The power distribution network is connected to each sub-array through (SMP-type) coaxial connectors.

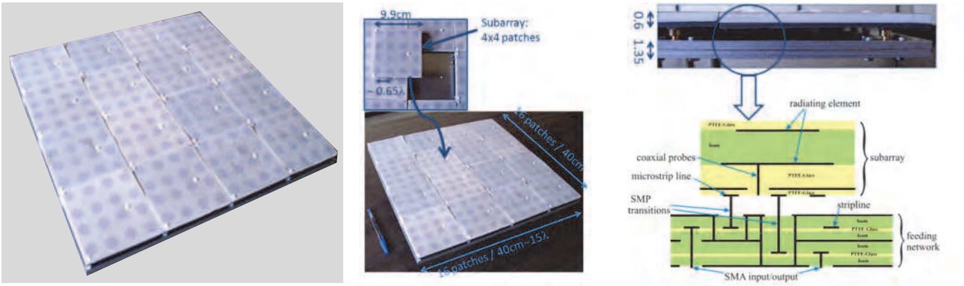

a b c

Fig. 4. a) Dual polarized portable printed antenna for satellite communication at X band, b) Sub-array perspective view, and c) Side view and multilayer scheme.

Fig. 5.a and Fig. 5.b show the sub-array unit cell. In order to obtain better polarization purity, each element is rotated 90° and excited by a 90° phase-shifted signal. Moreover, in Fig. 5.c is showed a miniaturized branch-line coupler (BLC) of three branches working as a wide band hybrid circuit (García et al., 2010; Tang & Chen, 2007).

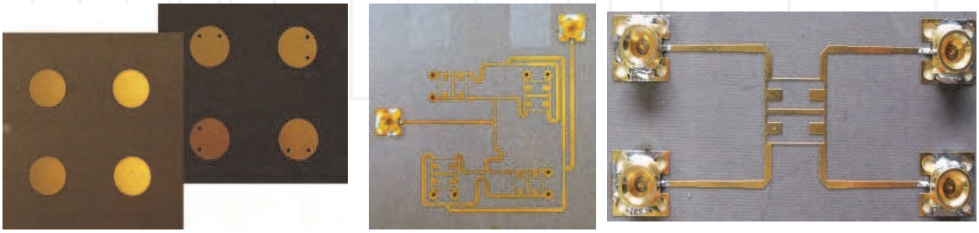

a b c

Fig. 5. Unit cell test board, a) Unit cell test board 2x2 stacked patches, b) Micro-strip feeding network, and c) Miniaturized BLC Prototype.

A conventional configuration takes up an area of 13.3 cm^2 which is big compared to the radiating element and the sub-array subsystem size. Therefore, a miniaturization of the BLC is needed using the equivalence between a $\lambda/4$ transmission line and a line with an open-ended shunt stub. An area reduction about 35% is achieved and the hybrid circuit behaves like a conventional BLC. In Fig. 6.b and Fig. 6.c measurement results for the BLC in Fig. 5.c are shown compared with simulations.

Fig. 7 depicts some sub-array measurements. The copular to crosspolar ratio is better than 25 dB and axial ratio is under 0.9 dB in the whole bandwidth.

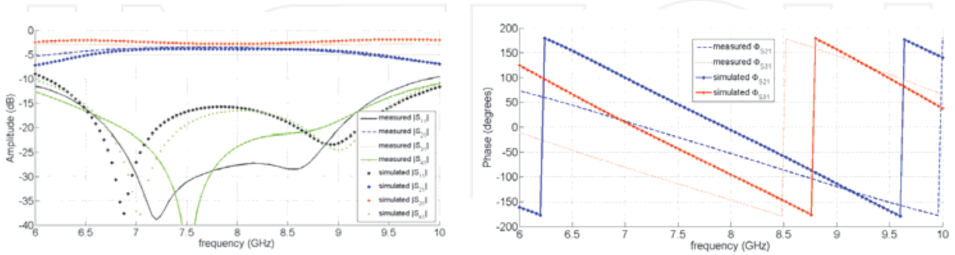

a b

Fig. 6. Miniaturized BLC, Measured and simulated S-parameters in: a) Amplitude, and b) Phase.

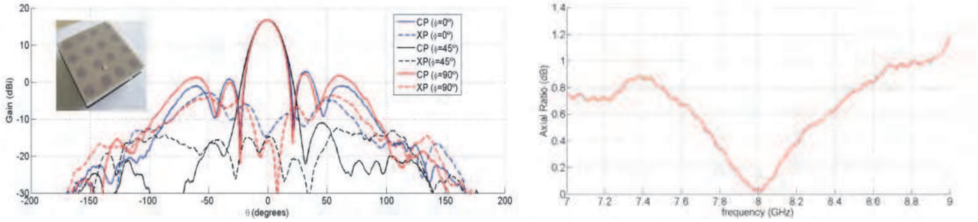

a b

Fig. 7. 4x4 patch sub-array measurements, a) Radiation pattern at 7.75 GHz, and c) Axial ratio for right-handed circular polarization.

3.2.3 Low losses power distribution network

The global feeding network presented in Fig. 8.a is a protected strip-line, where foam sheets of high thickness are used to get low losses. Such a kind of feeding network allows keeping a trade-off between the simplicity of exciting the radiating elements using printed circuits and the loss reduction when the distribution network is separated in a designed structure to have low losses. Losses in the structure are around 0.6 dB/m which yields to 0.3 dB of losses in the line. Two global inputs/outputs using SMA-type connectors, one for each polarization, excite the strip-line networks.

Vertical transitions have to be treated carefully and must be protected to avoid undesired higher order mode excitation. Thereby, it has been design a short-ended pseudo-waveguide, adding some extra losses about 0.3 dB, for two kinds of vertical transitions, as can be seen in Fig. 8.b and Fig. 8.c.

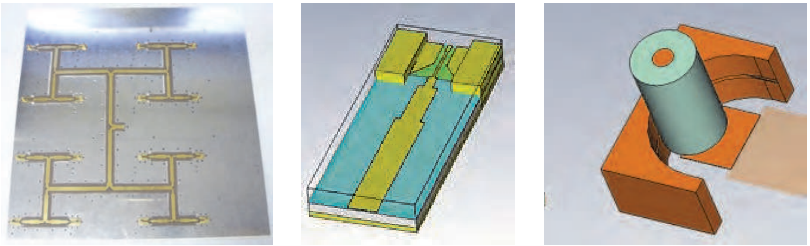

a b c

Fig. 8. a) Protected strip-line global corporative network for one polarization, b) Transitions from strip-line to SMA-type connector, and c) Transitions from strip-line to SMP-type connector.

3.2.4 Antenna performance

Fig. 9 depicts measured radiation pattern at 7.75 GHz, gain and axial ratio for the antenna system. It is shown a maximum gain of 25 dBi in the lower band and about 22 dBi in the upper band, and a SLL around 11 dB. Copolar to crosspolar ratio is better than 30 dB and axial ratio is under 0.7 dB. Total losses are about 4 dB in the working band.

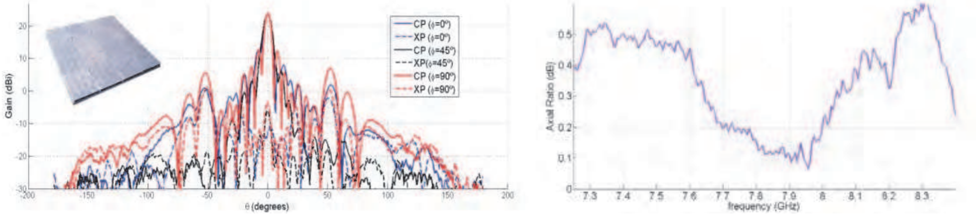

a b

Fig. 9. Antenna measurements results, a) Radiation pattern at 7.75 GHz, and c) Axial ratio for right-handed circular polarization.

3.3 Electronically steerable antennas for mobile and fixed portable systems

At present, two types of electric steerable antenna systems can be used to access the satellite communication services (Bialkwoski et al., 1996). These are: fixed position portable systems and mobile systems such as those installed on a land vehicle. The fixed portable antenna system is relatively easy to be accomplished by the antenna designer. The design involves standard procedures that concern the operational bandwidth, polarization and moderate gain (García et al., 2010). One drawback of the fixed position portable system is that they require the user to be stationary with respect to the ground. This inconvenience can be overcome with the mobile antenna system. A mobile user complicates the scenario since the ground mobile antenna needs to track the satellite (Alonso et al., 1996). The design of such a system is more challenging as new features associated with the mobility of the system have to be incorporated (Fernández et al., 2009). The requirement leads to a narrow beamwidth, for which satellite tracking is required as the vehicle moves around. Electronically steerable antennas enable the development of reconfigurable antennas for satellite applications.

3.3.1 Steerable antenna for fixed position portable systems

This antenna is a fixed satellite communication system with high gain at X band, consisting of an antenna array that integrates 32 2x2 sub-array modules in the complete antenna, as shown in Fig. 10.a. It is a planar and dual circular polarized antenna for Tx and Rx bands simultaneously. It is made up by a planar array of double stacked circular micro-strip patches, fed by 2 coaxial probes to generate circular polarization. A hybrid circuit allows the dual circular polarization as shown in Fig. 10.b.

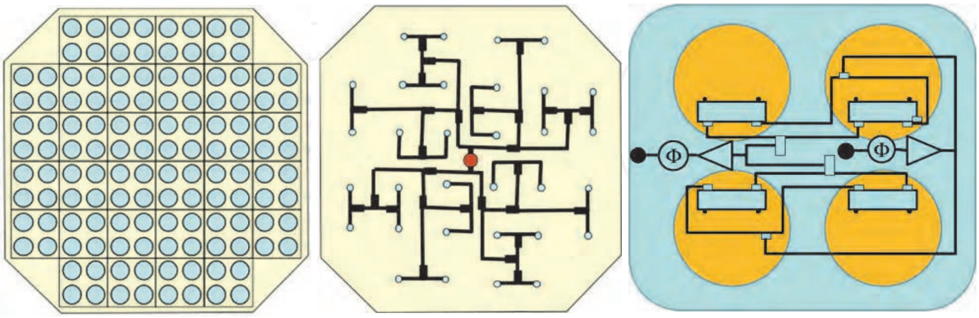

a b c

Fig. 10. Active multi-beam antenna, a) Top view, b) Feeding network of the complete antenna, and c) Beamforming network of the 2x2 sub-array module

The antenna has the same design parameters, structure and configuration as the antenna explained in Section 3.2 but with a different feeding network, as previously shown. In this case, the beamforming network requires changes in the feeding phase in the 2x2 sub-arrays, which can be achieved by phase shifters (ϕ) associated with different sub-arrays (Fig. 10.c). All these sub-arrays are connected to a feeding network, in Fig. 10.b, formed by transmission lines with low losses in strip-line. General specifications of the steerable antenna for fixed position portable systems are provided in Table 3.(a).

3.3.2 Automatic steerable antenna for mobile systems

A broadband circularly polarized antenna for satellite communication in X band is presented in Fig. 11 and specified in Table 3.(b). The arrangement features and compactness are required for highly integrated antenna arrays. It is desired to get a low-gain antenna for mobile satellite communications with low speed of transmission. In this system, the antennas are formed by 5 planar 4x4 arrays of antennas, which form a truncated pyramid with a pointing capability in a wide angular range, so that among the 5 planar arrays the complete antenna can cover any of the relative positions between the mobile system and the satellite in a practical way. The scheme of the active antenna can be seen in Fig. 11.

As it can be observed in Fig. 11.a, the antenna terminal is a multi-beam printed antenna shaped as a trunk pyramid capable of directing a main beam in the direction of the satellite. The antenna steering system consists of a multi-beam feeding structure with switches that lets combine the feed of each 4x4 arrays to form multiple beams. Switching the different 4x4 arrays, it is achieved different multiple beams and the variation of the steering direction.

The complete antenna consists of a Tx and Rx module that works independently in the 2 frequency bands.

The antenna has multiple beams covering the entire space to capture the satellite signal without moving the antenna. The signal detected in each of the beams is connected to a switch, which, by comparison, is chosen the most appropriate 4x4 array. The steering direction of the 4x4 array can vary between a range of directions that covers a cone angle range of 90º. To obtain the required gain and cover the indicated range, it is required around 15 beams, which can be obtained by integrating the beamforming networks with switches in the design as presented in (Fernández et al., 2009).

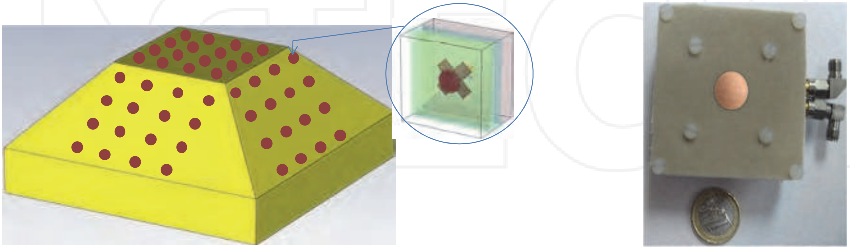
a b

Fig. 11. Complete antenna structure, a) Radiating element of the 4x4 arrays, and b) Prototype top view.

The radiating element of the 4x4 array is one 2 crossed dipoles with a stacked circular patch as shown in Fig. 11.a and Fig. 11.b. In Fig. 12 the cross-section of the radiating element structure is presented.

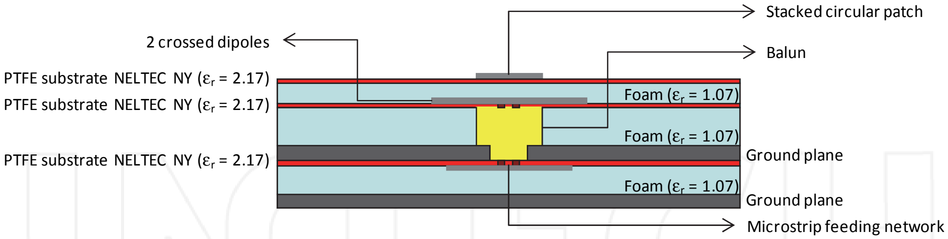
Stacked circular patch
2 crossed dipoles
Balun
PTFE substrate NELTEC NY ($\varepsilon_r = 2.17$)
PTFE substrate NELTEC NY ($\varepsilon_r = 2.17$)
Foam ($\varepsilon_r = 1.07$)
Foam ($\varepsilon_r = 1.07$)
PTFE substrate NELTEC NY ($\varepsilon_r = 2.17$)
Ground plane
Foam ($\varepsilon_r = 1.07$)
Ground plane
Microstrip feeding network

Fig. 12. Cross-section scheme of the radiating element.

The key element of the radiating element feeding structure (Fig. 14.b) is a resonant micro-strip feed ring that has been implemented, as well as a micro-strip 90º branch-line coupler to obtain the desired right hand or left hand circular polarizations (RHCP or LHCP) which ensures adequate port coupling isolation. The S-parameters in amplitude and phase of the micro-strip feeding structure are shown in Fig. 13.a and Fig. 13.b.

Fig. 14.a depicts the S-parameters of the radiating element with the micro-strip feed structure and they fulfill the specification, in Table 3.(b). In Fig. 14.c, the radiation pattern of the radiating element at 7.825 GHz is shown and in Fig. 14.d the radiation pattern of the 4x4 arrays is presented. It is shown a maximum gain of 19.4 dBi at the center frequency band (7.825 GHz). Copolar (CP) to crosspolar (XP) ratio is better than 17 dB and the axial ratio is under -3dB.

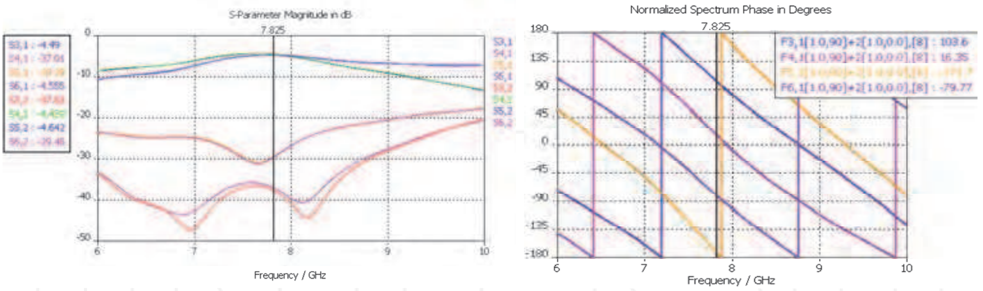

a b

Fig. 13. Micro-strip feeding structure, a) Amplitude of S-parameters, and b) Phase of S-parameters.

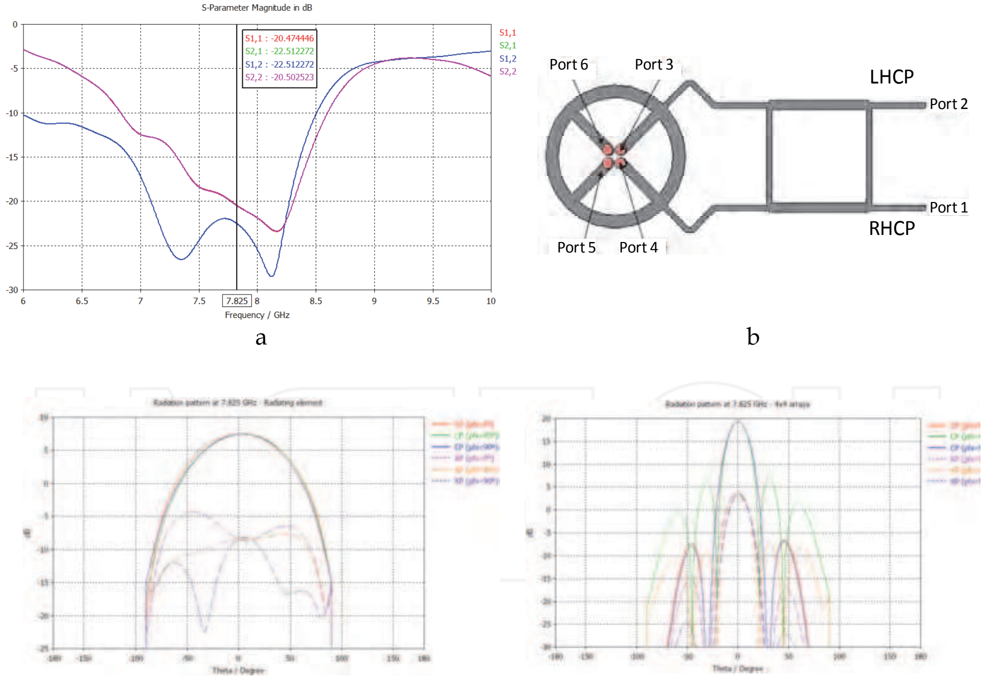

a b

c d

Fig. 14. a) S-parameters, b) Resonant ring + 90° branch-line coupler, c) radiation pattern at 7.825 GHz, and d) 4x4 array radiation pattern.

Parameter		Value (a)	Value (b)	Comments
Freq. range [GHz]	**Rx** **Tx**	7.25 - 7.75 7.9 - 8.4	7.25 - 7.75 7.9 - 8.4	Microwave applications.
G/T (in Rx) [dB/K]		7	7	
EIRP (in Tx) [dBW]		32	32	
Beamwidth at -3dB [deg.]		4	20	
Polarization		circular	circular	In both, reception and transmission.
Gain [dBi]		>28	>15	
Axial ratio [dB]		< 1	<3	**(a)** Between ±50°. **(b)** Between ±45°.
VSWR		< 1.4:1 (-15.6 dB)	< 1.5:1 (-13.9 dB)	
Isolation between ports [dB]		< -17	< -15	
Radiation pattern [deg.]		±35	±90	Steering direction tilt.
Dimensions [cm]		40x40x4	20x20x15	

Table 3. (a) General specifications of the steerable antenna for fixed position portable systems , and (b) General features of the automatic steerable antenna for mobile systems.

3.4 Transmit-array-type lens antenna for terrestrial and on board receivers

Technology in satellite communications has revealed an increasing interest in novel smart antenna designs. Phased-array based designs are basic in electronically reconfigurable devices for satellite applications, which are more and more demanding. The strict requirements in terms of architecture, shape and robustness are important constraints for the development of planar lens-type devices. Regarding the usage and location, lens-type devices are useful for either terrestrial or on board receivers, in vehicular technology. Some clear examples are satellite communications for aircrafts preserving the fuselage aerodynamics or for some other kind of vehicles such as trains, etc.

3.4.1 Introduction to lens-type structures

In a general view, in lens-type a particular signal is received (in our case, an electromagnetic wave with specific features in terms of frequency, wave-front, etc.), it is processed (either complex signal processing techniques or only phase correction tasks can be considered in this interface), and finally, the processed signal is retransmitted.

Regarding the lens configuration, a transmit-array lens consists of three well distinguished interfaces: the first one for signal reception, one interface for signal processing, and the last one for processed signal re-radiation, as depicted in Fig. 15.

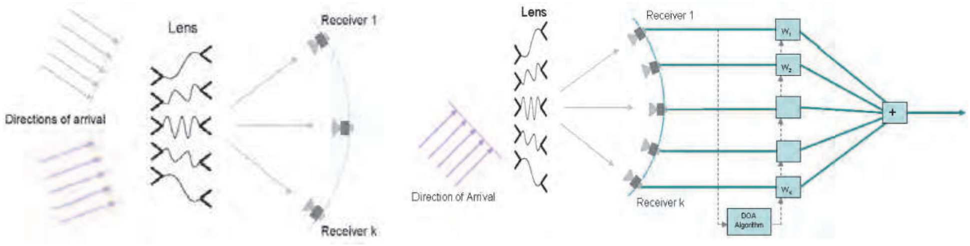

a b

Fig. 15. a) Multi-user scheme with different receivers and transmitters, and b) Adaptive scheme with DoA determination.

These structures are intimately related to reflect-array ones, where the reception and transmission interfaces are turned to be the same interface, with a reflection-type behavior (Encinar & Zornoza, 2001). Although in an equal output phase configuration a transmit-array device behavior would be similar to the one obtained with a reflect-array, the transmit-array offers the advantage of removing the feed blockage.

In a transmission scheme, depending on the transmitter position regarding the lens, a different steering direction is achieved and a different user is pointed. In the case of reception, the situation is the same: the user position configures the direction of arrival, which determines the receiver position around the lens (Padilla et al., 2010a). In adaptive schemes, applying the proper processing algorithm to the signal received in the different receivers around the lens, it is possible to develop an adaptive steering vector, in terms of the desired direction of arrival.

3.4.2 Transmit-array lens architecture and design

Lens-type structures provide two fundamental advantages. First, phase error correction due to spherical wave front coming from the feeding antenna. Fig. 16.a shows this effect. Second, new radiation patterns configuration. Fig. 16.b depicts this fact.

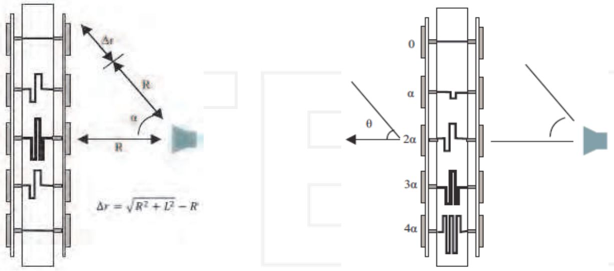

a b

Fig. 16. a) Phase error correction, and b) Radiation pattern reconfiguration.

3.4.3 Electronically reconfigurable devices for active transmit-array lenses

The addition of reconfigurability on transmit-array devices requires the possibility of controlling the phase response of the transmitted signal at each cell of the lens. Electronic control of phase signal may be added in two different ways: First, electronic tuning of the radiating element phase response (Padilla et al., 2010a): Modifications in the radiating element circuital behavior lead to changes in phase response (arg[S_{21}]). Fig. 17 shows an electronically reconfigurable microwave patch antenna for this purpose, along with the equivalent circuit and prototype outcomes in terms of phase.

Second, electronic tuning of phase shifters in transmission lines (Padilla et al., 2010c): Modifications in the phase response of the phase shifters lead to corresponding changes in phase response. Some options are applied for these devices, such as hybrid couplers, etc. Fig. 18 shows a microwave phase shifter prototype for this purpose, along with the working scheme and its outcomes in phase.

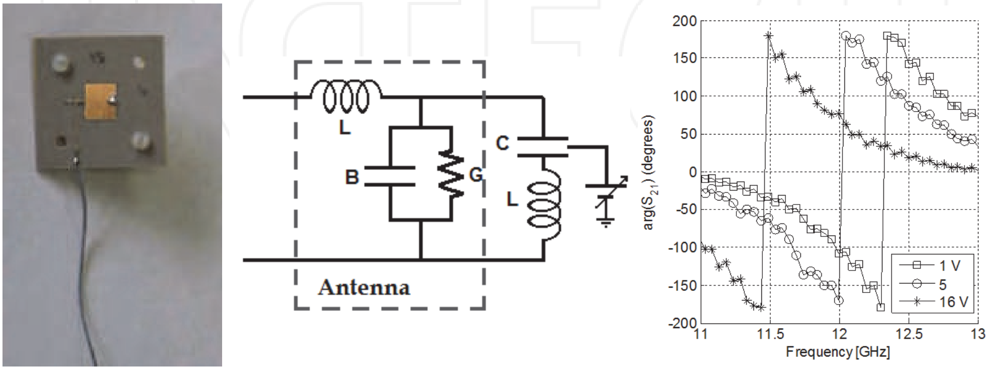

a b c

Fig. 17. Electronically reconfigurable antenna, a) Patch antenna prototypes, b) Equivalent circuit, and c) Phase behavior in frequency.

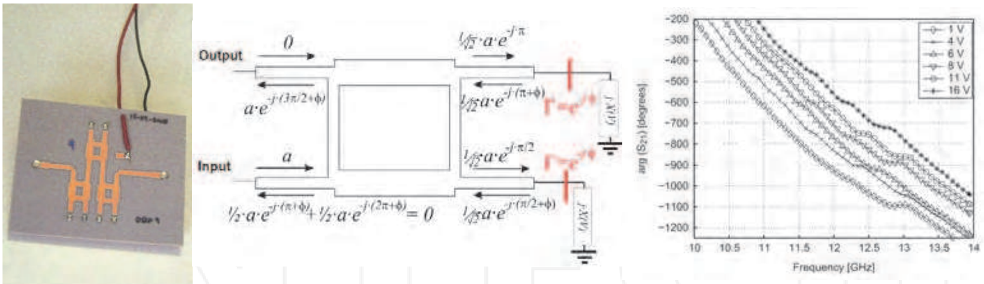

a b c

Fig. 18. Electronically reconfigurable phase shifter, a) Phase shifter prototype, b) Working scheme, and c) Phase behavior in frequency.

3.4.4 Electronically reconfigurable active transmit-array prototype

One electronically reconfigurable prototype is presented in Fig. 19 and detailed in this section. The prototype design implies the use of microwave phase shifters according to the design specified in section 3.4.3. This transmit-array lens prototype operates at 12 GHz. Main specifications are provided in Table 4.

Parameter	Value	Comments
Frequency range [GHz]	12 ± 0.5	Microwave applications.
Polarization	Linear	In both, reception and transmission.
Directivity [dBi]	>21	
Axial ratio [dB]	< 1	Between ±50° elevation.
S_{11} [dB]	< -20	
Radiation pattern [deg.]	±30	Steering direction tilt, for both H and V planes.
Feeding antenna [mm]	120	Corrugated horn linearly polarized
Phase shifters [deg.]	360	Full phase range variation.
Transmit-array elements	36	6x6 array topology.
Separation between elements	$0.7\lambda_0$	Related to the wavelength

Table 4. Main features of the electronically reconfigurable transmit-array prototype.

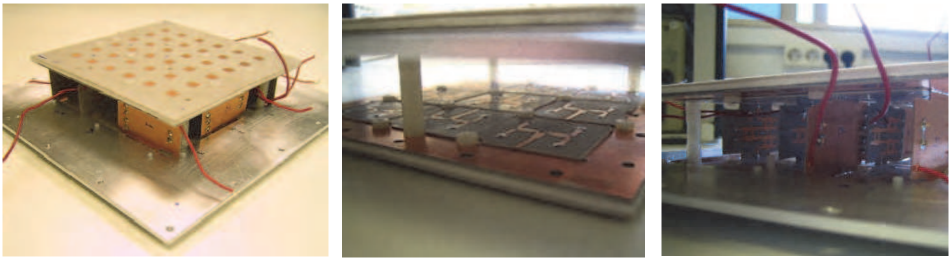
a b c

Fig. 19. Transmit-array core, a) Transmit-array prototype, b) Distribution networks, and c) Phase shifter integration.

The electronically controllable steering capabilities are tested and assured for a range of ± 30°in each main axis. An example of radiation pattern is provided in Fig. 20, for 9° tilt in one of the main axes.

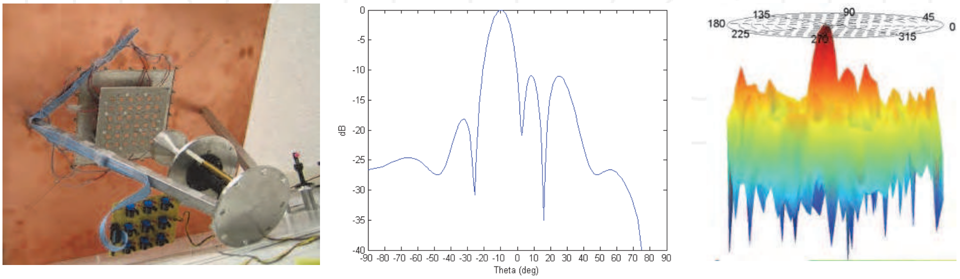

a b c

Fig. 20. a) Complete transmit-array with feeder and control circuits; and transmit-array measurement results for 9° tilt in one axis, b) H plane, and c) 3D plot.

4. Adaptive antenna array

Adaptive antennas can be described as systems usually based on three main parts: the antenna array, the receiver architecture and the beamforming scheme. Thus, adaptive antennas have those advantages owing to those three main parts. The system capabilities increase as complexity and development cost do. Furthermore, since signal processing is the basement of the adaptive antenna concept it is important to analyze the design challenges in terms of hardware architecture and components such as processors and embedded systems.

The antenna array provides the capability of performing the antenna pattern meeting the environment requirement under study. Besides, receiver architectures have some interesting advantages depending on the implemented receiver arraying technique such as signal to noise ratio (SNR) and bit error rate (BER) performance enhancement. Furthermore, symbol synchronization and carrier recovery can be used increasing the receiver complexity but providing higher performances. Finally, beamforming schemes use multiple antennas in order to maximize the strength of the signals being sent and received while eliminating, or at least reducing, interference as discussed in Section 4.3.

Adaptive antenna arrays are often called Smart Antennas because they have some key benefits over traditional antennas, by adjusting traffic patterns, space diversity or using multiple access techniques. The main four key benefits are: First, enhanced coverage through range extension by increasing the gain and steering capability of the ground station antenna; Second, enhanced signal quality through multi-target capability and reduction of interferences; finally, adaptive antennas improve the data download capacity in the ground segment of satellite communication by increasing the coverage range (Martínez et al., 2007).

4.1 Design and architecture based on software defined radio

For design there is the well known waterfall life cyclic model (Royce, 1970) that can be used to manage main aspects of the design of architectures. Thus, some tasks must be fulfilled subsequently as follow in Fig. 21.a.

Fig. 21.b shows the design schemes resulting of the requirement analysis stage corresponding software and hardware system specifications. In the depicted scheme, there are some system components such as the radiating element and RF circuits that are often designed under iterative prototyping model.

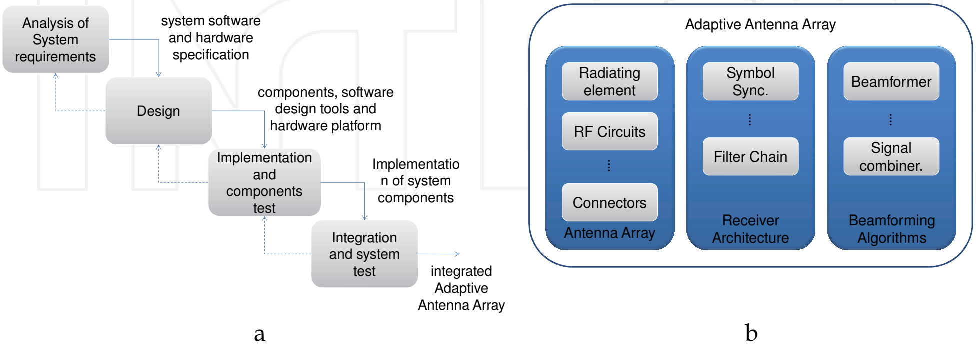
Analysis of System requirements
system software and hardware specification
Design
components, software design tools and hardware platform
Implementation and components test
Implementation of system components
Integration and system test
integrated Adaptive Antenna Array
a
Adaptive Antenna Array
Radiating element
RF Circuits
Connectors
Antenna Array
Symbol Sync.
Filter Chain
Receiver Architecture
Beamformer
Signal combiner.
Beamforming Algorithms
b

Fig. 21. a) Water life cyclic model of the adaptive antenna array design, and b) Simplified design scheme of adaptive antenna arrays.

Regarding the hardware implementation, tables presented in (Martínez et al., 2007) show the hardware resource consumption in the field programmable gate array (FPGA) Virtex-4 for the least mean squared (LMS) beamforming algorithm with full spectrum combining (FSC) receiver architecture and SIMPLE beamforming algorithm with symbol combining (SC) receiver architecture. Both scheme designs have an antenna array of 2 elements. The algorithm based on correlation requires less hardware. The main difference can be appreciated in the amount of digital signal processing oriented component (DSP48) resources, typically used for filtering applications (Martínez et al., 2007).

4.2 Receiver architectures based on algorithms type

Several receiver architectures can be implemented, and they are frequently based on the type of the beamforming algorithm used. When training signals are available in the transmitted frame, a time-based reference algorithm can be used. However, this solution is only valid when the earth station is capable of demodulating the received training sequence. Other algorithms used in deep space communications are based on signal correlation and they avoid performing the demodulating process. This kind of algorithms are blind techniques that do not require any additional signal demodulation before applying some beamforming technique and work better in low SNR conditions than time-based algorithms.
Several receiver architectures can be implemented exploiting the processing capabilities of the SDR, such as FPGA, application-specific integrated circuits (ASICS), and digital signal processing (DSPs). The design of the receiver architecture fundamentally depends on the selection of beamforming algorithms. An example of beamforming technique is the LMS algorithm whose estimation of coefficients or weights requires a temporal reference and is implemented through SC receiver architecture (Fig. 22.a). In the other hand, the SIMPLE algorithm (Rogstad, 1997) constitutes a beamforming technique that is implemented using FSC receiver architecture (Fig. 22.b) in order to perform the calculation of weights.

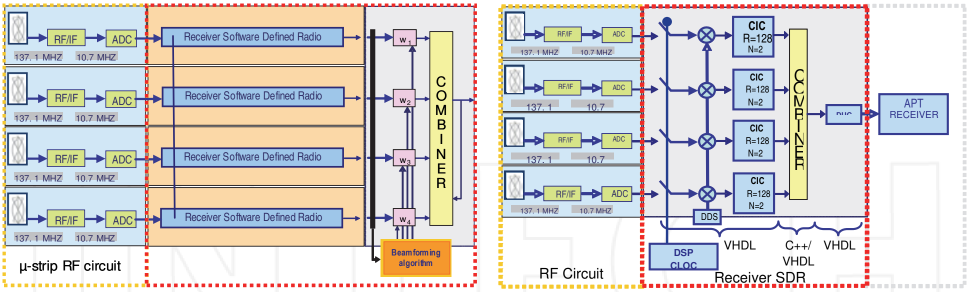

a b

Fig. 22. Comparison of receiver architectures. a) Symbol Combining (SC), and b) Full Spectrum Combining (FSC).

The SC architecture can be divided into two more sub-classes which work on a phase-recovery basis. The complex symbol combining (CSC) recovers the phase information with regard to a reference element using feed-forward and feedback algorithms. One of the advantages of this scheme is that the rate of data sent to the combining module has a rate slightly higher than the symbol rate. For most applications, the symbol rate is relatively low and is a multiple of the data rate. In this kind of schemes, there is an important cost consideration in real-time applications and the requirements of instrumental phase stability are very severe (Rogstad et al., 2003). Other type of SC architecture is the stream symbol combining (SSC). In this kind of scheme, data are sent to the combining module at a rate equal to the symbol rate. The symbol rate depends on the coding scheme and for most applications is relatively modest. Also, the requirements of instrumental phase stability are no severe, as in the case of CSC scheme. The disadvantage of the SSC is the additional hardware required for each antenna.

Furthermore, there are the baseband combining (BC) and carrier arraying (CA) architectures discussed in (Rogstad et al., 2003). In BC architectures the signal from each antenna is carrier locked and combining in baseband for further demodulation and synchronization. In effect, the carrier signal from the spacecraft is used as a phase reference so that locking to the carrier eliminates the radio-frequency phase differences between antennas imposed by the propagation medium. Besides, in CA architectures, one individual carrier-tracking loop is implemented on each array element. Then, the elements branches are coupled in order to increase the carrier-to-noise ratio (CNR), but losses of radio channel are far compensated (Rogstad et al., 2003).

In general, the selection of the beamforming algorithms is determined by the following aspects: Hardware and computational resources; Speed of convergence and residual error of adaptive algorithms; Calibration requirements and auto-compensation ability; and system signal-transmission characteristics.

4.3 Beamforming techniques for satellite tracking

Some satellites transmit useful information inside its frames for synchronization and tracking purposes. The gathering of satellite data requires the tracking operation along its earth orbit. To accomplish this goal with adaptive array architectures, some beamforming techniques should be implemented. Fig. 23 illustrates a simple example of a narrowband linear adaptive beamformer system.

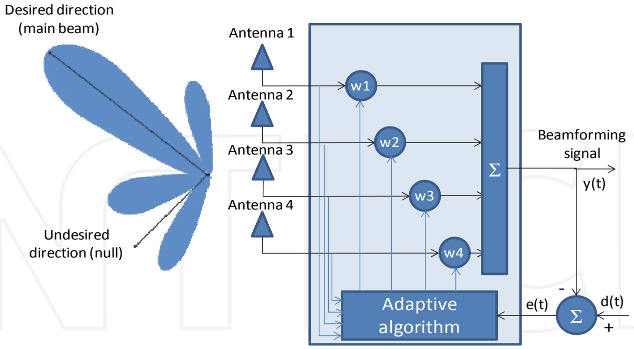

Fig. 23. Adaptive antenna system.

A linear beamformer combines signals according to some weights w_i, to produce a desired radiation pattern. The mathematical expression of a linear beamformer at the array output in vector notation can be expressed as $y = w^H x$, where x is the received signal vector to be combined, w are the weights computed by the beamforming algorithm and ^{H} denotes transposition and conjugate of $(\cdot)$.

In adaptive antennas design, weights are dynamically calculated with a certain algorithm in order to optimize some signal parameter like signal to interference-plus-noise ratio (SINR), SNR, or BER. An extended variety of algorithms exist in the literature for beamforming purpose and the most appropriated selection is done depending on the signal characteristics of the received signal.

4.3.1 Blind techniques

Blind beamformers make use of an inherent property of the received signal, such as the ciclo-stationarity of the constant modulus. In the latter, the algorithm eliminates the fluctuation of the signal amplitude and computes the weights to minimize the effect produced by those variations. The algorithms that make use of these methods are denoted as Constant Modulus Algorithms (CMA) (Biedka, 2001).

CMA algorithms present an important disadvantage: as the phase information is not considered, the constellation of quadrature phase shift keying (QPSK) signals commonly used in satellite communications appears rotated after beamforming, which imposes the need of an additional phase recovery subsystem in the array output.

4.3.2 Temporal-reference algorithms

Algorithms based on a temporal reference require a known reference included in the frame of the signal, such as training sequences, unique word (UW) or pilot bits. Thus, these schemes are normally used for digital signals. The aim of these beamformers is the minimization of the energy of an error signal integrated by interferences and noise. In order to reduce the order of the problem, the weight calculation is usually done iteratively.

The most popular adaptive filters are the LMS and Recursive Least Squares (RLS) algorithms (Haykin, 2002). Briefly, the main differences lie in the method to calculate and the final convergence behavior: while LMS has a linear complexity order with the number of antennas in the array, RLS makes use of matrix operation, so that the complexity order is quadratic, but the convergence is faster.

An interesting alternative to the LMS is the Normalized LMS (NLMS), which normalizes the adaptive step to avoid variation during the convergence process. The counterpart is the more intensive processing requirements to calculate signal power and normalization operation.

4.3.3 Correlation-based algorithm

In contrast to beamformers based on temporal reference, schemes based on signal correlation do not require the demodulation of any signal. These techniques are the most popular to extract the spatial information for beamforming, and we have focused on the use of the SIMPLE algorithm (Rogstad, 1997). This algorithm has been used by the Deep Space Network (DSN) of National Aeronautics and Space Administration (NASA) to combine the signals received from spatial probes in radio telescopes located in different sites around the Earth surface. The main disadvantage of correlation based schemes is the lack of ability to cancel interference signals.

4.4 Performance comparison

Some simulation comparisons between spatial and blind algorithms are presented to show benefits and drawbacks. Four algorithms have been selected with a 4-element uniform linear array (ULA). The spatial algorithms simulated are post-beamformer interference canceller – orthogonal interference beamformer (PIC-OIB) (Godara, 2004) and minimum power distortionless response (MPDR) (Van Trees, 2002). On the other hand, the blind algorithms are the matrix-free EIGEN and the SUMPLE (Rogstad, 1997). The convergence process is compared as a function of the input SNR as depicted in Fig. 24.

As it can be observed from the above results, spatial algorithms outperform blind ones at low SNR, and vice versa. On the other hand, with medium-low SNR and low or absence of interferences, the behavior of all algorithms is quite similar.

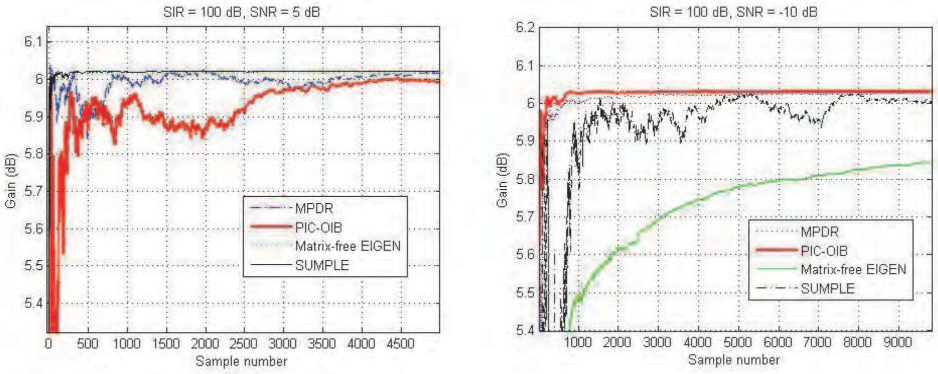

a b

Fig. 24. Convergence behavior of spatial versus blind algorithms in the absence of interferences with several input SNR. a) SNR = 5 dB, and b) SNR = -10 dB.

5. Experimental Test-Bed based on SDR platform

This section presents a test platform known as Adaptive Antenna Array Test-Bed - A3TB, where a comparative study of several beamforming algorithms can be performed and modularity of the architecture is a well proved advantage. The test bed is based on SDR technology and uses a novel architecture that can be used with both blind and spatial-based beamforming algorithms. The A3TB concept can be applied to a number of scenarios as the current version is independent of the signal properties. Simulation results using the A3TB with the APT channel from NOAA satellites show the performance of the concept and the feasibility of the proposed implementation.

The scope of the system development was is to prove the concept of antenna arrays applied to ground stations instead of reflectors for different applications, such as telemetry data downloading or end-user in mobile applications as discussed in the introduction section. In contrast to reflector antennas, antenna arrays offer the possibility of electronic beam-steering avoiding the use of complex mechanical parts and therefore reducing the cost of the antenna. It is also a good chance for Universities and Research Centers aiming to have their own ground station sited in their installations.

5.1 A3TB concept

The A3TB can be defined as *a software-defined radio beamformer applied to a ground station for tracking LEO satellites*. The novelty relies on the use of an antenna array to smartly combine the received signals from the satellite and its implementation based on SDR technology. The reason to use an antenna array instead of a single antenna is to electronically steer the beam in the direction of the satellite along its orbit without requiring a mechanical system for tracking. In addition to the advantages of the use of SDR technology and antenna array, it is the modularity and flexible architecture implemented in the A3TB. Fig. 25 shows the A3TB architecture where it is evident the feasibility to update or change during operation any of the main blocks. It is possible to change during operation the beamforming algorithm and to include new beamforming modules to the system. Furthermore, changes on the BENADC are possible to implement not during operation, but new receiver architecture at off-line such as those options discussed at follow.

In (Salas et al., 2007), the block diagram represents the software system implementation of the first version of the test-bed prototype and most of it is based on VHDL. Depending on the firmware, three options could be installed into the FPGA Virtex4. The option A is implemented with the signal processing on the PC, so the SIMPLE beamforming is done in the module developed in C++. The option B is implemented completely on VHDL and this option need to export the beamforming weights just to draw the array pattern diagram. Finally, in contrast to the option B, the option C is implemented for the LMS beamforming algorithm.

With the first version of the Test-Bed, the modularity on the selection of firmwares was proved switching between A, B or C receiver architectures, and an important result of the Test-Bed development is the hardware resources occupation presented in (Salas et al., 2007).

The advantage of the SDR implementation is that A3TB architecture can be used to process any received signal from a LEO satellite in the appropriate band imposed by the RF stages. Moreover, most of the processing tasks are performed on software, using appropriate routines to process any receive signal. There are 2 main schemes to implement the beamforming stage: SC and FSC [41]. Both schemes are compared in Section 4.2.

The current version of the A3TB in Fig. 25.a was updated to track NOAA satellites in the VHF band, in particular the APT channel. Previous versions of A3TB dealt with LRPT signals from MetOp-A, where a complete receiver with beamforming and synchronization stages has been implemented(Salas et al., 2007; Martínes et al., 2007).

5.2 Implementation of the A3TB

The A3TB prototype consists of 4 main parts as shown in Fig. 25.a. The first part is the antenna array, which has 4 crossed-dipole antennas as depicted in Fig. 25.b. The second part consists of RF-IF circuits which amplify and down convert to IF incoming signals. Furthermore, an automatic gain control (AGC) was implemented using two steps of variable attenuators in the IF domain.

The third part is the SDR platform which consists of the beamforming algorithms implemented on C++ and the FPGA firmware on VHDL, PC and BENADC blocks show in Fig. 25, respectively. The hardware resources occupation for this Test-Bed implementation is similar to one presented in (Martínes et al., 2007). The last part is the software from weather satellite signal to image decoder (WXtoImg) on the PC using the sound card output/input in order to get the weather satellite image.

Since the implemented architecture is FSC the demodulation is not required and the IF signal is digitized. For the signal processing hardware design the BenADC-v4 has been chosen. This solution includes a FPGA Xilinx Virtex4-SX55 with four 12-bit analog inputs at

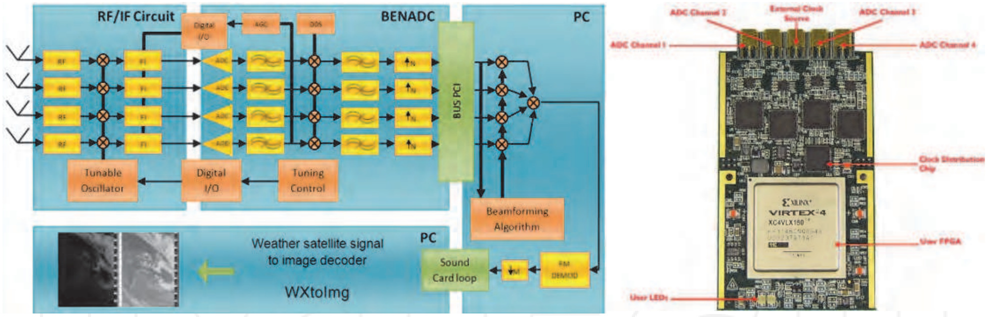
RF/IF Circuit
BENADC
PC
Digital I/O
AGC
DDS
Tunable Oscillator
Digital I/O
Tuning Control
BUS PCI
Beamforming Algorithm
Weather satellite signal to image decoder
WXtoImg
PC
Sound Card loop
FM DEMOD
ADC Channel 2
External Clock Source
ADC Channel 3
ADC Channel 1
ADC Channel 4
Clock Distribution Chip
User FPGA
User LEDs
XILINX
VIRTEX-4
XC4VLX160

a b

Fig. 25. a) Block diagram of the A3TB, and b) BenADC – Virtex 4-sx55.

250 Msps (Martínes et al., 2007). Digital samples are transferred to the PC where beamforming and subsequent APT demodulation of the array output are performed using C++ routines. This implementation design offers higher flexibility for testing different beamforming schemes. Finally, demodulated APR frames are sent to the WXtoImg software to show meteorological maps.

The A3TB is controlled by the PC for simulations and field trials. The graphical user interface allows presented in (Salas et al., 2008) the user to choose the beamforming algorithm and set all the parameters of the LEO satellite for tracking such as the number of antennas of the array, distance between the elements, direction of arrival and IF frequency. The C++ routine calculates the beamforming weights and plots the synthesized array factor. Subsequently, the reception of meteorological images has real time system requirements. Thus, it is necessary a data transfer from the FPGA to the C++ module to process the samples continuously, and give APT frames to the audio output of the PC. Since, the meteorological satellites often have a low baud rate, in the case of study with NOAA satellites the data transfer is made using two buffers controlled by a thread.

It is important to mention that the A3TB with SDR architecture can evaluate different beamforming algorithms and receiver schemes. The update of A3TB for larger arrays is immediate, as the basis for algorithms is independent of the number of elements in the array. The architecture of a new ground station concept to track LEO satellites based on software defined radio and antenna arraying as Test-Bed is a well proved choice to evaluate future antenna array architectures for satellite communication and benchmark features of the proposed system. As the A3TB VHF version is based on FSC scheme, the concept can be applied to a number of satellite tracing scenarios.

6. Conclusions

The performance analysis of different beamforming algorithms is an important issue in the new generation antenna array development and research. Thus, A3TB helps to analyze beamforming algorithms paving the way for testing and debugging for posteriori use in larger arrays, such as GEODA. Results obtained in real scenarios with A3TB state, for example, that spatial reference algorithms such as MPDR should be used in the absence of interferences, whereas blind algorithms are appropriate for low SNR conditions. Finally, the A3TB can also serve to validate the performance of calibration procedures.

In future work, the A3TB will deal with the system combining of full modularity with the capability of change firmwares based on the first version design of the Test-Bed, plus the flexible architecture of the current design of the Test-Bed based on VHDL, C++ and Antenna Arraying. Furthermore, the addition of more modules to increase the number of antenna array elements is evident in next generations.

7. Acknowledgment

Authors wish to thank MICINN (Ministerio de Ciencia e Innovación) for grants and CROCANTE project (ref: TEC2008-06736/TEC), INSA (Ingeniería y servicios Aeroespaciales) and Antenas Moyano S.L., for the partial funding of this work. Simulations in this work have been done using CST Studio Suite 2011 under a cooperation agreement between Computer Simulation Technology and Universidad Politécnica de Madrid. Substrates used in prototypes were kindly given by NELCO S.A.

8. References

Torre, A.; Gonzalo, J.; Pulido, M., & Martínez Rodríguez-Osorio, R. (2006). New generation Ground Segment Architecture for LEO satellites. *57th International Astronautical Congress*, pp. 221-226. Valencia, Spain, October 2006.

Tomasic, B.; Turtle, J. & Liu, S. (2002). A Geodesic Sphere Phased Array Antenna for satellite control and communication. *XXVII General Assembly of the International Union of Radio Science,* Maastricht, August 2002.

Godara, L. C. (1997). Application of Antenna Arrays to Mobile Communication, Part II: Beamforming and Direction of Arrival Considerations. *Proc. IEEE* , vol.85, No.8, (August 1997), pp. 1195-1245.

Martínez, R. & Salas Natera, M. A. (2010). On the use of Ground Antenna Arrays for Satellite Tracking: Architecture, Beamforming, Calibration and Measurements. *61st International Astronautical Congress*, pp. 1-7. Prague, 2010.

Liu, S.; Tomasic, B.; Hwang, S. & Turtle, J. (2006). The Geodesic Dome Phased Array Antenna (GDPAA) for Satellite Operations Support. *IEEE 18th International Conference on Applied Electromagnetics and Communications*, pp. 1-1, Dubrovnik, April 2006.

Tomasic, B. (1998). Analysis and Design Trade-Offs of Candidate Phased Array Architectures for AFSCN Application. *Second AFSCN Phased Array Antenna Workshop,* Hanscom, April 1998.

Sierra Pérez, M.; Torre, A.; Masa Campos, J. L.; Ktorza, D. & Montesinos, I. (2007). GEODA: Adaptive Antenna Array for Metop Satellite Signal Reception. *4th ESA International Workshop on Tracking, Telemetry and Command System for Space Application*, pp. 1-4, Darmstadt, Germany, September 2007.

Arias Campo, M.; Montesinos Ortego, I.; Fernández Jambrina, J. & Sierra Pérez, M. (2010). T/R Module Design for GEODA Antenna. *4th European Conference on Antenna and Propagation,* pp. 115-116, Barcelona, Spain, April 2010.

Franchi, A.; Howell, A. & Sengupta, J. (2000). Broadband mobile via satellite: Inmarsat BGAN. *IEEE Seminar on the Critical Success Factors - Technology, Services and Markets*, pp. 23-30, October 2000.

Evans, J. V. (1998). Satellite systems for personal communications. *Proc. IEEE* , Vol.86, No.7, (July 1998), pp. 1325-1341.

Wu, W. W. (1994). Mobile satellite communications. *Proc. IEEE* , Vol.82, No.9, (September 1994), pp. 1431-1448.

Fujimoto, K. & James, J. (2001). *Mobile Antenna Systems Handbook.* Artech House.

Garcia Aguilar, A.; Inclán Alonso, J. M.; Vigil Herrero, L.; Fernández González, J. M. & Sierra Pérez, M. (2010). Printed antenna for satellite communications. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 529-535, Boston, USA, October 2010.

Geissler, M.; Woetzel, F.; Bottcher, M.; Korthoff, S.; Lauer, A. & Eube, M., (2010). L-band phased array for maritime satcom. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 518-523, Boston, USA, October 2010.

Baggen, R.; Vaccaro, S. & Del Río, D. (2007). Design Considerations for Compact Mobile Ku-Band Satellite Terminals. *EuCAP The Second European Conference on Antennas and Propagation*, pp. 1-5. Edimburg, Scothland, September 2007.

Vaccaro, S.; Tiezzi, F.; Rúa, M. & De Oro, C. (2010). Ku-Band Low-Profile Rx-only and Tx-Rx antennas for Mobile Satellite Communications. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 536-542, Boston, USA, October 2010.

Schippers, H. (2008). Broadband Conformal Phased Array with Optical Beam Forming for Airborne Satellite Communication. *IEEE Aerospace Conference*, pp. 1-17, September 2008.

Kanno, M.; Hashimura, T. T.; Sato, M.; Fukutani, K. & Suzuki, A. (1996). Digital beam forming for conformal active array antenna. *IEEE International Symposium on Phased Array Systems and Technology*, pp. 37-40, October 1996.

Khalifa, I. & Vaughan, R. (2007). Optimal Configuration of Multi-Faceted Phased Arrays for Wide Angle Coverage. *IEEE 65th Vehicular Technology Conference*, pp. 304-308, Boltimore, USA, April 2007.

Salas Natera, M. A.; Martínez Rodríguez-Osorio, R.; Antón Sánchez, A.; García-Rojo, I. & Cuellar, L. (2008). A3TB: Adaptive Antenna Array test-bed for tracking LEO satellites based on software defined radio. *59th International Astronautical Congress*, pp. 313-317, Glasgow. September 2008.

Salas Natera, M. A.; Martínez Rodríguez-Osorio, R. & García-Rojo López, I. (2007). Design of an Adaptive Antenna Array Test-Bed based on Software Radio for Tracking LEO Satellites. *IEEE EuCAP.* Edinburgh, Scotchland, Npvember 2007.

Martínez, R.; Salas Natera, M.; Bravo, A.; García-Rojo, I.; de Haro, L.; Mateo, M. & Gómez, M. (2007). VHF Ground Station with increased angular coverage for reception of meteorological satellites with electronic beamforming. *4th ESA International Workshop on Tracking, Telemetry and Command Systems for Space Application.* Darmstadt, Germany, 2007.

Mailloux, R. (2005). *Phased Array Antenna Handbook,* Artech House, Norwood, Massachusetts, USA.

Josefsson, L. & Persson, P. (2006). *Conformal Array Antenna. Theory and Design*, John Wiley & Sons , Hoboken, New Jersey, USA.

Montesinos, I.; Sierra Pérez, M.; Fernández, J. L.; Martínez, R. & Masa, J. L. (2009). GEODA: Adaptive Antenna of Multiple Planar Arrays for Satellite Communications. *European Conference on Antenna and Propagation.* Berlin, Germany, 2009.

Salas Natera, M. A.; Martínez, R.; De Haro Ariet, L. & Fernández Jambrina, J. (2010). Automated System for Measurement and Characterization of Planar Active Arrays. *IEEE Internationa Symposium on Phase Array Systems and Technology*, pp. 1-6, Boston USA, October 2010.

Sierra-Castañer, M.; Vera-Isasa, M.; Sierra-Pérez, M. & Fernández-Jambrina, J. (2005). Double-Beam Parallel Plate Slot Antenna. *IEEE Transactions on Antennas and Propagation*, vol.53, No.3, (2005) pp. 977-984.

Garg, R.; Bhartia, P.; Bahl, I. & Ittipiboon, A. (2001). *Microstrip Antennas Design Handbook,* Artech House, Norwood, Massachusetts, USA.

Tang, C. & Chen, M. (2007). Synthesizing Microstrip Branch-Line Couplers With Predetermined Compact Size and Bandwidth. *IEEE Transactions on Microwave Theory and Techniques, vol. 55, No.9,* (September 2007), pp. 1926-1934.

Bialkowski, M.; Jellett, S. & Varnes, R. (1996). Electronically Steered Antenna System for the Australian Mobilesat. *IEEE Transaction on Antennas ans Propagation*, vol. 143, No. 4, (August 1996), pp. 347-352.

Alonso, J.; Blas, J.; Garcia, L.; Ramos, J.; Pablos, J. & Grajal, J. (1996). Low Cost Electronically Steered Antenna and Receiver System for Mobile Saltellite Communications. *Trans. IEEE MTT*, vol. 44, No. 12, (December 1996), pp. 2438-2449.

Fernández, J. M.; Rizzo, C. & Sierra-Pérez, M. (2009). Antena Impresa Multihaz con Polarización Circular Derechas/Izquierdas Para Comunicaciones por Satélite en Banda X. *Proccedings of XXIV Simposium Nacional de la Unión Científica Internacional de Radio (URSI).* Santander, Spain, September 2009.

Encinar, J. A. & Zornoza, J. (2003). Broadband design of three- layer printed reflectarrays. *IEEE Transactions on Antennas and Propagation*, vol. 51, No. 7, (July 2003), pp. 1662-1664.

Padilla, P.; Muñoz-Acevedo, A.; Sierra-Castañer, M. & Sierra-Pérez, M. (2010). Electronically reconfigurable transmitarray at Ku band for microwave applications. *IEEE Transactions on Antennas and Propagation*, vol. 58, No. 8, (August 2010), pp. 2571-2579.

Padilla, P.; Muñoz-Acevedo, A. & Sierra-Castañer, M. (2010). Low Loss 360º Ku Band electronically Reconfigurable Phase Shifter. *International Journal of Electronics and Communications,* vol. 64, No. 11, (November 2010), pp. 1100-1104.

Royce, W. W. (1970). Managing the Development of Large Software Systems. *IEEE WESCON*, pp. 328 – 338, 1970.

Rogstad, D. H. (1997). *The SUMPLE Algorithm for Aligning Arrays of Receiving Radio Antennas: Coherence Achivied with Less Hardware and Lower Combining Loss.* TDA Progress Report, Jet Propulsion Laboratory.

Rogstad, D. H.; Mileant, A. & Pham, T. (2003). *Antenna Arraying Techniques in the Deep Space Network.* Deep Space Communication and Navegation Series, JPL California Institute of Technology, Ref. 03-001, Pasadena, USA.

Biedka, T. E. (2001). *Analysis and Development of Blind Adaptive Beamforming Algorithms.* PdD Thesis, Faculty of Virginia Polytechnique Institute and State University, Virginia.

Haykin, S. (2002). *Adaptive Filter Theory* (4th ed.). Prentice-Hall.

Godara, L. C. (2004). *Smart Antennas* (1st ed.). CRC Press..

Van Trees, H. L. (2002). *Optimum Array Processing. Part IV of Detection, Estimation, and Modulation Theory*. Wiley.

INTECH

INTECH

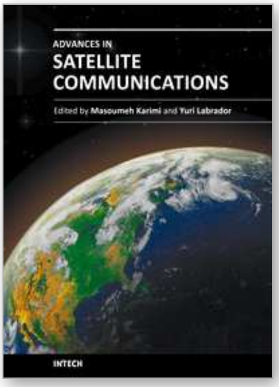

Advances in Satellite Communications

Edited by Dr. Masoumeh Karimi

ISBN 978-953-307-562-4
Hard cover, 194 pages
Publisher InTech
Published online 27, July, 2011
Published in print edition July, 2011

Satellite communication systems are now a major part of most telecommunications networks as well as our everyday lives through mobile personal communication systems and broadcast television. A sound understanding of such systems is therefore important for a wide range of system designers, engineers and users. This book provides a comprehensive review of some applications that have driven this growth. It analyzes various aspects of Satellite Communications from Antenna design, Real Time applications, Quality of Service (QoS), Atmospheric effects, Hybrid Satellite-Terrestrial Networks, Sensor Networks and High Capacity Satellite Links. It is the desire of the authors that the topics selected for the book can give the reader an overview of the current trends in Satellite Systems, and also an in depth analysis of the technical aspects of each one of them.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Miguel Alejandro Salas Natera, Andrés García Aguilar, Jonathan Mora Cueva, José Manuel Fernández, Pablo Padilla De La Torre, Javier García-Gasco Trujillo, Ramón Martínez Rodríguez-Osorio, Manuel Sierra-Perez, Leandro De Haro Ariet and Manuel Sierra Castañer (2011). New Antenna Array Architectures for Satellite Communications, Advances in Satellite Communications, Dr. Masoumeh Karimi (Ed.), ISBN: 978-953-307-562-4, InTech, Available from: http://www.intechopen.com/books/advances-in-satellite-communications/new-antenna-array-architectures-for-satellite-communications

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

Hindawi Publishing Corporation
International Journal of Antennas and Propagation
Volume 2012, Article ID 428284, 10 pages
doi:10.1155/2012/428284

Review Article

Some Recent Developments of Microstrip Antenna

Yong Liu, Li-Ming Si, Meng Wei, Pixian Yan, Pengfei Yang, Hongda Lu, Chao Zheng, Yong Yuan, Jinchao Mou, Xin Lv, and Housjun Sun

Department of Electronic Engineering, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

Correspondence should be addressed to Yong Liu, fatufo@bit.edu.cn

Received 12 August 2011; Revised 23 November 2011; Accepted 2 January 2012

Academic Editor: Zhongxiang Q. Shen

Copyright © 2012 Yong Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Although the microstrip antenna has been extensively studied in the past few decades as one of the standard planar antennas, it still has a huge potential for further developments. The paper suggests three areas for further research based on our previous works on microstrip antenna elements and arrays. One is exploring the variety of microstrip antenna topologies to meet the desired requirement such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so on. Another is to apply microstrip antenna to form composite antenna which is more potent than the individual antenna. The last is growing towards highly integration of antenna/array and feeding network or operating at relatively high frequencies, like sub-millimeter wave or terahertz (THz) wave regime, by using the advanced machining techniques. To support our points of view, some examples of antennas developed in our group are presented and discussed.

1. Introduction

The concept of microstrip antenna was first introduced in the 1950s [1]. However, this idea had to wait nearly 20 years to be realized after the development of the printed circuit board (PCB) technology in the 1970s [2, 3]. Since then, microstrip antennas are considered as the most common types of antennas due to their obvious advantages of light weight, low cost, low profile, planar configuration, easy of conformal, superior portability, suitable for arrays, easy for fabrication, and easy integration with microwave monolithic integrate circuits (MMICs) [4–7]. They have been widely employed for the civilian and military applications such as television, broadcast radio, mobile systems, global positioning system (GPS), radio-frequency identification (RFID), multiple-input multiple-output (MIMO) systems, vehicle collision avoidance system, satellite communications, surveillance systems, direction founding, radar systems, remote sensing, biological imaging, missile guidance, and so on [8].

Despite the many advantages of typical microstrip antennas, they also have three basic disadvantages: narrow bandwidth, low gain, and relatively large size. The narrow bandwidth is one of the main drawbacks of these types of antennas. A straightforward method of improving the bandwidth is increasing the substrate thickness. However, surface wave power increases and radiation power decreases with the increasing substrate thickness [7], which leads to poor radiation efficiency. Thus, various other techniques are presented to provide wide-impedance bandwidths of microstrip antennas, including impedance matching networks using stub [9, 10] and negative capacitor/inductor [11], microstrip slot antennas using the U, L, T, and inverted T slots in the ground plane (sometimes termed defected ground structures (DGSs)) [12, 13], surface wave suppressing using magneto-dielectric substrate [14] and electromagnetic bandgap (EBG) structures [15], and composite-resonator microstrip antennas using metamaterial resonators [16, 17]. Another problem to be solved is the low gain for conventional microstrip antenna element. Cavity backing has been used to eliminate the bidirectional radiation, thereby providing higher gain compared with conventional microstrip antenna [18]. Lens covering is an alternative way to achieve gain enhancement. The lens with canonical profile, like elliptical, hemielliptical, hyper-hemispherical, extended hemispherical, used to focus the radiation beam from the radiator elements. The integrated lens microstrip antenna can be treated as composite antenna combined by microstrip radiator elements and dielectric lens, which is very useful for high frequencies (mm, sub-mm, terahertz (THz), and optical waves) applications

[19]. It is also well known that antenna array is an effective means for improving the gain [20–25].

The last limitation of conventional microstrip antennas is the relatively large size, particularly at lower microwave frequencies, since their operation frequencies are related to the electrical size of antenna. In general, the size of the rectangular microstrip antenna should be of order of a half-guided wavelength. This limitation was mathematically studied by Wheeler [26] and Chu [27]. There have been numerous efforts to minimize the antenna size and obtain the electrically small microstrip antenna with the raised demand towards smaller and smaller wireless devices. Inductive or capacitive loading are effective ways to reduce the size of microstrip antennas [28]. In the former work, we demonstrated that the size of microstrip antenna can be miniaturized using composite metamaterial resonators [16, 17]. Magneto-dielectric substrates have been widely used to miniaturize microstrip antennas due to magnetic substrates and could provide wider bandwidths than dielectric substrates [29–32]. Fractal geometries, which are composed by self-similar structures, have opened an alternative way for antenna miniaturization [33].

From the above discussions, we see that many methods and materials are used to improve the properties of microstrip antennas. However, there should be a relationship among bandwidth, gain, and size of the microstrip antennas. Antenna engineers have recognized that the improvement in one antenna property is frequently accompanied by decline in its other performances. For example, the antenna size is reduced usually at the expense of its bandwidth and gain. Therefore, a more comprehensive consideration must be given on further developments of microstrip antennas.

In this paper, we will suggest three areas for further research based on our previous works on microstrip antenna elements and arrays [16–25, 34–41]. We first note that novel microstrip antenna topologies are proposed to meet the desired requirement of variety of potential wireless applications, such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so on. Next, we discuss the composite antennas based on microstrip antennas which have more potent than each individual antenna. Finally, with the development of micro-/nano-machining techniques, antennas/arrays with highly integration and with highly operating frequencies are discussed. We present some examples of antennas developed in our group to support our points of view.

2. Variety of Microstrip Antenna Topologies

Microstrip antennas have extensively used in commercial and military applications due to their attractive advantages. However, the traditional microstrip antennas have the impedance bandwidth of only a few percent and radiation pattern with omnidirection, which obviously does not meet the requirements of various wireless applications. To this end, a wide variety of microstrip antenna topologies, including different microstrip antenna element structures and different microstrip array arrangements, have been studied to meet the desired requirement such as ultrawide band (UWB), high gain, miniaturization, circular polarization, multipolarized, and so forth.

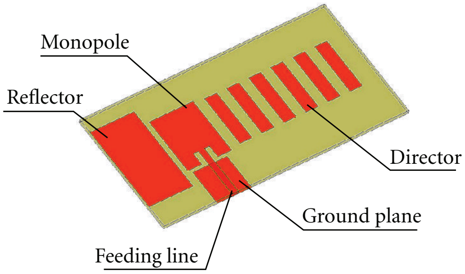

(a) The structure of the quasi-Yagi antenna

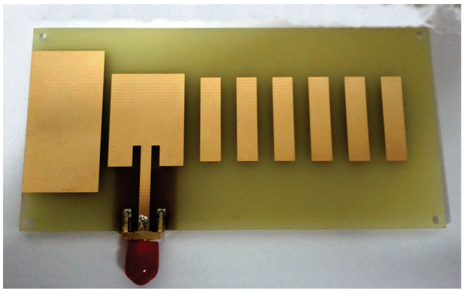

(b) The photograph of the quasi-Yagi antenna

FIGURE 1: Compact broad-band quasi-Yagi antenna.

As we know, microstrip antennas inherently have narrower bandwidth and lower gain compared to conventional bulky antennas. Some microstrip antennas with special topologies, like quasi-Yagi, planar reflector antenna, are proposed to replace the conventional bulky antennas. Here, we will take a quai-Yagi antenna as an example to show how to design a planar microstrip antenna with Yagi-Uda end-fire radiation pattern. In addition, a microstrip array with special array topology is designed to get dual-polarized property.

2.1. Compact Broad-Band Quasi-Yagi Antenna. A novel S-band compact quasi-Yagi antenna has been designed, fabricated and measured by our group, as shown in Figure 1. This antenna is composed of a printed monopole-driven element, a printed reflector element, and six printed director elements.

To explain the end-fire radiation behavior of the quasi-Yagi antenna, a comparison of radiation patterns, among (1) microstrip monopole only, (2) microstrip monopole and a reflector, (3) microstrip monopole and a director, (4) microstrip monopole and a reflector with one director, and (5) microstrip monopole and a reflector with six director, is shown in Figure 2. We can observe that both the reflector and the director can increase the end-fire radiation, and it could be substantially improved by increasing the number of directors.

The measured VSWR results are shown in Table 1. A bandwidth of 14% for VSWR less than 1.5 is achieved. The gain of the antenna is above 7.5 dBi, as shown in Table 2. In this design, we see that the microstrip antenna with special topology could be conveniently used to replace the bulky Yagi-Uda antenna.

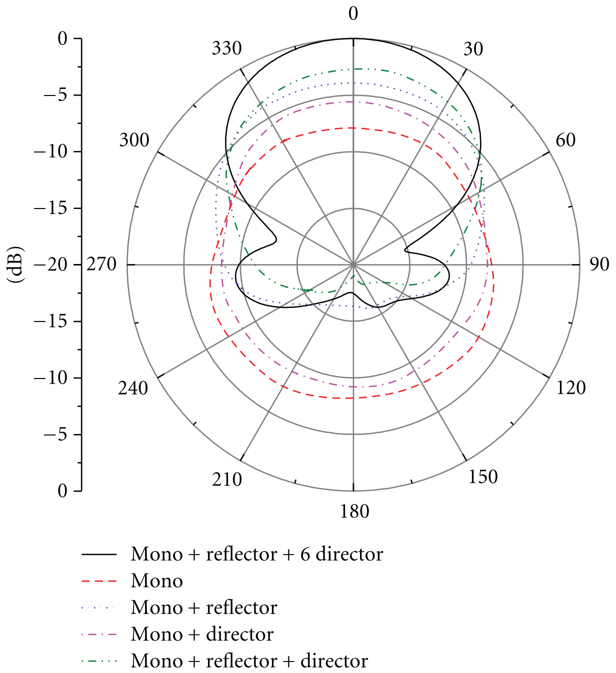

Figure 2: Radiation patterns of microstrip monopole only, microstrip monopole and a reflector, microstrip monopole and a director, microstrip monopole and a reflector with one director, and microstrip monopole and a reflector with six director.

Table 1: The measured VSWR of the quasi-Yagi antenna.

No.	Frequency (GHz)			
	3.25	3.5	3.75	Inband
1	1.36	1.34	1.47	<1.5
2	1.37	1.26	1.49	<1.5
3	1.36	1.25	1.48	<1.5

Table 2: The measured gain of the quasi-Yagi antenna (unit: dBi).

No.	Frequency (GHz)		
	3.25	3.5	3.75
1	7.57	8.73	8.35
2	7.58	8.55	8.37
3	7.56	8.77	8.51

2.2. Dual-Polarized Microstrip Antenna Array. The dual-polarized antenna is highly required for the radar, electronic countermeasure, and aerospace systems. It is known that the microstrip antenna can easily be integrated with microwave circuits and feeding network. Here, a novel Ku-band dual-polarization microstrip antenna array with a mixed feeding network, that is, the slot coupled feeding (V-port) and the co-plane feeding (H-port), is designed by our group, as shown in Figure 3. It is a three layers structure: top microstrip patch layer, middle stripline feeding network layer, and bottom coplane microstrip feeding network layer. Through proper array arrangement, very good isolation can be obtained.

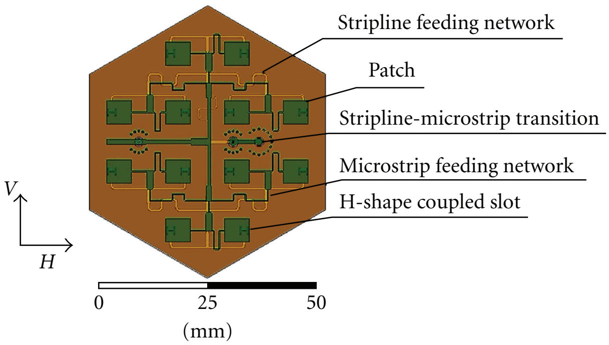
Stripline feeding network
Patch
Stripline-microstrip transition
Microstrip feeding network
H-shape coupled slot
V
H
0
25
50
(mm)

(a) The structure of the dual-polarized microstrip antenna array

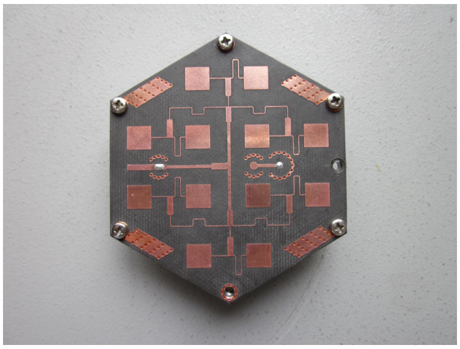

(b) The photograph of the dual-polarized microstrip antenna array

Figure 3: Dual-polarized microstrip antenna array.

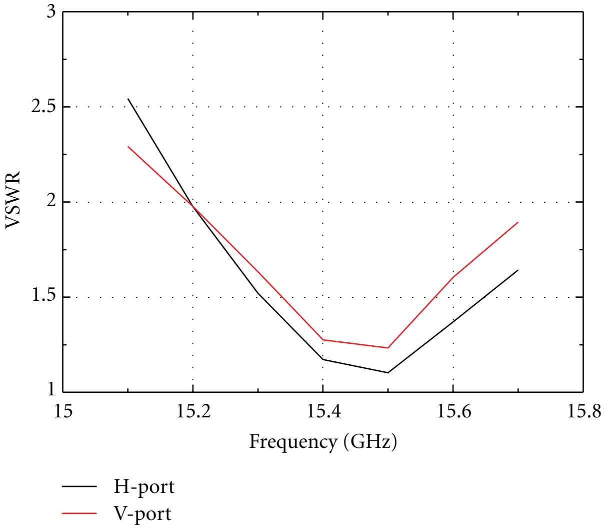

Figure 4: The VSWR of the dual-polarized microstrip antenna array.

The VSWR, radiation patterns, and the isolation between two polarizations of the proposed dual-polarized microstrip antenna array are shown in Figures 4, 5, and 6, respectively. The results indicate that this microstrip antenna array has a good impedance matching, good radiation performance, as well as very high isolation (less than −25 dB), which can be an idea candidate for the dual-polarized wireless systems.

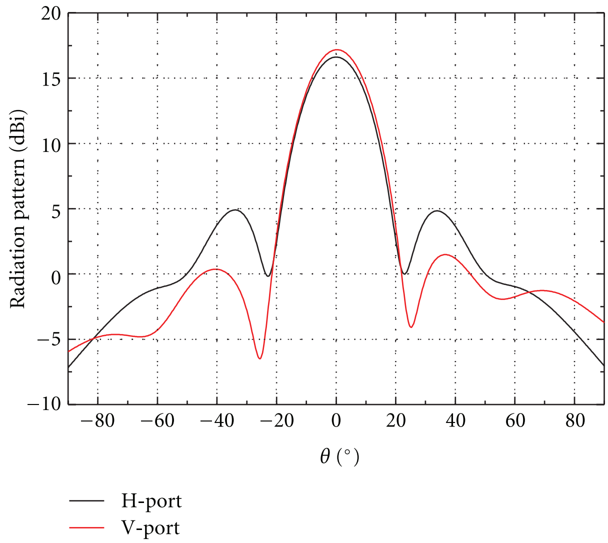

FIGURE 5: The radiation patterns of the dual-polarized microstrip antenna array at the center frequency.

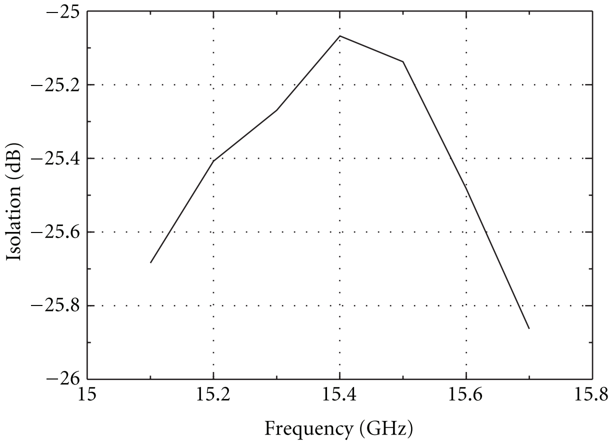

FIGURE 6: The isolation of the dual-polarized microstrip antenna array.

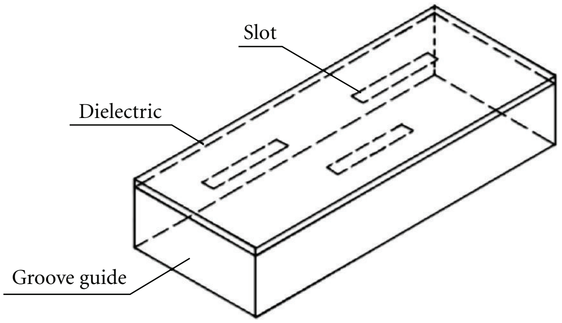

FIGURE 7: The structure of the DCWS.

3. Microstrip-Antenna-Based Composite Antenna

As many antenna designers have found, it is not easy to design an antenna to meet the user-defined stringent performance requirements demanded by special wireless applications like military radars, surveillances, and missile guidance, if only one type of antenna is considered. This difficulty may require the use of two more different types or structures of antenna elements with different characteristics. Composite antenna formed by two more types or structures of antennas is particularly suitable for these applications due to more advantages offered by different types or structures of antennas. For example, it is a challenging task to use single type of antenna to design a dual-band dual-polarization antenna for satellite digital multimedia broadcast (S-DMB) application [36]. A composite antenna composed with a left-handed circularly polarized (LHCP) microstrip antenna and a linear polarized omnidirectional biconical antenna is proposed by our group to meet this requirement [36]. Another example of composite antenna is comprised of a dielectric lens and microstrip log-period antenna, which has been widely applied to THz systems (this type of antenna will be further discussed in Section 4.2). Here, we will give an example of composite antenna with "structure composite" method.

3.1. Monopulse Circular-Polarized Dielectric Complex Waveguide Slot Antenna Array. Waveguide slot antenna array has been widely used for wireless system, due to its advantages of high radiation efficiency, high power capacity, and high reliability. However, it is hard to overcome the disadvantage of high cost of fabrication.

One composite antenna with waveguide slot antenna array property, termed dielectric complex waveguide slot (DCWS), is composed with slot microstrip line and groove guide, as shown in Figure 7. The slot microstrip line is formed by a metal clad dielectric substrate and slots etched in the metal. This composite antenna not only maintains the advantages of the traditional waveguide slot antenna array but also has the characteristics of high consistence, easy for fabrication, and low cost.

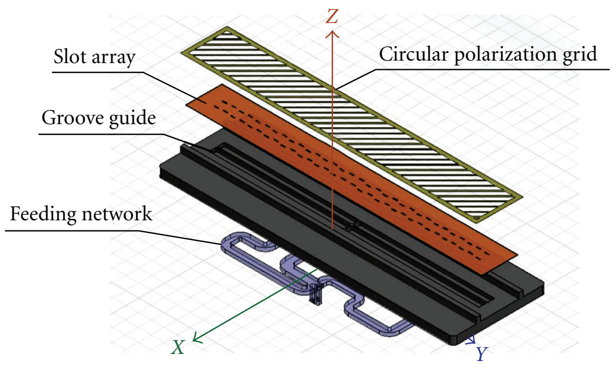

(a) The structure of the monopulse circular-polarized DCWS antenna array (separating view)

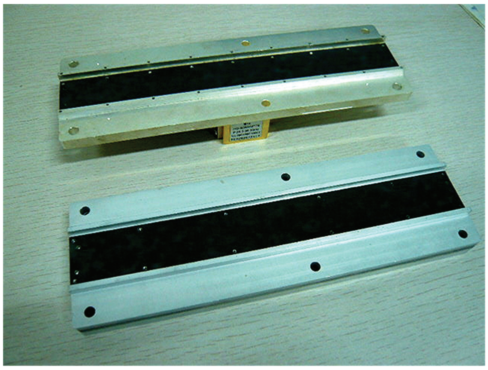

(b) The photograph of the monopulse circular-polarized DCWS antenna array.

Figure 8: Ka-band monopulse circular-polarized dielectric complex waveguide slot (DCWS) antenna array.

A Ka-band monopulse circular-polarized dielectric complex waveguide slot (DCWS) antenna array is designed, fabricated, and measured by our group, as shown in Figure 8. It consists of a circular polarization grid, a slot microstrip array, and a groove guide and feeding network. The slot microstrip array is fabricated on a Rogers 5880 film with dielectric constant of 2.2 and the thickness of 0.254 mm. The measured results of VSWR of sum and different port are shown in Figure 9. Figure 10 shows the measured radiation pattern at the center frequency. Some important array performance parameters such as gain, null depth and axial ratio (AR) are also given in Table 3. As shown in the measured results, very good performance can be obtained with the DCWS antenna array. The radiating efficiency of the DCWS antenna array is 80%, which is almost the same as the traditional waveguide slot antenna array. Moreover, the DCWS antenna array has 40% larger bandwidth than the traditional waveguide slot antenna array.

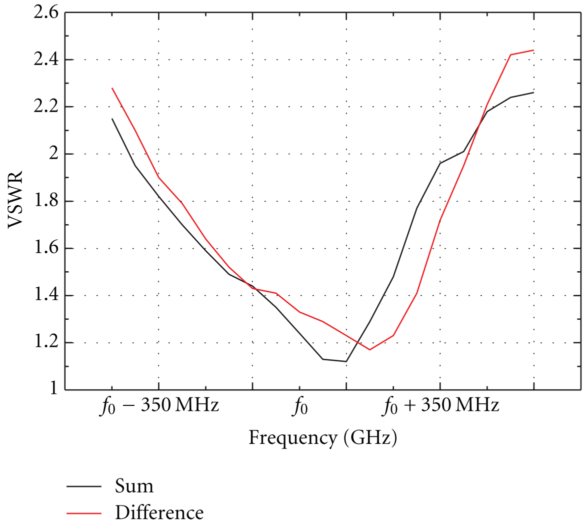

Figure 9: The VSWR of sum and difference port of the monopulse circular-polarized DCWS antenna array.

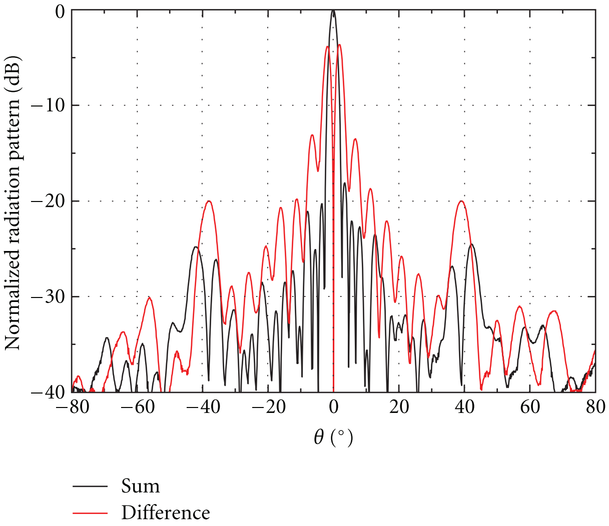

Figure 10: The radiation pattern of the monopulse circular-polarized DCWS antenna array at the center frequency.

4. Highly Integration and Highly Operating Frequency Antennas Based on Advanced Machining Techniques

It is known that the microstrip antenna was first fabricated using PCB technology in 1970s, nearly 20 years after its concept was first presented in 1950s [1–3]. Clearly, the development of microstrip antennas is closely related with the machining techniques. Recently, various machining techniques, including multilayer printed circuit board (MPCB), complementary metal oxide semiconductor (CMOS), low-temperature cofired ceramics (LTCC), and micro-electro-mechanical systems (MEMS), are highly developed, opening opportunities for innovative antennas, such as active antennas, reconfigurable antennas, metamaterial-based antennas, THz antennas, and so forth. With the availability of high-precision and high-speed advanced machining techniques, microstrip antennas are growing towards highly integration of antenna/array and feed network and operating at relatively high frequencies. Since they are all based on the advanced

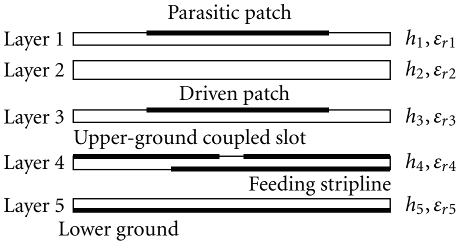

(a) Schematic side view of the structure of the high integrate broadband microstrip antenna array

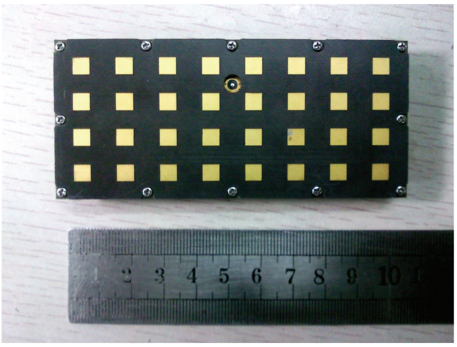

(b) The photograph of the high integrate broadband microstrip antenna array

Figure 11: Ku-band high integrate broadband microstrip antenna array using MPCB technology.

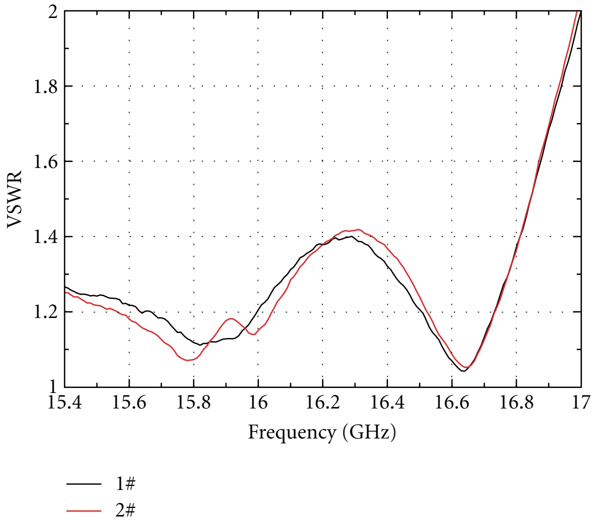

Figure 12: The VSWR of the high integrate broad-band microstrip antenna array using MPCB technology.

machining techniques, we suggest that a third research area of microstrip antennas is constantly introducing novel advanced machining techniques. In the following, two examples will be presented to show how important the advanced machining technique is to fabricate microstrip antennas. One is the highly integrate broad-band microstrip antenna array fabricated using MPCB technology. Another is THz wave planar integrated active microstrip antenna using MEMS technology.

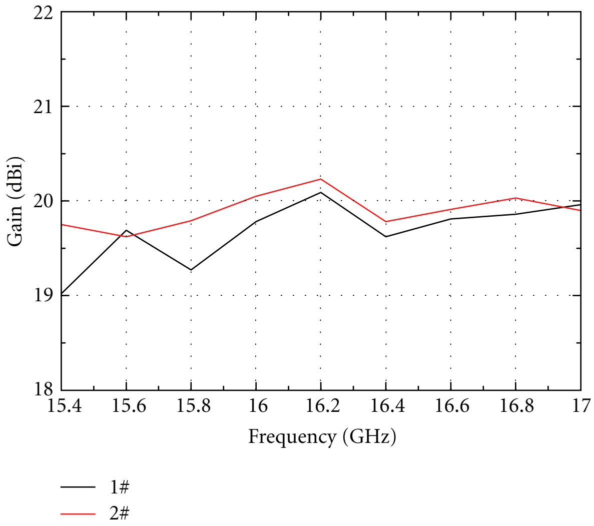

Figure 13: The gain of the high integrate broad-band microstrip antenna array using MPCB technology.

Table 3: The measured data of the monopulse circular-polarized DCWS antenna array.

Fre. (GHz)	Gain (dBi)	Null depth (dB)	AR (dB)
$f_0 - 0.2$	22.8	−37.3	3.8
f_0	21.9	−29.9	2.9
$f_0 + 0.2$	22.1	−26	4.1

4.1. High Integrate Broad-Band Microstrip Antenna Array Using Multilayer Printed Circuit Board (MPCB) Technology.

Recently, with the development of the multilayer printed circuit board (MPCB) technology, the microstrip antennas can be designed and fabricated from one-dimensional (1D) to 2D and even 3D structures.

Based on the MPCB technology, a high integrated broad-band Ku-band microstrip antenna array is designed, fabricated, and measured by our group, as shown in Figure 11. This antenna consists of a parasitic patch, a driven patch, a stripline feeding network, a broad-band coaxial line to stripline transition, some buried screw holes, and some via holes. The feeding network is integrated in the bottom of the substrate of the antenna. As all of the structures fabricated at once, the accuracy and the uniformity can be assured. Two antennas of this type are measured. The measured VSWR, gain, and radiation pattern at the center frequency are shown in Figures 12, 13, and 14, respectively. The measured results show that this antenna maintains good radiation and matching performances with relative bandwidth of 13%. They have also shown good uniformity by using MPCB technology.

4.2. THz Wave Planar Integrated Active Microstrip Antenna Using Micro-Electromechanical Systems (MEMSs) Technology.

THz waves typically include frequencies between 0.1 THz and 10 THz. THz technology is now becoming a promising technology which has potential applications in many fields, such as short-range communication, biosensor, imaging,

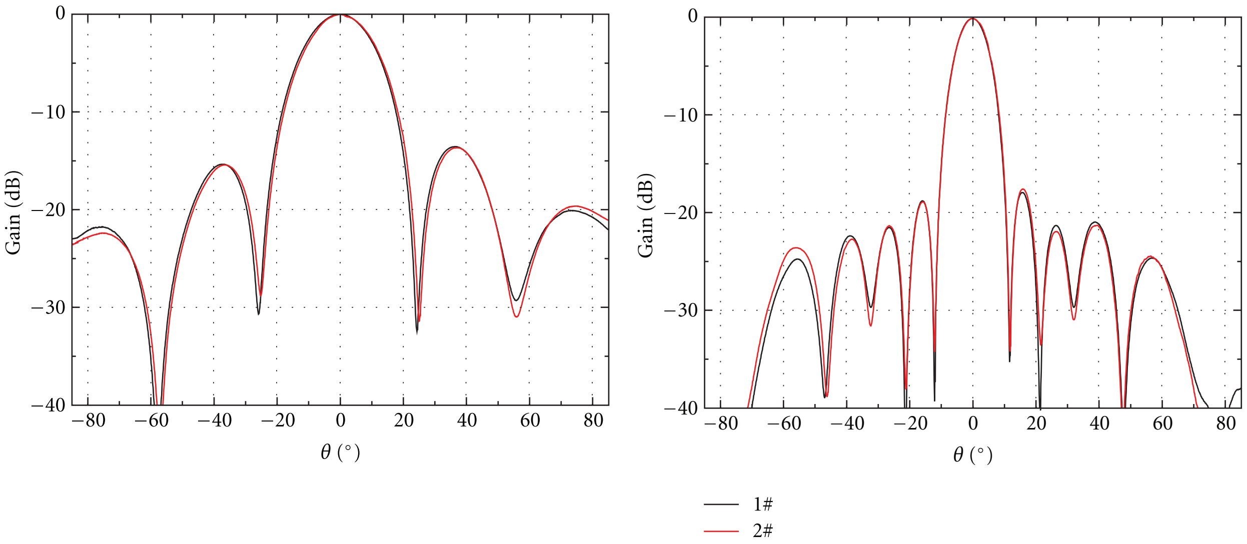

(a) E-plane (b) H-plane

Figure 14: The radiation pattern of the high integrate broad-band microstrip antenna array using MPCB technology at the center frequency.

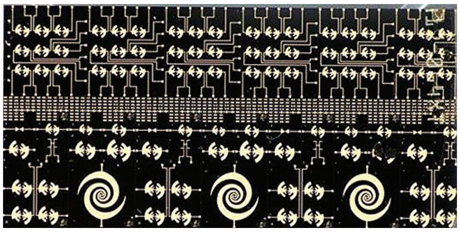

(a) The photograph of the THz monolithic antenna

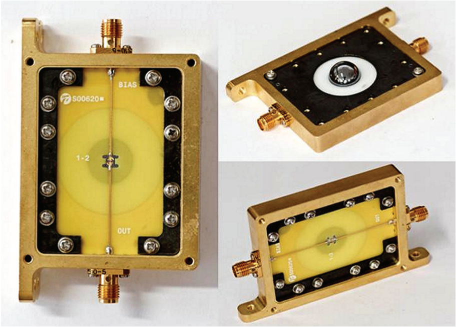

(b) The photograph of the THz monolithic antenna covered by a dielectric lens

Figure 15: THz wave planar integrated active microstrip antenna using micro-electromechanical systems (MEMSs).

national security, space exploration and communication, and so forth [39–46]. To realize THz transceiver system, antenna is an essential component. We often use horn antenna, lens antenna, and dielectric parabolic antenna, for THz systems. However, they are not easy to integrate with monolithic integrate circuits. Although the microstrip antenna has the merits of small volume, light weight, and easy integration with circuit, it is difficult to be processed in such high-frequency regions. MEMS technology opens the way to design of THz antennas, circuits, and systems. THz monolithic antenna fabricated using MEMS technology and covered by a dielectric lens, which can be considered a composite antenna, are designed, fabricated, and measured by our group, as shown in Figure 15.

Diodes have the functions of mixing and/or modulating the carrier-wave signal. It is an effective way to reduce the propagation path for detectors application by integrating the diode and microstrip antenna. The extended hyper-hemispherical dielectric lens is used to increase the gain of the microstrip antenna. An antenna-coupled detector integrated with a dielectric lens is designed and fabricated up to THz range by our group. The planar microstrip log-spiral antenna and log-period antenna have been fabricated using micro-electromechanical systems (MEMSs) technology. The photographs of the antennas are demonstrated in Figure 15. The measured responses of the antenna-coupled detector working at different frequency bands are shown in Figure 16, which can be considered to determine the effective operating frequencies [19, 40]. This detector gave a valid response from 12 GHz to 110 GHz frequencies. The results prove the validity and feasibility of the THz antenna designed using micro-electromechanical systems (MEMSs) technology.

5. Conclusion

The advantages and disadvantages of microstrip antennas are discussed in this paper. In particular, three areas for further development of microstrip antennas are presented based on our previous works on microstrip antenna elements and arrays. Variety of microstrip antenna topologies and microstrip-antenna-based composite antenna are discussed, and

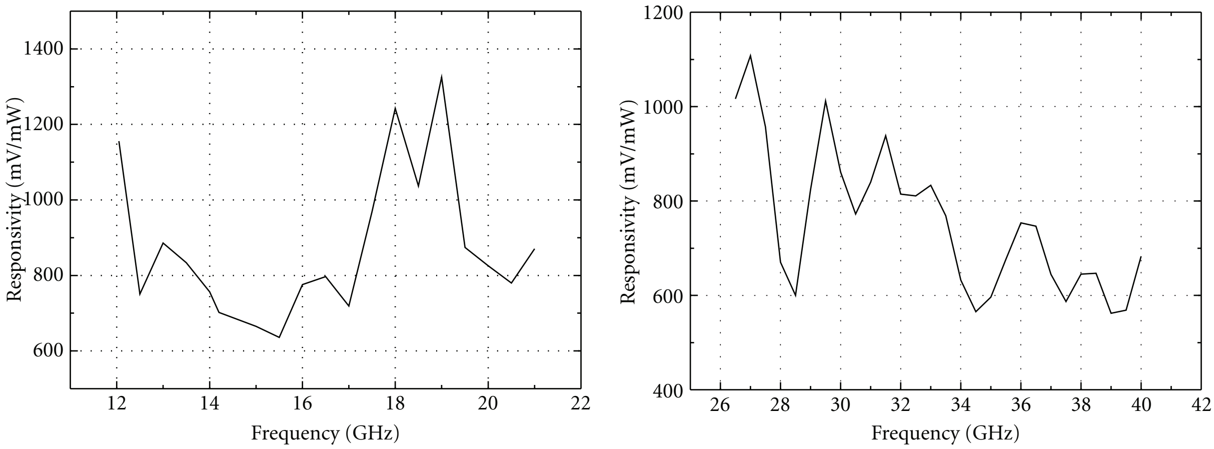

(a) Ku-band

(b) Ka-band

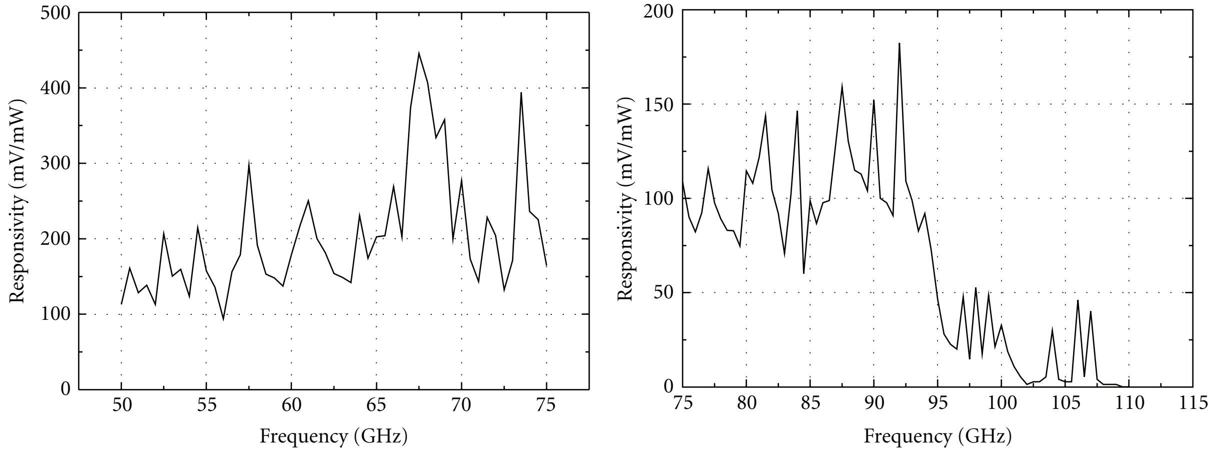

(c) V-band

(d) W-band

FIGURE 16: Frequency responses test results of the THz wave planar integrated active microstrip antenna covered by a dielectric lens.

the advanced machining techniques pushing the microstrip antennas towards the highly integration of antenna/array and feeding network and the highly operating frequencies are described. To demonstrate the distinctive features of novel microstrip antennas, various antenna elements and arrays for different applications are presented. This paper has shown that the microstrip antennas are still very promising paradigm for civilian and military wireless applications.

References

[1] G. A. Deschamps, "Microstrip microwave antennas," in *Proceedings of the 3rd USAF Symposium on Antennas*, 1953.

[2] E. V. Byron, "A new flush-mounted antenna element for phased array application," in *Proceedings of the Phased Array Antenna Symposium*, pp. 187–192, 1970.

[3] J. Q. Howel, "Microstrip antennas," in *Proceedings of the Digest of the International Symposium of the Antennas and propagation Society*, pp. 177–180, 1972.

[4] I. J. Bahl and P. Bhartia, *Microstrip Antennas*, Artech House, 1980.

[5] J. R. James, P. S. Hall, and C. Wood, *Microstrip Antenna Theory and Design*, Peter Peregrinus, 1981.

[6] K. R. Carver and J. W. Mink, "Microstrip antenna technology," *IEEE Transactions on Antennas and Propagation*, vol. 1, no. 1, pp. 2–24, 1981.

[7] D. M. Pozar, "Microstrip antennas," *Proceedings of the IEEE*, vol. 80, no. 1, pp. 79–91, 1992.

[8] R. Garg, I. Bhartia, I. Bahl, and A. Ittipiboon, *Micro-Strip Antenna Design Handbook*, Artech House, Boston, Mass, USA, 2001.

[9] D. M. Pozar and B. Kaufman, "Increasing the bandwidth of a microstrip antenna by proximity coupling," *Electronics Letters*, vol. 23, no. 8, pp. 368–369, 1987.

[10] H. F. Pues and A. R. Van de Capelle, "Impedance-matching technique for increasing the bandwidth of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 37, no. 11, pp. 1345–1354, 1989.

[11] A. Kaya and E. Y. Yüksel, "Investigation of a compensated rectangular microstrip antenna with negative capacitor and negative inductor for bandwidth enhancement," *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 5, pp. 1275–1282, 2007.

[12] K. L. Wong and N. H. Hsu, "A broad-band rectangular patch antenna with a pair of wide slits," *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 9, pp. 1345–1347, 2001.

[13] S. I. Latif, L. Shafai, and S. K. Sharma, "Bandwidth enhancement and size reduction of microstrip slot antennas," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 3, pp. 994–1003, 2005.

[14] P. M. T. Ikonen, S. I. Maslovski, C. R. Simovski, and S. A. Tretyakov, "On artificial magnetodielectric loading for improving the impedance bandwidth properties of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 6, pp. 1654–1662, 2006.

[15] R. Cocooli, F. R. Yang, T. P. Ma, and T. Itoh, "Apertu-recoupled patch antenna on UCPBG substrate," *IEEE Transactions on Microwave Theory and Techniaues*, vol. 1, no. 47, pp. 2123–2130, 1999.

[16] L.-M. Si and X. Lv, "CPW-FED multi-band omni-directional planar microstrip antenna using composite metamaterial resonators for wireless communications," *Progress in Electromagnetics Research*, vol. 83, pp. 133–146, 2008.

[17] L. M. Si, H. Sun, Y. Yuan, and X. Lv, "CPW-fed com-pact planar UWB antenna with circular disc and spiral split ring resonators," in *Progress in Electromagnetics Research Symposium*, pp. 502–505, 2009.

[18] Y. Yuan, L.-M. Si, Y. Liu, and X. Lv, "Integrated log-periodic antenna for Terahertz applications," in *International Conference on Microwave Technology and Computational Electromagnetics*, pp. 276–279, 2009.

[19] L. Xin, M. Jinchao, Y. Yong, Y. Weihua, N. Hongbin, and G. Yafen, "Integrated active antennas in terahertz focal plane imaging system," in *Proceedings of the International Workshop on Antenna Technology*, pp. 157–163, 2011.

[20] Q. Xu, W. Zhang, H. Sun, and X. Lu, "Millimetre wave multi-polarised microstrip antenna array and application example," *IET Microwaves, Antennas and Propagation*, vol. 4, no. 10, pp. 1525–1530, 2010.

[21] M. Wei, H. Deng, H. Sun, and Y. Liu, "Design of an X/Ka dual-band co-aperture broadband microstrip antenna array," in *Proceedings IEEE International Conference on Microwave Technology and Computational Electromagnetics*, pp. 217–220, 2011.

[22] Y. Liu, X. Lu, Y. Yuan, Y. Chen, and L. Shi, "Integrated design and research of Ka-band electronically large mono-pulse planar antenna and feed system," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 150–153, October 2006.

[23] Y. Liu, H. J. Sun, X. Lu et al., "Study of the millimeter microstrip monopulse antenna array in the dual-sensor system," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 157–160, October 2006.

[24] Y. An, X. Lv, and B. Gao, "Developing a kind of mi-crostrip array antenna with beam squint," in *Proceedings of the 5th International Symposium on Antennas, Propagation and EM Theory*, pp. 443–446, 2000.

[25] L. Yong, L. Xin, H. Zhao, and Y. Yuan, "Integrated design and research of novel KA-band circular polarized monopulse interrogator array antenna," in *Proceedings of the IET International Radar Conference 2009*, pp. 1–4, April 2009.

[26] H. A. Wheeler, "Fundamental limitations of small an-tennas," *Proceedings of the IRE*, vol. 35, pp. 1479–1484, 1947.

[27] L. J. Chu, "Physical limitations of omni-directional antennas," *Journal of Applied Physics*, vol. 19, no. 12, pp. 1163–1175, 1948.

[28] R. Azadegan and K. Sarabandi, "A novel approach for miniaturization of slot antennas," *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 3, pp. 421–429, 2003.

[29] H. Mosallaei and K. Sarabandi, "Magneto-dielectrics in electromagnetics: concept and applications," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 6, pp. 1558–1567, 2004.

[30] P. M. T. Ikonen, S. I. Maslovski, C. R. Simovski, and S. A. Tretyakov, "On artificial magnetodielectric loading for improving the impedance bandwidth properties of microstrip antennas," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 6, pp. 1654–1662, 2006.

[31] P. M. T. Ikonen, K. N. Rozanov, A. V. Osipov, P. Alitalo, and S. A. Tretyakov, "Magnetodielectric substrates in antenna miniaturization: potential and limitations," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 11, pp. 3391–3399, 2006.

[32] P. Ikonen and S. Tretyakov, "On the advantages of magnetic materials in microstrip antenna miniaturization," *Microwave and Optical Technology Letters*, vol. 50, no. 12, pp. 3131–3134, 2008.

[33] J. Anguera, C. Puente, C. Borja, and J. Soler, "Frac-tal-shaped antennas: a review," in *Wiley: Encyclopedia of RF and Microwave Engineering*, K. Chang, Ed., vol. 2, pp. 1620–1635, 2005.

[34] L. Shi, H. Sun, X. Lu, C. Deng, and Y. Liu, "Study of the improved MM-Wave omni-directional microstrip antenna," in *Proceedings of the 7th International Symposium on Antennas, Propagation and EM Theory (ISAPE '06)*, pp. 238–241, October 2006.

[35] L. Shi, H. Sun, and X. Lu, "A composite antenna with wide circularly polarized beamwidth," *Microwave and Optical Technology Letters*, vol. 51, no. 10, pp. 2461–2464, 2009.

[36] L. Shi, H. Sun, W. Dong, and X. Lv, "A dual-band multi-function carborne hybrid antenna for satellite communication relay system," *Progress in Electromagnetics Research*, vol. 95, pp. 329–340, 2009.

[37] Y. Chen, H. Sun, and X. Lv, "Novel design of dual-polarization broad-band printed L-shaped probe fed microstrip patch antenna for SAR applications," in *IET International Radar Conference 2009*, pp. 1–4, April 2009.

[38] L. M. Si, H. J. Sun, and X. Lv, "Numerical simulations of backward-to-forward leaky-wave antenna with composite right/left-handed coplanar waveguide," *Chinese Physics Letters*, vol. 27, no. 3, Article ID 034106, 2010.

[39] L.-M. Si, H.-J. Sun, and X. Lv, "THz leaky-wave antenna with high-directivity and beam-steering using CPW CRLH metamaterial resonators," in *Proceedings of the 3rd International Symposium on Photoelectronic Detection and Imaging*, Beijing, China, 2009.

[40] Y. Liu, L. M. Si, S. Zhu, and H. Xin, "Experimental realization of an integrated THz electromagnetic crys-tals (EMXT) H-plane horn antenna," *Electronics Letters*, vol. 47, no. 2, pp. 80–82, 2011.

[41] L. M. Si, Y. Liu, S. H. Zhu, and H. Xin, "Integrated THz horn antenna using EBG structures," in *Proceedings of the IEEE International Symposium on Antennas and Propagation and USNC/CNC/URSI National Radio Science Meeting*, University of Colorado at Boulder, 2011.

[42] Y. Liu, Y. Yuan, X. Lv, and L.-M. Si, "Design of 0.5 THz 2D square lattice EBG waveguide transmission line and power-divider using MEMS technology," in *Proceedings 3rd International Symposium on Photoelectronic Detection and Imaging*, vol. 7385, Beijing, China, 2009.

[43] L.-M. Si, Y. Yong, H.-J. Sun, and L. Xin, "Characterization and application of planar terahertz narrow bandpass filter with metamaterial resonators," in *Proceedings of the International*

Workshop on Metamaterials, pp. 351–354, Nanjing, China, 2008.

[44] L. M. Si and X. Lv, “Terahertz waves hairpin microstrip bandpass filter and its application to overlaid dielectric material detection,” *Modern Physics Letters B*, vol. 22, no. 29, pp. 2843–2848, 2008.

[45] L.-M. Si, H.-J. Sun, and X. Lv, “Theoretical investigation of terahertz amplifier by carbon nanotubes within transmission line metamaterials,” *Microwave and Optical Technology Letters*, vol. 53, no. 4, pp. 815–818, 2011.

[46] W. Hong, “Millimeter wave and THz communications in China,” in *Proceedings of the IEEE International Microwave Symposium Digest (MTT ’11)*, pp. 1–4, 2011.

Journal of
Engineering

The Scientific World Journal

International Journal of
Rotating Machinery

Journal of
Sensors

International Journal of
Distributed Sensor Networks

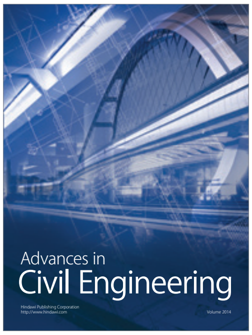
Advances in
Civil Engineering

Hindawi

Submit your manuscripts at
http://www.hindawi.com

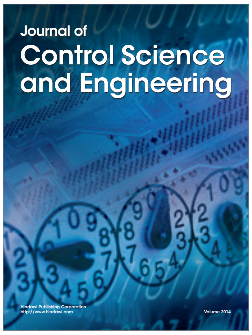
Journal of
Control Science and Engineering

Journal of
Robotics

Journal of
Electrical and Computer Engineering

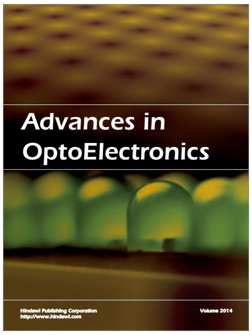
Advances in
OptoElectronics

VLSI Design

International Journal of
Navigation and Observation

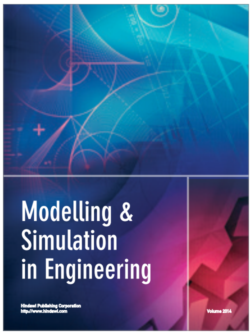
Modelling & Simulation in Engineering

International Journal of
Aerospace Engineering

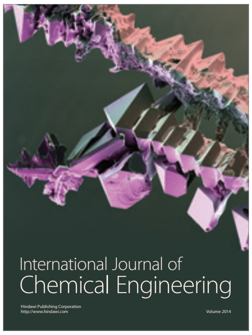
International Journal of
Chemical Engineering

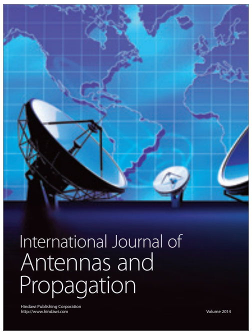
International Journal of
Antennas and Propagation

Active and Passive
Electronic Components

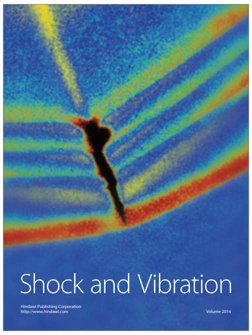
Shock and Vibration

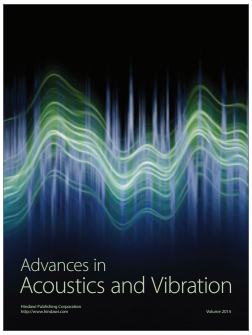
Advances in
Acoustics and Vibration

A 2.4 GHz Polarization-diversity Planar Printed Dipole Antenna for WLAN and Wireless Communication Applications

This article presents the design simulation, fabrication and measured performance of a 2.4 GHz polarization-diversity printed dipole antenna for wireless communication applications. Two orthogonal printed dipole antennas, each with a microstrip via-hole balun for vertical and horizontal polarization, are combined and fabricated on a PCB substrate. PIN diodes are used as switches to select the desired antenna polarization. The 3D finite-element-method (FEM) electromagnetic EM simulator, HFSS, is used in the design simulation of this planar antenna structure. Numerical and measured results of the antenna radiation characteristics, including input SWR, radiation pattern coverage and polarization-diversity, are presented and compared.

In wireless communication systems, such as wireless local area networks (WLAN), research and development efforts are aiming at smaller size and better performance. In addition to the use of signal processing techniques to improve communication channel capacity, the radiation characteristics of the portable antenna system is also very important for communication performance.

In urban or indoor environments, the radio wave will propagate through complicated reflection or scattering processes. The polarization of the radio wave may change significantly. In order to effectively receive the communications signal, a polarization-diversity antenna for wireless communications may become an important requirement. A polarization-diversity antenna may have a pair of linearly-polarized antennas, and the radio signal received on both antenna is sampled and compared at

Huey-Ru Chuang, Liang-Chen Kuo,
Chi-Chang Lin and Wen-Tzu Chen
National Cheng Kung University,
Tainan, Taiwan

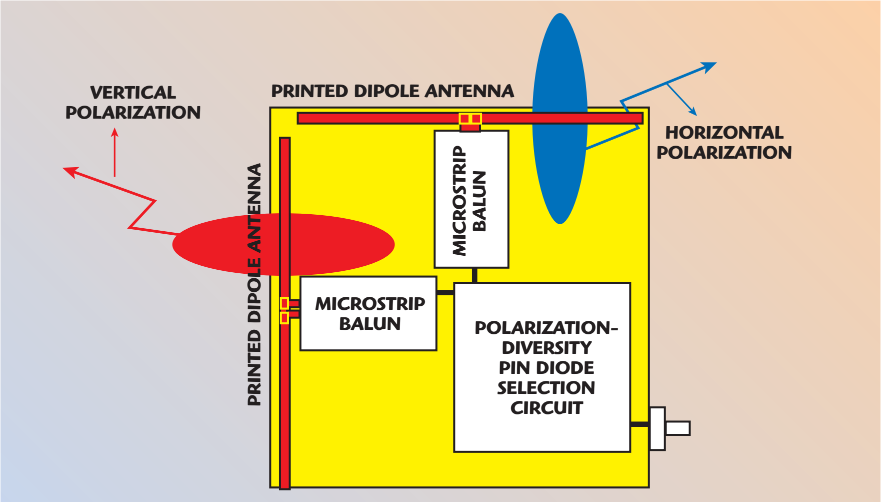
VERTICAL POLARIZATION
PRINTED DIPOLE ANTENNA
HORIZONTAL POLARIZATION
PRINTED DIPOLE ANTENNA
MICROSTRIP BALUN
MICROSTRIP BALUN
POLARIZATION-DIVERSITY PIN DIODE SELECTION CIRCUIT

▲ *Fig. 1 A 2.4 GHz polarization-diversity antenna.*

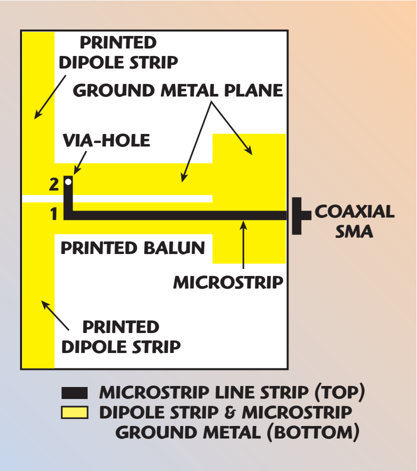
PRINTED DIPOLE STRIP
GROUND METAL PLANE
VIA-HOLE
2
1
COAXIAL SMA
PRINTED BALUN
MICROSTRIP
PRINTED DIPOLE STRIP
■ MICROSTRIP LINE STRIP (TOP)
▭ DIPOLE STRIP & MICROSTRIP GROUND METAL (BOTTOM)

▲ *Fig. 2 Printed dipole antenna with a microstrip balun.*

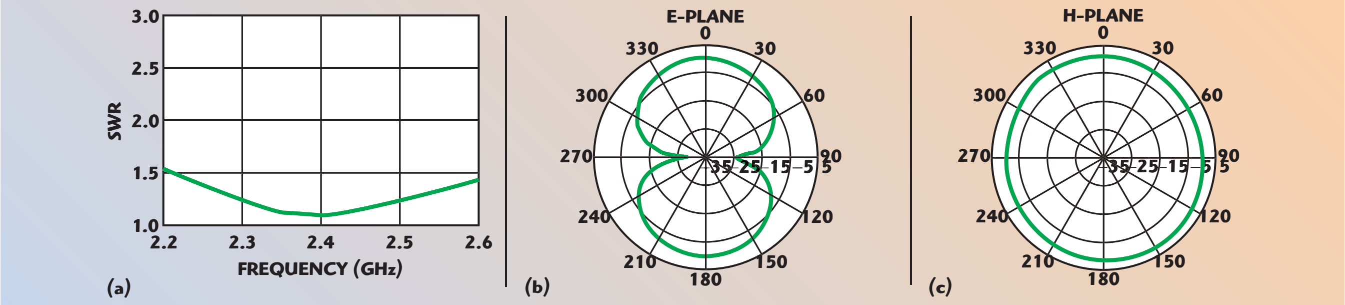

▲ *Fig. 3 Simulated performance of a 2.4 GHz printed dipole antenna placed horizontally; (a) input SWR, (b) E-plane pattern and (c) H-plane pattern.*

certain time intervals. Then the antenna with the best signal quality is selected.

A typical dipole antenna radiates a vertically polarized EM wave and has an omnidirectional antenna pattern. In order to have a preferred planar antenna structure for this 2.4 GHz polarization-diversity antenna, a printed dipole antenna with a microstrip via-hole balun is designed. As shown in ***Figure 1***, two orthogonal printed dipole antennas, for vertical and horizontal polarization, respectively, are combined and fabricated on a PCB substrate. PIN diodes are used to switch and select the desired antenna polarization.

In the antenna design, the high frequency structure simulator (HFSS), based on a 3D FEM, was employed for design simulation of the complete printed dipole structure. A printed dipole antenna and a polarization-diversity planar dipole antenna board (with a polarization-selection PIN diode circuit) have been fabricated on FR-4 PCB substrates. A complete 3D structure FEM simulation and the measured performance of the realized printed dipole-antenna are compared. The measured radiation characteristics of the polarization-diversity planar dipole antenna, including input SWR, radiation pattern coverage and polarization diversity, are presented.

PRINTED DIPOLE ANTENNA WITH MICROSTRIP BALUN

As shown in ***Figure 2***, a printed dipole antenna has a printed microstrip balun which acts as an unbalanced-to-balanced transformer from the feed coaxial line to the two printed dipole strips. The length of the dipole strip and the balun microstrip are both about 1/4 wavelength. The ground plane of the microstrip line and the dipole antenna strips are in the same plane. A via-hole permits the feed point 2 of a printed dipole strip to have the same phase as the feed point 1 of the other printed dipole strip. Due to the 180° phase difference between the top strip and the ground plane of the microstrip line, the feed point 2 of the printed dipole strip will have 180° phase difference with the other feed point 1. Accurate dimensions of the printed dipole strip and the microstrip balun structure are determined by numerical simulation, using HFSS.

The simulation results for a 2.4 GHz printed dipole antenna placed horizontally with a microstrip via-hole balun and fabricated on an FR-4 substrate are shown in ***Figure 3***. The input SWR is less than 1.5 from 2.2 to 2.6 GHz. The simulated E- and H-plane antenna patterns are very close to those of an ideal dipole antenna, where the H-plane pattern is omnidirectional. ***Figure 4*** is a photograph of a realized antenna. The measured input SWR and antenna patterns (measured with the dipole placed vertically) agree well with the simulation results, as shown in ***Figure 5***.

PLANAR POLARIZATION-DIVERSITY PRINTED DIPOLE ANTENNA

Figure 6 shows photographs of a realized 2.4 GHz planar polarization-diversity antenna consisting of two orthogonal printed dipole antennas with a polarization-switched PIN diode circuit. Each printed dipole has

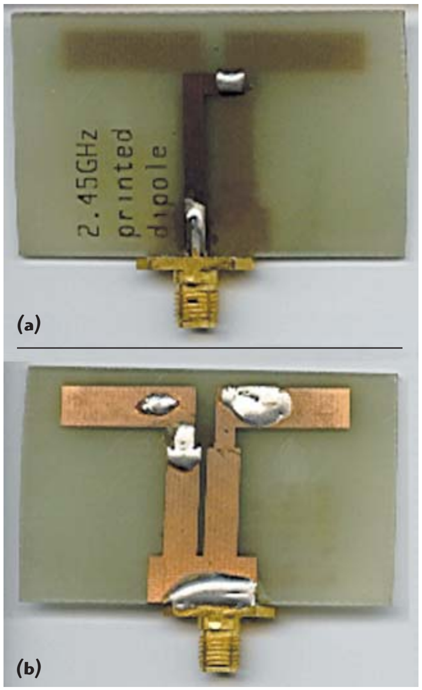

▲ *Fig. 4 A 2.4 GHz printed dipole antenna with a microstrip via-hole balun; (a) top view and (b) bottom view.*

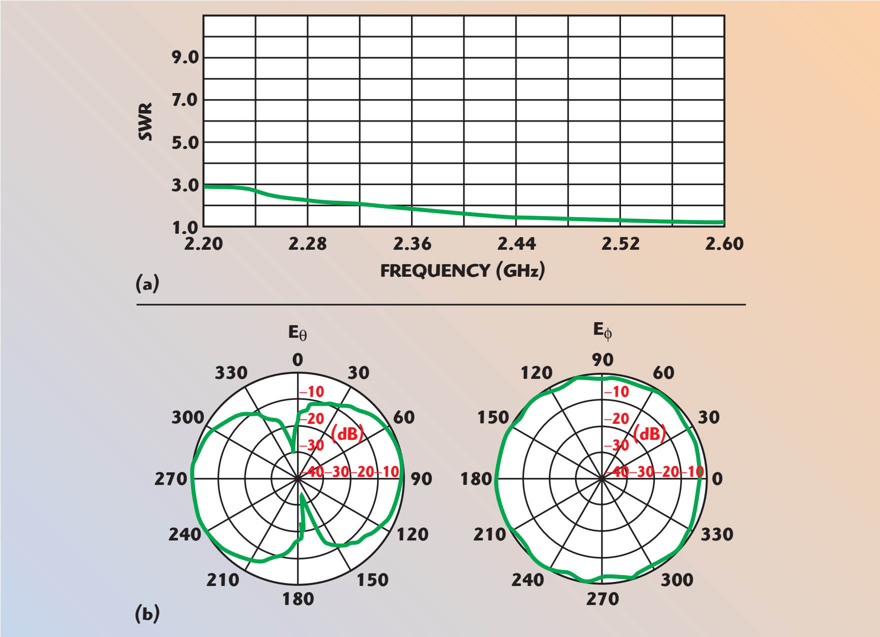

▲ *Fig. 5 Measured input SWR (a) and radiation patterns (b).*

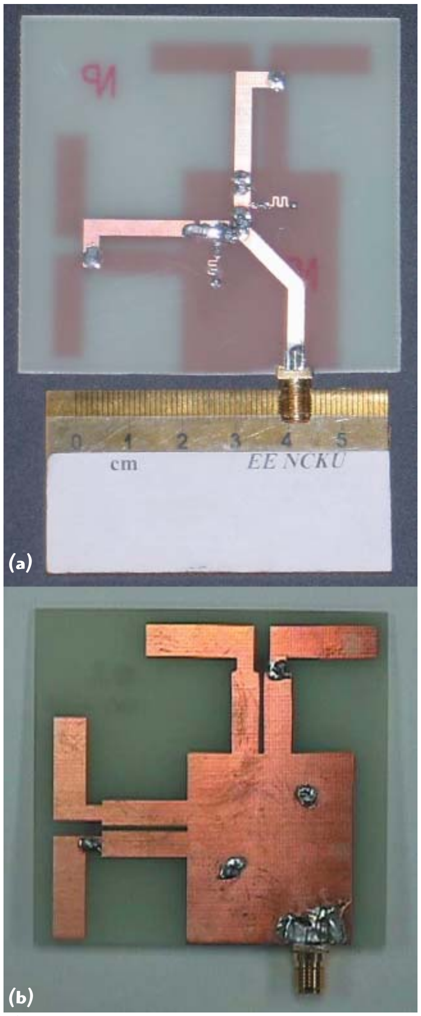

▲ *Fig. 6 A 2.4 GHz planar polarization-diversity antenna with a polarization-switched PIN diode circuit; (a) top view and (b) bottom view.*

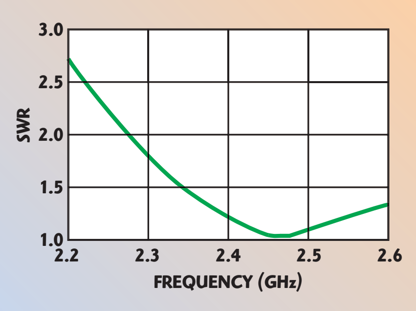

▲ *Fig. 7 Input SWR simulation of a 2.5 GHz polarization-diversity dipole antenna with the vertical dipole selected.*

a microstrip via-hole balun. The terminals of the two baluns are connected to a PIN diode selection circuit. Voltages from the transceiver circuit (±5.0V) are fed through a cable to the input of the PIN diode circuit section, to short or open-circuit the PIN diodes. Hence, either the vertical or horizontal printed dipole can be selected and connected to the transceiver.

Since the two dipoles are very close to each other and near the PIN diode circuit section, EM coupling will degrade the performance of each dipole. ***Figure 7*** shows the input SWR simulation results with the vertical dipole antenna selected (+5V to PIN diode switching circuit). The input SWR is less than 2 from 2.25 to 2.60 GHz. The simulated E- and H-plane antenna patterns are all very close to those of an ideal dipole antenna, of which the H-plane pattern is still omnidirectional, as shown in ***Figure 8***. Note that the dominant polarization is the vertical (E_θ) field, which agrees with the selection of the vertical dipole. The antenna pattern has some attenuation in the direction of the PIN diode circuit board. It can also be seen that a certain level of the input RF signal is induced to the horizontal antenna path by EM coupling, which generates some level of cross-polarization field. ***Figure 9*** shows the simulation results with the horizontal dipole antenna selected (–5V to PIN diode switching circuit). Results similar to the ones obtained for the vertical dipole antenna can be observed, except that the dominant polarization is the horizontal (E_ϕ) field, which agrees with the selection of the horizontal dipole.

The measured antenna input SWR with vertical or horizontal dipole selection confirms the input SWR of each dipole antenna (through the PIN diode selection circuit) is less than 1.5 from 2.2 to 2.6 GHz, which agrees with the HFSS simulation results. The measured antenna patterns with the selection of the vertical or horizontal dipole shows that for the selection of the vertical dipole, the H-plane pattern is still quite omnidirectional (as an ideal vertical dipole) with some attenuation in the direc-

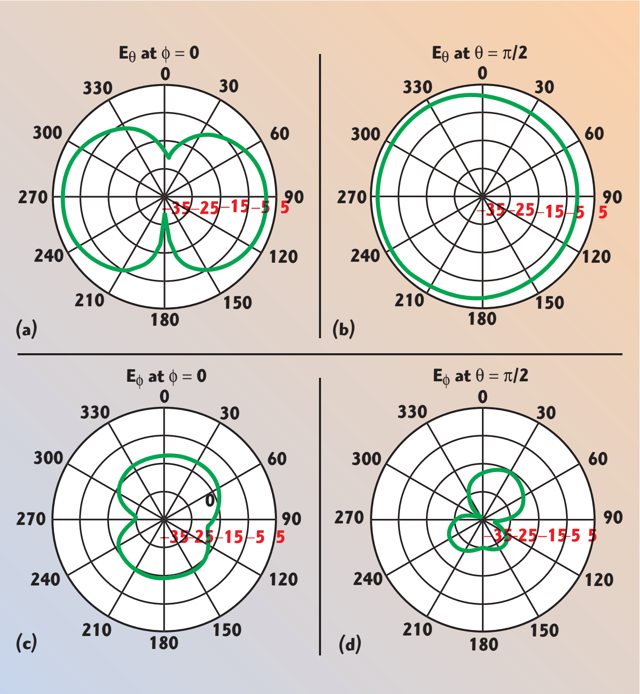

▲ *Fig. 8 Simulation of a 2.5 GHz polarization-diversity dipole antenna with the vertical dipole selected; (a) E_θ-field E-plane pattern, (b) E_θ-field H-plane pattern, (c) E_ϕ-field (cross-polarization) E-plane pattern and (d) E_ϕ-field (cross-polarization) H-plane pattern.*

tion of the PIN diode circuit board. ***Figures 10*** and ***11*** show the mesured SWR and antenna patterns, respectively. A certain level of the induced cross-polarization pattern is observed as predicted by the HFSS simulation due to the proximity of the horizontal dipole strip and the PIN diode circuit board. As for the selection of the horizontal dipole, the E-plane pattern is also close to that of an ideal horizontal dipole. Also, the induced cross-polarization pattern is observed, which is the same situation as the selection of the vertical dipole. The measured data shows a good agreement with the HFSS simulation results and how the antenna polarization-diversity is working.

CONCLUSION

3D FEM design simulation, realization and measurements of a 2.4 GHz printed dipole antenna (with a microstrip via-hole balun) and a planar polarization-diversity printed dipole antenna are presented. The planar polarization-diversity antenna consists of two orthogonal printed dipole antennas for vertical and horizontal polarization and is fabricated on a FR-4 PCB board. A PIN diode switching circuit is used to select the desired antenna polarization. Satisfactory agreement between simulation and measurements is observed. The measured input SWR of the realized printed dipole antenna is less than 1.5 from 2.2 to 2.6 GHz. The measured input SWR of the vertical and horizontal dipole (through the PIN diode switching circuit) of the realized planar polarization-diversity antenna is less than 1.5 from 2.3 to 2.6 GHz. The measured E- and H-plane patterns of the polarization-diversity antenna show that the selected vertical or horizontal dipole have a performance close to a single dipole antenna in a vertical or horizontal position. The designed planar polarization-diversity antenna can be used for wireless communication and WLAN applications.

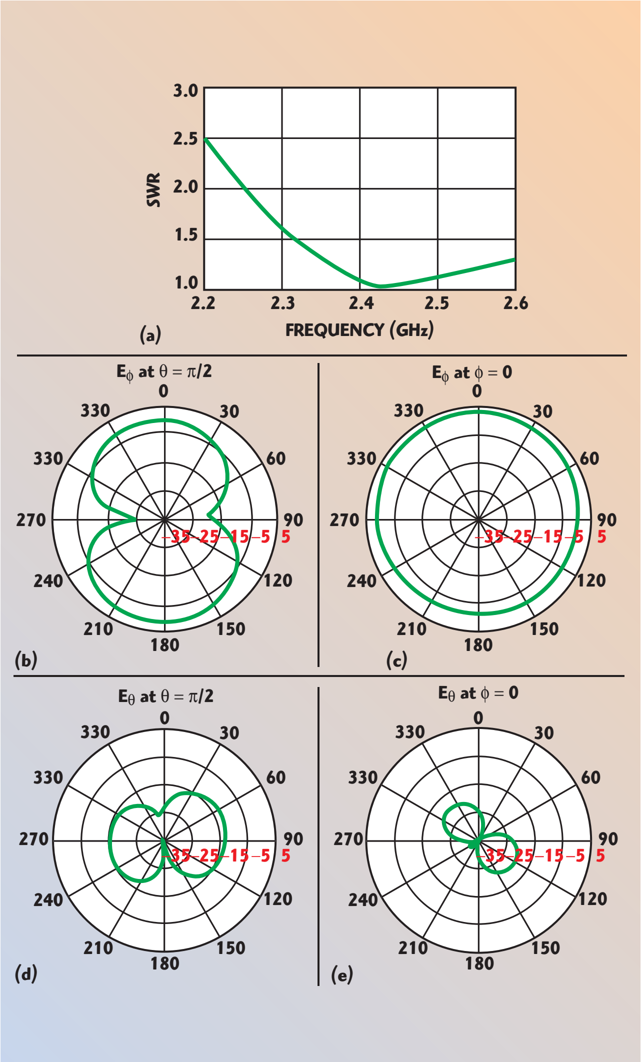

▲ *Fig. 9 Simulation of a 2.4 GHz polarization-diversity printed dipole antenna (with the horizontal dipole selected); (a) input SWR, (b) E_ϕ-field E-plane pattern, (c) E_ϕ-field H-plane pattern, (d) E_θ-field (cross-polarization) E-plane pattern and (e) E_θ-field (cross-polarization) H-plane pattern.*

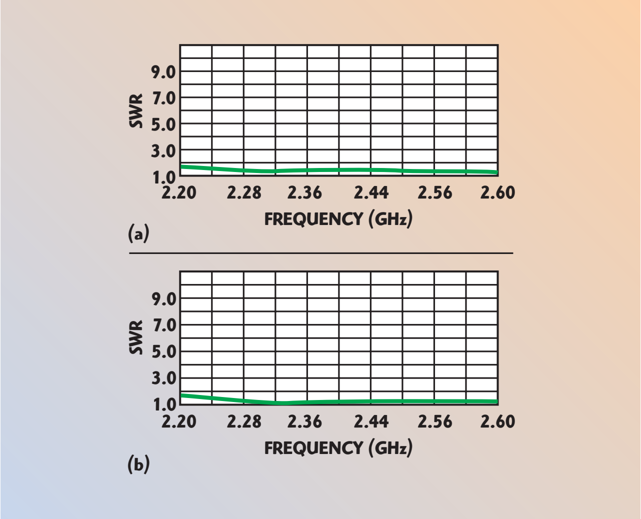

▲ *Fig. 10 Measured input SWR of a 2.4 GHz polarization-diversity printed dipole antenna; (a) vertical dipole selection and (b) horizontal dipole section.*

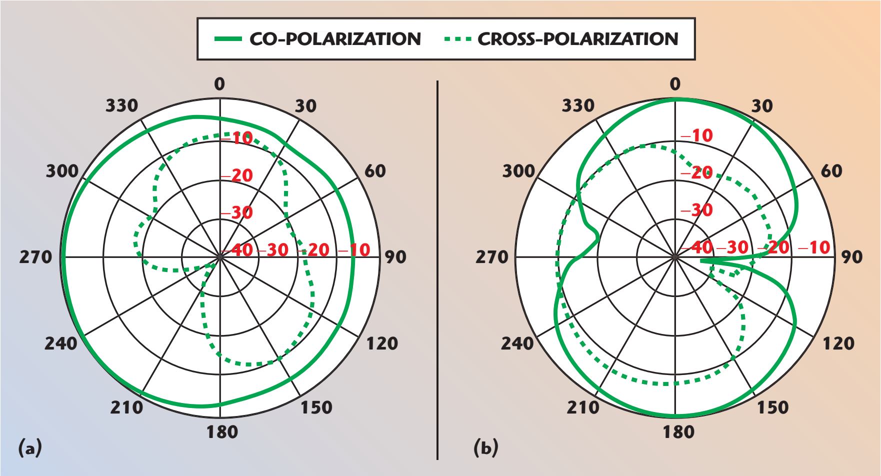

▲ *Fig. 11 Measured co-and cross-polarized patterns of a 2.4 GHz polarization-diversity printed dipole antenna; (a) vertical dipole selected and (b) horizontal dipole selected.*

ACKNOWLEDGMENT

The authors would like to thank Ansoft Inc. for its support of the HFSS software. ■

References

1. K. Fujimoto and J. R. James, *Mobile Antenna Systems Handbook*, Artech House Inc., Norwood, MA 1994.
2. B. Edward and D. Rees, "A Broadband Printed Dipole with Integrated Balun," *Microwave Journal*, May 1987, pp. 339–344.
3. K. Hettak, G.Y. Delisle and M.G. Stubbs, "A Novel Variant of Dual Polarized CPW Fed Patch Antenna for Broadband Wireless Communications," *IEEE Antennas and Propagation Society International Symposium Digest*, Vol.1, 2000, pp. 286–289.
4. L. Zhu and K. Wu, "Model-based Characterization of CPS-fed Printed Dipole for Innovative Design of Uniplanar Integrated Antenna," *IEEE Microwave and Guided Wave Letters*, Vol. 9, No. 9, September 1999, pp. 342–344.
5. N. Michishitai and H. Arai, "A Polarization-diversity Antenna by Printed Dipole and Patch With a Hole," *IEEE Antennas and Propagation Society International Symposium Digest*, 2001, pp. 368–371.

Huey-Ru Chuang *received his BSEE and MSEE degrees from National Taiwan University, Taipei, Taiwan, in 1977 and 1980, respectively, and his PhD degree in electrical engineering from Michigan State University, East Lansing, MI, in 1987. From 1987 to 1988, he was a post-doctoral research associate at the Engineering Research Center of Michigan State University. From 1988 to 1990, he was with the Portable Communication Division of Motorola Inc., Ft. Lauderdale, FL. He joined the department of electrical engineering of National Cheng Kung University, Tainan, Taiwan, in 1991, where he is currently a professor. His research interests include portable antenna design, RF/microwave circuits and RFIC/MMIC for wireless communications, electromagnetic computation of human interaction with mobile antennas, EMI/EMC, microwave communication and detection systems.*

Liang-Chen Kuo *received his BSEE degree from Nan-Tai Institute of Technology, Tainan, Taiwan, and his MSEE degree from Tatung Institute of Technology, Taipei, Taiwan, in 1987 and 1996, respectively. He is currently working toward his PhD degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan. His research interests include computational electromagnetics and antenna design.*

Chi-Chang Lin *received his BSEE and MSEE degrees from Tatung Institute of Technology, Taipei, Taiwan, in 1999 and 2001, respectively. He is currently working toward his PhD degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan. His research interests include EM simulation and microwave antenna design.*

Wen-Tzu Chen *received his PhD degree from National Cheng Kung University, Tainan, Taiwan, in 1998. He is currently an assistant professor at the Institute of Computer and Communication, Shu-Te University, Yen Chau, Taiwan. His research interests include numerical computation of EM interaction between the antenna and the human body, and microwave antenna design.*

射频和天线设计培训课程推荐

易迪拓培训(www.edatop.com)由数名来自于研发第一线的资深工程师发起成立，致力并专注于微波、射频、天线设计研发人才的培养；我们于 2006 年整合合并微波 EDA 网(www.mweda.com)，现已发展成为国内最大的微波射频和天线设计人才培养基地，成功推出多套微波射频以及天线设计经典培训课程和 ADS、HFSS 等专业软件使用培训课程，广受客户好评；并先后与人民邮电出版社、电子工业出版社合作出版了多本专业图书，帮助数万名工程师提升了专业技术能力。客户遍布中兴通讯、研通高频、埃威航电、国人通信等多家国内知名公司，以及台湾工业技术研究院、永业科技、全一电子等多家台湾地区企业。

易迪拓培训推荐课程列表： http://www.edatop.com/peixun/tuijian/

射频工程师
养成培训课程套装
易迪拓培训
www.edatop.com

射频工程师养成培训课程套装

该套装精选了射频专业基础培训课程、射频仿真设计培训课程和射频电路测量培训课程三个类别共 30 门视频培训课程和 3 本图书教材；旨在引领学员全面学习一个射频工程师需要熟悉、理解和掌握的专业知识和研发设计能力。通过套装的学习，能够让学员完全达到和胜任一个合格的射频工程师的要求…

课程网址：http://www.edatop.com/peixun/rfe/110.html

手机天线设计培训视频课程

该套课程全面讲授了当前手机天线相关设计技术，内容涵盖了早期的外置螺旋手机天线设计，最常用的几种手机内置天线类型——如 monopole 天线、PIFA 天线、Loop 天线和 FICA 天线的设计，以及当前高端智能手机中较常用的金属边框和全金属外壳手机天线的设计；通过该套课程的学习，可以帮助您快速、全面、系统地学习、了解和掌握各种类型的手机天线设计，以及天线及其匹配电路的设计和调试...

手机天线设计培训
中文视频课程
主讲：李明洋
易迪拓培训

课程网址： http://www.edatop.com/peixun/antenna/133.html

企业内训课程
WiFi天线和射频电路设计培训

WiFi 和蓝牙天线设计培训课程

该套课程是李明洋老师应邀给惠普 (HP)公司工程师讲授的 3 天员工内训课程录像，课程内容是李明洋老师十多年工作经验积累和总结，主要讲解了 WiFi 天线设计、HFSS 天线设计软件的使用，匹配电路设计调试、矢量网络分析仪的使用操作、WiFi 射频电路和 PCB Layout 知识，以及 EMC 问题的分析解决思路等内容。对于正在从事射频设计和天线设计领域工作的您，绝对值得拥有和学习！…

课程网址：http://www.edatop.com/peixun/antenna/134.html

CST 学习培训课程套装

示意图
DVD
CST视频培训课程套装
李明洋主讲
易迪拓培训
www.edatop.com

该培训套装由易迪拓培训联合微波 EDA 网共同推出，是最全面、系统、专业的 CST 微波工作室培训课程套装，所有课程都由经验丰富的专家授课，视频教学，可以帮助您从零开始，全面系统地学习 CST 微波工作的各项功能及其在微波射频、天线设计等领域的设计应用。且购买该套装，还可超值赠送 3 个月免费学习答疑…

课程网址：http://www.edatop.com/peixun/cst/24.html

HFSS 学习培训课程套装

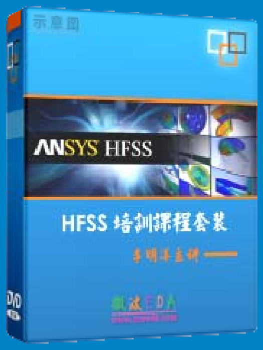
示意图
ANSYS HFSS
HFSS培训课程套装
李明洋主讲
微波EDA
DVD

该套课程套装包含了本站全部 HFSS 培训课程，是迄今国内最全面、最专业的 HFSS 培训教程套装，可以帮助您从零开始，全面深入学习 HFSS 的各项功能和在多个方面的工程应用。购买套装，更可超值赠送 3 个月免费学习答疑，随时解答您学习过程中遇到的棘手问题，让您的 HFSS 学习更加轻松顺畅…

课程网址：http://www.edatop.com/peixun/hfss/11.html

ADS 学习培训课程套装

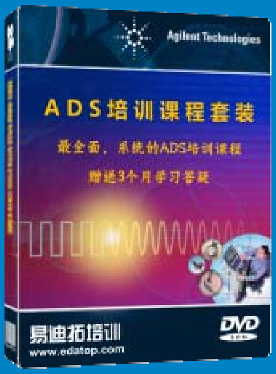
Agilent Technologies
ADS培训课程套装
最全面、系统的ADS培训课程
赠送3个月学习答疑
易迪拓培训
www.edatop.com
DVD

该套装是迄今国内最全面、最权威的 ADS 培训教程，共包含 10 门 ADS 学习培训课程。课程是由具有多年 ADS 使用经验的微波射频与通信系统设计领域资深专家讲解，并多结合设计实例，由浅入深、详细而又全面地讲解了 ADS 在微波射频电路设计、通信系统设计和电磁仿真设计方面的内容。能让您在最短的时间内学会使用 ADS，迅速提升个人技术能力，把 ADS 真正应用到实际研发工作中去，成为 ADS 设计专家...

课程网址： http://www.edatop.com/peixun/ads/13.html

我们的课程优势：

※ 成立于 2004 年，10 多年丰富的行业经验，

※ 一直致力并专注于微波射频和天线设计工程师的培养，更了解该行业对人才的要求

※ 经验丰富的一线资深工程师讲授，结合实际工程案例，直观、实用、易学

联系我们：

※ 易迪拓培训官网：http://www.edatop.com

※ 微波 EDA 网：http://www.mweda.com

※ 官方淘宝店：http://shop36920890.taobao.com

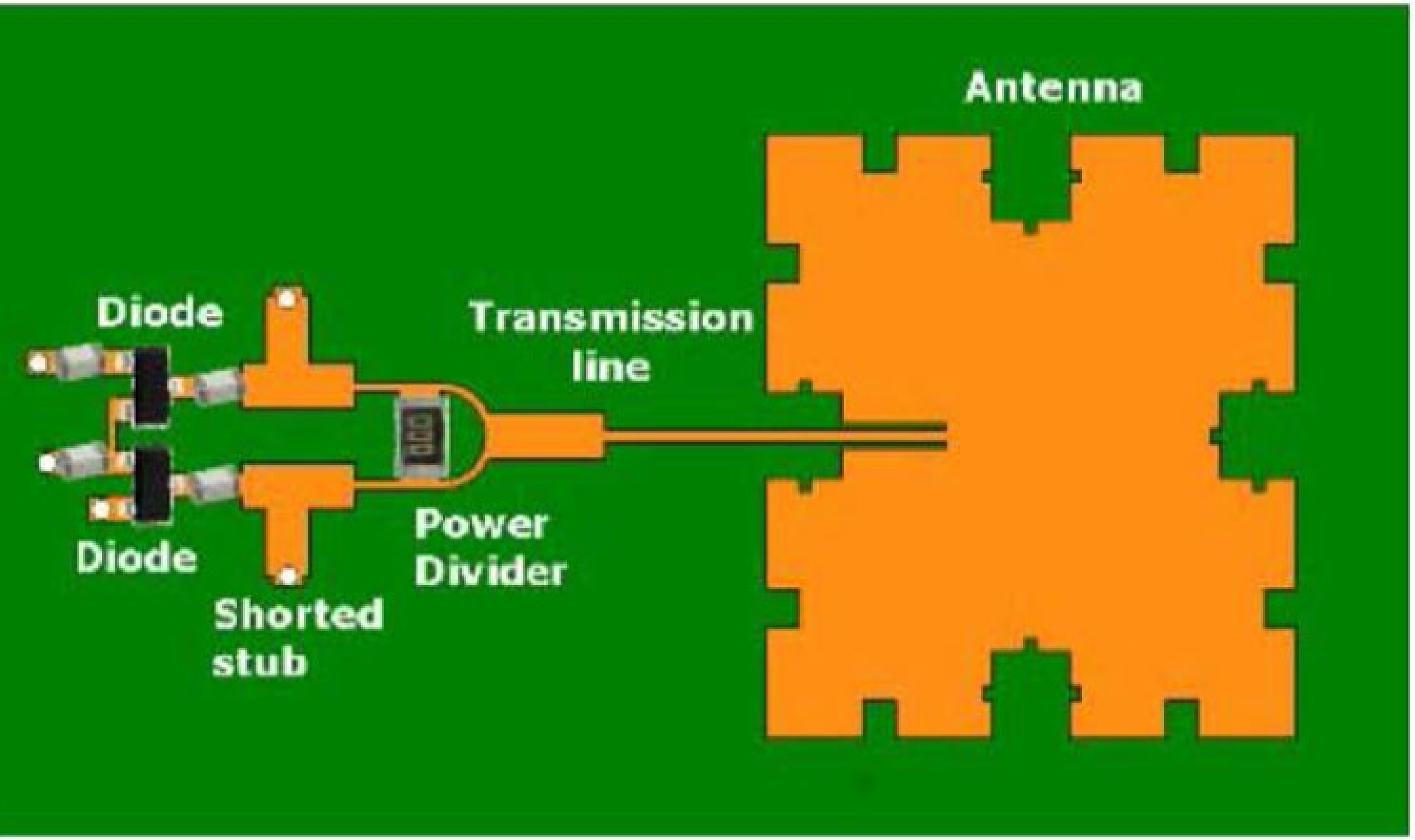
Antenna
Diode
Transmission line
Diode
Power Divider
Shorted stub

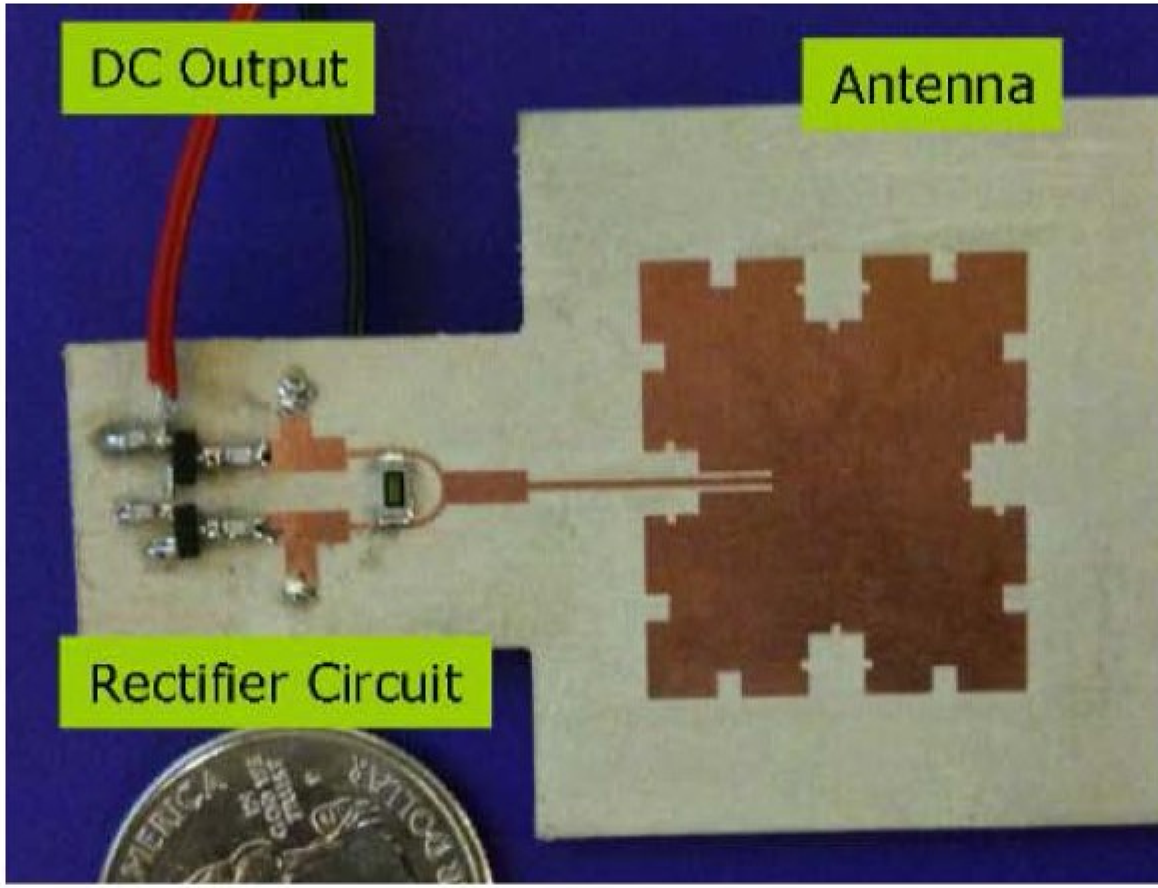
DC Output
Antenna
Rectifier Circuit

Giedrius Rakauskas

Follow

Phased Patch Antenna Array Design

Engineering

1473 17 0

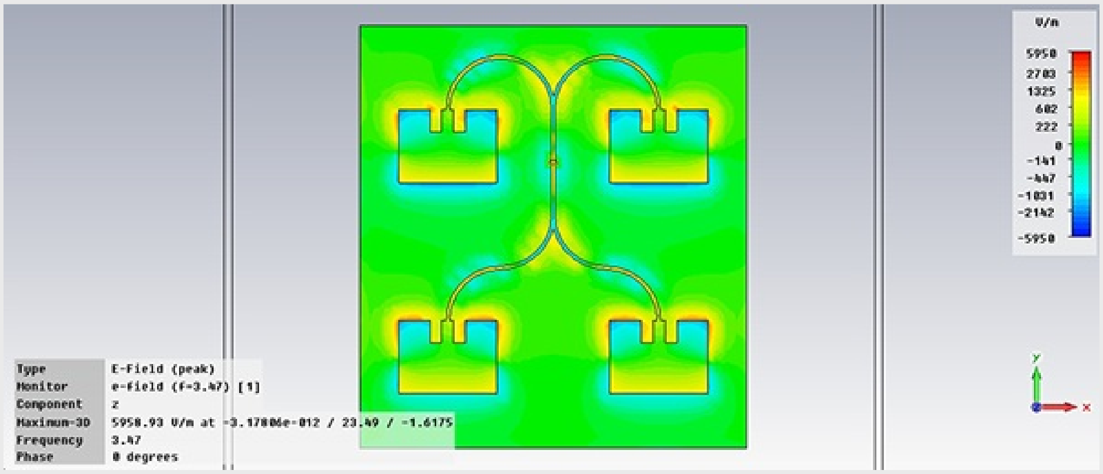

Type	E-Field (peak)
Monitor	e-field (f=3.47) [1]
Component	z
Maximum-3D	5950.93 V/m at -3.17806e-012 / 23.49 / -1.6175
Frequency	3.47
Phase	0 degrees

Third Year Project

The aim of this project was to design a directive phased patch antenna array that could be implemented in modern vehicles for distance measurement and autonomous cruise control.

The particular technology allows improved performance as well as reduced costs and size.

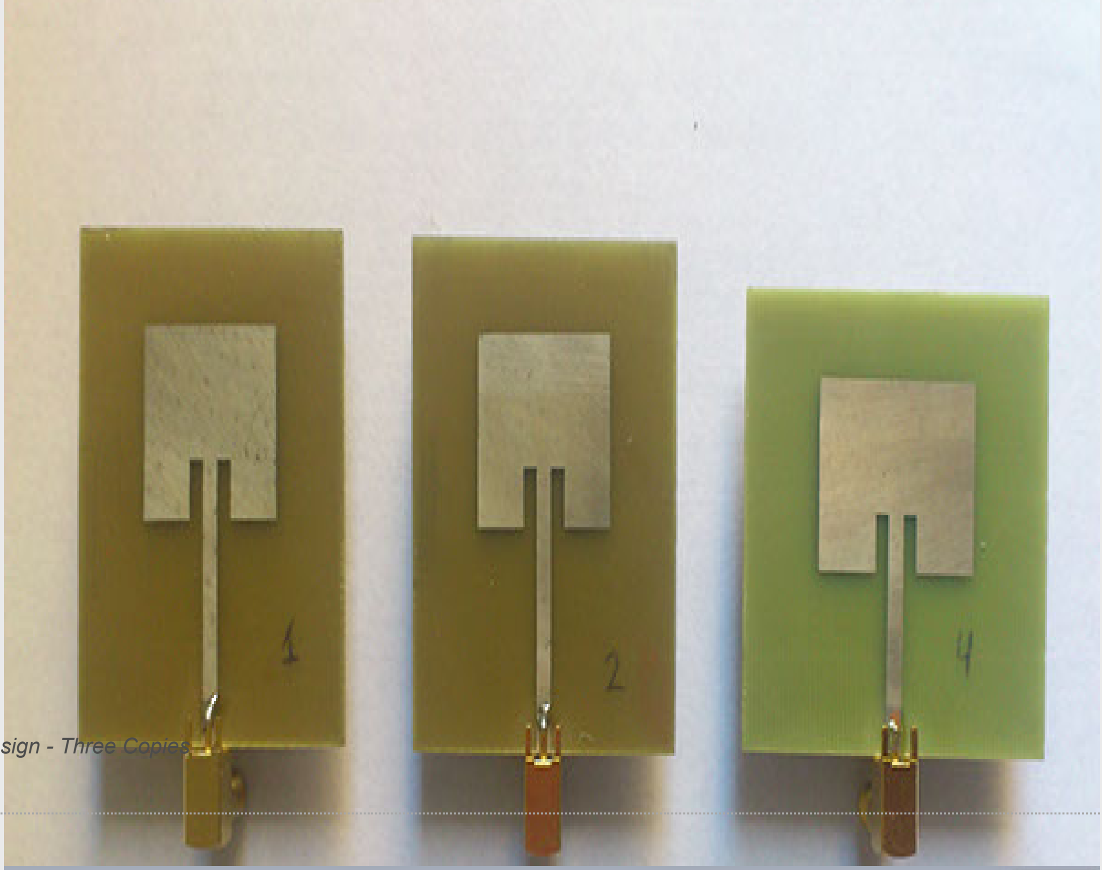

Initial Design - Three Copies

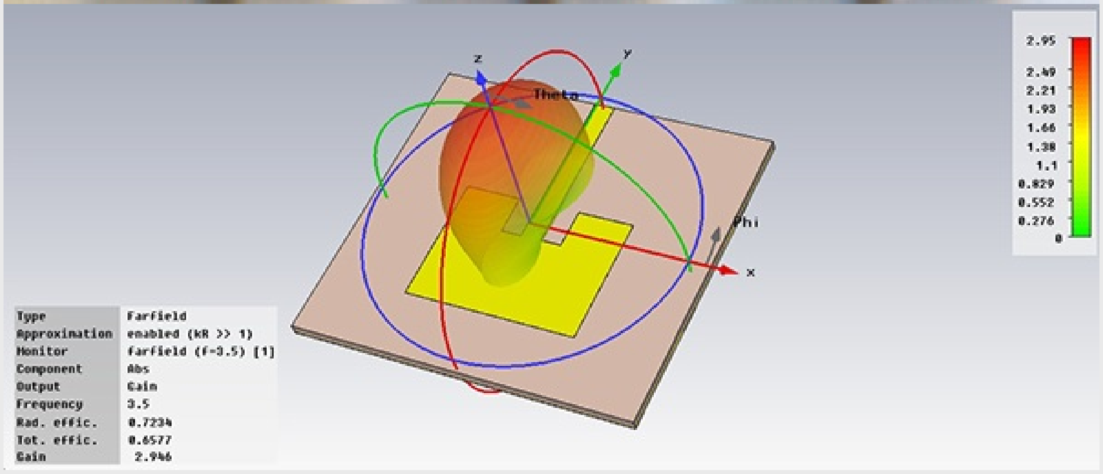
z
y
Theta
Phi
x
2.95
2.49
2.21
1.93
1.66
1.38
1.1
0.829
0.552
0.276
0

Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=3.5) [1]
Component	Abs
Output	Gain
Frequency	3.5
Rad. effic.	0.7234
Tot. effic.	0.6577
Gain	2.946

Square Patch Antenna - Farfield Simulation Results

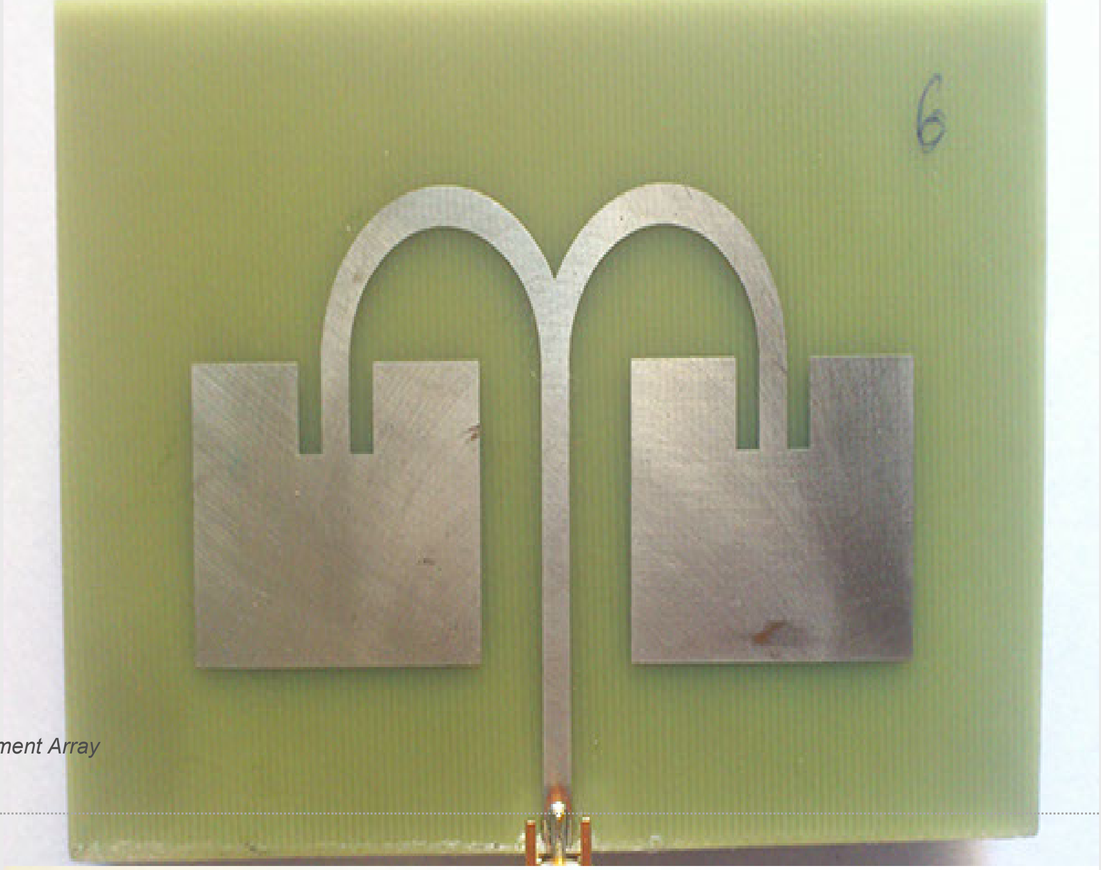

Two-Element Array

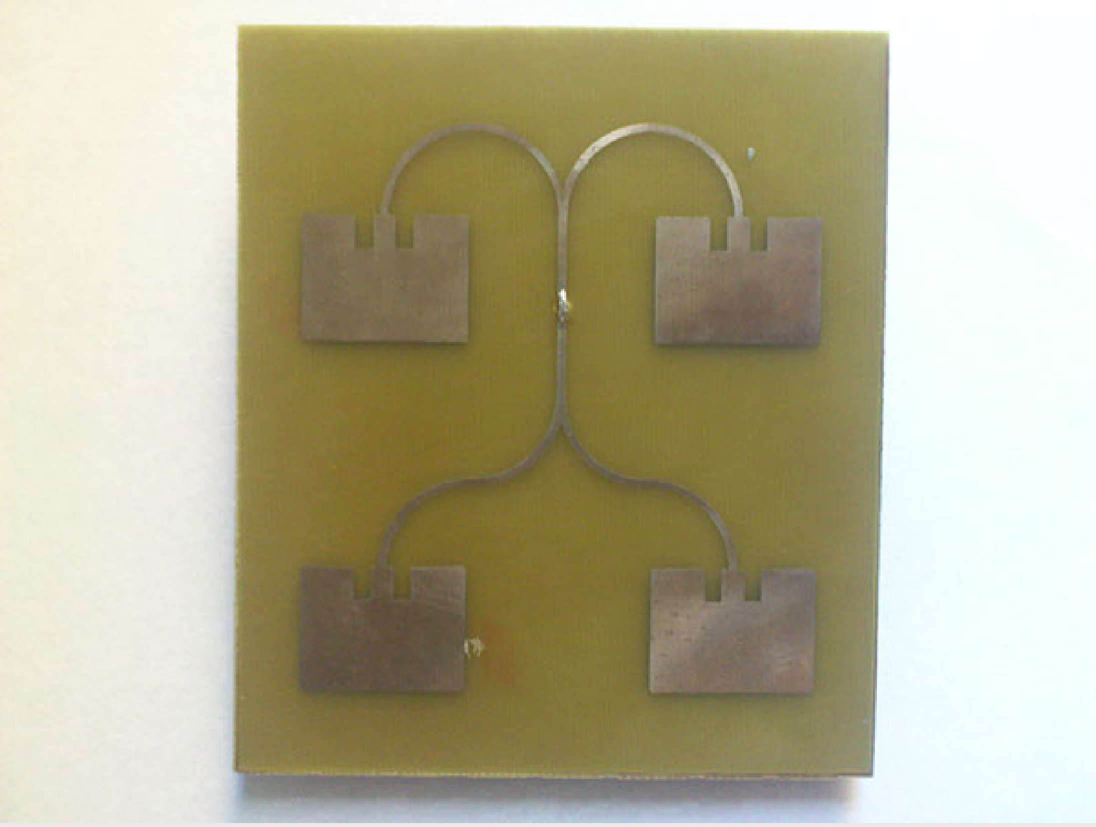

Four-Element Array

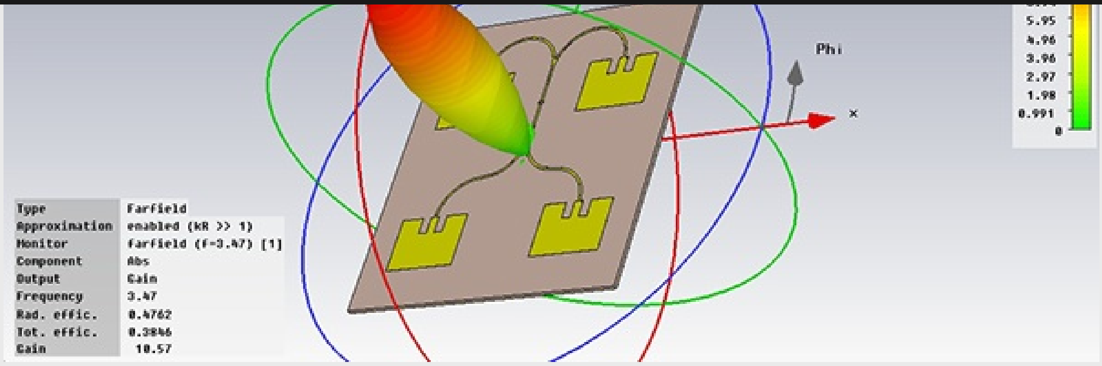

5.95
4.96
3.96
2.97
1.98
0.991
0
Phi
x

Type	Farfield
Approximation	enabled (kR >> 1)
Monitor	farfield (f=3.47) [1]
Component	Abs
Output	Gain
Frequency	3.47
Rad. effic.	0.4762
Tot. effic.	0.3846
Gain	10.57

Four-Element Array - CST Microwave Studio Simulation Results

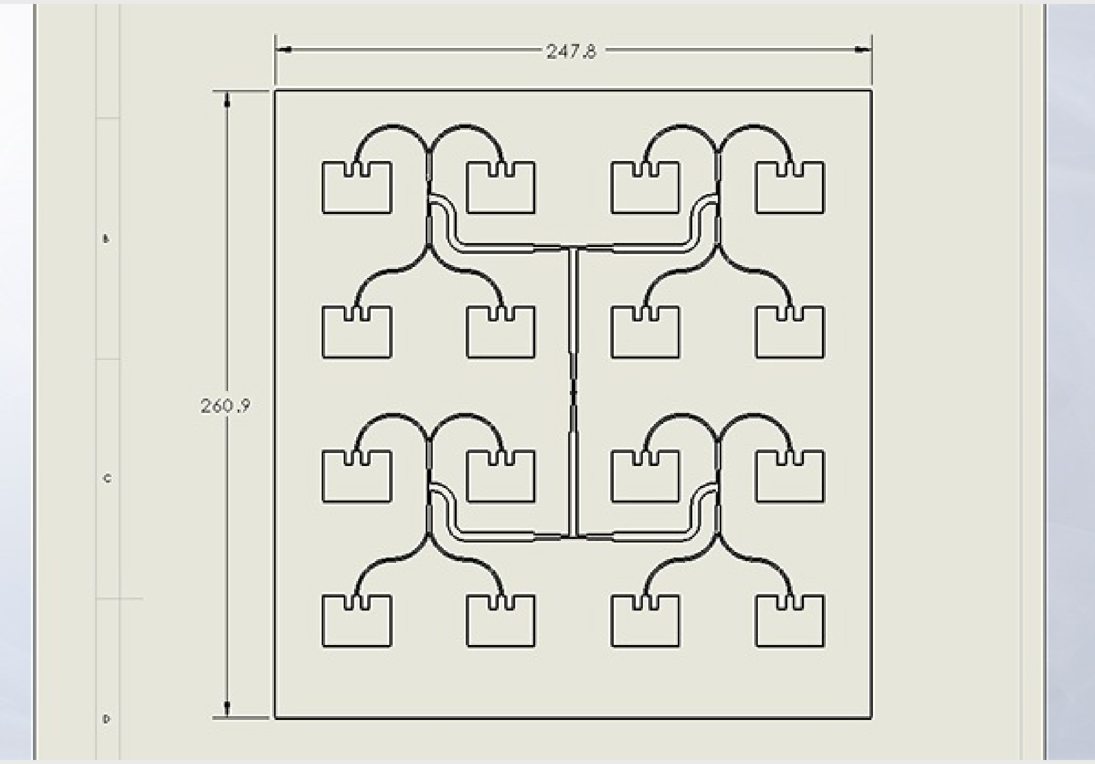

247.8
260.9
B
C
D

Sixteen-Element Array Design

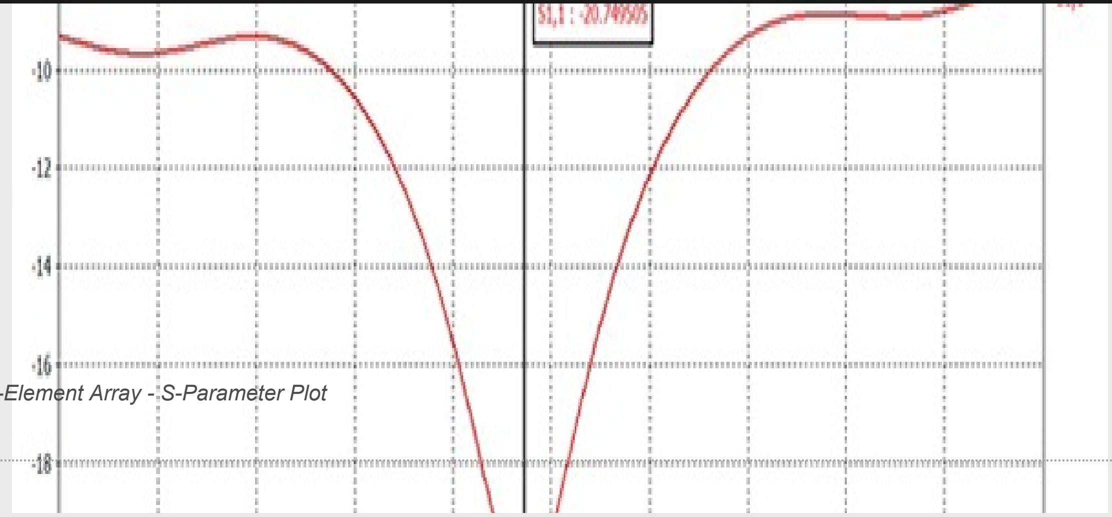

Sixteen-Element Array - S-Parameter Plot

Giedrius Rakauskas

Vilnius, Lithuania

Follow

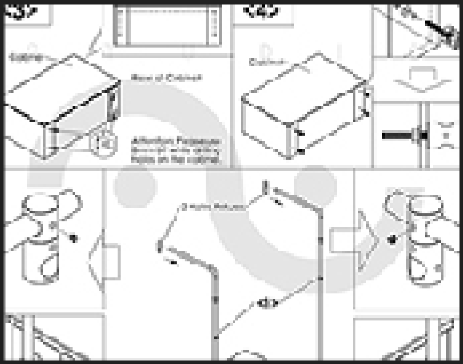

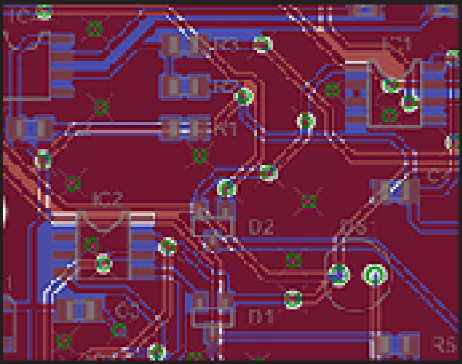

Behance

View Complete Profile →

Comments

Basic Info

Third Year BEng Project - Phased Patch Antenna Array

Credits

Giedrius Rakauskas
Vilnius, Lithuania

+ Follow

Tags

- patch antenna
- antenna array
- square patch antenna
- cst mws
- cst microwave studio

Tools Used

CST Microwave Studio, FR-4

Report

Bēhance

About Behance & Careers Adobe Portfolio Blog Behance API Portfolio Review Week
Creative Career Tips Apps

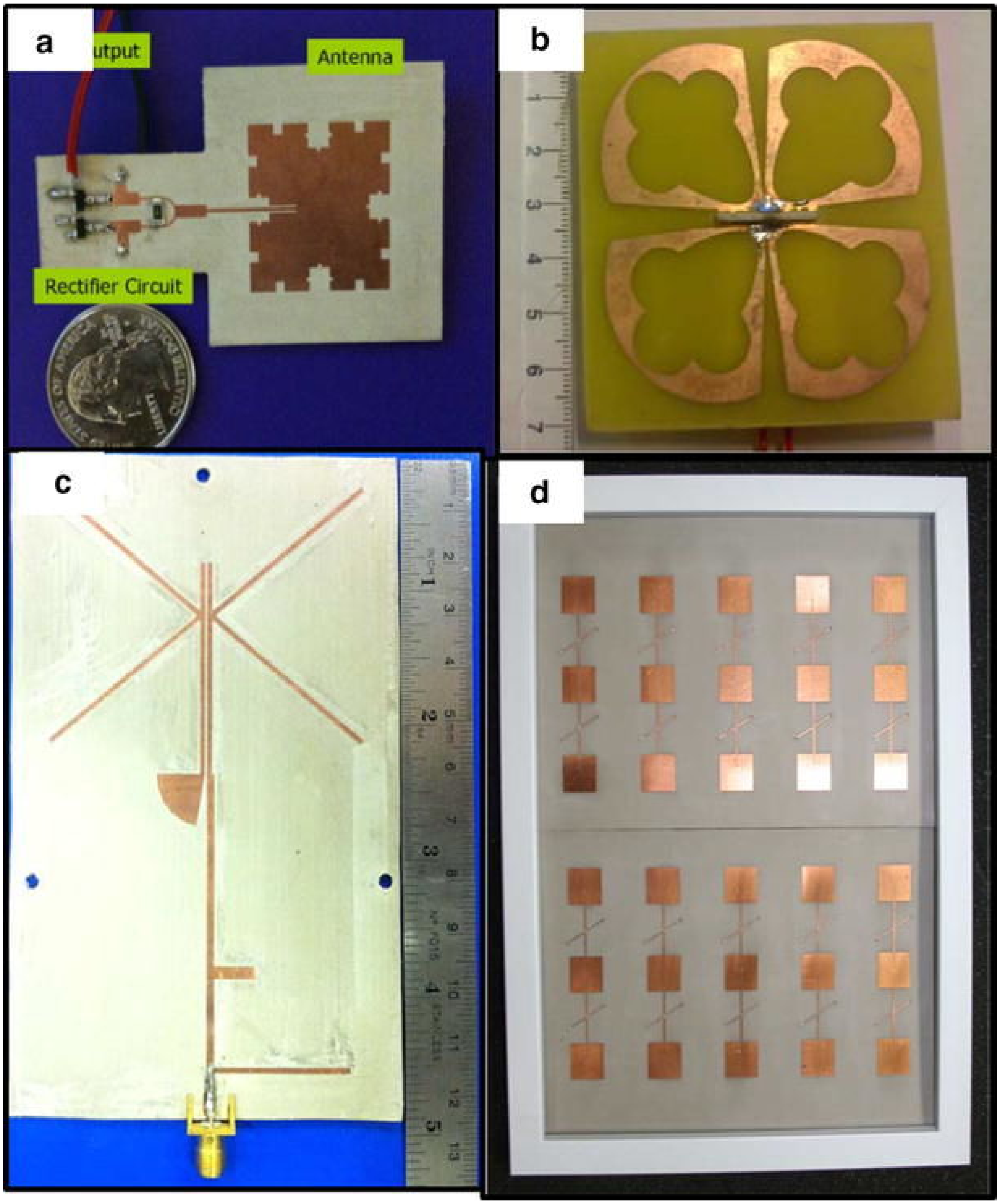

Fig. 4

Types of antennas **a** 2.45 GHz patch antenna with rectifier [38]. ©2010 IEEE. All rights reserved; **b** Planar dual-polarized antenna [24]. ©2015 IEEE. All rights reserved; **c** Microstrip antenna [37]. ©2012 IEEE. All rights reserved; **d** Array of stacked differential patch antenna [33]. ©2015 IEEE. All rights reserved

The plate antennas are popular and have many applications [27, 34, 38]; on-chip antennas are preferred for small and compact applications. Recently, many publications addressed wide-band and multi-band antennas. It has been proven that narrow-band antennas offer high energy conversion efficiencies but can only retrieve a limited amount of energy. On the other hand, wide-band or multi-band frequency antennas can retrieve more RF energy in space. However, the tradeoffs are low overall efficiency and large aperture. In [32], antennas with a resonance frequency of 4.9 and 5.9 GHz were designed with PCEs of 65.2 and 64.8%, respectively. Further work by Lu et al. [26] on polarization antennas supports the assertion that expanding the bandwidth of an antenna leads to increasing the amount of power harvested. In this work, the demonstration of broadband polarization antennas with three separate modes allows the antenna to operate in a wider range of frequencies. One common mechanism in the aforementioned works is the control of the antenna configuration by switching the diodes on and off, thus altering its resonant frequencies. However, since it uses separate modes for different frequencies, this antenna is not able to simultaneously resonate at two frequencies. On the other hand, the antenna presented in [39] is capable of operating at 2.45 and 5.8 GHz, simultaneously, providing 2.6 V output with a PCE of 65% and power density of 10 mW/cm^2.

The main purpose of aligning antennas in arrays is to enhance the antenna gain and obtain high voltage/current. Array antennas are preferred over large aperture antennas because they do not require large breakdown voltage diodes to operate. Antenna arrays can be connected before or after rectification. The first configuration enhances the retrieved power at the main beam while the second configuration expands the ability to retrieve power from various angles away from the main beam [40]. In case the RF waves are combined before rectification, the rectifier requires a large breakdown diode. If RF waves are combined after rectification, combining DC current becomes problematic. Antenna arrays can be connected in series or parallel to obtain high voltage or large current. Nonetheless, expanding the arrays yields better output but this might cause a deduction in conversion efficiency [33].

For demonstrating antenna arrays, Sun et al. [35] invented a T-junction to connect four quasi-Yagi antennas together. The advancement in this work was that the T-junction was flexible in changing from 1 × 4 array to 2 × 2 array topologies. Consequently, the system was able to operate at an ambient power level as low as 455 μW/cm^2 while obtaining 40% PCE.

Impedance matching network

In low-power consumption electrical systems, power leakage during transmission may lead to energy insufficiency. In these circumstances, adding an impedance matching network (IMN) ensures that the maximum power transfers between the RF source and load. For WPH applications, the receiving antenna is considered as the source while the rectifier/voltage multiplier is considered as the load. It is acknowledged that in DC, power transfer is optimum when the resistances of the source and load are indistinguishable. In an RF circuit, the impedance is referred to instead of resistance. An impedance mismatch between the source and load creates reflected power flow in the circuit that lowers the efficiency of the system. As its name indicates, the IMN ensures that the impedance of the source and load are identical by adding reactive components in between.

There are three basics matching configurations i.e. L, T and π matching networks (Fig. 5). The L matching is commonly used since it typically has two components, which simplifies the designing and controlling process. Additionally, the L matching networks do not alter the quality factor (Q) of the circuit.

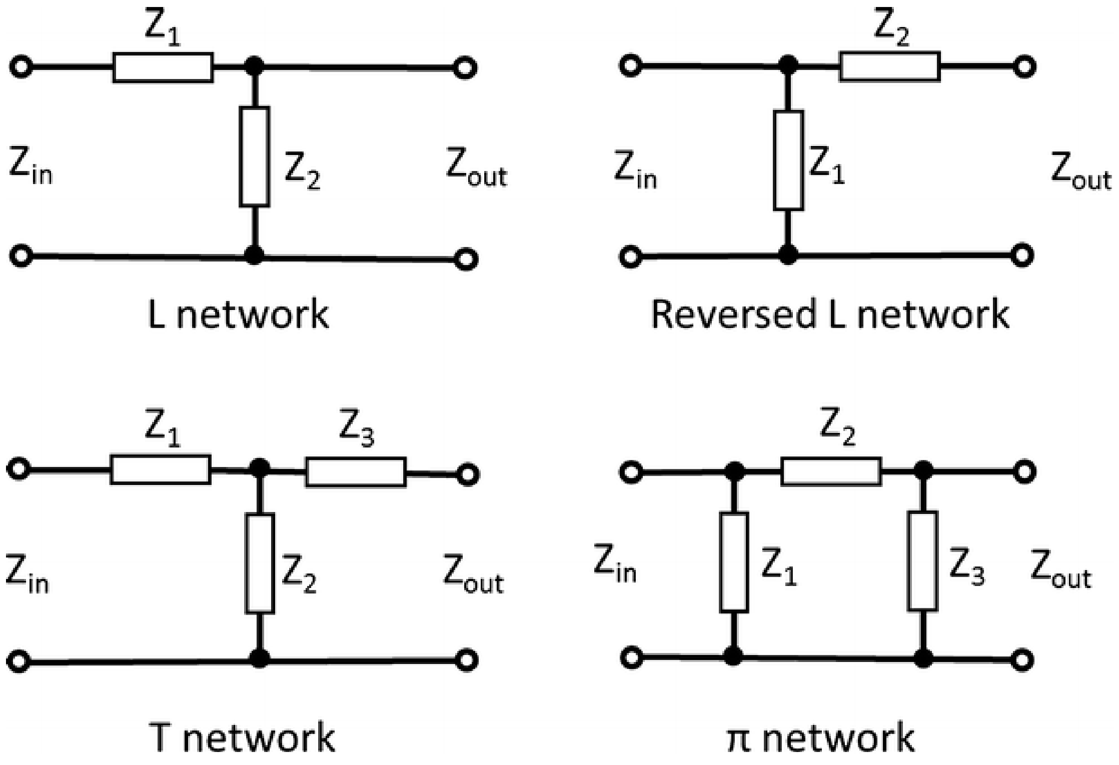

Fig. 5

Configuration of common impedance matching networks

The T and π matching configurations are more complex than the L network. Furthermore, organizing the T and π configurations into multiple stages will retain the final matching results but will change the Q factor. This strategy is useful in improving voltage boost.

There are tradeoffs between the attributes of an IMN, which include frequency, bandwidth, adjustability, and complexity. For instant, in [41], the suboptimal impedance matching and multiport ladder matching methods were introduced to enhance the matching performance and harvested power of antennas, respectively. However, the tradeoff was that implementation of these configurations required more components than traditional matching networks, thus, escalating the circuit's complexity.

Etor et al. [42] designed an IMN for THz frequencies applications using transmission lines and self-designed metal–insulator-metal diodes instead of lumped components. Moreover, fixed IMN and tunable IMN [43, 44, 45] were introduced as a technique for better matching with wide-band and multi-band antennas.

Rectifier/voltage multiplier

RF energy extracted from free space usually possesses low power density since the electric field power density decreases at the rate of $1/d^2$, where d is the distance from the RF source [46]. Therefore, a power amplifier circuit is required that yields enough DC energy from the electromagnetic waves to drive the loads. This gives rise to two possibilities, if the power consumption of the load is lower than the average power harvesting, the electronic devices at the load may work continuously; otherwise, if the load consumes more energy than the power harvesting circuit can generate, the devices cannot work continuously [47].

Rectifying is the most popular application of diodes, which refers to the conversion of AC current to DC current. In terms of power harvesting application, the RF signal retrieved in the antenna has a sinusoidal waveform. The signal after transformation through IMN would be rectified and boosted to meet the power requirements of the application.

The most fundamental topology of the rectifier is the half-wave rectifier that comprises of a single diode D1 (Fig. 6a). When AC voltage transfers through D1, only the positive cycle remains and the negative cycle is cutoff; thus, it diminishes half of the AC power. Moreover, the output V_{out} is discontinuous since the negative cycle is cutoff. Despite its simplicity, a half-wave rectifier is usually inadequate for common applications. Hence, a full-wave rectifier is more preferable. The circuit design of the full-wave rectifier is shown in Fig. 6b. During the first negative cycle of AC input, diode D1 is conductive and capacitor C1 is charged to the corresponding energy level of V_{peak} of the input. Then, at the next positive cycle, diode D1 is blocked, diode D2 is conductive so that capacitor C2 is also charged. In consequence, the output V_{out} would see two capacitors in series (each one is storing a voltage of $V_{peak)}$. Thus, V_{out} is twice V_{peak}. Therefore, this topology is more stable and efficient than the half-wave rectifier. There is also a bridge rectifier that rectifies both positive and negative cycles of the AC input but retains $V_{out} = V_{peak}$ by alternatively blocking pairs of diodes D1, D4 and D2, D3 (Fig. 6c).

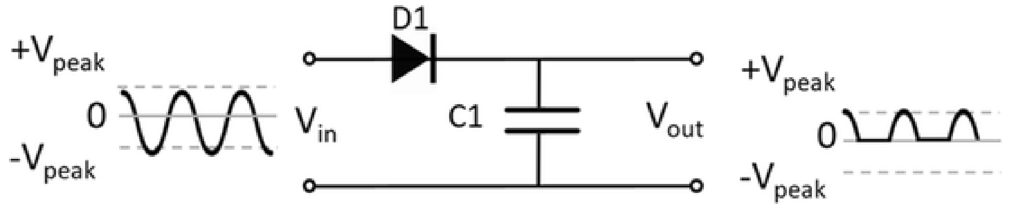

a Half-wave

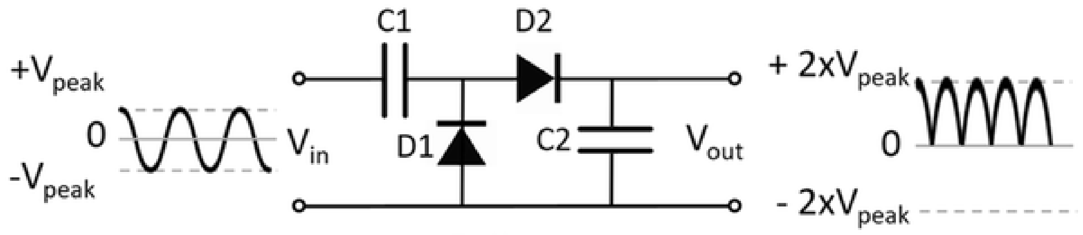

b Full-wave

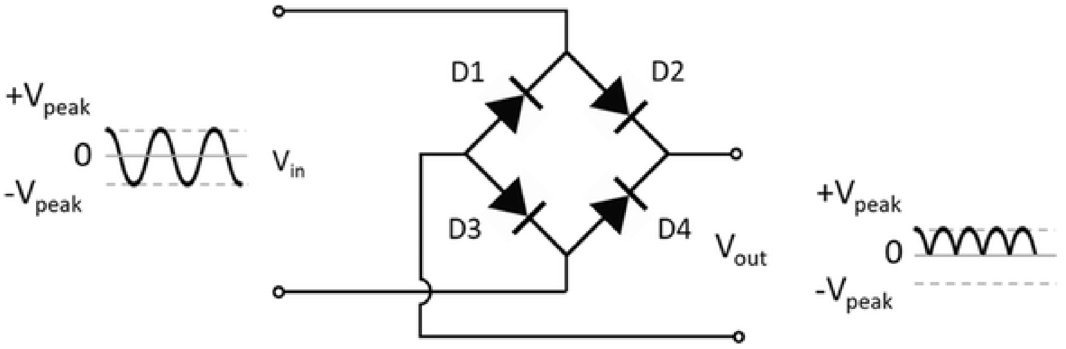

c Bridge

Fig. 6

Some common topologies of a rectifier

Voltage multiplier is a special type of rectifier circuit that converts and boosts AC input to DC output. In some case where the rectified power is inadequate for the application, there is a need for boosting the output DC by stacking single rectifiers into series, forming the voltage multiplier [48]. Several configurations of the voltage multiplier are shown in Fig. 7. The most fundamental configuration is the Cockcroft–Walton voltage multiplier (Fig. 7a). This circuit's operational principle is similar to the full-wave rectifier (Fig. 6b) but has more stages for higher voltage gain. The Dickson multiplier in Fig. 7b is a modification of Cockcroft–Walton's configuration with stage capacitors being shunted to reduce parasitic effects. Thus, the Dickson multiplier is preferable for small voltage applications. However, it is challenging to obtain high PCE due to the high threshold voltage among diodes creating leakage current, thus reducing the overall efficiency. Additionally, for high resistance loads, output voltage drops drastically leading to low current supply to the load. A summary of recent works related to voltage multiplier is shown in Table 3.

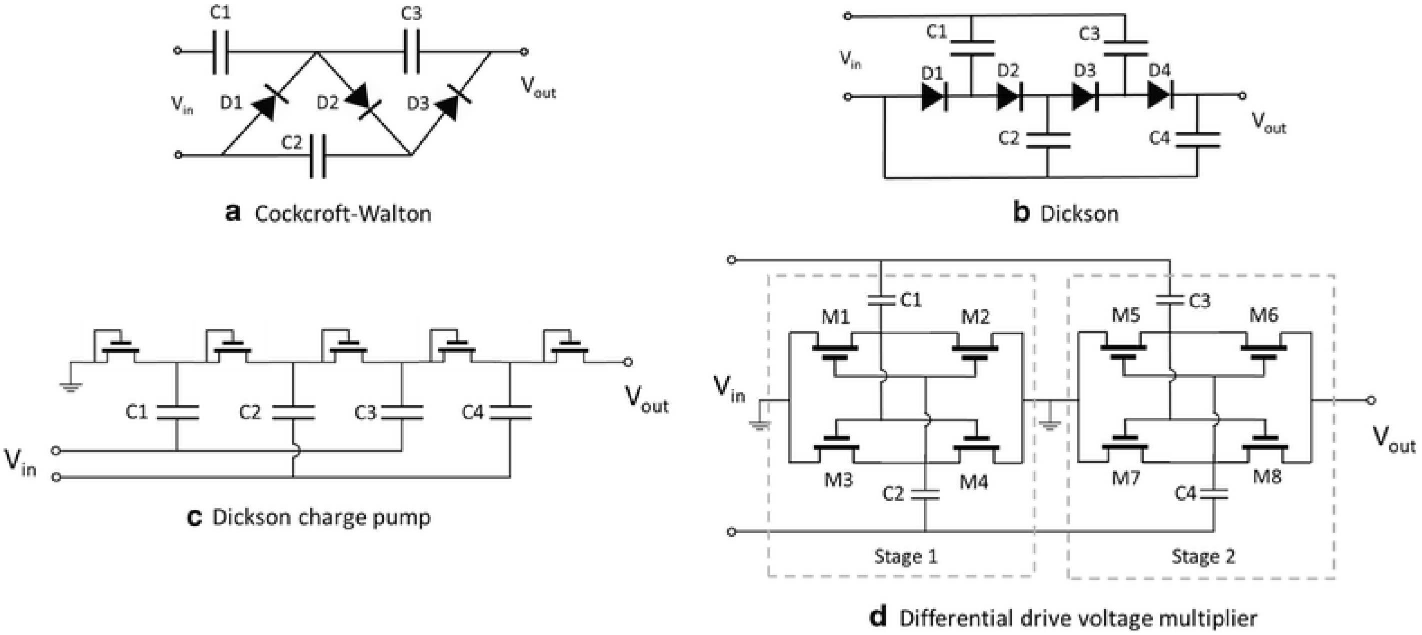
C1
C3
Vin
D1
D2
D3
Vout
C2
C1
C3
Vin
D1
D2
D3
D4
Vout
C2
C4
C1
C2
C3
C4
Vout
Vin
C1
M1
M2
C3
M5
M6
Vin
Vout
M3
C2
M4
M7
C4
M8
Stage 1
Stage 2

a Cockcroft-Walton

b Dickson

c Dickson charge pump

d Differential drive voltage multiplier

Fig. 7

Common voltage multiplier configurations: **a** Three stages Cockcroft–Walton voltage multiplier, **b** Four stages Dickson voltage multiplier, **c** Four stage Dickson voltage multiplier using CMOS technology, **d** Two stages voltage multiplier comprised of differential drive unit

Table 3

Published work comparison regarding voltage multiplier

Ref.	Rectifier topology	No. of stage	Freq. (MHz)	Tech.	Range (dBm)	Maximum PCE	Size
[51]	Differential	2	433	0.18 µm CMOS	0–20	74% @ 2 dBm	–
[52]	PMOS transistors	7	900	40 nm CMOS	–	44% @ > 390 mV	0.04 mm^2
[53]	Half-wave	4	900	0.18 µm CMOS	–	37.42% @ 390 mV	–
[54]	Comparator-based/active-diode	3	13.56	0.18 µm CMOS	8–15	67.9% @ 12.8 dBm	–
[55]	Dickson	3	13.56	250 nm CMOS	–	72%	0.13 mm^2
[56]	–	–	915	0.13 µm CMOS	−21.6	22.6% @ −16.8 dBm	0.186 mm^2

Prototyping Methods

There's more on this in Scherz, *Practical Electronics for Inventors.*

Solderless Breadboard or **Plugboard**

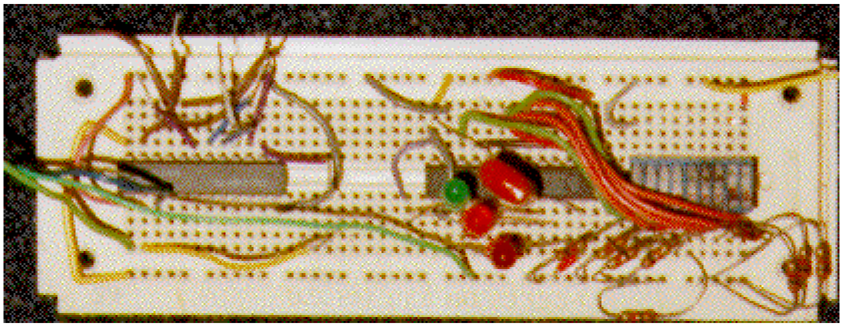

This is what we've been using in lab.

Advantages

- Very fast to build and make changes.
- Works well with DIP ICs

Disadvantages

- Limited reliability—lab testing only, for limited-size circuits.
- High capacitance between adjacent rows (~ 10 pF).
- Only for small-lead-size components (.032" or 0.82 mm max—a 1 A diode lead is just barely too big to meet the spec.)

Springboard

Used in ENGS 22

Advantages

- Fast to build and make changes.
- Accomodates large wire sizes.
- Secure, reliable connections.
- Can handle higher current/power.

Disadvantages

- Can't accommodate small lead spacings such as on ICs.
- Although it is reliable enough for long-term use, it's expensive for that purpose.

Wirewrap

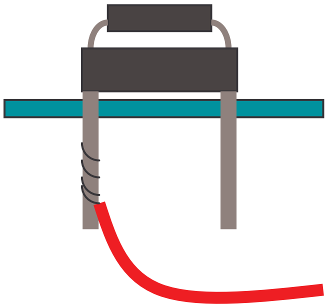

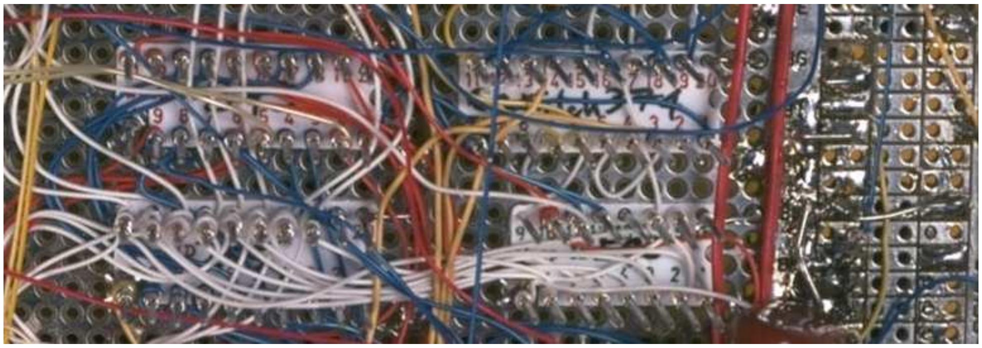

Connections made by fine wire wrapped tightly around square pins of special IC sockets. This is a great way to make a permanent version of a digital circuit almost as quickly as using a solderless breadboard. Used in ENGS 31.

Advantages

- Fast to build and make changes.
- Can make a complex circuit compact.
- Reliable.
- Can be used permanently.
- Inexpensive.

Disadvantages

- Low current—limited to under a few hundred mA.
- High inductance/resistance—on the order of 0.1 ohm, 0.25 µH per foot.
- Only makes reliable connections to square leads (as are on wire-wrap IC sockets). Discrete components (e.g., resistors, capacitors, transistors) need to be soldered to a "header" that goes in an IC socket or to individual wire-wrap pins.

Perfboard and Solder.

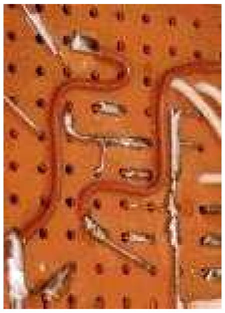

Using the same perfboard as used in wire-wrap work, it is possible to simply twist and solder leads, and run wires where needed.

Advantages

- Reliable, if done well.
- Can be used permanently.
- Inexpensive.
- Can handle any size components
- Convenient for working with discrete components.

Disadvantages

- o Slow, requires skill to do well.
- o Works for ICs, but not very easily.

Variations:

- Perfboard with individual copper pads on each hole so that the solder will better hold things in place. (Photo below)
- Perfboard with a perforated ground plane. Provides the shielding and grounding benefits of a ground plane, and makes ground connections easy, but requires care to avoid shorts. Some have the ground plane etched from around the holes; others require you to cut the copper away from the holes with a special tool. This is also possible with wire-wrap, as in the photo in that section.

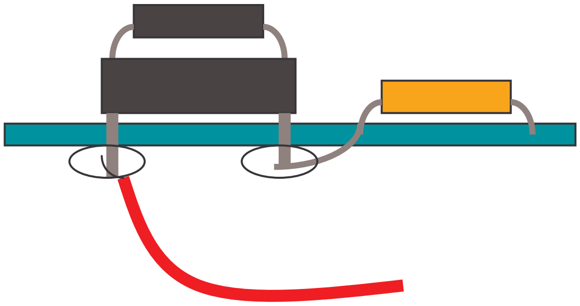

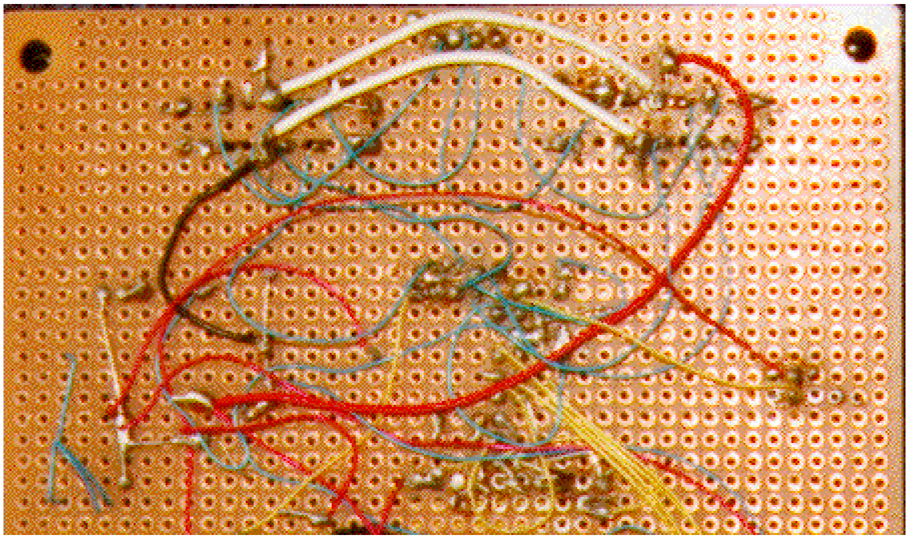

Generic Printed-Circuit Board—multiple hole-per-pad.

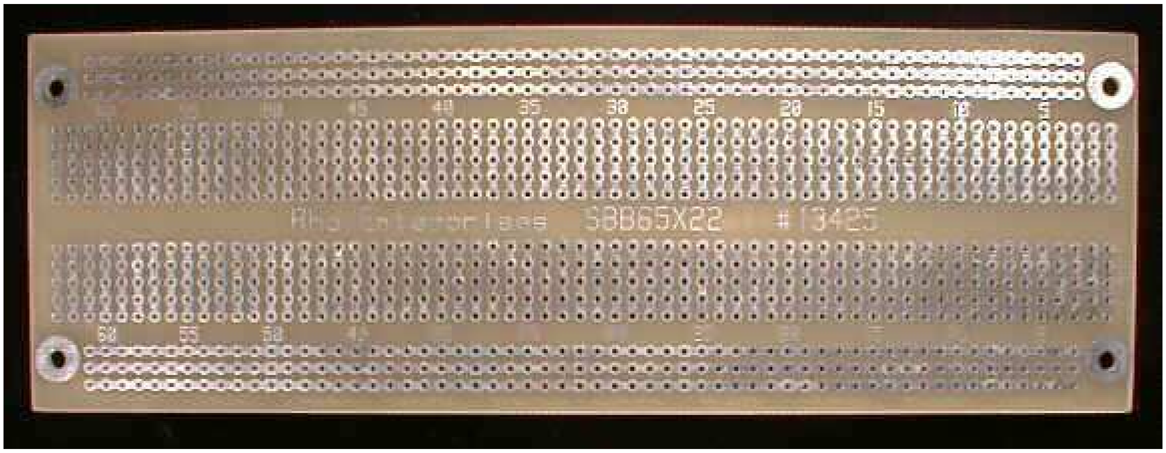

A printed circuit board with a pattern of holes and connections similar to a solderless breadboard.

Advantages

- Easier than using plain perfboard, especially for ICs.
- Reliable.
- Can be used permanently.
- Available with ground planes if needed.

Disadvantages

- Usually not as compact a final circuit as some alternatives, because you are constrained by layout.
- Pads bigger than needed can add capacitance, but not much.
- Can be expensive, especially the “vectorboard” brand.
- Sometimes awkward for odd sized or large components.

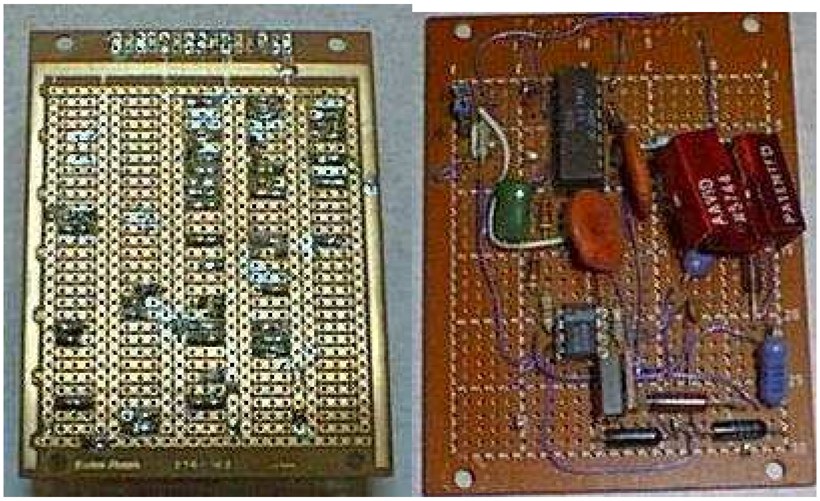

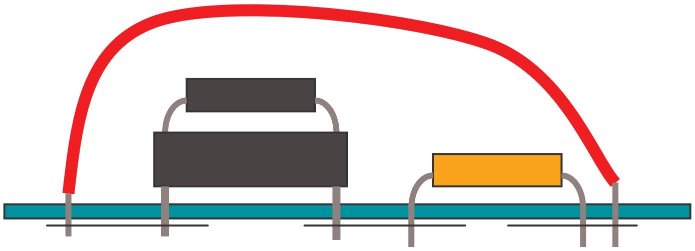

“Dead Bug,” or “Ugly-board”

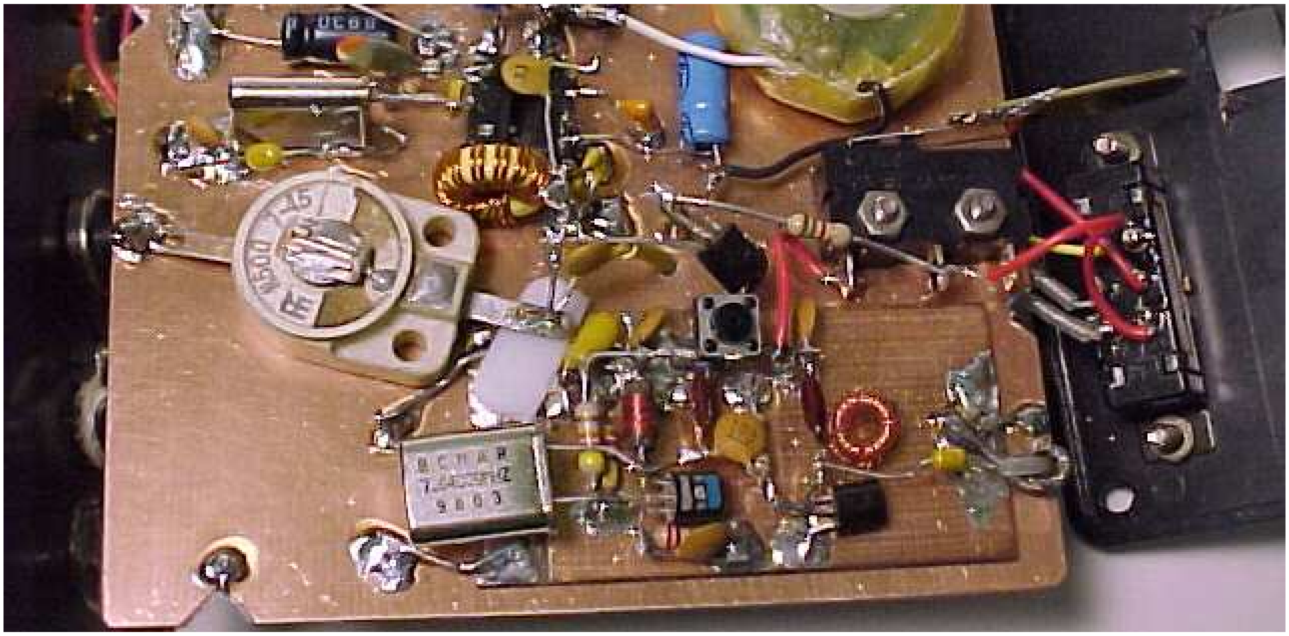

Start with a plain copper-clad board. Glue ICs down with the leads sticking up in the air. Then solder to them

Advantages

- Provides an excellent ground plane. Can be a high performance way to build sensitive and/or high-frequency analog circuits
- Can make a complex circuit compact.
- Reliable.
- Can be used permanently.
- Inexpensive.

Disadvantages

- Requires high soldering skill.
- Takes a long time to build.
- Mechanical support for components is marginal; can add glue (“RTV”) after debugging.
- Only makes reliable connections to square leads (as are on wire-wrap IC sockets). Other components (e.g., resistors, capacitors, transistors) need to be soldered to a “header” that goes in an IC socket or to individual wire-wrap pins.

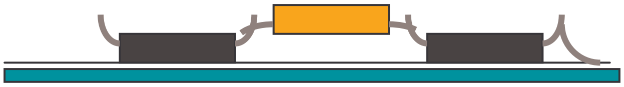

Variations:

- Manually cut the board with a dremel tool to isolate sections for purposes other than ground plane (use the back for ground plane). See photo below.
- Glue on little rectangles of cut-out board to add power busses, etc.

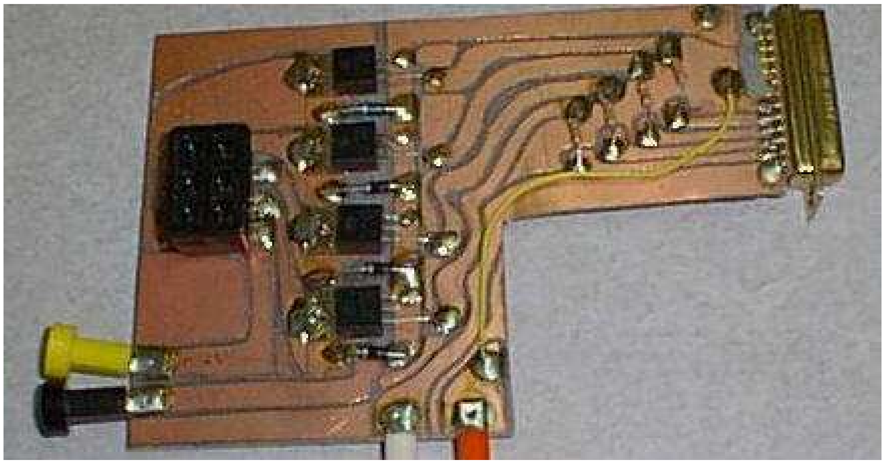

Printed circuit board (PCB)

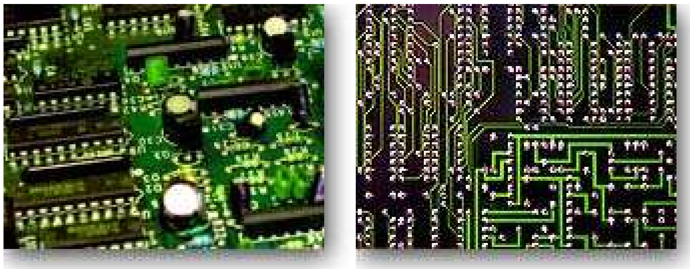

This is what is used virtually universally in production of electronics.

Advantages

- Easy to build in production.
- Repeatable, controllable stray L, C.
- Can handle virtually any component, power level.
- Highly reliable.
- Can make very compact.
- Design can be (somewhat) automated from a schematic you have entered.

Disadvantages

- Laying out the board and getting it fabricated takes time, although you can pay for fabrication in a few days if you can afford it.
- Expensive, on the order of hundreds of dollars for one, but with almost no increase in cost to make many.
- Hard to make changes, but making changes may be easier than building another type of prototype.

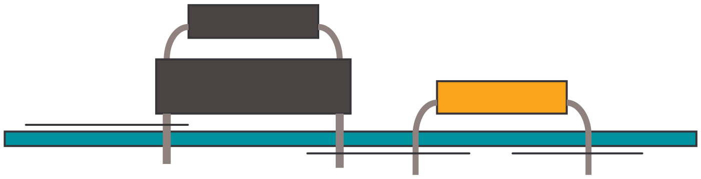

Notes on surface-mount components

Most modern production designs now use surface-mount components instead of “through hole” components. The circuits can be more compact, and the board layout is easier because different things can be done on each side (and in additional layers between sides) without through-holes interfering. But prototyping gets much more difficult!

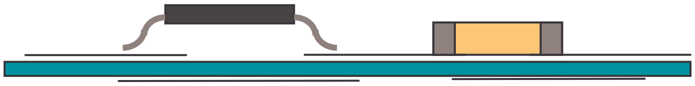

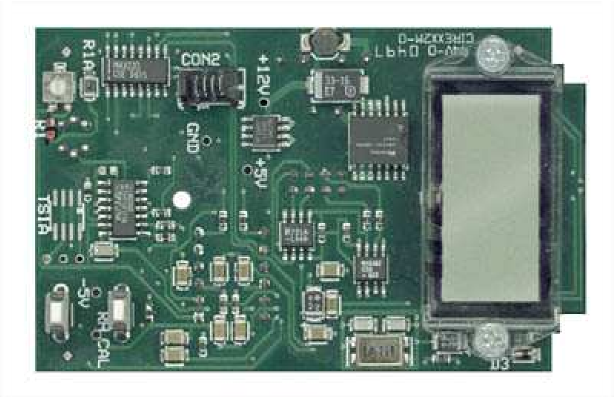

Prototyping options for surface mount

- Simulate, and then lay out a PCB. Don’t ever make a breadboard.
- Order DIP ICs and leaded passives for the prototype, and then switch to surface mount for production.
- Get adaptors that have pads to solder surface-mount ICs to, and then standard-spacing (0.1”) pins in DIP or SIP layout. Digikey carries “surfboards” made by Capital Advanced Technologies, and adaptors made by Aries Electronics. In addition to DIP and SIP adaptors, there are boards with solder pads for connecting larger wires; these work well for prototypes built in “dead bug” style.

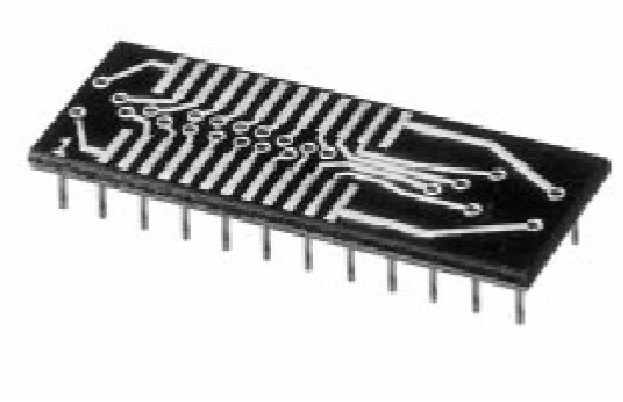

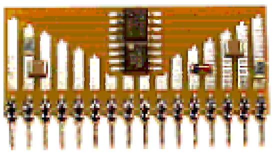

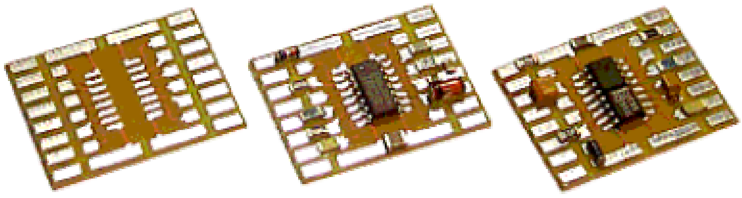

Crystal receiver set 11

Back to the index

This receiver has a single tuned circuit, the Q factor of the tuned circuit is quite low.
The receiver can be used well for reception of local stations, for reception of distant stations it is less suitable.

I tried to give this receiver a nice "old-fashioned" look, the reception performance was in this design of less importance.
This receiver is for sale, I made a series production of these, and will build more on request.
This receiver is also for sale as do-it-yourself kit.
See the shop for more information.

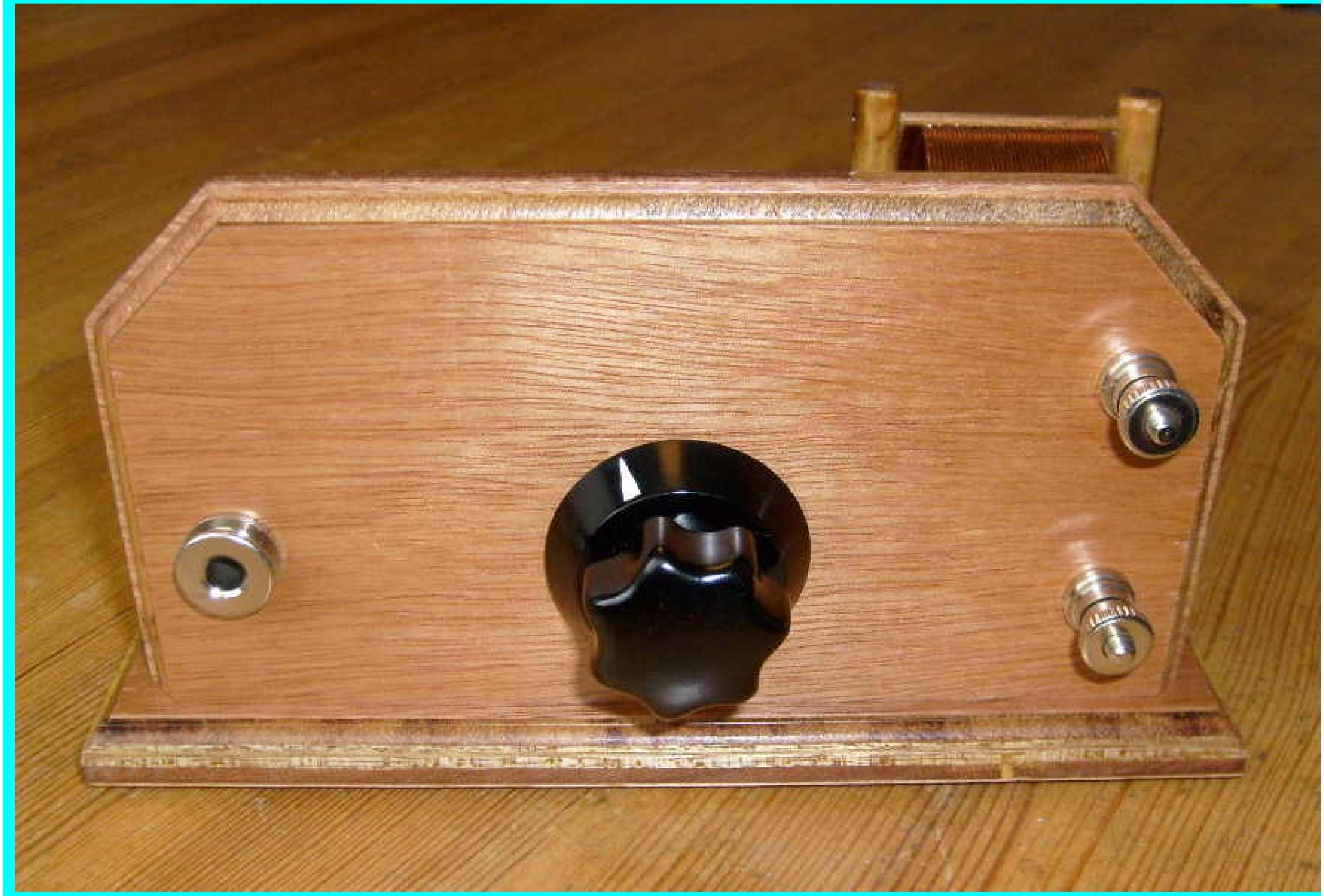

The front panel of the receiver.
On the left side the socket for the headphone.
In the middle the tuning knob.
On the right the connections for antenna (upper) and ground (lower).

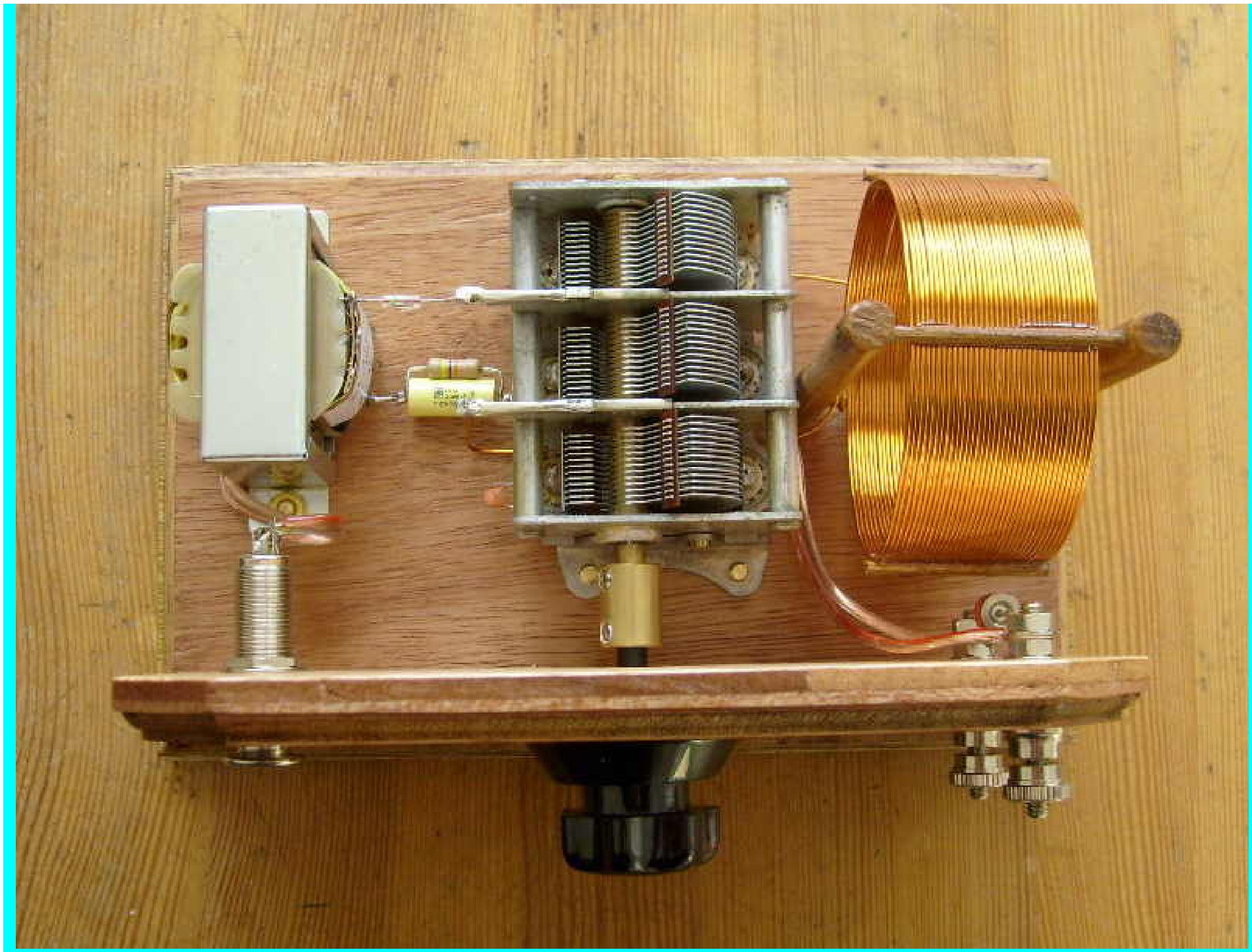

Top view of the receiver.

Circuit diagram of receiver set 11

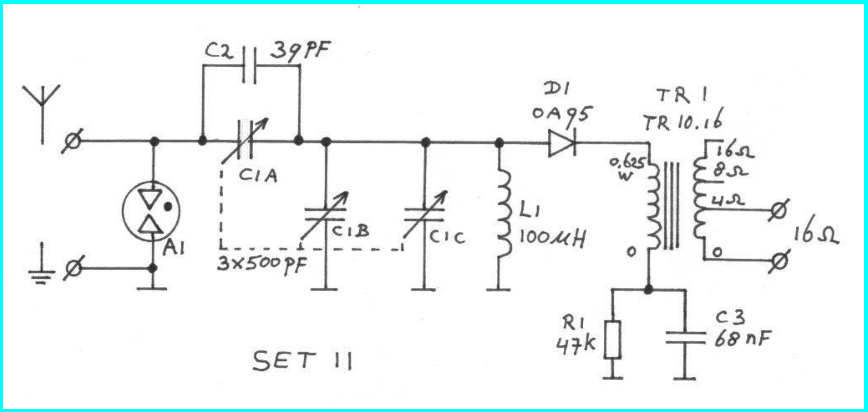

Circuit description.
The resonance circuit is formed by coil L1 and C1b and C1c (together 1000 pF).
Capacitor C1a and C2 provide the matching between antenna and tuned circuit.
The frame (rotor) of the tuning capacitor is carrying the RF signal, by this it is possible to tune simultaneously the tuned circuit (C1b and C1c) and the antenna matching (C1a).
Germanium diode D1 provides the signal detection.
Transformer TR1 is loaded with 16 Ω at its 4 Ω output, through this the input impedance increases from 16 kΩ to about 43 kΩ.
At the output a headphone of 2x 32 Ω can be connected, with the two speakers parallel, the impedance is 16 Ω.

Component A1 is a *gas discharge tube* (also called: *surge arrester*) with type number: N81-A90X.
The gas discharge tube protects the antenna input for too high voltages.

These high voltages can occur if the antenna picks up static charge from the air (especially occurs with long outdoor antennas from non insulated antenna wire).
As the voltage at the antenna is higher then 90 Volt, the gas discharge tube will start to conduct and short the high voltage to ground.
As soon as the charge has flown to ground, the conduction stops automatically.

Frequency range of the receiver.
Frequency range without antenna: 500 - 2500 kHz.
Frequency range with 10 meter antenna connected: 487 - 1860 kHz.
Both with and without antenna connected, the whole medium wave band can be tuned.

Q factor of the LC circuit (Without antenna and without diode connected):
600 kHz: Q= 83
900 kHz: Q= 81
1200 kHz: Q= 75
1500 kHz: Q= 65
The circuit Q is rather low, one reason for this is because the RF signal is in this design on the frame of the tuning capacitor.
Because the frame of the tuning capacitor is directly connected to the wooden bottom plate losses occur here.

You can find a complete building instruction op the following pages:

Step 1 Preparing the components

Step 2 Making the frame of the receiver

Step 3 Making the coil

Step 4 Placing the components

You will find here the part list of this receiver.

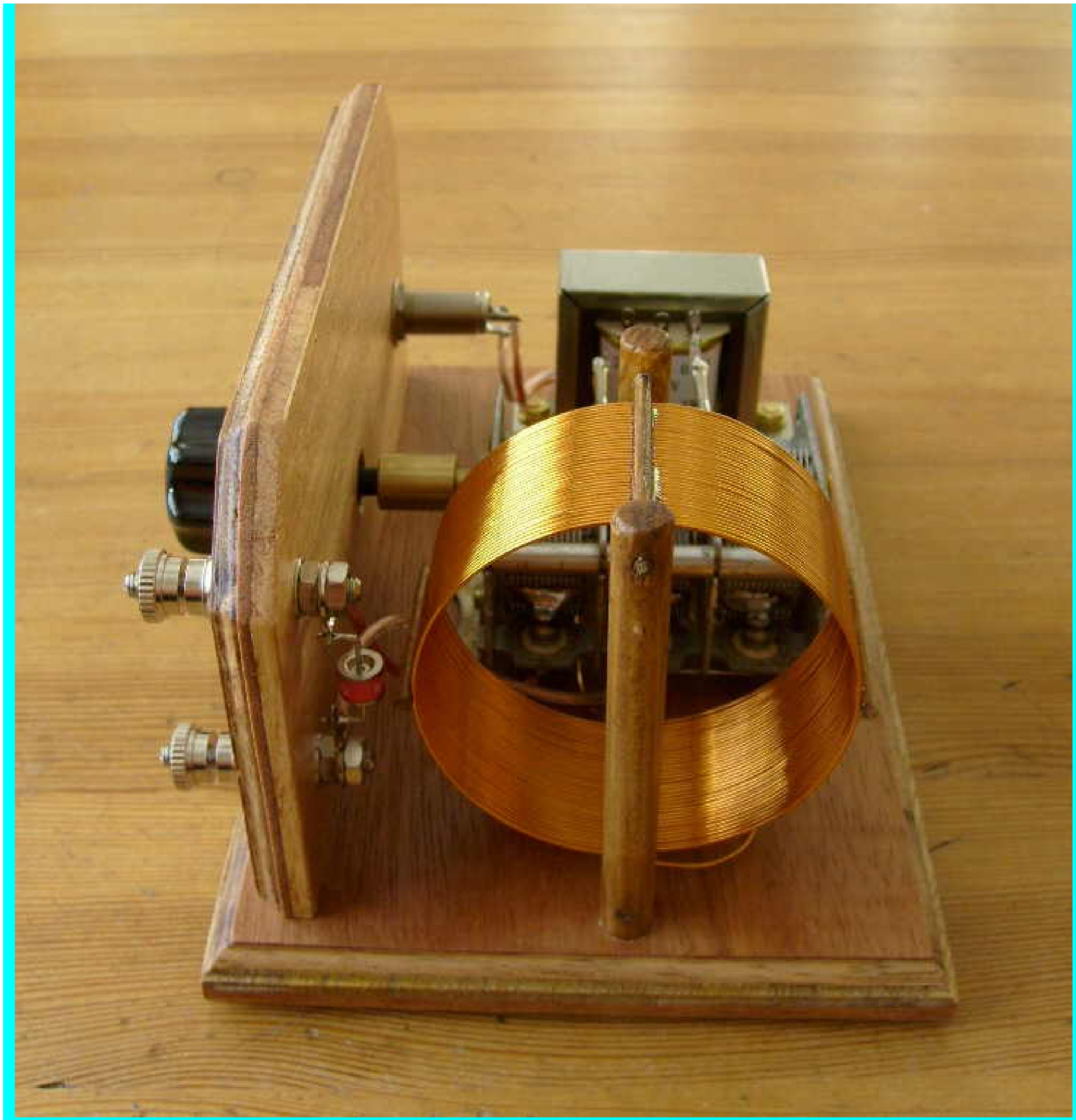

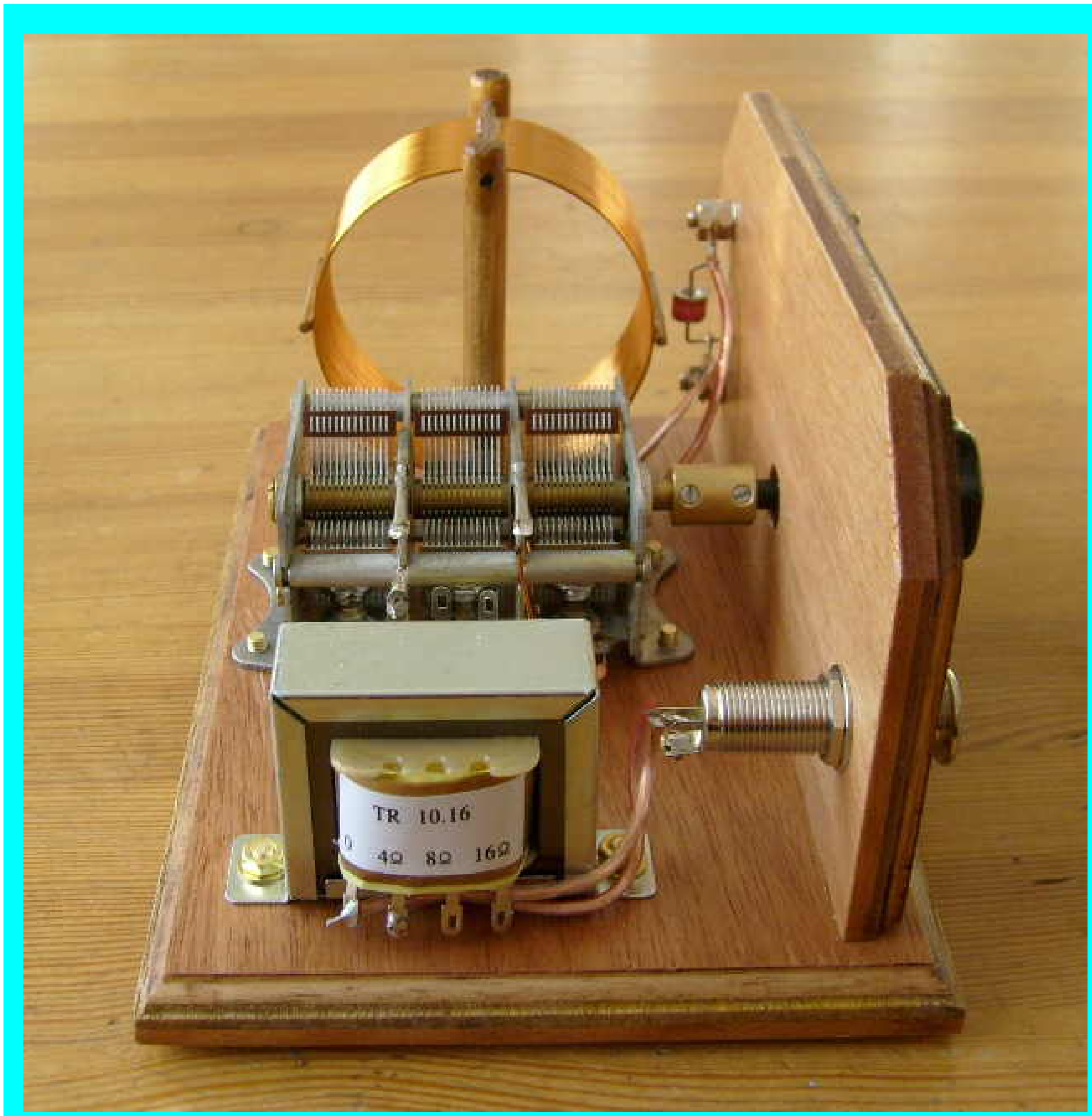
TR 10.16
Ω 4Ω 8Ω 16Ω

Back to the index

How to make a batteryless (crystal set) radio

By **Sagar Sapkota** - July 21, 2012

CRYSTAL SET

The simplest radio receiver, known as a Crystal Set, consists of nothing more than a coil, tuning capacitor, diode detector, and a pair of earphones. A typical circuit diagram for a Crystal Set Radio is given below where inductor or coil L1 is tuned by variable capacitor VC1 to the transmitter frequency. Diode D1 demodulates the signal, which is fed straight to the earphones. There is no amplification.

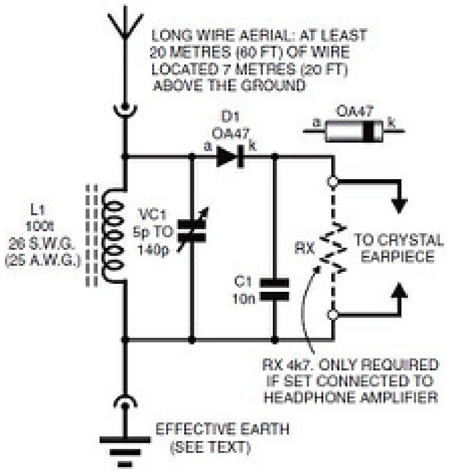
LONG WIRE AERIAL: AT LEAST 20 METRES (60 FT) OF WIRE LOCATED 7 METRES (20 FT) ABOVE THE GROUND
D1
OA47
a
k
OA47
a
k
L1
100t
26 S.W.G.
(25 A.W.G.)
VC1
5p TO
140p
RX
TO CRYSTAL EARPIECE
C1
10n
RX 4k7. ONLY REQUIRED IF SET CONNECTED TO HEADPHONE AMPLIFIER
EFFECTIVE EARTH
(SEE TEXT)

CLICK HERE FOR BIGGER PHOTO

** You can use 1N34 Germanium diode in place of OA47*

A long (at least 20 metres), high (17 metres or more) aerial and a good earth (a buried biscuit tin or a metre of copper pipe driven into damp ground) are required in order to ensure audible headphone reception. The earphones originally used with these receivers had an impedance of around 4000 ohms and were very sensitive (and heavy and uncomfortable). They are no longer available, but a crystal earpiece, which relies on the piezoelectric effect, will give acceptable results.

Low impedance "Walkman" type earphones are NOT suitable.

Component details:

Resistor- RX-4.7k, 0.25 W- only required if set is connect to audio amplifier

Capacitors- C1- 10nF disc ceramic VC1- 5p to 140p, polythene dielectric variable capacitor.

Semiconductors: D1: OA47 or 1N34- Germanium Diode

Miscellaneous: L1- Ferrite Rod, 100mm(4 inch)x9mm/10mm dia., with coil.

Crystal earpiece and jack socket to suit; plastic control knob; plastic insulated flexible cable for aerial wire, download and earth connection, 30 meters minimum; buried biscuit tin or 1 meter of copper pipe for earth system; 50gm reel of 26SWG enamelled copper wire, for tuning coil; card and glue for coil former; multistrand connecting wire; crocodile clips or terminals for aerial and earth lead connection; solder, etc.

READ MY EXPERIENCES- HOW EASILY I MADE A BATTERYLESS RADIO

DRAWBACKS

Quite apart from the absence of amplification, two factors seriously limit the performance of crystal receivers. Germanium diodes become increasingly reluctant to conduct as the applied voltage falls below 0.2V, and this makes the receiver insensitive to weak signals. Silicon diodes have a threshold of around 0.6V, and are, therefore, unsuitable for circuits of this kind.

** Silicon diodes like 1N4001 are not suitable*

The earphone loading imposes heavy damping on the tuned circuit, hence, reduces its ability to separate signals. With such low selectivity insensitivity can be a blessing, and crystal sets are normally only capable of receiving a single, strong transmission on the long and medium wavebands. They will sometimes receive more than one if a shortwave coil is fitted.

The aerial and diode can be connected to tappings on the tuning coil in order to reduce damping, but the improvement in selectivity is usually at the expense of audio output. When valves cost a week's wages and had to be powered by large dry batteries and lead/acid accumulators, the construction of simple receivers of this kind could be justified. With high performance transistors now costing only a few pence or cents, crystal sets are now regarded as 'nostalgic pieces". Some readers may, however, wish to build one out of curiosity, or for the novelty of having a receiver that **does not require a power supply.**

CIRCUIT DETAILS

Ferrite loop aerial L1 and polythene dielectric variable capacitor VC1 form the tuned circuit. Point contact germanium diode D1 (1N34 or OA47) demodulates the signal; capacitor C1 bypasses residual r.f. (radio frequency) to earth and also exhibits a reservoir action, enabling the a.f. (audio frequency) output to approach its peak value. The recovered audio signal is fed directly to a crystal earpiece. Signal voltages introduced in the ferrite loop aerial by the radiated magnetic field are much too much to produce an output from the detector, and the component is used here simply as a tuning coil. The ferrite core does, however, reduce the number of turns required for the coil winding, thereby reducing its resistance and increasing its audio quality.

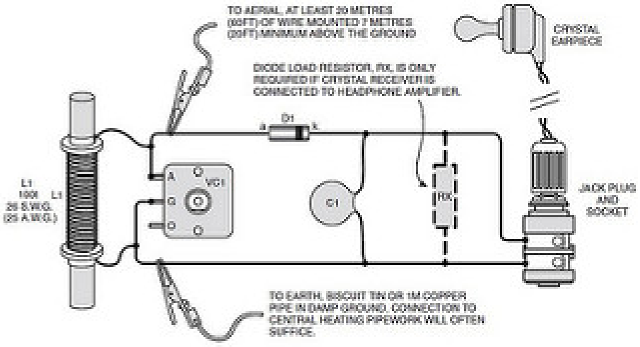
TO AERIAL, AT LEAST 20 METRES (66FT) OF WIRE MOUNTED 7 METRES (20FT) MINIMUM ABOVE THE GROUND
CRYSTAL EARPIECE
DIODE LOAD RESISTOR, RX, IS ONLY REQUIRED IF CRYSTAL RECEIVER IS CONNECTED TO HEADPHONE AMPLIFIER.
D1
a
k
A
VC1
G
O
L1
100t
L1
26 S.W.G.
(25 A.W.G.)
C1
RX
JACK PLUG AND SOCKET
TO EARTH, BISCUIT TIN OR 1M COPPER PIPE IN DAMP GROUND. CONNECTION TO CENTRAL HEATING PIPEWORK WILL OFTEN SUFFICE.

GET BIGGER PICTURE

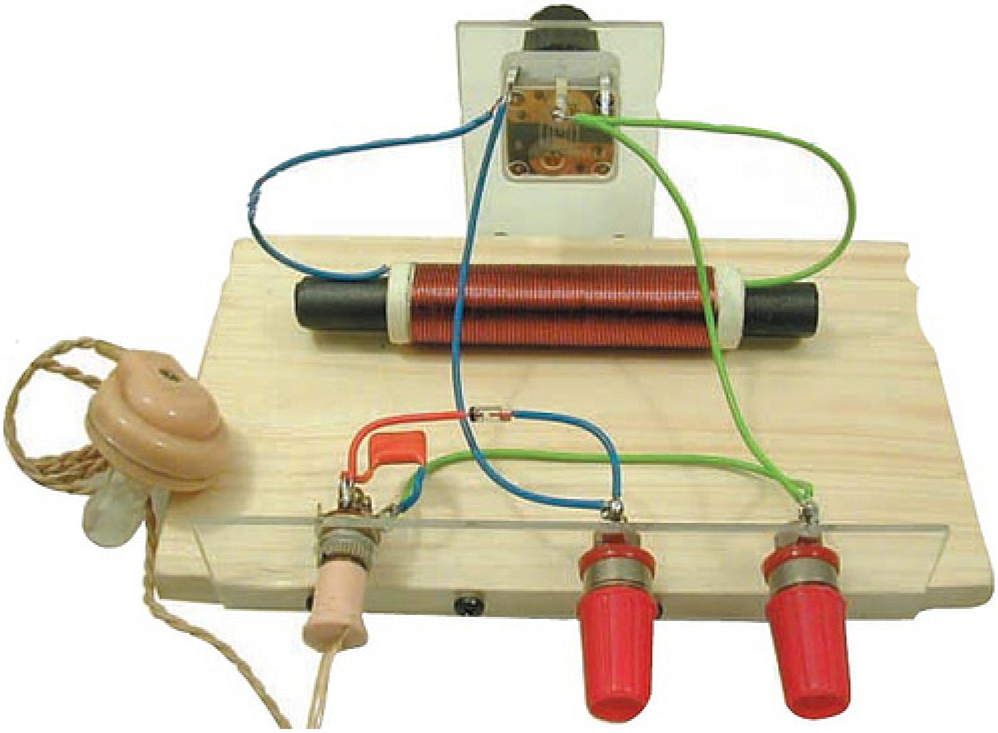

COIL DETAILS

Full construction and winding details for the ferrite tuning/ aerial coil L1 are shown in figure.

The coil is made from 26 SWG(25 AWG) enamelled copper wire, close wound on a cardboard former.

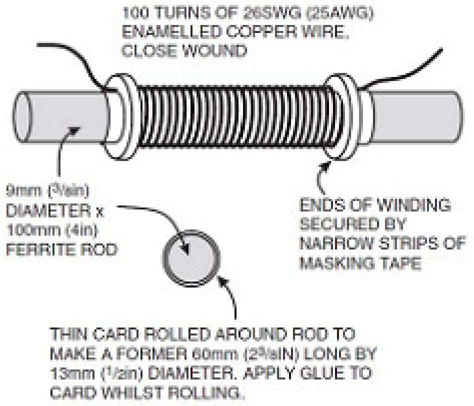
100 TURNS OF 26SWG (25AWG)
ENAMELLED COPPER WIRE,
CLOSE WOUND
9mm (3/8in)
DIAMETER x
100mm (4in)
FERRITE ROD
ENDS OF WINDING
SECURED BY
NARROW STRIPS OF
MASKING TAPE
THIN CARD ROLLED AROUND ROD TO
MAKE A FORMER 60mm (2 3/8IN) LONG BY
13mm (1/2in) DIAMETER. APPLY GLUE TO
CARD WHILST ROLLING.

The r.f. bypass capacitor C1 can in practice, be omitted with no noticeable reduction in performance. However, if the set is to be connected to either the headphone amplifier or speaker amplifier decribed next month, the component together with diode, load resistor, RX, must be included.

AMPLIFICATION Audio frequency amplification, after the diode detector will permit the use of low impedance Walkman type earphones or even loud speaker operation. It will do nothing, however, to overcome the diode’s insensitivity to weak signals. For this we must have radio frequency amplification of the signals picked up by the aerial before they reach the detector(The standard circuit for a transistor portable receiver has three stages of radio frequency amplification ahead of the diode).

For amplification, you can use a simple audio amplifier using LM386.

READ MY EXPERIENCES ON MAKING A BATTERYLESS RADIO

WATCH THESE VIDEOS:

- **JLCPCB: Prototype 10 PCBs for $2 (2-layer,100*100mm)**
- **China's Largest PCB Prototype Manufacturer, 290,000+ Customers & 8,000+ Online Orders Per Day**
- **PCB Pricing: https://jlcpcb.com/quote**

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/271597499

Fully printed flexible and disposable wireless cyclic voltammetry tag

Article *in* Scientific Reports · January 2015

DOI: 10.1038/srep08105 · Source: PubMed

CITATIONS
21

READS
121

11 authors, including:

Jinsoo Noh
Gumi Electronics and Information Technolog…
22 PUBLICATIONS **641** CITATIONS

SEE PROFILE

Myoungho Pyo
Sunchon National University
96 PUBLICATIONS **1,319** CITATIONS

SEE PROFILE

Ali Javey
University of California, Berkeley
296 PUBLICATIONS **22,530** CITATIONS

SEE PROFILE

Gyoujin Cho
Sunchon National University
95 PUBLICATIONS **2,833** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

Electrochemistry View project

All content following this page was uploaded by Gyoujin Cho on 02 February 2015.

The user has requested enhancement of the downloaded file.

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

ENVIRONMENTAL MONITORING

ELECTRICAL AND ELECTRONIC ENGINEERING

Received
22 October 2014

Accepted
6 January 2015

Published
29 January 2015

Correspondence and requests for materials should be addressed to G.C. (gcho@sunchon.ac.kr)

Fully printed flexible and disposable wireless cyclic voltammetry tag

Younsu Jung[1], Hyejin Park[1], Jin-Ah Park[1], Jinsoo Noh[1], Yunchang Choi[1], Minhoon Jung[2], Kyunghwan Jung[2], Myungho Pyo[1], Kevin Chen[3], Ali Javey[3] & Gyoujin Cho[1]

[1]Department of Printed Electronics Engineering, Sunchon National University, Maegok, Sunchon, Jeonnam 540-742 Korea, [2]Research Institute of Printed Electronics, PARU Co. Seomyeon, Sunchon, Jeonnam 540-813 Korea, [3]Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720.

A disposable cyclic voltammetry (CV) tag is printed on a plastic film by integrating wireless power transmitter, polarized triangle wave generator, electrochemical cell and signage through a scalable gravure printing method. By proximity of 13.56 MHz RF reader, the printed CV tag generates 320 mHz of triangular sweep wave from +500 mV to −500 mV which enable to scan a printed electrochemical cell in the CV tag. By simply dropping any specimen solution on the electrochemical cell in the CV tag, the presence of solutes in the solution can be detected and shown on the signage of the CV tag in five sec. 10 mM of N,N,N′,N′-tetramethyl-p-phenylenediamine (TMPD) was used as a standard solute to prove the working concept of fully printed disposable wireless CV tag. Within five seconds, we can wirelessly diagnose the presence of TMPD in the solution using the CV tag in the proximity of the 13.56 MHz RF reader. This fully printed and wirelessly operated flexible CV tag is the first of its kind and marks the path for the utilization of inexpensive and disposable wireless electrochemical sensor systems for initial diagnose hazardous chemicals and biological molecules to improve public hygiene and health.

Cyclic voltammetry (CV) has been used as a powerful tool for the study of electrochemical redox reactions between the electrodes and the solutions of all kind of solutes such as metals[1], organic molecules[2], proteins[3], bacteria[4], viruses[5], and DNA[6]. Because of the high sensitivity of the electron transfer redox reactions, CV could become a very promising ubiquitous electrochemical sensor protocol if the cost and size of the CV system were significantly reduced to commercially viable single-use disposable units for checking traces of hazardous materials such as lead[7], mercury[8], arsenic[9], e-*coli*[10], and pesticides[11] in water or to diagnose the level of glucose[12], cholesterol[13] and specific enzymes[14] in blood. This single-use disposable CV would dramatically reduce the cost of maintaining the public health system[15]. Therefore, inexpensive and disposable CV measurement system that can be operated wirelessly using a smartphone or RF (radio frequency) reader without complicate operation processes is in high demand for the realization of an ubiquitous sensor network system (Figure 1a). They would mainly be used as ubiquitous diagnostic and testing tools for detecting and monitoring the level of target specimens. However, there is no technology that is advanced enough yet to build a wireless, inexpensive and disposable CV system. In this paper, as a form of RF-tag, an extremely inexpensive, disposable and fully printed CV system is demonstrated for the first time by mimicking and combining basic CV concepts and wireless power transmission technologies of RF devices. To realize the fully printed CV tag, a key issue of wirelessly generating triangular waveform (±500 mV) that can scan the electrochemical cell at a low frequency (<1 Hz) to set up a redox reaction needs to be addressed through a minimum number of printed thin film transistors (TFTs). Fully printed 13.56 MHz rectenna and a ring oscillator with large trap charges in the channels of printed TFTs were respectively utilized to wirelessly generate triangular waveform by using only 10 printed TFTs.

Results

Based on two key units of the wireless power transmitter (Figure 1b-①) and the triangular wave generator (Figure 1b-②), the circuit layout of the fully printed wireless CV tag was designed by using a minimum number of printed thin film transistors (TFTs) to alleviate the issue of V_{th} shift and shown in Figure 1b as a platform of disposable CV system. The circuit of CV tag was fabricated using all scalable printing methods, a roll-to-roll gravure, a roll-to-plate gravure, a drop casting, and a screen printer (Figure S1 in Supplementary Information).

To provide a polarized DC voltage from the coupled 13.56 MHz AC signal of the reader, we modified our previously reported R2R gravure printed rectenna[16,17], as shown in the layout in Figure 2a and the resulting

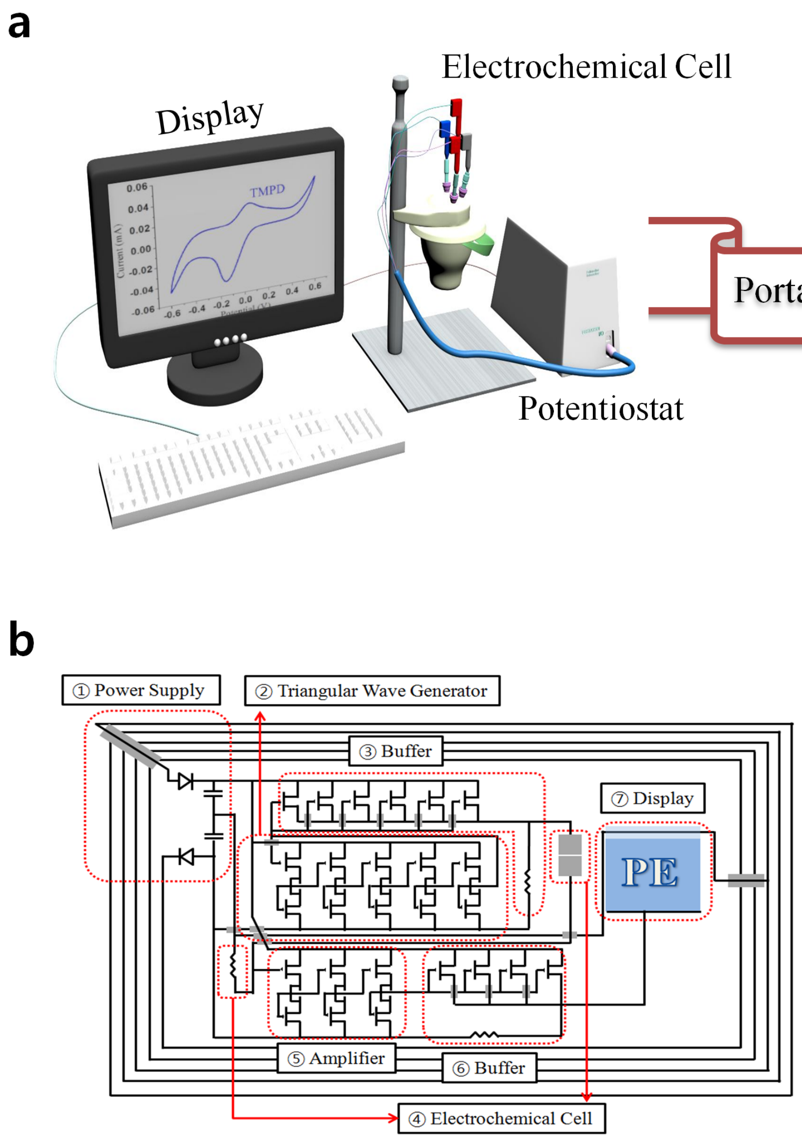

Figure 1 | (a) Informative illustration of typical CV system and disposable printed CV tag. (b) Schematic circuit diagram of the gravure printed wireless cyclic voltammetry (CV) tags. The circuit was designed to couple AC power from a 13.56 MHz reader and then convert the coupled AC to polarized DC ±10 V (① in Figure 1b). Polarized DC will operate the printed ring oscillator to generate a triangular voltage waveform (② in Figure 1b). The generated waveform will pass through the buffer to meet the impedance difference from the electrochemical cell (③ in Figure 1b). The electrochemical cell will run the redox reaction with a single drop of specimen solution (④ in Figure 1b) by the voltage triangular waveform. The output current of the electrochemical redox reaction will be amplified *via* an amplifier circuit (⑤ in Figure b) and the signal will pass through the buffer to meet the impedance difference from the printed signage (⑥ in Figure b). The signage will indicate the concentration level of specimen in the solution (⑦ in Figure b). It will indicate whether the concentration is above or below a pre-determined value.

printed diodes and capacitors are shown in Figure 2b. The printed diodes showed a rectifying ratio of 10^3, and capacitances of two printed capacitors were all about 8 nF/cm^2. The polarized DC voltages (Figure 2c) were measured by placing the printed rectenna on the RF (13.56 MHz) reader with a distance of 2 cm. We adapted a center tap transformer, consisted by divided antenna, 2 diodes and 2 capacitors, to provide the + and − DC voltage for printed triangle wave generator. Using a load of 1 MΩ, voltages of +9.4 V and −10.8 V DC were attained for the optimized printed antenna by monitoring the rectified polarized DC voltages and varying the values of the inductance of the antenna (Figure S2 in Supplementary Information). The polarized DC voltage was also decreased with decreasing load's resistance (Figure 2d).

To generate a cyclic waveform with a low frequency (<1 Hz) from the rectified polarized DC voltage, fully gravure printed *p*-type SWNT-based channel network thin film transistors (*cn*TFTs) were used to construct a five stage ring oscillator with *p*-type inverters. The *cn*TFTs can operate under less than 20 V and their mobility range (0.01 to 10 cm^2/Vs) can be controlled through the loading concentration of SWNT in the ink[18–21]. To print the five stage ring oscillator, we used the same silver and $BaTiO_3$ inks as in the printing of the rectenna for reducing a number of printing steps while SWNT ink was used for printing the channel of the *cn*TFTs. To manufacture the ring oscillator, roll-to-plate (R2P) gravure was used to print 10 *cn*TFTs with a gate width of 300 μm and channel length of 160 μm for both the drive and load *cn*TFTs (Figure S3 in Supplementary Information). The optimized cell depth of gravure plate was 37–40 μm with the cell wall thickness of 50–55 μm for the printing active layers of drive and load TFTs using low viscosity of SWNT based ink (~30 cp) (Figure S4 in Supplementary Information). By repeating the printing of active layers on load TFTs, a higher SWNT network density can be achieved to generate higher current in the

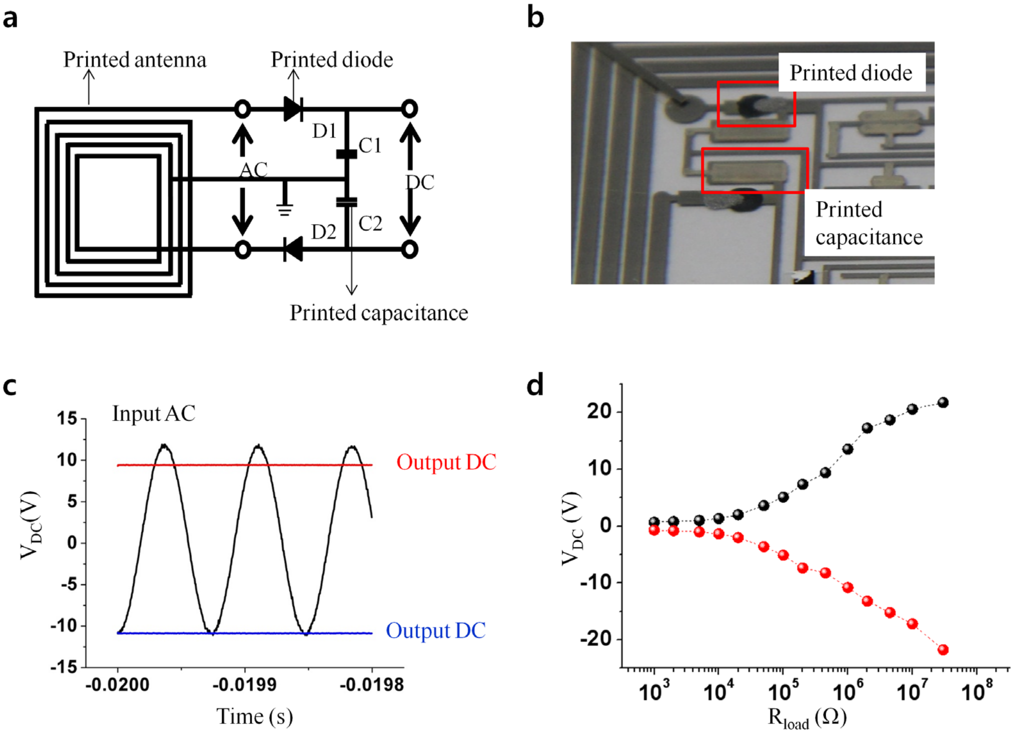

Figure 2 | (a) Circuit layout of the printed rectenna to provide polarized DC power, (b) a real image of printed diodes and capacitors, (c) input-output electrical characteristics of the rectifier at 13.56 MHz AC input and (d) output ± DC voltages under variations in load resistance.

channels while thicknesses of all printed gate, dielectric and drain-source electrodes are kept same (Figure S5 in Supplementary Information).

Their output and transfer characteristics are shown in Figure 3a and 3b respectively. Based on the attained transfer characteristics, the mobility (μ), transconductance (g_m), on-off current ratios from I_{DS}-V_{DS} and threshold voltages (V_{th}) for both five drive *cn*TFTs and five load *cn*TFTs were extracted and shown in Table S1 (Supplementary Information). The shift range of V_{th} from the fully gravure printed 10 *cn*TFTs were about ±0.4 V for the drive *cn*TFTs and ±0.6 V for load *cn*TFTs. Those shift ranges are narrow enough to run the printed ring oscillator for generating the triangular wave. Each inverter delivered a gain of 3 (Figure 3c), and the ring oscillator with five inverters generated a pseudo triangular waveform with a voltage amplitude of 7 V at 320 mHz (Figure 3d). In fact, the pseudo triangular wave can be only generated by the printed *cn*TFTs based ring oscillator because of the parasitic and trapped channel capacitances due to carbon nanotube network structures in *cn*TFTs. Although we only showed the frequency of 320 mHz at here, the frequency can be varied from 0.3 Hz to 2 Hz based on the printed network density of SWNT of printed *cn*TFTs. This is an advantage of our printed *cn*TFTs for scanning electrochemical cell in CV tag over other technologies such as amorphous silicon TFTs where only a sine wave with a designed single frequency can be generated because of no trap capacitance in the channel (Figure S6 in Supplementary Information). The frequency of the triangular waveform can be fine-tuned in the range of 0.3 Hz to 2 Hz as shown in the Figure S7 (Supplementary Information) by simply changing loading concentration of SWNTs.

When the generated triangular waveform is supplied directly into the printed electro-chemical cell (Figure S8 in Supplementary Information), wherein printed silver and carbon electrodes were used with a poly(ethylene oxide) and $LiCF_3SO_3$ gel type electrolyte, the output voltage range is reduced to zero because of improper impedance matching between the electrochemical cell (10 KΩ) and ring oscillator (1 MΩ). Therefore, a buffer unit is needed to match the impedance levels between the ring oscillator and electrochemical cell. The printed buffer unit (Figure 1b-③) with high on currents and low on-off current ratios (Figure 3e) is consists of 6 *cn*TFTs and a resistor (~7 KΩ). In the printed CV tag, three resistors were printed by screen printer using a carbon paste (DC-20, purchased from Dozen Co. Korea) for the buffer units (Figure 1b-③ and ⑥) and the electrochemical cell (Figure 1b-④). Their electrical parameters are listed in Table S1 (Supplementary Information). The 6 *cn*TFTs of buffer unit were printed using R2P gravure with high SWNT concentration in the semiconducting ink. Each *cn*TFTs exhibited on-currents in the range of 400 ~ 500 μA with a very low on-off ratio due to the enhanced metallic percolation in the dense SWNT networks. The buffer units with low on-off ratio were employed to reduce the input voltage to ±500 mV (Figure 3f) which is the appropriate scanning range of the electrochemical cell. The printed electrochemical cell with two electrodes was proven to be well operated under commercial CV (Figure S9 in Supplementary Information). The resulting current after scanning printed electrochemical cell in the CV tag can be converted to voltage by connecting the printed resistor (Figure 1b-④: ~5 × 10^3 ohm) which provides the output voltage for electrochromic indicator and re-plotted into the cyclic voltammogram.

Because these output voltage (±70 mV) and current (~40 μA) were low after scanning the electrochemical cell, they need to be amplified to turn on the electrochromic indicator for showing pre-determined concentration (10 mM) level of N,N,N′,N′-tetramethyl-p-phenylenediamine (TMPD, purchased from Aldrich) as a standard reference in this work (Figure 3g). After running the redox reaction of TMPD (Figure 3h) in the printed cell, the output voltage and current needed to be amplified to reach 1.5 V and 500 μA, respectively, to provide sufficient power to display the letters "PE" as an indicator of the presence of TMPD in the specimen. To construct the amplifier, *cn*TFTs based three inverters were printed with a gain of three to five (Figure S10 in Supplementary Information) to amplify the output signal as shown in Figure 3i. The electrical characteristics of the 6 *cn*TFTs in the inverters with extracted mobility, transcon-

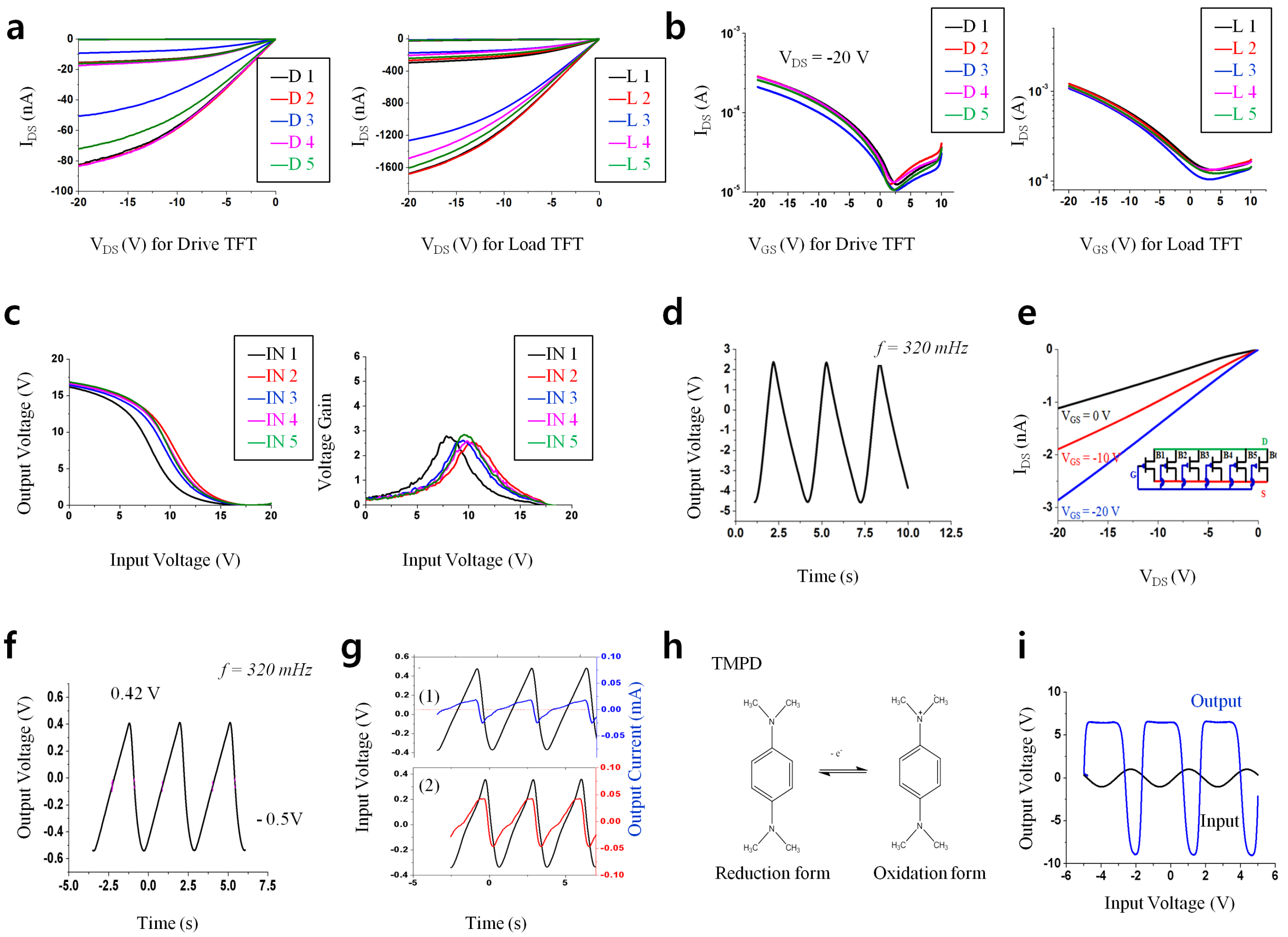

Figure 3 | (a and b) Output and transfer characteristics of the printed *cn*TFTs for 5 drive and 5 load TFTs respectively in the printed ring oscillator. (c) Electrical characteristics of inverters and (d) output characteristic of the ring oscillator. (e) Total output characteristics of the printed buffer unit consisting of 6 *cn*TFTs and a resistor (measured based on contacting gate and drain-source electrodes as shown in the inset circuit). (f) Modified triangular wave following the buffer unit. (g) Generated signals before scanning (black) and after scanning the electrochemical cells without (1, blue) and with TMPD (2, red) in a drop of solution. (h) TMPD structure for oxidation and reduction reaction. (i) The input and amplified output signals after passing through three amplifying inverters.

ductance, on-off current ratios and threshold voltage are shown in Table S2 (Supplementary Information). As the impedance was matched between the ring oscillator and the electrochemical cell using the buffer unit, another buffer unit (Figure 1b-⑥) is also needed between the electrochromic signage (20 KΩ) and the amplifier (1 MΩ) for impedance matching. The electrical parameters of the buffer are listed in Table S2 (Supplementary Information).

The amplified output signal was used to run the reduction and oxidation of the patterned conducting polymer in the electrochromic signage, and thus the blinking signage will be used to indicate the presence of chemicals in the specimens. The signage was fully printed and attached on printed CV tag. A clear concept of the printing sequence for the electrochromic signage was given in Figure S11 in Supplementary Information.

Discussion

The fully printed flexible CV tag was completed by assembling the printed circuits including ring oscillator, buffer, electrochemical cell, amplifier, and signage onto the previously R2R gravure printed rectenna. The resultant CV tag is shown in Figure 4a. The working concept of the wireless and flexible CV tag is demonstrated in the following sequences (watch the video file by clicking Figure S12 in Supplementary Information). After dropping 500 μl of TMPD solution (10 mM) on the printed electrochemical cell, the CV tag was placed on the custom made RF (13.56 MHz) reader (Figure S13 in Supplementary Information). The antenna was subsequently coupled to 13.56 MHz AC. The coupled AC was rectified into polarized DC (> ± 10 V) through two diodes and two capacitors which caused the ring oscillator to generate a pseudo triangular waveform with output voltage of 7 V at 320 mHz. This was then passed through the buffer unit to scan the electrochemical cell. The output signal after scanning the electrochemical cell was amplified to turn on the signage according to the concentration level of TMPD in the solution. In this work, whenever the concentration of TMPD was higher than 10 mM, the signage "PE" blinked (Figure 4b) while it did not show anything (Figure 4c) when it was lower than 10 mM. Furthermore, clear cyclic voltammograms for scanning the electrochemical cell with and without 10 mM of TMPD were obtained by re-plotting the output triangular voltage waveform (Figure 4d). The attained half potential ($E_{1/2}$ ~ 0.05 V, oxidation potential is 0.22 V and reduction potential is −0.17 V) of TMPD from the CV tag was nearly identical to the value obtained from a commercial CV instrument, SP-240, Biologic, which uses the same frequency for the triangular voltage waveform and printed electrochemical cell (Figure 4e). However, when the generated triangle wave frequency is higher than 0.6 Hz in the CV tag, the clear redox peaks cannot be

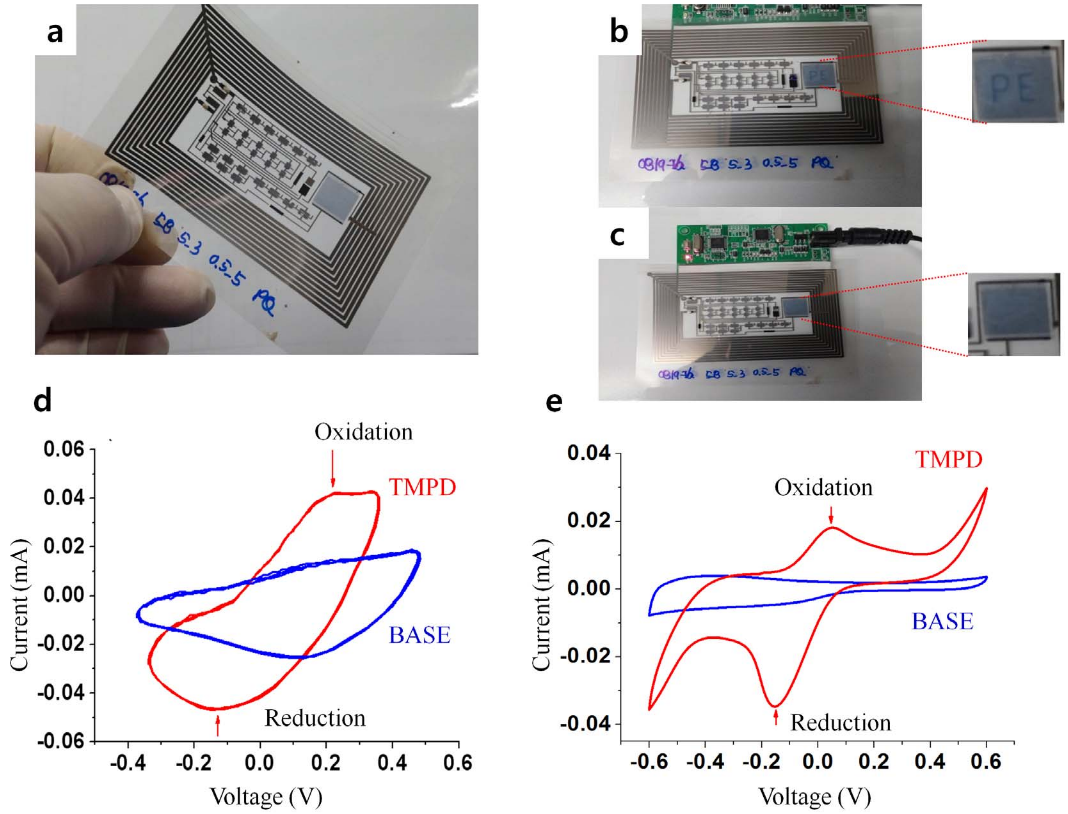

Figure 4 | (a) Optical image of fully printed wireless cyclic voltammetry tag. (b) Operation images of CV tag with and (c) without TMPD in the solution on 13.56 MHz reader. (d) Converted cyclic voltammogram from the printed wireless CV tag *vs* (e) a commercial CV instrument (please refer the interconnected video file for the demonstration of wireless CV tag operation in Figure S12 in Supplementary Information).

observed as shown in Figure S14 (Supplementary Information). Furthermore, as a typical example of testing a specimen in aqueous solution, 10 mM of $K_3(FeCN)_6$ aqueous solution was checked using the CV tag, and results of converted cyclic voltammogram was shown in Figure S15 (Supplementary Information). Those results support that the CV tag can examine specimens not only in organic solution, but also in aqueous one.

Up to present, there has been no commercial success in fully printed TFT-based electronic devices due to some kind of difficulties in integrating a number of printed TFTs through a scalable printing method. As increasing a number of integrated TFTs, the range of threshold voltage shits of TFTs needs to be controlled as narrow as possible to properly operate designed function of printed devices. However, the scalable printing method is not fully understood to integrate a number of TFTs on plastic foils while keeping the constant range of threshold voltage shifts because of difficulty in controlling a constant charge transport through the printed channels. Therefore, the printed TFT-based devices should be constructed by using a minimum number of printed TFTs while a function of printed device needs to be competitive over Si based one. Since the fully printed CV tag required the integration of only 26 *cn*TFTs, the acceptable range of threshold voltage shits of *cn*TFTs is controllable and achievable by using a scalable printing method such as a roll-to-roll gravure. Furthermore, the unique function of printed CV tags cannot be simply reproduced using a Si technology with a comparable cost to the R2R printing process. Therefore, the fully printed CV tag will be the first leading model to show a way of how printed TFT-based electronic devices should go for the successful launching in the IT markets.

In the core of the printed CV tag, fully gravure-printed *cn*TFTs were employed for constructing a circuit consisting of a ring oscillator to generate triangular waveforms, buffer units for an impedance matching, and an amplifier for enhancing output signal to turn on the signage. Those simple units can be combined with a variety of printable sensors and RF devices to create novel devices with an extremely low cost. In other words, the printed rectenna, ring oscillator, buffer and amplifier units in the printed CV tag will be key elements for the construction of variety of printed RF-sensors, this CV tag will be used as a platform for the further construction of a variety of RF based electrochemical sensors.

In conclusion, we have presented a fully printed flexible CV tag with the size of 9.5 × 5.5 cm^2 that can operate wirelessly using a 13.56 MHz RF reader. However, for the practical purpose, the size of the CV tag can be reduced up to the half (4 × 2 cm^2) of the current one by optimizing spaces between *cn*TFTs in the circuit of the CV tag and reducing the operation power to less than DC 5 V as the current size of 13.56 MHz antenna is designed to provide maximum coupled AC power using a sufficient length of printed antenna pattern (higher inductance). Furthermore, although we utilized R2P gravure, screen printer and drop casting method for the convenience of fabricating the CV tag in the Lab, the whole fabricating process can be repeated by R2R gravure because the rectenna, ring oscillator, buffers, electrochemical cell, amplifier and signage in the CV tag were all designed to compromise the limit (±20 μm) of a registration accuracy of a current R2R gravure system. Therefore, this technology could be used as a platform for disposable wireless electrochemical sensors to a variety of specimens in aqueous and organic solutions to monitor or diagnose pathogens and hazardous materials by modifying the electrodes of the electrochemical cells.

Methods

Roll-to-roll (R2R) gravure printed antenna, bottom electrodes, gate electrodes and wires. Two color units of R2R gravure system (Taejin Co. Korea: Figure S1 in the Supplementary Information) was employed to print antenna, bottom electrodes and wires on a roll of poly(ethylene terephthalate) (PET) film (width of 200 mm and thickness of 100 μm, purchased from SKC, Korea) with silver nanoparticle based conducting ink (PG-007 BB type, Paru Co. Korea) and the web transfer speed of 8 m/min under roll pressure of 0.5 MPa and web tension of 60 N. The silver ink was further formulated to meet the viscosity of 500 cp and surface tension of 47 mN/m using respectively ethylene glycol (Aldrich) and Dipropylene glycol methyl ether (Aldrich). The curing time of printed silver layer was 10 sec by passing through a heating chamber of 150°C, installed in R2R gravure system. The resulting printed silver patterns on PET roll showed the thickness of 450 (±50) nm without any defects and were shown in Figure S16a (Supplementary Information).

Roll-to-plate (R2P) gravure printed ring oscillator, buffer, and amplifier. Full R2P gravure system (Figure S1 in the Supplementary Information) was employed to print single walled carbon nanotube network based thin film transistors (*cn*TFTs), used as a building block to construct all the integrated circuit of printed cyclic voltammetry (CV) tag (Figure 1). 26 *cn*TFTs for the integration of circuit in printed CV tag were printed on the previously R2R gravure printed film (printed antenna, bottom electrodes and wires with thickness of 450 nm; Figure S16b in Supplementary Information) following printing sequences. First, $BaTiO_3$ nanoparticle based dielectric ink (PD-100, Paru Co., Korea) was further formulated to meet viscosity (75 cP) and surface tension (31.8 mN/m) using the ethyl 2-cynoacrylate (Aldrich). As shown in R2P gravure printing scheme (Figure S17 in Supplementary Information), the homogeneous and defect free dielectric layers were first R2P gravure printed with a printing speed of 200 mm/s and roll pressure 7.5 kg_f on previously printed gate electrodes and bottom electrodes (for capacitors) with a thickness of 2.3 (±0.1) μm and width of 940 (±2) μm (Figure S5 in Supplementary Information). The printed dielectric layers were cured for 1 min 20 sec at 150°C. Second, single walled carbon nanotube (SWNT) based semiconducting ink (PR-040, Paru Co., Korea) was diluted by diethylene glycol monobutyl ether (Daejung Co., Korea) to 1 : 1 volume ratio for printing active layers using R2P gravure on the printed dielectric layers with a printing speed of 350 mm/s and a roll pressure of 4 kg_f. The viscosity and surface tension of diluted SWNT ink were respectively 30 cP and 29 mN/m. The resulting SWNT printed films were dried 1 min at 150°C. For using *cn*TFTs in the buffer unit, we used the dilution volume ratio of 5 : 1 between SWNT ink (PR-040, Paru Co., Korea) and diethylene glycol monobutyl ether for printing only on the TFTs in the buffer unit to increase SWNT network density for providing the high current (Figure S5 in Supplementary Information). Finally, drain-source electrodes were R2P gravure printed using silver nanoparticle based ink (PG-007 AA type, Paru Co., Korea) on SWNT printed film with a viscosity of 800 cP and a surface tension of 42 mN/m. The R2P gravure printing was carried out under a printing speed of 200 mm/s and a role pressure of 5 kg_f.

Diode fabrication. Although Shottky diodes were able to print using R2R gravure based on our previous reported process[16,17], the drop casting method was used for the convenience in this work. First, dropping ZnO based hybrid semiconducting ink (PD-070, Paru Co., Korea) on printed silver electrode and then, Al ink (PA-009, Paru Co., Korea) was dropped on the top of ZnO ink. The resulting diodes were further cured for 5 min at 120°C.

Measurements. Formulated and diluted inks were characterized using DCAT21 (Dataphysics Co., Germany) and SV-10 Vibro viscometer (AND Co., Japan) for respectively measure surface tension and viscosity. All measurements in this work were carried out under ambient condition if there are no other comments. Resistance of printed antenna, bottom electrodes, gate electrodes and wires were measured using two probe methods with HP3457A multimeter. Both surface morphology and thickness of printed layers was characterized using surface profiler (NV-2200, Nanosystem, Korea). LCR meter (E4980A, Agilent, USA), semiconductor parameter analyzer (4155C, Agilent, USA) and digital phosphor oscilloscope (DPO 4054, Tektronix, USA) were respectively used to characterize printed *cn*TFT based circuits.

1. Sanderson, M. D., Kamplain, J. K. & Bielawski, C. W. Quinone-annulated N-heterocyclic carbene-transition-metal complexes: Observation of π-back bonding using FT-IR spectroscopy and cyclic voltammetry. *J. Am. Chem. Soc* **128**, 16514–16515 (2006).
2. Lai, R. Y., Fabrizio, E. F., Lu, L., Jenekhe, S. A. & Bard, A. J. Synthesis, cyclic voltammetrics studies, and electrogenerated chemiluminescence of a new donor-acceptor molecule: 3,7-[Bis[4-phenyl-2-quinolyl]]-10-methylphenothiazine. *J. Am. Chem. Soc* **123**, 9112–9118 (2001).
3. Armstrong, F. A., Hill, H. A. & Walton, N. J. Direct electrochemistry of redox proteins. *Acc. Chem. Res* **21**, 407–413 (1998).
4. Norouzi, P., Ganjali, M. R., Daneshgar, P., Alizadeh, T. & Mohammadi, A. Development of fast Fourier transformation continuous cyclic voltammetry as a highly sensitive detection system for ultra-trace monitoring of penicillin V. *Anal. Biochem* **360**. 175–184 (2007).
5. Koyama, J. *et al.* Correlation of redox potentials and inhibitory effects on Epstein-Barr virus activation of 2-azaanthraquinones. *Cancer Letters* **212**, 1–6 (2004).
6. Teijeiro, C., Perez, P., Marin, D. & Paleček, E. Cyclic voltammetry of mitomycin C and DNA. *Bioelectrochemistry and Bioenergetics* **38**, 77–83 (1995).
7. Zerihun, T. & Gründler, P. Electrically heated cylindrical microelectrodes. Determination of lead on Pt by cyclic voltammetry and cathodic stripping analysis. *J. Electroanal Chem* **415**, 85–88 (1996).
8. Bond, A. M. & Scholz, F. Electrochemical, Thermodynamic, and mechanistic data derived from voltammetric studies on insoluble metallocenes, mercury halide and sulfide compounds, mixed silver halide crystals, and other metal complexes following their mechanical transfer to a graphite electrode. *Langmuir* **7**, 3197–3204 (1991).
9. Brusciotti, F. & Duby, P. Cyclic voltammetry study of arsenic in acidic solutions. *Electrochim Acta* **52**, 6644–6649 (2007).
10. Ruan, C., Yang, L. & Li, Y. Immunobiosensor chips for detection of escherichia coli O157:H7 using electrochemical impedance spectroscopy. *Anal. Chem* **74**, 4814–4820 (2002).
11. Liu, G. & Lin, Y. Electrochmical sensor for orgnohposphate pesticides and nerve agents using zirconia nanoparticles as selective sorbents. *Anal. Chem* **77**, 5894–5901 (2005).
12. Ascaso, J. F. *et al.* Diagnosing insulin resistance by simple quantitative methods in subjects with normal glucose metabolism. *Diabetes Care* **26**, 12, 3320–3325 (2003)
13. Karl, T. *et al.* Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling. *J. Appl. Physiol* **91**, 762–770 (2001).
14. Armstrong, F. A. Recent developments in dynamic electrochemical studies of adsorbed enzymes and their active sites. *Chem. Biol* **9**, 110–117 (2005).
15. Mollah, M. Y., Schennach, R., Parga, J. R. & Cocke, D. L. Electrocoagulation (EC)-science and applications. *J. Hazard. Mater* **B84**, 29–41 (2001).
16. Park, H. *et al.* Fully roll-to-roll gravure printed rectenna on plastic foils for wireless power transmission at 13.56 MHz. *Nanotech* **23**, 344006–344012 (2012).
17. Kang, H. *et al.* Fully Roll-to-Roll Gravure Printable Wireless (13.56 MHz) Sensor-Signage Tags for Smart Packaging. *Sci. Rep*. **4**, 5387; DOI: 10.1038/srep05387 (2014).
18. Jung, M. *et al.* All-printed and roll-to-roll-printable 13.56 MHz-operated 1-bit RF tag on plastic foils. *IEEE Trans Electron Devices* **57**, 3, 571–580 (2010).
19. Noh, J. *et al.* Integrable single walled carbon nanotube (SWNT) network based thin film transistors using roll-to-roll gravure and inkjet. *Org. Electron* **12**, 2185–5191 (2011).
20. Noh, J. *et al.* Modeling of printed single walled carbon nanotube thin film transistors for attaining optimized clock signals. *J. Appl. Phys* **108**, 10, 102811–102817 (2010).
21. Noh, J. *et al.* Fully Gravure-Printed Flexible Full Adder Using SWNT-Based TFTs. *IEEE. Eletron. Dev. Lett.* **33**, 11, 1574–1576 (2012).

Acknowledgments

The authors would like to thank BK Plus program, Basic Science Research program (NRF-2014R1A1030419), and RIC in Sunchon National University for their financial support. G. C. thanks to Paru Co. for providing all electronic inks in this work.

Author contributions

Y.J., A.J. and G.C. designed the experiments. Y.J., H.P., J.P., Y.C. and M.J. carried out the experiments. J.N., K.J. and K.C. performed device simulations. Y.J. and K.C. performed mobility calculations. Y.J. and M.P. performed electrochemistry and analyzed electrochemical data. Y.J., H.P., M.P., A.J. and G.C. contributed to analyzing the data. M.J., M.P., A.J. and G.C. wrote the paper while all authors provided feedback.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: Prof. Cho was a consultant of PARU Co. Korea and received compensation from PARU Co. Both Dr. M. Jung and Mr. K Jung has been employee of PARU Co. Korea. Dr. Y. Jung, Miss H. Park, Ms. J. Park, Dr. J. Noh, Mr. Y. Choi, Prof. Pyo, Mr. K. Chen and Prof. Javey declare no competing financial interests.

How to cite this article: Jung, Y. *et al.* Fully printed flexible and disposable wireless cyclic voltammetry tag. *Sci. Rep.* **5**, 8105; DOI:10.1038/srep08105 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

View publication stats

AIP | Applied Physics Letters

A microwave metamaterial with integrated power harvesting functionality

Allen M. Hawkes, Alexander R. Katko, and Steven A. Cummer

Citation: Applied Physics Letters **103**, 163901 (2013); doi: 10.1063/1.4824473
View online: http://dx.doi.org/10.1063/1.4824473
View Table of Contents: http://scitation.aip.org/content/aip/journal/apl/103/16?ver=pdfcov
Published by the AIP Publishing

---

www.goodfellowusa.com

Goodfellow

metals • ceramics • polymers
composites • compounds • glasses

Save 5% • Buy online

70,000 products • Fast shipping

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

APPLIED PHYSICS LETTERS **103**, 163901 (2013)

A microwave metamaterial with integrated power harvesting functionality

Allen M. Hawkes,[a)] Alexander R. Katko, and Steven A. Cummer
Department of Electrical and Computer Engineering, Duke University, 101 Science Drive, Durham, North Carolina 27708-9976, USA

(Received 30 May 2013; accepted 22 September 2013; published online 14 October 2013)

We present the design and experimental implementation of a power harvesting metamaterial. A maximum of 36.8% of the incident power from a 900 MHz signal is experimentally rectified by an array of metamaterial unit cells. We demonstrate that the maximum harvested power occurs for a resistive load close to 70 Ω in both simulation and experiment. The power harvesting metamaterial is an example of a functional metamaterial that may be suitable for a wide variety of applications that require power delivery to any active components integrated into the metamaterial. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4824473]

Metamaterials are composed of sub-wavelength particles that exhibit bulk properties that are different from their individual components. Electromagnetic metamaterials are engineered materials that can achieve parameters not possible within naturally occurring materials, such as a negative index of refraction[1] or a zero index of refraction.[2] Exotic properties like these allow for a variety of interesting applications including a superlens device[3] and an invisibility cloak.[4] Integrating active and nonlinear functionality into metamaterials has been demonstrated in the form of dynamic resonant frequency tuning,[5,6] phase conjugation,[7] and wave mixing.[8,9] More specific functional behavior has also been demonstrated in metamaterials, including radio frequency (RF) limiting[10] and harmonic generation.[11]

Metamaterials are also well-suited for other functional behaviors, including electromagnetic power harvesting, the focus of this work. Power harvesting devices convert one type of energy to another, typically converting to a direct current (DC) signal. Many types of energy can be harvested, from acoustic (using a piezoelectric harvester)[12] to electromagnetic (using a rectenna).[13] Power harvesting devices require a method to couple to the energy that will be harvested as well as a device to convert the energy from one form to another. By their very nature, metamaterials are designed to couple to various types of energy, e.g., from acoustic[14] to optical,[15] and thus provide a natural platform for power harvesting. Electromagnetic metamaterials provide flexibility in design due to their electrically small, low-profile nature.[16] Since metamaterials are typically designed as infinite arrays, the resonant frequency and input impedance include coupling effects. Metamaterials can be adapted to various applications, such as flexible sheets to cover surfaces.[17] Moreover, many metamaterials that have been presented in the literature require some form of external signal. This could be a DC bias voltage[18] or a large external pump signal.[7] In general, metamaterials could be modified to harvest such an external signal that is already present for other purposes. With these design advantages, power harvesting metamaterials offer design flexibility for a large number of applications that general antenna-based microwave power harvesting devices may lack.

A recent simulation-based study[19] investigated the conversion efficiency between incident RF power and induced power in a split-ring resonator (SRR). Our work is focused on the experimental measurement of RF to DC efficiency based on the conventional effective area of the SRR. We demonstrate that metamaterials can also include embedded devices to convert the incident RF energy to a DC voltage, providing a platform for power harvesting that utilizes the advantages of metamaterial design.

An SRR is a canonical example of a resonant metamaterial particle and is used as the basis for the unit cells of the metamaterial power harvester designed here. By tuning the SRR parameters, we design an SRR (Fig. 1) to resonate at 900 MHz using an S-parameter simulation within Computer Simulation Technology (CST) Microwave Studio software. Using CST Microwave Studio, we can also simulate the effects of embedding devices within the SRR by retrieving its S-parameters using a lumped port. The retrieved S-parameters are loaded into Agilent Advanced Design System (ADS), allowing us to simulate both fullwave 3D effects and circuit-level nonlinear effects.

An SRR couples strongly to an incident magnetic field and can be loaded with a wide variety of circuit elements. In this work, we embed a rectifying circuit within an SRR to convert the incident RF power to DC power. A number of rectifying circuits could be chosen depending on the particular application for the power harvesting metamaterial. We choose to use a Greinacher[20] circuit because the output voltage is double the input voltage maximum, which allows for

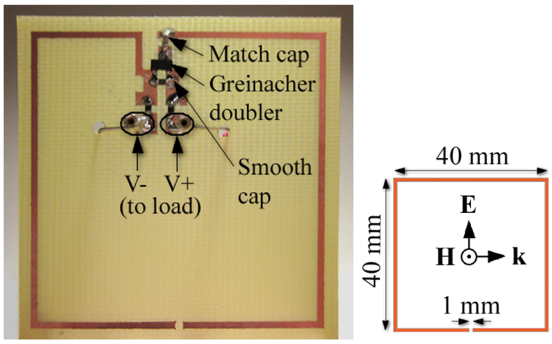

FIG. 1. (Left) Power harvesting SRR. (Right) Plain SRR with dimensions shown, 1 mm traces. Incident wave polarization is also shown.

[a)] allen.hawkes@duke.edu

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

easier power transmission and measurement. A Greinacher circuit has a lower effective capacitance in comparison to other rectification circuits (such as a bridge rectifier), allowing faster switching and thus a higher frequency of operation. A Greinacher circuit also has a low threshold voltage, allowing operation at lower incident power levels. A Greinacher voltage doubler can be placed across a gap (Fig. 1) in the top side of the copper trace to rectify the induced current present in the SRR. Schottky diodes are used for the voltage doubler due to their typically low open junction capacitance and fast switching capabilities, which allow for rectification of a high-frequency RF signal, as well as their typically low threshold voltage. A resistive load placed across the output of the voltage doubler is a simple way to determine DC power out using $P = V^2/R$. This DC power is maximized through ADS for matching capacitor and resistor values. The parameters of our selected Schottky diode (HSMS 2862) and the simulated S-parameters of the above SRR in a model of our actual parallel-plate waveguide are input into ADS for the simulation, shown in a schematic in Fig. 2.

The simulated components lead to a maximum efficiency of 61% for an input power of 24.25 dBM, the maximum available experimental power (at an incident power density of approximately 1.6 mW/cm^2). One important figure of merit for a power harvester is its RF-to-DC power conversion efficiency:

$$\eta = \frac{P_{DC}}{P_{RF}}.$$

We determine P_{RF} by measuring the total incident power in our measurement apparatus. For a large metamaterial sample, we assume that the total incident power to the measurement apparatus is incident on the metamaterial. For a single unit cell, it is necessary to use the effective area of the unit cell to determine the incident P_{RF}. The maximum effective area may be calculated by[21]

$$A_{e,max} = \frac{\lambda^2}{4\pi} D_0,$$

where $D_0 = 1.5$ since the SRR is effectively a small loop illuminated by a transverse electromagnetic (TEM) wave. For an SRR resonant at 900 MHz, the effective area is thus $A_{e,max} = 5.3A_{physical}$. The full waveguide area is approximately $6.8A_{physical}$ where $A_{physical}$ is the physical area of a single SRR unit cell. As $A_{e,max}$ is 78% of the open wave guide area, 78% of the incident power density is used as input power for simulation of the single cell.

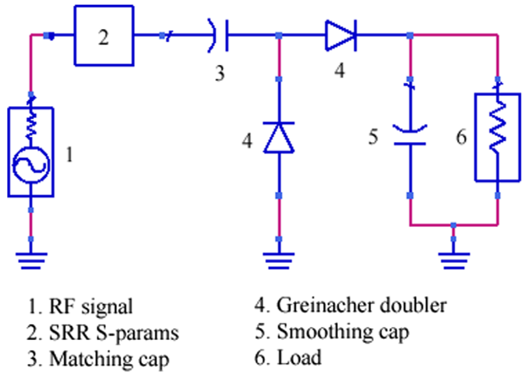

FIG. 2. ADS simulation schematic of SRR power harvester. CST Microwave Studio was used to determine the SRR S-parameters.

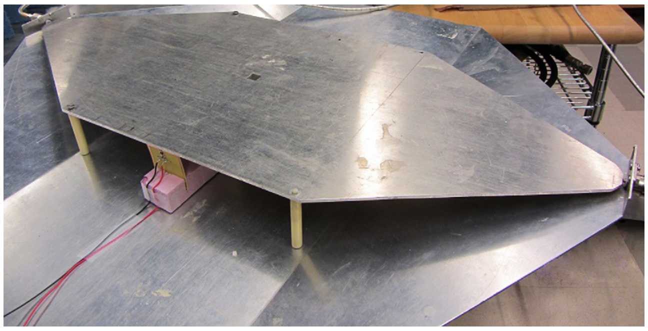

FIG. 3. Placement of SRR within open waveguide.

The designed voltage doubler and resistive load are added to an SRR as shown in Fig. 1, resulting in the power harvesting metamaterial unit cell. To observe power harvesting capabilities, the cell is placed in an open, TEM waveguide (Fig. 3) where input power is produced by a signal generator and amplifier, and output power is measured with an oscilloscope via leads placed across the resistive load (Fig. 4). The DC power harvested is determined by $P = V^2/R$ as previously mentioned, and input power is measured with a spectrum analyzer connected to the signal generator and amplifier via the open waveguide. By increasing the incident power from 13 to 24 dBm and measuring the DC output from the SRR, the normalized harvested power, $P_{DC}/P_{RF,incident}$, as a function of incident power $P_{RF,incident}$ and resulting efficiency are determined at each point. The maximum efficiency of the single cell is 14.2%, setting $P_{RF,incident}$ as 78% of the total input power from effective area calculations.

Multiple power harvesting SRR cells are then tested simultaneously to create the power harvesting metamaterial, which is accomplished through a 5 × 1 array shown in Fig. 5. Through a parallel connection of the leads from each SRR's resistive load, the total power harvested by the metamaterial is found in the same way as the single cell. The maximum efficiency for the 5 × 1 array is 36.8%, where P_{in} is the entire input power because the array spans the entire length. Measured efficiencies of both the single and the array of power harvesting SRRs are shown in Fig. 6. Also shown in Fig. 6 is the open circuit voltage, V_{OC}, a load-independent measure of the available voltage harvested by the array of SRRs.

The power harvesting metamaterial array is more efficient than the single unit cell. This is partially due to the

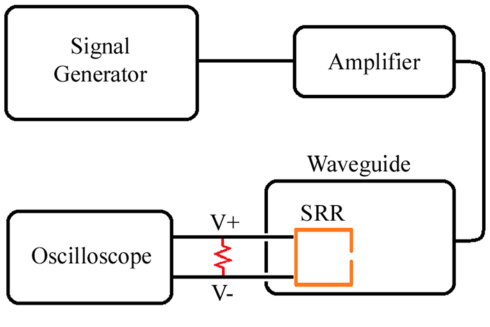

FIG. 4. Experimental test setup schematic.

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

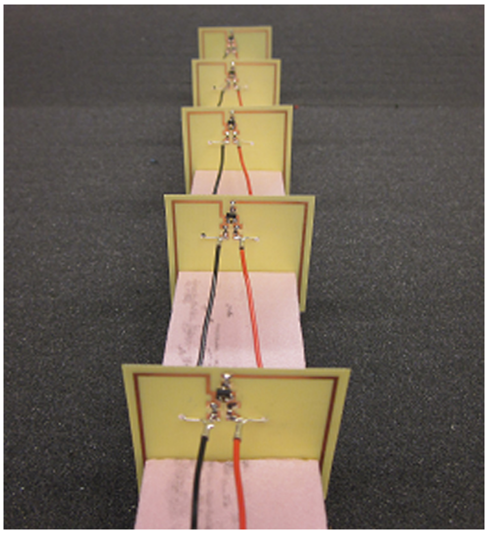

FIG. 5. 5 × 1 array of power-harvesting SRRs.

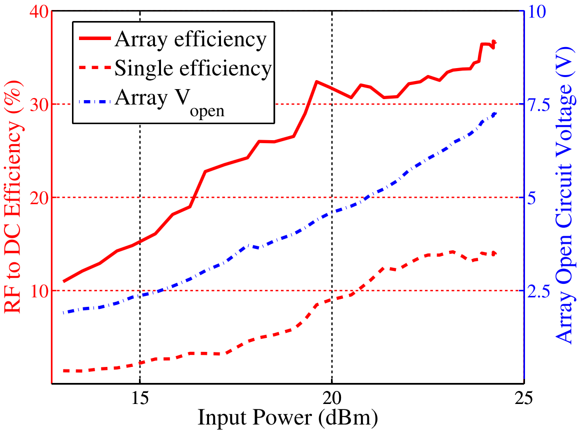

FIG. 6. RF to DC efficiency of power harvesters. The open-circuit voltage of the array is also shown.

larger effective area of the array. While the effective area of the single cell is 78% the entire width of the waveguide, some of this power is not harvested by the single unit cell due to fringing effects on the sides of the waveguide. For this reason, placement of multiple cells that together span the entire waveguide width results in a higher efficiency as the array captures more of the electromagnetic energy that undergoes the fringing effect.

Another important relationship is the efficiency of the power harvesting metamaterial as a function of load resistance. Simulated and experimental efficiencies for the 5 × 1 array are shown in Fig. 7. Though the experimental efficiencies do not match the values from simulation, the simulation does accurately predict the maximum harvested power as falling approximately within the range of 70–80 Ω, confirmed by the experimental data. The experimental efficiency maximum occurs at 70 Ω, and the simulated maximum occurs at 82 Ω, showing close correspondence.

In summary, we have designed, simulated, and experimentally measured a functional metamaterial power harvester capable of converting up to 36.8% of the incident RF power to DC power. Through a parallel connection of five SRRs, a V_{OC} of 7.3 V is achieved. Simulations match

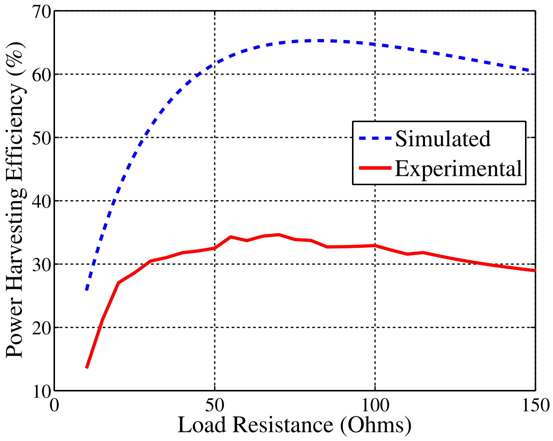

FIG. 7. RF to DC efficiency of SRR array as a function of load resistance, for both simulation and experiment. Note that both show maximum efficiency for around 70–80 Ω.

experimental results showing an optimal resistive load for DC power transfer of 70–80 Ω. The SRR power harvester is an example of functional metamaterial that may be suitable for a wide variety of RF applications that require power delivery to any active integrated components.

This work was supported by a Multidisciplinary University Research Initiative from the Army Research Office (Contract No. W911NF-09-1-0539).

[1] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz, Phys. Rev. Lett. **84**, 4184–4187 (2000).
[2] R. W. Ziolkowski, Phys. Rev. E **70**, 046608 (2004).
[3] N. Fang, H. Lee, C. Sun, and X. Zhang, Science **308**(5721), 534–537 (2005).
[4] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith, Science **314**(5801), 977–980 (2006).
[5] I. V. Shadrivov, S. K. Morrison, and Y. S. Kivshar, Opt. Express **14**(20), 9344–9349 (2006).
[6] H.-T. Chen, J. F. O'Hara, A. K. Azad, A. J. Taylor, R. D. Averitt, D. B. Shrekenhamer, and W. J. Padilla. Nature Photon. **2**(5), 295–298 (2008).
[7] A. R. Katko, S. Gu, J. P. Barrett, B.-I. Popa, G. Shvets, and S. A. Cummer, Phys. Rev. Lett. **105**, 123905 (2010).
[8] D. Huang, A. Rose, E. Poutrina, S. Larouche, and D. R. Smith, Appl. Phys. Lett. **98**(20), 204102 (2011).
[9] A. Rose, D. Huang, and D. R. Smith, Phys. Rev. Lett. **110**, 063901 (2013).
[10] A. R. Katko, A. M. Hawkes, J. P. Barrett, and S. A. Cummer, IEEE Antennas Wireless Propag. Lett. **10**, 1571–1574 (2011).
[11] I. V. Shadrivov, A. B. Kozyrev, D. W. van der Weide, and Y. S. Kivshar, Appl. Phys. Lett. **93**(16), 161903 (2008).
[12] M. Lallart, D. Guyomar, C. Richard, and L. Petit, J. Acoust. Soc. Am. **128**, 2739–2748 (2010).
[13] N. Zhu, R. W. Ziolkowski, and H. Xin, Appl. Phys. Lett. **99**(11), 114101 (2011).
[14] J. Li and C. T. Chan, Phys. Rev. E **70**(5), 055602 (2004).
[15] H. J. Lezec, J. A. Dionne, and H. A. Atwater, Science **316**(5823), 430–432 (2007).
[16] R. W. Ziolkowski and A. Erentok, IEEE Trans. Antennas Propag. **54**(7), 2113–2130 (2006).
[17] G. X. Li, S. M. Chen, W. H. Wong, E. Y. B. Pun, and K. W. Cheah, Opt. Express **20**(1), 397–402 (2012).
[18] J. Zhao, Q. Cheng, J. Chen, M. Q. Qi, W. X. Jiang, and T. J. Cui, New J. Phys. **15**, 043049 (2013).
[19] O. M. Ramahi, T. S. Almoneef, M. Alshareef, and M. S. Boybay, Appl. Phys. Lett. **101**, 173903 (2012).
[20] A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 6th edition (Oxford University Press, 2009).
[21] C. A. Balanis, *Antenna Theory: Analysis and Design* (Wiley-Interscience, New York, 2005).

This article is copyrighted as indicated in the abstract. Reuse of AIP content is subject to the terms at: http://scitation.aip.org/termsconditions. Downloaded to IP: 152.3.43.156 On: Fri, 08 Nov 2013 20:28:30

Ambient RF Energy Harvesting in Urban and Semi-Urban Environments

Manuel Piñuela, *Student Member, IEEE*, Paul D. Mitcheson, *Senior Member, IEEE*, and
Stepan Lucyszyn, *Senior Member, IEEE*

***Abstract*—RF harvesting circuits have been demonstrated for more than 50 years, but only a few have been able to harvest energy from freely available ambient (i.e., non-dedicated) RF sources. In this paper, our objectives were to realize harvester operation at typical ambient RF power levels found within urban and semi-urban environments. To explore the potential for ambient RF energy harvesting, a city-wide RF spectral survey was undertaken from outside all of the 270 London Underground stations at street level. Using the results from this survey, four harvesters (comprising antenna, impedance-matching network, rectifier, maximum power point tracking interface, and storage element) were designed to cover four frequency bands from the largest RF contributors (DTV, GSM900, GSM1800, and 3G) within the ultrahigh frequency (0.3–3 GHz) part of the frequency spectrum. Prototypes were designed and fabricated for each band. The overall end-to-end efficiency of the prototypes using realistic input RF power sources is measured; with our first GSM900 prototype giving an efficiency of 40%. Approximately half of the London Underground stations were found to be suitable locations for harvesting ambient RF energy using our four prototypes. Furthermore, multiband array architectures were designed and fabricated to provide a broader freedom of operation. Finally, an output dc power density comparison was made between all the ambient RF energy harvesters, as well as alternative energy harvesting technologies, and for the first time, it is shown that ambient RF harvesting can be competitive with the other technologies.**

***Index Terms*—Ambient RF, energy harvesting, maximum power point tracking (MPPT), multiband, rectenna, RF survey, RF-dc.**

I. Introduction

FOR ALMOST 50 years, far-field RF technology has been used to remotely power systems from relatively large unmanned helicopters [1] to very small smart dust sensors [2] and contact lenses that measure eye pressure [3]. With all these systems, a dedicated RF source is used, where the operator may have control over the effective isotropically radiated power (i.e., both transmit power and antenna characteristics), beam pointing and polarization of the RF source, ensuring optimal line-of-sight operation between the source transmitter (TX) and harvesting receiver (RX). It is important to highlight that this work will focus only in radiative power transfer and not inductive or near-field power transfer, as demonstrated in [4]. A more convenient solution, however, is to power these devices from ambient RF energy sources, such as television and mobile phone signals, thus removing the need for a dedicated source. As ambient RF levels are lower than those that can be provided by a dedicated RF source, the efficiency of the harvesting system, and its minimum startup power are of critical importance.

In order to assess the feasibility of deploying ambient RF energy harvesters, the available RF power needs to be measured in different locations. Such measurements, in conjunction with knowledge on harvester performance, can then be used to determine the locations at which RF harvester powered devices can be successfully deployed. Several RF spectral surveys, which measure ambient RF power levels from sources such as television and mobile phone base stations, have been previously reported. Many have been undertaken using personal exposimeters or spectrum analyzers, where the exact location of each measurement is unknown and with RF power levels only being reported under general scenarios (e.g., outdoor, indoor, street, bus, etc.) [5], [6]. While being of academic interest for health-related research [7], the lack of power level and specific time/location information limits their usefulness for exploitation in ambient RF energy harvesting applications.

Most rectennas (normally comprising an antenna, impedance matching network, rectifier, storage element, and load) presented in the open literature have been tested using dedicated sources rather than harvesting from ambient RF energy [8]. In recent years, efficiencies as high as 78% [9] and 90% [10] have been demonstrated with relatively high input RF power levels (i.e., $> +10$ dBm). Moderate efficiencies have also been achieved using dedicated TXs that provided relatively low input RF power levels; e.g., an efficiency of 60% was achieved with -22.6-dBm input power [11]. In one demonstrator [12], designed to operate over a broad range of input RF power levels, (-30 to $+30$ dBm), the efficiency increased from 5% at a low input RF power to a peak of 80% at $+25$ dBm.

Despite advancements in end-to-end (i.e., input RF to output dc) power conversion efficiencies at low input RF power levels (similar to those measured in the spectral surveys), only a few attempts at true ambient RF energy harvesting have been reported. For example, one relatively efficient rectenna, utilizing a modified omnidirectional patch antenna, has an efficiency of 18% with a single-tone input RF power of -20 dBm [13]. This dedicated signal source was meant to emulate the input RF power levels measured from a nearby digital TV (DTV) TX in Tokyo, Japan, but did not take into account the more realistic effect of harvesting from a modulated broadband signal. Another attempt

Manuscript received November 20, 2012; revised April 24, 2013; accepted April 26, 2013. Date of publication May 24, 2013; date of current version June 28, 2013. This work was supported in part by the U.K. Government and the Mexican National Council of Science and Technology (CONACYT).

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: m.pinuela09@imperial.ac.uk; paul.mitcheson@imperial.ac.uk; s.lucyszyn@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMTT.2013.2262687

0018-9480/$31.00 © 2013 IEEE

at harvesting ambient RF energy from a mobile phone base station at 845 MHz was reported in [14]. This prototype managed to power an LCD thermometer for 4 min, but only after harvesting for 65 h. In that work, when the authors used a dedicated signal source with a single-tone input RF power of -15 dBm, an efficiency of 3% was recorded. A batteryless location sensor has also been demonstrated [15], powered by a rectenna with a printed antenna on a flexible substrate and a solar cell, although no details for the RF-dc efficiency were reported. Finally, successful prototypes capable of harvesting energy using TV antennas were presented, but again no details of their efficiency were given [16], [17].

In order to demonstrate the feasibility for implementing ambient RF energy harvesting, here we first present the results of a citywide RF spectral survey, indicating suitable locations and associated RF bands with sufficient input RF power density levels for harvesting. Based on these results, rectennas were then fabricated and their efficiencies, under ambient RF energy harvesting operation, were calculated using *in-situ* field strength measurements. Furthermore, an investigation of multiband rectenna arrays is also presented, demonstrating the tradeoffs between series (voltage summing) and parallel (current summing) topologies with the aim of reducing the minimum input power required for harvester operation. Finally, a comparison between measured ambient RF energy harvesting and alternative forms of energy harvesting technologies is presented; highlighting, for the first time, the practical feasibility of exploiting existing freely available sources of RF energy.

II. LONDON RF SURVEY

In order to quantify input RF power density levels present in a typical urban and semi-urban environment, a citywide RF spectral survey within the ultrahigh frequency (0.3–3 GHz) part of the frequency spectrum was conducted within Greater London. A number of citywide RF spectral surveys have previously been conducted, but in general, only a few samples were taken, giving little insight into (semi-)urban environments [14], [18], [19]. Other surveys [20], [21] compare their measurements relative to the distance from the nearest TX. In a (semi-)urban environment, this may not provide enough information about the RF spectrum since there is likely to be local geographical variations in base-station density and propagation characteristics (e.g., multipath effects and diffraction around and attenuation through buildings).

Each station on the London Underground network was used as a survey point to provide a robust dataset for representing Greater London in terms of geographical distribution and population density, having a combination of urban (in the center) and semi-urban (in surrounding areas) characteristics. Measurements were taken at each of the 270 stations (from a randomly chosen exit, at street level and a height of 1.6 m). To provide traceability and for use as a historical reference, time stamps and GPS locations were recorded. In addition, measurements were taken inside a building at Imperial College London (ICL), to represent a typical office block within an urban environment.

A. Methodology

Mobile phone usage varies during the daytime, and hence, ambient RF energy in their bands is expected to be time dependant, with more energy available during the daytime than at night time. Therefore, in order to be able to make fair comparisons between locations, measurements were taken between 10:00 am and 3:00 pm on weekdays over a period of one month (between March 5, and April 4, 2012). Electric field strength was measured between 0.3–2.5 GHz using an Agilent N9912A FieldFox RF analyzer [22] with a calibrated Aaronia BicoLOG 20300 omnidirectional antenna [23]. It is important to note that the spectral measurements were undertaken during the analog-to-digital switchover period in the U.K. and so the measurements for DTV may represent an underestimate of present RF power levels measured now that the switch over is complete [24]. It should also be noted that this survey was conducted prior to the 4G network being switched on within the U.K.

A "panning method," which complies with international regulations for measuring exposure limits, was used [25]–[27]. Here, the calibrated antenna is rotated to three orthogonal axes while the spectrum analyser is set to "max-hold," ensuring that the maximum reading is recorded. For each measurement, more than 1 min was allocated to allow for more than three sweeps across the selected frequency range. Additionally, to maintain a comparable signal-to-noise (S/N) ratio, attenuation was introduced (with a minimum set at 5 dB) to avoid compression when high input RF power levels were detected. For all measurements, the resolution bandwidth (BW) was fixed at 100 kHz, the internal amplifier was turned on and the highest resolution of 1001 points was selected. These settings provide the ability to obtain a snapshot of the power density that can be expected in an urban or semi-urban environment from continuously variable sources.

B. Results

After inputting the manufacturers' frequency-banded antenna factors into the spectrum analyzer, to ensure a fully calibrated system, the electric field strength measurements were taken. The input RF power density (S) is then calculated from the electric field strength measurement. Fig. 1 shows the input RF power density measured outside the Northfields London Underground station, where the spectral bands for DTV, GSM900, GSM1800, 3G, and Wi-Fi can been clearly identified.

A well-designed rectenna should ideally be capable of harvesting energy across an entire band, and thus it is important to calculate the total band power. The banded input RF power density S_{BA} (nW/cm^2) is calculated by summing all the spectral peaks across the band (i.e., in a similar way, the spectrum analyzer calculates channel power). These levels provide a snapshot of source availability at the time and location of the measurement. Moreover, they are used as a harvester design starting point since the power density at each band will define the input impedance of a rectenna.

The exact frequencies for each band are set by the U.K.'s official frequency band allocation [28]; the GSM900, GSM1800 and 3G base transmit (BTx) bands were separated from the associated mobile transmit (MTx) bands. Table I shows average

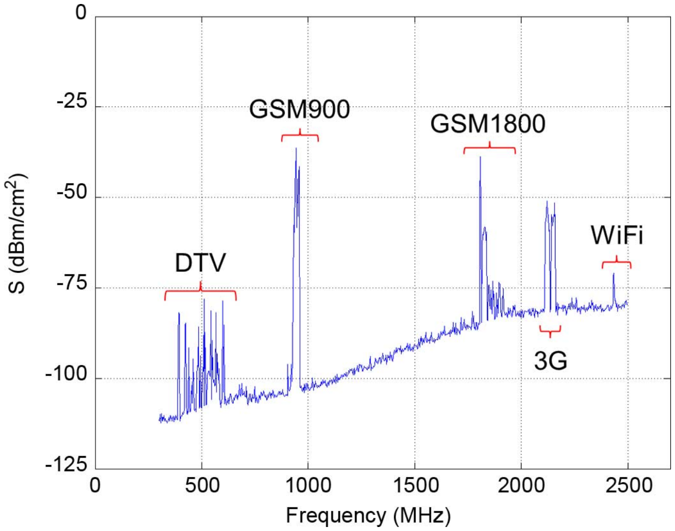

Fig. 1. Input RF power density measurements outside the Northfields London Underground station.

TABLE I
SUMMARY OF LONDON RF SURVEY MEASUREMENTS

Band	Frequencies (MHz)	Average S_{BA} (nW/cm^2)	Maximum S_{BA} (nW/cm^2)
DTV (during switch over)	470-610	0.89	460
GSM900 (MTx)	880-915	0.45	39
GSM900 (BTx)	925-960	36	1,930
GSM1800 (MTx)	1710-1785	0.5	20
GSM1800 (BTx)	1805-1880	84	6,390
3G (MTx)	1920-1980	0.46	66
3G (BTx)	2110-2170	12	240
WiFi	2400-2500	0.18	6

TABLE II
INPUT RF POWER DENSITY THRESHOLD: LONDON UNDERGROUND STATIONS WITHIN CENTRAL ZONES 1–3 (URBAN) AND OUTER ZONES 4–9 (SEMI-URBAN)

Band	S_{BA} *Threshold* (nW/cm^2)	Number of Stations	
		Urban	Semi-urban
DTV (during switch over)	40	10	0
GSM900	230	8	2
GSM1800	450	7	3
3G	62	6	4

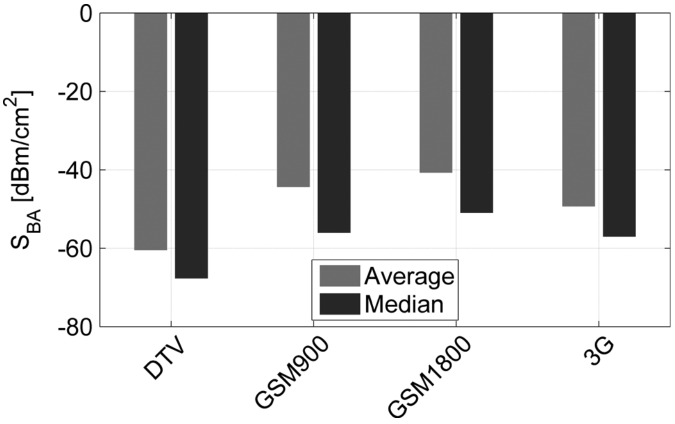

Fig. 2. Banded input RF power density measurements for the four largest ambient sources in Greater London.

RF power levels across all London Underground stations for the banded input RF power density measurements. It can be seen that all base-station transmit levels are between one and three orders of magnitude greater than the associated MTx levels. For this reason, and the fact that the population of transmitting mobile phones in close proximity of the harvester is highly variable, only base-station TXs will be considered further.

From our London RF survey, DTV, GSM900, GSM1800, 3G, and Wi-Fi were identified as potentially useful ambient RF energy harvesting sources, although DTV appears to be heavily dependent on line-of-sight and sudden changes in atmospheric conditions (e.g., temperature inversion) and Wi-Fi is very dependent on user traffic. It should be noted that the mobile phone base-station TXs employ vertically polarized antennas, placing a constraint on harvester orientation in deployment. With DTV, within the U.K., the main TXs have horizontally polarized antennas, while repeater TXs have vertically polarized antennas.

It is convenient to define the boundary between urban and semi-urban environments by the line that separates zones 3 and 4 on the London Underground map [29]. As one would expect, the central zones 1–3 host the highest density of base stations. As shown in Table II, a banded input RF power density threshold was selected to filter the ten London Underground stations with the highest measurements for each band. With DTV, the highest recorded measurements were all found within the urban environment. This is because Greater London's main DTV TX (at Crystal Palace) is located on the southeastern boarder of zones 3 and 4 and there are no London Underground stations further south. With mobile phones, more than 50% of the stations were inside the urban environment and those in a semi-urban environment were all located in close proximity to a cluster of base-station TXs.

Using the complete dataset from the London RF survey, Fig. 2 shows the average and median of the banded input RF power density measurements for the four largest ambient RF sources in Greater London. It can be seen that more than half of the locations have below average power levels. This is due to the fact that several stations had maximum levels that were considerably higher than the average because of their close proximity to TV TXs (e.g., Crystal Palace), extremes in base-station density and propagation characteristics.

In addition to the London RF survey, measurements within the Department of Electrical and Electronic Engineering building at ICL were taken on the 11th floor of the south stairwell. These are shown in Table III. As can be seen, DTV and GSM900 have a higher than average power level, due to a near line of sight from the TV TX and a close proximity to the 2G GSM900/1800 base stations.

The dataset from the London RF survey, with all relevant information (e.g., locations, timestamps, and banded input RF power density measurements), can be found at our interactive website: www.londonrfsurvey.org [30]. These measurements were used to design efficient harvesters and compared to ICL

TABLE III
MEASURED BANDED INPUT RF POWER DENSITIES AT ICL

	DTV	GSM900 (BTx)	GSM1800 (BTx)	3G (BTx)
S_{BA} [nW/cm^2]	18	48	50	3

measurements to identify locations in Greater London where the designed harvesters could operate. The design procedures and prototype test measurements will be presented in the following sections.

III. SINGLE-BAND AMBIENT RF ENERGY HARVESTERS

In order to implement efficient ambient RF energy harvesters, designed for the banded input RF power density levels measured at ICL, a set of single-band prototypes were realized and characterized; these will be compared to multiband array architectures in Section V.

A. Antenna Design and Measurements

Since our harvesters are intended to operate within a general (semi-)urban environment, where the exact location of the TX source is unknown, the rectennas' antennas need to be as close to omnidirectional as possible, avoiding the need for beam-pointing during deployment. This is at the obvious expense of limited antenna gain, and therefore, the corresponding levels of P_{RF} that the rectifier can receive. Conversely, if the location of the TX is known, then it may be tempting to use a high gain antenna, but this would require an appropriate level of beam-pointing and polarization matching that can be established and maintained.

Another requirement is that the antennas need to be easily scalable across all frequency bands since one important objective for this work is to compare and contrast different banded harvesters. Finally, the antennas need to be easily fabricated. For all these reasons, a linear polarized folded dipole was selected, although a monopole would also be suitable [31].

To simplify impedance matching between the antenna and rectifier, a modified folded dipole was used to obtain the required 50-Ω reference input impedance. A balun does not need to be employed, as there is no significant degradation in performance for this particular application, even with the use of an unbalanced microstrip rectifying circuit [32]. Furthermore, the antenna was not integrated onto a substrate, to give the additional freedom to embed the harvester on windows or within walls, furnishings, fixtures, or fittings. To this end, two different antennas were fabricated for each band; one made using a 560-μm diameter copper wire and the other with 75-μm-thick 25-mm-wide copper tape. The fabricated antennas are shown in Fig. 3. Since the copper tape was not rigid enough to retain its shape, it was placed on a Perspex substrate, to represent a flat panel.

To design the antennas, full-wave 3-D electromagnetic simulations were performed using CST Microwave Studio. As discussed previously, the antennas were designed to be as omnidirectional as possible, while covering as much of the ambient RF source BW as possible. Fig. 4 shows the typical simulated gain profile for the DTV tape antenna, having a front-to-back ratio of unity. Table IV shows the simulated gain and the 10-dB return-loss fractional BW for the optimized copper wire and tape antennas.

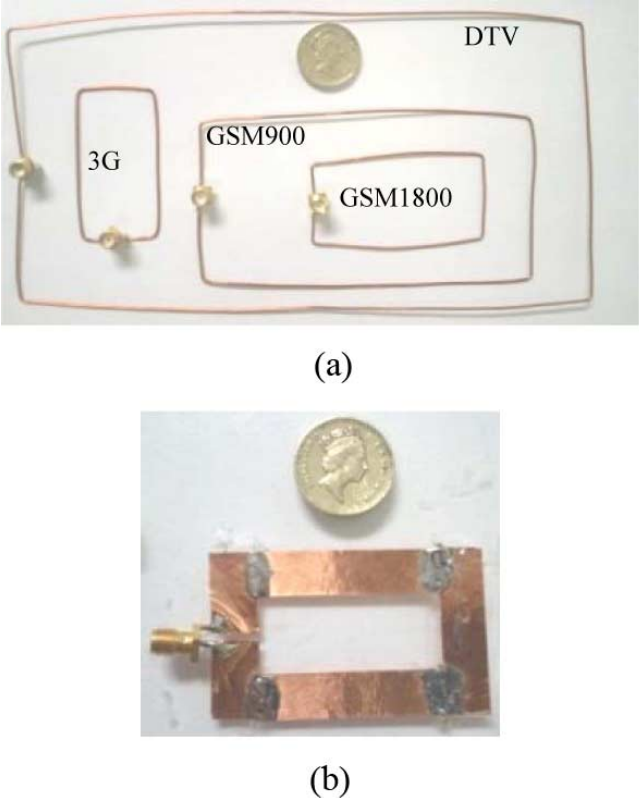

(a)

(b)

Fig. 3. 50-Ω folded-dipole antennas shown next to a British £1 coin. (a) DTV, GSM900 (BTx), GSM 1800 (BTx) and 3G (BTx) copper wire antennas. (b) 3G (BTx) copper tape antenna on Perspex.

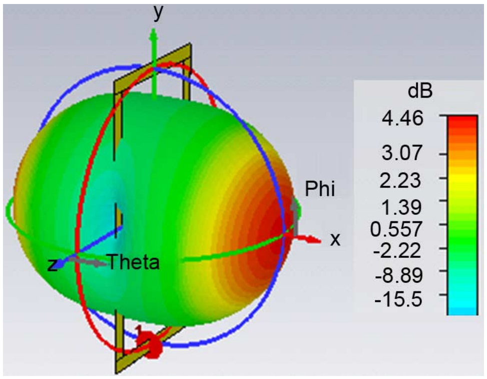

Fig. 4. Simulated beam profile for the DTV tape antenna.

Fig. 5 shows excellent agreement between predicted and measured return-loss results, within a 10-dB return-loss bandwidth, for the eight fabricated single-band antennas. As one would expect with such a simple classical antenna, the out-of-band performances (not shown) were also in good agreement. It was found that better return-loss measurements are achieved with our single-band antennas when compared to other reported single-band omnidirectional [13] and multiband [33] designs. The latter may be important, as it may be tempting to implement a more compact multiband rectenna, but which

TABLE IV
SIMULATED GAIN AND 10-dB RETURN LOSS FRACTIONAL BW FOR FOLDED-DIPOLE SINGLE-BAND ANTENNAS

Band		Wire		Tape	
	BW (%)	Gain (dBi)	BW (%)	Gain (dBi)	BW (%)
DTV	26	4.35	4	4.48	6
GSM900	3.7	4.42	4.3	4.73	4.3
GSM1800	4.1	4.32	5.3	4.73	10.7
3G	2.8	4.39	5.4	4.76	12

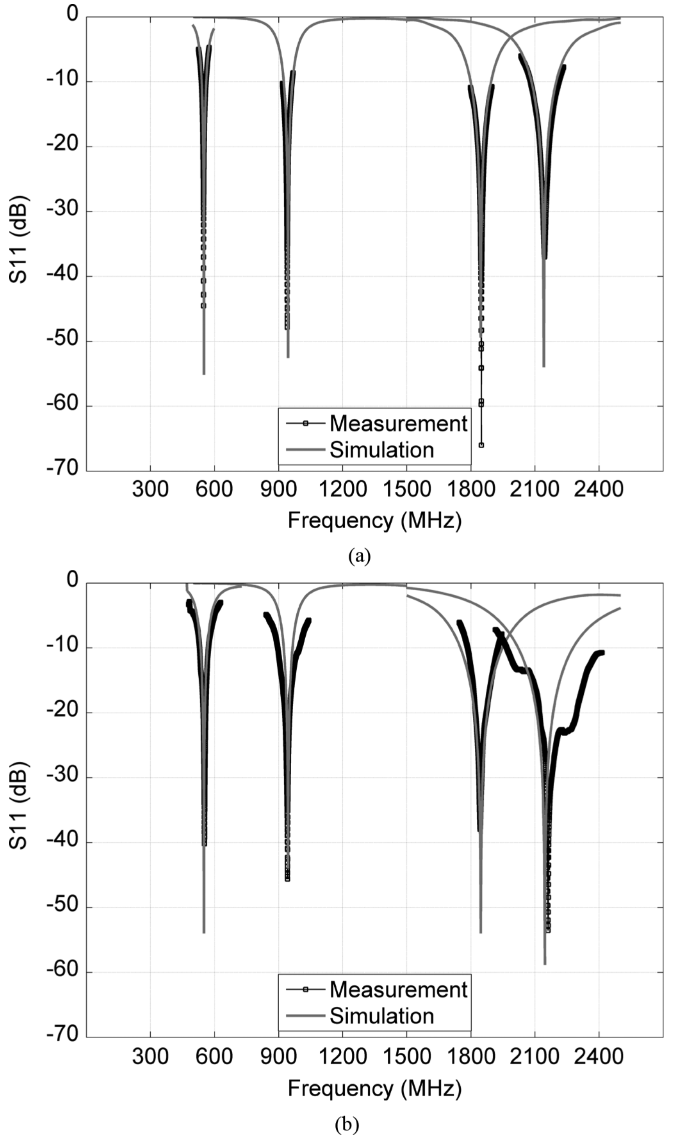

(a)

(b)

Fig. 5. Input return-loss predictions and measurements for all single-band folded-dipole antennas. (a) Wire. (b) Tape.

may ultimately not give better ambient RF energy harvesting performance.

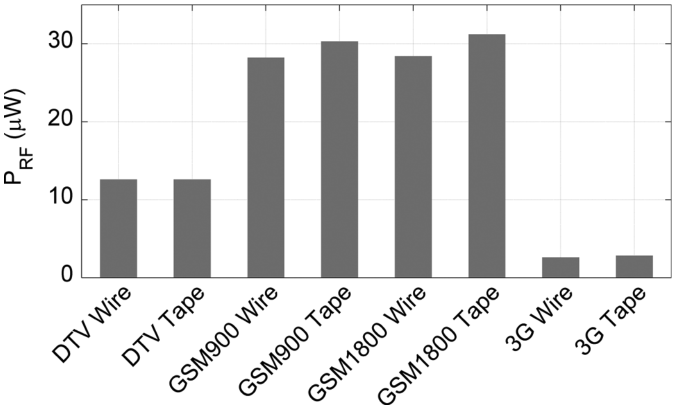

Fig. 6. Predicted input RF power levels for the four largest ambient sources at the ICL testing location.

Obtaining a minimum acceptable return loss over an antenna fractional bandwidth as large as that of the source is key to harvest as much input RF energy as possible. As can be seen in Table IV, where the fractional bandwidth is defined for a 10-dB return loss, our antennas have a fractional bandwidth greater than those of the sources, with the exception of DTV, which only covers approximately 35% of the target frequency range. In other work [13], 5-dB return-loss fractional bandwidth is adopted for RF energy harvesting applications.

The fractional bandwidth of the antennas having a minimum return loss of 5 dB is too great to assume a constant antenna gain over the whole band. Therefore, an additional advantage of using 10-dB return-loss fraction bandwidths is that (1) can be used to calculate the input RF power with the assumption that the midband antenna gain is constant with frequency [34]. Therefore, the time-averaged input RF power P_{RF} is given by

$$P_{\mathrm{RF}} = S_{BA} \cdot A_{\mathrm{real}} \text{ and } A_{\mathrm{real}} \approx G(f_o)\frac{\lambda_o^2}{4\pi} \qquad (1)$$

where A_{real} is the real aperture (or capture area) of the antenna, λ_o is the free-space wavelength at the midband frequency f_o, and $G(f_o)$ is the rectenna's antenna gain at f_o.

Substituting the measured banded input RF power densities recorded in Table III and the predicted midband antenna gains in Table IV into (1), realistic values for P_{RF} can be calculated for all four bands, with the results shown in Fig. 6. It can be seen that with all antenna gains being in the region of ~ 4.5 dBi, the 2G GSM900/1800 harvesters will generate the highest input RF power levels, due to the high-banded input RF power density levels measured *in situ*. At the other extreme, the 3G harvesters will be the worst performers. As only a small fraction of the required frequency range is covered by the DTV antennas, the predicted values for P_{RF} represent an overestimation.

B. Rectifier Design and Measurements

Based on a previously reported analysis [35] and the predicted input RF power levels presented in Fig. 6, the zero-bias SMS7630 diode (in a series configuration) was selected as the optimal solution for our ambient RF energy harvesters, as shown in Fig. 7. In a series configuration, the junction capacitance $(C_j(V))$ of the diode dominates the detector's impedance, as long as $C_{\mathrm{out}} > C_j(V)$, and thus C_{out} has little or no effect on the matching circuit. This allows C_{out} to be large enough to provide a ripple-free output voltage. In contrast, C_{out} must be less than 1 pF to achieve a good impedance match with a shunt configuration as C_{out} appears in parallel with $C_j(V)$ and the packaging parasitic capacitance. However, C_{out} is too small to

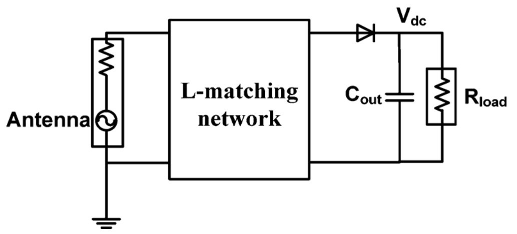

Fig. 7. Series detector configuration with L-matching network.

provide a ripple-free dc voltage to the load. This can be overcome with a matching network that will allow a good impedance match with a large output capacitor, but at the expense of introducing losses. Furthermore, as shown in [36], these issues start to become negligible with a shunt configuration as the shunt diode becomes more self-biased as the input power increases. Simulations were performed using Agilent Technologies' ADS software. Its Momentum package not only takes into account the losses from the low-cost FR4 substrate, but also calculates fringing fields, effects from which are passed on to the harmonic-balance package for simulating the nonlinear behavior of the rectifier.

A good impedance match was achieved by employing a simple matching network; a series lumped-element inductor was used to absorb part of the capacitive reactance from the series diode and an additional quarter-wavelength short-circuit shunt stub was employed to achieve the desired 50-Ω impedance [37]. Since the impedance of the diode varies with frequency and input RF power, impedance matching between the antenna and rectifier was first undertaken by finding the optimal output load resistance for an input RF power level of -20 dBm with a single-tone source at the midband frequency. After the optimal load was found, further broadband optimization was performed to the matching network and the load to ensure good impedance matching throughout the target frequency range and the measured $P_{\rm RF}$ for each band.

As with the antenna analysis, and unlike conventional RF circuits that adopt the more traditional half-power bandwidth definition, the rectifier should adopt the 10-dB-input return-loss bandwidth. The reason for this is that, for ambient RF energy harvesting applications, the input RF power is at a premium and so what little energy is available should not be wasted by being reflected back from avoidable impedance mismatches at either the antenna or rectifier.

Fig. 8 shows excellent agreement between predicted and measured input return loss results, within the -10-dB bandwidth, for the DTV and GSM900 rectifiers, having fractional bandwidths of 5.7% (below target) and 4.8%, respectively. With these lower frequency designs, the fundamental and higher order harmonics were below -55 dBm, ensuring a clean dc voltage at the load, without the need for any output filtering. Reasonable agreement was found with the GSM1800 and 3G rectifiers, having fractional bandwidths of 1.6% (below target) and 7.4%, respectively. It was found that with these two higher frequency designs, the higher order harmonics were -40 dBm at the output. This reduced performance, as illustrated in Fig. 8, is due to the higher series inductive reactance leads of the output shunt storage capacitor. For this reason, a second prototype version (v2) was designed for the 3G rectifier, using distributed-element components for the input impedance matching stage and an additional output filter stage. With the lumped-element matching network, it was found that in order to achieve good input impedance matching to 50 Ω, all the microstrip transmission lines had to have a characteristic impedance of 92 Ω. With the distributed-element matching network, a simple shunt quarter-wavelength open-circuit stub, designed for operation at the fundamental frequency, was employed. A microstrip line was added between the cathode of the diode and the stub to absorb the capacitive reactance of the diode. The stub effectively filters to < -50 dBm the higher order harmonics. The microstrip design can be seen in Fig. 13.

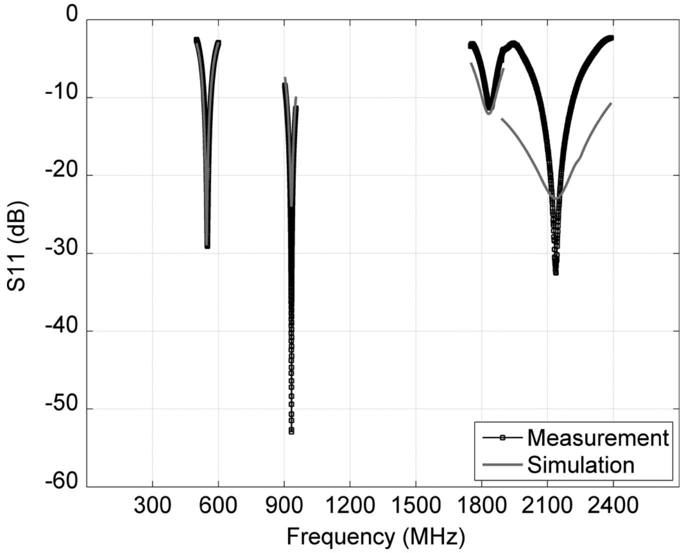

Fig. 8. Input return-loss predictions and measurements for all first prototype single-band rectifiers, with $P_{\rm RF} = -20$ dBm at the input and optimal load resistances at the output.

C. PMM

Since the input RF power from ambient sources can be represented as a multi-tone source, with power levels fluctuating across the target frequency range, the output impedance of the rectifier is time varying. A power management module (PMM) capable of performing maximum power point tracking (MPPT) is required.

For our work, a low-power integrated-circuit PMM from Texas Instruments Incorporated (BQ25504) was selected, due to its low quiescent current ($<$ 330 nA) and low input voltage operation ($\sim$ 80 mV hot-start and 330-mV cold-start) [38]. It is worth noting that its startup voltage is lower than PMMs previously reported and realized using hybrid circuits for RF energy harvesting [13]. The BQ25504 PMM includes a boost converter that steps up its input voltage (having a 350-mV average value during ambient operation) to useful levels between 2.4–5.3 V. The BQ25504 also has a built-in battery management module, which is used to control the duty cycle of the output power to the load.

MPPT operation on the BQ25504 is achieved by periodically sampling the open-circuit voltage (OCV) at the input of the converter, which then draws a current causing the converter

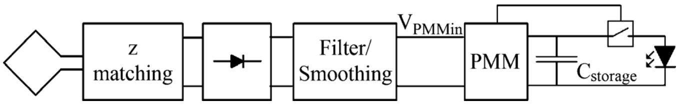

Fig. 9. System block diagram.

input voltage to fall and be held at a pre-programmed fraction of the OCV (set by a potential divider). In a simple dc circuit, with a resistive source impedance, the optimal ratio is 0.5. For the rectenna-based system, a ratio of 0.48–0.53 was found to maximize the power output of the system. The BQ25504 is designed to charge a storage element, and in this case, a capacitor C_{storage} was used. The programmed PMM continuously charges the storage capacitor, and the load (a low-power LED) was automatically connected to the storage capacitor when the capacitor voltage reaches an upper limit $V_{\text{high}} = 2.84$ V and automatically disconnected when it reaches a lower limit $V_{\text{low}} = 2.40$ V. The duty cycle of the LED can then be used to calculate the efficiency of the system, as will now be described. A diagram of the system is shown in Fig. 9.

IV. END-TO-END EFFICIENCY ANALYSIS

The efficiency η of an RF energy harvesting system is

$$\eta = \frac{P_{\text{dc}}}{P_{\text{RF}}} \tag{2}$$

where P_{dc} is the time-averaged output (i.e., equivalent dc) power into the storage element (e.g., battery or supercapacitor) and load and P_{RF} is as previously defined. Measurements for this type of system are usually performed in a controlled environment (e.g., an anechoic chamber or TEM cell), using a dedicated constant or variable amplitude single-tone RF signal source [32], [39]. However, the former is not suitable for evaluating ambient RF energy harvesting operation, which has a much broader spectrum of nonconstant input frequencies and where the instantaneous input RF power is time variant. The use of a constant single-tone dedicated source provides a convenient stable reference power to the harvester; while the latter reflects a more realistic signal source having fluctuating power levels across a nonzero bandwidth, multipath, and reflection effects which are very difficult to emulate in a controlled environment.

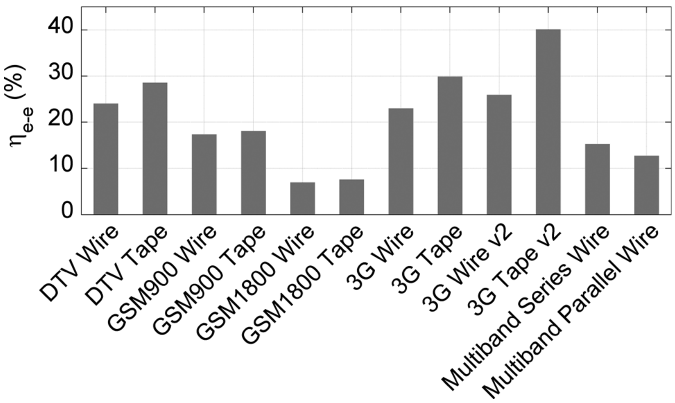

Fig. 10. End-to-end efficiencies for ambient RF energy harvesting at ICL.

Therefore, to determine the overall end-to-end efficiency $\eta_{\text{e-e}}$ for a complete ambient RF energy harvester, the input RF energy U_{RF} was calculated based on the harvester's antenna characteristics and the actual banded input RF power density measurements taken at the time of harvester operation, using the Agilent Fieldfox and the calibrated antenna. It is important to note that since the impedance mismatch between the antenna and detector is not taken into account, U_{RF} is higher than expected, providing an underestimate of end-to-end efficiency. The output dc energy U_{dc} was then calculated by measuring the charge–discharge cycle time, t_{cycle} of the storage capacitor between V_{high} and V_{low}, as the LED is repeatedly connected and disconnected. The output dc energy equation is already taking into account the efficiency of the PMM given the fact that the measurements are taken at its output voltage. The end-to-end efficiency of one charge–discharge cycle of C_{storage} is

$$\eta_{\text{e-e}} = \frac{U_{\text{dc}}}{U_{\text{RF}}} \tag{3}$$

where the input RF energy is given by integrating the time-averaged input RF power over a cycle time, as

$$U_{\text{RF}} = \int_{0}^{t_{\text{cycle}}} P_{\text{RF}} dt \tag{4}$$

and the output dc energy is given by the energy supplied to the load, as follows:

$$U_{\text{dc}} = C_{\text{storage}} \frac{({V_{\text{high}}}^2 - V_{\text{low}}^2)}{2}. \tag{5}$$

A. ICL Field Trials

Four single-band ambient RF energy harvesters were assembled by connecting the rectifiers to the wire/tape antennas and PMMs programmed for the optimal load. A 100-μF shunt capacitor was employed as the storage element, providing $U_{\text{dc}} = 115\ \mu$J. Our system is capable of cold-starting the boost converter and MPPT since the rectenna is capable of providing the minimum starting voltage of 330 mV. When the minimum voltage is reached, the boost converter and MPPT start to operate and the charge–discharge cycle at the load begins, causing the LED to flash. During field trials, t_{cycle} took up to 170 s for the harvester with lowest banded input RF power density, corresponding to 3G with the wire antenna. Table V summarizes the results where t_c and t_d are the charge and discharge times, respectively, and ΣV and ΣI are the multiband voltage and current summing array architectures, respectively. A detailed discussion on the multiband rectenna arrays will be presented in the following section. The end-to-end efficiency was calculated using (3) with data from Fig. 6 and measuring the charge–discharge cycle time during harvesting operation.

Fig. 10 shows the overall end-to-end efficiencies for all the harvester demonstrators, deployed and tested at ICL. As predicted by simulations, the improved 3G v2 demonstrator with

TABLE V
HARVESTERS CHARGE AND DISCHARGE TIMES (t_c, t_d, RESPECTIVELY) FOR A SPECIFIED LOAD

Band	Wire					Tape				
	t_c (s) load independent	t_d (s) load dependant	t_{cycle} (s) load dependant	P_{dc} (t_d) (μW)	P_{dc} (t_{cycle}) (μW)	t_c (s) load independent	t_d (s) load dependant	t_{cycle} (s) load dependant	P_{dc}(t_d) (μW)	P_{dc} (t_{cycle}) (μW)
DTV	26	12	38	9.6	3	14	18	32	8.2	3.6
GSM900	14	10	24	11.5	4.8	8	13	21	14.4	5.5
GSM1800	43	15	58	7.7	2	22	27	49	5.2	2.4
3G v2	167	3	170	38.4	0.7	96	5	101	1.2	1.1
Multiband ΣV	43	7	50	66	2.3	-	-	-	-	-
Multiband ΣI	55	5	60	92.2	2	-	-	-	-	-

tape antenna outperformed its original design by 11%; achieving an end-to-end efficiency of 40% with an input RF power of only −25.4 dBm.

It is believed that a much greater efficiency can be achieved for the DTV harvester if the fractional bandwidths for the first prototype circuits (i.e., 4.4/4.5% for the antennas and 5.8% for the rectifier) could be increased to match the much greater target value of 26%. Likewise, the reduced efficiency of the GSM1800 harvesters can be attributed to the detrimental effects of the narrowband input impedance matching of the rectifier (i.e., having a fractional bandwidth of only 1.6%, when compared to its target value of 4.1%). Finally, with all the harvesters, the end-to-end efficiencies can be enhanced through better antenna design and optimal polarization matching.

Table VI, shows the number of locations from the London RF survey that would be able to support our harvesters. Unlike the single-band 3G harvester, which can operate at 45% of the locations, our DTV harvester can only be used at two locations (one in zone 2 and the other in zone 3). Therefore, for the general deployment of an ambient RF energy harvester within an (semi-)urban environment, at street level, the single-band DTV harvester may not be practical

TABLE VI
NUMBER OF LOCATIONS FROM THE LONDON RF SURVEY CAPABLE OF SUPPORTING IDENTICAL HARVESTERS AT THE SAME EFFICIENCY LEVELS

	DTV	GSM900 (BTx)	GSM1800 (BTx)	3G (BTx)
Stations with higher S_{BA}	2	28	68	122

V. ARRAY ARCHITECTURES

Since ambient input RF power levels can be low (i.e., below −25 dBm) and dependent on both time and spatial considerations, harvesters could be designed to extract energy with spatial-diversity within the same frequency band or using different frequency bands. For example, with the former, at a particular location there may be only one band that has significant levels of RF energy worth harvesting. In this case, spatial-diversity array architectures may provide more usable output power. Alternatively, with the latter, multiband array architectures may provide more robust operation.

With both forms of parallel array architecture (i.e., spatial-diversity and multiband), a further classification can be seen through the use of either diversity/band switching or a summing node. With the former, physical switches automatically select whichever signal path delivers the highest input RF power level; with the latter, power from all signals is combined. Fig. 11 illustrates generic forms of parallel array architectures, showing that switching/summing can be performed electromagnetically at a single antenna or at the output from multiple antennas, rectifiers, or PMMs.

Multiband array architectures, similar to those in Fig. 11(c) and (d), capable of RF harvesting from the four previously identified bands, were selected as possible optimal solutions, given no size/cost constraints. Our objectives were to reach the minimum cold-start voltage at the lowest possible input RF power levels and increase the harvesters' operational capabilities within (semi-)urban environments.

To this end, two different multiple rectenna architectures were investigated. The first with a single shared PMM and the second with multiple PMMs, as illustrated in Fig. 12. To simplify assembly, the wire antennas were selected since they did not require a substrate. Unwanted coupling between the single-band antennas was minimized by placing them a minimum distance of $\lambda_L/5$ apart; where λ_L is the wavelength of the lowest frequency band antenna [40]. For example, the DTV and the 3G antennas were kept at least 11 cm apart, as shown in Fig. 13. This allowed S_{11} measurements to be the same as in Fig. 5 once all antennas were assembled into the array.

A. Multiple Rectennas With a Shared PMM

In order to improve the cold-start performance of the system, the outputs of multiple rectennas can be connected in series, as shown in Fig. 12(a). This increases the probability of the voltage on the input of the PMM reaching the cold-start level (330 mV for the BQ25504) under any given scenario. While cold-starting the PMM, each rectenna harvests (albeit not optimally). Once

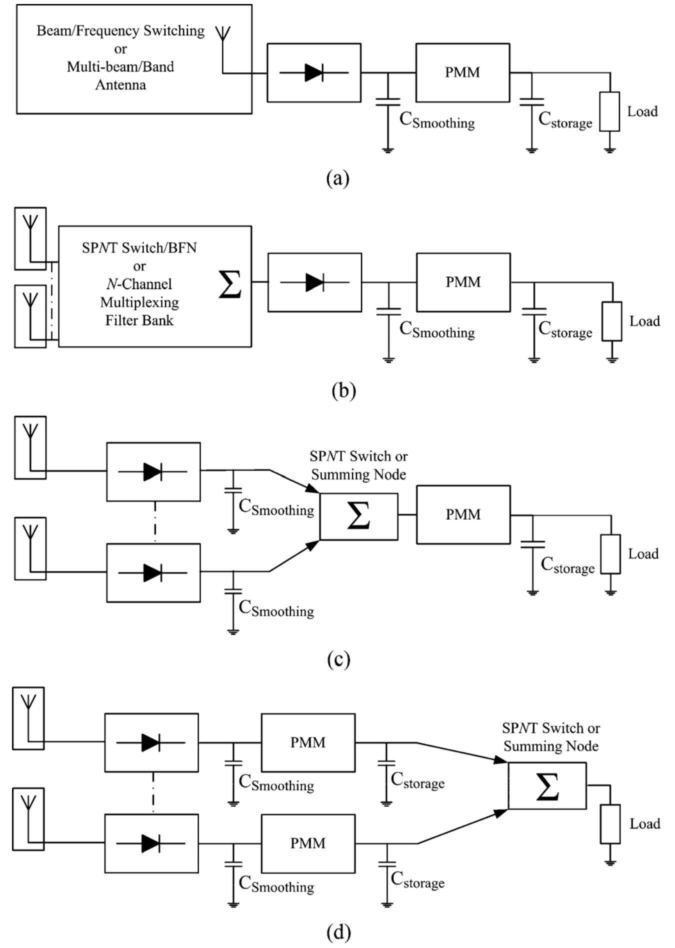
Beam/Frequency Switching
or
Multi-beam/Band
Antenna
PMM
C_Smoothing
C_storage
Load
(a)
SPNT Switch/BFN
or
N-Channel
Multiplexing
Filter Bank
Σ
PMM
C_Smoothing
C_storage
Load
(b)
SPNT Switch or
Summing Node
Σ
PMM
C_Smoothing
C_Smoothing
C_storage
Load
(c)
PMM
PMM
C_Smoothing
C_Smoothing
C_storage
C_storage
SPNT Switch or
Summing Node
Σ
Load
(d)

Fig. 11. Parallel array architectures with switching/summing at the: (a) antenna, (b) output of multiple antennas, (c) output of multiple rectifiers, and (d) output of multiple PMMs.

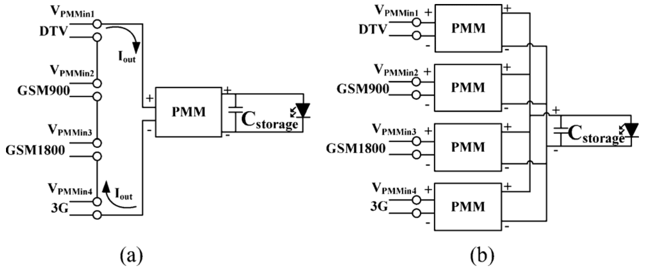
V_PMMin1
DTV
I_out
V_PMMin2
GSM900
V_PMMin3
GSM1800
V_PMMin4
3G
I_out
PMM
C_storage
(a)
V_PMMin1
DTV
PMM
V_PMMin2
GSM900
PMM
V_PMMin3
GSM1800
PMM
V_PMMin4
3G
PMM
C_storage
(b)

Fig. 12. Multiband array architectures (with $N = 4$ bands). (a) Voltage summing at the outputs of the single-band rectennas. (b) Current summing at the outputs of the single-band harvesters.

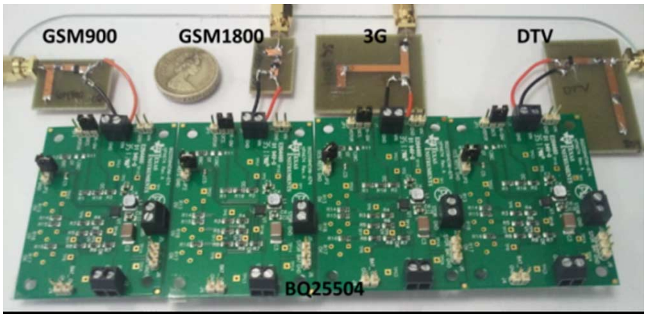
GSM900
GSM1800
3G
DTV
BQ25504

Fig. 13. Rectenna array architecture with individual PMMs for the four largest contributors with wire antennas.

the PMM circuit starts, with the MPPT operating, the harvested power level increases.

The behavior of the series rectenna topology with a shared PMM requires some discussion. As the output impedance and the OCV for each rectenna is different, since they operate at different frequencies and input RF power levels, the rectennas are forced to share the same output current in a series configuration, which does not allow them all to operate at their individual maximum power points. This causes the voltage on each rectenna output, except the one having the highest input RF power, to collapse. This operation is analogous to the partial shading problem with a series string of solar panels [41] sharing a common boost converter. With this photovoltaic system, bypass diodes placed around individual cells stop the poorly lit cells contributing a negative voltage (and power) to the string. In our case, the series circuit formed by the loop antennas and rectifying diodes performs the same task. This means that while all rectennas contribute to system startup, only the rectenna with the highest input RF power contributes significant power for continuous operation once the PMM starts. Fig. 10 shows the end-to-end efficiency for the voltage summing multiband harvester array when tested at ICL. An efficiency of only 15% was achieved with a combined input RF power of -12 dBm. The lower efficiency, when compared to a single-band harvester, is due to the imbalance of rectifier outputs, as discussed above. Here, the charge time was 43 s, compared to 167 s with the lowest contributing single-band 3G harvester with wire antenna.

B. Multiple Rectennas With Individual PMMs

In order to overcome the balancing issues when multiple rectennas share a common PMM, as discussed previously, each rectenna can have its own PMM, whose outputs can be connected to a common storage element (in this case, a 400-μF shunt capacitor, providing $U_{\rm dc} = 461\ \mu$J), as illustrated in Fig. 12(b) and shown in Fig. 13. Although not achieving cold-start as quickly as the series topology, this parallel topology has the advantage of being able to run each rectenna at its maximum power point. In addition, once one rectenna is able to harvest enough energy for a cold-start, all PMMs will start because they share a common storage element, allowing the rectennas with low-input RF power levels to harvest at levels below which they could not do so independently.

This parallel topology was tested and found to be capable of operating in many locations where the series array was unable to operate; e.g., if only one of the bands had $P_{\rm in} > -25$ dBm. As expected, the largest contributor hot-started the other PMMs, allowing them to harvest at an input RF power level down to -29 dBm.

However, as with the previous results for voltage summing, having a combined input RF power of -12 dBm, the efficiency using multiple PMMs is slightly lower, at 13%, as shown in Fig. 10. This is because useful dc output power from the cold-

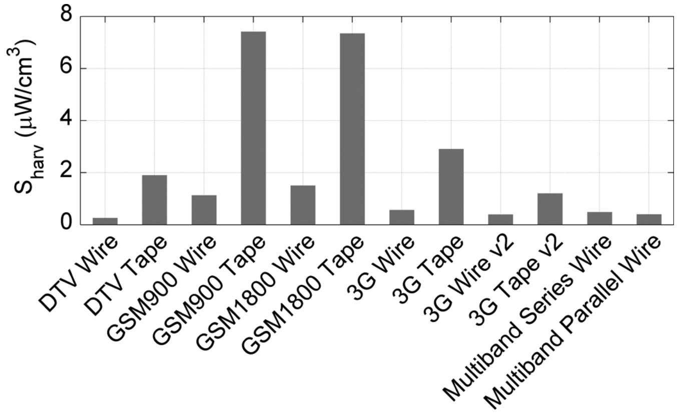

Fig. 14. Output dc power density for all harvesters at ICL.

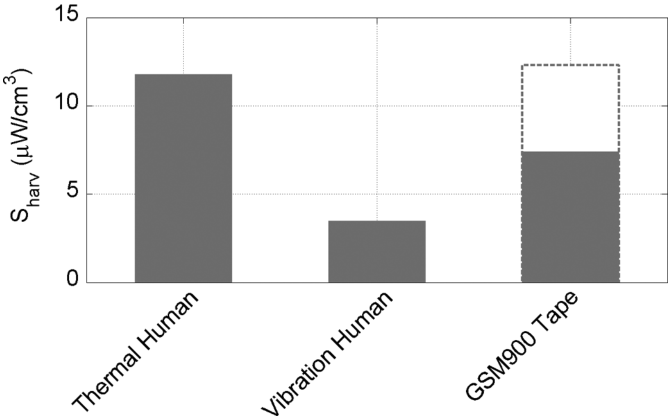

Fig. 15. Output dc power density comparison for alternative ambient harvesting technologies [40]–[43] against the best current generation of RF harvesters at ICL.

starting harvester is being supplied to the other harvesters for hot-starting, even though some of them may not actually be contributing any of their own harvested power.

VI. OUTPUT DC POWER DENSITY COMPARISON

The volumetric output dc power density $S_{\rm harv}$ $(\mu{\rm W/cm}^3)$ represents an important figure of merit for comparing alternative energy harvesting technologies. For ambient RF energy harvesting, the output dc power is calculated by multiplying the effective input RF power by the overall end-to-end-efficiency. The total volume (including that of the antenna, rectifier, and PMM; not including energy storage, as this does not directly affect the dc power output) must be determined. It is important to note that the volume for the antenna could effectively disappear if it is assembled onto a window or within a wall, furnishing, fixture, or fitting. Moreover, the required PMM printed circuit board (PCB) size used throughout these calculations was assumed to be ten times the size of the BQ25504 chip, to account for any necessary additional components.

Fig. 14 shows the output dc power density for all the harvesters demonstrated here. It can be seen that the 2G GSM900/1800 harvesters with tape antennas both have the highest value of $S_{\rm harv} = 7.4$ $(\mu{\rm W/cm}^3)$, when tested at ICL, due to the high-banded input RF power density S_{BA}. The value for the most efficient harvester (i.e., 3G v2 with tape antenna) was not the highest in terms of output RF power because S_{BA} in this band was more than an order of magnitude lower than with GSM900.

S_{harv} allows a direct and meaningful comparison to be made with other alternative energy harvesting technologies. Our best performing ambient RF energy harvester (i.e., GSM900 with tape antenna) was compared against alternative energy harvesting technologies, assuming they used the same PMM board size [42]–[44].

It can be seen in Fig. 15 that ambient RF energy harvesting has a low output dc power density when compared to alternative energy harvesting technologies, but only when the total volume of the first prototype demonstrator is considered. However, when the antenna is absorbed onto or into a background feature and when the PMM is fully integrated into the rectifier, it can outperform (as indicated by the dotted column) the alternative energy harvesting technologies, while providing a complimentary means of extracting energy from the environment. The RF harvesters, however, have the additional advantage in that they do not require a thermal gradient, and unlike vibration-driven devices, they have no moving parts.

VII. CONCLUSIONS

Our objectives were to reach the lowest possible ambient input RF power levels and extend the harvesters' operational capabilities within (semi-)urban environments. To this end, a comprehensive citywide RF spectral survey was undertaken, indicating that more than 50% of the 270 London Underground stations are suitable locations for the deployment of our ambient RF energy harvesters. It has been demonstrated that single-band harvesters can operate with efficiencies of up to 40% in a (semi-)urban environment, and can start to operate from power levels as low as -25 dBm.

To increase the freedom of operation, multiband array architectures were investigated. With the current summing harvester arrays, RF harvesting was achieved at an input RF power level as low as -29 dBm, without any external dc power supply to hot-start the PMM. Limitations on the multiband array architectures were discussed, highlighting the need for further work in balancing rectennas with voltage summing rectenna arrays when operating at lower input RF power levels.

Finally an output dc power density comparison against alternative energy harvesting technologies has shown that RF harvesting can represent a competitive solution within (semi-)urban environments, especially when the antenna can be absorbed into background features.

ACKNOWLEDGMENT

The authors would like to thank the members of the London RF survey team and F. Kusidlo, Agilent EEsof EDA, for his advice regarding simulations.

REFERENCES

[1] “Demonstration of microwave power transmission,” *Pract. Wireless*, pp. 835–835, 1965.

[2] T. Salter, G. Metze, and N. Goldsman, “Parasitic aware optimization of an RF power scavenging circuit with applications to smartdust sensor networks,” in *IEEE Radio Wireless Symp.*, 2009, pp. 332–335.

[3] “SENSIMED Triggerfish®—Continuous IOP monitoring,” SENSIMED, Lausanne, Switzerland, Mar. 2012. [Online]. Available: http://www.sensimed.com/S-Trig-glaucoma.htm, , “

[4] M. Pinuela, D. C. Yates, S. Lucyszyn, and P. D. Mitcheson, “Maximizing DC-to-load efficiency for inductive power transfer,” *IEEE Trans. Power Electron.*, vol. 28, no. 5, pp. 2437–2447, May 2013.

[5] W. Joseph, G. Vermeeren, L. Verloock, M. M. Heredia, and L. Martens, “Characterization of personal RF electromagnetic field exposure and actual absorption for the general public,” *Health Phys.*, vol. 95, no. 3, pp. 317–330, Sep. 2008.

[6] G. Thuroczy, F. Molnar, J. Szabo, G. Janossy, N. Nagy, G. Kubinyi, and J. Bakos, “Public exposure to RF from installed sources: Site measurements and personal exposimetry,” in *1st Eur. Antennas Propag. Conf.*, 2006, pp. 1–4.

[7] P. Elliott, M. B. Toledano, J. Bennett, L. Beale, K. de Hoogh, N. Best, and D. J. Briggs, “Mobile phone base stations and early childhood cancers: Case-control study,” *Br. Med. J.*, vol. 340, no. 1, pp. 1–7, Jun. 2010.

[8] N. Shinohara, “Power without wires,” *IEEE Microw. Mag.*, vol. 12, no. 7, pp. S64–S73, Dec. 2011.

[9] J. Y. Park, S. M. Han, and T. Itoh, “A rectenna design with harmonic-rejecting circular-sector antenna,” *IEEE Antennas Wireless Propag. Lett.* , vol. 3, pp. 52–54, 2004.

[10] “Dengyo rectennas,” Nihon Dengyo Kosaku Company Ltd., Tokyo, Japan, Jul. 2012. [Online]. Available: http://www.den-gyo.com/solution/solution01.html

[11] T. Le, K. Mayaram, and T. Fiez, “Efficient far-field radio frequency energy harvesting for passively powered sensor networks,” *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1287–1302, May 2008.

[12] V. Marian, B. Allard, C. Vollaire, and J. Verdier, “Strategy for microwave energy harvesting from ambient field or a feeding source,” *IEEE Trans. Power Electron.*, vol. 27, no. 11, pp. 4481–4491, Nov. 2012.

[13] C. Mikeka, H. Arai, A. Georgiadis, and A. Collado, “Dtv band micropower RF energy-harvesting circuit architecture and performance analysis,” in *IEEE Int. RFID-Technol. Appl. Conf.*, 2011, pp. 561–567.

[14] S. Kitazawa, H. Ban, and K. Kobayashi, “Energy harvesting from ambient RF sources,” in *IEEE MTT-S Int. Microw. Symp. Workshop*, 2012, pp. 39–42, Series on Innovative Wireless Power Transmission: Technol., Syst., Appl..

[15] R. Vyas, V. Lakafosis, and M. Tentzeris, “Wireless remote localization system utilizing ambient RF/solar power scavenging RFID tags,” in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 23–28, 2010, pp. 1764–1767.

[16] H. Nishimoto, Y. Kawahara, and T. Asami, “Prototype implementation of ambient RF energy harvesting wireless sensor networks,” *IEEE Sens. J.*, pp. 1282–1287, Nov. 2010.

[17] A. Sample and J. R. Smith, “Experimental results with two wireless power transfer systems,” in *IEEE Radio Wireless Symp.*, 2009, pp. 16–18.

[18] V. Nitu, G. Lojewski, and S. Nitu, “Electromagnetic field evaluation on an antennas shared site, EUROCON 2009,” in *IEEE EUROCON’09*, 2009, pp. 70–75.

[19] Y. Kawahara, K. Tsukada, and T. Asami, “Feasibility and potential application of power scavenging from environmental RF signals,” in *IEEE AP-S Int. Symp.*, 2009, pp. 1–4.

[20] H. J. Visser, A. C. F. Reniers, and J. A. C. Theeuwes, “Ambient RF energy scavenging: GSM and WLAN power density measurements,” in *38th Eur. Microw. Conf.*, 2008, pp. 721–724.

[21] T. G. Cooper, S. M. Mann, M. Khalid, and R. P. Blackwell, “Public exposure to radio waves near GSM microcell and picocell base stations,” *J. Radiol. Protection*, vol. 26, no. 2, pp. 199–211, 2006.

[22] “FieldFox RF Analyser N9912A 4/6 GHz,” Agilent Technol., Santa Clara, CA, USA, Tech. Overview, 2012.

[23] “Biconical EMC broadband antennas—BicoLOG Series,” Aaronia, Strickscheid, Germany, Data Sheet, 2012.

[24] “Digital TV switchover in the UK,” Digital UK Ltd., London, U.K., 2012.

[25] “International commission on non-ionizing radiation protection, guidelines for limiting exposure to time- varying electric, magnetic, and electromagnetic fields (up to 300 GHz),” *Health Phys.*, vol. 74, no. 4, pp. 494–522, 1998.

[26] M. Riederer, “EMF exposure due to GSM base stations: measurements and limits,” in *IEEE Int. Electromagn. Compat. Symp.*, 2003, vol. 1, pp. 402–405.

[27] “Revised ECC recommendation (02)04: Measuring non-ionising electromagnetic radiation (9 kHz—300 GHz),” ECC within the Eur. CEPT, Oct. 2003.

[28] Nat. Freq. Planning Group, “United Kingdom frequency allocation table, issue no. 16,” 2010, Cabinet Official Committee on U.K. Spectrum Strategy.

[29] Transport for London, “London underground map,” Aug. 2012. [Online]. Available: http://www.tfl.gov.uk/gettingaround/1106.aspx

[30] Imperial College London, “London RF Survey,” Aug. 2012. [Online]. Available: http://londonrfsurvey.org/

[31] R. Vyas, V. Lakafosis, M. Tentzeris, H. Nishimoto, and Y. Kawahara, “A battery-less, wireless mote for scavenging wireless power at UHF (470–570 MHz) frequencies,” in *IEEE AP-S Int. Symp./USNC/URSI Nat. Radio Sci. Meeting*, Jul. 3–8, 2011, pp. 1069–72.

[32] T. Ungan and L. M. Reindl, “Harvesting low ambient RF-sources for autonomous measurement systems,” in *Proc. IEEE Instrum. Meas. Technol. Conf.*, 2008, pp. 62–65.

[33] A. Costanzo, A. Romani, D. Masotti, N. Arbizzani, and V. Rizzoli, “RF/baseband co-design of switching receivers for multiband microwave energy harvesting,” *Sensors Actuators A, Phys.*, vol. 179, pp. 158–168, Jun. 2012.

[34] J. L. Volakis, *Antenna Engineering Handbook*. New York, NY, USA: McGraw-Hill, 2007.

[35] M. Pinuela, P. D. Mitcheson, and S. Lucyszyn, “Analysis of scalable rectenna configurations for harvesting high frequency ambient radiation,” in *Proc. Power MEMS*, Leuven, Belgium, 2010, pp. 41–44.

[36] J. O. McSpadden, L. Fan, and K. Chang, “Design and experiments of a high-conversion-efficiency 5.8-GHz rectenna,” *EEE Trans. Microw. Theory Techn.*, vol. 46, no. 12, pp. 2053–2060, Dec. 1998.

[37] “HSMS-285x series datasheet,” Avago Technol., San Jose, CA, USA, 2009.

[38] “Ultra low power boost converter with battery management for energy harvester applications,” Texas Instruments Incorporated, Dallas, TX, USA, BQ25540 Datasheet, 2011.

[39] A. Nimo, D. Grgic, and L. M. Reindl, “Ambient electromagnetic wireless energy harvesting using multiband planar antenna,” in *IEEE 9th Int. Syst., Signals, Devices Multi-Conf.*, Mar. 20–23, 2012, pp. 1, 6.

[40] H. Takhedmit, L. Cirio, B. Merabet, B. Allard, F. Costa, C. Vollaire, and O. Picon, “A 2.45-GHz dual-diode rectenna and rectenna arrays for wireless remote supply applications,” *Int. J. Microw. Wireless Technol.*, vol. 3, no. 3, pp. 251–258, Jun. 2011.

[41] S. Vemuru, P. Singh, and M. Niamat, “Modeling impact of bypass diodes on photovoltaic cell performance under partial shading,” in *IEEE Int. Electro/Information Technol. Conf.*, 2012, pp. 1–5.

[42] P. D. Mitcheson, E. M. Yeatman, G. K. Rao, A. S. Holmes, and T. C. Green, “Energy harvesting from human and machine motion for wireless electronic devices,” *Proc. IEEE*, vol. 96, no. 9, pp. 1457–1486, Sep. 2008.

[43] R. J. M. Vullers, R. V. Schaijk, H. J. Visser, J. Penders, and C. V. Hoof, “Energy harvesting for autonomous wireless sensor networks,” *IEEE Solid-State Circuits Mag.*, vol. 2, no. 2, pp. 29–38, Spring, 2010.

[44] V. Leonov, P. Fiorini, S. Sedky, T. Torfs, and C. Van Hoof, “Thermoelectric MEMS generators as a power supply for a body area network,” in *13th Int. Solid-State Sens., Actuators, Microsyst. Conf. Tech. Dig.*, vol. 1, pp. 291–294.

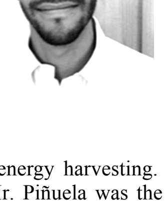

Manuel Piñuela (M’12) received the BSc. degree in electrical and electronic engineering from the National Autonomous University of Mexico (UNAM), Mexico City, Mexico, in 2007, and is currently working toward the Ph.D. degree at Imperial College London, London, U.K.

From 2006 to 2009, he worked in industry as an Electronic Design and Project Engineer for companies focused on oil and gas services in both Mexico and the US. His research interests are wireless power transfer, RF power amplifiers, and RF energy harvesting.

Mr. Piñuela was the recipient of the Gabino Barrera Medal.

Paul D. Mitcheson (SM'12) received the M.Eng. degree in electrical and electronic engineering and Ph.D. degree from Imperial College London, U.K., in 2001 and 2005, respectively.

He is currently a Senior Lecturer with the Control and Power Research Group, Electrical and Electronic Engineering Department, Imperial College London. His research interests are energy harvesting, power electronics, and wireless power transfer. He is involved with providing power to applications in circumstances where batteries and cables are not suitable. His work has been sponsored by the European Commission, EPSRC and several companies.

Dr. Mitcheson is a fellow of the Higher Education Academy.

Stepan Lucyszyn (M'91–SM'04) received the Ph.D. degree in electronic engineering from King's College London (University of London), London, U.K., in 1992, and the D.Sc. (higher doctorate) degree from Imperial College London, London, U.K., in 2010.

He is currently a Reader (Associate Professor) of millimeter-wave electronics and Director of the Centre for Terahertz Science and Engineering, Imperial College London. After working in industry as a Satellite Systems Engineer for maritime and military communications, he spent the first 12 years researching microwave and millimeter-wave RF integrated circuits (RFICs)/monolithic microwave integrated circuits (MMICs), followed by RF microelectromechanical systems (MEMS) technologies. He has coauthored approximately 140 papers and 11 book chapters in applied physics and electronic engineering. He has delivered many invited presentations at international conferences.

Dr. Lucyszyn was an associate editor for the JOURNAL OF MICROELECTROMECHANICAL SYSTEMS (2005–2009). In 2011, he was the chairman of the 41st European Microwave Conference, Manchester, U.K. In 2005, he was elected Fellow of the Institution of Electrical Engineers (IEE), U.K., and Fellow of the Institute of Physics, U.K. In 2008, was invited as a Fellow of the Electromagnetics Academy, USA. In 2009, he became an IEEE Distinguished Microwave Lecturer (2010–2012). He is currently an Emeritus DML for 2013 and a newly appointed European Microwave Lecturer (EML) for the European Microwave Association.

ELECTRICAL & ELECTRONIC ENGINEERING | RESEARCH ARTICLE

A dual-band RF energy harvesting circuit using 4th order dual-band matching network

Sachin Agrawal[1]*, Manoj S. Parihar[1] and P.N. Kondekar[1]

Received: 06 March 2017
Accepted: 12 May 2017
First Published: 20 May 2017

*Corresponding author: Sachin Agrawal, PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482005, India
E-mail: bitssachin.agrawal@gmail.com

Reviewing editor:
Wei Meng, Wuhan University of Technology, China

Additional information is available at the end of the article

Abstract: A novel compact rectifier for dual-band operation in the RF energy harvesting is presented. The circuit comprises a 4th order dual-band impedance matching and a single-series circuit with one double diode, both are integrating into a compact shape to occupy a small area of $30 \times 35\ \text{mm}^2$. The merit of the proposed rectifier circuit is that it can be extended to n number of the frequency band by using only $2 \times n$ matching elements. To validate the design method experimentally, a prototype of a dual-band rectifier is fabricated for two public telecommunication bands of GSM-900 and 1800. In order to reduce the circuit complexity and sensitivity arising due to lumped elements, the meander line and the open stub are used to realize the proposed circuit. A good agreement is obtained between the simulation and the measurement. The measured results show that the proposed rectifier circuit exhibits the conversion efficiency of 25.7 and 65% for an input power of −20 and 0 dBm, respectively. In addition, diode nonlinearity which affects the performance of the rectifier in terms of impedance matching is also investigated.

Subjects: Electromagnetics & Microwaves; Electronics; Circuits & Devices

Keywords: RF energy harvesting; dual band impedance matching; rectifier; RF-to-dc-conversion efficiency; frequency transformation

Sachin Agrawal

ABOUT THE AUTHOR

Sachin Agrawal did his BE and ME from Jiwaji University and Birla Institute of Technology Pilani, India in 2005 and 2009, respectively. Currently, he is pursuing his PhD from PDPM-Indian Institute of Information Technology Design and Manufacturing Jabalpur, India. His area of interest are; RF energy harvesting circuit, Antenna including planar and dielectric resonator antenna and Power amplifier.

PUBLIC INTEREST STATEMENT

With rapid growth in wireless communication, a huge amount of radio frequency (RF) energy broadcasted through billions of microwave sources such as mobile phones, handheld radios, and radio broadcast stations. Therefore, it is meaningful to collect and supply it to many electrical devices like mobile headsets, wearable medical sensors through RF energy harvesting. Since the ambient RF energy is distributed in multiple frequency bands, therefore the amount of energy harvested could increase if the circuit is designed for multiple frequency bands. In this work, we present a compact dual-band energy harvesting circuit to harvest energy from two most useful frequency bands, GSM-900 and 1800. The merit of the proposed rectifier circuit is it can be extended to n number of the frequency band by using only $2 \times n$ matching elements. A prototype is fabricated, and its performance is evaluated using Vector Network Analyzer (VNA). The total size of the rectifier is about $30 \times 35\ \text{mm}^2$.

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

1. Introduction

A revolutionary growth in wireless technology attracts huge attention from research community to make the self-sustainable device feasible through RF energy harvesting. It exploits ambient electromagnetic energy transmitted from different RF systems to remotely feed the electronic devices (Nintanavongsa, Muncuk, Lewis, & Chowdhury, 2012). Compared to other harvesting techniques, RF energy harvesting provides relatively predictable energy supply owing to the features of easy availability and less dependency on environmental variations. The typical block diagram of RF energy harvesting circuit is shown in Figure 1. It consists of three major blocks viz; antenna, matching network (MN), diode detector followed by an energy storage. The first element, antenna is employed to capture the RF signals of different frequencies and polarization, while second MN is for maximum power transfer, and the last rectifier is used to convert the RF energy to dc voltage. It means harvesting circuit performance can be evaluated in terms of accessible ambient RF energy and its conversion rate (Agrawal, Pandey, Singh,& Parihar, 2014). These parameters are heavily influenced by surrounding terrain conditions as the multiple reflection and dissipation certainly deteriorate the level of available ambient RF energy. As a result, conversion efficiency and dc output voltage may degrade. Previously, the majority of available RF energy harvesting circuits focused on single frequency band hence offer low dc output voltage. As the multiple RF energy sources of different frequency bands are available, thus from an ambient RF harvesting perspective, the output dc voltage could be increased if the circuit is designed for multiple frequency bands rather than a single band. A wide-band energy harvester can also promise a high output voltage by accumulating the number of RF signals at a time. However, due to nonlinear behavior of the diode, harvesting circuit itself exhibits nonlinearity i.e. its input impedance varies with the received RF power. Thus, it is quite difficult to retain the impedance match and high conversion efficiency over a large frequency range (Song, Huang, Zhou, & Carter, 2014). The losses due to impedance mismatch over a large bandwidth can be illustrated in Collado and Georgiadis (2013), where only 8% conversion efficiency is achieved at -20 dBm.

To address this, it is preferable to harvest energy from several narrow frequency bands rather than a single large one. In literature, numerous topologies have been proposed to accomplish the multiband energy harvesting (Bergès, Fadel, Oyhenart, Vigneras, & Taris, 2015; Hamano et al., 2016; Ho et al., 2016; Keyrouz, Visser, & Tijhuis, 2013; Kuhn, Lahuec, Seguin, & Person, 2015; Liu, Zhong, & Guo, 2015; Niotaki, Georgiadis, Collado, & Vardakas, 2014; Pinuela, Mitcheson, & Lucyszyn, 2013; Scheeler, Korhummel, & Popovic, 2014; Shariati, Rowe, Scott, & Ghorban, 2015; Sun, Guo, He, & Zhong, 2013). These topologies can be differentiated in terms of filter functionality i.e. how the antenna or source impedance is matched to the rectifier circuit. For instance, in Pinuela et al. (2013) and Keyrouz et al. (2013) several single-band rectennas (combination of antenna and rectifier circuit) were stacked to constitute a multi-band harvesting circuit. In this case, each rectenna was designed for a specific frequency band. Thus, for compact applications, this architecture is not suitable due to the number of antennas used. Moreover, in most of the reported works, the quality assessment of the output voltages combination was not taken into consideration. In Kuhn et al. (2015), the circuit complexity is reduced to a certain extent by replacing the multiple antennas with a single wide-band antenna. However, in this topology too, the number of rectifiers increases with the frequency bands, which leads to prolonging the circuit complexity.

Figure 1. Typical block diagram of RF energy harvesting circuit.

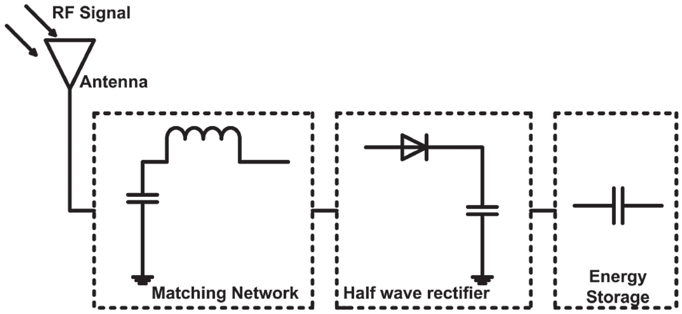

Besides, a multi-band harvesting circuit can also be formed by simply embedding a multi-band matching network between the multi-band antenna and the rectifying circuit (Bergès et al., 2015; Hamano et al., 2016; Ho et al., 2016; Liu et al., 2015; Niotaki et al., 2014; Scheeler et al., 2014; Shariati et al., 2015; Sun et al., 2013). The multi-band matching network can be designed either by distributed or by lumped element. In general, the multi-band rectifier circuit experiences two types of losses: first due to shift in resonance frequency from the optimum frequency point, and second due to the filter complexity. Because of the diode nonlinearity, the input impedance of the circuit varies as a function of power and frequency which causes a shift in resonance frequency. The difficulty due to diode nonlinearity can be observed in Sun et al. (2013) where the dual-band rectifier circuit exhibits its impedance matching for a small range of input power. The losses induced because of filter complexity can be observed in the recently reported works on dual band harvesting circuit (Niotaki et al., 2014; Scheeler et al., 2014; Shariati et al., 2015). In Niotaki et al. (2014), for $P_{\text{in}} = -15$ dBm, author achieved the conversion efficiency of 23% at the expense of increased filter complexity consisting two series and two shunt pairs of reactive elements. Thus, for more than dual band applications, the proposed circuit topology is not suitable due to excessive filtering components used. To obtain good conversion efficiency a dual-band rectenna reported in Scheeler et al. (2014). However, the rectenna was large in size and requiring a complex impedance tuning circuit. In Shariati et al. (2015) also, a dual-band matching network consisting nine reactive elements was employed to achieve the dual-band characteristics.

In order to reduce the filter complexity, this work proposed a compact dual-band harvesting circuit for GSM-900 and 1800. It consists of a 4th order dual-band matching network based on $1-n$ frequency transformation, which is optimized for the energy harvesting circuit to reduce the complexity up to $2 \times n$ reactive elements (n is the number of frequency bands). Similar to frequency transformation method, the proposed dual-band rectifier circuit can be extended to n number of frequency bands by using the $2 \times n$ number of reactive elements. The detailed analysis and design guidelines of dual band rectifier circuit are discussed in Section 2.

2. Dual band rectifier design and analysis

This section presents the design and analysis of a dual-band harvesting circuit in terms of impedance matching, DC output voltage and RF-to-dc conversion efficiency. The topology of the proposed dual-band RF energy harvesting circuit is shown in Figure 2(a). As seen, the low-cost Schottky diode is used to transform the input RF power to DC voltage. The impedance matching at two frequency is achieved using a series and parallel combination of the *LC* pair. The main idea underlying the suggested multi-band matching network is $1-n$ frequency transformation (one to many mapping of frequency), which transforms a single-band matching network to multi-band matching network (Nallam & Chatterjee, 2013). As the name ($1-n$) suggests that for designing a multi-band matching network, primarily a single-band matching network is required whose resonant frequency is dependent on the frequencies for which multi-band matching network proposed to designed.

Moreover, this frequency transformation method depends on the type of load impedance, whether it is series or parallel combination of *RC* or *RL*. Since the selected diode (HSMS-2852) has capacitive behavior throughout the frequency, it can be represented in a series or parallel combination of *R* and *C*. In the case of parallel *RC* load, the following equations are used to transform the single-band matching network into multi-band matching network.

$$\omega = \frac{\omega_t^{n} + a_m \omega_t^{n-m} + a_{m+2}\omega_t^{n-(m+2)} + \cdots}{\omega_t^{n-1} + a_{m+1}\omega_t^{n-(m+1)} + a_{m+3}\omega_t^{n-(m+3)} + \cdots} \tag{1}$$

where, n is the number of bands and m varies from 2 to n. After substituting the value of n, Equation (1) can be expanded in partial fraction form using the causal foster analysis as:

$$\omega = \omega_t + \frac{1}{\frac{\omega_t}{a_2 - a_3} + \frac{1}{\cdots}} \tag{2}$$

Figure 2. (a) Circuit diagram of the proposed dual-band rectifier circuit and (b) optimized dual-band rectifier circuit.

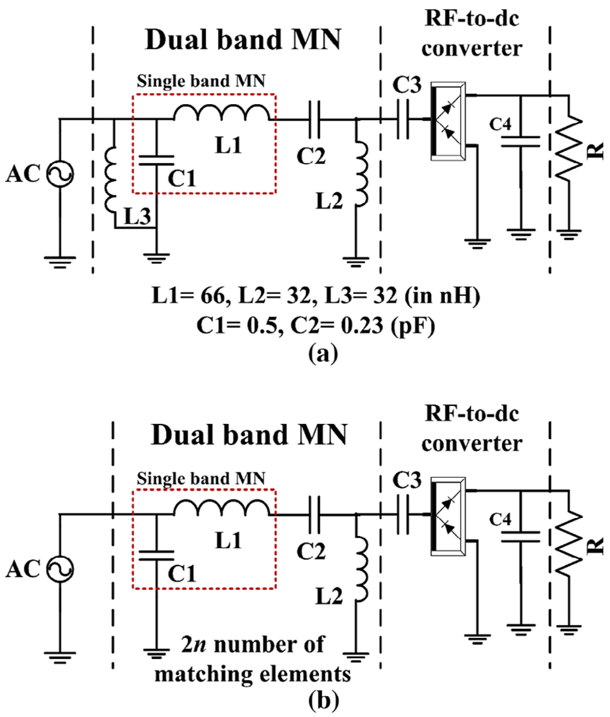
Dual band MN
RF-to-dc converter
Single band MN
C3
C4
R
L1
C2
AC
C1
L2
L3
L1= 66, L2= 32, L3= 32 (in nH)
C1= 0.5, C2= 0.23 (pF)
(a)
Dual band MN
RF-to-dc converter
Single band MN
C3
C4
R
L1
C2
AC
C1
L2
2n number of matching elements
(b)

The coefficients a_2 and a_3 can be calculated as:

$$\omega_m = \sum_{i=1}^{n} (-1)^{i-1} \omega_i \tag{3}$$

$$a_m = (-1)^m \sum_{i,j=1,1 \,\&\, i \neq j}^{n,n} (-1)^{i+j} \omega_i \omega_j \tag{4}$$

$$a_{m+1} = (-1)^{m+1} \sum_{i,j,k=1,1,1 \,\&\, i \neq j \neq k}^{n,n,n} (-1)^{i+j+k} \omega_i \omega_j \omega_k \tag{5}$$

$$a_{m+n} = (-1)^{m+n} \sum_{i,j,k\ldots=1,1,1\ldots i \neq j \neq k}^{n,n,n\ldots} (-1)^{i+j+k+\cdots} \omega_i \omega_j \omega_k \ldots \tag{6}$$

$$a_n = \prod_{i=1}^{n} (-1)^n \omega_i \tag{7}$$

Equation (6) is similar to that presented in Nallam and Chatterjee (2013), except the term $(-1)^{m+n}$, which is included here to realize the multi-band matching network for more than three frequency bands i.e. for $n \geq 3$.

With this transformation, the capacitor of the matching network is transformed to the combination of prototype capacitor parallel with inductor whereas, an inductor is transformed into a combination of the same inductor with a series capacitor. Figure 3 shows the circuit schematic of transformation of a single-band matching network to the dual-band matching network. It can be seen that C_1 is transformed to $C_1 \| L_2$ and L_1 transformed to L_1 series with C_2. After successful usage of (1)–(7), the resultant multi-band matching network requires $3n - 1$ and $4n - 1$ reactive elements for L and Π-type topology, respectively.

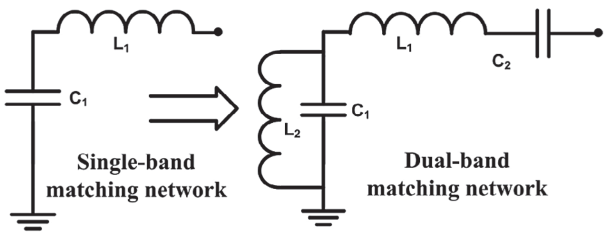

Figure 3. Conversion of a single-band matching network to dual-band matching network using $1-n$ frequency transformation.

As the aim is to design a dual-band harvesting circuit, therefore, we require here only 5 or 7 reactive elements with L and Π-type topologies, respectively. From Figure 2(a), it can be seen that the resultant matching network consists of five elements, where the inductor L_3 and capacitor C_2 results after the transformation of capacitor C_1 and inductor L_1, respectively. Besides, the inductor L_2 occurrs due to the diode reactive element, which is generally a capacitor.

In this work, two frequencies 0.9 and 1.8 GHz that correspond to the maximum signal strength are chosen for dual-band harvesting circuit. According to this method, it is necessary to assign the frequencies in descending order e.g. $\omega_1 = 1.8, \omega_2 = 0.9$. Therefore, from (3) single-band matching network frequency is equal to $\omega_1 - \omega_2 = 2\pi(1.8 - 0.9) \times 10^9 = 0.9 \times 2\pi \times 10^9$. In order to match the source impedance with the rectifier at the calculated frequency 0.9 GHz, the chosen matching topology is L-type as shown by the encircled portion in Figure 2(a). The corresponding element values can be approximated using the various methods some of which are described in Pozar (2010). Subsequently, this single-band matching network is transformed to dual-band matching network using (1)–(7). The detailed design steps of the dual-band rectifier circuit are summarized as follows:

(1) As we are interested in matching the diode to 50 Ω at two frequencies (0.9 and 1.8 GHz) so, the order of transformation is equal to 2 or $n = 2$.

(2) In the first step, single-band matching network is designed at the frequency f calculated as: $f = f_2 - f_1 = 1.8 - 0.9 = 0.9$ GHz. In this case, any matching topology that matches the diode to 50 Ω, at 0.9 GHz for an input power $P_{\text{in}} = -20$ dBm, and load resistance 4.7 kΩ can be used. The chosen single-band matching network is shown by the encircled portion in Figure 2(a).

(3) Afterwards, this single-band matching network is transformed into dual-band using the (1)–(7) as shown below:

Since $n = 2$ therefore, from (2)

$$\omega = \omega_t + \frac{1}{\frac{\omega_t}{a_2}} \tag{8}$$

From (4) a_2 can be calculated as:

$$a_2 = (-1)^{2+2} \sum (-1)^{1+2} \omega_1 \omega_2 = \omega_1 \omega_2 \tag{9}$$

$$a_2 = -1.62 \times 4\pi^2 \times 10^8 = 0.64 \times 10^{20} \tag{10}$$

Thus, inductor L_1 (=66 nH) is transformed to impedance as:

$$j66 \times 10^9 \omega = j66 \times 10^9 \omega_t + \frac{1}{j0.23 \times 10^{-12} \omega_t} \tag{11}$$

Similarly, capacitors (=0.5 pF) are transformed to the admittance as:

$$j5 \times 10^{-13} \omega = j5 \times 10^{-13} \omega_t + \frac{1}{j32 \times 10^{-9} \omega_t} \tag{12}$$

The circuit schematic of the dual-band harvesting circuit is shown in Figure 2(a). It can be seen that resultant matching network consists of five reactive elements according to $3n - 1$. In order to reduce the circuit complexity and sensitivity due to reactive elements, a parametric study has been carried out to eliminate the elements showing minimum influence on the circuit performance.

Figure 4 shows the simulated $|S_{11}|$ for the different combination of matching elements. The simulated results demonstrate that $|S_{11}|$ experiences maximum change when inductor L_2 and capacitor C_2 are removed from the circuit, whereas it remains almost unaffected when L_3 is not present in the circuit. Therefore, inductor L_3 can be extruded from the circuit and the resultant matching circuit requires only $2n$ and $3n$ reactive elements in place of $3n - 1$ and $4n - 1$ elements. In this way, for each topology, the proposed circuit reduces $n - 1$ elements compared to the conventional method. Figure 2(b) demonstrates the optimized circuit diagram of the dual-band rectifier. It can be observed that circuit requires large inductors value of 32 and 66 nH. Thus, it is quite difficult to realize the practical rectifier circuit whose response is similar to the response of simulated result. In order to avoid any impedance mismatch due to the small difference in elements value, the meander line inductor and open stub are used to realize the inductors and capacitors, respectively. In this case, not only fabrication and optimization process become so easy but the cost will also reduced.

Figure 5, shows the layout of the dual-band rectifier circuit. In Nintanavongsa et al. (2012) and Agrawal et al. (2014), it has been demonstrated that the number of rectifying diodes or equivalently voltage multiplier stages are very much sensitive to the RF-to-dc conversion efficiency. In low-power region (≤ -20 dBm), efficiency decreases if voltage multiplier stages increase, whereas in higher power region (≥ -20 dBm), an opposite effect occurs. As the demand is to harvest energy in low-power region, single-series circuit with a double diode is used to convert received RF energy into dc voltage. From the left side of the circuit, the first meander line corresponds to the inductor L_1, while the second meander line represents the inductor L_2, of the Figure 2(b). The shunt stub is accounted

Figure 4. Simulated $|S_{11}|$ vs. frequency for different combination of circuit elements.

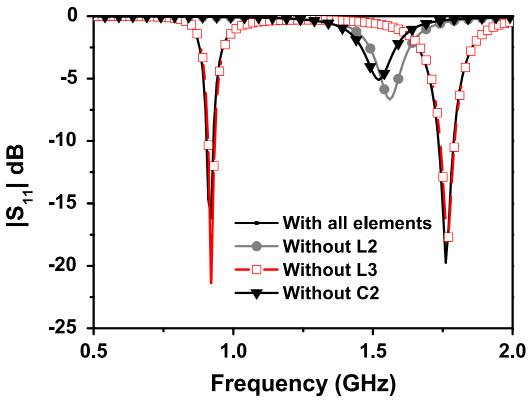

Figure 5. Layout of the proposed dual-band RF energy harvesting circuit.

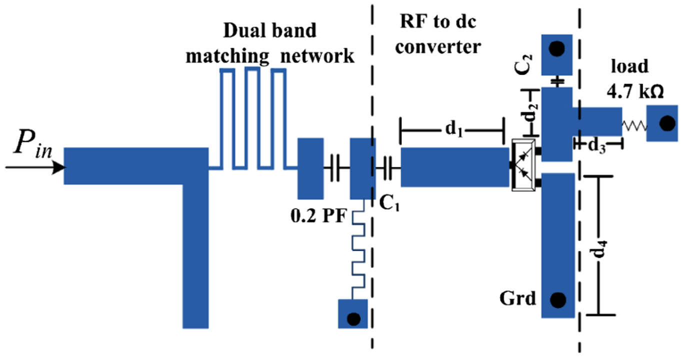

for the shunt capacitor C_1 of Figure 2(b). The dimensions of each element are calculated according to their respective reactive element value and the substrate on which circuit has to be fabricated. In Assimonis, Daskalakis, and Bletsas (2016), it has been demonstrated that traces (microstrips) connected to the rectifier terminals (e.g. distance between the diode and via and diode and load) are highly sensitive for RF-to-dc efficiency. Therefore, traces d_1, d_2 between diode and capacitor C_1 and C_2, d_3 between diode and load and d_4 between diode and ground are adjusted to optimize the impedance matching as well as the conversion efficiency of the rectifier. Due to nonlinear behavior of the diode, harvesting circuit itself exhibits nonlinearity i.e. its input impedance varies with received RF power, therefore harmonic-balance (HB) and large signal analysis (LSSP) were employed to take into consideration the nonlinear behavior of the rectifier.

The photograph of the fabricated dual-band rectifier is shown in Figure 6. It is fabricated on a 1.54 mm thick FR-4 substrate with a dielectric constant (ϵ_r) of 4.3 using chemical etching method. The rectifier performance is evaluated in terms of $|S_{11}|$ and output voltage using the Agilent vector network analyzer (VNA). The simulated and measured $|S_{11}|$ is illustrated in Figure 7(a). The measured result shows reasonable agreement with the simulated one; the slight difference can be accounted for the fabrication imperfections. It is well known that impedance matching is a function of frequency and input power, due to the nonlinearity of the diode. Such a characteristic is examined in Figure 7(b), where the measured $|S_{11}|$ is demonstrated as a function of input power level for three different load impedance values. From results, it is clear that impedance matching of the harvesting circuit is greatly affected by the input power and the load impedance. Figure 7(b) demonstrates that as power increases, the impedance matching at 0.9 GHz degraded drastically, while at 1.8 GHz, it improves. Moreover, it is noticed that the impedance matching at higher power level is more sensitive to the variation of load impedance (*RL*).

The measured RF-to-dc conversion efficiency and output voltage vs. input power for both frequencies are demonstrated in Figure 8(a). For 0.9 GHz, efficiency is equal to 25.7 and 65.1% for an input power of -20 and 0 dBm, respectively. However, at 1.8 GHz the efficiency is relatively small that might be due to the increased parasitic losses in the rectifier diode. Figure 8(b) shows the relation between the output voltage and frequency for various input power levels at fixed load resistance value of 4.7 kΩ. It can be seen that maximum output voltage is achieved in the frequency range of 860–900 and 1770–1800 MHz, showing the rectifier's capability to harvest RF energy in the GSM-900 and 1800 bands.

Figure 9, depicts the measured conversion for various values of load resistance. It can be noticed that the circuit yields maximum efficiency when the load impedance is 4.7 kΩ. It starts decreasing as the load impedance varies from 4.7 kΩ. Table 1 shows the comparison of the conversion efficiency and the size of the proposed rectifier with the similar works reported previously. Only measured results are compared in Table 1. It can be seen that in low-power condition maximum efficiency is achieved in Sun et al. (2013), but the expense of bulky circuit size. However, the proposed rectifier offers an optimal conversion efficiency with compact circuit size.

Figure 6. Photograph of the fabricated dual-band RF energy harvesting circuit.

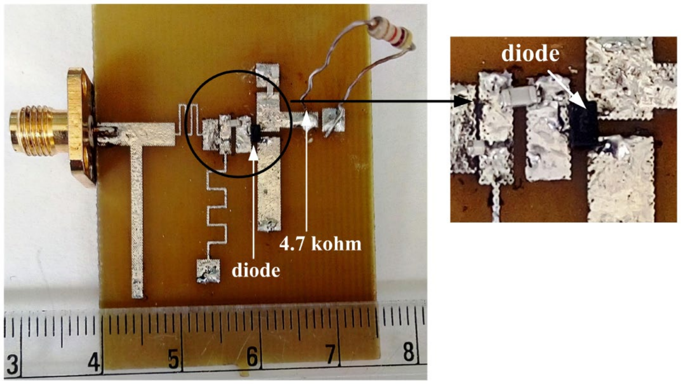

Figure 7. (a) Simulated and measured $|S_{11}|$ vs. frequency and (b) measured $|S_{11}|$ for various power levels.

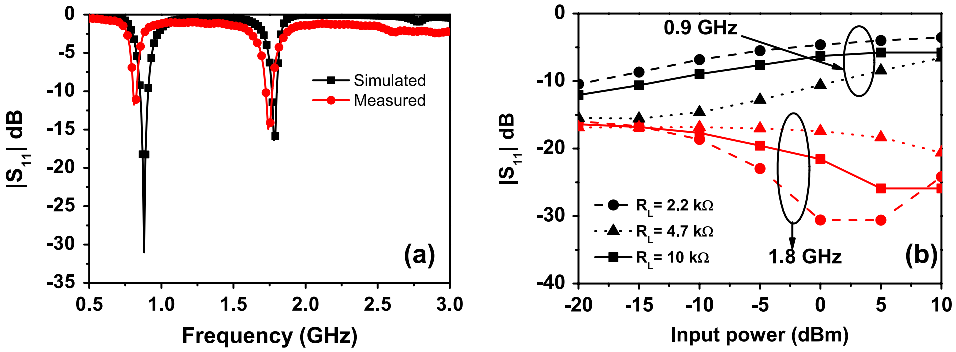

Figure 8. (a) Measured output dc voltage and RF-to-dc conversion efficiency and (b) measured dc voltage vs. frequency.

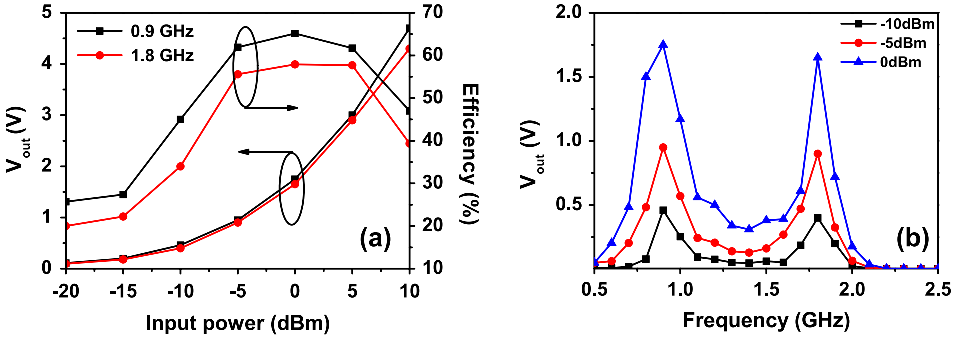

Figure 9. Measured RF-to-dc conversion efficiency for different load impedance.

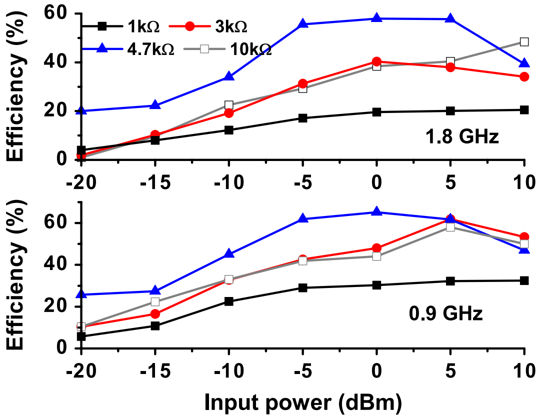

Table 1. Performance comparison of the proposed dual rectifier with recently published works

Ref.	Measured rectifier efficiency (%)	Input power (dBm)	Rectifier size
Ho et al. (2016)	15.8 @ 0.89 GHz	−20	100×65 mm^2
	11.2 @ 1.76 GHz		
Hamano et al. (2016)	10 @ 2.15 GHz	−10	37×71 mm^2
	15 @ 5.84 GHz		
Bergès et al. (2015)	27 @ 0.91/2.4 GHz	−16	78×88 mm^2
Sun et al. (2013)	30 @ 2.14 GHz	−20	145 mm
	35 @ 1.84 GHz		
Liu et al. (15)	20 @ (0.91+1.8) GHz	−20	23×37 mm^2
This work	27.5 @ 0.9 GHz	−20	30×35 mm^2
	20 @ 1.8 GHz		

3. Conclusion

A new compact 4th order dual-band rectifier has been designed to harvest the RF power of GSM-900 and 1800 bands. In order to reduce the circuit complexity and sensitivity due to reactive elements, the meander line and the open stub are used to fabricate the matching network. For $P_{\text{in}} = -20$ dBm, the measured RF-to-dc conversion efficiency of 27.5 and 20% is achieved at 0.9 and 1.8 GHz, respectively. Further, more than 45 and 34% conversion efficiency is maintained from -10 to 10 dBm for 0.9 and 1.8 GHz, respectively.

Funding
The authors received no direct funding for this research.

Author details
Sachin Agrawal[1]
E-mail: bitssachin.agrawal@gmail.com
Manoj S. Parihar[1]
E-mail: mparihar@iiitdmj.ac.in
P.N. Kondekar[1]
E-mail: pnkondekar@iiitdmj.ac.in
[1] PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur 482005, India.

Citation information
Cite this article as: A dual-band RF energy harvesting circuit using 4th order dual-band matching network, Sachin Agrawal, Manoj S. Parihar & P.N. Kondekar, *Cogent Engineering* (2017), 4: 1332705.

References

Agrawal, S., Pandey, S. K., Singh, J., & Parihar, M. S. (2014). *Realization of efficient RF energy harvesting circuits employing different matching technique* (pp. 754–761). Fifteenth International Symposium on Quality Electronic Design, Santa Clara, CA.

Assimonis, S. D., Daskalakis, S. N., & Bletsas, A. (2016). Sensitive and efficient rf harvesting supply for batteryless backscatter sensor networks. *IEEE Transactions on Microwave Theory and Techniques, 64*, 1327–1338.

Bergès, R., Fadel, L., Oyhenart, L., Vigneras, V., & Taris, T. (2015). A dual band 915MHz/2.44 GHz RF energy harvester. *Microwave Conference (EuMC), 2015 European*. IEEE.

Collado, A., & Georgiadis, A. (2013). Conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna. *IEEE Transactions on Circuits and Systems I: Regular Papers, 60*, p. 10.

Hamano, K., Tanaka, R., Yoshida, S., Miyachi, A., Nishikawa, K., & Kawasaki, S. (2016). Design of dual-band rectifier using microstrip spurline notch filter. *Radio-Frequency Integration Technology (RFIT), 2016 IEEE International Symposium on*. IEEE.

Ho, D. K., Kharrat, I., Ngo, V. D., Vuong, T. P., Nguyen, Q. C., & Le, M. T. (2016). *Dual-band rectenna for ambient RF energy harvesting at GSM 900 MHz and 1800 MHz* (pp. 306–310). 2016 IEEE International Conference on Sustainable Energy Technologies (ICSET), Hanoi.

Keyrouz, S., Visser, H., & Tijhuis, A. (2013, April). Multi-band simultaneous radio frequency energy harvesting. *In Antennas and Propagation (EuCAP), 2013 7th European Conference on*, 3058–3061.

Kuhn, V., Lahuec, C., Seguin, F., & Person, C. (2015). A multi-band stacked RF energy harvester with RF-to-DC efficiency up to 84\%. *IEEE Transactions on Microwave Theory and Techniques, 63*, 1768–1778.

Liu, Z., Zhong, Z., & Guo, Y.-X. (2015). Enhanced dual-band ambient RF energy harvesting with ultra-wide power range. *IEEE Microwave and Wireless Components Letters, 25*, 630–632.

Nallam, N., & Chatterjee, S. (2013, June). Multi-band frequency transformations, matching networks and amplifiers. *Circuits and Systems I: Regular Papers, IEEE Transactions on, 60*, 1635–1647.

Nintanavongsa, P., Muncuk, U., Lewis, D., & Chowdhury, K. (2012, March). Design optimization and implementation for rf energy harvesting circuits. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2*, 24–33.

Niotaki, K., Georgiadis, A., Collado, A., & Vardakas, J. S. (2014). Dual-band resistance compression networks for improved rectifier performance. *IEEE Transactions on Microwave Theory and Techniques, 62*, 3512–3521.

Pinuela, M., Mitcheson, P., & Lucyszyn, S. (2013, July). Ambient rf energy harvesting in urban and semi-urban environments. *Microwave Theory and Techniques, IEEE Transactions on, 61*, 2715–2726.

Pozar, D. M. (2010). *Microwave engineering* (3rd ed.). New Delhi: Wiley.

Scheeler, R., Korhummel, S., & Popovic, Z. (2014). A dual-frequency ultralowpower efficient 0.5-g rectenna. *IEEE Microwave Magazine, 15*, 109–114.

Shariati, N., Rowe, W. S., Scott, J. R., & Ghorban, K. (2015). Multi-service highly sensitive rectifier for enhanced RF energy scavenging. *Scientific Reports, 5*, 9655

Song, C., Huang, Y., Zhou, J., & Carter, P. (2014). Improved ultrawideband rectennas using hybrid resistance compression technique. *IEEE Transactions on Antennas and Propagation, 65*, 2057–2062.

Sun, H., Guo, Y. X., He, M., & Zhong, Z. (2013). A dual-band rectenna using broadband Yagi antenna array for ambient RF power harvesting. *IEEE Antennas and Wireless Propagation Letters, 12*, 918–921. doi:10.1109/LAWP.2013.2272873.

© 2017 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions
You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

***Cogent Engineering* (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com

Circuits and Systems, **2012, 3, 216-222**
http://dx.doi.org/10.4236/cs.2012.33030 Published Online July 2012 (http://www.SciRP.org/journal/cs)

Optimization of the Voltage Doubler Stages in an RF-DC Convertor Module for Energy Harvesting

Kavuri Kasi Annapurna Devi[1*], Norashidah Md. Din[2], Chandan Kumar Chakrabarty[2]
[1]Department of Electrical and Electronic Engineering, INTI International University, Nilai, Malaysia
[2]Department of Electronics and Communication Engineering, Universiti Tenega Nasional, Kajang, Malaysia
Email: {*kavurik.adevi., chandan}@newinti.edu.my, norashidah@uniten.edu.my

Received August 25, 2011; revised March 26, 2012; accepted April 4, 2012

ABSTRACT

This paper presents an optimization of the voltage doubler stages in an energy conversion module for Radio Frequency (RF) energy harvesting system at 900 MHz band. The function of the energy conversion module is to convert the (RF) signals into direct-current (DC) voltage at the given frequency band to power the low power devices/circuits. The design is based on the Villard voltage doubler circuit. A 7 stage Schottky diode voltage doubler circuit is designed, modeled, simulated, fabricated and tested in this work. Multisim was used for the modeling and simulation work. Simulation and measurement were carried out for various input power levels at the specified frequency band. For an equivalent incident signal of –40 dBm, the circuit can produce 3 mV across a 100 kΩ load. The results also show that there is a multiplication factor of 22 at 0 dBm and produces DC output voltage of 5.0 V in measurement. This voltage can be used to power low power sensors in sensor networks ultimately in place of batteries.

Keywords: Energy Conversion; RF; Schottky Diode; Villard; Energy Harvesting

1. Introduction

RF energy harvesting is one type of energy harvesting that can be potentially harvested such as solar, vibration and wind. The RF energy harvesting uses the idea of capturing transmitted RF energy at ambient and either using it directly to power a low power circuit or storing it for later use. The concept needs an efficient antenna along with a circuit capable of converting RF signals to DC voltage. The efficiency of an antenna mainly depends on its impedance and the impedance of the energy converting circuit. If the two impedances aren't matched then it will be unable to receive all the available power from the free space at the desired frequency band. Matching of the impedances means that the impedance of the antenna is the complex conjugate of the impedance of the circuit (voltage doubler circuit).

The concept of energy harvesting system is shown in **Figure 1**, which consists of matching network, RF-DC conversion and load circuits. The authors in [1], used a 2.4 GHz operating frequency with an integrated zero bias detector circuit using BiCMOS technology which produced an output voltage of 1 V into a 1 MΩ load at an input power level of 0 dBm. H. Yan and co-authors revealed that a DC voltage of 0.8 volts can be achieved from a –20 dBm RF input signal at 868.3 MHz through simulation results [2]. In [3], work was carried out on a firm frequency of 900 MHz by matching to a 50 Ω impedance and resonance circuit transformation in front of the Schottky diode which yields an output voltage of over 300 mV at an input power level of 2.5 µ. W. J. Wang, L. Dong and Y. Fu [4] used a Cockcroft-Walton multiplier circuit that produced a voltage level of 1.0 V into a 200 MΩ load for an input power level of less than –30 dBm at a fixed frequency of 2.4 GHz.

The energy conversion module designed in this paper is based on a voltage doubler circuit which can be able to output a DC voltage typically larger than a simple diode rectifier circuit as in [5], in which switched capacitor charge pump circuits are used to design two phase voltage doubler and a multiphase voltage doubler. The module presented in this can function as an AC to DC converter that not only rectifies the input AC signal but also elevates the DC voltage level. The output voltage of the

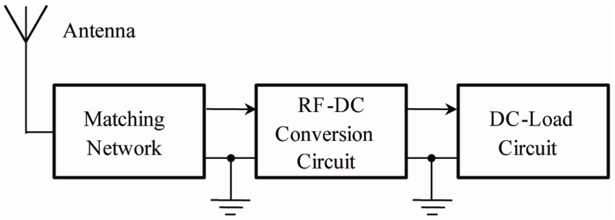

Figure 1. Schematic view of a RF energy harvesting system.

---

[*] Corresponding author.

Copyright © 2012 SciRes.

energy conversion module can be used to energize the low power devices for example sensors for a sensor network in application to agriculture.

Section 2 of this paper discusses on the theoretical background of the voltage doubler circuit. Section 3 presents the simulation study and implementation of the circuit design. Section 4 provides the results and analysis. Section 5 concludes with a discussion on the findings from the simulated and measured results.

2. Voltage Multiplier

There are various voltage multiplier circuit topologies. The design used in this module is derived from the function of peak detector or a half wave peak rectifier. The Villard voltage multiplier circuit was chosen in the circuit design of this paper because it produces two times of the input signal voltage towards ground at a single output and can be cascaded to form a voltage multiplier with an arbitrary output voltage and its design simplicity.

2.1. Diode Modeling

The voltage multiplier circuit in this design uses zero bias Schottky diode HSMS-2850 from Agilent. The attractive feature of these Schottky diodes are low substrate losses and very fast switching but leads to a fabrication overhead. This diode has been modeled for the energy harvesting circuit which comes in a one-diode configuration. The modeling parameters for these diodes are given by Agilent in their data sheets. These parameters are used in Multisim for its own modeling purposes. The modeling is done by transforming the diode into an equivalent circuit using passive components which are described by the SPICE parameters in **Table 1** [6].

The diode used in this design is shown in **Figure 2** and its equivalent model is shown in **Figure 3**. The special features of HSMS-2850 diode is that it provides a low forward voltage, low substrate leakage and uses the non symmetric properties of a diode that allows unidirectional flow of current under ideal condition.

Table 1. SPICE parameters.

Parameters	Units	HSMS 2850
B_V	V	3.8
C_{J0}	pF	0.18
E_G	Ev	0.69
I_{BV}	A	3E–4
I_S	A	3E–6
N	No unite	1.06
R_S	Ω	25
P_B (V_J)	V	0.35
P_T (XTI)	No units	2
M	No units	0.5

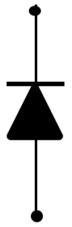

Figure 2. Schottky diode.

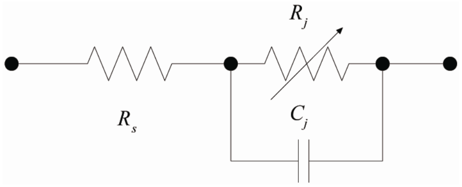

Figure 3. Linear circuit model of the Schottky diode [6].

The diodes are fixed and are not subject of optimization or tuning. This is described using the following derivations. By neglecting the effect of diode substrate, an equivalent linear model that can be used for the diode as shown in **Figure 3**. When C_j is the junction capacitance and R_j is the junction resistance, the admittance Y_z of the linear model is given by

$$Y_Z = Y_{C_j} + Y_{R_j} \tag{1}$$

Equation (1) related to the frequency of operation is given by

$$Y_Z = jwC_j + \frac{1}{R_j} \tag{2}$$

$$= \frac{jwC_jR_j + 1}{R_j} \tag{3}$$

The impedance Z of the linear model is given by

$$Z = \frac{R_j}{1 + jwR_jC_j} \tag{4}$$

The total impedance Z_T is given by

$$Z_T = R_S + \frac{R_j}{1 + jwR_jC_j} \tag{5}$$

where R_S is the series resistance of the circuit and R_j is given by

$$R_j = \frac{8.33 \times 10^{-5} \times N \times T}{I_b + I_S}$$

where:

I_b = bias current in μA;
I_s = saturation current in μA;
T = temperature (K);

Copyright © 2012 SciRes.

N = ideality factor.

In Equation (5), R_j and C_j are constants and the frequency of operation (w) is the only variable parameter. As the frequency increases, the value of Z is almost negligible compared to the series resistance R_S of the diode. From this it is concluded that the function of the diode is independent of the frequency of operation.

2.2. Single Stage Voltage Multiplier

Figure 4 represents a single stage voltage multiplier circuit. The circuit is also called as a voltage doubler because in theory, the voltage that is arrived on the output is approximately twice that at the input. The circuit consists of two sections; each comprises a diode and a capacitor for rectification. The RF input signal is rectified in the positive half of the input cycle, followed by the negative half of the input cycle. But, the voltage stored on the input capacitor during one half cycle is transferred to the output capacitor during the next half cycle of the input signal. Thus, the voltage on output capacitor is roughly two times the peak voltage of the RF source minus the turn-on voltage of the diode.

The most interesting feature of this circuit is that when these stages are connected in series. This method behaves akin to the principle of stacking batteries in series to get more voltage at the output. The output of the first stage is not exactly pure DC voltage and it is basically an AC signal with a DC offset voltage. This is equivalent to a DC signal superimposed by ripple content. Due to this distinctive feature, succeeding stages in the circuit can get more voltage than the preceding stages. If a second stage is added on top of the first multiplier circuit, the only waveform that the second stage receives is the noise of the first stage. This noise is then doubled and added to the DC voltage of the first stage. Therefore, the more stages that are added, theoretically, more voltage will come from the system regardless of the input. Each independent stage with its dedicated voltage doubler circuit can be seen as a single battery with open circuit output voltage V_0, internal resistance R_0 with load resistance R_L, the output voltage, V_{out} is expressed as in Equation (7).

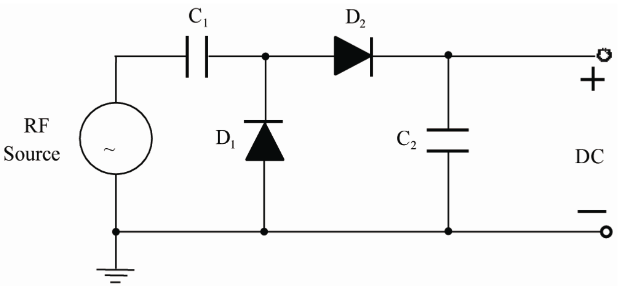

Figure 4. Single stage voltage multiplier circuit [7].

$$V_{\text{out}} = \frac{V_0}{R_0 + R_L} R_L \tag{6}$$

When n number of these circuits are put in series and connected to a load of R_L in Equation (6) the output voltage V_{out} obtained is given by this change in RC value will make the time constant longer which in turn retains the multiplication effect of two in this design of seven stage voltage doubler.

$$V_{\text{out}} = \frac{nV_0}{nR_0 + R_L} = V_0 \frac{1}{\frac{R_0}{R_L} + \frac{1}{n}} \tag{7}$$

The number of stages in the system has the greatest effect on the DC output voltage, as shown from Equations (6) and (7).

It is inferred that the output voltage V_{out} is determined by the addition of R_0/R_L and $1/n$, if V_0 is fixed. From this analysis it is observed that V_0, R_0 and R_L are all constants. Assume that $V_0 = 1$ V, $R_0/R_L = 0.25$, n = 2, 3, 4, 5, 6 and 7, the output voltage V_{out} = 1.33 V, 1.72 V, 2.0 V, 2.22 V, 2.43 V and 2.56 V respectively when substituted analytically in the Equation (7). This analysis can be compared with the results obtained in the circuit design of this module. In simulation at $n = 4$, $V_{\text{out}} = 1.42$ V, $n = 5$, $V_{\text{out}} = 1.67$ V; $n = 6$, $V_{\text{out}} = 1.92$; $n = 7$, $V_{\text{out}} = 2.15$ V; $n = 8$, $V_{\text{out}} = 1.92$ V; $n = 9$, $V_{\text{out}} = 1.81$ V. Also in measurement, for $n = 4$, $V_{\text{out}} = 2.1$ V; $n = 5$, $V_{\text{out}} = 2.9$ V; $n = 6$, $V_{\text{out}} = 3.72$ V; and $n = 7$, $V_{\text{out}} = 5$ V. As n increases, the increase in output voltage will be almost double the input voltage up to some number of stages. But at some point, *i.e.* beyond seven stages, in this circuit the output voltage gained (8 and 9 stages) will be negligible as shown in **Figure 5**.

The capacitors are charged to the peak value of the input RF signal and discharge to the series resistance (R_s) of the diode. Thus the output voltage across the capacitor of the first stage is approximately twice that of the input signal. As the signal swings from one stage to other, there is an additive resistance in the discharge path of the

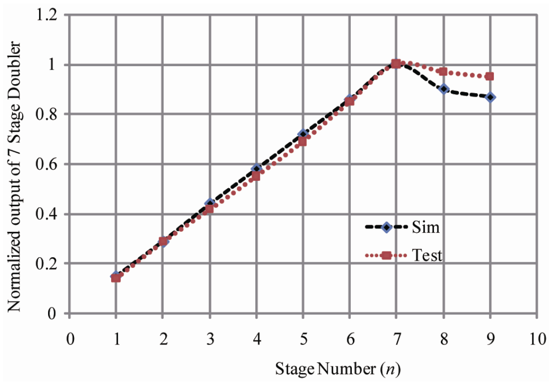

Figure 5. Normalized output voltage multiplier versus number of stages.

diode and increase of capacitance due to the stage capacitors.

2.3. Seven Stage Voltage Multiplier

The seven stage voltage multiplier circuit design implemented in this paper is shown in **Figure 6**. Starting on the left side, there is a RF signal source for the circuit followed by the first stage of the voltage multiplier circuit. Each stage is stacked onto the previous stage as shown in the **Figure 6**. Stacking was done from left to right for simplicity instead of conventional stacking from bottom to top.

The circuit uses eight zero bias Schottky surface-mount Agilent HSMS-285X series, HSMS-2850 diodes. The special features of these diode is that, it provides a low forward voltage, low substrate leakage and uses the non-symmetric properties of a diode that allows unidirectional flow of current under ideal conditions. The diodes are fixed and are not subject of optimization or tuning. This type of multiplier produces a DC voltage which depends on the incident RF voltage. Input to the circuit is a predefined RF source. The voltage conversion can be effective only if the input voltage is higher than the Schottky forward voltage.

The other components associated with the circuit are the stage capacitors. The chosen capacitors for this circuit are of through-hole type, which make it easier to modify for optimization, where in [8] the optimization was accomplished at the input impedance of the CMOS chip for a three stage voltage multiplier. The circuit design in this paper uses a capacitor across the load to store and provide DC leveling of the output voltage and its value only affects the speed of the transient response. Without a capacitor across the load, the output is not a good DC signal, but more of an offset AC signal.

In addition to the above, an equivalent load resistor is connected at the final node. The output voltage across the load decreases during the negative half cycle of the AC input signal. The voltage decreases is inversely proportional to the product of resistance and capacitance across the load. Without the load resistor on the circuit, the voltage would be hold indefinitely on the capacitor and look like a DC signal, assuming ideal components. In the design, the individual components of the stages need not to be rated to withstand the entire output voltage. Each component only needs to be concerned with the relative voltage differences directly across its own terminals and of the components immediately adjacent to it. In this type of circuitry, the circuit does not change the output voltage but increases the possible output current by a factor of two. The number of stages in the system is directly proportional to the amount of voltage obtained and has the greatest effect on the output voltage as explained in the Equation (7) and shown in **Figure 5**.

3. Simulation and Implementation

Multisim software was chosen for modeling and simulation which is a circuit simulation tool by Texas Instruments. The simulation and practical implementation were carried out with fixed RF at 945 MHz ± 100 MHz, which are close to the down link center radio frequency (947.5 MHz) of the GSM-900 transmitter. The voltages obtained at the final node V_{DC} of the multiplier circuit were recorded for various input power levels from –40 dBm - +5 dBm with power level interval (spacing) of 5 dBm.

The simulations were also carried out using same stage capacitance value (3.3 nF) and then with a varied capacitance value for all stages from 4 stages through 9 stages [9]. The capacitance value was varied in such a way that, from one stage to the next, it was halved. For example, if the first stage was 3.3 nF, the second stage was 1.65 nF, third stage was 825 pF, fourth stage was 415 pF and so on. But keeping in view of testing, the capacitance values were chosen to have a close match with the standard available values in the market.

Simulation was carried out through 4 to 9 voltage doubler stages. Based on results obtained a 7 stage doubler is best to implemented for this application.

The design of the printed circuit board (PCB) was carried out using DipTrace software. The material used to

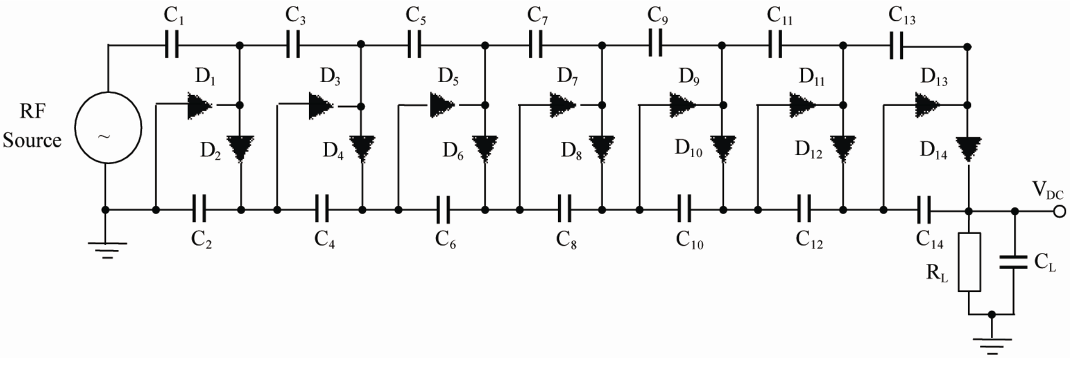

Figure 6. Schematic of 7 stage voltage multiplier.

manufacture the printed circuit board (PCB) is the standard Fiberglass Reinforced Epoxy (FR4), with the thickness of 1.6 mm and dielectric constant of 3.9. The topology is constructed on the PCB with the dimensions of 98 mm × 34 mm (W × H). The Sub Miniature version A (SMA) connectors are used at the input and output of PCB to carry out the measurements. The circuit components consist of active and passive components. The component used in circuit is shown in **Table 2**.

Special handling precautions have been taken to avoid Electro Static Discharge (ESD), while assembling of the surface-mount zero bias Schottky diodes. Also special attention has been given to mount other components and the SMA connectors on to the PCB. The Photograph of Assembled circuit board I shown in **Figure 7**.

4. Results and Analysis

The simulated and measured results at the output voltage of voltage multiplier circuit are shown graphically in **Figure 8**. From the graph analysis, the simulated and the measured results agree considerably with each other. The measured results are shown to be better than the simulation results. The reason behind this may be due to the uncertainty in series resistance value of the diode obtained from SPICE parameters in modeling as explained in Equation (5). This resistance vale of diodes in practical circuit may be lower than in the model, which provides fast discharge path, in turn rise in voltage as passes through the stages and reaches to final output. In this work, the DC output voltages obtained through simulation and measurement at 0 dBm re 2.12 V and 5.0 V respectively. These results are comparatively much better than in ref. [9], where in at 0 dBm, 900 MHz they achieved 0.5 V and 0.8 V through simulation and measurement respectively.

Figures 9 and **10** show the result of a 4 stage voltage doubler circuit with equal and varied capacitance values between the stages as described in Section 3.

From the analysis of these two simulations, it can be observed that the resulting output voltages are equal. The only difference between these two graphs is the rise time of the circuit with varied capacitance value is a little bit slower. But, overall result on the performance of rise time is still under 20 µs to 24 µs and the difference is negligible. From these results, the use of equal stage capacitance of each being 3.3 nF was hence considered for the design of the multiplier.

The results from **Figure 11**, shows that the output voltage reaches to 1.0 V within 20 µS and then uniformly increasing to 1.4 V, 1.67 V, 1.87 V and 2.12 V for 4, 5, 6 and 7 stages respectively compared to 2 mS as shown in [10]. **Figure 12** shows that the conversion ratio of 22 is achieved at 0 dBm input power and drops to 2.5 at –40 dBm. The highest value at 0 dBm is due to the innate characteristics of the zero bias Schottky diodes which conduct fairly well at higher input voltages.

5. Conclusion

From the experimental results, it is found that the pro-

Table 2. Component used in 7 stage voltage multiplier.

Name of component	Label	Value
Stage capacitors	C_1 - C_{14}	3.3 nF
Stage diodes	D_1 - D_{14}	HSMS 2850
Filter capacitor	C_L	100 nF
Load resister	R_L	100 kΩ

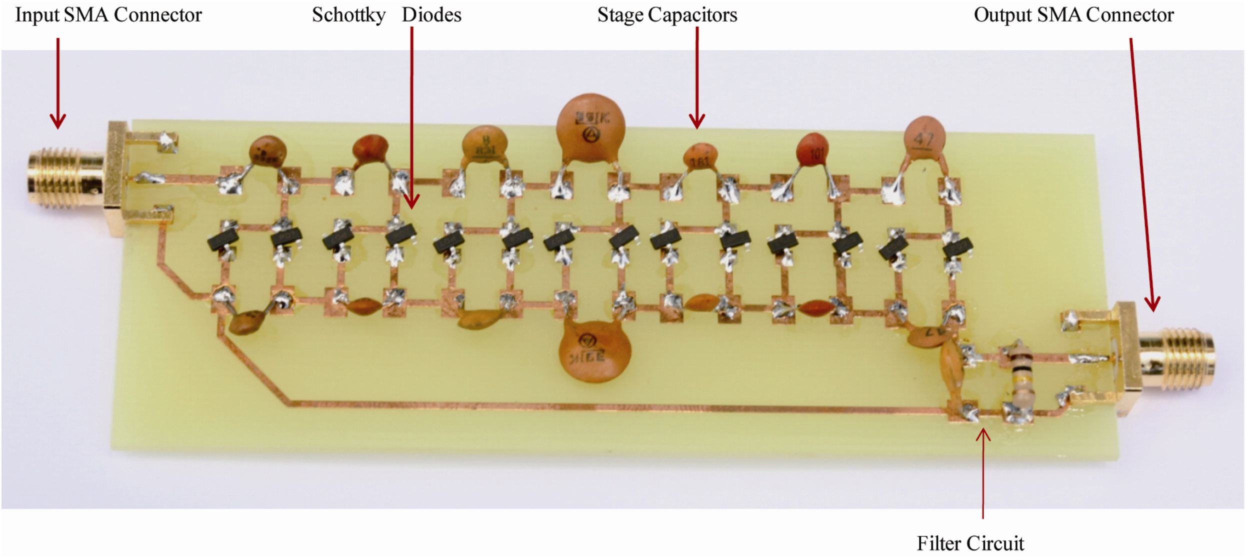

Figure 7. Photograph of assembled circuit board.

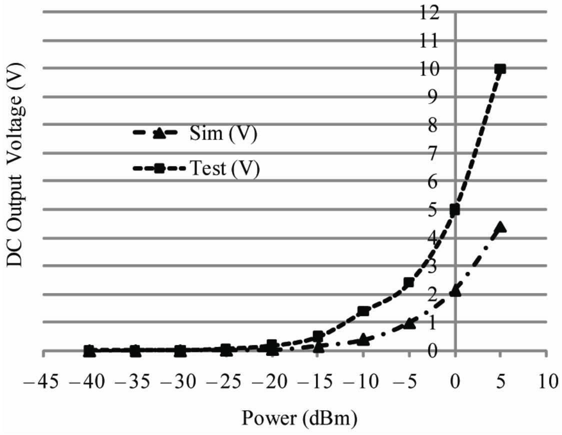

Figure 8. Simulated and test DC output voltage of multiplier as a function of input power.

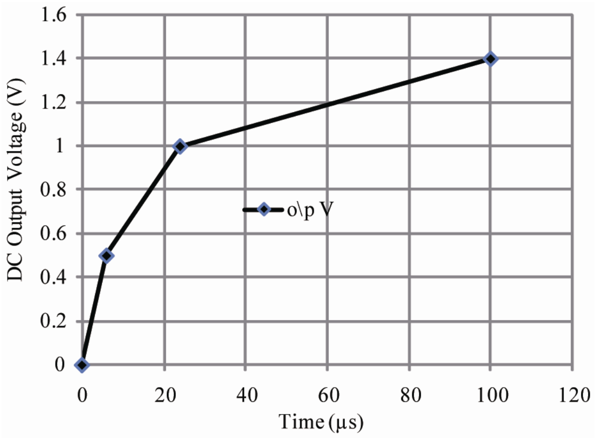

Figure 9. DC output voltage verses rise time of 4 stage voltage doubler circuit with equal stage capacitance [8].

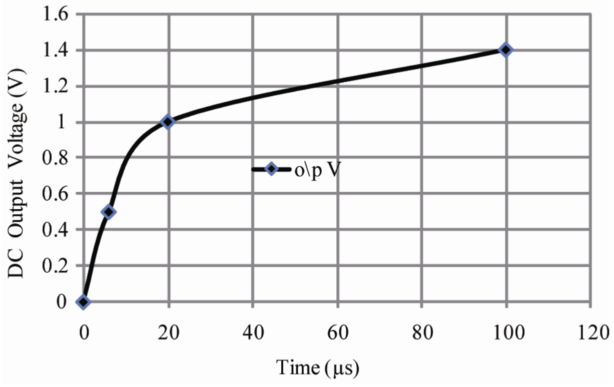

Figure 10. DC output voltage verses rise time of 4 stage voltage doubler with varied stage capacitance [8].

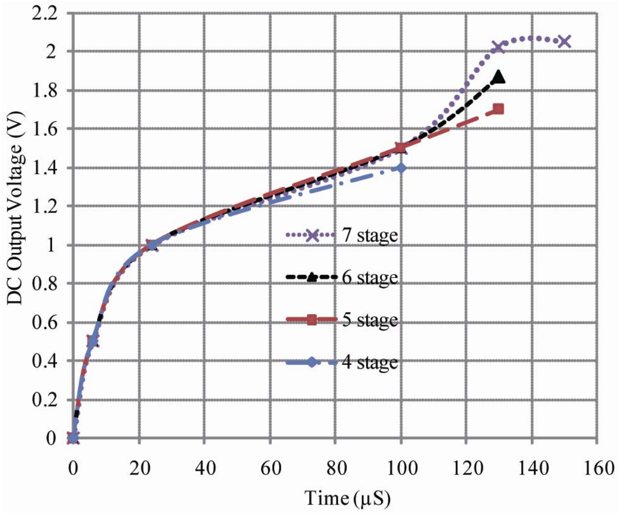

Figure 11. DC output voltage verses rise time of voltage doubler circuit through 4 - 7 stages with equal stage capacitance.

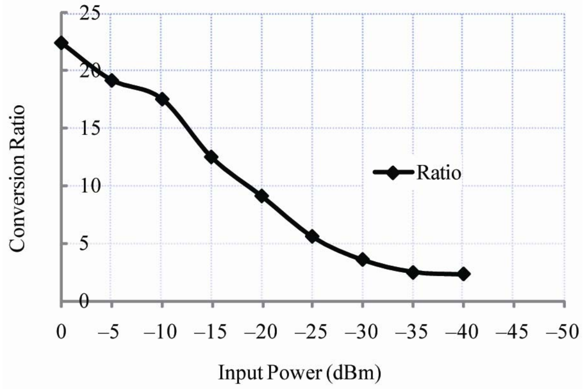

Figure 12. Conversion ratio as a function of input power.

posed voltage multiplier circuit operates at the frequency of 945 MHz with the specified input power levels. The results have shown that there is multiplication of the input voltage. From **Figure 12**, it is shown that at 0 dBm input power, the multiplication factor is 22. This is significant, as the work shows that RF energy in the GSM-900 band can be harvested from the ambient RF source using the Villard circuit topology. The power density levels from a GSM base station is expected from 0.1 mW/m^2 to 1 mW/m^2 for a distance ranging from 25 m - 100 m. These power levels may be elevated by a factor between one and three for the GSM-900 downlink frequency bands depending on the traffic density [10]. The next phase of the research work is to interface the voltage multiplier circuit through a matching network to the antenna at the input side and a low power device to power from the system at the output side to complete the RF energy harvesting system.

6. Acknowledgements

We would like to acknowledge and thank the Ministry of Higher Education Malaysia for funding this project under Fundamental Research Grant: FRGS/1/10/TK/UNITEN/02/13.

REFERENCES

[1] S. von der Mark and G. Boeck, "Ultra Low Power Wakeup Detector for Sensor Networks," *SBMO/IEEE MTT-S International Microwave & Optoelectronics Conference* (*IMOC* 2007), Berlin, 29 October-1 November 2007, pp. 865-868. doi:10.1109/IMOC.2007.4404394

[2] H. Yan, J. G. M. Montero, A. Akhnoukh, L. C. N. de Vreede and J. N. Burghartz, "An Integration Scheme for RF Power Harvesting," *Proceedings of STW Annual Workshop on Semiconductor Advances for Future Electronics and Sensors*, Veldhoven, November 2005, pp. 64-66. doi:10.1.1.126.5786

[3] T. Ungan and L. M. Reindl, "Concept for Harvesting Low Ambient RF-Sources for Microsystems," 2007. http://www.imtek.de/content/pdf/public/2007/powermems_2007_paper_ungan.pdf

[4] J. Wang, L. Dong and Y. Z. Fu, "Modeling of UHF Voltage Multiplier for Radio-Triggered Wake-Up Circuits," *International Journal of Circuit Theory and Application*, Vol. 39, No. 11, 2010, pp. 1189-1197. doi:10.1002/cta.692

[5] J. A. Starzyk, Y.-W. Jan and F. J. Qiu, "A DC-DC Charge Pump Design Based on Voltage Doublers," *IEEE Transactions on Circuits and Systems-I*: *Fundamental Theory and Applications*, Vol. 48, No. 3, 2001, pp. 350-359.

[6] HSMS-2850, "Surface Mount Zero Bias Schottky Detector Diodes." http://www.crystal-radio.eu/hsms285xdata.pdf

[7] D. W. Harrist, "Wireless Battery Charging System Using Radio Frequency Energy Harvesting," M.S. Thesis, University of Pittsburgh, Pittsburgh, 2004.

[8] E. Bergeret, J. Gaubert, P. Pannie and J. M. Gaultierr, "Modeling and Design of CMOS UHF Voltage Multiplier for RFID in an EEPROM Compatible Process," *IEEE Transactions on Circuits and Systems-I*, Vol. 54, No. 10, 2007, pp. 833-837. doi:10.1109/TCSII.2007.902222

[9] B. Emmanuel, J. Gaubert, P. Pannier and J. M. Gaultier, "Conception of UHF Voltage Multiplier for RFID Circui," *IEEE North-East Workshop on Circuits and Systems*, Gatineau, June 2006, pp. 217-220. doi:10.1109/NEWCAS.2006.250961

[10] H. J. Visser, A. C. F. Reniers and J. A. C. Theeuwes, "Ambient RF Energy Scavenging: GSM and WLAN Power Density Measurements," *Proceedings of the* 38*th European Microwave Conference*, Eindhoven, 27-31 October 2008, pp. 721-724. doi:10.1109/EUMC.2008.4751554

Copyright © 2012 SciRes.

Homebrew at ACØC
Combo Analog+DSP Filter
A Few Good CW Keyers
Preamps
RF Current Meter
100W Speaker Amp
The Ultimate Audio Filter
2xSi570 RF Gen
SB200 Sleeper
SB-200 Tank Mods
High Perf Xtal Osc
CW-Skimmer Array
Leakage Tester
XR2206 Sig Gen
Fixed Bench PS
Dual Variable Bench PS
Station Interface
Speaker/Paddle Switching

RF Current Meter

Stop Guessing and Built This Simple Meter Making Measurements Easy to Take

Why a Hand Held RF Current Meter

0
20
40
60
80
100
μA
SD-670
CLASS 2.0

There are two things that I wanted to do which would be made easy with a current meter... Both are related to the unusual nature of the antenna here at AC0C.

1. Checking common mode surface current on coax and control cables up in the attic antennas.

With the high cost of ferrites, placing them with the aid of a current meter is very helpful in that you can compare the before (no ferrite) and after (with ferrites attached) - and know that you have addressed the current problem on a given bit of line.

Otherwise, it seemed to me, that ferrite placement was more of a guess than a science.

2. Measuring relative antenna currents along an element

With the W8WWV RVM system, I can measure element currents where ever a current sensor is placed. This is typically at the center of a dipole element.

However, the center may or may not be the actual high-current point for the element depending on the frequency. My EZNEC models can indicate where the current peaks are - and by looking along these lines manually - comparing them with readings taken from the driven element, I can make current checks even in non-standard location of the array.

The meter described below accomplishes these two tasks very nicely.

Design Idea

The design follows an old RSGB design that I found on G3SEK's excellent web site:

http://www.ifwtech.co.uk/g3sek/clamp-on/clamp-on.htm

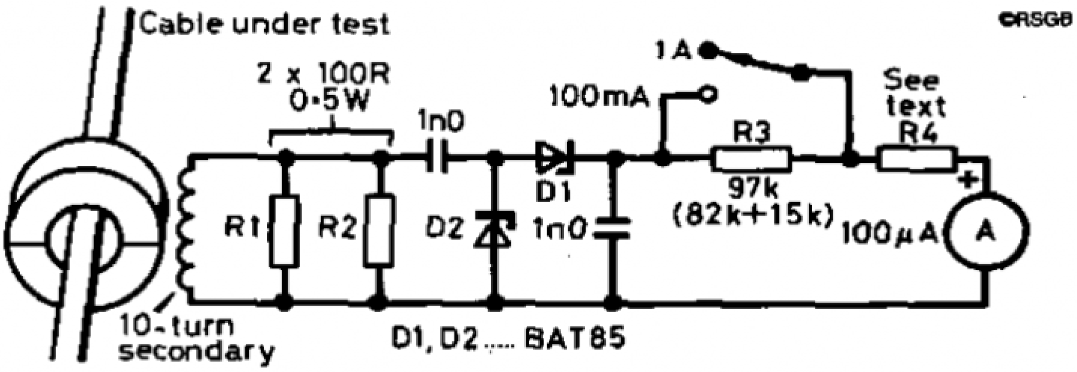

I followed this design pretty close except for the R3/R4 which I used variable trim pots. Due to the voltage-doubler nature of the circuit, it's capable of exceptional sensitivity.

The meter was calibrated for full-scale @ 1A on the higher scale, but I have left the lower scale set on a very high sensitivity setting which I find very handy for the common mode checking. Here, a sensitive meter providing a relative current indication is needed.

Construction Details

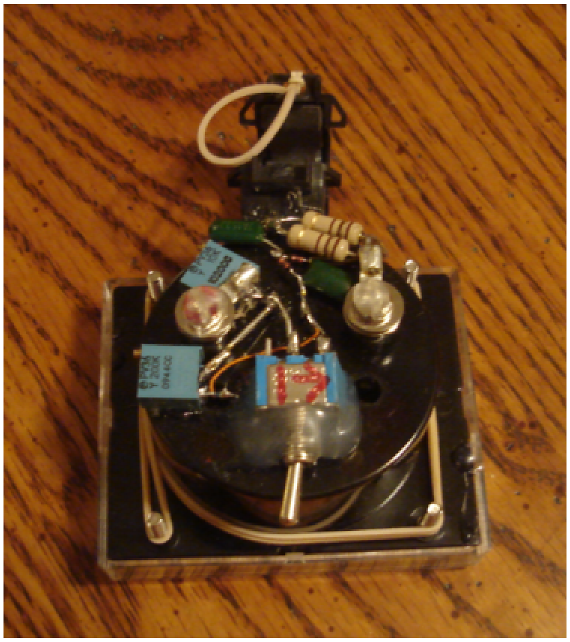

Once parts were in hand, construction was very easy and took perhaps an hour.

No attempt was made to beautify the work - it's 100% orientated toward functionality and utility.

I had a 100 uA meter in the junk box along with the other parts needed and orienting them on the back-side of the meter was really the most time consuming aspect.

The switch is held in position with hot-glue. And a bit of hot glue is applied to the meter terminations as a safety precaution - just in case I were to brush the meter up against something carrying high levels of current.

Wraps around the toroid were made with wire-wrap type wire, 30 Gage.

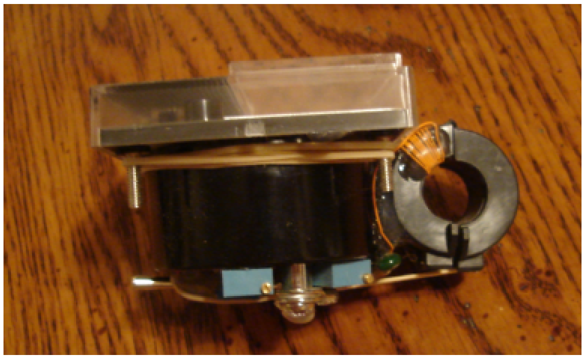

Selection of the toroid used was based on what would physically fit around the RG-213 sized cable and could be easily opened and closed.

The rubber band shown in the picture makes for a very convenient closing mechanism - and in a lot of applications, I simply hold the toroid closed with my finger pressure which makes moving the meter along a wire - and moving from cable to cable - very quick and easy.

For more ideas, check out Frank N4SPP's very nice RF current meter found HERE.

© 2011-2016 AC0C - All Rights Reserved

Created with the QTH.com SiteBuilder.

Bread board Arrangement:

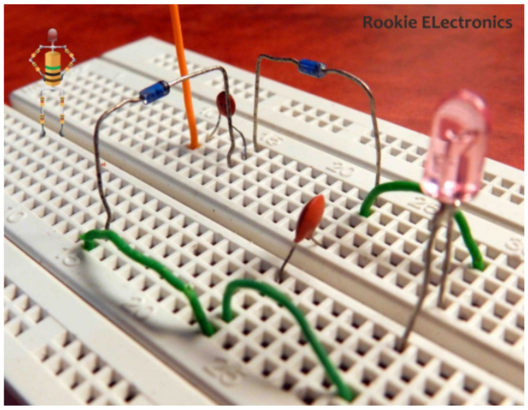
Rookie ELectronics

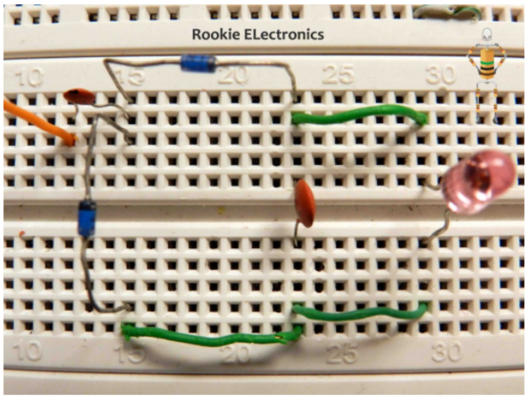
Rookie ELectronics

Strip Board:

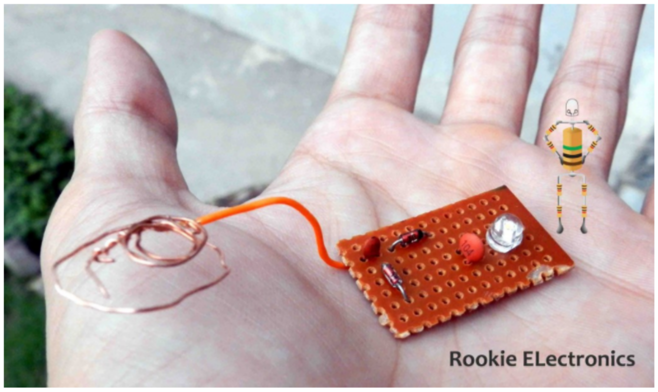
Rookie ELectronics

RF Diagnostics, LLC

WELCOME TO RF DIAGNOSTICS

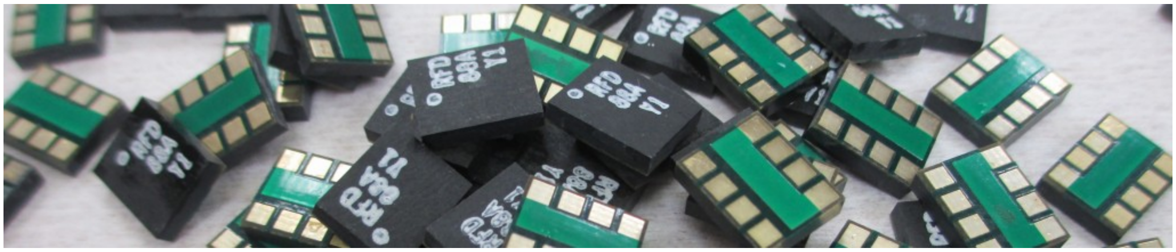

Energy Harvesting Modules

WIRELESS ENERGY HARVESTING PRODUCTS RF Diagnostics, LLC designs and sells surface mount RF to DC converters and other building block modules for energy harvesting systems. With our products you can harvest ambient AC/AM/FM/Wireless Energy to power your IoT battery free system. We also provide custom designs for RF/microwave and general electrical engineering products and design contracting services to clients in the commercial wireless and defense electronics businesses. We strive to be a high quality and cost effective design resource for your design team.

OUR PRODUCTS

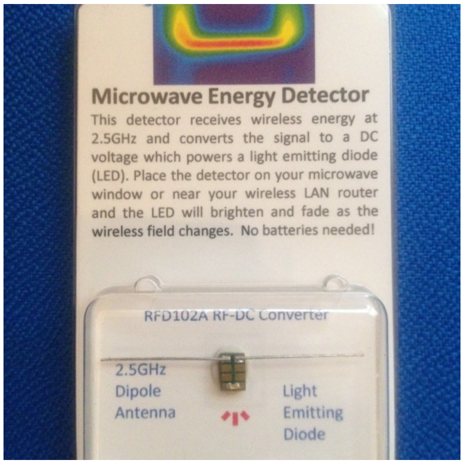
Microwave Energy Detector
This detector receives wireless energy at 2.5GHz and converts the signal to a DC voltage which powers a light emitting diode (LED). Place the detector on your microwave window or near your wireless LAN router and the LED will brighten and fade as the wireless field changes. No batteries needed!
RFD102A RF-DC Converter
2.5GHz Dipole Antenna
Light Emitting Diode

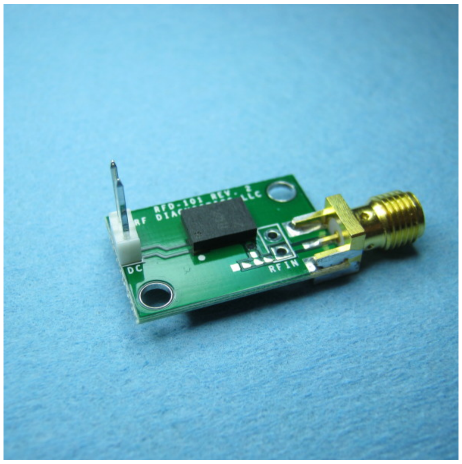
RFD-101 REV. 2
RF DIAGNOSTICS, LLC
RF
DC
RF IN

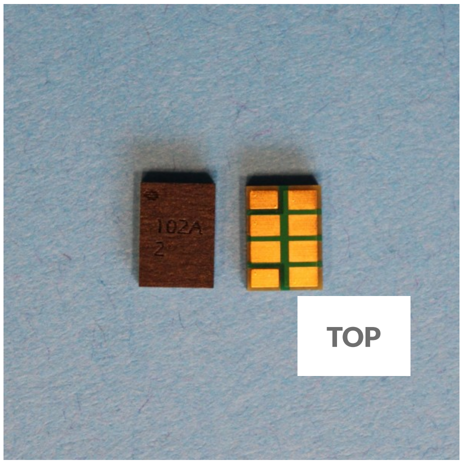

TOP

RFD102A-DET: Microwave Energy Detector

$18.00

Add to cart

RFD102A-TB: 60Hz…6GHz Energy Harvesting Test Board

$33.00

Add to cart

RFD102A: Wireless Energy Harvesting Module

$12.00

Add to cart

OUR DESIGN SERVICES

Energy Harvesting Module & System Design

Any of our standard products can be customized for your application. We can support your application with matching, testing, antenna design, assembly and many other tasks. Have a new idea for a product? We can help you realize it. We are developing new products and are always interested in your feedback to guide our product development.

MMIC Design

We use Agilent's Advanced Design System for MMIC design and PCB design. Previous designs completed are quad-band GSM power amplifier modules, 2.4-2.5GHz/5-6GHz wireless LAN power amplifiers, broadband power amplifiers, switch-filter designs, RFID detectors, RF/DC converters and filter designs.

Front End Module Design

We use Agilent's Advanced Design System and Orcad PCB Editor for Front-End Module designs. Previous design examples are high-volume switch filter modules, dual-band 802.11a/b/g front end modules, 76GHz car radar module, RFID card design.

RFID Detector System

On an NSF Phase I & II grant to one of our clients, operating as a subcontractor, the RF Diagnostics design team created a complete RFID MEMS resonator detection system. Both hardware and software were developed under this contract. The work was published at an SPIE conference in July 2010.

LEARN MORE

NISKAYUNA, NY 12309 USA
518-810-6560
INFO@RFDIAGNOSTICS.COM

TOP

RF DIAGNOSTICS, LLC COPYRIGHT © 2017.

Sebs World

- Home
- DA14 Asteriod
- Secret space plane
- Solar storm
- End of the world?
- UK monitoring
- Tracking
- Save money
- Get out of the EU
- UFO's
- Magnetic energy
- RF energy
- Radiant energy
- Perpetual
- Easy Steam!
- Irony- defined
- Funny Pics
- Internet Problems
- Star Wars
- Chat room
- Live music
- Kevzone
- Facebook Cheats
- Crashes

About us
Private Policy

RF energy..

This can extract some electricity from the air!

FREE ENERGY FROM AIR CIRCUIT PART 8/24

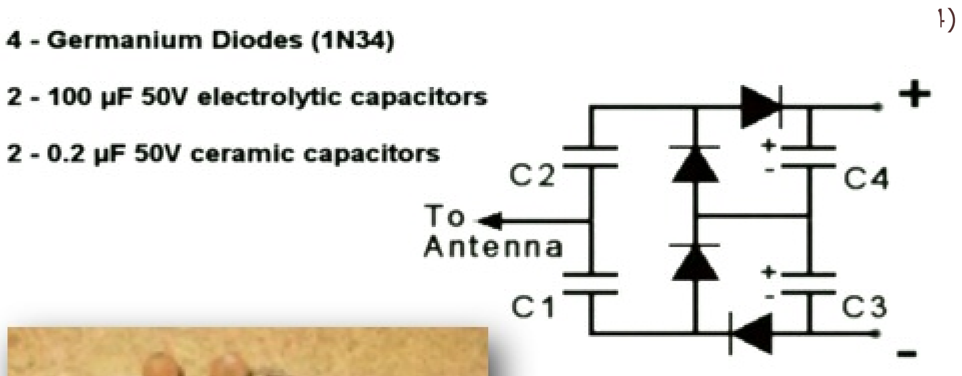

4 - Germanium Diodes (1N34)

2 - 100 µF 50V electrolytic capacitors

2 - 0.2 µF 50V ceramic capacitors

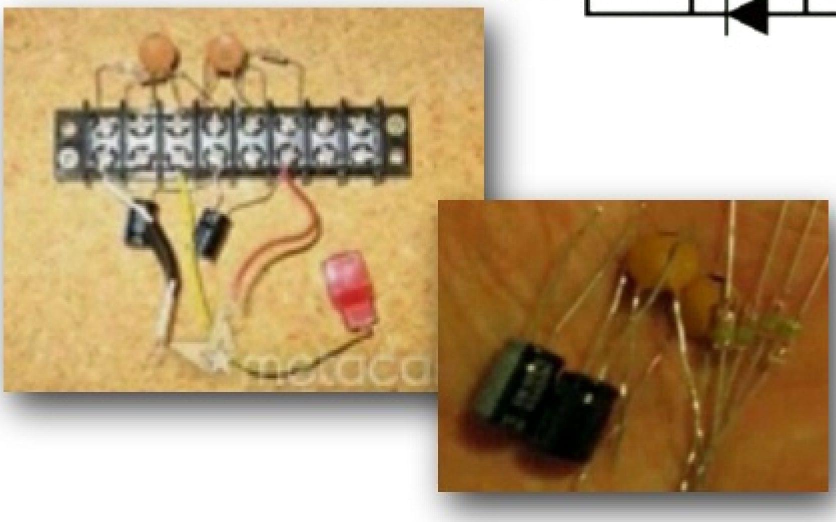

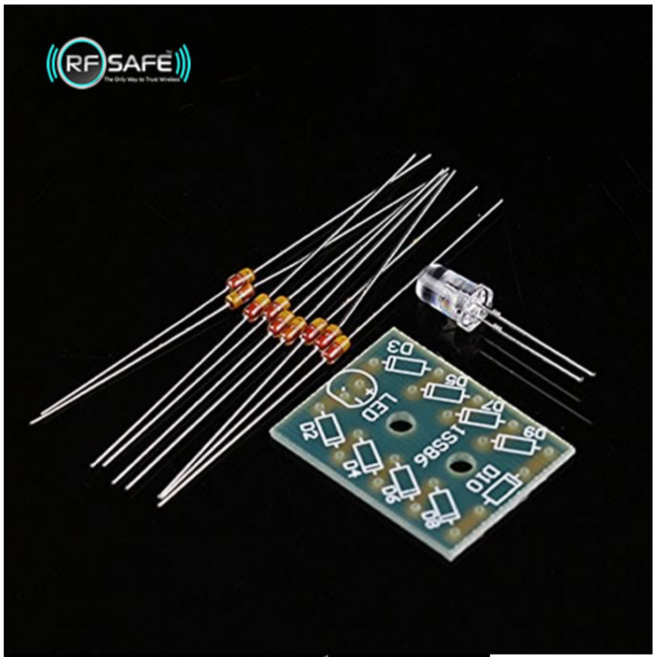

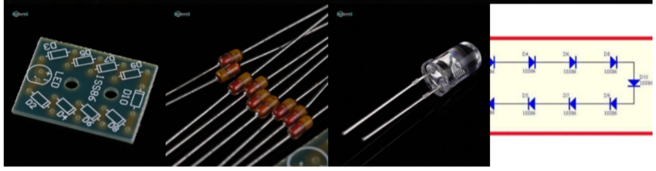

DIY RF METER: LED IS 100% POWERED BY RF RADIATION USING DETECTOR DIODES

$8.99

1 Add to cart

SAFE KIDS: LED Cell Phone Radiation Detector For DIY Science Project & Experiments || Test Phone Shields

NO batteries. it uses ONLY RF radiation power from cell phone. Yes, see with your own eyes how powerful cell phone radiation is when it powers a light from thin air. RF Safe's DIY circuit is the simplest detector for microwave radiation. The RF diodes convert microwaves from a cell phone to DC current to power an LED which emits visible light warning you of possible danger.

SKU: 636684233801 Categories: Cell Phone Radiation Safety, D-I-Y Tags: 1SS86 detector diode, Cell Phone Radiation Meter, germanium diode, LED RF Meter

A concurrent 915/2440 MHz RF energy harvester

L Fadel, L Oyhenart, R Bergès, V Vigneras, T Taris

To cite this version:

L Fadel, L Oyhenart, R Bergès, V Vigneras, T Taris. A concurrent 915/2440 MHz RF energy harvester. International Journal of Microwave and Wireless Technologies, Cambridge University Press/European Microwave Association 2016, 8, pp.405 - 413. <10.1017/S1759078716000179>. <hal-01474918>

HAL Id: hal-01474918
https://hal.archives-ouvertes.fr/hal-01474918

Submitted on 7 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A concurrent 915 MHz / 2440 MHz RF Energy Harvester

L. Fadel[1], L. Oyhenart[1], R. Bergès[1], V. Vigneras[1], T.Taris[1]

[1] Laboratoire IMS, UMR - 5218, Université de Bordeaux, 33405 Talence Cedex, France.

This paper presents the development of two dual-band RF harvesters optimized to convert far-field RF energy to DC voltage at very low received power. The first one is based on a patch antenna and the second on a dipole antenna. They are both implemented on a standard FR4 substrate with commercially off-the-shelf (COTS) devices. The two RF harvesters provide a rectified voltage of 1V for a combined power respectively of -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz and of -20 dBm at 915 MHz and -15 dBm at 2.44 GHz. The remote powering of a clock consuming 1V/5µA is demonstrated, and the rectenna yields a power efficiency of 12 %.

Keywords: Antennas and Propagation for Wireless Systems, Applications and Standards (mobile, Wireless, networks), Circuit Design and Applications.

Corresponding author: email: ludivine.fadel@ims-bordeaux.fr; phone: +33 540 002 615

I. INTRODUCTION

Today's the society is evolving toward creating smart environments where a multitude of sensors and devices are interacting to deliver an abundance of useful information. Essential to the implementation of this Internet Of Things (IOT) is the design of energy efficient solutions aiming toward a low-carbon-emission, namely green, society. Within this context, the energy harvesting appears as an alternative to provide systems with self-sustained operation. Many electronic devices operate in conditions where it is costly, inconvenient, or impossible to replace the battery. Examples include sensors for health monitoring of patients [1],[2], aircraft or building structural monitoring [3],[4], sensors in natural, industrial or hazardous environments, etc. The scavenging of natural ambient energy requires some specific conditions such as: daylight for solar energy [5], breeze for wind energy or motion for kinetic energy [6] to name a few. As consequences the exploitation of natural source does not fit with many cases of applications. On the other hand the electromagnetic (EM) [7], or Radio-Frequency (RF), energy is a human made source that is not dependent of weather conditions nor the daytime. It is so very attractive for wireless powering of remote devices. Furthermore the ever growing of commercial and personal wireless installations opens up to a 24 hour a day available energy in the vicinity of any human activity areas. The schematic of a general Wireless RF Power Transmission (WPT) system is shown in Fig.1. We talk here about far-field RF energy transmission [8], which is different from near-field RF energy transmission [9]. This later including inductive, capacitive or resonant coupling is a close contact transmission and is not relevant for remote devices. In Fig. 1, the receiver antenna collects the EM energy radiated by a RF source, and converts it into a RF signal. This RF signal is transferred to the rectifier by an impedance matching network, to be converted into DC power, which is further accumulated in a storage element. The main purpose in the deployment of WPT systems is the development of compact and efficient solutions. Most of the challenge concerns the implementation of harvesting modules, especially the antenna as its design defines the scavenging capability and the size of the RF harvester. At low frequency the transfer of energy is efficient, but the antenna footprint is large. To address the trade off between the efficiency of the WPT and the size of the modules, the frequency band located in the 433 MHz to 6 GHz frequency spectrums are preferred.

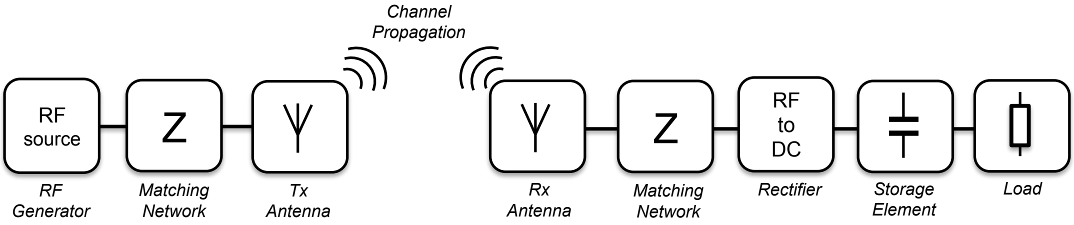

Fig. 1: Schematic of a Wireless RF Power Transmission (WPT) system

Over the last decade the research effort has focused on the development of WPT systems according two scenarios: the RF energy scavenging [10] and the RF energy transfer [11]. The two RF energy scavenging is an opportunistic collection of the RF ambient energy from the surrounding communication traffic. To improve the harvesting capability the scavenging RF harvesters are of wide-band type [12] and cover popular standards such as: DTV (470-610 MHz), GSM900, GSM1800, 3G (2.1GHz) and WiFi (2.4 GHz). Unfortunately these standards dedicated to convey wireless communications do not radiate a large RF power. As consequences the collected energy is weak, unpredictable and out of control. The RF energy scavenging remains a promising solution in the future as the increase of communication traffic could make it more reliable, and consistent with IOT applications. The second concept, namely RF energy transfer, assumes an identified source that is dedicated to perform the WPT. The amount of transmitted power is controlled by the source and the collected energy is larger than in scavenging approach. The licence-free Industry-Science-Medical (ISM) frequency bands located at 0.9, 2.4 and 5.8 GHz are usually exploited to support such a WPT scenario. Today the RF energy transfer in ISM Bands is not only promising, it becomes a reality as some pioneer companies propose some full kits: Powercast Corporation, AnSem and MicroChip to name a few. However there is still a lot of work to make the RF energy transfer an appropriate, low cost and easy-to-use solution for remote powering. One of the most critical point concerns the harvesting capability of the RF modules. So far the commercial kits referenced above only explore the 900 MHz ISM allocations to perform the WPT. This work proposes to demonstrate the interest of a concurrent harvesting at 915 MHz and 2.44 GHz. The design and implementation of a modified 4-stage doubler RF to DC converter, including a concurrent matching network, is first presented. The section III details the design of two types of multi-band antenna. The comparison between a single frequency and a multi-band WPT is exposed and the demonstration of the remote powering of a clock is reported as a case of application. To conclude a comparison of our results with the state of the art is exposed.

II. CONCURRENT RF TO DC CONVERTER

The Radio-Frequency IDentification (RFID) applications are the most popular systems exploiting the principle of RF energy transport. In passive RFID applications the reader transmits the RF power to the tag, and also sets up the communication. The RF to DC converter is designed to yield a maximum of power efficiency to the tag. Most of the time the reader and the tag are in line of sight and close to each other, these conditions improve the transmission of RF energy, the amount of power available at the tag antenna is large, typically between -15 dBm and -20 dBm. In RF energy harvesting the scenario is different. The distance between the RF source and the RF harvester ranges from 0.5 meter to 10 meters. The amount of collectable power is low, from -10 dBm to -25 dBm, and the remote powering is difficult. The RF harvesters are supposed to collect and to store the energy during a long period of time. Once the level of stored energy is large enough, it can be released to the application. For these reasons a rectifier dedicated to RF energy harvesting is first designed to yield a maximum of sensitive to increase its scavenging time and capability.

A) Rectifier Architecture

The rectifier architecture is based on voltage multipliers to provide an adequate output DC voltage. The architecture of the RF to DC converter, reported in Fig.2, includes a matching network based on a L-section, and a N-stage voltage multiplier based on Schottky diodes from Avago (HSMS285). The choice of the Schottky diode is very important in the design of the rectifier. A key parameter is its threshold voltage $V_{TH.}$ When only low power levels are available in the environment, the amplitude of the incident signal may be close to or even below this voltage. Below this voltage value, the diode will no longer conduct and the losses become predominant. For COTS devices the two Schottky diodes performing the best conversion efficiency in a 2.4 GHz range are HSMS-2850 from Avago and SMS-7630 from Skyworks [13].

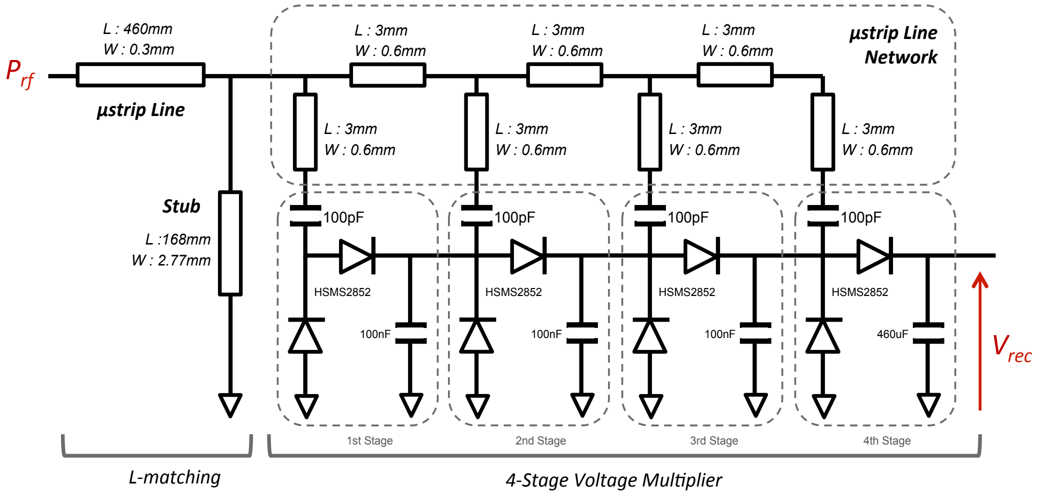

Fig. 2: Architecture of the RF to DC converter

Focusing on the sensitivity, the RF to DC converter is designed to maximize the rectified voltage for an input power close to -20 dBm. The optimum number of stage is fixed to four according [14]. The footprint of each voltage doubler imposes the micro-strip line network. The micro-strip lines namely "junction" is set to minimum length, the micro-strip lines "access" are used as an additional degree to tune the input-matching network. Indeed, the L section in combination with the micro-strip distributed network is equivalent to a T section (Fig.3). Many combinations of Z_1, Z_2, Z_3 can achieve input matching at 900 MHz or 2.4 GHz. Some of them are very close for each frequency, so we choose one that allows a return loss (< - 10 dB at least) both at 900 MHz and 2.4 GHz.

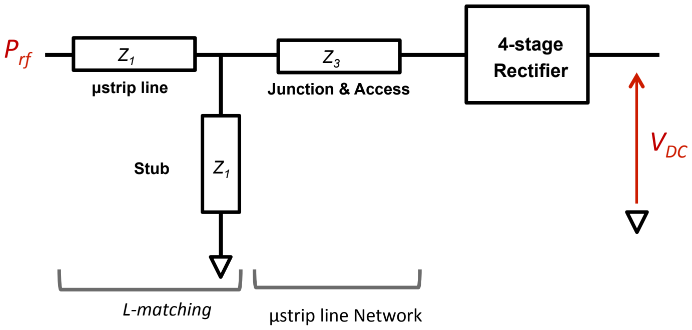

P_{rf}
Z_1
µstrip line
Z_3
Junction & Access
4-stage Rectifier
Stub
Z_1
V_{DC}
L-matching
µstrip line Network

Fig. 3: Topology of the input-matching network

The equivalent narrow band model of the matching network is proposed for each frequency (Fig.4). At 915 MHz, the voltage multiplier, including the rectification stages and the micro-strip line network, is modelled with a shunt capacitor (5pF) and a shunt resistor of 270 Ω (Fig.4a). The stub, (Fig.3) is equivalent to an inductor (Fig.4a), which compensates the shunt capacitor. The input micro-strip line, (Fig.3), is a quarter wave impedance transformer, (Fig.4a) it converts the 270 Ω into 50 Ω. At 2.44 GHz the micro-strip line network distributing the RF signal to the voltage doublers (Fig.3), becomes inductive (Fig.4b). The stub is equivalent to a shunt capacitor of 120fF, its effect is negligible. The impedance transformation is actually performed by the input micro-strip line, which is modelled by a shunt capacitor (0,6 pF) and a series inductor of 5.6 nH.

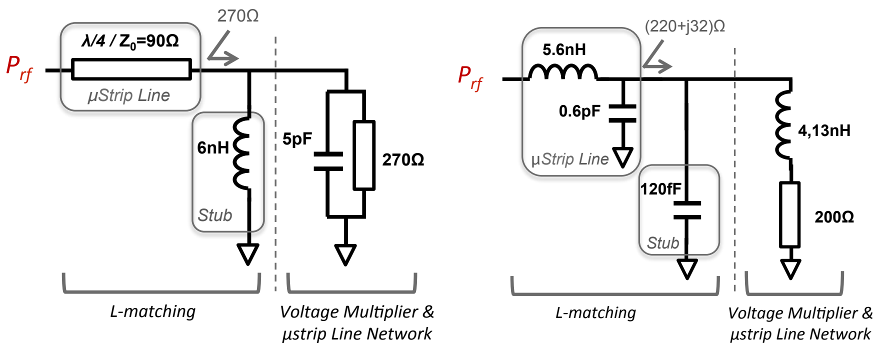

P_{rf}
$\lambda/4$ / Z_0=90Ω
µStrip Line
270Ω
6nH
Stub
5pF
270Ω
L-matching
Voltage Multiplier & µstrip Line Network
P_{rf}
5.6nH
0.6pF
µStrip Line
(220+j32)Ω
120fF
Stub
4,13nH
200Ω
L-matching
Voltage Multiplier & µstrip Line Network

(a) *(b)*

Fig. 4: Equivalent model of the input matching network at 915 MHz (a) at 2,44 GHz (b)

To study the impact of the power on the diode, and input matching, behaviour the return loss of the 4-stage rectifier has been measured and plotted (Fig.5) for various input power P_{rf} at 9000MHz.

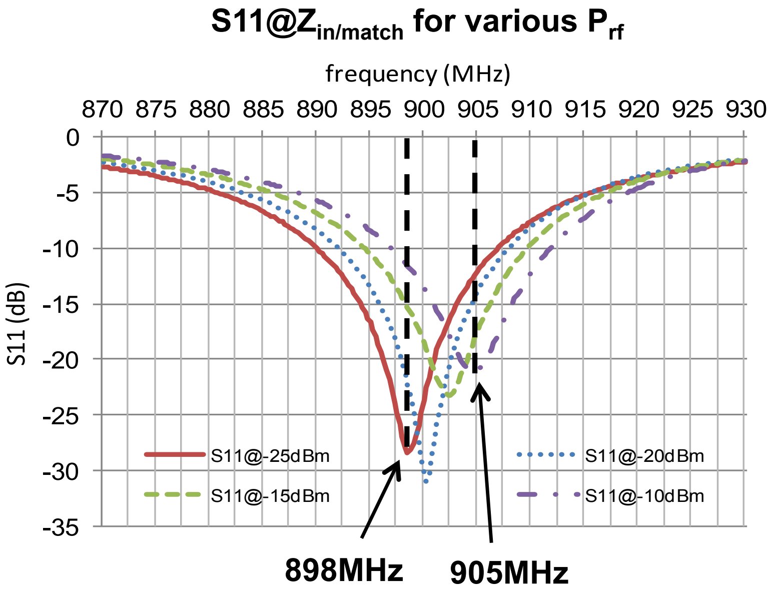

Fig. 5: Measured S_{11} of a 4-stage voltage multiplier with a L section at 900 MHz for various input power P_{rf}

As illustrated in the figure 5, the input return loss is not strongly affected by the input power if $P_{rf} < -15$ dBm. The RF harvesters developed in this work are dedicated to collect power from – 15 to – 25 dBm. Over this range the diode model can be considered as stable, and the slight frequency shift is still covered by the antenna bandwidth.

B) Rectifier Characterization

The power efficiency and the power sensitivity are two conversion characteristics of importance in RF harvesters. However, the RF harvester operating at low power level accumulates the energy in a storage element, to further release it to the application. In such accumulation mode the power sensitivity becomes more important than the power efficiency.

For the characterization the rectifier is not connected to a load. The load represents the equivalent impedance of the application (clock, sensor) to power. The effectiveness of RF-DC conversion of the rectenna and its DC ouput voltage varies depending on the load value. The rectifier is first characterized in a single tone mode, 915 MHz and 2.44 GHz respectively, and then in a dual-band mode. Measurements of the unloaded rectified voltage versus various input power P_{rf} are reported in Figure 6.

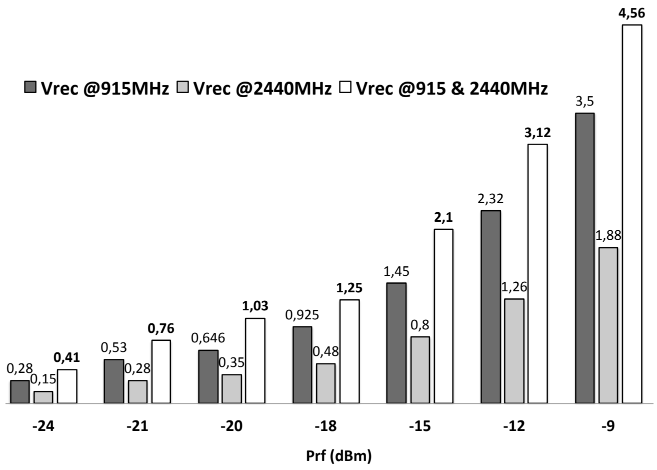

Fig. 6: Unloaded rectified voltage for various input power

To compare the results of the two considered tones, the target is fixed to a value of 1V. In a single tone mode, the required P_{rf} to rectify 1V is close to -18 dBm at 915 MHz, and would be larger than -15 dBm at 2.44 GHz. In a dual-band mode the circuit only needs a power P_{rf} of -20 dBm at each frequency. The dual-band rectification significantly improves the power sensitivity. The reverse breakdown voltage of the HSMS285 Schottky diode limits the input power to - 9 dBm, for which V_{rec} is 4,56 V.

III. ANTENNA DESIGN

To meet the low-cost constraints, the RF energy harvester will be implemented on a single low cost substrate, an FR4 PCB. For the antenna, there are more efficient substrates with a higher permittivity to reduce the size of the antenna or a lower losses but their cost is much higher than the improved performance. These powerful substrates fail to build low cost energy harvesters. This section exposes the design of two dual-band antennas implemented on a 1.6 mm FR4 printed circuit board. The fabrication uses a mechanical etching process with a 200µm resolution. We have chosen two complementary antenna topologies with a directional and omnidirectional radiation pattern.

A) Dual - band patch antenna

Emitting and receiving antennas do not usually meet the same constraints. Mobile devices such as smartphones and tablets use compact antennas (ifa, pifa…) to address the trade off between performance and size. Base stations can afford large efficient radiating elements (omnidirectional or directional antennas depending on the application). For energy harvesting purpose, micro-strip patch antennas are commonly used [15-17]. A rectangular micro-strip patch antenna (RMPA) is first developed to suit with both low cost technology of implementation and co-integration with the rectifier. Based on the cavity-model approximation, the resonant frequencies of the RMPA for the TM_{mn} mode is described in (1).

$$f_{mn} = \frac{c}{2\sqrt{\varepsilon_r}}\sqrt{\left(\frac{m}{L}\right)^2 + \left(\frac{n}{W}\right)^2} \qquad (1)$$

W, L are the patch dimensions, $c = 3.10^8$m/s

The antenna dimensions, 68 cm^2 (8.8 x 7.8 cm), described Fig.7a are dependent to the frequency bands and the feed location is selected to only excite the fundamental modes TM_{01} and TM_{10}. Those modes permit to obtain a large aspect ratio (W/L = 2,7) but reduce the performance of the RMPA. On the other hand, TM_{01} and TM_{30} modes require an aspect ratio close to one but offer beneficial radiation patterns for our application. The RMPA is fed by a probe whose position (x,y) adjusts the matching both at 915 MHz and 2.44 GHz. This two operating bands of the proposed antenna are on cross polarization planes. The geometric parameters of RMPA have been optimized with an approximate model, the TL model [18], and with a full wave method. Details of the two approaches have been studied in [19]. The return loss of the RMPA is better at 915 MHz than 2.44 GHz because the maximum impedance of TM30 mode is 31 Ω [19]. The TM30 mode does not achieve 50 Ω because it is not a fundamental mode. This antenna has a maximum gain of 1.3 dB at 915 MHz (Fig.7b) and 2.5 dB at 2.44 GHz (Fig.7c). This two operating bands of the proposed antenna are on cross polarization planes.

The realized gains of the dual band patch antenna are lower than the classical patch antenna because the radiating efficiency is low, 60% for the TM_{01} mode and 30% for the TM_{30} mode. The FR4 substrate has a loss tangent of 0.02. The radiation efficiency of the dual band patch antenna is highly dependent of the substrate losses.

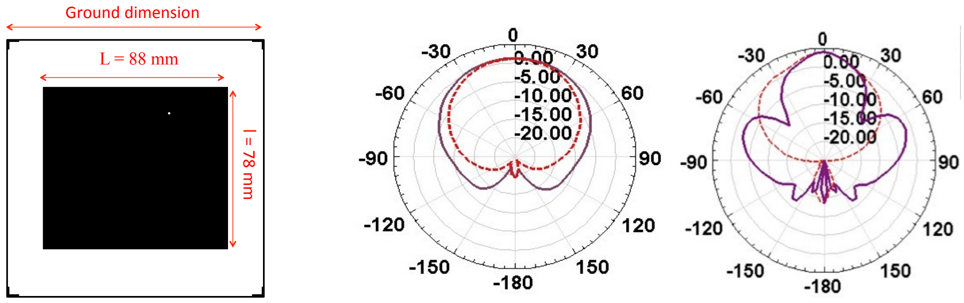

(a) *(b)* *(c)*

Fig. 7. Layout of the RMPA antenna (a) - Radiation pattern at 915MHz (b) and 2.44 GHz (c) Solid and dashed lines correspond to E-plane and H-plane respectively

B) Multi-band arm dipole antenna

The second antenna is a multi-band dipole type composed of three arms. Its dimension is about 23 cm^2 (11.1x2.1 cm), Fig.8a. Each arm is designed to work at one band of frequency. The longer one is for the 915 MHz, the middle one, not useful in our case, is for the 1.4 GHz and the last one, the smaller, is dedicated to 2.4 GHz [20].

All the geometric parameters have been optimized with a full-wave method in order to be matched both at 915/ 2440MHz. On Fig.6b and 6c, the radiation pattern is plotted for the elevation plane (orthogonal to substrate) of the simulated antenna. The maximum gain is 0.5 dB at 915 MHz and 3.4 dB at 2.44 GHz at 90°, on the substrate plane. The radiation efficiency is 99% at 915 MHz and 95% at 2.44 GHz.

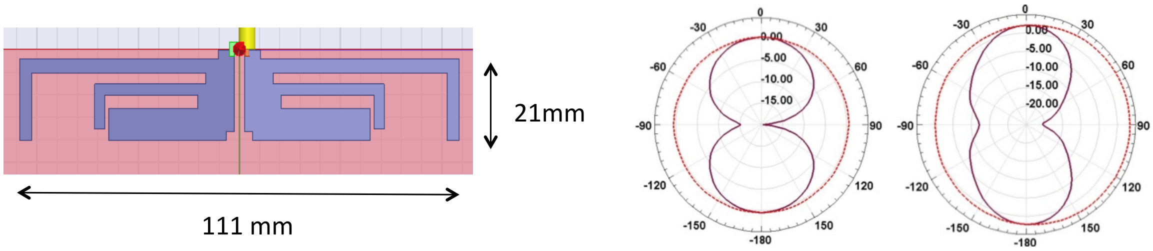

(a) *(b)* *(c)*

Fig. 8. Layout of the multi-band arms dipole antenna (a) - Radiation pattern at 915 MHz (b) and 2.44 GHz (c). Solid and dashed lines correspond to E-plane and H-plane respectively

It is interesting to compare the characteristics and performances of the two types of antennas. Although the radiation efficiency of the dipole antenna is better than the patch antenna, the antenna gains are similar because the high directivity of the RMPA antenna compensates the low values of radiation efficiency. When there are no cost constraints, it is interesting to use high performance substrates for the design of RMPA antennas because they improve the radiation efficiency and consequently antenna gain.

Moreover, the integration of the antenna with the rectifier will not be made in the same way. Considering the patch antenna, the rectifier can be integrated on the ground plane allowing a more compact solution. The dipole antenna, which is ground plane free, is less sensitive to the surrounding environment in our case. The performance of the dipole antenna and especially the radiation efficiency are very weakly dependent of the substrate characteristics. The design of a dipole antenna can be easily reuse with other material such as Kapton®, paper, Plexiglas to name a few.

IV. WIRELESS POWER TRANSMISSION

This part presents the measurement results of the assembled RF harvesters in the context of Wireless Power transfer. The two dual-band harvesters are realized with COTS devices such as HSMS diodes and capacitors. Those elements are reported by heat-treating. The RF to DC converter board, including the matching network and the rectifier, is reported on the backside and connected to the radiation part, on the front side, through a via (Fig.9a). The dipole antenna is connected to the rectifier circuit using SMA connector (Fig.9b).

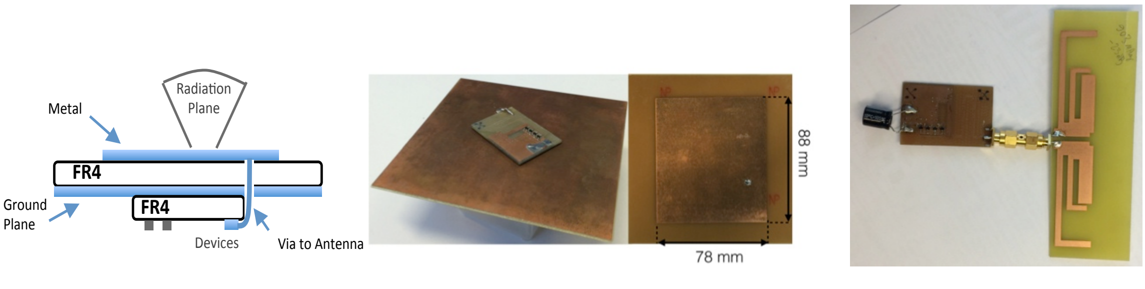

(a) *(b)*

Fig. 9. Dual-band RF harvesters based on patch antenna (a) – arms dipole antenna (b)

For the dual-band RF harvester based on patch antenna, the return loss, S_{11}, is measured for an input power of -20 dBm with a HP8720 network analyser. The patch antenna, the rectifier and the dipole antenna are centered at 915 MHz and 2.44 GHz with a low return loss ($S_{11} < -15$ dB), Fig.10. The return loss of the RMPA antenna is better at 915 MHz than 2.44 GHz because the maximum impedance of TM30 mode is 31 Ω (Fig. 6 and Fig. 7 of [16]). The TM30 mode does not achieve 50 Ω because it is not a fundamental mode.

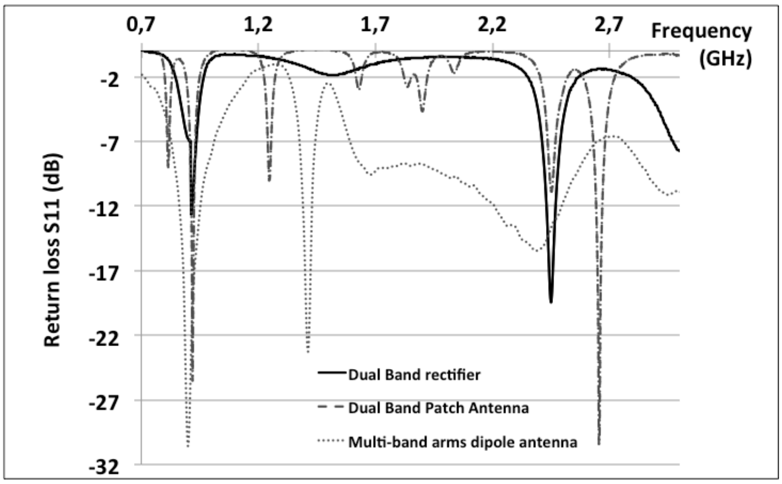

Fig. 10. Measured return loss S_{11} of the dual-band rectfier, patch and arms dipole antenna

A) Remote Powering and Power Efficiency

The rectenna is connected to a clock, which mimics a low power application. The remote powering of this clock is performed in a furnished room of the lab according the schematic of Fig.11. The distance between the source and the antenna is fixed to 2 m. The clock is turned on for different scenarios of transmitted power. For each combination of power proposed in Fig.12, the RF power is first measured with a calibrated antenna and a power meter. Then, the rectenna is measured and P_{eff} is the ration between the power delivered to the load (here the clock) and the power available at the antenna.

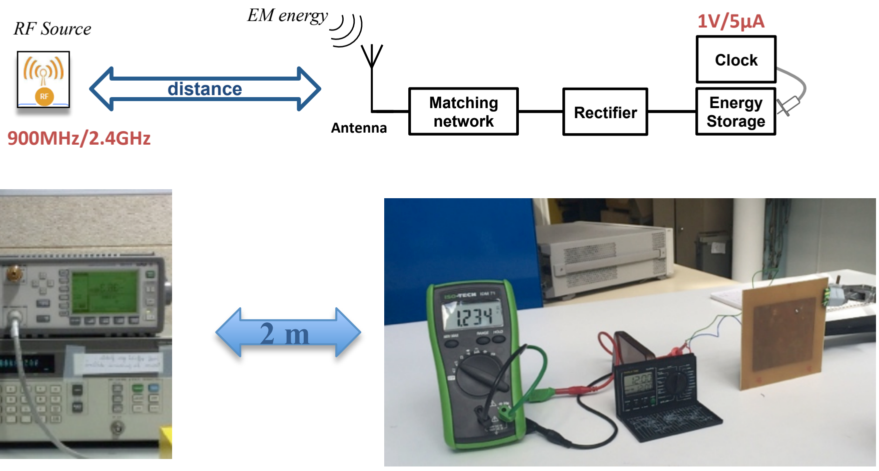

Fig. 11. Schematic and picture of the scene of remote powering of a clock

The power efficiency of the patch and the dipole rectenna is worked out from these experiments and reported in Fig.12. The power efficiency η is defined as the ratio between the DC power delivered to the clock and the RF power collected by the antenna.

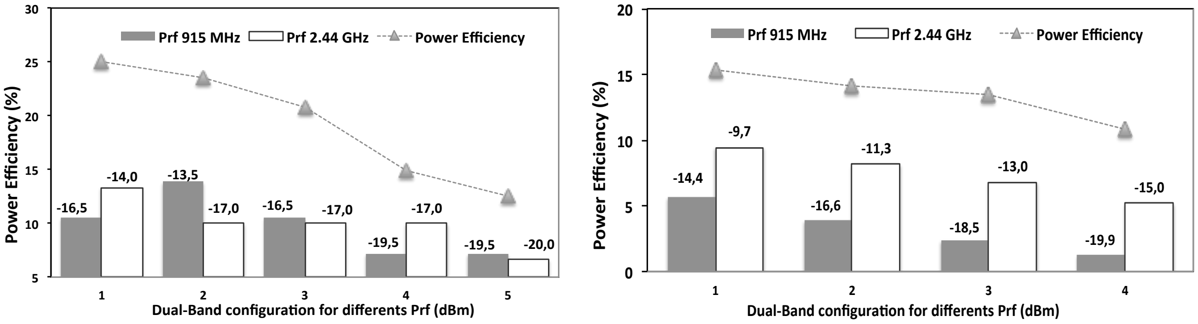

(a) *(b)*

Fig. 12. Power efficiency of the dual-band RF harvester based on the patch (a) and the arms dipole (b) antenna

The minimum power required to turn on the clock with the patch-based harvester, Fig. 10.a, is a two tone signal featuring: -19.5 dBm at 915 MHz and -20 dBm at 2.44 GHz. At this point, the power efficiency is 12.5 %, which corresponds to a DC output power of 2.7 µW/1V.

A maximum efficiency of 24 % occurs for a combined power of -16.5 dBm at 915 MHz and -14 dBm at 2.44 GHz. The harvester is able to deliver a DC ouput power of 15 µW. The harvester based on the arm dipole antenna needs a minimum power of -19.9 dBm at 915 MHz and -15 dBm at 2.44 GHz at the antenna to turn on the clock. For these conditions of remote powering, the efficiency of the harvester is 11 %. It delivers a DC power of 3.8 µW/1.15 V. The maximum power efficiency, 15.5 %, yields for an input power of -14.4 dBm at 915 MHz and -9.7 dBm at 2.44 GHz, the DC output power is 21 µW.

This scenario of remote powering figures out that the harvester based on the patch antenna exhibits a better power efficiency than the harvester combined with the dipole element. This difference is due to the antenna gains. Referring to Fig. 7 and Fig. 8, the gain of the patch antenna is larger (+0.8dB) at 915 MHz and lower (-0.9 dB) at 2.44 GHz than the dipole element. However the rectifier, referenced in [16], achieves a power efficiency of 17% at 915 MHz and only 5% at 2.44 GHz for an input signal of -15 dBm. As consequences the patch-based harvester is able to extract more power from a 915 MHz signal than the dipole-based harvester can do at 2.44 GHz. For this reason the overall efficiency of the patch harvester is better.

B) Power Sensitivity

The power sensitivity is measured with the same scenario of Fig.11 but the clock is disconnected. The output voltage is reported for different combination of collectable power at the antenna in a dual-band configuration.

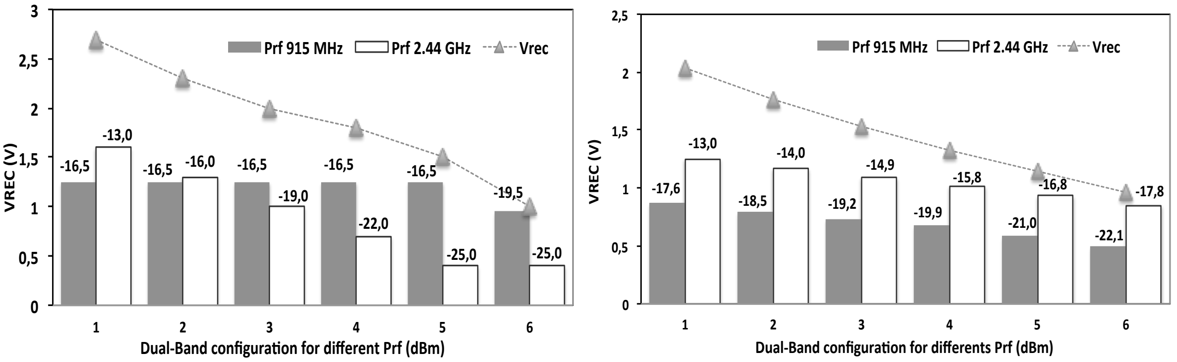

(a) *(b)*

Fig. 13. Rectified voltage of the RF harvester based on patch (a) and arm dipole (b) antenna

To rectify a 1V DC voltage, the patch-based harvester, Fig.13a, requires a two dual-band configuration: -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz, which is equivalent to an input power of -18.4 dBm (or 14 µW). For the same purpose the dipole-based harvester, Fig.13b, needs a dual-tone of -22.1 dBm at 915 MHz and -17.8 dBm at 2.44 GHz. The equivalent input power of this 2-tone signal is -16.5 dBm (or 22 µW). The patch-based harvester exhibits a better sensitivity than the dipole harvester for the same reason exposed in the part A of this section. In Fig.6, which reports the power sensitivity of the rectifier part only, the overall sensitivity is almost the same for the dipole harvester. It is improved by 1.8 dB for the patch harvester due to the additional gain of the antenna at 915 MHz.

C) Discussion and Comparison with the state of the art

An important characteristic of a remote powered device is its size. Indeed it is expected to be as small as possible to make it unobtrusive to our closest environment. In a scenario of RF harvesting the antenna footprint determines the compactness of a harvester operating Ultra High Frequency (UHF) bands. To complete the comparison between the two harvesting modules developed in this work, two figures of merit, FOM_{sens} and FOM_{eff}, including the size of the antenna, are proposed in (2) and (3).

$$FOM_{sens} = \frac{V_{REC@Psens}(V)}{\frac{P_{sens}(\mu W)}{100\mu W} \cdot \frac{A_{ant}(cm^2)}{100cm^2}} \tag{2}$$

With: P_{sens} the input RF power required to provide $V_{REC@Psens}$ the unloaded rectified output DC voltage, and A_{ant} the area of the antenna.

$$FOM_{eff} = \frac{\eta(\%)}{\frac{P_{eff}(\mu W)}{100\mu W} \cdot \frac{A_{ant}(cm^2)}{100cm^2}} \tag{3}$$

With: P_{eff} the input RF power required to achieve η the overall power efficiency.

FOM_{sens} and FOM_{eff} do not represent the same scenario of application. The FOM_{sens} illustrates the capability of the rectenna to start collecting energy and store it in an element such as a capacitor or a battery to further release it. FOM_{eff} demonstrates the capability of the RF harvester to yield "on time powering": the rectenna is connected to an application and supply it on time. Both are reported in the Table I, which also includes some references of the state of the art. The ability of the proposed rectenna to simultaneously operate in two frequency bands, significantly improves the power sensitivity.

Table I. Comparison with the state of the art

Ref.	Freq (GHz)	Efficiency (%@P_{rf})	Sensitivity (V_{rec}@Prf)	Number of stage	Schottky diodes	Size (cm^2)	FOM Sens	FOM Eff
[21]	0.9	15% @ -10dBm	0.75V @ -10dBm	1	SMS-7630	15×15	1.9	6.6
[21]	2.4	9% @ -13dBm	0.9V @ -13dBm	1	SMS-7630	15×15	1.25	8
[22]	2.45	10.5% @ -20dBm	0.075V @ -20dBm	1	SMS-7630	3.4×3.4	6.5	905
[23]	0.915/ 2.45	14%@ -20/-20	0,36V@ -10/-10	1	SMS-7630	6×6	0.5	185
[24]	1.8/2.2 /2.5	55%@ -10dBm	300mV@ -32dBm 3tones	1	SMS-7630	7×7	20	112
This work Dipole	0.915/ 2.44	**11% @ -20/-15dBm**	**1V @ -22/-18dBm**	4	HSMS-2850	2.1×11	**19.8**	**136**
This work Patch	0.915/ 2.44	**12.5% @ - 19.5 dBm/ -20dBm**	**1V @ -19.5dBm/ -25dBm**	4	HSMS-2850	7.8×8.8	**10.5**	**100**

According the Table I, the patch-based harvester exhibits the highest sensitivity to rectify 1V with a dual-tone featuring: -19.5 dBm at 915 MHz and only -25 dBm at 2440 MHz. The FOM_{sens} represents the trade-off between the sensitivity performances of a rectenna and the antenna area. The FOM_{eff} rates the efficiency performances to the antenna area. For these two figures of merit, the rectenna based on the multi-arm dipole element yields the best trade-off, both for FOM_{sens} and FOM_{eff}, compared to the patch-based solution. This dual tone and multi-arm dipole harvester is close to the work proposed in [24] which exhibits the highest FOM_{sens} reported so far in the literature to our knowledge.

V. CONCLUSION

The range of power collectable in a scenario of RF harvesting varies from -15dBm to -25dBm. To address this purpose the rectenna proposed in this work are optimized to operate at a RF input power close to -20 dBm (or 10 µW). To further improve the ability to collect the RF energy, these rectenna, developed with Schottky diodes HSMS285 from Avago, perform a concurrent harvesting in the 915 MHz and 2.44 GHz ISM bands. The harvester including a patch antenna implemented on a 1.6mm FR4 PCB achieves the highest sensitivity. It provides a 1V-rectified voltage for a dual-tone excitation of -19.5 dBm at 915 MHz and -25 dBm at 2.44 GHz. For these conditions of operation the rectenna yields a power efficiency of 12.5%. To take into account the dimensions of the haverster, two figures of merit, FOM_{sens} and FOM_{eff} including the size of the antenna, respectively related to the power sensitivity and the power efficiency are proposed. The rectenna developed with the arm dipole element exhibits the highest figures of merit. A case of application is proposed with the remote powering of a digital clock consuming 1V/5µA. The patch based harvester turns on the device with a dual tone excitation at the antenna of -19.5 dBm at 915 MHz and -20 dBm at 2.44 GHz. For the same scenario the harvester connected to the multi-arm dipole element needs a power of -22.1 dBm at 915 MHz and -17.8 dBm at 2.44 GHz.

REFERENCES

[1] T. Paing, A. Dolgov, J. Shin, J. Morroni, J. Brannan, R. Zane, and Z. Popovic, "Wirelessly powered wireless sensor platform'' in Eur. Microw. Conf. Dig, Munich, Germany, Oct. 2007, pp. 241-244.

[2] J. Bernhard, K. Hietpas, E. George, D. Kuchima, and H. Reis, ''An interdisciplinary effort to develop a wireless embedded sensor system to monitor and assess corrosion in the tendons of pre-stressed concrete girders,'' in Proc. IEEE Top. Conf. Wireless Commun, 2003, pp. 241-243.

[3] C. Walsh, S. Rondineau, M. Jankovic, G. Zhao, and Z. Popovic, ''A conformal 10-GHz rectenna for wireless powering of piezoelectric sensor electronics,'' in IEEE MTT-S Int. Microw. Symp. Dig, Jun. 2005, pp. 143-146.

[4] X. Zhao, T. Qian, G. Mei, C. Kwan, R. Zane, C. Walsh, T. Paing, and Z. Popovic, ''Active health monitoring of an aircraft wing with an embedded piezoelectric sensor/actuator network,'' Smart Mater. Struct, vol. 16, no. 2007, Jun. 2007, pp. 1218-1225.

[5] K. Lin, J. Yu, J. Hsu, S. Zahedi, D. Lee, J. Friedman, A. Kansal, V. Raghunathan, and M. Srivastava, "Heliomote: Enabling long-lived sensor networks through solar energy harvesting," in 3rd Int. Conf. Embedded Networked Sensor Syst., Nov. 2005, p. 309.

[6] J.A. Paradiso and T. Starner, "Energy scavenging for mobile and wireless electronics," IEEE Pervasive Computing, Vol. 4, and Issue: 1, Jan-March 2005, pp. 18- 27.

[7] K. R. Foster, "A world awash with wireless devices", IEEE Microwave magazine, vol. 14, Issue 2, March 2013, pp. 73-84.

[8] T. Le, K. Mayaram, and T. Fiez, "Efficient far-field radio frequency energy harvesting for passively powered sensor networks," IEEE J. Solid-State Circuits, vol. 43, no. 5, May 2008, pp. 1287-1302.

[9] J. Hirai, T.W. Kim, and A. Kawamura, ''Wireless transmission of power and information for cableless linear motor drive,'' IEEE Trans. Power Electron, vol. 15, 2000, pp. 21-27.

[10] H. Nishimoto, Y. Kawahara, and T. Asami, "Prototype implementation of ambient RF energy harvesting wireless sensor networks," IEEE Sens. J., Nov. 2010, pp. 1282-1287.

[11] BN. Shinohara, "Power without wires," IEEE Microw. Mag, vol. 12, no. 7, Dec. 2011, pp. S64-S73.

[12] M. Pinuela, P.D. Mitcheson, S. Lucyszyn, “Ambient RF Energy Harvesting in Urban and Semi-Urban Environments”, IEEE Trans. On Microwave Theory and Techniques, vol. 61, N°7, July 2013.

[13] Simon Hemour, Ke Wu “Radio-Frequency Rectifier for Electromagnetic Energy Harvesting : Developement Path and Future Outlook”, Proceeding of the IEEE, Vol.102, No.11, November 2014.

[14] T. Taris, L.Fadel, L.Oyhenart, V. Vigneras, “COTS-Based Modules for Far-Field Radio Frequency ”, IEEE NEWCAS, Paris, France, June 2013, pp. 1-4.

[15] Sika Shrestha, Sun-Kuk Noh, and Dong-You Choi, « Comparative Study of Antenna Designs for RF Energy Harvesting » International Journal of Antennas and Propagation Volume 2013, 10 pages.

[16] Z. W. Sim, R. Shuttleworth, M. J. Alexander, B. D. Grieve, « Compact patch antenna design for outdoor RF energy harvesting in wireless sensor networks, Progress In Electromagnetics Research, Vol. 105, 273–294, 2010.

[17] Naimul Hasan, Santu Kumar Giri, « Design of low power RF to DC generator for energy harvesting application », International Journal of Applied Sciences and Engineering Research, Vol 1, Issue 4, 2012.

[18] H. J. Visser, "Approximate Antenna Analysis for CAD", Wiley, 2009.

[19] R. Berges, L.Fadel, L. Oyhenart, V. Vigneras, T. Taris, “A dual Band 915MHz/2.44GHz RF Energy Harvester”, EUMW, Paris France, Sept 2015.

[20] Y-Y.Lu, J. Guo, H-C. Huang, “Design of Triple Symmetric Arms Dipole Antenna for 900/1800/2450MHz Applications”, Conf on Intelligent Information Hiding and Multimedia Signal Processing”, 2014.

[21] D. Masotti, A. Costanzo, M. Del Prete, V. Rizzoli, “Genetic-based design of a tetra-band high-efficiency radio-frequency energy harvesting system,” IET Microwaves, Antennas Propagation, vol.7, no. 15, June 2013, pp. 1254-1263.

[22] G. Vera, A. Georgiadis, A. Collado, and S. Via, “Design of a 2.45GHz rectenna for electromagnetic (EM) energy scavenging”, Proc. IEEE Radio and Wireless Symp, 2010, pp. 61-64.

[23] K. Niotaki, S. Kim, S. Jeong, A. Collado, A. Georgiadis, and M. Tentzeris, “A compact dual-band rectenna using slot-loaded dual band folded dipole antenna,” IEEE Antennas and Wireless Propagation. Lett. vol. 12, pp. 1634–1637, 2013.

[24] C. Song, Y. Huang, J. Zhou, J. Zhang, S. Yuan, and P. Carter, “A High-efficiency Broadband Rectenna for Ambient Wireless Energy Harvesting”, IEEE Transactions on Antennas and Propagation, accepted May 2015.

Bibliographies

Ludivine FADEL received the Ph.D in electrical engineering from the University of Bordeaux, France, in 2004. She is an Associate Professor at Bordeaux University and her main research interests include wireless power transfer and energy harvesting, materials and printing technologies for RF application.

Laurent Oyhenart received the Ph.D. degree in electrical engineering from the University of Bordeaux, Bordeaux, France, in 2005. He is currently an Associate Professor with the University of Bordeaux and develops his research activities in the IMS laboratory. His research interests include computational electromagnetics, photonic crystals and antenna synthesis.

Romain Bergès is a PhD student at IMS Laboratory. He received his Master from Bordeaux University in Electronic, in 2014. His current research focuses on RF Energy Harvesting.

Valérie Vigneras is a full professor in the Polytechnic Institute of Bordeaux and conducts research in the IMS Laboratory on complex structures (mixtures, structured media, self-assembled media) both by simulation to predict or optimize their properties and by characterization before their integration in high frequency systems (antennas, electromagnetic compatibility, radar shieldness…).

Thierry Taris is a full professor at the Bordeaux Institute of Technolgy (Bx-INP). He joined the IMS lab in 2005 where is research interests are related to the Radio-frequency Integrated Circuits in Silicon technologies.

List of figures and tables

Fig. 1: Schematic of a Wireless RF Power Transmission (WPT) system.
Fig. 2: Architecture of the RF to DC converter.
Fig. 3: Topology of the input-matching network.
Fig. 4: Equivalent model of the RF to DC converter, 915 MHz (a) - 2,44 GHz (b).
Fig. 5: Measured S_{11} of a 4-stage voltage multiplier with a L section at 900 MHz for various input power P_{rf}.
Fig. 6: Unloaded rectified voltage for various inpout power.
Fig. 7. Layout of the RMPA antenna (a) - Radiation pattern at 915MHz (b) and 2.44 GHz (c) Solid and dashed lines correspond to E-plane and H-plane respectively.
Fig. 8. Layout of the multi-band arms dipole antenna (a) - Radiation pattern at 915 MHz (b) and 2.44 GHz (c). Solid and dashed lines correspond to E-plane and H-plane respectively.
Fig. 9. Dual-band RF harvesters based on patch antenna (a) – arms dipole antenna (b).
Fig. 10. Measured return loss S_{11} of the dual-band rectfier, patch and arms dipole antenna.
Fig. 11. Schematic and picture of the scene of remote powering of a clock.
Fig. 12. Power efficiency of the dual-band RF harvester based on the patch (a) and the arms dipole (b) antenna.
Fig. 13. Rectified voltage of the RF harvester based on patch (a) and arm dipole (b) antenna.
Table I. Comparison with the state of the art.

Fun with Less Kilowatts: The Lectenna

By Vernon Trollinger, March 14, 2017, Energy Efficiency, Events & Fun, Family

Welcome to Fun with Less Kilowatts! We believe that science experiments at home can be a creative way to engage kids in learning while having fun. They can be educational AND great activities to keep your kids busy and away from the television. Each month, we'll feature a new science experiment that can be a great resource for parents and teachers.

The Lectenna

Did you know you that the energy from radio waves can be converted back into electric energy? It's true and you can make a *simple* two-component circuit that you and your kids can use to find 2.45 Ghz radio waves around your home. These are the same radio waves used by microwaves, WiFi network routers, bluetooth devices, smart phones, cordless phones, smart meters, and smart home systems that use Zigbee.

YouTuber pjaffeva posted this brilliant project showing how. Plus, there's also a link to a PDF set of instructions on Google Drive that you can download.

The Materials

- One HLMP-D150 Avago low current LED. Costs about 50¢/each. Available from: Amazon, Digikey, Mouser
- One Hitachi 1SS106 low capacitance Schottky diode. Pjaffeva recommends this Hitachi-made diode as being the only one in this shape that will work. But they're also hard to find. Currently, you can get them from LittleDiode's ebay store in the UK. Costs $6.38 but the bulk of that price is shipping from the UK.

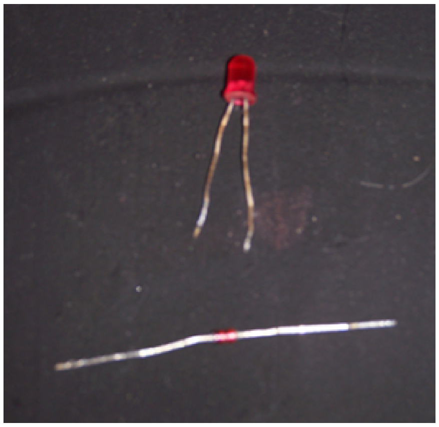

You'll also need:

- An operating WiFi router
- A bamboo kabob skewer
- Two pieces of tape

The Directions

1) Look on the base of the LED and find the side with the flat edge. This marks the negative or cathode side. Carefully splay the two wire leads away from each other as far as they will go. Make sure the one adjacent to the negative side gets bent over the flat edge so identify it. You can also mark it with a little spot of colored nail polish.

Categories

- Careers/Jobs
- Charity & Community
- Energy Efficiency
- Events & Fun
- Family
- FAQs
- Green
- Home Improvement
- Hurricane Prep
- Moving
- News
- Recipes
- Save Money
- Small Biz
- Special Offers

Subscribe via Email

Subscribe

Subscribe via Email

Subscribe

Subscribe

Subscribe in a reader

Archives

- December 2017
- November 2017
- October 2017
- September 2017
- August 2017
- July 2017
- June 2017
- May 2017
- April 2017
- March 2017
- February 2017
- January 2017
- *all archives*

Featured Articles

Prepare for Christmas With Green Cleaning Methods

Holiday Kitchen Safety – Cooking with Kids
How to Shop for Old Window Repairs
How Do I Clean A Clothes Dryer?
Gobble Up These Convenient Thanksgiving Deals on Meals

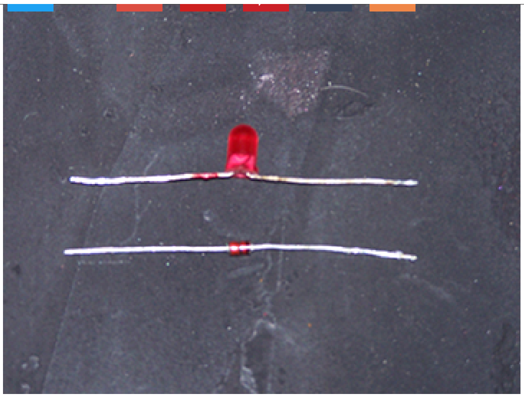

2) Next, find the band that runs around the Schottky diode. This identifies the negative end of this diode. The other end is the positive or anode end.

Using your fingers:

3) Twist the end of the negative or cathode wire from the LED onto to the positive or anode end of the Schottky diode.

4) Then twist the end of the negative or cathode end of the Schottky diode onto to the positive or anode end of the LED.

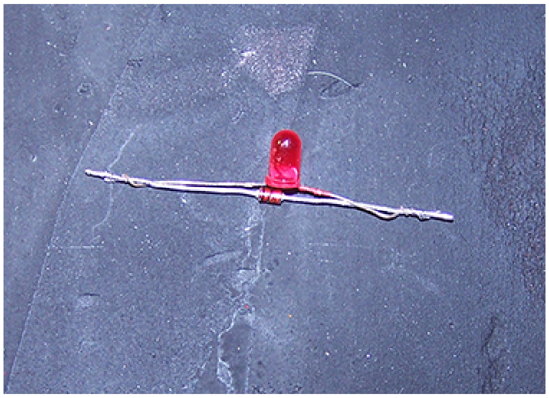

You have just made a "rectenna" and you'll need to keep the assembly and the wires straight as possible. The length of the wires should be about 6 cm long (a hair short of 2 3/8 inches).

5) Tape the two diodes to one end of the bamboo skewer.

6) With a WiFi router on and running, hold the diode near an antenna on the WiFi router.

The Result

The LED should light up as it draws closer to the antenna (signal). The further away it is, the weaker the signal and the less it can grab to light up.

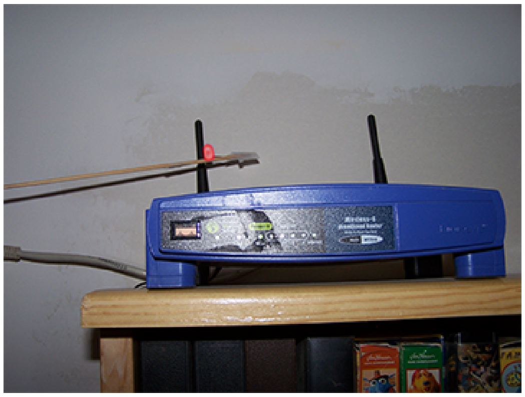

The Science

The WiFi routers uses radio waves to transmit data to and from computers and other devices in your home. Radio waves behave like water in a glass, they ripple or oscillate at a certain number of times or frequency per second. A radio wave's wavelength is the distance covered by one complete cycle of a radio wave — from peak to valley.

Radio waves used by WiFi routers, bluetooth devices, microwaves, etc., broadcast signals in the 2.45 gigaherz (Ghz) range or 2.45 billion oscillations per second. That means the signal wavelength is about 12 cm. To pick up signal, however, you only need an antenna that's half as big— 6 cm. When the radio waves strike this antenna, called a 1/2 dipole ("di"=two), the energy excites the antenna's

The "Rectenna": Because the radio waves cause voltage to flow one direction and then the opposite direction, the current is called an alternating current (AC). By connecting a diode to both poles of the antenna it converts the AC current to flow only in one direction, called direct current (DC). The antenna becomes a "rectenna".

How the diodes work: The diodes work as one-way valves for electricity, allowing voltage to flow forward but not back. In the case of the rectenna, AC current caused by the radio wave sends a positive charge down one of the antenna poles and then the other. By putting the Hitachi 1SS106 diode across both poles, the diode only lets the positive charge coming down one of the antenna poles through. The LED is actually a Light Emitting Diode and works the same way as a regular diode except that when it's on, it lights up. Meanwhile, Schottky diodes turn on at a lower voltage than other diodesand because the rectenna can only catch a fraction of the WiFi's broadcast power, the Schottky diode can rectify that little bit of AC to DC.

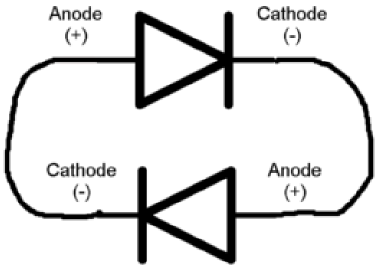

When we connected our diodes together, their polarity formed a loop that only allowed voltage to flow one way and thus light the LED. Brilliant job, pjaffeva!

Do you have any fun and kid-friendly science experiments you'd like to see us try for Fun with Less Kilowatts? Share with us in the comments!

Be Sociable, Share!

Tweet Like 0

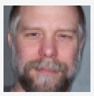

About Vernon Trollinger
A native of Wyomissing Hills, PA, Vernon Trollinger studied writing and film at the University of Iowa, later earning his MA in writing there as well. Following a decade of digging in CRM archaeology, he now writes about green energy technology, home energy efficiency, DIY projects, the natural gas industry, and the electrical grid.

Tags: family fun, Fun with Less Kilowatts, kids, science experiments

Like 338K people like this. Sign Up to see what your friends like.

Follow @BounceEnergy 28.2K followers

Share 1 Tweet 0 Pin 0 G+

Follow Bounce: Share this on:

0 Comments Bounce Energy Blog

1 Login

Recommend Share

Sort by Best

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

ALSO ON **BOUNCE ENERGY BLOG**

Choosing the Best Home Generator for Your Emergency Needs

1 comment • 3 years ago

T&B Electric LTD — Nice article covering the difference between portable and permanent generators. Some homeowners "back feed" …

Career Opportunity with Bounce Energy | PHP/Integration Developer

1 comment • 3 years ago

Chris Jones — Sounds like a great opportunity.

Take Care of Your Favorite Gadgets with Great Waterproof Tech

1 comment • 2 years ago

erna — Vermon thanks for your post, great as always! Coolest is cool I got one of this :) what's funny I also have swimmo watch for swimmers …

5 Quick DIY Home Improvement Fixes to Get Your Home Through Winter

1 comment • 3 years ago

Steve Veach — Excellent and informative blog post! I am a HVAC Contractor in Chicago and have found a new site to check in on. Thank …

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy DISQUS

Copyright © 2017 Bounce Energy. Learn more with Bounce Energy Articles | Charity and Community Involvement

Follow Bounce: Share this on:

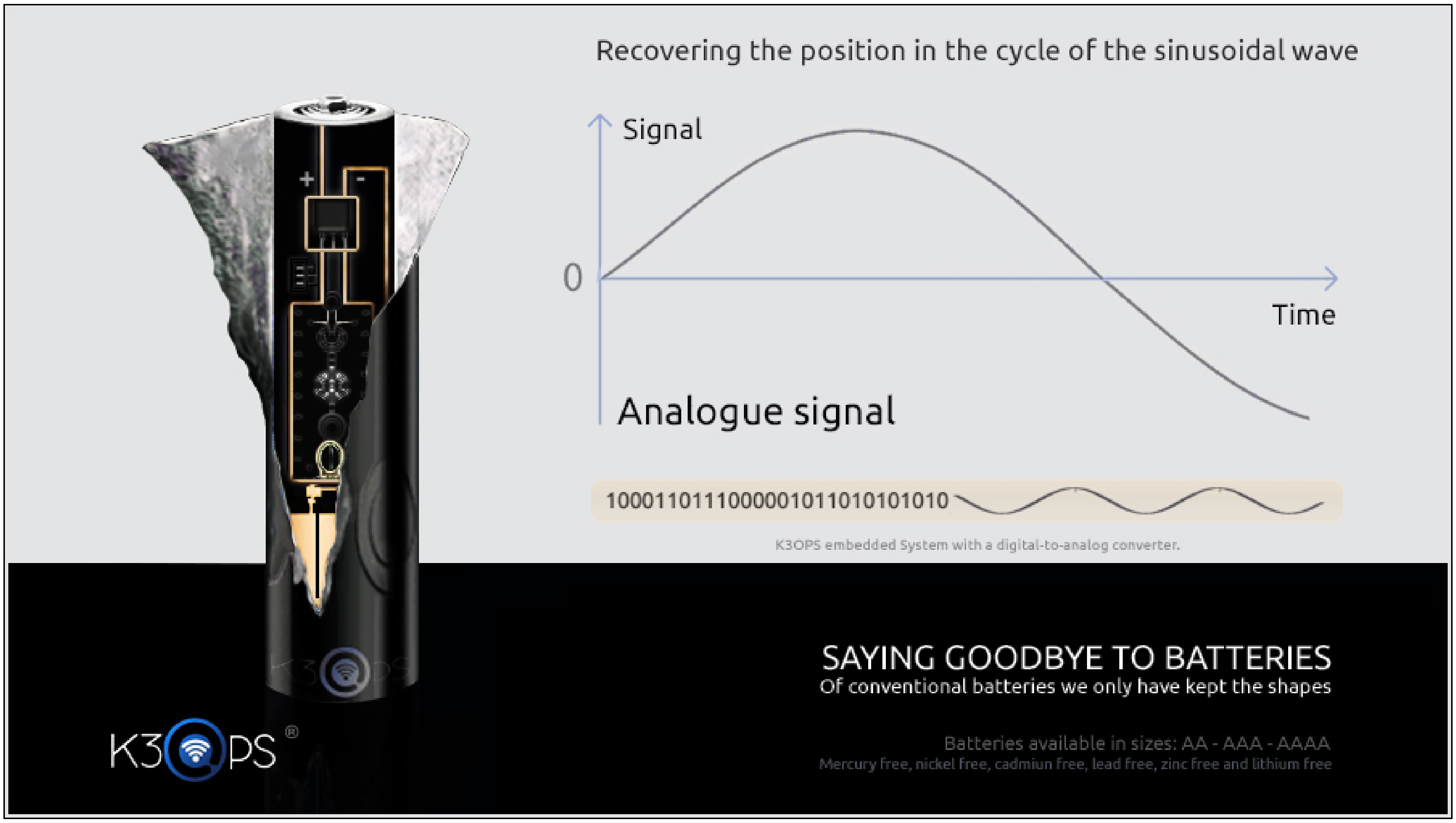
Recovering the position in the cycle of the sinusoidal wave
Signal
0
Time
Analogue signal
10001101110000010110101010101010
K3OPS embedded System with a digital-to-analog converter.
+
-
K3OPS
K3OPS®
SAYING GOODBYE TO BATTERIES
Of conventional batteries we only have kept the shapes
Batteries available in sizes: AA - AAA - AAAA
Mercury free, nickel free, cadmiun free, lead free, zinc free and lithium free

An electromagnetic field uses the photon as an elementary particle to transmit force. It combines:

- **A magnetic field** *force resulting from the movement of loads μT.*
- **An electric field** *force created by the attraction of repulsion loads, measured in volts per meter - V/m.*

With an energy determined according to the speed of light, the RF are by far the best medium to transmit any kind of information.

The multiplication of wireless communications systems in our environment ensures sufficient microwave leakages to harvest from the ambiant and enough energy to convert into DC electricity. Electromagnetic fields are everywhere and since they carry energy, they became the best candidate to deliver an endless source of renewable energy.

$\vec{B}$ is the magnetic induction expressed in T refered to **Nikola Tesla**, "Father of Free Energy", which is at the origin of the electromagnetism.

Using meta-materials combined with nanotechnology has deeply increased the performance and miniaturization of rectennas embedded in K3OPS system. Our products operate autonomously, offering an endless supply of green energy in a respectful and environment-friendly approach.

principles and beautiful sentences. At that point in time I would have simply been Xin, a child like any other...

But one day I dreamed. I woke up far away, somewhere else, in another past. I decided to change my destiny and even if I was supposed to become a mathematician, I eventually decided to create, because already as a child my heart was chasing the stars.

Of the hundreds of directions shown to me after graduating, only one captured my attention: a single goal... Build the impossible for a safer world. So, over the past 3 years, our real challenge to overcome for all RF Energy Harvesting technics was to optimize electricity conversion. The massive proliferation of wireless telecommunication systems since the past two decades brought a saturation of the electromagnetic fields with a constant growth of 15% every year in our environments. As a result, this situation reversed the base problem that makes today Harvesting RF Energy a game changer. The key was the Power Management System.

We are far beyond the conversion constraint and performance by controlling "RF-interferences", by harvesting different frequencies *from near and far*, by using Metamaterials combined with nanotechnologies. We dramatically have improved power conversion efficiency and reduced the size of our Energy Harvesting systems embedded in all K3OPS' products.

Thanks to Nikola Tesla, my inspiring mentor, K3OPS' products have reached by far their objectives in terms of converting and performance, offering an endless efficient source of green energy, reliable in an environmentally friendly approach.

Xin WEI Co-Founder of K3OPS technology with Alexandre Despallieres

| HOME |

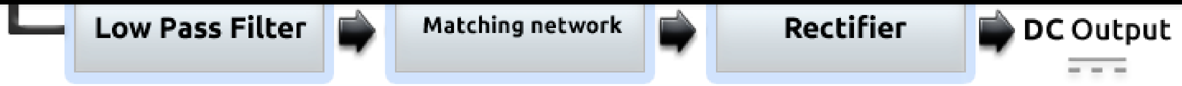
Low Pass Filter
Matching network
Rectifier
DC Output

The Rectenna was invented in 1964 by William C. Brown, patented in 1969. It is a rectifying antenna used to convert microwave energy into DC. A simple Rectenna consists of a dipole antenna with an RF diode connected across the dipole elements. The diode rectifies the AC current induced in the antenna to produce DC power.

K3OPS
RF Energy Harvesting Battery

K3OPS batteries
K3-Case

K3OPS Copyright © 2017 Alexandre Despallieres and Xin Wei - All rights reserved.

Terms of use | Privacy Policy | Press | Login

Centre for Collaboration in Electromagnetics and Antenna Engineering is a centre for research collaboration established to foster research in electromagnetics and engineering conducted by Macquarie University researchers and prominent collaborators from external institutions, including people from industry and overseas ins We have state of the art research ficilities including a NSI-700S-50 spherical nearfield anechoic chamber, 3 Vector Network Analyzers (capable up to 22GHz, 50GHz, anc respectively), state-of-the-art embroidery machine DreamCreator-XE-VM5100 for embroided antennas, Dielectric Characterization Kit - High Temperature Probe (2 20GHz), as well as several in-house and commercial licenced softwares.

Our recent and ongoing research projects are:

- Wireless Implantable Bio-Telemetry System and Miniature Antenna Design
- Wireless Freedom for Lab Rats
- Flexible and Wearable Antennas for Biomedical and Healthcare Applications
- Characterizing Properties of Carbon nano Tubes at Microwave Frequencies
- Electromagnetic Band Gap (EBG) Resonator Antennas
- Leaky-Wave Antennas for Advanced Wireless Systems
- Shared Aperture Arrays for Space Borne Applications
- Novel Dielectric Resonator Antennas
- Super Wideband Antennas
- Focal Plane Arrays for Radio Astronomy
- Archimedian Spiral Metamaterials
- Frequency Selective Surfaces for Energy-Saving Glass Panels
- Dual-Band Artificial Magnetic Conductor (AMC) Surfaces
- Antenna Technologies for Ultrawideband (UWB) Systems
- High Gain Antennas with Planar Surface-Mounted Short Horns
- Photonic Crystal/EBG Based Horn Antennas
- Broadband Microstrip Patch Antennas for Wireless Computer Networks
- Theoretical Analysis of Photonic Crystal Structures
- New Closed-form Green's Functions for Microstrip Circuits and other Layered Structures
- Integrated-design of Hybrid-resonator Antennas for Broadband Wireless and other Communication Systems
- Singularity-enhanced Finite-different Time-domain (FDTD) Method for Diagonal Metal Edges, Strips and Films
- Low-profile Dielectric-resonator (DR) Antennas

Wireless Implantable Bio-Telemetry System and Miniature Antenna Design

The two major challenges associated with the conversion of a wireless system operating in air to an implantable version, antenna detuning and biocompatibility, are a in a coherent way. An RFID-based biomedical telemetry system designed for free-space operation was chosen as the starting reference. A new, pin-compatible, spa antenna with a ground plane was designed, fabricated and tested, to replace the original “free-space” antenna in the active RFID tag without making any other chang tag circuit, such that the tag would function well when it is placed under rat skin and fat. Biocompatibility and potential antenna detuning due to rat tissue variati addressed in the design process, without significantly increasing the tag physical height, by applying a thin coating of biocompatible material directly over the ante operation of the medical telemetry system was successfully demonstrated, with the tag placed under rat skin and fat, and its range of 60-72 cm was found to be suf support medical research experiments conducted with rats in cages. Due to the biocompatible coating over the antenna, antenna matching is very insensitive to ch tissue dielectric constants and thickness. The footprint of the new antenna is 33% less than that of the original antenna, its measured 10 dB return-loss bandwidth is 1 11%, and overall efficiency is 0.82% at 920 MHz.

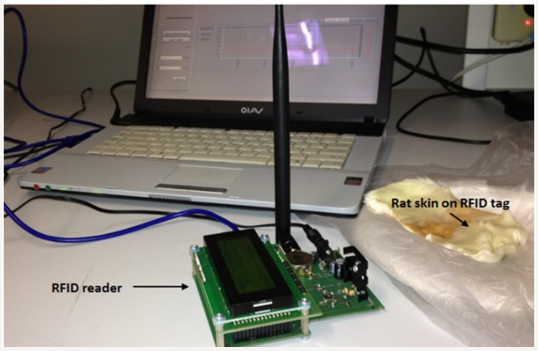

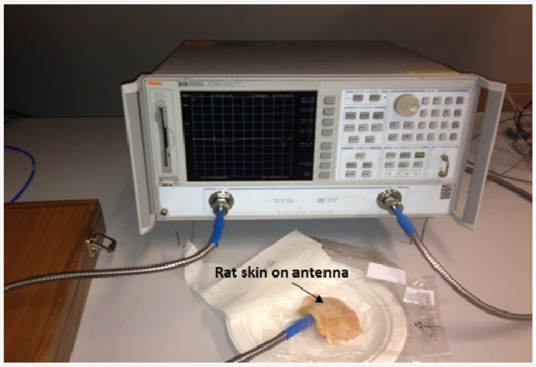

Wireless Freedom for Lab Rats

We are developing a fully implantable wireless telemetry system. This is a joint research project with BCS Innovations and the Australian School of Advanced Medicine will be first used in the research conducted in ASAM, with rats, on hypertension. To date the major method of controlling hypertension is through the use o pharmaceuticals. The pathway to approval for most drugs for human use involves pre-clinical (animal) trials. Lab rats are considered biologically similar to humans, pa in terms of their social behaviour. Therefore, it is very important to not compromise the pharmaceutical trials by unnecessarily stressing the rats by harnessing the monitoring equipment. One of the technical challenges of developing an implantable system that monitors the various signals, is the relatively small size requ implantable telemetry system is a miniature transceiver implanted in an animal that senses, processes and transmits data via a wireless link to monitor vital signals of c freely moving laboratory animals. This is crucial in giving researchers flexibility and reliability, especially in studies with special experimental settings using mazes wheels and treadmills. We have plans to develop a fully implantable telemetry system for subcutaneous or intraperitoneal placement in rats that monitors th parameters as well as blood pH and chemistry, nerve activity and circadian respiratory rate rhythms. The aim of this project is to eventually develop a system with e capabilities that costs less than what is currently available, to provide more universities and researchers with the opportunity to use this technology.

In this sequel, initially, when a module of our original system was placed under the skin of a rat, the wireless link failed completely. It could not send a temperature read a centimetre! The point of failure was the commercial antenna in the module that had been designed to work outside the body, in air. Such antennas do not work unde because the electrical characteristics of skin (rat or human) are significantly different from that of air. Hence the main challenge was to design an antenna that works v placed under the skin. In addition, it was necessary to cover the module and the antenna by biocompatible material, which also affects antenna performance. Possible v of skin characteristics from one rat to another or one person to another were considered. Unlike the commercial antenna, we wanted our antenna to radiate less into th the rat/person and more away from the body because that not only increases the quality of the wireless link, maximum range (distance between implanted mo monitoring station) and battery life but also reduces the exposure of the body to radio-frequency waves. Indeed we had to consider the electromagnetic effects of fat a material around the antenna. We were able to meet all these requirements with a novel compact antenna design that is approx. one third the size of the original co antenna. We successfully demonstrated wireless telemetry transmission of temperature with the new module placed under rat skin and the monitoring station pl distance of up to 80 cm! This range is sufficient for our immediate target application supporting new medical research by Professor Paul Pilowsky's team. If necessary, further extended by increasing the power level at the monitoring station.

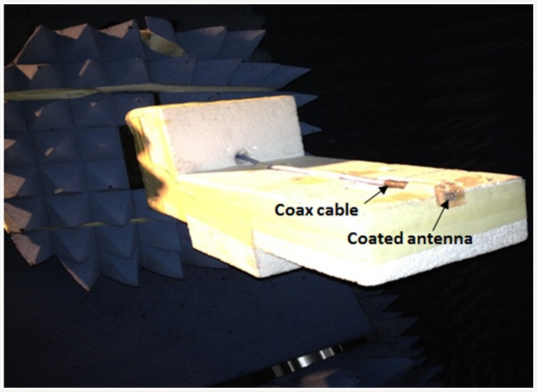

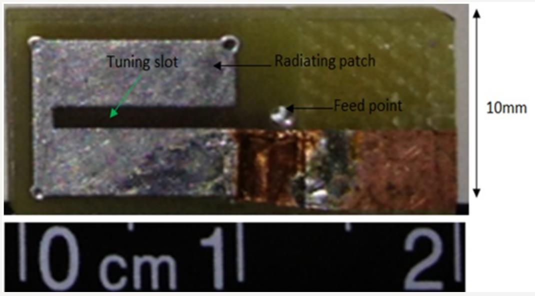

Flexible and Wearable Antennas for Bio-Medical and Healthcare Applications

Body Centric Wireless Communication is a rapidly growing research area targeted for medical, healthcare, public safety and defense applications. The need to address transceiver specifications and real-time scenarios in close proximity to human body is continuously evolving antenna system research. Several novel miniature antennas havi dual- and wide-band operations have been designed and tested for Wireless Body Area Network (WBAN). They have significant advantages of small size, wide radiation pat the human body for maximum coverage and are less sensitive to the gap variation between human body and antenna.

A compact ultra wideband antenna is shown below with strong notch-band rejections up to VSWR = 26, that is tunable over a wide frequency range from 3.55GHz to 6.8 been designed. To estimate the stub length to notch frequency for a given interfering application, analytical expressions for the normalized stub length which is indep substrate dielectric constant is also presented. This helps to avoid hit-an-trail method and gives a good estimate of initial design parameters for notch. Proposed antenna radiation patterns and yields a measured 10dB return-loss bandwidth from 3GHz to 10.5GHz.

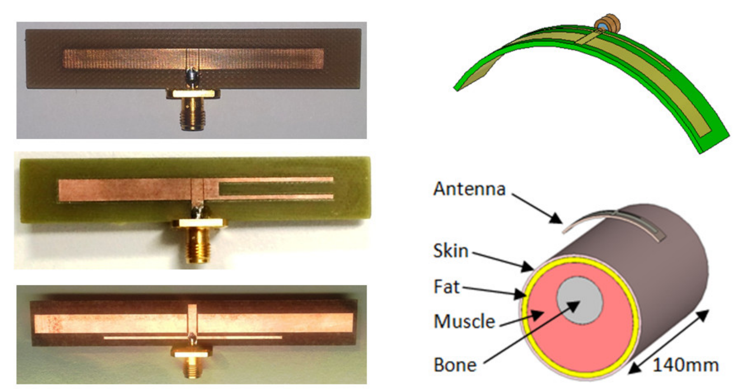

Characterizing Properties of Carbon Nano Tubes at Microwave Frequencies

Carbon Nanotube (CNT) yarns are novel CNT-based materials that extend the advantages of CNT from the nano-scale to macro-scale applications. We have modelled CNT potential data transmission lines. Test structures have been designed to measure electrical properties of CNT yarns, which are attached to these test structures using gold testing and microwave S-parameter measurements have been conducted for characterisation. The observed frequency independent resistive behaviour of the CNT yarn promising indicator that this material, with its added values of mechanical resilience and thermal conductivity, could be invaluable for a range of applications such as B Networks (BAN). A model is developed for CNT yarn, which fits the measured data collected and agrees in general with similar data for non-yarn CNTs.

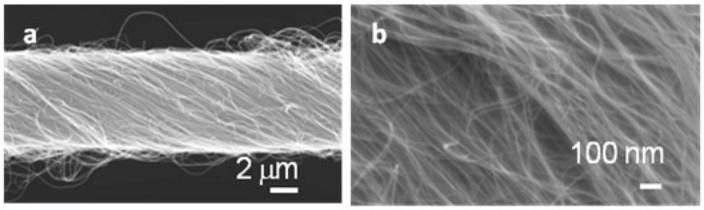

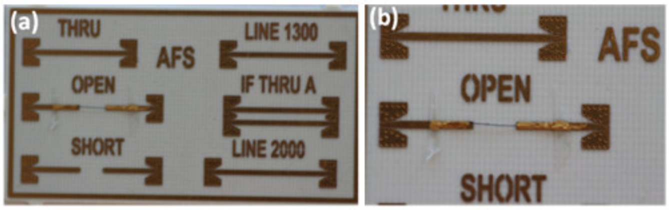

Electromagnetic Band Gap (EBG) Resonator Antennas (ERA)

We have designed, fabricated, and measured antennas based on 3D, planar and 1D EBG structures (i.e. photonic crystals). These flat microwave antennas, known as EBG antennas or Fabry-Perot cavity antennas, can give gains of about 20dB and very good efficiencies.

Enhancing Radiation Characteristics of ERAs by Improving Aperture Phase Distribution

This work focuses on achieving superior radiation characteristics of ERAs by improving their aperture phase distributions. A unique method, utilizing full-wave simula analytical analysis, has been developed to design Phase Correcting Structures (PCS) for ERAs. This method uses actual phase distribution on the physical aperture of ERAs relying on geometric optics. Several Phase Correcting Structures (PCSs) have been designed, which were later validated with the measurements of their fabricated p Tremendous improvement in the radiation performance is witnessed in both simulated and measured results. These exciting initial results validate our proposed method indicate an existence of a great potential to be explored. The figure below shows the field propagation above an ERA: (a) without PCS, and (b) with PCS. It is clear that the the PCS is much more uniform and the energy is focussed towards broadside direction, thus, resulting in increased directivity.

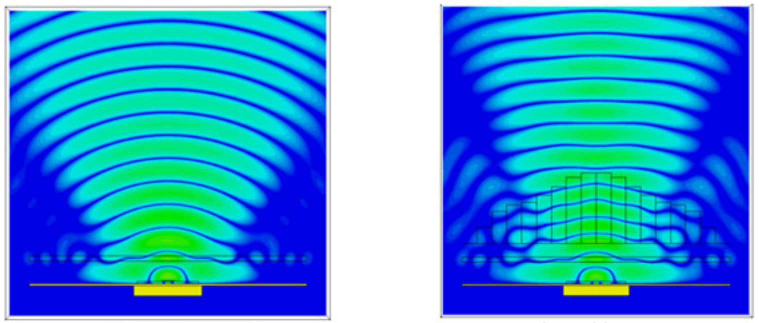

a. without PCS b. with PCS

Extremely Wideband High-Gain ERAs (Gain~15-20dBi, Bandwidths >50%)

In 2014-15, we invented an innovative class of electromagnetic band-gap (EBG) resonator antennas which provide high gain and wide bandwidth with an extremely r footprint. One of the prototype developed has only 8% of the area as compared to conventional EBG resonator antennas but its performance (gain bandwidth) is a record hi for this class of antennas, while providing gain in the range of 15-20 dBi. This represents an improvement of nearly two orders of magnitude in the bandwidth compared t EBG resonator antennas. Thanks to its practical advantages of flat shape and low-cost manufacturability, it can be easily attached to a wall of a building, for example, to co building wirelessly to the National Broadband Network.

Our most recent results on such wideband high gain antennas can be found in: “Achieving high gain-bandwidth through flat GRIN superstrates in Fabry-Perot cavity antenn 2014 IEEE International Symposium on Antennas and Propagation (AP-S/USNC-URSI), pp. 1748 – 1749, Memphis, Tennessee, USA, July 6-11, 2014.

Detailed antenna design along with experimental data of another of our wideband EBG resonator antenna having composite multi-layer superstrate, is published in the pa "Wideband high-gain EBG resonator antenna with a small footprint and all-dielectric superstructures" in IEEE Transactions on Antennas and Propagation, vol. 62, no. 6, 201

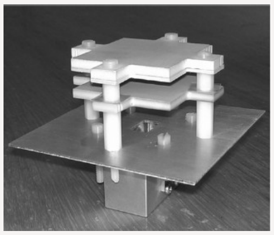

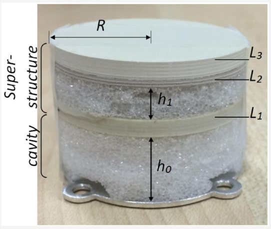

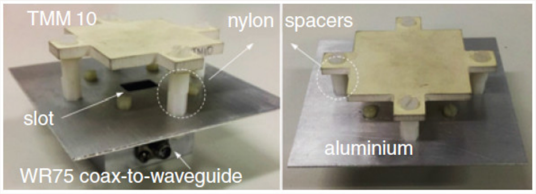

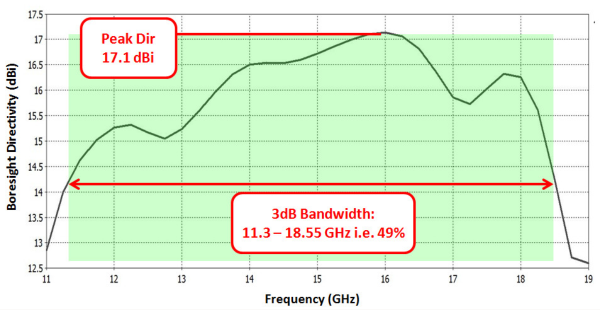

Frequency response of one of our Wideband ERA Prototypes

Simple Dual-Band ERAs

We developed a new method to obtain dual-band operation from a simple electromagnetic band gap resonator antenna. The antenna is based on a one-dimensional EBG made out of two low-cost unprinted dielectric slabs. The EBG structure is implemented as the antenna superstrate, which has been designed to provide a locally-inverted reflection phase gradient with high reflectivity, in two pre-determined frequency bands. The linearly polarised antenna design and experimental results are described in entitled: "A simple dual-band electromagnetic band gap resonator antenna based on inverted reflection phase gradient" published in IEEE Transactions on Antennas and Pro vol. 60, no. 10, pp. 4522-4529, 2012.

We have extended this concept for dual-band, dual-polarised and circularly-polarised antennas. We also designed a tri-band antenna following this concept. It needs only t cost unprinted dielectric slabs.

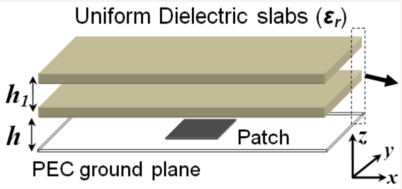

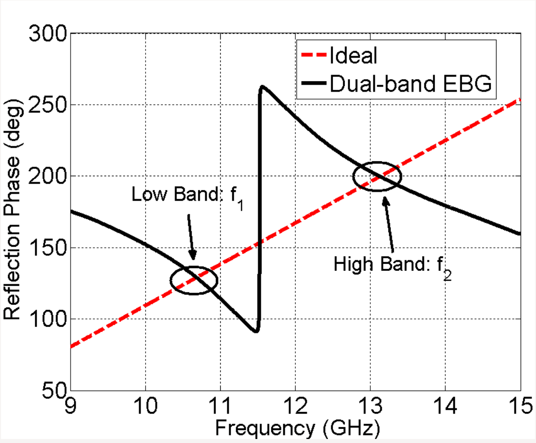

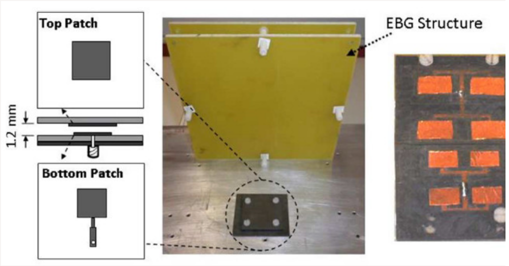

Low-Profile Wideband ERAs

We have designed and successfully tested wideband low-profile (thin) EBG resonator antennas. The breakthrough that contributed to this success is our design of a partially surface (PRS) with a positive reflection phase gradient. Thin single-dielectric-slab PRSs with printed patterns on both sides were investigated to minimise the PRS thickn simplify fabrication. Three such surfaces, each with printed dipoles on both sides, have been designed to obtain different positive reflection phase gradients and reflection r levels in the operating frequency bands. These surfaces, and the EBG resonator antennas formed from them, were analysed theoretically and experimentally to highlight t compromises involved and to reveal the relationships between the antenna peak gain, gain bandwidth, the reflection profile (i.e. positive phase gradient and magnitude) of th and the relative dimensions of dipoles. A small feed antenna, designed to operate in the cavity field environment, provides good impedance matching (|S11|< -10 dB) a operating frequency bands of all three EBG resonator antennas. Experimental results confirmed the wideband performance of a simple, low-profile EBG resonator antenna thickness is only 1.6mm, effective bandwidth is 12.6%, measured peak gain is 16.2 dBi at 11.5 GHz and 3dB gain bandwidth is 15.7%. Please find details in a paper entitled of Simple Thin Partially Reflective Surfaces with Positive Reflection Phase Gradients to Design Wideband, Low-Profile EBG Resonator Antennas," published in IEEE Tra on Antennas and Propagation in 2012.

Low-Profile Dual-Band ERAs

We have achieved dual-band operation in a low-profile EBGRA using a single dielectric superstate with a printed pattern only on one side. This also made use of our r inverting the gradient of the PRS reflection coefficient.

Woodpile EBG Material and Antennas

In 2003, we designed and built a woodpile 3D photonic crystal, also known as a EBG crystal, operating in microwave frequencies, and demonstrated experimentally the ex the electromagnetic band gap. The crystal is made out of cermain material.

Then we designed a planar EBG resonator antenna. Shown below, it has a resonant cavity between a ground place and a 3D woodpile photonic crystal. We employed both m and slots to feed the cavity, and investigated both linearly and circularly polarised antennas. The linearly polarised antenna design and experimental results are described in entitled "A planar resonator antenna based on a woodpile EBG material," published in IEEE Transactions on Antennas and Propagation,vol. 53, no. 1, pp. 216-223, January 2

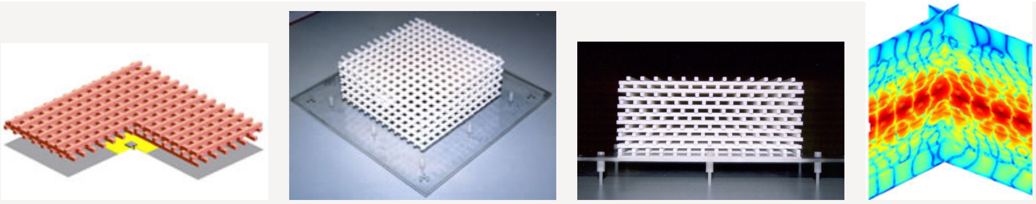

Leaky-Wave Antennas for Advanced Wireless Systems

Antenna beam steering can bring significant benefits to advanced wireless systems. Microstrip leaky-wave antennas (MLWAs) are of particular practical interest becaus planar low-profile configuration, ease of fabrication, and beam-scanning capabilities. In this research several planar MLWAs and arrays are developed to radiate at bores conical beam around the boresight, with simultaneous dual-side-beam scanning, dual-band forward and backward beam-scanning, and continuous beam scanning from the to the forward direction. Moreover, methods and antenna designs are proposed to steer the beam at a fixed frequency, shifting beam-steering range, and fixed-frequency bean in forward and backward directions.

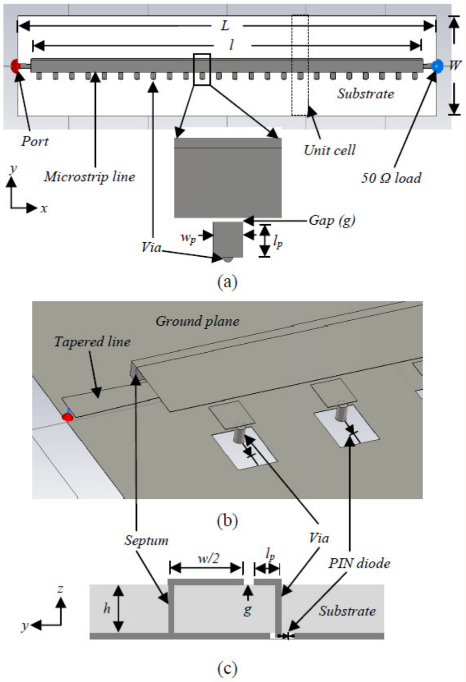
L
l
W
Substrate
Port
Unit cell
Microstrip line
50 Ω load
y
x
Gap (g)
w_p
l_p
Via
(a)
Ground plane
Tapered line
(b)
Septum
Via
PIN diode
w/2
l_p
z
h
g
Substrate
y
(c)

Figure 1. Proposed reconfigurable HW-MLWA: (a) top view, (b) perspective view, and (c) side view.

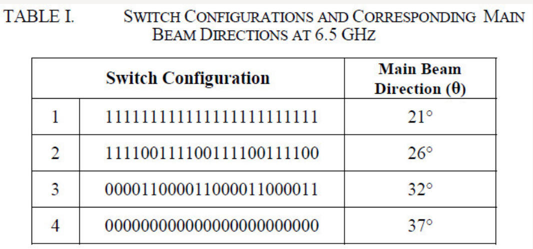

TABLE I. SWITCH CONFIGURATIONS AND CORRESPONDING MAIN BEAM DIRECTIONS AT 6.5 GHZ

	Switch Configuration	**Main Beam Direction (θ)**
1	111111111111111111111111	21°
2	111100111100111100111100	26°
3	000011000011000011000011	32°
4	000000000000000000000000	37°

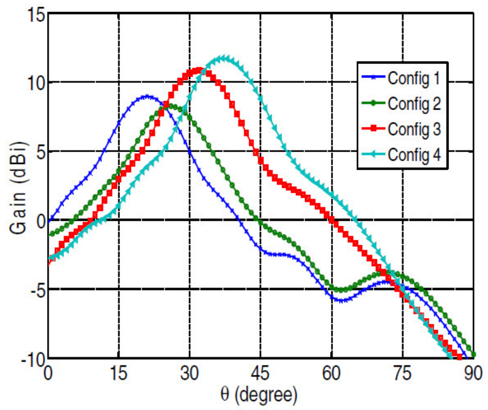

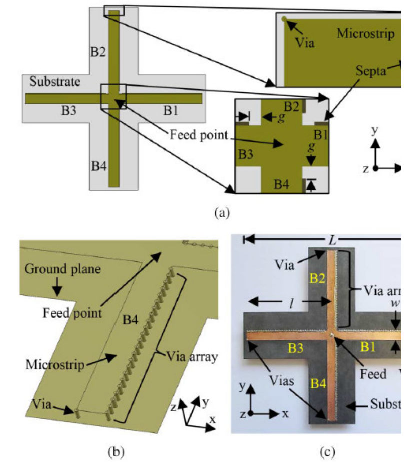
B2
Substrate
B3
B1
B4
Feed point
Via
Microstrip
Septa
B2
g
B1
g
B3
B4
y
z
(a)
Ground plane
Feed point
B4
Microstrip
Via array
Via
z
y
x
(b)
L
Via
B2
Via arr
l
w
B3
B1
Vias
Feed
B4
Subst
y
z
x
(c)

Fig. 1. HW-MLWA array design steps: (a) top view of the initial de septa, (b) perspective view of one branch of the final design with v (substrate omitted, not to scale), and (c) top view of the fabricated pro

Multi-band Dual-Polarized Shared Aperture Array

Multi-band dual-polarized shared-aperture (MBDP-SA) arrays are antenna arrays that operate in two (or more) frequency bands with dual-polarization in each band, a elements are integrated together into a common physical space by sharing the single aperture. The MBDP-SA array is of great interest in space-borne SAR system, bec technique can effectively reduce the payload and size of the antenna sub-system. In this research project, main efforts are focussed on three aspects: 1) improve the specifi current Dual-Band Dual-Polarized Shared-Aperture (DBDP-SA) array; 2) construct Tri-Band Dual-Polarized Shared-Aperture (TBDP-SA) array; and 3) explore some new for DBDP-SA antenna.

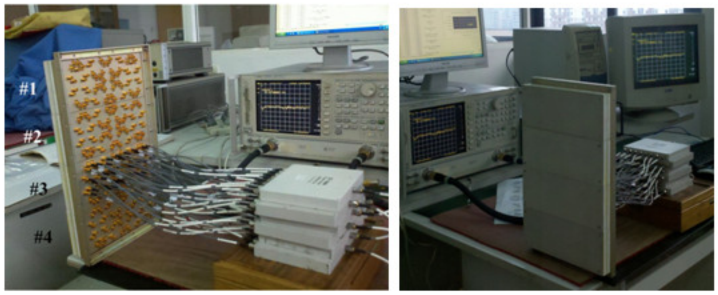

(a) Rear (b) Front

Fig.1 Improved DBDP-SA array.

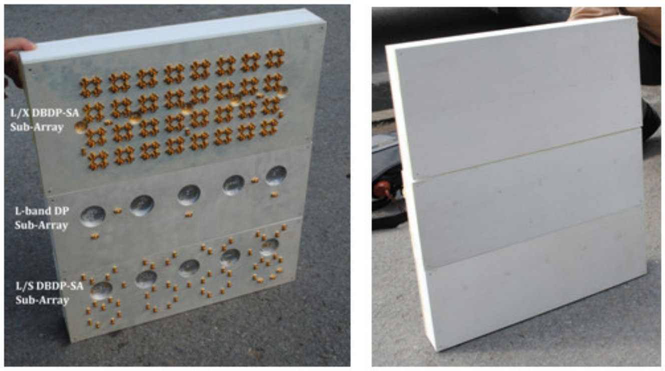

(a) Rear (b) Front

Fig.2 Photos of the TBDP shared aperture prototype array.

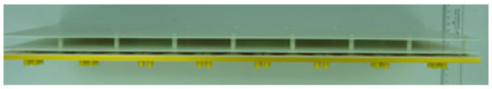

Side view

Fig.3 Novel structure DBDP-SA array

Compact Dielectric Resonator Antennas with Ultra-Wide 60%-110% Bandwidths

We have recently made a significant achievement in the emerging ultra wide-band (UWB) wireless communication systems, which require antennas with bandwidths gr 106%. In the past, the only way antenna engineers knew how to get such a bandwidth from a thin antenna was by removing the metal sheet underneath the antenna (known a plane”). This is not an acceptable solution for practical systems because the lack of it allows the antenna to radiate both upwards and (unnecessarily) downwards (i.e. electronic device on which the antenna is installed), wasting about half of its power. Saving power is crucial in UWB systems due to the severe power limits imposed by regu 2011, we made a breakthrough in dielectric-resonator (DR) antenna research, by inventing a novel DR antenna with a full ground plane and a 110% bandwidth. This disp myth that such bandwidths cannot be achieved with full ground planes. This antenna, published in the prestigious IEEE Transactions on Antennas and Propagation in Decem is 29% smaller but has a 30% greater bandwidth than the next-best DR antenna, which does not even have a full ground plane. Hence it is ideal for next-generation ultra-fas and sensor applications.

We theoretically and experimentally demonstrated that, by introducing a lower-permittivity full-length insert between the ground plane and a higher-permittivity dielectric dielectric resonator antennas (DRA) with ultra-wide bandwidths, in the range of 60%-110%, can be designed. Furthermore, the volume of such DRAs is reduced by appr 50% using a finite planar conducting wall. Unlike vertical monopole-type hybrid UWB DRAs, these antennas radiate sufficiently in the upward direction. Unlike in printed UWB antennas, the power radiated into the lower hemisphere is significantly less. An example prototype antenna, designed to operate in the FCC UWB band, has a dielectr of 12 x 8 x 15.2 mm3 (or 0.124 x 0.083 x 0.157 lambda3 at 3.1 GHz), and an average measured gain of 5 dBi from 3.1 to 10.6 GHz. These antennas exploit multiple low with overlapping bandwidths to form an ultra-wide contiguous bandwidth. With the proposed dielectric arrangement, it is possible to efficiently couple a sufficient numbe overlapping modes to a 50 ohm feedline using a single, simple feed.

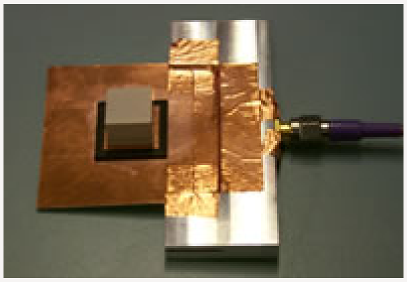

The antenna has a remarkably small footprint of 12x8 mm2 at 3.1 GHz - the lowest frequency of the FCC UWB band. Its dielectric volume is 1459 mm3, or 1.7x10-3 lamb lowest operating frequency of 3.1 GHz, and overall height is 15.2 mm or 0.157 lambda0. To place these results in perspective, it is worth comparing the new designs with wideband DR designs available in the literature. To the best of our knowledge, prior to this, the widest bandwidth ever obtained from a pure DRA design is 84% . The volu DR in that design is 0.225 x 0.172 x 0.062 lambda3 (= 2.4x10-3 lambda3) at its lowest operating frequency of 3.69 GHz. In that DRA, the DR is positioned in a non-traditi close to the edge of an orthogonal, “vertical” ground plane, which does not block radiation towards the lower hemisphere. The widest bandwidth demonstrated by a pure DI traditional “horizontal” ground plane, (which can be employed to shield the rest from the antenna, as discussed previously) is 78%. The dielectric volume of that design is 0. x 0.21 lambda3 (= 5.8x10-3 lambda3) at its lowest operating frequency of 6.7 GHz.

Super Wide-band Antennas

We have demonstrated that extremely wide bandwidths (ratio-bandwidths up to 1:25) can be obtained from a specially designed printed antenna with a tapered semi-ring design is described in “A Printed Elliptical Monopole Antenna with Modified feeding Structure for Bandwidth Enhancement,” in IEEE Transactions on Antennas and Pro vol. 59, no. 2, pp. 667-670, Feb. 2011.

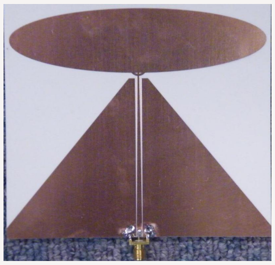

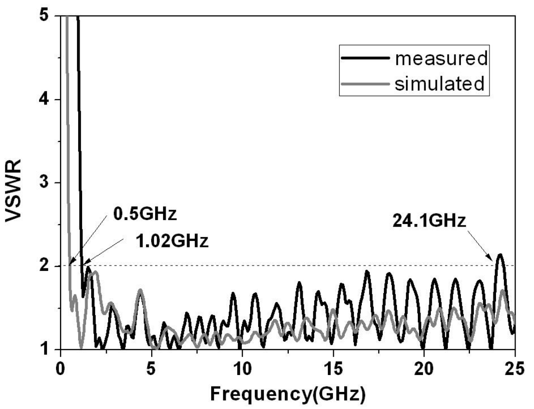

Focal Plane Arrays for Radio Astronomy

Dense focal plane arrays (FPAs) are a key technology for a new generation of Radio-telescopes. Their primary benefit is the rapid survey speed facilitated by the wide fiel
provided by multiple beams. Recent advances have brought dense FPAs within reach of radio astronomy applications. A number of institutions have significant research pr
this field. This technology is being considered for the Square Kilometre Array (SKA) (www.skatelescope.org). The PhD project of Douglas Hayman, conducted with CS
Centre and Division of Astronomy and Space Science, investigated beamforming aspects of FPAs and evaluated their performance in Radio Astronomy. A prototype interf
radiotelescope, built at CSIRO's Radiophysics Laboratory in Sydney, is used to demonstrate a suite of techniques forFPA beamforming and evaluation for this thesis. Bea
solutions were experimentally demonstrated in our paper in IEEE Transactions on Antennas and Propagation, entitled “Experimental Demonstration of Beamforming Sol
Focal Plane Arrays”.The THEA tile, shown below, is designed by ASTRON and used for the experimental component of this research.

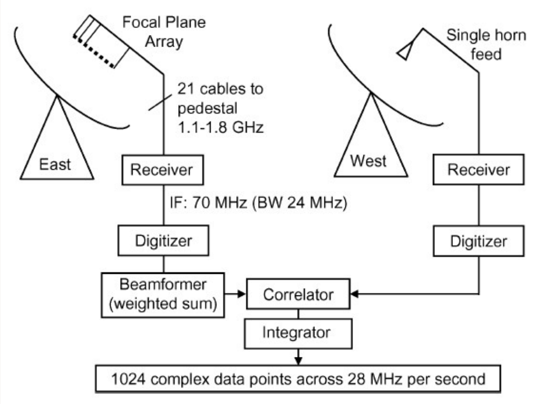

Negative Permeability of Spiral Metamaterials

Archimedean spirals and complementary Archimedean spirals are super-compact metamaterial particles. Thanks to their convoluted geometry, unit cells can be made electri
small. We theoretically analysed monofilar, bifilar, trifilar and quadrifilar Archimedean spiral metamaterial particles using point group theory and crystallography. From the s
properties electromagnetic response was determined. Magnetic, electric and magnetoelectric modes of the particles were identified along with their isotropy characteristics.
shown that all the particles, except monofilar spiral, are nonbianisotropic. Further, effective medium theory was applied to extract the effective permeability of the spiral med
results indicated negative values for permeability in certain frequency ranges. Detailed theory and numerical simulation results are available in the paper entitled “Analysis
metamaterials by use of group theory,” published in the Metamaterials Journal, vol. 3, no. 1, pp. 33-43, March 2009.

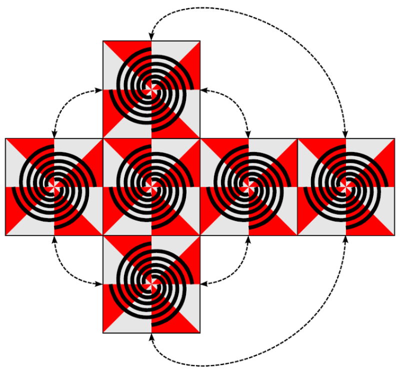

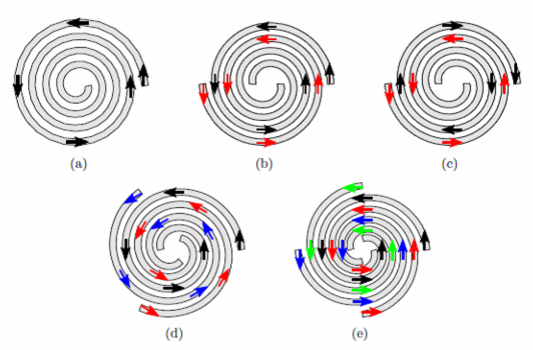

Archimedean Spiral Metamaterials and Backward Waves

We have shown backward wave propagation and double negative parameters over a 19% bandwidth in a microstrip line loaded with series gap discontinuities and super complementary Archimedean spiral resonator metamaterial particles. Moreover, our equivalent-circuit model for such unit cells almost perfectly described the structu practically important frequencies (by comparison with full-wave results). We also fabricated and tested compact filter circuits with only one or two complementary spiral met particles. Our results sre summarized in the paper entitled “Backward Wave Microstrip Lines with Complementary Spiral Resonators,” published in IEEE Transactions on and Propagation, Vol. 56, Issue: 10, pp. 3173-3178, Oct 2008.

We derived design equations for Archimedean spiral resonators and tested them against full-wave simulations. The details are in the paper entitled “Design of monofilar a Archimedean spiral resonators for metamaterial applications,” published in IET Microwaves, Antennas & Propagation, vol. 3, no.6, p. 929-935, Sep.2009.

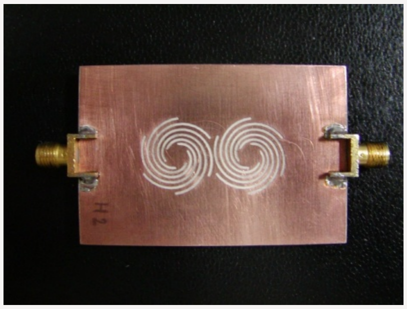

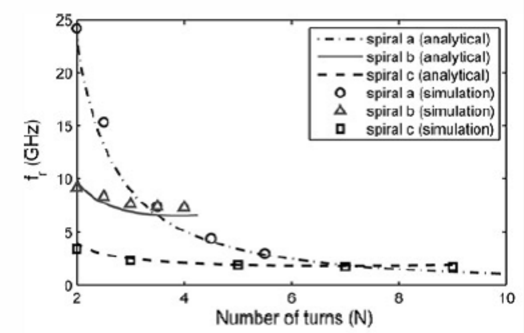

Frequency Selective Surfaces for Energy-Saving Glass Panel

Energy-saving glass is becoming very popular in building design due to their effective shielding of building interior against heat entering the building with infrared (IR) wav obtained by depositing a thin layer of metallic-oxide on the glass surface using special sputtering processes. This layer attenuates IR waves and hence keeps buildings summer and warmer in winter. However, this resistive coating also attenuates useful microwave/RF signals required for mobile phone, GPS and personal communication sy by as much as 30 dB. To overcome this drawback, we designed and tested a bandpass aperture type cross-dipole frequency selective surface (FSS), etched in the coatings c saving glass to improve the transmission of useful signals while preserving IR attenuation as much as possible. With this FSS, 15-18 dB peak transmission improveme achieved, for waves incident with 45 degrees from normal for both TE and TM polarizations.

Measurements and other results of this research, conducted in collaboration with the Lund University in Sweden, are available in the paper entitled “Cross-Dipole Frequency Selective Surface for Energy-Saving Glass Used in Buildings,” published in IEEE Transactions on Antennas and Propagation, vol. 59, no. 2, pp. 520-525, Feb. effect of these FSSs on the transmission of infrared and visible wavelengths through energy-saving glass was investigated theoretically and experimentally in another pap Microwaves, Antennas & Propagation, Vol. 4, Iss. 7, pp. 955–961, 2010.

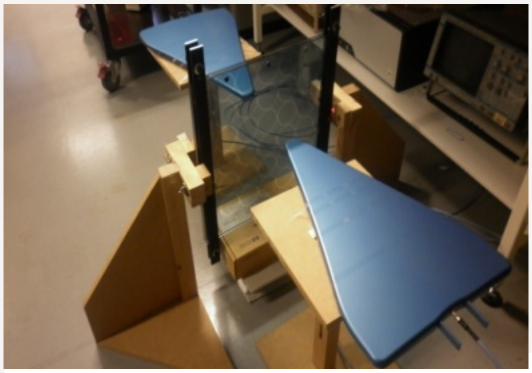

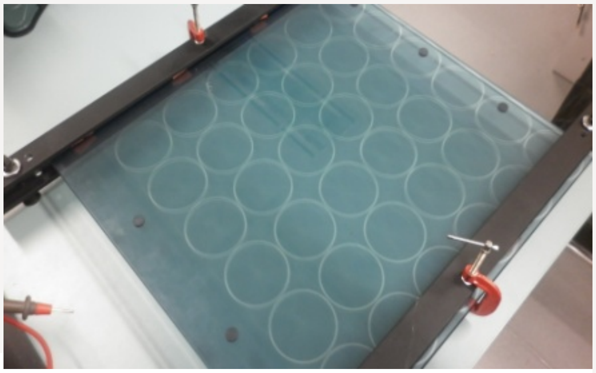

Switchable Frequency Selective Surfaces to Reconfigure Electromagnetic Architecture of Buildings

In large buildings and offices, frequency re-use methods will be required to enhance the spectral efficiency and capacity of wireless communication systems. This observatio to the concept of electromagnetic architecture of buildings. Passive bandstop FSSs can be used to enhance the electromagnetic architecture of a building, and hence to spectral efficiency and system capacity, but switchable FSSs can provide a better reconfigurable solution. If switchable FSSs are placed in strategic locations of a building, th reconfigured remotely and rapidly, which is not possible with passive FSSs.

With collaborators in UK and Sweden, we designed and successfully tested a single-layer active Frequency Selective Surface (FSS) that is electronically switchable between and transparent states. It can be used to provide a spatial filter solution to reconfigure the electromagnetic architecture of buildings. The FSS measurements show that the response of the filter does not change significantly when the wave polarization changes or the angle of incidence changes up to ±45º from normal. The FSS is based on sq aperture geometry, with each unit cell having four PIN diodes across the aperture at 90 degree intervals. Experiments demonstrated that almost 10 dB additional transmissio be introduced on average at the resonance frequency, for both polarizations, by switching PIN diodes to ON from OFF state.

For details, please refer to “Switchable Frequency Selective Surface for Reconfigurable Electromagnetic Architecture of Buildings,” in IEEE Transactions on Ante Propagation, Vol. 58, Issue 2, pp 581-584, February 2010.

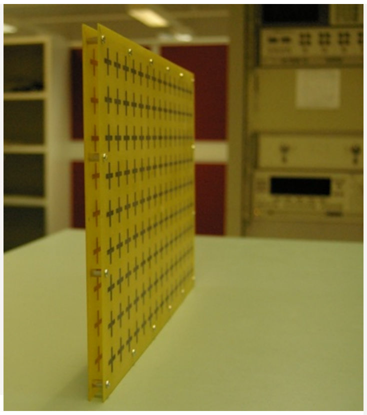

Absorb/Transit FSS

We designed and tested a novel absorb/transmit frequency selective surface (FSS) for 5-GHz wireless local area network (WLAN) applications. The novelty of the design i capable of absorbing, as opposed to rejecting, WLAN signals while passing mobile signals. The absorption of the WLAN signal is important to reduce additional multipa spread and resultant fading caused by typical reflect/transmit FSSs. Our FSS consists of two layers, one with conventional conducting cross dipoles and the other with resis dipoles. The FSS has good transmission characteristics for 900/ 1800/1900-MHz mobile bands and performs well for both horizontal and vertical polarizations.

Later we modified the FSS to obtain even better performance, for example, for both horizontal and vertical polarizations at oblique angles of incidence. The distance betwee layers has been successfully reduced to one eighth of free-space wavelength. This small distance makes it more compact as compared to the conventional Salisbury screen achieving an acceptable absorption in the stopband.

The details of our designs and test results can be found at "Oblique Incidence Performance of a Novel Frequency Selective Surface Absorber," in IEEE Transactions on Ant Propagation, Vol. 55, no. 10, pp. 2931 – 2934, Oct. 2007, and "A Novel Absorb/Transmit FSS for Secure Indoor Wireless Networks with Reduced Multi-path Fadin Microwave and Wireless Component Letters, Vol. 16 (6), pp. 378 - 380, June 2006.

Dual-Band Artificial Magnetic Conductor (AMC) Surfaces

AMC surfaces have many advantages and interesting properties due to their unique reflection characteristics, with near zero reflection phase. We have designed and pro novel dual-band AMC surface, which has a very wide upper AMC band and a narrow lower AMC band, and therefore suitable for multi-band wireless/mobile applications.

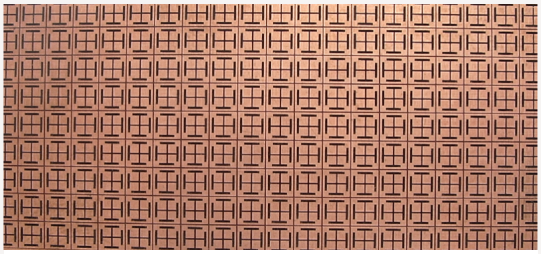

Fully Printed Quad-band Antennas for Wi-Fi IEEE802.11 and other WLAN Applications

Our fully-printed antennas can be fabricated and integrated to WLAN systems at almost zero cost by printing them on the same circuit board (e.g. FR4) with the radio circuit same standard fabrication methods. They are extremely compact: an antenna with a radiating element of 1cmx1cm covers all four IEEE standards (802.11a, 802.11b, 80 802.11n) as well as HiperLAN2 with a VSWR less than 2. We have also developed a packaging solution where the rest of the circuit can be shielded to satisfy EMC regulati leaving the antenna (on the same board) open for radiation. The advantages of the antennas based on this technology are: Lightweight; Radiates almost every direction in shadow region); Microstrip and co-planar waveguide (CPW) designs available; Compatible with all printed microwave circuits, including stripline circuits; Excellent because no cables, connectors, soldering or any mechanical attachments are required to connect the antenna to the radio; No protruding parts that are likely to break; Covers WLAN bands with one antenna (e.g. Wi-Fi IEEE 802.11 a, b, g, n, HiperLAN2 etc.); Excellent matching, i.e. input reflection < -10 dB (VSWR < 2) in all WLAN bands (e. GHz, 4.9-5.1 GHz, 5.15-5.35 GHz, 5.725 – 5.825 GHz); High efficiency (~60-70% on FR4; can be increased to over 90% on Duroid.); Ideal for internal mounting in a c device, to achieve very wide beam coverage; diversity reception or MIMO arrangements available; can be dsigned to cover other multiple bands (e.g. multiple mobile/c bands) and other applications.

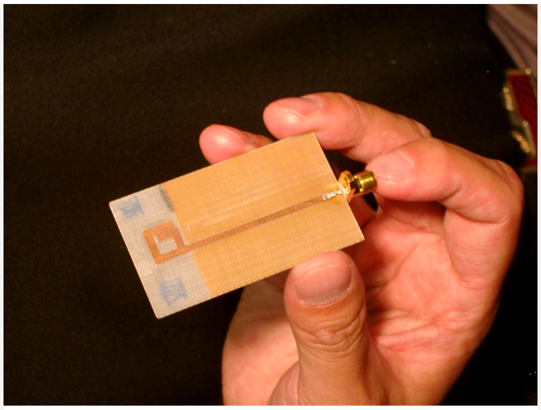

How can I optimise a UWB System to operate well over a range of directions?

Emerging ultra-wideband (UWB) communication devices will need to operate well, not just in one direction but over a range of directions. However previous UWB system optimisation techni been limited to one direction only. A system optimised in one direction may not work well in other directions due to pattern instability, or direction-dependent transfer function, of the an developed the concept of frequency-domain correlation patterns and proposed a figure of merit, the pattern stability factor (PSF), which can be determined from simulation of experimen With these new concepts, we have demonstrated the optimisation of UWB systems to operate well over a range of directions. The concepts and their applications are available in the PhD papers by Tharaka Dissanayake.

Antenna Techniques and Technologies for Ultrawideband (UWB) Systems

We developed antenna solutions for emerging UWB systems with very high data rates. Our research includes investigation of band-notching techniques to reduce UWB inference wi wireless systems and studies of antenna dispersion and pattern stability in the time domain.

High Gain Antennas with Planar Surface-Mounted Short Horns

In collaboration with University of Delhi, we have developed compact rectangular dielectric resonator antennas with surface mounted planar horns for gain enhancement. One configuration an aperture-fed rectangular dielectric resonator antenna and a planar surface mounted horn. We have achieved 10 dBi gain and a good efficiency from this configuration. The surface mou increases the gain of the standard dielectric resonator antenna by 4.9 dB. Total height of the prototype antenna shown in the figure is only 0.172 lambda and the aperture size is 0.96 lamb antennas can be easily adapted to low-profile, high-gain and array applications.

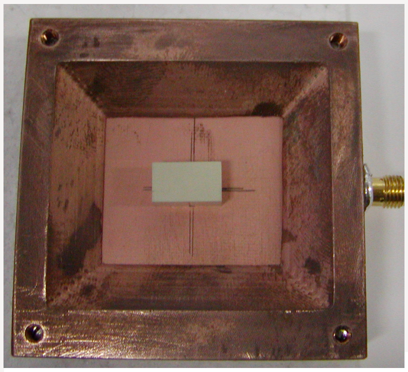

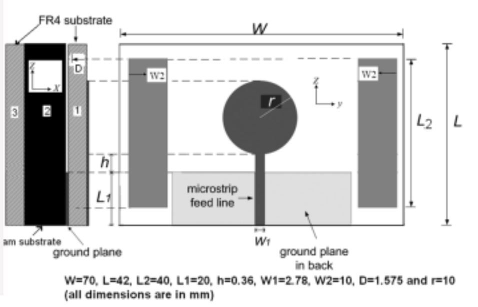

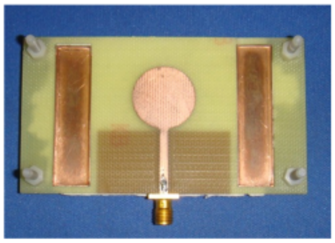

Photonic Crystal/EBG Based Horn Antennas

Waveguides and antennas made out of dielectric, as opposed to metal, are expected to perform better at Terahertz frequencies, as they do not suffer from the skin-effect loss of metal investigated horn antennas, horn arrays, waveguides, bends and junctions made out of 3D woodpile photonic crystals, which can be implemented at THz frequencies in high-resistivity S other materials. Our concepts have been tested by fabricating and measuring scaled-up prototypes operating in microwave frequencies. We have also designed a special broadband couple a photonic crystal waveguide to a rectangular waveguide or a coaxial cable, for example for testing using a vector network analyser. We demonstrated > 6% bandwidth both theore experimentally.

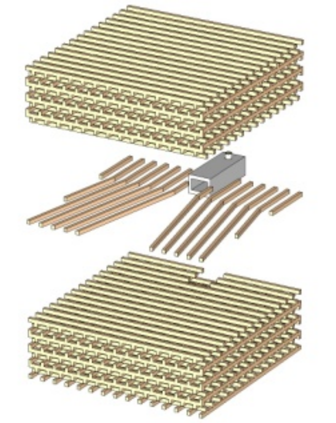

Broadband Microstrip Patch Antennas for Wireless Computer Networks

We have designed and tested broadband, compact E-shaped patch antennas for wireless communication networks, operating in frequencies from 4.9 GHz to 6.0 GHz. We made two achievements. First, we designed an E-shaped antenna, which is intrinsically compatible with a printed microstrip circuit. It can be made out of a single sheet of metal and can be mou microstrip circuit without expensive coaxial connector. Second, we designed a unique E-shaped antenna with corrugation to reduce the width of antenna. We successfully miniaturised the fit it inside a thin (4mm) PC (or PCMCIA) card extension. In fact, we were able to squeeze in two of these antennas into a single PC card of standard width (54mm), for diversity communica achieve excellent isolation (>20 dB) and matching (<-10 dB) over the whole WLAN frequency band.

Theoretical Analysis of Photonic Crystal Structures

We have developed and successfully implemented, in both personal computers and supercomputers, efficient theoretical methods to analyse and design complex electromagnetic (micro optical) circuits based on photonic crystals (PC). Among our novel techniques is a PC-based Perfectly Matched Layer (PML) absorbing boundary for use with the finite difference time doma method in the analysis of waveguides in 3D photonic crystals.

We have applied these techniques to model wave propagation in various guiding structures such as bends, junctions and tapers in 2D and 3D crystals. From this analysis, we can transmission and reflection coefficients, propagation characteristics, phase and delay response, etc., of a component or a system.

New Closed-form Green's Functions for Microstrip Circuits and other Layered Structures

When combined with the spatial-domain Method of Moments (MoM), our new closed-form functions now enable efficient and accurate analysis of microstrip circuits and antennas with approximations. The key feature in this new MoM is that the four-dimensional integrations in MoM matrix elements can be solved analytically, completely eliminating the need for expensive integrations. Our new closed-form functions are simpler and more flexible than previous such functions, and (unlike previous ones) they do not require additional (Taylor series approximations. The computer time required for the analysis of an example microstrip line using the new MoM was three times less than the next best method, which required some integrations.

Integrated-design of Hybrid-resonator Antennas for Broadband Wireless and other Communication Systems

Shown below is the first baby of our recent project on broadband hybrid-resonator antennas. This first prototype of a Dielectric Resonator on Patch (DRoP) antenna demonstrated a ba 24%. This design also proved, both theoretically and experimentally, that the electromagnetic fields in a dielectric resonator can be efficiently coupled to the fields in a patch resona perturbing the radiation characteristics of each resonator.

Singularity-enhanced Finite-different Time-domain (FDTD) Method for Diagonal Metal Edges, Strips and Films

We developed the world's first and still the only singularity-enhanced FDTD method for metal edges not parallel to the grid. The edges were assumed to be diagonal to cell faces. Compare conventional spit-cell model, the computer memory required for an FDTD analysis of a structure could be reduced by up to 27 times and the computing time could be reduced by up to without sacrificing the accuracy of results. On the other hand, when the same grid was used, the accuracy of results improved by a factor of more than 3 compared to the split-cell model, a of more than 7 compared to the staircase model. The new equations were stable in all tests, and even in most demanding tests when the computational speed was further maximised by the time step (Dt) to the maximum allowed by the FDTD method! (i.e. stability factor of 1!)

The key to this success was the derivation of enhanced FDTD equations for nodes near the edges by considering rigorously the singular nature of the electromagnetic fields. The table here improvement of accuracy achieved by using the enhanced FDTD equations. The new equations are simple to implement in a standard FDTD code. They are ideal for the analysis and microstrip components and high-speed digital circuits where thin metal films or strips with diagonal edges are encountered.

Low-profile Dielectric-resonator (DR) Antennas

We designed and tested the world's first low-profile, circularly polarised, rectangular dielectric-resonator antenna (DRA) in 1995. The radiating element of this antenna is shown in the photo have also designed many other dielectric-resonator antennas, including a low-profile linearly polarised DRA, for various applications (see publications). We pioneered the FDTD analy antennas and published the first radiation patterns of a DR antenna obtained using the FDTD method in 1995.

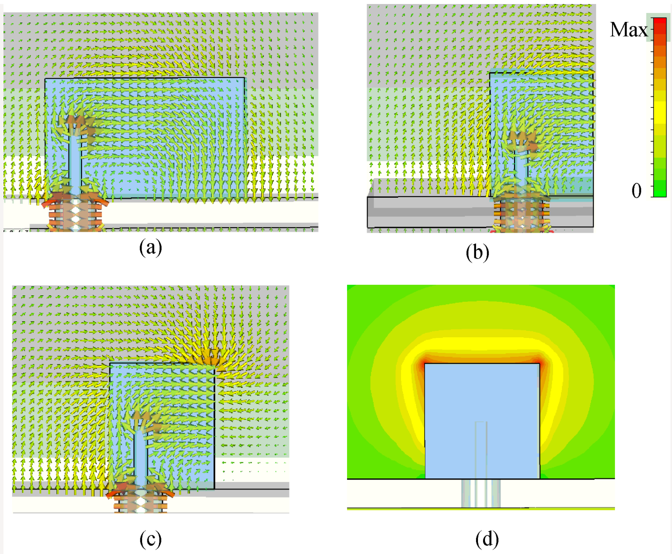

(a) (b) (c) (d)

Australian Antenna Measurement Facility (AusAMF)

AusAMF is operated by a consortium of Australian Universities, Industry and CSIRO. The facility provides access to a shielded anechoic chamber offeri spherical near-field measurement capability for small antennas operating in the frequency range of 1-20 GHz, primarily for research purposes.

The facility was established under an ARC LIEF (Australian Research Council Linkage-Infrastructure Equipment and Facilities) grant and is hosted by C ICT Centre located in the Sydney suburb of Marsfield.

The facility currently consists of a spherical near-field turn-key NSI-700S-50 system within a 6m x 3.3m x 3.3m anechoic chamber. It has been designed to operate up to fre of 20 GHz, and uses an Agilent PNA (E8362B) as a receiver. The measurement system is capable of supporting a ~40kg antenna under test (AUT) up to a diameter of approximately 1.2m. I the standard gain horns and the probes available at AusAMF are as follows:

Waveguide Band	Frequency (GHz)	Probe	Standard Gain Horn
WR975	0.75 - 1.12	x	x
WR650	1.12 – 1.70	x	x
WR430	1.70 – 2.60	x	x
WR340	2.20 - 3.30	x	x
WR229	3.30 – 4.90	x	x
WR159	4.90 – 7.05	x	x
WR112	7.05 – 10.00	x	x
WR90	8.20 – 12.40	x	
WR75	10.00 - 15.00	x	x
WR51	15.00 - 22.00	x	x

COPYRIGHT (C) 2013 KARU ESSELLE ALL RIGHTS RESERVED

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/273396385

Smart Surfaces: Large Area Electronics Systems for Internet of Things Enabled by Energy Harvesting

Article *in* Proceedings of the IEEE · November 2014

DOI: 10.1109/JPROC.2014.2357493

CITATIONS
24

READS
362

9 authors, including:

Luca Roselli
Università degli Studi di Perugia
273 PUBLICATIONS **2,132** CITATIONS

SEE PROFILE

Federico Alimenti
Università degli Studi di Perugia
200 PUBLICATIONS **1,263** CITATIONS

SEE PROFILE

Paolo Mezzanotte
Università degli Studi di Perugia
144 PUBLICATIONS **1,211** CITATIONS

SEE PROFILE

Giulia Orecchini
Università degli Studi di Perugia
24 PUBLICATIONS **347** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

Development of ultra wide band analog predistorsion of LASER diodes for Radio over Fiber (RoF) systems View project

Project

ADAHELI View project

All content following this page was uploaded by Luca Roselli on 30 July 2015.

The user has requested enhancement of the downloaded file.

INVITED PAPER

Smart Surfaces: Large Area Electronics Systems for Internet of Things Enabled by Energy Harvesting

This paper focuses on large-area "smart surfaces," RFID systems, and wearable RF electronics that could substantially benefit from multisource energy harvesting.

By LUCA ROSELLI, *Senior Member IEEE*, NUNO BORGES CARVALHO, *Senior Member IEEE*,
FEDERICO ALIMENTI, *Senior Member IEEE*, PAOLO MEZZANOTTE, *Member IEEE*,
GIULIA ORECCHINI, MARCO VIRILI, *Student Member IEEE*,
CHIARA MARIOTTI, *Student Member IEEE*,
RICARDO GONÇALVES, *Student Member IEEE*, AND PEDRO PINHO, *Member IEEE*

ABSTRACT | Energy harvesting is well established as one of the prominent enabling technologies [along with radio-frequency identification (RFID), wireless power transfer, and green electronics] for the pervasive development of Internet of Things (IoT). This paper focuses on a particular, yet broad, class of systems that falls in the IoT category of large area electronics (LAE). This class is represented by "smart surfaces." The paper, after an introductory overview about how smart surfaces are collocated in the IoT and LAE scenario, first deals with technologies and architectures involved, namely, materials, antennas, RFID systems, and chipless structures; then, some exemplifying solutions are illustrated to show the present development of these concurrent technologies in this area and to stimulate further solutions. Conclusions and future trends are then drawn.

KEYWORDS | Energy harvesting; green electronics; Internet of Things (IoT); large area electronics (LAE); radio-frequency identification (RFID) systems

I. INTRODUCTION

Looking at the telecommunication (TLC) market development worldwide in the last five years, we see an average decrease of about 5% per year, in both gross domestic product (GDP) contribution and employment [1]. Against this trend, the global information and communication technology (ICT) market has remained grossly constant. Some compartments, in fact, are experiencing an opposite trend. Beside some sectors related to new consumer products, such as tablets and smartphones, new Internet of Things (IoT) related products are growing at two digits per year and some estimations from big players report a market value in the order of trillions of dollars in the next decade [2], [3].

The vision behind IoT is in fact to connect objects directly to the Internet so as to allow them to provide information directly to the web without any human intermediation [4]. This vision will have a great impact on several electrical and electronic technologies, ranging from the basic physical layers (technology platforms) to the highest ones: communication protocols, software, human interface, and information management.

Manuscript received May 4, 2014; revised September 1, 2014; accepted September 5, 2014. Date of publication October 13, 2014; date of current version October 16, 2014.

L. Roselli, **F. Alimenti**, **P. Mezzanotte**, **G. Orecchini**, **M. Virili**, and **C. Mariotti** are with the Department of Engineering, University of Perugia, Perugia 06123, Italy (e-mail: roselli@ieee.org).
N. Borges Carvalho and **R. Gonçalves** are with the Instituto de Telecomunicações-Dep. Electrónica Telecomunicações e Informática, Universidade de Aveiro, Aveiro 3810-193, Portugal.
P. Pinho is with the Instituto de Telecomunicações, Instituto Superior de Engenharia de Lisboa, Lisboa 1959-007, Portugal.

Digital Object Identifier: 10.1109/JPROC.2014.2357493

0018-9219 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

In terms of physical layer, IoT actually inherits the technologies developed for wireless sensor networks (WSNs), bringing the distributed nature of that to the extreme. WSN, in fact, can be synthetically seen as a mesh of sensor nodes purposely conceived to build and communicate a sensed image of a monitored area. Within IoT, the sensing nodes are not purposely deployed to monitor something specific; instead, they are simply hosted by "objects" in order to provide the information they are able to collect and make it available on the web. Conceptually speaking, there is not a big difference, but in terms of related technological challenges, there is. Nodes must be hostable by objects; objects are of many types, and the better the nodes fit themselves transparently to the hosting objects, the denser and more reliable the information they are able to provide.

Several challenges can be envisioned in this evolution: first, electronic systems must be mechanically flexible, thin, and miniaturized in order to conform to the shape of as many objects as possible; second, the adopted materials have to be as recyclable as the hosting objects in order to avoid pollution from guest apparatuses; last but not least, hosted nodes must be autonomous, because they cannot be either connected to the grid or powered by life batteries.

A great technological paradigm shift is going to be pulled by the development of IoT; green materials, autonomous systems, ultralow-power circuits, energy saving protocols, and energy harvesting (EH) are concurrently mandatory.

Conversely, this technological evolution is pushing new solutions and architectures, so far constrained by the limits of conventional technologies. The development of inkjet printing techniques, the introduction of very cheap and eco-friendly materials compatible with roll-to-roll (R2R) circuit realization techniques [5] and so forth, open new horizons also to large area electronics (LAE).

LAE, first applied to printed photovoltaic and organic screens [6], is actually at the onset of its development. The development of R2R techniques and related materials, in fact, is allowing for tremendously increasing the dimensions of LAE circuits and systems, from the present tens of centimeters to meters and beyond. Along this evolutionary scenario, new configurations and architectures, based on massive integration of large circuits over conformable surfaces, can be envisioned, opening the way to what can be called the smart surfaces (SSs) approach.

SS, in turn, can be seen as a branch of IoT evolution. Large 2-D arrays of autonomous sensor nodes, for instance, make possible granular tracking of whatever parameter, ultimately providing augmented imaging of the environment; large 2-D arrays of tags can provide a very low-cost platform for precise localization and location-based services (LBS), thus enabling the realization, for instance, of smart floors [7] and smart wallpapers [8]; inheriting quasi-optical approach [9] contact-less electromagnetic (EM) wave processing (filtering, frequency conversion, selective shielding, etc.) can be conceived even at low frequencies.

On the one hand, SS is thus an approach stimulated by technologies pulled by IoT; on the other hand, it contributes to a class of architectures, within LAE, that can even widen the huge horizons of IoT applications in a sort of virtuous circle.

According to this wide vision, the paper is organized as follows. A review of the technologies suitable for distributed systems implementation to a large extent is given in Section II; then, the explanation about how radio-frequency identification (RFID) can be considered one of the most suitable technologies for the implementation of IoT architecture, and how it can be naturally integrated with SSs, is provided (Section III). In order to deal with the implications of the development of RFID systems compatible with SS and LAE evolution of IoT, two specific sessions have been provided: the first relates to antenna implications (Section IV), and the second relates to electronic architectures for RFID tags (mainly chipless ones; Section V). In order to smoothly bring the reader from relevant technologies to applications, quasi-optics, as an example of general approach concurrently exploiting the mentioned technologies to provide a platform suitable for LAE, is described in Section VI. After this, some application examples of how SS concept can be articulated, according to the IoT paradigm, are described; specifically, smart floor (Section VII), smart shoes as useful subsystems for the implementation of smart floors (Section VIII), and energy skin (Section IX). In order to have a vision of the logical links behind the many topics dealt with in this paper, a synoptic picture can be found in Fig. 1.

II. TECHNOLOGY FOR LAE DISTRIBUTED SYSTEMS

Given the described scenarios of SS, it is clear that the technologies involved have to be compatible with LAE

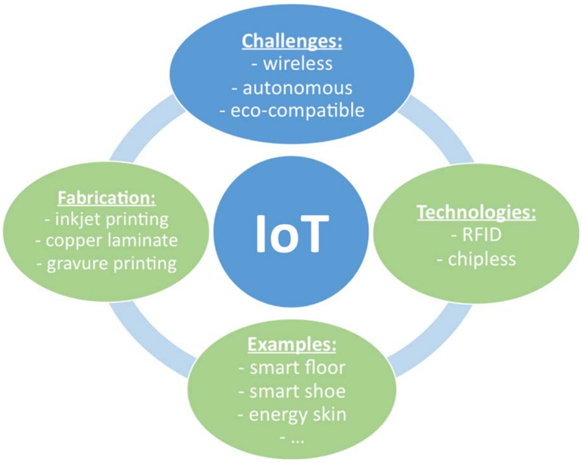

Fig. 1. *Synoptic view of the paper structure reflecting the IoT vision described in the Introduction.*

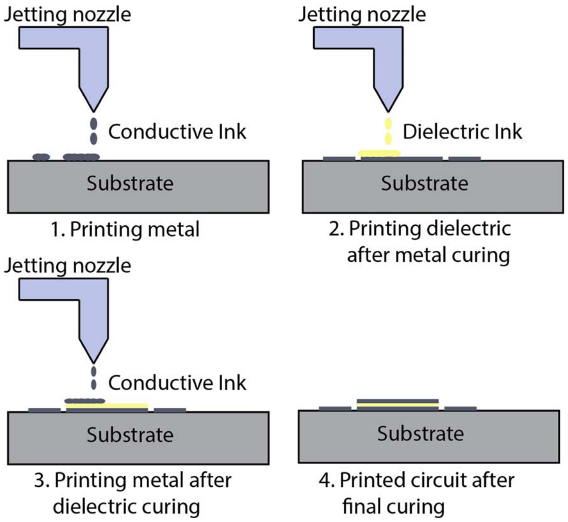

Fig. 2. ***Inkjet printing technique, synoptic description. The ink is deposited on the substrate with the desired pattern and after a curing procedure the next layer can be printed. The thickness of each film depends on the number of passes.***

approach, as described in the Introduction; according to [6], for example, the eco-compatibility of the materials, the mechanical flexibility of the substrates to make the device conformable to the hosting surface, and the life cycle (operability and biodegradability) have to be accounted for.

The new techniques proposed from different research groups [10]–[12] focus mostly on the printing techniques that can be industrialized easily with R2R methods, already investigated for the traditional printing on paper. Within the printing methods, one of the most attractive and investigated in the last decade is the inkjet printing of conductive and dielectric layers (Fig. 2). This technique allows circuits to be printed on almost any kind of substrate: from the photographic paper to liquid crystalline polymers (LCP) or Kapton, that are flexible, thin, and eco-compatible; to glass or poly methyl meta acrylate (PMMA, commonly referred to as Plexiglas) that are usually thicker but not flexible.

A first proof of concept for the inkjet printing technique at radio frequency (RF) and microwaves has been obtained on photographic paper by printing simple structures, such as antennas and other planar circuits and devices [10], [13]–[22], with a technology stack simply composed by the substrate and the conductive layer (usually realized by means of nanoparticle silver ink).

Recently, this method has been improved thanks to the development of new inks, usually composed by a polymer and a solvent (i.e., SU-8, PVP, PMMA, PEDOT, etc.); more complex fabrication procedures of multilayer structures are now possible, as shown in [23] and [24]. Another noticeable feature of this technology is represented by the resolution: conductive tracks of 50 μm of width and space gaps of 50 μm can be printed, allowing the design of millimeter-wave (mm-wave) frequency devices [25]. The combination of the quite high resolution (considering the simplicity of the technology) with the multilayer featuring also gives the opportunity to manufacture very easily matrix of passive and active devices that can be used in the SS development. It is worth noting that the process can be developed in a few steps based on the design and mostly on the number of layers needed. For example, a metal–insulator–metal (MIM) device can be realized by printing silver on the substrate [26], curing it in the oven in order to create a surface that, with the proper treatment (for instance, UV–Ozone exposure or preheating), is ready for the insulator printing. Then, the dielectric is cured as well and the last metal layer can be printed on top of it. The entire structure can be then cured at temperatures ranging between 130 °C and 200 °C.

In a perspective of industrialization of the inkjet printing technology, the hypothesis of R2R manufacturing of circuits is being investigated and some examples are already reported in literature for solar cells and other devices [27], [28].

Today, many pros have been mentioned for the inkjet printing method, and others can be found in the non-use of wasted chemicals (as is for the traditional lithographic technologies), in the no-need of clean-room environment for the fabrications, in the low-cost and rapidity of manufacture, and in a R2R perspective compatibility. However, the necessity of a curing procedure after the printing of a layer still represents the biggest inkjet limit for two main reasons: first, the time and type of curing are dependent on the inkjeted materials and on the material stack-up; second, especially on a laboratory level, the curing can cause misalignment issues, given the fact that the sample is removed and then replaced in the printer after the last layer curing.

Currently, the platform mostly used worldwide by researchers is the Dimatix 2800 DMP. In terms of inks, it is possible either to buy printable solutions or mix solutions in labs; for example, the combination of polymers and solvents allows layers to be printed with different electrical properties and thicknesses.

Besides the inkjet printing method, in [11], a new technique, suitable for LAE, has been proposed. It uses a metal (copper in that case) adhesive laminate technology based on the application of the standard etching process to an adhesive copper laminate material. This technique was already adopted to produce mm-wave circuits and to characterize the resulting compound substrates, as reported in [29]–[31]. A brief illustration of the metal laminate technique is here reported referring to Fig. 3.

The production process can be divided in five steps. The first step consists of the deposition of a (positive) photoresist film on the copper surface; then, the circuit layout is transferred to the photoresist by means of a photomask and ultraviolet exposure. After that, the unimpressed film of photoresist is removed using a NaOH

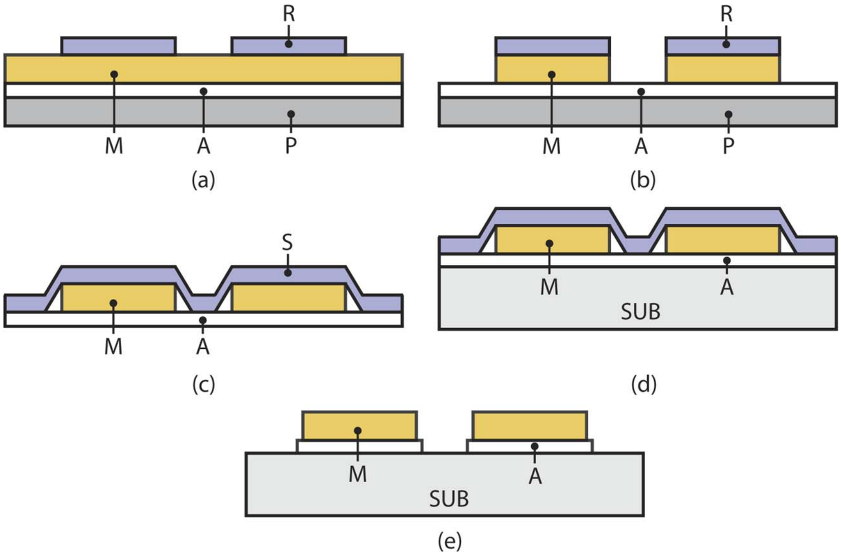

Fig. 3. ***Process steps for the circuit fabrication using the adhesive metal laminate. M: metal, A: adhesive, P: protection, R: photoresist, S: sacrificial layer, SUB: hosting substrate. After the standard lithographic procedure, the copper pattern can be transferred on top of any substrate by means of a sacrificial layer and exploiting the adhesive face of the copper tape.***

developer solution [Fig. 3(a)]. In the second step, the copper tape is wet etched. As can be seen in Fig. 3(b), in this way, the adhesive layer is exposed where the copper was removed, while it remains everywhere else covered by the original copper tape that serves as protection for the adhesive underneath.

The first two steps are, in this example, similar to those adopted in standard photolithographic technology. Moreover, different ways to remove the not needed metal can be used, as, for example, by means of numerical control pattern cutting plotters.

In the third step, depicted in Fig. 3(c), a sacrificial layer is stuck on the top copper side and, finally, the protection layer on the bottom is removed. The sacrificial layer is very important because it keeps the relative distances among the layout features constant even when these are not physically connected.

The fourth step is characterized by the transfer of the etched metal to the hosting paper substrate and, finally, the sacrificial layer can be removed [see Fig. 3(d) and (e)]. The last step is also useful to remove most of the exposed adhesive material.

With this method, quasi-fully-organic circuits and devices can be realized. The performance in terms of tracks width and pitch are, at present, a bit lower than what can be obtained with the inkjet printing, with the advantages of no curing processes, the possibility to fix devices on the circuit using standard soldering techniques, and a better conductivity (the conductivity of copper laminate, in fact, is one order of magnitude higher than that of the cured nanoparticle silver ink: 5.8×10^7 S/m versus about $6 \times 10^6 - 1 \times 10^7$S/m obtained with at least five layers of silver).

To verify the validity of the metal laminate technology at microwave frequencies, a 50-Ω microstrip line was manufactured exploiting the Mitsubishi photografic paper as the substrate (thickness 250 μm, relative permittivity $\varepsilon_r = 3.2$, and loss tangent $\tan\delta = 0.08$). The line is 30 mm long, and the measured scattering parameters are shown in Fig. 4. The same graph also compares the performance of a similar transmission line manufactured with an inkjet printing process. The performance of the

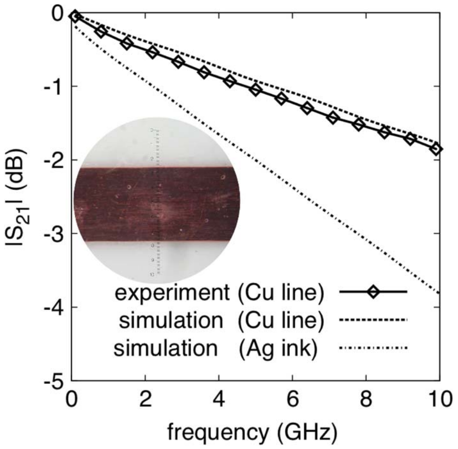

Fig. 4. ***Measured scattering parameters versus frequency for a 50-Ω microstrip line on paper substrate. The line, shown in the inset, is 30 mm long; the ruler division corresponds to 17.2 μm. The graph also reports a comparison with a microstrip line of equal dimensions made with an inkjet printing process (Ag ink, 3-μm thickness, $\sigma_{\text{ink}} = 1.1 \times 10^7$ S/m after curing). After [11].***

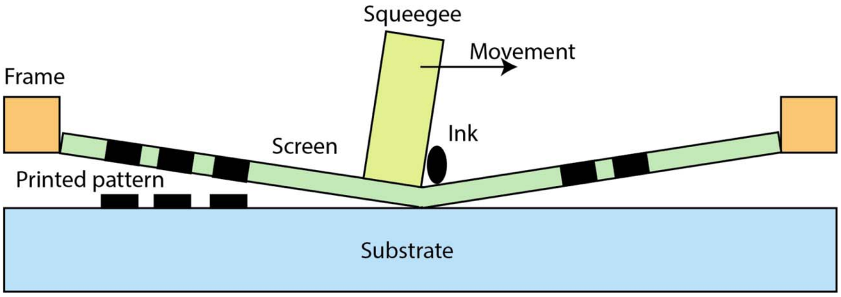

Fig. 5. ***Screen printing technique. The squeegee is used to press the screen with impressed pattern and transfer it on top of the substrate.***

metal laminate structure is superior compared to that of the inkjet printed line. At 10 GHz, the measured specific losses for the metal laminate microstrip are about 0.6 dB/cm [11]. Results at higher frequencies are also reported in [31]; in particular, an insertion loss of about 1.8 dBm/cm at 30 GHz is demonstrated.

The tradeoff is the adoption of a photolithographic step; it is worth mentioning, however, that the process is still compatible with R2R implementation.

Other possibilities for LAE are the screen printing [12] and the gravure printing [12], both of them adaptable to R2R industrialization techniques.

Screen printing consists of dragging a layer of ink across the surface of a screen and squeezing it through the open pores of the patterned mesh onto the substrate (Fig. 5). In general, the thickness of the printed layer and the achievable resolution depend on the density of the mesh and on the ink properties. Usually, the ink viscosity is in a range of 1–50 Pa × s, and this allows for printing with a resolution of about 100 μm and a thickness up to 100 μm. Until now screen printing has been adopted to realize low-resistance structures, solar cells, and field effect transistors (FETs), exploiting the possibility of printing very thick layers.

Gravure printing, also known as rotogravure, is considered a very high volume printing (Fig. 6) process, and it is being adopted to produce very conductive structures as, for example, capacitors, antennas, and organic devices such as pentacene-based diodes, organic light-emitting diodes (OLEDs), organic field effect transistors (OFETs), and organic thin film transistors (OTFTs).

The resolution can be of about 20 μm. The method is implemented by engraving the patterns into a metallic cylinder by a laser, by chemical etching, or electromechanically (as separate cells or intaglio trenches). Typically, the print pressure is high (1–5 MPa) in order to achieve a good ink transfer and to reduce the percentage of unprinted dots caused by the surface roughness of the flexible substrate. However, this high pressure of printing makes this technique suitable only for robust substrates with no soft, previously printed, layers that could be damaged.

To summarize, the technologies described in this section are characterized by some common features, such as: flexibility and conformability, compatibility with large area realization of circuits and interconnections, use of additive or mixed (subtractive/additive) deposition techniques, easy use of eco-friendly materials and, in some cases, compatibility with classical bond wiring as well as soldering techniques for electronic device placement.

Table 1 summarizes the main features of the described technologies.

III. RFID: A MIX OF CONCURRENT TECHNOLOGIES FOR SS IMPLEMENTATION

As mentioned in the outline paragraph at the end of the Introduction, after describing the technologies enabling SS development (Section II), here we illustrate RFID as a suitable means to transfer information through EM transmission between tags and readers to the Internet.

An RFID tag has a unique identification code and a memory used to store information, while the reader can

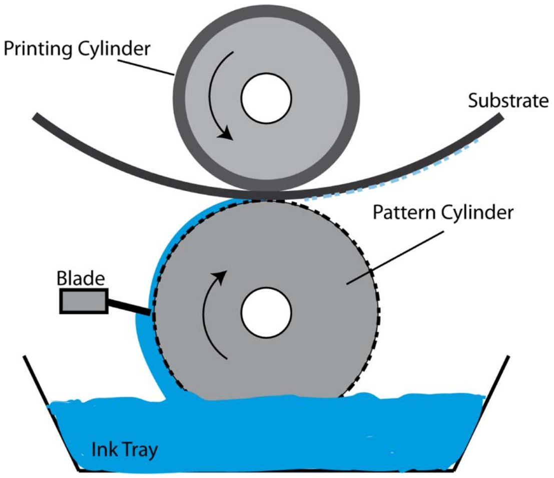

Fig. 6. ***Gravure printing process. Note that these machines can be from 20 cm to 3 m wide with a diameter ranging from 5 cm to 5 m.***

Table 1 Technologies Comparison

Technology	Resolution	Cost	Waste	Speed
Inkjet	50 μm	Low	Low	Fast
Copper	100 μm	Low	High	Medium
Gravure	20 μm	High	Medium	Very fast
Screen	100 μm	Medium	Low	Very fast

write or read data on and from the tag through RF transmissions. The tag is usually attached to an object that needs to be identified and monitored, or contains information to be read. In typical RFID applications [32], the main goal is to provide mapping of physical objects that are equipped with an RFID tag attached or embedded. In this way, RFID technology inherently leads to identifying, tracking, and localizing, but it can also enable the storage and transmission of information regarding object status and surrounding environments by embedding sensing capabilities.

As mentioned previously, IoT is most commonly described as a structured system of technologies concurrently working to capture meaningful data from objects and communicate them through networks to a decision-taking level. Useful keywords, whenever a definition of IoT is given, include: smart objects, autonomous logistics, machine-to-machine communication, RF technologies, centralized information, and so on.

In order to implement these functionalities, an IoT platform, an SS for instance, has to provide at least unique object identification together with wireless communication for data management.

The IoT physical layer can thus be naturally thought of as meshes of RFID and sensor nodes. The RFID tag that embeds a sensor can use the same working principles and protocols of the conventional RFID tags; a differentiation can be made, considering the way sensor tags are powered up, in active, semipassive, and passive tags; actually, the difference consists of the way they are powered.

Active tag sensors may use customarily conceived communication protocols or rely on RFID standards enabling the tag to be easily integrated into the existing wireless infrastructures so that they will not require expensive readers. In [33], a system architecture was developed to integrate WSNs and the RFID systems. Bluetooth and ZigBee technologies are adopted as the communication protocol of WSNs to meet the requirements of a large number of sensor nodes, large areas, and low cost.

Active tags with integrated sensors are used in several applications, including temperature and position monitoring, vibration detection, blood pressure, heart beat monitoring, and more. This type of sensor tags, besides having larger amount of energy provided by batteries, affords both a large range and multiple functionalities; some of them have also external buses that enable the use of connected external sensors [34].

Passive tags with an integrated sensor operate without battery, collecting the needed operational energy from external environment sources. The main requirements of this class are high energy conversion efficiency, large storage capability, and overall low-power consumption. Passive systems have usually low operating range and limited processing capability.

In this field, research activities are directed toward ultralow-power design of integrated sensor tags [35], antenna design for improving reading range [36], sensor overall performance optimization [37], and development of optimum protocols allowing for additional power saving [38], [39].

Typical applications are, for instance, temperature monitoring [40], photodetection [41], motion detections [42], etc.

Semipassive tags use both batteries and energy sources coming from the environment; they can integrate more operational capabilities than the passive ones, since they can exploit a higher amount of energy. For the same reason, semipassive RFID tags are more suitable to integrate a sensor; the operational methodology is similar to the passive one, using the reader signal to interrogate and cause a response from the tag. The primary difference is that the semipassive tag does have a battery used not to generate a response, but only to power electronics, like sensors incorporated in the tag, while exploiting RF circulation to reply to the reader interrogation. Just like passive tags, semipassive ones are limited in terms of slow reading speeds and short reading distances. In [43], the design, realization, and experimental validation of a battery-assisted RFID tag with sensing and computing capabilities, conceived to explore heterogeneous RFID-based sensor network applications, was presented.

Dealing with battery-less devices that are, by far, the most interesting ones for IoT, LAE, and SS applications, we can now focus on the typical energy sources adopted. No battery devices can exploit solar scavenging [44], [45], vibrational [46], [47], or RF radiations [48]–[52].

In many cases, wirelessly powered systems can harvest energy from the incident RF waves, thus generating the required direct current (dc) voltage to power the system. The EM energy present in the environment may be significant in some regions (some urban areas or indoor sites); here the use of proper conversion devices may allow enough energy to be scavenged. RF power harvesters usually consist of an antenna, an impedance transformation network, a rectifier, and a storage element. In [53], a design example of integrated rectifier antennas (rectennas) for wireless powering at low incident power densities, from 25 to 200 μW/cm^2, is given.

Regarding mechanical energy scavengers, they may be categorized upon the kind of energy source exploited; there are scavengers providing energy from a constant motion over extended periods of time (e.g., turbine air flow) or others based on intermittent motion (e.g., human step). Particular attention is given to vibrational energy sources; in this case, the amount of energy that can be harvested depends mainly on the vibration amplitude and frequency and on the vibrating mass (mass of the harvesting device) [54]–[56].

In order to power the systems by using energy sources available in the working environment, proper energy transducers thus have to be chosen. Focusing on distributed systems (IoT, LAE, and SS), it is worth referring to typical parameters for optimum transducer design since, in these applications, the different kinds of energy available in the various scenarios may lead to the need of integrating more than one energy conversion system into the same device, in order to take advantage of all the accessible energies. The main design parameters are the available power output, the electrical impedance, and the operating voltage. Table 2 summarizes them for the most common energy transductions [57], [58].

To conclude this section, the mentioned reasons why RFID can be considered a useful technology for the implementation and development of LAE and, more specifically for SSs, are summarized. RFID systems, especially in the passive version of tags, are communication systems characterized by the following characteristics: compatibility with very low-power communication protocols, compatibility with the constrains posed by the hosting objects, especially when they are implemented by adopting the technologies described in Section II, inherent capability to gather information from the environment (sensor tags) and from the hosting objects, a low number of electronic devices (inherent circuital simplicity) and, finally, proven compatibility with printing technologies.

For all these reasons, RFID technology has to be considered the leading one to pursue the development of LAE and SSs as a communication platform for the evolution of IoT.

IV. ANTENNA DESIGN FOR NODES

Distributed systems for ubiquitous electronics (UE) and WSNs [59], [60] are, today, increasingly employed for monitoring and sensing applications. These systems can be seen as networks of nodes, massively distributed in the environment and communicating wirelessly with a base station. This massive deployment can be designed and structured to monitor a set of specified parameters by means of purposely conceived sensor nodes, as in the classical WSNs, or it can be obtained by integrating nodes into existing objects, which is the vision of IoT evolution. In general, the nodes can have different functionalities, however they have to provide information collected by embedded sensors (such as, for instance, identification codes, position and monitored environment parameters as well as object status) whenever it is required.

In this scenario, the RFID technology, summarized in Section III, is attractive. When RFID technology meets sensor nodes, information goes from the node to the Internet via a question-and-reply protocol between nodes, hosted by tagged objects, and readers.

Thinking about distributed systems and specifically SS, it is natural to think about radiative coupling in the available industrial–scientific–medical (ISM) frequency bands, including mainly the ultrahigh frequency (UHF) and microwave frequencies (around 0.9, 2.4, 5.8 GHz, and higher), exploiting the communication standard such as the WiFi [61]. Recently, also the ultrawideband (UWB) has been considered in order to minimize the limitations of narrowband systems (mainly localization accuracy and sensitivity to interference) [62].

Multistandard systems are often adopted for large compatibility and higher use flexibility. An example of this recent approach is given in [61] where near-field UHF nodes are combined with far-field transceivers. This allows application fields with compromised radiation capability,

Table 2 Transducer Characteristics for Several Forms of Ambient EH [57], [58]

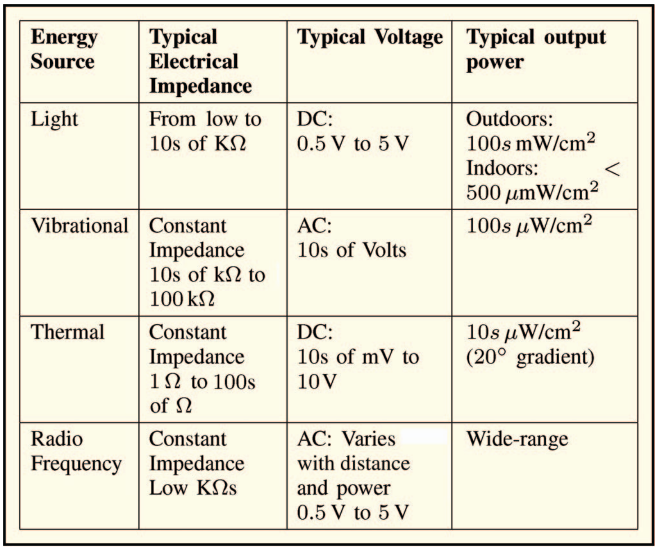

Energy Source	Typical Electrical Impedance	Typical Voltage	Typical output power
Light	From low to 10s of KΩ	DC: 0.5 V to 5 V	Outdoors: 100*s* mW/cm^2 Indoors: < 500 μmW/cm^2
Vibrational	Constant Impedance 10s of kΩ to 100 kΩ	AC: 10s of Volts	100*s* μW/cm^2
Thermal	Constant Impedance 1 Ω to 100s of Ω	DC: 10s of mV to 10 V	10*s* μW/cm^2 (20° gradient)
Radio Frequency	Constant Impedance Low KΩs	AC: Varies with distance and power 0.5 V to 5 V	Wide-range

such as in proximity of metal and liquids at short ranges, to be improved.

It is clear that antenna systems are key elements in these apparatuses. The foundation of antenna design rules can be found in [63]. The antenna design specific for distributed systems, in turn, has to take into account several instances. Beyond the typical design criteria, other aspects are: the communication standard, the form and the size in agreement with the application, the communication range as a function of the equivalent isotropic radiated power (EIRP) limits, of the environment and of the orientation, the reliability under several conditions and processes, and the material cost considering massive production [64]–[67]. Among them, the operating environment is the most critical aspect that influences the node behavior [68].

As a consequence, antenna topology and materials for SS and IoT systems, in general, have to be chosen in order to make the antennas capable of operating in variable environmental conditions and to be integrated into objects. As an example, let us think about wearable antennas for body area networks (BANs): they must be washable, flexible, and able to shield the body from radiation [69].

In general, antennas for SS should be flexible, planar, and low profile; moreover, by adopting low-cost materials and technologies (see Section II), cost per unit surface is low, thus large arrays and big elements can be adopted more easily, ultimately allowing for easy developments of LAE distributed systems and SS, even at low frequency.

Materials compliant with SS implementations must be, in turn, mechanically robust, flexible, and very low cost, while exhibiting acceptable EM performance. These materials must also exhibit high tolerance levels in terms of bending repeatability. Flexible materials, often used, are polymers and polymer-ceramics composites such as the poly-di-methyl-siloxane (PDMS) and the titanate-based ceramic [68], which provide flexibility, robustness, and resistance to harsh environments in order to protect antenna and circuit. Moreover, when eco-compatibility has to be guaranteed, recyclable materials and related technologies have to be adopted; among them, cellulose substrates in combination with inkjet printing of conductive inks [18], [70] or with conductive adhesive laminate technology [11] are increasingly frequently adopted. The aforementioned techniques, compatible with the well-known industrial printing and the R2R processes [5], are in agreement with the low-cost fabrication and are expected to facilitate widespread and very low-cost electronics for distributed systems. These fabrication technologies can also be adopted for electronic devices on flexible polymeric substrates. An example of cellulose-based antenna for smart floor applications, such as indoor localization, is documented in [71]: a low-profile planar loop antenna is realized on cellulose substrate exploiting adhesive copper tape.

Although beyond the scope of this paper, which focuses on SS, it is worth mentioning here that this evolution toward the development of more and more sophisticated "unconventional" materials for electronics has a great impact on other IoT related developments such as, for instance, implantable systems.

When applications require devices with soft visual impact (think about SS, such as smart windows, glass integrated inside and outside buildings and glass for automotive applications, or smart skins [72]), transparent and conductive materials, such as transparent conducting oxides (TCOs), and in particular indium-tin-oxide (ITO) and ITO-based multilayers, have been proposed [73]. These materials afford a good compromise between minimum electrical resistivity and high optical transmittivity in the visible light spectrum when patterned on see-through materials such as glass or transparent polymers. A technique to improve the efficiency of the transparent antennas, without affecting their transparency, is the application of a layer of highly conductive coating or metallization strips on the antenna edges where the current density is high [72].

Among SSs, or in general distributed systems within the IoT vision, a fast growing field of research is represented by wearable electronics for BANs [74]. In this case, the integration of electronic devices in the garments requires the development of wearable and washable antennas and antenna arrays on fabrics. Washable antennas can be realized by covering textile antennas by a breathable thermoplastic polyurethane coating in order to protect the device against water absorption and corrosion [69].

Although dimension constraints are not very critical for distributed systems on cheap materials, they cannot be neglected thoroughly. At RF, for example, high permittivity materials [68] and magnetic composite substrates [75] can be adopted. Artificial magnetic conductors (AMC) can be employed alternatively to the magnetic composites with similar results [76]. These approaches are simply demonstrated from the wavelength equation: wavelength is a function of the relative permittivity and permeability and adopting suitable materials to increase these values, the effective wavelength, and therefore the antenna size, decreases. Anyway, the use of these materials introduces a tradeoff between dimensions and performance: generally the substrate losses decrease the antenna radiation efficiency, thus affecting the communication range of the overall wireless system, as documented in [68], [75], and [76].

In terms of design methodology, the introduction of unconventional materials as well as the environmental constraints posed by distributed systems and systems for IoT influence and drive the design strategies. Some solutions are proposed in the literature to cope simultaneously with these constraints; in [68], shielded configurations have been analyzed to mitigate environmental sensitivity. In [69], an example of robust patch antenna for BAN application that reduces the effect of the body on the performance is illustrated; in [77], the use of broadband antennas, such as the Chebyshev monopoles with a fractional bandwidth of about 40%, is proposed to reduce the frequency shift introduced by the background materials or by the bending effects.

In SSs, and in general in IoT applications, antennas are used both to transmit and receive information and to harvest the necessary energy from the surrounding environment. Regardless of the fact that the available energy is randomly released in the environment by existing services (GSM, WiFi, and so on, in urban and semi-urban environments [78], [79]) or purposely delivered to empower a distributed population of nodes [80], antennas have to be able to harvest EM energy. This kind of "double effect" antenna is widely referred to in the literature as "rectenna": it combines rectifying circuitry with EM to electrical energy transduction.

The fundamental codesign techniques of antennas and rectifiers, focusing on the improvement of the conversion efficiency exploiting matching circuits, are described in [53]. Harmonic-balance and load-pull simulations are often used in order to find the optimal impedances that allow best efficiency to be obtained as a function of the EM incident power.

Multiband, broadband, and circularly polarized antennas are often adopted in EH applications to reduce the overall sensitivity to the environment. An example of planar antenna array for EH applications, in agreement with the SS principle, is documented in [81]: spiral antennas are proposed in order to provide a wide operating frequency band, and the rectification circuitry is optimized and characterized for low incident power levels. Similar antenna array for UWB EH is shown in Fig. 7.

A rectenna can be used in addition to the communication antenna or, alternatively, the same antenna can be used for both EH and data transmission. The use of different antennas is generally adopted if harvesting and communication are carried out at different frequencies and, mainly, in order to enhance harvesting efficiency without decreasing communication performance. In any case, this requires a codesign and can affect the node dimension and complexity. Alternatively, a single antenna can be a feasible compromise if communication and harvesting share the same frequency band and if the antenna harvesting efficiency is not the main issue; examples are the passive RFID tags empowered by the reader with a standard RFID antenna able to provide both EH and communication capability, as documented in [64] and [65].

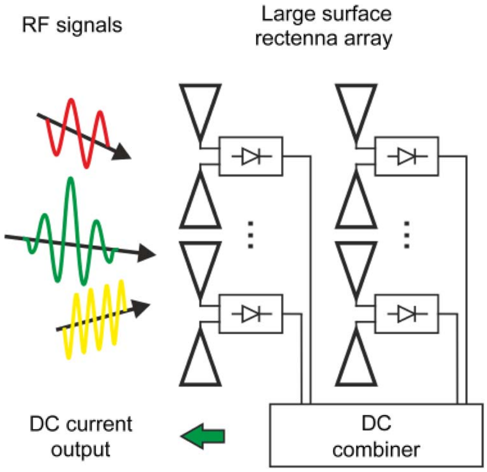

Fig. 7. ***Scheme of the rectenna array able to collect RF energy even in a broadband (depending of the kind of receiving antenna element), according to the solution adopted in [81].***

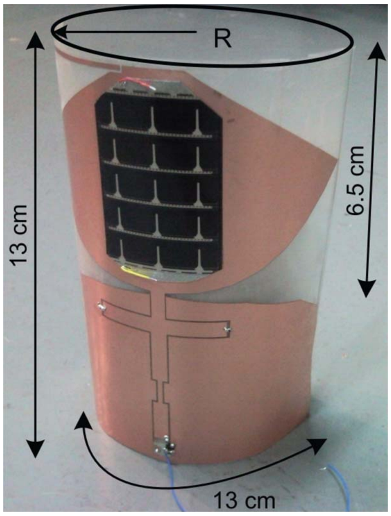

Fig. 8. ***Hybrid rectenna for solar and EM–EH. The hybrid rectenna consists of a broadband antenna and a solar cell combined in the same structure. After [82].***

However, the EM–EH is not the only solution to energize nodes and SS elements: hybrid EH, in fact, can be performed by exploiting different energy sources; that is, combining the rectennas with other sources of energy.

Recently, a low-cost and conformal structure for hybrid EH, based on low-cost and flexible poly-ethylene terephthalate (PET) substrate and an amorphous silicon solar cell, has been proposed in [82]; other examples are discussed in [20] and [83]. The use of low-cost, printable photovoltaics deposited on flexible substrates to form part of the antenna radiating structure is documented in [84]. In all these contributions, the main challenge appears to be the codesign of the antenna and solar cell in order to avoid reciprocal detrimental effects. The antenna is designed with a solar cell integrated on top of the radiating surface to obtain good radiation characteristics and to incorporate connections for the extraction of the dc provided by the solar cell. In [82], it was demonstrated that, by adopting thin film solar cells and paying attention to not exceeding the perimeter of the RF antenna, the solar cell can be integrated so as not to significantly affect the performance of the antenna. The hybrid rectenna for solar and EM–EH documented in [82] is shown in Fig. 8.

In terms of industrial assembly, the use of new materials and the need for low-cost and large-scale fabrication techniques would be fostered by architectures more insensitive against geometrical errors and placement misalignments. Following this direction, EM coupling can be exploited instead of ohmic contact to connect antennas on a flexible substrate to Si chips or, more generally, to the active circuitry [85]. This technique exploits a planar heterogeneous transformer, with the primary and secondary windings implemented on the antenna substrate and the Si chip, respectively. In this way, the galvanic contacts and the soldering process are avoided, and the final structures have been proven much more robust to alignment errors than traditional ohmic techniques without any significant performance degradation. By artificially applying a misalignment of 150 μm, which is the common dimension of a soldering pad, the technology proposed in [85] causes a loss increment of only 0.1 dB, due to the applied misalignment, while the same misalignment, when adopting ball grid arrays, would have implied missing the contact. This approach is quite attractive especially in combination with nonrigid materials where mechanical stress can result in significant substrate deformations. Recent contributions show the application of the aforementioned approach to inkjet printed antenna on PET substrate coupled to the chip for wireless nodes [86] and a textile patch antenna, magnetically coupled to the active circuitry, suitable for garment integration (see Fig. 9) [87], the reflection coefficient of which is shown in Fig. 10.

In order to testify the robustness of the proposed solution against uncertainties inherent to textile and garment electronics assembly, the radiation efficiency of the antenna, as a function of horizontal and vertical misalignment of the two transformer windings, used to perform the magnetic coupling, is plotted in Fig. 11(a) and (b), respectively. This technique can be extended also to interlayer communications of multilayer devices, and an example is documented in [88], where the circuitry fabricated on different garments communicates by means of overlapped coils.

V. CHIPLESS APPROACH

It has already been discussed that SS represents a possible technology for the development of killer applications within the IoT paradigm. Thanks to the convergence of several technologies, ranging from EH and sensing to RFID and LAE, SS can afford complex functions. For example, SSs can be used in intelligent buildings for structural monitoring and alarms (fire, for instance). To achieve these goals, SS must be equipped with a huge number of tags, each one monitoring a small portion of the surface itself.

The passive RFID systems can be divided into two families, namely, chip-based and chipless tags. The first family usually exploits a low-power complementary metal–oxide–semiconductor (CMOS) chip to implement the main tag functions [RF carrier rectification, dc voltage regulation, amplitude shift-keying (ASK) demodulation, IDentification code (ID) decoding, load modulation, etc.] and thus can be adapted to sensor applications by a few additional circuit blocks [signal conditioning,

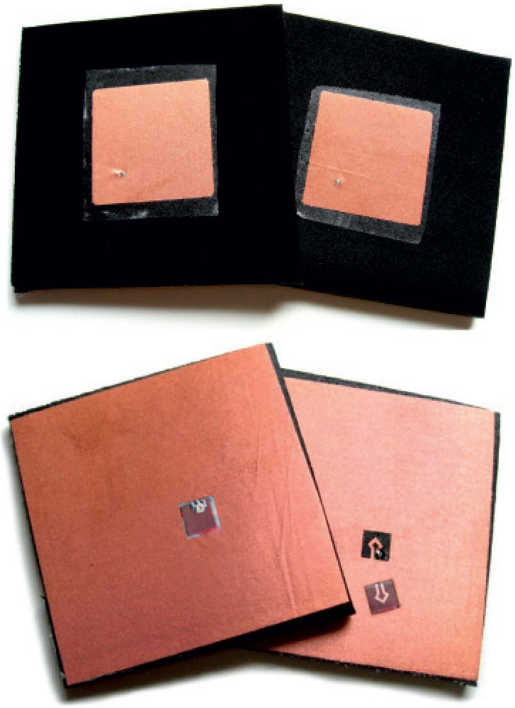

Fig. 9. ***Example of textile patch antennas operating at 2.4 GHz. The antenna and the ground plane are made of Flectron, a conductive textile, and the substrate is black foam. These materials are flexible and integrable in the garments. Magnetic coupling is exploited to connect the antenna and the active circuitry on flexible substrate. After [87].***

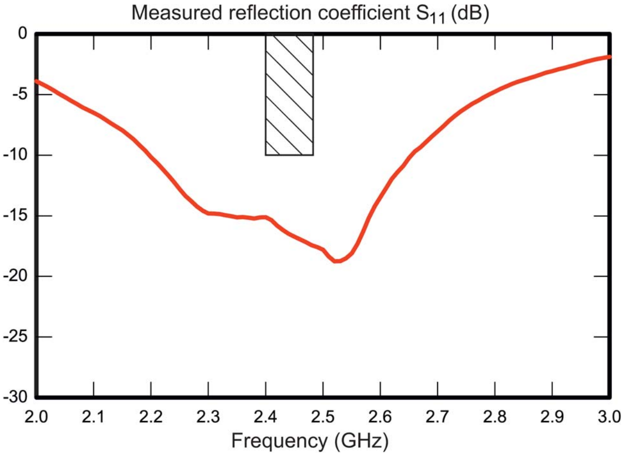

Fig. 10. ***Measured reflection coefficient of the previous textile antenna in the frequency band 2.4–2.48 GHz. After [87].***

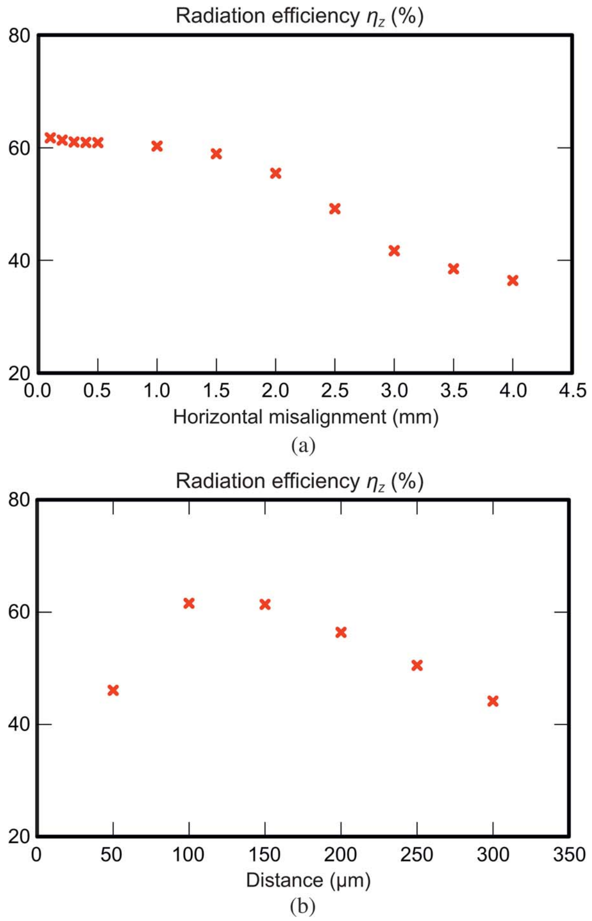

(a)

(b)

Fig. 11. ***Variation of the radiation efficiency evaluated versus (a) horizontal and (b) vertical misalignment between the antenna and the circuit windings. This test allows the robustness of the antenna performances against mounting uncertainties inherent to textile electronics to be evaluated at glance. After [87].***

analog-to-digital converter (ADC), etc.]. The main advantage of such an approach is the digital modulation of the transmitted signal, and thus the flexibility of the related data treatment, as shown in [89] and [90]. The production costs of chip-based tags are mostly associated with the heterogeneous integration of the silicon chip with the antenna [85], [91], the latter being manufactured typically on a flexible substrate or a textile substrate [92]–[97].

In order to reduce the aforementioned costs and to save as much energy as possible, solutions requiring minimum amount of electronics must be pursued. For this purpose, the chipless tag family was introduced and applied to wireless sensing in the recent years [98]–[103]. It is worth recalling here that the introduced buzzword "chipless" stands for a family of tags that are actually able to implement their functionalities by using only a few lumped components and passive distributed elements, without requiring any electronic integrated circuit (IC), the chip indeed, to be mounted.

The standard chipless tags (i.e., only coding and no sensing) exploit, basically, two mechanisms, namely, time-domain and frequency-domain scattering [64]. In the first case, the elapsed time between multiple reflections is used for coding. These reflections are obtained, at tag level, by connecting the antenna to a transmission line structure with several discontinuities. In the second case, instead, the tag has a specific frequency signature that is decoded at the reader level; to implement such a behavior, a coded series of resonators is realized on the tag antenna in order to reflect or not the corresponding frequency. Both approaches (time and frequency domain) lead to tag circuits that are not necessarily easy to miniaturize; nevertheless, it is worth noting that they are compliant with the LAE paradigm.

When dynamic sensor information has to be added to the static identification, we talk about chipless RFID sensor tags; these exploit an antenna, the electrical properties of which are controlled by the change of the physical parameter to be detected. There are, mostly, two proposed approaches for this. The first one is to induce a permanent change in the antenna property when a certain critical threshold (acceleration, temperature, fluid level, etc.) is exceeded [104]; this is useful for alarm type operations. The second one exploits a sensing load, i.e., a load the impedance of which is controlled by the sensed variable, connected to an antenna [105]. In both cases, the whole wireless sensor system (tag and reader) needs to have absolute accuracy, thus limiting the system performance with respect to both distance and fabrication tolerances.

A different method has been recently introduced in [106]. In this paper, a novel sensing principle is associated to the generation of an intermodulation signal from a tag, the latter being illuminated by two waves at different frequencies. The advantage of this idea is that the tag response is generated at a known frequency, thus the presence or the absence of such a signal can hardly be misinterpreted. Similar techniques have been used in harmonic radar systems [107] and in one-bit frequency doubling tags [108].

In line with the above ideas, a novel and completely original approach was proposed in [20] and [109] to solve the issue of absolute accuracy of the most of the passive chipless RFID sensors. The chipless tag described in [109] is based on the harmonic radar concept [110], [111], i.e., on a tag that, being illuminated by a carrier at frequency f_0, is capable of generating the second harmonic $2f_0$.

A simplified block diagram of the tag is illustrated in Fig. 12. For this purpose, the sensor information is encoded as the phase difference between two signals, one acting as the reference signal for the other one. First, the tag receives a carrier at frequency f_0. Then, two equal signals at frequency $2f_0$ are generated by means of a diode-based frequency doubler and a power divider. At this point, one of the two signals is phase-shifted using a passive sensing element. Finally, the $2f_0$ signals are reirradiated by

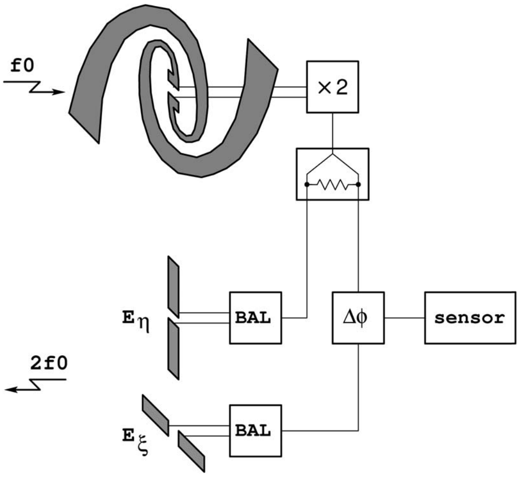

Fig. 12. ***Simplified block diagram of a harmonic chipless tag with sensing capabilities. After [109].***

exploiting two orthogonally polarized antennas. With this approach, the sensor information can be extracted by a suitable reader equipped with two complex (I/Q) receivers.

It must be observed that chipless tags, due to their inherent simplicity and extremely low number of lumped components, are easily realizable on flexible substrates by means of metal laminate [29], [30] and inkjet printing technologies. In particular, antennas [112], diodes [113], and passive sensing elements [114]–[117], i.e., all the main tag components, have already been experimented in cellulose-based materials. This will soon make SSs, equipped with sensing chipless tags, feasible at affordable prices.

VI. QUASI-OPTICS

"Quasi-optics concerns the propagation of EM radiation when the size of the wavelength is comparable to the size of the optical components (e.g., lenses, mirrors, and apertures) and hence diffraction effects become significant" [118].

In optics, operations are usually performed by using lenses. At EM frequencies, by following the quasi-optics analogy, lenses become large arrays, several wavelengths in size, of unconnected elements, each of them being able to get part of the signal incident on the array (usually focused by using a Gaussian beam), to carry out an operation and to irradiate the modified signal. The sum of the radiated signals modified by the array elements ends up in a radiated beacon elaborated with respect to the incident one [118]–[120]. From a historical point of view, quasi-optics was first introduced experimentally by Heinrich Hertz in the late 1880s [120] when he demonstrated the possibility of collimating EM signals on a multiwavelength surface by using cylindrical reflectors and studying effects that, until that time, were observed only at infrared and visible EM spectrum.

After more than 70 years, owing to the increasing demand for bandwidth and to the development of optical wavelength systems, quasi-optics experienced its palingenesis at the millimeter and submillimeter wavelengths.

We can classify the quasi-optical components into three main categories: frequency independent, frequency selective, and active devices. Frequency-independent surfaces include: delay lines, polarizing grids, hybrid junctions [121], attenuators [122], power dividers [123] and combiners [124], nonreciprocal devices [125], [126], absorbers, and calibration loads. Frequency-selective surfaces [127]–[129] include: inductive grids, capacitive grids, resonant grids, thick structures (perforated plates), and interferometers. Active devices include: oscillators [129], [130], amplifiers [131], [132], mixers [133], [134], phase shifters [135], [136], multipliers [108], [137]–[139], and switches [140]. This approach has been used so far mainly at millimeter and submillimeter frequencies because, being the size of the apparatuses proportional to the wavelength of the radiated wave, lower frequencies would have been prohibitive due to the very large size required. With the present development of technologies for LAE, however, quasi-optics has become a feasible approach even at RF and microwave bands, where large arrays of operational elements can be conceived.

Inheriting the quasi-optical approach, contact-less EM wave processing (filtering, frequency conversion, selective shielding, etc.) can be implemented even at low frequencies (in the order of megahertz). Two pioneer examples are switches [140] and the cross-dipole frequency doubler, proposed in [108], and implemented in [141], where, among various operating principles, the generation of harmonics is chosen to demonstrate the feasibility of such a component at microwave frequencies.

The layout of the proposed quasi-optical frequency doubler is shown in Fig. 13. The structure is inkjet printed on a cellulose-based substrate (a piece of photografic paper from Kodak). It consists of two crossed $\lambda/2$ dipoles. The longest dipole receives the incoming power at the fundamental frequency $f_0 = 3.5$ GHz, while the shortest one transmits the generated power at the doubled frequency $2f_0$ in an orthogonally polarized orientation. The length of the dipole operating at f_0 is 32 mm, while the $2f_0$ dipole is 16 mm long. The width of the tracks used to implement the dipoles is about 2 mm. The multiplication is provided by four diodes in a bridge configuration, thus forming a fully balanced multiplier unit.

Due to the omnidirectional nature of the dipole antenna, the harmonic signal is irradiated with the same intensity in the azimuthal plane (small differences are due to both the dielectric substrate and the planar nature of the dipole conductors). This means that the harmonic signal is reflected toward the reader and the transmitter toward other directions at the same time. A measurement of the received power at $2f_0 = 7$ GHz is reported in Fig. 13(b). Here the interrogation distance is 10 cm and the

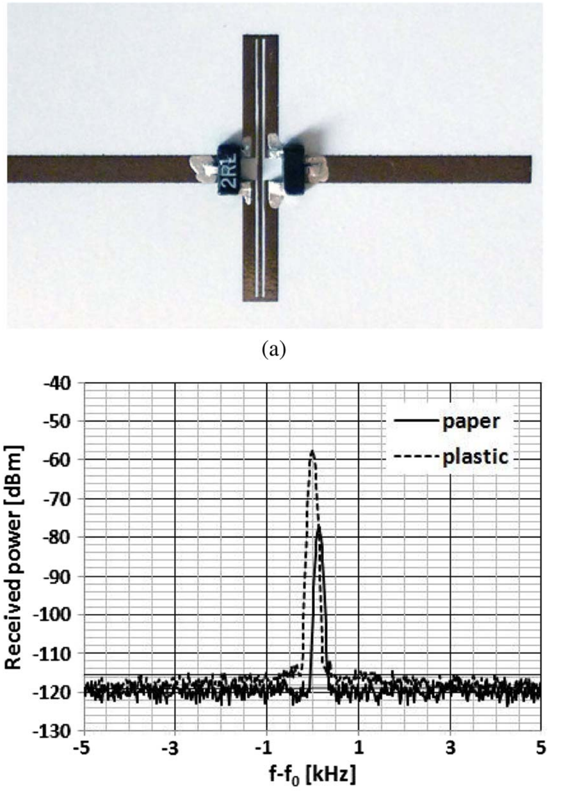

(a)

(b)

Fig. 13. ***Cellulose-based prototype of the (a) crossed-dipole frequency doubling tag and (b) measured second-harmonic response. The measurements have been carried out interrogating the tag at $f_o = 3.5$ GHz and receiving the $2f_o = 7$ GHz frequency component. The interrogation distance is 10 cm, and the transmitted power is equal to 20 dBm. The transmitter is equipped with a two-element Yagi antenna at f_o, while the receiver uses a helix antenna at $2f_o$. The fundamental frequency dipole is 32 mm long. After [141].***

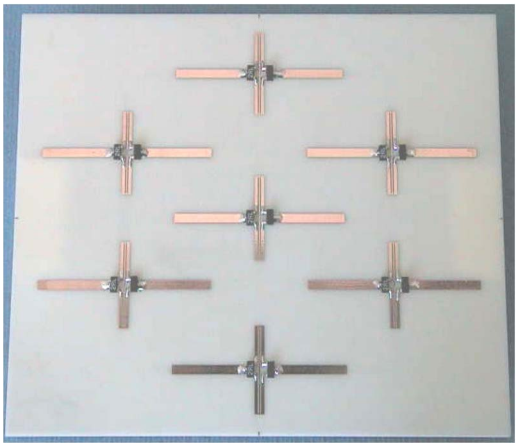

Fig. 14. ***Example of quasi-optical frequency doubler based on an array of crossed dipoles. This structure is implemented on a Rogers substrate, but this can be substituted with a cellulose-based material to reduce costs and obtain mechanical flexibility.***

transmitted power is equal to 20 dBm. The transmitter (reader side) is equipped with a two-element Yagi antenna at f_0, while the receiver (reader side) uses a helix antenna at $2f_0$. In these conditions, the power received from the cellulose-based (paper) tag prototype at $2f_0$ is about -80 dBm. The same structure implemented on a plastic substrate produces more power, mainly because of a better frequency tuning. An improved efficiency could be achieved by: 1) adopting a frequency multiplier with better harmonic terminations; and 2) using directive antennas in order to address the power only toward the reader.

In the frame of the IoT, the crossed-dipole tag is a very simple one-bit tag that can alarm a system when it is placed within the range of the reader. However, with very simple modifications, this idea can be used to implement a variety of on/off sensors (the paper substrate can be easily torn when a certain mechanical strain is exceeded).

The crossed-dipole structure can also be used as a building block to form arrays, an example being shown in Fig. 14. These arrays are mainly intended for LAE applications where a complete surface can be interrogated and answered by generating a harmonic signal. The fabricated prototype uses a Rogers substrate, thus it is not flexible. However, adopting a cellulose-based substrate, a flexible structure can be built. The bending capability of cellulose-based circuits is quite good and only limited by the discrete components mounted on them.

The second example consists of a paper-based contactless frequency doubler for harmonic RFID applications [142]. The doubler, realized on paper substrate, generates the harmonic signal by means of a single Schottky diode. The system operates at 7.5 and 15 MHz, and these frequencies are chosen, without lack of generality, to accomplish the realization of a fully organic frequency doubler exploiting paper printed coils and organic diode (pentacene-based), the present frequency limit of which is around 15 MHz [143]. Fig. 15 shows a picture of the organic tag, while Fig. 16 shows the doubling efficiency of the organic tag versus transmitting/receiving (TX/RX) distance, assuming TX power as a parameter (for the sake of completeness, it is worth noting that TX and RX antennas of the reader are equal to the RX and TX antennas of the tag, respectively).

As a final remark, it is worth noting that, to the authors' knowledge, only planar developments of quasi-optical arrays have been proposed so far; reasonably, combining quasi-optics with LAE related technologies can lead, in the near future, to new SS solutions only in a limited planar extent.

VII. SMART FLOOR

The new concept of smart cities is generating a group of new scenarios that will impose a new way people interact with the environment. This immersion brings the IoT to a new level of relationship with people and environment,

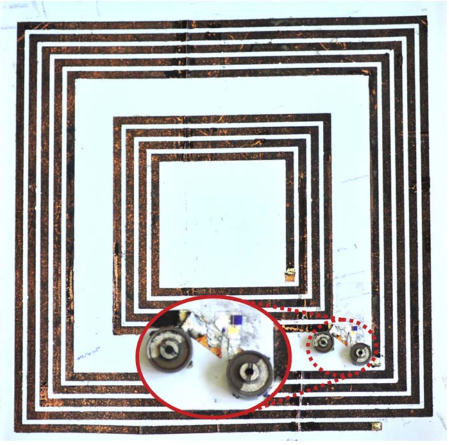

Fig. 15. ***Fully organic tag prototype of the frequency doubling tag operating at 7.5–15 MHz. After [142].***

where the scenario around is aware of people's needs and can interact with their way of living.

This new concept brings IoT to areas that cover broad topics, including transportation, energy distribution, data communications, and all technologies in a sense that they will become transparent to the normal users and citizens.

The growth of smart cities will significantly increase the quality of life, the reduction of energy waste, and the availability of information by using ICT. All these developments have, as a motivation, the overall inclusion of people into the city in a way that ICT should become completely embedded into our surroundings.

These new concepts are viable only if the smart environments are enabled by IoT. IoT is actually taking shape and is growing everyday through the number of devices that get connected to each other.

Nevertheless, most of these devices that team up with each other to build up the IoT world are powered by some kind of power sources. This is actually the main limitation of present IoT solutions. In order to be able to create real IoT environments, RFID, wireless power transfer (WPT), and EH devices appear as a major enabling technology for the IoT, due to their inherent simplicity and ability to provide sensing to remote objects without the need for constant powering.

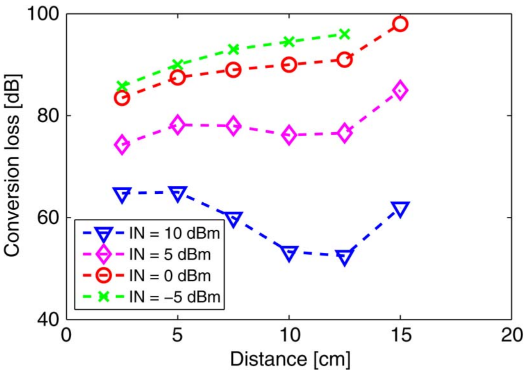

Fig. 16. ***Conversion (doubling) loss of the reader tag harmonic RFID system as a function of the reader tag distance, assuming the output reader power as a parameter. After [142].***

In this scenario, the concept of a smart floor becomes real. Smart floors should be capable of interacting with the environment seamlessly and be aware of persons and objects on top of them.

The focus of this section is going to be smart floors as an information system, where RFID capabilities are used to provide an identification and information source embedded into the floor. This massive RFID immersion will allow to create low-power high accuracy location, navigation, and, generally speaking, an information system that can team up with passers in a noninvasive way.

In this approach, the system uses passive RFID tags that are spread beneath a flooring in order to create a map that can be read with a mobile unit that might be self-powered through energy harvested from the movements of the subject to be located [144], [145].

This new location system based on smart floors can actually be extremely competitive when compared with existing solutions; among them, we can take into consideration those based on image processing [146]; they provide a high level of information at the price of high cost, power consumption, and intrusiveness. Another approach is based on pressure sensors [146], [147], providing actually a different type of smart floors; pressure sensors are nonintrusive and can be less expensive in terms of equipment when compared to the imaging systems; however, they can still be quite expensive to install into the floors; moreover, they do not allow for seamless identification of an individual in a given space. Eventually, ultrasounds [148] and RF localization system detection [149] have to be listed; although very flexible, they require dedicated communication infrastructures, additional complexity and costs, and they are mostly sensitive to environmental changes, both in terms of accuracy and functionalities.

The just mentioned smart-floor-based system tackles most of these challenges as it is very low cost, able to have a good, and precisely predictable, accuracy, and able to detect different users in a given space.

A. RFID-Based Smart Floor Including Material Aspects

Smart environments and floors should be completely seamless to the users. This entails the need to embed electronic devices into the environment in a noninvasive way. Nevertheless, this embedding of electronic devices, mainly wireless transceivers into the floor, is not a simple matter. Floors, in fact, have been made by almost the same kind of materials for centuries; tile industries are weakly available to introduce new materials to match

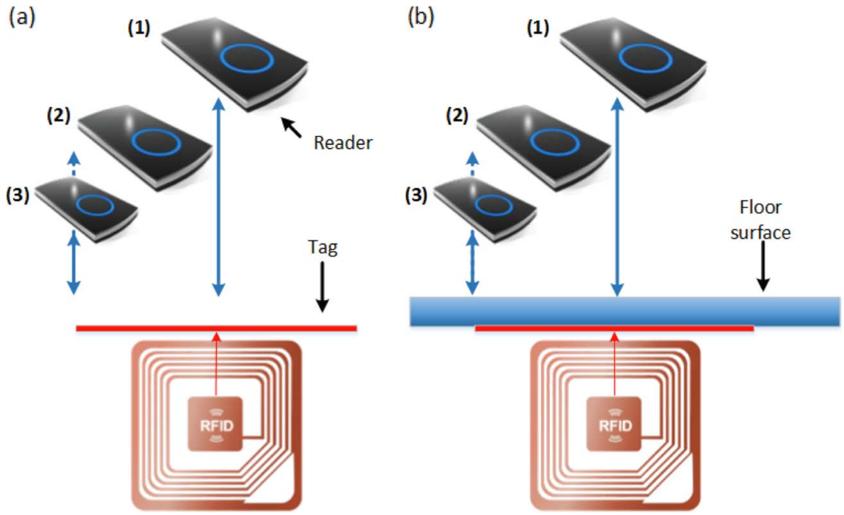

Fig. 17. ***Readable distance test scheme (a) without and (b) with tile. After [144].***

requirements posed by embedded electronics. Similar considerations can be made about the consolidated techniques to build apartments and houses: they are quite mature and whatever adaptation of them to embedded electronic systems is unrealistic. The consequence of this is that constitutive elements of SSs embedded into the floors must be suitable to materials and technologies already present: materials available are those adopted for tiles and SSs must be adapted to building constraints.

In this section, we focus on the most important approaches for the characterization of the EM properties of the materials, commonly adopted in floors, and their interactions with our electronic radio systems. As an example, ceramic tiles are going to be selected as the floor element where electronic tags have to be embedded.

Ceramic tiles have been used for ages to pave floors. These make up a great deal of most floor types in public buildings, from airports to banks, therefore, they are the most interesting flooring elements to investigate.

As the first approach, the study of the influence of some materials for floor surfaces, such as ceramic, cork laminates, and wood, in the communication between the transmitter and receiver devices has to be made; in this case, low-frequency (13 MHz) inductors are going to be used as the wireless interface.

In order to do the measurements, traditional RFIDs are used, which simplify the measurement approach significantly. A test setup is shown in Fig. 17, and experimental results are summarized in Table 3. What can be seen from the results presented in the table is that none of the tested materials affect the electrical characteristics of the RFID propagation significantly.

Another approach to the measurement is to evaluate the impact of the presence of metal inclusions beneath the ground; these can appear from pipes or from buildings supporting structures. In order to test this scenario, the presence of a metal sheet close to the coil was tested to verify its impact on the RFID reading distance. The test setup is presented in Fig. 18. In this case, when the reader is aligned (1) with the center of the coil beneath the tile, a minimum distance of 10 mm between the metallic sheet and the tag must be guaranteed in order to allow for the proper reading of the tag. In the other approach, the edge aligned scenario (2), the minimum distance between the metallic sheet and the tag increases to 30 mm to allow for communication.

Table 3 Results From Average Readable Distance Test

Tile type	Aligned (1)	Edge aligned (2)	Unaligned (3)
Without tile	4.5 cm	4.3 cm	2.4cm
Ceramic tile 4 mm thick	4.5 cm	4.3 cm	2.4cm
Ceramic tile 6 mm thick	4.5 cm	4.3 cm	2.4cm
Ceramic tile 7 mm thick	4.5 cm	4.3 cm	2.4cm
Cork laminate 5 mm thick	4.5 cm	4.3 cm	2.2cm
Wooden floor 8 mm thick	4.3 cm	4.2 cm	2.2cm

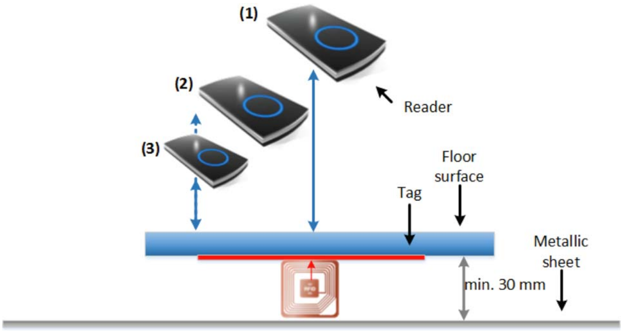

Fig. 18. ***Readable distance test scheme with metallic sheet presence. After [144].***

This test allows us to conclude that, in a real scenario, when considering the typical distance from the antenna to the floor, where the reader is usually embedded into a mobile device, a minimum distance from the tile to any metal inclusions in the ground has to be guaranteed. As seen, this safe distance should be at least 30 mm.

B. Navigation Approach Based on Smart Floors

A meaningful example of the applications of the smart floor concept is the demonstrator developed at the University of Aveiro, Aveiro, Portugal. It allows a nomadic RFID reader on top of it to be identified and localized by means of a smart floor.

The prototype in the present state is composed of four main components: a set of tagged tiles forming the smart floor, the nomadic reader, a wireless communication device to connect the reader to the information system, and a computer on which the location and navigation software is installed. A potential scenario of application for the monitoring of elderly housing is shown in Fig. 19. In the demonstrator presented, the RFID reader is connected to the radio unit and inserted into the user's shoes in order to read the passive tags as the user walks around a given area. The tile IDs are detected by the reader and sent to a centralized information system using, without lack of generality, a nonstandard wireless communication system, purposely developed in the laboratory.

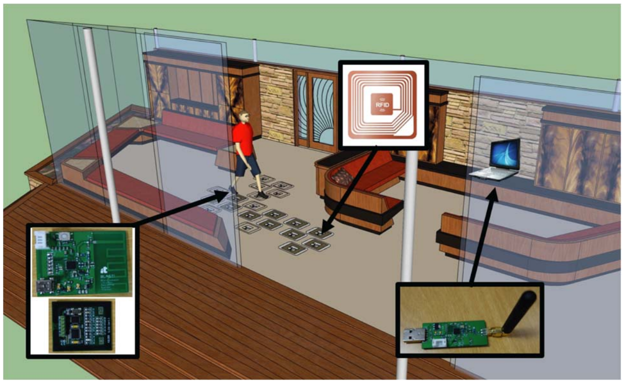

Fig. 19. ***Example of smart floor application scenario.***

The software in the computer then decodes the ID of the smart tile, and correlates it with a preloaded referenced map, thus being able to provide the precise position of the walker in the environment.

The location can be subsequently translated to a navigation system, and combined with artificial intelligence to allow navigation, situation awareness, etc.

Plenty of applications and related solutions can be envisioned on the basis of this smart floor approach: robot control, automatic pallet managements, prevention in harsh environments, and so forth.

VIII. SMART SHOES AND BAN

The aforementioned smart floor (Section VII), initially proposed to localize moving subjects in a given area, actually allows for the implementation of a very low-cost, accurate, and reliable system able to provide each kind of detectable information from an equipped area to the Internet.

To exploit the potentiality of a smart floor, it is obvious that developing purposely conceived communication units acting, on the one hand, as readers for tag tiles and, on the other hand, as communication systems with the web is needed.

One of the investigated systems for this is the so-called smart shoe [21]. Smart shoes are intended as shoes incorporating a reader able to illuminate the tag embedded into the tile and a wireless interface to communicate data to the web. To this extent, smart shoes can be seen as hubs of BANs able, on the one hand, to get information from the surrounding environment (body included) and, on the other hand, to communicate the collected information to the Internet. According to common buzzwords, they can be seen as the second technological layer, just above the first physical one consisting of tagged objects, in a hierarchical picture of IoT. Alternatively, they provide the cloud of nomadic readers with networking capabilities required for the implementation of the so-called networked RFID (N–RFID) [150] strategies. In the referred structure [21], the necessary antenna system is inkjet printed on paper, thus exploiting the process described in Section II, and it is a trademark-logo-shaped dual-band antenna working at 900 MHz and 2.4 GHz. A block diagram and its equivalent realization are shown in Fig. 20: it is worth noting that the reader itself works at 13.5 MHz exploiting the near-field communication (NFC) concept, according to what is stated in the smart floor section; as a consequence, there are no issues about the interference between the dual-band antenna and the NFC reader. Moreover, an approach similar to [145] is proposed to power the reader: the idea here is as well to harvest the energy produced by the human walking by means of piezoelectric materials embedded into the sole of the shoe.

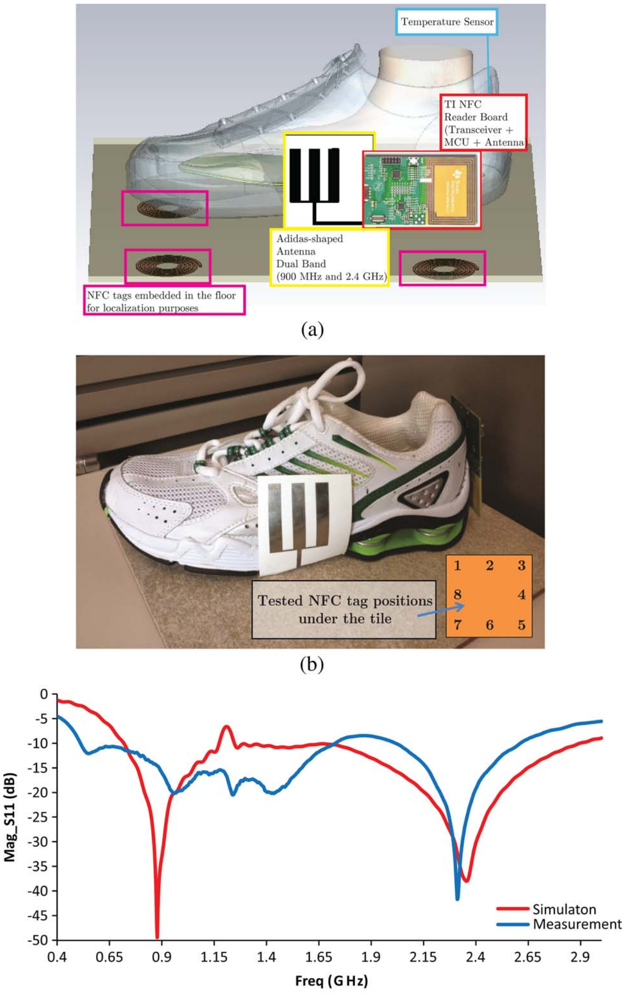

(c)

Fig. 20. ***(a) Block diagram, (b) photo, and (c) dual-band antenna results of the smart shoe that works as an energy autonomous reader for smart floors able to collect the localization data from the smart tiles and send them to a centralized system. After [21].***

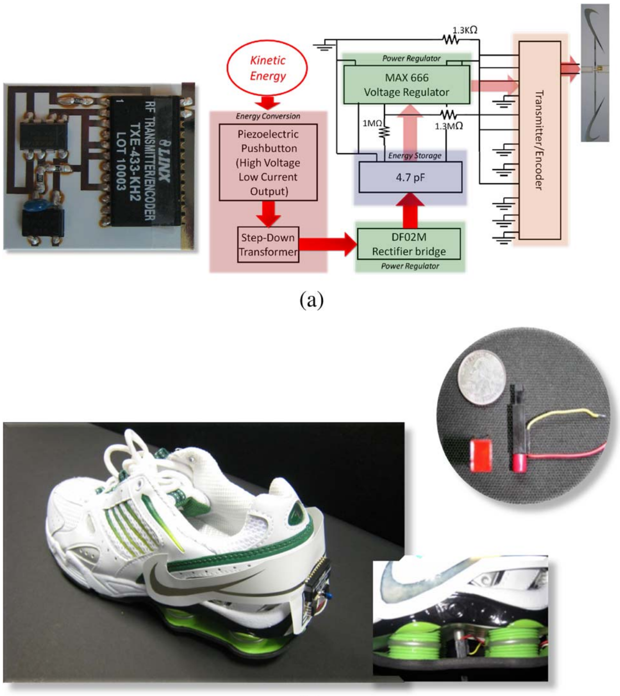

(b)

Fig. 21. ***(a) Mounted electronics on shoe with piezoelectric energy harvester and block diagram of the transmitting tag, driven by the EH unit. (b) Photograph of the assembled prototype showing the key components packaged on an organic flexible substrate. After [145].***

Fig. 21(a) shows the shoe with the mounted electronics consisting of a trademark-logo antenna inkjet printed on a flexible paper substrate together with the circuits in which the components [Fig. 21(b)] are fixed by conductive epoxy glue. A piezoelectric push button was embedded in the shoe to scavenge energy from human walking strike. This energy is then collected by a simple electronic circuit based on off-the-shelf components and used to supply the RFID transmitter. As reported in [145], the system is composed by the power-generator/energy-conversion device, the energy storage device, the power regulator circuit, and the RF transmitter that can broadcast the tag ID and the stored information.

When the push button is pressed, an inner spring is compressed and when the pressure exceeds a fixed threshold, the spring-loaded hammer will be released to deliver the dynamic mechanical force to the piezoelectric component; once the hammer strikes the piezoelectric element, a pressure wave is generated and reflected a few times between the hammer and the element, creating a mechanical resonance [151]. As a result, the generated output voltage follows closely an alternating current (ac) signal course. The signal out of the piezoelectric element is characterized by high voltage and low current; therefore, since this RF circuitry requires lower voltages at higher currents, a step-down transformer is used for better impedance matching to the following circuitry.

An energy storage device is adopted to collect the electrical energy and to provide it to the transmitter even when the external power source is temporarily unavailable; in fact, the RF transmitter takes tens of milliseconds to transmit one complete word, while the piezoelectric push button harvests energy in a transient time. To store the collected energy, a 4.7-μF tank capacitor was chosen. An ac–dc full wave diode bridge rectifier is used to convert the ac voltage from a piezoelectric element into a dc source and a dc linear regulator is added to adapt the dc voltage across the capacitor to 3-V transmitter operating voltage. For this purpose, a MAX666 low-dropout linear regulator, providing a stable 3-V supply until the tank capacitor charge is drained below the reference level of the dc linear regulator itself, is chosen. The transmitter used, the LINX

Table 4 Shoe Energy Performance Summary

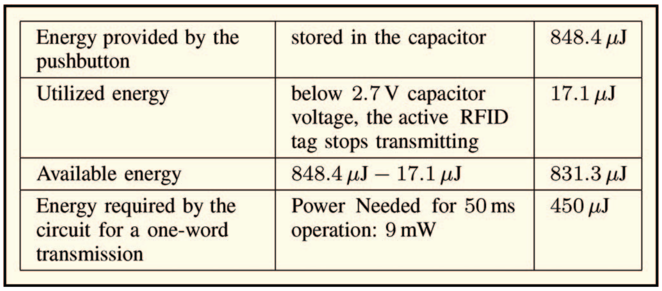

Energy provided by the pushbutton	stored in the capacitor	848.4 μJ
Utilized energy	below 2.7 V capacitor voltage, the active RFID tag stops transmitting	17.1 μJ
Available energy	848.4 μJ − 17.1 μJ	831.3 μJ
Energy required by the circuit for a one-word transmission	Power Needed for 50 ms operation: 9 mW	450 μJ

TXE-433-KH2 [152], is able to combine a highly optimized RF transmitter with an onboard encoder.

The maximum voltage across the capacitor is 19 V; therefore, the energy stored in the capacitor can be calculated to be close to 850 μJ. After the strike, the regulator becomes active, providing power to the tag circuitry; its 3-V output remains constant before it starts to drop to zero according to the capacitor discharge rate. When the voltage across the capacitor drops below 2.7 V, the active tag stops transmitting. The resulting energy left in the capacitor and not used is about 17 μJ. The required time for successful completion has been measured to be close to 50 ms; the total circuit energy consumption for a 50-ms transmission is then approximately 9 mW (Section IV). Consequently, the energy required for a one word transmission is approximately 450 μJ, much less than the power collected by the energy conversion and regulation unit. The energy performance of the smart shoe is summarized in Table 4.

"Smart" solutions like smart shoes are suitable for smart floor interaction and for collecting information (position included) from them. The smart shoe has actually a double functionality: on the one hand, it is a means to collect information from the environment (smart floor included) and from the on-body sensors and apparatuses; on the other hand, it transfers data to the Internet, via WiFi, for instance, while likely receiving control and feedback signals. To this extent, it contributes to the development of wearable electronics for body-centric communication systems, according to the scheme of Fig. 22. This class of devices is based on the BAN concept (see Fig. 22): a network of wearable computational units, sensors, energy harvesters, and transceivers networked and implemented around the human body. The information is acquired by means of several sensors and, after processing and digital modulation, is transmitted by means of off-body transceivers. The most attractive application of BANs and wearable systems, so far, is probably health monitoring [88], but the number of foreseeable applications is constantly increasing (personal fitness monitoring, personal audio system, personal alarm set, and so on). Today, a big contribution to the BANs is being given by the interaction with portable devices such as smartphones which can be exploited as gateways.

Low-power ICs, sensors, and wireless communications are key elements of the BANs, and the use of EH techniques and mixed technologies such as RFID, as in the above described smart shoe, can contribute to the development of autonomous BANs. According to the notation adopted in [74] and [153], BANs can exploit wireless communication not only for off-body communications with other systems, but also for on-body and in-body communications to interconnect networked devices [wireless body area networks (WBANs)]. Generally, the proposed WBAN solutions are based on existing communication standards; but, recently, the IEEE 802.15.6 working group has provided a standard for short-range, wireless communications in the vicinity of, or inside, a human body based on ISM bands for BANs [154]. In this rapidly evolving scenario, smart shoes can be seen, in turn, as a promising mean to interface information coming from smart floors to the BANs and ultimately to the Internet, directly or via portable devices.

IX. ENERGY SKIN

Within SSs, another area is emerging, which requires dc energy. The availability of devices that can actually increase the battery life of our electronic gadgets [155] is the holy grail of IoT; the idea is actually to convert energy from

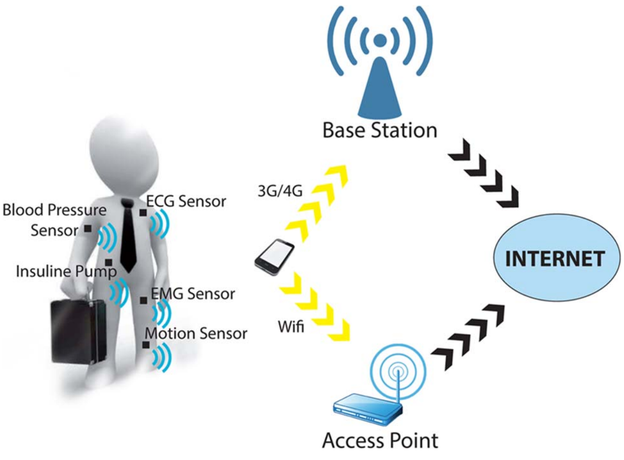

Fig. 22. ***BAN integrated with IoT: the data sensed on the human body and collected by smart devices are sent to the Internet through the standard protocols like third generation/fourth generation (3G/4G) and WiFi.***

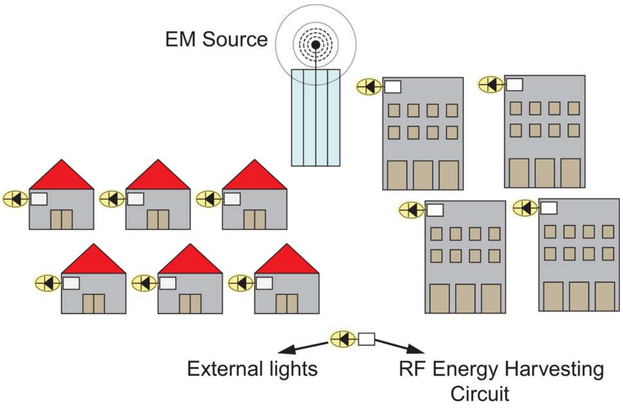

Fig. 23. ***EM source, likely supplied by solar energy in order to keep eco-friendliness, is used to broadcast energy to the urban neighborhood where houses are equipped with local RF–dc converters able to power supply local devices (external lamps in the example).***

different sources, as solar to dc [156], EM radio signals to dc, thermal to dc [156], vibration to dc [157], etc. [158].

Portable converters from solar to dc are today commonly available on the market at very low prize. Similar portable scavengers can be easily envisioned even for other energy converters. On this basis, plenty of sophisticated and more complex solutions are being proposed. Just as an example, we report the solution in Fig. 23, where the solar energy is used to supply an EM source that broadcasts RF energy in the surrounding urban environment; this energy is converted to dc energy at building level and, in turn, exploited to supply external local lamps. This approach was conceived for space applications [159], [160], but could be used in certain conditions in earth stations as well. The concept can actually move forward and be integrated into a smart skin, in the sense that all devices could be powered up by the skin energy conversion that a person can carry (vibrational, thermal, RF, etc.). This will become possible by integrating paper and textile radio electronic circuits in the smart skin surfaces of our clothes.

X. CONCLUSION

This paper focused on smart surfaces concept. This concept was introduced as a way to provide solutions to societal instances by means of technologies and techniques that today are gaining great momentum thanks to IoT development. The foreseeable IoT evolution, in fact, calls for systems and subsystems that are more and more energy conservative, fully autonomous (battery-less and not connected to the grid), as recyclable as possible (ultimately 100% biodegradable), flexible, and extremely low cost to be embedded in as many objects as possible. In order to meet these expectations, several technologies have to be implemented concurrently. This paper illustrates some of them, namely, recyclable and reliable materials, printing and R2R compatible techniques, conformal antennas on unconventional materials, and chipless RFID.

Subsequently, it has been shown how these technologies enable a fairly new branch of IoT development, represented just by SSs.

SSs are in their infancy; nevertheless, some examples, enabled by the aforementioned technologies, can be cited. Among them, the way to inherit "quasi-optics" and extend them to RF frequencies even below gigahertz is shown, and the smart floor implementation as a means to provide precise localization as well as assisted paths platforms has been reported. Energy skin as a means to convert several energy sources (solar, thermal, vibrational, EM, etc.) in electric one is described as well.

The reported examples must be researched as a new paradigm for potential promising developments aiming at providing new LAE solutions for the future networked society. ■

REFERENCES

[1] European Union (EU), "The ICT sector and RDI," Digital Agenda Scoreboard, 2012. [Online]. Available: http://ec.europa.eu/digital-agenda/sites/digital-agenda/files/scoreboard_ict_sector_and_rdi.pdf

[2] J. Manyika *et al.*, "How the Internet of Things is more like the Industrial Revolution than the Digital Revolution," MGI Rep., 2013. [Online]. Available: http://www.mckinsey.com/insights/business_technology/disruptive_technologies

[3] D. Evans, "The Internet of Things: How the next evolution of the internet is changing everything," CISCO White papers, 2011. [Online]. Available: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[4] I. G. Smith, "The Internet of Things 2012 New Horizons," IERC European Research Cluster on the Internet of Things, 2012.

[5] J. Morse, "Nanofabrication technologies for roll-to-roll processing," in *Proc. NIST-NNN Workshop*, Boston, MA, USA, Sep. 2011, pp. 1–31.

[6] A. Arias, J. MacKenzie, I. McCulloch, J. Rivnay, and A. Salleo, "Materials and applications for large area electronics: Solution-based approaches," *Chem. Rev.*, vol. 110, no. 1, pp. 3–24, Jan. 2010.

[7] T. Kampke, B. Kluge, and M. Strobel, *Field and Service Robotics*. New York, NY, USA: Springer-Verlag, 2008, ser. Springer Tracts in Advanced Robotics.

[8] M. Shur, "Giant area and flexible electronics," in *Proc. IEEE Nanotechnol. Mater. Devices Conf.*, Oct. 2006, vol. 1, pp. 264–265.

[9] M. Steer *et al.*, "Global modeling of spatially distributed microwave and millimeter-wave systems," *IEEE Trans. Microw. Theory Tech.*, vol. 47, no. 6, pp. 830–838, Jun. 1999.

[10] A. Rida, L. Yang, R. Vyas, and M. Tentzeris, "Conductive inkjet-printed antennas on flexible low-cost paper-based substrates for RFID and WSN applications," *IEEE Antennas Propag. Mag.*, vol. 51, no. 3, pp. 13–23, Jun. 2009.

[11] F. Alimenti, P. Mezzanotte, M. Dionigi, M. Virili, and L. Roselli, "Microwave circuits in paper substrates exploiting conductive adhesive tapes," *IEEE Microw. Wireless Compon. Lett.*, vol. 22, no. 12, pp. 660–662, Dec. 2012.

[12] D. Tobjörk and R. Österbackaa, "Paper electronics," *Adv. Mater.*, vol. 23, no. 17, pp. 1935–1961, 2011.

[13] P. Calvert, "Inkjet printing for materials and devices," *Chem. Mater.*, vol. 13, no. 10, pp. 3299–3305, 2001.

[14] G. Orecchini *et al.*, "Inkjet printed organic transistors for sustainable electronics," in *Proc. 60th Electron. Compon. Technol. Conf.*, Jun. 2010, pp. 985–989.

[15] L. Yang, G. Orecchini, G. Shaker, H.-S. Lee, and M. Tentzeris, "Battery-free RFID-enabled wireless sensors," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2010, pp. 1528–1531.

[16] G. Shaker, S. Safavi-Naeini, N. Sangary, and M. Tentzeris, "Inkjet printing of ultrawideband (UWB) antennas on paper-based substrates," *IEEE Antennas*

Wireless Propag. Lett., vol. 10, pp. 111–114, 2011.

[17] R. Vyas *et al.*, "Inkjet printed, self powered, wireless sensors for environmental, gas, and authentication-based sensing," *IEEE Sensors J.*, vol. 11, no. 12, pp. 3139–3152, Dec. 2011.

[18] G. Orecchini *et al.*, "Design and fabrication of ultra-low cost radio frequency identification antennas and tags exploiting paper substrates and inkjet printing technology," *IET Microw. Antennas Propag.*, vol. 5, no. 8, pp. 993–1001, Jun. 2011.

[19] S. Kim *et al.*, "Monopole antenna with inkjet-printed EBG array on paper substrate for wearable applications," *IEEE Antennas Wireless Propag. Lett.*, vol. 11, pp. 663–666, 2012.

[20] S. Kim *et al.*, "No battery required: Perpetual RFID-enabled wireless sensors for cognitive intelligence applications," *IEEE Microw. Mag.*, vol. 14, no. 5, pp. 66–77, Jul. 2013.

[21] C. Mariotti, V. Lakafosis, M. Tentzeris, and L. Roselli, "An IPv6-enabled wireless shoe-mounted platform for health-monitoring," in *Proc. IEEE Top. Conf. Wireless Sensors Sensor Netw.*, Jan. 2013, pp. 46–48.

[22] K. Sangkil *et al.*, "Inkjet-printed antennas, sensors and circuits on paper substrate," *IET Microw. Antennas Propag.*, vol. 7, no. 10, pp. 858–868, 2013.

[23] B. Cook, J. Cooper, and M. Tentzeris, "Multi-layer RF capacitors on flexible substrates utilizing inkjet printed dielectric polymers," *IEEE Microw. Wireless Compon. Lett.*, vol. 23, no. 7, pp. 353–355, Jul. 2013.

[24] B. Cook, J. Cooper, S. Kim, and M. Tentzeris, "A novel inkjet-printed passive microfluidic RFID-based sensing platform," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2013, DOI: 10.1109/MWSYM.2013.6697592.

[25] B. Cook, B. Tehrani, J. Cooper, and M. Tentzeris, "Multilayer inkjet printing of millimeter-wave proximity-fed patch arrays on flexible substrates," *IEEE Antennas Wireless Propag. Lett.*, vol. 12, pp. 1351–1354, 2013.

[26] C. Mariotti, B. Cook, L. Roselli, and M. Tentzeris, "High-performance inkjet-printed metal-insulator-metal (MIM) capacitors on lossy silicon substrate *IEEE Microw. Wireless Compon. Lett.*, 2014.

[27] D. Angmo, T. T. Larsen-Olsen, M. Jorgensen, R. R. Sazndergaard, and F. C. Krebs, "Roll-to-roll inkjet printing and photonic sintering of electrodes for ITO free polymer solar cell modules and facile product integration," *Adv. Energy Mater.*, vol. 3, no. 2, pp. 172–175, 2013.

[28] R. Sazndergaard, M. Hessel, D. Angmo, T. T. Larsen-Olsen, and F. C. Krebs, "Roll-to-roll fabrication of polymer solar cells," *Mater. Today*, vol. 15, no. 1–2, pp. 36–49, 2012.

[29] F. Alimenti *et al.*, "A 1.2 v, 0.9 mW UHF VCO based on hairpin resonator in paper substrate and Cu adhesive tape," *IEEE Microw. Wireless Compon. Lett.*, vol. 23, no. 4, pp. 214–216, Apr. 2013.

[30] F. Alimenti *et al.*, "24 GHz single-balanced diode mixer exploiting cellulose-based materials," *IEEE Microw. Wireless Compon. Lett.*, vol. 23, no. 11, pp. 596–598, Nov. 2013.

[31] C. Mariotti *et al.*, "Modeling and characterization of copper tape microstrips on paper substrate and application to 24 GHz branch-line couplers," in *Proc. Eur. Microw. Conf.*, Oct. 2013, pp. 794–797.

[32] L. Guoyu and W. Zhenkai, "Summarize of rfid technology and typical application," in *Proc. Cross Strait Quad-Regional Radio Sci. Wireless Technol. Conf.*, Jul. 2011, vol. 2, pp. 1032–1036.

[33] L. Wang, L. D. Xu, Z. Bi, and Y. Xu, "Data cleaning for RFID and WSN integration," *IEEE Trans. Ind. Inf.*, vol. 10, no. 1, pp. 408–418, Feb. 2014.

[34] H. Deng *et al.*, "Design of sensor-embedded radio frequency identification (SE-RFID) systems," in *Proc. IEEE Int. Conf. Mechatron. Autom.*, Jun. 2006, pp. 792–796.

[35] S. Mohamad, F. Tang, A. Amira, A. Bermak, and M. Benammar, "A low power oscillator based temperature sensor for RFID applications," in *Proc. 5th Asia Symp. Quality Electron. Design*, Aug. 2013, pp. 50–54.

[36] D. Hotte, R. Siragusa, Y. Duroc, and S. Tedjini, "Performance optimization for the choice of impedances in the design of passive RFID tags," in *Proc. IEEE-APS Top. Conf. Antennas Propag. Wireless Commun.*, Sep. 2013, pp. 1349–1352.

[37] M. Donelli, "Guidelines for the design and optimization of wireless sensors based on the modulated scattering technique," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 7, pp. 1824–1833, Jul. 2014.

[38] A. Ricci, M. Grisanti, I. De Munari, and P. Ciampolini, "Improved pervasive sensing with RFID: An ultra-low power baseband processor for UHF tags," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 12, pp. 1719–1729, Dec. 2009.

[39] L. Chen *et al.*, "Range extension of passive wake-up radio systems through energy harvesting," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2013, pp. 1549–1554.

[40] S. Sauer, A. Turke, A. Weder, and W.-J. Fischer, "Inkjet printed humidity threshold monitoring sensor solution with irreversible resistance change for passive RFID applications," in *Proc. IEEE Sensors*, Nov. 2013, DOI: 10.1109/ICSENS.2013.6688307.

[41] N. Cho, S.-J. Song, S. Kim, S. Kim, and H.-J. Yoo, "A 5.1-uW UHF RFID tag chip integrated with sensors for wireless environmental monitoring," in *Proc. 31st Eur. Solid-State Circuits Conf.*, Sep. 2005, pp. 279–282.

[42] M. Philipose *et al.*, "Battery-free wireless identification and sensing," *IEEE Perv. Comput.*, vol. 4, no. 1, pp. 37–45, Jan. 2005.

[43] D. De Donno, L. Catarinucci, and L. Tarricone, "A battery-assisted sensor-enhanced RFID tag enabling heterogeneous wireless sensor networks," *IEEE Sensors J.*, vol. 14, no. 4, pp. 1048–1055, Apr. 2014.

[44] N. Guilar, T. Kleeburg, A. Chen, D. Yankelevich, and R. Amirtharajah, "Integrated solar energy harvesting and storage," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 5, pp. 627–637, May 2009.

[45] D. Brunelli, C. Moser, L. Thiele, and L. Benini, "Design of a solar-harvesting circuit for batteryless embedded systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 11, pp. 2519–2528, Nov. 2009.

[46] D. Kwon, G. Rincon-Mora, and E. Torres, "Harvesting ambient kinetic energy with switched-inductor converters," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 7, pp. 1551–1560, Jul. 2011.

[47] S. Mehraeen, S. Jagannathan, and K. Corzine, "Energy harvesting from vibration with alternate scavenging circuitry and tapered cantilever beam," *IEEE Trans. Ind. Electron.*, vol. 57, no. 3, pp. 820–830, Mar. 2010.

[48] T. Le, K. Mayaram, and T. Fiez, "Efficient far-field radio frequency energy harvesting for passively powered sensor networks," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1287–1302, May 2008.

[49] G. Papotto, F. Carrara, and G. Palmisano, "A 90-nm CMOS threshold-compensated RF energy harvester," *IEEE J. Solid-State Circuits*, vol. 46, no. 9, pp. 1985–1997, Sep. 2011.

[50] S. Rahimizadeh, S. Korhummel, B. Kaslon, and Z. Popovic, "Scalable adaptive wireless powering of multiple electronic devices in an over-moded cavity," in *Proc. IEEE Wireless Power Transfer*, May 2013, pp. 84–87.

[51] S. Scorcioni, L. Larcher, A. Bertacchini, L. Vincetti, and M. Maini, "An integrated RF energy harvester for UHF wireless powering applications," in *Proc. IEEE Wireless Power Transfer*, May 2013, pp. 92–95.

[52] Z. Popovic, "Far-field wireless power delivery and power management for low-power sensors," in *Proc. IEEE Wireless Power Transfer*, May 2013, DOI: 10.1109/WPT.2013.6556867.

[53] E. Falkenstein, M. Roberg, and Z. Popovic, "Low-power wireless power delivery," *IEEE Trans. Microw. Theory Tech.*, vol. 60, no. 7, pp. 2277–2286, Jul. 2012.

[54] S. Roundy, P. K. Wright, and J. Rabaey, "A study of low level vibrations as a power source for wireless sensor nodes," *Comput. Commun.*, vol. 26, no. 11, pp. 1131–1144, 2003.

[55] C. D. Richards, M. J. Anderson, D. F. Bahr, and R. F. Richards, "Efficiency of energy conversion for devices containing a piezoelectric component," *J. Micromech. Microeng.*, vol. 14, no. 5, 2004, DOI: 10.1088/0960-1317/14/5/009.

[56] S. R. Anton and H. A. Sodano, "A review of power harvesting using piezoelectric materials (2003–2006)," *Smart Mater. Struct.*, vol. 16, no. 3, 2007, DOI: 10.1088/0964-1726/16/3/R01.

[57] Cymbet Corporation, "Energy harvesting and storage." [Online]. Available: http://www.cymbet.com/design-center/energy-harvesting.php#

[58] T. Armstrong, "Energy harvesting applications are everywhere," RTC, May 2011. [Online]. Available: http://www.rtcmagazine.com/articles/view/102133

[59] D. Puccinelli and M. Haenggi, "Wireless sensor networks: Applications and challenges of ubiquitous sensing," *IEEE Circuits Syst. Mag.*, vol. 5, no. 3, pp. 19–31, 2005.

[60] L. Roselli *et al.*, "WPT, RFID and energy harvesting: Concurrent technologies for the future networked society," in *Proc. Asia-Pacific Microw. Conf.*, Nov. 2013, pp. 462–464.

[61] H. Liu, M. Bolic, A. Nayak, and I. Stojmenovic, "Taxonomy and challenges of the integration of RFID and wireless sensor networks," *IEEE Network*, vol. 22, no. 6, pp. 26–35, Nov. 2008.

[62] D. Dardari, R. D'Errico, C. Roblin, A. Sibille, and M. Win, "Ultrawide bandwidth RFID: The next generation?" *Proc. IEEE*, vol. 98, no. 9, pp. 1570–1582, Sep. 2010.

[63] C. Balanis, *Antenna Theory: Analysis and Design*. New York, NY, USA: Wiley, 2005, ser. Knovel Library.

[64] K. Finkenzeller and D. Müller, *RFID Handbook: Fundamentals and Applications*

in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication. New York, NY, USA: Wiley, 2010.

[65] E. Perret, S. Tedjini, and R. Nair, "Design of antennas for UHF RFID tags," *Proc. IEEE*, vol. 100, no. 7, pp. 2330–2340, Jul. 2012.

[66] K. V. S. Rao, P. Nikitin, and S. Lam, "Antenna design for UHF RFID tags: A review and a practical application," *IEEE Trans. Antennas Propag.*, vol. 53, no. 12, pp. 3870–3876, Dec. 2005.

[67] H. Chaabane, E. Perret, and S. Tedjini, "A methodology for the design of frequency and environment robust UHF RFID tags," *IEEE Trans. Antennas Propag.*, vol. 59, no. 9, pp. 3436–3441, Sep. 2011.

[68] A. Babar, A. Elsherbeni, L. Sydanheimo, and L. Ukkonen, "RFID tags for challenging environments: Flexible high-dielectric materials and ink-jet printing technology for compact platform tolerant RFID tags," *IEEE Microw. Mag.*, vol. 14, no. 5, pp. 26–35, Jul. 2013.

[69] M. Scarpello, I. Kazani, C. Hertleer, H. Rogier, and D. Vande Ginste, "Stability and efficiency of screen-printed wearable and washable antennas," *IEEE Antennas Wireless Propag. Lett.*, vol. 11, pp. 838–841, 2012.

[70] V. Lakafosis *et al.*, "Progress towards the first wireless sensor networks consisting of inkjet-printed, paper-based RFID-enabled sensor tags," *Proc. IEEE*, vol. 98, no. 9, pp. 1601–1609, Sep. 2010.

[71] C. Mariotti, G. Orecchini, M. Virili, F. Alimenti, and L. Roselli, "Wireless localization in buildings by smart tiles," in *Proc. IEEE Workshop Environ. Energy Struct. Monitoring Syst.*, Sep. 2012, pp. 7–11.

[72] H. J. Song *et al.*, "A method for improving the efficiency of transparent film antennas," *IEEE Antennas Wireless Propag. Lett.*, vol. 7, pp. 753–756, 2008.

[73] F. Colombel *et al.*, "Ultra-thin metal layer, ITO film and ITO/Cu/ITO multilayer towards transparent antenna," *IET Sci. Meas. Technol.*, vol. 3, no. 3, pp. 229–234, May 2009.

[74] D. Zito and D. Pepe, "System-on-a-chip radio transceivers for 60-GHz wireless body-centric communications," in *Ultra-Wideband and 60 GHz Communications*. New York, NY, USA: Springer-Verlag, 2014, pp. 177–187.

[75] L. Martin *et al.*, "Effect of permittivity and permeability of a flexible magnetic composite material on the performance and miniaturization capability of planar antennas for RFID and wearable wireless applications," *IEEE Trans. Compon. Packag. Technol.*, vol. 32, no. 4, pp. 849–858, Dec. 2009.

[76] D. Kim and J. Yeo, "Low-profile RFID tag antenna using compact AMC substrate for metallic objects," *IEEE Antennas Wireless Propag. Lett.*, vol. 7, pp. 718–720, 2008.

[77] C. Kakoyiannis and P. Constantinou, "Compact WSN antennas of analytic geometry based on Chebyshev polynomials," in *Proc. Antennas Propag. Conf.*, Nov. 2012, DOI: 10.1109/LAPC.2012.6402976.

[78] H. Visser and R. Vullers, "RF energy harvesting and transport for wireless sensor network applications: Principles and requirements," *Proc. IEEE*, vol. 101, no. 6, pp. 1410–1423, Jun. 2013.

[79] M. Pinuela, P. Mitcheson, and S. Lucyszyn, "Ambient RF energy harvesting in urban and semi-urban environments," *IEEE Trans. Microw. Theory Tech.*, vol. 61, no. 7, pp. 2715–2726, Jul. 2013.

[80] K. M. Farinholt, G. Park, and C. Farrar, "RF energy transmission for a low-power wireless impedance sensor node," *IEEE Sensors J.*, vol. 9, no. 7, pp. 793–800, Jul. 2009.

[81] J. Hagerty, F. Helmbrecht, W. McCalpin, R. Zane, and Z. Popovic, "Recycling ambient microwave energy with broad-band rectenna arrays," *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 3, pp. 1014–1024, Mar. 2004.

[82] A. Collado and A. Georgiadis, "Conformal hybrid solar and electromagnetic (EM) energy harvesting rectenna," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 8, pp. 2225–2234, Aug. 2013.

[83] T. Wu, R. Li, and M. Tentzeris, "A scalable solar antenna for autonomous integrated wireless sensor nodes," *IEEE Antennas Wireless Propag. Lett.*, vol. 10, pp. 510–513, 2011.

[84] A. Sample, J. Braun, A. Parks, and J. Smith, "Photovoltaic enhanced UHF RFID tag antennas for dual purpose energy harvesting," in *Proc. IEEE Int. Con. RFID*, Apr. 2011, pp. 146–153.

[85] F. Alimenti *et al.*, "A new contactless assembly method for paper substrate antennas and UHF RFID chips," *IEEE Trans. Microw. Theory Tech.*, vol. 59, no. 3, pp. 627–637, Mar. 2011.

[86] W. Pachler *et al.*, "A silver ink-jet printed UHF booster antenna on flexible substratum with magnetically coupled RFID die on-chip antenna," in *Proc. 7th Eur. Conf. Antennas Propag.*, Apr. 2013, pp. 1730–1733.

[87] M. Virili, H. Rogier, F. Alimenti, P. Mezzanotte, and L. Roselli, "Wearable textile antenna magnetically coupled to flexible active electronic circuits," *IEEE Antennas Wireless Propag. Lett.*, vol. 13, pp. 209–212, 2014.

[88] H.-J. Yoo, "Your heart on your sleeve: Advances in textile-based electronics are weaving computers right into the clothes we wear," *IEEE Solid-State Circuits Mag.*, vol. 5, no. 1, pp. 59–70, Mar. 2013.

[89] X. Jingtian *et al.*, "Low-cost low-power UHF RFID tag with on-chip antenna," *J. Semicond.*, vol. 30, no. 7, pp. 075012/1–075012/6, Jul. 2009.

[90] M. Law, A. Bermak, and H. Luong, "A sub-μW embedded CMOS temperature sensor for RFID food monitoring application," *IEEE J. Solid-State Circuits*, vol. 45, no. 6, pp. 1246–1255, Jun. 2010.

[91] Alien Technology Corporation, "Fluidic self assembly," White Paper, 1999. [Online]. Available: http://www.alientechnology.com

[92] C. Hertleer, H. Rogier, L. Vallozzi, and L. V. Langenhove, "A textile antenna for off-body communication integrated into protective clothing for firefighters," *IEEE Trans. Antennas Propag.*, vol. 57, no. 4, pp. 919–925, Apr. 2009.

[93] X. Li, J. Liao, Y. Yuan, and D. Yu, "Eye-shaped segmented reader antenna for near-field UHF RFID applications," *Progr. Electromagn. Res.*, vol. 114, pp. 481–493, 2011.

[94] J.-J. Tiang, M. Islam, N. Misran, and J. S. Mandeep, "Circular microstrip slot antenna for dual-frequency RFID application," *Progr. Electromagn. Res.*, vol. 120, pp. 499–512, 2011.

[95] Y. Amin, Q. Chen, H. Tenhunen, and L.-R. Zheng, "Performance-optimized quadrate bowtie RFID antennas for cost-effective and eco-friendly industrial applications," *Progr. Electromagn. Res.*, vol. 126, pp. 49–64, 2012.

[96] Y. Amin, Q. Chen, L.-R. Zheng, and H. Tenhunen, "Development and analysis of flexible UHF RFID antennas for "green" electronics," *Progr. Electromagn. Res.*, vol. 130, pp. 1–15, 2012.

[97] F. Viani, M. Salucci, G. Olivieri, F. Robol, and A. Massa, "Design of UHF RFID/GPS fractal antenna for logistic management," *J. Electromagn. Waves Appl.*, vol. 26, no. 4, pp. 480–492, 2012.

[98] A. Springer, R. Weigel, A. Pohl, and F. Seifert, "Wireless identification and sensing using surface acoustic wave devices," *Mechatronics*, vol. 9, no. 7, pp. 745–756, Oct. 1999.

[99] K. Chang, Y. Kim, Y. Kim, and Y. Yoon, "Functional antenna integrated with relative humidity sensor using synthetised polymide for passive RFID sensing," *Electron. Lett.*, vol. 43, no. 5, pp. 1918–1923, May 2007.

[100] S. Shrestha, M. Balachandran, M. Agarwal, V. Phoha, and K. Varahramyan, "A chipless RFID sensor system for cyber centric monitoring applications," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 5, pp. 1303–1309, May 2009.

[101] C. Occhiuzzi, S. Cippitelli, and G. Marrocco, "Modeling, design and experimentation of wearable RFID sensor tag," *IEEE Trans. Antennas Propag.*, vol. 58, no. 8, pp. 2490–2498, Aug. 2010.

[102] A. Ramos, A. Lazaro, D. Girbau, and R. Villarino, "Time-domain measurement of time-coded UWB chipless RFID tags," *Progr. Electromagn. Res.*, vol. 116, pp. 313–331, 2011.

[103] R. Nair, E. Perret, and S. Tedjini, "Temporal multi-frequency encoding technique for chipless RFID applications," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Montreal, QC, Canada, Jun. 2012, DOI: 10.1109/MWSYM.2012.6259483.

[104] R. Bhattacharyya, C. Floerkemeier, and S. Sarma, "Low-cost, ubiquitous RFID-tag-antenna-based sensing," *Proc. IEEE*, vol. 98, no. 9, pp. 1593–1600, Sep. 2010.

[105] R. Potyrailo *et al.*, "Development of radio-frequency identification sensors based on organic electronic sensing materials for selective detection of toxic vapors," *J. Appl. Phys.*, vol. 106, no. 12, pp. 124 902–124 902-6, Dec. 2009.

[106] V. Viikari and H. Seppa, "RFID MEMS sensor concept based on intermodulation distortion," *IEEE Sensor J.*, vol. 9, no. 12, pp. 1918–1923, Dec. 2009.

[107] J. Riley *et al.*, "Tracking bees with harmonic radar," *Nature*, vol. 379, pp. 29–30, Jan. 1996.

[108] S. Helbing *et al.*, "Design and verification of a novel crossed dipole structure for quasi-optical frequency doublers," *IEEE Microw. Guided Wave Lett.*, vol. 10, no. 3, pp. 105–107, Mar. 2000.

[109] F. Alimenti and L. Roselli, "Theory of zero-power RFID sensors based on harmonic generation and orthogonally polarized antennas," *Progr. Electromagn. Res.*, vol. 134, pp. 337–357, 2013.

[110] G. Lovei, I. Stringer, C. Devine, and M. Cartellieri, "Harmonic radar—A method using inexpensive tags to study invertebrate movement on land," *New Zealand J. Ecology*, vol. 21, no. 2, pp. 187–193, 1997.

[111] B. Colpitts and G. Boiteau, "Harmonic radar transceiver design: Miniature tags for insect tracking," *IEEE Trans. Antennas Propag.*, vol. 52, no. 11, pp. 2825–2832, Nov. 2004.

[112] G. Orecchini *et al.*, "Green technologies and RFID: Present and future," *Appl. Comput. Electromagn. Soc. J.*, vol. 25, no. 3, pp. 230–238, Mar. 2010.

[113] S. Steudel *et al.*, "Comparison of organic diode structures regarding high-frequency rectification behavior in radio-frequency identification tags," *J. Appl. Phy.*, vol. 99, no. 11, Jun. 2006, 114519.

[114] L. Valentini and J. Kenny, "Novel approaches to developing carbon nanotube based polymer composites: Fundamental studies and nanotech applications," *Polymer*, vol. 46, no. 17, pp. 6715–6718, Aug. 2005.

[115] V. Marinov *et al.*, "Direct-write vapor sensors on FR4 plastic substrates," *IEEE Sensor J.*, vol. 7, no. 6, pp. 937–944, Jun. 2007.

[116] T. Unander and H.-E. Nilsson, "Characterization of printed moisture sensors in packaging surveillance applications," *IEEE Sensor J.*, vol. 9, no. 8, pp. 922–928, Aug. 2009.

[117] S. Couderc, B. Kim, and T. Someya, "Cellulose-based composite as a raw material for flexible ANS ultra-lightweight mechanical switch devices," in *Proc. IEEE 22nd Int. Conf. Micro Electro Mech. Syst.*, Sorrento, Italy, Jan. 2009, pp. 646–649.

[118] P. F. Goldsmith, "Quasioptics in millimeterwave systems," in *Proc. 12th Eur. Microw. Conf.*, Sep. 1982, pp. 16–24.

[119] P. Goldsmith, "Quasi-optical techniques," *Proc. IEEE*, vol. 80, no. 11, pp. 1729–1747, Nov. 1992.

[120] P. F. Goldsmith, *Quasioptical Systems*. New York, NY, USA: IEEE Press, 1998.

[121] T. Djerafi, K. Wu, and S. Tatu, "3 dB 90° hybrid quasi-optical coupler with air field slab in SIW technology," *IEEE Microw. Wireless Compon. Lett.*, vol. 24, no. 4, pp. 221–223, Apr. 2014.

[122] A. Gonzalez, Y. Fujii, K. Kaneko, and Y. Uzawa, "Quasi-optical attenuators for alma band 10 receiver," in *Proc. Asia-Pacific Microw. Conf. Proc.*, Dec. 2011, pp. 1977–1980.

[123] K. Song, F. Zhang, S. Hu, and Y. Fan, "Millimeter-wave-wave quasi-optical low-loss power combiner based on dipole antenna," *Electron. Lett.*, vol. 49, no. 18, pp. 1160–1162, Aug. 2013.

[124] S. Kawasaki and T. Itoh, "Quasi-optical planar arrays with fets and slots," *IEEE Trans. Microw. Theory Tech.*, vol. 41, no. 10, pp. 1838–1844, Oct. 1993.

[125] S. Hollung, W. Shiroma, M. Markovic, and Z. Popovic, "A quasi-optical isolator," *IEEE Microw. Guided Wave Lett.*, vol. 6, no. 5, pp. 205–206, May 1996.

[126] N. Gagnon and J. Shaker, "Accurate phase measurement of passive non-reciprocal quasi-optical components," *IEE Proc.—Microw. Antennas Propag.*, vol. 152, no. 2, pp. 111–116, Apr. 2005.

[127] R. Dickie, R. Cahill, V. Fusco, H. Gamble, and N. Mitchell, "THz frequency selective surface filters for earth observation remote sensing instruments," *IEEE Trans. Terahertz Sci. Technol.*, vol. 1, no. 2, pp. 450–461, Nov. 2011.

[128] M. Euler and V. Fusco, "Quasi optical circulator X-band scale model utilizing a circular polarization frequency selective surface," in *Proc. 4th Eur. Conf. Antennas Propag.*, Apr. 2010, pp. 1–4.

[129] W. Shiroma, S. Bundy, S. Hollung, B. Bauernfeind, and Z. Popovic, "Cascaded active and passive quasi-optical grids," *IEEE Trans. Microw. Theory Tech.*, vol. 43, no. 12, pp. 2904–2909, Dec. 1995.

[130] J. Birkeland and T. Itoh, "A 16 element quasi-optical fet oscillator power combining array with external injection locking," *IEEE Trans. Microw. Theory Tech.*, vol. 40, no. 3, pp. 475–481, Mar. 1992.

[131] M. Forman, T. Marshall, and Z. Popovic, "Two Ka-band quasi-optical amplifier arrays," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 1999, vol. 4, pp. 1829–1832.

[132] S. Hollung, A. Cox, and Z. Popovic, "A bi-directional quasi-optical lens amplifier," *IEEE Trans. Microw. Theory Tech.*, vol. 45, no. 12, pp. 2352–2357, Dec. 1997.

[133] J. Hacker, R. Weikle, M. Kim, P. De Lisio, and D. Rutledge, "A 100-element planar Schottky diode grid mixer," *IEEE Trans. Microw. Theory Tech.*, vol. 40, no. 3, pp. 557–562, Mar. 1992.

[134] G. Chattopadhyay, D. Miller, H. LeDuc, and J. Zmuidzinas, "A dual-polarized quasi-optical SIS mixer at 550 GHz," *IEEE Trans. Microw. Theory Tech.*, vol. 48, no. 10, pp. 1680–1686, Oct. 2000.

[135] V. Bezborodov, V. Kiseliov, Y. Kuleshov, and M. Yanovsky, "A terahertz quasi-optical polarization phase shifter with the crystal quartz phase sections," in *Proc. 5th Int. Kharkov Symp. Phys. Eng. Microw. Millimeter Submillimeter Waves*, Jun. 2004, vol. 2, pp. 603–605.

[136] G. Denisov *et al.*, "Fast quasi-optical phase shifter based on induced photoconductivity in silicon," in *Proc. Joint 32nd Int. Conf. Infrared Millimeter Waves/15th Int. Conf. Terahertz Electron.*, Sep. 2007, pp. 795–796.

[137] N. Camilleri, "A quasi-optical multiplying slot array," *IEEE Trans. Microw. Theory Tech.*, vol. 33, no. 11, pp. 1189–1195, Nov. 1985.

[138] A. Moussessian *et al.*, "A terahertz grid frequency doubler," *IEEE Trans. Microw. Theory Tech.*, vol. 46, no. 11, pp. 1976–1981, Nov. 1998.

[139] M. Cryan *et al.*, "Simulation and measurement of quasi-optical multipliers," *IEEE Trans. Microw. Theory Tech.*, vol. 49, no. 3, pp. 451–464, Mar. 2001.

[140] F. Jiang *et al.*, "High-speed monolithic millimeter-wave switch array," *IEEE Microw. Guided Wave Lett.*, vol. 8, no. 3, pp. 112–114, Mar. 1998.

[141] F. Alimenti *et al.*, "Crossed dipole frequency doubling RFID TAG based on paper substrate and ink-jet printing technology," in *Proc. IEEE MTT-S Int. Microw. Symp. Dig.*, May 2010, pp. 840–842.

[142] M. Virili *et al.*, "7.5–15 MHz organic frequency doubler made with pentacene-based diode and paper substrate," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Tampa, FL, USA, Jun. 2014, DOI: 10.1109/MWSYM.2014.6848395.

[143] L. Roselli *et al.*, "Feasibility study of a fully organic frequency doubler for harmonic RFID applications," in *Proc. IEEE 12th Top. Meeting Silicon Monolithic Integr. Circuits RF Syst.*, Jan. 2012, pp. 203–206.

[144] R. Gonalves *et al.*, "Smart floor: Indoor navigation based on RFID," in *Proc. IEEE Wireless Power Transfer*, May 2013, pp. 103–106.

[145] G. Orecchini, L. Yang, M. Tentzeris, and L. Roselli, "Wearable battery-free active paper printed RFID tag with human-energy scavenger," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2011, DOI: 10.1109/MWSYM.2011.5973290.

[146] C.-R. Yu, C.-L. Wu, C.-H. Lu, and L.-C. Fu, "Human localization via multi-cameras and floor sensors in smart home," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, Oct. 2006, vol. 5, pp. 3822–3827.

[147] S. Chang, S. Ham, S. Kim, D. Suh, and H. Kim, "Ubi-floor: Design and pilot implementation of an interactive floor system," in *Proc. 2nd Int. Conf. Intell. Human-Mach. Syst. Cybern.*, Aug. 2010, vol. 2, pp. 290–293.

[148] J. Morgado and A. Konig, "Low-power concept and prototype of distributed resistive pressure sensor array for smart floor and surfaces in intelligent environments," in *Proc. 9th Int. Multi-Conf. Syst. Signals Devices*, Mar. 2012, DOI: 10.1109/SSD.2012.6198020.

[149] J. Zhao and Y. Wang, "Autonomous ultrasonic indoor tracking system," in *Proc. Int. Symp. Parallel Distrib. Process. Appl.*, Dec. 2008, pp. 532–539.

[150] R. Imura, "The role of networked RFID for driving the ubiquitous," in *Proc. IEEE Conf. Emerging Technol. Factory Autom.*, Sep. 2006, pp. 1115–1118.

[151] Y. Tan, K. Y. Hoe, and S. Panda, "Energy harvesting using piezoelectric igniter for self-powered radio frequency (RF) wireless sensors," in *Proc. IEEE Int. Conf. Ind. Technol.*, Dec. 2006, pp. 1711–1716.

[152] Linx Technologies, "KH2 series transmitter/encoder data guide." [Online]. Available: https://www.linxtechnologies.com/resources/data-guides/txe-xxx-kh2.pdf

[153] D. Smith, D. Miniutti, T. Lamahewa, and L. Hanlen, "Propagation models for body-area networks: A survey and new outlook," *IEEE Antennas Propag. Mag.*, vol. 55, no. 5, pp. 97–117, Oct. 2013.

[154] *IEEE Standard for Local and Metropolitan Area Networks—Part 15.6: Wireless Body Area Networks*, IEEE Std. 802.15.6-2012, Feb. 2012.

[155] J. Thomas and M. Qidwai, "The design and application of multifunctional structure-battery materials systems," *JOM*, vol. 57, no. 3, pp. 18–24, Mar. 2005.

[156] A. Bonfiglio and D. De Rossi, *Wearable Monitoring Systems*. New York, NY, USA: Springer-Verlag, 2011.

[157] S. Lee and B. Youn, "A design and experimental verification methodology for an energy harvester skin structure," *Smart Mater. Struct.*, vol. 20, no. 5, May 2011, DOI: 10.1088/0964-1726/20/5/057001.

[158] R. Vullers, R. van Schaijk, I. Doms, C. van Hoof, and R. Mertens, "Microwave energy harvesting," *Solid-State Electron.*, vol. 53, no. 4, pp. 684–693, Apr. 2009.

[159] P. Jaffe and J. McSpadden, "Energy conversion and transmission modules for space solar power," *Proc. IEEE*, vol. 101, no. 6, pp. 1424–1437, Jun. 2013.

[160] J. McSpadden and J. Mankins, "Space solar power programs and microwave wireless power transmission technology," *IEEE Microw. Mag.*, vol. 3, no. 4, pp. 46–57, Dec. 2002.

ABOUT THE AUTHORS

Luca Roselli (Senior Member, IEEE) was born in Florence, Italy, in 1962. He received the ''Laurea'' degree in electronic engineering from the University of Florence, Florence, Italy, in 1988.

From 1988 to 1991, he worked at the University of Florence on SAW devices. In November 1991, he joined the University of Perugia, Perugia, Italy, where he is currently working as an Associate Professor and where he has been teaching several classes on electronic devices, microwave electronics, high-frequency (HF) electronic components, and applied electronics. Since 2000, he has been coordinating research activity at the High Frequency Electronics (HFE) Laboratory. In that year, he also founded the spin-off company WiS (Wireless Solutions) Srl, operating in the field of microwave electronic systems with which he cooperated as a consultant until it joined the new company ART Srl. in 2008. From 2008 to 2012, he was the Director of the Technical and Scientific Committee of ART Srl and a member of the Board of Directors. In 2005, he founded a second spin-off company: DiES (Digital Electronic Solutions) Srl. His research interests mainly focus on the design of high-frequency electronic circuits and systems, including the development of numerical methods for electronic circuit analysis with special attention to RFID–NFC systems, new materials (including organic and recyclable ones), and far-field wireless power transfer. In these fields, he published more than 220 contributions in international reviews and peer-reviewed conferences, the interest in which is testified by an HF index of 20 (source Google Scholar) and more than 1350 citations.

Mr. Roselli was the Chairman of the VII Computational Electromagnetic in Time Domain Workshop in 2007. In 2013, he was the Chairman of the First IEEE Wireless Power Transfer Conference (WPTC). Currently, he is member of the list of experts of Italian Ministry of Research and University (MIUR); member of several IEEE Technical Committees [MTT-24 RFID Technologies (past chair), MTT-25 RF nanotechnolgies, MTT-26 Wireless Power Transfer]; member of the Sub Committee 32 RFID Technologies of International Microwave Symposium (IMS) (past chair); member of the European Research Council (ERC) Panel PE7; member of the Advisory Committee of IEEE WPTC. He is involved in the boards of several international conference (RWCOM, RFID-TA, EuCAP, MAREW). He is a reviewer for many international conferences and journals (the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, IEEE MICROWAVE AND WIRELESS COMPONENTS LETTERS, *ACES Journal*, *Radioengineering Journal*, Hindawi publishing corporation, *Elsevier Organic Electronics*, *ASP Nanoscience and Nanotechnology Letters*).

Nuno Borges Carvalho (Senior Member, IEEE) was born in Luanda, Angola, in 1972. He received the Diploma and Doctoral degrees in electronics and telecommunications engineering from the University of Aveiro, Aveiro, Portugal, in 1995 and 2000, respectively.

He is a Full Professor at the Universidade de Aveiro, Aveiro, Portugal and a Senior Research Scientist with the Instituto de Telecomunicações (IT), Universidade de Aveiro, where he coordinates the wireless communication thematic area and the Radio Systems Group. He coauthored *Intermodulation in Microwave and Wireless Circuits* (Reading, MA, USA: Artech House, 2003) and *Microwave and Wireless Measurement Techniques* (Cambridge, U.K.: Cambridge Univ. Press, 2013). He has been a reviewer and author of over 200 papers in magazines and conferences. He coholds four patents. His main research interests include software-defined radio front-ends, wireless power transmission, nonlinear distortion analysis in microwave/wireless circuits and systems, and measurement of nonlinear phenomena. He has recently been involved in the design of dedicated radios and systems for newly emerging wireless technologies.

Dr. Borges Carvalho is the Chair of the IEEE MTT-11 Technical Committee. He is a member of IEEE MTT-20, MTT-24, and MTT-26 and the IEEE Portugal Section Chair. He is an Associate Editor for the IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, IEEE MICROWAVE MAGAZINE, *Cambridge Journal on Wireless Power Transmission* and the Chair of the URSI-Portugal Metrology Group.

Federico Alimenti (Senior Member, IEEE) was born in Foligno, Italy, in 1968. He received the Laurea degree (*magna cum laude*) and the Ph.D. degree from the University of Perugia, Perugia, Italy, in 1993 and 1997 respectively, both in electronic engineering.

In 1993, he held an internship at Daimler-Benz Aerospace, Ulm, Germany. In 1996, he was recipient of the URSI young scientist award and visiting scientist at the Technical University of Munich, Munich, Germany. Since 2001, he has been with the Department of Engineering, University of Perugia, as an Aggregate Professor, teaching the class of microwave electronics. Between 2006 and 2010, he was a member of the Academic Senate in the same university. He authored more than 150 papers in refereed journals and conference proceedings. He coauthored several book chapters and a European patent (EP13161946.2) on zero-power wireless sensors. His scientific interests concern the design and the experimental characterization of radio-frequency (RF)-integrated circuits in complementary metal-oxide-semiconductor (CMOS) and BiCMOS technologies.

Prof. Alimenti was a recipient of the IET Premium (Best Paper) Award in 2013 and was the Technical Program Committee (TPC) Chair of the IEEE Wireless Power Transfer (WPT) Conference. He has been a local coordinator of the ARTEMOS project, ENIAC, call 3, 2010. His H-index is equal to 14 (source: Google Scholar).

Paolo Mezzanotte (Member, IEEE) was born in Perugia, Italy, in 1965. He received the Ph.D. degree in electrical engineering from the University of Perugia, Perugia, Italy, in 1997.

Since 1992, he has been involved with FDTD analysis of microwave structures in cooperation with the Department of Electronic and Information Engineering (DIEI), University of Perugia. In 1999, he was appointed Research Associate. Since January 2007, he has been an Associate Professor with the same university, teaching the classes on radio-frequency engineering. His research activities concern numerical methods and computer-aided design (CAD) techniques for passive microwave structures and the analysis and design of microwave and millimeter-wave circuits. More recently his research interests have mainly focused on the study of advanced technologies such as LTCC, RF-MEMS, and microwave circuits printed on green substrates. These research activities are testified by more than 100 publications in the most important specialized journals and at the main conferences of the microwave scientific community.

Prof. Mezzanotte serves as a Reviewer of a number of IEEE journals and belongs to the Top Amplifier Research Group in a European Team (TARGET) and to the Advanced MEMS for RF and Millimeterwave Communications (AMICOM) Networks of Excellence. He is a member of the IMS Technical Program Review Committee and of the Technical Committee MTT-24 RFID Technologies. He serves as an Associate Editor of the *ACES Journal*. He was the Co-Chairman of the Seventh International Workshop on Computational Electromagnetics in Time Domain (CEM-TD 2007). His present H-index (ISI journals) is equal to 13.

Giulia Orecchini received the Laurea degree in electronic engineering and the Ph.D. degree in electronic and information engineering from the University of Perugia, Perugia, Italy, in 2008 and 2012, respectively.

During her Ph.D. studies she joined the ATHENA Research Group, Georgia Institute of Technology, Atlanta, GA, USA. Her research interests concern the development of radio-frequency identification (RFID) electronic systems and technologies. She is currently working as a Research Assistant at the University of Perugia.

Marco Virili (Student Member, IEEE) was born in Terni, Italy, in 1983. He received the Laurea degree in electronic engineering from the University of Perugia, Perugia, Italy, in 2009, where he is currently working toward the Ph.D. degree at the Department of Engineering.

His scientific interests concern the design and the experimental validation of radio-frequency (RF) and microwave circuits. He has a strong experience in using the main electromagnetic computer-aided design (CAD) tools for characterization and design of microwave devices. He is focused on chipless radio-frequency identification (RFID) tags and green electronics.

Chiara Mariotti (Student Member, IEEE) was born in Assisi, Italy, in 1987. She received the Laurea degree (*magna cum laude*) in electronic and telecommunication engineering from the University of Perugia, Perugia, Italy, in 2011, where she is currently working toward the Ph.D. degree at the High Frequency Electronics (HFE) Laboratory, working on green technologies for passive and energetically autonomous radio-frequency identification (RFID) tags.

In Spring 2012, she joined the ATHENA Research Group, Georgia Institute of Technology, Atlanta, GA, USA, for six months, working on eco-compatible indoor localization systems and other types of sensors fabricated by means of inkjet printing technology. At the end of 2013, she again joined the ATHENA Research Group for six months to work on multilayer inkjet printed systems and devices such as passives and microfluidics sensors.

Ricardo Gonçalves (Student Member, IEEE) was born in Lisbon, Portugal, in 1988. He received the B.Sc. and M.Sc. (*magna cum laude*) degrees in electronics and telecommunications engineering from the Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal, in 2010 and 2012, respectively. He is currently working toward the Ph.D. degree in electrical engineering at the University of Aveiro, Aveiro, Portugal.

He is a Researcher at the Instituto de Telecomunicações, Aveiro, Portugal. In 2011, he was awarded with merit scholarship for exceptional academic performances. From 2010 to 2012, he worked at Dailywork I&D as a Developer Engineer in Embedded Electronic Systems for Vehicular Location Systems and Healthcare Platforms. His main research interests include wireless power transfer systems, radio-frequency identification (RFID), and wireless passive sensor networks. His current focus is on the development of printed antennas and microwave circuits in nonconventional materials such as textiles, ceramics, plastic, paper, and cork, for the former applications.

Pedro Pinho (Member, IEEE) was born in Vale de Cambra, Portugal, in 1974. He received the Licenciado and M.S. degrees in electrical and telecommunications engineering and the Ph.D. degree in electrical engineering from the University of Aveiro, Aveiro, Portugal, in 1995 and 2000, respectively. from the University of Aveiro, Aveiro, Portugal, in 1997, 2000, and 2004, respectively.

He is currently a Professor Adjunto at the Department of Electrical Telecommunications and Computers Engineering, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Lisbon, Portugal, and a Member of the Research Staff at the Institute for Telecommunications, Aveiro, Portugal (since 1997). He authored or coauthored more than 90 papers in conferences and international journals and four book chapters. His current research interest is in antennas for location systems, reconfigurable antennas, and antenna design for passive sensors in nonconventional materials.

View publication stats

COUPONS AVAILABLE ACTIVATE COUPONS

advertisement

RFID READER DETECTOR AND TILT-SENSITIVE RFID TAG

Technology (/technology/) > Science | by nmarquardt (/member/nmarquardt/) Follow

286,565 531 157

Posted Nov. 10, 2008 | BY-NC-SA

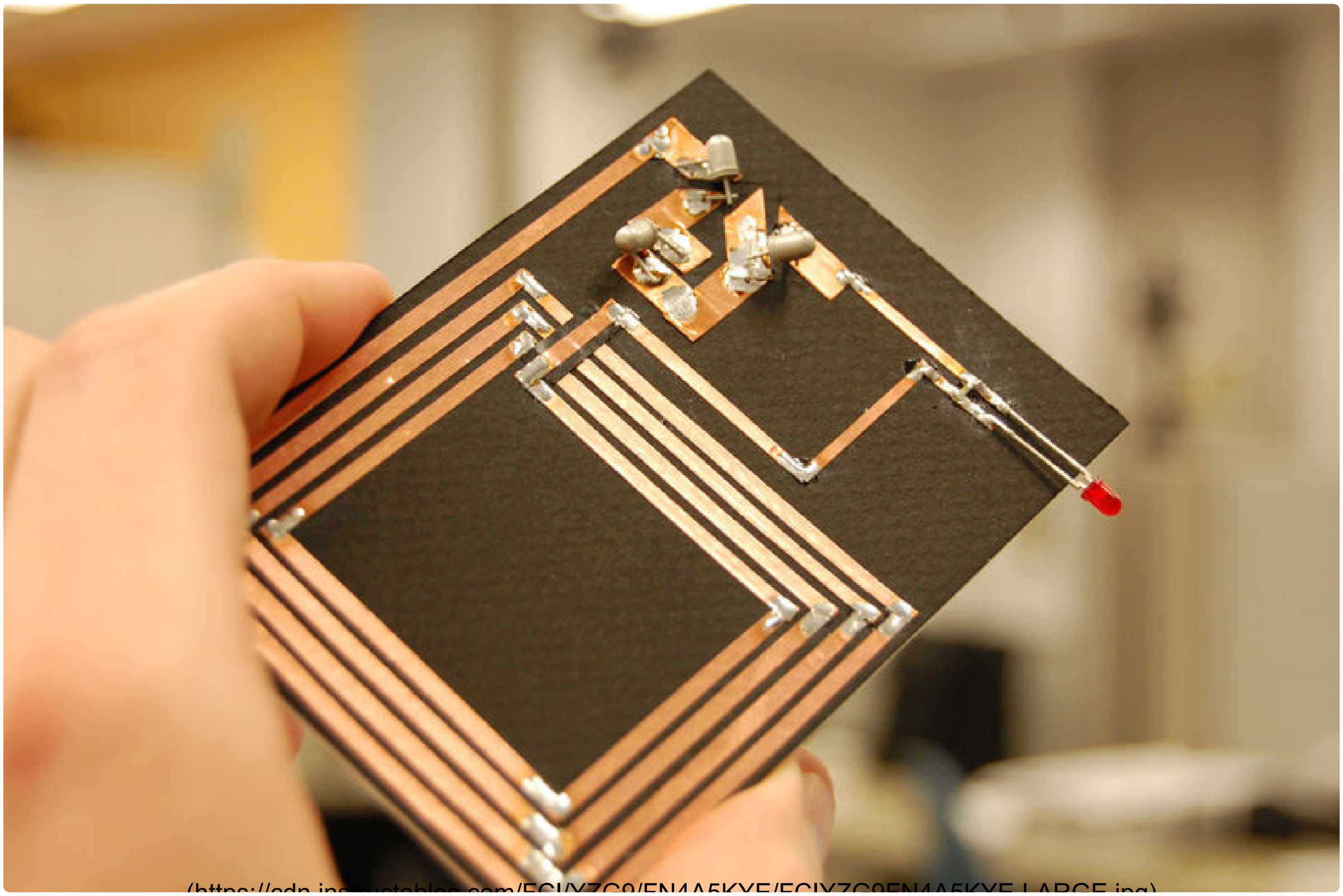

(https://cdn.instructables.com/FCLYZG9/FN4A5KYE/FCLYZG9FN4A5KYE.LARGE.jpg)

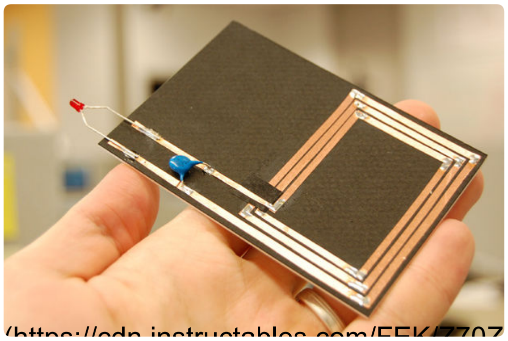

(https://cdn.instructables.com/FEK/7Z0Z

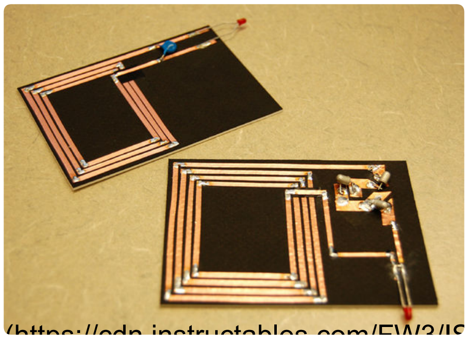

(https://cdn.instructables.com/FW3/I

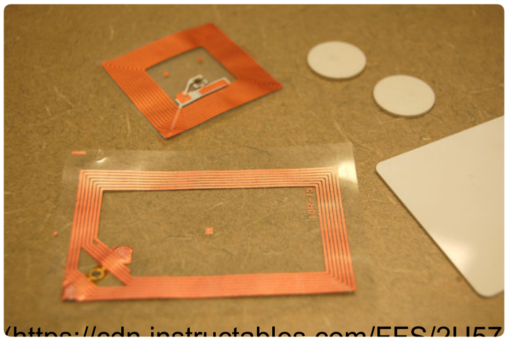

(https://cdn.instructables.com/FES/2U57

The 'rub' (http://www.usingenglish.com/reference/idioms/there%27s+the+rub.html)
Want to detect the presence of RFID readers? Want to control when a RFID tag is active or readable? We describe how to do both using bits of copper and card, and some readily available electronics hardware.

Longer preamble
Radio frequency identification (RFID (http://en.wikipedia.org/wiki/RFID)) is rapidly growing in popularity. RFID *tags* are found everywhere. They're attached to container freight, in those funny-looking white labels you find in newly purchased books, embedded in many corporate ID cards and passports, etc. The tags have a few common properties: they transmit a unique ID number, are optimized to be 'read' from predefined distances, and are usually small so they can remain unobtrusive or hidden.

RFID *readers* are used to track nearby tags by wirelessly reading a tag's unique ID (see Figure 4); a tag simply has to be brought into physical proximity with a reader to be read. Readers are mostly used for industrial or commercial purposes, e.g. asset tracking or electronic payment. Wal-mart use RFID tags and readers in their supply chain. The technology is also used in mass transit systems in cities like London (http://en.wikipedia.org/wiki/Oyster_card) and Hong Kong (http://en.wikipedia.org/wiki/Octopus_card). In Japan, many mobile phones incorporate readers to enable e-money payments in shops and vending machines.

For those of us who want to experiment with RFID, the problem is that the technology is almost always □black boxed (http://en.wikipedia.org/wiki/Black_box).

That is, the inner workings of a tag and its interaction with a reader is hidden from view, and thus difficult to have much control over.

In the two exercises that follow (building a RFID reader detector (https://www.instructables.com/id/SX28QZ1FN4H8QCG/) and a tilt-sensitive RFID tag (https://www.instructables.com/id/SGW3J0SFN49WUIF/)), we offer an example of how you can start revealing some of the workings of RFID and thus gain some control over the technology. The two exercises also hopefully show that the technology is relatively simple and how it can be extended to support some interesting interactions. We offer some other possibilities that build on our examples at the end (https://www.instructables.com/id/SQKIRD4FN49ZUS1/).

Add Tip Ask Question

Step 1: Material and Tools

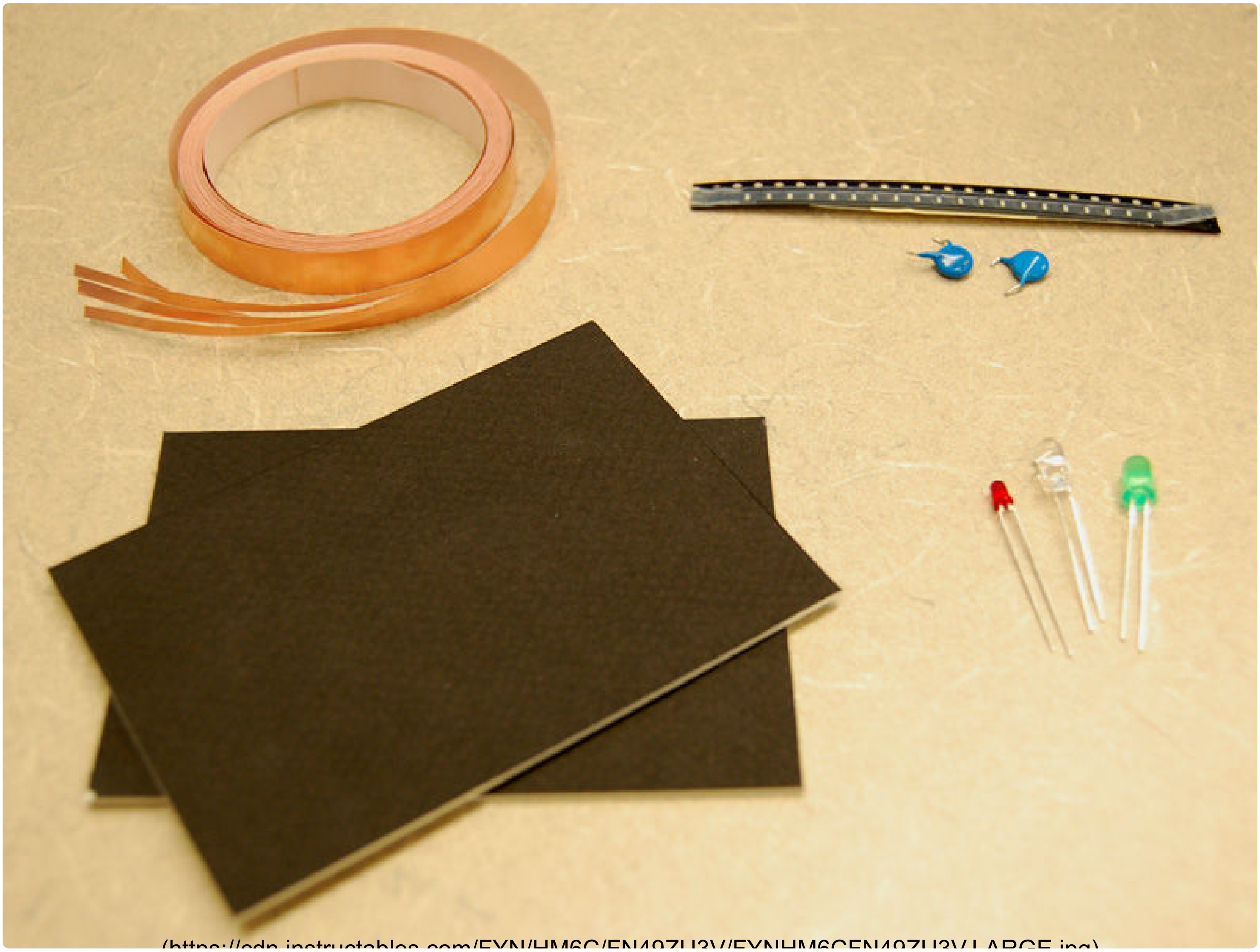

(https://cdn.instructables.com/FYN/HM6C/FN49ZU3V/FYNHM6CFN49ZU3V.LARGE.jpg)

(https://cdn.instructables.com/FIM/0KGJ/FN49ZU3U/FIM0KGJF

Weller®
WES50
Output For PES50 Iron Only. Use Only Original Weller Tips.
ESD SAFE
°Fx10

(https://cdn.instructables.com/F5C/GNJD/FN49ZU3T/F

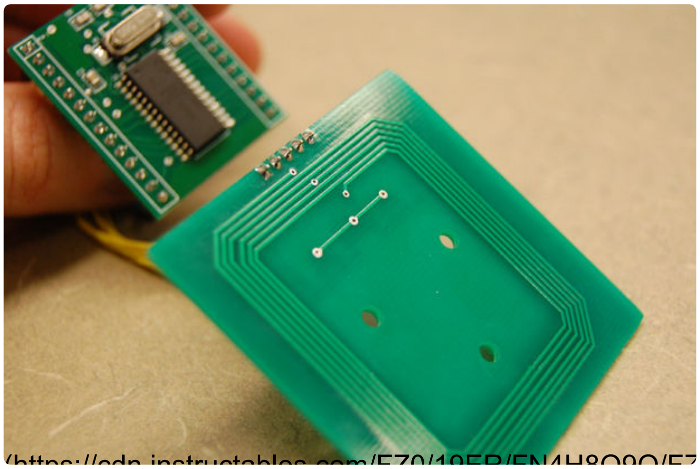

(https://cdn.instructables.com/FZ0/19FB/FN4H8O9O/FZ

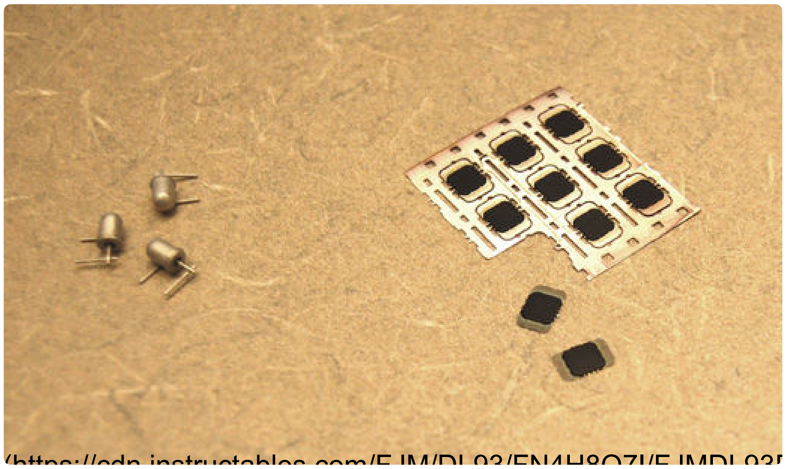

(https://cdn.instructables.com/FJM/DL93/FN4H8O7I/FJMDL93F

This section provides an overview of the necessary materials and tools.

Materials (see Figure 1):
We need the following material to built the basic RFID reader detector.
- Cardboard (around 100x70 mm)
- Conductive copper tape (e.g., order number 1218478 at www.farnell.com)
- Capacitor 82 pF (picofarad) (e.g., order number 1138852 at www.farnell.com)
- Low current LED (light-emitting diode) (e.g., order number 1003207at www.farnell.com)

Tools (see Figure 2 and 3):
- Craft knife and scissors
- Insulating tape (e.g., order number 1373979 at www.farnell.com)
- Soldering iron and solder

RFID reader for testing (see Figure 4):
To test our RFID tags we need an RFID reader that can operate at a frequency of 13.56 MHz.
There many readers for this widely used RFID standard, for instance the Sonmicro MIFARE USB reader (http://www.sonmicro.com/).
Note: The Phidget RFID reader does *not* work with the tags created in this project, as it uses a different frequency for communication with the tags (125 kHz).

Advanced material (see Figure 5):

The following material is necessary to build the second part of the project: the tilt-sensitive RFID tag.

- Micro tilt switches (e.g., www.digikey.com)
- RFID ICs (e.g., MIFARE Standard 1k, part no. 568-2219-1-ND at www.digikey.com)

Add Tip Ask Question

Step 2: Building the RFID Antenna

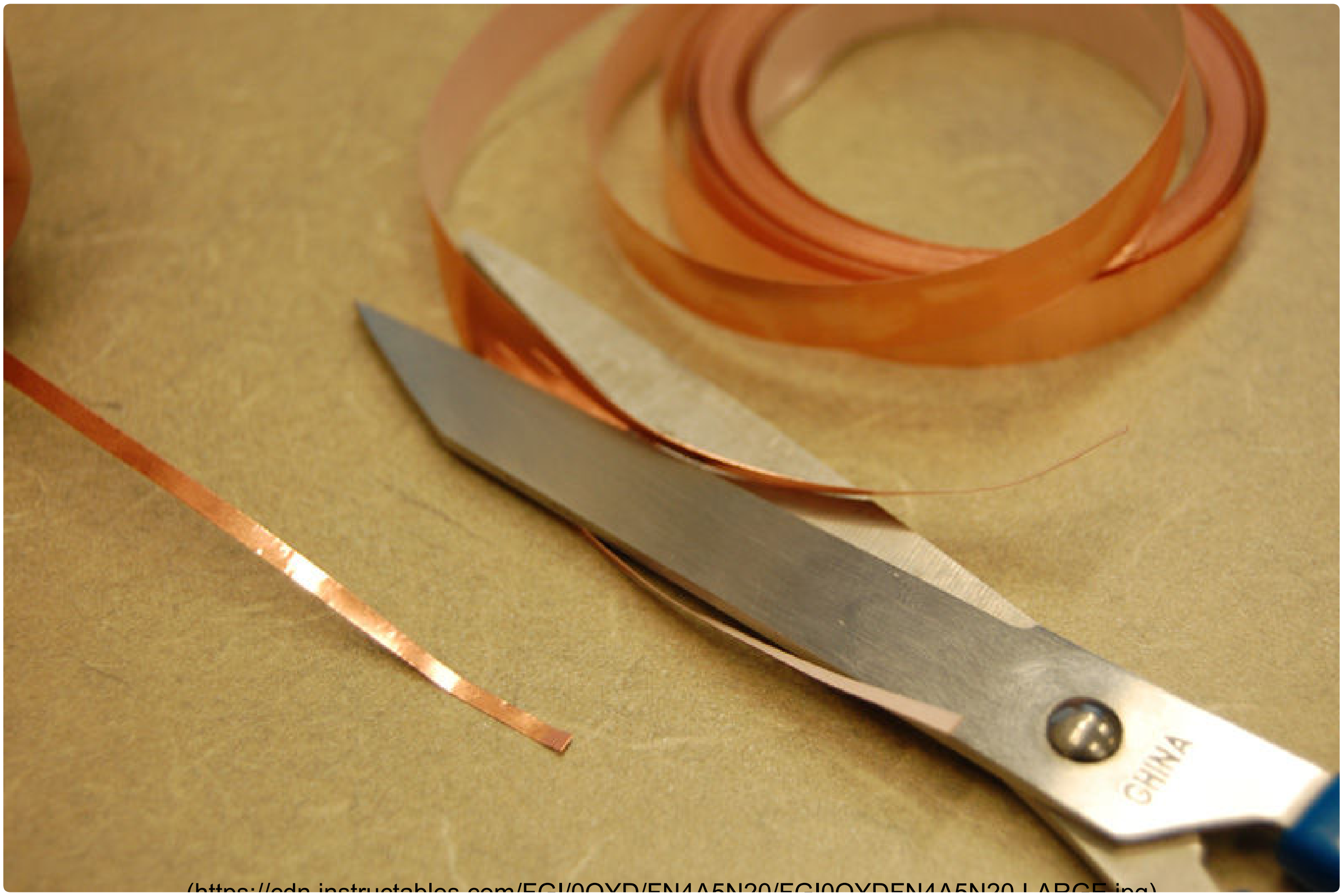

(https://cdn.instructables.com/FGI/0OXD/FN4A5N20/FGI0OXDFN4A5N20.LARGE.jpg)

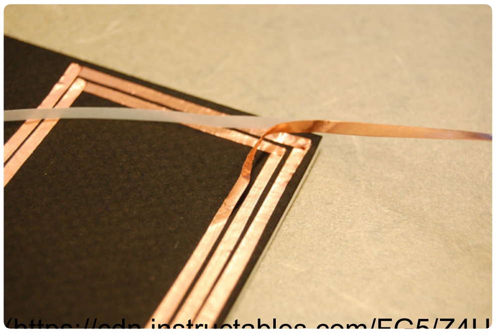
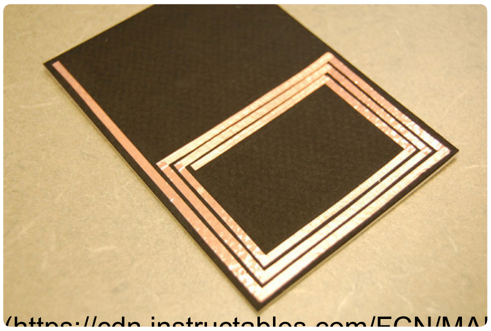
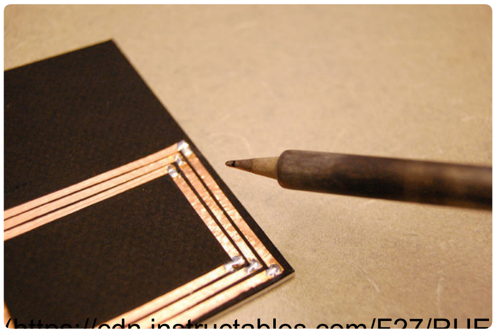

This step describes how to build the antenna for the RFID tag.

Building the RFID tag antenna
To build the tag's antenna follow these three steps.
1. Cut the conductive copper tape into thin stripes of around 2mm (see Figure 1).
2. Tape these stripes (see Figure 2) in loops around one half of the cardboard (see Figure 3 for the layout of the antenna). The tag should have between 3-4 loops for the antenna.
3. Solder all the connections between the copper tape. Sometimes, this isn't necessary as the tape's adhesive backing is conductive, but solder the connections if you want to be on the safe side.

Now we have created our RFID tag antenna, and we will add the "RFID reader detection" functionality in the following step.

A little background
RFID readers transmit an electromagnetic (EM) field with their reader antenna. This EM field induces a current in the antenna for all RFID tags within reading distance. This induced current activates the RFID chip that is connected to the tag's antenna. This chip then modulates a response (usually the unique ID number) that is transmitted back to the reader. The antenna of an RFID tag is usually a thin copper wire that is arranged in loops. The loops allow the emitted EM field of the RFID reader to induce current to the antenna of the tag.

Add Tip Ask Question

Step 3: RFID Reader Detection

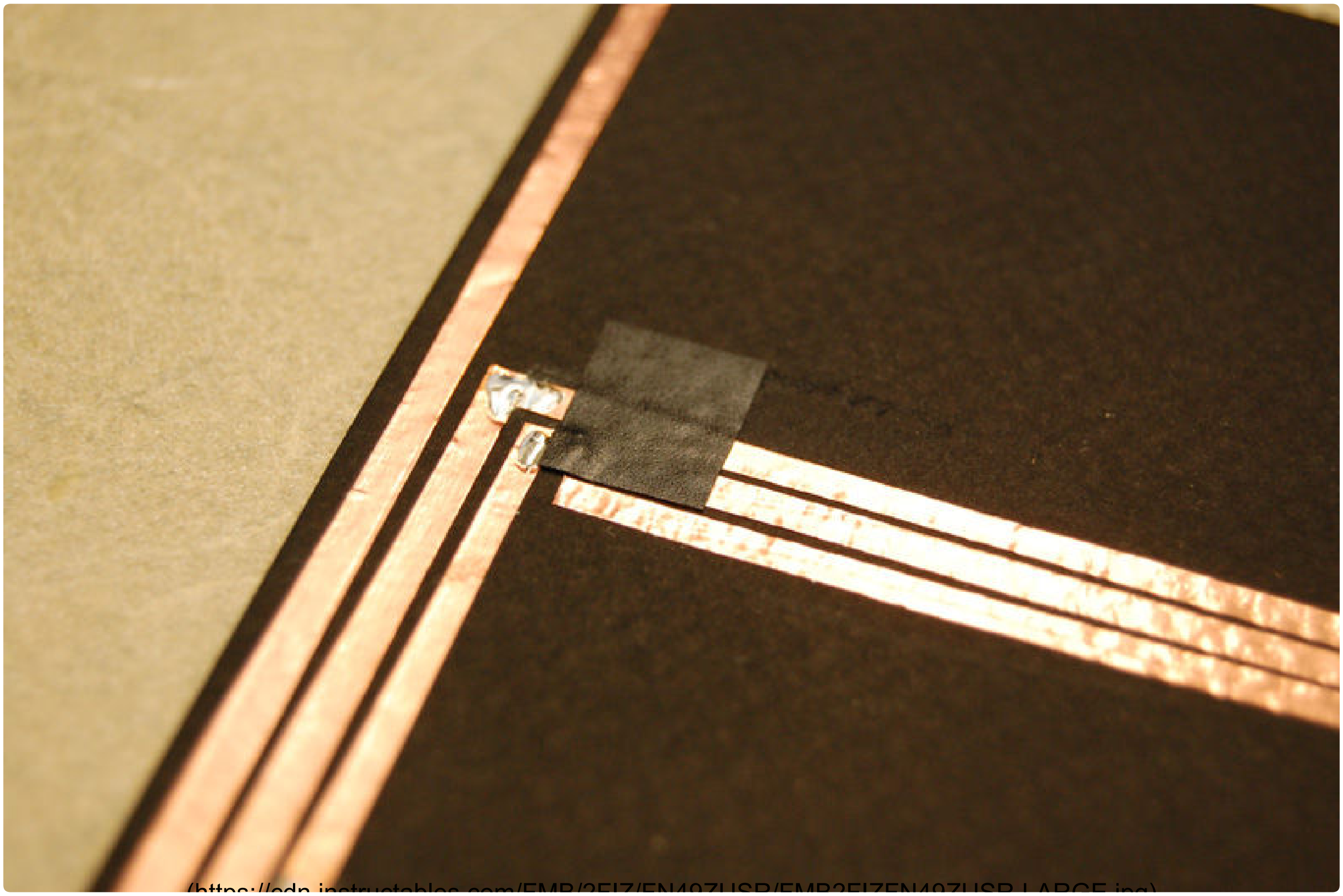

(https://cdn.instructables.com/FMB/3FIZ/FN49ZUSR/FMB3FIZFN49ZUSR.LARGE.jpg)

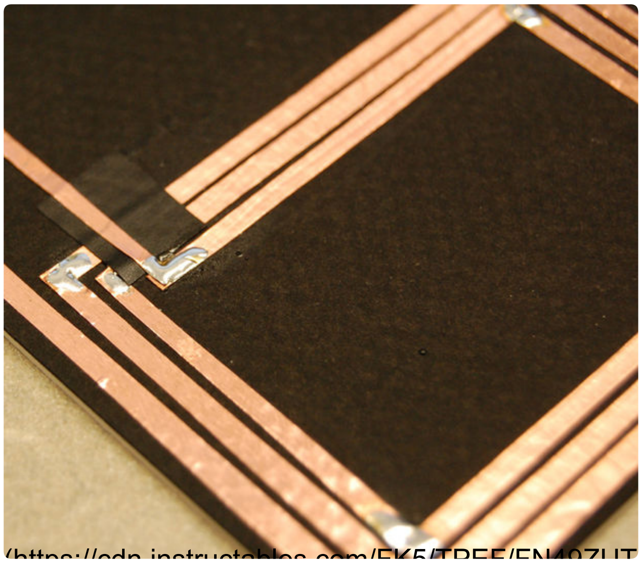

(https://cdn.instructables.com/FK5/TPEE/FN49ZUT

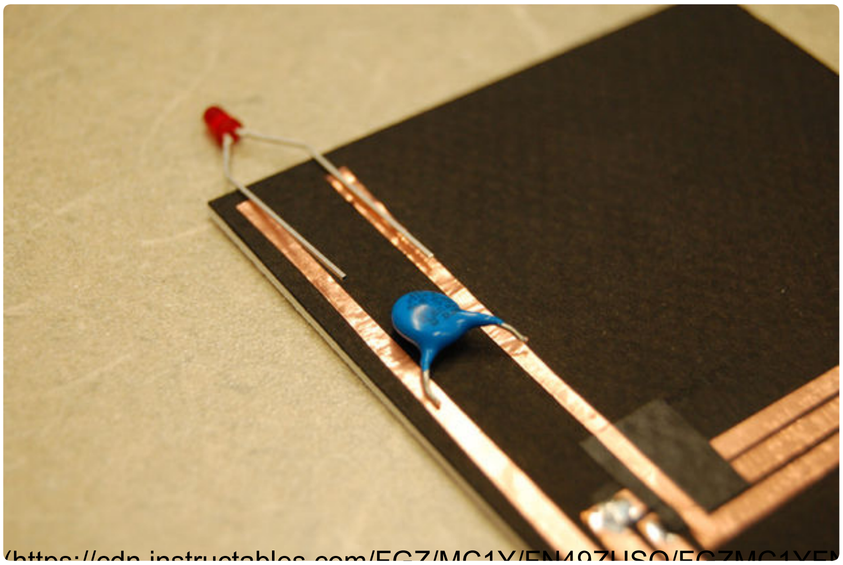

(https://cdn.instructables.com/FGZ/MC1Y/FN49ZUSQ/FGZMC1YFN

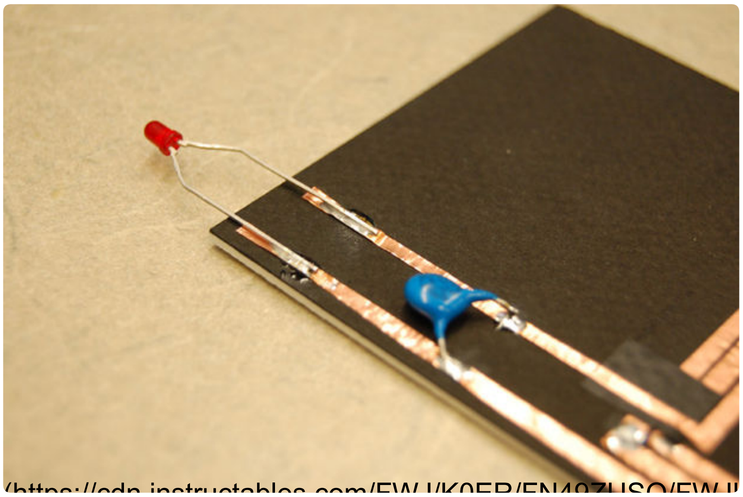
(https://cdn.instructables.com/FWJ/K0FB/FN49ZHSO/FWJ

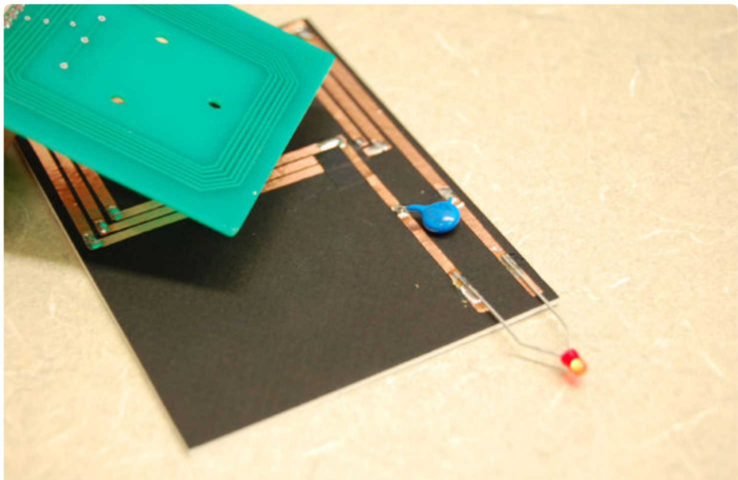
(https://cdn.instructables.com/F33/GPGW/FN49ZHSN/F33

This step describes how to add a simple mechanism to the RFID tag antenna that allows us detect nearby RFID readers.

Antenna connection

First, we add a small piece of insulation tape for the connection of the inner end of the antenna loop (as illustrated in Figure 1). This is to insulate the outer loops. Then we add another copper tape strip to the inner end of the antenna as shown in Figure 2. Here again we solder the two ends of the conductive copper tape together.

Capacitor and LED

Next, we add the capacitor (82 pF) and the low current LED to the tag as shown in Figure 3. They are connected in parallel. We also solder these two components to the copper tape (see Figure 4).

Testing

With these simple steps, our RFID reader detector is finished! By bringing our DIY RFID detector close to an RFID reader (as shown in Figure 5), the connected LED lights up. With the Sonmicro reader hardware the distance to the reader has to be below 8-10 cm; however, there are RFID readers available with a stronger EM field and therefore a higher maximum reading distance.

In the next step of the instructable we will show how to extend a basic RFID tag and make it tilt-sensitive.

Add Tip

Ask Question

Step 4: Tilt-Sensitive RFID Tag

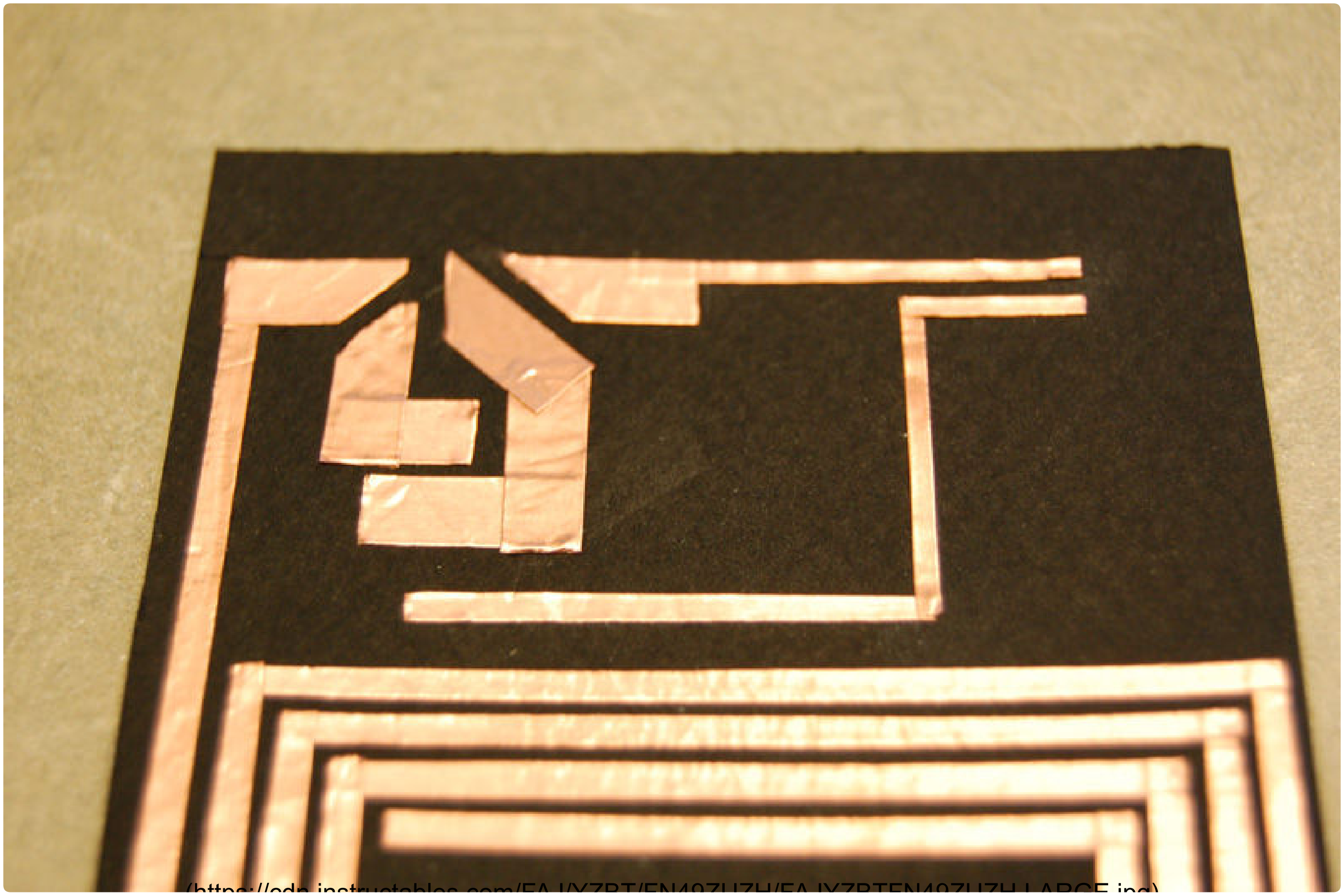

(https://cdn.instructables.com/FAJ/YZBT/FN49ZUZH/FAJYZBTFN49ZUZH.LARGE.jpg)

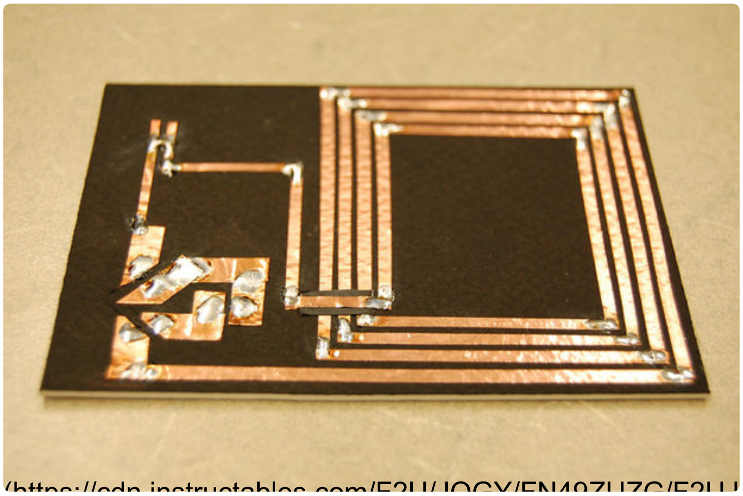

(https://cdn.instructables.com/F2U/JOGY/FN49ZUZG/F2U

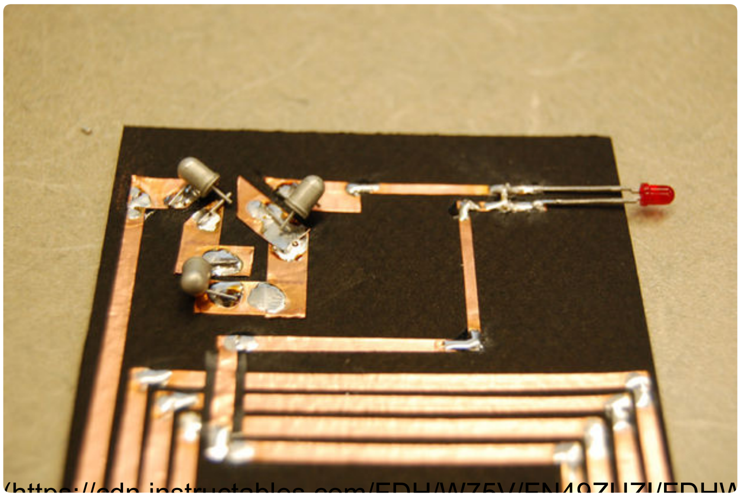

(https://cdn.instructables.com/FDH/W75V/FN49ZUZI/FDHW

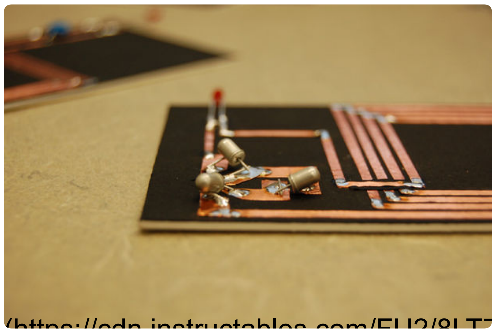

(https://cdn.instructables.com/FU2/8LT7

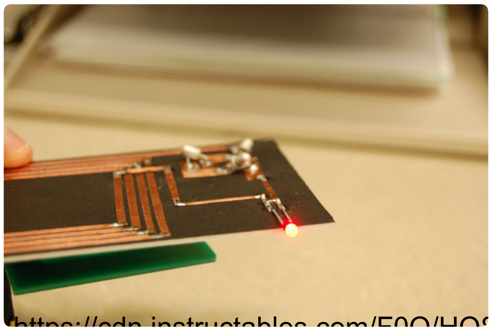

(https://cdn.instructables.com/F0Q/HQS

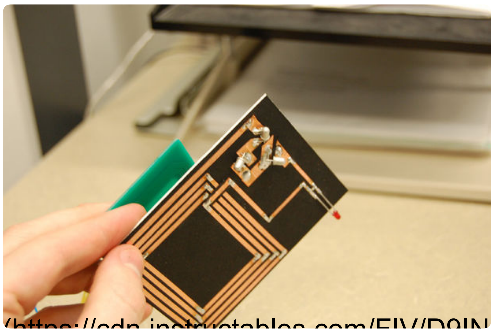

(https://cdn.instructables.com/FIV/D9IN

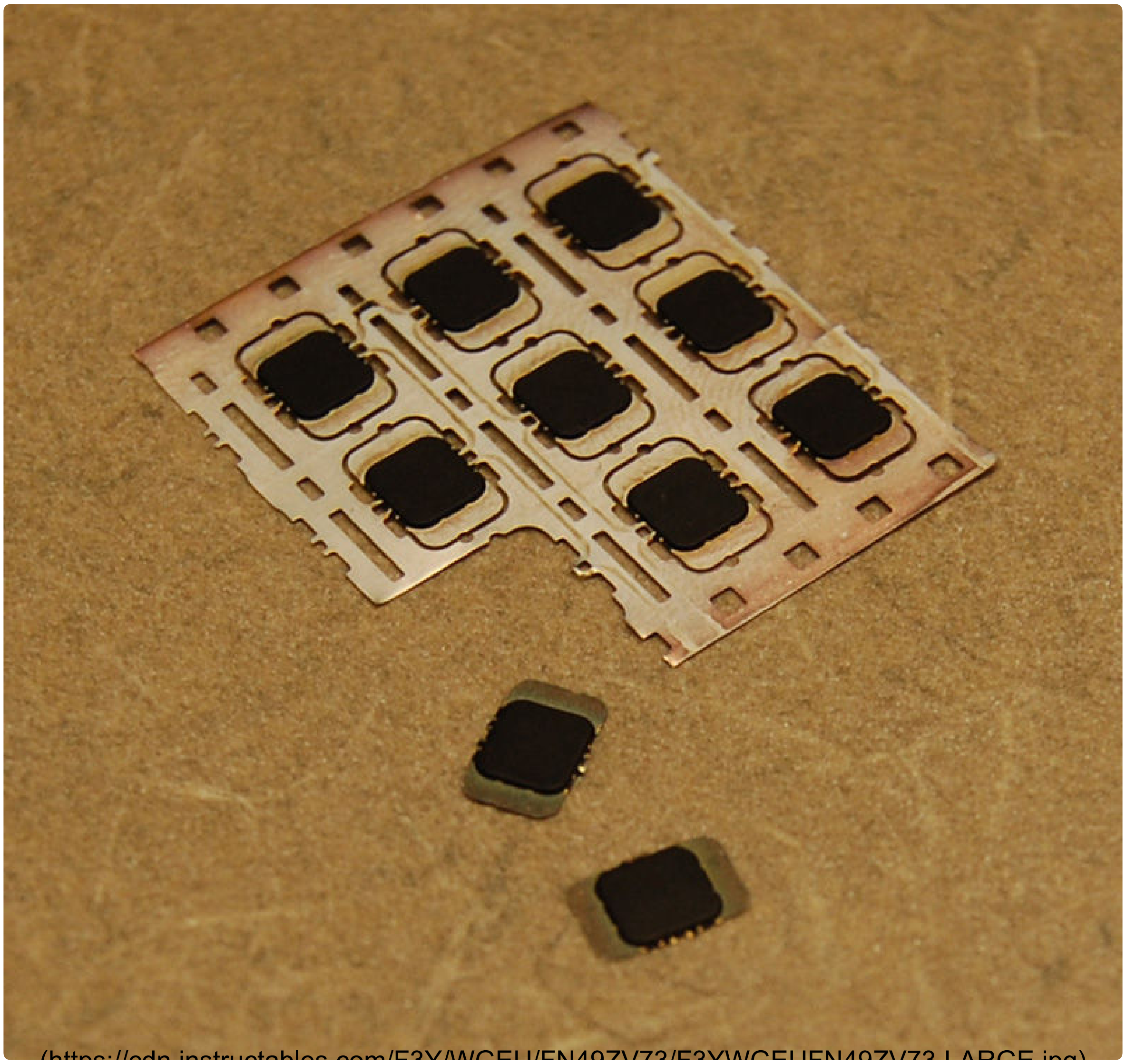

(https://cdn.instructables.com/F3Y/WGEU/FN49ZV73/F3YWGEUFN49ZV73.LARGE.jpg)

We now describe the process of how to build a tilt-sensitive RFID tag. This extends the previous exercise.

Antenna

The antenna for this second RFID tag is similar to the first antenna we built. We thus need another piece of cardboard and to repeat the steps described earlier in STEP 2 (https://www.instructables.com/id/S2YO8NWFN4H8Q8A/) of this instructable.

Tilt-sensitive tag

Next, we add additional copper tape connections to the tag, as shown in Figure 1. These connections allow us to connect three tilt switches, a capacitor, and the LED to the antenna. Again, all the connections of the copper tape are soldered together. We add the three tilt switches to the tag as shown in Figure 3. The tilt switches are soldered to the copper tape, and it is important to connect them in a slight angle (around 5-10 degrees) as shown in Figure 4. This makes sure that the silt switches are in a *closed* state while the RFID tag is in a horizontal position, and in a *open* state while the tag is in a vertical position.

Again, we also add an LED and a capacitor to the antenna as shown in Figure 3 (we use a different form factor of the capacitor here just to illustrate the alternative options).

Testing the tilt-sensitive tag

We can now use our Sonmicro RFID reader again to test our new tilt-sensitive RFID tag. The tag is activate while in a horizontal position as in Figure 5, and is inactive when in a vertical position as in Figure 6.

Using RFID chips

We can now replace the connected capacitor and LED from our tag with an RFID

MICROWAVE AND RADIO FREQUENCY ENGINEERING

INDEX

admittance 4
AM 20
Ampere's law 8
a_n parameter 15
anisotropic 21
attenuation constant 6
$\mathscr{B}$ magnetic flux density 8
B susceptance 4
beta 2
b_n parameter 15
capacitance 4
carrier 7
CDMA 20
cellular 20
characteristic admittance 3
characteristic impedance 2, 3
circulator 16
communications frequencies 20
complex permittivity 12, 21
complex propagation constant 6, 10
conductivity 12
conductor loss factor 13
copper cladding 13
cosmic rays 20
coupling factor 16
$\mathscr{D}$ electric flux density 8
dB 16
dBm 18
del 18
dielectric 21
dielectric constant 12
dielectric loss factor 13
dielectric relaxation frequency 8
directional coupler 16
directivity 16
div 18
divergence 18
$\mathscr{E}$ electric field 8
effective permittivity 13
EHF 20
electric conductivity 12
electric permittivity 12
electromagnetic spectrum 20
ELF 20
empirical 21
envelope 7
evanescent 17, 21
excitation port 14, 16
Faraday's law 8
Fourier series 3
frequency domain 8
frequency spectrum 20
gamma rays 20
Gauss' law 8
general math 18
glossary 21
grad operator 18
gradient 18
graphing 19
group velocity 7
GSM 20
$\mathscr{H}$ magnetic field 8
HF 20
high frequency 9
high frequency resistance 11
hybrid ring 15
hyperbolic functions 19
impedance 6
    intrinsic 10
    waves 10
incident wave amplitude 15
internally matched 14
intrinsic impedance 10
isotropic 21
j 18
$\mathscr{J}$ current density 8
k wave number 10
k of a dielectric 12
lambda 6
Laplacian 19
LF 20
light 20
linear 21
loss tangent 9
    complex 12
lossless network 15
low frequency 8
magnetic permeability 11
Maxwell's equations 8
MF 20
microstrip conductors 13
mode number 17
modulated wave 7
nabla operator 18
network theory 14
normalize 4
observation port 14, 16
omega-beta graph 7
overdamped 21
parallel plate capacitance 4
PCS 20
permeability 11
permittivity 12
    complex 12
    effective 13
    relative 12
phase constant 2, 6, 8, 10
phase velocity 2, 10
phasor notation 18
plane waves 10
polar notation 18
power 12, 18
    network 15
propagation constant
    complex 6, 10
quarter-wave section 6
quasi-static 9
radar 20
rat race 15
reciprocity 15
reflected wave amplitude 15
reflection coefficient 3, 10
relative permittivity 12
resistance
    high frequency 11
resistivity 12
scattering matrix 14, 16
scattering parameter 14, 15, 16
self-matched 14
separation of variables 17
series stub 5
sheet resistance 11
SHF 20
shunt stub 5
signs 2
S_{ij} scattering parameter 14
single-stub tuning 5
skin depth 7
SLF 20
Smith chart 4, 5
space derivative 18
spectrum 20
square root of j 18
stripline conductor 13
stub length 5
susceptance 4
tan δ 9
Taylor series 19
TE waves 17
telegrapher's equations 2
TEM assumptions 9
TEM waves 9
thermal speed 12
time domain 8
time of flight 3
time variable 21
time-harmonic 8
TM waves 17
transmission coefficient 3
transmission lines 2
transverse 21
transverse electromagnetic waves 9
transverse plane 17
TV 20
UHF 20
ULF 20
ultraviolet 20
underdamped 21
uniform plane waves 9
unitary matrix 15
vector differential equation 18
velocity of propagation 2, 10
v_g group velocity 7
VHF 20
VLF 20
v_p velocity of propagation 2
wave analogies 10
wave equation 2, 17
wave impedance 10
wave input impedance 11
wave number 10, 21
wavelength 6
Wheeler's equation 14
X-ray 20
Y admittance 4
y_0 characteristic admittance 3
Z_0 characteristic impedance 3
α attenuation constant 6
α_c conductor loss factor 13
α_d dielectric loss factor 13
β phase constant 6
δ loss tangent 9
δ skin depth 7
ε permittivity 12
ε_c complex permittivity 12
ε_r relative permittivity 12
γ complex propagation constant 6
η intrinsic wave impedance 10
λ wavelength 6
$\lambda/4$ 6
μ permeability 11
ρ reflection coefficient 3
ρ_v volume charge density 8
σ conductivity 12
τ transmission coefficient 3
∇ del 18
∇ divergence 18
∇ gradient 18
∇^2 Laplacian 19
∇^2 Laplacian 17

TRANSMISSION LINES

TELEGRAPHER'S EQUATIONS

$$(1)\quad \frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t} \qquad (2)\quad \frac{\partial I}{\partial z} = -C\frac{\partial V}{\partial t}$$

By taking the partial derivative with respect to *z* of equation 1 and partial with respect to *t* of equation 2, we can get:

$$(i)\quad \frac{\partial^2 V}{\partial z^2} = LC\frac{\partial^2 V}{\partial t^2} \qquad (ii)\quad \frac{\partial^2 I}{\partial z^2} = LC\frac{\partial^2 I}{\partial t^2}$$

SOLVING THE EQUATIONS

To solve the equations (i) and (ii) above, we guess that $F(u) = F(z \pm vt)$ is a solution to the equations. It is found that the unknown constant v is the wave propagation velocity.

$$\boxed{V_{total} = V_+(z - vt) + V_-(v + vt)}$$ where:

z is the position along the transmission line, where the load is at z=0 and the source is at z=-l, with l the length of the line.

v is the **velocity of propagation** $1/\sqrt{LC}$ or ω/β, the speed at which the waveform moves down the line; see p 2

t is time

THE COMPLEX WAVE EQUATION

The general solutions of equations (i) and (ii) above yield the complex wave equations for voltage and current. These are applicable when the excitation is sinusoidal and the circuit is under steady state conditions.

$$\boxed{V(z) = V^+e^{-j\beta z} + V^-e^{+j\beta z}}$$

$$\boxed{I(z) = I^+e^{-j\beta z} + I^-e^{+j\beta z}}$$

$$\boxed{I(z) = \frac{V^+e^{-j\beta z} + V^-e^{+j\beta z}}{Z_0}}$$ where:

$e^{-j\beta z}$ and $e^{+j\beta z}$ represent wave propagation in the $+z$ and $-z$ directions respectively,

$\beta = \omega\sqrt{LC} = \omega / v$ is the **phase constant,**

$Z_0 = \sqrt{L/C}$ is the **characteristic impedance** of the line. These equations represent the voltage and current **phasors**.

+/- WATCHING SIGNS

By convention z is the variable used to describe position along a transmission line with the origin z=0 set at the load so that all other points along the line are described by **negative** position values.

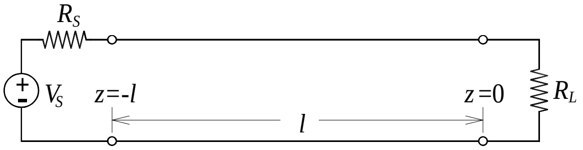

Ohm's law for right- and left-traveling disturbances:

$$V_+ = I_+Z_0 \qquad V_- = -I_-Z_0$$

v_p VELOCITY OF PROPAGATION [cm/s]

The velocity of propagation is the speed at which a wave moves down a transmission line. The velocity approaches the speed of light but may not exceed the speed of light since this is the maximum speed at which information can be transmitted. But v_p may exceed the speed of light mathematically in some calculations.

$$\boxed{v_p = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\varepsilon\mu}} = \frac{\omega}{\beta}}$$ where:

L = inductance per unit length [H/cm]
C = capacitance per unit length [F/cm]
ε = permittivity of the material [F/cm]
μ = permeability of the material [H/cm]
ω = frequency [radians/second]
β = phase constant

Phase Velocity The velocity of propagation of a TEM wave may also be referred to as the *phase velocity*. The phase velocity of a TEM wave in conducting material may be described by:

$$\boxed{v_p = \omega\delta = \frac{\omega}{k} = c\frac{2\pi\delta}{\lambda_0} = c\frac{1}{\sqrt{\varepsilon_{r\ \mathrm{eff}}}}}$$ where:

δ = skin depth [m]
c = speed of light 2.998×10^8 m/s
λ_0 = wavelength in the material [m]

Z_0 CHARACTERISTIC IMPEDANCE [Ω]

The **characteristic impedance** is the resistance initially *seen* when a signal is applied to the line. It is a physical characteristic resulting from the materials and geometry of the line.

Lossless line: $$\boxed{Z_0 \equiv \sqrt{\frac{L}{C}}} = \frac{V_+}{I_+} = -\frac{V_-}{I_-}$$

Lossy line: $$\boxed{Z_0 \equiv \sqrt{\frac{R+j\omega L}{G+j\omega C}} = |Z_0| e^{j\phi_z}}$$

L = inductance per unit length [H/cm]
C = capacitance per unit length [F/cm]
V_+ = the forward-traveling (left to right) voltage [V]
I_+ = the forward-traveling (left to right) current [I]
V_- = the reverse-traveling (right to left) voltage [V]
I_- = the reverse-traveling (right to left) current [I]
R = the line resistance per unit length [Ω/cm]
G = the conductance per unit length [Ω^{-1}/cm]
ϕ = phase angle of the complex impedance [radians]

y_0 CHARACTERISTIC ADMITTANCE [Ω^{-1}]

The **characteristic admittance** is the reciprocal of the characteristic impedance.

$$\boxed{y_0 \equiv \sqrt{\frac{C}{L}}} = \frac{I_+}{V_+} = -\frac{I_-}{V_-}$$

r REFLECTION COEFFICIENT

The **reflection coefficient** is the ratio of reflected voltage to the forward-traveling voltage, a value ranging from –1 to +1 which, when multiplied by the wave voltage, determines the amount of voltage reflected at one end of the transmission line.

$$\boxed{\rho \equiv \frac{V_-}{V_+} = -\frac{I_-}{I_+}}$$

A reflection coefficient is present at each end of the transmission line:

$$\boxed{\rho_{\text{source}} = \frac{R_S - z_0}{R_S + z_0}} \qquad \boxed{\rho_{\text{load}} = \frac{R_L - z_0}{R_L + z_0}}$$

t TRANSMISSION COEFFICIENT

The **transmission coefficient** is the ratio of total voltage to the forward-traveling voltage, a value ranging from 0 to 2.

$$\boxed{\tau \equiv \frac{V_{total}}{V_+}} = 1 + \rho$$

TOF TIME OF FLIGHT [s]

The **time of flight** is how long it takes a signal to travel the length of the transmission line

$$\boxed{TOF \equiv \frac{l}{v}} = l\sqrt{LC} = \sqrt{L_{TOT}C_{TOT}}$$

l = length of the transmission line [cm]
v = the velocity of propagation $1/\sqrt{LC}$, the speed at which the waveform moves down the line
L = inductance per unit length [H/cm]
C = capacitance per unit length [F/cm]
L_{TOT} = total inductance [H]
C_{TOT} = total capacitance [F]

DERIVED EQUATIONS

$$V_+ = z_0 I_+ = \left(V_{TOT} + I_{TOT} z_0\right)/2$$
$$V_- = -z_0 I_- = \left(V_{TOT} - I_{TOT} z_0\right)/2$$
$$I_+ = y_0 V_+ = \left(I_{TOT} + V_{TOT} y_0\right)/2$$
$$I_- = -y_0 V_- = \left(I_{TOT} - V_{TOT} y_0\right)/2$$

C_n FOURIER SERIES

The function $x(t)$ must be periodic in order to employ the Fourier series. The following is the exponential Fourier series, which involves simpler calculations than other forms but is not as easy to visualize as the trigonometric forms.

$$\boxed{C_n = \frac{1}{T}\int_{t_1}^{t_1+T} x(t)\, e^{-jn\omega_0 t}\, dt}$$

C_n = amplitude	n = the harmonic (an integer)
T = period [s]	w_0 = frequency $2\pi/T$ [radians]
t = time [s]	

The function $x(t)$ may be delayed in time. All this does in a Fourier series is to shift the phase. If you know the C_ns for $x(t)$, then the C_ns for $x(t\text{-}\alpha)$ are just $C_n e^{-jn\omega 0\alpha}$. (Here, C_ns is just the plural of C_n.)

C CAPACITANCE [F]

$$v(t)=\frac{1}{C}\int_0^t i\,d\tau+v(0) \qquad I_{cap}=C\frac{dV_{cap}}{dt}$$

$$\boxed{v(t)=v_f+(v_0-v_f)e^{-t/t}} \qquad i(t)=i_f+(i_0-i_f)e^{-t/t}$$

$$P(t)=i_0{}^2Re^{-2t/\tau}$$

$v(t)$ = voltage across the capacitor, at time t [V]

v_f = final voltage across the capacitor, steady-state voltage [V]

v_0 = initial voltage across the capacitor [V]

t = time [s]

τ = the time constant, RC [seconds]

C = capacitance [F]

Natural log: $\ln x=b \Leftrightarrow e^b=x$

C PARALLEL PLATE CAPACITANCE

$$C=\frac{\varepsilon A}{h} \qquad C_{\text{per unit length}}=\frac{\varepsilon A}{lh}=\frac{\varepsilon wl}{lh}=\frac{\varepsilon w}{h}$$

ε = permittivity of the material [F/cm]

A = area of one of the capacitor plates [cm^2]

h = plate separation [cm]

w = plate width [cm]

l = plate length [cm]

C = capacitance [F]

CAPACITOR-TERMINATED LINE

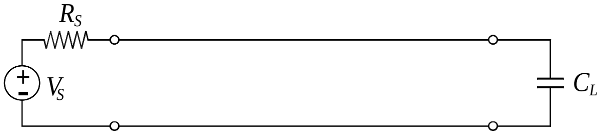

Where the incident voltage is $V_+=V_0\left(1-e^{-t/\tau_0}\right)$,

$$V_{cap}=V_++V_-=V_0\left(2+\frac{2\tau_1}{\tau_0-\tau_1}e^{-t/\tau_1}-\frac{2\tau_0}{\tau_0-\tau_1}e^{-t/\tau_0}\right)$$

V_0 = final voltage across the capacitor [V]

t = time [s]

τ_0 = time constant of the incident wave, RC [s]

τ_1 = time constant effect due to the load, Z_0C_L [s]

C = capacitance [F]

SMITH CHART

First **normalize the load impedance** by dividing by the characteristic impedance, and find this point on the chart.

When working in terms of **reactance** X, an inductive load will be located on the top half of the chart, a capacitive load on the bottom half. It's the other way around when working in terms of **susceptance** B [Siemens].

Draw a straight line from the center of the chart through the normalized load impedance point to the edge of the chart.

Anchor a compass at the center of the chart and draw a circle through the normalized load impedance point. Points along this circle represent the normalized impedance at various points along the transmission line. Clockwise movement along the circle represents movement from the load toward the source with one full revolution representing 1/2 wavelength as marked on the outer circle. The two points where the circle intersects the horizontal axis are the voltage maxima (right) and the voltage minima (left).

The point opposite the impedance (180° around the circle) is the **admittance** Y [Siemens]. The reason admittance (or susceptibility) is useful is because admittances in parallel are simply added. (**Admittance** is the reciprocal of impedance; **susceptance** is the reciprocal of reactance.)

$$\boxed{\Gamma(z)=\Gamma_L e^{j2\beta z}}$$

$$e^{j2\beta z}=1\angle 2\beta z$$

$$\mathbf{G}(z)=\frac{\mathbf{Z}(z)-1}{\mathbf{Z}(z)+1}$$

$$\mathbf{Z}_L=\frac{\Gamma_L-1}{\Gamma_L+1} \qquad \boxed{\mathbf{Z}=\frac{Z_L}{Z_0}}$$

z = distance from load [m]

$j=\sqrt{-1}$

ρ = magnitude of the reflection coefficient

β = phase constant

Γ = reflection coefficient

Z = normalized impedance [Ω]

SINGLE-STUB TUNING

The basic idea is to connect a line stub in parallel (shunt) or series a distance d from the load so that the imaginary part of the load impedance will be canceled.

Shunt-stub: Select d so that the admittance Y looking toward the load from a distance d is of the form $Y_0 + jB$. Then the stub susceptance is chosen as $-jB$, resulting in a matched condition.

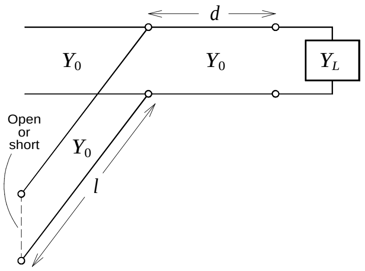

d
Y0
Y0
YL
Open
or
short
Y0
l

Series-stub: Select d so that the admittance Z looking toward the load from a distance d is of the form $Z_0 + jX$. Then the stub susceptance is chosen as $-jX$, resulting in a matched condition.

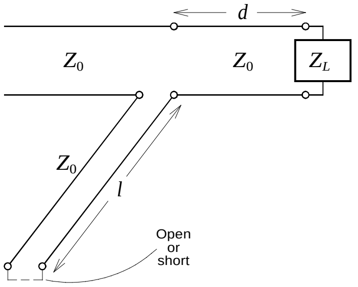

d
Z0
Z0
ZL
Z0
l
Open
or
short

FINDING A STUB LENGTH

Example: Find the lengths of open and shorted shunt stubs to match an admittance of 1-j0.5. The admittance of an open shunt (zero length) is Y=0; this point is located at the left end of the Smith Chart x-axis. We proceed clockwise around the Smith chart, i.e. away from the end of the stub, to the +j0.5 arc (the value needed to match –j0.5). The difference in the starting point and the end point on the wavelength scale is the length of the stub in wavelengths. The length of a shorted-type stub is found in the same manner but with the starting point at Y=∞.

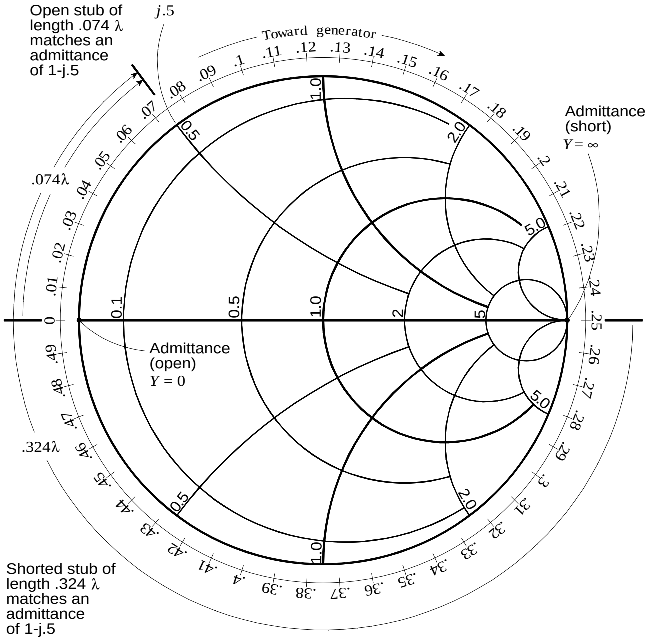
Open stub of length .074 λ matches an admittance of 1-j.5

Shorted stub of length .324 λ matches an admittance of 1-j.5

j.5
Toward generator
Admittance
(short)
Y = ∞
.074λ
.324λ
Admittance
(open)
Y = 0
0
.01
.02
.03
.04
.05
.06
.07
.08
.09
.1
.11
.12
.13
.14
.15
.16
.17
.18
.19
.2
.21
.22
.23
.24
.25
.26
.27
.28
.29
.3
.31
.32
.33
.34
.35
.36
.37
.38
.39
.4
.41
.42
.43
.44
.45
.46
.47
.48
.49
0.1
0.5
1.0
2
5
0.5
1.0
2.0
5.0
5.0
2.0
1.0
0.5

In this example, all values were in units of admittance. If we were interested in finding a stub length for a series stub problem, the units would be in impedance. The problem would be worked in exactly the same way. Of course in impedance, an open shunt (zero length) would have the value Z=∞, representing a point at the right end of the x-axis.

LINE IMPEDANCE [Ω]

The impedance seen at the source end of a lossless transmission line:

$$Z_{in} = Z_0 \frac{1+\rho}{1-\rho} = Z_0 \frac{Z_L + jZ_0 \tan(\beta l)}{Z_0 + jZ_L \tan(\beta l)}$$

For a lossy transmission line:

$$Z_{in} = Z_0 \frac{Z_L + Z_0 \tanh(\gamma l)}{Z_0 + Z_L \tanh(\gamma l)}$$

Line impedance is periodic with spatial period λ/2.

$Z_0 = \sqrt{L/C}$, the characteristic impedance of the line. [Ω]

ρ = the reflection coefficient

Z_L = the load impedance [Ω]

β = 2π/λ, phase constant

γ = α+*j*β, complex propagation constant

l WAVELENGTH [cm]

The physical distance that a traveling wave moves during one period of its periodic cycle.

$$\boxed{\lambda = \frac{2\pi}{\beta} = \frac{2\pi}{k} = \frac{v_p}{f}}$$

$\beta = \omega\sqrt{LC}$ = 2π/λ, phase constant

$k = \omega\sqrt{\mu\varepsilon}$ = 2π/λ, wave number

v_p = velocity of propagation [m/s] see p 2.

f = frequency [Hz]

l/4 QUARTER-WAVE SECTION

A quarter-wave section of transmission line has the effect of inverting the normalized impedance of the load.

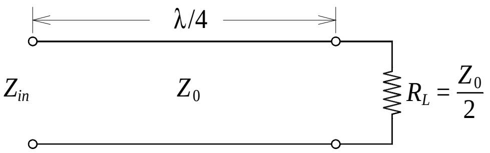

To find Z_{in}, we can normalize the load (by dividing by the characteristic impedance), invert the result, and "unnormalize" this value by multiplying by the characteristic impedance.

In this case, the normalized load is $\frac{Z_0}{2} \div Z_0 = \frac{1}{2}$

so the normalized input impedance is $\left(\frac{1}{2}\right)^{-1} = 2$

and the actual input impedance is $Z_{in} = 2Z_0$

g COMPLEX PROPAGATION CONSTANT

The propagation constant for lossy lines, taking into account the resistance along the line as well as the resistive path between the conductors.

$$\gamma = \alpha + j\beta = \sqrt{ZY} = \sqrt{(R + j\omega L)(G + j\omega C)}$$

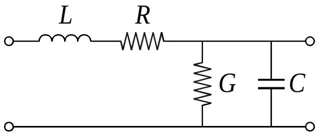

α = $\sqrt{RG}$ **attenuation constant**, the real part of the complex propagation constant, describes the loss

β = 2π/λ, **phase constant**, the complex part of the complex propagation constant

Z = **series impedance** (complex, inductive) per unit length [Ω/cm]

Y = **shunt admittance** (complex, capacitive) per unit length [Ω^{-1}/cm]

R = the resistance per unit length along the transmission line [Ω/cm]

G = the conductance between conductors per unit length [Ω^{-1}/cm]

L = inductance per unit length [H/cm]

C = capacitance per unit length [F/cm]

MODULATED WAVE

Suppose we have a disturbance composed of two frequencies:

$$\sin\left[\left(\omega_0-\delta\omega\right)t-\left(\beta_0-\delta\beta\right)z\right]$$

and $$\sin\left[\left(\omega_0+\delta\omega\right)t-\left(\beta_0+\delta\beta\right)z\right]$$

where ω_0 is the average frequency and β_0 is the average phase.

Using the identity $2\cos\left(\frac{A-B}{2}\right)\sin\left(\frac{A+B}{2}\right)=\sin A+\sin B$

The combination (sum) of these two waves is

$$\boxed{2\underbrace{\cos\left(\delta\omega t-\delta\beta z\right)}_{\text{envelope}}\underbrace{\sin\left(\omega_0 t-\beta_0 z\right)}_{\text{carrier}}}$$

The envelope moves at the group velocity, see p 7.

δ = "the difference in"…

ω_0 = carrier frequency [radians/second]

ω = modulating frequency [radians/second]

β_0 = carrier frequency phase constant

β = phase constant

So the sum of two waves will be a modulated wave having a **carrier** frequency equal to the average frequency of the two waves, and an **envelope** with a frequency equal to half the difference between the two original wave frequencies.

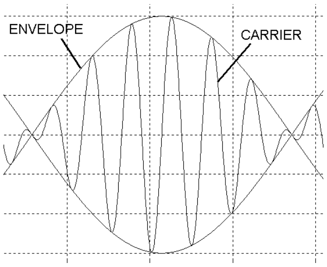

v_g GROUP VELOCITY [cm/s]

The velocity at which the envelope of a modulated wave moves.

$$\boxed{v_g=\frac{\delta\omega}{\delta\beta}=\frac{1}{\sqrt{LC_P}}\sqrt{1-\frac{\omega_c^{2}}{\omega^2}}}$$ where:

L = inductance per unit length [H/cm]

C_P = capacitance per unit length [F/cm]

ε = permittivity of the material [F/cm]

μ = permeability of the material, dielectric constant [H/cm]

ω_c = carrier frequency [radians/second]

ω = modulating frequency [radians/second]

β = phase constant

Also, since β may be given as a function of ω, remember

$$v_g=\left(\frac{d\beta}{d\omega}\right)^{-1}$$

OMEGA - BETA GRAPH

This representation is commonly used for modulated waves.

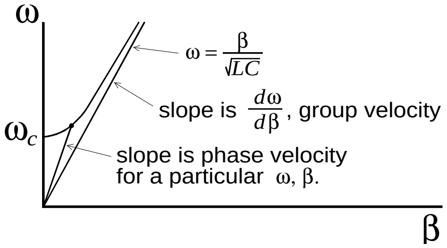
ω
$\omega=\frac{\beta}{\sqrt{LC}}$
slope is $\frac{d\omega}{d\beta}$, group velocity
ω_c
slope is phase velocity for a particular ω, β.
β

d SKIN DEPTH [cm]

The depth into a material at which a wave is attenuated by 1/e (about 36.8%) of its original intensity. This isn't the same δ that appears in the *loss tangent*, tan δ.

$$\boxed{\delta=\frac{1}{\alpha}=\sqrt{\frac{2}{\omega\mu\sigma}}}$$ where:

$\alpha = \sqrt{RG}$ **attenuation constant**, the real part of the complex propagation constant, describes loss

μ = permeability of the material, dielectric constant [H/cm]

ω = frequency [radians/second]

σ = (*sigma*) conductivity [Siemens/meter] see p12.

Skin Depths of Selected Materials

	60 Hz	1 MHz	1 GHz
silver	8.27 mm	0.064 mm	0.0020 mm
copper	8.53 mm	0.066 mm	0.0021 mm
gold	10.14 mm	0.079 mm	0.0025 mm
aluminum	10.92 mm	0.084 mm	0.0027 mm
iron	0.65 mm	0.005 mm	0.00016 mm

MAXWELL'S EQUATIONS

Maxwell's equations govern the principles of guiding and propagation of electromagnetic energy and provide the foundations of all electromagnetic phenomena and their applications. The time-harmonic expressions can be used only when the wave is sinusoidal.

	STANDARD FORM (Time Domain)	**TIME-HARMONIC** (Frequency Domain)
Faraday's Law	$\nabla\times\vec{\mathscr{E}} = -\frac{\partial\vec{\mathscr{B}}}{\partial t}$	$\nabla\times\vec{E} = -\mathrm{j}\omega\vec{B}$
Ampere's Law*	$\nabla\times\vec{\mathscr{H}} = \vec{\mathscr{J}} + \frac{\partial\vec{\mathscr{D}}}{\partial t}$	$\nabla\times\vec{H} = \mathrm{j}\omega\vec{D}+\vec{J}$
Gauss' Law	$\nabla\cdot\vec{\mathscr{D}} = \rho_v$	$\nabla\cdot\vec{D} = \rho_v$
no name law	$\nabla\cdot\vec{\mathscr{B}} = 0$	$\nabla\cdot\vec{B} = 0$

$\mathscr{E}$ = electric field [V/m]

$\mathscr{B}$ = magnetic flux density [W/m^2 or T] $\mathscr{B} = \mu_0\mathscr{H}$

t = time [s]

$\mathscr{D}$ = electric flux density [C/m^2] $\mathscr{D} = \varepsilon_0\mathscr{E}$

ρ = volume charge density [C/m^3]

$\mathscr{H}$ = magnetic field intensity [A/m]

$\mathscr{J}$ = current density [A/m^2]

*Maxwell added the $\frac{\partial\vec{\mathscr{D}}}{\partial t}$ term to Ampere's Law.

ELECTROMAGNETIC WAVES

MODELING MAXWELL'S EQUATIONS

This is a model of a wave, analogous to a transmission line model.

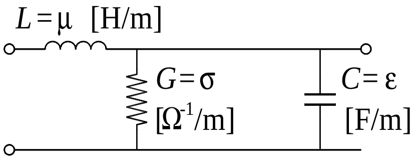

L = inductance per unit length [H/cm]

μ = permeability of the material, dielectric constant [H/cm]

G = the conductance per unit length [Ω^{-1}/cm]

σ = (*sigma*) conductivity [Siemens/meter]

C = capacitance per unit length [F/cm]

ε = permittivity of the material [F/cm]

propagation constant: $\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon+\sigma)}$

LOW FREQUENCY

At low frequencies, more materials behave as **conductors**. A wave is considered low frequency when

$\boxed{\omega \ll \frac{\sigma}{\varepsilon}}$ $\frac{\sigma}{\varepsilon}$ is the dielectric relaxation frequency

$\boxed{\eta = \frac{1}{\sigma\delta}(1+j)}$ intrinsic wave impedance, see p 12.

What happens to the complex propagation constant at low frequency? From the wave model above, gamma is

$$\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon+\sigma)} = \sqrt{j\omega\mu\sigma}\sqrt{1+\frac{j\omega\varepsilon}{\sigma}}$$

Since both ω and ε/σ are small

$$\gamma = \sqrt{j\omega\mu\sigma}\left(1+\frac{1}{2}j\omega\frac{\varepsilon}{\sigma}\right) = \sqrt{j\omega\mu\sigma}\,(1)$$

Since $\sqrt{j} = \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}$

$$\gamma = \sqrt{\omega\mu\sigma}\left(\frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}\right) = \sqrt{\frac{\omega\mu\sigma}{2}} + j\sqrt{\frac{\omega\mu\sigma}{2}}$$

So that, with $\gamma = \alpha + j\beta$

we get $\boxed{\alpha = \sqrt{\frac{\omega\mu\sigma}{2}}}$, $\boxed{\beta = \sqrt{\frac{\omega\mu\sigma}{2}}}$ or $\boxed{\gamma = \frac{1}{\delta}(1+j)}$

α = **attenuation constant**, the real part of the complex propagation constant, describes the loss

β = **phase constant**, the complex part of the complex propagation constant

σ = (*sigma*) conductivity [Siemens/cm]

δ = skin depth [cm]

So the wave is attenuating at the same rate that it is propagating.

HIGH FREQUENCY

At high frequencies, more materials behave as **dielectrics**, i.e. copper is a dielectric in the gamma ray range. A wave is considered high frequency when

$$\boxed{\omega \gg \frac{\sigma}{\varepsilon}} \qquad \frac{\sigma}{\varepsilon}$$ is the dielectric relaxation frequency

$$\boxed{\eta = \sqrt{\frac{\mu}{\varepsilon}}}$$ intrinsic wave impedance, see p 12.

What happens to the complex propagation constant at high frequency?

$$\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon + \sigma)} = \sqrt{j\omega\mu\ j\omega\varepsilon\left(1 + \frac{\sigma}{j\omega\varepsilon}\right)}$$

Since both $1/\omega$ and σ/ε are small

$$\gamma = j\omega\sqrt{\mu\varepsilon}\left(1 + \frac{1}{2}\frac{\sigma}{j\omega\varepsilon}\right) \qquad \boxed{\gamma = \frac{\sigma}{2}\sqrt{\frac{\mu}{\varepsilon}} + j\omega\sqrt{\mu\varepsilon}}$$

With $\gamma = \alpha + j\beta$

we get $\boxed{\alpha = \frac{\sigma}{2}\sqrt{\frac{\mu}{\varepsilon}}}$, $\boxed{\beta = \omega\sqrt{\mu\varepsilon}}$,

tan d LOSS TANGENT

The loss tangent, a value between 0 and 1, is the loss coefficient of a wave after it has traveled one wavelength. This is the way data is usually presented in texts. This is not the same δ that is used for skin depth.

$$\boxed{\tan\delta = \frac{\sigma}{\omega\varepsilon}}$$

Graphical representation of loss tangent:

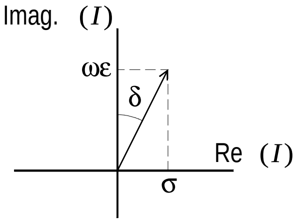

For a dielectric, $\tan\delta \ll 1$.

$$\alpha \approx \frac{1}{2}(\tan\delta)\beta = \frac{\pi}{\lambda}\tan\delta$$

$\omega\varepsilon$ is proportional to the amount of current going through the capacitance C.

σ is proportional to the amount current going through the conductance G.

TEM WAVES

Transverse Electromagnetic Waves

Electromagnetic waves that have single, orthogonal vector electric and magnetic field components (e.g., $\mathscr{E}_x$ and $\mathscr{H}_y$), both varying with a single coordinate of space (e.g., z), are known as *uniform plane waves* or *transverse electromagnetic (TEM) waves*. TEM calculations may be made using formulas from electrostatics; this is referred to as *quasi-static* solution.

Characteristics of TEM Waves

- The velocity of propagation (always in the *z* direction) is $v_p = 1/\sqrt{\mu\varepsilon}$, which is the speed of light in the material
- There is no electric or magnetic field in the direction of propagation. Since this means there is no voltage drop in the direction of propagation, it suggests that no current flows in that direction.
- The electric field is normal to the magnetic field
- The value of the electric field is η times that of the magnetic field at any instant.
- The direction of propagation is given by the direction of $\mathbf{E}\times\mathbf{H}$.
- The energy stored in the electric field per unit volume at any instant and any point is equal to the energy stored in the magnetic field.

TEM ASSUMPTIONS

Some assumptions are made for TEM waves.

$\mathscr{E}_z = 0$ $\qquad$ $\mathscr{H}_z = 0$

$\sigma = 0$ $\qquad$ time dependence $e^{j\omega t}$

WAVE ANALOGIES

Plane waves have many characteristics analogous to transmission line problems.

Transmission Lines	**Plane Waves**
Phase constant $\beta = \omega\sqrt{LC} = \frac{\omega}{v_p} = \frac{2\pi}{\lambda}$	Wave number $k = \omega\sqrt{\mu\varepsilon} = \frac{\omega}{v_p} = \frac{2\pi}{\lambda}$
Complex propagation const. $\gamma = \alpha + j\beta$ $= \sqrt{(R + j\omega L)(G + j\omega C)}$	Complex propagation constant $\gamma = \sqrt{(j\omega\mu)(j\omega\varepsilon + \sigma)}$
Velocity of propagation $v_p = \frac{1}{\sqrt{LC}} = \frac{\omega}{\beta}$	Phase velocity $v_p = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{\omega}{k} = \omega\delta = c\frac{2\pi\delta}{\lambda}$
Characteristic impedance $Z_0 = \sqrt{\frac{L}{C}} = \frac{V_+}{I_+}$	Intrinsic impedance $\eta = \sqrt{\frac{\mu}{\varepsilon}} = \frac{E_{x+}}{H_{y+}}$
Voltage $V(z) = V_+ e^{-j\beta z} + V_- e^{j\beta z}$	Electric Field $E_x(z) = E_+ e^{-jkz} + E_- e^{jkz}$
Current $I(z) = \frac{1}{Z_0}\left[V_+ e^{-j\beta z} - V_- e^{j\beta z}\right]$	Magnetic Field $H_y(z) = \frac{1}{\eta}\left[E_+ e^{-jkz} - E_- e^{jkz}\right]$
Line input impedance $Z_{in} = Z_0 \frac{Z_L + jZ_0 \tan(\beta l)}{Z_0 + jZ_L \tan(\beta l)}$ $Z_{in} = Z_0 \frac{Z_L + Z_0 \tanh(\gamma l)}{Z_0 + Z_L \tanh(\gamma l)}$	Wave input impedance $\eta_{in} = \eta_0 \frac{\eta_L + j\eta_0 \tan(kl)}{\eta_0 + j\eta_L \tan(kl)}$ $\eta_{in} = \eta_0 \frac{\eta_L + \eta_0 \tanh(\gamma l)}{\eta_0 + \eta_L \tanh(\gamma l)}$
Reflection coefficient $\rho = \frac{Z_L - Z_0}{Z_L + Z_0}$	Reflection coefficient $\rho = \frac{\eta_L - \eta_0}{\eta_L + \eta_0}$

k WAVE NUMBER [rad./cm]

The phase constant for the uniform plane wave; the change in phase per unit length. It can be considered a constant for the medium at a particular frequency.

$$\boxed{k = \frac{\omega}{v} = \omega\sqrt{\mu\varepsilon} = \frac{2\pi}{\lambda}}$$

k appears in the phasor forms of the uniform plane wave

$E_x(z) = E_1 e^{-jkz} + E_2 e^{jkz}$, etc.

k has also been used as in the "*k* of a dielectric" meaning ε_r.

h (eta) INTRINSIC WAVE IMPEDANCE [Ω]

The ratio of electric to magnetic field components. Can be considered a constant of the medium. For free space, $\eta = 376.73\Omega$. The units of η are in ohms.

$$\eta = \frac{E_{x+}}{H_{y+}} = -\frac{E_{y+}}{H_{x+}} \qquad -\eta = \frac{E_{x-}}{H_{y-}} = -\frac{E_{y-}}{H_{x-}}$$

at low frequencies	**at high frequencies**
$\eta = \frac{1}{\sigma\delta}(1 + j)$	$\eta = \sqrt{\frac{\mu}{\varepsilon}}$

When an electromagnetic wave encounters a sheet of conductive material it sees an impedance. K is the direction of the wave, **H** is the magnetic component and **E** is the electrical field. **E** × **H** gives the direction of propagation K.

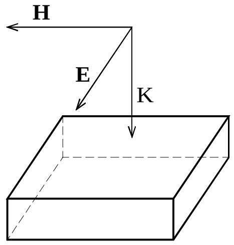

SHEET RESISTANCE [Ω]

Consider a block of material with conductivity σ.

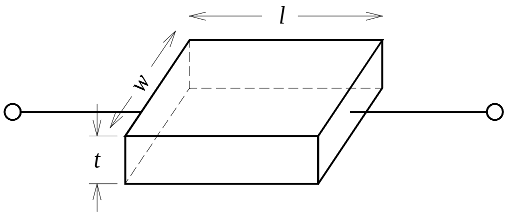

It's resistance is $R = \frac{l}{wt\sigma}$ Ω.

If the length is equal to the width, this reduces to

$$R = \frac{1}{t\sigma} \ \Omega.$$

And this is sheet resistance.

HIGH FREQUENCY RESISTANCE [Ω]

When a conductor carries current at high frequency, the electric field penetrates the outer surface only about 1 skin depth so that current travels near the surface of the conductor. Since the entire cross-section is not utilized, this affects the resistance of the conductor.

Cross-section of a conductor showing current flow near the surface:

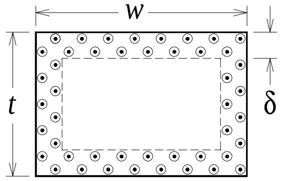

$$\boxed{R \approx \frac{1}{\sigma\delta(\text{perimeter})} = \sqrt{\frac{\omega\mu_0}{2\sigma}}\,\frac{1}{2w+2t}}$$

σ = (*sigma*) conductivity (5.8×10^5 S/cm for copper) [Siemens/meter]
ω = frequency [radians/second]
δ = skin depth [cm]
μ_0 = permeability of free space $\mu_0 = 4\pi\times10^{-9}$ [H/cm]
w = width of the conductor [cm]
t = thickness of the conductor [cm]

h$_{in}$ WAVE INPUT IMPEDANCE [Ω]

The impedance seen by a wave in a medium.

For example, the impedance of a metal sheet in a vacuum:

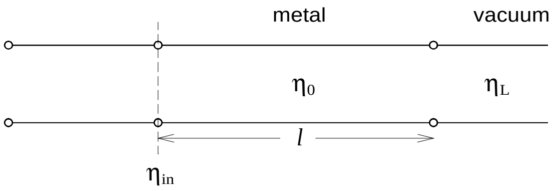

Note that a transmission line model is used here because it is analogous to a wave traveling in a medium. The "load" is the element most remote in the direction of propagation.

The input impedance is $\eta_{in} = \eta_0 \frac{\eta_L + \eta_0 \tanh(\gamma l)}{\eta_0 + \eta_L \tanh(\gamma l)}$ Ω.

In this example, *l* is the thickness of a metal sheet. If the metal thickness is much greater than the skin depth, then

$$\tanh(\lambda l) = \tanh\left[\frac{1}{\delta}(1+j)l\right] = \tanh[(\text{big number})(1+j)] \approx 1$$

If *l* is much less than the skin depth δ, then

$$\tanh(\lambda l) = \tanh\left[\frac{1}{\delta}(1+j)l\right] = \tanh[(\text{small number})(1+j)]$$

$$= (\text{same small number})(1+j) = \frac{l}{\delta}(1+j)$$

m MAGNETIC PERMEABILITY [H/m]

The relative increase or decrease in the resultant magnetic field inside a material compared with the magnetizing field in which the given material is located. The product of the permeability constant and the relative permeability of the material.

$\mu = \mu_0\mu_r$ where $\mu_0 = 4\pi\times10^{-7}$ H/m

Relative Permeabilities of Selected Materials

Air	1.00000037	Mercury	0.999968
Aluminum	1.000021	Nickel	600
Copper	0.9999833	Oxygen	1.000002
Gold	0.99996	Platinum	1.0003
Iron (99.96% pure)	280,000	Silver	0.9999736
Iron (motor grade)	5000	Titanium	1.00018
Lead	0.9999831	Tungsten	1.00008
Manganese	1.001	Water	0.9999912

e ELECTRIC PERMITTIVITY [F/m]

The property of a dielectric material that determines how much electrostatic energy can be stored per unit of volume when unit voltage is applied, also called the ***dielectric constant***. The product of the constant of permittivity and the relative permittivity of a material.

$$\varepsilon = \varepsilon_0 \varepsilon_r \quad \text{where } \varepsilon_0 = 8.85\times10^{-14}\text{ F/cm}$$

e$_c$ COMPLEX PERMITTIVITY

$$\varepsilon_c = \varepsilon' - j\varepsilon'' \quad \text{where} \quad \frac{\varepsilon''}{\varepsilon'} = \tan\delta_c$$

In general, both ε' and ε'' depend on frequency in complicated ways. ε' will typically have a constant maximum value at low frequencies, tapering off at higher frequencies with several peaks along the way. ε'' will typically have a peak at the frequency at which ε' begins to decline in magnitude as well as at frequencies where ε' has peaks, and will be zero at low frequencies and between peaks.

e$_r$ RELATIVE PERMITTIVITY

The permittivity of a material is the relative permittivity multiplied by the permittivity of free space

$$\varepsilon = \varepsilon_r \times \varepsilon_0$$

In old terminology, ε_r is called the "*k* of a dielectric". Glass (SiO_2) at ε_r = 4.5 is considered the division between low *k* and high *k* dielectrics.

Relative Permittivities of Selected Materials

Air (sea level)	1.0006	Polystyrene	2.6
Ammonia	22	Polyethylene	2.25
Bakelite	5	Rubber	2.2-4.1
Glass	4.5-10	Silicon	11.9
Ice	3.2	Soil, dry	2.5-3.5
Mica	5.4-6	Styrofoam	1.03
most metals	~1	Teflon	2.1
Plexiglass	3.4	Vacuum	1
Porcelain	5.7	Water, distilled	81
Paper	2-4	Water, seawater	72-80
Oil	2.1-2.3		

NOTE: Relative permittivity data is given for materials at **low or static frequency conditions**. The permittivity for most materials varies with frequency. The relative permittivities of most materials lie in the range of 1-25. At high frequencies, the permittivity of a material can be quite different (usually less), but will have resonant peaks.

s CONDUCTIVITY [S/m] or [1/(Ω·m)]

A measure of the ability of a material to conduct electricity, the higher the value the better the material conducts. The reciprocal is *resistivity*. Values for common materials vary over about 24 orders of magnitude. Conductivity may often be determined from skin depth or the loss tangent.

$$\sigma = \frac{n_c q_e^2 \bar{l}}{m_e v_{th}} \text{ S/m} \quad \text{where}$$

n_c = density of conduction electrons (for copper this is 8.45×10^{28}) [m^{-3}]

q_e = electron charge? 1.602×10^{-23} [C]

$\bar{l}$ = $v_{th}t_c$ the product of the thermal speed and the mean free time between collisions of electrons, the average distance an electron travels between collisions [m]

m_e = the effective electron mass? [kg]

v_{th} = thermal speed, usually much larger than the drift velocity v_d. [m/s]

Conductivities of Selected Materials [1/(Ω·m)]

Aluminum	3.82×10^7	Mercury	1.04×10^6
Carbon	7.14×10^4	Nicrome	1.00×10^6
Copper (annealed)	5.80×10^7	Nickel	1.45×10^7
Copper (in class)	6.80×10^7	Seawater	4
Fresh water	$\sim10^{-2}$	Silicon	$\sim4.35\times10^{-4}$
Germanium	~2.13	Silver	6.17×10^7
Glass	$\sim10^{-12}$	Sodium	2.17×10^7
Gold	4.10×10^7	Stainless steel	1.11×10^6
Iron	1.03×10^7	Tin	8.77×10^6
Lead	4.57×10	Titanium	2.09×10^6
		Zinc	1.67×10^7

P POWER [W]

Power is the time rate of change of energy.

Power reflected at a discontinuity: $\%\text{ power} = |\rho|^2 \times 100$

Power transmitted at a discontinuity: $\%\text{ power} = \left(1 - |\rho|^2\right)\times 100$

MICROSTRIP CONDUCTORS

How fast does a wave travel in a microstrip? The question is complicated by the fact that the dielectric on one side of the strip may be different from the dielectric on the other side and a wave may travel at different speeds in different dielectrics. The solution is to find an **effective relative permittivity** $\varepsilon_{r\,\mathrm{eff}}$ for the combination.

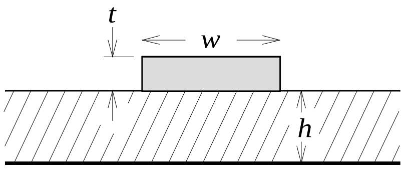

Some Microstrip Relations

$$Z_0^{\mathrm{air}} = Z_0\sqrt{\varepsilon_{r\,\mathrm{eff}}} \qquad C^{\mathrm{air}} Z_0^{\mathrm{air}} = \sqrt{\varepsilon_0\mu_0}$$

$$L = Z_0^{\mathrm{air}}\sqrt{\varepsilon_0\mu_0} = C^{\mathrm{total}}\left(Z_0\right)^2 \qquad L\,C^{\mathrm{air}} = \varepsilon_0\mu_0$$

$$Z_0 = \sqrt{\frac{L}{C^{\mathrm{total}}}} \qquad Z_0^{\mathrm{air}} = \sqrt{\frac{L}{C^{\mathrm{air}}}}$$

$$\gamma = j\beta = j\omega\sqrt{\varepsilon_0\mu_0}\sqrt{\varepsilon_{r\,\mathrm{eff}}} \qquad \varepsilon_{r\,\mathrm{eff}} = \frac{C^{\mathrm{total}}}{C^{\mathrm{air}}}$$

$$v_p = \frac{1}{\sqrt{\varepsilon_0\mu_0\varepsilon_{r\,\mathrm{eff}}}} = \frac{1}{\sqrt{L\,C^{\mathrm{total}}}}$$

It's difficult to get more than 200Ω for Z_0 in a microstrip.

Microstrip Approximations

$$\varepsilon_{r\,\mathrm{eff}} = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2\sqrt{1+12h/w}}$$

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}}\ln\left[\dfrac{8h}{w}+\dfrac{w}{4h}\right], & \text{for } \dfrac{w}{h}\le 1 \\[2ex] \dfrac{120\pi}{\sqrt{\varepsilon_{r\,\mathrm{eff}}}\left[\dfrac{w}{h}+1.393+0.667\ln\left(\dfrac{w}{h}+1.444\right)\right]}, & \text{for } \dfrac{w}{h} > 1 \end{cases}$$

$$\frac{w}{h} = \begin{cases} \dfrac{8e^{A}}{e^{2A}-2}, & \dfrac{w}{h} < 2 \\[2ex] \dfrac{2}{\pi}\left\{B-1-\ln\left(2B-1\right)+\dfrac{\varepsilon_r-1}{2\varepsilon_r}\left[\ln\left(B-1\right)+0.39-\dfrac{0.61}{\varepsilon_r}\right]\right\}, & \dfrac{w}{h} > 2 \end{cases}$$

where $A = \dfrac{Z_0}{60}\sqrt{\dfrac{\varepsilon_r+1}{2}} + \dfrac{\varepsilon_r-1}{\varepsilon_r+1}\left(0.23+\dfrac{0.11}{\varepsilon_r}\right),\ B = \dfrac{377\pi}{2Z_0\sqrt{\varepsilon_r}}$

STRIPLINE CONDUCTOR

Also called *shielded microstrip*. The effective relative permittivity is used in calculations.

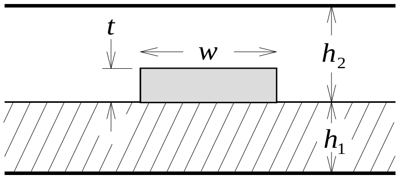

assuming $w \ge 10h$, $\quad \varepsilon_{r\,\mathrm{eff}} = \dfrac{\varepsilon_{r1}h_1+\varepsilon_{r2}h_2}{h_1+h_2}$ where

ε_{r1} = the relative permittivity of the dielectric of thickness h_1.

ε_{r2} = the relative permittivity of the dielectric of thickness h_2.

COPPER CLADDING

The thickness of copper on a circuit board is measured in ounces. 1-ounce cladding means that 1 square foot of the copper weighs 1 ounce. 1-ounce copper is 0.0014" or 35.6 µm thick.

a$_d$ DIELECTRIC LOSS FACTOR [dB/cm]

$$\alpha_d = 8.68\frac{\beta_0\varepsilon_r\left(\varepsilon_{r\,\mathrm{eff}}-1\right)}{2\sqrt{\varepsilon_{r\,\mathrm{eff}}}\left(\varepsilon_r-1\right)}\tan\delta$$

a$_c$ CONDUCTOR LOSS FACTOR [dB/cm]

$$\alpha_c = 8.68\frac{R}{2Z_0}, \qquad R = \frac{1}{\sigma\delta(\mathrm{perimeter})} = \sqrt{\frac{\omega\mu_0}{2\sigma}}\frac{1}{(\mathrm{perimeter})}$$

WHEELER'S EQUATION

Another approximation for microstrip calculations is Wheeler's equation.

$$Z_0 = \frac{42.4}{\sqrt{1+\varepsilon_r}} \ln\left(1+\left\{\frac{4h}{w'}\left[\frac{14+\frac{8}{\varepsilon_r}}{11}\times\frac{4h}{w'}+\sqrt{\left(\frac{14+\frac{8}{\varepsilon_r}}{11}\times\frac{4h}{w'}\right)^2+\pi^2\frac{1+\frac{1}{\varepsilon_r}}{2}}\right]\right\}\right)$$

where $$w' = \frac{8h\sqrt{\frac{7+\frac{4}{\varepsilon_r}}{11}\left[\exp\left(\frac{Z_0}{42.4}\sqrt{\varepsilon_r+1}\right)-1\right]+\frac{1+\frac{1}{\varepsilon_r}}{0.81}}}{\exp\left(\frac{Z_0}{42.4}\sqrt{\varepsilon_r+1}\right)-1}$$

NETWORK THEORY

S_{ij} SCATTERING PARAMETER

$$\underset{\text{observation port}\nearrow\ \ \nwarrow\text{excitation port}}{S_{\mathrm{i\,j}}}$$

A scattering parameter, represented by S_{ij}, is a dimensionless value representing the fraction of wave amplitude transmitted from port j into port i, provided that all other ports are terminated with matched loads and only port j is receiving a signal. Under these same conditions, S_{ii} is the reflection coefficient at port i.

To experimentally determine the scattering parameters, attach an impedance-matched generator to one of the ports (***excitation port***), attach impedance-matched loads to the remaining ports, and observe the signal received at each of the ports (***observation ports***). The fractional amounts of signal amplitude received at each port i will make up one column j of the **scattering matrix**. Repeating the process for each column would require n^2 measurements to determine the scattering matrix for an n-port network.

S_{ij} SCATTERING MATRIX

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1N} \\ S_{21} & S_{22} & \cdots & S_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ S_{N1} & S_{N2} & \cdots & S_{NN} \end{bmatrix}$$

The scattering matrix is an $n\times n$ matrix composed of scattering parameters that describes an n-port network.

The elements of the diagonal of the scattering matrix are reflection coefficients of each port. The elements of the off-diagonal are transmission coefficients, under the conditions outlined in "SCATTERING PARAMETER".

If the network is **internally matched** or **self-matched**, then $S_{11} = S_{22} = \cdots = S_{NN} = 0$, that is, the diagonal is all zeros.

The sum of the squares of each column of a scattering matrix is equal to one, provided the network is lossless.

a_n, b_n INCIDENT/REFLECTED WAVE AMPLITUDES

The parameters a_n and b_n describe the incident and reflected waves respectively at each port n. These parameters are used for power and scattering matrix calculations.

The amplitude of the wave incident to port n is equal to the amplitude of the incident voltage at the port divided by the square root of the port impedance.

$$a_n = \frac{V_n^+}{\sqrt{Z_{0n}}}$$

Amplitude of the wave reflected at port n is equal to the amplitude of the reflected voltage at the port divided by the square root of the port impedance.

$$b_n = \frac{V_n^-}{\sqrt{Z_{0n}}}$$

The scattering parameter is equal to the wave amplitude output at port i divided by the wave amplitude input at port j provided the only source is a matched source at port j and all other ports are connected to matched loads.

$$S_{\mathrm{ij}} = \frac{b_{\mathrm{i}}}{a_{\mathrm{j}}}$$

The relationship between the S-parameters and the a- and b-parameters can be written in matrix form where **S** is the scattering matrix and **a** and **b** are column vectors.

$$\mathbf{b} = \mathbf{Sa}$$

Power flow into any port is shown as a function of a- and b-parameters.

$$P = \frac{1}{2}\left(|a|^2 - |b|^2\right)$$

The ratio of the input power at port j to the output power at port I can be written as a function of a- and b-parameters or the S-parameter.

$$\frac{P_{in\,\mathrm{j}}}{P_{out\,\mathrm{i}}} = \frac{|a_{\mathrm{j}}|^2}{|b_{\mathrm{j}}|^2} = \frac{1}{|S_{\mathrm{ij}}|^2}$$

RECIPROCITY

A network is reciprocal when $S_{\mathrm{ij}} = S_{\mathrm{ji}}$ in the scattering matrix, i.e. the matrix is symmetric across the diagonal. Also, $Z_{\mathrm{ij}} = Z_{\mathrm{ji}}$ and $Y_{\mathrm{ij}} = Y_{\mathrm{ji}}$. Networks constructed of "normal materials" exhibit reciprocity.

Reciprocity Theorem:

$$\oint_S \vec{E}_a \times \vec{H}_b \cdot ds = \oint_S \vec{E}_b \times \vec{H}_a \cdot ds$$

E_a and H_b are fields from two different sources.

LOSSLESS NETWORK

A network is lossless when

$$\underline{S}\;\underline{S}^{\dagger} = /$$

† means to take the complex conjugate and transpose the matrix. If the network is reciprocal, then the transpose is the same as the original matrix.

/ = a unitary matrix. A **unitary matrix** has the properties:

$$\sum_{k=1}^{N} S_{ki} S_{ki}^{*} = 1 \qquad \sum_{k=1}^{N} S_{ki} S_{kj}^{*} = 0$$

In other words, a column of a unitary matrix multiplied by its complex conjugate equals one, and a column of a unitary matrix multiplied by the complex conjugate of a different column equals zero.

RAT RACE OR HYBRID RING NETWORK

The rat race or hybrid ring network is lossless, reciprocal, and internally matched.

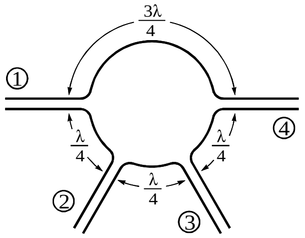

The signal splits upon entering the network and half travels around each side. A signal entering at port 1 and exiting at port 4 travels ¾ of a wavelength along each side, so the signals are in phase and additive. From port 1 to port 3 the signal travels one wavelength along one side and ½ wavelength along the other, arriving a port 3 out of phase and thus canceling. From port 1 to port 2 the paths are ¼ and 5/4 wavelengths respectively, thus they are in phase and additive.

DIRECTIONAL COUPLER

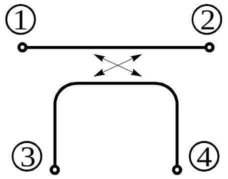

The directional coupler is a 4-port network similar to the rat race. It can be used to measure reflected and transmitted power to an antenna.

An input at one port is divided between two of the remaining ports. The **coupling factor**, measured in dB, describes the division of signal strength at the two ports. For example if the coupler has a coupling factor of –10 dB, then a signal input at port 1 would appear at port 4 attenuated by 10 dB with the majority of the signal passing to port 2. In other words, 90% of the signal would appear at port 2 and 10% at port 4. (-10 dB means "**10 dB down**" or 0.1 power, -6 dB means 0.25 power, and –3 dB means 0.5 power.) A reflection from port 2 would appear at port 3 attenuated by the same amount. Meters attached to ports 3 and 4 could be used to measure reflected and transmitted power for a system with a transmitter connected to port 1 and an antenna at port 2. The **directivity** of a coupler is a measurement of how well the coupler transfers the signal to the appropriate output without reflection due to the coupler itself; the directivity approaches infinity for a perfect coupler. $\text{directivity} = 10\log\left(p_3/p_1\right)$, where the source is at port 1 and the load is at port 2.

The directional coupler is **lossless** and **reciprocal**. The scattering matrix looks like this. In a real coupler, the off-diagonal zeros would be near zero due to leakage.

$$\begin{bmatrix} 0 & p & 0 & -q \\ p & 0 & q & 0 \\ 0 & q & 0 & p \\ -q & 0 & p & 0 \end{bmatrix}$$

CIRCULATOR

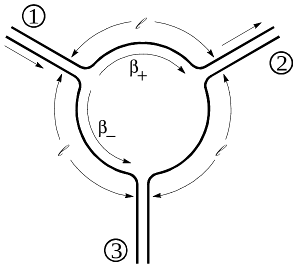

The circulator is a 3-port network that can be used to prevent reflection at the antenna from returning to the source.

Port 3 is terminated internally by a matched load. With a source at 1 and a load at 2, any power reflected at the load is absorbed by the load resistance at port 3. A 3-port network cannot be both lossless and reciprocal, so the circulator is not reciprocal.

Schematically, the circulator may be depicted like this:

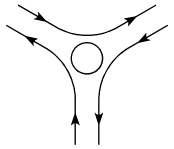

The circulator is **lossless** but is not **reciprocal**. The scattering matrix looks like this:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

MAXWELL'S EQUATIONS, TIME HARMONIC FORM

$$\nabla \times \bar{\mathscr{E}} = -j\omega\mu\mathbf{H} \qquad \text{"curl on E"}$$

$$\nabla \times \bar{\mathscr{H}} = -j\omega\mu\mathbf{E} \qquad \text{"curl on H"}$$

$$\bar{\mathscr{E}} = \left[E_x(x,y)\hat{\mathbf{x}} + E_y(x,y)\hat{\mathbf{y}} + E_z(x,y)\hat{\mathbf{z}}\right]e^{j\omega t-\gamma z}$$

$$\bar{\mathscr{H}} = \left[H_x(x,y)\hat{\mathbf{x}} + H_y(x,y)\hat{\mathbf{y}} + H_z(x,y)\hat{\mathbf{z}}\right]e^{j\omega t-\gamma z}$$

From the curl equations we can derive:

(1) $\dfrac{\partial E_z}{\partial y} + \gamma E_y = -j\omega\mu H_x$ (4) $\dfrac{\partial H_z}{\partial y} + \gamma H_y = j\omega\varepsilon E_x$

(2) $-\dfrac{\partial E_z}{\partial x} - \gamma E_x = -j\omega\mu H_y$ (5) $-\dfrac{\partial H_z}{\partial x} - \gamma H_x = j\omega\varepsilon E_y$

(3) $\dfrac{\partial E_y}{\partial x} - \dfrac{\partial E_x}{\partial y} = -j\omega\mu H_z$ (6) $\dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = j\omega\varepsilon E_z$

From the above equations we can obtain:

(1) & (5) $H_x = \dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(j\omega\varepsilon\dfrac{\partial E_z}{\partial y} - \gamma\dfrac{\partial H_z}{\partial x}\right)$

(2) & (4) $H_y = \dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(j\omega\varepsilon\dfrac{\partial E_z}{\partial x} - \gamma\dfrac{\partial H_z}{\partial y}\right)$

(2) & (4) $E_x = -\dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(-\gamma\dfrac{\partial E_z}{\partial x} + j\omega\mu\dfrac{\partial H_z}{\partial y}\right)$

(1) & (5) $E_y = -\dfrac{1}{\gamma^2 + \omega^2\mu\varepsilon}\left(-\gamma\dfrac{\partial E_z}{\partial y} + j\omega\mu\dfrac{\partial H_z}{\partial x}\right)$

This makes it look like if E_z and H_z are zero, then H_x, H_y, E_x, and E_y are all zero. But since $\infty \times 0 \neq 0$, we could have non-zero result for the TEM wave if $\gamma^2 = -\omega^2\mu\varepsilon \Rightarrow \gamma = j\omega\sqrt{\mu\varepsilon}$. This should look familiar.

WAVE EQUATIONS

From Maxwell's equations and a vector identity on curl, we can get the following wave equations:

$$\nabla^2 \overline{\mathscr{E}} = -\omega^2 \mu\varepsilon \overline{\mathscr{E}} \qquad \text{"del squared on E"}$$

$$\nabla^2 \overline{\mathscr{H}} = \omega^2 \mu\varepsilon \overline{\mathscr{H}} \qquad \text{"del squared on H"}$$

The z part or "del squared on E_z" is:

$$\nabla^2 E_z = \frac{\gamma^2 E_z}{\partial x^2} + \frac{\gamma^2 E_z}{\partial y^2} + \frac{\gamma^2 E_z}{\partial z^2} = -\omega^2 \mu\varepsilon E_z$$

Using the separation of variables, we can let:

$$E_z = X(x) \cdot Y(y) \cdot Z(z)$$

We substitute this into the previous equation and divide by $X \cdot Y \cdot Z$ to get:

$$\underbrace{\frac{1}{X}\frac{d^2 X}{dx^2}}_{-k_x{}^2} + \underbrace{\frac{1}{Y}\frac{d^2 Y}{dy^2}}_{-k_y{}^2} + \underbrace{\frac{1}{Z}\frac{d^2 Z}{dz^2}}_{-k_z{}^2} = \underbrace{-\omega^2 \mu\varepsilon}_{\text{a constant}}$$

Since X, Y, and Z are independent variables, the only way the sum of these 3 expressions can equal a constant is if all 3 expressions are constants.

So we are letting $\frac{1}{Z}\frac{d^2 Z}{dz^2} = -k_z{}^2 \Rightarrow \frac{d^2 Z}{dz^2} = -Zk_z{}^2$

A solution could be $Z = e^{-\gamma z}$

so that $\gamma^2 e^{-\gamma z} = -k_z{}^2 e^{-\gamma z}$ and $-k_z{}^2 = \gamma^2$

Solutions for X and Y are found

$$\frac{1}{X}\frac{d^2 X}{dx^2} = -k_x{}^2 \Rightarrow X = A\sin(k_x x) + B\cos(k_x x)$$

$$\frac{1}{Y}\frac{d^2 Y}{dy^2} = -k_y{}^2 \Rightarrow Y = C\sin(k_y y) + D\cos(k_y y)$$

giving us the general solution $k_x{}^2 + k_y{}^2 - \gamma^2 = \omega^2 \mu\varepsilon$

For a particular solution we need to specify initial conditions and boundary conditions. For some reason, initial conditions are not an issue. The unknowns are k_x, k_y, A, B, C, D. The boundary conditions are

$$E_{\tan} = 0 \qquad \frac{\partial H_{\tan}}{\partial n} = 0$$

$E_{\tan}$ = the electric field tangential to a conducting surface
$H_{\tan}$ = the magnetic field tangential to a conducting surface
n = I don't know

TM, TE WAVES IN PARALLEL PLATES

TM, or transverse magnetic, means that magnetic waves are confined to the *transverse plane*. Similarly, TE (transverse electric) means that electrical waves are confined to the transverse plane.

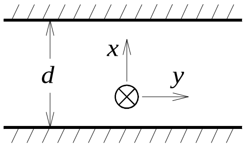

(z direction is into page)

Transverse plane means the plane that is transverse to (perpendicular to) the direction of propagation. The direction of propagation is taken to be in the z direction, so the transverse plane is the x-y plane. So for a TM wave, there is no H_z component (magnetic component in the z direction) but there is an E_z component.

$$\boxed{E_z = A\sin(k_x x)e^{-\gamma z}}$$

A = amplitude [V]

$k_x = \frac{m\pi}{d}$ The magnetic field must be zero at the plate boundaries. This value provides that characteristic. [cm^{-1}]

x = position; perpendicular distance from one plate. [cm]

d = plate separation [cm]

γ = propagation constant

z = position along the direction of propagation [cm]

m = mode number; an integer greater than or equal to 1

$$\boxed{\gamma = \sqrt{-\omega^2 \mu\varepsilon + (kx)^2}}$$

Notice than when $(kx)^2 \geq \omega^2 \mu\varepsilon$, the quantity under the square root sign will be positive and γ will be purely real. In this circumstance, the wave is said to be evanescent. The wavelength goes to infinity; there is no oscillation or propagation. On the other hand, when $(kx)^2 < \omega^2 \mu\varepsilon$, γ is purely imaginary.

The magnitude of E_z is related to its position between the plates and the mode number m. Note that for m = 2 that $d = \lambda$.

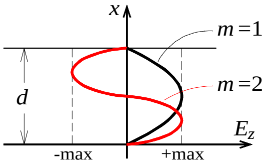

GENERAL MATHEMATICAL

COMPLEX TO POLAR NOTATION

j in polar notation:

$$j = e^{j\frac{\pi}{2}}$$

So we can find the square root of *j*:

$$\sqrt{j} = \sqrt{e^{j\frac{\pi}{2}}} = e^{j\frac{\pi}{4}} = \frac{1}{\sqrt{2}} + j\frac{1}{\sqrt{2}}$$

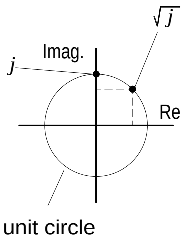

dBm DECIBELS RELATIVE TO 1 mW

The decibel expression for power. The logarithmic nature of decibel units translates the multiplication and division associated with gains and losses into addition and subtraction.

0 dBm = 1 mW
20 dBm = 100 mW
-20 dBm = 0.01 mW

$$P(\text{dBm}) = 10\log\left[P(\text{mW})\right]$$

$$P(\text{mW}) = 10^{P(\text{dBm})/10}$$

PHASOR NOTATION

To express a derivative in phasor notation, replace $\frac{\partial}{\partial t}$ with $j\omega$. For example, the

Telegrapher's equation $\frac{\partial V}{\partial z} = -L\frac{\partial I}{\partial t}$

becomes $\frac{\partial V}{\partial z} = -Lj\omega I$.

Ñ NABLA, DEL OR GRAD OPERATOR

Compare the ∇ operation to taking the time derivative. Where $\partial/\partial t$ means to take the derivative with respect to time and introduces a s^{-1} component to the units of the result, the ∇ operation means to take the derivative with respect to distance (in 3 dimensions) and introduces a m^{-1} component to the units of the result. ∇ terms may be called space derivatives and an equation which contains the ∇ operator may be called a vector differential equation. In other words $\nabla\mathbf{A}$ is how fast **A** changes as you move through space.

in rectangular coordinates: $\nabla\mathbf{A} = \hat{x}\frac{\partial A}{\partial x} + \hat{y}\frac{\partial A}{\partial y} + \hat{z}\frac{\partial A}{\partial z}$

in cylindrical coordinates: $\nabla\mathbf{A} = \hat{r}\frac{\partial A}{\partial r} + \hat{\phi}\frac{1}{r}\frac{\partial A}{\partial \phi} + \hat{z}\frac{\partial A}{\partial z}$

in spherical coordinates: $\nabla\mathbf{A} = \hat{r}\frac{\partial A}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial A}{\partial \theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial A}{\partial \phi}$

∇ GRADIENT

$$\boxed{\nabla\vec{\Phi} = -\mathbf{E}}$$

"The gradient of the vector Φ" or "del Φ" is equal to the negative of the electric field vector.

$\nabla\Phi$ is a vector giving the direction and magnitude of the maximum spatial variation of the scalar function Φ at a point in space.

$$\nabla\vec{\Phi} = \hat{\mathbf{x}}\frac{\partial\,\Phi}{\partial\,x} + \hat{\mathbf{y}}\frac{\partial\,\Phi}{\partial\,y} + \hat{\mathbf{z}}\frac{\partial\,\Phi}{\partial\,z}$$

Ñx DIVERGENCE

$\nabla\cdot$ is also a vector operator, combining the "del" or "grad" operator with the dot product operator and is read as "the divergence of". In this form of Gauss' law, where **D** is a density per unit area, with the operators applied, $\nabla\cdot\mathbf{D}$ becomes a density per unit volume.

$$\boxed{\text{div}\,\mathbf{D} = \nabla\cdot\mathbf{D} = \frac{\partial\,D_x}{\partial\,x} + \frac{\partial\,D_y}{\partial\,y} + \frac{\partial\,D_z}{\partial\,z} = \rho}$$

D = electric flux density vector $\mathbf{D} = \varepsilon\mathbf{E}$ $[\text{C/m}^2]$
ρ = source charge density $[\text{C/m}^3]$

Ñ^2 THE LAPLACIAN

∇^2 is a combination of the divergence and del operations, i.e. $\mathrm{div}(\mathrm{grad}\ \Phi) = \nabla\cdot\nabla\ \Phi = \nabla^2\ \Phi$. It is read as "the LaPlacian of" or "del squared".

$$\boxed{\nabla^2\mathbf{F} = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2}}$$

F = electric potential [V]

GRAPHING TERMINOLOGY

With *x* being the horizontal axis and *y* the vertical, we have a graph of ***y* versus *x*** or ***y* as a function of *x***. The *x*-axis represents the **independent variable** and the *y*-axis represents the **dependent variable**, so that when a graph is used to illustrate data, the data of regular interval (often this is time) is plotted on the *x*-axis and the corresponding data is dependent on those values and is plotted on the *y*-axis.

HYPERBOLIC FUNCTIONS

$$j\sin\theta = \sinh(j\theta)$$

$$j\cos\theta = \cosh(j\theta)$$

$$j\tan\theta = \tanh(j\theta)$$

TAYLOR SERIES

$$\sqrt{1+x} \approx 1 + \frac{1}{2}x, \quad x \ll 1$$

$$\frac{1}{1-x^2} \approx 1 + x^2 + x^4 + x^6 + \cdots, \quad |x| < 1$$

$$\frac{1}{1\pm x} \approx 1 \mp x, \quad x \ll 1$$

ELECTROMAGNETIC SPECTRUM

FREQUENCY	WAVELENGTH (free space)	DESIGNATION	APPLICATIONS
< 3 Hz	> 100 Mm		Geophysical prospecting
3-30 Hz	10-100 Mm	ELF	Detection of buried metals
30-300 Hz	1-10 Mm	SLF	Power transmission, submarine communications
0.3-3 kHz	0.1-1 Mm	ULF	Telephone, audio
3-30 kHz	10-100 km	VLF	Navigation, positioning, naval communications
30-300 kHz	1-10 km	LF	Navigation, radio beacons
0.3-3 MHz	0.1-1 km	MF	AM broadcasting
3-30 MHz	10-100 m	HF	Short wave, citizens' band
30-300 MHz 54-72 76-88 88-108 174-216	1-10 m	VHF	TV, FM, police TV channels 2-4 TV channels 5-6 FM radio TV channels 7-13
0.3-3 GHz 470-890 MHz 915 MHz 800-2500 MHz 1-2 2.45 2-4	10-100 cm	UHF "money band"	Radar, TV, GPS, cellular phone TV channels 14-83 Microwave ovens (Europe) PCS cellular phones, analog at 900 MHz, GSM/CDMA at 1900 L-band, GPS system Microwave ovens (U.S.) S-band
3-30 GHz 4-8 8-12 12-18 18-27	1-10 cm	SHF	Radar, satellite communications C-band X-band (Police radar at 11 GHz) K$_u$-band (dBS Primestar at 14 GHz) K-band (Police radar at 22 GHz)
30-300 GHz 27-40 40-60 60-80 80-100	0.1-1 cm	EHF	Radar, remote sensing K$_a$-band (Police radar at 35 GHz) U-band V-band W-band
0.3-1 THz	0.3-1 mm	Millimeter	Astromony, meteorology
10^{12}-10^{14} Hz	3-300 μm	Infrared	Heating, night vision, optical communications
3.95×10^{14}- 7.7×10^{14} Hz	390-760 nm 625-760 600-625 577-600 492-577 455-492 390-455	Visible light	Vision, astronomy, optical communications Red Orange Yellow Green Blue Violet
10^{15}-10^{18} Hz	0.3-300 nm	Ultraviolet	Sterilization
10^{16}-10^{21} Hz		X-rays	Medical diagnosis
10^{18}-10^{22} Hz		γ-rays	Cancer therapy, astrophysics
> 10^{22} Hz		Cosmic rays	Astrophysics

GLOSSARY

anisotropic materials materials in which the electric polarization vector is not in the same direction as the electric field. The values of ε, μ, and σ are dependent on the field direction. Examples are crystal structures and ionized gases.

complex permittivity e The imaginary part accounts for heat loss in the medium due to damping of the vibrating dipole moments.

dielectric An insulator. When the presence of an applied field displaces electrons within a molecule away from their average positions, the material is said to be polarized. When we consider the polarizations of insulators, we refer to them as *dielectrics*.

empirical A result based on observation or experience rather than theory, e.g. *empirical data*, *empirical formulas*. Capable of being verified or disproved by observation or experiment, e.g. *empirical laws*.

evanescent wave A wave for which $\beta=0$. α will be negative. That is, γ is purely real. The wave has infinite wavelength—there is no oscillation.

isotropic materials materials in which the electric polarization vector is in the same direction as the electric field. The material responds in the same way for all directions of an electric field vector, i.e. the values of ε, μ, and σ are constant regardless of the field direction.

linear materials materials which respond proportionally to increased field levels. The value of μ is not related to $\mathbf{H}$ and the value of ε is not related to $\mathbf{E}$. Glass is linear, iron is non-linear.

overdamped system in the case of a transmission line, this means that when the source voltage is applied the line voltage rises to the final voltage without exceeding it.

time variable materials materials whose response to an electric field changes over time, e.g. when a sound wave passes through them.

transverse plane perpendicular, e.g. the x-y plane is *transverse* to z.

underdamped system in the case of a transmission line, this means that after the source voltage is applied the line voltage periodically exceeds the final voltage.

wave number k The phase constant for the uniform plane wave. k may be considered a constant of the medium at a particular frequency.

LEARNING MADE EASY

Ixia Special Edition

5G

for dummies®

A Wiley Brand

Innovations that make 5G possible

Achieve scale through network virtualization

See the future of the IoT and New Radio

Compliments of

ixia

Kalyan Sundhar

Lawrence C. Miller

About Ixia

Ixia, recently acquired by Keysight Technologies, provides testing, visibility, and security solutions, strengthening applications across networks and cloud environments for enterprises, service providers, and network equipment manufacturers. Ixia offers companies trusted environments in which to develop, deploy, and operate. Customers worldwide rely on Ixia to verify their designs, optimize their performance, and ensure protection of their networks and cloud environments to make their applications stronger. Learn more at **www.ixiacom.com**.

About Keysight Technologies

Keysight Technologies is a leading technology company that helps its engineering, enterprise, and service provider customers optimize networks and bring electronic products to market faster and at a lower cost. Keysight's solutions go where the electronic signal goes, from design simulation, to prototype validation, to manufacturing test, to optimization in networks and cloud environments. Customers span the worldwide communications ecosystem, aerospace and defense, automotive, energy, semiconductor, and general electronics end markets. Keysight generated revenues of $2.9 billion in fiscal year 2016. In April 2017, Keysight acquired Ixia, a leader in network test, visibility, and security. More information is available at **www.keysight.com**.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

5G

Ixia Special Edition

by Kalyan Sundhar and Lawrence C. Miller

for dummies®
A Wiley Brand

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

5G For Dummies®, Ixia Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
`www.wiley.com`

Copyright © 2017 by John Wiley & Sons, Inc., Hoboken, New Jersey

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at `http://www.wiley.com/go/permissions`.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. Ixia and the Ixia logo are trademarks or registered trademarks of Ixia Corporation. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact `info@dummies.biz`, or visit `www.wiley.com/go/custompub`. For information about licensing the *For Dummies* brand for products or services, contact `BrandedRights&Licenses@Wiley.com`.

ISBN 978-1-119-42415-4 (pbk); ISBN 978-1-119-42416-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. Some of the people who helped bring this book to market include the following:

Project Editor: Martin V. Minner

Senior Acquisitions Editor: Amy Fandrei

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Production Editor: Vasanth Koilraj

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Table of Contents

INTRODUCTION 1
About This Book 2
Foolish Assumptions 2
Icons Used in This Book 3
Beyond the Book 3
Where to Go from Here 3

CHAPTER 1: **Understanding the Journey to a 5G Future** 5
Tracing the Evolution of Wireless Communications 5
Focusing on the 5G Vision 8
Ordering Up 5G in Five Easy Pieces 9

CHAPTER 2: **Achieving Faster Speeds and Larger Feeds** 11
Fattening the Data Pipe 11
We All Bundle — with CA 13
Eeny, Meeny, Miny, MIMO 14
No Qualms About QAM 15

CHAPTER 3: **Tapping into Unlicensed Spectrum** 17
Giving a “High 5(G)” to Wi-Fi Advancements 17
LTE in Unlicensed Spectrum 18
License Assisted Access 19
MulteFire 20

CHAPTER 4: **Enabling Massive IoT** 21
Key Connectivity Requirements for IoT Devices 21
NarrowBand IoT (NB-IoT) 22
Long Term Evolution for Machines (LTE-M) 23
LoRaWAN and Sigfox 23

CHAPTER 5: **Getting Real About the Need to Virtualize** 25
Driving 5G and IoT with Virtualization 25
    Recognizing the IoT explosion 26
    Focusing on service 26
    Managing the migration 30

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Virtualizing Network Components 30
    Virtual EPC (vEPC) 30
    Centralized Radio Access Network (C-RAN) 31
    Mobile Edge Computing (MEC) 31

CHAPTER 6: Creating a New 5G World Order with New Radio (NR) 33
5G NR Basics 33
More Spectrum — mmWave Bands 35

CHAPTER 7: Exploring 5G Use Cases 37
Fixed Wireless Broadband Service 37
Entertainment Everywhere 38
Virtual Reality (VR) and Augmented Reality (AR) 39
Connected and Autonomous Vehicles 40
The Connected World 41
Augmented Humans 42

CHAPTER 8: Ten Myths About 5G — Debunked 43

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Introduction

The next-generation mobile network (NGMN) is on the horizon. 5G, the next iteration of 4G Long Term Evolution (LTE) networks, will enable significantly greater mobile speeds — as much as 20 gigabits per second (Gbps) with less than one millisecond (ms) latency — to enable real-time connectivity for mission-critical and potentially lifesaving devices and applications. 5G will also provide truly ubiquitous connectivity in the most challenging and remote areas of the world whether on land, in the air, or at sea — even on the 42nd floor of an office building in downtown Chicago! Finally, 5G networks will connect billions of Internet of Things (IoT) devices with a wide variety of speed and data volume requirements.

But 5G is an ambitious goal. Work on key technologies to enable 5G has already begun. In much the same way that 4G LTE was rolled out in 2008, but is only now achieving the 4G LTE goal of 1 Gbps speeds with the 4G LTE Advanced standard, 5G will be a steady evolution that begins with commercial availability expected in 2020. Many technologies that have emerged in the evolution of 4G LTE, such as carrier aggregation (CA) and multiple input multiple output (MIMO), will continue to develop to achieve the massive speed and scale required in 5G. Innovative new technologies will leverage unlicensed spectrum — where Wi-Fi operates — to offload certain traffic from the carrier networks to create more capacity in their data pipes. Low-power technologies such as NarrowBand IoT (NB-IoT), LTE for Machines (LTE-M), Long Range Wide Area Network (LoRaWAN), Sigfox, and others, will be used in billions of IoT devices, and a new radio interface technology — 5G New Radio (5G NR) — will be developed for connections between User Equipment (UE) and carrier enhanced Node B (eNodeB) stations. Finally, carriers will fully embrace virtualization technologies in their core networks to enable massive scale and efficiency.

In this book, you learn about the technological innovations that are being developed today to enable a 5G future. You also learn about potential use cases that will transform entire businesses and industries, and create new business models and opportunities.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

About This Book

5G For Dummies, Ixia Special Edition, consists of eight short chapters that explore

- How wireless communications technology has evolved and where it's going next (Chapter 1).
- How developments in today's networks are blazing the 5G trail to higher speeds (Chapter 2).
- Which technologies in unlicensed spectrum will be leveraged in the 5G networks of the future (Chapter 3).
- Why the Internet of Things requires 5G connectivity (Chapter 4).
- Where virtualization in mobile networks can help address the need for scale and elasticity (Chapter 5).
- What 5G New Radio (5G NR) is and how it will help create a 5G future (Chapter 6).
- How 5G will be used in various use case scenarios (Chapter 7).
- Ten common myths and the reality of 5G (Chapter 8).

Foolish Assumptions

It's been said that most assumptions have outlived their uselessness, but we assume a few things nonetheless!

Mainly, we assume that you either work in a technology profession or you're an avid user of wireless communications technology — if you have a smartphone within arm's distance of you, we're talking about you!

Beyond a basic knowledge of wireless communications and mobile technology in general, we don't assume you have a particularly strong technical background. As such, this book is written primarily for nontechnical readers — we explain any technical terms and concepts that come up in this book.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

If any of these assumptions describe you, this book is for you! If none of these assumptions describe you, keep reading anyway. It's a great book, and when you finish reading it, you'll know enough about 5G to be dangerous!

Icons Used in This Book

Throughout this book, we occasionally use special icons to call attention to important information. Here's what to expect:

REMEMBER

This icon points out information you should commit to your non-volatile memory, your gray matter, or your noggin' – along with anniversaries and birthdays!

TECHNICAL STUFF

You won't find a map of the human genome here, but if you seek to attain the seventh level of NERD-vana, perk up! This icon explains the jargon beneath the jargon.

TIP

Tips are appreciated, never expected — and we sure hope you'll appreciate these tips! This icon points out useful nuggets of information.

Beyond the Book

There's only so much we can cover in 48 short pages, so if you find yourself at the end of this book thinking "gosh, this is a great book; where can I learn more?" just go to `www.ixiacom.com`.

Where to Go from Here

With our apologies to Lewis Carroll, Alice, and the Cheshire cat:

"Would you tell me, please, which way I ought to go from here?"

"That depends a good deal on where you want to get to," said the Cat — er, the Dummies Man.

"I don't much care where . . . ," said Alice.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

“Then it doesn’t matter which way you go!”

That’s certainly true of *5G For Dummies*, which, like Alice in Wonderland, is also destined to become a timeless classic!

If you don’t know where you’re going, any chapter will get you there — but Chapter 1 might be a good place to start! However, if you see a particular topic that piques your interest, feel free to jump ahead to that chapter. Each chapter is written to stand on its own, so you can read this book in any order that suits you (though we don’t recommend upside down or backward).

We promise you won’t get lost falling down the rabbit hole!

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- **Recapping a century of innovation in wireless communications**
- **Addressing speed, scale, and responsiveness with 5G networks**
- **Unlocking the five key fundamentals of 5G**

Chapter 1

Understanding the Journey to a 5G Future

In this chapter, you take a glimpse back at the evolution of wireless communications and a look ahead to the 5G future.

Tracing the Evolution of Wireless Communications

For more than a century, radio technology has been enabling wireless communications over ever greater distances and with ever greater capabilities.

In the late nineteenth century, Guglielmo Marconi built the first wireless telegraphy system, capable of transmitting Morse code via radio signals up to one-half mile. Today, more than seven billion mobile devices enable us to communicate with anyone, anywhere in the world.

The first truly mobile two-way radio was developed in 1923 and used in Australian police cars — although it took up the entire back seat of a patrol car. Hand-held radios — "walkie-talkies" — were first used in World War II.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

In 1973, the first call on a hand-held cellular phone was made — the cellular phone was described as a "brick" weighing nearly two pounds, with just 30 minutes of talk time and a ten-hour battery recharge time. Ten years later, Motorola introduced the DynaTAC phone, weighing just one pound and costing $3,500.

To support modern wireless communications, cellular networks have evolved over several generations, as follows:

- **1G (analog cellular):** The first analog cellular service was launched in Japan in 1979. In 1983, the Advanced Mobile Phone Service (AMPS) was launched in North America. Analog cellular signals permitted only voice traffic and were not encrypted, so they could be easily intercepted.

  TECHNICAL STUFF

  1G service consumed lots of spectrum and used the frequency division multiple access (FDMA) channel access method. FDMA allocates one or more frequency bands (or channels) to a user for communication.

- **2G (digital cellular):** The second generation of cellular technology was launched in 1991 with the commercial release of the Global Standard for Mobile Communications (GSM) in Finland. Major innovations in 2G networks included:
  - **Digital:** Digital signals generally have less static and background noise, and they use available spectrum more efficiently than do analog signals.
  - **Encryption:** 2G digital calls can be encrypted to make eavesdropping and intercept more difficult.
  - **Data:** Short message service (SMS) text messages were first introduced in 2G networks — O-M-2G!

  TECHNICAL STUFF

  2G technologies use either time division multiple access (TDMA) or code division multiple access (CDMA) channel access methods. TDMA divides a signal into different time slots, enabling multiple callers to share the same frequency channel. CDMA assigns a code to each caller and uses spread-spectrum technology to create a signal with a wider bandwidth.

  In 2000, the European Telecommunications Standards Institute (ETSI) created the General Packet Radio Service (GPRS), which implemented packet-switched domains, in addition to existing circuit-switched domains. GPRS was

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

dubbed “2.5G” and had nothing to do with *Two and a Half Men,* introduced by CBS three years later.

- **3G (data driven):** Apple and Google brought smartphones to the masses with iPhones and Android devices, respectively, in the early 21st century. These powerful devices — and the mobile apps installed on them (including Global Positioning System or GPS, location-based services, and on-demand video) — created an insatiable appetite for faster download speeds. The first 3G networks, introduced in 1998, provided minimum information transfer rates of 200 kilobits per second (Kbps). The International Telecommunication Union (ITU) has never formally defined a standard for 3G data rates, so downlink data speeds vary widely — from 384 Kbps in a moving vehicle for Wideband Code Division Multiple Access (W-CDMA) to 42.2 megabits per second (Mbps) for Evolved High Speed Packet Access (HSPA+), also known as 3.5G, and 168 Mbps for Advanced HSPA+.
- **4G (Long Term Evolution):** Commercially available 4G mobile networks were rolled out in 2008, and 4G LTE followed in 2010. However, unlike 3G, the ITU Radiocommunication Sector (ITU-R) defined minimum 4G standards — but neither “4G” nor “4G LTE” meets those standards! The ITU-R International Mobile Telecommunications Advanced (IMT-Advanced) requirements include (among other things):
  - Packet-switched all-IP core networks
  - Peak data rates of approximately 100 Mbps for high mobility (such as moving vehicles)
  - Peak data rates of approximately 1 gigabit per second (Gbps) for low mobility (such as walking — or your authors sprinting)

  With the introduction of LTE Advanced, true 4G speeds of up to 1 Gbps finally arrived. LTE Advanced Pro is the next evolution of LTE technology, and it establishes the foundation for 5G. LTE Advanced Pro will deliver speeds in excess of 3 Gbps with less than 2 milliseconds (ms) of latency.

REMEMBER

At this point you may be thinking, “Long Term Evolution — no kidding! Will 5G ever get here?” However, the trend has been for each generation of mobile technology innovation to take about a decade — most of us just didn’t pay attention before we got our first smartphones midway through the 3G era.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Focusing on the 5G Vision

The vision for the 5G future is bold: It is much more than just the next iteration of mobile networks. 5G will achieve three main goals:

- **Speed (ultra-high speed radio access):** 5G will provide download speeds of up to 20 Gbps. If you're wondering "Why would anyone ever need that much speed?" first answer this question: When have you ever heard anyone complain that their phone was *too fast*? It's also important to remember that bandwidth is shared by all the users on a cell tower. Today, if a few users are streaming a video at the airport or watching replays of a touchdown in a stadium, chances are that the download is choppy with lots of buffering, and the experience isn't so great. With 5G, you could theoretically download a 40 gigabyte (GB) 4K Ultra-High-Definition (UHD) movie (like *Jaws*) in less than a minute — you're gonna need a bigger data plan!
- **Responsiveness (ultra-low latency):** 5G networks will be used to control autonomous cars and high precision, mission-critical industrial devices in real-time. High reliability and availability at all times is a necessity for these use cases. For this to happen safely, end-to-end latency — the time it takes for data or commands to travel across the network — has to be extremely low. Latency in 5G networks will be five times faster than today's networks — less than 1ms.
- **Scale (massive connectivity):** By 2020, Gartner conservatively forecasts that there will be more than 21 billion connected devices in the Internet of Things (IoT). Some estimates predict more than 50 billion connected IoT devices by 2020. That's anywhere from three to seven connected devices for every person on the planet in 2020 — not including smartphones, tablets, and computers! These devices will have widely varying network requirements — from environmental sensors for agricultural applications installed in remote areas that might send a few bits of data every few days or weeks, to extremely high-precision, low-latency devices in nanobiotechnology, autonomous cars, and mission-critical industrial environments that rely on real-time communication for potentially lifesaving functions. 5G networks will need to handle the massive scale and

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

challenging connectivity requirements for all these diverse devices, sensors, and applications.

Ordering Up 5G in Five Easy Pieces

5G will be a quantum leap beyond today's networks and will require many technological innovations. We serve them up here in five easy pieces (easier than making wheat toast for Jack Nicholson — "hold the butter, the lettuce, the mayonnaise . . . and the chicken") in this section and the chapters that follow (see Figure 1-1):

- **Speeds and feeds.** Speeds of up to 20 Gbps will be achieved using a combination of innovations such as carrier aggregation (CA), massive multiple input multiple output (MIMO), and quadrature amplitude modulation (QAM). You learn about speeds and feeds in Chapter 2.
- **Unlicensed spectrum:** MNOs are increasingly using unlicensed spectrum in the 2.4 and 5 Gigahertz (GHz) frequency bands. 5G networks will need to tap into the vast amount of spectrum available in these unlicensed bands to offload traffic in heavily congested areas and provide connectivity for billions of IoT devices. Advancements in Wi-Fi, LTE in Unlicensed spectrum (LTE-U), License Assisted Access (LAA), and MulteFire, among others, provide better quality and regulated access to unlicensed spectrum. You learn about these advancements in Chapter 3.
- **Internet of Things (IoT):** IoT devices pose a diverse set of requirements and challenges for 5G networks. It's only fair that IoT should likewise pose a diverse set of solutions as well! You learn about a few of these solutions — including NarrowBand IoT (NB-IoT), LTE Category M1 (LTE-M), Long Range (LoRa) and Sigfox — in Chapter 4.
- **Virtualization:** Network functions virtualization (NFV) enables the massive scale and rapid elasticity that MNOs will require in their 5G networks. Virtualization enables a virtual evolved packet core (vEPC), centralized radio access network (C-RAN), mobile edge computing (MEC), and network slicing — all explained in Chapter 5.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

- **New Radio (NR):** Although the other 5G innovations introduced in this section all have strong starting points in LTE Advanced Pro, 5G NR is a true 5G native technology that has yet to be standardized. 5G NR addresses the need for a new radio access technology that will enable access speeds up to 20 Gbps — and you learn about it in Chapter 6!

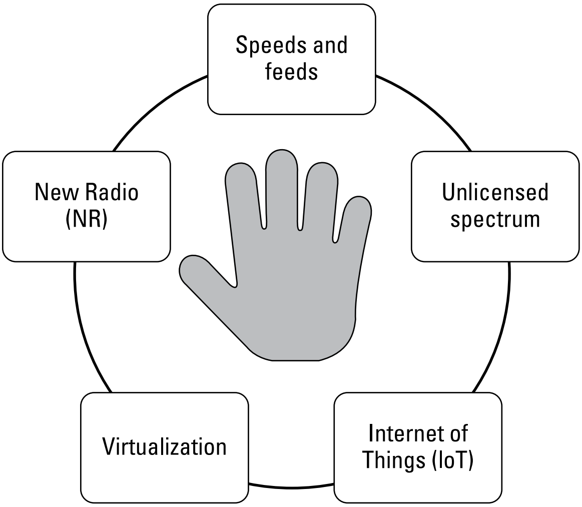
Speeds and feeds
Unlicensed spectrum
Internet of Things (IoT)
Virtualization
New Radio (NR)

FIGURE 1-1: The five keys to 5G.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- **Issuing a license to thrill — with higher speeds**
- **Cobbling together spectrum with carrier aggregation**
- **Going well beyond “rabbit ears” with massive antenna arrays**
- **Seeing stars with quadrature amplitude modulation**

Chapter 2

Achieving Faster Speeds and Larger Feeds

Over the past decade, the need for speed in mobile networks has increased dramatically. To address this need, 5G will increase the speeds of today’s most advanced Long Term Evolution (LTE) networks by an order of magnitude — from a few gigabits per second (Gbps) to as much as 20 Gbps. In this chapter, you get a glimpse of the engine — or, more correctly, the parts and components of the engine — that will power the 5G networks of the future.

Fattening the Data Pipe

Wireless spectrum is limited and highly regulated throughout the world. The International Telecommunication Union (ITU) allocates frequency spectrum worldwide, and governing bodies within the respective countries then license that spectrum for use by individual mobile network operators (MNOs). The ITU has identified three International Telecommunication Regions

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

throughout the world with distinct frequency bands allocated to each region (see Figure 2-1).

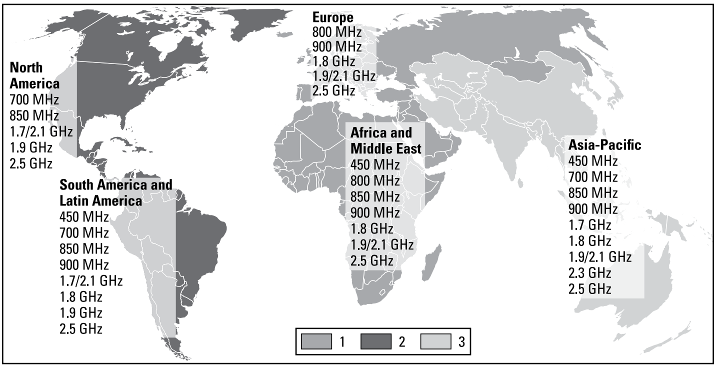
Europe
800 MHz
900 MHz
1.8 GHz
1.9/2.1 GHz
2.5 GHz
North America
700 MHz
850 MHz
1.7/2.1 GHz
1.9 GHz
2.5 GHz
Africa and Middle East
450 MHz
800 MHz
850 MHz
900 MHz
1.8 GHz
1.9/2.1 GHz
2.5 GHz
Asia-Pacific
450 MHz
700 MHz
850 MHz
900 MHz
1.7 GHz
1.8 GHz
1.9/2.1 GHz
2.3 GHz
2.5 GHz
South America and Latin America
450 MHz
700 MHz
850 MHz
900 MHz
1.7/2.1 GHz
1.8 GHz
1.9 GHz
2.5 GHz
1
2
3

FIGURE 2-1: ITU International Telecommunication Regions and spectrum allocations.

MNOs must pay for the right to transmit and receive data using this shared medium in each country in which they operate. To fatten a wireless data pipe, an MNO must procure additional licenses to use a designated frequency spectrum.

In the U.S., the Federal Communications Commission (FCC) auctions spectrum to MNOs — and we aren't talking about eBay auctions here! In the U.S. alone, spectrum auctions have raised more than 60 billion dollars since 1994. Another challenge for MNOs is that available spectrum is limited and going once . . . going twice . . . sold!

Thus, MNOs must find new ways to use existing spectrum more efficiently. LTE Advanced and LTE Advanced Pro are pioneering the path to 5G, using several innovations to get more data through existing spectrum, including:

- Carrier aggregation (CA)
- Multiple input multiple output (MIMO)
- Quadrature amplitude modulation (QAM)

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

We All Bundle — with CA

As you might imagine, the process of buying and selling spectrum, over time, causes spectrum to be sliced and diced in some pretty creative ways. MNOs also come and go, or get merged, acquired, and divested (for example, Cingular Wireless and MCI). All of this causes yet another challenge — contiguous spectrum is hard to find and MNOs must cobble together different bands to maximize their available bandwidth. It's sort of like a business being so unreasonable as to want a continuous range of direct inward dialing (DID) phone numbers to simplify its phone switch programming, company directories, and business cards! Except that in the case of frequency spectrum, non-contiguous bands aren't just a messy inconvenience — they limit available bandwidth.

For LTE networks, including LTE Advanced and LTE Advanced Pro, four carrier bandwidths (or sizes) are available for transporting data:

- 5 megahertz (MHz)
- 10 MHz
- 15 MHz
- 20 MHz

Larger bandwidths can transport more data. For example, 10 MHz can transport data at 37.5 megabits per second (Mbps) and 20 MHz can transport data at 75 Mbps. These data transfer rates assume a single antenna on the user equipment (UE) side and on the Evolved Node B (eNodeB) side. This is known as single input single output (SISO).

TECHNICAL STUFF

User equipment (UE) refers to an end-user device in a mobile network, such as a smartphone. *Evolved Node B (eNodeB)* is the MNO hardware — for example, a base transceiver station (BTS) — that wirelessly communicates directly with the UE.

Carrier aggregation (CA) is a technique that allows an MNO to use more than one component carrier (CC) — known as the *secondary carrier* — as an additional data pipe. For example, 2CA enables any of the four carrier bandwidths (5, 10, 15, or 20 MHz) to be used as the primary carrier and any of the other four carrier bandwidths to be used as the secondary carrier.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

In the simplest deployment scenario, known as *intra-band contiguous CA*, a 20 MHz primary and a 20 MHz secondary carrier (totaling 40 MHz) would provide twice the maximum possible bandwidth and throughput of a single carrier.

For SISO, a 2CA of 20+20 MHz would provide 150 Mbps of data throughput.

TIP

MNOs in the U.S. are currently deploying 3CA and some are already moving to 4CA (see Figure 2-2). The challenge for MNOs now is to find three or more CCs that they own the license to operate in and can aggregate. The LTE Advanced standard specifies 5CA totaling 100 MHz, while LTE Advanced Pro calls for 32CA totaling 640 MHz of aggregated carrier bandwidth.

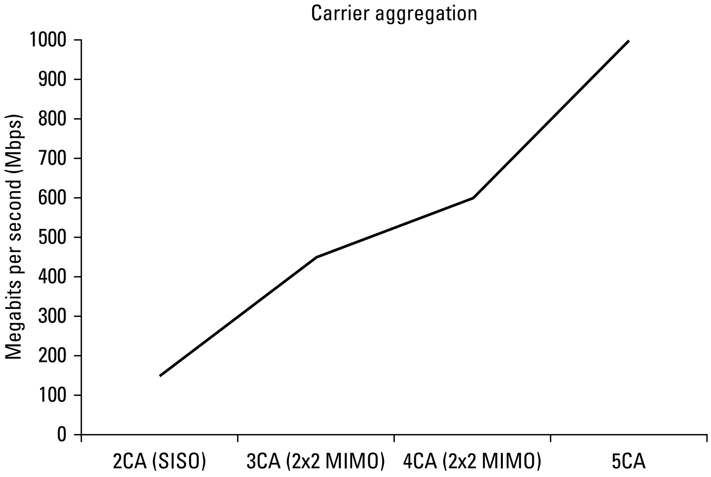

FIGURE 2-2: Carrier aggregation (CA) increases available bandwidth and data throughput.

Eeny, Meeny, Miny, MIMO

One of the techniques for fattening the data pipe defined from day one in 4G LTE networks is multiple input multiple output (MIMO) — using multiple antennas on the transmit and receive side in the wireless network. Using a technique called *spatial multiplexing*, it's possible to send a different data stream on each antenna, thereby increasing the throughput to the cell tower and to the user device.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

While 2x2 MIMO deployments (two antennas in, two antennas out) are common today, 4x4 MIMO (four antennas in, four antennas out) is becoming more practical. 4x4 MIMO effectively enables four independent data streams, thereby increasing the throughput by four times. Thus, the 2CA deployment example (described in the previous section) with two 20 MHz channels and a 2x2 MIMO configuration would increase the data throughput from 150 Mbps to 300 Mbps. The same scenario with a 4x4 MIMO configuration would theoretically increase data throughput to 600 Mbps.

LTE Advanced specifications for MIMO accommodate 8-, 16-, and 32-antenna configurations. LTE Advanced Pro specifications increase MIMO configurations to 64 antenna elements. Realistically, using 64 antenna streams to increase the throughput of a single UE isn't practical. However, using a technique called *beamforming,* the eNodeB can focus energy (that is, steer beams of data) to a particular UE, thereby increasing the throughput of that UE while simultaneously handling other UEs through a different set of beams. Massive MIMO will utilize massive antenna arrays — comprised of hundreds of antennas — to provide efficient signal coverage and higher data rates with lower latency in 5G networks.

No Qualms About QAM

Quadrature amplitude modulation (QAM) is a technique widely used to vary data signals on a carrier used for radio communication. When used for digital transmission of radio communication applications, QAM can carry higher data rates than ordinary amplitude modulated and phase modulated schemes. In QAM, the constellation points are normally arranged in a square grid with equal vertical and horizontal spacing. As a result, the most common forms of QAM use a constellation with the number of points equal to a power of 2 (such as 4, 16, and 64). Thus, 16-QAM uses a 16-point constellation while 256-QAM uses a 256-point constellation.

By using higher order modulation formats (that is, more points on the constellation), it is possible to transmit more bits per symbol. So, 64-QAM has six bits per symbol (more data transmitted), whereas 16-QAM uses only four bits per symbol (less data transmitted). However, the points for a higher QAM are closer together and are therefore more susceptible to noise and data errors. 256-QAM has been used for data in digital cable

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

communications and is now starting to be used for radio communications. 256-QAM is included as part of the LTE Advanced Release 12 standard from the Third Generation Partnership Project (3GPP) because it is likely to work with small cell towers. T-Mobile has achieved 400Mbps downlink speeds in trials through a combination of 4x4 MIMO and 256-QAM.

REMEMBER

5CA and beyond, massive MIMO, and higher-order QAM are all techniques defined as part of the LTE Advanced and LTE Advance Pro standards. With LTE Advanced deployments in full swing, 1 Gbps download speeds are getting closer to becoming a reality in operational networks. 5G will use all these techniques to achieve 20 Gbps data speeds. For example, these are a few of the developments on the 5G horizon:

- CA on very large bandwidths, leading to aggregate totals of 800 MHz (and even 1 GHz) carrier bandwidth
- Massive MIMO antenna arrays of 128 elements or more
- 1024-QAM, and even 4096-QAM

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- » **Expanding cellular traffic to unlicensed spectrum**
- » **Getting LTE-U and LAA to coexist with Wi-Fi in unlicensed spectrum**
- » **Going it alone in unlicensed spectrum with MulteFire**

Chapter 3

Tapping into Unlicensed Spectrum

Unlicensed spectrum is used by low-power devices to transmit and receive wireless signals over short distances — typically a few meters. Although specific devices are permitted to operate only in specific bands, the process of getting certifications is not time-consuming or costly, compared to getting a cellular operator license. Some common devices in this category include garage door openers, nursery monitors, home security systems, cordless phones, and Bluetooth speakers/headsets. In this chapter, you learn about the role of unlicensed spectrum in the 5G future.

Giving a "High 5(G)" to Wi-Fi Advancements

Wi-Fi operates in the unlicensed 2.4 gigahertz (GHz) and 5 GHz spectrums. There are fewer rules on who can access these bands and more available spectrum compared to licensed frequency bands. Wi-Fi devices must therefore compete to use the same spectrum as other devices.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Over the past few years, improvements in speed and the advancement of new capabilities have made Wi-Fi viable (and lucrative) for many mobile network operators (MNOs).

MNOs now regularly use Wi-Fi to offload their cellular networks wherever possible. Newer Wi-Fi standards enable massive speed increases. The standards utilize many of the latest multiple input multiple output (MIMO) beamforming techniques and higher quadrature amplitude modulation (QAM) schemes as those used in the licensed spectrum (discussed in Chapter 2).

For example, 802.11ac delivers Wi-Fi speeds of up to several gigabits per second (Gbps) operating in the 5 GHz band, using 80 or 160 megahertz (MHz) wide channels, advanced beamforming techniques, eight spatial streams (MIMO), multi-user MIMO (MU-MIMO), and 256-QAM (which produces four times the spectral efficiency of the previous 802.11n Wi-Fi standard).

The next wave in Wi-Fi is 802.11ax. 802.11ax will use orthogonal frequency-division multiple access (OFDMA) — the same technique used in Long Term Evolution (LTE) networks — in which different subcarriers within a carrier can be used to transport data for different users. As a result, more than one user equipment (UE) device can access the same medium at a given instant without having to back off or concede the medium to another UE device.

TIP

In case you're wondering, other standards exist between 802.11ac and 802.11ax, but they define Wi-Fi standards in unlicensed spectrum other than 2.4 GHz and 5 GHz — and they don't sound as cool as "A-X"!

LTE in Unlicensed Spectrum

LTE in Unlicensed spectrum (LTE-U) is a proposal that was originally developed by Qualcomm for LTE to co-exist with Wi-Fi in shared unlicensed spectrum. In LTE-U, calls are initially set up using licensed LTE spectrum. Additional carriers (for data) can then be aggregated from the unlicensed spectrum. This method will allow the operators to use unlicensed spectrum to "fatten the data pipe."

The Wi-Fi Alliance developed a Wi-Fi co-existence test plan to ensure that Wi-Fi and LTE could peacefully co-exist, and MNOs (including T-Mobile, AT&T, and Verizon) began experimenting with unlicensed spectrum in addition to their licensed spectrum.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

LTE-U has also generated a lot of interest from a small cell perspective because this approach potentially makes small cells a viable alternative to Wi-Fi hot spots.

As you might imagine, strong reservations about LTE-U exist, because of concerns that it will unfairly use the unlicensed spectrum and potentially interfere with other Wi-Fi users. Cable companies, such as Comcast, Charter Communications, and Time Warner Cable (TWC), as well as Google and Microsoft, are opposed to LTE-U. As of February 2017, the U.S. Federal Communications Commission (FCC) has approved Ericsson and Nokia equipment as LTE-U certified.

License Assisted Access

Unlike LTE-U, License Assisted Access (LAA) is a standard defined in the Release 13 specification from the Third Generation Partnership Project (3GPP). LAA attempts to resolve contention issues with Wi-Fi in unlicensed spectrum using a protocol many of us learned as children — Listen Before Talk (LBT). The LBT contention protocol requires LAA to listen to the unlicensed channel first, and then, if there is no other active Wi-Fi user (the channel is silent), the LTE user can use the channel.

TIP

A recent report by Strategy Analytics states that only 16 percent of operators can achieve gigabit LTE without using unlicensed spectrum. Using LTE-U and LAA, 64 percent of operators can achieve gigabit LTE.

Qualcomm's over-the-air trials with Deutsche Telecom in Germany showed that LAA had increased coverage and capacity compared to Wi-Fi on the same spectrum. By respecting the LBT maxim (uh, protocol), the behavior of LAA is more closely modeled to the principle "love thy neighbor" and is a better neighbor to Wi-Fi than Wi-Fi is to LTE. Wi-Fi was more of an imp during its formative years and ignores LBT, trying to access the same medium at the same time as others.

Enhanced LAA (eLAA), which aggregates the uplink on unlicensed spectrum, is the next step in the evolution of LAA. eLAA will be defined in the 3GPP Release 14 specification.

REMEMBER

Both LAA and eLAA require the initial call to be set up on LTE licensed spectrum as the primary channel. Unlicensed spectrum is only used to "fatten" the data pipes.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

MulteFire

MulteFire goes beyond LTE-U and LAA by enabling the LTE primary channel on unlicensed spectrum. In fact, MulteFire uses unlicensed spectrum exclusively. This means MulteFire can be deployed for LTE by anyone — without owning licensed spectrum — such as Internet service providers (ISPs) and commercial enterprises. MulteFire also benefits MNOs by providing new deployment opportunities for offloading and augmenting their cellular networks.

In January 2017, the MulteFire Alliance released version 1.0, defining an LTE-like network that can run entirely on unlicensed spectrum and, in some cases, may be an alternative to Wi-Fi with more capacity, better security, and easier handoffs across carrier networks.

REMEMBER

Today's 4G spectrum has around 500 MHz of unlicensed spectrum available to deploy LTE-U, LAA, and MulteFire. In July 2016, the FCC opened up 5G spectrum in the U.S., which has 10.85 GHz of total spectrum (3.85 GHz of licensed spectrum and 7 GHz of unlicensed spectrum). The techniques that are now evolving in LTE Advanced Pro (Releases 13 and 14) — LTE-U, LAA (and eLAA), MulteFire, and others (see Figure 3-1) — will be vital as even more unlicensed spectrum becomes available in 5G. The key goal will be to provide users with a seamless experience, irrespective of whether they are operating on a licensed or unlicensed band.

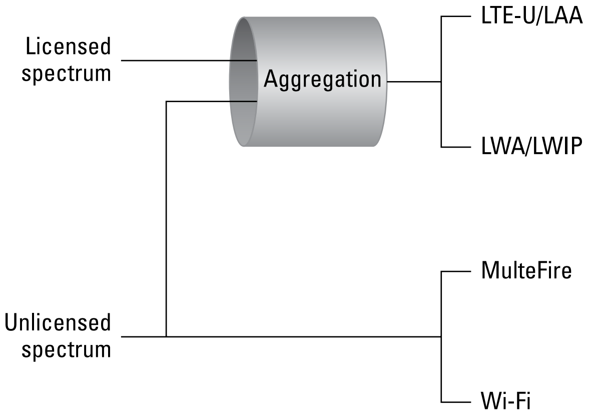

FIGURE 3-1: Leveraging licensed and unlicensed spectrum.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- **Defining IoT connectivity requirements**
- **Handling massive scale with NB-IoT**
- **Slicing up spectrum for the rise of the machines**
- **Exploring LoRaWAN and Sigfox for low-power networks**

Chapter 4

Enabling Massive IoT

When 4G networks were introduced in 2008, there were close to 700 million mobile subscribers worldwide. Today, there are more than 7 billion mobile subscribers worldwide. By 2020, according to the most conservative estimates, there will be approximately 20 billion Internet of Things (IoT) devices, in addition to the 9 billion mobile subscribers that are expected worldwide.

Thus, 5G networks must not only deliver extreme speeds and feeds, they must provide massive scale, predictability, and reliability to eventually support as many as a trillion connected devices, including mission-critical and potentially lifesaving applications and scenarios. In this chapter, you learn about the key requirements for IoT device connectivity and the innovations that are being developed to support IoT.

Key Connectivity Requirements for IoT Devices

Unlike smartphones and other cellular devices, IoT device communications can be sporadic in nature. Many of these devices "sleep" (to extend battery life for ten or more years in some cases)

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

for long periods of time — hours, days, or weeks — before transmitting a few bytes of data, and thus needn't always be connected to the network. 5G networks must be designed to handle infrequent, but important communications from these types of IoT devices. Although the amount of data these devices send may be significantly lower, they may still be of a time-critical nature. For example, a sensor that detects a hazardous condition may instruct an Evolved Node B (eNodeB) element to shut down equipment in an industrial plant or building. These types of communication, though extremely rare, must be handled with the utmost responsiveness and reliability.

Additionally, eNodeB elements will require massive capacity to scale to support tens of thousands of IoT devices — all with different communications requirements and characteristics — in a single cell.

Finally, a wide variety of security threats and new attack vectors will be surfaced in IoT devices. Unlike many of today's threats — identity theft and credit card fraud, for example — many IoT security threats will be potentially life threatening. Future threats may include hacked control systems in autonomous vehicles and smart grids, or compromised medical devices such as insulin pumps and pacemakers. Thus, 5G technologies will need to provide secure end-to-end communications throughout the network.

NarrowBand IoT (NB-IoT)

The Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) Advanced Pro Release 13 specification defines a new standard — NarrowBand IoT (NB-IoT) — for handling low volumes of data (similar to 2G) from tens of thousands of devices in a single cell (tower, not amoeba).

NB-IoT extends LTE to even narrower bandwidths optimized for low data rate, latency-tolerant IoT applications. NB-IoT reduces device complexity, enables multiyear battery life, and provides deeper coverage to reach sensors in challenging locations, such as remote rural areas or inside buildings.

Perhaps one of the most attractive features of NB-IoT is its ability to leverage already ubiquitous LTE networks, in addition to providing high quality of service (QoS) and comprehensive security.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

NB-IoT can be deployed in three different modes:

- **Standalone as a dedicated carrier:** Can use GSM frequencies in a bandwidth of 200 kHz. This does not require LTE.
- **Guard band:** Uses a free resource block within the LTE guard band. This allows the IoT devices to not compete with other LTE devices for the resource blocks within the carrier.
- **LTE in-band:** Uses a resource block within the LTE frequency band. The rest of the blocks are used by the regular LTE devices.

Long Term Evolution for Machines (LTE-M)

Like NB-IoT, Long Term Evolution for Machines (LTE-M or LTE Category M1) leverages a narrow slice of existing LTE spectrum to send and receive data for IoT devices. LTE-M has the same benefits as NB-IoT, but uses a larger network slice than NB-IoT (1.4 MHz compared to 180 KHz in NB-IoT) and leverages the LTE protocol more than the NB-IoT in terms of reusing the same control, data, and transport channels.

TIP

Verizon launched the first LTE-M network in the U.S. on March 31, 2017, and ATT Wireless was expected to follow shortly after.

LoRaWAN and Sigfox

Long Range Wide Area Network (LoRaWAN) is a Low Power Wide Area Network (LPWAN) specification for wireless, battery-operated IoT devices. LoRaWAN operates in the sub-1 GHz unlicensed spectrum band. This limits the volume and frequency of traffic, as well as the ability of the base station to control the network and send traffic down. However, LoRaWAN has great advantages in terms of battery life and cost, and communication is bi-directional.

Sigfox is a French company that created a technology similar to LoRaWAN for IoT device communication. Sigfox technology uses

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

very low bandwidth connections. It is not bi-directional and is only used for sending sparse uplink data with very limited downlink. Like LoRaWAN, Sigfox operates in the sub-1 GHz space and thus uses very low power.

LoRaWAN and Sigfox will be used with certain types of sensors, smart meters, and other low data IoT devices. A disadvantage for these technologies is that, unlike NB-IoT, which is built on top of existing LTE infrastructure, LoRaWAN and Sigfox are not integrated with LTE. However, the cost of deploying a LoRaWAN or Sigfox based IoT device is far less than for an NB-IoT or LTE-M device — by an order of magnitude, since NB-IoT and LTE-M devices must integrate LTE modules into their devices.

REMEMBER

As countries and mobile network operators (MNOs) make IoT technology decisions, it is likely that many standards, protocols, and technologies will need to co-exist. Factors such as cost, coverage, battery life, and integration with older 4G networks will determine which standards, protocols, and technologies are best for each use case.

The 3GPP specifications for IoT technologies are just starting to come out as part of the LTE-A Pro standards. Many IoT devices will need to operate at very low power, ideally suited for sub-1 GHz spectrum rather than the millimeter wave (mmWave) spectrum (discussed in Chapter 6). That said, there are some compelling 5G techniques, like resource spread multiple access (RSMA) waveforms that allow grant-free transmissions for "things" to send their data without prior scheduling by the eNodeB. As a result, the scheduling algorithm becomes less complex. 5G will also enable multi-hop mesh for these low-power devices, allowing out-of-coverage devices to relay to other connected devices in order to send data to the eNodeB.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- » Virtualizing the network for 5G and IoT
- » Knowing which network elements to virtualize

Chapter 5

Getting Real About the Need to Virtualize

Virtualization has been a hot topic in the technology industry for many years and its advantages transcend the cellular industry. In this chapter, you learn about the network elements that can be virtualized and the essential role of virtualization in 5G.

Driving 5G and IoT with Virtualization

Telecommunications companies and mobile network operators (MNOs) have invested heavily in their 4G and Long Term Evolution (LTE) networks. These organizations have embraced virtualization to enable faster, more agile, and scalable deployments that can keep pace with the explosion in subscriber data traffic and consumption, all while keeping control of their overall capital and operating costs. Even though these investments have already given users mobile connectivity of unprecedented speed and pervasiveness, they have only laid the foundation of what's to come, as 5G and smart, connected devices start to roll out.

A recent study by a division of Nokia Bell Labs provides a glimpse of what's coming. The study found that the number of Internet of Things (IoT) connected devices is expected to expand from 1.6 billion in 2014 to well over 20 billion by 2020. Also, by 2020, global consumption demand for digital content and services on portable devices will see an average increase of 30 to 45 times the

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

levels seen in 2014. Thus, MNOs will need to further accelerate their technology investments to meet ever-increasing consumer and business connectivity demands.

Recognizing the IoT explosion

Although MNOs in the U.S. are starting to embrace “pre-standard” 5G, vying for the first mover’s advantage, the standards are not expected until 2018. This has echoes of the early days of 3G in the late 1990s and early 2000s, before proprietary implementations were standardized. It isn’t surprising that some operators in the U.S. have already announced their plans to deploy 5G later this year, and some of the early adopters in Korea, Japan, and China are also ready to roll out 5G. Because of the rapid deployment, many 5G concepts are being solidified quickly.

5G is sometimes narrowly classified as a higher bandwidth radio access technology. But it is much more than that. It will also be the network for low-power devices and sensors that are classified as IoT devices, as well as low-latency applications. One example is the low latency required for some mission-critical devices, such as autonomous vehicles. The need for low latency will dictate that key LTE base station functions are distributed, with some moving closer to the edge and others being pooled in the cloud.

Another example is that vertical sectors will require different types of services from the 5G network. Some will need high bandwidth, while others will need low power. Some will need very low latency, and others will need very high availability. The vast volume of IoT devices will range from those that send multiple gigabits of data per second to those that will only send a few bits every month. This flexibility and elasticity can be supported only by advanced network virtualization.

Focusing on service

To support such varying 5G use cases across multiple verticals, MNOs need to shift from being network-centric to being more service-oriented.

To understand this shift, consider the concept of network slicing (see Figure 5-1). Here are a few typical cases:

- Mobile broadband with higher bandwidth video requiring high availability everywhere

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

- Low-power sensors that can operate on a pair of AA batteries for 10 years
- Autonomous vehicles that can zip through crowded city streets but brake within microseconds when they sense potential obstructions in their paths

TECHNICAL STUFF

Each of these scenarios requires a different configuration of the requirements and parameters in the network, potentially including:

- Home Subscriber Server (HSS)
- Mobility Management Entity (MME)
- User equipment (UE)
- Enhanced Node B (ENB or eNodeB)
- Serving Gateway (SGW)
- Packet Data Network Gateway (PGW)
- Policy and Charging Rules Function (PCRF)
- LTE Evolved Packet Core (EPC) interfaces (S1-MME, S11, S1-U, SGi)

In essence, each use case requires its own network slice. Networks must be built in a manner that allows speed, availability, capacity, and coverage to be allocated in logical slices to meet the demands of each use case.

The best way to implement network slices will be via virtualization — to use service provider software-defined networking (SDN), network functions virtualization (NFV), and network orchestration (see Figure 5-2). In each of these cases, the SDN controller will configure and build network slices that include service chains such as deep packet inspection (DPI), emails, and security scans per user or per service. The network services — including the Centralized Radio Access Network (C-RAN), Virtual Evolved Packet Core (vEPC), and Virtual IP Multimedia Subsystem (vIMS) — are virtualized as virtual network functions (VNFs) that allow MNOs to set up services rapidly, and scale them in response to network and service demands.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

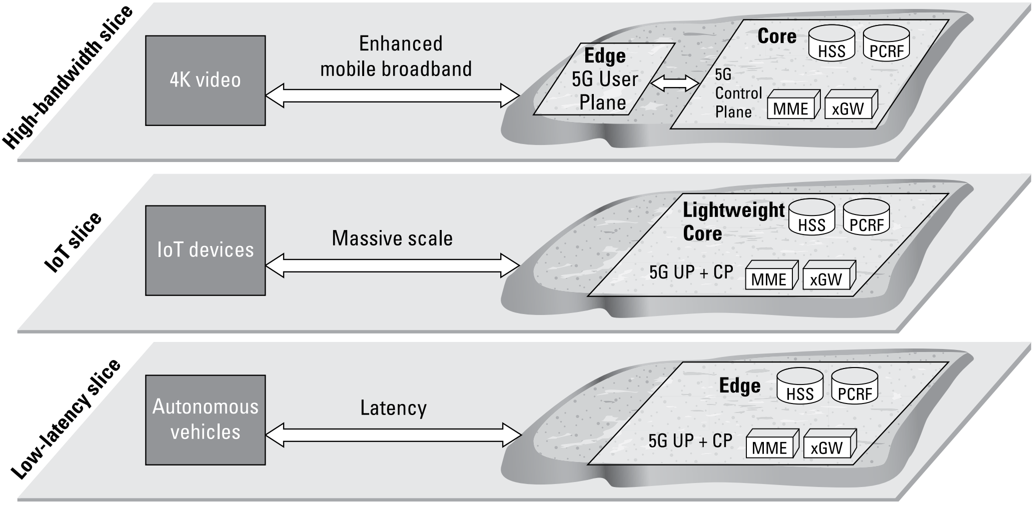
High-bandwidth slice
4K video
Enhanced mobile broadband
Edge
5G User Plane
Core
HSS
PCRF
5G Control Plane
MME
xGW
IoT slice
IoT devices
Massive scale
Lightweight Core
HSS
PCRF
5G UP + CP
MME
xGW
Low-latency slice
Autonomous vehicles
Latency
Edge
HSS
PCRF
5G UP + CP
MME
xGW

FIGURE 5-1: 5G network slices.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

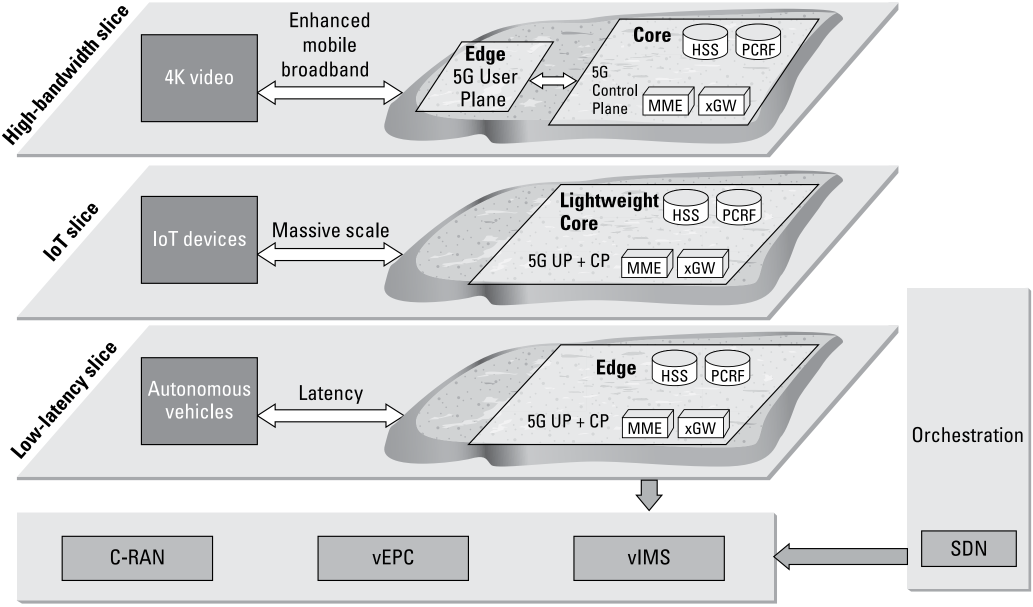
High-bandwidth slice
4K video
Enhanced mobile broadband
Edge
5G User Plane
Core
HSS
PCRF
5G Control Plane
MME
xGW
IoT slice
IoT devices
Massive scale
Lightweight Core
HSS
PCRF
5G UP + CP
MME
xGW
Low-latency slice
Autonomous vehicles
Latency
Edge
HSS
PCRF
5G UP + CP
MME
xGW
Orchestration
SDN
C-RAN
vEPC
vIMS

FIGURE 5-2: Network slices, VNFs, and management and orchestration in a 5G architecture.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Managing the migration

Thus, virtualization will be a vital piece of the puzzle as 5G is rolled out and the IoT grows. But the speed and agility that virtualization enables must be balanced against the need for network visibility, resilience, manageability and security, throughout the development, roll-out, and mass usage of each service.

MNOs and service providers must weigh the trade-offs between

- Quality and cost.
- Flexibility and control.
- Moving too quickly and not moving fast enough.

REMEMBER

As a result, the demand for simple, end-to-end solutions that can efficiently test and validate the effectiveness and benefits of virtualization at every stage is only going to increase. Network virtualization will be a key driver of the IoT ecosystem as it develops — and full-lifecycle reliability and performance testing will be a key element of this shift.

Virtualizing Network Components

When LTE networks were conceptualized, one notable difference from earlier 3G architectures was the idea of a central "brain" — the evolved Node B (eNodeB). From the eNodeB — the point at which the cellular network wirelessly connects directly to user equipment (UE) — to the Internet, the connectivity would all be Internet Protocol (IP) based.

When data centers and other wired network components (such as routers, switches, and firewalls) started being virtualized on standard off-the-shelf servers, it was only natural for the wireless, IP-based 4G evolved packet core (EPC) to follow.

Virtual EPC (vEPC)

Because packet gateways, policy servers, and subscriber databases are all IP-based in the LTE network, the EPC is an excellent

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

candidate for replacing proprietary hardware with virtualized functions hosted on standard, off-the-shelf servers. The benefit for MNOs is the elasticity this model enables — no more dedicated hardware and no need to oversubscribe every part of the network. Servers can be quickly and easily provisioned or de-provisioned based on real-time subscriber capacity needs. With SDN controllers and EPC functions migrated to NFV, and orchestration schemes added to manage it all, the wireless core could be fully virtualized and cloud ready!

Centralized Radio Access Network (C-RAN)

Virtualization in the radio access network (RAN) requires separating out latency-sensitive elements and the radio itself, then pooling the rest of the baseband functions into centralized baseband units (BBUs). Which functions should stay with the remote radio head (RRH) and which ones should move to centralized BBUs has been extensively studied, and a few options exist.

The C-RAN model, which requires a fiber connection between the RRH and the BBUs, is popular in some regions of the world, such as China and Korea. This front-haul interface is challenging because it has distance limitations and is bandwidth limited. It uses a protocol known as the Common Public Radio Interface (CPRI). China Mobile has been in the forefront of the C-RAN architecture. Other countries such as Japan and Korea have also adopted C-RAN. In the U.S., Verizon is looking at implementing C-RAN for small cells.

Mobile Edge Computing (MEC)

Mobile Edge Computing (MEC) offers an important balance as cellular networks move to a central model. MEC allows certain latency-sensitive components to be moved to the network edge. For example, even though it may make business sense for MNOs to centralize several functions into VNFs running on standard off-the-shelf servers, some critical functions may need to be

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

closer to the edge of the network to ensure high availability, low latency, and higher levels of security.

TIP

The network edge is also a good place to host application servers that require local context. For example, if you are driving around and you receive restaurant ads on your mobile device based on your current location, you need the application service to be close to the edge to avoid a lengthy delay in the app that could render the information obsolete.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- **Tuning in to the basic functions of 5G New Radio**
- **Catching the millimeter wave**

Chapter 6

Creating a New 5G World Order with New Radio (NR)

In this chapter, you learn about a completely new technology that creates the foundation for 5G — the 5G New Radio (NR) physical air interface.

5G NR Basics

Unlike the other 5G elements (covered in previous chapters), that advance the industry toward the goals of 5G networks, 5G New Radio (NR) is not an evolution of 4G Long Term Evolution (LTE) innovations. 5G NR is a completely new technology specification — a physical air interface — that is required to achieve the extreme bandwidth, low latency, and massive scalability requirements of 5G.

REMEMBER

The other four areas of 4G LTE innovation that are being further developed for 5G include:

- **Speeds and feeds:** Discussed in Chapter 2, innovations such as carrier aggregation (CA), massive multiple input multiple output (MIMO), and quadrature amplitude modulation (QAM), among others, enables carriers to “fatten” the data pipe.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

- **Unlicensed spectrum:** Discussed in Chapter 3, LTE in unlicensed spectrum (LTE-U), License Assisted Access (LAA and enhanced LAA, or eLAA), and MulteFire enable carriers to leverage unlicensed spectrum as additional data pipes.
- **Internet of Things (IoT):** Discussed in Chapter 4, IoT developments include NarrowBand IoT (NB-IoT), LTE for Machines (LTE-M or LTE Category M1), and Low Power Wide Area Network (LPWAN) specifications such as Long Range WAN (LoRaWAN) and Sigfox. These developments enable wireless communication with a diverse array of billions of IoT devices.
- **Virtualization:** Discussed in Chapter 5, software-defined networking (SDN) and network functions virtualization (NFV) enable mobile network operators (MNOs) to achieve cost-effective scalability and elasticity in their core networks with innovations that include network slicing, virtual evolved packet core (vEPC), centralized radio access network (C-RAN), and mobile edge computing (MEC).

The 5G NR specification enables the following goals of 5G:

- **Extreme bandwidths:** 5G NR aggregates eight component carriers (CCs). The carrier width specifications are still in development, but will likely be 100 megahertz (MHz) or greater, providing a total of approximately one gigahertz (GHz) of aggregated bandwidth — wide enough to carry 20 gigabits of data per second.
- **Low latency:** In LTE Advanced, each subframe (there are 10 subframes) is handled in one millisecond (ms). 5G will be five to ten times faster — each subframe is handled in 100 to 200 microseconds (µs). 5G NR will use new channel coding techniques, such as low density parity check (LDPC), that are more efficient than existing techniques, resulting in shorter transmission time intervals (TTIs).

TECHNICAL STUFF

A technique called *scalable orthogonal frequency-division multiplexing (OFDM) numerology* will be used to support different use cases and scenarios within the same frame. Effectively, shorter TTIs will be used for low latency, high reliability use cases, and longer TTIs for higher spectral efficiency, higher bandwidth use cases. 5G NR can also effectively multiplex between short and long TTIs, thereby allowing a diverse set of users to simultaneously use the system.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

- **Massive MIMO:** 5G NR extends MIMO up to 256 antenna elements and enables massive MIMO. This is a key enabler for higher spectrum bands. The antenna elements in this case are smaller, thereby making it less complex to build a massive array.
- **Massive IoT:** 5G NR will use resource spread multiple access (RSMA) on the uplink to enable grant-free transmission of data on the uplink. A device does not need an enhanced Node B (eNodeB) to give it a grant (or slot) in the pipe to transmit data. This capability eliminates the need for signaling and allows devices to send small packets asynchronously. 5G NR will also address distance and location challenges in low-power IoT devices, using a technique called *multi-hop mesh* to relay uplink data via nearby devices.

5G NR is part of the Release 15 specification from the Third Generation Partnership Project (3GPP). Two versions of 5G NR exist:

- **Non-standalone (NSA) 5G NR:** NSA will use the existing LTE radio and core network as the control plane anchor for mobility management and coverage, while adding a new 5G carrier. Early 2019 deployments of 5G NR will use this configuration. In this mode, the connection is anchored in 4G LTE while 5G NR carriers are used to boost data rates and reduce latency.
- **Standalone (SA) 5G NR:** SA will use the new 5G core network (5GCN) architecture, including the full control and user plane offered by 5G.

An interim release of the 5G NR NSA specification was accelerated to the end of 2017 to enable large-scale trials and deployments for enhanced mobile broadband (eMBB) use cases to begin in 2019. The 5G NR SA specification is expected in mid-2018.

More Spectrum — mmWave Bands

In the U.S., the Federal Communications Commission (FCC) has opened a total of 10.85 GHz of spectrum above 24 GHz to enable 5G use cases. The new spectrum includes 3.85 GHz of licensed spectrum from 27.5 to 28.35 GHz and 37 to 40 GHz, and 7 GHz of unlicensed spectrum from 64 to 71 GHz.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

These bands — between 30 GHz and 300 GHz — are known as millimeter wave bands (mmWave). These high frequency bands translate to narrow wavelengths in the range of one to ten millimeters. There are some challenges with mmWave, including:

- Short transmission paths and high propagation losses over long distances and anything that is not line of sight
- Weakened signals in gases and precipitation

However, the short transmission paths and propagation loss characteristics of mmWave enable spectrum reuse, by limiting the interference between adjacent cells. Additionally, because of the extremely short wavelengths of mmWave signals, transmission paths can be extended using small, multi-element dynamic beamforming antennas that can be installed in user equipment (UE) such as smartphones.

TIP

5G NR, operating at very high frequencies with small coverage areas (because of propagation and line-of-sight issues), is ideal for dense urban locations, where cell sizes are generally small (approximately 200 meters).

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- Bringing 5G home
- Staying entertained on the move
- Getting a dose of virtual and augmented reality
- Driving 5G for smart cars and autonomous vehicles
- Living in a connected world
- Assimilating humans and machines

Chapter 7

Exploring 5G Use Cases

In this chapter, you learn about some of the key use cases that will be enabled by 5G.

Fixed Wireless Broadband Service

5G will begin its commercial drive with the fixed wireless broadband use case in the U.S. AT&T and Verizon are racing to be the first to deliver fixed wireless broadband service to their customers.

Verizon acquired XO Communications, which included the 28 gigahertz (GHz) spectrum, in 2016. One of the primary drivers for Verizon will be to offer 5G as an alternative to cable/advanced digital subscriber line (ADSL) for high-speed broadband access to residential customers for gigabit speed Internet, 4K content, virtual reality (VR), and more. The goal is to achieve greater than one gigabit per second (Gbps) speeds without digging trenches to get fiber to your home.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

AT&T is also pursuing 5G on fixed wireless in both the 28 GHz and 37 GHz spectrum bands. 5G video trials are currently underway for some residential customers, allowing them to stream DIRECTV NOW video service over a fixed wireless 5G connection. In its internal testing, AT&T Wireless claims it can now achieve 13 Gbps over a 5G connection. Like Verizon, AT&T is also looking at 5G to replace home broadband connections delivered by cable companies.

If successful, these use case trials could signal the end of the "cable guy" coming out "sometime between 1 p.m. and 4 p.m." to route a coaxial cable through your home. In areas where there is currently only one option for home broadband or television service, this technology could also lead to more competition, better service, and lower rates for subscribers.

Cable companies, such as Charter Communications, are also exploring 5G technologies. Charter is planning to conduct 28 GHz 5G experiments in the near future with antennas mounted on a mobile trailer and van with hydraulic masts, which will be moved to each of the test locations.

TIP

The motivation for a company like Charter is that 5G will enable a host of new products that take advantage of low-latency, high-capacity networks, including virtual reality (VR) and augmented reality (AR) applications, many of which will be used in a fixed home or office location.

Entertainment Everywhere

The natural progression of the fixed wireless broadband use case (discussed in the previous section) enables mobile broadband at extreme data rates "everywhere" — not only to a fixed home or office location.

Although some of the envisioned mobile broadband experience started in Long Term Evolution (LTE) networks, 5G takes it to a different level with 4K (and subsequently 8K) television, high dynamic range video streaming, 3D videos, and more. All these applications require very high bandwidth, and many require always-on connectivity to push real-time information to the users. The transformation from fixed wireless to wireless everywhere

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

requires several mobility-specific factors to be addressed, such as fast handoffs, seamless connectivity at very high speeds and low latency, high reliability when mobile, and others.

REMEMBER

Cloud storage is also driving the growth of uplink data rates. In the past, content was mostly downloaded and thus required a "fat" pipe in only one direction: the downlink from the base station to the mobile device. However, content is increasingly being uploaded to different cloud storage platforms, requiring robust data pipes in both the downlink and uplink directions. Cloud gaming is another 5G driver that requires fast responsiveness and high broadband capacity.

Virtual Reality (VR) and Augmented Reality (AR)

Virtual reality (VR) technology creates a fully immersive, computer-generated experience that simulates or re-creates real-life situations and environments. In contrast to VR, augmented reality (AR) layers computer-generated images and enhancements onto a real-world situation or environment to provide a more meaningful context for user interaction.

Although current 4G networks are sufficient for some early-adopter VR and AR experiences, the introduction of 5G will enable more novel VR and AR experiences and make them available for mass adoption by consumers. Offering much more capacity, lower latency, and a more uniform experience, 5G will not only improve, but will also be a requirement for some of the most exciting AR and VR use cases, including:

- Sharing live streaming content on social media from event venues along with 50,000 other people in a stadium becomes even more challenging with 4K 360-degree video because each user is uploading 25 Mbps at the same time.
- Next-generation VR and AR experiences will have "six degrees of freedom" (6DoF) — the next level of immersion — allowing users to move within and intuitively interact with the environment. 6DoF content is an order of magnitude richer in naturalness and interactivity than current "three degrees of freedom" (3DoF) video. 3DoF experiences, such as 360-degree

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

video, allow the user to look around rotationally from a fixed position. 6DoF experiences, which are available in video games today, allow the user to move spatially through the environment just by walking or leaning their head forward.

6DoF head-motion tracking is required to enjoy 6DoF content in an intuitive manner. Many industries such as tourism, education, and other forms of immersive video will flourish as 6DoF technologies evolve. Most components of the video delivery pipeline are currently ill-suited for 6DoF video, including capture devices, production software, codecs, compression algorithms, the network, and players. 6DoF video also demands bit rates in the range of 200 Mbps to 1Gbps, depending on the end-to-end latency.

Connected and Autonomous Vehicles

Smart, connected cars are already here — and some security risks have already been notoriously demonstrated — and self-driving autonomous vehicles are beginning to appear on our roads. Other applications that 5G will enable for smart, connected cars include:

- **Traffic safety:** This includes the ability to detect hazardous road conditions, such as inclement weather or nearby accidents, and provide real-time guidance for appropriate courses of action to enable safer driving and reduce the risk of accidents. For example, imagine a situation in which your driverless car gets a real-time message (in microseconds) that a truck is rapidly approaching the intersection that you are about to cross. Your car then automatically slows down to let the truck pass the intersection, thereby avoiding a possible accident.
- **Entertainment (for passengers):** Live video and music streaming (including 4K ultra high-definition movies), interactive video games, cloud connectivity, and data exchange elaborate the need for high capacity and high mobility mobile broadband.
- **Augmented reality (AR):** Displaying key information in near real-time for drivers requires low latency so that the information is timely and relevant. In the earlier truck

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

example, a live situation is being tracked by the involved IoT devices and infrastructure, and relayed in real-time to the users — vehicle and human.

- **Self-driving, autonomous vehicles:** These will require ultra-reliable, high-speed communication between different driverless cars, and between cars and infrastructure.

The Connected World

Smart homes, smart cities, multiple industries (such as health, retail, smart grids, and remote factories) have a common thread — they are using more and more devices and sensors that communicate with one another and the rest of the world.

Many of these devices are mission critical; others may send high-definition video, requiring high availability and very low latency. Yet another set of devices may send small data packets relatively infrequently (for example, every few hours, days, or weeks). Some examples of common use cases with diverse connectivity requirements include:

- **Logistics and freight:** These devices and sensors typically require lower data rates, but need wide coverage and reliable location information.
- **Smart grid:** A smart grid requires low latency sensors to regulate the use of utilities such as electricity, natural gas, and water. Leveraging digital information, such as the behaviors of suppliers and consumers, allows the smart grid to improve the efficiency, reliability, economics, and sustainability of the production and distribution of these resources.
- **Remote medical:** Collaborating about a medical case with other surgeons located thousands of miles away was a use case scenario discussed as part of the Long Term Evolution (LTE) rollout. It can become a reality with the extreme bandwidth, low latency, and high availability of 5G networks.
- **Hazardous areas:** The ability to remotely explore mining areas or shut down a nuclear power plant during an emergency — in a fraction of the time required for human interaction, and without risk to human life — is possible with 5G.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Augmented Humans

Looking ahead to the more distant future — perhaps 30 years from now — Google futurist Ray Kurzweil predicts humans will be able to upload their entire minds to computers and become "digitally immortal" — an event called *singularity.*

In 2011, IBM's Watson beat former winners on the television game show *Jeopardy!*, proving that computers can outperform the best of humans when it comes to synthesizing information and beating them to the buzzer. Google Home and Amazon Echo are becoming more common in homes — perhaps a bit unnerving at first as they start entering our private lives, and increasingly taking center stage when we want them to tell us more about the weather, play our favorite song, read us our favorite *5G For Dummies* book, or tell a few jokes. As machine learning and artificial intelligence develop further, these intelligent devices will better understand human behavior and evolve beyond databases of information. Eventually, they will also become more mobile and intrinsic in our lives, rather than sitting on a shelf in our living rooms.

REMEMBER

5G — with its extreme bandwidths, very low latency, and massive scale support — will be a critical component in creating these real-time experiences, initially as an assistant to humans, and eventually even "thinking" for humans (in some cases). Shopping for clothes with your kids or selecting your next car will become a more immersive experience. 5G technology might also assist humans as an augmented partner, offloading some of the tasks humans don't want to "expend their own brain cells" on. Although it may sound like science fiction, 5G might be a starting point for the next evolution of humans and machines.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

IN THIS CHAPTER

- **Going beyond speed and latency**
- **Looking at drivers and timelines**
- **Overcoming misconceptions about limitations**
- **Recognizing complementary technologies**
- **Transforming entire industries**

Chapter 8

Ten Myths About 5G — Debunked

In this chapter, we expose ten common myths about 5G — and we clue you in to the reality behind them. Keep these points in mind:

- **5G is all about higher speeds to the user.** Although one of the key goals of 5G is to provide extreme bandwidth (high-speed data) to users, low latency and massive scale are other key goals of 5G. So, 5G isn't just about speed!
- **5G requires less than one millisecond latency.** Although less than one millisecond of latency is a goal of 5G, 5G networks will be deployed before that target is actually achieved.
- **Smartphones will lead the charge to 5G.** The iPhone and Android were born in the 3G era and virtually exploded (literally, in some cases) during the 4G LTE era. However, 5G will not only enable faster and better smartphones — it will also lead to mass-market consumer VR and AR devices, sensors and applications for smart homes and cars, industrial robots, and billions of other Internet of Things (IoT) devices yet to be conceived.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

- **5G will be commercially available in 2017.** Although Verizon plans to roll out 5G fixed mobility in 2017 and Olympic 5G trials will be underway during the 2018 Winter Olympics, 5G standards will not be finalized until the end of 2017. Thus, although some early, pre-standard versions of 5G will be coming soon, the 5G standard is not complete. Standards-compliant 5G will be coming much later.
- **5G is only for short-range, line-of-sight communication.** In addition to other frequency bands, 5G uses mmWave bands, which are ideal for very short ranges. However, plenty of ongoing experiments demonstrate how techniques such as beamforming can achieve greater ranges to users in challenging environments beyond line-of-sight.
- **5G will be used only in very high bands.** Although 5G will be deployed in very high millimeter wave (mmWave) bands, it will also re-use spectrum in lower bands, both licensed and unlicensed.
- **5G will replace 4G LTE.** 5G will coexist with 4G LTE for a long time to come. 4G has plenty to offer for many current applications such as voice, data, and even IoT.
- **5G will be a revolution, not an evolution.** Although 5G brings in a new physical layer (in 5G New Radio, or 5G NR), there is plenty of evolution from LTE-A Pro technologies such as carrier aggregation (CA), massive multiple input/multiple output (MIMO), quadrature amplitude modulation (QAM), unlicensed spectrum (LTE in unlicensed spectrum or LTE-U, License Assisted Access or LAA, and MulteFire, among others), IoT, and virtualization.
- **5G will be required to drive IoT.** IoT will initially be driven by LTE-A Pro where NarrowBand IoT (NB-IoT) is specified. In addition, other low-power technologies, such as Long Range Wide Area Network (LoRaWAN) and Sigfox, have been defined for IoT.
- **The 5G winners will be the operators and vendors.** Mobile network operators (MNOs), network equipment manufacturers (NEMs), and smartphone manufacturers were the primary business beneficiaries of 4G LTE. However, 5G will transform many industries, including car manufacturing, agriculture, health and medicine, transportation and logistics, and many more.

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

5 Keys to 5G

1

5G-NR (New Radio)

Next-Gen XAir

2

5G Speeds/Feeds

Ixia Multi-Gig LTE with XAir2

3

Wi-Fi Coexistence

Ixia LTE Unlicensed

4

Massive IoT

Ixia IoT

5

Testing NFV Deployments

IxLoad Wireless VE

Read more about about enabling the 5G revolution at www.ixiacom.com/solutions/5g-wireless-test

These materials are © 2017 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Dream up new opportunities in the 5G future

The next-generation mobile network is on the horizon. 5G, the successor to 4G LTE networks, will enable significantly greater mobile speeds — as much as 20 gigabits per second — with very low latency. 5G will support a massive array of devices, from smart phones to virtual reality devices to small sensors. In this book, you learn about the technological innovations being developed today to enable a 5G future. You also learn about potential use cases that will transform entire industries and create new business models and opportunities.

Inside…

- Where wireless communication is headed
- Why higher speeds are the future
- How the IoT will employ 5G connectivity
- Why the unlicensed spectrum matters
- Why New Radio is a starting point for 5G
- How 5G will enable new business models

ixia

Kalyan Sundhar is Vice President — Mobility, Virtualization, and Applications at Ixia. He has more than 25 years experience in the industry, building cutting-edge products. **Lawrence C. Miller** has worked in information technology for more than 25 years. He has written more than 60 *For Dummies* books.

Go to Dummies.com®
for videos, step-by-step photos,
how-to articles, or to shop!

for dummies®
A Wiley Brand

ISBN: 978-1-119-42415-4
Ixia part number:
915-8156-01
Not for resale

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.

AUGUST 2017

LTE to 5G: Cellular and Broadband Innovation

This white paper was written for 5G Americas by Rysavy Research (http://www.rysavy.com) and utilized a composite of statistical information from multiple resources.

Table of Contents

INTRODUCTION **4**
BROADBAND TRANSFORMATION **7**
EXPLODING DEMAND **11**
  Application Innovation 11
  Internet of Things 12
  Video Streaming 12
  Cloud Computing 13
  5G Data Drivers 13
  Global Mobile Adoption 14
ALMOST AT 5G **18**
  Expanding Use Cases 18
  1G to 5G Evolution 20
  4G LTE Advances 23
  5G Use Cases (ITU and 3GPP) 27
  5G Technical Objectives 29
  5G Concepts and Architectures 30
  5G New Radio (NR) 33
  Expected 5G Performance 35
  Release 15 Non-Standalone and Standalone Options 35
  5G Schedule 36
  5G Core Network 37
  5G Device Availability 37
  mmWave 38
  Network Slicing 40
  Information-Centric Networking 41
  3GPP Releases 42
SUPPORTING TECHNOLOGIES AND ARCHITECTURES **45**
  Multiple Cell Types 45
  Smalls Cells and Heterogeneous Networks 46
  Neutral-Host Small Cells 49
  Unlicensed Spectrum Integration 49
  Internet of Things and Machine-to-Machine 53
  Smart Antennas and MIMO 55
  Virtualization 56
  Multi-Access Edge Computing 58
  Multicast and Broadcast 58
  Remote SIM Provisioning 58
VOLTE, RCS, WEBRTC, AND WI-FI CALLING **60**
  Voice Support and VoLTE 60
  Rich Communications Suite 60
  WebRTC 61
  Wi-Fi Calling 62
PUBLIC SAFETY **63**
  LTE Features for Public Safety 63
  Deployment Approaches 65

Device Considerations for Public Safety 66
EXPANDING CAPACITY 68
SPECTRUM DEVELOPMENTS 72
AWS-3 74
Broadcast Incentive Auction (600 MHz) 74
3550 to 3700 MHz “Small-Cell” Band 75
2.5 GHz Band 76
5G Bands 76
Harmonization 78
Unlicensed Spectrum 80
Spectrum Sharing 81
CONCLUSION 85
APPENDIX: TECHNOLOGY DETAILS 86
3GPP Releases 86
Data Throughput Comparison 88
Latency Comparison 94
Spectral Efficiency 95
Data Consumed by Streaming and Virtual Reality 102
Spectrum Bands 104
5G 107
  Architecture 107
  Architecture Options 108
  LTE-NR Coexistence 110
  Integrated Access and Backhaul 113
  Performance 114
LTE and LTE-Advanced 117
  LTE-Advanced Terminology 118
  OFDMA and Scheduling 119
  LTE Smart Antennas 121
  LTE-Advanced Antenna Technologies 125
  Carrier Aggregation 129
  Coordinated Multi Point (CoMP) 133
  Cellular V2X Communications 135
  User-Plane Congestion Management (UPCON) 136
  Network-Assisted Interference Cancellation and Suppression (NAICS) 136
  Multi-User Superposition Transmission (MUST) 136
  IPv4/IPv6 137
  TDD Harmonization 137
  SMS in LTE 138
  User Equipment Categories 138
  LTE-Advanced Relays 139
  Proximity Services (Device-to-Device) 140
  LTE Throughput 141
  VoLTE and RCS 148
  LTE Ultra-Reliable and Low-Latency Communications 153
  Evolved Packet Core (EPC) 153
Heterogeneous Networks and Small Cells 156
  Enhanced Intercell Interference Coordination 160
  Dual Connectivity 165

Internet of Things and Machine-to-Machine 167
Cloud Radio-Access Network (RAN) and Network Virtualization 168
Unlicensed Spectrum Integration 173
  Release 6 I-WLAN 174
  Release 8 Dual Stack Mobile IPv6 and Proxy Mobile IPv6 174
  Release 11 S2a-based Mobility over GTP 174
  Multipath TCP 175
  ANDSF 175
  Bidirectional Offloading Challenges 176
  Other Integration Technologies (SIPTO, LIPA, IFOM, MAPCON) 178
  Hotspot 2.0 178
Self-Organizing Networks (SON) 180
IP Multimedia Subsystem (IMS) 182
Broadcast/Multicast Services 183
Backhaul 185
UMTS-HSPA 186
  HSDPA 187
  HSUPA 190
Evolution of HSPA (HSPA+) 191
  Advanced Receivers 191
  MIMO 192
  Continuous Packet Connectivity 193
  Higher Order Modulation 193
  Multi-Carrier HSPA 193
  Downlink Multiflow Transmission 194
  HSPA+ Throughput Rates 195
UMTS TDD and TD-SCDMA 199
EDGE/EGPRS 200
ABBREVIATIONS AND ACRONYMS 204
ADDITIONAL INFORMATION 214

Copyright ©2017 Rysavy Research, LLC. All rights reserved. http://www.rysavy.com

Introduction

Mobile computing with wireless communications has already changed how people socialize and how companies do business. Yet, we are still in the nascent stages of the transformation that ubiquitous connectivity is enabling. Early examples of this exciting future include virtual and augmented reality, autonomous driving, smart cities, wearable computers, and connected devices throughout our environment. While no one can predict the full extent of innovation that the new global broadband fabric, aided by complementary innovations such as AI, will unleash, one thing is certain: We are rushing toward an extraordinary time.

The step from 3G to 4G was dramatic, and industry advances occurring now in LTE and 5G will be even greater. Standards bodies are standardizing 5G, a process that will continue through the 2020 timeframe, with ongoing enhancements continuing during the next decade. Some operators are deploying pre-standard 5G in trials this year, and initial standards-based networks could be deployed as early as 2019. 5G will not replace LTE; in most deployments, the two technologies will be tightly integrated and co-exist through at least the late-2020s.

Many of the capabilities that will make 5G so effective are appearing in advanced forms of LTE. With carrier aggregation, for example, operators have not only harnessed the potential of their spectrum holdings to augment capacity and performance, but the technology is also the foundation for entirely new capabilities, such as operating LTE in unlicensed bands.

The computing power of today’s handheld computers rivals that of past mainframe computers, powering intuitive operating systems and millions of applications. Coupled with affordable mobile broadband connectivity, these devices provide such unprecedented utility that billions of people are using them.

With long term growth in smartphone and other mobile device use limited by population, innovators are increasingly turning their attention to the Internet of Things (IoT), which already encompasses a wide array of applications. Enhancements to LTE, followed by 5G IoT capabilities, will connect wearable computers, sensors, and other devices, leading to better health, economic gains, and other advantages. 5G addresses not only IoT deployments on a massive scale but also enables applications previously not possible that depend on ultra-reliable and low-latency communications, sometimes called “mission-critical applications.” Although a far more fragmented market than smartphones, the benefits will be so great that the realization of IoT on a massive scale is inevitable.

Regulatory policies are striving to keep pace, addressing complex issues that include how best to allocate and manage new spectrum, network neutrality, and privacy. Policy decisions will have a major impact on the evolution of mobile broadband.

These are exciting times for both people working in the industry and those who use the technology—which today is nearly everybody. This paper attempts to capture the scope of what the industry is developing, beginning with Table 1, which summarizes some of the most important advances.

Table 1: Most Important Wireless Industry Developments in 2017

Development	Summary
5G Research and Development Accelerates	5G, in early stages of definition through global efforts and many proposed technical approaches, could be deployed in non-standalone versions as early as 2019. Deployment will continue through 2030. Some operators will deploy pre-standard networks for fixed-wireless access in 2017. 5G is being designed to integrate with LTE, and some 5G features may be implemented as LTE-Advanced Pro extensions prior to full 5G availability.
5G New Radio (NR) Being Defined	Key aspects of the 5G NR have been determined, such as radio channel widths and use of OFDMA. The first version, specified in Release 15, will support low-latency, beam-based channels, massive Multiple Input Multiple Output (MIMO) with large numbers of controllable antenna elements, scalable-width subchannels, carrier aggregation, Cloud Radio-Access Network (RAN) capability, and dynamic co-existence with LTE.
LTE Has Become the Global Cellular Standard	A previously fragmented wireless industry has consolidated globally on LTE. LTE is being deployed more quickly than any previous-generation wireless technology.
LTE-Advanced Provides Dramatic Advantages	LTE capabilities continue to improve with carrier aggregation, 1 Gbps peak throughputs, higher-order MIMO, multiple methods for expanding capacity in unlicensed spectrum, new IoT capabilities, vehicle-based communications, small-cell support including Enhanced Inter-Cell Interference Coordination (eICIC), lower latency, Self-Organizing Network (SON) capabilities and Enhanced Coordinated Multi Point (eCoMP).
Internet of Things Poised for Massive Adoption	IoT, evolving from machine-to-machine (M2M) communications, is seeing rapid adoption, with tens of billions of new connected devices expected over the next decade. Drivers include improved LTE support, such as low-cost and low-power modems, enhanced coverage, higher capacity, and service-layer standardization, such as oneM2M.
Unlicensed Spectrum Becomes More Tightly Integrated with Cellular	The industry has developed increasingly sophisticated means for integrating Wi-Fi and cellular networks, such as LTE-WLAN Aggregation (LWA) and LTE-WLAN Aggregation with IPSec Tunnel (LWIP), making the user experience ever more seamless. The industry has also developed and is now deploying versions of LTE that can operate in unlicensed spectrum, such as LTE-Unlicensed (LTE-U), LTE-Licensed Assisted Access (LTE-LAA), and

Development	Summary
	MulteFire. Cellular and Wi-Fi industry members are successfully collaborating to ensure fair spectrum co-existence.
Spectrum Still Precious	Spectrum in general, and in particular licensed spectrum, remains a precious commodity for the industry. Recently added spectrum in the United States includes the 600 MHz band, auctioned in 2017, and the 3.5 GHz Citizens Broadband Radio Service (CBRS) "small-cell" band, which could see initial deployments in 2018. 5G spectrum will include bands above 6 GHz, including "mmWave" (30 GHz to 100 GHz), with the potential of ten times (or more) as much spectrum as is currently available for cellular. Radio channels of 200 MHz and 400 MHz, and even wider in the future, will enable multi-Gbps peak throughput.
Small Cells Take Baby Steps, Preparing to Stride	Operators have begun installing small cells, which now number in the tens of thousands. Eventually, hundreds of thousands if not millions of small cells will lead to massive increases in capacity. The industry is slowly overcoming challenges that include restrictive regulations, site acquisition, self-organization, interference management, power, and backhaul.
Network Function Virtualization (NFV) Emerges	Network function virtualization (NFV) and software-defined networking (SDN) tools and architectures are enabling operators to reduce network costs, simplify deployment of new services, reduce deployment time, and scale their networks. Some operators are also virtualizing the radio-access network, as well as pursuing a related development called cloud radio-access network (cloud RAN). NFV and cloud RAN will be integral components of 5G.

The main part of this paper covers the transformation of broadband, exploding demand for wireless services, the path to 5G including planned and expected capabilities, new LTE innovations, supporting technologies and architectures, voice over LTE (VoLTE), Wi-Fi Calling, LTE for public safety, options to expand capacity, and spectrum developments.

The appendix delves into more technical aspects of the following topics: 3GPP Releases, Data Throughput, latency, 5G, LTE, LTE-Advanced, LTE-Advanced Pro, HetNets and small cells, IoT, cloud RANs, Unlicensed Spectrum Integration, self-organizing networks, the IP Multimedia Subsystem, broadcast/multicast services, backhaul, UMTS/WCDMA,[1] HSPA, HSPA+, UMTS TDD, and EDGE/EGPRS.

---

[1] Although many use the terms "UMTS" and "WCDMA" interchangeably, in this paper "WCDMA" refers to the radio interface technology used within UMTS, and "UMTS" refers to the complete system. HSPA is an enhancement to WCDMA.

Broadband Transformation

Broadband networks rely on a fiber core with various access technologies, such as fiber to the premises, coaxial cable, digital subscriber line (DSL), or wireless connections. LTE provides a broadband experience, but capacity limitations prevent it from being the only broadband connection for most users. As a result, a majority of consumers in developed countries have both mobile broadband and fixed broadband accounts.

Two developments will transform the current situation:

1. **Fiber Densification**. Multiple companies are investing to extend the reach of fiber, decreasing the distance from the fiber network to the end node.
2. **5G Standardization and Deployment.** As 5G mmWave technology, including massive MIMO and beamforming, becomes commoditized, it will increasingly be a viable alternative to fixed-access technologies such as coaxial, DSL, and even fiber connections. 5G commercial services will enable a new innovation cycle. The ability to create new applications and services with fewer limitations will take the connected society to a new level.

Consequently, the companies that provide broadband service may change, and eventually, fixed and mobile broadband services may converge. For a more detailed discussion of trends in broadband, including the disruptive role of mmWave, refer to the 2017 Datacomm Research and Rysavy Research report, *Broadband Disruption: How 5G Will Reshape the Competitive Landscape*.[2]

As shown in Figure 1, the emerging broadband network is one with denser fiber and competing access technologies in which wireless connectivity plays a larger role.

---

[2] Details at https://datacommresearch.com/reports-broadband/.

Figure 1: Fiber Densification with Multiple Access Technologies, including mmWave

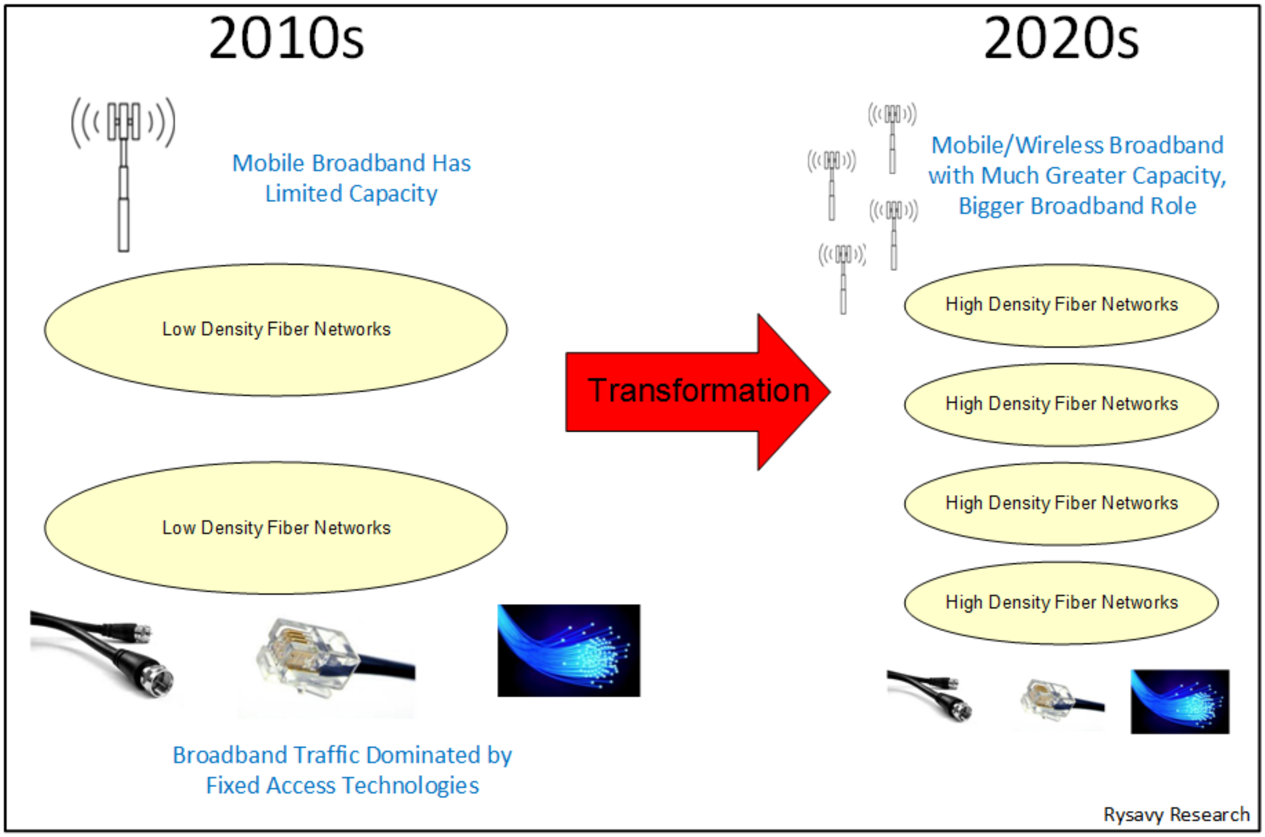
2010s
2020s
Mobile Broadband Has Limited Capacity
Mobile/Wireless Broadband with Much Greater Capacity, Bigger Broadband Role
Low Density Fiber Networks
Low Density Fiber Networks
Transformation
High Density Fiber Networks
High Density Fiber Networks
High Density Fiber Networks
High Density Fiber Networks
Broadband Traffic Dominated by Fixed Access Technologies
Rysavy Research

Many elements are interacting to transform mobile broadband, but the factors playing the most important roles are emerging capabilities for IoT, radio advances granting access to far more spectrum, small cells, new network architectures that leverage network function virtualization and software-defined networking, and new means to employ unlicensed spectrum. Except for access to high-band spectrum, a 5G objective, these advances apply to both LTE and 5G.

Figure 2: Fundamental Mobile Broadband Transformational Elements

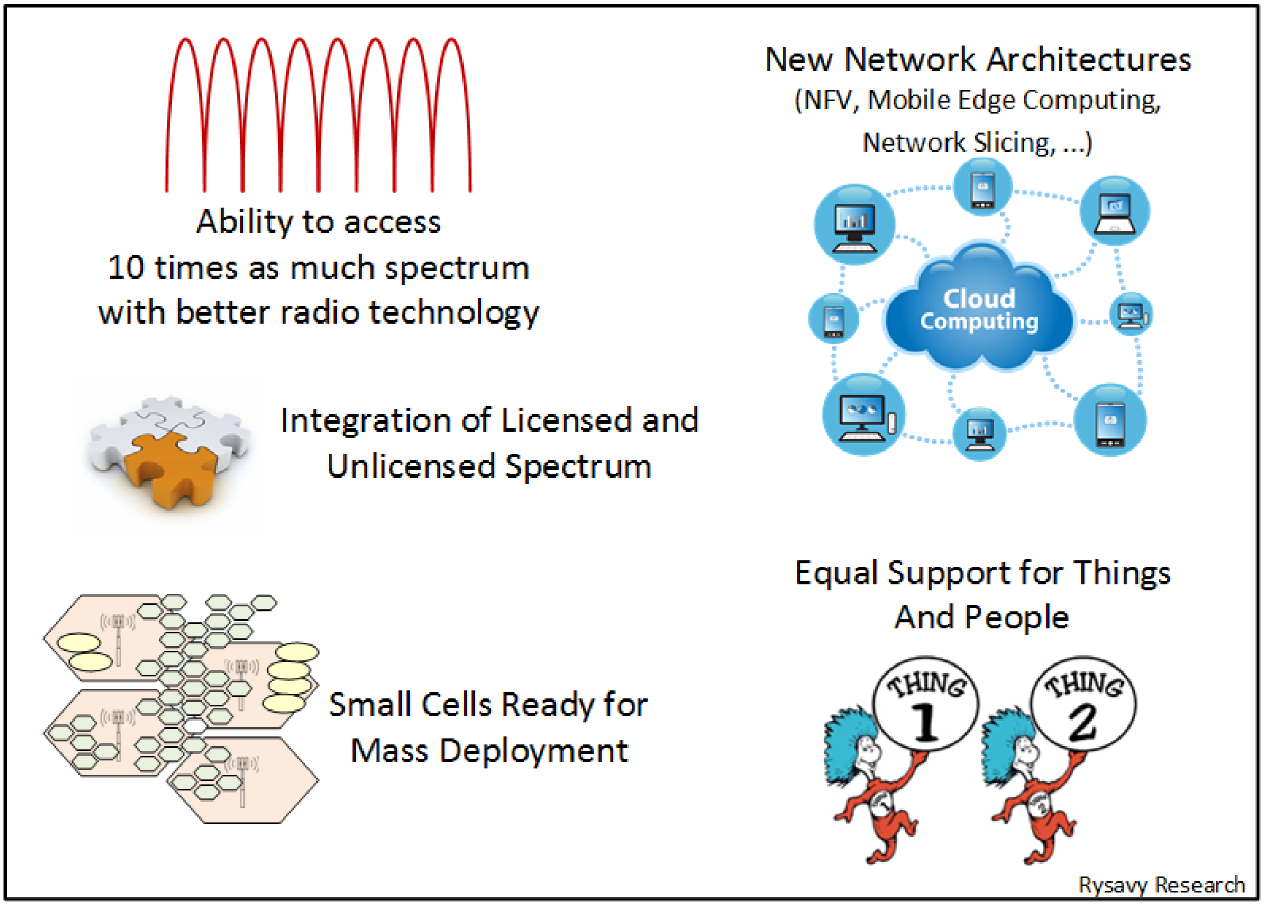
Ability to access
10 times as much spectrum
with better radio technology
New Network Architectures
(NFV, Mobile Edge Computing,
Network Slicing, ...)
Cloud Computing
Integration of Licensed and
Unlicensed Spectrum
Equal Support for Things
And People
THING 1
THING 2
Small Cells Ready for
Mass Deployment
Rysavy Research

In the past, developers used modems and networks designed for human communication for machine-type applications. But now, new modes of network operation, initially in LTE then enhanced further in 5G, will cater to the unique needs of a wide variety of machine applications, addressing low-cost, long battery life, long communications range, and a wide variety of throughputs. For instance, some IoT applications need only low-throughput communications, some sending only a small number of bits per day.

As for spectrum, throughout radio history, technology has climbed up a ladder to use higher frequencies. What were called “ultra-high frequencies” when made available for television are now considered low-band frequencies for cellular. Frequencies above 6 GHz, particularly mmWave frequencies, are the new frontier. Networks will ultimately take advantage of ten times as much spectrum as they use now, and likely even more over time. Although challenging to use because of propagation limitations and higher penetration loss, massive MIMO, beam steering, beam tracking, dual connectivity, carrier aggregation, small-cell architectures with self-backhauling, and other methods will help mitigate the challenges at these frequencies. The result: massive increases in capacity in localized areas.

In addition to accessing higher bands, cellular technologies are integrating unlicensed spectrum more efficiently, using technologies such as LTE-U, LAA, MulteFire, LWA, and LWIP. This integration will immediately augment small-cell capacity, improving the business case for small cells.

Small cells, on the roadmap for many years but held back by implementation difficulties, such as backhaul, are on the verge of large scale deployment, leading ultimately to densities as high as ten small cells or more for every macro cell. Paving the way are better wireless backhaul solutions, neutral-host capabilities enabled by new technologies, and soon, access to mmWave bands.

Facilitating the capabilities listed above, networks are becoming programmable. Using a distributed, software-enabled network based on virtualization and new architectural approaches such as Multi-access Edge Computing (MEC) and network slicing, operators and third parties will be able to deploy new services and applications more rapidly, and in a more scalable fashion. Centralizing RAN signal processing will also play a huge role; depending on the deployment scenario, such centralization will increase RAN efficiency and decrease deployment cost.

This paper lists the dozens of other innovations also fueling mobile and cellular technology transformation. Together, these transformed networks will mean that for millions, and ultimately billions, of people, wireless connections will be the only connections that they need. These networks will also provide the foundation for entire new industries, ones not yet even conceived.

Exploding Demand

Two technology trajectories have created critical mass: handheld computing and fast wireless connections. This combined computing and communications platform inspires the innovation that has produced millions of mobile applications.

IoT is a third trajectory. LTE, and eventually 5G, will connect tens of billions of devices. And fixed-wireless access could be a fourth. 5G, with expected network capacity a hundred times greater than 4G networks, will make wireless connections not only a viable substitute for wireline broadband connections for tens of millions of users over the next ten years, but make that broadband connection the only connection that many users need.

Figure 3: Exploding Demand from Critical Mass of Factors

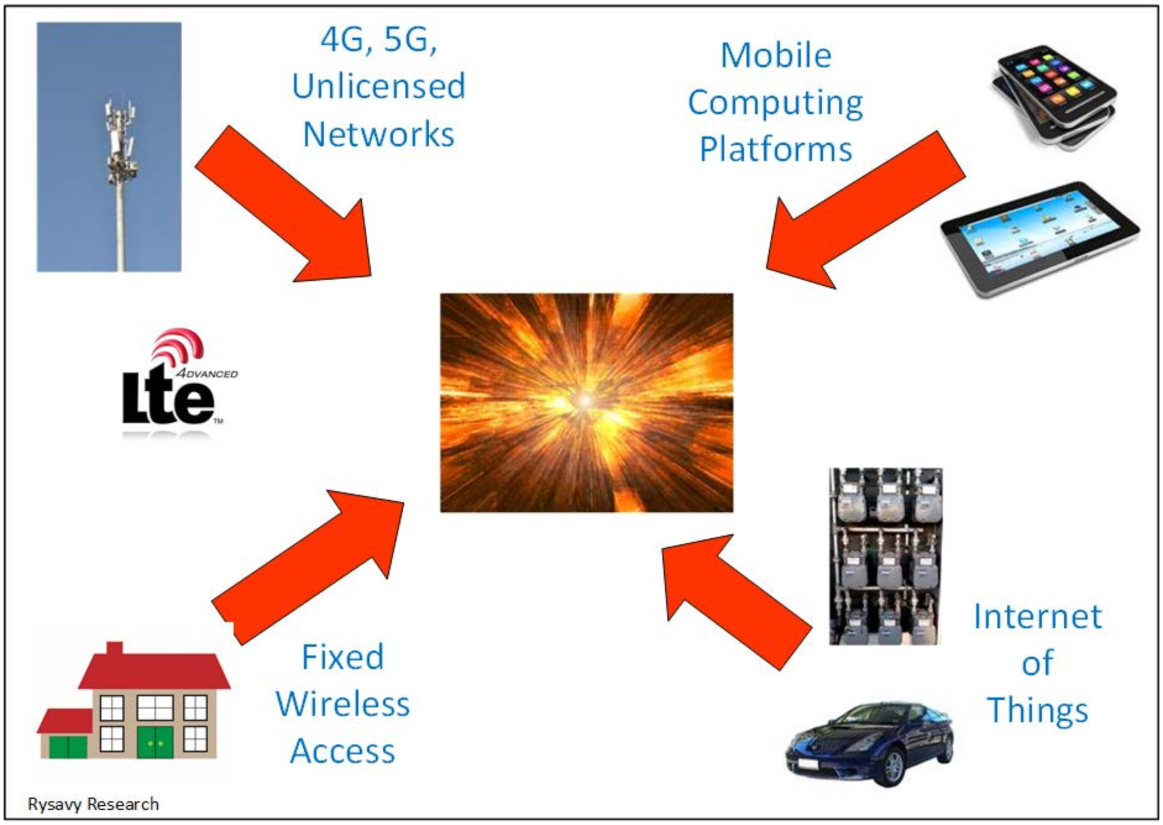

This section analyzes some of these demand factors.

Application Innovation

When planning 4G network technology, who could have predicted applications such as Uber and Lyft, which combine location information with mapping and online payment and are now disrupting the taxi industry and even challenging notions of private vehicle ownership? While some applications of new technology can be predicted, many cannot.

More efficient technology not only addresses escalating demand, it also provides higher performance, thus encouraging new usage models and further increasing demand.

Today's smartphones and tablets, dominated by the iOS and Android ecosystems, in combination with sophisticated cloud-based services, provide a stable, well-defined application environment, allowing developers to target billions of users. Developers have rich platform-specific development tools; web-based tools such as HTML5; application programming interfaces (APIs) for mobile-specific functions, such as WebRTC (Web Real-Time Communications); and cloud-based services for applications and application services, such as notifications, IoT support, and mobile-commerce.

Of concern to some companies in the wireless industry, however, are current network neutrality rules that could possibly hamper innovation. By restricting prioritization, for example, the rules seem to fail to recognize that traffic streams from different applications have inherently different quality of service requirements.[3]

Internet of Things

Current M2M and Internet of Things applications include vehicle infotainment, connected healthcare, transportation and logistics, connected cars, home security and automation, manufacturing, construction and heavy equipment, energy management, video surveillance, environmental monitoring, smart buildings, wearable computing, object tracking, and digital signage. Municipalities, evaluating the concept of "smart cities," are exploring how to optimize pedestrian and vehicular traffic, connect utility meters, and deploy trash containers that can report when they need emptying.

Although promising, the IoT market is also challenging, with varying communications requirements, long installation lifetimes, power demands that challenge current battery technology, cost sensitivity, security and data privacy concerns, unsuitability of conventional networking protocols for some applications, and other factors that developers must address. Consequently, the IoT opportunity is not uniform; it will eventually comprise thousands of markets. Success will occur one sector at a time, with advances in one area providing building blocks for the next.

Cloud-based support platforms and standardized interfaces are essential for development and deployment of IoT applications. For example, oneM2M has developed a service-layer that can be embedded in hardware and software to simplify communications with application servers.[4]

To address the IoT opportunity, 3GPP is defining progressive LTE refinements that will occur over multiple 3GPP releases. These refinements include low-cost modules that approach 2G module pricing and enable multi-year battery life. See the section "Internet of Things and Machine-to-Machine" in the appendix for more details.

Video Streaming

Video represents the greatest usage of data on smartphones. Just an hour a day of mobile video at 1.0 Mbps throughput, common with YouTube or Netflix, consumes 13.5 GB per month. Many streams, now available in high-definition (HD), consume even more data.

---

[3] For further discussion, see Rysavy Research, *How "Title II" Net Neutrality Undermines 5G*, April 19, 2017. Available at http://www.rysavy.com/Articles/2017-04-How-Title-II-Net-Neutrality-Undermines-5G.pdf. Also see *How Wireless is Different – Considerations for the Open Internet Rulemaking*, September 12, 2014. Available at http://www.rysavy.com/Articles/2014-09-Wireless-Open-Internet.pdf.

[4] OneM2M home page: http://onem2m.org/.

See the Appendix section “Data Consumed by Video” for a quantification of data consumed by video for multiple usage scenarios.

An increasing number of video applications adapt their streaming rates based on available bandwidth. By doing so, they can continue to operate even when throughput rates drop. Conversely, they take advantage of higher available bandwidth to present video at higher resolution. Fortunately, application developers are becoming sensitive to bandwidth constraints and are offering options for users to reduce consumption.

Nevertheless, video can consume so much data that cutting the broadband cord at the same time as the television cord is not possible for most consumers. With 5G’s massive increase in capacity, however, an increasing percentage of users will use wireless connections as their only broadband connections, including for all their entertainment needs.

Cloud Computing

Cloud computing inherently increases data consumption because it requires communications for all operations. Examples include data synchronization and backup, cloud-based applications (such as email, word processing, and spreadsheets), automatic photo uploads, and music and video streaming.

5G Data Drivers

Futurists can predict some 5G applications, but many others will arise as industries evolve or come into existence to take advantage of new network capabilities. Some potential applications of 5G include:

- Ultra-high-definition, such as 4K and 8K, and 3D video.
- Augmented and immersive virtual reality. Ultra-high-fidelity virtual reality can consume 50 times the bandwidth of a high-definition video stream.
- Realization of the tactile internet—real-time, immediate sensing and control, enabling a vast array of new applications.
- Automotive, including autonomous vehicles, driver-assistance systems, vehicular internet, infotainment, inter-vehicle information exchange, and vehicle pre-crash sensing and mitigation.
- Monitoring of critical infrastructure, such as transmission lines, using long battery life and low-latency sensors.
- Smart transportation using data from vehicles, road sensors, and cameras to optimize traffic flow.
- Mobile health and telemedicine systems that rely on ready availability of high-resolution and detailed medical records, imaging, and diagnostic video.
- Public safety, including broadband data and mission-critical voice.
- Sports and fitness enhancement through biometric sensing, real-time monitoring, and data analysis.
- Fixed broadband replacement.

Some of these applications are already being addressed by 4G, but 5G's lower costs, higher throughputs, and lower latency will hasten realization of their potential.

Specific industries expected to take advantage of 5G include manufacturing, healthcare, media and entertainment, financial services, public safety, automotive, public transport, and energy utilities.[5]

Global Mobile Adoption

Figure 4 shows the often-cited Cisco projection of global mobile data consumption through 2021, measured in exabytes (billion gigabytes) per month, demonstrating traffic growing at a compound annual rate of 47%.

Figure 4: Global Mobile Data 2016 to 2021[6]

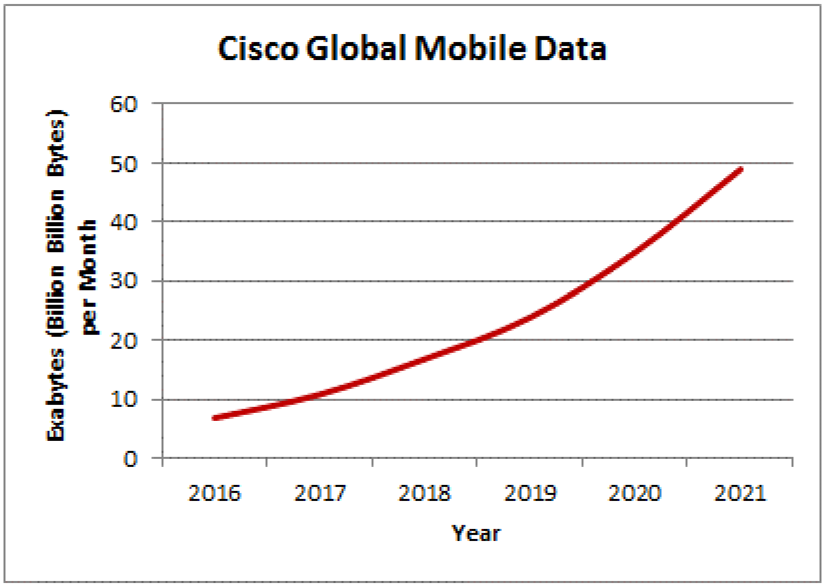

Figure 5 shows another data projection, predicting 45% annual growth in data for the 2016-to-2022 period, resulting in eight-fold growth.

---

[5] For further insight, refer to the Ericsson white paper, *The 5G Business Potential,* February 2017. Available at https://www.ericsson.com/en/networks/insights/the-5g-business-potential.

[6] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021*, February 2017.

Figure 5: Global Mobile Voice and Data (Exabytes/Month) 2012 to 2022[7]

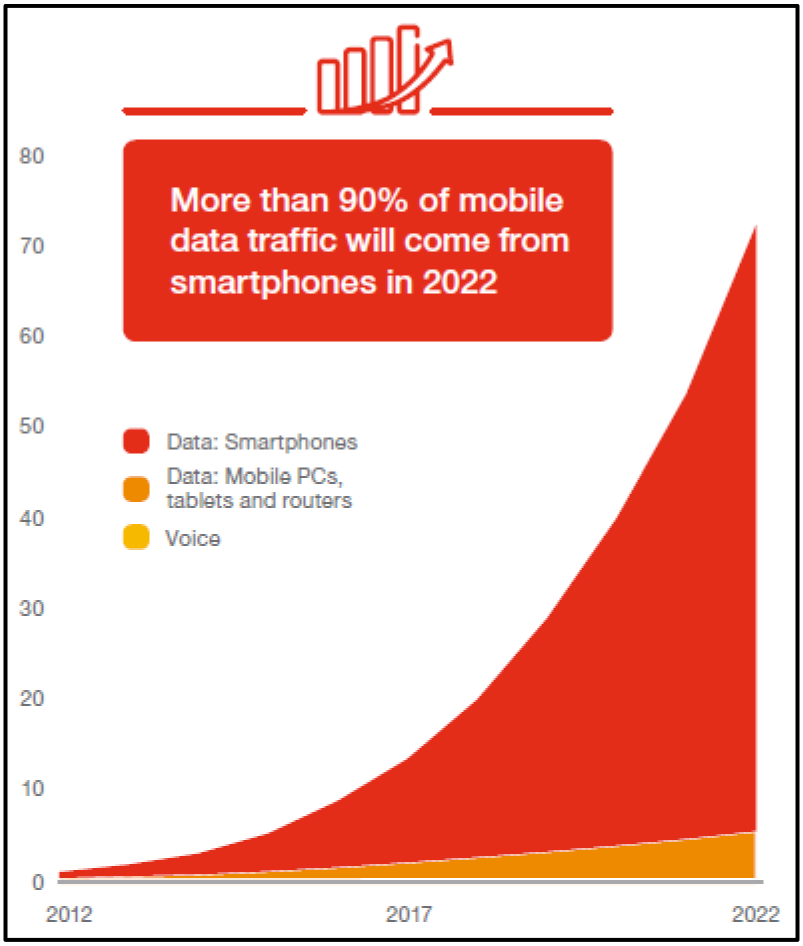

Figure 6 shows how the Internet of Things will drive far more connections than are launched by people.

---

[7] Ericsson, *Ericsson Mobility Report*, June 2017.

Figure 6: Connections of Places Versus People Versus Things[8]

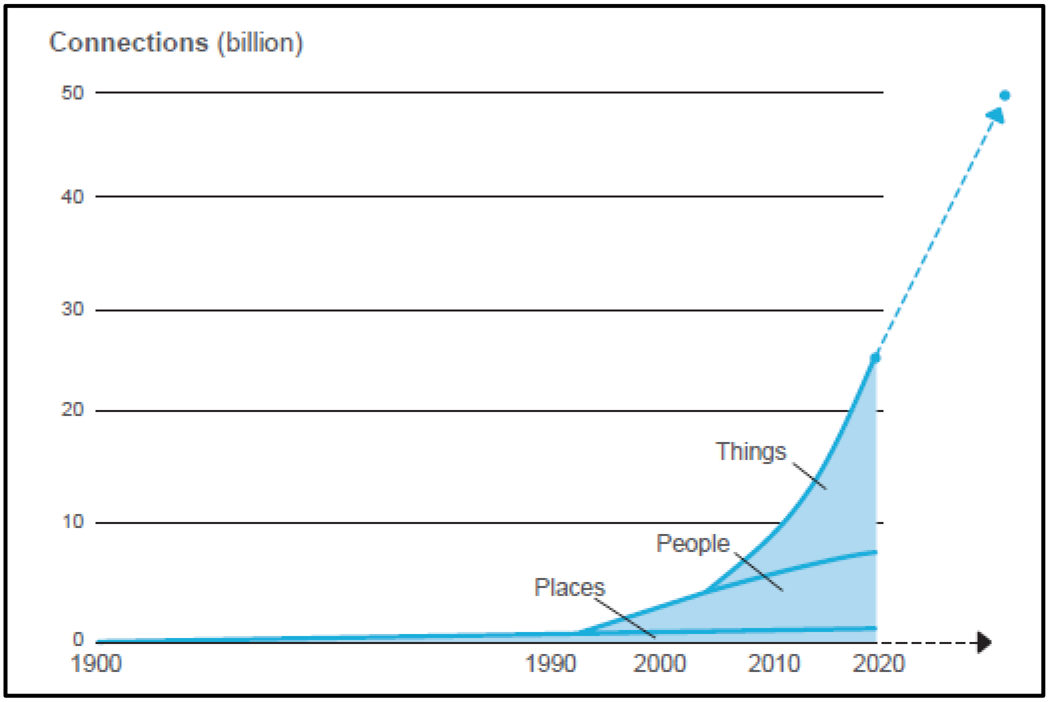

In June 2017, more than 7.56 billion GSM-HSPA-LTE connections were in effect[9]—greater than the world's 7.40 billion population.[10] By the end of 2022, the global mobile broadband market is expected to include nearly 9.4 billion subscribers, with 9.3 billion using 3GPP technologies, representing more than 99% market share.[11]

LTE has experienced faster deployment than any mobile technology ever developed. All major U.S. operators now offer nationwide LTE coverage. LTE has also been chosen by U.S. national public-safety organizations as their broadband technology of choice.

As shown in Figure 7, 2G GSM has peaked and is now declining, as is CDMA. Both HSPA and LTE subscriptions will continue to rise through the rest of the decade.

---

[8] Ericsson, *IoT Security*, February 2017, available at https://www.ericsson.com/assets/local/publications/white-papers/wp-iot-security-february-2017.pdf.

[9] Ovum, July 2017.

[10] U.S. Census Bureau, "U.S. and World Population Clock," http://www.census.gov/popclock/, accessed July 18, 2017.

[11] Ovum, July 2017. Note that the 2018 mobile broadband market figures include GSM/EDGE, because most GSM networks are likely to include Evolved EDGE, which provides mobile broadband capability.

Figure 7: Global Adoption of 2G-4G Technologies 2010 to 2020[12]

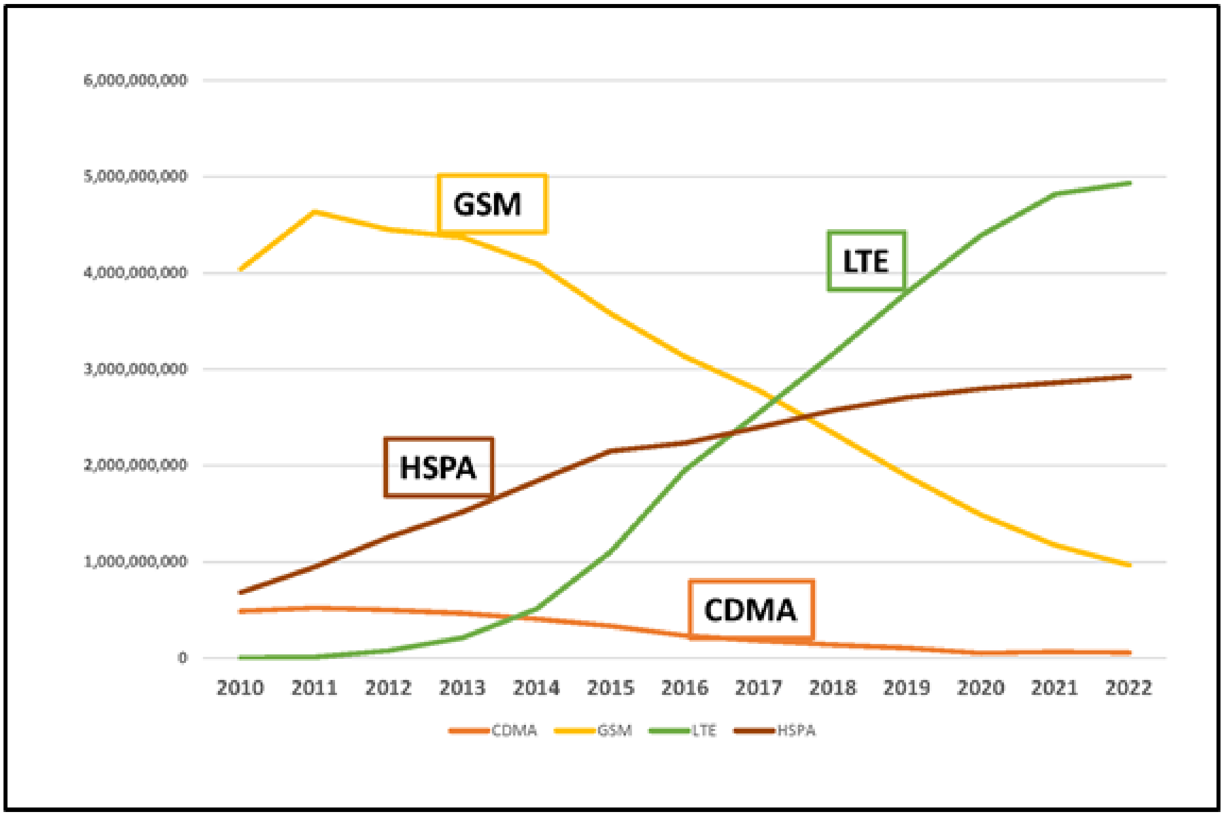

IoT is growing rapidly as well, although connections so far are well below phone-type connections. Cisco projects 3.3 billion IoT connections by 2021, with 6% on 2G cellular, 16% on 3G, and 46% on 4G.[13]

---

[12] Ovum (WCIS), July 2017.

[13] Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021*, February 2017.

Almost at 5G

With intensive development occurring within 3GPP, the first 5G specification will be completed in early 2018, enabling the first standards-based networks to be deployed in the 2019-2020 timeframe. This section includes expanding use cases, 1G-to-5G evolution, 4G LTE advances, 5G radio and architectures, Information-Centric Networking, phasing of 5G releases, and an overview of 3GPP releases.

Expanding Use Cases

Many wireless technology discussions focus on radio capabilities, but other equally important aspects include use cases, the services built on top of the technology, how different networks integrate with one another, and the topology of the networks. As summarized in Figure 8, these aspects are expanding, making mobile/wireless technology the foundation for other enterprises, including business-process optimization, consumer electronics, M2M, connected devices, and a multitude of vertical industries.

Figure 8: Expanding Use Cases

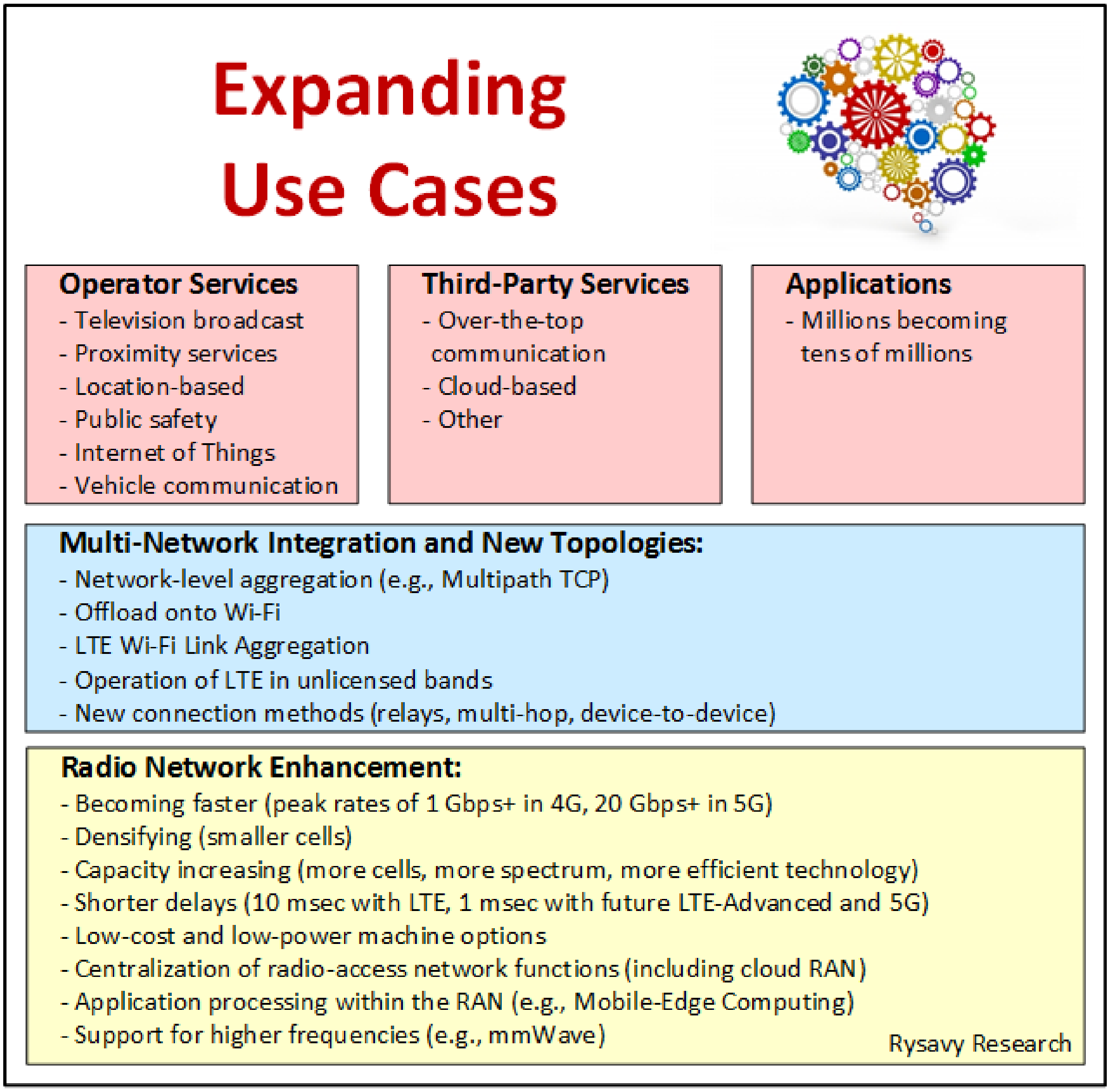

Expanding Use Cases

Operator Services
- Television broadcast
- Proximity services
- Location-based
- Public safety
- Internet of Things
- Vehicle communication

Third-Party Services
- Over-the-top communication
- Cloud-based
- Other

Applications
- Millions becoming tens of millions

Multi-Network Integration and New Topologies:
- Network-level aggregation (e.g., Multipath TCP)
- Offload onto Wi-Fi
- LTE Wi-Fi Link Aggregation
- Operation of LTE in unlicensed bands
- New connection methods (relays, multi-hop, device-to-device)

Radio Network Enhancement:
- Becoming faster (peak rates of 1 Gbps+ in 4G, 20 Gbps+ in 5G)
- Densifying (smaller cells)
- Capacity increasing (more cells, more spectrum, more efficient technology)
- Shorter delays (10 msec with LTE, 1 msec with future LTE-Advanced and 5G)
- Low-cost and low-power machine options
- Centralization of radio-access network functions (including cloud RAN)
- Application processing within the RAN (e.g., Mobile-Edge Computing)
- Support for higher frequencies (e.g., mmWave)

Rysavy Research

Table 2 summarizes the requirements of the expanding number of use cases that employ wireless technology.

Table 2: Requirements for Different Use Cases[14]

Use Cases	Requirements	Desired Value
Autonomous vehicle control	Latency	5 msec
	Availability	99.999 percent
	Reliability	99.999 percent

---

[14] Ericsson, *5G Systems – Enabling the Transformation of Industry and Society*, January 2017. Available at https://www.ericsson.com/assets/local/publications/white-papers/wp-5g-systems.pdf. Adapted from Table 1.

Use Cases	Requirements	Desired Value
Emergency communication	Availability	99.9 percent victim discovery rate
	Energy efficiency	One-week battery life
Factory cell automation	Latency	Down to below 1ms
	Reliability	Down to packet loss of less than 10^{-9}
High-speed train	Traffic density	Downlink (DL): 100Gbps/km^2, uplink (UL): 50 Gbps/km^2
	User throughput	DL: 50Mbps, UL: 25Mbps
	Mobility	500 km/h
	Latency	10ms
Large outdoor event	User throughput	30Mbps
	Traffic density	900Gbps/km^2
	Connection density	Four devices/m^2
Massive IoT	Connection density	1,000,000 devices/km^2
	Availability	99.9 percent coverage
	Energy efficiency	10-year battery life
Remote surgery and examination	Latency	Down to 1ms
	Reliability	99.999 percent
Smart city	User throughput	DL: 300Mbps, UL: 60Mbps
	Traffic density	700 Gbps/km^2
	Connection density	200,000 devices/km^2
Virtual and augmented reality	User throughput	4-28Gbps
	Latency	< 7msec
Broadband to the home	Connection density	4,000 devices/km^2
	Traffic density	60Gbps/km^2

The economic role that wireless technology plays keeps increasing. One study anticipates that in 2035, 5G will enable $12.3 trillion of global economic output.[15]

1G to 5G Evolution

5G is almost upon us; it is a technology that will be constructed from millions of ideas, methods, algorithms, and processes. Just as 4G LTE became available when previous technologies, such as HSPA, could be further improved, 5G enters the stage when the roadmap for LTE has not been exhausted. And just as 2G coexists today with 3G and 4G, 5G will co-exist with previous generations of technology.

For historical context, “1G” refers to analog cellular technologies that became available in the 1980s. “2G” denotes initial digital systems that became available in the 1990s and that introduced services such as short messaging and lower-speed data. 3G requirements were specified by the International Telecommunication Union (ITU) as part of the International Mobile Telephone 2000 (IMT-2000) project, for which significant voice capacity improvement was a focus and digital networks had to provide 144 Kbps of throughput at mobile speeds, 384 Kbps at pedestrian speeds, and 2 Mbps in indoor environments. UMTS-

---

[15] HIS Economics and HIS Technology, *The 5G Economy: How 5G Technology will contribute to the global economy,* January 2017. Commissioned by Qualcomm Technologies.

HSPA and CDMA2000 are the primary 3G technologies. 3G technologies began to be deployed early last decade.

In 2008, the ITU issued requirements for IMT-Advanced, which many people initially used as a definition of 4G. The focus on 4G was to improve data coverage, capacity, and quality of experience. Requirements included operation in up to-40 MHz radio channels and extremely high Spectral Efficiency. The ITU required peak spectral efficiency of 15 bps/Hz and recommended operation in up-to-100 MHz radio channels, resulting in a theoretical throughput rate of 1.5 Gbps. In 2009 and 2010, the term “4G” became associated with mobile broadband technologies deployed at the time, such as HSPA+, WiMAX, and initial LTE deployments. Today, 4G usually refers to HSPA+ or LTE.

Although the industry is preparing for 5G, LTE capabilities will continue to improve in LTE-Advanced Pro through the rest of the decade. Many of these enhancements will come through incremental network investments. Given the scope of global wireless infrastructure, measured in hundreds of billions of dollars, offering users the most affordable service requires operators to leverage investments they have already made. 5G will eventually play an important role, but it must be timed appropriately so that the jump in capability justifies the new investment.

5G groups researching next generation wireless architecture and requirements include, among others, the International Telecommunication Union (ITU),[16] the European Union’s 5G Infrastructure Public-Private-Partnership (5G PPP), which is the framework for several projects, including METIS II (Mobile and wireless communications Enablers for the Twenty-twenty Information Society), and Next Generation Mobile Networks (NGMN). Finally, 5G Americas is actively involved in developing the vision and requirements of 5G for North, Central, and South America. 5G Americas signed an MoU to collaborate with 5G-PPP.[17]

The ITU, the standardization group of the United Nations, has set the following standardization timetable in its IMT-2020 project:[18]

- ❑ 2016-2017: Definition of technical performance requirements, evaluation criteria and methods, and submission templates.
- ❑ 2018-2019: Submission of proposals.
- ❑ 2019: Evaluation of proposed technologies.
- ❑ 2019 and 2020: Publication of IMT-2020 specifications.

---

[16] International Telecommunication Union, “Working Party 5D (WP 5D) - IMT Systems,” http://www.itu.int/ITU-R/index.asp?category=study-groups&rlink=rwp5d&lang=en, accessed May 16, 2017.

[17] 5GPPP, “5G-PPP MoU with 4G Americas,” March 2, 2015. http://5g-ppp.eu/5g-ppp-mou-with-4g-amerericas/

[18] ITU, “ITU towards ‘IMT for 2020 and beyond’,” http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx, viewed May 16, 2017.

Wireless technology has progressed to the extent that significant new capabilities are inevitable, making 5G a possible alternative to wireline broadband for many subscribers.[19]

Table 3 summarizes the generations of wireless technology.

Table 3: 1G to 5G

Generation	Requirements	Comments
1G	No official requirements. Analog technology.	Deployed in the 1980s.
2G	No official requirements. Digital technology.	First digital systems. Deployed in the 1990s. New services such as SMS and low-rate data. Primary technologies include IS-95 CDMA (cdmaOne), IS-136 (D-AMPS), and GSM.
3G	ITU's IMT-2000 required 144 Kbps mobile, 384 Kbps pedestrian, 2 Mbps indoors.	First deployment in 2000. Primary technologies include CDMA2000 1X/EV-DO and UMTS-HSPA. WiMAX.
4G (Initial Technical Designation)	ITU's IMT-Advanced requirements include the ability to operate in up-to-40-MHz radio channels and with very high spectral efficiency.	First deployment in 2010. IEEE 802.16m and LTE-Advanced meet the requirements.
4G (Current Marketing Designation)	Systems that significantly exceed the performance of initial 3G networks. No quantitative requirements.	Today's HSPA+, LTE, and WiMAX networks meet this requirement.
5G	ITU IMT-2020 has defined technical requirements for 5G, and 3GPP is developing specifications.	First standards-based deployments in 2019 and 2020.

The interval between each significant technology platform has been about ten years. Within each platform, however, innovation surges unabated. For example, with 2G technology,

---

[19] Rysavy Research, "How will 5G compare to fiber, cable or DSL?" Fierce Wireless, May 2014. Available at http://www.rysavy.com/Articles/2014-05-5G-Comparison-Wireline.pdf.

EDGE significantly improved data performance compared with initial General Packet Radio Service (GPRS) capabilities. Similarly, HSPA hugely increased data speeds compared with initial 3G capabilities. LTE and LTE-Advanced are also acquiring continual improvements that include faster speeds, greater efficiency, and the ability to aggregate spectrum more flexibly.

Figure 9 presents the timeline of technology generations, including past and future, showing initial deployment, the year of the peak number of subscribers, and decline. Each cellular generation spans multiple decades, with peak adoption occurring some 20 years after initial deployment. 6G deployment in 2030, though highly speculative, is consistent with deployment of previous generations.

Figure 9: Timeline of Cellular Generations

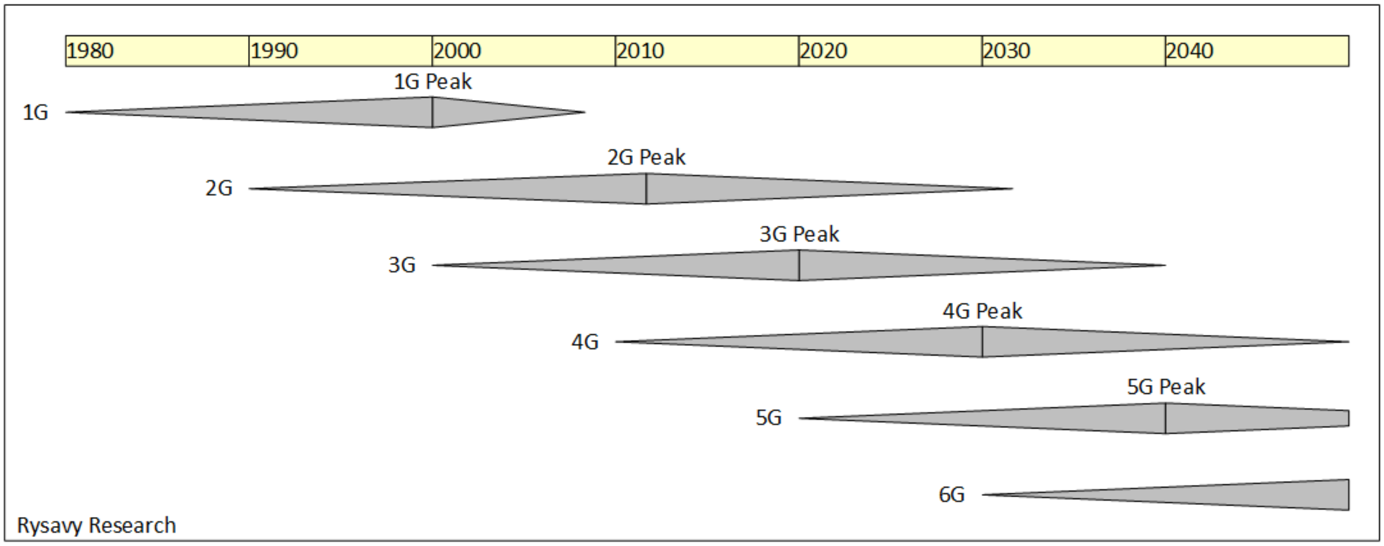

4G LTE Advances

As competitive pressures in the mobile broadband market intensified, and as demand for capacity persistently grew, LTE became the favored 4G solution because of its high data throughputs, low-latency, and high spectral efficiency. Specifically:

- **Wider Radio Channels.** LTE can be deployed in wide radio channels (for example, 10 MHz or 20 MHz) with carrier aggregation now up to 640 MHz, although inter-band aggregation of four or five carriers (up to 100 MHz) represents a practical upper limit.
- **Easiest MIMO Deployment.** By using new radios and antennas, LTE facilitates MIMO Deployment, in contrast to the logistical challenges of adding antennas for MIMO to existing legacy technologies. Furthermore, MIMO gains are maximized because all user equipment supports it from the beginning.
- **Best Latency Performance.** For some applications, low latency (packet traversal delay) is as important as high throughput. With a low transmission time interval (TTI) of 1 millisecond (msec) and a flat architecture (fewer nodes in the core network), LTE has the lowest latency of any cellular technology.

In the same way that 3G coexists with 2G systems in integrated networks, LTE systems co-exist with both 3G and 2G systems, with devices capable of 2G, 3G, and 4G modes.

Beyond radio technology, the Evolved Packet Core (EPC) provides a core architecture that integrates with both legacy GSM-HSPA networks and other wireless technologies, such as CDMA2000 and Wi-Fi. The combination of EPC and LTE is referred to as the Evolved Packet System (EPS).

LTE is available in both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes. Many deployments are based on FDD in paired spectrum. The TDD mode, however, is important for deployments in which paired spectrum is unavailable. Instances of TDD deployment include China, Europe at 2.6 GHz, U.S. Broadband Radio Service (BRS) spectrum at 2.6 GHz, and the 3.5 GHz band.

The versions of LTE most widely deployed today are just the first in a series of innovations that will increase performance, efficiency, and capabilities. Enhancements in the 2013 to 2016 period were the ones defined in 3GPP Releases 10, 11, and 12 and are commonly referred to as LTE-Advanced.[20] Subsequent releases, including Releases 13 to 15, specify LTE-Advanced Pro.

Keeping in mind that different operators have varying priorities, the following list roughly ranks the most important features of LTE-Advanced and LTE-Advanced Pro for the 2017 to 2020 timeframe:

1. **Carrier Aggregation.** With this capability, already in use, operators can aggregate radio carriers in the same band or across disparate bands to improve throughputs (under light network load), capacity, and efficiency. Carrier aggregation can also combine FDD and TDD and is the basis of LTE-U and LTE-LAA. As examples, in 2015, AT&T aggregated 700 MHz with AWS, and 700 MHz with PCS. T-Mobile aggregated 700 MHz with AWS, and AWS with PCS.[21] Operators are now deploying three-carrier aggregation and eventually may aggregate four carriers.[22] Release 13 introduced support for carrier aggregation of up to 32 carriers, addressing primarily the opportunity to aggregate multiple unlicensed channels. Release 14 specifies interband carrier aggregation for up to five downlink carriers and 2 uplink carriers.
2. **VoLTE.** Initially launched in 2015 and with widespread availability in 2017, VoLTE enables operators to roll out packetized voice for LTE networks, resulting in greater voice capacity and higher voice quality.
3. **Tighter Integration of LTE with Unlicensed Bands.** LTE-U became available for testing in 2016, and 3GPP completed specifications for LAA in Release 13, with deployment expected around 2018. MulteFire, building on LAA, will operate without requiring a licensed carrier anchor. LTE/Wi-Fi Aggregation through LWA and LWIP are other options for operators with large Wi-Fi deployments.

---

[20] From a strict standards-development point of view, the term "LTE-Advanced" refers to the following features: carrier aggregation, 8X8 downlink MIMO, and 4XN uplink MIMO with N the number of receive antennas in the base station.

[21] AT&T band combinations are 3GPP Band 13 + Band 4, Band 17 + Band 4, and Band 17 + Band 2. T-Mobile band combinations are Band 12 + Band 4, Band 12 + Band 2, and Band 4 + Band 2.

[22] For carrier aggregation to operate, both the network and the device have to support the particular band combination. Legacy devices typically do not support new network aggregation capabilities.

4. **Enhanced Support for IoT.** Release 13 brings Category M1, a low-cost device option, along with Narrowband-IoT (NB-IoT), a version of the LTE radio interface specifically for IoT devices, called Category NB1.
5. **Higher-Order and Full-Dimension MIMO.** Deployments in 2017 use up to 4X4 MIMO. Release 14 specifies a capability called Full-Dimension MIMO, which supports configurations with as many as 32 antennas at the base station. See the section "Smart Antennas and MIMO" and Appendix section "LTE Smart Antennas" for further detail.
6. **Dual Connectivity.** Release 12 introduced the capability to combine carriers from different sectors and/or base stations (i.e. evolved Node Bs [eNBs]) through a feature called Dual Connectivity. Two architectures were defined: one that supports Packet Data Convergence Protocol (PDCP) aggregation between the different eNBs and one that supports separate S1 connections on the user-plane from the different eNBs to the EPC.
7. **256 QAM Downlink and 64 QAM Uplink.** Defined in Release 12 and already deployed in some networks, higher-order modulation increases user throughput rates in favorable radio conditions.
8. **1 Gbps Capability.** Using a combination of 256 QAM modulation, 4X4 MIMO, and aggregation of three carriers, operator networks can now reach 1 Gbps peak speeds. See below for more information.
9. **V2X Communications.** Release 14 specifies vehicle-to-vehicle and vehicle-to-infrastructure communications. See the section "Cellular V2X Communications" for more information.
10. **Coordinated Multi Point.** CoMP (and enhanced CoMP [eCoMP]) is a process by which multiple base stations or cell sectors process a User Equipment (UE) signal simultaneously, or coordinate the transmissions to a UE, improving cell-edge performance and network efficiency. Initial usage will be on the uplink because no user device changes are required. Some networks had implemented this feature in 2017.
11. **HetNet Support.** HetNets integrate macro cells and small cells. A key feature is enhanced inter-cell interference coordination (eICIC), which improves the ability of a macro and a small-cell to use the same spectrum. This approach is valuable when the operator cannot dedicate spectrum to small cells. Operators are currently evaluating eICIC, and at least one operator has deployed it.[23] Further enhanced ICIC (feICIC) introduced in Rel-11 added advanced interference cancellation receivers into devices.
12. **Ultra-Reliable and Low-Latency Communications.** Being specified in Release 15, URLLC in LTE will shorten radio latency to a 1 msec range using a combination of shorter transmission time intervals and faster hybrid automatic repeat request (HARQ) error processing. See the Appendix section "LTE Ultra-Reliable and Low-Latency Communications" for further details.
13. **Self-Organizing Networks.** With SON, networks can automatically configure and optimize themselves, a capability that will be particularly important as small cells begin

---

[23] Fierce Wireless, "SK Telecom teams with Nokia Networks on eICIC," January 2015.

to proliferate. Vendor-specific methods are common for 3G networks, and trials are now occurring for 4G LTE standards-based approaches.

Other key features that will become available in the 2016-to-2020 timeframe include full-dimension MIMO, enhanced Multimedia Broadcast/Multicast Services (eMBMS), User-Plane Congestion Management (UPCON), and device-to-device communication (targeted initially at public-safety applications).

The appendix explains these features and quantifies performance gains, and Figure 10 illustrates the transition from LTE to LTE-Advanced and LTE-Advanced Pro, which include these features.

Figure 10: LTE to LTE-Advanced Pro Migration[24]

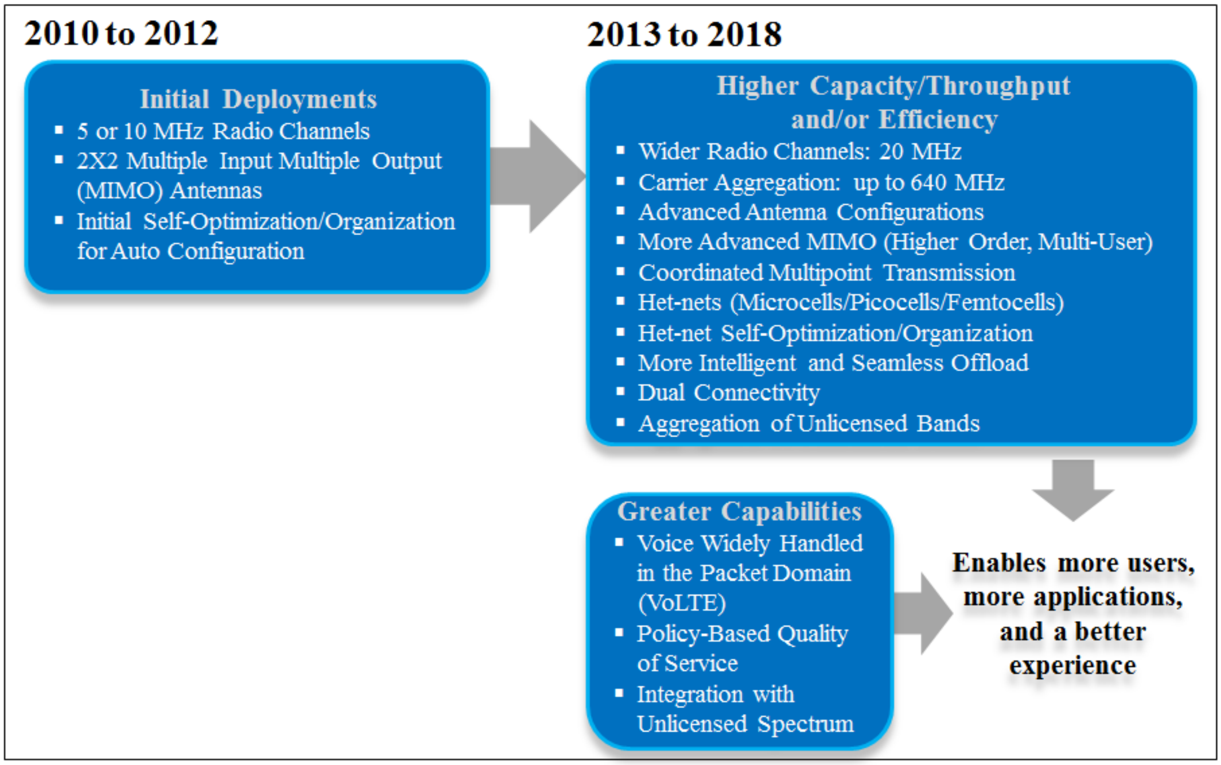

2010 to 2012

Initial Deployments

- 5 or 10 MHz Radio Channels
- 2X2 Multiple Input Multiple Output (MIMO) Antennas
- Initial Self-Optimization/Organization for Auto Configuration

2013 to 2018

Higher Capacity/Throughput and/or Efficiency

- Wider Radio Channels: 20 MHz
- Carrier Aggregation: up to 640 MHz
- Advanced Antenna Configurations
- More Advanced MIMO (Higher Order, Multi-User)
- Coordinated Multipoint Transmission
- Het-nets (Microcells/Picocells/Femtocells)
- Het-net Self-Optimization/Organization
- More Intelligent and Seamless Offload
- Dual Connectivity
- Aggregation of Unlicensed Bands

Greater Capabilities

- Voice Widely Handled in the Packet Domain (VoLTE)
- Policy-Based Quality of Service
- Integration with Unlicensed Spectrum

Enables more users, more applications, and a better experience

Figure 11 illustrates gains in peak downlink speeds through carrier aggregation, higher-order MIMO, and higher-order modulation.

---

[24] 5G Americas/Rysavy Research

Figure 11: Successive Gains in Peak LTE Downlink Throughput[25]

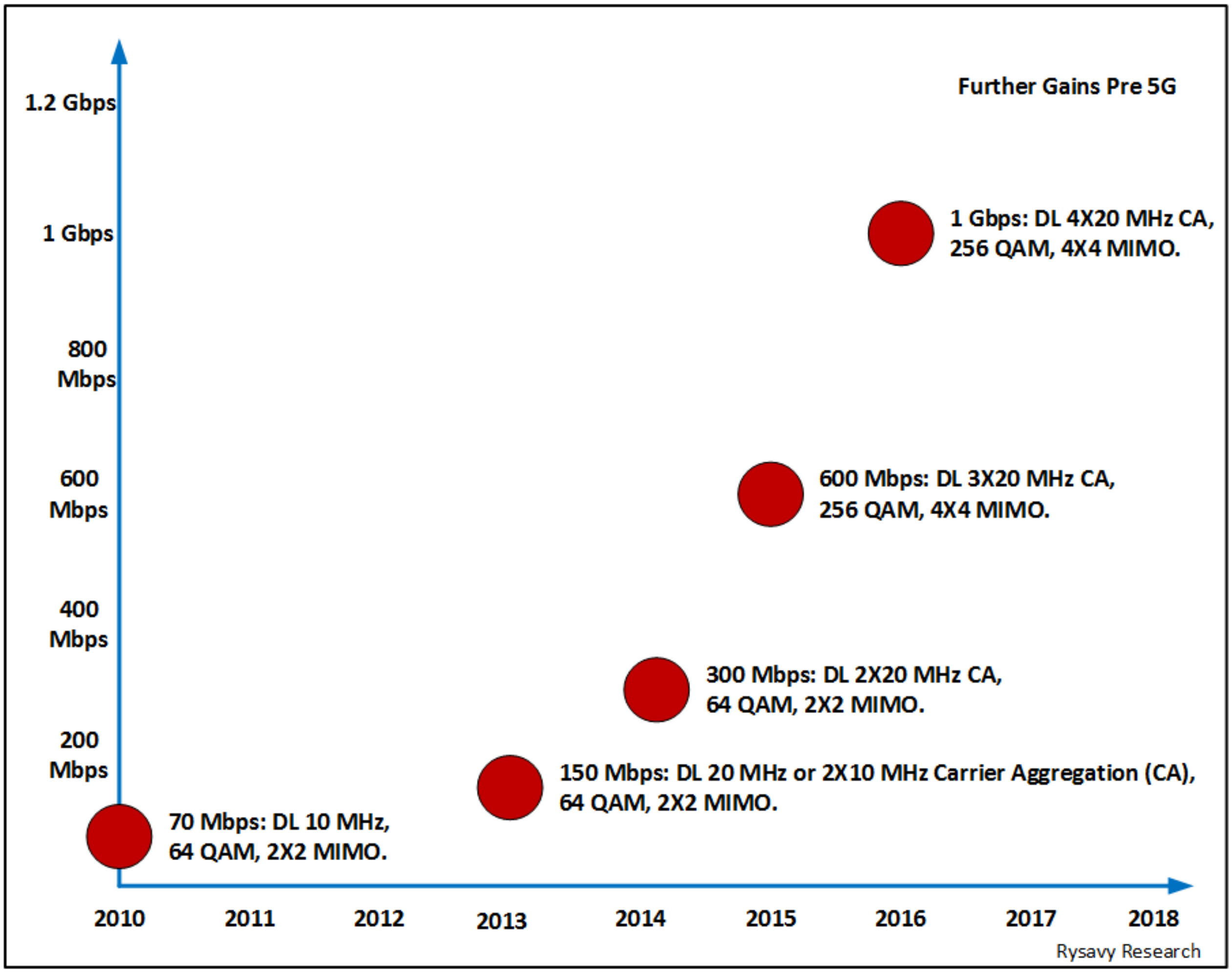

5G Use Cases (ITU and 3GPP)

The ITU, in its 5G recommendations, divides use cases into three main categories, as shown in Figure 12.

1. **Enhanced Mobile Broadband (eMBB).** eMBB is the most obvious extension of LTE capability, providing higher speeds for applications such as streaming, Web access, video conferencing, and virtual reality. Highest speeds will occur in small cells with limited movement speed of end users, such as with pedestrians.
2. **Massive Machine-Type Communications (mMTC).** Massive machine-type communications extends LTE Internet of Things capabilities—for example, NB-IoT—

---

[25] Rysavy Research analysis based on information about commercially available chipsets. Not all LTE features are available simultaneously, and thus throughputs do not necessarily scale linearly with items such as bandwidth. See also the following article about achieving gigabit speeds with LTE: Ars Technica, “Qualcomm’s new LTE modem will make gigabit download speeds easier to hit,” February 21, 2017. Available at https://arstechnica.com/gadgets/2017/02/qualcomms-new-lte-modem-will-make-gigabit-download-speeds-easier-to-hit/#p3.

to support huge numbers of devices with lower costs, enhanced coverage, and long battery life. As shown in the ITU objectives, below, 5G will support ten times as many devices in an area as LTE.

3. **Ultra-Reliable and Low Latency Communications (URLLC).** Of the three categories, URLLC enables wireless applications never before possible. Driven by high dependability and extremely short network traversal time, URLLC, also referred to as "mission-critical" communications, will enable industrial automation, drone control, new medical applications, and autonomous vehicles. These types of applications are potentially the ones that will deliver the greatest societal benefits, yet unfortunately, at least in the United States, they could possibly be undermined by current network neutrality regulations. This category is also referred to as critical machine-type communications (cMTC).

Figure 12: ITU Use Case Model[26]

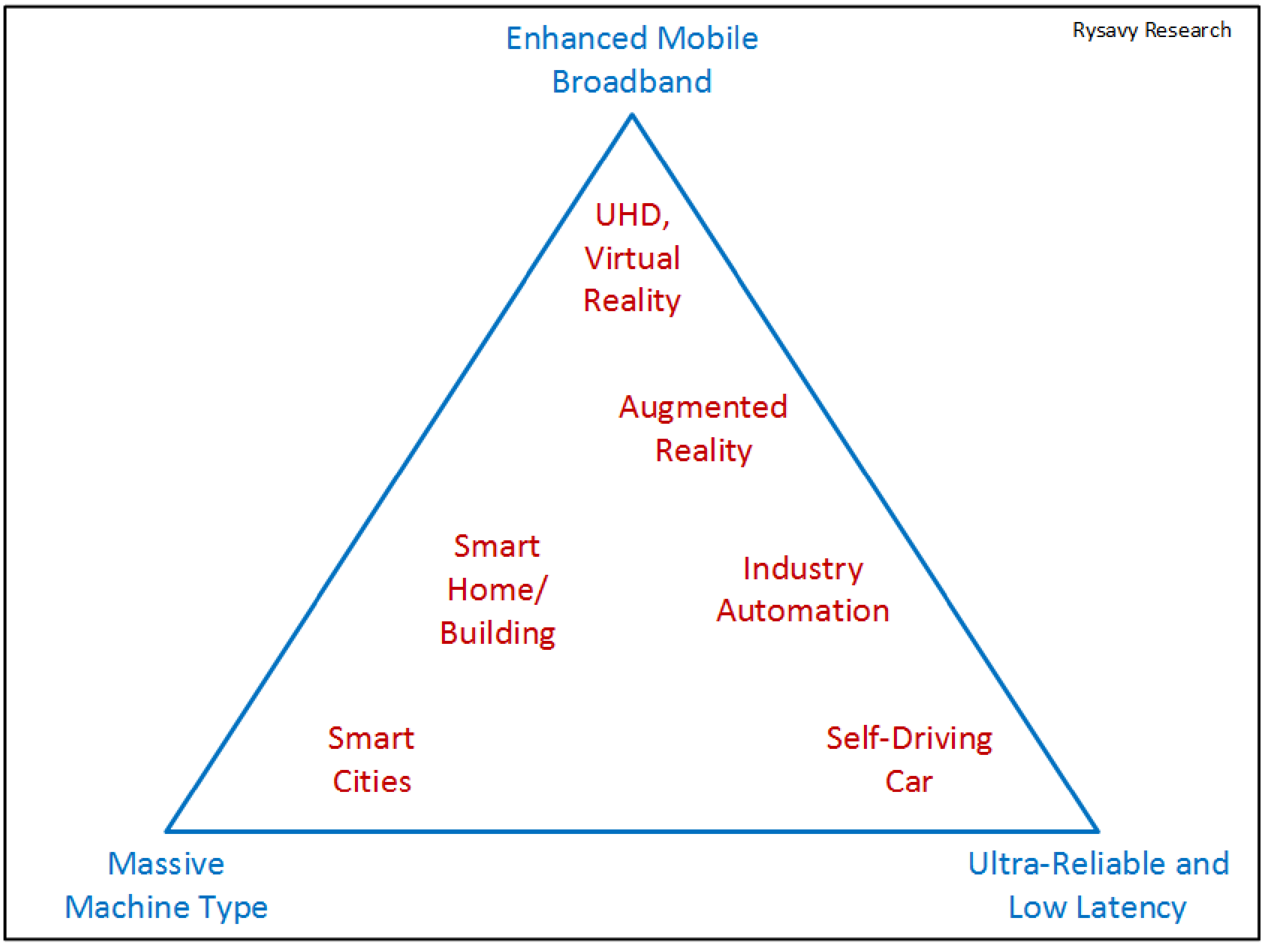
Rysavy Research
Enhanced Mobile Broadband
UHD, Virtual Reality
Augmented Reality
Smart Home/ Building
Industry Automation
Smart Cities
Self-Driving Car
Massive Machine Type
Ultra-Reliable and Low Latency

---

[26] For background, see ITU, *IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*, Recommendation ITU-R M.2083-0, Sep. 2015.

3GPP, in studying 5G, has methodically identified multiple specific use cases in a project called “SMARTER,” which are consistent with ITU’s model.[27] 3GPP’s service dimensions include:

- ❑ Massive Internet of Things (eHealth, wearables, eCity, eFarm).
- ❑ Critical Communications (e.g., vehicles, drones, industry robots).
- ❑ Enhanced Mobile Broadband (e.g., augmented reality, virtual reality, ultra-high definition).
- ❑ Network Operation (e.g., network slicing, connectivity/routing, migration/interworking).

Using small cells and mmWave radio channels, a 5G network built for capacity will deliver 1 Tbps/km^2 or higher, enabling 5G to compete with wireline broadband services.[28]

5G Technical Objectives

Table 4 shows the ITU’s objectives for IMT-2020 (5G) relative to IMT-Advanced (4G).

Table 4: ITU Objectives for IMT-2020 compared with IMT-Advanced[29]

	IMT-Advanced	IMT-2020
Peak Data Rate	DL: 1 Gbps UL: 0.05 Gbps	DL: 20 Gbps UL: 10 Gbps
User Experienced Data Rate	10 Mbps	100 Mbps[30]
Peak Spectral Efficiency	DL: 15 bps/Hz UL: 6.75 bps/Hz	DL: 30 bps/Hz UL: 15 bps/Hz
Average Spectral Efficiency		DL eMBB indoor: 9 bps/Hz DL eMBB urban: 7.8 bps/Hz DL eMBB rural: 3.3 bps/Hz UL eMBB indoor: 6.75 bps/Hz UL eMBB urban: 5.4 bps/Hz UL eMBB rural: 1.6 bps/Hz

---

[27] 3GPP TR22.891, *Feasibility Study on New Services and Markets Technology Enablers*; TR22.861 (Massive Internet of Things); TR22.862 (Critical Communications); TR 22.863 (Enhanced Mobile Broadband); TR22.864 (Network Operation).

[28] The hotspot capacity requirement of 10 Mbps/sq. m. is equivalent to 10 Tbps/sq. km. See also *Nokia, Ten key rules of 5G deployment, Enabling 1 Tbit/s/km^2 in 2030*, 2015.

[29] ITU Working Party 5D, *Minimal Requirements Related to Technical Performance for IMT-2020 Radio Interfaces,* Feb 22, 2017. See also 3GPP TR 38.913, *Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14),* V14.2.0, Mar. 2017.

[30] Per ITU, “User experienced data rate is the 5% point of the cumulative distribution function (CDF) of the user throughput.”

	IMT-Advanced	IMT-2020
Mobility	350 km/h	500 km/h
User Plane Latency	10 msec	1 msec[31]
Connection Density	100 thousand devices/sq.km.	1 million devices sq./km.
Network Energy Efficiency	1 (normalized)	100X over IMT-Advanced
Area Traffic Capacity	0.1 Mbps/sq. m.	10 Mbps/sq. m. (hot spots)
Bandwidth	Up to 20 MHz/radio channel (up to 100 MHz aggregated)	Up to 1 GHz (single or multipole RF carriers)

In supporting different usage scenarios, not all of these objectives will necessarily be simultaneously available. For example, an IoT application may need to support a large number of devices but at lower throughput rates, while a vehicular application may need high mobility and low latency.

Other expected enhancements include:

- ❑ Deep coverage for machines buried within environments.
- ❑ Extremely low energy demands for ten years or more of battery operation.
- ❑ Low complexity options for inexpensive machine communications.
- ❑ Auto-awareness through discovery and self-optimization.

5G Concepts and Architectures

Standards bodies have not yet defined 5G requirements, but various groups are analyzing the possibilities of what might constitute 5G for network deployments in 2020 or beyond. Often stated goals of 5G include:

- ❑ Being able to support a greater number of end-systems, including IoT applications, at lower average revenue than 4G systems.
- ❑ Peak data rates of multi Gbps (see Table 4 above).
- ❑ Rather than emphasizing peak rates, a more uniform user experience across the coverage area.
- ❑ Support for many frequencies, including existing cellular bands and frequencies above 6 GHz.

---

[31] Per 3GPP TR 38.913 (V14.2.0, Mar. 2017), 0.5 msec for DL and 0.5 msec for UL for URLCC and 4 msec for UL and 4 msec for DL for eMBB.

- ❑ Availability of TDD and FDD modes for all bands.
- ❑ Hierarchical/planned and ad hoc deployment models.
- ❑ Use of licensed and unlicensed bands.
- ❑ Equal support for human-type and machine-type communications. Includes highly efficient small data transmission.
- ❑ Advanced Spectrum Sharing, possibly based on spectrum sharing approaches being developed for the 3.5 GHz band.

Figure 13 shows the transformation of networks, moving from today's LTE-Advanced networks to future LTE-Advanced and eventually 5G networks.

Figure 13: Network Transformation[32]

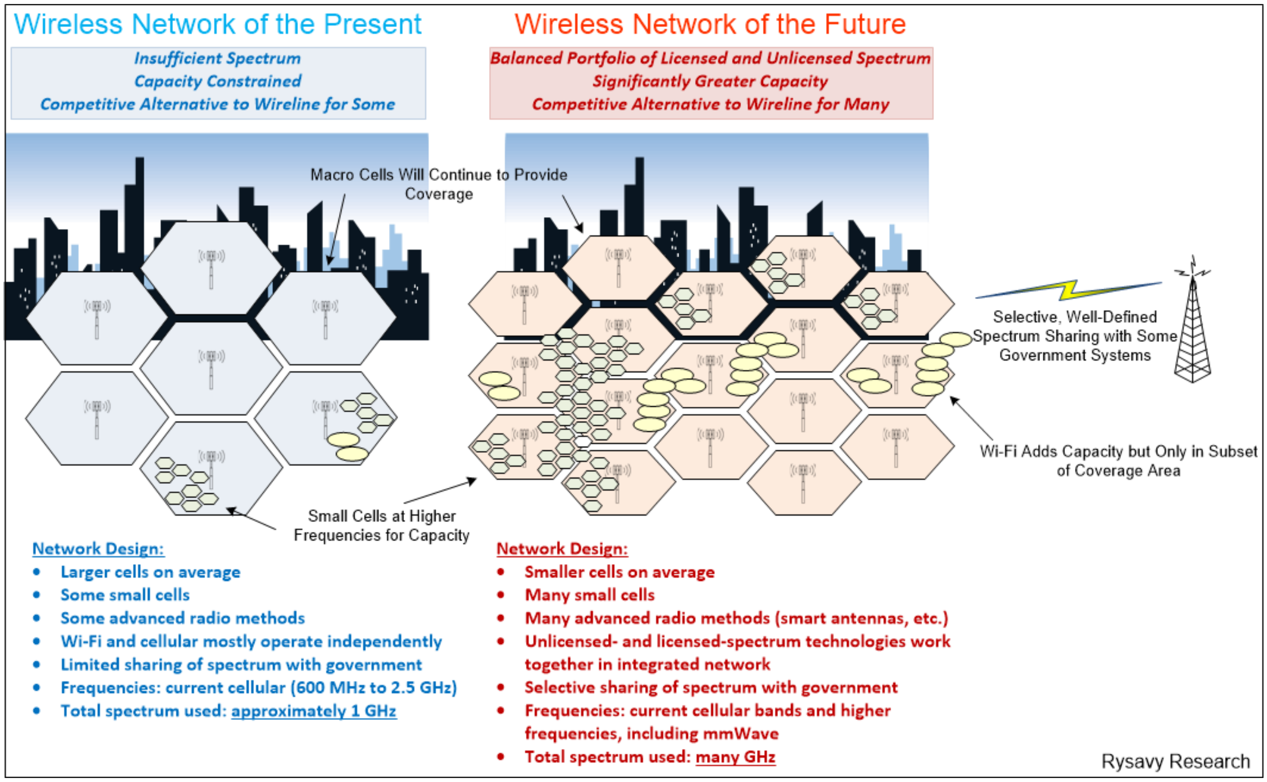
Wireless Network of the Present
Insufficient Spectrum
Capacity Constrained
Competitive Alternative to Wireline for Some
Wireless Network of the Future
Balanced Portfolio of Licensed and Unlicensed Spectrum
Significantly Greater Capacity
Competitive Alternative to Wireline for Many
Macro Cells Will Continue to Provide Coverage
Selective, Well-Defined Spectrum Sharing with Some Government Systems
Wi-Fi Adds Capacity but Only in Subset of Coverage Area
Small Cells at Higher Frequencies for Capacity

Network Design:

- Larger cells on average
- Some small cells
- Some advanced radio methods
- Wi-Fi and cellular mostly operate independently
- Limited sharing of spectrum with government
- Frequencies: current cellular (600 MHz to 2.5 GHz)
- Total spectrum used: approximately 1 GHz

Network Design:

- Smaller cells on average
- Many small cells
- Many advanced radio methods (smart antennas, etc.)
- Unlicensed- and licensed-spectrum technologies work together in integrated network
- Selective sharing of spectrum with government
- Frequencies: current cellular bands and higher frequencies, including mmWave
- Total spectrum used: many GHz

Rysavy Research

The fundamental decision for 5G is how to best leverage existing technology investments in LTE while exploiting new spectrum and new technology capabilities. 5G design emphasizes ways to combine existing 4G LTE networks with capabilities provided by 5G. One approach likely to be used by many operators is to use LTE in existing frequency bands and the 5G NR in new bands, such as mmWave, as shown in Figure 14. 5G NR, however, will operate in all frequencies, and just as 2G and 3G spectrum has been re-farmed for LTE, existing cellular bands will eventually be re-farmed for 5G.

---

[32] See also Rysavy Research infographic, "Mobile Broadband Networks of the Future," April 2014. Available at http://www.rysavy.com/Articles/2014-05-Networks-of-the-Future-Infographic.pdf.

Figure 14: 5G Combining of LTE and New Radio Technologies

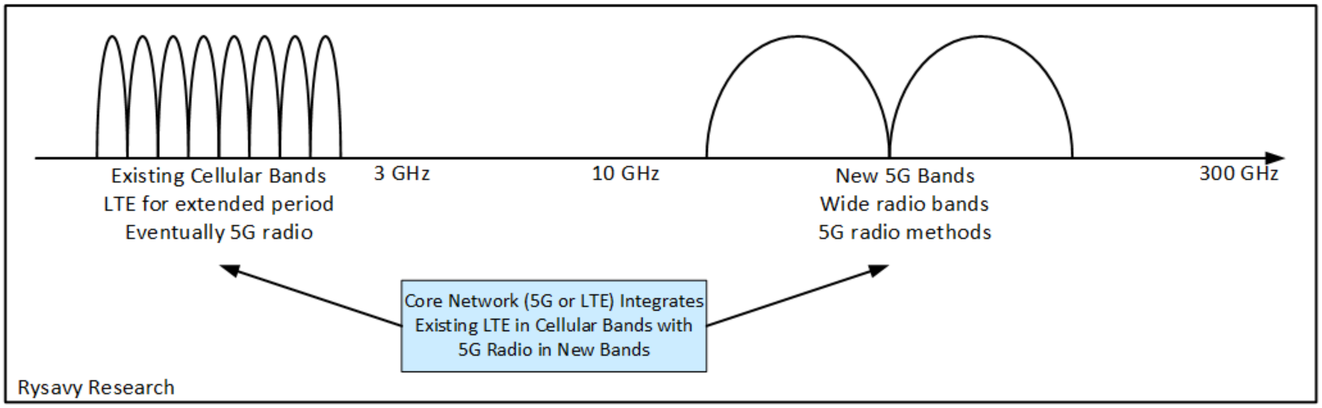
Existing Cellular Bands
LTE for extended period
Eventually 5G radio
3 GHz
10 GHz
New 5G Bands
Wide radio bands
5G radio methods
300 GHz
Core Network (5G or LTE) Integrates Existing LTE in Cellular Bands with 5G Radio in New Bands
Rysavy Research

As shown in Figure 15, higher frequency bands in 5G will provide capacity with smaller cells, and lower bands will provide coverage with larger cells. This is similar to the approach taken in 4G.

Figure 15: Characteristics of Different Bands[33]

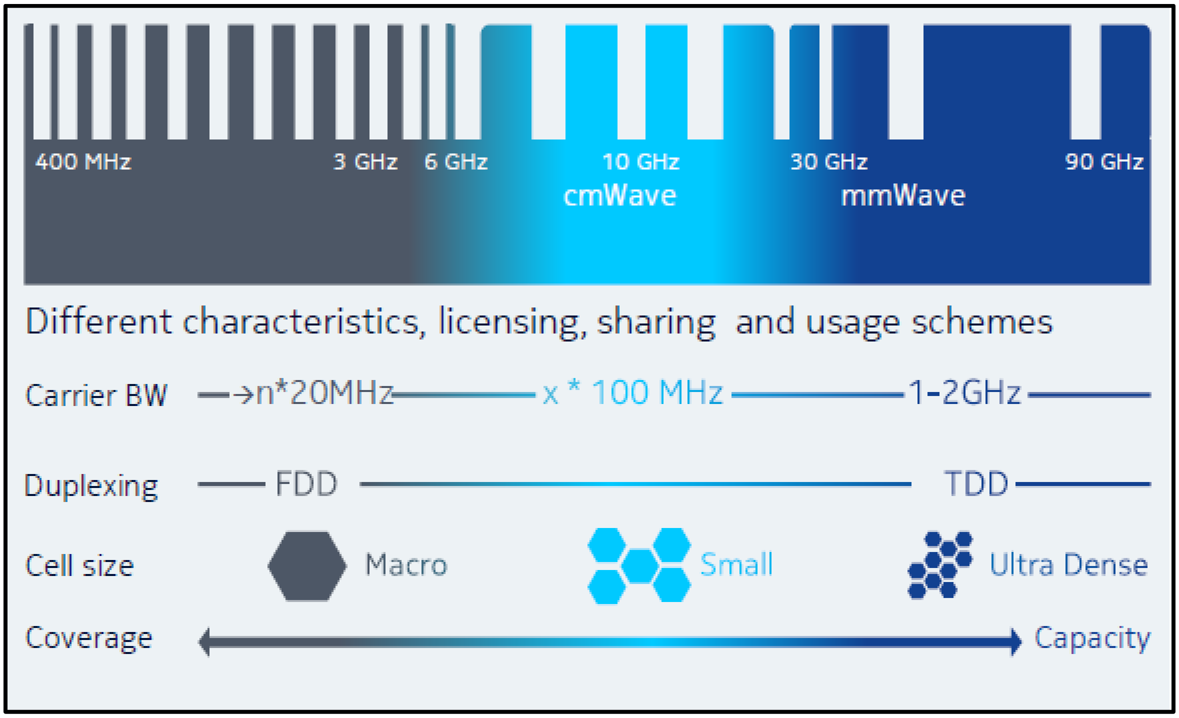
400 MHz
3 GHz
6 GHz
10 GHz
cmWave
30 GHz
mmWave
90 GHz
Different characteristics, licensing, sharing and usage schemes
Carrier BW
→n*20MHz
x * 100 MHz
1-2GHz
Duplexing
FDD
TDD
Cell size
Macro
Small
Ultra Dense
Coverage
Capacity

---

[33] Nokia, *Vision & Priorities for Next Generation Radio Technology*, 3GPP RAN workshop on 5G, Sep. 17-18, 2015.

New radio methods in NR will boost performance and will co-exist with future versions of LTE. 3GPP, in Release 14, completed a study item on 5G NR and will specify NR in Release 15.

Figure 16: Evolution to 5G Including LTE Improvements and New 5G Radio Methods

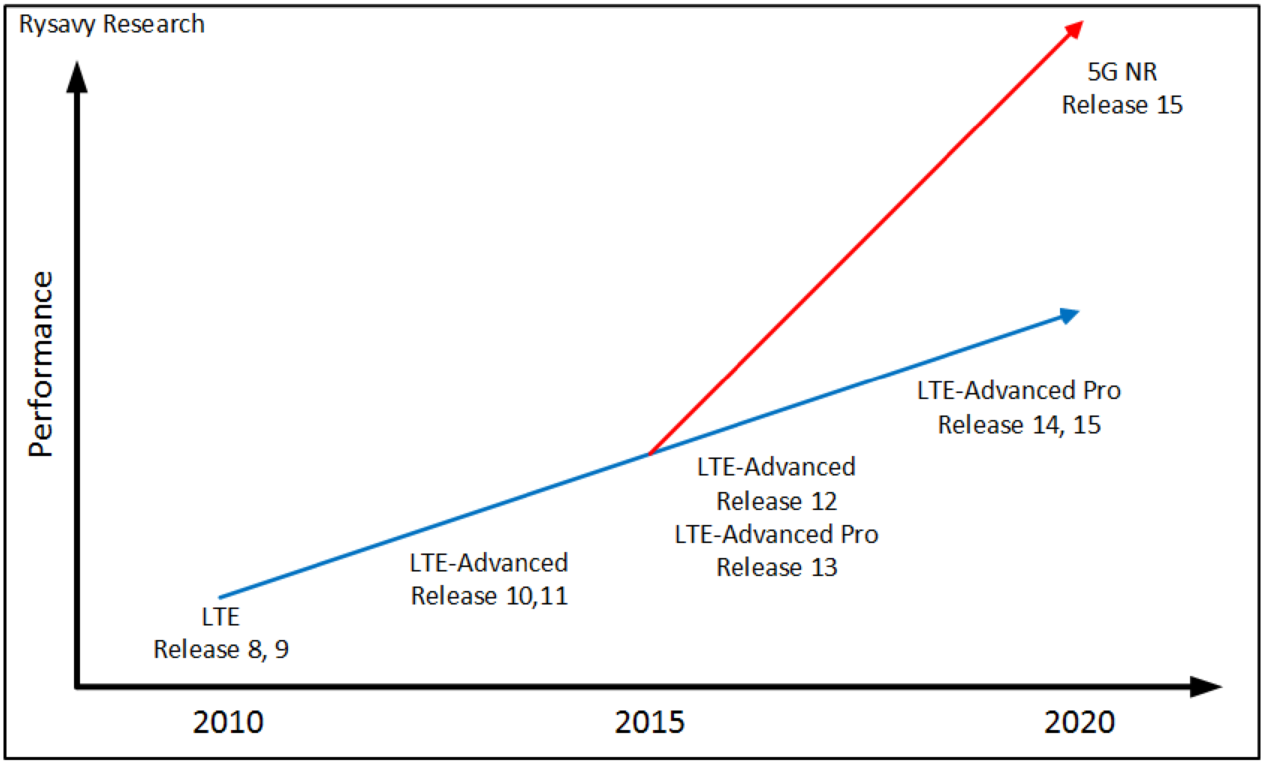

While many deployments will integrate LTE and NR, some operators could choose NR-only deployments.

5G New Radio (NR)

3GPP is defining 5G NR in Release 15, with an initial release for non-standalone operation scheduled for March 2018 and full release for standalone version scheduled for September 2018. Release 16 will add additional capabilities to the radio, such as support for URLLC. Despite specifications not being finalized, the following are planned capabilities and features of NR in Release 15:

- Ability to operate in any frequency band, including low, mid, and high bands.
- Network can support both LTE and 5G NR, including dual connectivity with which devices have simultaneous connections to LTE and NR.
- Carrier aggregation for multiple NR carriers.
- 5 Gbps peak downlink throughput in initial releases, increasing to 50 Gbps in subsequent versions.
- OFDMA in downlink and uplink, with optional Single Carrier Frequency Division Multiple Access (SC-FDMA) for uplink. Radio approach for URLLC to be defined in

Release 16, but Release 15 will provide physical layer frame structure and numerology support.

- Massive MIMO and beamforming.
- Ability to support either FDD or TDD modes for 5G radio bands.
- Numerologies of 2^N X 15 kHz for subcarrier spacing up to 120 kHz or 240 kHz. This approach, depicted in Figure 17, supports both narrow radio channels (for example, 1 MHz), or wide ones (for example, 400 MHz). Phase 1 likely to support a maximum of 400 MHz bandwidth with 240 kHz subcarrier spacing.
- Error correction through low-density parity codes (LDPC), which are computationally more efficient than LTE turbo codes at higher data rates.
- Standards-based cloud RAN support, compared with proprietary LTE approaches, that specifies a split between the PDCP and Radio Link Control (RLC) protocol layers.
- Self-contained integrated subframes that combine scheduling, data, and acknowledgement.
- Future-proofing by providing a flexible radio framework that has forward compatibility to support future, currently unknown services, such as URLLC to be specified in Release 16.
- Scalable transmission time intervals with short time intervals for low latency and longer time intervals for higher spectral efficiency.
- QoS support using a new model.
- Dynamic co-existence with LTE in the same radio channels. (See the Appendix section “LTE-NR Co-existence” for more details.)

According to one 5G Americas member company in 2017, operators globally have expressed interest in deploying NR in a wide variety of bands, including current cellular bands, 3.5 GHz, and mmWave bands.

Release 16 will add support for:

- URLCC.
- Shared spectrum operation.
- 5 GHz unlicensed spectrum operation based on current LTE approaches such as LAA, LWA, LWIP, and RAN Controlled LTE Wireless LAN Interworking (RCLWI).
- Possibly integrated access and backhaul based on a Release 15 study item. See the appendix section, “Integrated Access and Backhaul,” for more details.
- Other, as yet unknown, features.

The ability to simultaneously transmit and receive on the same frequency has been stated in the past as an objective of 5G, and although such capability remains of interest, it is not currently being specified.

Figure 17: Hypothetical Example of 5G Numerology

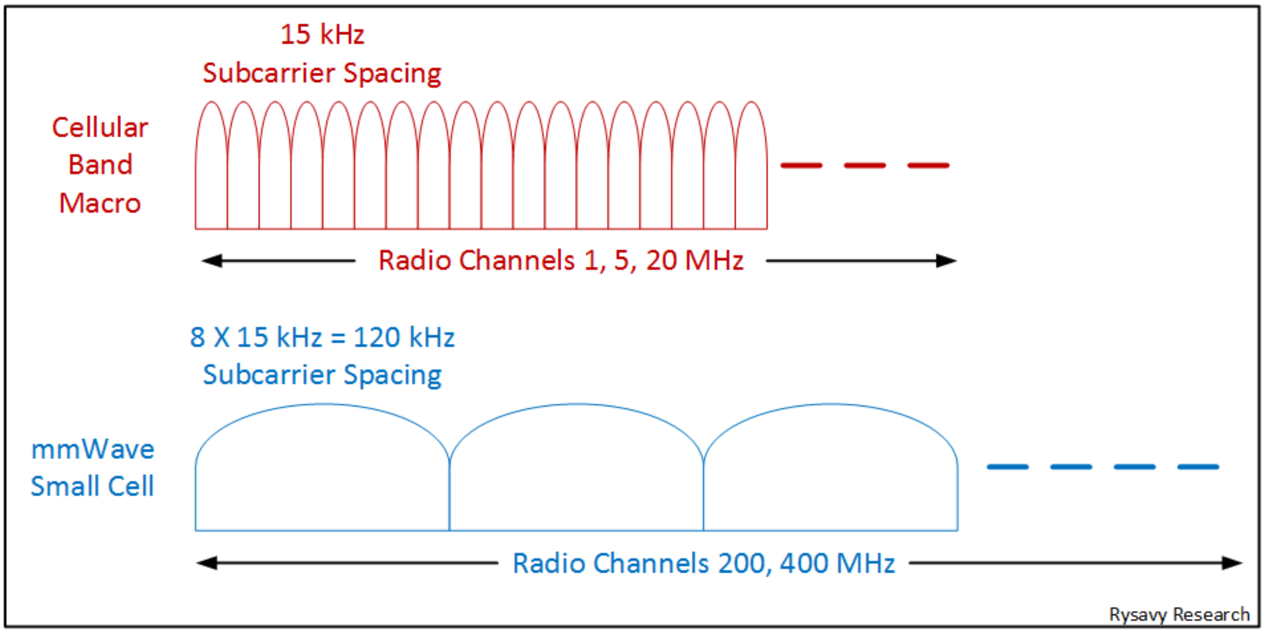

Expected 5G Performance

Predicting the 5G user experience is challenging because 5G will be deployed in many configurations, including different bands and with varying width radio channels. In addition, the throughput rates a user experiences depend on signal quality, device capability, and network loading. Some early predictions, however, are possible by examining the ITU objectives, the expected width of radio channels, and results from simulation studies.

The ITU objectives are that in a 5G network, 95% or more users should be able to experience at least 100 Mbps, which is also termed as cell-edge throughput. Examining the simulation results in the appendix section on 5G, "Performance," such results appear feasible using 200 MHz or wider radio channels. The results also suggest that average or typical rates could reach 500 Mbps using 200 MHz radio channels and 1 Gbps using 400 MHz radio channels.

Just as LTE throughputs have increased significantly over this decade, 5G performance will keep improving over the next ten years. One 5G Americas member expects that the 5G theoretical peak rate could be 5 Gbps in 2020, rising to 50 Gbps by 2024.

Release 15 Non-Standalone and Standalone Options

3GPP is specifying the first phase of 5G in Release 15. So that operators can deploy 5G sooner, 3GPP divided Release 15 into two sets of specifications. The first and earlier set of specifications define how a 5G RAN can integrate with an LTE network in what 3GPP calls a non-standalone option. In this earliest version (architecture option 3), NR will rely on an existing LTE network, both in the RAN and in the core.

The complete Release 15 specifications will also define a 5G core network. Figure 18 shows some of the architecture options. Options 3, 4 and 7 are the non-standalone options and options 1, 2, and 5 are standalone. The appendix section, "5G Architecture Options," discusses all of the deployment options in greater detail.

Figure 18**: Release 15 Non-Standalone and Standalone Options**

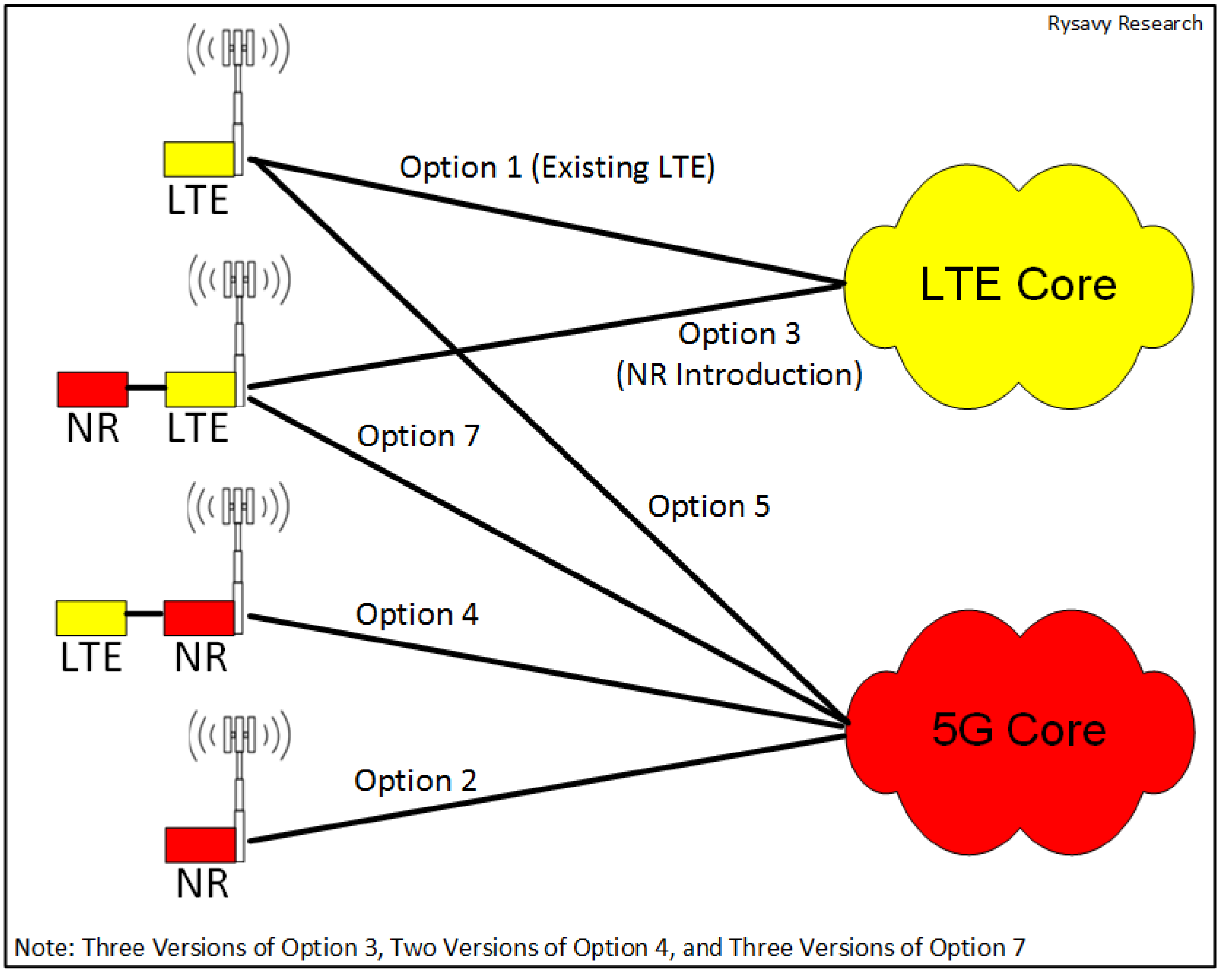
Rysavy Research
LTE
Option 1 (Existing LTE)
LTE Core
NR
LTE
Option 3
(NR Introduction)
Option 7
Option 5
LTE
NR
Option 4
5G Core
Option 2
NR

Note: Three Versions of Option 3, Two Versions of Option 4, and Three Versions of Option 7

5G Schedule3GPP is currently standardizing 5G in Release 15, which will complete the non-standalone version of 5G in March 2018. Based on a typical minimum period of 18 months to build and deploy the technology, initial 5G NSA deployments could occur toward the end of 2019 or the beginning of 2020. 3GPP will complete the full Release 15 specifications in September 2018, enabling deployments in 2020. Release 16, which is the second phase of 5G, will be complete at the end of 2019, and Release 16 deployments could occur in 2021. In 2020, 3GPP will begin work on Release 17 which will include as yet unknown capabilities.

Figure 19Figure 19 shows the current schedule for 5G development and deployment. 3GPP is currently standardizing 5G in Release 15, which will complete the non-standalone version of 5G in March 2018. Based on a typical minimum period of 18 months to build and deploy the technology, initial 5G NSA deployments could occur toward the end of 2019 or the beginning of 2020. 3GPP will complete the full Release 15 specifications in September 2018, enabling deployments in 2020. Release 16, which is the second phase of 5G, will be complete at the end of 2019, and Release 16 deployments could occur in 2021. In 2020, 3GPP will begin work on Release 17 which will include as yet unknown capabilities.

Figure 19: 5G Schedule[34]

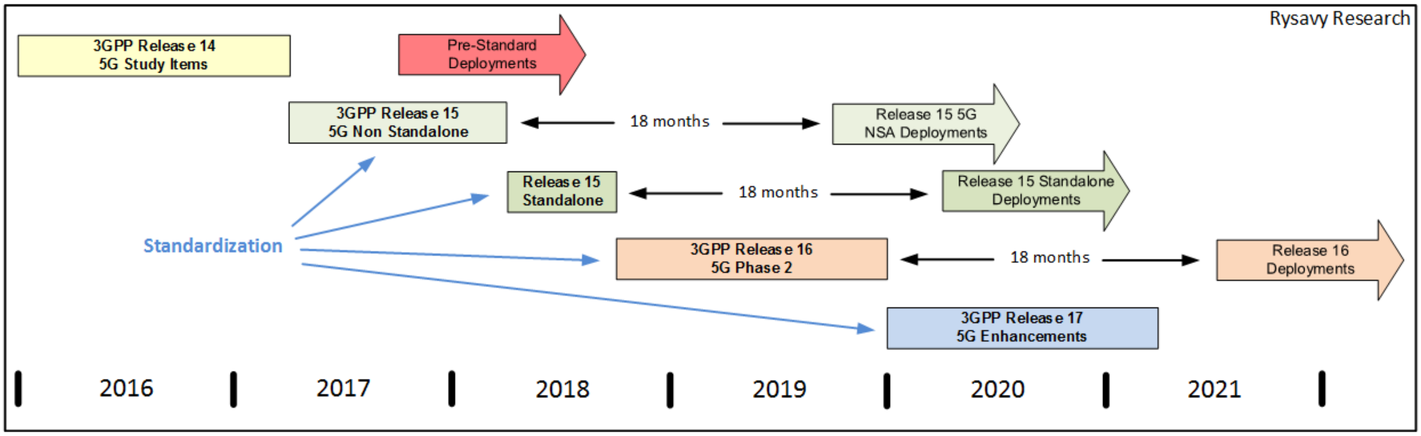
Rysavy Research
3GPP Release 14
5G Study Items
Pre-Standard
Deployments
3GPP Release 15
5G Non Standalone
18 months
Release 15 5G
NSA Deployments
Release 15
Standalone
18 months
Release 15 Standalone
Deployments
Standardization
3GPP Release 16
5G Phase 2
18 months
Release 16
Deployments
3GPP Release 17
5G Enhancements
2016
2017
2018
2019
2020
2021

5G Core Network

Release 15 also defines initial core network capabilities that support QoS and network slicing. See the 5G appendix section, "Architecture," for more details. A study item in Release 15 will study integrated access and backhaul, with which base stations can use the 5G radio channel for backhaul. This capability could become a work item for Release 16. See the appendix section, "Integrated Access and Backhaul," for more details.

Many operators will virtualize their 5G core networks, just as they have for LTE, but such virtualization is outside the scope of 3GPP specifications.

5G Device Availability

User devices capable of 5G operation have not yet been announced, but availability will likely follow the trends of previous generations of networks. Initial devices, possibly in the 2019 timeframe, will likely include routers that have a 5G radio and use Wi-Fi for local Hotspot capability and USB modems. Handset vendors are in the early stages of designing mmWave support into smartphones.[35] These devices could come online in the 2021 timeframe, although this estimate could tighten or lengthen depending on chipset availability and handset vendor plans.

---

[34] 3GPP and 5G Americas member contribution. Completion based on anticipated ASN.1 freeze dates.

[35] For example, see "Wireless Week, Apple Wants to Test Performance at 28 GHz, 39 GHz for 5G Devices," May 24, 2017, https://www.wirelessweek.com/news/2017/05/apple-wants-test-performance-28-ghz-39-ghz-5g-devices, viewed May 29, 2017.

mmWave

Use of higher frequencies, such as above 6 GHz, represents one of the greatest opportunities for higher throughputs and higher capacity. This benefit derives from the potential availability of ten times the amount of spectrum as is currently available, with multiple GHz of contiguous spectrum. But these higher frequencies, especially mmWave frequencies (above 30 GHz), are suitable only over short distances, as explained below. The combination of lower and higher frequencies is therefore crucial for 5G operation. Lower bands can be devoted to coverage and control, while higher bands can provide opportunistic access for high data rates. The lower and higher spectrum bands can operate in a carrier aggregation or dual connectivity model. Initial 5G specifications include such dual-connectivity capability.

Compared with lower frequencies, mmWave frequencies suffer from worse propagation characteristics, even in line-of-sight conditions, because the comparatively smaller aperture area of the receiver's antenna requires some form of beamforming at the transmit side, and potentially even at the receive side. Fortunately, the smaller form factors of mmWave antennas allow for dense packing of antenna arrays. Experimental systems using antenna arrays have demonstrated reliable communications at 28 GHz, even in dense, urban, non-line-of-sight conditions, for distances up to 200 meters.[36] Arrays at the terminal side are space-constrained, but some basic beamforming at the terminal is possible. On the base station side, the arrays may include hundreds of antennas in an approach called "massive MIMO."

Figure 20 shows how an increasing number of antenna elements can extend coverage through tighter beams. A 77 X 77 antenna array (6,000 elements) can exceed a kilometer at 3.5 GHz (33 dBm transmit power) and reach over 800 meters, even at 30 GHz.

---

[36] Samsung, *5G Vision,* February 2015. Available at http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf.

Figure 20: Range Relative to Number of Antenna Elements[37]

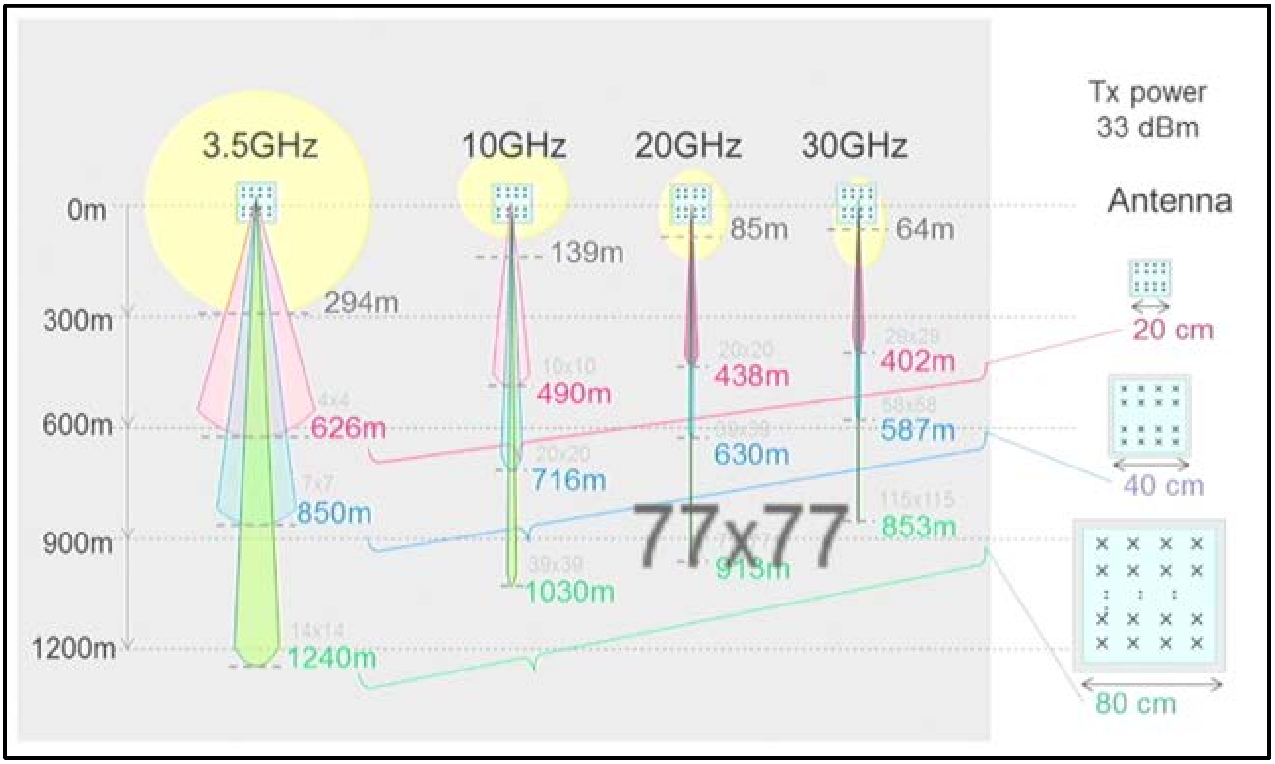

More typically, mmWave cells will employ shorter ranges of 50 to 200 meters. Extreme densification is another way that 5G networks will achieve massive increases in capacity. 3G networks reached densities of four to five base stations per sq. km, 4G networks eight to 10, but 5G networks may reach densities of 40 to 50.[38] A likely 5G architecture will use the macro cell for control information, coverage, and fallback, but small cells, often operating at higher frequencies, for high-bandwidth data communication. Either wireless connections or fiber will provide backhaul. Figure 21 shows how such an approach could also employ beamforming and beam tracking when using mmWave bands in the small cells.

---

[37] Dr. Seizo Onoe, NTT DOCOMO, presentation at Brooklyn 5G Summit, Apr. 21, 2016. Used by permission.

[38] IEEE Wireless Communications, *5G Ultra-Dense Cellular Networks,* Feb. 2016.

Figure 21: 5G Architecture for Low-Band/High-Band Integration

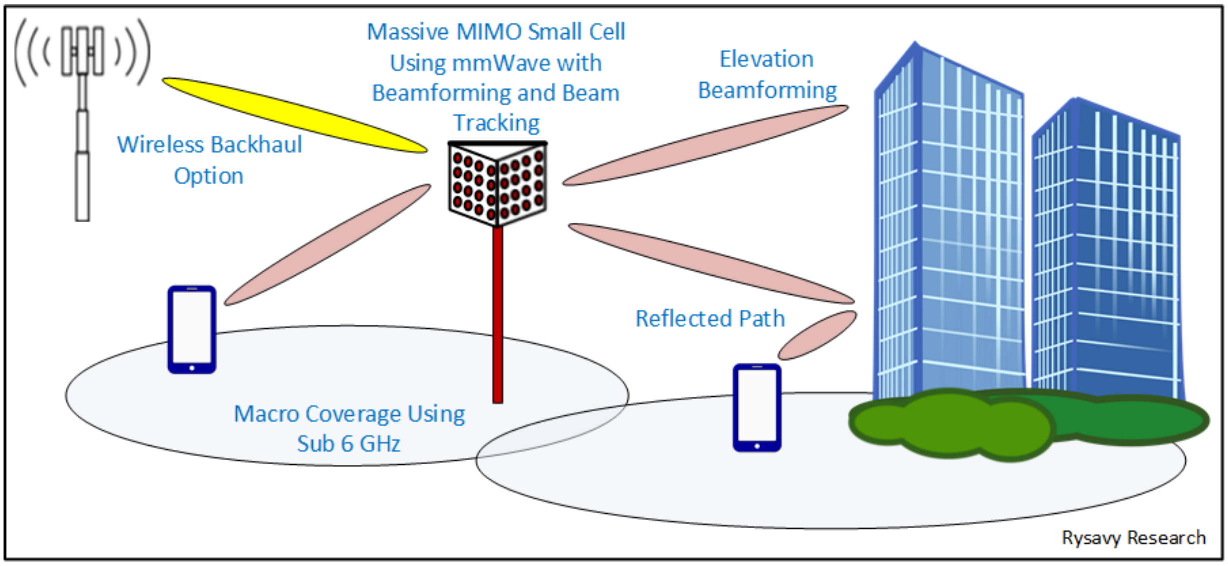
Massive MIMO Small Cell Using mmWave with Beamforming and Beam Tracking
Elevation Beamforming
Wireless Backhaul Option
Reflected Path
Macro Coverage Using Sub 6 GHz
Rysavy Research

In combination, the various methods expected in 5G will provide users in mmWave band hotspot coverage at least a 100-fold increase in throughput over LTE, achieved by:

- ❑ Five- to 10-fold gains due to fewer users in each small cell. (Five to ten times as many cells.)
- ❑ 10-fold gains from access to much larger amounts of spectrum.
- ❑ Three-fold gains or more from improved spectral efficiency.

It is this huge increase in capacity, combined with Gbps performance, that will allow 5G to compete with wireline networks.[39]

Network Slicing

Not only will 5G networks include a new radio, but thanks to virtualization, these networks will be able to present multiple faces for different use cases using a concept called network slicing. This architecture allows an operator to provide multiple services with different performance characteristics. Each network slice operates as an independent, virtualized version of the network. For an application, the network slice is the only network it sees. The other slices, to which the customer is not subscribed, are invisible and inaccessible. The advantage of this architecture is that the operator can create slices that are fine-tuned for specific use cases. One slice could target autonomous vehicles, another enhanced mobile broadband, another low-throughput IoT sensors, and so on.

Figure 22 shows the network slicing architecture, with devices having access to only the slice or slices for which they have subscriptions. Each slice has radio resources allocated, with specific QoS characteristics. Within the core network, virtualized core network functions support each slice and provide connections to external networks.[40]

---

[39] For a further discussion of 5G capacity and ability to compete with wireline networks, refer to Datacomm Research and Rysavy Research, *Broadband Disruption: How 5G Will Reshape the Competitive Landscape*, 2017, available at https://datacommresearch.com/reports-broadband/.

Figure 22: Network Slicing Architecture

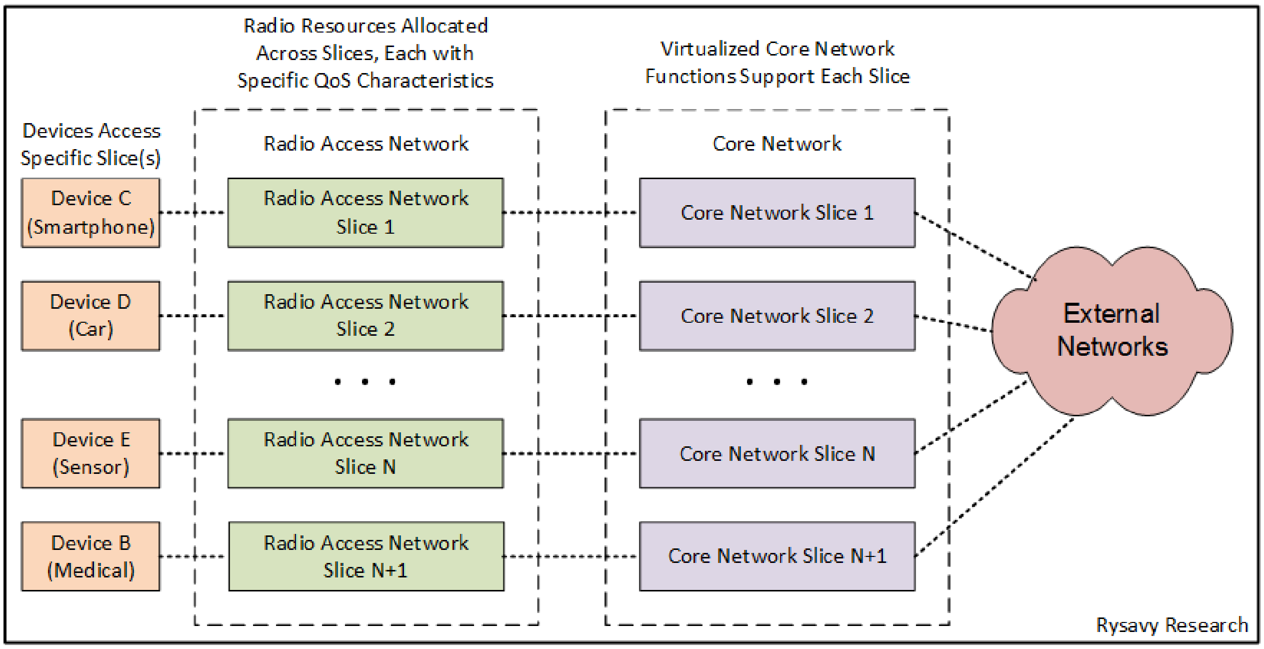
Radio Resources Allocated Across Slices, Each with Specific QoS Characteristics
Virtualized Core Network Functions Support Each Slice
Devices Access Specific Slice(s)
Radio Access Network
Core Network
Device C (Smartphone)
Device D (Car)
Device E (Sensor)
Device B (Medical)
Radio Access Network Slice 1
Radio Access Network Slice 2
Radio Access Network Slice N
Radio Access Network Slice N+1
Core Network Slice 1
Core Network Slice 2
Core Network Slice N
Core Network Slice N+1
External Networks
Rysavy Research

Information-Centric Networking

For many usage scenarios, wireless networks provide broadband access to the internet, a network that itself is evolving. The internet is based on a node-centric design developed forty years ago. The point-to-point method of communication the internet uses has functioned well for a vast array of applications but is not optimal for the way content is developed and distributed today. Industry and academic organizations are researching a concept called “Information-Centric Networking.” ICN seeks a new approach of in-network caching that distributes content on a large scale, cost-efficiently and securely.

Most internet content uses Uniform Resource Identifiers (URIs) to locate objects and define specific location-dependent IP addresses. This approach, however, causes problems when content moves, sites change domains, or content is replicated and each copy appears as a different object. Developments such as peer-to-peer overlays and content distribution networks (such as Akamai) that distribute cached copies of content are a first step toward an information-centric communication model.

ICN is built from the ground up on the assumption of mobility, so it eliminates the mobility overlays on which current mobile broadband networks depend. The approach will be able to place information anywhere in the network with immediate and easy retrieval.

Key principles of ICN include:

---

[40] For more details, see 5G Americas, *Network Slicing for 5G Networks & Services*, November 2016. Available at: http://www.5gamericas.org/files/3214/7975/0104/5G_Americas_Network_Slicing_11.21_Final.pdf.

- The architecture inherently supports user mobility.
- Network operations are name-based instead of address- or node-based.
- The network itself stores, processes, and forwards information.
- Intrinsic security guarantees the integrity of every data object.

The goal of ICN is to simplify the storage and distribution of gigantic amounts of content while reducing the amount of traffic and latency users face when accessing the content. The internet cannot just be replaced, however, so in initial stages, ICN would operate as an overlay, and over time would assume an increasing percentage of the functions within the internet. ICN would not discard IP; rather, it seeks to generalize the routing concept to enrich networking with new capabilities.

Some technology aspects of ICN include:

- Information retrieval from multiple sources without needing to know the location of the information.
- Multipath communications that improves user performance and traffic load balancing.
- Subsequent requests for the same data will be served locally without needing to fetch it from original repository.
- Elimination of the name-to-location indirection associated with Domain Name Service (DNS).

Because mobility is such a central aspect of ICN, mobile network operators are in a unique position to participate in ICN-related research and development, and to do so as part of 5G development. ICN has not progressed to a level at which 3GPP specification work could include it, so instead promoters are ensuring that 5G specification work does not preclude it. With this approach, operators in the 2020s will have the option of overlaying ICN capability on their 5G networks. ICN could even be implemented as a 5G network slice for mobile and end-systems capable of ICN.

3GPP Releases

3GPP standards development falls into three principal areas: radio interfaces, core networks, and services. Progress in the 3GPP family of technologies has occurred in multiple phases, first with GSM, then GPRS, EDGE, UMTS, HSPA, HSPA+, LTE, LTE-Advanced, and now LTE-Advanced Pro. Underlying radio approaches have evolved from Time Division Multiple Access (TDMA) to CDMA to Orthogonal Frequency Division Multiple Access (OFDMA), which is the basis of LTE. 3GPP is also involved in standardization of 5G technology.

Table 5 summarizes the key 3GPP technologies and their characteristics.

Table 5: Characteristics of 3GPP Technologies

Technology Name	Type	Characteristics	Typical Downlink Speed	Typical Uplink Speed
HSPA[41]	WCDMA	Data service for UMTS networks. An enhancement to original UMTS data service.	1 Mbps to 4 Mbps	500 Kbps to 2 Mbps
HSPA+	WCDMA	Evolution of HSPA in various stages to increase throughput and capacity and to lower latency.	1.9 Mbps to 8.8 Mbps in 5+5 MHz[42] 3.8 Mbps to 17.6 Mbps with dual-carrier in 10+5 MHz	1 Mbps to 4 Mbps in 5+5 MHz or in 10+5 MHz
LTE	OFDMA	New radio interface that can use wide radio channels and deliver extremely high throughput rates. All communications handled in IP domain.	6.5 to 26.3 Mbps in 10+10 MHz[43]	6.0 to 13.0 Mbps in 10+10 MHz
LTE-Advanced	OFDMA	Advanced version of LTE designed to meet IMT-Advanced requirements.	Significant gains through carrier aggregation, 4X2 and 4X4 MIMO, and 256 QAM modulation.	

User-achievable rates and additional details on typical rates are covered in the appendix section "Data Throughput."

3GPP develops specifications in releases, with each release addressing multiple technologies. For example, Release 8 defined dual-carrier operation for HSPA but also introduced LTE. Each release adds new features and improves performance of existing functionality in different ways. Table 6 summarizes some key features of different 3GPP releases.

---

[41] HSPA and HSPA+ throughput rates are for a 5+5 MHz deployment.

[42] "5+5 MHz" means 5 MHz used for the downlink and 5 MHz used for the uplink.

[43] 5G Americas member company analysis for downlink and uplink. Assumes single user with 50% load in other sectors. AT&T and Verizon are quoting typical user rates of 5-12 Mbps on the downlink and 2-5 Mbps on the uplink for their networks. See additional LTE throughput information in the section below, "LTE Throughput."

Table 6: Key Features in 3GPP Releases[44]

Release	Year	Key Features
99	1999	First deployable version of UMTS.
5	2002	High Speed Downlink Packet Access (HSDPA) for UMTS.
6	2005	High Speed Uplink Packet Access (HSUPA) for UMTS.
7	2008	HSPA+ with higher-order modulation and MIMO.
8	2009	Long Term Evolution. Dual-carrier HSDPA.
10	2011	LTE-Advanced, including carrier aggregation and eICIC.
11	2013	Coordinated Multi Point (CoMP).
12	2015	Public safety support. Device-to-device communications. Dual Connectivity. 256 QAM on the downlink.
13	2016	LTE-Advanced Pro features. LTE operation in unlicensed bands using LAA. Full-dimension MIMO. LTE-WLAN Aggregation. Narrowband Internet of Things.
14	2017	LTE-Advanced Pro additional features, such as eLAA (adding uplink to LAA) and cellular V2X communications. Study item for 5G “New Radio.”
15	2018	Additional LTE-Advanced Pro features, such as ultra-reliable low-latency communications. Phase 1 of 5G. Will emphasize enhanced mobile broadband use case and sub-40 GHz operation. Includes Massive MIMO, beamforming, and 4G-5G interworking, including ability for LTE connectivity to a 5G CN.
16	2019	Phase 2 of 5G. Full compliance with ITU IMT-2020 requirements. Will add URLLC, spectrum sharing, unlicensed operation, and multiple other enhancements.
17	2021	Further LTE and 5G enhancements.

Refer to the Appendix section “3GPP Releases” for a more detailed listing of features in each 3GPP Release.

---

[44] After Release 99, release versions went to a numerical designation beginning with Release 4, instead of designation by year.

Supporting Technologies and Architectures

Network architects design networks using a deep and wide toolkit, including multiples types of cell sizes, integration with unlicensed spectrum, smart antennas, converged services, and virtualization.

Multiple Cell Types

Operators have many choices for providing coverage. Lower frequencies propagate further and thus require fewer cells for coverage. The resulting network, however, has lower capacity than one with more cells, so operators must continually evaluate cell placement with respect to both coverage and capacity.

Table 7 lists the many types of cells. Note that the distinctions, such as radius, are not absolute—perhaps one reason the term "small cell" has become popular, as it encompasses picocells, metrocells, femtocells, and sometimes Wi-Fi.

With "plug-and-play" capability derived from self-configuring and self-organizing features, small cells will increasingly be deployed in an ad hoc manner, anywhere power and backhaul are available, yet will operate in tight coordination with the rest of the network.

A proliferation of small cells inside buildings will also provide coverage from inside to outside, such as in city streets, the reverse of traditional coverage that extends from outdoor cells to inside.

Table 7: Types of Cells and Typical Characteristics (Not Formally Defined)

Type of Cell	Characteristics
Macro cell	Wide-area coverage. LTE supports cells up to 100 km in range, but typical distances are .5 to 5 km radius. Always installed outdoors.
Microcell	Covers a smaller area, such as a hotel or mall. Range to 2 km, 5-10W, and 256-512 users. Usually installed outdoors.
Picocell	Indoor or outdoor. Outdoor cells, also called "metrocells." Typical range 15 to 200 meters outdoors and 10 to 25 meters indoors, 1-2W, 64-128 users. Deployed by operators primarily to expand capacity.
Consumer Femtocell	Indoors. Range to 10 meters, less than 50 mW, and 4 to 6 users. Capacity and coverage benefit. Usually deployed by end users using their own backhaul.
Enterprise Femtocell	Indoors. Range to 25 meters, 100-250 mW, 16-32 users. Capacity and coverage benefit. Deployed by operators.
Distributed antenna system	Expands indoor or outdoor coverage. Same hardware can support multiple operators (neutral-host) since antenna can support broad frequency range and multiple technologies. Indoor deployments are typically in larger spaces such as airports. Has also been deployed outdoors for coverage and capacity expansion.

Type of Cell	Characteristics
Remote radio head (RRH)	Uses baseband at existing macro site or centralized baseband equipment. If centralized, the system is called “cloud RAN.” Requires fiber connection.
Wi-Fi	Primarily provides capacity expansion. Neutral-host capability allows multiple operators to share infrastructure.

Smalls Cells and Heterogeneous Networks

Historically, increasing the number of cell sites has been the primary method for increasing capacity, providing gains far greater than what can be achieved by improvements in spectral efficiency alone. The next wave of densification is by using what the industry calls “small cells.”

Central to small-cell support is the heterogeneous network architecture, with multiple types of cells serving a coverage area, varying in frequencies used, radius, and even radio technology used.

HetNets offer significant increases in capacity and improvements, including:

1. Smaller cells, such as open femtocells (home-area coverage) and picocells (city-block-area coverage), inherently increase capacity because each cell serves a smaller number of users.

2. Strategic placement of picocells within the macro cell provides the means to absorb traffic in areas where there are higher concentrations of users. Locations can include businesses, airports, stadiums, convention centers, hotels, hospitals, shopping malls, high-rise residential complexes, and college campuses.

3. Smaller cells can also improve signal quality in areas where the signal from the macro cell is weak.

Essential elements for practical HetNet deployment are self-optimization and self-configuration, especially as the industry transitions from tens of thousands of cells to hundreds of thousands, and eventually to millions. The appendix covers technical aspects of HetNets in the sections, “Heterogeneous Networks and Small Cells” and “Self-Organizing Networks.”

While promising in the long term, one immediate challenge in deploying a large number of small cells is backhaul, since access to fiber is not necessarily available and line-of-sight microwave links are not always feasible. Site acquisition and the need for multiple operators to deploy their own cells in a coverage area are additional challenges.[45] Figure 23 depicts the challenges.

---

[45] For further discussion of this topic, refer to 5G Americas and Small Cell Forum, *Small cell siting challenges,”* February 2017.

Figure 23: Small-Cell Challenges

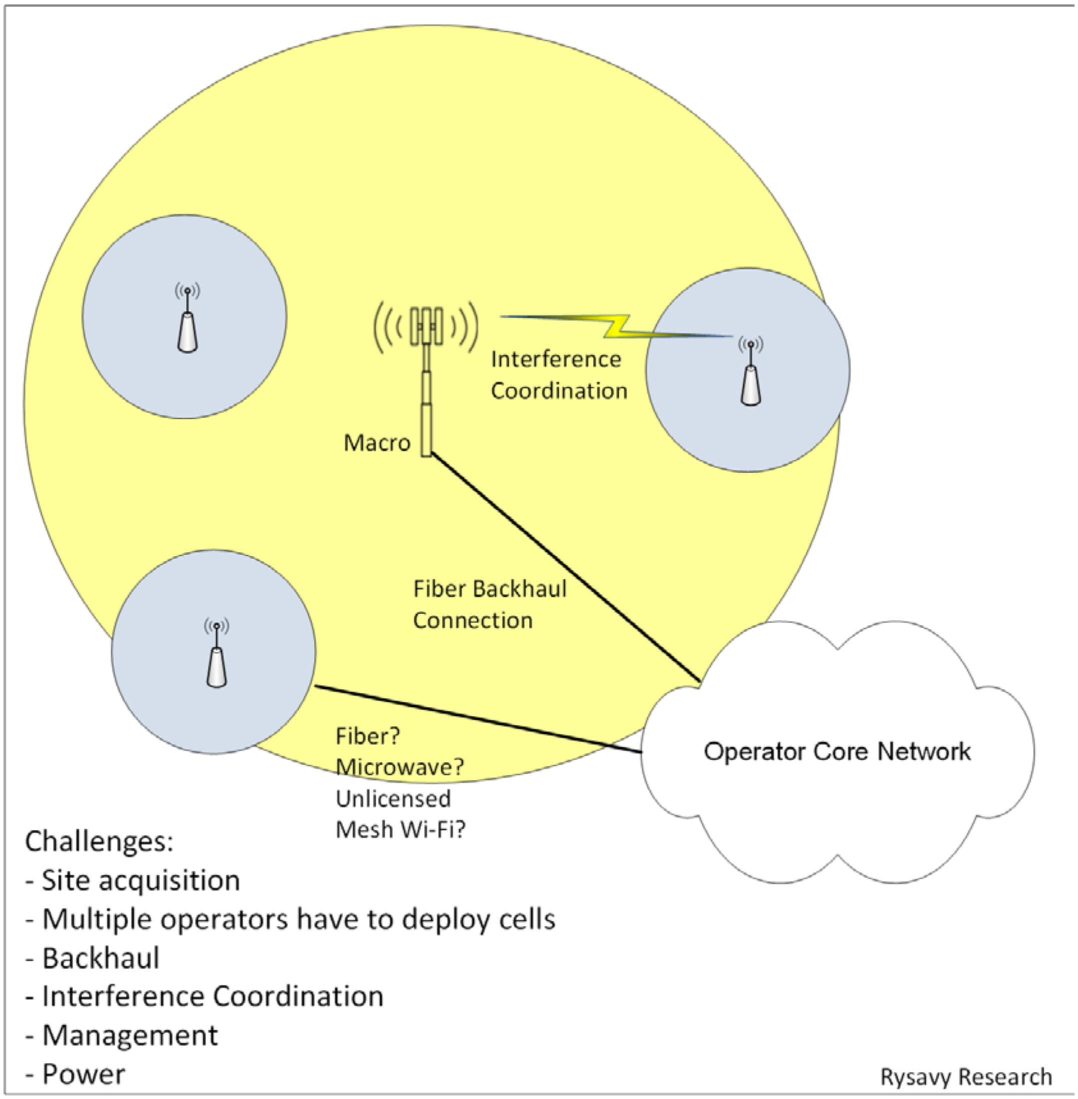
Macro
Interference
Coordination
Fiber Backhaul
Connection
Fiber?
Microwave?
Unlicensed
Mesh Wi-Fi?
Operator Core Network
Challenges:
- Site acquisition
- Multiple operators have to deploy cells
- Backhaul
- Interference Coordination
- Management
- Power
Rysavy Research

Despite these challenges and the relatively small number of small cells deployed today, many believe small-cell deployments will accelerate.[46]

In the United States, the FCC in April 2017 issued a Notice of Proposed Rulemaking, titled "Accelerating Wireless Broadband Deployment by Removing Barriers to Infrastructure

---

[46] For example, see Fierce Wireless, "Crown Castle expects its small cell business to double in next 2 years," April 25, 2017. Available at http://www.fiercewireless.com/wireless/crown-castle-expects-its-small-cell-business-to-double-next-two-years.

Investment," that seeks to streamline items such as site acquisition and consistent pricing.[47]

5G small-cell considerations include:

- ❑ Due to limited propagation at mmWave frequencies, 5G small-cell deployments will be dense and involve large numbers of sites. Inter-site distances (ISDs) will range from 100 to 300 meters in many deployments, with 200 meters a typical value.[48]
- ❑ The high-capacity of mmWave small cells will require multi-Gbps backhaul connections using an expected combination of fiber, mmWave radio in point-to-point connections, and 5G self-backhaul.
- ❑ The expected use of cloud RAN and centralized base station facilities will simplify equipment at the site, facilitating dense deployments.
- ❑ Dense deployments will motivate neutral-host (multi-tenant) approaches, but these are outside the scope of specification efforts.

The effective range of a mmWave small cell depends on multiple factors, including whether line-of-sight is available, extent of foliage, pole height, whether user equipment is indoors or outdoors, and the types of building materials the signal must pass through to reach indoor equipment.

Despite the challenges, small cells will ultimately contribute greatly to increased network capacity. Table 8 lists possible configurations. Note that many of these approaches can be combined, such as using picos and Wi-Fi offload.

Table 8: Small-Cell Approaches

Small-Cell Approach	Characteristics
Macro plus small cells in select areas.	Significant standards support. Femtocells or picocells can use the same radio carriers as macro (less total spectrum needed) or can use different radio carriers (greater total capacity).
Macro in licensed band plus LTE operation in unlicensed bands.	Promising approach for augmenting LTE capacity in scenarios where operator is deploying LTE small cells.[49] See discussion below in the section on unlicensed spectrum integration.

---

[47] FCC, *Wireless Infrastructure NPRM and NOI*, April 21, 2017. Available at https://www.fcc.gov/document/wireless-infrastructure-nprm-and-noi.

[48] 5G Americas member contributions.

[49] See Rysavy Research, *Accelerating Innovation in Unlicensed Spectrum,* Fierce Wireless, November 2016. Available at http://www.rysavy.com/Articles/2016-11-Innovation-Unlicensed-Spectrum.pdf. See also Rysavy Research, *Will LTE in Unlicensed Spectrum Unlock a Vast Store of Mobile Broadband Capacity?* MIMO World, June 2014. Available at http://www.mimoworld.com/?p=2377.

Small-Cell Approach	Characteristics
Macro (or small-cell) cellular in licensed band plus Wi-Fi.	Extensively used today with increased use anticipated. Particularly attractive for expanding capacity in coverage areas where Wi-Fi infrastructure exists but small cells with LTE do not. LTE Wi-Fi Aggregation (being specified in Release 13) is another approach, as are MP-TCP and MP-QUIC.
Wi-Fi only.	Low-cost approach for high-capacity mobile broadband coverage, but impossible to provide large-area continuous coverage without cellular component.

Neutral-Host Small Cells

Multi-operator and neutral-host solutions could accelerate deployment of small cells.[50] Currently, nearly all small-cell deployments are operator-specific, but in the future, deployments supporting multiple operators could reduce the cost per operator to provide coverage.

A candidate band for neutral-host small cells is 3.5 GHz, using LTE TDD and MulteFire as potential technologies. Wi-Fi technology also addresses neutral-host configurations at the access level, but it has roaming and authentication challenges. HotSpot 2.0 (covered in the appendix) addresses roaming and authentication.

Unlicensed Spectrum Integration

Unlicensed spectrum is becoming ever more important to mobile broadband networks. Initial use was rudimentary offload onto Wi-Fi networks, but now, Wi-Fi networks are becoming more tightly integrated into cellular networks. Efforts are also underway to use LTE in unlicensed spectrum.

Unlicensed spectrum adds to capacity in two ways. First, a large amount of spectrum (approximately 500 MHz) is available across the 2.4 GHz and 5 GHz bands, with the 3.5 GHz band adding further spectrum in the future. A significant amount of unlicensed spectrum is also available in mmWave bands, with 7 GHz already in use in the United States (57 to 64 GHz) and an additional 7 GHz in 5G spectrum allocations. Second, unlicensed spectrum is mostly used in small coverage areas, resulting in high-frequency re-use and much higher throughput rates per square meter of coverage versus typical cellular deployments.

The IEEE 802.11 family of technologies has experienced rapid growth, mainly in private deployments. The latest 802.11 standard, 802.11ac, offers peak theoretical throughputs in excess of 1 Gbps and improved range through use of higher-order MIMO. 802.11ac Wave 2 products include a Multi-user MIMO capability that further increases capacity and throughput. IEEE is developing a subsequent version, 802.11ax, that emphasizes capacity improvements as well as higher throughputs. In the mmWave frequencies, IEEE has developed 802.11ad, which operates at 60 GHz, and the standards body is currently working on a successor technology, 802.11ay.

---

[50] 5G Americas and Small Cell Forum, *Multi-operator and neutral host small cells; Drivers, architectures, planning and regulation,* December 2016. Report available at http://www.5gamericas.org/files/4914/8193/1104/SCF191_Multi-operator_neutral_host_small_cells.pdf.

Integration between mobile broadband and Wi-Fi networks can be either loose or tight. Loose integration means data traffic routes directly to the internet and minimizes traversal of the operator network. This is called “local breakout.” Tight integration means data traffic, or select portions thereof, may traverse the operator core network. An example is Wi-Fi calling, which uses IP Multimedia Subsystem.

Although offloading onto Wi-Fi can reduce traffic on the core network, the Wi-Fi network does not necessarily always have greater spare capacity than the cellular network. The goal of future integrated cellular/Wi-Fi networks is to intelligently load balance between the two. Simultaneous cellular/Wi-Fi connections will also become possible. For example, in Release 13, 3GPP introduces link aggregation of Wi-Fi and LTE through LWA and LWIP.

Successfully offloading data and providing users a good experience mandates measures such as automatically provisioning subscriber devices with the necessary Wi-Fi configuration options and automatically authenticating subscribers on supported public Wi-Fi networks. Many stakeholders are working toward tighter integration between Wi-Fi and cellular networks.

Another approach for using unlicensed spectrum employs LTE as the radio technology, initially in a version referred to as LTE-Unlicensed, which works with Releases 10-12 of LTE, as defined in the LTE-U Forum. In Release 13, 3GPP specified LAA, which implements listen-before-talk capability, a requirement for unlicensed operation in Europe and Japan. Initially, carrier aggregation combines a licensed carrier with an unlicensed 20 MHz carrier in the 5 GHz band as a supplemental channel, with the aggregation of multiple unlicensed channels an eventual possibility. Operating LTE in unlicensed bands could decrease the need for handoffs to Wi-Fi. Up to 32 unlicensed carriers (of 20 MHz each) can be aggregated to theoretically access 640 MHz of unlicensed spectrum. LAA may also be deployed in 3.5 GHz bands. Enhanced LAA (eLAA), specified in Release 14, adds uplink use of unlicensed spectrum. Carriers are now conducting LAA trials.[51]

A concern with using LTE in unlicensed bands is whether it will be a fair neighbor to Wi-Fi users. LTE-U based on Release 10-12 addresses this concern by selecting clear channels to use and measuring the channel activity of Wi-Fi users, then using an appropriate duty cycle for fair sharing. License-Assisted Access in Release 13 adds listen-before-talk (LBT) and implements other regulatory requirements that exist in some countries. 3GPP conducted a study and concluded that, “A majority of sources providing evaluation results showed at least one LBT scheme for LAA that does not impact Wi-Fi more than another Wi-Fi network.”[52]

To address co-existence, the cellular industry worked with the Wi-Fi Alliance in 2016 to develop a test plan for LTE-U. The testing goal was to verify that, in a laboratory environment, an LTE-U base station does not impact a Wi-Fi network any more than another Wi-Fi access point.[53]

---

[51] For example, see AT&T, “AT&T Reaches Wireless Speeds of More than 750 Mbps with LTE Licensed Assisted Access (LTE-LAA) Field Trials,” available at http://about.att.com/story/lte_licensed_assisted_access_field_trials.html.

[52] 3GPP, *Technical Specification Group Radio Access Network; Study on Licensed-Assisted Access to Unlicensed Spectrum; (Release 13). 36.889.* See section 9, “Conclusions.”

MulteFire, specified by the MulteFire Alliance, is an application of LTE in unlicensed bands that does not require an anchor in licensed spectrum, opening up the possibility of deployments by non-operator entities, including internet service providers, venue operators, and enterprises. Under a roaming arrangement with cellular operators, LTE customers could roam into MulteFire networks. Figure 24 shows the evolution of the different versions of LTE for unlicensed bands.

Figure 24: Timeline Relationship of LTE-U, LAA, eLAA, and MulteFire

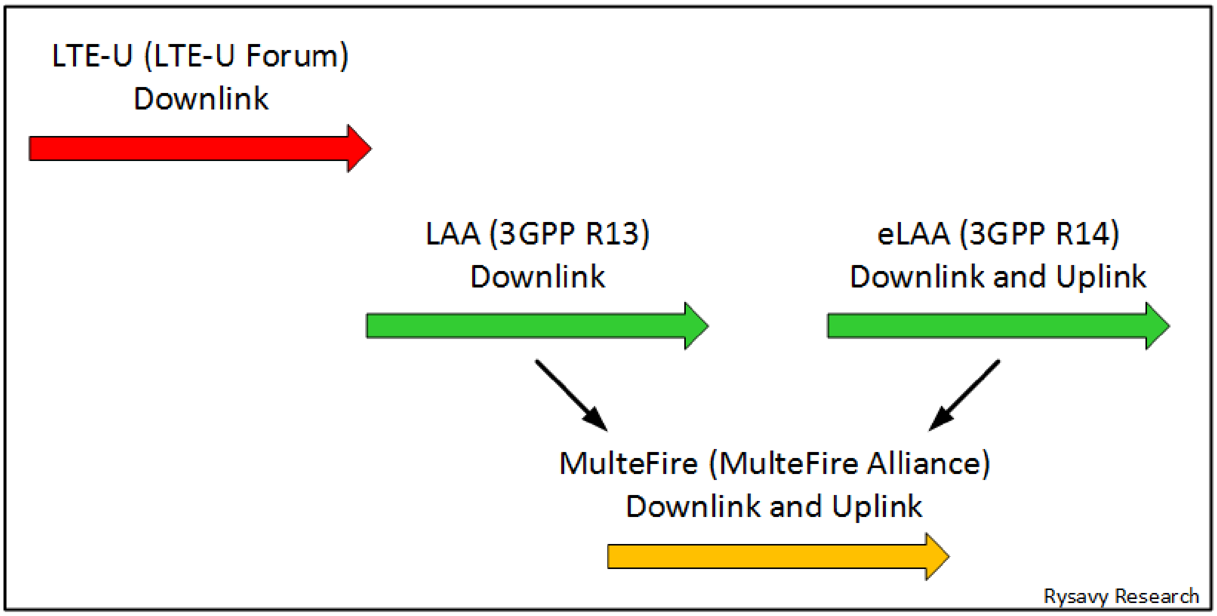

An alternative approach for integrating Wi-Fi is LWA. LTE handles the control plane, but connections occur over separate LTE base stations and Wi-Fi access points. LWA benefits operators that wish to emphasize Wi-Fi technology for harnessing capacity in unlicensed spectrum. LWIP is a variation of LWA that also integrates LTE and Wi-Fi, but by integrating at a higher level of the protocol stack (IP instead of PDCP), it facilitates use of existing Wi-Fi equipment and devices, with integration typically occurring at the eNodeB.

Figure 25 shows how the different technologies exploit licensed and unlicensed spectrum.

---

[53] See Wi-Fi Alliance, "Unlicensed Spectrum," http://www.wi-fi.org/discover-wi-fi/unlicensed-spectrum.

Figure 25: How Different Technologies Harness Spectrum

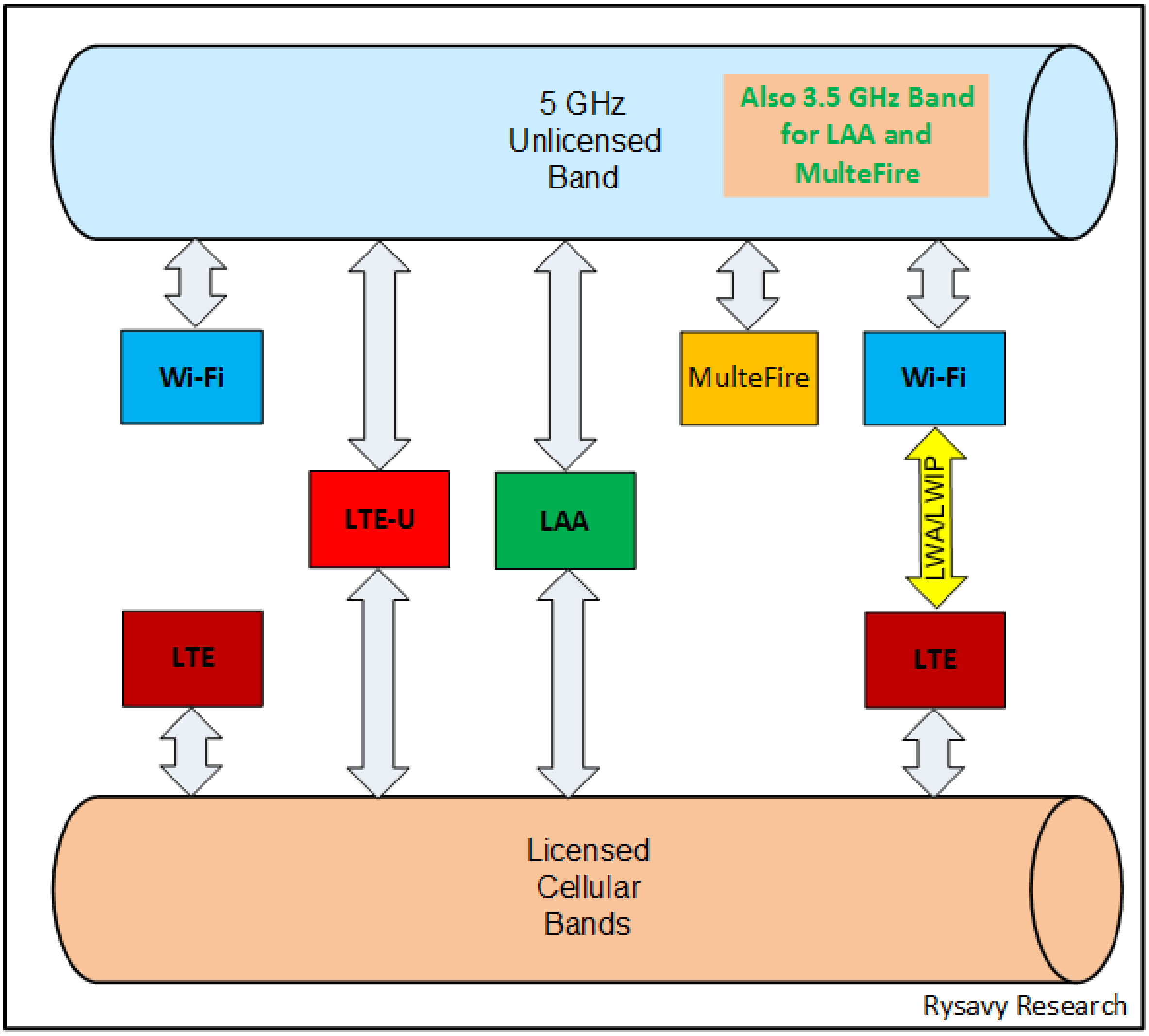
5 GHz Unlicensed Band
Also 3.5 GHz Band for LAA and MulteFire
Wi-Fi
MulteFire
Wi-Fi
LTE-U
LAA
LWA/LWIP
LTE
LTE
Licensed Cellular Bands
Rysavy Research

Table 9 summarizes the different uses of unlicensed spectrum for public mobile broadband networks.

Table 9: Approaches for Using Unlicensed Spectrum.

	Technology	Attributes
Wi-Fi	Ever-more-sophisticated means to integrate Wi-Fi in successive 3GPP Releases.	Combining Wi-Fi with cellular increases capacity.
Release 13 RAN Controlled LTE WLAN Interworking	Base station can instruct the UE to connect to a WLAN for offload.	Available in late 2017 or 2018 timeframe.

Release 10-12 LTE-U Based on LTE-U Forum Specifications	LTE-U Forum-specified approach for operating LTE in unlicensed spectrum.	Available in 2017. More seamless than Wi-Fi. Cannot be used in some regions (e.g., Europe, Japan).
Release 13 Licensed-Assisted Access	3GPP-specified approach for operating LTE in unlicensed spectrum. Downlink only.	Available in late 2017 or 2018 timeframe. Designed to address global regulatory requirements.
Release 14 Enhanced Licensed-Assisted Access	Addition of uplink operation.	Available in 2019 or 2020 timeframe.
5G Unlicensed Operation	Release 15 has a study item for approaches that Release 16 will standardize.	Potentially available in 2021.
MulteFire	Does not require a licensed anchor.	Potentially creates a neutral-host small cell solution.
LWA	Aggregation of LTE and Wi-Fi connections at PDCP layer.	Part of Release 13. Available in late 2017 or 2018 timeframe.
LWIP	Aggregation of LTE and Wi-Fi connections at IP layer.	Part of Release 13. Available in late 2017 or 2018 timeframe.

Cellular operators are currently emphasizing simple offload to Wi-Fi or LTE-U/LAA. Aggregation techniques, such as LWA and LWIP, do not currently have market traction.

Refer to the appendix section “Unlicensed Spectrum Integration” for further technical details.

Internet of Things and Machine-to-Machine

Machine-to-machine communications, now evolving into the Internet of Things, is a vast opportunity for wireless communications, with all 3GPP technologies potentially playing roles.

The lowest-cost cellular devices enabling M2M communications today are GPRS modems, which risk becoming obsolete as operators sunset their GSM systems. HSPA is also used for M2M communications, as is LTE, which has been optimized to efficiently communicate small bursts of information, making it particularly well suited for M2M.

Low-cost GSM (through Enhanced Coverage GSM IoT [EC-GSM-IoT]) and LTE modem options in 3GPP releases 10 through 13 reduce cost, improve communications range, and extend battery life. See the appendix section “Internet of Things and Machine-to-Machine” for details.

3GPP had a Release 13 study item on how LTE technologies can operate for vehicle communications, including vehicle-to-vehicle and vehicle-to-infrastructure, leveraging device-to-device communications capabilities already specified for LTE in Releases 12 and 13.[54] This work is now being standardized in Release 14.

Developers will use 3GPP wireless technologies for many IoT applications. In other instances, developers will use local area technologies, such as Wi-Fi, Bluetooth Low Energy, and ZigBee. New Low-Power Wide-Area (LPWA) wireless technologies emerging specifically to support IoT include Ingenu, LoRa, and Sigfox. The low-power operation of some of these technologies, including LTE, will permit battery operation over multiple years. Table 10 summarizes the various technologies.

Table 10: Wireless Networks for IoT

Technology	Coverage	Characteristics	Standardization / Specifications
GSM/GPRS/EC-GSM-IoT	Wide area. Huge global coverage.	Lowest-cost cellular modems, risk of network sunsets. Low-throughput.	3GPP
HSPA	Wide area. Huge global coverage.	Low-cost cellular modems. Higher power, high throughput.	3GPP
LTE, NB-IoT	Wide area. Increasing global coverage.	Wide area, expanding coverage, cost/power reductions in successive 3GPP releases. Low to high throughput options.	3GPP
Wi-Fi	Local area.	High throughput, higher power.	IEEE
ZigBee	Local area.	Low throughput, low power.	IEEE
Bluetooth Low Energy	Personal area.	Low throughput, low power.	Bluetooth Special Interest Group
LoRa	Wide area. Emerging deployments.	Low throughput, low power. Unlicensed bands (sub 1 GHz, such as 900 MHz in the U.S.)	LoRa Alliance[55]
Sigfox	Wide area. Emerging deployments.	Low throughput, low power. Unlicensed bands (sub 1 GHz such as 900 MHz in the U.S.)	Sigfox[56]

---

[54] 3GPP, *3GPP TR 36.885, Technical Specification Group Radio Access Network; Study on LTE-based V2X Services; (Release 14).*

[55] For details, see LoRa Alliance, https://www.lora-alliance.org/.

[56] For details, see Sigfox, https://www.sigfox.com/en.

Technology	Coverage	Characteristics	Standardization / Specifications
Ingenu (previously OnRamp Wireless)	Wide area. Emerging deployments.	Low throughput, low power. Using 2.4 GHz ISM band. Uses IEEE 802.15.4k.	Ingenu[57]
Weightless	Wide area. Expected deployments.	Low throughput, low power. Unlicensed bands (sub 1 GHz such as TV White-Space and 900 MHz in the U.S.)	Weightless Special Interest Group[58]

Security is of particular concern to both developers and users of IoT technology. An increasing amount of network-connected infrastructure will result in new security vulnerabilities that are being addressed by concerted effort from the industry.[59]

Smart Antennas and MIMO

Smart antennas, defined with progressively greater capabilities in successive 3GPP releases, provide significant gains in throughput and capacity. By employing multiple antennas at the base station and the subscriber unit, the technology either exploits signals traveling through multiple paths in the environment or does beam steering, in which multiple antennas coordinate their transmissions to focus radio energy in a particular direction.

Initial low-band LTE deployments used 2X2 MIMO on the downlink (two base station transmit antennas, two mobiles receive antennas) and 1X2 on the uplink (one mobile transmit antenna, two base station receive antennas). In the higher bands, 2X2 downlink MIMO has been deployed, but it is more common to employ four antennas for uplink reception in a 1X4 configuration. LTE deployments are now using 4X2 MIMO and 4X4 MIMO on the downlink (four base station transmit antennas). LTE specifications encompass higher-order configurations, such as 4X4 MIMO, 8X2 MIMO, and MU-MIMO on the downlink and 1X4 on the uplink. Practical considerations, such as antenna sizes that are proportional to wavelength, dictate MIMO options for different bands.

Engineers are now experimenting with what are called massive MIMO systems, which employ a far larger number of antenna elements at the base station—64, 128, and eventually even more. Use in 5G of cmWave and mmWave bands, with their short wavelengths, will facilitate massive MIMO, but even before then, 3GPP is developing specifications for massive MIMO for 4G systems in what it calls full-dimension MIMO (FD-MIMO). Release 14 specifies configurations with up to 32 antennas at the base station.

---

[57] For details, see Ingenu, https://www.ingenu.com/.

[58] For details, see http://www.weightless.org/.

[59] For further insight, refer to the Ericsson white paper, *IoT Security,* February 2017, available at https://www.ericsson.com/assets/local/publications/white-papers/wp-iot-security-february-2017.pdf.

Massive MIMO is practical even in cellular frequencies. For example, at 2.5 GHz, an 8X8 array using half wavelength spacing would produce a form factor of 50 cm X 50 cm. Applications of such arrays include beamforming along a horizontal direction as well as beamforming in a vertical direction, such as to serve different levels of high-rise buildings.

See the appendix section “LTE Smart Antennas” and “LTE-Advanced Antenna Technologies” for further details.

Virtualization

Virtualization refers to implementing the functions of infrastructure nodes in software on commercial “off-the-shelf” computing equipment. The approach promises lower capital expenditures, lower operating costs, faster deployment of new services, energy savings, and improved network efficiency. With NFV, multiple tenants will be able to share the same infrastructure, facilitating, for example, mobile virtual network operator (MVNO) and multi-operator virtualized RAN arrangements. NFV, however, also constitutes an entirely new way of building and managing networks, so widespread adoption will occur over a long period.

Both the core network and portions of the radio-access network can be virtualized. The core network, consisting of fewer nodes, is an easier starting point. Virtualizing RAN elements, although more complex, could eventually provide the greatest network efficiency gains, particularly for small-cell deployments where it can facilitate coordination among cells and use of methods such as CoMP and interference coordination. Unlike the core, virtualizing the entire RAN is not possible because a Physical Network Function must terminate the radio interface. As operators virtualize their core networks, they put in place the systems and know-how to extend virtualization to the RAN.

The European Telecommunications Standards Institute (ETSI) is standardizing a framework, including interfaces and reference architectures for virtualization. Other standards and industry groups involved include 3GPP, the Open Networking Foundation, OpenStack, OpenDaylight, and OPNFV.

Figure 26 shows the ETSI framework, in which virtualized network functions are the nodes or applications by which operators build services.

Figure 26: ETSI NFV High-Level Framework

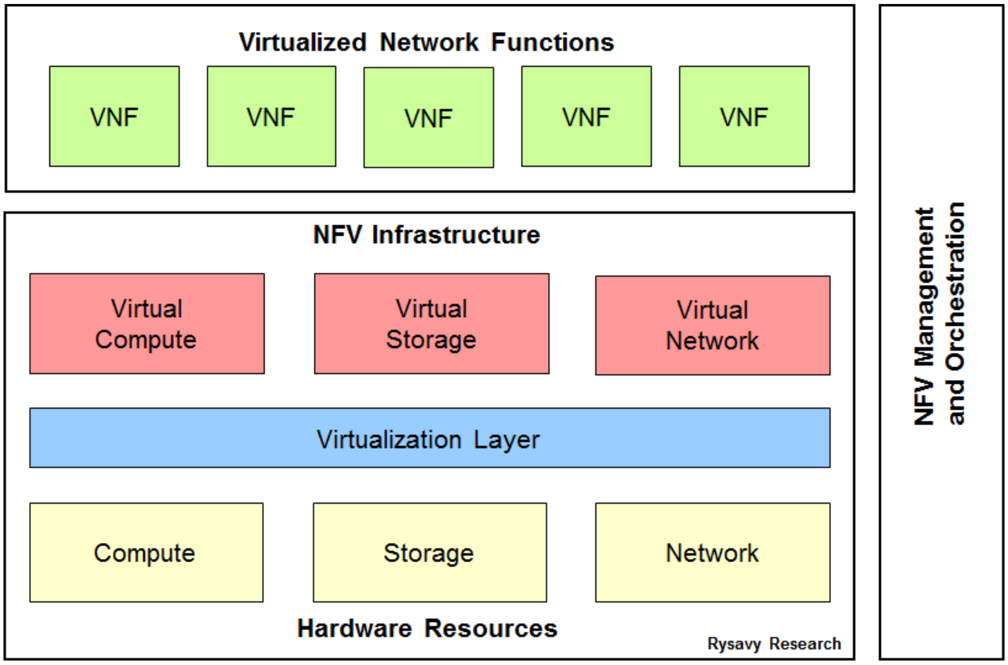
Virtualized Network Functions
VNF
VNF
VNF
VNF
VNF
NFV Infrastructure
Virtual Compute
Virtual Storage
Virtual Network
Virtualization Layer
Compute
Storage
Network
Hardware Resources
Rysavy Research
NFV Management and Orchestration

Some specific use cases for NFV include:

- **5G.** 5G networks will likely be fully virtualized.
- **IMS and VoLTE.** IMS is necessary for VoLTE, but an NFV approach could reduce the complexity associated with the multiple nodes and interfaces in the IMS architecture.
- **Virtualized EPC (VEPC).** The Evolved Packet Core, consisting of the Serving Gateway (SGW), the Packet Gateway (PGW), and Mobile Management Entity (MME), can be virtualized, but doing so will require meeting operator bandwidth, latency, and control plane service requirements.
- **New VEPC Services.** With a virtualized EPC, an operator can more easily create MVNO services, each with its own virtualized MME, SGW, and PGW. An M2M virtualized service is another example of offering a more finely tuned service for the target application. Because the PGW connects to external networks, further opportunities exist for virtualized services to augment networking functions, including video caching, video optimization, parental controls, ad insertion, and firewalls.
- **Cloud RAN.** Pooling of baseband processing in a cloud RAN can, but does not necessarily, use virtualization techniques. Separating the radio function from baseband processing typically requires transporting digitized radio signals across high-bandwidth (multi-Gbps) fiber connections, sometimes referred to as

fronthauling. Refer to the appendix section “Cloud Radio-Access Network (RAN) and Network Virtualization” for a more detailed technical discussion.[60]

Because of higher investment demands, RAN virtualization will take longer to deploy than core network virtualization and likely will occur selectively for small-cell deployments.

For additional details, refer to the 5G Americas white paper, *Bringing Network Function Virtualization to LTE*.[61]

Multi-Access Edge Computing

ETSI is standardizing Multi-access Edge Computing, previously known as Mobile-Edge Computing, a technology that empowers a programmable application environment at the edge of the network, within the RAN.[62] Goals include reduced latency, more efficient network operation for certain applications, and an improved user experience. Although MEC emphasizes 5G, it can also be applied to 4G LTE networks.

Applications that will benefit are ones that require server-side processing but are location specific. Examples include:

- ☐ Augmented reality.
- ☐ Intelligent video processing, such as transcoding, caching, and acceleration.
- ☐ Connected cars.
- ☐ Premise-based IoT gateways.

Multicast and Broadcast

Another important new service is video streaming via multicast or broadcast functions. 3GPP has defined multicast/broadcast capabilities for both HSPA and LTE. Mobile TV services have experienced little business success so far, but broadcasting uses the radio resource much more efficiently than having separate point-to-point streams for each user. For example, users at a sporting event might enjoy watching replays on their smartphones. The technology supports these applications; it is a matter of operators and content providers finding appealing applications. The appendix covers technical aspects in more detail.

Remote SIM Provisioning

The GSM Association (GSMA) is developing specifications that make it possible for consumers to purchase unprovisioned devices, select the operator of their choice and then download the subscriber identity module (SIM) application into the device.[63] This capability benefits devices such as watches, health bands, health monitors, and other small connected items. Initial products may be available in 2017.

---

[60] For further details, see “Network Functions Virtualisation,” http://www.etsi.org/technologies-clusters/technologies/nfv. Viewed May 17, 2017.

[61] Available at http://www.4gamericas.org/files/1014/1653/1309/4G_Americas_-_NFV_to_LTE_-_November_2014_-_FINAL.pdf.

[62] For further details, see ETSI, “Multi-access Edge Computing,” http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing. Accessed May 17, 2017.

[63] For details, see GSMA, “A New SIM,” available at https://www.gsma.com/rsp/, viewed June 8, 2017.

VoLTE, RCS, WebRTC, and Wi-Fi Calling

Voice has evolved from a separate circuit-switched service in 2G and 3G networks to a packet-switched service in 4G LTE networks that can integrate with other services and applications, such as messaging and video calling. Elements that make these capabilities possible include the quality-of-service mechanisms in LTE, the IMS platform discussed above, implementation of Rich Communications Suite, compliance with GSMA IR.92 guidelines, and optional support for WebRTC.

Voice Support and VoLTE

While 2G and 3G technologies were deployed from the beginning with both voice and data capabilities, LTE networks can be deployed with or without voice support. Moreover, there are two methods available: circuit-switched fallback (CSFB) to 2G/3G and VoIP. Most operators deploying LTE used CSFB initially but are now migrating to VoIP methods with VoLTE that uses IMS. Initial VoLTE deployments occurred in 2012. Because VoLTE needs new software in phones, the transition from circuit-switched voice to VoLTE on a large scale will occur over a number of years as users upgrade their devices.

For the time being, 3GPP operators with UMTS/HSPA networks will continue to use circuit-switched voice for their 3G connections.

Using VoLTE, operators can offer high-definition (HD) voice using the new Adaptive Multi-Rate Wideband (AMR-WB) voice codec. HD voice not only improves voice clarity and intelligibility, it suppresses background noise. AMR-WB extends audio bandwidth to 50-7000 Hz compared with the narrowband codec that provides audio bandwidth of 80-3700 Hz. HD voice will initially function only between callers on the same network. 3GPP has also developed a new voice codec, called “Enhanced Voice Services” (EVS), which will be the successor to AMR and AMR-WB codecs.

Other advantages of LTE’s packetized voice include being able to combine it with other services, such as video calling and presence; half the call set-up time of a 3G connection; and high voice spectral efficiency. With VoLTE’s HD voice quality, lower delay, and higher capacity, operators can compete against OTT VoIP providers. Due to traffic prioritization, VoLTE voice quality remains high even under heavy loads that cause OTT-voice service to deteriorate.

Applications based on WebRTC will also increasingly carry voice sessions. See the section “VoLTE and RCS” in the appendix for more details on LTE voice support.

Rich Communications Suite

An initiative called “Rich Communications Suite” (RCS), supported by many operators and vendors, builds on IMS technology to provide a consistent feature set as well as implementation guidelines, use cases, and reference implementations. RCS uses existing standards and specifications from 3GPP, Open Mobile Alliance (OMA), and GSMA and enables interoperability of supported features across operators that support the suite. RCS supports both circuit-switched and packet-switched voice and can interoperate with LTE packet voice.

Core features include:

- A user capability exchange or service discovery with which users can know the capabilities of other users.

- Enhanced (IP-based) messaging (supporting text, IM, and multimedia) with chat and messaging history.
- Enriched calls that include multimedia content (such as photo or video sharing) during voice calls. This could become the primary way operators offer video calling.

The primary drivers for RCS adoption are the ability to deploy VoLTE in a well-defined manner and to support messaging in the IP domain. RCS addresses the market trend of users moving away from traditional text-based messaging and provides a platform for operator-based services that compete with OTT messaging applications. Figure 27 shows the evolution of RCS capability, including the addition of such features as messaging across multiple devices, video calling, video sharing, and synchronized contact information across multiple devices.

Figure 27: Evolution of RCS Capability.[64]

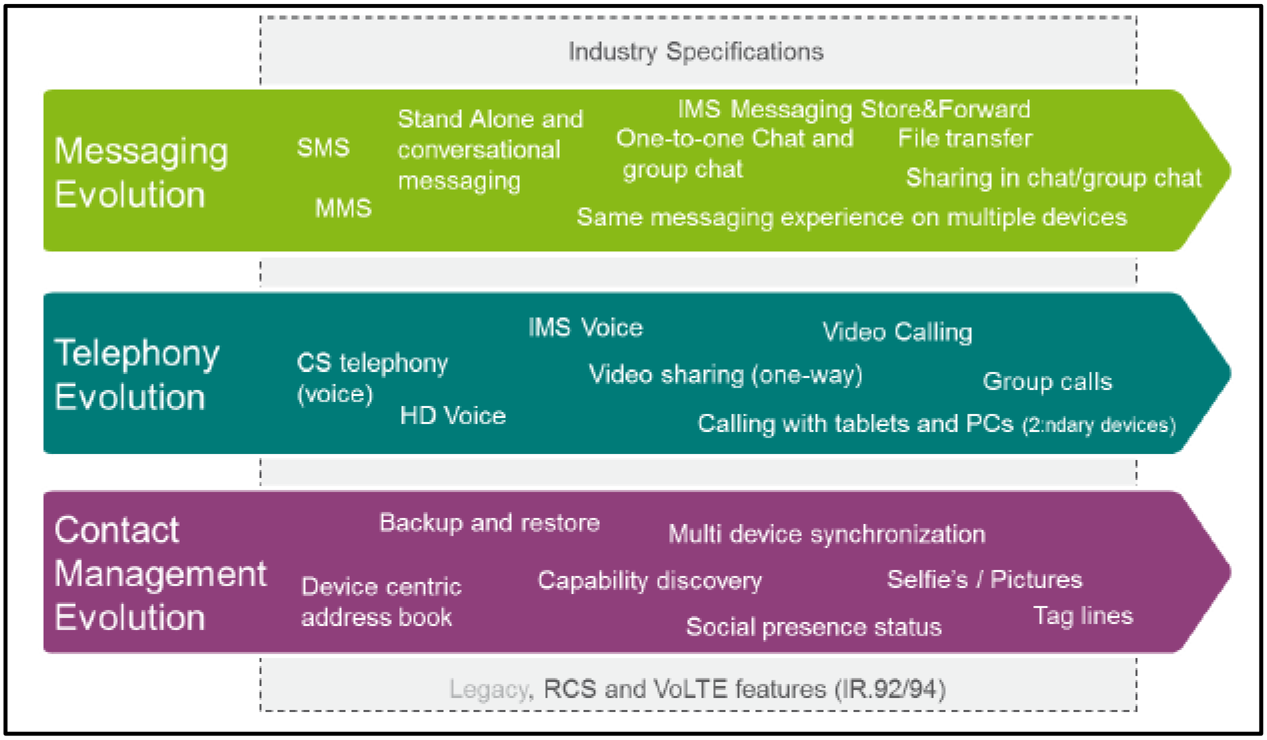
Industry Specifications
Messaging Evolution
SMS
MMS
Stand Alone and conversational messaging
IMS Messaging Store&Forward
One-to-one Chat and group chat
File transfer
Sharing in chat/group chat
Same messaging experience on multiple devices
Telephony Evolution
CS telephony (voice)
HD Voice
IMS Voice
Video sharing (one-way)
Video Calling
Group calls
Calling with tablets and PCs (2:ndary devices)
Contact Management Evolution
Device centric address book
Backup and restore
Capability discovery
Multi device synchronization
Social presence status
Selfie's / Pictures
Tag lines
Legacy, RCS and VoLTE features (IR.92/94)

WebRTC

WebRTC is an open project supported by Google, Mozilla, and Opera within the Internet Engineering Taskforce (IETF) that enables real-time communications in Web browsers via JavaScript APIs. 3GPP Release 12 specifications define how WebRTC clients can access IMS services, including packet voice and video communication. WebRTC operating over IMS gains the additional benefit of seamless transition across transport networks, for example, LTE to Wi-Fi.

Operators can integrate WebRTC with RTC, facilitating development of vertical applications such as telemedicine and customer service. WebRTC and RCS are more complementary than competitive. Both, through application interfaces, can provide access to underlying network functions.

---

[64] 4G Americas, *VoLTE and RCS Technology - Evolution and Ecosystem*, Nov. 2014.

Wi-Fi Calling

Another advantage of the VoLTE/IMS/RCS architecture is that it is agnostic to the user connection, meaning voice and video service can extend to Wi-Fi connections as easily as LTE connections. Wi-Fi calling can be advantageous in coverage areas were the Wi-Fi signal has better quality than an LTE signal. For video calling, use of Wi-Fi will also reduce data consumption over the cellular connection. By implementing a standards-based approach, as opposed to OTT-voice approaches, called parties see the same phone number regardless of network and can reach the subscriber using that phone number.

Previous technical approaches, such as Generic Access Network (GAN, initially called Unlicensed Mobile Access [UMA]), did not include as robust a handover mechanism as is provided by VoLTE/IMS.

For the best-quality voice in a Wi-Fi network, the device and Wi-Fi network should implement Wi-Fi Multimedia (WMM), which gives voice packets higher priority than other data traffic. WMM is especially necessary in congested networks. In addition, the Access Network Discovery and Selection Function (ANDSF) and cellular-WLAN enhancement features in 3GPP Release 12 have policies for enabling voice handover between LTE and Wi-Fi.

Roaming with Wi-Fi calling will need to address whether the visited network's IMS infrastructure handles the Wi-Fi call or whether the home network's IMS does.

Public Safety

An important LTE application is for public safety, initially as a broadband data service and eventually for mission-critical voice service. Current public safety networks use technologies, such as Terrestrial Trunked Radio (TETRA) in Europe and Project 25 (P25) in the United States, that provide mission-critical voice but only narrowband data.

In the United States, the government has made 20 MHz of spectrum available at 700 MHz in Band 14 and created the First Responder Network Authority (FirstNet), an independent authority within the National Telecommunications and Information Administration (NTIA) to provide a nationwide public-safety broadband network. AT&T will build and deploy the network.[65]

Another country driving the use of LTE for public safety is the United Kingdom, where the UK Home Office has a program for the Emergency Service Network.

Using LTE for public safety is a complex undertaking because public-safety needs differ from those of consumers. Addressing these needs requires both different features, which 3GPP is incorporating in multiple releases of LTE specifications, and different network deployment approaches. Public safety also has different device-level needs than consumers.

LTE Features for Public Safety

Some broadband applications for public safety can use standard LTE capability. For example, sending email, accessing a database, or streaming a video may not require any special features. Other applications, however, require new capabilities from 3GPP standards, including:

Group Communication

Available in Release 12, the Group Communication Service (GCS) application server, using one-to-one (unicast) and one-to-many communications (broadcast), will be able to send voice, video, or data traffic to multiple public-safety devices. The broadcast mode will employ eMBMS to use radio resources efficiently, but if coverage is weak, a unicast approach may deliver data more reliably. The system will be able to dynamically switch between broadcast and unicast modes. Release 14 adds single-cell point-to-multipoint transmission.

Proximity-Based Services (Device-to-Device)

With proximity-based services, defined in Release 12, user devices can communicate directly, a capability that benefits both consumers and public safety. This type of communication is called sidelink communication. Consumer devices can find other devices only with assistance from the network, but for public safety, devices will be able to communicate directly with other devices independently of the network.

With Release 13, devices will be able to act as relays for out-of-coverage devices, such as inside a building.

---

[65] For details, see "FirstNet AT&T," https://www.corp.att.com/public-safety/att-firstnet/. See also RCR Wireless, *Editorial Report: Public Safety LTE*, March 2017. Available at http://content.rcrwireless.com/20170322_Public_Safety_Report.

The appendix section “Proximity Services (Device-to-Device)” discusses this feature in greater detail.

Mission-Critical Push-to-Talk

MCPTT, being defined in Release 13, provides one-to-one and one-to-many push-to-talk communications services. With this feature, available in the 2018 timeframe, public-safety organizations will be able to consider retiring legacy voice-based systems.

Mission-Critical Video over LTE and Mission-Critical Data over LTE

Release 14 adds Mission-Critical Video over LTE and Mission-Critical Data over LTE, designed to work with Mission-Critical Push-to-Talk, giving first responders more communications options.

Prioritization

To prevent interference with public-safety operations in emergency situations experiencing high load, the network can prioritize at multiple levels. First, the network can bar consumer devices from attempting to access the network, thus reducing signaling load. Second, the network can prioritize radio resources, giving public-safety users higher priority. Third, using a new capability called “Multimedia Priority Service” (MPS), the network can prioritize a connection between an emergency worker and a regular subscriber. Finally, the network can assign specific quality-of-service (QoS) parameters to specific traffic flows, including guaranteed bit rate. 3GPP has defined specific QoS quality-class identifiers for public safety.

High Power

Release 11 defines higher power devices for the public safety band that can operate at 1.25 Watts, improving coverage and reducing network deployment costs.

Isolated operation

With Release 13, a base station can continue offering service even with the loss of backhaul, a capability that will benefit public-safety personnel in disaster situations.

Relays

Figure 28 summarizes the more than eighteen features in 3GPP relays that apply to public safety.

Figure 28: Summary of 3GPP LTE Features to Support Public Safety[66]

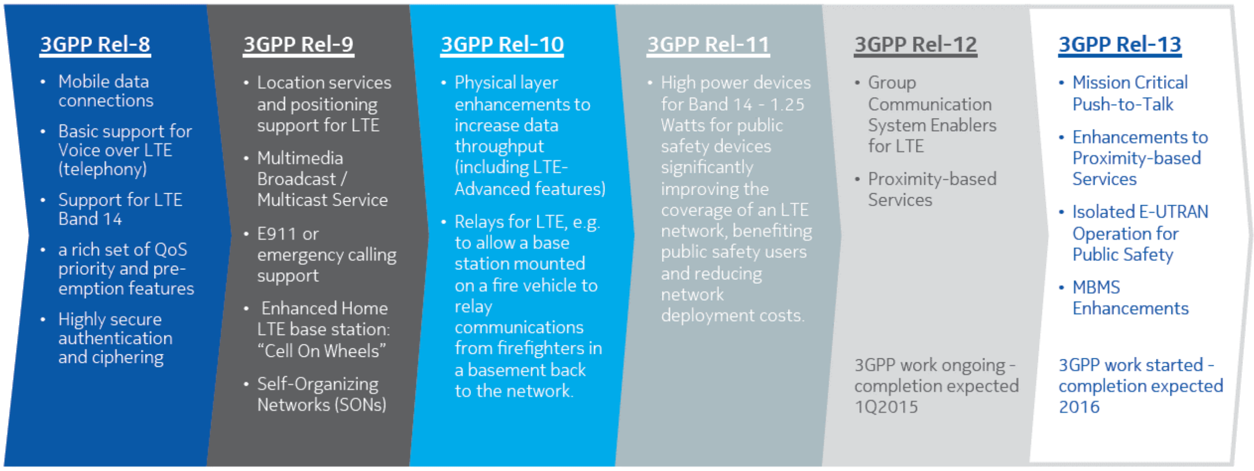

3GPP Rel-8

- Mobile data connections
- Basic support for Voice over LTE (telephony)
- Support for LTE Band 14
- a rich set of QoS priority and pre-emption features
- Highly secure authentication and ciphering

3GPP Rel-9

- Location services and positioning support for LTE
- Multimedia Broadcast / Multicast Service
- E911 or emergency calling support
- Enhanced Home LTE base station: "Cell On Wheels"
- Self-Organizing Networks (SONs)

3GPP Rel-10

- Physical layer enhancements to increase data throughput (including LTE-Advanced features)
- Relays for LTE, e.g. to allow a base station mounted on a fire vehicle to relay communications from firefighters in a basement back to the network.

3GPP Rel-11

- High power devices for Band 14 - 1.25 Watts for public safety devices significantly improving the coverage of an LTE network, benefiting public safety users and reducing network deployment costs.

3GPP Rel-12

- Group Communication System Enablers for LTE
- Proximity-based Services

3GPP work ongoing - completion expected 1Q2015

3GPP Rel-13

- Mission Critical Push-to-Talk
- Enhancements to Proximity-based Services
- Isolated E-UTRAN Operation for Public Safety
- MBMS Enhancements

3GPP work started - completion expected 2016

Deployment Approaches

Because huge infrastructure investments would be required for a network dedicated solely to public safety, industry and government are evaluating approaches in which public-safety uses can leverage existing commercial network deployments. One caveat is that public-safety networks have more stringent resilience and security requirements than commercial networks.

Shared Network

As depicted in Figure 29, multiple sharing approaches are possible:

1. In this scenario, a public-safety entity owns and operates the entire network, an approach that gives public-safety organizations the greatest control over the network but at the highest cost.

2. A commercial operator shares its radio-access network, including cell sites and backhaul, but the public-safety entity manages core network functions including gateways, the Mobile Management Entity, the Home Subscriber Server (HSS), and public-safety application servers. Because the radio-access network is the costliest part of the network, this approach significantly reduces the amount public safety must invest in the network. Even though the RAN is shared, public safety still can use its dedicated spectrum.

3. In an MVNO approach, the operator shares its cell sites and backhaul as well as some core network functions, such as the MME and Serving Gateway. Public safety manages a small number of network functions, such as the Packet Gateway, HSS, and its application servers.

---

[66] Nokia, *LTE networks for public safety services*, 2014. Available at http://networks.nokia.com/sites/default/files/document/nokia_lte_for_public_safety_white_paper.pdf.

4. A final approach, not shown in the figure, is one in which the mobile operator hosts all of the elements shown in the figure and public safety manages only its application servers.

Figure 29: Sharing Approaches for Public-Safety Networks

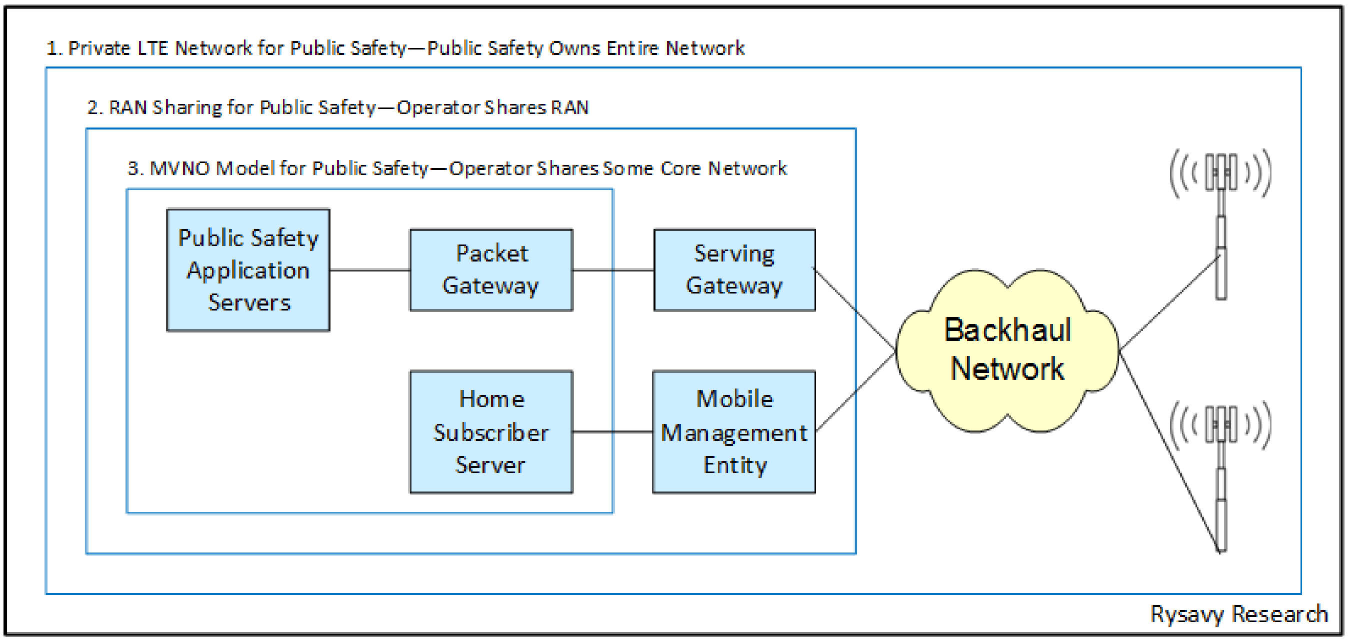
1. Private LTE Network for Public Safety—Public Safety Owns Entire Network
2. RAN Sharing for Public Safety—Operator Shares RAN
3. MVNO Model for Public Safety—Operator Shares Some Core Network
Public Safety Application Servers
Packet Gateway
Serving Gateway
Home Subscriber Server
Mobile Management Entity
Backhaul Network
Rysavy Research

Resilience

Public safety may need greater resilience than found in commercial networks, including hardware redundancy, geographic redundancy, load balancing, fast re-routing in IP networks, interface protection, outage detection, self-healing, and automatic reconfiguration.

Security

Public-safety networks may have higher security requirements than commercial networks, including physical security of data centers, core sites, and cell sites. Whereas LTE networks do not have to encrypt traffic in backhaul and core networks, public-safety applications may choose to encrypt all IP traffic using virtual private networking approaches.

Coverage

A number of approaches can ensure the broadest possible coverage for public-safety networks. First, public-safety frequencies at 700 MHz already propagate and penetrate well. Next, public-safety devices will be able to transmit at higher power. In addition, base stations can employ four-way receiver diversity and higher-order sectorization. For disaster situations, public safety can also use rapidly deployable small cells, such as on trailers. Finally, proximity-based services operating in a relay mode, as discussed above, can extend coverage.

Device Considerations for Public Safety

Public-safety devices will have unique requirements, including guaranteed network access under all conditions and guidelines for how devices are shared among users.

Access to Commercial Networks

Public-safety devices could be designed to also communicate on commercial LTE networks, providing an alternative communications avenue when the device cannot connect to a public-safety network. Subscriptions to all major commercial networks would make this approach the most effective. Wi-Fi capability further extends this concept.

Device Sharing

Because public-safety devices may be shared among personnel, user profiles cannot be stored on USIM cards stored in the devices. Bluetooth-based remote SIMs are one approach to address this problem.

Expanding Capacity

Wireless technology is playing a profound role in networking and communications, even though wireline technologies such as fiber have inherent capacity advantages.

Over time, wireless networks will gain substantial additional capacity through the methods discussed in the next section. While they will compete with copper twisted pair and coax, they will never catch up to fiber. The infrared frequencies used in fiber-optic communications have far greater bandwidth than radio. As a result, one fiber-optic strand has greater bandwidth than the entire usable radio spectrum to 100 GHz, as illustrated in Figure 30.[67]

Figure 30: RF Capacity vs. Fiber-Optic Cable Capacity

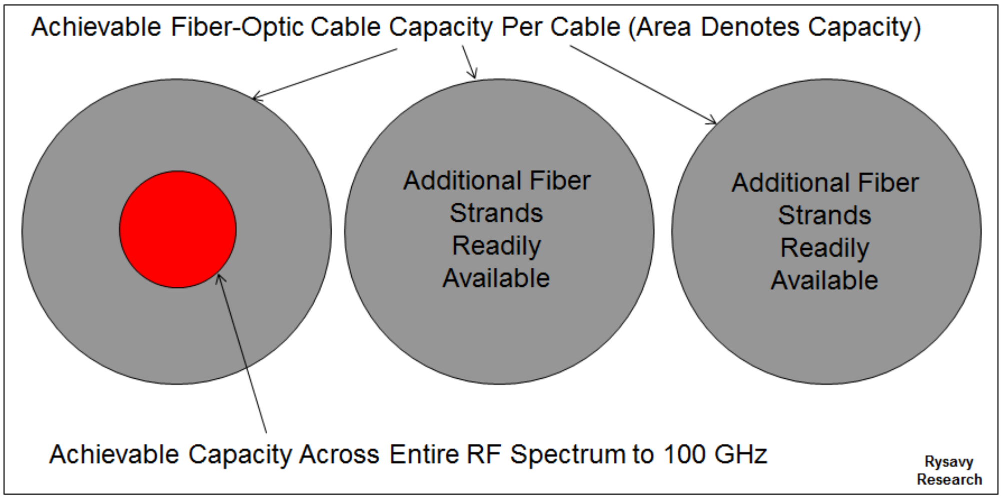
Achievable Fiber-Optic Cable Capacity Per Cable (Area Denotes Capacity)
Additional Fiber Strands Readily Available
Additional Fiber Strands Readily Available
Achievable Capacity Across Entire RF Spectrum to 100 GHz
Rysavy Research

A dilemma of 4G mobile broadband is that it *can* provide a broadband experience similar to wireline, but it *cannot* do so for all subscribers in a coverage area at the same time. Hence, operators must carefully manage capacity, demand, policies, pricing plans, and user expectations. Similarly, application developers must become more conscious of the inherent constraints of wireless networks. 5G, with its far greater capacity, at least in small deployments, will increase the percentage of users that use wireless connections as their only broadband.

Three factors determine wireless network capacity, as shown in Figure 31: the amount of spectrum, the spectral efficiency of the technology, and the size of the cell. Because smaller cells serve fewer people in each cell and because there are more of them, small cells are a major contributor to increased capacity.

---

[67] One fiber-optic cable can transmit over 10,000 Gbps compared with all wireless spectrum to 100 GHz, which, even at an extremely high spectral efficiency 10 bps/Hz, would have only 1,000 Gbps of capacity.

Figure 31: Dimensions of Capacity

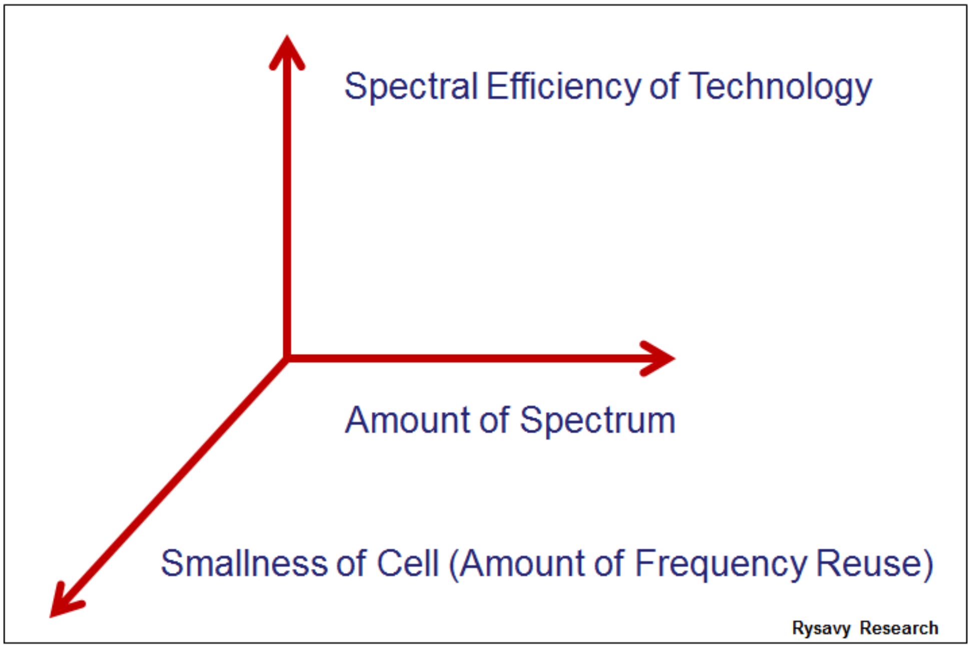
Spectral Efficiency of Technology
Amount of Spectrum
Smallness of Cell (Amount of Frequency Reuse)
Rysavy Research

Given the relentless growth in usage, mobile operators are combining multiple approaches to increase capacity and managing congestion:

- **More spectrum.** Spectrum correlates almost directly to capacity, and more spectrum is becoming available globally for mobile broadband. In the U.S. market, the 2010 FCC National Broadband Plan is in the process of adding an additional 500 MHz of spectrum by 2020. mmWave band spectrum for 5G will provide far more spectrum, but propagation characteristics will restrict its use to small cells. Multiple papers by Rysavy Research and others[68] argue the critical need for additional spectrum.
- **Unpaired spectrum.** LTE TDD operates in unpaired spectrum. In addition, technologies such as HSPA+ and LTE permit the use of different amounts of spectrum between downlink and uplink. Additional unpaired downlink spectrum can be combined with paired spectrum to increase capacity and user throughputs.
- **Supplemental downlink.** With downlink traffic five to ten times greater than uplink traffic, operators often need to expand downlink capacity rather than uplink capacity. Using carrier aggregation, operators can augment downlink capacity by combining separate radio channels.
- **Spectrum sharing.** Policy makers are evaluating how spectrum might be shared between government and commercial entities. Although a potentially promising approach for the long term, sharing raises complex issues, as discussed further in the section "Spectrum Developments."
- **Increased spectral efficiency.** Newer technologies are spectrally more efficient, meaning greater aggregate throughput using the same amount of spectrum. Wireless technologies such as LTE, however, are reaching the theoretical limits of spectral

---

[68] See multiple papers on spectrum and capacity at http://www.rysavy.com/writing.

efficiency, and future gains will be quite modest, allowing for a possible doubling of LTE efficiency over currently deployed versions. See the section "Spectral Efficiency" for a further discussion.

- **Smart antennas.** Through higher-order MIMO and beamforming, smart antennas gain added sophistication in each 3GPP release and are the primary contributor to increased spectral efficiency (bps/Hz). Massive MIMO, beginning in Release 13, will support 16-antenna-element systems and in 5G, will expand to potentially hundreds of antenna elements.
- **Uplink gains combined with downlink carrier aggregation.** Operators can increase network capacity by applying new receive technologies at the base station (for example, large-scale antenna systems such as massive MIMO) that do not necessarily require standards support. Combined with carrier aggregation on the downlink, these receive technologies produce a high-capacity balanced network, suggesting that regulators should in some cases consider licensing just downlink spectrum.
- **Small cells and heterogeneous networks.** Selective addition of picocells to macrocells to address localized demand can significantly boost overall capacity, with a linear increase in capacity relative to the number of small cells. HetNets, which also can include femtocells, hold the promise of increasing capacity gains by a factor of four and even higher with the introduction of interference cancellation in devices. Distributed antenna systems (DAS), used principally for improved indoor coverage, can also function like small cells and increase capacity. Actual gain will depend on a number of factors, including number and placement of small cells,[69] user distribution, and any small-cell selection bias that might be applied.
- **Offload to unlicensed spectrum.** Using unlicensed spectrum with Wi-Fi or LTE operation in unlicensed spectrum offers another means of offloading heavy traffic. Unlicensed spectrum favors smaller coverage areas because interference can be better managed, so spectral re-use is high, resulting in significant capacity gains.
- **Higher level sectorization.** For some base stations, despite the more complex configuration involved, six sectors can prove advantageous versus the more traditional three sectors, deployed either in a 6X1 horizontal plane or 3X2 vertical plane.[70]

Strategies to manage demand include:

- **Quality of service (QoS) management.** Through prioritization, certain traffic, such as non-time-critical downloads, could occur with lower priority, thus not affecting other active users. Current network neutrality rules, however, may constrain use of traffic prioritization.[71]
- **Off-peak hours.** Operators could offer user incentives or perhaps fewer restrictions on large data transfers during off-peak hours.

---

[69] With small-cell range expansion using a large selection bias, small cells can be distributed uniformly.

[70] An example of vertical layering would be a 3X1 layer at ground level and a separate 3X1 layer for higher levels of surrounding buildings.

[71] For a discussion of this issue, see Rysavy Research, *LTE Congestion Management – Enabling Innovation and Improving the Consumer Experience*, January 2015. Available at http://www.rysavy.com/Articles/2015-01-Rysavy-LTE-Congestion-Management.pdf.

Given a goal of increasing capacity by a factor of 1,000, 50X could roughly be achieved through network densification; 10X through more spectrum, including higher frequencies such as mmWave; and 2X by increases in spectral efficiency.

Based on historical increases in the availability of new spectrum, technologies delivering better spectral efficiency, and increases in the number of cell sites, Rysavy Research has calculated that, over the last thirty-year period, aggregate network capacity has doubled every three years. Rysavy Research expects this trend to continue into the future.

Rysavy Research Analysis:

Aggregate Wireless Network Capacity
Doubles Every Three Years.

Spectrum Developments

Licensed spectrum scarcity continues to challenge the industry. Tactics to make the best use of this limited resource include deploying technologies that have higher spectral efficiency; adapting specifications to enable operation of UMTS-HSPA and LTE in all available bands; designing both FDD and TDD versions of technology to take advantage of both paired and unpaired bands; designing carrier aggregation techniques in HSPA+ and LTE-Advanced Pro; and deploying as many new cells, large and small, as is economically and technically feasible. Although all of these industry initiatives greatly expand capacity, they do not obviate the need for additional spectrum. 5G technology will be able to employ frequencies not previously used in cellular systems, spanning 6 GHz to 100 GHz.

An important aspect of UMTS-HSPA and LTE deployment is for infrastructure and mobile devices to accommodate the expanding number of available radio bands. The fundamental system design and networking protocols remain the same for each band; only the frequency-dependent portions of the radios must change. As other frequency bands become available for deployment, standards bodies adapt UMTS-HSPA and LTE for these bands as well.

3GPP has specified LTE for operation in many different bands, and initial use will be more fragmented than the four bands (850 MHz, 900 MHz, 1.8 GHz, 1.9 GHz) that enable global roaming on 2G and the additional two bands (1.7 GHz and 2.1 GHz) that enable 3G roaming. In the Americas, LTE roaming may occur in the 1.7/2.1 GHz (AWS) bands, and globally, LTE roaming may occur in the 1.8 and 2.6 GHz bands. Longer term, operators will re-farm spectrum used for 2G and 3G and apply it to LTE. Unfortunately, the process of identifying new spectrum and making it available for the industry is a lengthy one, as shown in Figure 32.

Figure 32: Spectrum Acquisition Time[72]

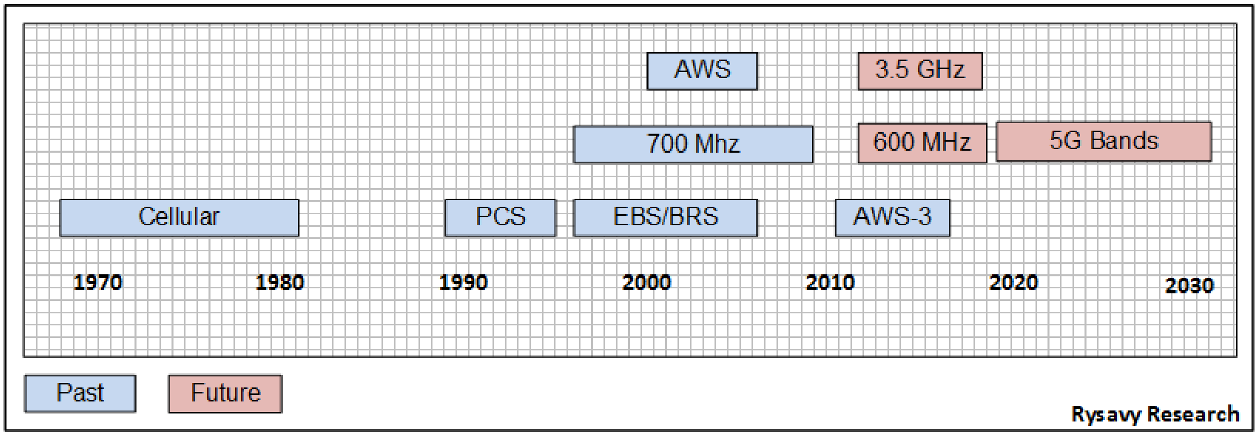

New short-term spectrum opportunities in the United States include the “small-cell” band from 3550 to 3700 MHz and 5G spectrum.

---

[72] Source for historical data, FCC, *National Broadband Plan*, Chapter 5. Available at http://www.broadband.gov/plan/5-spectrum/, accessed May 18, 2017. Future based on Rysavy Research analysis.

Table 11 summarizes current and future spectrum allocations in the United States.[73]

Table 11: United States Current and Future Licensed Spectrum Allocations

Frequency Band	Amount of Spectrum	Comments
600 MHz	70 MHz	Ultra-High-Frequency (UHF).
700 MHz	70 MHz	Ultra-High Frequency (UHF).
850 MHz	64 MHz	Cellular and Specialized Mobile Radio.
1.7/2.1 GHz	90 MHz	Advanced Wireless Services (AWS)-1.
1695-1710 MHz, 1755 to 1780 MHz, 2155 to 2180 MHz	65 MHz	AWS-3. Uses spectrum sharing.
1.9 GHz	140 MHz	Personal Communications Service (PCS).
2000 to 2020, 2180 to 2200 MHz	40 MHz	AWS-4 (Previously Mobile Satellite Service).[74]
2.3 GHz	20 MHz	Wireless Communications Service (WCS).
2.5 GHz	194 MHz	Broadband Radio Service. Closer to 160 MHz deployable.
	FUTURE	
3.55 to 3.70 GHz	150 MHz	Small-cell band with spectrum sharing and unlicensed use.
Above 6 GHz	Multi GHz	Anticipated for 5G systems beginning in 2017. Based on wavelengths, 3 GHz to 30 GHz is referred to as the cmWave band and 30 GHz to 300 GHz is referred to as the mmWave band.

Today’s licensed spectrum networks operate most efficiently and are deployed most cost-effectively using a combination of low-band spectrum, below 1 GHz, for coverage and 1 GHz to 3 GHz for capacity. As technology improves, bands in 3 GHz to 100 GHz, and eventually higher, will supplement capacity.

---

[73] For international allocations, refer to Wik-Consult, Study for the European Commission, *Inventory and review of spectrum use: Assessment of the EU potential for improving spectrum efficiency*, September 2012. Available at http://ec.europa.eu/digital-agenda/sites/digital-agenda/files/cion_spectrum_inventory_executive_summary_en.pdf.

[74] Supported in 3GPP Band 70, which adds 1995-2000 MHz, pairing it with 1695-1710 MHz in AWS-3 band.

The subsections below provide additional information about the recently completed AWS-3 auction, incentive auctions, the 3.5 GHz, 5G, spectrum harmonization, unlicensed spectrum, and spectrum sharing.

AWS-3

In early 2015, the FCC received close to $45 billion for the U.S. Treasury in the AWS-3 auction, more than twice the amount of any previous auction, demonstrating the value of mid-band spectrum.[75] The auction adds 65 MHz of desirable spectrum to the mobile-broadband industry. The plan is to employ spectrum sharing among commercial networks and select government systems. Eventually, most of these government systems will migrate to another spectrum. 3GPP has specified use of both AWS-3 and AWS-4 spectrum in what it refers to as "Band 66."

This band is asymmetrical, with a downlink of 90 MHz and an uplink of 70 MHz. An operator can use the upper-most 20 MHz of the downlink only with carrier aggregation.

Broadcast Incentive Auction (600 MHz)

The broadcast incentive auction completed in 2017 will reallocate 84 MHz of UHF channels in the 600 MHz band that are currently used by TV broadcasters, with 70 MHz of licensed spectrum and 14 MHz of unlicensed spectrum. The auction was more complicated than past spectrum auctions, when the FCC simply reassigned or designated spectrum for commercial mobile use and then conducted an auction.

In the first stage, the FCC conducted a reverse auction to determine how much spectrum broadcasters might wish to relinquish in exchange for how much compensation. In the second stage, mobile operators bid for spectrum in a forward auction, similar to past spectrum auctions.

Figure 33 shows the final band plan.

Figure 33: 600 MHz Band Plan[76]

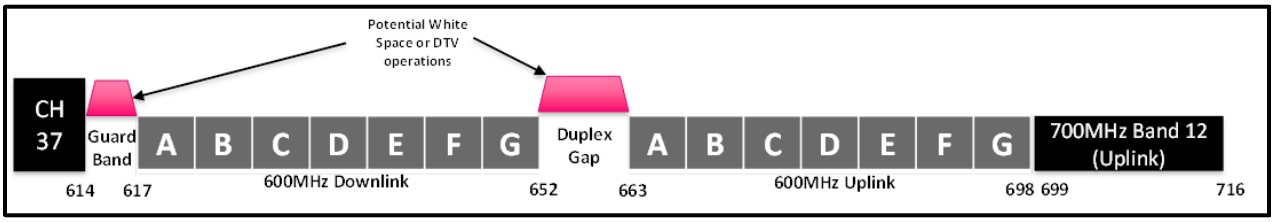

Part of the auction process reorganizes and repacks relinquished channels, as well as channels needed for broadcasters that want to keep broadcasting, to make useful blocks of spectrum for mobile broadband. The FCC's goal is to design an auction that will result in a uniform nationwide band plan, but in some markets, there may be deviations from that plan.

---

[75] For further details, see Rysavy Research, "Latest FCC auction shatters spectrum myths," January 2015. Available at https://gigaom.com/2015/01/17/latest-fcc-auction-shatters-spectrum-myths/.

[76] 5G Americas member contribution.

With a 39-month schedule for winning bidders to move into their new spectrum, the 600 MHz band will be fully available by 2020. However, some operators, will begin using this spectrum in advance of this date, possibly as soon as the second half of 2017, as the spectrum becomes available for use.[77] Operators will use the band for LTE and 5G.

3550 to 3700 MHz "Small-Cell" Band

In the United States, the FCC is in the process of opening the 3550 to 3700 MHz CBRS band. The best use of this band will be small cells and backhaul. The FCC is implementing a three-tier model with incumbent access, priority access with priority access licenses (PALs), and General Authorized Access (GAA) for unlicensed users.[78] Incumbent access will include government radar systems.

Two industry organizations, the Wireless Innovation Forum[79] and the CBRS Alliance,[80] are working for the realization of 3.5 GHz systems.

The mobile broadband industry will mostly use LTE TDD in small-cell configurations, in either a licensed mode or an unlicensed mode using LAA or MulteFire.

Although this band represents a significant amount of new spectrum, potential adopters of the band have expressed concern about proposed rules for use of the band. For example:

- ☐ IEEE has stated it will not pursue a version of 802.11 for this band because exclusion zones are so large that the percentage of the U.S. population that could be served is below what is required for a successful market.[81]
- ☐ Licensed users will only be able to obtain three-year licenses with no automatic renewal.[82]

Assuming LAA auctions in early 2018, initial 3.5 GHz deployments could occur in 2018. Because an Environmental Sensing Capability (ESC) has not yet been developed and approved, these initial deployments may not come to fruition in coastal areas where the ESC is needed for co-existence with military systems using those frequencies.

---

[77] For example, see FierceWireless, "T-Mobile confirms speedy 600 MHz rollout for 2017, covering 1 million+ square miles," August 8, 2017, available at http://www.fiercewireless.com/node/334911.

[78] For further details, see Official FCC Blog, "Innovation in the 3.5 GHz Band: Creating a New Citizens Broadband Radio Service," March 2015, available at http://www.fcc.gov/blog/innovation-35-ghz-band-creating-new-citizens-broadband-radio-service. See also FCC, *Further Notice of Proposed Rulemaking--Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550- 3650 MHz Band*, April 23, 2014.

[79] See http://www.wirelessinnovation.org/.

[80] See https://www.cbrsalliance.org/.

[81] IEEE 802.11, *Comments of IEEE 802.11, In the Matter of Amendment of the Commission's Rules with Regard to Commercial Operations in the 3550-3650 MHz Band,"* Jul, 2015. Available at http://apps.fcc.gov/ecfs/document/view?id=60001115064.

[82] See Rysavy article about this band, "Scary Experimentation at 3.5 GHz," Jun. 2016, available at http://www.rysavy.com/Articles/2016-06-Scary-Experimentation-3-5-GHz.pdf. See also AT&T ex-parte FCC communication, http://apps.fcc.gov/ecfs/document/view?id=60001569272.

See the section “Spectrum Sharing” for further details of this band.

2.5 GHz Band

In the United States, the primary operator at 2.5 GHz is Sprint. In 2017, Sprint announced that they will deploy 5G in its 2.5 GHz spectrum as soon as the second half of 2019.[83] Having an average of 160 MHz of spectrum at 2.5 GHz in its top 100 markets, Sprint is in the process of deploying LTE-Advanced. In addition, Sprint is in an LTE deployment mode that includes the following features:

- ☐ Carrier aggregation capability between 2.5 GHz and lower bands (800MHz and 1900 MHz) and aggregating three carriers in a single band.
- ☐ TDD operation at 2.5 GHz and FDD operation in lower bands.
- ☐ 8 transmit and 8 receive radios at the base station that enables 4X2 MIMO in combination with beamforming.[84]

5G Bands

As radio technology progresses, it can handle higher frequencies, and it occupies greater bandwidth. 1G systems used 30 kHz radio carriers, 2G in GSM uses 200 kHz carriers, 3G in UMTS uses 5 MHz carriers, and 4G in LTE uses carriers of up to 640 MHz through carrier aggregation.

3GPP is specifying 5G NR to be band agnostic. 5G will use low-band, mid-band, and high-band spectrum, and 3GPP is in the process of identifying candidate bands for 5G across all frequency ranges. 3GPP Technical Services Group - Radio Access Networks (TSG-RAN) agreed to a process of efficiently adding LTE/NR band combinations and carrier-aggregated NR/NR band combinations. Initial bands being considered by the 3GPP specification process include low band bands 5, 20, 28 and 71, mid-band bands 1, 3, 7, 66 and 41, and key mmWave bands. Just as it has done with LTE, over time, 3GPP will specify additional 5G bands spanning multiple frequencies.

Although 5G research and development is in its infancy, to achieve the 20 Gbps or higher throughput rates envisioned for 5G will require radio carriers such as 200 MHz or 400 MHz, bandwidths available only at frequencies above 5 GHz. Researchers globally are studying high-frequency spectrum options including both cmWave frequencies (3 GHz to 30 GHz) and mmWave (30 GHz to 300 GHz). Ten times as much spectrum, or more, could be available in these higher frequencies than in all current cellular spectrum.

---

[83] For more details, see Sprint press release, “Qualcomm, SoftBank and Sprint Announce Collaboration on 2.5 GHz 5G,” May 10, 2017, available at http://newsroom.sprint.com/qualcomm-softbank-and-sprint-announce-collaboration-on-25-ghz-5g.htm.

[84] For more details, see Sprint article at: http://newsroom.sprint.com/blogs/sprint-perspectives/introducing-the-sprint-lte-plus-network--faster-stronger-more-reliable-than-ever-before.htm

5G NR is being designed to operate in any frequency band. These bands encompass 600 MHz, current cellular bands, 3.5 GHz, and higher bands such as mmWave. Because 5G will be the first cellular technology to be able to operate in mmWave bands, industry and government are working to make such bands available for 5G.

During the World Radiocommunication Conference (WRC) 15, the ITU proposed a set of global frequencies for 5G[85], which it intends to finalize at the next conference in 2019 (WRC 19):

- ☐ 24.25–27.5GHz
- ☐ 31.8–33.4GHz
- ☐ 37–40.5GHz
- ☐ 40.5–42.5GHz
- ☐ 45.5–50.2GHz
- ☐ 50.4–52.6GHz
- ☐ 66–76GHz
- ☐ 81–86GHz
- ☐ In 2014, the FCC published a Notice of Inquiry into use of spectrum bands above 24 GHz for Mobile Radio Services,[86] followed by a Notice of Proposed Rulemaking in October 2015. [87]

The FCC issued adopted rules to identify and open up 5G mmWave spectrum allocation in July 2016 that identify 3.85 GHz of licensed spectrum and 7 GHz of unlicensed spectrum: licensed use in 28 GHz, 37 GHz, and 39 GHz bands; unlicensed use in 64-71 GHz; and shared access in the 37-37.6 GHz band.

On July 24, 2016, the FCC adopted a report and order and further notice of proposed rulemaking that creates a new Part 30 for rules governing 28 GHz, 36 GHz, and 39 GHz bands, summarized in Table 12.

Table 12: United States 5G mmWave Bands[88]

Bands	Details
28 GHz Band (27.5-28.35 GHz)	Currently licensed for Local Multipoint Distribution Service (LMDS). Could be licensed using either 200 or 400 MHz wide channels.

---

[85] 5G Americas Webcast, “LTE-Steps to 5G,” Feb 12, 2016.

[86] FCC, *Notice of Inquiry, Use of Spectrum Bands above 24 GHz for Mobile Radio Services*, Oct. 17, 2014.

[87] FCC, *Notice of Proposed Rulemaking, Use of Spectrum Bands Above 24 GHz for Mobile Radio Services, GN Docket No. 14-177,* Oct 23, 2015.

[88] For more details, refer to FCC, *Report and Order and Further Notice of Proposed Rulemaking, Use of Spectrum Bands Above 24 GHz for Mobile Radio Services*, July 14, 2016. See also 5G Americas, *5G*

Bands	**Details**
39 GHz Band (38.6-40 GHz)	Currently licensed for fixed microwave in 50 MHz channels. Segment will be auctioned in 200 MHz blocks.
37 GHz Band (37-38.6 GHz)	Lower 37-37.6 GHz segment will be shared between federal and non-federal users. Upper 37.6-38.6 GHz segment will be auctioned in 200 MHz blocks.
64-71 GHz Band	Available for unlicensed use with same Part 15 rules as existing 57-64 GHz band.

In the 28 GHz band, satellite operations are secondary, but these operations are co-primary in the 37/39 GHz bands. Spectrum sharing may be required in some 5G bands, including 38.6 to 40 GHz, such as with fixed satellite service.

In the further notice of proposed rulemaking, the FCC is investigating additional bands, including 24 GHz, 32 GHz, 42 GHz, 47 GHz, 50 GHz, 70/80 GHz, and bands above 95 GHz.

The complex ITU harmonization process may mean that some regions, or even countries, pursue 5G bands that are not globally harmonized. For example, U.S. operators, along with operators in Korea and Japan, are planning 5G trials in the 28 GHz band, even though it is not one of the ITU bands.

Harmonization

Spectrum harmonization delivers many benefits, including higher economies of scale, better battery life, improved roaming, and reduced interference along borders.

As regulators make more spectrum available, it is important that they follow guidelines such as those espoused by 5G Americas:[89]

- ☐ Configure licenses with wider bandwidths.
- ☐ Group like services together.
- ☐ Be mindful of global technology standards.
- ☐ Pursue harmonized/contiguous spectrum allocations.
- ☐ Exhaust exclusive use options before pursuing shared use.
- ☐ Because not all spectrum is fungible, align allocation with demand.

Emerging technologies such as LTE benefit from wider radio channels. These wider channels are not only spectrally more efficient, they also offer greater capacity. Figure 34 shows increasing LTE spectral efficiency obtained with wider radio channels, with 20 MHz on the

---

Spectrum Recommendations, April 2017, available at http://www.5gamericas.org/files/9114/9324/1786/5GA_5G_Spectrum_Recommendations_2017_FINAL.pdf. This report covers global 5G spectrum developments.

[89] 4G Americas, *Sustaining the Mobile Miracle – A 4G Americas Blueprint for Securing Mobile Broadband Spectrum in this Decade*, March 2011.

downlink and 20 MHz (20+20 MHz) on the uplink comprising the most efficient configuration.

Figure 34: LTE Spectral Efficiency as Function of Radio Channel Size[90]

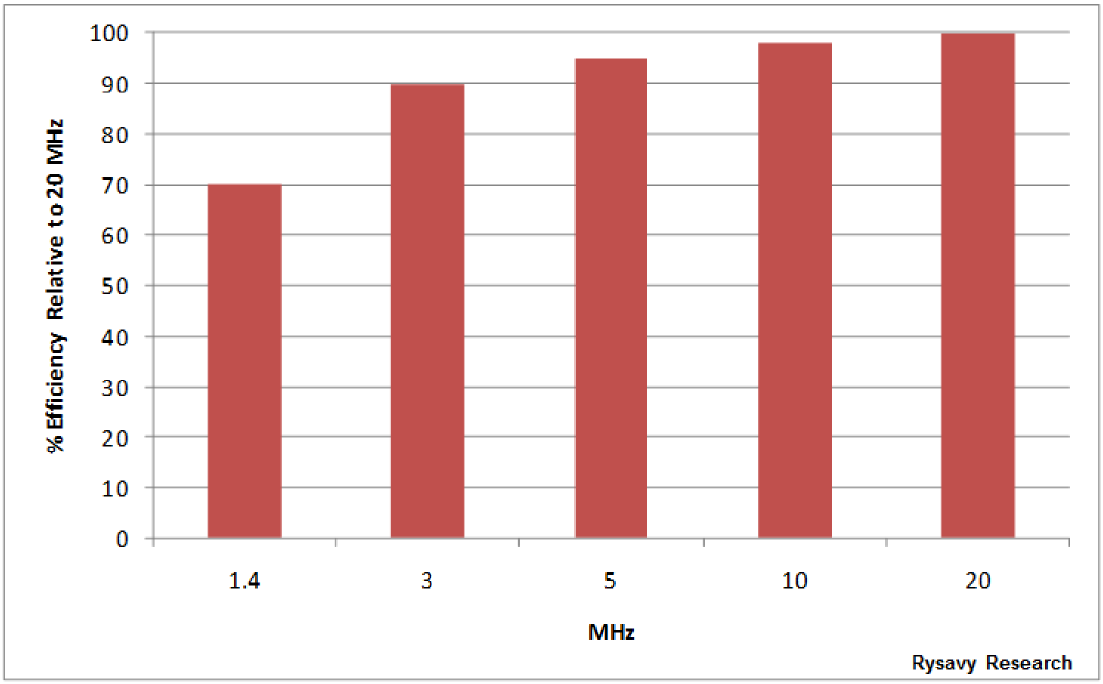

The organization tasked with global spectrum harmonization, the International Telecommunication Union, periodically holds World Radiocommunication Conferences.[91]

Harmonization occurs at multiple levels:

- ❑ Allocation of radio frequencies to a mobile service in the ITU frequency allocation table.
- ❑ Establishment of global or regional frequency arrangements, including channel blocks and specific duplexing modes.
- ❑ Development of detailed technical specifications and standards, including system performance, RF performance, and coexistence with other systems in neighboring bands.
- ❑ Assignment for frequency blocks with associated technical conditions and specifications to appropriate operators and service providers.[92]

---

[90] 5G Americas member company analysis.

[91] International Telecommunication Union, "World Radiocommunication Conferences (WRC)," http://www.itu.int/ITU-R/index.asp?category=conferences&rlink=wrc&lang=en, viewed May 18, 2017.

[92] International Telecommunication Union Radiocommunication Study Groups, *Technical Perspective on Benefits Of Spectrum Harmonization for Mobile Services and IMT*, Document 5D/246-E, January 2013.

Unlicensed Spectrum

Wi-Fi uses spectrum efficiently because its small coverage areas result in high-frequency reuse and high data density (bps per square meter). Less efficient are white-space unlicensed networks, sometimes called "super Wi-Fi," that, because of large coverage areas, have much lower throughput per square meter. While white-space networks may be a practical broadband solution in rural or undeveloped areas, they face significant challenges in urban areas that already have mobile and fixed broadband available.[93] See the section on "White Space Networks" in the appendix for further details.

Advocates argue that unlicensed spectrum unleashes innovation and that government should allocate greater amounts of unlicensed spectrum. Although Wi-Fi has been successful, the core elements that make unlicensed spectrum extremely successful are also the source of inherent disadvantages: local coverage and its unlicensed status. Local coverage enables high data density and high frequency reuse but makes widespread continuous coverage almost impossible. Similarly, unlicensed operation facilitates deployment by millions of entities but results in overlapping coverage and interference.

Networks built using unlicensed spectrum cannot replace networks built using licensed spectrum, and vice versa. The two are complementary and helpful to each other, as summarized in Table 13.[94]

Table 13: Pros and Cons of Unlicensed and Licensed Spectrum

Unlicensed Spectrum		Licensed Spectrum	
Pros	**Cons**	**Pros**	**Cons**
Easy and quick to deploy	Potential of other entities using same frequencies	Huge coverage areas	Expensive infrastructure
Low-cost hardware	Difficult to impossible to provide wide-scale coverage	Able to manage quality of service	Each operator has access to only a small amount of spectrum

Some operators offer a "Wi-Fi first" capability with which devices always attempt to use a Wi-Fi connection and fall back to a cellular connection only if no Wi-Fi is available. Such cellular backup is essential because Wi-Fi, due to low-power operation in many bands, is inherently unsuited for providing continuous coverage. The sharp drop-off in signal strength due to low transmit power makes coverage gaps over large areas inevitable, especially outdoors.

---

[93] For further analysis, see Rysavy Research, "White spaces networks are not 'super' nor even Wi-Fi," Gigaom, March 2013. Available at http://gigaom.com/2013/03/17/white-spaces-networks-are-not-super-nor-even-wi-fi/.

[94] For further analysis, see Rysavy Research, "It's Time for a Rational Perspective on Wi-Fi," Gigaom, April, 2014. Available at http://gigaom.com/2014/04/27/its-time-for-a-rational-perspective-on-wi-fi/.

Figure 35: Propagation Losses of Cellular vs. Wi-Fi[95]

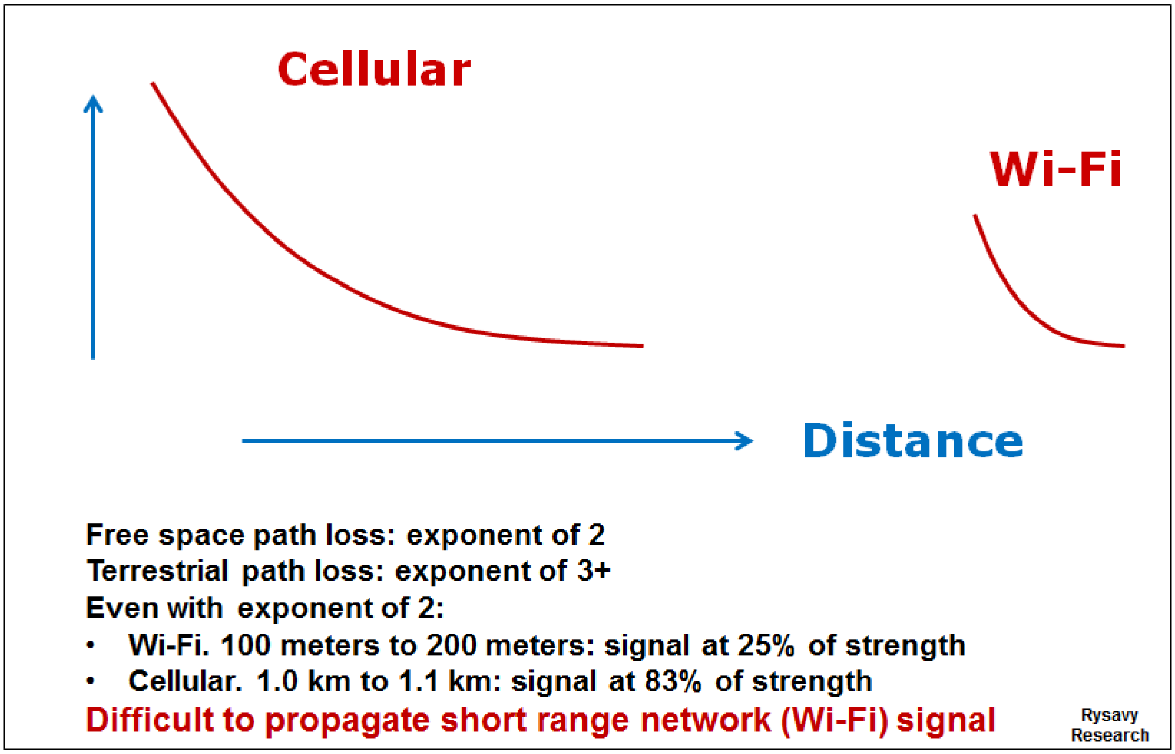

Free space path loss: exponent of 2
Terrestrial path loss: exponent of 3+
Even with exponent of 2:

- **Wi-Fi. 100 meters to 200 meters: signal at 25% of strength**
- **Cellular. 1.0 km to 1.1 km: signal at 83% of strength**

Difficult to propagate short range network (Wi-Fi) signal

Spectrum Sharing

In 2012, the President's Council of Advisors on Science and Technology issued a report titled, "Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth." The report recommended spectrum sharing between government and commercial entities.

The U.S. government can designate spectrum for exclusive, shared, or unlicensed use, as shown in Figure 36. Shared use can be opportunistic, as with TV white spaces; two-tier with incumbents and licensed users; or three-tier, which adds opportunistic access. The bands initially targeted for spectrum sharing include AWS-3 (two tiers on a temporary basis) and the 3.5 GHz band (three tiers).

The three-tier plan envisioned by the U.S. government for the 3.5 GHz band gives more entities access to the spectrum but at the cost of increased complexity.

---

[95] Assumes 1.0 km radius for cellular and 100-meter radius for Wi-Fi.

Figure 36: Spectrum Use and Sharing Approaches

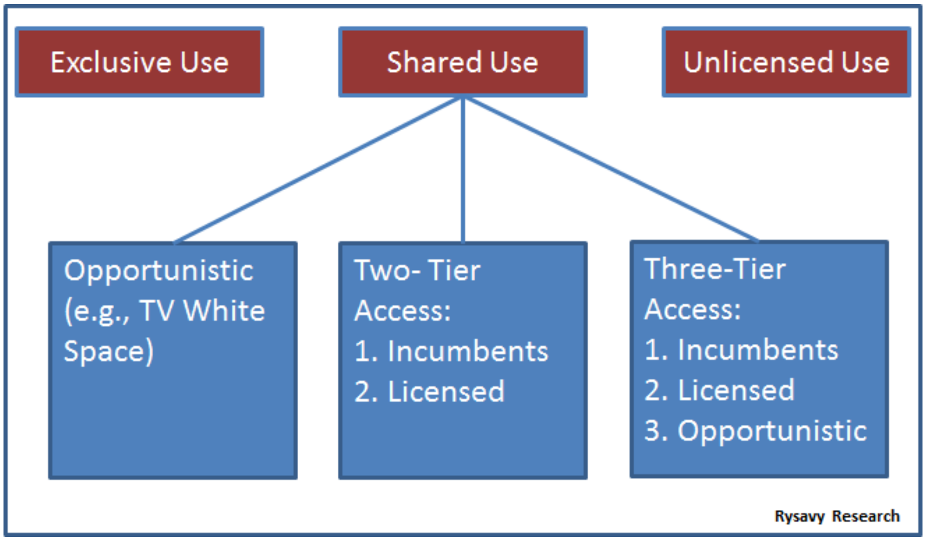
Exclusive Use
Shared Use
Unlicensed Use
Opportunistic (e.g., TV White Space)
Two- Tier Access: 1. Incumbents 2. Licensed
Three-Tier Access: 1. Incumbents 2. Licensed 3. Opportunistic
Rysavy Research

The European Telecommunications Standards Institute (ETSI) is the leading organization standardizing cognitive radios. The most relevant effort is called “Licensed Shared Access” (LSA), a two-tier spectrum sharing system that includes incumbents and licensed secondary users that access shared spectrum via a database, as depicted in Figure 37.

Figure 37: Licensed Shared Access (LSA)

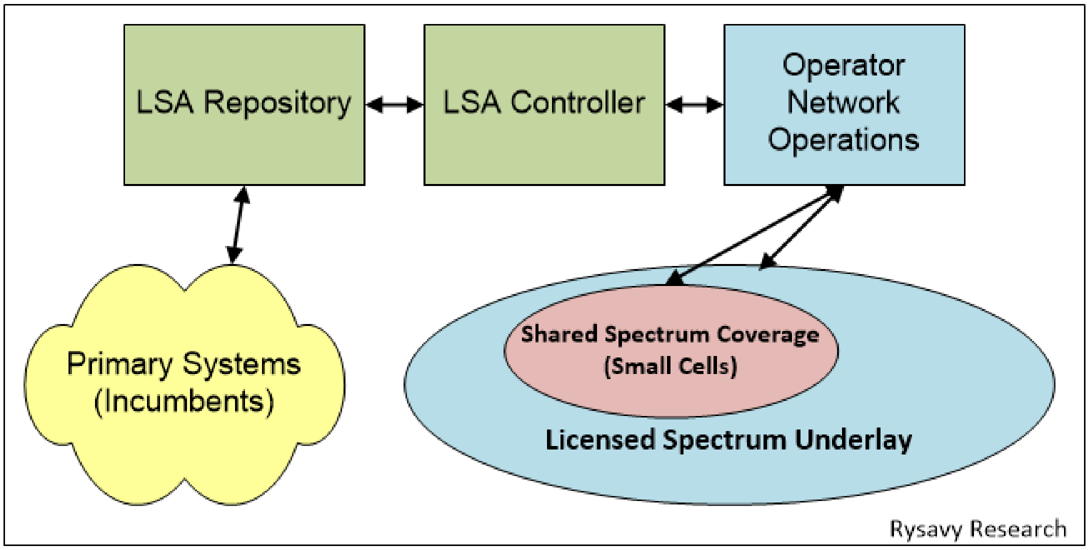
LSA Repository
LSA Controller
Operator Network Operations
Primary Systems (Incumbents)
Shared Spectrum Coverage (Small Cells)
Licensed Spectrum Underlay
Rysavy Research

The three-tier system expected for the 3.5 GHz band in the United States will be complex, necessitating a real-time Spectrum Access System, the SAS, the design and development of which will encompass:

- Algorithms and methods;
- Methods of nesting hierarchical SAS entities (federal secure SAS versus commercial SASs);

- Coordination among multiple, competing commercial SAS managing entities;
- Interface definitions;
- Communication protocol definitions;
- Database and protocol security;
- Policy enforcement;
- Speed of channel allocation/reallocation;
- Time intervals for spectrum allocation;
- Effectively managing large numbers of Tier 3 users; and
- Data ownership, fees, rules, fairness, and conflict resolutions, all of which have policy, regulatory, and business implications.

TVWS databases available today address only a tiny subset of these requirements.[96]

Figure 38 shows the proposed architecture of the 3.5 GHz system. The system consists of incumbents (government systems), Priority Access Licenses, and General Authorized Access. Government systems include military ship-borne radar, military ground-based radar, fixed satellite service earth stations (receive only), and government broadband services (3650 to 3700 MHz). PAL licenses will be used by entities such as cellular operators and will be available for three-year periods. GAA users are licensed by rule and must protect both incumbents and PALs. Incumbents are protected by an Environmental Sensing Capability that detects incumbents and informs the SAS. Some examples of GAA use cases are small-business hotspots, campus hotspots, and backhaul.

Citizens Broadband Radio Service Devices (CBSDs) are the base stations operating under this service; they can operate only under the authority and management of the SAS, either by direct communications or a proxy node.

---

[96] For further analysis, see Rysavy Research and Mobile Future, *Complexities of Spectrum Sharing: How to Move Forward*, April 2014. Available at http://www.rysavy.com/Articles/2014-04-Spectrum-Sharing-Complexities.pdf.

Figure 38: United States 3.5 GHz System Currently Being Developed

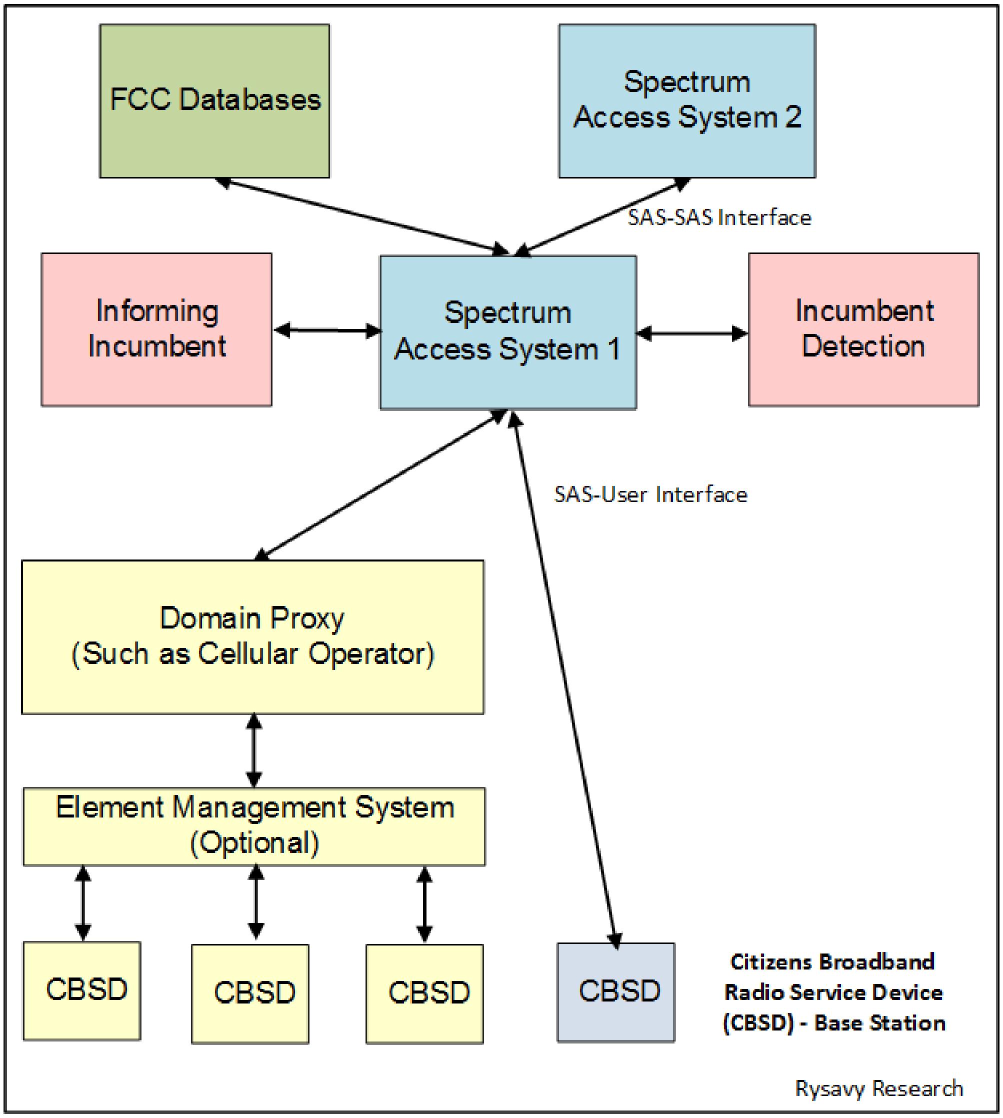
FCC Databases
Spectrum Access System 2
SAS-SAS Interface
Informing Incumbent
Spectrum Access System 1
Incumbent Detection
SAS-User Interface
Domain Proxy (Such as Cellular Operator)
Element Management System (Optional)
CBSD
CBSD
CBSD
CBSD
Citizens Broadband Radio Service Device (CBSD) - Base Station
Rysavy Research

Conclusion

Mobile broadband remains at the forefront of innovation and development in computing, networking, and application development. As users, and now machines, consume ever more wireless data, the industry is responding with more efficient, faster, and higher capacity networks. Even as excitement builds about 5G, LTE, through ongoing advances, has become the global standard.

LTE-Advanced and LTE-Advanced Pro innovations include VoLTE, 1 Gbps peak rate capability, higher-order MIMO, carrier aggregation, LAA/LWA/LWIP, IoT capabilities in Narrowband-IoT and Category M-1, V2X communications, small-cell support, URLLC, SON, dual connectivity, and CoMP—all capabilities that will improve performance, efficiency, and capacity and enable support for new vertical segments. Carriers are implementing NFV and SDN to reduce network costs, improve service velocity, and simplify deployment of new services. Such improvements also facilitate cloud RANs that promise further efficiency gains.

5G research and development efforts have accelerated; standards-based deployment will begin in 2019 and continue through 2030. Operators have announced pre-standard fixed-wireless deployments for 2017. 3GPP has implemented a 5G standardization process that began in Release 14 with a study of the new 5G radio, continues now with a first phase of specifications in Release 15 that provides both non-standalone and standalone options, then moves ahead with a second phase of complete specifications in Release 16.

By harnessing new spectrum, such as mmWave bands above 24 GHz, 5G will eventually be able to access ten times as much spectrum as is currently available for cellular operation. Using radio bands 200 MHz or wider will result in multi-Gbps throughput capabilities. 5G will be designed to integrate with LTE networks, and many 5G features may be implemented as LTE-Advanced Pro extensions prior to full 5G availability.

Small cells will play an ever-more-important role in boosting capacity, benefiting from a number of technologies and developments, including SON, eICIC, Dual Connectivity, LWA/LWIP, LTE-U, LTE-LAA, MulteFire, improved backhaul options, and spectrum ideal for small cells, such as the 3.5 GHz and mmWave bands.

Obtaining more spectrum remains a priority globally. In U.S. markets, the recently completed television incentive auctions for 600 MHz spectrum supplied 70 MHz of new licensed spectrum; the 3.5 GHz small-cell band will provide 150 MHz; and the 5G spectrum bands, with both licensed and unlicensed bands, will add about 10 GHz of new spectrum.

The future of mobile broadband, including both LTE-Advanced and 5G, is bright, with no end in sight for continued growth in capability, nor for the limitless application innovation that mobile broadband enables.

Appendix: Technology Details

The 3GPP family of data technologies provides ever increasing capabilities that support ever more demanding applications. Services obviously need to provide broad coverage and high data throughput. Less obvious for users, but as critical for effective application performance, are the need for low latency, QoS control, and spectral efficiency. Higher spectral efficiency translates to higher average throughputs (and thus more responsive applications) for more active users in a coverage area. The discussion below details the progression of capability for each technology, including throughput, security, latency, QoS, and spectral efficiency.

This appendix provides details on 5G, UMTS/HSPA, and EDGE.

3GPP Releases

- ❑ **Release 99**: Completed. First deployable version of UMTS. Enhancements to GSM data (EDGE). Provides support for GSM/EDGE/GPRS/WCDMA radio-access networks.

- ❑ **Release 4**: Completed. Multimedia messaging support. First steps toward using IP transport in the core network.

- ❑ **Release 5**: Completed. HSDPA. First phase of Internet Protocol Multimedia Subsystem (IMS). Full ability to use IP-based transport instead of just Asynchronous Transfer Mode (ATM) in the core network.

- ❑ **Release 6**: Completed. HSUPA. Enhanced multimedia support through Multimedia Broadcast/Multicast Services (MBMS). Performance specifications for advanced receivers. Wireless Local Area Network (WLAN) integration option. IMS enhancements. Initial VoIP capability.

- ❑ **Release 7**: Completed. Provides enhanced GSM data functionality with Evolved EDGE. Specifies HSPA+, which includes higher-order modulation and MIMO. Performance enhancements, improved spectral efficiency, increased capacity, and better resistance to interference. Continuous Packet Connectivity (CPC) enables efficient “always-on” service and enhanced uplink UL VoIP capacity, as well as reductions in call set-up delay for Push-to-Talk Over Cellular (PoC). Radio enhancements to HSPA include 64 Quadrature Amplitude Modulation (QAM) in the downlink and 16 QAM in the uplink. Also includes optimization of MBMS capabilities through the multicast/broadcast, single frequency network (MBSFN) function.

- ❑ **Release 8**: Completed. Comprises further HSPA Evolution features such as simultaneous use of MIMO and 64 QAM. Includes dual-carrier HSDPA (DC-HSDPA) wherein two downlink carriers can be combined for a doubling of throughput performance. Specifies OFDMA-based 3GPP LTE. Defines EPC and EPS.

- ❑ **Release 9**: Completed. HSPA and LTE enhancements including HSPA dual-carrier downlink operation in combination with MIMO, Multimedia Broadcast Multicast Services (MBMS), HSDPA dual-band operation, HSPA dual-carrier uplink operation, EPC enhancements, femtocell support, support for regulatory features such as emergency user equipment positioning and Commercial Mobile Alert System (CMAS), and evolution of IMS architecture.

- ❑ **Release 10**: Completed. Specifies LTE-Advanced that meets the requirements set by ITU’s IMT-Advanced project. Key features include carrier aggregation, multi-antenna enhancements such as enhanced downlink eight-branch MIMO and uplink MIMO, relays, enhanced LTE Self-Organizing Network capability, Evolved Multimedia Broadcast Multicast Services (eMBMS), HetNet enhancements that include eICIC, Local IP Packet Access, and new frequency bands. For HSPA, includes quad-carrier operation and additional MIMO options. Also includes femtocell enhancements, optimizations for M2M communications, and local IP traffic offload.

- **Release 11**: Completed. For LTE, emphasizes Coordinated Multi Point (CoMP), carrier-aggregation enhancements, devices with interference cancellation, development of the Enhanced Physical Downlink Control Channel (EPDCCH), and further enhanced eICIC including devices with CRS (Cell-specific Reference Signal) interference cancellation. The release includes further DL and UL MIMO enhancements for LTE. For HSPA, provides eight-carrier on the downlink, uplink enhancements to improve latency, dual-antenna beamforming and MIMO, CELL Forward Access Channel (FACH) state enhancement for smartphone-type traffic, four-branch MIMO enhancements and transmissions for HSDPA, 64 QAM in the uplink, downlink multipoint transmission, and noncontiguous HSDPA carrier aggregation. Wi-Fi integration is promoted through S2a Mobility over GPRS Tunneling Protocol (SaMOG). An additional architectural element called "Machine-Type Communications Interworking Function" (MTC-IWF) will more flexibly support machine-to-machine communications.

- **Release 12**: Completed. Enhancements include improved small cells/HetNets for LTE, LTE multi-antenna/site technologies (including Active Antenna Systems), Dual Connectivity, 256 QAM modulation option, further CoMP/MIMO enhancements, enhancements for interworking with Wi-Fi, enhancements for MTC, SON, support for emergency and public safety, Minimization of Drive Tests (MDT), advanced receivers, device-to-device communication (also referred to as Proximity Services), group communication enablers in LTE, addition of Web Real Time Communication (WebRTC) to IMS, energy efficiency, more flexible carrier aggregation, dynamic adaptation of uplink-downlink ratios in TDD mode, further enhancements for HSPA+, small cells/HetNets, Scalable-UMTS, and FDD-TDD carrier aggregation.

- **Release 13**: Completed. LTE features include Active Antenna Systems (AAS) with support for as many as 16 antenna elements (full-dimension MIMO) and beamforming, Network-Assisted Interference Cancellation and Suppression (NAICS), radio-access network sharing, carrier aggregation supporting 32 component carriers,[97] carrier aggregation of up to four carriers on the downlink and two carriers on the uplink, LAA for operation in unlicensed bands, LTE Wi-Fi Aggregation including LWIP, RCLWI, isolated operation and mission-critical voice communications for public safety, application-specific congestion management, User-Plane Congestion Management, enhancement to WebRTC interoperability, architecture enhancement for dedicated core networks, enhancement to proximity-based services, Mission-Critical Push-to-Talk, group communications, CoMP enhancements, small cell enhancements, machine-type communications enhancements including NB-IoT and Extended Coverage GSM (EC-GSM), VoLTE enhancements, SON enhancements, shared network enhancements, indoor positioning based on WLAN access points, Bluetooth beacons and barometric pressure, and enhanced circuit-switched fallback. HSPA+ features include support for dual-band uplink carrier aggregation.

---

[97] This level of aggregation refers to signaling capabilities. The number of carriers that can be combined in an actual deployment is smaller and depends on RAN co-existence studies. Refer to the appendix section on "Carrier Aggregation" for additional details.

- **Release 14**: Expected completion June 2017. Contemplated features include uplink operation for LAA (enhanced LAA), full-dimension MIMO enhanced with up to 32 antenna elements, dual-connectivity of licensed and unlicensed carriers across non-collocated nodes, vehicle-to-vehicle and vehicle-to-infrastructure (V2X) communications built on Release 12 Proximity Services, shared LTE broadcast in which different operators broadcast the same content on the same frequency, non-IP operation for IoT, Downlink Multi-user Superposition Transmission (MUST), enhanced LWA, VoLTE enhancements, LWIP/LWA enhancements, eMBMS enhancements, NB-IoT enhancements, and LTE latency reduction.

- **Release 15**: Expected completion September 2018. Non-standalone (using LTE core network) option expected March 2018. Specifies phase 1 of 5G. NR radio, 4G-5G interworking, and MIMO/beamforming. Further LTE enhancements include ultra-reliable low-latency communications, NB-IoT enhancements, LAA enhancements, V2X enhancements, DL 1024 QAM, CoMP enhancements, AAS enhancements, and LTE/5G core network capability.

- **Release 16**: Expected completion end of 2019. Specifies phase 2 of 5G. Adds URLLC, spectrum sharing, unlicensed spectrum operation and integration, and multiple other enhancements. Further LTE enhancements.

Data Throughput Comparison

Data throughput is an important metric for quantifying network throughput performance. Unfortunately, the ways in which various organizations quote throughputs vary tremendously, often resulting in misleading claims. The intent of this paper is to realistically represent the capabilities of these technologies.

One method of representing a technology's throughput is what people call "peak throughput" or "peak network speed," which refers to the fastest possible transmission speed over the radio link and is generally based on the highest-order modulation available and the least amount of coding (error correction) overhead. Peak network speed is also usually quoted at layer 2 of the radio link. Because of protocol overhead, actual application throughput may be up to 10% lower than this layer-2 value.

Another method is to disclose throughputs actually measured in deployed networks with applications such as File Transfer Protocol (FTP) under favorable conditions, which assume light network load (as low as one active data user in the cell sector) and favorable signal propagation. This number is useful because it demonstrates the high-end, actual capability of the technology in current deployments, referred to in this paper as the "peak user rate." Average rates are lower than this peak rate and are difficult to predict because they depend on a multitude of operational and network factors. Except when the network is congested, however, the majority of users should experience throughput rates higher than one-half of the peak achievable rate.

Some operators, primarily in the United States, also quote typical throughput rates, which are based on throughput tests the operators have done across their operating networks and incorporate a higher level of network load. Although the operators do not disclose the precise methodologies they use to establish these figures, the values provide a good indication of what users can realistically expect.

Table 14 presents the technologies in terms of peak network throughput rates, peak user rates (under favorable conditions), and typical rates. It omits values that are not yet known, such as for future technologies.

The projected typical rates for HSPA+ and LTE show a wide range because these technologies exploit favorable radio conditions to achieve high throughput rates, but under poor radio conditions, throughput rates are lower.

Table 14: Throughput Performance of Different Wireless Technologies (Blue Indicates Theoretical Peak Rates, Green Typical)

	Downlink		Uplink	
	Peak Network Speed	Peak and/or Typical User Rate	Peak Network Speed	Peak and/or Typical User Rate
5G in mmWave, early versions[98]	5 Gbps	500 Mbps	2 Gbps	250 Mbps
5G in mmWave, later versions	50 Gbps	5 Gbps	25 Gbps	2 Gbps
LTE (2X2 MIMO, 10+10 MHz, DL 64 QAM, UL 16 QAM)	70 Mbps	6.5 to 26.3 Mbps[99]	35 Mbps[100]	6.0 to 13.0 Mbps
LTE-Advanced (2X2 or 4X4 MIMO, 20+20 MHz or 40+20 MHz with Carrier Aggregation [CA], DL 64 QAM, UL 16 QAM)	300 Mbps	N/A	71 Mbps[101]	N/A
LTE Advanced (4X4 MIMO, 60+20MHz, CA, 256 QAM DL, 64 QAM UL)	600 Mbps		150 Mbps	
LTE Advanced (4X4 MIMO, 80+20 MHz, CA, 256 QAM DL, 64 QAM UL)	> 1 Gbps		150 Mbps	

---

[98] Speculative values, Rysavy Research estimates. Assumes 200 MHz radio channel, 2:1 TDD. Throughput rates would double using 400 MHz.

[99] 5G Americas member company analysis for downlink and uplink. Assumes single user with 50% load in other sectors. AT&T and Verizon are quoting typical user rates of 5-12 Mbps on the downlink and 2-5 Mbps on the uplink for their networks. See additional LTE throughput information in the section below, “LTE Throughput.”

[100] Assumes 64 QAM. Otherwise 22 Mbps with 16 QAM.

[101] Assumes 64 QAM. Otherwise 45 Mbps with 16 QAM.

	Downlink		Uplink	
	Peak Network Speed	Peak and/or Typical User Rate	Peak Network Speed	Peak and/or Typical User Rate
LTE Advanced (8X8 MIMO, 20+20 MHz, DL 64 QAM, UL 64 QAM)	1.2 Gbps	N/A	568 Mbps	N/A
LTE Advanced, 100 MHz + 100 MHz	3 Gbps		1.5 Gbps	
LTE Advanced 32 Carriers	>> 3 Gbps			
EDGE (type 2 MS)	473.6 Kbps	Not Applicable (N/A)	473.6 Kbps	N/A
EDGE (type 1 MS) (Practical Terminal)	236.8 Kbps	200 Kbps peak 160 to 200 Kbps typical[102]	236.8 Kbps	200 Kbps peak 80 to 160 Kbps typical[103]
HSDPA Initial Devices (2006)	1.8 Mbps	> 1 Mbps peak	384 Kbps	350 Kbps peak
HSDPA	14.4 Mbps	N/A	384 Kbps	N/A

---

[102] Assumes four-to-five downlink timeslot devices (each timeslot capable of 40 Kbps).

[103] Assumes two-to-four uplink timeslot devices (each timeslot capable of 40 Kbps).

	Downlink		Uplink	
	Peak Network Speed	**Peak and/or Typical User Rate**	**Peak Network Speed**	**Peak and/or Typical User Rate**
HSPA[104] Initial Implementation	7.2 Mbps	> 5 Mbps peak 700 Kbps to 1.7 Mbps typical[105]	2 Mbps	> 1.5 Mbps peak 500 Kbps to 1.2 Mbps typical
HSPA	14.4 Mbps	N/A	5.76 Mbps	N/A
HSPA+ (DL 64 QAM, UL 16 QAM, 5+5 MHz)	21.6 Mbps	1.9 Mbps to 8.8 Mbps typical[106]	11.5 Mbps	1 Mbps to 4 Mbps typical
HSPA+ (2X2 MIMO, DL 16 QAM, UL 16 QAM, 5+5 MHz)	28 Mbps	N/A	11.5 Mbps	N/A
HSPA+ (2X2 MIMO, DL 64 QAM, UL 16 QAM, 5+5 MHz)	42 Mbps	N/A	11.5 Mbps	N/A
HSPA+ (DL 64 QAM, UL 16 QAM, Dual Carrier, 10+5 MHz)	42 Mbps	Approximate doubling of 5+5 MHz rates - 3.8 to 17.6 Mbps.	11.5 Mbps	1 Mbps to 4 Mbps typical
HSPA+ (2X2 MIMO DL, DL 64 QAM, UL 16 QAM, Dual Carrier, 10+10 MHz)	84 Mbps	N/A	23 Mbps	N/A

---

[104] High Speed Packet Access (HSPA) consists of systems supporting both High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA).

[105] Typical downlink and uplink throughput rates based on AT&T press release, June 4, 2008

[106] 5G Americas member company analysis. Assumes Release 7 with 64 QAM and F-DPCH. Single user. 50% loading in neighboring cells. Higher rates expected with subsequent 3GPP releases.

	Downlink		Uplink	
	Peak Network Speed	Peak and/or Typical User Rate	Peak Network Speed	Peak and/or Typical User Rate
HSPA+ (2X2 MIMO DL, DL 64 QAM, UL 16 QAM, Quad-Carrier,[107] 20+10 MHz)	168 Mbps	N/A	23 Mbps	N/A
HSPA+ (2X2 MIMO DL and UL, DL 64 QAM, UL 16 QAM, Eight-Carrier, 40+10 MHz)	336 Mbps	N/A	69 Mbps	N/A
HSPA+ (4X2 MIMO DL, 2X2 MIMO UL, DL 64 QAM, UL 16 QAM, 8 carrier, 40+10 MHz)	672 Mbps	N/A	69 Mbps	N/A
EDGE (type 2 MS)	473.6 Kbps	Not Applicable (N/A)	473.6 Kbps	N/A
EDGE (type 1 MS) (Practical Terminal)	236.8 Kbps	200 Kbps peak 160 to 200 Kbps typical[108]	236.8 Kbps	200 Kbps peak 80 to 160 Kbps typical[109]
CDMA2000 EV-DO Rel. 0	2.4 Mbps	> 1 Mbps peak	153 Kbps	150 Kbps peak

---

[107] No operators have announced plans to deploy HSPA in a quad (or greater) carrier configuration. Three carrier configurations, however, have been deployed.

[108] Assumes four-to-five downlink timeslot devices (each timeslot capable of 40 Kbps).

[109] Assumes two-to-four uplink timeslot devices (each timeslot capable of 40 Kbps).

	Downlink		Uplink	
	Peak Network Speed	**Peak and/or Typical User Rate**	**Peak Network Speed**	**Peak and/or Typical User Rate**
CDMA2000 EV-DO Rev. A	3.1 Mbps	> 1.5 Mbps peak 600 Kbps to 1.4 Mbps typical[110]	1.8 Mbps	> 1 Mbps peak 300 to 500 Kbps typical
CDMA2000 EV-DO Rev. B (3 radio channels 5+5 MHz)	14.7[111] Mbps	Proportional increase of Rev A typical rates based on number of carriers.	5.4 Mbps	N/A
CDMA2000 EV-DO Rev B Theoretical (15 radio channels 20+20 MHz)	73.5 Mbps	N/A	27 Mbps	N/A

Additional information about LTE throughput appears below in the section "LTE Throughput."

Latency Comparison

As important as throughput is network latency, defined as the round-trip time it takes data to traverse the network. Each successive data technology from GPRS forward reduces latency, with LTE networks having latency as low as 15 msec. Ongoing improvements in each technology mean that all of these values will go down as vendors and operators fine-tune their systems. Figure 39 shows the latency of different 3GPP technologies.

---

[110] Typical downlink and uplink throughput rates based on Sprint press release Jan. 30, 2007.

[111] Assuming use of 64 QAM.

Figure 39: Latency of Different Technologies[112]

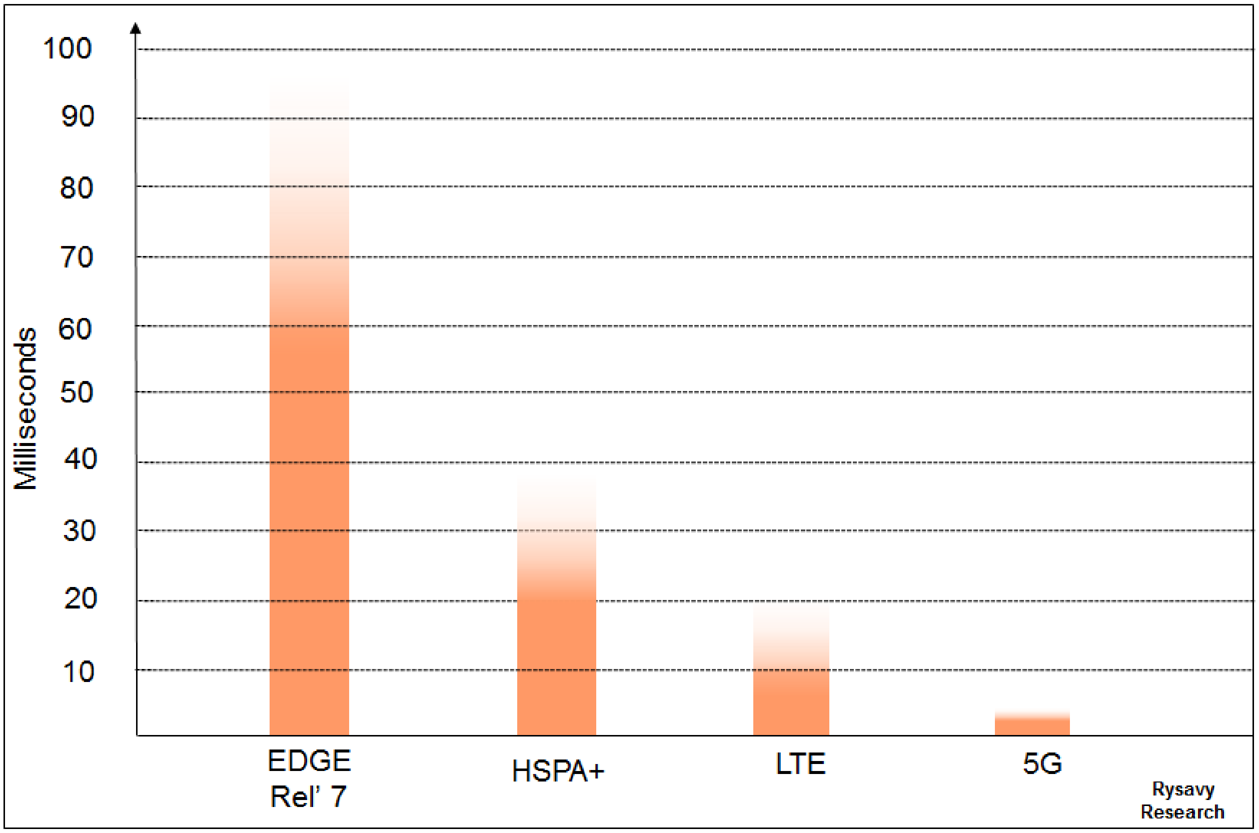

The values shown in Figure 39 reflect measurements of commercially deployed technologies, with EDGE Release 7 achieving 70 to 95 msec, HSPA+ 25 to 30 msec, and LTE 15 to 20 msec. A latency goal for 5G is less than 4 msec for broadband and 0.5 msec for mission-critical applications.

Spectral Efficiency

The evolution of data services is characterized by an increasing number of users with ever-higher bandwidth demands. As the wireless data market grows, deploying wireless technologies with high spectral efficiency is of paramount importance. Keeping all other things equal, including frequency band, amount of spectrum, and cell site spacing, an increase in spectral efficiency translates to a proportional increase in the number of users supported at the same load per user—or, for the same number of users, an increase in throughput available to each user.

Increased spectral efficiency, however, comes at a price because it generally involves greater complexity for both user and base station equipment. Complexity can arise from the increased number of calculations performed to process signals or from additional radio components. Hence, operators and vendors must balance market needs against network and equipment costs. OFDMA technologies, such as LTE and planned 5G approaches, achieve higher spectral efficiency with lower overall complexity, especially in larger bandwidths.

---

[112] 5G Americas member companies. Measured between subscriber unit and a node immediately external to wireless network. Does not include internet latency. Note that there is some variation in latency based on network configuration and operating conditions.

As shown in Figure 40, the link-layer performance of modern wireless technologies is approaching the theoretical limits as defined by the Shannon bound. (The Shannon bound is a theoretical limit to the information transfer rate [per unit bandwidth] that can be supported by any communications link. The bound is a function of the SNRs of the communications link.) Figure 40 also shows that HSDPA, 1xEV-DO, and IEEE 802.16e-2005 are all within 2 to 3 decibels (dB) of the Shannon bound, indicating that there is not much room for improvement from a link-layer perspective.

Figure 40: Performance Relative to Theoretical Limits for HSDPA, EV-DO, and WiMAX (IEEE 802.16e-2005)[113]

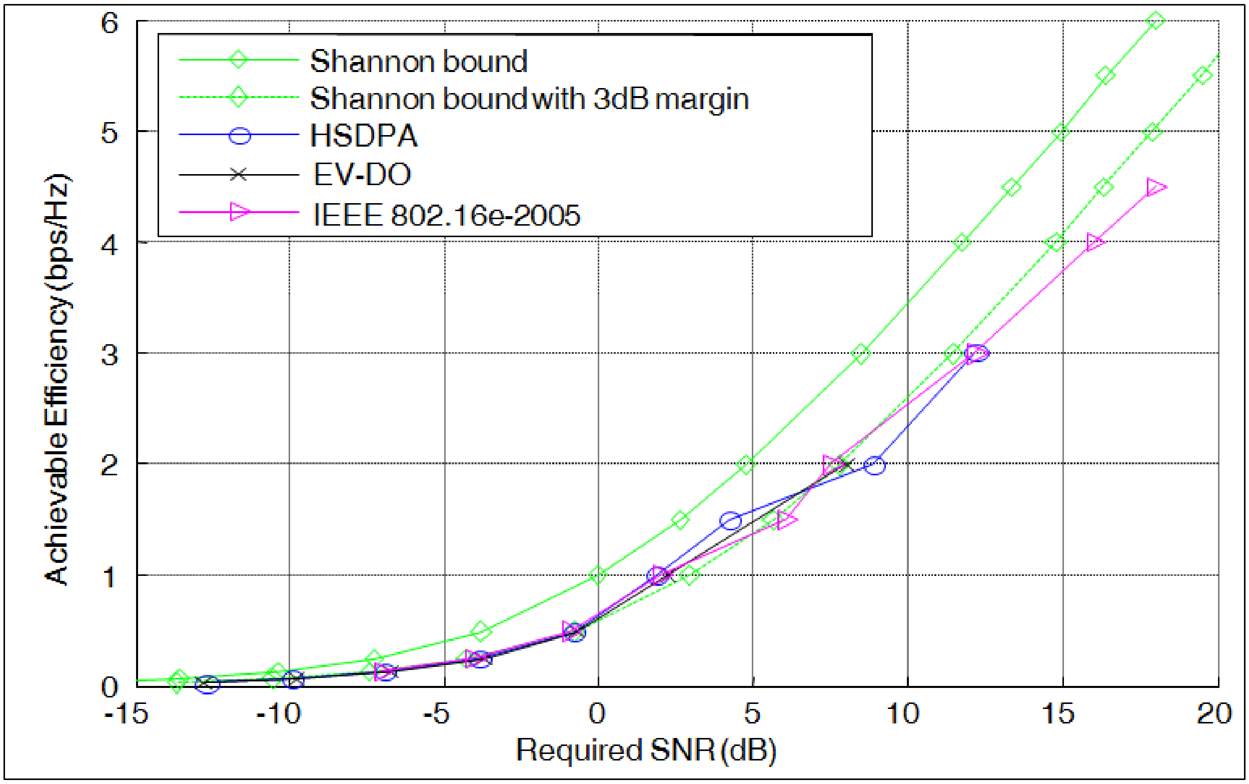

The curves in Figure 40 are for an Additive White Gaussian Noise Channel (AWGN). If the channel is slowly varying and the frame interval is significantly shorter than the coherence time, the effects of fading can be compensated for by practical channel estimation algorithms—thus justifying the AWGN assumption. For instance, at 3 km per hour and fading at 2 GHz, the Doppler spread is about 5.5 Hz. The coherence time of the channel is thus 1 second (sec)/5.5 or 180 msec. Frames are well within the coherence time of the channel, because they are typically 20 msec or less. As such, the channel appears

---

[113] 5G Americas member contribution.

“constant” over a frame, and the Shannon bound applies. Furthermore, significantly more of the traffic in a cellular system is at slow speeds (for example, 3 km/hr. or less) rather than at higher speeds. The Shannon bound is consequently also relevant for a realistic deployment environment.

As the speed of the mobile station increases and the channel estimation becomes less accurate, additional margin is needed. This additional margin, however, would impact the different standards fairly equally.

The focus of future technology enhancements is on improving system performance aspects that reduce interference to maximize the experienced SNRs in the system and antenna techniques (such as MIMO) that exploit multiple links or steer the beam rather than on investigating new air interfaces that attempt to improve link-layer performance.

MIMO techniques using spatial multiplexing to increase the overall information transfer rate by a factor proportional to the number of transmit or receive antennas do not violate the Shannon bound because the per-antenna transfer rate (that is, the per-communications link transfer rate) is still limited by the Shannon bound.

Figure 41 compares the spectral efficiency of different wireless technologies based on a consensus view of 5G Americas contributors to this paper. It shows the continuing evolution of the capabilities of all the technologies discussed. The values shown are reasonably representative of real-world conditions. Most simulation results produce values under idealized conditions; as such, some of the values shown are lower (for all technologies) than the values indicated in other papers and publications. For instance, 3GPP studies indicate higher HSPA and LTE spectral efficiencies. Nevertheless, there are practical considerations in implementing technologies that can prevent actual deployments from reaching calculated values. Consequently, initial versions of technology may operate at lower levels but then improve over time as designs are optimized. Therefore, readers should interpret the values shown as achievable, but not as the actual values that might be measured in any specific deployed network.

Figure 41: Comparison of Downlink Spectral Efficiency[114]

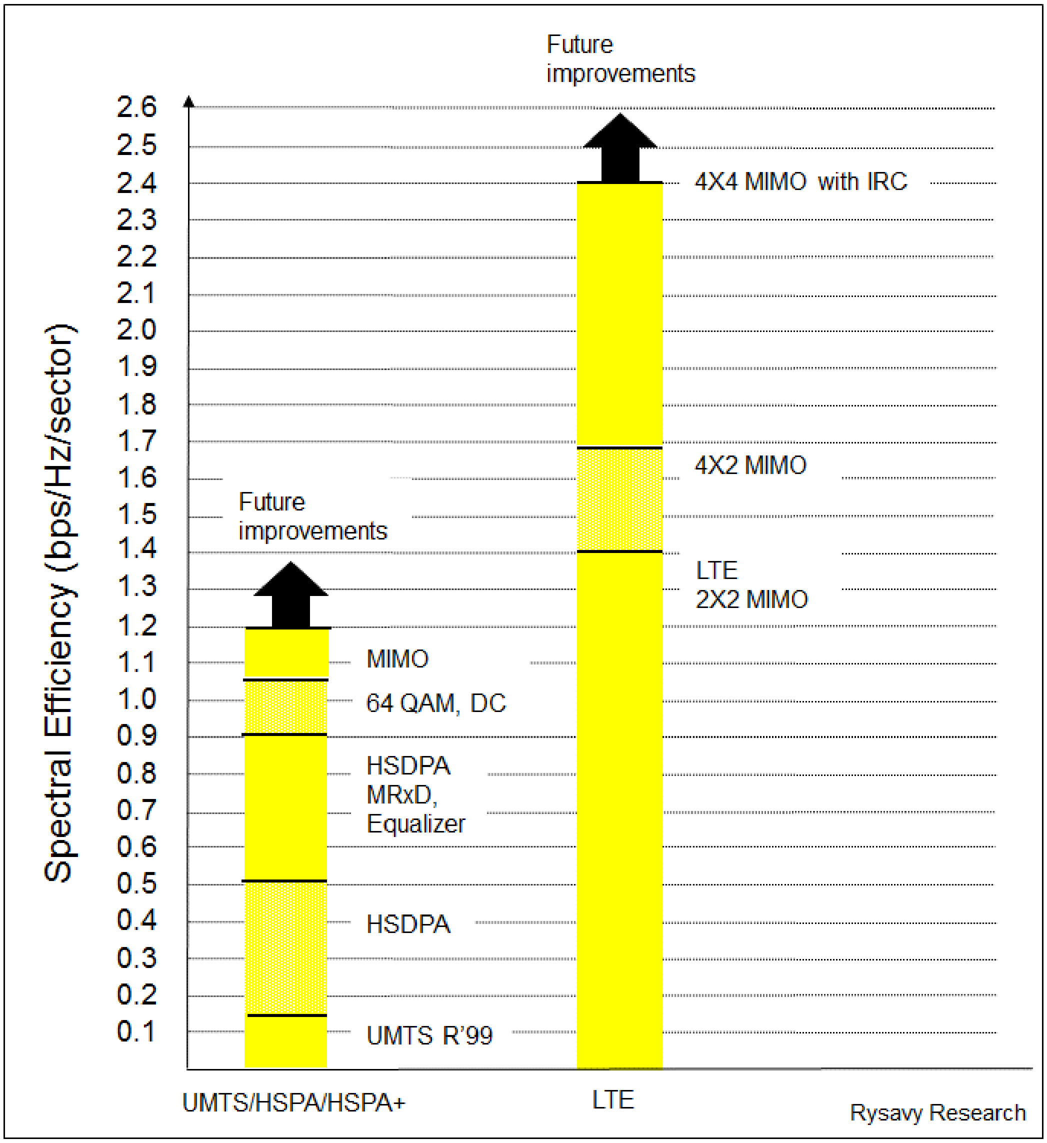

The values shown in Figure 41 are not all possible combinations of available features. Rather, they are representative milestones in ongoing improvements in spectral efficiency. For instance, terminals may employ Mobile Receive Diversity but not equalization.

The figure does not include EDGE, but EDGE itself is spectrally efficient at 0.6 bps/Hz using mobile receive diversity and, potentially, 0.7 bps/Hz with MIMO. Relative to WCDMA Release 99, HSDPA increases capacity by almost a factor of three. Type 3 receivers that include MMSE equalization and Mobile Receive Diversity (MRxD) effectively double HSDPA spectral efficiency. The addition of dual-carrier operation and 64 QAM increases spectral efficiency by about 15%, and MIMO can increase spectral efficiency by another 15%, reaching 1.2 bps/Hz. Dual-carrier HSPA+ offers a gain in spectral efficiency from cross-carrier scheduling with possible gains of about 10%.[115]

---

[114] Joint analysis by 5G Americas members. 5+5 MHz FDD for UMTS-HSPA/LTE. Mix of mobile and stationary users.

Some enhancements, such as 64 QAM for HSPA, are simpler to deploy than other enhancements, such as 2X2 MIMO. The former can be done as a software upgrade, whereas the latter requires additional hardware at the base station. Thus, the figure does not necessarily show the actual progression of technologies that operators will deploy to increase spectral efficiency.

Beyond HSPA, 3GPP LTE results in further spectral efficiency gains, initially with 2X2 MIMO, then 4X2 MIMO, and then 4X4 MIMO. The gain for 4X2 MIMO will be 20% more than LTE with 2X2 MIMO; the gain for 4X4 MIMO in combination with interference rejection combining (IRC) will be 70% greater than 2X2 MIMO, reaching 2.4 bps/Hz. This value represents a practical deployment of 4X4 MIMO, with random phase and some timing-alignment error included in each of the four transmit paths. CoMP, discussed below in the appendix, provides a minimal contribution to spectral efficiency.

LTE is even more spectrally efficient when deployed using wider radio channels of 10+10 MHz and 20+20 MHz, although most of the gain is realized at 10+10 MHz. LTE TDD has spectral efficiency that is within 1% or 2% of LTE FDD.[116]

Figure 42 compares the uplink spectral efficiency of the different systems.

---

[115] 5G Americas member analysis. Vendor estimates for spectral-efficiency gains from dual-carrier operation range from 5% to 20%. Lower spectral efficiency gains are due to full-buffer traffic assumptions. In more realistic operating scenarios, gains will be significantly higher.

[116] Assumes best-effort traffic. Performance between LTE-TDD and FDD differs for real-time traffic for the following reasons: a.) The maximum number of HARQ process should be made as small as possible to reduce the packet re-transmission latency. b.) In FDD, the maximum number of HARQ process is fixed and, as such, the re-transmission latency is 7ms. c.) For TDD, the maximum number of HARQ process depends on the DL:UL configurations. As an example, the re-transmission latency for TDD config-1 is 9ms. d.) Because of higher re-transmission latency, the capacity of real-time services cannot be scaled for TDD from FDD based on the DL:UL ratio.

Figure 42: Comparison of Uplink Spectral Efficiency[117]

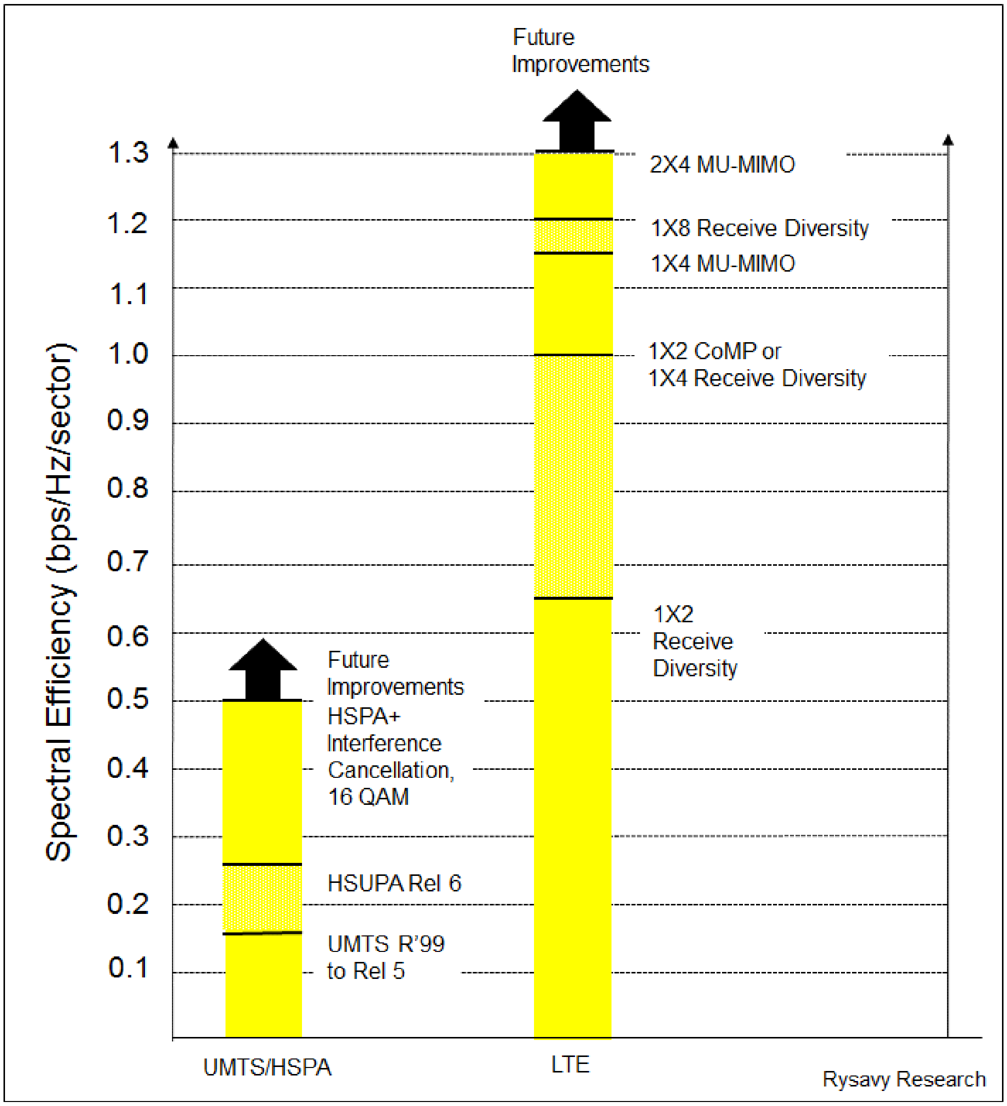

The implementation of HSUPA in HSPA significantly increases uplink capacity.

With LTE, spectral efficiency increases by use of receive diversity. Initial systems will employ 1X2 receive diversity (two antennas at the base station). 1X4 diversity will increase spectral efficiency by 50%, to 1.0 bps/Hz, and 1X8 diversity will provide a further 20% increase, from 1.0 bps/Hz to 1.2 bps/Hz.

It is also possible to employ Multi-User MIMO (MU-MIMO), which allows simultaneous transmission by multiple users on the same physical uplink resource to increase spectral efficiency. MU-MIMO will provide a 15% to 20% spectral efficiency gain, with actual increases depending on how well link adaptation is implemented. The figure uses a conservative 15% gain, showing MU-MIMO with a 1X4 antenna configuration increasing spectral efficiency by 15%, to 1.15 bps/Hz, and 2X4 MU-MIMO a further 15%, to 1.3 bps/Hz.

---

[117] Joint analysis by 5G Americas members. 5+5 MHz for UMTS-HSPA/LTE. Mix of mobile and stationary users.

In Release 11, uplink CoMP using 1X2 increases efficiency from .65 bps/Hz to 1.0 bps/Hz. Many of the techniques used to improve LTE spectral efficiency can also be applied to HSPA since they are independent of the radio interface.

Figure 43 compares voice spectral efficiency.

Figure 43: Comparison of Voice Spectral Efficiency[118]

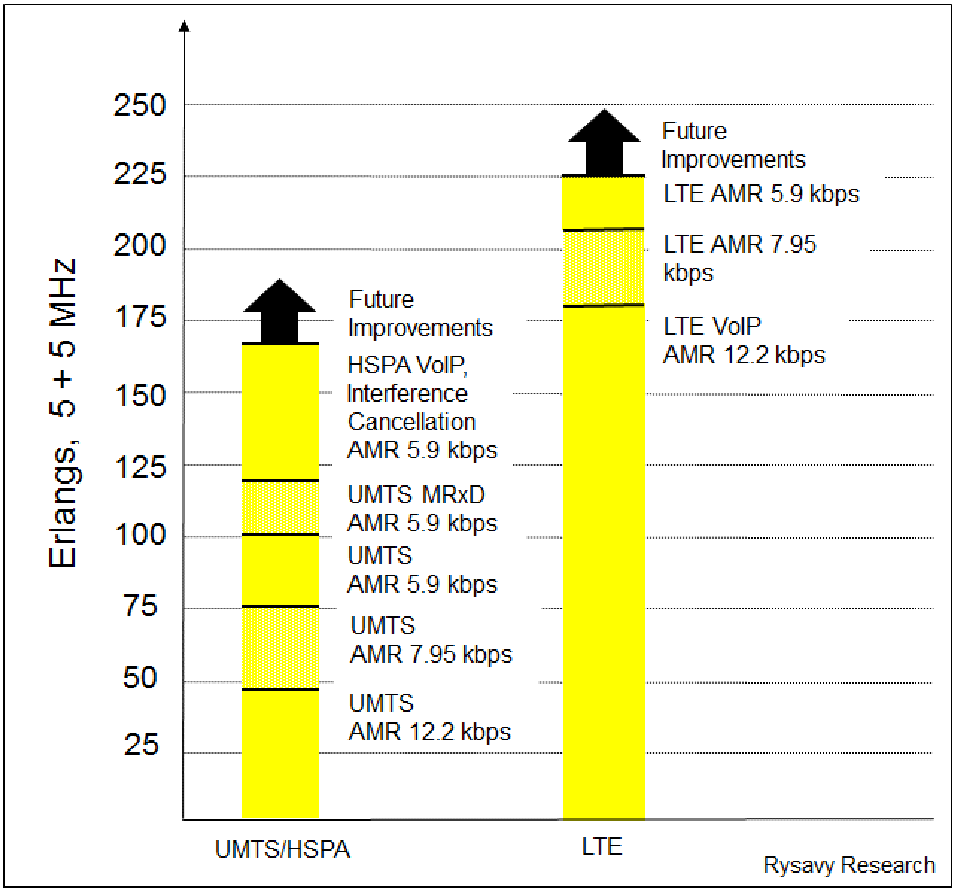

Figure 43 shows UMTS Release 99 with AMR 12.2 Kbps, 7.95 Kbps, and 5.9 Kbps vocoders. The AMR 12.2 Kbps vocoder provides superior voice quality in good (for example, static and indoor) channel conditions.

UMTS has dynamic adaptation between vocoder rates, enabling enhanced voice quality compared with EVRC at the expense of capacity in situations that are not capacity limited. With the addition of mobile receive diversity, UMTS circuit-switched voice capacity could reach 120 Erlangs in 5+5 MHz.

---

[118] Joint analysis by 5G Americas members. 5 + 5 MHz for UMTS-HSPA/LTE. Mix of mobile and stationary users.

VoIP Erlangs in this paper are defined as the average number of concurrent VoIP users that can be supported over a defined period of time (often one hour) assuming a Poisson arrival process and meeting a specified outage criteria (often less than 2% of the users exhibiting greater than 1% frame-error rate). Depending on the specific enhancements implemented, voice capacity could double over existing circuit-switched systems. These gains do not derive through use of VoIP, but rather from advances in radio techniques applied to the data channels. Many of these same advances may also be applied to current circuit-switched modes.

LTE achieves very high voice spectral efficiency because of better uplink performance since there is no in-cell interference. The figure shows LTE VoIP spectral efficiency using AMR at 12.2 Kbps, 7.95 Kbps, and 5.9 Kbps.

VoIP for LTE can use a variety of codecs. The figures show performance assuming specific codecs at representative bit rates. For Enhanced Variable Rate Codecs (EVRCs), the figure shows the average bit rate.

The voice efficiency of the wideband AMR voice codec, operating at 12.65 Kbps, is similar to the AMR codec at 12.2 Kbps, with a value of 180 Erlangs for both since both codecs operate at approximately the same bit rate. 1xRTT has voice capacity of 85 Erlangs in 5+5 MHz with EVRC-A and reaches voice capacity of 120 Erlangs with the use of Quasi-Linear Interference Cancellation (QLIC) and EVRC-B at 6 Kbps.

Data Consumed by Streaming and Virtual Reality

Table 15 quantifies usage based on advanced video compression schemes such as H.264 and H.265, the type of application, and usage per day.

Table 15: Data Consumed by Streaming and Virtual Reality[119]

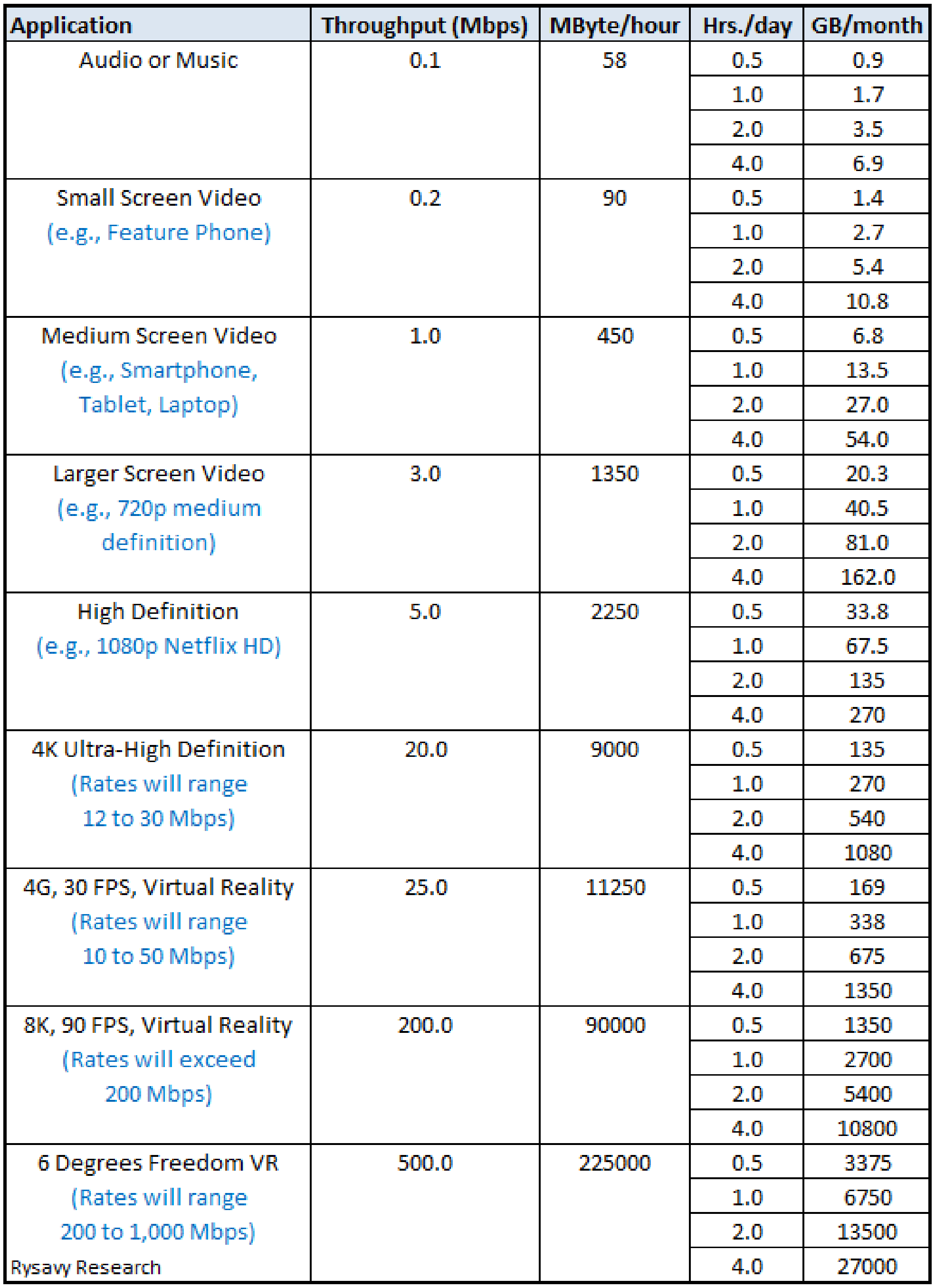

Application	Throughput (Mbps)	MByte/hour	Hrs./day	GB/month
Audio or Music	0.1	58	0.5	0.9
			1.0	1.7
			2.0	3.5
			4.0	6.9
Small Screen Video (e.g., Feature Phone)	0.2	90	0.5	1.4
			1.0	2.7
			2.0	5.4
			4.0	10.8
Medium Screen Video (e.g., Smartphone, Tablet, Laptop)	1.0	450	0.5	6.8
			1.0	13.5
			2.0	27.0
			4.0	54.0
Larger Screen Video (e.g., 720p medium definition)	3.0	1350	0.5	20.3
			1.0	40.5
			2.0	81.0
			4.0	162.0
High Definition (e.g., 1080p Netflix HD)	5.0	2250	0.5	33.8
			1.0	67.5
			2.0	135
			4.0	270
4K Ultra-High Definition (Rates will range 12 to 30 Mbps)	20.0	9000	0.5	135
			1.0	270
			2.0	540
			4.0	1080
4G, 30 FPS, Virtual Reality (Rates will range 10 to 50 Mbps)	25.0	11250	0.5	169
			1.0	338
			2.0	675
			4.0	1350
8K, 90 FPS, Virtual Reality (Rates will exceed 200 Mbps)	200.0	90000	0.5	1350
			1.0	2700
			2.0	5400
			4.0	10800
6 Degrees Freedom VR (Rates will range 200 to 1,000 Mbps)	500.0	225000	0.5	3375
			1.0	6750
			2.0	13500
			4.0	27000

Rysavy Research

---

[119] Rysavy Research analysis. For virtual reality-data requirements, refer to ABI Research/Qualcomm, *Augmented and Virtual Reality: the First Wave of 5G Killer Apps*, 2017. See also Netflix discussion of usage, "How can I control how much data Netflix uses?" https://help.netflix.com/en/node/87. Viewed May 3, 2016.

Spectrum Bands

3GPP technologies operate in a wide range of radio bands. As new spectrum becomes available, 3GPP updates its specifications for these bands. Although the support of a new frequency band may be introduced in a particular release, 3GPP specifies ways to implement devices and infrastructure operating on any frequency band, according to releases previous to the introduction of that particular frequency band. For example, although band 5 (US Cellular Band) was introduced in Release 6, the first devices operating on this band were compliant with the release 5 of the standard.

Table 16 shows the UMTS FDD bands.

Table 16: UMTS FDD Bands[120]

Operating Band	UL Frequencies UE Transmit, Node B Receive	DL Frequencies UE Receive, Node B Transmit
I	1920 - 1980 MHz	2110 -2170 MHz
II	1850 -1910 MHz	1930 -1990 MHz
III	1710-1785 MHz	1805-1880 MHz
IV	1710-1755 MHz	2110-2155 MHz
V	824 - 849MHz	869-894MHz
VI	830-840 MHz	875-885 MHz
VII	2500 - 2570 MHz	2620 - 2690 MHz
VIII	880 - 915 MHz	925 - 960 MHz
IX	1749.9 - 1784.9 MHz	1844.9 - 1879.9 MHz
X	1710-1770 MHz	2110-2170 MHz
XI	1427.9 - 1447.9 MHz	1475.9 - 1495.9 MHz
XII	699 - 716 MHz	729 - 746 MHz
XIII	777 - 787 MHz	746 - 756 MHz
XIV	788 - 798 MHz	758 - 768 MHz
XV	Reserved	Reserved
XVI	Reserved	Reserved
XVII	Reserved	Reserved
XVIII	Reserved	Reserved
XIX	830 – 845 MHz	875 -890 MHz
XX	832 - 862 MHz	791 - 821 MHz
XXI	1447.9 - 1462.9 MHz	1495.9 - 1510.9 MHz
XXII	3410 – 3490 MHz	3510 – 3590 MHz
XXV	1850 -1915 MHz	1930 -1995 MHz
XXVI	814-849 MHz	859-894 MHz
XXXII (NOTE 1)	N/A	1452 – 1496 MHz
NOTE 1: Restricted to UTRA operation when dual-band is configured (e.g., DB-DC-HSDPA or dual band 4C-HSDPA). The down link frequenc(ies) of this band are paired with the uplink frequenc(ies) of the other FDD band (external) of the dual band configuration.		

---

[120] 3GPP, *Base Station (BS) radio transmission and reception (FDD) (Release 14)*, December 2016, Technical Specification 25.104, V14.1.0.

Universal Mobile Telecommunications System (UMTS) Time Division Duplex (TDD) bands are the same as the LTE TDD bands. Table 17 details the LTE Frequency Division Duplex (FDD) and TDD bands.

Table 17: LTE FDD and TDD bands[121]

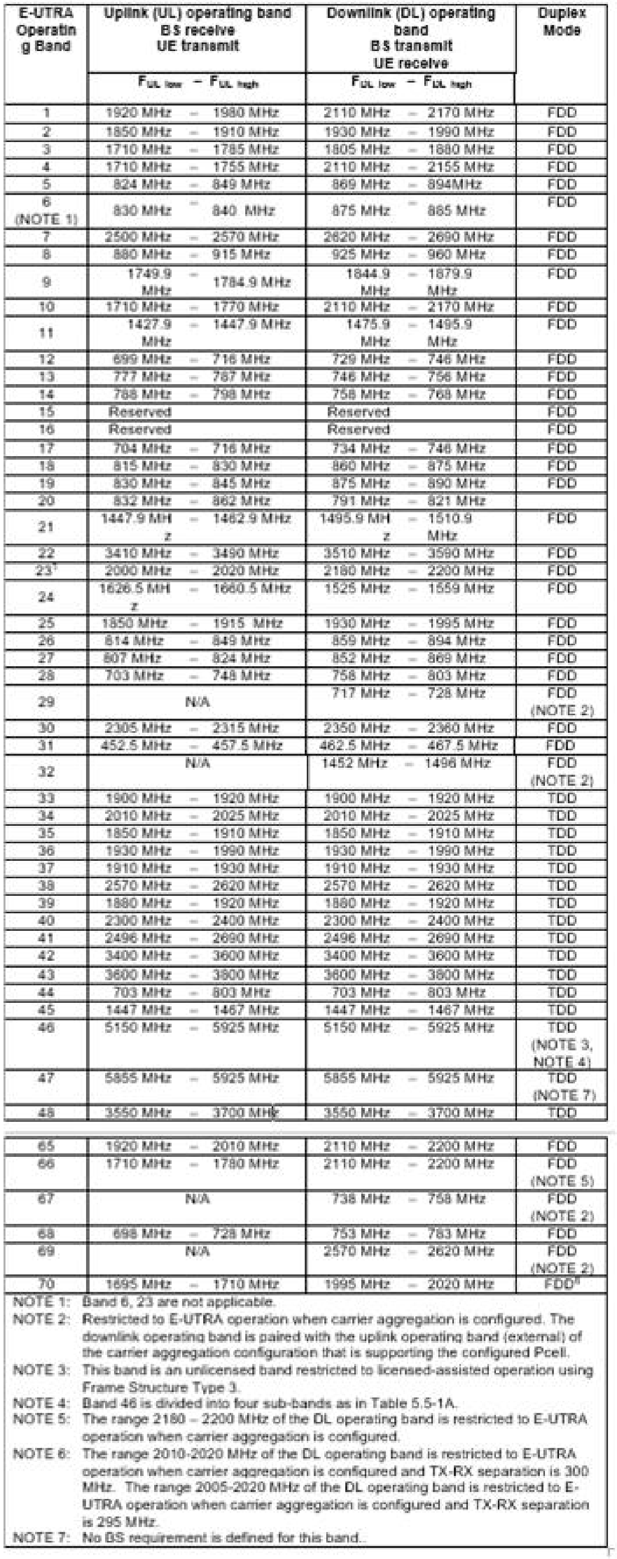

E-UTRA Operating Band	Uplink (UL) operating band BS receive UE transmit	Downlink (DL) operating band BS transmit UE receive	Duplex Mode
	$F_{UL\ low}$ – $F_{UL\ high}$	$F_{DL\ low}$ – $F_{DL\ high}$	
1	1920 MHz – 1980 MHz	2110 MHz – 2170 MHz	FDD
2	1850 MHz – 1910 MHz	1930 MHz – 1990 MHz	FDD
3	1710 MHz – 1785 MHz	1805 MHz – 1880 MHz	FDD
4	1710 MHz – 1755 MHz	2110 MHz – 2155 MHz	FDD
5	824 MHz – 849 MHz	869 MHz – 894MHz	FDD
6 (NOTE 1)	830 MHz – 840 MHz	875 MHz – 885 MHz	FDD
7	2500 MHz – 2570 MHz	2620 MHz – 2690 MHz	FDD
8	880 MHz – 915 MHz	925 MHz – 960 MHz	FDD
9	1749.9 MHz – 1784.9 MHz	1844.9 MHz – 1879.9 MHz	FDD
10	1710 MHz – 1770 MHz	2110 MHz – 2170 MHz	FDD
11	1427.9 MHz – 1447.9 MHz	1475.9 MHz – 1495.9 MHz	FDD
12	699 MHz – 716 MHz	729 MHz – 746 MHz	FDD
13	777 MHz – 787 MHz	746 MHz – 756 MHz	FDD
14	788 MHz – 798 MHz	758 MHz – 768 MHz	FDD
15	Reserved	Reserved	FDD
16	Reserved	Reserved	FDD
17	704 MHz – 716 MHz	734 MHz – 746 MHz	FDD
18	815 MHz – 830 MHz	860 MHz – 875 MHz	FDD
19	830 MHz – 845 MHz	875 MHz – 890 MHz	FDD
20	832 MHz – 862 MHz	791 MHz – 821 MHz	
21	1447.9 MHz – 1462.9 MHz	1495.9 MHz – 1510.9 MHz	FDD
22	3410 MHz – 3490 MHz	3510 MHz – 3590 MHz	FDD
23[1]	2000 MHz – 2020 MHz	2180 MHz – 2200 MHz	FDD
24	1626.5 MHz – 1660.5 MHz	1525 MHz – 1559 MHz	FDD
25	1850 MHz – 1915 MHz	1930 MHz – 1995 MHz	FDD
26	814 MHz – 849 MHz	859 MHz – 894 MHz	FDD
27	807 MHz – 824 MHz	852 MHz – 869 MHz	FDD
28	703 MHz – 748 MHz	758 MHz – 803 MHz	FDD
29	N/A	717 MHz – 728 MHz	FDD (NOTE 2)
30	2305 MHz – 2315 MHz	2350 MHz – 2360 MHz	FDD
31	452.5 MHz – 457.5 MHz	462.5 MHz – 467.5 MHz	FDD
32	N/A	1452 MHz – 1496 MHz	FDD (NOTE 2)
33	1900 MHz – 1920 MHz	1900 MHz – 1920 MHz	TDD
34	2010 MHz – 2025 MHz	2010 MHz – 2025 MHz	TDD
35	1850 MHz – 1910 MHz	1850 MHz – 1910 MHz	TDD
36	1930 MHz – 1990 MHz	1930 MHz – 1990 MHz	TDD
37	1910 MHz – 1930 MHz	1910 MHz – 1930 MHz	TDD
38	2570 MHz – 2620 MHz	2570 MHz – 2620 MHz	TDD
39	1880 MHz – 1920 MHz	1880 MHz – 1920 MHz	TDD
40	2300 MHz – 2400 MHz	2300 MHz – 2400 MHz	TDD
41	2496 MHz – 2690 MHz	2496 MHz – 2690 MHz	TDD
42	3400 MHz – 3600 MHz	3400 MHz – 3600 MHz	TDD
43	3600 MHz – 3800 MHz	3600 MHz – 3800 MHz	TDD
44	703 MHz – 803 MHz	703 MHz – 803 MHz	TDD
45	1447 MHz – 1467 MHz	1447 MHz – 1467 MHz	TDD
46	5150 MHz – 5925 MHz	5150 MHz – 5925 MHz	TDD (NOTE 3, NOTE 4)
47	5855 MHz – 5925 MHz	5855 MHz – 5925 MHz	TDD (NOTE 7)
48	3550 MHz – 3700 MHz	3550 MHz – 3700 MHz	TDD
65	1920 MHz – 2010 MHz	2110 MHz – 2200 MHz	FDD
66	1710 MHz – 1780 MHz	2110 MHz – 2200 MHz	FDD (NOTE 5)
67	N/A	738 MHz – 758 MHz	FDD (NOTE 2)
68	698 MHz – 728 MHz	753 MHz – 783 MHz	FDD
69	N/A	2570 MHz – 2620 MHz	FDD (NOTE 2)
70	1695 MHz – 1710 MHz	1995 MHz – 2020 MHz	FDD[6]

NOTE 1: Band 6, 23 are not applicable.

NOTE 2: Restricted to E-UTRA operation when carrier aggregation is configured. The downlink operating band is paired with the uplink operating band (external) of the carrier aggregation configuration that is supporting the configured Pcell.

NOTE 3: This band is an unlicensed band restricted to licensed-assisted operation using Frame Structure Type 3.

NOTE 4: Band 46 is divided into four sub-bands as in Table 5.5-1A.

NOTE 5: The range 2180 – 2200 MHz of the DL operating band is restricted to E-UTRA operation when carrier aggregation is configured.

NOTE 6: The range 2010-2020 MHz of the DL operating band is restricted to E-UTRA operation when carrier aggregation is configured and TX-RX separation is 300 MHz. The range 2005-2020 MHz of the DL operating band is restricted to E-UTRA operation when carrier aggregation is configured and TX-RX separation is 295 MHz.

NOTE 7: No BS requirement is defined for this band..

5G

This section provides early details on aspects of 5G, including architecture, LTE-NR coexistence, integrated access and backhaul, and performance.

Architecture

The overall 5G architecture consists of what 3GPP calls the New Generation Radio-Access Network (NG-RAN) and the 5G Core (5GC), as shown in Figure 44. The figure shows the Access and Mobility Management Function (AMF); the User Plane Function (UPF); the NR NodeB (gNB), which is the 5G base station; and the NG and Xn interfaces.

Figure 44: 5G Architecture[122]

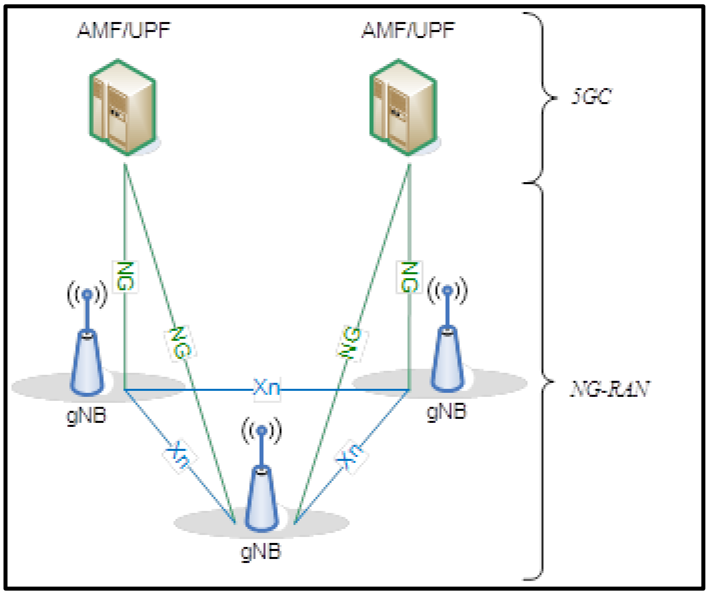

Figure 45 shows the functional split between the NG-RAN and 5GC.

---

[121] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception (Release 14)*, March 2017, Technical Specification 36.104, V14.3.0.

[122] 3GPP, *3GPP TS 38.300, NR; NR and NG-RAN Overall Description; Stage 2 (Release 15), V0.3.0 (2017-05).*

Figure 45: Functional Split between NG-RAN and 5GC[123]

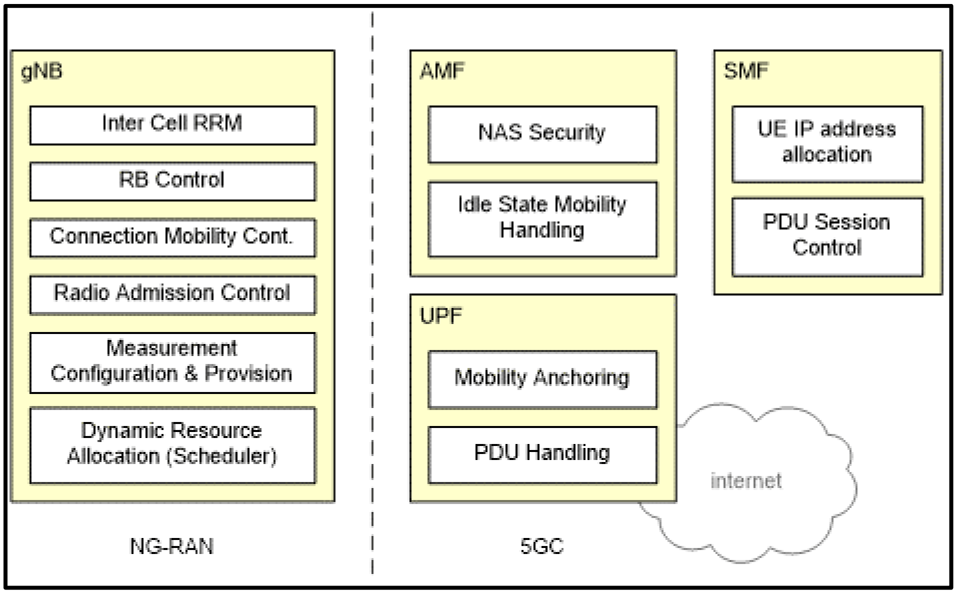
gNB
Inter Cell RRM
RB Control
Connection Mobility Cont.
Radio Admission Control
Measurement Configuration & Provision
Dynamic Resource Allocation (Scheduler)
AMF
NAS Security
Idle State Mobility Handling
SMF
UE IP address allocation
PDU Session Control
UPF
Mobility Anchoring
PDU Handling
internet
NG-RAN
5GC

The main body of this paper summarizes the features being specified in Releases 15 and 16 for NR and the core network. Additional capabilities that will be part of Release 15 include:

- A PDCP packet duplication function to allow redundant transmission of signaling or user data on two bearer paths.
- A new protocol layer called Service Data Adaptation Protocol (SDAP) that offers 5GC QoS flows.
- A new Radio Resource Control (RRC) inactive state designed for low-latency communications.
- A new system information broadcast model that allows on-demand system information instead of always having to broadcast system information (to reduce overhead in 5G beam sweeping).

Architecture Options

This topic was introduced in the main part of the paper and is covered here in more detail. In Release 15, 3GPP defines a number of different architecture options, shown in the following three figures. In many of these options, although not all, the 5G network integrates with LTE.

---

[123] Ibid.

Figure 46: 5G Network Architecture Options in 3GPP Release 15[124]

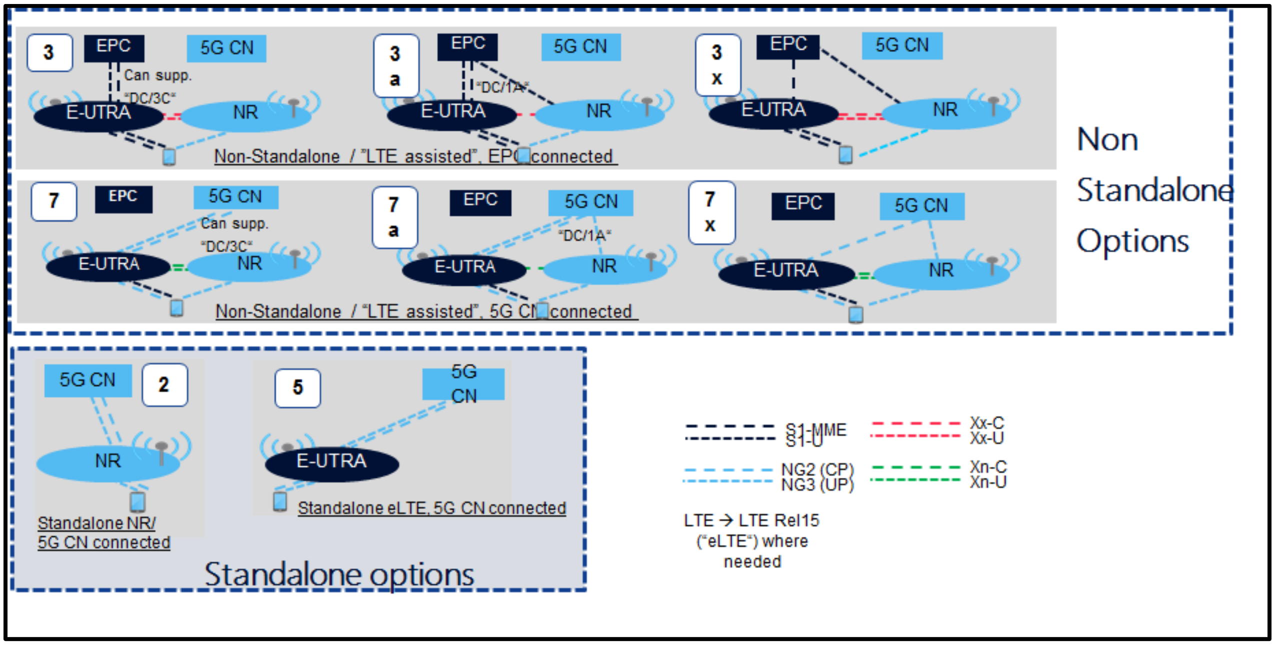
3
EPC
5G CN
Can supp.
"DC/3C"
E-UTRA
NR
3a
EPC
5G CN
"DC/1A"
E-UTRA
NR
3x
EPC
5G CN
E-UTRA
NR
Non-Standalone / "LTE assisted", EPC connected
7
EPC
5G CN
Can supp.
"DC/3C"
E-UTRA
NR
7a
EPC
5G CN
"DC/1A"
E-UTRA
NR
7x
EPC
5G CN
E-UTRA
NR
Non-Standalone / "LTE assisted", 5G CN connected
Non Standalone Options
5G CN
2
NR
Standalone NR/ 5G CN connected
5
5G CN
E-UTRA
Standalone eLTE, 5G CN connected
Standalone options
S1-MME
S1-U
Xx-C
Xx-U
NG2 (CP)
NG3 (UP)
Xn-C
Xn-U
LTE → LTE Rel15 ("eLTE") where needed

Figure 47: De-Prioritized 5G Network Architecture Options in 3GPP Release 15

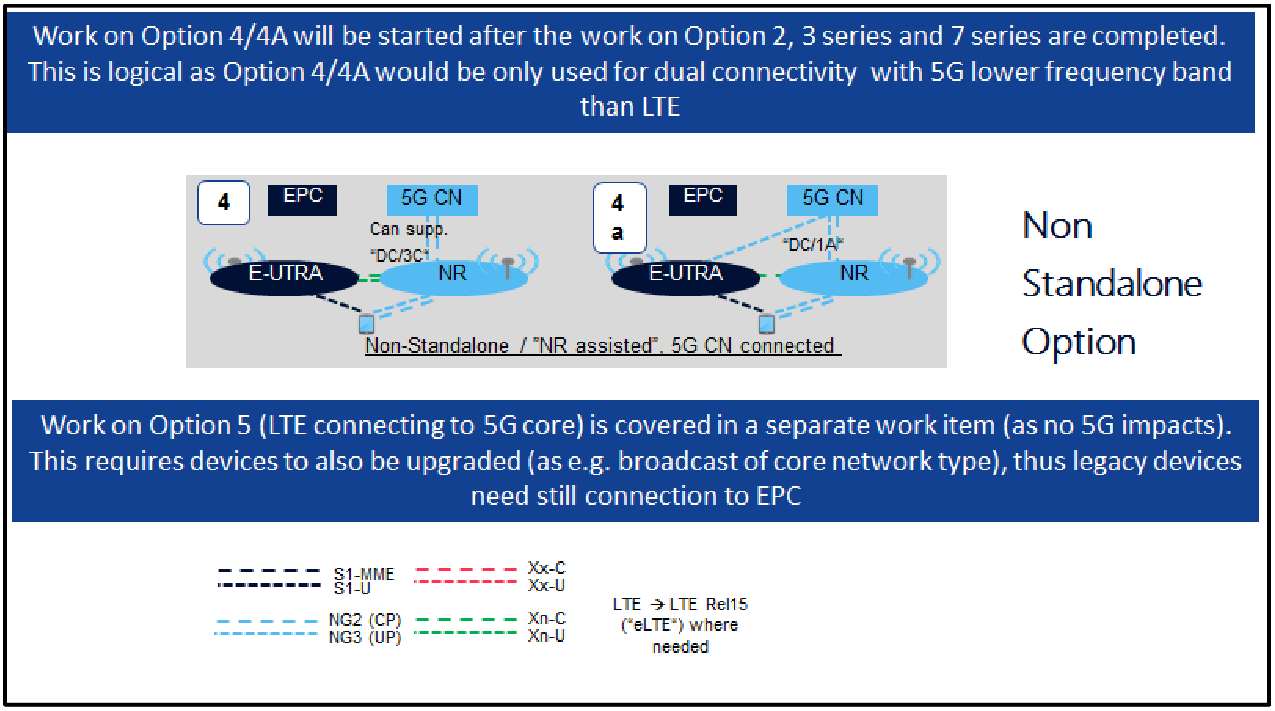

Work on Option 4/4A will be started after the work on Option 2, 3 series and 7 series are completed. This is logical as Option 4/4A would be only used for dual connectivity with 5G lower frequency band than LTE

4
EPC
5G CN
Can supp.
"DC/3C"
E-UTRA
NR
4a
EPC
5G CN
"DC/1A"
E-UTRA
NR
Non-Standalone / "NR assisted", 5G CN connected
Non Standalone Option

Work on Option 5 (LTE connecting to 5G core) is covered in a separate work item (as no 5G impacts). This requires devices to also be upgraded (as e.g. broadcast of core network type), thus legacy devices need still connection to EPC

S1-MME
S1-U
Xx-C
Xx-U
NG2 (CP)
NG3 (UP)
Xn-C
Xn-U
LTE → LTE Rel15 ("eLTE") where needed

---

[124] Nokia contribution, including subsequent two figures. For further details, refer to section 7.2, "5G Architecture Options," 3GPP TR 38.801, "Radio access architecture and interfaces."

Figure 48: Dual-Connectivity Options with LTE as Master

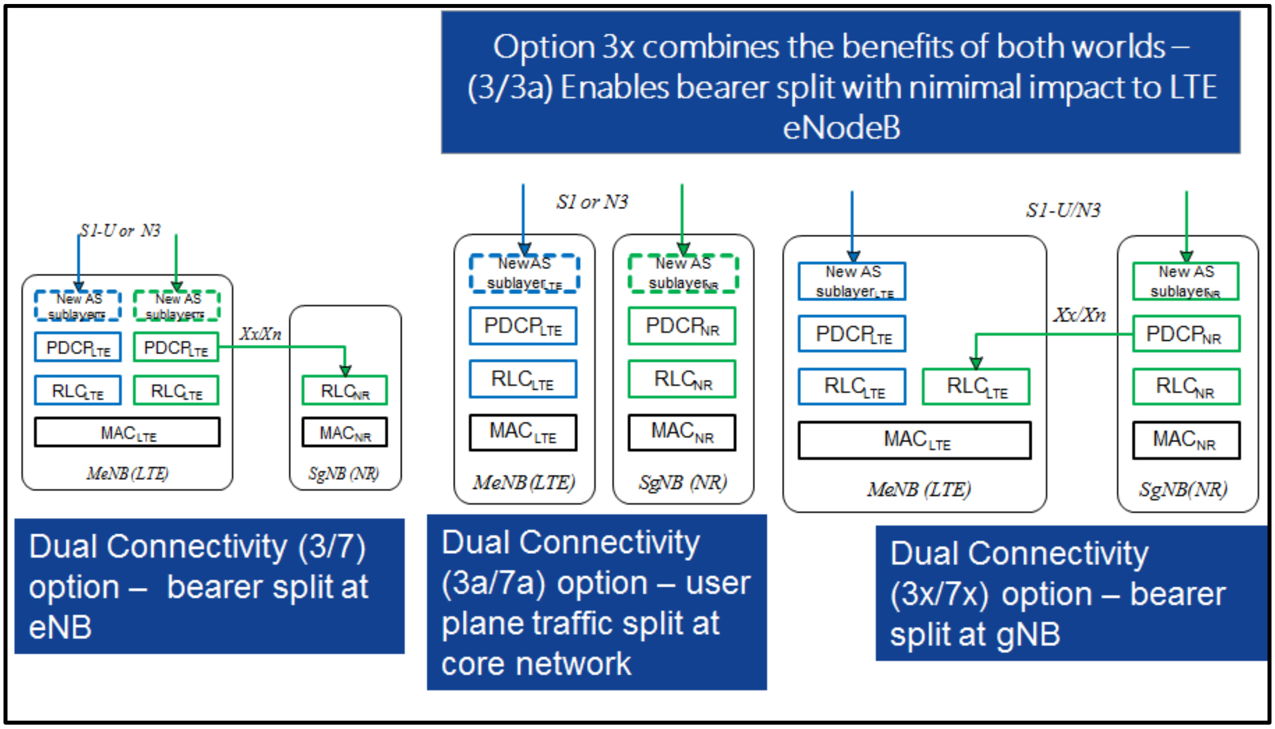
S1-U or N3
New AS sublayer$_{LTE}$
New AS sublayer$_{LTE}$
PDCP$_{LTE}$
PDCP$_{LTE}$
Xx/Xn
RLC$_{LTE}$
RLC$_{LTE}$
RLC$_{NR}$
MAC$_{LTE}$
MAC$_{NR}$
MeNB (LTE)
SgNB (NR)
S1 or N3
New AS sublayer$_{LTE}$
New AS sublayer$_{NR}$
PDCP$_{LTE}$
PDCP$_{NR}$
RLC$_{LTE}$
RLC$_{NR}$
MAC$_{LTE}$
MAC$_{NR}$
MeNB (LTE)
SgNB (NR)
S1-U/N3
New AS sublayer$_{LTE}$
New AS sublayer$_{NR}$
Xx/Xn
PDCP$_{LTE}$
PDCP$_{NR}$
RLC$_{LTE}$
RLC$_{LTE}$
RLC$_{NR}$
MAC$_{LTE}$
MAC$_{NR}$
MeNB (LTE)
SgNB(NR)

Option 3x combines the benefits of both worlds – (3/3a) Enables bearer split with nimimal impact to LTE eNodeB

Dual Connectivity (3/7) option – bearer split at eNB

Dual Connectivity (3a/7a) option – user plane traffic split at core network

Dual Connectivity (3x/7x) option – bearer split at gNB

LTE-NR Coexistence

LTE-NR coexistence is a Release 15 work item. This section describes how such co-existence may be achieved. Different LTE-NR co-existence cases include the following: time domain LTE/NR adjacent channel coexistence; LTE Secondary Cell on/off for LTE/NR adaptation; in-carrier LTE+NR coexistence in downlink, and in-carrier LTE+NR coexistence in uplink.

NR coexistence is required for LTE UEs of all releases. Because carrier aggregation was not introduced into LTE until LTE Release 10, CA-based techniques cannot be used as the sole means to achieve LTE/NR coexistence. However, CA techniques can be used for both time domain coexistence and frequency domain coexistence. For time domain coexistence, on a given carrier, LTE and NR are time-multiplexed by means of Secondary Cell (SCell) activation or deactivation. For frequency domain coexistence, the network configures a carrier, such as a 20MHz carrier, into multiple carriers, with, for example, 10MHz allocated to LTE and the remaining 10MHz to NR. Note that frequency domain coexistence can be accomplished without using carrier aggregation.

Figure 49 illustrates the frequency domain technique. Note that when splitting the 20MHz carrier into two allocations of 10MHz, the LTE carrier remains centered at the same frequency and the NR allocation is not consecutive.

Figure 49: Frequency Domain Coexistence of LTE and NR[125]

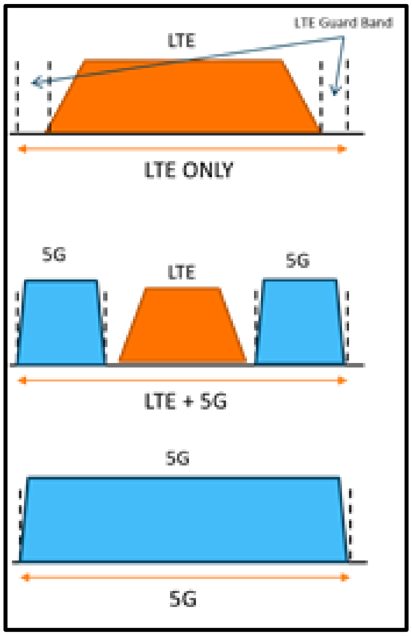

Time Domain LTE/NR Coexistence Techniques

Time domain coexistence of LTE and NR can be dynamic (subframe level) or semi-static (MAC/RRC). In the latter case, spectrum resources are configured as SCell for a LTE UE, and the network can turn these resources on or off by means of SCell activation or deactivation using MAC control elements or by adding and removing the SCell via RRC signaling. Whenever the SCell is deactivated or not configured, the spectrum resources can be used for NR transmissions. In LTE Rel. 12, small-cell enhancements were introduced that allow a UE to measure discovery reference signals (DRS) on a deactivated SCell. In that case, further coordination between LTE and NR may be required even when the SCell is deactivated, as DRS may still be transmitted periodically. Generally, though, this kind of coexistence can be achieved by network implementation.

For the case of dynamic coexistence, LTE and NR co-exist in the same spectrum, and the network can multiplex the two on a subframe level. Because LTE transmits Cell-Specific Reference Signals (CRS) in all DL subframes and in the Downlink Pilot Time Slot (DwPTS) and non-Multicast-Broadcast Single-Frequency network (non-MBSFN) region of special and MBSFN subframes, respectively, dynamic coexistence is not as straightforward as semi-static coexistence.

Similar to the case of time domain coexistence based on CA techniques, whenever OFDM symbols do not carry CRS, coexistence can be achieved by a gNB scheduler implementation. In particular, the gNB can schedule mini-slot-based transmissions in the Uplink Pilot Time Slot (UpPTS) region of a special subframe and in the MBSFN region of an MBSFN subframe, neither of which carry CRS. In LTE UL subframes, the gNB can schedule NR transmissions using either slots or mini-slots. For example, when Sounding Reference Signal (SRS) is transmitted at the end of a subframe, all 14 OFDM symbols may not be available for NR, and mini-slots can be used. Otherwise, slots can be used to transmit NR signals and channels in LTE UL subframes. Even in normal downlink subframes, mini-slots could be used to transmit NR channels and signals on OFDM symbols not carrying CRS. This, however, may leave almost half the resources of a normal DL subframe unusable for NR, so other techniques may be preferable. For example, symbols carrying CRS could also puncture NR transmissions, similar to URLLC transmissions that pre-empt NR transmissions. The same mechanisms specified for eMBB/URLLC coexistence could be used for LTE/NR coexistence.

---

[125] AT&T contribution, including explanatory text.

Frequency Domain LTE/NR Coexistence Techniques

Frequency domain coexistence between LTE and NR can also be dynamic or semi-static. Semi-static FDM-based coexistence is illustrated in Figure 49. Dynamic frequency domain coexistence is possible when the (e/g) NB schedules both LTE and NR in the same subframe on a Physical Resource Block (PRB) level.

There also exists the possibility of mixing semi-static and dynamic schemes as well as time division multiplexing (TDM)- and frequency division multiplexing (FDM)-based schemes based on the duplex direction. UL resources could be dynamically shared in a TDM fashion, whereas DL resources would be semi-statically configured and frequency division multiplexed between LTE and NR. For example, LTE could operate in paired spectrum, and NR could use LTE UL resources for NR UL transmissions but be configured with a separate DL or dynamic TDD carrier, such as at a higher frequency band. In this scenario, the LTE DL would be semi-statically frequency division multiplexed with NR, but LTE UL resources would need to be dynamically shared between LTE and NR. The semi-statically frequency division multiplexed NR resources could be for DL only or for both DL and UL. For example, it could be beneficial to allow for NR SRS transmissions on the frequency division multiplexed NR-only carrier.

Several issues need to be addressed for the shared LTE UL carrier. For example, if the non-shared NR carrier operates in mmWave spectrum while the shared NR/LTE carrier operates below 6GHz, the UE does not receive NR DL signals that can be used for power control and timing advance of the NR UL transmissions in the shared LTE UL resources. In this case, NR signals may have to be sent in the LTE-only DL resources or, alternatively, the NR-only UE needs to receive and process LTE signals in the LTE-only DL carrier. To avoid NR UEs processing LTE signals or LTE eNBs transmitting NR signals, 3GPP will need to investigate whether the aforementioned problem could be solved by signaling mechanisms. Regardless, further studies are needed to address these issues.

Coordination Requirements for LTE/NR Coexistence

While semi-static techniques identified for coexistence may require minimal coordination, dynamic (for example, per-TTI) sharing can be done by coordinating the LTE and NR transmissions via three different mechanisms:

1. Co-locating the NR and LTE scheduling.
2. Via the X2 interface (or the evolved version of the X2 interface in the new RAN architecture).
3. Over-the-air.

Options 1 and 2 do not impact any RAN1 specification, whereas Option 3 requires RAN1 specifications. Also, over-the-air coordination is desirable because it does not require LTE and NR scheduling and transmission to be handled by a single eNodeB, nor does it require an ultra-low-latency transport between them, thereby providing much more deployment flexibility. This can even allow NR and LTE to be deployed on different tiers (for example, macro and pico) and share the same channel.

Integrated Access and Backhaul

A study item for Release 15 with the likelihood of standardization in Release 16 is integrated access and backhaul, with which the 5G radio can provide both access and backhaul functions, as shown in Figure 50.

Figure 50: 5G Integrated Access and Backhaul

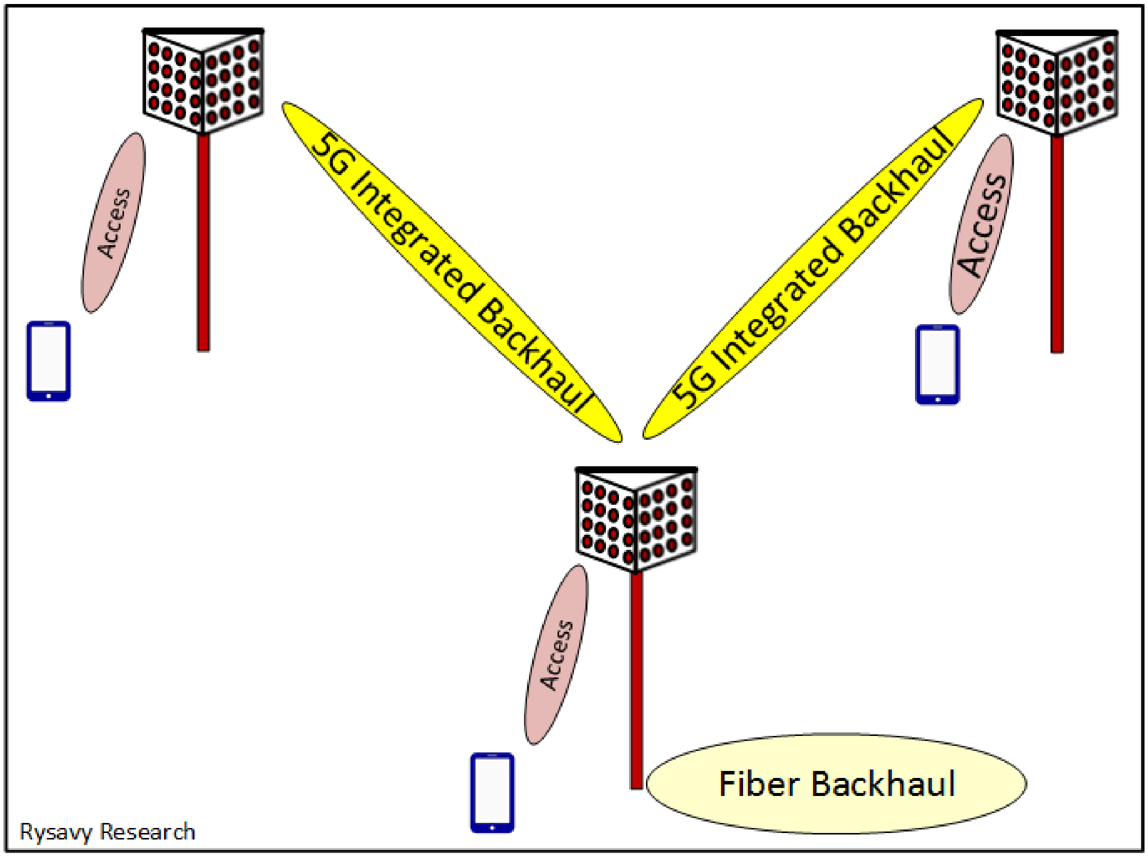

This 3GPP Study item, which has not yet begun, could address the following capabilities:[126]

- ❑ Wireless backhaul using 5G or LTE.
- ❑ Both indoor and outdoor scenarios.
- ❑ Flexible partitioning between access and backhaul functions.
- ❑ Autonomous configuration of access and wireless self-backhaul functions.
- ❑ Multi-hop wireless self-backhauling to enable flexible extension of range and coverage area.

---

[126] 3GPP, “Study on Integrated Access and Backhaul for NR,” RP-170831, March 2017.

- Autonomous adaptation on wireless self-backhaul to minimize service disruptions.
- Topological redundant connectivity on the wireless self-backhaul to enhance reliability and capacity and to reduce latency.

Performance

5G, with the ability to use wider radio channels than LTE, will be able to deliver much higher peak and average speeds, with initial estimates listed above in the section, “Data Throughput Comparison.” In the absence of deployed networks to measure, companies have performed simulations, concentrating initially on one of the first uses cases of fixed wireless access.

Figure 51 shows downlink performance for a network using different base station ISDs, and with and without foliage. The Nokia simulation used base stations with 512 antenna elements, outdoor-mounted user equipment, 28 GHz, and an 800-MHz radio channel, mostly allocated to the downlink.

Figure 51: Downlink Performance, Different ISDs, Foliage vs. None[127]

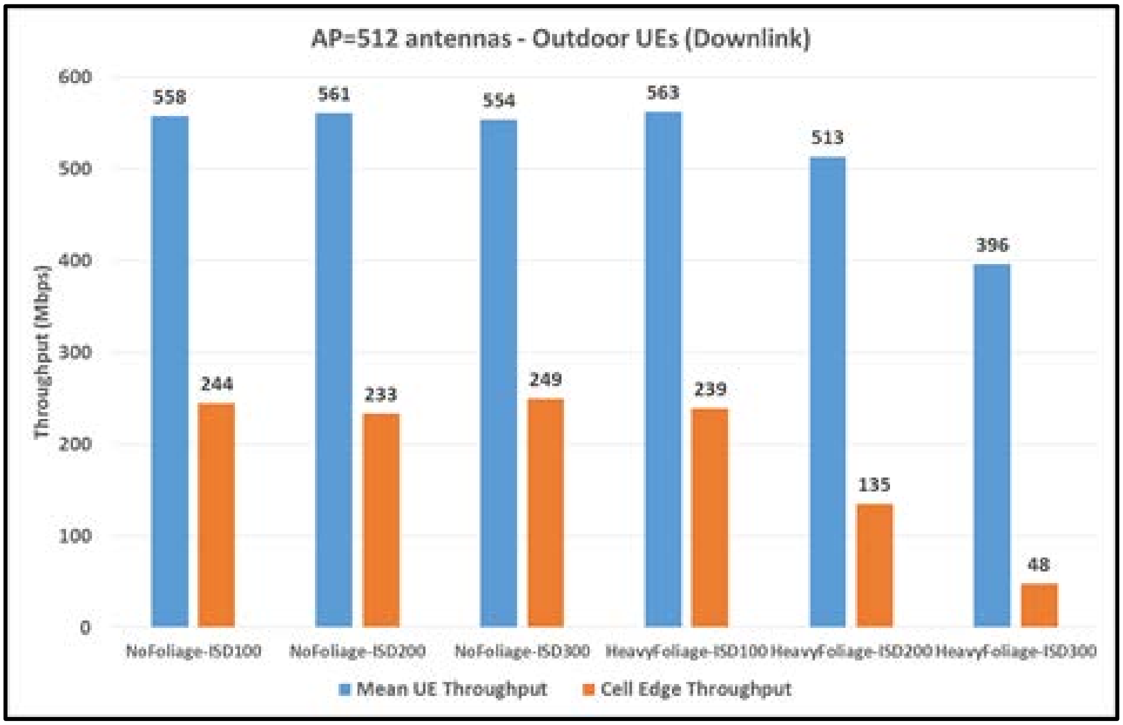

The conclusion of this simulation is that a minimum performance of 100 Mbps at the cell-edge, a 5G objective, is possible at ISDs up to 200 meters, with and without foliage.

The following three figures are from another simulation study by Ericsson, this one also for fixed-wireless access, with the following key assumptions: 350-meter ISD, 96-antenna base stations, 200 MHz radio channels, 57% allocated to downlink, 1000 homes per sq.

---

[127] Nokia contribution. For a full discussion, refer to the associated paper by Frederick W. Vook, Eugene Visotsky, Timothy A. Thomas, and Amitava Ghosh, Nokia Bell Labs, *Performance Characteristics of 5G mmWave Wireless-to-the-Home*, available at http://ieeexplore.ieee.org/document/7869558/.

km., 25% of homes using 4K UHD video service at 15 Mbps, building heights of 4 to 10 meters, and trees from 5 to 15 meters.

Figure 52 shows the throughputs available across the coverage area, with many locations able to receive close to 1 Gbps.

Figure 52: Throughput Map of Suburban Area at Low Load[128]

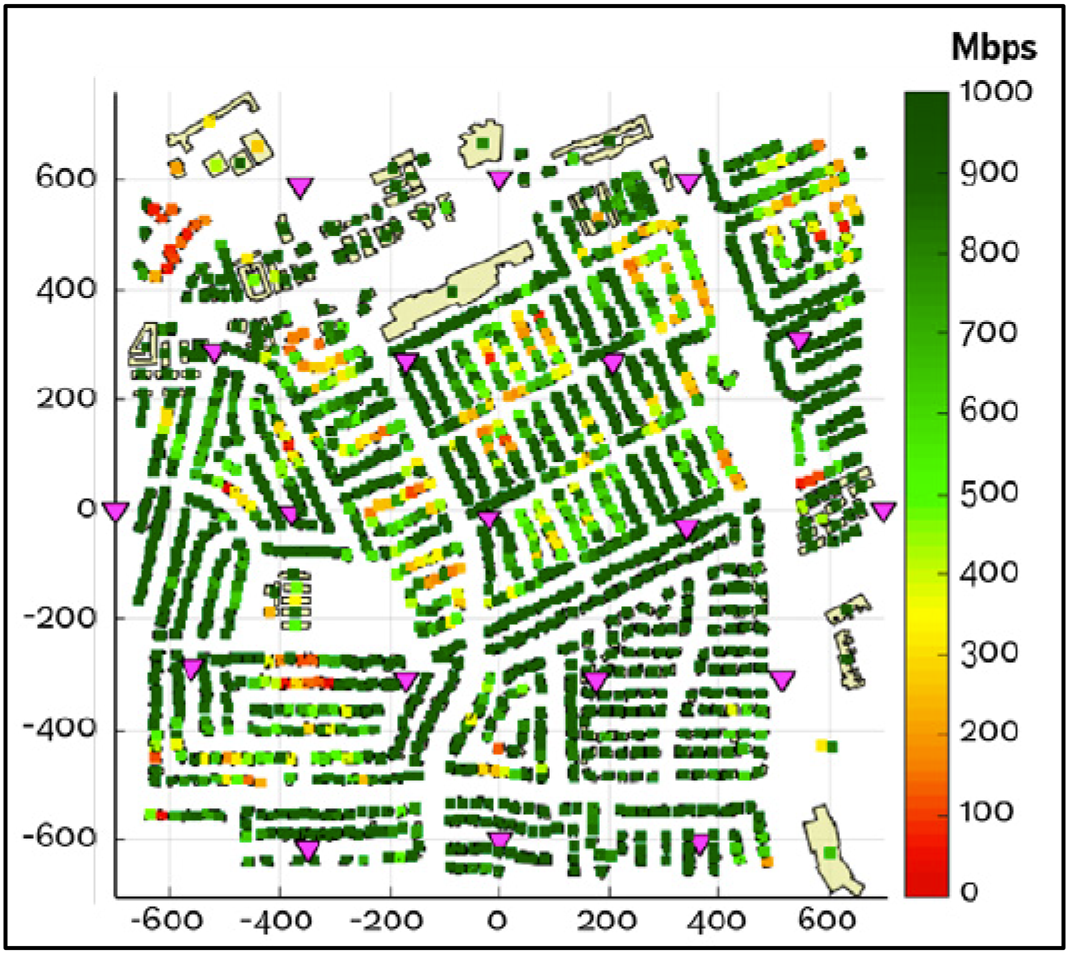
Mbps
1000
900
800
700
600
500
400
300
200
100
0
600
400
200
0
-200
-400
-600
-600
-400
-200
0
200
400
600

Figure 53 shows the proportion of users that can obtain 15 Mbps and 100 Mbps service relative to monthly traffic volume. Note that the system supports thousands of GBs of service per subscriber per month.

---

[128] Ericsson contribution, Ericsson Technical Review, *5G and Fixed Wireless Access*, 2016, available at https://www.ericsson.com/assets/local/publications/ericsson-technology-review/docs/2016/etr-5g-and-fixed-wireless-access.pdf.

Figure 53: Proportion of Satisfied Users Relative to Monthly Usage[129]

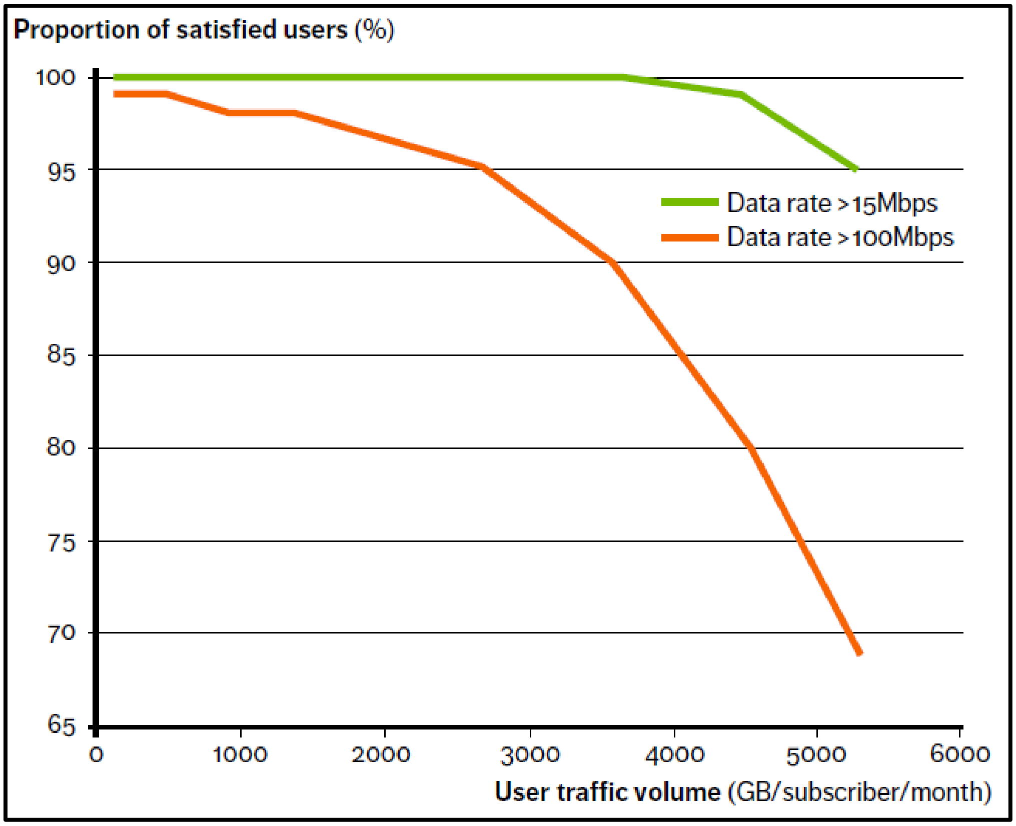

Figure 54 shows that an ISD of 350 can be used with a combination of indoor, wall-mounted, and rooftop antennas. A large percentage of users, 78%, can use indoor antennas, facilitating deployment.

Figure 54: Breakdown of Indoor, Wall-Mounted, and Rooftop Antennas[130]

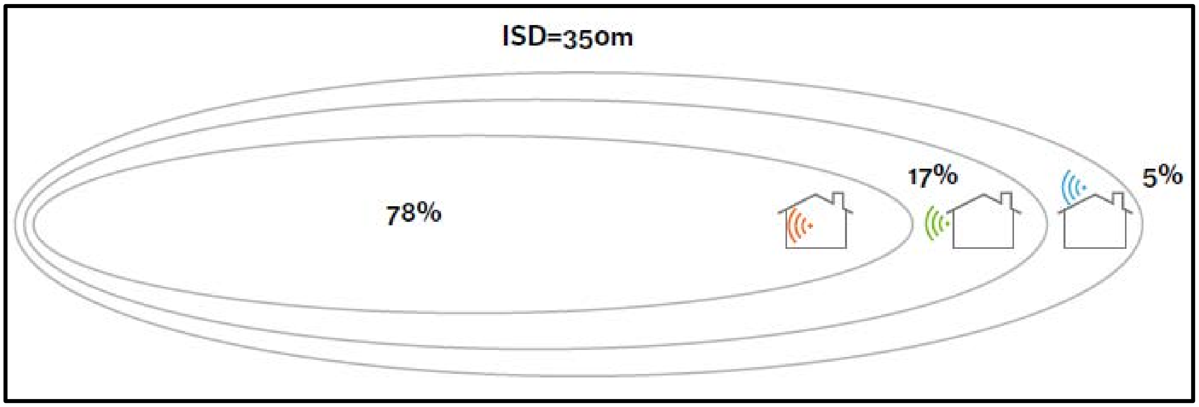

---

[129] Ibid.

[130] Ibid.

The simulation study shows that 5G fixed wireless access deployments using a larger ISD of 350 meters, translating to 11 base stations per square kilometer, can provide competitive broadband service.

In this environment, handsets with 5G mmWave capability will also be able to access the networks, but the antennas they use may not be as effective as the fixed-wireless equipment, so handsets may need to fall back to 4G, depending on their precise locations. For this reason, the dual-connectivity being planned for 5G will play an important role.

Figure 55 shows another simulation study, this one from Intel, using the following assumptions: 28 GHz operation, 2:1 DL:UL ratio, 25% control overhead, 10 bps/Hz maximum downlink spectral efficiency, CPEs placed either north or south side of house and one with best SNR chosen, and indoor CPE equipment with 30dB outdoor-to-indoor penetration loss. Scenario 1 is 60 access points per sq. km. Scenario 2 is 120 access points per sq. km. (Base grid of 40 houses in a 250x200m area with four rows of 10 houses per row, APs placed along streets and alleys, single-family homes, 4 sectors per AP, and 4.5-meter pole height.)

Figure 55: 5G Fixed Wireless Simulation with Different Loading and Densities[131]

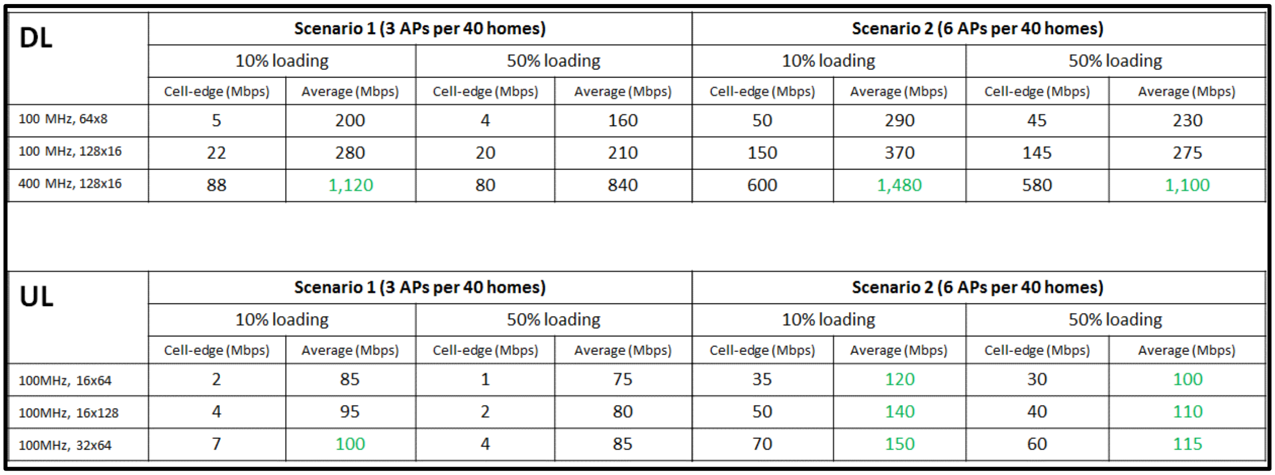

DL	**Scenario 1 (3 APs per 40 homes)**				**Scenario 2 (6 APs per 40 homes)**			
	10% loading		50% loading		10% loading		50% loading	
	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)
100 MHz, 64x8	5	200	4	160	50	290	45	230
100 MHz, 128x16	22	280	20	210	150	370	145	275
400 MHz, 128x16	88	1,120	80	840	600	1,480	580	1,100

UL	**Scenario 1 (3 APs per 40 homes)**				**Scenario 2 (6 APs per 40 homes)**			
	10% loading		50% loading		10% loading		50% loading	
	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)	Cell-edge (Mbps)	Average (Mbps)
100MHz, 16x64	2	85	1	75	35	120	30	100
100MHz, 16x128	4	95	2	80	50	140	40	110
100MHz, 32x64	7	100	4	85	70	150	60	115

Using 400 MHz and six access points per 40 homes, and 50% loading, the average throughput was more than 1 Gbps.

LTE and LTE-Advanced

Although HSPA and HSPA+ offer a highly efficient broadband-wireless service that will enjoy success for the remainder of this decade and well into the next, 3GPP completed the specification for Long Term Evolution as part of Release 8. LTE offers even higher peak throughputs in wider spectrum bandwidth. Work on LTE began in 2004 with an official work item started in 2006 and a completed specification early 2009. Initial deployments began in 2010.

LTE uses OFDMA on the downlink, which is well suited to achieve high peak data rates in high-spectrum bandwidth. WCDMA radio technology is basically as efficient as OFDM for delivering peak data rates of about 10 Mbps in 5 MHz of bandwidth. Achieving peak rates in the 100 Mbps range with wider radio channels, however, would result in highly complex terminals, and it is not practical with current technology, whereas OFDM provides a practical implementation advantage. Scheduling approaches in the frequency domain can also minimize interference, thereby boosting spectral efficiency. The OFDMA approach is also flexible in channelization: LTE operates in various radio channel sizes ranging from 1.4 to 20 MHz.

---

[131] Intel contribution.

On the uplink, however, a pure OFDMA approach results in high Peak to Average Ratio (PAR) of the signal, which compromises power efficiency and, ultimately, battery life. Hence, LTE uses an approach called “SC-FDMA,” which is somewhat similar to OFDMA, but has a 2 to 6 dB PAR advantage over the OFDMA method used by other technologies such as WiMAX.

LTE capabilities include:

- Downlink peak data rates up to 300 Mbps with 20+20 MHz bandwidth in initial versions, increasing to over 1 Gbps in subsequent versions through carrier aggregation, higher-order modulation, and 4X4 MIMO.
- Uplink peak data rates up to 71 Mbps with 20+20 MHz bandwidth in initial versions, increasing to over 1 Gbps in subsequent versions.
- Operation in both TDD and FDD modes.
- Scalable bandwidth up to 20+20 MHz covering 1.4, 3, 5, 10, 15, and 20 MHz radio carriers.
- Increased spectral efficiency over HSPA by a factor of two to four.
- Reduced latency, to 15 msec round-trip times between user equipment and the base station, and to less than 100 msec transition times from inactive to active.
- Self-organizing capabilities under operator control and preferences that will automate network planning and will result in lower operator costs.

LTE-Advanced Terminology

LTE-Advanced, as specified in Release 10, is a term used for the version of LTE that addresses IMT-Advanced requirements. The ITU ratified LTE-Advanced as IMT-Advanced in November 2010. LTE-Advanced is both backward- and forward-compatible with LTE, meaning LTE devices operate in newer LTE-Advanced networks, and LTE-Advanced devices operate in older, pre-Release 10 LTE networks.

The following lists at a high level the most important features of LTE-Advanced, as well as other features planned for subsequent releases, including Release 11:

- Carrier aggregation.
- Higher-order downlink MIMO (up to 8X8 in Release 10).
- Uplink MIMO (two transmit antennas in the device).
- Coordinated multipoint transmission (CoMP) in Release 11.
- Heterogeneous network (HetNet) support including Enhanced Inter-cell Interference Coordination (eICIC).
- Relays.

3GPP, from Release 13, has referred to LTE as LTE-Advanced Pro, which includes features such as LAA, LWA, low latency, and massive MIMO.

OFDMA and Scheduling

LTE implements OFDM in the downlink. The basic principle of OFDM is to split a high-rate data stream into a number of parallel, low-rate data streams, each a narrowband signal carried by a subcarrier. The different narrowband streams are generated in the frequency domain, and then combined to form the broadband stream using a mathematical algorithm called an "Inverse Fast Fourier Transform" (IFFT) that is implemented in digital signal processors. In LTE, the subcarriers have 15 kHz spacing from each other. LTE maintains this spacing regardless of the overall channel bandwidth, which simplifies radio design, especially in supporting radio channels of different widths. The number of subcarriers ranges from 72 in a 1.4 MHz radio channel to 1,200 in a 20 MHz radio channel.

The composite signal obtained after the IFFT is extended by repeating the initial part of the signal (called the Cyclic Prefix [CP]). This extended signal represents an OFDM symbol. The CP is basically a guard time during which reflected signals will reach the receiver. It results in an almost complete elimination of multipath-induced Intersymbol Interference (ISI), which otherwise makes extremely high data rate transmissions problematic. The system is called orthogonal because the subcarriers are generated in the frequency domain (making them inherently orthogonal), and the IFFT conserves that characteristic.

OFDM systems may lose their orthogonal nature as a result of the Doppler shift induced by the speed of the transmitter or the receiver. 3GPP specifically selected the subcarrier spacing of 15 kHz to avoid any performance degradation in high-speed conditions. WiMAX systems that use a lower subcarrier spacing (~11 kHz) are more impacted in high-speed conditions than LTE.

Figure 56: OFDM Symbol with Cyclic Prefix

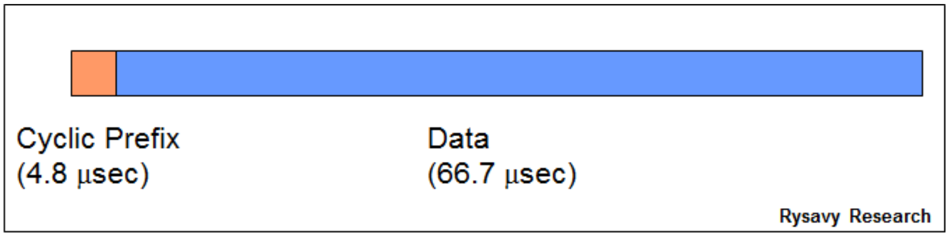

The multiple access aspect of OFDMA comes from being able to assign different users different subcarriers over time. A minimum resource block that the system can assign to a user transmission consists of 12 subcarriers over 14 symbols in 1.0 msec. Figure 57 shows how the system can assign these resource blocks to different users over both time and frequency.

Figure 57: LTE OFDMA Downlink Resource Assignment in Time and Frequency

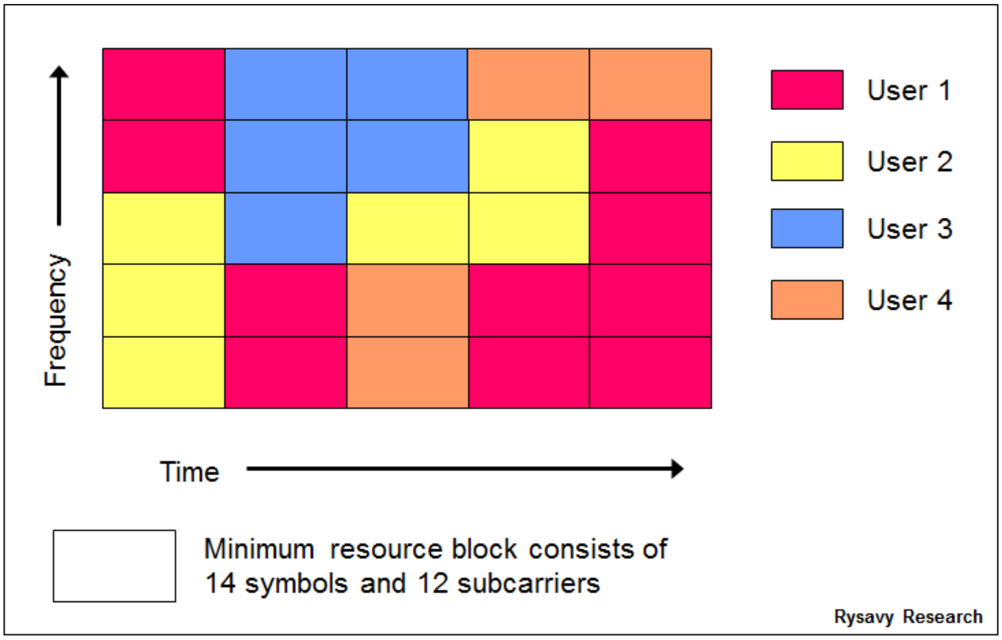
User 1
User 2
User 3
User 4
Frequency
Time
Minimum resource block consists of
14 symbols and 12 subcarriers
Rysavy Research

By controlling which subcarriers are assigned in which sectors, LTE can easily control frequency reuse. Using all the subcarriers in each sector, the system would operate at a frequency reuse of 1; but by using a different one third of the subcarriers in each sector, the system can achieve a looser frequency reuse of 1/3. The looser frequency reduces overall spectral efficiency but delivers high peak rates to users.

Beyond controlling frequency reuse, frequency domain scheduling, as shown in Figure 58 can use those resource blocks that are not faded, not possible in CDMA-based systems. Since different frequencies may fade differently for different users, the system can allocate those frequencies for each user that result in the greatest throughput. This results in up to a 40% gain in average cell throughput for low user speed (3 km/hour), assuming a large number of users and no MIMO. The benefit decreases at higher user speeds.

Figure 58: Frequency Domain Scheduling in LTE[132]

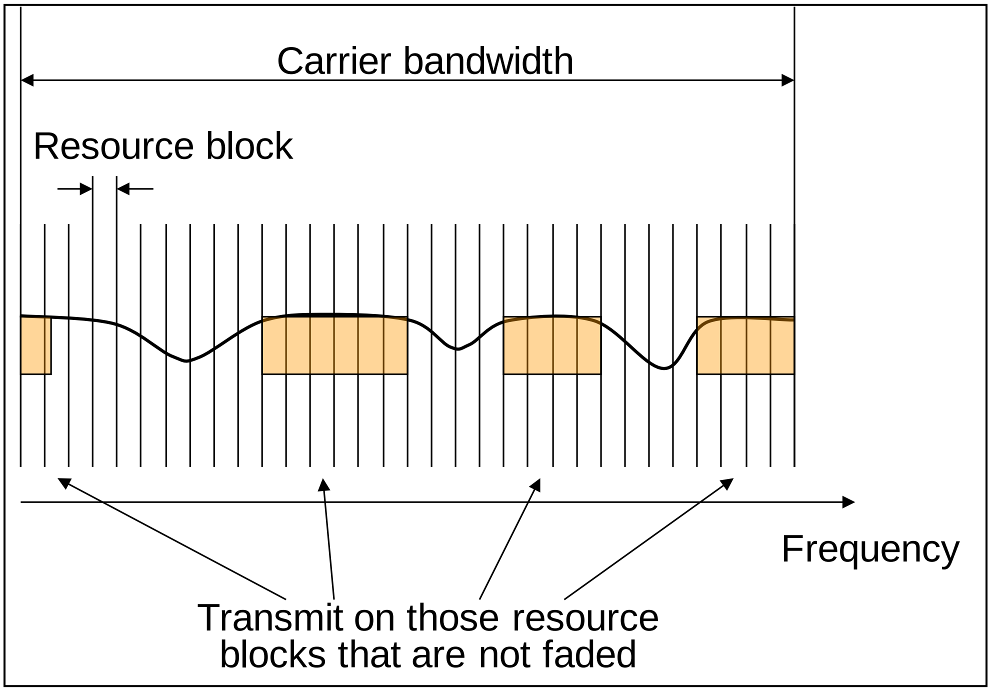

LTE Smart Antennas

Wireless networks can achieve significant gains by employing multiple antennas, either at the base station, the mobile device, or both. LTE uses multiple antennas in three fundamentally different ways:

- ❑ **Diversity.** So long as the antennas are spaced or polarized appropriately, the antennas provide protection against fading.
- ❑ **Beamforming.** Multiple antennas can shape a beam to increase the gain for a specific receiver. Beamforming can also suppress specific interfering signals. Beamforming is particularly helpful for improving cell-edge performance.
- ❑ **Spatial Multiplexing.** Often referred to as MIMO antenna processing, spatial multiplexing creates multiple transmission paths through the environment, effectively sending data in parallel through these paths, thus increasing both throughput and spectral efficiency.

Table 18 shows the varous antenna transmission modes.

---

[132] 5G Americas member contribution.

Table 18: LTE Transmission Modes[133]

Transmission Mode	Description
1	Single antenna transmission.
2	Transmit Diversity.
3	Transmit diversity for one layer, open-loop codebook-based precoding if more than one layer.
4	Closed-loop codebook-based precoding.
5	Multi-user MIMO version of transmission mode 4.
6	Special case of closed-loop codebook-based precoding limited to single layer transmission.
7	Beamforming. (Non-codebook-based precoding supporting one layer.)
8	Dual-layer beamforming. (Release 9. Non-codebook-based precoding supporting up to two layers.)
9	8-layer transmission. (Release 10. Non-codebook-based precoding supporting up to eight layers.)
10	8-layer transmission with support for CoMP. (Release 11.)

Being able to exploit different antenna modes based on local conditions produces huge efficiency and performance gains, and is the reason that 3GPP is developing even more advanced antenna modes in subsequent LTE releases.

Precoding refers to a mathematical matrix operation performed on radio symbols to determine how they are combined and mapped onto antenna ports. The precoder matrix can operate in either open-loop or closed-loop modes. For each transmission rank for a given number of transmission ports (antennas), there is a limited set of precoder matrices defined, called the codebook. This helps limit the amount of signaling needed on uplink and downlink.

Fundamental variables distinguish the different antenna modes:

- ❑ **Single base station antenna versus multiple antennas.** Single antennas provide for Single Input Single Output (SISO), SIMO, and planar-array beamforming. (Multiple Output means the UE has multiple antennas.) Multiple antennas at the base station provide for different MIMO modes such as 2X2, 4X2, and 4X4.

---

[133] Erik Dahlman, Stefan Parkvall, Johan Skold, *4G - LTE/LTE Advanced for Mobile Broadband*, Academic Press, 2011.

- **Single-user MIMO versus multi-user MIMO.** Release 8 only provides for single-user MIMO on the downlink. Release 10 includes multi-user MIMO.

- **Open-Loop versus Closed-Loop.** High vehicular speeds require open-loop operation whereas slow speeds enabled closed-loop operation in which feedback from the UE modifies the transmission. In closed-loop operation, the precoder matrix is based on this feedback.

- **Rank.** In a MIMO system, the channel rank is formally defined as the rank of the channel matrix and is a measure of the degree of scattering that the channel exhibits. For example, in a 2x2 MIMO system, a rank of one indicates a low-scattering environment, while a rank of two indicates a high-scattering environment. The rank two channel is highly uncorrelated, and is thus able to support the spatial multiplexing of two data streams, while a rank one channel is highly correlated, and thus can only support single stream transmission (the resulting multi-stream interference in a rank one channel as seen at the receiver would lead to degraded performance). Higher Signal to Interference plus Noise Ratios (SINR) are typically required to support spatial multiplexing, while lower SINRs are typically sufficient for single stream transmission. In a 4x4 MIMO system channel rank values of three and four are possible in addition to values of one and two. The number of data streams, however, or more specifically codewords in LTE is limited to a value of two. Thus, LTE has defined the concept of layers, in which the DL transmitter includes a codeword-to-layer mapping, and in which the number of layers is equal to the channel rank. An antenna mapping or precoding operation follows, which maps the layers to the antenna ports. A 4x2 MIMO system is also possible with LTE Release 8, but here the channel rank is limited to the number of UE antennas, which is equal to two.

The network can dynamically choose between different modes based on instantaneous radio conditions between the base station and the UE. Figure 59 shows the decision tree. The antenna configuration (AC) values refer to the transmission modes. Not every network will support every mode. Operators will choose which modes are the most effective and economical. AC2, 3, 4, and 6 are typical modes that will be implemented.

Figure 59: Decision Tree for Different Antenna Schemes[134]

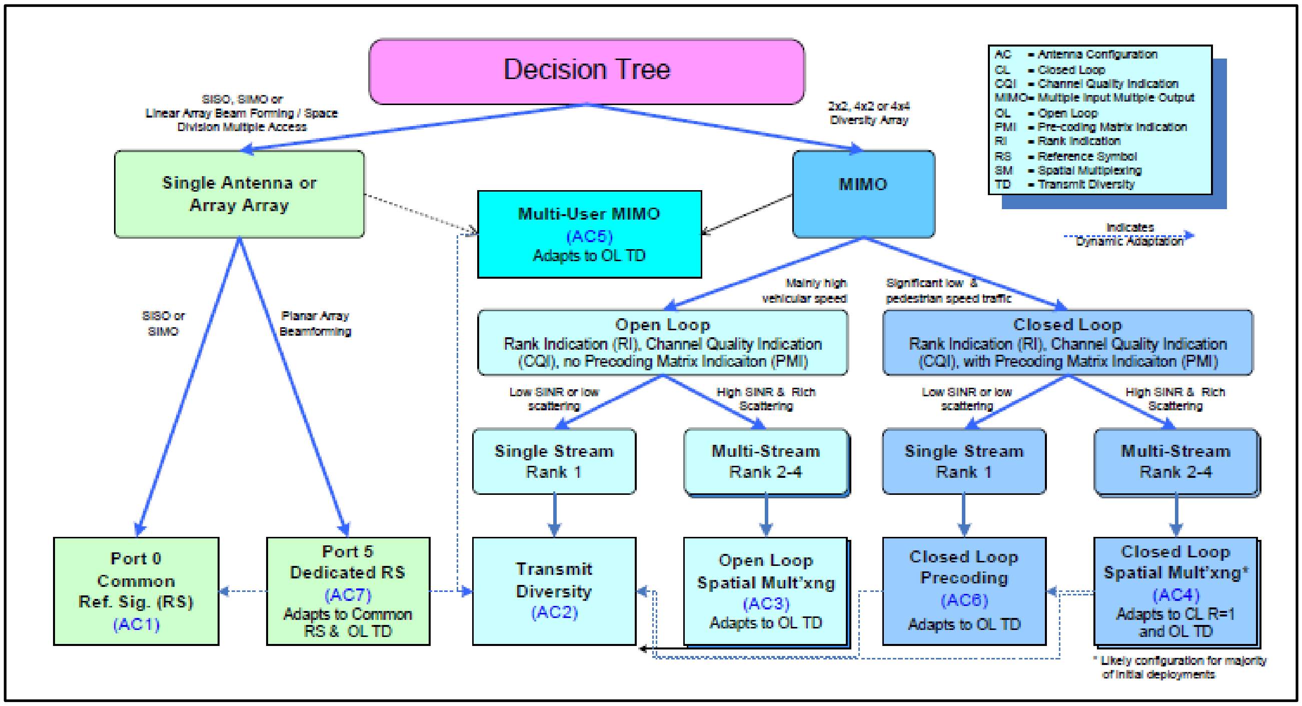
Decision Tree
AC = Antenna Configuration
CL = Closed Loop
CQI = Channel Quality Indication
MIMO= Multiple Input Multiple Output
OL = Open Loop
PMI = Pre-coding Matrix Indication
RI = Rank Indication
RS = Reference Symbol
SM = Spatial Multiplexing
TD = Transmit Diversity
SISO, SIMO or Linear Array Beam Forming / Space Division Multiple Access
2x2, 4x2 or 4x4 Diversity Array
Single Antenna or Array Array
MIMO
Multi-User MIMO (AC5) Adapts to OL TD
Indicates Dynamic Adaptation
Mainly high vehicular speed
Significant low & pedestrian speed traffic
SISO or SIMO
Planar Array Beamforming
Open Loop Rank Indication (RI), Channel Quality Indication (CQI), no Precoding Matrix Indicaiton (PMI)
Closed Loop Rank Indication (RI), Channel Quality Indication (CQI), with Precoding Matrix Indicaiton (PMI)
Low SINR or low scattering
High SINR & Rich Scattering
Low SINR or low scattering
High SINR & Rich Scattering
Single Stream Rank 1
Multi-Stream Rank 2-4
Single Stream Rank 1
Multi-Stream Rank 2-4
Port 0 Common Ref. Sig. (RS) (AC1)
Port 5 Dedicated RS (AC7) Adapts to Common RS & OL TD
Transmit Diversity (AC2)
Open Loop Spatial Mult'xng (AC3) Adapts to OL TD
Closed Loop Precoding (AC6) Adapts to OL TD
Closed Loop Spatial Mult'xng* (AC4) Adapts to CL R=1 and OL TD
* Likely configuration for majority of initial deployments

The simplest mode is AC2, referred to as Transmit Diversity (TD) or sometimes Space Frequency Block Code (SFBC) or even Open Loop Transmit Diversity. TD can operate under all conditions, meaning it works under low SINR, high mobility, and low channel rank (rank = 1). This rank means that the channel is not sufficiently scattered or de-correlated to support two spatial streams. Thus, in TD, only one spatial stream or what is sometimes referred as a single codeword (SCW) is transmitted. If the channel rank increases to a value of two, indicating a more scattered channel, and the SINR is a bit higher, then the system can adapt to AC3 or Open-Loop Spatial Multiplexing (OL-SM), also referred to as large-delay Cyclic Delay Diversity (CDD). This mode supports two spatial streams or two codewords. This mode, also called multiple codeword (MCW) operation, increases throughput over SCW transmission.

If the rank of the channel is one, but the device is not moving very fast or is stationary, then the system can adapt to AC6, called closed-loop (CL) precoding (or CL-rank 1 or CL-R1). In this mode, the network receives from the device with Precoding Matrix Indication (PMI) bits that inform the base station what precoding matrix to use in the transmitter to optimize link performance. This feedback is only relevant for low-mobility or stationary conditions since in high mobility conditions the feedback will most likely be outdated by the time the base station can use it.

Another mode is AC4 or Closed Loop Spatial Multiplexing (CL-SM), which is enabled for low-mobility, high SINR, and channel rank of two. This mode theoretically provides the best user throughput. The figure above shows how these modes can adapt downwards to either OL TD, or if in CL-SM mode, down to either OL TD or CL R1.

---

[134] 4G Americas *MIMO and Smart Antennas for 3G and 4G Wireless Systems – Practical Aspects and Deployment Considerations*, May 2010.

For a 4x4 MIMO configuration, the channel rank can take on values of three and four in addition to one or two. Initial deployment at the base station, however, will likely be two TX antennas and most devices will only have 2 RX antennas, and thus the rank is limited to 2.

AC5 is MU-MIMO, which is not defined for the downlink in Release 8.

AC1 and AC7 are single antenna port modes in which AC1 uses a common Reference Signal (RS), while AC7 uses a dedicated RS or what is also called a user specific RS. AC1 implies a single TX antenna at the base station. AC7 implies an antenna array with antennal elements closely spaced so that a physical or spatial beam can be formed toward an intended user.

LTE operates in a variety of MIMO configurations. On the downlink, these include 2X2, 4X2 (four antennas at the base station), and 4X4. Initial deployment will likely be 2x2 whereas 4X4 will be most likely used initially in femtocells. On the uplink, there are two possible approaches: single-user MIMO (SU-MIMO) and multi-user MIMO (MU-MIMO). SU-MIMO is more complex to implement as it requires two parallel radio transmit chains in the mobile device, whereas MU-MIMO does not require any additional implementation at the device but relies on simultaneous transmission on the same tones from multiple mobile devices.

The first LTE Release thus incorporates MU-MIMO with SU-MIMO deferred for subsequent LTE releases. An alternate form of MIMO, originally called network MIMO, and now called CoMP, relies on MIMO implemented (on either the downlink or uplink or both) using antennas across multiple base stations, as opposed to multiple antennas at the same base station. This paper explains CoMP in the section on LTE Advanced below.

Peak data rates are approximately proportional to the number of send and receive antennas. 4X4 MIMO is thus theoretically capable of twice the data rate of a 2X2 MIMO system. The spatial multiplexing MIMO modes that support the highest throughput rates will be available in early deployments.

For a more detailed discussion of 3GPP antenna technologies, refer to the 5G Americas white paper “MIMO and Smart Antennas for 3G and 4G Wireless Systems – Practical Aspects and Deployment Considerations,” May 2010.

For advancements in LTE Smart Antennas, see the next section.

LTE-Advanced Antenna Technologies

Release 10 added significant enhancements to antenna capabilities, including four-layer transmission resulting in peak spectral efficiency exceeding 15 bps/Hz. Uplink techniques fall into two categories: those relying on channel reciprocity and those that do not. With channel reciprocity, the eNB determines the channel state by processing a Sounding Reference Signal from the UE. It then forms transmission beams accordingly. The assumption is that the channel received by the eNB is the same as the UE. Techniques that use channel reciprocity are beamforming, SU-MIMO, and MU-MIMO. Channel reciprocity works especially well with TDD since both forward and reverse links use the same frequency.

Non-reciprocity approaches apply when the transmitter has no knowledge of the channel state. Techniques in this instance include open-loop MIMO, closed-loop MIMO, and MU-MIMO. These techniques are more applicable for higher speed mobile communications.

For the downlink, the technology can transmit in as many as eight layers using an 8X8 configuration for a peak spectral efficiency of 30 bps/Hz. This exceeds the IMT-Advanced requirements, conceivably supporting a peak rate of 1 Gbps in just 40+40 MHz, and even higher rates in wider bandwidths. This would require additional reference signals for channel estimation and for measurements, including channel quality, to enable adaptive, multi-antenna transmission.

Release 10 supports a maximum of two codewords, the same as previous LTE releases. The release specifies a new transmission mode (TM-9) that supports SU-MIMO up to Rank 8 (up to eight layers), as well as the ability to dynamically switch between SU-MIMO and MU-MIMO.

Figure 60 shows the different forms of single-user MIMO in Releases 8, 9, and 10. Release 8 supports only a single layer, whereas two-layer beamforming is possible in Release 9, and eight layers are possible in Release 10 with eight antennas at the base station.

Figure 60: Single-User MIMO[135]

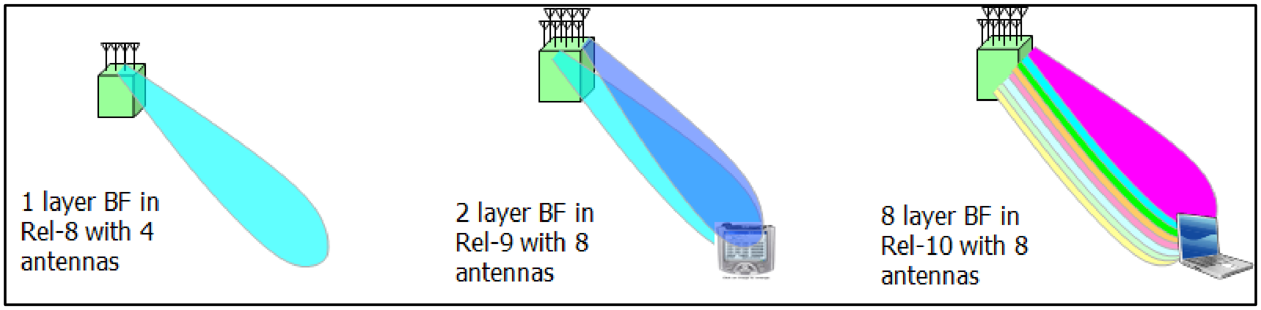
1 layer BF in Rel-8 with 4 antennas
2 layer BF in Rel-9 with 8 antennas
8 layer BF in Rel-10 with 8 antennas

Figure 61 shows multi-user MIMO options across different releases. Release 8 supports two simultaneous users, each with one layer using four antennas, while Releases 9 and 10 support four simultaneous users, each with one layer.

Figure 61: Multi-User MIMO[136]

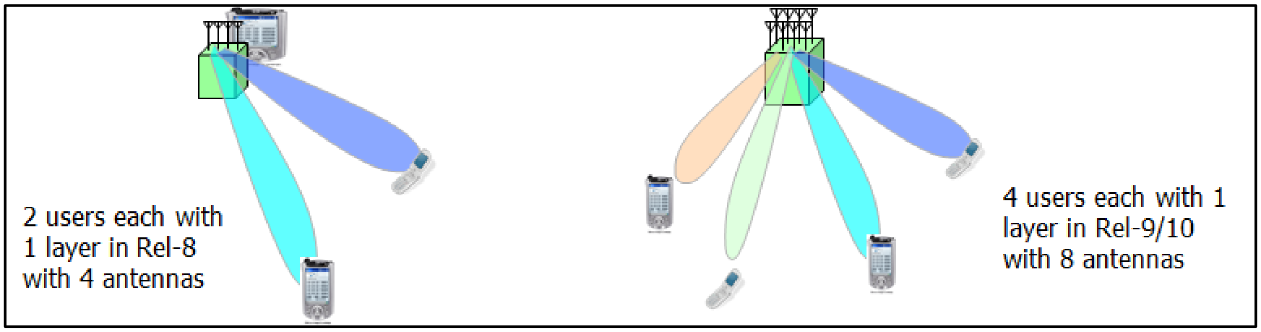
2 users each with 1 layer in Rel-8 with 4 antennas
4 users each with 1 layer in Rel-9/10 with 8 antennas

For four-antenna configurations at the base station, Release 12 improves throughput by adding a feedback mode, called mode 3-2, in which sub-band precoders and sub-band channel quality indicators (CQIs) are included in the UE's feedback to the eNodeB. Release 12 also adds a new codebook that further improves throughput.

---

[135] 5G Americas member contribution.

[136] 5G Americas member contribution.

As depicted in Figure 62 and Figure 63, compared with the Release 8 codebook, the new Release 12 codebook provides a 10% gain for both median and cell-edge throughputs. Compared with feedback mode 3-1, feedback mode 3-2 provides an 18% to 20% gain in median and cell-edge throughput. Jointly, the two methods provide a 28% to 30% gain.

Figure 62: Median Throughput of Feedback Mode 3-2 and New Codebook.[137]

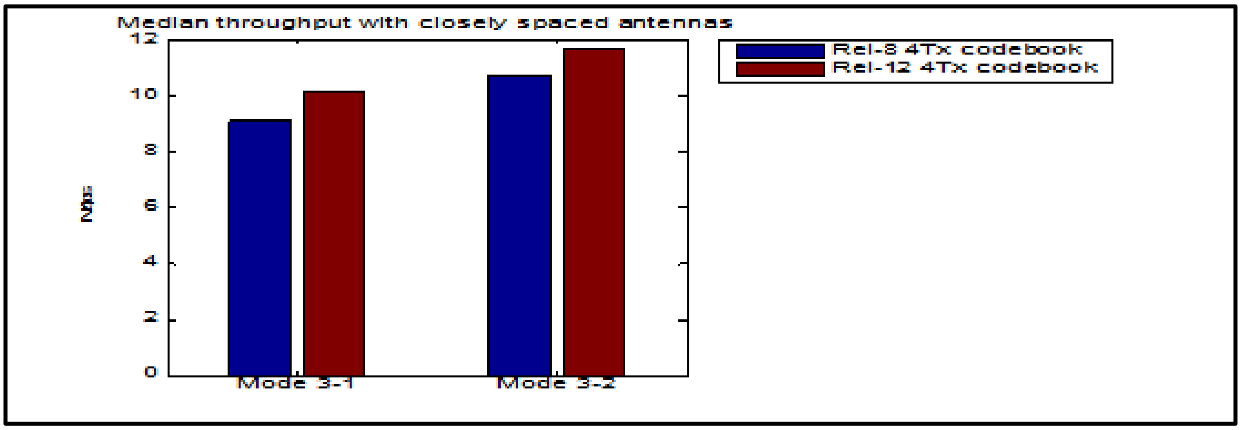

Figure 63: Cell-Edge Throughput of Feedback Mode 3-2 and New Codebook[138]

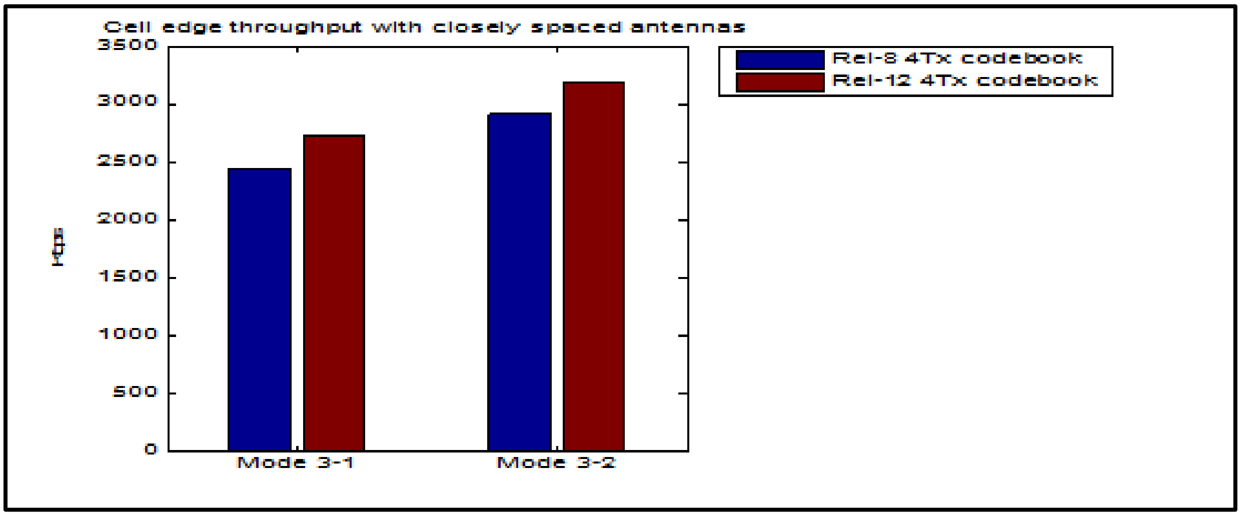

Release 12 also defines how Active Antenna Systems can use multiple transceivers on an antenna array to dynamically adjust a radiation pattern.

---

[137] 5G Americas member contribution. Assumptions include: cellular layout of 19 sites hexagonal grid with three sectors per site and 500-meter inter-site distance; simulation case ITU uMa for macro; carrier frequency 2 GHz, deployment scenario A homogenous macro; SU-MIMO with maximum two layers per UE; proportional fair scheduler; and bursty traffic model.

[138] 5G Americas member contribution. Same assumptions as previous figure.

Release 13 is likely to define full-dimension MIMO, which adds a large number of antenna elements, potentially as many as 64 elements.

A practical consideration with antennas is that many towers today already support multiple operators, with tower companies having to manage interference placement, spectrum allocations, and wind and snow load. At higher frequencies, a single radome (antenna enclosure) can support 4X2 MIMO, but higher-order MIMO may prove impractical for many deployments.

5G systems operating at much higher frequencies will have an advantage since the antenna arrays will be much smaller due to the much smaller wavelengths.

Initial massive MIMO techniques applied to LTE, such as full-dimension MIMO using 8, 16, and 64 transmit antennas, can provide dramatic performance gains, particularly in dense deployments, as shown in Figure 64.

Figure 64: Performance Gains with FD-MIMO Using 200 Meter ISD[139]

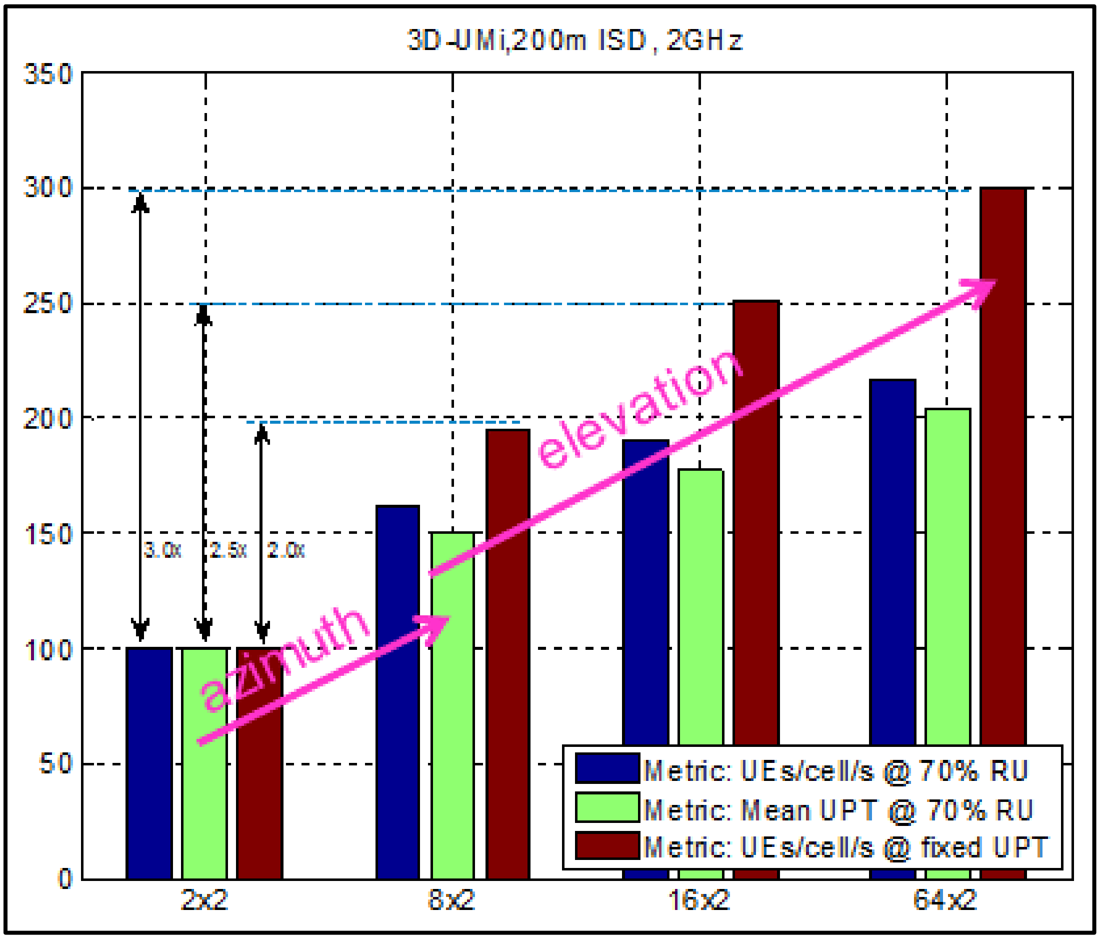

This figure compares 8X2, 16X2, and 64X2 MIMO performance relative to 2X2 MIMO (normalized to value 100). The blue bars (case 1) show the supported number of users per sector (referred to as “cell” in the figure) at a fixed resource utilization (RU) of 70%; the green bars (case 2) show mean user throughput (UPT) at a fixed RU of 70%; and the red bars (case 3) show system capacity in terms of supported number of users for a given user throughput. Resulting gains are:

---

[139] 5G Americas member contribution.

- Case 2 (green bars): 1.5X with 8X2, 1.75X with 16X2, and 2X with 64X2 MIMO.
- Case 3 (red bars): 2X with 8X2, 2.5X with 16X2, and 3X with 64X2 MIMO.

The primary gains are from azimuth (horizontal dimension) in going from 2X2 to 8X2, and from elevation in going to 16X2 and 64X2. FD-MIMO gains are lower with larger ISD values, such as 500 meters.

3GPP has also studied FD-MIMO and conducted a field trial showing impressive throughput gains, particularly in a high-rise scenario.[140]

Carrier Aggregation

Carrier aggregation, first available in Release 10, plays an important role in providing operators maximum flexibility for using all of their available spectrum. By combining spectrum blocks, LTE can deliver much higher throughputs than otherwise possible. Asymmetric aggregation (for example, different amounts of spectrum used on the downlink versus the uplink) provides further flexibility and addresses the greater demand on downlink traffic.

Specific types of aggregation include:

- Intra-band on adjacent channels.
- Intra-band on non-adjacent channels.
- Inter-band (700 MHz, 1.9 GHz).
- Inter-technology (for example, LTE on one channel, HSPA+ on another). This approach is not currently specified nor being developed. While theoretically promising, a considerable number of technical issues would have to be addressed.[141] See Figure 65.

---

[140] 3GPP, *3D-MIMO Prototyping and Initial Field Trial Results*, TSG RAN WG1 Meeting #80, Agenda Item: 7.2.4.4, Document R1-150451.

[141] For further details, see 4G Americas, *HSPA+ LTE Carrier Aggregation*, June 2012.

Figure 65: Inter-Technology Carrier Aggregation[142]

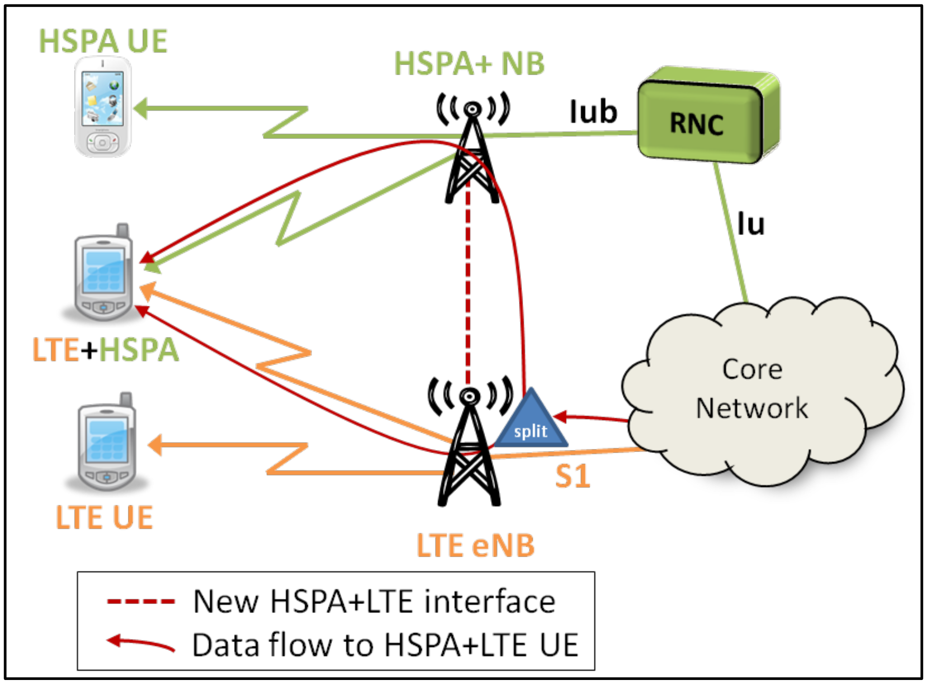
HSPA UE
HSPA+ NB
Iub
RNC
Iu
LTE+HSPA
Core Network
split
S1
LTE UE
LTE eNB
New HSPA+LTE interface
Data flow to HSPA+LTE UE

Figure 66 depicts the carrier-aggregation capabilities of different 3GPP releases.

---

[142] 5G Americas member contribution.

Figure 66: Carrier Aggregation Capabilities across 3GPP Releases[143]

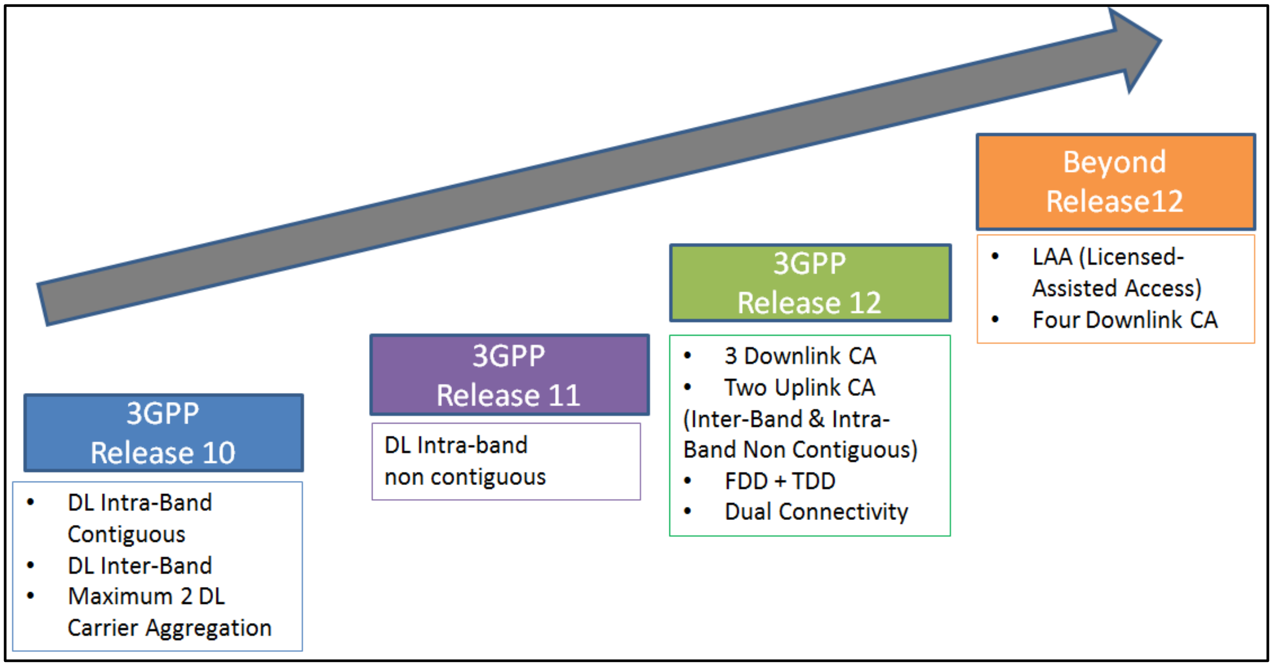
3GPP
Release 10
DL Intra-Band Contiguous
DL Inter-Band
Maximum 2 DL Carrier Aggregation
3GPP
Release 11
DL Intra-band non contiguous
3GPP
Release 12
3 Downlink CA
Two Uplink CA
(Inter-Band & Intra-Band Non Contiguous)
FDD + TDD
Dual Connectivity
Beyond
Release12
LAA (Licensed-Assisted Access)
Four Downlink CA

One anticipated benefit of inter-band aggregation stems from using the lower-frequency band for users who are at the cell edge, to boost their throughput rates. Though this approach improves average aggregate throughput of the cell by only a small amount (say, 10%), it results in a more uniform user experience across the cell coverage area.

Figure 67 shows an example of intra-band carrier aggregation using adjacent channels with up to 100+100 MHz of bandwidth supported. Radio-access network specifications, however, limit the number of carriers to two in Release 10 and Release 11.

Figure 67: Release 10 LTE-Advanced Carrier Aggregation[144]

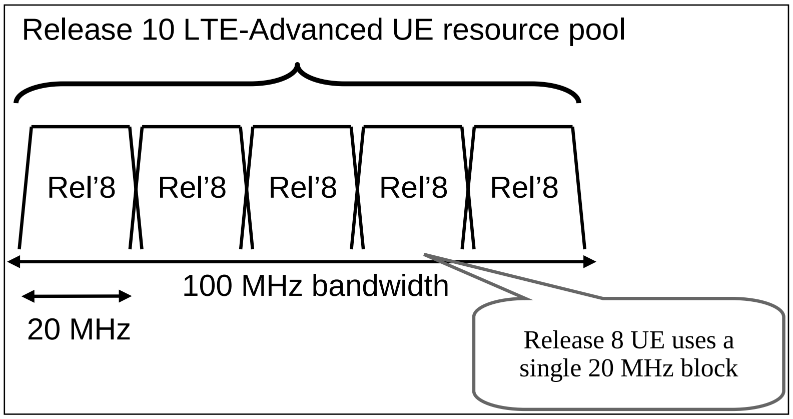
Release 10 LTE-Advanced UE resource pool
Rel'8
Rel'8
Rel'8
Rel'8
Rel'8
100 MHz bandwidth
20 MHz
Release 8 UE uses a single 20 MHz block

---

[143] 4G Americas, *Mobile Broadband Evolution: Rel-12 & Rel-13 and Beyond*, 2015.

[144] Harri Holma and Antti Toskala, *LTE for UMTS, OFDMA and SC-FDMA Based Radio Access*, Wiley, 2009.

Figure 68 shows the carrier aggregation operating at different protocol layers.

Figure 68: Carrier Aggregation at Different Protocol Layers[145]

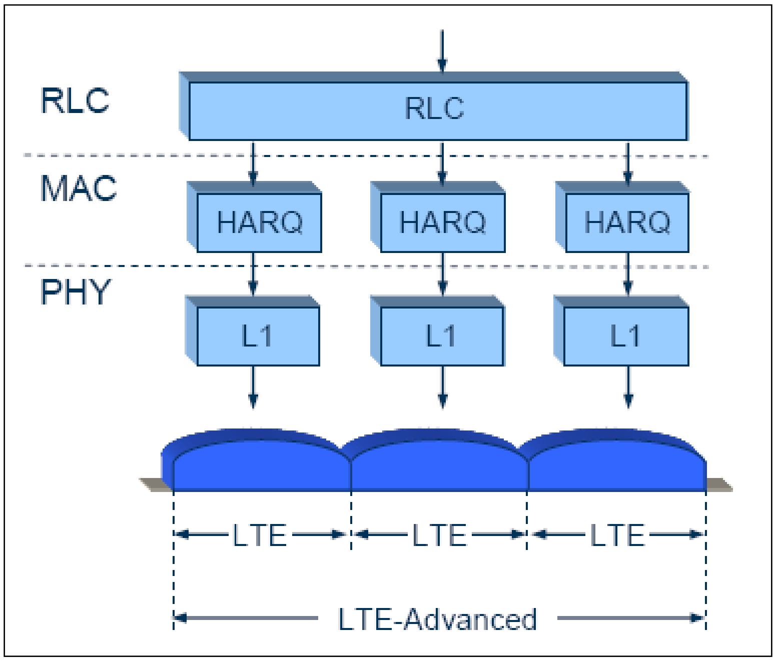

For a list of band combinations, refer to the 5G Americas white paper, *Wireless Technology Evolution Towards 5G: 3GPP Release 13 to Release 15 and Beyond,* February 2017, at section 3.4.3.

Figure 69 shows the result of one simulation study that compares download throughput rates between the blue line, which shows five user devices in 700 MHz and five user devices in AWS not using CA, and the pink line, which shows ten user devices that have access to both bands. Assuming a lightly loaded network with CA, 50% or more users (the median) experience 91% greater throughput, and 95% or more users experience 50% greater throughput. These trunking gains are less pronounced in heavily loaded networks.

---

[145] Stefan Parkvall and David Astely, Ericsson Research, "The Evolution of LTE towards IMT-Advanced," Journal of Communications, Vol. 4, No. 3, April 2009. Available at http://www.academypublisher.com/jcm/vol04/no03/jcm0403146154.pdf.

Figure 69: Gains from Carrier Aggregation[146]

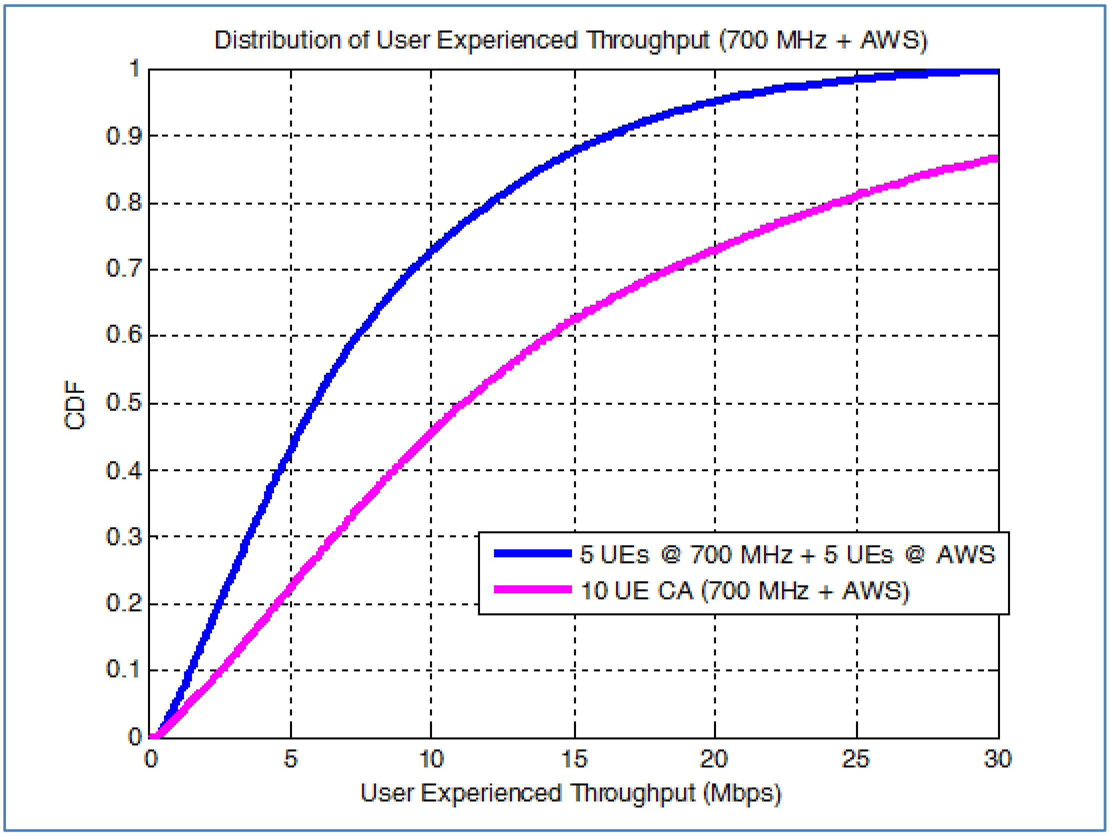

Work in Release 12 is investigating aggregation of joint TDD and FDD carriers.

Coordinated Multi Point (CoMP)

Coordinated Multi Point (CoMP) is a communications technique that can improve coverage, cell-edge throughput, and/or system spectrum efficiency by reducing interference. This technique was thoroughly studied during the development of LTE-Advanced Release 10 and was standardized in Release 11.

CoMP coordinates transmissions at different cell sites, thereby achieving higher system capacity and improving cell-edge data rates.

The main principle of CoMP is that a UE at a cell edge location can receive signals from multiple transmission points, and/or its transmitted signal can be received by multiple reception points. Consequently, if these multiple transmission points coordinate their transmissions, the DL throughput performance and coverage can improve.

For the UL, signals from the UE received at multiple reception points can significantly improve the link performance. Techniques can range from simple interference avoidance methods, such as Coordinated Beam Switching (CBS) and Coordinated Beam Forming

---

[146] 5G Americas member contribution. Assumptions: lightly-loaded network, 2.0 site-to-site distance, file size is 750 Kbytes, traffic model bursty with mean inter-arrival time of five seconds.

(CBF), to complex joint processing techniques that include Joint Transmission (JT), Joint Reception (JR), and Dynamic Point Selection (DPS).

CoMP architectures include inter-site CoMP, intra-site CoMP, as well as CoMP with distributed eNBs (i.e., an eNB with distributed remote radio heads). Figure 70 shows two possible levels of coordination.

Figure 70: Different Coordination Levels for CoMP[147]

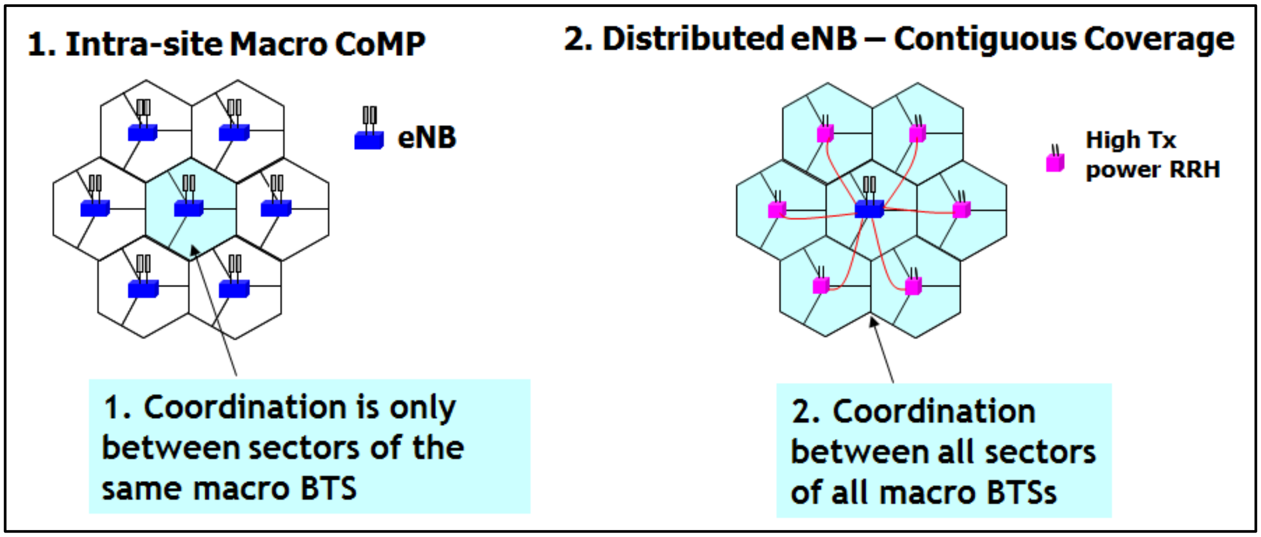
1. Intra-site Macro CoMP
eNB
1. Coordination is only between sectors of the same macro BTS
2. Distributed eNB – Contiguous Coverage
High Tx power RRH
2. Coordination between all sectors of all macro BTSs

In one CoMP approach, called coordinated scheduling and shown in Figure 71, a single site transmits to the user, but with scheduling, including any associated beamforming, coordinated between the cells to reduce interference between the different cells and to increase the served user's signal strength. In Joint Transmission, another CoMP approach also shown in Figure 71, multiple sites transmit simultaneously to a single-user. This approach can achieve higher performance than coordinated scheduling, but it has more stringent backhaul communications requirements. One simpler form of CoMP that will be available in Release 10, and then further developed in Release 11, is ICIC. Release 11 of LTE defines a common feedback and signaling framework for enhanced CoMP operation.

Figure 71: Coordinated Scheduling / BF and Joint Processing CoMP Approaches[148]

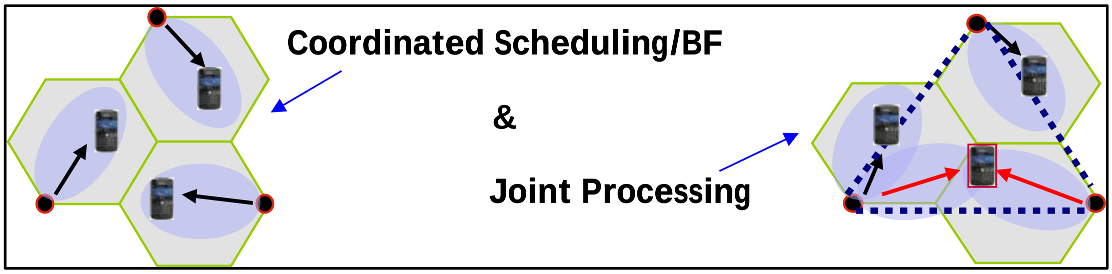
Coordinated Scheduling/BF
&
Joint Processing

---

[147] 5G Americas member contribution.

[148] 5G Americas member contribution.

Release 11 also implements CoMP on the uplink, by which multiple base stations receive uplink transmissions and jointly process the signal, resulting in significant interference cancellation and improvements to spectral efficiency.

The performance gains expected from CoMP are under discussion in the industry. According to 3GPP document TR 36.819, for the case of resource utilization below 35%, CoMP may provide a 5.8% performance gain on the downlink for the mean user and a 17% gain for cell-edge users relative to HetNets without eICIC. For resource utilization of more than 35%, CoMP may provide a 17% mean gain and a 40% cell-edge gain.[149] CoMP can also be used in combination with eICIC for additional gains.

In the same 3GPP TR 36.819 document, 3GPP estimates the downlink CoMP gain in spectral efficiency, defined as average sector throughput for full buffer traffic using JT and 4x2 MU-MIMO as defined in R11, compared with 4x2 MU-MIMO based on R10, to be about 3% for intra-eNodeB CoMP. That gain drops to about 9% for inter-eNodeB CoMP in the case of no delay in the backhaul used to exchange information between eNodeBs. The corresponding gains in cell-edge user throughput are 20% and 31%, respectively.

When increasing the backhaul latency to a more realistic value of 10 msec for inter-eNodeB, spectral efficiency decreases to zero, and the cell edge gain decreases to 10%.

The gains for DL CoMP based on Coordinated Scheduling/Coordinated Beamforming (CS/CB) and intra-eNodeB are less than that provided by JT, with spectral efficiency at 1% and cell edge gains at 4%.

All of the above gains are for FDD networks with cross-polarized antennas at the eNodeBs. For TDD networks, the gains are higher by virtue of being able to invoke channel reciprocity and thus infer the DL channel directly from the UL channel. For example, for intra-eNodeB CoMP with JT 4x2 MU-MIMO, the respective gains in spectral efficiency and cell-edge throughput are 14% and 29%, respectively.

The gains for UL CoMP based on Joint Reception (JR) are greater than the DL gains. For intra-eNodeB CoMP, the average and cell-edge throughputs are increased to 22% and 40%, assuming two receive antenna paths with SU-MIMO. These respective gains increase to 31% and 66% for inter-eNodeB CoMP. In addition, UL CoMP does not require standardization and thus facilitates vendor implementation.

Uplink CoMP assists VoLTE because it improves cell-edge performance, making voice handover more reliable when traversing between cells. The benefit is analogous to CDMA soft handover; in both cases, the mobile device communicates with two sites simultaneously.

Cellular V2X Communications

In Release 14, 3GPP is specifying cellular vehicle-to-X (C-V2X) communications with two complementary transmission modes: direct communications between vehicles and network communications.

Direct communications will use bands such as the Intelligent Transportation Systems (ITS) 5.9 GHz band, using the PC5 interface specified for LTE device-to-device communications, and will not require a Universal Integrated Circuit Card (UICC) SIM (USIM) card. By operating on different channels in the ITS band, direct cellular V2X will be able to co-exist with IEEE 802.11p, another automotive communications protocol.

---

[149] 3GPP, *Coordinated Multi-Point Operation for LTE Physical Layer Aspects, TR 36.819 v11.1.0,* Tables 7.3.1.2-3 and 7.3.1.2-4, September 2011.

In the network communications mode, the system will use traditional cellular licensed spectrum.

Use cases include do-not-pass warnings, blind-curve hazard warnings, road-works warnings, blind-intersection assistance, bicyclist and pedestrian alerts, and left-turn assistance.

C-V2X is being designed to be compatible with other automotive standards, such as those from ETSI and the Society of Automotive Engineers.

User-Plane Congestion Management (UPCON)

With User-Plane Congestion Management, specified in Release 13, operators have additional tools to mitigate network congestion in specific coverage areas. Mechanisms include traffic prioritization by adjusting QoS for specific services; reducing traffic by, for example, compression; and limiting traffic, such as by prohibiting or deferring certain traffic.

3GPP specifications add a new architectural entity, called the “RAN Congestion Awareness Function” (RCAF), that determines whether a cell is congested, determines the UEs supported by that cell, and informs the Policy Control and Charging Rules Function (PCRF), which can subsequently apply different policies to mitigate the congestion.[150]

Network-Assisted Interference Cancellation and Suppression (NAICS)

NAICS, a Release 13 capability, enhances the interference cancellation and suppression capability of UEs by using more information from the network. The fundamental goal of NAICS is to identify and cancel the dominant interferer, not an easy task when the dominant interferer can be on or off and can change in time and frequency. One analysis estimates an average performance gain of 7.4% relative to Release 11 Interference Rejection Combining and 11.7% at the cell edge.[151] 5G Americas members expect even higher performance gains, for example 20%, with implementation-specific scheduling and as NAICS methods are refined.

Multi-User Superposition Transmission (MUST)

MUST, a study item in Release 13 and tentatively planned for Release 14 uses simultaneous transmissions of data for more than one UE within a cell without time, frequency, or spatial layer separation. The concept relies on a UE close to the base station having low propagation loss and a UE far from the base station having high propagation loss. The far UE is not aware of, nor interfered by the near UE transmission. The near UE cancels the far UE interference. The capacity gain grows with the SNR/SINR difference between the close and far UEs.

---

[150] For further details, see 3GPP TR 23.705, “Study on system enhancements for user plane congestion management (Release 13).”

[151] Harri Holma, Antti Toskala, Jussi Reunanen, *LTE Small Cell Optimization: 3GPP Evolution to Release 13*, Jan 2016, Wiley, ISBN: 978-1-118-91257-7.

IPv4/IPv6

Release 8 defines support for IPv6 for both LTE and UMTS networks. An Evolved Packet System bearer can carry both IPv4 and IPv6 traffic, enabling a UE to communicate both IPv4 and IPv6 packets (assuming it has a dual stack) while connected through a single EPS bearer. It is up to the operator, however, whether to assign IPv4, IPv6, or both types of addresses to UE.

Communicating between IPv6-only devices and IPv4 endpoints will require protocol-conversion or proxies. For further details, refer to the 5G Americas white paper, "IPv6 – Transition Considerations for LTE and Evolved Packet Core," February 2009.

TDD Harmonization

3GPP developed LTE TDD to be fully harmonized with LTE FDD including alignment of frame structures, identical symbol-level numerology, the possibility of using similar Reference Signal patterns, and similar synchronization and control channels. Also, there is only one TDD variant. Furthermore, LTE TDD has been designed to co-exist with TD-SCDMA and TD-CDMA/UTRA (both low-chip rate and high-chip rate versions). LTE TDD achieves compatibility and co-existence with TD-SCDMA by defining frame structures in which the DL and UL time periods can be time aligned to prevent BTS to BTS and UE to UE interference to support operation in adjacent carriers without the need for large guardbands between the technologies. This will simplify deployment of LTE TDD in countries such as China that are deploying TD-SCDMA. Figure 72 demonstrates the synchronization between TC-SCDMA and LTE-TDD in adjacent channels.

Figure 72: TDD Frame Co-Existence between TD-SCDMA and LTE TDD[152]

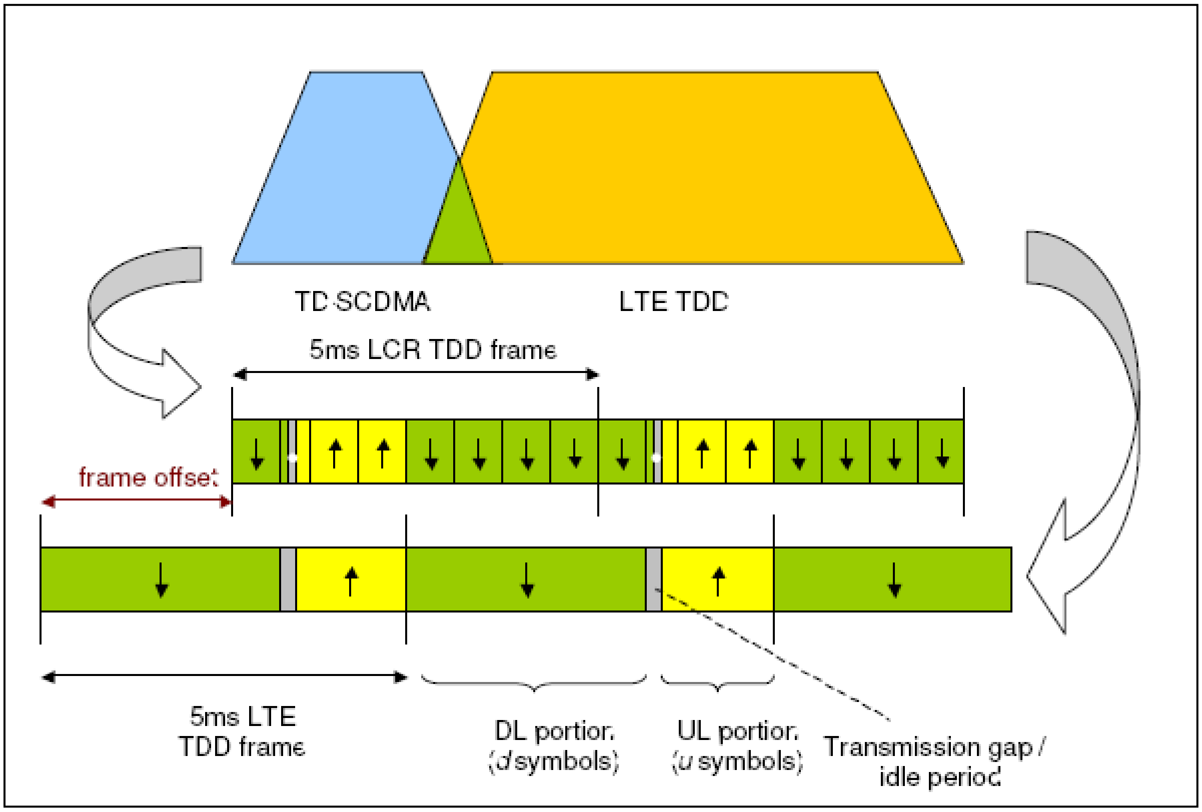
TD-SCDMA
LTE TDD
5ms LCR TDD frame
frame offset
5ms LTE TDD frame
DL portion (d symbols)
UL portion (u symbols)
Transmission gap / idle period

For LTE FDD and TDD to co-exist, large guardbands will be needed to prevent interference.

SMS in LTE

Even if an LTE network uses CSFB for voice, LTE devices will be able to send and receive SMS messages while on the LTE network. In this case, the 2G/3G core network will handle SMS messaging, but will tunnel the message to the MME in the EPC via the SGs interface. Once an LTE network uses IMS and VoLTE for packet voice service, SMS will be handled as SMS over IP and will use IMS infrastructure.[153]

User Equipment Categories

LTE specifications define categories of UE, which mainly determine the maximum throughputs of devices but also govern the number of downlink MIMO layers, as shown in Table .

Higher throughput capabilities are possible with 64 QAM and 256 QAM modulation. 3GPP is also defining Category 0 and Category M devices for M2M, as discussed in the section “Internet of Things and Machine-to-Machines.”

---

[152] 5G Americas member company contribution.

[153] For further details, see 4G Americas, *Coexistence of GSM, HSPA and LTE*, May 2011, 35.

Table 19: UE Categories[154]

UE Category	Max DL Throughput	Maximum DL MIMO Layers	Maximum UL Throughput
1	10.3 Mbps	1	5.2 Mbps
2	51.0 Mbps	2	25.5 Mbps
3	102.0 Mbps	2	51.0 Mbps
4	150.8 Mbps	2	51.0 Mbps
5	299.6 Mbps	4	75.4 Mbps
6	301.5 Mbps	2 or 4	51.0 Mbps
7	301.5 Mbps	2 or 4	102.0 Mbps
8	2998.6 Mbps	8	1497.8 Mbps
9	452.3 Mbps	2 or 4	51.0 Mbps
10	452.3 Mbps	2 or 4	102.0 Mbps
11	603.0 Mbps	2 or 4	51.0 Mbps
12	603.0 Mbps	2 or 4	102.0 Mbps
13	391.6 Mbps	2 or 4	150.8 Mbps
14	3916.6 Mbps	8	9587.7 Mbps
15	798.8 Mbps	2 or 4	226.1 Mbps
16	1051.4 Mbps	2 or 4	105.5 Mbps
17	2506.6 Mbps	8	2119.4 Mbps
18	1206.0 Mbps	2 or 4 (or 8)	211.0 Mbps
19	1658.3 Mbps	2 or 4 (or 8)	13563.9 Mbps

LTE-Advanced Relays

Another capability being planned for LTE-Advanced is relays, as shown in Figure 73. The idea is to relay frames at an intermediate node, resulting in much better in-building penetration, and with better signal quality, user rates will improve. Relay nodes can also improve cell-edge performance by making it easier to add picocells at strategic locations.

---

[154] 3GPP, *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio access capabilities,* Technical Specification 36.306 V14.2.0, March 2017.

Relays provide a means for lowering deployment costs in initial deployments in which usage is relatively low. As usage increases and spectrum needs to be allocated to access only, operators can then employ alternate backhaul schemes.

Figure 73: LTE-Advanced Relay[155]

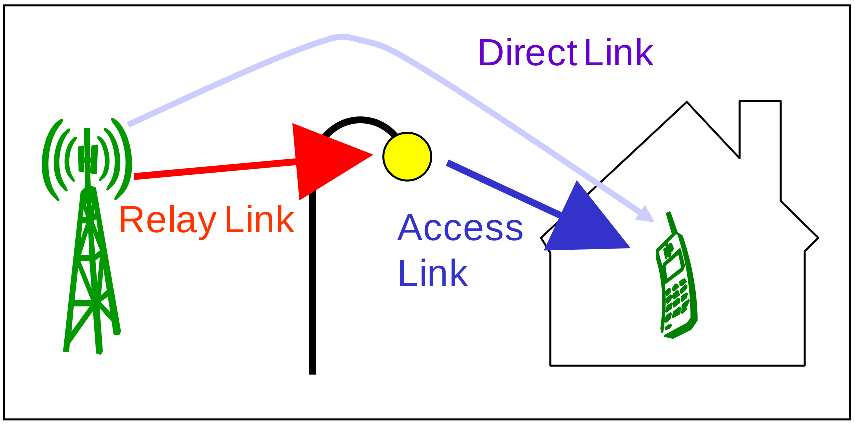

Proximity Services (Device-to-Device)

Release 12 defined a capability for devices to communicate directly with one another using LTE spectrum, a feature also called "operator-enabled proximity services." With this capability, devices can autonomously discover nearby relevant devices and services in a battery-efficient manner. Devices broadcast their needs and services and can also passively identify services without user intervention. The communication between devices is called "sidelink communications" and uses an interface called "PC5." Release 12, emphasizing public-safety applications, supports only one-to-many sidelink communications, whereas Release 13 supports one-to-one sidelink communications between two group member UEs and between a remote UE and a relay UE.

Initial emphasis of this capability, in both Release 12 and Release 13, is on public safety. Examples of potential consumer or commercial applications include discovering friends and family (social matching), push advertising for relevant notifications, tourist bulletins, venue services, crime alerts, home automation, vehicle-to-vehicle communication, and detecting children leaving the vicinity of their homes. The service is designed to work during infrastructure failures, even in emergencies and natural disasters. As a new means of communicating, proximity services could result in innovative types of applications.

The LTE network performs configuration and authentication; however, communication can be either via the network or directly between devices. To minimize battery consumption, devices synchronously wake up for brief intervals to discover services. The impact on LTE network capacity is minimal.

As with other location-based services, operators and application developers will need to address privacy concerns.

---

[155] 5G Americas member contribution.

LTE Throughput

The section “4G LTE Advances” above in the main section of the paper and “Data Throughput Comparison” in the appendix provide an overview of LTE throughputs. This section provides additional details.

Table shows initial (Release 8) LTE peak data rates based on different downlink and uplink designs.

Table20: LTE Peak Throughput Rates

LTE Configuration	Downlink (Mbps) Peak Data Rate	Uplink (Mbps) Peak Data Rate
Using 2X2 MIMO in the Downlink and 16 QAM in the Uplink, 10+10 MHz	70.0	22.0
Using 4X4 MIMO in the Downlink and 64 QAM in the Uplink, 20+20 MHz	300.0	71.0

LTE is not only efficient for data but, because of a highly efficient uplink, is extremely efficient for VoIP traffic. As discussed in the “Spectral Efficiency” section above, in 10+10 MHz of spectrum, LTE VoIP capacity will reach 500 users.[156]

Table 19 analyzes LTE median and average throughput values in greater detail for different LTE configurations.

Table 19: LTE FDD User Throughputs Based on Simulation Analysis[157]

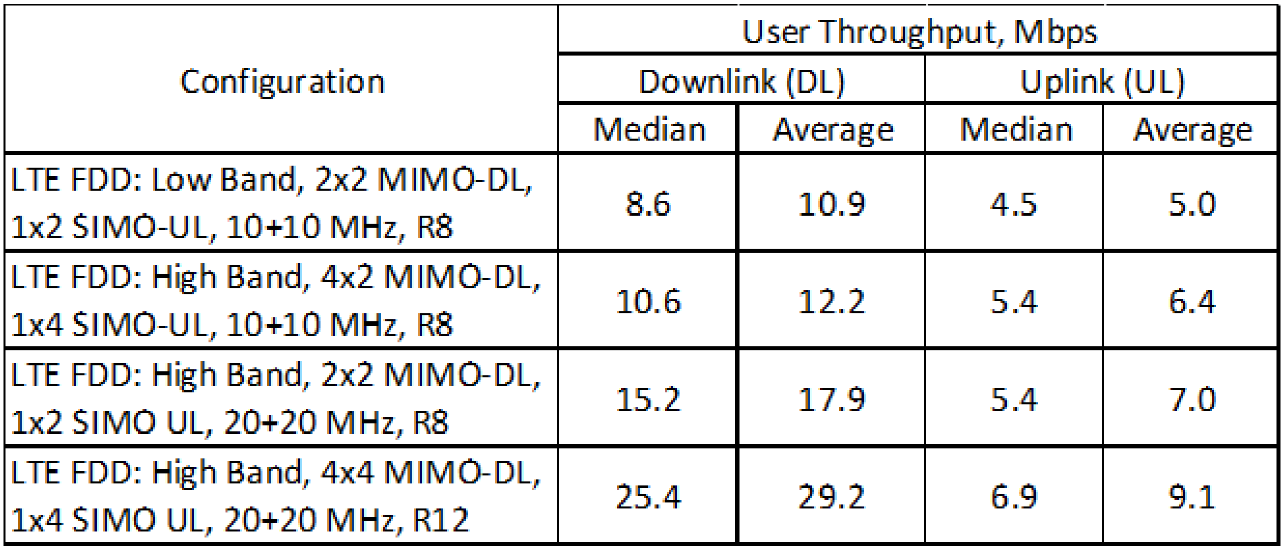

Configuration	User Throughput, Mbps			
	Downlink (DL)		Uplink (UL)	
	Median	Average	Median	Average
LTE FDD: Low Band, 2x2 MIMO-DL, 1x2 SIMO-UL, 10+10 MHz, R8	8.6	10.9	4.5	5.0
LTE FDD: High Band, 4x2 MIMO-DL, 1x4 SIMO-UL, 10+10 MHz, R8	10.6	12.2	5.4	6.4
LTE FDD: High Band, 2x2 MIMO-DL, 1x2 SIMO UL, 20+20 MHz, R8	15.2	17.9	5.4	7.0
LTE FDD: High Band, 4x4 MIMO-DL, 1x4 SIMO UL, 20+20 MHz, R12	25.4	29.2	6.9	9.1

---

[156] 3GPP Multi-member analysis.

[157] 5G Americas member contribution. SIMO refers to Single Input Multiple Output antenna configuration, which in the uplink means one transmit antenna at the UE and multiple receive antennas at the eNodeB.

The simulation results represent a consensus view of 5G Americas members working on this white paper project. The goal of the analysis was to quantify LTE throughputs in realistic deployments. Simulation assumptions include:

- Traffic is FTP-like at a 50% load with a 75/25 mix of indoor/outdoor users.
- Throughput is at the medium-access control (MAC) protocol layer. (Application-layer throughputs may be 5 to 8 percent lower due to protocol overhead.)
- The 3GPP specification release numbers shown correspond to the infrastructure capability.
- The configuration in the first row corresponds to low-frequency band operation, representative of 700 MHz or cellular, while the remaining configurations assume high-frequency band operation, representative of PCS, AWS, or WCS. (Higher frequencies facilitate higher-order MIMO configurations and have wider radio channels available.)
- The downlink value for the first row corresponds to Release 8 device-receive capability (Minimum Mean Square Error [MMSE]), while the values in the other rows correspond to Release 11 device-receive capability (MMSE – Interference Rejection Combining [IRC]).
- The uplink value for the first row corresponds to a Maximal Ratio Combining (MRC) receiver at the eNodeB, while the remaining values correspond to an IRC receiver.
- Low-band operation assumes 1,732-meter inter-site distance, while high-band operation assumes 500-meter ISD. The remaining simulation assumptions are listed in Table 20.

Table 20: LTE FDD User Throughput Simulation Assumptions[158]

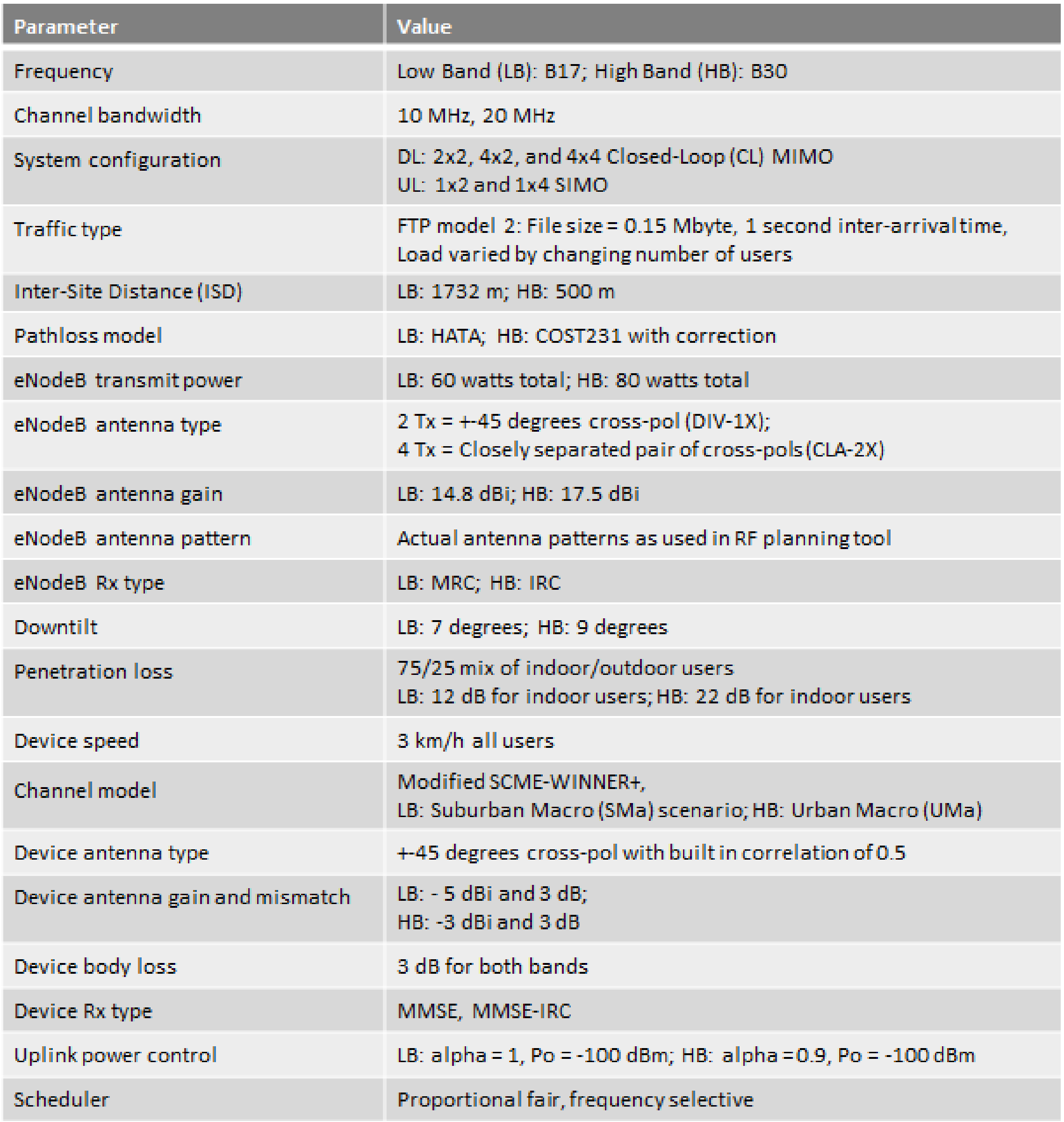

Parameter	Value
Frequency	Low Band (LB): B17; High Band (HB): B30
Channel bandwidth	10 MHz, 20 MHz
System configuration	DL: 2x2, 4x2, and 4x4 Closed-Loop (CL) MIMO UL: 1x2 and 1x4 SIMO
Traffic type	FTP model 2: File size = 0.15 Mbyte, 1 second inter-arrival time, Load varied by changing number of users
Inter-Site Distance (ISD)	LB: 1732 m; HB: 500 m
Pathloss model	LB: HATA; HB: COST231 with correction
eNodeB transmit power	LB: 60 watts total; HB: 80 watts total
eNodeB antenna type	2 Tx = +-45 degrees cross-pol (DIV-1X); 4 Tx = Closely separated pair of cross-pols (CLA-2X)
eNodeB antenna gain	LB: 14.8 dBi; HB: 17.5 dBi
eNodeB antenna pattern	Actual antenna patterns as used in RF planning tool
eNodeB Rx type	LB: MRC; HB: IRC
Downtilt	LB: 7 degrees; HB: 9 degrees
Penetration loss	75/25 mix of indoor/outdoor users LB: 12 dB for indoor users; HB: 22 dB for indoor users
Device speed	3 km/h all users
Channel model	Modified SCME-WINNER+, LB: Suburban Macro (SMa) scenario; HB: Urban Macro (UMa)
Device antenna type	+-45 degrees cross-pol with built in correlation of 0.5
Device antenna gain and mismatch	LB: - 5 dBi and 3 dB; HB: -3 dBi and 3 dB
Device body loss	3 dB for both bands
Device Rx type	MMSE, MMSE-IRC
Uplink power control	LB: alpha = 1, Po = -100 dBm; HB: alpha = 0.9, Po = -100 dBm
Scheduler	Proportional fair, frequency selective

The assumptions, emphasizing realistic deployments, do not necessarily match assumptions used by other organizations, such as 3GPP, so results may differ.

Additional insight into LTE performance under different configuration comes from a test performed on a cluster of cells in an LTE operator’s network, comparing downlink performance of 4X2 MIMO against 2X2 MIMO, and uplink performance of 1X4 SIMO against 1X2 SIMO. The test employed LTE category 4 devices.[159]

---

[158] 5G Americas member contribution.

[159] 5G Americas member contribution.

These tests, which were performed in a 20+20 MHz cluster, show significant improvements in cell edge uplink and downlink throughput, in addition to an overall increase in uplink and downlink throughputs. Specific results include:

- A 100% increase in uplink throughput at the cell edge with 1X4 SIMO compared to 1x2 SIMO.
- A 40% increase in downlink throughput at the cell edge with 4x2 closed-loop MIMO compared to 2x2 open-loop MIMO.
- A 50 to 75% increase in downlink throughput with closed loop MIMO compared to transmit diversity modes.
- Up to 6dB gains in uplink transmit power with 1X4 SIMO, which directly translates into UE battery savings.
- Peak speeds of 144 Mbps with 4X2 MIMO in the downlink and 47 Mbps with 1X4 SIMO in the uplink.

Another LTE operator's testing results for LTE in a TDD configuration, using 20 MHz channels, 3:2 DL to UL ratio, and category 3 devices, showed:

- Peak speeds of 55 Mbps.
- Typical speeds of 6 to 15 Mbps.[160]

Figure 74 shows the result of a drive test in a commercial LTE network with a 10 MHz downlink carrier demonstrating 20 Mbps to 50 Mbps throughput rates across much of the coverage area. Throughput rates would double with a 20+20 MHz configuration.

---

[160] 5G Americas member contribution.

Figure 74: Drive Test of Commercial European LTE Network (10+10 MHz)[161]

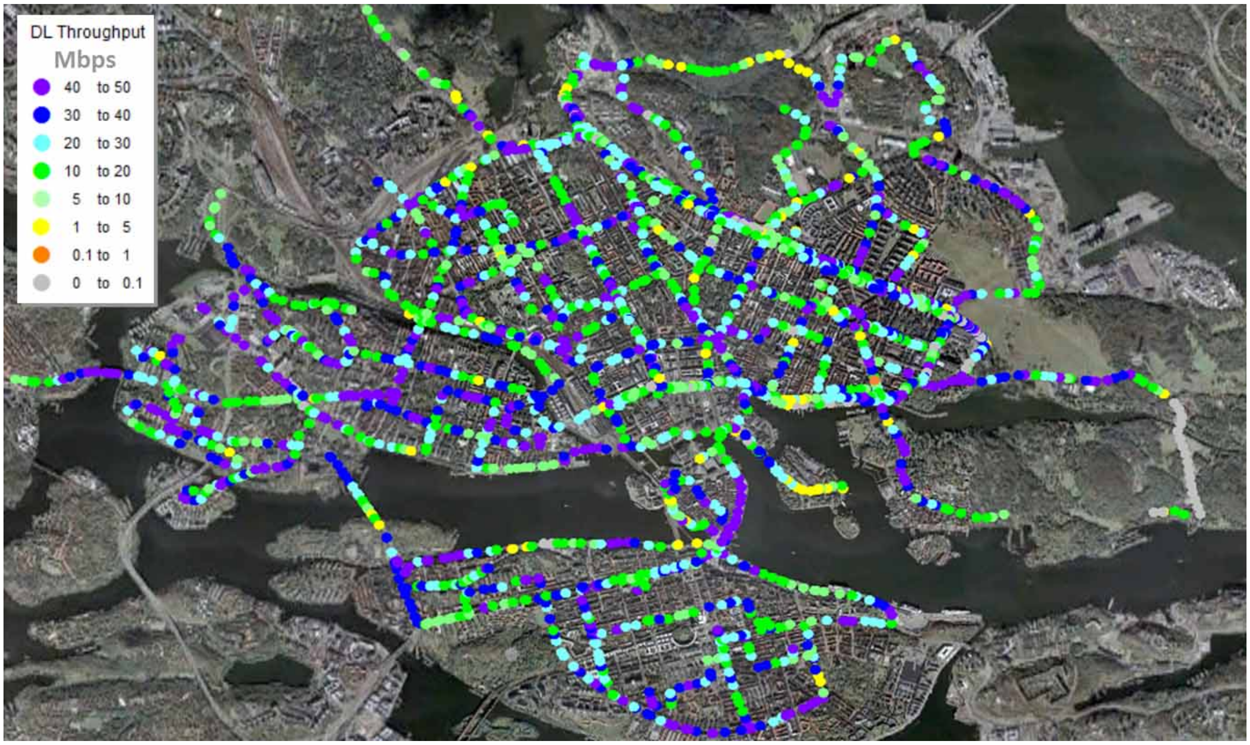
DL Throughput
Mbps
40 to 50
30 to 40
20 to 30
10 to 20
5 to 10
1 to 5
0.1 to 1
0 to 0.1

---

[161] Ericsson contribution.

Figure 75 provides additional insight into LTE downlink throughput, showing Layer 1 throughput simulated at 10 MHz bandwidth using the Extended Vehicular A 3 km/hour channel model. The figure shows the increased performance obtained with the addition of different orders of MIMO. Note how throughput improves based on higher signal to noise ratio (SNR).

Figure 75: LTE Throughput in Various Modes[162]

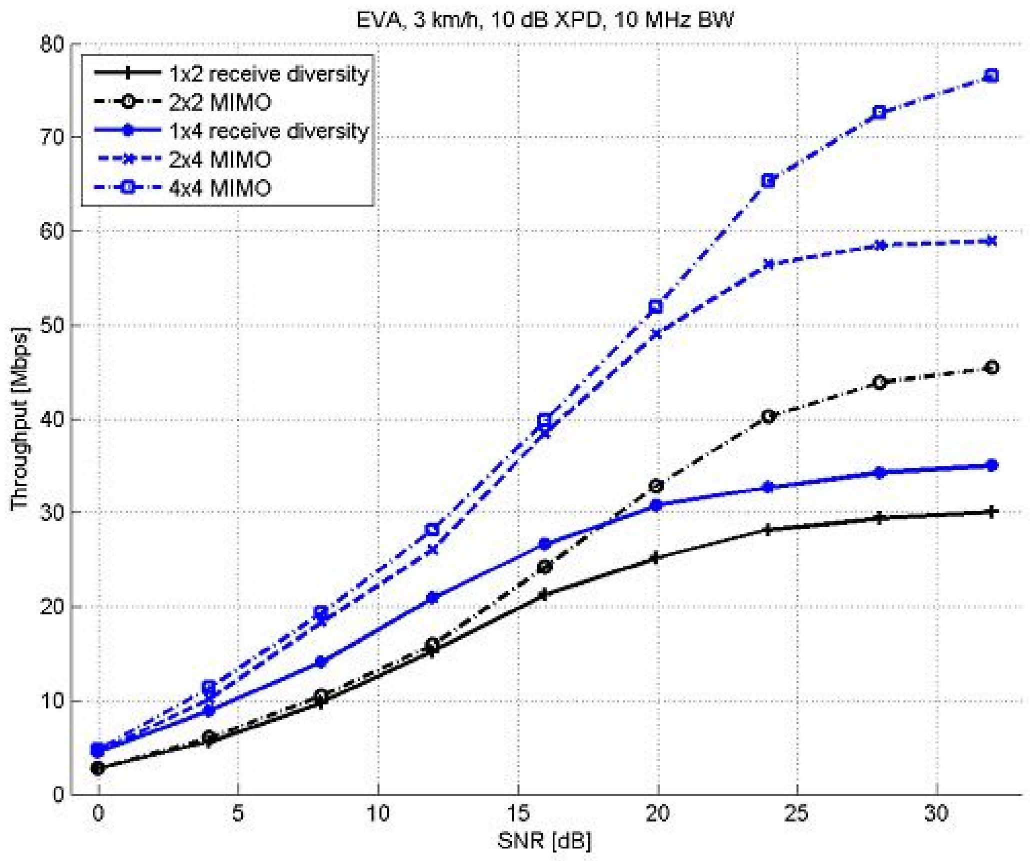

Actual throughput rates that users experience are lower than the peak rates and depend on a variety of factors:

- **RF Conditions and User Speed.** Peak rates depend on optimal conditions. Suboptimal conditions include being at the edge of the cell or moving at high speed, resulting in lower throughput.
- **Network Loading.** Like all wireless systems, throughput rates go down as more devices simultaneously use the network. Throughput degradation is linear.

---

[162] Jonas Karlsson, Mathias Refback, "Initial Field Performance Measurements of LTE," *Ericsson Review*, No. 3, 2008.

Figure 76 shows how dramatically throughput rates can vary by number of active users and radio conditions. The higher curves are for better radio conditions.

Figure 76: LTE Actual Throughput Rates Based on Conditions[163]

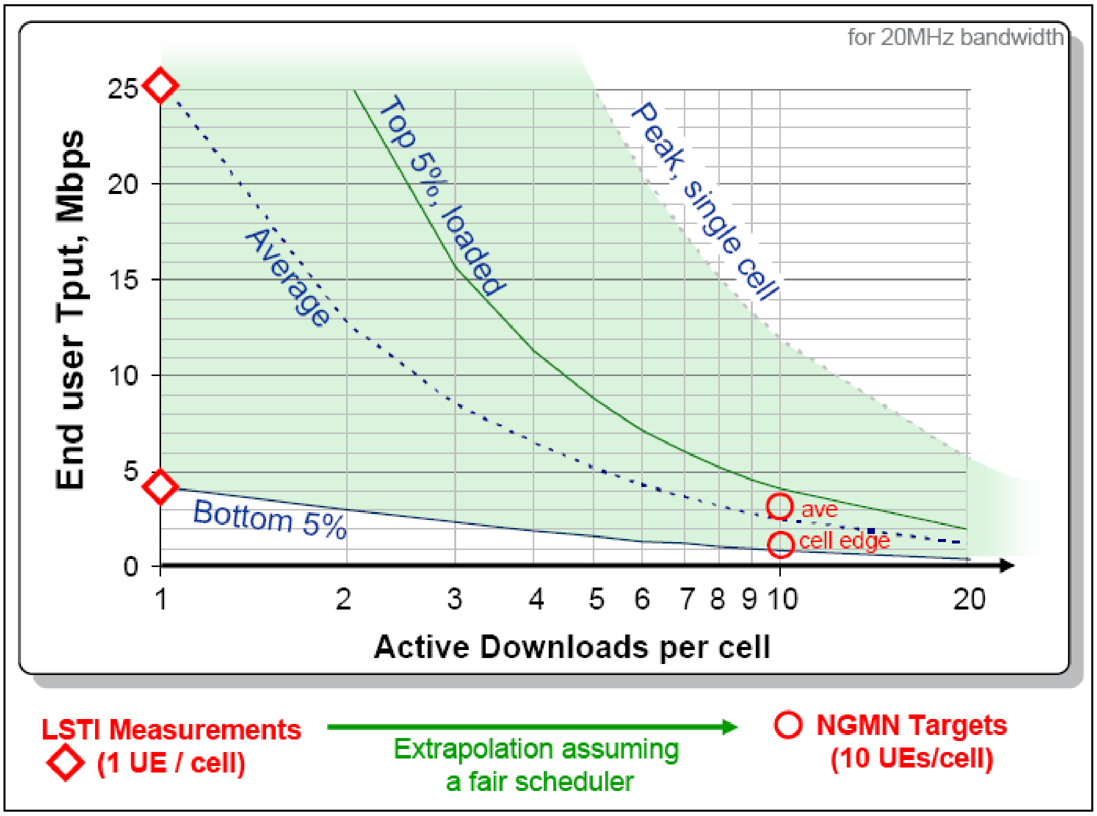

---

[163] LTE/SAE Trial Initiative, "Latest Results from the LSTI, Feb 2009," http://www.lstiforum.org.

VoLTE and RCS

This paper introduced VoLTE and voice support in the earlier section, "VoLTE, RCS, WebRTC, and Wi-Fi Calling." This section in the appendix provides additional technical detail about the operation of VoLTE and RCS.

Voice in LTE can encompass: no voice support, voice implemented in a circuit-switched fallback (CSFB) mode using 2G or 3G, and VoIP implemented with IMS.

Initial LTE network deployments used CSFB, with which the LTE network carries circuit-switched signaling over LTE interfaces, allowing the subscriber to be registered with the 2G/3G MSC even while on the LTE network. When there is a CS event, such as an incoming voice call, the MSC sends the page to the LTE core network, which delivers it to the subscriber device. The device then switches to 2G/3G operation to answer the call.

Voice over LTE using VoIP requires IMS infrastructure. To facilitate IMS-based voice, vendors and operators created the One Voice initiative to define required baseline functionality for user equipment, the LTE access network, the Evolved Packet Core, and the IMS. GSMA adopted the One Voice initiative in what it calls VoLTE, specified in GSMA reference document IR.92.[164] GSMA specifies interconnection and international roaming among LTE networks through the IR.88[165] specification. Another specification, IR.94, provides the IMS Profile for Conversational Video Service, a service referred to as "Video over LTE" (ViLTE).[166]

For a phone to support VoLTE, it needs software implementing the IMS protocol stack. For example, the iPhone 6 was the first iPhone to implement such software. Additional software implementing RCS application programming interfaces can provide applications with access to IMS-based services, such as voice, messaging, and video. The Open Mobile Alliance has defined RESTful network APIs for RCS that support the following functions: notification channel, chat, file transfer, third-party calls, call notification, video sharing, image sharing, and capability discovery. As shown in Figure 77, over time, new profile releases will broaden the scope of these APIs.

---

[164] GSMA, "IMS Profile for Voice and SMS," Document IR.92. Available at http://www.gsma.com/newsroom/wp-content/uploads/2013/04/IR.92-v7.0.pdf.

[165] GSMA, "LTE Roaming Guidelines," GSMA Document IR.88. Available at http://www.gsma.com/newsroom/wp-content/uploads/2013/04/IR.88-v9.0.pdf.

[166] GSMA, "IMS Profile for Conversational Video Service," Document IR.94. Available at http://www.gsma.com/newsroom/all-documents/ir-94-ims-profile-for-conversational-video-service/.

Figure 77: Evolution of RCS API Profiles[167]

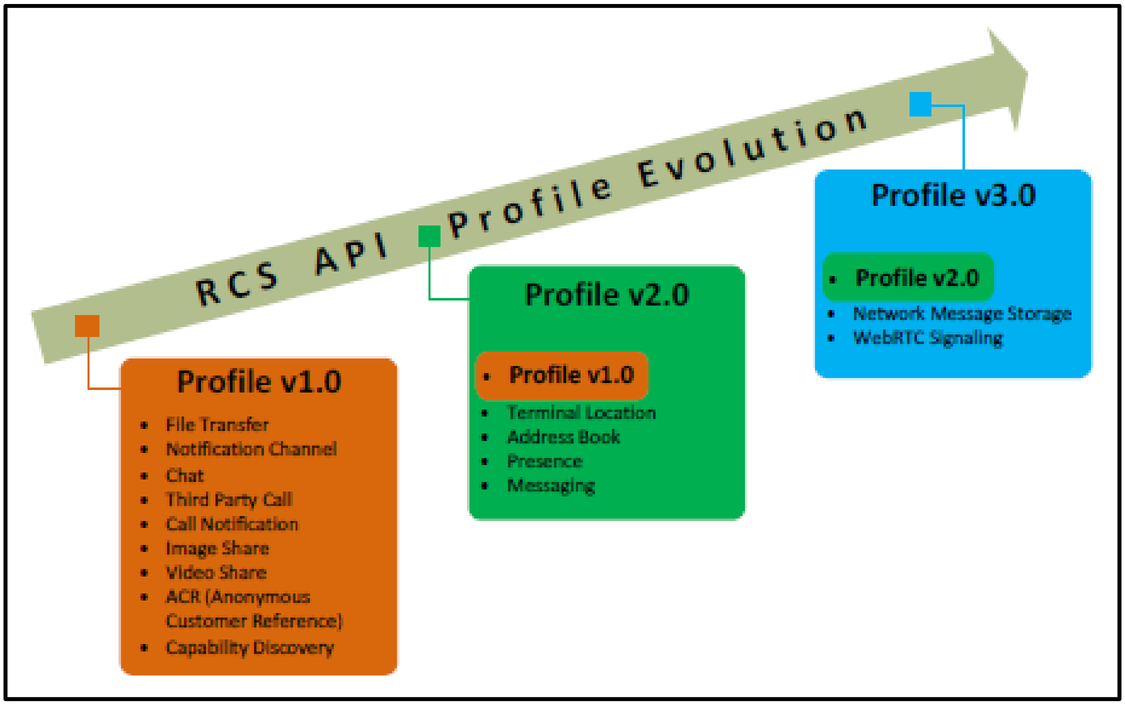
RCS API Profile Evolution
Profile v1.0
File Transfer
Notification Channel
Chat
Third Party Call
Call Notification
Image Share
Video Share
ACR (Anonymous Customer Reference)
Capability Discovery
Profile v2.0
Profile v1.0
Terminal Location
Address Book
Presence
Messaging
Profile v3.0
Profile v2.0
Network Message Storage
WebRTC Signaling

LTE VoIP leverages the QoS capabilities defined for EPC, which specify different quality classes. Features available in LTE to make voice operation more efficient include Semi-Persistent Scheduling (SPS) and TTI bundling. SPS reduces control channel overhead for applications (like VoIP) that require a persistent radio resource. Meanwhile, TTI bundling improves subframe utilization by reducing IP overhead, while in the process optimizing uplink coverage.

Another way to increase voice capacity in LTE and to support operation in congestion situations is vocoder rate adaptation, a mechanism with which operators can control the codec rate based on network load, thus dynamically trading off voice quality against capacity.

VoLTE roaming across operators will require network-to-network interfaces between their respective IMS networks. Such roaming and interconnect will follow initial VoLTE deployments. Different IMS stack implementations between vendors will also complicate roaming.

One roaming consideration is how operators handle data roaming. LTE roaming can send all visited network traffic back to the home network, which for a voice call, increases voice latency. For voice calls, the local breakout option would mitigate this latency.

Using Single-Radio Voice Call Continuity (SR-VCC) and Enhanced SR-VCC (eSRVCC), user equipment can switch mid-call to a circuit-switched network, in the event that the user moves out of LTE coverage. Similarly, data sessions can be handed over in what is called “Packet-Switched Handover” (PSHO).

---

[167] 4G Americas, *VoLTE and RCS Technology – Evolution and Ecosystem*, Nov. 2014.

Figure 78 shows how an LTE network might evolve in three stages. Initially, LTE performs only data service, and the underlying 2G/3G network provides voice service via CSFB. In the second stage, voice over LTE is available, but LTE covers only a portion of the total 2G/3G coverage area. Hence, voice in 2G/3G can occur via CSFB or SR-VCC. Eventually, LTE coverage will match 2G/3G coverage, and LTE devices will use only the LTE network.

Figure 78: Evolution of Voice in an LTE Network[168]

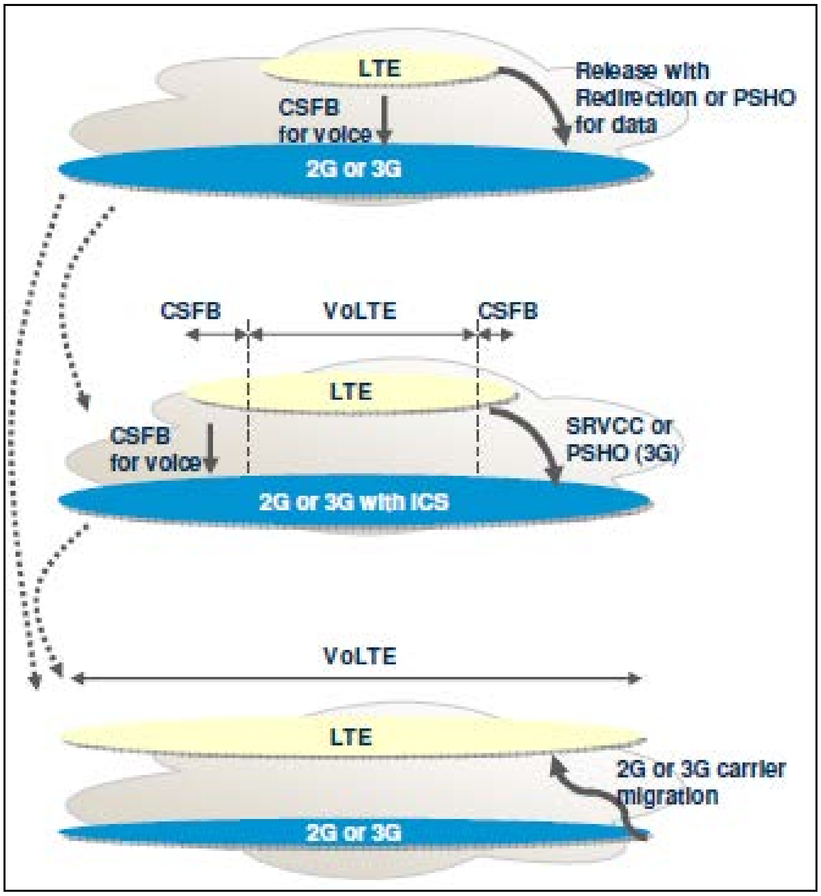
LTE
Release with Redirection or PSHO for data
CSFB for voice
2G or 3G
CSFB
VoLTE
CSFB
LTE
SRVCC or PSHO (3G)
CSFB for voice
2G or 3G with ICS
VoLTE
LTE
2G or 3G carrier migration
2G or 3G

Another voice approach, called “Voice over LTE via Generic Access” (VoLGA), defined circuit-switched operation through an LTE IP tunnel. 3GPP, however, has stopped official standards work that would support VoLGA.

3GPP has developed a new codec, called “Enhanced Voice Services” (EVS), which will include super-wideband voice capability. For the same bit rate, EVS provides higher voice quality than the other codecs.[169] Table 21 summarizes the features and parameters of the three 3GPP codecs used in LTE.

---

[168] 5G Americas member contribution.

[169] See Figure 9.2. 3GPP, *TR 26.952 V12.1.0, Codec for Enhanced Voice Services (EVS); Performance Characterization*, March 2015.

Table 21: Comparison of AMR, AMR-WB and EVS Codecs[170]

Features	AMR	AMR-WB	EVS
Input and output sampling frequencies supported	8KHz	16KHz	8KHz, 16KHz, 32KHz, 48 KHz
Audio bandwidth	Narrowband	Wideband	Narrowband, Wideband, Super-wideband, Fullband
Coding capabilities	Optimized for coding human voice signals	Optimized for coding human voice signals	Optimized for coding human voice and general-purpose audio (music, ringtones, mixed content) signals
Bit rates supported (in kb/s)	4.75, 5.15, 5.90, 6.70, 7.4, 7.95, 10.20, 12.20	6.6, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05, 23.85	5.9, 7.2, 8, 9.6 (NB and WB only), 13.2 (NB, WB and SWB), 16.4, 24.4, 32, 48, 64, 96, 128 (WB and SWB only)
Number of audio channels	Mono	Mono	Mono and Stereo
Frame size	20 ms	20 ms	20 ms
Algorithmic Delay	20-25 ms	25 ms	Up to 32 ms

Figure 79 shows mean opinion scores (MOS) for different codecs at different bit rates, illustrating the advantage of EVS, particularly for bit rates below 32 kbps that cellular networks use.

---

[170] 4G Americas, *Mobile Broadband Evolution: Rel-12 & Rel-13 and Beyond*, 2015. See also T-Mobile 2016 EVS announcement: https://newsroom.t-mobile.com/news-and-blogs/volte-enhanced-voice-services.htm.

Figure 79: Combined Mean Opinion Score Values[171]

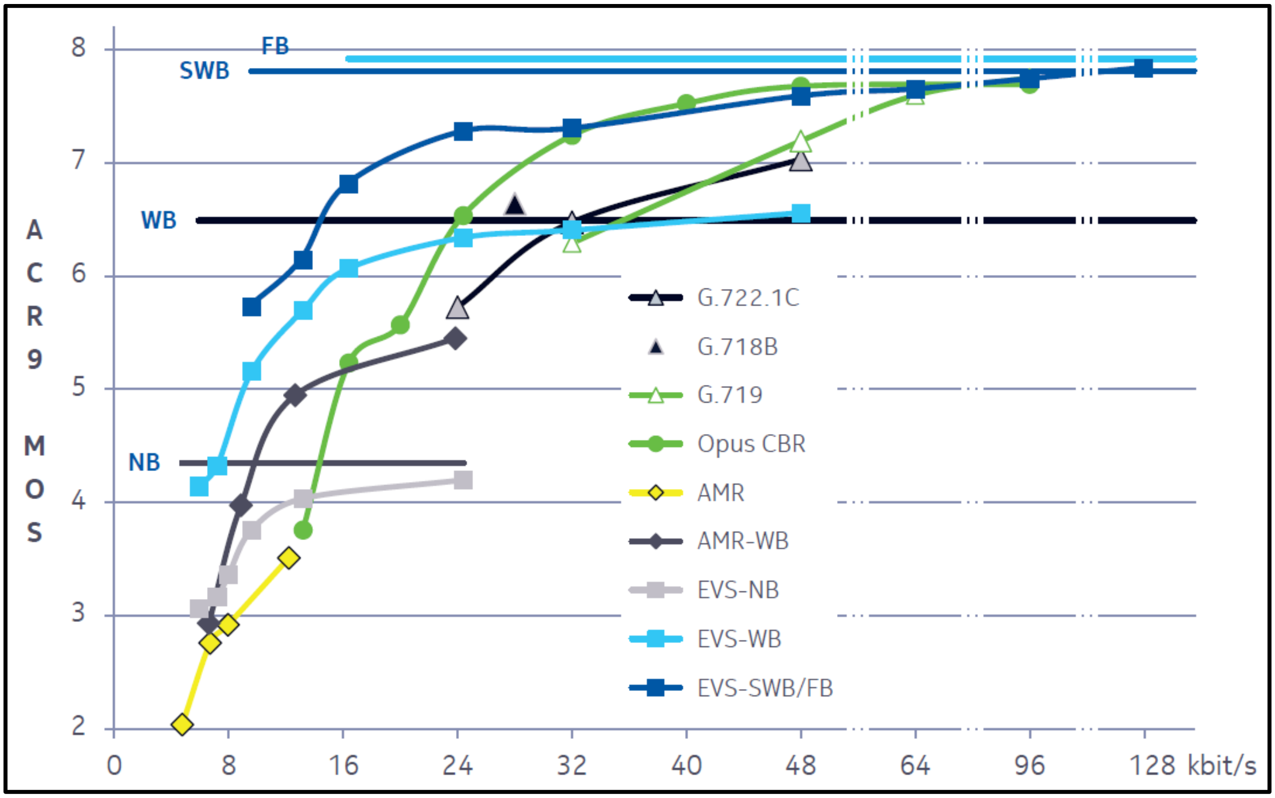

Table 22 shows EVS (narrowband, wideband, super-wideband) audio bandwidths and bitrates that create subjective quality equal to or better than AMR or AMR-WB for typical conversational voice scenarios.

Table 22: EVS Compared to AMR and AMR-WB[172]

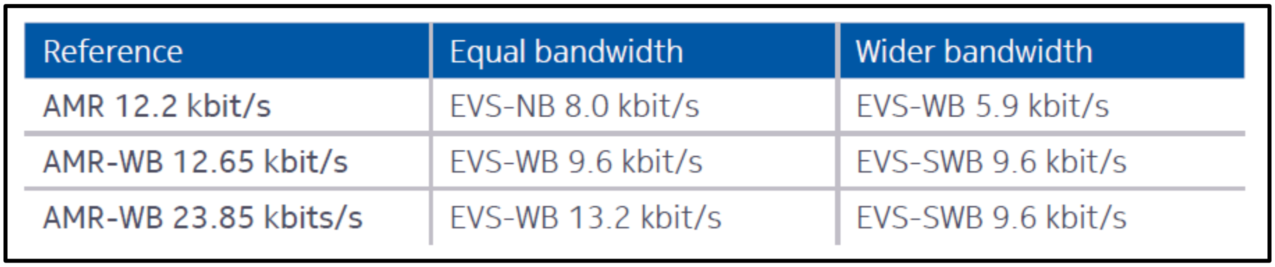

Reference	Equal bandwidth	Wider bandwidth
AMR 12.2 kbit/s	EVS-NB 8.0 kbit/s	EVS-WB 5.9 kbit/s
AMR-WB 12.65 kbit/s	EVS-WB 9.6 kbit/s	EVS-SWB 9.6 kbit/s
AMR-WB 23.85 kbits/s	EVS-WB 13.2 kbit/s	EVS-SWB 9.6 kbit/s

Figure 80 compares EVS capacity gains over AMR and AMR-WB for the references cases shown in Table 22. EVS-SWB at 9.6 kbps almost doubles voice capacity compared to AMW-WB at 23.85 kbps.

---

[171] Nokia*, The 3GPP Enhanced Voice Services (EVS) codec*, 2015.

[172] Ibid.

Figure 80: EVS Voice Capacity Compared to AMR and AMR-WB[173]

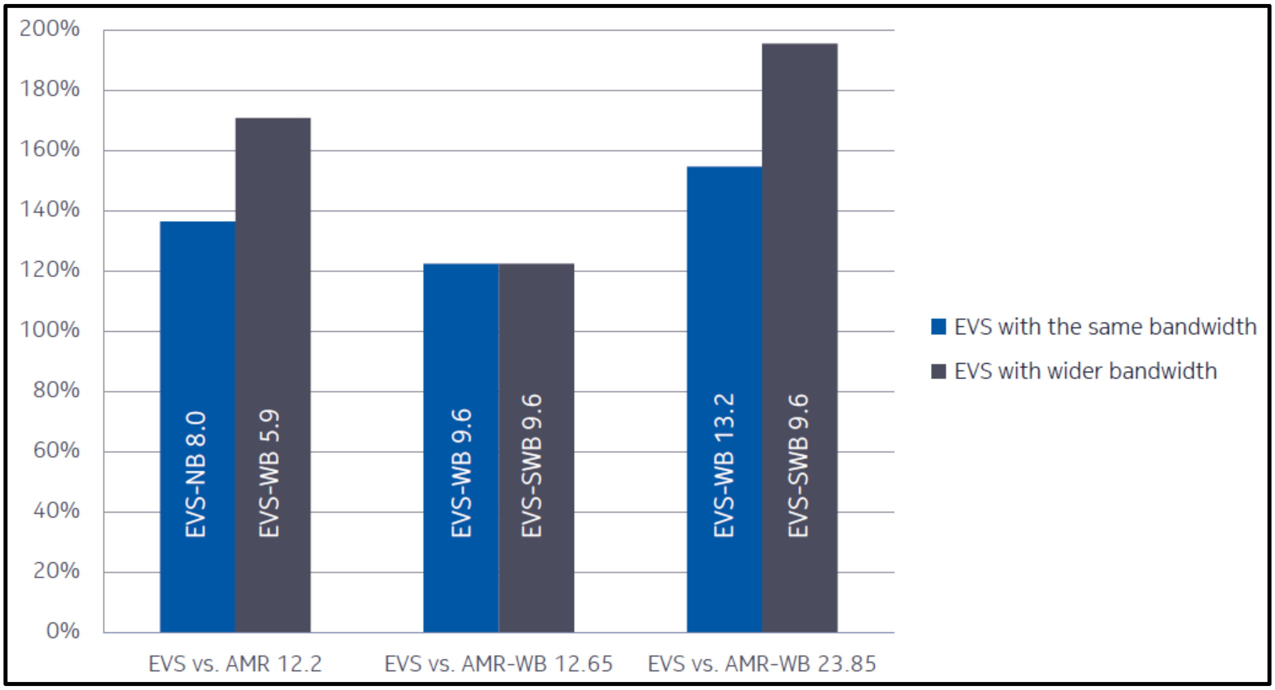

LTE Ultra-Reliable and Low-Latency Communications

The 3GPP work item for this feature states, "3GPP LTE technology enhancements are needed to serve such new use cases and to remain technologically competitive up to and beyond 2020. As a candidate technology for ITU IMT-2020 submission, it is motivated to further enhance the LTE system such that it can meet the key IMT-2020 requirements including those for URLLC in terms of reliability ($1\text{-}10^{-5}$ reliability for small data packets within a latency of 1ms) as well as latency (≤1ms one way user plane latency)."[174]

Evolved Packet Core (EPC)

3GPP defined the Evolved Packet Core (EPC) in Release 8 as a framework for an evolution or migration of the network to a higher-data-rate, lower latency, packet-optimized system that supports multiple radio-access technologies including LTE, as well as and legacy GSM/EDGE and UMTS/HSPA networks. EPC also integrates CDMA2000 networks and Wi-Fi.

EPC is optimized for all services to be delivered via IP in a manner that is as efficient as possible—through minimization of latency within the system, for example. It also provides service continuity across heterogeneous networks, which is important for LTE operators who must simultaneously support GSM-HSPA customers.

One important performance-enhancing aspect of EPC is a flatter architecture. For packet flow, EPC includes two network elements, called "Evolved Node B" (eNodeB) and the Access Gateway (AGW). The eNodeB (base station) integrates the functions traditionally performed by the radio network controller, which previously was a separate node controlling multiple Node Bs. Meanwhile, the AGW integrates the functions traditionally performed by the SGSN and GGSN. The AGW includes both control functions, handled through the Mobile Management Entity (MME), and user plane (data communications) functions. The user plane functions consist of two elements: A serving gateway that addresses 3GPP mobility and terminates eNodeB connections, and a Packet Data Network (PDN) gateway that addresses service requirements and also terminates access by non-3GPP networks. The MME serving gateway and PDN gateways can be collocated in the same physical node or distributed, based on vendor implementations and deployment scenarios.

---

[173] Ibid.

[174] RP-170796, 3GPP Work Item Description, "Ultra Reliable Low Latency Communication for LTE," March 2017.

EPC uses IMS as a component. It also manages QoS across the whole system, an important enabler for voice and other multimedia-based services.

Figure 81 shows the EPC architecture.

Figure 81: EPC Architecture

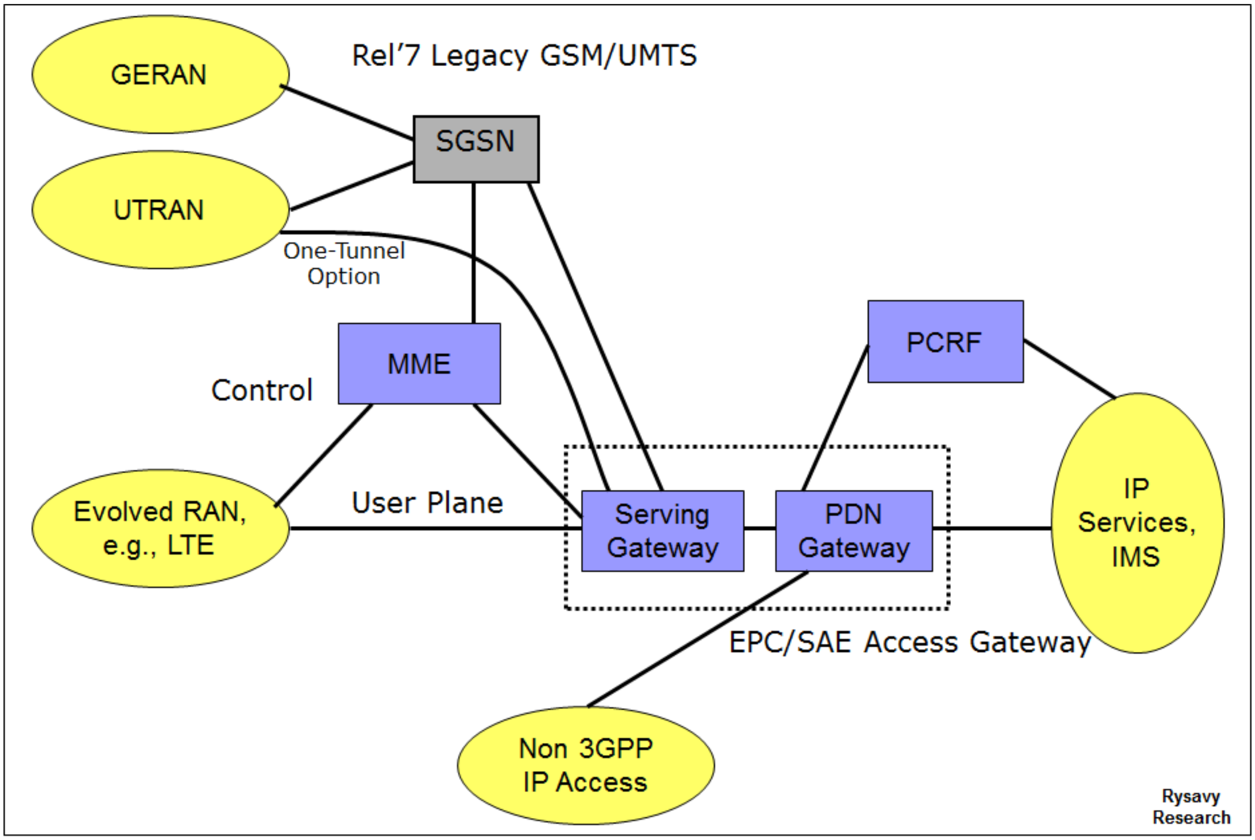
Rel'7 Legacy GSM/UMTS
GERAN
UTRAN
SGSN
One-Tunnel Option
MME
Control
PCRF
Evolved RAN, e.g., LTE
User Plane
Serving Gateway
PDN Gateway
IP Services, IMS
EPC/SAE Access Gateway
Non 3GPP IP Access
Rysavy Research

Elements of the EPC architecture include:

- Support for legacy GERAN and UTRAN networks connected via SGSN.
- Support for new radio-access networks such as LTE.
- Support for non-3GPP networks such as EV-DO and Wi-Fi. (See section below on Wi-Fi integration).
- The Serving Gateway that terminates the interface toward the 3GPP radio-access networks.

- The PDN gateway that controls IP data services, does routing, allocates IP addresses, enforces policy, and provides access for non-3GPP access networks.
- The MME that supports user equipment context and identity, as well as authenticating and authorizing users.
- The Policy Control and Charging Rules Function that manages QoS aspects.

QoS in EPS employs the QoS Class Identifier (QCI), a number denoting a set of transport characteristics (bearer with/without guaranteed bit rate, priority, packet delay budget, packet error loss rate) and used to infer nodes specific parameters that control packet forwarding treatment (such as scheduling weights, admission thresholds, queue management thresholds, or link-layer protocol configuration). The network maps each packet flow to a single QCI value (nine are defined in the Release 8 version of the specification) according to the level of service required by the application. Use of the QCI avoids the transmission of a full set of QoS-related parameters over the network interfaces and reduces the complexity of QoS negotiation. The QCI, together with Allocation Retention Priority (ARP) and, if applicable, Guaranteed Bit Rate (GBR) and Maximum Bit Rate (MBR), determines the QoS associated to an EPS bearer. A mapping between EPS and pre-Release 8 QoS parameters permits interworking with legacy networks.

The QoS architecture in EPC enables a number of important capabilities for both operators and users:

- **VoIP support with IMS.** QoS is a crucial element for providing LTE/IMS voice service. (See section below on IMS).
- **Enhanced application performance.** Applications such as gaming or video can operate more reliably.
- **More flexible business models.** With flexible, policy-based charging control, operators and third parties will be able to offer content in creative new ways. For example, an enhanced video stream to a user could be paid for by an advertiser.
- **Congestion control.** In congestion situations, certain traffic flows (bulk transfers, abusive users) can be throttled down to provide a better user experience for others.

Table 23 shows the nine QCI used by LTE.

Table 23: LTE Quality of Service

QCI	Resource Type	Priority	Delay Budget	Packet Loss	Examples
1	GBR (Guaranteed Bit Rate)	2	100 msec.	10^{-2}	Conversational voice
2	GBR	4	150 msec.	10^{-3}	Conversational video (live streaming)
3	GBR	3	50 msec.	10^{-3}	Real-time gaming
4	GBR	5	300 msec.	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100 msec.	10^{-6}	IMS signaling

QCI	Resource Type	Priority	Delay Budget	Packet Loss	Examples
6	Non-GBR	6	300 msec.	10^{-6}	Video (buffered streaming), TCP Web, email, and FTP
7	Non-GBR	7	100 msec.	10^{-3}	Voice, video (live streaming), interactive gaming
8	Non-GBR	8	300 msec.	10^{-6}	Premium bearer for video (buffered streaming), TCP Web, e-mail, and FTP
9	Non-GBR	9	300 msec.	10^{-6}	Default bearer for video, TCP for non-privileged users

Heterogeneous Networks and Small Cells

A fundamental concept in the evolution of next-generation networks is the blending of multiple types of networks to create a "network of networks" characterized by:

- ❑ Variations in coverage areas, including femtocells (either enterprise femtos or home femtos, called HeNBs), picocells (also referred to as metro cells), and macro cells. Cell range can vary from 10 meters to 50 kilometers.
- ❑ Different frequency bands.
- ❑ Different technologies spanning Wi-Fi, 2G, 3G, 4G, and 5G.
- ❑ Relaying capability in which wireless links can serve as backhaul.

Figure 82 shows how user equipment might access different network layers.

Figure 82: Load Balancing with Heterogeneous Networks[175]

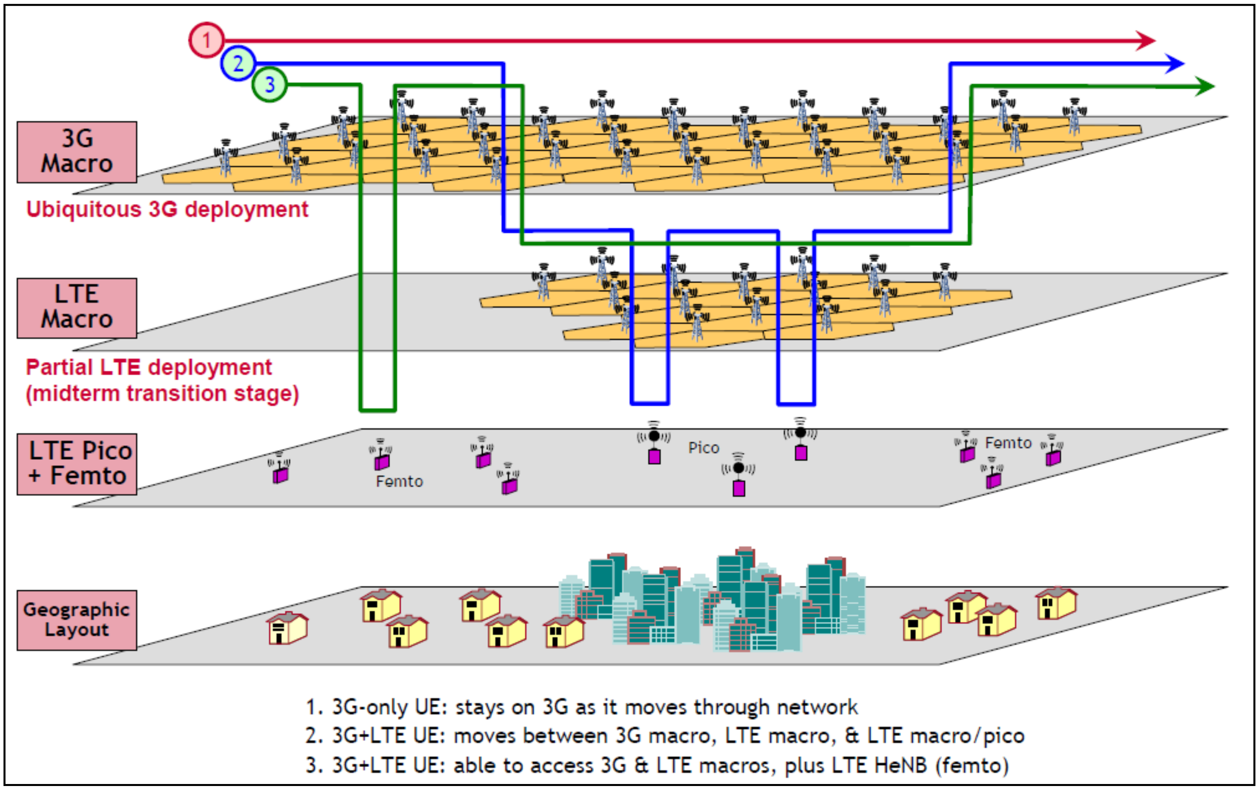
1
2
3
3G Macro
Ubiquitous 3G deployment
LTE Macro
Partial LTE deployment (midterm transition stage)
LTE Pico + Femto
Femto
Pico
Femto
Geographic Layout
1. 3G-only UE: stays on 3G as it moves through network
2. 3G+LTE UE: moves between 3G macro, LTE macro, & LTE macro/pico
3. 3G+LTE UE: able to access 3G & LTE macros, plus LTE HeNB (femto)

HetNets will allow significant capacity expansion in configurations in which operators can add picocells to coverage areas served by macrocells, particularly if there are hot spots with higher user densities.

Small cells differentiate themselves from macrocells according to the parameters shown in Table 24.

Table 24: Small Cell Vs. Macro Cell Parameters: Typical Values

Parameter	Small Cell	Macro Cell
Transmission Power	24 dBm (0.25 W)	43 dBm (20 W)
Antenna Gain	2 dBi	15 dBi
Users	Tens	Hundreds
Mobility	30 km/hr	350 km/hr

---

[175] 5G Americas member contribution.

Whether or not the small cell uses the same radio carriers as the macro cell involves multiple tradeoffs. In Figure 83 Scenario 1, the small cells and macro cell use different radio carriers, the two not interfering with each other. Although this configuration requires more spectrum, the small cells are able to cover larger areas than if they were deployed using the same radio carrier as the macro. This configuration supports medium-to-high penetration levels of small cells, allowing the network to reach huge capacity.

In Scenario 2, the small cells and macro cells use the same radio carrier, accommodating operators with more limited spectrum, but the network must manage interference using the techniques discussed below. Operators must carefully manage small-cell transmission power in this configuration.

Figure 83: Scenarios for Radio Carriers in Small Cells

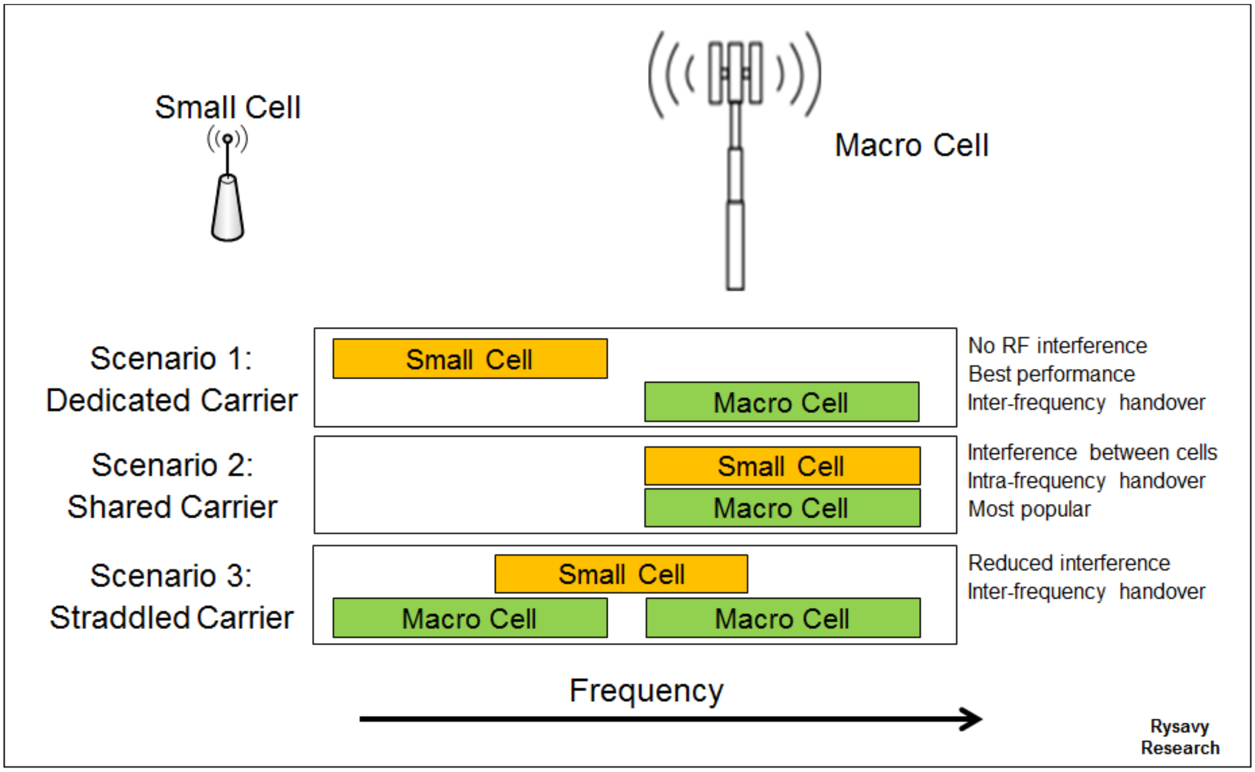
Small Cell
Macro Cell
Scenario 1:
Dedicated Carrier
Small Cell
Macro Cell
No RF interference
Best performance
Inter-frequency handover
Scenario 2:
Shared Carrier
Small Cell
Macro Cell
Interference between cells
Intra-frequency handover
Most popular
Scenario 3:
Straddled Carrier
Small Cell
Macro Cell
Macro Cell
Reduced interference
Inter-frequency handover
Frequency
Rysavy
Research

In Scenario 3, the small cells use a straddled radio carrier, accommodating operators with more spectrum, but the network still needs to manage interference using techniques discussed below. Compared with a shared carrier configuration, this configuration has benefits similar to dedicated carriers in terms of radio-parameter planning and reduced interference.

Figure 84 shows two different traffic distribution scenarios, with a uniform distribution of devices in the first and higher densities serviced by picocells in the second. The second scenario can result in significant capacity gains as well as improved user throughput.

Figure 84: Different Traffic Distributions Scenarios

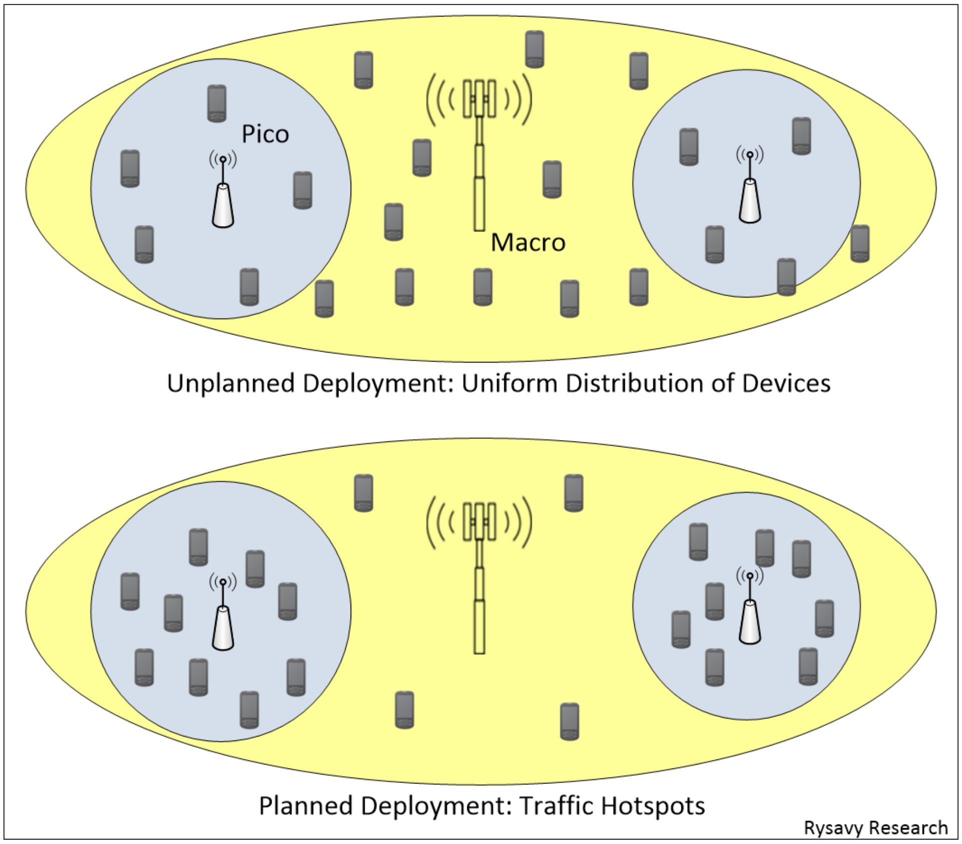
Pico
Macro
Unplanned Deployment: Uniform Distribution of Devices
Planned Deployment: Traffic Hotspots
Rysavy Research

One vendor calculated expected HetNet gains assuming no eICIC, no picocell range extension, and no eICIC. For the case of four picocells without picocell range extension and uniform user distribution, the median-user-throughput gain compared with a macro-only configuration was 85%. For a similar case of four picocells but using a hotspot user distribution, the gain was much higher, 467%.[176] Additional gains will occur with picocell range extension.

Expected picocell gains rise proportionally to the number of picocells, so long as a sufficient number of UEs connect to the picocells.

Release 10 and Release 11 added enhanced support to manage the interference in the HetNet scenario in the time domain with Enhanced Inter-cell Interference Coordination (eICIC) and Further Enhanced Intercell Interference Coordination (feICIC), as well as in the frequency domain with carrier-aggregation-based ICIC.

HetNet capability keeps becoming more sophisticated through successive 3GPP releases as summarized in Table 25.

---

[176] 5G Americas member contribution. Further assumes 2X1 W picocell transmit power, cell-edge placement (planned picocell deployment), 67% of all the users within 40m of the pico locations, and 3GPP Technical Report 36.814 adapted to 700 MHz.

Table 25: 3GPP HetNet Evolution

3GPP Release	HetNet Feature
8	Initial SON capabilities, most for auto configuration. Initial intercell interference coordination (ICIC) available.
9	More mobility options (for example, handover between HeNBs), operator customer subscriber group (SCG) lists, load-balancing, coverage and capacity improvements.
10	An interface for HeNBs, called “Iurh,” that improves coordination and synchronization, LTE time domain eICIC. Carrier-aggregation-based ICIC also defined.
11	Improved eICIC, further mobility enhancements.

Enhanced Intercell Interference Coordination

Significant challenges must be addressed in these heterogeneous networks. One is near-far effects, in which local small-cell signals can easily interfere with macro cells if they are using the same radio carriers.

Interference management is of particular concern in HetNets since, by design, coverage areas of small coverage cells overlap with the macro cell. Beginning with Release 10, eICIC introduces an approach of almost-blank subframes by which subframe transmission can be muted to prevent interference. Figure 85 illustrates eICIC for the macro layer and pico layer coordination. If a UE is on a picocell but in a location where it is sensitive to interference from the macro layer, the macro layer can mute its transmission during specific frames when the pico layer is transmitting.

Figure 85: Example of Enhanced Intercell Interference Coordination[177]

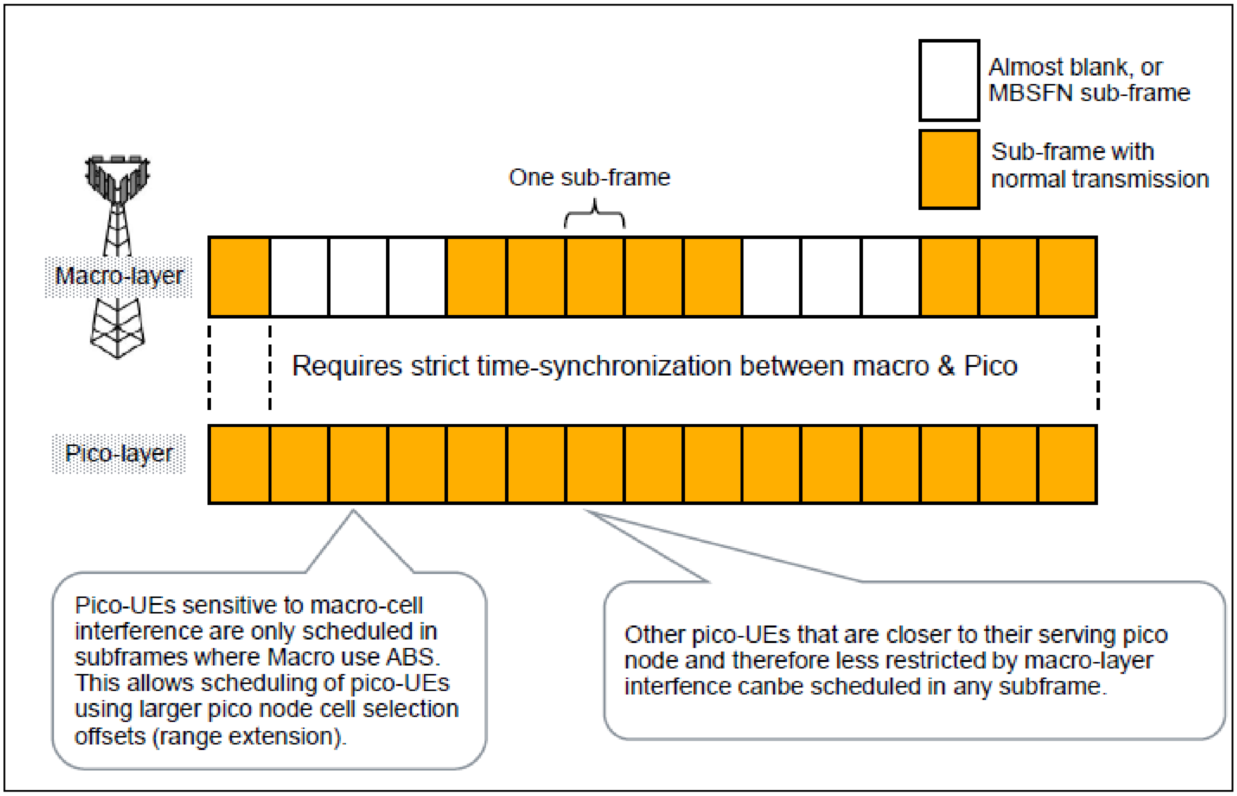
Almost blank, or MBSFN sub-frame
Sub-frame with normal transmission
One sub-frame
Macro-layer
Requires strict time-synchronization between macro & Pico
Pico-layer
Pico-UEs sensitive to macro-cell interference are only scheduled in subframes where Macro use ABS. This allows scheduling of pico-UEs using larger pico node cell selection offsets (range extension).
Other pico-UEs that are closer to their serving pico node and therefore less restricted by macro-layer interfence canbe scheduled in any subframe.

LTE can also combine eICIC with interference-cancellation-based devices to minimize the harmful effects of interference between picocells and macro cells.

Figure 86 shows one 4G America member's analysis of anticipated median throughput gains using picocells and Release 11 Further Enhanced ICIC.

---

[177] 5G Americas member contribution.

Figure 86: Median Throughput Gains in Hotspot Scenarios[178]

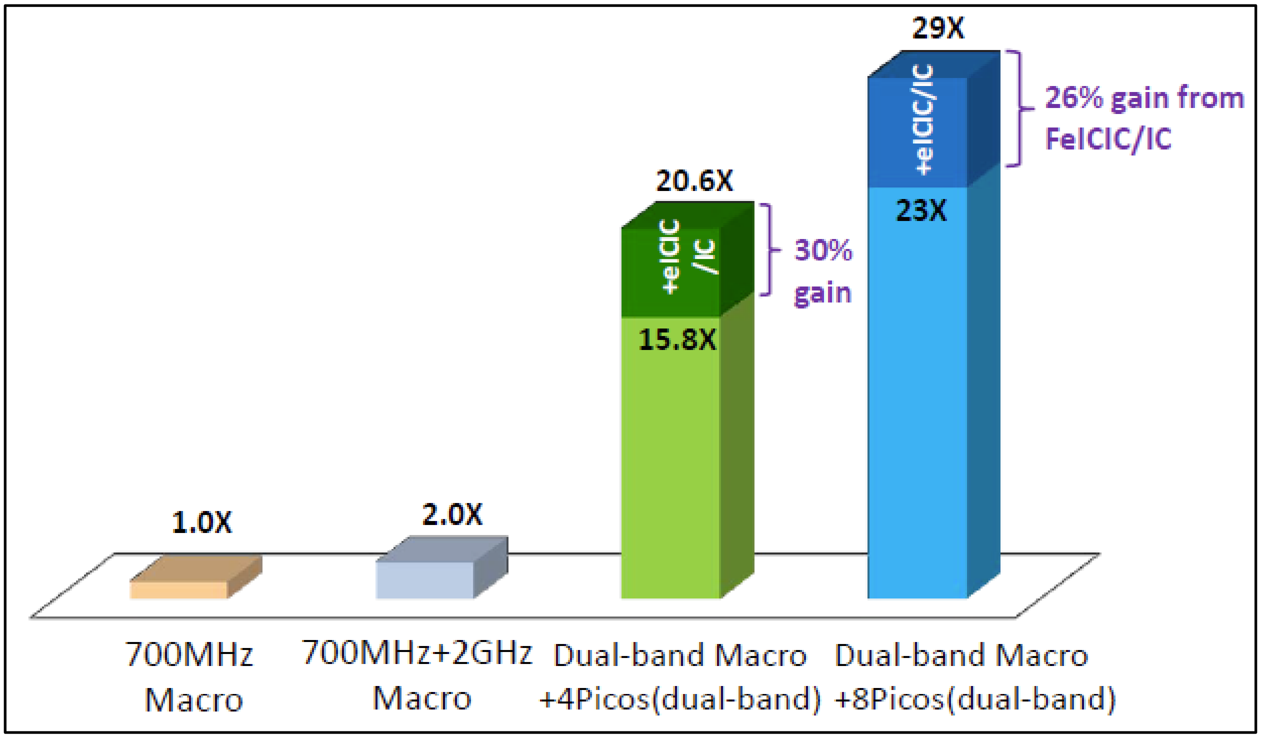

FeICIC is also beneficial in non-hotspot scenarios. In the case of a uniform distribution of picocells, this same 5G Americas member estimates a 130% gain from FeICIC for an eight picocell per macro-cell scenario, increasing capacity from a factor of 3.3 for the picocells alone to a factor of 7.6 with the addition of FeICIC.[179]

Further insight is available from Figure 87, which shows 5 percentile and 50 percentile throughput with and without eICIC under different conditions of range extension and almost blanked subframes.

---

[178] 5G Americas member contribution. Assumes 3GPP evaluation methodology TR 36.814, carrier-aggregation UEs, macro ISD = 1732m, 700 MHz and 2GHz carrier frequency, full-buffer traffic, FDD 10+10 MHz per carrier, 6-degree antenna downtilt, 4 or 8 Picos and 30 UEs per Macro cell, hotspot distribution with 20 of 30 UEs near picos, PF scheduler, 2x2 MIMO, TU3 channel, NLOS, local partitioning algorithm.

[179] Assumes 3GPP evaluation methodology TR 36.814, macro ISD = 1732m, 700 MHz and 2GHz carrier frequency, full-buffer traffic, 6-degree antenna downtilt, 30 carrier-aggregation UEs per Macro cell, uniform random layout, PF scheduler, FDD, 10+10 MHz per carrier, 2x2 MIMO, TU3 channel, NLOS, local partitioning algorithm. Additional information is available at ftp://ftp.3gpp.org/tsg_ran/WG1_RL1/TSGR1_66b/Docs/R1-113383.zip.

Figure 87: User Throughput Performance With / Without eICIC for Dynamic Traffic vs. Average Offered Load per Macro Cell Area[180]

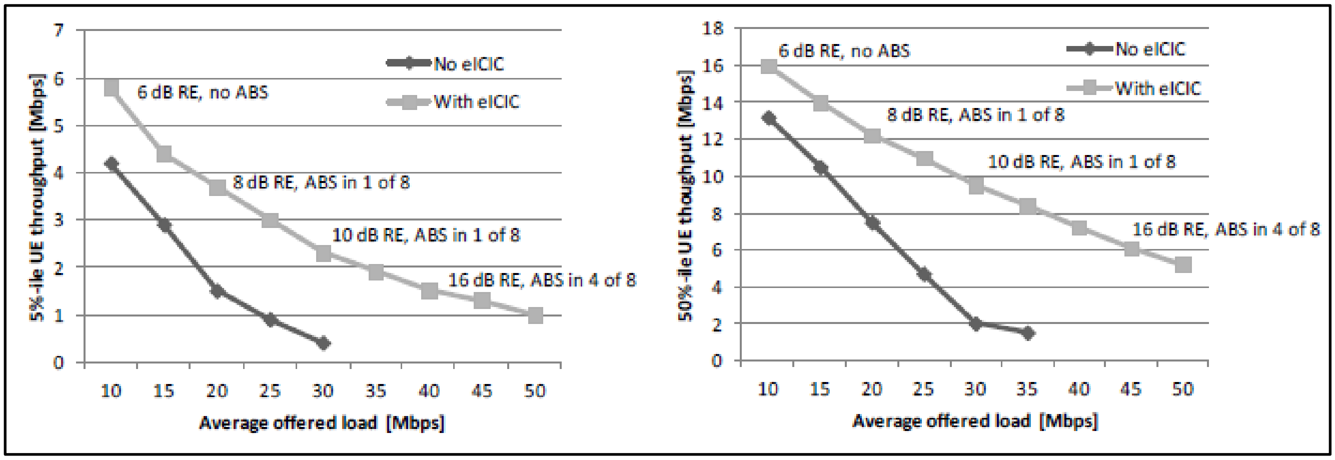

The muting of certain subframes in eICIC is dynamic and depends on identifying, on a per user basis, whether an interfering cell's signal exceeds a threshold relative to the serving cell signal. Coordinating muting among small cells can be complicated because a small cell can simultaneously be an interferer while serving a UE that is a victim of another cell. The network must therefore coordinate muting among multiple small cells.

Figure 88 below at left shows user throughput gains of time domain interference relative to network load. Throughput gains are higher at higher network loads because of more active users and the higher likelihood of interference between the small cells.

Figure 88 below at right shows the maximum muting ratio, which increases with higher network load.

---

[180] 5G Americas member contribution. Assumes 3GPP evaluation methodology TR 36.814, 500 meter ISD, 4 picos per macro-cell area, Poisson call arrival, finite payload for each call, and termination of call upon successful delivery.

Figure 88: Throughput Gain of Time-Domain Interference Coordination[181]

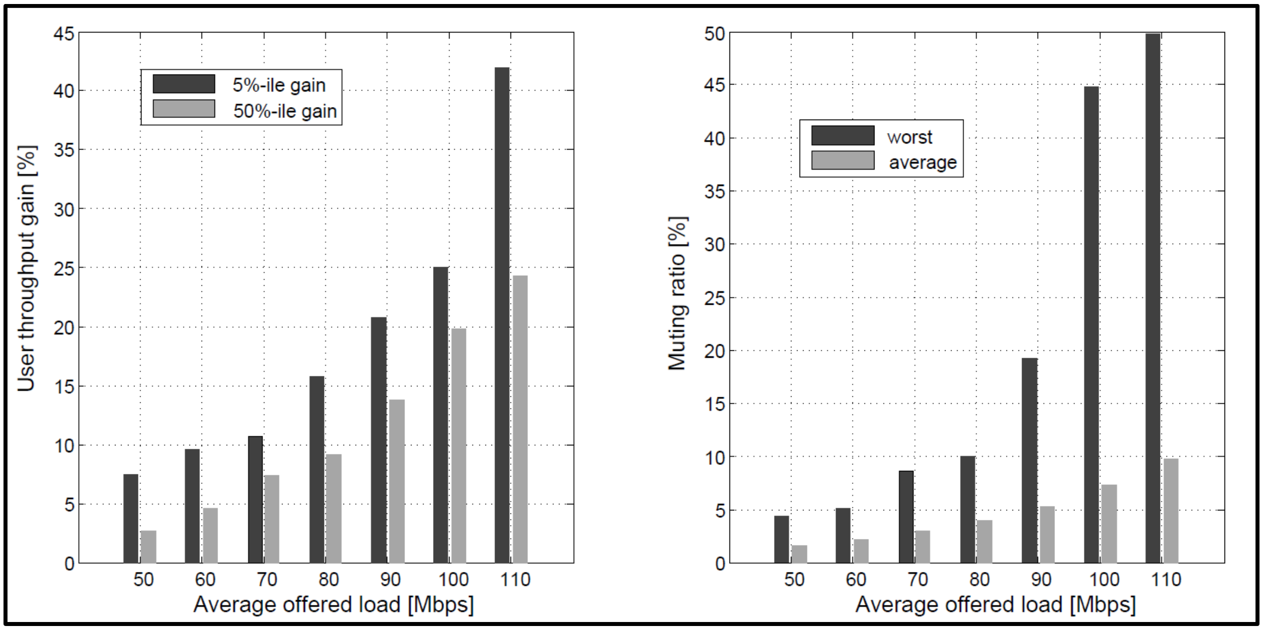

Another approach for addressing inter-layer interference cancellation in HetNets can come from carrier aggregation with no further additions or requirements and realizable with Release 10 LTE networks. Consider the scenario in Figure 89, in which both the macro eNB and the pico eNB are allocated two component carriers (namely CC1 and CC2). The idea is to create a "protected" component carrier for downlink control signals and critical information (Physical Downlink Control Channel, system information, and other control channels) while data can be conveniently scheduled on both component carriers through cross-carrier scheduling.

Figure 89: Carrier-Aggregation Based ICIC[182]

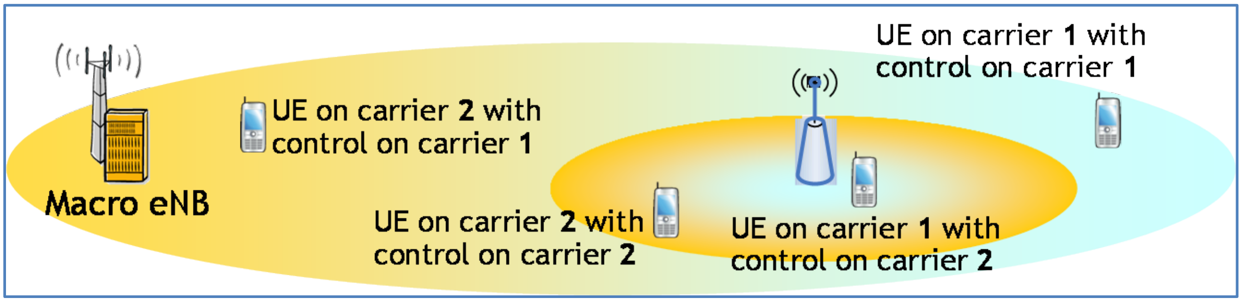
Macro eNB
UE on carrier **2** with control on carrier **1**
UE on carrier **1** with control on carrier **1**
UE on carrier **2** with control on carrier **2**
UE on carrier **1** with control on carrier **2**

CC1 is the primary component carrier for the macro cell, while CC2 is the primary for the picocell; hence the protected carriers are CC1 for the macro cell and CC2 for the picocell. The macro cell allocates a lower transmission power for its secondary CC in order to reduce interference to the picocell's primary component carrier. The network can schedule data on both the primary and secondary component carriers. In the figure, users in the cell range expansion (CRE) zone can receive data via cross-carrier scheduling from the secondary CC at subcarrier frequencies on which interference from the other cell can be reduced if the cells exchange appropriate signaling over what is called an "X2 interface." Users operating close to the eNodeBs can receive data from both component carriers as their interference levels will hopefully be lower. Therefore, a CA-capable receiver will enjoy the enhanced throughput capabilities of carrier aggregation, while simultaneously receiving extra protection for control and data channels at locations with potentially high inter-layer interference.

---

[181] 5G Americas member contribution. Simulations based on 12 densely deployed small cells at 3.5 GHz and 3GPP Release 12 simulation assumptions in TR 36.842.

[182] 5G Americas member contribution.

Thus, carrier aggregation can be a useful tool for deployment of heterogeneous networks without causing a loss of bandwidth. These solutions, however, do not scale well (in Release 10 systems) to small system bandwidths (say, 3+3 MHz or 1.4+1.4 MHz radio carriers) because control channels occupy a high percentage of total traffic. Additionally, interference between the cell reference signals (CRS) would also be significant.

Dual Connectivity

A major enhancement in Release 12 is a UE being served at the same time by both a macro cell and a small cell operating at different carrier frequencies, a capability called dual connectivity and illustrated in Figure 90. Data first reaches the macro eNodeB and is split, with part of it transmitted from the macro and the balance sent via an X2 interface to the small cell for transmission to the UE.

Figure 90: Dual Connectivity[183]

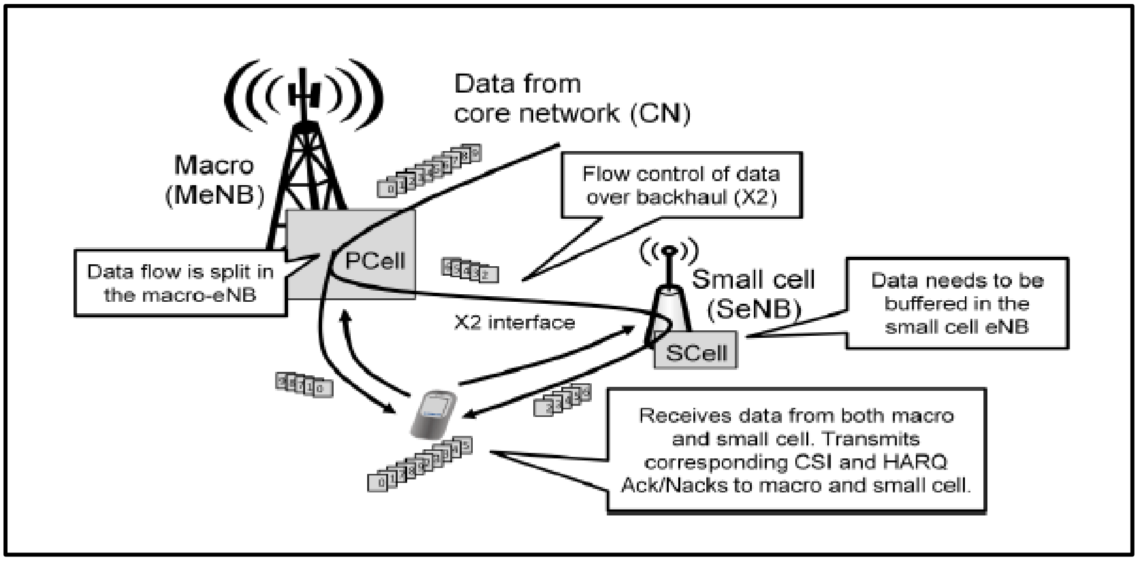
Data from
core network (CN)
Macro
(MeNB)
Flow control of data
over backhaul (X2)
PCell
Data flow is split in
the macro-eNB
Small cell
(SeNB)
Data needs to be
buffered in the
small cell eNB
X2 interface
SCell
Receives data from both macro
and small cell. Transmits
corresponding CSI and HARQ
Ack/Nacks to macro and small cell.

Figure 91 shows throughput gains of dual connectivity at 5 percentile and 50 percentile (median) levels relative to the load on the network and different degrees of latency in the X2 interface. Benefits are higher with lower network load and with lower X2 latency.

---

[183] Source: 5G Americas member contribution.

Figure 91: Dual Connectivity User Throughput[184]

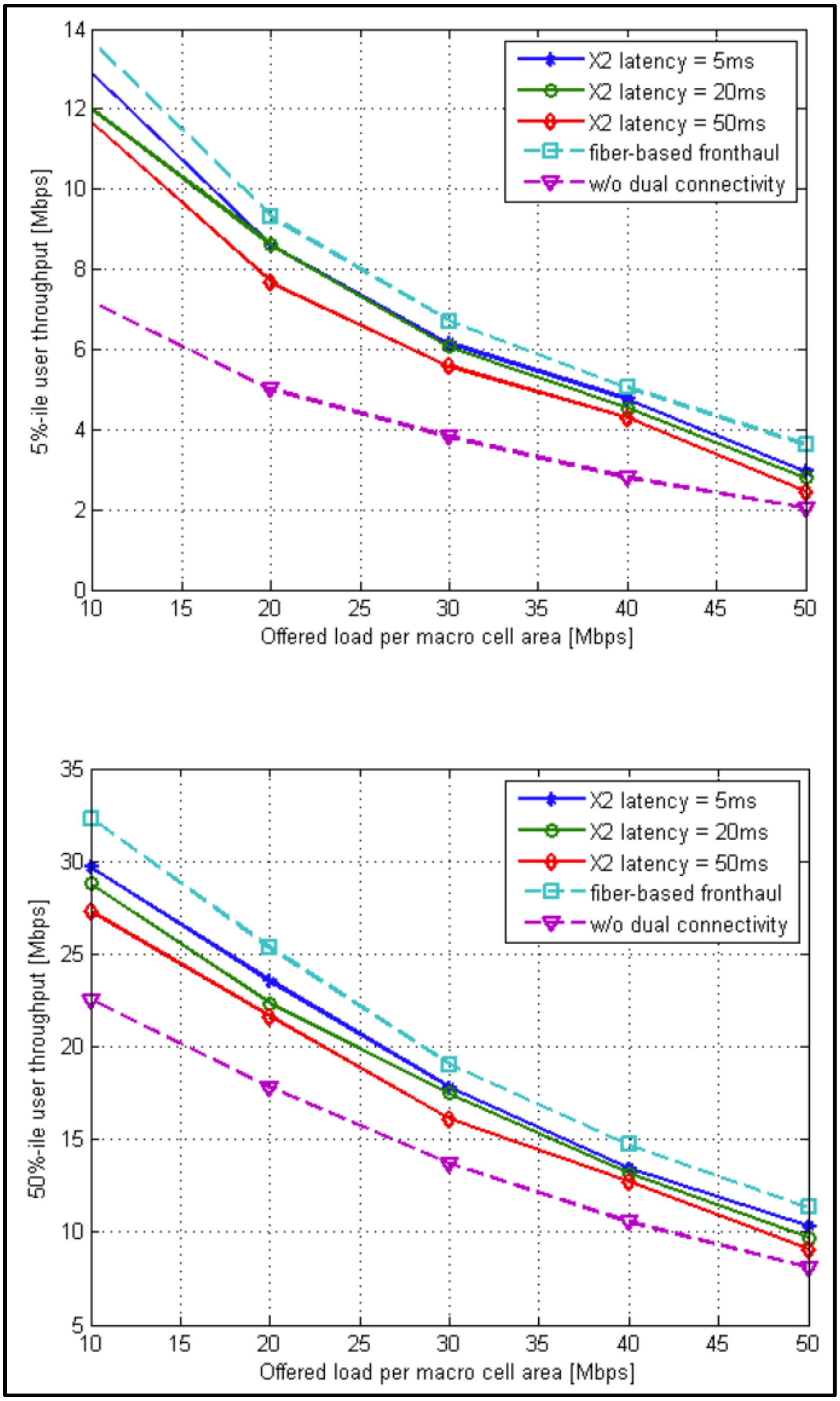

---

[184] 5G Americas member contribution.

Internet of Things and Machine-to-Machine

Anticipating huge growth in machine-to-machine communications, Release 11 added a Machine Type Communications (MTC) Interworking Function and Service Capability Server. Release 12 defined a category 0 device designed to deliver low cost through a single antenna design and other simplifications.[185] Release 13 went even further, with a category M-1 architecture that further reduces cost, improves range, and extends battery life. Category 13 also added Narrowband-IoT capability with Category NB-1 and an IoT solution for GSM, called "EC-GSM-IoT," that extends coverage by 20 dB. Category M-1 and NB-IoT devices could achieve battery life as high as 10 years.

Figure 92 depicts the methods used to reduce cost in a Category M device compared with a Category 4 device.

Figure 92: Means of Achieving Lower Cost in IoT Devices[186]

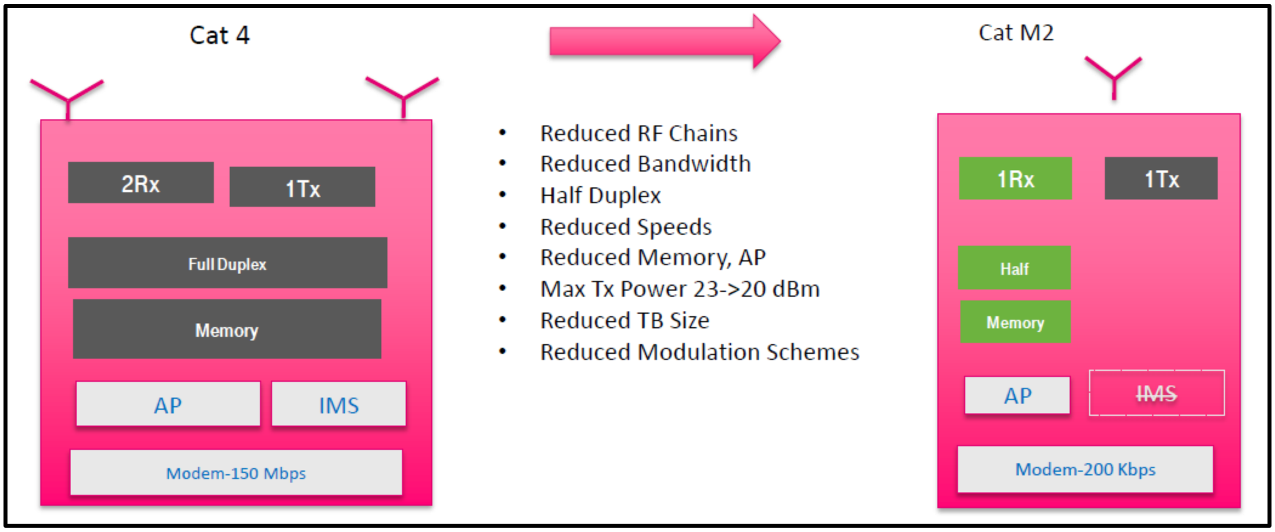

- Reduced RF Chains
- Reduced Bandwidth
- Half Duplex
- Reduced Speeds
- Reduced Memory, AP
- Max Tx Power 23->20 dBm
- Reduced TB Size
- Reduced Modulation Schemes

Table 26 summarizes the features of different LTE IoT devices based on 3GPP Release.

Table 26: Summary of IoT Features in LTE Devices

Device Category	Category 3	Category 1	Category 0	Category M-1	Category NB-1	EC-GSM-IoT
3GPP Release	10	11	12	13	13	13

---

[185] 3GPP, *Access System for Ultra Low Complexity and Low Throughput Internet of Things based on Cellular*, GP-140301, May 2014.

[186] 5G Americas member contribution.

Device Category	Category 3	Category 1	Category 0	Category M-1	Category NB-1	EC-GSM-IoT
Max. Data Rate Downlink	100 Mbps	10 Mbps	1 Mbps	1 Mbps	200 Kbps	74 Kbps
Max. Data Rate Uplink	50 Mbps	5 Mbps	1 Mbps	1 Mbps	200 Kbps	74 Kbps
Max. Bandwidth	20 MHz	20 MHz	20 MHz	1.08 MHz	0.18 MHz	0.2 MHz
Duplex	Full	Full	Optional half-duplex	Optional half-duplex	Half	Half
Max. Receive Antennas	Two	Two	One	One	One	One
Power		Power Save Mode[187]	Power Save Mode	Power Save Mode		
Sleep				Longer sleep cycles using Idle Discontinuous Reception (DRX)		
Coverage				Extended through redundant transmissions and Single Frequency Multicast		

Cloud Radio-Access Network (RAN) and Network Virtualization

Still in the early stages of development, cloud RAN (C-RAN) is a distributed architecture in which multiple remote radio heads connect to a “cloud” that consists of a farm of baseband processing nodes. This approach can improve centralized processing, as is needed for

---

[187] Power Save Mode specified in Release 12, but applicable to Category 1 device configured as Release 12.

CoMP, centralized scheduling, and Multiflow, without the need to exchange information among many access nodes. The performance of both LTE and HSPA technologies could be enhanced by the application of cloud RAN architectures. The term "fronthauling" has been used to describe the transport of "raw" radio signals to central processing locations, such as between the Physical Network Function (PNF) and a Virtual Network Function (VNF).

This architecture, shown in Figure 93, comes at the cost of requiring high-speed, low-latency backhaul links between these radio heads and the central controller. One vendor states that carrying 10+10 MHz of LTE with 2X2 MIMO requires 2.5 Gbps of bandwidth and imposes less than 0.1 msec of delay.[188] A standard called "Common Public Radio Interface" (CPRI) addresses generic formats and protocols for such a high-speed link. ETSI has also developed the Open Radio Equipment Interface (ORI). The feasibility of cloud RAN depends to a large extent on the cost and availability of fiber links between the remote radio heads and the centralized baseband processing location.

Unlike virtualizing the EPC, in which the entirety of the function can be virtualized, cloud RAN needs a PNF that terminates the RF interface. Cloud RAN therefore requires a split to be defined within the RAN. As a consequence, initial deployments of cloud RAN have looked to ruse the CPRI interface between the RRH and the baseband unit.

Figure 93: Potential Cloud RAN Approach

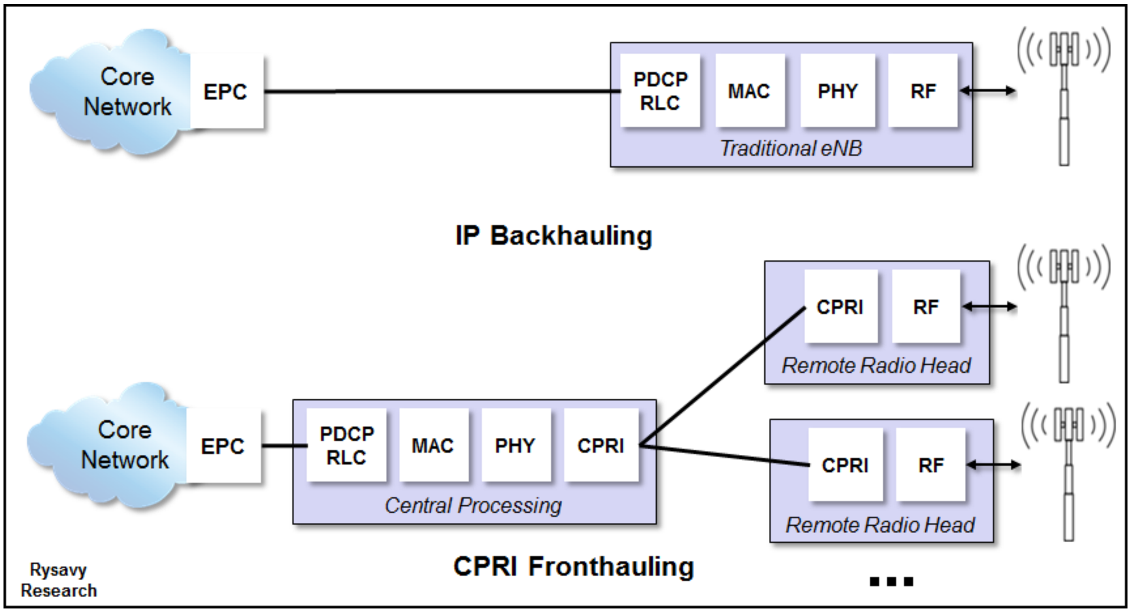

The next evolutionary step after centralizing baseband processing is to virtualize the processing by implementing the functions in software on commodity computing platforms, thus abstracting the functions from any specific hardware implementation.

C-RANs can vary by the extent of coverage, ranging from being highly localized and operating across a small number of sites to metropolitan-wide solutions. Other variables include existing deployments versus greenfield situations, new LTE and 5G technologies versus integrating legacy 2G and 3G technologies, and integrating Wi-Fi. Greater scope increases complexity but yields benefits including better load-balancing and greater flexibility in spectrum re-farming.

---

[188] Dudu Bercovich, Ran Avital, "Holistic HetNet Hauling (3H)," Ceragon, February 2013. Available at http://www.ceragon.com/images/Reasource_Center/White_Papers/Ceragon_Holistic_Hetnet_Hauling_White_Paper.pdf.

Another design choice, as detailed in Table 27, is whether to centralize Layer 1 and Layer 2 functions (an RF-PHY split), or whether to keep Layer 1 at the base stations and centralize only Layer 2 (a PHY-MAC split).

Table 27: Partially Centralized Versus Fully Centralized C-RAN

	Fully Centralized	Partially Centralized
Transport Requirements	Multi-Gbps, usually using fiber	20 to 50 times less
Fronthaul Latency Requirement	Less than 100 microseconds	Greater than 5 milliseconds
Applications	Supports eICIC and CoMP	Supports centralized scheduling
Complexity	High	Lower
Benefit	Capacity gain	Lower capacity gain

Figure 94 analyzes the different possible RAN decompositions in greater detail.

Figure 94: Costs and Benefits of Various RAN Decompositions[189]

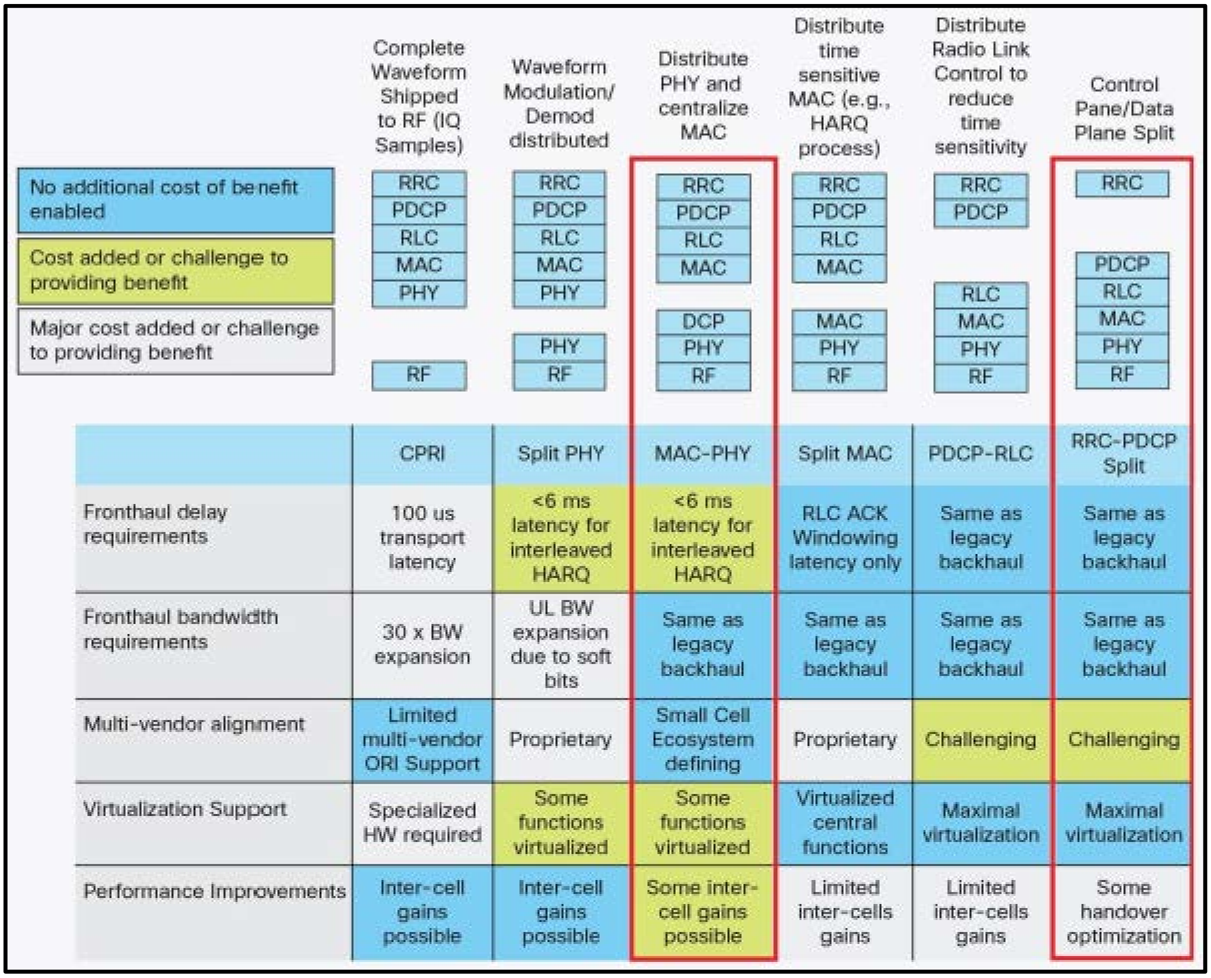

RRC
PDCP
RLC
MAC
PHY
RF
DCP

🟦 No additional cost of benefit enabled

🟨 Cost added or challenge to providing benefit

⬜ Major cost added or challenge to providing benefit

	Complete Waveform Shipped to RF (IQ Samples)	Waveform Modulation/ Demod distributed	Distribute PHY and centralize MAC	Distribute time sensitive MAC (e.g., HARQ process)	Distribute Radio Link Control to reduce time sensitivity	Control Pane/Data Plane Split
	CPRI	Split PHY	MAC-PHY	Split MAC	PDCP-RLC	RRC-PDCP Split
Fronthaul delay requirements	100 us transport latency	<6 ms latency for interleaved HARQ	<6 ms latency for interleaved HARQ	RLC ACK Windowing latency only	Same as legacy backhaul	Same as legacy backhaul
Fronthaul bandwidth requirements	30 x BW expansion	UL BW expansion due to soft bits	Same as legacy backhaul	Same as legacy backhaul	Same as legacy backhaul	Same as legacy backhaul
Multi-vendor alignment	Limited multi-vendor ORI Support	Proprietary	Small Cell Ecosystem defining	Proprietary	Challenging	Challenging
Virtualization Support	Specialized HW required	Some functions virtualized	Some functions virtualized	Virtualized central functions	Maximal virtualization	Maximal virtualization
Performance Improvements	Inter-cell gains possible	Inter-cell gains possible	Some inter-cell gains possible	Limited inter-cells gains	Limited inter-cells gains	Some handover optimization

Although some operators in dense deployments with rich fiber assets may centralize all functions, Figure 94 uses the red rectangles to show the two most likely functional splits for LTE-Advanced and 5G:

1. **Distributed PHY and Centralized MAC.** This approach relaxes the fronthaul delay requirement to 6 msec, compared with the CPRI requirement of 250 microseconds. Fronthaul bandwidth requirement is only 10-20% greater than conventional backhaul.
2. **Control Plane/Data Plane Split.** This approach further relaxes fronthaul requirements to 30 msec and is the approach used for dual-connectivity, such as a macro and small cell simultaneously connecting to a user. For 5G, 3GPP Release 15 specifications standardize a split for cloud RAN between the PDCP and RLC layers.

---

[189] Cisco, *Cisco 5G Vision Series: Small Cell Evolution*, 2016.

Next Generation Mobile Networks studied the pros and cons of different fronthauling interfaces and published the results in March 2015.[190]

Longer-term, perhaps in the 5G context, virtualized C-RANs may take away the very concept of cells. With methods such as beamforming and device-to-device communication, coverage may extend dynamically from a multitude of sources based on instantaneous load notifications and the radio resources available at different nodes.

In the past, RAN and core networks have been distinct entities, but over the next decade, the two may merge with more centralized, virtualized, and cloud-driven approaches.

Another form of virtualization is software-defined networking, an emerging trend in both wired and wireless networks. For cellular, SDN promises to reduce OPEX costs, simplify the introduction of new services, and improve scalability; all major infrastructure vendors are involved. The Open Networking Foundation explains that an SDN decouples the control and data planes, centralizing network state and intelligence, while abstracting the underlying network infrastructure from applications.[191] Virtualization of network functions will be a complex, multi-year undertaking and will occur in stages, as shown in Figure 95.

---

[190] Next Generation Mobile Networks, *Further Study on Critical C-RAN Technologies, Version 1.0*, March 2015. See sections 2.2 and 2.3. Available at https://www.ngmn.org/uploads/media/NGMN_RANEV_D2_Further_Study_on_Critical_C-RAN_Technologes_v1.0.pdf.

[191] Open Networking Foundation, "Software-Defined Networking: The New Norm for Networks," http://www.opennetworking.org/sdn-resources/sdn-library/whitepapers, accessed June 20, 2014.

Figure 95: Software-Defined Networking and Cloud Architectures[192]

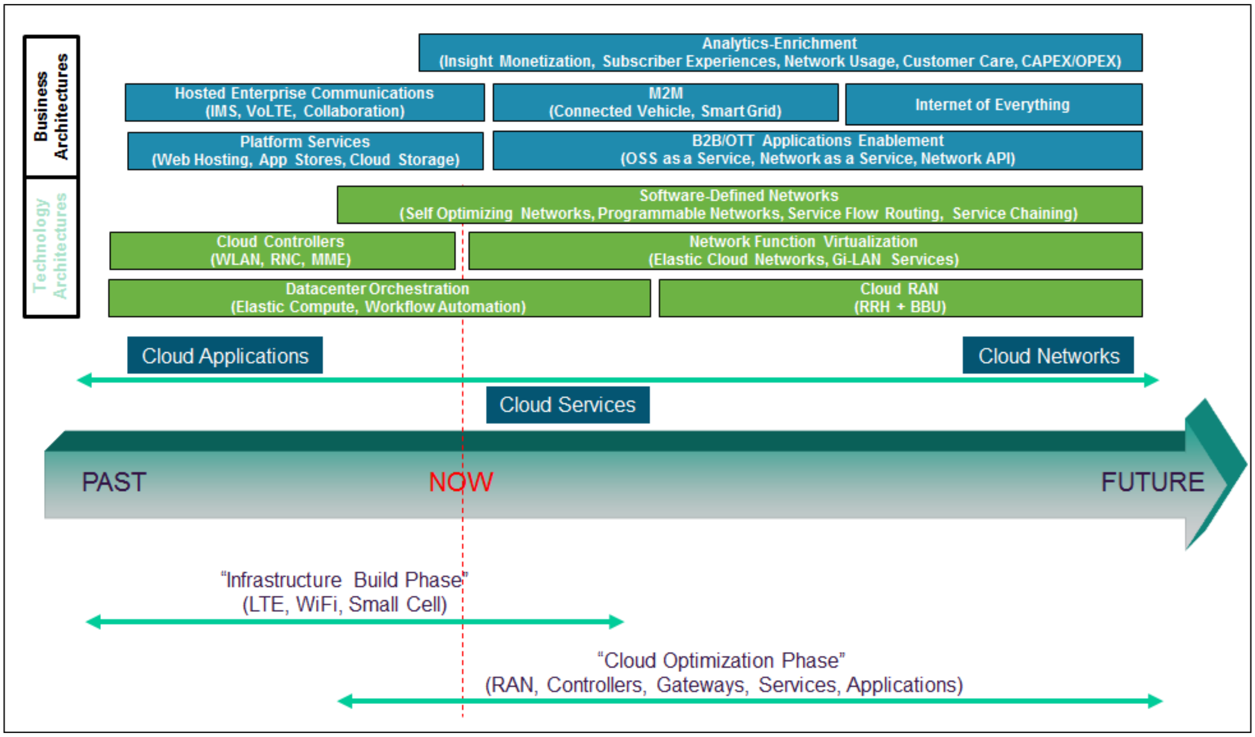
Business Architectures
Technology Architectures
Analytics-Enrichment
(Insight Monetization, Subscriber Experiences, Network Usage, Customer Care, CAPEX/OPEX)
Hosted Enterprise Communications
(IMS, VoLTE, Collaboration)
M2M
(Connected Vehicle, Smart Grid)
Internet of Everything
Platform Services
(Web Hosting, App Stores, Cloud Storage)
B2B/OTT Applications Enablement
(OSS as a Service, Network as a Service, Network API)
Software-Defined Networks
(Self Optimizing Networks, Programmable Networks, Service Flow Routing, Service Chaining)
Cloud Controllers
(WLAN, RNC, MME)
Network Function Virtualization
(Elastic Cloud Networks, Gi-LAN Services)
Datacenter Orchestration
(Elastic Compute, Workflow Automation)
Cloud RAN
(RRH + BBU)
Cloud Applications
Cloud Networks
Cloud Services
PAST
NOW
FUTURE
"Infrastructure Build Phase"
(LTE, WiFi, Small Cell)
"Cloud Optimization Phase"
(RAN, Controllers, Gateways, Services, Applications)

Unlicensed Spectrum Integration

See the earlier section in this report on unlicensed spectrum integration, which includes a discussion of LTE-U, LTE-LAA, MulteFire, LWA, LWIP, and RCLWI. This section covers integration approaches other than these.

3GPP has evolved its thinking on how best to integrate Wi-Fi with 3GPP networks. At the same time, the Wi-Fi Alliance and other groups have also addressed hotspot roaming, namely the ability to enable an account with one public Wi-Fi network provider to use the services of another provider that has a roaming arrangement with the first provider.

The multiple attempts to make Wi-Fi networks universally available have made for a confusing landscape of integration methods, which this section attempts to clarify. Most integration today is fairly loose, meaning that either a device communicates data via the cellular connection or via Wi-Fi. If via Wi-Fi, the connection is directly to the internet and bypasses the operator core network. In addition, any automatic handover to hotspots occurs only between the operator cellular network and operator-controlled hotspots. The goals moving forward are to:

- Support roaming relationships so that users can automatically access Wi-Fi hotspots operated by other entities.
- Enable automatic connections so that users do not have to enter usernames and passwords. In most cases, this will mean authentication based on SIM credentials.

---

[192] 5G Americas member contribution.

- ❑ Provide secure communications on the radio link as provided by the IEEE 802.11i standard.
- ❑ Allow policy-based mechanisms that define the rules by which devices connect to various Wi-Fi networks.
- ❑ Enable simultaneous connections to both cellular and Wi-Fi, with control over which applications use which connections.
- ❑ Support different types of Wi-Fi deployments, including third-party access points and carrier access points.

Release 6 I-WLAN

3GPP Release 6 was the first release to offer the option of integrating Wi-Fi in a feature called "Interworking WLAN" (I-WLAN), using a separate IP address for each network type.

Release 8 Dual Stack Mobile IPv6 and Proxy Mobile IPv6

3GPP Release 8 specified Wi-Fi integration with the EPC using two different approaches: host-based mobility with Dual Stack Mobile IPv6 (DSMIPv6) in the client, and network based mobility with Proxy Mobile IPv6 (PMIPv6) using an intermediary node called an "Enhanced Packet Data Gateway" (ePDG).[193] This method is intended for untrusted (non-carrier-controlled) Wi-Fi networks.

Release 11 S2a-based Mobility over GTP

Release 11, however, implements a new and advantageous approach as shown in Figure 96, one that eliminates the ePDG. Called "S2a-based Mobility over GTP" (SaMOG), a trusted WLAN Access Gateway connects to multiple 3GPP-compliant access points. Traffic can route directly to the internet or traverse the packet core. This method is intended for trusted (carrier-controlled) Wi-Fi networks.

---

[193] 3GPP, *System Architecture Evolution (SAE); Security aspects of non-3GPP accesses*. TS 33.402.

Figure 96: Release 11 SaMOG-based Wi-Fi Integration

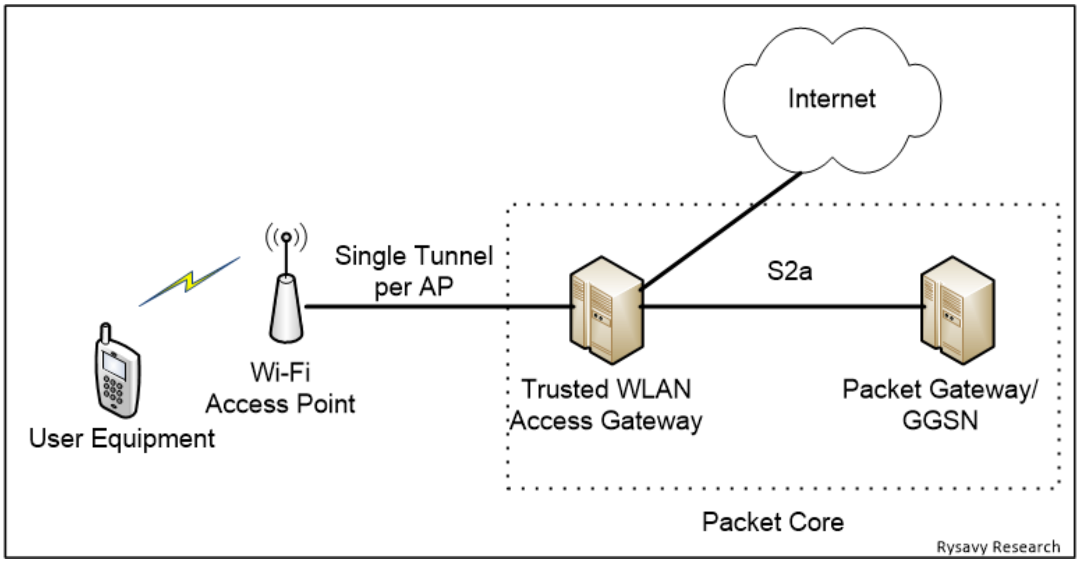

Release 12 improves SaMOG capabilities in Enhanced SaMOG (eSaMOG), in which UEs can:

- ☐ Request the connectivity type
- ☐ Indicate the Access Point Name (APN) to establish PDN connectivity
- ☐ Request to hand over an existing PDN connection
- ☐ Establish multiple PDN connections in parallel over the WLAN
- ☐ Establish a non-seamless WLAN offload connection in parallel to a Packet Data Network connection over WLAN.

Multipath TCP

A new method for potentially integrating Wi-Fi and 3GPP networks is based on work by the Internet Engineering Taskforce (IETF). Called “Multipath TCP,” the approach allows a TCP connection to occur simultaneously over two different paths. The advantages of this approach include higher speeds by aggregating links and not requiring any special provisions for link-layer handovers.

The IETF has published an experimental specification, *Request for Comments 6824: CP Extensions for Multipath Operation with Multiple Addresses*, which explains this approach. The IETF is also specifying Multipath QUIC.

ANDSF

Another relevant specification is 3GPP Access Network Discovery and Selection Function (ANDSF), which provides mechanisms by which mobile devices can know where, when, and how to connect to non-3GPP access networks, such as Wi-Fi.[194] ANDSF operates independently of SaMOG or other ways that Wi-Fi networks might be connected.

---

[194] 3GPP, *Architecture enhancements for non-3GPP accesses, Technical Specification 23.402*.

ANDSF functionality increases with successive 3GPP versions, as summarized in Table 28.

Table 28: ANDSF Policy Management Objects and 3GPP Releases[195]

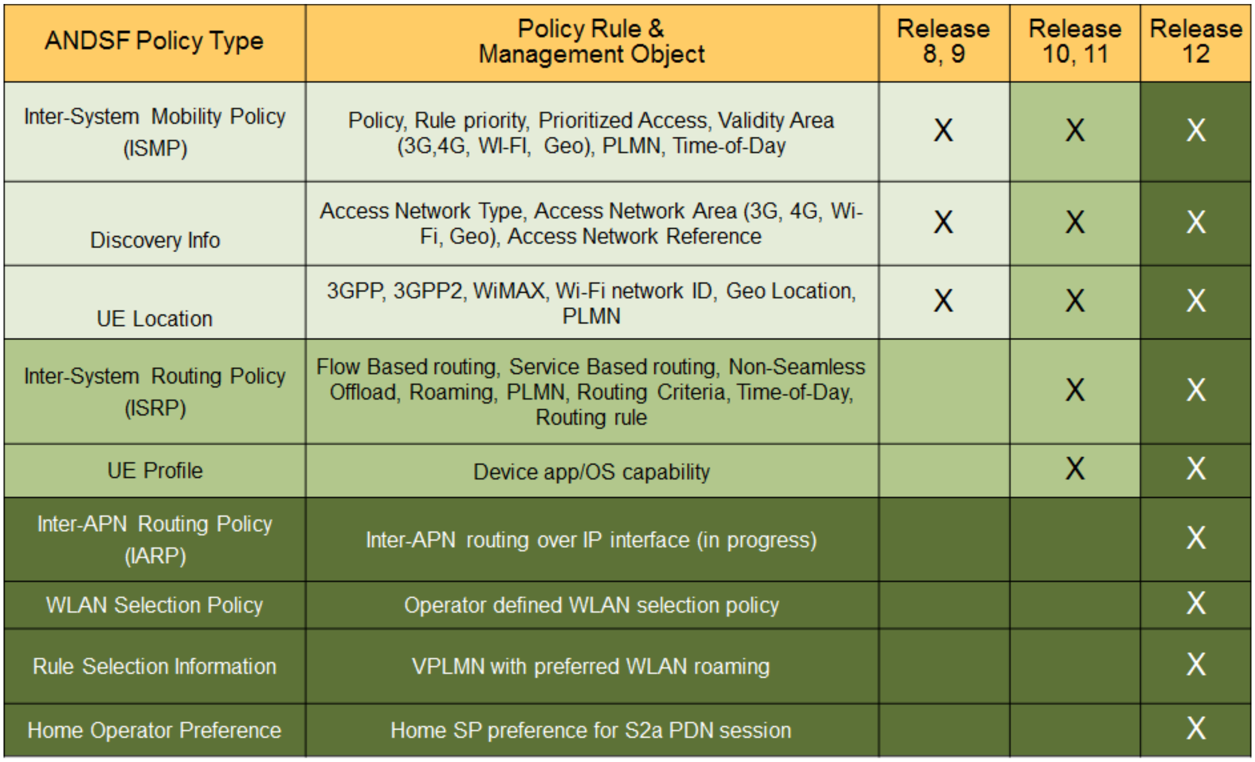

ANDSF Policy Type	Policy Rule & Management Object	Release 8, 9	Release 10, 11	Release 12
Inter-System Mobility Policy (ISMP)	Policy, Rule priority, Prioritized Access, Validity Area (3G,4G, WI-FI, Geo), PLMN, Time-of-Day	X	X	X
Discovery Info	Access Network Type, Access Network Area (3G, 4G, Wi-Fi, Geo), Access Network Reference	X	X	X
UE Location	3GPP, 3GPP2, WiMAX, Wi-Fi network ID, Geo Location, PLMN	X	X	X
Inter-System Routing Policy (ISRP)	Flow Based routing, Service Based routing, Non-Seamless Offload, Roaming, PLMN, Routing Criteria, Time-of-Day, Routing rule		X	X
UE Profile	Device app/OS capability		X	X
Inter-APN Routing Policy (IARP)	Inter-APN routing over IP interface (in progress)			X
WLAN Selection Policy	Operator defined WLAN selection policy			X
Rule Selection Information	VPLMN with preferred WLAN roaming			X
Home Operator Preference	Home SP preference for S2a PDN session			X

Bidirectional Offloading Challenges

Eventually, operators will be able to closely manage user mobile broadband and Wi-Fi connections, dynamically selecting a particular network for a user based on real-time changes in loads and application requirements. Work is occurring in Release 12 to define parameters that would control switching from LTE to Wi-Fi or from Wi-Fi to LTE.[196]

Bidirectional offloading, however, creates various challenges, as shown in Figure 97 and discussed below.

---

[195] Courtesy Smith Micro Software, 2014. http://www.smithmicro.com.

[196] 3GPP, *Study on Wireless Local Area Network (WLAN) - 3GPP radio interworking (Release 12)*, TR 37.834.

Figure 97: Bidirectional Offloading Challenges

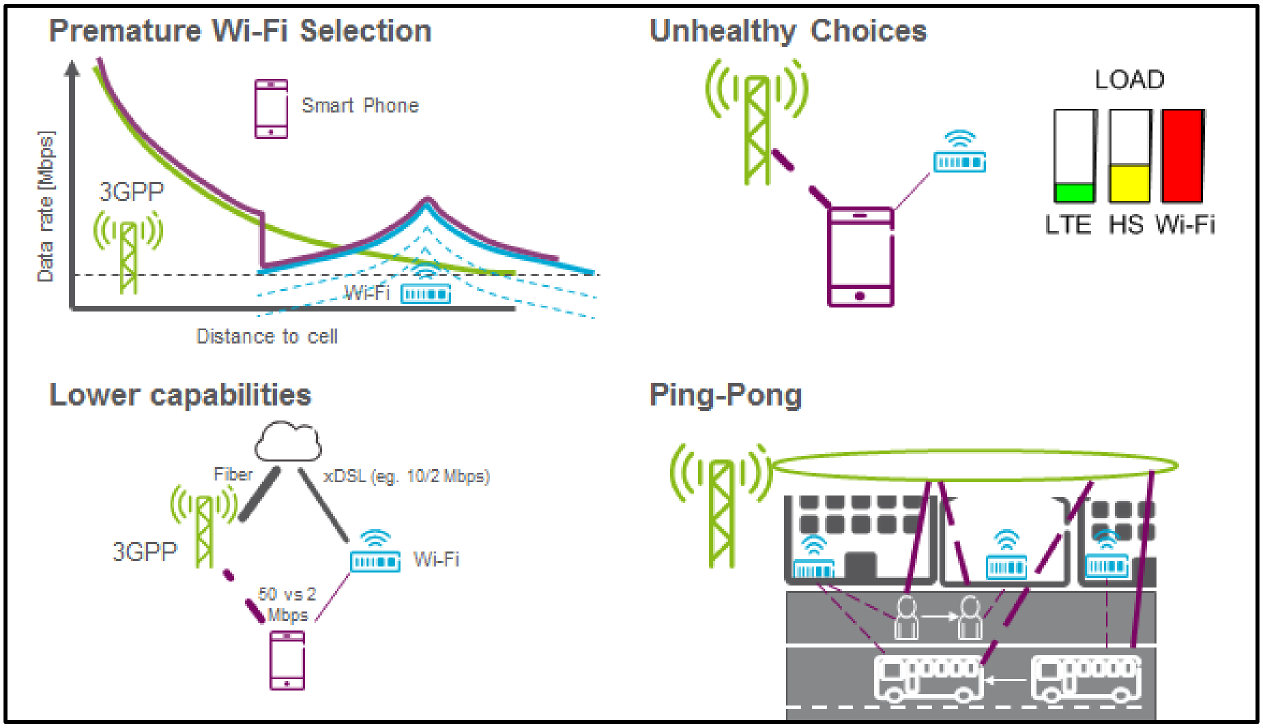
Premature Wi-Fi Selection
Smart Phone
3GPP
Data rate [Mbps]
Wi-Fi
Distance to cell
Unhealthy Choices
LOAD
LTE
HS
Wi-Fi
Lower capabilities
Fiber
xDSL (eg. 10/2 Mbps)
3GPP
Wi-Fi
50 vs 2 Mbps
Ping-Pong

- **Premature Wi-Fi Selection.** As Wi-Fi-capable devices move into Wi-Fi coverage, they can prematurely reselect to Wi-Fi without comparative evaluation of existing cellular and incoming Wi-Fi capabilities, possibly resulting in the degradation of the end user experience. Real-time throughput-based traffic steering can mitigate this effect.
- **Unhealthy choices.** In a mixed network of LTE, HSPA, and Wi-Fi, reselection can occur due to a strong Wi-Fi network signal even though the network is under heavy load. The resulting “unhealthy” choice degrades the end user experience because the performance on the cell edge of a lightly loaded cellular network may be superior to that of the heavily loaded Wi-Fi network. Real-time load-based traffic steering can be beneficial in this scenario.
- **Lower capabilities.** In some cases, selection to a Wi-Fi network may result in reduced performance even if it offers a strong signal because of other factors, such as lower-bandwidth backhaul. Evaluation of criteria beyond wireless capabilities prior to access selection can improve this circumstance.
- **Ping-Pong.** Ping-ponging between Wi-Fi and cellular, especially if both offer similar signal strengths, can also degrade the user experience. Hysteresis approaches, similar to those used in cellular inter-radio transfer, can better manage transfer between Wi-Fi and cellular accesses.

3GPP RAN2 is discussing real-time or near-real-time methods to address the challenges discussed above.

Other Integration Technologies (SIPTO, LIPA, IFOM, MAPCON)

Release 10 defines additional options for Wi-Fi integration, including Selected IP Traffic Offload (SIPTO), Local IP Access (LIPA), Multi-Access PDN Connectivity (MAPCON), and IP Flow and Seamless Offload (IFOM).

SIPTO is mostly a mechanism to offload traffic that does not need to flow through the core, such as internet-destined traffic. SIPTO can operate on a home femtocell, or it can operate in the macro network.

Local IP Access (LIPA) provides access to local networks, useful with femtocells that normally route all traffic back to the operator network. With LIPA, the UE in a home environment can access local printers, scanners, file servers, media servers, and other resources.

IFOM, as shown in Figure 98, enables simultaneous cellular and Wi-Fi connections, with different traffic flowing over the different connections. A Netflix movie could stream over Wi-Fi, while a VoIP call might flow over the cellular-data connection. IFOM requires the UE to implement Dual Stack Mobile IPv6 (DSMIPv6).

Figure 98: 3GPP IP Flow and Seamless Mobility

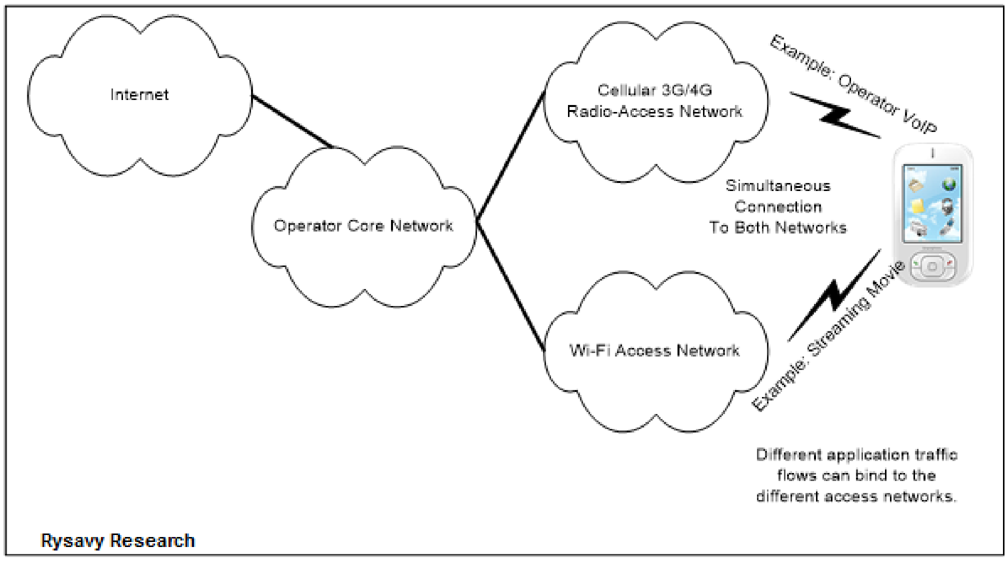
Internet
Operator Core Network
Cellular 3G/4G Radio-Access Network
Wi-Fi Access Network
Example: Operator VoIP
Simultaneous Connection To Both Networks
Example: Streaming Movie
Different application traffic flows can bind to the different access networks.
Rysavy Research

Similar to IFOM, Release 10 feature MAPCON allows multiple simultaneous PDN connections (each with a separate APN), such as Wi-Fi and 3GPP radio access. The UE uses separate IP addresses for each connection but does not need Dual Stack Mobile IPv6 (DSMIPv6).

Hotspot 2.0

Developed by the Wi-Fi Alliance, Hotspot 2.0 specifications, also called “Next Generation Hotspot,” facilitate Wi-Fi roaming. Using the IEEE 802.11u standard that allows devices to determine what services are available from an access point, Hotspot 2.0 simplifies the process by which users connect to hotspots, automatically identifying roaming partnerships and simplifying authentication and connections, as shown in Figure 99.[197] It also provides for encrypted communications over the radio link.[198]

Figure 99: Roaming Using Hotspot 2.0

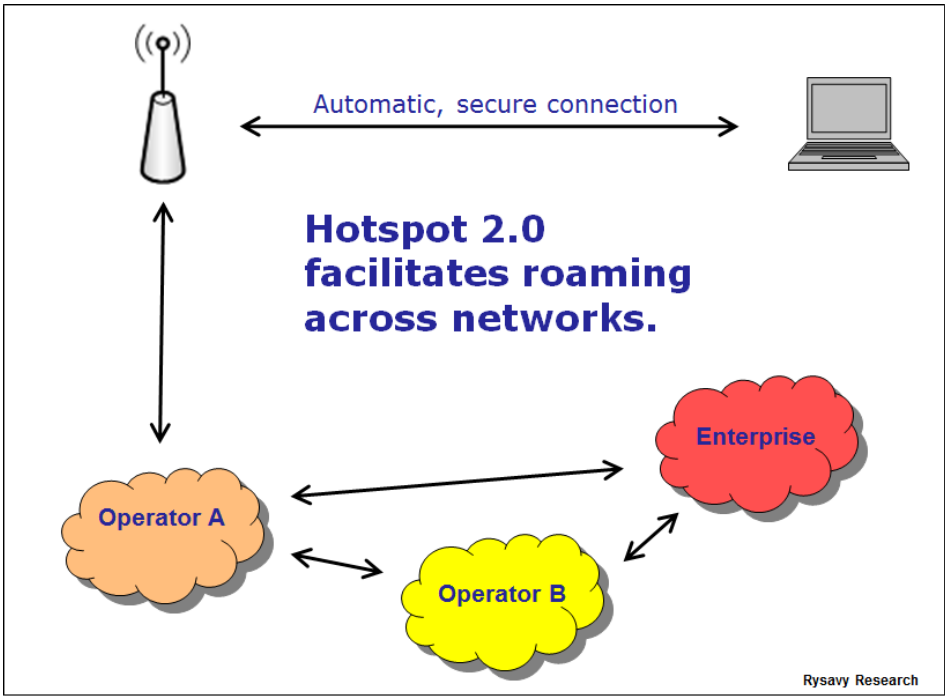
Automatic, secure connection
Hotspot 2.0 facilitates roaming across networks.
Enterprise
Operator A
Operator B
Rysavy Research

Using IEEE 802.11u, devices can determine what roaming relationships an access point supports and can then securely connect to the Wi-Fi network using one of these roaming arrangements, as shown in Figure 100. Hotspot 2.0 authentication is based on the Extended Authentication Protocol (EAP) using SIM credentials. There are plans to enhance the Hotspot 2.0 protocols in Phase 2, which will define online signup to enable non-SIM-based devices to easily and securely register for services. The Wi-Fi Alliance began a Hotspot 2.0 certification process for devices and access points in June 2012 and uses the designation “Wi-Fi Certified Passpoint” for compliant devices.

---

[197] For example, user devices can be authenticated based on their SIM credentials. Or, users can register or click through an agreement and then not need to redo that with future associations.

[198] The IEEE 802.11i standard has provided encryption for 802.11 communications for many years; however, most hotspots have not implemented this encryption, whereas Hotspot 2.0 does.

Figure 100: Hotspot 2.0 Connection Procedure

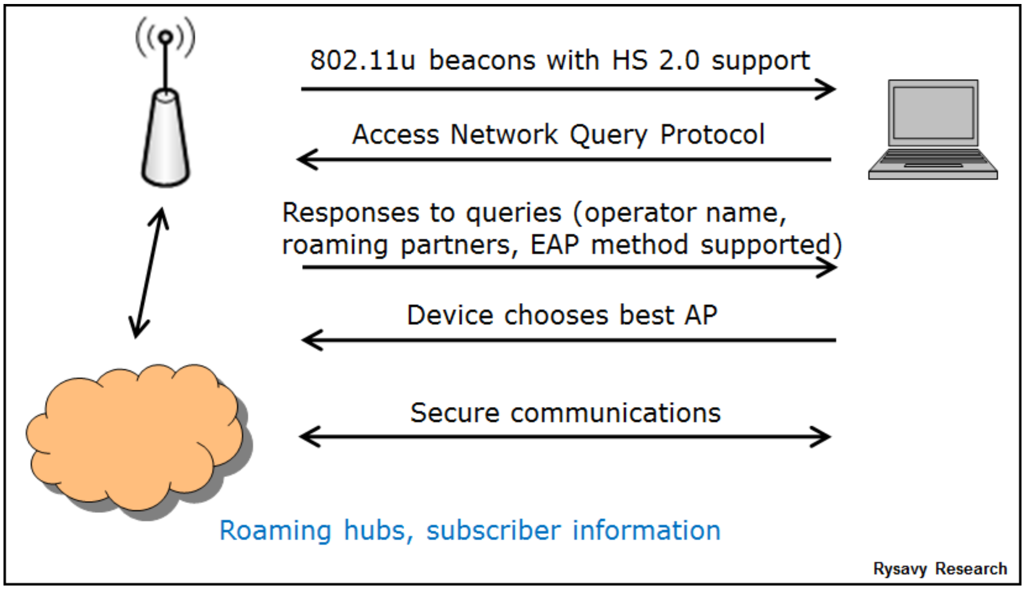
802.11u beacons with HS 2.0 support
Access Network Query Protocol
Responses to queries (operator name, roaming partners, EAP method supported)
Device chooses best AP
Secure communications
Roaming hubs, subscriber information
Rysavy Research

Release 2 of Passpoint, available in 2014, added immediate account provisioning, which facilitates a user establishing an account at the point of access. The new version also provides for policies to be downloaded from the network operator; these policies control network selection priorities when multiple networks are available.

Self-Organizing Networks (SON)

As the number of base stations increase through denser deployments and through deployment of femtocells and picocells, manual configuration and maintenance of this infrastructure becomes impractical. With SON, base stations organize and configure themselves by communicating with one another and with the core network. SONs can also self-heal in failure situations.

3GPP began standardization of self-optimization and self-organization in Releases 8 and 9, a key goal being support of multi-vendor environments. Successive releases have augmented SON capabilities.

Features being defined in SON include:

- Automatic inventory;
- Automatic software download;
- Automatic neighbor relation;
- Automatic physical Cell ID assignment;
- Mobility robustness/handover optimization;
- Random access channel optimization;
- Load-balancing optimization;
- Inter-cell interference coordination (ICIC) management;

- Enhanced inter-cell interference coordination (eICIC) management;
- Coverage and capacity optimization;
- Cell outage detection and compensation;
- Self-healing functions;
- Minimization of drive testing;
- Energy savings; and
- Coordination among various SON functions.

3GPP categorizes SON as centralized, distributed, or hybrid, which is a combination of centralized and distributed approaches.

In a centralized architecture, SON algorithms operate on a central network management system or central SON server. In contrast, in a distributed approach, the SON algorithms operate at the eNBs, which make autonomous decisions based on local measurements as well as from other nearby eNBs received via an X2 interface that interconnects eNBs.

The distributed architecture permits faster and easier deployment but is not necessarily as efficient or as consistent in operation, especially in multi-vendor infrastructure deployments.

In a hybrid approach, shown in Figure 101, SON algorithms operate both at the eNB and at a central SON server, with the server supplying values of initial parameters, for example. The eNBs may then update and refine those parameters in response to local measurements.

The hybrid approach resolves deployment scenarios that cannot be resolved by dSON, for example, cases such as:

- No X2 interface between the eNBs.
- Multi-vendor deployment with different dSON algorithms.
- Multi-technology load balancing and user steering.

Figure 101: Hybrid SON Architecture[199]

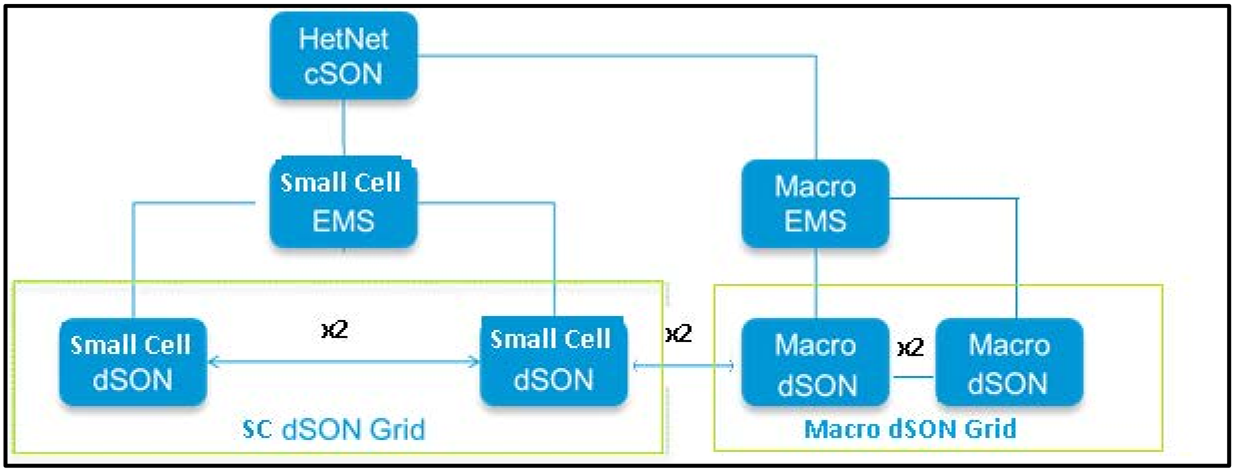

With increasing numbers of macro cells and small cells, interference opportunities increase as well. Optimizing power settings through intelligent power management algorithms is crucial for maximum efficiency with the least amount of interference, including pilot pollution. Pilot pollution can result in low data rates and ping-pong handovers due to channel fading. A hybrid SON approach is well suited for optimized power management.

IP Multimedia Subsystem (IMS)

IP Multimedia Subsystem (IMS) is a service platform for IP multimedia applications: video sharing, PoC, VoIP, streaming video, interactive gaming, and others. IMS by itself does not provide all these applications. Rather, it provides a framework of application servers, subscriber databases, and gateways to make them possible. The exact services will depend on cellular operators and the application developers that make these applications available to operators. The primary application in 2016, however, is VoLTE.

The core networking protocol used within IMS is Session Initiation Protocol (SIP), which includes the companion Session Description Protocol (SDP) used to convey configuration information such as supported voice codecs. Other protocols include Real Time Transport Protocol (RTP) and Real Time Streaming Protocol (RTSP) for transporting actual sessions. The QoS mechanisms in UMTS will be an important component of some IMS applications.

Although originally specified by 3GPP, numerous other organizations around the world are supporting IMS. These include the IETF, which specifies key protocols such as SIP, and the Open Mobile Alliance, which specifies end-to-end, service-layer applications. Other organizations supporting IMS include the GSMA, ETSI, CableLabs, 3GPP2, The Parlay Group, the ITU, ANSI, the Telecoms and Internet Converged Services and Protocols for Advanced Networks (TISPAN), and the Java Community Process (JCP).

IMS is relatively independent of the radio-access network and can, and likely will, be used by other radio-access networks or wireline networks. Other applications include picture and video sharing that occur in parallel with voice communications. Operators looking to roll out VoIP over networks will use IMS. For example, VoLTE depends on IMS infrastructure.

---

[199] 5G Americas member contribution.

3GPP initially introduced IMS in Release 5 and has enhanced it in each subsequent specification release.

As shown in Figure 102, IMS operates just outside the packet core.

Figure 102: IP Multimedia Subsystem

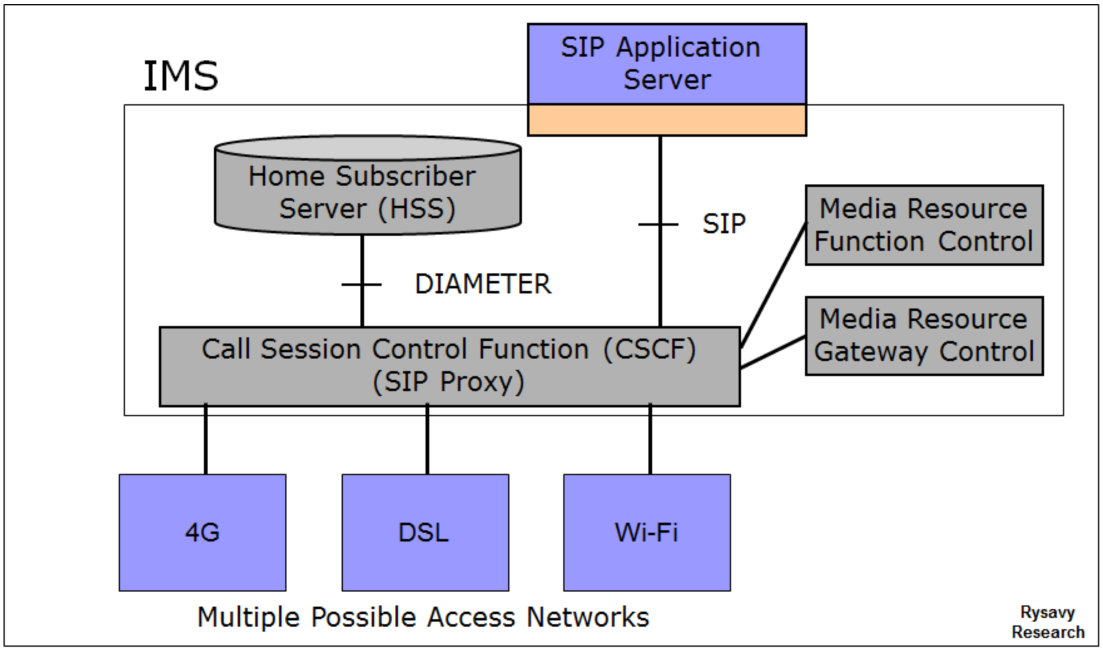
IMS
SIP Application Server
Home Subscriber Server (HSS)
SIP
DIAMETER
Media Resource Function Control
Media Resource Gateway Control
Call Session Control Function (CSCF) (SIP Proxy)
4G
DSL
Wi-Fi
Multiple Possible Access Networks
Rysavy Research

The benefits of using IMS include handling all communication in the packet domain, tighter integration with the internet, and a lower cost infrastructure based on IP building blocks for both voice and data services.

IMS applications can reside either in the operator's network or in third-party networks including those of enterprises. By managing services and applications centrally—and independently of the access network—IMS can enable network convergence. This allows operators to offer common services across 3G, Wi-Fi, and wireline networks.

Service Continuity, defined in Release 8, provided for a user's entire session to continue seamlessly as the user moves from one access network to another. Release 9 expanded this concept to allow sessions to move across different device types. For example, the user could transfer a video call in midsession from a mobile phone to a large-screen TV, assuming both have an IMS appearance in the network.

Release 8 introduced the IMS Centralized Services (ICS) feature, which allows for IMS-controlled voice features to use either packet-switched or circuit-switched access.

Given that LTE operators will integrate their 5G networks with their current LTE networks, operators are likely to keep using IMS in conjunction with LTE for their voice and other services that use IMS, even as they begin deploying 5G.

Broadcast/Multicast Services

An important capability for 3G and evolved 3G systems is broadcasting and multicasting, wherein multiple users receive the same information using the same radio resource. This creates a more efficient approach to deliver video when multiple users desire the same content simultaneously. In a broadcast, every subscriber unit in a service area receives the information, whereas in a multicast, only users with subscriptions receive the information. Service areas for both broadcast and multicast can span either the entire network or a specific geographical area. Potential applications include sporting events, select news, venue-specific (shopping mall, museum) information, and even delivery of software upgrades. Giving users the ability to store and replay select content could further expand the scope of applications.

3GPP defined highly efficient broadcast/multicast capabilities for UMTS in Release 6 with MBMS. Release 7 defined optimizations through a feature called multicast/broadcast, single-frequency network operation that involves simultaneous transmission of the exact waveform across multiple cells. This enables the receiver to constructively superpose multiple MBMS Single Frequency Network (SFN), or MBSFN, cell transmissions. The result is highly efficient, WCDMA-based broadcast transmission technology that matches the benefits of OFDMA-based broadcast approaches.

LTE also has a broadcast/multicast capability called eMBMS. OFDM is particularly well suited for efficient broadcasting, as shown in Figure 103, because the mobile system can combine the signal from multiple base stations, also an MBSFN approach, and because of the narrowband nature of OFDM. Normally, these signals would interfere with one another. The single frequency network is a cluster of cells that transmit the same content synchronously with a common carrier frequency.

Figure 103: OFDM Enables Efficient Broadcasting

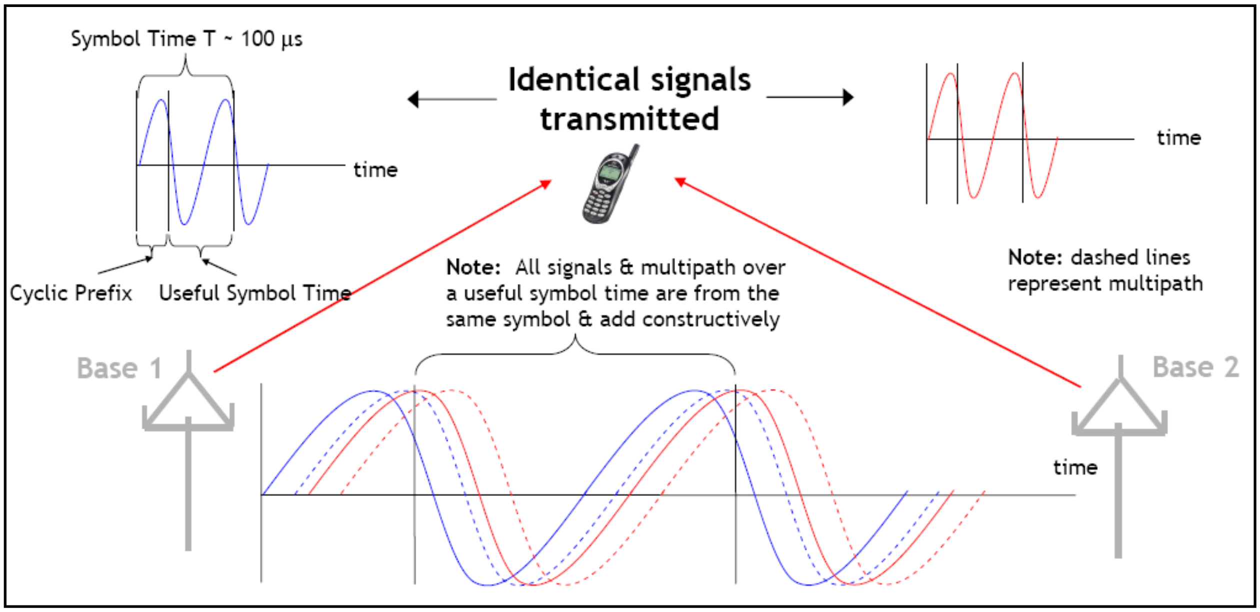
Symbol Time T ~ 100 μs
time
Cyclic Prefix
Useful Symbol Time
Identical signals transmitted
time
Note: dashed lines represent multipath
Note: All signals & multipath over a useful symbol time are from the same symbol & add constructively
Base 1
Base 2
time

Despite various broadcast technologies being available, market adoption to date has been relatively slow. Internet trends have favored unicast approaches, with users viewing videos of their selection on demand, but there is increasing interest in using eMBMS with LTE to alleviate capacity demands.

Backhaul

Connecting sites to core networks remains a challenge, whether for small cells or macro cells, especially as networks need to deliver higher bandwidth. Fiber is the gold standard, but it is not available everywhere and can be expensive, so operators use a combination of wired and wireless links.

Today's backhaul requirements for LTE can range from 1 to 10 Gbps. By 2020, backhaul requirements could exceed 10 Gbps.[200]

Table 29 and Table 30 summarize the methods and capabilities of the various available approaches.

Table 29: Wired Backhaul Methods and Capabilities[201]

Technology	Distance	Throughput Speed
Direct Fiber	80 km	Hundreds of Mbps to Gbps
Bonded VDSL2	To 5,000 feet	75 Mbps down, 12 Mbps up
FTTX	Most urban areas	Up to 2.5 Gbps down, 1.5 Gbps up
DOCSIS	Most urban areas	Up to 285 Mbps down, 105 Mbps up

Table 30: Wireless Backhaul Methods and Capabilities[202]

Technology	Distance	Line-of-Sight	Throughput Speed
5G Integrated Access and Backhaul	1 km	Yes	1 to 10 Gbps
Millimeter Wave (60 GHz)	1 km	Yes	1 Gbps
Millimeter Wave (70-80 GHz)	3 km (with speed tradeoff)	Yes	10 Gbps
Microwave (6-60 GHz)	Varies by frequency: 2-4 km typical at 30-42 GHz	Yes	1 Gbps+

---

[200] Arthur D. Little, *Creating a Gigabit Society – The Rule of 5G; A report by Arthur D. Little for Vodafone Group,* 2017. See Figure 6.

[201] Small Cell Forum, "Backhaul Technologies for Small Cells," February 2013.

[202] Ibid.

Technology	Distance	Line-of-Sight	Throughput Speed
Licensed sub 6 GHz	1.5 to 10 km	No	170 Mbps (20 MHz TDD), 400 Mbps+ with new technology
Unlicensed sub-6 GHz	Up to 250 meters	No	450 Mbps (IEEE 802.11n 3X3 MIMO)
TV White Space (802.11af-based)	1 to 5 km max throughput, 10 km+ possible	Depends on deployment model	80 Mbps in 6 MHz TDD with 4X4 MIMO
Satellite	Available everywhere	Yes	Up to 50 Mbps downlink, 15 Mbps uplink

UMTS-HSPA

UMTS technology is mature and benefits from research and development that began in the early 1990s. It has been thoroughly trialed, tested, and commercially deployed. UMTS employs a wideband CDMA radio-access technology. The primary benefits of UMTS include high spectral efficiency for voice and data, simultaneous voice and data capability, high user densities that can be supported with low infrastructure costs, and support for high-bandwidth data applications. Operators can also use their entire available spectrum for both voice and high-speed data services.

Additionally, operators can use a common core network, called the UMTS multi-radio network as shown in Figure 104, which supports multiple radio-access networks including GSM, EDGE, WCDMA, HSPA, and evolutions of these technologies.

Figure 104: UMTS Multi-radio Network

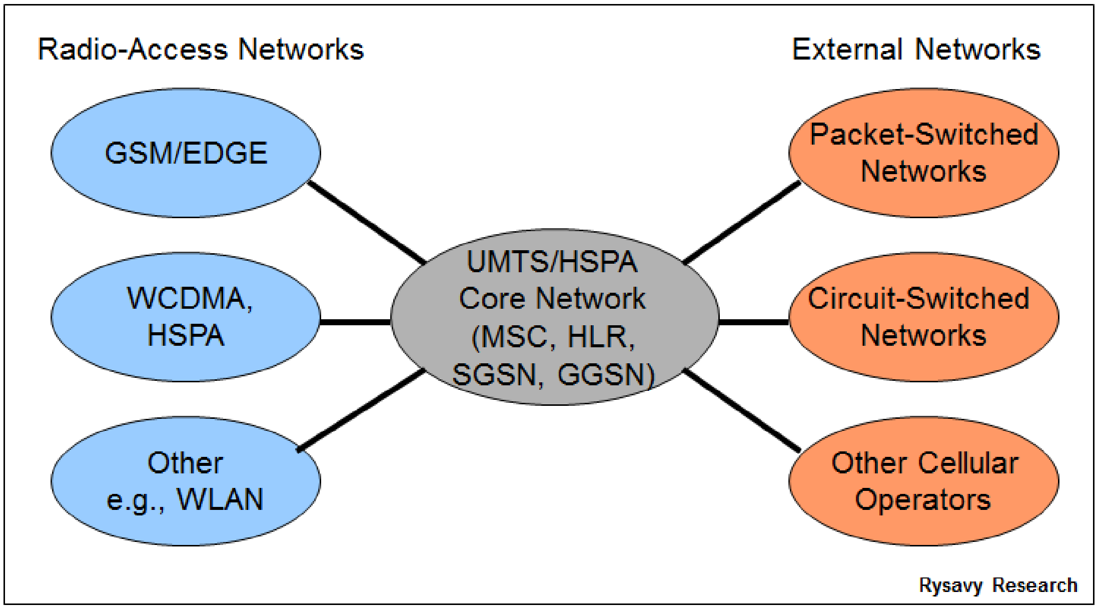
Radio-Access Networks
External Networks
GSM/EDGE
WCDMA, HSPA
Other e.g., WLAN
UMTS/HSPA Core Network (MSC, HLR, SGSN, GGSN)
Packet-Switched Networks
Circuit-Switched Networks
Other Cellular Operators
Rysavy Research

HSPA refers to networks that support both HSDPA and HSUPA. All new deployments today are HSPA, and many operators have upgraded their HSDPA networks to HSPA. For example, in 2008, AT&T upgraded most of its network to HSPA. By the end of 2008, HSPA was deployed throughout the Americas.

The UMTS radio-access network consists of base stations referred to as Node B (corresponding to GSM base transceiver systems) that connect to RNCs (corresponding to GSM base station controllers [BSCs]). The RNCs connect to the core network as do the BSCs. When both GSM and WCDMA access networks are available, the network can hand users over between these networks. This is important for managing capacity, as well as in areas in which the operator has continuous GSM coverage, but has only deployed WCDMA in some locations.

Whereas GSM can effectively operate like a spread-spectrum system[203], based on time division in combination with frequency hopping, WCDMA is a direct-sequence, spread-spectrum system. WCDMA is spectrally more efficient than GSM, but it is the wideband nature of WCDMA that provides its greatest advantage—the ability to translate the available spectrum into high data rates. This wideband technology approach results in the flexibility to manage multiple traffic types including voice, narrowband data, and wideband data.

HSDPA

HSDPA, specified in 3GPP Release 5, saw the introduction of high-performance, packet data service that delivers peak theoretical rates of 14 Mbps. Peak user-achievable throughput rates in initial deployments are well over 1 Mbps and as high as 4 Mbps in some networks. The same radio carrier can simultaneously service UMTS voice and data users, as well as HSDPA data users.

HSDPA achieves its high speeds through techniques similar to those that push EDGE performance past GPRS including higher order modulation, variable coding, and soft combining, as well as through the addition of fast scheduling and other techniques.

HSDPA achieves its performance gains from the following radio features:

- ☐ High-speed channels shared in both code and time domains
- ☐ Short TTI
- ☐ Fast scheduling and user diversity
- ☐ Higher order modulation
- ☐ Fast link adaptation
- ☐ Fast HARQ

These features function as follows:

High-Speed Shared Channels and Short Transmission Time Interval: First, HSDPA uses high-speed data channels called “High Speed Physical Downlink Shared Channels” (HS-PDSCH). Up to 15 of these channels can operate in the 5 MHz WCDMA radio channel. Each uses a fixed spreading factor of 16. User transmissions are assigned to one or more of these channels for a short TTI of 2 msec. The network can then readjust how users are assigned to different HS-PDSCH every 2 msec. Resources are thus assigned in both time

---

[203] Spread spectrum systems can either be direct sequence or frequency hopping.

(the TTI interval) and code domains (the HS-PDSCH channels). Figure 105 illustrates different users obtaining different radio resources.

Figure 105: High Speed–Downlink Shared Channels (Example)

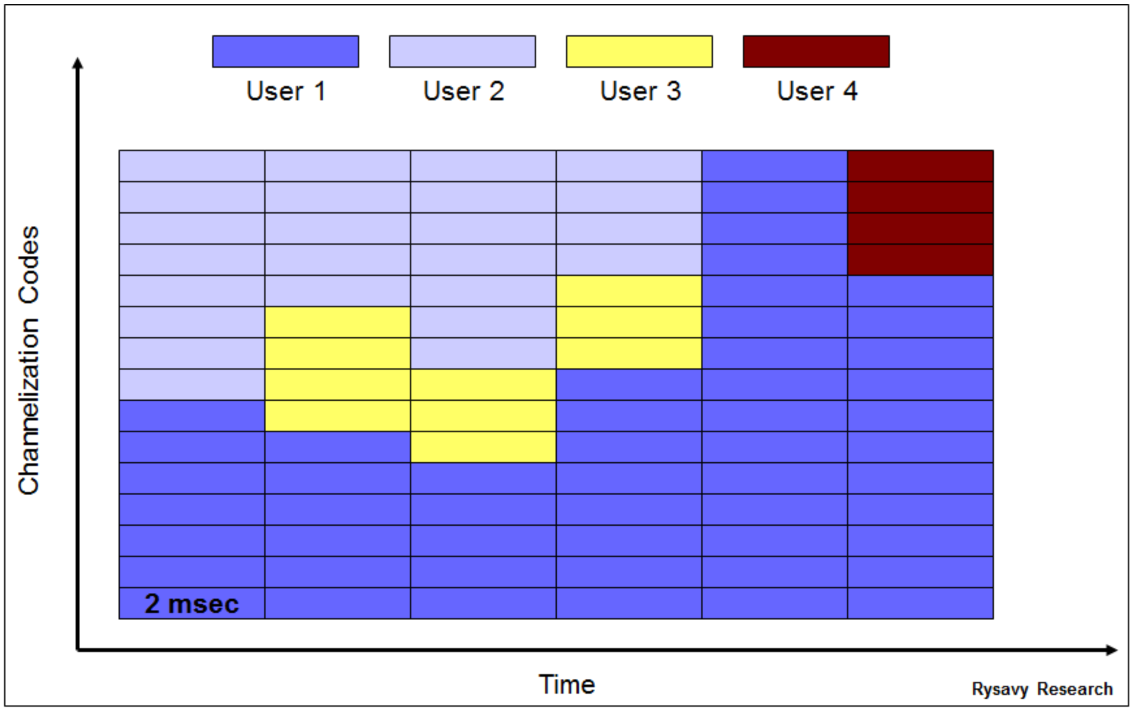

Fast Scheduling and User Diversity: Fast scheduling exploits the short TTI by assigning users channels that have the best instantaneous channel conditions, rather than in a round-robin fashion. Because channel conditions vary somewhat randomly across users, most users can be serviced with optimum radio conditions and thereby obtain optimum data throughput. Figure 106 shows how a scheduler might choose between two users based on their varying radio conditions to emphasize the user with better instantaneous signal quality. With about 30 users active in a sector, the network achieves significant user diversity and much higher spectral efficiency. The system also ensures that each user receives a minimum level of throughput, an approach called proportional fair scheduling.

Figure 106: User Diversity

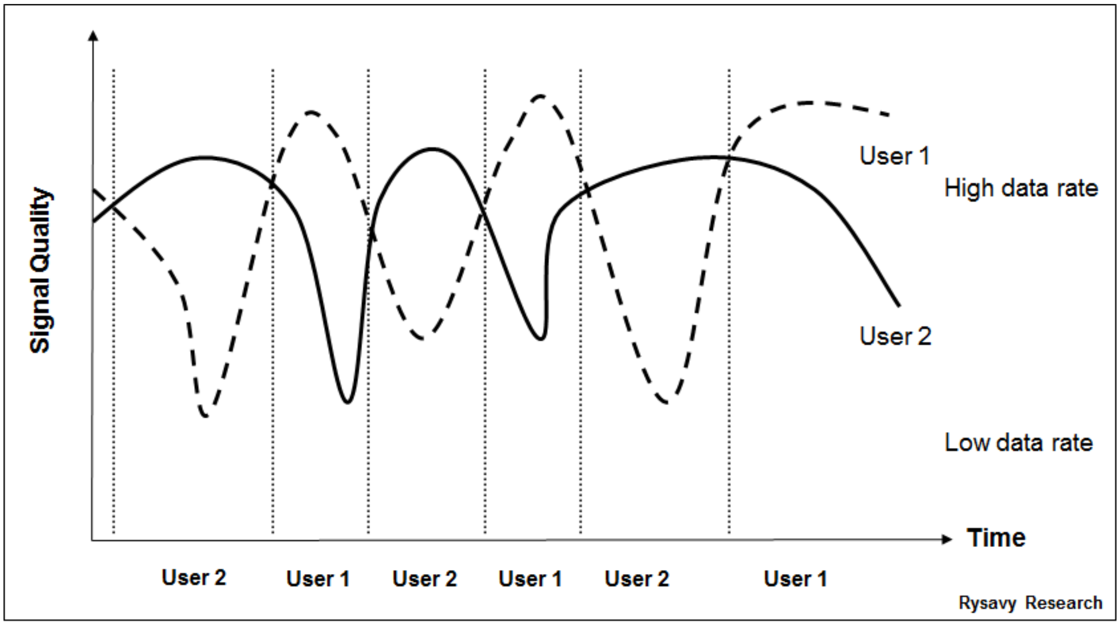

Higher Order Modulation: HSDPA uses both the modulation used in WCDMA—namely QPSK—and, under good radio conditions, an advanced modulation scheme—16 QAM. 16 QAM transmits 4 bits of data in each radio symbol compared to 2 bits with QPSK. Data throughput is increased with 16 QAM, while QPSK is available for adverse radio conditions. HSPA Evolution adds 64 QAM modulation to further increase throughput rates. 64 QAM became available in Release 7, and the combination of MIMO and 64 QAM became available in Release 8.

Fast Link Adaptation: Depending on the condition of the radio channel, different levels of forward-error correction (channel coding) can also be employed. For example, a three-quarter coding rate means that three quarters of the bits transmitted are user bits, and one quarter are error-correcting bits. Fast link adaptation refers to the process of selecting and quickly updating the optimum modulation and coding rate and occurs in coordination with fast scheduling.

Fast Hybrid Automatic Repeat Request: Another HSDPA technique is Fast Hybrid Automatic Repeat Request (Fast Hybrid ARQ). “Fast” refers to the medium-access control mechanisms implemented in Node B (along with scheduling and link adaptation), as opposed to the BSC in GPRS/EDGE, and “hybrid” refers to a process of combining repeated data transmissions with prior transmissions to increase the likelihood of successful decoding. Managing and responding to real-time radio variations at the base station, as opposed to an internal network node, reduces delays and further improves overall data throughput.

Using the approaches just described, HSDPA maximizes data throughputs and capacity and minimizes delays. For users, this translates to better network performance under loaded conditions, faster application performance, and a greater range of applications that function well.

Field results validate the theoretical throughput results. With initial 1.8 Mbps peak rate devices, vendors measured consistent throughput rates in actual deployments of more than 1 Mbps. These rates rose to more than 2 Mbps for 3.6 Mbps devices and then close to 4 Mbps for 7.2 Mbps devices.

In 2008, typical devices supporting peak data rates of 3.6 Mbps or 7.2 Mbps became available. Many operator networks support 7.2 Mbps peak operation, and some even support the maximum rate of 14.4 Mbps.

HSUPA

Whereas HSDPA optimizes downlink performance, HSUPA—which uses the Enhanced Dedicated Channel (E-DCH)—constitutes a set of improvements that optimizes uplink performance. Networks and devices supporting HSUPA became available in 2007. These improvements include higher throughputs, reduced latency, and increased spectral efficiency. HSUPA was standardized in Release 6. It results in an approximately 85% increase in overall cell throughput on the uplink and more than a 50% gain in user throughput. HSUPA also reduces packet delays, a significant benefit resulting in much improved application performance on HSPA networks

Although the primary downlink traffic channel supporting HSDPA serves as a shared channel designed for the support of services delivered through the packet-switched domain, the primary uplink traffic channel defined for HSUPA is a dedicated channel that could be used for services delivered through either the circuit-switched or the packet-switched domains. Nevertheless, by extension and for simplicity, the WCDMA-enhanced uplink capabilities are often identified in the literature as HSUPA.

HSUPA achieves its performance gains through the following approaches:

- An enhanced dedicated physical channel.
- A short TTI, as low as 2 msec, which allows faster responses to changing radio conditions and error conditions.
- Fast Node B-based scheduling, which allows the base station to efficiently allocate radio resources.
- Fast Hybrid ARQ, which improves the efficiency of error processing.

The combination of TTI, fast scheduling, and Fast Hybrid ARQ also serves to reduce latency. HSUPA can operate with or without HSDPA in the downlink, although use the two approaches together. The improved uplink mechanisms also translate to better coverage and, for rural deployments, larger cell sizes.

HSUPA can achieve different throughput rates based on various parameters including the number of codes used, the spreading factor of the codes, the TTI value, and the transport block size in bytes.

Initial devices enabled peak user rates of close to 2 Mbps as measured in actual network deployments, while current devices have throughputs of more than 5 Mbps. Future devices could have network rates as high as 69 Mbps, as discussed further below.

Beyond throughput enhancements, HSUPA also significantly reduces latency.

Evolution of HSPA (HSPA+)

The goal in evolving HSPA is to exploit available radio technologies—largely enabled by increases in digital signal processing power—to maximize CDMA-based radio performance. This evolution has significantly advanced HSPA and extends the life of sizeable operator infrastructure investments.

Wireless and networking technologists have defined a series of enhancements for HSPA, beginning in Release 7 and now continuing through Release 14. These include advanced receivers, multi-carrier operation, MIMO, Continuous Packet Connectivity, Higher-Order Modulation, One-Tunnel Architecture, HetNet support, and advanced voice capabilities both in circuit- and packet-switched domains.

Taking advantage of these various radio technologies, 3GPP has standardized a number of features, beginning in Release 7 including higher order modulation and MIMO. Collectively, these capabilities are referred to as HSPA+. Release 8 through Release 12 include further enhancements.

The goals of HSPA+ were to:

- Exploit the full potential of a CDMA approach.
- Provide smooth interworking between HSPA+ and LTE, thereby facilitating the operation of both technologies. As such, operators may choose to leverage the EPC planned for LTE.
- Allow operation in a packet-only mode for both voice and data.
- Be backward-compatible with previous systems while incurring no performance degradation with either earlier or newer devices.
- Facilitate migration from current HSPA infrastructure to HSPA+ infrastructure.

HSPA improvements have continued through successive 3GPP releases, including Release 14, which has downlink interference mitigation. Release 15 has work items for quality of experience, multi-carrier enhancements, and various protocol enhancements.

The following sections discuss specific enhancements that have already been implemented in HSPA.

Advanced Receivers

3GPP has specified a number advanced-received designs including: Type 1, which uses mobile-receive diversity; Type 2, which uses channel equalization; and Type 3, which includes a combination of receive diversity and channel equalization. Type 3i devices, which became available in 2012, employ interference cancellation. Note that the different types of receivers are release-independent. For example, Type 3i receivers will work and provide a capacity gain in an earlier Release 5 network.

The first approach is mobile-receive diversity. This technique relies on the optimal combination of received signals from separate receiving antennas. The antenna spacing yields signals that have somewhat independent fading characteristics. Hence, the combined signal can be more effectively decoded, which almost doubles downlink capacity when done in combination with channel equalization. Receive diversity is effective even with smaller devices such as like PC Card modems and smartphones.

Current receiver architectures based on rake receivers are effective for speeds up to a few megabits per second. But at higher speeds, the combination of reduced symbol period and multipath interference results in Intersymbol Interference and diminishes rake receiver performance. This problem can be solved by advanced-receiver architectures with channel equalizers that yield additional capacity gains over HSDPA with receive diversity. Alternate advanced-receiver approaches include interference cancellation and generalized rake receivers (G-Rake). Different vendors are emphasizing different approaches. The performance requirements for advanced-receiver architectures, however, were specified in 3GPP Release 6. The combination of mobile-receive diversity and channel equalization (Type 3) is especially attractive, because it results in a large capacity gain independent of the radio channel.

What makes such enhancements attractive is that the networks do not require any changes other than increased capacity within the infrastructure to support the higher bandwidth. Moreover, the network can support a combination of devices including both earlier devices that do not include these enhancements and later devices that do. Device vendors can selectively apply these enhancements to their higher-end devices.

MIMO

Another standardized capability is MIMO, a technique that employs multiple transmit antennas and multiple receive antennas, often in combination with multiple radios and multiple parallel data streams. The most common use of the term “MIMO” applies to spatial multiplexing. The transmitter sends different data streams over each antenna. Whereas multipath is an impediment for other radio systems, MIMO—as illustrated in Figure 107—actually exploits multipath, relying on signals to travel across different uncorrelated communications paths. The multiple data paths effectively operate in parallel and, with appropriate decoding, in a multiplicative gain in throughput.

Figure 107: MIMO Using Multiple Paths to Boost Throughput and Capacity

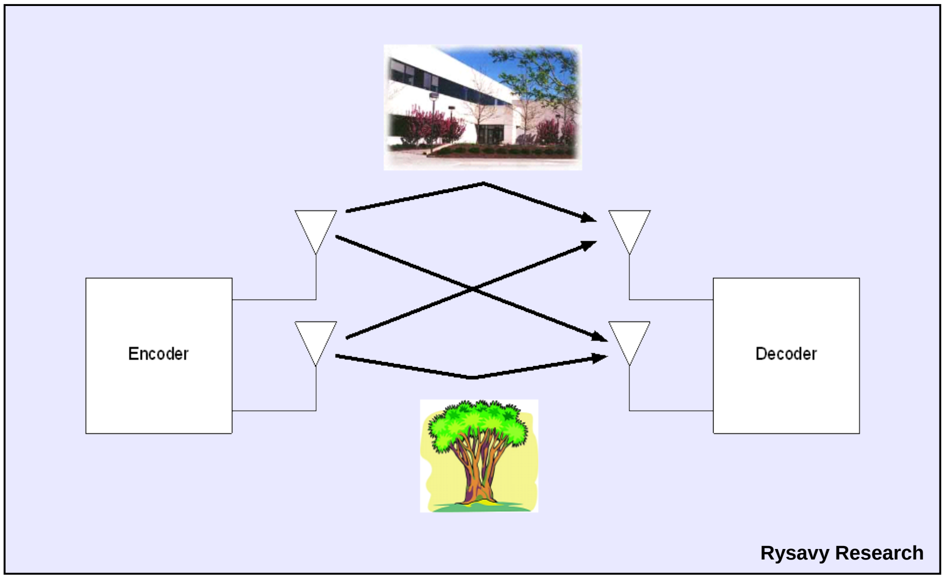

Tests of MIMO have proven effective in WLANs operating in relative isolation where interference is not a dominant factor. Spatial multiplexing MIMO can also benefit HSPA “hotspots” serving local areas including airports, campuses, and malls. In a fully loaded network with interference from adjacent cells, however, overall capacity gains will be more modest—in the range of 20% to 33% over mobile-receive diversity. Relative to a 1x1 antenna system, however, 2X2 MIMO can deliver cell throughput gains of about 80%. 3GPP has standardized spatial multiplexing MIMO in Release 7 using Double Transmit Adaptive Array (D-TxAA).

Release 9 provides for a means to leverage MIMO antennas at the base station when transmitting to user equipment that does not support MIMO. The two transmit antennas in the base station can transmit a single stream using beam forming. This is called “single stream MIMO” or “MIMO with single-stream restriction” and results in higher throughput rates because of the improved signal received by the user equipment.

3GPP designed uplink dual-antenna beamforming and 2X2 MIMO for HSPA+ in Release 11.

Continuous Packet Connectivity

Continuous Packet Connectivity (CPC) specified in Release 7 reduces the uplink interference created by the dedicated physical control channels of packet data users when those channels have no user data to transmit, which increases the number of simultaneously connected HSUPA users. CPC allows both discontinuous uplink transmission and discontinuous downlink reception, wherein the modem can turn off its receiver after a certain period of HSDPA inactivity. CPC is especially beneficial to VoIP on the uplink because the radio can turn off between VoIP packets, as shown in Figure 108.

Figure 108: Continuous Packet Connectivity

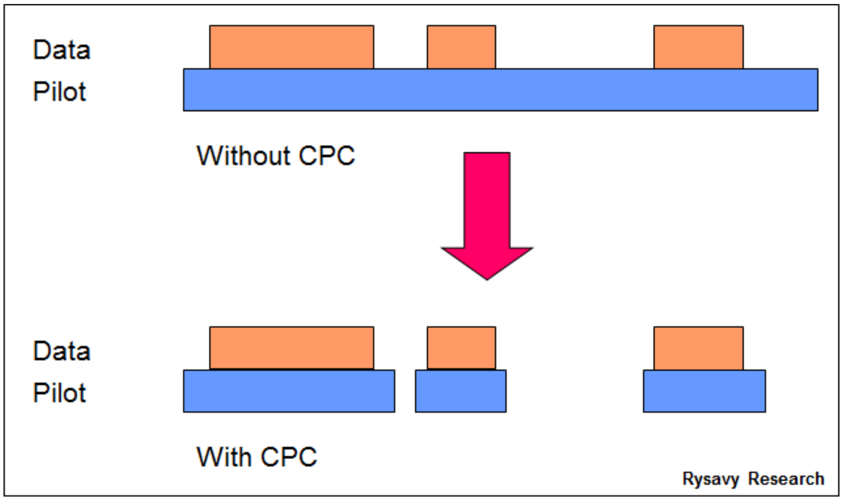

Higher Order Modulation

Another way of increasing performance is with higher order modulation. HSPA uses 16 QAM on the downlink and QPSK on the uplink, but HSPA+ adds 64 QAM to the downlink and 16 QAM to the uplink. 3GPP has also introduced 64 QAM to the uplink for HSPA+ in Release 11. Higher order modulation requires a better SNR, achieved through receive diversity and equalization.

Multi-Carrier HSPA

3GPP defined dual-carrier HSPA operation in Release 8, which coordinates the operation of HSPA on two adjacent 5 MHz carriers so that data transmissions can achieve higher throughput rates, as shown in Figure 109. The work item assumed two adjacent carriers, downlink operation and no MIMO. This configuration achieves a doubling of the 21 Mbps maximum rate available on each channel to 42 Mbps.

Figure 109: Dual-Carrier Operation with One Uplink Carrier[204]

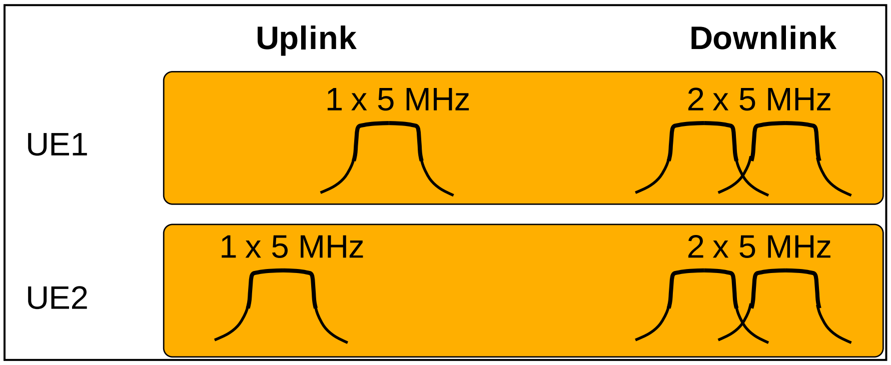

Benefits include:

- An increase in spectral efficiency of about 15%, comparable to what can be obtained with 2X2 MIMO.
- Significantly higher peak throughputs available to users, especially in lightly-loaded networks.
- Same maximum-throughput rate of 42 Mbps as using MIMO, but with a less expensive infrastructure upgrade.

Scheduling packets across two carriers is a more efficient use of resources, resulting in what is called "trunking gain." Multi-user diversity also improves from an increased number of users across the two channels.

Release 9 also supports dual-carrier operation in the uplink. Release 10 specifies the use of up to four channels, resulting in peak downlink data rates of 168 Mbps. Release 11 supports eight radio channels on the downlink, resulting in a further doubling of theoretical throughput to 336 Mbps. On the uplink, devices can transmit using two antennas for either rank 1 (single stream beamforming) or rank 2 (dual-stream MIMO) transmission modes. Rank 1 beamforming helps with coverage (approximately 40%), while rank 2 MIMO helps with throughput speeds (approximately 20% median and 80% at cell edge). In addition, 64 QAM will be possible on the uplink, enabling uplink speeds to 69 Mbps in dual-carrier operation.

Downlink Multiflow Transmission

Release 11 specifies means by which two cells can transmit to the mobile station at the same time. The two cells transmit independent data, in effect a spatial multiplexing approach, improving both peak and average data.

---

[204] Harri Holma and Antti Toskala, *LTE for UMTS, OFDMA and SC-FDMA Based Radio Access*, Wiley, April 2009.

Multiflow transmission with HSPA+ also enhances HetNet operation in which picocell coverage can be expanded within a macrocell coverage area, as shown in Figure 110.

Figure 110: HSPA+ HetNet Using Multipoint Transmission[205]

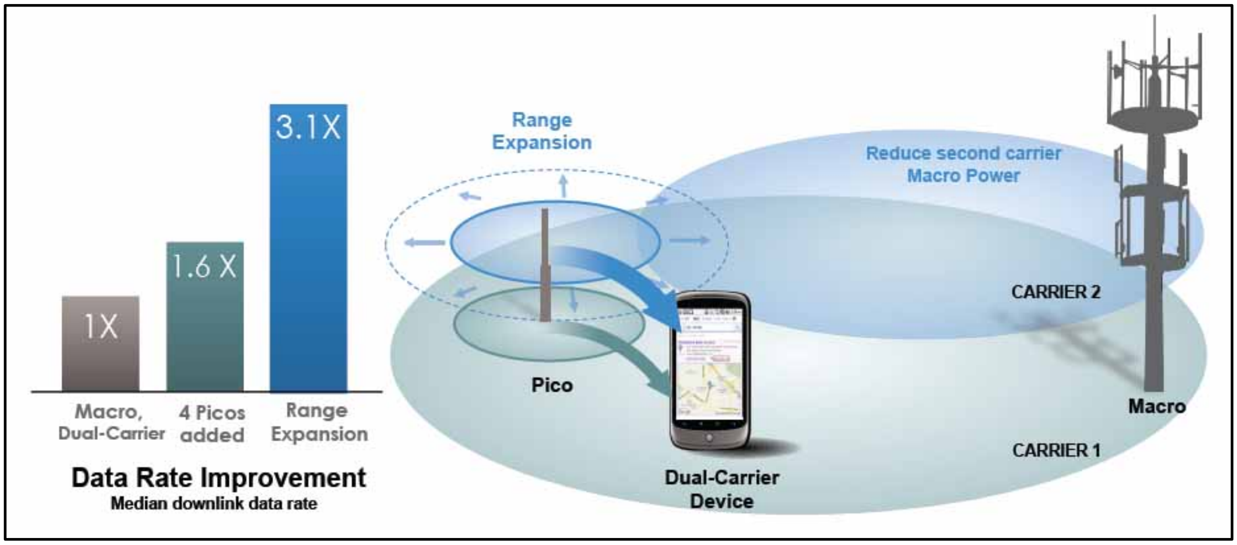
3.1X
1.6 X
1X
Macro, Dual-Carrier
4 Picos added
Range Expansion
Data Rate Improvement
Median downlink data rate
Range Expansion
Reduce second carrier Macro Power
CARRIER 2
Pico
Macro
CARRIER 1
Dual-Carrier Device

Multiflow enhances HSPA+ network operation using the following approaches:

- **Single Frequency Dual Cell.** The UE communicates with two different cells using the same frequency, improving cell-edge performance and providing network load balancing.
- **Dual Frequency Three Cell.** The UE communicates with two different cells using the same frequency. In addition, it communicates with one other cell on a different frequency.
- **Dual Frequency Four Cells.** The UE communicates using two instances of Single Frequency Dual Cell operation as described above.

In Release 12, 3GPP is considering the following enhancement to Multiflow operation, which is primarily targeted towards HetNet operation:

- **Dual Frequency Dual Carrier.** The UE aggregates cells on two different frequencies from two different sites.

HSPA+ Throughput Rates

Table 31 summarizes the capabilities of HSPA and HSPA+ based on the various methods discussed above.

---

[205] Qualcomm, "HSPA+ Advanced: Taking HSPA+ to the Next Level," February 2012, http://www.qualcomm.com/media/documents/hspa-advanced-taking-hspa-next-level-whitepaper, accessed June 20, 2014.

Table 31: HSPA Throughput Evolution

Technology	Downlink (Mbps) Peak Data Rate	Uplink (Mbps) Peak Data Rate
HSPA as defined in Release 6	14.4	5.76
Release 7 HSPA+ DL 64 QAM, UL 16 QAM, 5+5 MHz	21.1	11.5
Release 7 HSPA+ 2X2 MIMO, DL 16 QAM, UL 16 QAM, 5+5 MHz	28.0	11.5
Release 8 HSPA+ 2X2 MIMO DL 64 QAM, UL 16 QAM, 5+5 MHz	42.2	11.5
Release 8 HSPA+ (no MIMO) Dual Carrier, 10+5 MHz	42.2	11.5
Release 9 HSPA+ 2X2 MIMO, Dual Carrier DL and UL, 10+10 MHz	84.0	23.0
Release 10 HSPA+ 2X2 MIMO, Quad Carrier[206] DL, Dual Carrier UL, 20+10 MHz	168.0	23.0
Release 11 HSPA+ 2X2 MIMO DL and UL, 8 Carrier DL, Dual Carrier UL, 40+10 MHz	336.0	69.0

Release 13 enables aggregation of two UL carriers across bands.

Figure 111 shows the cumulative distribution function of throughput values in a commercially deployed Release 8 HSPA+ network in an indoor coverage scenario. The figure shows significant performance gains from higher-order modulation and MIMO.

---

[206] No operators have announced plans to deploy HSPA in a quad (or greater) carrier configuration. Three carrier configurations, however, have been deployed.

Figure 111: HSPA+ Performance Measurements Commercial Network (5+5 MHz)[207]

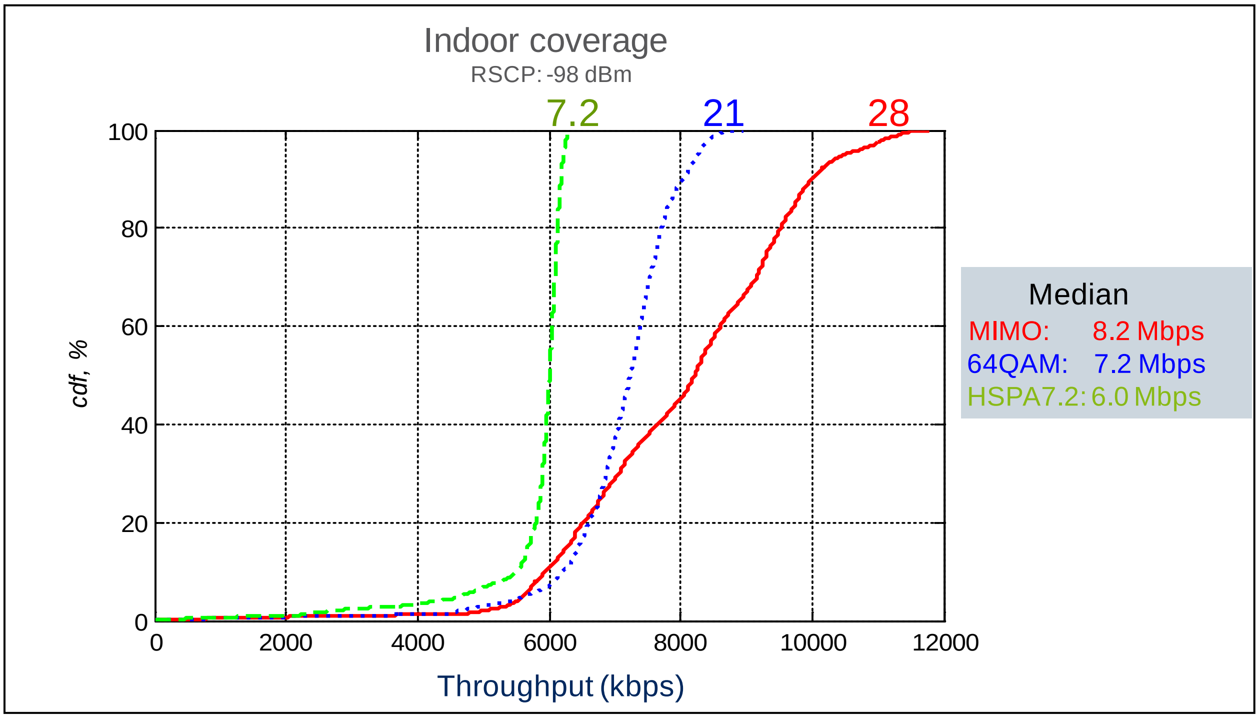

The figure shows a reasonably typical indoor scenario in a macro-cell deployment. Under better radio conditions, HSPA+ will achieve higher performance results.

Figure 112 shows the benefit of dual-carrier operation (no MIMO employed), which essentially doubles throughputs over single carrier operation.

---

[207] 5G Americas member company contribution.

Figure 112: Dual-Carrier HSPA+ Throughputs[208]

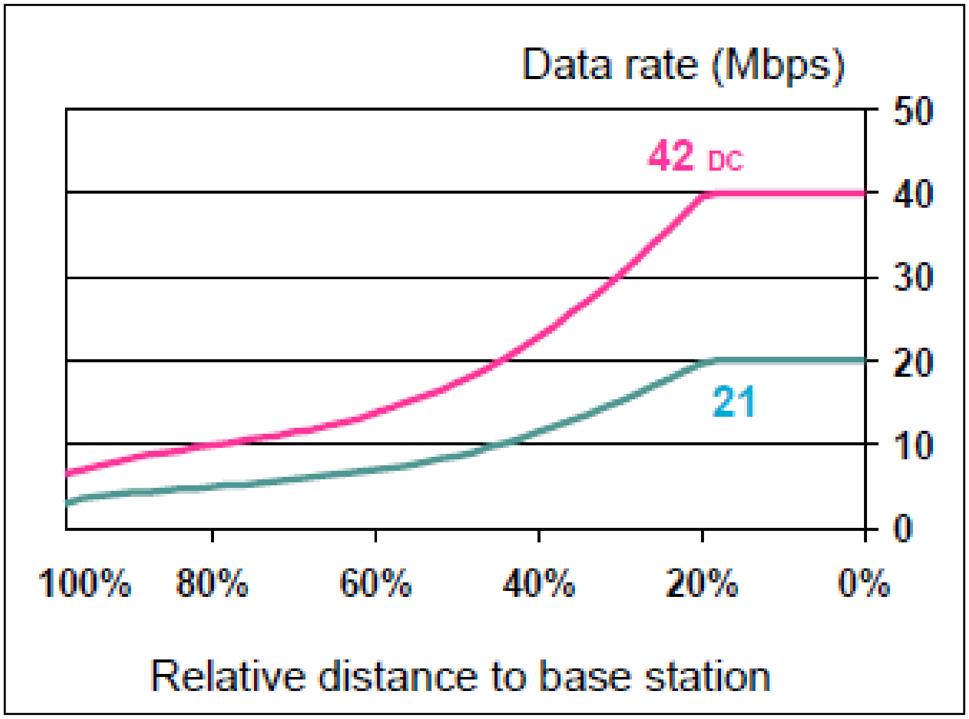

HSPA+ also has improved latency performance of as low as 25 msec and improved packet call setup time of below 500 msec.

Figure 113 summarizes the key capabilities and benefits of the features being deployed in HSPA+.

---

[208] 5G Americas member company contribution. 64 QAM.

Figure 113: Summary of HSPA Functions and Benefits[209]

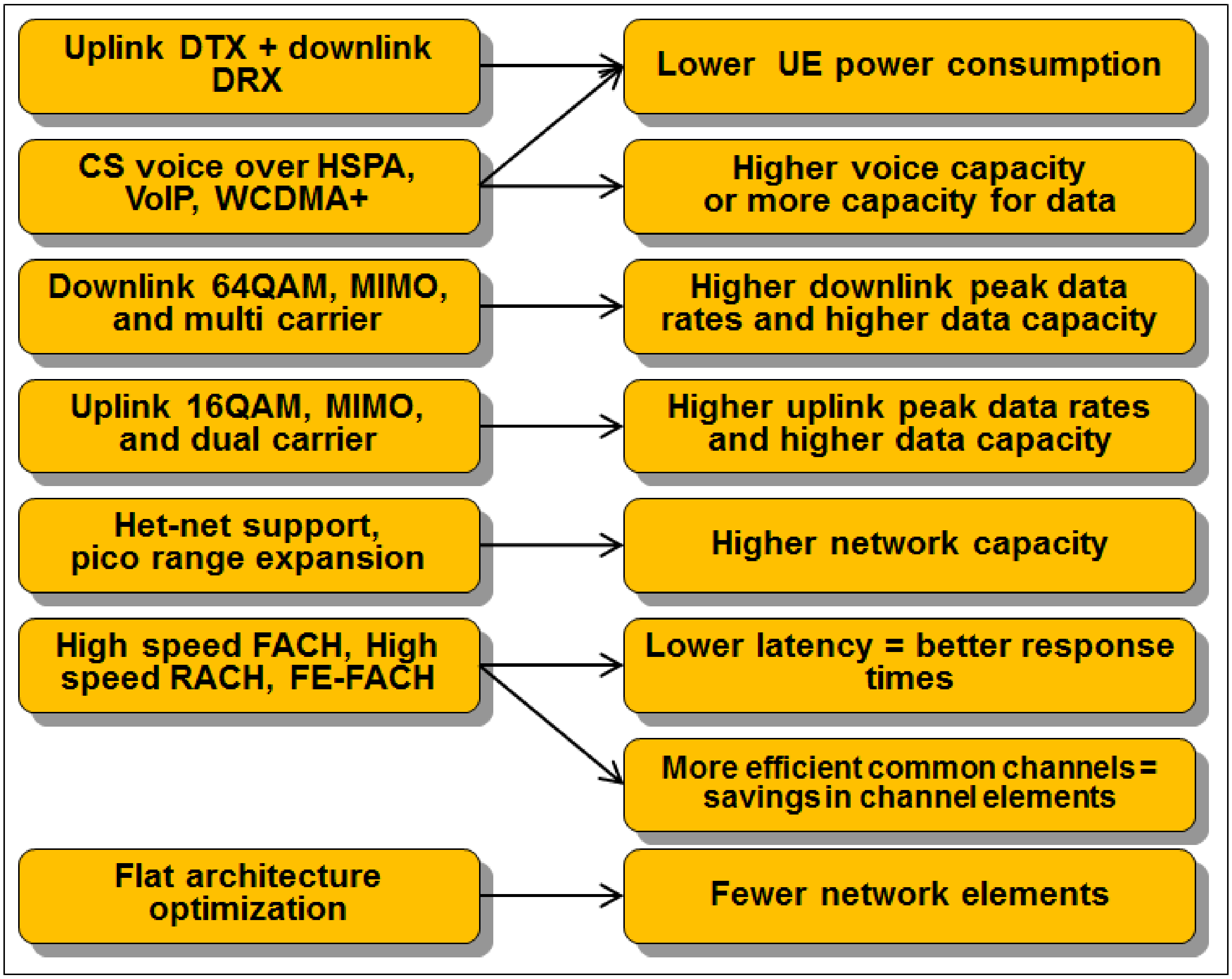
Uplink DTX + downlink DRX
CS voice over HSPA, VoIP, WCDMA+
Downlink 64QAM, MIMO, and multi carrier
Uplink 16QAM, MIMO, and dual carrier
Het-net support, pico range expansion
High speed FACH, High speed RACH, FE-FACH
Flat architecture optimization
Lower UE power consumption
Higher voice capacity or more capacity for data
Higher downlink peak data rates and higher data capacity
Higher uplink peak data rates and higher data capacity
Higher network capacity
Lower latency = better response times
More efficient common channels = savings in channel elements
Fewer network elements

UMTS TDD and TD-SCDMA

Most WCDMA and HSDPA deployments are based on FDD, which uses different radio bands for transmit and receive. In the alternate TDD approach, transmit and receive functions alternate in time on the same radio channel. 3GPP specifications include a TDD version of UMTS, called “UMTS TDD.”

TDD does not provide any inherent advantage for voice functions, which need balanced links—namely, the same amount of capacity in both the uplink and the downlink. Many data applications, however, are asymmetric, often with the downlink consuming more bandwidth than the uplink. A TDD radio interface can dynamically adjust the downlink-to-uplink ratio accordingly, hence balancing both forward-link and reverse-link capacity. Note that for UMTS FDD, the higher spectral efficiency achievable in the downlink versus the uplink addresses the asymmetrical nature of average data traffic.

---

[209] 5G Americas member contribution.

The UMTS TDD specification also includes the capability to use joint detection in receiver-signal processing, which offers improved performance.

One consideration, however, relates to available spectrum. Various countries around the world including those in Europe, Asia, and the Pacific region have licensed spectrum available specifically for TDD systems. TDD is also a good choice for any spectrum that does not provide a duplex gap between forward and reverse links.

In the United States, there is limited spectrum specifically allocated for TDD systems, the major band being BRS at 2.5 GHz used by Clearwire for WiMAX and now LTE TDD.[210] UMTS TDD is not a good choice in FDD bands; it would not be able to operate effectively in both bands, thereby making the overall system efficiency relatively poor.

TDD systems require network synchronization and careful coordination between operators or guardbands, which may be problematic in certain bands.

There has not been widespread deployment of UMTS TDD.

Time Division Synchronous Code Division Multiple Access (TD-SCDMA) is one of the official 3G wireless technologies, mostly for deployment in China. Specified through 3GPP as a variant of the UMTS TDD System and operating with a 1.28 megachips per second (Mcps) chip rate versus 3.84 Mcps for UMTS TDD, TD-SCDMA's primary attribute is that it supports very high subscriber densities, making it a possible alternative for wireless local loops. TD-SCDMA uses the same core network as UMTS, and it is possible for the same core network to support both UMTS and TD-SCDMA radio-access networks.

Although there are no planned deployments in any country other than China, TD-SCDMA could theoretically be deployed anywhere unpaired spectrum is available—such as the bands licensed for UMTS TDD—assuming appropriate resolution of regulatory issues.

EDGE/EGPRS

Today, most GSM networks support EDGE, an enhancement to GPRS, which is the original packet data service for GSM networks.[211] GPRS provides a packet-based IP connectivity solution supporting a wide range of enterprise and consumer applications. GSM networks with EDGE operate as wireless extensions to the internet and give users internet access, as well as access to their organizations from anywhere. Peak EDGE user-achievable[212] throughput rates are up to 200 Kbps. Figure depicts the system architecture.

---

[210] The 1910-1920 MHz band targeted unlicensed TDD systems but has never been used.

[211] GSM technology also provides circuit-switched data services, which are not described in this paper since they are seldom used.

[212] "Peak user-achievable" means users, under favorable conditions of network loading and signal propagation, can achieve this rate as measured by applications such as file transfer. Average rates depend on many factors and will be lower than these rates.

Figure 114: GSM/GPRS/EDGE Architecture

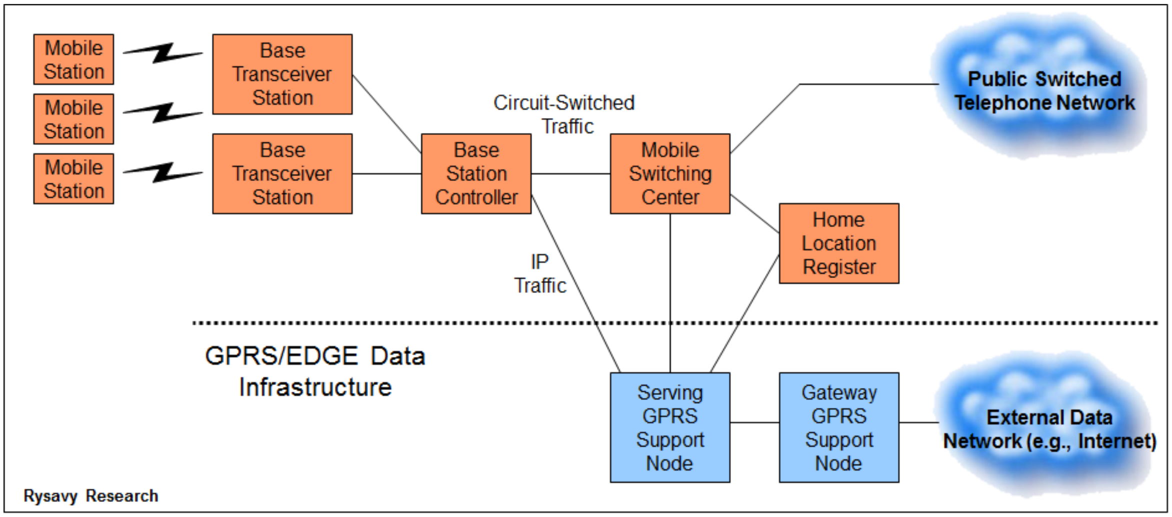
Mobile Station
Mobile Station
Mobile Station
Base Transceiver Station
Base Transceiver Station
Base Station Controller
Circuit-Switched Traffic
Mobile Switching Center
Public Switched Telephone Network
Home Location Register
IP Traffic
GPRS/EDGE Data Infrastructure
Serving GPRS Support Node
Gateway GPRS Support Node
External Data Network (e.g., Internet)

Rysavy Research

EDGE is essentially the addition of a packet-data infrastructure to GSM. In fact, this same data architecture is preserved in UMTS and HSPA networks, and the data architecture is technically referred to as GPRS for the core-data function in all of these networks. The term GPRS may also be used to refer to the initial radio interface, now supplanted by EDGE. Functions of the data elements are as follows:

- The base station controller directs/receives packet data to/from the Serving GPRS Support Node (SGSN), an element that authenticates and tracks the location of mobile stations.
- The SGSN performs the types of functions for data that the Mobile Switching Center (MSC) performs for voice. Each serving area has one SGSN, and it is often collocated with the MSC.
- The SGSN forwards/receives user data to/from the Gateway GPRS Support Node (GGSN), which can be viewed as a mobile IP router to external IP networks. Typically, there is one GGSN per external network (for example, the internet). The GGSN also manages IP addresses, dynamically assigning them to mobile stations for their data sessions.

Another important element is the Home Location Register (HLR), which stores users' account information for both voice and data services. Of significance is that this same data architecture supports data services in GSM and in UMTS-HSPA networks, thereby simplifying operator network upgrades.

In the radio link, GSM uses radio channels of 200 kilohertz (kHz) width, divided in time into eight timeslots comprising 577 microseconds (μs) that repeat every 4.6 msec, as shown in Figure . The network can have multiple radio channels (referred to as transceivers) operating in each cell sector. The network assigns different functions to each timeslot such as the Broadcast Control Channel (BCCH), circuit-switched functions like voice calls or data calls, the optional Packet Broadcast Control Channel (PBCCH), and packet data channels. The network can dynamically adjust capacity between voice and data functions, and it can also reserve minimum resources for each service. This scheduling approach enables more data traffic when voice traffic is low or, likewise, more voice traffic when data traffic is low, thereby maximizing overall use of the network. For example, the PBCCH, which expands the capabilities of the normal BCCH, may be set-up on an additional timeslot of a Time Division Multiple Access (TDMA) frame when justified by the volume of data traffic.

Figure 115: Example of GSM/EDGE Timeslot Structure[213]

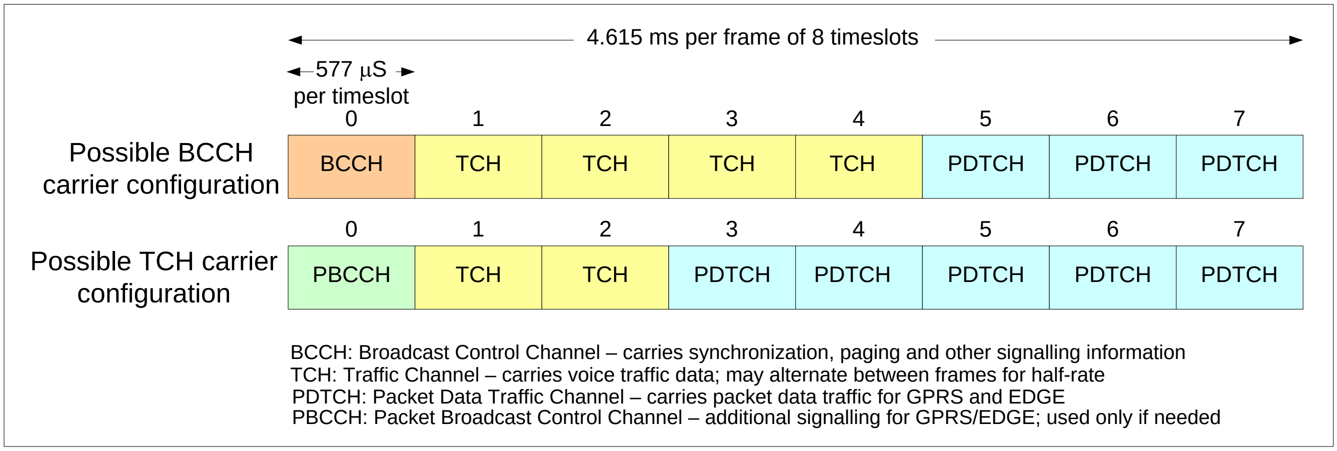

BCCH: Broadcast Control Channel – carries synchronization, paging and other signalling information
TCH: Traffic Channel – carries voice traffic data; may alternate between frames for half-rate
PDTCH: Packet Data Traffic Channel – carries packet data traffic for GPRS and EDGE
PBCCH: Packet Broadcast Control Channel – additional signalling for GPRS/EDGE; used only if needed

EDGE offers close coupling between voice and data services. In most networks, while in a data session, users can accept an incoming voice call, which suspends the data session, and then resume their data session automatically when the voice session ends. Users can also receive SMS messages and data notifications[214] while on a voice call, as described below.

With respect to data performance, each data timeslot can deliver peak user-achievable data rates of up to about 40 Kbps. The network can aggregate up to five timeslots on the downlink and up to four timeslots on the uplink with current devices.

If multiple data users are active in a sector, they share the available data channels. As demand for data services increases, however, an operator can accommodate customers by assigning an increasing number of channels for data service that is limited only by that operator’s total available spectrum and radio planning.

EDGE is an official 3G cellular technology that can be deployed within an operator's existing 850, 900, 1800, and 1900 MHz spectrum bands. EDGE capability is now largely standard in new GSM deployments. A GPRS network using the EDGE radio interface is technically called an “Enhanced GPRS” (EGPRS) network, and a GSM network with EDGE capability is referred to as GSM Edge Radio Access Network (GERAN). EDGE has been an inherent part of GSM specifications since Release 99. It is fully backward-compatible with older GSM networks, meaning that GPRS devices work on EDGE networks and that GPRS and EDGE terminals can operate simultaneously on the same traffic channels.

Dual Transfer Mode (DTM) devices can simultaneously communicate voice and data. DTM is a 3GPP-specified technology that enables new applications like video sharing while providing a consistent service experience (service continuity) with UMTS. Typically, a DTM end-to-end solution requires only a software upgrade to the GSM/EDGE radio network. There are a number of networks and devices supporting DTM.

---

[213] 5G Americas member company contribution.

[214] Example: WAP notification message delivered via SMS.

A feature in Release 9 that applies to EDGE is the Enhanced Flexible Timeslot Assignment (EFTA), which allows for more efficient adaptation to varying uplink versus downlink transmission needs. The network allocates uplink and downlink timeslots that overlap in time, and the mobile station may either use the corresponding uplink timeslots for transmission or receive on the overlapping downlink time slot, if it has nothing to transmit. In addition, alternative EFTA multi-slot classes enable the support of as many as eight timeslots per downlink carrier (instead of five or six timeslots with multi-slot classes 30 to 45).

Abbreviations and Acronyms

The following abbreviations are used in this paper. Abbreviations are defined on first use.

1G – First Generation
1xEV-DO – One Carrier Evolution, Data Optimized
1xEV-DV – One Carrier Evolution, Data Voice
1XRTT – One Carrier Radio Transmission Technology
2G – Second Generation
3G – Third Generation (meeting requirements set forth by the ITU IMT project)
3GPP – Third Generation Partnership Project
3GPP2 – Third Generation Partnership Project 2
4G – Fourth Generation (meeting requirements set forth by the ITU IMT-Advanced project)
5GC – 5G Core
8-PSK – Octagonal Phase Shift Keying
AAS – Adaptive Antenna Systems
ABR – Allocation Retention Priority
AGW – Access Gateway
AMF – Access and Mobility Management Function
AMR – Adaptive Multi Rate
AMR-WB – Adaptive Multi-Rate Wideband
ANDSF – Access Network Discovery and Selection Function.
ANSI – American National Standards Institute
APCO – Association of Public Safety Officials
API – Application Programming Interface
APN – Access Point Name
ARP – Allocation Retention Priority
ARPU – Average Revenue per User
ARQ – Automatic Repeat Request
ATM – Asynchronous Transfer Mode
AWGN – Additive White Gaussian Noise Channel
AWS – Advanced Wireless Services
BCCH – Broadcast Control Channel
bps – bits per second
BRS – Broadband Radio Service
BSC – Base Station Controller
BTS – Base Transceiver Station
C/I – Carrier to Intermodulation Ratio
CAPEX- Capital Expenditure
CBF – Coordinated Beam Forming
CBRS – Citizens Broadband Radio Service

CBS – Coordinated Beam Switching

CSS3 – Cascading Style Sheets 3 (CSS3)

CDD – Cyclic Delay Diversity

CDF – Cumulative Distribution Function

CDMA – Code Division Multiple Access

CL – Closed Loop

CL-SM – Closed Loop Spatial Multiplexing

CMAS – Commercial Mobile Alert System

CMOS – Complementary Metal Oxide Semiconductor

CoMP – Coordinated Multi Point

cMTC – Critical Machine Type Communications

CP – Cyclic Prefix

CPC – Continuous Packet Connectivity

CPRI – Common Public Radio Interface

CQI - Channel Quality Indicators

C-RAN – Cloud Radio Access Network

CRM – Customer Relationship Management

CRS – Cell-specific Reference Signal

CS – Convergence Sublayer

CSFB – Circuit-Switched Fallback

CTIA – Cellular Telephone Industries Association

C-V2X – Cellular Vehicle-to-X

D-AMPS – Digital Advanced Mobile Phone Service

DAS – Distributed Antenna System

DAS – Downlink EGPRS2-A Level Scheme

dB – Decibel

DBS – Downlink EGPRS2-B Level Scheme

DC-HSPA – Dual Carrier HSPA

DFT – Discrete Fourier Transform

DL – Downlink

DNS – Domain Name Service

DPCCH – Dedicated Physical Control Channel

DPS – Dynamic Point Selection

DSL – Digital Subscriber Line

DSMIPv6 – Dual Stack Mobile IPv6

DTM – Dual Transfer Mode

DRX – Discontinuous Reception

D-TxAA – Double Transmit Adaptive Array

DVB-H – Digital Video Broadcasting Handheld

E-DCH – Enhanced Dedicated Channel
EBCMCS – Enhanced Broadcast Multicast Services
EC-GSM – Extended Coverage GSM
eCoMP – enhanced CoMP
EDGE – Enhanced Data Rates for GSM Evolution
EFTA – Enhanced Flexible Timeslot Assignment
EGPRS – Enhanced General Packet Radio Service
eICIC – Enhanced Inter-Cell Interference Coordination
eMBMS – Enhanced Multimedia Broadcast Multicast Services
eNodeB – Evolved Node B
EAP – Extensible Authentication Protocol
eLAA – Enhanced Licensed-Assisted Access
eNB – Evolved Node B
EPC – Evolved Packet Core
EPDCCH – Enhanced Physical Downlink Control Channel
eMBB – Enhanced Mobile Broadband
ePDG – Enhanced Packet Data Gateway
EPS – Evolved Packet System
ERP – Enterprise Resource Planning
eSaMOG – Enhanced S2a-based Mobility over GTP
ESC – Environmental Sensing Capability
eSRVCC – Enhanced Single-Radio Voice Call Continuity
ETRI – Electronic and Telecommunications Research Institute
ETSI – European Telecommunications Standards Institute
E-UTRAN – Enhanced UMTS Terrestrial Radio Access Network
EVS – Enhanced Voice Services (codec)
FE-FACH – Further Enhanced Forward Access Channel
EV-DO –Evolution, Data Optimized
EV-DV –Evolution, Data Voice
EVRC – Enhanced Variable Rate Codec
FBMC – Filter-Bank Multi-Carrier
FCC – Federal Communications Commission
FDD – Frequency Division Duplex
FeCoMP – Further Enhanced Coordinated Multi Point
feICIC – Further enhanced ICIC
FirstNet – First Responder Network Authority
Flash OFDM – Fast Low-Latency Access with Seamless Handoff OFDM
FLO – Forward-Link Only
FMC – Fixed Mobile Convergence

FP7 – Seventh Framework Programme
FTP – File Transfer Protocol
GAA – General Authorized Access
GAN – Generic Access Network
GB – Gigabyte
Gbps – Gigabits Per Second
GBR – Guaranteed Bit Rate
GByte – Gigabyte
GCS – Group Communication Service
GERAN – GSM EDGE Radio Access Network
GFDM – Generalized Frequency Division Multiplexing
GGSN – Gateway GPRS Support Node
GHz — Gigahertz
GMSK – Gaussian Minimum Shift Keying
gNB – NR NodeB
GPRS – General Packet Radio Service
G-Rake – Generalized Rake Receiver
GSM – Global System for Mobile Communications
GSMA – GSM Association
HARQ – Hybrid Automatic Repeat Request
HD – High Definition
HetNet – heterogeneous network
HLR – Home Location Register
Hr – Hour
HSDPA – High Speed Downlink Packet Access
HS-FACH – High Speed Forward Access Channel
HS-PDSCH - High Speed Physical Downlink Shared Channels
HS-RACH – High Speed Reverse Access Channel
HSPA – High Speed Packet Access (HSDPA with HSUPA)
HSPA+ – HSPA Evolution
HSS – Home Subscriber Server
HSUPA – High Speed Uplink Packet Access
Hz – Hertz
ICIC – Inter-Cell Interference Coordination
ICN – Information-Centric Networking
ICS – IMS Centralized Services
ICT – Information and Communication Technologies
IEEE – Institute of Electrical and Electronic Engineers
IETF – Internet Engineering Taskforce

IFFT – Inverse Fast Fourier Transform

IFOM – IP Flow and Seamless Offload

IM – Instant Messaging

IMS – IP Multimedia Subsystem

IMT – International Mobile Telecommunications

IMT-Advanced - International Mobile Telecommunications-Advanced

IRC – Interference Rejection Combining

IoT – Internet of Things

IPR - Intellectual Property Rights

IP – Internet Protocol

IPTV – Internet Protocol Television

IR – Incremental Redundancy

ISD – Inter-site Distance

ISI – Intersymbol Interference

ISP – Internet Service Provider

ITU – International Telecommunication Union

JCP – Java Community Process

JR – Joint Reception

JT – Joint Transmission

Kbps – Kilobits Per Second

kHz — Kilohertz

km – Kilometer

LAA – License-Assisted Access

LBT – Listen-Before-Talk

LDPC - Low-Density Parity Code

LIPA – Local IP Access

LMDS – Local Multipoint Distribution Service

LPWA – Low-Power Wide-Area

LTE – Long Term Evolution

LTE-A – LTE-Advanced

LTE-TDD – LTE Time Division Duplex

LTE-U – LTE-Unlicensed

LSTI – LTE/SAE Trial Initiative

LWA – LTE Wi-Fi Aggregation

LWIP – LTE WLAN Radio Level Integration with IPsec Tunnel

M2M – Machine-to-machine

MAC – Medium-Access Control

MAPCON – Multi-Access PDN Connectivity

MB - Megabyte

MBMS - Multimedia Broadcast/Multicast Service

Mbps – Megabits Per Second

MBR – Maximum Bit Rate

MBSFN – Multicast/broadcast, Single Frequency

MCPA – Mobile Consumer Application Platform

Mcps – Megachips Per Second

MCPTT – Mission-Critical Push-to-Talk

MCS – Modulation and Coding Scheme

MCW – Multiple Codeword

MDT – Minimization of Drive Tests

MEAP – Mobile Enterprise Application Platforms

MEC – Multi-access Edge Computing

MediaFLO – Media Forward Link Only

METIS – Mobile and wireless communications Enablers for the Twenty-twenty Information Society

MHz – Megahertz

MID – Mobile Internet Devices

MIMO – Multiple Input Multiple Output

MMSE – Minimum Mean Square Error

mITF – Japan Mobile IT Forum

MMDS – Multichannel Multipoint Distribution Service

MME – Mobile Management Entity

mMTC – Massive Machine Type Communications

MOS – Mean Opinion Score

MP-QUIC – Multipath Quick UDP Internet Connections

MP-TCP – Multipath TCP

MRxD – Mobile Receive Diversity

ms – millisecond

MS – Mobile Station

MSA – Mobile Service Architecture

MSC – Mobile Switching Center

MTC – Machine Type Communications

MTC-IWF – Machine-Type Communications Interworking Function (MTC-IWF)

msec – millisecond

MU-MIMO – Multi-User MIMO

MUST – Downlink Multiuser Superposition Transmission

NAICS – Network-Assisted Interference Cancellation and Suppression

NB-IoT – Narrowband Internet of Things

NENA – National Emergency Number Association

NGMC – Next Generation Mobile Committee

NGMN – Next Generation Mobile Networks Alliance
NG-RAN – New Generation Radio Access Network
NOMA – Non-Orthogonal Multiple Access
NR – New Radio
NTIA – National Telecommunications and Information Administration
OFDM – Orthogonal Frequency Division Multiplexing
OFDMA – Orthogonal Frequency Division Multiple Access
OL-SM – Open Loop Spatial Multiplexing
OMA – Open Mobile Alliance
ORI – Open Radio Equipment Interface
PA – Priority Access
PAL – Priority Access License
PAR – Peak to Average Ratio
PBCCH – Packet Broadcast Control Channel
PCH – Paging Channel
PCRF – Policy Control and Charging Rules Function
PCS – Personal Communications Service
PDCP – Packet Data Convergence Protocol
PDN – Packet Data Network
PGW – Packet Gateway
PHY – Physical Layer
PMI – Precoding Matrix Indication
PMIPv6 – Proxy Mobile IPv6
PNF – Physical Network Function
PoC – Push-to-Talk Over Cellular
PSH – Packet Switched Handover
PSK – Phase-Shift Keying
QAM – Quadrature Amplitude Modulation
QCI – Quality of Service Class Identifier
QLIC – Quasi-Linear Interference Cancellation
QoS – Quality of Service
QPSK – Quadrature Phase Shift Keying
QUIC – Quick UDP Internet Connections.
RAB – Radio Access Bearer
RAN – Radio Access Network
RCAF – RAN Congestion Awareness Function
RCLWI - RAN Controlled LTE WLAN Interworking
RCS – Rich Communications Suite
REST – Representational State Transfer

RF – Radio Frequency
RLC – Radio Link Control
RNC – Radio Network Controller
ROHC – Robust Header Compression
RRC – Radio Resource Control
RRH – Remote Radio Head
RRU – Remote Radio Unit
RTP – Real Time Transport Protocol
RTSP – Real Time Streaming Protocol
SAE – System Architecture Evolution
SaMOG – S2a-based Mobility over GTP
SAS – Spectrum Access System
SC-FDMA – Single Carrier Frequency Division Multiple Access
SCMA – Sparse Coded Multiple Access
SCRI – Signaling Connection Release Indication
SCW – Single Codeword
SDAP – Service Data Adaptation Protocol
SDMA – Space Division Multiple Access
SDN – Software-Defined Networking
SDP – Session Description Protocol
sec – Second
SFBA – Space Frequency Block Code
SFN – Single Frequency Network
SGSN – Serving GPRS Support Node
SGW – Serving Gateway
SIC – Successive Interference Cancellation
SIM – Subscriber Identity Module
SIMO – Single Input Multiple Output
SINR – Signal to Interference Plus Noise Ratio
SIP – Session Initiation Protocol
SIPTO – Selected IP Traffic Offload
SISO – Single Input Single Output
SMS – Short Message Service
SNR – Signal to Noise Ratio
SON – Self-Organizing Network
SPS – Semi-Persistent Scheduling
SRVCC – Single-Radio Voice Call Continuity
SU-MIMO – Single User MIMO
SVDO – Simultaneous 1XRTT Voice and EV-DO Data

SVLTE – Simultaneous Voice and LTE
TCH – Traffic Channel
TCP/IP – Transmission Control Protocol/IP
TD – Transmit Diversity
TDD – Time Division Duplex
TDMA – Time Division Multiple Access
TD-SCDMA – Time Division Synchronous Code Division Multiple Access
TD-CDMA – Time Division Code Division Multiple Access
TETRA – Terrestrial Trunked Radio
TIA/EIA – Telecommunications Industry Association/Electronics Industry Association
TISPAN – Telecoms and Internet Converged Services and Protocols for Advanced Networks
TSG-RAN – Technical Services Group Radio Access Network
TTI – Transmission Time Interval
UAS – Uplink EGPRS2-A Level Scheme
UBS – Uplink EGPRS2-B Level Scheme
UE – User Equipment
UFMC – Universal Filtered Multi-Carrier
UICC – Universal Integrated Circuit Card
UL – Uplink
UMA – Unlicensed Mobile Access
UMB – Ultra Mobile Broadband
UMTS – Universal Mobile Telecommunications System
UPCON – User-Plane Congestion Management
UPF – User Plane Function
URA-PCH – UTRAN Registration Area Paging Channel
URI – Uniform Resource Identifier
URLLC – Ultra-Reliable and Low Latency Communications
us – Microsecond
USIM – UICC SIM
UTRAN – UMTS Terrestrial Radio Access Network
V2X – vehicle -to-infrastructure
VAMOS – Voice Services over Adaptive Multi-User Channels on One Slot
VDSL – Very-High-Bit-Rate DSL
VEPC – Virtualized EPC
ViLTE – Video Over LTE
VoIP – Voice over Internet Protocol
VoHSPA – Voice over HSPA
VoLGA – Voice over LTE Generic Access
VoLTE – Voice over LTE

VNF- Virtual Network Function
VPN – Virtual Private Network
WAP – Wireless Application Protocol
WBA – Wireless Broadband Alliance
WCDMA – Wideband Code Division Multiple Access
WCS – Wireless Communication Service
WebRTC – Web Real-Time Communication
Wi-Fi – Wireless Fidelity
WiMAX – Worldwide Interoperability for Microwave Access
WLAN – Wireless Local Area Network
WMAN – Wireless Metropolitan Area Network
WMM – Wi-Fi Multimedia
WRC – World Radiocommunication Conference

Additional Information

5G Americas maintains market information, LTE deployment lists, and numerous white papers, available for free download on its web site: http://www.5gamericas.org.

If there are any questions regarding the download of this information, please call +1 425 372 8922 or e-mail Anushka Bishen, Public Relations Coordinator at info@5gamericas.org.

This white paper was written for 5G Americas by Rysavy Research (http://www.rysavy.com) and utilized a composite of statistical information from multiple resources.

The contents of this paper reflect the independent research, analysis, and conclusions of Rysavy Research and may not necessarily represent the comprehensive opinions and individual viewpoints of each particular 5G Americas Board member company.

Rysavy Research provides this document and the information contained herein to you for informational purposes only. Rysavy Research provides this information solely on the basis that you will take responsibility for making your own assessments of the information.

Although Rysavy Research has exercised reasonable care in providing this information to you, Rysavy Research does not warrant that the information is error-free. Rysavy Research disclaims and in no event shall be liable for any losses or damages of any kind, whether direct, indirect, incidental, consequential, or punitive arising out of or in any way related to the use of the information.

Copyright © 2017 Rysavy Research, LLC. All rights reserved.

http://www.rysavy.com

Select Language ▼

Robert P. Duncan A.B., S.M., M.B.A., Ph.D.: Government warfare and surveillance system architect, author, and independent investigator..

Presents

ASK ME ANYTHING ON REDDIT. POST QUESTION&VOTE UP. INDEFINITE Q&A. ALSO LATEST CLASSIFIED INFO.

SEE YOU AT NOVEMBER 5th 2017 WHITE HOUSE MMM!!!! BE THERE!

Project Soul Catcher Vol. 2: Secrets of Cyber and Cybernetic Warfare Revealed (click book image to buy)

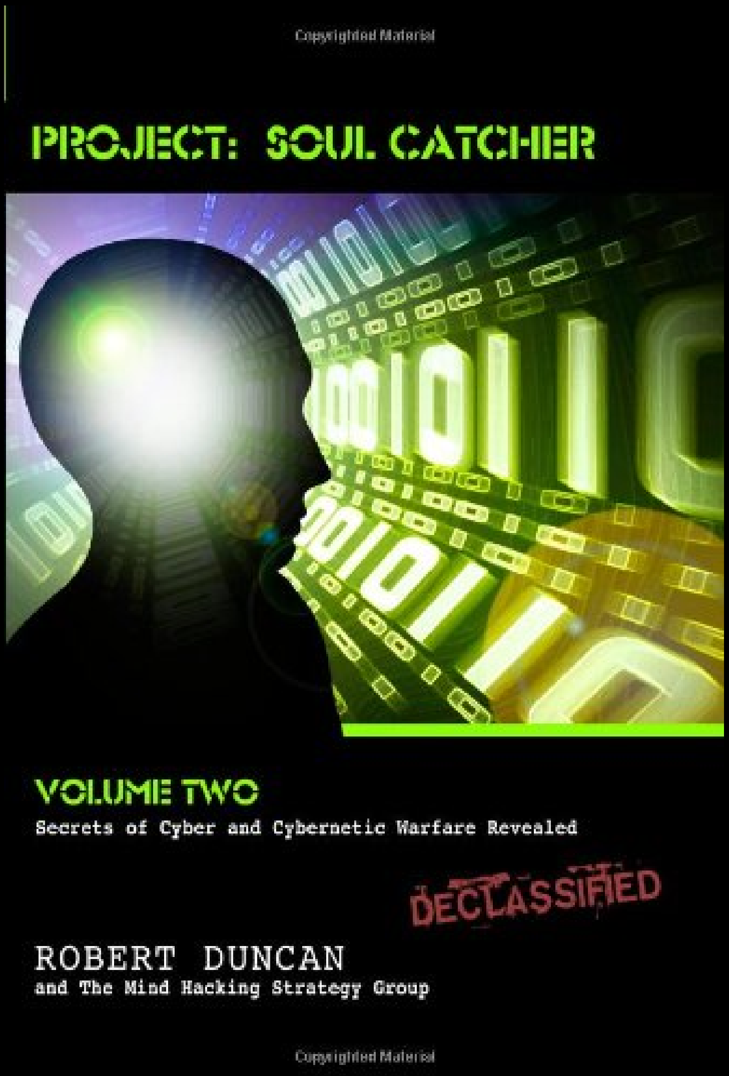

Copyrighted Material

PROJECT: SOUL CATCHER

VOLUME TWO

Secrets of Cyber and Cybernetic Warfare Revealed

DECLASSIFIED

ROBERT DUNCAN

and The Mind Hacking Strategy Group

Copyrighted Material

Volume 2 details the CIA's practices of interrogation and cybernetic mind control in their pursuit to weaponize neuropsychology. It covers the art of bio-communication war. Human beings are complex machines but their inner workings have been deciphered. Mind control and brainwashing have been perfected in the last 60 years. Hacking computers and hacking into individual minds are similar. The 21st century will be known as the age of spiritual machines and soulless men.

Check out Dr. Robert Duncan's other book, How to Tame a Demon: A short practical guide to organized intimidation stalking, electronic torture, and mind control.

More information on Dr. Robert Duncan is available at drrobertduncan.com.

new SSL secure chat #targets

DRROBERTDUNCAN.COM GUEST BOOK & FORUM. PROFESS YOUR LOVE, DROP A LINE, SHARE YOUR STORY, WHATEVER.

Free eBook: The Matrix Deciphered by Dr. Robert Duncan, A.B., S.M., M.B.A., Ph.D., the book discloses secret details on government operations and gives a hint to what it's like to live like a victim targeted..

Bio of Dr. Robert Duncan, from his book The Matrix Deciphered:

"True nobility is exempt from fear".
– King Henry the Sixth, Part II (Act IV, Scene I).

> Call me The Saint. I am the all American - prep school, Harvard College graduating with honors in computer science and a minor in premedical studies, and advanced degrees from Harvard and Dartmouth in business and science. My famous ancestors are President Lincoln, King Duncan of Scotland, and Governor William Bradford, the first governor of Massachusetts.
>
> My research interests have been neural networks, virtual reality, and EEG controlled robotics. Before graduate school I worked for the Department of Defense, Navy, NATO, and various intelligence agencies computer science projects. I have done business consulting and computer consulting for the largest companies in the world. I have been a professor, inventor, artist, and writer. I am one of the last Renaissance men.
>
> My projects have included algorithms for Echelon and CIA natural language parsing and classification of document content, IRS formula for red flagging audits, writing the artificial intelligence code to automate tracking of the Soviet Nuclear Submarine Fleet and all water vessels, work integrating HAARP with SIGINT SIGCOM and SPAWAR. I have worked on projects for the Justice Department connecting local, state, and federal databases for the tracking of terrorists. I developed a system for the FBI to track license plates past toll booths and other locations. I worked on the soldier 2000 program to create body networks for reading vital signs and other information. A system I worked on called Snyper is operational in Iraq which triangulates on intercity conflict gun shots. I have been to a couple secret bases in the so called free world. I have developed telemedicine robotic surgery and virtual reality applications for the Army. For DARPA, I have worked on satellite computer vision target tracking applications and tank simulation as well as integration of the land, sea, and air surveillance systems like SOSSUS, towed arrays, and others.
>
> Projects that I have worked on outside of government contracts include my thesis on computer generated holography, a project making paralyzed people walk again using choreographed stimulated muscles movements, face recognition, voice identification and recognition, finger print recognition, and neural network robotic controller. My research interests moved to enhanced reality heads up displays and wearable computing systems. My current research involves finding a cure for the mind control directed energy weapons fiasco. The integrated global surveillance grid is actually part of the holy grail of weapons and human control systems.
>
> My apologies to the human race for my contribution to tyranny. I was tricked into thinking it would not be misused by corrupt government especially in my beloved country. I was wrong. The Joint Non-Lethal Weapons Directorate has Skip Green on the governing board. One of my old colleagues at a technology think tank in Cambridge partly in charge of the radio frequency weapons testing for neurological disruption now torturing and killing people worldwide. Several other US Navy and UK Navy scientists have been knocked-off and that is why I have come forward. I know my time maybe near. I am currently a professor in computer science and business, helping to educate the public on government corruption, greed, and stupidity. Like my ancestor Lincoln, I am fighting against mental slavery in a new Civil War engineered by the same useless elements in over 80 government agencies who have tested radiological, drug, and viral weapons on unsuspecting Americans for over 45 years.
>
> I have great pride in the fundamental and constitutional values of America and mean no disrespect by my blasting the incompetence, apathy, and stupidity of those involved in the conspiracy of involuntary biological, chemical, and psychotronic human effects testing. There are so many brave men and women serving in the armed forces who fight and protect us for the American values they believe in. But because of the silos of information called compartmentalization in the security agencies and the lack of accountability and oversight of black operations and some top secret projects, diabolical elements become rogue or worse destroying the very country they are tasked with protecting turning it into a hypocracy rather than a democracy.
>
> My goal is to awaken Americans to the continued horrific acts of military and CIA weapons testers in this country and the other branches of government's inability to stop them and hold them accountable.
>
> All I ask from the reader is to listen to testimony and validate facts presented here and come to their own conclusion. Then fight to win back America from this silent overthrow. This has been my project for about two years, investigating government corruption, incompetence, and cover-up upon which I accidentally stumbled one day while looking at a reverse MRI scheme to inject electromagnetic signals into the brain for virtual reality applications. I have interviewed over two hundred people and worked on the highest level of military projects for the U.S., NATO, and U.K. and have given videotaped testimony to senators and representatives on this topic resulting in lip service since they have no real power to enforce. Two high level FBI agents and a couple CIA agents have come forward to validate the existence of a MKULTRA like project that continues to grab random people for mind control experimentation. Two of these have since become part of the program and endure daily psychotronic tortures. All the torture can be done using directed energy psychotronic weapons with the so called continental ballistic missile surveillance defense grids.
>
> I apologize to the human race for any contribution to these 4th generation weapons that I may have worked on that are more horrific than the nuclear bomb and whose cover-up is more pervasive than the Manhattan Project. And because of the horrific acts of violence being committed on as many as two thousand Americans as far as my research has uncovered and many others in other countries, I understand the extreme risk to my own welfare that publishing this material will have. But freedom is so important to me that I know full heartedly that the human race

must have an open discussion on these weapons to decide their own fate before the point of decision is gone, that I am willing to risk divulging so called national secrets. All I offer you is the truth.

All information presented in this book was received through legitimate channels such as the Freedom of Information Act, military documents, victim testimony, and turned agents. I still hold valid the oath I took to keep secret the details of the projects that I worked on under DoD budgets. The majority of the proceeds of this book will go towards helping the psychotronic experiment survivors and surviving families of those that have perished. When the government fails, business and the citizens must look out for each other.

Dr. Robert Duncan in Energy Weapons & Testing, Coast to Coast AM broadcast December 5th 2006 (2.75+ million listeners weekly, syndicated in most cities and areas of the United States, and other places around the world.)..

0:57 / 1:53:43

Download - Dr. Robert Duncan, Coast to Coast AM interview - Energy Weapons & Testing.

Energy Weapons & Testing synopsis:

Independent investigator Robert Duncan discussed directed energy and neurological weapons and his contention that they've been tested on the public at large. While directed energy is used in microwaves, to remove kidney stones, and in non-invasive surgery, it's also been developed extensively for military purposes, he reported.

The civilian population was targeted for experiments, in programs such as MK-ULTRA, starting after WW II, when Nazi scientists were imported to the US, some working on scalar or gravity weapons, said Duncan. The town of Taos, NM, where a hum was heard by many citizens, was the subject of a directed energy experiment by the U.S. Navy, he claimed. And most recently the Active Denial System (see article below) was tested on human subjects.

Duncan said he interviewed over 600 mind control victims (Dec 5th 2006 report), and found some validity to their allegations. There are weapons that can project voices into people's heads such as one system known as "The Voice of God," he detailed. Blocking techniques include jamming the signals with electronic scramblers, and using shields with metal alloys or mylar. He recommended the following websites for further information/assistance: mindjustice.org, raven1.net, freedomfchs.com.

The Goodbye Weapon

A new non-lethal weapon, the Air Force's Active Denial System, or ADS, has been certified for use in Iraq, after extensive testing. The ADS shoots a beam of waves that causes extreme (but temporary) pain and induces what experimenters call the "Goodbye effect," or "prompt and highly motivated escape behavior." Wired News obtained documents about the weapon and has published two reports (article(1)/documentation(2)). Additionally, the report's author David Hambling has posted a commentary(3) on the subject at Defense Tech.

1. http://www.wired.com/news/technology/0,72134-0.html
2. http://www.wired.com/news/technology/0,72236-0.html
3. http://www.defensetech.org/archives/003036.html

Coast to Coast AM guest bio:

Dr. Robert Duncan holds multiple degrees from Harvard University and Dartmouth College amongst others. He has had the most expensive American education money can buy. He is an investigator, author, and soon be movie producer on the topics of directed energy, neurological weapons, psychological, and information warfare. His movie is called "The Enemy Within - Psychic Warfare". A book he is co-authoring will be out in a few months called "Hacking the Human Mind".

Dr. Duncan has worked as a business and information technology consultant to the Fortune 500. He has worked for companies like Oracle Corporation, BEA systems, HP, BBN, and as a professor at a college. For the department of defense he has written the artificial intelligence code to track the Soviet nuclear submarine fleet with passive and active acoustical arrays and has been to a couple secret NATA Navy underground bases in Europe.

-Todd Giffen
Target since birth as I was born into American enslavement, surveillance, and control, under attack with directed energy since 2008
503-967-5202
case@oregonstatehospital.net
obamasweapon.com

Copyright © 2015 Todd Giffen

"The Use of Psychotronic Weapons"
"What everyone should know"

(courtesy : http://www.mindjustice.org/cherkova.htm)

Definition:

Definition of psychotronic (psycho-physical) weapons.

Psychotronic Weapons (PF- weapons) this is the totality of all possible methods and means (techno-genic, suggestive, pharmacological, paranormal, complexes, and others) of hidden, forced influences on the psyche of a person for the purpose of modifying his consciousness, behavior and health for what is desired in the way of influencing aspects of control..."

Second definition of psychotronic weapons (as an epigraph): "Applications in military and for antisocial purposes are possibilities also for knowledge of psychotronics, its means, methods, systems, 'generators', and there is the path for the transformation of the humanitarian, essential attributes of life and society in the weapons. These are also psychotronic weapons. This is not only dangerous, this is deadly!"

Types of Weapons:

- Electromagnetic Weapons
- Microwave Weapons
- Non-Lethal Weapons
- ELF (Extremely Low Frequency) Weapons
- Directed Energy Weapons
- Acoustic Weapons
- Psychotronic Weapons
- RF (Radio Frequency) Weapons
- Soft Kill Weapons
- Less-Than-Lethal Weapons

How it affects (Physical & psycho-physical damages):

Physical damage, visible results achieved at the moment of influence.

* burns, pricks, cuts, wounds (ray weapons, laser)
* bleeding into the brain (laser ray with energy pulses of 40 joules) and on the skin energy in pulses: 3 - 100 joules)
* Blindness in the eyes, loping of eyelashes (laser ray)
* loss of memory (influence of frequency of up to 100Hz. and a reduction to 5 Hz.)
* destruction of the vestibulary [inner ear] apparatus, spatial orientation
* symptoms of [***] [with a large exposure and strength, paralysis (UBCh) in the area of the beginning of the spinal column
* irregular sleeping, heart-beating
* numbness of the hands; at night, cramping of the legs
* sharp decrease in hemoglobin, micro-burns to the skin, ionization of the blood and body fluids, an increase in the deterioration rate of clothing material, photographic material [illuminating], (thermal-electronic emissions from lamps in the room with the feeding units of electrical networks and fittings of the wall, [...]
* vibrations (separate organs can vibrate: the heart, kidneys, etc.), [...] (from

* vibrational technologies)
* pressure, coolness of the skin
* feelings of non-specific fear and panic (from the weapons) [...] tremors, tapping, destruction of fragile objects from the waves, from the impulses and from other manifestations of technological poltergeist

Psycho-physical damage depending on intensity and duration of effects. Damage in the physical sphere (up to programmed premature death).

Stage I.

* changes in the blood (early signs)
* overheating substances especially and increase in crystallization, leading to cataracts (early sign)
* ill reactions along the path of the bio-energetic potential from the effects [...]
* prickly sensations, tremors to the muscles of the feet
* aggravation of chronic illnesses
* internal bleeding (nose, stomach, gynecological)
* auditory tones of various frequencies
* sharp pain in the internal ear
* dull pains in the area behind the ears
* internal trembling, vibrations
* head pain in the temple and a warmness in that area, pain in the eye-sockets
* likelihood of miscarriage, an increase of 80% for chances of premature births, anomalous
* development of the fetus, often with defective brains

Stage II.

* "radio-sound" effect (before this is long-conducted the resonance-loosening of spinal and brain fluids at the frequency of 2-10 Hz. for maintaining the fluids of theskull and brain in stimulation
* steady reactions to all aspects of coding
* lowering of energy (observed with irradiation at the frequency carrying frequencies of ~780 Hz.and amplitude modulation ~71 Hz.)
* aggravation of dermatosis
* eye pains (cataracts, asymmetry in lenses of the eyes, puffiness [?] of the retinas)
* aggravation of the course of chronic illnesses
* destruction of the functions of the kidneys, liver, organs of digestion, circulatory system, bleeding in gingivae, gums, peridontal and others
* gauntness
* "sand" in the eyes, burning sensation, loss of hair, brittleness of nails
* compression pain in the heart fits of coughing
* changes of the color of the face (bronze shading of the skin)
* asymmetrical puffiness of the face
* disfiguration of the face, the intentional emaciation of separate muscles, cartilage
* syndromes of destruction to the external nerves of the thighs
* impotence

Stage III.

* destruction of the brain, irreversible changes to the structure of the brain (academician V.P. Vekhtereva)

* irreversible processes in the muscle material, in the bones of the skull, and in the structures of neurons "..minimalization of life" --external reaction to EMP appears as an illness, up to heavy, sharp, chronic, organic pathological processes (academician V.P. Kaznacheev)
* "..the entire speeding up of the aging process, sclerosis, oncogenesis (academician V.P. Kaznacheev)
* destruction of bio-chemical processes in the cells with distrophic changes in the organs (academician V.P. Kaznacheev)
* destruction of the cell membranes, chemical connections in the cells -leading to leukemia, cancer of the brain and prostate gland, breast cancer, (even with weak EMP, but of long duration "...the genetic apparatus is altered..." (academician V.P. Kaznacheev)
* the reflex activities are destroyed
* the functioning of the hormonal system is destroyed
* [....] Parkinson's disease (trembling of the hands) and Alzheimer's disease (from changes in the production of the hormone melanin ([...] gland) that causes degenerative changes in substances at the molecular level disfunctions of the sino-atrial node of the heart (even with a low frequency)
* infarction, cerebral thrombosis, osteo-[***], peridontal, sugar diabetes
* allergies
* epilepsy (especially caused by the multiple-rapid stimulation by a weak, low frequency waves)
* hypertension, hypotension
* programmed or instantaneous (with the subjection to extreme effects -up to the border of deadly conditions) premature death --it is well known by the experimentors that cases of death do occur (academician V.P. Kaznacheev)
* a decrease and dying off of the population, humanity, mankind

Damage in the mental sphere (up to the point of insanity).
Stage I.

* manifestations of depression
* lowering in the interest in perceived information
* shortness of temper (irritability)

Stage II.

* steady depression
* lowering of abilities for analytical thinking, some worsening of the memory
* symptoms of chronic fatigue
* superficial nature of sleep
* higher incidents of conflicts

Stage III.

* confusion in the accounting of events in real life
* loss of ability for analytical thinking
* lowering and a loss of memory
* feelings of the lack of sleep
* indifference to personal external appearance [...]
* lowering of self-control from the deficit of the hormone seratonin due to SVCh irradiation
* suicide (as the experimentors "practice" provocation techniques)
* "...invisible weak fields ...their ignoring that they are carrying with them a deadly danger"... "...people simply start going crazy and nobody understands why..." (academician V.P. Kaznacheev)

"Consequences of effects of micro-leptonic generators are described by the leading developer in this area [from the] V. Lab. of Micro-leptonic Technologies D. Okhatryym: in 1982 there began extremely unpleasant things with the health our co-workers. At first, there arose their feeling unwell, then the loss of logic, then after this, the loss of spatial orientation. Finally, a breakdown of diseased [affected] organs. The coming out of this condition -and this includes myself -was achieved with great difficulty."

(in the book by Ju. Vorob'evskii" Knock at the Golden Gate, Moscow, 1999.

Manipulation with the memory.

* Erasure
* Blockage
* Recording of information needed by the experimentor
* training -"prompting" (after erasure)
* comparison,
* reproduction of picture a from a prior testing as though bifurcated [?]
* a shifting of the effective frequency up to 100 Hz. or a lowering to 5 Hz., with memory loss

Control of Stimulation in the brain.

* by states of consciousness (up to the loss of these)
* in sleep
* emotions
* intellectual capabilities
* attention
* speech
* activeness
* muscle tone
* stimulation of one and the same zone with stimulations of:
o 30 -40 mV- activization of attention, memory, movement;
o 60-70 mV- "emotiono-genic" effects

Different methods:

Types of Psychotronic, psychotropic and other Psy-Influences.

* radio-waves
* laser-irradiation
* infra-red irradiation
* ultra-violet irradiation UF
* x-ray irradiation
* gravitational fields
* torsional fields (hypothetical)
* leptonic and other fields

Acoustic Waves.

* infra-sonic (IZ) from 0 Hz. to 15-20 KHz.
* ultra-sonic(UZ) from 15-20 KHz. to 10**9 Hz. (ranges for IZ and UZ)

see more similar information in -Table 3

Formation of fields:

* electro-magnetic charge
* gravitational-mass
* torsional (surrounding spin fields) -spin, with any rotating body
* micro-leptonic energy (in free aspect) -[over-abundance of stable particles]?

Operator is secretly brought in, or in a secret installation nearby

BG: Biogenerator (could be tomography)
A: Antenna (irradiation and reception)
D: operator for the gathering of information, transmission and control
ZU: reminding sytem on a magnetic strip, holographic
BIP: bio-informational [...]
KC: channel connections (IK, radio, telephone, telegraph)
E: screen display
O: operator
SP: light pen with which the operator indicates on the screen, the point of influence

Stress systems and methods of harassment.

* slander
* gestures, hand movements, expressions, tests, practical "jokes"
* taking off what is needed for the eyes (vision) [?]
* psychological pressure through family, neighbors or surrounding people, vehicle drivers
* discrediting: along 5 directions (often with the involvement of the police)
o irresponsibility
o spy
o drug dealer
o vagrant
o accomplice in a crime (in an alleged crime)
* torture, assault
* blackmail, threats
* attempts on their lives with PTOs
* death to someone close or in immediate surrounding life
* murder

Stages of Zombification (Substitution of consciousness with external control).

* programming, modeling of a cybernetic double
* conducting of programs according to non-differentiation of a proper and artificial "I"-"Not-I"
* extreme influencing, stress, harassment, irradiation, special-preparations [substances], chemicals
* taking to the limits of mental and physical capacities (survivability)
* loss of will, memory
* shutting off of his own consciousness
* replacement of consciousness with the control from his cybernetic twin

IDEAS
IN ORBIT
EXPLORE THE TECH
Raytheon

Did You Miss It? Get Breaking Defense Coverage Of SNA Symposium Here →

ACQUISITION, **AIR**, **INTEL & CYBER**

Pentagon Studies Weapons That Can Read Users' Mind

By SYDNEY J. FREEDBERG JR.
on July 14, 2017 at 3:15 PM
15 Comments

28
SHARES

NEWSEUM: The troops of tomorrow may be able to pull the trigger using only their minds. As artificially intelligent drones, hacking, jamming, and missiles accelerate the pace of combat, some of the military's leading scientists are studying how mere humans can keep up with the incredible speed of cyber warfare, missiles and other threats.

One option: Bypass crude physical controls — triggers, throttles, keyboards — and plug the computer directly into the human brain. In one DARPA experiment, a quadriplegic first controlled an artificial limb and then flew a flight simulator. Future systems might monitor the users' nervous system and compensate for stress, fatigue, or injury. Is this the path to what the Pentagon calls human-machine teaming?

This is an unnerving scenario for those humans, like Stephen Hawking, who mistrust artificial intelligence. If your nightmare scenario is robots getting out of control, "let's teach them to read our minds!" is probably *not* your preferred solution. It sounds more like the beginning of a movie where cyborg Arnold Schwarzenegger goes back in time to kill someone.

Navy X-47B drone.

But the Pentagon officials who talked up this research yesterday at Defense One's annual tech conference emphasized the objective was to *improve* human control over artificial intelligence. Teaching AI to monitor its user's level of stress, exhaustion, distraction, and so on helps the machine adapt itself to better serve the human — instead of the other way around. Teaching AI to instantly detect its user's *intention* to give a command, instead of requiring a relatively laborious push of a button, helps the human keep control — instead of having to let the AI off the leash because no human can keep up with it.

Official Defense Department policy, as then-Secretary Ash Carter put it, is that the US will "never" allow an artificial intelligence to decide for itself whether or not to kill a human being. However, no less a figure than the Carter's undersecretary of acquisition and technology, Frank Kendall, fretted publicly that making our robots wait for human permission would slow them down so much that enemy AI without such constraints would beat us. Vice-Chairman of the Joint Chiefs, Gen. Paul Selva, calls this the "Terminator Conundrum." Neuroscience suggests a way out of this dilemma: Instead of slowing the AIs down, make the humans' orders come faster.

DARPA's Revolutionizing Prosthetics program is devising new kinds of artificial limbs — and new ways to control them.

Accelerate Humanity

"We will continue to have humans on the loop, we will have human input in decisions, but the way we go about that is going to have to shift, just to cope with the speed and the capabilities that autonomous systems bring," said Dr. James Christensen, portfolio manager at the Air Force Research Laboratory's 711th Human Performance Wing. "The decision cycle with these systems is going to be so fast that they have to be sensitive to and responsive to the state of the individual (operator's) *intent*, as much as overt actions and control inputs that human's providing."

In other words, instead of the weapon system responding to the human operator *physically* touching a control, have it respond to the human's brain cells forming the *intention* to use a control. "When you start to have a direct neural interface of this type, you don't necessarily need to command and control the aircraft using the stick," said Justin Sanchez, director of DARPA's Biological Technologies Office. "You could potentially re-map your neural signatures onto the different control surfaces" — the tail, the flaps — "or maybe any other part of the aircraft" — say landing gear or weapons. "That part hasn't really been explored in a huge amount of depth yet."

Reading minds, even in this limited fashion, will require deep understanding and close monitoring of the brain, where thoughts take measurable form as electrical impulses running from neuron to neuron. "Can we develop precise neurotechnologies that can go to those circuits in the brain or the peripheral nervous system in real time?" Sanchez asked aloud. "Do we have computational systems that allow us to understand what the changes in those signals (mean)? And can we give meaningful feedback, either to the person or to the machine to help them to do their job better?"

DARPA's Revolutionizing Prosthetics program hooked up the brain of a quadriplegic — someone who could move neither arms nor legs — to a prosthetic arm, allowing the patient to control it directly with their thoughts. Then, "they said, 'I'd like to try to fly an airplane,'" Sanchez recounted. "So we created a virtual flight simulator for this person, allowed this neural interface to interface with the virtual aircraft, and that person flew."

"That was a real wake-up call for everybody involved," Sanchez said. "We didn't initially think you could do that."

timesofindia
May
21
19:40
Tuesday
100%
21 Tuesday May 2013

Tony Stark (Robert Downey Jr.) relies on the JARVIS artificial intelligence — exquisitely adapted to his personal strengths and weaknesses — to help pilot his Iron Man suit. (Marvel Comics/Paramount Pictures)

Adapting To The Human

Applying direct neural control to real aircraft — or tanks, or ships, or network cybersecurity systems — will require a fundamental change in design philosophy. Today, said Christensen, we give the pilots tremendous information on the aircraft, then expect them to adapt to it. In the future, we could give the aircraft tremendous information on its pilots, then have it use artificial intelligence to adapt itself to *them*. The AI could customize the displays, the controls, even the mix of tasks it took on versus those it left up to the humans — all exquisitely tailored not just to the preferences of the individual operator but to his or her current mental and physical state.

When we build planes today, "they're incredible sensor platforms that collect data on the world around them and on the aircraft systems themselves, (but) at this point, very little data is being collected on the pilot," Christensen said. "The opportunity there with the technology we're trying to build now is to provide a continuous monitoring and assessment capability so that the aircraft knows the state of that individual. Are they awake, alert, conscious, fully capable of performing their role as part of this man-machine system? Are there things that the aircraft then can do? Can it unload gees (i.e. reduce g-forces)? Can it reduce the strain on the pilot?"

Air Force MQ-9 Reaper operators in training. Today's drones require extensive human oversight and are hardly user friendly.

"This kind of ability to sense and understand to the state and the capabilities of the human is absolutely critical to the successful employment of highly automated systems," Christensen said. "The way all of our systems are architected right now, they're fixed, they're predictable, they're deterministic" — that is, any given input always produces the exact same output.

Predictability has its advantages: "We can train to that, they behave in very consistent ways, it's easier to test and evaluate," Christensen said. "What we lose in that, though is the real power of highly automated systems, autonomous systems, as *learning systems*, of being able to adapt themselves. "

"That adaptation, though, it creates unpredictability," he continued. "So the human has to adapt alongside the system, and in order to do that, there has to be some mutual awareness, right, so the human has to understand what is the system doing, what is it trying to do, why is that happening; and vice versa, the system has to has some understanding of the human's intent and also their state and capabilities."

This kind of synergy between human and artificial intelligence is what some theorists refer to as the "centaur," after the mythical creature that combined human and horse — with the human part firmly in control, but benefiting from the beast's strength and speed. The centaur concept, in turn, lies at the heart of the Pentagon's ideas of human-machine teaming and what's known as the Third Offset, which seeks to counter (offset) adversaries' advancing technology with revolutionary uses of AI.

The neuroscience here is in its infancy. But it holds the promise of a happy medium between hamstringing our robots with too-close control or letting them run rampant.

28
SHARES

Topics: AFRL, air force, Air Force Research Laboratory AFRL, artificial intelligence, darpa, DefenseOne Tech Summit 2017, James Christensen, Justin Sanchez, neuroscience, offset strategy, War Algorithm

SPONSORED CONTENT
Raytheon's trusted missile defense solutions help the U.S. and its allies make the world a safer place

Breaking
Defense
Click here to view gallery »

EDITORIAL STAFF

Colin Clark
Editor

Sydney J. Freedberg, Jr.
Deputy Editor

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/313112342

Internet of Things (IoT) in 5G Wireless Communications

Article *in* IEEE Access · January 2016

DOI: 10.1109/ACCESS.2016.2646120

CITATIONS
2

READS
680

7 authors, including:

Waleed Ejaz
Ryerson University
69 PUBLICATIONS **397** CITATIONS
SEE PROFILE

Muhammad Ali Imran
University of Glasgow
402 PUBLICATIONS **4,102** CITATIONS
SEE PROFILE

Minho Jo
Korea University
98 PUBLICATIONS **1,145** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project IU-ATC View project

Project Visible Light Communication View project

All content following this page was uploaded by Minho Jo on 27 March 2017.

The user has requested enhancement of the downloaded file.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2016.2646120

GUEST EDITORIAL

Internet of Things (IoT) in 5G Wireless Communications

During the past decade, the Internet of Things (IoT) has revolutionized the ubiquitous computing with multitude of applications built around various types of sensors. A vast amount of activity is seen in IoT based product-lines and this activity is expected to grow in years to come with projections as high as billions of devices with on average 6-7 devices per person by year 2020. With most of the issues at device and protocol levels solved during the past decade, there is now a growing trend in integration of sensors and sensor based systems with cyber physical systems and device-to-device (D2D) communications. 5th generation wireless systems (5G) are on the horizon and IoT is taking the center stage as devices are expected to form a major portion of this 5G network paradigm. IoT technologies such as machine to machine communication complemented with intelligent data analytics are expected to drastically change landscape of various industries. The emergence of cloud computing and its extension to fog paradigm with proliferation of intelligent 'smart' devices is expected to lead further innovation in IoT. These developments excite us and form a motivation to survey existing work, design new techniques, and identify new applications of IoT. Researchers, scientists, and engineers face emerging challenges in designing IoT based systems that can efficiently be integrated with the 5G wireless communications.

We received enthusiastic response to our special issue call for papers. A total of nine high quality papers were received out of which only seven were selected after a thorough review process. Invited articles were sought from two highly cited and accomplished researchers. Dr. Mischa Dohler (Fellow IEEE, Kings College London, UK) and Dr. Mung Chiang (Fellow IEEE, Princeton University, USA).

The first article "MIMO-NOMA design for small packet transmission in the Internet of Things", written by a collaborative team of researchers Ding et al.,, opportunistic serving mechanism is designed as part of the effort in the novel paradigm of Multiple Input Multiple Output Non-Orthogonal Multiple Access transmission scheme. Under the proposed method, one user is completely served with their quality of service requirements completely taken care of whereas second user is served opportunistically under the NOMA paradigm. The main contribution of this article is a design with two sets of system parameters, precoding and power allocation coefficients, in order to ensure that the potential of NOMA can be realized even if the users' channel conditions are similar. Two types of power allocation policies are developed in this paper. One is to meet first user's QoS requirements in the long term. For example, in order to satisfy its targeted outage probability. The other is in which to realize second user's QoS requirements instantaneously, e.g., the power allocation coefficients are designed to realize its targeted data rate for each channel realization.

Machine to machine communication has a significant role to play in emerging internet of things paradigm in years and decades to come. The emerging IoT-5G scenario extends sensor based IoT capabilities to robots, actuators and drones for distributed coordination and low-latency reliable execution of tasks at hand. In the invited work titled "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications" by Condoluci et al., core attention is focused on the end-to-end reliability, latency, and energy consumption comprising both up and downlinks for 5G-IoT communication. The authors propose the definition, design, and analysis of machine-type multicast service (MTMS). They recommend different procedures that need to be redesigned for MTMS and derive the most appropriate design drivers by analyzing different performance indicators, such as scalability, reliability, latency, and energy consumption. Overall, a very interesting read complemented by open problems and future research directions to pursue.

Security is one of the biggest challenges faced by Internet of Things. With devices becoming ubiquitous and pervasive in day to day lives necessitate reliable and secure algorithms. The third article, "Security enhancement for IoT communications exposed to eavesdroppers with uncertain locations" by Xu et al., develops a secure framework for eavesdroppers with Uncertain Locations in IoT. With the assumption that the locations of eavesdroppers change independently from hop to hop, authors derive an expression for the secrecy outage probability of the two-hop transmission, which is shown to be the upper bound of the outage probability when the locations of eavesdroppers remain unchanged. Following this expression, the end users formulate a secrecy rate maximization problem with the secrecy-outage probability constraint. The optimal rate design for codebooks and power allocation between the source and relay are derived. By studying the performance of the optimal scheme in some special cases, we obtain several insights concerning the setting of system parameters.

In the article, "Enabling massive IoT in 5G and beyond systems: PHY radio frame design considerations" by Ayesha Ijaz et al., the authors propose a flexible frame structure and design for massive Internet of Things (IoT) devices working in 5G wireless network. The authors also discussed

2169-3536 © 2017 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

the interdependence of different frame design parameters, service requirements and characteristics of radio environment. Based on these interdependency, they provide guidelines for radio numerology design and elaborated on the frame design for IoT communications in 5G networks to support massive connection density of low-rate, low-power devices. The article concludes with some key research findings and challenges massive IoT in 5G wireless network.

It is estimated that in year 2020, 20 to 40 billion devices will be connected to the Internet as part of the Internet of Things. A critical bottleneck for realizing the efficient IoT is the pressure it puts on the existing communication infrastructures, requiring transfer of enormous data volumes. In the article ''CONDENSE: A reconfigurable knowledge acquisition architecture for future 5G IoT'', by Dejan Vukobratovic et al., the authors propose a architecture named 'Condense' which integrates the acquisition of IoT-generated data within the 3GPP MTC (machine type communications) systems. The proposed Condense architecture introduces a service within 3GPP MTC systems – computing linear and non-linear functions over the data generated by MTC devices This service brings about the possibility that the underlying communication infrastructure communicates only the desired function of the MTC- generated data (as required by the given application at hand), and not the raw data in its entirety. This transformational approach has the potential to dramatically reduce the pressure on the 3GPP MTC communication infrastructure. The article concludes by discussing challenges, provides insights, and identifies future research directions for implementing function computation and function decomposition within practical 3GPP MTC systems.

In the article, ''Frequency-domain oversampling for cognitive CDMA systems: Enabling robust and massive multiple access for Internet of Things'' by Su Hu et al., the authors utilize the concept of cognitive radio with dynamic non-continuous spectrum bands and code division multiple access to tackle the challenge of massive spectrum resource management in IoT. In order to suppress multiple access interference resulting from the non-orthogonality of partial available spectrum bins, carrier frequency offset, and spectrum sensing mismatch, the authors propose an enhanced receiver design that combines the frequency-domain oversampling scheme (FDO) and linear minimum mean square error (MMSE) method. The simulation results show that the cognitive-CDMA with FDO-MMSE receiver outperforms that with conventional per-user MMSE receiver in the presence of multipath fading channels, carrier frequency offset, and spectrum sensing mismatch.

In the invited article, ''A survey of client-controlled HetNets for 5G'' by Michael Wang et al., a comprehensive review is provided on spectrum of client-controlled HetNets for 5G networks: from the fully devolved distributed local control approach to the hybrid control approach where clients may make the decisions given some global information provided by the network. After giving a thorough review, the authors also provide future research directions and recommendations for evolution of 5G heterogeneous networks as an enabler for next generation internet of things.

To conclude, the Editors would like to thank all authors who submitted their manuscripts against this Special issue. There had been some very exciting submissions though we could only accept best seven high quality submissions. The editorial staff also thanks the reviewers for their timely comments on the manuscript which helped us improve the presentation and quality of the content manifold. All the associate editors thank the Editor in Chief Dr. Michael Pecht for his assistance in making this special issue a reality. We would also like to thank the publications editor Ms. Kimberly Shumard for her continuous follow up on timelines and management of articles to ensure timely completion of the review process. We complete this note by letting the readers know to stay tuned with IEEE Access as we expect even more exciting issues in months to come.

WALEED EJAZ
Ryerson University,
Toronto, ON, Canada

ALAGAN ANPALAGAN
Ryerson University,
Toronto, ON, Canada

MUHAMMAD ALI IMRAN
University of Glasgow,
Glasgow, U.K.

MINHO JO
Korea University,
Sejong, South Korea

MUHAMMAD NAEEM
Ryerson University,
Toronto, ON, Canada

SAAD BIN QAISAR
National University of Sciences and Technology,
Islamabad, Pakistan

WEI WANG
Zhejiang University,
Hangzhou, China

WALEED EJAZ (S'12–M'14–SM'16) received the B.Sc. and M.Sc. degrees in computer engineering from the National University of Sciences and Technology, Islamabad, Pakistan, and the University of Engineering and Technology, Taxila, Pakistan, respectively, and the Ph.D. degree in information and communication engineering from Sejong University, South Korea, in 2014. He was with top engineering universities in Pakistan and Saudi Arabia as a Faculty Member. He was a Post-Doctoral Fellow with Queen's University, Kingston, Canada. He is currently a Senior Research Associate with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. He also completed certificate courses on teaching and learning in higher education from the Chang School, Ryerson University. His current research interests include Internet of Things, energy harvesting, 5G cellular networks, and mobile cloud computing. He is currently serving as an Associate Editor of the *Canadian Journal of Electrical and Computer Engineering* and the IEEE Access. In addition, he is handling the special issues in *IET Communications*, the IEEE Access, and the *Journal of Internet Technology*.

ALAGAN ANPALAGAN (S'98–M'01–SM'04) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto. He is currently a Professor with the Department of Electrical and Computer Engineering, Ryerson University, where he directs the research group involved in radio resource management and radio access and networking areas within the WINCORE Lab. He served as an Editor of the IEEE Communications Surveys and Tutorials from 2012 to 2014, the IEEE Communications Letters from 2010 to 2013, the *Wireless Personal Communications* (Springer) from 2011 to 2013, and the EURASIP *Journal of Wireless Communications and Networking* from 2004 to 2009. He has co-authored three edited books, *Design and Deployment of Small Cell Networks* (Cambridge University Press, 2014), *Routing in Opportunistic Networks* (Springer, 2013), and *Handbook on Green Information and Communication Systems* (Academic Press, 2012).

Dr. Anpalagan is a registered Professional Engineer in ON, Canada, and a fellow of the Institution of Engineering and Technology. He served as the TPC Co-Chair of the IEEE Globecom'15: SAC Green Communication and Computing, the IEEE WPMC'12 Wireless Networks, the IEEE PIMRC'11 Cognitive Radio and Spectrum Management, and the IEEE CCECE'04/08. He served as the Vice Chair of the IEEE SIG on Green and Sustainable Networking and Computing with Cognition and Cooperation (2015–), the IEEE Canada Central Area Chair (2012–2014), the IEEE Toronto Section Chair (2006–2007), the ComSoc Toronto Chapter Chair (2004–2005), and the IEEE Canada Professional Activities Committee Chair (2009–2011). He was a recipient of the Dean's Teaching Award in 2011, the Faculty Scholastic, Research, and Creativity Award in 2010 and 2014, the Faculty Service Award in Ryerson University in 2011 and 2013. He also received the Exemplary Editor Award from the IEEE ComSoc in 2013 and an Editor-in-Chief Top10 Choice Award in the IEEE Transactions on Emerging Telecommunications Technology in 2012.

MUHAMMAD ALI IMRAN (M'03–SM'12) received the M.Sc. (Hons.) and Ph.D. degrees from the Imperial College London, U.K., in 2002 and 2007, respectively. He is currently a Professor in Communication Systems with the University of Glasgow, the Vice Dean of the Glasgow College UESTC, and the Program Director of Electrical and Electronics with Communications. From 2007 to 2016, he was an Adjunct Associate Professor with The University of Oklahoma, USA, and a Visiting Professor with the 5G Innovation Centre, University of Surrey, U.K.

He has led a number of multimillion-funded international research projects encompassing the areas of energy efficiency, fundamental performance limits, sensor networks, and self-organizing cellular networks. He also led the new physical layer work area for 5G Innovation center at Surrey. He has a global collaborative research network spanning both academia and key industrial players in the field of wireless communications. He has supervised 21 successful Ph.D. graduates and published over 200 peer-reviewed research papers, including over 20 IEEE Transaction papers. He secured first rank in his B.Sc. and a distinction in his M.Sc. degree along with an award of excellence in recognition of his academic achievements conferred by the President of Pakistan. He received the IEEE Comsoc's Fred Ellersick Award 2014, the Sentinel of Science Award 2016, and the FEPS Learning and Teaching Award 2014 and twice nominated for Tony Jean's Inspirational Teaching Award. He is a shortlisted finalist for The Wharton-QS Stars Awards 2014, the Reimagine Education Award for innovative teaching, and the VC's learning and teaching award in the University of Surrey. He is a Senior Fellow of the Higher Education Academy, U.K.

Dr. Imran has given an invited TEDx talk (2015) and over ten plenary talks, several tutorials and seminars in international conferences and other institutions. He has taught international short courses in USA and China. He is the Co-Founder of the IEEE Workshop BackNets 2015 and chaired several tracks/workshops of international conferences. He is an Associate Editor of the IEEE Communications Letters, the IEEE Access, and the *IET Communications* journal.

MINHO JO (SM'16) received the B.A. degree from the Department of Industrial Engineering, Chosun University, South Korea, in 1984, and the Ph.D. degree from the Department of Industrial and Systems Engineering, Lehigh University, USA, in 1994. He is currently a Professor with the Department of Computer Convergence Software, Korea University, Sejong, South Korea. He is one of the founders of the Samsung Electronics LCD Division. He was the Vice President of the Institute of Electronics and Information Engineers. He is currently the Vice President of the Korean Society for Internet Information. His current research interests include Internet of Things, AI/big data in Internet of Things, LET-Unlicensed, cognitive radio, mobile cloud computing, network security, heterogeneous networks in 5G, wireless energy harvesting, optimization in wireless communications, and massive MIMO. He is the Founder and an Editor-in-Chief of the *KSII Transactions on Internet and Information Systems* [SCI (ISI) and SCOPUS]. He served as an Editor of the IEEE Network. He is an Editor of the IEEE Wireless Communications, and an Associate Editor of the IEEE Internet of Things Journal, *Security and Communication Networks*, *Ad-Hoc & Sensor Wireless Networks*, and *Wireless Communications and Mobile Computing*.

MUHAMMAD NAEEM received the B.S. and M.S. degrees in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2000 and 2005, respectively, and the Ph.D. degree from Simon Fraser University, BC, Canada, in 2011. From 2000 to 2005, he was a Senior Design Engineer with Comcept (pvt.) Ltd. At the design department of Comcept (pvt) Ltd., he participated in the design and development of smart card-based GSM and CDMA pay phones. From 2012 to 2013, he was a Post-Doctoral Research Associate with the WINCORE Laboratory, Ryerson University, Toronto, ON, Canada. Since 2013, he has been an Assistant Professor with the Department of Electrical Engineering, COMSATS Institute of IT, Wah Campus, Pakistan, and a Research Associate with the WINCORE Laboratory, Ryerson University. He is also a Microsoft Certified Solution Developer. His research interests include optimization of wireless communication systems, nonconvex optimization, resource allocation in cognitive radio networks, and approximation algorithms for mixed integer programming in communication systems. He was a recipient of NSERC CGS Scholarship.

SAAD BIN QAISAR (S'04–M'09–SM'14) received the master's and Ph.D. degrees in electrical engineering from Michigan State University, USA, in 2005 and 2009, respectively. He is currently serving as an Associate Professor with the National University of Sciences and Technology, Pakistan, and also leading Bitsym, an Internet of Things and data-analytics enterprise. He has authored over 60 publications, inventor of various patents, has completed multiple national and international projects and winner of various awards, including the University Best Researcher Award in 2012 and the UNIDO Cleantech Award in 2015, to name a few. His company, Bitsym, has been named among the 50 Most Promising Start-ups for 2012 by Global Entrepreneurship Week, globally.

WEI WANG (S'08–M'10–SM'15) received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2009 respectively. From 2007 to 2008, he was a Visiting Student with the University of Michigan, Ann Arbor, USA. From 2013 to 2015, he was a Hong Kong Scholar with The Hong Kong University of Science and Technology, Hong Kong. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. He has authored over 90 international journal and conference papers in the area of wireless communications and networks. His research interests mainly focus on cognitive radio networks, green communications, and radio resource allocation for wireless networks.

Dr. Wang is the Associate Secretary-General of the Zhejiang Institute of Communications. He is the Editor of the book *Cognitive Radio Systems* (Intech, 2009) and serves as an Editor of the IEEE Access, Transactions on Emerging Telecommunications Technologies, and the *KSII Transactions on Internet and Information Systems*. He also serves as the Co-Chair or a TPC Member for several international conferences.

• • •

View publication stats

ERICSSON WHITE PAPER

Uen 284 23-3204 Rev C | April 2016

5G RADIO ACCESS

CAPABILITIES AND TECHNOLOGIES

The capabilities of 5G wireless access must extend far beyond previous generations of mobile communication. Examples of these capabilities include very high data rates, very low latency, ultra-high reliability, energy efficiency and extreme device densities, and will be realized by the development of LTE in combination with new radio-access technologies. Key technology components include extension to higher frequency bands, access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

WHAT IS 5G?

5G radio access technology will be a key component of the Networked Society. It will address high traffic growth and increasing demand for high-bandwidth connectivity. It will also support massive numbers of connected devices and meet the real-time, high-reliability communication needs of mission-critical applications.

5G will provide wireless connectivity for a wide range of new applications and use cases, including wearables, smart homes, traffic safety/control, critical infrastructure, industry processes and very-high-speed media delivery. As a result, it will also accelerate the development of the Internet of Things.

The overall aim of 5G is to provide ubiquitous connectivity for any kind of device and any kind of application that may benefit from being connected.

5G networks will not be based on one specific radio-access technology. Rather, 5G is a portfolio of access and connectivity solutions addressing the demands and requirements of mobile communication beyond 2020.

The specification of 5G will include the development of a new flexible air interface, NX, which will be directed to extreme mobile broadband deployments. NX will also target high-bandwidth and high-traffic-usage scenarios, as well as new scenarios that involve mission-critical and real-time communications with extreme requirements in terms of latency and reliability.

In parallel, the development of Narrow-Band IoT (NB-IoT) in 3GPP is expected to support massive machine connectivity in wide area applications. NB-IoT will most likely be deployed in bands below 2GHz and will provide high capacity and deep coverage for enormous numbers of connected devices.

Ensuring interoperability with past generations of mobile communications has been a key principle of the ICT industry since the development of GSM and later wireless technologies within the 3GPP family of standards. In a similar manner, LTE will evolve in a way that recognizes its role in providing excellent coverage for mobile users, and 5G networks will incorporate LTE access (based on Orthogonal Frequency Division Multiplexing (OFDM)) along with new air interfaces in a transparent manner toward both the service layer and users.

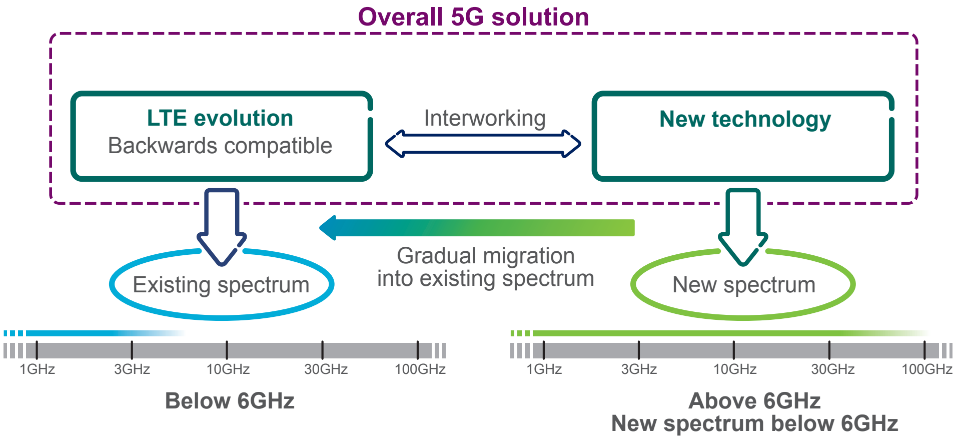

Figure 1: The overall 5G wireless-access solution consisting of LTE evolution and new technology.

Around 2020, much of the available wireless coverage will continue to be provided by LTE, and it is important that operators with deployed 4G networks have the opportunity to transition some – or all – of their spectrum to newer wireless access technologies. For operators with limited spectrum resources, the possibility of introducing 5G capabilities in an interoperable way – thereby allowing legacy devices to continue to be served on a compatible carrier – is highly beneficial and, in some cases, even vital.

At the same time, the evolution of LTE to a point where it is a full member of the 5G family of air interfaces is essential, especially since initial deployment of new air interfaces may not operate in the same bands. The 5G network will enable dual-connectivity between LTE operating within bands below 6GHz and the NX air interface in bands within the range 6GHz to100GHz. NX should also allow for user-plane aggregation, i.e. joint delivery of data via LTE and NX component carriers.

This paper explains the key requirements and capabilities of 5G, along with its technology components and spectrum needs.

5G – REQUIREMENTS AND CAPABILITIES

In order to enable connectivity for a very wide range of applications with new characteristics and requirements, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communication. These capabilities will include massive system capacity, very high data rates everywhere, very low latency, ultra-high reliability and availability, very low device cost and energy consumption, and energy-efficient networks.

MASSIVE SYSTEM CAPACITY

Traffic demands for mobile-communication systems are predicted to increase dramatically [1] [2]. To support this traffic in an affordable way, 5G networks must deliver data with much lower cost per bit compared with the networks of today. Furthermore, the increase in data consumption will result in an increased energy footprint from networks. 5G must therefore consume significantly lower energy per delivered bit than current cellular networks.

The exponential increase in connected devices, such as the deployment of billions of wirelessly connected sensors, actuators and similar devices for massive machine connectivity, will place demands on the network to support new paradigms in device and connectivity management that do not compromise security. Each device will generate or consume very small amounts of data, to the extent that they will individually, or even jointly, have limited impact on the overall traffic volume. However, the sheer number of connected devices seriously challenges the ability of the network to provision signaling and manage connections.

VERY HIGH DATA RATES EVERYWHERE

Every generation of mobile communication has been associated with higher data rates compared with the previous generation. In the past, much of the focus has been on the peak data rate that can be supported by a wireless-access technology under ideal conditions. However, a more important capability is the data rate that can actually be provided under real-life conditions in different scenarios.

- 5G should support data rates exceeding 10Gbps in specific scenarios such as indoor and dense outdoor environments.
- Data rates of several 100Mbps should generally be achievable in urban and suburban environments.
- Data rates of at least 10Mbps should be accessible almost everywhere, including sparsely-populated rural areas in both developed and developing countries.

VERY LOW LATENCY

Very low latency will be driven by the need to support new applications. Some envisioned 5G use cases, such as traffic safety and control of critical infrastructure and industry processes, may require much lower latency compared with what is possible with the mobile-communication systems of today.

To support such latency-critical applications, 5G should allow for an application end-to-end latency of 1ms or less, although application-level framing requirements and codec limitations for media may lead to higher latencies in practice. Many services will distribute computational capacity and storage close to the air interface. This will create new capabilities for real-time communication and will allow ultra-high service reliability in a variety of scenarios, ranging from entertainment to industrial process control.

ULTRA-HIGH RELIABILITY AND AVAILABILITY

In addition to very low latency, 5G should also enable connectivity with ultra-high reliability and ultra-high availability. For critical services, such as control of critical infrastructure and traffic safety, connectivity with certain characteristics, such as a specific maximum latency, should not merely be ‘typically available.’ Rather, loss of connectivity and deviation from quality of service requirements must be extremely rare. For example, some industrial applications might need to guarantee successful packet delivery within 1 ms with a probability higher than 99.9999 percent.

VERY LOW DEVICE COST AND ENERGY CONSUMPTION

Low-cost, low-energy mobile devices have been a key market requirement since the early days of mobile communication. However, to enable the vision of billions of wirelessly connected sensors, actuators and similar devices, a further step has to be taken in terms of device cost and energy consumption. It should be possible for 5G devices to be available at very low cost and with a battery life of several years without recharging.

ENERGY-EFFICIENT NETWORKS

While device energy consumption has always been prioritized, energy efficiency on the network side has recently emerged as an additional KPI, for three main reasons:

- Energy efficiency is an important component in reducing operational cost, as well as a driver for better dimensioned nodes, leading to lower total cost of ownership.
- Energy efficiency enables off-grid network deployments that rely on medium-sized solar panels as power supplies, thereby enabling wireless connectivity to reach even the most remote areas.
- Energy efficiency is essential to realizing operators’ ambition of providing wireless access in a sustainable and more resource-efficient way.

The importance of these factors will increase further in the 5G era, and energy efficiency will therefore be an important requirement in the design of 5G wireless access.

MACHINE-TYPE COMMUNICATION

Fundamentally, applications such as mobile telephony, mobile broadband and media delivery are about information for humans. In contrast, many of the new applications and use cases that drive the requirements and capabilities of 5G are about end-to-end communication between machines. To distinguish them from the more human-centric wireless-communication use cases, these applications are often termed machine-type communication (MTC).

Although spanning a wide range of applications, MTC applications can be divided into two main categories – massive MTC and critical MTC – depending on their characteristics and requirements.

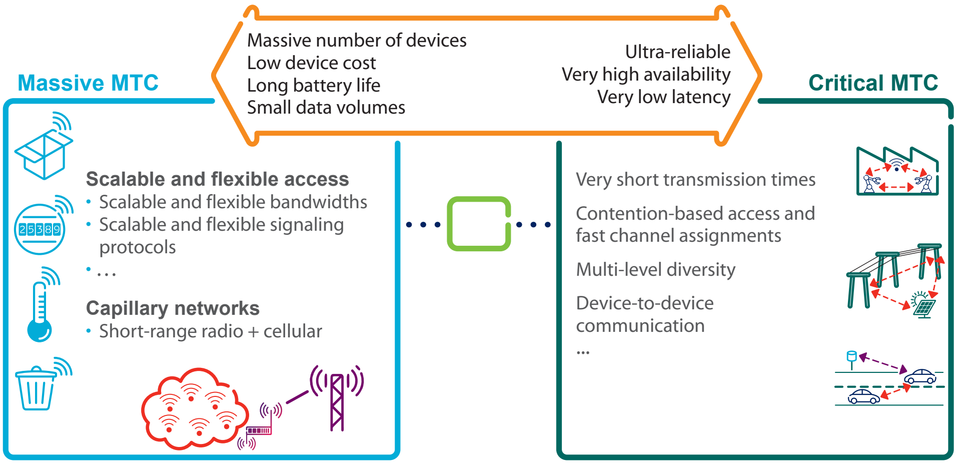
Massive number of devices
Low device cost
Long battery life
Small data volumes
Ultra-reliable
Very high availability
Very low latency
Massive MTC
Critical MTC
25389
Scalable and flexible access
• Scalable and flexible bandwidths
• Scalable and flexible signaling protocols
• …
Capillary networks
• Short-range radio + cellular
Very short transmission times
Contention-based access and fast channel assignments
Multi-level diversity
Device-to-device communication
…

Figure 2: Massive MTC and critical MTC.

Massive MTC refers to services that typically span a very large numbers of devices, usually sensors and actuators. Sensors are extremely low cost and consume very low amounts of energy in order to sustain long battery life. Clearly, the amount of data generated by each sensor is normally very small, and very low latency is not a critical requirement. While actuators are similarly limited in cost, they will likely have varying energy footprints ranging from very low to moderate energy consumption.

Sometimes, the mobile network may be used to bridge connectivity to the device by means of capillary networks. Here, local connectivity is provided by means of a short-range radio access technology, for example Wi-Fi, Bluetooth [3] or 802.15.4/6LoWPAN [4]. Wireless connectivity beyond the local area is then provided by the mobile network via a gateway node.

Critical MTC refers to applications such as traffic safety/control, control of critical infrastructure and wireless connectivity for industrial processes. Such applications require very high reliability and availability in terms of wireless connectivity, as well as very low latency. On the other hand, low device cost and energy consumption is not as critical as for massive MTC applications. While the average volume of data transported to and from devices may not be large, wide instantaneous bandwidths are useful in being able to meet capacity and latency requirements.

There is much to gain from a network being able to handle as many different applications as possible, including mobile broadband, media delivery and a wide range of MTC applications by means of the same basic wireless-access technology and within the same spectrum. This avoids spectrum fragmentation and allows operators to offer support for new MTC services for which the business potential is inherently uncertain, without having to deploy a separate network and reassign spectrum specifically for these applications.

SPECTRUM FOR 5G

In order to support increased traffic capacity and to enable the transmission bandwidths needed to support very high data rates, 5G will extend the range of frequencies used for mobile communication. This includes new spectrum below 6GHz, as well as spectrum in higher frequency bands.

Specific candidate spectrum for mobile communication in higher frequency bands is yet to be identified by the ITU-R or by individual regulatory bodies. The mobile industry remains agnostic about particular choices, and the entire frequency range up to approximately 100GHz is under consideration at this stage, although there is significant interest in large contiguous allocations that can provide dedicated and licensed spectrum for use by multiple competing network providers.

The lower part of this frequency range, below 30GHz, is preferred from the point of view of propagation properties. At the same time, very large amounts of spectrum and the possibility of wide transmission frequency bands of the order of 1GHz or more are more likely above 30GHz.

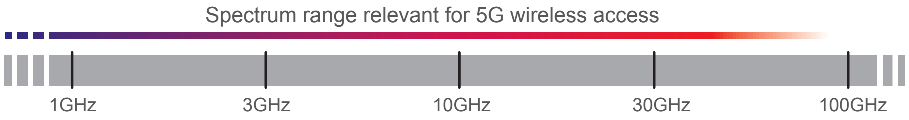

Figure 3: Spectrum relevant for 5G wireless access.

Spectrum relevant for 5G wireless access therefore ranges from below 1GHz up to approximately 100GHz, as Figure 3 shows.

It is important to understand that high frequencies, especially those above 10GHz, can only serve as a complement to lower frequency bands, and will mainly provide additional system capacity and very wide transmission bandwidths for extreme data rates in dense deployments. Spectrum allocations at lower bands will remain the backbone for mobile-communication networks in the 5G era, providing ubiquitous wide-area connectivity.

The World Radio Conference (WRC)-15 discussions have resulted in an agreement to include an agenda item for IMT-2020, the designated ITU-R qualifier for 5G, in WRC-19. The conference also reached agreement on a set of bands that will be studied for 5G, with direct applicability to NX. Many of the proposed bands are in the millimeter wave region and include:

- 24.25GHz to 27.5GHz, 37GHz to 40.5GHz, 42.5GHz to 43.5GHz, 45.5GHz to 47GHz, 47.2GHz to 50.2GHz, 50.4GHz to 52.6GHz, 66 GHz to 76GHz and 81GHz to 86GHz, which have allocations to the mobile service on a primary basis; and
- 31.8GHz to 33.4GHz, 40.5GHz to 42.5GHz and 47GHz to 47.2GHz, which may require additional allocations to the mobile service on a primary basis.

The mobile industry will strive to gain access to spectrum in the 6GHz to 20GHz range, but the policy directions being followed by regulators seem to be focused on frequency bands above 30GHz. In the US, the FCC has issued two Notices of Public Rule Making (NPRM) on bands above 24GHz. Ofcom has likewise indicated a preference for bands above 30GHz within the mobile industry.

The capacity needs of the mobile industry will continue to be served by licensed spectrum, although novel sharing arrangements for spectrum will become progressively more important as restricted opportunities for new spectrum start to impact incumbent services such as satellite communication and radio location. Two examples of sharing arrangements include LSA planned in Europe for the 2.3GHz band and the Citizens Band Radio Service for 3.5GHz in the US.

5G TECHNOLOGY COMPONENTS

Beyond extending operation to higher frequencies, there are several other key technology components relevant for the evolution to 5G wireless access. These components include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

ACCESS/BACKHAUL INTEGRATION

Wireless technology is already frequently used as part of the backhaul solution. Such wireless-backhaul solutions typically operate under line-of-sight conditions using proprietary radio technology in higher frequency bands, including the millimeter wave (mmW) band.

In the future, the access (base-station-to-device) link will also extend to higher frequencies. Furthermore, to support dense low-power deployments, wireless backhaul will have to extend to cover non-line-of-sight conditions, similar to access links.

In the 5G era, the wireless-access link and wireless backhaul should not therefore be seen as two separate entities with separate technical solutions. Rather, backhaul and access should be seen as an integrated wireless-access solution able to use the same basic technology and operate using a common spectrum pool. This will lead to more efficient overall spectrum utilization as well as reduced operation and management effort.

DIRECT DEVICE-TO-DEVICE COMMUNICATION

The possibility of limited direct device-to-device (D2D) communication has recently been introduced as an extension to the LTE specifications. In the 5G era, support for D2D as part of the overall wireless-access solution should be considered from the start. This includes peer-to-peer user-data communication directly between devices, but also, for example, the use of mobile devices as relays to extend network coverage.

D2D communication in the context of 5G should be an integral part of the overall wireless-access solution, rather than a stand-alone solution. Direct D2D communication can be used to offload traffic, extend capabilities and enhance the overall efficiency of the wireless-access network. Furthermore, in order to avoid uncontrolled interference to other links, direct D2D communication should be under network control. This is especially important for the case of D2D communication in licensed spectrum.

FLEXIBLE DUPLEX

Frequency Division Duplex (FDD) has been the dominating duplex arrangement since the beginning of the mobile communication era. In the 5G era, FDD will remain the main duplex scheme for lower frequency bands. However, for higher frequency bands – especially above 10GHz – targeting very dense deployments, Time Division Duplex (TDD) will play a more important role.

In very dense deployments with low-power nodes, the TDD-specific interference scenarios (direct base-station-to-base-station and device-to-device interference) will be similar to the 'normal' base-station-to-device and device-to-base-station interference that also occurs for FDD.

Furthermore, for the dynamic traffic variations expected in very dense deployments, the ability to dynamically assign transmission resources (time slots) to different transmission directions may allow more efficient utilization of the available spectrum.

To reach its full potential, 5G should therefore allow for very flexible and dynamic assignment of TDD transmission resources. This is in contrast to current TDD-based mobile technologies, including TD-LTE, for which there are restrictions on the downlink/uplink configurations, and for which there typically exist assumptions about the same configuration for neighbor cells and also between neighbor operators.

FLEXIBLE SPECTRUM USAGE

Since its inception, mobile communication has relied on spectrum licensed on a per-operator basis within a geographical area. This will remain the foundation for mobile communication in the 5G era, allowing operators to provide high-quality connectivity in a controlled-interference environment.

However, per-operator licensing of spectrum will be complemented with the possibility to share spectrum. Such sharing may be between a limited set of operators, or may occur in license-exempt scenarios. The

Citizens Band Radio Service in the US in the 3.5GHz band and the 5GHz unlicensed spectrum are examples of managed and unlicensed sharing regimes respectively.

New air interfaces like NX will likely be well served by more conventional licensed allocations of spectrum, mainly due to the need to establish a basic foundation for the technology to operate in an independent manner while interoperability is established with technologies like LTE. At some point, further allocations of spectrum for 5G may leverage the mobile industry's experience of sharing approaches in lower cellular bands.

MULTI-ANTENNA TRANSMISSION

Multi-antenna transmission already plays an important role in current generations of mobile communication and will be even more central in the 5G era, due to the physical limitations of small antennas. Path loss between a transmitter and receiver does not change as a function of frequency, as long as the effective aperture of the transmitting and receiving antennas does not change. The antenna aperture does reduce in proportion to the square of the frequency, and that reduction can be compensated by the use of higher antenna directivity. The 5G radio will employ hundreds of antenna elements to increase antenna aperture beyond what may be possible with current cellular technology.

In addition, the transmitter and receiver will use beamforming to track one another and improve energy transfer over an instantaneously configured link. Beamforming will also improve the radio environment by limiting interference to small fractions of the entire space around a transmitter and likewise limiting the impact of interference on a receiver to infrequent stochastic events. The use of beamforming will also be an important technology for lower frequencies; for example, to extend coverage and to provide higher data rates in sparse deployments.

ULTRA-LEAN DESIGN

Ultra-lean radio-access design is important to achieve high efficiency in 5G networks. The basic principle of ultra-lean design can be expressed as: minimize any transmissions not directly related to the delivery of user data. Such transmissions include signals for synchronization, network acquisition and channel estimation, as well as the broadcast of different types of system and control information.

Ultra-lean design is especially important for dense deployments with a large number of network nodes and highly variable traffic conditions. However, lean transmission is beneficial for all kinds of deployments, including macro deployments.

By enabling network nodes to enter low-energy states rapidly when there is no user-data transmission, ultra-lean design is an important component in delivering high network energy performance. Ultra-lean design will also enable higher achievable data rates by reducing interference from non-user-data-related transmissions.

USER/CONTROL SEPARATION

Another important design principle for 5G is to decouple user data and system control functionality. The latter includes the provisioning of system information; that is, the information and procedures needed for a device to access the system.

Such a decoupling will allow separate scaling of user-plane capacity and basic system control functionality. For example, user data may be delivered by a dense layer of access nodes, while system information is only provided via an overlaid macro layer on which a device also initially accesses the system.

It should be possible to extend the separation of user data delivery and system control functionality over multiple frequency bands and RATs. As an example, the system control functionality for a dense layer based on new high-frequency radio access could be provided by means of an overlaid LTE layer.

User/control separation is also an important component for future radio-access deployments relying heavily on beamforming for user data delivery. Combining ultra-lean design with a logical separation of user-plane data delivery and basic system connectivity functionality will enable a much higher degree of device-centric network optimization of the active radio links in the network. Since only the ultra-lean signals related to the system control plane need to be static, it is possible to design a system where almost everything can be dynamically optimized in real time.

An ultra-lean design combined with a system control plane logically separated from the user data delivery function also provides higher flexibility in terms of evolution of the RAT as, with such separation, the user plane can evolve while retaining system control functionality.

CONCLUSION

5G is the next step in the evolution of mobile communication and will be a key component of the Networked Society. In particular, 5G will accelerate the development of the Internet of Things. To enable connectivity for a wide range of applications and use cases, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communications.

These capabilities include very high achievable data rates, very low latency and ultra-high reliability. Furthermore, 5G wireless access needs to support a massive increase in traffic in an affordable and sustainable way, implying a need for a dramatic reduction in the cost and energy consumption per delivered bit.

5G wireless access will be realized by the evolution of LTE for existing spectrum in combination with new radio access technologies that primarily target new spectrum. Key technology components of 5G wireless access include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

REFERENCES

[1] ICT-317669 METIS project, Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations, Deliverable D1.5, April 2015, available at:
https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D1.5_v1.pdf

[2] Ericsson, Ericsson Mobility Report, November 2015, available at:
http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf

[3] Bluetooth Special Interest Group, accessed March 2016, available at:
http://bluetooth.org

[4] IETF, IPv6 over Low power WPAN (6lowpan), accessed March 2016, available at:
http://datatracker.ietf.org/wg/6lowpan

GLOSSARY

D2D	device-to-device
FDD	Frequency Division Duplex
mmW	millimeter wave
MTC	Machine-Type Communication
NPRM	Notices of Public Rule Making
OFDM	Orthogonal Frequency Division Multiplexing
TDD	Time Division Duplex
WRC	World Radio Conference

© 2016 Ericsson AB – All rights reserved

ERICSSON WHITE PAPER
Uen 284 23-3278 | January 2016

CELLULAR NETWORKS FOR MASSIVE IOT

ENABLING LOW POWER WIDE AREA APPLICATIONS

With new standards specifically targeting the connectivity requirements of Massive Internet of Things (IoT) applications, cellular networks can deliver reliable, secure and diverse IoT services using existing network infrastructure.

DELIVERING NEW VALUE IN THE NETWORKED SOCIETY

In the future, all devices that benefit from an internet connection will be connected. In this Networked Society, every person and every industry will be empowered to reach their full potential. Internet of Things (IoT) technology is a key enabler of this vision by delivering machine-to-machine (M2M) and machine-to-person communications on a massive scale.

As shown in Figure 1, Ericsson predicts there will be around 28 billion connected devices by 2021, of which more than 15 billion will be connected M2M and consumer-electronics devices [1]. A large share of these will be applications served by short-range radio technologies such as Wi-Fi and Bluetooth, while a significant proportion will be enabled by wide area networks (WANs) that are primarily facilitated by cellular networks.

THE NEW IOT LANDSCAPE

The IoT revolution offers huge potential value in terms of improved efficiency, sustainability and safety for industry and society. Analysts predict that the total added value of the IoT will be USD 1.9 trillion by 2020 [2].

The variety of applications and solutions designed for individuals, business and industry is spurring the rapid expansion of the IoT market. The IoT is playing a major role across a variety of vertical sectors, generating cost savings, new revenue streams and other benefits.

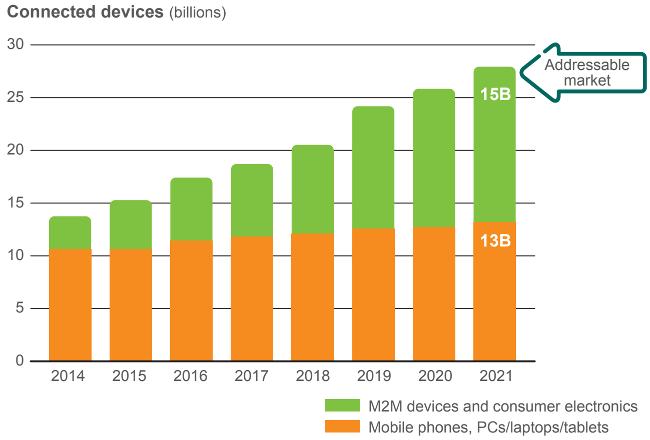

Figure 1: Growth in connected devices.

Each IoT application needs a clear value proposition and business logic in line with the prevailing ecosystem, business models and value chains of the various stakeholders. For all applications, solutions need to be integrated on platforms that can scale and handle millions of devices efficiently. Business processes for administration, provisioning and charging will have to be streamlined to minimize costs and enhance the business case.

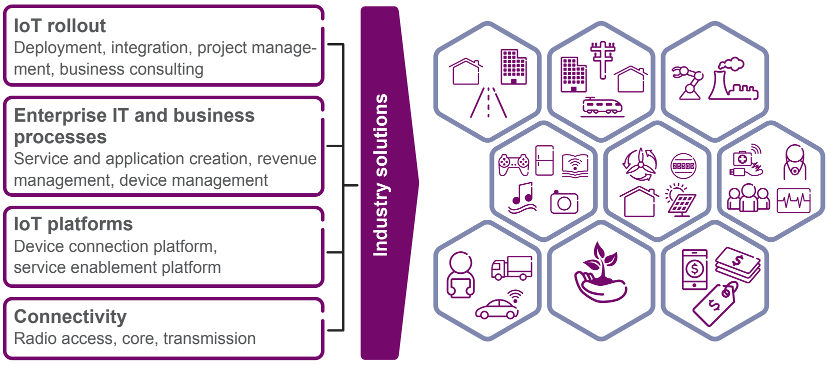
IoT rollout
Deployment, integration, project management, business consulting
Enterprise IT and business processes
Service and application creation, revenue management, device management
IoT platforms
Device connection platform, service enablement platform
Connectivity
Radio access, core, transmission
Industry solutions

Figure 2: The new IoT landscape.

As they are largely responsible for wireless connectivity on a global scale, operators are in an excellent position to capture a share of the added value generated by the emerging IoT market. The size of this share will depend on the role that operators adopt in the value chain. This could range from being a straightforward connectivity provider (monetizing connectivity in new ways), all the way to being an end-to-end solution provider of turnkey solutions to vertical markets [3].

DIFFERENT IOT CONNECTIVITY ALTERNATIVES

Connectivity is the foundation for IoT, and the type of access required will depend on the nature of the application. Many IoT devices will be served by radio technologies that operate on unlicensed spectrum and that are designed for short-range connectivity with limited QoS and security requirements typically applicable for a home or indoor environment. Currently, there are two alternative connectivity tracks for the many IoT applications that depend on wide-area coverage:

Cellular technologies: 3GPP technologies like GSM, WCDMA, LTE and future 5G. These WANs operate on licensed spectrum and historically have primarily targeted high-quality mobile voice and data services. Now, however, they are being rapidly evolved with new functionality and the new radio access technology narrowband IoT (NB-IoT) specifically tailored to form an attractive solution for emerging low power wide area (LPWA) applications.

Unlicensed LPWA: new proprietary radio technologies, provided by, for example, SIGFOX and LoRa, have been developed and designed solely for machine-type communication (MTC) applications addressing the ultra-low-end sensor segment, with very limited demands on throughput, reliability or QoS.

One way to segment IoT applications is to categorize them according to coverage needs and performance requirements (such as data speed or latency demands).

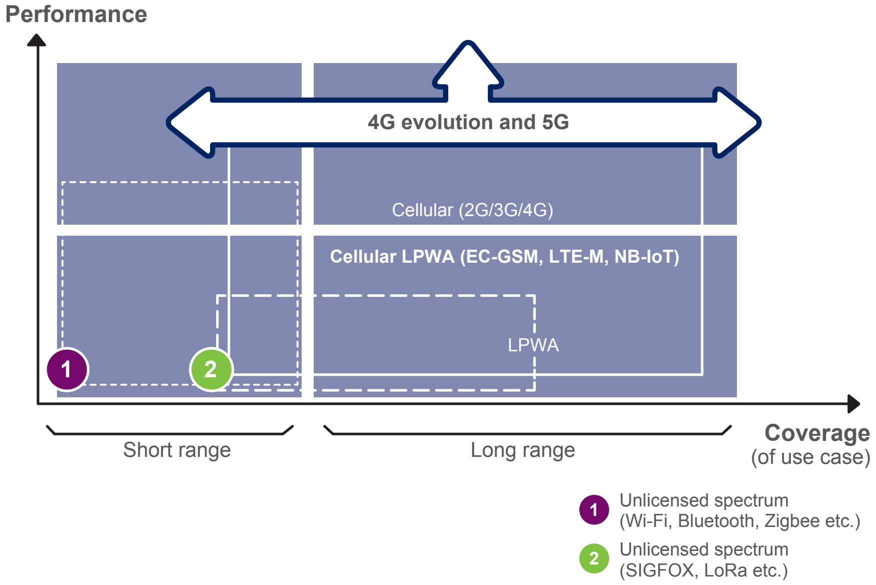

Figure 3: Technologies addressing different IoT segments.

The coverage needs of a particular use case may be highly localized (such as a stationary installation within a building), while other use cases require global service coverage (such as container tracking). 3GPP technologies already dominate use cases with large geographic coverage needs and medium- to high-performance requirements.

With new feature sets specifically tailored for LPWA IoT applications, 3GPP technologies are taking a large leap forward to cover segments with low-cost, low-performance requirements too.

CAPILLARY NETWORKS COMBINING CELLULAR AND UNLICENSED STRENGTHS

Even when existing 3GPP end-to-end connectivity is not feasible, cellular technology can still provide key benefits when used as a bridging option, i.e. as an aggregation and routing solution. This capillary network approach allows end devices to utilize varying access solutions from either the short range or LPWA domain and access the cellular networks via a gateway device. Capillary networks enable the reuse of cellular functions and assets such as security, device management, billing and QoS without requiring each end device to be cellular-enabled.

A WIDE RANGE OF IOT REQUIREMENTS

There will be a wide range of IoT use cases in the future, and the market is now expanding toward both Massive IoT deployment as well as more advanced solutions that may be categorized as Critical IoT, as shown in Figure 4.

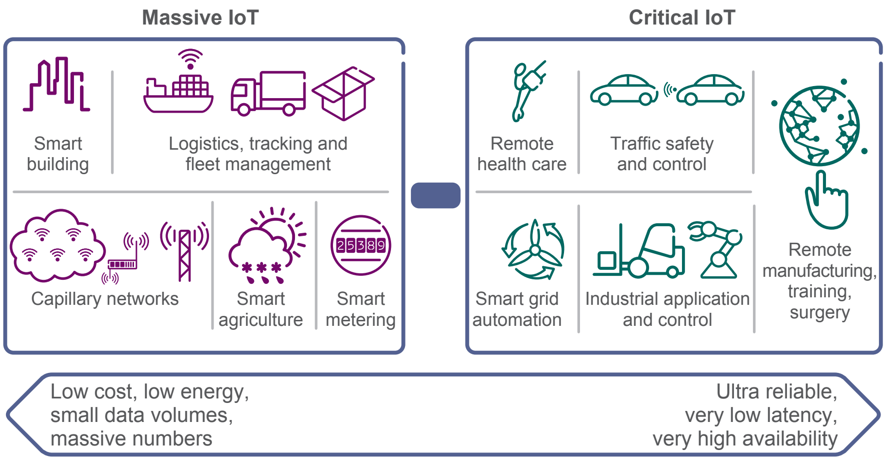
Massive IoT
Critical IoT
Smart building
Logistics, tracking and fleet management
25389
Capillary networks
Smart agriculture
Smart metering
Remote health care
Traffic safety and control
Remote manufacturing, training, surgery
Smart grid automation
Industrial application and control
Low cost, low energy, small data volumes, massive numbers
Ultra reliable, very low latency, very high availability

Figure 4: Differing requirements for Massive and Critical IoT applications.

At one end of the scale, in Massive IoT applications – typically sensors that report to the cloud on a regular basis – the end-to-end cost must be low enough for the business case to make sense. Here, the requirement is for low-cost devices with low energy consumption and good coverage.

At the other end of the scale, Critical IoT applications will have very high demands for reliability, availability and low latency. These use cases are enabled by LTE or 5G capabilities. Here, the volumes are typically much smaller, but the business value is significantly higher.

There are, however, many other use cases between the two extremes, which today rely on 2G, 3G or 4G connectivity.

MASSIVE IOT USE CASE DIVERSITY

The Massive IoT market segment includes several applications widely used in industries and societies, as shown in Figure 5.

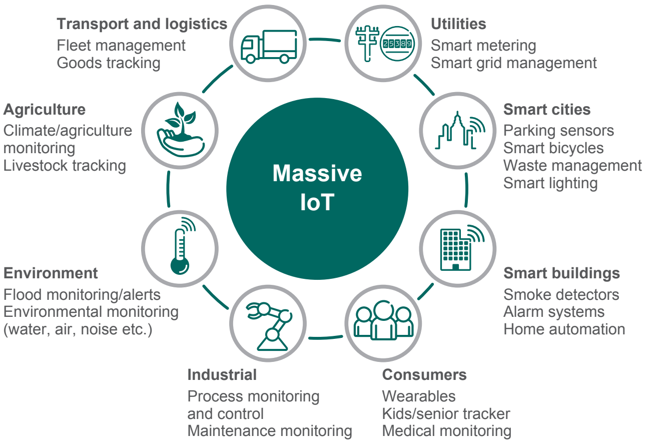
Transport and logistics
Fleet management
Goods tracking
Utilities
Smart metering
Smart grid management
Agriculture
Climate/agriculture monitoring
Livestock tracking
Smart cities
Parking sensors
Smart bicycles
Waste management
Smart lighting
Massive IoT
Environment
Flood monitoring/alerts
Environmental monitoring (water, air, noise etc.)
Smart buildings
Smoke detectors
Alarm systems
Home automation
Industrial
Process monitoring and control
Maintenance monitoring
Consumers
Wearables
Kids/senior tracker
Medical monitoring

Figure 5: Industry and society applications enabled by LPWA.

The potential applications for the IoT run into the millions, with a huge variety of requirements regarding cost, battery life, coverage, connectivity performance (throughput and capacity), security and reliability. Figure 6 illustrates some such applications and their requirements regarding devices and connectivity.

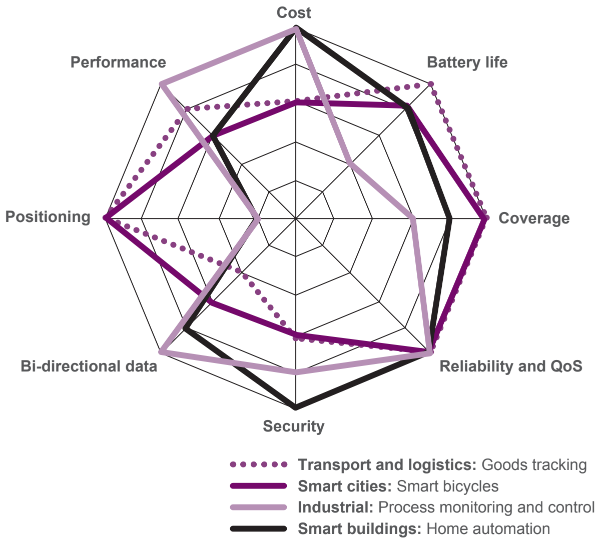

Figure 6: Device and connectivity requirements for sample IoT use cases.

Some devices will only send a few messages per day – such as status indicators for temperature – while others may need to transmit a video stream to guide a remote repair technician, for example. The difference in throughput requirements is huge. If operators or service providers handle several applications, it may be of great benefit to be able to harmonize communication modules, so that they all use the same underlying radio solution to reduce operational and fault management effort and complexity.

Many higher-value applications will require two-way communications – in other words, an uplink as well as a downlink – to enable monitoring and control of devices in systems like heating, ventilation and cooling plants. The long lifetime of many IoT applications makes it invaluable to be able to perform over-the-air device updates for new functionality or parameter settings. The amount of data sent for such updates can often be more demanding for the network than the monitoring or control application itself.

Relatively simple uplink-focused applications can benefit greatly from a bi-directional link to provide robustness. For example, a connected smoke detector must deliver a smoke alarm with absolute certainty. The ability of a network to provide acknowledgements of a received message enables better fault management and the required level of reliability. Positioning can be used to locate the sensor at the point it failed and simplify operations. For tracking applications, location information is essential.

In applications like building security, sensitive information could be reported over the air, which will require strict security. Furthermore, in the case of a break-in, it is crucial that the alarm information reaches the control center in time – making QoS and two-way communication vital.

KEY CHALLENGES FOR MASSIVE IOT

The key challenges to enabling large-scale uptake of Massive IoT include:

Device cost – clearly a key enabler for high-volume, mass-market applications, enabling many of the use cases.

Battery life – many IoT devices will be battery-powered, and often the cost of replacing batteries in the field is not viable.

Coverage – deep indoor connectivity is a requirement for many applications in the utility area. Furthermore, regional (or even national or global) coverage is a prerequisite for many use cases, especially within the transport area.

Scalability – in order to enable a Massive IoT market, networks need to scale efficiently. The initial investment required for supporting a limited number of devices has to be manageable, while on the other hand, the network capacity must be easy to scale to handle thousands – or millions – of devices.

Diversity – connectivity should be able to support diverse requirements from different use cases. One network supporting everything from simple static sensors to tracking services, to applications requiring higher throughput and lower latency is essential in terms of total cost of ownership (TCO).

THE ADVANTAGES OF CELLULAR TECHNOLOGIES

Each of the technologies available for IoT connectivity has its own advantages and disadvantages. However, the range of IoT connectivity requirements – both technical and commercial – means cellular technologies can provide clear benefits across a wide variety of applications, as summarized in Figure 7.

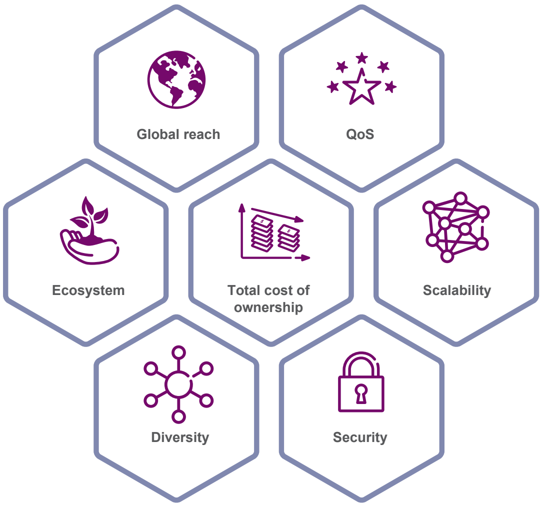

Figure 7: Advantages of cellular technologies.

In terms of global reach, cellular networks already cover 90 percent of the world's population. WCDMA and LTE are catching up, but GSM will offer superior coverage in many markets for years to come. Cellular networks have been developed and deployed over three decades, and they will be around for the foreseeable future.

The cellular mobile industry represents a huge and mature ecosystem, incorporating chipset, device and network equipment vendors, operators, application providers and many others. The global cellular ecosystem is governed by the 3GPP standardization forum, which guarantees broad industry support for future development.

When it comes to scalability, cellular networks are built to handle massive volumes of mobile broadband traffic; the traffic from most IoT applications will be relatively small and easily absorbed. Operators are able to offer connectivity for IoT applications from the start-up phase and grow this business with low TCO and only limited additional investment and effort. Operation in licensed spectrum also provides predictable and controlled interference, which enables efficient use of the spectrum to support massive volumes of devices.

Cellular connectivity offers the diversity to serve a wide range of applications with varying requirements within one network. While competing unlicensed LPWA technologies are designed solely for very low-end MTC applications, cellular networks can address everything from Massive to Critical IoT use cases.

QoS mechanisms will be essential for many IoT applications. Cellular systems have mature QoS functionality, and this enables critical MTC applications to be handled together with traffic from sensors, voice and mobile-broadband traffic on the same carrier. QoS, along with licensed spectrum as described above, provides a foundation for long-term Service Level Agreements with a specific grade of service.

Traditionally, the security mechanisms of cellular networks have been based on a physical SIM attached to the device, referred to as a Universal Integrated Circuit Card (UICC). This has also enabled roaming between operators, which has been one of the main factors behind the huge success of mobile networks. The SIM will also be essential in future IoT applications, with SIM functionality embedded in the chipset (eUICC) or handled as a soft-SIM solution running in a trusted run-time environment of the module.

With a straightforward rollout of new software, cellular networks will be able to support the full breadth of applications, ranging from low-end use cases in the LPWA segment, to the high-end segments of in-car entertainment and video surveillance. One network connecting the whole diversifying IoT market will guarantee the lowest possible TCO as well as fast time to market.

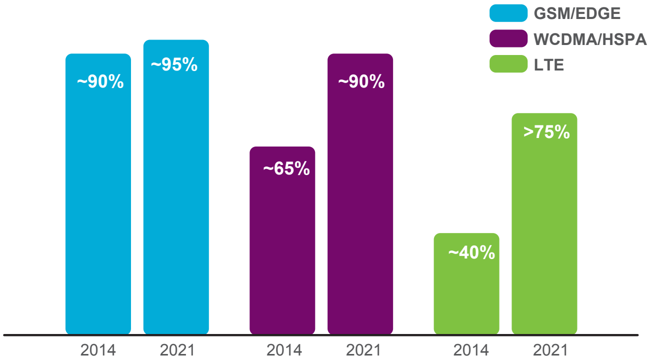

Figure 8: World population coverage by cellular 3GPP technology.

EVOLVING STANDARDS

To meet the new connectivity requirements of the emerging Massive IoT segment, 3GPP has taken evolutionary steps on both the network side and the device side.

The key improvement areas addressed in 3GPP up to Release 13 are:

- Lower device cost – cutting module cost for LTE devices by reducing peak rate, memory requirement and device complexity. The LTE module cost-reduction evolution started in Release 8 with the introduction of LTE for machine-type communication (LTE-M) Cat 1 devices with reduced peak rate to a maximum of 10Mbps, and continued in Releases 12 and 13 with reduced device complexity for lower performance and using less bandwidth or a narrowband IoT carrier to cut costs further.
- Improved battery life – more than 10 years of battery life can be achieved by introducing Power Saving Mode and/or extended discontinuous reception (eDRX) functionality. These features allow the device to contact the network – or to be contacted – on a per-need basis, meaning that it can stay in sleep mode for minutes, hours or even days.
- Improved coverage – an improvement of 15dB on LTE-M and of 20dB on NB-IoT and GSM, which translates into a seven-fold increase in the outdoor coverage area and significantly improved indoor signal penetration to reach deep indoors. This supports many IoT devices like smart meters, which are often placed in a basement.
- Support for massive numbers of IoT connections – specifically, one LTE cell site can support millions of IoT devices, depending on the use case. Core network enhancements include software upgrades for service differentiation handling, signaling optimization and high-capacity platforms (more than 30 million devices per node).

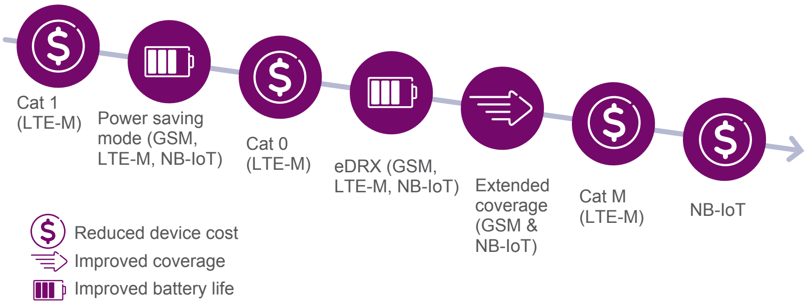

Figure 9: 3GPP evolution steps for Massive IoT.

A FULL RANGE OF CELLULAR LPWA SOLUTIONS

No single technology or solution is ideally suited to all the different potential Massive IoT applications, market situations and spectrum availability. As a result, the mobile industry is standardizing several LPWA technologies, including Extended Coverage GSM (EC-GSM), LTE-M and NB-IoT.

LTE-M, NB-IoT and EC-GSM are all superior solutions to meet Massive IoT requirements as a family of solutions, and can complement each other based on technology availability, use case requirements and deployment scenarios. LTE-M consisting of Cat 1, Cat 0 and Cat M supports a wide range of IoT applications, including those that are content-rich; NB-IoT covers ultra-low-end IoT applications with a cost and coverage advantage over LTE-M; and EC-GSM serves IoT services for all GSM markets.

For example, a smart city application such as waste management may use EC-GSM technology to provide LPWA connectivity in markets where it can be deployed on existing 2G networks; NB-IoT technology may be used for water-metering applications, which have some of the most extreme coverage requirements in underground locations. On the other hand, asset-tracking applications that can support a relatively high number of messages triggered by certain events may employ LTE-M.

EC-GSM – GLOBAL CELLULAR IOT FOR ALL GSM MARKETS

GSM is still the dominant mobile technology in many markets, and the vast majority of cellular M2M applications today use GPRS/EDGE for connectivity. GSM is likely to continue playing a key role in the IoT well into the future, due to its global coverage footprint, time to market and cost advantages.

Recognizing this – and identifying the requirements for Massive IoT discussed earlier in this paper – an initiative was undertaken in 3GPP Release 13 to further improve GSM.

The resulting EC-GSM functionality enables coverage improvements of up to 20dB with respect to GPRS on the 900MHz band.

This coverage extension is achieved for both the data and control planes by utilizing the concept of repetitions and signal combining techniques. It is handled in a dynamic manner with multiple coverage classes to ensure optimal balance between coverage and performance.

EC-GSM is achieved by defining new control and data channels mapped over legacy GSM. It allows multiplexing of new EC-GSM devices and traffic with legacy EDGE and GPRS. No new network carriers are required: new software on existing GSM networks is sufficient and provides combined capacity of up to 50,000 devices per cell on a single transceiver.

Initially part of EC-GSM but now a separate 3GPP item, eDRX improves the power efficiency – and therefore the battery life – for many use cases. eDRX improves the idle mode behavior by allowing the use of a number of inactivity timers, where the device can choose to tune in to the network and listen for downlink pages and traffic.

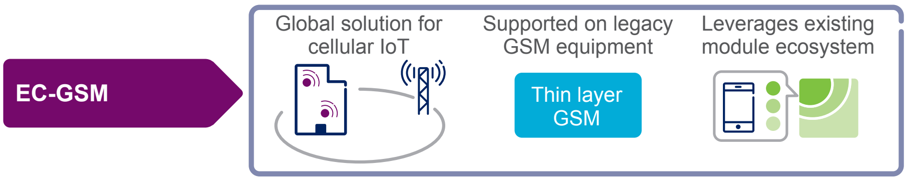
EC-GSM
Global solution for cellular IoT
Supported on legacy GSM equipment
Thin layer GSM
Leverages existing module ecosystem

Figure 10: EC-GSM.

EC-GSM extends the data handling and power efficiency advantages that GSM/GPRS technology already offers for MTC, and it will help operators extend the service life of their huge 2G legacy base.

LTE-M – SUPPORTING A WIDE RANGE OF MASSIVE IOT USE CASES

LTE is the leading mobile broadband technology and its coverage is expanding rapidly. So far, the focus has been on meeting the huge demand for mobile data with highly capable devices that utilize new spectrum. With features like Carrier Aggregation, MIMO and Lean Carrier, the gigabit per second performance for LTE cell throughput is now reaching levels that result in an excellent mobile broadband user experience.

The advent of LTE-M signifies an important step in addressing MTC capabilities over LTE. LTE-M brings new power-saving functionality suitable for serving a variety of IoT applications; Power Saving Mode and eDRX extend battery life for LTE-M to 10 years or more. LTE-M traffic is multiplexed over a full LTE carrier, and it is therefore able to tap into the full capacity of LTE. Additionally, new functionality for substantially reduced device cost and extended coverage for LTE-M are also specified within 3GPP.

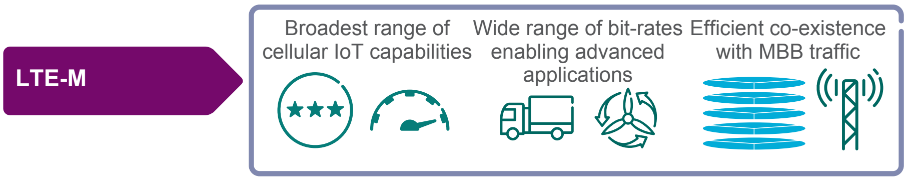
LTE-M
Broadest range of cellular IoT capabilities
Wide range of bit-rates enabling advanced applications
Efficient co-existence with MBB traffic

Figure 11: LTE M.

NB-IOT – SUPPORTING ULTRA-LOW-END MASSIVE IOT APPLICATIONS

In addition to LTE-M, NB-IoT technology is being standardized in time for 3GPP Release 13. NB-IoT is a self-contained carrier that can be deployed with a system bandwidth of only 200kHz, and is specifically tailored for ultra-low-end IoT applications. It is enabled using new network software on an existing LTE network, which will result in rapid time to market.

NB-IoT
Scalable ultra-low-end cellular IoT solution
Ultra-low bit rates and extreme coverage
789
Native narrowband solution
In-band
LTE
200kHz

Figure 12: NB-IoT.

NB-IoT provides lean setup procedures, and a capacity evaluation indicates that each 200kHz NB-IoT carrier can support more than 200,000 subscribers. The solution can easily be scaled up by adding multiple NB-IoT carriers when needed. NB-IoT also comes with an extended coverage of up to 20dB, and battery saving features, Power Saving Mode and eDRX for more than 10 years of battery life.

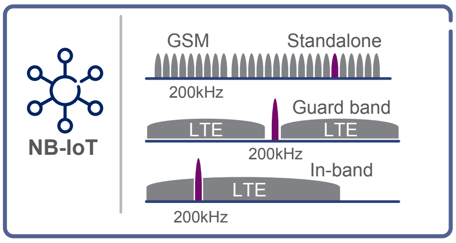
NB-IoT
GSM
Standalone
200kHz
Guard band
LTE
LTE
200kHz
In-band
LTE
200kHz

Figure 13: NB-IoT deployment.

NB-IoT is designed to be tightly integrated and interwork with LTE, which provides great deployment flexibility. The NB-IoT carrier can be deployed in the LTE guard band, embedded within a normal LTE carrier, or as a standalone carrier in, for example, GSM bands.

- Standalone deployment in a GSM low band: this is an option when LTE is deployed in a higher band and GSM is still in use, providing coverage for basic services.
- Guard band deployment, typically next to an LTE carrier: NB-IoT is designed to enable deployment in the guard band immediately adjacent to an LTE carrier, without affecting the capacity of the LTE carrier. This is particularly suitable for spectrum allocations that do not match the set of LTE system bandwidths, leaving gaps of unused spectrum next to the LTE carrier.
- Efficient in-band deployment, allowing flexible assignment of resources between LTE and NB-IoT: it will be possible for an NB-IoT carrier to time-share a resource with an existing LTE carrier. The in-band deployment also allows for highly flexible migration scenarios. For example, if the NB-IoT service is first deployed as a standalone deployment in a GSM band, it can subsequently be migrated to an in-band deployment if the GSM spectrum is re-farmed to LTE, thereby avoiding any fragmentation of the LTE carrier.

NB-IoT reduces device complexity below that of LTE-M with the potential to rival module costs of unlicensed LPWA technologies, and it will be ideal for addressing ultra-low-end applications in markets with a mature LTE installed base.

CONCLUSION

Uptake of Massive IoT is set to take off, and operators have a unique opportunity to drive the implementation of new IoT applications by offering affordable connectivity on a global scale.

For IoT applications, existing cellular networks offer distinct advantages over alternative WAN technologies, such as unlicensed LPWA. The global reach, QoS, ecosystem, TCO, scalability, diversity and security of cellular networks are all vital factors that can support the fast uptake and success of IoT. Enabled by new software in existing legacy networks, cellular networks can support a diverse range of IoT applications – ensuring the lowest possible TCO.

3GPP standardization work for GSM and LTE, and the recent addition of NB-IoT, is further improving the ability of cellular networks to address the Massive IoT market, where ultra-low end-to-end cost is a prerequisite.

GSM/GPRS, which already serves the majority of cellular-based MTC applications, is evolving with a new EC-GSM standard, which delivers significantly better energy efficiency and increased coverage. EC-GSM enhancements will cement GSM's position as a highly relevant connectivity platform for low-end, Massive IoT applications globally.

New downsized NB-IoT and LTE-M chipsets, designed for MTC, and features that improve both coverage and device battery life will boost the ability of LTE infrastructure to address the IoT market. One network that supports all applications – from advanced mobile broadband services, VoIP and all kinds of low- to high-end IoT use cases – creates a very strong value proposition.

Whether operators choose the GSM, NB-IoT or LTE-M track – or a combination of these – will depend on several factors such as technology coverage, future technology strategies and targeted market segments. Whichever path they take, they have a huge opportunity to benefit from the emerging IoT revolution. Operators can choose to continue offering telecom-grade connectivity as they do today, or they can evolve to become a platform or fully-fledged IoT service provider targeting a larger slice of future IoT revenues.

GLOSSARY

EC-GSM	Extended Coverage GSM
eDRX	extended discontinuous reception
eUICC	embedded Universal Integrated Circuit Card
IoT	Internet of Things
LPWA	low power wide area
LTE-M	LTE for machine-type communication
M2M	machine-to-machine
MIMO	multiple-input, multiple-output
MTC	machine-type communication
NB-IoT	narrowband IoT
TCO	total cost of ownership
UICC	Universal Integrated Circuit Card
WAN	wide area network

REFERENCES

[1] Ericsson, Ericsson Mobility Report, November 2015, available at: http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf

[2] Gartner, Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020, December 2013, available at: http://www.gartner.com/newsroom/id/2636073

[3] Ericsson, White Paper: Machine-to-Machine – exploring potential operator roles, September 2014, available at: http://www.ericsson.com/res/docs/whitepapers/wp-m2m.pdf

© 2016 Ericsson AB – All rights reserved

THE Next Wave

The National Security Agency's review of emerging technologies

Vol. 21 | No. 3 | 2017

[Photo credit: StuPorts, AlexLMX/iStock/Thinkstock]

Research Directorate
Advancing Intelligence Through Science

THE
Next Wave
The National Security Agency's review of emerging technologies

Editor's column

While the global mobile telecommunications industry has been attempting to keep pace with ever-changing technology and consumer needs, it has not had a strong development roadmap like the International Technology Roadmap for Semiconductors (ITRS). Instead, mobile technology's development has been characterized by multinational companies pushing proposed standards for adoption by international standards bodies. These standards are subsequently adopted in a haphazard, nation-by-nation process. However, consumer desire for faster, more fully featured mobile devices has proven to be as strong a driving force as the ITRS, and has led the industry to roll out generations of new technology on a roughly 10-year basis (approximate dates: 1G - 1981, 2G - 1992, 3G - 2001, and 4G - 2009). This decade-by-decade introduction of new mobile communication technology has led to the common prediction for fifth-generation (5G) mobile to appear around 2020, in keeping with the observed linear cadence.

But why create 5G mobile anyway? The push for 5G is not just a mad rush to keep pace and provide more bandwidth to services that can already stream high-definition video. The Internet of Things (IoT; see TNW Vol. 21 No. 2) is a major driving force (among several) behind technologies being developed for 5G mobile. As personal mobile devices are more enmeshed into machine-to-machine (M2M) communications and the number of IoT sensors explodes, 5G technologies must address several needs: high-speed data rates for many more users, increased density of users, greatly increased simultaneous connections, and reduced latencies. These needs will propel many of the technologies that Dr. Farroha et al. describe in their introduction to 5G article (page 2).

One of those technologies, virtualization, has been instrumental in making efficient use of servers (virtual machine or VM) and computer networking (software-defined networking or SDN). Over the course of continuing improvement to 4G implementations, the networks are evolving into an all-digital Internet protocol packet-switched system. This evolution means that those efficiencies developed for SDN and used in computing can be applied to mobile networking. We can see this application in more depth in the article on 5G virtualization (page 16).

Several markets have already taken advantage of 4G mobile technology, most notably media. This market is expected to expand with 5G as consumer desire for high-resolution video and augmented/virtual reality increases. Automotive, energy, health, and public safety are several areas of M2M that will be greatly enhanced by increased bandwidth and, most importantly, low-latency networks. On 13 December 2016, the US Department of Transportation proposed a rule for vehicle-to-vehicle communication and announced a vehicle-to-infrastructure communication rule to be proposed. The article on 5G and the automotive industry (page 20) provides insight into how this market segment is enhanced by 5G connectivity.

Although mobile technological developments are typically categorized into different generations, each generation covers a vast array of individual technologies and protocols that roll out as they mature. In practice, mobile service providers often work across a range of generations. This is highlighted by the fact that in many cases voice calling is handled by the 3G network while data is handled on the 4G network. Domestically, T-Mobile and AT&T did not enable voice over LTE (VoLTE) until May 2014, and did not meet the International Telecommunication Union (ITU) standard until Verizon launched LTE-Advanced in August 2016.

Many factors can impact the broad global deployment of 5G networks, such as existing technologies, geography, spectrum, and national interests. Some countries may jump to the current network generation while others may delay deployment because their current networks are considered to be "good enough." The decreased per-customer costs of updating networks in densely populated countries makes it more attractive to keep on the leading edge. The increased data rates and simultaneous connections require increased spectrum allocations, which are handled by national regulatory organizations. Several countries, most notably South Korea, have tied their technological identity to the increased connectivity of next-generation wireless technologies. All these factors give rise to a patchwork international environment for 5G implementation.

Current forecasts for the rollout of 5G networks are in the 2020 time frame. Recent news indicates that it may be earlier than that. Samples of Qualcomm's new Snapdragon X50 4G/5G modem that uses the 27.5 gigahertz (GHz) to 28.35 GHz band—part of the spectrum opened by the Federal Communications Commission for 5G—will be available in the second half of 2017. Samsung and KT, one of South Korea's mobile providers, have announced that they will be the first to provide mobile 5G trial service at the PyeungChang 2018 Winter Olympic Games. Undoubtedly, this trial 5G service will not be fully compliant with 5G standards, which have yet to be adopted. However, we should expect things to advance quickly following the Korean introduction of 5G.

RECENT NEWS

[1] "AT&T launches first 5G business customer trial with Intel and Ericsson." AT&T. 5 Dec 2016. Available at: http://about.att.com/story/att_launches_first_5g_business_customer_trial_with_intel_and_ericsson.html.

[2] "Government must take action now to secure our connected future so we are ready for 5G, and essential services are genuinely available where they are needed." GOV.UK. 14 Dec 2016. Available at: https://www.gov.uk/government/news/government-must-take-action-now-to-secure-our-connected-future-so-we-are-ready-for-5g-and-essential-services-are-genuinely-available-where-they-are-n.

[3] Qualcomm Incorporated. "Qualcomm, Ericsson and SK Telecom announce collaboration on 5G NR trials to accelerate wide-scale 5G deployments." *PR Newswire*. 19 Dec 2016. Available at: http://www.prnewswire.com/news-releases/qualcomm-ericsson-and-sk-telecom-announce-collaboration-on-5g-nr-trials-to-accelerate-wide-scale-5g-deployments-300381570.html.

[4] Reichert C. "Huawei and NTT DoCoMo reach 11Gbps speeds in 5G Japanese field trial. 16 Nov 2016. Available at: http://www.zdnet.com/article/huawei-and-ntt-docomo-reach-11gbps-speeds-in-5g-japanese-field-trial/.

Contents

2 Introducing 5G
Dr. Sam Farroha,
Jared S. Everett,
Jason J. Uher,
Jason R. Harper,
Jessica K. Bridgland, and
Pamela M. Patton

16 Virtualizing the 5G Architecture
Staff Writer

20 5G and the Automotive Industry
Staff Writer

25 FROM LAB TO MARKET

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NSA/CSS.

This publication is available online at http://www.nsa.gov/thenextwave. For more information, please contact us at TNW@tycho.ncsc.mil.

Vol. 21 | No. 3 | 2017

Introducing 5G

Dr. Sam Farroha, Jared S. Everett (Johns Hopkins University Applied Physics Laboratory; JHU APL), Jason J. Uher (JHU APL), Jason R. Harper (JHU APL), Jessica K. Bridgland (JHU APL), and Pamela M. Patton (JHU APL)

[Photo credit: Juli_Rose/iStock/Thinkstock]

The evolution of modern cellular communications has been marked by a series of technology generations. Although the technology itself tends to evolve continuously, a new generation of standards marks a revolutionary step forward, with a substantial increase in system requirements to drive fundamentally new applications. With fourth-generation (4G) networks now widely deployed, the industry has turned its sights on "the next big thing." Fifth-generation (5G) networks are expected to enable a seamlessly connected society in the time frame beyond 2020 for both people and things, including vehicles, homes, smart cities, sensor networks, and the power grid. While the Long-Term Evolution (LTE) standard will continue to evolve and play a critical role in the wireless ecosystem, 5G represents an opportunity to architect a new system that is fundamentally different without the constraint of backward compatibility with existing technologies.

What is 5G?

Although there are as yet no standards for 5G mobile networks, a number of key technology trends have emerged. This article describes seven major technology trends that will pave the way to the next generation of 5G networks.

- **New Flexible Radio Access Technology (RAT):** A new, non-backward-compatible RAT will be defined for 5G that is distinct from previous generations, such as 4G LTE and its evolution. New multiple access schemes under consideration include various modified Orthogonal Frequency-Division Multiplexing (OFDM)-based solutions with improved spectral efficiency. The new RAT must be flexible enough to accommodate a variety of traffic types with often conflicting radio requirements. The concept of a unified air interface has been proposed for multiplexing multiple physical layer (PHY) regions with different characteristics [e.g., transmission time interval (TTI), subcarrier spacing] on a contiguous block of spectrum [1, 2, 3]. Spectrum for the new RAT will include existing bands below 6 gigahertz (GHz), as well as new centimeter-wave (cmWave) and millimeter-wave (mmWave) bands in the 6- to 100 GHz range [4]. The new RAT must also support significantly reduced latency, with as low as 100 microsecond (µs) transmission time interval (TTI) at the PHY for the ultra-reliable and low-latency communications (URLLC) use case [5]. Lastly, to further improve spectral efficiency, full-duplex transmission schemes have been proposed, potentially allowing the same time-frequency resources to be used for uplink and downlink transmissions simultaneously [6].
- **Virtualization:** Software-Defined Networking (SDN) and Network Function Virtualization (NFV) are two key architecture concepts in development to support the flexibility and mobility demands of the 5G network infrastructure [7, 8, 9]. Virtualization of network functions, which were traditionally implemented in hardware, will pave the way for commercial telecommunications operators and service providers to introduce new features and integrate new standards releases at an accelerated rate. NFV enables providers to move toward a decentralized network to increase flexibility, pushing core functions toward the edge to reduce latency, and virtualizing those functions on cloud-based servers. The proposed Cloud Radio Access Network (C-RAN) architecture, a specific use case of NFV applied to the RAN, uses a pooled architecture of baseband resources to increase scalability, physical layer flexibility, and spectral efficiency [7, 10].
- **Millimeter Wave (mmWave) Communications:** The term mmWave refers to carrier frequencies in the International Telecommunication Union (ITU) extremely high-frequency (EHF) band, from 30 to 300 GHz. Within the context

of 5G, the term has recently been loosely used by industry to refer to the higher frequencies from 6 to 100 GHz that are under consideration for new mobile spectrum [11, 12]. mmWave technologies are becoming an increasingly attractive solution to the problems of frequency reuse, cell density, raw data throughput, and antenna array size. This has led to a synergy between mmWave, small cell deployments, and massive multiple-input, multiple-output (MMIMO) techniques [13, 14].

- **Massive Multiple-Input, Multiple-Output (MMIMO) Techniques:** MMIMO is a new concept in antenna arrays that provides a number of advantages over traditional MIMO arrays currently deployed in 4G networks. Traditional MIMO arrays use only a few antenna elements (i.e., 2 to 16), whereas MMIMO uses a large number of elements in the array, currently considering a range of 128 to 512 at a minimum. Highly directional beamforming to multiple users simultaneously allows for increased user density and higher aggregate cell throughput [15]. So-called hybrid MMIMO has also been proposed; it combines beam steering with array processing techniques, such as spatial multiplexing, to increase single-user throughput [16].
- **Heterogeneous Networks (HetNets):** HetNets expand the mobile access network capacity by coordinating small cells with larger macro cells or offloading traffic to wireless local area network (WLAN) access points. There are two types of heterogeneity: 1) various cell sizes (e.g., macro, pico, femto) and 2) heterogeneous RATs [e.g., third-generation (3G), 4G, 5G, WLAN]. Small cells may include femto, pico, and micro cells, which can range in capacity from less than 10 to several hundred simultaneous active users. While HetNet deployments have already been introduced in 4G networks, network densification through the aggressive deployment of small cells is expected to increase significantly in future 5G networks [17, 18].
- **Native Machine-Type Communications (MTC) Support:** 5G networks are expected to incorporate a new model for connectivity specifically designed for MTC [19]. With the significant increase in connected machines over the last several years, a new 5G standard is seen as a prime opportunity to ensure new RATs can efficiently support a large number of connected devices with their own unique access constraints. Two categories of MTC are discussed: 1) general MTC and 2) vehicle-to-everything (V2X) MTC. General MTC devices have a few unique design and deployment considerations—namely, lower bandwidth needs, stringent power budgets, and relaxed latency requirements. V2X MTC devices, in contrast, require low-latency communications, out-of-coverage networks, and limited operation on a subscription-free basis [20].
- **Device-centric Architectures:** New network architectures will focus on a uniform quality of experience (QoE) for the user device, in contrast to traditional base-station-centric architectures. A number of new device-centric approaches are under consideration: decoupling the user plane and control plane, decoupling the uplink and downlink, and device-to-device communications [4, 21]. Another novel proposal is the user-centric cell or virtual cell model, which uses distributed beamforming and decoupled user/control planes to create a virtual cell around each user [2, 3]. Because the virtual cell follows the user, QoE variations are reduced and the cell-edge problem is mitigated. New device-centric architectures may significantly alter the traditional concept of cell handovers or eliminate it entirely.

Many of these technologies are already being added to the evolution of existing technologies beyond 4G, such as LTE-Advanced Pro [22].

In 2015, there was a significant increase in industry activities surrounding 5G networks. Major standards bodies, including the ITU and the Third-Generation Partnership Project (3GPP), reached important milestones in the early development of the eventual 5G standards. In September, the ITU published its vision for 5G networks [19]. The vision for International Mobile Telecommunications for 2020 and beyond (IMT-2020) defines three future-looking, high-level use cases for 5G:

- **Enhanced Mobile Broadband (eMBB):** This is generally a human-centric use case driven by the exponential increase in demand for mobile access to multimedia content, services, and data.

The eMBB use case will come with new application areas and requirements that go beyond existing mobile broadband applications for improved performance and increasingly seamless user experience. This use case covers a range of scenarios, including wide-area coverage and localized high-throughput spot coverage, which will have different requirements.

- **Massive Machine-Type Communications (mMTC):** This use case is characterized by a large number of connected devices typically transmitting a relatively low volume of nondelay-sensitive data. Devices are intended to be low cost and have a very long battery life.
- **Ultra-Reliable and Low-Latency Communications (URLLC):** This use case is characterized by stringent requirements for latency, throughput, and availability. Examples include wireless control of industrial manufacturing processes, remote medical surgery, distributed smart grid automation, and transportation safety [e.g., vehicle-to-vehicle (V2V) or vehicle-to-everything (V2X) communication]. Many companies have referred to this use case as critical MTC (cMTC) or ultra-reliable MTC (uMTC). However, based on the ITU definition in [19], this use case is not strictly limited to MTC applications.

It is important to consider that the applications that will use 5G technology do not necessarily correspond to a single use case but are more accurately described as a combination of multiple use cases. Figure 1 illustrates some examples of currently envisioned 5G applications and their relationship to these three IMT-2020 use cases [19]. Figure 2 illustrates eight key capabilities identified by ITU for IMT-2020 and their relative importance to the same three use cases [19]. Furthermore, additional future use cases are expected to emerge but cannot be accurately predicted (i.e., what will be the "killer app" in 2025?). Therefore, it is desired that 5G standards will provide the flexibility to adapt to new use cases.

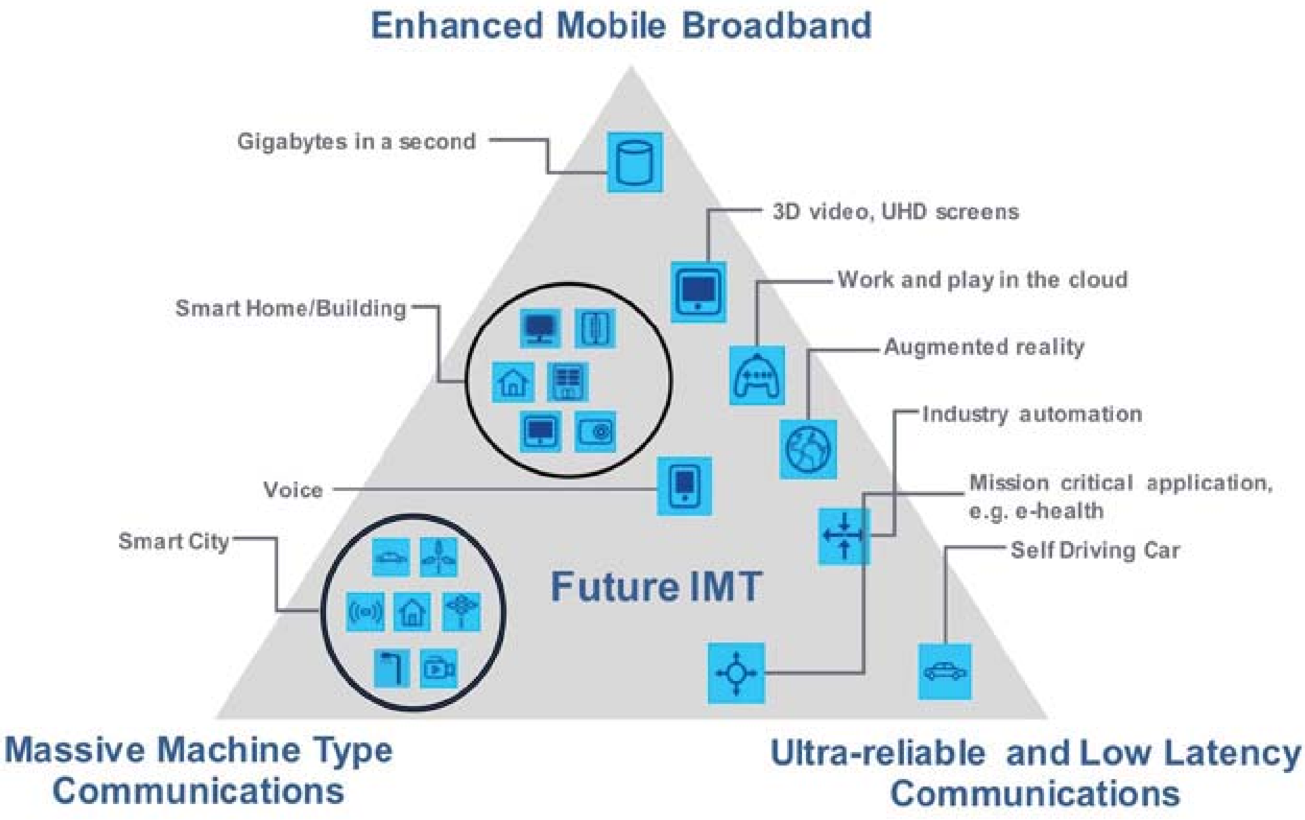
Enhanced Mobile Broadband
Gigabytes in a second
3D video, UHD screens
Work and play in the cloud
Smart Home/Building
Augmented reality
Industry automation
Voice
Mission critical application, e.g. e-health
Smart City
Self Driving Car
Future IMT
Massive Machine Type Communications
Ultra-reliable and Low Latency Communications

FIGURE 1. 5G use cases as defined by ITU for IMT-2020 [19].

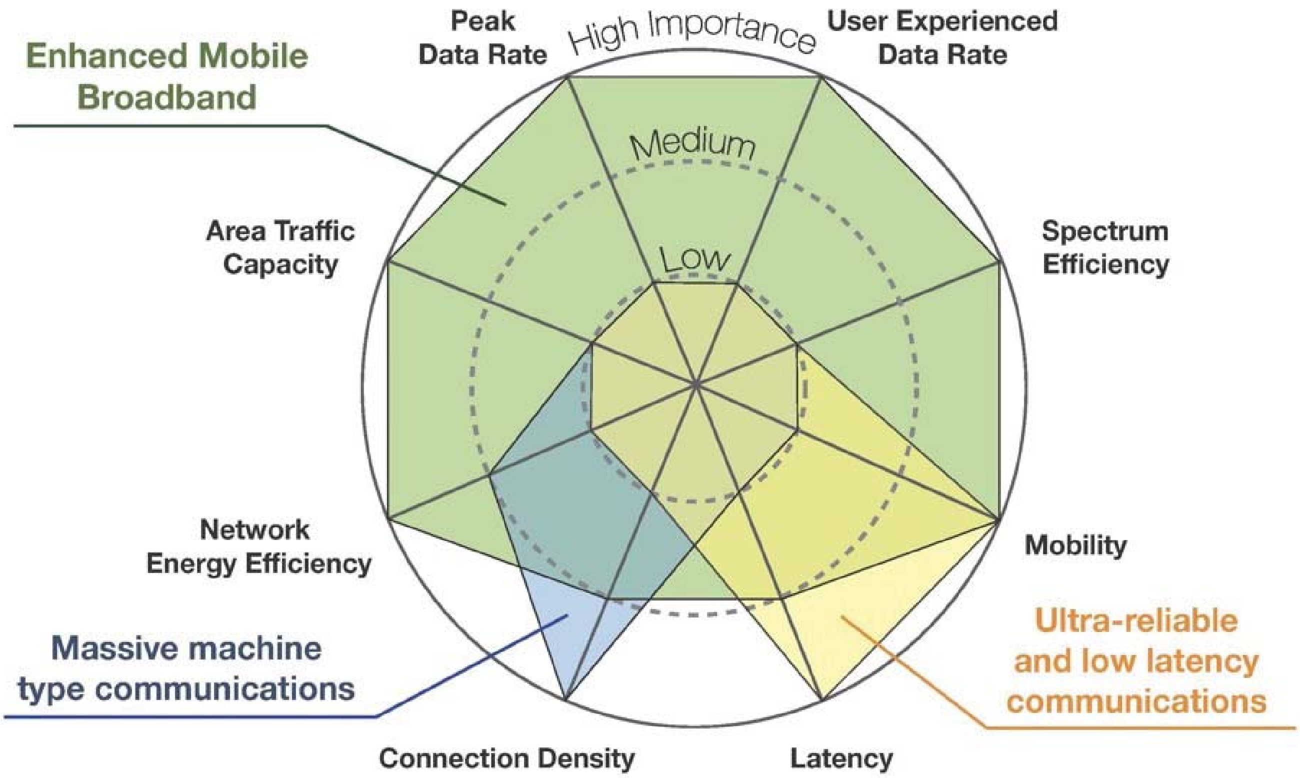
Peak Data Rate
High Importance
User Experienced Data Rate
Enhanced Mobile Broadband
Medium
Area Traffic Capacity
Low
Spectrum Efficiency
Network Energy Efficiency
Mobility
Massive machine type communications
Ultra-reliable and low latency communications
Connection Density
Latency

FIGURE 2. Eight key capabilities and their relative importance to 5G use cases [19].

5G Standardization

Development work toward 5G is well under way. Standards bodies are actively working on new 5G mobile technologies to be deployed in the 2020 time frame. This section summarizes the activities and corresponding 5G development timelines for three major standards bodies: ITU, 3GPP, and the Institute of Electrical and Electronics Engineers (IEEE).

Standardization in ITU

ITU is the United Nations agency responsible for promoting worldwide improvement and rational use of information and communication technology. Its members include industry, academia, and standards organizations from more than 190 member nations. The ITU Radiocommunication Sector (ITU-R) works toward worldwide consensus in the use of terrestrial and space radiocommunication services, including mobile communication technologies. Although compliance with ITU-R recommendations is not mandatory, they nevertheless have a high degree of adoption worldwide and hold the status of international standards [23].

International Mobile Telecommunications framework

ITU-R Working Party 5D (WP 5D) is responsible for overall radio aspects of terrestrial mobile systems, referred to as International Mobile Telecommunications (IMT). The purpose of IMT is to provide high-quality mobile services with a high degree of interoperability worldwide. Since 2000, the ITU has developed the IMT standards framework in a manner that parallels cellular generations from an industry perspective. Although ITU-R WP 5D defines the requirements for IMT, it does not develop the actual radio technologies. Rather, candidate radio technologies are submitted for inclusion by external standards bodies, such as 3GPP and IEEE. For this reason, ITU-R WP 5D maintains strong cooperation with the major global standards bodies.

The first family of standards derived from the IMT concept (IMT-2000) aligned with 3G cellular. Radio technologies accepted into IMT-2000 included 3GPP Wideband Code-Division Multiple Access (WCDMA), 3GPP2 cdma2000, and IEEE 802.16 [i.e., Mobile Worldwide Wireless Interoperability for Microwave Access (WiMAX)]. The next generation of IMT standards (IMT-Advanced) aligned with 4G cellular. Radio technologies accepted into IMT-Advanced included 3GPP LTE-Advanced and IEEE 802.16m [i.e., Wireless Metropolitan Area Network (WMAN)-Advanced].

Timeline for IMT-2020

In 2012, ITU embarked on a program to develop “IMT for 2020 and beyond,” setting the stage for emerging 5G research activities around the world. The program has since adopted the name IMT-2020 and forms the framework for the next generation of mobile broadband standards. The timeline for the development of IMT-2020 is shown in figure 3 [24]. The IMT 2020 timeline will essentially follow the same process used in the development of IMT-Advanced.

The IMT-2020 program is well under way, with a number of key milestones completed. In September 2015, ITU published its vision of the 5G mobile broadband connected society [19]. This document defined three high-level use cases for 5G, described earlier in this article, which have already been widely adopted by 3GPP and industry in general. In the next phase, the 2016–2017 time frame, ITU-R WP 5D will define in detail the performance requirements, evaluation criteria, and methodology for the assessment of the new IMT radio interfaces. It is anticipated that the time frame for proposals will be focused in 2018. In the 2018–2020 time frame, independent, external groups will evaluate proposals and the definition of the new radio interfaces to be included in IMT-2020 will take place. ITU-R WP 5D also plans to hold a workshop in late 2017 to discuss the performance requirements and evaluation criteria for candidate technologies for IMT-2020, as well as to provide an opportunity for presentations by potential proponents for IMT-2020 in an informal setting. The whole process is planned to be completed in 2020, when a new draft of the ITU-R recommendation with detailed

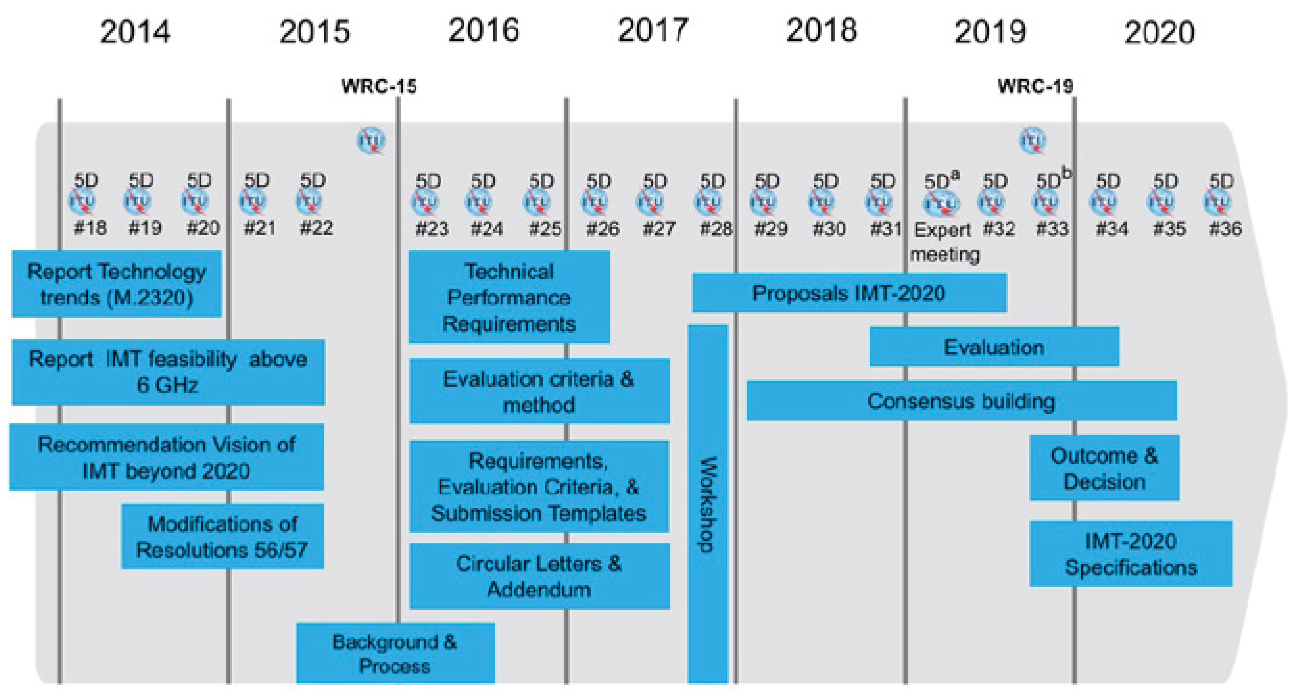
2014
2015
2016
2017
2018
2019
2020
WRC-15
WRC-19
5D #18
5D #19
5D #20
5D #21
5D #22
5D #23
5D #24
5D #25
5D #26
5D #27
5D #28
5D #29
5D #30
5D #31
5Da Expert meeting
5D #32
5Db #33
5D #34
5D #35
5D #36
Report Technology trends (M.2320)
Report IMT feasibility above 6 GHz
Recommendation Vision of IMT beyond 2020
Modifications of Resolutions 56/57
Background & Process
Technical Performance Requirements
Evaluation criteria & method
Requirements, Evaluation Criteria, & Submission Templates
Circular Letters & Addendum
Workshop
Proposals IMT-2020
Evaluation
Consensus building
Outcome & Decision
IMT-2020 Specifications

(a) – if needed focus meeting towards WRC-19 (non-Technology), (b) – focus meeting on Evaluation (Technology)

Note: While not expected to change, details may be adjusted if warranted.

FIGURE 3. Detailed timeline and process for IMT-2020 in ITU-R [24].

specifications for the new radio technologies will be submitted for approval within ITU-R [24].

Standardization in 3GPP

3GPP is the international standards body responsible for the development and maintenance of major second-generation (2G), 3G, and 4G cellular standards. The purpose of the organization is to produce interoperable cellular communications standards, as well as studies and reports that define 3GPP technologies. The following technologies are currently maintained and evolved by 3GPP:

- Global System for Mobile Communications (GSM), General Packet Radio Service (GPRS), and Enhanced Data Rates for GSM Evolution (EDGE);
- Universal Mobile Telecommunications System (UMTS), WCDMA, High-Speed Packet Access (HSPA), and HSPA Evolution (HSPA+); and
- LTE, LTE-Advanced, and LTE-Advanced Pro.

These 3GPP technologies are constantly evolving through a series of backward-compatible releases. Since the completion of the first LTE and Evolved Packet Core (EPC) specifications, 3GPP has become the focal point for mobile systems beyond 3G. Therefore, 3GPP is expected to be a critical player in the development of 5G, and their timeline will have a direct influence on the timeline of the emerging 5G market.

3GPP is currently defining a new 5G RAT and corresponding network architecture. These are being developed within 3GPP under the working names “new radio (NR)” and “next-generation (NextGen) architecture,” respectively [25]. In October 2016, 3GPP announced that the new 3GPP system will officially be known by the name “5G” from Release 15 onward [26]. Some initial standardization steps that have been taken to date include the following:

- **SMARTER study item:** In March 2015, 3GPP Technical Specification Group (TSG) System Aspects (SA) began a study item on technology enablers for new 5G services and markets, known as the SMARTER study item [27]. The objective of this study was to develop high-level use cases and identify the related high-level potential requirements to enable 3GPP network operators to support new services and markets in 5G. Phase 1 of the SMARTER study item was completed in March 2016; results are documented in 3GPP Technical Report (TR) 22.891 to be included in Release 14 [28]. A total of 74 use cases were identified. This work prompted four building block studies that grouped the use cases into families with common requirements: massive Internet of Things (IoT), critical communications, eMBB, and network operation. The building block studies were completed in June 2016; results are documented in 3GPP TRs 22.861, 22.862, 22.863, and 22.864 to be included in Release 14 [29]. The results of the SMARTER study will form the basis for a work item to define normative stage 1 requirements for the next-generation 5G system. The work item is scheduled for completion in March 2017; results will be documented in 3GPP Technical Specification (TS) 22.261 to be included in Release 15 [29].
- **Study item on channel model for frequency spectrum above 6 GHz:** The first 5G study conducted by TSG RAN focused on developing new channel models to support high-frequency spectrum from 6 GHz to 100 GHz. The models consider a variety of scenarios including urban, rural, and indoor, as well as the impact of line-of-sight (LOS) versus non-LOS (NLOS). The study was completed in June 2016, and results are documented in 3GPP TR 38.900 to be included in Release 14 [30].
- **Study item on architecture for next-generation system:** In December 2015, 3GPP TSG SA approved a study item to design a system architecture for the next generation of mobile networks. The new architecture will support at least the new 5G RAT(s), the evolution of LTE, and non-3GPP access types and will minimize access dependencies. The study considers new approaches such as NFV and network slicing. The study item was scheduled for completion in December 2016; results will be documented in 3GPP TR 23.799 to be included in Release 14 [31].
- **Study item on scenarios and requirements for next-generation access technologies:** In December 2015, 3GPP TSG RAN approved a study item to develop deployment scenarios and requirements of next-generation access

technologies. The study identifies 12 deployment scenarios that are more diverse than those originally envisioned for legacy RATs, such as LTE and its predecessors. It also identifies key performance indicators (KPIs) and other requirements for 5G NR. The bulk of the study was completed in September 2016 to provide guidance to the ongoing technical work being performed in the RAN working groups. However, the study item will remain open until March 2017 to match the IMT-2020 timeline and ensure all IMT-2020 requirements are captured. Final results are documented in 3GPP TR 38.913 to be included in Release 14 [32, 4].

- **Study item on NR access technology:** In March 2016, 3GPP TSG RAN approved a study item to develop the 5G NR access technology capable of meeting the broad range of use cases defined for 5G. The study seeks to develop a single technical framework capable of addressing all usage scenarios and requirements defined in TR 38.913 for eMBB, mMTC, and URLLC, with an emphasis on forward compatibility. The study is scheduled for completion in March 2017; results will be documented in 3GPP TR 38.912 to be included in Release 14 [1].

5G standardization activities in 3GPP will continue through 2020 and beyond, as described next.

Emerging 3GPP standardization timeline

In March 2015, 3GPP announced a tentative standardization timeline for 5G based on the ITU work plan timeline for IMT-2020 [33]. Since then, a more detailed timeline has come into focus as study items have commenced and completed, and as the 3GPP TSGs coordinate for the initial release of 5G. The timeline shown in figure 4 is a composite from multiple sources.

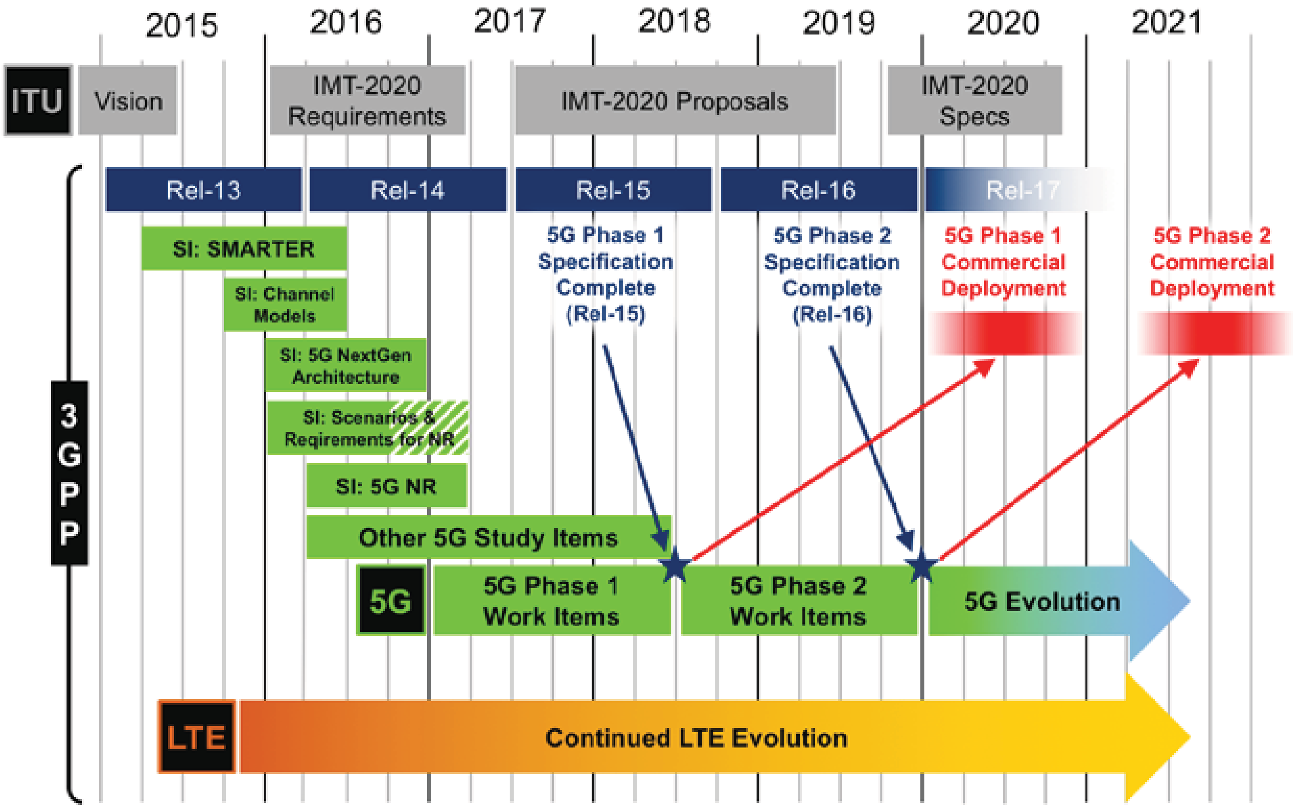
2015
2016
2017
2018
2019
2020
2021
ITU
Vision
IMT-2020 Requirements
IMT-2020 Proposals
IMT-2020 Specs
3GPP
Rel-13
Rel-14
Rel-15
Rel-16
Rel-17
SI: SMARTER
SI: Channel Models
SI: 5G NextGen Architecture
SI: Scenarios & Reqirements for NR
SI: 5G NR
Other 5G Study Items
5G Phase 1 Specification Complete (Rel-15)
5G Phase 2 Specification Complete (Rel-16)
5G Phase 1 Commercial Deployment
5G Phase 2 Commercial Deployment
5G
5G Phase 1 Work Items
5G Phase 2 Work Items
5G Evolution
LTE
Continued LTE Evolution

FIGURE 4. Emerging 5G standardization timeline for 3GPP. (Figure is a composite from [34] and [35] that includes additional data from various sources.)

The initial 5G study items in TSG SA and TSG RAN commenced in 2015 and 2016, as described previously. These initial 5G study items will be included in 3GPP Release 14. This work is carried out in parallel with ongoing LTE work. At the 3GPP plenary meeting in June 2016, the TSGs agreed on a work plan for the first release of 5G in 3GPP Release 15, including a clear work division between the TSGs [25].

5G work items were scheduled to begin in December 2016 for TSG SA and March 2017 for TSG RAN. The Phase 1 5G work items will fall into Release 15, with planned completion in June 2018. Additional 5G study items will continue during Release 15 in support of Phase 2. Subsequently, the Phase 2 5G work items will fall into Release 16, which will be completed around December 2019 in time for the final submission to ITU for IMT-2020. Phase 1 commercial deployments are expected to begin in 2020, followed by Phase 2 deployments in the 2021–2022 time frame. However, early pre-5G mmWave deployments may emerge in limited markets, such as South Korea or the United States, before 2020.

For example, Verizon Wireless has announced its plans to pilot a 28-GHz mmWave deployment in the United States for fixed wireless applications starting in 2017 [36]. To support this effort, the Verizon 5G Technology Forum (V5GTF)—an industry consortium led by Verizon—published an open radio interface specification in July 2016 [37]. The Verizon specification uses an OFDM-based PHY similar to time-division LTE (TD-LTE) with enhanced beamforming for operation in 28 and 39 GHz mmWave spectrum. However, with the initial focus on fixed wireless, the first release does not support user mobility. The Verizon specification can be considered pre-5G in the sense that it supports new mmWave capabilities beyond 4G but does not address all the use cases and associated requirements for 5G. The Verizon specification is expected to be incompatible with the eventual 3GPP 5G standard, potentially leading to market fragmentation [38].

3GPP phased approach to 5G standardization

3GPP TSG RAN will take a two-phased approach to developing the new 5G RAT [35]. The Phase 1 standard will define a new, non-backward-compatible 5G RAT. A subset of prioritized features and use cases will be addressed in Phase 1 to allow for early commercial deployments targeted for the year 2020. The Phase 2 standard will implement the full set of features and use cases necessary to meet the requirements for 5G. An initial proposal will be submitted to ITU as a candidate radio interface technology for IMT-2020 by the June 2019 submission deadline. The Phase 2 standard will later form the final submission around December 2019. The Phase 1 standard will be designed for forward compatibility with Phase 2 [35]. Forward compatibility means that Phase 1 must be designed from the beginning to optimally accommodate all of the features and use cases expected to be added later in Phase 2, even though those features are not yet fully implemented. Although the forward-compatibility requirement may sound straightforward, it represents a fundamental shift from the normal 3GPP standardization process, which historically has focused on a series of backward-compatible releases.

While prioritization of features between the two phases has been a topic of much debate, it is clear that the 5G Phase 1 standard will support tight interworking with LTE to simplify initial rollout. The phased approach and tight interworking with LTE means that elements of the LTE system architecture may persist in 5G deployments for some time to come. This implies that current and future work on LTE, LTE-Advanced, and LTE-Advanced Pro networks and technologies may have direct applicability to eventual 5G network deployments.

Standardization in IEEE

Initial 5G standards activities within the IEEE suggest that they do not intend to be a direct competitor with organizations like 3GPP on the radio interface between the RAN and the user equipment. Instead, IEEE has begun developing complementary technologies to support other communications requirements within the 5G ecosystem. In 2016, IEEE established two new working groups related to 5G: IEEE 1914 and IEEE 1918.

IEEE 1914 is the Next Generation Fronthaul Interface Working Group. This working group is currently developing two standards: the 1914.1 standard for packet-based fronthaul transport networks and the 1914.3 standard for radio over Ethernet encapsulations and mappings [39, 40]. These standards focus on the fronthaul interface within the RAN between baseband units (BBUs) and remote radio heads (RRHs) to support novel RAN architectures like C-RAN, and antenna techniques like MMIMO and coordinated multi-point (CoMP) transmission and reception. The projected completion dates for these standards are August 2018 for 1914.1 and October 2017 for 1914.3.

IEEE 1918 is the Tactile Internet Working Group. This working group is currently developing the 1918.1 standard, which defines a framework for the Tactile Internet [41]. The purpose of this framework is to establish a basis for the rapid development of the Tactile Internet as a 5G and beyond application, with the expectation of additional IEEE 1918 standards to follow. The projected completion date for the 1918.1 standard is October 2018.

With respect to IMT-2020, IEEE may seek to expand the role of WLAN in 5G as a complementary radio interface for next-generation HetNets. In September 2016, the IEEE 802.11 working group sent a liaison statement to 3GPP TSG RAN and TSG SA inviting them to consider the use of IEEE 802.11-based WLAN in unlicensed spectrum as a complementary means of meeting the performance requirements of IMT-2020, potentially leading to inclusion in a joint submission to IMT-2020 [42]. This approach would be a logical extension of the increasing level of interworking between LTE and WLAN in recent standards releases. WLAN is already widely used in 3GPP networks for high data rate offloading.

Recent enhancements in radio-level interworking have increased the efficiency of these networks. Enhancements include LTE-WLAN Aggregation (LWA) and LTE WLAN Radio Level Integration with IPsec Tunnel (LWIP) in 3GPP Release 13, with further enhancements in 3GPP Release 14. Although 3GPP declined to make a decision at the September 2016 plenary meeting, the concept of a potential joint submission could represent a novel approach to IMT-2020. In contrast, previous generations of IMT saw IEEE in competition with 3GPP, with the submission of the IEEE 802.16 WiMAX family of standards to IMT-2000 and IMT-Advanced as a direct competitor in the 3G and 4G markets.

Conclusion

This article provided an introduction to major technology trends in the emergence of next-generation 5G mobile networks. These networks are expected to see initial commercial deployment starting around the year 2020. Early 5G standardization activities in the ITU, 3GPP, and IEEE were addressed.

For further information on the latest developments in 5G, the interested reader is directed to the following resources. ITU publications for IMT-2020 can be found on the IMT-2020 web page [24]. Notable documents include the ITU vision for IMT-2020 [19] as well as the technical performance requirements for IMT-2020 (scheduled for completion in February 2017). The 3GPP web site (www.3GPP.org) is the most direct source for 3GPP-related technical information. 3GPP press releases provide high-level summaries of ongoing standards activities and often include links to more detailed further reading. The latest versions of the 3GPP TR and TS documents mentioned in this article can be accessed there as well. Information on the IEEE 1914.1, 1914.3, and 1918.1 standards can be found in the approved project authorization request (PAR) documents [39, 40, 41] and the corresponding working group web pages. Lastly, the Verizon mmWave specification is available on the V5GTF website (www.5GTF.org). This pre-5G specification defines layers 1 to 3 of an open radio interface using a document structure similar to that of LTE.

References

[1] 3GPP TSG RAN. "Revision of SI: Study on new radio access technology." September 2016. 3GPP Document TD RP-161596. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_73/Docs/RP-161596.zip.

[2] Huawei. "Vision on 5G radio access technologies." September 2015. 3GPP Document RWS 150006. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150006.zip.

[3] ZTE. "Considerations on 5G key technologies and standardization." September 2015. 3GPP Document dRWS 150024. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150024.zip.

[4] 3GPP. "3GPP TSG RAN study on scenarios and requirements for next generation access technologies." October 2016. 3GPP Specification TR 38.913, Release 14, Version 14.0.0. Available at: http://www.3gpp.org/DynaReport/38913.htm.

[5] Fettweis G, Alamouti S. "5G: Personal mobile internet beyond what cellular did to telephony." *IEEE Communications Magazine*. February 2014; 52(2): 140-145. doi: 10.1109/MCOM.2014.6736754.

[6] Zhang Z, Chai X, Long K, Vasilakos AV, Hanzo L. "Full duplex techniques for 5G networks: Self-interference cancellation, protocol design, and relay selection." *IEEE Communications Magazine*. May 2015; 53(5): 128-137. doi: 10.1109/MCOM.2015.7105651.

[7] ITU-R. "Future technology trends of terrestrial IMT systems." November 2014. ITU-R Report M.2320-0. Available at: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2320-2014-PDF-E.pdf.

[8] 4G Americas. "5G: 4G Americas' recommendations on 5G requirements and solutions." October 2014. Available at: http://www.4gamericas.org/files/2714/1471/2645/4G_Americas_Recommendations_on_5G_Requirements_and_Solutions_10_14_2014-FINALx.pdf.

[9] 3GPP. "3GPP NFV study: Management of mobile networks that include virtualized network functions (MANO)" [Press release]. 5 Oct 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1738-sa5_nfv_study.

[10] Checko A, Christiansen H, Yan Y, Scolari L, Kardaras G, Berger M, Dittman L. "Cloud RAN for mobile networks—A technology overview." *IEEE Communication Surveys and Tutorials*. January 2015; 17(1): 405-426. doi: 10.1109/COMST.2014.2355255.

[11] ITU-R. "Technical feasibility of IMT in bands above 6 GHz." June 2015. ITU-R Report M.2376-0. Available at: www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2376-2015-PDF-E.pdf.

[12] Rangan S, Rappaport T, Erkip E. "Millimeter-wave cellular wireless networks: Potentials and challenges." In: *Proceedings of the IEEE;* 2014 January; 102(3). doi: 10.1109/JPROC.2014.2299397.

[13] Swindlehurst AL, Ayanoglu E, Heydari P, Capolino F. "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Communications Magazine*. September 2014; 52(9): 56-62. doi: 10.1109/MCOM.2014.6894453.

[14] Straight Path Communications. "A straight path toward 5G." September 2015. 3GPP Document RWS150087. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS150087.zip.

[15] Hoydis J, ten Brink S, Debbah M. "Massive MIMO: How many antennas do we need?" In: *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*; 2011 Sep 27-28; Monticello, IL; p. 545–550. doi: 10.1109/Allerton.2011.6120214.

[16] InterDigital. "Roadmap for 5G radio standardization in 3GPP." September 2015. 3GPP Document RWS 150030. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150030.zip.

[17] Wannstrom J, Mallinson K. "Heterogeneous networks in LTE." Available at: http://www.3gpp.org/technologies/keywords-acronyms/1576-hetnet.

[18] Bangerter B, Talwar S, Arefi R, Stewart K. "Networks and devices for the 5G era." *IEEE Communications Magazine*. February 2014; 52(2): 90–96. doi: 10.1109/MCOM.2014.6736748.

[19] ITU-R. "IMT vision—Framework and overall objectives of the future development of IMT for 2020 and beyond." September 2015. ITU-R Recommendation M.2083-0. Available at: http://www.itu.int/rec/R-REC-M.2083.

[20] General Motors Research. "Towards 5G: An automotive perspective." September 2015. 3GPP Document RWS 150011. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150011.zip.

[21] Boccardi F, Heath Jr. R, Lozano A, Marzetta T, Popovski P. "Five disruptive technology directions for 5G." *IEEE Communications Magazine*. February 2014; 52(2): 74-80. doi: 10.1109/MCOM.2014.6736746.

[22] 3GPP. "LTE-advanced pro ready to go" [Press release]. 28 Oct 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1745-lte-advanced_pro.

[23] ITU-R. "ITU-R study groups." May 2013. Available at: http://www.itu.int/dms_pub/itu-r/opb/gen/R-GEN-SGB-2013-PDF-E.pdf.

[24] ITU-R. "ITU toward 'IMT for 2020 and beyond.'" Available at: http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Pages/default.aspx.

[25] 3GPP. "3GPP on track to 5G" [Press release]. 27 Jun 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1787-ontrack_5g.

[26] 3GPP. "New partners at London 3GPP meeting" [Press release]. 24 Oct 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1800-pcg_37.

[27] 3GPP TSG SA. "New WID study on new services and markets technology enablers (FS_SMARTER) from S1-150300." March 2015. 3GPP Document SP-150142. Available at: ftp://ftp.3gpp.org/tsg_sa/TSG_SA/TSGS_67/Docs/SP-150142.zip.

[28] 3GPP. "3GPP TSG SA feasibility study on new services and markets technology enablers stage 1." June 2016. 3GPP Specification TR 22.891, Release 14, Version 14.1.0. Available at: http://www.3gpp.org/DynaReport/22891.htm.

[29] 3GPP. "SA1 completes its study into 5G requirements" [Press release]. 23 Jun 2016. Available at: http://www.3gpp.org/news-events/3gpp-news/1786-5g_reqs_sa1.

[30] 3GPP. "3GPP TSG RAN study on channel model for frequency spectrum above 6 GHz." June 2016. 3GPP Specification TR 38.900, Release 14, Version 14.0.0. Available at: http://www.3gpp.org/DynaReport/38900.htm.

[31] 3GPP TSG SA. "New WID study on architecture for next generation system." December 2015. 3GPP Document TD SP-150853. Available at: http://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGS_70/Docs/SP-150853.zip.

[32] 3GPP TSG RAN. "New study item proposal: Study on scenarios and requirements for next generation access technologies." December 2015. 3GPP Document TD RP-152257. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_70/Docs/RP-152257.zip.

[33] Flore D, Bertenyi B. "Tentative 3GPP timeline for 5G" [Press release]. 17 Mar 2015. Available at: http://www.3gpp.org/news-events/3gpp-news/1674-timeline_5g.

[34] Qualcomm. "5G: Views on technology and standardization." September 2015. 3GPP Document RWS 150012. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150012.zip.

[35] Flore D. "RAN workshop on 5G: Chairman summary." September 2015. 3GPP Document RWS 150073. Available at: ftp://ftp.3gpp.org/workshop/2015-09-17_18_RAN_5G/Docs/RWS 150073.zip.

[36] Alleven M. "Verizon's Shammo: 5G pilot in 2017 is all about fixed wireless, not mobility." *Fierce Wireless*. 2016 Apr 21. Available at: http://www.fiercewireless.com/tech/story/verizons-shammo-5g-pilot-2017-all-about-fixed-wireless-not-mobility/2016-04-21.

[37] V5GTF. "V5GTF Network and Signaling Working Group; Verizon 5G radio access; Overall description." June 2016. V5GTF Specification TS V5G.300, Release 1, Version 1.0. Available at: http://www.5gtf.org/V5G_300_v1p0.pdf.

[38] Alleven M. "Verizon's version of 5G not compatible with 3GPP's current specs—or easily upgradeable: Report." *Fierce Wireless*. 2016 Oct 28. Available at: http://www.fiercewireless.com/tech/verizon-s-version-5g-not-compatible-3gpp-s-current-specs-or-easily-upgradeable-report.

[39] IEEE. "IEEE P1914.1: Standard for packet-based fronthaul transport networks." February 2016. IEEE Approved PAR P1914.1. Available at: http://sites.ieee.org/sagroups-1914/files/2016/06/P1914.1.pdf.

[40] IEEE. "IEEE P1904.3: Standard for radio over ethernet encapsulations and mappings." June 2016. IEEE Approved PAR P1904.3. Available at: http://sites.ieee.org/sagroups-1914/files/2016/07/P1904.3.pdf.

[41] IEEE. "IEEE P1918.1: Tactile Internet: Application scenarios, definitions and terminology, architecture, functions, and technical assumptions." March 2016. IEEE Approved PAR P1918.1. Available at: http://grouper.ieee.org/groups/1918/1/files/P1918.1.pdf

[42] 3GPP TSG RAN. "IEEE 802.11 Working group liaison on the role of WLAN in IMT-2020." September 2016. 3GPP Document TD RP-161804. Available at: http://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_73/Docs/RP-161804.zip.

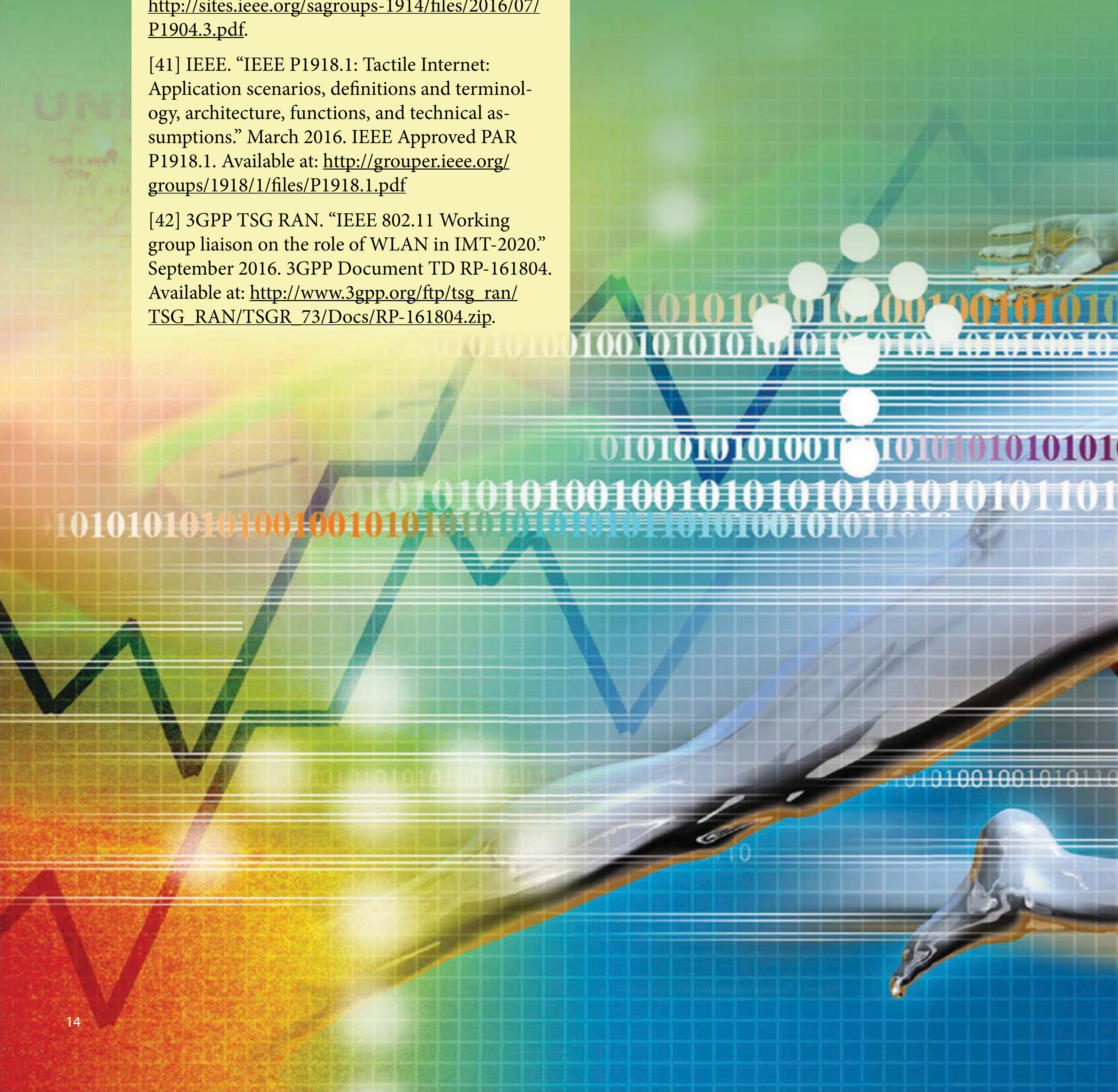

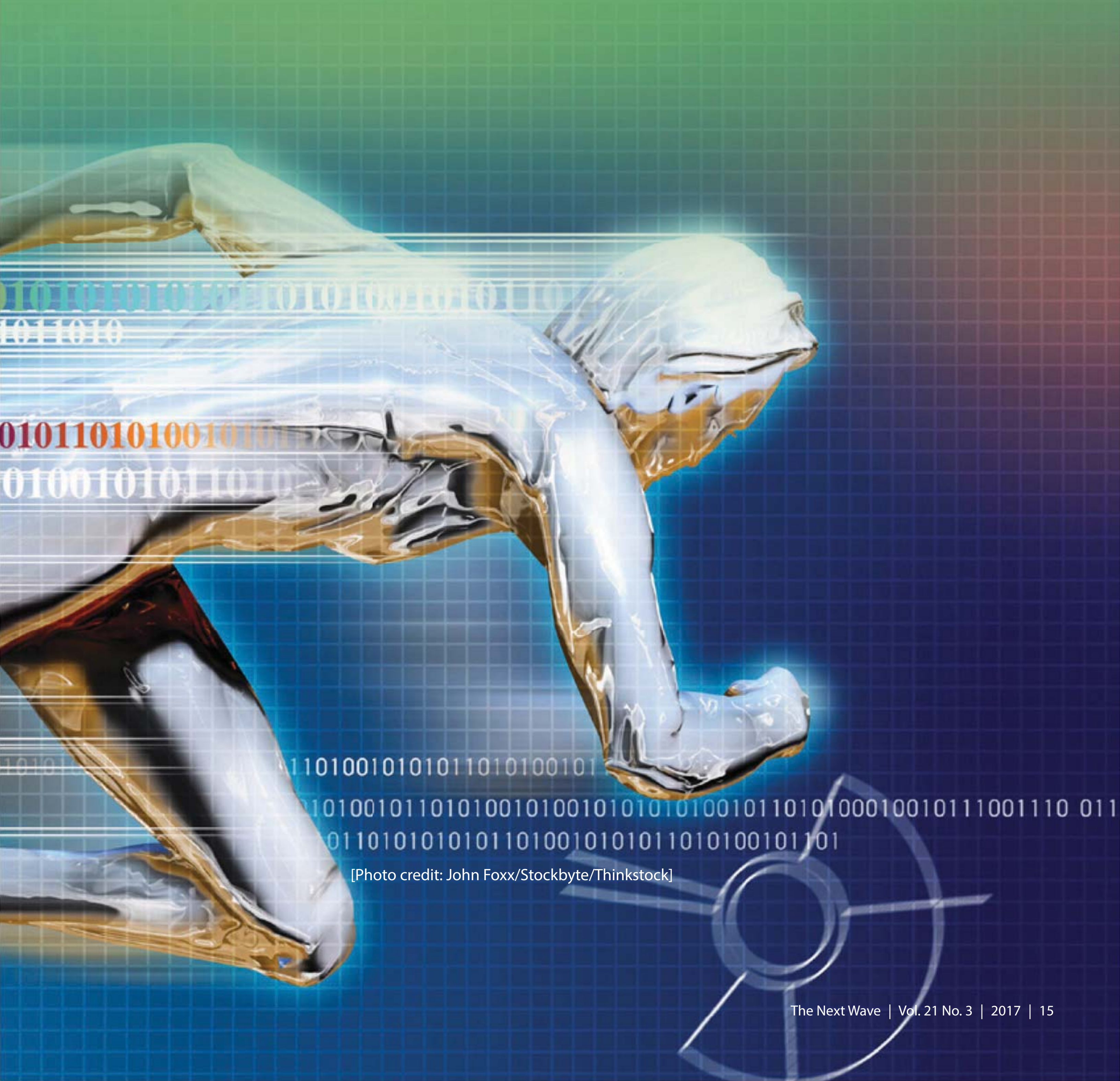

[Photo credit: John Foxx/Stockbyte/Thinkstock]

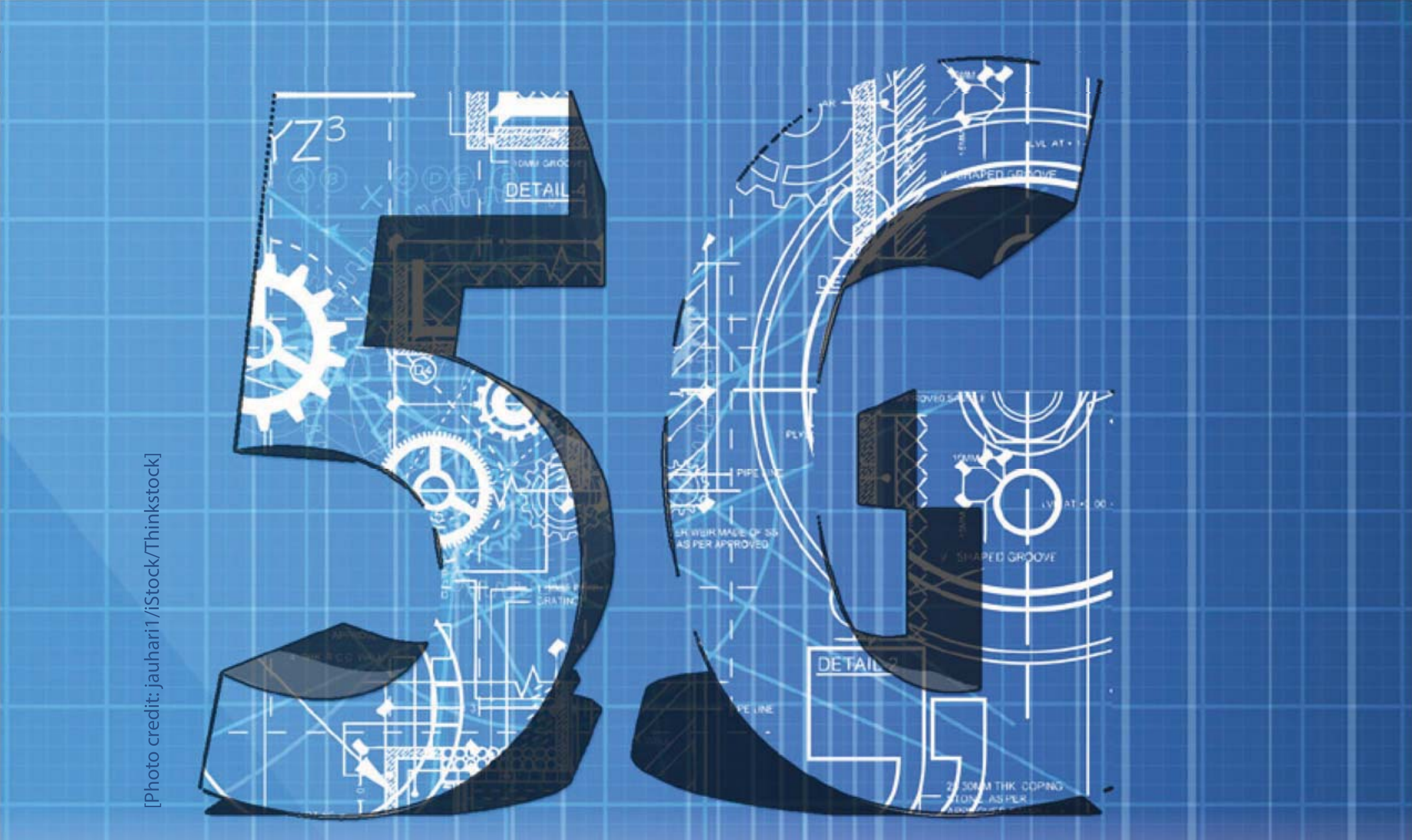

[Photo credit: jauhari1/iStock/Thinkstock]

Virtualizing the 5G Architecture

Staff Writer

Virtually changing the 5G architecture

Virtualization is set to play a major role in the evolution of the fifth-generation (5G) core network. According to industry experts, 5G will use software-centric networking technologies such as software-defined networking (SDN) and network functions virtualization (NFV), and will be natively cloud based. If correct, this will represent a major transition in system architecture and will require much greater collaboration across the networking ecosystem. The push to incorporate more cloud- or software-based components is driven by the need for greater flexibility and scalability to respond to the demands of radio access technologies that offer more bandwidth, reduced latency, and stringent quality of service (QoS) requirements. The new 5G core network must be adaptable and better equipped to handle various devices and manage capacity in near-real time. Mobile network operators see the advances in cost and efficiency that virtualization brings to other market segments and will use the emerging 5G technology to determine if these same advances can benefit the mobile market [1, 2].

SDN and NFV are two key architecture concepts in development to support the flexibility and mobility demands of the 5G network infrastructure. Virtualizing network functions that were previously implemented in hardware will allow providers to introduce new features and integrate new standards at a faster rate. SDN/NFV provides an avenue for providers to decentralize their networks, thereby increasing flexibility and reducing latency. Two areas where SDN/NFV will benefit 5G networks, and in some cases even fourth-generation (4G) mobile technology, are network slicing and cloud-radio access network (C-RAN).

Network slicing

Network slicing would promote end-to-end mobile network virtualization by "slicing" the network into virtual channels. These virtual channels would be autonomous and encompass a set of resources—physical or virtual—including bandwidth on a network link, processing capacity of servers, processing capacity of network elements, as well as operations support system (OSS) and business support system (BSS) processes. Operators could then use these channels to dynamically devote the appropriate network resources to create a "lane" in the network specifically designed for a particular use or service. This would accommodate the many use cases being put forth for 5G. The operator-led Next Generation Mobile Networks (NGMN) Alliance has sought to define categories of 5G use cases (i.e., service types) that have distinct performance characteristics and commercial potential. In a 2015 white paper, the NGMN listed eight application categories for 5G [3, 4, 5]:

1. Broadband access in dense areas,
2. Broadband access everywhere,
3. Higher user mobility,
4. Massive Internet of Things,
5. Extreme real-time communications,
6. Lifeline communications,
7. Ultra-reliable communications, and
8. Broadcast-like services.

Each of these service types demand different network requirements that are determined by the types of traffic being sent and even the types of devices sending the traffic. For example, someone downloading cat videos will not have the same bandwidth or low latency requirements as a doctor in Los Angeles performing surgery virtually on a patient in Mumbai. The end-to-end notion of network slicing could be key to 5G's ability to effectively accommodate all of these disparate use cases.

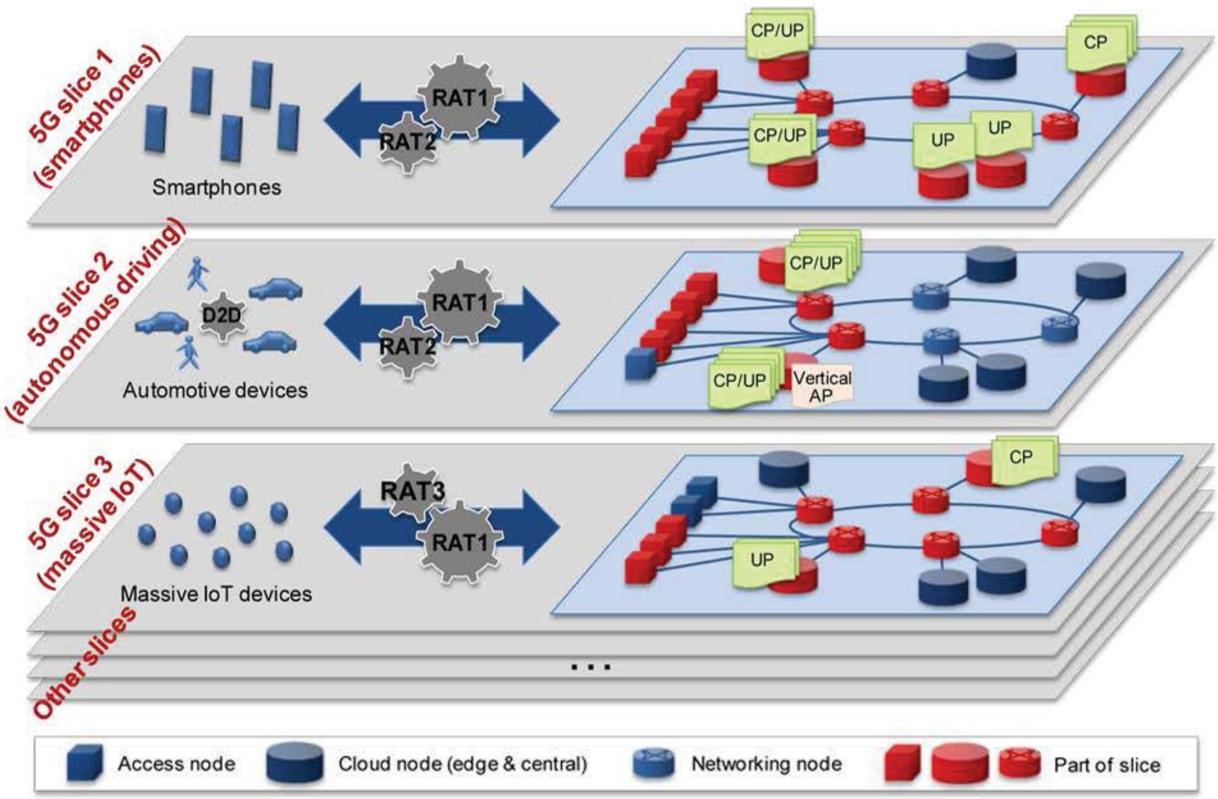
5G slice 1 (smartphones)
Smartphones
RAT1
RAT2
CP/UP
CP
CP/UP
UP
UP
5G slice 2 (autonomous driving)
D2D
Automotive devices
RAT1
RAT2
CP/UP
CP/UP
Vertical AP
5G slice 3 (massive IoT)
Massive IoT devices
RAT3
RAT1
CP
UP
Other slices
...
Access node
Cloud node (edge & central)
Networking node
Part of slice

FIGURE 1. 5G's incorporation of SDN and NFV would allow network slices to be created dynamically and deployed as needed to accommodate a variety of scenarios [2].

The Open Networking Foundation, in an April 2016 white paper, called out SDN's ability to support multiple services over a common architecture as a key enabler for network slicing. SDN also allows for slices to be created dynamically and deployed as needed to accommodate a variety of scenarios (see figure 1). Currently, 4G mobile networks prioritize traffic to get a similar effect, but with more rigidity and limitations. However, as 4G networks incorporate SDN and NFV, network slicing will become an option.

C-RAN

There are more than six million base stations (excluding small cells) deployed worldwide across approximately five million different cell sites serving close to four billion users. The surge in demand for connectivity has network operators searching for ways to shrink their network footprint, lower operational expenditures (OPEX), and still meet users' demand for access. C-RAN meets these requirements and has either been implemented or trialed by several operators including Verizon, AT&T, KT (South Korea), and China Mobile. Radio base stations currently depend on special purpose-built hardware deployed at the cell site. The baseband processing unit (BBU) is the part of the RAN that is responsible for managing the radio functions (or all functions that require an antenna). The BBU is one of the parts of the RAN that can be moved to a central location, creating a pool of BBUs to serve multiple base stations. C-RAN aims to centralize and virtualize baseband processing to reduce cell site costs and enable coordinated scheduling of resource blocks across a coverage area [6].

Figure 2 illustrates the evolution from the classic RAN model to a C-RAN setup. On the left, the classic model has the BBU deployed at the cell site connecting to the core network over IP/Ethernet transport. Any coordination between cell sites takes place over the X2 interface, which allows two sites to communicate. In a C-RAN architecture, the BBUs are pooled at a location away from the cell site. Pooling BBUs negates the need for the X2 interface as communications between cell sites now takes place internally. This is one of the reasons for the increased performance in C-RANs. The C-RAN model also makes updating the waveform and protocols easier as it only requires a software upgrade at the centralized BBU and not at each individual cell site

The move to a C-RAN architecture increases the flexibility of the network by allowing providers to, in theory, even change the types of RANs used—from 3G to 4G. For instance, in an area that has a mix of 3G and 4G users, operators can rebalance radio frequency resources by shifting more resources to 4G when 4G

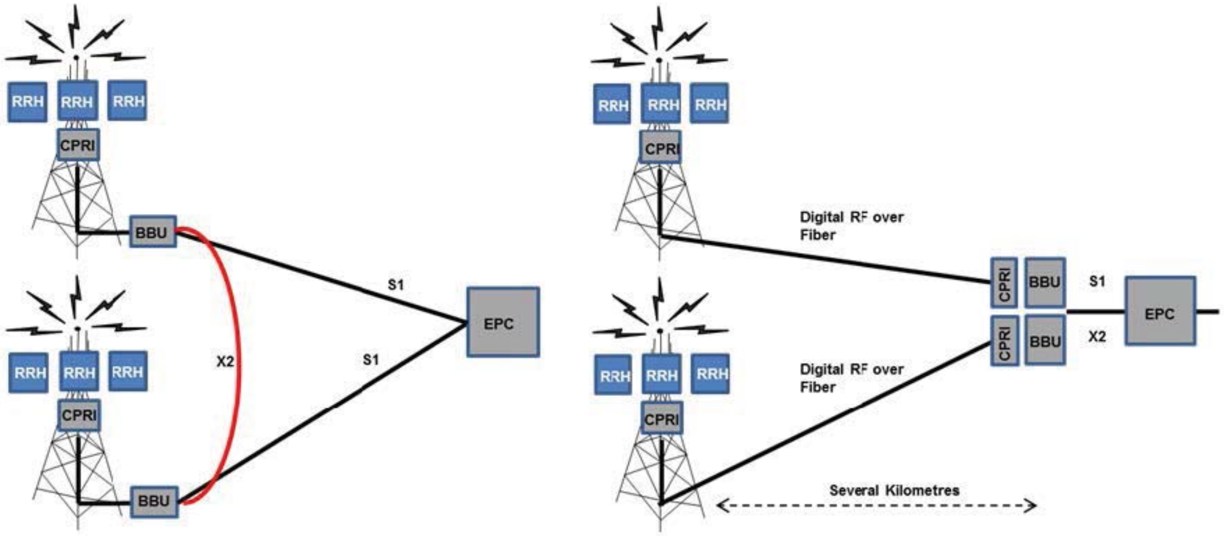

FIGURE 2. Unlike the traditional RAN model (on the left), C-RAN technology (on the right) pools the BBUs at a location away from the cell site offering an increase in performance and simpler upgrade path [6].

users are prevalent in the covered area. This shuffling of resources ensures that there is sufficient capacity for subscribers. C-RAN will also enable RAN-as-a-Service (RaaS), which will allow operators to rent RAN capacity to other operators.

Conclusion

On the face of it, virtualizing the 5G architecture seems like a necessary step towards preparing for the low latency requirements associated with 5G services. However, there are many uncertainties around 5G and a lack of visibility into what 5G will actually become through the standardization process, and at which phase different performance requirements will be supported or required. Phase II, the phase that is expected to meet International Mobile Telecommunication system for the year 2020 (IMT-2020) requirements, is expected in 2020. However, early versions or "pre-5G" offerings could be on the market before the final standard is approved. So, while these pre-5G solutions will have some 5G functionality, they will not field the full complement of improvements offered by an approved 5G system. For this reason, it is unclear to what degree virtualization techniques like C-RAN and network slicing will need to be implemented. It may come down to what the industry leaders in 5G implement in their "pre-5G" networks that will decide what a virtual 5G network will look like and when we can expect to see one.

References

[1] Farroha B, Everett JS, Uher JJ, Harper JR, Bridgland JK, Patton PM. "The status of 5G and the impact on current cellular architecture," 2016. College Park (MD): Laboratory for Telecommunications Sciences and The Johns Hopkins University Applied Physics Laboratory.

[2] Brown G. "5G use cases & candidate technologies," 2016. New York (NY): Heavy Reading.

[3] Hedman P. "Description of network slicing concept," 2016. NGMN Alliance. Version 1.0. Available at: https://www.ngmn.org/uploads/media/160113_Network_Slicing_v1_0.pdf.

[4] Marek S. "Dynamic network slicing is the 'secret sauce' behind 5G." SDxCentral. 2016 Aug 30. Available at: https://www.sdxcentral.com/articles/news/dynamic-network-slicing-secret-sauce-behind-5g/2016/08/.

[5] Paul M, Schaller S, Betts M, Hood D, Shirazipour M, Lopez D, Kaippallimalil J. "Applying SDN architecture to 5G slicing," 2016. Palo Alto (CA): Open Networking Foundation. Report No. TR-526. Available at: https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/Applying_SDN_Architecture_to_5G_Slicing_TR-526.pdf.

[6] Brown G. "Cloud RAN: New network architectures & industry structures," 2016. New York (NY): Heavy Reading.

5G and the Auto

Staff Writer

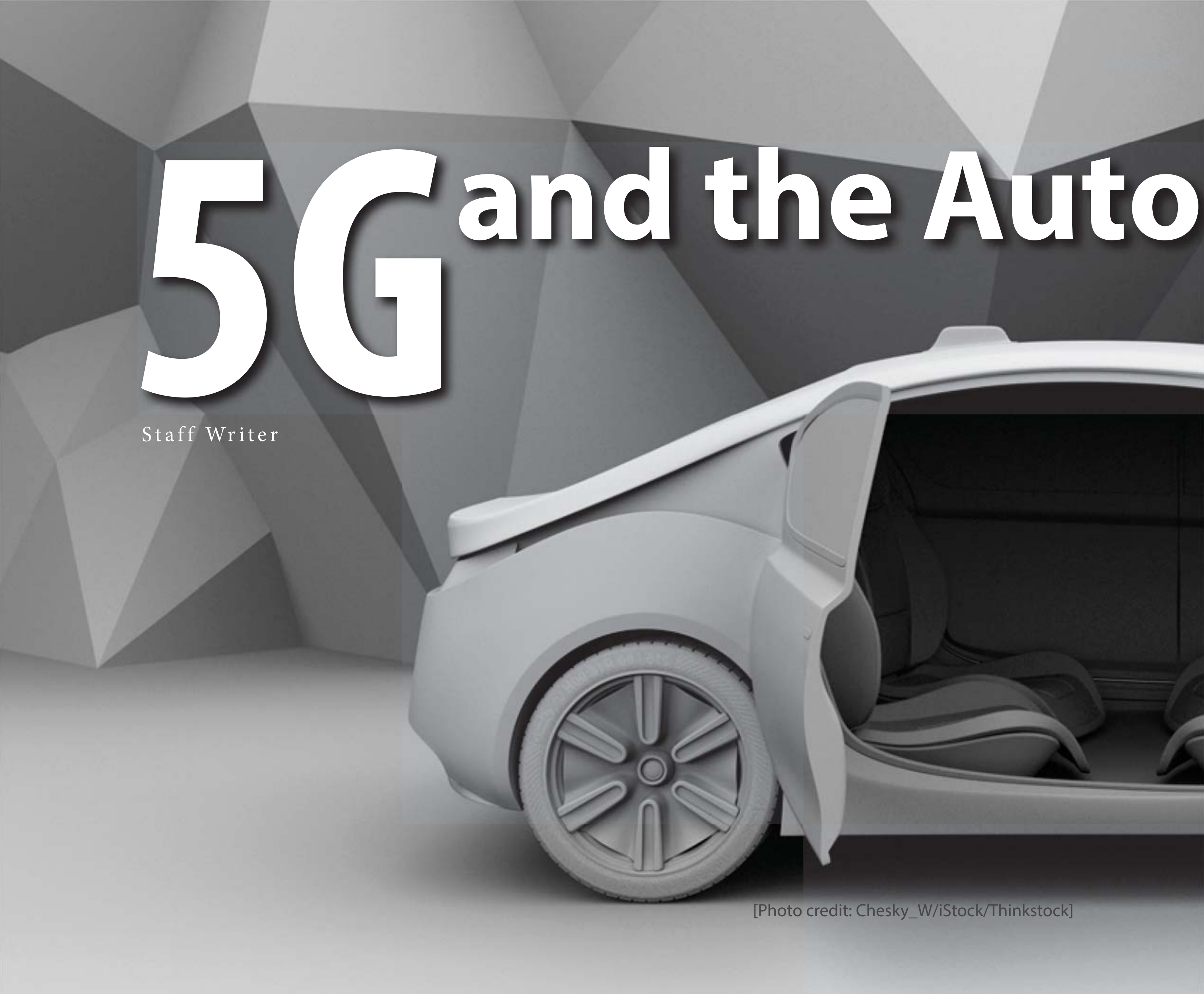

[Photo credit: Chesky_W/iStock/Thinkstock]

Industry stakeholders envision 5G as a key enabler that allows network connectivity in vehicles to shift in status from an optional accessory to a core feature that supports not only the individual vehicle, but also communication with other vehicles and sensors that inform traffic, parking, and navigation—while also ensuring passenger safety and data security. For consumers, a connected vehicle provides a growing number of features and services that make the driving experience safer, convenient, and less costly. 5G connectivity will enable information from in-car sensors to continuously be passed to the cloud. By sharing information and alerts about micro-level weather, road temperature, surface conditions, and violent breaking ahead, more efficient and consistent traffic flows will be achieved that reduce congestion and emissions. The aggregated and interpreted data will provide more informed driving information, as well as alert and activate onboard safety systems to prevent accidents [1].

motive Industry

Autonomous vehicles

Autonomous vehicles—also referred to as driverless or piloted vehicles—are predicted to hit the market by 2020, but stable 5G infrastructure will play a key role, according to a white paper by ABI Research [2]. Estimates indicate that 5G latency could be as low as one millisecond (ms) over-the-air, and 5 ms end-to-end, enabling the following automotive use cases:

- Broadband multimedia streaming (driverless vehicles as mobile living rooms).
- Cloud services for vehicle lifecycle management, apps, security, and over-the-air updates (cloud-to-vehicle).
- Capturing or uploading huge volumes of sensor data for real-time traffic, weather, parking, and mapping services (vehicle-to-cloud).
- Cooperative mobility: low latency vehicle-to-vehicle and vehicle-to-infrastructure for active safety and autonomy [redundancy for advanced driver assistance systems (ADAS)] [2].

The current ADAS being delivered on 2016 vehicles already facilitate SAE Level 1 (see figure 1) and are beginning to incorporate features that would be considered Level 2. However, Level 4 and 5 capabilities may not be that far away. In October 2015, Robot Taxi, a joint venture between Japanese mobile Internet company DeNA and vehicle technology developer ZMP, announced that it would offer driverless transportation to about 50 people in an area near Tokyo. Its goal is to commercialize the service by 2020, in time for the Tokyo Olympics [1]. While Robot Taxi is shooting for full automation (Level 5) in time for the games, it is likely that the use of such vehicles will initially be limited to shuttling passengers between Olympic venues. However, this venture is particularly notable because the technology is brand-agnostic and can be retrofitted to any vehicle [4].

Clearly, experimentation with autonomous vehicles is increasing, and 2020 as a date for some form of commercial implementation is certainly feasible from a technology perspective. Probably the best-known self-driving car project belongs to Google, which was started in 2009 and has clocked more than two million miles on public roads to date [5]. However, Google is not alone; Tesla, BMW, Audi, Mercedes, and most recently GM, have all showcased self-driving concept cars and demonstration projects. In mid-October 2016, Tesla announced that all cars currently being produced in their factories would include the hardware needed for full self-driving capability at a level of safety far greater than that of a human driver. However, the company added that the technology first needed to be tested and calibrated via “millions of miles of real-world driving” before the hardware would be activated on consumer vehicles [6]. In addition to traditional car manufacturers, companies such as Uber and Chinese search giant Baidu are also working on autonomous technology and self-driving cars [1].

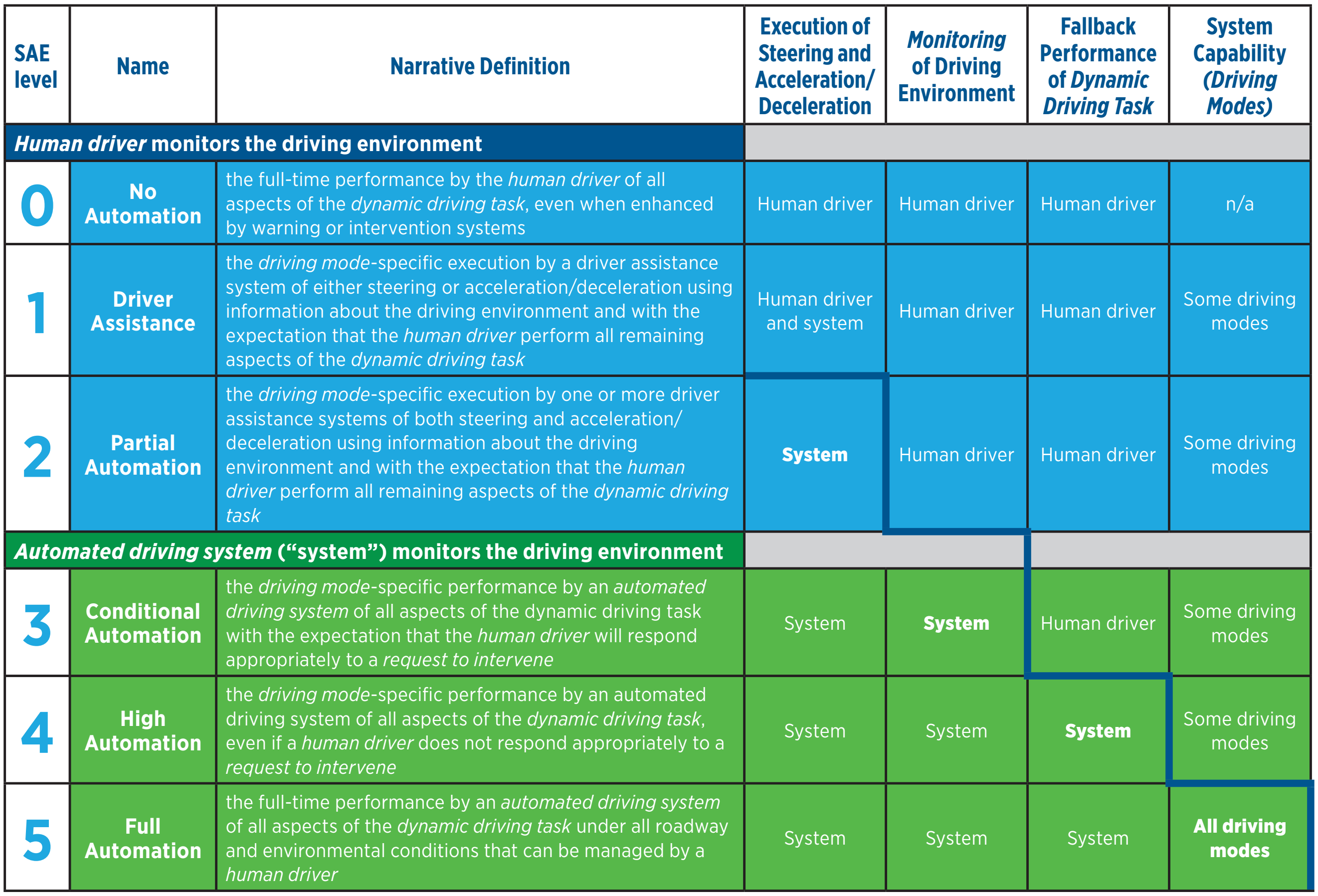

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-specific execution by one or more driver assistance systems of both steering and acceleration/ deceleration using information about the driving environment and with the expectation that the *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* (“system”) monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

FIGURE 1. SAE Levels of automation. The Society of Automotive Engineers (SAE) has defined levels of automation to clarify what role (if any) drivers have in operating a vehicle while a driving automation system is engaged. These levels are intended to establish a consistent framework that can be used across industries as the dialogue about autonomous vehicles continues. (Figure credit: SAE International J3016 [3].)

The motivations for creating an autonomous vehicle are beyond just technology. It's about reducing emissions through better fuel consumption, as well as addressing the demographic changes of an aging population that increase, rather than decrease, the potential for human error-induced accidents. It's also about leveraging the convergence of the shared economy and urban living, where young and old people no longer feel the need to own a car if there is a cost-effective and convenient alternative, such as Zip Car rentals, or on-demand ride-sharing services such as Uber [1].

Looking forward

Geographical coverage will be a key condition for 5G to have any relevance in the automotive sector. Initial 5G coverage can be supplemented by 4G and Wi-Fi connectivity on phones and other devices while the infrastructure is being built up, but these multimode, multiconnectivity solutions will not suffice for critical automotive use cases relying on the unique capabilities of 5G in terms of latency, reliability, and security [2]. Even once 5G is fully deployed, the adoption of self-driving technology will likely play out differently in the various markets in different regions since the forces shaping it are diverse at both the global and local level [1].

Government and industry cooperation

The continuous progression of ADAS-enabled cars and the gradual adoption of the autonomous vehicle will significantly reduce, and possibly eliminate, the number of crashes. This could, in turn, allow the removal of some regulations that relate to safety considerations, such as crumple zones, bumpers, and airbags. It also means that a review of laws relating to driving age, drunk driving, and speed restriction enforcement may be required, but not until all vehicles are compliant [1].

In September 2016, the US Department of Transportation (DOT) released the Federal Automated Vehicles Policy for highly automated vehicles (HAVs), or those intended to operate at Levels 3 to 5 as defined by SAE. The document—which is currently intended as guidance rather than formal policy—lays out standards for safe design, development, and testing of HAVs before they are commercially sold or operated on public roads. It also proposes guidelines for state governments to ensure a consistent national framework for regulation of motor vehicles with all levels of automated technology [7]. DOT's National Highway Traffic Safety Administration released additional nonbinding guidance in October 2016, outlining best cybersecurity practices for motor vehicle manufacturers and individuals and organizations involved in developing self-driving technology. The guidance aims to make cybersecurity a top priority for the automotive industry and proposes layered solutions to ensure that automated driving systems are designed to take appropriate and safe actions, even when an attack is successful [8].

A challenge yet to be addressed is that, historically, car manufacturers have completely controlled the design and development of vehicles. The advent of computers and software that effectively become the "mind" of the car means that manufacturers could lose control to technology and software companies, and yet still remain liable for any issues or catastrophes related to the car. It also remains to be seen how much and what type of data car manufacturers and network providers would be expected to share to improve overall safety and security of connected vehicles [1].

References

[1] Bell S. "IOT & the automotive industry," 2015. New York (NY): Heavy Reading.

[2] Bonte D, Hodgson J. "5G in automotive," 2016. Oyster Bay (NY): ABI Research.

[3] SAE International. "SAE International technical standard provides terminology for motor vehicle automated driving systems" [Press release]. 2 Oct 2014. Available at: http://www.sae.org/misc/pdfs/automated_driving.pdf.

[4] Collins K. "Driverless taxis could be the bullet train of the Tokyo 2020 Olympics." CNET. 22 Aug 2016. Available at: https://www.cnet.com/news/driverless-taxis-could-be-the-bullet-train-of-the-tokyo-2020-olympics/.

[5] Higgins T. "Google's self-driving car program odometer reaches 2 million miles." *The Wall Street Journal.* 2016 Oct 5. Available at: http://www.wsj.com/articles/googles-self-driving-car-program-odometer-reaches-2-million-miles-1475683321.

[6] Tesla. "All Tesla cars being produced now have full self-driving hardware" [Blog post]. 2016 Oct 19. Available at: https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware/?utm_campaign=GL_Blog_101916.

[7] US Department of Transportation. “Federal automated vehicles policy.” September 2016. Available at: https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016.

[8] National Highway Traffic Safety Administration. “Cybersecurity best practices for modern vehicles.” October 2016. Report No. DOT HS 812 333. Available at: http://www.nhtsa.gov/staticfiles/nvs/pdf/812333_CybersecurityForModernVehicles.pdf.

[Photo credit: moodboard/Thinkstock]

FROM LAB TO MARKET

News from the Technology Transfer Program

CREATING & GROWING ECONOMIC IMPACT

NSA's Technology Transfer Program (TTP) is igniting innovation through collaborations with industry, academia, and other government agencies that introduce NSA technologies to the commercial marketplace. By fulfilling the legislative mandate to share government-patented technologies with industry, NSA's TTP plays a role in strengthening national security by contributing to the nation's economic growth.

Technology Transfer Program
Office of Research & Technology Applications
NSA Research Directorate

tech_transfer@nsa.gov
www.nsa.gov/techtransfer
866-680-4539

2000–2014 REPORT

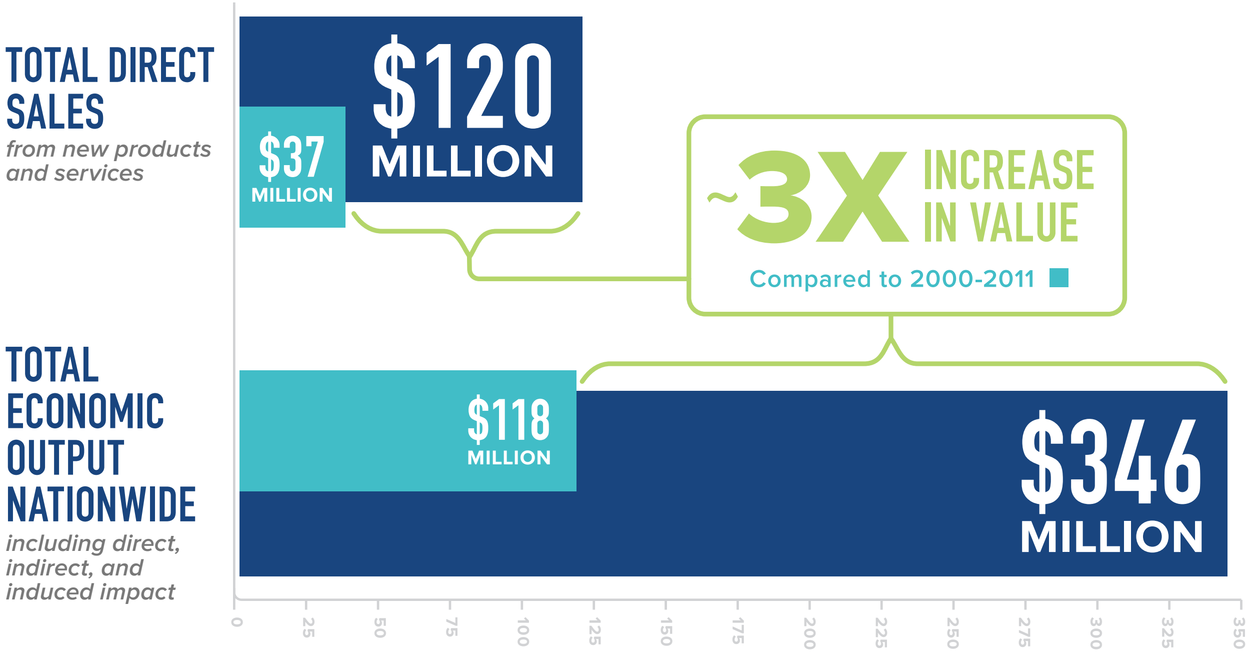

JOBS CREATED
or retained

1,330

SOURCE: *National Economic Impacts from DoD License Agreements With U.S. Industry*, TechLink and UC Business Research Division.

THE RESULTS

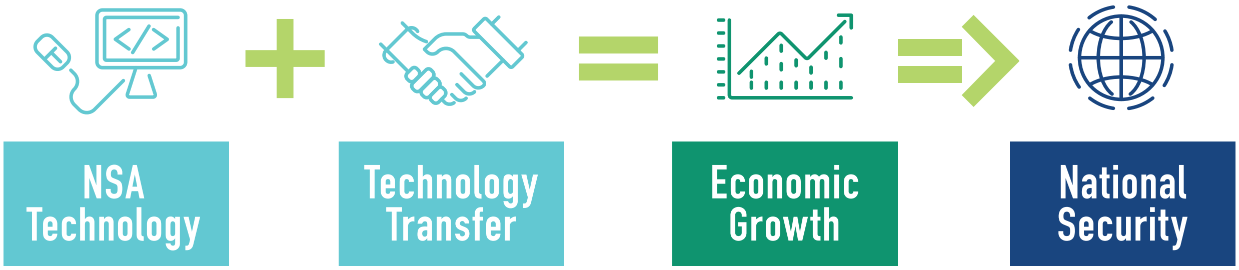

1G
2G
3

symmetry

MDPI

Article

IoT's Tiny Steps towards 5G: Telco's Perspective

Enida Cero [1], Jasmina Baraković Husić [1,2,*] and Sabina Baraković [3,4]

[1] BH Telecom, Joint Stock Company, Sarajevo, Franca Lehara 7, 71000 Sarajevo, Bosna i Hercegovina; enida.cero@bhtelecom.ba
[2] Faculty of Electrical Engineering, University of Sarajevo, Zmaja od Bosne bb, 71000 Sarajevo, Bosna i Hercegovina
[3] Faculty of Transport and Communications, University of Sarajevo, Zmaja od Bosne 8, 71000 Sarajevo, Bosna i Hercegovina; barakovic.sabina@gmail.com
[4] Department for Computer Engineering, American University in Bosnia and Herzegovina, Trg solidarnosti 10, 71000 Sarajevo, Bosna i Hercegovina
* Correspondence: Jasmina.BarakovicHusic@bhtelecom.ba; Tel.: +387-61-265-500

Received: 10 August 2017; Accepted: 28 September 2017; Published: 2 October 2017

Abstract: The numerous and diverse applications of the Internet of Things (IoT) have the potential to change all areas of daily life of individuals, businesses, and society as a whole. The vision of a pervasive IoT spans a wide range of application domains and addresses the enabling technologies needed to meet the performance requirements of various IoT applications. In order to accomplish this vision, this paper aims to provide an analysis of literature in order to propose a new classification of IoT applications, specify and prioritize performance requirements of such IoT application classes, and give an insight into state-of-the-art technologies used to meet these requirements, all from telco's perspective. A deep and comprehensive understanding of the scope and classification of IoT applications is an essential precondition for determining their performance requirements with the overall goal of defining the enabling technologies towards fifth generation (5G) networks, while avoiding over-specification and high costs. Given the fact that this paper presents an overview of current research for the given topic, it also targets the research community and other stakeholders interested in this contemporary and attractive field for the purpose of recognizing research gaps and recommending new research directions.

Keywords: 5G mobile communication; Internet of Things; applications classification; performance requirements; enabling technologies

---

1. Introduction

Telecom operators (telcos) had the most power and influence over business in the information and communication technology (ICT) industry during past decades. This dominance was the result of owning and provisioning communication infrastructures which nowadays have become more of a commodity than a luxury. Telco's revenue streams were mainly based on the provisioning of traditional services, such as voice calls and short message services (SMS). Recent work of regulatory agencies and the appearance of alternative service providers have led telecom operators to form the opinion that alternative service providers have conflicting interests and provide competitive services, thus decreasing telco's revenue from traditional services [1]. However, the latest econometric analysis presented in [2] has shown that telecom operators and alternative service providers have aligned interests and their collaboration could be favorable to both parties. The interests of alternative service providers and telcos are not inevitably conflicting, since the economic growth of the alternative service providers are positively correlated with telco revenues and vice versa.

The development of a digital society has changed the traditional value chain and introduced new issues for telcos as they seek a way how to monetize new digital services [3]. These services are diversified across various domains of industry, such as agriculture, construction, utilities, transportation, healthcare, finance, etc., and delivered to customers through connected devices building thereby the concept of Internet of Things (IoT).

IoT can be defined as an interconnected network of things/objects that are able to interact with each other and cooperate with other things/objects through wireless and wired connections in order to create new services/applications for the benefit of society. In this way, IoT brings almost limitless benefit's, which have the potential to radically change our daily life by saving time and resources while creating possibilities for growth and innovation [4]. IoT has disruptive potential in almost every application domain, which has created many challenges that need to be faced when implementing IoT-based solutions, such as prevailing over obstacles generated by the fragmentation of IoT, both in terms of application domains and in terms of technologies. Therefore, recognizing all potential IoT applications while having in mind the development of technology and the requirements of individuals, businesses, and society as a whole is quite challenging.

Telcos now have the opportunity to seize a share of the value that is generated by IoT implementation. The size of this share will depend on telco's role in the value chain that ranges from being a traditional provider of communication infrastructure to being an end-to-end solution provider [3]. In order to monetize IoT, telcos will have to address many challenges which can be summarized as follows: (1) strategic challenges relating to decisions on future directions; (2) business challenges relating to successful management, investment, partnerships; and (3) technical challenges relating to changing connectivity and performance requirements. This paper focuses on the technical challenges since connectivity and performance requirements of IoT objects/things cannot be fulfilled using existing cellular networks that limit numerous IoT applications. In order to overcome issues associated with the current cellular networks, new types of technologies are being introduced leading towards the fifth generation (5G) network [5–7].

The architecture of the 5G network has to seamlessly integrate the requirements of diverse IoT applications: from delay-sensitive video applications to ultra-low latency, from high-speed entertainment applications in a vehicle to mobility on demand for connected objects and from best effort applications to reliable and ultra-reliable ones for health and safety [8]. A full understanding of emerging IoT applications and the variability of their performance requirements can serve telecom operators as input for specifying 5G enabling technologies. These technologies should be flexible and scalable to meet the aforementioned requirements. Since IoT applications sometimes demand extreme requirements, the 5G network must simultaneously satisfy all of them, which can lead to over-specification and high cost. In order to avoid this, telecom operators first have to adequately classify IoT applications to facilitate the selection of 5G enabling technologies being capable to efficiently meet their performance requirements.

The 5G classification concept includes three different service classes [9], i.e., (1) extreme mobile broadband (xMBB); (2) massive machine-type communications (mMTC); and (3) ultra-reliable machine-type communications (uMTC). Nevertheless, this classification concept can be considered insufficient to properly select 5G enabling technologies to meet the diverse requirements of IoT applications. Furthermore, it will be challenging to classify the emerging IoT applications given that there is many criteria for their categorization. The definition of the IoT domain in many cases overlaps with the definition of IoT application, which may at the same time belong to another IoT domain. For example, Smart Buildings may either be considered as a standalone IoT application or as an element needed to form a Smart City application domain. IoT applications are usually classified according to spheres of human life [4,10–12] or performance requirements [9,13–15]. However, these classifications are not suitable for telecom operators which have to fulfil the performance requirements of particular IoT applications, since it is difficult to choose the most important one from a class containing a wide range of IoT applications.

Telecom operators can easily specify performance requirements for current communication services since traffic patterns generated by these services are driven by predictable activities, such as making calls, receiving email, surfing the web, and watching videos. However, traffic patterns generated by emerging IoT applications are driven by less predictable activities. Hence, the idea is to determine a common set of activities based on customer service requests which describe the reasonably foreseeable function/purpose of emerging IoT applications. Then these activities can be used by telcos as a new classification criterion of IoT applications. The assumption is that such a classification can help telecom operators meet performance requirements of future IoT applications and select appropriate 5G enabling technologies. Therefore, the aims of this paper are to classify IoT applications according to relevant activities, specify and prioritize the performance requirements of each IoT application class, and consider enabling technologies used to accomplish specified demands on the radio access part of 5G networks, all according to telco's perspective. In addition, the intention is to provide a literature review which covers 258 references dominantly published in the period between 2012 and 2017. As this paper provides a state-of-the-art review of IoT applications, their performance requirements and 5G enabling technologies, it can also be of use to the research community and other stakeholders interested in this contemporary and attractive field in order to recognize the research gaps and recommend new research directions.

The remainder of the paper is thusly organized. In order to answer research questions, Section 2 describes the methodological approach in regard to the conducted research. Section 3 identifies a common set of activities as a new classification criterion for IoT applications and assigns them the relevant 5G service classes. This facilitates the determining of performance requirements of activity-based IoT application classes, which are summarized and prioritized in Section 4. The enabling technologies used to accomplish these requirements in 5G networks are contemplated in Section 5. Finally, Section 6 discusses the outcomes of the paper by clarifying the contribution of our research in identifying open issues for future work, while Section 7 concludes this paper.

2. Research Approach and Design

The main objectives of this paper can be summarized as follows: (1) to identify activities relevant to IoT customer service requests and use them as a new classification criterion of IoT applications; (2) to specify and prioritize performance requirements of such IoT application classes; (3) to analyze relevant literature in order to provide insight into 5G radio technologies used to fulfil the requirements of activity-based IoT application classes; (4) to recognize research gaps and directions.

The research questions posed in this study are: (1) Which activities can be used as classification criteria for IoT applications? (2) What are the performance requirements of such IoT application classes? (3) Which enabling technologies in the radio access part of the 5G network can be used to meet the requirements of IoT application classes?

Investigation of different criteria for the classification of IoT applications underlines to what extent various classification concepts have been covered by existing literature and provides a basis for the introduction of a new classification criterion called the activity. We thus surveyed and compared different studies contributing to the understanding of the scope and classification of a wide range of IoT applications. This approach has been motivated by the challenge to propose an activity-based classification of IoT applications, the determination and prioritization of performance requirements of such IoT application classes and the review of 5G enabling technologies used to meet them. These findings can be utilized by telecom operators and other interested parties (e.g., the research community, software network function providers, network infrastructure manufactures, etc.) depending on their interest and potential to utilize these findings towards 5G implementation and commercialization.

The methodological approach to research conducted in this paper is illustrated in Figure 1 and includes four phases.

Phase I included searching, identifying, and extracting papers from three categories, i.e., IoT in 5G service classification, IoT in 5G performance requirements and IoT in 5G enabling technologies. The keywords used to search relevant scholar databases are shown in Figure 2. This search has resulted in the selection of 258 recently published papers, which can be categorized into three groups (i.e., review papers, technical papers, research effectiveness), as shown in Table 1. Review papers summarize the status of knowledge and outline future directions in a given area of research. Technical papers describe the process, progress, or result of research, whereas research effectiveness directly provides an answer to the research question raised for this study. The reference distribution by year of publication is shown in Figure 3a, while the total percentage of references per publication type is presented in Figure 3b.

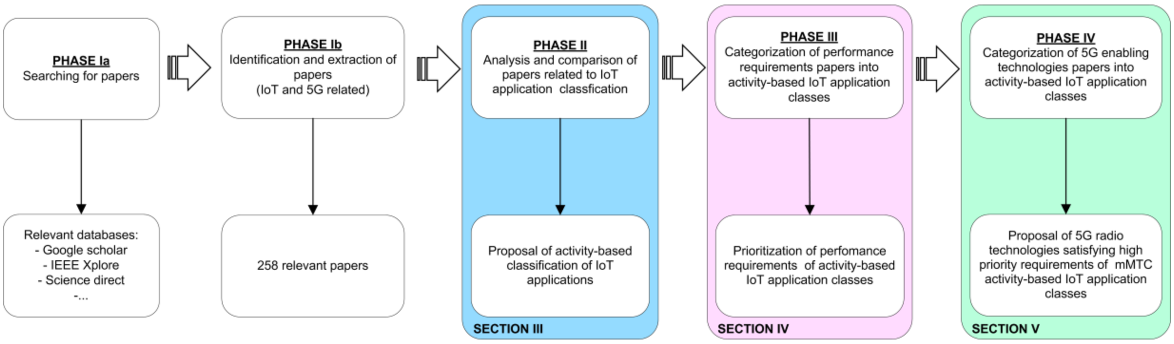
PHASE Ia
Searching for papers
PHASE Ib
Identification and extraction of papers
(IoT and 5G related)
PHASE II
Analysis and comparison of papers related to IoT application classfication
PHASE III
Categorization of performance requirements papers into activity-based IoT application classes
PHASE IV
Categorization of 5G enabling technologies papers into activity-based IoT application classes
Relevant databases:
- Google scholar
- IEEE Xplore
- Science direct
-...
258 relevant papers
Proposal of activity-based classification of IoT applications
Prioritization of perfomance requirements of activity-based IoT application classes
Proposal of 5G radio technologies satisfying high priority requirements of mMTC activity-based IoT application classes
SECTION III
SECTION IV
SECTION V

Figure 1. Methodological approach to research.

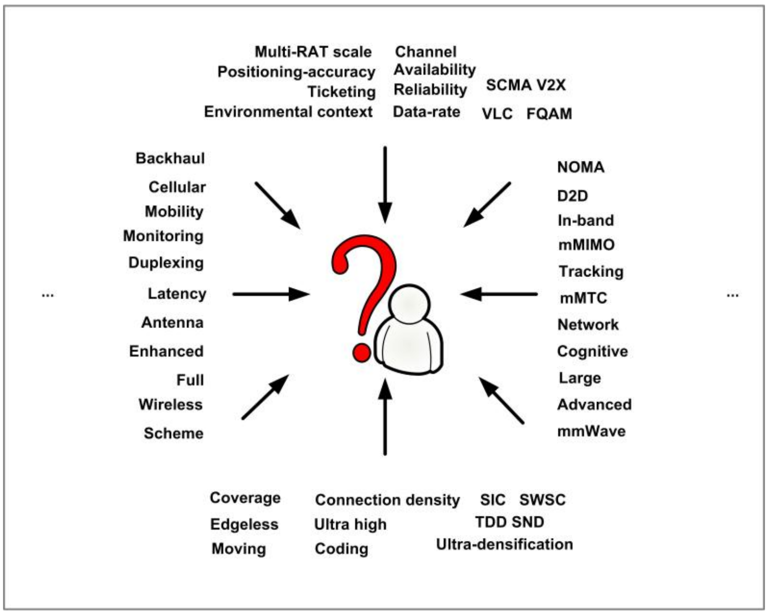
Multi-RAT scale
Positioning-accuracy
Ticketing
Environmental context
Channel
Availability
Reliability
Data-rate
SCMA V2X
VLC FQAM
Backhaul
Cellular
Mobility
Monitoring
Duplexing
Latency
Antenna
Enhanced
Full
Wireless
Scheme
NOMA
D2D
In-band
mMIMO
Tracking
mMTC
Network
Cognitive
Large
Advanced
mmWave
...
...
Coverage
Edgeless
Moving
Connection density
Ultra high
Coding
SIC SWSC
TDD SND
Ultra-densification

Figure 2. Keywords for searching relevant databases.

Table 1. Reference categorization.

Reference Type	Reference Number
Review paper	[1,2,10,12,14–88]
Technical paper	[8,89–197]
Research effectiveness	[3–7,9,11–13,29,52,198–258]

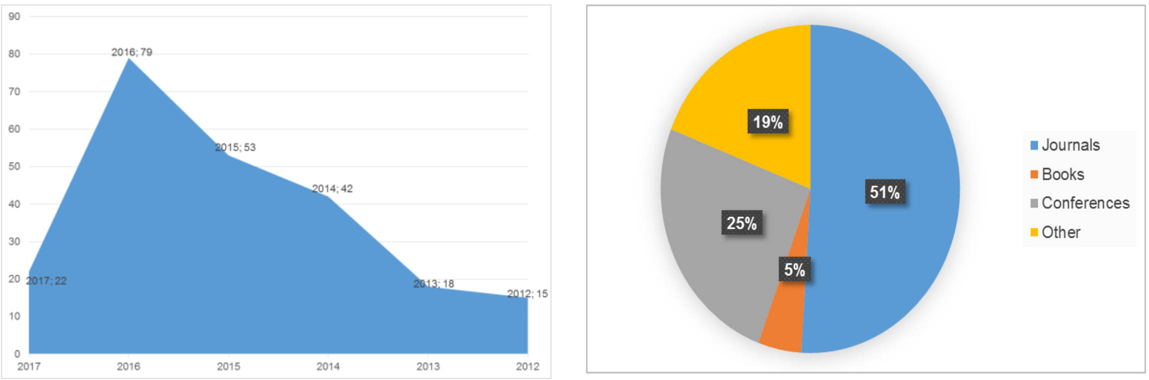

(**a**) (**b**)

Figure 3. Reference distribution by (**a**) year of publication; (**b**) type of publication.

Phase II included the analysis and comparison of papers related to the category of IoT in 5G service classification. Papers were first sorted according to the existing classification criteria of IoT applications in order to underline their deficiencies from telco's perspective. Since emerging IoT applications generate traffic driven by less predictable activities, specifying their performance requirements and enabling technologies is not an easy task from telco's point of view. Therefore, we have proposed a common set of activities to serve as classification criteria of IoT applications. Four activities were selected based on IoT customer service requests denoting the function/purpose of existing IoT applications. According to the authors' knowledge, existing IoT applications can be grouped around following activities: ticketing, tracking, monitoring, and managing/controlling. Such a classification makes it easier for telecom operators to associate particular IoT applications to the relevant 5G service class. Finally, Phase II has resulted in the proposal of an activity-based classification of IoT applications which was associated with the 5G service classification.

Phase III included the analysis of papers related to the category of IoT in 5G performance requirements. Papers were grouped according to eight key performance indicators, i.e., data rate, mobility, latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance requirements were prioritized indicating high, medium, and low importance of each requirement for specific activity-based IoT application classes proposed in Phase II. Finally, Phase III resulted in the identification of highly important performance requirements for each activity-based IoT application class which can be utilized by telcos to identify 5G enabling technologies used to meet them.

Phase IV provided a review of papers from the category of IoT in 5G enabling technologies. Papers were first sorted according to eight technological groups used in the radio access part of 5G networks, i.e., wide and flexible bandwidth technology, advanced modulation and coding, duplexing, multiple access and waveform, advanced interface management, access architecture related radio technologies, energy related technologies, and other technologies. These technologies have been discussed in terms of their possibility to meet the high priority performance requirements of mMTC activity-based IoT application classes, since this area is more mature from telco's perspective. The results of Phase III were reflected in a proposal of technologies in the radio access part of 5G networks that can be used to meet the high priority performance requirements of the mMTC activity-based IoT class, as well as the identification of research gaps and directions for future work.

3. IoT in 5G Service Classification

As mentioned in the Introduction, various challenges awaiting telcos can be identified on the basis of the role that they will play in the IoT value chain. Regardless of this role, telcos will face many technical challenges, such as the necessity for global deployment, the need for rapid infrastructure scaling, unpredictable IoT application behavior, etc. Even now, existing IoT applications accelerate the growth rate of traffic driven by less predictable activities which require new strategies towards 5G networks. These networks will be the key enabler for IoT applications by providing a unified infrastructure capable of meeting the high variability of their performance requirements [8,16]. Various stakeholders have recently described what 5G networks may be and grouped the major types of 5G services into three different classes: (1) xMBB—Requiring extremely high data rates and low-latency communication; (2) mMTC—Requiring scalable connectivity for an extremely large number of devices; and (3) uMTC—Requiring ultra-reliable low-latency and resilient communication [198]. This 5G service classification can be used by telcos to identify performance requirements of various IoT applications, and, as such, was used in this research study. However, emerging IoT applications will have extreme requirements, so an analysis based solely on 5G service classifications may not be sufficient. Therefore, telcos need a more precise classification of IoT applications which is an important precondition for meeting their diverse requirements.

IoT covers human-to-human (H2H), human-to-machine (H2M) and machine-to-machine (M2M) communication, which will be the main driving force towards 5G networks. In addition, the terms M2M communication (M2MC) and machine type communication (MTC) are used interchangeably as in [17–19,89–93,199,200]. Although IoT is a broader concept which evolves from M2M, this paper assumes that IoT and M2M are synonyms as in [94].

3.1. Existing Classifications of IoT Applications

In order to meet the requirements of a wide range of IoT applications, they have to be classified in an appropriate manner. The existing approaches to the classification of IoT applications are summarized in Table 2. IoT serves different user categories, including individuals, businesses, and society as a whole, and may span through a broad range of application domains [4,10–12], such as transportation and logistics, healthcare, smart environment, personal and social, futuristic applications, food sustainability, smart living, smart manufacturing, smart energy, smart city, etc. These application domains are created to be human-centric, which means that they cover different domains of human life. IoT applications belonging to these domains have diverse requirements for 5G networks, and should not be treated equally. Therefore, the acceptance of these domains as a classification criterion raises the issue of assigning IoT application to a particular domain. Moreover, the emergence of new applications of IoT may require defining new application domains, which makes this classification inappropriate from telco's perspective.

The M2M applications may be classified by the mobility and the amount of dispersion that needs to be supported into four categories [13]: (1) fixed and concentrated; (2) fixed and dispersed; (3) mobile and concentrated; and (4) mobile and dispersed. However, the mobility and amount of dispersion present very rough classification criteria and cannot meet the precise network requirements of individual IoT applications.

Additionally, M2M applications can be grouped according to delay tolerance into four categories [14]: (1) elastic (delay tolerant); (2) hard real-time (delay constraint); (3) delay-adaptive (delay sensitive but tolerant); and (4) rate-adaptive application (adjust their transmission rates according to available radio resource). However, the main drawback of this classification is the lack of consideration of other IoT application requirements except delay tolerance, although its importance as a classification criterion has been recognized.

According to data reporting mode, the M2M applications can be classified into five categories [91]: (1) time-driven; (2) query-driven; (3) event-driven; (4) continuous-based; and (5) hybrid-driven. This classification is specific for the former IoT concept, which is narrower in nature than the definition adopted in this study.

Another study considered the reliability, availability, and end-to-end latency in order to classify IoT applications into two groups [15]: (1) monitoring-based and mission-critical; (2) monitoring-based and non-mission critical; (3) control-oriented and mission-critical; and (4) control-oriented and non-mission critical. Monitor-based IoT applications periodically collect sensor data from smart objects and transmit them. The majority of monitoring-based IoT applications are not mission-critical. Control-oriented IoT applications use sensor data to control actuators in real-time, and rely on mission critical communication. This classification is based on multiple criteria and represents the precursor of the Mobile and wireless communications Enablers for Twenty-twenty Information Society (METIS) 5G service classification.

Table 2. Internet of Things (IoT) application classification—summary.

Criteria for Classification	IoT Classes	References
Domains	(1) Transportation and logistics (2) Healthcare (3) Environment (4) Personal and social (5) Futuristic applications (6) Food/water monitoring (7) Living (8) Manufacturing (9) Energy (10) Building (11) Industry (12) City (13) Security and safety (14) Communication (15) e-society (16) Vehicular (17) Sport and leisure	[4,10–12]
Mobility and amount of dispersion	(1) Fixed and concentrated (2) Fixed and dispersed (3) Mobile and concentrated (4) Mobile and dispersed	[13]
Delay tolerance	(1) Elastic (2) Hard real time (3) Delay-adaptive (4) Rate-adaptive	[14]
Data reporting mode	(1) Time-driven (2) Query-driven (3) Event-driven (4) Continuous-based (5) Hybrid-driven	[198]
Reliability, availability, and end-to-end latency	(1) Monitoring-based and mission critical (2) Monitoring-based and non-mission critical (3) Control-oriented and mission critical (4) Control-oriented and non-mission critical	[15]
Characteristics and requirements	mMTC and uMTC	[12]

Legend: IoT (Internet of Things), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication).

Although it spans through a wide range of different applications, MTC can be divided in two main categories, i.e., massive and ultra-reliable MTC, which depend on their characteristics and requirements [89]. As mentioned above, this categorization is a part of 5G service classification which was used in this study as a basis to identify performance requirements of IoT applications and will be described in more detail in Section 4. Massive MTC (mMTC) typically involves a very large number of devices (tens of billions [9]), such as sensors, actuators, and similar devices [12], different in complexity and cost [9], and with varying quality of service (QoS) requirements. These devices should be of very low cost with very low energy consumption, enabling very long battery life [12]. At the same time, the amount of data generated by each device is normally very small, and very low latency is not a critical requirement [12]. Ultra-reliable MTC (uMTC) requires very high reliability and availability, and very low latency [9,12]. Low device cost and energy consumption are not as critical as they are for mMTC applications [12], and the number of devices and required data rates are relatively low [198].

The vision of a pervasive IoT requires the integration of various domains into a single and unified domain. It addresses the enabling technologies needed for these domains while taking into account the elements that form the third dimension like security, privacy, trust, and safety [11]. The current classification criteria and IoT application classes do not clearly differentiate IoT domains from IoT applications. IoT domains are usually viewed as a specific area of IoT applications that support a huge variety of use cases across industries impacting businesses and customers. Therefore, IoT domains overlap with IoT applications within existing IoT classifications. Moreover, IoT classification based on certain QoS parameters is difficult to apply from telco's perspective, because it is challenging to choose the most important parameter for the wide range of IoT applications. However, these problems need to be solved for upcoming 5G networks.

3.2. Activity-Based Classification of IoT Applications

The aforementioned issues motivated us to propose a new approach to the classification of IoT applications. This approach is based on "**the activity**", which primarily characterizes specific IoT application. In this sense, the activities are defined as new classification criteria, which denote a main function/purpose of specific IoT application observed from telco's point of view. Telecom operators observe the technical challenges of IoT through systems, tools, devices, and platforms because their availability and integration complexity determines the opportunity to capture a share of the value that is generated by IoT implementation. Based on the literature review, it was found that the most commonly mentioned terms in this sense were related to: (1) ticketing system [201]; (2) monitoring: devices [199,201–204]/services [199,201,202,205,206]/tools [207]/systems [11,12,199,201,205,206,208–210]/data [95,201]/framework [12,201]/solutions [20,205]/networks [211]/process [7]/activity [203,204,207]; (3) tracking technologies [4]/devices [204]/systems [204]/applications [11,210]/system [201]/services [202,205]; (4) managing/controlling applications [15,210,212]/operation [211]/services [4,208]/tools [208,211]/devices [204]/system [95,198,204,210]/concepts [211]/solutions [95]/platforms [95,199]. This has led us to identify four activities, i.e., ticketing, monitoring, tracking, and managing/controlling, as new classification criteria of IoT applications. According to the authors' knowledge, it was found that these activities can cover reasonably foreseeable functions/purposes of IoT in existing application domains. Using activities as classification criteria, telcos can specify IoT application performance requirements more easily, which are in that case dictated by the specific activity, not by the IoT application domain, and determine enabling technologies in the radio access part of 5G networks.

In addition, our classification approach allows better service differentiation and service delivery closer to customer expectations. According to BH Telecom's experience, customers typically come up with the following service requests when it comes to IoT: (1) they need to track their products and determine products' distribution across different regions based on their own data analysis; (2) they want to monitor their products since they do not have a department for supervision and analysis; (3) they need information panels about their outlets and working time to be downloaded by a scanning tab-ticket; (4) they want to manage/control their products according to market needs. This way of expressing customer needs has inspired us to propose the activity-based classification of IoT applications as it allows telcos to define performance requirements more precisely, and thereby improve the customer experience.

The activity-based classification of IoT applications is presented in Table 3. Each activity is associated with the application domain where that activity may be applied. The requirements of identified activities and associated application domains are then mapped to 5G service classes in order to be further able to identify their performance requirements and enabling technologies necessary to meet them. For example, according to Table 3, the managing/controlling activity can be realized in several domains, such as healthcare, food, energy, transportation and logistics. For each domain, we have identified an application example, such as remote surgery [213] in the healthcare domain, food processing facilities [96] in the food domain, energy distribution [12] in the energy domain and traffic/driving [12,213] in the transportation and logistics domain. According to diverse performance requirements of these IoT applications [214], managing/controlling activity in healthcare, food, transportation, and logistic domains can be associated with the uMTC service class, while the energy domain can be mapped to the mMTC service class. A more detailed description of performance requirements for each activity-based class of IoT applications is provided in Section 4, while in Section 5 it is discussed how these requirements can be accomplished in the radio access part of 5G networks.

Table 3. Activity-based classification of IoT applications.

Activity	Domain Examples	5G Service Classification	Application Examples
Ticketing	Smart Transportation and Logistics	mMTC	POS Terminal [4]
Monitoring	Smart Healthcare	uMTC	Health condition [97]
	Smart Buildings	uMTC	Structures (buildings, tunnels, etc.) [12,25,213]
	Smart Buildings/Smart City	mMTC	Parking spaces
	Smart Buildings/Smart Environment	mMTC	Home video [12]
	Smart food/water monitoring	mMTC	Food growth condition [10]
Tracking	Smart Healthcare/Sport and Leisure	mMTC	Medical assets, wearables [200,213]
	Smart Transportation and Logistics	mMTC	Transport fleet [200,213]
	Smart Industry/Social Networking	mMTC	Shipping of products
	Smart Healthcare	mMTC	People in science museum
Managing/ controlling	Smart Healthcare	uMTC	Remote surgery [213]
	Smart food/water monitoring	uMTC	Food processing facilities [96]
	Smart Transportation and Logistics	uMTC	Traffic, driving [12,213]
	Smart Energy	mMTC	Energy distribution [12,98]

Legend: 5G (Fifth Generation), POS (Point Of Sale), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication).

As such, the activity-based classification of IoT applications can be used for the creation of new business models, which represent the stakeholder's plan to generate revenue and make a profit from operations, and thereby include many components and functions of the business [21]. However, there is no common opinion which components constitute a business model. The business model architecture can be illustrated by four dimensions [22]: (1) who, identifying the definition of the target customer as one central dimension in designing a new business model; (2) what, describing what is offered to the target customer; (3) how, referring to the construction and distribution of the value proposition; (4) value, explaining why the business model is financially viable. Answering these four questions allows the creation of IoT business models. According to this business model definition, the activity-based classification of IoT applications affects the who and the how dimensions in the following manner. In terms of the who dimension, it directly allows customer segmentation according to considered activities as a way to express their requirements (e.g., customers that require tracking or monitoring of their products). Being aware of the current customer requirements, the proposed classification of IoT applications around four activities (i.e., ticketing, monitoring, tracking, and managing/controlling) can be considered complete. At the same time, the activity-based classification of IoT applications is flexible since additional activities as classification criterion can be concerned with emerging customer requirements. On the other hand, the activity-based classification of IoT applications indirectly impacts the how dimension of new business models which among others includes relevant resources and capabilities in the focal stakeholder's internal value chain. This indicates that the proposed activity-based classification of IoT applications can be easily applied to IoT business models.

Various categories of IoT business models can be identified according to eight IoT architectural layers [23], i.e., collaboration and processes layer, application layer, service layer, abstraction layer, storage layer, processing layer, network communication layer, physical layer. Along with IoT architectural layer, the IoT business model needs also to address the IoT value proposition [24] and

IoT stakeholders that can participate in more than one layer. Based on their role in the IoT ecosystem, telecom operators usually take a part in the network communication layer. In this regard, they follow four evolutionary business models [23]: (1) selling connectivity services only; (2) selling third-party products; (3) selling internal products; (4) providing a broad menu of IoT products. Each of these IoT business models includes traditional telcos strength, i.e., connectivity which determines the how dimension of the IoT business model. In this sense, our activity-based classification of IoT applications indirectly allows the construction and distribution of performance as IoT value proposition using enabling technologies as IoT resources in telco's internal value chain.

3.3. Summary of IoT in 5G Service Classification

The previous discussion has shown that the classification of IoT applications is a complex task due to their numerosity and diversity. The existing classifications of IoT applications pose some drawbacks that can be summarized as follows. The domain-based classification does not allow clear differentiation between IoT domains and IoT applications due to either an imprecise classification criterion or the diversity and unpredictability of IoT applications. On the other hand, QoS-based classifications need to identify the common and most important performance metric for a broad range of IoT applications. Therefore, on the basis of the literature review and IoT customer service requests, we have proposed a new approach to the classification of IoT applications. It is based on the activity as new classification criterion, which denotes a main function/purpose of specific IoT application observed from telco's point of view. According to the authors' knowledge, four activities, i.e., ticketing, monitoring, tracking, and managing/controlling, have been identified to cover reasonably foreseeable functions/purposes of existing IoT applications. The resulting activity-based IoT application classes have been associated with 5G service classes in order to determine and prioritize their performance requirements as described in next section. Finally, the proposed classification of IoT applications was discussed in terms of its completeness, flexibility, and applicability to new business models.

4. IoT in 5G Performance Requirements

This section provides an insight into performance requirements of activity-based classes of IoT applications proposed in Section 3. The analysis is based on eight key performance indicators identified in [5,207] as shown in Table 4: data rate, mobility, latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance indicators are usually well described for specific IoT applications. However, one of the main challenges of 5G is to support a variety of performance requirements for numerous IoT applications in a flexible, reliable, and cost-effective way [15]. Hence, there is a need for a comprehensive understanding of these requirements for activity-based IoT application classes. The 5G service classification defined the performance requirements for mMTC and uMTC [9]. We have assigned these requirements to the activity-based IoT application classes introduced in Section 3 for the purpose of proposing priorities of each requirement for a specific class. Three levels of priorities (high, medium, low) are associated with the performance requirements of activity-based classes of IoT applications as shown in Table 4. The prioritization of performance requirements is inspired by analysis of related work undertaken in [213]. Some activity-based classes of IoT applications may demand optimization of multiple performance requirements. Table 4 illustrates the main differences between activity-based IoT application classes, and therefore, the need for a 5G network that enables support of optimal configurations for a variety of, sometimes opposite, requirements. For example, the mMTC tracking-based IoT application class requires support for high mobility, high positioning accuracy, and high connection density, while the mMTC monitoring-based IoT application class also requires high connection density but low mobility and low positioning accuracy. A more detailed description of considered performance requirements is contained in the following subsections.

Table 4. Activity-based classes of IoT applications—performance requirements.

5G Service Classification	Activity	User Experienced Data Rate [Gbps] Outdoor: 0.1 [5] Indoor: 1 [5]	Mobility [km/h] Required: 500 [5]	Latency [ms] Control Plane: 50 [5] User Plane: 1 [5]	Connection Density [Connections/km^2] Required: 10^6 [5]	Reliability [%] Required: 99.999 [207]	Positioning Accuracy Required: A Few cm [5]	Coverage/ Availability [%] Required: 99.999 [207]	Energy Efficiency [bits/J]
mMTC	Ticketing	L to M	L	L	H	M	H	H	H
	Tracking	M	H	L	H	M	H	H	M to H
	Monitoring	M to H	L	L	H	L to M	L	H	M
	Managing/controlling	L	L	L	H	M	L	M to H	M
uMTC	Monitoring	L	L to M	H	H	H	M to H	H	M
	Managing/controlling	L	M to H	H	L	H	H	M to H	M

Legend: 5G (Fifth Generation), mMTC (massive Machine Type Communication), uMTC (ultra-reliable Machine Type Communication), L (Low), M (Medium), H (High).

4.1. Data Rate

Data rate is the most important evaluation factor for generations of wireless communication networks [6]. It is contemplated in two ways: (1) peak data rate—defined as the maximum achievable data rate by the user; and (2) minimum guaranteed user data rate—defined as the minimum experience data rate by the user [16]. New mobile technologies are primarily driven by users' needs for higher data rates, as discussed in [8,16,99,215,216]. The expected values in 5G networks are 10 Gbps for minimum peak data rate and 100 Mbps as minimum guaranteed user data rate [5]. High data rate requirements are mainly posed by xMBB related use cases [9] like hotspots [214] or dense urban areas where in 95% of locations and time experience data rate by user should be 300 Mbps in downlink and 60 Mbps in uplink [100]. High data rate is also important in some activity-based classes of IoT applications, as shown in Table 4. Medium (e.g., monitoring of parking spaces) to high (e.g., monitoring of home video) data rate is needed in cases of mMTC monitoring activities. High data rate can be achieved by using a millimeter wave (mmWave) spectrum to share multi-gigabit data in the surrounding environment (i.e., mMTC monitoring with high data rate requirement) and to recognize an object via cloud in real time to find the optimal driving strategy instantaneously for IoT—autonomous vehicles application [101]. Data rates generated during transmissions of tracked medical assets, transport fleets, and ticked point of sale (POS) terminals is low (e.g., medium importance for ticketing and tracking activities in mMTC). In addition, low data rates are needed in uMTC monitoring and controlling (e.g., monitoring of health condition, control of driving).

4.2. Mobility

Mobility is defined as relative velocity between the receiver and the transmitter [16]. The applications of IoT pose very diverse requirements for mobility in 5G networks, which range from static to high mobile, even up to 500 km/h [16,102,207]. Use cases in which except ultra-high mobility, ultra-high traffic volume density, and ultra-high connection density are needed may be quite challenging for 5G networks [215], like V2X communication [5]. High mobility is a very important requirement for mMTC tracking activity (e.g., tracking of assets in high speed trains [214]). Moreover, the support for high mobility is needed in uMTC management activities, if the monitor/control object is moving (e.g., high speed trains [207]). Low mobility is needed for ticketing, monitoring, and managing/controlling activities in the mMTC class, as shown in Table 4. This is the reason why 5G networks should not assume mobility for all devices and services but rather provide mobility on demand [8].

4.3. Latency

Latency requirements are usually expressed in terms of end-to-end (E2E) latency perceived by the end user [8]. 5G networks should enable "zero latency" [103] represented by the millisecond level of E2E latency [16,26,27,213] through significant enhancements and new technology in architecture aspects [6,28,214], such as device-to-device (D2D) communication [5]. The IoT applications associated with uMTC monitoring and managing/controlling activities require low latency, as they tend to be real time. Required latency levels depend on the particular IoT application [8,16,28], being the narrowest in uMTC managing/controlling activity with a value that should not exceed 1 ms [8]. For example, the tolerable delay for use case mobile health [215] and remote surgery application is in order of sub-milliseconds [213]. For V2X communications, latencies should be ultra-low for some warning signals [8].

4.4. Connection Density

Connection density is defined as the number of connected devices per unit area [16]. It can usually be expressed in terms of an extremely high number of simultaneous active connections, such as 1 million connections per square meter [207,215], or 10 to 100 times higher number of connected devices [29]. The performance values of connection and traffic density for various 5G services are listed in [8,212]. This performance requirement is highly correlated with identified activity-based classes of IoT applications, since it is the main characteristic of mMTC. High device density brings the need for rapid internet protocol version 6 (IPv6) deployment, and high quality security algorithms and techniques, which lead to new system design and implementation, as described in [104].

4.5. Reliability

Reliability is the maximum tolerable packet loss rate at the application layer [105]. 5G must bring a reliability of 99,999% [27,30,31,212], or higher for specific use cases [212], (e.g., tele-protection in a smart grid network [207] or driverless cars [106]). Reliability is the main characteristic of uMTC monitoring and managing/controlling activities. Reliability will be a particularly challenging task in high-speed trains because of speed, load, and cell distance [16,107].

4.6. Position Accuracy

Position accuracy is the maximum positioning error tolerated by the application [105]. 5G should ensure accurate positioning of the device outdoors [108] with accuracy from 10 m to <1 m on 80% of occasions and better than 1 m in indoor deployment [212]. Accuracy positioning is very important in uMTC monitoring-based activities (e.g., monitoring remote cameras), and uMTC managing/controlling-based activities (e.g., driving) [213]. Moreover, the mMTC tracking-and ticketing-based activities pose high performance requirements in terms of position accuracy.

4.7. Coverage

Coverage requirements in 5G should provide connectivity anytime and anywhere with a minimum user experience data rate of 1 Gbps [32]. However, the perception of 100% coverage is rather a technical decision than a business one [33], which could be extended through ultra-cell deployment [31] and vehicle-to-infrastructure (V2I) communications [30]. Almost every activity-based IoT application class requires very high levels of coverage (99,999% availability) [207]. Total coverage will enable new unmanned aerial vehicles (UAV) to use single network connection, instead of connection steering mechanism, as one described in [109].

4.8. Energy Efficiency

Energy efficiency is defined as the number of bits that can be transmitted per joule of energy [216]. Compared with current wireless technologies, the energy efficiency (measured in b/J) of the 5G network may need to be improved by a factor of 1000 [4,7,34,110,217]. High energy efficiency is important in case of ticketing- and in some cases of tracking-based activities (e.g., smart industry [205] or implantable medical devices [35]). Other activities, such as both mMTC and uMTC monitoring- and managing/controlling-based activities (e.g., health condition [205]) and some tracking-based activities (e.g., sports wearables [205]) require medium energy efficiency. Energy efficiency is very important design objective for the reduction of operating costs of telecom operators, as well as for minimizing the environmental impact of the wireless domain [217]. On a higher layer of the network protocol stack, adaptive base station switch on/off algorithms use renewable energy sources to save energy [218] along with an energy scheduler, as in the with heating, ventilation and air conditioning (HVAC) [111]. However, at the physical layer, adaptively switching off unused carriers is a key strategy that can be used to save energy from the radio-frequency (RF) transceiver chain of base stations [218].

4.9. Spectrum Efficiency

Spectrum efficiency is defined as the data throughput per unit of spectrum resource per cell or per unit area (bps/Hz/cellar bps/Hz/km^2) [216]. In order to achieve network sustainability, required for 5G networks [216], spectrum efficiency needs to be improved 3–5 times [20,32,216]. Minimum peak spectrum efficiency is 30 bps/Hz for downlink and 15 bps/Hz for uplink [32]. This is mostly important for the xMBB services class [212,214]. Since this performance requirement is not relevant for activity-based IoT application classes which are associated with the mMTC and uMTC service class, it has not been further considered herein.

4.10. Summary of IoT in 5G Performance Requirements

Activity-based classes of IoT applications are associated with performance requirements of 5G service classes, i.e., mMTC and uMTC. On the basis of the literature review, each activity-based class (ticketing-, tracking-, monitoring-, and managing/controlling-based class) is associated with mMTC service class, while the monitoring- and managing/controlling-based classes are the only ones linked to the uMTC service class. According to the authors' best knowledge, there is no available literature concerning the uMTC ticketing- and tracking-based classes of IoT applications or such IoT customer service requests.

Activity-based IoT application classes pose many performance requirements, which have been discussed in terms of data rate, mobility (speed), latency, connection density, reliability, positioning accuracy, coverage, and energy efficiency. These performance requirements have been assigned three-level priorities (high, medium, low) for the purpose of facilitating identification of the enabling technologies used to fulfill them. Based on the literature review, it was found that each activity-based IoT application class poses high requirements in terms of connection density except the uMTC managing/controlling-based IoT application class. Moreover, ticketing- and tracking-based IoT application classes have high requirements in terms of positioning accuracy, coverage, and energy efficiency. Additionally, positioning accuracy is a highly important performance requirement for the uMTC managing/controlling-based IoT application class, whereas coverage is important for both uMTC and mMTC monitoring-based IoT application classes. Finally, latency and reliability represent highly important performance requirements for uMTC monitoring- and managing/controlling-based IoT application classes. A deep and comprehensive understanding of performance requirements of each activity-based IoT application class may facilitate the selection of 5G enabling technologies needed to meet them as described in the following section.

5. IoT in 5G Enabling Technologies

Activity-based classification of IoT applications proposed in Section 3 allows telecom operators to identify performance requirements of each class relying on 5G service classification, as discussed in Section 4. However, these performance requirements cannot be satisfied for many IoT applications with current cellular (2nd, 3rd, and 4th generation) network technologies, since they limit their potential due to many issues [3], i.e., protocol implementation complexity, poor coverage in non-urban environments, high cost of networking equipment and data transmission. In order to overcome these issues, many dedicated communication technologies are being installed.

From the very beginning of IoT, many proprietary technologies, such as radio-frequency identification (RFID), wireless highway addressable remote transducer (WirelessHART) or Z-Wave, have first appeared, and then, more generic ones, such as Bluetooth, IEEE 802.15.4, IPv6 over low-power wireless personal area networks (6LoWPAN). However, none of these technologies have become a market leader mainly because of technology shortcomings and business model uncertainty [219]. Hence, new solutions, such as low-power wireless fidelity (LP Wi-Fi), low-power wide area (LPWA) or several improvements for cellular M2M communications have become serious candidates for IoT implementation. LP Wi-Fi is an IEEE 802.11ah standard designed to extend the application area of Wi-Fi networks in order to meet IoT requirements (i.e., large number of devices, large coverage range, energy constrains). First performance studies indicate that this standard will support a broad range of M2M scenarios with a required QoS level, and enable scalable and cost-effective solutions. In addition, LPWA networks have been deployed for some time in the form of many different proprietary solutions (e.g., Amber Wireless, Coronis, Huawei's CIoT, LoRa, M2M Spectrum Networks, Sigfox, Weightless, etc.), but only the LoRa Alliance, Sigfox and Weightless are involved in LPWA standardization activities. Despite some drawbacks which are mainly related to use of an unlicensed spectrum, LPWA networks are expected to become a key enabler for IoT deployment in early market rollouts and for limited IoT applications [219].

The appearance of these IoT communication technologies was considered as competition to current cellular networks from the telco's point of view from the start. But, in the meanwhile, telcos have realized that the aforementioned communication technologies can be utilized to meet the changing connectivity and performance requirements of IoT applications. Therefore, many dedicated communication technologies have already been deployed in various IoT applications. For example, Orange, Swisscom and South Korea (SK) Telecom have built nationwide networks based on LoRa [3], whereas Deutsche Telekom (DT), Vodafone and all three Chinese operators have completed the rollout of narrowband IoT (NB-IoT), as another LPWA standard utilizing existing long term evolution (LTE) networks [36]. In this context, standardization and interoperability becomes critical because there is a need to consider a broad range of connectivity solutions as presented in Figure 4 [37]. One of those solutions includes satellite technologies which integrated with 5G radio technologies form the 6th generation (6G) standard for providing global coverage [38].

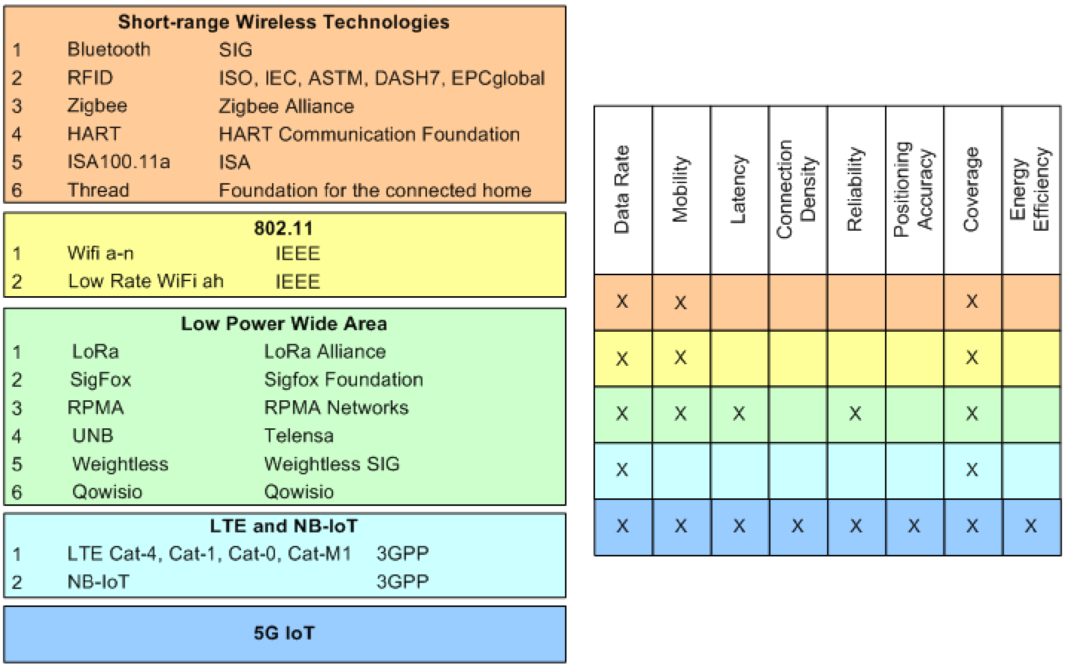

Short-range Wireless Technologies

1	Bluetooth	SIG
2	RFID	ISO, IEC, ASTM, DASH7, EPCglobal
3	Zigbee	Zigbee Alliance
4	HART	HART Communication Foundation
5	ISA100.11a	ISA
6	Thread	Foundation for the connected home

802.11

1	Wifi a-n	IEEE
2	Low Rate WiFi ah	IEEE

Low Power Wide Area

1	LoRa	LoRa Alliance
2	SigFox	Sigfox Foundation
3	RPMA	RPMA Networks
4	UNB	Telensa
5	Weightless	Weightless SIG
6	Qowisio	Qowisio

LTE and NB-IoT

1	LTE Cat-4, Cat-1, Cat-0, Cat-M1	3GPP
2	NB-IoT	3GPP

5G IoT

Data Rate	Mobility	Latency	Connection Density	Reliability	Positioning Accuracy	Coverage	Energy Efficiency
X	X					X	
X	X					X	
X	X	X		X		X	
X						X	
X	X	X	X	X	X	X	X

Figure 4. IoT performance requirements and enabling technologies.

In this regard, standardization activities of 5G and beyond are being undertaken by several standards bodies, such as the Institute of Electrical and Electronics Engineers (IEEE), the 3rd Generation Partnership Project (3GPP), the Internet Engineering Task Force (IETF), ITU Radiocommunication Sector (ITU-R), and ITU Telecommunication Standardization Sector (ITU-T). An overview on these standards bodies and their effort to develop communication standards for 5G and beyond are provided in [112]. Due to the envisioned 5G applicability, these standards have defined an air interface that can significantly improve the performance, and network architectures that allow the deployment and coexistence of various 5G technologies. A detailed discussion of these technologies regarding both technological and standardization aspects indicates that 5G can be perceived as the main driving force for enabling the vision of a truly global IoT [219].

In this sense, the technologies described in this section will allow 5G networks to form a unified communication infrastructure for the realization of a wide range of IoT applications. These technologies will bring performance enhancement which will trigger a complete revolution in almost all spheres of human life creating a new "everything connected" era. Future 5G networks in such an era will face a serious problem regarding a huge number of different service types [220]. To meet the diverse requirements for a huge number of IoT applications, 5G introduces the concept of network slicing to offer programmable network instances [220]. Using network slicing, multiple independent and dedicated network instances can be created within the same infrastructure to run services that have completely different requirements for latency, reliability, throughput, and mobility [220]. This technology primarily targets a portion of the 5G core network, but also indicates that 5G radio access network (RAN) functionalities can be different for each network slice [221]. To deploy network slices, network functions need to be enabled on an on-demand basis, which has driven the use of virtualization and softwarization in 5G core network [221]. Network function virtualization (NFV) is a paradigm that enables that 5G network functions to run in a virtual environment instead of dedicated hardware [221]. The software define networking (SDN) paradigm facilitates isolation of network slices avoiding the traffic of one slice affecting the performance of another slice [221]. These two paradigms make the network much more dynamic, agile, on-demand, and flexible [222,223]. Since the aforementioned technologies are used to deal with a variety of 5G applications, they will not be further discussed. This paper will focus on technologies used to meet performance requirements for specific applications (inter slice performance) which are implemented in the radio access part of 5G networks.

These technologies are briefly described and discussed in terms of their advantages, disadvantages, and research gaps. Given the fact that the realization of particular technology affects multiple performance requirements, tiny steps towards its implementation will allow telcos to deploy 5G networks, and thereby provide a performance improvement in terms of more capacity, lower latency, more mobility, more position accuracy, increased reliability, and availability [6,214,224]. In other words, 5G networks will accommodate many more users and devices while delivering more data [113,114] to each user requiring high data rates [108] in a more energy-efficient way [115,116].

Table 5 summarizes considered radio technologies used to implement future 5G networks being capable of achieving performance requirements identified in Section 4. On the basis of the literature review, one may conclude that the 5G radio access network is crowded with multiple technologies, and there seems to be a duplication of technologies, all eager to grab telco's attention and convince them to buy into the particular choice. As Table 5 shows, multiple technologies and techniques can be used to meet each performance requirement. Since the mMTC activity-based IoT classes are more mature, we have primarily focused on technologies used to fulfil their high performance requirements (i.e., connection density, positioning, and coverage). We have selected 5G enabling technologies considered from telco's perspective as the most representative and promising candidates to meet these requirements. The implementation of these technologies will also significantly affect the fulfillment of uMTC service class requirements, which will lead to complete automation in all spheres of human life. Therefore, the following subsections will discuss these technologies in more detail with the final goal of identifying research gaps and providing recommendations for future work, which are summarized in Table 6.

Table 5. IoT in 5G enabling radio technologies.

Enabling Technologies		**IoT in 5G Service Requirements**								
		Data Rate	**Mobility**	**Latency**	**Connection Density**	**Reliability**	**Positioning**	**Coverage**	**Energy Efficiency**	**Spectrum Efficiency**
Wide and flexible bandwidth technology	mmWave Band Communication and large-scale antenna	[7,202,225–227]		[202]			[233]	[5]	[202]	[7,29,214]
	Heterogeneous Multi-RAT Integration	[5,217,228–232]						[5]		
	Cognitive Radio and Spectrum Sharing							[5]		
Advanced modulation and coding	Advanced Modulation	[52,234–236]	[237,238]	[239]		[237,239]		[237]		
	Advanced Channel Coding	[235]		[235]						[216]
Duplexing	In-band FD			[124]				[5]		[7,29,214]
	Dynamic TDD	[240]		[240]						[9]
Multiple access and waveform	Multiple access			[216]					[218]	[7,9,29,216,256,257]
	New waveform				[5]					[7,9,29,216,256,257]
Advanced interface management	SND and SWSC	[5]								
Access architecture related radio technologies	Advanced small cell	[7,202]					[5]	[5,251]	[7,202,217]	[202,257]
	MN	[202]	[202]					[5]	[202]	
	Enhanced wireless backhaul							[5]		
	D2D	[202,248]	[7,241,244]					[245–247]	[7,202,215]	[7,9,29,202,252]
Energy related technologies	Energy harvesting						[100,253]			
	UAV	[254]				[254]		[183–185]		
Other important technologies	mMIMO	[7,202,249]		[13,202]		[249]		[200]	[217]	[7,9,214,216,257]
	VLC	[202,250]		[202]					[202,258]	
	SIC		[202]	[202,236]						
	V2X			[255]				[255]		

Legend: IoT (Internet of Things), 5G (Fifth Generation), mmWave (Milimeter-wave), RAT (Radio Access Technology), FQAM (Frequency Quadrature Amplitude Modulation), FD (Full Duplexing), TDD (Time Division Duplexing), SND (Simultaneous Non-unique Decoding), SWSC (Sliding Window Superposition Coding), MN (Moving Network), D2D (Device-to-Device), mMIMO (massive Multiple Input Multiple Output), VLC (Visible Light Communication), SIC (Self Interference Cancelation), V2X (Vehicle to Everything), UAV (Unmanned Aerial Vehicles).

Table 6. 5G enabling technologies—research directions.

5G Enabling Technologies		Research Gaps and Directions	References
Wide and flexible bandwidth technology	mmWave band communication and large-scale antenna	(1) 3D channel modeling; (2) dynamic power control; (3) user scheduling and congestion control; (4) hardware limitation and adaptive beam-steering technique; (5) design of mobility management and admission control for mmWave-based dense HetNet; (6) design of frequency management schemes for mmWave; (7) Tactile Internet; (8) effective and efficient mmWave implementation in HetNets (access and networking).	[7,87,217]
	Heterogeneous multi-RAT integration	(1) cell-association; (2) traffic-offloading algorithms; (3) interference management schemes in case of inter user and inter cell interference; (4) cross-tier handover, access admission, and mobility management schemes of a multi-tier HetNets.	[196,217]
	Cognitive radio and spectrum sharing	(1) Spectrum sensing (design of cooperative frameworks, choose cooperative secondary users and transmit cooperative information); (2) develop framework and algorithms for group handoff of secondary users and security; (3) simulation of different attacks and scenarios to enhance security; (4) in-depth performance analysis between GFDM and UFMC in CR settings.	[29,110]
Duplexing	In-band full duplexing	(1) Redesign of network and management; (2) antenna and circuit design and development of the theoretical foundation; (3) analyze the throughput of a network of randomly deployment terminals sing stochastic geometry; (4) characterize the capacity advantage due to IFDB in various network scenarios; (5) guidelines to practical design: coding, modulation, power allocation, beamforming, channel estimation, equalization, digital interference cancellation and decoding, (6) design of a MAC layer.	[55,197]
Multiple access and waveform	New waveform	(1) Performance of SIC cancelation or filtering on f-OFDMA; (2) balance of time and frequency dispersion and design an efficient filter prototype for UFMC.	
Access architecture related radio technologies	Advanced small cell	(1) Expect of wireless backhauling on user experience; (2) exploitation of location data and fingerprints in optimizing small cell discovery in terms of time and energy- efficiency; (3) interference management when integrating D2D and small cells.	[34,80]
	Enhanced wireless backhaul	(1) TDD multi-flow coordination schemes to avoid bottlenecks in the downlink backhaul; (2) Backhaul aware association in ultra-dense deployment; (3) reliability and security of the backhaul.	[34,88,198]
	Moving network	(1) Resource allocation and interference in the mobile relay when trains are moving from opposite directions; (2) handover decision of users (more than one train arrive or depart, stop or pass); (3) group mobility for users on board very high-speed vehicles; (4) design od cooperative communication schemes; (5) deployment of moving networks in various vehicle environments, not just on fixes route railways.	[5]
	D2D communication	(1) Interference management (mode selection, resource allocation and power control); (2) integration of novel reputation-based mechanism for identify and avoid malicious users from multiple users in multi-hop D2D communications; (3) testing of D2D interference management schemes in 5G scenarios (mmWave, cell densification).	[29,208,209]
Energy related technologies	Energy harvesting	(1) Improving energy harvesting schemes; (2) simulation of proposed models; (3) integration with other 5G technologies.	[100,182,253]
	UAV	(1) Optimal deployment, mobility and energy-efficient use of UAVs; (2) integration with other 5G technologies.	[81]
Other technologies	mMIMO	(1) Performance of practical mMIMO.	

Legend: IoT (Internet of Things), 5G (Fifth Generation), mmWave (Milimeter-wave), RAT (Radio Access Technology), D2D (Device-to-Device communication), mMIMO (massive Multiple Input Multiple Output), 3D (Three Dimensional), HetNet (Heterogeneous Network), GFDM (Generalized Frequency Division Multiplexing), UFMC (Universal Filtered Multi Carrier), CR (Cognitive Radio), IFDB (In-band Full Duplexing), MAC (Medium Access Control), f-OFDMA (filtered Orthogonal Frequency Division Multiple Access), TDD (Time Division Duplexing), UAV (Unmanned Aerial Vehicle).

5.1. Wide and Flexible Bandwidth Technology

mmWave band communication and large-scale antennas are promising technologies for future 5G networks. The mmWave band covers frequencies from 30 GHz to 300 GHz [202], and from an industry and wireless academia point of view, it is a unique solution for solving 5G capacity requirements [96]. mmWave band communications will provide high data rates [7,202,225–227] utilizing a much larger spectrum bandwidth that can reach up to 5 GHz [39] and by using directional antennas and high attenuation [40]. To provide sufficient antenna gain mmWave requires implementation of large-scale antennas at the transmitter and receiver side [96]. Deployment of large antenna arrays with mmWave will also bring high spectral efficiency, high throughput and channel gain [7]. Despite all of its advantages, mmWave needs line of sight (LOS) operation [41], while the effective communication distance of mmWave signals is within 200 m due to the propagation characteristic of this frequency band [29]. Open problems associated with mmWave include three dimensional (3D) channel modeling, dynamic power control, user scheduling and congestion control, hardware limitation and the adaptive beam-steering technique, as described in [7]. The problem of high power consumption of a large number of antennas in an array [7], high efficiency low complexity adaptive antenna array processing algorithms [29] and innovative hardware architecture of large-scale antenna transmitters [29] still remain unsolved. A proposal of design guidelines in architectures and protocols for mmWave communications is presented in [42]. It is demonstrated that new mmWave technologies, which are under investigation for 5G communications systems, will be able to provide indoor centimeter (cm)-accuracy localization in a robust manner, ideally suited for Assisted Living (AL) [228].

Heterogeneous multi-radio access technologies (multi-RAT) integration is specific to 5G networks radio design that coexists with existing networks. Since 5G networks will not be developed to replace current wireless networks, but rather to advance and integrate existing network infrastructures with new ones [217], we refer to 5G as a heterogeneous network. Multi-RAT is defined as the capability of a mobile network to support multiple radio access technology with seamless interworking among them [229]. When deployed in heterogeneous networks (HetNets), with traffic offloading among different RATs, multi-RAT improves capacity [230], supports better communication rates [34,117–119,231,232], better energy efficiency [34,117–119,231,232], and ensures seamless connectivity with higher QoS [230]. In designing a heterogeneous RAT, researches are finding solutions for cell-association and traffic-offloading algorithms [217].

Cognitive radio with spectrum sharing is a new software defined technology, which is expected to improve the utilization of the congested radio frequency (RF) spectrum [43]. In 5G networks, it is used for designing multi-tier architectures, removing interference among cells, and minimizing energy consumption in the network [44–47,120,233]. Moreover, a spectrum sharing technique can be used along with the CR technology to integrate the 5G spectrum [48,121]. During the practical implementation of the CR and spectrum sensing (SS) technique researchers had to design cooperative frameworks, choose cooperative secondary users and transmit cooperative information during the spectrum sensing, further developing the framework and algorithms for the group handoff of secondary users and enhancing security [29]. Researchers are already working on designs of protocols for different IoT application, based on cognitive radio, and some of them are presented in [49–51,122]. Technologies in this group are used together to optimize 5G performance requirements (e.g., a prototype of mmWave integrated HetNet in [123] and HetNet that incorporates massive multiple-input and multiple-output (mMIMO) and mmWave technologies [217]).

5.2. Advanced Modulation and Coding

This technological group involves advanced modulation (i.e., frequency and quadrature amplitude modulation (FQAM)) and advanced channel coding schemes.

Advanced modulation schemes group includes FQAM, Amplitude and Phase Shift Keying (APSK), Unitary Space-Time Modulation (USTM), Spatial Modulation (SM), Wave Modulation

(WAM), and Orthogonal Time Frequency and Space (OTFS). The FQAM is a combination of frequency shift keying (FSK) and quadrature shift keying (QAM) [124–126]. This modulation can achieve a higher transmission rate for cell edge users reducing interference at cell edge [52,234–236]. FQAM also improves energy efficiency, which makes it adequate for MTC devices with stringent energy consumption requirements [124]. Due to performance advantages in terms of Frame Error Rate (FER) [53,236], FQAM is an ideal candidate for services with high coverage and reliability requirements [237]. APSK is another modulation which draws a lot of attention. It is shown that its main performance gain (i.e., to achieve a channel capacity very close to Shannon's) relies on advanced channel coding and demodulation algorithms [237]. This adds more complexity at the APSK transmitter and receiver. Another modulation, which does not need Channel State Information (CSI) to enable high throughput, is called USTM [237]. USTM and its extension (see [238]) is very useful for 5G services with high mobility [127] or latency and reliability constraints [239]. Other modulation schemes involve SM [128], and proprietary WAM and OTFS [237]. On the basis of the foregoing, researches have analyzed and compared many modulation forms, but still have not found a practical guide on how to choose modulation in any of the 5G use cases.

Advanced channel coding techniques are used for correcting the communication errors caused by noise, interference, and poor signal strength [129]. Authors in [54,129,130] compared turbo, low density parity check (LDPC) and polar codes in decoders in contrast to 5G requirements: (1) maturity; (2) throughput and latency; (3) error correction capability; (4) flexibility; (5) computation complexity; (6) interconnect complexity; (7) high-performance flexible implementation complexity; and (8) backward compatibility. This comparison showed that turbo codes hold the greatest promise for offering high performance throughputs, latencies and error correction capabilities, as well as high degrees of flexibility at the lowest implementation complexity [129] in most 5G use cases. However, further work is needed to implement a decoder based on this code followed by detail analysis.

5.3. Duplexing

This technological group includes in-band full duplexing (FD) and dynamic time division duplexing (TDD).

In-band FD or simultaneous data transmission and reception will provide a 1000-fold increase in throughput [131,240], double spectral efficiency [132,240], and reduce the air interface delay [240]. The central research problem for the practical implementation of an in-band full-duplex radio is the attenuation of the self-interference signal by an adequate amount [132]. For the practical implementation of an in-band full-duplex radio, many aspects of network design and management need to be restructured, where terminals antenna and circuit design and the development of theoretical foundation are in focus [55]. One practical implementation of the full-duplex radio is shown in [133]. Authors in [134] commented that for the design of a full-duplex radio, it is necessary to unify researches from three domains, i.e., RF circuit and system design, digital signal processing and networking.

Dynamic TDD is the predecessor of FD transmission technology and a candidate for 5G [133,135]. It represents a scheduling technique in which every base station (BS) is free to choose its own uplink/downlink (UL/DL) split [136–139]. This technique is used to adapt the allocation of network resources to variable traffic requirements [139,140], often found in ultra-densely deployed networks [141]. The dynamic TDD can significantly increase bandwidth efficiency [142] and provide higher throughput and low latency [141]. However, it is characterized by the severe co-channel interference (CCI) [141–143]. Dynamic TDD could be used in combination with D2D communication and in the self-backhauling scenario, as described in [55].

5.4. Multiple Access and Waveforms

Multiple access techniques are becoming an important technology in 5G because of their ability to support mMTC activity-based IoT classes with urgent deploy demand [144]. They include several non-orthogonal multiple access (NOMA) forms: multi-user shared multiple access (MUSA) [145–148], resource spread multiple access (RSMA) [149], sparse code multiple access (SCMA) [150–152], pattern division multiple access (PDMA) [153–155], interleave-division multiple access (IDMA) [156,157], and NOMA by power domain [158]. The NOMA is a radio access technology design for enabling greater spectrum efficiency [56,159–162,207], higher cell-edge throughput, relaxed channel feedback, and low transmission latency [99]. NOMA can be employed to enhance user fairness and to support massive connections with diverse QoS requirements [163]. In NOMA, there are several challenges and open issues, such as dynamic user pairing, the impact of transmission distortion, the impact of interference, resource allocation, NOMA with multiple antennas, heterogeneous networks, outage probability analysis, practical channel model, uniform fairness, NOMA with antenna selection, carrier aggregation and other challenges as discussed in [216]. There are still many challenging issues for SCMA, which need to be solved in future work. For example, there are several open issues in SCMA transceiver design [57], the optimization of algorithms for user grouping and power allocation [164], and further enhancement of the SCMA and MIMO combination.

New waveforms have become a serious candidate for 5G being studied in terms of: (1) modulation based on pulse shaping: filter bank multicarrier (FBMC) [58], generalized frequency division multiplexing (GFDM) [59], pulse shaped OFDM [165] and QAM-FBMC [241]; (2) modulation based on sub-band filtering: universal filtered multi carrier (UFMC) [242], filtered OFDM (f-OFDM) [243] and resource block f-OFDM (RB-f-OFDM) [244]; other modulation format: guard interval discrete Fourier transform spread OFDM (GI DFT-s-OFDM) [245], spectrally-precoded OFDM (SP-OFDM) [246] and orthogonal time frequency and space (OTFS) [247]. The f-OFDM is seen as a potential candidate for IoT applications [60]. Due to narrow sub-bands and thereby pure detection performance, additional processing is needed [60]. Performance of self-interference cancellation (SIC) or filtering on f-OFDMA [60] remains a topic for further study. Additionally, it is interesting to consider the balance of time and frequency dispersion in UFMC, as well as the design of an efficient prototype filter according to application scenarios [60].

5.5. Advanced Interface Management

This technological group includes receiver advances in terms of simultaneous non-unique decoding (SND) and sliding-window superposition coding (SWSC).

SND follows a rule that implies that each receiver attempts to recover the code words from intended and interfering senders [248]. The combination of advanced receivers and joint scheduling provides an improvement of over 50% in cell edge throughput without sacrificing the cell average throughput [5]. This gain demonstrates that if 5G networks incorporate advanced interference management, they will provide a virtually edgeless end-user experience [5].

SWSC [61,62] combines the theory concept from superposition coding without rate splitting [63], block Markov coding [166,167], successive cancellation decoding [63,64,168], and sliding-window decoding [65,169,249]. The sliding-window coded modulation (SWCM) aims to mitigate inter cell interference at the physical layer by achieving simultaneous decoding performance with point-to-point channel codes, low-complexity decoding, and minimal coordination overhead [170]. The realization of the theoretical concept to the practical transmission coding scheme is an important research direction in SWSC [5].

5.6. Access Architecture Related Radio Technology

This technological group includes ultra-densification, enhanced wireless backhaul, moving networks, and D2D communication.

Advanced small cell [66,123] deployment is considered to be one of the key enablers for achieving many of the requirements currently envisioned for 5G [67], with higher data rates [68,171,250] and more capacity [171–174], as most important. Higher data rates and smaller battery consumption can be achieved using short distances in the small cell [69]. Solving the capacity and data rate challenge with network densification could be very expensive in terms of equipment, maintenance, and operations [200]. With denser cell deployment come significant challenges in the design of a high-performance backhauling system for RAN and the impact of backhaul on radio resource management (RRM) [70]. However, there are in existence many interesting topics to be further investigated which can be grouped around several problem areas [34]: user association, backhauling, interference management, energy efficiency, and propagation modelling.

Moving networks (MN) are the combination of multi-hop and vehicular communications concepts [69,71]. The MN deploys one or several moving relay node(s) (MRNs) [175,176,251] on vehicles that form their own cell(s) inside the vehicle to serve vehicular users [72]. Challenges in using MRN are efficient backhauling, design of efficient resource allocation and interference management techniques, as well as mobility management schemes to exploit the benefit of group handovers for vehicular user equipment (UE) devices served by the same MRN [72]. The main research challenges related to MN concept are associated with complexity management due to the mobility of access points and providing a high-rate wireless backhaul link from the moving cell to the fixed network [69].

Enhanced wireless backhaul is one of the main challenges in hyper-dense 5G networks [34,217]. The performance of wireless backhaul is dictated by the environment and traffic profile of the intended use case [9]. In UDN except mmWave, new approaches are needed, e.g., interface aware routing and intelligent resource allocation [9], as these links may occupy part of the spectrum used in the access network [73]. Latency and reliability of this link are important issues to be considered being prone to blocking and fading [69]. In any case, further research is necessary to explore TDD multi-flow coordination schemes to avoid bottlenecks in the downlink backhaul [9]. The guidelines for deploying future 5G wireless backhaul networks in economical and highly energy-efficient ways are provided in [74].

D2D communication is defined as a direct route of data traffic between spatially closely located mobile UE [75,177,178]. The advantages of D2D communication compared with the traditional cellular method include the reduction of transmission latency [73,75] and power consumption [75], improvements in coverage [71,76,179], spectral and energy efficiency [77], and throughput, when power control and resource allocation methods are used [78]. The use of D2D communications has an overall positive impact on system capacity in cellular environments. D2D communication brings challenges and complexities related to interference management [79,80], QoS requirements [79], and resource allocation [34,215]. The design of D2D direct communication link is still a hot topic [29].

5.7. Energy Related Technologies

IoT applications span a broad range of domains including home automation, healthcare, surveillance, transportation, smart environments, etc. One of the most important obstacles for implementing such an impressive scheme is supplying adequate energy to operate the network in a self-sufficient way without compromising QoS [180]. Since energy efficiency is of most importance to battery constrained IoT devices, researches have focused their work on the development of the device energy saving mechanisms.

Energy harvesting is a new paradigm which uses solar, thermal, wind and kinetic energy sources to power sensor nodes and consequently prolong network lifetime [181]. Among different energy harvesting methods, wireless energy harvesting (WEH) has proven to be one of the most promising solution for energy aware IoT devices, because of its simplicity, ease of implementation, and availability [180]. This concept will improve some of the 5G high network communication requirements, like reliability of IoT communications, as presented in [182]. Authors in [252] have shown that the integration of social awareness and energy harvesting in D2D communication results in higher system capacity. Network architectures proposed in [100,253] have shown better energy efficiency in 5G wireless network. Researchers could work on improving this architecture and integration with various 5G technologies in the future.

UAV implies a flying vehicle without the driving presence of a human pilot to control it on-board [113]. Drones are a possible example of UAVs, although many other robotics related applications could be part of this application group [113]. UAVs can play a vital role in IoT scenarios where devices are unable to transmit over a long distance due to their energy constraints. In this case UAVs play the role of moving aggregators which fly toward one IoT devices, collect the data, and transmit it to other devices [81]. In [183] authors presented UAV–based floating relay (FR) for cell dynamic and coverage improvement. Deployment of unmanned aerial base stations in 5G heterogeneous architecture can improve throughput [254], coverage [183–185,254], connectivity [184,185], and 5th percentile spectral efficiency [254]. On-demand wireless systems with low-altitude UAVs are in faster to deploy, more flexibly reconfigured, and likely to have better communication channels due to the presence of short-range line-of-sight links [186] when compared to terrestrial communications or those based on high-altitude platforms. The utilization of highly mobile and energy-constrained UAVs for wireless communication introduces many new challenges [186], some of them are listed in [82]. To effectively use UAVs for IoT, several challenges must be addressed such as optimal deployment, mobility and energy-efficient use of UAVs [81].

5.8. Other Technologies

Along with the aforementioned technologies, a lot of research attention is devoted to technologies, such as mMIMO, visible light communication (VLC), SIC, vehicle-to-everything (V2X), etc.

mMIMO technology [73,123] promises significant gains in data rate and link reliability [187], reduces latency and energy [13], simplifies media access control (MAC) layer [13], shows robustness against intentional jamming [13], unintended man-made interface [83], and increases capacity [127,198] due to spatial multiplexing [123]. Reliable links are provided by benefiting from spatial diversity and the mitigating effects of fast fading, beamforming, and zero forcing caused by multi-user interference [203]. The mMIMO can be exploited to extend the coverage of higher frequency bands by relying on beamforming gains [200]. Other specific benefits of mMIMO system are: increased capacity 10 times or more with simultaneous improvement in radiated energy efficiency in the order of 100 times, the possibility to use inexpensive, low-power components, the reduction of latency on the air interface, and multiple access layer simplification [83]. The low complexity mMIMO uplink scheme for IoT lightweight devices is presented in [188]. The performance of practical mMIMO system needs to be investigated, since the research community pays attention to analyzing the performance of mMIMO in an ideal channel state information assumption [73].

VLC is a growing technology for short range, high capacity LOS optical links [189]. It uses a visible range of the electromagnetic spectrum (370–780 nm) which provides data transmission and room illumination using light emitting diodes (LEDs). Prominent features of VLC are an abundant license-free spectrum, the ability to provide multiple gigabit-per-second data rates, low energy consumption, and low implementation costs [190]. VLC is sensitive to sunlight and is not able to work long range without LOS. Since VLC coverage is LOS limited [84,191], this is known as LOS blocking [192]. With poor performance in non-line-of-sight scenarios, VLC networks fail to provide convenient UL coverage at the current state-of-the-art, and each AP illuminates only a small confined cell compared to cellular RF networks [192]. VLC applications are expected to include IoT, wireless Internet access, and vehicle-to-vehicle (V2V) communications, broadcast from LED, M2M communications, positioning systems, and navigation. The problem of how to provide mobile applications over VLC is still quite an open one [189].

SIC implementation enables low-latency applications in a cost effective manner [132]. SIC can be used to increase link capacity, spectrum virtualization, any-division duplexing, novel relay solutions, and enhanced interface coordination [132]. Practical SIC implementation is still in the investigation phase.

V2X is the communication between a vehicle and everything with which it might interact (e.g., other vehicles, traffic operators and service providers). It is used in conjunction with D2D for coverage extension and latency reduction in 5G networks [255]. V2X communication might play a vital role in terms of implementing smart and efficient traffic solutions [193], improving road safety, traffic efficiency, and driver convenience. To reduce traffic congestion one may use an intelligent route management system based on V2X communication [7]. Since it has big potential in autonomous driving application, the localization of vehicles in dense urban areas is difficult, so several algorithms are proposed to solve the problems [194]. Many V2X applications are described in existing literature [85,195]. However, there is still a need to design vehicular mobility management strategies in next UDN networks and to find a way to enhance existing connected vehicle services with V2X [86].

These technologies are highlighted as the main 5G driver in existing literature [5–7,96,207]. Authors compare these technologies and try to implement them through prototypes, or test the combination of these technologies. However, other technologies exist, such as cloud-RAN [7], Coordinated Multi-Point (CoMP), White Space Spectrum [208], which are not in the scope of this paper.

5.9. Summary of IoT in Enabling Technologies

Performance requirements of activity-based IoT application classes have served to identify the enabling technologies in the radio access part of 5G networks, which are summarized in Table 5. Various 5G enabling technologies have been sorted into eight technological groups:

1. Wide and flexible bandwidth technology: mmWave band communication and large-scale antennas, heterogeneous multi-RAT integration, cognitive radio, and spectrum sharing;
2. Advanced modulation and coding: advanced modulation schemes, advanced channel coding;
3. Duplexing: in-band FD, dynamic TDD;
4. Multiple access and waveform: multiple access, new waveforms;
5. Advanced interface management: SND, SWSC;
6. Access architecture related radio technologies: advanced small cell, MN, enhanced wireless backhaul, D2D;
7. Energy related technologies: energy harvesting, UAV;
8. Other technologies: mMIMO, VLC, SIC, V2X.

These technologies have been discussed in terms of their possibilities to meet performance requirements of activity-based IoT application classes and identify research gaps and directions, which are summarized in Table 6.

We have highlighted technologies only used to satisfy high priority performance requirements of mMTC activity-based IoT application classes, i.e., connection density, positioning accuracy, and coverage. We have focused on these technologies from telco's perspective, since the area of mMTC activity-based IoT classes is more mature, as its development has already started within LTE, while uMTC poses new research questions to be answered in order to achieve unprecedented levels of reliability needed for new applications in 5G.

The new waveform technology may be used to fulfil the requirements of almost every activity-based IoT application class in terms of connection density. This technology affects the connection density by definition. Analysis of different modulation formats indicates that f-OFDM is the most suitable candidate for IoT applications with high connection density requirements. In this regard, a deeper and more comprehensive analysis of f-OFDM is needed in the context of different mMTC activities (i.e., ticketing, tracking, monitoring, and managing/controlling). The result of the analysis to be performed should be the discovery of an optimal waveform for each activity-based IoT class which has to be integrated with other 5G service classes.

Large scale antennas and advanced small cell technologies can be used to satisfy the requirements of ticketing-, tracking-, and uMTC managing/controlling-based IoT application classes in terms of positioning accuracy. Fulfilling these performance requirements is rather a business than a technical decision. The size of the 5G cell must be planned to meet the desired position accuracy since it is very important for future IoT applications, such as automated driving, remote surgery, robotics, taxis, etc.

Wide and flexible bandwidth technologies, in-band full duplex technologies, access architecture related radio technologies, and mMIMO can be used to achieve the high requirements of ticketing-, tracking-, and monitoring-based IoT application classes in terms of coverage. The combination of these technologies should be tested to find the optimal solution to meet these performance requirements. Since small cell deployment has an impact on required position accuracy, they should serve as a basis for achieving the desired level of coverage. Along with small cell deployment, mMIMO and mmWave technologies are present in almost every 5G prototype.

Since this section discusses radio technologies used to accomplish performance requirements of mMTC activity-based IoT application classes, it can serve telecom operators and other interested parties depending on their interest and potential to utilize these findings towards IoT implementation and monetization. It describes each technology in terms of definition, advantages, disadvantages, and the possible impact on the performance requirements of a broad range of IoT applications that drive the deployment of 5G networks. Moreover, this section can be useful to the research community interested in this attractive field to address recognized research gaps and directions.

6. Discussion

Telecom operators have the opportunity to capture a share of the revenue that is generated by IoT implementation depending on their role in the IoT value chain. Regardless of their role, telcos have to face many technical challenges in order to meet the changing connectivity and performance requirements. Since current cellular networks limit numerous IoT applications, new technologies are being introduced leading towards 5G networks. Therefore, telcos have to focus on deploying these network technologies in order meet the changing requirements necessary to achieve success in IoT.

In this regard, writing this paper was motivated by the challenge of providing an enhanced understanding of the scope and classification of the broad range of IoT applications in order to determine and prioritize their performance requirements with the goal of specifying the enabling technologies towards 5G networks. The aim has been to propose a new classification of IoT applications, define and prioritize the performance requirements of such IoT application classes, and give insight into state-of-the-art technologies used to meet these requirements from telco’s point of view. The motivation that led to the focus of this paper being on IoT applications classification, performance requirements, and 5G enabling technologies could be explained by telco’s need for added value from IoT services. Hence, an analysis of IoT customer service requests inspired us to propose an activity-based classification of IoT applications as it can, according to authors’ best knowledge, has allowed telcos to more precisely specify their performance requirements and 5G enabling technologies, thereby improving customer experience.

In this regard, the paper fulfilled the following four objectives: (1) the identification of activities relevant to IoT applications and their usage as a new criterion for IoT application classification; (2) the specification and prioritization of performance requirements of such IoT applications classes; (3) the analysis of the radio technologies used to accomplish IoT application requirements; and (4) the identification of the research gaps and the recommendation of new research directions.

Through fulfilling these objectives, we reviewed literature from the fields of IoT in 5G service classifications, IoT in 5G performance requirements, and IoT in 5G enabling technologies. Since the first aim of this paper was to propose a new classification of IoT applications, the existing approaches have been summarized to serve as a basis to identify their drawbacks and formulate the appropriate solution to overcome them. In this context, we proposed a new approach to IoT applications classification, which was based on the activity as a new classification criterion denoting the main function/purpose of specific IoT application. This approach enabled a clearer and more precise positioning of particular application in the IoT application spectrum, as well as the determination of performance requirements and enabling technologies from telco’s point of view. The activity-based IoT application classification facilitated the specification of the performance requirements and determination of technologies which enable these requirements to be fulfilled. In this context, the proposed approach served as a basis for the simplification of the realization of particular IoT application. For example, although the monitoring- and tracking-based applications seem to be similar, they set different performance requirements on the network, thereby allowing the proposed approach to precisely grade the IoT applications from the same IoT application domain.

Furthermore, we associated the activity-based IoT application classes with the 5G service classes, i.e., mMTC and uMTC, in order to specify and prioritize their performance requirements. It was determined that almost each activity-based IoT application class poses high requirements in terms of connection density, whereas ticketing-based and tracking-based IoT application classes additionally require high positioning accuracy and coverage. In order to analyze these performance requirements, we have summarized the radio technologies used to implement the future 5G networks. The focus was on the technologies used to meet the aforementioned performance requirements of mMTC activity-based IoT application classes, since mMTC is more mature, as its deployment has already started, while uMTC requires further research to achieve incomparable levels of reliability needed to enable new applications in 5G.

The analysis showed that new waveform technology can be used to meet the requirements in terms of connection density; large-scale antennas and advanced small cell technologies can be used for the purpose of satisfying the requirements in terms of positioning accuracy; while wide and flexible bandwidth technologies, in-band full duplexing technologies, and access architecture related radio technologies (i.e., advanced small cell, enhanced wireless backhaul, moving network, and D2D communication) can be used to achieve high requirements in terms of coverage. In addition, the conducted research study allowed us to highlight a number of open research issues that could serve the research community and other stakeholders interested in this contemporary and attractive field. In this context, we have recognized research gaps and directions which mostly relate to network redesign and optimization in order to accommodate large-scale IoT applications.

7. Conclusions

The IoT paradigm has the potential to revolutionize all areas of daily life of individuals, businesses, and society as a whole. Telcos enjoy a central role in the paradigm of IoT because of owing communication infrastructure which is exposed to the numerous technical challenges due to changing connectivity and performance requirements of various IoT applications. These requirements cannot be met with the current cellular networks which create the need to introduce new types of technologies leading toward 5G networks as the main driver for enabling numerous IoT applications. In this context, this paper proposed the activity-based classification of IoT applications and specification of their performance requirements in order to identify 5G radio technologies used to meet them, all from telco's perspective. Activity-based classification of IoT applications indirectly allowed the construction and distribution of performance, as IoT value proposition, while using 5G enabling technologies, as IoT resources in telco's internal value chain. In this context, high performance requirements of each activity-based IoT application class (i.e., connection density, positioning accuracy, coverage) served as a basis to analyze various enabling technologies in radio access part of 5G network in terms of advantages, disadvantages, and research gaps.

On the basis of the conducted analysis, we concluded that the following technologies can meet the high performance requirements of mMTC activity-based IoT application classes: new waveform technology (in terms of connection density), large scale antennas (in terms of positioning accuracy), wide and flexible bandwidth technologies, in-band full duplexing technologies, and access architecture related radio technologies (in terms of coverage). According to the identified research gaps, one may conclude that the optimal solutions for each technology are still in its infancy. This implies that any 5G enabling technology should be first implemented and tested through prototype construction before the deployment of specific IoT application. Thereafter, the technology combination that meets the performance requirements of each activity-based IoT class should first be found. Finally, the practical implementation of different technological combinations may further lead to a deep and comprehensive analysis of QoS and QoE in the context of IoT applications.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Author Contributions: Enida Cero and Jasmina Baraković Husić conceived and designed the experiments (research study); Enida Cero performed the experiments (research study); Sabina Baraković analyzed the data and contributed materials; Enida Cero and Jasmina Baraković Husić wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Antonopoulos, A.; Perillo, C.; Verikoukis, C. Internet Service Providers vs. Over-the-Top Companies: Friends or Foes?—Short talk. *SIGMETRICS Perform. Eval. Rev.* **2016**, *44*, 37. [CrossRef]
2. Antonopoulos, A.; Kartsakli, E.; Perillo, C.; Verikoukis, C. Shedding Light on the Internet: Stakeholders and Network Neutrality. *IEEE Commun. Mag.* **2017**, *55*, 216–223. [CrossRef]
3. IoT-Ignite. Role of Telcos in Internet of Things. Available online: https://devzone.iot-ignite.com/wp-content/uploads/2017/01/Role-of-Telcos-in-IoT.pdf (accessed on 5 September 2017).
4. Kaur, S.; Singh, I. A Survey Report on Internet of Things Applications. *Int. J. Comput. Sci. Trends Technol.* **2016**, *4*, 330–335.
5. 5G Forum. 5G Vision, Requirements, and Enabling Technologies. Available online: http://kani.or.kr/5g/whitepaper/5G%20Vision,%20Requirements,%20and%20Enabling%20Technologies.pdf (accessed on 10 July 2017).
6. Saha, R.K.; Saengudomlert, P.; Aswakul, C. Evolution towards 5G Mobile Networks—A survey on Enabling Technologies. *Eng. J.* **2016**, *20*, 87–112. [CrossRef]
7. Akyildiz, I.F.; Nie, S.; Lin, S.C.; Chandrasekaran, M. 5G roadmap: 10 Key Enabling Technologies. *Comput. Netw.* **2014**, *106*, 17–48. [CrossRef]
8. NGMN Alliance. 5G White Paper-Executive Version. Available online: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
9. Popovski, P.; Mange, G.; Gozalves-Serrano, D.; Rosowski, T.; Zimmermann, G.; Agyapong, P.; Fallgren, M.; Hoglund, A.; Queseth, O.; Tullberg, H.; et al. Deliverable D6.6 Final Report on the METIS 5G System Concept and Technology Roadmap. Available online: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D6.6_Summary.pdf (accessed on 10 July 2017).
10. Aly, H.; Elmogy, M.; Barakat, S. Big Data on Internet of Things: Applications, Architecture, Technologies, Techniques, and Future Directions. *Int. J. Comput. Sci. Eng.* **2015**, *4*, 300–313.
11. Vermesan, O.; Friess, P. *Internet of Things—From Research and Innovation to Market Deployment*; Riverbed: Aalborg, Denmark, 2014.
12. Amaral, L.A.; de Matos, E.; Tiburski, R.T.; Hessel, F.; Lunardi, W.T.; Marczak, S. Middleware Technology for IoT Systems: Challenges and Perspectives Toward 5G. In *Internet of Things (IoT) in 5G Mobile Technologies*; Mavromoustakis, C., Mastorakis, G., Batalla, J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 333–367.
13. OECD. Machine-to-Machine Communications: Connecting Billions of Devices. Available online: http://www.oecd-ilibrary.org/docserver/download/5k9gsh2gp043-en.pdf?expires=1499694889&id=id&accname=guest&checksum=F403637E723693581F03EA9D6F3A3C3B (accessed on 10 July 2017).
14. Zheng, K.; Hu, F.; Wang, W.; Xiang, W. Radio Resource Allocation in LTE Advanced Cellular Networks with M2M Communications. *IEEE Commun. Mag.* **2012**, *50*. [CrossRef]
15. Zhang, Q.; Fitzek, F.H.P. Mission Critical IoT Communication in 5G. In *Future Access Enablers for Ubiquitous and Intelligent Infrastructures*; Springer International Publishing: Cham, Switzerland, 2015; Volume 159, pp. 35–41.
16. Oughton, E.J.; Frias, Z. Exploring the Cost, Coverage and Rollout Implications of 5G in Britain. Available online: http://www.itrc.org.uk/wp-content/uploads/Exploring-costs-of-5G.pdf (accessed on 10 July 2017).
17. Chen, Y.; Wang, W. Machine-to-Machine Communication in LTE-A. In Proceedings of the 2010 IEEE Vehicular Technology Conference (VTC 2010-Fall), Ottawa, ON, Canada, 6–9 September 2010.
18. Booysen, M.J.; Gilmore, S.J.; Zeadally, S.; van Rooyen, G.J. Machine-to-Machine (M2M) Communications in Vehicular Networks. *KSII Trans. Internet Inf. Syst.* **2012**, *6*, 529–546. [CrossRef]

19. Niyato, D.; Xiao, L.; Wang, P. Machine-To-Machine Communications for Home Energy Management System in Smart Grid. *IEEE Commun. Mag.* **2011**, *49*, 53–59. [CrossRef]
20. De Matos, W.D.; Gondim, P.R.L. M-Health Solutions Using 5G Networks and M2M Communications. *IT Prof.* **2016**, *18*, 24–29. [CrossRef]
21. Investopedia. Business Model. Available online: http://www.investopedia.com/terms/b/businessmodel.asp (accessed on 3 September 2017).
22. Gassmann, O.; Frankenberger, K.; Csik, M. Revolutionizing the Business Model. In *Management of the Fuzzy Front End of Innovation*; Springer: New York, NY, USA, 2014.
23. Vermesan, O.; Bahr, R.; Gluhak, A.; Boesenberg, F.; Hoeer, A.; Osella, M. D02.01 IoT Business Models Framework. Available online: http://www.internet-of-things-research.eu/pdf/D02_01_WP02_H2020_UNIFY-IoT_Final.pdf (accessed on 3 September 2017).
24. Dijkmana, R.M.; Sprenkelsa, B.; Peetersa, T.; Janssenb, A. Business Models for the Internet of Things. *Int. J. Inf. Manag.* **2015**, *35*, 672–678. [CrossRef]
25. Farooq, M.O.; Sreenan, C.J.; Brown, K.N.C. Research Challenges in 5G Networks: A HetNets Perspective. In Proceedings of the 19th International ICIN Conference Innovations in Clouds, Internet and Networks, Paris, France, 1–3 March 2016.
26. Hu, F. 5G Overview: Key Technologies. In *Opportunities in 5G Networks*, 1st ed.; Hu, F., Ed.; CRC Press: Boca Raton, FL, USA, 2016; pp. 1–557.
27. Ford, R.; Zhang, M.; Mezzavilla, M.; Duttam, S.; Rangap, S.; Zorzi, M. Achieving Ultra-Low Latency in 5G Millimeter Wave Cellular Networks. *IEEE Commun. Mag.* **2017**, *55*, 196–203. [CrossRef]
28. Prasad, R. IMT for 2020 and Beyond. In *5G Outlook—Innovations and Applications*; Riverbed Publishers: Gistrup, Denmark, 2016; pp. 1–260.
29. Ma, Z.; Zhang, Z.; Ding, Z.; Fan, P.; Li, H. Key Techniques for 5G Wireless Communications: Network Architecture, Physical Layer, and MAC Layer Perspectives. *Sci. China Inf. Sci.* **2015**, *50*, 1–20. [CrossRef]
30. Rappaport, T.S.; Daniels, R.C.; Heath, R.W.; Murdock, J.N. Introduction. In *Millimeter Wave Wireless Communications*; Pearson Education: Upper Saddle River, NJ, USA, 2014.
31. Ge, X.; Chen, J.; Ying, S.; Chen, M. Energy and Coverage Efficiency Trade-off in 5G Small Cells Network. *arXiv* **2016**, arXiv:1612.04459.
32. Hossain, S. 5G Wireless Communication Systems. *Am. J. Eng. Res.* **2013**, *2*, 344–353.
33. Artiga, X.; Nunez-Martinez, J.; Perez-Neira, A.; Lendrino Vela, G.J.; Fare Garcia, J.M.; Ziaragkas, G. Terrestrial-satellite Integration in Dynamic 5G Backhaul Networks. In Proceedings of the 2016 8th Advanced Satellite Multimedia Systems Conference and the 14th Signa (ASMS/SPSC), Palma de Mallorca, Spain, 5–7 September 2016.
34. Kamel, M.; Hamouda, W.; Youssef, A. Ultra-Dense Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2522–2545. [CrossRef]
35. Kartsakli, E.; Lalos, A.S.; Antonopoulos, A.; Tennina, S.; Di Renzo, M.; Alonso, L.; Verikoukis, C. A Survey on M2M Systems for mHealth: A Wireless Communications Perspective. *Sensors* **2014**, *10*, 18009–18052. [CrossRef] [PubMed]
36. Morris, A. DT, Chinese Operators take NB-IoT to Market. Available online: http://www.lightreading.com/iot/nb-iot/dt-chinese-operators-take-nb-iot-to-market/d/d-id/734022 (accessed on 6 September 2017).
37. Vermesan, O.; Eisenhauer, M.; Sunmaeker, H.; Guillemin, P.; Serrano, M.; Tragos, E.Z.; Valino, J.; van der Wees, A.; Gluhak, A.; Bahr, R. Internet of Things Cogitive Transformation Technology Research Trends and Applications. In *Cognitive Hyperconnected Digital Transformation*; Vermesan, O., Bacquet, J., Eds.; River Publishers: Delft, The Netherlands, 2017; pp. 17–95.
38. Gawas, A.U. An Overview on Evolution of Mobile Wireless Communication Networks: 1G–6G. *Int. J. Recent Innov. Trends Comput. Commun.* **2015**, *3*, 3130–3133. [CrossRef]
39. Qi, Y.; Hunukumbure, M.; Nekovee, M.; Sgardoni, V. Quantifying Data Rare and Bandwith Requirements for Immersive 5G Experience. In Proceedings of the 2016 IEEE International Conference on Communications Workshops, Kuala Lumpur, Malaysia, 23–27 May 2016.
40. Sim, G.H.; Asadi, A.; Loch, A.; Widmer, J. Opp-Relay: Managing Directionaly and Mobility Issues of Millimeter-Wave via D2D Communication. In Proceedings of the I9th International Conference on COMSNETS 2017, Bangalore, India, 4–8 January 2017.

41. Mer, S.B. Smart Vehicle-to-Vehicle Communication with 5G Technology. *Int. J. Recent Innov. Trends Comput. Commun.* **2015**, *3*, 3241–3244.
42. Niu, Y.; Li, Y.; Jin, D.; Su, L.; Vasilakos, A.V. A Survey of Millimeter Wave Communications (Mmwave) for 5G: Opportunities and Challenges. *Wirel. Netw.* **2016**, *21*, 2657–2676. [CrossRef]
43. Rusek, F.; Persson, D.; Lau, B.K.; Larsson, E.G.; Marzetta, T.L.; Edfors, O.; Tufvesson, F. Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays. *IEEE Signal Process. Mag.* **2012**, *30*, 1053–5888. [CrossRef]
44. Hong, X.; Wang, J.; Wang, C.X.; Shi, J. Cognitive Radio in 5G: A Perspective on Energy-Spectral Efficiency Trade-Off. *IEEE Commun. Mag.* **2014**, *52*, 46–53. [CrossRef]
45. Huang, L.; Zhu, G.; Du, X. Cognitive Femtocell Networks: An Opportunistic Spectrum Access for Future Indoor Wireless Coverage. *IEEE Wirel. Commun.* **2013**, *20*, 44–51. [CrossRef]
46. Lee, C.H.; Shih, C.Y. Coverage Analysis of Cognitive Femtocell Networks. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 177–180. [CrossRef]
47. Wang, W.; Yu, H.; Huang, A. Cognitive Radio Enhanced Interference Coordination for Femtocell Networks. *IEEE Commun. Mag.* **2013**, *51*, 37–43. [CrossRef]
48. Menon, R.; Buehrer, R.M.; Reed, J.H. On the Impact of Dynamic Spectrum Sharing Techniques on Legacy Radio Systems. *IEEE Trans. Wirel. Commun.* **2008**, *7*, 4198–4207. [CrossRef]
49. Zhang, Y.; Yu, R.; Nekovee, M.; Liu, Y.; Xie, S.; Gjessing, S. Cognitive Machine-to-Machine Communications: Visions and Potentials for the Smart Grid. *IEEE Netw.* **2012**, *26*, 6–13. [CrossRef]
50. Aijaz, A.; Aghvami, A.H. Cognitive Machine-To-Machine Communications for Internet-of-Things: A Protocol Stack Perspective. *IEEE Internet Things J.* **2015**, *2*, 103–112. [CrossRef]
51. Li, R.; Lu, B.; McDonald-Maier, K.D. Cognitive Assisted Living Ambient System: A Survey. *Digit. Commun. Netw.* **2015**, *1*, 229–252. [CrossRef]
52. Andrews, J.G.; Buzzi, S.; Choi, W.; Hanly, S.V.; Lazano, A.; Soong, A.C.K.; Zhang, J.C. What Will 5G Be? *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1065–1082. [CrossRef]
53. Samsung Electronics. 5G Vision, Version 2. Available online: http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf (accessed on 10 July 2017).
54. Iscan, O.; Lentner, D.; Xu, W. A Comparison of Channel Coding Schemes for 5G Short Message Transmission. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016.
55. Sabharwal, A.; Schniter, P.; Guo, D.; Bliss, D.W.; Rangarajan, S.; Wichman, R. In-Band Full-Duplex Wireless: Challenges and Opportunities. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1637–1652. [CrossRef]
56. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chin-Lin, I.; Vincent Poor, H. Application of Non-Orthogonal Multiple Access in LTE and 5G Networks. *IEEE Commun. Mag.* **2006**, *55*, 185–191. [CrossRef]
57. Lei, W. Non-Orthogonal Multiple Access Schemes for Future Cellular Systems. Ph.D. Thesis, University of Surrey, Surrey, UK, 17 May 2017.
58. Farhang-Boroujeny, B. OFDM Versus Filter Bank Multicarrier. *Signal Process. Mag.* **2011**, *28*, 92–112. [CrossRef]
59. Bellanger, M. FS-FBMC: An Alternative Scheme for Filter Bank Multicarrier Transmission. In Proceedings of the 2012 5th International Symposium on Communications Control and Signal Processing, Rome, Italy, 2–4 May 2012.
60. Cai, X.; Qin, Z.; Cui, F.; Li, G.Y.; McCann, J.A. Modulation and Multiple Access for 5G Network. Available online: https://arxiv.org/pdf/1702.07673.pdf (accessed on 10 July 2017).
61. Wang, L.; Sasoglu, E.; Kim, Y.H. Filtered OFDM: Sliding-Window Superposition Coding for Interference Networks. In Proceedings of the International Symposium Information Theory, Honolulu, HI, USA, 29 June–4 July 2014.
62. Wang, L. Channel Coding Technique for Network Communication. Ph.D. Thesis, University of California, San Diego, CA, USA, December 2015.
63. Cover, T.M. Broadcast Channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 2–14. [CrossRef]
64. Liao, H.H.J. Multiple Access Channels. Ph.D. Thesis, University of Hawaii, Honolulu, HI, USA, September 1972.
65. Kramer, G.; Gastpar, M.; Gupta, P. Cooperative Strategies and Capacity Theorems for Relay Networks. *IEEE Trans. Inf. Theory* **2005**, *51*, 3037–3063. [CrossRef]

66. Bhushan, N.; Li, J.; Malladi, D.; Gilmore, R.; Brenner, D.; Damnjanovic, A.; Sukhavasi, R.; Patel, C.; Geirhofer, S. Network Densification: The Dominant Theme for Wireless Evolution into 5G. *IEEE Commun. Mag.* **2014**, *52*, 82–89. [CrossRef]
67. Chin, W.H.; Fan, Z.; Haines, R. Emerging Technologies and Research Challenges for 5G Wireless Networks. *IEEE Wirel. Commun.* **2014**, *21*, 106–112. [CrossRef]
68. Biral, A.; Centenaro, M.; Zanellan, A. The Challenges of M2M Massive Access in Wireless Cellular Networks. *Digit. Commun. Netw.* **2015**, *1*, 1–19. [CrossRef]
69. Barreto, A.N.; Faria, B.; Almeida, E.; Rodriguez, I.; Lauridsen, M.; Amorim, R.; Vieira, R. 5G-Wireless Communications for 2020. *J. Commun. Inf. Syst.* **2016**, *31*, 146–163. [CrossRef]
70. Wang, N.; Hossain, E.; Bhargava, V.K. Backhauling 5G Small Cells: A Radio Resource Management Perspective. *IEEE Wirel. Commun.* **2015**, *22*, 41–49. [CrossRef]
71. Chen, L.; Huang, Y.; Xie, F.; Gao, Y.; Chu, H.; Li, Y.; Liang, F.; Yuan, Y. Mobile Relay in LTE-Advanced Systems. *IEEE Commun. Mag.* **2013**, *51*, 144–151. [CrossRef]
72. Svensson, T. Moving Cells/Networks. Available online: http://snow.itn.liu.se/asbtr/SNOWTommySvensson.pdf (accessed on 11 July 2017).
73. Zheng, K.; Leung, V.C.M.; Yang, L.L.; Chatzimisios, P. Guest Editorial Special Issue on 5G Wireless Systems with Massive MIMO. *IEEE Syst. J.* **2017**, *11*. [CrossRef]
74. Ge, X.; Cheng, H.; Guizani, M.; Han, T. 5G Wireless Backhaul Networks: Challenges and Research Advances. *IEEE Netw.* **2014**, *28*, 6–11. [CrossRef]
75. Liu, J.; Kato, N.; Ma, J.; Kadowaki, N. Device-to-Device Communication in LTE-Advanced Networks: A Survey. *IEEE Commun. Surv. Tutor.* **2014**, *17*, 1923–1940. [CrossRef]
76. Pérez-Romero, J.; Sallent, O.; Agusti, R. Enhancing Cellular Coverage through Opportunistic Networks with Learning Mechanisms. In Proceedings of the IEEE Global Telecommunication Conference, Atlanta, GA, USA, 9–13 December 2013.
77. Asadi, A.; Wang, Q.; Mancuso, V. A Survey on Device-to-Device Communication in Cellular Networks. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 1801–1819. [CrossRef]
78. Chen, T.; Kunnari, E.; Tapani, R. Device-to-Device Communication for LTE-Advanced Network System. Available online: https://tel.archives-ouvertes.fr/tel-00983507/document (accessed on 10 July 2017).
79. Hicham, M.; Abghour, N.; Ouzzif, M. Device-to-Device (D2D) Communication under LTE-Advanced Networks. *Int. J. Wirel. Mob. Netw.* **2016**, *8*, 11–22. [CrossRef]
80. Noura, M.; Nordin, R. A Survey on Interface for Device-to-Device (D2D) Communication and its Challenges in 5G Networks. *J. Netw. Comput. Appl.* **2016**. [CrossRef]
81. Mozaffari, M.; Saad, W.; Bennis, M.; Debbah, M. Mobile Unmanned Aerial Vehicles (UAVs) for Energy-Efficient Internet of Things Communications. Unpublished work, 2017.
82. Hayat, S.; Yanmaz, E.; Muzaffar, R. Survey on Unmanned Aerial Vehicle Networks for Civil Applications: A Communications Viewpoint. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2624–2661. [CrossRef]
83. Larsson, E.G.; Edfors, O.; Tufvesson, F.; Marzetta, T.L. A Massive MIMO for Next Generation Wireless Systems. *IEEE Commun. Mag.* **2016**, *52*, 186–195. [CrossRef]
84. Jovanovic, A.; Li, J.; Richardson, T. Visible light Communication: Opportunities, Challenges and the Path to Market. *IEEE Commun. Mag.* **2013**, *51*, 130–135. [CrossRef]
85. Popovicki, G.; Lauridsen, M.; Soret, B.; Pedersen, K.I.; Mogensen, P. Automation for on-Road Vehicles: Use Cases and Requirements for Radio Design. In Proceedings of the IEEE 82nd Vehicular Technology Conference (VTC Fall), Boston, MA, USA, 6–9 September 2015.
86. Abboud, K.; Omar, H.A.; Zhuang, W. Interworking of DSRC and Cellular Network Technologies for V2X Communications: A Survey. *IEEE Trans. Veh. Technol.* **2016**, *65*, 9457–9470. [CrossRef]
87. Zheng, K.; Zhao, L.; Mei, J.; Dohler, M.; Xiang, W.; Peng, Y. 10 Gb/s HetSNets with Millimeter-Wave Communications: Access and Networking-Challenges and Protocols. *IEEE Commun. Mag.* **2015**, *53*, 222–231. [CrossRef]
88. Jaber, M.; Imran, M.A.; Tafazolli, R.; Tukmanov, A. 5G Backhaul Challendes and Emerging Research Direction: A Survey. *IEEE Access* **2014**, *4*, 1743–1766. [CrossRef]
89. Abdul Salam, S.; Mahmud, S.A.; Khan, G.M.; Raweshidy, H.S. M2M Communication in Smart Grids: Implementation Scenarios and Performance Analysis. In Proceedings of the 2012 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Paris, France, 1 April 2012.

90. 3GPP. Technical Specification Group Services and System Aspects (Release 8). Available online: http://www.arib.or.jp/IMT-2000/V720Mar09/5_Appendix/Rel8/23/23521-800.pdf (accessed on 10 July 2017).
91. Ratasuk, R.; Tan, J.; Ghosh, A. Coverage and capacity analysis for machine type communications in LTE. In Proceedings of the IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 6–9 May 2012.
92. Wu, G.; Talwar, S.; Johnsson, K.; Himayat, N.; Johnson, K.D. M2M: From Mobile to Embedded Internet. *IEEE Commun. Mag.* **2011**, *49*, 36–43. [CrossRef]
93. Chen, M.; Wan, J.; Li, F. Machine-To-Machine Communications: Architectures, Standards and Applications. *KSII Trans. Internet Inf. Syst.* **2012**, *6*, 480–497. [CrossRef]
94. Marjanovic, M.; Skorin-Kapov, L.; Pripužić, K.; Antonić, A.; Podnar Žarko, I. Energy-Aware and Quality-Driven Sensor Management For Green Mobile Crowd Sensing. *J. Netw. Comput. Appl. Arch.* **2015**, *59*, 96–108. [CrossRef]
95. 5GPPP. View on 5G Architecture. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-For-public-consultation.pdf (accessed on 10 July 2017).
96. 5GPP. 5G Empowering Vertical Industries. Available online: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf (accessed on 10 July 2017).
97. 4G Americas. Cellular Technologies Enabling the Internet of Things. Available online: http://www.5gamericas.org/files/6014/4683/4670/4G_Americas_Cellular_Technologies_Enabling_the_IoT_White_Paper_-_November_2015.pdf (accessed on 10 July 2017).
98. Carlesso, M.; Antonopoilos, A.; Granelli, F.; Verikoukis, C. Uplink Scheduling form Smart Metering and Real-Time Traffic Coexistence in LTE Networks. In Proceedings of the International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 820–825.
99. Islam, R.; Avazov, N.; Dobre, O.A.; Kwak, K.S. Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 721–742. [CrossRef]
100. Liu, Y.; Zhang, Y.; Yu, R.; Xie, R. Integrated Energy and Spectrum Harvesting for 5G Wireless Communications. *IEEE Netw.* **2015**, *29*, 75–81. [CrossRef]
101. Kong, L.; Khan, M.K.; Wu, F.; Chen, G.; Zeng, P. Millimeter-Wave Wireless Communications for IoT-Cloud Supported Autonomous Vehicles: Overview, Design, and Challenges. *IEEE Commun. Mag.* **2017**, *55*, 62–68. [CrossRef]
102. Costa-Perez, X.; Garcia-Saavedra, A.; Li, X.; de la Oliva, A.; Iovanna, P.; Deiß, T.; di Giglio, A.; Mourad, A. 5G-Crosshaul: An SDN/NFV Integrated Fronthaul/Backhaul Transport Network Architecture. *IEEE Wirel. Commun.* **2017**, *24*, 38–45. [CrossRef]
103. Trivisonno, R.; Guerzoni, R.; Vaishnavi, I.; Soldani, D. Towards Zero Latency Software Defined 5G Networks. In Proceedings of the IEEE International Conference on Communication Workshop (ICCW), London, UK, 8–12 June 2015.
104. Esaki, H.; Nakamura, R. Overlaying and Slicing for Iot Era Based on Internet's End-To-End Discipline. In Proceedings of the IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), Osaka, Japan, 12–14 June 2017.
105. 5GPP. 5G and Media & Entertainment. Available online: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-Media-Entertainment-Vertical-Sector.pdf (accessed on 10 July 2017).
106. GSA. The Road to 5G: Frivers, Applications, Requirements and Technical Development. Available online: http://www.huawei.com/minisite/5g/img/GSA_the_Road_to_5G.pdf (accessed on 10 July 2017).
107. Erman, B.; Yiu, S. Modeling 5G Wireless Network Service Reliability Prediction with Bayesian Network. In Proceedings of the 2016 IEEE International Workshop Technical Committee on Communications Quality and Reliability, Stevenson, WA, USA, 10–12 May 2016.
108. NOKIA. 5G Use Cases and Requirements White Paper. Available online: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2015/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
109. Motlagh, N.H.; Bagaa, M.; Taleb, T.; Song, J. Connection Steering Mechanism between Mobile Networks for Reliable UAV's Iot Platform. In Proceedings of the IEEE International Conference on Communications (ICC), Paris, France, 21–25 July 2017.
110. Kaltenberger, F.; Knopp, R.; Vitiello, C.; Danneberg, M.; Festag, A. Experimental Anaysis of 5G Candidate Waveform and Their Coexistence with 4G Systems. In Proceedings of the Joint NEWCOM/COST Workshop on Wireless Communications JNCW, Barcelona, Spain, 14–15 October 2016.

111. Serra, J.; Pubill, D.; Antonopoulos, A.; Verikoukis, C. Smart HVAC Control in IoT: Energy Consumption Minimization with User Comfort Constraints. *Sci. World J.* **2014**, *2014*. [CrossRef] [PubMed]
112. Nikolich, P.; Lin, C.; Korhonen, J.; Marks, R.; Tye, B.; Li, G.; Ni, J.; Zhang, S. Standards for 5G and Beyond: Their Use Cases and Applications. *IEEE 5G Tech Focus* **2017**, *1*. Available online: https://5g.ieee.org/tech-focus/june-2017/standards-for-5g-and-beyond (accessed on 9 September 2017).
113. Kusume, K.; Fallgren, M.; Queseth, O.; Braun, V.; Gozalvez-Serrano, D.; Korthals, I.; Zimmermann, G.; Schubert, M.; Hossain, M.I.; Widaa, A.A.; et al. Updated Scenarios, Requirements and KPIs for 5G Mobile and Wireless System with Recommendations for Future Investigations. METIS_D1.5_v1. METIS. Available online: https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D1.5_v1.pdf (accessed on 10 July 2017).
114. 5GPP. 5G Vision and Requirements. Available online: https://5g-ppp.eu/imt-2020-5g-promotion-group-and-5g-ppp-announce-memorandum-of-understanding-for-5g/ (accessed on 10 July 2017).
115. Yang, C.; Li, J. *Interference Mitigation and Energy Management in 5G Heterogeneous Cellular Networks*; ICI Global: Harshey, PA, USA, 2017; pp. 1–362.
116. Ericsson. 5G Energy Performance. Available online: http://www.5gamericas.org/files/3114/3898/7332/Ericsson_wp-5g-energy-performance.pdf (accessed on 10 July 2017).
117. Shen, K.; Yu, W. Distributed Pricing-Based User Association for Downlink Hterogeneous Cellular Networks. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1100–1113. [CrossRef]
118. Nguyen, H.V.; Le, L.B. Distributed Base Station Association and Power Control for Heterogeneous Cellular Networks. *IEEE Trans. Veh. Technol.* **2014**, *63*, 282–296. [CrossRef]
119. Singh, S.; Andrews, J.G. Joint Resource Portitioning and Offloading in Heterogeneous Cellular Networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 888–901. [CrossRef]
120. Elsawy, H.; Hossain, E.; Kim, D.I. HetNets with Cognitive Small Cells: User Offloading and Distributed Channel Access Technique. *IEEE Commun. Mag.* **2013**, *51*, 28–36. [CrossRef]
121. Ji, Z.; Li, K. Cognitive Radios for Dynamic Spectrum Access-Dynamic Spectrum Sharing: A Game Theoretical Overview. *IEEE Commun. Mag.* **2007**, *45*, 88–94. [CrossRef]
122. Aijazand, A.; Aghvami, A.H. PRMA-Based Cognitive Machine-to-Machine Communications in Smart Grid Networks. *IEEE Trans. Veh. Technol.* **2015**, *64*, 3608–3623. [CrossRef]
123. Okasaka, S.; Weiler, R.J.; Keusgen, W.; Pudeyev, A.; Maltsev, A.; Karls, I.; Sakaguchi, K. Proof-of-Concept of a Millimeter-Wave Integrated Heterogeneous Network for 5G Cellular. *Sensors* **2016**, *9*. [CrossRef] [PubMed]
124. Qi, S.; Tesanovic, M. FQAM-FBMC Design and Its Application to Machine Type Communication. In Proceedings of the IEEE 27th Annual International Symposium on PIMRC, Valencia, Spain, 4–8 September 2016.
125. Domouchtsidis, S.G.; Ntouni, G.D.; Kapinas, V.M.; Karagiannidis, G.K. OFDM-IM vs. FQAM: A Comparative Analysis. In Proceedings of the International Conference on Telecommunications, Thessaloniki, Greece, 16–18 May 2016.
126. Hong, S.; Sagong, M.; Lim, C.V.; Cheun, K.; Cho, S. FQAM: A Modulation Scheme for Beyond 4G Cellular Wireless Communication Systems. In Proceedings of the IEEE GC Wkshps, Atlanta, GA, USA, 9–13 December 2013.
127. Perlman, S.; Forenza, A. pCell Wireless Reinvented. Arthemis Networks. Available online: http://www.rearden.com/DIDO/DIDO_White_Paper_110727.pdf (accessed on 10 July 2017).
128. Mesleh, R.Y.; Hass, H.; Sinanovic, S.; Ahn, C.W.; Yun, S. Spatial Modulation. *IEEE Trans. Veh. Technol.* **2008**, *57*, 2228–2241. [CrossRef]
129. Maunder, R.G. The 5G Channel Code Contenders. Available online: https://www.accelercomm.com/sites/accelercomm.com/files/WhitePaper09Aug2016.pdf (accessed on 10 July 2017).
130. Scholl, S.; Weithoffer, S.; When, N. Advanced Iterative Channel Coding Schemes: When Shannon meets Moore. In Proceedings of the 9th International Symposium on Turbo Codes & Iterative Information, Brest, France, 5–9 September 2016.
131. Korpi, D.; Tamminen, J.; Tarunen, M.; Huusari, T.; Choi, Y.-S.; Anttila, L.; Talwar, S.; Valkama, M. Full-Duplex Mobile Device: Pushing the Limit's. *IEEE Commun. Mag.* **2016**, *54*, 80–87. [CrossRef]
132. Hong, S.S.; Brand, J.; Choi, J.I.; Jain, M.; Mehl, J. Applications of Self-Interference Cancellation in 5G and Beyond. *IEEE Commun. Mag.* **2014**, *52*, 114–121. [CrossRef]

133. Duarte, M.; Dick, C.; Sabharwal, A. Experiment-Driven Characterization of Full-Duplex Wireless Systems. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 4296–4307. [CrossRef]
134. Bharadia, D.; McMilin, E.; Katti, S. Full Duplex Radios. In Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, Hong Kong, China, 12–16 August 2013.
135. Čelik, H.; Sung, K.W. On the Feasibility of Blind Dynamic TDD in Ultra-Dense Wireless Networks. In Proceedings of the 2015 IEEE 81st VehicularTechnology Conference (VTC Spring), Glasgow, UK, 11–14 May 2015.
136. Osti, P.; Aalto, S.; Lassila, P. Flow-Level Modeling and Optimization of Intercell Coordination with Dynamic TDD. In Proceedings of the 10th ACM symposium on QoS and security for wireless and mobile networks Q2SWinet '14, Montreal, QC, Canada, 21–26 September 2014.
137. Kerttula, J.; Marttinen, A.; Ruttik, K.; Jäntti, R.; Alam, N. Dynamic TDD on LTE Small Cells. *EURASIP J. Wirel. Commun. Netw.* **2016**, *194*. [CrossRef]
138. Gupta, A.K.; Kulkarni, M.N.; Visotsky, E.; Vook, F.W.; Ghosh, A.; Andrews, J.G.; Heath, R.W. Rate Analysis and Feasibility of Dynamic TDD in 5G Cellular Systems. In Proceedings of the 2016 IEEE International Conference on Communications, Cuala Lumpur, Malaysia, 22–27 May 2016.
139. Sun, H.; Wildemeersch, M.; Sheng, M.; Quek, T.K. D2D Enchanced Heterogeneus Cellular Networks with Dynamic TDD. *IEEE Trans. Wirel. Commun.* **2015**, *14*, 4204–4218. [CrossRef]
140. Sun, H.; Wildemeersch, M.; Sheng, M.; Quek, T.K.; Li, J. Traffic Adaptation and Energy Efficiency for Small Cell Networks with Dynamic TDD. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 3234–3251. [CrossRef]
141. Čelik, H.; Song, K.W. Joint Transmission with Dummy Symbols for Dynamic TDD in Ultra-Dense Deployment. Available online: https://aps.arxiv.org/abs/1702.06045 (accessed on 10 July 2017).
142. Li, J.; Farahvash, S.; Kavehrad, M.; Valenzuel, R. Dynamic TDD and Fixed Cellular Networks. *IEEE Commun. Lett.* **2002**, *4*, 218–220. [CrossRef]
143. Shen, Z.; Khoryae, A.; Eriksson, E.; Pan, X. Dynamic Uplink-Downlink Configuration and Interference Management in TD-LTE. *IEEE Commun. Mag.* **2012**, *50*, 51–59. [CrossRef]
144. Chunlin, Y.; Zhifeng, Y.; Weimin, L.; Yifei, Y. Non-Orthogonal Multiple Access Schemes for 5G. *ZTE Commun.* **2016**, *14*, 11–16. [CrossRef]
145. 3GPP. Discussion on Multiple Access for Radio Interface. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-84b--31661.htm (accessed on 10 July 2017).
146. Yuan, Z.; Yu, G.; Li, W.; Yuan, Y.; Wang, X.; Xu, J. Multi-user Shared Access for Internet of Things. In Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, China, 15–18 May 2016.
147. 3GPP. Receiver Implementation for MUSA. 3GPP. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-85--31662.htm (accessed on 10 July 2017).
148. 3GPP. Contetion-Based Non-Orthogonal Multiple Access for UL mMTC. Available online: http://www.3gpp.org/DynaReport/TDocExMtg--R1-85--31662.htm (accessed on 10 July 2017).
149. 3GPP. Resource Spread Multiple Access. Available online: http://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_85/Docs/ (accessed on 10 July 2017).
150. Taherzadeh, M.; Nikopour, H.; Bayesteh, A.; Baligh, H. SCMA Codebook Design. In Proceedings of the IEEE International Symposium on Personal, Indoor And Mobile Radio Communications, London, UK, 14–17 September 2014.
151. Nikopour, H.; Baligh, H. Sparse Code Multiple Acccess. In Proceedings of the 2014 IEEE 80th Vehicular Technology Conference, Vancouver, BC, Canada, 14–17 September 2014.
152. Future Mobile Communication Forum. 5G Whitepaper v2.0, Part d-Alternative Multiple Access v1. Available online: http://www.future-forum.org/dl/151106/whitepaper.rar (accessed on 10 July 2017).
153. 3GPP. Candidate Solution for New Multiple Access. 3GPP. Available online: http://portal.3gpp.org/ngppapp/CreateTdoc.aspx?mode=view&contributionId=697587 (accessed on 10 July 2017).
154. Dai, X.; Chen, S.; Sun, S.; Kang, S.; Wang, Y.; Shen, Z.; Xu, Z. Successive Interference Cancelation Amenable Multiple Access (SAMA) for Future Wireless Communications. In Proceedings of the IEEE International Conference on Communication Systems, Macau, China, 19–21 November 2014.
155. Dai, X. Successive Interference Cancellation Amenable Space-Time Codes with Good Multiplexing-Diversity Tradeoff. *Wirel. Pers. Commun.* **2010**, *55*, 645–654. [CrossRef]

156. Li, P.; Liu, L.; Wu, K.; Leung, W.K. On Interleave-Division Multiple-Access. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
157. Li, P.; Liu, L.; Wu, K.; Leung, W.K. Interleave Division Multiple-Access. *IEEE Trans. Wirel. Commun.* **2006**, *5*, 938–947. [CrossRef]
158. Rabee, F.A.; Dayaslioglu, K.; Gitlin, R. The Optimum Received Power Levels of Uplink Non-Orthogonal Multiple Access (NOMA) Signals. In Proceedings of the 2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON), Cocoa Beach, FL, USA, 24–25 April 2017.
159. Saito, Y.; Kishiyama, Y.; Benjebbour, A.; Nakamura, T.; Li, A.; Higuchi, K. Non-orthogonal Multiple Access (NOMA) for Cellular Future Radio Access. In Proceedings of the 77th IEEE Vehicular Technology Conference, Dresden, Germany, 2–5 June 2013.
160. Choi, J. On Multiple Access Using H-ARQ with SIC Techniques for Wireless Ad Hoc Networks. *Wirel. Pers. Commun.* **2013**, *69*, 187–212. [CrossRef]
161. Choi, J. Non-orthogonal Multiple Access in Downlink Coordinated Two-point Systems. *IEEE Commun. Lett.* **2014**, *18*, 313–316. [CrossRef]
162. Sedtheetorn, P.; Chulajata, T. Accurate Spectral Efficiency Analysis for Non Orthogonal Multiple Access. *Trans. Adv. Commun. Technol.* **2016**, *5*, 844–850.
163. Wild, T.; Schaich, F.; Chen, Y. 5G Air Interface Design Based on Universal Filtered (UF-) OFDM. In Proceedings of the 19th International Conference in DSP, Hong Kong, China, 20–23 August 2014.
164. Ding, Z.; Yang, Z.; Fan, P.; Poor, H.V. On the Performance of Non-orthogonal Multiple Access in 5G Systems with Randomly Deployed Users. *IEEE Signal Process. Lett.* **2014**, *21*, 1501–1505. [CrossRef]
165. Zhao, Z.; Schellmann, M.; Wang, Q.; Gong, X.; Boehnke, R.; Xu, W. Pulse shaped OFDM for Asynchronous Uplink Access. In Proceedings of the Asilomar Conference Signals, Systems and Computers, Pacific Grove, CA, USA, 8–11 November 2015.
166. Cover, T.M.; El Gamel, A. Capacity Theorems for The Relay Channel. *IEEE Trans. Inf. Theory* **1979**, *25*, 572–584. [CrossRef]
167. Cover, T.M.; Leung, C.S.K. An Achievable Rate Region for the Multiple-access Channel with Feedback. *IEEE Trans. Inf. Theory* **1981**, *27*, 292–298. [CrossRef]
168. Ahlswede, R. Multiway Communication Channels. In Proceedings of the 2nd International Symposium on Information Theory, Tsahkadsor, Armenian SSR, 2–8 September 1971.
169. Xie, L.L.; Kumar, P.R. An Achievable Rate for the Multiple-level Relay Channel. *IEEE Trans. Inf. Theory* **2005**, *51*, 1348–1358. [CrossRef]
170. Kim, K.T.; Ahn, S.-K.; Kim, Y.-H.; Park, H.; Wang, L.; Chen, C.-Y.; Park, J. Adaptive Sliding-Window Coded Modulation in Cellular Networks. In Proceedings of the IEEE GLOBECOM, San Diego, CA, USA, 6–10 December 2015.
171. Thurfjell, M.; Ericsson, M.; de Bruin, P. Network Densification Impact on System Capacity. In Proceedings of the 2015 IEEE 81st Vehicular Technology Conference, Glasgow, Scotland, 11–14 May 2015.
172. Darwish, A.; Ibrahim, A.S. Capacity Improvements via Indoor Small Cells. In Proceedings of the 2014 International Wireless Communications and Mobile Computing Conference (IWCMC), Nicosia, Cyprus, 4–8 August 2014.
173. Hoadley, J.; Maveddat, P. Enabling Small Cell Deployment with Hetnet. *IEEE Wirel. Commun.* **2012**, *19*, 4–5. [CrossRef]
174. Hughes, M.; Jovanovic, V.M. Small Cells-Effective Capacity Relief Option for Heterogeneous Networks. In Proceedings of the 2012 IEEE Vehicular Technology Conference, Quebec City, QC, Canada, 3–6 September 2012.
175. Sui, Y.; Guvenc, I.; Svensson, T. Interference Management for Moving Networks in Ultra-Dense Urban Scenarios. *EURASIP J. Wirel. Commun. Netw.* **2015**, *111*. [CrossRef]
176. Sui, Y.; Vihrälä, J.; Papadogiannis, A.; Sternad, M.; Svensson, T. Moving Cells: A Promising Solution to Boost Performance for Vehicular Users. *IEEE Commun. Mag.* **2013**, *51*, 62–68. [CrossRef]
177. Sambo, Y.A.; Shakir, M.Z.; Héliot, F.; Imran, M.A.; Mumtaz, S.; Qaraqe, K.A. Device-to-Device Communication in Heterogeneous Networks. In *Smart Device to Smart Device Communication*; Mumtaz, S., Rodriguez, J., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 219–235.
178. Liu, X.; Andrews, J.G.; Ghosh, A. Spectrum Sharing for Device- to-Device Communication in Cellular Networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 6272–6740. [CrossRef]

179. Zafar, B.; Gherekhloo, S.; Haardt, M. Analysis of Multihop Relaying Networks: Communication Between Range-Limited and Cooperative Nodes. *IEEE Veh. Technol. Mag.* **2012**, *7*, 40–47. [CrossRef]
180. Kamalinejad, P.; Mahapatra, C.; Sheng, Z.; Mirabbasi, S.; Leung, V.C.M.; Guan, Y.L. Wireless Energy Harvesting for the Internet of Things. *IEEE Commun. Mag.* **2015**, *53*, 102–108. [CrossRef]
181. Mekikis, P.-V.; Antonopoulos, A.; Kartsakli, E.; Lalos, A.S.; Alonso, L.; Verikoukis, C. Information Exchange in Randomly Deploys Dense WSNs with Wireless Energy Harvesting Capabilities. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 3008–3018. [CrossRef]
182. Mekikis, P.-V.; Antonopoulos, A.; Alonso, L.; Verikoukis, C. Stochastic Modeling of Wireless Charged Wearables for Reliable Health Monitoring In Hospital Environments. In Proceedings of the International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
183. Li, Y.; Cai, L. UAV-Assisted Dynamic Coverage in a Heterogeneous Cellular System. *IEEE Netw.* **2017**, *31*, 56–61. [CrossRef]
184. Messous, M.-A.; Senouci, S.-M.; Sedjelmaci, H. Network Connectivity and Area Coverage for UAV Fleet Mobility Model with Energy Constraint. In Proceedings of the Wireless Communications and Networking Conference (WCNC), Doha, Qatar, 3–6 April 2016.
185. Mekikis, P.-V.; Kartsakli, E.; Alonso, L.; Verikoukis, C. Flexible Aerial Relay Nodes for Communication Recovery and D2D Relaying. In Proceedings of the IEEE 5th Global Conference on Consumer Electronics, Kyoto, Japan, 11–14 October 2016.
186. Zeng, Y.; Zhang, R.; Lim, T.J. Wireless Communications with Unmanned Aerial Vehicles: Opportunities and Challenges. *IEEE Commun. Mag.* **2016**, *54*, 36–42. [CrossRef]
187. Yang, X.; Lu, W.-J.; Wang, N.; Nieman, K.; Jin, S.; Zhu, H.; Mu, X.; Wong, I.; Huang, Y.; You, X. Design and Implementation of a TDD-Based 128-Antenna Massive MIMO Prototyping System. Available online: http://adsabs.harvard.edu/cgi-bin/bib_query?arXiv:1608.07362 (accessed on 11 July 2017).
188. Chen, B.; Zhu, C.; Shu, L.; Su, M.; Wei, J.; Leung, V.C.M.; Rodriques, J.J.P.C. Uplink Transmission for Lightweight Single-antenna UEs in the Presence of a Massive MIMO eavesdropper. *IEEE Access* **2016**, *99*, 2169–3536. [CrossRef]
189. Wu, S.; Wang, H.; Youn, C.H. Visible Light Communications for 5G Wireless Networking Systems: From Fixed to Mobile. *IEEE Access* **2014**, *28*, 41–45. [CrossRef]
190. Feng, L.; Hu, R.Q.; Wang, J.; Xu, P.; Qian, Y. Applying VLC in 5G Networks: Architectures and Key Technologies. *IEEE Access* **2016**, *30*, 77–83. [CrossRef]
191. Grobe, L.; Paraskevopoulos, A.; Hilt, J.; Schulz, D.; Lassak, F.; Hartlieb, F.; Kottke, C.; Jungnickel, V.; Langer, K.-D. High-Speed Visible Light Communication Systems. *IEEE Commun. Mag.* **2013**, *51*, 60–66. [CrossRef]
192. Li, X.; Zhang, R.; Wang, J.; Hanzo, L. Cell-Centric and User-Centric Multi-User Scheduling in Visible Light Communication Aided Networks. In Proceedings of the IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015.
193. Katriniok, A.; Kleibaum, P.; Res, C.; Eckstein, L. Automation of Road Vehicles Using V2X: An Application to Intersection Automation. Available online: http://papers.sae.org/2017-01-0078/ (accessed on 10 July 2017).
194. Ghods, A.; Severi, S.; Abreu, G. Localization in V2X Communication Networks. In Proceedings of the IEEE Intelligent Vehicles Symposium, Gothenburg, Sweden, 19–22 June 2016.
195. Goebel, N.; Bialon, R.; Mauve, M.; Graffi, K. Coupled Simulation of Mobile Cellular Networks, Road Traffic and V2X Applications Using Traces. In Proceedings of the International Conference on Communications, Kuala Lumpur, Malaysia, 22–27 May 2016.
196. Xiao, Z.; Liu, H.; Havyarimana, V.; Li, T.; Wang, D. Analytical Study on Multi-Tier 5G Heterogeneous Small Cell Networks: Coverage Performance and Energy Efficiency. *Sensors* **2016**, *11*. [CrossRef] [PubMed]
197. Choi, J.I.; Jain, M.; Srinivasan, K.; Levis, P.; Katti, S. Achieving Single Channel, Full Duplex Wireless Communication. In Proceedings of the sixteenth annual international conference on Mobile computing and networking, Chicago, IL, USA, 20–24 September 2010.
198. Song, Q.; Nuaymi, L.; Lagrange, X. Survey of Radio Resource Management Issues and Proposals for Energy-Efficient Cellular Network That Will Cover Billions of Machines. *Wirel. Commun. Netw.* **2016**, *140*. [CrossRef]

199. Costa, G.; Miao, G. Context-Aware Machine-To-Machine Communications. In Proceedings of the 33rd Annual IEEE International Conference on Computer Communications (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014.
200. Agyapong, P.K.; Iwamura, M.; Staehle, D.; Kiess, W.; Benjebbour, A. Design Considerations for a 5G Network Architecture. *IEEE Commun. Mag.* **2014**, *52*, 65–75. [CrossRef]
201. Vermesan, O.; Friass, P.; Guillemin, P.; Giaffreda, R.; Grindvoll, H.; Eisenhauer, M.; Serrano, M.; Moessner, K.; Spirito, M.; Blystad, L.-C.; et al. *Building the Hypercconnected Society—IoT Research and Innovation Value Chains, Ecosystems and Markets*; Vermesan, O., Friass, P., Eds.; River Publishers: Aalborg, Denmark, 2015; Volume 43.
202. Panwar, N.; Sharma, S.; Singh, A.K. A Survey on 5G: The Next Generation of Mobile Communication. *Elsevier Phys. Commun.* **2016**, *18*, 64–84. [CrossRef]
203. Thuemmler, C.; Gavrasm, A.; Jumelle, A.K.L.; Paulin, A.; Sadique, A.; Schneider, A.; Fedell, C.; Abraham, D.; Trossen, D. 5G and e-Health. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf (accessed on 10 July 2017).
204. Rose, K.; Eldridge, S.; Chaplin, L. The Internet of Things: An Overview. Available online: https://www.internetsociety.org/sites/default/files/ISOC-IoT-Overview-20151014_0.pdf (accessed on 10 July 2017).
205. Huawei. 5G Opening up New Business Opportunities. Available online: http://www-file.huawei.com/-/media/CORPORATE/PDF/x-lab/10-5g-opening-up-new-business-opportunities-en.pdf?la=en (accessed on 10 July 2017).
206. Lin, H.; Shao, J.; Zhang, C.; Fang, Y. CAM: Cloud-Assisted Privacy Preserving Mobile Health Monitoring. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 985–997. [CrossRef]
207. NGMN Alliance. NGMN 5G White Paper. Available online: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf (accessed on 10 July 2017).
208. Le, N.T.; Hossain, M.A.; Islam, A.; Kim, D.-Y.; Choi, Y.-J.; Jang, Y.M. Survey of Promising Technologies for 5G Networks. *Hindawi* **2016**, *2016*. [CrossRef]
209. Militano, L.; Araniti, G.; Condoluci, M.; Farris, I.; Iera, A. Device-to-Device Communications for 5G Internet of Things. *EAI Endorsed Trans. Internet Things* **2015**, *1*. [CrossRef]
210. Shariatmadari, H.; Ratasuk, R.; Iraji, S.; Laya, A.; Taleb, T.; Ghosh, A. Machine-type Communications: Current Status and Future Perspective toward 5G Systems. *IEEE Commun. Mag.* **2015**, *53*, 10–17. [CrossRef]
211. 5GPPP. 5G and the Factories of the Future. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future-Vertical-Sector.pdf (accessed on 10 July 2017).
212. Elayoubi, S.E.; Fallgren, M.; Spapis, P.; Zimmermann, G.; Martín-Sacristán, D.; Yang, C.; Jeux, S.; Agyapong, P.; Campoy, L.; Qi, Y.; et al. 5G Service Requirements and Operational Use Cases: Analysis and METIS II Vision. In Proceedings of the 2016 European Conference on Networks and Communications (EuCNC), Athens, Greece, 27–30 June 2016.
213. Blanco, B.; Fajardo, H.O.; Giannis, I.; Kafetzakis, E.; Pneg, S.; Perez-Romero, J.; Trajkovska, I.; Khodashenas, P.S.; Goratti, L.; Paolino, M.; et al. Technology Pillars in the Architecture of Future 5G Mobile Networks: NFV, MEC and SDN. *Comput. Stand. Interfaces* **2017**, *54*, 216–228. [CrossRef]
214. ITU-R. Recommendation ITU-R M.2083-0. IMT Vision: Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf (accessed on 10 July 2017).
215. Le, L.B.; Lau, V.; Jorswieck, E.; Dao, N.-D.; Haghighat, A.; Kim, D.I.; Le-Ngoc, T. Enabling 5G Mobile Wireless Technologies. *EURASIP J. Wirel. Commun. Netw.* **2015**, *218*. [CrossRef]
216. Liu, G.; Jiang, D. 5G: Vision and Requirements for Mobile Communication System towards Year 2020. *Chin. J. Eng.* **2016**, *2016*, 5974586. [CrossRef]
217. Bogale, T.E.; Le, L.B. Massive MIMO and mmWave for 5G Wireless HetNet: Potential Benefit's and Challenges. *IEEE Veh. Technol. Mag.* **2016**, *11*, 64–75. [CrossRef]
218. Banelli, P.; Buzzi, S.; Colavolpe, G.; Modenini, A.; Rusek, F.; Ugolini, A. Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency. *IEEE Signal Process. Mag.* **2014**, *31*, 80–93. [CrossRef]
219. Palattella, M.R.; Dohler, M.; Grieco, A.; Rizzo, G.; Torsner, J.; Engel, T.; Ladidm, L. Internet of Things in the 5G Era: Enablers, Architecture and Business Models. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 510–527. [CrossRef]

220. Pries, R.; Morper, H.-J.; Galambosi, N.; Jarschel, M. Network as a Service—A Demo on 5G Network Slicing. In Proceedings of the 28th International Teletraffic Congress (ITC 28), Würzburg, Germany, 12–16 September 2016.
221. Caraguay, A.L.V.; Peral, A.B.; Lopez, L.I.B. SDN: Evolution and Opportunities in the Development IoT Applications. *Int. J. Distrib. Sens. Netw.* **2014**, *10*. [CrossRef]
222. Bizanis, N.; Kuipers, F.A. SDN and Virtualization Solutions for the Internet of Things: A Survey. *IEEE Access* **2016**, *4*, 5591–5606. [CrossRef]
223. Stancu, A.; Vulpe, A.; Fratu, O.; Halunga, S. Default Values Mediator Used for a Wireless Transport SDN Proof of Concept. In Proceedings of the IEEE Conference on Standards for Communications and Networking (CSCN), Berlin, Germany, 31 October–2 November 2016.
224. Warren, D.; Dewar, C. Understanding 5G: Perspectives on Future Technological Advancements in Mobile. Available online: https://www.gsmaintelligence.com/research/?file=141208-5g.pdf&download (accessed on 10 July 2017).
225. Rappaport, T.S.; Heath, R.V.; Daniels, R.C.; Murdock, J.N. *Millimeter Wave Wireless Communications*, 1st ed.; Pearson: Harshey, PA, USA, 2017; pp. 1–657.
226. Thakare, S. Research Challenges of Millimeter Wave Technology in 5G Cellular System. In Proceedings of the 4th International Conference on Recent Trends in Engineering & Technology (ICRTET), Nashik, India, 2–4 July 2015.
227. Ye, J.; He, J.; Ge, X.; Chen, M. Energy Efficiency Analysis of 5G Ultra-dense Networks Based on Random Way Point Mobility Models. In Proceedings of the 2016 19th International Symposium on Wireless Personal Multimedia Communications, Shenzhen, China, 14–16 November 2016.
228. Witrisal, K.; Meissner, P.; Leitinger, E.; Shen, Y.; Gustafson, C.; Tufvesson, F.; Haneda, K.; Dardari, D.; Molisch, A.F.; Conti, A.; et al. High-Accurany Localization for Assisted Living: 5G Systems Will Turn Multipath Channels from Foe to Friend. *IEEE Signal Process. Mag.* **2016**, *33*, 59–70. [CrossRef]
229. 5GPP. 5G Automotive Vision. Available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf (accessed on 10 July 2017).
230. Siaud, I.; Umner-Moll, A.M. Green-Oriented Multi-Techno Link Adaptation Metrics for 5G Heterogeneous Networks. *EURASIP J. Wirel. Commun. Netw.* **2016**, *92*. [CrossRef]
231. Le, L.B.; Niyato, D.; Hossain, E.; Kim, D.I.; Hoang, D.T. QoS-Aware and Energy-Efficient Resource Management in OFDMA Femtocells. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 180–194. [CrossRef]
232. Ye, Q.; Rong, B.; Chen, Y.; Al-Shalash, M.; Caramanis, C.; Andrews, J.G. User Association for Load Balancing in Heterogeneous Cellular Networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 2706–2716. [CrossRef]
233. Mavromoustakis, C.X.; Bourdena, A.; Mastorakis, G.; Pallis, E.; Kormentzas, G. Coverage An Energy-Aware Scheme for Efficient Spectrum Utilization in A 5G Mobile Cognitive Radio Network Architecture. *Telecommun. Syst.* **2015**, *59*, 63–75. [CrossRef]
234. Wu, S.; Wang, Y.; Al Imari, M.; Nekovee, M. FQAM: Frequency and Quadrature Amplitude Modulation for 5G Networks. In Proceedings of the 2016 European Conference on Networks and Communications (EuCNC), Athens, Greece, 27–30 June 2016.
235. Al-Imari, M.; Mouhouche, B.; Al.Imari, M.; Nekovee, M.F. Non-Orthogonal FQAM for Multiple Access in the Uplink of 5G Wireless Networks. In Proceedings of the 2016 International Symposium on Wireless Communication Systems (ISWCS), Poznań, Poland, 20–23 September 2016.
236. Hong, S.; Sagong, M.; Lim, C.; Cho, S.; Cheun, K.; Yang, K. Frequency and Quadrature Amplitude Modulation for Downlink Cellular OFDMA Networks. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1256–1267. [CrossRef]
237. Nekovee, M.; Yue, W.; Tesanovic, M.; Wu, S.; Qi, Y.; Al-Imari, M. Overview of 5G Modulation and Waveforms Candidates. *J. Commun. Inf. Netw.* **2016**, *1*, 44–60. [CrossRef]
238. Hochwald, B.M.; Sweldens, W. Differential Unitary Space-Time Modulation. *IEEE Trans. Commun.* **2000**, *48*, 2041–2052. [CrossRef]
239. Ostman, J.; Yang, W.; DUrisim, G.; Koch, T. Diversity Versus Multiplexing at Finite Blocklength. In Proceedings of the 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, Spain, 26–29 August 2014.
240. Devaillie, B.; van Liempd, B.; Hershberg, B.; Craninckx, J.; Rikkinen, K.; van den Broek, D.J.; Klumperink, E.A.M.; Nauta, B. In-Band Full-Duplex Transceiver Technology for 5G Mobile Networks. In Proceedings of the ESSCIRC Conference, Graz, Austria, 4–8 December 2016.

241. Kim, C.; Kim, K.; Yun, Y.H.; Ho, Z.; Lee, B.; Seol, J.-Y. QAM-FBMC: A New Multi-carrier System for Post-OFDM Wireless Communications. In Proceedings of the IEEE Global Communication Conference GLOBECOM, San Diego, CA, USA, 6–10 December 2015.
242. Fettweis, G.; Krondorf, M.; Bittner, S. GFDM-generalized Frequency Division Multiplexing. In Proceedings of the IEEE 69th VTC, Barcelona, Spain, 26–29 April 2009.
243. Abdoli, G.; Jia, M.; Ma, J. Filtered OFDM: A New Waveform for Future Wireless Systems. In Proceedings of the IEEE 16th International Workshop SPAWC, Stockholm, Sweden, 28 June–1 July 2015.
244. Li, J.; Kearney, K.; Bala, E.; Yang, R. Filtered OFDM: A Resource Block Based Filtered OFDM Scheme and Performanse Comparison. In Proceedings of the 2013 20th International Conference on Telecommunications, Casablanca, Morocco, 6–8 May 2013.
245. Kumar, U.; Ibars, C.; Bhorkar, A.; Jung, H. Filtered OFDM: A Waveform for 5G: Guard Interval DFT-S-OFDM. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), San Diego, CA, USA, 6–10 December 2015.
246. Chung, C.D. Spectrally Precoded OFDM. *IEEE Trans. Commun.* **2006**, *54*, 2173–2185. [CrossRef]
247. Monk, A.; Hadani, R.; Tsatsanis, M.; Rakib, S. OTFS—Orthogonal Time Frequency Space. Available online: https://arxiv.org/ftp/arxiv/papers/1608/1608.02993.pdf (accessed on 10 July 2017).
248. Bandemer, B.; Gamal, A.E.; Kim, Y.H. Optimal Achievable Rates for Interference Networks with Random Codes. *IEEE Trans. Inf. Theory* **20125**, *61*, 6536–6549. [CrossRef]
249. Carleial, A.B. Multiple-access Channels with Different Generalized Feedback Signals. *IEEE Trans. Inf. Theory* **1982**, *28*, 841–850. [CrossRef]
250. Prasad, A.; Uusitalo, M.A.; Li, Z.; Lunder, P. Reflection Environment Maps for Enhanced Reliability in 5G Self-Organizing Networks. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA, 4–8 December 2016.
251. Sui, Y.; Papadogiannis, A.; Sternad, M. The Potential of Moving Relays-A Performance Analysis. In Proceedings of the 2012 IEEE 75th Vehicular Technology Conference (VTC Spring), Yokohama, Japan, 6–9 May 2012.
252. Jiang, L.; Tian, H.; Xing, K.; Zhang, K.; Maharjan, S.; Cjessing, S.; Zhang, Y. Social-aware Energy Harvesting Device-to-Device Communications in 5G Networks. *IEEE Wirel. Commun.* **2016**, *23*, 20–27. [CrossRef]
253. Huang, X.; Yu, R.; Kang, J.; Gao, Y.; Maharjan, S.; Gjessing, S.; Zhang, Y. Software Defined Energy Harvesting Networking for 5G Green Communications. *IEEE Wirel. Commun.* **2017**, *24*, 38–45. [CrossRef]
254. Merwaday, A.; Guvenc, I. UAV Assisted Heterogeneous Networks for Public Safety Communications. In Proceedings of the IEEE Wireless Communications and Networking Conference Workshops (WCNCW), New Orleand, LA, USA, 9–12 March 2015.
255. Osseiran, A. The 5G Radio-Access Technologies. In *5G Mobile and Wireless Communications Technology*, 1st ed.; Osseiran, A., Monserrat, J.F., Marsch, P., Dohler, M., Nakamura, T., Eds.; Cambridge University Press: Cambridge, UK, 2016.
256. Sahin, A.; Yang, R.; Bala, E.; Beluri, M.C.; Olesen, R.L. Flexible DFT-S-OFD: Solutions and Challenges. *IEEE Commun. Mag.* **2016**, *54*, 106–112. [CrossRef]
257. Shi, S.; Yang, W.; Zhang, J.; Chang, Z. Review of Key Technologies of 5G Wireless Communication System. *MATEC Web Conf.* **2015**, *22*. [CrossRef]
258. Wang, C.-X.; Haider, F.; Gao, X.; You, X.-H.; Yang, Y.; Yuan, D.; Aggoune, H.; Haas, H.; Fletcher, S.; Hepsaydir, E. Cellular Architecture and Key Technologies for 5G Wireless Communication Networks. *IEEE Commun. Mag.* **2014**, *52*, 122–130. [CrossRef]

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

5G VISION

DMC R&D Center, Samsung Electronics Co., Ltd.

Date : February 2015

Samsung envisions the fifth Generation (5G) mobile communication era to be the beginning of a full scale Internet of Things (IoT). Billions of connected devices autonomously interconnect with one another while ensuring personal privacy. The unprecedented latencies offered by 5G Networks will enable users to indulge in gigabit speed immersive services regardless of geographical and time dependent factors. This white paper introduces you to future services, key requirements, and enabling technologies that will herald in the 5G era that is expected to revolutionize the way we experience mobile services.

Dawn of the 5G Era

Fuelled by the unprecedented growth in the number of connected devices and mobile data, and the ever-fast approaching limits of the 4G technologies to address this enormous data demand, industry efforts and investments to define, develop and deliver the systems and specifications for the fifth-Generation (5G) mobile system and services are well under way - signaling the dawn of the 5G Era.

As shown in Figure 1, the number of connected Internet of Things (IoT) is estimated to reach 50 Billion by 2020 [1], while the mobile data traffic is expected to grow to 24.3 Exabytes per month by 2019 [2].

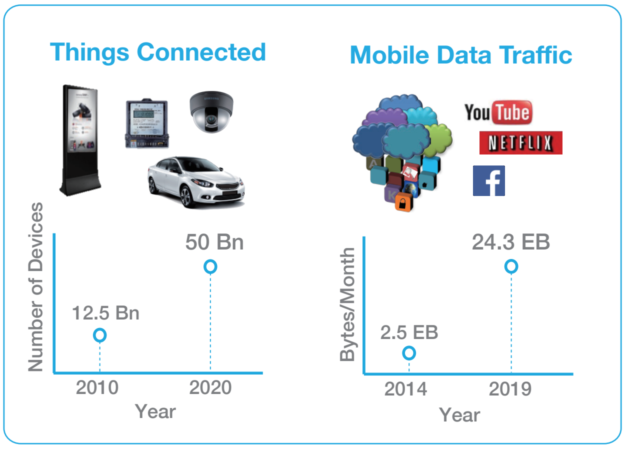

Figure 1 Growth in Mobile Traffic and Connected Devices

Add to it, the impact of higher cell capacity and end-user data rate requirements due to emerging new services such as Ultra-High-Definition (UHD) multimedia streaming and extremely low latency requirements for cloud computing and storage/ retrieval, and it soon becomes evident that the current 4G systems, which are already stretched to near-breaking points (despite massive Wi-Fi offload), will be stretched too thin to deliver the quality of experience (QoE) necessary to support mobile experience that 5G is set to deliver.

5G Era can be expected to revolutionize the way we communicate by supporting immersive applications that demand exceptionally high-speed wireless connections, a fully-realized IoT, experience lower latency and promote both spectrum and energy efficiency. To realize these benefits, 5G systems will differ fundamentally from their predecessors fueling a series of groundbreaking innovations. Let's look at the services and the requirements that 5G is expected to address.

5G Service Vision

5G services have the potential to revolutionize the mobile experience. Here's how:

Internet of Things

5G will make the "Internet of Things" a reality. With 5G technology, a device will be able to maintain network connectivity regardless of time and location, and open the possibility to connect all the connected devices without human intervention. For this, the basic fabric of the 5G system design is expected to provide support for up to a million simultaneous connections per square kilometer, enabling a variety of machine-to-machine services including wireless metering, mobile payments, smart grid and critical infrastructure monitoring, connected home, smart transportation, and telemedicine. Intelligent devices will communicate with each other autonomously in the background and share information freely. This ubiquitous connectivity - a basic tenet of the 5G services, will truly enable IoT services which in turn is expected to profoundly change human lives by connecting virtually everything.

– Smart Home

Dishwashers will fix themselves using information from peers of the same model while home appliances at home and in the neighboring homes may cooperate to extinguish a fire. A smart refrigerator, recommending a recipe of cuisine to be cooked with ingredients that are already in your refrigerator, is yet another plausible scenario.

– Fitness & Healthcare

Connected Health and Fitness related wearable devices such as The Samsung Gear^{TM} Fit will record your athletic performance while you exercise and recommend the type of exercise, its duration and frequency per day. These connected healthcare devices will also send vital signs such as brainwave, blood pressure and heartbeat to an expert system in the hospital in real-time to prevent medical emergencies before they occur. Such time critical applications put unprecedented requirements on latency.

– Smart Store

In large shopping malls, while many people walk around window shopping, vicinity to products is continuously tracked, usually by a server somewhere in the cloud. Customized alerts for low priced product can be sent to the user's device as the user is detected in the vicinity of that low-priced product, or information of other lower prices in the nearby stores can be sent to the device as the customer spends more time in the vicinity of a certain product class.

Such a system can be tailored to deliver a highly customized experience thereby greatly enhancing a user's shopping experience. To support such a scenario, massive connectivity and low latency technologies are inevitable.

– Smart Office

In smart office environments, office appliances are connected with one another and will share information. Nearby computers and input/output devices can recognize a user and change the settings using the user's preferences stored in the IoT cloud. Printers will print out the relevant documents when the user passes by the printer. Almost all the office appliances will connect wirelessly, while exchanging massive data through wireless medium without noticeable delay. Alerts on the upcoming meetings, materials and documents relevant to that meeting will instantaneously become available to the user's device, while documents and tasks that are modified will be automatically updated.

– Connected Car

Many of us use navigation services via in-car navigation systems or our smart phones to find the most efficient route to our destination. Vehicle diagnostic services are becoming attractive to obtain the information such as battery level, fuel level or engine status on our smartphones. The 'eCall' system that automatically calls emergency services in case of an emergency exemplifies such a service. By 2020 and beyond, more attractive services that wirelessly connect 'cars' around us with 'things' will emerge and make people in the vehicles very comfortable, and provide an enjoyable driving experience. For safer driving, sensor and camera data in a vehicle as well as supplementary information from the neighboring vehicles will be collected using mobile networks so that a potential emergency situation can be reliably informed to a driver in real-time and timely steps can be taken to avoid an actual emergency situation. This operation will eventually be applied to self-driving cars, which can be viewed as an important type of 'things'.

Immersive Multimedia Experience

In a 5G environment, users will experience life-like multimedia streams anytime and anywhere. Users will feel as if they are part of the scene when they watch videos on their smart devices. To provide such an immersive experience, many obstacles will need to be overcome. Agility to instantaneously respond to the user's thought and behavior will be necessary. An upcoming service that is expected to provide life-like experience in 5G system is UHD video streaming with its greatly enhanced resolution and clarity. Currently, UHD services over terrestrial broadcast are already being standardized in some countries. In addition, some smart phones in the market are now equipped with a camera module

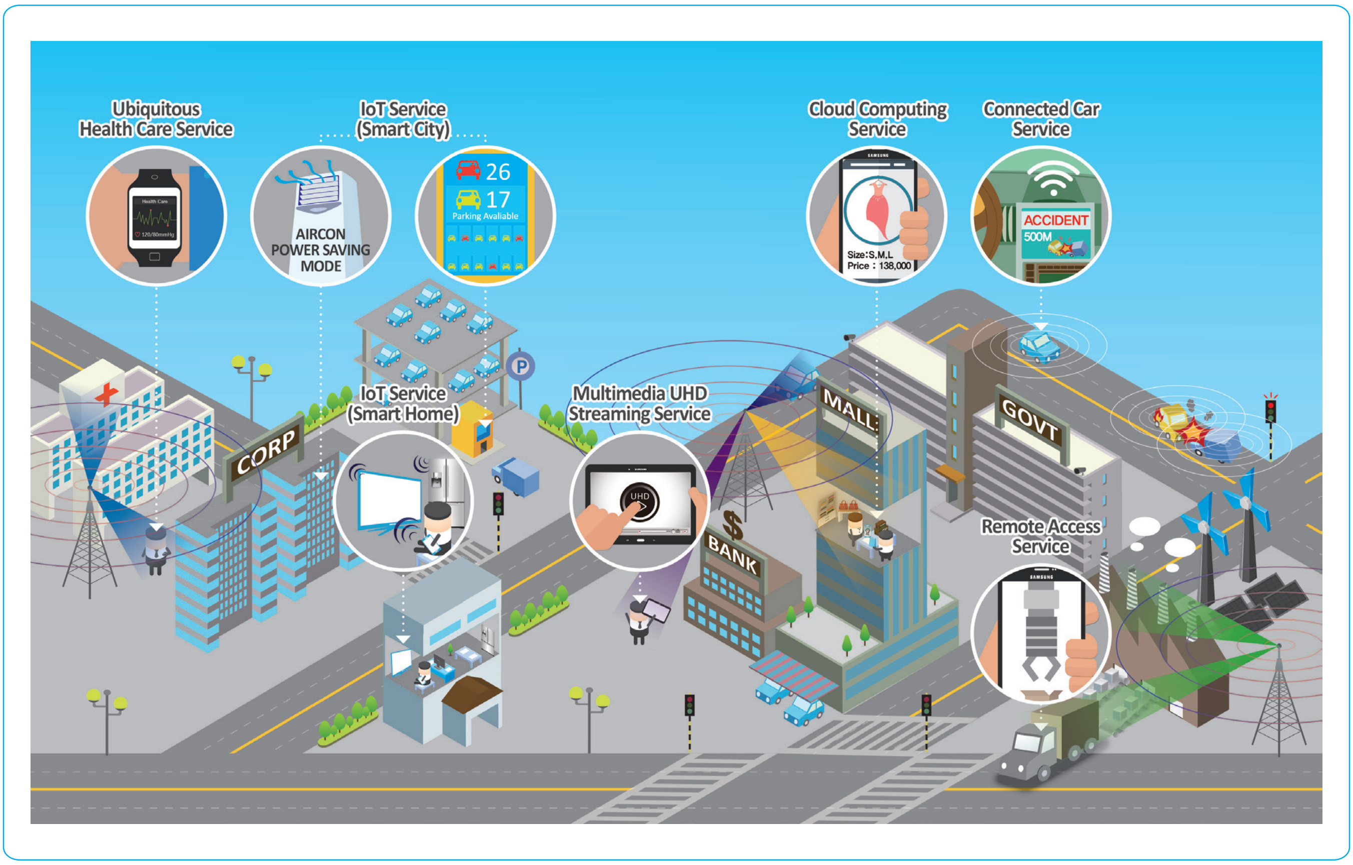
Ubiquitous Health Care Service
Health Care
120/80mmHg
IoT Service (Smart City)
AIRCON POWER SAVING MODE
26
17
Parking Avaliable
Cloud Computing Service
SAMSUNG
Size : S,M,L
Price : 138,000
Connected Car Service
ACCIDENT
500M
P
IoT Service (Smart Home)
Multimedia UHD Streaming Service
UHD
CORP
MALL
GOVT
BANK
Remote Access Service
SAMSUNG

Figure 2 Major Service Scenarios with 5G

that can record video with 4K UHD quality. It is expected that UHD services are likely to go mainstream by 2020 raising an acute need for performance enhancement of cellular systems to support such services.

Other examples of “immersive” services that will fundamentally revolutionize entertainment, health, education, and other industry sectors are Virtual Reality (VR) and Augmented Reality (AR).

VR provides a world where physical presence is simulated by computer graphics, and the user can actively interact with the simulated elements, as in immersive sports broadcasting for instance (See Figure 3). Other interesting VR service scenarios are interactive 360° movies, online games, remote education, and virtual orchestra. Samsung’s recently launched VR headset called (called Gear VR), virtual reality video service platform (called Milk VR) and ‘Project Beyond,’ (a 360° 3D camera with 17 Full HD (FHD) camera modules, optimal for generating contents for Milk VR) indicate the humble beginnings of the truly immersive experience that is to come in the 5G Era.

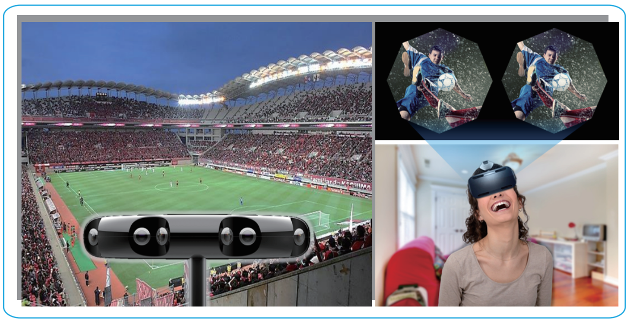

Figure 3 Watching Sport Events with VR

In an AR service scenario, computer-aided real-time information based on user context is graphically augmented to the display, delivering added value for the user. In the future the service desk attendant no longer needs to memorize the tiresome details of the products. Instead, AR services will help to inform the price, popularity, and details of a given product. Figure 4 illustrates another service scenario - AR navigation on windshield, where navigation information and other helpful notifications (refuel reminder and nearby shop location) are displayed on the windshield of a car, so that the user can continue to focus on driving while getting subtle context-aware notifications about potential services at the same time.

Samsung is actively involved in the development of 5G technologies to support these immersive VR and AR services, which will entertain users and provide a truly life-like experience on the move.

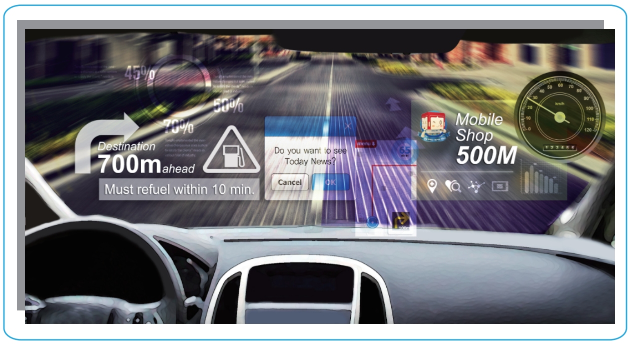

Figure 4 Driving a Car with AR Navigation

Everything on the Cloud

5G will provide users with a desktop-like experience based on cloud computing. With everything stored and processed on the cloud and immediately accessed with low latency, only simple input and output interfaces on mobile devices are needed, making them lighter, thinner, fancier, and more eco-friendly.

As an example, when you go shopping, the smart device can notify you about the arrival of new dresses that you might like, or let you know how well a dress in the newly incoming inventory matches with your liking based on your purchase history. This notification can be triggered, for instance, as you step into a shop, or take a picture of the dress, while the necessary computation to come up with the dress recommendation through crowd sourcing is done on a cloud server that is potentially half a world away.

Intuitive Remote Access

Users will be able to control remote machines and appliances as if they are right before them, even from thousands of miles away. Thanks to the reliable connections and near-zero latency of 5G, users will be able to control heavy industrial machines remotely, or access hazardous site remotely. It will also help mankind in exploring areas on the earth that are as yet unexplored, such as the Polar Regions or parts of the ocean floor.

5G Requirements

In order to realize such a demanding and unprecedented service vision, Samsung proposes the 5G rainbow requirements consisting of 7 Key Performance Indices (KPIs) as shown in Figure 5.

5G systems will be required to deliver an order of magnitude cell capacities and per-user data rate compared to its predecessors. Specially, 5G systems are expected to support data rates of 10-50 Gbps for low-mobility users. As a baseline, 5G sys-

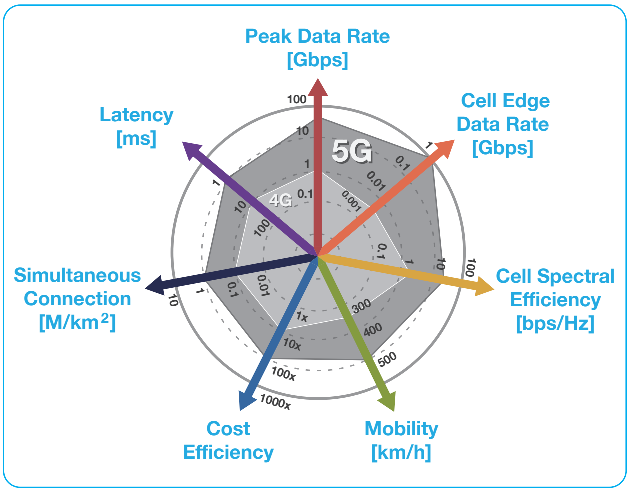

Figure 5 5G Rainbow Requirements

tems will provide gigabit-rate data services regardless of a user's location as shown in Figure 6 and Figure 7. To provide this uniform QoE, 5G network deployments are expected to be much denser compared to 4G networks, so making cost-effective deployment is a very important requisite.

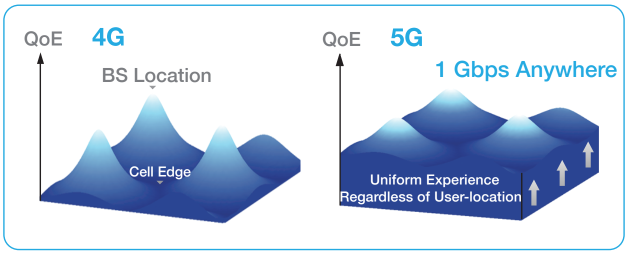

Figure 6 Edgeless RAN - 1 Gbps Anywhere

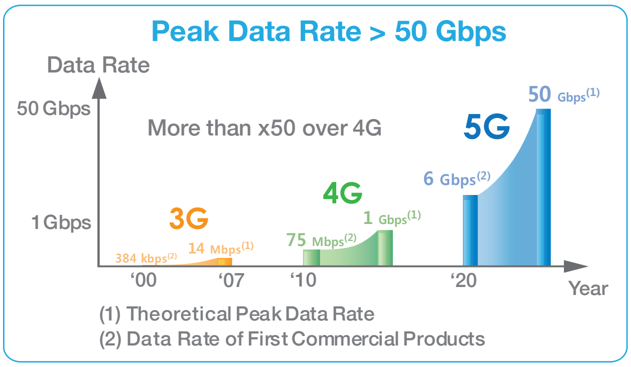

Figure 7 Data Rate Comparison of 5G with 3G and 4G

To fundamentally support the cloud storage/computing infrastructure of the future, 5G networks will deliver an end-to-end latency of less than 5 milliseconds and over-the-air latency of less than one millisecond (see Figure 8) - which is one-tenth compared to the latency of a 4G network. Critical infrastructure monitoring, for example, currently requires service levels achievable only on dedicated wireline networks while 5G technologies offer the promise of making these service levels achievable over wireless networks. Likewise, low-latency networks will allow pre-crash sensing, enabling vehicles to sense imminent collisions and exchange relevant data that could salvage the situation and/or mitigate adverse impact of the collision. Other challenging low-latency services that could be enabled by 5G could include self-driving cars, public safety communications systems, augmented reality, and "tactile internet" [3].

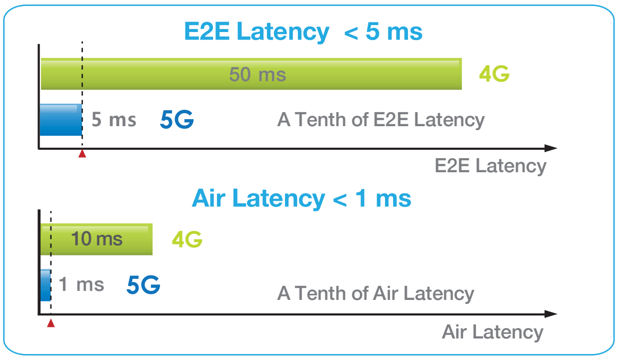

Figure 8 Ultra Low Latency of 5G

With spectral efficiency requirements set to 10 bps/Hz levels (in contrast to the 1-3 bps/Hz on 4G networks), 5G is also expected to deliver an efficient use of the spectrum by using MIMO, advanced coding and modulation schemes and new waveform design (more on this in the enabling technology section)

To address the ever-widening revenue gap that the operators and service providers are experiencing, 5G systems are targeted to be 50 times more efficient than 4G by delivering reduced cost and energy usage per bit. This sequentially requires low-cost network equipment, lower deployment costs, and enhanced power saving functionality on the network and user equipment sides.

5G technologies will be required to cope efficiently with all degrees of mobility by providing "mobility on demand" based on each device's and service's needs. On one hand, the mobility of user equipment should be guaranteed to be at least the same level as the current 4G system - that is the baseline. On the other hand, Samsung envisions that specialized 5G systems will support mobility at speeds ranging from 300 to 500 km/h.

To make the IoT Vision come true, the number of simultaneous connections in the 5G system is expected to be over 10^6 per unit square kilometer, which is much higher than that of the legacy system.

5G Key Enabling Technologies

Groundbreaking innovations will drive 5G technologies to meet the unprecedented speeds, near-wireline latencies, ubiquitous connectivity with uniform QoE, and the ability to connect massive amounts of devices with each other, all working in unison to provide the user with an immersive experience, even while the user is on the move. Future 5G systems will encompass fundamentally new designs to boost wireless capacity utilizing new frequency bands, advanced spectrum efficiency enhancement methods in the legacy band, and seamless integration of licensed and unlicensed bands.

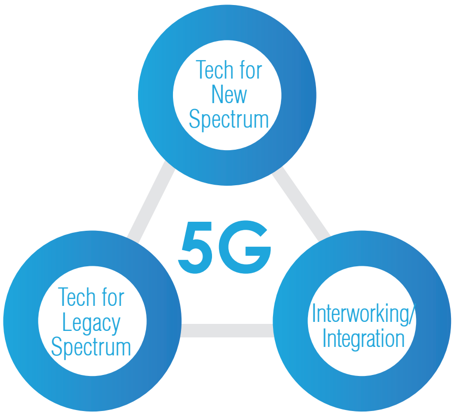
Tech for New Spectrum
5G
Tech for Legacy Spectrum
Interworking/ Integration

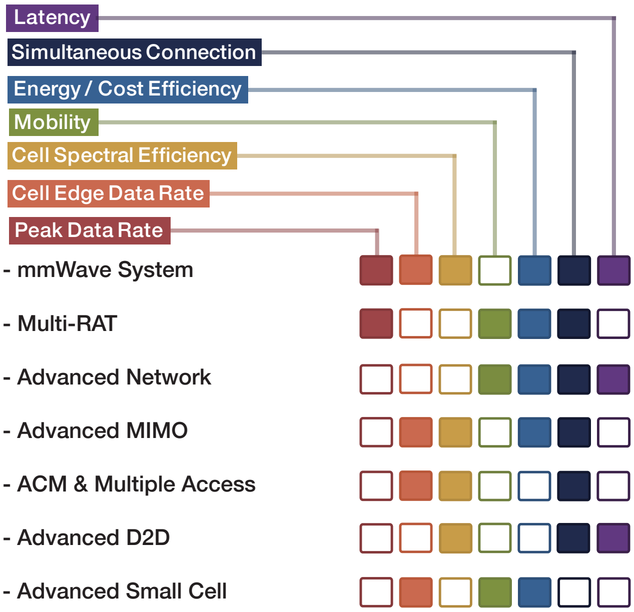

	Peak Data Rate	Cell Edge Data Rate	Cell Spectral Efficiency	Mobility	Energy / Cost Efficiency	Simultaneous Connection	Latency
- mmWave System	■	■	■	□	■	■	■
- Multi-RAT	■	□	□	■	■	■	□
- Advanced Network	□	□	□	■	■	■	■
- Advanced MIMO	□	■	■	□	■	■	□
- ACM & Multiple Access	□	■	■	□	□	■	□
- Advanced D2D	□	□	■	□	□	■	■
- Advanced Small Cell	□	■	□	■	■	□	□

Figure 9 shows an overview of the 5G key enabling technologies. The massively higher capacity needs of the 5G systems will be addressed by new mmWave systems, high-density small cells, advanced Multiple-Input and Multiple-output (MIMO) and new multiple access schemes like Filter-Bank Multi-Carrier (FBMC). Adaptive Coding and Modulation like Frequency and Quadrature Amplitude Modulation (FQAM) can significantly improve the cell edge performance and combined with higher density deployments with multi-BS cooperation will help to deliver on the promise of "Gbps anywhere" and Uniform QoE. Multi-Radio Access Technology (Multi-RAT) integration including carrier

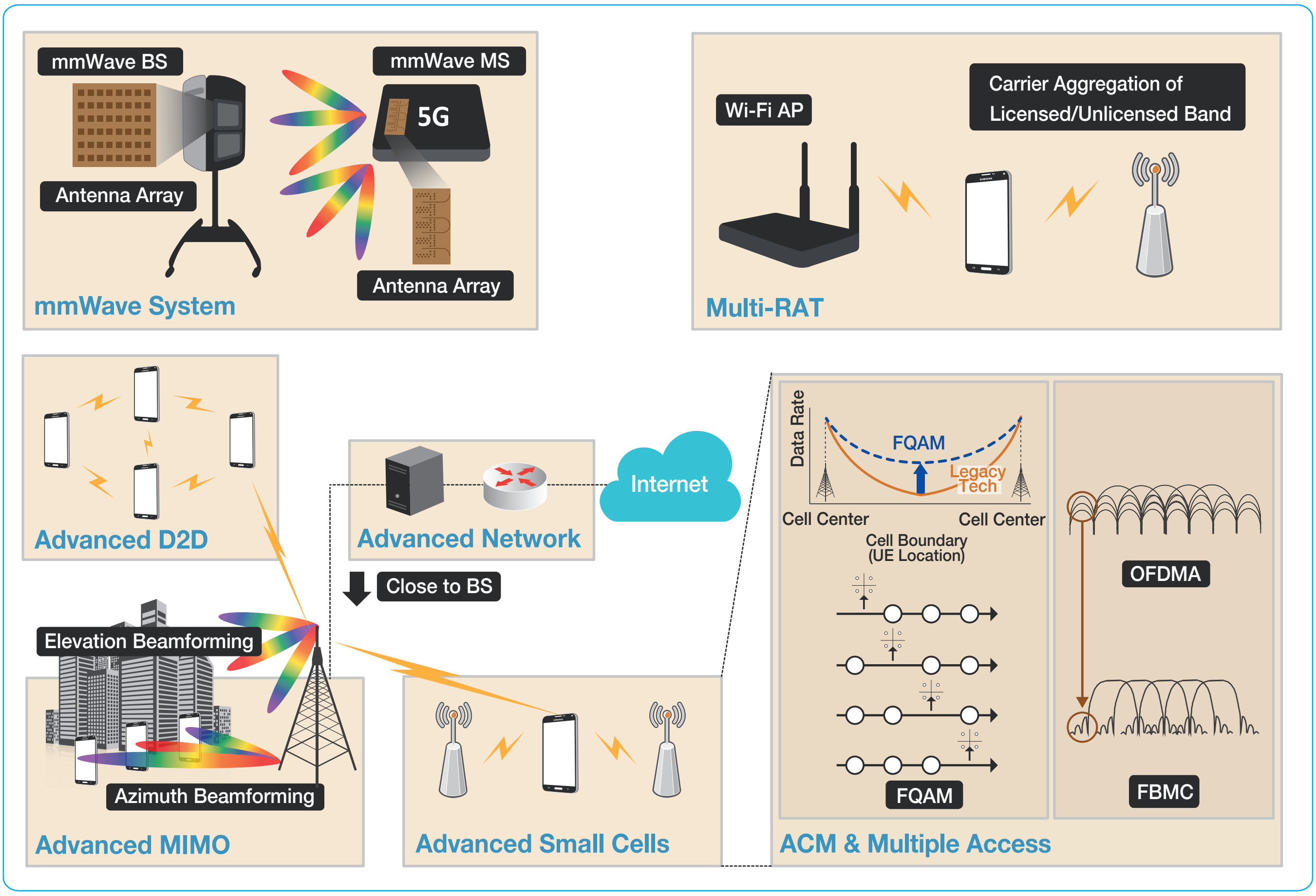
mmWave BS
mmWave MS
5G
Antenna Array
Antenna Array
mmWave System
Wi-Fi AP
Carrier Aggregation of Licensed/Unlicensed Band
Multi-RAT
Advanced D2D
Internet
Advanced Network
Close to BS
Data Rate
FQAM
Legacy Tech
Cell Center
Cell Center
Cell Boundary (UE Location)
OFDMA
FQAM
FBMC
ACM & Multiple Access
Elevation Beamforming
Azimuth Beamforming
Advanced MIMO
Advanced Small Cells

Figure 9 Overview of 5G Key Enabling Technologies

aggregation of licensed and unlicensed bands will inevitably help in increasing the available system bandwidth. On the network side, novel topologies including application servers placed closer to the network edge will contribute to significantly reducing the network latency. Advanced Device-to-Device (D2D) technology can help reduce the communications latency and support larger number of simultaneous connections in a network.

These 5G key enabling technologies are described in more detail in the following sections.

mmWave System

The mmWave bands provide 10 times more bandwidth than the 4G cellular-bands, as illustrated in Figure 10. Therefore, the mmWave bands can support the higher data rates required in future mobile broadband access networks.

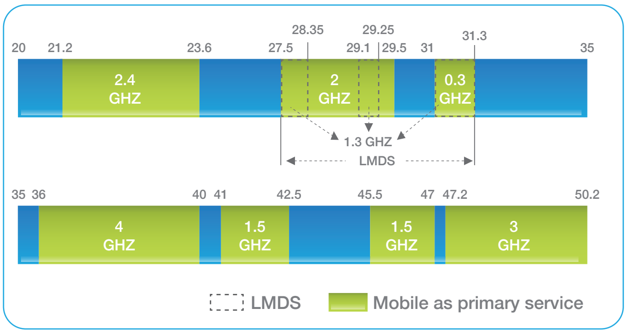
20
21.2
23.6
27.5
28.35
29.1
29.25
29.5
31
31.3
35
2.4 GHZ
2 GHZ
0.3 GHZ
1.3 GHZ
LMDS
35
36
40
41
42.5
45.5
47
47.2
50.2
4 GHZ
1.5 GHZ
1.5 GHZ
3 GHZ
LMDS
Mobile as primary service

Figure 10 Potential Bands in 20-50 GHz (US)

Unlike the below 6GHz bands, in the mmWave bands we have to carefully consider the wireless conditions such as rainfalls, snowfalls, and fogs. However, the amount of loss is mild for the expected range of a 5G communication link. For example, atmospheric absorption loss due to H_2O and O_2 at 28 GHz would be around 0.02 dB at a distance of 200 meters. Intensive rainfall (110 mm/h) results in about 4 dB loss at 200 meters distance. Loss due to heavy snow (10 mm/h) or heavy fog (visibility distance under 50 meters) will be less than 0.1 dB, at all distance of 200 meters [4][5].

In Figure 11, we show mmWave channel measurements from extensive experiments in Daejeon, Korea [6][7]. These channel measurements were carried out at 28 GHz with a channel bandwidth of 250 MHz, transmission power of 29 dBm and the horn antenna gain of 24.5 dBi for both the transmitter and the receiver [6][7].

The path loss exponents (based on 8 meters distance free space path loss) is 3.53 in Non Line of Sight (NLoS) links [6]. The mild path loss exponent from measurement results strongly suggests that mmWave communication links can be supported over 200 meters of distance, even in dense urban outdoor NLoS environments. Similar results are also reported in ray-tracing-based mmWave propagation model [8], confirming the fact that mmWave

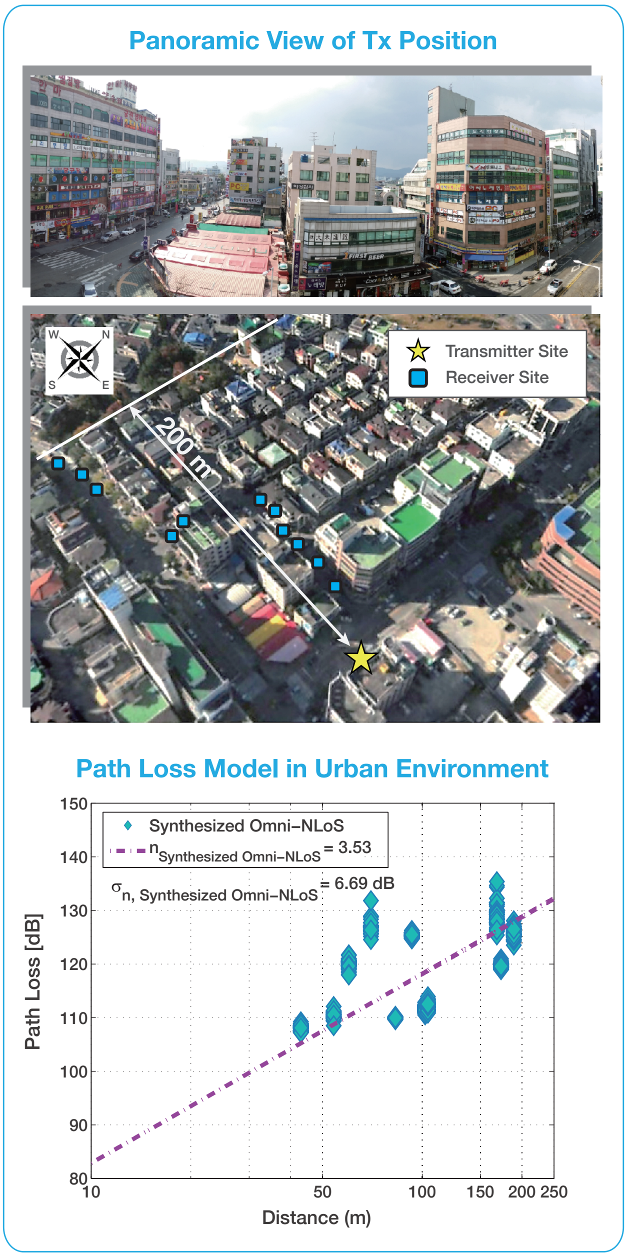

Figure 11 mmWave Channel Measurements in Urban Area, Daejeon (Korea)

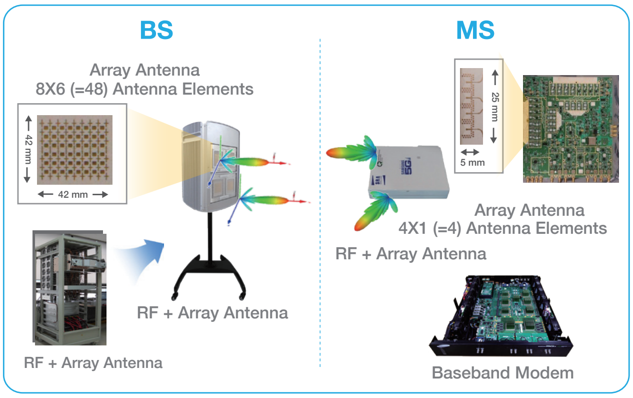
BS
MS
Array Antenna
8X6 (=48) Antenna Elements
42 mm
42 mm
RF + Array Antenna
RF + Array Antenna
25 mm
5 mm
Array Antenna
4X1 (=4) Antenna Elements
RF + Array Antenna
Baseband Modem

Figure 12 Adaptive Pencil Beamforming (Example)

bands can be successfully utilized for future IMT systems. Moreover, based on the ray-tracing, mmWave 3D-channel model in urban scenario is proposed in [9]. Higher path-loss and the resulting fragile link is largely due to weak diffractions in the mmWave bands. We need to overcome these challenges in order to make outdoor mmWave communication a reality [10][11]. Fortunately, the small mmWave wavelength also means highly directional beamforms can be obtained using a large number of antenna elements in a smaller form factor. These adaptive directional beams with large antenna array gain are key in combating the large propagation loss in the mmWave [12][13][14][15], as illustrated in Figure 12.

We have developed a mmWave beamforming prototype at the DMC R&D Center, Samsung Electronics, Korea, in order to demonstrate the feasibility of using mmWave bands for cellular services. We showed that mmWave system can meet the two key requirements of cellular services: sufficiently large geographical coverage and support for mobility in NLoS environments.

With the prototype, we tested outdoor coverage to demonstrate the service availability in a typical outdoor environment for both LoS and NLoS sites. The tests were performed at sites surrounded by tall buildings where various channel propagation effects such as reflection, diffraction, and penetration were present, as shown in Figure 13. We observed that reliable communication links are formed even for NLoS sites that are more than 200 meters away from the base station.

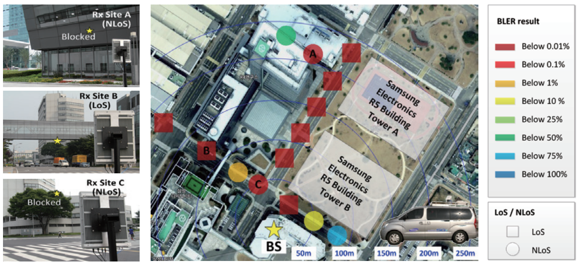
Rx Site A (NLoS)
Blocked
Rx Site B (LoS)
Rx Site C (NLoS)
Blocked
A
B
C
BS
Samsung Electronics R5 Building Tower A
Samsung Electronics R5 Building Tower B
50m
100m
150m
200m
250m
BLER result
Below 0.01%
Below 0.1%
Below 1%
Below 10 %
Below 25%
Below 50%
Below 75%
Below 100%
LoS / NLoS
LoS
NLoS

Figure 13 Exemplary Outdoor Coverage Test Results

Advances in semiconductor technology have made commercial mmWave systems readily available. Recent technologies of Silicon-based Complementary Metal Oxide Semiconductor (CMOS) processes are capable of integrating mixers, Low Noise Amplifiers (LNAs), Power Amplifiers (PAs), and Inter-Frequency (IF) amplifiers, all in a single package. A good example is the 60 GHz commercialized products with the label of Wireless Gigabit Alliance (WiGig) [16], and it is well recognized that cost effective implementations of sub-100nm CMOS process made it possible to utilize the 60 GHz bands [17].

GaAs Monolithic Microwave Integrated Circuit (MMIC) technologies are maturing fast, and they are becoming a dominant choice for components in the RF chain including PAs, LNAs, switches for digital attenuators and phase shifters, Voltage Controlled Oscillators (VCOs), and passive components from a few GHz to 100 GHz. As illustrated in Figure 14, it is projected that the Power Added Efficiency (PAE) of a 5G Front End Modules (FEM) will soon match the PAE of today's commercial 4G system. Meanwhile, a good Effective Isotropic Radiated Power (EIRP) can be achieved with the help of high antenna gains from a large number of antenna elements [18][19].

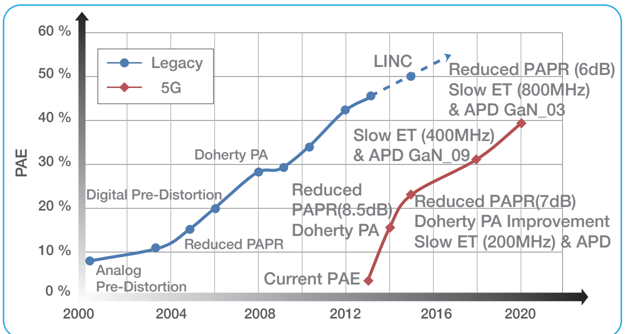

Figure 14 Projections of 5G FEM PAE

High performance antenna solution is an important piece of the 5G puzzle. The constantly varying LoS and NLoS propagation environment demands a novel antenna which exhibits high gain as well as wide spatial coverage capacity. The antenna array comprising of multiple antenna elements must fit within the small form factor of a 5G mobile handset.

Samsung has been developing innovative 5G phased array antennas that have near zero-footprint and reconfigurable antenna modes, as shown in Figure 15.

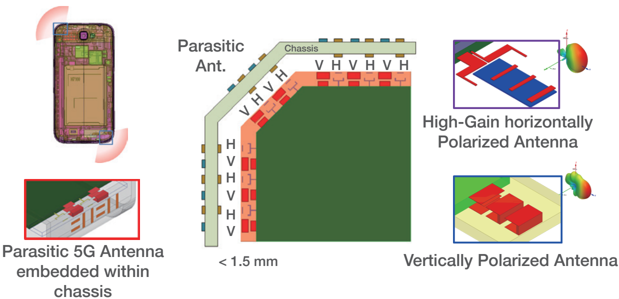
Parasitic Ant.
Chassis
V H V H V H
V H V H
H V H V H V
< 1.5 mm
Parasitic 5G Antenna embedded within chassis
High-Gain horizontally Polarized Antenna
Vertically Polarized Antenna

Figure 15 The Reconfigurable 5G Phased-array Antenna

Characterizing the biological implications on a user's body imposed by cellular handset devices is one of the most important aspects that need to be verified prior to its commercial release. The Signal Absorption Rate (SAR) regulated by worldwide governmental bodies is used as a guideline to assess the effect of mmWave bands on a user's body. The SAR of an envisioned 5G (28 GHz) cellular handset is analyzed and illustrated in Figure 16.

When optimally configured, the 5G beamforming technology enables the peak radiation of the Mobile Station (MS) antenna to steer away from the user's head. As a result, the average SAR can be reduced by more than a factor of 10 compared to that of the present day 4G cellular handsets.

The maximum SAR is expected to be further minimized in the future, as 5G antenna technology continues to evolve.

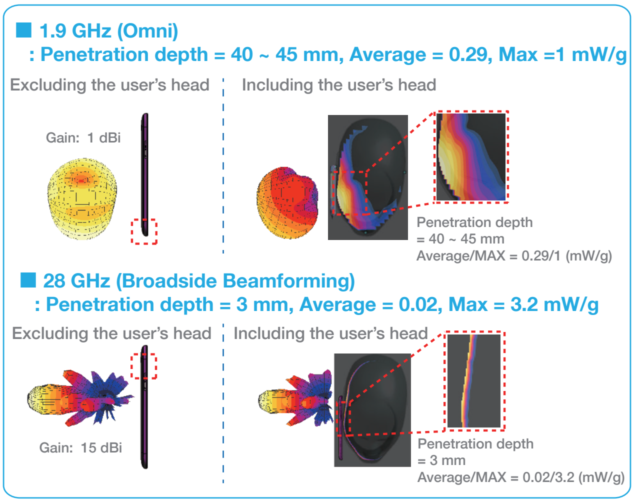

Figure 16 Biological Implications on the User's Body

Multi-RAT

To realize the envisioned 5G services, significant enhancement of per-user throughput and overall system capacity is required, compared with those of the 4G system. Such an enhancement can be achieved through advanced PHY/MAC/network technologies and efficient methods of cell deployment and spectrum management. In particular, utilizing a larger amount of system bandwidth guarantees an increase in the capacity by allocating more frequency resources to each user in the system. Therefore, utilizing the spectrum where huge bandwidth is available can be considered to be the most critical issue for the 5G system.

Currently, the 4G system specifies its operating frequency bands and some of them are assigned to cellular operators. These deployments can be beneficial since the existing 4G system will not interfere with other RATs.

However, obtaining the licensed spectrum requires not only considerable financial investment, but also a significantly long period of time spent on regulatory procedures. More importantly, a substantial portion of the licensed spectrum around 2 GHz is already being used. Therefore, finding a bandwidth that is wide enough to support the 5G system is very challenging.

The recent trend in spectrum management is to aggregate both the licensed and unlicensed spectrums to extend available system bandwidth, as shown in Figure 17.

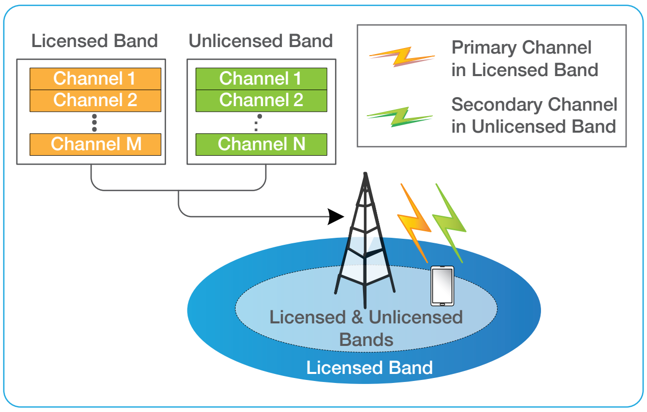
Licensed Band
Channel 1
Channel 2
Channel M
Unlicensed Band
Channel 1
Channel 2
Channel N
Primary Channel in Licensed Band
Secondary Channel in Unlicensed Band
Licensed & Unlicensed Bands
Licensed Band

Figure 17 Integration of Licensed and Unlicensed Bands

The unlicensed spectrum has plenty of bandwidth. For instance, approximately 500 MHz and 7 GHz bandwidths are available in the 5 GHz and 60 GHz bands, respectively. In order for the 5G system to utilize the unlicensed spectrum, regulations imposed on each frequency band should be carefully reconsidered. Representative examples of the regulations are the Transmit Power Control (TPC), Dynamic Frequency Selection (DFS), and Listen Before Talk (LBT).

To efficiently utilize the unlicensed spectrum, we will develop the 5G system with the following characteristics.

First, we will design PHY/MAC/network algorithms suitable for the nature of the unlicensed spectrum. Since a wide range of frequency bands are included in the unlicensed spectrum, each frequency band has its own characteristics. Hence, band-specific solutions will be provided.

Second, efficient coexistence mechanisms that take into account other RATs (e.g., WiFi or WiGig) operating in the unlicensed spectrum will be suggested.

Finally, techniques for interworking and integrating the 5G system with other RATs will be developed. By taking advantage of multiple RATs, the 5G system will be able to take advantage of the unique characteristics of each RAT and improve the practicality of the system as a whole. For instance, the 4G system is used for exchanging the control messages to maintain the connection, to perform handover, and to provide real-time services such as VoLTE. The technology operating in mmWave unlicensed frequency band would support the gigabit data rate service. Multiple mmWave cells can be overlaid on top of the underlying 4G macro cells, as shown in Figure 18.

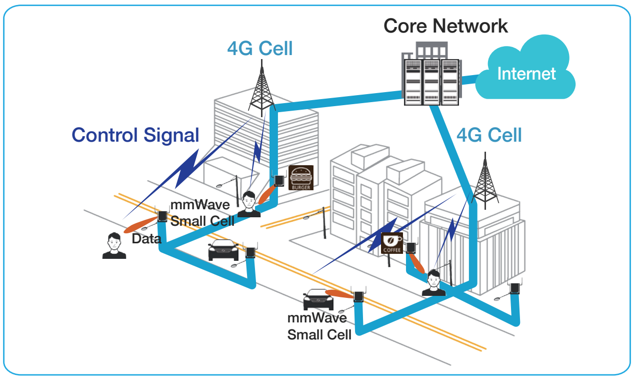

Figure 18 Overlaid Network of mmWave Small Cell Integrated with the Underlay 4G System

Advanced Network

In order to meet the 5G key requirements such as latency and the large number of simultaneous connections, and to support new business models and scenarios for operators, technologies above the radio access level should also be considered. Therefore, we need to develop technologies at the system architecture level from the network point of view. To accomplish the 5G key requirements and support the increased data rate of new 5G radio access, 5G network will have to evolve towards a distributed and flat architecture.

Current mobile network architectures designate a dedicated node in the core network (e.g., PGW - Packet Data Network Gateway in 3GPP) as a mobility anchor that allocates an IP address to the terminal, tracks terminal location in the IP topology, and ensures terminal's reachability by tunneling its traffic to wherever it goes. All terminal traffic is tunneled through the centralized node in the mobile core network. However, the undesirable consequences of this design include the following:

- Increase in end-to-end transmission latency due to elongated data paths.
- Additional load of backhauling and network processing in the core networks.
- Low network reliability due to introduction of a single point of failures.

In the 5G flat network architecture, as illustrated in Figure 19, user mobility is managed efficiently and in a dynamically scalable fashion by pushing the functionality to the edges of the network and even onto the mobile terminals [20].

The three important benefits of this approach include the followings:

First, such a distributed mobility management always provides the shortest data path between MS and the Internet without traversing the core network. This distributed mobility management leads to a significant reduction of signaling and data transmission delay. Also, low end-to-end latency requirements of 'less than 5 ms' for new 5G services such as immersive UHD video streaming, cloud gaming, and virtual reality, cannot be archived solely by reducing radio access latency but would also require a flatter network architecture design. In the flat network architecture, services which require low latency transmission can be provided by Edge Servers and they can benefit from advanced network features which utilizes network information for optimal operations.

Second, it provides a highly scalable solution compared to the centralized architecture, in which a single core network gateway maintains the whole traffic from MSs or to MSs.

Third, it easily avoids the risk of having a single point of failure. In flat network architecture, the breakdown of one network gateway would not significantly interfere with the operations of the other gateways.

Flexibility is considered as another key requirement of 5G network architecture. Software-Defined Networking (SDN) and Network Function Virtualization (NFV) provide promising examples of programmable design technologies for realizing a flexible 5G network architecture.

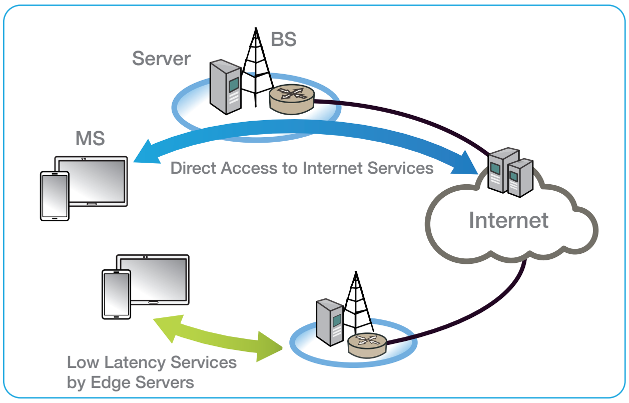

Figure 19 5G Flat Network Architecture

Advanced MIMO

One promising technology for meeting the future demands is massive MIMO transmission/reception [21]. When used with multi-user precoding schemes such as Maximum Ratio Transmission (MRT) precoding, also known as channel conjugate precoding, massive MIMO systems experience small inter-user and inter-cell interferences, and consequently achieve significantly higher throughput than the state-of-the-art MIMO systems.

One of the main challenges to build massive MIMO systems in practice is the limitation in the number of antennas that can be equipped at a BS, caused by the BS form factors and operating carrier frequencies.

For example, to horizontally install a large number of antenna elements (e.g., > 8) at the top of a BS tower operating with the lowest 4G system frequency bands of 700 MHz, eight antenna elements with 0.5λ spacing require up to 1.7 m width, where λ is the carrier wavelength. For the typical 4G system frequency bands of 2.5 GHz, fitting 32 antenna elements with 0.5λ spacing require up to 1.9 m width, which is still not feasible in many BSs that have only limited room on the tower. This practical limitation has motivated Full-Dimension MIMO (FD-MIMO) cellular communication systems, which place a large number of active antenna elements in a two dimensional grid at the BSs.

A typical FD-MIMO deployment scenario is illustrated in Figure 20, for a macro BS with 3 sectors equipped with 2D Active Antenna Array (AAA) panels.

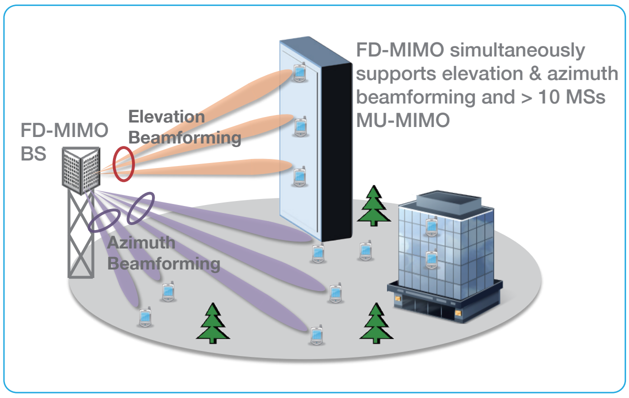
FD-MIMO simultaneously supports elevation & azimuth beamforming and > 10 MSs MU-MIMO
FD-MIMO BS
Elevation Beamforming
Azimuth Beamforming

Figure 20 Example of FD-MIMO Deployment

FD-MIMO system can support high-order Multi-User MIMO (MU-MIMO), while fully exploiting the elevation as well as the azimuth dimension, thereby generating improved system throughput.

In full buffer system-level evaluations, it has been found that 64-antenna-port FD-MIMO system achieves 243% average-cell and 244% cell-edge performance gain, compared to those of the 8-antenna-port legacy MIMO system. In order to achieve the promising gain of FD-MIMO in practice, we need accurate beam steering and tracking in three Dimensions (3D).

To steer and track the MU beams toward serving MSs, FD-MIMO BS should be equipped with multiple transceivers (TRX) feeding 2D array elements, in which the number of TRX doubles or even quadruples compared to that of the conventional BS.

Having a large number of TRX poses new challenges, such as antenna calibration and complexity issues associated with Channel State Information (CSI) acquisition and precoding.

On the other hand, high-order MU-MIMO introduces another set of new challenges, such as scheduling complexity and link adaptation. Furthermore, in Frequency Division Duplex (FDD) systems, other new challenges emerge such as pilot overhead, CSI estimation complexity, CSI quantization and feedback overheads.

ACM & Multiple Access

– FBMC

As cellular IoT has been one of key driving forces to 5G, spectrally efficient support for heterogeneous services that have quite different requirements is becoming ever so important. Accordingly, several enabling methods such as multi-RAT coexistence and flexible spectrum sharing have been actively investigated.

Recently, FBMC has drawn much attention as an enabling technology for enhancing the fundamental spectral efficiency, though its theory has a long history similar to that of Orthogonal Frequency Division Multiplexing (OFDM).

Because of the well-localized time/frequency traits adopted from a pulse shaping filter per subcarrier, the FBMC system can reduce the overhead of guard band required to fit in the given spectrum bandwidth, while meeting the spectrum mask requirement.

Furthermore, the effectively increased symbol duration is suitable for handling the multi-path fading channels even without Cyclic Prefix (CP) overhead. Consequently, the FBMC system can reduce the inherent overheads such as CP and guard-bands in CP-OFDM. FBMC is also attractive in specific asynchronous scenarios, including Coordinated Multi-Point Transmission and Reception (CoMP) and Dynamic Spectrum Access (DSA) in a fragmented spectrum.

However, to maintain the transmission symbol rate, the conventional FBMC system generally doubles the lattice density either in time or in frequency compared with OFDM while adopting Offset Quadrature Amplitude Modulation (OQAM). In OQAM, in-phase and quadrature-phase modulation symbols are mapped separately with half symbol duration offset. Thus, so-called OQAM-FBMC or Staggered Multi-Tone (SMT) causes intrinsic interference that makes it difficult to apply conventional pilot designs and corresponding channel estimation algorithms as well as MIMO schemes as in CP-OFDM systems [22].

With a base-filter that takes the spectrum confinement and the orthogonality among adjacent subcarriers into consideration, the QAM-FBMC system performs comparable to the CP-OFDM system even without the CP overhead, while the guardband overhead reduction is also available from the well-confined spectrum. Sophisticated receiver algorithms including channel estimation and equalization can further mitigate the multi-path fading channel impact without the CP.

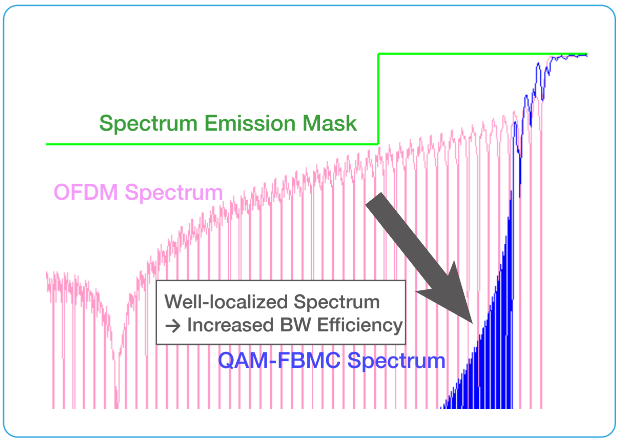

Figure 21 One QAM-FBMC Symbol Generation

– FQAM

One of key requirements for 5G is enhancement of cell-edge performance, which means that every user should be supported with Gigabit experience anywhere. Conventional approaches to enhance the cell-edge performance mainly focus on managing interference (e.g., interference cancellation, interference avoidance), by dealing with interference as a Gaussian. However, it is proved that the worst-case additive noise in wireless networks with respect to the channel capacity has a Gaussian distribution. From this observation, one can expect that the channel capacity can be increased by a non-Gaussian interference design which makes Inter-Cell Interference (ICI) non-Gaussian. The distribution of ICI depends on the modulation schemes of the interfering BSs. Therefore, an active interference design for improved cell-edge performance can be achieved by applying a new type of modulation.

FQAM, a combination of Frequency Shift Keying (FSK), and Quadrature Amplitude Modulation (QAM) can be used as an active interference design scheme. Figure 22 shows the signal constellation of 16-ary FQAM that is a combination of 4-ary FSK and 4-ary QAM.

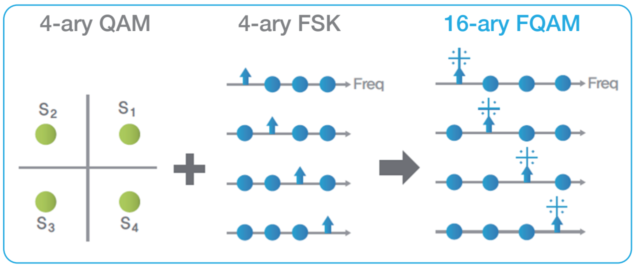

Figure 22 Example of 16-ary FQAM

With FQAM, the statistical distribution of ICI is likely to be non-Gaussian, especially for cell-edge users. As a result, the transmission rates for the cell-edge users can be significantly improved.

The statistics of ICI and the performance enhancement possibility have been proven by practical implementation of a system which uses FQAM. FQAM system environment for cellular downlink OFDMA networks is shown in Figure 23 and 24. Our experimental results show that the transmission rates for interference-limited users in FQAM-based OFDMA networks are around 300% higher than those in QAM-based OFDMA networks [23][24].

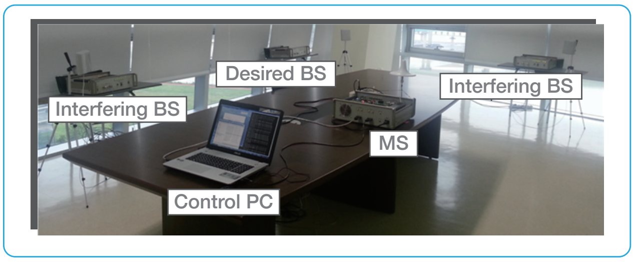

Figure 23 3 Cell Structure with MS and BSs

Figure 24 Implemented MS and BS

Advanced D2D

Advanced D2D communication is an attractive technology that enhances spectral efficiency and reduces end-to-end latency for 5G. Not entirely depending on the cellular network, D2D devices can communicate directly with one another when they are in close proximity. Hence, D2D communication will be used for offloading data from network so that the cost of processing those data and related signaling is minimized. Moreover, enhanced version of D2D communication was suggested lately for being used as special purpose such as Mission Critical Push-To-Talk (MCPTT) communication and Vehicle-to-Anything (V2X) communication.

In advanced D2D communication, a single radio resource can be reused among multiple groups which want to communicate with each other if the interference incurred between groups is tolerable. Hence, we can increase the spectral efficiency and the number of the simultaneous connection by utilizing D2D communication in 5G. Moreover, since the data is directly transmitted and not going through the core network, the end-to-end latency can be considerably reduced. Therefore, advanced D2D communication is quite well aligned with IoT services as shown in Figure 25. The cars can communicate with each other to exchange the information for safety alarm and infotainment without cellular base station. The home appliances communicate with each other for home automation service. Many objects in proximity region can be connected to each other so that IoT services can be accomplished.

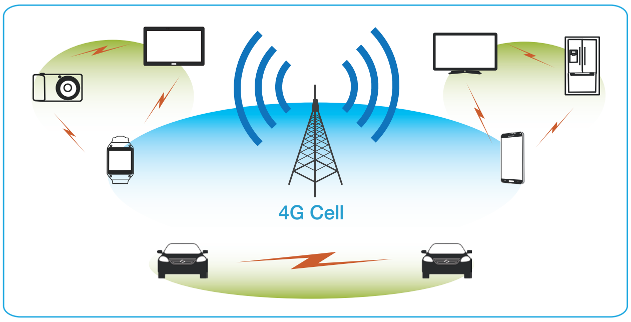

Figure 25 Advanced D2D Communications

Advanced Small Cell

To achieve significant throughput enhancement in a practical manner, it is necessary to deploy a large number of cells in a given area and to manage them intelligently. The 5G system is expected to utilize higher frequencies to take advantage of the vast bandwidth in the mmWave bands. Hence, the considerably high propagation loss of mmWave makes it suitable for dense small cell deployment, which leads to higher spatial reuse.

Moreover, Figure 26 shows the concept of a user-centric virtual cell. Conventional static network topologies with a central controller have an “edge”, the reach of the central controller. However, a user-centric virtual cell that consists of a group of cooperating BSs is continuously reformed so that any user will always find himself/herself at the “center” of the cell.

Distributed and self-configuring network technologies will make it easy to deploy many small BSs in urban and suburban areas. In-band wireless backhaul can be used between BSs for cooperative communication, reducing the cost and complexity of backhaul network deployment.

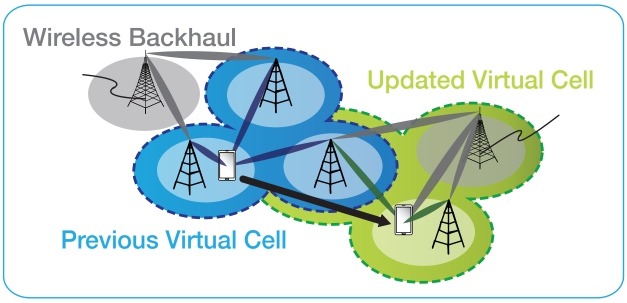

Figure 26 User Centric Virtual Cellular Network

5G Deployment Scenario

In order to provide the ubiquitous high QoE gigabit accessibility, we envision an overlaid deployment of 5G in conjunction with the existing 4G macro cells. The 5G small cells can be coupled with the overlaid 4G macro cells, and the 4G macro cells will control the operations of the associated 5G small cells, as illustrated in Figure 27. In the figure, 5G BS primarily provides multi-gigabit per second throughput with high QoE to mobile users over the legacy spectrum or higher spectrum like the 5 GHz unlicensed band and bands above 6 GHz. Meanwhile, the 4G BS can serve as a control channel to 5G MSs for supporting seamless connection anywhere over the legacy 4G spectrum.

5G system will need enhanced RAN technologies that not only utilize the legacy frequency bands assigned to IMT and IMT-Advanced systems, but also use new frequency bands. At the same time, 5G system will also need to support seamless interworking with other RATs operating in unlicensed bands.

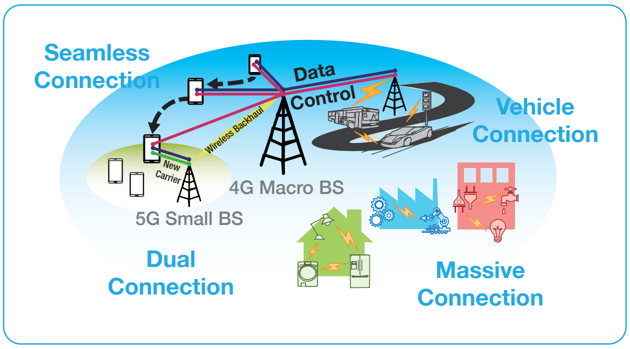

Figure 27 5G Deployment Scenario

Standardizations and Regulations

As shown in Figure 28, 5G is planned to be commercialized in year 2020. ITU-R WP 5D is preparing timelines in the vision document for standardization, spectrum allocation, and commercialization, in order to be on time. From the perspective of commercialization, standards for 5G should be ready by the year 2017 to allow 2 or 3 years for the development of 5G products. Considering average periods of previous standardization, 5G standards need to get started in 2015 to make 5G standards available by the end of 2017.

New frequency bands may be required to achieve the target performance of 5G. World Radio communication Conference (WRC)-15 agenda item 1.1 made in WRC-12 indicates consideration for additional spectrum allocations to mobile services on a primary basis and for identification of additional frequency bands for IMT purposes.

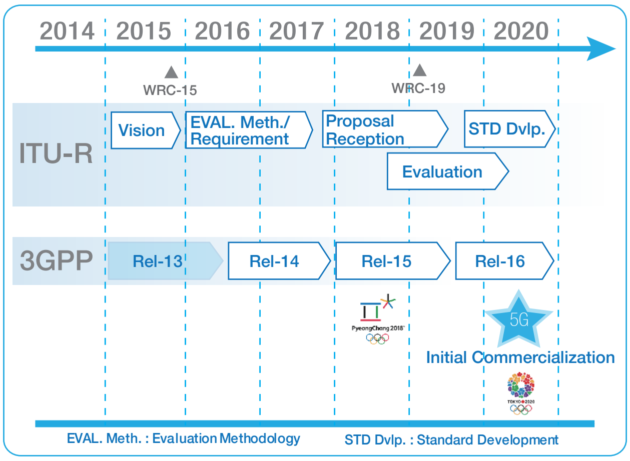
2014
2015
2016
2017
2018
2019
2020
WRC-15
WRC-19
ITU-R
Vision
EVAL. Meth./ Requirement
Proposal Reception
STD Dvlp.
Evaluation
3GPP
Rel-13
Rel-14
Rel-15
Rel-16
5G
Initial Commercialization
EVAL. Meth. : Evaluation Methodology
STD Dvlp. : Standard Development

Figure 28 5G Timelines

In light of this agenda item, various frequency ranges had been proposed and discussed in ITU-R WP 5D in July 2013, and Joint Task Group (JTG) 4-5-6-7 had performed compatibility studies among various technologies in each candidate frequency band until July 2014.

Although bands above 6 GHz were not included in JTG 4-5-6-7's candidate frequency bands, WP 5D has been developing a new report on IMT systems in bands above 6 GHz, including channel characteristics and the potential usage of the bands for IMT systems.

WP 5D may also provide JTG 4-5-6-7 with the necessary information regarding bands above 6 GHz in order to be referenced as appropriate in the draft text of Conference Preparatory Meeting (CPM). The possibility and further necessity to study the bands are captured in the report from JTG 4-5-6-7 to CPM as follows:

"The demand of high bit rate, especially in densely populated area could be accommodated in higher frequency bands (e.g. above 6 GHz) than those currently being considered in studies, however the technical information required for compatibility studies has yet to be developed and these studies and proposals are being explored for future work, beyond WRC-15."

This report implies that new frequency bands can be added to the IMT frequency bands toward WRC-19, and it is likely that bands above 6 GHz can be one of candidate frequency bands.

In addition, Licensed Assisted Access (LAA) study item has been accepted in 3GPP LTE in the second half of 2014, acknowledging the use of unlicensed bands for cellular system.

Recently, as FCC relieves 5GHz band regulations for various unlicensed band usage, related coexistence technology between LTE and WLAN will be standardized in both standards.

Further, as shown in Figure 29, Samsung is actively engaging in the most of global 5G research initiatives, such as European 5G PPP projects of Horizon 2020, 5G Innovation Centre (5GIC) in UK, NYU Wireless Center in US, Giga KOREA project and Chinese 836 project.

Samsung is leading various collaborations with industries and academics over the world. In particular, Samsung has played an important role as the full member of 5G PPP Infrastructure Association, the executive board member of 5G Forum in Korea and the chair of vision sub-working group for Future IMT (5G) in ITU-R WP5D.

In order to have a consistent perspective on 5G with those of other academic institutes, we are vigorously developing 5G core technologies with several outstanding universities around the world.

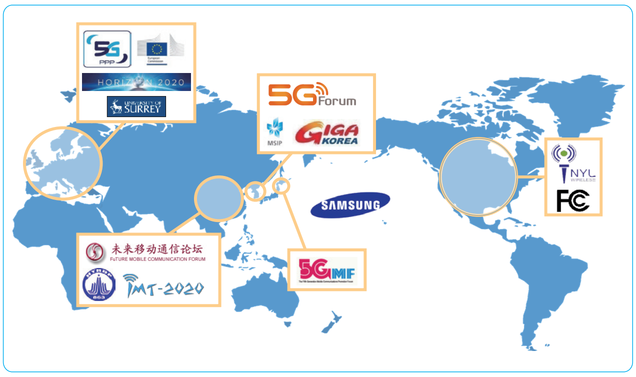
5G PPP
HORIZON 2020
UNIVERSITY OF SURREY
5G Forum
MSIP
GIGA KOREA
SAMSUNG
NYU WIRELESS
FCC
未来移动通信论坛
FUTURE MOBILE COMMUNICATION FORUM
IMT-2020
5GMF

Figure 29 Global 5G R&D Activities

Conclusion

5G will usher in a revolutionary generation of mobile communication that provides ubiquitous multi-Gbps data rate regardless of the user's location.

Significantly increased system capacity and real-time responsiveness of the 5G system will introduce life-changing services providing the users with a truly immersive and rich experience. The profoundly life-alternating world of billions of interconnected devices, the IoT, will require something as revolutionary as 5G to be fully realized in its fullest extent. And finally 5G promises to reverse the widening revenue gap and make it worthwhile for operators and service providers to invest again in innovative new services, and continue to propel increased productivity and efficiency.

Reference

1 UMTS, "Mobile Traffic Forecasts 2010-2020 Report," UMTS Forum, Jan. 2011.

2 Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update: 2013-2018," Cisco, Feb. 2014.

3 Calum Dewar et al., "Understanding 5G: Perspectives on future technological advancements in mobile," GSMA Intelligence Report, Dec. 2014.

4 ITU-R P.840-5, "Attenuation Due to Clouds and Fog," ITU-R, Feb. 2012.

5 DARPA R-1335-ARPA, "Atmospheric Effects on Terrestrial Millimeter-Wave Communications," DARPA Report, Mar. 1974.

6 Sooyoung Hur et al., "Wideband Spatial Channel Model in an Urban Cellular Environments At 28 GHz," in Proc. EuCAP'15, April. 2015.

7 Sooyoung Hur et al., "28 GHz Channel Modeling Using 3D Ray-Tracing in Urban Environments," in Proc. EuCAP'15, April. 2015.

8 Youngbin Chang et al., "A Novel Dual-Slope mm-Wave Channel Model Based on 3D Ray-Tracing in Urban Environments," in Proc. IEEE PIMRC'14, Sep. 2014.

9 Sooyoung Hur et al., "Synchronous Channel Sounder Using Horn Antenna and Indoor Measurements on 28 GHz," in Proc. IEEE BlackSeaCom'14, pp. 83-87, May 2014.

10 Zhouye Pi et al., "An Introduction to Millimeter-Wave Mobile Broadband Systems," IEEE Communications Magazine, vol.49, no.6, pp.101-107, Jun. 2011.

11 Cheol Jeong et al., "Random Access in Millimeter-Wave Beamforming Cellular Networks: Issues and Approaches," IEEE Communications Magazine, vol. 53, no. 1, pp.180-185, Jan. 2015.

12 Wonil Roh et al., "Millimeter-wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," IEEE Communications Magazine, vol. 52, no. 2, pp. 106-113, Feb. 2014.

13 Samsung, "Technologies for Rel-12 and Onwards," RWS-120021, 3GPP TSG RAN Workshop on Rel-12 and Onwards, Jun. 2012.

14 Taeyoung Kim et al., "Tens of Gbps Support with mmWave Beamforming Systems for Next Generation Communications," in Proc. IEEE GLOBECOM'13 Workshop, pp.3685-3690, Dec. 2013.

15 Chanhong Kim et al., "Multi-Beam Transmission Diversity with Hybrid Beamforming for MIMO-OFDM Systems," in Proc. IEEE GLOBECOM'13 Workshop, pp. 61-65, Dec. 2013.

16 WiGig Alliance, "Defining the Future of Multi- Gigabit Wireless Communications," WiGig White Paper, Jul. 2010.

17 R12-WP5D-C-0258, "Technical Feasibility of IMT in the Bands above 6 GHz," ITU, Jan. 2013.

18 Theodore S. Rappaport et al., "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," IEEE Access, vol.1, pp.335-349, May 2013.

19 Theodore S. Rappaport et al., "28GHz Propagation Measurements for Outdoor Cellular Communications Using Steerable Beam Antennas in New York City," in Proc. IEEE ICC'13, pp.5143-5147, Jun. 2013.

20 Alper Yegin et al., "Terminal-centric Distribution and Orchestration of IP Mobility for 5G Networks," IEEE Communications Magazine, vol. 52, no. 11, pp.86-92, Nov. 2014.

21 Young-Han Nam et al., "Full Dimension MIMO (FD-MIMO) for Next Generation Cellular Technology," IEEE Communications Magazine, vol. 51, no. 6, pp.172-179, Jun. 2013.

22 Maurice Bellanger et al., "FBMC Physical Layer: a Primer," PHYDYAS Tech. Rep., Jun. 2010.

23 Sungnam Hong et al., "A Modulation Technique for Active Interference Design under Downlink Cellular OFDMA Networks," in Proc. IEEE WCNC'14, pp.683-688, Apr. 2014.

24 Sungnam Hong et al., "Frequency and Quadrature-Amplitude Modulation for Downlink Cellular OFDMA Networks," IEEE Journal on Selected Areas in Communication, vol. 32, no. 6, pp.1256-1267, Jun. 2014.

Samsung Demonstrates World's First 5G Data Transmission at Highway Speeds

Record-breaking 1.2Gbps data transmission at over 100km/h, and 7.5Gbps in stationary conditions using 28GHz spectrum

SEOUL, Korea – October 15, 2014 – Samsung Electronics Co., Ltd today announced the record-breaking demonstration of super-fast data transmission targeted for fifth generation (5G) mobile networks using 28 GHz spectrum. Samsung researchers confirmed the world's first data rate of 1.2 Gbps, or 150 MB per second on a vehicle cruising at over 100 km/h. This marks a significant step towards the utilization of millimeter wave frequency bands for 5G mobile networks

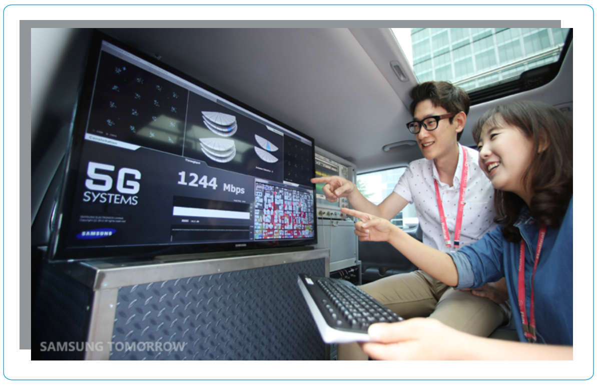

Figure A Achieved Data Rate in Mobile Condition (1.2Gbps)

The achievement was bolstered by another record-breaking demonstration in which Samsung achieved data transmission speed up to 7.5 Gbps, or 940 MB per second when the vehicle came to a complete stop. The peak data rate is more than thirty times faster compared to the state-of-the-art 4G LTE technology.

Back in May, 2013, Samsung revealed the world's first 28 GHz based 5G data transmission speed of 1Gbps at pedestrian speeds. Since then researches at Samsung were able to increase the maximum data rate by more than 7-fold and support mobility up to highway speeds. The continued ground breaking success underlines the company's leadership in next-generation mobile communications.

"The expectations for 5G communications will continue to grow based on the rising demands for smart devices, cloud services, smart home technology, and Internet of Things" , said ChangYeong Kim, Head of DMC R&D Center of Samsung Electronics. "We are committed to developing innovative 5G technologies and will continue to utilize our exceptional R&D capabilities as well as our diverse partnerships with leading companies and research centers around the world".

While 5G standard has yet to be ratified, 5G networks are expected to feature data rates and capacity that are hundreds times larger compared to 4G LTE.

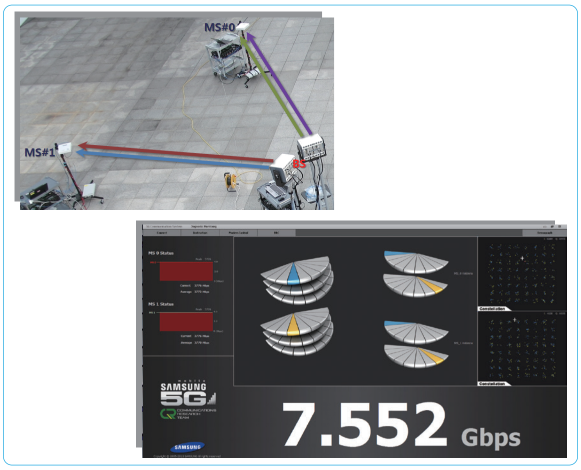

Figure B Achieved Data Rate in Stationary Condition (7.5Gbps)

In order to achieve such high data rates, 5G networks will inevitably exploit frequencies much higher and less congested than those of which are currently used for cellular networks (typically under 3 GHz). However, difficulties such as large propagation loss at these frequency bands have prohibited the industry from utilizing these bands for cellular applications. To address these challenges, Samsung has applied 'Hybrid Adaptive Array Technology' at 28 GHz frequency bands.

In addition, Samsung has revealed its '5G Rainbow' – seven technical requirements which are pillars to ensure a truly differentiating 5G user experience. The list comprises of maximum data rate, spectral efficiency, speed of mobility, data transmission rate at the cell boundary, the number of simultaneous connections, communication delays, and cost. To meet these requirements, Samsung has been developing key technologies including transmission technologies for high frequency bands, multiple access schemes and low latency networks among others.

The company is committed to developing innovative technologies that will enable the 5G era and plans to demonstrate its capabilities at the upcoming 2018 PyeongChang and 2020 Tokyo Olympics.

Australian Government
Department of Communications and the Arts

5G—Enabling the future economy

October 2017

Disclaimer

The material in this paper is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this paper.

This paper has been prepared for consultation purposes only and does not indicate the Commonwealth's commitment to a particular course of action. Additionally, any third party views or recommendations included in this paper do not reflect the views of the Commonwealth, or indicate its commitment to a particular course of action.

Copyright

© Commonwealth of Australia 2017

The material in this paper is licensed under a Creative Commons Attribution—4.0 Australia license, with the exception of:

- the Commonwealth Coat of Arms
- this Department's logo
- any third party material
- any material protected by a trademark, and
- any images and/or photographs.

More information on this CC BY license is set out at the creative commons website: https://creativecommons.org/licenses/by/4.0/. Enquiries about this license and any use of this paper can be sent to: Department of Communications and the Arts, GPO Box 2154, Canberra, ACT, 2601.

Attribution

Use of all or part of this paper must include the following attribution:

© Commonwealth of Australia 2017

Using the Commonwealth Coat of Arms

The terms of use for the Coat of Arms are available from the It's an Honour website (see www.itsanhonour.gov.au and click 'Commonwealth Coat of Arms').

Contents

Australia's 5G vision **1**

What is 5G? **3**

The 5G economics case **3**
Enhanced Mobile Broadband (eMBB) 3
Massive machine type communications (mMTC) 4
Critical communications 4

What is different about 5G? **6**
Network slicing 6
Mesh networks 6
Spectrum sharing 7
Antenna technology and network topology 7
Business case 8

How are countries preparing for 5G? **9**
Industry 5G preparations in Australia 9

The Government's direction for 5G **10**
Making spectrum available in a timely manner 10
Actively engaging in international spectrum harmonisation activities 11
Streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly 11
Reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose ... 12

Next steps **12**

Australia's 5G vision

5G is the next step in the evolution of mobile wireless communications technology, promising improved connectivity, greater network speeds and bandwidth, and very low latency. It is the fifth generation in mobile technology which, at each step, has seen significant developments in communications networks:

- 1G—The first generation of mobile phone networks were deployed in the early 1980s, providing a basic voice service using analogue transmission.
- 2G—In 1991, second generation networks were deployed, making the switch to digital standards with improved voice messaging and the introduction of the short message service (SMS).
- 3G—The third generation launched in 2001 and introduced data services in addition to voice and SMS.
- 4G—In 2009, the fourth generation protocol, Long Term Evolution (LTE), was introduced, supporting improved mobile broadband which saw increased capacity and speed for data.

New capabilities of mobile communications networks enabled by 5G technology will allow for a variety of 'use cases':

- higher quality and more video services provided to multiple users with full mobility, even at high speed
- massive scale automation delivered through widespread sensor networks and multiple connected devices
- delivery of critical communications assured by low latency and ultra-reliable networks, and
- improved productivity assisted by high quality, real time data analytics.

Unlike existing mobile communications networks, 5G networks have the potential to allow tailoring of requirements for each of these different use cases within the same network.

The Government considers that 5G is more than an incremental change for mobile communications. Instead, it provides the underlying architecture that will enable the next wave of productivity and innovation across different sectors of the Australian economy. Efficient rollout of 5G and uptake of the services it supports has the potential to produce far-reaching economic and social benefits and support growth of Australia's digital economy. This will be supported by the rollout of the National Broadband Network (NBN) allowing greater capabilities for the seamless delivery of services across high speed mobile, fixed line and fixed wireless networks.

The Government wants to create an environment that allows Australia's telecommunications industry to be at the forefront of seizing the benefits of 5G across the economy. The communications sector will lead the rollout of 5G networks in Australia. However, the Government can create the policy and regulatory environment to support a more efficient rollout, given its potential benefits to the economy.

The Government's direction will be to support the timely rollout of 5G in Australia to enable the next wave of broad-based industry productivity, and support the growth of Australia's digital economy.

This includes immediate actions by Government that enable the communications market to introduce new 5G technologies in line with international developments. These include:

- making spectrum available in a timely manner
- actively engaging in the international standardisation process
- streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly, and
- reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose.

The Government recognises that as 5G continues to develop, other issues relating to the technology will likely emerge which may require future Government action. In particular, while there are opportunities for 5G to create economy-wide transformation, this will require a broader examination of sectoral regulatory frameworks.

To that end, the Government will work collaboratively with industry to foster an ongoing dialogue on 5G beyond the launch of this paper to identify and remove sectoral barriers to its successful and timely rollout. Through this dialogue, the Government will also look at opportunities to build on other Government activities, such as the national Digital Economy Strategy which will more broadly focus on building the productivity of sectors across the economy.

What is 5G?

The International Telecommunication Union (ITU) is the United Nations specialised agency for information and communications technologies. This body decides global spectrum allocation frameworks and harmonises international spectrum to ensure networks and connected devices can communicate seamlessly. The ITU will undertake the formal, international process to identify bands for 5G by 2020. It has developed draft technical specifications for 5G which include:[1]

- high data rates (1 Gbps for hotspots, 100 Mbps download and 50 Mbps upload for wide-area coverage)
- massive connectivity (1 million connections per square kilometre)
- ultra-low latency (1 millisecond)
- high reliability (99.999% for mission critical ‘ultra-reliable’ communications), and
- mobility at high speeds (up to 500 km/h i.e. high speed trains).

In working towards these specifications, 5G represents a significant leap from the capabilities of previous generations and introduces a range of new technological possibilities. The success of 5G in delivering new technologies and services will be supported by existing communications infrastructure, including the NBN. This convergence of high-speed fixed-line and mobile services will collectively produce a consistent and ubiquitous user experience.

The 5G economics case

Unlike early generations of mobile networks, 5G will represent a significant shift in the telecommunications industry’s focus away from voice and more towards mobile broadband and increased industrial applications. These new use cases are expected to create benefits across a range of sectors—including transportation, health, manufacturing and agriculture—and have varying networking requirements. These use cases, as identified by the industry, can be divided into the following categories:

- enhanced Mobile Broadband
- massive Machine Type Communications, and
- critical communications.

Enhanced Mobile Broadband (eMBB)

eMBB will deliver improved capacity to a greater number of devices. This will enable higher rates and volumes of data transmission per device and improve coverage to a broader range of locations. eMBB will likely be the focus of early 5G deployments as it can immediately support the growing communications requirements for the digital economy.

An improved mobile experience for consumers

5G networks will give consumers a better mobile experience in more locations. Increased network capacity will support more users, even in crowded areas, such as large public events, and at peak times. Faster network speeds will also enable consumers to view rich content in more places, supporting the streaming of live events and high resolution media.

---

[1] ITU, 23 February 2017, ‘Press Release: ITU agrees on key 5G performance requirements for IMT-2020’.

Massive machine type communications (mMTC)

As 5G networks mature, they will support the widespread and dense deployment of sensors and other network-connected devices by significantly reducing their power requirements and providing flexible coverage across different spectrum bands. This proliferation of the Internet of Things (IoT) across industries is expected to produce significant productivity benefits and support integration between sectors.

Supporting productivity and innovation

The term Industry 4.0 describes the next step in the advancement of the manufacturing sector (the ‘fourth industrial revolution’). Industry 4.0 introduces autonomous systems supported by a combination of technologies such as IoT, artificial intelligence, continued technological improvements and digitalisation in manufacturing.

Australia stands to benefit from Industry 4.0, given our world-class manufacturing sector, which includes several high-value industries such as medical technology and aerospace. Australian manufacturers can improve their productivity and international competitiveness through Industry 4.0 processes by supporting their participation in global value chains. This is of particular benefit to SME manufacturers, opening them up to new markets and opportunities. 5G can support Industry 4.0, by providing communications infrastructure that is more accessible and flexible to suit specific industry needs.

5G can enable innovation in other sectors such as agriculture. A challenge for Australia’s agricultural sector is identifying how to improve productivity while balancing environmental and commercial constraints. Precision agriculture, which focuses on improving yields and minimising economic risks, seeks to provide more control in the management of agricultural production. While precision agriculture requires a range of enablers—including data analysis, sensor networks and geographical information systems—5G can provide the supporting infrastructure for these technologies.

Critical communications

Low latency and ‘ultra-reliable’ communications networks will support the delivery of critical communications, i.e. to support public safety use and playing role in the technology ecosystem supporting autonomous vehicles. In addition to automation, critical communications will also help to support technological advancement in areas including robotics and artificial intelligence.

The social benefits from autonomous vehicles

5G networks are expected to play a role in the technology ecosystem supporting the development of autonomous vehicles, which will enable a number of social benefits for transportation. Traffic congestion, which is estimated to cost Australia $53 billion by 2031,[2] could be proactively reduced by smart city traffic management systems that are informed by machine-to-machine communications with autonomous vehicles.

Improved road safety is also expected to be a key outcome of autonomous vehicles, as the majority of car accidents involve human error. In the 12 months to July 2017, there were 1,235 deaths on Australian roads with road trauma costing the Australian community an estimated $27 billion annually.[3] Autonomous vehicles can have a valuable role not just in terms of financial savings, but in saving human lives.

---

[2] Infrastructure Australia, 2015, ‘Australian Infrastructure Audit’.

[3] Bureau of Infrastructure, Transport and Regional Economics, August 2017, ‘Road deaths—monthly bulletins’.

Figure 1—Outline of the relationship between the technical requirements of 5G and expected services and applications it will deliver

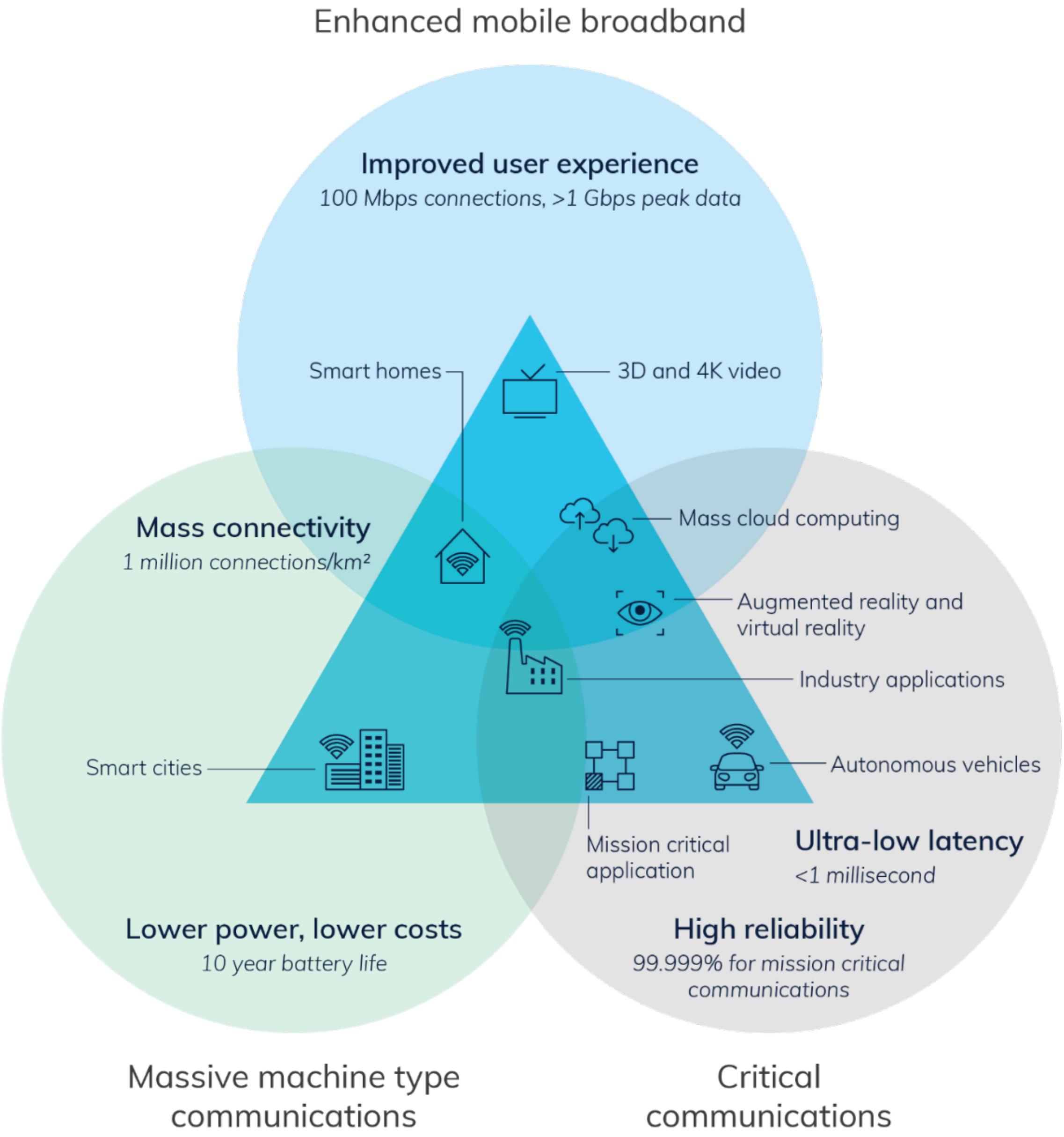
Enhanced mobile broadband
Improved user experience
100 Mbps connections, >1 Gbps peak data
Smart homes
3D and 4K video
Mass cloud computing
Mass connectivity
1 million connections/km²
Augmented reality and virtual reality
Industry applications
Smart cities
Autonomous vehicles
Mission critical application
Ultra-low latency
<1 millisecond
Lower power, lower costs
10 year battery life
High reliability
99.999% for mission critical communications
Massive machine type communications
Critical communications

Source: Based on figure 2 from the ITU paper, '[IMT Vision—Framework and overall objectives of the future development of IMT for 2020 and beyond].'

What is different about 5G?

5G is expected to enable productivity outcomes across key verticals of the economy as a result of a range of characteristics, such as **network slicing** and **mesh networks**. However, it will need to overcome deployment challenges such as **spectrum acquisition** and **coverage**. This is expected to lead to new approaches for deployment to enable widespread 5G coverage in Australia. The industry will also need to develop a robust **business case** for 5G rollout, which will likely be demand driven, to support the significant capital expenditure required.

Network slicing

Network slicing allows operators to split their network into separate sub-networks (also referred to as slices), enabling them to dedicate network resources to different users and applications. Sub-networks can 'slice' the infrastructure resources from the physical network to create virtual independent networks. This is a significant development from previous mobile network generations, as it enables an operator to deliver many different capabilities by creating slices that can be tailored for the intended usage. For example, an operator could create a network slice for IoT devices, or alternatively, a network slice for higher security for a higher quality of service for government or public safety uses.

Network slicing also provides a model for infrastructure sharing between network operators. In this case, a single transmitter's network could be shared with more than one operator. Given the capital expenditure that will be associated with 5G rollout, network slicing can provide a cost effective, short-term solution for operators.

Figure 2—5G network slicing enables operators to create separate virtual networks to meet application requirements

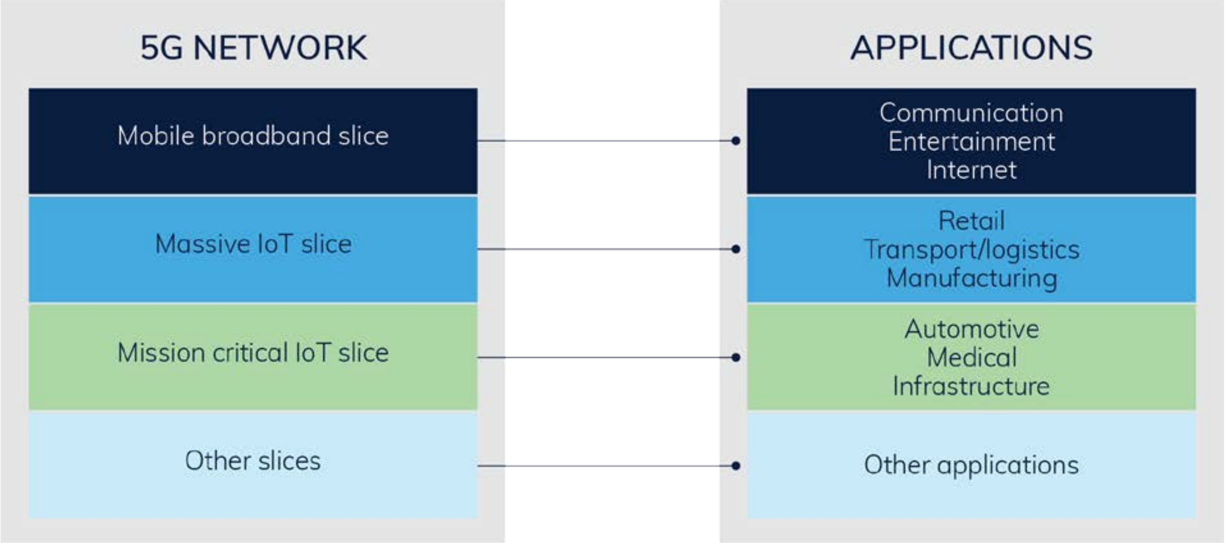

Source: Based on figure 2 from the ITU article, 'Why end-to-end network slicing will be important for 5G.'

Mesh networks

Mesh networks can be utilised to increase the range of coverage, where the 'mesh' is an interconnection among a network of devices. Only one device in the mesh needs to be connected to the network, which can then relay data to other nearby devices.

Mesh networks not only provide the opportunity to support 5G deployment but also have the additional benefit of providing efficient network speeds. Through dynamic routing, devices on a mesh network are able to seek the fastest and most reliable pathway to send and receive data. As such, this architecture can provide a cost effective solution for coverage in more remote areas. For example, primary industries would benefit from this network approach by simplifying network connections and costs associated with deploying and managing an IoT sensor network. However, the application of mesh networks is still highly speculative as the industry considers how they will operate in practice and in different environments. This may contribute to the development of new business models for 5G.

Spectrum sharing

Spectrum is a critical enabler of Australia's current and future communications infrastructure. The specific spectrum bands and quantity of spectrum required for 5G are still being considered. However, 5G will likely require a mix of low, medium and high frequency spectrum to meet different scenarios relating to coverage, connectivity, and latency:

- Low frequency (less than 1GHz)**—**providing widespread coverage across urban, suburban and rural areas and supporting IoT for low data rate applications.
- Medium frequency (1–6GHz)**—**providing good coverage and high speeds, and including the expected initial 5G range of 3.3–3.8GHz which has been identified as the most likely band for launching 5G globally.
- High frequency (above 6GHz)**—**providing ultra-high broadband speeds for advanced mobile broadband applications, and most suitable for applications in dense traffic hotspots.

5G technologies can be expected to deliver improvements in spectral efficiency (the data rate that can be supported per unit of spectrum). However, the use of 5G networks for applications such as widespread industrial applications is likely to require significantly more amounts of contiguous spectrum to be made available.

While some of the potential bands for 5G currently have unused spectrum, other bands would need to be 'refarmed', noting that it is likely that bands currently used for 2G, 3G and 4G in Australia will transition over time to 5G. Refarming enables spectrum to be transitioned to the highest value use as required. The refarming of spectrum already held by mobile broadband operators is a commercial decision for those operators. 5G is also expected to provide the opportunity for 'soft-refarming' where 4G and 5G technologies can both be supported simultaneously, minimising the impact to legacy devices during transition periods.

Spectrum sharing, that is spectrum accessed by numerous users on a shared basis, has also been identified as an option for 5G technology. Spectrum sharing encompasses a range of different aspects of spectrum management. Spectrum can be shared by geography, time, economic priority schemes, code modulation, polarisation, directionality or power. Access to spectrum is divided between users so it can be used without interference issues.

5G opens up new opportunities for increased spectrum sharing, through mechanisms such as network slicing. 5G technology is also designed to support shared arrangements, and allows for the sharing of the same spectrum ('unlicensed Wi-Fi' spectrum) with other technologies. Operators can augment their holdings in situations where existing exclusive holdings are insufficient to meet customer needs. Spectrum sharing in the 5G context is also supported by the expected use of highly directional antenna technology which would enable operators to operate in closer proximity without interference.

Antenna technology and network topology

5G will require radically different structures of networks if it is to achieve successful deployment in Australia. As 5G will likely utilise different frequencies, new equipment will be necessary. Additionally, the higher frequency 5G spectrum can only travel a small distance and will need more cells to ensure adequate coverage. However, antennas and equipment will be smaller, making it easier to attach these cells to existing infrastructure such as street lights and buildings.

The more dense deployment of cells will also give rise to other approaches that improve the reliability of data transmission across a 5G network. Data may be divided into individual streams and transmitted through multiple antenna segments in a process called Multiple Input Multiple Output (MIMO) which allows for more information to be transmitted simultaneously. This technique is further empowered by 'beamforming' which allows base stations to direct focused beams of energy to a specific area rather than dissipate the available power of a larger area. These developments will enable more efficient transmission and increase overall throughput.

5G is also expected to assist the adoption of IoT by further reducing power consumption through extended discontinuous reception. In this scenario, IoT devices shift between active and inactive cycles, transmitting only when required. This will allow connected devices to operate for extended periods on a single charge, reducing operational costs.

5G will also increase the support for a greater density of devices that would have otherwise been limited by the capacity on 3G and 4G networks. They will also enhance IoT deployment through the use of network slicing to create virtual network configurations that are optimised for the low power and coverage requirements of IoT networks.

Business case

For Australians to experience the benefits of 5G, the communications sector will need to explore and develop new business cases to attract investment and support the rollout of 5G services. As 5G becomes an integral part of the communications ecosystem, the sector will need to be agile to respond to the needs and expectations of other sectors which will be seeking to take advantage of these next generation networks.

While residential consumers will inevitably be attracted to the enhanced mobile broadband services offered by 5G, it is the industrial applications for 5G where industry expects to see the greatest opportunity for new business models.

Compelling industrial applications are still to be developed. For example, industry sectors can be expected to seek tailored solutions for their business needs such as enhancing their local area network or enabling autonomous systems, and small businesses will look for low cost deployments of IoT.

It is therefore expected that the model for 5G will be demand driven and will require the communications industry to foster new business opportunities.

How are countries preparing for 5G?

In preparation for 5G, many countries have been taking steps to test the technology and review their spectrum arrangements.

Examples of 5G work being undertaken by other countries

The United States is clearing the 600MHz band through an incentive auction for the potential early deployment of 5G. The United States has also identified reforms to infrastructure deployment as a priority for 5G rollout. Mobile network operators are also conducting trials of low-band spectrum for use in 5G services.

The Asia-Pacific Telecommunity (APT), the regional arm of the ITU, approved a recommendation for the use of the 700MHz band for 5G, with 26 countries in the Asia-Pacific region identifying this band for this use, including Australia, Japan, South Korea and New Zealand.

Korea and Japan have stated their intention to use some or all of the 26.5 to 29.5GHz range to trial enhanced mobile broadband applications ahead of the 2018 Winter and 2020 Summer Olympics respectively. It is expected that these trials will lead to commercial availability of 5G services.

The European Communications Commission (ECC) identified the 3.4–3.8 GHz and 26GHz band (from 24.25 to 27.5GHz) bands for the deployment of 5G in Europe. All European countries are expected to select at least part of this range to launch 5G by 2020.

In the United Kingdom, the government launched its 5G strategy in March 2017. It is funding testbeds to understand the different deployment requirements and security considerations for 5G. The United Kingdom is also working to make suitable spectrum available in the high (24.25 GHz–27.5 GHz, and other bands above 30 GHz), medium (3.4–3.8 GHz) and low frequency (700 MHz) bands.

Industry 5G preparations in Australia

Australia is well positioned to harness the opportunities of 5G. Australia has an effective and competitive mobile communications market, with voice and data coverage available to more than 99 per cent of the population.[4] It is the top performer internationally in terms of having in place effective enablers—infrastructure, affordability, consumer readiness and content availability—to support mobile internet adoption.[5]

5G trials have already commenced in Australia, with each of the main carriers working with mobile equipment suppliers in testing the application and limits of the technology. These trials will continue and will inform the communications sector on how 5G can be effectively deployed for the Australian environment.

---

[4] Department of Communications and the Arts, accessed 1 September 2017, 'Mobile phone towers'.

[5] GSM Association, 24 June 2016, 'Global Mobile Connectivity Index'.

The Government's direction for 5G

The Government has made significant investments to improve telecommunications infrastructure in Australia, through the NBN and the Mobile Black Spot Program. The Government is also working to create a policy and regulatory environment that supports a competitive and innovative communications market.

The Government recognises that 5G will enable innovation and productivity across industry sectors and can significantly contribute to Australian's growth and future prosperity. Therefore, the Government will focus on enabling the early deployment of this new generation of mobile networks in Australia and encourage its use in delivering new services and applications.

The Government will support the timely rollout of 5G in Australia to enable the next wave of broad-based industry productivity, and support the growth of Australia's digital economy.

Industry expects and needs to lead the deployment of 5G. However, the Government has a role in supporting network rollout by modernising policy and regulatory frameworks and removing barriers that would delay rollout and adoption unnecessarily.

In the first instance, the Government will support the early deployment of 5G in Australia by:

- making spectrum available in a timely manner
- actively engaging in international spectrum harmonisation activities
- streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly, and
- reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose.

Making spectrum available in a timely manner

A clear, efficient and flexible regulatory framework governing spectrum access will be essential to support the timely deployment of 5G networks in Australia.

The Government is currently undertaking work to modernise Australia's spectrum management framework to ensure it remains fit-for-purpose. In May 2017, it outlined its proposed reforms to the framework which are designed to simplify and streamline the processes for spectrum allocation and provide a transparent, efficient and flexible spectrum management framework. This will be the most significant change to the Australian spectrum management framework in the last 25 years.

The reforms will remove barriers between licence types, and enable flexible licensing issue and allocation processes. This strategic approach will remove outdated processes and support the Australian Communications and Media Authority (the ACMA) to more effectively respond to market demands and new technologies, such as 5G. This will help Australia remain internationally competitive with a modern, innovative economy over the coming decades.

The Government will put in place its new spectrum management framework by 2019.

In addition, the ACMA will continue to work on making spectrum available for 5G. The ACMA has been investigating the use of 1.5GHz and 3.6GHz and high frequency mmWave bands in considering additional spectrum for mobile broadband services. The ACMA has decided to prioritise refarming of the 3.6GHz band over the 1.5GHz band, citing industry submissions noting this band is likely to be a pioneer band for early 5G deployments and the need to provide greater clarity and investment certainty for incumbents and potential new band entrants alike. The ACMA is currently engaging with industry on which parts of the 3.6GHz band should be reallocated and on what terms. This approach also follows international trends which have seen the 3.6GHz band commonly used for 5G trials.

The ACMA will work to bring 3.6GHz spectrum to auction in 2018.

Actively engaging in international spectrum harmonisation activities

There is already significant work underway globally with several countries trialling 5G, but standards for this new generation technology are yet to be finalised. The formal, international process to define 5G is led by the ITU. The ITU's Working Party 5D is responsible for shaping the standard for "futuristic mobile technologies" to support International Mobile Telecommunications (IMT) for 2020 and beyond.[6] This process is known as IMT-2020.

Stakeholders such as regulatory and policy setting bodies, hardware manufacturers and governments of countries in which they are based will be seeking to influence the international dialogue. A key body is the industry-driven 3GPP which undertakes technology standardisation. Collectively, the ITU and 3GPP will drive spectrum harmonisation activities: the ITU, led by administrations, will focus on the spectrum requirements; and the 3GPP, led by industry, will concentrate on equipment and device standards.

Industry is well-placed to lead the standardisation process particularly given its role identifying the application of 5G technologies. However, there is also an important role for the Government in these processes. This is particularly the case in the harmonisation of international spectrum arrangements which will have significant impact on the availability and cost of 5G devices in Australia and can be strongly contested.

In the past, Australia and the Asia-Pacific region have been influential in contributing to standards and spectrum plans that have been adopted across the world. This ensured that Australia was able to adopt new technologies quickly and that the Australian market could take advantage of the economies of scale and have a greater choice of mobile handset equipment.

The Government will ensure strong participation by Australian in domestic and international discussions about 5G spectrum harmonisation. Our continued involvement provides the opportunity to contribute to this dialogue and secure outcomes that will benefit the adoption of 5G in Australia.

Streamlining arrangements to allow mobile carriers to deploy infrastructure more quickly

The design and deployment of mobile networks will be radically different from those of today. 5G is expected to require additional infrastructure in new forms, including smaller cells and more densely-located antennas, particularly in the use of high-band spectrum.

Carriers have specific powers and immunities relating to telecommunications infrastructure deployment and installation. These laws help carriers to rollout telecommunications infrastructure quickly in a nationally-uniform way, rather than having to follow state, territory and local government requirements. These laws have existed in their current form since 1997.

The Government has recently consulted on proposed new arrangements that take account of technology developments and changes in operating practices as well as identifying opportunities to streamline deployment processes. Under these new proposed arrangements, mobile carriers would be able to rollout new communications technologies such as 5G more efficiently.

The Government understands that some members of the community have expressed concerns about the impact on public amenity from increased telecommunications infrastructure. The Government considers that telecommunications providers should work with local communities to address concerns about their infrastructure plans and is encouraging industry to consider consultation requirements for future 5G networks.

The Government continues to work with stakeholders and will implement the first tranche of changes to carrier powers and immunities following the conclusion of the consultation process.

---

[6] ITU, accessed 1 September, ITU towards "IMT for 2020 and beyond."

Reviewing existing telecommunications regulatory arrangements to ensure they are fit-for-purpose

The pace of change in the communications sector is impacting on the effectiveness of the existing telecommunications regulatory framework. In response, the Government is progressively working to modernise the regulatory architecture to ensure that the regulatory and policy settings flexibly respond to the current and future needs of the communications sector.

In undertaking reform, the Government is cognisant that the communications regulatory framework will need to be sufficiently flexible to address the emergence of new technologies and business models.

5G deployment will benefit from the Government's reform efforts, including updates to the regulatory framework for telecommunications and radio spectrum and modernising the ACMA. In these cases, the Government has been revising regulation to shift it away from the traditional vertical telecommunications sectors to principles based, flexible arrangements. This will provide the versatility necessary to account for 5G developments and services.

The Government will continue to work with industry to modernise current telecommunications regulatory arrangements to ensure they encourage competition and innovation in the sector.

As with previous mobile networks, cybersecurity will be a critical consideration as 5G is deployed. Security will be even more so a challenge for 5G, as the reliability of communications will be pivotal in the technology's ability to deliver benefits, particularly in the case of critical communications.

Additionally, in providing the architecture for automation, 5G networks will trigger an ever-increasing volume of data. While the bulk of data will be machine-to-machine communications, users will want assurance that their personal information is protected. User consent will be an area of growing complexity due to the intersection between autonomous systems and the individual.

Industry has strong incentives to address cybersecurity risks in 5G's new types of network deployments and systems. However, **the Government will continue to assess cybersecurity and privacy issues as they evolve to ensure Australians have confidence in using 5G.**

Next steps

The Government recognises the opportunities presented by 5G for economy-wide transformation, creating productivity benefits in sectors such as transportation, health, manufacturing and agriculture. However, realising the benefits of 5G in sectors other than communications will need the right sectoral regulatory settings.

The Government will work to ensure that sectoral regulatory frameworks are updated to take advantage of 5G. The communications portfolio is well-placed to facilitate an ongoing strategic dialogue on 5G that will support sectors to identify and work towards unlocking the potential benefits of the technology. This dialogue would provide a starting point for greater engagement across Government and with industry and the community.

To that end, **the Government will establish a 5G working group that will bring together representatives from across Government and industry.** The working group will create a platform for this strategic dialogue with a mandate to seek out opportunities and emerging issues on 5G. This will provide better coverage across Government of the evolving policy and regulatory challenges associated with 5G.

Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

1-1-2016

Energy Efficient Switched Parasitic Array Antenna for 5G Networks and IoT

Ahmed Kausar
Boise State University

Hani Mehrpouyan
Boise State University

Mathini Sellathurai
Boise State University

Rongrong Qian
Heriot Watt University

Shafaq Kausar
National University of Sciences and Technology

© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. doi: 10.1109/LAPC.2016.7807569

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

Energy Efficient Switched Parasitic Array Antenna for 5G Networks and IoT

Ahmed Kausar*, Hani Mehrpouyan *, Mathini Sellathurai†, Rongrong Qian†, and Shafaq Kausar‡
*Department of Electrical and Computer Engineering
Boise State University, Boise, USA
Email: ahmedkausar@boisestate.edu, hanimehrpouyan@boisestate.edu
† School of Engineering and Physical Sciences
Heriot Watt University, Edinburgh, U.K
Email: m.sellathurai@hw.ac.uk, rq31@hw.ac.uk
‡Department of Electrical Engineering
National University of Sciences and Technology, Islamabad, Pakistan
Email: shafaqkausar@nust.edu.pk

***Abstract*—This paper includes design and implementation result of an adaptive beam forming antenna for upcoming 5G and Internet of Things (IoT). Switched parasitic array antennas are low cost, small sized and compact circular array antennas that steer beam in a desired direction by variation in switching pattern of parasitic elements. The proposed antenna design has an active center element, which is surrounded by several symmetrically placed parasitic elements. The designed antenna has a gain of 8 dB and is capable of 360 degrees beam steering in steps of 60 degrees each. Simulations are validated with results of the fabricated antenna. Antenna beam is steered by controlling parasitic elements. Future application of Electronically Steerable Parasitic Array Radiator (ESPAR) antennas and switched parasitic array antennas in next generation communication networks and methods for reducing size of the antenna are also highlighted.**

***Index Terms*—Internet of Things, 5G, circular array, reconfigurable, smart antenna, adaptive beamforming, Switched-Parasitic Array**

I. INTRODUCTION

Smart antennas are capable of adaptive beam forming in accordance with the environment. Over the last three decades wireless communication has made rapid progress due to better modulation schemes and better error correction schemes. Wireless communication systems are replacing the wired ones. Antenna is an integral part of wireless devices. To meet the ever-increasing demand of higher data rate, smart antennas capable of adaptive beam forming are introduced as a potential solution to increasing data rate and improving signal-to-noise ratio (SNR) in upcoming 5G and Internet of Things (IoT) [1]. By using smart antennas Equivalent Isotropically Radiated power (EIRP) can be increased. EIRP denotes product of the transmit power (P_t) and transmit antenna gain (G_t), i.e.

$$EIRP = P_t \times G_t \tag{1}$$

From (1), it is observed that by using a smart antenna G_t increases, which in turn leads to the increase in the EIRP. The beam of an antenna can be steered by means of mechanical movement of a directional antenna or by electronic beam steering. Electronic beam steering is resilient and provides more accurate approximation of the desired signal source. Moreover electronic beam steering antenna does not involve moving parts. Therefore, there is less wear and tear on the antenna as compared to mechanical beam steering. Generally electronic beam forming is achieved by using phase shifters coupled with an array of antenna elements. The phase of an individual phase shifter is varied such that there is net constructive interference in the intended direction of the antenna beam and destructive interference in all other directions [2]. Hence, electronic beam steering is preferred over mechanical beam steering.

Electronically Steerable Parasitic Array Radiator (ESPAR) antennas use mutual coupling between antenna elements to steer the beams instead of phase shifters. Design of parastic array radiator antennas is discussed in [3], [4] in which there is a center active element surrounded by a circular array of passive elements, where each passive element is loaded with a specific reactance. ESPAR antennas are smaller in size as compared to phase array antennas because elements have to be placed at a quarter of wavelength or less distance from each other for effective mutual coupling. Hence, ESPAR antennas are cost effective and space efficient. In such antennas we have only one active element, therefore, only one feed is required. Consequently, ESPAR antennas have lower losses (due to single feed) as compared to phase array antennas [3].

Previously designed antenna Electronically Steerable Parasitic Raidator(ESPAR) antennas have all cylindrical element, including the feed element [6], [7]. In this paper we have proposed design of seven element ESPAR antenna with conical center(active element)[1]. Length of conical element is optimized using optimetrics in HFSS. Our proposed antenna has higher bandwidth as compared to previously designed ESPAR antennas. [6], [7]. In multiple-input-multiple-output (MIMO)

---

[1] This work was supported by National Science Foundation (NSF) on Enhancing Access to Radio Spectrum (EARS)

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

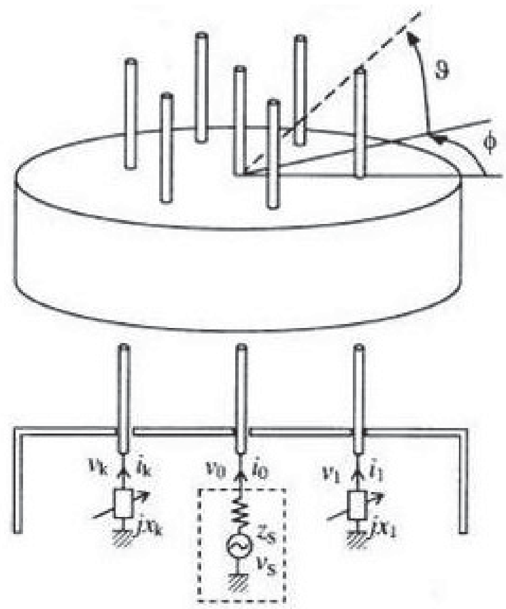

Fig. 1. Seven Element switched parasitic array antenna, ϕ is azimuth angle and θ is elevation angle.

systems there are multiple transmit and and receive antennas. Each receive antenna receives signal over different propagation channels yielding diversity gain. The MIMO technique can increase system throughput and reduce propagation losses. Similarly diversity gain can also be achieved by using smart antennas. In particular, we can transmit the same signal in multiple directions through directional beams, such that they experience different propagation channels [4]. Our proposed antenna is more suited for MIMO applications compared to previously designed model because of its higher gain and improved bandwidth efficiency.

This paper focuses on the design of a seven element switched parasitic array antenna. The antenna is first simulated and then it is fabricated. Simulation results are validated with the hardware results. Previously done work on ESPAR and switched array antennas involve cylindrical monopole elements mounted on a ground plain [8] due to cylindrical nature bandwidth of active monopole is confined, subsequently bandwidth of switched array and ESPAR antennas is limited. Instead on cylindrical monopole element we have used tapered conical monopole. Due to tapered design bandwidth is enhanced.

II. SWITCHED PARASITIC ARRAY ANTENNA CONCEPT

The conceptual overview of the switched parasitic array antenna is given in Fig. 1. We have first simulated the antenna with seven monopole elements. There is one element at center which is surrounded by a circular array. Feed is applied to the center element and it is termed as active element, whereas elements in the circular array are parasitic elements. Parasitic array antennas do not involve phase shifters, so that the manufacturing cost is less as compared to the phase array antennas. Each element is separated from other element by a quarter of wavelength, thus mutual coupling phenomena comes into account and there is a shift in phase of current in each element because of this mutual coupling phenomenon. Since switched parasitic array antennas offer adaptive beamforming solution with compact size and low cost, they are likely to be used in future communication devices to increase data-rates [6], [7]. Parasitic array antennas have vast potential for usage in laptops, cell phones, WLANs etc.

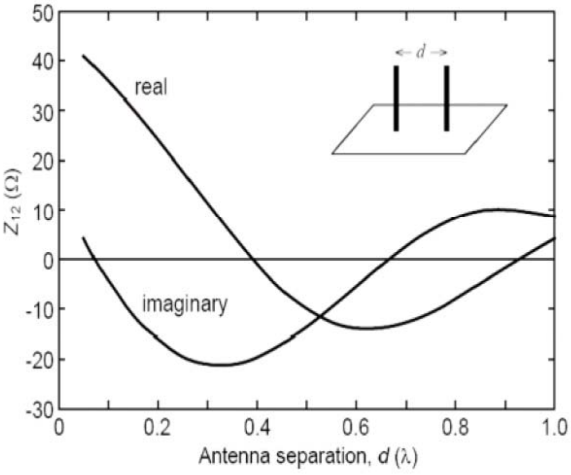

Fig. 2. Z_{21} versus separation distance.

A. Mutual Coupling

Mutual Coupling is basically the interaction between antenna elements in an array. Since parasitic elements are placed closer to each other (at a distance of λ/4) there will be mutual coupling.This mutual coupling phenomenon is responsible for a shift in phase of current induced in each element [8].The current induced by one antenna element in another antenna element [9] is given by (3)

$$I_2 = \frac{Z_{21}}{Z_{22}} \times I_0 \tag{2}$$

$$Z_{21} = \left|\frac{V_2}{I_1}\right|_{I_2=0} \tag{3}$$

$$Z_{21} = \left|\frac{V_2}{I_2}\right|_{I_1=0} \tag{4}$$

Here Z_{22} is constant as it is open-circuit output impedance. Z_{21} is open-circuit transfer impedenace from port 2 to port 1 and it can be approximated by graph shown in Fig. 2. From equation (3), beam can be steered in the desired direction by varying the phase of current I_2 [10].

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

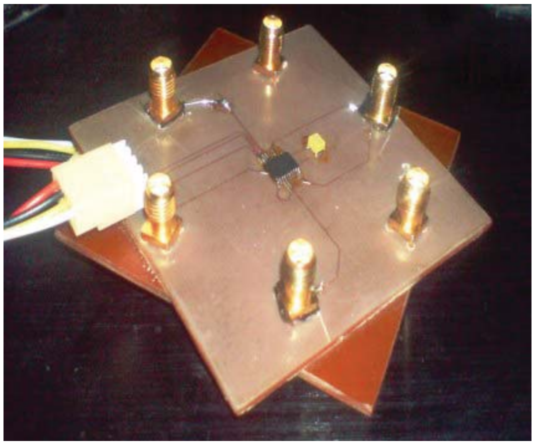

Fig. 3. RF PCB with Analog devices ADG 904 RF switches.

TABLE I
DESIGN PARAMETERS

Parameters	Values
Ground Skirt length	lambda/4
Monople Length	lambda/4
Frequency	2.45 GHz
Ground Radius	lambda/2
Monopole Radius	lambda/200
Ground Thickness	3 mm

B. *Beam Steering Mechanism*

In this work the designed antenna is capable of 360^0 beam steering in steps of 60^0. The beam steering is achieved by using two sets of 4:1 Analog devices multiplier switch ADG 904. Fig.3 shows RF PCB designed for beam steering. Beam is steered towards an element which is open. The open element behaves as a director and shorted elements act as reflectors. In order to achieve a maxima at 0^0, we make the element 1 open by using RF switch IC while other elements (i.e., elements 2,3,4,5 and 6) are shorted. This mechanism is similar to Yagi-Uda antenna where elements with shorter lengths in front of feed element act as directors and elements at rear of feed element acts as reflectors [11]. In Yagi-Uda antenna feed element is often a dipole or folded dipole.

III. PARASITIC ARRAY ANTENNA DESIGN

Design parameters and mechanical design of antenna is included in this section.

A. *Design Parameters*

Antenna design parameters are tabulated and the results are optimized using antenna simulation tools. Ground is skirted, since skirted ground provides mechanically sound model for PCB assembly to mount at back of antenna [12]. Skirt is λ/4 in length similarly each monopole element is lambda/4 in length and λ/200 is its radius.Table I shows the design parameters for the seven-element switched parasitic array antenna.

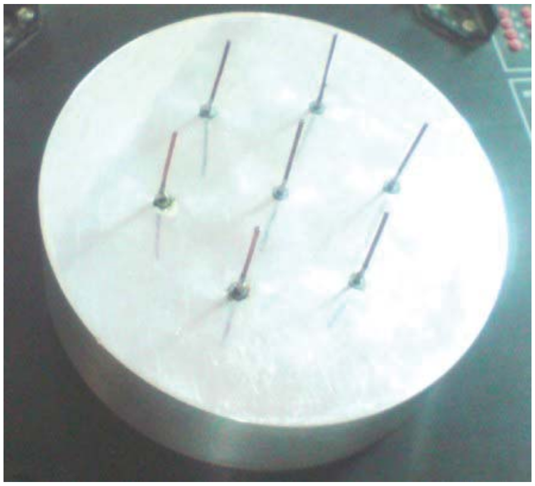

Fig. 4. Designed seven element parasitic switched array antenna.

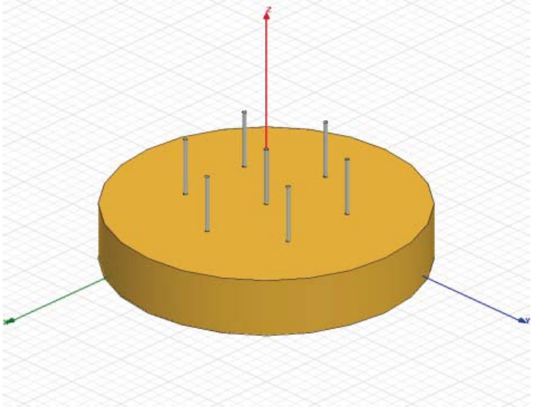

Fig. 5. Simulated antenna model.

B. *Mechanical Design*

Monopole elements are made by cutting copper wire of desired gauge i.e lambda/200 in radius [13] and these elements are soldered on pin of SMA connectors.Aluminum is used as ground plane. Electromagnetic waves are reflected by the ground. Instead of using seven dipole elements we have used seven monopole elements mounted on the ground plane [14]. Ground is made hollow in order to create space for mounting control circuitry [15]. Fig. 4 shows designed hardware.

IV. ANTENNA SIMULATION AND PLOTS

The antenna is first simulated and then simulated results are compared with the measured results of fabricated antenna. Simulated results were in line with the hardware results and 8 dB gain was observed in simulated results as well as in fabricated antenna. In simulation each of the parasitic element is loaded with lumped port excitation, simulating opening and shorting [16] of monopole elements. Fig. 5 shows simulated model.

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

TABLE II
MAXIMA PATTERN

Azimuth Angle	Element 1	Element 2	Element 3	Element 4	Element 5	Element 6
0 degree	Open	Short	Short	Short	Short	Short
30 degree	Open	Open	Short	Short	Short	Short

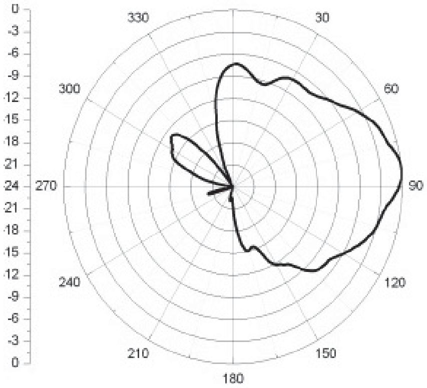

Fig. 6. Anechoic chamber results for θ=90^0 (Beam in elevated because we are using monopole elements with finite ground).

Two dimension polar plots depicts the beam steering in the direction of given elevation angle, theta (θ) and azimuth angle, phi (ϕ). Anechoic chamber results for θ=90 and ϕ=0 and 30 degrees are shown in Fig. 7 and 8 respectively. Using large ground plane angle of elevation is reduced closer to 90^0. Fig. 6 shows anechoic chamber result for θ=90 one can observe small elevation form desired x-y axis because of ground skirt [17].In simulations elevation angle of 85^0 is observed. Fig. 6(a) shows antenna simulated plot for elevation angle θ and Fig. 6(b) shows anechoic chamber results.Anechoic chamber plot of Fig. 7 shows maxima at ϕ=0^0, in order to get maxima at 0^0 elements from element number 2-6 are shorted and the element 1 is opened. Similarly Anechoic chamber plot of Fig. 8 shows maxima at ϕ=30^0, in order to get maxima at 30^0 elements from element number 3-6 are shorted and element 1 and element 2 are open.

Traditional ESPAR antenna and switched parasiric array antenna have all cylindrical elements, including the active element [3], [4]. We have made center element conical instead of cylindrical. Due to conical nature of active element response our frequency range and improved. In order to validate the gain enhancement both models were simulated and results were compared. Fig 9. (a) shows conventional design and Fig. 9(b) shows simulated antenna with conical center element. Fig. 10 shows S_{11} plot for cylindrical active element. Considering -8_{dB} as cut off point bandwidth range from 2.358 GHz to 2.51 GHz (0.152 GHz). Fig. 11 shows $S-11$ plot for centre conical element, considering 8_{dB} as cut off point bandwidth ranges from 2.36 GHz to 2.555 GHz (0.195 GHz). Therefore my making center element conical we have improved bandwidth as compared to traditional switch array parasitic antennas. Low profile design for switched parasitic array antennas in also highlighted in [18].

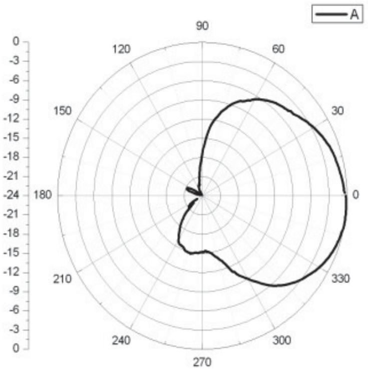

Fig. 7. Anechoic chamber results for ϕ=0^0.

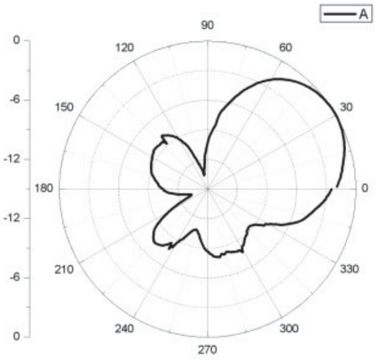

Fig. 8. Anechoic chamber results for ϕ=30^0.

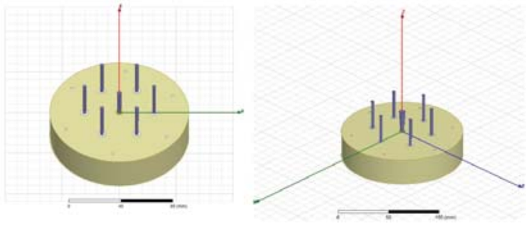

Fig. 9. (a) Conventional ESPAR antenna model with cylindrical Center Element, Fig. 9 (b) Designed ESPAR antenna with conical element for Bandwidth Enhancement.

V. FUTURE WORK AND CONCLUSION

5G networks will be designed for device to device, device to human and human to human interactions. High data rates are required to meet demands of 5G systems. One of proposed bands for 5G is in mm-Waves range i.e from 30GHz to 60GHz for such high frequencies path losses are considerable and beam steering antennas are way forward to overcome path

This is an author-produced, peer-reviewed version of this article. The final, definitive version of this document can be found online at *2016 Loughborough Antennas & Propagation Conference (LAPC)*, published by IEEE. Copyright restrictions may apply. doi: 10.1109/LAPC.2016.7807569

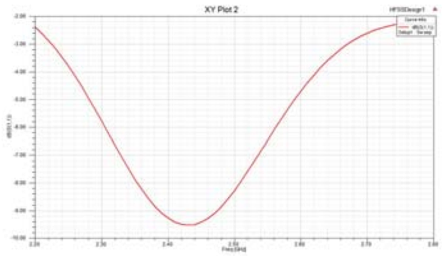

Fig. 10. S_{11} plot for conventional switched array antenna.

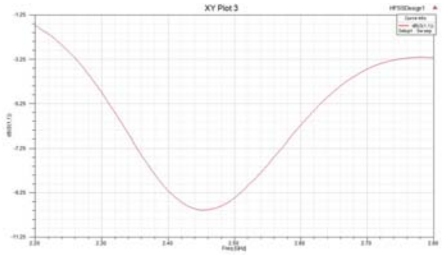

Fig. 11. S_{11} plot for center conical element switched array antenna.

losses [19]. ESPAR and switched parasitic array antennas provide a cost effective way of incorporating adaptive beamforming in future 5G and internet of Things (IoT) devices [20]. However for the efficient use of ESPAR antennas in mobile terminals the size of the antenna system must be reduced, for this further research on dielectric switched parasitic array antennas is underway. Seven element switched parasitic array antenna is designed that is capable of dynamic beam steering. Gain of 8dB is achieved practically. Simulated results are in line with the anechoic chamber results. Designed antenna increase the channel capacity by improving signal to noise ratio. Gain of 8dB was achieved in specified directions of elevation angle ,theta (θ) and azimuth angle, phi (ϕ). Antenna is optimized in terms of interference reduction for its use in wireless ad hoc networks, operating at frequency of 2.45 GHz. Bandwidth of designed antenna is 450 MHz and beam is steered in 360^0.

ACKNOWLEDGEMENT

Thanks to EPSRC Project EP/M014126/1, Large Scale Antenna Systems Made Practical: Advanced Signal Processing for Compact Deployments [LSAS-SP] and National Science Foundation (NSF), Enhancing Access to Radio Spectrum (EARS) program for funding.

REFERENCES

[1] M. Rzymowski, P. Woznica, and L. Kulas, "Single-anchor indoor localization using espar antenna," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 1183–1186, 2016.

[2] L. Zhou, F. A. Khan, T. Ratnarajah, and C. B. Papadias, "Achieving arbitrary signals transmission using a single radio frequency chain," *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 4865–4878, Dec 2015.

[3] V. I. Barousis and C. B. Papadias, "Arbitrary precoding with single-fed parasitic arrays: Closed-form expressions and design guidelines," *IEEE Wireless Communications Letters*, vol. 3, no. 2, pp. 229–232, April 2014.

[4] M. R. Islam and M. Ali, "Elevation plane beam scanning of a novel parasitic array radiator antenna for 1900 mhz mobile handheld terminals," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 10, pp. 3344–3352, Oct 2010.

[5] S. Kausar, H. U. Rahman, T. Hassan, and A. Kausar, "Miniaturization of espar antenna using folded monopoles and conical central element," in *Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2015 International Conference on*. IEEE, 2015, pp. 87–91.

[6] H. T. Liu, S. Gao, and T. H. Loh, "Electrically small and low cost smart antenna for wireless communication," *IEEE Transactions on Antennas and Propagation*, vol. 60, no. 3, pp. 1540–1549, March 2012.

[7] T. H. Loh, H. Liu, and S. Gao, "A cost-effective direct magnitude measurement methodology for smart antennas," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 4, pp. 2043–2050, April 2013.

[8] F. Khan, K. Ali, A. Kausar, and S. Kausar, "Energy generation capacity analysis of a canal based hydro project," in *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation*, Sept 2013, pp. 277–281.

[9] A. Kausar, M. A. Cheema, S. Kausar, H. Rehman, and T. Hassan, "Espar antenna system designing & simulation," in *Proceedings of the 2014 First International Conference on Systems Informatics, Modelling and Simulation*, ser. SIMS '14. IEEE Computer Society, 2014, pp. 193–196. [Online]. Available: http://dx.doi.org/10.1109/SIMS.2014.37

[10] S. Akkar and A. Gharsallah, "Reactance domains unitary music algorithms based on real-valued orthogonal decomposition for electronically steerable parasitic array radiator antennas," *IET Microwaves, Antennas Propagation*, vol. 6, no. 2, pp. 223–230, January 2012.

[11] S. Kausar, H. u. Rahman, A. Kausar, and T. Hassan, "Espar antenna system for dynamic tracking of active targets," in *Modelling Symposium (EMS), 2013 European*, Nov 2013, pp. 533–535.

[12] A. Umar Kausar, F. Aslam Khan, T. Hassan, A. Nasser, and S. Kausar, "Adaptive espar antenna system modelling & design." *International Journal of Simulation–Systems, Science & Technology*, vol. 14, no. 4, 2013.

[13] D. v. B. M. Trindade, C. Mller, M. C. F. D. Castro, and F. C. C. D. Castro, "Metamaterials applied to espar antenna for mutual coupling reduction," *IEEE Antennas and Wireless Propagation Letters*, vol. 14, pp. 430–433, 2015.

[14] S. Kausar, Ahmed Kausar, "Smart adaptive beam forming antenna for interference cancellation," in *Wireless Days (WD), 2013 IFIP*. IEEE, 2013, pp. 1–3.

[15] A. Kausar, S. Kausar, and T. Hassan, "Smart adaptive beam forming antenna for interference minimization," in *Future Generation Communication Technology (FGCT), 2013 Second International Conference on*. IEEE, 2013, pp. 6–9.

[16] F. Aslam, E. Ulhaq, A. Nasser, and A. Umar, "Intelligent modeling scheme for detection of line losses in power distribution system," in *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on*. IEEE, 2013, pp. 218–223.

[17] R. Qian, M. Sellathurai, and D. Wilcox, "A study on mvdr beamforming applied to an espar antenna," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 67–70, Jan 2015.

[18] A. Pal, A. Mehta, D. Mirshekar-Syahkal, and H. Nakano, "A low-profile switched-beam dual-band capacitively coupled square loop antenna," in *Antennas and Propagation Conference (LAPC), 2013 Loughborough*, Nov 2013, pp. 563–566.

[19] A. Kausar, S. Kausar, T. Hassan *et al.*, "Interference cancellation using adaptive beam forming for 3 d beam steering." *International Journal of Simulation–Systems, Science & Technology*, vol. 14, no. 1, 2013.

[20] M. R. Nikkhah, P. Loghmannia, J. Rashed-Mohassel, and A. A. Kishk, "Theory of espar design with their implementation in large arrays," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 6, pp. 3359–3364, June 2014.

[21] T. Hassan, A. Kausar, H. Umair, and M. A. Anis, "Gain optimization of a seven element espar antenna using quasi-newton method," in *Microwave technology & computational electromagnetics (ICMTCE), 2011 IEEE international conference on*. IEEE, 2011, pp. 293–296.

978-1-5090-0783-7/16/$31.00 ©2016 IEEE

Recent Antenna System Technologies for Next-generation Wireless Communications

Koichi Tsunekawa†

Abstract

This paper describes the requirements of the antennas or antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying them. It also explains the antennas and antenna systems being researched and developed for next-generation wireless communication systems. The antenna in such a system requires high gain, high-efficiency antenna elements and a multi-antenna design. This paper introduces a high-efficiency antenna created by integrating it with a monolithic microwave integrated circuit (MMIC) in a millimeter-wave frequency-band system, a multi-antenna system for the multiple-input multiple-output (MIMO) technique that enables multiple data transmission for land mobile communication systems, and a beam-scanning antenna for satellite communication systems. It also explains the general research and development trend of antenna systems and technical problems in each wireless system as an introduction to the other selected papers in this issue.

1. Introduction

Wireless access technologies have advanced rapidly [1] and broadband wireless communication systems have become popular in our daily lives. An antenna is a very important component of a wireless system because it acts as the input and output interface for wireless equipment, as shown in **Fig. 1**. However, system performance is determined by the total performance of each component, such as the antenna and the radio-frequency (RF), intermediate-frequency (IF), and baseband (BB) units. For recent high-performance wireless systems, overall system design has been essential. Accordingly, the antenna must not

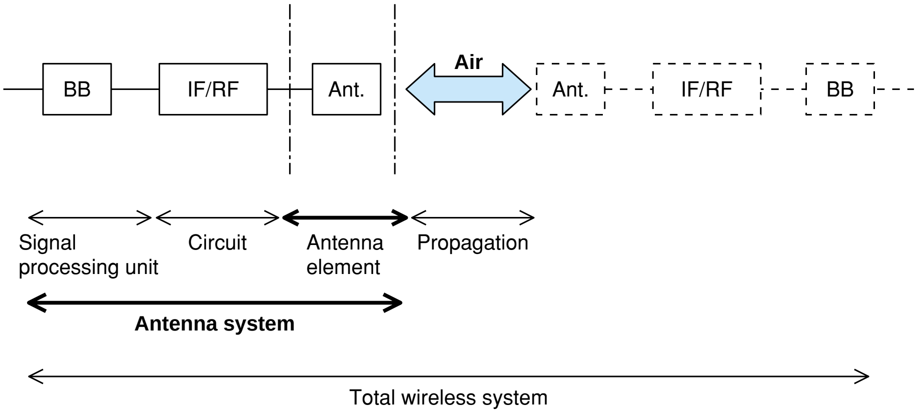

Fig. 1. Wireless and antenna system.

---

† NTT Network Innovation Laboratories
Yokosuka-shi, 239-0847 Japan
E-mail: tsunekawa.koichi@lab.ntt.co.jp

only have antenna elements with high-performance characteristics, but must also perform effectively as an antenna system in conjunction with the RF, IF, and BB units. New functions can be added through antenna system design: for example, appropriate radiation pattern control provides a large system capacity by using baseband signal processing and an antenna element with an integrated RF circuit provides high gain without feeder loss.

The selected papers in this issue describe the latest antenna and antenna system technologies for achieving high performance. Two of the papers present beam-scanning antenna systems in satellite broadband communication. Beam scanning is carried out using by two methods: one [2] uses a variable active device or tunable reactance device incorporated into the antenna element, which has high-speed electronic control while the other [3] uses mechanically controlled antenna movement, which has a very precise direction control function. Another paper [4] introduces an active integrated antenna system with an RF circuit, which provides high gain and high efficiency, for a high-capacity millimeter-wave transmission system. And two papers describe a multiple-input multiple-output (MIMO) antenna system for obtaining high-speed transmission and high system capacity in the land mobile communication system from the viewpoints of beamforming techniques [5] and implementation methods [6].

As an introduction to these advanced antenna techniques, this paper provides a general explanation of the requirements of antennas and antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying these requirements. It also describes current antenna systems and NTT's antenna research results for typical wireless systems.

2. Antenna requirements in recent wireless communication systems

The trend of wireless systems is shown in **Fig. 2**. This indicates the typical system transmission speed and number of antennas needed to achieve a particular system performance for each operating frequency. Wireless communication systems are moving towards high transmission speeds and high capacity by using higher system frequencies to enable a wide frequency

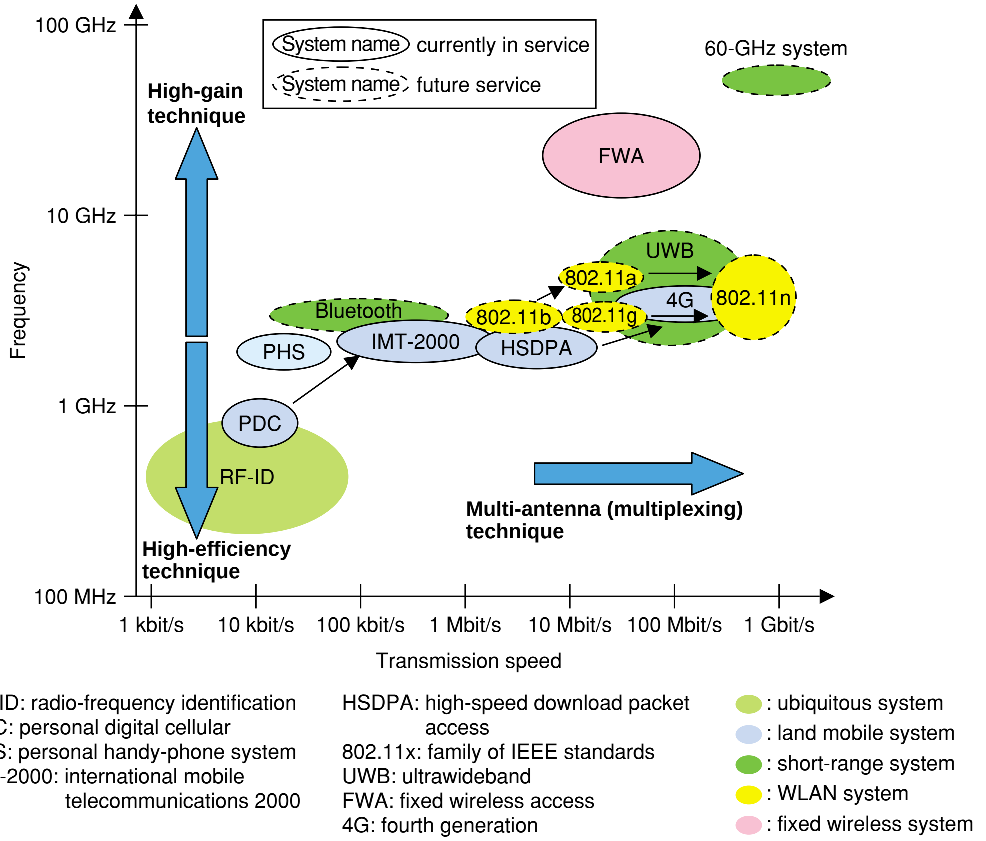

Fig. 2. Advances of terrestrial wireless communication systems and their relationship to antenna technologies.

band and using advanced system technology to achieve high quality and signal multiplexing. The antenna requirements for these system properties are:

- High-gain and high-efficiency technology
- Multi-antenna technology

Increases in the gain and efficiency of an antenna are offset by higher propagation loss at high frequencies, so the received signal level needs to be higher. High gain is an age-old problem in antenna research and development. Multi-antenna technology improves the receiving signal quality and provides multiplexed signal transmission by using an antenna system that includes RF circuits, BB units, and antenna elements.

2.1 High-gain and high-efficiency antenna

Theoretical and experimental antenna gain and efficiency characteristics are shown in **Fig. 3**. The dependences of gain and beamwidth on horn dimension L(λ) in a relatively large antenna are explained in Fig. 3(a) [7]. When L is more than 0.5 times the wavelength, the antenna's efficiency is about 100%; a larger antenna provides higher gain. However, the high-gain radiation pattern leads to a narrow beamwidth. Therefore, a mobile terminal needs to use beam-scanning or beam-tracking techniques to follow the transmitter while the terminal moves. On the other hand, the efficiency decreases if the antenna size is less than 0.5 wavelengths. The bandwidth of a linear dipole antenna versus its efficiency for various antenna sizes (D: diameter) is shown in Fig. 3(b). This graph was estimated from theoretical limits [8] and many experimental results. The antenna efficiency becomes approximately –5 dB when antenna bandwidth of 10% is provided using a linear dipole antenna with a length of 0.3 wavelengths. This figure shows that the efficiency and bandwidth have a tradeoff relationship and that the efficiency drops greatly as the antenna size decreases. Therefore, some clever ideas for the antenna element structure are needed to overcome these weak points.

2.2 Multi-antenna system

A multi-antenna system has several antenna elements, but it does not merely work as an ordinary array antenna. The performance requirements are satisfied because the antenna system includes RF circuits, BB units, and antenna elements. The antenna element technologies and the multi-antenna system are shown in **Fig. 4**. In the antenna element, a high gain can be achieved by improvements to the antenna element structure, for example, by using a new material, adding a parasitic element, or integrating an RF circuit (Figs. 4(a)–(c)). The antenna shown in Fig. 4(c) is called an integrated antenna. It can control the radiation pattern as well as provide high gain by eliminating feeder loss. In a multi-antenna system (Fig. 4(d)), the radiation pattern is controlled by baseband signal processing, so the pattern of each wireless channel can be controlled by each baseband channel. Furthermore, it is possible to get multiplexed signal transmission in the space domain by baseband signal processing. Multi-antenna systems are generally called smart antennas or active antennas [9].

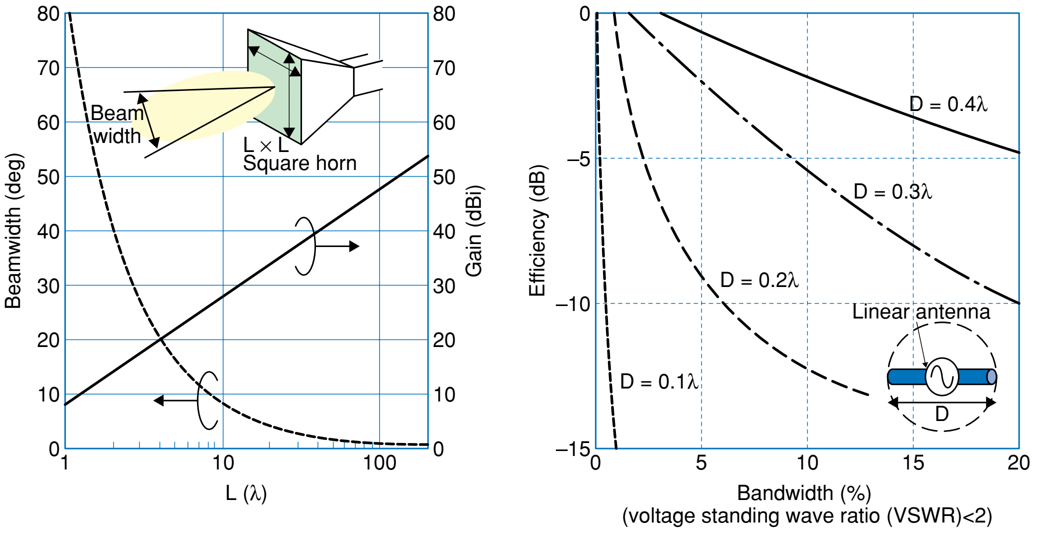

(a) Aperture antenna gain characteristics

(b) Dipole antenna efficiency characteristics

Fig. 3. Antenna gain and efficiency characteristics.

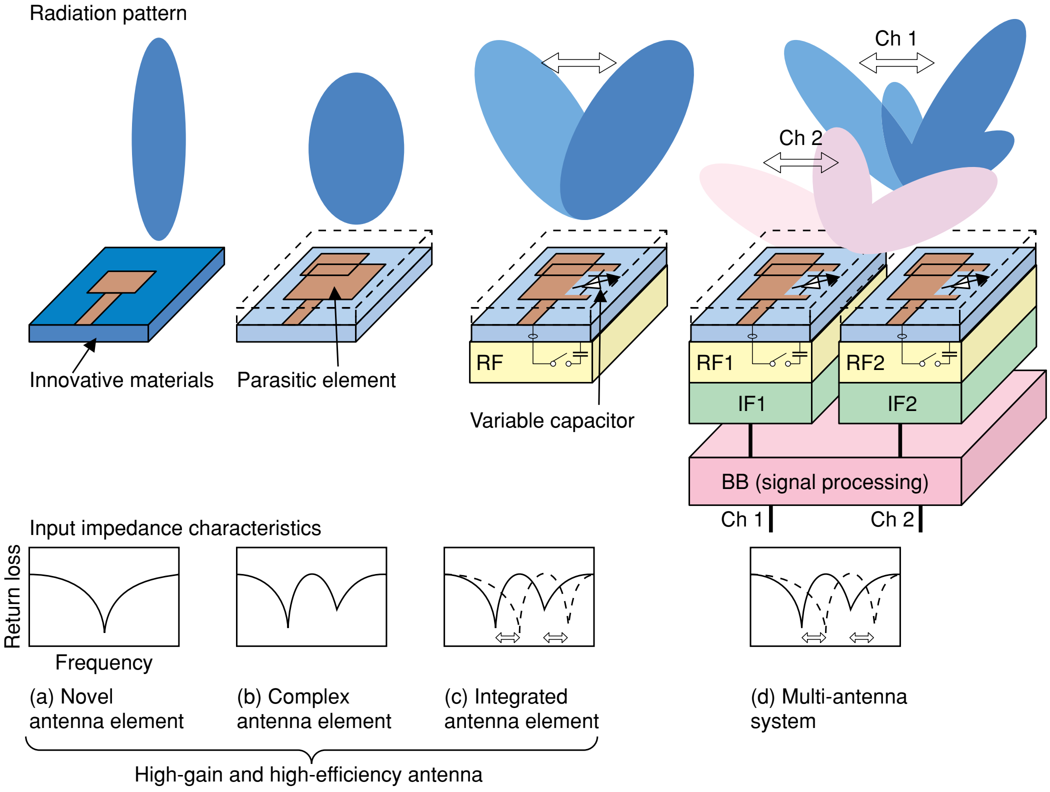
Radiation pattern
Ch 1
Ch 2
RF
RF1
RF2
IF1
IF2
BB (signal processing)
Ch 1
Ch 2
Innovative materials
Parasitic element
Variable capacitor
Input impedance characteristics
Return loss
Frequency
(a) Novel antenna element
(b) Complex antenna element
(c) Integrated antenna element
(d) Multi-antenna system
High-gain and high-efficiency antenna

Fig. 4. Antenna elements and system.

An example of a mobile communication system using a smart antenna in the base station is shown in **Fig. 5**. Independent information can be sent to the terminals, which are spatially separated, using the same frequency at the same time because each beam is generated for a specific terminal by using the space division multiple access (SDMA) technique. Furthermore, N multiplexed transmissions can be sent to a terminal with N antennas by utilizing the slight difference in antenna locations in the terminal. MIMO is representative of this kind of technology. The latest smart antenna systems can increase the antenna gain even in a mobile system and expand the service area. The designed system capacity can be increased by using multiplexing techniques based on baseband signal processing. Multi-antenna systems can achieve very high performance with high-level functions corresponding to the wireless system requirements.

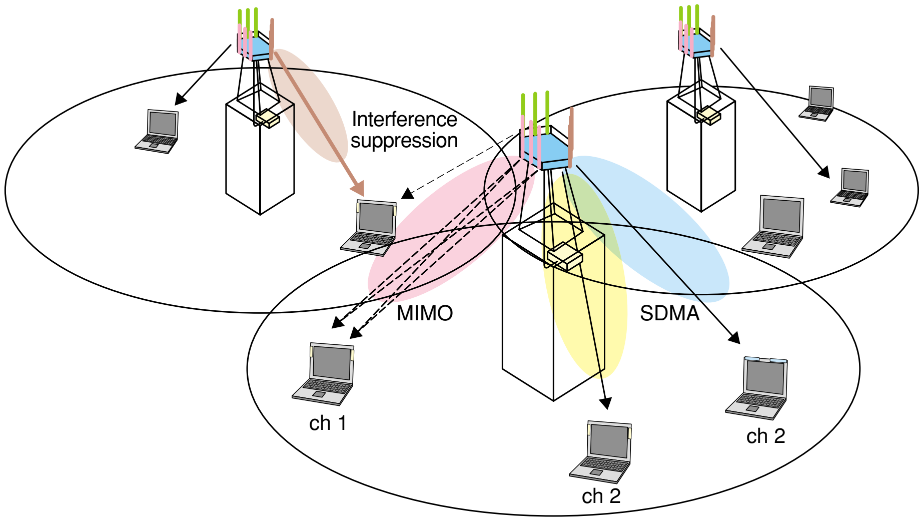
Interference suppression
MIMO
SDMA
ch 1
ch 2
ch 2

Fig. 5. Smart antenna system (land mobile telecommunication).

2.3 Key antenna technologies

The key technologies for achieving a high-gain, high-efficiency, multi-antenna system are shown in **Table 1**. Antennas for the latest wireless communication systems are required to achieve high gain and high efficiency. The multi-antenna systems are designed to make maximum use of system resources, especially space.

3. Wireless systems and suitable antenna systems

The kinds of wireless systems that we can expect to find around us in the near future are illustrated in **Fig. 6**. RF-ID (radio-frequency identification) tags will be attached to various kinds of objects in a ubiquitous network and a large quantity of information will be gathered to provide useful knowledge to each person. High-speed data transmission ranging from a few megabits per second to 1 Gbit/s will be possible in mobile satellite systems, land mobile systems, and wireless local area network (WLAN) systems. In home networks, individual items of electronic equipment will be connected to each other at very high

Table 1. Key technologies to meet antenna requirements.

Requirements	Key technologies
1) High gain and high efficiency	a. Innovative materials (meta-material, high magnetic material, …) b. Element design (parasitic element, fractal element, …) c. Integrated antenna (antenna with RF circuit or RF device, …) d. Active antenna (antenna within variable reactance or switch, ….)
2) Multi-antenna techniques	a. Beam scanning antenna b. Sector antenna (fixed narrow multi-beam system) c. Diversity antenna d. SDMA antenna e. MIMO antenna

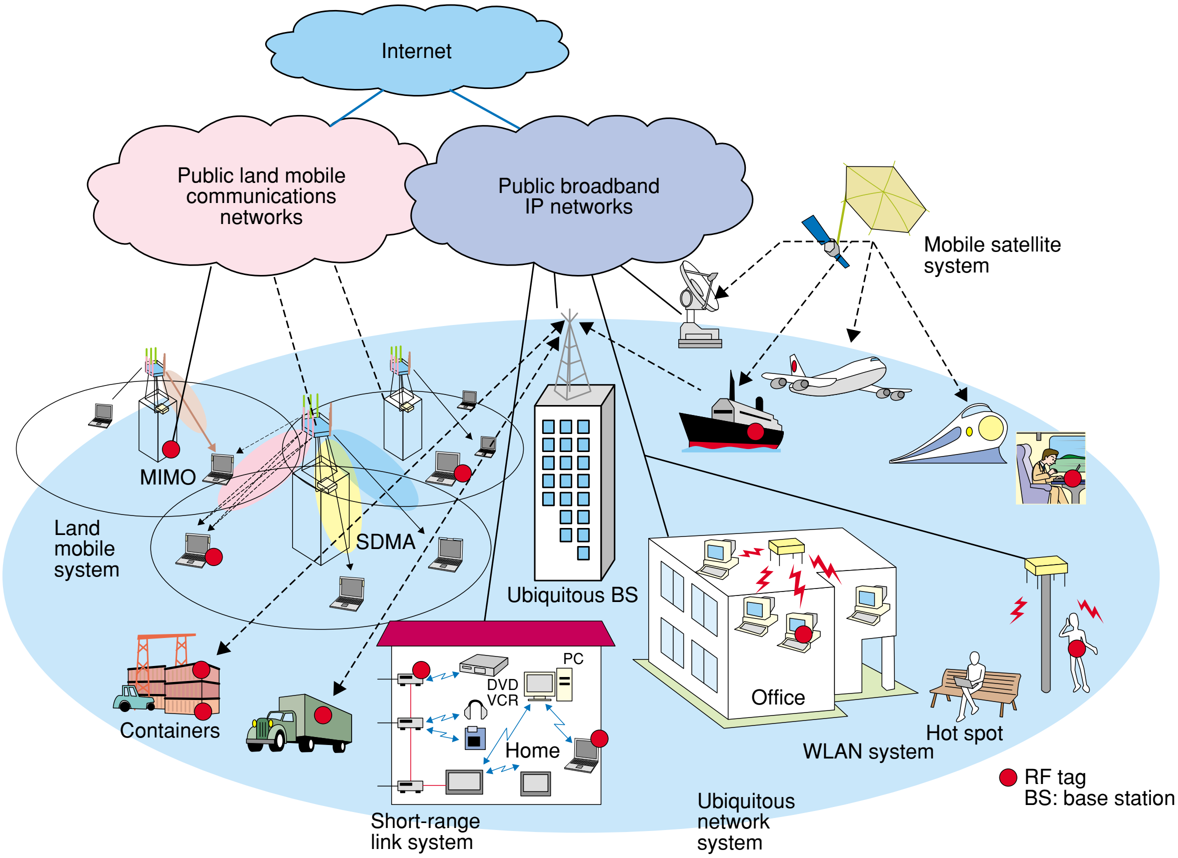
Internet
Public land mobile communications networks
Public broadband IP networks
Mobile satellite system
MIMO
Land mobile system
SDMA
Ubiquitous BS
Containers
PC
DVD
VCR
Home
Office
Hot spot
WLAN system
Short-range link system
Ubiquitous network system
● RF tag
BS: base station

Fig. 6. Wireless communication systems.

Table 2. Wireless systems and NTT's antenna system approaches.

	System features	Main technical trends (example system, standard, or equipment)	NTT's antenna approaches (using technologies in Table 1)
Ubiquitous RF-ID	Multiple access Low rate	Active tags Sensor network Mesh network (802.15.5)	Compact high-efficiency tag antenna High-gain beam tilting base station antenna (1a, 1b, 1c)
Short-range communication	Short range Ultrahigh speed Link transmission	High frequency band (millimeter-wave system) Ultrawideband (UWB) Infrared rays (IrDA)	Active integrated antenna system (millimeter wave) (1b, 1c, 2a, 2b)
Land mobile and WLAN	Medium range High speed/capacity Many subscribers	MIMO-OFDM (802.11n,802.16a/e) MC-CDMA (4G, Beyond 3G) HSDPA (3.5G,1xEV-DO/V)	MIMO-OFDM antenna system Compact mobile terminal antenna system (1b, 1d, 2b, 2c, 2d, 2e)
Mobile satellite	Very long range Medium to high speed Various mobile stations	S/Ku band high-speed mobile comm. (N-Star/ETS-VIII) Broadband (Internet) (WINS) Millimeter wave/optical comm. (MILSTAR, GeoLITE)	Multi-beam antenna Large deployable antenna Beam-tracking antenna system (1c, 1d, 2a, 2e)

(discussed in papers in this issue)

speed (several gigabits per second) wirelessly instead of by cables. In these wireless systems, highly functional and high-performance antennas will be essential. **Table 2** explains the main features and technical trend of the systems and also shows NTT's approaches to antennas and antenna systems for each high-performance wireless system. The technologies in Table 1 that are used are indicated in parentheses. Each antenna system must satisfy severe specifications and also provide high functionality, as shown in Figs. 4(c) and (d). The antennas discussed in the selected papers in this issue are shown in blue in Table 2. The present status of each antenna system and research and development in NTT are explained below.

3.1 Ubiquitous network antenna system

Tag antenna elements have the same problems as pagers or portable telephone systems that use the LF, UHF, and VHF bands. The antenna element must be as small as possible and have high efficiency in a small volume and in the neighborhood of various objects. Most antennas for passive tags using the LF band (13 MHz) are spiral, helical, N-turn loop, or meander line antennas [10] for achieving an electromagnetic induction area and their gains are considered to be very low. Bent and spiral monopole or dipole antennas are used in personal computer card wireless communication systems or in small square film tags using UHF-band systems [11]. The approach is to design the antenna and RF-IC as one system; in other words, the connection impedance between the antenna and RF-IC is designed as a freely selectable value instead of being restricted to 50 Ω. The same kind of technique is currently being investigated in NTT in a study of active integrated antennas [12]. On the other hand, the base station antenna can be designed by applying techniques developed for pagers or land mobile antennas because it has high-gain and beam-tilting characteristics.

3.2 Millimeter-wave frequency band communication antenna system

A high-speed wireless communication system is achieved by using wideband technology, so millimeter-wave frequencies, which provide a very wide bandwidth, are attracting attention. The use of high frequencies also requires high output power to compensate for the huge propagation loss and passive circuit losses, so high-gain and high-efficiency antennas are very important in system design [13]. Millimeter-wave systems mainly use horn, reflector [14], or lens antennas [15]. Horn or reflector antennas are used in relatively long-range communication where the range is several hundred meters. A lens type antenna, which is integrated with monolithic microwave integrated circuits (MMICs) to achieve high efficiency without a feeding line, is mainly used for short-range transmission, e.g., for a home link system. Active integrated antenna technology in which an amplifier for transmitting and receiving is mounted without using a connector or cables is suitable for a high-frequency-band system [16]. Typical antenna types for millimeter-wave frequency systems are shown in **Table 3**. Horn or reflector antennas provide very high gain but cannot be integrated with MMICs. The lens antenna provides high gain and can be made into an integrated antenna but requires adjustment to correct its

Table 3. Antenna types for millimeter-wave systems.

Antenna	Gain	Integration with MMIC	Mounting adjustment
Horn or reflector	Very high (over 20 dBi)	No	(Feeder)
Lens	High (10 to 20 dBi)	Yes	Needed
Single patch	Low (2 dBi)	Yes	No
Parasitic element patch	High (10 dBi)	Yes	No

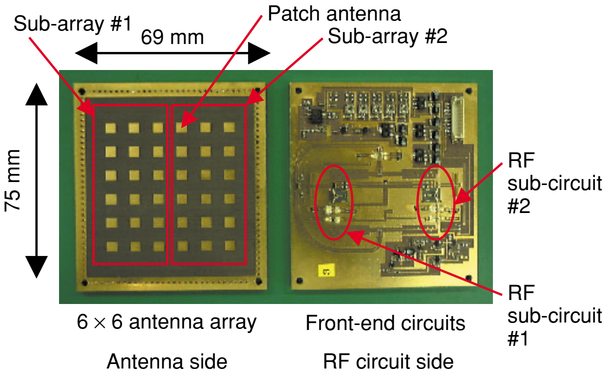

Fig. 7. Active integrated antenna for 25-GHz system.

mounting error. The patch antenna is a good element for integration with MMICs because it can be made on the same material substrates, so it does not need antenna adjustment. We have developed an active integrated antenna designed for broadband mobile wireless access systems using the 25-GHz band, as shown in **Fig. 7** [12]. It exhibited output power of 14.6 dBm and a noise figure of less than 5 dB. The only weakness of this integrated antenna is that high gains cannot be provided with one patch antenna. The third selected paper in this issue describes a high-gain integrated active antenna and high-efficiency RF circuit for the millimeter-wave frequency band for a millimeter-wave communication system [4]. A high-gain integrated patch antenna is investigated as a parasitic element to obtain a wide antenna aperture. The millimeter-wave frequency-band link system was first proposed for application to the transmission of uncompressed HDTV (high-definition television) signals in the home or in a studio.

3.3 Land mobile and WLAN antenna systems

The number of subscribers to land mobile systems, especially cellular systems, has increased rapidly, and users are now demanding high-speed transmission for broadband Internet access. To meets these demands, many improvements to the land mobile system have been proposed such as wideband code division multiple access (W-CDMA), adaptive modulation, and *ad hoc* transmission. Recently, the utilization of space has become as desirable as frequency and time utilization [17]. Space utilization is achieved using a multi-antenna system with a radiation control mechanism, which is accomplished by a smart antenna system.

3.3.1 Typical multi-antenna techniques for mobile systems

The application areas of multi-antenna techniques depend qualitatively on the system characteristics, as shown in **Fig. 8**. A fixed high-gain beam antenna does not always need antenna elements, but the other techniques are achieved through various combinations of antenna elements, so they are provided by adaptive or integrated multi-antenna systems, as shown in Figs. 4(c) and (d). A narrow-beam or high-gain antenna is used in the line-of-sight situation (where there are no obstacles between the transmitting and receiving antennas and there are few paths in the multipath transmission), so a beam-scanning mechanism is needed when a terminal moves fast, and multiple beams or sector beams are required when there are many terminals in the service area. The diversity antenna technique can provide a stable signal level in a multipath mobile environment. Beam scanning (with a sector beam or multiple beams) and diversity antenna techniques have been used fairly extensively from the early days of mobile communication service because they use only simple signal processing. However, with the rapid progress of signal processing technology, a desired signal can now be distinguished from a mixture of many signals in a multipath environment. The SDMA technique distributes signals among the terminals communicating in the service area. MIMO techniques enable very-high-data-rate communication with multiplexed transmission in each antenna of the multi-antenna system of the base station or mobile terminal. These techniques require more calculation as the movement speed of the mobile terminal increases. In these wireless systems, signal processing ability and superior algorithms are more important than the performance of antenna elements.

3.3.2 Research results and developed antenna systems

A very small and thin planar six-sector antenna using a patch Yagi-Uda array with common directors for mobile terminals has been proposed (**Fig. 9**) [18],

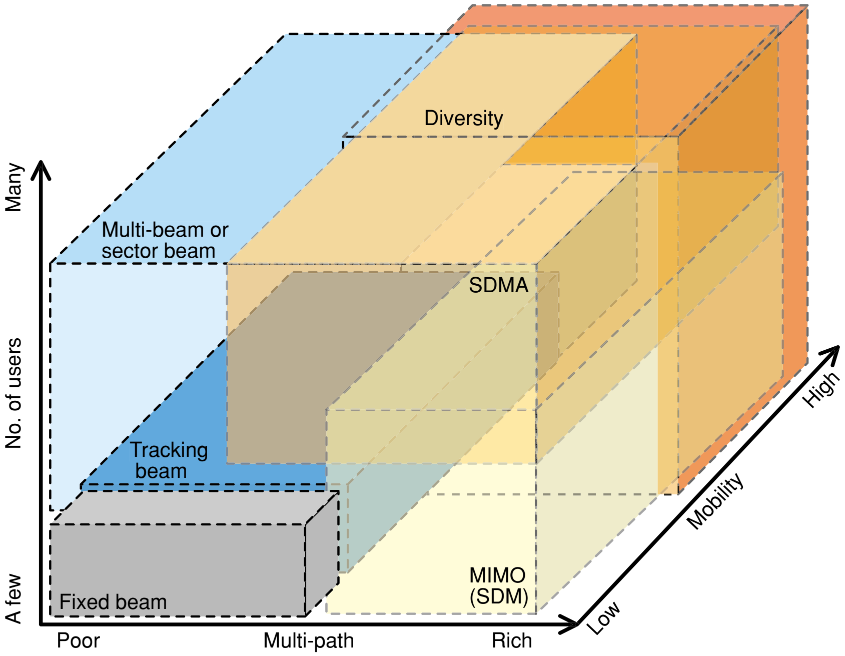
Many
No. of users
A few
Multi-beam or
sector beam
Diversity
SDMA
Tracking
beam
Fixed beam
MIMO
(SDM)
Poor
Multi-path
Rich
Low
Mobility
High

Fig. 8. Dependence of multi-antenna techniques on system characteristics.

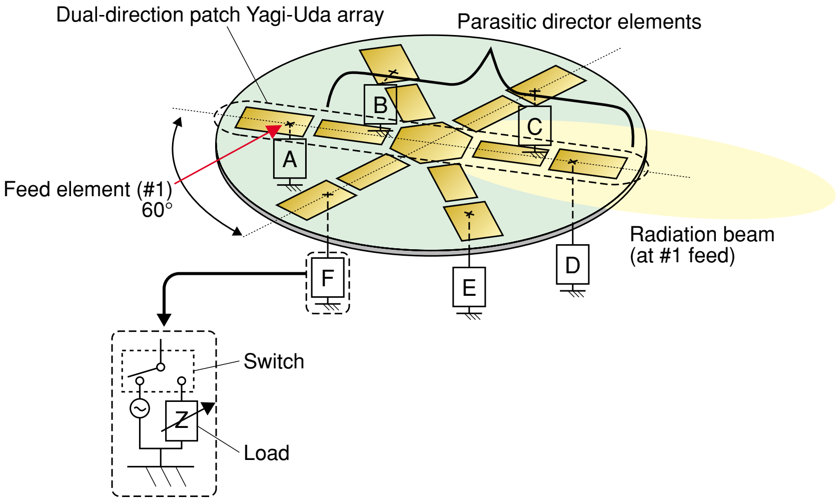
Dual-direction patch Yagi-Uda array
Parasitic director elements
Feed element (#1)
60°
A
B
C
D
E
F
Radiation beam
(at #1 feed)
Switch
Z
Load

Fig. 9. Thin and small sector beam antenna.

[19]. It has square parasitic director elements that are shared by two sector arrays that face in opposite directions, and it has a hexagonal central director element that is shared by all six sectors. These elements form three line arrays, which intersect at 60°. These shared elements enable the sector antenna to be compact. Experimental results showed that the fabricated antenna had a conical plane beam width of 63° and high front-back ratio of 17 dB. The creation of adaptive (smart) antenna hardware has been investigated [20]. A novel beam combination method for wideband digital indoor communication systems such as WLANs has been proposed [21] in NTT. Crossed fan beams are used at the base station and mobile terminal. The beam combination method provides better transmission quality than the traditional pencil-beam combination and eliminates the need for complex beam control in the base station. Maximum dispersion distributions in the transmission frequency bandwidth have been calculated using a three-dimensional indoor propagation delay simulation algorithm. A method of designing adaptive antennas including an

antenna configuration that is suitable for street microcells, considering the propagation environment, has been described. An automatic calibration method at the transmitters and receivers using the transmitting signal (ACT), which enables realtime calibration, has been presented [22]. An SDMA technique using polarization has also been studied [23]. The transmission quality of an SDMA scheme that uses vertical pattern and polarization control was investigated in an actual cellular environment. For a sector cell system, this configuration was shown to be effective based on an evaluation of the spatial correlation characteristics. We also found that the output carrier-to-interference-plus-noise power ratio (CINR) of the proposed configuration was from 10 to 17 dB higher than that of the conventional SDMA configuration when the number of users was greater than four. The proposed configuration can reduce the antenna size compared with conventional SDMA. Moreover, MIMO techniques enable very-high-data-rate transmission using a limited frequency band. We have studied technologies based on the eigenmode-space division multiplexing (E-SDM) [24] scheme because it provides the theoretically maximum channel capacity and reduces the terminal's calculation load. The selected papers in this issue introduce both theoretical and practical technologies. The fourth paper describes new MIMO methods with beamforming to overcome problems of calculation complexity and low MIMO effect in the line-of-sight scenario [5]. The fifth paper describes the configuration of an 8 × 4 MIMO-OFDM transceiver and presents experimental results for the propagation characteristics, bit error rate (BER), and frequency utilization [6].

3.4 Satellite broadband communication system

If we want to create a seamless and flexible wireless network system, then a satellite communication system is very important because it provides a wide service area. However, it has the inherent problem that the distance between the transmitter and receiver is long, so the terminal needs a larger antenna than ones in the land mobile system to provide high-speed transmission. The effective solution is to make the satellite dish larger to provide high gain and multibeam transmission. We have studied multibeam communication satellite technology and onboard large deployable antennas for creating a broadband mobile satellite communication system [25]. An MMIC-based beamforming network equipment for a multibeam feeder whose weight is only 1/3 that of the conventional type was developed. We are also developing a large, highly precise deployable antenna with a diameter in the over-10-m range. The mobile terminal acting as an earth station is also an important component of a broadband mobile satellite communication system. Typical existing antennas in earth stations are shown in **Table 4**. Various mobile terminals are accommodated in the satellite communication system by different transmission rates, and these terminals use various antenna methods and locations. A comparatively large terminal antenna is required to handle high data transmission rates in a satellite system to mitigate the huge propagation loss. The high gain and narrow beam of a large antenna make a beam-scanning mechanism necessary to enable the antenna to follow the satellite direction as the terminal moves. The sixth selected paper describes a highly accurate and cost-effective auto-tracking antenna system for earth stations onboard vessels, using a novel inclinometer that is highly accurate in acceleration disturbance environments and a new systematic stabilization controller design process based on H_∞ control [3].

Table 4. Typical antennas in earth stations.

Satellite category	Typical satellite system	Current typical earth station antenna (size or length, gain)
LEO	Iridium Globalstar	Helical (0.1 m, 2 dBi) Linear whip (0.1 m, 0 dBi) Patch (0.1 m^2, 3 dBi)
MEO GEO	ICO N-STAR INMARSAT	Parabola (0.6 – 0.9 m, 18 – 21 dBi) Phased patch array (0.5 m^2, 14 dBi) Helical (0.3 m, 2 dBi) Patch array (0.3 m^2, 12 dBi) Patch (0.1 m^2, 3 dBi)

LEO: low earth orbit
MEO: medium earth orbit
GEO: geostationary earth orbit

4. Conclusion

This paper gave an overview of the requirements of antennas and antenna systems in recent high-speed and high-capacity wireless communication systems and the key technologies for satisfying these requirements to create next-generation wireless communications. It also described current antenna systems and NTT's antenna research results for typical wireless systems.

Antennas for next-generation wireless communications require high gain, high efficiency, and a multi-antenna design. A high-gain antenna can be made using a relatively large antenna aperture, but a beam-scanning mechanism is needed for mobile communications. A high-efficiency antenna can be made using an active integrated antenna, which integrates an MMIC with the antenna element. Multi-antenna technology is the innovative technology that enables multiple data transmission in space. In MIMO-type multi-antenna systems, the most important key technologies are the beamforming algorithm and the implementation method. The solution technologies for NTT's advanced antenna systems are explained in detail in the following selected papers in this issue. The technological background and the current research and development situation are also described to make the papers easy to understand.

References

[1] "Next-Generation Broadband Wireless Access Technologies," NTT Technical Review, Vol. 2, No. 1, pp. 12-63, 2004.

[2] N. Honma, T. Seki, and K. Tsunekawa, "Planar Beam-Scanning Microstrip Antenna Using Tunable Reactance Devices for Satellite Communication Mobile Terminal," NTT Technical Review, Vol. 3, No. 9, pp. 24-32, 2005.

[3] T. Yoshida, K. Ohata, and M. Ueba, "Highly Accurate and Cost-effective Auto-tracking Antenna System for Satellite Broadband Communication on Vessels," NTT Technical Review, Vol. 3, No. 9, pp. 60-71, 2005.

[4] T. Seki, K. Nishikawa, I. Toyoda, and K. Tsunekawa, "Millimeter-wave High-Efficiency Multilayer Parasitic Microstrip Antenna Array for System-on-package," NTT Technical Review, Vol. 3, No. 9, pp. 33-41, 2005.

[5] Y. Takatori, R. Kudo, K. Nishimori, and K. Tsunekawa, "Beamforming Method for Broadband MIMO-OFDM Systems," NTT Technical Review, Vol. 3, No. 9, pp. 42-49, 2005.

[6] K. Nishimori, R. Kudo, Y. Takatori, and K. Tsunekawa, "Experimental Evaluation of 8×4 Eigenmode SDM Transmission in Broadband MIMO-OFDM Systems," NTT Technical Review, Vol. 3, No. 9, pp. 50-59, 2005.

[7] J. D. Kraus, "Antennas, Sec. 2: Basic Antenna Concepts," McGraw-Hill Book Company, 1988.

[8] R. C. Hansen, "Fundamental limitations in antennas," Proceedings of the IEEE, Vol. 69, Issue 2, Feb. 1981.

[9] A. Alexiou and M. Haardt, "Communications smart antenna technologies for future wireless systems: trends and challenges," Magazine, IEEE, Vol. 42, Issue 9, pp. 90-97, Sep. 2004.

[10] S. S. Basat, K. Lim, J. Laskar, and M. M. Tentzeris, "Design and Modeling of Embedded 13.56 MHz Rfid Antennas," IEEE AP-S Int. Antennas Propagat. Symp. Dig.,Vol. 4, Session 130, pp. 64-67, 2005.

[11] U. Karthaus and M. Fischer, "Fully integrated passive UHF RFID transponder IC with 16.7-/spl mu/W minimum RF input power," Solid-State Circuits, IEEE Journal of Vol. 38, Issue 10, pp. 1602-1608, Oct. 2003.

[12] T. Seki, F. Nuno, K. Nishikawa, and K. Cho, "Active Integrated Antenna Technique," NTT Technical Review, Vol. 2, No. 1, pp. 36-43, 2004.

[13] T. Ihara and K. Fujimura, "Research and development of millimeter-wave short-range application systems," Trans., IEICE, Vol. E79-B, No. 12, pp. 1741-1753, Dec. 1996.

[14] W. Menzel, D. Pilz, and M. Al-Tikriti, "Millimeter-wave folded reflector antennas with high gain, low loss, and low profile," IEEE Antennas and Propagation magazine, pp. 24-29, June 2002.

[15] M. J. M. van der Vorst, P. J. I. de Maagt, A. Neto, R. M. Heeres, W. Luinge, and M. H. A. J. Herben, "Effect of internal reflections on the radiation properties and input impedance of integrated lens antennas-comparison between theory and measurements," Microwave Theory and Techniques, IEEE Trans. on 49, Issue 6, Part 1, pp. 1118-1125, June 2001.

[16] T. Seki, K. Uehara, and K. Kagoshima, "A Three-Dimensional Active Antenna for a High-Speed Wireless Communication Application," IEEE MTT-S Int. Microwave Symp., Denver, USA, Vol. 2, pp. 975-978, June 1997.

[17] G. J. Foschini, G. D. Golden, R. A. Valenzuela, and P. W. Wolniansky, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," IEEE Journal on Vol. 17, Issue 11, pp. 1841-1852, Nov. 1999.

[18] N. Honma, F. Kira, T. Maruyama, K. Cho, and H. Mizuno, "Compact six-sector antenna employing patch Yagi-Uda array with common director," Antennas and Propagation Society International Symposium 2002, IEEE, Vol. 1, pp. 26-29, June 2002.

[19] N. Honma, T. Seki, K. Tsunekawa, F. Kira, and K. Cho, "Compact Multi-Sector Antenna Employing Microstrip Yagi-Uda Array Antenna with Common Director Elements," IEICE Trans. B, Vol. J87-B, No. 9, pp. 1363-1371, Sep. 2004 (in Japanese).

[20] T. Hori, K. Cho, Y. Takatori, and K. Nishimori, "Key techniques for actualizing smart antenna hardware," Antennas and Propagation Society International Symposium 2000, IEEE, Vol. 2, pp. 948-952, July 2000.

[21] K. Uehara, Y. Takatori, and K. Kagoshima, "Proposal for a novel beam combination method for indoor high-speed wireless communication systems," Vehicular Technology, IEEE Transactions on Vol. 48, Issue 5, pp. 1511-1517, Sep. 1999.

[22] K. Nishimori, K. Cho, Y. Takatori, and T. Hori, "A Novel Configuration for Realizing Automatic Calibration of Adaptive Array Using Dispersed SPDT Switches for TDD Systems," IEICE Vol. E84-B, No. 9, pp. 2516-2522, Sep. 2001.

[23] K. Nishimori, and K. Cho, "Evaluation of SDMA employing vertical pattern and polarization control in actual cellular environment measurement," Vehicular Technology Conference 2004, VTC 2004-Spring, 2004 IEEE 59th Vol. 1, pp. 244-248, May 2004.

[24] A. Paulraj, R. Nabar, and D. Gore, "Introduction to space-time wireless communications," Cambridge, U.K., Cambridge Univ. Press, 2003.

[25] M. Ueba, K. Ohata, J. Mitsugi, and M. Umehira, "Broadband and scalable mobile satellite communication system for future access networks," 22nd AIAA ICSSC, AIAA 2004-3154.

Koichi Tsunekawa

Senior Manager, Executive Research Engineer, Wireless Systems Innovation Laboratory, NTT Network Innovation Laboratories.

He received the B.S., M.S., and Ph.D. degrees in engineering science from the University of Tsukuba, Tsukuba, Ibaraki in 1981, 1983, and 1992, respectively. He joined the Electrical Communications Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT), Tokyo in 1983. Since 1984, he has been engaged in R&D of portable telephone antennas in land mobile communication systems. From 1993 to 2003, he was with NTT DoCoMo, Inc. where he worked on radio propagation research, intelligent antenna systems for wireless communications, and the development of IMT-2000 antenna systems. His research interests are antenna systems for MIMO transmission, millimeter-wave broadband access, and ubiquitous wireless system. He is a member of IEEE and IEICE.

Communications and Network, 2016, 8, 9-21
Published Online February 2016 in SciRes. http://www.scirp.org/journal/cn
http://dx.doi.org/10.4236/cn.2016.81002

The Roles of 5G Mobile Broadband in the Development of IoT, Big Data, Cloud and SDN

Bao-Shuh Paul Lin[1,2], Fuchun Joseph Lin[1,2], Li-Ping Tung[1]

[1]Microelectronics & Information Research Center, National Chiao Tung University, Taiwan
[2]College of Computer Science, National Chiao Tung University, Taiwan
Email: bplin@mail.nctu.edu.tw, fjlin@nctu.edu.tw, lptung@nctu.edu.tw

Received 3 November 2015; accepted 22 February 2016; published 25 February 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

The fast technology development of 5G mobile broadband (5G), Internet of Things (IoT), Big Data Analytics (Big Data), Cloud Computing (Cloud) and Software Defined Networks (SDN) has made those technologies one after another and created strong interdependence among one another. For example, IoT applications that generate small data with large volume and fast velocity will need 5G with characteristics of high data rate and low latency to transmit such data faster and cheaper. On the other hand, those data also need Cloud to process and to store and furthermore, SDN to provide scalable network infrastructure to transport this large volume of data in an optimal way. This article explores the technical relationships among the development of IoT, Big Data, Cloud, and SDN in the coming 5G era and illustrates several ongoing programs and applications at National Chiao Tung University that are based on the converging of those technologies.

Keywords

5G, Internet of Things (IoT), Software Defined Networks (SDN), Big Data Analytics, Cloud Computing

1. Introduction

The roles of SDN, Cloud, IoT, and Big Data in 5G Networks have raised great interest recently [1]. According to the IEEE Computer Society (IEEE CS 2022 Report [2]), Cloud Computing (Cloud), Big Data Analytics (Big Data), Internet of Things (IoT), and Software Defined Networks (SDN) are among 4 of 20+ emerging technologies as illustrated in **Table 1**. In the meantime, based on the report in IEEE Communications Society (IEEE ComSoc Technology News [3]), among top 10 trends in 2015 as listed in **Table 2**, 5G, Virtualization (SDN & NFV), everywhere connectivity for IoT & IoE, and Big Data are also included. Combining both reports, we can identify 5G, Cloud, IoT/IoE, Big Data, and SDN as the five most worthwhile ICTs (information & communica-

How to cite this paper: Lin, B.-S.P., Lin, F.J. and Tung, L.-P. (2016) The Roles of 5G Mobile Broadband in the Development of IoT, Big Data, Cloud and SDN. *Communications and Network*, **8**, 9-21. http://dx.doi.org/10.4236/cn.2016.81002

Table 1. IEEE computer society 2022 report.

Technology & Indexing Terms		
1. Security Cross-Cutting Issues	9. Multicore	17. 3D Printing
2. Open Intellectual Property Movement	10. Photonics	18. Big Data and Analytics
3. Sustainability	11. Networking and Interconnectivity	19. Machine Learning and Intelligent Systems
4. Massively Online Open Courses	12. Software Defined Networks	20. Life Sciences
5. Quantum Computing	13. High Performance Computing	21. Computational Biology and Bioinformatics
6. Device and Nanotechnology	14. Cloud Computing	22. Robotics
7. 3D Integrated Circuits	15. Internet of Things	
8. Universal Memory	16. Natural User Interfaces	

Table 2. IEEE communications society top 10 trends 2015.

1) 5G
2) FIBER EVERYWHERE
3) VIRTUALIZATION, SDN & NFV
4) EVERYWHERE CONNECTIVITY FOR IoT & IoE
5) BIG DATA, COGNITIVE NETWORKS
6) CYBERSECURITY
7) GREEN COMMUNICATIONS
8) SMARTER SMARTPHONES, CONNECTED SENSORS
9) NETWORK NEUTRALITY, INTERNET GOVERNANCE
10) MOLECULAR COMMUNICATIONS

tions technologies) to watch out up to 2020 in term of their potential, convergence and applications. **Figure 1** depicts the overlapping of these five major ICT trends. It also illustrates how the infrastructure deployment of 5G mobile broadband and the architectural integration of Cloud Computing strongly impact the development of IoT, Big Data, and SDN. This paper explores the technical relationships of these five areas and discusses several ongoing programs and applications currently under development at NCTU based on these technologies.

2. The Roles of 5G, IoT, Big Data, Cloud, and SDN till 2020

Although so far the 5G mobile broadband requirements and standard specifications are not ready yet, 5G technology research & development are already started and some 5G features or subsystems are readily available. By Year 2020 the commercial 5G will be available and IoT applications will be deployed everywhere with mobile broadband technology. Moreover, the Big Data generated by IoT applications will become a norm and Cloud will be largely utilized to compute, store and virtualize network functions (NFV). Also, the underlying network infrastructure will adopt SDN to reduce both capital expense (CAPEX) and operational expense (OPEX). **Figure 2** further illustrates the roles of 5G, IoT, Big Data, Cloud and SDN and their relationships. This figure is modified based on the reference [4] by inserting 5G mobile broadband in the center.

3. Technical Relationships among IoT, Big Data, Cloud, & SDN in 5G Era

Based on **Figure 2**, we develop **Figure 3** to better explain the technical relationships among IoT, Big Data, Cloud, and SDN in the 5G mobile broadband services (5G MBS) [5]. First, IoT is capable of generating Big Data with four Vs: volume, velocity, variety and veracity. Then, Cloud is brought in for Big Data storage and processing. Finally, SDN is employed to provide more efficient and flexible networks for inter-Cloud data transport. Out of Big Data, Cloud, and SDN, advanced technologies such as machine learning analytics, Cloud

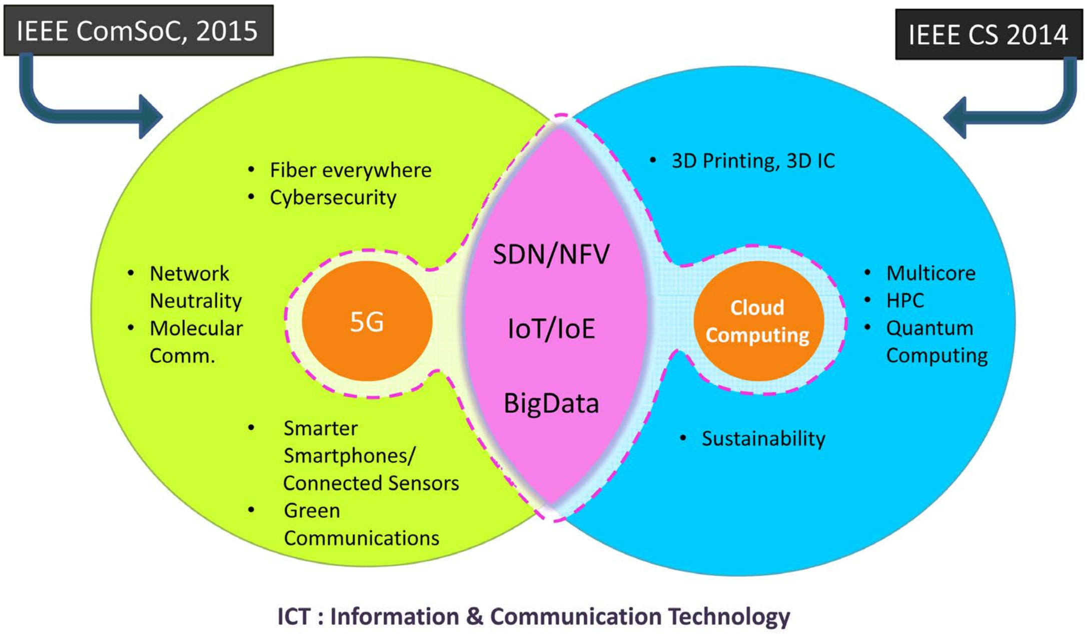
IEEE ComSoC, 2015
IEEE CS 2014
Fiber everywhere
Cybersecurity
Network Neutrality
Molecular Comm.
5G
Smarter Smartphones/ Connected Sensors
Green Communications
SDN/NFV
IoT/IoE
BigData
3D Printing, 3D IC
Multicore
HPC
Quantum Computing
Cloud Computing
Sustainability
ICT : Information & Communication Technology

Figure 1. ICT major trends for 2015-2020.

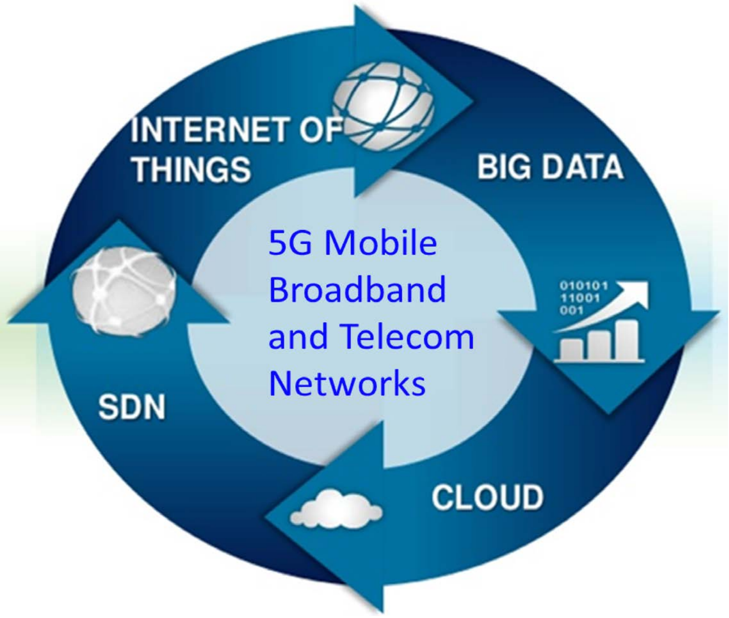
INTERNET OF THINGS
BIG DATA
5G Mobile Broadband and Telecom Networks
010101
11001
001
SDN
CLOUD

Figure 2. Relationship among 5G, IoT, Big Data, Cloud and SDN based on [4].

RAN and softwarized 5G then are developed.

As illustrated in **Figure 4**, 5G will serve as a better gateway and transport network for IoT applications so that IoT data can be delivered more efficiently and economically. In addition, IoT will become one of the major sources of Big Data by producing large volume, fast velocity, and many varieties of data [6] as illustrated in **Figure 5**.

Finally, **Figure 6** shows that Cloud can be adopted in the 5G Radio Access Network (RAN) and turns it to a Cloud-based RAN (C-RAN). Both SDN and NFV have been applied to data center in the cloud to enable better load balance and resource allocation of the cloud. SDN has also been applied to 5G mobile broadband core networks to enable smart routing, better traffic management and improve network resource utilization.

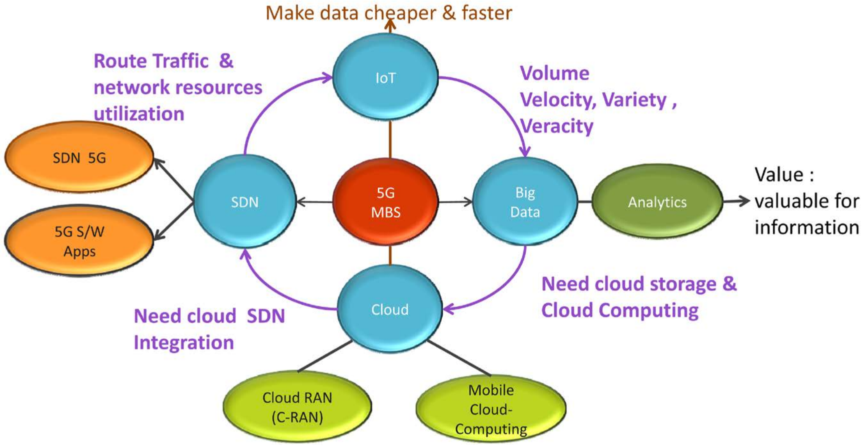
Make data cheaper & faster
Route Traffic & network resources utilization
IoT
Volume Velocity, Variety , Veracity
SDN 5G
5G S/W Apps
SDN
5G MBS
Big Data
Analytics
Value : valuable for information
Need cloud storage & Cloud Computing
Cloud
Need cloud SDN Integration
Cloud RAN (C-RAN)
Mobile Cloud-Computing

Figure 3. Technical relationships among 5G & IoT, Big Data, Cloud, and SDN.

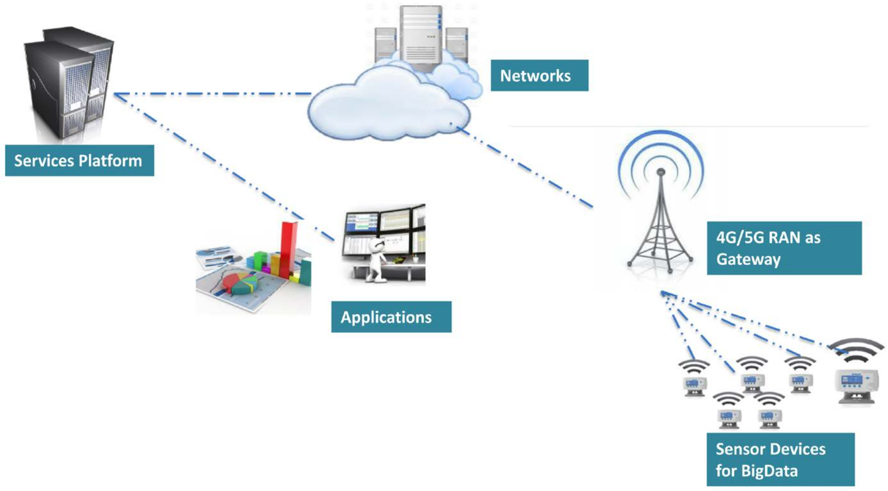
Services Platform
Networks
4G/5G RAN as Gateway
Applications
Sensor Devices for BigData

Figure 4. 4G/5G as the gateway for IoT applications.

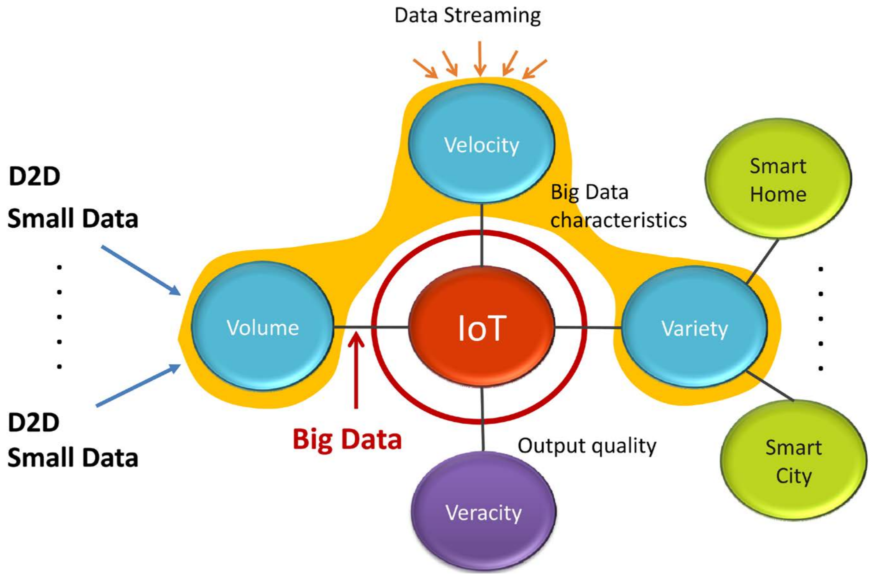
Data Streaming
Velocity
D2D
Small Data
Big Data characteristics
Smart Home
Volume
IoT
Variety
D2D
Small Data
Big Data
Output quality
Veracity
Smart City

Figure 5. Relationships between IoT and Big Data.

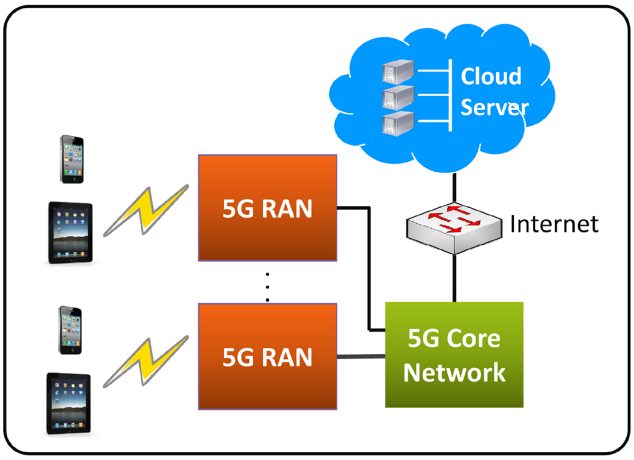

Figure 6. C-RAN architecture in 5G mobile network.

4. Ongoing Programs and Applications at National Chiao Tung University (NCTU)

In this section, we describe five ongoing programs and applications at NCTU that are closely related to the SDN, Cloud, Big Data, IoT, and 5G technologies [7]-[17].

4.1. Program on SDN-Enabled Cloud-Based Broadband and Wireless Network Technologies and Services

This program is a national program to support the SDN Industry-Academia Cooperation, led by both NCTU and CHT (Chunghwa Telecom, the largest mobile & telecom operator in Taiwan) with 5 other networking and communications companies in Taiwan. This is a multi-year program with the target to set up a testbed for end-to-end testing and to help establish an ecosystem for local SDN industry.

Figure 7 shows the scope of this program that covers mobile access, WiFi access, broadband core and data center cloud end-to-end application systems.

Figure 8 further shows both the structure of the program and its network configuration across NCTU, CHT and another university National Tsing Hwa University (NTHU).The program includes research and development involving 4G/LTE, B4G/5G, SDN, Cloud, SDN for Wi-Fi, and SDN for WAN technologies. There are five sub-programs: 1) SDN broadband network technologies and services, 2) SDN mobile/wireless network technologies and services, 3) SDN and cloud integration services and management, 4) SDN switch-related devices and systems and 5) SDN system integration and field trials.

The program is intended to construct an SDN network in NCTU, CHT and NTHU, respectively. These SDN networks then will be interconnected to form a wide area SDN network in Taiwan. Eventually, we plan to have this end-to-end system connected to the global network. The SDN switches deployed in this testbed will be mostly small or medium scale due to their experimental nature. Also included in this testbed are SDN-based Wi-Fi access points. Cloud-based data centers will be deployed in each location of NCTU, CHT and NTHU. Three types of SDN controllers including OpenDaylight, Ryu and Floodlight are currently under trial in the experimental network. Our architecture assumes the scenario of multiple SDN controllers with a hierarchical topology on which useful SDN APPs such as load balancer, network optimizer for video delivery, network co-ordinator (including visualization), dynamic flow configuration, end-to-end service configuration, dynamic policy-based traffic engineering, and multitenant network automation can be developed and deployed.

4.2. Program on Big Data Analytics for Network Traffic and Management Data

In this program, we address the network performance issues with two experimental networks: BML at NCTU campus and ITRINET of ITRI.

The architecture of the BML experimental network is illustrated in the lower part of **Figure 9** that consists of a 4G RAN, a 4G Core and a Cloud environment. For the 4G RAN, both indoor and outdoor environment are taken into consideration. The left upper part of **Figure 9** shows the scope of the ITRINET experimental network

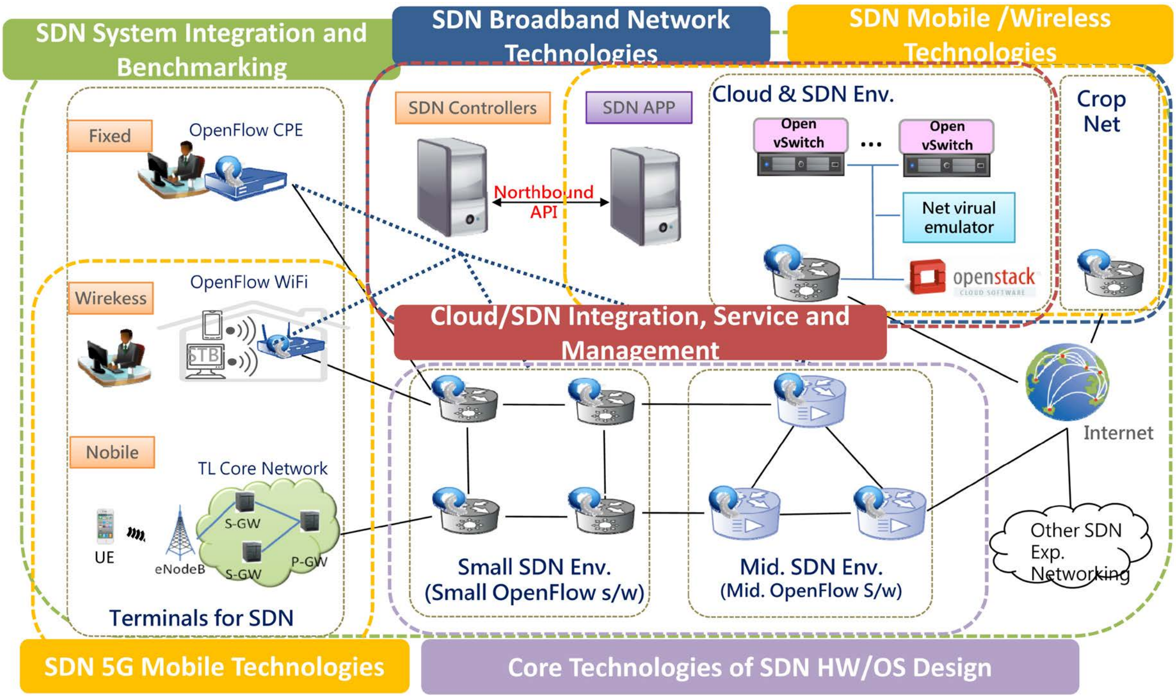
SDN System Integration and Benchmarking
SDN Broadband Network Technologies
SDN Mobile /Wireless Technologies
Fixed
OpenFlow CPE
SDN Controllers
SDN APP
Northbound API
Cloud & SDN Env.
Open vSwitch
Open vSwitch
Net virual emulator
openstack
Crop Net
Wirekess
OpenFlow WiFi
STB
Cloud/SDN Integration, Service and Management
Nobile
TL Core Network
S-GW
S-GW
P-GW
UE
eNodeB
Terminals for SDN
Small SDN Env. (Small OpenFlow s/w)
Mid. SDN Env. (Mid. OpenFlow S/w)
Internet
Other SDN Exp. Networking
SDN 5G Mobile Technologies
Core Technologies of SDN HW/OS Design

Figure 7. Applications of SDN & cloud in 4G/LTE & 5G wireless networks.

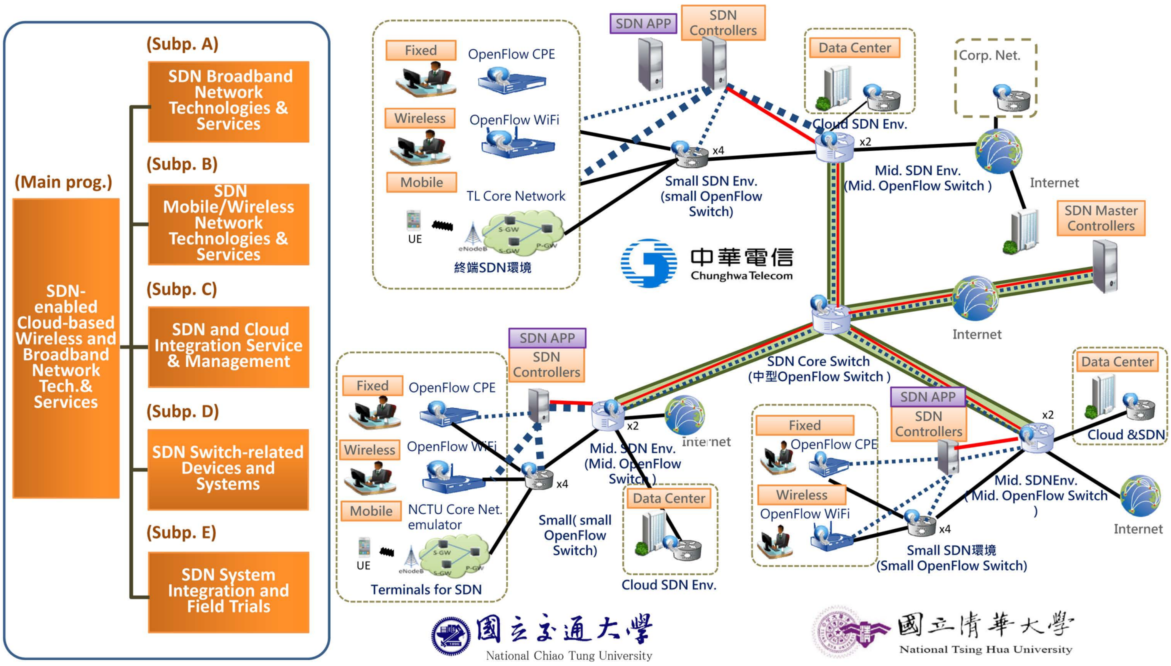
(Main prog.)
SDN-enabled Cloud-based Wireless and Broadband Network Tech.& Services
(Subp. A)
SDN Broadband Network Technologies & Services
(Subp. B)
SDN Mobile/Wireless Network Technologies & Services
(Subp. C)
SDN and Cloud Integration Service & Management
(Subp. D)
SDN Switch-related Devices and Systems
(Subp. E)
SDN System Integration and Field Trials
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
Mobile
TL Core Network
UE
eNodeB
S-GW
S-GW
P-GW
終端SDN環境
SDN APP
SDN Controllers
Data Center
Corp. Net.
Cloud SDN Env.
x2
x4
Small SDN Env. (small OpenFlow Switch)
Mid. SDN Env. (Mid. OpenFlow Switch)
Internet
SDN Master Controllers
中華電信
Chunghwa Telecom
Internet
SDN Core Switch (中型OpenFlow Switch)
SDN APP
SDN Controllers
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
Mobile
NCTU Core Net. emulator
UE
eNodeB
S-GW
S-GW
P-GW
Terminals for SDN
x2
x4
internet
Mid. SDN Env. (Mid. OpenFlow Switch)
Small(small OpenFlow Switch)
Data Center
Cloud SDN Env.
Fixed
OpenFlow CPE
Wireless
OpenFlow WiFi
SDN APP
SDN Controllers
Data Center
Cloud &SDN
x2
x4
Mid. SDNEnv. (Mid. OpenFlow Switch)
Internet
Small SDN環境 (Small OpenFlow Switch)
國立交通大學
National Chiao Tung University
國立清華大學
National Tsing Hua University

Figure 8. SDN technical R&D center and projects at NCTU.

that covers a much greater area of Hsinchu County than NCTU and includes additionally ITRI (Industrial Technologies Research Institute), HSIP (Hsinchu Science Industrial Park) and THR (Taiwan High-speed Rail, Hsinchu station). The right upper part of **Figure 9** illustrates how ITRINET covers various R&D buildings in the corporate campus of ITRI such as B11, B12, B51, etc.

Figure 10 illustrates how Big Data analytics based on InfoSphere or Spark is performed on ITRINET for the purpose of network optimization. First, network traffic measurement and network management data are collected.

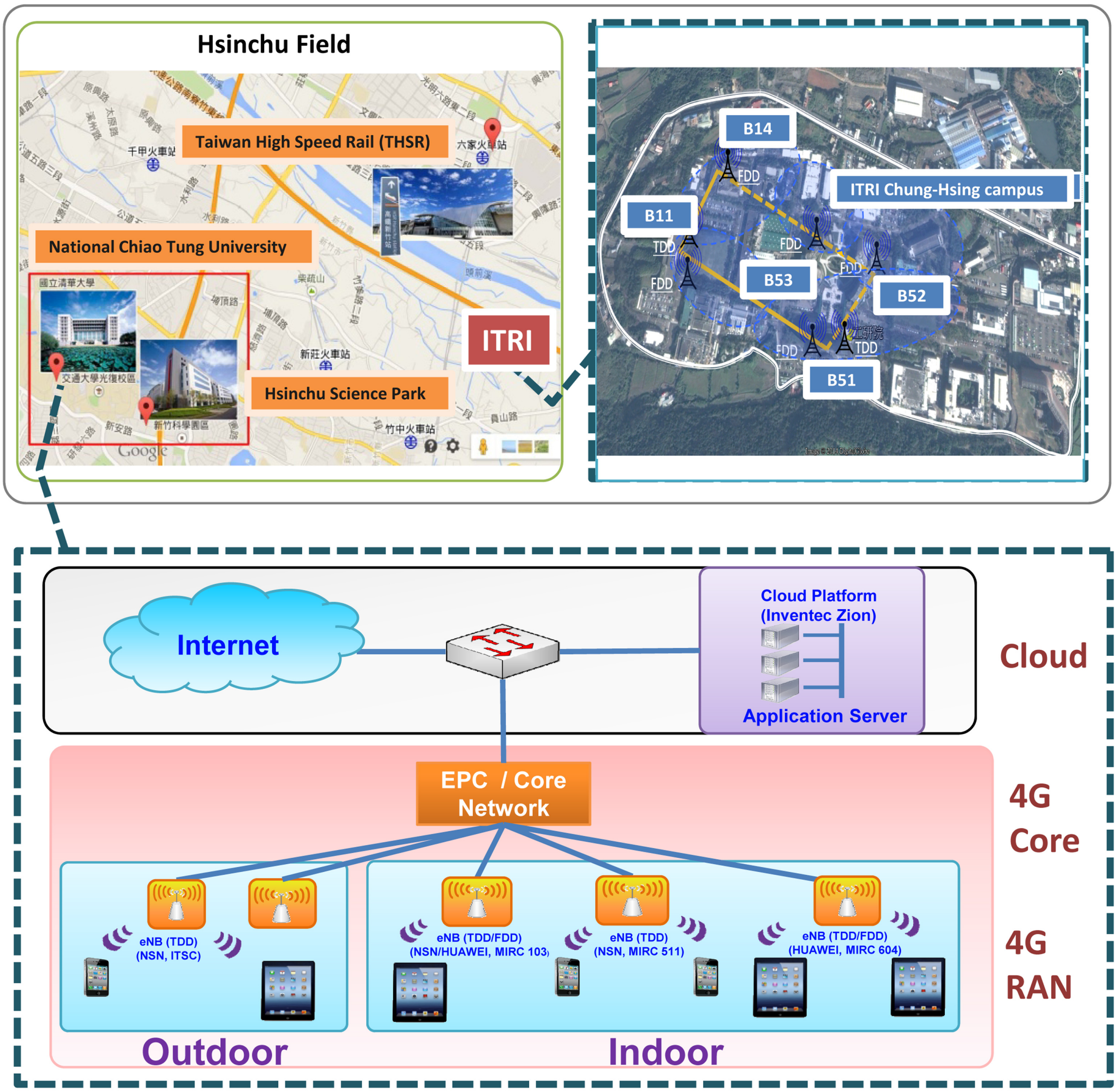
Hsinchu Field
Taiwan High Speed Rail (THSR)
National Chiao Tung University
ITRI
Hsinchu Science Park
國立清華大學
交通大學光復校區
新竹科學園區
千甲火車站
六家火車站
新莊火車站
竹中火車站
Google
B14
B11
B53
B52
B51
ITRI Chung-Hsing campus
FDD
TDD
Internet
Cloud Platform
(Inventec Zion)
Application Server
Cloud
EPC / Core
Network
4G
Core
eNB (TDD)
(NSN, ITSC)
eNB (TDD/FDD)
(NSN/HUAWEI, MIRC 103)
eNB (TDD)
(NSN, MIRC 511)
eNB (TDD/FDD)
(HUAWEI, MIRC 604)
4G
RAN
Outdoor
Indoor

Figure 9. Greater Hsinchu 4G/LTE & Future 5G experimental network ITRINET (include ITRI, NCTU, THR & HSIP).

Then, data analytics methods based on machine learning, data mining and statistical modeling are applied to analyze the collected data. Finally, we apply the results thus generated to network performance evaluation and optimization by providing a feedback loop for system re-configuration. The whole operational cycle includes the technologies of 4G/LTE, B4G/5G, Big Data Analytics, Cloud, and Traffic Engineering.

4.3. Application on IoT Platform Integrated with Data Generation and Data Analytics

This is an application where we set up an IoT platform integrated with data generator and data analytics capabilities as illustrated in **Figure 11**. A common challenge for IoT/M2M service providers is how to test their large scale IoT/M2M applications with the near realistic data that the system will handle in a production environment. As such tests may involve not only a large number but also a large variety of sensors, deploying a testing environment that contains all the necessary sensors turns out to be an infeasible, if not impossible job. To tackle this problem, we develop a data generation method (illustrated in the lower part of **Figure 11**) based on streams generation capabilities of IBM InfoSphere and Spark to emulate data from a large number and a large variety of sensors. Such generated data will be sent into and processed by the applications residing on the IoT/M2M

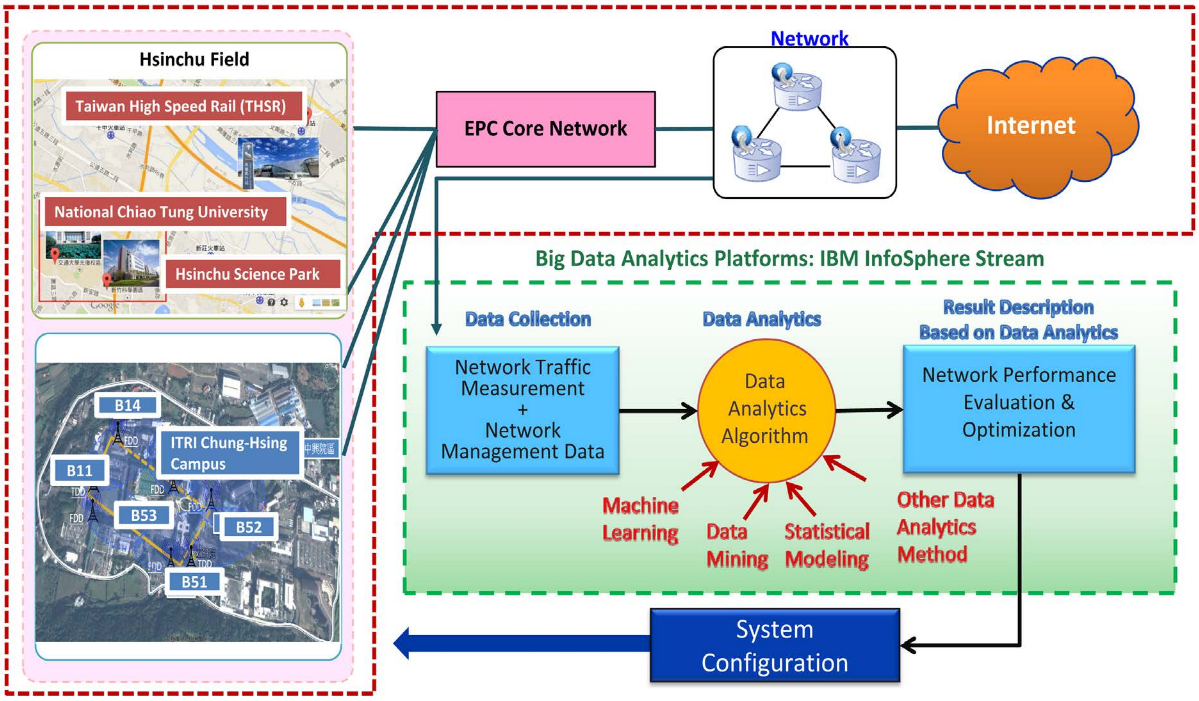
Hsinchu Field
Taiwan High Speed Rail (THSR)
National Chiao Tung University
Hsinchu Science Park
B14
ITRI Chung-Hsing Campus
B11
B53
B52
B51
EPC Core Network
Network
Internet
Big Data Analytics Platforms: IBM InfoSphere Stream
Data Collection
Network Traffic Measurement + Network Management Data
Data Analytics
Data Analytics Algorithm
Result Description Based on Data Analytics
Network Performance Evaluation & Optimization
Machine Learning
Data Mining
Statistical Modeling
Other Data Analytics Method
System Configuration

Figure 10. Big Data analytics for Traffic engineering of ITRINET.

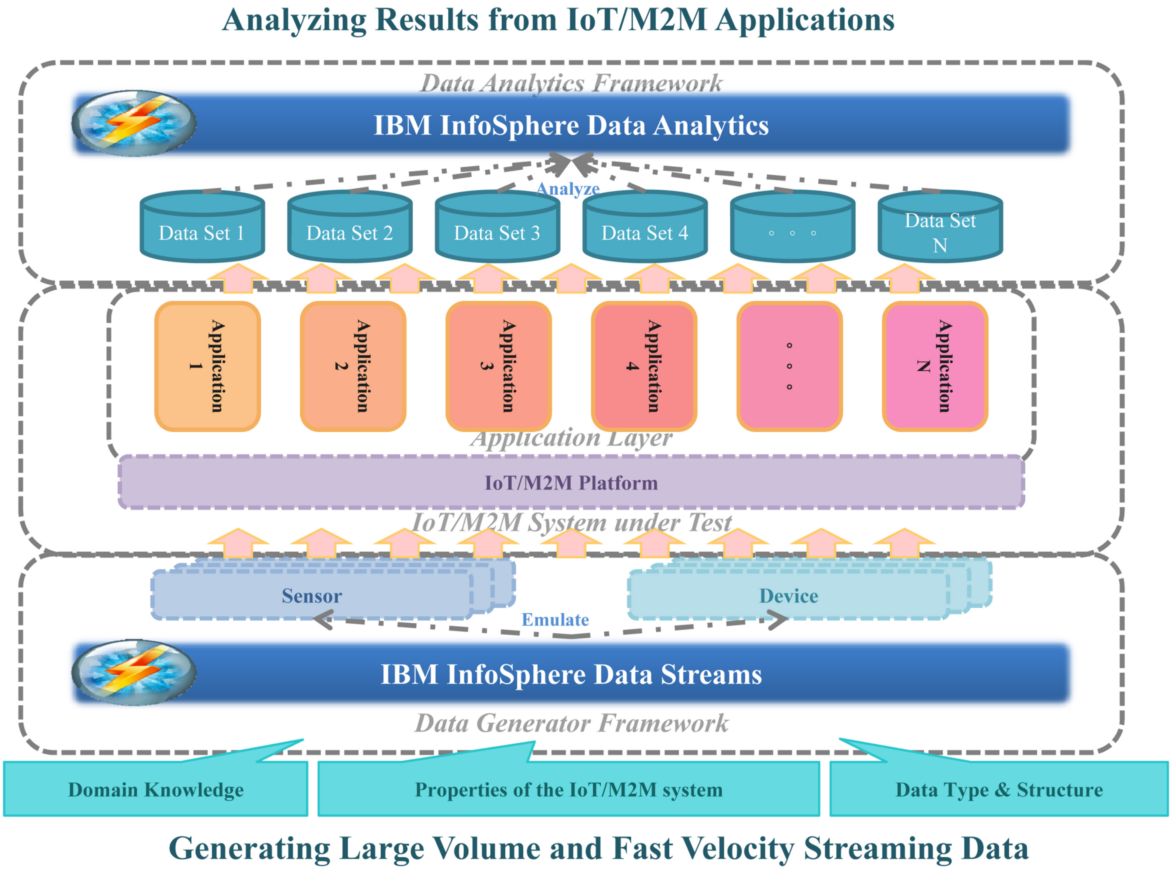
Analyzing Results from IoT/M2M Applications
Data Analytics Framework
IBM InfoSphere Data Analytics
Analyze
Data Set 1
Data Set 2
Data Set 3
Data Set 4
。。。
Data Set N
Application 1
Application 2
Application 3
Application 4
。。。
Application N
Application Layer
IoT/M2M Platform
IoT/M2M System under Test
Sensor
Device
Emulate
IBM InfoSphere Data Streams
Data Generator Framework
Domain Knowledge
Properties of the IoT/M2M system
Data Type & Structure
Generating Large Volume and Fast Velocity Streaming Data

Figure 11. IoT platform integrated with data generation and data analytics.

platform (illustrated in the middle part of **Figure 11**). Finally, the data sets produced by the IoT/M2M applications are analyzed by data analytics engines in IBM InfoSphere and Spark (illustrated in the upper part of **Figure 11**).

The effectiveness of our data generation methodology is first verified in a people management system of a large factory environment. With our methodology, we are able to test our factory management IoT/M2M application more efficiently without actually deploying a large number and a large variety of sensors. This application development uses both IoT and Big Data Analytics.

4.4. Application on Data Analytics for Traffic Generated by Testing & Measurement

The automatic Testing & Measurement (T&M) of DUT (Device under Test, including 4G/LTE SD, UE/CPE) for 4-Stage (namely, conformance, interoperability, operator-IOT and field trials) testing is shown in **Figure 12**.

We apply the data analytics learning techniques for the classification of the DUT types by series numbers or id. Once the type is identified, the parameters for feature extraction can be measured by the testing and measurement equipment. An example for DUT classification is shown in **Table 3**. In this example, the DUT type relational table is a 4 × 4 table with 4 fields: 1) DUT Device Type, 2) UE capability Information (FGI), 3) DNS (TTL), and 4) HTTP (User Agent). This device relational table is created by applying multivariate statistical analytics on the network traffic flow and LTE UE connection related information. Then by extracting specific features, we are able to identify the types of DUT such as Dongle, Smartphone, CPE, etc.

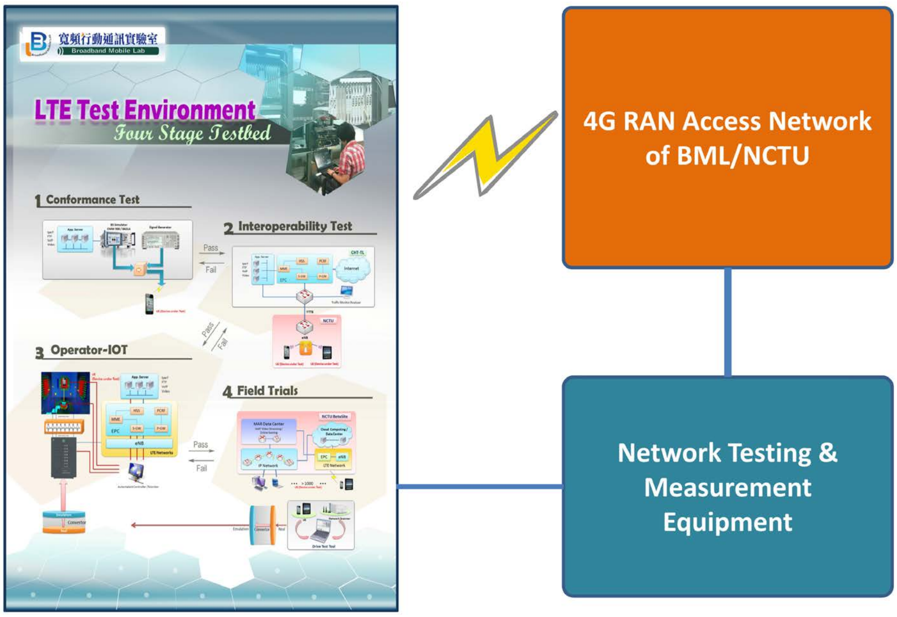

Figure 12. Data generation by endpoints from testing & measurement.

Table 3. DUT relational table.

DUT Device Type	UE FGI	DNS-TTL	HTTP-User Agent
Dongle A	0100 0110 0000 0101 0001 1000 0000 0000	128	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n
Dongle B	0101 1111 0000 1111 1111 1100 1000 0000	63	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n
Smart Phone	0101 1110 0000 1101 1101 1000 1000 0000	43	iPhone6,2/8.3 (12F70)\r\n
CPE	0101 1111 0000 1111 1111 1100 1000 0000	127	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.134 Safari/537.36\r\n

Another example of applying data analytics in T&M traffic is to evaluate LTE QoS. When evaluating LTE QoS, it is required to setup testing scenarios with different QoS levels. Traffic Flow Template (TFT) (as shown in **Table 4**) is designed to filter packets into correspondent bearers, either a default bearer or dedicated bearer, in LTE network. Each bearer has its own QoS level. We can establish a dedicated bearer of guarantee bit rate (GBR) for applications such as VoIP, or a default bearer of basic QoS level for applications like file transfer.

What TFT does is to filter packets into correspondent bearers according to packets' IP address, port number, protocol, direction. The information, however, is insufficient to differentiate among web file download, FB news browsing, and Line chat. We are unable to give different QoS levels to these applications. To tackle this problem, we propose a new architecture which integrates deep packet inspection (DPI) with TFT to provide a higher granularity of QoS levels for applications. When some traffic flow has been identified as a certain application type, TFT will be informed to update rules of packet filters. With this improved method, the traffic flow can be delivered through the bearer of suitable QoS level.

4.5. Applications on Data Analytics for Traffic Flow Created by APP's of SD

To fulfill the Quality of Service (QoS) requirements from users, it is important to make effective use of the network resource. We can optimize the performance of a network by applying data analytics to traffic engineering. In particular, it is important to classify mobile applications traffic intended by the user with data analytics.

We propose a HMM-based (Hidden Markov Model) model to classify the mobile applications. By surveying related work, we have realized that there are different handshake patterns of well-known application protocols. Also, according to the observation of some specific mobile applications, we discover that every mobile Internet service has its unique negotiation process at the beginning when service starts. HMM was widely used to recognize time-dependent sequences and find out unknown patterns of data, such as speech recognition, handwriting recognition.

Table 4. Traffic Flow Template (TFT): Use to specify the packet filters associated.

UL Packet Filter ID	DRB	Packet Filter Evaluation Precedence	Protocol Number (IPv4)/ Next Header (IPv6)	Remote Address and Subnet Mask	Single Local Port (UE)	Local Port Range (UE)	Single Remote Port Range (NW)	Remote Port Range (NW)	IPSec SPI Range	Type of Service (IPv4)/ Traffic Class (IPv6) and Mask	Flow Label (IPv6)
1	DRB2	6	6 (TCP)	IPv4: 172.168.8.0 [255.255.255.0] IPv6: 2001:0ba0::[ffff:ffff::]	60051	-	-	-	-	-	-
2	DRB2	2	17 (UDP)	-	-	-	60201	-	-	-	-
3	DRB3	7	6 (TCP)	IPv4: 172.168.8.0 [255.255.255.0] IPv6: 2001:0ba0::[ffff:ffff::]	-	60100:60200	-	-	-	-	-
4	DRB3	3	17 (UDP)	-	-	-	-	60300: 60400	-	-	-
5	DRB3	5	50 IPSec (ESP)	-	-	-	-	-	0x0F80F0000	-	-
6	DRB3	1	-	-	-	-	-	-	-	00101000, Mask = 11111100	-
7	DRB3	6	-	-	-	-	-	-	-	-	-
8	DRB1 (default bearer)	255	6 (TCP)	-	-	-	-	-	-	-	-

In our method, we extract the packet size sequence and packet transmission direction sequence of the first 20 packets to train the HMM model. **Figure 13** shows our designed model structure of HMM. X_n is the hidden variable. It represents the transmission states which cannot be observed directly. Because of the unknown transmission states, we need to use observation variable O_n^1 and O_n^2 as training features to build the mobile application models. Here, "*n*" is the sequence number of packets in a traffic flow. O_n^1 is the observation symbol of packets transmission direction. O_n^2 is the observation of packet size that quantizes to a certain range of packet size. We quantize smallest packet size to number 1 and largest packet size to number 8. The rest of packet sizes are then divided into six groups.

Figure 14 shows the main process of our classification system. First, we process the collected packets by reading from original PCAP files, and extract the necessary field of packet header, including source IP, destination IP, source port number, destination port number, packet timestamp, and packet size. Second, we use packet

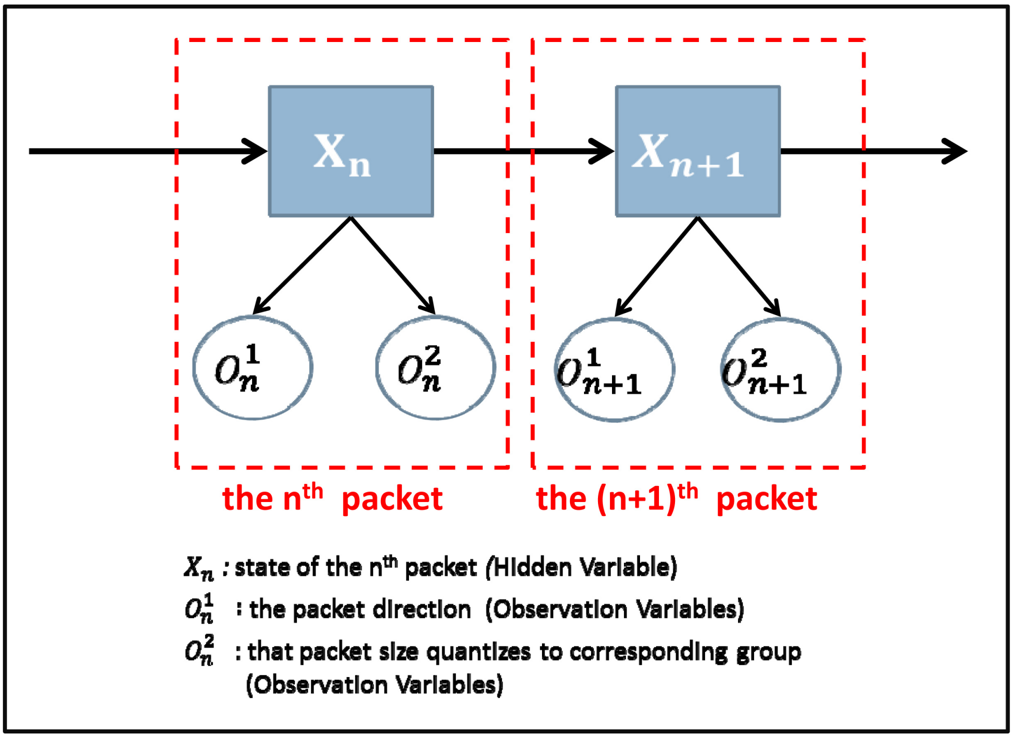
X_n
X_{n+1}
O_n^1
O_n^2
O_{n+1}^1
O_{n+1}^2
the n^{th} packet
the $(n+1)^{th}$ packet
X_n : state of the n^{th} packet (Hidden Variable)
O_n^1 : the packet direction (Observation Variables)
O_n^2 : that packet size quantizes to corresponding group (Observation Variables)

Figure 13. Classification model of Internet application traffic.

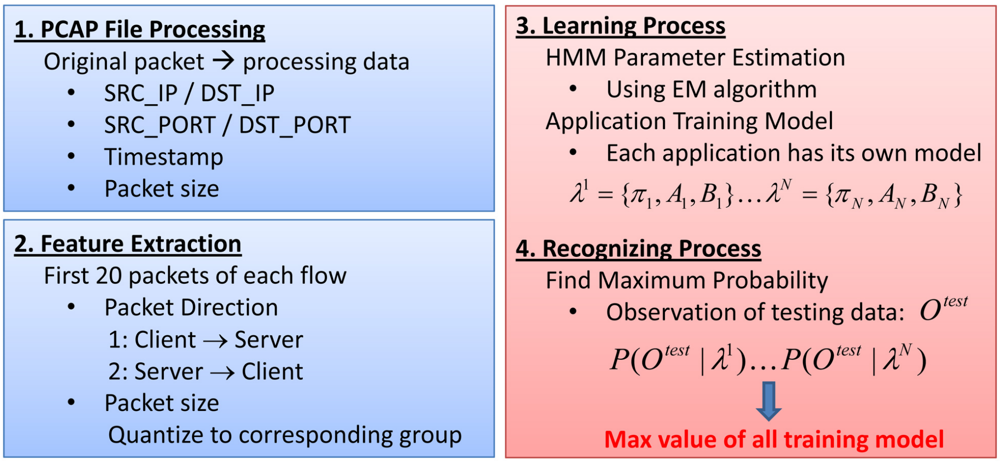

1. PCAP File Processing

Original packet → processing data

- SRC_IP / DST_IP
- SRC_PORT / DST_PORT
- Timestamp
- Packet size

2. Feature Extraction

First 20 packets of each flow

- Packet Direction
  1: Client → Server
  2: Server → Client
- Packet size
  Quantize to corresponding group

3. Learning Process

HMM Parameter Estimation

- Using EM algorithm

Application Training Model

- Each application has its own model

$$\lambda^1 = \{\pi_1, A_1, B_1\} \ldots \lambda^N = \{\pi_N, A_N, B_N\}$$

4. Recognizing Process

Find Maximum Probability

- Observation of testing data: O^{test}

$$P(O^{test} \mid \lambda^1) \ldots P(O^{test} \mid \lambda^N)$$

$$\Downarrow$$

Max value of all training model

Figure 14. The process of our proposed classification system.

header information (5-tuple: source IP, destination IP, source port number, destination port number, and protocol) to process packets with the same 5-tuple information into a unit of mobile application traffic flow, and extract the features of each application flow. Third, in learning process, we use the extracted features which are quantized into corresponding symbol sequence to training the HMM-based application models for different applications. Finally, we can identify the new traffic flow by finding out the maximum value of log likelihoods derived from different application models.

5. Conclusion

This paper presents a forward looking view of the convergence of IoT, big data, cloud, SDN technologies along with the arrival of 5G mobile broadband networks. We intend to demonstrate the technical relationships of those technologies and the compelling programs and applications that can be created when two, three or more of those technologies converge. Due to the nature of fast evolution of ICT and the ongoing innovation of those five technologies, this paper should be updated on annual basis to keep the related information up to day with the ICT major trends.

Acknowledgements

This work was supported by the Ministry of Science and Technology (MOST) of Taiwan and National Chiao Tung University under grants: MOST 103-2622-E-009-012-, 103-2221-E-009-138-, 103-2218-E-009-032-, 104-2221-E-009-023- and NCTU-ICTL 104Q707.

References

[1] Lin, B-S.P. (2015) The Roles of SDN, Cloud, IoT, & Big Data, in 5G Networks and Their Technical Relationships & Applications. Invited Talk, IEEE VTS APWCS 2015, Singapore.

[2] Alkhatib, H., *et al*. (2015) IEEE CS 2022 Report. IEEE Computer Society, IEEE.

[3] Neira, E.M. (2015) IEEE ComSoc CTN Special Issue on Ten Trends That Tell Where Communication Technologies are Headed in 2015.

[4] Tucker, L. (2013) On the Cloud of the Future. Cloud Computing Summit, San Jose.

[5] Kleiner, T. (2015) 5G Research in Horizon 2020.

[6] Thadani, A. and Andreoli, P. (2014) Realizing Business Value from Convergence of IoT + Big Data Technologies. 2014 *Midwest Architecture Community Collaboration* (*MACC*) *Conference*, Minneapolis, 13 November 2014.

[7] Dinh, H.T., Lee, C., Niyato, D. and Wang, P. (2013) A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches. *Wireless Communications and Mobile Computing*, **13**, 1587-1611. http://dx.doi.org/10.1002/wcm.1203

[8] Lin, B-S.P., *et al.* (2013) The Design of Cloud-based 4G/LTE for Mobile Augmented Reality with Smart Mobile Device. *Proceedings of* 2013 *IEEE International Symposium on Mobile Cloud, Computing, and Service Engineering* (*IEEE Mobile Cloud* 2013), Redwood City, 25-28 March 2013, 561-566. http://dx.doi.org/10.1109/sose.2013.57

[9] Lin, B.-S.P., Tung, L-P., Hsieh, I-C., Liu, T-H. and Chou, S-Y. (2015) The Design of Big Data Analytics for Testing & Measurement and Traffic Flow on an Experimental 4G/LTE Network. *Proceedings of IEEE WOCC* 2015, Taipei, 23-24 October 2015, 40-44. http://dx.doi.org/10.1109/wocc.2015.7346173

[10] Lin, B-S.P., Tung, L-P., Hsieh, I-C. and Chou, S-Y. (2015) Performance Estimation of MAR for Outdoor Navigation Applications on an 5G Mobile Broadband by Mobile Smart Devices. *Proceedings of IEEE APWCS* 2015, Singapore, 19-20 August 2015, 1-6.

[11] Lin, F.J., Ren, Y. and Cerritos, E. (2013) A Feasibility Study on Developing IoT/M2M Applications over ETSI M2M Architecture. *Proceedings of the* 1st *International Workshop on Internet of Things Technologies* (*IoTT*), Seoul, 15-18 December 2013, 558-563. http://dx.doi.org/10.1109/icpads.2013.100

[12] Chen, H.W. and Lin, F.J. (2014) Converging MQTT Resources in ETSI Standards Based M2M Platform. *Proceedings of IEEE iThings*, Taipei, 1-3 September 2014, 292-295. http://dx.doi.org/10.1109/ithings.2014.52

[13] Cerritos, E. and Lin, F.J. (2014) M2M-Enabled Real-Time Trip Planner. *Proceedings of the* 2nd *International Workshop on Internet of Things Technologies* (*IoTT*), Hsinchu, 16-19 December 2014, 886-891. http://dx.doi.org/10.1109/padsw.2014.7097902

[14] Lusiarta Putera, C.A. and Lin, F.J. (2015) Incorporating OMA Lightweight M2M Protocol in IoT/M2M Standard Architecture. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 559-564.

http://dx.doi.org/10.1109/WF-IoT.2015.7389115

[15] Adrianto, D. and Lin, F.J. (2015) Analysis of Security Protocols and Corresponding Cipher Suites in ETSI M2M Standards. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 777-782. http://dx.doi.org/10.1109/wf-iot.2015.7389152

[16] Lin, F.J. and Chen, H. (2015) Improving Utilization and Customer Satisfaction of Parking Space with M2M Communications. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 465-470.

[17] Lin, F.J., Tsai, W., Cerritos, E., Lin, B.-T. and Hu, W.-H. (2015) Identification and Analysis of Charging Factors in M2M Communications. *Proceedings of IEEE World Forum on IoT*, Milan, 14-16 December 2015, 160-165. http://dx.doi.org/10.1109/wf-iot.2015.7389045

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 6309509, 20 pages
https://doi.org/10.1155/2018/6309509

Research Article

A Model of Socially Connected Web Objects for IoT Applications

Sajjad Ali,[1] **Muhammad Golam Kibria**,[1] **Muhammad Aslam Jarwar**,[1]
Hoon Ki Lee,[2] **and Ilyoung Chong**[1]

[1] *Department of Information and Communications Engineering, Hankuk University of Foreign Studies, Yongin-si, Republic of Korea*
[2] *Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea*

Correspondence should be addressed to Ilyoung Chong; iychong@hufs.ac.kr

Received 2 September 2017; Revised 26 November 2017; Accepted 19 December 2017; Published 28 January 2018

Academic Editor: Nathalie Mitton

Copyright © 2018 Sajjad Ali et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Internet of Things (IoT) is evolving with the connected objects at an unprecedented rate, bringing about enormous opportunities for the future IoT applications as well as challenges. One of the major challenges is to handle the complexity generated by the interconnection of billions of objects. However, Social Internet of Things (SIoT), emerging from the conglomeration of IoT and social networks, has realized an efficient way to facilitate the development of complex future IoT applications. Nevertheless, to fully utilize the benefits of SIoT, a platform that can provide efficient services using social relations among heterogeneous objects is highly required. The web objects enabled IoT environment promotes SIoT features by enabling virtualization using virtual objects and supporting the modularity with microservices. To realize SIoT services, this article proposes an architecture that provides a foundation for the development of lightweight microservices based on socially connected web objects. To efficiently discover web objects and reduce the complexity of service provisioning processes, a social relationship model is presented. To realize the interoperable service operations, a semantic ontology model has been developed. Finally, to evaluate the proposed design, a prototype has been implemented based on a use case scenario.

1. Introduction

The Internet of Things (IoT) envisions billions of objects connected to the Internet that continuously generate data about the physical environment. Several state-of-the-art applications can be built using the information and services provided by the pervasive and heterogeneous IoT objects. However, building applications based on these objects raises some challenges. Examples include how the growing number of objects will interact or coordinate to render valuable services, how the complexity generated by the coordination of objects can be handled, and how a large amount of diverse data sensed from the surroundings can be managed properly.

Numerous promising solutions have been provided to handle the complexity and the heterogeneity of IoT objects. One of the prominent approaches is to develop virtual environments to resolve the issues caused by the heterogeneous objects. Virtualization has become a key component in many IoT architectures, in the form of either virtual objects or virtual entities. A virtual object (VO) is defined as a digital representation of a real-world object [1] and provides a way for the IoT services to discover and compose features that cannot be achieved directly with a real-world object (RWO). It has been realized that issues such as heterogeneity of dissimilar objects and scalability can be effectively resolved using VOs. Hence, a crucial role of a VO is to bridge the gap between the physical world and the virtual world by hiding the physical characteristics of an IoT device and acquiring and interpreting its data and context.

Another way to reduce the complexity caused by a large number of objects is to link these objects in a social network [2], which provides a more exact solution than using a sole individual object. This principle has given birth to the concept of Social Internet of Things (SIoT), which is emerging as a new paradigm with the merger of social networks and IoT [3–7]. It has been realized as an efficient way to facilitate the development of complex future IoT applications. It supports sharing of information generated by the people and the devices based on the social relationships. These relationships further promote efficient discovery of the objects and effective service compositions.

Let us have a glimpse of the application scenarios in which SIoT will make a profound effect on our daily life in the near future. Imagine a group of vehicles that form a mobile social network on the move when heading towards the same location or when they are bound in a common relationship. Vehicles may form diverse social communities; for example, small cars create a social network to share available parking space information or bicycles share information about the vacant paths using a network [8]. Vehicles commuting from home to office can join a social network to share traffic congestion experience, accident warnings, or other common information (e.g., a meeting plan) with colleagues by accessing their vehicles' social network. Let us analyze another SIoT use case where a user enters a public museum with an IoT environment, which incorporates smart objects that exploit social relationships to share information. At the entrance to the premises, the user's smartphone is prompted with a beacon signal to recommend a service app installation. When the user grants the permission, an app is installed and a social agent acquires his profile and friendship details. The app incorporates a microservice which uses a smartphone object to establish a social relationship with the museum visit service. The museum is equipped with a network of smart objects that monitor the user's location and provide services to enhance his/her experience such as personalizing the displays, assisting in the navigation, and recommending the relevant services.

To support the above-mentioned use cases, most of the current IoT architectural approaches are not suitable as they are based on the traditional monolithic approaches, which further hinder the development of efficient, modular, and independent services that cannot scale well with the increasing user's service demands, as already witnessed in many studies [9–12]. Using monolithic approaches, we cannot fully achieve the benefits offered by SIoT including efficient information discovery, improved scalability, and simplified interconnection of objects.

These limitations motivate the current paper to define a design for the development of SIoT based services with efficient and lightweight mechanisms to exploit SIoT features for improved service provisioning. The proposed design is based on the microservices concept which promises a more solid practice of SOA. There is no particular definition of microservices architectural style; however, it is explained in [13] by Martin Fowler as "an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API." Following the microservices based proposed design, SIoT applications can be developed, deployed, and maintained more efficiently and independently, which will promote system modularity and interoperability.

Moreover, a SIoT design needs to facilitate the efficient information discovery based on the social relationships among objects. To achieve this, incorporating the existing social relationship models in the proposed design can be useful. However, applications with intelligent service features require dynamic object selection. Therefore, objects need to acquire the ability to infer new relationships for interconnection with other objects in the system. This view further motivates the present article to develop a social relationship model based on the semantic ontology. The utilized ontology-based approach is highly useful to dynamically infer new social relationships for efficient service provisioning.

Furthermore, to fully achieve the benefits of both, the SIoT and microservices, selecting an IoT environment is necessary where not only can the social interconnection among heterogeneous objects be utilized but also the lightweight services can be developed to apply these relationships for effective IoT service provisioning. To fulfill this goal, the Web of Objects (WoO) [14] platform has been selected. According to the ITU recommendation (ITU-T Y.4452) [15], "the Web of Objects (WoO) is a realization way of the IoT services, where virtualized objects (i.e., virtual objects (VOs) and composite virtual objects (CVOs)) are connected, controlled, and incorporated with resources to facilitate the development, deployment, and operation of IoT services on the World Wide Web."

To exploit the opportunities provided by the above discussed technological advancements, the main contributions are highlighted as follows:

(i) This article contributes a novel design based on the microservices concept with lightweight and modular services to support the development of SIoT applications. The proposed design involves mechanisms to enhance object discovery and to reduce the magnitude of complexity generated due to the huge number of objects.

(ii) A social relationship ontological model is developed, which helps identify the hidden and unidentified relationships among objects with reasoning mechanisms.

(iii) A use case with the implementation prototype based on the WoO platform is developed to exploit the social relationship model, supporting the proposed design for efficient IoT service provisioning.

Further, to support the above contributions, virtual objects, serving on real-world objects, have been developed, which can be reusable in other services as well. CVOs are implemented, where each update of the information in the real-world object is reflected in the VO and propagated to the CVO. Also, the Social IoT notion is incorporated into virtual objects to provide easy discovery and efficient execution while maintaining collaboration using social connections. Moreover, an experimental analysis has been carried out to evaluate the discovery time of VOs and the time required for service execution, with or without social relationships among virtual objects.

The remainder of this article is organized as follows. In Section 2, the related work is described. Section 3 presents the proposed design of the social web objects accompanied by the social interaction model. In Section 4, an IoT environment use case is explained with the details of the prototype implementation and discussion of the performance evaluation. Finally, Section 5 provides a conclusion of this paper.

2. Related Work

This section presents the related work with respect to the SIoT and microservices. Further, the significance of both technologies pertaining to the IoT environment has been discussed.

2.1. Social Internet of Things (SIoT).

The concept behind SIoT is to enable smart objects to create a social network similar to a human social network. The objective is to exploit the social relationships among objects in an IoT environment to facilitate the effective information discovery, to promote the scalability, to enhance the interactive communication between objects that are friends, and to achieve trustworthiness [3].

The initial notion of a social network of objects was proposed by [16], which discussed how the wireless devices form social connections on the temporary bases that are controlled by the node owners. As the work was carried out before the introduction of IoT, it did not consider the SIoT concept. Later studies [5] investigated the objects' social interaction where objects formed social networks and communicated with each other based on the rules already set.

The idea of using the Web of Things in a social ecosystem was presented by [17, 18], where things were shared using the social network infrastructures, for example, Twitter, Facebook, and LinkedIn. These implementations consisted of objects that communicated either through the built-in embedded servers or through smart gateways. Further, in these settings, the web-enabled things owned by a person provided an interface to his/her social group of friends so that they can interact with the things using a social network.

Lysis [19] is another good example of Social IoT platform based on cloud infrastructure. It uses the platform as a service (PaaS) model and focuses on the deployment of applications in the cloud space. However, dynamic inferencing of new social relationships from existing relationships is not supported by this system. Moreover, making a modular interoperable design by employing concepts like microservices and semantic web technologies is not considered in this work.

Paraimpu [20] provided a social Web of Things platform to connect virtual and physical things to the web. In this system, the social concept only relates to humans by providing them with the capability to share things with each other using a human social network. Also, virtual things are not considered as VOs but rather viewed as services on other IoT platforms.

Using SIoT at the edge of the network was investigated by [21]. In this system, an approach was proposed to cope with the communication delay due to the objects being remotely located in the cloud. This approach exploits the computing resources at the network edge to deploy virtual objects.

The ThingSpeak solution was used as the basis for the implementation of an early SIoT platform [22]. This platform provided the object social behavior functions incorporated at the centralized server. In this implementation, functions such as the creation and the management of the social relationships were developed. However, VO concept was limited only to the records in the remote database.

Another featured SIoT framework was contributed by [4]. This work presented how to combine services, devices, users, and their social interaction to enable interactivity, discovery, and recommendation of services. Social relationships have also been investigated in communities [23], where a proposed framework enables the identification of communities in the social networks.

Although most of the research contributions in the SIoT domain incorporate a notion of the social association of objects, however, limited information has been provided about the social relationship model and the details on how that can be used by IoT services. Furthermore, existing approaches are lacking lightweight and modular mechanisms which can be utilized by services to exploit social relationships efficiently. In our approach, we have developed a semantic ontology-based social relationship model that provides the capability to infer new connections in a network of social objects. Another distinction of the proposed system is a microservices based design that utilizes the developed social model to provide lightweight, modular, simple, and interoperable IoT services.

2.2. Microservices.

The microservices architecture is emerging as a new trend among the practices of developing distributed web applications. To develop SOA based systems, microservices have become a prominent approach in real-world settings [24]. In the microservices based architectural pattern, each application incorporates a collection of small services which execute independently and use lightweight communication mechanisms [9]. The microservices are targeted for doing one thing well at a time based on the principle of single responsibility [25], as defined by Robert C. Martine: "Gather together those things that change for the same reason and separate those things that change for different reasons" [26]. Also, the microservices are conceived as autonomous entities, which means they can be changed and deployed independently of each other without requiring their consumers to change [24].

In the last few years, several IoT architectural designs were proposed and developed. Some of these designs provided innovative features such as the European FP7 project iCore [27] which proposed a cognitive framework for IoT application development. This project provided an architecture with the concept of virtual objects and their composites. However, in iCore, services are designed using a traditional SOA based monolithic style, whereas the microservices based architecture provides a better option to develop IoT services. Furthermore, in iCore, VOs lack the mechanisms to semantically represent real-world objects (RWOs), which limits extensibility and interoperability.

To support IoT applications, microservices based architectures are being proposed in many research initiatives and applied in several research projects. In [28], microservices are used for M2M applications, realizing the fact that monolithic approaches for M2M cannot provide a real solution. Another work [10] provided an analysis of using microservices architecture for large-scale distributed applications. This work demonstrated that significant benefits can be achieved as compared to generic SOA approaches when using microservices in an IoT platform for smart city applications.

Moreover, Almanac FP7 EU Project [29] provides smart city services based on the IoT environment. The project uses microservices to employ scalability in a horizontal and vertical fashion. Also, in the industry, microservices have become a solution for developing large-scale applications. Netflix, Pivotal [30], and Amazon use microservices in their software bases.

The increased importance of microservices for the IoT applications is due to the fact that they simplify complex systems. By dividing a system into smaller parts, higher cohesion and lower coupling can be achieved, which makes it highly scalable. In scenarios where requirements keep changing continuously, microservices can help make a system easily modifiable. Moreover, the major benefits that are envisaged when using microservices in an IoT environment are as follows: Microservices leverage scalability to provide a highly decoupled pattern and can allow scaling individually. Their usefulness can be realized as when the demand for the requirement increases the system complexity also rises. In this case, a system can be supported by creating new instances of existing services. Microservices scalability [9] fits in three-dimensional space, that is, horizontal scalability (typical scalability), vertical scalability (splitting different individual microservices), and z-axis scalability (splitting similar things, such as DB partitioning). Microservices architecture supports a plug-and-play behavior where system components become loosely coupled. To acquire new functionality or replace a failed service, pull the plug from one microservice and plug into a new microservice [31]. One of the major problems that IoT is facing today is how to deal with the heterogeneity of incompatible solutions. To harmonize the heterogeneity, semantic web technologies with microservices are leveraged to provide interoperable exchange and communication of data. Microservices realize a decentralized and autonomous behavior, operating on their own priorities and schedule. This way, they provide several benefits for being utilized in an IoT design.

3. The Proposed Design of Social Web Objects

Recently, there have been several research efforts [3–6, 17, 20, 32] for defining the Social IoT. This new aspect minimizes the complexity generated by the communication of billions of real-world objects and enables sharing of information. The social notion can be incorporated into the WoO platform to provide a semantically rich base for IoT applications that can utilize real-world object relationships for efficient information discovery. Moreover, this section provides a brief introduction of WoO reference architecture in Section 3.1. A classification of social relationships among web objects (virtual objects and composite virtual objects) is discussed in Section 3.2 and a social relationship model is presented in Section 3.3. Further, the details of functional components for the proposed architecture are explained in Section 3.4.

3.1. Web of Objects (WoO) Reference Architecture. To enable the deployment of IoT services on the World Wide Web, WoO provides a reference framework. In other words, it realizes the IoT services in such a way that virtualized objects are interwoven with resources to support the development, deployment, and operations of IoT services [33, 34]. The WoO platform provides service functionality by merging VOs and features of web applications. Moreover, the WoO platform uses semantic web technologies to enable interoperability among heterogeneous resources. This realization provides the basis for the composition and harmonization of objects to provide smart services in an IoT domain. Building services based on WoO platform using diverse technologies, including microservices [35], social networks, and semantic web, helps reduce the complexity and fosters the easy and efficient development of IoT applications.

Virtualization has become a major concept in IoT to address the heterogeneity of diverse types of physical objects. VO is a digital representation of a real-world object that can be anything, living or nonliving, mobile or stationary, concrete or immaterial. In WoO, VO is defined as domain-specific semantic ontology based on the VO information model [36]. It is uniquely identified using URI and provides information updates on the representative real-world object. On the other hand, WoO also provides a notion of CVO which aggregates one or more VOs to enable service features that satisfy the application requirements. CVOs chain semantically interoperable VOs together. The WoO platform incorporates features to efficiently reuse existing VOs and handles the complexity generated by self-management and control mechanisms [37]. In the WoO, service level decides which CVOs and VOs will take part in creating an object mashup to fulfill service objectives. The service logic is saved in the form of templates and stored in the template repository [38–40]. The domain expert or knowledge engineer defines the service templates that are used to instantiate new services.

WoO reference architecture is shown in Figure 1. In this architecture, the service layer handles requests and provides several management functions, whereas the CVO level consists of functions to manage and instantiate CVOs or reuse the existing ones based on the provided service execution logic. New objects such as sensors are registered in the WoO platform using a registration function and their templates are generated to make them digitally available in the form of VOs. Another most important aspect is the semantic representation of data at three-layered WoO architecture [23, 39]. The database at each layer is supported with semantic web technologies.

3.2. Types of Social Relationships. Previous research [3] on SIoT has derived some basic relationship types among objects. These relationships are characterized as follows. *Parental Object relationship* (POR) exists between the objects that belong to the same batch, such as objects that are created in the same production process or at the same time by a common manufacturer. *Cowork object relationship* (CWOR) is the relationship among objects that are grouped together based on some commonly shared job to be done by them. Similarly, *colocation object relationship* (CLOR) is established

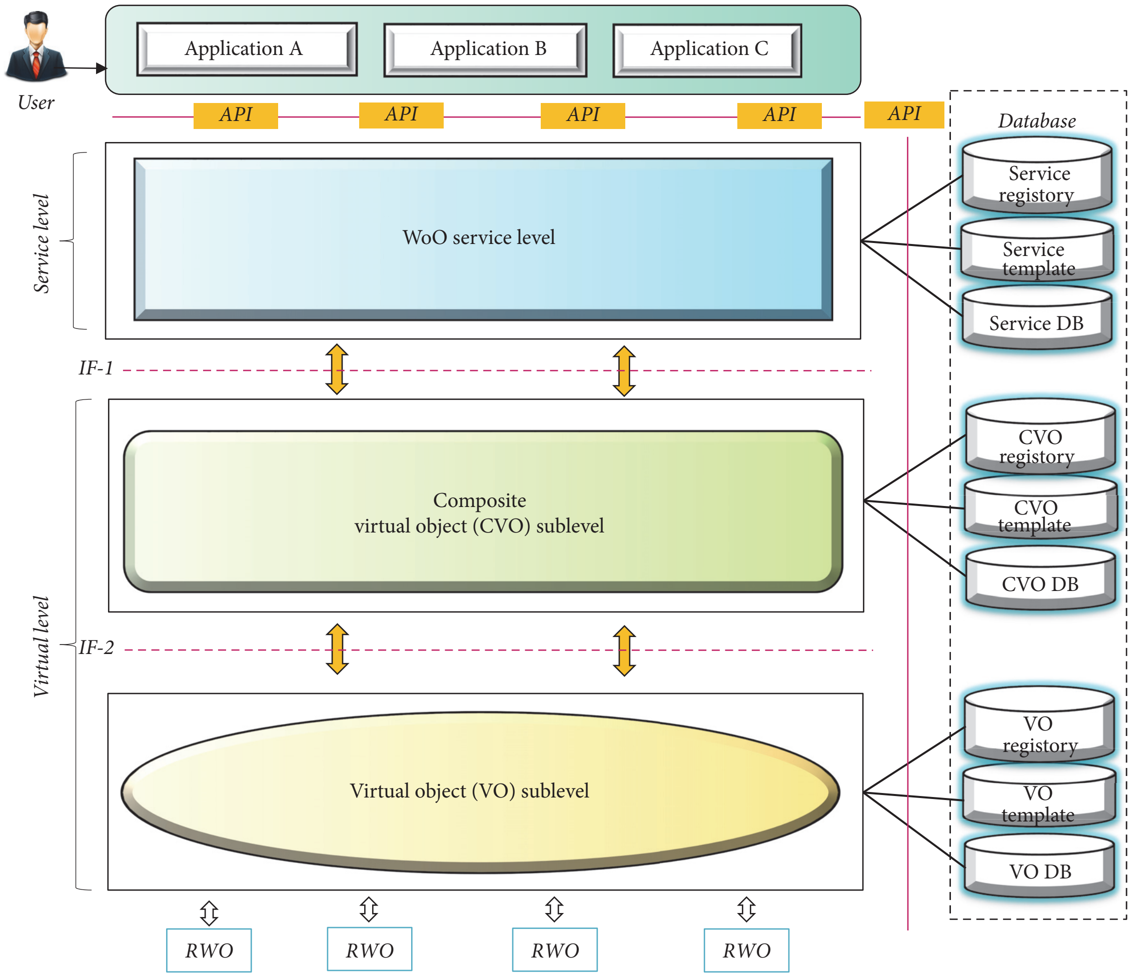
User
Application A
Application B
Application C
API
API
API
API
API
Database
Service registory
Service template
Service DB
Service level
WoO service level
IF-1
Composite virtual object (CVO) sublevel
CVO registory
CVO template
CVO DB
Virtual level
IF-2
Virtual object (VO) sublevel
VO registory
VO template
VO DB
RWO
RWO
RWO
RWO

Figure 1: Web of Objects reference architecture.

between objects belonging to the same location such as home or office. Different or same types of objects can be combined based on the common location or premises. On the other hand, *Ownership Object Relationship* (OOR) is formed among objects belonging to the same owner. *Social Object Relationship* (SOR) is another kind of relationship that is created among objects when they come close to each other, either random in time or periodically. This relationship can be envisaged when the owners of objects come into contact with each other. Apart from the above relationships, in [41], the authors define the *Guardian Object Relationship* (GOR) in the Internet of Vehicles (IoV) scenario where on-board units of vehicles become children in relation to the super nodes of Road Side Units. This child and parent association gives a special meaning to a new hierarchical relationship.

In addition to the relationships defined in other studies, we have identified few more relationships suitable for some diverse scenarios. These include *Sibling Object Relationship* (SIBOR) that is created among objects that belong to a group of friends or family members. This relationship among objects extracts the property of trustworthiness based on the relationship of their owners. Another relationship type is the *guest object relationship* (GSTOR) which is formed between objects that belong to the users in the guest role, for example, when a person visits a friend's home and gets the privilege as a guest. The same can be applied to the objects that move in guest relation (GSTOR) from one place to another; they will have some privilege of accessibility of information as compared to other objects. *Stranger Object Relation* (STGOR) applies to objects that encounter the presence of each other in an anonymous environment such as on the go or in the public environment. Similar to people who meet each other sometimes regularly but are anonymous to each other or are not fully aware of each other, this relationship can be used among objects to form different trust levels or to form strict restrictions to promote secure connections. Moreover, in *service object relationship* (SVOR), objects form a relationship while coordinating in the same service composition to fulfill a service request. Generally, in the proposed scheme, service request execution generates a mashup of objects, and objects that belong to the same mashup are assumed to be in SVOR.

We believe that the above defined new social relationships are beneficial in several distinctive scenarios. For example, SIBOR is more useful in scenarios where trust is highly important for establishing a connection. Unlike the more generalized relation POR, SIBOR is established

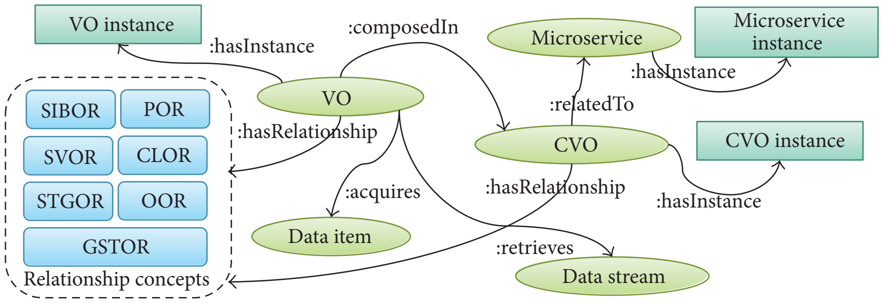

(i) Sibling Object Relationship (SIBOR)
(ii) Service Object Relationship (SVOR)
(iii) Stranger Object Relationship (STGOR)
(iv) Guest Object Relationship (GSTOR)
(v) Parental Object Relationship (POR) [3]
(vi) Colocation Object Relationship (CLOR) [3]
(vii) Ownership Object Relationship (OOR) [3]

Figure 2: Ontological model representing web objects' relationships.

among objects from the same manufacturer having similar features but different behaviors. Similarly, GSTOR is suitable in scenarios where objects use relationships for privileged services. Imagine a person as a guest speaker in a conference with his smart devices, where he can connect to the network easily, and his personal devices are recognized as a guest (with GSTOR) in the venue network to avail free services like high-speed Internet, special notifications, and others. Stranger Object Relation (STGOR) is used in scenarios where objects are on the go, and to get some information, they have to compromise with a limited trust level. For instance, in a public transit system, one mobile node on the move wants to use crowdsourced information from another mobile node which is not fully trusted. Also, SVOR is useful in service composition specific scenarios, where object mashups are formed to facilitate service requirement, and it is also useful in a particular orchestration of objects.

To model information about microservices, social web objects, and their relationships, a semantic ontological model has been developed as depicted in Figure 2. Conceptualization of objects and their social relationships in semantic structures are highly beneficial to infer new connections among a network of objects with the help of reasoning techniques. The ontology facilitates the representation of objects, as well as what they measure in terms of observations, their processing, and functions. Web objects' relationship ontology contains the concepts that define each object in terms of a class. The data properties are used to represent object values and the object properties are defined to identify the link between two objects in the ontology. Major concepts in the ontology are microservices, CVOs, VOs, and their social relationships. Further, microservices have an individual instance that uniquely defines a microservice object and associated data with it. Similarly, CVOs and VOs are also instantiated and contain data properties to retain their values. With the increasing number of objects, a semantic ontology model is used to perform reasoning function that helps identify new connections. Further, to analyze the social relationship among objects (such as VOs and CVOs), their properties are provided in Table 1.

3.3. Social Relationship Model for Web Objects.

At each level of service life cycle, objects form a social relationship with each other to accomplish a service task. Codifying the relationship among web objects can provide many benefits, such as efficient information discovery from related objects and better composition and reuse. To perceive social interaction between web objects, we assume every service is based on one or more microservices and other objects including CVOs and VOs. At VO level, VOs create several relationships with each other based on the RWO they represent. CVOs associated with microservices are incorporated in relationships as well.

Object-to-object associations are categorized as either vertical associations (interobject relations) or horizontal associations (intraobject relations) as shown in Figure 3. The relationships that flow from the bottom to the top in WoO based architecture are known as interobject relations (i.e., from VO to CVO or vice versa). On the other hand, relationships are known as intraobject relations if they are generated within objects at each level such as between one VO and another, or within one CVO to a relative CVO. The newly defined relationships among objects are maintained at different levels, which enables information discovery more efficiently. It is considered that not only will the relationships be maintained when two objects (i.e., VOs) are used in the same CVO or service, but also they will be maintained if objects are used in different CVOs or services.

As depicted in Figure 3, VO1 and VO2 bind in the SIBOR relationship, considering the fact that they belong to the same family of CVO1. Similarly, VO4, VO5, and VO6 are combined in a CWOR, as they are used in the same service. However, all of them do not have a common family relation that makes their trustworthiness restrictive to some level as compared to SIBOR. Moreover, though VO7 and VO8 are not used together in a service setup, still they form a STGOR relationship. This relationship also helps codify those VOs that have similarity in premises during a certain point in time regularly but they had never combined in any service scenario before. For example, if a person visits a subway regularly but stays very shortly, in such scene, the user's smartphone VO and subway station point

Table 1: Social relationship attributes.

Social object relationships types	Attributes/properties
Sibling object relation (SIBOR)	O_x *has relationship* with O_y → SIBOR, if objects' *ownership* is defined as: == SameFamily
Service object relationship (SVOR)	O_x *has relationship* with O_y → SVOR, if objects' *serviceStatus* is defined as: == SameServiceComposition
Stranger object relation (STGOR)	O_x *has relationship* with O_y → STGOR, if objects' *authentication* is defined as: == Anonymous
Guest object relationship (GSTOR)	O_x *has relationship* with O_y → GSTOR, if objects' *ownership* is defined as: !=SameFamily && == GuestFriend
Parental object relationship (POR)	O_x *has relationship* with O_y → POR, if objects' *creation* is defined as: == SameBatch ‖ == SameProduction
Colocation object relationship (CLOR)	O_x *has relationship* with O_y → CLOR, if objects' *proximity* is defined as: == Neighborhood ‖ *location* is defined as: == SameLocation
Ownership object relationship (OOR)	O_x *has relationship* with O_y → OOR, if objects' *proprietorship* is defined as: == SameOwner

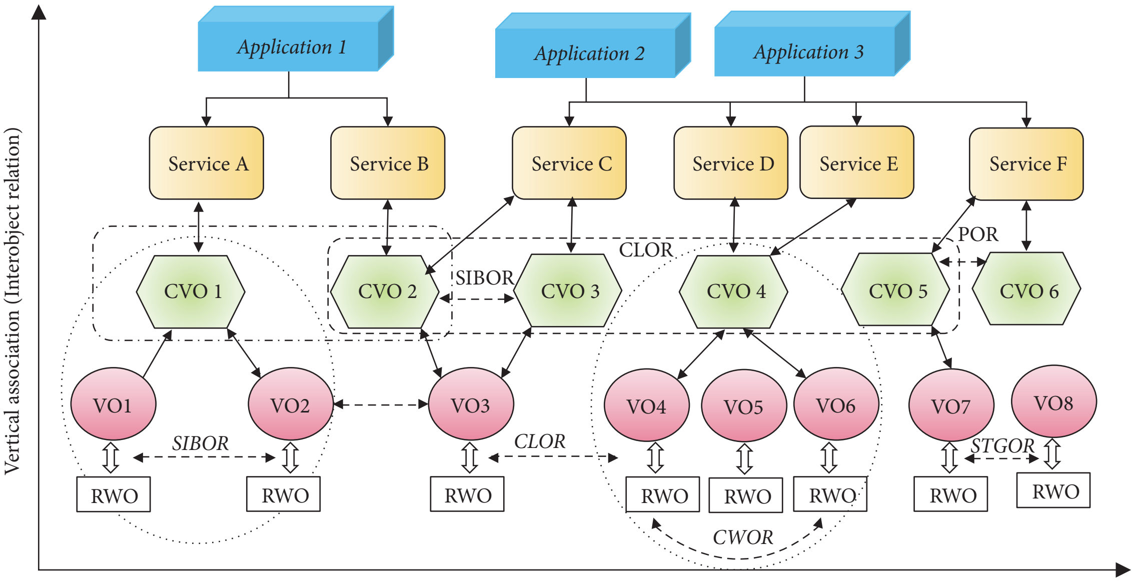

Figure 3: Hierarchical social relationships of web objects.

VO or subway security camera VO may establish STGOR. This may help reuse VOs in service scenarios where only for short duration a VO object is required. At CVO level, CVO2 forms two different relationships, which increases the level of connectivity of the CVO. CVO2 first shares a CLOR with CVO1 and CVO3, CVO4, and CVO5 together. Secondly, CVO2 is combined in a SIBOR with CVO3 and becomes in the same service family. Further, in the selected use case scenario described in Section 4, several other social interactions have also been elaborated.

3.4. Social Web Objects Architecture. SIoT envisages a system where the social framework will bring smart devices and people to interact with each other. By incorporating web technologies like SOA based microservices, IoT services can be rendered on top of the social framework. The proposed social web objects architecture (as shown in Figure 4) has been decomposed into three levels: service level, object virtualization level, and aggregated object virtualization level. At the service level, to support social relationships within web objects, several microservices have been designed, which are discussed further in the following section. Moreover, the data management function in the proposed design helps each layer to interconnect with semantic databases. The SPARQL endpoints have been defined at the service and the object virtualization levels that expose interfaces to retrieve, store, and modify RDF graphs. Several interfaces allow the knowledge engineer, domain expert, and the developer to create service templates and VOs, as well as update the RWK model, user profile, and policies.

3.4.1. Service Level Functional Components. The service level functions handle operations from the request inception to the execution of services. This level is supported by some core functions common to be used in each service life cycle. These include the management function to handle service

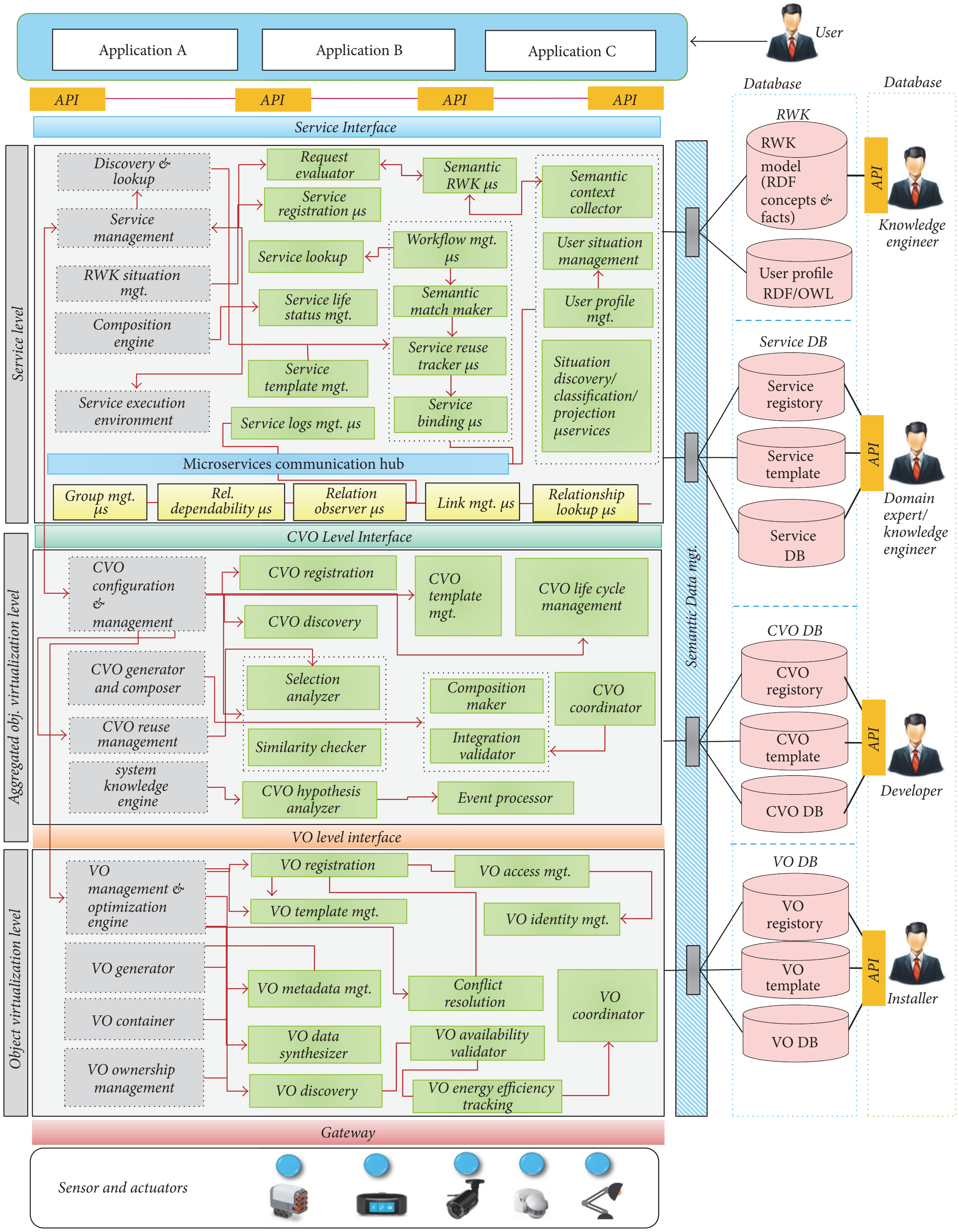
Application A
Application B
Application C
User
API
API
API
API
Service Interface
Service level
Discovery & lookup
Service management
RWK situation mgt.
Composition engine
Service execution environment
Request evaluator
Service registration µs
Service lookup
Service life status mgt.
Service template mgt.
Service logs mgt. µs
Semantic RWK µs
Workflow mgt. µs
Semantic match maker
Service reuse tracker µs
Service binding µs
Semantic context collector
User situation management
User profile mgt.
Situation discovery/ classification/ projection µservices
Microservices communication hub
Group mgt. µs
Rel. dependability µs
Relation observer µs
Link mgt. µs
Relationship lookup µs
CVO Level Interface
Aggregated obj. virtualization level
CVO configuration & management
CVO generator and composer
CVO reuse management
system knowledge engine
CVO registration
CVO discovery
CVO template mgt.
CVO life cycle management
Selection analyzer
Similarity checker
Composition maker
Integration validator
CVO coordinator
CVO hypothesis analyzer
Event processor
VO level interface
Object virtualization level
VO management & optimization engine
VO generator
VO container
VO ownership management
VO registration
VO template mgt.
VO metadata mgt.
VO data synthesizer
VO discovery
VO access mgt.
VO identity mgt.
Conflict resolution
VO availability validator
VO energy efficiency tracking
VO coordinator
Gateway
Sensor and actuators
Semantic Data mgt.
Database
Database
RWK
RWK model (RDF concepts & facts)
User profile RDF/OWL
API
Knowledge engineer
Service DB
Service registry
Service template
Service DB
API
Domain expert/ knowledge engineer
CVO DB
CVO registry
CVO template
CVO DB
API
Developer
VO DB
VO registry
VO template
VO DB
API
Installer

Figure 4: Social web objects functional architecture.

requests and initiate service provisioning processes, the service discovery function to select the available services, the composition engine to enable service mashups, and the Real-World Knowledge (RWK) and situation management function to acquire RWK and continuously maintain it. Besides these core functions, several microservices have been used to provide the plug-and-play feature in the proposed design; that is, in case of service failure, it is replaced with another one. These services include the Request Evaluator microservice which interprets a service request and matches a request query to the service template. The registration microservice registers newly added service objects to the registry. When a new template is instantiated, objects are put in service, and to know the existence of available service instances, they are recorded in the service registry. The lookup microservice is defined to search for the required service objects using the semantic representation. The representations available in RDF/OWL format are queried from the registry based on the request parameters. Service life cycle management keeps track of the states of service objects during their life cycle, whereas RWK management microservice is a semantic data processing service which processes facts about user preferences, profile, and situational information.

To realize service provisioning, it is always not possible to satisfy the service requirement with a single service function. Service composition and harmonization features are required to fulfill a demanding service request. Service composition is not a single operation; many microservices are used to compose the individual features. One of the major components of the composition is the Workflow Management (WFM) microservice which is responsible for decomposing the service requirement if no single service is available to satisfy the task. WFM microservice identifies the service objects using input and output interfaces. In this concern, it uses the matchmaking microservice to identify the approximate matches that can serve the required service functionality. Here, service objects are semantically represented in the form of ontologies. The semantic ontology alignment algorithm is used to match the service up to a certain threshold level that is defined by the matchmaking microservice. The reuse tracker microservice is used to identify the current service instance usage by the application; it incorporates a coordination mechanism before instantiating any new instance, and this helps in service reuse by many clients. Binding microservice enables tying of services selected by WFM; this function synchronizes input and output of services in a workflow to produce a single service output based on the generated composition plan.

Furthermore, user characterization is another important element to support the IoT services by matching user requirements more closely. The User Profile Manager microservice handles many facts related to the human user. These include user profile, preference, context, and policies. Semantic Context Collector microservice organizes the services requested by the user and the context in which these services were requested. This helps in identifying user interest with respect to the service context. User context or situation helps in service selection more efficiently if a user requests again with a similar service requirement. Some microservices have been designed to handle user situation information; these microservices include situation discovery, classification, and recognition functions. Situation discovery is used to process sensor data that is provided by CVOs to prompt a situation, whereas machine learning algorithms are applied to remove the false values and data normalization is performed in case inappropriate triggers in the data exist. Situation classification incorporates the reasoning methods and rules to generate a relationship over detected events and refine the missing knowledge between the events. Situation projection analyzes the events data with the help of machine learning methods and provides an output in the form of predicted facts that are related to the situation being detected. Moreover, all the microservices coordinate using a communication hub to share information with each other.

A flow of service inception to service execution in the proposed architecture is depicted in Figure 5 where service management is responsible for initiating the above-mentioned functionalities and instantiating the microservices. This process further results in the generation of CVO/VO mashup graphs to be handled by CVO management (discussed in Section 3.4.2). Situational information has also been used to improve request analysis and service provisioning, whereas VO management (discussed in Section 3.4.3) acquires and aggregates data from sensors and other RWOs.

Social Relationship Management Microservices. To support the social relationships among web objects at each layer of the proposed system, relationship management microservices are defined, which support the codification of relationships among web objects (VOs and CVOs) and their management. These microservices are defined as follows. Group management microservice identifies the grouping of objects into specific sets based on the type and interaction of objects; it builds a social graph of the object relationships. Link management microservice incorporates mechanisms to maintain relationships among different objects and it includes several other subfunctions. The first subfunction is object selection which involves choosing candidate objects that are likely to form a particular association. The second subfunction is matchmaking service which provides exact relationship match based on the type of objects. The third subfunction is the association life cycle manager, which maintains the relationship status between objects; it checks whether they are in active relation or not. It keeps track of relationship validity and also the duration of relationships. Relationship observer microservice enables the object activities to be observed, and it also handles the mutual sharing, which defines what information an object is allowed to share or use of another object based on the relationship among them. Relationship dependability recognizer microservice identifies, maintains, and continuously improves the level of reliability among objects. If objects A and B are frequently interacting and always perform a task together that leads to successful execution of an application to serve the requirements, then they are more likely to form a reliable relationship with each other. The consistency of the relationship allows objects to share their information with fewer constraints as compared to unpredictable association among objects. Relationship

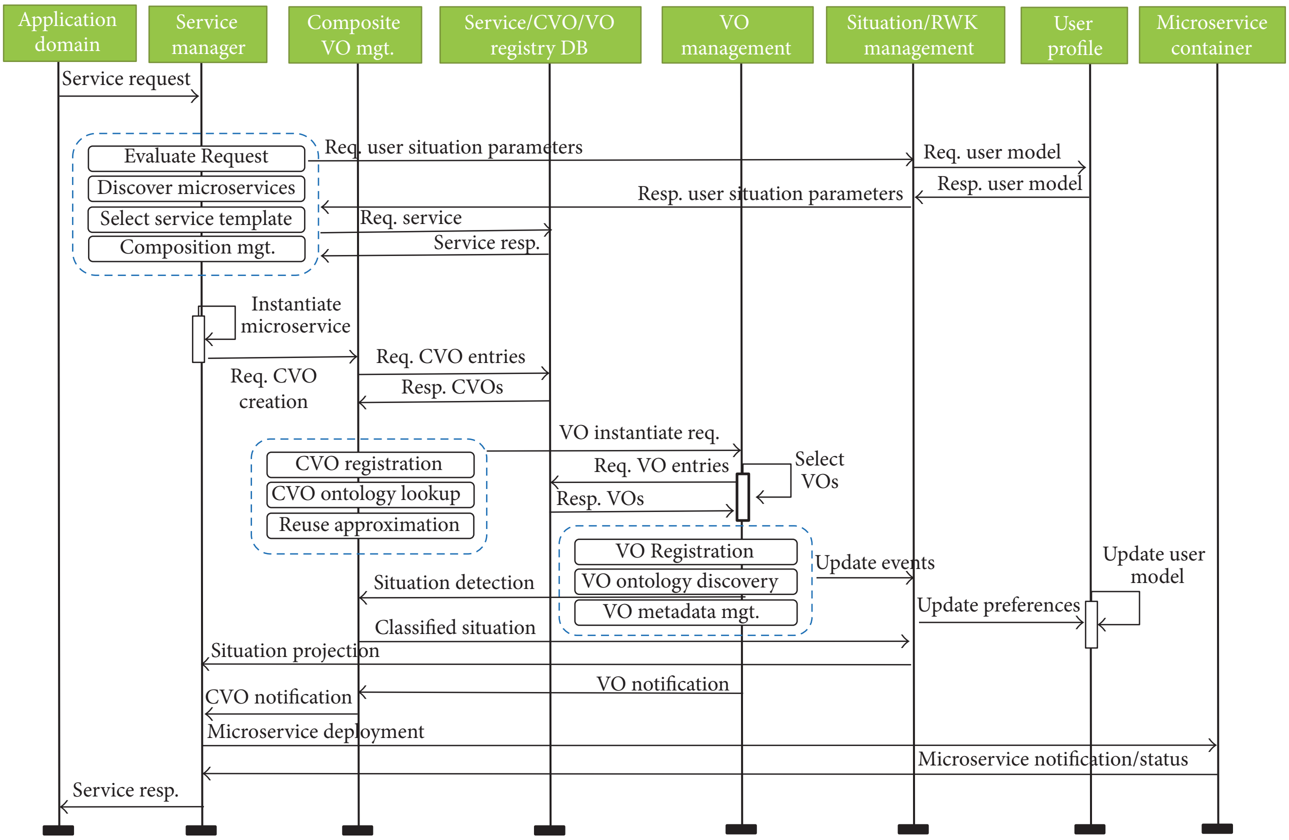
Application domain
Service manager
Composite VO mgt.
Service/CVO/VO registry DB
VO management
Situation/RWK management
User profile
Microservice container
Service request
Evaluate Request
Discover microservices
Select service template
Composition mgt.
Req. user situation parameters
Req. user model
Resp. user model
Resp. user situation parameters
Req. service
Service resp.
Instantiate microservice
Req. CVO creation
Req. CVO entries
Resp. CVOs
VO instantiate req.
CVO registration
CVO ontology lookup
Reuse approximation
Req. VO entries
Select VOs
Resp. VOs
VO Registration
VO ontology discovery
VO metadata mgt.
Update events
Update user model
Update preferences
Situation detection
Classified situation
Situation projection
VO notification
CVO notification
Microservice deployment
Microservice notification/status
Service resp.

Figure 5: Request initiation and service execution flow in the proposed architecture.

lookup microservice incorporates the discovery algorithm (as given in listing of Algorithm 1) to discover the social relationships and maintains the list of all objects and their potential association rating. The rating defines the opportunity for objects to form a relationship with each other.

Moreover, the operation flow of social microservices has been illustrated in Figure 6, where VO management retrieves the social relationship among objects this is queried using link management, which uses a lookup service. Associations are maintained in a social relationship graph using semantic ontology and queried using the relation observer service. The social graph is also updated and maintained continuously with respect to changing links among objects, on the basis of which VOs are selected.

3.4.2. Functional Components in Aggregated Object Virtualization Level. Aggregated object virtualization (AOV) level involves all the necessary functions to instantiate, manage, and continuously monitor CVOs. Besides, it includes the interface required by the service level to interact with CVO components. Service level provides the mashup graphs of CVOs and VOs that are required to fulfill a service request. Aggregation level also supports some core functions; these are categorized as CVO configuration and management to decompose requests from service level and manage CVO instances, CVO generator to instantiate CVOs, composition function to form CVO and VO mashups, reuse management to select existing CVOs that approximately match service requirement, and system knowledge engine to grow knowledge for the optimal use of CVO resources. Further, other aggregation level functionalities are as follows: registration function records the CVO entry into the semantic CVO registry in the form of RDF/JSON format. CVO includes ontologies that specify conditions to be applied on the VO data. Also, at this level, templates are created by a domain expert or knowledge engineer, which consist of functionalities associated with each CVO type. Template management module provides an interface to include templates to the CVO template repository and also supports the modification mechanism. CVOs are discovered using their semantic annotation via the discovery function. CVO selection analyzer function provides estimated selection of a CVO for reuse if there is no single CVO that can exactly provide the required service. This is supported by similarity checking mechanism to analyze the level of similarity before recommending for reuse. At aggregation level, another important function is event processing; event streams generated by object virtualization level are highly valuable to infer situational awareness and building RWK that is used by high-level functions. Event processing module facilitates processing of events and provides generated facts as output. Event processing is supported by CVO hypothesis analyzer that incorporates several hypotheses which are trained over data to build system level knowledge. Aggregation level also

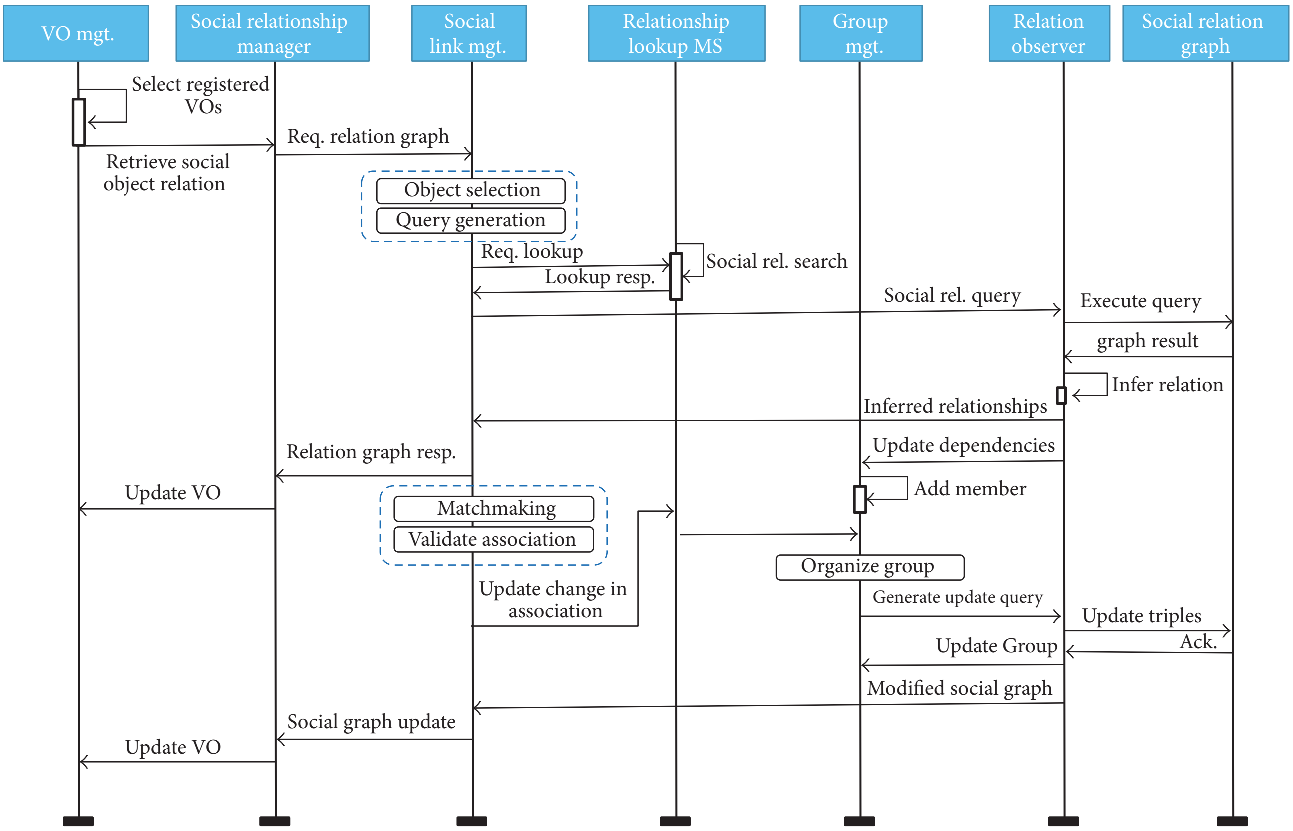
VO mgt.
Social relationship manager
Social link mgt.
Relationship lookup MS
Group mgt.
Relation observer
Social relation graph
Select registered VOs
Req. relation graph
Retrieve social object relation
Object selection
Query generation
Req. lookup
Social rel. search
Lookup resp.
Social rel. query
Execute query
graph result
Infer relation
Inferred relationships
Update dependencies
Relation graph resp.
Add member
Update VO
Matchmaking
Validate association
Organize group
Update change in association
Generate update query
Update triples
Ack.
Update Group
Modified social graph
Social graph update
Update VO

Figure 6: Sequence of operation required to establish a social relationship among objects.

provides composition function which provides the basis for merging multiple VOs to generate single service feature and validation function to validate the compositions of multiple VOs. Moreover, the coordinator helps in resolving conflicts in case of simultaneous access of single CVO by multiple services together.

3.4.3. Functional Components in Object Virtualization Level. Although IoT middleware virtualizes sensors and actuators in some form of virtual entities, however, in WoO, every living or nonliving thing can be represented in the virtual world. VOs are supported by an information model [14, 15] which includes the representation of RWO that can be ICT or non-ICT. In the proposed architecture, object virtualization level includes functions to create, maintain, and coordinate VOs such as VO generator, VO management, and optimization engine. VO registration function provides an interface to register VOs with the help of VO ownership management function where the VO ownership is maintained. Template management function handles templates or loads updated ones as provided by the domain expert. Metadata management maintains data associated with VOs; as VO is a digital representative of real-world entities, there is a high need to store metadata information to recognize the class of sensed data. Use of metadata provides the real value of the data in the form of the context stored within it. Furthermore, VO data synthesizer service handles the discrepancies if found in the data corresponding to VO; this includes data manipulation and missing value rectifications. Conflict resolution function helps in case the same VO is used by multiple service instances, to resolve the conflicts. Also, VO access management function controls access for VOs in case access right information has been defined by the VO owner, whereas identification management module maintains identification information about VOs so that each VO can be uniquely identified in the system.

3.4.4. Social Relationship Discovery and Composition. To identify associations among web objects within a social relationship graph, an algorithm is elaborated briefly. In the following algorithm, the microservices that are in execution are retrieved first, represented as $\Sigma M\mu$; next, all the CVOs associated with each of those microservices are recovered in ΣMc and, further, the corresponding VO model ΣMv is iterated. In the next step, executing the query statement routine returns the results denoted as $\overline{\mathrm{T}}$. The process of finding the relationships is carried out through lines (6)–(19) in multiple nested iterations as shown in the pseudocode of Algorithm 1 listing.

Algorithm 2 facilitates the composition of objects based on social association types. This algorithm takes as input the set of object social relationships denoted by $\mathbb{Z}$ and registry entries of the objects that are currently available in the system, represented as $\mathbb{R}e$. After the object model denoted

```
Require: ($\Sigma\phi$), ($\mathbb{R}_e$)
(1) Output: $\hat{S}$ (social relationship graph of objects)
(2)  $⅀M_{\mu}$ ← Load Microservices in model
(3)  $⅀M_c$ ← Load CVOs in model
(4)  $⅀M_v$ ← Load VOs in model
(5)  ₸ ← executeQueryStatement ($M_{\mu}$, $M_c$, $M_v$, $Q$)
(6)  for All $\mu$ in $\mu$₸ do (iterate Microservice instance being used)
(7)      for All $O \in \mu$ do (check CVO and VOs used by the Services)
(8)          if $O_i$ == output of $C_R$ in a relationship set $R_c$ then
(9)              add $C_R$ to $\mathbb{L}_L$
(10)         else
(11)           if $O_i$ == output of $V_R$ in relationship set $R_v$ then
(12)             add $V_R$ to $\mathbb{L}_L$
(13)             $\mathbb{N}_i$ ← tag (Assign relationship tag to each object entry)
(14)           else
(15)             add $O_i$ to $\mathbb{L}_U$
(16)           end if
(17)         end if
(18)     end for
(19) end for
```

Algorithm 1: Social relationship discovery.

```
Require: ($\Sigma\phi$), ($\mathbb{R}_e$), ($\mathbb{Z}$)
(1) Output: $\mathbb{C}$ (Composite Service)
(2) ⅀M ← List all objects available (Microservices, CVOs, VOs) in model
(3) ₸ ← executeQueryStatement (⅀M, $Q$)
(4) $\mathbb{Z}$ ← executeQueryStatement (⅀M, $Q'$)
(5) for All $\sigma i \in$ ₸ do
(6)     for All $\lambda i \in \sigma$ do
(7)         if $\lambda i$ == any $z$ in $\mathbb{Z}$ then
(8)             add $\lambda i$ to $\mathbb{L}_{\mathrm{Lm}}$
(9)             store $\mathbb{L}_{\mathrm{Lm}}$
(10)        end if
(11) end for
(12) for All $\varpi \mathrm{i} \in \mathbb{L}_{\mathrm{Lm}}$ do
(13)     if $\varpi \mathrm{i}$ == any $\varpi \mathrm{i}$ in $\mathbb{L}_{\mathrm{Lm}}$ then
(14)        $\partial$ ← ranking (Assign rank to each service)
(15)        add $\varpi \mathrm{i}$ to $\mathbb{L}_{\mathrm{LR}}$
(16)        store $\mathbb{L}_{\mathrm{LR}}$
(17)     end if
(18) end for
(19) for All $\sigma j \in \mathbb{L}_{\mathrm{LR}}$ do
(20)     if $\sigma j(R\partial) > \kappa$ then
(21)         add $\sigma j$ to Wf
(22)         Pq ← assign to queue
(23)         Sort Pq
(24)         Formulate Wf
(25)     end if
(26) end for
```

Algorithm 2: Composition of objects based on social relationship.

as ⅀M is loaded, the first step is to compose objects based on their relationship types. This way, the SPARQL queries are executed; the first query is denoted as Q to retrieve the available service templates and the second query is represented as Q' to extract the relationships associated with the services. From line (5) to (11), first, the social relationship is searched among objects and then the found relative objects are composed in groups. The design philosophy employed over here is to assign a ranking to each selected group based on the relationship type that helps in a more efficient

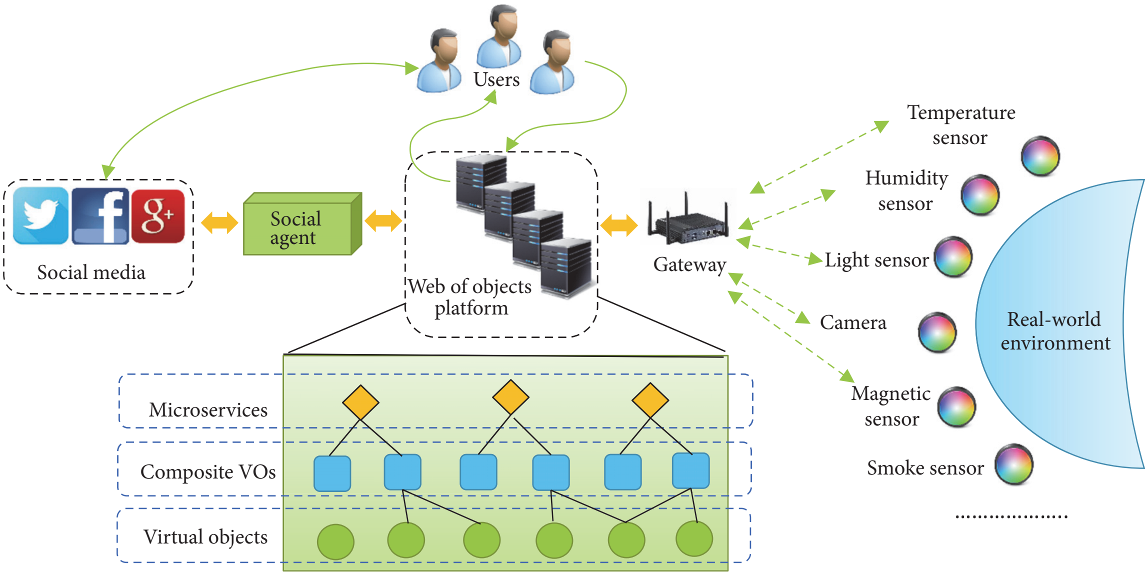

Figure 7: Use case scenario.

composition of service objects (from line (12) to (18)). The last part of Algorithm 2 will continue to iterate to compose the object composition workflow by checking first the assigned relationship ranking with the defined threshold, and then the selected composition workflow graphs are sorted and stored in queues for service execution.

4. Use Case and Prototype Implementation

In this section, first, to realize IoT service provisioning based on social web objects, a use case about IoT enabled museum environment has been presented and then the details on the prototype implementation are furnished.

4.1. Use Case Scenario. This use case is based on a user experience in a public museum where the IoT environment is already set up with the WoO platform incorporating web objects that use social relationship model and a social agent which obtains the user's profile and his friendship details from social media (as shown in Figure 7). The user enters the premises of the museum and his smartphone is prompted with a beacon signal to recommend service app installation where he accepts the request. The phone app incorporates a microservice which establishes a social relationship (GSTOR) with museum visit service objects. The system monitors the interaction of the user in the museum environment; the installed camera network and motion detectors detect the user presence and this feed is collected by the museum visit service. The user shares his current location with the system, which is used by the location navigator service to assist him in moving around the museum by providing the direction to the locations where different museum items are situated. It is worth noticing that the VOs for museum visit service and location navigator service are bound in CLOR relationship due to the same location points. The user shares his social network profile which is used to acquire the list of friends and choices for things the user is interested in. This information is acquired by the recommendation service which suggests the best things to do with respect to the user's interests, making his visit most enjoyable. Recommendation service suggests facilities available in the museum such as places a user should visit, cafeteria food menu, and prices for the day. This service also shares with the user the past experiences of his friends at the museum to help him get most out of his visit. Moreover, personalization service provides special features such as adjusting contents on displays or optimizing the luminosity, to support users with disabilities. To enable these service features, objects exploit the SVOR relationship. Also, the fire management service maintains the status of all VOs that can detect an unusual fire situation in the museum. The temperature VOs acquire readings from temperature sensors and smoke detector VOs get a feed from smoke detection sensors. These VOs form a CLOR relationship with each other in the common proximity. Fire situation monitoring CVO acquires VOs data. One or more CVOs are managed by the fire management service. In case of a fire breakout, the fire management service executes with the data from the fire detection CVOs and the temperature and smoke detector VOs. Meanwhile, services incorporate high-level functionality to react on the data provided by virtual entities in the system. It is important to notice that sharing of information based on social relationships has become easy in the proposed use case. That is because the user information and device information are communicated with social links of web objects and are utilized by the services.

4.2. Proof-of-Concept Details. To evaluate the proposed design, a proof of concept has been furnished, which realizes services based on web objects as considered in the above use case. These objects include, for instance, fire situation monitoring CVO to handle emergency fire breakout situations, location navigator CVO to find the shortest path to the user

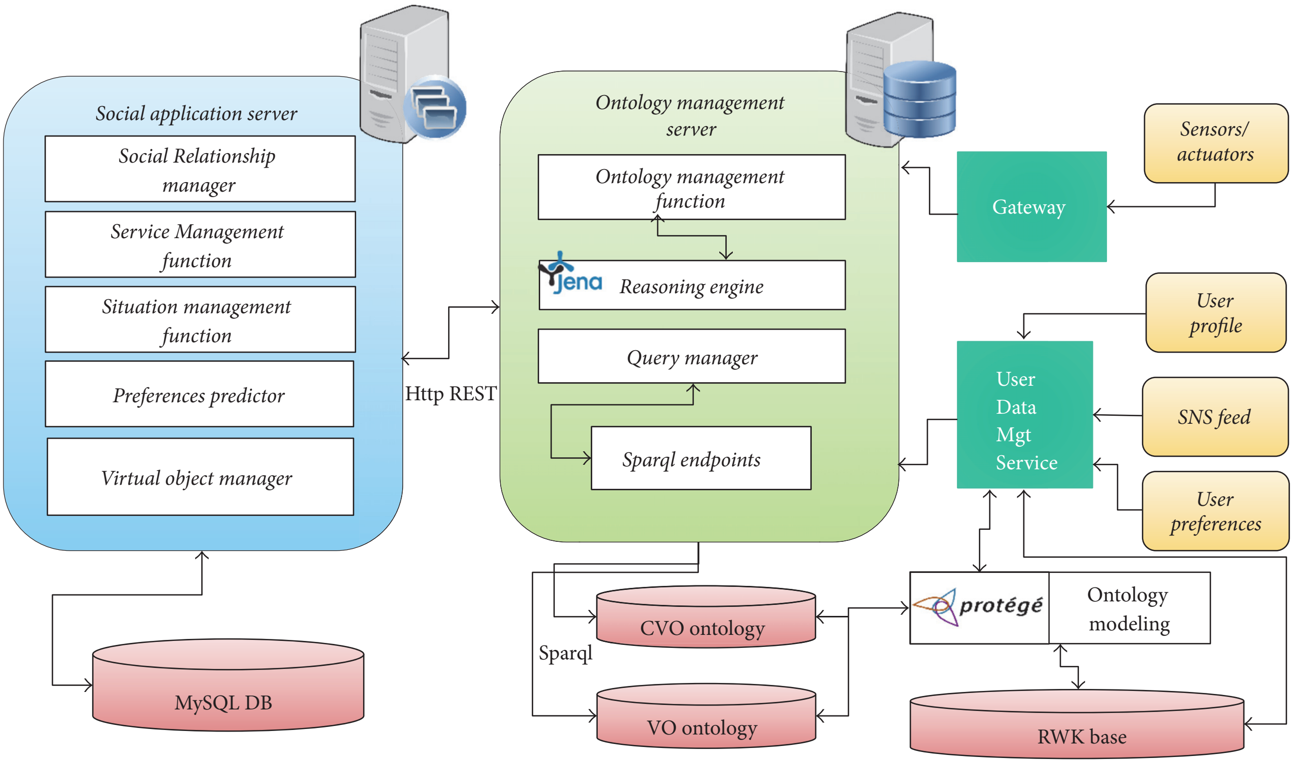
Social application server
Social Relationship manager
Service Management function
Situation management function
Preferences predictor
Virtual object manager
Http REST
Ontology management server
Ontology management function
Jena
Reasoning engine
Query manager
Sparql endpoints
Gateway
Sensors/ actuators
User profile
User Data Mgt Service
SNS feed
User preferences
MySQL DB
Sparql
CVO ontology
VO ontology
protégé
Ontology modeling
RWK base

FIGURE 8: Prototype implementation components.

intended destinations, and user profile reader CVO which acquires user personal information from the social media profile. Several other CVOs have been developed that acquire diverse data and metadata from VOs where each represents a different RWO. These VOs include a user VO, smartphone VO, camera VO, wristband VO, motion detector VO, and others that are associated with the chosen environment (such as CO_2 sensor VO, humidity VO, temperature VO, and light VO). Each VO has several instances; for example, temperature sensor instances are represented as temperatureSesor0001 and temperatureSesor0002. Further VO data is collected and analyzed by CVOs with the rules incorporated in the ontologies. CVOs composed in microservices help analyze different situations and initiate inference over data using inference engines. Considering the system requirement, several aspects have been evaluated, including identification of web objects for effective IoT service provisioning, finding the social relationships among web objects, efficient discovery of web objects based on these relationships, and composition and harmonization of web objects.

4.3. Prototype Implementation Details. To analyze the feasibility of the proposed architecture, a prototype on the discussed use case has been implemented. Figure 8 illustrates the components of the prototype, which constitute social application server (SAS), ontology management server (OMS), and databases to hold VO and CVO graphs that are supported with SPARQL endpoints. The gateway connecting sensors and actuators has been incorporated in the implementation. Additionally, user data management services are implemented to get user profile and preferences history along with social network data.

The application server comprises five major components: social relationship manager (SRM), service management function (SMF), situation/context management function (CMF), preference predictor, and virtual object manager (VOM). The SMF is an entry point for the service request; it interprets and evaluates the service request. It is also responsible for matching the service request with the available template. SMF communicates with VOM to check the VO or CVO availability. A compiled list of VOs is provided to SMF which generates object mashups to satisfy the service request. The execution of VO mashup graph is done using microservices. The second function, VOM, maintains instances of CVOs and associated VOs and coordinates with OMS to use available ontologies in the system. It records the association of services and VOs and enables the reuse of VOs by more than one service; it also resolves conflicts using conflict resolution mechanism. VOM is implemented using RESTful web services. Further, SRM is implemented to maintain object relationships in the form of RDF concepts represented in ontologies. It also involves several microservices, each implemented for a specific task. For example, group management microservice is developed to group objects that have similar characteristics. Similarly, other tasks are distributed in different microservices; for example, link management microservice is used to identify the dependency among objects. Relationship observer microservice is implemented to monitor the objects when they change their state at runtime, whereas to search a web

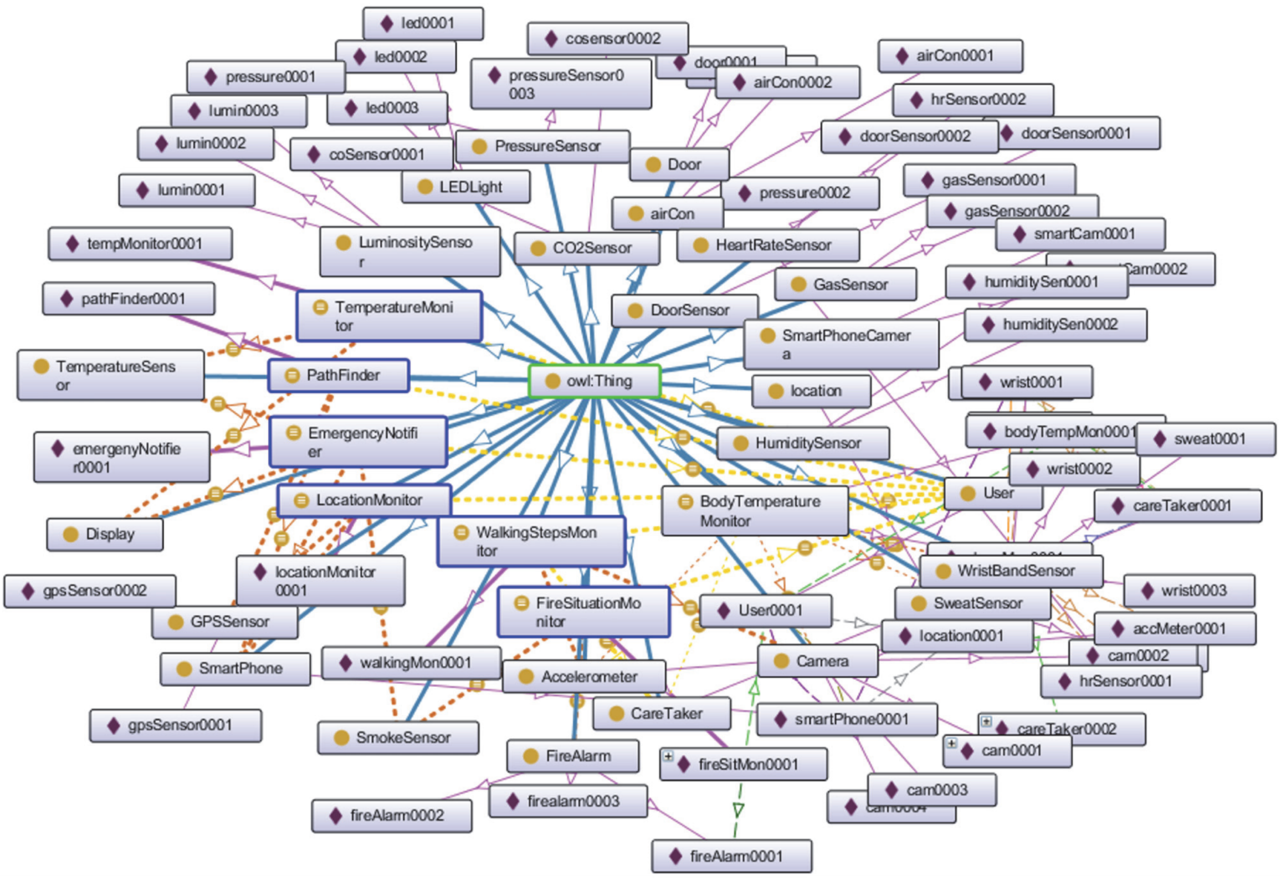
led0001
cosensor0002
led0002
door0001
airCon0001
pressure0001
pressureSensor0003
airCon0002
lumin0003
led0003
hrSensor0002
lumin0002
coSensor0001
PressureSensor
doorSensor0002
doorSensor0001
Door
lumin0001
LEDLight
pressure0002
gasSensor0001
airCon
gasSensor0002
tempMonitor0001
LuminositySensor
CO2Sensor
HeartRateSensor
smartCam0001
Cam0002
humiditySen0001
GasSensor
pathFinder0001
TemperatureMonitor
DoorSensor
humiditySen0002
SmartPhoneCamera
TemperatureSensor
PathFinder
owl:Thing
location
wrist0001
EmergencyNotifier
HumiditySensor
bodyTempMon0001
sweat0001
emergenyNotifier0001
wrist0002
User
LocationMonitor
BodyTemperatureMonitor
careTaker0001
Display
WalkingStepsMonitor
WristBandSensor
locationMonitor0001
gpsSensor0002
wrist0003
FireSituationMonitor
SweatSensor
User0001
GPSSensor
location0001
accMeter0001
SmartPhone
walkingMon0001
Camera
cam0002
Accelerometer
hrSensor0001
CareTaker
smartPhone0001
careTaker0002
gpsSensor0001
SmokeSensor
FireAlarm
cam0001
fireSitMon0001
cam0003
fireAlarm0002
firealarm0003
fireAlarm0001

Figure 9: Ontological model of social web objects.

object relationship, the lookup microservice is used. The CMF function has been developed to process the situational information and it selects the most suitable VOs based on the context. The situational information contains the current location of the user, location of the RWOs, and the area under observation. CMF helps form a VO configuration specific to the user current situation. Additionally, to incorporate the real-world facts about the user and the situation, RWK base has been developed which is updated continuously. This also helps in adapting services according to the current situation. VOs provided data is accumulated in RWK model and the reasoning engine is used to infer new facts based on the existing data. The last function of social application server is the preference predictor which provides predictions by using machine learning methods. To implement the learning function, Weka Library [42] has been used. For the objective of demonstration, user preference patterns are collected from the social data and user preference history is used to predict the future patterns.

On the other hand, OMS deals with the semantic representation of VOs. The first functional component of OMS is the ontology management service, which is used to handle VO and CVO ontologies and provides an interface to use VO graphs. Further, to manage ontologies in the developed system, we have used Apache Jena framework [43], and to model VO and CVO ontologies, Protégé [44] has been used. Models generated in Protégé are utilized by ontology management service and stored in OWL format in semantic data stores. To take full advantage of ontology-based implementation, a reasoning engine has also been used to get extra facts related to the modeled concepts. Reasoning engine helps to grow RWK facts stored in VO and CVO ontologies. For the proof of concept, the prototype has used Apache Jena inference engine along with Jena provided reasoner. For the deployment of persistent storage of the RDF data, Jena TDB has been used in our prototype, which is the Jena native high-performance triple store with API support [43]. Query manager is another important component of OMS which provides an interface to query CVO and VO ontologies saved in RDF/XML format, using SPARQL. Also, web services are implemented for the execution of SPARQL queries such as extracting and updating RDF triples in the database. Other than SPARQL query processing, semantic matching of existing concepts is also performed.

Moreover, to achieve interoperable relationships among web objects, a semantic ontology model has been developed as shown in Figure 9. The model constitutes the VO and CVO ontologies and their relationships. Several CVOs have been instantiated, for example, the location navigator, fire situation monitor, emergency notification, and many others that are

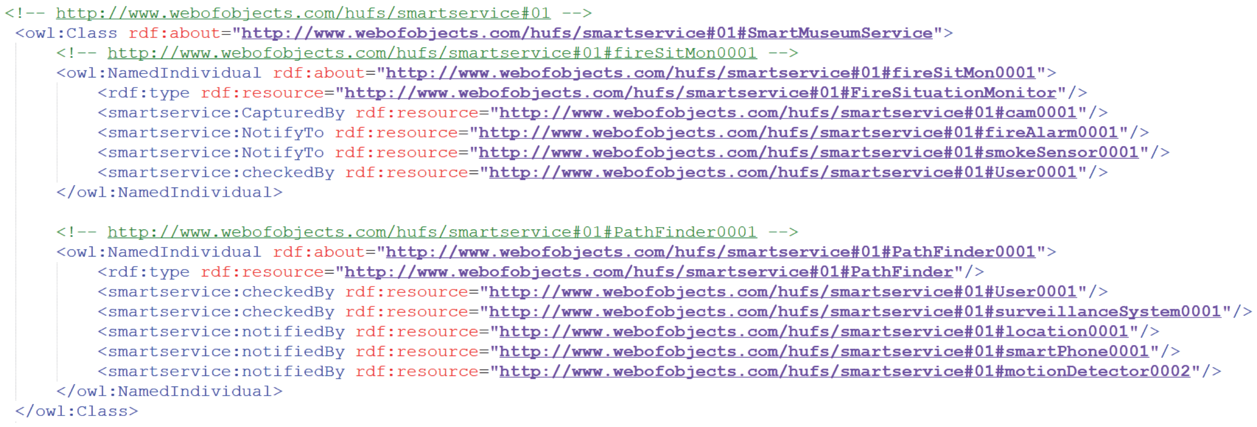

```
<!-- http://www.webofobjects.com/hufs/smartservice#01 -->
 <owl:Class rdf:about="http://www.webofobjects.com/hufs/smartservice#01#SmartMuseumService">
     <!-- http://www.webofobjects.com/hufs/smartservice#01#fireSitMon0001 -->
     <owl:NamedIndividual rdf:about="http://www.webofobjects.com/hufs/smartservice#01#fireSitMon0001">
         <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#FireSituationMonitor"/>
         <smartservice:CapturedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#cam0001"/>
         <smartservice:NotifyTo rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#fireAlarm0001"/>
         <smartservice:NotifyTo rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#smokeSensor0001"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#User0001"/>
     </owl:NamedIndividual>

     <!-- http://www.webofobjects.com/hufs/smartservice#01#PathFinder0001 -->
     <owl:NamedIndividual rdf:about="http://www.webofobjects.com/hufs/smartservice#01#PathFinder0001">
         <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#PathFinder"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#User0001"/>
         <smartservice:checkedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#surveillanceSystem0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#location0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#smartPhone0001"/>
         <smartservice:notifiedBy rdf:resource="http://www.webofobjects.com/hufs/smartservice#01#motionDetector0002"/>
     </owl:NamedIndividual>
 </owl:Class>
```

Figure 10: CVOs and VOs used in the ontology.

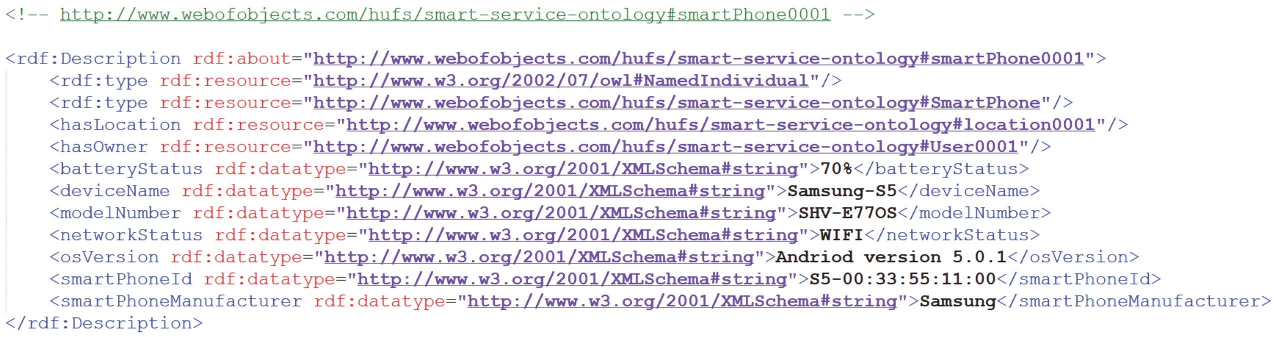

```
<!-- http://www.webofobjects.com/hufs/smart-service-ontology#smartPhone0001 -->

<rdf:Description rdf:about="http://www.webofobjects.com/hufs/smart-service-ontology#smartPhone0001">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NamedIndividual"/>
    <rdf:type rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhone"/>
    <hasLocation rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#location0001"/>
    <hasOwner rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#User0001"/>
    <batteryStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">70%</batteryStatus>
    <deviceName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Samsung-S5</deviceName>
    <modelNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string">SHV-E770S</modelNumber>
    <networkStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">WIFI</networkStatus>
    <osVersion rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Andriod version 5.0.1</osVersion>
    <smartPhoneId rdf:datatype="http://www.w3.org/2001/XMLSchema#string">S5-00:33:55:11:00</smartPhoneId>
    <smartPhoneManufacturer rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Samsung</smartPhoneManufacturer>
</rdf:Description>
```

Figure 11: Smartphone VO used in the ontology.

based on the composition of numerous VOs. This development has been supported by OMS and semantic web tools such as Jena. Using OMS, CRUD operations are performed like create, update, and delete on CVOs and VOs using SPARQL. In the developed semantic relationships, two CVO instances, namely, “FireSitMon0001” and “PathFinder0001,” have been used in the code snippet provided in Figure 10, for fire situation monitoring and location navigation, respectively. We can see the data properties and object properties associated with each object to maintain the relationship with other objects in the system. Further, to understand the VO representation, a smartphone VO “SmartPhone0001” is shown in Figure 11 with the defined properties. A list of VOs used in our prototype has been given in OWL description as shown in Figure 12.

In the test bed environment, the sensor to the gateway connectivity is enabled using ZigBee and BLE. Another source of information that has been incorporated to provide user-centric service capability is SNS feed, which consists of Twitter and Facebook data collected using social media APIs. On the other hand, user profile and preferences are updated using data management service, and this information is used to grow RWK base.

To test the prototype, an android application has been developed. After the user logs in to the app for WoO based assisted services, the app starts the configuration of services according to the user profile. A list of services is presented on the user’s screen from which the user can select any service. These services include museum visit service, location navigator service, recommendation service, and emergency notification. The selected service starts collecting information with the interaction of the user or when the devices start pushing values about any change in the observed environment. Based on these observed parameters, alerts are generated by the app; for example, the museum visit service generates alerts about the available facilities in the museum and the environmental conditions including the indoor and outdoor temperatures. On the other hand, the location navigator service helps the user to navigate around the museum by showing directions to several visitor attractions. Moreover, the emergency notification service triggers alerts when a temporary fire is created to mimic an unusual situation such as fire breakout. The resulting fire increases the temperature reading which is detected by the temperature sensor. This change in reading is forwarded through the gateway and analyzed by the fire situation monitor CVO which checks its threshold value and generates an emergency notification on the user’s app screen. Also, the recommendation service is used to generate suggestions about the museum events and the places to visit.

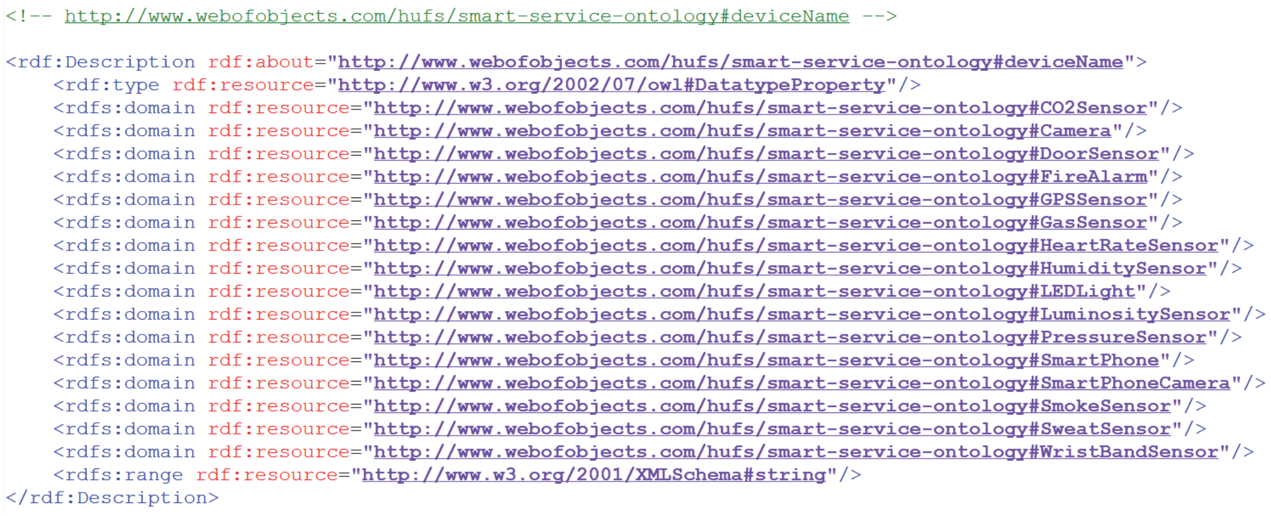

```
<!-- http://www.webofobjects.com/hufs/smart-service-ontology#deviceName -->

<rdf:Description rdf:about="http://www.webofobjects.com/hufs/smart-service-ontology#deviceName">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#CO2Sensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#Camera"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#DoorSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#FireAlarm"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#GPSSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#GasSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#HeartRateSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#HumiditySensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#LEDLight"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#LuminositySensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#PressureSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhone"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmartPhoneCamera"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SmokeSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#SweatSensor"/>
    <rdfs:domain rdf:resource="http://www.webofobjects.com/hufs/smart-service-ontology#WristBandSensor"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</rdf:Description>
```

Figure 12: List of VOs used in the ontology.

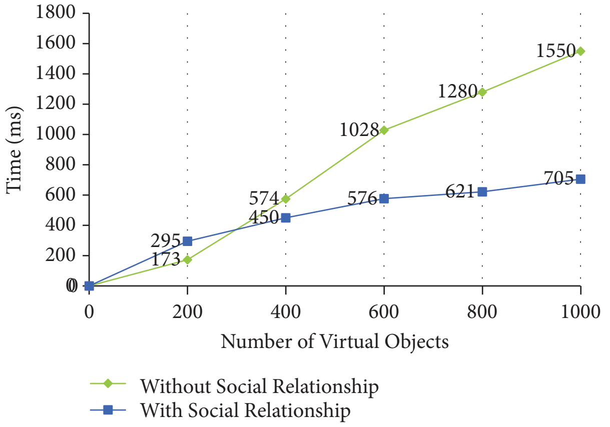

Figure 13: VO discovery time with the increasing number of virtual objects.

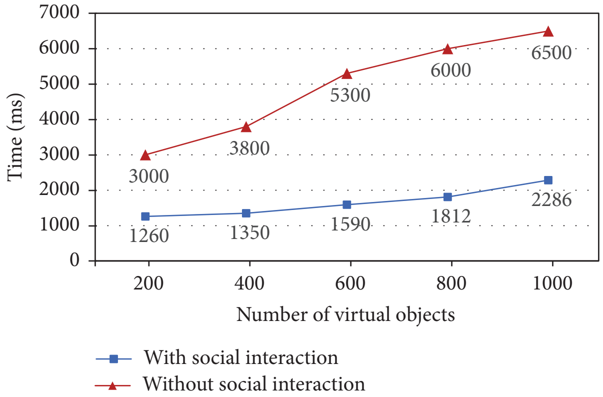

Figure 14: Service execution time with the increasing number of virtual objects.

4.4. Performance Evaluation. The proposed architecture has been analyzed from two different perspectives: system scalability and resource consumption. The results with respect to the experiments and the performance criterion are discussed in this section. The first experiment involves the analysis of the time required to discover the virtual objects in the proposed system with or without social relationship criterion as shown in Figure 13. As it is apparent from the results, initially, to form a social graph, VO discovery process incurs delay due to the communication for establishing social relationship links. However, direct lookup for VOs in the registry at the start of the discovery process comparatively requires less time, but as the number of VOs increases, the delay rises. Traversing the social links to discover objects as the number of virtual objects grows considerably reduces the total lookups in the registry database.

The second experiment provides an analysis of the execution time required for the CVOs as shown in Figure 14. To test this in a real environment, an android application has been developed. The app allows access to the parameters from the available VOs. When the app requires retrieving observations, it iterates through VOs in the repository using a query interface. This requires high CPU utilization with more execution time required to retrieve the sensor values against each app generated request. This is due to the fact that the decision for the selection of VOs has been done at runtime; it requires more time to select the VO graphs, execute them, and get the relevant data from the VOs. However, when object interaction is based on the social relationship, it requires less time for the retrieval as friends in the social network share the data on other friends' requests, which results in less computation as compared with the previous scheme.

Moreover, Figure 15 depicts the time to discover VOs in four different types of services. These services are the museum visit service, the location navigator service, the emergency notification service, and the recommendation service. The discovery time varies from one service to another depending on the number of VOs used. However, as compared with the impact of social connections, it can be

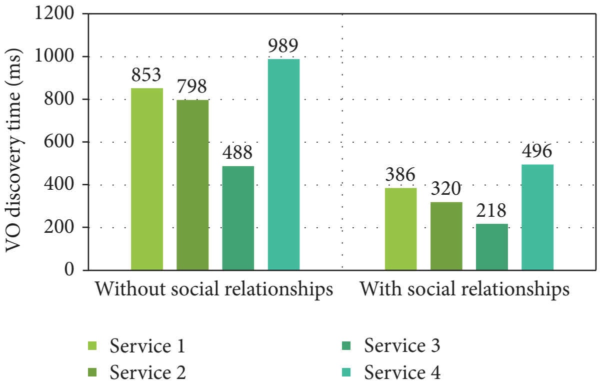

FIGURE 15: VO discovery time for four different services with or without social relationship utilization.

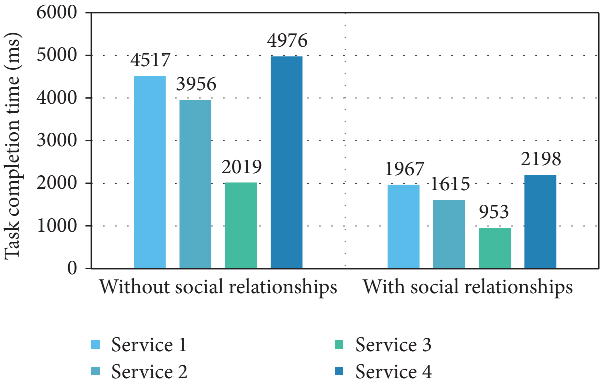

FIGURE 16: Task completion time required for each service with or without social relationship exploitation.

realized that the services can considerably reduce the time of object discovery and execution by using the relationship model. This is due to the fact that with social association among web objects a service does not have to query the registry for all VOs and rather VOs in the same social graph are utilized. Also, VOs collected observations are shared to save the service time for probing each VO. To view this, Figure 16 shows the overall time required for services to execute their tasks. An important factor to notice here is that the overall task completion time is also affected when social links are used to collect observations from VOs that are used in the service execution.

5. Conclusion

The Internet of Things is bringing the next technological revolution by connecting billions of objects on the earth and providing intelligent IoT services. However, it also carries two most important challenges: the first one is the complexity of handling a huge number of heterogeneous objects and the second one is how to deal with the monolithic approaches for providing services over existing IoT infrastructure. To address these challenges, we have proposed an integrated design based on the principles of SIoT, where the complexity has been handled using socially connected web objects; to overcome the monolithic approaches, microservices that can compose new service features independently have been adopted. In the proposed design, a social relationship model has been presented, which enables the efficient discovery of web objects and reduces the complexity of service provisioning process with the algorithms to discover and compose web objects. A semantic ontology model has also been developed to realize the interoperable social interaction among heterogeneous objects. Finally, a prototype implementation based on a use case scenario has been demonstrated; to evaluate the system with respect to the performance issues, experimental results pertaining to the VO discovery and service execution time have been rendered.

Details of Notations Used in Algorithms

$\Sigma\phi$: Service object ontologies based on semantic representation where ϕ is a replacement for M_μ which represents microservices M_c which represents CVO and M_v are VO ontologies

$\mathbb{I}\mathbb{I}_L$: Collection or list of all relationships that are retrieved from service objects

$\mathbb{I}\mathbb{I}_U$: List of the objects that do not have a relationship or are not associated

$\hat{S}$: Social relationship graph of objects

U: Correlation union of templates

$\underline{\mu}$₸: All service objects returned in response to the query request

₸: All objects returned in response to the query request

μ: The single instance iterated from the collection of microservices

M: Data model based on the specified ontology, where ΣM_s represents service data model, ΣM_c represents context data model, and user profile model is represented as ΣM_u

Q: SPARQL query to retrieve the available service templates for user rating

Q': SPARQL query to retrieve the relationship associated with services

$\mathbb{R}_e$: Registry entries of service objects

C_R: Relationship associated with CVO

V_R: Relationship associated with VO

R_c: All possible relationships that can be associated with CVO

R_v: All possible relationships that can be associated with VO

O_i: Iterator object instance for iterating the list of CVO and VO objects

$\mathbb{Z}$: Set of objects' social relationships

σi: Iteration item of the list of service objects

λi: An instance of iteration items in the list of service objects

$\mathbb{I}\mathbb{I}_{\mathrm{Lm}}$: List of all matched service items

ϖi: An instance of iteration items in the list of matched service items

$\mathbb{III}_{\text{LR}}$: List of all matched ranking
∂: Ranking value assigned to a service object
σj: Iterator item for the ranking list
κ: Threshold to rank a service object
Wf: Workflow for the composition of service objects
Pq: Priority queue to store ranked object instances
R∂: Assigned relationship ranking.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by Korea Institute for Advancement of Technology (KIAT) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (N040800001, Development of Web Objects Enabled EmoSpaces Service Technology).

References

[1] M. Nitti, V. Pilloni, G. Colistra, and L. Atzori, "The virtual object as a major element of the internet of things: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1228–1240, 2016.

[2] J. Surowiecki, *The Wisdom of Crowds*, Anchor Books, 2005.

[3] L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The social internet of things (SIoT)—when social networks meet the internet of things: concept, architecture and network characterization," *Computer Networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

[4] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between internet of things and social networks: Review and research challenges," *IEEE Internet of Things Journal*, vol. 1, no. 3, pp. 206–215, 2014.

[5] L. Atzori, D. Carboni, and A. Iera, "Smart things in the social loop: Paradigms, technologies, and potentials," *Ad Hoc Networks*, vol. 18, pp. 121–132, 2014.

[6] L. Atzori, I. Antonio, and G. Morabito, "Making things socialize in the Internet -Does it help our lives?" in *Kaleidoscope 2011: The Fully Networked Human?-Innovations for Future Networks and Services (K-2011), Proceedings of ITU. IEEE*, 2011.

[7] L. Atzori, A. Iera, and G. Morabito, "SIoT: giving a social structure to the internet of things," *IEEE Communications Letters*, vol. 15, no. 11, pp. 1193–1195, 2011.

[8] A. M. Vegni and V. Loscrí, "A survey on vehicular social networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, article no. A3, pp. 2397–2419, 2015.

[9] L. Krause, *Microservices: Patterns and Applications: Designing Fine-Grained Services by Applying Patterns*, 2015.

[10] A. Krylovskiy, M. Jahn, and E. Patti, "Designing a smart city internet of things platform with microservice architecture," in *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud (FiCloud '15)*, pp. 25–30, Rome, Italy, August 2015.

[11] B. Butzin, F. Golatowski, and D. Timmermann, "Microservices approach for the internet of things," in *Proceedings of the 21st IEEE International Conference on Emerging Technologies and Factory Automation, ETFA 2016*, Germany, September 2016.

[12] N. Dragoni, I. Lanese, S. T. Larsen, M. Mazzara, R. Mustafin, and L. Safina, "Microservices: How To Make Your Application Scale," *Computing Research Repository*, 2017.

[13] "Microservices Guide Pages - Martin Fowler," 2017, https://martinfowler.com/articles/microservices.html.

[14] "Web of Objects. ITEA 3 · Project · 10028 , D2.1: State-of-the-Art Relevant to the Web of Objects," 2017, https://itea3.org/project/web-of-objects.html.

[15] "Y.4452: Functional framework of web of objects," 2017, http://www.itu.int/rec/T-REC-Y.4452-201609-P.

[16] L. E. Holmquist, F. Mattern, B. Schiele, P. Alahuhta, M. Beigl5, and H.-W. Gellersen, "Smart-its friends: a technique for users to easily establish connections between smart artefacts," in *Ubicomp 2001: Ubiquitous Computing*, pp. 116–122, Springer, Berlin, Germany, 2001.

[17] D. Guinard, M. Fischer, and V. Trifa, "Sharing using social networks in a composable Web of Things," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2010*, pp. 702–707, Germany, April 2010.

[18] M. Kranz, L. Roalter, and F. Michahelles, "Things that twitter: social networks and the internet of things.' What can the Internet of Things do for the Citizen (CIoT)," in *Proceedings of the 8th International Conference on Pervasive Computing (Pervasive 2010)*, 2010.

[19] R. Girau, S. Martis, and L. Atzori, "Lysis: A platform for iot distributed applications over socially connected objects," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 40–51, 2017.

[20] A. Pintus, D. Carboni, and A. Piras, "Paraimpu: A platform for a social Web of Things," in *Proceedings of the 21st Annual Conference on World Wide Web, WWW'12*, pp. 401–404, France, April 2012.

[21] I. Farris, R. Girau, L. Militano et al., "Social Virtual Objects in the Edge Cloud," *IEEE Cloud Computing*, vol. 2, no. 6, pp. 20–28, 2015.

[22] R. Girau, M. Nitti, and L. Atzori, "Implementation of an experimental platform for the social internet of things," in *Proceedings of the 7th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2013*, pp. 500–505, July 2013.

[23] M. A. Jarwar, R. A. Abbasi, M. Mushtaq et al., "CommuniMents: A framework for detecting community based sentiments for events," *International Journal on Semantic Web and Information Systems*, vol. 13, no. 2, pp. 87–108, 2017.

[24] S. Newman, "Building microservices: designing fine-grained systems," 2015.

[25] "The Single Responsibility Principle - Programmer 97-things," 2017, http://programmer.97things.oreilly.com/wiki/index.php/The_Single_Responsibility_Principle.

[26] F. Viktor, "The DevOps 2.0 Toolkit: Automating the Continuous Deployment Pipeline with Containerized Microservices," 2016.

[27] "iCore: Internet Connected Objects for Reconfigurable Ecosystems, European FP7 Project," 2017, http://cordis.europa.eu/project/rcn/100873_en.html.

[28] D. Namiot and M. Sneps-Sneppe, "On microservices Architecture," *International Journal of Open Information Technologies*, vol. 2, no. 9, 2014.

[29] D. Bonino, M. T. D. Alizo, A. Alapetite et al., "ALMANAC: internet of things for smart cities," in *Proceedings of the 3rd*

International Conference on Future Internet of Things and Cloud (FiCloud '15), pp. 309–316, IEEE, Rome, Italy, August 2015.

[30] "Developing Microservices for PaaS with Spring and Cloud Foundry," 2017, https://www.infoq.com/presentations/microservices-pass-spring-cloud-foundry.

[31] "Microservices in action, Part 2: Containers and microservices — a perfect pair," 2017, https://www.ibm.com/developerworks/cloud/library/cl-bluemix-microservices-in-action-part-2-trs/index.html.

[32] J. E. Kim, X. Fan, and D. Mosse, "Empowering End Users for Social Internet of Things," in *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation - IoTDI '17*, pp. 71–82, Pittsburgh, Pa, USA, April 2017.

[33] Z. U. Shamszaman, S. S. Ara, I. Chong, and Y. K. Jeong, "Web-of-Objects (WoO)-based context aware emergency fire management systems for the Internet of Things," *Sensors*, vol. 14, no. 2, pp. 2944–2966, 2014.

[34] S. Ali, M. G. Kibria, and I. Chong, "WoO enabled IoT service provisioning based on learning user preferences and situation," in *Proceedings of the 2017 International Conference on Information Networking (ICOIN)*, pp. 474–476, Da Nang, Vietnam, January 2017.

[35] M. A. Jarwar, S. Ali, M. G. Kibria, S. Kumar, and I. Chong, "Exploiting interoperable microservices in web objects enabled Internet of Things," in *Proceedings of the Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 49–54, Milan, Italy, July 2017.

[36] M. Kibria, S. Ali, M. Jarwar, S. Kumar, and I. Chong, "Logistic Model to Support Service Modularity for the Promotion of Reusability in a Web Objects-Enabled IoT Environment," *Sensors*, vol. 17, no. 10, p. 2180, 2017.

[37] M. Kibria, S. Fattah, K. Jeong, I. Chong, and Y. Jeong, "A User-Centric Knowledge Creation Model in a Web of Object-Enabled Internet of Things Environment," *Sensors*, vol. 15, no. 9, pp. 24054–24086, 2015.

[38] M. A. Jarwar and . Ilyoung Chong, "Exploiting IoT services by integrating emotion recognition in Web of Objects," in *Proceedings of the 2017 International Conference on Information Networking (ICOIN)*, pp. 54–56, Da Nang, Vietnam, January 2017.

[39] S. Ali, H.-S. Kim, and I. Chong, "Implementation model of WoO based smart assisted living IoT service," in *Proceedings of the International Conference on Information and Communication Technology Convergence, ICTC 2016*, pp. 816–818, October 2016.

[40] S. Kumar, M. G. Kibria, S. Ali, M. A. Jarwar, and I. Chong, "Smart spaces recommending service provisioning in WoO platform," in *Proceedings of the International Conference on Information and Communications (ICIC)*, pp. 311–313, Hanoi, Vietnam, June 2017.

[41] K. M. Alam, M. Saini, and A. El Saddik, "Toward social internet of vehicles: concept, architecture, and applications," *IEEE Access*, vol. 3, pp. 343–357, 2015.

[42] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java," 2017, https://www.cs.waikato.ac.nz/ml/weka/.

[43] "Jena, Apache. &quot;A free and open source Java framework for building Semantic Web and Linked Data applications.&quot;," 2017, http://jena.apache.org/.

[44] Protégé, "A free, open-source ontology editor and framework for building intelligent systems," 2017, https://protege.stanford.edu/.

Journal of Engineering
The Scientific World Journal
International Journal of Rotating Machinery
Journal of Sensors
International Journal of Distributed Sensor Networks
Advances in Civil Engineering
Journal of Control Science and Engineering
Journal of Robotics
Journal of Electrical and Computer Engineering
Advances in OptoElectronics
VLSI Design
International Journal of Navigation and Observation
Modelling & Simulation in Engineering
International Journal of Aerospace Engineering
International Journal of Chemical Engineering
International Journal of Antennas and Propagation
Active and Passive Electronic Components
Shock and Vibration
Advances in Acoustics and Vibration

Hindawi

Submit your manuscripts at
https://www.hindawi.com

Multiple Antenna Technologies

Manar Mohaisen | YuPeng Wang | KyungHi Chang

The Graduate School of Information Technology and Telecommunications

INHA University

ABSTRACT

Multiple antenna technologies have received high attention in the last few decades for their capabilities to improve the overall system performance. Multiple-input multiple-output systems include a variety of techniques capable of not only increase the reliability of the communication but also impressively boost the channel capacity. In addition, smart antenna systems can increase the link quality and lead to appreciable interference reduction.

I. Introduction

Multiple antennas technologies proposed for communications systems have gained much attention in the last few years because of the huge gain they can introduce in the communication reliability and the channel capacity levels. Furthermore, multiple antenna systems can have a big contribution to reduce the interference both in the uplink and the downlink by employing smart antenna technology.

To increase the reliability of the communication systems, multiple antennas can be installed at the transmitter or/and at the receiver. Alamouti code is considered as the simplest transmit diversity scheme while the receive diversity includes maximum ratio, equal gain and selection combining methods. Recently, cooperative communication was deeply investigated as a mean of increasing the communication reliability by not only considering the mobile station as user but also as a base station (or relay station). The idea behind multiple antenna diversity is to supply the receiver by multiple versions of the same signal transmitted via independent channels.

On the other hand, multiple antenna systems can tremendously increase the channel capacity by sending independent signals from different transmit antennas. BLAST spatial multiplexing schemes are a good example of such category of multiple antenna technologies that boost the channel capacity.

In addition, smart antenna technique can significantly increase the data rate and improve the quality of wireless transmission, which is limited by interference, local scattering and multipath propagation. Through shaping the antenna radiation pattern and adaptively adjusting the antenna weight vector, smart antennas improve the communication link quality by increasing the received signal power and suppressing the interference.

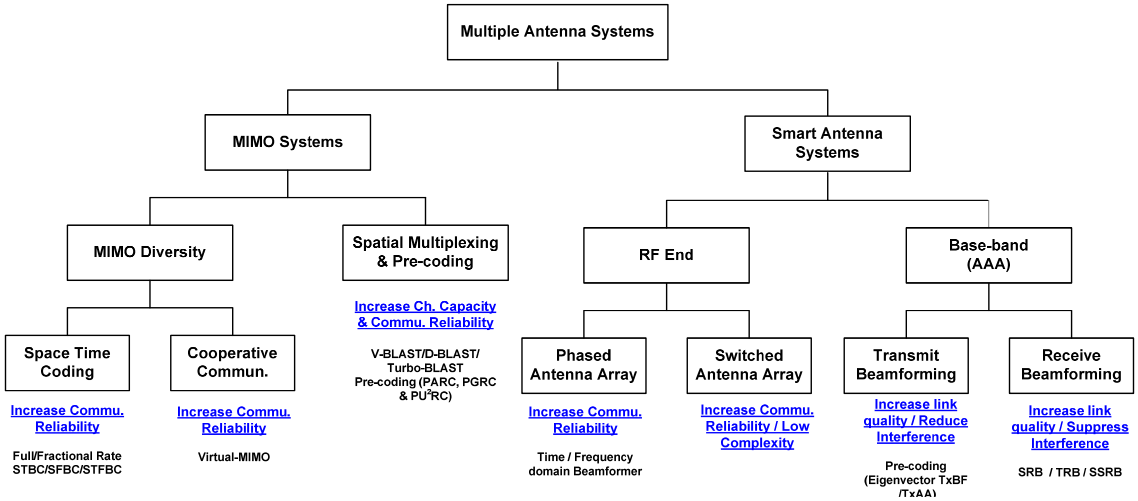
Multiple Antenna Systems
MIMO Systems
Smart Antenna Systems
MIMO Diversity
Spatial Multiplexing & Pre-coding
RF End
Base-band (AAA)
Increase Ch. Capacity & Commu. Reliability
V-BLAST/D-BLAST/ Turbo-BLAST Pre-coding (PARC, PGRC & PU²RC)
Space Time Coding
Cooperative Commun.
Phased Antenna Array
Switched Antenna Array
Transmit Beamforming
Receive Beamforming
Increase Commu. Reliability
Full/Fractional Rate STBC/SFBC/STFBC
Increase Commu. Reliability
Virtual-MIMO
Increase Commu. Reliability
Time / Frequency domain Beamformer
Increase Commu. Reliability / Low Complexity
Increase link quality / Reduce Interference
Pre-coding (Eigenvector TxBF /TxAA)
Increase link quality / Suppress Interference
SRB / TRB / SSRB

Fig. 1. Multiple antenna technologies.

Besides, on-line calibration technique is also adopted to correct the errors due to the distortions and nonlinearity of the radio frequency components in the antenna array system.

Fig. 1 summarizes the different multiple antenna technologies and gives some examples of these technologies.

This paper is organized as follows: in section II we present multiple antenna diversity schemes employed at the transmitter or/and at the receiver. Spatial multiplexing presented by BLAST schemes is detailed in section III. Section IV is dedicated to some advanced multiple input multiple-output (MIMO) systems including multi-user MIMO and cooperative communications. While techniques related to the smart antennas such as phased antenna array, switched beam antenna array, and adaptive antenna array are described in Section V. Finally, we conclude in Section VI.

II. MIMO Diversity

In communication systems, we have to increase the reliability of the communication operation between transmitter and receiver while maintaining a high spectral efficiency. The ultimate solution relies in the use of *diversity*, which can be viewed as a form of redundancy [1]. There are many diversity techniques that can be applied to communication systems; we mention herein time diversity, frequency diversity, and spatial diversity or any combination of these three diversities. In *time diversity*, the same information-bearing signal is transmitted in different time slots where a good gain can be achieved when the duration between the two slots, in which the same symbol is transmitted, is greater than the coherence time of the channel. In *frequency diversity*, the same information-bearing signal is transmitted on different subcarriers where a good diversity gain can be achieved when the separation between subcarriers is greater than the coherence bandwidth.

Finally, in *spatial diversity*, the same information-bearing signal is transmitted or received via different antennas where the maximum gain can be achieved when the fading occurring in the channel is independent (or low correlated). In the receiver, diversity gain can be achieved by combining the redundant signals arriving via independent (or lowly correlated) channels.

Fig. 2 shows some possible combinations of transmit diversity which can be achieved when employing multiple transmit antennas.

In the following section, we present some famous space-time block codes applied at the transmitter side. We present also the combining techniques used when different versions of the information-bearing signal are received. Finally, we present a scheme that includes transmit and receive diversities.

2.1 Space Diversity at the Transmit Side

The basic idea of the use of transmit diversity is to reduce the mobile station (MS) receiver complexity while improving the detection performance.

The pioneering work in the transmit diversity was done by Alamouti where he proposed his famous 2×1 space-time code. Alamouti scheme achieves diversity gain while requiring only a linear decoder.

Later on, Tarokh et al. proposed a generalized theory of the complex orthogonal space-time codes. Based on Tarokh work, more than two antennas can be used and the code rate can be fractional.

In the following we present the two different types of space-time codes.

2.1.1 Complex Orthogonal Space-Time Codes

For this type of space time codes, the following conditions must be satisfied

- Square transmission matrix (number of transmit antennas N_t equal to number of used time slots m)
- A unity code rate (number of used time slots m equals to number of transmitted symbols l)
- Orthogonality of the transmission matrix in the time and space domains ($\mathrm{SS}^{\mathrm{H}} = \mathrm{S}^{\mathrm{H}}\mathrm{S}$) where S^{H} is the conjugate transpose of S.

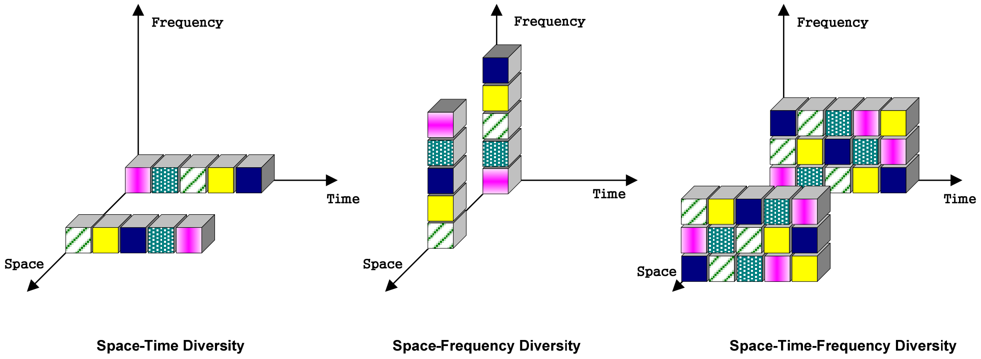

Fig. 2. Transmit diversity.

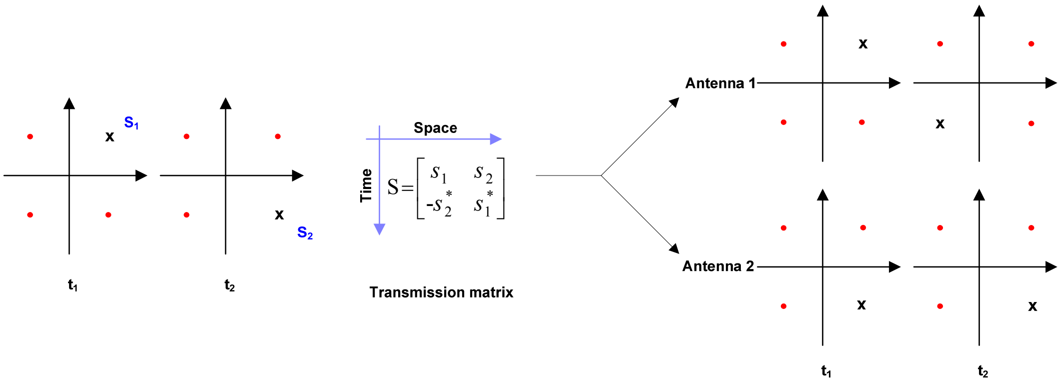
S1
x
S2
x
t1
t2
Space
Time
$S=\begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$
Transmission matrix
Antenna 1
Antenna 2
x
t1
t2

Fig. 3. Alamouti scheme example with QPSK modulation.

As said before, the simplest complex orthogonal space-time code is the Alamouti code which uses two transmit antennas and one receive antenna. Furthermore, Alamouti scheme requires that the fading channel envelope remains constant over two time slots.

Fig. 3 shows an example of the encoding process of Alamouti scheme with QPSK modulation [2], [3].

Fig. 4 shows the receiver structure used for decoding the combined received symbols.

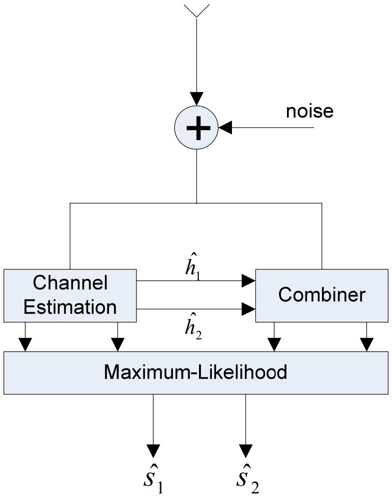
noise
+
Channel Estimation
Combiner
$\hat{h}_1$
$\hat{h}_2$
Maximum-Likelihood
$\hat{s}_1$
$\hat{s}_2$

Fig. 4. Alamouti code receiver.

At the receiver the following signals are received (with applying the complex conjugate to the received signal at t_2)

$$\begin{bmatrix} y_1 \\ y_2^* \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \end{bmatrix} \qquad (1)$$

The linear combiner multiplies the received symbols by the Hermitian transpose of the channel matrix (for simplicity, we consider that channel is perfectly estimated). The output of the linear combiner is then given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_1^2 + h_2^2 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad (2)$$

Maximum-Likelihood (ML) decoder is then applied to get the transmitted symbols. As one can see, the simplicity of the receiver is due to the spatio-temporal orthogonality of the transmission matrix.

A complex orthogonal space-time code using 4 or 8 antennas was proposed by Tarokh et al. in [4].

2.1.2 Generalized Complex Orthogonal Space-Time Codes

The search for space-time codes with more than two antennas was started by Tarokh, Jafarkhani, and Calderbank. Their work has built the basis for a theory of generalized complex orthogonal designs.

Generalized complex orthogonal designs are distinguished from Alamouti code by the following

- A non-square transmission matrix (number of used time slots ≠ number of Tx antennas)
- A fractional code rate (number of transmitted symbols < number of used time slots)
- Orthogonality of the transmission matrix is only guaranteed in the time sense.

As a consequence of these characteristics, the spectral efficiency is reduced and the number of time slots over which the channel should be constant is increased.

The transmission matrix of a generalized complex space-time code with 3 antennas, 4 transmitted symbols and 8 used time slots is given by [5]

$$\mathbf{G}_3 = \begin{bmatrix} s_1 & s_2 & s_3 \\ -s_2 & s_1 & -s_4 \\ -s_3 & s_4 & s_1 \\ -s_4 & -s_3 & s_2 \\ s_1^* & s_2^* & s_3^* \\ -s_2^* & s_1^* & -s_4^* \\ -s_3^* & s_4^* & s_1^* \\ -s_4^* & -s_3^* & s_2^* \end{bmatrix} \qquad (3)$$

In the literature, more research was done to increase the rate of the space-time codes. For more details refer to [6].

Table 1 summarizes the difference between Alamouti and space-time code characterized by the transmission matrix $\mathbf{G}_3$.

These shown coding schemes can be transmitted in the space-time domains, space-frequency domains or in space-frequency-Time domains. These coding schemes are thus known as ST, SF, and STF coding, respectively [7].

2.1.3 Cyclic Delay Diversity (CDD)

CDD can be considered as a very simple transmit diversity scheme. CDD can achieve transmit artificial frequency diversity by selecting appropriate transmit delays. In this method, multiuser diversity, obtained by scheduling based on frequency domain channel response, can be improved by adjusting the delay spread (at the transmitter) which is done by controlling the delay values dependent on the channel condition [8], [9].

2.2 Space Diversity at the Receive Side

In space diversity at the receive side, multiple antennas are used in the receiver with sufficient spacing between antennas in such a way mutual correlation between antennas is reduced and as consequence diversity gain is increased [10]. To get diversity gain at the receiver, received signals from different antennas are combined. There are four combining methods, namely, select combining (SC), maximal-ratio combining (MRC), equal-gain combining (EGC), and square-law combining. The first three schemes are linear while the last requires a non-linear receiver.

Fig. 5 shows a simplified block diagram of the linear combining schemes which differ in the weighting vector **w.** In SC, the signal at the branch with maximum signal to noise ratio (SNR) is selected and other received signals are discarded.

The weighting vector $\mathbf{w} = (w_1, w_2, \ldots, w_M)$ is the N^{th} column of the identity matrix of size M where the N^{th} branch has the maximum SNR.

Table 1. Comparison between Alamouti and generalized complex space-time codes.

Space-time code	**Number of Tx antennas**	**Number of transmitted symbols, *l***	**Number of used time slots, *m***	**Orthogonality of Tx matrix**	**Rate = *l/m***
S	2	2	2	Spatio-temporal sense	1
G_3	3	4	8	Only temporal sense	1/2

Table 2. Comparison between diversity combining schemes.

Scheme	**Requiring CSI**	**Outage Probability F(x)**	**Application**
SC	No	$\left[1-e^{-x/\bar{y}_c}\right]^M$	No constraints
MRC	Yes	$1-e^{-x/\bar{y}_c}\sum_{k=0}^{M-1}\frac{1}{k!}\left(\frac{x}{\bar{y}_c}\right)^k$	No constraints
EGC	Yes	No closed form for M > 2	No constraints
SLC	No	-	FSK or DS-CDMA

As one can see, the SC scheme does not require any channel information except that of SNR. On the other hand, MRC and EGC schemes require the channel state information (CSI) or a part of it (channel envelope, phase, delay). MRC scheme weights the received signals according to their reliability; a more reliable signal has a high weight while a less reliable signal has a small weight. Also, the channel phase distortion is compensated. Finally, signals are aligned then combined. On the other hand, EGC scheme can be viewed as a simplified version of MRC where signals are weighted equally (i.e. the weighting vector **w** = $[1, 1, \ldots, 1]_M$) then aligned before being combined coherently. In practice, the phase at different branches can't be often estimated. So, EGC and MRC can't be employed. In such situations, square law combining (SLC) can be applied to obtain spatial diversity without requiring phase estimation. Unlike linear combining schemes, SLC scheme can only be applied to modulation schemes which preserve some sort of orthogonality including frequency-shift keying (FSK) or direct-sequence CDMA [5], [11].

Table 2 summarizes a comparison between combining schemes. It is known that, from a performance point of view, MRC is optimum and gives the best performance among the pre-described combining schemes.

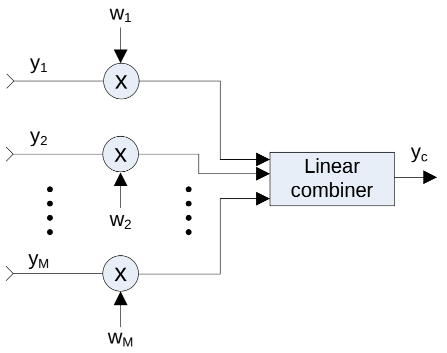

Fig. 5. Simple block diagram of linear combining schemes.

2.3 Combined Transmit/Receive Diversity

Spatial diversity schemes explained in the previous two sections can be combined together to achieve diversity at both receive and transmit sides. An example of such a hybrid spatial diversity scheme is the 2×2 Alamouti/MRC MIMO system [12].

III. Spatial Multiplexing

As shown in the previous section, MIMO diversity can be used in the transmitter or the receiver sides or in both to increase the reliability of the communication. In this section we talk about spatial multiplexing schemes which are for goal to increase the channel capacity.

The most known spatial multiplexing schemes are the BLAST family which includes Vertical-BLAST, Diagonal-BLAST, and Turbo-BLAST. The acronym BLAST stands for "*Bell Laboratories Layered Space-Time*".

3.1 Diagonal-BLAST

D-BLAST was originally proposed by Foschini [13]. In D-BLAST, the symbols to be transmitted are arranged on the diagonals of the space-time transmission matrix where elements under the diagonal are padded with zeros. Fig. 6-a depicts the structure of the D-BLAST transmitter for four transmit antennas. At first, the bit stream is de-multiplexed into four parallel streams which are encoded and modulated independently. Encoded-modulated streams are cycled over time. Equation (4) is an example of the transmission matrix when using four transmit antennas.

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & s_{1,3} & s_{1,4} & \cdots & s_{1,K-1} & s_{1,K} & 0 & 0 & 0 \\ 0 & s_{2,1} & s_{2,3} & s_{2,4} & \cdots & s_{2,K-2} & s_{2,K-1} & s_{2,K} & 0 & 0 \\ 0 & 0 & s_{3,1} & s_{3,2} & \cdots & s_{3,K-3} & s_{3,K-2} & s_{3,K-1} & s_{3,K} & 0 \\ 0 & 0 & 0 & s_{4,1} & \cdots & s_{4,K-4} & s_{4,K-3} & s_{4,K-2} & s_{4,K-1} & s_{4,K} \end{bmatrix} \qquad (4)$$

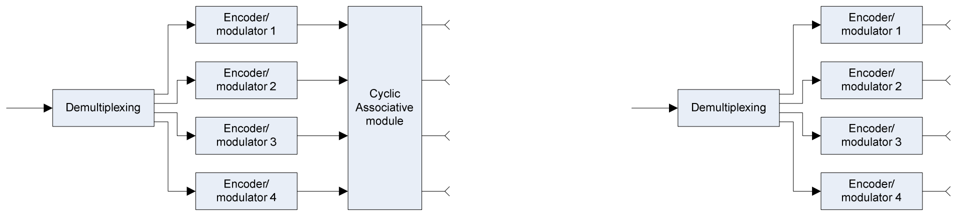
Demultiplexing
Encoder/ modulator 1
Encoder/ modulator 2
Encoder/ modulator 3
Encoder/ modulator 4
Cyclic Associative module
Demultiplexing
Encoder/ modulator 1
Encoder/ modulator 2
Encoder/ modulator 3
Encoder/ modulator 4

(a) D-BLAST

(b) V-BLAST

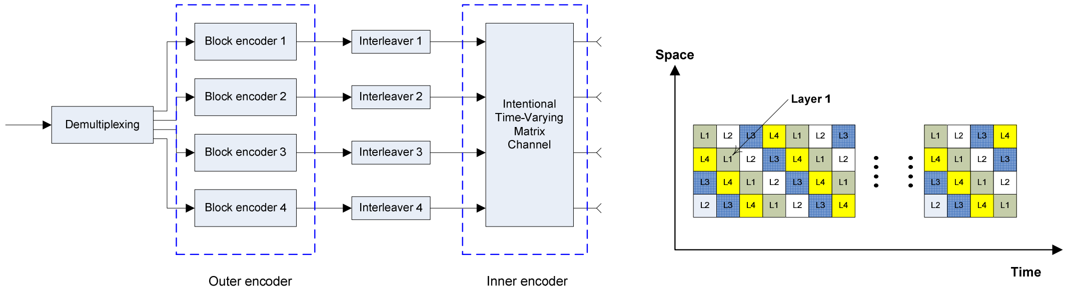
Demultiplexing
Block encoder 1
Block encoder 2
Block encoder 3
Block encoder 4
Interleaver 1
Interleaver 2
Interleaver 3
Interleaver 4
Intentional Time-Varying Matrix Channel
Outer encoder
Inner encoder
Space
Layer 1
L1 L2 L3 L4 L1 L2 L3
L4 L1 L2 L3 L4 L1 L2
L3 L4 L1 L2 L3 L4 L1
L2 L3 L4 L1 L2 L3 L4
L1 L2 L3 L4
L4 L1 L2 L3
L3 L4 L1 L2
L2 L3 L4 L1
Time

(c) Turbo-BLAST

(d) RLST code

Fig. 6. Transmitter block diagrams for BLAST family using four transmit antennas.

Table 3. MIMO systems diversity orders.

MIMO Configuration	**Diversity order**
STBC	$N_t \times N_r$
BLAST	$N_r - N_t + 1$

The first diagonal of S is transmitted via the first antenna; the second diagonal is transmitted via antenna 2, and so on.

3.2 Vertical-BLAST

A simplified version of D-BLAST was proposed by Wolniansky known as Vertical-BLAST or V-BLAST [14]. In V-BLAST, incoming data stream is de-multiplexed into N_t streams each of which is encoded and modulated independently and sent on an antenna of its own. V-BLAST high-level diagram is depicted in Fig. 6-b where four antennas are used at the transmit side. Compared to D-BLAST, V-BLAST does not include cycling over time, the complexity is significantly reduced. In addition, unlike D-BLAST, V-BLAST does not include any space-time wastage. At the receiver, transmitted symbols can be decoded using *ordered serial interference-cancellation (OSIC) detector*. For the OSIC to work properly, the number of receive antennas N_r must be at least as large as the number of transmit antennas.

3.3 Turbo-BLAST

Turbo-BLAST was first described by Sellathurai and Haykin [15]. The Turbo-BLAST transmitter structure is depicted in Fig. 6-c. The data stream bits are firstly demultiplexed into N_t parallel streams which are encoded independently using the block encoder (outer encoder) (i.e. channel coding). The output streams of the outer encoder are interleaved independently and passed to the inner encoder. The mission of the outer encoder is to achieve *random-layered space-time* (RLST) coding.

The structure of the RLST encoder, with periodical cyclic space-time interleaving is depicted in Fig. 6-d. For optimal performance of the RLST code, the receiver should employ the *maximum a posteriori probability (MAP)* decoding algorithm. Nevertheless, the complexity of the MAP decoding algorithm is very high (increases exponentially with N_t). To decrease the complexity of the receiver, the near-optimal turbo-like receiver can be used. This near-optimal turbo-like receiver is known as *iterative detection and decoding (IDD) receiver*.

Before going further, we list in Table 3 a comparison between diversity order of the different space-time coding and the BLAST family schemes.

IV. Advanced Topics

4.1 Single and Multi CodeWord MIMO

In single codeword (SCW) MIMO, an encoded packet is distributed across many streams to form the MIMO transmission. Feedback is used to control the rank of the MIMO transmission (number of streams used) as well as the overall rate of transmission. In multiple codeword MIMO, several separately encoded packets are transmitted independently over the multiple streams. Here the rate of each stream can be controlled with feedback [16] and [17].

4.2 Single-User MIMO and Multi-User MIMO

In single-user MIMO, already explained techniques in previous sections are used where the channel capacity grows linearly with *min(Nt, Nr)* [18].

For multi-user MIMO, which is of high interest research topic, it was shown that for N_t transmitting antennas (at the base station) and N_r users, the same overall capacity can be achieved. This later work was encouraged by applying dirty paper coding [19] where results showed that if the transmitter knows the interfering signal, then the channel capacity will not be affected by the presence of the interference [20]. On the other hand, multi-user MIMO can integrate beamforming to apply spatial division multiple access (SDMA).

4.3 Cooperative Communication and Virtual MIMO

In cooperative communication, a mobile can act as both a user and relay. As consequence, mobile sends to the base station its own data bits and some of other mobiles (sometimes called partner) information bits. Fig. 7 shows a cooperative cellular system where for simplicity we consider three cooperative users and one base station [21]. As depicted in Fig. 7, user 1 cooperates with users 2 and 3 to send its own information. As a result, the overall cooperative system can be seen as virtual-MIMO (V-MIMO) and in the above example it is $3\times N_{BS}$ MIMO system (for the uplink) where N_{BS} is the base station's number of receive antennas. Users 2 and 3 can simply amplify and forward user 1 received information or detect and forward [22]. Another method of cooperation is the coded cooperation where different coded portions are sent via different fading channels [23].

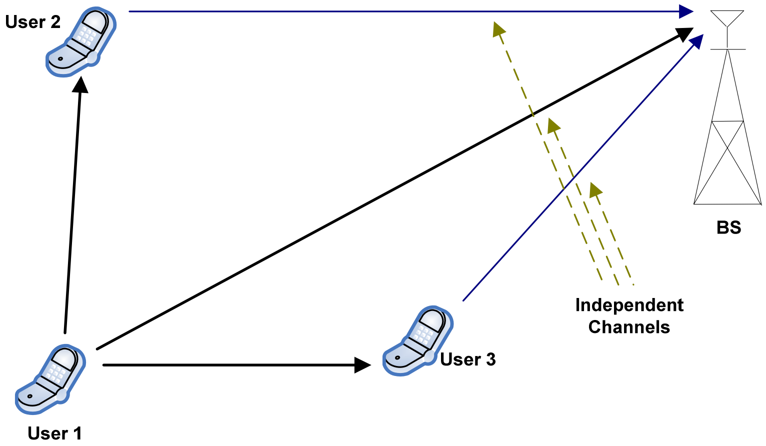

Fig. 7. Cooperative communication and virtual-MIMO.

4.4 Pre-Coded MIMO with Rank Adaptation

4.4.1 Per Antenna Rate Control (PARC)

PARC can be considered as a closed-loop MIMO system where transmitter uses *channel quality indication* (CQI) fed by the receiver to select the best modulation and coding schemes per antenna. Fig. (8) shows a general PARC transmitter structure with 4 transmit antennas [24].

4.4.2 Per Group Rate Control (PGRC)

In PARC, a CQI feedback is necessary for each transmit antenna. This increases the uplink overhead.

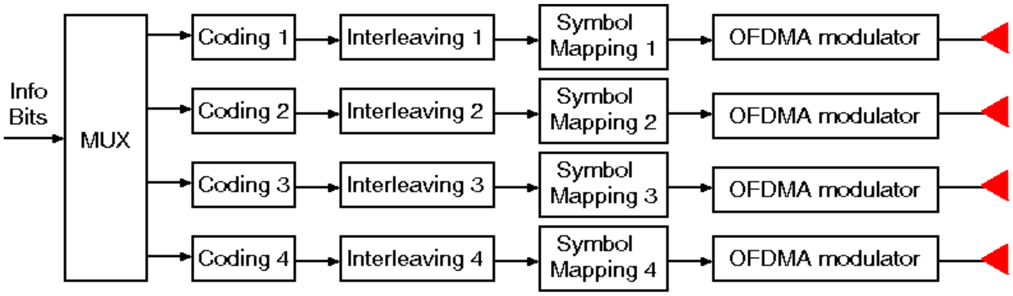

Fig. 8. Per Antenna Rate Control (PARC).

To solve this problem, PGRC is used where a feedback is required per group of antennas. This reduces the feedback information while maintaining almost the same performance of PARC [24].

4.4.3 Per User Unitary Rate Control (PU^{2}RC)

PU^{2}RC is a multi-user closed-loop MIMO system. Each user feeds back the CQI to the base station. The base station uses the CQI to determine the modulation and coding schemes per user. In addition, base station can apply unitary pre-coding and adaptively select the number of transmit antennas (rank adaption) [25].

V. Smart Antennas

5.1 Introduction

Smart antenna was born in the early 1990s when well developed adaptive antenna arrays originate from Radar system. Later, Smart antenna technique is applied in wireless communications system. Recently, Smart antenna technique has been proposed as a promising solution to the future generations of wireless communication systems, such as the Fourth-Generation mobile communication systems, broadband wireless access networks, where a wide variety of services through reliable high-data rate wireless channels are expected. Smart antenna technique can significantly increase the data rate and improve the quality of wireless transmission, which is limited by interference, local scattering and multipath propagation [26], [27]. Smart antennas offer the following main applications in high data-rate wireless communication systems [28], [29]:

- Spatial Diversity
- Co-channel interference reduction
- Angle reuse or space division multiple access (SDMA)
- Spatial multiplexing

Smart antenna system can be categorized into three main groups: Phased antenna array system, switched beam systems, and adaptive antenna array system. To match the characteristics in each radio frequency chain of the transmitter and receiver, on-line calibration is required in smart antenna systems. On-line calibration technique can compensate the errors such as the distortions of radio frequency components due to small environment changes, the nonlinear characteristics of mixer, amplifier and attenuator, I/Q imbalance errors, etc.

5.2 Phased Antenna Array System

Phased antenna array is a group of antennas in which the relative phases of the respective signals feeding the antennas are varied in such a way that the effective radiation pattern of the array is reinforced in a desired direction and suppressed in undesired directions.

Phased antenna array system is usually utilized in *radio frequency* (RF) or *intermediate frequency* (IF) with the system central frequency larger than 10 GHz, such as satellite communication system [30]. There are two main different types of phased arrays, also called beamformers. There are time domain beamformers and frequency domain beamformers.

5.3 Switched Beam System

The switched beam method is considered as an extension of the current sectorization scheme. In the switched beam approach, the sector coverage is achieved by multiple predetermined fixed beam patterns with the greater gain placed in the centre of a beam [30]. When a mobile user is in the vicinity of a beam, then the signals at the output ports will be given as in (5). This enables the switched beam system to select the signal from the output port corresponding to that beam. As the mobile moves to the coverage of another beam during the call, the system monitors the signal strength and switches to other output ports as required. A basic switched beam antenna architecture is shown in Fig. 9. And Fig. 10 illustrates the produced antenna pattern with 4 antennas.

$$y_i(t) = s_i(t)G_i(t)\sum_{l=1}^{L} I_l(t)G_{li}(\theta_l) \qquad (5)$$

where $y_i(t)$ is the total signal appearing at port i, $s_i(t)$ is the signal source, $I_l(t)$ is the interfering signal source located at arbitrary angles θ_l, G_i is the transfer function between signal source along the main beams and their corresponding output ports, G_{li} is the transfer function between interference signal l and port i.

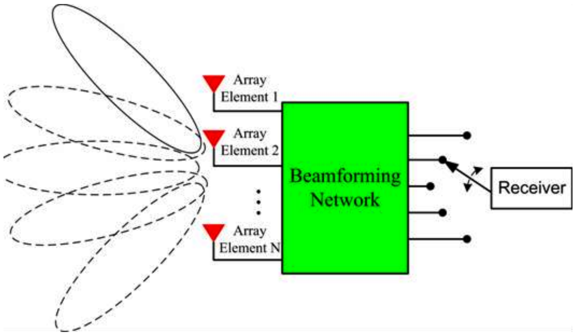

Fig.9. Functional block diagram of switched beam system.

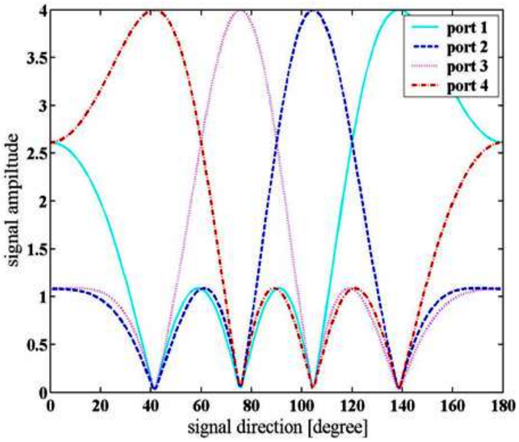

Fig.10. Produced antenna pattern of switched beam system with 4 antennas.

Switched beam systems can offer several advantages, including

- Low complexity and cost. Since switched beam system only requires a beamforming network, RF switches, and simple control logic, they are relatively easy and cheap to implement.
- Moderate interaction with base station receivers. In practice, switched beam system can simply replace conventional sector antennas without requiring significant modifications to the radio base station antenna interface or the baseband algorithms implemented at the receiver.
- Coverage extension. The antenna array aperture gain will boost the link

budget, which could be translated to a coverage extension.

5.4 Adaptive Antenna Array System (AAA)

Adaptive antennas date back to 1959. The original work was attributed to L. C. Van Atta's work, Electromagnetic Reflection. Since then, adaptive beamforming techniques have been employed to remove unwanted noise and jamming from the output, mainly in military applications. With the thriving commercial wireless communication industry and the advancing microprocessor technologies, the adaptive beamforming techniques have found their applications in commercial wireless communications. With powerful *digital signal processing* (DSP) hardware at the base-band, algorithms could control antenna beam patterns adaptively to the real signal environment, forming beams towards the desired signals while forming nulls to co-channel interferers. Thus, the system performance is optimized in terms of link quality and system capacity [31].

Adaptive antenna array can be utilized in the transmitter side, which is known as *transmit beamforming* (TxBF) or in the receiver side, which is called *receive beamforming* (RxBF).

5.4.1 Transmit Beamforming (TxBF)

The implementation of adaptive antenna array technique in a handset is difficult with today`s hardware due to its limitations in size, cost, and energy storage capability, while it is feasible to adopt antenna arrays at base stations. Transmit beamforming provides a powerful method for increasing downlink capacity [32]-[35]. The idea of TxBF is similar to the pre-coded MIMO technique but with different strategies to calculate the transmit weight vector. TxBF adjusts the antenna main lobe towards to the desired user and reduce the interference to other users. A simple illustration of TxBF is shown in Fig. 11.

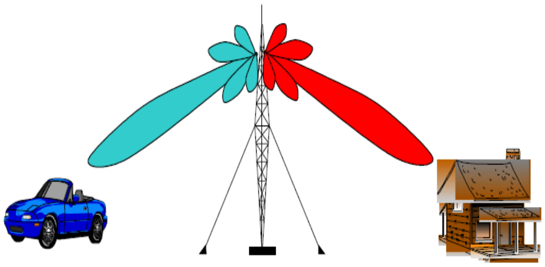

Fig.11. An illustration of TxBF.

Eigenvector TxBF Algorithm

Eignenvector TxBF algorithm is widely used for TxBF. The eigenvector of the spatial covariance channel matrix is calculated as

$$\mathbf{R}_{ss} = \boldsymbol{\lambda}\mathbf{H} \quad (6)$$

where $\mathbf{R}_{ss}$ is the autocovariance matrix of the desired user`s signal, and $\mathbf{H}$ is the spatial covariance channel matrix. The eigenvector $\boldsymbol{\lambda}_{max}$ which corresponds to the largest eigenvalue will be selected as the weight vector [36]. One example of beam pattern for 4 uniform linear array elements is shown in Fig 12.

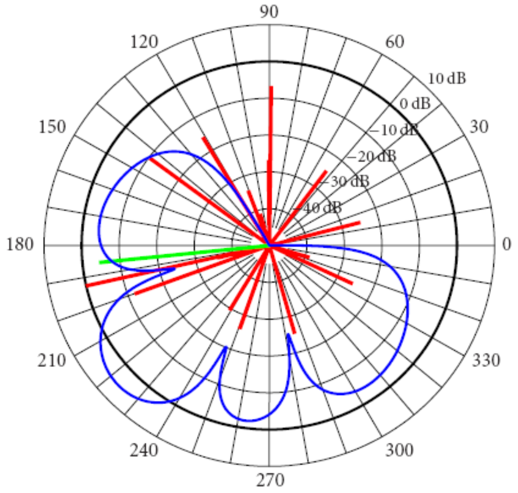

Fig. 12. Example beam pattern of 4 antenna elements in a sectorized system for a single sector (main beam direction is 240°).

Transmit Adaptive Array (TxAA) Algorithm

Transmit adaptive array (TXAA) is a technique in which the user periodically sends quantized estimates of the optimal transmit weights to the BS via a feedback channel. The transmitter weights are optimized to deliver maximum power to the user. The optimal transmit weights are given by

$$\mathbf{w} = \mathbf{H}^{H} / \mathbf{H}\mathbf{H}^{H} \qquad (7)$$

where **w** is the transmit weight vector and **H** is the channel matrix.

The weights are normalized so that the total transmitted power is not altered. In the case of multipath channels emanating from each antenna, the optimal weights will be given by the principal eigenvector of the channel correlation matrix $\mathbf{H}^{H}\mathbf{H}$.

5.4.2 Receive Beamforming (RxBF)

Beamforming also can be applied in the uplink to improve the link quality and suppress the co-channel interference, which is known as *receive beamforming* (RxBF). Through RxBF, smart antenna system can receive predominantly from a desired direction (direction of the desired source) compared to some undesired directions (direction of interfering sources). This implies that the digital processing has the ability to shape the radiation pattern to adaptively steer beams in the direction of the desired signals and put nulls in the direction of the interfering signals. This enable low co-channel interference and large antenna gain to the desired signal.

Based on the reference signals adopted in the beamforming algorithms, RxBF can be classified into *spatial reference beamforming* (SRB), *temporal reference beamforming* (TRB), and *signal structure reference beamforming* (SSRB).

Spatial Reference Beamforming (SRB)

Spatial reference beamforming method is sometimes referred as *direction of arrival* (DoA) method. SRB estimates the direction of arrival of the signal based on the spatial reference signal, using any of the techniques like *multiple signal classification* or *estimation of signal*

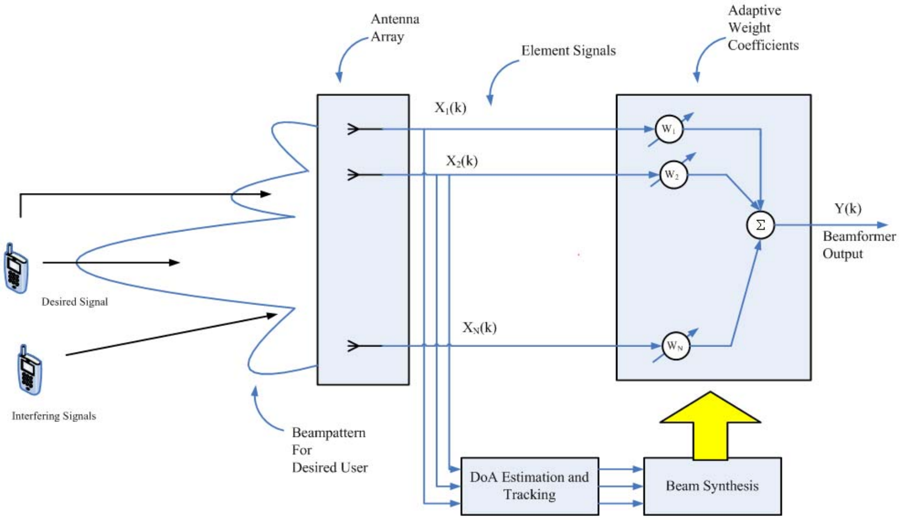
Antenna Array
Element Signals
Adaptive Weight Coefficients
$X_1(k)$
$X_2(k)$
$X_N(k)$
W_1
W_2
W_N
Σ
Y(k)
Beamformer Output
Desired Signal
Interfering Signals
Beampattern For Desired User
DoA Estimation and Tracking
Beam Synthesis

Fig. 13. A general structure of SRB.

parameters via rotational invariance techniques algorithms or their derivatives. They involve finding a spatial spectrum of the antenna/sensor array, and calculating the DoA from the peaks of this spectrum [37]. A general architecture of SRB algorithm is shown in Fig. 13. The general steps of SRB method are shown as follows:

- DoA Estimation
    - Arbitrary Array: MUSIC, etc.
    - Linear Array: ESPRIIT, etc.
- Beam Synthesis
    - Gram-Schmidt, etc.
- Combining
    - EGC, MRC, Wiener Filter, etc.

Multiple signal classification (MUSIC) algorithm estimates the DoA of the desired signal by using an eigen-space method based on a spatial reference signal. MUSIC requires intensive calculation of eigenvalues and eigenvectors of an autocorrelation matrix of the input vectors from the receiving antenna array. A general step of MUSIC algorithm is shown below:

- Collect received samples and estimate the covariance matrix of the received samples.
- Perform eigen-decomposition of the covariance matrix.
- Calculate spatial spectrum.
- Estimate DoA by locating peaks in the spectrum.

Estimation of signal parameters via rotational invariance techniques (ESPRIT) is also well known for the SRB method. In addition, ESPRIT has many important advantages over MUSIC algorithm [38]:

- No knowledge of the array geometry and element characteristics are required.
- Much less complex on computation.
- No calibration of the array is required.
- The algorithm simultaneously estimates the number of sources and DoA`s

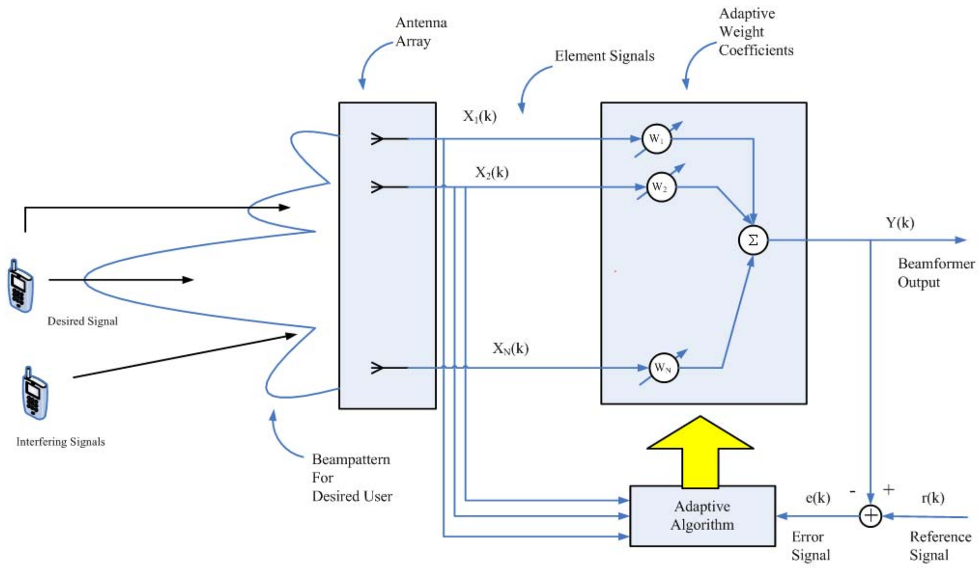
Antenna Array
Element Signals
Adaptive Weight Coefficients
X₁(k)
X₂(k)
X_N(k)
W₁
W₂
W_N
Σ
Y(k)
Beamformer Output
Desired Signal
Interfering Signals
Beampattern For Desired User
Adaptive Algorithm
e(k)
-
+
r(k)
Error Signal
Reference Signal

Fig. 14. A general structure of TRB.

Temporal Reference Beamforming (TRB)

Temporal Reference Beamforming shown in Fig. 14, is a method used to create the radiation patter of the antenna array by adding constructively the phases of the signals in the DoA of the desired user, and nulling the pattern of the interfering users based on the temporal reference signal [39]. Based on the temporal reference signal and some predefined adaptive weight calculation criterion, some adaptive algorithms such as LMS (Least Mean Square), RLS (Recursive Least Squares), and DSMI (Direct Sample Matrix Inverse) algorithms, are used to adjust the weight vector of the antenna array to improve the link quality. The general characteristics of TRB are as follows:

- Good performance in multipath channel environment
- Computationally inexpensive
- Requiring Training sequence
- Difficult to apply TxBF because of the absence of DoA information

Criterion of Adaptive Weight Calculation

In the *minimum mean-square error* (MMSE) criterion, the weights are chosen to minimize the *mean-square error* (MSE) between the beamformer output and the temporal reference signal. While in the *maximum signal-to-interference ratio* (MSIR) criterion, the weights are chosen to directly maximize the signal-to-interference ratio (SIR). And The *minimum variance* (MV) criterion chooses the weights that minimize the variance of the output power. All the above three criterions has the same form of

$$\mathbf{w}_{opt} = \beta \mathbf{R}_i^{-1} \mathbf{v} \qquad (8)$$

where $\mathbf{R}_i^{-1}$ is the inverse of the covariance matrix of the interference signals received in the antenna array and **V** is the antenna array propagation vector [40].

Let us assume that *d(t)* is the transmitted temporal reference signal and $\mathbf{R}_u$ is the covariance matrix of interference signals at the output of the beamformer. The calculation of β for MMSE, MSIR

Table 4. β Calculation.

Criterion	MMSE	MSIR	MV
β	$\frac{E\{d^2(t)\}}{1+E\{d^2(t)\}\mathbf{v}^H\mathbf{R}_i^{-1}\mathbf{v}}$	$\frac{E\{d^2(t)\}}{SIR}\mathbf{v}^H\mathbf{w}_{opt}$	$\frac{g}{\mathbf{v}^H\mathbf{R}_u^{-1}\mathbf{v}}$

and MV criterions are summarized in Table 4.

Adaptive Beamforming Algorithm

The *least mean square* (LMS) algorithm uses the temporal reference signal to update the weights at each iteration. In the LMS algorithm, we are searching for the optimal weight that would make the array output either equal or as close as possible to the reference signal, which is the weight that minimizes the MSE. Since the MSE has a quadratic form, moving the weights in the negative direction of the gradient of the MSE should lead us to the minimum of the error surface. The weight update equation is shown in (9) [30].

$$\mathbf{w}(t+1)=\mathbf{w}(t)-\mu\mathbf{x}(t+1)\varepsilon^* \quad (9)$$

where μ is a constant, called the step size, which determines how close the weights approach the optimum value after each iteration and it controls the convergence speed of the algorithm. And ε is the error signal between the temporal reference signal and the received signal at the beamformer output. $\mathbf{x}$(t+1) is the received signal vector at the antenna array at time *t+1*.

The main drawback of the LMS algorithm is that it is sensitive to the scaling of its input. This makes it very hard (if not impossible) to choose a step size μ that guarantees stability of the algorithm. The *normalized least mean square* (NLMS) algorithm is a variant of the LMS algorithm that solves this problem by normalizing with the power of the input. The weights updating function of NLMS algorithm is shown as

$$\mathbf{w}(t+1)=\mathbf{w}(t)+\frac{\hat{\mu}}{a+\|x(t)\|^2}\mathbf{x}(t)\varepsilon^* \quad (10)$$

Recursive least square (RLS) algorithm is derived to overcome the drawback of slow convergence speed in the LMS algorithm, when the eigenvalue spread of the correlation matrix R of received signal vector **x** is large. RLS algorithm replaces the step size μ with the inverse of R. The weights are then updated using (11).

$$\mathbf{w}(t+1)=\mathbf{w}(t)-\mathbf{R}^{-1}\mathbf{x}(t+1)\varepsilon^* \quad (11)$$

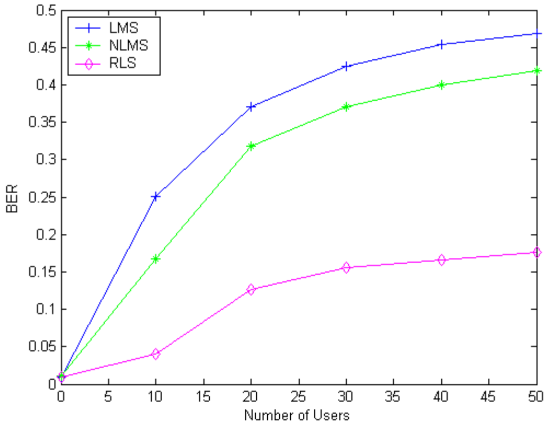

Fig. 15. Performance comparison among LMS, NLMS, and RLS algorithms.

Fig. 15 shows a simple performance comparison of the above three algorithms in OFDMA system under the Rayleigh fading channel with 8 antenna elements [41]. From this figure, we see that RLS algorithm performs best due to its faster convergence speed than LMS and NLMS algorithms.

Signal Structure Reference Beamforming (SSRB)

SSRB method is based on inherent structure of the transmit signal of the implicit kind reference signal. Algorithms such as blind beamforming, least squares, and constant modulus algorithms, are based on the SSRB method. SSRB method is robust against different propagation conditions and does not require the array manifold knowledge. But the convergence problem becomes the main drawback of the SSRB method.

VI. Conclusions

In this paper we introduced the multi antenna technologies which can be considered as one of the most vivid area of research. Multiple antenna technologies were categorized into two main groups where in the first group we introduced some techniques related to spatial diversity and spatial multiplexing by outlining the gain achieved by these schemes. Furthermore, we introduced the smart antenna techniques and the up-to-date research progress in this field.

The advantages of multiple antenna systems make of them a very strong candidate to increase link reliability, increase channel capacity and reduce interference in both uplink and downlink.

References

[1] A. Goldsmith, *Wireless Communications*, Cambridge Univ. Press, 2005.

[2] M. Alamouti, “A simple transmit diversity technique for wireless communications”, *IEEE J. Sel. Area Commun.,* 16, pp. 1451-1458, 1998.

[3] G. Tsoulos, *MIMO System Technology for Wireless Communications*, Taylor and Francis, 2006.

[4] V. Tarokh et al., “Space-time block codes from orthogonal designs”, *IEEE Trans. Inf. Theory,* vol. 45, no. 5, pp. 1456-1467, 1999.

[5] S. Haykin and M. Mohr, *Modern Wireless Communications,* Pearson Prentice Hall, 2005.

[6] W. Su and X-G. Xi, “Two generalized complex orthogonal space-time block codes of rates 7/11 and 3/5 for 5 and 6 transmit antennas”, *IEEE Trans. Inf. Theory,* vol. 49, no. 1, pp. 313-316, Jan 2003.

[7] K. Suto, and T. Ohtsuki, “Space-time-frequency block codes over frequency selective fading channels,” *IEICE Trans. Commun.,* vol. E87-B, no. 7, pp. 1939-45, July 2004.

[8] NTT DoCoMo, “Multi-degree cyclic delay diversity with frequency-domain channel dependent scheduling”, *3GPP TSG-RAN WG1 meeting #44bis*, R1-060991, Mar. 2003.

[9] Samsung, “Adaptive cyclic delay diversity,” *3GPP TSG-RAN1 43#*, R1-051354, 7th – 11st, Nov. 2005, Seoul, Korea.

[10] G. Stüber, *Principles of Mobile Comm-unication,* Kluwer, 2001.

[11] M. K. Simon and M-S. Alouini, *Digital Communication over Fading Channels*, Wiley, 2005.

[12] E. Biglieri et al., “Diversity, interference cancellation and spatial

multiplexing in MIMO mobile WiMAX systems", *to appear in IEEE WiMAX '07*, 2007.

[13] G. Foschini, "Layered space-time architecture for wireless communication in fading environment when using multi-element antennas", *Bell Labs Technical Journal,* Autumn 1996.

[14] P. Wolniansky et al., "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel", *URSI International Symposium on Signals,* Systems and Electronics, 1998.

[15] M. Sellathurai and S. Haykin, "TURBO-BLAST for wireless communications: theory and experiments", *IEEE Trans. Signal Processing,* vol. 50, no., 10 pp. 2538-2546, Oct. 2002.

[16] A. Hottinen et al., "Industrial embrace of smart antennas MIMO," *IEEE Wireless Commun. Mag.*, Aug. 2006.

[17] A. Jette et al. "UMBFDD candidate proposal for IEEE 802.20," IEEE C802.20-07/09, Mar. 2007.

[18] A. Goldsmith et al., "Capacity limits of MIMO channels," *IEEE J. on Sel. Areas in Commun.*, vol. 21, no. 5, pp. 684-702, Jun. 2003.

[19] M. Costa, "Writing on dirty paper," *IEEE Trans. Info. Theory*, vol. 29, no. 3, pp. 439-441, May 1983.

[20] Q. Spencer et al., "An introduction to the multi-user MIMO downlink," *IEEE Commun. Mag.,* pp. 60-67, Oct. 2004.

[21] A. Nosratinia and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Communications Magazine,* pp. 74-80, Oct. 2004.

[22] J. Laneman et al., "An efficient protocol for realizing cooperative diversity in wireless networks," *Proc. IEEE ISIT*, June 2001, pp.294.

[23] T. Hunter and A. Nosratinia, "Diversity through coded cooperation," *IEEE Trans. on Wireless Commun.,* vol. 5, no. 2, pp. 1-7, Feb. 2006.

[24] Texas Instruments, "MIMO OFDMA E-UTRA proposal for different antenna configurations," *3GPP TSG RAN1 WG1 #43*, R1-051315, Nov. 2005.

[25] Samsung, "Downlink MIMO for EUTRA," *3GPP TSG RAN1 WG1 #43*, R1-051353, Nov. 2005.

[26] S. Ohmori, Y. Yamao, and N. Nakajima, "The future generationsof mobile communications based on broadband access technologies," *IEEE Commun. Mag.,* vol. 38, pp. 134–142, Dec. 2000.

[27] K. Sheikh, D. Gesbert, D. Gore, and A. Paulraj, "Smart antennas for broadband wireless access networks," *IEEE Commun. Mag.,* vol. 37, pp. 134–142, Nov. 1999.

[28] A. Lozano, F.R. Farrokhi, and R.A. Valenzuela, "Lifting the limits on high-speed wireless data access using antenna arrays," *IEEE Commun. Mag.,* pp. 156–162, Sept. 2001.

[29] Jeffrey H. Reed," Smart antennas: A system level overview for software defined radios for creating an API", SDRF-04-I-0057-V0.00, Software Defined Radio Forum, Jan. 2004.

[30] Ahmed EI Zooghby, *Smart Antenna Engineering*, Artech House, 2005.

[31] Michael Chryssomallis, "Smart antennas", *IEEE Antennas and Propagation Mag.,* vol. 42, pp.129-136, , June 2000.

[32] H. Boche and M. Schubert, "Theoretical and experimental comparison of optimization criteria for downlink beamforming," *European*

Trans. on Telecomm., vol. 12, no. 5, pp. 417–426, 2001.

[33] R. M. Buehrer, A. G. Kogiantis, S.-C. Liu, J. Tsai, and D. Uptegrove, “Intelligent antennas for wireless communications -uplink,” *Bell Labs Technical Journal,* vol. 4, no. 3, pp. 73–103, 1999.

[34] H. Holma and A. Toskala, *WCDMA for UMTS*, John Wiley & Sons, 2000.

[35] A. Yener, R. D. Yates, and S. Ulukus, “Interference management for CDMA systems through power control, multiuser detection, and beamforming,” *IEEE Trans. on Comm.,* vol. 49, no. 7, pp. 1227–1239, 2001.

[36] S.J. Ko, J. Heo, and K.H. Chang, "An effective downlink resource allocation for supporting heterogeneous traffic data in an OFDM/SDMA-based cellular system," *in Proc. IEEE GLOBECOM*, Nov. 2006, WLC27-5.

[37] Lal C. Godara, “Application of antenna arrays to mobile communications, part II: beam-forming and direction-of-arrival considerations”, *IEEE Proc.,* vol .85, pp.1195-1245, Aug. 1997.

[38] Haardt, M., Nossek, J.A., “Unitary ESPRIT: how to obtain increased estimation accuracy with a reduced computational burden,” *IEEE Trans. on Signal Processing,* vol.43, pp. 1232 -1242, 1995.

[39] B. L. P. Cheung, Simulation of Adaptive Array Algorithms for OFDM and Adaptive Vector OFDM Systems, Master of Science in Electrical Engineering of Virginia Polytech., Sept. 2002.

[40] John Litva and Titus Kwok-Yeung Lo, *Beamforming in Wireless Communication*, Artech House, 1996.

[41] J. Heo and K.H. Chang, "Transmit and receive beamforming for OFDMA/TDD system," *in Proc. ISAP*, Aug. 2005, pp.27-30.

Manar Mohaisen

July, 2001: BS, Communication and Control, University of Gaza, Gaza, Palestine

Sep., 2005: MS，The School Polytechnic of Nice University, Sophia-Antipolis, France

Feb., 2006 ~ Present: Ph.D. student, The Graduate School of IT & T, INHA University

2001 ~ 2003: The Palestinian Telecommunication Company (JAWWAL)

<Rresearch Interests> MIMO Detection and Co-Channel Interference Cancellation

Yupeng Wang

July, 2004: BS, Communication Engineering, Northeastern University, Shenyang, China

July, 2006: MS, The Graduate School of IT & T, INHA University

Sep., 2006 ~ Present: Ph.D. student, The Graduate School of IT & T, INHA University

<Research Interests> 3GPP LTE Systems, Radio Resource Management, MIMO Techniques, and UWB

KyungHi Chang

Feb., 1985: BS, Electronics Engineering, Yonsei University

Feb., 1987: MS, Electronics Engineering, Yonsei University

Aug., 1992: Ph.D., EE Dept., Texas A&M Univ.

1989 ~ 1990: A Member of Research Staff, Samsung Advanced Institute of Technology (SAIT)

1992 ~ 2003: A Principal Member of Technical Staff (Team Leader), Electronics and Telecommunications Research Institute (ETRI)

2003 ~ Present: Associate Professor, The Graduate School of IT & T, INHA University

<Research Interests> RTT design for IMT-Advanced & 3GPP LTE Systems, WMAN System Design, Cognitive Radio, Cross-layer Design, and Cooperative Relaying System

autorité de régulation
des communications électroniques
et des postes

RÉPUBLIQUE FRANCAISE

5G: ISSUES & CHALLENGES

March 2017

The issues and challenges surrounding 5G

The telecoms industry is currently in the process of designing the technologies that are due to take over from 4G, which is still being deployed today. A great deal of work is thus underway to prepare these “5G” technologies.

To prepare for the arrival of this new generation of technologies, Arcep wanted to take a detailed look at the industry to better understand what is in the works. This report is the fruit of the interviews and research that Arcep conducted over the course of 2016, and which the Authority wanted to publish as a way to contribute to the public debate over 5G.

Its aim it to provide as objective and exhaustive an overview as possible, and deliver a concise, informative snapshot of the work that is currently underway on the future generation of mobile networks.

This report reflects the views of the stakeholders who were interviewed, but in no way represents Arcep’s positions on or roadmap for 5G. Arcep awarded in 2015 the 700 MHz band and is currently working towards licensing the 3.5 GHz band, both of which have been identified as 5G bands. Arcep is also working with the Direction générale des entreprises and the Agence nationale des fréquences towards enabling spectrum for 5G.

Arcep would like to thank all of the entities (listed on the last page) who agreed to take part in this process, and who were willing to contribute to the regulator’s investigation into the development of 4G’s successor. Arcep is dedicated to sustaining an ongoing dialogue with the market, and invites all interested parties to provide information on the matter or to share their reactions to this report, by contacting Arcep’s Mobile and Innovation Directorate at the following address: 5G@arcep.fr.

Table of Contents

INTRODUCTION: THE VISION FOR 5G, CREATING AN ULTRA-CONNECTED SOCIETY 4

1 5G OBJECTIVES 7

1.1 TECHNICAL SPECIFICATIONS OF 5G.. 8
1.2 5G USE CASES 10
1.3 JUMP IN PERFORMANCE COMPARED TO 4G NETWORKS 12
1.4 NETWORK SLICING AND SOFTWARE-DEFINED NETWORKS.. 14
1.5 TECHNOLOGICAL BUILDING BLOCKS TO ACHIEVE THE OBJECTIVES 15
1.5.1 AIR INTERFACE.. 15
1.5.2 NETWORK ARCHITECTURE.. 17
1.5.3 5G: A MULTI-TECHNOLOGY GENERATION COHABITATING WITH EXISTING NETWORKS................................ 19

2 5G DEVELOPMENT INITIATIVES 21

2.1 GOVERNMENT INITIATIVES 21
2.1.1 IN EUROPE.. 21
2.1.2 WORLDWIDE (EXAMPLES IN THE US, SOUTH KOREA, JAPAN AND CHINA) 23
2.2 A HOST OF PRIVATE INITIATIVES – A FEW EXAMPLES 24

3 THE CHALLENGES OF 5G 26

3.1 NEW BUSINESS MODELS FOCUSED ON VERTICAL MARKETS 26
3.1.1 THE AUTOMOTIVE SECTOR.. 26
3.1.2 INDUSTRY 4.0 27
3.2 SPECTRUM HARMONISATION 28
3.2.1 MILLIMETRE WAVE FREQUENCIES 28
3.2.2 FREQUENCY BANDS BELOW 6 GHZ... 29
3.3 INCREASINGLY SMALL CELLS 30
3.3.1 TAXATION.. 31
3.3.2 ACCESS TO ELEVATED AND "SEMI-ELEVATED" LOCATIONS 31
3.3.3 5G NETWORKS' REGIONAL COVERAGE AND BACKHAUL 32
3.4 NET NEUTRALITY ISSUES.. 32

ANNEX 1 DEFINITION AND STANDARDISATION WORK.. 35

ANNEX 2 ENTITIES ARCEP MET WITH.. 40

ANNEX 3 FIGURES.. 41

Introduction: the vision for 5G, creating an ultra-connected society

5G wants to be the disruptive generation, the generation that no longer only caters to the needs of mobile operators and consumer communications, but which opens up new prospects and enables an extremely wide diversity of applications and use, unified within a single technology. 5G is setting itself up as an enabler of the digitisation of society and the economy.

The idea that is starting to take shape behind the notion of 5G is that it will not mean just an increase in transmission speeds, as has been the case with previous generations.

Consumer mobile communications, video downloads and the use of mobile apps account for the bulk of 4G networks' radio resources usage today. With 5G, the goal will be to enable a much broader spectrum of uses and a much greater diversity of users.

5G is targeting a wide variety of sectors, which will not necessarily have anything other than this technology in common, but which are central pillars in a society: energy, healthcare, media, industry and transportation.

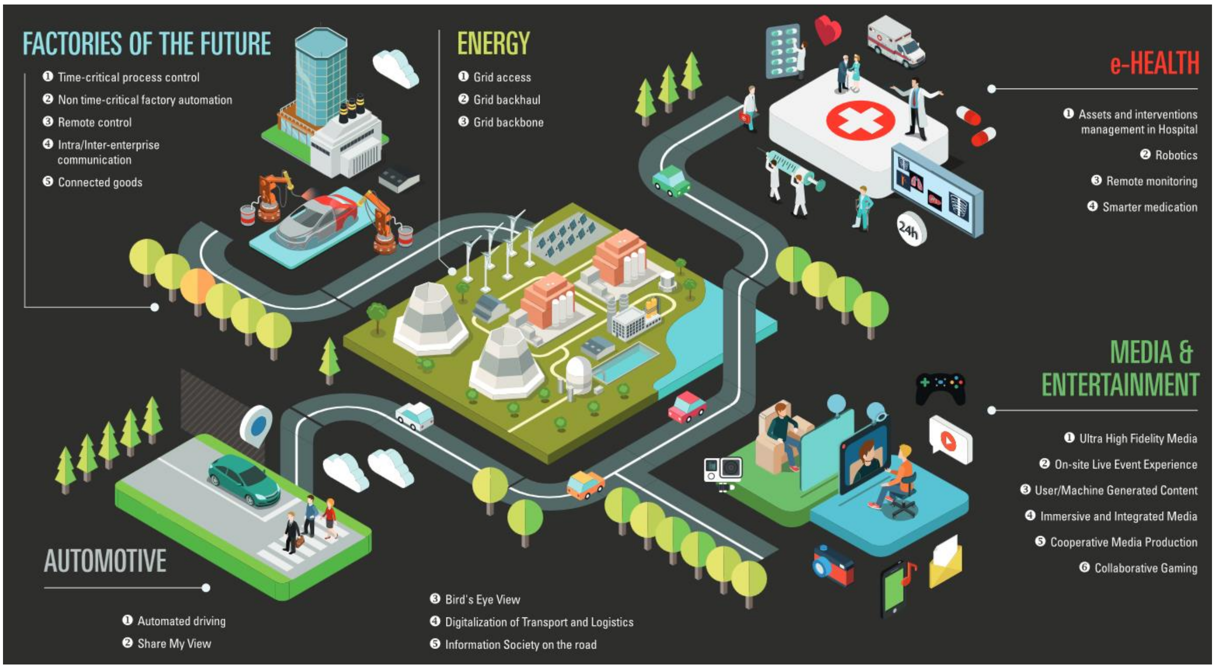
FACTORIES OF THE FUTURE
❶ Time-critical process control
❷ Non time-critical factory automation
❸ Remote control
❹ Intra/Inter-enterprise communication
❺ Connected goods
ENERGY
❶ Grid access
❷ Grid backhaul
❸ Grid backbone
e-HEALTH
❶ Assets and interventions management in Hospital
❷ Robotics
❸ Remote monitoring
❹ Smarter medication
24h
MEDIA & ENTERTAINMENT
❶ Ultra High Fidelity Media
❷ On-site Live Event Experience
❸ User/Machine Generated Content
❹ Immersive and Integrated Media
❺ Cooperative Media Production
❻ Collaborative Gaming
AUTOMOTIVE
❶ Automated driving
❷ Share My View
❸ Bird's Eye View
❹ Digitalization of Transport and Logistics
❺ Information Society on the road

Figure 1. 5G driving industrial and societal changes[1]

The energy sector, for instance, has undergone a great many changes and developments over the past several years in terms of energy production, storage and transport. Rising fossil fuel costs, the introduction of renewable energy sources and markets being opened up to competition have all helped usher in new kinds of energy product and new stakeholders – both independent companies and ordinary citizens – to the energy marketplace. The objective with 5G will thus be to enable better management of these networks (smart grids, smart agriculture, factories of the future) and their interconnections, to achieve more efficient and more agile distribution.

---

[1] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

Healthcare, transportation and media are sectors that are bound up with our fellow citizens' daily lives: any improvements in these sectors thus have a clearly visible impact. The introduction of robots to perform specialised surgical operations, of video on demand or the advent of new connected features in cars have already improved our daily lives. 5G is promising to go several steps further in all of these areas: it would enable remote medical diagnoses and operations in real time, it would democratise streaming of 360° 3D video, it would provide users with a vast selection of video content with a picture quality better than ultra high-definition (4K, 8K…). The automotive universe could rely on these new networks to help cars make decisions without human involvement, and also communicate with one another (this is already possible, for instance, with the first experimental fleet of fully autonomous taxis being tested on the streets of Singapore by the firms nuTonomy and Grab[2]), with reaction times that are compatible with the demands of high-speed travel.

In terms of factories of the future, the improvements brought by 5G are primarily targeting the introduction of new generations of connected robots, the interconnection of production sites and the much heavier use of smart sensors to improve industrial processes. Generally speaking, the aim is to achieve ubiquitous communication between machines, a process that is already well underway.

One thing that it is crucial to understand, through the few examples listed above whose demands might seem mutually incompatible, is that depending on the sector or the application, the network properties and the functionalities required will not be the same. Service providers – whether they are today's mobile operators or other market players – will need to be capable of adapting their network to demand, occasionally in real time. 5G will thus be not so much a universal technology as a polymorphous, or multi-faceted technology, capable of adapting to any use, up to and including the most demanding ones.

---

[2] http://nutonomy.com/

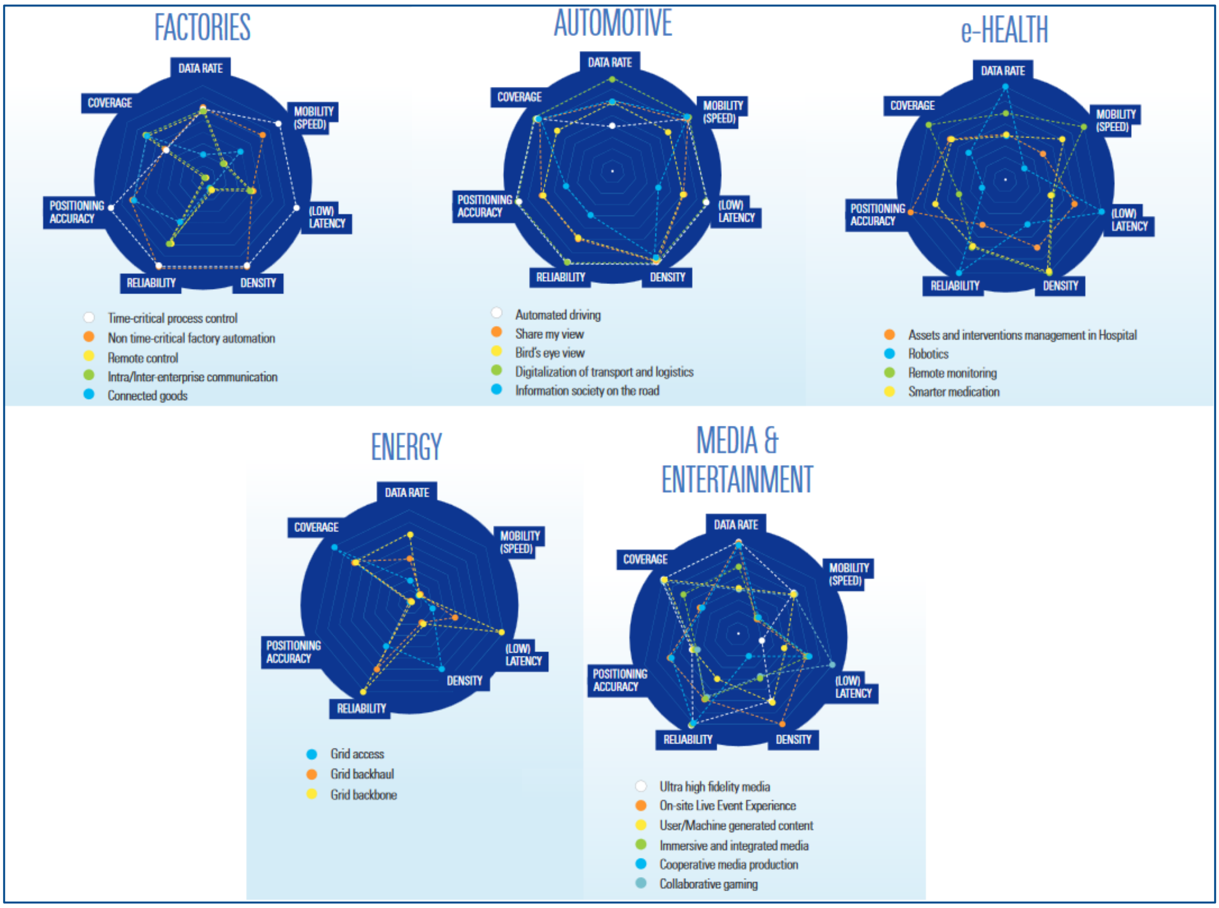

Figure 2. Performances required by vertical sectors[3]

---

[3] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

1 5G objectives

Ever since the first real mobile telephone call, 44 years ago[4], mobile technologies have been evolving steadily and their performances have improved exponentially: as mentioned earlier, calling services and later texting and finally the mobile internet and the use of multi-service mobile applications have characterised the evolution of mobile networks and the transition from one generation to the next. The birth of LTE technology and the fourth generation (4G), coupled with the widespread use of smartphones and tablets, have driven a massive increase in the amount of mobile data traffic being relayed over the networks.

The use of a mobile handset and its applications is now an integral part of our fellow citizens' daily habits. Portable connected devices are increasingly powerful: in many instances they have replaced users' landline telephones, cameras, computers and even televisions. Today, 5 million videos are watched on YouTube and 67,000 images uploaded to Instagram every minute (see Figure 3).

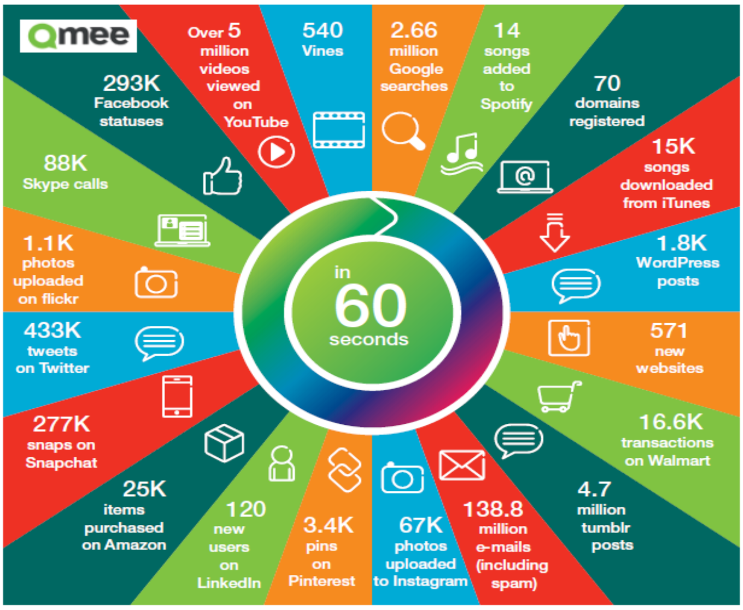

Figure 3. What we do over the network today in one minute[5]

The latest *Mobility Report*[6] from Ericsson indicates that traffic on mobile networks almost doubled in a single year, and that over the next five years it will have increased to 10 times what it is today. New solutions must therefore be found to meet this demand, and to optimise how resources are used.

The increase in the number of applications available, their diversification and the improved quality of mobile networks have all contributed to driving up demand, the emergence of new uses (connected objects, drones, etc.) and new users.

---

[4] On 3 April 1973, Motorola's Martin Cooper made a call from the corner of 56th street in New York City with the first mobile phone.

[5] https://www.ericsson.com/res/docs/2014/5g-what-is-it-for.pdf

[6] https://www.ericsson.com/assets/local/mobility-report/documents/2016/ericsson-mobility-report-november-2016.pdf

5G is at the crossroads of these news uses; it aims to better and simultaneously satisfy this tremendous variety of needs and these new demands, through a unified technology that takes this diversity into account at the design stage.

The advent of 5G could have a significant impact not only in the technical realm, but also on different countries' economic and social development. As indicated in the introduction, 5G targets a very large number of sectors and, through society's digitisation, is expected to contribute to countries' economic growth.

To give an example, according to a report produced by InterDigital Europe, Real Wireless, Tech4i2 and Connect (Trinity College Dublin)[7], this new technology will require a great deal of money and a great deal of work, but will generate €113.1 billion in profits per annum for the European economy by 2025.

1.1 Technical specifications of 5G

The specifications for a new generation of mobile telephony are set primarily by two bodies: ITU (International Telecommunication Union) and 3GPP (3rd Generation Partnership Project).

ITU is the United Nations agency devoted to information and communications technologies. It carries out research and studies through its Working Party 5D, the sub-group responsible for the overall radio system aspects of international mobile telecommunications (IMT). In 2013, this group began working on defining the characteristics of the new IMT standard, IMT-2020 (5G) (cf. 1.3), as it had done back in the early 2000s to define 4G (IMT-Advanced). Further details on the roadmap established for this work can be found in Annex 1, the objective set by ITU-R being to complete its analyses by 2020.

Parallel to the work being done by ITU are the studies being conducted by 3GPP. The 3rd Generation Partnership Project was created in 1998, and its members include seven standardisation bodies, several hundred industry players, associations and public organisations. It is responsible for developing and maintaining technical specifications for mobile telephony standards[8]. When a new standard is being defined by ITU, 3GPP works on the technical solutions that make it possible to achieve the objectives set by ITU.

Although 5G is one of the most widely debated topics inside international and European bodies today, no standard has yet been defined by 3GPP. The draft of Release 15, the first 5G standard from 3GPP, is still in the works: definition of the new architecture began in December 2016 and work on the New Radio (NR) interface is set to begin in March 2017. A first Release of the standard should be validated in September 2018 to meet the more urgent demands; a second Release (3GPP Release 16) will then be published in March 2020. All of these elements are addressed in more detail in Annex 1.

It should also be said that the transition from one generation to the next takes place gradually. LTE will continue to evolve alongside NR, and these two standards will likely be very complementary initially. In particular, for pioneer 5G rollouts LTE shall probably remain the master of the network and control the NR antennae. Moreover, some of the objectives set for 5G could be achieved thanks to functionalities or technologies introduced in 3GPP Releases 13, 14 and 15 that will not be proper to 5G, but rather evolutions of 4G (which some refer to as 4.9G or LTE Advanced Pro).

---

[7] http://ir.interdigital.com/file/Index?KeyFile=36051369

[8] GSM for 2G, UMTS for 3G and LTE for 4G

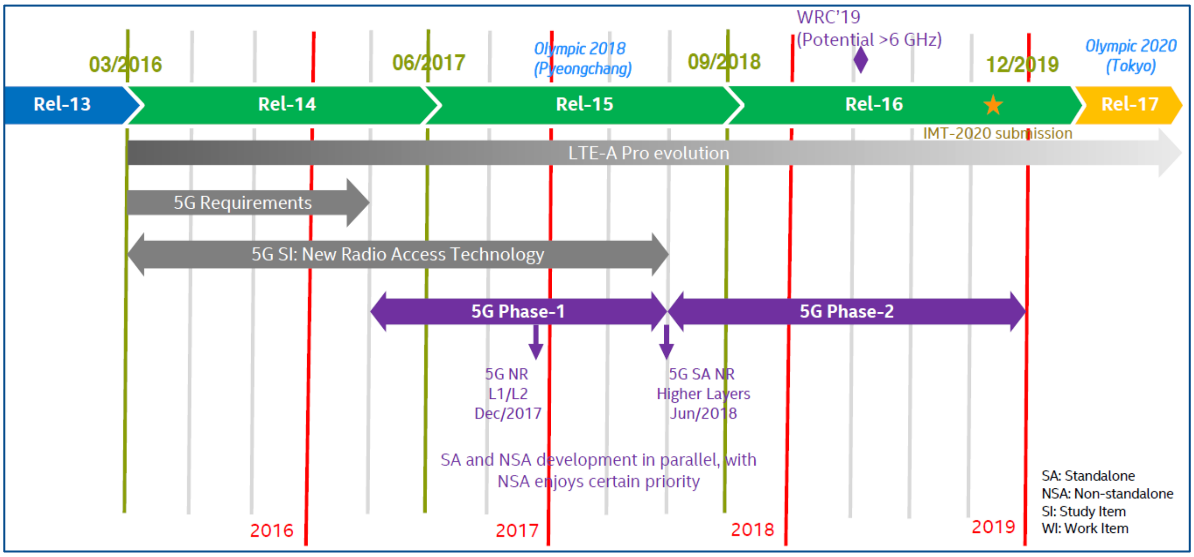
WRC'19
(Potential >6 GHz)
03/2016
06/2017
Olympic 2018
(Pyeongchang)
09/2018
12/2019
Olympic 2020
(Tokyo)
Rel-13
Rel-14
Rel-15
Rel-16
Rel-17
IMT-2020 submission
LTE-A Pro evolution
5G Requirements
5G SI: New Radio Access Technology
5G Phase-1
5G Phase-2
5G NR
L1/L2
Dec/2017
5G SA NR
Higher Layers
Jun/2018
SA and NSA development in parallel, with
NSA enjoys certain priority
2016
2017
2018
2019
SA: Standalone
NSA: Non-standalone
SI: Study Item
WI: Work Item

Figure 4. 3GPP timeline for 5G[9]

This same momentum occurred when making the transition from 3G (Release 4) to 4G (Release 10, first Release of IMT-Advanced): the first LTE specifications (Releases 8 & 9) do not achieve all of the ITU targets for 4G, and reprised many of the characteristics of the most advanced form of 3G at the time.

The synergies between the successive generations, the absence in most instances of a great initial jump in performances between generations, and the race between competitors to outperform one another and be the most innovative, can often lead operators and equipment suppliers to give a "commercial" name to each generation of mobile telephony. For instance, in the United States AT&T had called its HSPA network 4G, while in France it is a 3G+ network.

So in all likelihood the first 5G networks deployed on a large scale will be 4.9G systems, using carrier aggregation, massive MIMO (Multiple-Input Multiple-Output) and Network Function Virtualization, or NFV (cf. trials conducted in in France, detailed in Section 1.5.3). These technologies, which will be explored in greater detail further on, represent more of an evolution of the fourth generation than an actual transition to 5G, which will occur when disruptive technologies such as NR carriers in millimetre bands, non orthogonal multiple access (NOMA*)* and mobile edge computing (MEC) can be put into place.

In addition to this commercial race, equipment suppliers – often backed by their respective government at home, notably in the United States, Japan and South Korea – want to get a head start in defining and testing 5G technologies, in the hope of establishing themselves as the technological leaders in international standardisation bodies. Here, the precocity of the first rollouts announced in South Korea (5G deployment for the Winter Olympics in 2018 in Pyeongchang) and Japan (5G deployment for the Summer Olympics in Tokyo) would justify the supposition that only a small portion of 5G technologies will be used, and that these rollouts will be based more on 4.9G or pre-5G technologies.

---

[9] http://gsacom.com/paper/intel-5g-technology-mm-wave-frequencies/

1.2 5G use cases

Three main use cases (defined by ITU, under IMT-2020), with their respective – and potentially mutually incompatible – demands are in the process of taking shape, and will make it possible to meet the sector-specific needs referred to in the introduction.

1. **mMTC – Massive Machine Type Communications**: A very large number of connected devices with disparate quality of service requirements. The objective of this category is to provide a response to the exponential increase in the density of connected objects;
2. **eMBB – Enhanced Mobile Broadband**: ultra high-speed connection indoors and outdoors, with uniform quality of service, even on the edges of a cell;
3. **uRLLC – Ultra-reliable and Low Latency Communications**: this use case has stringent requirements for capabilities such as latency and packet-loss, to ensure increased reactivity.

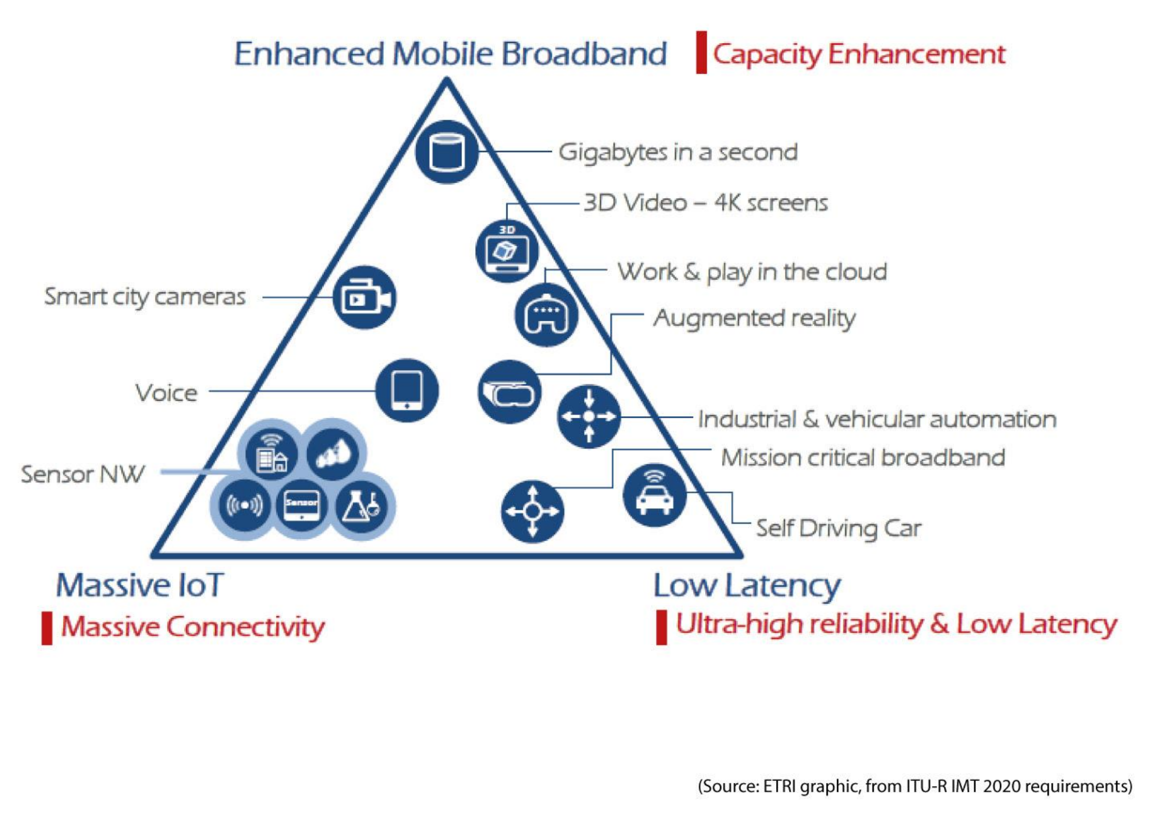
Enhanced Mobile Broadband
Capacity Enhancement
Gigabytes in a second
3D Video – 4K screens
Work & play in the cloud
Augmented reality
Smart city cameras
Voice
Industrial & vehicular automation
Mission critical broadband
Sensor NW
Self Driving Car
Massive IoT
Massive Connectivity
Low Latency
Ultra-high reliability & Low Latency

(Source: ETRI graphic, from ITU-R IMT 2020 requirements)

Figure 5. 5G use cases[10]

The first group (mMTC) primarily encompasses all Internet of Things related uses. These services require broad coverage, lower energy consumption and relatively slow transmission speeds. What 5G will deliver compared to existing technologies is the ability to connect objects that are spread out in a very dense fashion across a given area.

Enhanced mobile broadband (eMBB) concerns all of the applications and services that require increasingly fast connections, for instance to watch ultra high-definition (8K) videos or to stream virtual or augmented reality applications wirelessly.

---

[10] http://5g.ieee.org/standards

Ultra-reliable and Low Latency Communications (uRLLC) include all of the applications that require an extreme reactivity and very strong message transmission guarantees. The sectors where these requirements are particularly prevalent are transportation (reaction time when an accident occurs, for instance) and medicine (telesurgery), and when digitising manufacturing processes in general.

To implement these three use case, ITU established eight key performance indicators[11] (KPI) to specify, quantify and measure the characteristics of IMT-2020 (5G) systems:

- Peak data rate (Gbit/s);
- User experienced data rate (Mbit/s);
- Spectrum efficiency (bit/Hz);
- Device mobility (km/h);
- Latency (ms);
- Connection density (number of connected/accessible objects per km²);
- Network's energy efficiency;
- Area traffic capacity (Mbit/s/m²).

---

[11] Added to these eight classes, five new indicators were defined and are currently being examined: reliability, mobility interruption time, bandwidth, maximum spectrum efficiency, 5th percentile spectrum efficiency.

1.3 Jump in performance compared to 4G networks

As mentioned in Paragraph 1.1, the deployment of 5G is likely to take place in two stages:

1. The first 5G rollouts will deliver better performances, but as a continuation of what is being provided at the time by 4G systems which will continue to evolve as well (4.5G, 4.9G);
2. Performances will continue to improve with the gradual introduction of disruptive technologies, such as the use of millimetre wave frequencies.

This gradual rollout is very similar to the way in which 4G was introduced, as the performances obtained with pioneer deployments were relatively similar to those supplied by existing 3G networks.

ITU defines the 4th generation of technologies, called IMT-Advanced, by the values for the eight KPI listed above, as indicated in the following chart. The 5th generation, called IMT-2020, is represented as well. Here, it should be noted that the official ITU definition of 4G in fact corresponds to what commercial rollouts supplied later on, under the name 4G+ or LTE-Advanced.

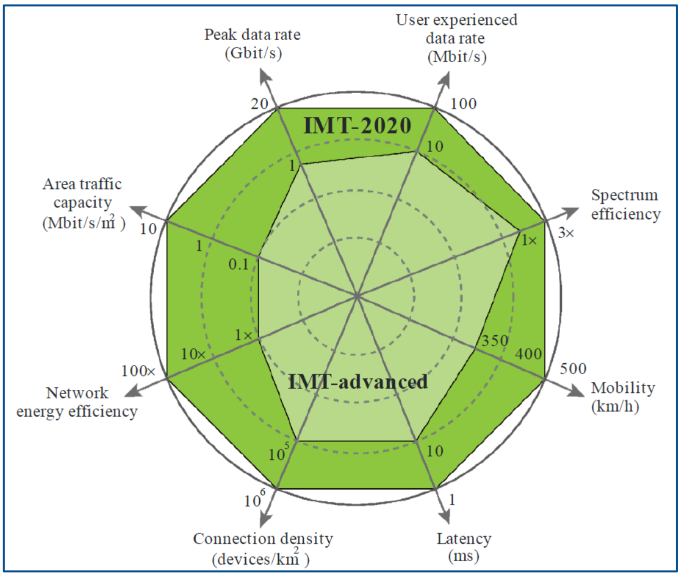

Figure 6. Comparison between 4G and 5G with respect to the eight key performance indicators[12]

According to these objectives, 5G must be able to provide a user experienced data rate and a peak data rate that is respectively 10 and 20 times higher than what is currently available. The maximum connection density will be multiplied by 10 and latency divided by at least 10 (the target point-to-point latency is 1 ms, compared to 30 to 40 ms today).

---

[12] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

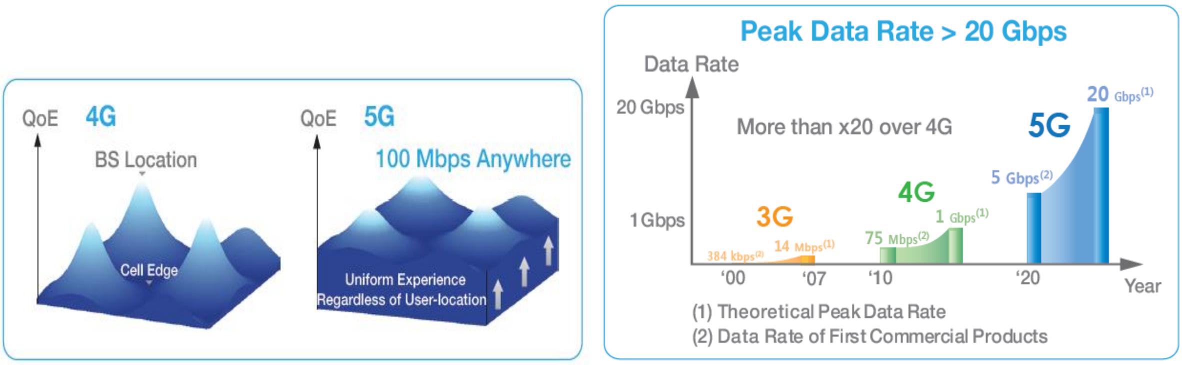

Figure 7. Representation of the "average user experience data rate" KPI for 4G and 5G and "Peak data rate" KPI for 3G, 4G and 5G[13]

As the successor to 4G, the objective for 5G will also be to provide:

- An extremely reliable network, with more consistent performances regardless of the user's position with respect to the base station;
- A stable connection, even when travelling (at speeds of up to 500 km/h);
- Greater network energy efficiency (with batteries that consume up to 100 times less power).

The following table summarises the target performances for 5G and those currently available with 4G:

Performances/Generation	4G	5G
1. Peak data rate (Gbit/s)	1	20
2. User experience data rate (Mbit/s)	10	100
3. Spectrum efficiency	1x	3x
4. Speed (km/h)	350	500
5. Latency (ms)	10	1
6. Connection density (number of objects/km²)	10^5	10^6
7. Network energy efficiency	1x	100x
8. Area traffic capacity (Mbit/s/m²)	0.1	10

Table 1. Comparisons between 4G and 5G performances

---

[13] http://www.samsung.com/global/business-images/insights/2015/Samsung-5G-Vision-0.pdf

1.4 Network slicing and software-defined networks

It is vital to understand that the set of indicators listed in Section 1.3 determine a set of peak performances for 5G. It will not, however, be possible to achieve all these peak values simultaneously: not every requirement or use case is compatible, so a trade-off will need to be made when defining categories of use that each have their own performance envelope, notably for the use cases described in Section 1.2 (mMTC, eMBB and uRLLC). This is the principle of network slicing: each slice has its own set of KPI, which is a trade-off tied to the target use. On a 5G system, the network properties will need to adapt to the chosen environment.

The following diagram positions the three main use cases listed in Section 1.2. with respect to the eight key performance indicators listed above.

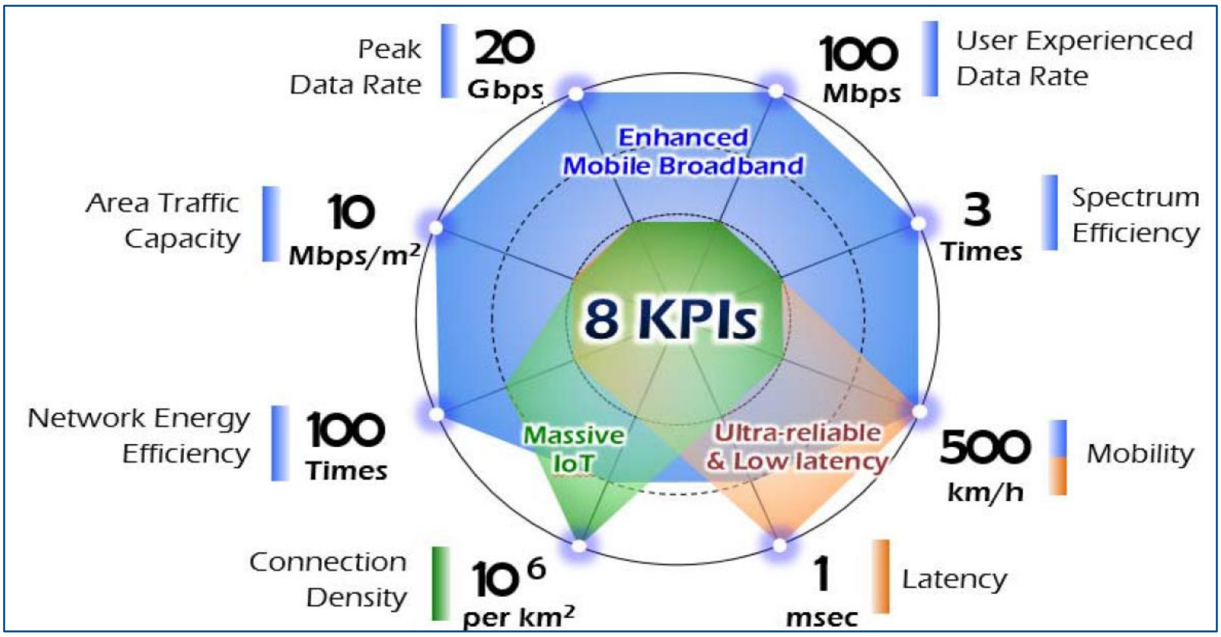
Peak Data Rate
20 Gbps
100 Mbps
User Experienced Data Rate
Enhanced Mobile Broadband
Area Traffic Capacity
10 Mbps/m²
3 Times
Spectrum Efficiency
8 KPIs
Network Energy Efficiency
100 Times
Massive IoT
Ultra-reliable & Low latency
500 km/h
Mobility
Connection Density
10⁶ per km²
1 msec
Latency

Figure 8. Key performance indicators for the three 5G use cases[14]

So for applications that require enhanced mobile broadband (eMBB), such as 4K, 8K or 3D video or virtual reality, a certain number of performance indicators, such as spectrum efficiency, peak data rate and area traffic capacity, can be reached only at the expense of others, such as latency or connection density.

On the flipside, when a massive simultaneous connection of connected objects (mMTC) needs to be managed, the network will concentrate its resources and use the technologies required to achieve this task, but will not be able, for instance, to use spectrum as efficiently or to guarantee low latency.

Lastly, when ultra-reliable and low latency communications (uRLLC) are required, the number of simultaneous connections, data rates and spectrum efficiency may be reduced.

This flexibility, or ability to adapt, that network slicing brings can only be achieved thanks to the softwarisation and virtualisation of a sizeable number of network components (cf. 1.5.2) – a process referred to as Software-Defined Networking (SDN) and Network Function Virtualisation (NFV). Behind these acronyms is a common idea, namely to use as many generic and reconfigurable components as possible, rather than bespoke ones that are permanently dedicated to very specific tasks. This evolution towards software-based systems has been in the works for several years, but is now becoming possible thanks to improved performances from all of these reconfigurable components, including those that are the closest to the elementary tasks of wireless communications (detection, baseband coding, bitstream management, frequency handover, signal processing, etc.).

---

[14] https://5g-ppp.eu/wp-content/uploads/2016/11/06_10-Nov_Session-3_Lee-JunHwan.pdf

1.5 Technological building blocks to achieve the objectives

1.5.1 Air interface

Several, sometimes competing radio access technologies are currently being examined. Some have already been pre-implemented by equipment manufactures and can be used in trials, notably massive MIMO and NFV. Others, such as NOMA modulation and mobile edge computing (MEC), will no doubt take longer before they are ready to be used. In any event, a consensus – which could be painful for certain suppliers whose investments will be lost – will need to be found when defining 5G standards, to ensure the systems' interoperability.

The technologies currently being examined are the following:

- **Millimetre wave frequencies**: the use of millimetre wave frequencies constitutes one of the disruptive 5G technologies. The term refers to the frequencies above 6 GHz which have never been considered for mobile fronthaul network rollouts, for reasons of technological maturity and propagation quality. To meet the demand for ever increasing data rates and traffic volumes, new bands with very wide channels (over 100 MHz per user) will need to be employed: millimetre wave frequencies could provide this spectrum resource, and in certain cases their use would make it possible to achieve the data rates listed in Table 1. Comparisons between 4G and 5G). In exchange, to be able to use these frequencies all of the required, miniaturised low-cost technologies will need to be developed, and ensure a level of energy consumption that is compatible with portable devices (amplifiers, coders, signal processing, antennae, etc.). In particular, because of millimetre waves' poor propagation quality, each cell will have limited coverage and so require the use of beamforming (described below) to better focus the power transmitted by the antenna.
- **Massive MIMO** (Multiple Inputs – Multiple Outputs): this technology involves the use of a large number of smart micro-antennae, located on the same panel (between eight and 128 today, but the number will increase with the use of frequencies above 6 GHz). The appeal of using massive MIMO is twofold: first, the technology makes it possible to increase data rates, thanks to spatiotemporal multiplexing; second, it makes it possible to focus energy on a device to improve its link budget, thanks to beamforming.

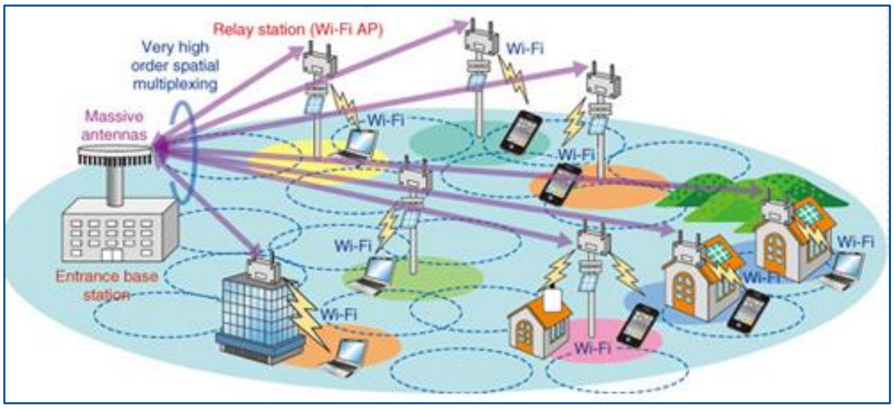
Relay station (Wi-Fi AP)
Very high order spatial multiplexing
Massive antennas
Entrance base station
Wi-Fi

Figure 9. Example of the use of a beamforming antenna to connect Wi-Fi access points[15]

- **Full Duplex**: in classic systems, transmission and reception takes place either on different frequency bands, i.e. frequency division duplexing (FDD) used on all mobile network bands in France, or at different times: time division duplexing (TDD), the top contender for LTE wireless local loop networks in France. The full duplex is intended to enable the simultaneous transmission and reception of data, on the same frequencies, at the same time and in the same location.[16]

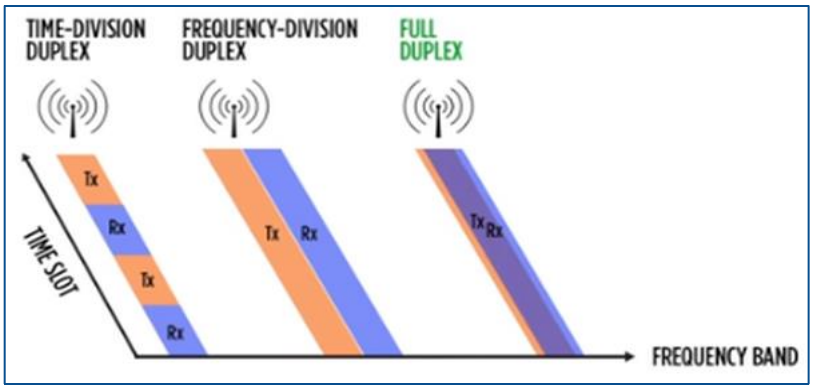
TIME-DIVISION DUPLEX
FREQUENCY-DIVISION DUPLEX
FULL DUPLEX
TIME SLOT
Tx
Rx
Tx
Rx
Tx
Rx
TX RX
FREQUENCY BAND

Figure 10. Illustration of full-duplex, compared to FDD and TDD[17]

- **NOMA Multiplexing** (Non Orthogonal Multiple Access): LTE uses what is referred to as orthogonal multiplexing, with each device using a portion of the resource blocks in a unique fashion at any given time. For 5G to provide improved spectrum efficiency compared to 4G, the plan is to use non-orthogonal multiplexing methods, whereby several users can use the same frequencies at the same time. A distinction can be made between several users by assigning different codes to each user – referred to as SCMA or sparse code multiple access – a combination of 3G's code division multiple access (CDMA) and 4G's orthogonal frequency division multiple access (OFDMA) or by playing on the difference in users' signal to noise ratios (power domain NOMA[18], illustrated below). These are the two methods chosen for NR.

---

[15] https://www.slideshare.net/100001290086432/massive-mimo

[16] The basic operating principle is the following: an antenna sends a signal at the same time as it is receiving signals coming from devices in the cell. However, the signal received by the antenna is a combination of the signal sent by itself and signals coming from the devices (everyone is "talking" at once). As it knows which signal it itself has sent, the antenna can subtract it during digital processing from the ones it received. So only the signals received from the devices remain.

[17] http://compeng.columbia.edu/biggest-component-full-duplex-wi-fi-radio-antenna

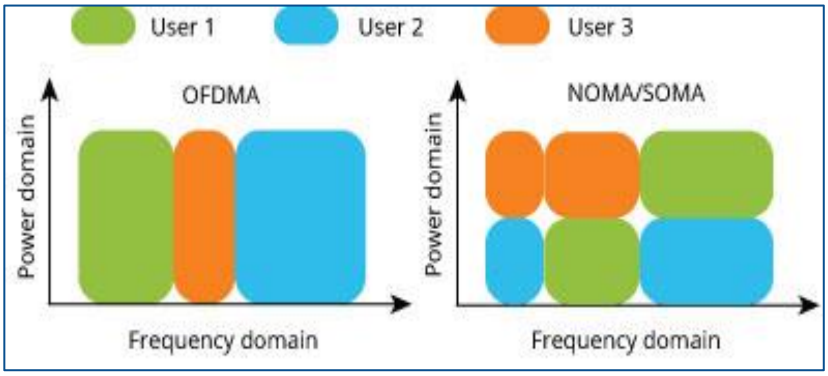

Figure 11. Illustration de multiplexing *power domain* NOMA[19]

- **QAM256**: as with many modern communication systems, 4G uses quadrature amplitude modulation (QAM). In 4G this modulation can achieve QAM64, which means that six bits of information are being transmitted ($2^6 = 64$) at any given time. On wireless systems, the main limitation on QAM order is the signal to noise ratio: when a large amount of information is sent all at once, its transmission will be very sensitive to disruptions (a bit like trying to talk in a noisy environment: it is easy to understand "yes" or "no" but harder to understand more complex sentences). Thanks to an improved link budget, via antenna or signal processing technologies, 5G modulation could reach QAM256, i.e. eight bits of information being transmitted at any given time, which translates into a 33% increase in maximum capacity under ideal conditions. This improved modulation will also be deployed on advanced 4G systems.
- **IoT waveforms**: new waveforms are being explored for the future deployment of 5G IoT in mobile bands. But although mass market IoT is one of the main challenges put forth for 5G, no concrete results have yet been made public. Operators[20] are starting to deploy new standards (EC-GSM or Extended Coverage GSM, LTE-(e)MTC or enhancements for Machine-Type Communications, NB-IoT or NarrowBand IoT) which were defined by 3GPP in Release 13 but, as they are based on 2G and 4G, they do not deliver the performance levels, notably in terms of autonomy, coverage and density, that are compatible with the targets set for future 5G networks.

1.5.2 Network architecture

As with air interfaces, new network architectures are also being explored:

- **Software-defined networking (SDN) and network functions virtualisation (NFV)**: these two functionalities extend beyond the scope of 5G networks per se. They are part of an overall process of network upgrades taking place today and already available with 4G technology (4.9G). The nevertheless remain a key enabler of 5G:

---

[18] The basic operating principle is the following: either User 1 (U1) with a good signal to noise ratio and User 2 (U2) with a less good signal to noise ratio. The antenna sends a high power Signal 2 to U2, and superimposes a weaker Signal 1 aimed at U1. U2 will only see S2 as S1 is drowned out by the noise. U1 will decrypt S2 then delete it from the signal received, to create a higher quality S1, thanks to a better signal to noise ratio.

[19] https://www.anritsu.com/en-AU/test-measurement/technologies/5g-everything-connected/5g-everything-connected-detail

[20] http://www.vodafone.com/content/index/what/technology-blog/nbiot-commercial-launch-spain.html

  - SDN (Software-Defined Network) is designed to disassociate the network's control plane from its data plane, these two planes traditionally being linked and distributed in a set fashion in the network (see diagram below). Controlling the network, a task previously assigned to specialised and unscalable hardware components, is centralised in the form of software on more powerful servers and, in theory, free of equipment manufacturer specifications. This enables the deployment of high value-added services (load balancing, smart routing, dynamic configuration, etc.) in disparate environments.

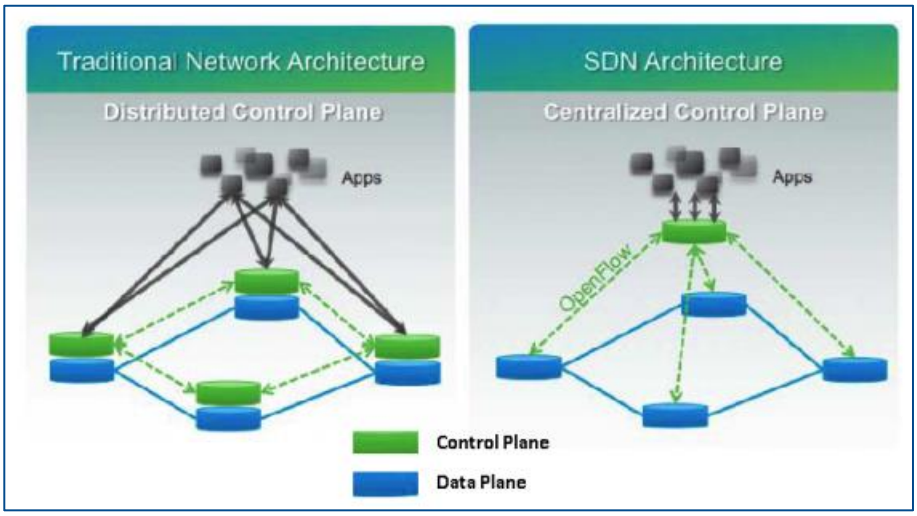
Traditional Network Architecture
Distributed Control Plane
Apps
SDN Architecture
Centralized Control Plane
Apps
OpenFlow
🟩 Control Plane
🟦 Data Plane

Figure 12. Centralisation of the control plane in a software-defined network[21]

  - NFV, which builds upon SDN, is used to virtualise, in other words to replace hardware designed specially to perform certain key network functions (firewall, network core, interfaces between different systems…) with software on a server, to accelerate rollouts and enable rapid changes and upgrades.

- **CloudRAN**: this functionality, also know as centralized-RAN, requires a very different network architecture to what we find today. It is an evolution of SDN: the base stations' signal processing units, currently installed at the base station level, are moved to the cloud and centralised. They communicate with the network radio heads, located closer to the antenna, over an optical fibre network (Radio over fibre technology). This centralisation makes it possible to obtain a complete overview of all of the stations deployed and to coordinate signal processing and manage interference between cells and devices

---

[21] http://www-igm.univ-mlv.fr/~dr/XPOSE2014/software-defined_networking/sdn.html

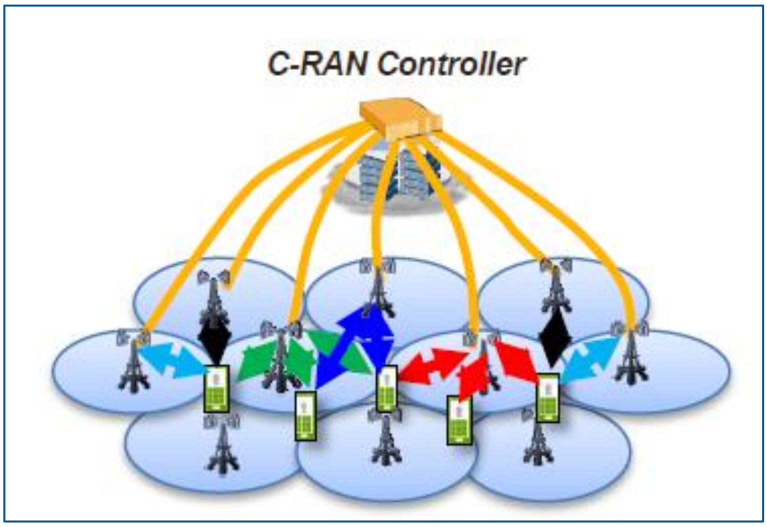

Figure 13. Illustration of CloudRAN[22] network architecture

- **Optimised content delivery, using a mobile content delivery network (mobile CDN)**: corresponds to a set of servers working together in a transparent fashion to optimise the delivery of content to end users over wireless (mobile or Wi-Fi) networks, with high availability and performance. With 5G, the objective for these CDN is to cache content close to users, notably thanks to predictive algorithms, to offload traffic from the networks and decrease latency.
- **MEC (mobile edge computing)**: MEC is an evolution of mobile CDN whose purpose, in addition to bringing data closer to devices, is to provide devices with an accessible computing power with very low latency, within a very specific area for demanding applications. This technology makes it possible to locate a portion of the network's intelligence (managing local critical applications and performance analysis) at the base station level. The "antennae" will be capable of analysing a certain number of data, and so to make decisions very quickly.
- **Device-to-device**: D2D is a direct form of communication between two nearby devices, which does not require the data to travel over the cellular network. Device-to-device communication is not new, as technologies such as Bluetooth and Wi-Fi direct already enable it. But a new mesh networking technology will be introduced with 4.9G and later 5G network rollouts: LTE-direct. Far more energy-efficient than its predecessors, this technology will have a range of up to 500 metres and geolocation capabilities to enable communications. This technology will be very useful for low latency V2V (vehicle-to-vehicle) or V2X (vehicle-to-everything) communications and for certain public-security related uses.

1.5.3 5G: a multi-technology generation cohabitating with existing networks

As stated earlier, 5G is not meant to replace 4G overnight. In practice, the devices will undoubtedly be multi-modal: still connected to the 4G network, which will provide extended coverage for pioneer rollouts, then transitioning to 5G networks when they become available.

While 4G and 5G frequencies will probably be initially segregated, in all likelihood devices will rapidly become capable of aggregating 4G and 5G carriers and, further down the road, 4G carriers will be encapsulated in 5G ones.

---

[22] https://5g-ppp.eu/wp-content/uploads/2016/11/04_10-Nov_Session-3_Takaharu-Nakamura.pdf

4G is still being deployed, and its technological evolution and certain building blocks will be used by both the first 5G networks and advanced 4G networks. To wit, the latest trials conducted in France are allowing 4G networks to perform better thanks to the use of pre-5G technologies:

- Bouygues Telecom, in partnership with Huawei, managed to achieve a peak data rate of 1 Gbps thanks to the simultaneous use of four-carrier aggregation (800 MHz, 1800 MHz, 2100 MHz, 2600 MHz) and more powerful modulation (256 QAM)[23];
- In the coming weeks, Orange will be launching a massive MIMO (16x16) trial with Nokia[24].

In addition to the strong integration between 4G and 5G, the new generation will no doubt also continue convergence efforts between frequency bands governed by exclusive licences – i.e. bands that are allocated exclusively to an operator, such as mobile operators – and unlicensed frequency bands, governed by a system of general authorisation (e.g. Wi-Fi bands), which already began in 4G with LTE-LAA (Long Term Evolution – License Assisted Access) and LTE-LWA (Long Term Evolution – Wi-Fi Link Aggregation).

LTE-LAA is characterised by the aggregation of one or several LTE carriers, used in licensed bands, with other LTE carriers employing unlicensed 5 GHz Wi-Fi bands. To guarantee cohabitation with very widespread Wi-Fi networks whose deployment pattern is unpredictable, it uses LBT (Listen Before Talk) technology to listen to the radio channel before transmitting, in order to determine whether or not a frequency is available.

LTE-LWA consists of an aggregation of LTE carriers in licensed bands with Wi-Fi traffic. To achieve this, the LTE cell and Wi-Fi access point need to be connected: the Wi-Fi traffic is sent to the cell which sends back the entire aggregated link on the 4G core network. This technology is particularly well suited to indoor environments with small cells, a system which, as detailed in paragraph 3.3 below, will lend itself well to 5G deployments.

---

[23] https://www.corporate.bouyguestelecom.fr/wp-content/uploads/2016/05/Communiqu%C3%A9-Bouygues-Telecom-et-Huawei-Technologies-r%C3%A9alisent-un-test-de-d%C3%A9bit-4G-%C3%A0-plus-de-1-Gbps-pour-la-premi%C3%A8re-fois-en-Europe-de-lOuest.pdf

[24] https://networks.orange.fr/actualites/actualites-des-networks/orange-et-nokia-partners-pour-developper-les-futurs-services-5g

2 5G development initiatives

With a view to keeping the many promises being made for 5G, a number of initiatives are currently underway around the globe to promote its development. Below, we detail the foremost initiatives put forth by the players with which Arcep met as part of its investigation.

2.1 Government initiatives

The prospect of significant socio-economic repercussions generated by 5G (possible revenue of close to $225 billion a year[25] by 2025) combined with many countries' desire to establish themselves as technological leaders and make their companies more competitive, have propelled a multitude of government initiatives around the globe, aimed at encouraging the mobile ecosystem to begin the work and make the investments required to drive the rapid construction of the first 5G networks.

A selection of the largest initiative is detailed below.

2.1.1 In Europe

5G-PPP

The 5G Public Private Partnership (5G-PPP) is dedicated to 5G research and development, created on the initiative of the European Commission in 2013, with a budget de €700 million in public funding. The main objectives set by 5G-PPP are:

- Create stronger ties between the economic players and academic bodies devoted to the telecommunications sector over R&D projects, along the entire value chain;
- Reduce technological dependence on the United States and Asia while sustaining a strong global market;
- Regain technological leadership, notably in disruptive technologies, by promoting standards in international bodies;
- Allow innovative business models to emerge;
- Facilitate large-scale experimentation.

The results of this work will help clarify the 5G action plan (see below), and fuel the standardisation work that is currently underway.

To achieve its ambitions, 5G-PPP has initiated three stages of work, financed by the European Union, whose roadmap is in sync with the main international initiatives (3GPP and ITU, cf. Figure 14):

- The first stage which is currently underway, will last until mid-2017;
- A second stage focused on systems optimisations, from the end of 2017 to mid-2019;
- And finally a full-size trial stage from 2019 to 2020.

Deploying 5G by 2020 will require Europe to develop leading edge technologies, globally approved standards and especially to achieve consensus over the most suitable frequency bands. This funding and these projects – involving a great many researchers from more than a hundred companies and the finest R&D centres in Europe – are thus vital.

---

[25] https://www.abiresearch.com/press/abi-research-projects-5g-worldwide-service-revenue

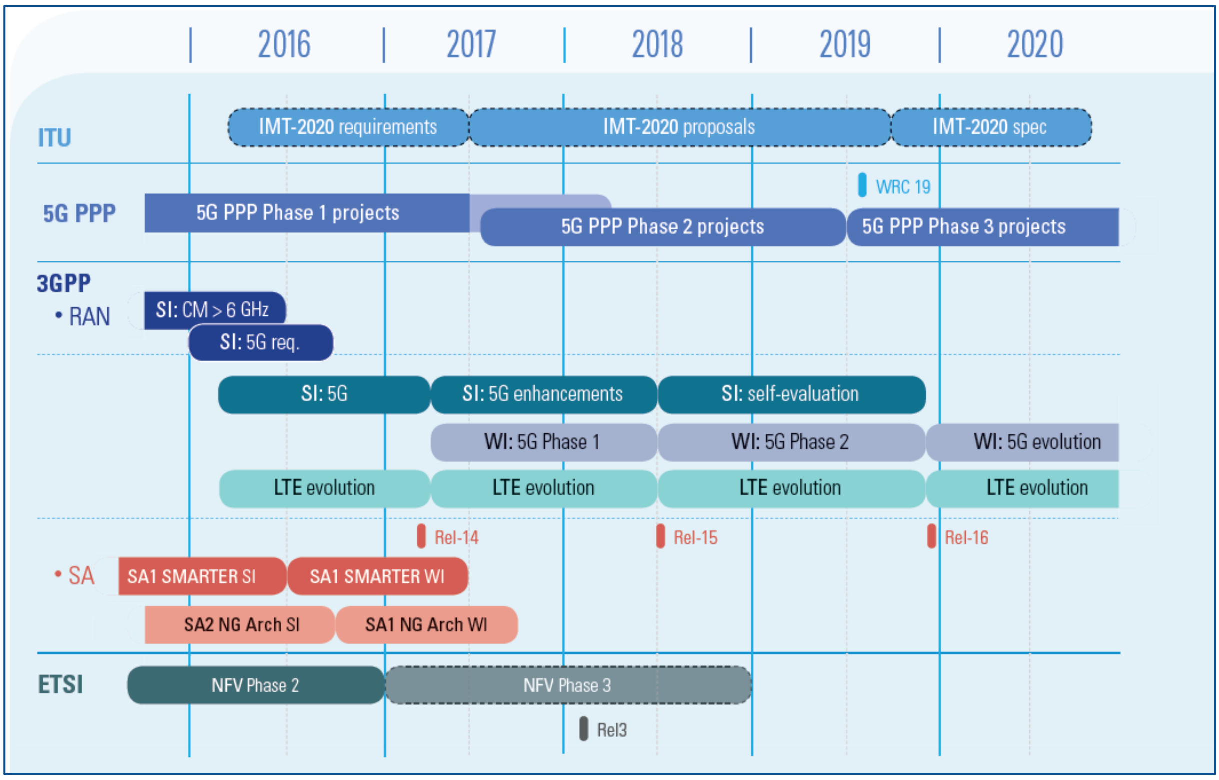
2016
2017
2018
2019
2020
ITU
IMT-2020 requirements
IMT-2020 proposals
IMT-2020 spec
WRC 19
5G PPP
5G PPP Phase 1 projects
5G PPP Phase 2 projects
5G PPP Phase 3 projects
3GPP
• RAN
SI: CM > 6 GHz
SI: 5G req.
SI: 5G
SI: 5G enhancements
SI: self-evaluation
WI: 5G Phase 1
WI: 5G Phase 2
WI: 5G evolution
LTE evolution
LTE evolution
LTE evolution
LTE evolution
Rel-14
Rel-15
Rel-16
• SA
SA1 SMARTER SI
SA1 SMARTER WI
SA2 NG Arch SI
SA1 NG Arch WI
ETSI
NFV Phase 2
NFV Phase 3
Rel3

Figure 14. 5G PPP vs. 3GPP and ITU roadmaps[26]

5G action plan

As a complementary measure, on 14 September 2016 the European Commission launched its *5G for Europe Action Plan* to bolster investments in 5G infrastructure and service rollout efforts in the Digital Single Market between now and 2020. This action plan sets out a clear roadmap for public and private 5G investments inside the EU.

The Commission has proposed the following measures to achieve this plan:

- Align roadmaps and priorities for a coordinated 5G deployment across all EU Member states, targeting early network introduction by 2018, and moving towards commercial large scale introduction by the end of 2020 at the latest.
- Make provisional spectrum bands available for 5G ahead of the 2019 World Radio Communication Conference (WRC-19), to be complemented by additional bands as quickly as possible, and work towards a recommended approach for the authorisation of the specific 5G spectrum bands above 6GHz.
- Promote early deployment in major urban areas and along major transport paths.
- Promote pan-European multi-stakeholder trials as catalysts to turn technological innovation into full business solutions.
- Facilitate the implementation of an industry-led venture fund in support of 5G-based innovation.
- Unite leading actors in working towards the promotion of global standards.

---

[26] 5G Empowering vertical industries. White Paper, 2016, https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

The European Commission has given every EU country a certain number of ambitious, numerical targets. One core objective for 5G is thus to have at least one major city in every European country outfitted with this new generation mobile system by 2020, and coverage of every city, motorway and high-speed railway lines by 2025. This comes in response to announcements from South Korea and Japan which are both promising large-scale 5G demonstrations, respectively, at the Winter Olympics in Pyeongchang in 2018 and the Summer Olympics in Tokyo in 2020.

2.1.2 Worldwide (examples in the US, South Korea, Japan and China)

The race is already underway between countries to be the first to begin large-scale 5G trials, and later to introduce commercially available services, so much so that they are willing to employ pre-standard technical specifications.

Several countries are therefore working on standards and seeking to reach a consensus with one another to achieve international backing for the technical specifications that will satisfy their needs, and earn them the best possible return on the investments made thus far. This in turn is creating a certain turmoil, for instance, over the frequency bands that would be the best candidates for 5G.

The United States[27]

5G is seen as an unprecedented opportunity for economic growth, with a tremendous impact on education, job, transportation, etc. According to the US federal regulator, the FCC (Federal Communications Commission), the following three elements in particular need to be the prime focus of attention: spectrum, infrastructure and the backhaul network:

- **Spectrum**: in July 2016, the FCC voted to free up and to open up nearly 11 GHz of high frequency spectrum to be used for fixed and mobile broadband applications: 3.85 GHz of licensed spectrum in the 27.5 – 28.35 GHz and 37 - 40 GHz bands, as well 7 GHz of unlicensed spectrum from the 64 - 71 GHz band. The FCC's stated goal was to provide the assurance and clarity for investments in the telecoms sector. Verizon welcomed this decision, and plans on achieving the first rollouts in 2017, with trials already underway in several cities around the country (cf. 3.3).
- **Infrastructures**: The FCC believes that 5G must be underpinned by a robust infrastructure network capable of handling already heavy traffic that could potentially increase exponentially in the coming years. It will interact in a hybrid fashion between traditional towers with macro cells and small cell deployments, as well as distributed antenna systems. With an eye on the deployment of new towers, the FCC resolved to reduce, or eliminate in certain cases, the regulatory restrictions on installing facilities and antennae that will have very little impact "on historic properties" (tower re-use, small antennae, indoor installations, etc.)[28].
- **Backhaul network**: Because 5G systems will require very high-power backhaul networks, the FCC is in the process of reforming and updating the regulations governing data services markets, including wireless backhaul solutions. The FCC is seeking to protect this market while also creating competition, so that competitive and high quality backhaul network solutions can emerge.

---

[27] Forging Our 5G Future – Federal Communications Commission: https://www.fcc.gov/5G

[28] https://apps.fcc.gov/edocs_public/attachmatch/DA-16-900A1_Rcd.pdf

Japan

Japan wants to demonstrate its 5G leadership by deploying the first commercial 5G network that complies with international technical specifications, in time for the Summer Olympic Games in Tokyo in 2020.

According to the *Radio Policies Towards 2020*[29] report, published in June 2016 by Japan's Ministry of Internal affairs and Communications (MIC), the 3600 - 4200 MHz, 4400 - 4900 MHz and 27.5 – 29.5 GHz bands were selected to be the 5G candidate bands on a national scale. Although other bands are also being investigated, with a view to WRC19, 5G rollouts are already planned in the 3600 - 4100 MHz, 4405 - 4895 MHz and 27.5 - 28.28 GHz bands as early as 2017 in Tokyo, and will be built out in 2018 and 2019.

China

Like Japan, China too wants to prove its leadership with the first commercial rollouts in 2020. The first trials will be conducted that year in the 3400 – 3600 MHz band. The 3300 - 3400 MHz, 4400 - 4500 MHz and 4800 - 4990 MHz bands are also under consideration and currently being investigated. For higher speeds, the country plans on using spectrum around the 25 GHz and 40 GHz frequencies.

South Korea

South Korea, meanwhile, has set its sights on a pre-commercial 5G service that would be ready in time for the 2018 Winter Olympics in Pyeongchang. Several trials are already underway to prepare for full scale demonstrations in several South Korean cities, including Pyeongchang and Seoul. The country's three national mobile operators are hoping for spectrum in the 26.5 – 29.5 GHz bands.

South Korea's largest mobile operator, SK Telecom, has announced a plan for conducting interoperability tests with Qualcomm and Ericsson, as well as outdoor trials of the new air interface based on the 3GPP NR standards that are being developed today. These trials and experiments will be carried out in the second half of 2017.

If the goal of these trials is to help accelerate the specification of the new NR air interface, which is part of the work being done on 3GPP Release 15, it has not yet been specified which frequency bands will be used for these trials.

2.2 A host of private initiatives – a few examples

5G Open Trial Specification Alliance

In early 2016, South Korean operators KT and SK Telecom, Japanese carrier NTT DoCoMo and American carrier Verizon formed the 5G Open Trial Specification Alliance to carry out collaborative 5G trials.

Due to be conducted between 2016 and 2018, the aim for these trials is to provide a common platform for operators, for exchanging results and shared assessments of the different 5G network components and elements. One of the operators' objectives is to help accelerate the definition of standards and specs, and to enable economies of scale. The findings of these evaluations will provide input for 3GPP discussions, adding simulations to the experimental data produced by the partnership.

---

[29] http://www.gsma.com/spectrum/wp-content/uploads/2016/08/MIC_Spectrum-for-5G-MIC-Kuniko-OGAWA.pdf

The alliance's founders hope to attract a number of industrial partners to the platform, including other carriers, equipment suppliers, chipset makers, etc. Several of the founding operators' equipment suppliers are already involved in the scheduled trials.

The trials will cover multiple frequency bands, both above and below 6 GHz.

Verizon 5G Technical Forum

The Verizon 5G Technology Forum (V5GTF) was created in late 2015 by Verizon, in cooperation with its partners Cisco, Ericsson, Intel, LG, Nokia and Qualcomm. The goal of the collaboration is to provide a platform for testing radio interface specifications in the 28 and 39 GHz bands.

The first fruit of this collaboration, in July 2016 Verizon announced the completion of the first specifications of their 5G radio interface. These specs should enable the different parties involved to develop interoperable solutions and so to help drive forward the definition pre-standards.

Even though they run the risk of developing solutions that are not compatible with 3GPP or ITU standards, Verizon believes it has a solid enough grasp of the overall concepts being discussed in those bodies. The tests carried out in several US cities were thus able to validate substantially greater capacities than 4G.

Orange/Ericsson

Orange and Ericsson have been working together since October 2016 on developing 5G use cases and services, as well as demonstrations. The purpose is to develop technological building blocks, conduct trials and pilot projects for a range of use cases, including multi gigabits/s wireless internet access in suburban and rural environments, the Internet of Things and connected cars.

Their collaboration made it possible to achieve 15 Gb/s in a laboratory environment, notably thanks to the use of massive MIMO and beamtracking using centimetre waves. The collaboration will also focus on the transition from 4G to 5G solutions, notably in terms of reducing costs and improving energy efficiency, and on the use of SDN and NFV technologies.

Since January 2017, the partnership has included PSA (Peugeot) to carry out trials on connected cars.

5G-ConnectedMobility

5G-ConnectedMobility is a consortium formed by Ericsson, BMW, Deutsche Bahn, Germany's three mobile operators, Deutsche Telekom, Telefonica Deutschland and Vodafone, the 5G TU Dresden laboratory, the BAST research institute and BnetzA, with the goal of galvanising and accelerating 5G R&D in Germany.

5G-ConnectedMobility thus plans on supplying a digital motorway infrastructure and real application environment to be able to test V2V (vehicle-to-vehicle) technologies as well as solutions for digitising railway infrastructure.

To this end, 5G-ConnectedMobility operates with the help of an independent network infrastructure that is not connected to any commercial network. The Ericsson mobile network devoted to this project makes it possible to run live tests: Ericsson thus obtained permission from regulator BNetzA to employ spectrum in the 700 MHz band in the Nuremberg area.

3 The challenges of 5G

In this part we explore the different challenges that lie ahead for 5G, which emerged from the interviews that Arcep conducted over the past several months.

3.1 New business models focused on vertical markets

3G and especially 4G technologies were designed primarily for the superfast mobile internet. 5G continues on in this direction, but also wants to target what are known as vertical markets, which encompass several segments, including:

- Connected vehicles, not only to deliver entertainment and information to passengers, but also to guarantee safety via communications both between vehicles and between vehicles and infrastructure;
- Factories of the future;
- Smart cities with requirements in the areas of public transportation (similar to the needs of connected vehicles), the environment, managing buildings and energy consumption;
- Medicine, healthcare and robot-assisted telesurgery;
- Smart grid flow monitoring and management (electricity, gas, water, etc.).

This section will focus in particular on the connected car and factories of the future segments. Because of their current and future macro-economic context, along with the plurality and effervescence of the pioneer work being done in these areas, these two segments constitute the main avenues of 5G development in vertical markets. Smart cities and smart grids have already begun to soar through existing Internet of Things (IoT) technologies.

3.1.1 The automotive sector

The car is an extremely common form of transportation, and safe driving is a fundamental consideration: human error is the number one cause of all transport accidents. The transportation sector wants to use technological innovations to tackle this problem, and to continue to make transportation more efficient, more sustainable and safer.

There are three areas in which technological progress could help improve automotive transportation. 5G could have a role to play in all three, but particularly in the first two:

- Provide in car internet connectivity, to deliver entertainment to passengers;
- Provide access to driver assistance information, to reduce accidents and improve the fluidity of traffic;
- And, finally, the ability to make cars autonomous, thanks to artificial intelligence algorithms.

The first area is just an extension of the developments occurring today around 4G. The aim is to give passengers access to their messaging services, to the internet, to multimedia content, online gaming, etc. The increased connection speeds promised by 5G will improve the use of all of these services.

The second area aims to make cars more intelligent by using information that was not previously available to them. This in turn will help improve the safety and efficiency of the networks as well as help drivers make the right decisions, and adapt to driving conditions. Such connected vehicles could have access to information about dangers on the road (slow or stopped cars, traffic jam warnings, indications of where construction is taking place on the roads, weather conditions, emergency braking, emergency service vehicles approaching, etc.) or regarding signage (signalling/signage on-board vehicles, speed limits on vehicles, failure to heed signage/safety precautions at crossings, request for right-of-way at traffic lights for designated vehicles, green light optimal speed advisory, etc.). Other services, such as information on refuelling or recharging stations, vulnerable road user protection, managing street parking and traffic information, and smart guidance, could also prove useful. It is still not clear whether, to achieve this, vehicles will simply exchange information with each other or if connectivity with infrastructure will be required along the roads to optimise vehicles' behaviour. In both instances, 5G could have a role to play.

Of the many initiatives that are already underway we can begin with an example from France: the SCOOP@F[30] (cooperative intelligent transport systems) project coordinated by the Ministry of the Environment, Energy and the Sea, and which unites local authorities and R&D centres. Launched in February 2014, new partners subsequently joined the initiative, including Orange and Austrian, Spanish and Portuguese partners. As it is a European project, it receives 50% of its financing from the European Commission, and cross-tests are conducted with Austria, Spain and Portugal. SCOOP@F is a pilot rollout project for cooperative intelligent transport systems; it aims to deploy 3,000 vehicles on 2,000 km of roadway spread across five locations: Ile-de-France, the A4 motorway, the Isère, and ring roads in Bordeaux and Brittany. Its main objectives are to improve road safety and the safety of roadway workers, achieve more efficient traffic management, reduce pollution, streamline infrastructure management costs and participate in defining the car of the future.

In addition, in early 2017, mobile carrier Orange, equipment supplier Ericsson and car-maker PSA Group signed a partnership agreement[31], as part of the "*Towards 5G*" initiative, to conduct technical trials relating to 5G. The aim of this alliance is to test the different paths of technological evolution from 4G to 5G to meet the needs of connected cars, notably in terms of intelligent transport systems (ITS), for safer driving and new on-board services.

The third area concerns the emergence of autonomous vehicles. A number of projects are underway in this area. The first step is to outfit vehicles with algorithms that enable it to make decisions quickly based on their environment. This requires a large number of sensors to deliver a full "understanding" of what is happening around the vehicle. Without pre-judging the technologies that will ultimately be employed to achieve this, as with a human driver (see above), the car could take advantage of connections with other vehicles on the road and with an infrastructure, to have access to all of the aforementioned information.

3.1.2 Industry 4.0

Competitiveness does not depend solely on innovation and the evolution of products, but also on modernising businesses and their means of production. According to certain studies carried out in 2015[32], the digital transition in Europe will enable enterprises to increase their revenue by more than 110 billion euros a year over a five-year period.

A great many countries have set out a strategy for modernising their manufacturing infrastructure (*l'industrie du futur* in France[33], Industry 4.0 in Germany…) of which one aspect will involve digitising processes and trade. The European Commission itself introduced measures in 2016 for strengthening competitiveness in Europe, in which 5G could play a significant role. This included earmarking 500 million euros for the Horizon-2020[34] research programme.

---

[30] http://www.scoop.developpement-durable.gouv.fr/spip.php?page=sommaire

[31] http://www.orange.com/fr/Press-Room/communiques-2017/Ericsson-Orange-et-le-groupe-PSA-partners-pour-la-voiture-connectee-en-5G

[32] PwC, "Industry 4.0: Opportunities and challenges of the industrial internet" (2015), and Boston Consulting Group, "Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries" (2015)

[33] http://www.economie.gouv.fr/files/files/PDF/industrie-du-futur_dp.pdf

The advent of new technologies (4G, fibre and soon 5G) and new services (the Internet of Things, the cloud, big data) should facilitate businesses' digital transition. 5G in particular is expected to be a very versatile technology, capable of undergirding a very wide array of uses, and could go a long way in furthering companies' transition to digital technologies and solutions.

3.2 Spectrum harmonisation

5G is emerging as a technology that will use both low frequencies (f < 1 GHz), high frequencies (1 GHz < f < 6 GHz) and, for the first time ever in consumer networks, very high frequencies referred to as "millimetre wave" frequencies (f > 6 GHz).

This spectrum diversity is entirely bound up with the promises of 5G: extended coverage (low frequencies), ultra high speeds (very large channels in very high frequency bands), low power consumption. Furthermore, satellite services will also contribute to the development of this new technology, especially in areas that are difficult to cover and to provide backhaul solutions. The satellite industry is thus taking an interest in 5G, and wants to be involved in defining these new generation network.

3.2.1 Millimetre wave frequencies

The "millimetre" band, also referred to as millimetre wave spectrum, aka frequencies above 6 GHz, are essential to enabling 5G to mark a departure from 4G, for the reasons cited in Section 1.5.1.

At the latest World Radiocommunications Conference (WRC-15 in Geneva), a conference under the aegis of ITU whose objective is to change the way frequencies are allocated between users, discussions over the definition of future mobile bands made it possible to focus future 5G studies, for millimetre wave frequencies, on a certain number of bands situated between 24 GHz and 86 GHz (33.25 GHz identified in total): 24.25 - 27.5 GHz, 31.8 - 33.4 GHz, 37 – 43.5 GHz, 45.5 - 50.2 GHz, 50.4 - 52.6 GHz, 66 - 76 GHz, 81 - 86 GHz.

It is important to stress that, even if the above-listed bands have been identified as "5G bands", at this stage there is no way to know whether they can actually be used to deploy this new generation system: only the results of technical studies will make it possible to establish the constraints and rules of compliance, and to validate the feasibility of these hypotheses.

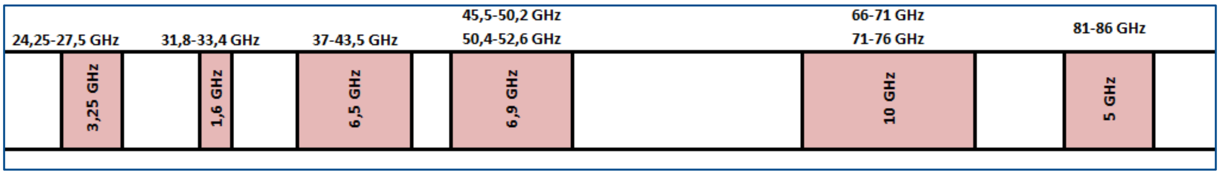

Figure 15. Millimetre wave frequencies identified at WRC-15

Contrary to the conclusions of the Conference, which reflect European recommendations, the United States and certain Asian countries (South Korea, Japan) have decided to perform 5G trials in the 28 GHz band, and equipment suppliers such as Qualcomm and Samsung, have begun manufacturing 28 GHz- band products.

---

[34] https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

Meanwhile Europe, following the publication of an RSPG (Radio Spectrum Policy Group within which France is represented by ANFR) opinion[35], decided to focus its first studies on the 26 GHz band (pioneer band), then on the 32 GHz and 42 GHz bands. Later, studies will be carried out on introducing 5G in all of the other bands identified by WRC-15.

The rapid choice of the 26 GHz band as the pioneer band was made to enable economies of scale for equipment production, since it is very likely that dual-mode equipment, i.e. compatible with both the 26 GHz and 28 GHz band, will be available for pioneer rollouts[36].

In France, the 26 GHz band is already being employed for a variety of uses: mobile operators' wireless fixed links (4G infrastructure links), fixed satellite service systems and ground stations for space services. Studies will therefore need to be conducted to take these services into account, and define either the cohabitation or migration of applications to other millimetre wave frequencies.

3.2.2 Frequency bands below 6 GHz

a) The 3.4-3.8 GHz band

It will not be possible for 5G to run entirely on millimetre wave frequencies: the propagation qualities of these bands make it difficult to achieve widespread coverage, particularly in more sparsely populated areas. In addition, these bands are still lacking in technological maturity when it comes to delivering consumer market communication services. As a result, a "core" band below 6 GHz needs to be identified: one that provides sufficiently large channels to enable future 5G operators to provide innovative services and a higher quality of service than with 4G.

The 3400 - 3800 MHz band appears to be a good candidate.

First, it has already been harmonised for ultrafast mobile services inside the European Union. Initially the 3400 - 3600 MHz band and later the 3600 - 3800 MHz band were identified as the "IMT" bands (for high-speed mobile). Second, they have a great deal of available spectrum (up to 400 MHz). Lastly, technological advancements (antenna and signal processing) make these frequencies compatible with their use to establish macro cells, and not only microcells. To give an example, it emerged from the interviews conducted by Arcep that coverage with this band could be similar to coverage with the 2.6 GHz band, the core 4G frequency band.

This analysis has been confirmed by the European Commission (in its 5G mandate to RSCOM[37]) and by RSPG[38] which consider the 3.4 - 3.8 GHz band as the only credible 5G band for deployments taking place before the end of 2020.

In France, the band is assigned to Arcep, in a priority fashion, for the 3400 - 3600 MHz frequencies (and in a non-priority fashion to the Ministry of the Interior and the Ministry of Defence), and exclusively for the 3600 - 3800 MHz frequency bands. It is used for wireless local loop and satellite applications.

The compatibility of future uses with current applications is being examined by the Electromagnetic Compatibility Advisory Committee, which is chaired by the National Frequency Agency and of which

---

[35] http://rspg-spectrum.eu/wp-content/uploads/2013/05/RPSG16-032-Opinion_5G.pdf

[36] The tuning range is vital here: a system capable of regulating its frequency from 26 to 28 GHz appears entirely technologically achievable within the timeframe for the first 5G millimetre wave deployments. This would have been less true if a switch between 28 and 32 GHz were required.

[37] https://circabc.europa.eu/sd/a/448dc765-51de-4fc8-b6e0-56ed6a1d0bca/RSCOM16-40rev3%205G%20draft_mandate_CEPT.pdf

[38] http://rspg-spectrum.eu/wp-content/uploads/2013/05/RPSG16-032-Opinion_5G.pdf

Arcep is a member: exclusion zones, protecting certain locations tied to long-term use of satellite solutions, need to be set up. Military radiolocation systems operating below 3.4 GHz also need to be protected.

b) Other bands below 6 GHz

The bands being used today for 2G, 3G and 4G could be used for future 5G deployments.

Refarming 2G, 3G and 4G could be a delicate undertaking, however, because of the duplexing methods they use. Most mobile communications in Europe use FDD (Frequency Division Duplexing)[39] to exchange information. Technical discussions over 5G are nevertheless anticipating that TDD (Time Division Duplexing)[40] will be the chief, if not sole, form of duplexing used for this new generation, notably because it makes it possible to adjust bandwidth to data rates and because it is particularly efficient when beamforming is used.

For 5G networks, using already harmonised mobile bands will thus require in-depth technical studies to define the terms and conditions of use and sharing with existing services. On this point, the ECC (Electronic Communications Committee[41]*)* decided to assess the potential of certain already harmonised bands, notably the 700 MHz and the L band (1427-1492 MHz).

In France, Arcep allocated the 700 MHz band to mobile operators in late 2015. Even if the country’s four operators all obtained frequencies in this band, only Free Mobile, which has no blocks of 800 MHz spectrum, decided to begin its rollouts in the band using LTE (4G) technology.

Arcep is currently analysing the contributions to its public consultation, addressing among other the 3.5 GHz band, that ended recently.

The L band (1427 - 1518 MHz), which was defined to be used exclusively in SDL (Supplemental DownLink) mode, could be considered to meet the constant demand for ever higher data rates and the greater increase in downlink traffic, compared to uplink traffic. In France the band is used by wireless fixed links authorised by Arcep, by the Ministry of Defence for mobile services (excluding aeronautical) and by the Ministry of the Interior. Using it for 4G or 5G would therefore suppose moving the above-mentioned uses over to other bands.

3.3 Increasingly small cells

Today, mobile network rollouts are essentially based on the use of macro base stations: installations that are equipped with high-power antennae deployed to guarantee coverage for a relatively wide area, providing good quality of service. Network configurations are evolving constantly: new radio base stations are installed on a regular basis to increase the networks’ capacity, to better meet users’ needs in terms of indoor and outdoor coverage, and to improve quality of service.

However, this continually growing demand for capacity already requires operators to increase the density of their networks with smaller and smaller cells.

5G – which will probably bring about a sizeable increase in data traffic, and which will use millimetre wave frequencies whose propagation capabilities are weak – will no doubt require the widespread deployment of low-power base stations (small cells).

---

[39] A technique that consists of sending and receiving data simultaneously, but on two different frequency bands.

[40] A technique that consists of sending and receiving data on the same frequency band but at different times.

[41] The European body that specifies the technical terms and conditions governing the use of frequency bands, which are then set by European Commission decisions.

To satisfy demand and enable the introduction of 5G, estimates indicate at least 10 small cells per macro base station in urban settings[42], where cells are already today relatively tightly meshed to handle traffic density.

Here, the players that met with Arcep raised several questions that will no doubt need to be answered when considering the ubiquitous deployment of small cells.

3.3.1 Taxation

The current regulatory framework stipulates that base stations whose power requires an opinion, an agreement or a statement from the National Frequency Agency (ANFR) are subject to a flat tax on network companies (IFER[43]). The amount of the tax varies according to transmission power, the type of installation and the deployment location. For a deployment in an urban area, the tax stands at €1,607/year/installation for a base station with an effective isotropic radiated power (EIRP) of more than 5W (a COMSIS agreement from ANFR is required to be able to transmit) and €160.70 €/ year/installation for an EIRP of between 1W and 5W (declaration to ANFR is required to be able to transmit).

5G small cells will likely use variable transmission powers of between 1W and 25W. In light of rollout density forecasts, hence the number of small cells to install, some of the stakeholders that Arcep met with raised the question of possibly adapting this tax in such a way as to enable massive small cell deployments without generating an equally massive increase in the total tax amount.

This process is already underway, notably with a view to decreasing taxes on base stations in locations that are hard to cover: Act No. 2016-1888 of 28 December 2016 on the modernisation, development and protection of mountain regions thus exonerates mobile base stations built in mountain regions between 1 January 2017 and 31 December 2020 from paying the IFER tax.

3.3.2 Access to elevated and “semi-elevated” locations

To perform their deployments successfully, mobile operators have traditionally needed to install their base stations in elevated locations (towers, rooftops, etc.). This will continue to be true with 5G networks, but will be even more challenging for two main reasons:

1. 5G antennae will probably be larger (in the m^2 range for some) than current 2G, 3G or 4G antennae, because of the above-mentioned massive MIMO processing that will require the use of a very large number of radiating elements. Moreover, additional antennae compatible with new 5G bands will no doubt also need to be deployed. So the re-use of existing masts could very well be problematic, and new (possibly collocated) transmission sites will need to be found.
2. This search for new sites will also need to be carried out to install small cells in semi-elevated locations, but with extra care as the density of these installations will undoubtedly be high: operators will thus be required to deploy their equipment on urban furniture and infrastructures such as bus shelters, lampposts, public buildings, billboards, etc.

As a result, public authorities will need to keep a close eye on the matter and, if necessary, adopt measures that will facilitate 5G rollouts.

---

[42] http://www.lemag-numerique.com/wp-content/uploads/2015/10/WP_-Souverainete_Telecoms_PetitesCells_FINAL.pdf

[43] IFER = *Imposition forfaitaire sur les enterprises de réseaux*

3.3.3 5G networks' regional coverage and backhaul

Ensuring regional connectivity will be one of the challenges for this new generation of mobile networks.

The diversity of use-cases envisioned for future 5G networks, the geographical distribution of which is not yet precisely known, must be factored in when addressing coverage issues.

The higher frequency bands that are being put forth for future 5G networks, along with the potentially very substantial bandwidth consumed by these new uses, pose the question of the regional foothold of these networks. As a matter of fact, mobile networks have never before employed such high frequencies whose use will require a large number of relay antennae to be installed.

Moreover, connecting the 5G installations to the network will push to the fore the question of the cost of connecting them via optical fibre, which will no doubt be necessary in most instances to ensure the expected quality of service. The industry needs to design the technologies that will make it possible to minimise the cost of 5G rollouts in rural areas.

3.4 Net neutrality issues

European regulation on safeguarding an Open Internet[44], adopted by the European Parliament and Council on 25 November 2015, for which European regulators required an additional nine months to specify the rules governing its application, introduces the principle of net neutrality as one of the top priorities in the standardisation hierarchy.

Net neutrality is an overriding principle that guarantees equal treatment for all data traffic on the internet. In particular, it excludes any form of discrimination with respect to the source, the destination or the content of data flows.

On 30 August 2016, BEREC (the Body of European Regulators for Electronic Communications) published guidelines for national regulators on the enforcement of the European Open Internet regulation[45].

During the BEREC public consultation[46] on its draft net neutrality guidelines, several enterprises and electronic communications sector stakeholders took the opportunity to deliver a clear-cut point of view in their *"5G manifesto for timely deployment of 5G in Europe"*[47]. This manifesto aims at warning public authorities against a too restrictive approach to traffic management, and especially of the supposedly negative effects that, in their opinion, an overly strong enforcement of net neutrality could have on the 5G rollout roadmap.

---

[44] Regulation (EU) 2015/2120: http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R2120&from=EN

[45] Internet service providers (ISP) can employ reasonable traffic management measures for certain categories of traffic, under non-discriminatory conditions and provided they are transparent, proportionate and justified by objective technical requirements, and not used to serve their own commercial interests. ISPs can also distinguish certain services, called specialised services, from their internet access service, without them affecting the latter, provided these services have specific, objective transit requirements.

[46] http://www.berec.europa.eu/

[47] http://telecoms.com/wp-content/blogs.dir/1/files/2016/07/5GManifestofortimelydeploymentof5GinEurope.pdf

BEREC's opinion[48], in response to the contributions to the aforementioned public consultation, recalls that the rules of net neutrality are technology neutral and therefore apply to 5G networks. Moreover, BEREC guidelines clarify that network slicing which networks will need to use to respond to the different use cases (mMTC, eMBB, uRLLC) planned for this new generation (cf. 1.4) system, may be used to deliver specialised services[49].

The subject of net neutrality as it applies to future 5G networks is still very much open and unexplored, but new analyses could be performed parallel to the work being done on defining 5G. Arcep is open to discussion and will also work, within BEREC, towards bringing the required clarity to the ecosystem.

---

[48] http://berec.europa.eu/eng/document_register/subject_matter/berec/download/0/6161-berec-report-on-the-outcome-of-the-publi_0.pdf

[49] http://berec.europa.eu/eng/document_register/subject_matter/berec/download/0/6160-berec-guidelines-on-the-implementation-b_0.pdf - footnote 26

Annexes

Annex 1 Definition and standardisation work

A new generation of mobile telephony is defined primarily by two bodies: ITU (International Telecommunication Union) and 3GPP (3rd Generation Partnership Project). These two, respectively public and private, organisations are dedicated to defining the objectives, standards and technical specifications of the new technology in question.

ITU

The first large-scale 5G commercial rollouts are expected to take place in 2020. As mentioned earlier, the exploratory phase – which provides an opportunity to determine demands and identify the most promising techniques and technologies for these future 5G networks – has already begun. Although a number of organisations and consortia are involved in defining 5G, 3GPP will very probably be the central standardisation body for its technical specifications.

Whatever the case may be, ITU (International Telecommunication Union) is vital to defining the technologies and standards that govern any new generation of IMT (International Mobile Telecommunications) at the global level.

These IMT standards are established with the involvement of public authorities and industry players, and have provided the framework for the evolution of mobile communication services around the world, since the beginning of IMT standardisation, with IMT-2000 (3G, UMTS), then IMT-Advanced (4G, LTE-A) and more recently IMT-2020 (5G).

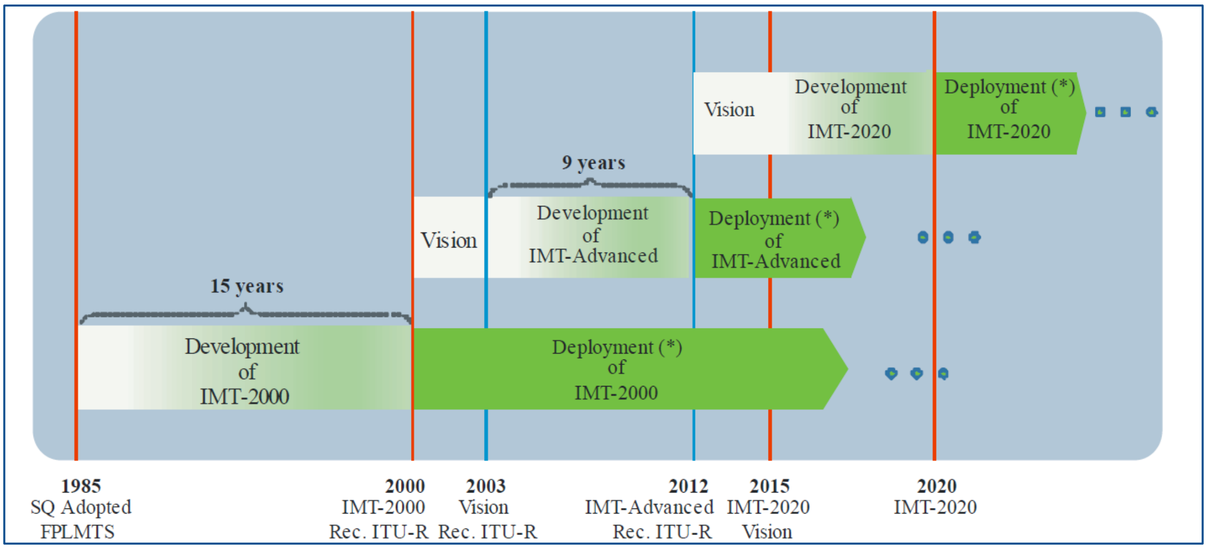
Vision
Development of IMT-2020
Deployment (*) of IMT-2020
9 years
Vision
Development of IMT-Advanced
Deployment (*) of IMT-Advanced
15 years
Development of IMT-2000
Deployment (*) of IMT-2000
1985 SQ Adopted FPLMTS
2000 IMT-2000 Rec. ITU-R
2003 Vision Rec. ITU-R
2012 IMT-Advanced Rec. ITU-R
2015 IMT-2020 Vision
2020 IMT-2020

Figure 16. Timeline of IMT developments and deployments[52]

The advent of an IMT standard typically occurs in three main stages : a vision stage, a development stage and a deployment stage.

The vision stage – whose completion is marked by a document that is usually called a Vision Recommendation – is the stage during which ITU sets the framework and objectives for the future technology. The aim of this document is generally to define what the new technology will be in a more or less concrete fashion, what its characteristics will be, the uses it will enable, etc. Whether for

---

[52] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

IMT-Advanced, the commonly used technical abbreviation attributed to the definition of 4G, or IMT-2020, the abbreviation attributed to definition of 5G, the vision stage lasts around three years.

Recommendation ITU-R M.2083-0 was published in September 2015, bringing to a close three years of work performed by ITU-R (Working Party 5D) on defining the framework and objectives for IMT for 2020 and beyond. It is this document that today serves as the basis of the different 5G research and standardisation work being done around the globe.

Next comes the standards development stage, based on the conclusions of the vision stage. Regarding IMT-2020, this development and standardisation work is already underway, and ITU plans on having completed it by 2020 to be able to satisfy the most pressing needs of the ITU members and organisations that want to deploy 5G as quickly as possible. This is all the more challenging as the development stage will last only five years, compared to 15 years for IMT-2000 and nine years for IMT-Advanced.

The different deadlines set for IMT-2020 within ITU can be found in Figure 17. The spectrum identification phases (indicated by the black triangles) coincide with the World Radiocommunications Conferences, of which the latest was WRC-2015 and the next is WRC-2019 (see 3.2).

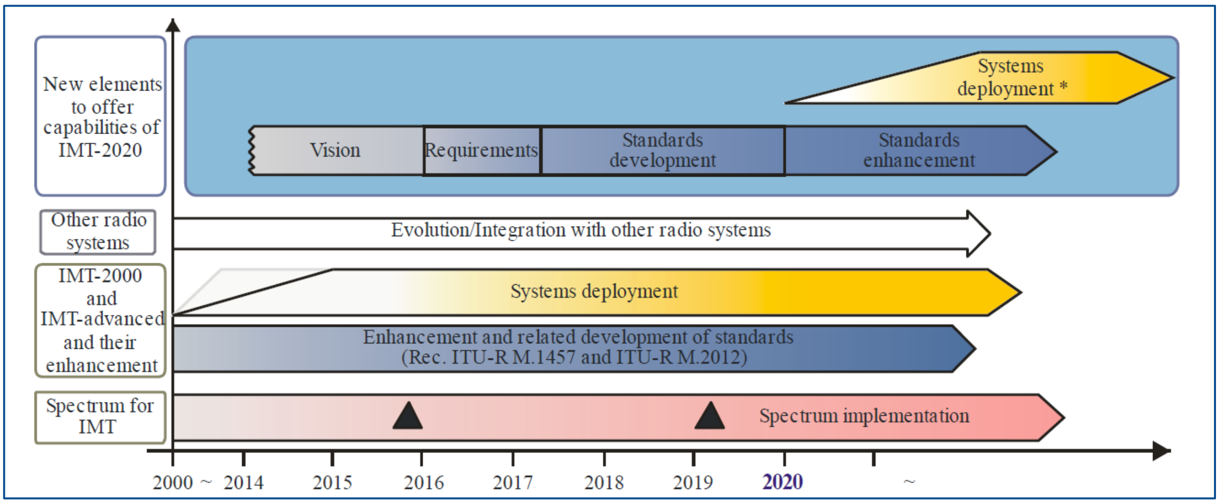
New elements to offer capabilities of IMT-2020
Systems deployment *
Vision
Requirements
Standards development
Standards enhancement
Other radio systems
Evolution/Integration with other radio systems
IMT-2000 and IMT-advanced and their enhancement
Systems deployment
Enhancement and related development of standards (Rec. ITU-R M.1457 and ITU-R M.2012)
Spectrum for IMT
Spectrum implementation
2000 ~ 2014
2015
2016
2017
2018
2019
2020
~

Figure 17. Stages and expected deadlines for IMT-2020[53]

All of the work that ITU is conducting on IMT-2020 is following a roadmap that is detailed in Figure 18, with the completed stages indicated in green and those to come in blue. The work that is underway today concerns 5G spectrum aspects, prerequisites and assessment criteria, along with the different technical studies and proposals, which are a prerequisite to the standardisation phase of the work.

---

[53] Recommendation ITU-R M.2083-0 (09/2015), https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

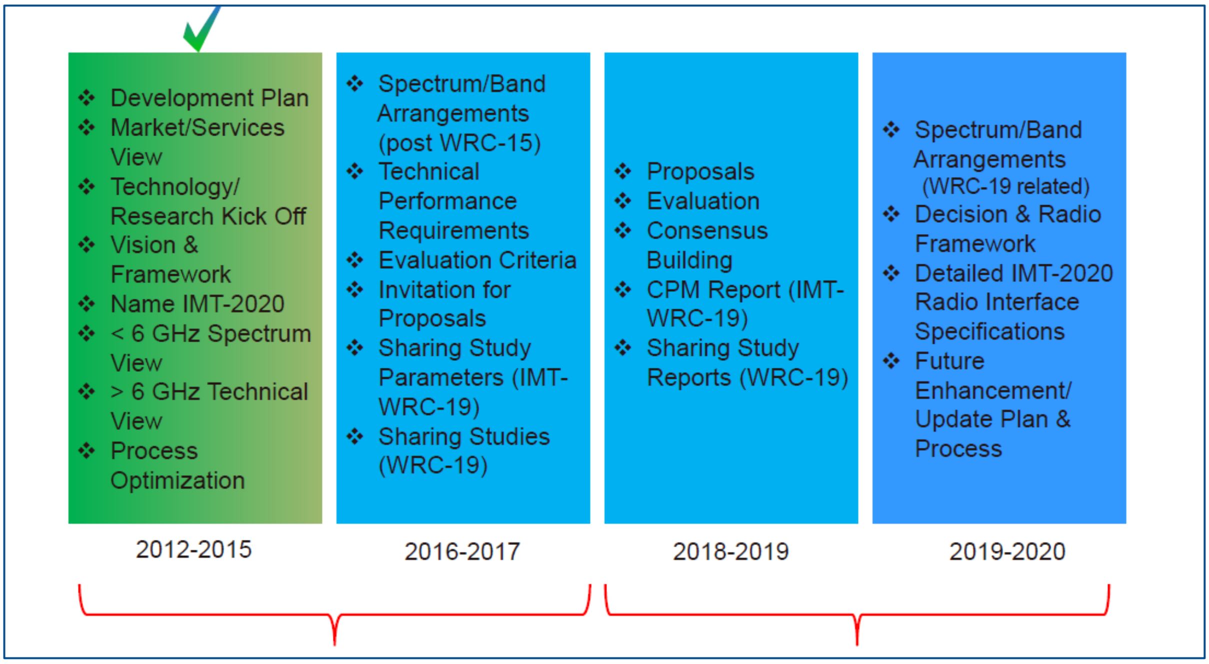
✓
❖ Development Plan
❖ Market/Services View
❖ Technology/ Research Kick Off
❖ Vision & Framework
❖ Name IMT-2020
❖ < 6 GHz Spectrum View
❖ > 6 GHz Technical View
❖ Process Optimization
2012-2015
❖ Spectrum/Band Arrangements (post WRC-15)
❖ Technical Performance Requirements
❖ Evaluation Criteria
❖ Invitation for Proposals
❖ Sharing Study Parameters (IMT-WRC-19)
❖ Sharing Studies (WRC-19)
2016-2017
❖ Proposals
❖ Evaluation
❖ Consensus Building
❖ CPM Report (IMT-WRC-19)
❖ Sharing Study Reports (WRC-19)
2018-2019
❖ Spectrum/Band Arrangements (WRC-19 related)
❖ Decision & Radio Framework
❖ Detailed IMT-2020 Radio Interface Specifications
❖ Future Enhancement/ Update Plan & Process
2019-2020

Figure 18. WP 5D[54] work programme

3GPP

Based on ITU Recommendations, 3GPP has played a major role in the success of LTE over the past several years, which has been the fastest growing cellular technology to date: never before has a new radio standard been adopted and deployed as rapidly and as widely from the finalisation of the first version of its standardisation (for LTE Release 8, in December 2008).

For the first time, under the LTE acronym, the entire mobile industry agreed on a single new technology (contrary to 3G where 3GPP and 3GPP2 co-exist, each backing their version of 3G standardisation that complies with IMT-2000 criteria), and so enabling unprecedented economies of scale and momentum in the ecosystem.

After Release 8, the work performed by 3GPP has been centred on the following strategic areas:

- Enhancing LTE radio standards to further improve capacity and performance;
- Enhancing system standards to make LTE and EPC (Evolved Packet Core, the core LTE network technology) available to new business segments;
- Introducing improvements for system robustness, especially for handing exponential smartphone traffic growth.

These areas of focus have made it possible to map out the general path of evolution from LTE to LTE-Advanced (Releases 10 to 12) and later LTE-Advanced Pro (Releases 13 and 14) while awaiting 5G.

The ambitiousness of 5G requirements, the tight timeline “imposed” by the market, along with the different national tendencies have pushed 3GPP to define two stages of specification work:

1. A first stage that will end in the second half of 2018, with the termination of Release 15, and which will address the most urgent matters with respect to commercial requirements;

---

[54] https://www.itu.int/en/membership/Documents/missions/GVA-mission-briefing-5G-28Sept2016.pdf

2. A second stage that will end in December 2019, which marks the end of Release 16, which will address the other use cases and prerequisites identified in the IMT-2020 Vision Recommendation.

The Work Plan for Release 15, the first set of 5G specifications, was ratified at the plenary meeting of the TSG#72 (Technical Specification Group #72) 3GPP working groups in June 2016. This plan includes a set of intermediate tasks and checkpoints to steer ongoing work in the different groups. This Work Plan sets out the transition from current studies to the standardisation phase of the work:

- Starting in December 2016: beginning of standardisation work by TSG SA2 (Technical Specification Group System Architecture);
- Starting in March 2017: specification begins for the 5G NR (New Radio) interface in the TSG RAN (Technical Specification Group Radio Access Network).

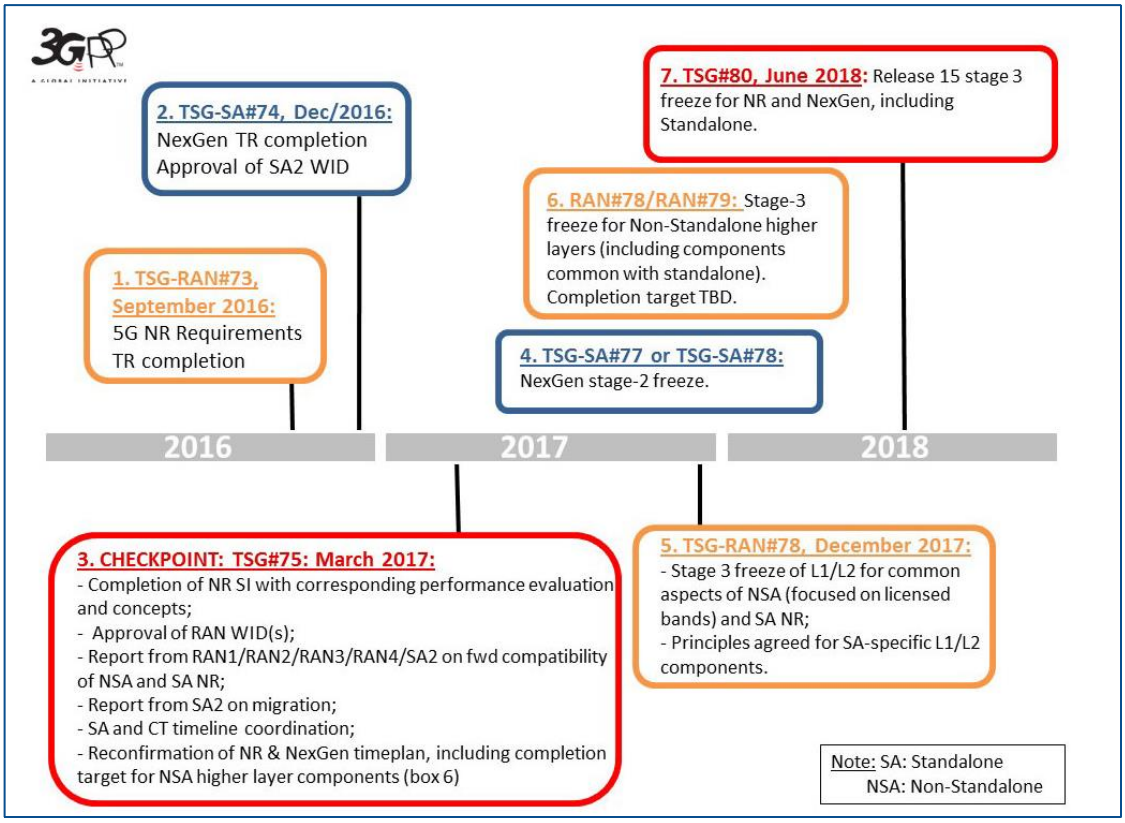
3GPP
A GLOBAL INITIATIVE
2. TSG-SA#74, Dec/2016:
NexGen TR completion
Approval of SA2 WID
7. TSG#80, June 2018: Release 15 stage 3 freeze for NR and NexGen, including Standalone.
6. RAN#78/RAN#79: Stage-3 freeze for Non-Standalone higher layers (including components common with standalone). Completion target TBD.
1. TSG-RAN#73, September 2016:
5G NR Requirements TR completion
4. TSG-SA#77 or TSG-SA#78:
NexGen stage-2 freeze.
2016
2017
2018
3. CHECKPOINT: TSG#75: March 2017:
- Completion of NR SI with corresponding performance evaluation and concepts;
- Approval of RAN WID(s);
- Report from RAN1/RAN2/RAN3/RAN4/SA2 on fwd compatibility of NSA and SA NR;
- Report from SA2 on migration;
- SA and CT timeline coordination;
- Reconfirmation of NR & NexGen timeplan, including completion target for NSA higher layer components (box 6)
5. TSG-RAN#78, December 2017:
- Stage 3 freeze of L1/L2 for common aspects of NSA (focused on licensed bands) and SA NR;
- Principles agreed for SA-specific L1/L2 components.
Note: SA: Standalone
NSA: Non-Standalone

Figure 19. Roadmap for 3GPP standardisation groups[55]

If 5G is viewed as the technology capable of transforming society, and even of ushering in the next industrial revolution by impacting multiple (vertical) sectors with new business models and in a way that benefits the economy, notably in Europe, in all likelihood it will initially be driven forward by eMBB (Enhanced Mobile Broadband, cf. 1.2).

With this in mind, 3GPP defined the framework for Release 15, stage 1 specifications of 5G standardisation, as detailed earlier, whose work will focus on the following:

- Standalone and Non-Standalone NR operation: the Standalone version to work independently of LTE technology, and the Non-standalone NR being highly interoperable with LTE[56].

---

[55] https://5g-ppp.eu/wp-content/uploads/2016/11/01_10-Nov_Session-3_Dino-Flore.pdf

- Target use cases: starting with Enhanced Mobile Broadband (eMBB) as well as Low Latency and High Reliability to enable some URLLC (Ultra-reliable and Low Latency Communications, cf. 1.2) use cases.
- Simultaneous examination of frequency ranges below 6 GHz and above 6 GHz.

Release 16, the second phase of 5G specifications, will then focus more on other segments such as mMTC (Massive Machine Type Communications, cf. 1.2), for instance, to tackle IoT use cases that require higher quality of service than what LPWAN *(*Low-Power Wide-Area Network)[57] can provide.

---

[56] In its non-standalone version, the NR control plane is LTE's. in other words, the 4G network controls the 5G carriers and spreads users across the different bands and technologies.

[57] Sigfox and LoRa LPWAN

Annex 2 Entities Arcep met with

ERICSSON
QUALCOMM®
HUAWEI
SAMSUNG
NOKIA
AIRBUS DEFENCE & SPACE
THALES
RENAULT
Systematic
Paris Region Systems & ICT Cluster
TELECOM ParisTech
leti
ceatech
Google
ETSI
World Class Standards
bouygues TELECOM
free mobile
numericable THD
SFR
orange™
GSMA™

Annex 3 Figures

Figure 1. 5G driving industrial and societal changes 4

Figure 2. Performances required by vertical sectors 6

Figure 3. What we do over the network today in one minute 7

Figure 4. 3GPP timeline for 5G 9

Figure 5. 5G use cases 10

Figure 6. Comparison between 4G and 5G with respect to the eight key performance indicators 12

Figure 7. Representation of the "average user experience data rate" KPI for 4G and 5G and "Peak data rate" KPI for 3G, 4G and 5G 13

Figure 8. Key performance indicators for the three 5G use cases 14

Figure 9. Example of the use of a beamforming antenna to connect Wi-Fi access points 16

Figure 10. Illustration of full-duplex, compared to FDD and TDD 16

Figure 11. Illustration de multiplexing *power domain* NOMA 17

Figure 12. Centralisation of the control plane in a software-defined network 18

Figure 13. Illustration of CloudRAN network architecture 19

Figure 14. 5G PPP vs. 3GPP and ITU roadmaps 22

Figure 15. Millimetre wave frequencies identified at WRC-15 28

Figure 16. Timeline of IMT developments and deployments 35

Figure 17. Stages and expected deadlines for IMT-2020 36

Figure 18. WP 5D work programme 37

Figure 19. Roadmap for 3GPP standardisation groups 38

HEAVY
READING

White Paper

Designing 5G-Ready Mobile Core Networks

Prepared by

Gabriel Brown
Senior Analyst, Heavy Reading
www.heavyreading.com

on behalf of

affirmed®

www.affirmednetworks.com

and

JUNIPER NETWORKS

www.juniper.net

September 2016

5G Core Service & Performance Requirements

An industry "Vision for 5G" is now established and supported worldwide. Services and performance requirements have been identified, and the industry is moving into the development and implementation phases with a view to commercial launch of 5G mobile service around 2019 and fixed wireless access using mmWave radio from 2017. To meet these schedules, leading operators – especially Tier 1s with high expectations and aggressive deployment timelines – are working to make their networks 5G-ready ahead of commercial deployment.

This white paper argues that with a 5G-ready technology strategy, operators can prepare for rapid 5G service launch in a way that optimizes their investment in next-generation IP and mobile core platforms over the next three years. Specifically, it discusses 1) how 5G services drive a requirement for an IP services fabric to connect the distributed data centers that will host 5G network functions, content and applications and 2) the development of cloud-native, service-orientated core networks for advanced 4G and 5G networks.

5G Services Dimensions

Development of the 5G system architecture, and associated core network, is being led by 5G service requirements. **Figure 1** shows the output from the 3GPP SMARTER study group that has investigated the service dimensions of 5G.

Figure 1: Service Dimensions in Future 5G Networks

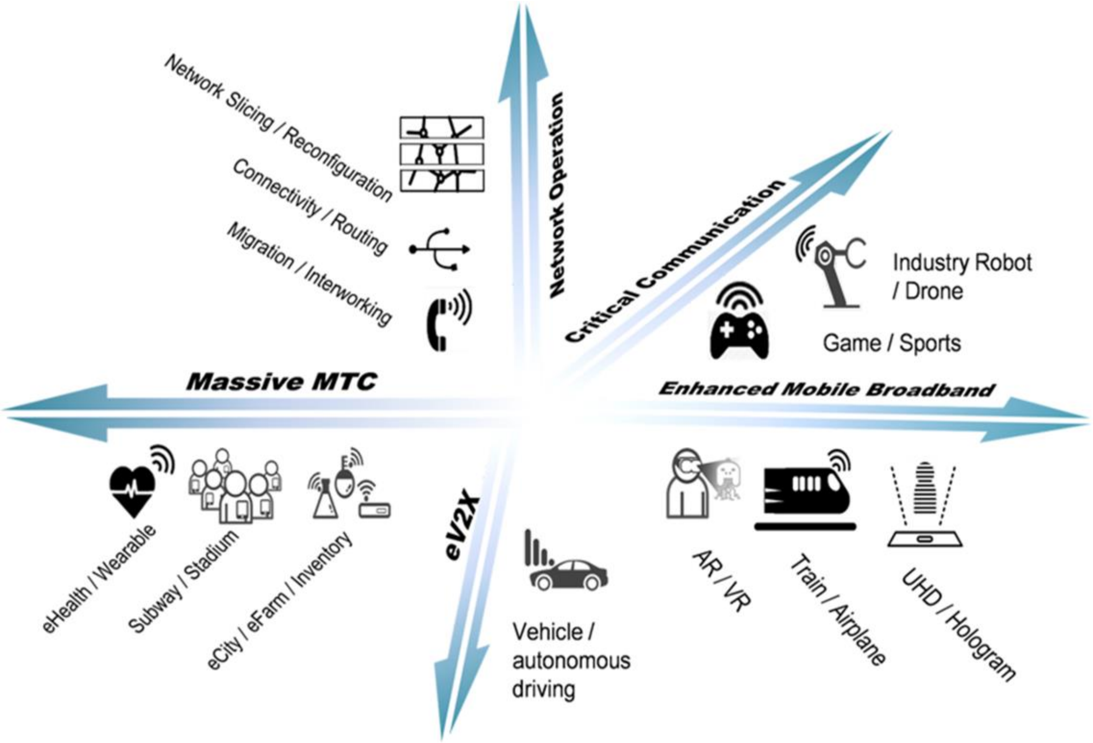
Network Slicing / Reconfiguration
Connectivity / Routing
Migration / Interworking
Network Operation
Critical Communication
Industry Robot / Drone
Game / Sports
Massive MTC
Enhanced Mobile Broadband
eHealth / Wearable
Subway / Stadium
eCity / eFarm / Inventory
eV2X
Vehicle / autonomous driving
AR / VR
Train / Airplane
UHD / Hologram

Source: 3GPP SMARTER

The chart maps closely to the now famous triangle representation of 5G services made up of enhanced mobile broadband (eMBB), low latency and mission critical communications and massive machine-type services (mMTC). However, it adds two important axes: a dedicated eV2X axis for vehicular services and, significantly, a network operations axis.

The dedicated operations axis is important because it makes explicit that automation is fundamental to 5G. To address new commercial opportunities, operators need the ability to create service-specific "network slices" that extend end to end across the infrastructure. Automating set up and management of these "network slices," including configuration of the underlying IP network, is critical to the 5G business case.

These service dimensions incorporate many different performance requirements. **Figure 2** shows IMT-2020 performance targets, developed by the ITU, for 5G. To the left, these are shown in comparison to IMT Advanced (4G). To the right, performance requirements are mapped to the three major use-case categories.

Figure 2: IMT-2020 Performance Targets

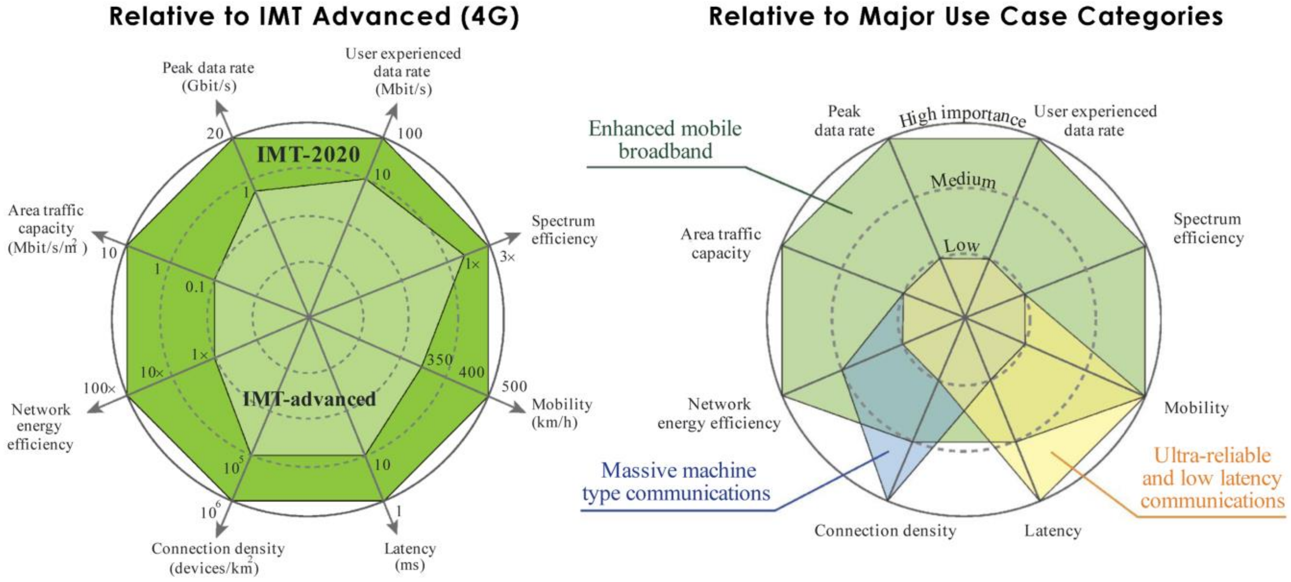

Relative to IMT Advanced (4G)

Relative to Major Use Case Categories

Source: ITU Recommendation ITU-R M.2083-0, September 2015

Probably the most influential factor on 5G physical network deployment will be the low-latency requirement. To deliver services over a wide-area network (WAN) with a consistent 5-10ms roundtrip time is extremely challenging and will drive a new system architecture deployed on a distributed cloud infrastructure.

A 5G-Ready Network Investment Strategy

Operators now have a good idea of what 5G will look like in scope and a reasonable view into the initial deployments models. Development of normative standards for 5G new radio (NR) and the next-generation core network (NG Core) is underway with the first specifications – and the first "compatible" pilot deployments – expected in mid 2018. Although there is uncertainty about the details of the specifications under development, the industry has already undertaken significant R&D work on 5G candidate technologies, making it is possible for operators to pursue a 5G-ready network strategy with a reasonable degree of confidence that they will be able to rapidly deploy 5G, at scale, as the technology becomes commercially available.

From an IP and core network perspective, five tiers of a 5G-ready investment strategy are, from the ground up, as follows:

1. **Develop a distributed data center footprint**. The Central Office Re-architected as Data Center (CORD) model provides a good reference. The idea is for operators to develop physical assets located close to the edge and transform them into distributed data centers. The edge location is critical for low-latency 5G services and, as Mobile Edge Computing (MEC) shows, the edge-cloud is valuable in many networking scenarios – it is not solely a 5G investment. In time, the distributed data center will also come to support Cloud RAN (C-RAN) and Virtual RAN (V-RAN) hub sites and will host core network functions, such as distributed user-plane nodes.
2. **Create an "IP Services Fabric" for 5G with software-defined networking (SDN) control**. Edge-cloud locations running 5G network functions and services will require high-performance, secure connectivity. Investment in wide-area SDN is already underway to create an "IP services fabric" that connects centralized data centers, distributed cloud locations and cell sites. This IP services fabric provides routing, security, service chaining, redundancy and orchestration. Again, this investment is, in many aspects, not unique to 5G, but common to high-performance networking at the cloud edge.
3. **Invest in cloud-based Evolved Packet Core (EPC) and NG Core.** With a distributed cloud infrastructure, it is logical to redesign core network functions to take advantage of this asset to meet the strict latency requirements of 5G services and efficiently manage data traffic. Some aspects of NG Core are already reasonably well understood, (for example, control- and user- plane separation or "CUPS"), while others remain works in progress (for example, mobility and session management). In general, expect control-plane separation to apply across the 5G architecture, from radio to core.
4. **Use network slicing for a service-optimized core.** Network slicing is an important bridge from a 4G core to a new 5G core and is a key component of a 5G-ready network investment strategy. In this model, virtual network functions (VNFs) are created per service and data is routed via an optimized processing path across the network. By isolating services into virtual network slices, operators can offer better security, more efficient transport, optimized core network processing and appropriate service quality. A network slice created for an Internet of Things (IoT) customer, for example, may support both 4G and 5G access networks.
5. **Automation and service orchestration.** Automating network operations provides the ability to launch new services rapidly, to reach new market segments, to evaluate success/failure quickly, and then to modify, scale and repeat. Significant progress is being made in applying service orchestration to 4G-Long Term Evolution (LTE) networks, especially where software-based core networks are deployed. Network slicing is a good example of the need for automation because there will be many more logical entities to commission, provision and manage. "Automated service orchestration" is a good, tangible example of 5G-ready functionality.

The 5G Core Network

The new 5G core network is now reasonably well understood: It will be "cloud-native," it will make extensive use of network slicing, and it will operate in concert with a new model-driven service orchestration layer. The industry can therefore start to develop NG Core products that can be adapted as standards emerge and solidify.

From a practical perspective, for the initial deployment, operators can support 5G radio access on a "5G-ready" EPC and then migrate to a new NG Core over time. In both cases, the "IP services fabric" provides SDN-controlled connectivity and related IP services, across a distributed cloud infrastructure.

Virtualized & Cloud-Native Mobile Core

Operators have made good progress on virtual EPC over the past couple of years. The largest networks now support more than 15 million subscribers (AT&T has discussed this publicly); some progressive operators have started to refresh their main EPC networks using multivendor network functions virtualization (NFV) (e.g., Docomo, Etisalat); and others (e.g., Vodafone) have deployed virtual core elements for IoT services.

Figure 3 shows two migration paths to a "cloud-native" 5G core. The blue line shows a virtualized 4G core as a stepping stone to a "5G-ready" core and then a full 5G core. The grey line shows a more conservative option where the operator scales the classic EPC in the near-term and then makes a bigger leap to 5G later.

Figure 3: A Preliminary Cloud-Based 5G Architecture

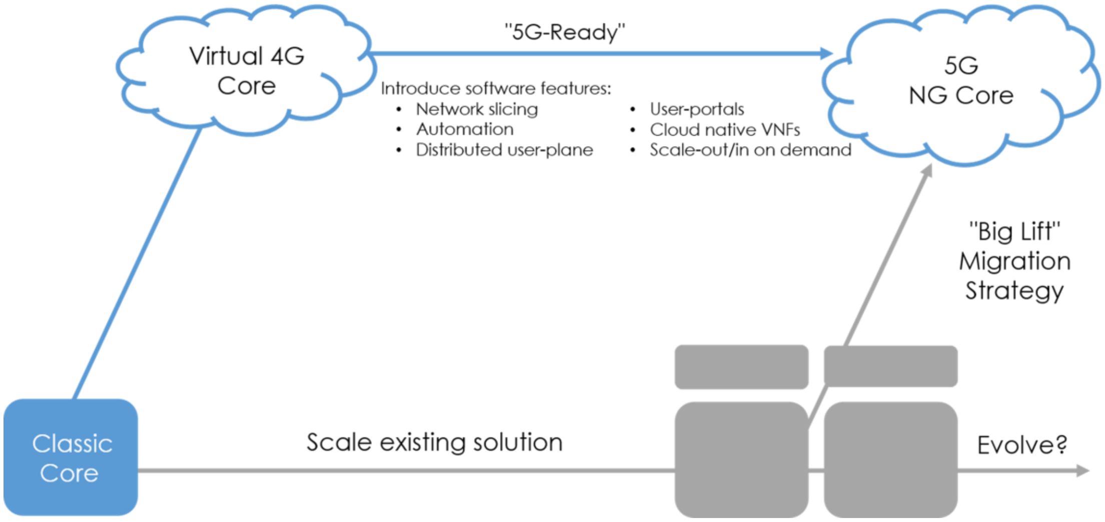
Virtual 4G Core
"5G-Ready"
5G NG Core
Introduce software features:
• Network slicing
• Automation
• Distributed user-plane
• User-portals
• Cloud native VNFs
• Scale-out/in on demand
"Big Lift" Migration Strategy
Classic Core
Scale existing solution
Evolve?

Source: Heavy Reading

There is debate about what exactly "cloud native" means, particularly in the context of stateful telecom network functions. Nevertheless, Heavy Reading's established view is that operators that deploy virtual EPC – even at small scale – gain experience in how these systems work, how to operate them, and how to evolve them, and that in turn will give them a sustainable advantage in the longer-term transition to software-centric 5G mobile networks.

Migrating 5G RAN From EPC to NG Core

A 5G-ready core strategy is determined, in part, by how the operator plans to introduce 5G radio. There are two basic possibilities: operate 5G RAN using an EPC or using a new NG Core. In practice, operators with plans to launch 5G early are likely to start with an EPC and migrate to NG Core over time, as shown in **Figure 4** below.

Figure 4: Migrating 5G RAN from EPC to NG Core

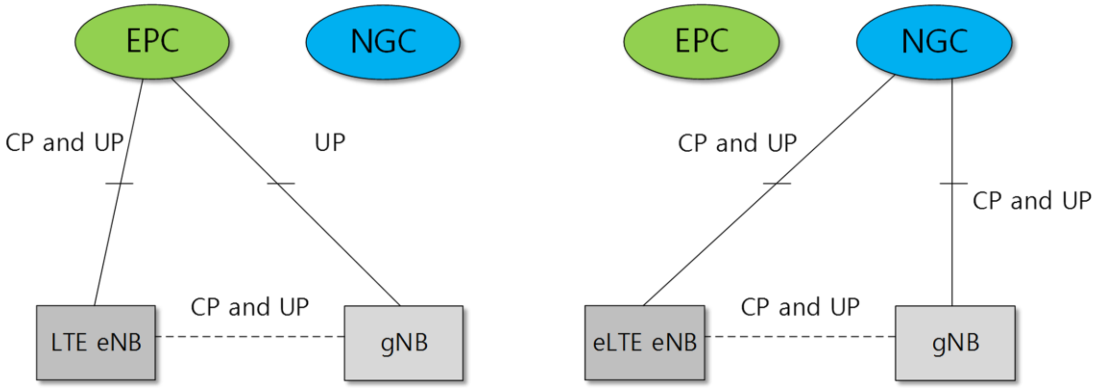

Source: 3GPP TR 38.801 v0.4.0

The diagram is taken from the most recent version of the 3GPP "Study on New Radio Access Technology" (TR 38.801), which will inform development of standards in Release 15. Several permutations of the architecture are under consideration; however, in simple terms, to the left, the new 5G base station (gNB) user-plane interface connects directly to a 4G EPC, while the control-plane functions, such as tracking, paging, etc., are provided by an evolved 4G base station (eNB), which in turn also connects to the EPC. In this scenario, the EPC requires little or no modification, making this a fast and simple way to deploy 5G radio from a core perspective.

Over time, both 4G and 5G base stations can migrate to a new NG Core, which will provide both control- and user-plane functions. At this stage, NG Core becomes the primary core network for 4G and 5G access, as shown to the right in the diagram. This is conceptually similar to how EPC supports 3G and 4G access networks.

Note that in cases where 5G radio is deployed for fixed wireless access, there is no need for an LTE RAN to provide control-plane functions to the 5G user device; a standard EPC is sufficient, although it would need to provide session management for the 5G access. Some of the first 5G deployments are expected to use this model.

Network Slicing: A Key Bridge to 5G

Network slicing is one of the key bridges between the 4G and 5G core. To support diverse service types, operators will use multiple core networks deployed as "network slices" on a common IP services infrastructure. The idea, shown in **Figure 5**, is to create virtual core network instances (or "slices") dedicated to different services. Each slice can be optimized for the traffic profile and the commercial context of the associated service – for example, IoT, public safety, mobile virtual network operator (MVNO), connected car, voice over WiFi or enterprise services. Network slices can be two dimensional in the sense that they can be both service- and customer-specific.

Figure 5: Dedicated Packet Core Network Slices

- Virtualization enables multi-tenant and single-tenant private core networks
- "Slices" configured according to service type & traffic profile
- Multiple options to steer user traffic into a core network processing path
- Mechanisms in 3G/4G include APN routing, MoCN, DECOR

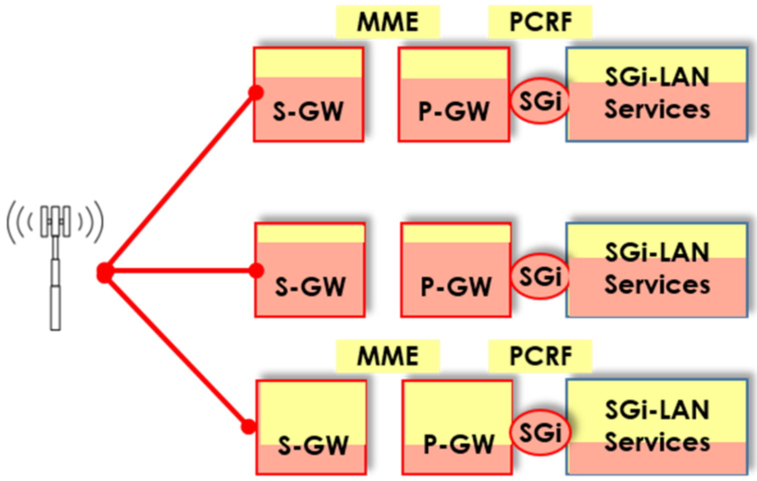

Source: Heavy Reading

IoT is an example of how operators can use "slicing" to support a 5G-ready core strategy. Since many of the core network design parameters are similar between 4G and 5G, investment in a software-based IoT core network can be made with the expectation that the same core (with software updates) will also support IoT services on 5G in future. Moreover, because devices from both access types will connect to a common IoT core, operators will be able to develop integrated 4G/5G IoT strategies that optimize investment and enable them to go to market with narrowband IoT services before 5G specifications are released and radio equipment is deployed.

5G will, naturally, bring more capabilities to network slicing. Most importantly, it is expected that the slice will run end to end across the RAN, core and transport network. Radio is typically deployed as a shared resource, which means that 5G slicing will involve advanced self-organizing network (SON) capabilities and hierarchical, slice-aware scheduling on the air interface. On the network side, the operator can use SDN to reserve resources for the network slice on the IP services fabric.

NG Core for 5G

A new core network offers important benefits, particularly for services with demanding performance requirements, and is an important part of the 5G architecture. The high-level view of the NG core network architecture, as it looks in September 2016, according to TR 23.799 (the Technical Report on the NextGen System Architecture) is shown in **Figure 6**.

The chart shows the primary elements of the new 5G system architecture, including the device (UE), the radio access network (NG-RAN) and the core network that comprises the control-plane (NG-CP) and user-plane (NG-UP) functions. The separation of control and user planes is a direct extension of the "CUPS" concept being developed for advanced EPCs in 4G networks and is another example of how operators can invest in 5G-ready core network designs in the near term.

Figure 6: High Level View of NG Core Architecture (Proposed)

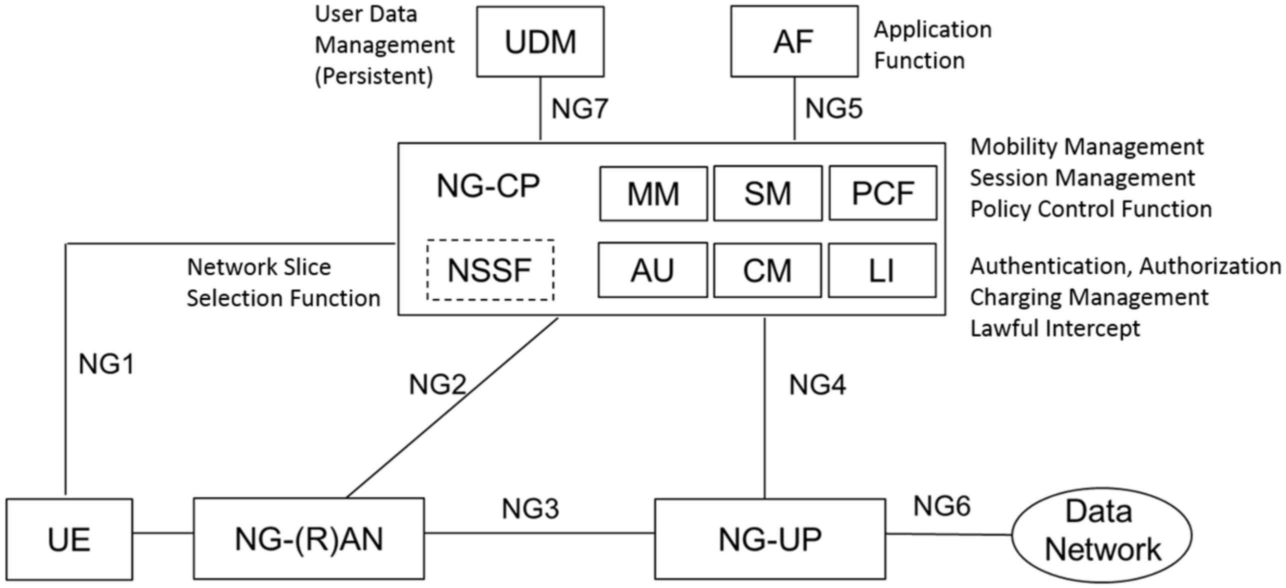
User Data
Management
(Persistent)
UDM
AF
Application
Function
NG7
NG5
NG-CP
MM
SM
PCF
Mobility Management
Session Management
Policy Control Function
Network Slice
Selection Function
NSSF
AU
CM
LI
Authentication, Authorization
Charging Management
Lawful Intercept
NG1
NG2
NG4
UE
NG-(R)AN
NG3
NG-UP
NG6
Data
Network

Source: 3GPP, Adapted from TR 23.799 v0.8.0, September 2016 (Figure 7.2.4-1)

It is interesting to compare this new functional architecture with the existing EPC. **Figure 7** maps the new NG Core elements to their equivalent in EPC. It shows significant overlap between the EPC and NG Core and provides confidence that today's state-of-the-art EPC is reasonably close to the next-generation core in terms of functionality, albeit that the interfaces and protocols will change/evolve.

Figure 7: Mapping NG Core & EPC Functions

NG Core Network Function	Approximate Equivalent EPC Function
UDM (User data management)	SPR
AF (Application function)	AF
NSSF (Network slice selection function)	MME (DECOR) + HSS
MM (Mobility management)	MME + SGW / MME + SGW-C
SM (Session management)	PGW/PGW-C
PCF (Policy control function)	PCRF
AU (Authentication & Authorization)	HSS/AAA
CM (Charging management)	PGW + PCRF, OCS, OFCS
LI (Lawful intercept)	LI
NG-UP (User-plane function)	SGW + PGW / SGW-U + PGW-U

Source: Heavy Reading, Affirmed, Juniper

Key Capabilities of the Next-Gen System Architecture

In addition to proposed architectures, the Technical Report provides guidance on the features and services NG Core should support. It identifies 19 key issues for the Next-Gen System Architecture, which can be grouped into four categories:

- **Flexible Deployment**: Including network slicing, user-plane network selection, network function granularity (e.g., decomposition of VNFs and service chains), interworking and migration, a policy and charging framework.
- **Flexible Access Support**: Including variable core/access splits, a flexible authentication framework (with varying credentials according to device-type, use case and policy), support for relays and multi-hopping, improved network discovery and selection mechanisms.
- **Connectivity**: Including session management, mobility management, a quality-of-service (QoS) framework and service and session continuity (e.g., across accesses).
- **Adapting Existing Capabilities**: Including network/service capability exposure, multicast and broadcast, IP Multimedia Subsystem (IMS) support, off-network communications (e.g., *ad hoc* networks for public safety).

Implementation of these capabilities is tightly linked to both the design of the NG Core and the underlying IP service infrastructure. The 3GPP specifications do not reference connectivity services, but, in practice, there is a close relationship.

A Distributed 5G Services Fabric

The infrastructure onto which 5G will be deployed should support multiple, demanding use cases. This drives a need for a high-performance wide-area IP services fabric controlled (or "orchestrated") by SDN. This IP services fabric should provide connectivity between many different distributed data centers in a meshed architecture that provides resiliency and scalability, and should be programmable such that it can support dynamic, service-specific network slices. In effect, the IP services fabric makes distributed centers act in a unified manner – i.e., behave as one integrated data center. The concept is shown in **Figure 8**.

The design of the distributed IP services fabric is formed by the variety of use cases, the requirement for low latency and high availability, and the need to scale efficiently. In tandem, the mobile network architecture will evolve to a more distributed model. To meet 5G service requirements, NG Core will be deployed using the "CUPS" model, with user-plane nodes hosted at distributed data center locations, and control-plane nodes at more centralized locations, interconnected by the IP services fabric.

Service orchestration will generally be domain specific, with SDN-controlled IP services, cloud resource management, NFV lifecycle management and end-user service orchestration, all operating quasi-independently in a layered architecture. Coordination between the layers will use "cross domain orchestration" to ensure a network slice contains all the networking components needed to deliver the service.

Depending on how the architecture evolves, there will be a need to support highly accurate, 1588v2-based timing (frequency and phase) within this IP services fabric. Virtual RAN/Cloud RAN is an example on the network side that benefits from accurate network timing. On the customer side, it is expected that some market segments, such as the financial industry, will also require timing to be incorporated into network slices.

Figure 8: IP Services Fabric for 5G Networks

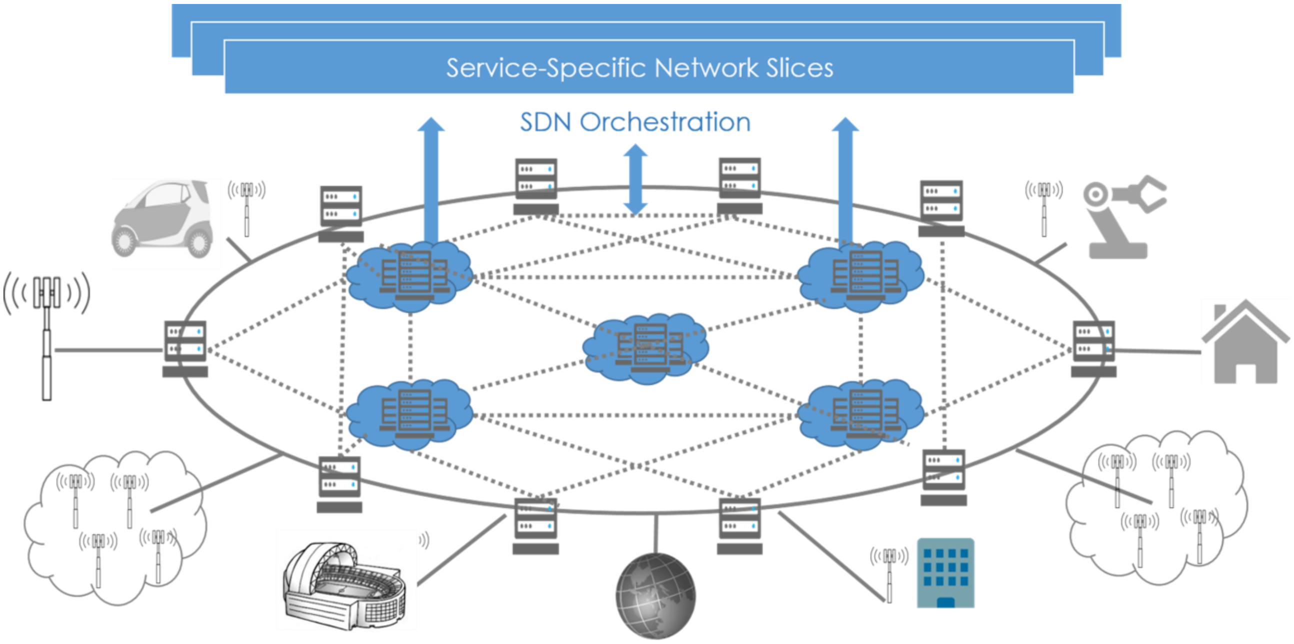

Source: Heavy Reading

Distributed Data Centers

To provide access network services, telecom operators have a physical footprint – in the form of central offices, base station controller sites and transport aggregation sites – that they can convert into distributed or micro data centers.

Depending on the service requirements, these data centers can be used to terminate access connections using virtualized edge functions and become the obvious place to host latency-sensitive content and applications. For centralized cloud providers, this capability is harder to replicate unless they build out an extensive physical footprint or partner with an access provider, and, therefore, this strategy provides operators with a potentially important and sustainable competitive advantage.

The "edge cloud" model will be critical to 5G and is the subject of detailed R&D across the industry. As an example, a preliminary architecture developed by the SuperFluidity project (a European 5G research project) is shown in **Figure 9**. Work on this particular project, which is academic in focus but contains several industrial partners, is scheduled to finish by the end of 2017.

As the name implies, the project aims to achieve superfluidity in the network: the ability to instantiate services on the fly, run them anywhere in the network (core, aggregation, edge) and shift them transparently to different locations. This is the essence of the "IP Services Fabric" for 5G. In this new network, operations will also need to move from a "red light, take action" model to an automated model where the orchestration and associated next-gen operations support system (OSS) detect and fix problems with the IP and NFV layers, and re-optimize the network accordingly.

Figure 9: A Preliminary Converged Cloud-Based 5G Architecture

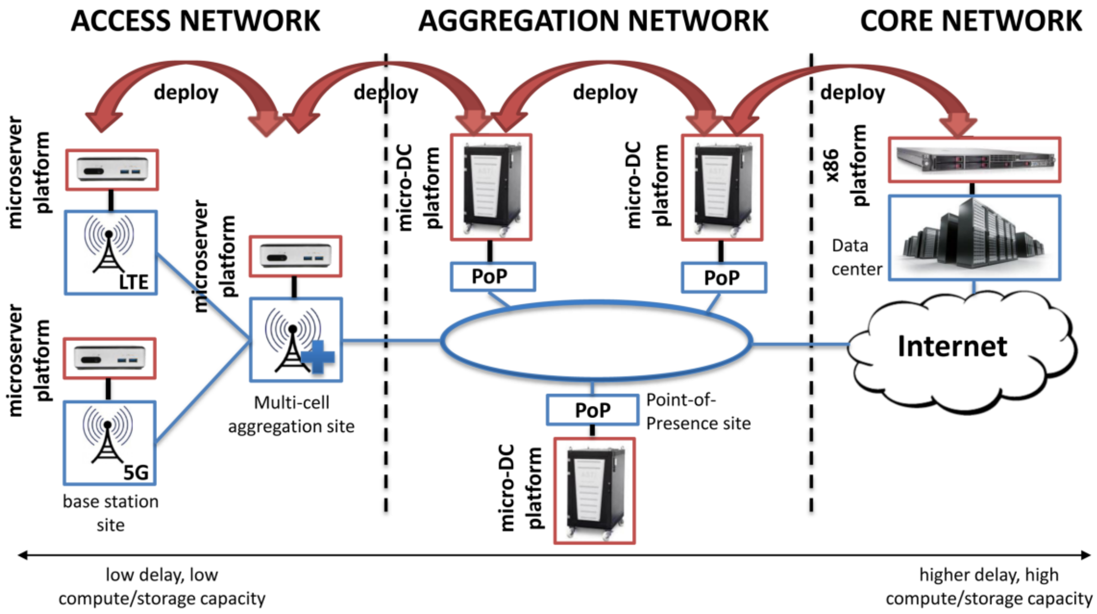
ACCESS NETWORK
AGGREGATION NETWORK
CORE NETWORK
deploy
deploy
deploy
deploy
microserver platform
LTE
microserver platform
5G
base station site
microserver platform
Multi-cell aggregation site
micro-DC platform
PoP
micro-DC platform
PoP
PoP
Point-of-Presence site
micro-DC platform
x86 platform
Data center
Internet
low delay, low compute/storage capacity
higher delay, high compute/storage capacity

Source: SuperFluidity

Central Office Re-architected as a Data Center (CORD)

There are several industry and service provider initiatives to define the architecture for SDN-enabled distributed data centers. The Central Office Re-architected as a Data Center (CORD) initiative hosted by the Linux Foundation is one example of how operators aim to make better use of these assets.

According to the CORD initiative, AT&T alone operates 4,000 to 5,000 central offices, each serving 10,000 to 100,000 residential, enterprise and mobile customers. These central offices contain fragmented vendor hardware with multiple physical appliances installed per site (AT&T said it has 300+ unique appliances deployed in its central offices nationwide). The opportunity is to re-architect these central offices to support edge cloud infrastructure and deploy VNFs in place of appliances.

Each CORD location will be connected using an SD-WAN, making it possible to load-balance content and NFV workloads across the distributed cloud using the same SuperFluidity concepts discussed above. In the CORD case, the focus is on the ONOS controller, but in the sense that this is a generic architecture, multiple SDN controller options are viable and attractive. Note that for I/O-intensive workloads, the selection of an edge location should consider the physical transport resources available at the site. It would not make sense, for example, to deploy a 5G user-plane node for fixed wireless access in a central office that does not have a high bandwidth connection to the Internet.

In March 2016, the Linux Foundation announced the M-CORD initiative for mobile operators, backed by AT&T, SK Telecom, Verizon, China Unicom and NTT Communications. It highlighted three key aspects of the architecture, saying that it:

- Will use the same CORD principles of elastic commodity cloud and SDN to bring data center economies and cloud agility to the mobile edge.
- Will demonstrate integration of disaggregated/virtualized RAN, disaggregated/virtualized EPC and mobile edge services.
- Will partner with the SDN controller groups to accelerate adoption of open source SDN and NFV solutions and realize the benefits of the cloud.

Given the timing of the work, we expect M-CORD and similar initiatives to turn their focus toward "5G-ready" core networks as the architecture, interfaces and protocols for NG Core become more clear. A common CORD and M-CORD implementation, with the same architecture and foundational technologies, will create a good foundation for fixed-mobile network convergence, enabling access agnostic services – an important objective of many operators pursuing 5G.

Mapping Mobile Core to Distributed Cloud

The NG Core for 5G and EPC for advanced 4G networks must be mapped to the distributed cloud architecture. One approach would be to simply deploy more packet gateways (and mobility controllers) at the edge of the network to meet capacity and performance demands.

The challenge with this is that today's centralized packet core deployments are characterized by complex integration with surrounding network functions, such as policy, charging, IMS, SGi-LAN and routing services. By moving this model to the edge, the operator would, in effect, have to "distribute complexity," which is costly to deploy and, in particular, to manage. To meet 5G performance, scalability and automation requirements, a new architecture for packet core is needed that will make operation in the edge cloud infrastructure simpler and faster.

There are many aspects to this new architecture. Part of the solution is CUPS, as is currently being developed for 4G-LTE core networks. This involves extracting the control-plane functions from the gateway to leave a simpler, user-plane node. The gateway thus is "split" into S/PGW-U and S/PGW-C components that can that can scale independently, as shown in **Figure 6** above. A key benefit of the architecture is that the control plane, and all the associated complex interactions, can be centralized, while the user plane is distributed across the IP services fabric and scaled as required by the traffic load. This is shown in **Figure 10**.

Depending on the scaling needs, the S/PGW-U functions deployed at the edge data center can be implemented in several different ways: on a router, on a white box switch (potentially), on an existing PGW platform, as a virtualized function, or as part of a vRouter. Virtualized user-plane nodes can more easily be placed at the optimal location, as determined by the use case and this flexibility is a strong argument to deploy S/PGW-U nodes as VNFs rather than as hardware functions.

The S/PGW-C components would similarly be deployed as virtualized functions on a cloud platform, typically at a more centralized location. There may be an opportunity to collapse MME, SGW-C, PGW-C functions and perhaps other 3GPP control-plane functions into some form of unified "mobility controller" node. This has the potential to simplify operations and this same model is being pursued for NG Core.

In this architecture, virtualized SGi service LAN components can be placed where appropriate for the traffic coming from the distributed user-plane nodes. For example, some SGi functions could be on router-based compute blades, or on COTS x86 servers deployed at the distributed site, while others could be in the central data center or close to Internet peering points. The selection of the site for the SGi functions is carried out by the central orchestration system, which steers traffic into service chains on a per subscriber, application, bearer, device or combination basis.

Figure 10: Distributed User-Plane at Edge Data Centers

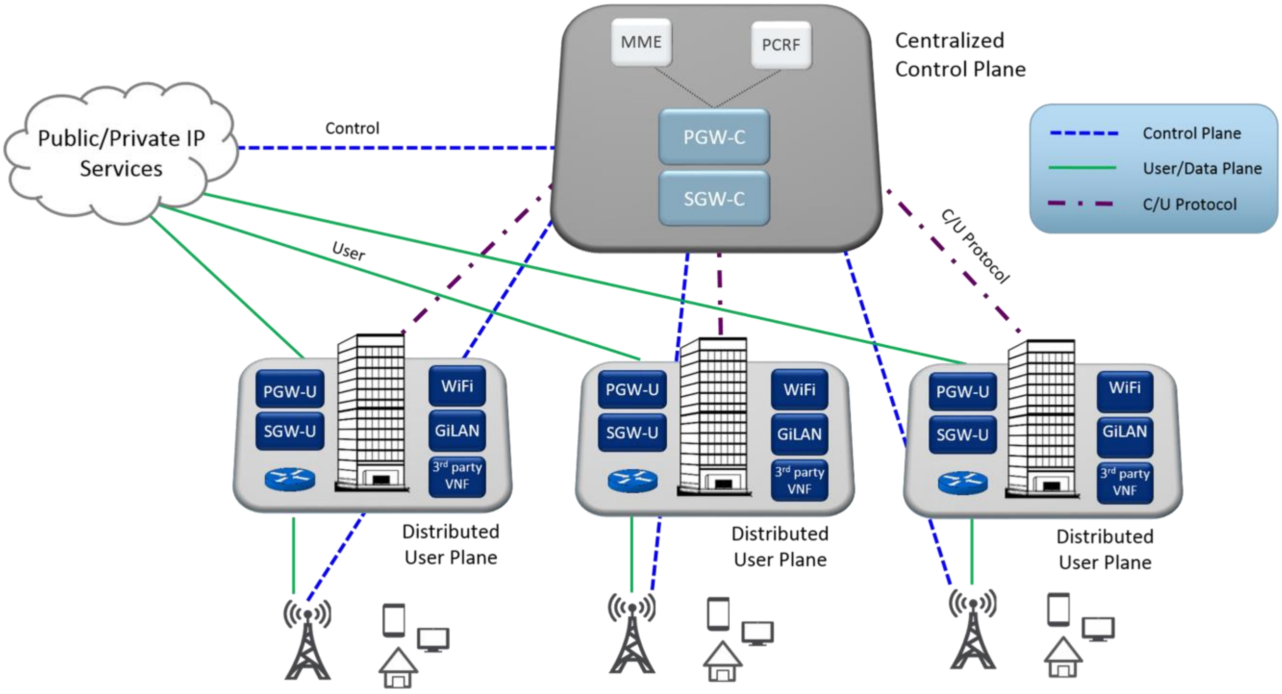
MME
PCRF
Centralized Control Plane
PGW-C
SGW-C
Public/Private IP Services
Control
User
C/U Protocol
Control Plane
User/Data Plane
C/U Protocol
PGW-U
SGW-U
WiFi
GiLAN
3rd party VNF
Distributed User Plane

Source: Affirmed Networks

This model of discrete control- and user-plane functions is expected to be fundamental to 5G and the NG Core. In this sense, CUPS can be viewed as an important part of a 5G-ready investment strategy.

A 5G-Ready Core Is Now a Priority

The industry now has a reasonable view of 5G service requirements, and progressive operators with aggressive deployment timelines are now working to prepare their networks for rapid deployment of 5G when equipment is available. This is driving investment in the critical IP services network needed to connect the edge cloud locations that will run 5G network functions, content and services.

On the mobile core side, development of cloud-native, service-orientated core networks for advanced 4G and 5G networks is underway. Network slicing provides a conceptual bridge between 4G and 5G investment and facilitates a faster insertion of new services into the network. Similarly, CUPS provides a reference for NG Core and the new 5G network architecture. With a 5G-ready technology strategy, operators can prepare for 5G service launch in a way that optimizes their investment in next-generation IP and mobile core platforms over the next three years.

About Affirmed Networks

Affirmed Networks' NFV solution has become the standard for the world's top mobile operators, who are embracing new business models and building new revenue streams by making the transition to virtualized architectures. The company's technology portfolio includes the Affirmed Mobile Content Cloud, the Affirmed Wi-Fi Gateway (serving as a TWAG/TWAP and an ePDG), Affirmed Service Automation Platform (ASAP) and Affirmed Virtual Probe and Analytics Solution. These virtualized solutions have come to represent the present and the future of virtualized mobile networks with extreme scalability, remarkable flexibility, comprehensive network orchestration and future-proof solutions for a 5G-ready architecture. Please find more information at www.affirmednetworks.com.

About Juniper Networks

Juniper Networks (NYSE: JNPR) is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at www.juniper.net, or connect with Juniper on Twitter and Facebook.

CONSULTING - SPECIFYING engineer

CFE Edu is an interactive, multi-media online platform that offers exclusive content covering engineering topics that are critical to your work. Catapult your career for

Research Webcasts Newsletters Magazine

LOGIN Username *******
Forgot password?
Subscribe
Manage Account
Search
Sponsored by:

Trending New Products Power Fire, Life Safety Electrical, Lighting Building Types Codes and Standards HVAC

PE New Products for Engineers Database
Find An Innovative Product

PE
Searching for a product?

Energy harvesting power for the Internet of Things

Energy harvesting wireless technology easily interconnects thousands of individual devices in a system, opening up unlimited processing and monitoring applications.

Laurent Giai-Miniet, EnOcean, Oberhaching (Munich), Germany
09/30/2013

Share

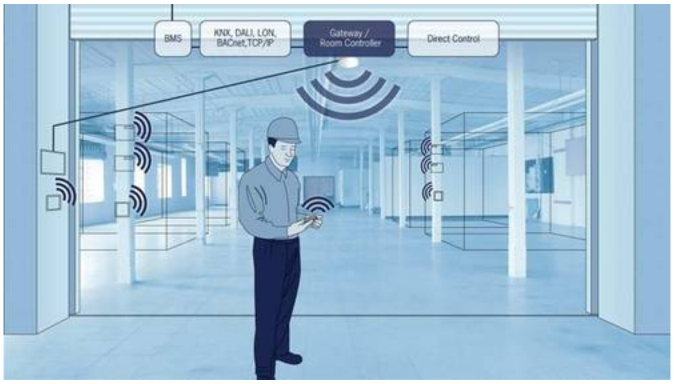

Energy is everywhere within reach; it just needs to be harvested. This is the principle of energy harvesting. Today, energy harvesting wireless solutions are already well established in the commercial building automation sector. But the technology is just getting started. New application fields for batteryless, wireless communication will be found to further enhance the world around us.

Based on energy harvesting wireless technology, a wide range of energy-autonomous applications are currently available for connected buildings that use motion, light, or temperature differences as their energy source such as batteryless switches; intelligent window handles; temperature, moisture, and light sensors; presence detectors, heating valves; and smart home systems. However, building automation is by no means where energy harvesting wireless ends.

Multiple interconnections

Everybody's talking about the Internet of Things (IoT). But how should billions of communicating devices be powered? The answer is by energy harvesting, and the reason is simple: Liberating sensors from external power, making them energy-autonomous, opens up unlimited processing and monitoring applications where cables or batteries represent an insurmountable hurdle. These features make energy harvesting wireless technology the ideal communication standard to easily and reliably interconnect thousands of individual devices in a system, as well as network them with other wireless protocols.

at AireTechnologies, Inc
HVAC Human Fire Life Safety Devices
Delivered As Early As Tomorrow!
Learn More
JAN 22-24, 2018 AHR EXPO CHICAGO
Visit us at the AHR Expo Booth 5341

Today, energy harvesting wireless technology is widespread, providing M2M solutions in the building automation sector and bridging the control of light, HVAC, and other fields of building technology to smart buildings, smart metering, and energy management systems. This is the starting point to actuate further applications that lead to the IoT in the long term. The following four categories show what this could look like.

Monitoring and control

Wireless and batteryless technology significantly eases energy monitoring and control in buildings with little intervention into the existing systems. The wireless devices are highly flexible to install so that individual components, wall switches, sensors, and relay receivers can be easily networked to form an intelligent system without complex cabling. In addition, dispensing with batteries eliminates the burden of maintaining the devices' energy supply in a regular time period, which can be up to each year.

An example for such a flexible automation system is HVAC control. Here, a thermostat, VAV (variable air volume), or fan coil controller receives information related to occupancy, temperature, humidity, window position, or CO_2 from the respective batteryless sensors and controls the opening and closing of valve actuators for radiators, or dampers for VAV systems. At the same time, the controller sends status information to a central building automation system, and receives control messages from the BAS. This enables the building to be monitored from a central location that can be remote from the building itself, and building-wide settings, such as holiday shutdown, to be implemented. Enormous progress is also being made on the product side, leveraging advancements in energy harvesting. Self-powered radiator valves generate energy from the difference in temperature between the hot water and the surrounding air. This energy powers the communication with a controller or BAS system, and turns the valve itself. Without cables or batteries, these wireless devices are especially easy to install, and they require no maintenance.

In further optimized systems, central equipment such as boilers or air handling units are integrated into the wireless communication system, enabling scalable HVAC generation, visible and controllable over the Internet on a PC, tablet, or

PE
Searching for a product?

smartphone.

Performing tasks

16:56
Auto
enocean

Alarm systems are a second field that batteryless wireless technology is opening up, due to its specific features. Here, the reliability requirements are much more stringent than those required for lighting controls. A system failure not only means a malfunction but also can cause much more serious consequences for other systems that depend upon the equipment being monitored. It's a fact that more malfunctions are caused by battery failures than by the electronics, especially in large systems. Energy harvesting overcomes this issue.

There are already various batteryless wireless water detectors available that use miniaturized solar cells or motion energy converters to power wireless signals that report water leaks in areas such as water supply networks in spacious industrial facilities. In the AFRISO universal module, for example, the EnOcean wireless signal immediately sends the leakage information to a gateway controller or directly to a valve, causing the main water pipeline or the affected supply line to be shut off. A notification is sent to the user's smartphone or smartpad at the same time to inform the user about the incident. In addition, the water valve can be opened and closed, independent of leakage notifications, by GSM connection via smartphone or smartpad.

Embedded processing

A major requirement of today's and the future energy supply is the Smart Grid. It's intended to network centralized and decentralized energy suppliers, including private homes producing electricity by photovoltaic installations, to an intelligent system that provides energy only when needed, updating in real time. This requires continuous data flow and processing from all involved parties, which means from millions of information points.

A key component is smart metering systems. To work reliably and cost-efficiently, interoperability between the meters is supplied by different manufacturers; this is why smart metering calls for standardized technologies. Consequently, the members of the EnOcean Alliance have defined a specific device communication protocol, the Automated Meter Reading (AMR) profile for batteryless wireless devices. Smart meter systems based on this open protocol are already available from a number of manufacturers. For example, Eltako meter components read and transmit the current electricity, water, and gas consumption, including accumulated meter figures, by means of energy harvesting wireless technology located at a variety of points inside a building. BSC software monitors and displays the current meter readings and compares them against default values. This makes all relevant data available for systems processing for intelligent energy management on demand.

Bridge to the cloud

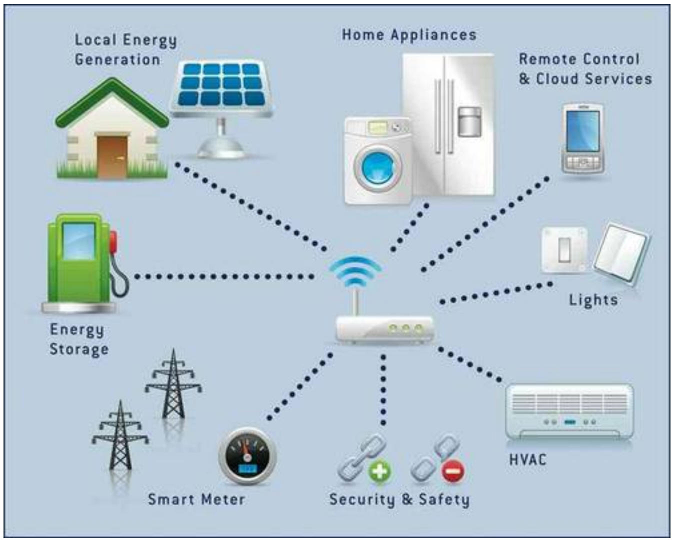

Via similar gateways, the standard-based energy harvesting technology can also communicate with Ethernet, Wi-Fi, GSM/UMTS/CDMA, and other networks for integration in cloud services. Here, all data collected by batteryless wireless sensors is encrypted and transmitted to a cloud service over the Internet. The gateways connected to a control and visualization software by TCP/IP that can be used to control all relay receivers and sensors bidirectionally. Some manufacturers have developed a cloud solution that offers energy management as a service. Therefore, facility managers, building owners, and businesses can monitor important inventory, equipment, assets, and energy-related information from anywhere at any time, via the cloud. Critical building-related data is automatically pushed to the cloud, freeing owners and managers from the often-challenging coordination and expense of hosting on-site servers.

One of the major advantages of such a cloud-based solution is that the management system arrives completely precommissioned from the manufacturer and ongoing device commissioning is expertly done on behalf of the client and pushed out from the cloud. The users are granted unlimited access to their remote, dedicated virtual server with their own IP address, accessible from a desktop or smartphone — the perfect precondition for a deeply connected world of an IoT.

The energy harvesting market is growing and multiplies on a year-by-year basis. Forecasts show that this trend will continue, especially as the next generation of energy harvesting wireless solutions is just around the corner to pave the way to the Internet of Things.

Laurent Giai-Miniet is CEO of energy harvesting wireless solutions provider EnOcean and previously spent 20 years with Texas Instruments (TI), where he held several management positions including General Manager for Low Power RF Products (LPRF).

Related News:

- Energy harvesting for Internet of things - 02.07.2014 01:48
- Consumer market will drive industrial Internet of things for manufacturing - 06.03.2014 11:43
- How the Internet of Things will affect consulting engineers - 02.01.2014 11:35
- IP-all-the-way is the way to go - 23.07.2013 01:28
- Integration: Lighting and HVAC systems - 16.04.2013 01:30

More than 80 photos

The Basic Soldering Guide

Learn to solder successfully!

Alan Winstanley

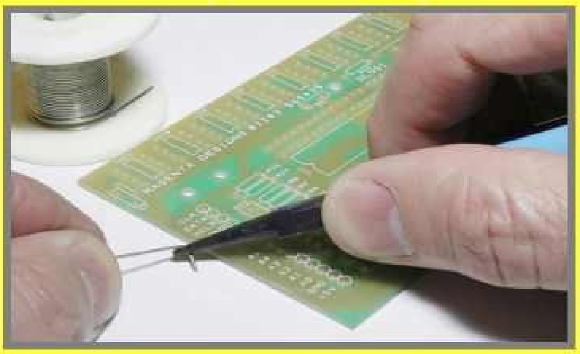
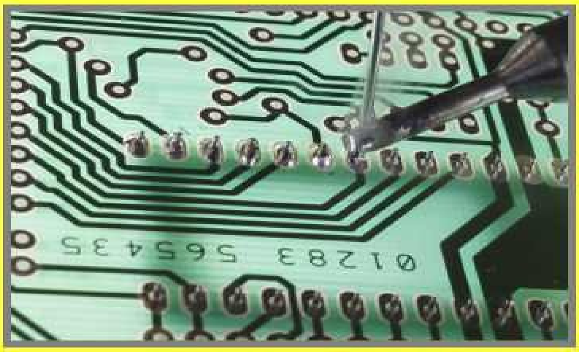
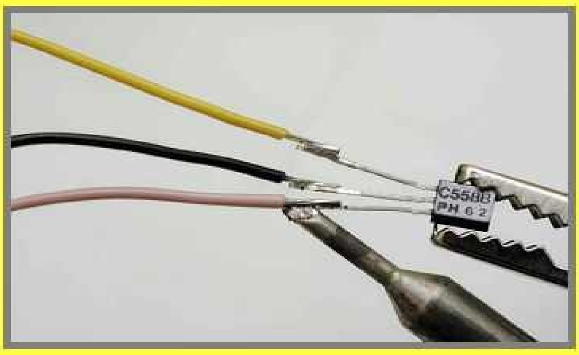
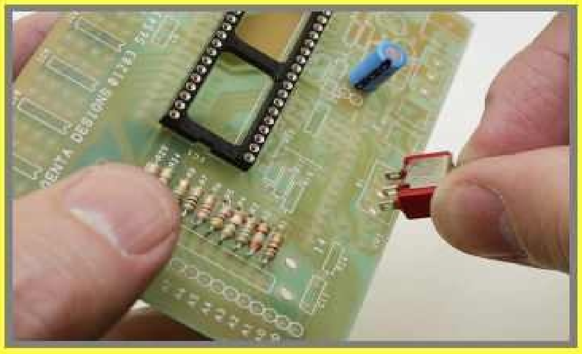
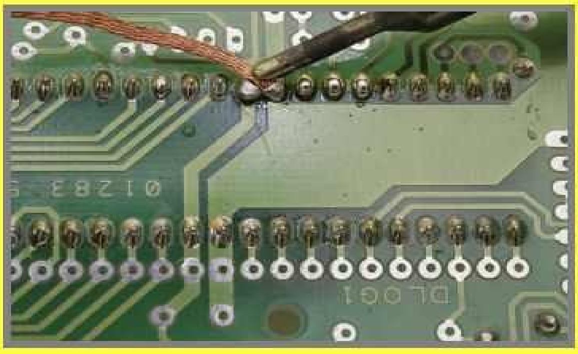

In association with

ANTEX

ANTEX (Electronics) Limited

The Basic Soldering Guide

by Alan Winstanley

***EPE Magazine* Online Editor**

In association with

Antex (Electronics) Ltd.

ANTEX

ANTEX (Electronics) Limited

Copyright Notice and Disclaimer

All text, diagrams and photographs are Copyright © 1996-2013 A. R. Winstanley. All Rights Reserved.

All text, diagrams and photographs are fully protected by copyright and may not be reproduced in any commercial or non-commercial publication or medium, mirrored on any commercial or non-commercial website nor may they be appended to or amended, monetised or used or re-distributed for any commercial reason, without the prior permission of the writer.

Every care has been taken to ensure that the information and guidance given is accurate and reliable, but since conditions of use are beyond the writer's control no legal liability or consequential claims will be accepted for any errors contained herein.

Where mentioned, the U.K. mains voltage supply is 230V a.c. and you should amend ratings for your local conditions.

Preface

In 1996 when the world wide web was still very young, I launched the first and most detailed website ever describing the practical skills of electronic soldering, and my **Basic Soldering Guide** quickly became the #1 web site of its kind in the Google search engine.

Thanks to its in-depth reference text and the unequalled high-quality close-up photography showing soldering step by step, many quickly learned the essential stages needed to make a solder joint successfully. Even novices who had never tried soldering before, gained the skills and confidence needed to acquire this skill.

My Basic Soldering Guide became a key *go-to* online guide for soldering, and I've enjoyed receiving encouraging feedback ever since from the likes of the US Air Force, US Marines, US Coastguard, Honeywell trainers, Atomic Energy authorities, Australian aeronautical suppliers, UK colleges and universities, trainees and many more around the world.

In association with Antex (Electronics) Ltd., the leading UK manufacturer of electronic soldering equipment, I'm delighted to bring you this updated Basic Soldering Guide containing over 80 all-new colour photographs, more background, more detailed information and lots more practical hints and tips.

I've revisited various areas of the guide and refreshed them, taking onboard readers' queries and nearly 17 years of online feedback, experiences and answering reader's questions. With the help of all-new photography, I'm sure you'll master the skills needed to solder electronics successfully using this updated guide.

I welcome feedback and comments, and readers can reach me via my website **www.alanwinstanley.com** or Email me at **alan@epemag.demon.co.uk**

Alan Winstanley

July 2013

Table of Contents

Copyright Notice and Disclaimer

Preface

Introduction

First steps

Soldering iron choices

How to solder

Next steps

Cleanliness and “Tinning” the bit

Getting Ready

Solder and Fluxes

Flux dispensers and Colophony

Temperature flow

Now is the time…

Soldering Step by Step

Interwiring – get hooked up!

Wire joints

Tidying up

Reflow technique

Fatigue & Breakage

Faults & Desoldering techniques

Desolder Braid

Quick Summary Guide

Troubleshooting Guide

Possible Hazards and simple First Aid

Useful resources

Acknowledgements

Photography

Conclusion

Introduction

The first and most important aspect of assembling any electronic project is that of *soldering,* which is a delicate and precise skill that can be mastered with experience. Sometimes called “soft soldering”, there’s no shortcut to acquiring the necessary expertise, and producing a consistently satisfactory solder joint takes a little practice. However, like riding a bicycle, soldering is an art which once learnt is never forgotten, and the purpose of this new and updated guide is to explain the techniques of soldering and desoldering for beginners, which I hope will set the hobbyist or trainee technician firmly on the road to successful electronic assembly or repairs in the future.

Soldering is the least “aggressive” way of joining non-ferrous metals together, and is used universally in electronics, air conditioning and refrigeration circuits, household plumbing and more besides – applications where the precise joining together of components at fairly moderate temperatures is needed. Further up the scale, **brazing** involves using higher temperatures to melt brazing rods onto larger metal parts, perhaps to repair a metal chair, lawnmower or to fabricate metal components or jewellery into intricate shapes. Lastly, **welding** is a very aggressive way of fabrication using welding rods or wire; steel girders, oil rigs and ships are all welded together, or robotic spot-welding is used for the mass production of, say, washing machines or car bodyshells using sheet steel to make strong rigid assemblies.

Due to the lower temperatures used and the need to make consistently good electrically conductive and mechanically sound joints with precision, soldering is used to connect components together when manufacturing electronic circuits. Small components would quickly be destroyed by brazing or welding, although tiny spot-welding joints do appear in electronics, perhaps to weld a metal tag onto a button battery.

This guide therefore deals with the soldering techniques used in electronics at hobbyist or trainee educational level. It explains what to look for before buying a soldering iron, describes ways of making various solder joints on circuit boards and other electronic components, and also how to *desolder* – removing solder in order to repair a circuit board or replace an electronic component.

You’ll also find more details of other aspects of soldering, including an outline of typical solder types and fluxes. In short, everything you need to get started in electronics soldering is here, so let’s get started!

First steps

The principle behind soldering sounds quite simple: the idea is to join components together to form an electrical connection, by using a mixture of lead and tin solder or alternatively “lead-free” solder (an alloy of tin and copper), which is melted onto the joint using a *soldering iron*. If you have never picked up a soldering iron before, then this guide will show you everything to help you start soldering with confidence. I also hope that the guidance will help those working in other areas — computer technicians or audio enthusiasts, for example, who may be faced with electronic repairs or modifications using a soldering iron for the first time.

If you’re an electronics hobbyist or trainee, before embarking on any form of ambitious electronic project, it is recommended that you practice your soldering technique on some *brand new* components using *clean* strip board (or protoboard) or a printed circuit board, and select a simple and straightforward constructional design as a starting point. Become acquainted and comfortable with your chosen soldering iron, which likely to become as familiar to you as a favourite pen. Learn how to balance it and handle it with precision. Try soldering an assortment of resistors, capacitors, diodes, transistors and integrated circuits with it, and then try your hand at *desoldering* – removing the solder again to make a repair or modification.

A really good place to start learning is by building a simple electronics kit, such as a Velleman kit that contains a good quality printed circuit board. You’ll learn some of the basic skills of successful soldering this way, and it’ll be a great confidence booster too.

Did you know? In the USA and elsewhere, the letter L in “solder” is silent and they say “soda” or “sodder” – but here in Britain we do pronounce the L and we say “sole-der”.

Soldering iron choices

Search any electronics catalogue or website and you'll see a bewildering array of soldering equipment on sale, including irons, controllers, work stations and desoldering equipment too. A large range of soldering irons is readily obtainable - which one is suitable for you depends on your budget and how serious your interest in electronics is, but there's something for every pocket distributed by a variety of retail, industrial and mail-order outlets.

The Antex range of soldering equipment has been very popular with industry, education and the electronics hobbyist for 60 years; I grew up with an Antex iron and a trusty 15W Antex iron was an everyday part of my hobby electronics all the way through the 1980's. An industrial user or a more dedicated hobbyist – with a bigger budget! – will be interested in a soldering station instead and again Antex offers a range of British-made products for industry or home use.

A very basic mains electric soldering iron can cost from under £5 (US$ 8), but I find that these very cheap irons, as sold on auction websites, are pretty crude and imprecise. They are best suited for simple electrical repairs and DIY rather than precision electronics or printed circuit boards discussed here. They tend to be bulky and uncomfortable for extended use, and they may not have suitable "bits" or tips of various sizes to suit different tasks.

^^ This classic Antex XS25 "pencil style" 25W mains-powered soldering iron has exchangeable tips or "bits" and is great for general hobbyist or educational use. Stands are also available to store them safely in between use.

A quality pencil-style electric soldering iron such as the Antex XS25 (photo) will be approximately £20 (US$18 tax free) - though it's possible to spend into three figures on a soldering iron “station” if you’re really serious about the subject! Don’t be tempted to over-spend on an elaborate workstation though, unless you are really very serious about becoming involved in electronics. You will usually obtain perfectly satisfactory results using a fairly modest “pencil” iron model, and you can upgrade to something more sophisticated should your needs change in the future.

^^ For enthusiasts or industry, this Antex 660TC soldering station has a separate mains-powered control unit and a matching low-voltage soldering iron rated at 24 Volts, 50 Watts so it’s suited to a wider range of tasks than a lower powered one.

When choosing your soldering iron, certain factors to bear in mind include:

Voltage: for the British market, “mains” electric irons run directly from the mains at 230V a.c. or will be set for other voltages (110V a.c.) depending on the country. However, **low voltage types** (e.g. 12V or 24V) usually form part of a “soldering station” for use with a matching controller made by the same manufacturer. Some low-voltage irons run off batteries (e.g. a car battery or Ni-cads) but these are uncommon.

Wattage: this is an extremely important factor to think about when choosing your iron. Typically, irons for general electronics work may have a power rating of between 15-25 watts or so, which is fine for most electronic assembly tasks, printed circuit boards and inter-wiring. It’s important to note that a higher wattage does *not* mean that the iron runs hotter - it simply means that there is more power “in reserve” for coping with larger joints. The maximum electric iron wattage generally available is about 100W, which is OK for DIY electrical repairs but is far too high for general electronics or circuit board use.

A higher wattage iron offers you more flexibility for tackling a wider range of tasks. It has a better “recovery rate” which makes it more “unstoppable” when it comes to heavier-duty work, because it won't be drained of its heat so quickly. So check the power ratings carefully, and anything between 15–40W is fine for general electronics soldering.

Temperature Control - the simplest and cheapest types don't have any form of temperature "regulation". Simply plug them into the mains and switch them on! Thermal regulation is "designed in" (by physics, not electronics!). Sometimes they are described as "thermally balanced" as they have some degree of temperature "matching" – in other words, they warm up as quickly as they lose heat during use, so in a primitive way they maintain roughly a constant temperature. This type of iron is perfectly acceptable for hobby or less demanding professional use. It's essential to use the manufacturer's specified tips (see later) to maintain proper temperature matching, otherwise the iron may not heat up enough – or it may overshoot in temperature.

These unregulated irons form an ideal general-purpose iron for most users, and they cope reasonably well with printed circuit board soldering and general interwiring. However, most of these "miniature" types of iron will be of little use when attempting to solder large joints (e.g. very large terminals or very thick copper wires) because the components being soldered will draw or "sink" heat away from the tip of the iron, cooling it down too much and preventing solder from flowing properly. That's when a higher wattage iron is needed.

A proper **temperature-controlled iron** will be quite a lot more expensive - retailing at say £40 (US$ 60) or more - and will have some form of built-in thermostatic control, to ensure that the temperature of the "bit" (the tip of the iron) is maintained at a fixed level within reasonable limits. This is desirable during frequent use, since it helps to ensure that the temperature will be relatively stable regardless of the workload. Some irons have a bimetallic strip thermostat built into the handle which gives an audible "click" in use, and some may include an adjustable screwdriver control within the handle as well. Others may have electronic controls built in.

^^ *An Antex 660TC soldering station with matching iron and bench stand.*

More expensive still, **soldering stations** cost from £70 (US$ 115) upwards (the iron may be sold separately, so you can pick the type you prefer), and consist of a complete bench-top control unit into which a special *low-voltage* soldering iron is plugged. Some versions might have a built-in digital temperature readout, and a control knob to vary the setting. The temperature could be boosted for soldering larger joints, for example, or for using higher melting-point solders (e.g. lead-free or silver solder). These are designed for the discerning user, or for continuous production line or professional use. A thermocouple will be built into the tip or shaft, which monitors temperature.

The best soldering stations have irons which are well balanced, with comfort-grip handles which remain cool all day, and silicone-based cables which are burn proof. Antex produces a range of irons with silicone cables specially for education use, to help avoid accidents caused by careless use by students.

Anti-static protection: if you need to solder a lot of static-sensitive parts (e.g. CMOS chips or MOSFET transistors), more advanced and expensive soldering iron stations use static-dissipative materials in their construction to avoid static charges accumulating on the iron itself, which could otherwise damage or “zap” some semiconductors. Such irons are listed as “ESD safe” (electro-static discharge proof). The cheapest irons are not classed as ESD-safe but they still perform well enough in most hobby or educational applications provided you take the usual anti-static precautions when handling these components. The iron would need to be well earthed (grounded) in these circumstances, to carry away any static.

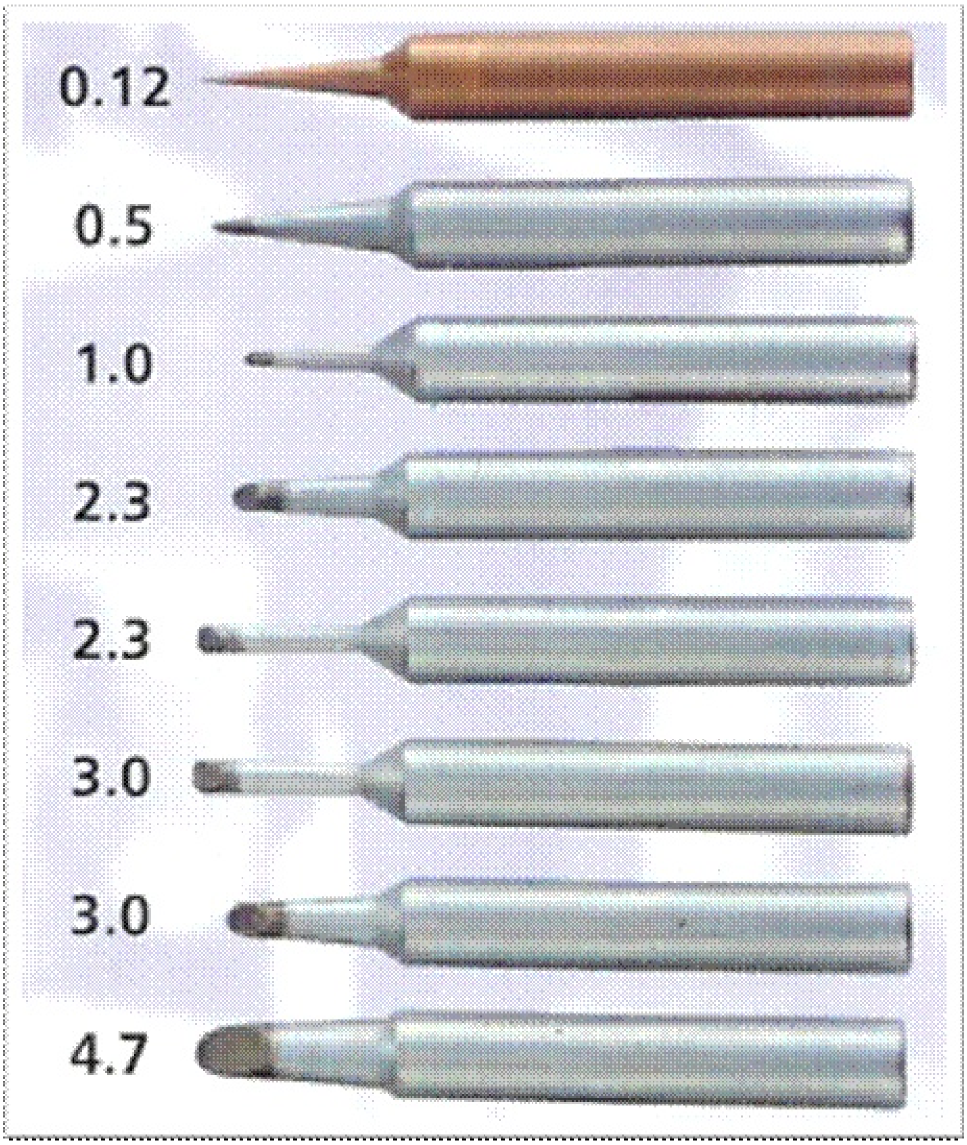

^^ Just part of the range of spare tips or "bits" produced by Antex for their soldering irons. Use only tips intended for a particular iron, to avoid thermal matching problems.

Tips or Bits: it's often useful to have a small selection of manufacturer's bits (soldering iron tips) available with different diameters or shapes, which can be changed depending on the type of work in hand. You will probably find that you become accustomed to, and work best with, one particular shape of tip for the majority of your work. Usually, tips are iron-coated or plated to preserve their life and to maintain good tip "hygiene". Be sure only to use tips that are specifically designed for your iron, otherwise thermal problems may arise. I show separately some typical tips, courtesy of Antex.

^^ Useful accessories from Antex including solder, tips and tip cleaner, desoldering braid and a heatshunt.

Spare parts: it's always reassuring to know that spares are available in the future if required, so if the element blows, you don't need to replace the entire iron. This is especially the case with expensive irons. Check some websites or mail-order catalogues to see whether spare parts are listed. One drawback is that you may need another soldering iron when exchanging a broken heating element!

ANTEX Gascat 40

^^ Gascat 40 gas-powered soldering iron kit by Antex, with spare tips and cleaning sponge.

Gas or electric? So far I've discussed electric soldering irons, but **gas-powered** soldering irons are sold which use butane propellant rather than mains electricity to operate. They have a flint lighter or (better still) a built-in piezo for ignition, and have a catalytic element which, once warmed up, continues to glow hot when gas passes over them. They tend to be big and bulky compared to electric pencil irons.

Field service engineers use gas-powered irons for working on repairs where there may be no power available, or where a joint is tricky to reach with a normal electric iron, so they are really for occasional "on the spot" use, rather than for mainstream construction or assembly work. I use one when I just need to do a quick repair and can't be bothered getting the electric soldering iron going!

Gas irons can have higher power equivalents than electric ones (eg 125 watts or more) but some gas-powered irons are nothing more than miniaturised blowtorches, which may or may not be useful for occasional heavier duty soldering. In the author's experience they can be difficult to use in confined areas. *Extreme care is needed at all times to ensure hot gas emitted from the iron's exhaust port doesn't damage nearby materials, plastics or wiring.* Gas irons can have useful accessories to convert them into e.g. a hot knife for sealing nylon rope, or a hot-air blower for use with heatshrink tubing. Almost every electronics constructor uses an electric-powered iron though.

A **solder gun** is a pistol-shaped iron, typically running at 100W or more, and is completely unsuitable for soldering modern electronic components: they're too hot, heavy and unwieldy for micro-electronics use, nor are they designed for that. Plumbing or DIY, maybe..!

^^ *A heat resistant soldering iron stand with cleaning sponge. (Antex)*

Soldering irons are best used along with a **heat-resistant bench stand**, where the hot iron can be safely stored in between use (photo). It is extremely important that a hot soldering iron is safely “parked” ready for action, and a bench stand is really a necessity. Soldering stations usually have such a feature, otherwise a separate soldering iron stand is essential, ideally one that’s supplied with a tip-cleaning sponge. You can make your own cleaning sponges using *cellulose* sponge only.

^^ *A benchtop fume extractor fan for hobbyist use. A replaceable carbon filter helps remove particles and air is vented out the back.*

Other equipment worth considering includes the use of **fume extractors**, which are compulsory in the industrial workplace. Solder fumes and flux smoke are not known to be toxic but they can cause irritation. A basic fume extractor (photo) consists of a small bench-top fan which draw fumes and irritating smoke away from the operator’s face and filters out some of the smoke particles, before exhausting the air back out through the fan. The carbon-impregnated foam filters are replaceable. Such devices are quite effective and users soon find them indispensable, even though they can be a bit noisy at close range.

Professional fume extraction systems draw the smoke and fumes directly from the work area via a clip-on tube fitted to the soldering iron, then vent the fumes away through a large filter pump. It is definitely worth considering a small bench top unit for regular hobby or occasional professional use, as having decent ventilation can only be a good thing.

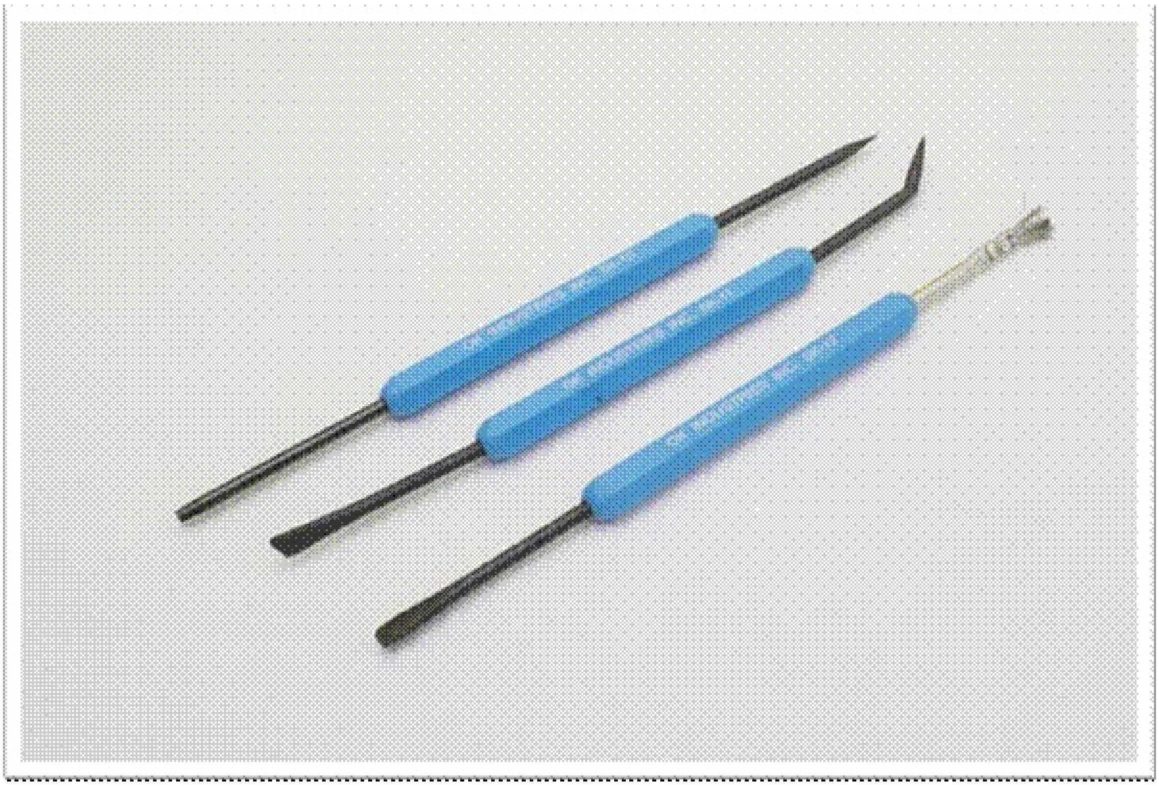

^^ *A soldering accessory toolkit by OK Industries, including probes, scrapers and wire brush*

A variety of specialist **hand tools** are available that assist with soldering, and a good supplier's catalogue will offer a range of small brushes, scrapers and cleaning tools in a handy kit, together with the usual types of wire cutters, pliers and so forth, which are necessary for handling components and tidying up as required. Some specialist service aids (aerosols etc.) are also described later.

Now let's look at how to use soldering irons properly, and later on we will describe the techniques for putting things right when a joint somehow goes wrong — and don't worry, even the experts get it wrong sometimes!

How to solder

This guide will show you step by step how to solder successfully and plenty of photographs are provided to help explain the techniques needed.As you read through the guide I'll explain all the stages in more detail, but let's look at the basics first.

First of all, successful soldering requires that the items being soldered together are held with as little movement as possible. So it's best to *secure the work* as needed, so that your accuracy isn't affected should the workpiece move accidentally.

In the case of a printed circuit board, various holding frames are useful especially when densely populated boards are being soldered: the idea is to insert all the parts on one side (a process called "stuffing the board"), hold them in place with a suitable pad to prevent them falling out again, turn the board over and then snip off the wires with cutters before soldering the joints. The frame saves an awful lot of turning the board over and back again, especially with large boards: all the soldering can be performed in one pass.

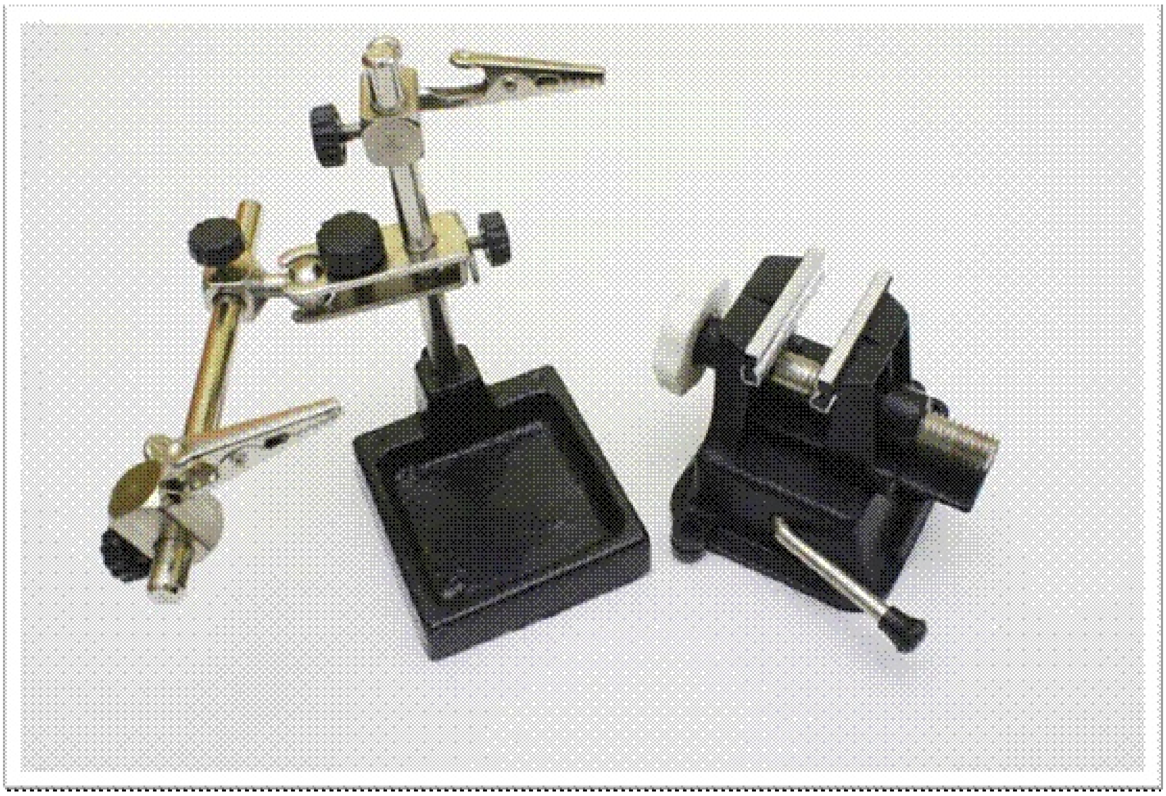

^^ The ever-popular "Helping Hands" (left) helps support sundry parts, wires etc. during soldering.
A modeller's vice (right) holds parts firmly. A vacuum base fixes it onto smooth surfaces.

Only the more serious constructor will purchase a holding frame, and hobbyists can retain parts in place by improvising in a variety of ways – the ever popular "Helping Hands" stands cost a few pounds and is widely sold. They have adjustable crocodile clips to grip parts, and maybe a magnifying glass or soldering iron stand as well. The cast iron base provides stability. Other parts could be held firm in a modeller's small vice, for example.

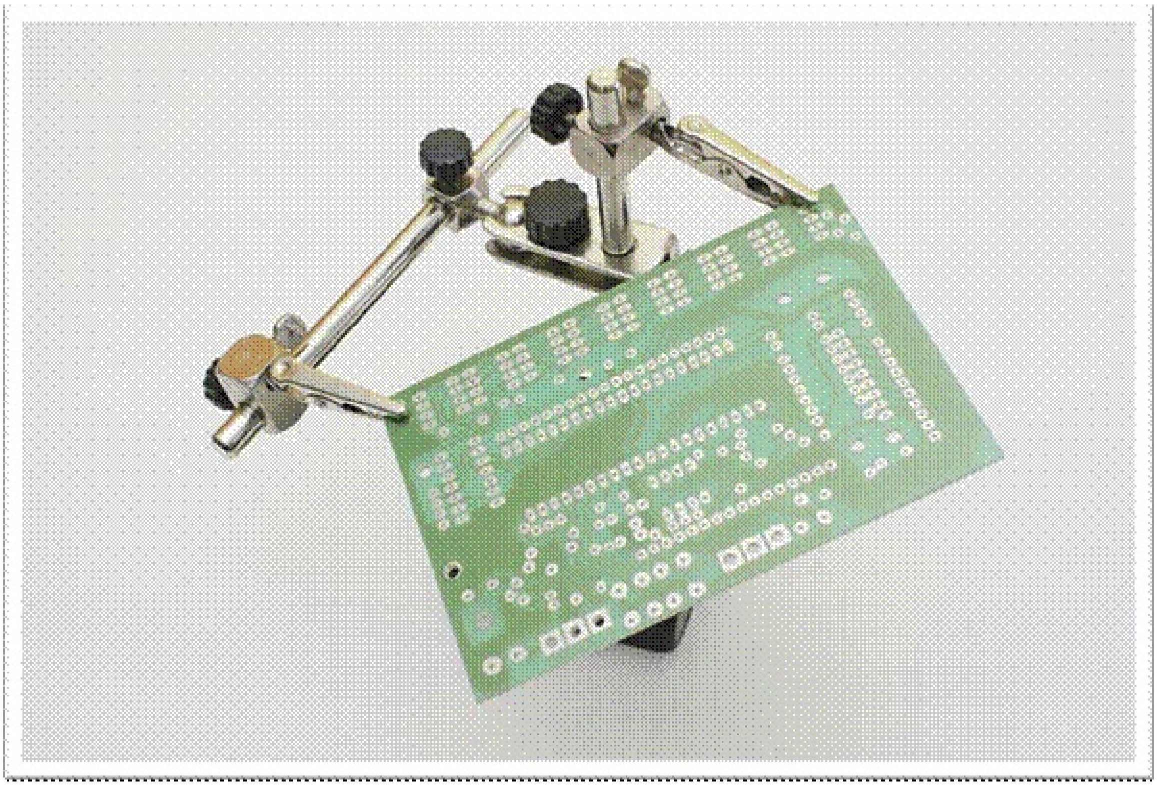

^^ *"Helping Hands" uses crocodile clips to grip parts during soldering. Or just place parts flat on the bench.*

When soldering parts onto an ordinary circuit board, components' wires can simply be bent to the correct pitch (distance apart) to fit through the board, insert the part flush down against the board's surface, splay the wires a little so that the component grips the board under spring tension, and then solder it. This technique is universally used in manual **through-hole soldering,** which is explained in full later.

In the author's view - opinions vary – it's generally better to snip off the surplus wires leads first, to make the joint and neighbouring joints more accessible and also to reduce the mechanical shock transmitted to the p.c.b. copper foil. However, in the case of diodes and transistors the author tends to leave the snipping until after the joint has been made, since the excess wire will help to "sink" heat away from the heat-sensitive semiconductor junction.

A special clip-on heatsink is available which also helps stop excess heat from reaching temperature-sensitive semiconductors like these. I've always managed without one but beginners might find them reassuring. Integrated circuits can either be soldered directly into place if you are confident enough, or better, use a dual-in-line socket to prevent heat damage. The chip can then be swapped out at a later date if needed.

Parts which become hot in operation (e.g. some power resistors), should be raised above the board slightly to allow air to circulate. Some components, especially large electrolytic capacitors, may require a mounting clip to be screwed down to the board first, otherwise the part may eventually break off due to vibration. It's a good idea to bolt such components firmly into place before soldering their terminals, in order not to strain the soldered joints or the components when fasteners are tightened.

By securing the workpiece as much as possible to prevent movement, you have a much higher chance of producing good-quality reliable solder joints.

Next steps

Let’s get to grips with actual soldering techniques in more detail. The soldering of electronics components utilises **lead/ tin** or **lead-free solder** and the process is compatible with many non-ferrous metals. You can solder copper, lead, brass, gold plate, silver, nickel, tin and tin plate, zinc and more besides but some metals such as nichrome, galvanised or stainless steel require a highly specialist “flux” (see later) to solder them and aren’t discussed here. Some materials such as beryllium, chromium, magnesium and titanium are non-solderable in any case, according to solder manufacturers Multicore.

In electronics we’re mainly concerned with soldering parts or wires onto printed circuit boards or terminals that are usually already “tinned” with solder or plated, ready for soldering with flux-cored solder. The key factors affecting the quality of a solder joint are:

- **Cleanliness** – dirt or impurities drastically hinder good solder coverage.
- **Temperature** – the right level to enable the solder to flow freely!
- **Time** – apply heat for just the right amount of time!
- **Adequate solder coverage** – enough to form a good joint without touching neighbouring areas.

A little effort spent now in soldering the perfect joint may save you — or somebody else —a considerable amount of time in troubleshooting a defective joint in the future. Let’s discuss the basic principles outlined above in more depth.

Cleanliness and "Tinning" the bit

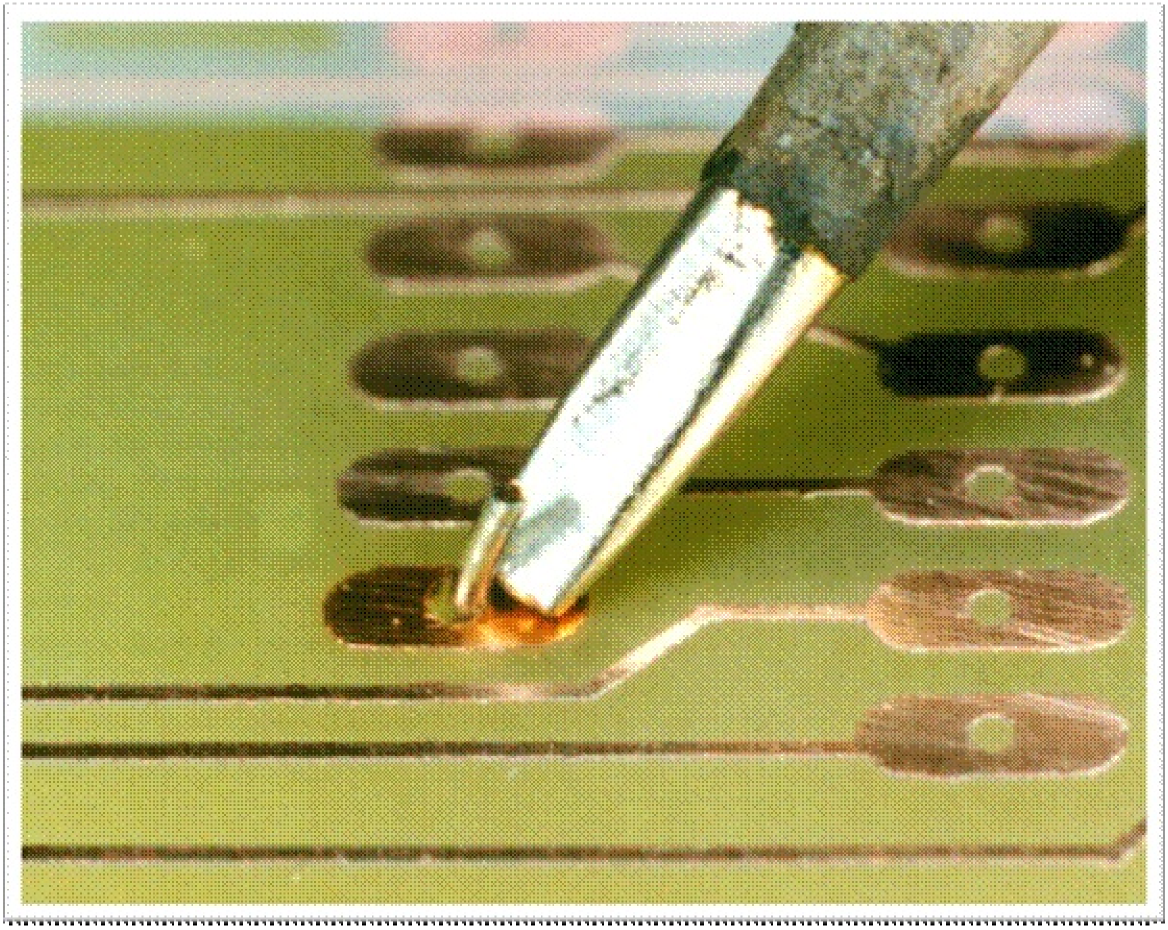

^^ A clean and shiny soldering tip or "bit" is essential for each and every solder joint you make.

Firstly, and without exception, all parts - including the soldering iron's tip itself - must be clean and free from contamination. Fact is, solder just will not "take" to dirty parts! Old components or copper board can be impossible to solder because of oxidation that builds up on the surface of the leads. This repels the molten solder and will soon become evident because the solder will "bead" into shiny globules looking like mercury, going everywhere except where you need it. *Dirt and contamination are the enemies of a good quality soldered joint!*

When all the conditions are right for soldering, materials are said to be "wettable" and will accept molten solder, which should flow readily over their surfaces. Hence, it's **really necessary** to ensure that parts are free from grease, oxidation and other contaminants. Note that in any case, some incompatible materials or surface finishes just cannot be soldered using ordinary lead-free solder, no matter how hard you try – e.g. aluminium parts would require special aluminium solder and fluxes (see later) to be used.

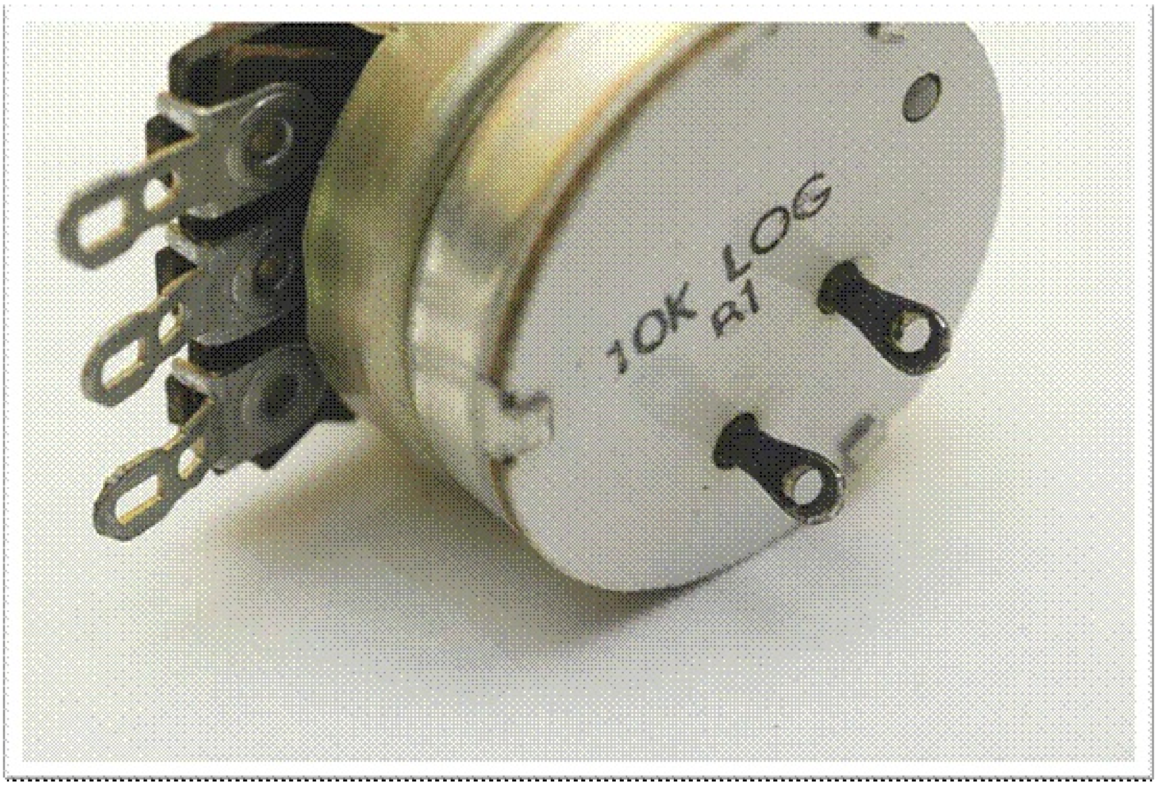

^^ The two tags on this switched potentiometer have blackened with age and oxidation. They must be cleaned before they can be soldered…

In the case of old components that have been stored a long time, where the leads have blackened and oxidised, use a small hand-held file or perhaps fine emery cloth to reveal fresh metal underneath. Stripboard and copper printed circuit board will generally oxidise after a few months, especially where it has tarnished due to fingerprints, so the copper strips could be cleaned using e.g. an abrasive rubber block made for the purpose. Also available is a fibre-glass filament brush (photo), which is used propelling-pencil-like to remove contamination. They're OK for general cleaning but best avoided on fine surfaces (e.g. gold plated switch contacts). These produce tiny particles which can irritate the skin, so avoid accidental contact with any debris.

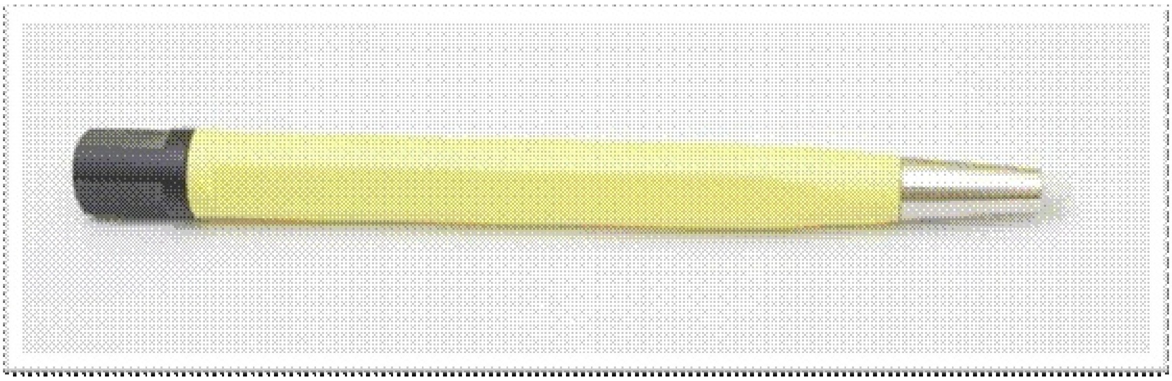

^^ A glass-fibre filament brush like this is useful for cleaning oxidised parts.

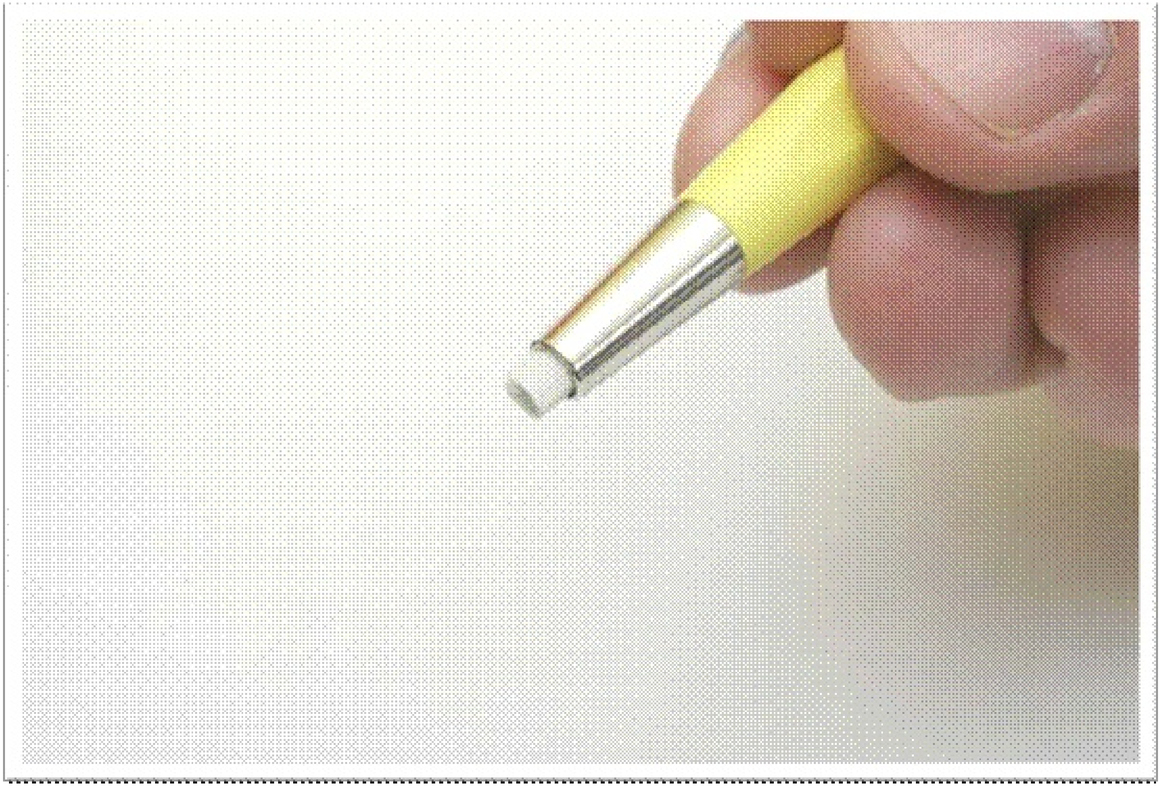

^^ The glass-fibre brush works like a propelling pencil and produces irritating dust.

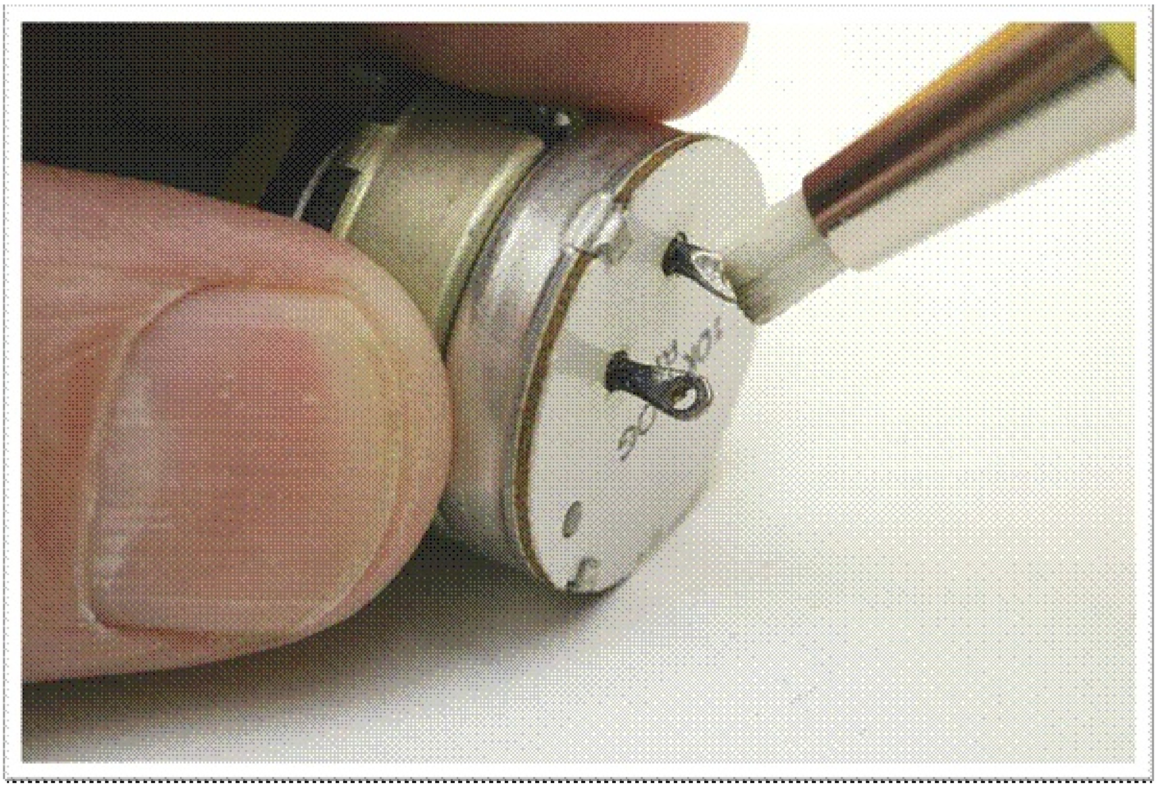

^^ Cleaning oxidised tags with a glass fibre brush to make them nice and shiny, ready for soldering

Afterwards, a wipe with a rag soaked in cleaning solvent will remove most grease marks and fingerprints. After preparing the surfaces, avoid touching any parts as far as possible.

Another side effect of trying to solder contaminated or incompatible materials is the tendency to apply more heat and "force the solder to take". As the materials aren't wettable the molten solder won’t flow where you want it to flow. This can do more harm than good, because it may be impossible to burn off any contaminants anyway, and the component or the printed circuit board may be overheated and damaged in the process. Semiconductors may be harmed by applying excessive heat for more than a few seconds, and extreme heat applied to printed circuit board tracks can cause irreparable damage, because the tracks will be lifted away from the substrate especially on a delicate or badly designed board. You can avoid trouble by ensuring the surfaces to be soldered are clean and wettable to begin with.

Getting Ready

As already explained, cleanliness in soldering is a major factor in obtaining successful results. A soldering iron stand usually has a sponge that is dampened – it wants to be quite damp but not running wet, so squeeze out any excess.

^^ Dampen the sponge but don’t soak it unduly. Note the hole, which gives an edge to wipe tip with.

^^ An Antex soldering iron tip. It simply slides onto the iron's heating element. Only the extreme end is wettable and will accept solder: the tip should be kept nice and shiny.

Soldering iron bits are typically iron-plated to resist wear, then chrome-plated to prevent molten solder being deposited on them, with the extreme tip being "wettable" to work with molten solder. Before using the iron to make a joint, the hot tip must be "tinned" with a few millimetres of solder*: you should always flood a brand new tip with plenty of solder to tin it immediately, when using it for the first time*.

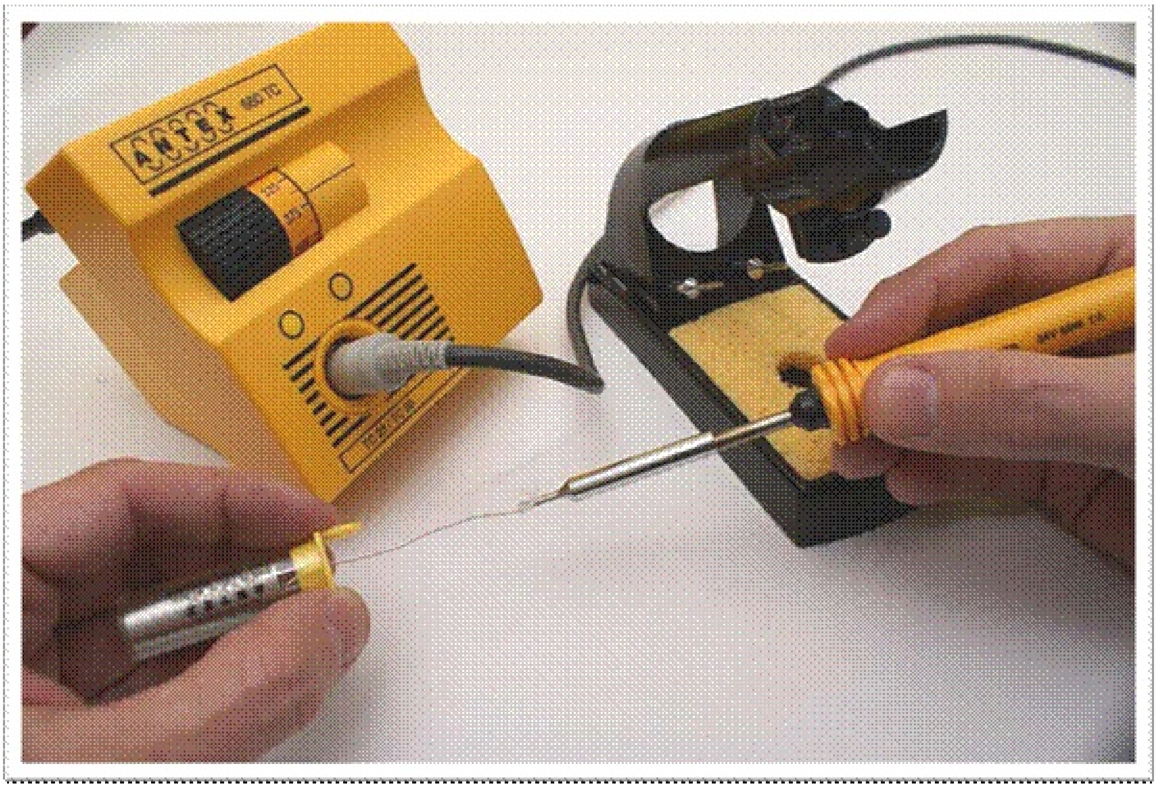

^^ Tinning the tip ready for use: apply a few millimetres of solder to the hot tip.

Wipe off excess solder using a damp sponge and it's ready to use. That's why sponges have a hole or well in them – the edge acts as a wiper and the hole catches excess solder.

A useful tip: after tinning the bit, just before using the iron it helps to re-apply a small amount of solder to a clean tip, to improve the thermal contact between the tip and the joint. The molten solder fills the void between the materials and the iron tip, to help transfer heat better so that the solder flows more quickly and easily.

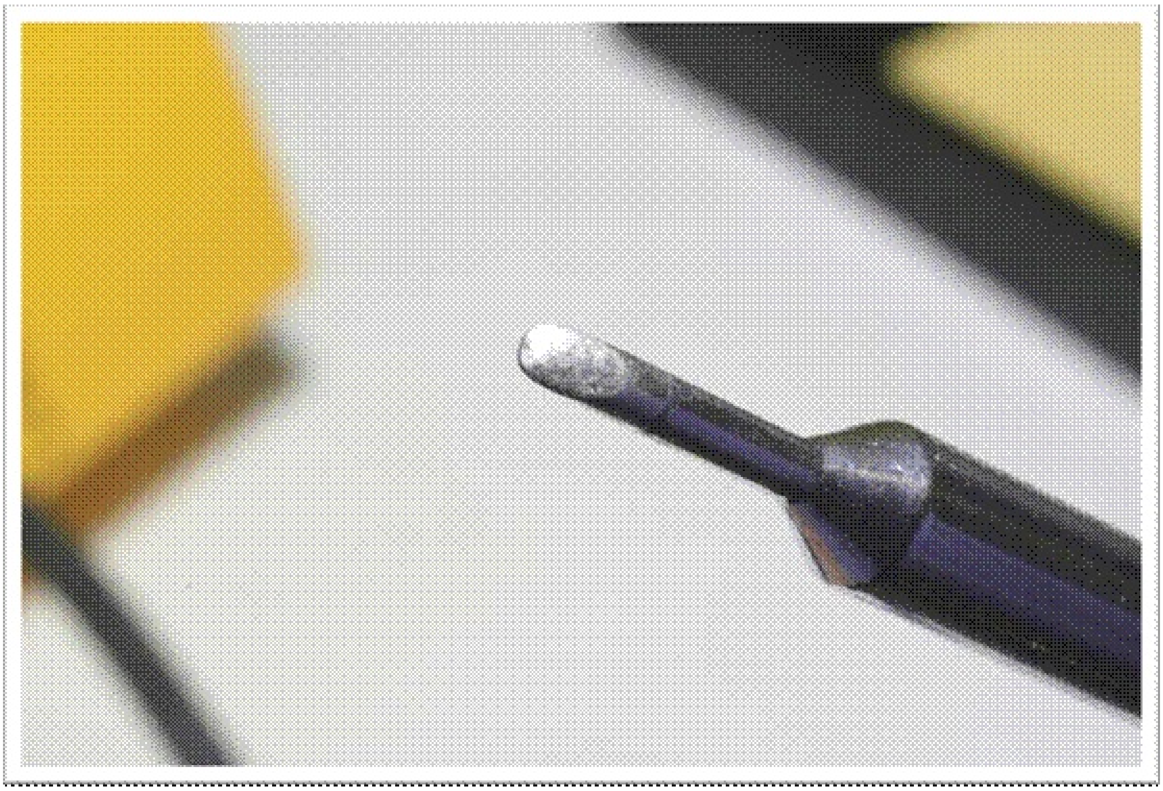

^^ *A tinned tip should look like this, ready to make the next joint.*

It’s sometimes better to tin larger parts as well before making the joint itself, but this is not generally necessary with p.c.b. work. Small tinlets of special paste are available onto which you dab a hot iron - the product cleans and tins the iron ready for use. It’s useful in extreme cases where the bit has some stubborn contamination.

I find special Tip Tinner &Cleaner products to be very useful: gently abrasive in action, they help to clean dirty bits and keep them in good condition. Use them for removing stubborn contaminants, but don’t overdo it.

Multicore
Tip Tinner/Cleaner
TTC1
15g / 0.5oz
Made in Malaysia
LBL 738

^^ Tip Tinner and Cleaner is a useful aid to maintaining a soldering iron bit …

^^ Just press the hot iron onto the solid paste and scrub it around a little. The tip will be cleaned, tinned, and made ready for use.

The move to lead-free solders (see next chapter) has had some effect on the life of soldering iron bits, with increased wear and corrosion noted due to the higher temperatures and the fluxes found in tin-based solders. You can therefore expect bits to wear out over time. Once the iron-plating is damaged due to oxidation or erosion, the bit is fast approaching its end of life. Never use an abrasive or file to sand down a tip: the iron-coating will be damaged and the iron's core exposed, so the tip will soon be made useless due to erosion.

Having prepared the soldering iron tip ready for use, in the next chapter solder and fluxes are discussed.

Solder and Fluxes

In recent years there's been a move towards using more environmentally-friendly materials in electronic products. EU legislation such as *Restriction of the Use of Certain Hazardous Substances in Electrical and Electronic Equipment* (RoHS) aims to reduce toxic heavy metals being sent to landfill. (Look for the RoHS symbol on equipment to indicate compliance.) Due to RoHS compliance, the electronics industry had to change the type of solder it uses in electronic production.

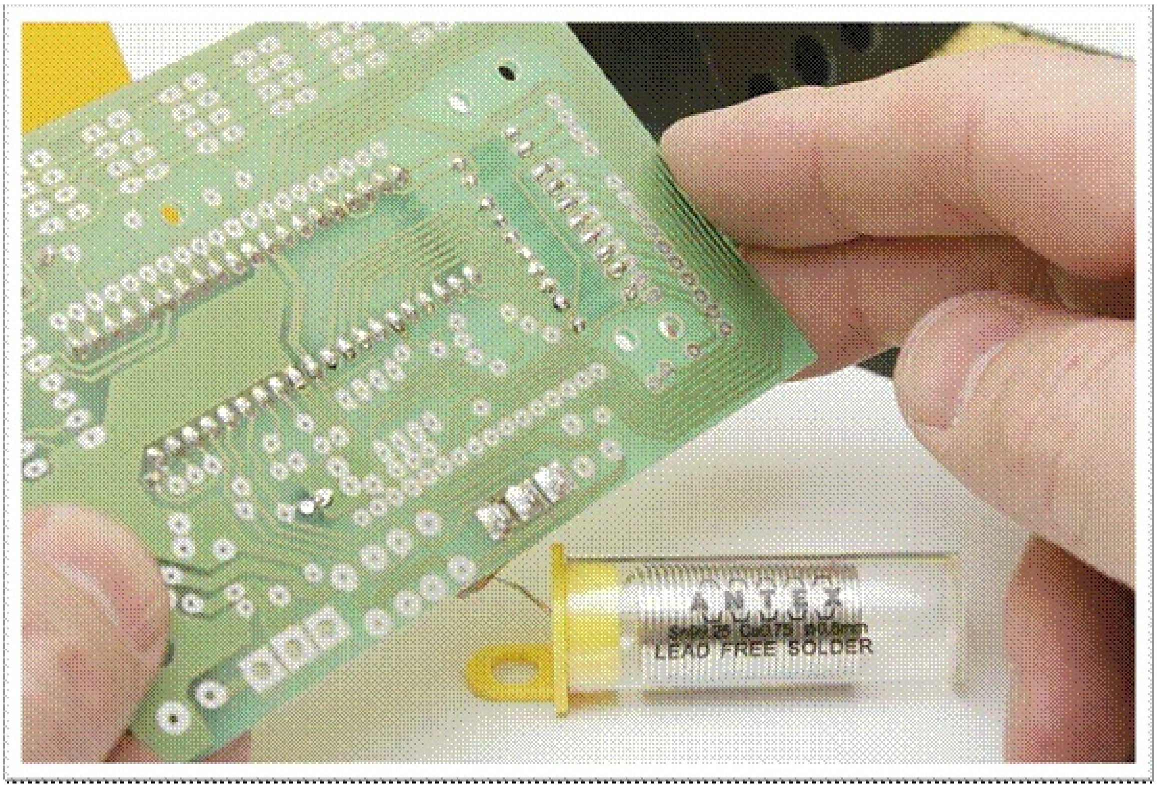

^^ *Electronics solder is supplied in reels or handy dispensers like this.*

Solder comes in various forms including solid bars or pellets for melting in small electric 'solder pots' used for treating the ends of wires with solder. Traditional general-purpose electronics-grade solder is in wire form – starting with so-called "60/40" which contains 60% tin (symbol Sn) and 40% lead (symbol Pb) and is sold in handy dispensers or reels. Although tin-lead solder is now banned in industry, there's nothing to stop the hobbyist from using it but best practice is to use lead-free solder in our work: my advice is to try both, and see which you prefer to work with. "40/60" tin-lead produces lower quality results but is slightly cheaper and perfectly acceptable in hobby circles.

Various diameters of solder wire are marketed. In the UK they're sold in Standard Width Gauge (SWG) sizes, typically as 18SWG (1.2mm) or 22SWG (0.7mm). The latter is fine for almost all hand-soldering of printed circuit boards or general electronics. For larger solder joints (e.g. larger switch or motor terminals), 18SWG solder would be better as more solder can be dispensed more quickly.

Lead-free solder is universally available and contains typically 99.7% pure tin and 0.3% copper (symbol Cu). It needs a higher melting point which makes it slightly more difficult to work with, but standard soldering irons will cope with it well. Antex lead-free solder (Sn 99.25 / Cu 0.75) is a good compromise at 0.8mm diameter and is sold in small dispensers.

Other solders are produced for specialist work, including aluminium solder (Alu-Sol®) and another solder variant used by professionals is Multicore “Smart” wire which contains a small amount of pure silver (symbol Ag). It produces very clean results and is often associated with SMD (surface mount devices), though some engineers also use it for routine p.c.b. work for producing the best possible finish by hand. As “Smart” wire contains lead it is not RoHS compliant.

An interesting variant is **Eutectic** solder, which is 63/37 Tin/Lead. It goes instantly from solid to liquid when melted and is particularly good for hand-soldering. An almost-equivalent lead-free product would be Stannol Flowtin TC or TSC solder.

The low-down on fluxes

When melting solder with a soldering iron, oxides of metal are produced as a result of the high temperatures involved. Unfortunately, these oxides contaminate the metal surfaces being soldered, which interferes with the flow of molten metal and the production of a good quality solder joint.

All electronics-grade solder wire therefore contains an additive called “flux” which helps the molten solder to flow more easily over the joint. It does this by scrubbing away the oxides which arise naturally during heating, and it will often be seen as a pungent brown fluid bubbling away on the joint, accompanied by some fumes.

Those coming into electronics from other industries should note that flux is already contained within “cored” electronics solder and on no account should any acidic flux be applied separately before using the soldering iron. Plumbers, for example, apply flux paste to copper pipes before soldering them, but electronics-grade solder wire already contains a flux and extra flux is almost never needed. Electronics is no place for acid fluxes!

^^ *A close-up of electronics-grade multi-cored solder. Five cores of rosin flux can be seen running through it.*

For almost all electronic hand assembly, solder wire containing “Rosin flux” is used. Cores of flux run through the solder wire like letters running through seaside candy, and they prevent the hot area from being contaminated by oxides, otherwise solder would never flow properly and the result would be an incomplete and unreliable joint.

Flux dispensers and Colophony

Flux dispenser pens are sold that allows special liquid flux to be applied separately onto a work area. These might be handy for difficult or challenging jobs to help solder to flow better: adding more flux this way won't do any harm and may help a solder joint to be made more quickly and reliably. In my hobby electronics, there's hardly ever been a time when I felt the need to apply extra flux but it's useful for some very tricky or demanding jobs.

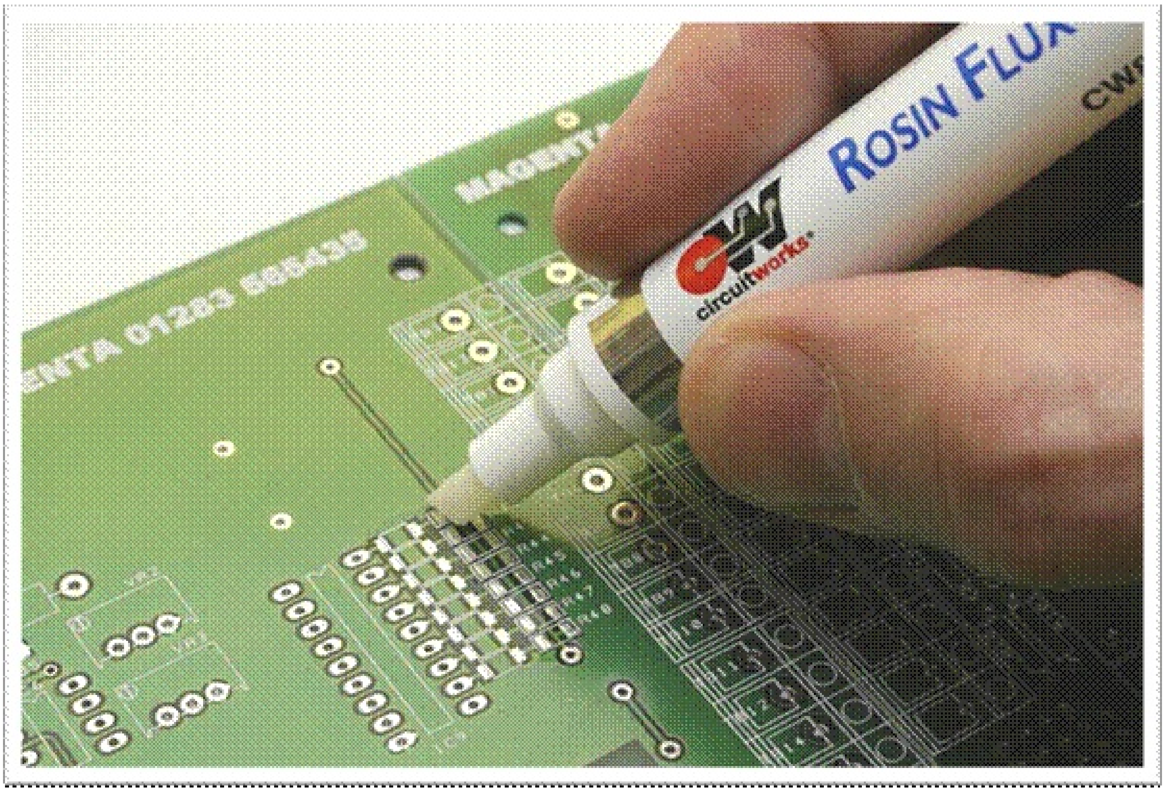

^^ *For more demanding work, a flux dispenser pen allows additional flux to be applied*

For example I've used specialist Chemtronics flux dispenser pens on tricky, extra-large solder joints involving very thick wires for lead-acid battery connections where I really struggled to make the solder flow properly. You might also use them in micro-electronic surface-mount work. The extra flux can only help, but for the rest of the time rosin-flux core solder wire contains sufficient flux and that's all that you'll need.

^^ *Colophony or rosin flux is available in small tinlets if you want to prepare your own liquid flux.*

One thing that many seasoned electronics enthusiasts will recognise is the distinctive smell of rosin flux: its intense woody pine smell is not unpleasant, and flux fumes themselves are not known to be harmful but solder smoke can be an irritant, especially if you suffer from asthma or other respiratory conditions. (I dealt with solder fume extraction earlier when soldering irons were outlined.)

Rosin flux is also known as **Colophony.** It's an amber resin that's glassy and brittle like sugar candy, distilled from the resins of conifers (mainly pine trees) and it's worth knowing about. For electronics use, Colophony is available in individual tinlets of solid resin (e.g. *Kolophonium* in 20g tins by Donau Elektronik GmbH, sold by Westfalia or Conrad).

As an aside, the resins of various trees are used in incense burners: placing a few fragments on a charcoal burner produces intense delicious fragrances of pine, frankincense or forest but you still wouldn't want to inhale the smoke or fumes directly.

Solids of Colophony can be dissolved to make a semi-liquid flux, for dipping or applying manually prior to soldering. To make your own rosin flux from solids of Colophony, chip off some fragments and crush them into a small tinfoil dish, then apply some isoproponal alcohol (start with 2 parts solvent to 1 part Colophony) and let it dissolve over about 20-30 minutes or more.

^^ Chip off some solids of colophony and put them in a small tin foil dish

^^ Add e.g. x 2 volume isopropanol alcohol and allow the solids to dissolve. Experiment as needed.

^^ *The resulting flux can be very sticky so handle with care and do not spill.*

In this form it can be applied directly with a brush or dropper to areas prior to soldering. In an open dish, after a day or two the solvent will mostly have evaporated to leave an extremely sticky resin that should be handled carefully: it's no co-incidence that Colophony is also used in adhesives and varnishes, and different grades of Colophony resins are used to treat violin or double-bass bows to add friction to bow hairs! Try experimenting with different ratios of solvent and making some up and storing it in an old nail varnish bottle with brush, or generally do what works for you.

Unused Colophony crystals can be reformed into their storage tinlet by warming carefully with a hot air gun, but don't overdo it. You can also use flux cleaners or PCB solvent cleaners if you want to remove traces of excess rosin flux after soldering: these are left behind as a brown deposit and are otherwise harmless.

The subject of using your soldering iron to raise the temperature of the materials is discussed next.

Temperature flow

This aspect takes a bit more understanding, but with practice you'll soon understand how temperature flows within a workpiece being soldered and how you have to harness it properly. After ensuring all parts are clean and the soldering iron is ready to go, the next step to successful soldering requires that the **temperature** of **all** the parts is raised to roughly the same level, before solder can be applied.

Imagine the most basic task of soldering a simple resistor onto a printed circuit board: the copper p.c.b. and the resistor lead should both be heated *together* so that the solder will flow readily over the joint. Later I'll show the precise stages step by step.

A beginner will often mistakenly just heat one part of the joint (e.g. a resistor's wire protruding through a printed circuit board) and hope that the resultant "blob" of solder will be enough to tack everything together. That's completely wrong, because the remaining metal in the joint is cold when molten solder is flooded on to it. The joint will be weak, incomplete or unreliable. Flux will not have flowed properly either, so the joint could be contaminated internally.

Another beginner's mistake is to use a soldering iron to carry blobs of molten solder over to the joint, as if to daub solder over it. The secret of success is to control the iron accurately and apply the hot tip onto the workpiece so that it's in contact with *all the parts*. Within a fraction of a second, heat will conduct from the iron and raise the temperature of the entire joint, after which solder can be melted over it. Remove the iron and let the joint cool naturally. It takes some practice with your chosen soldering iron tip to obtain the best heating action, making sure the tip is clean and tinned properly to begin with.

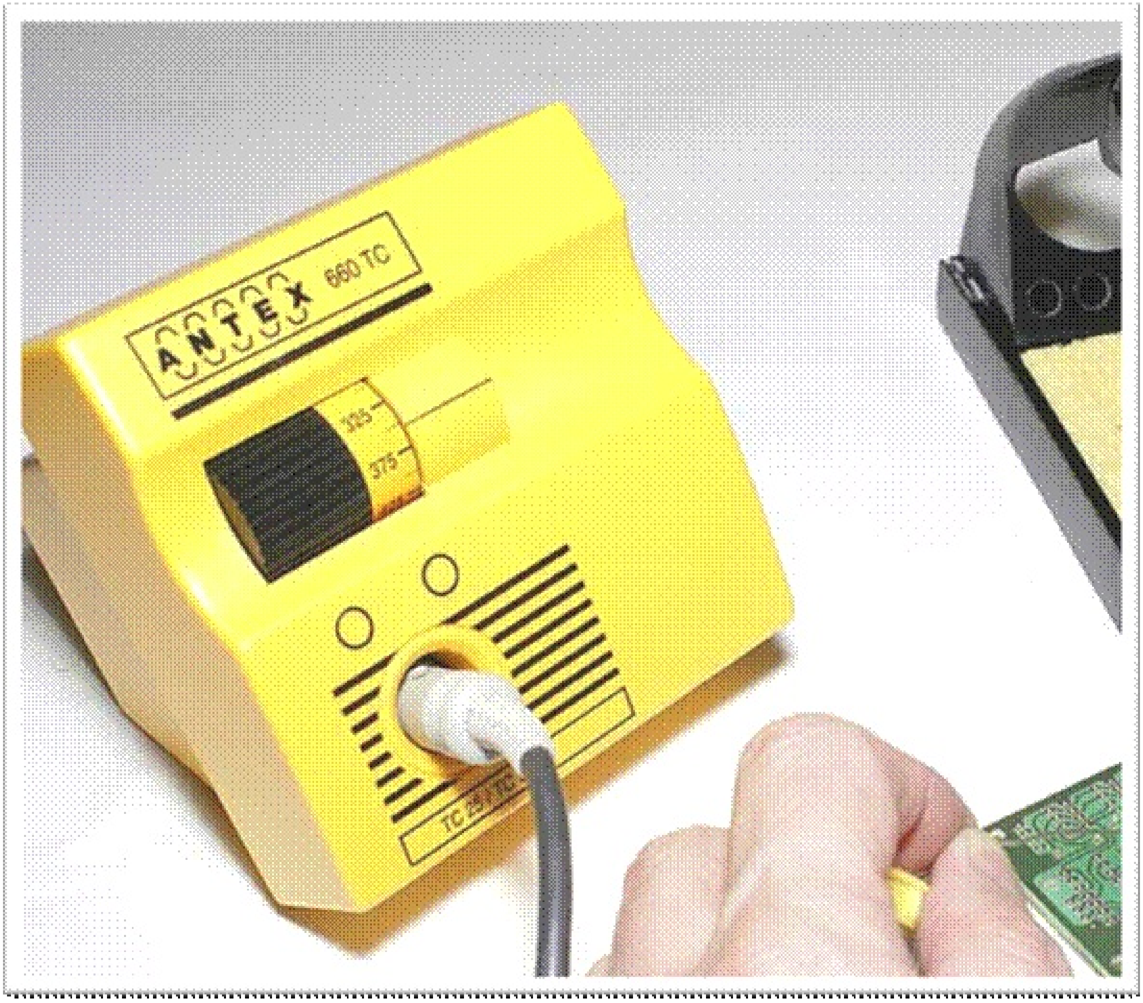

^^ *Set a typical temperature of 330-350 ºC (626-662 ºF) or more, if you have a soldering station.*

The melting point of traditional tin-lead solders is about 190ºC (374ºF) but lead-free tin-based solders require higher temperatures, having melting points of typically 201-227 ºC. As Antex reminds us, the melting point is *not* the temperature of the soldering iron tip: instead you should set a temperature that ensures the solder melts *instantly* onto the tip. Fixed-temperature soldering irons have no adjustability but they'll cope just fine with either type of solder. A soldering station usually has a variable control that gives more control in different circumstances. In practice the iron tip temperature should be set for typically 330-350 ºC (626-662 ºF) or maybe a little more if using lead-free solder.

Now is the time...

The next diagram shows what would happen if you applied a hot soldering iron to an imaginary metal block. In step (1) heat travels out of the iron into the cold metal block, which starts to warm up starting at the edges. Gradually (2), the whole metal block's temperature rises until the middle of the block finally rises in temperature as well. In effect heat is now travelling back "towards" the iron tip, until finally (3) the whole block is at the same temperature as the iron. At this point solder could now be applied.

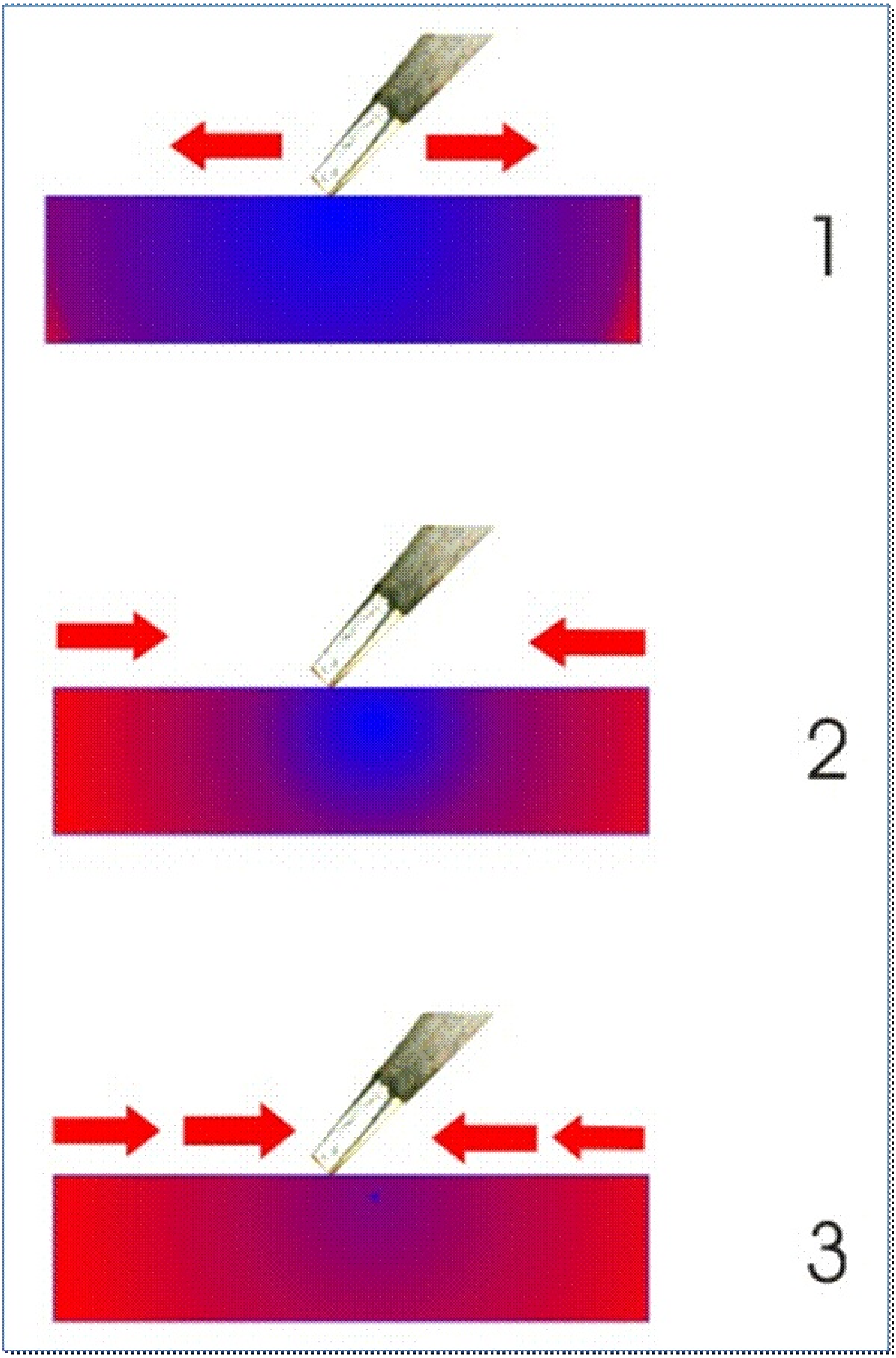

^^ How metal in the joint actually "sinks" heat away from the tip to begin with. Then heat moves back towards the tip until the solder's melting point is finally reached.

You'll notice this effect as a time delay when soldering any joint. The "working area" of the iron's tip gets **cooled** to begin with, because the metal in the rest of the joint is sinking heat away from where it's most wanted. Only after the whole joint has risen in temperature can solder be melted onto it.

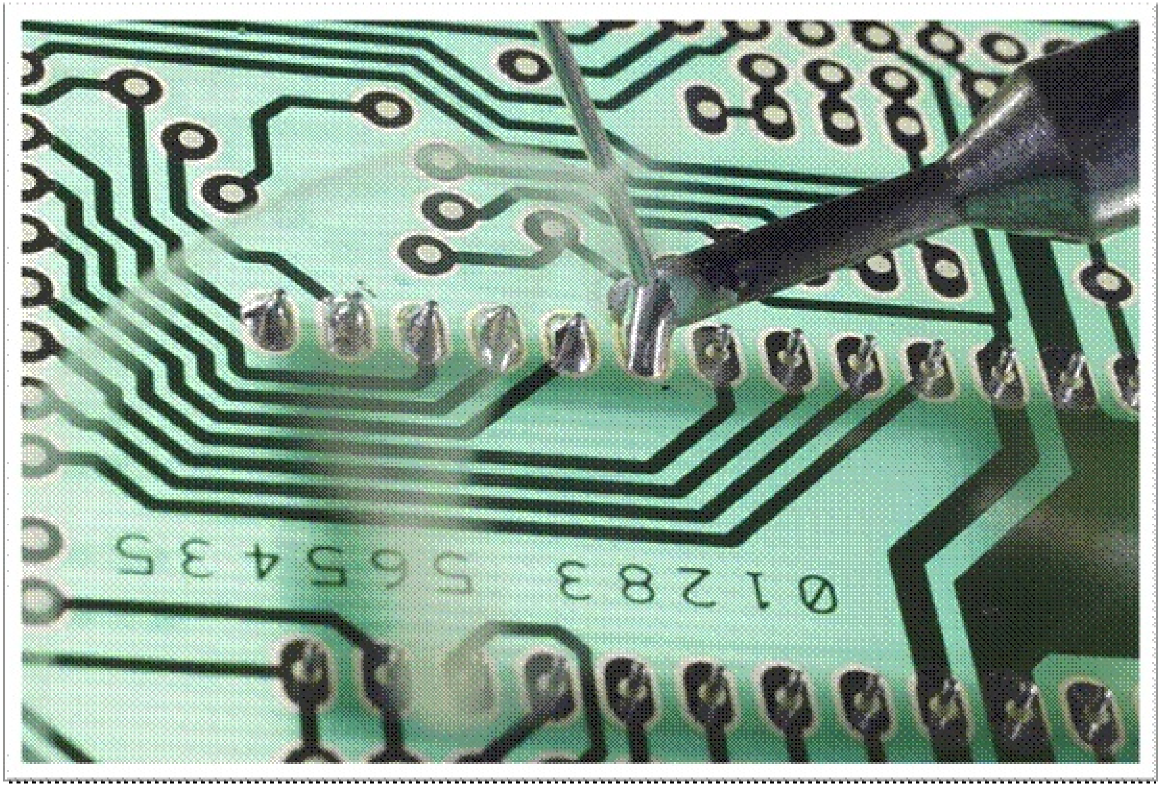

^^ *Soldering a small p.c.b. joint*

With experience, you'll get a feel for how long it takes before you can apply solder. The more metal that is present in the joint, the longer period that heat must be applied for. A small p.c.b. joint takes well under one second to complete. A large metal terminal could take quite a few seconds or more to heat up. As I explained earlier, higher power (wattage) irons cope better with larger workpieces because they recover more quickly and are more "unstoppable", making it easier for them to heat larger workpieces without cooling down so much.

If you apply solder too early, it won't melt properly and the result will be a grey, crystalline joint caused by the solder's melting point temperature not being reached and the flux not having flowed properly. Semiconductors must be soldered as rapidly as possible as they are heat-sensitive, but they're a lot more robust than they used to be.

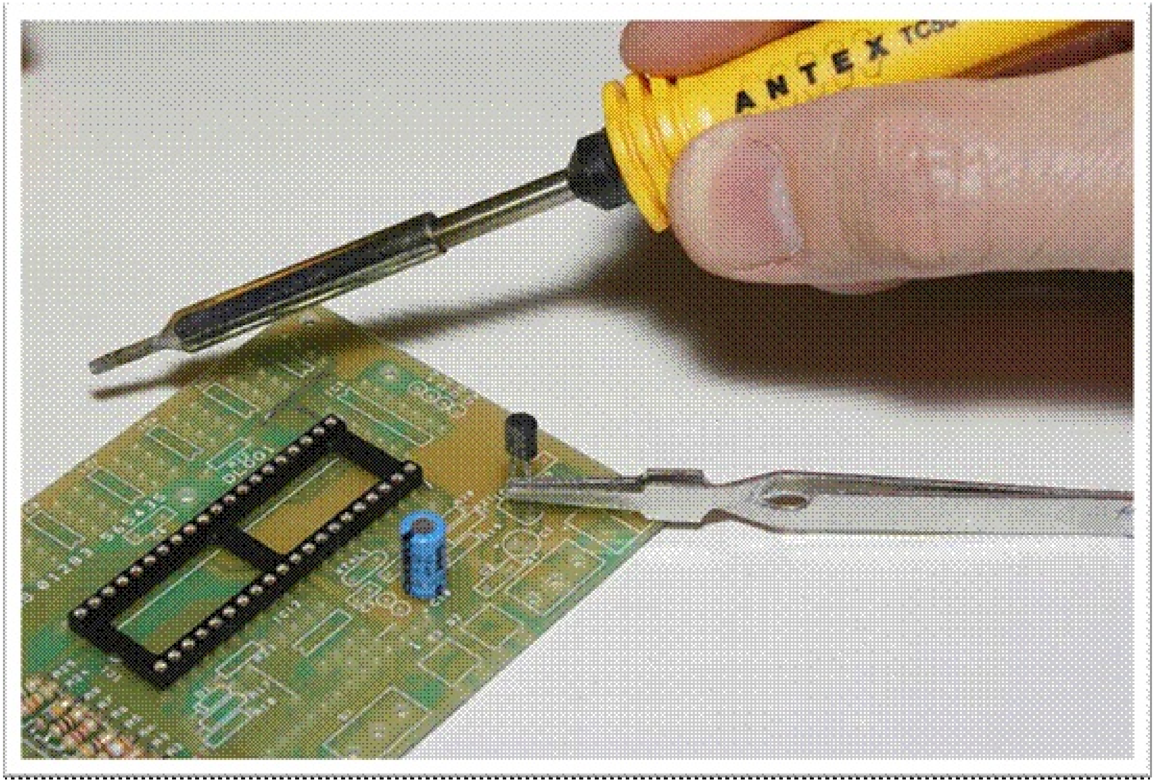

^^ *A clip-on Antex heat shunt fitted to a transistor leg, helps prevent thermal damage due to overheating when soldering it in place. Less essential these days, but beginners find them re-assuring.*

Until they have gained some practice, novices sometimes buy a small clip-on heat-shunt (photo) which resembles a pair of aluminium tweezers. In the example of, say, a transistor, the shunt is attached to one of the leads near to the transistor's body. Any excess heat then diverts up the heat shunt instead of into the transistor junction, thereby avoiding the risk of thermal damage. Applying far too much heat may destroy the part or damage the p.c.b. foil which may lift away from the board.

In due course constructors learns to judge how much solder should be applied to any particular joint. An ideal p.c.b. joint is very slightly concave in shape. If not enough solder is used, the result may be an incomplete joint which may cause an intermittent fault later on. An excess of solder – shaped like a ball bearing - is an unnecessary waste and in extreme cases may cause short circuits, especially on densely-populated boards. There is no need to add more solder "for luck". Professionally-produced p.c.b.s have a green *solder resist* coating which helps to ensure that solder does not stray onto adjacent pads. As a finishing touch, I usually spray the solder-side of a circuitboard with aerosol spray lacquer afterwards. It keeps the solder joints nice and shiny and helps prevent corrosion.

Some components can create hazards during a soldering operation:

Coin cells and button batteries are commonly used as power, clock or memory backups. If heated excessively they can explode without warning due to the build-up of internal pressure. Spot-welders are used in industry to connect tags to them, but if you need to solder wires to such a cell then it should be done as quickly as possible.

Some **memory back-up capacitors** or **electrolytic capacitors** remain energised for a while even when the circuit is powered down. Molten solder is a perfect electrical conductor and in some cases the component’s contacts could be shorted during the soldering (or desoldering) operation. If molten solder shorts it out then the arcing may cause globules of solder to be spattered outwards without warning, potentially risking eyesight damage.

Always ensure that powered components are electrically inert and discharged before soldering. Cells, batteries and battery packs should not be accidentally shorted during the soldering process, to avoid arcing and solder spatter. Note that electrolytic capacitors can also explode after a while if reverse-connected, so observe polarity closely.

Let’s now consider the practical stages of soldering components and wires successfully.

Soldering Step by Step

Earlier I explained the individual factors that affect the quality of a solder joint. These are:

- **Cleanliness** – dirt or impurities drastically hinder good solder coverage.
- **Temperature** – the right level to enable the solder to flow freely!
- **Time** – apply heat for just the right amount of time!
- **Adequate solder coverage** – enough to form a good joint without touching neighbouring areas.

These rules apply whether soldering a p.c.b. or performing other tasks such as interwiring (hooking everything together with connecting wire).

We’ll now summarise the stages of making a typical solder joint – soldering components onto a printed circuit board (through-hole soldering). Most people insert components into the circuit board and simply splay the wires out to hold them in place under spring tension. I find it best to snip excess wire leads off at this stage, to improve accessibility.

- All parts must be bright, clean and free from dirt and grease.
- Try to secure the work firmly to stop parts moving around.
- “Tin” the soldering iron tip with a small amount of solder. Do this immediately, with new tips being used for the first time.
- Wipe the tip of the hot soldering iron on a damp cellulose sponge to remove excess solder or contamination.
- Many people then add a tiny amount of fresh solder to the cleansed tip just before using it.
- Heat all parts of the joint with the iron typically for under a second or so, until it’s heated throughout.
- While heating, then apply sufficient rosin-core solder to form an adequately-covered joint.
- It only takes a second or two at most, to solder the average p.c.b. joint this way.
- Do not move parts until the solder has cooled.
- Remove and return the iron safely to its stand.

This special photo sequence illustrates these stages. It’s best to start with the smallest, fiddliest parts first when soldering a blank p.c.b., because that’s when you’ve got the most access on the board. Accessibility will be reduced as more components are added, so we’ll start with a simple wire link on a professionally designed p.c.b.

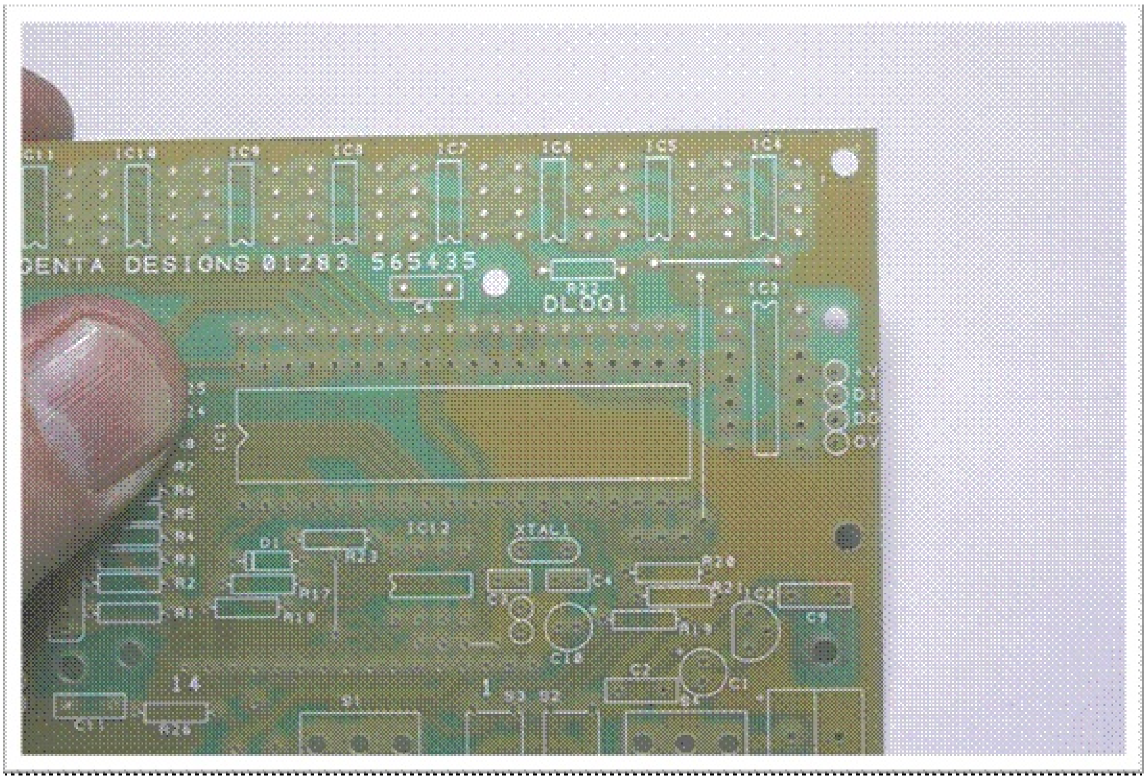

^^ A typical professional blank p.c.b. – silk-screen printing shows what goes where. The underside has been treated with a green solder resist coating, and the solder pads are ready-tinned to help with soldering.

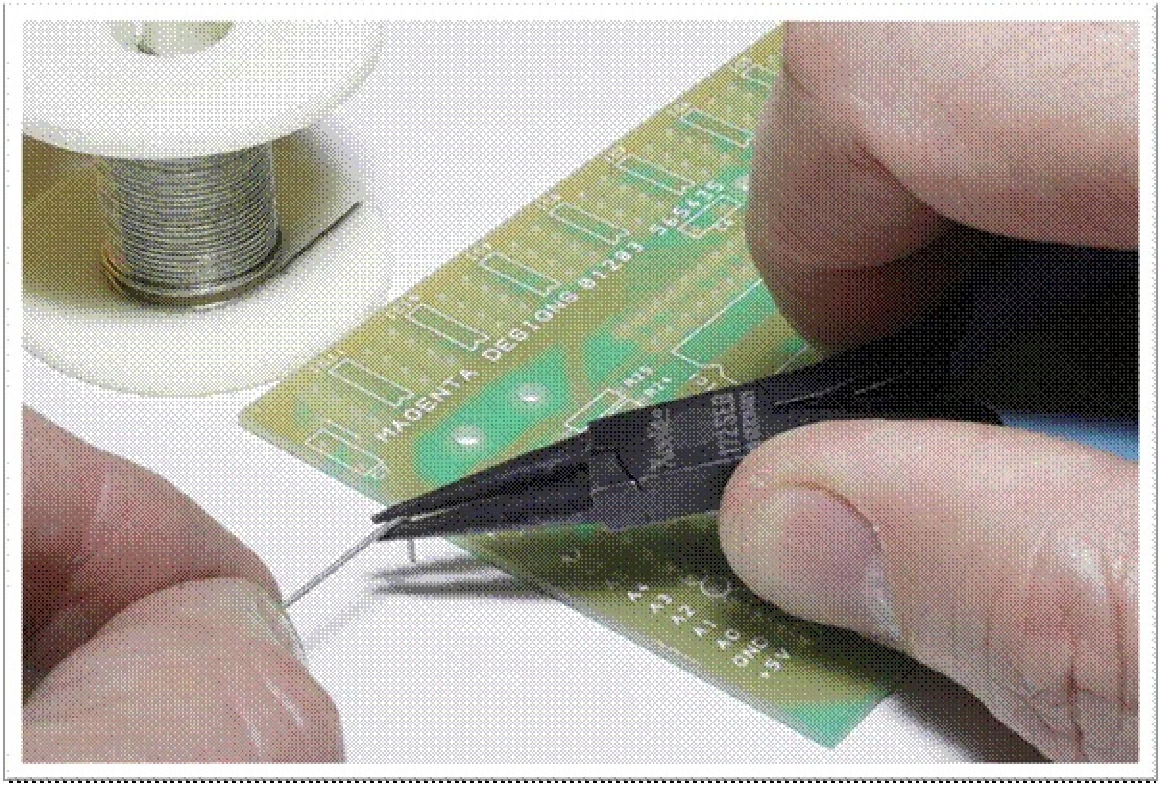

^^ Preparing a wire link for soldering – cut off some tinned copper wire and bend the ends to fit the p.c.b. correctly. Round-nose pliers (shown) are perfect for this, but ordinary electronics or "radio" pliers will do.

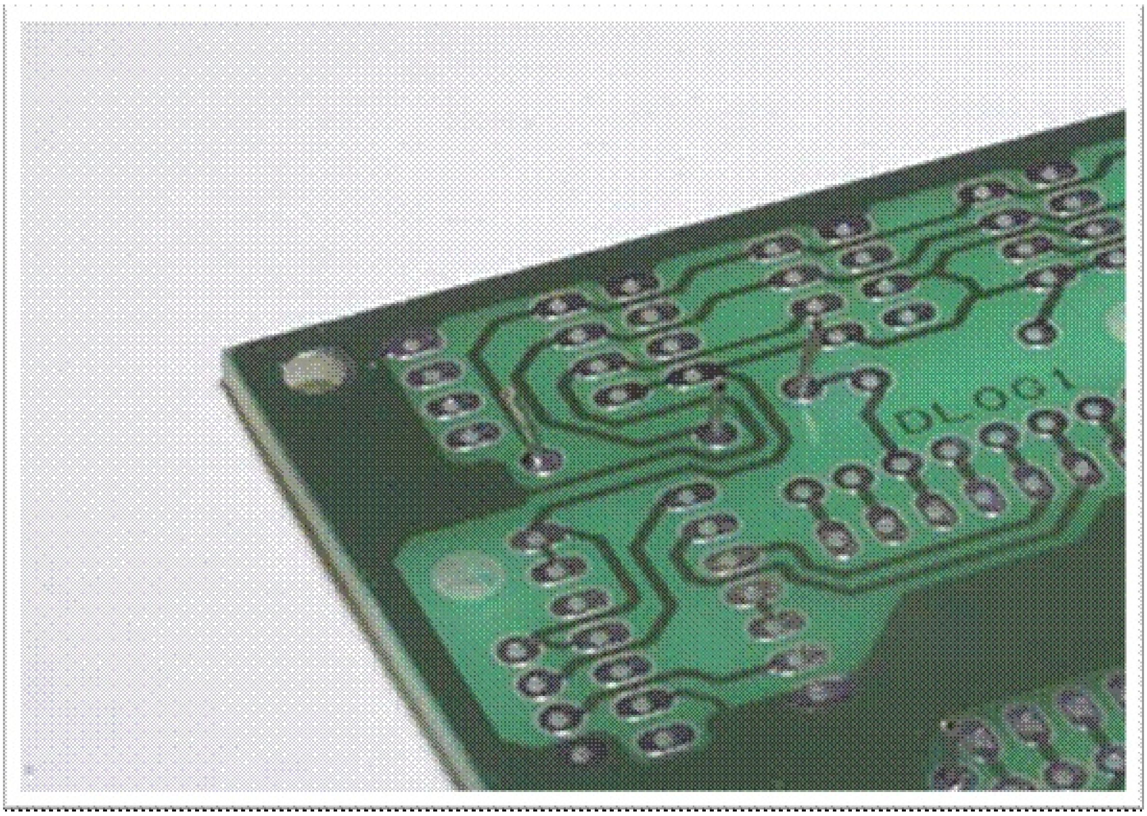

^^ Wires are prodded through the holes in the board, then turn it upside down to view the solder side. You can then "spring" or splay the ends apart slightly, so they are held in place while you solder them.

^^ Wipe the hot soldering iron on the damp sponge to clean the tip. Do this periodically when contamination, flux deposits etc. build up on the iron to keep the tip nice and shiny. Tip Tinner & Cleaner helps too.

ANTEX
660 TC

^^ It often helps to dab a tiny amount of solder wire onto the tip, to improve heat transfer.

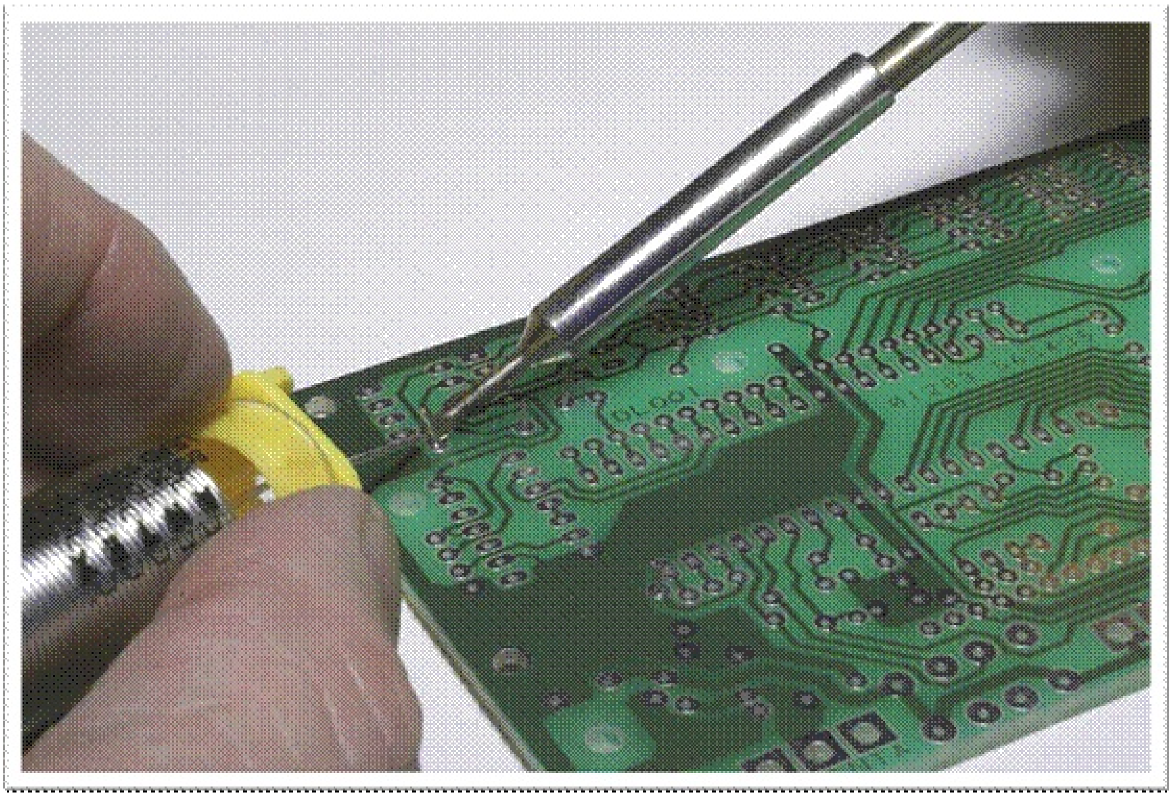

^^ Then apply the soldering iron to heat both the solder "pad" and the wire end at the same time (say <1 second). Apply a few millimetres only, of solder. Then remove the soldering iron immediately and allow the joint to cool down by itself. The green solder-resist coating ensures solder doesn't flow onto neighbouring pads.

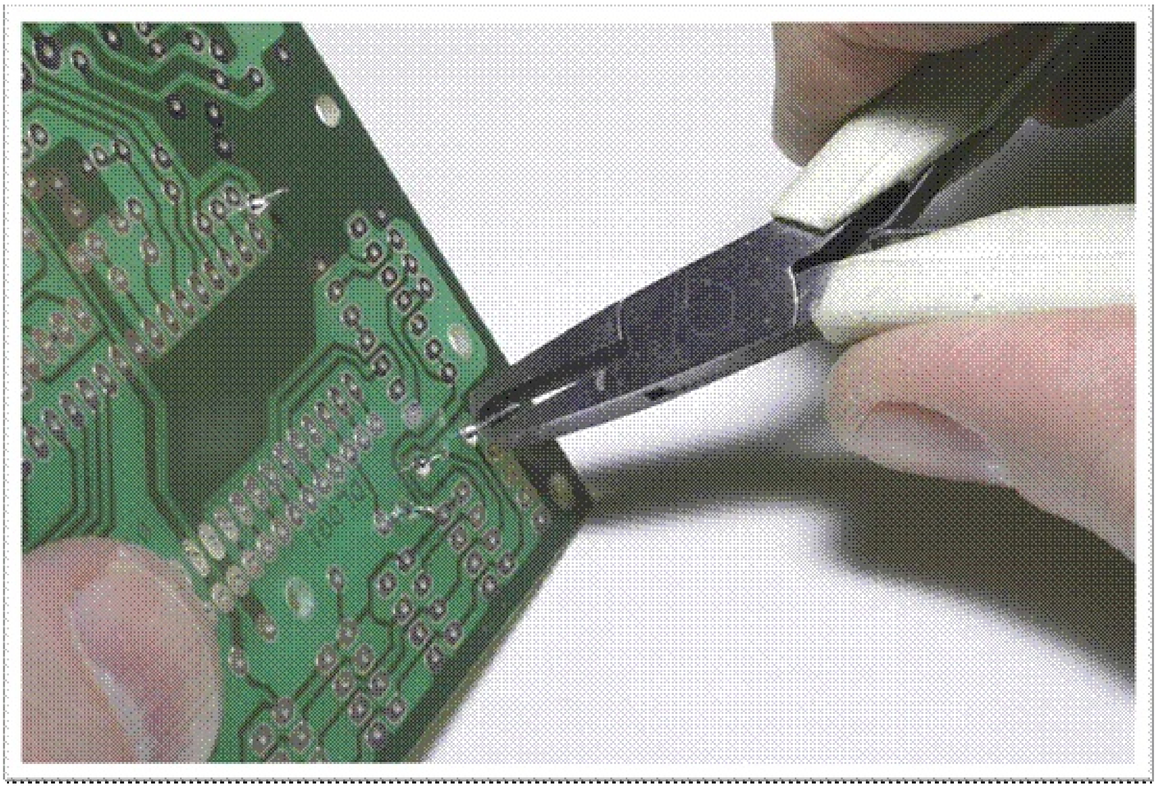

^^ These "end cutters" can then be used to snip off excess wire after soldering….

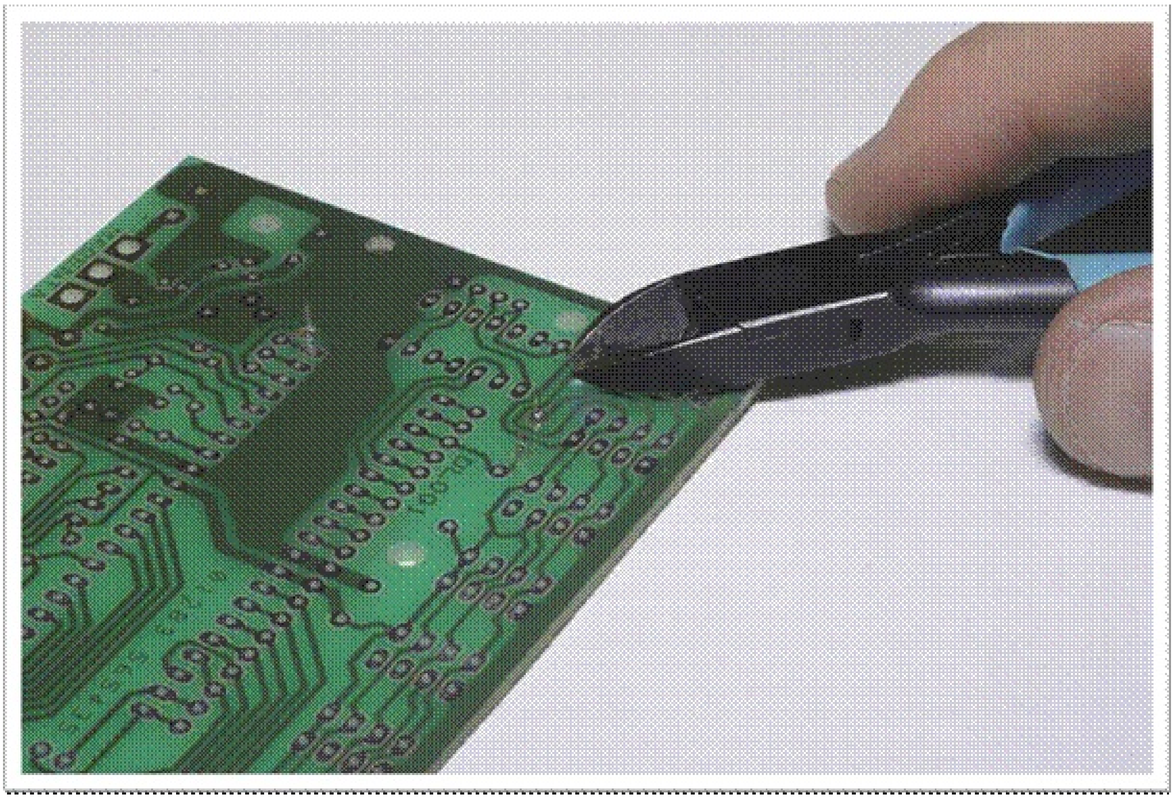

^^ … but ordinary electronics "side cutters" are fine for trimming or snipping wires.

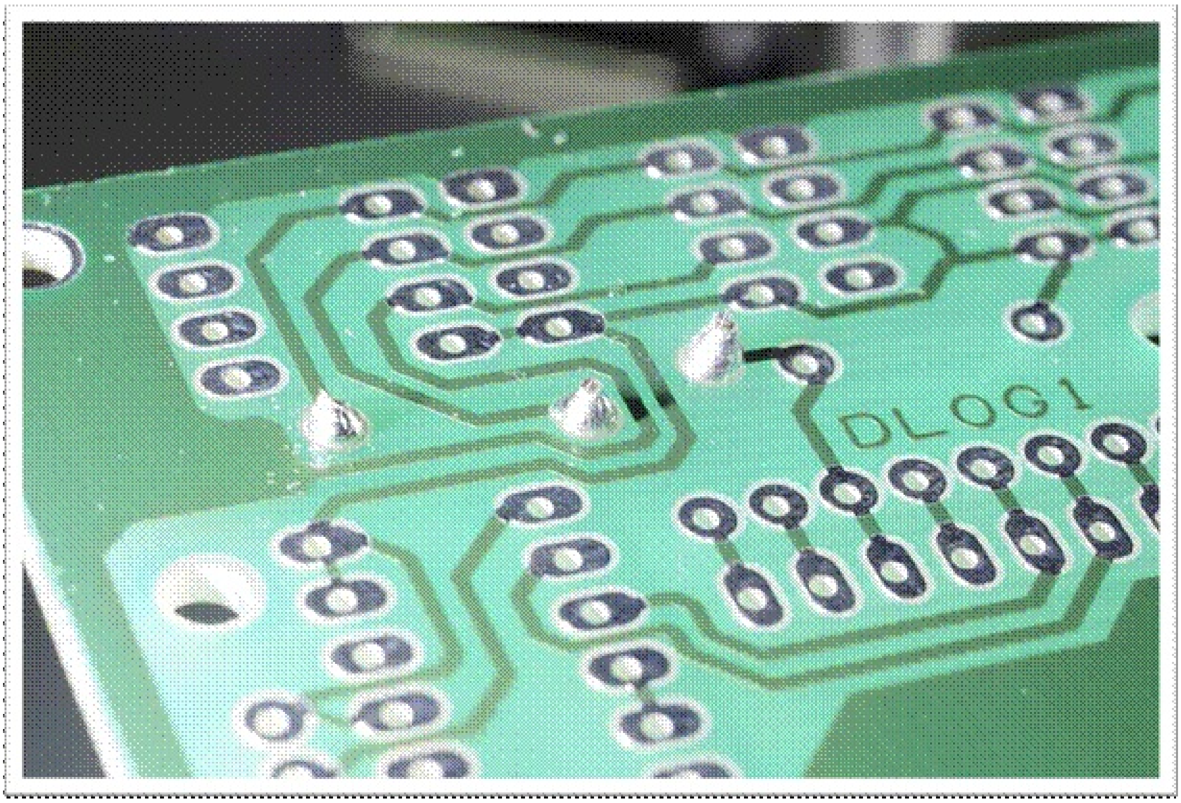

^^ The ideal solder joint should be smooth and slightly concave, quite shiny, not dull or crystalline-looking nor ball-bearing shaped.

Try to get consistent results in your soldering, but don't worry too much about the lack of uniformity – this is soldering by hand after all, not by precision machinery, and some slight variation is OK provided the solder coverage is good and the soldering is generally clean and effective. You'll improve with practice, that's for sure!

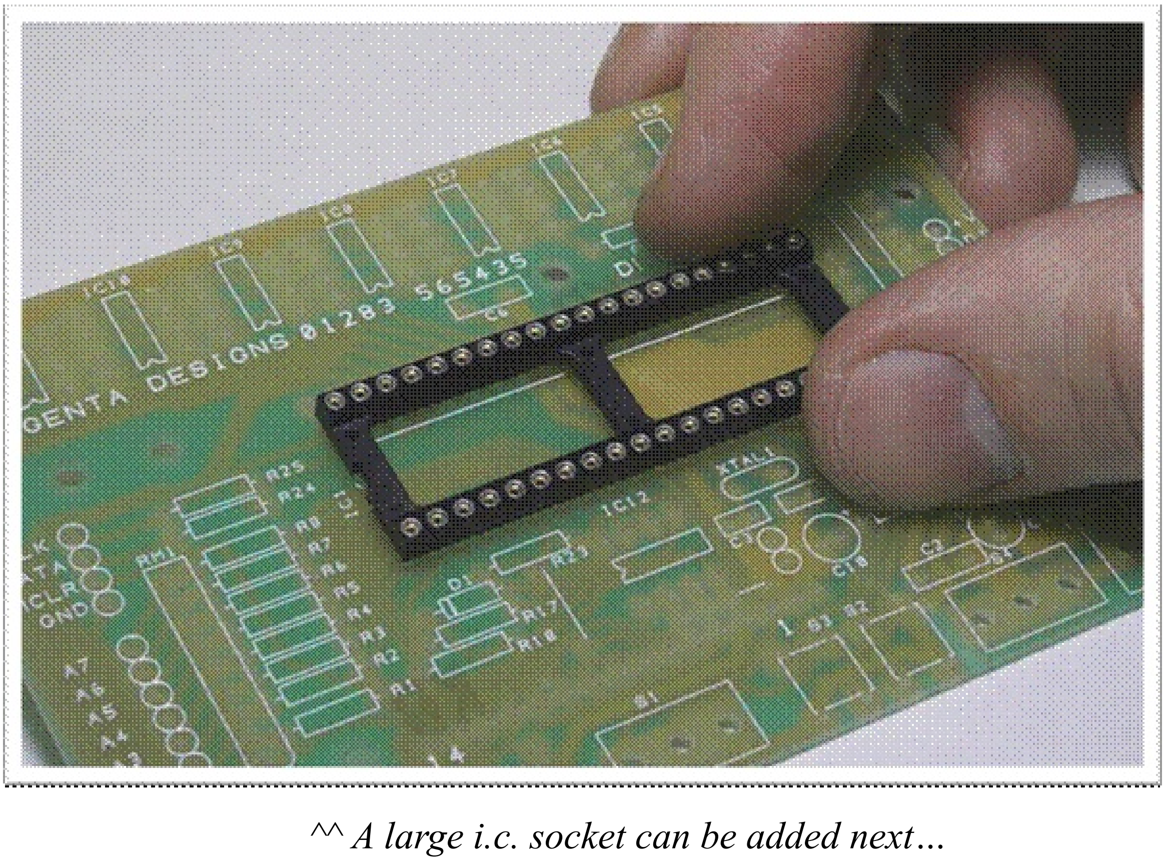

^^ A large i.c. socket can be added next…

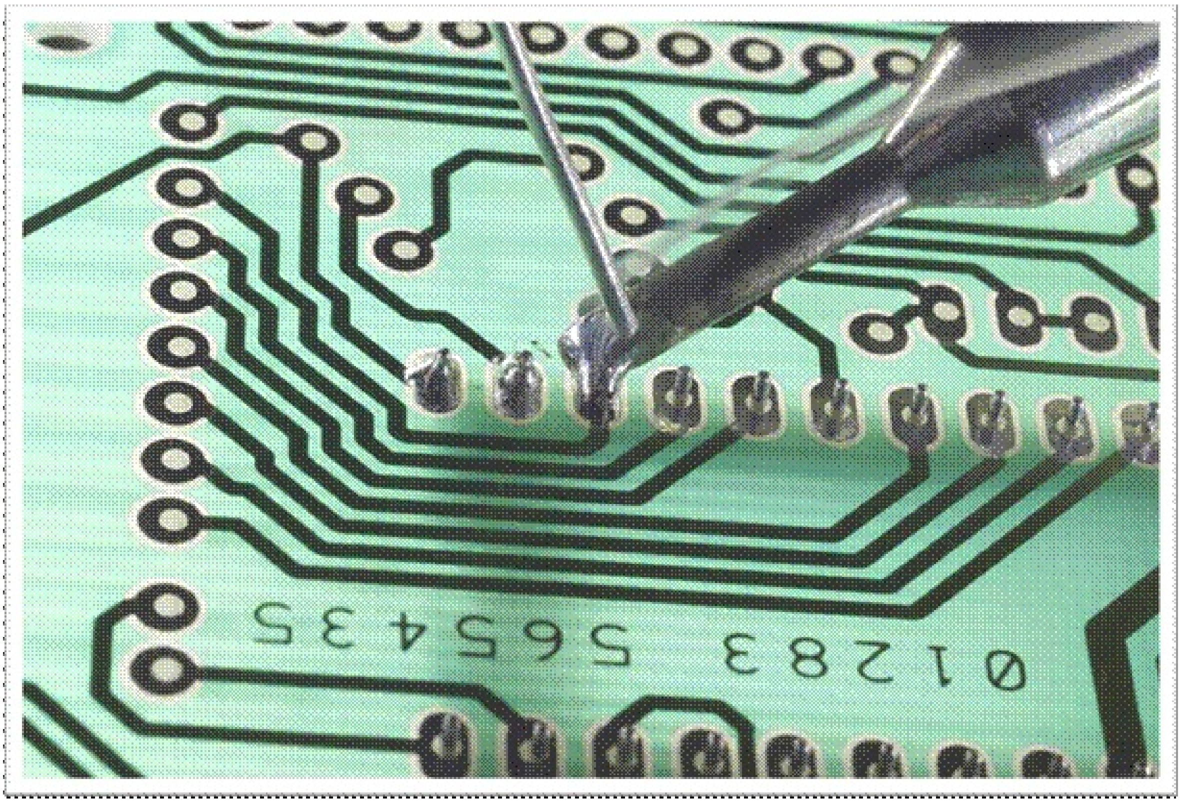

^^ Go along the rows of pins and solder them one at a time. See how simple it is! Heat both the solder pad and pin with the iron, and dab on a little solder wire to ensure full coverage. Ensure the socket is flush against the board.

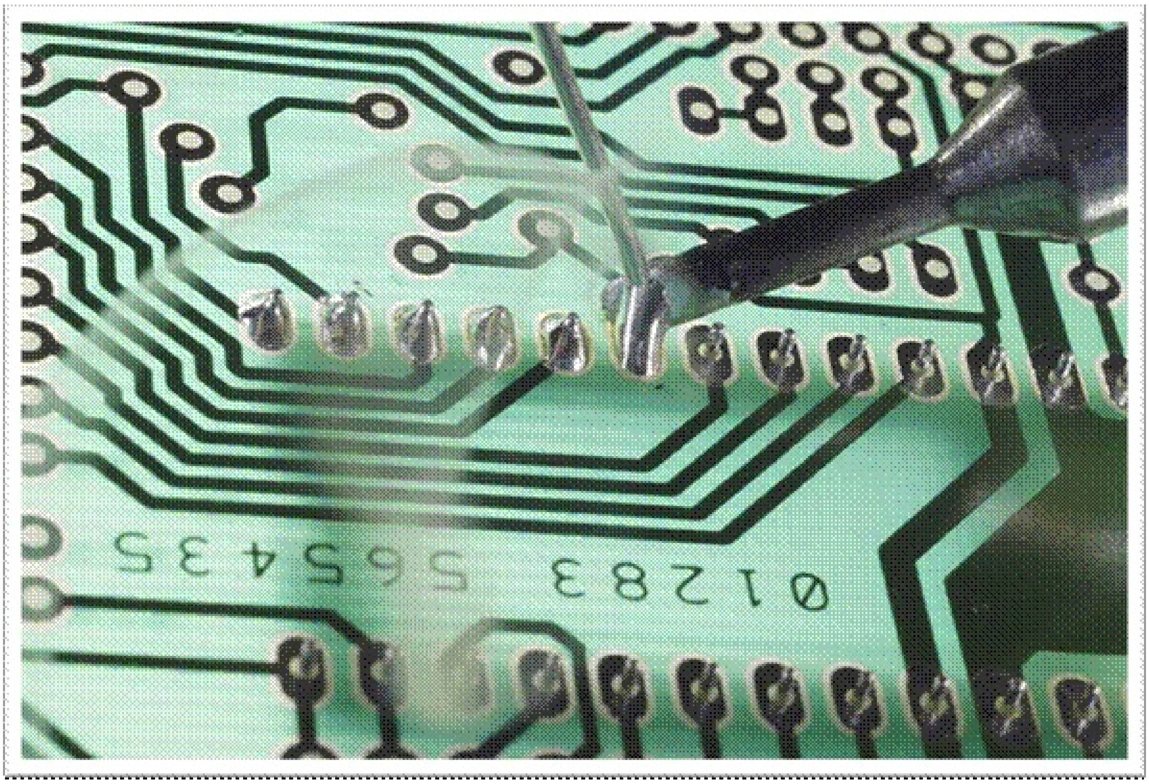

^^ As hardly any metal is involved each pin should take well under a second to solder at most.

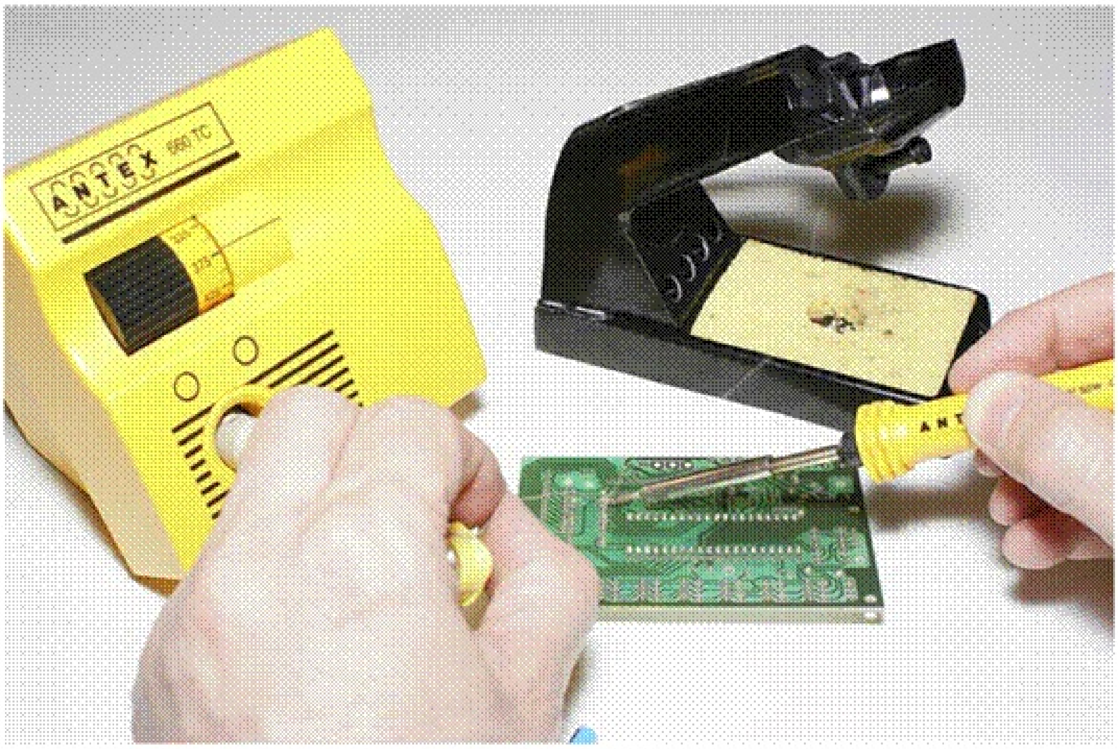

^^ Continue until the i.c. socket is soldered.

Once the smallest parts are soldered into place, you can continue to solder the remaining components. It's easiest to handle the smallest so-called "discrete" parts first while you still have plenty of room on the board. I usually solder resistors and capacitors next. The principle of soldering them is just the same as a simple wire link: insert them from above till they are flush on the board, then splay their wires a little to hold them in place, and preferably snip off at least some of the wire to give you more access.

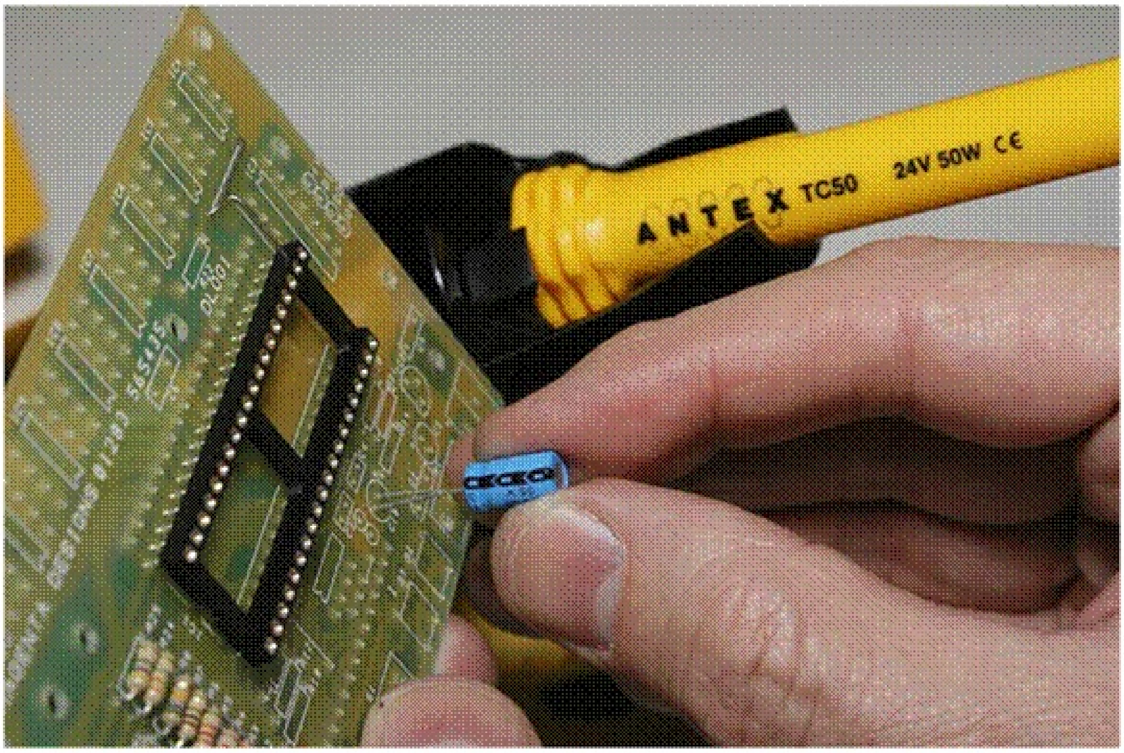

^^ Continue by inserting "discrete" components from above, splay their wires out

underneath to secure them in place, then snip and solder the joints exactly as before. Some parts like this blue electrolytic capacitor are polarity sensitive (note the – sign), and must be inserted the right way round. Same is true of every semiconductor.

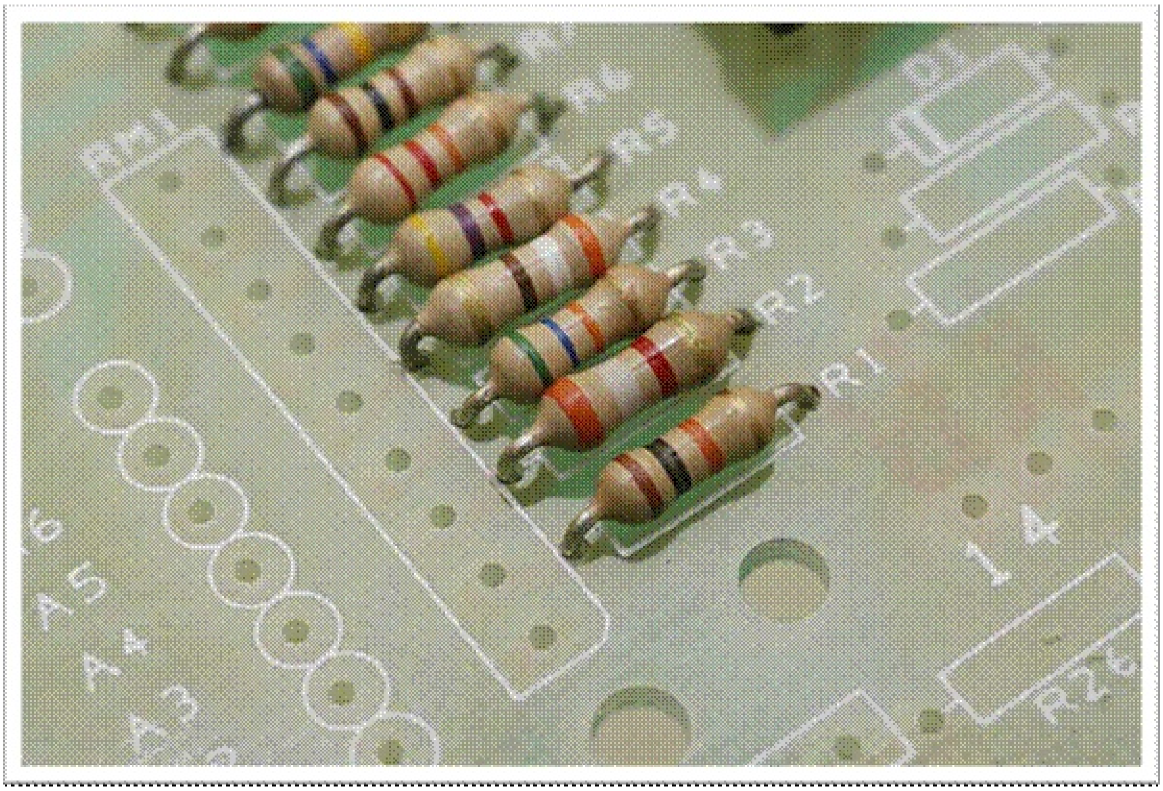

^^ This row of ¼ watt resistors was next. I like to show the colour codes all the same way round, with gold or silver (tolerance bands) on the right for consistency.

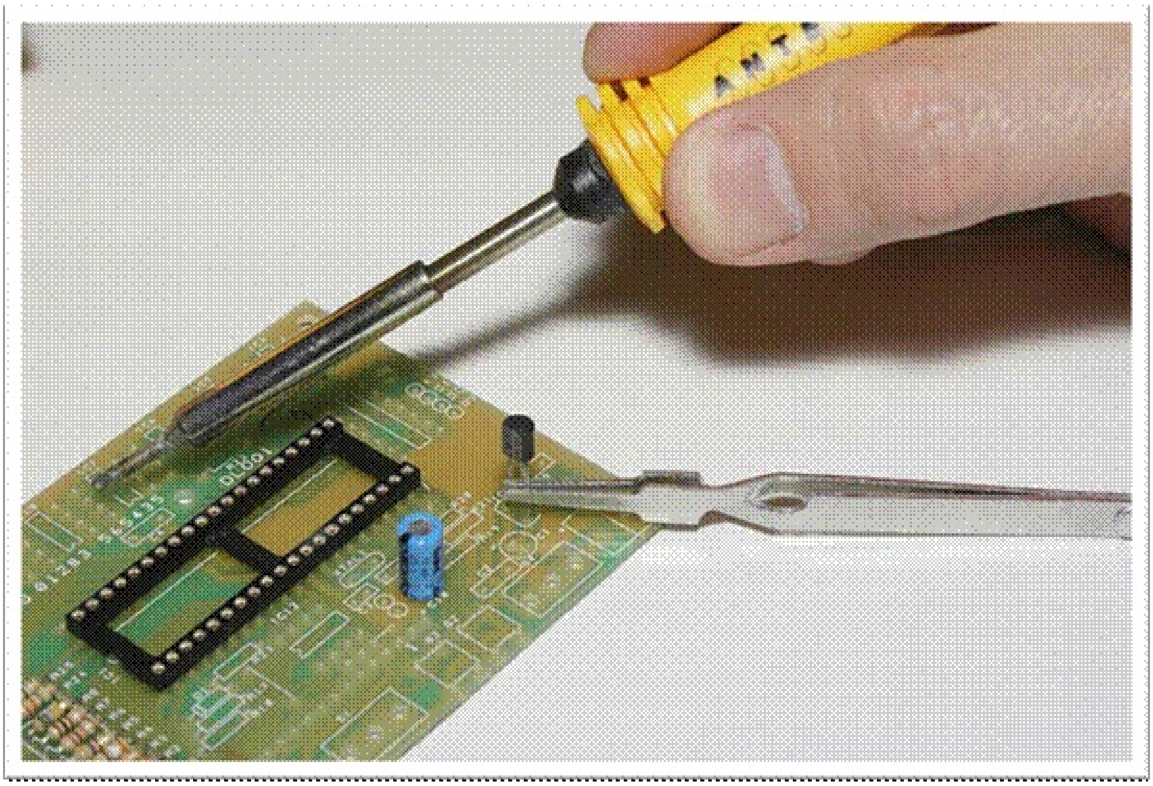

^^ This transistor was next. Solder it quickly to avoid damage, and observe polarity correctly, so it is the right way round on the p.c.b. The heatshunt is optional.

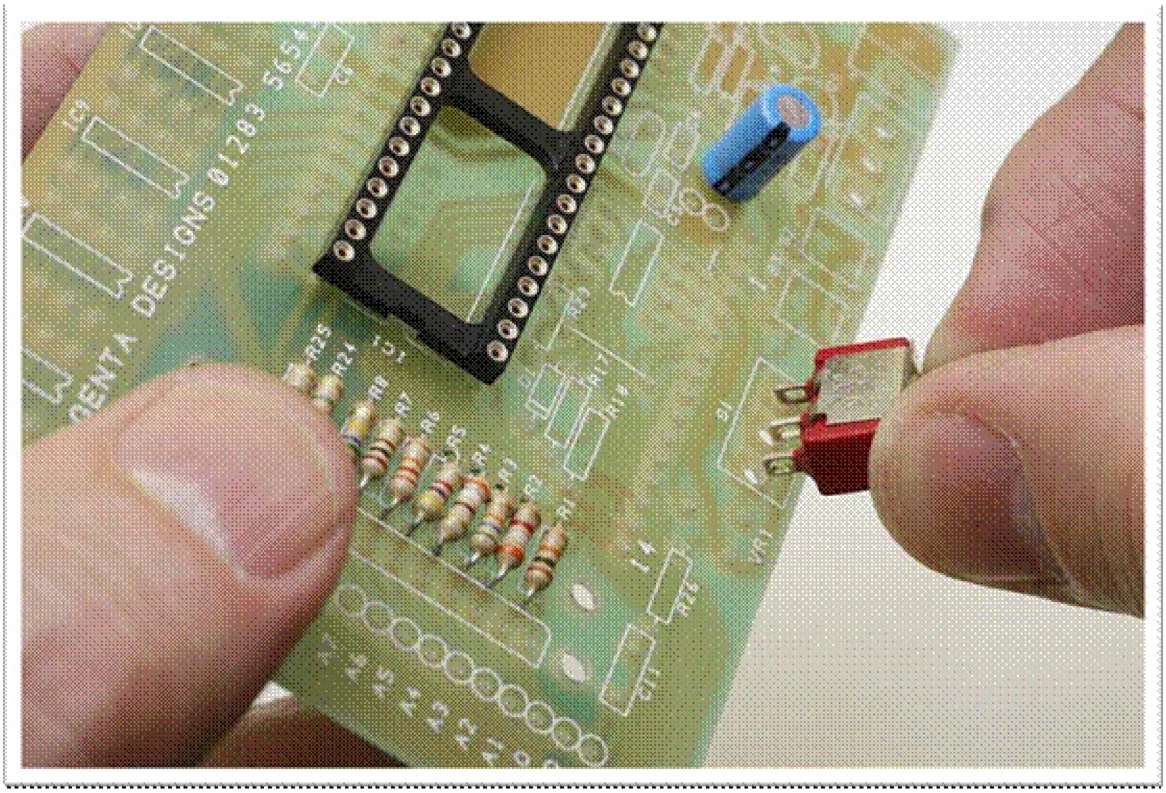
GENTA DESIGNS 01283 5654
IC1

^^ Unusually, this toggle switch fitted directly onto the board as well which saved wiring it up. Neat idea!

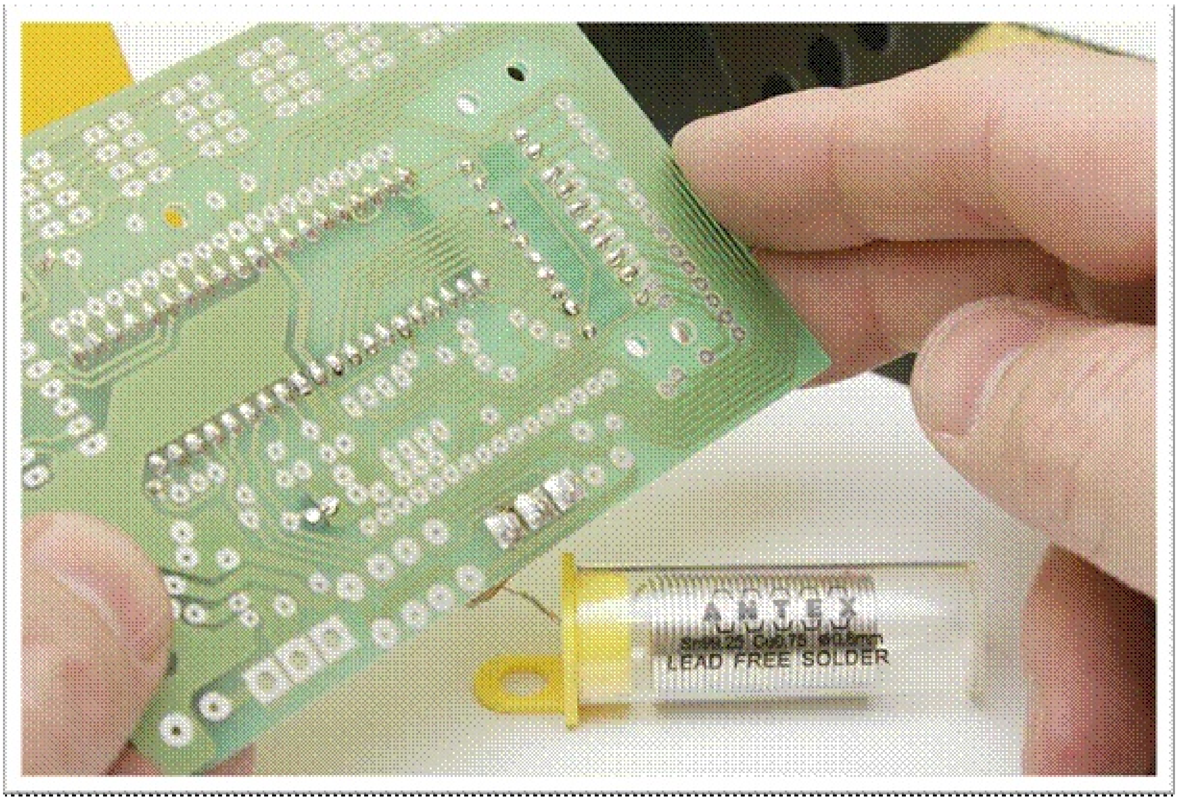
ANTEX
LEAD FREE SOLDER

^^ *As there is comparatively more metal to heat up, it'll take longer to solder the switch terminals, and you'll need more solder as well. Thicker gauge solder is useful at such times. Allow say 2-3 seconds to solder each terminal.*

Don't forget to clean the soldering iron tip on its damp sponge every now and then, to ensure the bit is kept clean and shiny. Later on I'll show you how to correct any problems by desoldering using various techniques. Next, we'll move on from printed circuit board "through-hole" soldering and look at how to handle wires and leads.

Interwiring – get hooked up!

With practice, through-hole soldering of p.c.b.s will become second nature. There's no substitute for tackling some soldering jobs though, particularly trying a simple kit based on a quality p.c.b. which will boost your confidence enormously.

Most electronic devices need connecting up to external components such as battery packs, speakers, l.e.d.s or switches. Usually, multi-stranded connecting wire is used to connect circuit boards and external parts together. Unlike single solid-core wire, *multi-stranded wire* is flexible and vibration-resistant. Hobbyists mainly use 7/0.2mm wire (7 strands, each 0.2mm diameter) for low-voltage hook-ups although much Chinese equipment uses much thinner wires than this. So let's look at some aspects of soldering this type of wire.

In a separate photo sequence I show how a potentiometer (a panel-mounted variable resistor) and a light-emitting diode (l.e.d.) are connected using multi-stranded wire. The same principles of soldering apply to most other components including panel-mounted switches, loudspeakers, buzzers, audio sockets and more.

Components usually have terminals or "tags" to which wires can be soldered. Start by ensuring the component's tags are clean: otherwise solder will not wet properly and the joint will be impossible to solder, so all contamination must be removed. This is especially true of parts that have been in storage a long time. The connections often oxidise or blacken, so clean the solder tags with e.g. an abrasive glass-fibre brush, or a needle file or abrasive paper. Using a glass fibre brush was shown earlier in "Cleanliness and Tinning the Bit".

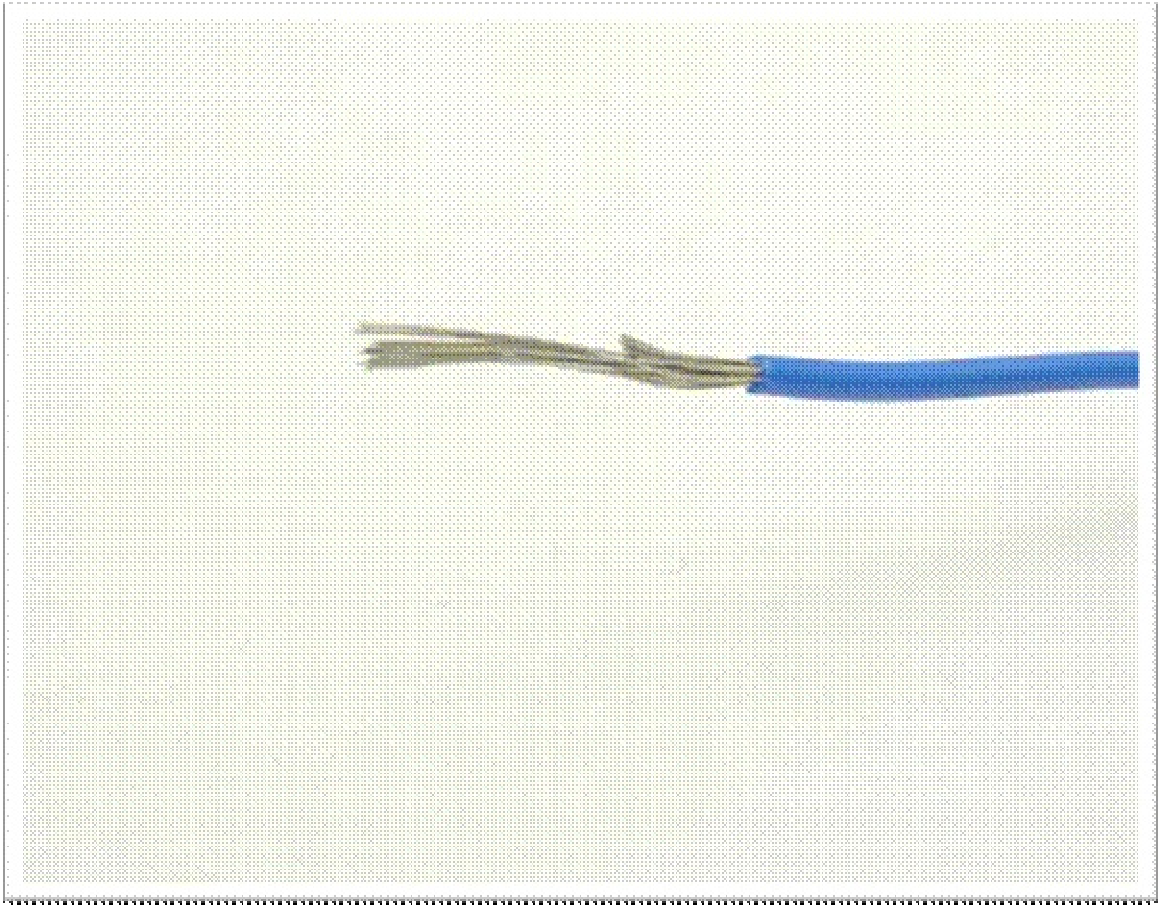

^^ How not to strip insulation off wire: some of the cores have also been cut – avoid doing this!

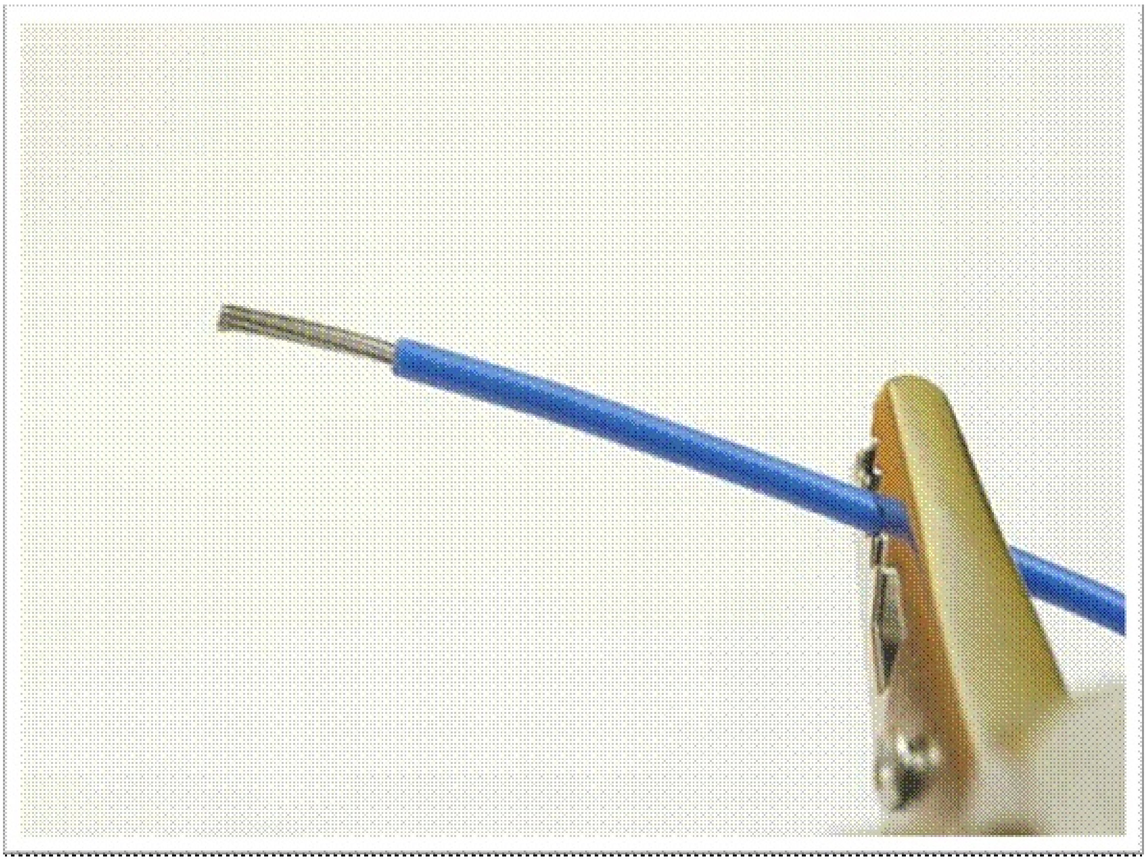

^^ Gripping a wire end in a "Helping Hand" croc clip can help with soldering.

After stripping a short length of insulation from the connecting wire, there are two ways to solder it to a component's solder tag. The first way is to "*tin" the stripped wire end to solidify it* - just heat it with the iron and melt a little solder on it, and let it cool. Poke the wire end through the solder tag, apply the hot iron to both parts and solder them together using a few millimetres of solder.

Although the assembly doesn't hold itself together so well during soldering (consider a Helping Hands jig if needed) this is quicker and easier to make and also easy to desolder again, and is perfectly adequate for most joints of this kind. The majority of commercial wire joints seem to be made this way.

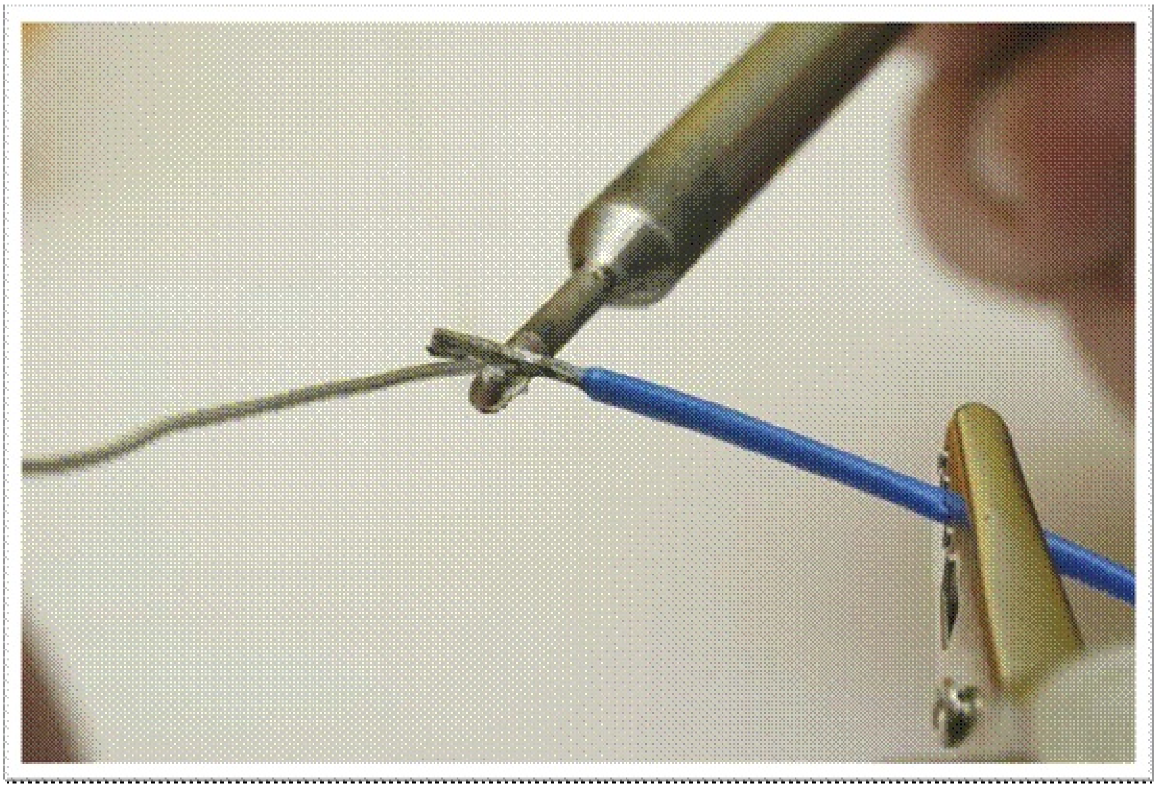

^^ Apply a hot iron and solder, in order to "tin" the wire ends. This makes them into a solid.

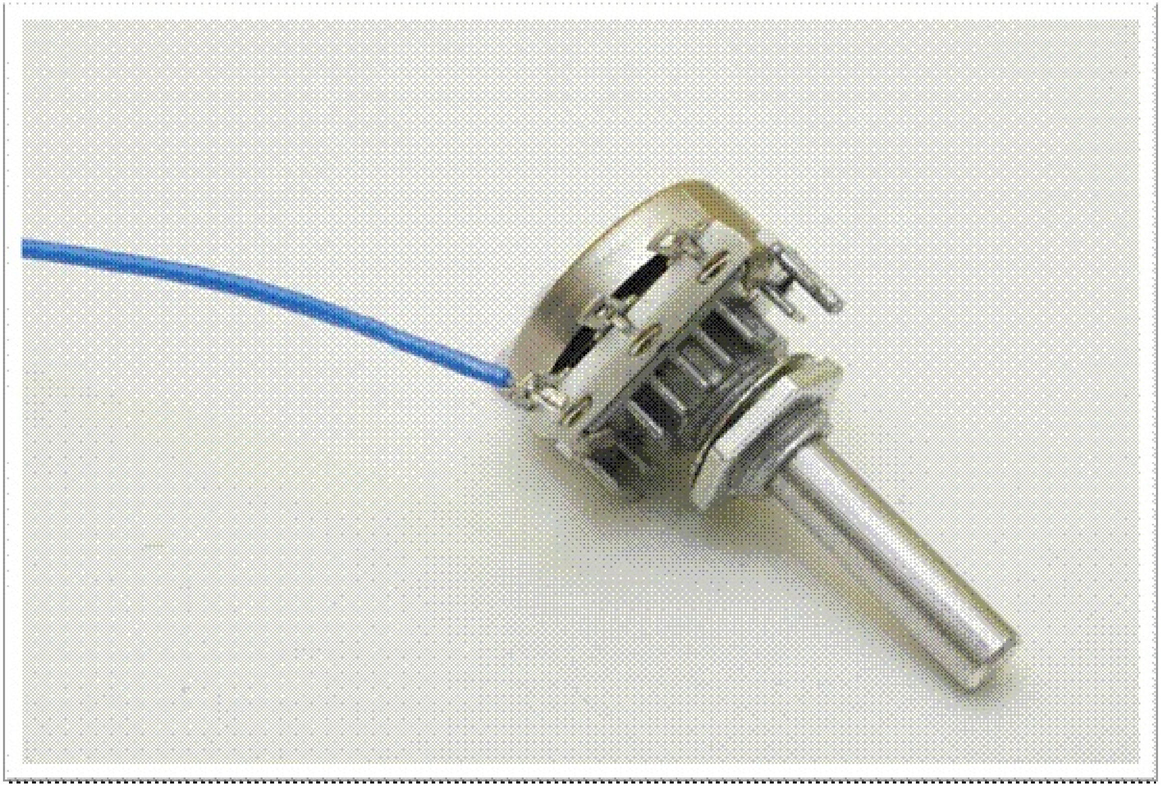

^^ Then feed the tinned wire through the hole in the solder tag. Crop the wire with cutters if needed…

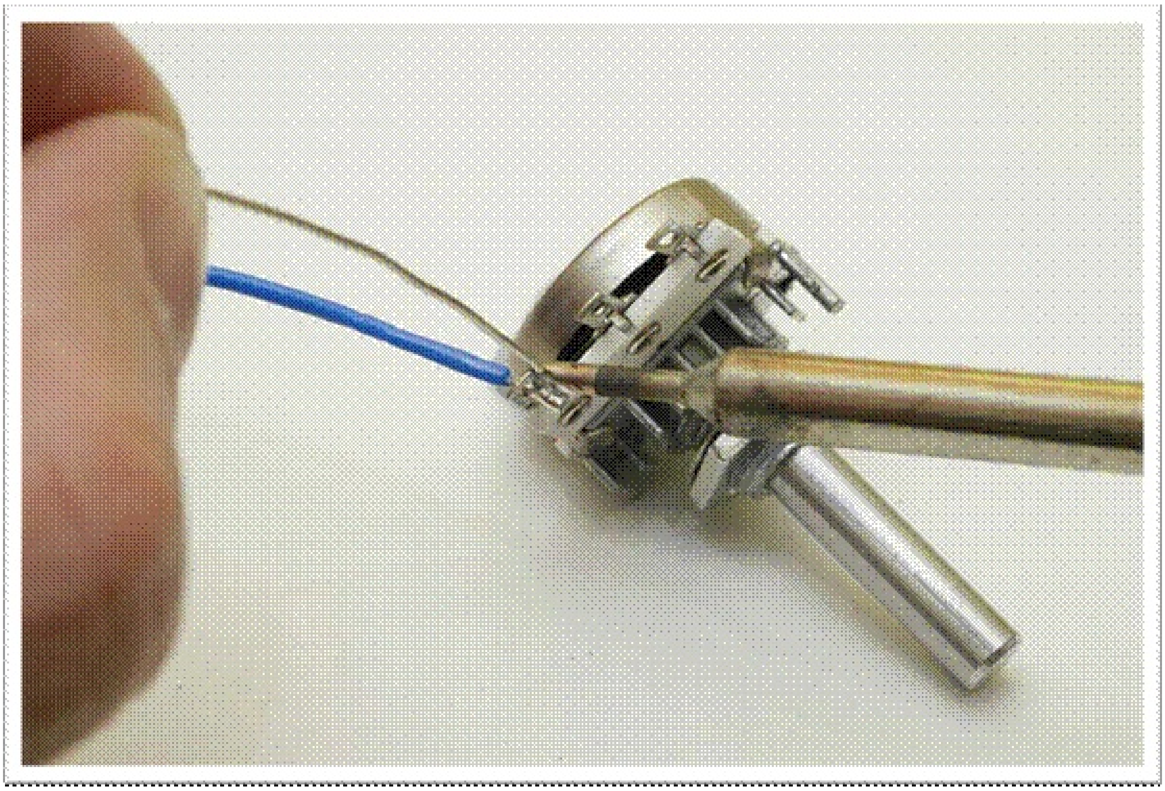

^^ … and apply the iron to heat the joint. As there's quite a lot of metal to heat, allow several seconds to heat it up and then simply solder the wire and solder tag together with a dab of solder wire.

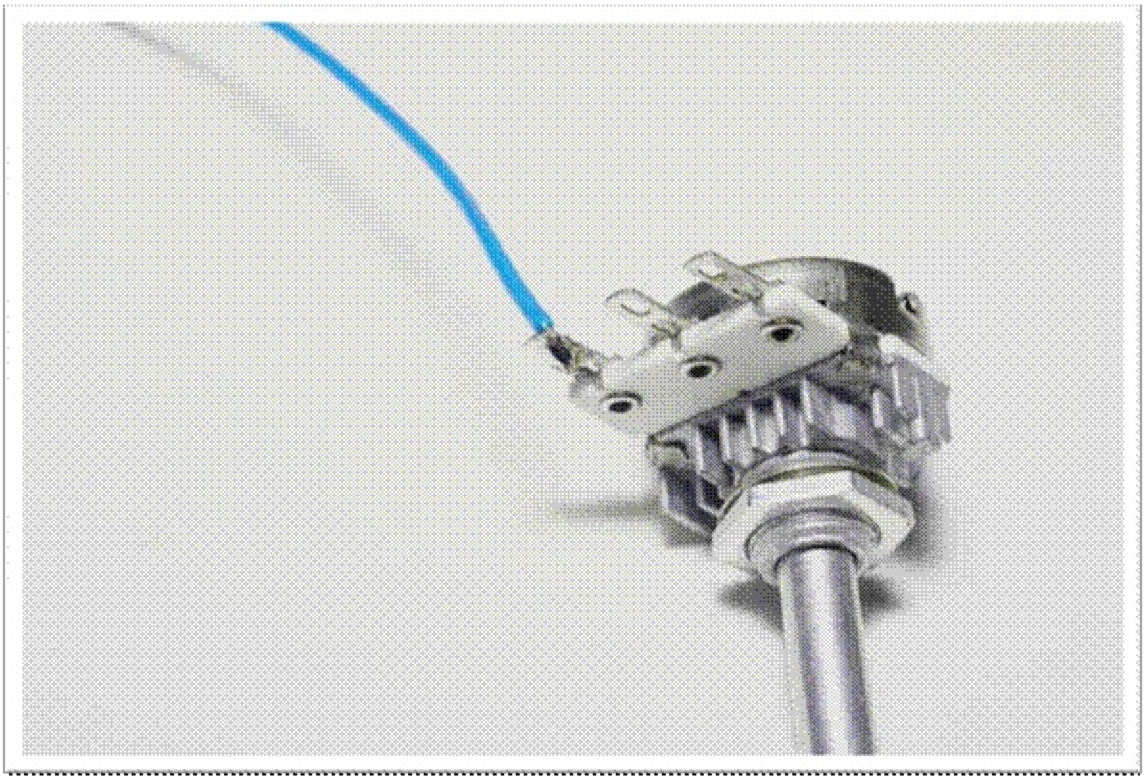

^^ The result is a perfectly satisfactory solder joint.

The second way is to loop the **untinned** wire through the tag a few turns and then solder it. This secures the wire during soldering, but it's messier to desolder if things go wrong (see later):

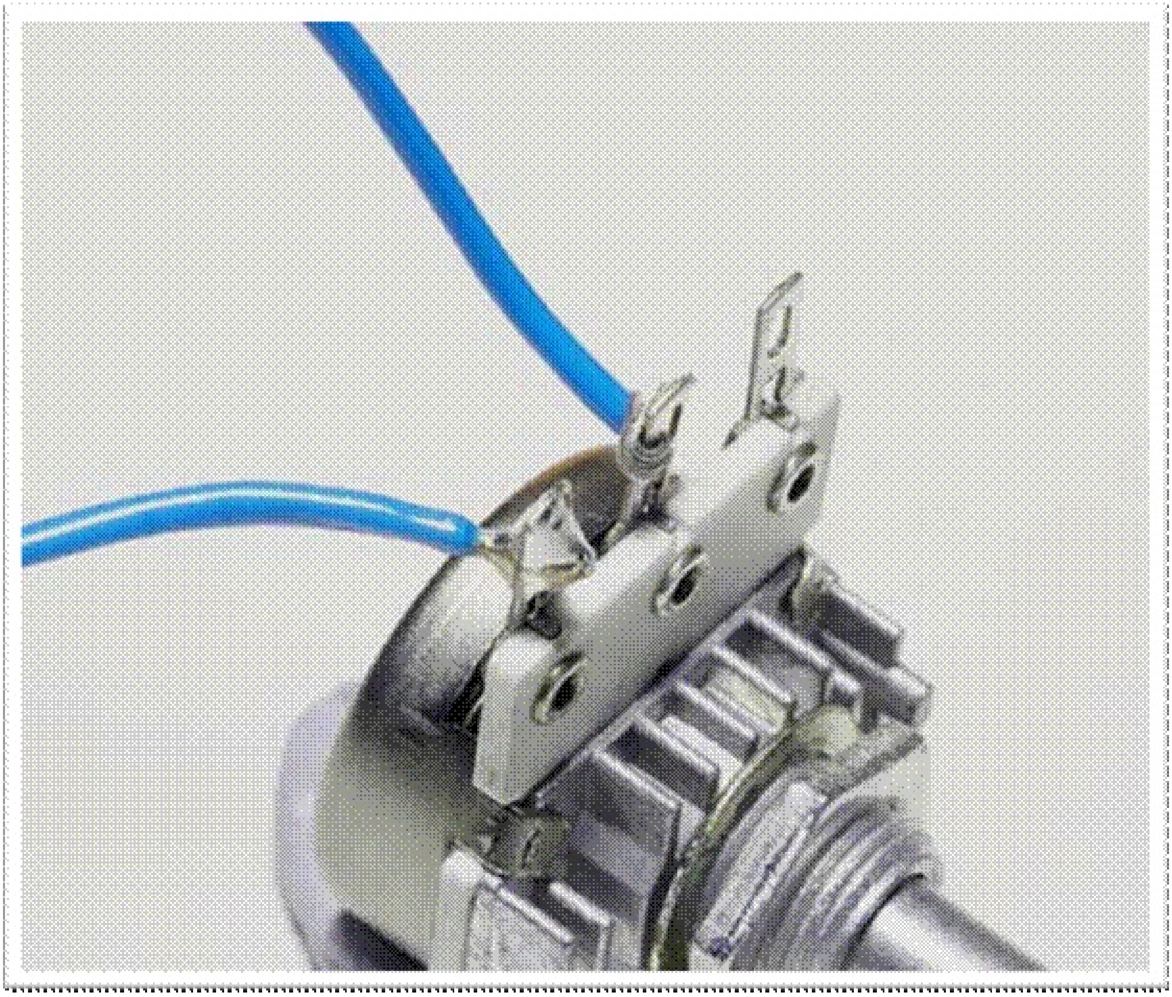

^^ *The centre wire (potentiometer wiper terminal) has been wrapped around to secure it before soldering.*

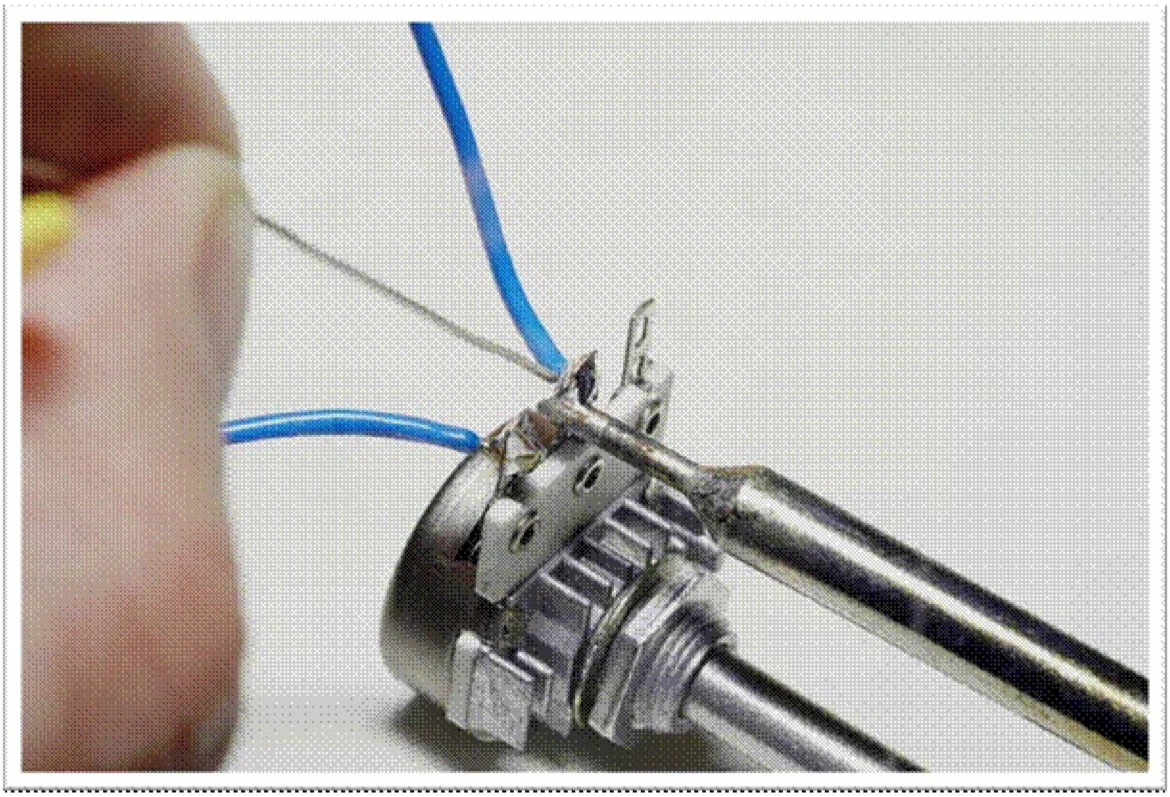

^^ *Solder the joint in just the same way – heat it up and dab on some solder wire.*

Much electronic equipment is connected or inter-wired this way and soldering everything together is a key stage in assembling any electronic project. I’ll explain some refinements you can try, later.

Wire joints

Soldering isn't always the best solution to some problems, for reasons that I explain in the next section called Fatigue and Breakage. However you can solder wires together to join them simply and cheaply, and next I show a classic wire joint of this kind, sometimes called a **Western Union** joint.

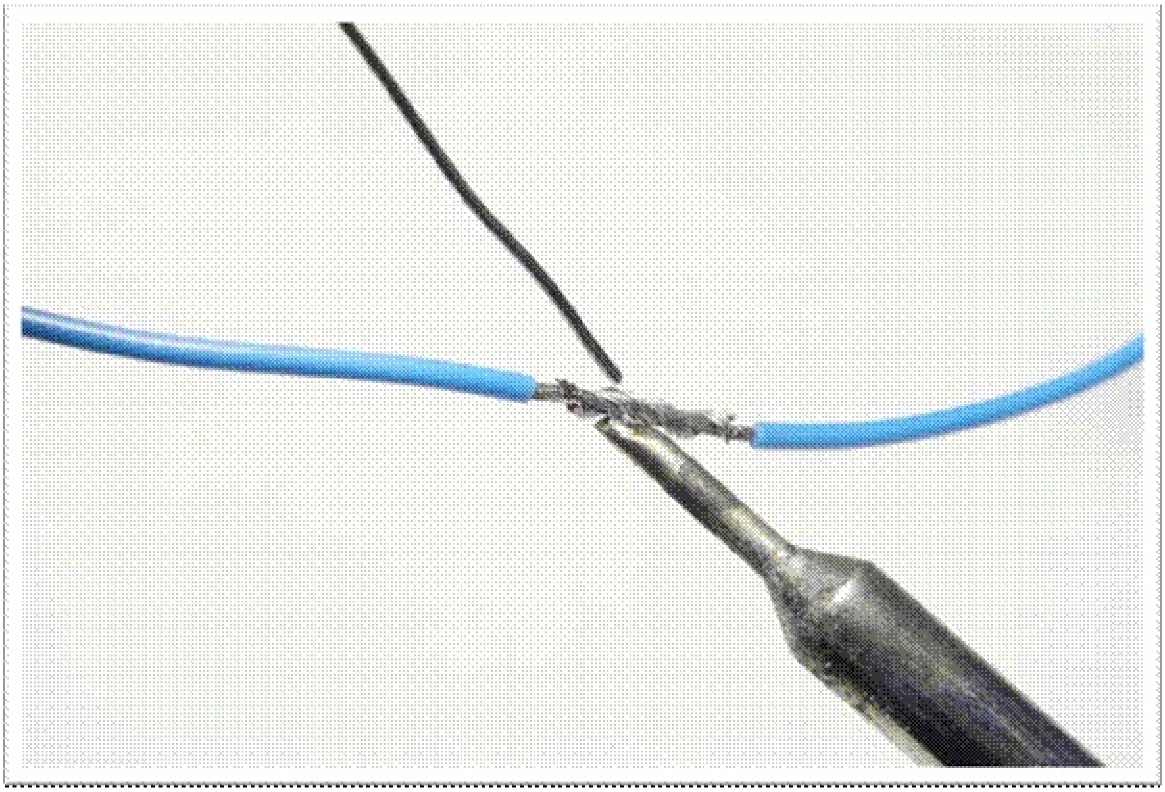

^^ *Soldering a Western Union wire joint.*

Strip the ends from each wire using wire strippers, and try not to cut any of the stranded cores or this will create a weakness. (Stray strands of wire can also be a hazard.) Then twist the wires securely together several times, ideally four or five. Apply a soldering iron to the exposed joint and dab on some solder until it flows fully over the joint, then remove the iron and allow the joint to cool.

The wire's plastic insulation may shrink back a little due to the heat, but try not to overheat it excessively. If you have the resources you can finish off by insulating the joint with some heatshrink tubing (slide it over a wire before you make the solder joint!). Otherwise use ordinary PVC tubing or insulation tape if you have any.

A solder tag can be soldered easily enough, and it's probably best to feed a stripped wire end through the tag and twist it over several times before soldering it, so it's held together when soldering.

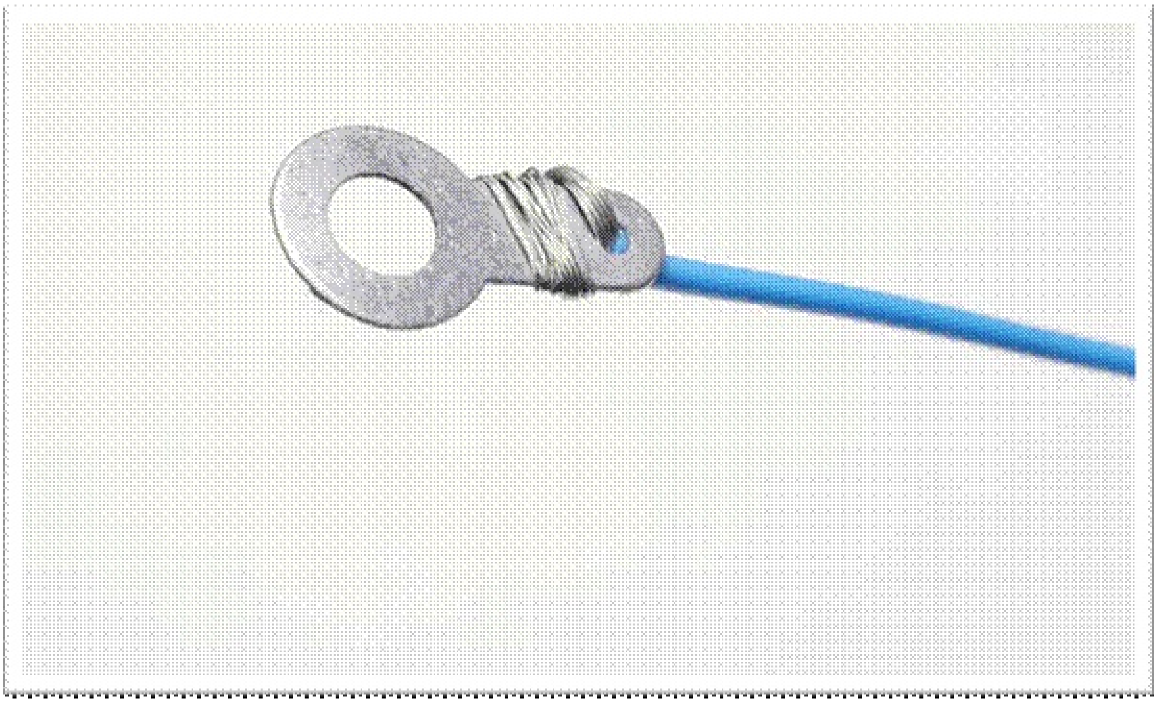

^^ *A solder tag is soldered by wrapping a stripped wire end through the hole and round again, to secure it.*

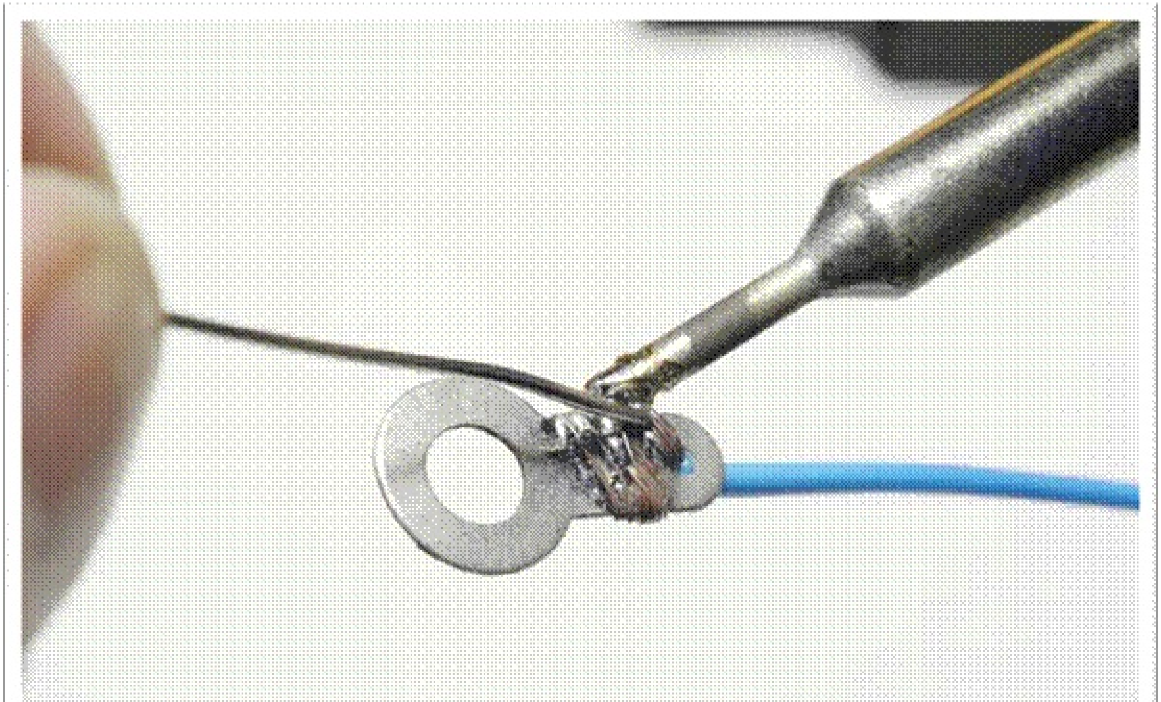

^^ *Apply plenty of heat and solder for a few seconds.*

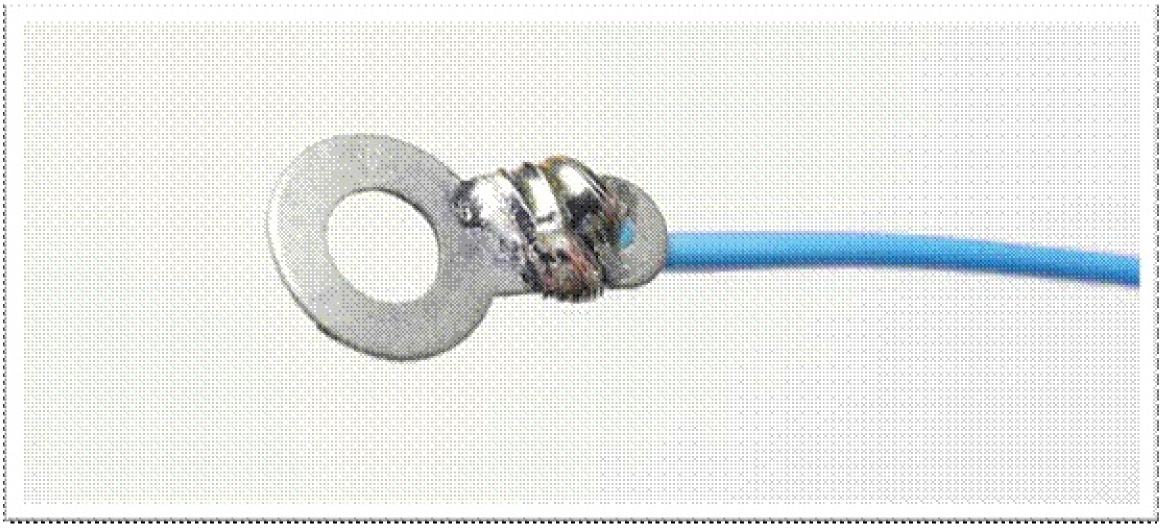

^^ *The completed solder tag*

The next sequence show how the same basic principles of soldering wires are used to connect up a device such as a panel-mounting l.e.d.. A series resistor can be wrapped around and soldered directly to one of the l.e.d. solid wires, and then two multi-stranded connecting leads are soldered to the l.e.d. and resistor. It's best to use heatshrink tubing or PVC sleeve to insulate the solder joints and prevent short circuits afterwards.

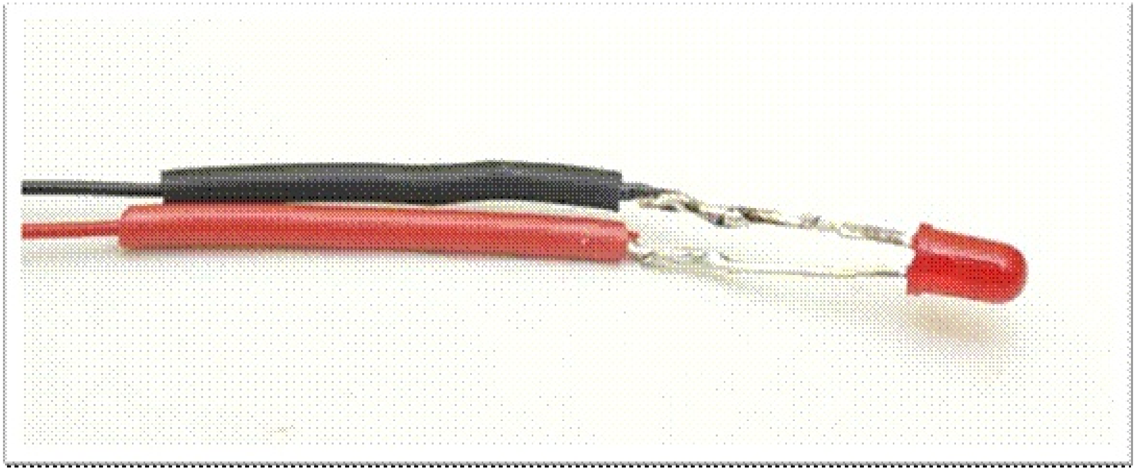

^^ *Multi-stranded wires can be soldered to the solid leads of an l.e.d.: insulate them afterwards. Different coloured wires identify the polarity of anode and cathode.*

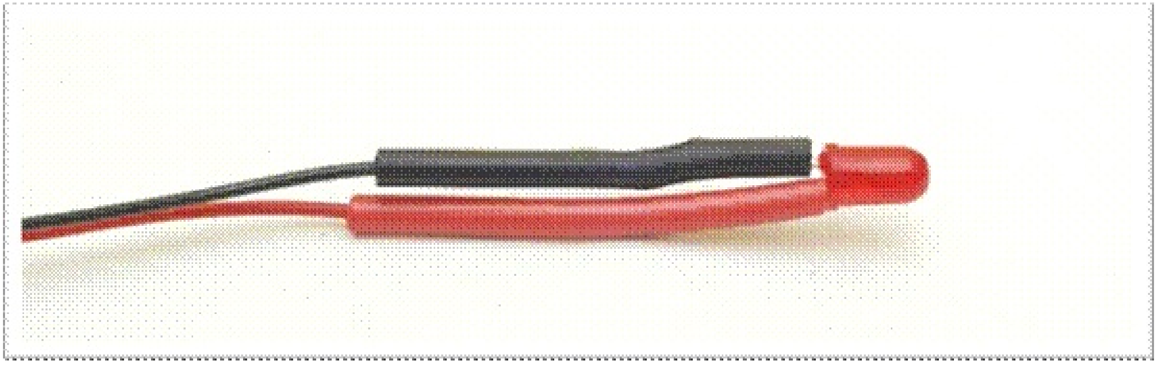

^^ *PVC sleeving or heatshrink is needed to prevent short circuits.*

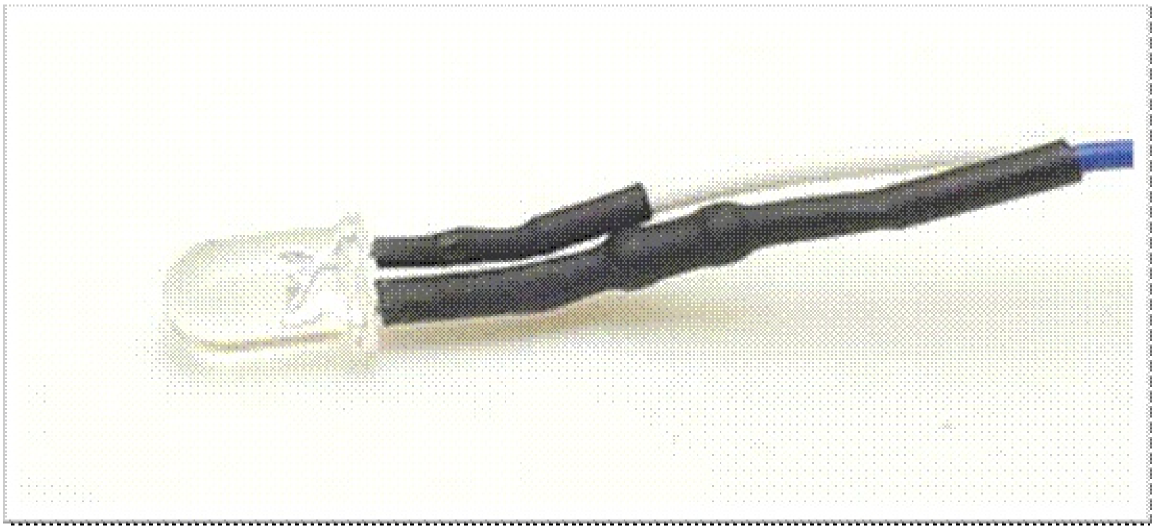

^^ A commercial l.e.d.– the series resistor can be made out in the sleeving.

Wires can also be soldered directly to printed circuit boards by stripping and tinning the ends, provided they fit through the holes in the p.c.b. It's a very common and cheap way of hooking up a board using "flying leads" and you'll see this all the time in consumer electronic equipment. For convenience so-called "solder pins" can be used, onto which flying leads may be soldered from the component-side instead.

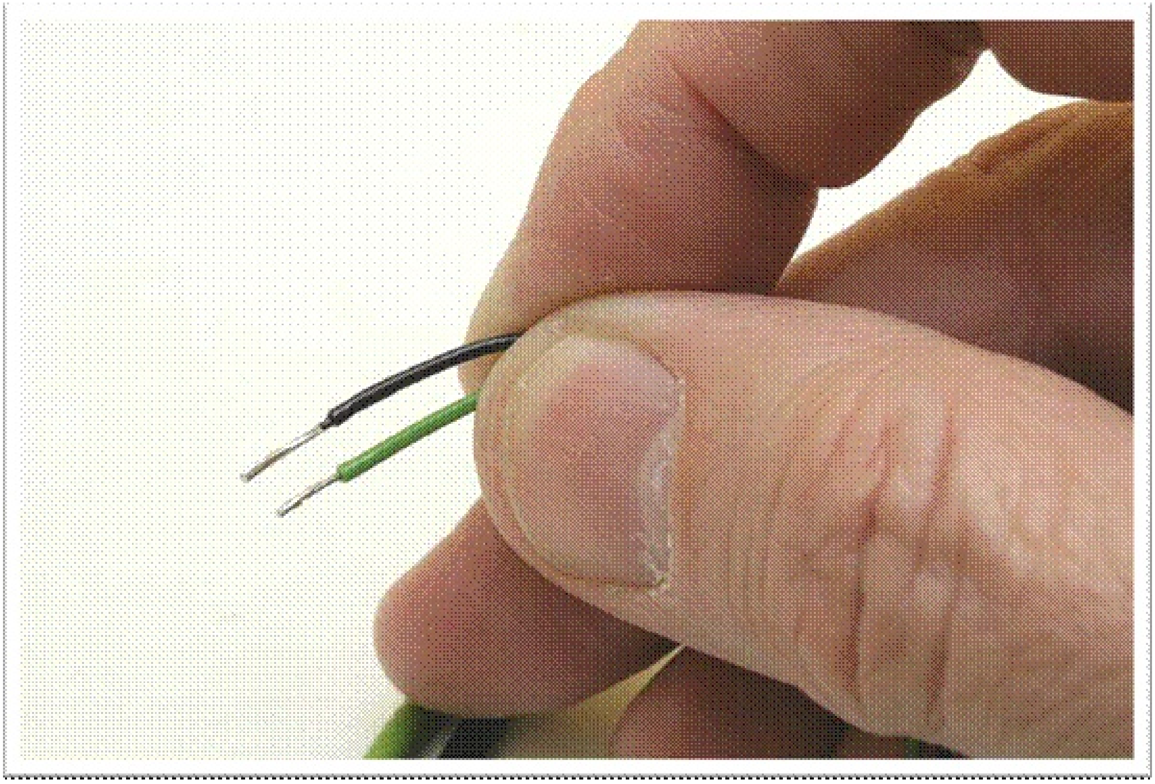

^^ A set of l.e.d.s connecting wires, stripped, tinned and ready to be soldered to the p.c.b.

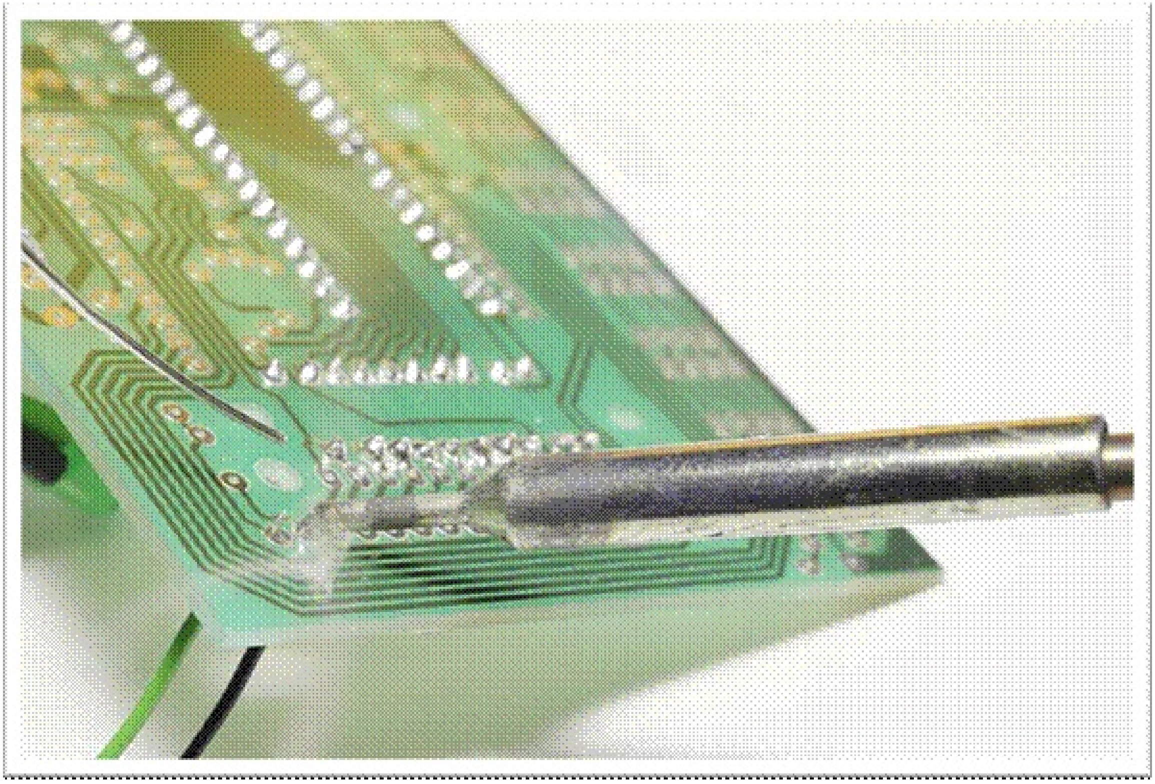

^^ Solder the two "flying leads" to the copper pads in the usual way. Hopefully the holes are big enough for the wires!

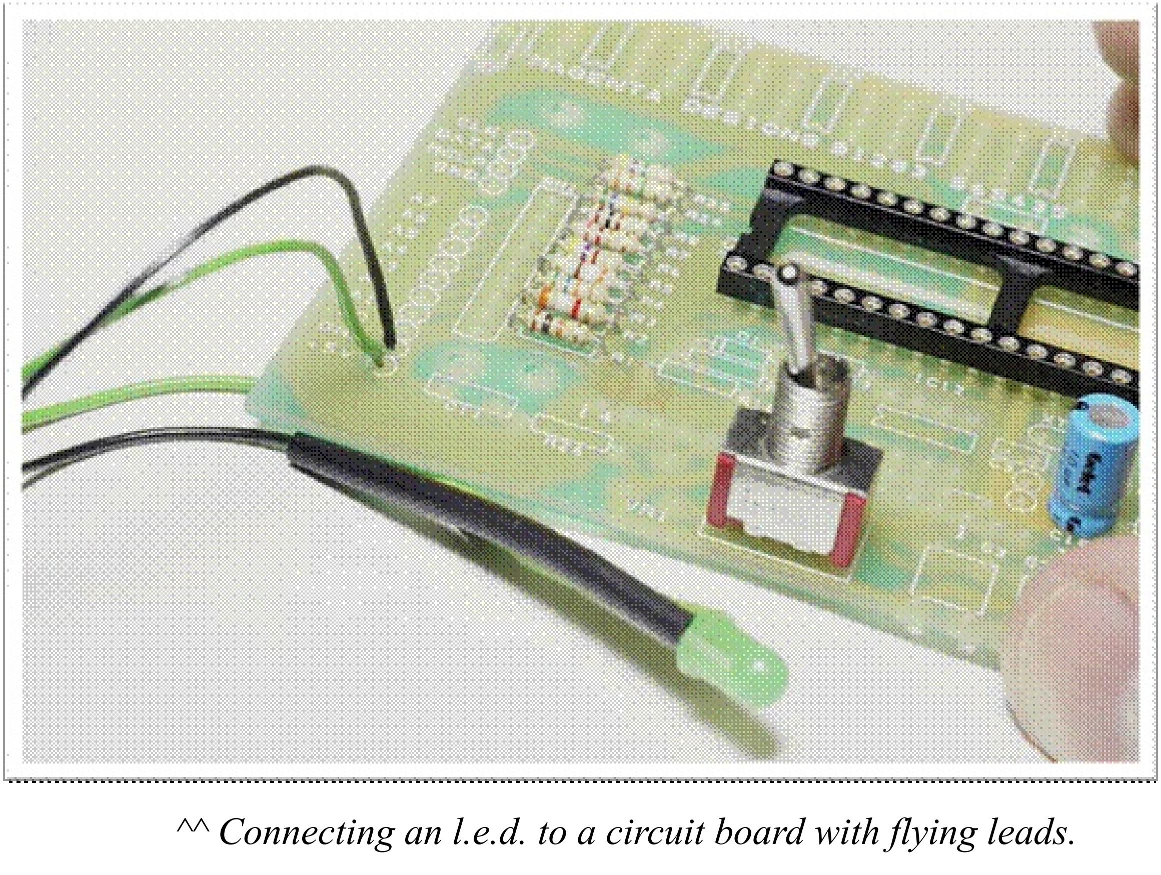

^^ Connecting an l.e.d. to a circuit board with flying leads.

Wires can also be attached to the underside (solder-side) of circuit boards, tacking them onto existing solder joints by re-melting them and absorbing the tinned wire end into it ("reflow soldering" – see later). This is a cheap and cheerful, semi-reliable way of rigging wires to a circuit board, and it's used all the time in imported consumer electronics.

Tidying up

I’ll leave it to readers to decide whether they should cut off excess wire before or after soldering. After the soldering is complete, I prefer to tidy up the joint by snipping any excess wire from the joint using a pair of “end cutters” shown earlier. These expensive hand tools have specially angled blades that snip the joint neatly down to the top of the solder joint. Ordinary side cutters will do fine.

It’s worth taking time out to inspect the work closely, looking for any missing solder joints, whiskers of solder or swarf shorting out any solder pads, and all such potential problem areas should be dealt with prior to testing the board. Faultfinding goes beyond the scope of this guide, but it’s true to say that almost always, any problems noticed after powering up the circuit are due to soldering faults or wrong components being used.

Reflow technique

Another technique often used is “reflow” soldering. This is used to “tack” devices or wires together, especially if very small, sensitive or fiddly parts are involved. There might be no room to make a “proper” sturdy joint, or it might just not be necessary to have any mechanical strength in the joint, especially if tiny parts are used.

As an example, imagine a small temperature sensor (I used a transistor) for use in a thermometer project. It could be quite tricky to solder flying leads onto the sensor’s leadouts, so a good approach is to tin both the flying leads and the sensor’s leads, and then simply touch them together and re-melt the solder with the iron. There’s no need to add any more solder, because the solder that’s already there will re-melt and the joint will be made. Sometimes this is called a butt joint.

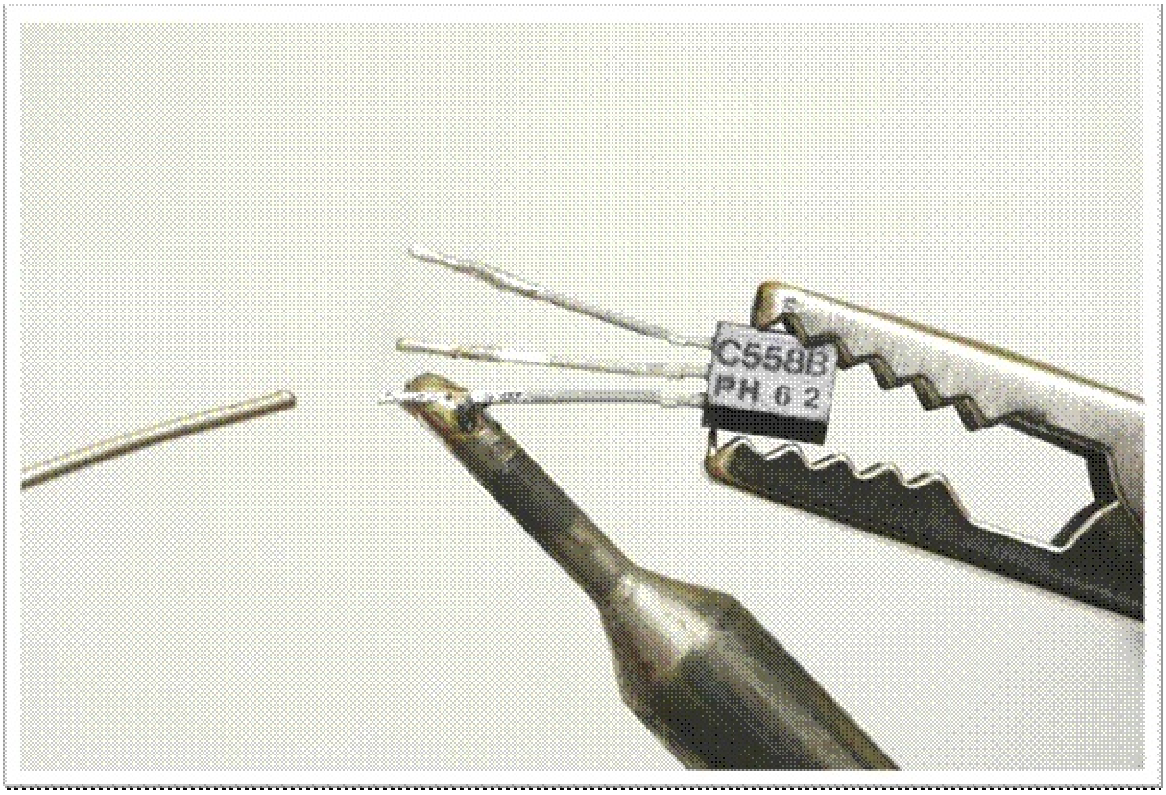

^^ *The leads of this sensor (transistor) have been tinned, ready for flying leads to be tacked onto them with a reflow method. A Helping Hands jig might help!*

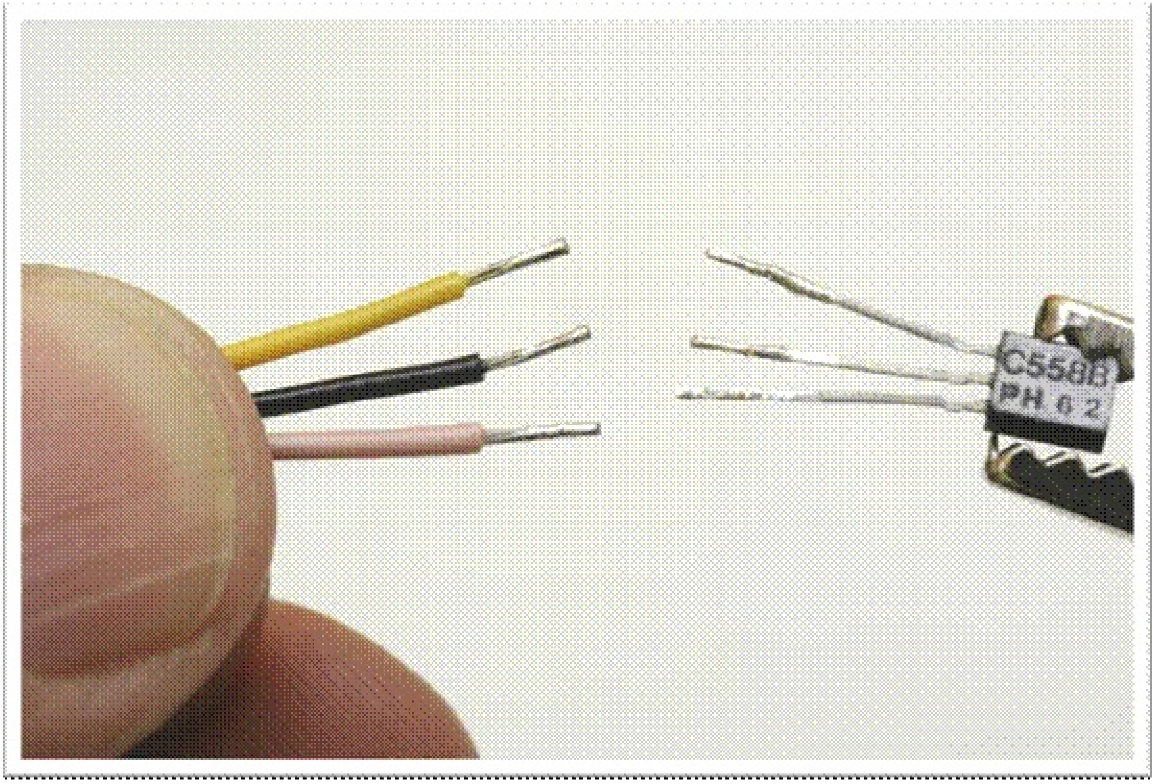

^^ The three flying leads stripped, tinned and ready to be reflow-soldered onto the device.

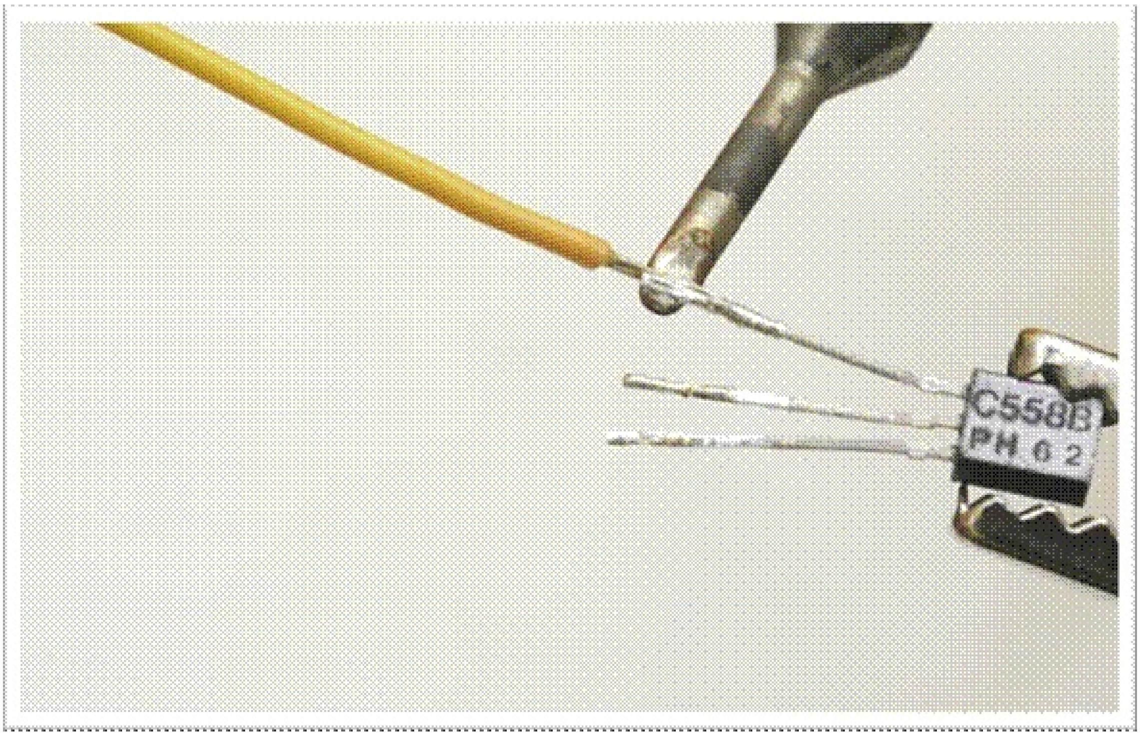

^^ To reflow solder them, simply hold the two leads together while re-melting the solder with the iron.

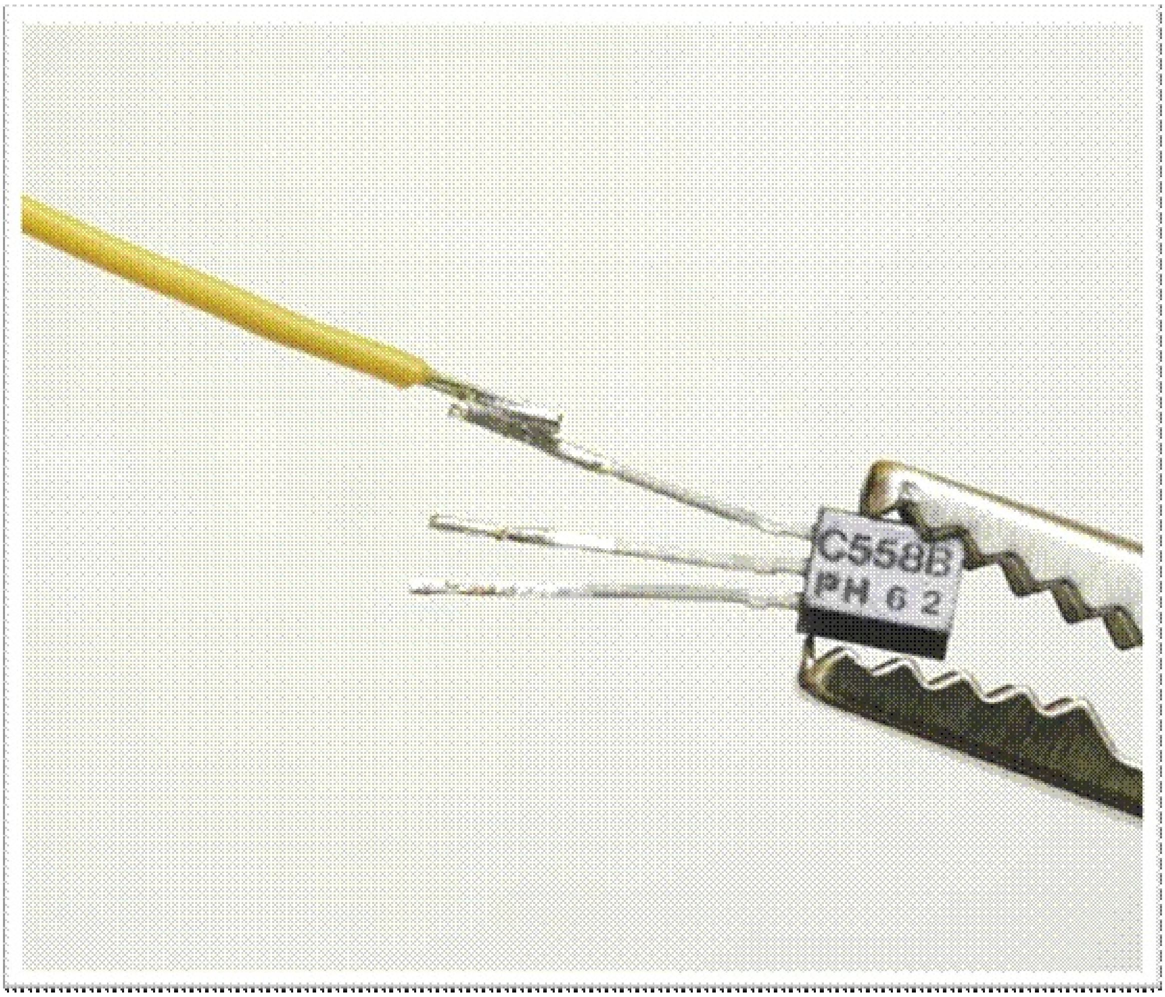

^^ Remove the iron and let the solder cool. The wire is tacked on.

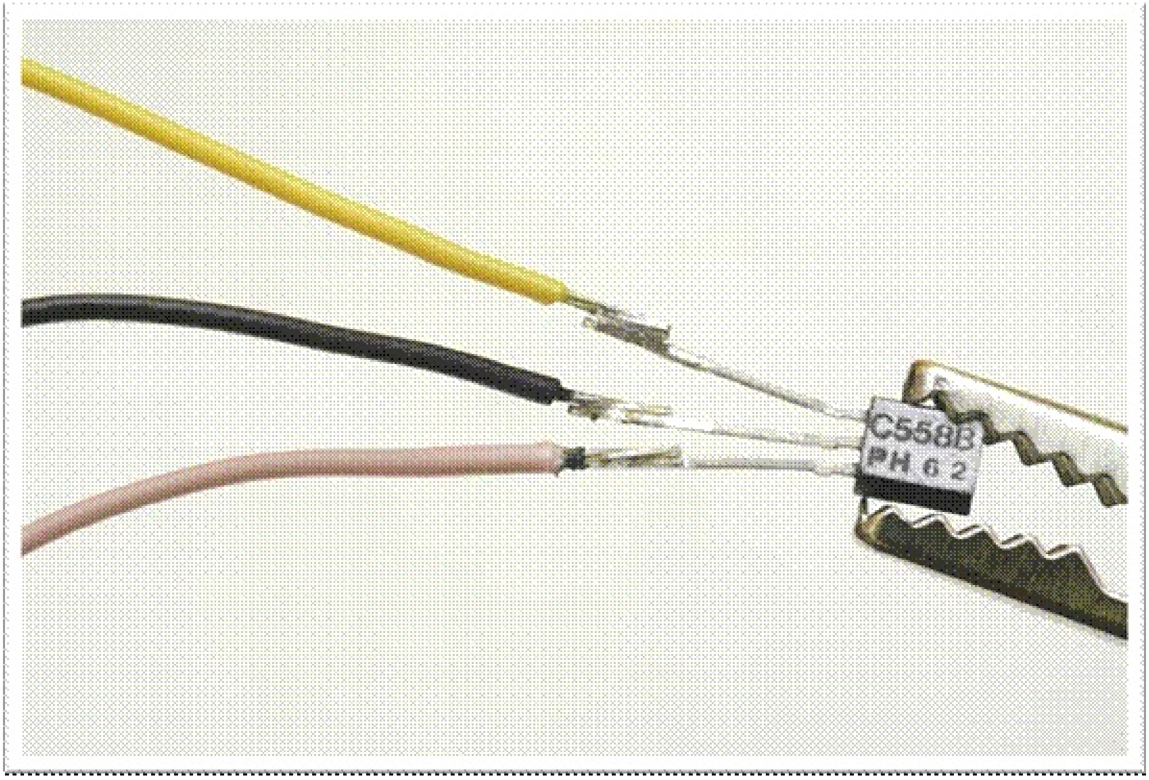

^^ Repeat for the other leads. They just need insulating with sleeving, then job done!

Depicted next is a typical "tag strip", an insulated panel with metal solder tags used for making sundry connections. (Entire TVs and radios used to be hand-built with them, in the early-mid 20th Century!) The principles of soldering are exactly the same, but more time is needed when applying the soldering iron because more metal is present, which needs more heat. You'll need more solder for bigger joints like these, so larger diameter solder makes a quicker job of it. Consider adding more flux (see earlier) to see if it helps.

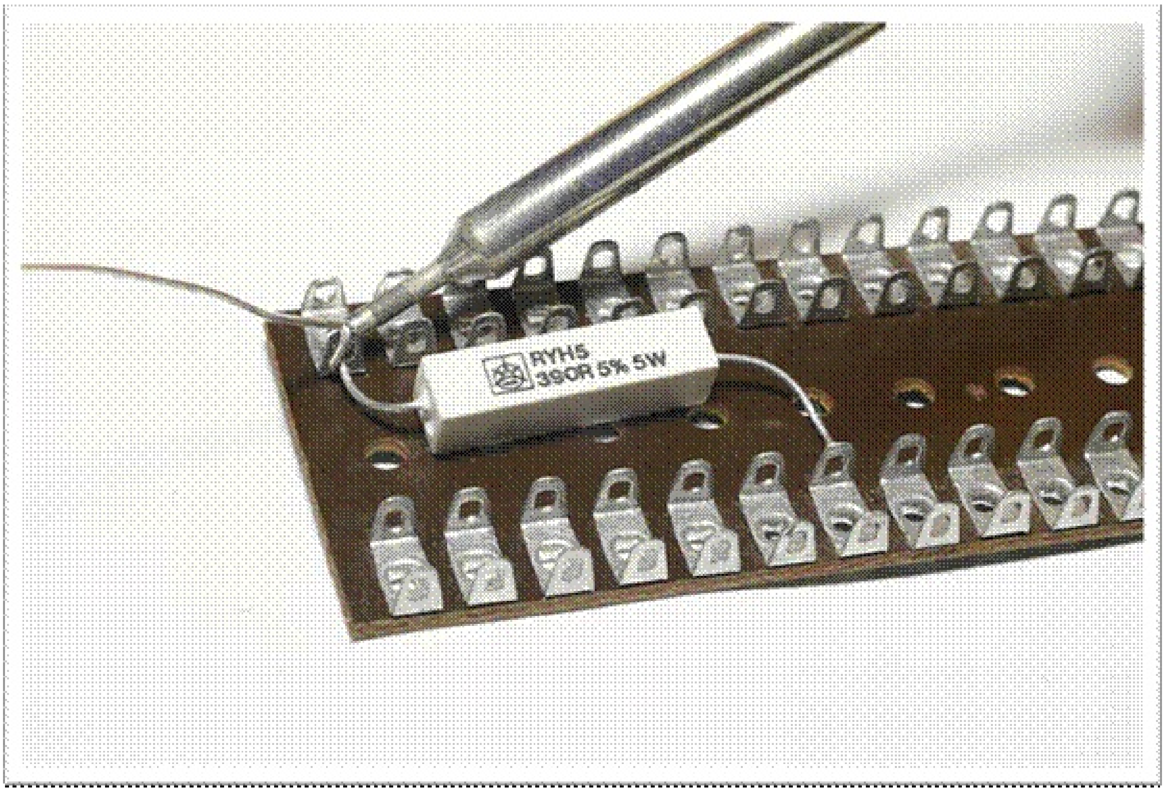

^^ Wrap wires through the terminals and arrange everything neatly, then solder as normal.

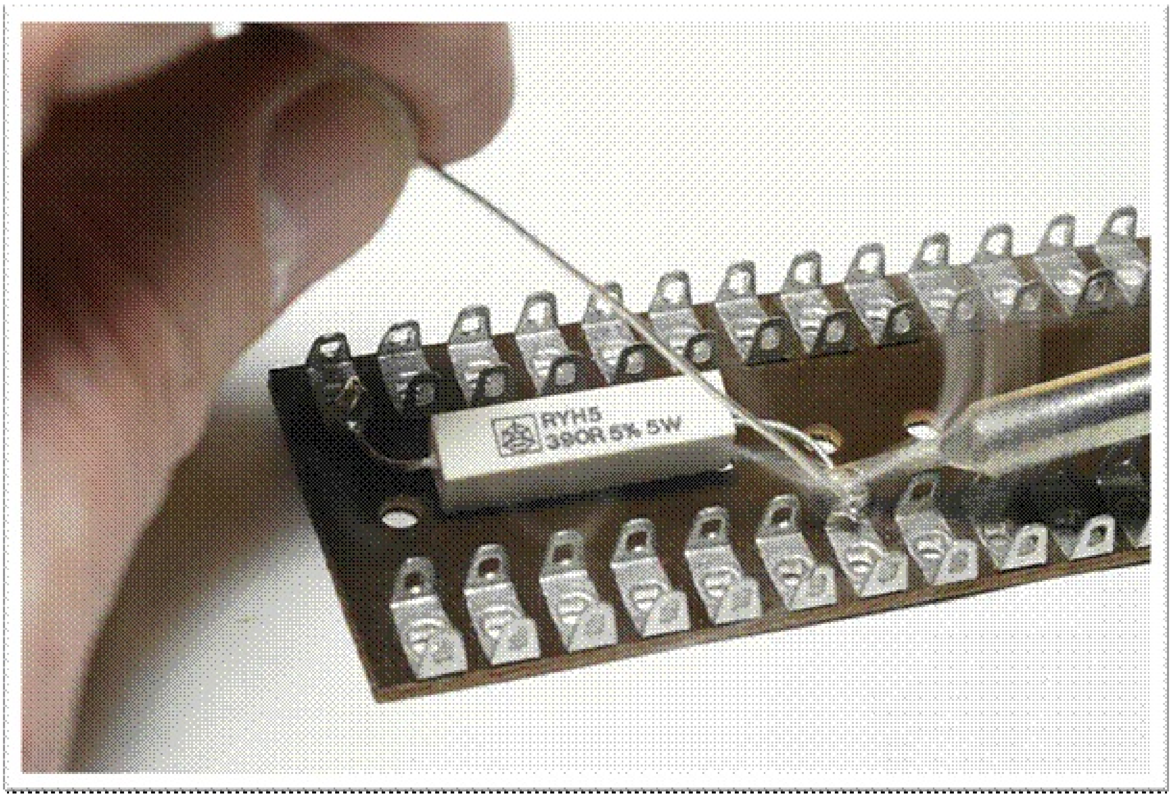

^^ Don't be afraid to apply more heat with larger assemblies like these: they contain more metal than, say, an ordinary p.c.b so allow more time for the solder to flow over

everything properly.

Fatigue & Breakage

Earlier I explained how multi-stranded hookup wire is flexible, which also makes it vibration proof – that's why a car's electrics are full of multi-stranded wiring, and test equipment probes use ultra-flexible multi-stranded wire for the same reason. Solid-core wire can be bent and will stay in shape (called "plate wiring") but if it's repeatedly bent or vibrated then it may eventually break somewhere due to fatigue.

The same is true of wires that have been soldered or tinned. No longer is the wire 100% multi-cored and flexible – instead it's been turned into a *single core wire* at the point where it's been soldered. This is potentially a weak spot and could eventually fracture due to fatigue, if subjected to continued vibration (e.g. in a car engine bay or in motorised equipment).

In a lot of equipment problems can be avoided by adding *strain reliefs* of some sort, to stop the wire being flexed where it's been soldered. Heatshrink tubing, or a dab of hot-melt glue, are ways of taking the pressure off the joint and ensuring soldered wires won't snap off due to vibration.

One of the reasons that higher-quality equipment and cars etc. use crimp terminals and connectors is that crimped (as opposed to soldered) connections retain all the flexibility of multi-core stranded wire from end to end, avoiding problems of wires breaking off. A *ferrule* is a very neat way of tidying wire ends and preventing stray strands of wire doing damage. Ferrules can be used to connect wires into screw terminal blocks etc. Simply clamp the terminal block down onto it and the screw will grip the ferrule.

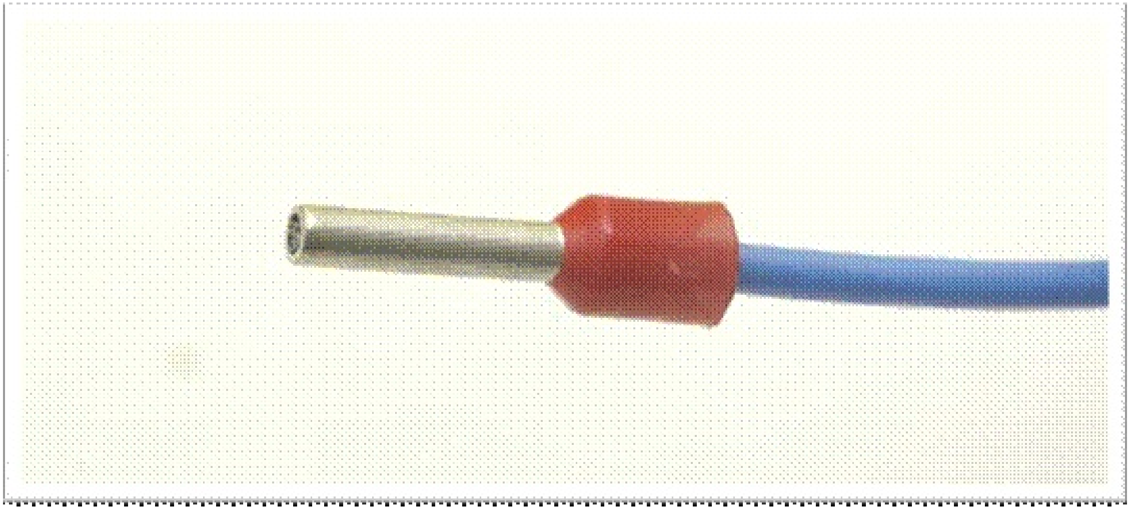

^^ Instead of soldering them, wires can be terminated with ferrules prior to fitting to e.g. screw terminal blocks. This makes them vibration-proof and also avoids any problem of stray wires poking out.

Faults & Desoldering techniques

By putting into practice the guidelines in my *Basic Soldering Guide* practice, there’s no reason at all why you should not obtain perfect results and eliminate any potential problems. Hopefully the information gives you plenty of guidance to tackle various soldering projects with confidence. There’s no substitute for getting some hands-on experience though, so I’d repeat the advice to try assembling a simple high quality electronic kit or two, such as those produced by Velleman and see how you get on. Powering up your first project successfully is a great thrill as every electronics hobbyist knows.

^^ *Popular desoldering products, a pump and various widths of desolder braid.*

Let’s now look at reversing the soldering procedure – what to do if things go wrong, or maybe you have to repair a circuit by replacing a faulty component.

A solder joint which is badly made is likely to be electrically “noisy”, unreliable and will probably worsen over time. Expansion and contraction of the joint due to heating and cooling can also throw up intermittent problems later down the line. The joint may look OK but underneath it may have a poor electrical connection, or could work initially and then cause the equipment to fail at a later date! These intermittent problems can be maddening to fix. TV repair technicians have an uncanny ability to go straight to a faulty solder joint because they see the same problem all the time, especially on equipment that has a “reputation”.

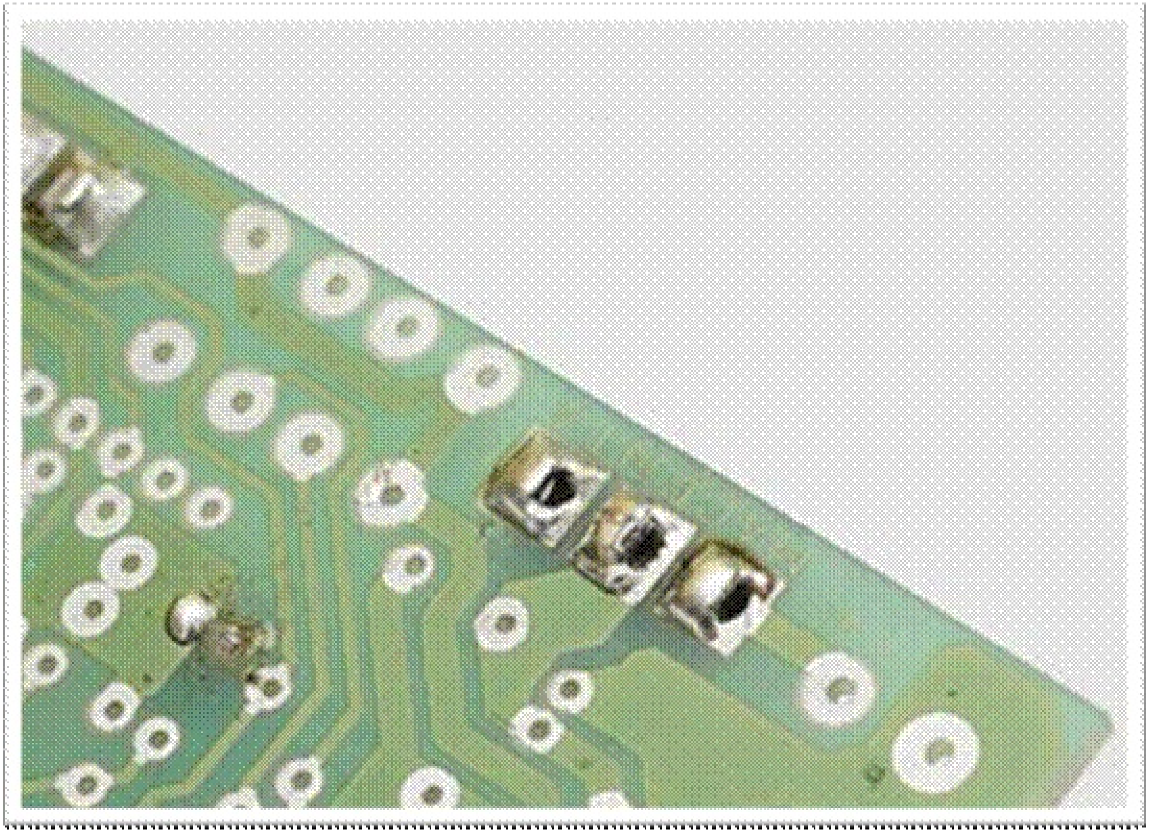

^^ These dry or grey/ gray joints were caused by dirty parts, poor heating and inadequate solder coverage

A solder joint that's poorly formed is called a "dry joint" or "cold joint" or a "grey/ gray" joint. Usually it results from inadequate heating, dirt or grease preventing the solder from melting onto the parts properly, and is often noticeable because the solder tends not to wet the surface properly. Instead it forms beads or globules. Alternatively, if it seems to take a very long time for the solder to wet the joint, that's another sign of contamination and that the joint may be a dry one, or the material is incompatible anyway. A matt, crystalline appearance instead of a shiny joint points to inadequate heating: the solder cooled down far too quickly and didn't flow properly.

Whether you want to replace a faulty component or fix a dry or poor-quality solder joint, it's usually necessary to remove the troublesome solder and then re-solder it afresh. Naturally, there are tools and techniques that make the job easy. It's very bad practice to simply re-melt the joint and then lash out with the board, whiplash style, hoping that the molten solder will be flicked off the board.

The usual way of removing solder from a joint is to use a *desoldering pump*. These work like a small spring-loaded bicycle pump, only in reverse! A plunger is pressed down until it locks into position. It's released by pressing a button which sucks air back through a pointed nozzle, carrying any molten solder with it. It may take one or two attempts to clean up a joint this way, but a small desoldering pump is an invaluable tool especially for p.c.b. work and they are widely available now.

^^ *A hobby desoldering pump primed for use*

Desoldering pumps often have a heatproof P.T.F.E. nozzle which may need replacing occasionally. Each time the button is pressed, a plunger clears the nozzle but sometimes solder particles and swarf will be ejected in the process; when you prime the pump, point the nozzle into a small pot or old aerosol top to catch any debris. Remove the spout and clean out the pump from time to time.

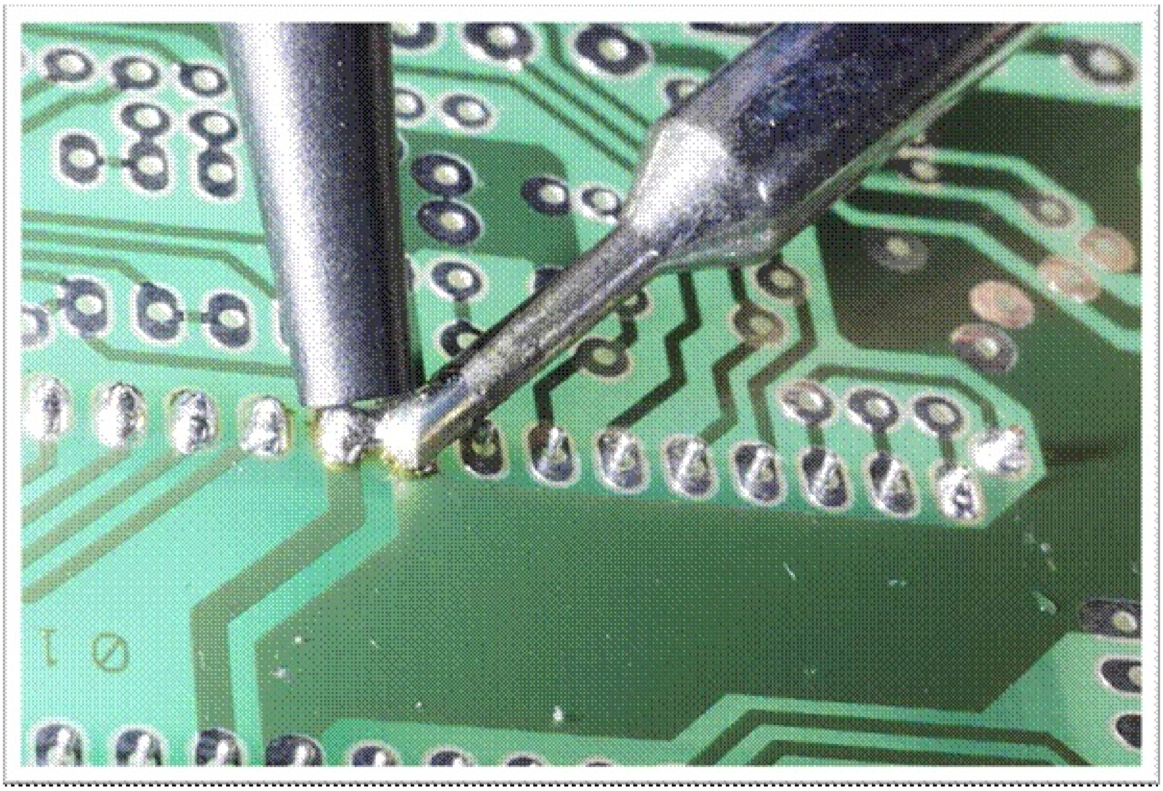

^^ *Suck up molten solder using a desoldering pump*

With very stubborn joints where the last traces of molten solder just can't be shifted, it sometimes helps to actually *add* more solder and then desolder the whole lot again with a pump.

Desolder Braid

An alternative to a pump is to use *desoldering braid* which arrives in small dispenser reels. It's a flux-impregnated fine copper braid which is applied to the molten joint, and the solder is then drawn up into the wick by capillary action. It's remarkably effective and for certain tasks, it can be more thorough than a pump. I recommend that a small reel is bought (start with 1.5mm width) for the toolbox, to tackle larger or difficult joints which would take several attempts with a pump.

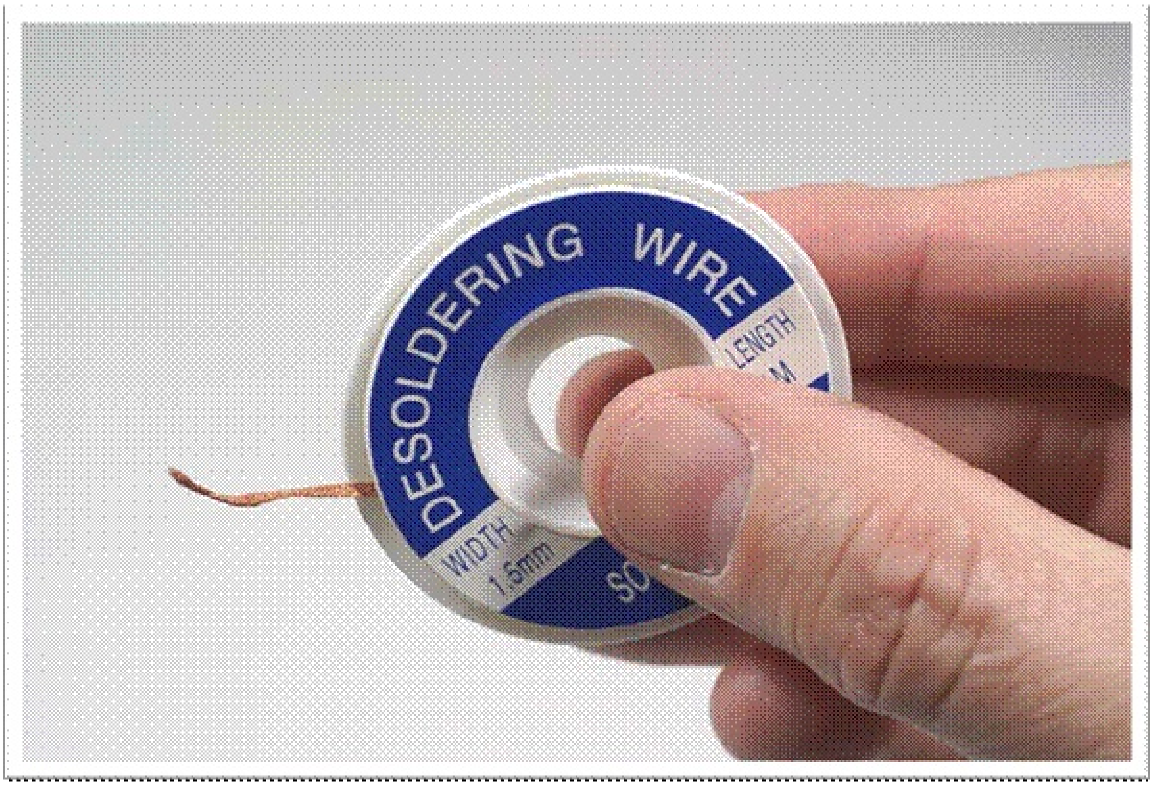

^^ *Desolder braid is also handy and can sometimes be more effective than a desolder pump. It comes in various widths to suit the scale of work being tackled.*

To use desolder braid, press the end of the braid down onto the joint using the tip of an iron, and let the solder melt underneath: the braid will then absorb the solder. The braid becomes hot so beware of burns. Once the solder's solidified on the braid, cut it off and discard.

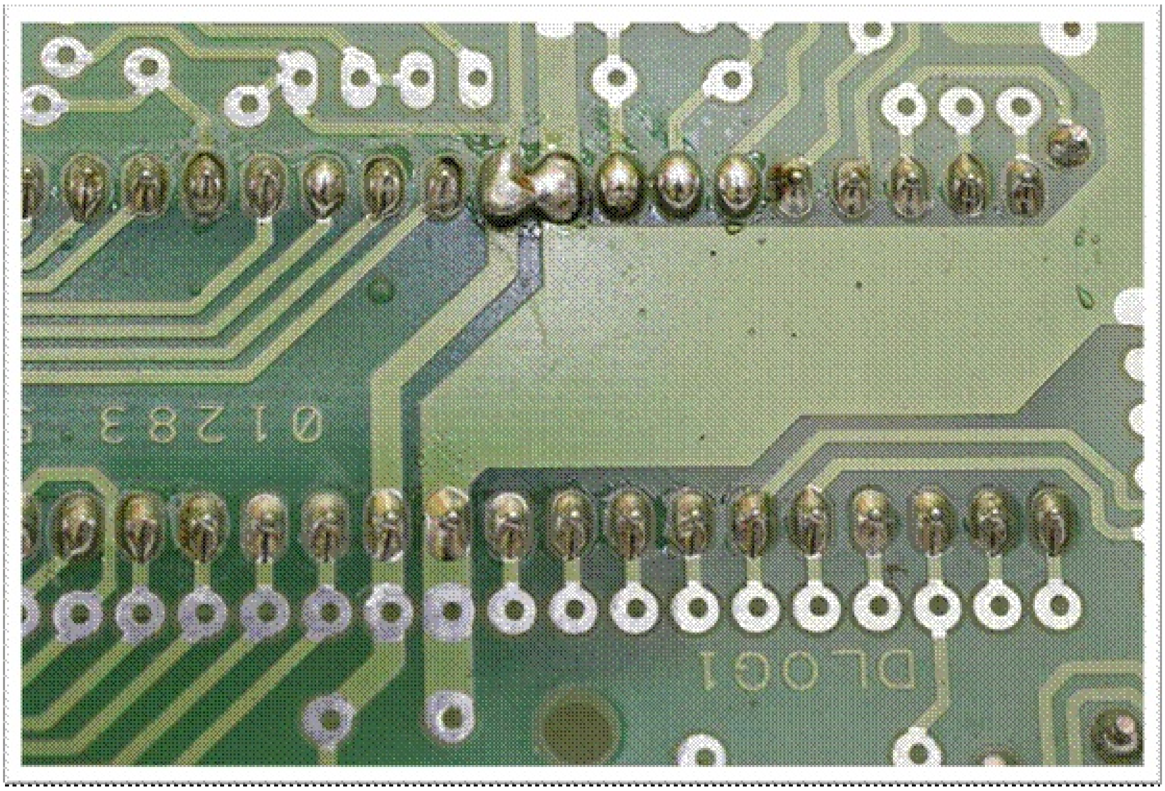
01283 5
DLOG1

^^ Desolder braid can also be used to remove excess solder, e.g. two i.c. pins shorted together.

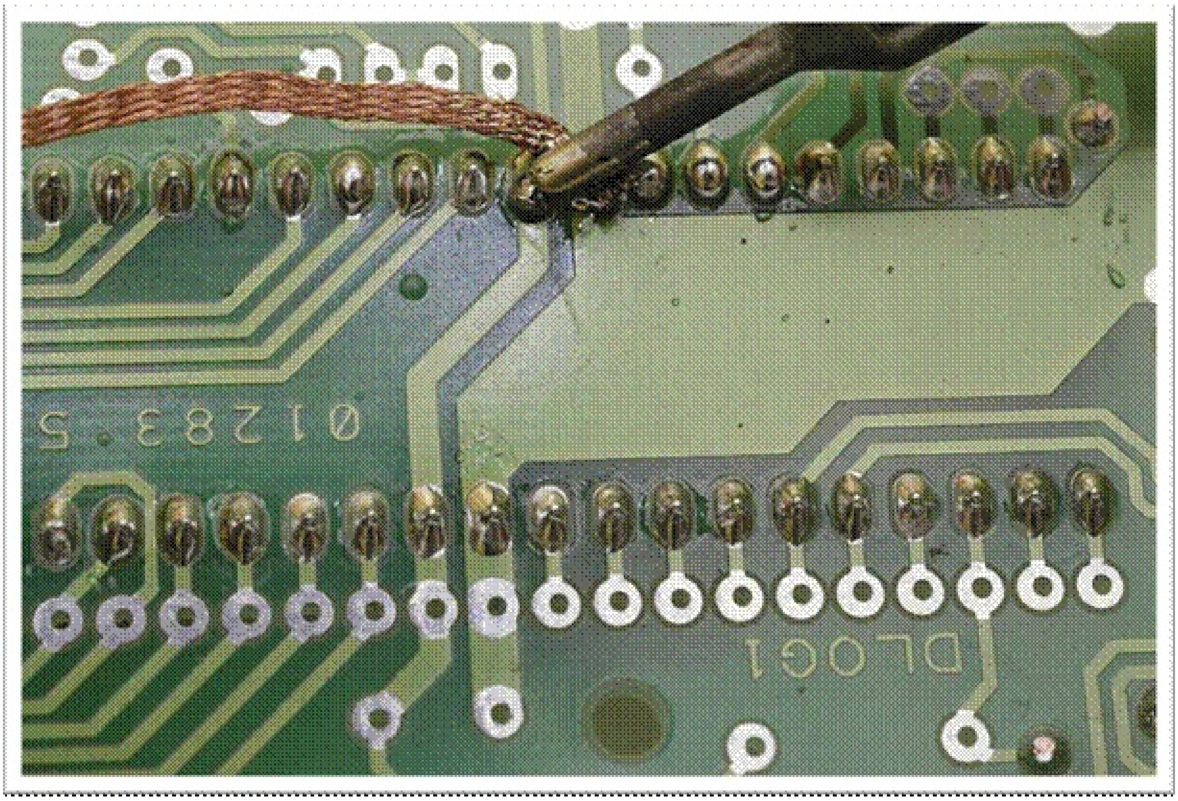
01283 5
DLOG1

^^ Press some desolder braid over the joint, then melt it with an iron to draw molten solder up the braid.

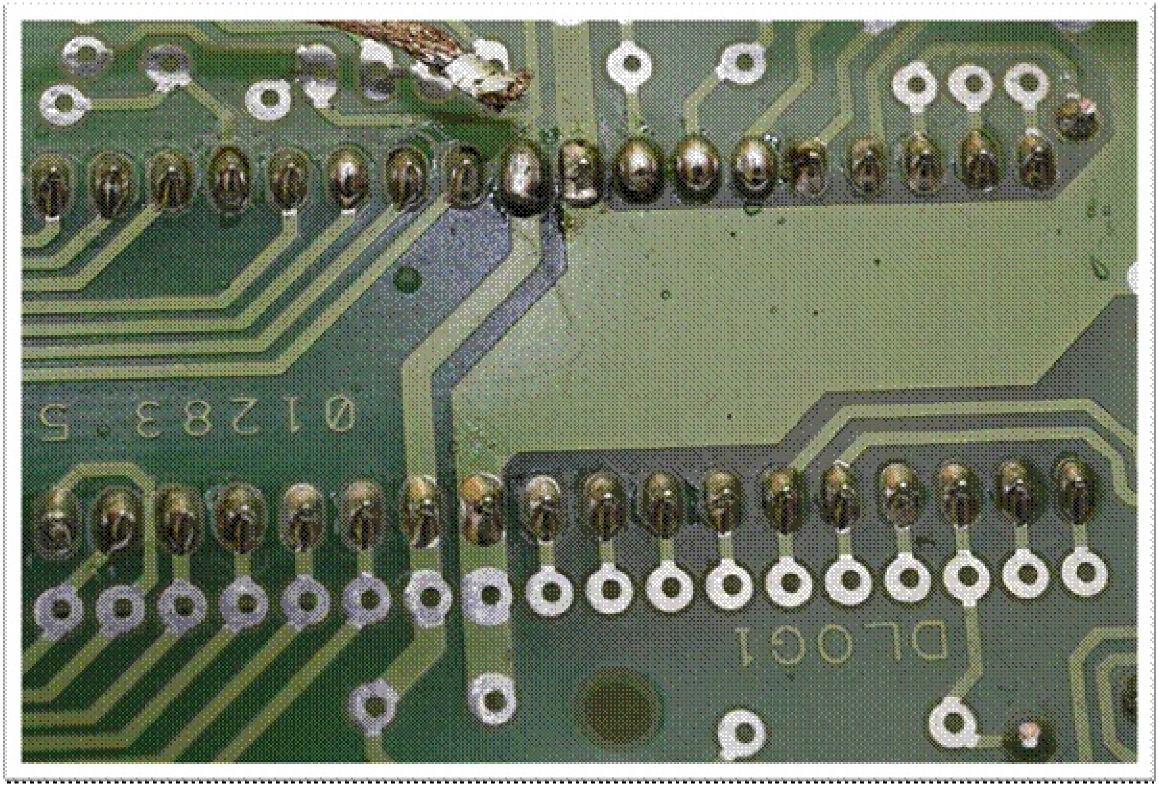

^^ Remove the braid immediately and don't drag "whiskers" of molten solder around. The excess solder is absorbed by the braid, which is snipped off ready for the next job.

Be aware that you can damage a printed circuit board accidentally when removing the desolder braid if it's not removed *quickly enough*. The solder will soon solidify, which effectively solders the braid to the printed circuit board! A careless tug may yank copper tracks or pads away from the board, stuck to the braid. You can also drag solder "whiskers" onto neighbouring pads unless the braid is removed cleanly. Why not practice using desolder braid with an old circuit board?

Whichever desoldering method you use, care is needed to ensure that the boards and parts are not damaged by excessive heat. It's not that difficult to apply so much heat when desoldering that the adhesive holding the copper foil tracks on the p.c.b. eventually fails, causing the copper track to lift away – everyone's worst nightmare.

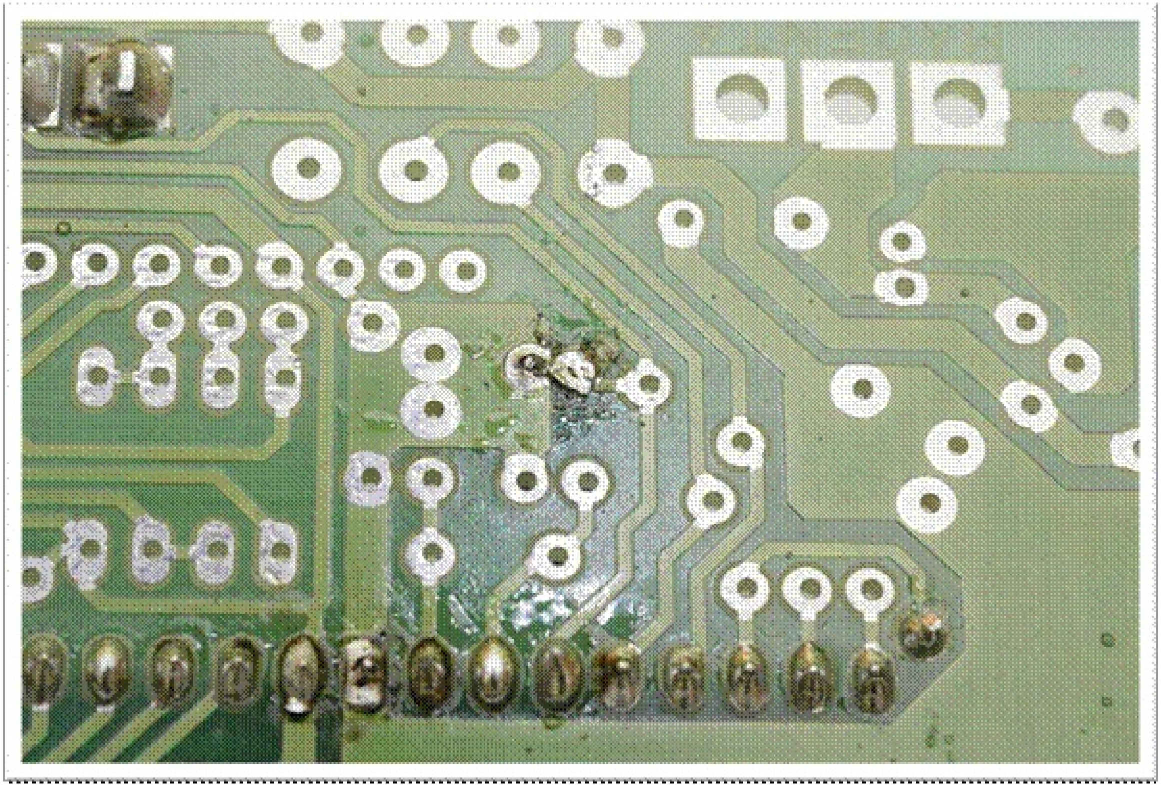

^^ Typical copper track damage (centre) caused by overheating during soldering or desoldering. The track has lifted off, but you can try repairing it by adding extra wiring or SuperGlue it if the track isn't broken.

If this should ever happen, remove the iron immediately and permit the area to cool (a freezer aerosol is valuable at such times). If you're lucky, you can maybe repair the lifted track using a droplet or two of Super Glue, or add "jumper wires" to bypass the damage.

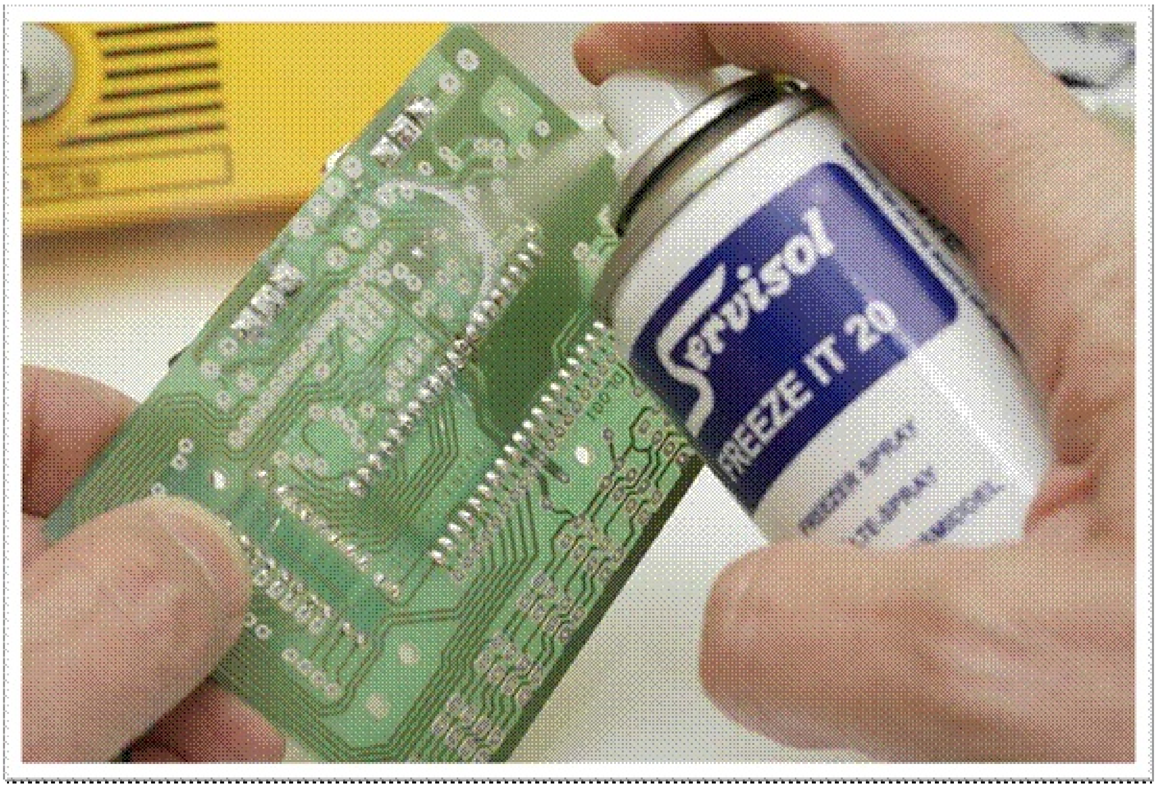

^^ A Freezer aerosol can give rapid cooling where excess heat has been applied

during soldering. Also used in circuit faultfinding to identify overheating parts.

You now know everything you need to know about making the ideal solder joint, and desoldering it in case you need to make a repair. Just to remind you, a Quick Summary guide follows.

Quick Summary Guide

To round off the Basic Soldering Guide, let’s summarise how to make the perfect solder joint.

- Ensure materials to be soldered are compatible with tin/ lead or lead-free solder.
- All parts must be clean and free from dirt and contaminants.
- Try to secure the workpiece firmly during soldering.
- Brand new soldering iron tips must be flooded with solder immediately, the first time they are used.
- Wipe the tip of the hot soldering iron on a damp cellulose sponge at frequent intervals. Then “tin” the iron tip by applying a small amount of solder.
- Aim to heat all parts of the joint with the iron for under a second or so, to bring them up to the same temperature.
- Continue heating and apply sufficient rosin-core tin/ lead or lead-free solder to form a complete joint.
- It only takes a second at most, to solder the average p.c.b. joint. It should be smooth and shiny, and through-hole joints should be slightly convex in shape.
- Remove the iron and return it safely to its stand.
- Do not move parts until the solder has cooled.
- Tin the soldering iron tip and clean it well, when switching it off, ready for next time.
- Consider using e.g. electronics flux dispenser pens or Colophony (rosin) to help with difficult joints.

Sometimes solder joints don’t go quite to plan, and sooner or later everyone is faced with the need to problem-solve or troubleshoot, so a simple Troubleshooting Guide follows next.

Troubleshooting Guide

This troubleshooting guide may help fix common problems encountered with troublesome solder joints.

SYMPTOMS	LIKELY CAUSES	REMEDY
Solder won't "take" (wet) and won't flow properly over the joint — molten solder forms beads or "ball bearings" instead of flowing properly.	Grease or contaminants present; Material may not be suitable for soldering with standard lead/tin or lead-free solder, e.g chromium.	Treat contaminated parts with abrasive cleaners etc. as required to expose base metal. Some metals can't be soldered with electronics-grade solder.
Solder doesn't melt or flow very well — the joint is crystalline or grainy-looking - a grey or dry joint.	Joint has been moved before being allowed to cool naturally, or: Joint was not heated adequately. Too large a joint – too much metal present – and/ or the soldering iron temperature or power rating are too low.	Desolder and remake. Apply heat for a longer period, or use a higher power soldering iron, or check the temperature setting and raise it if possible.
Solder joint forms a "spike" and applying the iron again makes it even worse!	Probably overheated, burning away the flux. The iron, when removed, would cause the solder to stand up in a spike.	It is usually best to desolder and remake the joint freshly again.
The copper foil of my p.c.b. has lifted off the circuit board!	Excessive use of heat has damaged the adhesive. Provided the track hasn't broken, it may be repairable.	You can sometimes repair it with Super Glue, or re-wire the board with jumper wires.
Brown varnish-like deposits are left behind after I finish soldering.	These are the remains of rosin flux and are nothing to worry about.	It can be removed with PCB cleaners or some solvents, if you want to tidy up the board and inspect your work.

Possible Hazards and simple First Aid

It’s extremely rare that soldering iron operators receive any burns or other injuries from the use of hot soldering irons. Soldering is perfectly safe provided that common sense precautions are taken during the soldering operation. Here are some of them:

- Components are very hot after soldering, so let them cool before handling them to avoid skin burns.
- Beware of splashes of molten solder caused by careless handling of a hot soldering iron.
- Beware of energised components (capacitors, batteries etc.) being shorted by molten solder and ejecting solder splashes due to arcing.
- Always park a hot iron safely on a stand in between use — never hang it vertically next to the bench.
- Keep a hot soldering iron away from its mains cable (silicone cables reduce the risk of accidental damage).
- Beware of wire offcuts flying off (danger to eyesight) when snipping wires to length before or after soldering.
- Avoid inhalation of solder and flux fumes as this can irritate the respiratory tracts, especially in sensitive cases (e.g. asthma).

Should you receive a more serious skin burn which requires attention, then:

- **Cool the affected area immediately.** Use plenty of cooler running water – but avoid ice cubes etc. as they can cause nerve damage after a time or inhibit the flow of blood to the affected area.
- Remove any **objects** which may be constrictive, before any swelling starts (rings, watches, bracelets).
- Do not prick blisters nor apply ointments, salves or lotions at this stage.
- Local pain relief for small burns can be obtained by spraying *Burneze* aerosol onto unbroken skin.
- Seek medical attention for more serious burns.

Eyesight problems are exceptionally rare, e.g. pieces of wire offcuts or solder splashes lodging in the eye area, and should be treated by a qualified first-aider or A&E. The best you can do is bathe the affected area with e.g. a first-aid eyewash bottle or fresh water. Then seek professional medical help straight away.

Useful resources

Details of the Antex range of soldering equipment, solder tips and spare parts can be obtained from www.antex.co.uk. UK and international distributors are also listed on the Antex (Electronics) web site.

If you’re interested in hobby electronics then why not try **EPE Magazine**, the No. 1 UK magazine for hobby electronics enthusiasts, students, trainees and technicians around the world. You can buy a printed edition from newsstands, download a PDF version (for Windows) or try a tablet/ smartphone version from Pocketmags. More details at www.epemag.com

You can learn more about the writer at his website http://www.alanwinstanley.com/

Some mail order suppliers of soldering equipment and electronic parts include:

- ESR Electronic Components Ltd. (soldering equipment, Velleman kits)
- Bowood Electronics (Antex irons and spares, electronic parts)
- Rapid Electronics (components, tools, equipment)
- Cricklewood Electronics (CCTV, Antex soldering irons and spares, components)
- Maplin Electronics (UK electronics retailer and mail order)
- Farnell Electronic Components (major UK industrial supplier)
- RS (UK electronics industrial supplier)
- Brewsters Ltd (soldering equipment mail order specialists)
- Velleman UK (electronic kits)
- Quasar Electronics (Velleman kit mail order retailers)
- Kemo Kits (Germany, trade only)
- Hobbytronics (UK hobby mail order supplier)
- Multicore Solders (now a Henkel brand)

These links were correct at the time of going to press, July 2013.

Acknowledgements

Antex (Electronics) UK generously provided samples of their ever popular British-made soldering equipment, materials and accessories for use in the Basic Soldering Guide. When starting out as a 1970's teenage schoolboy electronics hobbyist, I and my trusty Antex soldering iron shared many adventures in electronics together, and it has been a pleasure to use the latest Antex soldering irons throughout this guide.

Also I would like to thank Brian Brooks of Magenta Electronics Ltd. (www.magenta2000.co.uk) for supplying me with a variety of their professionally-designed printed circuit boards and a range of components which were the subjects used in the photographs.

Photography

All photographs were taken by the author using Sony Alpha DSLRs and 50mm macro lens with extension tubes, and a Sony macro flash. High-resolution versions of these and more technical photos are available for commercial or educational royalty-free use. Please contact me with any enquiries.

Conclusion

I really hope the Basic Soldering Guide will give you the confidence to try your hand at electronic soldering. It's really a lot easier than it sounds, and armed with the guide's advice and photographs, the next thing to do is invest in a decent-quality soldering iron such as the excellent range manufactured by Antex that will serve you well for years to come.

Soldering an electronics kit together such as the professional designs produced by Velleman is a great way of testing out your new skills. Start with a small, simple kit costing a few pounds to gain confidence and experience, and avoid the temptation to tackle something too complex until you're ready to extend yourself further.

Feedback is welcomed by email to alan@epemag.demon.co.uk

You can learn more about me at www.alanwinstanley.com or read my column in EPE Magazine.

Good luck with your soldering!

Alan Winstanley

Text and Photographs © Copyright © Alan Winstanley 1997-2013

Play fair and support writers who share their skills and expertise

Preparing a resource like this takes months and is based on four decades of practical experience which I gladly share with readers. It costs a small fortune in time, experience, IT and photographic equipment to put an ebook like this together, and the Royalties that you pay to buy the book puts food directly on my table. You paid for your copy (thank you!) so if you see any unauthorised versions of my work anywhere in circulation, I’d really appreciate being told about it. Thank you! A.R.W.

Interested in electronics?
Read EPE - the No. 1 UK hobby electronics magazine.
Ask your newsagent, download or subscribe now!
www.epemag.com

BSG/ ARW/V1.1/0813

Project Book

MAKE A
WEARABLE
SOLDER SKILL
BADGE

Learn to Solder

Tools and Techniques for Assembling Electronics

Brian Jepson,
Tyler Moskowite
& Gregory Hayes

O'REILLY®

Maker
PRESS™

Project Book

Learn to Solder

Learn the fundamentals of soldering—and pick up an essential skill for building electronic gadgets. You'll discover how to preheat and tin your iron, make a good solder joint, desolder cleanly (when things don't quite go right), and how to use helping hands to hold components in place.

This concise book is part of MAKE's Getting Started with Soldering Kit. Using the tools in the kit and some electronic components, you can practice soldering while making fun blinky objects. Then show the world you just learned a new skill by wearing the Learn to Solder Skill Badge.

- **Learn how to prepare your workspace**
- **Get to know the components you'll work with**
- **Use the best methods for soldering components in place**
- **Experience the perfect solder joint**
- **Know how to desolder when things don't go right**

Heat up the iron and start soldering today!

US $3.99 CAN $4.99
ISBN: 978-1-449-33724-7
50399
9 781449 337247

O'REILLY®
Twitter: @oreillymedia
facebook.com/oreilly
oreilly.com

Maker PRESS™
makerpress.com

Learn to Solder

Brian Jepson, Tyler Moskowite, and Gregory Hayes

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Learn to Solder

by Brian Jepson, Tyler Moskowite, and Gregory Hayes

Copyright © 2012 O'Reilly Media. All rights reserved.
Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editors: Shawn Wallace and Brian Jepson
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Gregory Hayes

May 2012: First Edition.

Revision History for the First Edition:
April 27, 2012 First release
See *http://oreilly.com/catalog/errata.csp?isbn=9781449337247* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Learn to Solder* and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-33724-7
[LSI]
1335542415

Contents

Welcome v

1/Getting the Workspace Ready 1
- 1: Attach the Tip 2
- 2: Secure the Tip 2
- 3: Wet the Soldering Station Sponge 4
- 4: Set the Station’s Temperature 4
- 5: Wipe the Iron’s Tip 6
- 6: Tin the Iron’s Tip 6
- Go Make Your Learn to Solder Skill Badge 7

2/How to Solder 9
- Working with Solder 9
- Keeping the Circuit Board from Moving 10
- Tinning Solder Pads 10
- Placing a Component in the Board 11
- Knowing Which Way a Component Goes In 13
- Stabilizing and Straightening Components 14
- Safely Soldering Sensitive Components 15
- Soldering a Component in Place 16
- Trimming Your Leads 18
- Going From Solderless Breadboard to PCB 20
- Soldering Jumper Wire 22
- Bridging Joints with Solder 22
- Relieving Strain on Cables 24
- Getting the Perfect Solder Joint 25

3/Desoldering 29
- The Desoldering Wick 29
- Preparing the Wick 30
- Desoldering with the Wick 31
- Desoldering with the Solder Sucker 32

4/What’s Next? 33
- Teach the World to Solder 33
- Make More Things 34
  - Supercap Racer Kit 34

Wee Blinky Kit **35**
MintyBoost **36**

Welcome

Welcome to MAKE’s Getting Started with Soldering Kit, and congratulations on taking a big step into the world of DIY electronics. Once you get the hang of soldering, you can put together some of the many great kits that are available, fix electronics that are broken, and build inventions of your own. With this kit, you’ll learn how to:

- Prepare and clean your soldering iron
- Assemble electronic circuits from kits
- Transfer circuit designs from a solderless breadboard to a prototyping PCB
- Correct soldering mistakes you’ve made

Basic Tools

There’s a lot of great stuff in the box, and before you start using it, here’s a tour of what you’ll find in there. Figure 0-1 shows the soldering tools you’ll be using most of the time.

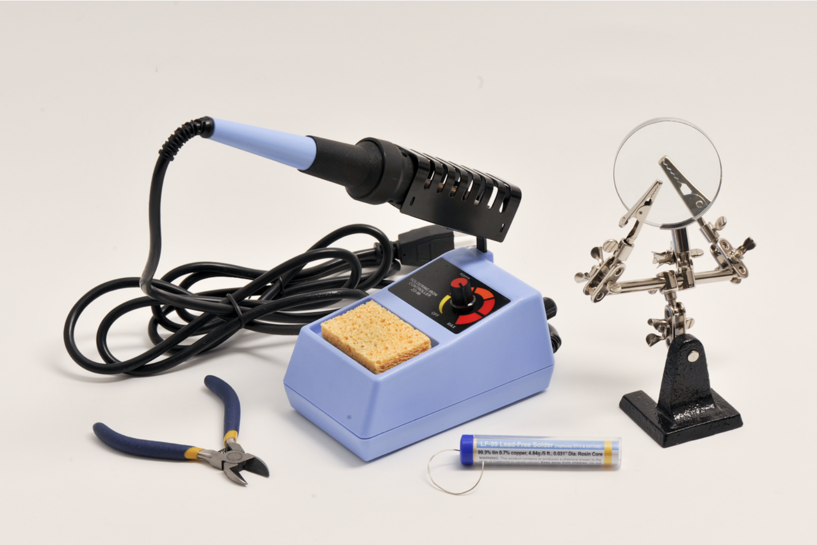

Figure 0-1. *Basic soldering tools*

Deluxe soldering iron
: (Top left) This soldering station includes a variable temperature controller, a cleaning sponge, soldering iron, and a ringed holder.

Crosscut pliers
: (Left) You'll use these to trim away excess leads after you solder components in place.

Solder tube
: (Bottom) This is enough solder to get you started and keep you busy over many projects.

Helping hands
: (Right) For those times when you need a third or fourth hand, the helping hands let you hold items steady while you solder.

Advanced Tools

Most of the time, you'll only need the basic tools to get things done. But when you need to replace your soldering iron tip, correct a mistake you made while soldering, or need a tool to help keep components from overheating, you'll need the items shown in Figure 0-2.

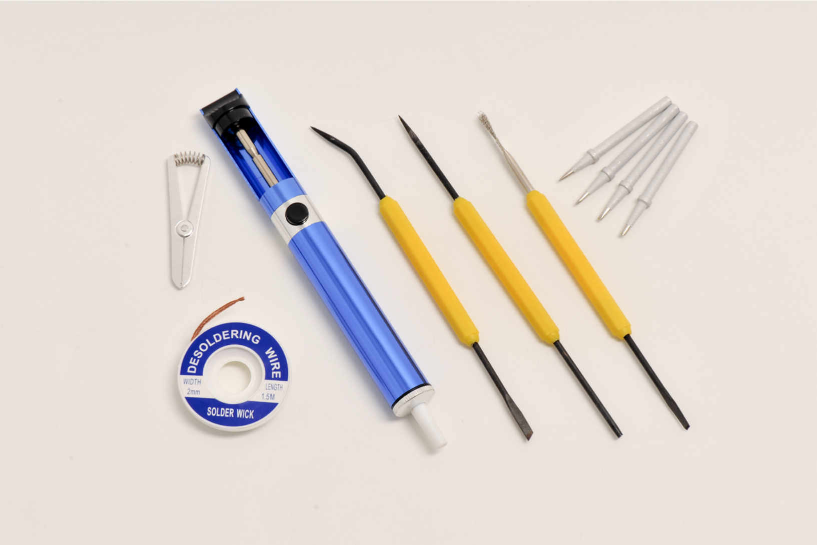

Figure 0-2. *Advanced soldering tools*

Heat sink

(Left) Clip this to sensitive components to help dissipate heat.

Desoldering wick

(Bottom left) Use this to wick away excess molten solder.

Desolder pump

(Center left) This pump will suck up molten solder when you have a lot of solder to remove.

Soldering tools

(Center right) The scrapers, brush, and slotted probe come in handy when you need to move solder around or precisely position a component.

Replacement tips

(Right) Tips don't last forever. When you've worn out your tip, use one of these as a replacement.

Project 1: Learn to Solder Skill Badge

Our Learn to Solder Skill Badge Kit (the 2011 model is shown in Figure 0-3) has been used to teach thousands of people of all ages how to solder at Maker Faires across the country. It's a simple, fun way to learn how to solder and also how to teach others to solder. After you build the one included in the box, you can order more from *makershed.com* and teach others how to solder.

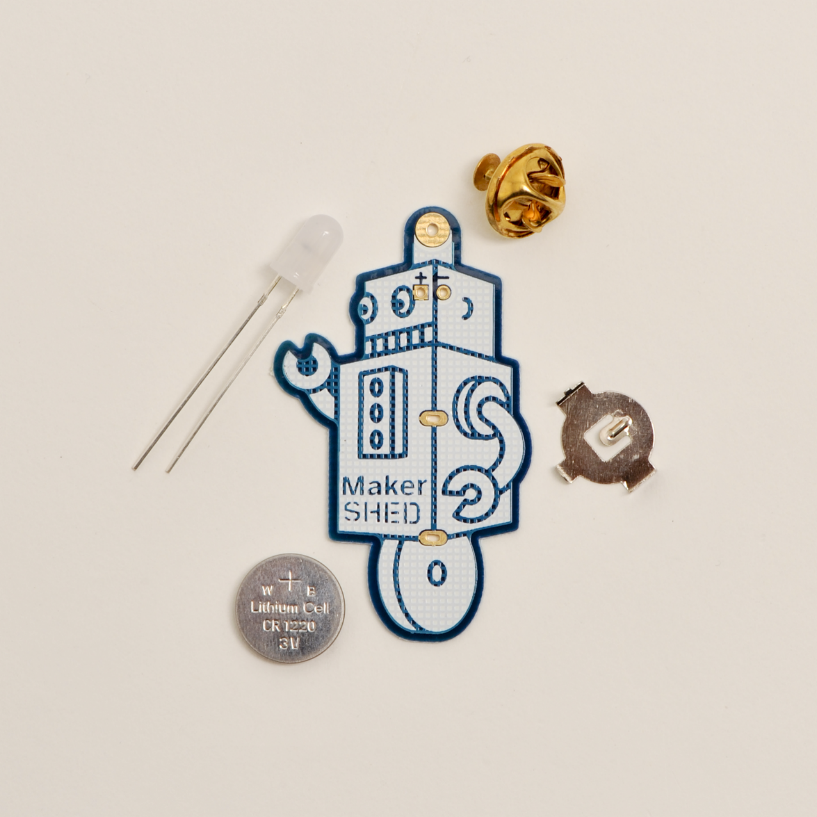

Figure 0-3. *Learn to Solder Skill Badge, 2011 model*

Blinking or Color Change LED

(Top left) This is an LED (Light-Emitting Diode) with a twist. Normal LEDs give off a single color, and keep shining as long as you give them power. Also, normal LEDs require a specific voltage to operate: don't give it enough power, it won't light at all; give it even a little too much, and you could burn it out.

This LED is different in both respects: it has three elements (red, green, and blue) that are under the control of a small integrated circuit (IC) embedded within the LED. The IC causes the elements to change color in a repeating pattern. Because the IC controls the voltage that it delivers to the individual color elements, it's more tolerant of variations in the voltage you give it. For example, even though red LED color elements typically operate at around 2 volts, you're able to use a 3 volt battery with the pin.

The 2012 model uses two self-blinking LEDs instead of one self-color changing RGB LED.

CR1220 battery

(Lower left) This "coin cell" battery supplies power to the pin.

Printed Circuit Board (PCB)

(Center) The 2011 Learn to Solder Skill Badge features a friendly robot. The 2012 model will also have a robot, every bit as friendly.

Pin and clutch

(Top right) This is what holds the pin to your clothing.

Battery holder

This keeps the battery on the PCB.

Project 2: 555 Timer Blinky

The 555 timer blinky is a simple circuit that makes two LEDs flash in an alternating pattern. When one is on, the other is off. Figure 0-4 shows the components for this project, and Figure 0-5 shows the finished project. This is a slightly more challenging project than the skill badge.

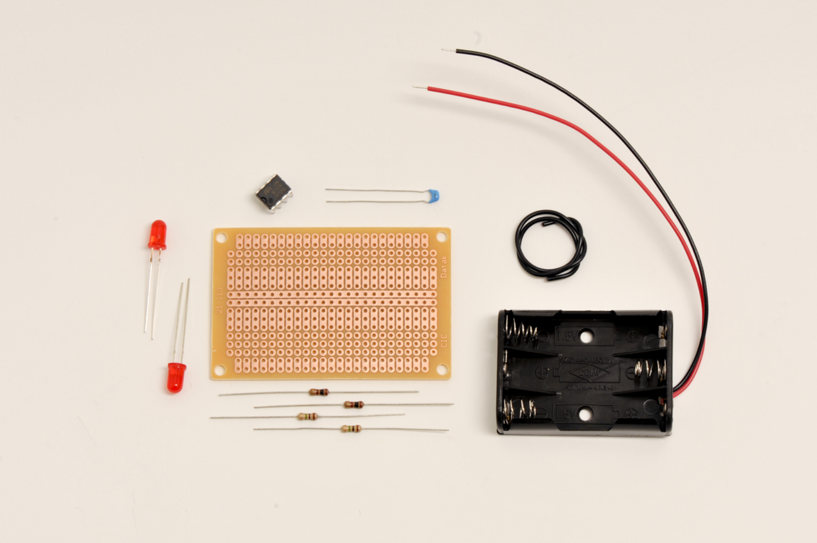

Figure 0-4. *555 blinky parts*

LEDs

(Left) Unlike the LEDs used in the skill badge, these are your normal, run of the mill LEDs.

Protoboard

(Center) Look back at the skill badge PCB. Notice how it was custom designed for a single purpose. Now look at this protoboard PCB: it's got a lot of different holes in convenient locations. You can build many kinds of simple electronics projects on a board like this.

555 Timer

(Top left) The 555 timer is a special type of integrated circuit that can turn electrical current on and off in a repeating pattern. It's just what you need to blink an LED.

Capacitor

(Top right) To customize the 555 timer, you need to connect certain components to it. The value (capacitance or resistance) of the component determines how quickly the 555 turns things on and off.

Hookup/Jumper Wire
(Top right) You'll need this to connect one thing to another.

Resistors
(Bottom) Resistors serve two purposes here: two are used to make sure the LEDs don't get too much current; the other two work with the capacitor to control the behavior of the 555 timer.

Battery Box
(Right) This holds 3 AAA batteries.

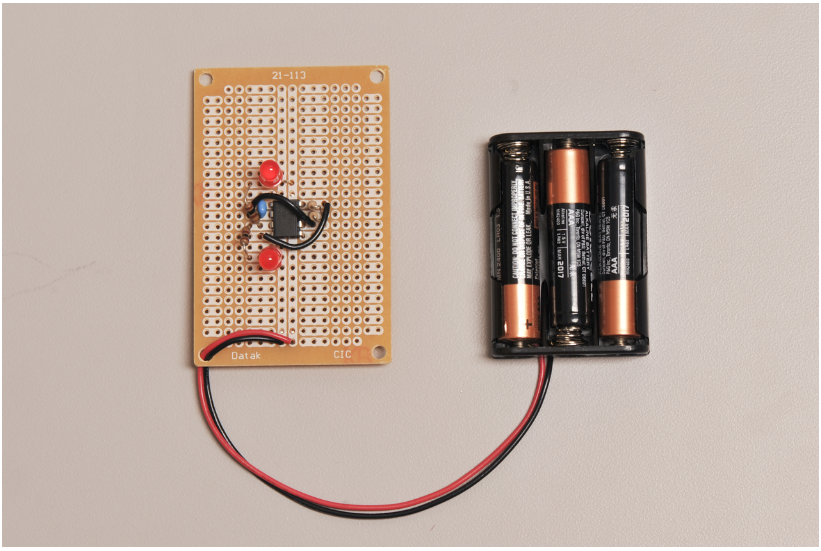

Figure 0-5. *Finished 555 blinky*

Contact Us

Maker Shed is here Monday to Friday, 7:30 AM to 5:00 PM Pacific Time

Email: *help@makershed.com*
Phone: 800-889-8969 or 707-827-7019

Answers to our most commonly asked questions: *http://www.makershed.com/help.asp*

Got a kit or product you think we should sell? Send a note to *kits@makezine.com*.

Want to carry MAKE books or kits in your store? Send a note to *resellers@makershed.com*.

1/Getting the Workspace Ready

Before you can start soldering, you need to get your iron ready. This involves a few steps. First, you'll have to attach the tip to the soldering iron. Then, you'll need to prepare the station: make sure your sponge is moistened, set the temperature, and "tin" the tip by applying some solder to it.

At the end of this chapter, your soldering iron will be assembled, clean, and hot enough to begin soldering.

Solder contains harmful chemicals. Wash your hands after soldering as well as after handling solder or the solder station sponge. Though it may be tempting to snack while you're working, keep any food or drink away from your soldering area.

1: Attach the Tip

Locate the soldering iron (make sure it's unplugged and cool), a soldering iron tip, and the tip nut that holds the tip in place. Insert the tip into the soldering iron as shown in Figure 1-1, and push it in as far as it will go. It will not require much force to insert it.

WARNING: Make sure the soldering iron is unplugged and cool.

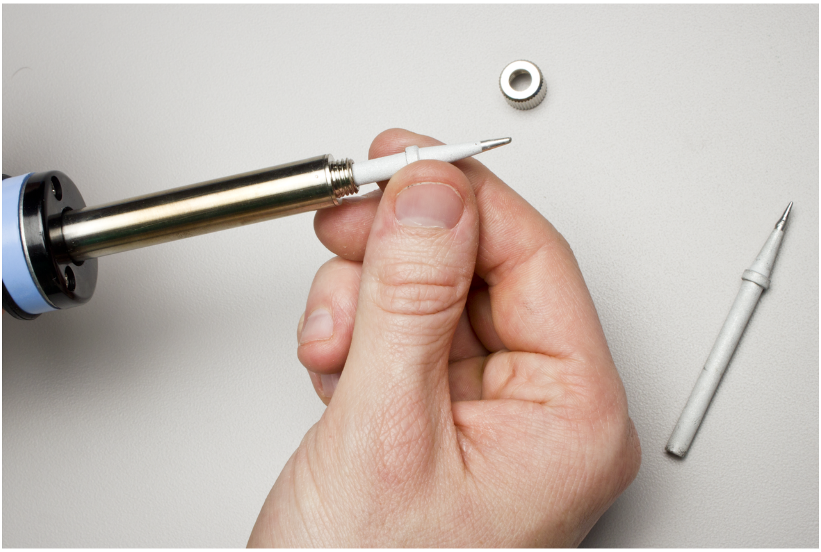

Figure 1-1. *Inserting the tip into the tube*

2: Secure the Tip

Slide the tip nut over the tip, and screw it in place. You can use pliers to tighten it (see Figure 1-2). Be careful not to overtighten it, since you will eventually need to replace it when the tip gets worn from all the projects you'll be making with it.

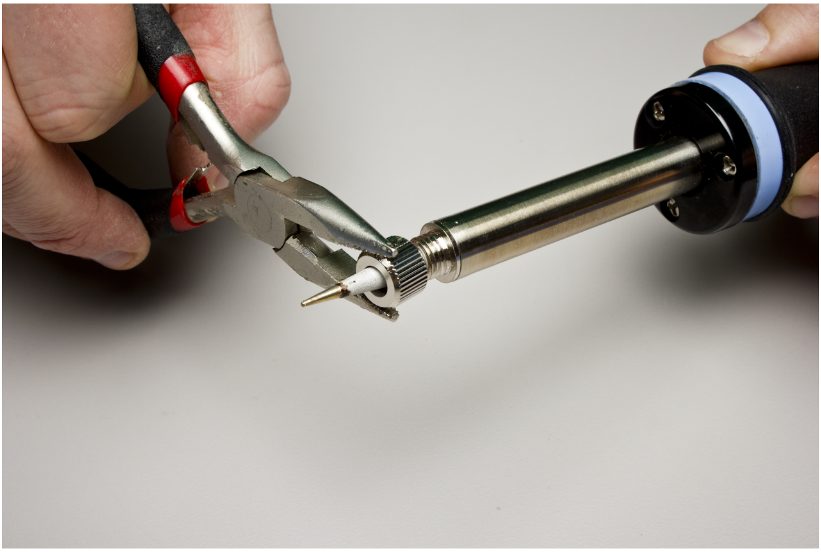

Figure 1-2. *Making sure the tip stays put*

WARNING: Over time, the tip nut might become loose. Let the iron cool down before you tighten it.

3: Wet the Soldering Station Sponge

You'll need to perform this step (and the remaining ones in this chapter) at the beginning of each soldering session.

As you use the soldering iron, you're going to be wiping the tip a lot, so you need to keep the solder station sponge moist. Pour a little clean, fresh water on the sponge as shown in Figure 1-3. Don't soak it, but make sure it's completely moistened.

NOTE: If your water has a high mineral content, use distilled water for wetting the sponge.

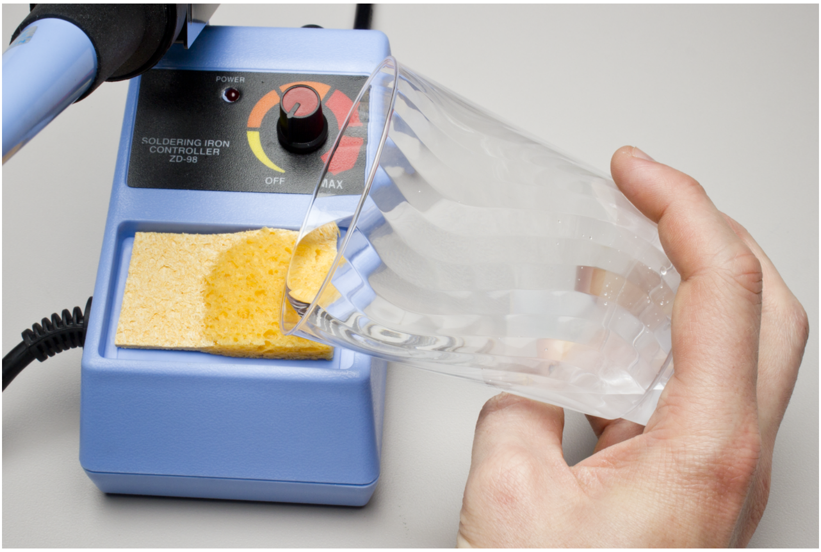

Figure 1-3. *Wetting the sponge*

4: Set the Station's Temperature

Plug your soldering iron in, and set the heat setting to the mark shown in Figure 1-4 (pointing at the 3 o'clock position on the dial). Give the iron a couple minutes to warm up.

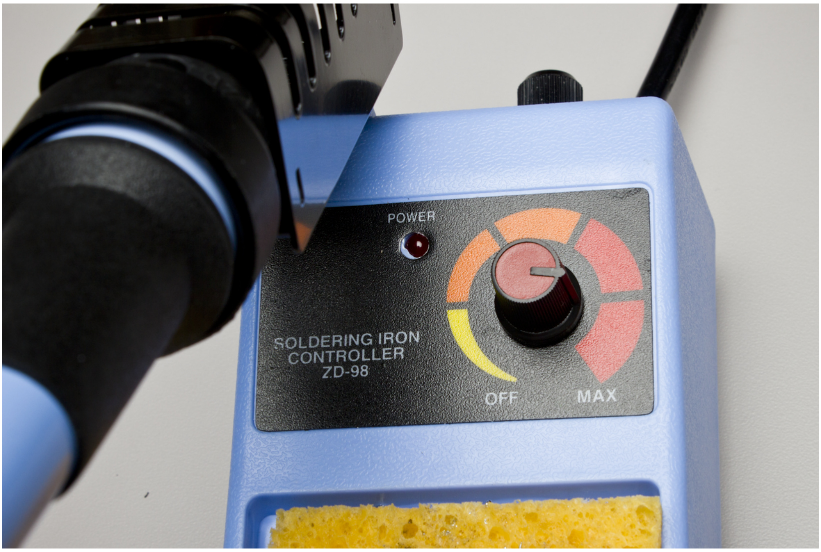
POWER
SOLDERING IRON
CONTROLLER
ZD-98
OFF
MAX

Figure 1-4. *Setting the temperature*

WARNING: The iron will get very hot, and certainly hot enough to burn you. Don't touch the tip and don't touch the tip to anything other than solder or components you are soldering together.

5: Wipe the Iron's Tip

Wipe the tip on the sponge as shown in Figure 1-5, being sure to turn the iron a few times as you wipe it, so you wipe all of its surfaces. It will make a sizzling sound as the water in the sponge comes into contact with the tip. The figure shows a sponge that's been used a few times. Notice the little blobs of solder that have come off the tip over time. A clean tip is critical to successful soldering. Even a small layer of oxidized material or other crud will limit the amount of heat that's transferred to the component you're soldering. Wipe the tip often, and keep the tip clean.

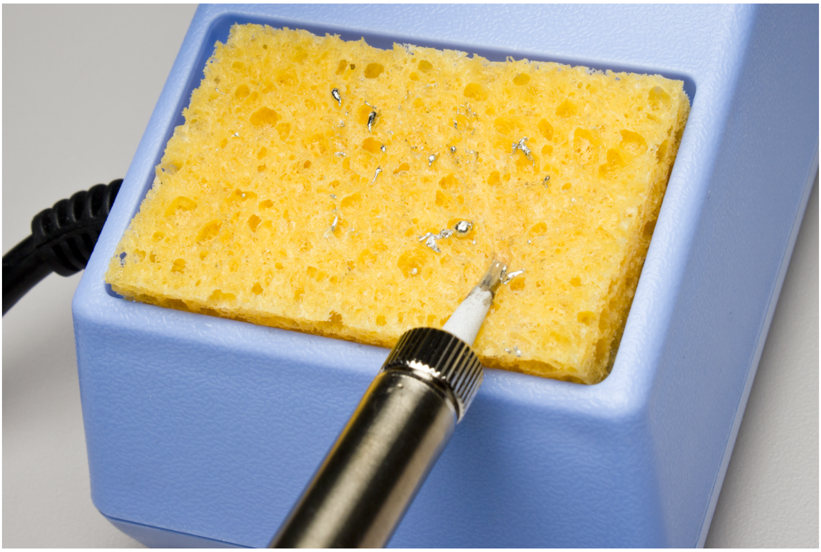

Figure 1-5. *Wiping the tip clean*

6: Tin the Iron's Tip

Before you begin soldering components, you should put a thin layer of solder on the tip as shown in Figure 1-6. This will have two effects: burning off any crud that shouldn't be there, but also providing a thin layer of liquid solder. This layer makes for a very effective heat transfer between the soldering iron and whatever you're soldering. Give the tip one more quick wipe on the iron and place the soldering iron into the holder until you need it.

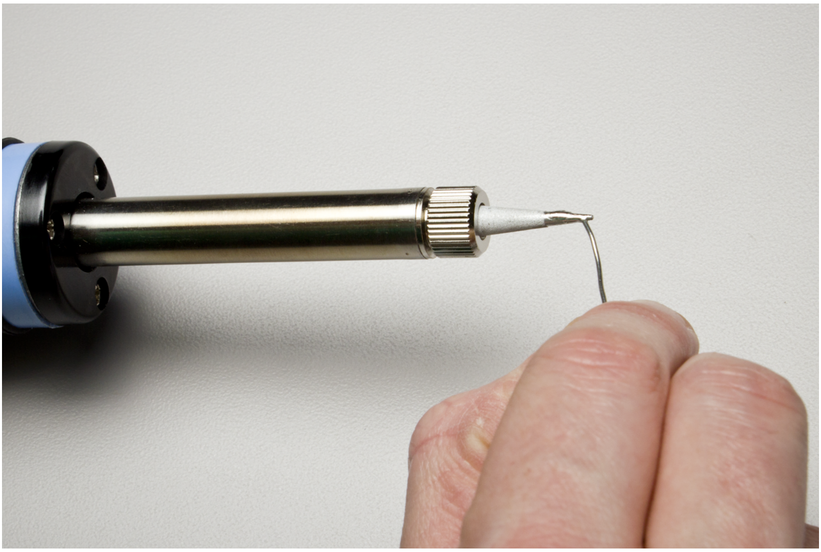

Figure 1-6. *Tinning the soldering iron tip*

NOTE: A small puff of smoke will appear as the *rosin flux* in the center of the solder is activated. The rosin serves two purposes: to clean the soldering iron (and the joint that you are soldering) and to help the solder flow freely. Avoid inhaling this smoke.

Go Make Your Learn to Solder Skill Badge

With your workstation all set up, you're ready to try your hand at soldering. The Learn to Solder Skill Badge ("Project 1: Learn to Solder Skill Badge" on page viii) is a great place to start. It's so simple, we think you can even try it before you read any more. But before you try the second project ("Project 2: 555 Timer Blinky" on page x), you should definitely give the rest of this guide a read.

2/How to Solder

With your soldering iron prepped, you're ready to begin soldering. This chapter has many helpful tips and techniques for soldering effectively, safely, and neatly.

Working with Solder

Solder is made of metal with a core that contains flux, which cleans the connection as you solder. When the metal melts, the flux begins to flow onto the joint. Some solder lead-based, but the solder in this kit is lead-free. To melt solder, heat it with the soldering iron as shown in Figure 2-1.

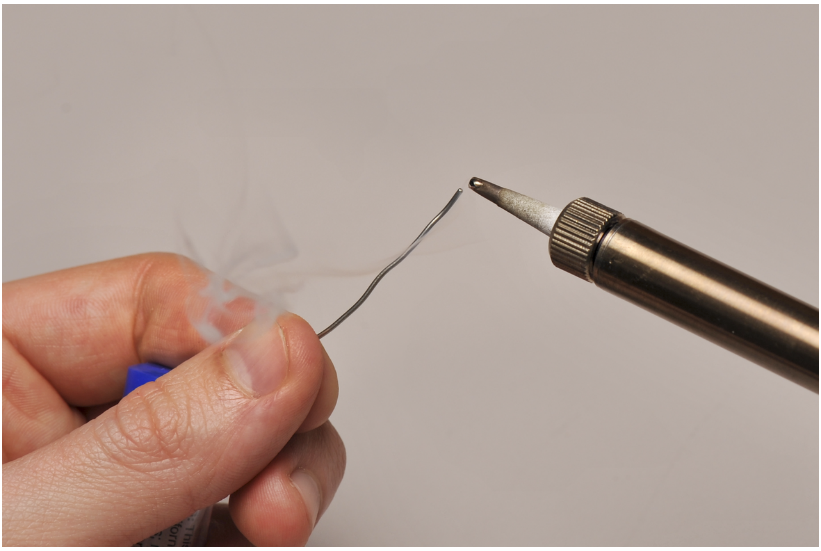

Figure 2-1. *Melting some solder*

Try melting some solder, but be careful not to drip it on yourself or anything other than a work surface. You can try pushing tiny balls of solder around on your work surface to see how it flows. If you have some bare hookup wire or similar metal, try heating it up by holding the soldering iron to it (don't hold the wire while you're heating it). Touch a piece of solder to the wire, but choose a point that's an inch or so from where the soldering iron is touching. How close do you have to get to the soldering iron tip before it melts?

Keeping the Circuit Board from Moving

When you're soldering, you've got a lot to juggle: the soldering iron, the solder, and the two things you are connecting to each other. The helping hands let you place the items you're soldering in a stable position. Use the clips to hold the item in place as shown in Figure 2-2. But once you start inserting items into the board, you need some way to keep them from moving. "Placing a Component in the Board" on page 11 shows you how to do that.

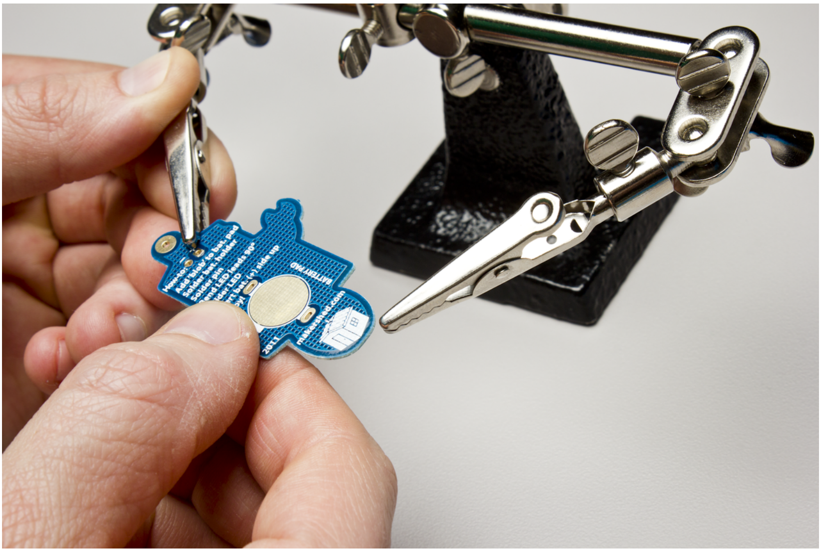

Figure 2-2. *Holding a PCB in the helping hands*

Tinning Solder Pads

You'll sometimes come across things that need a layer of solder to work right. For example, on the Learn to Solder Skill Badge, you need to put a bump of solder in place to give the battery a snug fit. With solder pads that are this big, you need to heat the pad with the iron really well; the pad is so large that it's going to take longer to heat. It's best if you melt the solder by pushing it onto the pad rather than pushing it directly against the iron. This makes sure that the solder flows thoroughly over the pad. If the solder doesn't melt, try tinning the tip again first.

Start with a little blob as shown in Figure 2-3, and spread it around evenly. When you've got the desired thickness, first pull the solder away, then pull the soldering iron away. Doing it in this order will avoid leaving chunks of solder behind.

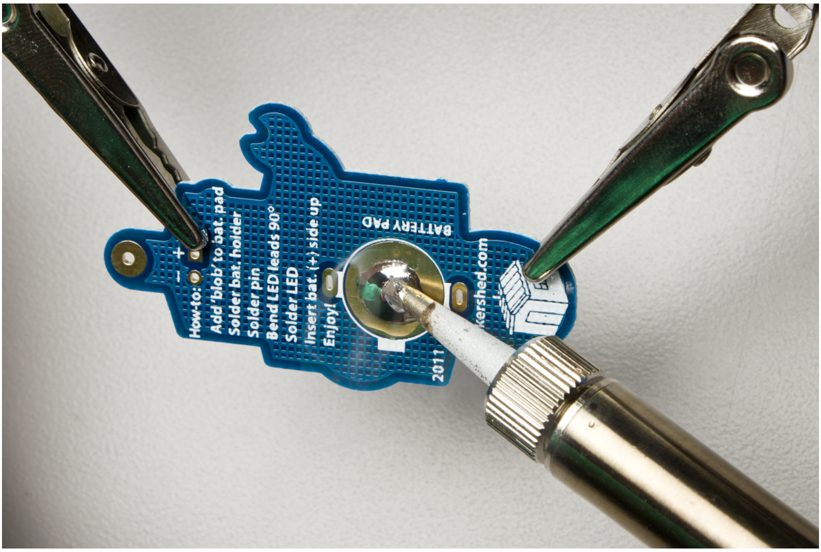
How-to:
Add blob to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
BATTERY PAD
kershed.com
2011

Figure 2-3. *Laying down some solder*

Placing a Component in the Board

You don't want components falling out of the board while you're soldering. In fact, you'll often insert a component with the board upside down (see Figure 2-4), then flip the board over when you put it in the helping hands. So the components have plenty of opportunity to fall out. To keep the component in place, bend the leads out as shown in Figure 2-5.

In some cases, you may need to quickly tack a component in place to keep it from moving. See "Stabilizing and Straightening Components" on page 14.

Don't try to place every component at once. Start with low-profile (shorter) components, solder them in one at a time, and move on to higher-profile components.

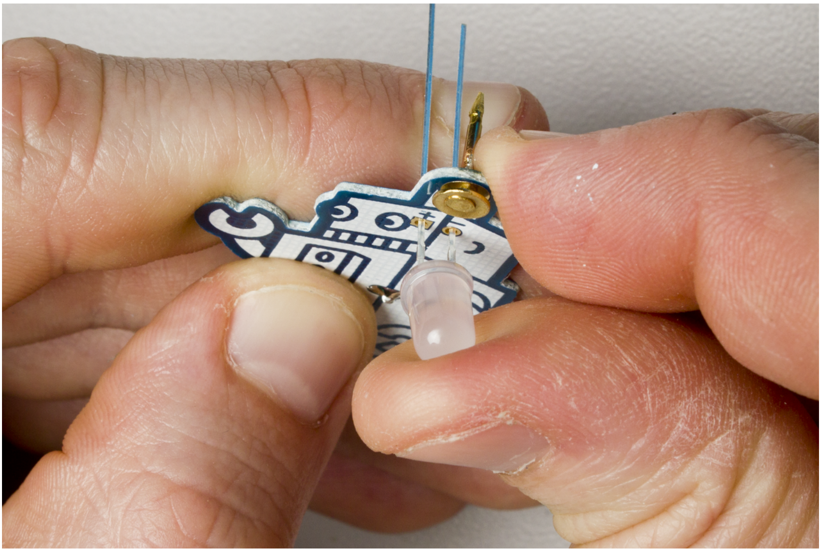

Figure 2-4. *Inserting an LED*

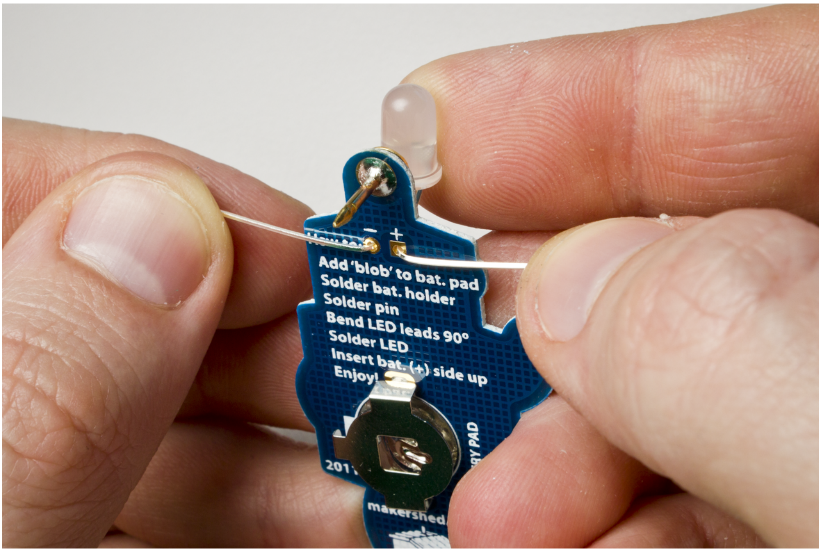
– +
Add 'blob' to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
makershed

Figure 2-5. *Bending out the leads*

Knowing Which Way a Component Goes In

Some components don't care which way you put them in: resistors, some capacitors (such as the one that comes with this kit), and many other components fall into that category. On the other hand, LEDs are *polarized*; electrical current will only flow through them in one direction. So if you put your LED in backwards, it won't light up at all. You can determine an LED's polarity in a couple of ways: First, the longer of the two leads is the positive (+), and the shorter is negative (-).

Second, look closely at the bottom of the LED and examine the ring that bulges out around it. There is a flattened part of that ring that indicates the side of the LED that's negative (-). You can see this flattened side on the right in Figure 2-6.

With ICs, this also matters. Every pin has a specific function, and there will be one or more pins for positive (+) and one or more pins for negative (-).

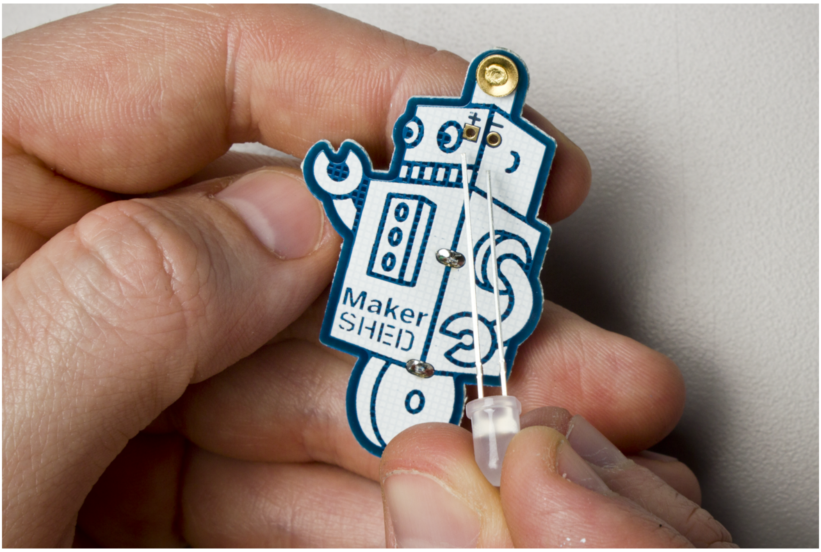

Figure 2-6. *Polarity—LEDs have it*

Stabilizing and Straightening Components

ICs present a unique problem. They don't have long enough leads for you to bend out effectively, which means they are more likely to slip and slide while you're soldering them. So do the best you can with that (see Figure 2-7), but flip the board back over after you solder the first joint. In most cases, you'll find that the IC has shifted. But since you only have one joint soldered, you can easily move the IC into the position you want. Once you've done this, flip the board over and solder the remaining connections.

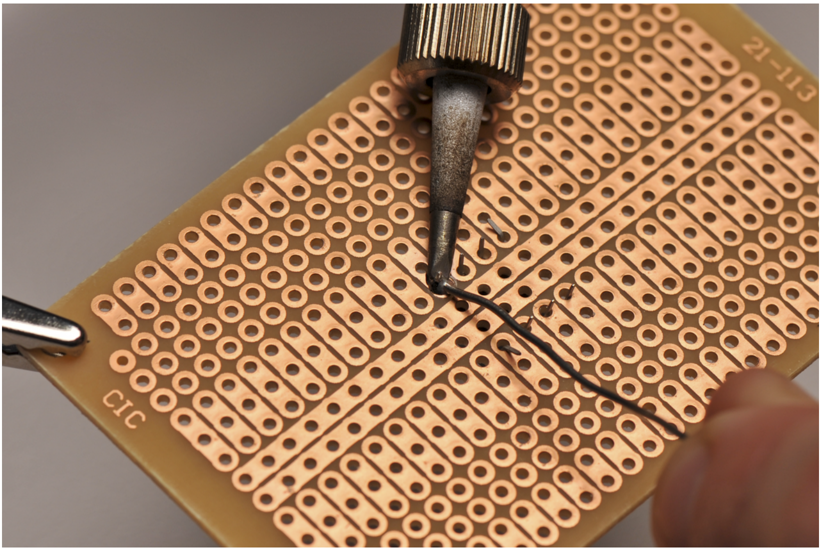

Figure 2-7. *Tacking down an IC*

Safely Soldering Sensitive Components

You need to avoid applying too much heat to a component, or you might damage it. One way to help avoid this is to clamp the heatsink to the component as shown in Figure 2-8. This will draw heat away from the component, protecting it somewhat.

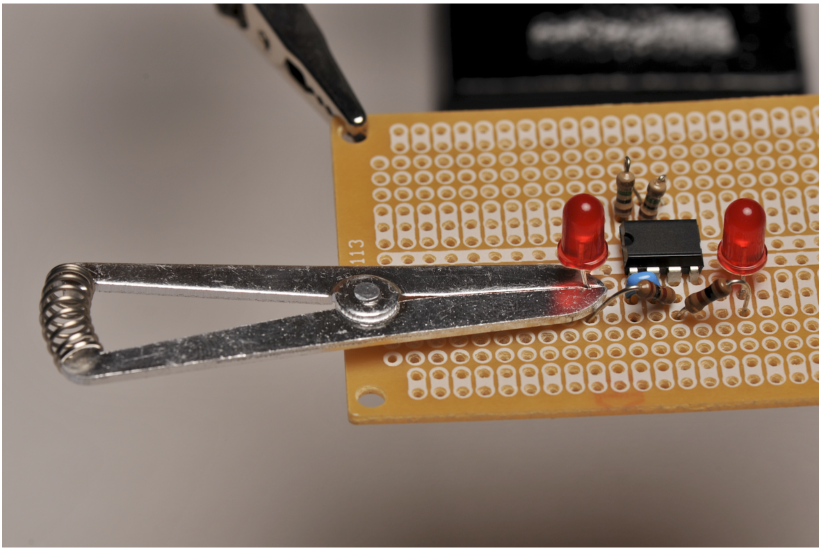

Figure 2-8. *Keeping heat away from the component*

Soldering a Component in Place

Wipe the soldering iron on the sponge and hold it so it's touching both the solder pad and the lead that protrudes through. Avoid using the soldering iron to melt the solder. Instead, heat up both the lead and the pad so that the solder melts when you touch it to them.

This must all be done quickly: within a second of touching the iron to the joint, push a small amount of solder into the joint, and let it flow around the joint as shown. Give it a second to flow, take the solder away, then wait a second, and take the iron away. With practice, you should be able to do all of this in three to four seconds per joint. Figure 2-9 shows how you'd solder the pin to the skill badge. Figure 2-10 shows an LED being soldered.

If you hold the iron to the component too long, you run the risk of damaging either the component or the solder trace on the PCB.

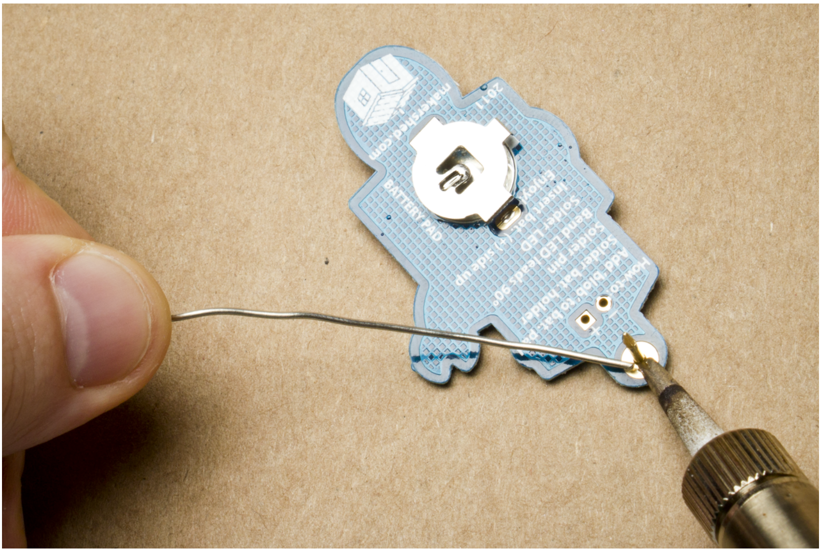

Figure 2-9. *Soldering a pin*

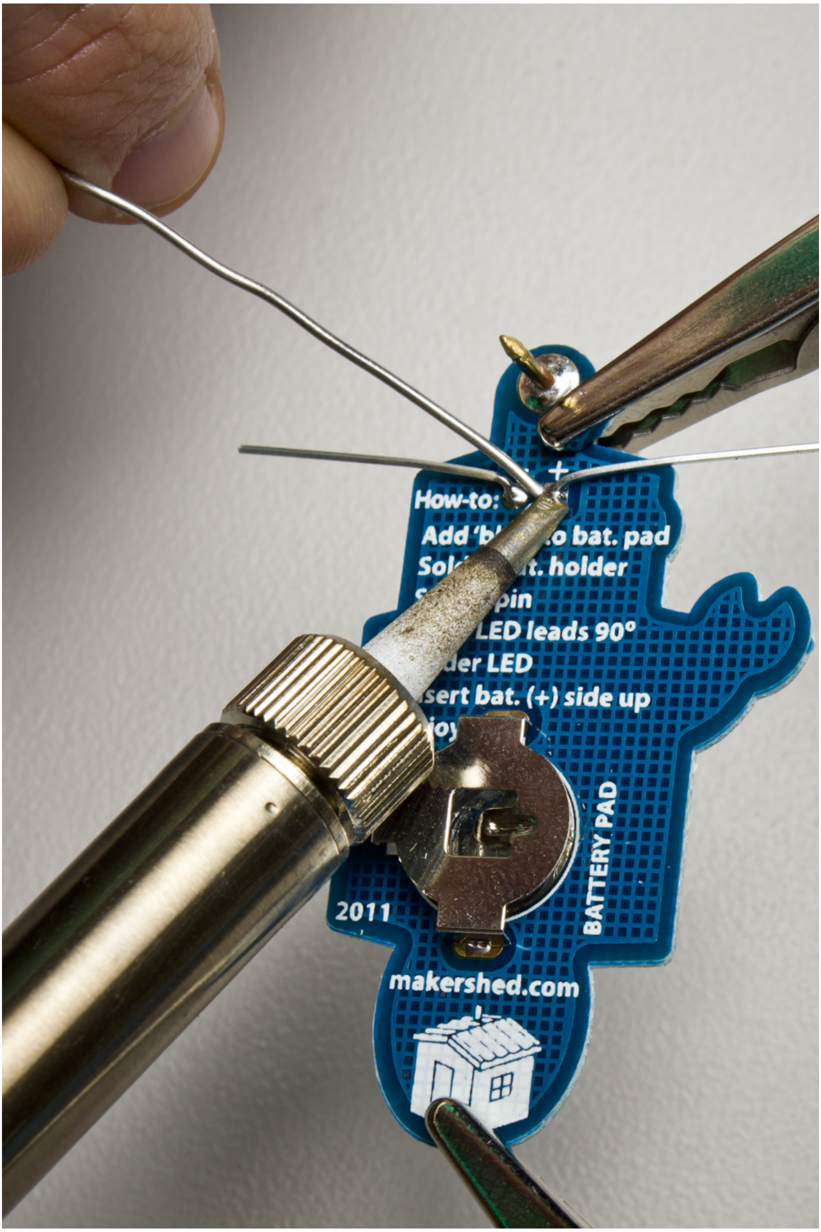
How-to:
Add 'b to bat. pad
Sol t. holder
pin
LED leads 90°
der LED
sert bat. (+) side up
joy
BATTERY PAD
2011
makershed.com

Figure 2-10. *Soldering an LED*

Trimming Your Leads

After you solder a component in place, you need to tidy things up. Grab the cutters from your kit, and trim the leads as close to the solder joint as possible (see Figure 2-11 and Figure 2-12). Don't cut into the solder joint; the idea is to trim the excess wire lead.

When you clip it, the wire lead will fly away from the clippers very fast, and could injure someone. The best way to keep this from happening is to use one or two fingers to hold the lead you're clipping. With one finger, you can put enough pressure on it to keep it from flying away.

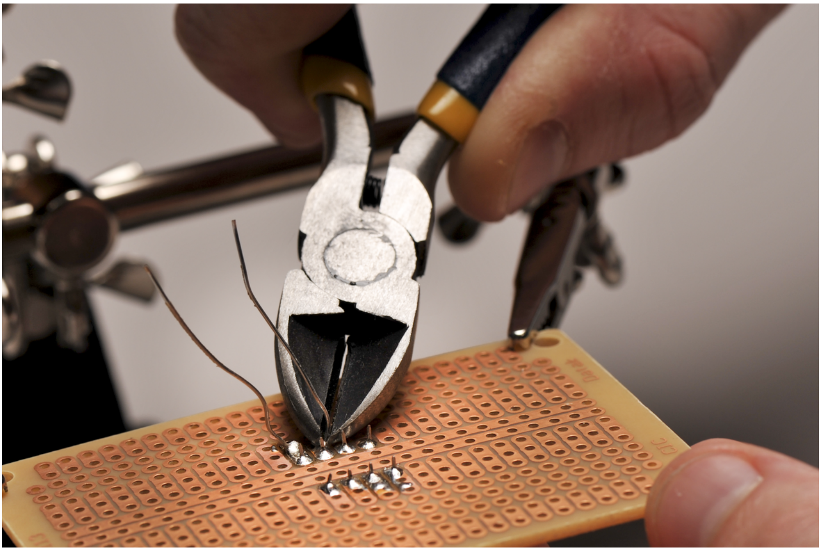

Figure 2-11. *Trimming leads on the 555 blinky*

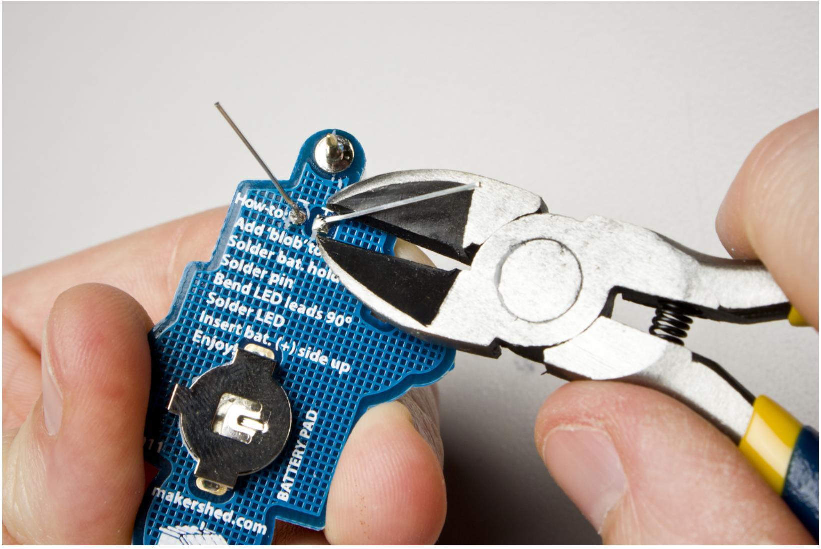
How-to:
Add 'blob' to
Solder bat. hold
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up
Enjoy!
BATTERY PAD
makershed.com

Figure 2-12. *Trimming leads on the Learn to Solder Skill Badge*

Going From Solderless Breadboard to PCB

The beauty of protoboard is that it mirrors the layout of a solderless breadboard, which is used for prototyping electronic circuits. Once you have your project laid out and working on a breadboard, you can easily transfer it to the protoboard. As with a solderless breadboard, the protoboard has rows and columns that are tied together. The layout is slightly different, though. Where the breadboard has two rails (one for positive, one for negative) on both the top and bottom as shown in Figure 2-13, the protoboard that comes with this kit has its rails in the center of the board.

Figure 2-14 shows the project from Figure 2-13, but laid out on a protoboard instead of the solderless breadboard.

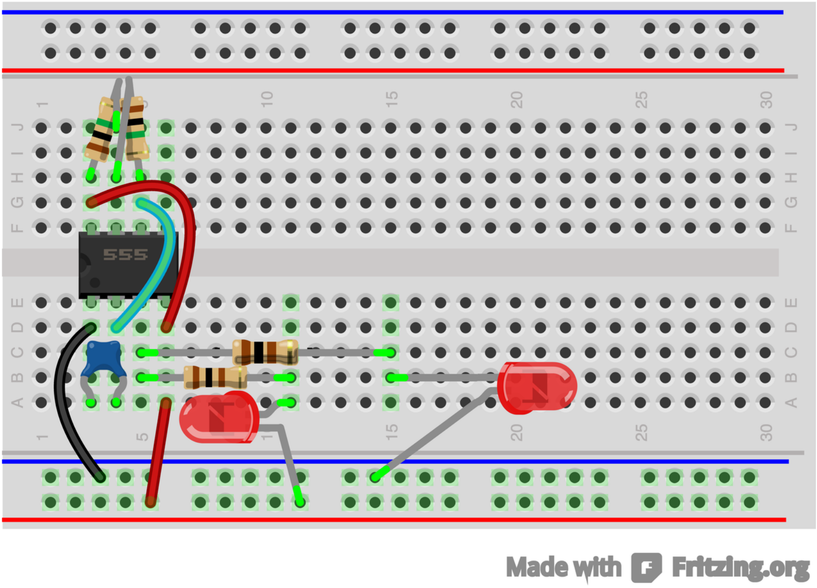

Figure 2-13. *The 555 blinky project on a breadboard*

NOTE: The breadboard diagram in Figure 2-13 was made with Fritzing, an open-source initiative to support designers, artists, researchers and hobbyists to work creatively with interactive electronics. For more information, see *http://fritzing.org/*.

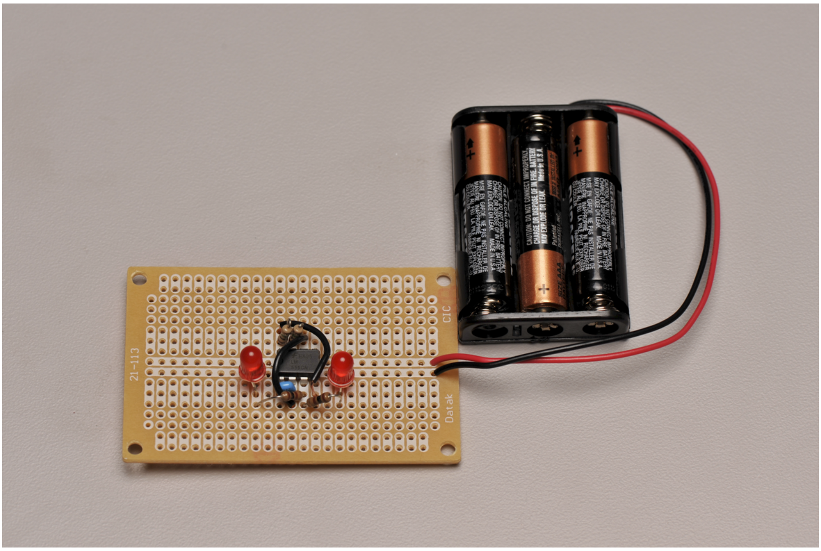

Figure 2-14. *The 555 blinky project on a protoboard*

Soldering Jumper Wire

Jumper wire can be treated a lot like other components (see "Placing a Component in the Board" on page 11): make sure there's enough bare wire poking through so you can bend it outward enough to hold it in place. The notched probe included with this kit can be helpful with placing jumper wire in the right place (see Figure 2-15).

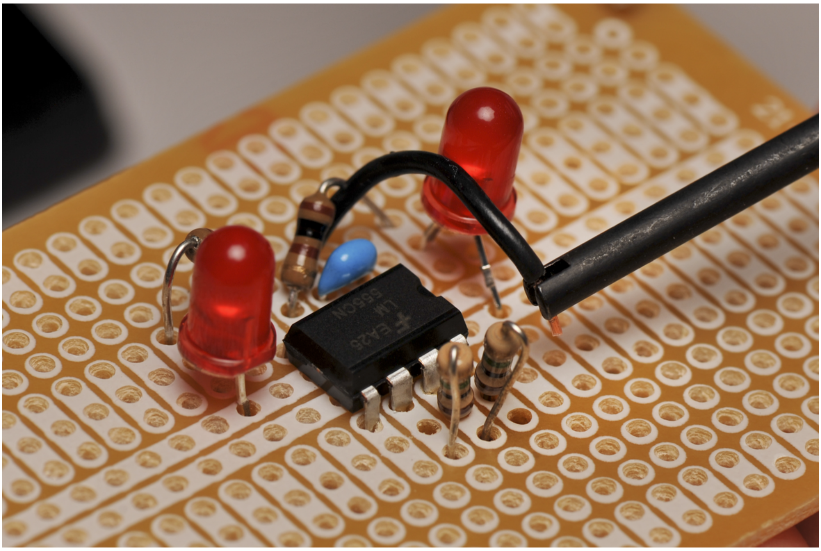

Figure 2-15. *Placing the jumper wire*

Bridging Joints with Solder

When you're working with the protoboard, you will sometimes need to bridge joints that are close to each other. This is preferable to using jumper wire since your board remains uncluttered on the top. It doesn't always look great on the bottom, though, because you end up with a lot of solder in some places. Figure 2-16 shows the start of a bridge: feeding the solder into the gap between solder joints. Figure 2-17 shows a bridge being formed.

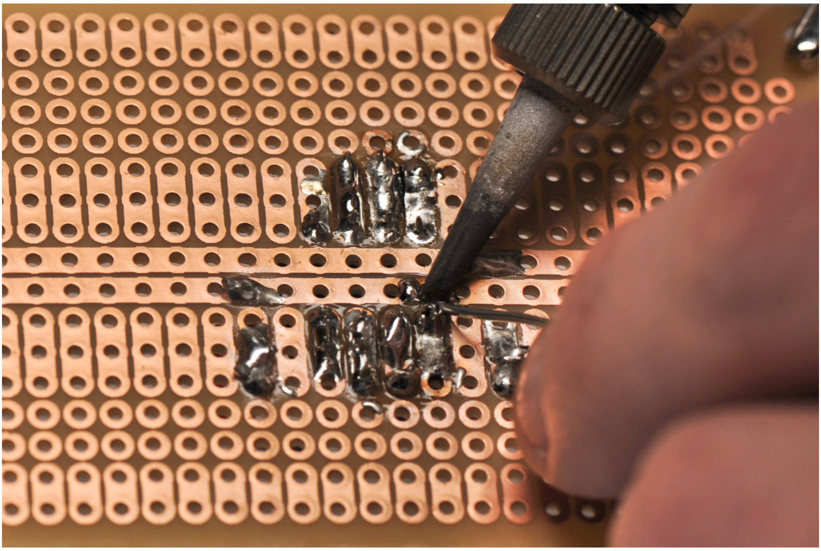

Figure 2-16. *Feeding solder into a gap*

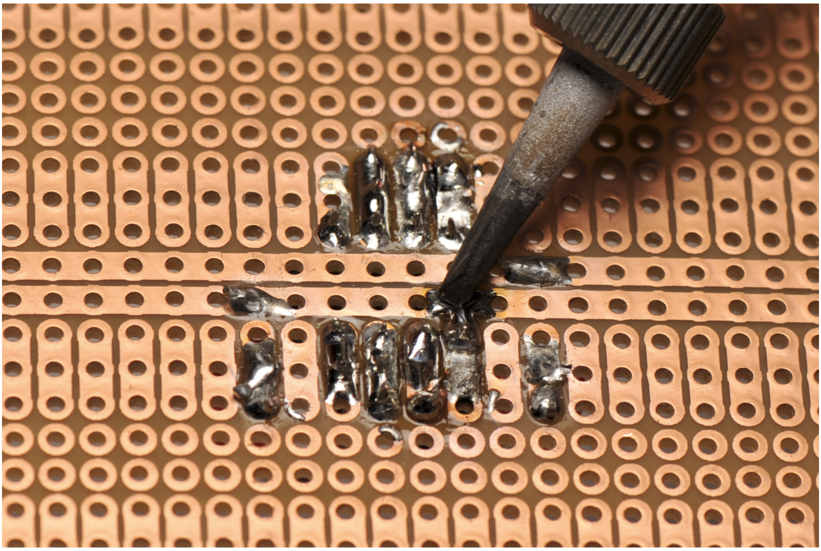

Figure 2-17. *Forming a bridge across the gap*

Relieving Strain on Cables

If you have something heavy dangling off your board, such as a battery box, it's very likely that the solder joints will come undone. If you pass the wires through the large holes on the board before soldering them down (Figure 2-18), you can add some strain relief (Figure 2-19) that will help prevent this from happening. If you have a small drill, you can also drill out a couple of holes close to where you are going to solder the connection, and loop the wire through there. This will be even more rugged than using the larger holes.

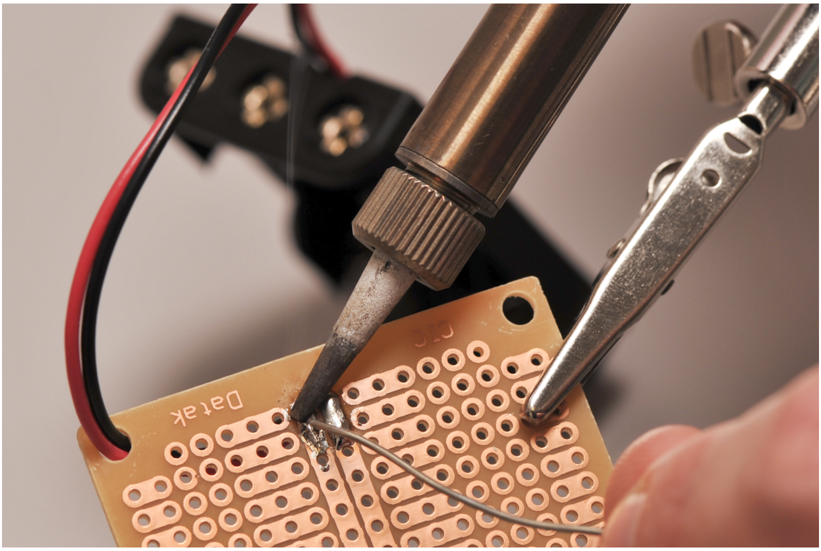

Figure 2-18. *Soldering the battery holder wires*

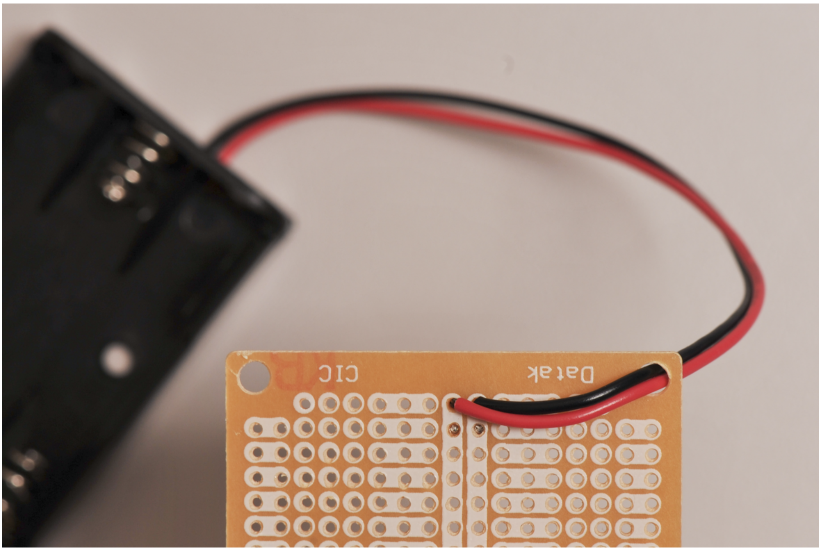

Figure 2-19. *Strain relief*

Getting the Perfect Solder Joint

It's not hard to get a great solder joint, but it's also easy to be sloppy. Here are some pictures of good and bad joints to help you out.

Figure 2-20
Here's a good solder joint on the positive (+) pad: solder surrounds the joint and makes a little peak. This joint had just the right amount of solder and just the right amount of heat.

Figure 2-21
The joint on the negative (-) pad is sloppy. The solder didn't flow around the joint, and some of it is spread out across the lead. You can try to fix it by holding the iron to the joint for a couple seconds; the solder should flow down the lead and onto the joint.

Figure 2-22
The joint on the positive (+) pad has too much solder, and it's balled up. You can use the desoldering wick or solder sucker to remove the excess. After you do that, hold the iron to the joint for a couple seconds to get the solder to flow around it.

Figure 2-23
Here are two joints for comparison: one good, one blobby.

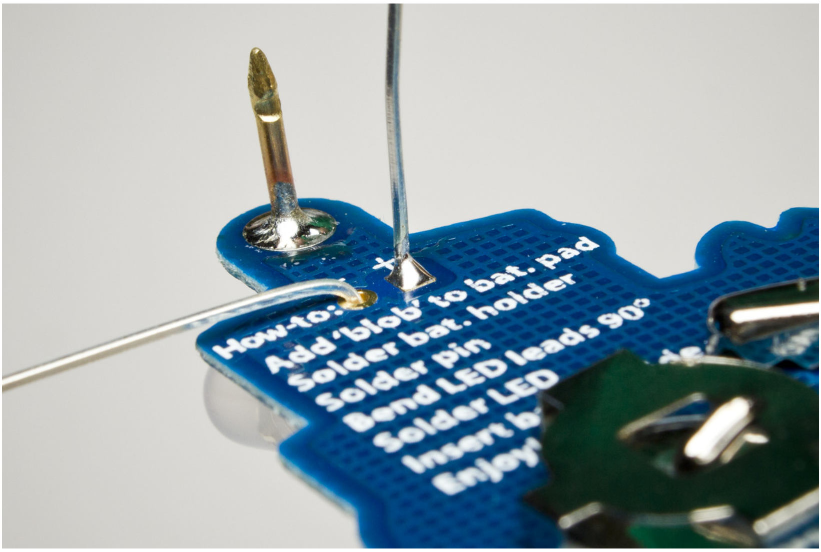
+
How-to:
Add 'blob' to bat. pad
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert b
Enjoy

Figure 2-20. *A good solder joint*

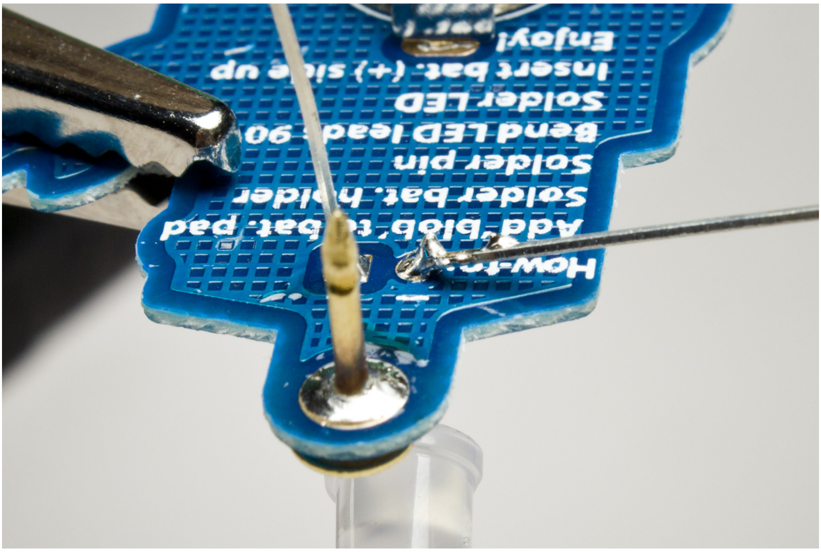
Enjoy!
Insert bat: (+) side up
Solder LED
Bend LED leads 90
Solder pin
Solder bat. holder
Add 'blob' to bat. pad
How-to:

Figure 2-21. *A messy solder joint*

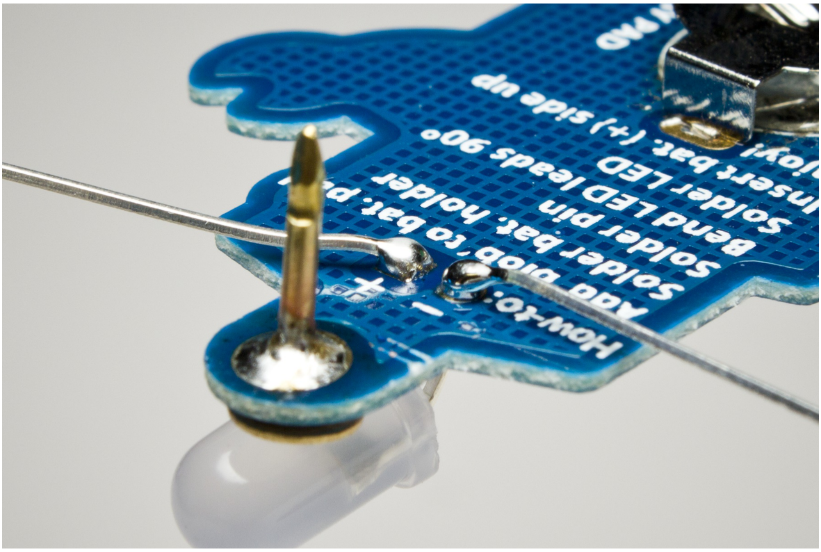
How-to:
Solder bat. holder
Solder pin
Bend LED leads 90°
Solder LED
Insert bat. (+) side up

Figure 2-22. *A blobby solder joint*

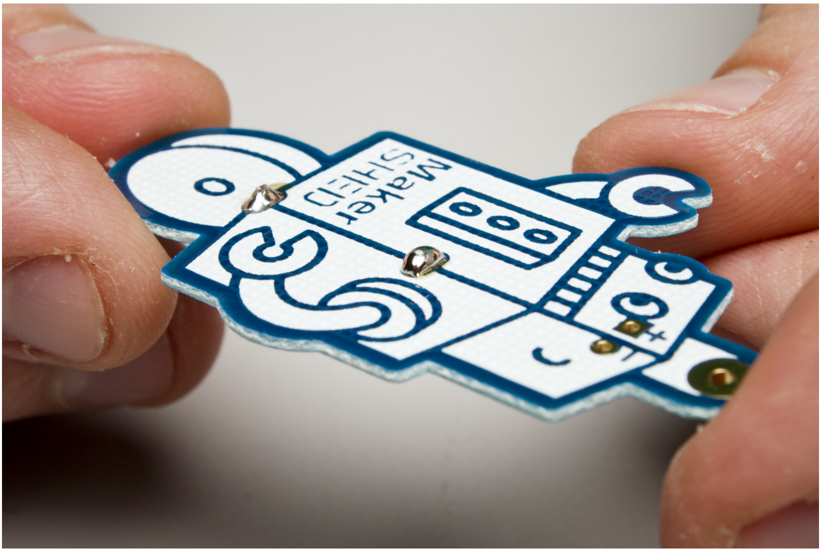

Figure 2-23. *Comparing two joints*

3/Desoldering

If everything went right the first time, you'd never learn anything. This section shows you how to recover from little soldering mistakes, and also includes a gallery of various solder joints so you can compare your work.

Desoldering is the "undo" command for soldering. With this technique, you can remove solder from a joint, allowing you to free the component so you can reorient it as needed. There are two tools for this: the *desoldering wick* (sometimes called *desoldering braid*) and the *solder sucker*.

The Desoldering Wick

Pull a small length of wick out of its spool (Figure 3-1). You don't need much, but if you don't use it all, you can wind the wick back into its spool.

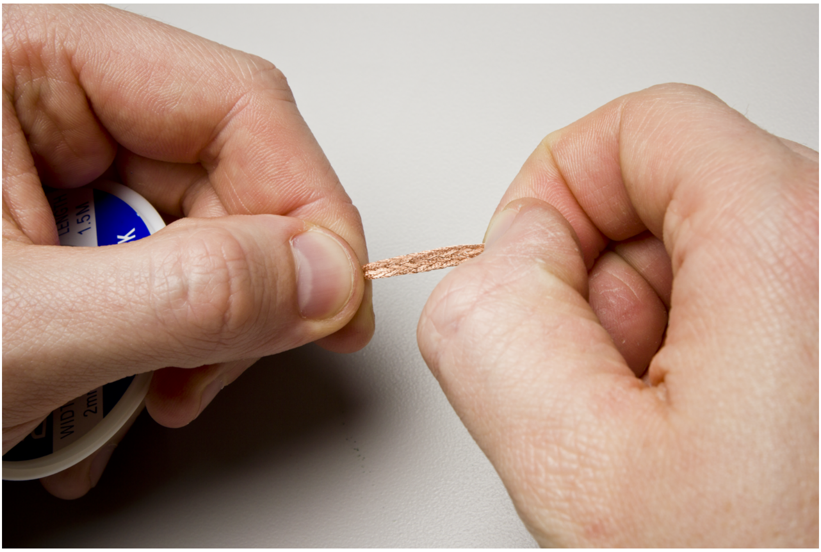

Figure 3-1. *Pulling out some wick*

Preparing the Wick

Before you use the wick, hold it as shown in Figure 3-2, and push inward. This will cause the wick to spread out, giving you more surface area to work with.

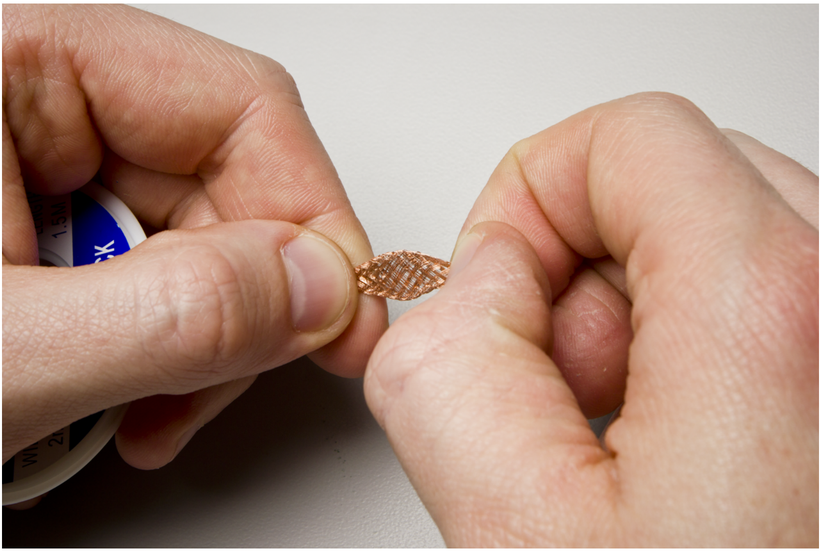

Figure 3-2. *Spreading the wick*

Desoldering with the Wick

Position the wick over the joint, and press down on the wick with the soldering iron as shown in Figure 3-3. The heat will pass through the wick and melt the solder. As this happens, the wick will soak up the solder, removing it from the joint. When you're done, you can trim off the used portion of wick with your wire snips.

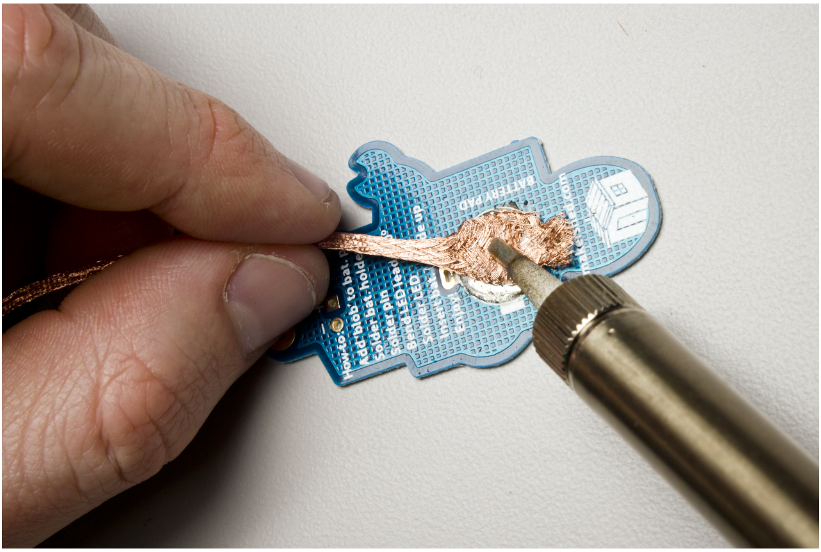

Figure 3-3. *Desoldering*

WARNING: The wick will get very hot, so take your hands off it as soon as you press the soldering iron to it (the soldering iron will hold it in place).

Desoldering with the Solder Sucker

The solder sucker is another tool for removing solder. It's best for removing excess solder, and you can use it to clean things up a bit before you use the wick. To use the solder sucker, press the plunger down until it locks. Use the soldering iron to melt the solder around the joint, and quickly bring the solder sucker's nozzle to the joint (Figure 3-4). Press the button on the sucker, and it will pull the solder right up into the tube. When you press the plunger again, any solder in the tube will be ejected.

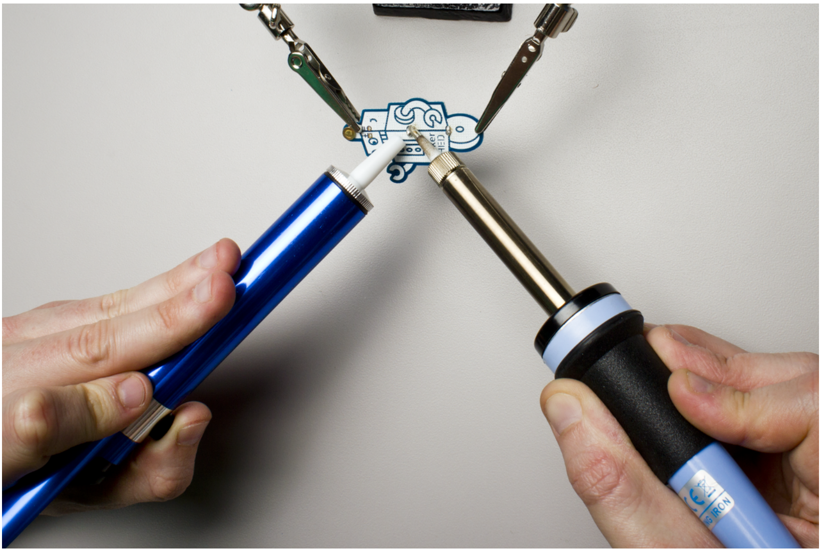

Figure 3-4. *Using the solder sucker*

4/What's Next?

Now that you've learned the ins and outs of soldering, it's time to do more. You can share this book with a friend so they can learn, or you can teach them yourself. And the world is full of wonderful things you can make now that you know how to solder.

Teach the World to Solder

Before you can teach someone to solder, you need to figure out what you want them to try making. You can purchase more Learn to Solder Skill Badge Kits from *http://makershed.com*. Here are a few tips for teaching soldering to others:

Keep Your Hands Off the Tools

You're there to teach, not to solder. And your students will learn best if they always have their hands on the tools. If you grab the tools from them and do part of the work, they won't learn from it.

Be Patient

Try to remember what you went through the first time you tried to solder. It takes time, and you did things incorrectly, and maybe you even made mistakes that required desoldering. Let your students learn from their mistakes, and let them take their time.

No Food or Beverages

Even with lead-free solder, there are chemicals in the flux that aren't good for you. If you allow food or drink near the work area, it could get contaminated. Keep food and drink away from where you're working.

Make Sure There's Plenty of Light

The components you work with while you solder are small, and hard to see. Make sure to use a well-lit workspace for teaching.

Keep Spare Parts Around

Beginners will break things. Pick up some spare resistors, LEDs, and other common components, and have them on hand in case you need to replace a damaged component.

Make More Things

Maker Shed (*http://makershed.com*) carries many electronic kits that you can apply your newly-learned soldering skills to. Here's just a few of them:

Supercap Racer Kit

One of the latest in our popular Mintronics line of DIY kits is this tiny race car designed by George Albercook. Once assembled, this little racer (Figure 4-1) will run around getting its juice from a supercapacitor. Capacitors are like little batteries. Supercaps store and release power faster than ordinary ones, so they're great for making quick-charge gadgets and power-on-demand circuits.

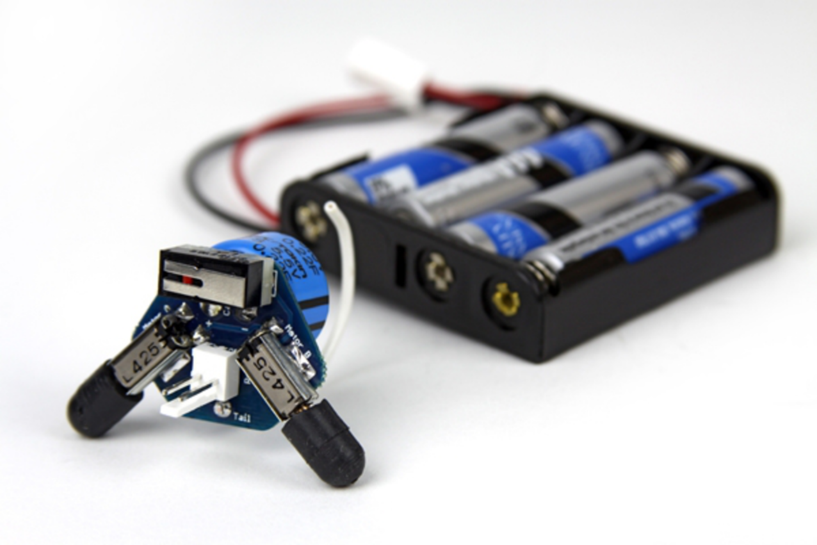

Figure 4-1. *Supercap Racer Kit*

Wee Blinky Kit

The Wee Blinky kit (Figure 4-2) is an easy-to-solder two (2) LED blinker circuit. It comes with a 9V battery snap, but will work with almost any voltage from 3V to 12V. A 9V battery is required but not included. It's tiny, it blinks, and it's a great kit to hone your soldering skills since it's cheap too!

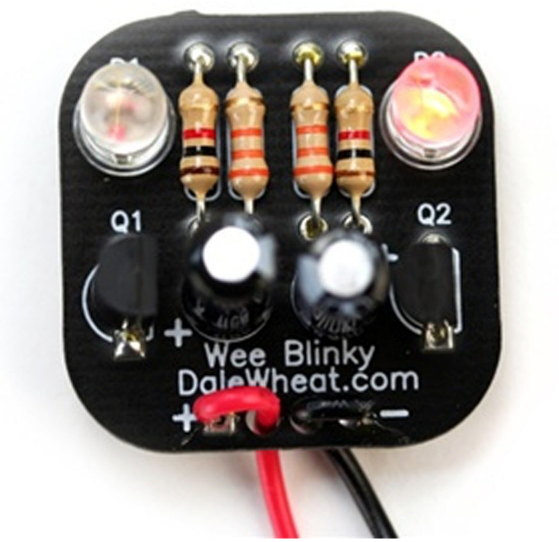

Figure 4-2. *Wee Blinky Kit*

MintyBoost

The MintyBoost (Figure 4-3) is a small and simple USB charger for your iPod or other MP3 player, your camera, cellphone, and any other gadget you can plug into a USB port to charge.

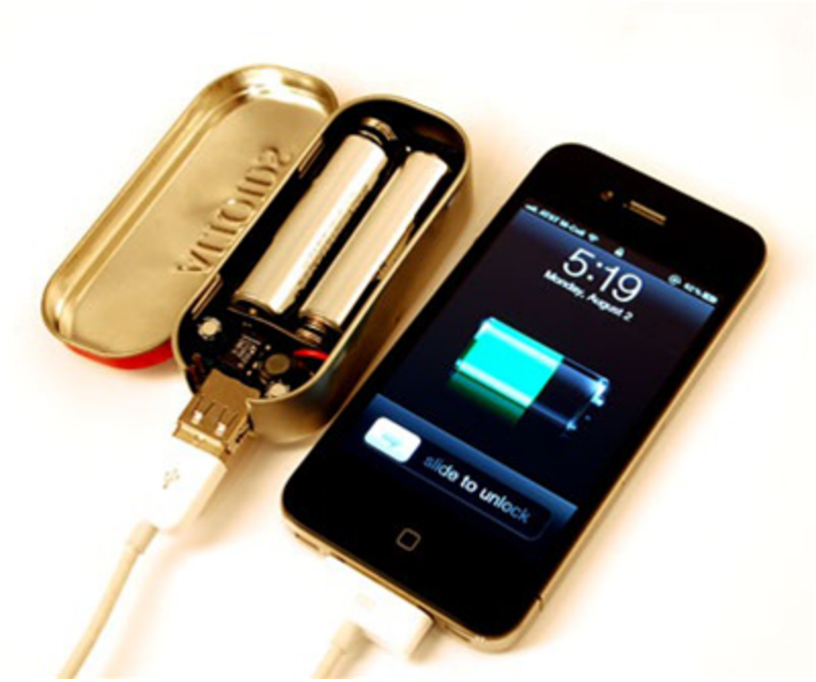

Figure 4-3. *MintyBoost USB Charger Kit*

About the Authors

Brian Jepson is an O'Reilly editor, hacker, and co-organizer of Providence Geeks and the Rhode Island Mini Maker Faire. He's also been involved in various ways over the years with AS220, a non-profit arts center in Providence, Rhode Island. AS220 gives Rhode Island artists uncensored and unjuried forums for their work and also provides galleries, performance space, fabrication facilities, and live/work space.

Tyler Moskowite, a programmer, engineering intern at Make Magazine, and student at Santa Rosa Junior College, has been tinkering with electronics for almost half his life. He picked up Arduino, then Android, and with the release of the ADK he has found his niche.

As a photographer for Make Magazine, Gregory Hayes has ruined more clothes than he ever did as a handyman, hiked more miles with a heavier load than he did as a backpacker, done more research than he did as a writer, and gotten closer to more human hands than advised by any epidemiologist. Taught to solder at the age of seven and forced to solder for his supper at the age of nine, he's now content to let others enjoy the lion's share while he stands by watching safely from behind glass.

How to solder

This pdf refers on informations from the site: http://store.curiousinventor.com/guides/
Have also a look to the 7min. Tutorial video on: http://store.curiousinventor.com/guides/How_to_Solder

Select a Soldering Iron

A 25 or 30 Watt iron should suffice for most small electronics work.

1. Most **soldering "guns" are vastly overpowered for electronics soldering** and can easily overheat components or expose them to harmful voltages. However, some people cleverly use them to solder multiple leads on surface mount devices. Soldering "guns" are for plumbing and much heavier duty applications, and are usually over 100 Watts. The "guns" work by passing high currents through the tips, and these currents can generate voltages that damage electronic components. Also, magnetic fields from guns with transformers can damage some electronics.By forming the heating element in the shape of of the chip, a soldering gun can be used to heat many leads simultaneously.

2. **How much wattage do you need for a particular application and how does wattage relate to tip temperature?**

   A loose analogy: Imagine a car tire has a leak, but you're trying to keep it inflated by pumping air into the tire at the same time it's escaping out the leak. The bigger the leak, the more air you have to pump into it to keep the pressure up. If the tire pressure represents tip temperature and the air lost through the leak represents heat lost through the tip, then wattage represents the maximum amount of air your pump could supply. Once more air escapes through the leak than your pump can replace, the tire pressure (or tip temperature) starts to drop.

   If you had a very small leak and a huge pump (say a 100 Watt iron equivalent), you might be afraid that the pump would cause the tire to explode since so much more air is going in and so little going out. But if you have a nozzle to regulate the pump's air, you could only allow just the right amount of air in to replace what's lost through the leak. This is how "temperature controlled" soldering irons work. As long as you aren't losing more heat out of the tip than the iron can replace (up to its rated wattage), it will automatically regulate just the right amount of heat into the tip to maintain the same temperature.

   However, typical plug-in irons have no such regulation. A 15 Watt iron always delivers 15 Watts of heat to the tip, and the tip temperature stops increasing only when 15 Watts of heat escape through the air. When the tip touches a part, its temperature drops, and if the part you're soldering can dissipate enough heat, the temperature will keep dropping until it won't melt solder any more. After the iron is pulled away from the joint, the temperature will climb again. There is some amount of natural regulation: as the tip gets hotter, it dissipates more heat, and as it gets cooler, it dissipates less.

   Usually, the bigger the component the more heat it can absorb and dissipate, so the general rule is that you need more wattage for larger parts. If you're just soldering small resistors and ICs, 15 Watts will probably suffice, but you may have to wait a bit in between joints for the tip to recover. If you're soldering larger components, especially ones with heat sinks (like voltage regulators), or doing a lot of soldering, you'll probably want a 25 or 30 Watt iron. For soldering larger things like 10 gauge copper wire, motor casings, or large heat sinks, you may need upwards of a 50 Watt iron or more. The following video shows what happens to tip temperature as 15, 25, and 40 Watt irons solder various sizes of wires and components. For cheap irons, higher wattage does indeed mean higher temperatures!

- What is the difference between cheap RadioShack® irons and more expensive ones like Wellers®? What do $100+ and $400+ soldering "stations" have over the cheaper kinds that plug straight into the wall? expand Among the irons that plug straight into a wall and don't have a separate station, the dirt cheap kinds will work satisfactorily for many applications. From personal experience, the tips on RadioShack® irons often come loose and sometimes can be impossible to remove. The irons can also get uncomfortably hot to hold after several hours of use. The more professional Weller (or other) lines are made for longer, continuous use and have insulation on the handles that keeps them cooler. They

can also take a wider variety of tips.
Soldering iron "stations" usually provide some control over the heat being supplied to the iron tip. Ones that are temperature controlled automatically control the amount of heat delivered to the tip so that it remains at a set temperature. In every iron, when the tip touches a component, some heat is lost and the temperature drops. One measure of quality is the time needed for the tip to regain its temperature. A nice feature of many soldering stations is that the tip heats up in seconds after you turn it on.
Many stations also allow you to hot-swap the iron tip, which can be very helpful if you're alternating between surface mount joints and larger components.

- If standard tin-lead solder melts below 400 °F (and lead free below 500 °F), why do most soldering irons have tip temperatures between 600 and 800 °F? Just what is the right **soldering temperature**?
  expand The basic reason that tips are so much hotter than solder's melting point is because that difference helps to transfer heat faster to the joint. What is the "correct" temperature is a debatable topic, but a common rule of thumb is to start off at 600 °F and increase from there until acceptable results are achieved. Typical Kester (a solder manufacturer) datasheets recommend 600-700 °F for lead-based solder, and 700-800 °F for lead-free solder. "No-clean" or "low solids" fluxes will burn off before a joint can be made with higher temperatures, so low temperatures (below 700) may be essential for these fluxes.
  From Kester's hand-soldering knowledge base: "When hand soldering with a rosin flux such as the Kester #44 or the # 285 the recommended iron tip temperature is 750°F. If you are soldering with a low residue no clean solder such as the #245or # 275 we recommend a tip temperature of 600-650°F.
  What are acceptable results? The goal is to heat up the parts enough so that solder will adhere to them and form a good bond. The higher the iron temperature, the faster it will heat up the parts, so why not set it extremely high to work faster?
  Besides the obvious increased risk of overheating components and the board, higher temperatures cause the iron tip to oxidize faster and can significantly reduce its life. Some claim a 10 °C rise reduces tip life by half (ref p.33). For occasional use, though, tip life may not be much of a factor, especially if the tip is kept covered with solder at all times.

- **Tip size and shape:** a basic guide is to pick a tip that's slightly smaller than the pad you're soldering to. From there, you want a tip with a large thermal mass and short stroke (why?) In most soldering irons, the tip is not actually the heater, but sits in between your work and the heater. You can think of it like a heat bucket that empties into your work and gets filled again by the heater. Typically touching a component empties heat out of the tip much faster than the iron can replace it, and if you have a small bucket (tip), the temperature will quickly drop to an ineffective level.
  Especially if you have a small wattage iron (15 Watts or less), the temperature will drop before you can heat up a larger part, or you'll have to wait a bit in between joints for the tip temperature to recover. With a bigger bucket (tip), you can handle larger joints with smaller wattage, but eventually you'll need to step up the wattage.
  The "stroke", or length of the tip should be minimized to get the heater closer to the work; it takes some time for heat to transfer through the tip. This is balanced with the need to get into tight places where you need a longer tip.
- What do common tip shapes look like and what applications are they best for?

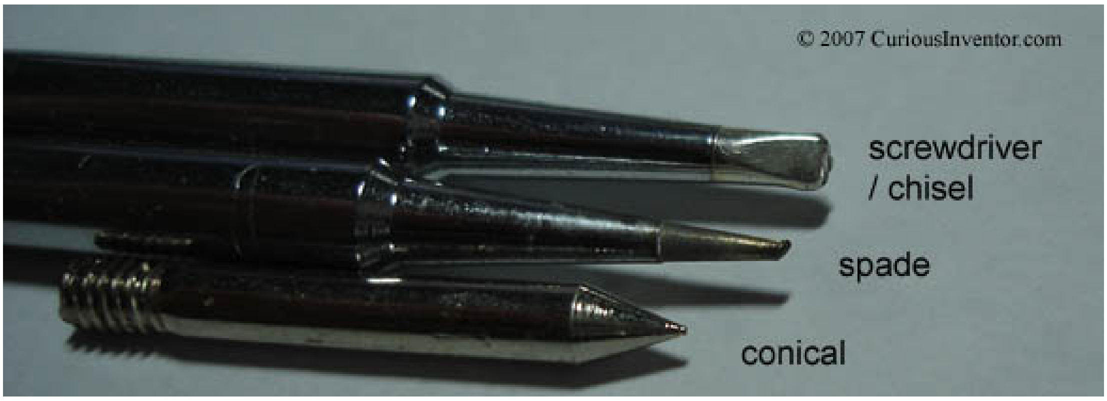

Screwdriver, spade, and conical are some of the more common tip shapes. Personal preference is the biggest factor when choosing a tip, but the goal is to get as much surface area contact between the tip and work as possible. Chisel and spade tips have more surface area at their ends, and also "hold" solder at their tips more readily than conical tips, which have a tendency to draw solder away. Even for fine pitch surface mount soldering, having a small flat at the end can be helpful.

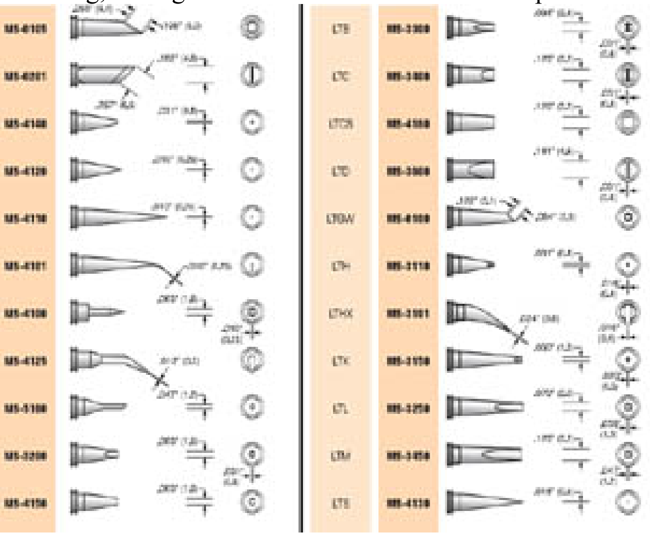

- There are myriad other tip shapes and sizes. The picture to the right shows one Plato catalog page of many. Some other non-standard shapes include a knife-blade (useful for fine pitch leads) and a surface mount desoldering tip.

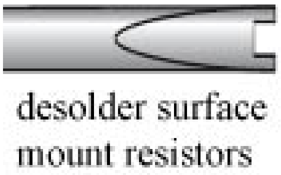

desoldering a surface mount soic chip © 2007 CuriousInventor.com

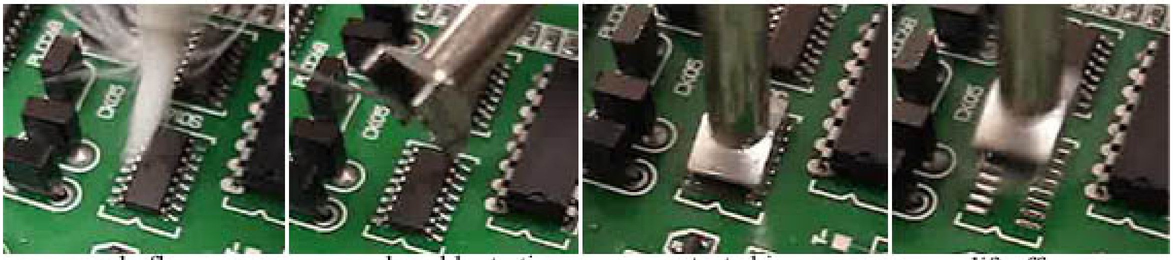

apply flux apply solder to tip contact chip lift off

To preserve tip life, the number one thing you can do is reduce the tip temperature (if your iron allows this). After that, ALWAYS keep a layer of solder on the tip to prevent the tip itself from oxidizing, and clean it in between uses. Put a glob of solder whenever you put it back in the stand, and before you turn it off. When heating up a new tip for the first time, hold solder against it so the tip can be covered as soon as the iron gets hot enough.

The longer flux residues and oxides are left on the tip, the harder they are to clean off. They also can drastically reduce the tip's ability to heat up a part, and prevent solder from "wetting" the tip. Regular cleaning of the tip before use is one of the best ways to prolong tip life and make soldering easier. It's important that solder "wet" or cling to the surface of the iron--without solder in between the tip and work the tip's ability to heat is drastically reduced.

- What about gas powered irons and the Cold Heat® iron that is supposedly touchable 1 sec. after use? Butane (and other gas) powered irons are mainly used in situations where electrical power isn't

available. Weller sells some battery powered irons as well.
Everyday Practical Electronics gives a pretty damning review of the Cold Heat iron here, in addition to having one of the better how-to guides out there. To summarize, the Cold Heat® iron has a forked end that you must bridge with the work or solder to turn on the iron, so it can be hard to hold it in a place that keeps it on and also effectively heats the part. Many people complain about pushing harder to make a good connection and then having the brittle tips break. Running power through your work to heat it may not be the greatest idea with some parts. Finally, the iron doesn't get hot enough for a lot of jobs, or cool enough to do anything like throwing it in your pocket right after use. But for something that's portable and cordless, heats up and down in under a few seconds, maybe it's worth the price ($20). Weller's battery powered ($20) iron doesn't have a forked end and supposedly heats up in under 15 seconds, but I don't know about cool-down time.

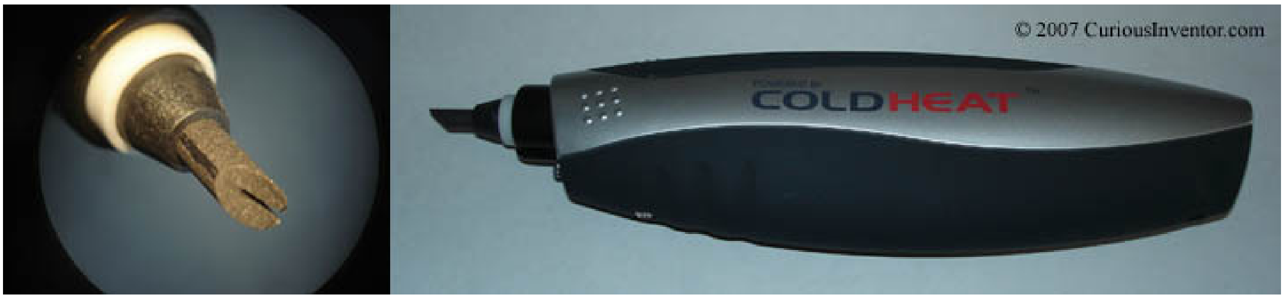

What kind of solder (rosin cored, etc. lead-free)? What is flux and when is it necessary?

As a starting place, for most small electronics soldering, 1/32 inch (.03) rosin-cored, 60/40 (tin-lead) or 63/37 solder should work fine. Rosin-cored lead-free is fine, too. Unless you have reason otherwise, don't use "no-clean" solder--it's very likely that you don't need to clean the regular rosin-cored solder. The solder should be thin enough to prevent accidentally applying too much (and causing a solder bridge), but thick enough so that more doesn't have to be gathered from the coil too often.

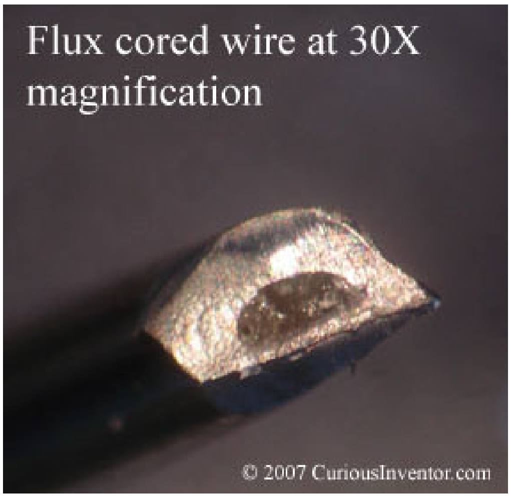
Flux cored wire at 30X magnification
© 2007 CuriousInventor.com

Besides affecting your feed-rate and convenience, the solder thickness also relates to the amount of flux that is delivered. Flux is basically a weak acid that removes oxides so that solder can adhere to the metal, and is so essential to the soldering process that it's built into the core of common wire-solder. It also helps the solder spread out (reduces surface tension), transfer heat, and acts as a protective blanket to keep oxygen away from the metal until solder displaces it.

For the most part, manufacturers include a sufficient amount of flux in the wire, but if you use an extremely thin wire there may not be enough to clean the joint OR the iron tip. Consider using a thicker gauge for cleaning the tip periodically if you're using especially thin solder. Liquid flux is helpful for SMD soldering, too.

When picking a wire-solder, there are 4 features to decide on: flux type and amount (% weight), alloy (tin-lead, lead free, silver bearing, etc.), thickness and total amount (1oz, 1lb?).

- **Flux:** Just what is flux, what kinds are there, and when do I need liquid flux?

  Why it's needed: Solder doesn't just freeze on a joint, it actually forms a metallurgical bond by dissolving and chemically reacting with the base material. Unfortunately, almost all metals oxidize in air and form an oxidized layer that prevents solder from wetting and bonding to them. What is oxidation?

  Oxidation, is when Oxygen (or other oxidizing agents like sulfur) atoms combine with base materials, stripping loosely attached electrons and forming new compounds like iron(III) oxide. It is what happens when apple slices turn brown, iron rusts, copper turns black / green and unsolderable. (read references for more accurate/complete explanation)

  Results vary significantly. When iron rusts, the oxides flake off until no iron is left. Alternatively, aluminum oxidizes extremely rapidly, but is then protected from additional oxidation by the oxide layer. This layer makes Al impossible to solder without using special solder and extremely aggressive flux, or plating the surface with a solderable metal, like nickel. The chromium in stainless steel fulfills the same function, oxidizing to form a protective barrier that is difficult to solder. Gold remains shiny because it doesn't oxidize, and is easy to solder to, but forms brittle joints. Heat, moisture, and salt all increase the rate of oxidation.

Oxidation can add a hidden cost to components and boards that may have been sitting on the shelf for long periods of time, or have been exposed to hot, humid environments. Copper pads on PCBs (printed circuit boards) are covered with solder or are plated to prevent oxidation, but given enough time, oxygen can still penetrate these barriers. Surplus parts in particular may need a bit of steel wool.

Some interesting references:
rusting chemical reaction
fairly understandable explanation of oxidation
wiki/Corrosion
wiki/Oxidation

Oxidation occurs much faster at higher temperatures, so even if you somehow had clean metals to start with, you would still need flux to prevent new oxides from forming while soldering.

The main choice to make when deciding on a flux, whether it comes in a cored wire or a liquid or paste form, is how aggressive it should be. The more aggressive or "active" the flux, the harder the oxides it can remove, and the faster it can remove them. Going from weakest to strongest, typical choices for hand soldering applications include: "no clean", RMA (Rosin Mildly Activated), RA (Rosin Activated), and water soluble. A newer classification system (J-STD-004) has recently been adopted and classifies fluxes not by rosin content, but by activity, material, and halide presence.

The new system classifies flux by material (RO=rosin, RE=resin, OR=organic, IN=inorganic), activity level (Low, Moderate, High), and halide presence (0 or 1). No-clean, rosin-based no-clean fluxes might be labeled ROL0 or ROL1. Although there is no direct translation between the old system and new, most R and RMA fluxes fall under Low activity level, RA are generally labeled as Moderate activity, and water soluble are High activity. (source IPC-HDBK-001 www.ipc.org)

The downside to choosing a more aggressive flux is that the residues left over after soldering MAY be corrosive, conductive, or enable fern-like growths called "dendrites" to grow between connections. A brief description (p.29) of dendrite growth and some great pictures at the end of this paper.

Because of the risk of corrosion and dendrite growth, most manufacturers clean off the residue from RMA and RA fluxes, and some even clean "no-clean" residues. The question of what flux to use and whether / how to clean it is quite involved.

Rosin flux is quite an interesting animal. Made from pine tree sap, at room temperature, it's an excellent insulator and non-corrosive. When it hits 226 °F it begins to turn acidic and attack oxides, but then when it cools it supposedly leaves residues that are again inert. Kester's data sheet for "44" flux (classified as RA and ROM1) claims no cleaning whatsoever is necessary. I haven't read of any manufacturers that would use a RA (or even an RMA) flux and not clean it--the military won't even use a RA flux WITH cleaning because of the risk that some will be left behind. This Chemtronics author recommends cleaning even the "no-clean" fluxes. He also points out that even if the residue is non-corrosive and non-conductive, it might be tacky and attract dust that causes a short.

To add one more piece to the puzzle, flux generally gets used up during the soldering process. This is why no-clean fluxes are oftentimes ineffectual for lead-free soldering, which can require slightly higher temperatures and longer heating because the lead-free solder "wets" slower. The no-clean flux can burn off before the joint is complete. Alternatively, if you apply liquid flux far from a joint, it may still be active (corrosive) if it never got heated.

I'm not making missile guidance electronics, I'm making a robot that pours a beer, what flux should I use and do I really need to clean? Even manufacturers of non-life critical electronics have much more stringent reliability requirements than an individual. They must ensure that tens of thousands of products will work for multiple years, not a single project.

The safe advice is to use the least aggressive flux that enables solder to quickly wet or cling to the surfaces, and then clean off the residues with alcohol and lint-free wipes (don't just spread the flux around). Try starting with a rosin-based mildly activated flux: RMA. I am inclined to trust Kester's spec sheet for "44" (RA) flux that says it does not actually require cleaning. Other flux manufacturers may have RA or RMA level fluxes that do indeed need to be cleaned, so if you don't know what you're using, cleaning is probably prudent. If you are going to clean rosin fluxes, do it soon after soldering because they quickly harden (see pics under 'Cleaning'). Finally, I would personally avoid no-clean fluxes and solder unless you have a critical application and very clean parts.

Lead-free solder generally requires a made-for-lead-free flux designed to be used under slightly higher temperatures.

Liquid flux can greatly help with surface mount soldering and desoldering components, but the flux inside cored-solder should be sufficient for through hole components. When soldering SMD components and desoldering pretty much anything, liquid flux acts as a blanket that helps to spread heat and also keep oxygen away from the metals. Finally, flux lowers the surface tension of solder, helping it to spread out and wick into connections.

A water-soluble flux may be necessary for heavily oxidized parts or difficult metals like nickel. Without question, clean these fluxes. Special fluxes and solders exist for aluminum and stainless steel and these also certainly require cleaning. Never use acid-core solder; it deposits zinc chloride into the solder that cannot be cleaned out. A final reason to clean flux residues is if you'd like to apply a conformal coating and aren't sure whether it will adhere to those residues.

Some more references:

- Never use acid-core and how to solder to stainless steel (Kester)
- White residue and all about rosin (more Kester)
- Good mini-class on fluxes: Bolton University
- To clean or not to clean and a brief history of electronics cleaning: more Bolton
- The Kester catalog provides good info on the solderability and flux requirements of various materials (see p.14).
- Flux residues and what to do about them. This explains a bit about cleaning options and the health risks of rosin and non-rosin fluxes.

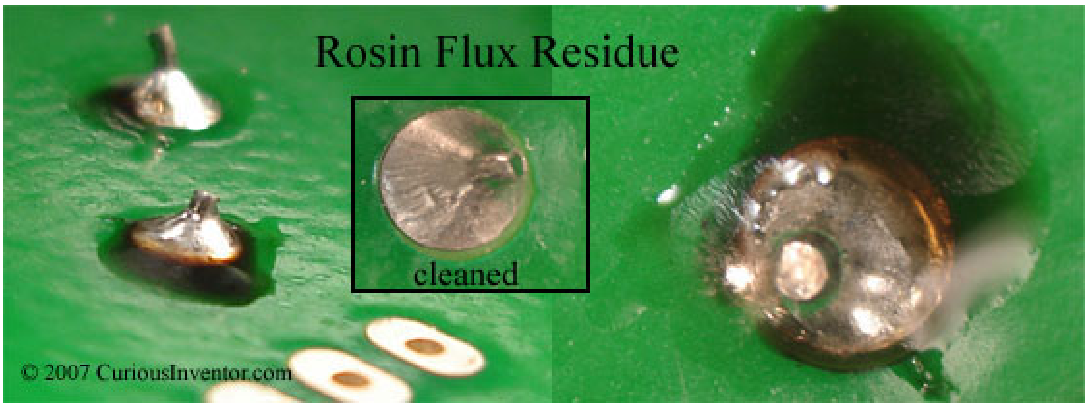

- **Alloy:** 60/40, 63/37, tin-lead, lead-free, silver bearing, RoHS, eutectic, oh my...

- **Standard lead-based solder** is made of tin and lead. When you see 60/40 or 63/37, it means 60% tin by weight, 40% lead. Either once of these alloys should be fine for typical small electronics soldering. 63/37 in bulk is slightly more expensive because of additional tin, but has the special property of being a "eutectic" alloy, which transitions from liquid to solid at one temperature (like water) instead of range of temperatures. Basically, in non-eutectic alloys like 60/40, there is a "pasty" region of temperature where portions of the solder are frozen and other portions are liquid. What does this mean for soldering and is 63/37 really that much better?

  Alloy metals have some interesting properties that are different from the metals comprising them. In tin-lead solder, the mixture has a lower melting point than either lead or tin alone, and the melting point varies depending on the portions. The mixture that yields the lowest melting point is called eutectic. This is also the only mix where all the constituents melt and freeze at the same temperature.

  If the tin-lead alloy isn't eutectic (ie, if it is not 63% tin), it will go through a "pasty" phase while it freezes. Unlike water, which freezes entirely at 0 °C, some parts of a non-eutectic mixture of tin-lead freeze at higher temperatures than other parts. For a somewhat simplified explanation, if you held the temperature of 60/40 slightly above 361 °F, the "extra lead" would solidify and be floating in a liquid 63/37 eutectic mix. For a more exact and great explanation of this process, look here.

  How is it that the mixture of two elements somehow lowers the melting point? And I quote: "increased entropy." Chew on this. (Another great phase-diagram explanation with a bonus of why ice and salt can get almost 30 °F below freezing--enough to freeze ice cream) And one last great explanation--talks a bit about grain structure and how solder isn't a simple homogeneous mixture of tin and lead.

  This compares the difference between eutectic freezing in tin-lead and eutectic freezing in more complex 3 element lead-free alloys.

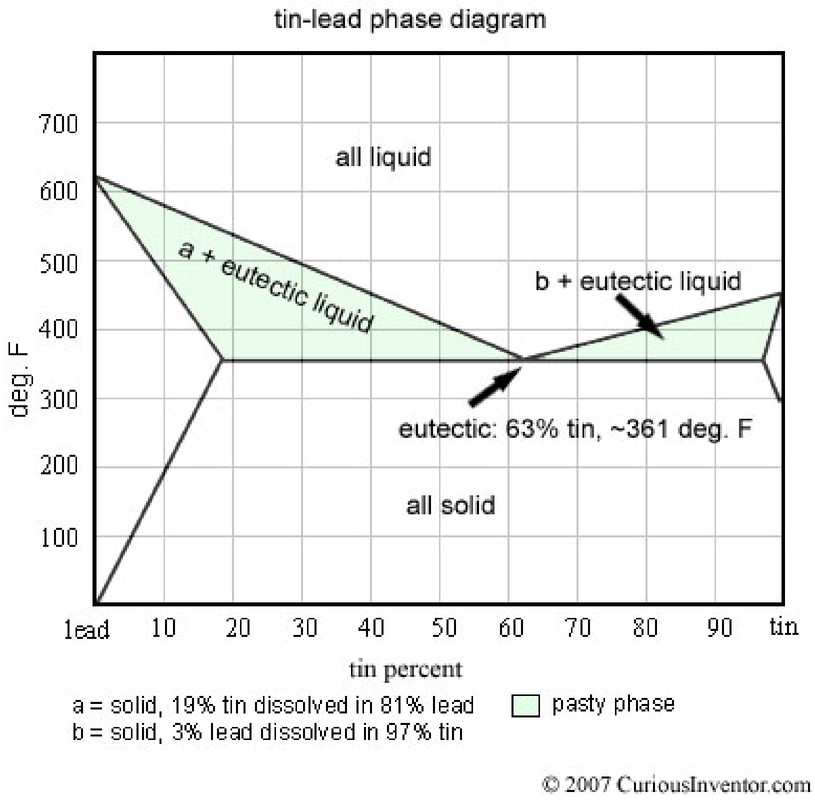

So which is better, 60/40 or 63/37? A decade or so back (before the lead-free movement), most manufacturers incurred about a 5% increase in cost to switch from 60/40 to 63/37. 63/37 flows slightly easier, makes shinier joints, and has a faster total freezing time which means there is less risk the joints will be "distrubed," which is what happens when the joint moves during solidification. This can lead to internal fractures that cause poor electrical connections and unreliable mechanical joints. Note that 63/37 doesn't freeze instantly (just like water)--it still has a window of time during which the joint can be disturbed, too.

I think most would say that the enhanced properties of 63/37 really only matter for mass soldering operations like wave or reflow soldering, and that there is little difference for hand soldering. A dull joint is more often caused by insufficient heat, dirt or oxides, or lack of flux rather than alloy makeup. (A no-clean flux may burn off before the joint is complete) Holding everything else constant, the difference in shininess between 60/40 and 63/37 is completely cosmetic. If you'd like to see a shiny 60/40 joint, try using Kester 60/40 with #44 flux.

- **Silver bearing solder:** (that is, contains silver, not for roller bearings) Silver is used in one of the leading alloys for lead free solder (An96.5% Ag3.0% Cu.5%) and also as an addition to tin-lead solder, usually in the 2-4% range (when you se 62/36/2 this means Sn64Pb36Ag2).

  People claim that it flows better, has a lower melting point, is stronger, and has a higher conductivity. According to Indium's solder wire data sheet, their 2% silver solder has an electrical conductivity that is 11.9% of Cu compared to 11.5% of 63/37 tin-lead solder, a shear strength of 7540psi vs. 6200psi, and a tensile strength of 7000psi vs. 7500psi for 63/37. So, yes, the claims are true, and also mostly insignificant. Silver was initially added to solder to prevent silver platings on component leads from dissolving into the solder ("silver migration") and forming brittle joints. Having silver in the solder reduces migration, so you may want to use it on silver joints. (Note: this logic doesn't entirely make sense to me. If silver getting in the solder caused embrittlement, how does adding *more* silver prevent this?)

  Audiophiles seem to be enamored by 4% silver bearing solder, namely some from WBT, Cardas, and WonderSolder. Are these really better for audio?

  The superior claims include things like higher purity, eutectic alloys, higher conductivity, and better flux. I haven't found any controlled studies showing that a group of people can actually hear the difference, so I'm skeptical. Although additional silver does increase conductivity, the increase is small and the joint distance over which that conductivity applies is also extremely small. For what it's worth, here's a forum discussion that discusses a bit about solder in high-end audio and also a FAQ (scroll down)concerning solder on Cardas Audio's site. Another decent discussion.

  Bad joints made with any solder can create a high-resistance connection, especially if the underlying components were heavily oxidized initially. My advice: if it makes you feel better, get it, but be wary of sellers that don't provide spec sheets.

- **Lead-free Solder:** As of July 1st, 2006, European laws mandated that new electronics be almost entirely lead free. As of yet, there are no US laws (outside CA) mandating the removal of lead, but most manufacturers are switching over for competitive reasons. More on RoHS, WEEE, and lead risks:

  The European Union passed directives in 2003 stating that no equipment sold in Europe should, by July 2006, have more than .1% lead in any homogenous component (like a solder joint). The directives are known as the WEEE (Waste Electrical and Electronic Equipment) and RoHS (Restriction of Hazardous Substances) (wiki). There are corresponding laws in China, Korea, and California. Japan manufacturers actually voluntarily begin switching to lead free years before RoHS or WEEE for competitive reasons. IPC's informative lead-free website: here.

Although only .5% of lead used in the US gets embedded in electronics (verses 80% in batteries), there is concern that the lead from those electronics will leach into ground water supplies from landfills. Why the concern over solder joints when batteries contain so much more lead? For the most part, lead containing batteries are recycled and regulated, whereas electronics are routinely just thrown away. The EPA claims that 1% of municipal waste is electronics. Interestingly, according to this publication by IPC, no studies have found any evidence of lead getting into the ground water from landfills.

Lead health risks: Lead does not get absorbed through the skin, and is actually not present in solder fumes to any appreciable degree (fumes are still bad for you, see fumes section below). The greatest risk of hand soldering with lead comes from ingesting lead by eating or smoking without first washing. Health risks include increased blood pressure, fertility problems, nerve disorders, muscle and joint pain, irritability, and memory or concentration problems. The latest health data indicates that there is no amount of lead that will not be detrimental to health. Google lead or start with this link.

Some great lead-free joint pictures: here.

The most popular lead-free alloy seems to be Tin 96.5% Silver 3.0% Copper .5%. The wiki page on solder mentions several different lead-free varieties.
-AIM lead-free solders.
-Huge list of Indium lead-free solders and their properties.

Lead-free solder generally melts at a higher temperature, and doesn't wet as quickly to metals. (Eutectic tin-lead solder melts at 361 °F and the SnAg3Cu.5 melts at 423 °F.) Manufacturers generally recommend setting soldering iron temperatures between 700-800 °F for lead-free instead of 600-700 °F for tin-lead soldering. The 15 Watt RadioShack® iron I had operated a bit below 500 °F, so soldering should be possible with it, but maybe slow. Technique wise, since lead-free wets slower, joints will take longer (upwards of 4-7 seconds), but this doesn't mean the soldering iron temperature should be turned up excessively--patience is better than higher temperatures. If you're going to use a lead-free solder, get a flux that's designed for the higher temperatures--the regular no-clean fluxes will likely burn off before doing their job.

A quick word on reliability. Some say that lead-free joints are stronger, and while the material is indeed stronger, it's less flexible than lead-based solder, so expansion and contraction due to temperature change has been shown to break components held by lead-free solder. It seems true mechanical reliability of lead-free vs. lead depends heavily on the situation (see p.30). There is also concern of something called "tin whiskers." These are extremely thin crystalline growths that grow perpendicularly out from surfaces. These took down some space systems and NASA has a great page here. These are different from dendrites (which grow on the surface) and appear to be more likely on bright all-tin platings. Most component platings used to consist of a tin-lead mixture, and since all-tin platings are a common lead-free replacement, people are concerned. I have yet to find any literature that points to TinSilverCopper solder as a risk factor, though. I believe it is a plating issue, only.

One more link: -Why tin or silver?

- **Thickness and Amount:** As a general guide, .032" thick solder (21 gauge) should be suitable for through hole soldering and some surface mount soldering. For finer pitch surface mount devices, use .02" or .015", and if you're soldering a lot of switch terminals, or tinning thick gauge wire you may want .05". If you use .015" solder consider having some thicker solder on hand to re-tin your tip, since the amount of flux in .015" may not be enough to remove tip oxides. The picture below shows how the various thicknesses compare next to the standard .1" spaced DIP pins.

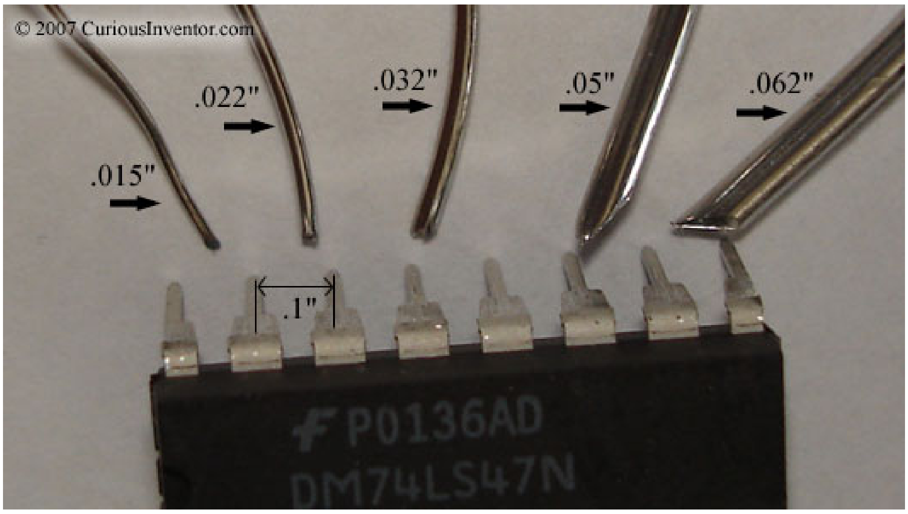
© 2007 CuriousInventor.com
.015"
.022"
.032"
.05"
.062"
.1"
P0136AD
DM74LS47N

Expand to see how .032" and .015" solder compare to a SOIC surface mount chip and fine pitch (.02") device.

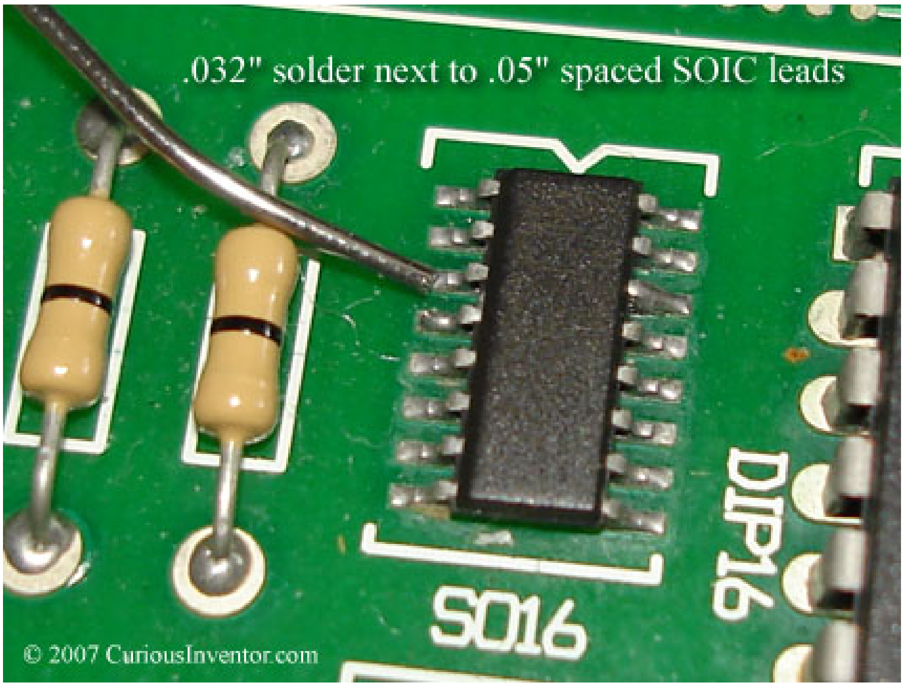
.032" solder next to .05" spaced SOIC leads
SO16
DIP16
© 2007 CuriousInventor.com

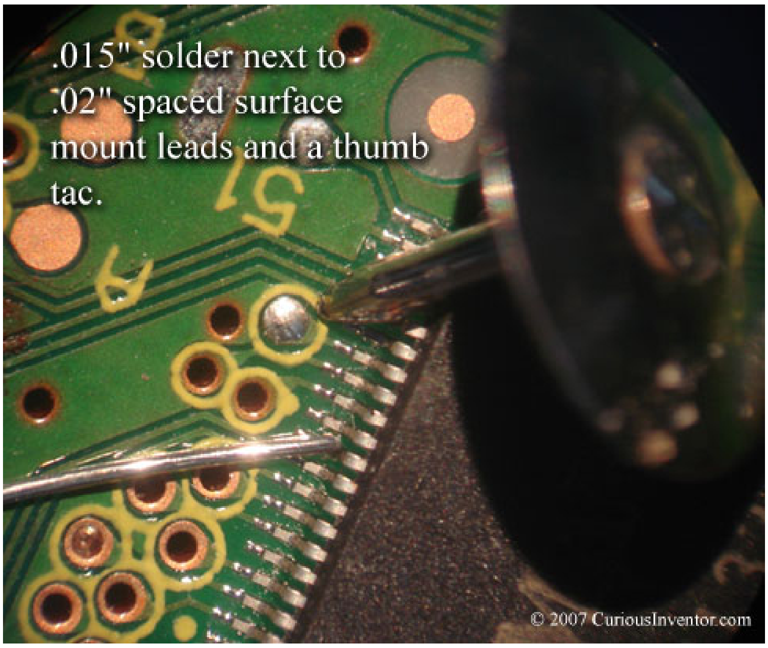
.015" solder next to .02" spaced surface mount leads and a thumb tac.
© 2007 CuriousInventor.com

How much solder do I really need? An ounce? A pound? How long will a pound last?

To get a very approximate feel for just how much solder is in an ounce vs. a pound, I measured how much .032" solder it took to attach ten 16 pin DIPs. Trying to provide an upper bound, the soldering is excessive, and a generous glob was placed on the tip in between each chip to account for solder used to tin and protect the tip during normal use.

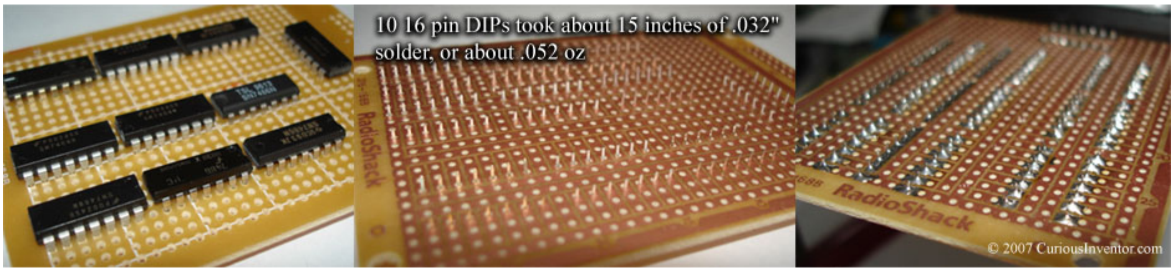
10 16 pin DIPs took about 15 inches of .032" solder, or about .052 oz
RadioShack
RadioShack
© 2007 CuriousInventor.com

The ten chips took about 15" of generously applied .032" solder that weighed in approximately at .052oz. This particular solder weighed about .00348oz/in. At about 3.27e-4 oz/joint, a 1 oz spool should last 3060 joints, a half pound should last 25 thousand joints, and a pound should last about 50 thousand joints. Mileage will certainly vary with different sized solders, joints, tinning wires, and highway vs. city driving, but if you're not in a production environment, a half pound should last a while.

Solder Fumes:

What is exactly in solder fumes? Am I safer using lead-free solder?

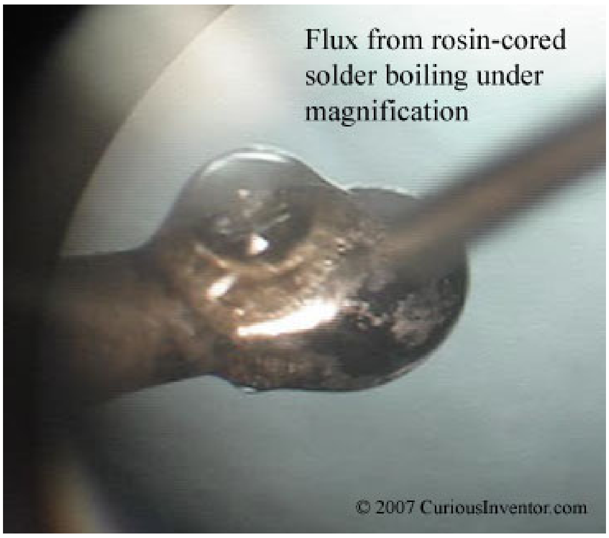

Lead boils at over 3000 °F, and in most cases soldering tips should be kept below 750 °F, so it is highly unlikely that gaseous lead is present in the fumes. The fumes are actually from the flux boiling, which still isn't great for you--many of the chemicals found in cigarette smoke are found in flux fumes: formaldehyde, toluene, alcohols, and hydrochloric acid to name a few. Most of the public health literature indicates that asthma is the major health risk from soldering fumes (not cancer or lead poisoning). When acquired, it is permanent and can cause hyper sensitivity so that even small amounts of fumes bring on attacks. Surprisingly, scientists have not been able to determine what exactly in the fumes cause the health defects, nor what amounts are harmful. Yet, the British health department has set exposure limits of .05 mg/m^3 over 8 hours and .15 mg/m^3 over 15 minutes. I believe these limits have been shown to provide a safe work environment and also one for which the necessary systems / filters are financially reasonable.

Some informative links:

- Solder Fumes and You A British health department pamphlet explaining the health hazards of rosin-based flux fumes (irritation, headaches, dermatitis, asthma) and what precautions employees and employers should take. Note the total lack of any mentioning of lead poisoning.
- Workplace Exposure to Rosin-based Solder Flux Fume During Hand Soldering A study done by the UK Health and Safety Laboratory measuring exposure levels and also the effectiveness of various exhaust, ventilation, and filter systems.
- Measurement of the Performance of Air Cleaners Against the Particulate Element of Rosin-based Solder Flux Fume Another UK Health study investigating the effectiveness of various fume extraction and filter systems. Most interesting finding: although activated carbon filters can remove gaseous hazards, they are largely ineffective for fine particulate in the fumes which they believe to cause much of the harm. Some combination of carbon and HEPA filter is needed, and even these are useless without sufficient air flow.

Returning to the topic of lead, it is widely agreed that eating, smoking and drinking without first washing is the greatest risk factor. Despite the high boiling point of lead, there is also agreement that at least a small amount of lead particles are indeed present in the fumes. The conspicuous lack of emphasis on lead poisoning in all the research done by the UK health department implies that these particles are of little concern. Sentry Air Systems has a brief page that is one of the very few sources I found to claim that lead particles under normal soldering conditions are harmful. The company sells fume extraction technology.

The material safety data sheet for Kester #44 cored solder says under the fire fighting section: "Melted solder above 1000 °F will liberate toxic lead and/or antimony fumes."

According to IPC's DVD-11, "General Safety in Electronic Assembly," when solder is heated past 850 °F the lead can become atomized and end up in the fumes. video link (if link breaks, you may have to search for DVD-11 at www.ipc.org).

Useful comments from someone in the manufacturing world regarding lead.

It would seem that, for typical lead-based, rosin cored solder, the risks are probably not that great from the fumes if you only solder occasionally, don't use abnormally high temperatures, and are in a well ventilated area. If ventilation isn't too good, and you're soldering for long periods of time, the cheaper foam-type carbon filters may not be good enough.

But what about lead-free solder? Lead-free solder often requires higher temperatures and more active fluxes, and both of these factors lead to significantly worse fumes.

Fume Extraction Becomes More Important in a Leadfree Environment - from the Weller blog

Another excellent article on the increased risk of lead-free fumes from OK International.

Instructables has all sorts of home-made fume extractors.

Prepare the Work

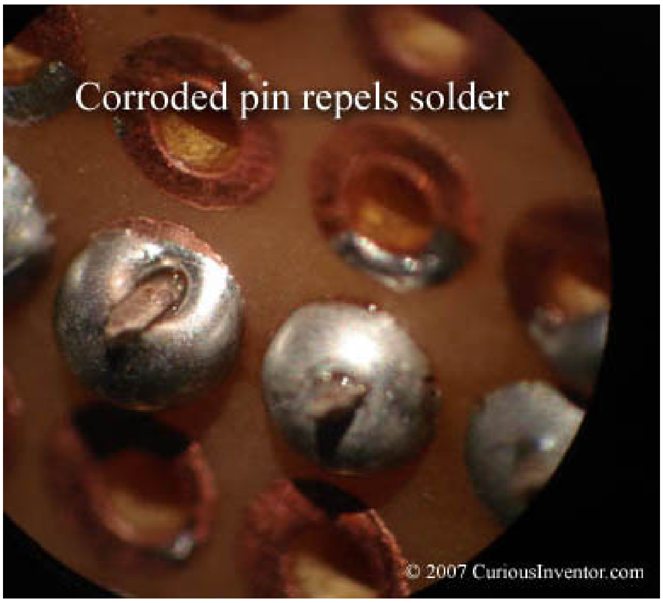

- **Start with clean components:** Flux can remove small amounts of oxides, but will be of little help for heavy oxidation, grease, oil or dirt. Notice how the solder in the adjacent picture has been repelled by the heavily oxidized pin. It may be necessary to lightly use steel wool or fine grit sand paper to remove especially bad oxides. Some people say that you should not do this because it creates scratches that can promote future oxidation... sand at your own risk. Use Silicon Carbide sandpaper (black) as opposed to Garnet (brown, for woodworking) sandpaper because the Garnet paper will shatter and become embedded in the metal. An effective and gentle alternative is to use a pink eraser, especially for copper traces.
- **Clamp your work:** PanaVise makes a popular clamp that accepts several different attachments for holding different sized circuit boards. It's by far the most popular clamp and is also very sturdy. Having the work held in place is especially helpful for desoldering when it's necessary to push or pull a bit. The alligator hands are a cheap alternative.

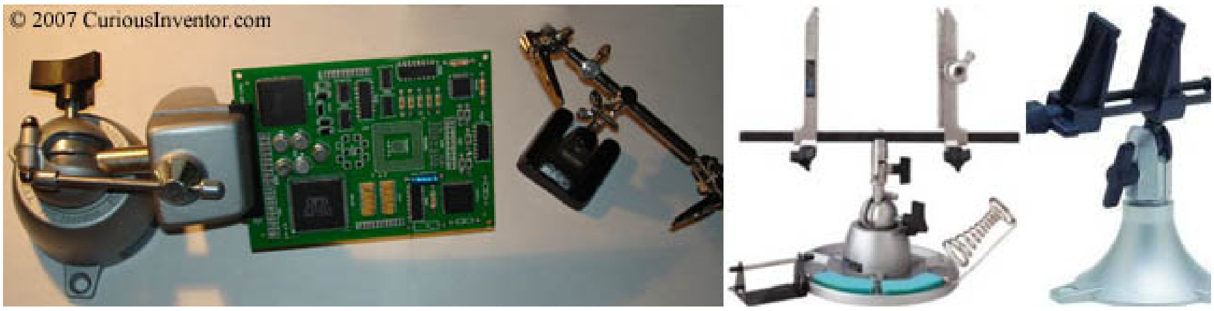

- **Wire preparation:** Tin stranded wires so they don't "bird-cage," or bend out from their original lay. Expand for instructions on the correct way to strip a wire manually, use an automatic stripper, and tin wire.

- **Manual wire stripping:** The natural way you would think to use the wire strippers is to first cut the insulation, and then use the same cutting hole to push off the insulation slug. In production environments where reliability is critical, this is not allowed because it's very easy to nick the wires this way. The pliers at the tip of the stripper are actually meant to remove the insulation slug. First score the insulation with the cutting hole, and then pull it off with either the pliers or by hand. If removing by hand, twist slightly while pulling to keep the strands in place. This is the "right" way to strip a wire, if you're careful, pulling the insulation off with the cutting hole is probably fine.

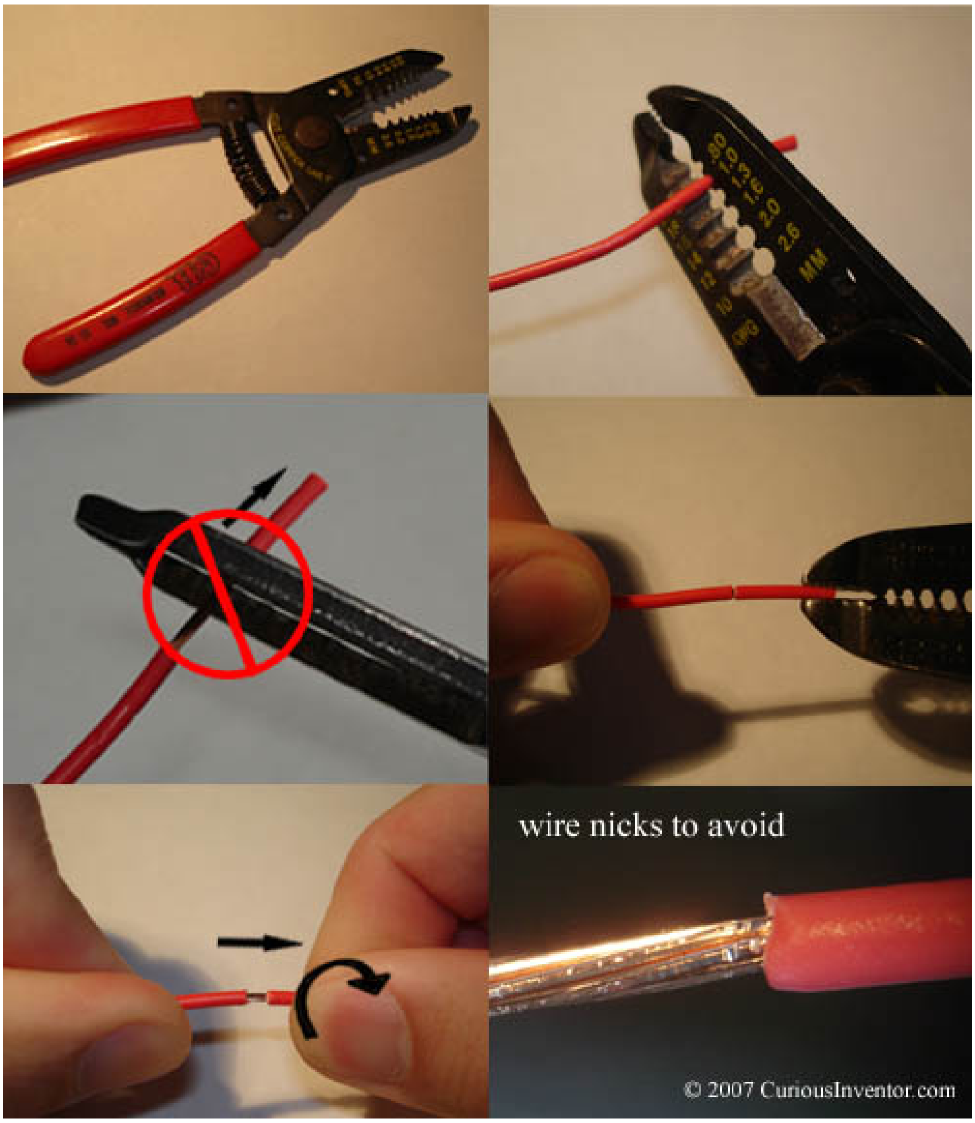

- **Automatic stripping:** With one squeeze these cut and pull off insulation.

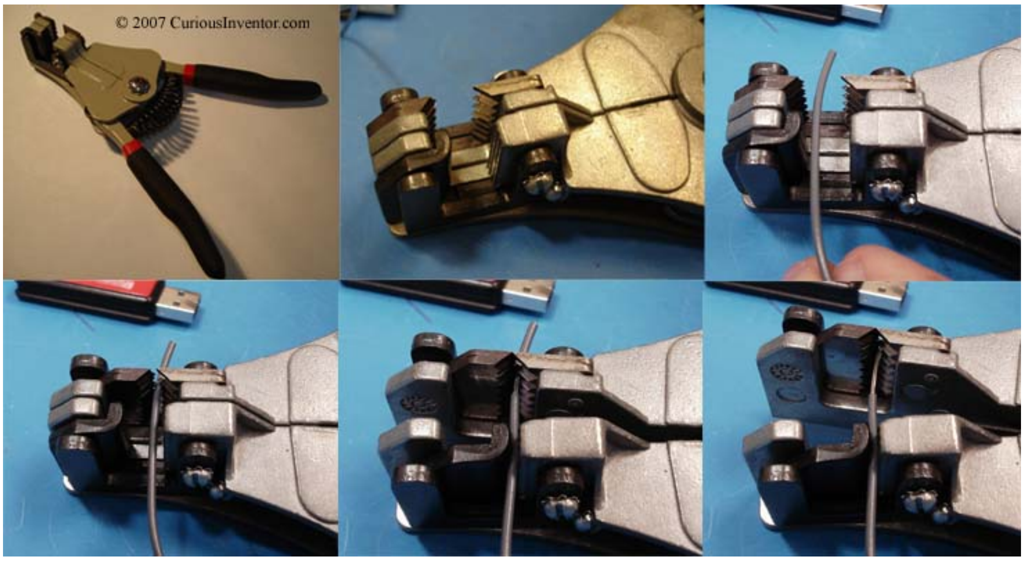

- **Tinning:** Just like soldering a joint, the key here is to apply solder to the wire, not the iron tip. If the wire is hot enough to melt the solder directly, then it's hot enough to form a good bond with the solder. Hold the iron against one side of the wire and apply a small amount of solder in between the tip and wire to form a heat bridge (if there isn't already enough solder from tinning the tip). Now apply solder to the opposite side. The solder should wick into the strands. Move the iron and solder as necessary to coat the surfaces. You should still be able to clearly see the individual strands when it's complete.

  Don't tin wires that are going to be held in clamping terminals since the solder will slowly deform under pressure and cause the connection to come loose.

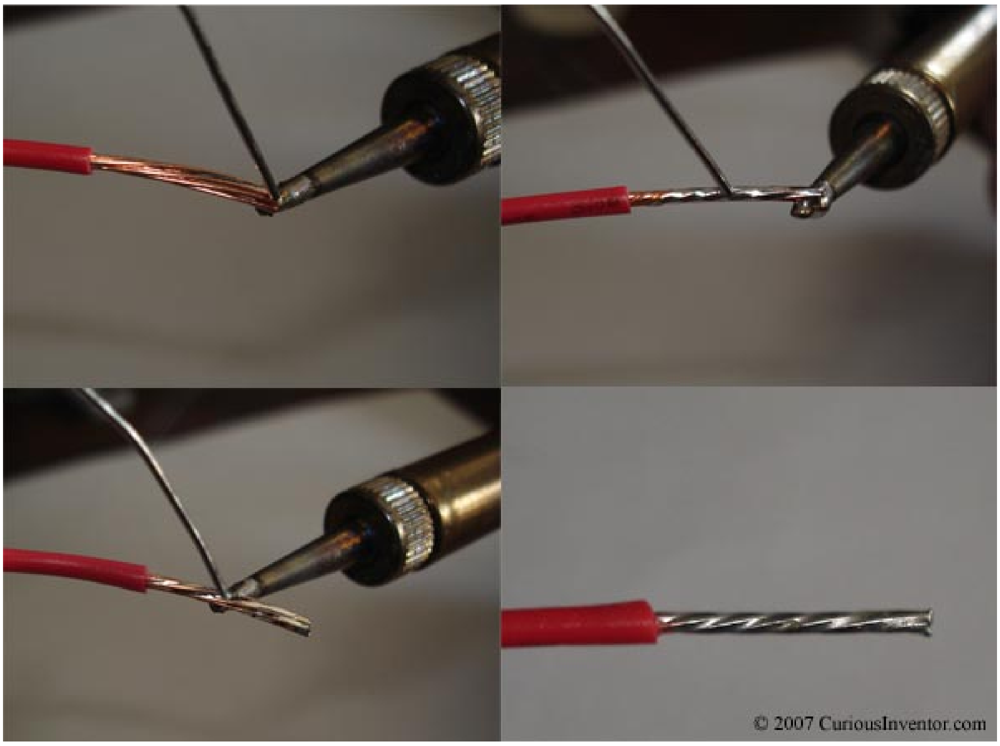

  Got a lot of tinning to do? For under $300 you can get a vat of molten solder to speed up the process. The only catch is that there's no flux in the vat, so you need to dip the end of the wires in liquid flux before dunking them in the molten solder. Only dip the very end of the wire in the flux; capillary action will draw it further up the strands. If you dipped more than just the end into the flux, the capillary action would draw flux beyond the insulation where it can't be cleaned, and would eventually corrode the wire.

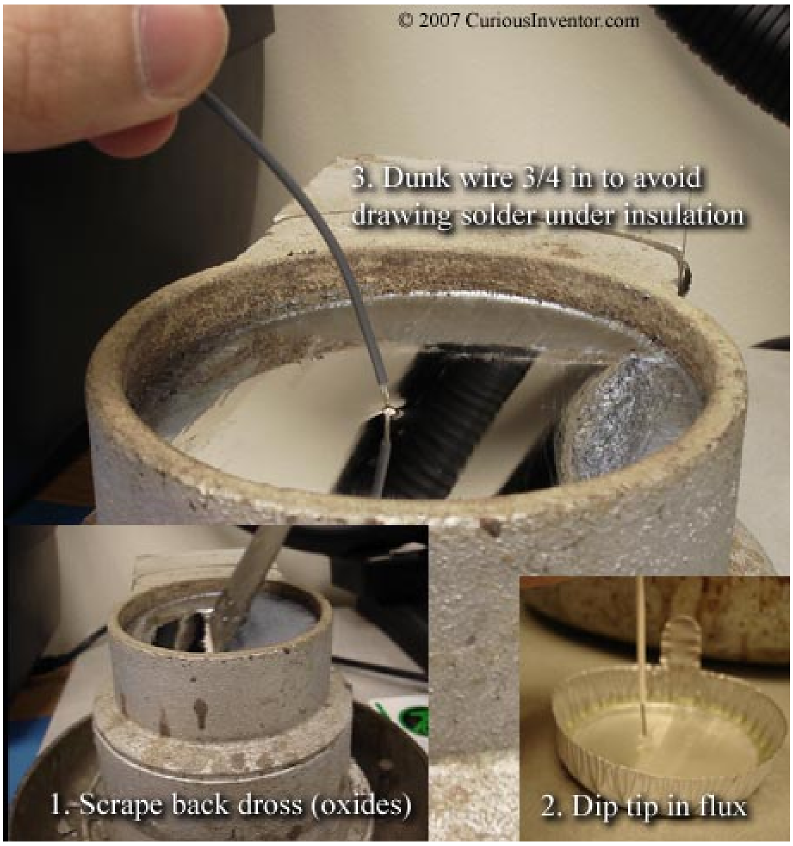
© 2007 CuriousInventor.com
3. Dunk wire 3/4 in to avoid drawing solder under insulation
1. Scrape back dross (oxides)
2. Dip tip in flux

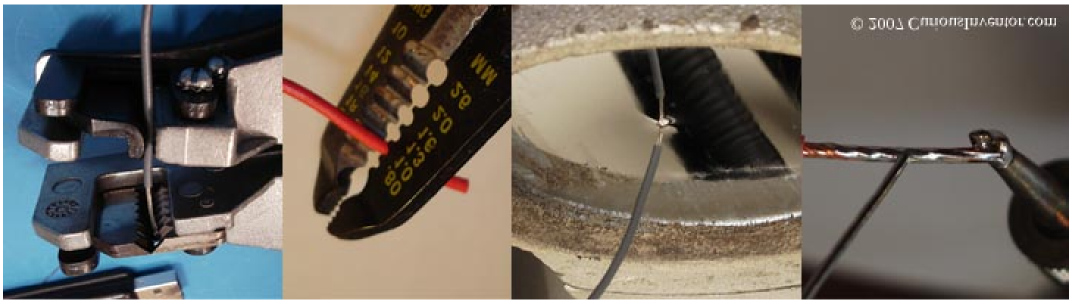

- **Insert, clinch and trim components:** First, make bends *before* inserting the components. Avoid stressing the connection between lead and component by bracing the lead with pliers while bending. Pliers with serrated tips aren't used in high-reliability production because the grooves can create nicks in the leads that eventually cause a break after a lot of vibration and thermal changes. Round nose pliers make it easy to make any sized radius.

  Unless the component has a metal casing or needs clearance for air flow to keep cool, insert it until it's flush with the board. This doesn't apply to some transistors, and also capacitors that have plastic coverings that need to be kept out of the solder joint. Clinch or bend out the leads so the component is held in place during soldering, and finally trim the leads to about the radius of the pad. Trim the leads before soldering since doing so afterwards can shock and crack the joint. Wearing safety glasses for this process is not at all excessive--those leads can get you. Everything else about proper component installation: NASA guide.

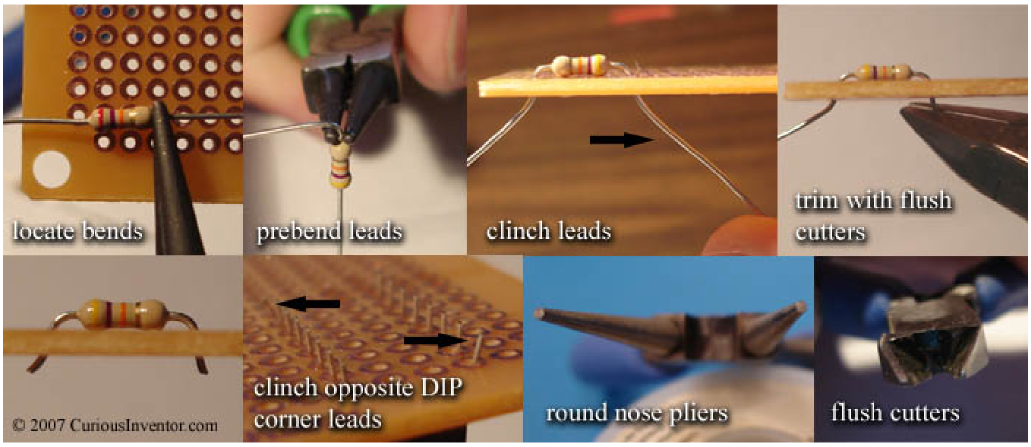
locate bends
prebend leads
clinch leads
trim with flush cutters
© 2007 CuriousInventor.com
clinch opposite DIP corner leads
round nose pliers
flush cutters

- **Add heat sink:** Some semi-conductors (some transistors and diodes) are especially heat sensitive. This clip acts a heat shunt to keep the transistor protected.

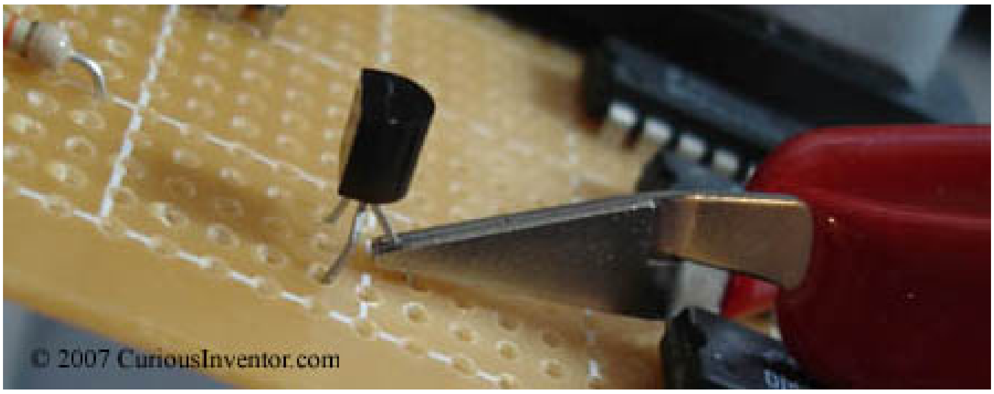
© 2007 CuriousInventor.com

Clean and tin the Tip

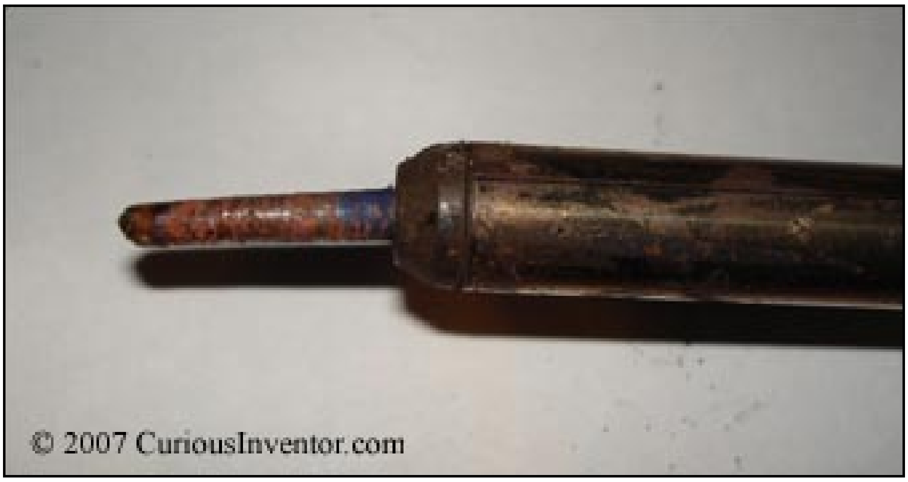
© 2007 CuriousInventor.com

- **Regular cleaning = easier soldering and longer tip life:**The iron tip's ability to transfer heat is drastically reduced when it gets covered in oxides and burnt flux residues. Not only does heat not transfer as well through this debris, but the contaminants also prevent solder from wetting or sticking to the tip. Most heat transfer actually goes through a fluid solder "heat bridge" that lies between the iron tip and components, so an iron tip that repels solder will be very ineffective.

The longer oxides and charcoaled flux residues remain on the tip, the harder they become to remove, so it's a good idea to clean the tip every time you pick up the iron.

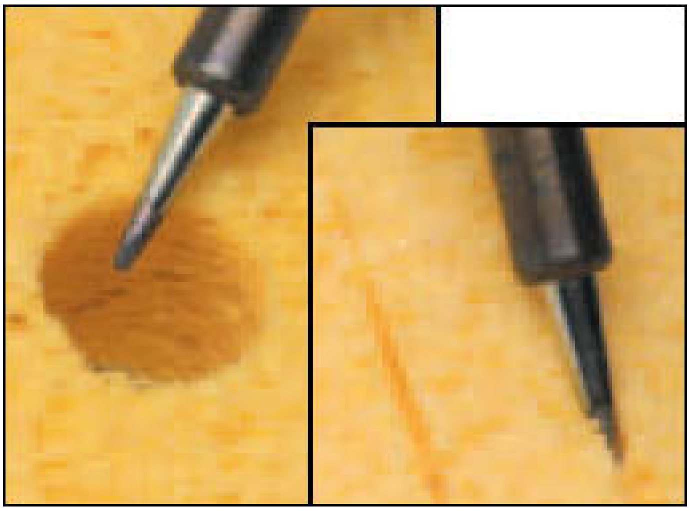

Wiping the iron on an edge of a hole cut into a sponge can help to remove oxides easier, and also allows waste to fall away. A dry cleaner can also be used. It consists of soft metal shavings that are coated with flux. You clean by thrusting the iron into the shaving a few times. By avoiding the thermal shock of touching a damp sponge, these cleaners help to increase tip life, and in our opinion, do a better, faster job.

HAKKO 599B
TIP CLEANER
© 2007 CuriousInventor.com

Usually touching the tip with rosin-cored solder will supply enough flux so that oxides can be removed with a damp sponge. If this isn't sufficient, you can purchase "tip tinners and cleaners" that are a mixture of solder paste and flux. The flux is oftentimes stronger (more activated) to help remove oxides.

64-020
RadioShack
Tip Tinner and Cleaner
WARNING: KEEP OUT OF REACH OF CHILDREN. Read carefully other warnings on back of card.
Custom Assembled in USA with Foreign and Domestic Components for RadioShack Corporation, Fort Worth, TX 76102
0.5 oz. (14.2gm)
© 2007 CuriousInventor.com

Finally, when that doesn't work, special polishing bars to can be used to salvage extremely bad tips. Another last resort is to gently rub the oxides off with an emery cloth or soft steel brush. Cover the tip immediately with solder after cleaning to prevent further oxidation. Never file the tip to clean it or form a different shape. The tips are mostly copper with a protective iron plating, and once that plating is pierced, the tip will die quickly. Copper is used because it's an excellent heat conductor, but if exposed to solder, it will quickly dissolve into the solder.

Weller®
WPB1
POLISHING BAR FOR
SOLDERING TIPS

© 2007 CuriousInventor.com

Tin the tip: Add a small amount of solder back onto the tip. This helps to protect the newly cleaned and exposed tip, and also helps to transfer heat to components.

Heat and Solder the Joint

- **Heat the joint:** Place the iron tip so that it touches both the component lead and pad--the goal is to get as much surface area contact between the iron tip and joint as possible. Almost no heat will travel through the point.

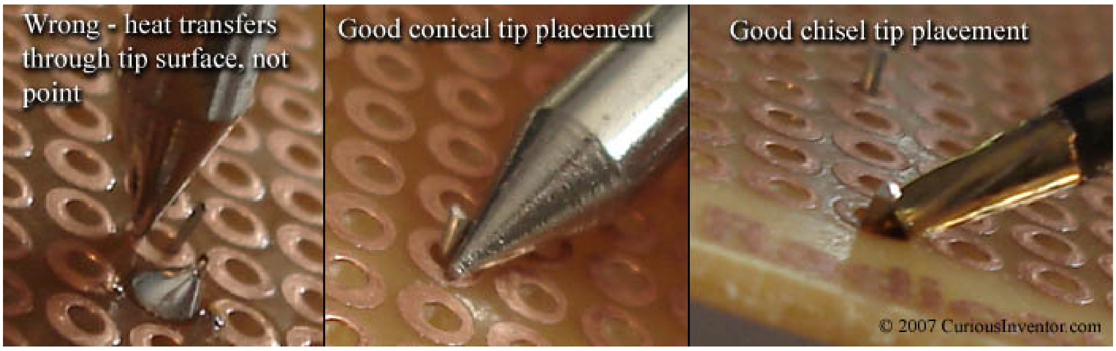
Wrong - heat transfers through tip surface, not point
Good conical tip placement
Good chisel tip placement
© 2007 CuriousInventor.com

- **Make heat bridge:** Add a small amount of solder between the tip and the work--heat transfers much faster through the liquid solder than dry surface contact. This is why a tip that won't "wet" is so difficult to use. Pressing hard should not be necessary. This step may not be necessary if there's enough solder already on the tip from tinning it after cleaning.

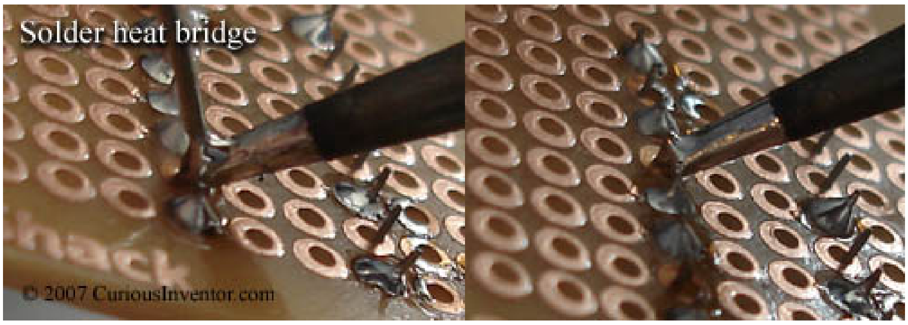
Solder heat bridge
© 2007 CuriousInventor.com

- **Apply solder to opposite side:** Apply solder to the parts, *not* the iron. By doing this, you ensure the parts are hot enough for the solder to "wet" and bond with them. Also, solder will run towards the heat source, so applying solder opposite from the iron helps to spread it out and cover the joint.

  For larger joints, rather than dumping in all the solder quickly, continuously pulse in small amounts to keep a fresh supply of active flux available.

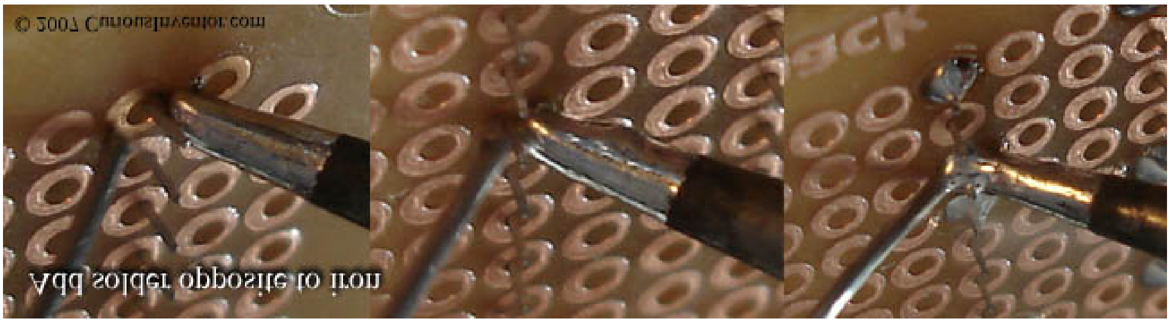
© 2007 CuriousInventor.com
Add solder opposite to iron

- **Time:** The joint should take about 2-5 seconds total time for standard 60/40, 63/37 lead based solder and a non- no-clean flux, and up to 7 seconds for lead-free solder. Lead-free solder just takes longer to "wet" the metal.

  Wetting is how easily and quickly solder spreads out over a surface. A water droplet on a freshly waxed car shows poor wetting, as does solder on a heavily oxidized soldering iron tip. It basically comes down to how attracted the liquid molecules are to each other verses the surface (see surface tension). In industry, tests are done to determine the "solderability" of materials by measuring the time it takes solder to spread out over a surface, or measuring how much force a pot of liquid solder will pull down on a component partially submerged.

Good pictures and description of wetting and surface tension. Contains a video of a razor blade floating on the surface of water until a drop of soap is added.

A brochure for a solderabiliy testing station. Scroll down a few pages to see some great pictures of this machine holding one lead of a surface mount chip in a drop of solder.

Another related wiki: wiki/Surface_energy

In general, the goal is to make the joint as quickly as possible. Longer times can char and damage the board, lift pads, overheat components, burn off and polymerize flux (making it harder to remove), and finally lead to a more brittle joint. Solder doesn't just freeze on a joint, tin in the solder dissolves and chemically reacts with copper in the connection to form a new bonding material, called an "intermetallic layer". While this layer is what makes an excellent thermal and electrical bond, it is also extremely brittle; a doubling of its thickness reduces joint tensile strength by half (ref 1). Since this layer grows faster with higher temperatures, joints should be made using the coolest temperature and shortest soldering time possible. This layer is also why re-heating joints has been shown to weaken them. Having said all this, I have to admit that I don't know just how long is *too* long for projects that don't need to operate for 30 years with 100% reliability. After 10 seconds there's a good chance the flux has been used up.

- **Remove solder, then iron:** Pull the iron out fairly quickly to avoid leaving a solder spike.

© 2007 CuriousInventor.com

- **Good and bad joint gallery:** The solder should smoothly ramp to meet surfaces and be shiny in appearance if it's lead-based. Lead-free solder will have a duller and grainier surface, but will still be a good joint as long as there are no signs of non-wetting. The important thing to look for is any solder that looks like it didn't cling to a surface, or is just sitting on top or next to a surface.

Gallery of joints:

A NASA gallery of every possible joint / board defect you could ever imagine. Here's another great gallery of defects.

Great comparison pictures between lead-based and lead-free joints. p. 34

Good joints:

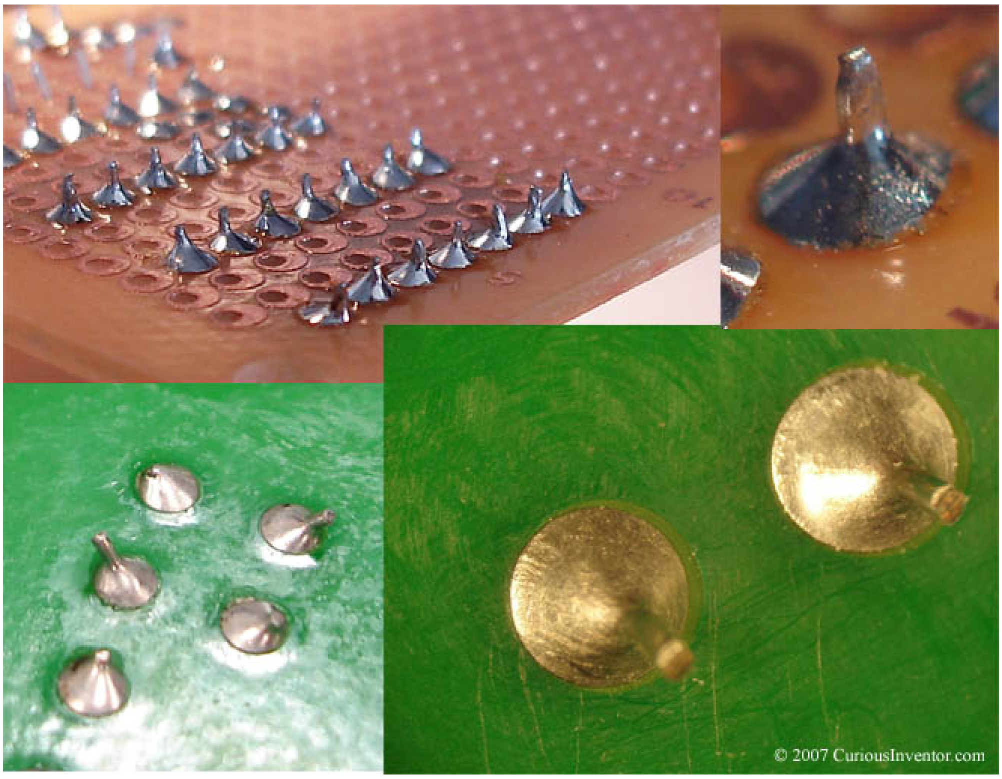
© 2007 CuriousInventor.com

Great joints:

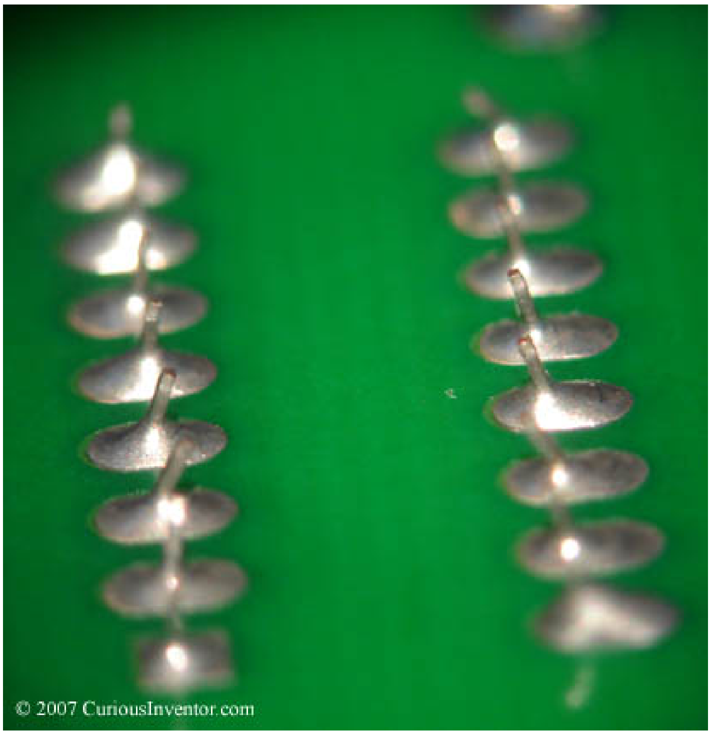
© 2007 CuriousInventor.com

What's the right amount? A large amount of solder is not needed, just enough to cover the pad and lead without any gaps (actually, only 270° is required by IPC J-STD-001). The measure for too much solder is whether or not you can see the outline of the lead in the solder. This is important because you need to be able to see whether the solder adhered or "clung" to the lead (indicating a true bond), and didn't just freeze around it. This applies to all types of joints: tinning wires, soldering to connectors, surface mount components... the solder should never completely hide the underlying wires or leads.

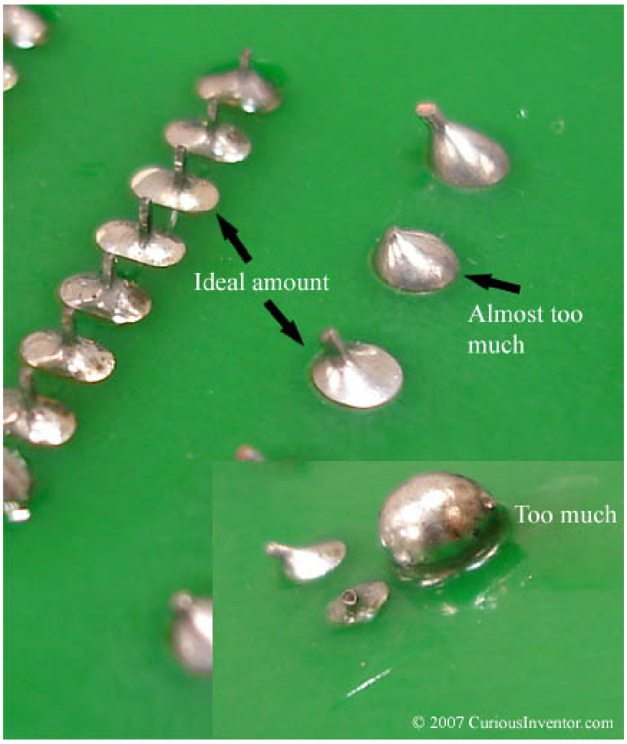
Ideal amount
Almost too much
Too much
© 2007 CuriousInventor.com

Disturbed joint (bad): If the component moves during solidification, the internal structure of the solder will have fractures, leading to a high resistance or unreliable electrical connection, as well as a fragile mechanical one. The solder also appears dull and grainy--typical signs of a "cold" joint that doesn't actually bond with the underlying surfaces.

"Cold" joints are often formed when the underlying pad or lead didn't get hot enough for the solder to wet it.

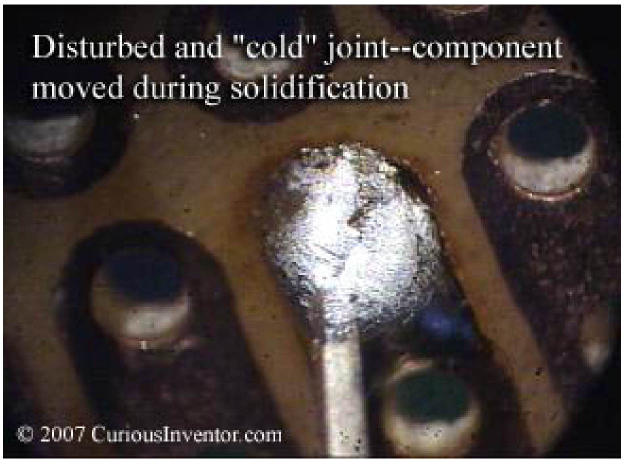
Disturbed and "cold" joint--component moved during solidification
© 2007 CuriousInventor.com

Awful joints:

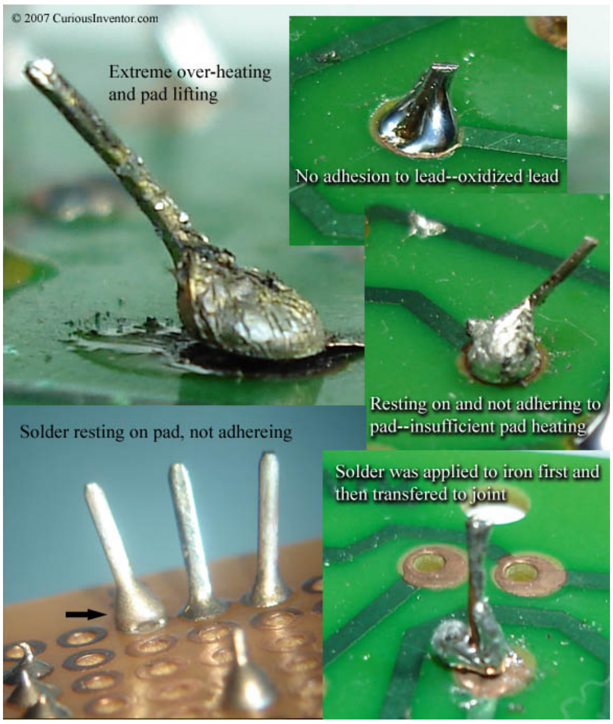
© 2007 CuriousInventor.com
Extreme over-heating and pad lifting
No adhesion to lead--oxidized lead
Resting on and not adhering to pad--insufficient pad heating
Solder resting on pad, not adhereing
Solder was applied to iron first and then transfered to joint

Clean Up

Most manufacturers will clean off residues from any flux that isn't labeled "no-clean," despite flux datasheets like Kester's that say even some of the more active fluxes do not need to be cleaned. For short life-span hobby projects, it probably doesn't matter unless you're using a solder/flux labeled "organic" or "water-soluble"--these fluxes leave behind very aggressive acids that will quickly eat away circuits. Cleaning may be necessary if you're applying a protective coating that won't adhere to flux residues. Finally, some rosin residues are tacky and may attract dust that can short a circuit.

The fact that a flux is made from rosin doesn't tell you much about how strong it is or whether it should be cleaned. What matters is how concentrated the mix is and how much acidic (halides) activators were added. RadioShack® doesn't supply any information on the flux in their standard rosin-cored solder, but it's probably weak enough that the residues do not need to be cleaned off.

Isopropyl alcohol works decently on rosin-based residues, but clean shortly after soldering because the residues quickly harden. Use water for water-soluble fluxes. This pump containing bottle dispenses a little alcohol when you push down on the top with a brush, and keeps the rest from evaporating. If you are going to clean, make sure you wipe up the remnants with a lint-free cloth--don't just spread them around the board with a brush and alcohol.

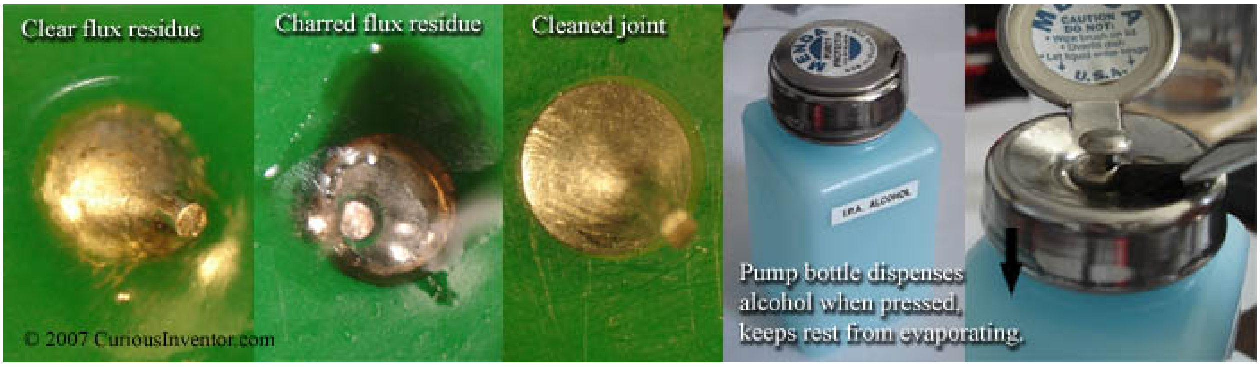
Clear flux residue
Charred flux residue
Cleaned joint
I.P.A. ALCOHOL
CAUTION DO NOT!
U.S.A.
Pump bottle dispenses alcohol when pressed, keeps rest from evaporating.
© 2007 CuriousInventor.com

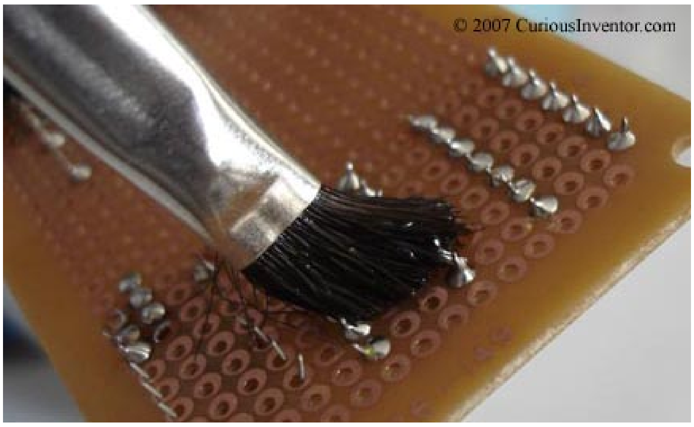
© 2007 CuriousInventor.com

Desoldering

The three most common inexpensive ways to remove solder include a "solder sucker", solder wick, and an iron with an attached desoldering bulb.

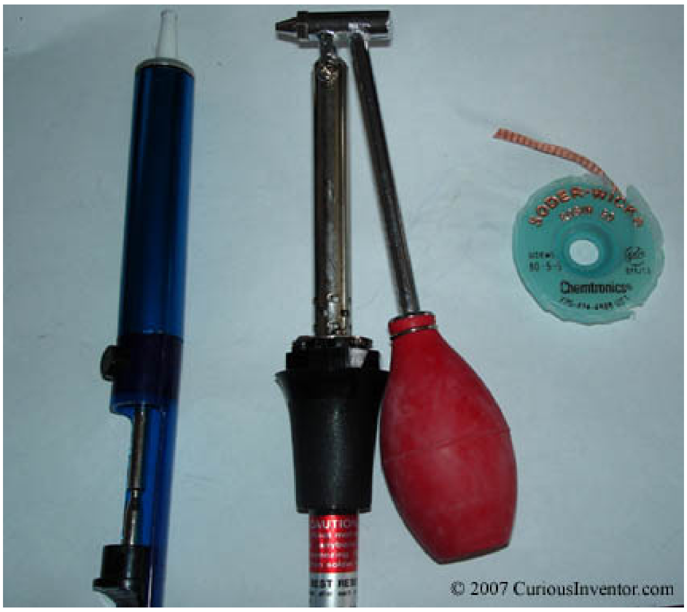

© 2007 CuriousInventor.com

- **Solder Sucker:** To use, press in the plunger, heat up the joint, and as quickly as possible, pull out the iron, place the sucker's nozzle over the joint and press the release button. It may take a few iterations to clean out a joint enough to wiggle free a component. On plated through holes it may actually be easier to add *more* solder to the joint before sucking, since the sucker won't be able to create a vacuum with too little solder remaining.

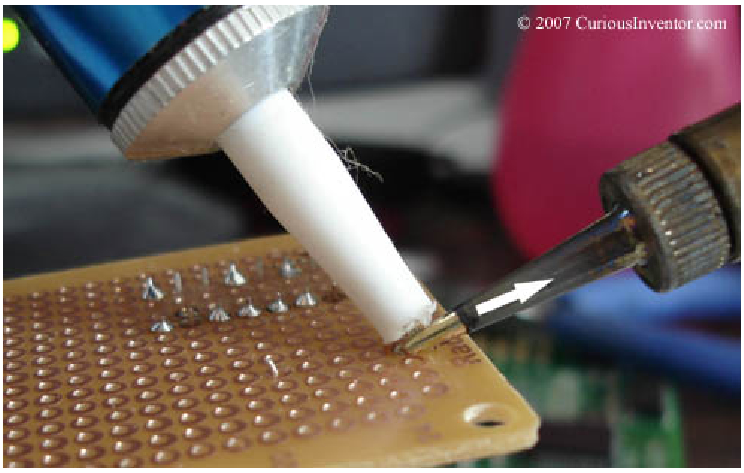

© 2007 CuriousInventor.com

- **Desoldering Iron:** This works just like the solder sucker, but is much easier to use because there's no race to suck out the solder before it freezes, as with the "solder sucker".

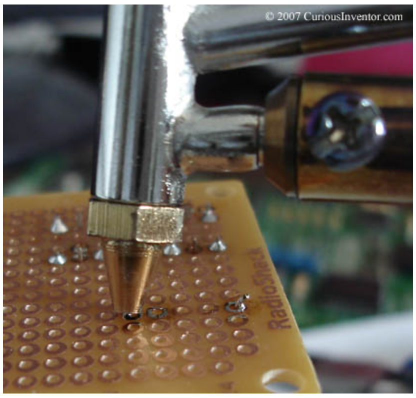
© 2007 CuriousInventor.com
RadioShack

- **Solder Wick (desoldering braid):** The above two tools will often not be able to completely remove all the solder. Many people use the sucker devices above to remove most of the solder, and then get the rest with solder wick, which absorbs solder through capillary action. It's a fine weave of copper wires that are coated in flux. Place it over a joint and then heat from the opposite side with a soldering iron. It may help to have a bit of solder already on the iron's tip to speed heat transfer.

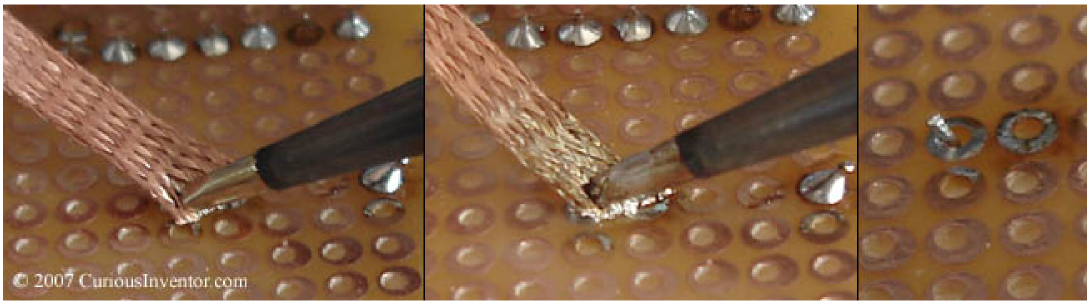
© 2007 CuriousInventor.com

Surface mount desoldering: Surface mount chips are especially hard to desolder because it's very difficult to completely remove all the solder pin by pin, and avoid overheating the board and lifting a pad. Professional shops use expensive hot air guns or special tips (shown below) to heat all the joints at once.

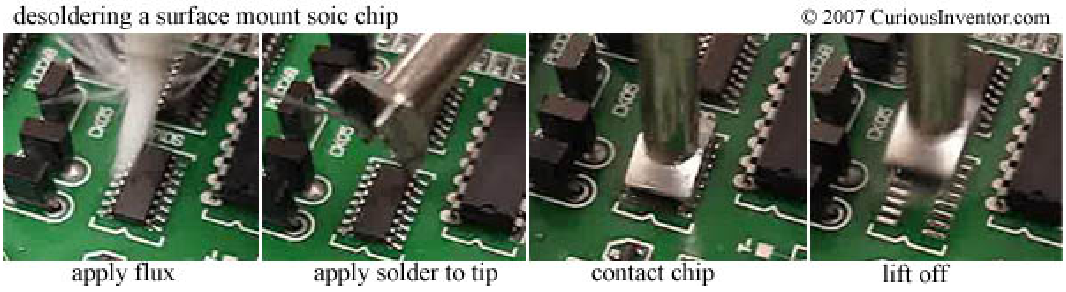
desoldering a surface mount soic chip
© 2007 CuriousInventor.com
apply flux
apply solder to tip
contact chip
lift off

There are fortunately a few cheap ways to desolder surface mount chips.

- ChipQuik provides an interesting solder that when melted over existing joints produces a new low-melting point alloy (under 200 °F) with a much longer solidification time. The longer solidification time enables you to melt all the joints at once and then flick off the chip.
- There are many guides for making a DIY hot air gun with the RadioShack® desoldering iron and a fish tank air pump. Engadget has one of the better how-to guides. Here and here are some more plus the supposed original.

The Ersa soldering primer

Soldering made easy

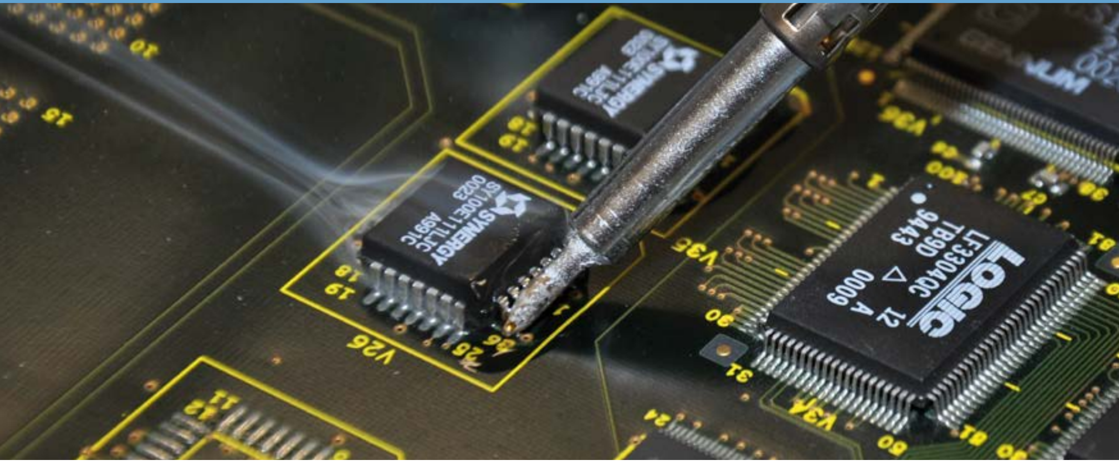

kurtz ersa

Soldering – a never-ending story

for more than 5,000 years already

Man had scarcely learned how to use metals for his purposes when the desire to join them arose in him. Many of the pieces of jewelry, tools and weapons we know from the Bronze Age owe their usefulness and beauty to the art of soldering.

ersa
RDS 80
Power
350 °C
150
300
400

Today, it is difficult to say who first discovered how to "glue" metals. One thing is certain - the goldsmiths of ancient Egypt knew how to join gold and silver already more than 5,000 years ago. Their colleagues in Troy were also master craftsmen long before the ancient Teutons could even dream of such handicraft. Soldering really "came of age" when tin was discovered as a joining metal. And that was, after all, already 4,000 years ago!

From then on, soldering technology was on its way. It first spread around the Mediterranean. The Cretans showed it to the Etruscans who then taught it to the Romans, Tunisians, Spaniards, followed by many others, including the less developed cultures of the time – the Swiss, Bohemians, Hungarians, Teutons and Scandinavians. From culture to culture, from generation to generation, the craft of soldering was continuously improved and refined.

The ancient Romans already laid down and soldered 400 km of leaden water pipes, conjured up stoves and bathtubs from bronze sheets, not to mention the excellent craftsmanship of their armorers and goldsmiths. Apart from craftsmanship in soldering, our understanding of the science of soldering has grown and has been refined over the centuries.

Today, soft soldering in the electronics industry has developed into a full-fledged production technology, encompassing the fields of mechanics, chemistry, physics and metallurgy. Ernst Sachs, the founder of Ersa (the company name consists of the initial letters of his first and last name) contributed to this development. In 1921 he developed the first electrically operated soldering iron for the industry that was manufactured in series. Since that time, Ersa has committed itself to the further development and perfection of soldering technology with great passion and extending its full power of innovation.

Today, Ersa stands for the most comprehensive product range in the soft soldering technology worldwide and for more than 90 years of industry experience and innovation, know-how and highest product quality.

The Ersa soldering iron product range covers ultra-fine soldering tips, classical soldering irons powered from the standard power net and special soldering tools up to the 550 W hammer soldering iron. Ersa's electronically temperature-controlled soldering stations represent the industry standard, as does the extensive range of rework and

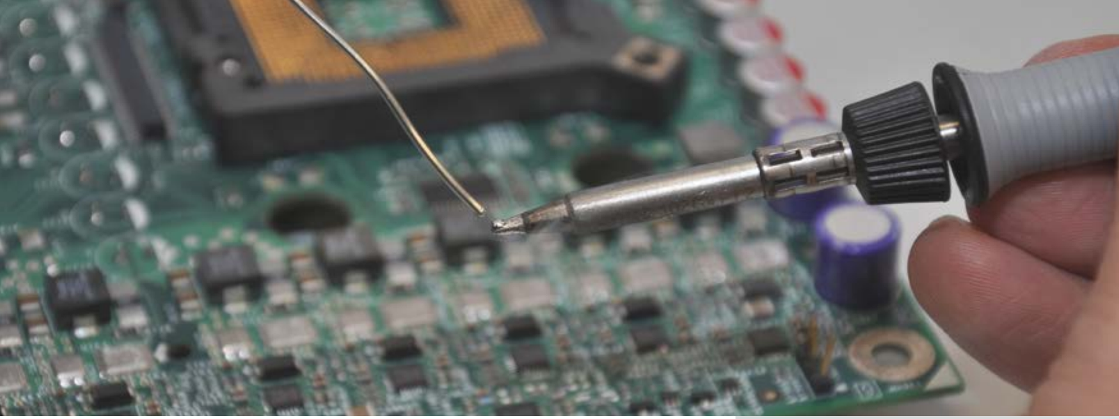

inspection systems, wave-, reflow- and selective soldering systems. The line of Ersa screen printers complements the product range.

Ersa's quality soldering tools are used in the hobby area, such as, for example, in model-making or tiffany soldering, in the craft sector, in laboratories and in industrial electronics manufacturing.

New challenges for the soldering technology were raised by the ban on certain hazardous substances (RoHS) in 2006. Since 1 July 2006 electric and electronic equipment may not contain any lead, mercury, hexavalent chromium, PBB (Polybrominated biphenyls) or PBDE (Polybrominated diphenyl ethers).

In many cases this restriction entailed having to depart from the use of the well-known soft solders based on tin and lead.

By publication of this small primer, Ersa would like to facilitate your entry into the "World of Soldering", and raise your enthusiasm for a modern technology with a long history.

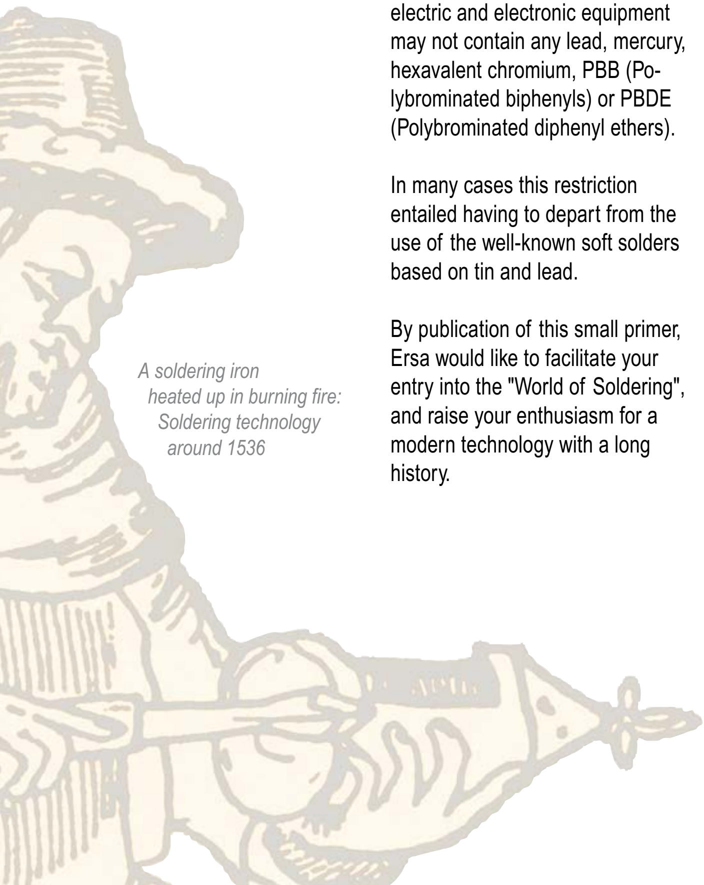

A soldering iron heated up in burning fire: Soldering technology around 1536

Content

Page
The history of soldering 2
The art of soldering 5
What is needed to solder? 6
Health 9
THT soldering 10
SMD soldering 13
BGA soldering / rework 16
Miniature and microsoldering irons 18
Universal soldering irons 19
Standard, hammer and high-speed soldering irons 20
Power and gas soldering irons 21
Soldering stations 22
Multifunctional soldering and desoldering stations 23
Hybrid rework station 26
Solder fume extractions 27
Tiffany soldering 28
Accessories 30
Seminars, references 31

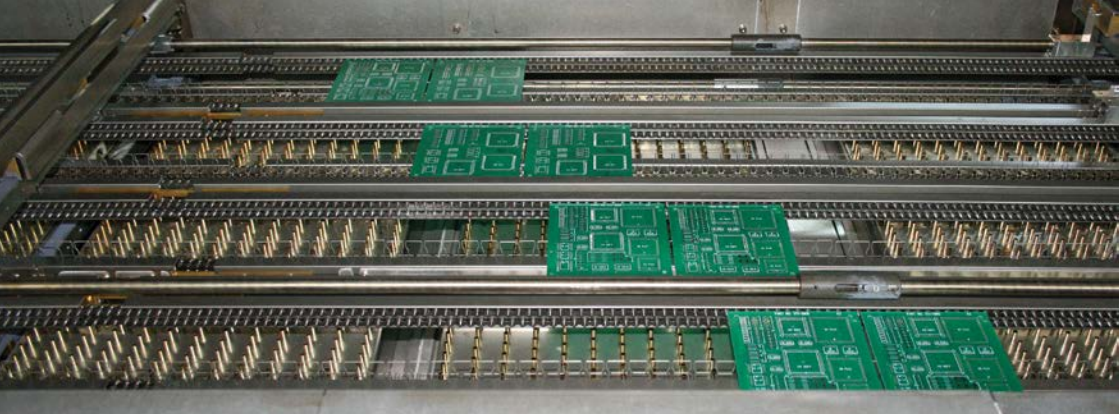

Solder joints hold the world together

Connecting pacemakers and solar parks

To imagine today's world without soft soldering is not possible. It is the means to fabricate safely functioning, electrically conducting connections. Regardless of whether we talk about power technology, drive technology, telecommunications, automation or electronic controls – in all those fields soldered connections have a decisive share that everything functions in a way as has been foreseen and planned by the developers and visionaries of the products. Today, soft soldering is such a common place occurrence, that no one wastes any further thought on it. We take the daily use of our computers, mobile phones and play stations for granted, the modern comforts provided for by electronics found in modern automobiles is expected as a matter of course, and we fly – privately or on business – to the farthest spots in the world.

Consequential damage because of the failure of a solder joint in an iPod is relatively limited. It is a different matter altogether, though, if the electronics in an airplane full of vacationers, in a space shuttle, or in an implanted pacemaker fails. Such failures are immanently life-threatening. But not to worry – the highest quality demands apply for those applications, and rightfully so!

Aside from soldering in consumer electronic products, there are numerous other applications such as, to name but a few, alternative power generation with wind turbines or solar parks, R&D departments and in work performed by craftsmen such as electricians and plumbers. Let us not forget the many part-time and hobby users for whom there are no limits curtailing their phantasies and artistic freedom when handling a soldering iron and solder.

The art of soldering

What's behind it?

In a soldering process two metal parts are joined by means of a molten metallic bonding agent (solder), whereby the melting point of the bonding agent is always lower than that of the metal parts to be joined. If the melting point is below 450 °C, then it is a soft soldering process, if it is above, it is called hard soldering or brazing. Welding, on the other hand, is the process where two metals will be heated up to their melting point, at which time they will, together with a filler material, form a pool of molten material causing coalescence. In soft soldering, the seams between the metals to be joined will be filled with a tin alloy. It is important that the alloy does not simply stick to the foreign metals' surface after cooling but unites with the metal. For this purpose, a small quantity of the foreign metals must dissolve and unite with the tin alloy forming a mix of crystals – the so called diffusion zone. That is the task of the tin, whereas the alloy's other components are responsible for the solder liquefaction and the joint's mechanical stability. A solder joint consists of the following layers:

① Base metal
② Diffusion zone
③ Solidified solder
② Diffusion zone
① Base metal

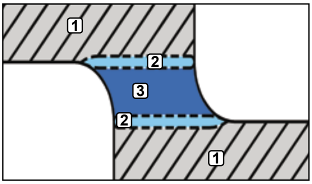

Cross section of a solder joint

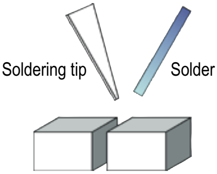

The prepared parts and the solder are heated

The liquid solder flows into the gap and fills it

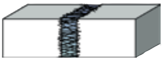

The solidified solder joins the parts together

To achieve the highest mechanical stability, i.e. to assure the durability of the solder joint, the diffusion zone may neither be too thick nor too thin. Its ideal thickness is 0.5 µm. The formation of the diffusion zone depends on the temperature, the solder time and the alloy used. If the diffusion zone is too thick, the solder joint will be brittle and porous, whereas the formation of a zone which is too thin indicates that an insufficient connection or no mechanical connection at all has been formed.

Reaction of the flux during manual soldering on a PCB using flux cored solder wire

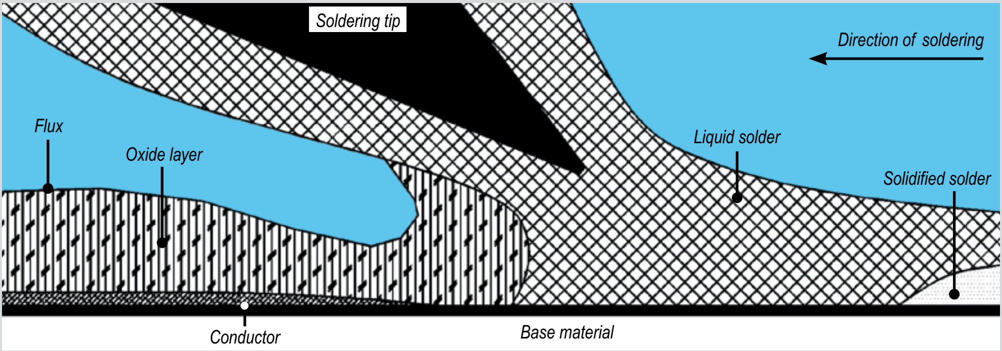

What do you need for soldering?

The basics – the 5 essential factors

1. The soldering iron

to provide the heat required

Heat is required to melt the solder. This is the soldering iron's job.

Temperatures of 200 °C – 450 °C are required depending on solder joint and solder alloy. In the field of electronics, the usual temperature lies between 250 °C and 375 °C.

In order to have the proper temperature for any soldering application, the soldering iron's thermal output and an efficient heat transfer to the solder joint to be made is decisive. One either selects a soldering iron that performs within the temperature range required, or a soldering station with temperature control is used. Temperature-controlled soldering stations enable the user to work on different applications without loss of solder joint quality, because of the precise control of the preset temperature at the soldering iron tip.

Both the soldering station's registration of the actual tip temperature should be highly precise, and the heating element should be powerful and recover quickly in order to avoid over-heated or cold solder joints.

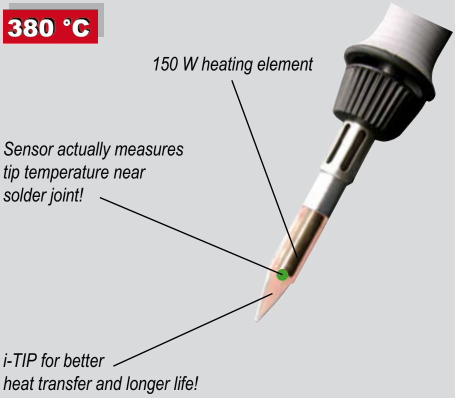
380 °C
150 W heating element
Sensor actually measures tip temperature near solder joint!
i-TIP for better heat transfer and longer life!

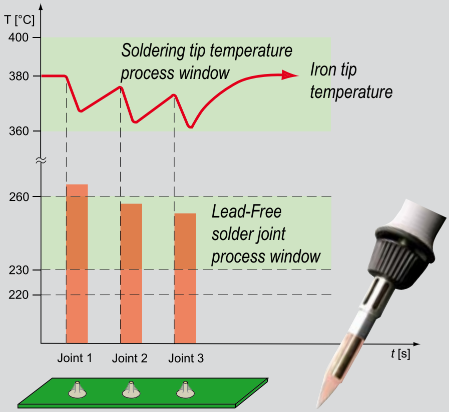

Temperature curve of controlled tip temperature when soldering multiple joints

2. The soldering tip

to transfer the heat from the heating element to the tip

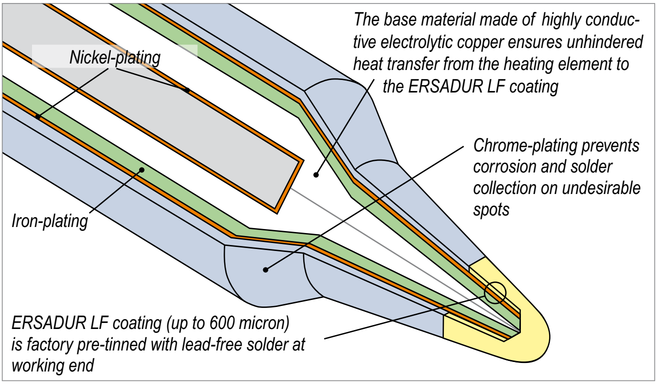
Nickel-plating
The base material made of highly conductive electrolytic copper ensures unhindered heat transfer from the heating element to the ERSADUR LF coating
Chrome-plating prevents corrosion and solder collection on undesirable spots
Iron-plating
ERSADUR LF coating (up to 600 micron) is factory pre-tinned with lead-free solder at working end

Cross-section of an ERSADUR soldering tip, non-scale representation

The soldering tip is the "heart" of the soldering iron and responsible for the heat transfer from the heating element via the solder to the solder joint. Depending on the soldering iron and the application, different types of tips are available. Prerequisites for good solder joints are the correct soldering iron tip shape, perfect heat transfer, a good condition of the tip and a reliable performance over time. In addition, the soldering tip has to convey also the necessary amount of sensitivity back to the operator.

How to take solder joint quality to its peak – ERSADUR long-life soldering tips

Traditional soldering tips are made of copper which conducts heat well and is inexpensive, but the tip oxidizes heavily when heated and releases copper particles into the solder until it has been "corroded" entirely. To maintain the tip in operational shape, it requires intensive care. Today, only coated soldering tips are used in electronics production – the largest field of application for soft soldering.

ERSADUR long-life tips have been conceived for continuous operation and for high-quality results. They are galvanically plated with an iron coating and protected against corrosion and oxidation by an additional chrome layer in a very special manufacturing process, developed and used exclusively by Ersa. And their perfect thermal conductivity protects the heating element from overheating and premature wear. Ersa offers a comprehensive range of soldering tips for the diverse requirements.

Proper tip care increases tip lifetime considerably:

- Never clean the long-life tip before putting the soldering iron into its holder, since the solder remaining on the tip prevents oxidation of the solder track.
- Always keep the long-life soldering tip covered with solder, as otherwise it becomes passive and will no longer wet properly.

Passive tips can be reactivated by the application of the lead- and halogen-free Ersa TIP REACTIVATOR. All that is needed is to wipe the hot tip on the surface of the regeneration compound. Furthermore, the hot tips should regularly be cleaned with a moist viscose sponge before soldering. Alternatively the tips can be dry cleaned using the Ersa "dry sponge", a sponge made of special metal wool. Dry cleaning has proven to be advantageous in lead-free soldering. The soldering tips are not cooled abruptly, and contaminated tips resulting from dirty sponges are avoided. Due to the slightly abrasive properties of the special wire mesh, passive layers that accumulated on the tip can easily be removed. Tip life is thus increased considerably in lead-free hand soldering.

Ersa TIP REACTIVATOR. A comprehensive range of accessories is available at ***www.ersa.com***

What do you need for soldering?

The basics – the 5 essential factors

3. The solder

for the connection

Prerequisites for good soldering results are:

- Suitable soldering iron
- Clean surfaces of where to solder
- Sound soldering tip
- Suitable fluxes
- Suitable solder alloy
- Correct soldering time

Metallic bonding agents (solders), mostly in the shape of a wire or a bar, are available as diverse alloys.

Soft solders consisted mostly of a mix of tin (Sn) and lead (Pb). Since the implementation of the RoHS directives on 01 July 2006 the use of solders containing lead is prohibited. Lead-free solders are usually alloys containing silver (Ag) and copper (Cu).

The alloy's composition determines melting temperature and physical properties of the joint. Criteria for the choice of an alloy are: production process, specification of the product, field of application, cost of the alloy.

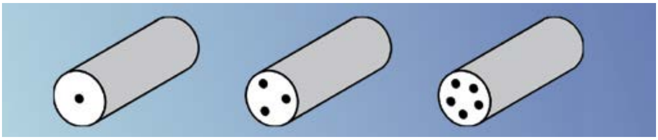

Solder wire with one or with multiple flux cores

4. The flux

for the ability to wet

Fluxes are used to attain the best possible bonding between solder and metal. They provide for metallically clean surfaces of the parts to be soldered, remove the oxides as well as other flow-inhibiting contaminations and prevent the formation of new oxides during the soldering process. A difference is made between acidic (as used in plumbing) and acid-free products (as used in electrical and electronic applications). It is most common to use solder wire with one or more flux cores in electronics production, whereas bar solder is the form of choice in plumbing as well as in the radiator and auto body work.

Examples of some common alloys

Alloy	Flux type	Melting point / range
L-Sn60Pb40	EN 29454/1.1.2 (F-SW 26/DIN 8511)	183 °C – 190 °C
L-Sn60Pb38Cu2	EN 29454/1.1.2 (F-SW 26/DIN 8511)	183 °C – 190 °C
L-Sn63Pb37	EN 29454/1.1.3 (F-SW 32/DIN 8511), free of halogen	183 °C eutectic
L-Sn62Pb36Ag2	EN 29454/1.1.3 (F-SW 32/DIN 8511), free of halogen	178 °C – 190 °C
Alloy – lead-free / complying RoHS-WEEE		
L-Sn95,5Ag3,8Cu0,7	EN 29454/1.1.2 (F-SW 26/DIN 8511)	217 °C eutectic
L-Sn96,5/Ag3,5	EN 29454/1.2.3 (F-SW 33/DIN 8511), free of halogen	221 °C eutectic
L-Sn99,3Cu0,7	EN 29454/1.2.3 (F-SW 33/DIN 8511), free of halogen	227 °C eutectic

5. A safe and clean working environment

to ensure quality and health

Safety first, especially while soldering. The VDE and GS emblems ensure the electrical safety of soldering equipment. The use of the VDE sign obligates the manufacturer to monitor production accordance with test guidelines and to conduct tests according to the regulations determined by the VDE testing institute.

Noxious gases develop in every soldering process

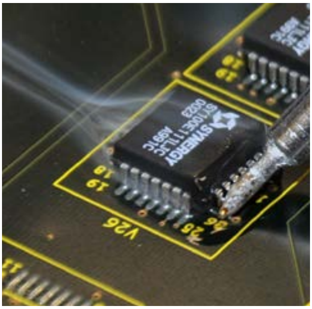

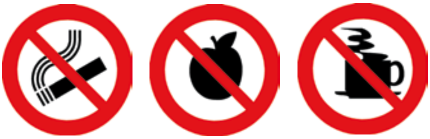

Health protection during soldering

- The breathing zone is very close to the soldering process during manual soldering, and the solder is added by hand. Thus there is the danger of contaminated air and hands or items which may have been touched
- Flux vapors can be damaging to the operator's health and should be kept out of the breathing area. Suitable devices for this purpose are solder fume extractions, which extract the smoke and associated vapors from the workplace and remove particles and gases. Modern fume extractions can be programmed to operate only when the soldering process is taking place, thus saving energy.
- One should never eat, drink or smoke in rooms where soldering is preformed. Contaminations which remain on the hands can enter the human organism through food or cigarette smoke.
- Hands should be washed thoroughly after soldering.

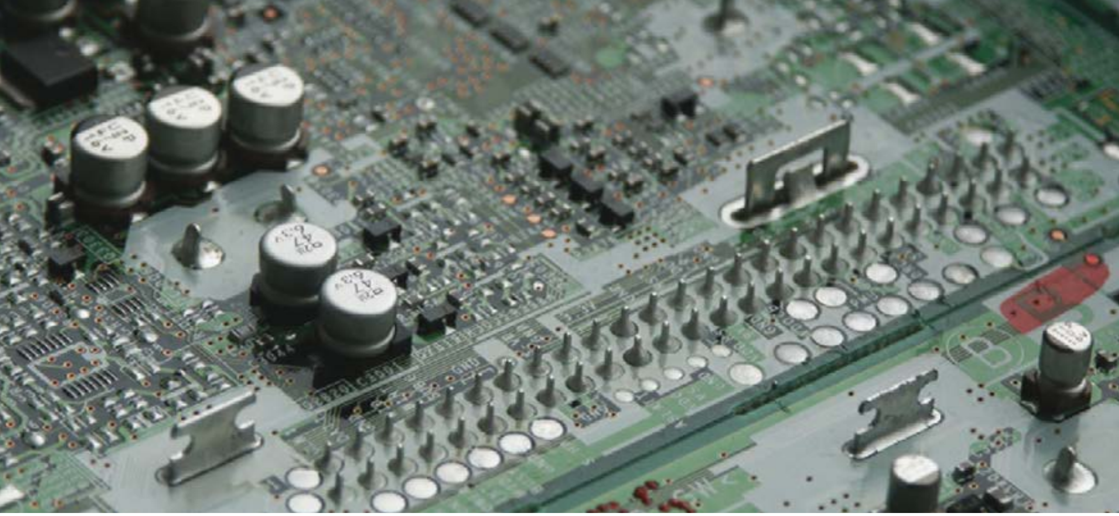

Best solder quality

due to the right preparation and correct soldering parameters

Preparation

The most important prerequisite for a good solder joint is absolute cleanliness. Conductor and component must be free of dirt, oil and oxides. They can be removed with solvents or flux.

Prior to soldering ERSADUR soldering tips should be cleaned while hot with a moist sponge or with a metallic dry sponge. Do not file the tip as you would copper tips, because this would damage the protective coating and render the tip unusable.

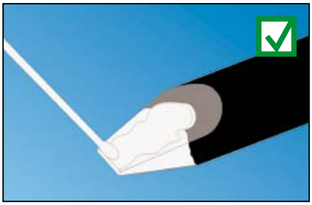

Wetting the soldering iron tip

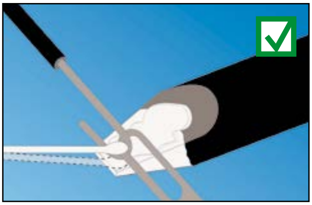

Soldering a joint

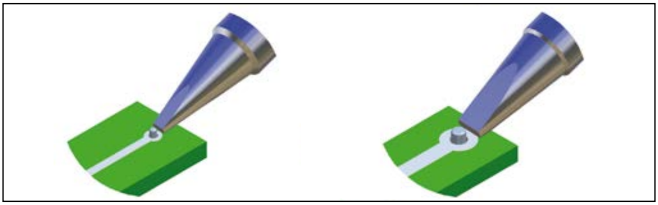

The selection of the right soldering tip is decisive for good soldering results

Soldering process

The soldering process has three phases: wetting, flowing, and bonding, whereas the working temperature is the most important criterion. It is best to work at the lowest temperature at which the three phases can progress smoothly. This requires some experience. A temperature-controlled soldering station will definitely facilitate this work. Place the tip on the joint to be soldered after cleaning and heat up the joint. Then feed the flux-cored solder wire between the soldering tip and the joint and melt as much solder as is required to wet the complete joint. Then remove the solder wire first and right after that the soldering tip to prevent overheating the solder. Allow the solder to solidify, avoiding any vibrations or jarring during this time.

Soldering time

The soldering process should be completed within 2 to maximum 5 seconds with a correctly dimensioned soldering tip. When soldering electronic components with lead-free solders, experience shows that more time is required. But even here requiring more than 5 seconds is not permissible, and it indicates that either the temperature setting is too low or the soldering iron lacks the necessary power.

30
2
4
6
8
10
12
14
16
18
20
22
24
26
28

Solder joint quality

When the leads of the components mounted on the board are crimped, a good solder joint has been formed when the contour of the soldered lead is still visible. This will not be the case, if an excessive amount of solder has been used to form the joint.

A further quality attribute is the wetting angle. This consideration is based on the fact that good wetting of the pad, discernible through a small wetting angle, has given rise to the formation of a diffusion zone (intermetallic zone). Wetting angles of up to 25º identify a good joint, wetting angles of up to 50º are still tolerable in manual soldering.

Another quality indicator is what the solder surface actually looks like. It should be smooth and shiny, without any porous areas visible. Grainy surfaces indicate either overheating of the solder or an excessive soldering time. Using lead-free solders, especially silver loaded alloys, matt surfaces may form.

The only absolute quality indicator for a sound and strong solder joint is the formation of the diffusion zone. In the diffusion zone, intermetallic compounds of copper and tin are formed, whose presence is the final proof of quality (mixed crystals, see page 5). Unfortunately, the only way to make this zone visible is through a destructive test (sectioning of the joint). If the diffusion zone is too thick, the solder joints have no tensile strength and become brittle. The higher the soldering temperature and the longer the soldering duration, the thicker the diffusion zone will be. Therefore the joint should be made at the lowest suitable temperature and within shortest soldering times.

As soon as the last solder joint has been made, the soldering iron is placed securely into the holder. At this time, the tip should not be cleaned, because the remaining solder on the tip prevents oxidization.

Correct!

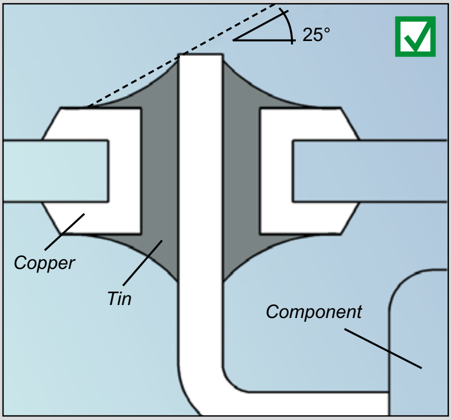

Wrong!

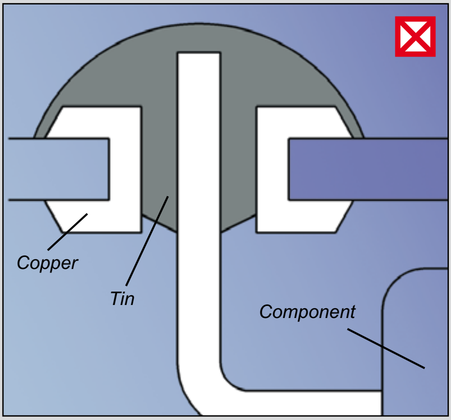

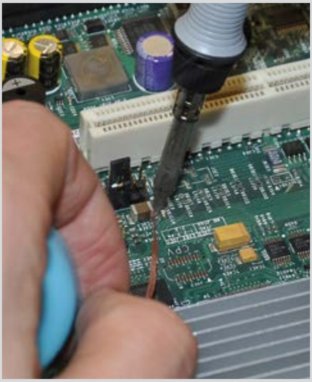

Desoldering with a solder wick

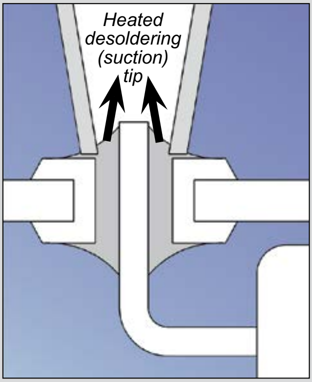

Desoldering with a temperature-controlled soldering iron

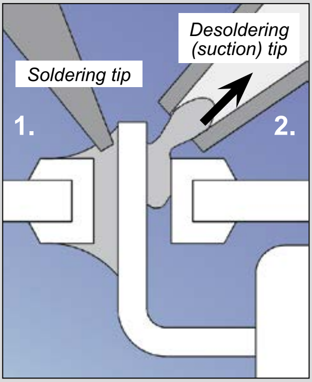

Desoldering with a desoldering pump

To achieve good results in desoldering, it is essential to select the right equipment. One can choose between desoldering wicks (desoldering using the principle of capillary action), mechanical desoldering pumps or electronically temperature-controlled desoldering systems. These are divided in desoldering systems with conductive heat or those with hot air.

Correct desoldering made easy

The right tool for each application

Desoldering

Reheating is not recommended for repairing a faulty solder joint. It is better to remove the solder and to resolder the joint. When using a desoldering pump, the solder joint has to be heated with the soldering iron until the solder has molten. Then the tip is removed and the desoldering pump is placed on the joint to extract the solder.

Using a heated desoldering tool, the hollow desoldering tip is placed on the joint to be repaired, making certain that there is good thermal contact. Once the solder has molten, it is extracted.

Desoldering is also dependent on proper tip selection. For example, the desoldering tip's inner diameter should be the same size as the diameter of the through-hole or even slightly larger (by max. 0.3 mm, see above drawing). The best desoldering results with least damage to PCB or the components can be achieved with temperature-controlled desoldering irons (also see pages 24/25). Please note that basically it has to be distinguished between the desoldering of through-hole components and the desoldering of SMD components.

Temperature-controlled desoldering iron

Sucking up the old, molten solder remnants with a mechanical solder sucker (desoldering pump)

Ever smaller, ever finer

SMD technology, a true challenge

SMD Soldering

SMD technology (Surface Mount Device Technology) is currently the standard process in electronics manufacturing. Ever smaller and highly integrated surface mount components place growing demands on SMD soldering equipment.

Single solder joints, for example on resistors, are made with ERSADUR long-life soldering tips starting with a diameter of 0.2 mm. Fine-pitch connections with a high pin-out, i.e. IC's, are most efficiently soldered with Ersa PowerWell soldering tips.

Ersa PowerWell technology for soldering fine pitch components within seconds

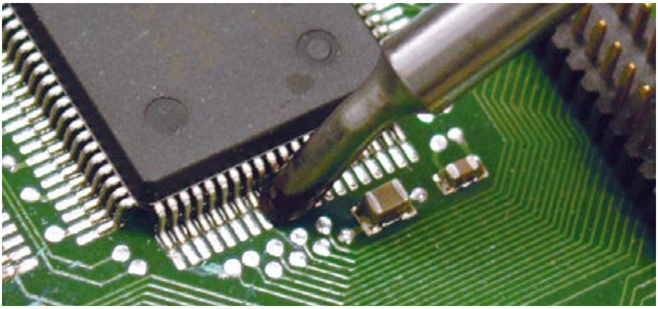

Some examples of SMD components →

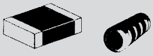

C.CAPAC./MELF

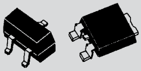

SOT&DPAK

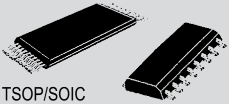

TSOP/SOIC

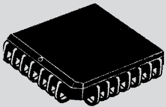

PLCC/LCCC

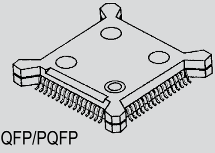

BQFP/PQFP

Soldering fine-pitch components

Process guide description

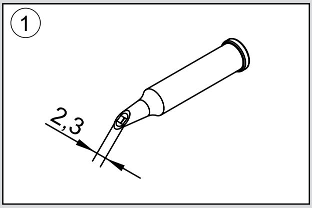

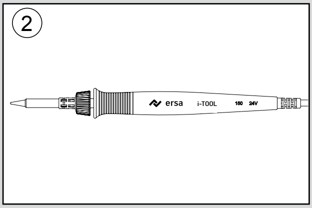

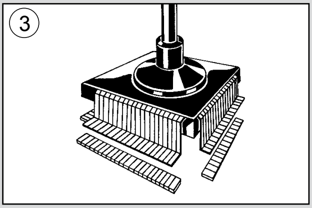

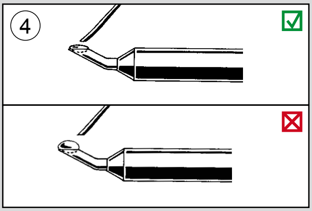

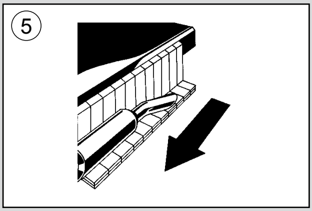

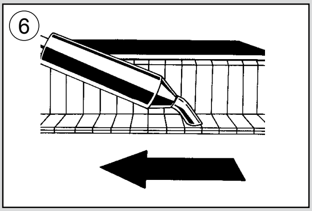

At first glance, soldering fine-pitch components by hand seems to be a tough job. Yet it is easy, with the right equipment at hand:

Insert an Ersa PowerWell i-TIP (1) into the i-TOOL soldering iron (2) and set a tip temperature of 285 °C to 360 °C (depending on the alloy used – tin/lead or lead-free).

Then position the component (3) and fix two corner pins.

Add flux cream (also see page 30) to the pins on all 4 sides. Clean the front and concave portion of the PowerWell tip with a damp sponge or the Ersa dry sponge.

Fill the concave portion with solder to slightly above the rim by melting the solder wire, until a small dome occurs (4). Take care not to add too much solder.

Place the i-TOOL lightly on the flat section of the pins (5), and pull the tip across the pins towards you (6) without exercising pressure.

Repeat steps (4) to (6) to solder the remaining sides.

Good solder joint, sufficient solder and good fillet

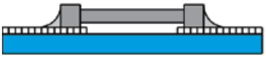

Wrong: insufficient solder

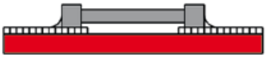

Wrong: too much solder

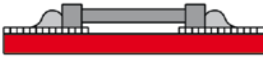

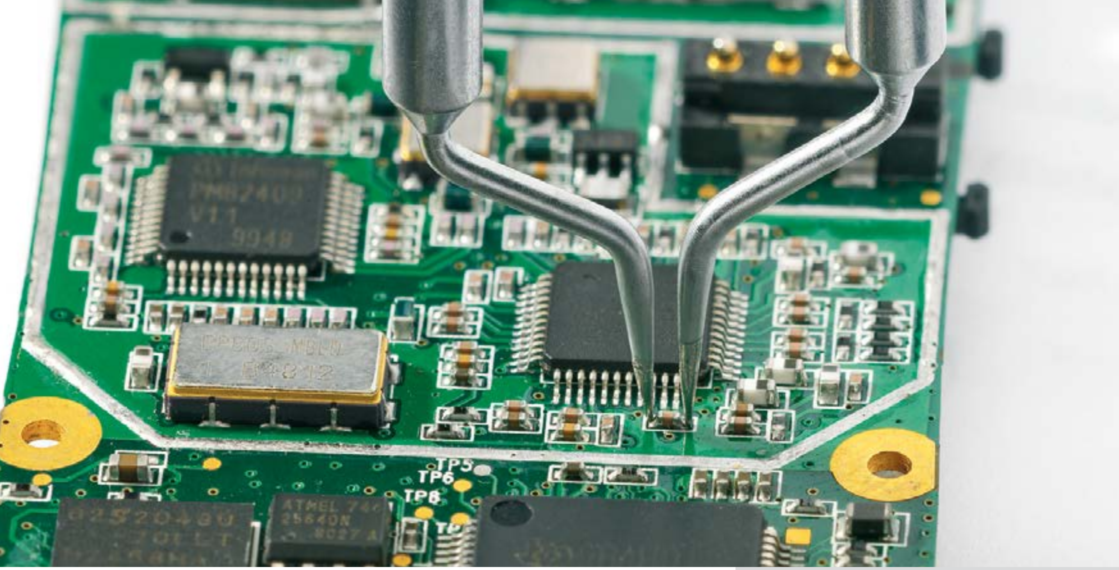

SMD desoldering in 3 seconds – with the Ersa desoldering tweezers

Some examples of different desoldering tip shapes: →

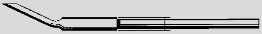

to desolder and remove MELF's

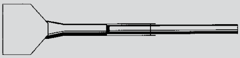

to desolder and remove SOIC's

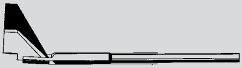

to desolder and remove QFP and PLCC components

SMD Desoldering

To desolder or rework a damaged SMD component, the suitable tools are required to remove the component from the board. When using desoldering tweezers, it is extremely important to select the proper pair of desoldering tips. After having desoldered the component, the pad has to be cleared of the residual solder (e. g. with a suitable soldering tip and a no-clean desoldering wick). Afterwards the new component can be positioned and soldered. An optional IR heating plate is a very useful addition – particularly in lead-free hand soldering applications.

More comprehensive instructions on SMD desoldering is available in the process description "SMD desoldering" on our website at www.ersa.com. For the soldering and desoldering of BGAs or other high pin-out SMDs, particularly those with hidden joints, we recommend the semi-automatic Ersa rework stations (see page 31).

Solder melting and removal with a desoldering wick

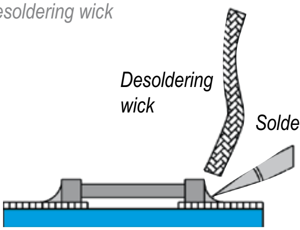

Ersa desoldering tweezers

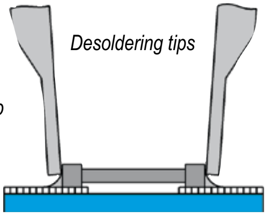

The compact CHIP TOOL VARIO desoldering tweezers are perfect for precise desoldering of very small SMDs.

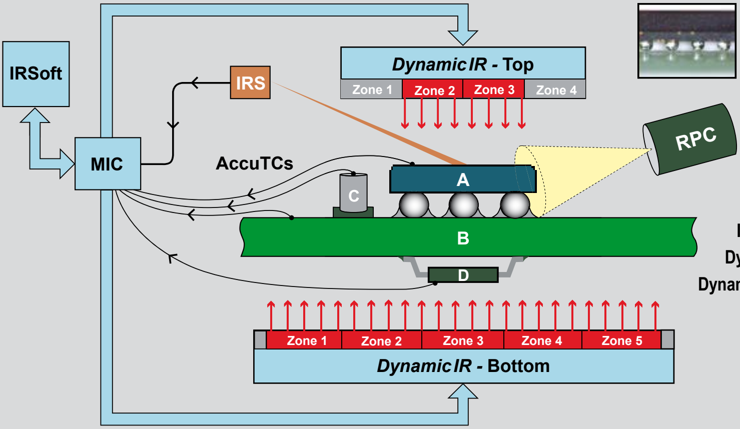

A:	Component for rework
B:	PCB
C:	Adjacent component (top side)
D:	Adjacent component (bottom side)
AccuTCs:	Thermocouples (4 each)
IntelligentIRS:	Non-contact infrared sensor
DynamicIR Top:	Top IR radiator (4 zones)
DynamicIR Bottom:	Bottom IR radiator (5 zones)
RPC:	Reflow Process Camera
IRSoft:	Control & documentation software
MIC:	Microprocessor controls the DynamicIR heating system

Principle of a rework station

Rework or repair soldering

The rework process demonstrated on the example of a BGA (Ball Grid Array)

Rework designates the repair or touch-up of electronic components such as Ball Grid Arrays (BGAs) or SMTs. This chapter describes the process steps for BGA rework:

1. Desoldering the BGA

The rework station heats the printed circuit board from the bottom, whereas the BGA body itself is heated from the top.

The real-time temperature of the component controls the pre-set temperature curve so that all solder joints melt at the same time. Then the vacuum suction cup is placed on the BGA and once all joints have molten the BGA is lifted off.

2. Removal of residual solder, cleaning

Solder remaining on the pads is removed with a soldering iron. To do so the residual solder is coated with flux. Then a flat soldering tip (e. g. 0102ADLF40 or 0102ZDLF150) is moved over the pad without applying any force. The solder adheres to the larger surface area of the tip, thereby levelling off the remaining solder of the connection pads. Flux residues are finally cleaned off (e. g. using Ersa FLUX-REMOVER).

3. Reballing – Reusing the BGA

Desoldered BGA's can be refit with new solder balls and re-attached to the board. This process is referred to as reballing. Residual solder is removed from the BGA by means of a soldering station. Lying on its back, the component is coated with flux first. Then the new solder balls are attached, for example by means of a stencil. These solder balls are heated up to the melting point with the rework station to firmly connect with the BGA body. Now the BGA is ready to be re-attached.

4. Application of flux or solder paste

The component and the connection pads are now fluxed, or, as the case may be, solder paste is added through stencil printing. The type of technique applied depends on the application, components and the skill level of the operators. For the commonly used PBGA's the application of flux is mostly sufficient.

5. Placing the component

Once the pad has been prepared, the component has to be placed. Since all solder joints of a BGA are hidden under the component body, a component placement unit is required. Successful visual placement by hand requires an operator with extensive experience and excellent skills. If the component is placed on solder paste deposits, great care has to be taken not to squash the solder depot, since doing so may lead to shorts after soldering.

6. Resoldering the BGA

The component is heated to the melting temperature of the solder alloy used via a controlled temperature curve. The heating continues until all solder joints have melted and remained so for some seconds. During this time durable and lasting solder joints are formed. After resoldering, the board is cooled in a similar fashion as had been done after desoldering.

These process steps are generally applicable for all surface mount components. Subject to the type of connections (wired, hidden) they may slightly vary.

Rework – Repair of high-terminal count IC's
to successfully repair SMT assemblies,
the several points must be observed:

- Dimensions and properties of the assembly have an influence on its temperature requirement
- PCB holders and supports keep the assembly flat and prevent warpage
- A gentle and controlled heating process, continuously monitored, prevents damage of components or the board
- Accurate component placement is a prerequisite for a good soldering result
- Operators that are well trained will understand the process and ensure good results

Rework process steps

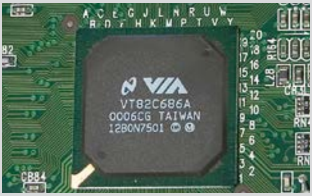
1. Remove BGA

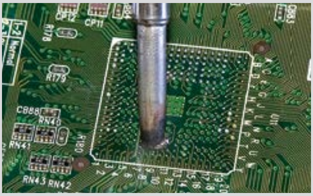
2. Remove residual solder from the pads of the board

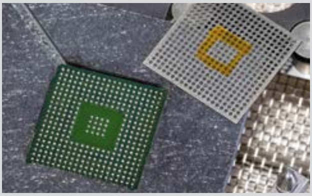
3. Reballing – addion of new solder balls

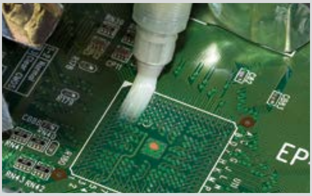
4. Application of flux or solder paste

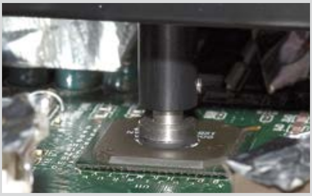
5. Placement of the new or reballed component

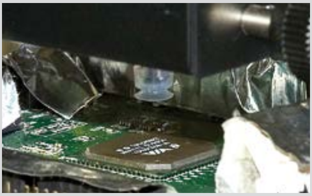
6. Resolder the BGA on the prepared surfaces of the board

Miniature soldering iron

Ersa MINOR S

MINOR S, 5 W

Ersa MINOR S (5 W), a featherweight at 6 g for ultrafine soldering applications

The **MINOR S** miniature soldering iron with a rating of 5 W and a maximum tip temperature of 440 °C is an ideal tool for ultrafine soldering applications on micro IC's and under a microscope. It can either operate with a 6 V transformer or a 6 V battery.

Besides electronics the MINOR S can also be used in watch repair, in the photographic industry and in dental technology.

Micro soldering irons

Ersa MULTITIP Series

MULTITIP series, 15/25 W

Ersa MULTITIP (15/25 W) internally heated soldering iron for small to medium solder joints

The **MULTITIP** soldering irons are especially short, light and easy-to-handle soldering irons with minimal distance between soldering tip and the handle's front part. They are ideally suited for small solder joints. Its internally heated soldering tip provides an enhanced degree of efficiency.

The MULTITIP is available for 15 W and 25 W which makes it suitable for micro soldering joints and medium-sized joints, as for example, on distributor strips or in the hobby sector.

Universal soldering iron

ERSA 30 S, MULTI-PRO, ERSA 15+ and ERSA 25+

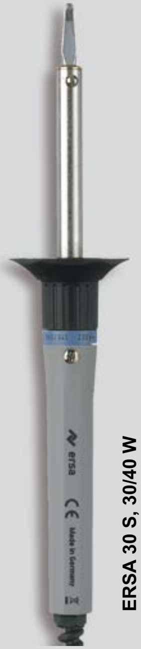

ERSA 30 S, 30/40 W

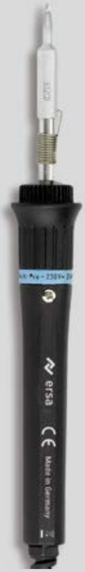

Ersa MULTI-PRO, 20 W

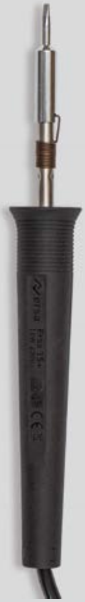

ERSA 15+, 15 W

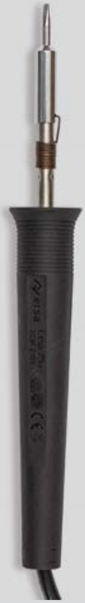

ERSA 25+, 25 W

The proven **ERSA 30 S,** available with a 30 or 40 W rating, is a very robust and durable soldering iron. Its improved ergonomics is further enhanced by the practical stick-on rubber support disk. This universal soldering iron is designed for multi-purpose use in the crafts, service and hobby sector.

The **MULTI-PRO** with its heat-resistant power supply cord is extremely flexible, as it can be operated with a wide range of different tips. Therefore, it is suitable for almost all conventional soldering applications.

The **ERSA 15+** universal soldering iron is the ideal tool for cost-effective soldering. Internally heated soldering tips guarantee an outstanding performance. The ERSA 15+ can be operated with diverse fine tips of the 832 and 842 tip series. Its ergonomic handle assists in the safe processing of fine soldering joints.

The **ERSA 25+** is perfect for soldering applications with increased heat requirements. The internally heated, larger soldering tips of the 832 and 842 tip series ensure good heat transfer at a high thermal performance. The ergonomic handle provides a safe grip for every soldering application.

Ersa universal soldering irons – for almost all applications

Standard and hammer soldering irons

ERSA 50 S, 80 S and 150 S / ERSA 200, 300 and 550

ERSA 50 S, 80 S, 150 S, 50/80/150 W

ERSA 200/300, 200/300 W

ERSA 550, 550 W

The **ERSA 50 S/80 S** or **150 S** series is designed for applications where a large amount of heating capacity is required, as, for example, on copper conductors with cross-sections of 2.5 mm to 6 mm.

The **ERSA 200, 300** and **550** hammer soldering iron series are particularly suited for sheet metal processing and installation work and for soldering commutators and copper bus bars.
Hammer soldering irons have also proven their merit in leveling applications during body work and lead glazing.

High-speed soldering iron

Ersa MULTI-SPRINT

The **Ersa MULTI-SPRINT** is an ergonomically designed and extremely light-weight, pistol-type soldering iron with up to 150 W power which does not require a transformer. Its PTC heating element, together with the internally heated ERSADUR long-life soldering tip, ensures its exceptionally high efficiency. Due to its very short heat-up time, it is ideally suited for making quick single soldering joints. The MULTI-SPRINT is powered only as long as the button is pressed. Depending on the heat requirement of the solder joint to be made, additional energy can be supplied by periodically pressing the button. The large variety of soldering tips of the 832 / 842 series make the MULTI-SPRINT suitable for more than only the use in repair and service applications.

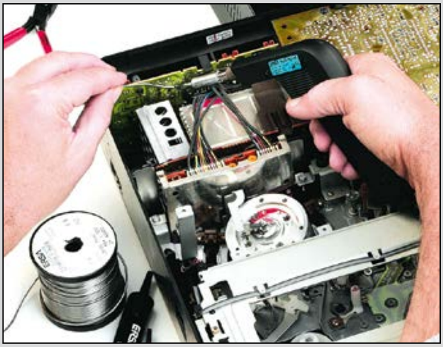

Ersa MULTI-SPRINT, 150/75 W

Power soldering iron

Ersa MULTI-TC and PTC 70

Ersa MULTI-TC, 75 W

Ersa PTC 70, 75 W

The **MULTI-TC** and **PTC 70** are powerful and robust, temperature-controlled soldering irons. Both tools offer an outstanding heat-up rating.

Due to their high thermal performance, and because of the large selection of soldering tips, the MULTI-TC and PTC 70 are suitable for fine soldering joints in electronics as well as for joints with medium heat requirements. The MULTI-TC is furthermore also used anywhere else where standard irons with 150 W power are in use, including, for example, lead glass and Tiffany soldering.

Gas soldering irons

Ersa INDEPENDENT 75 and INDEPENDENT 130

Ersa INDEPENDENT 75, 15 – 75 W

Ersa INDEPENDENT 130, 25 – 130 W

Ersa gas soldering irons are fueled with commercially available lighter gas and are fired up through the piezo ignition. Compared with electrical soldering irons, the **INDEPENDENT 75** has between 15 – 75 W, and the **INDEPENDENT 130** between 25 – 130 W performance. Both irons are available in the BASIC-SET and PROFI-SET versions.

Aside from handling the usual types of electronic components, the selection of tips available enables the INDEPENDENT to also handle SMD soldering, micro-welding, forming and cutting of synthetic materials and the processing of shrink sleeves.

The basic soldering station

Ersa ANALOG 60

The electronically temperature-controlled ANALOG 60 is Ersa's basic soldering station model. It has the tried and proven RESISTRONIC temperature control technology with the PTC heating element serving as the temperature sensor. The high heat-up rating of 190 W guarantees the immediate supply of heat and a heat-up from room temperature to 280 ºC within 60 seconds. The BASIC TOOL 60 soldering iron uses the internally heated ERSADUR long-life soldering tips of the 832/842 series and provides very high performance.

Due to the wide range of 832/842 soldering tips, the Ersa ANALOG 60, which is also available as an antistatic version, covers a wide range of applications with the most varied soldering requirements.

Digital soldering station

Ersa RDS 80 – high performance for low cost

The digital soldering station Ersa **RDS 80** offers the proven and tested Ersa RESISTRONIC temperature control technology with a strong heating power of 80 W. The ceramic PTC heating element (Positive Temperature Coefficient) acts as the temperature sensor in this control system. Due to its very high ramp-up capability of up to 190 W, the station reaches operating temperature very fast.

RT 80: sleek soldering iron with a large selection of soldering tips

Multifunctional soldering stations

Ersa i-CON NANO and i-CON PICO – small, powerful and versatile

Micro-SD card with SD-card and USB adapter

The soldering stations **i-CON NANO** and **i-CON PICO**, two models of the i-CON product family, fulfill all the needs of today's electronics manufacturing while requiring minimal space. They are designed for continuous operation in the electronic manufacturing environment as well as for special applications in laboratories and R&D departments.

Due to the simple and user-friendly operating concept, the factory settings provide for a variable adjustment of operating temperature as well as the setting of stand-by time and calibration value.

Further adjustments such as fixed temperatures, power level, interlock and shutdown functions are available with the free PC software and an optionally available micro smart SD card.

The concept of the Ersa i-CON stations ensures that each application is processed with the optimal parameters. They stand for the highest level of process safety and quality control at low investment and operating cost.

Ersa i-CON PICO

- Small footprint (145 x 80 mm) – saves valuable space
- Antistatic as per MIL-SPEC/ESA (only i-CON NANO)
- Three fixed temperature settings or continuously adjustable temperature settings from 150 °C to 450 °C
- Three selectable power levels
- Ultra-light and ergonomic soldering tool with max. 80 W power
- Wide range of low-cost exchangeable long-life soldering tips
- Automatic stand-by and sleep function for low energy consumption and longer tip life
- Password interlock for maximum process control
- Calibration function for a precise tip temperature
- Complete parameterization through PC software and Micro-SD card

High-end soldering and desoldering stations

The Ersa i-CON product family for highest productivity and process safety

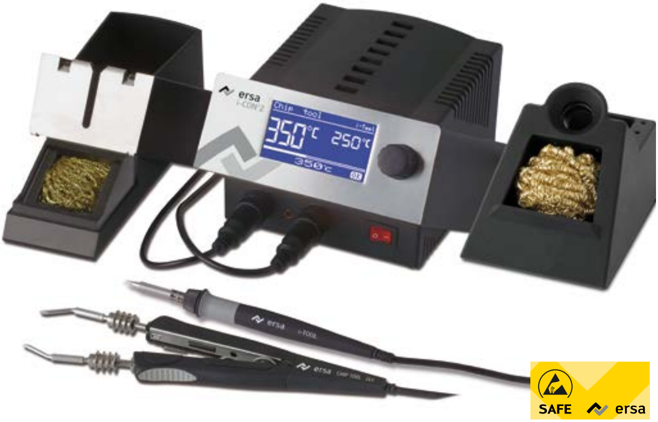

Ersa *i*-CON2 with *i*-TOOL and CHIP TOOL

The stations of the i-CON product family are Ersa's innovative solution for intelligent manual soldering. The need to be able to cope with higher working temperatures and with progressively smaller process windows when working with lead-free solders, poses no problem whatsoever for the i-CON product family.

The i-CON is available as a single station or as double iron station. The single station **i-CON1** is delivered together with the i-TOOL soldering iron. The i-TOOL is extremely small, ultra-light and ergonomic. It is powered by a 150 W micro heating element, which realizes short heat-up times (within 9 seconds up to 350 ºC!) and rapid heat recovery.

The microprocessor which stores the temperature calibration of the iron is located in the PCB which is installed in the handle. This now allows for each individual i-TOOL to be calibrated independent of the soldering station.

In contrast to the concept followed by the cartridge-type soldering tips, only the tip itself is exchanged at the i-TOOL. The cost-intensive heating element remains.

The double iron soldering station **i-CON2** can be operated either with a second i-TOOL or with the SMD desoldering tweezers CHIP TOOL

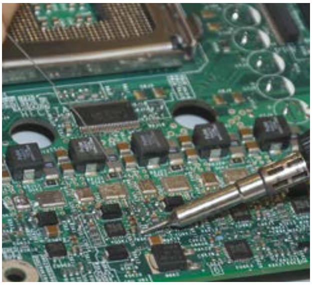

Soldering miniature and densely placed SMD components with the i-TOOL and i-TIPS, starting at a 0.2 mm diameter.

i-CON1 with i-TOOL soldering iron with high-performance micro heating element

- i-TOOL soldering iron with 150 W micro heating element technology
- Low-cost exchangeable long-life tips of the i-TIP series
- User friendly “One-Touch” operation
- Three power levels – no overshoot
- Process window and alarms
- Interface to control peripheral equipment on the workplace, such as heating plate and solder fume extraction systems
- Stand-by control for tools, heating plate and solder fume extraction system
- i-TOOL calibration
- Tools for SMT and conventional soldering applications
- Automatic tool recognition

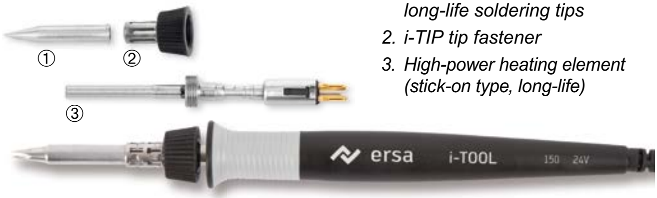

1. *Low-cost i-TIP long-life soldering tips*
2. *i-TIP tip fastener*
3. *High-power heating element (stick-on type, long-life)*

Ersa i-TOOL: The ideal soldering iron – ultra light (only 30 g), ultra short (only 155 mm), and ultra short tip-to-grip (only 45 mm) and extremely powerful.

respectively with the desoldering irons X-TOOL or X-TOOL VARIO.

The CHIP TOOL facilitates safe and quick desoldering, of the smallest chips up to large PLCC's. To remove the residual solder and to desolder wired components – also on multilayer boards – the X-TOOL or X-TOOL VARIO desoldering iron is the right tool.

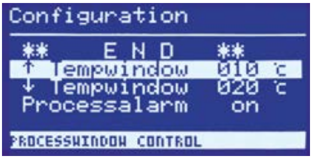

```
Configuration
 **      E N D      **
 ↑ Tempwindow    010 °C
 ↓ Tempwindow    020 °C
 Processalarm     on

PROCESSWINDOW CONTROL
```

User-friendly controls: quick programming & lock, huge multi-functional display with i-Op controls, menu in 7 languages, online help

The C-line of the i-CON stations was developed, so that peripheral equipment could be controlled or to communicate with them. Via a serial interface, the i-CON1 C or the i-CON2 C controls the Ersa IR heating plates or the Ersa solder fume extraction systems.

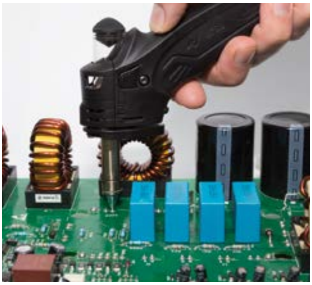

High-mass through-hole desoldering with the X-TOOL VARIO

i-CON VARIO 4 multi-channel station with X-TOOL VARIO desoldering iron, i-TOOL soldering iron, i-TOOL AIR S hot-air iron and CHIP TOOL VARIO desoldering tweezers.

Top Tips

i-TIPS for all applications

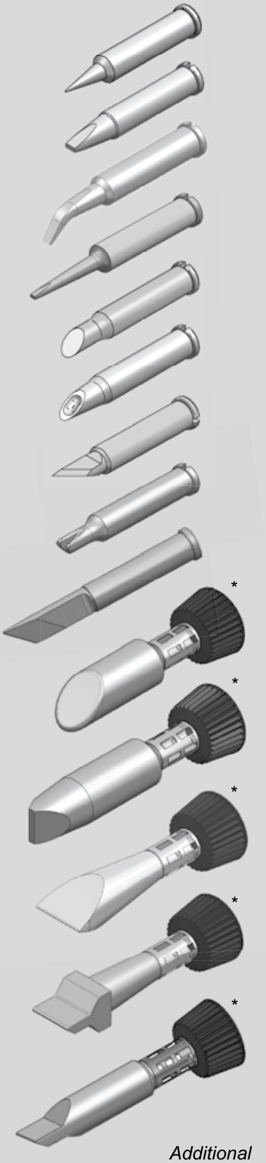

**Figure with i-TIP tip fastener*

Additional soldering tips for the i-TOOL and the i-CON soldering stations is available on our web site:
www.ersa.com

Hybrid Rework Station

Ersa HR 100 A – The innovative combination of hot air and infrared radiation

ersa
HR 100
Ad1

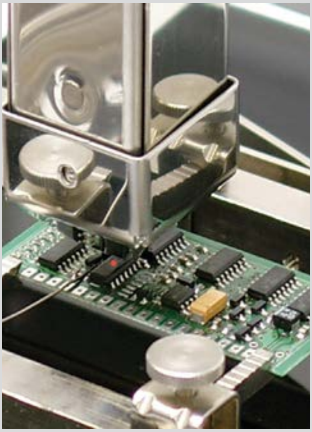

The Ersa **HR 100 A** applies the Ersa Hybrid Rework Technology for a safe desoldering and replacement of small SMD components. Its medium wave length IR radiation, combined with a safe stream of hot air ensures optimal heat transfer to the component.

The HYBRID TOOL offers a gentle and homogeneous warming of all sizes of components, from 0201 chips up to 20 x 20 mm SMD's and larger. Exchangeable hybrid adapters target the thermal energy available (up to 200 W) on to the component, all the while protecting neighboring areas and not blowing away or moving adjacent chips.

Its user-friendly handling permits even operators with little experience to safely and efficiently operate the HR 100. More experienced operators, on the other hand, can not only variably adjust the air flow as well as the heat output, but also record and run profiles. A positioning laser whose laser point makes it possible to keep track of the component worked on through the complete process is integrated in the grip of the ergonomic hybrid tool.

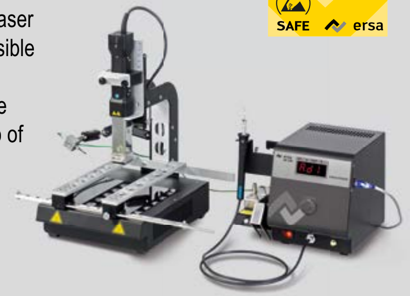

Complete system with HR 100 A, IR heating plate, stand for HYBRID TOOL and PCB holder.

- HYBRID TOOL with 200 W heating element, positioning laser integrated in grip
- 3 exchangeable hybrid adapters (6 x 6 mm, 10 x 10 mm and 20 x 20 mm)
- Silent rework fan (below 40 dB)
- Integrated vacuum pump and vacuum pen; tool holder and K-type thermocouple input socket; USB interface; LED display
- 2-channel temperature recording:
  TC & IRS; AccuTC and Flexpoint thermocouple holder
- Tool holder with z-axis height adjustment
- x-y PCB holder (290 mm x 250 mm)
- 800 W IR heating plate with glass cover:
  125 mm x 125 mm IR high-performance heating element
- Rework profile and documentation software Ersa IRsoft

CLEAN-AIR solder fume extractions

Ersa EASY ARM 1 and EASY ARM 2 – protecting environment, health and resources

ersa
EASY ARM 1

Ersa EASY ARM 1

ersa
EASY ARM 2

Ersa EASY ARM 2

The Ersa **EASY ARM 1** is a compact yet powerful filter unit to efficiently clean the process air at the workplace. Both i-CON1 C, i-CON2 C and i-CON VARIO can be connected to the EASY ARM 1 with an interface cable. The solder fume extraction has three filtration levels to remove noxious gases.

With the solder fume extraction **EASY ARM 2**, Ersa offers the user a further compact and highly efficient fume extraction unit for either one or two workplaces. One or two units i-CON1 C / i-CON2 C / i-CON VARIO can be connected to the EASY ARM 2 via interface cable.

Both EASY ARM systems are provided with a stand-by mode and operate only when the soldering stations they are connected to are being used. The EASY ARM 1 and 2 are easy to install and can be placed very flexibly. Due to their very low noise level, they can be operated in virtually all environments, be it repair shops, development facilities or laboratories.

To accommodate different working conditions and applications, a variety of extraction arms and nozzle shapes which can easily and quickly be exchanged are available to meet different working conditions.

Noxious gases develop in every soldering process.

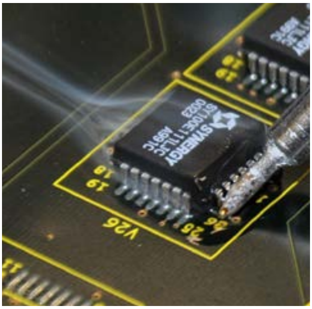

Q&A 1: So, how can you get rid of them? By using a solder fume extraction unit to clear the air from particles and gases.

Q&A 2: How does that work? Active carbon in the filters absorbs harmful gas molecules; result: a clean working environment.

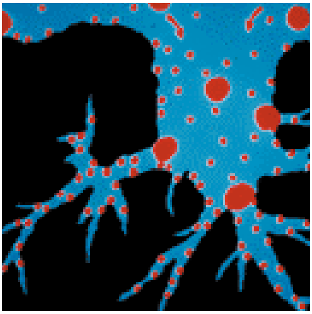

Tiffany soldering

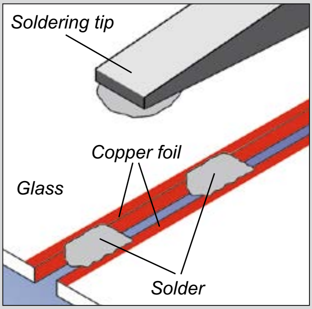
Soldering tip
Copper foil
Glass
Solder

Step 1: spot soldering

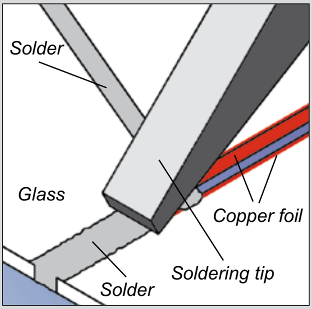
Solder
Glass
Copper foil
Soldering tip
Solder

Step 2: rough soldering

Tiffany or sheet metal soldering

Soldering beyond the field of electronics

Tiffany soldering (Lead glass soldering)

Soldering a Tiffany object generally involves three individual steps:

- Spot soldering
- Rough soldering
- Finishing

Prior to the actual soldering process, copper foil is glued on the glass edges. The next step is spot soldering, i. e. the glass parts are fastened or connected by taking a drop of solder with the tip of the soldering iron and carefully applying it to the solder joint. Each spot soldering operation should only take about a second. During rough soldering the gap between the glass is filled with solder after flux has been applied. For this purpose the tip and solder are moved together along the joint. Always drag the soldering iron, never push it. Only if this procedure is accurately followed, and if a sufficient amount of solder is applied, the desired, half-round and convex seams are achieved. The lack of visual quality of the seam at this stage is optimized during the finishing soldering step, in which the soldering tip is dragged slowly and from the beginning of the seam until the end at an even speed. The seam worked on should always lie flat on the bench.

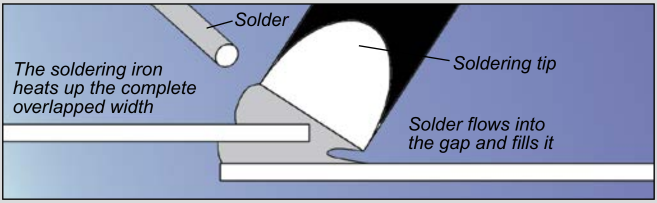
Solder
The soldering iron heats up the complete overlapped width
Soldering tip
Solder flows into the gap and fills it

Guiding the soldering iron when soldering a broad seam

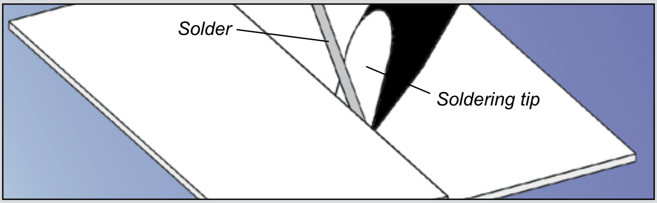
Solder
Soldering tip

Guiding the soldering iron when soldering a narrow seam

Plumbing and tinsmith work

For the joining of sheet metal and metal pipes, the joints to be soldered have to be bare. This calls for a good prior cleaning.

After this, the flux – either solder grease or solder fluid, a zinc chloride solution – is applied. Then the area to be soldered is heated with the tip of the soldering iron.

Once the soldering area is sufficiently hot, solder is added and the solder gap is filled.

The aggressive flux residues are removed after cooling to avoid the risk of future corrosion.

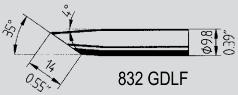

832 GDLF

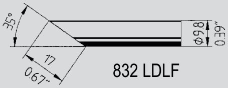

832 LDLF

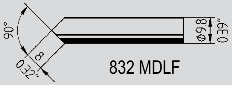

832 MDLF

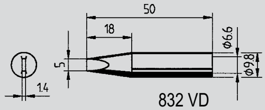

832 VD

ERSADUR Tiffany soldering tips

The types VD, GDLF, LDLF and MDLF of the comprehensive 832 tip series are particularly suited for Tiffany work. On account of their shapes and their high mass (excellent because of the ensuing heat retention capacity), the seams between the glass parts can quickly and easily be filled with solder. And the ERSADUR finish warrants a long tip lifetime.

ERSADUR Tiffany soldering tips

Ersa DIGITAL 2000A, Ersa MULTI-TC and ERSADUR soldering tips

Tiffany lamp

The **MULTI-TC** soldering iron is very light, robust and powerful. The Ersa SENSOTRONIC Control with the PT-1000 temperature sensor in the tip reacts immediately when heat is withdrawn.

The MULTI-TC is provided with a 2.2 mm wide, chisel-shaped soldering tip. Together with the Tiffany soldering tips, the slim MULTI-TC is superbly suited for the use in Tiffany work.

Due to its precise temperature control, the **DIGITAL 2000 A** soldering station completely eliminates the possibility to overheat any glass components or the copper foil, and the station has tremendous power reserves, making it comparable and putting it on an equal level with unregulated 150 W soldering irons.

The Ersa SENSOTRONIC control with its internally heated soldering tip, where the temperature sensor is mounted directly below the tip, provides for precise temperature control and ensures uniform temperature levels at the soldering joint.

The long-life ceramic PTC heating element provides up to 290 W heat-up rating which means that the powerful soldering iron is ready for use in only 60 seconds.

The DIGITAL 2000 A is provided with a 2.2 mm wide, chisel-shaped soldering tip. Particularly good results are achieved with the optionally available tip versions 832 VD, 832 GDLF, 832 CDLF and 832 MDLF, which are specifically offered for Tiffany applications.

Ersa MULTI-TC, 75 Watt

Ersa DIGITAL 2000 A with POWER TOOL

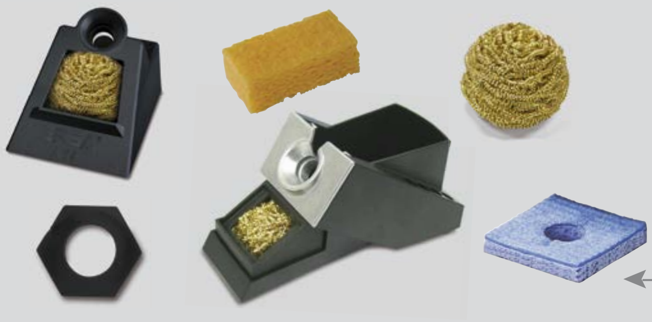

Ersa tool holders or stick-on rubber supports to safely and ergonomically put down the soldering iron during work stoppages or during heat-up.

Ersa viscose cleaning sponges for moist or special metal wool for dry cleaning of the hot soldering tip prior to the soldering process.

Auxiliary products and practical accessories

Everything for your soldering needs from one source

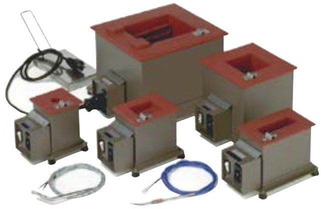

Ersa solder baths

Ersa solder wire and solder wire dispenser

Ersa DTM 100 temperature measuring device

Ersa tip exchanger

Ersa desoldering wicks

Ersa SOLDAPULLT desoldering pump

Ersa VAC X desoldering pump

Ersa SMD flux and flux cream

FLUX-REMOVER

Ersa FLUX-REMOVER

Ersa TIP-REACTIVATOR

Further peripherals and accessories can be obtained from your local Ersa distributor and at **www.ersa.com**. Or just ask for our catalog!

IR rework stations

Ersa IR/PL 550, IR/PL 650 and HR 600 – from mobile phone boards up to XXL boards

Ersa HR 600

More than 5,000 customers worldwide use Ersa's patented IR-Rework technology.

Ersa rework stations deploy the DynamicIR Heating Technology, programmable top and bottom heating zones as well as the accurate, user-friendly and motorized "Auto Pick & Place" feature. System control, process documentation and visualization is handled by the Ersa rework software IRSoft/HRSoft.

Optical solder joint inspection

ERSASCOPE inspection systems – the patented original

ERSASCOPE 2

Since its introduction over 10 years ago more than 3,000 customers worldwide profit from the possibility to non-destructively inspect hidden solder joints.

Regardless of whether the joints are under a flip-chip, or whether areas are to be inspected where other microscopes have come to their limits – the ERSASCOPE technology offers a considerable added value to any quality assurance program!

Ersa Know-How Seminars

Expert knowledge for professional users comprehensively imparted

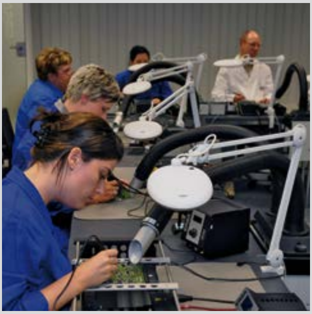

Today more than ever, cost-effective production and highest quality in the electronics manufacturing industry is the basis for profitable competitiveness. A key ingredient for being competitive is a well-qualified staff, which is up-to-date on current technology and process.

Ersa Know-How Seminars are ideal for transmitting this knowledge to those of your staff members with process responsibility in electronic manufacturing. In these seminars, theoretical and practical knowledge is presented in a neutral fashion in workshops and presentations by reputed experts in their field.

For further information go to our website **www.ersa.com** and click "Events".

Ersa Electronics Production Equipment

Katalog
Ersa Lötkolben, Löt- und Entlötstationen, Lötrauchabsaugungen, Hybrid-Rework & Zubehör
kurtz ersa

Technical data and further detailed information on our range of hand soldering equipment is available in our "Soldering Tools" catalog. This is available in printed form and as PDF, as are the "Rework & Inspection" catalog.

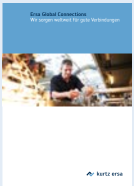
Ersa Global Connections
Wir sorgen weltweit für gute Verbindungen
kurtz ersa

Aside from the extensive selection of hand soldering tools, rework stations and inspection systems for non-destructive inspection, Ersa, as Europe’s largest manufacturer of soldering systems, also offers a complete range of selective, wave and reflow soldering systems, as well as stencil printers for the industrial electronic manufacturing industry.

Please refer to our brochures and our website **www.ersa.com** for further information.

Ersa GmbH
Leonhard-Karl-Straße 24
97877 Wertheim/Germany
Tel. +49 9342 800-0
Fax. +49 9342 800-127
info@ersa.de
www.ersa.com

Ersa France
Chevigny Saint Sauveur, France
info-efr@kurtzersa.com

Ersa North America
Plymouth, USA
info-ena@kurtzersa.com

Ersa Asia Pacific
Hong Kong, China
info-eap@kurtzersa.com

Ersa Shanghai
Shanghai, China
info-eap@kurtzersa.com

kurtz ersa

97596 - 02/2015 • subject to changes • © by Ersa GmbH, Germany

Soldering Guide

Safety Precautions

- **Never touch the element or tip of the soldering iron.**
  They are very hot (about 400°C) and will give you a nasty burn.
- **Take great care to avoid touching the mains flex with the tip of the iron.**
  The iron should have a heatproof flex for extra protection. Ordinary plastic flex melts immediately if touched by a hot iron and there is a risk of burns and electric shock.
- **Always return the soldering iron to its stand when not in use.**
  Never put it down on your workbench, even for a moment!
- **Allow joints a minute or so to cool down before you touch them.**
- **Work in a well-ventilated area.**
  The smoke formed as you melt solder is mostly from the flux and quite irritating. Avoid breathing it by keeping you head to the side of, not above, your work.
- **Wash your hands after using solder.**
  Solder contains lead.

Treatment for minor burns

Most burns from soldering are likely to be minor and treatment is simple:

- **Immediately cool the affected area under gently running cold water.**
  Keep the burn in the cold water for at least 5 minutes (15 minutes is recommended). If ice is readily available this can be helpful too, but do not delay the initial cooling with cold water.
- **Do not apply any creams or ointments.**
  The burn will heal better without them. A dry dressing, such as a clean handkerchief, may be applied if you wish to protect the area from dirt.
- **Seek medical attention if the burn covers an area bigger than your hand.**

Preparing the soldering iron

- **Place the soldering iron in its stand and plug in.**
  The iron will take a few minutes to reach its operating temperature of about 400°C.
- **Dampen the sponge in the stand.**
  The best way to do this is to lift it out the stand and hold it under a cold tap for a moment, then squeeze to remove excess water. It should be damp, not dripping wet.
- **Wait a few minutes for the soldering iron to warm up.**
  You can check if it is ready by trying to melt a little solder on the tip.
- **Wipe the tip of the iron on the damp sponge.**
  This will clean the tip.
- **Melt a little solder on the tip of the iron.**
  This is called 'tinning' and it will help the heat to flow from the iron’s tip to the joint. It only needs to be done when you plug in the iron, and occasionally while soldering if you need to wipe the tip clean on the sponge.
- **You are now ready to start soldering!**
  Please turn the page for further instructions...

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Making soldered joints

- **Hold the soldering iron like a pen, near the base of the handle.**
  Imagine you are going to write your name!
  Remember to never touch the hot element or tip.
- **Touch the soldering iron onto the joint to be made.**
  Make sure it touches both the component lead and the track.
  Hold the tip there for a few seconds and...
- **Feed a little solder onto the joint.**
  It should flow smoothly onto the lead and track to form a volcano shape as shown in the diagram below. Make sure you apply the solder to the joint, not the iron.
- **Remove the solder, then the iron, while keeping the joint still.**
  Allow the joint a few seconds to cool before you move the circuit board.
- **Inspect the joint closely.**
  It should look shiny and have a ‘volcano’ shape. If not, you will need to reheat it and feed in a little more solder. This time ensure that both the lead and track are heated fully before applying solder.

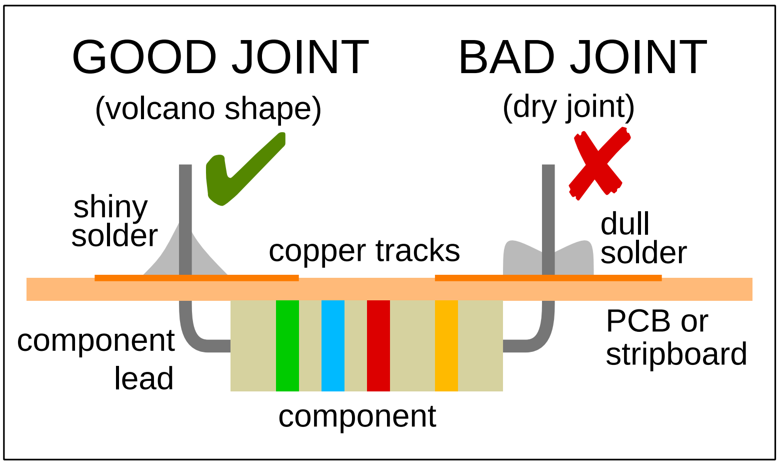

Using a heat sink

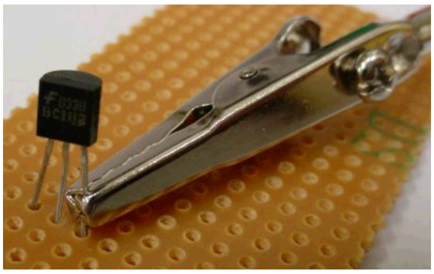

Some components, such as transistors, can be damaged by heat when soldering. It is wise to use a heat sink clipped to the lead between the joint and the component body, as shown in the picture. You can buy a special tool, but a standard crocodile clip works just as well and is cheaper!

Soldering advice for components

Some components require special care when soldering. Many must be placed the correct way round and a few are easily damaged by the heat from soldering. Appropriate warnings are given in the table on the next page, together with other advice which may be useful when soldering.

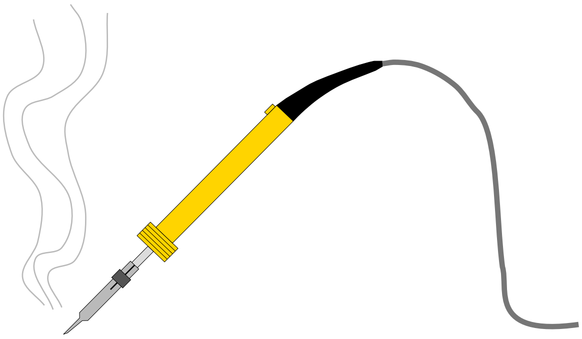

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Components	Pictures	Soldering advice
Resistors		No special precautions are required. Connect either way round.
Diodes	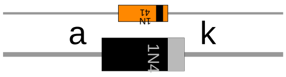	Diodes must be connected the correct way round: a = anode, k = cathode. Use a heat sink with germanium diodes.
IC holders (DIL sockets)	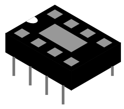	Ensure the notch is at the correct end. Do **not** insert the IC at this stage to prevent it being damaged by heat.
Presets (small variable resistors)	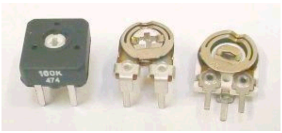	No special precautions are required. On stripboard take care to ensure you insert them the correct way round.
Capacitors, non-polarised (less than 1µF)	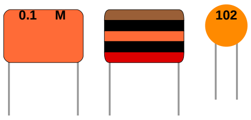	No special precautions are required. Connect either way round. Take care to identify their value.
Capacitors, electrolytic (1µF and greater)	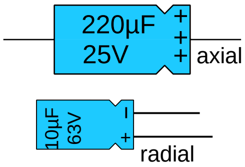	Electrolytic capacitors must be connected the correct way round, they are marked with + or - near one lead.
LEDs (Light Emitting Diodes)	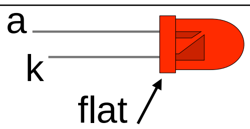	LEDs must be connected the correct way round: a = anode, k = cathode. Use a heat sink with small (3mm) LEDs.
Transistors	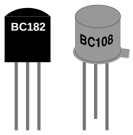	Transistors have three leads and must be connected the correct way round. Use a heat sink clipped to each lead in turn between the joint and the transistor.
Wire links between points on the board		Use tinned copper wire (such as the offcut from a resistor lead) or single-core plastic-coated wire.
Other parts mounted on the board		No special precautions are required for most parts, but make sure they are the correct way round.
Battery clips, buzzers and other parts with wires		Red (+) and black (-) wires must be connected the correct way round.
Wires to parts off the board such as switches		Use plastic-coated stranded wire which is flexible, single-core wire is likely to break at the joint.
Integrated Circuits (ICs or 'chips')	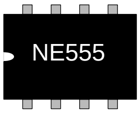	When all soldering is complete, carefully insert ICs the correct way round in their holders. Make sure all the pins are lined up before pushing in firmly.

For further information about electronic components please see: www.kpsec.freeuk.com

THE ELECTRONICS CLUB

© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Desoldering

At some stage you will probably need to desolder a joint to remove or re-position a wire or component. There are two ways to remove the solder:

1. With a desoldering pump (solder sucker)

- Set the pump by pushing the spring-loaded plunger down until it locks.
- Apply both the pump nozzle and the tip of your soldering iron to the joint.
- Wait a second or two for the solder to melt.
- Then press the button on the pump to release the plunger and suck the molten solder into the tool.
- Repeat if necessary to remove as much solder as possible.
- The pump will need emptying occasionally by unscrewing the nozzle.

2. With solder remover wick (copper braid)

- Apply both the end of the wick and the tip of your soldering iron to the joint.
- As the solder melts most of it will flow onto the wick, away from the joint.
- Remove the wick first, then the soldering iron.
- Cut off and discard the end of the wick coated with solder.

After removing most of the solder from the joint(s) you may be able to remove the wire or component lead straight away (allow a few seconds for it to cool). If the joint will not come apart easily apply your soldering iron to melt the remaining traces of solder at the same time as pulling the joint apart, taking care to avoid burning yourself.

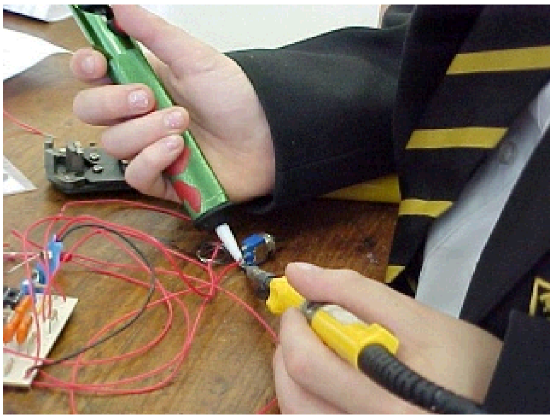

Using a desoldering pump (solder sucker)

What is solder?

Solder is an alloy (mixture) of tin and lead, typically 60% tin and 40% lead. It melts at a temperature of about 200°C. Coating a surface with solder is called ‘tinning’ because of the tin content of solder. Lead is poisonous and you should always wash your hands after using solder.

Solder for electronics use contains tiny cores of flux, like the wires inside a mains flex. The flux is corrosive, like an acid, and it cleans the metal surfaces as the solder melts. This is why you must melt the solder actually on the joint, not on the iron tip. Without flux most joints would fail because metals quickly oxidise and the solder itself will not flow properly onto a dirty, oxidised, metal surface.

The best size of solder for electronic circuit boards is 22swg (swg = standard wire gauge).

For plugs, component holders and other larger joints you may prefer to use 18swg solder.

THE ELECTRONICS CLUB
© John Hewes 2006, The Electronics Club, **www.kpsec.freeuk.com**
This document may be copied for educational purposes.

Safety Rules for Soldering

Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island

1. Wear goggles for eye protection. When trimming off leads or excess solder dross, be careful of the flyaway that could injury yourself as well as other people nearby.
2. Work in a well-ventilated area and use a fume extractor. Do not inhale fumes from the soldering processes.
3. Always return the soldering iron to its stand when not in use. Never put it down on your workbench. The soldering iron tip is very hot (about 400°C). Avoid touching plastic, wire insulator, or any flammable material in the working area with the soldering iron. Turn the soldering station to standby or off if not used for more than few minutes. Turn unit off or unplug it when done.
4. Use a third hand, a circuit board vice, pliers, tweezers, or clamps for holding components to avoid burns. Legs and arms should be covered to avoid burns from splashed hot solder.
5. Do not have food or drink near the working area. The solder is usually a tin/lead alloy and lead is toxic. The flux is a chemical used to help metal parts soldered together. It is acidic and toxic. Clean up spilled flux immediately. Wash hands after soldering. Flux can cause acid burns to the skin or damage clothing. In case of acid burns, flush immediately with water.
6. Wash hands thoroughly after handling flux and solder containing lead. Use lead-free solder whenever possible.
7. Clean up the area when finished. Discard lead and silver solder and dross in a container with a lid. Label the container: "Lead/Silver Solder Waste for Recycling". Used solder sponges and contaminated rags must be disposed of as hazardous waste.

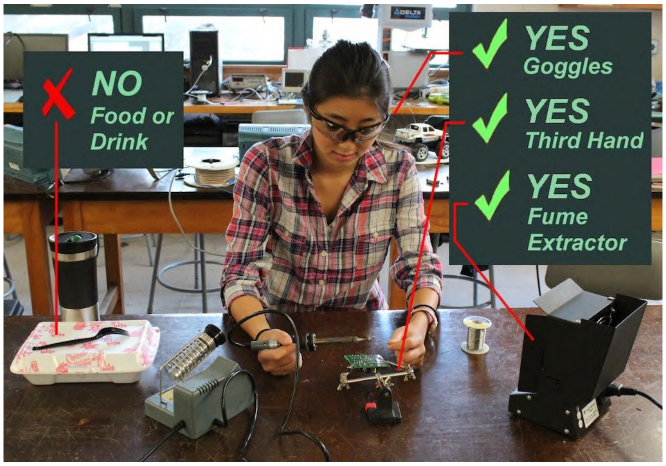

Soldering Techniques

1. Basic tools include an electric soldering station (a temperature-adjustable type preferred), solder, a pair of needle-nose pliers, a wire stripper, a flush cutter, a desoldering bulb, a utility blade, and a third hand or a vice.
2. Workspace should be as uncluttered as possible. Clear the path between the soldering station stand and the project to be soldered.
3. Prepare the soldering iron tip. Scrape off oxides with a utility blade if necessary. Adjust to an appropriate temperature. It won't work if you can't get the solder melt onto the soldering tip. Tinning the soldering tip by coating it with a thin coat of solder. This helps heat transfer between the tip and the components. The cleaning sponge should be soaking wet. Quickly run the soldering tip over the wet sponge to get rid of excess solder.
4. Clean the contact surfaces of the metal parts if necessary. Scrape off oxides with a utility blade. In some cases, it is easier to tin the individual components first before joining them.
5. Position the components using a third hand or a vice. Avoid holding the metal part of a component with your fingers.
6. Timing is important for soldering the components together. The solder needs to be completely melt at the joint to avoid a "cold solder point." However, too much heat could damage a sensitive electronic component or melt the insulator of a wire. The required time depends on the heat transfer, which is affected by the temperature of the soldering tip, the mass of the components, the presence of flux or rosin, and how clean the contact surfaces are.
7. Start over when an attempt fails. Old solder has impurities. Remove old solder completely with a desoldering bulb or pump. Clean up flux/rosin residuals. Try again.

Examples: www.leadsdirect.co.uk/technical/perfectsolderjoint.html

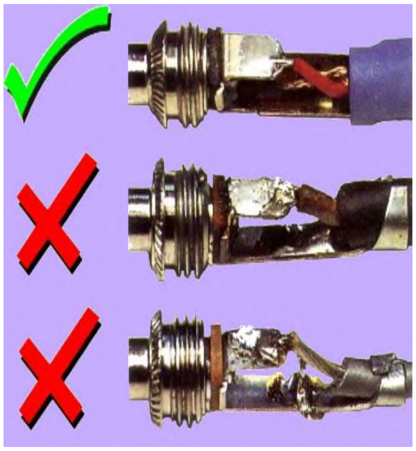

Good – clean, shiny, and just the right amount of solder.

Bad – too much solder which is uneven, has sharp points in places, and was probably overheated.

Worse – uneven texture, and the flux remaining on the joint.

Soldering

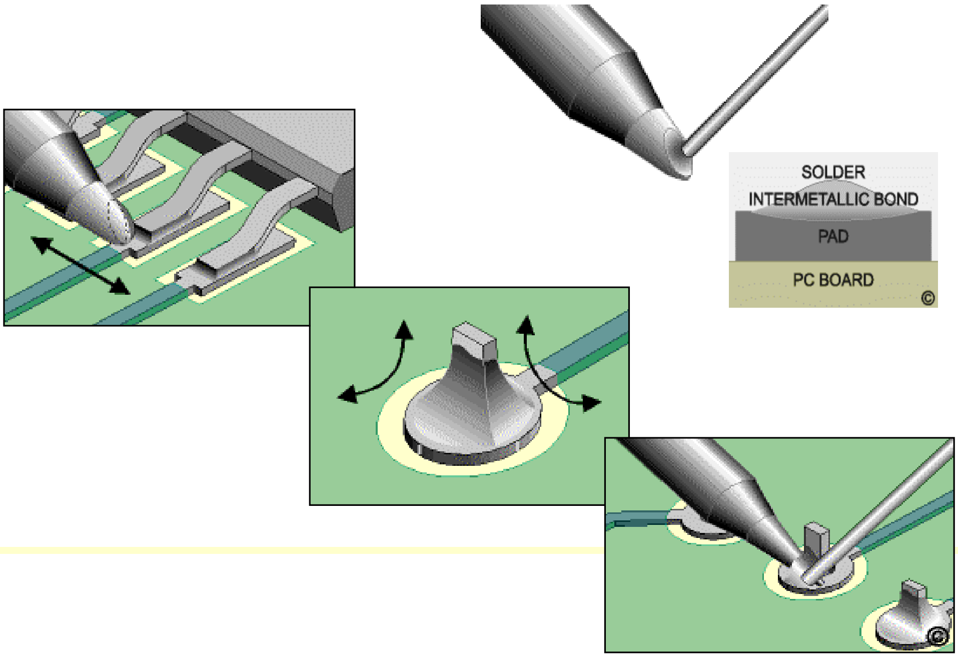
SOLDER
INTERMETALLIC BOND
PAD
PC BOARD
©

TABLE OF CONTENTS

page

1. Introduction 1
2. Basic metals 1
3. Flux 1
4. Solder 2
5. Heating, generation of soldering connection 3
6. Soldering with a soldering iron:
    a) Cleaning surfaces 3
    b) Temperature & time 3
    c) Adding solder & quantity 4
    d) Soldering : 7 steps 5
    e) Soldering of a Flat-package IC 6
    f) Incorrect soldering 7
7. Desoldering with warm air and soldering paste 8+9
8. Lead free soldering 10
9. Temperature specifications of the soldering bit 10
10. Types of soldering bits 11
11. Soldering devices 11

SOLDERING

Theoretical background for technicians

1. Introduction :

Progress within the electronics industry would not have resulted in the mass production of electronic appliances as we know them today, without a similar progress in the field of soldering technology.

To be able to understand how soldering techniques are used in the electronics industry, we first have to get familiar with the materials used during the soldering process, and we have to learn how a soldering connection is created on a molecular level.

Soldering is an old technique for creating permanent electrical and mechanical connections between metals. In contrast to ordinary glueing, which makes only a physical connection between materials (although electrically conducting glue also exists) soldering causes a chemical reaction with other materials, creating a new alloy.

There are several soldering techniques, which all have a few things in common: basic metals, soldering flux, solder and heating.

2. Basic metals :

When we place electronic components on a PCB, the legs of the components and the soldering lands on the PCB, are the basic metals that will connect with the solder.
Several metals like copper, bronze, silver and some types of steel react with solder and make a strong chemical and physical connection.
Other materials like aluminium, titanium, etc. are difficult or even impossible to solder.
The fact that materials exist which can not be soldered is very important, as these materials can be used for making soldering machines and soldering irons. These materials can also be used for covering components during soldering.

There is a direct relationship between how strongly the surface of the basic metal oxidizes and how the soldering reacts on this: When there is more oxidation present, the soldering connection will be weaker.
The fact that most metals oxidize very quickly when heated, is a specific problem, because if one wants to obtain a chemical soldering reaction, a high temperature is necessary…
Flux is used to counteract this oxidation problem.

3. Soldering flux :

Although the soldering surface looks clean, there is always a thin layer of oxide on the metals. This layer always arises as soon as the metal is exposed to air. Reliable soldering connections can only be made on "clean" surfaces. "Clean" surfaces could be accomplished using cleaning product, but this would be insufficient as metal oxidizes incredibly fast when heated. To prevent the creation of the layer of oxide, the use of soldering flux is necessary.

Using soldering flux on the basic metals before the soldering starts has various reasons and advantages:

The main reason of the use of flux is to stop the oxidation of the basic metals during heating. The flux makes sure that the air can't reach the soldering surface, and therefore prevents the formation of oxidation during warming up.
Most of the different flux types have acid components which are used to remove the oxidation that is already present on the basic metal.
If a very strong acid is active in the flux, theoretically it would be possible to remove all the oxides. This however is in practice not possible, as the acid would also affect the PCB and the components, which of course is not the intention.

Therefore there will always be a search for the balance between a flux that contains strong acids, and as a consequence also removes the oxidates more effectively (but also affects the PCB), and a flux that contains a less strong acid and therefore removes less oxides. (but does not affect the PCB).

Most fluxes though use types of acid. Therefore it generally is necessary to remove the flux residues after soldering.

When solder is applied, the flux needs to flow off so that the solder can make direct contact with the basic metal. During this process it is however impossible to avoid the flux mixing with the solder.
Flux developers make use of this fact to develop a flux that has the property to reduce the surface tension of the solder on contact, as a result of which better "wetting" arises.

Flux consists of natural or synthetic resin and chemical additives, called activators.
In PEM a "no clean" flux of Tamura is used (for the placement of Flat-Package-IC's). Consequently it is not necessary to remove the flux after soldering.
Furthermore it is not necessary to add extra flux when doing easy manual soldering because the soldering wire contains flux (3.3 %), like a resin core.

4. Solder :

There several existing metals and metal alloys that can be used for soldering.
The choice for using a certain metal is based on the properties of the metal:
Is the metal flexible or fragile? How well does it conduct heat?
Does it expand rapidly when heated? What is the electrical resistance of the metal?
How strong is it? Is it toxic? And not unimportant: what does it cost?
Although it is not perfect, an alloy of tin and lead is most often used traditionally.
Pure tin melts at 203°C and pure lead melts at 330°C. A mixture of both metals has a lower melting point.
By using a different mixture the melting temperature of the soldering alloy can be changed.
A alloy of tin and lead (e.g. 60/40) has a relatively low melting point (183°C) and can be produced at a relatively low price, in comparison to other alloys with similar properties.
Lead is a very cheap and can be found in a lot of places. Therefore the cost of the tin-lead alloy is mainly determined by the cost of tin.

When this alloy is warmed up, it goes through several phases:
the alloy goes from a "fixed" condition, to a "paste" condition (semi-liquid), and finally becomes totally liquid. It is difficult to solder with an alloy that is in a "paste" condition: because there is little movement or vibration during this "paste" condition, this can cause bad soldering.
If the alloy is of 50% tin and 50 % lead, it starts to melt at a temperature of 193°C, and will be totally melted at a temperature of 216°C.

To prevent bad soldering during the "paste" condition, "eutectical" solder is used.
An "eutectical" alloy is an alloy that, when heated, goes immediately from a fixed condition to a liquid condition: the "paste" stage is effectively skipped.
An "eutectical" tin-lead alloy contains 63% tin and 37% lead.
Often a tin-lead alloy with 60% tin and 40% lead is used. This alloy can be produced cheaper than the eutectical solder because it contains less tin. But because this alloy is almost "eutectical", it can effectively be used for soldering.
During soldering, be careful to make no sudden movements whilst the solder is cooling.
This can cause "interrupted connections" and therefore unreliable soldering.
An "interrupted" soldering connection has a rough and irregular surface and appears dim instead of clear and shiny.
It is also very important that no impurities get into the solder.
If small metal particles from the components get into the soldering "cup" of the soldering machine.
This can result in a serious change to surface tension, and consequently bad soldering connections.
These metal impurities could even change the melting temperature of the solder.

5. Heat, generation of soldering connection

If the solder gets in contact with a copper surface, the solder dissolves and penetrates the copper layer. The molecules of the solder and the copper get mixed and a new alloy is created, consisting partially of copper and partially of solder. This reaction is called "wetting" and has as a result an "inter-metallic" connection between the components. "Wetting" can only take place if the surface of the copper does not contain oxides. The solder and the surface of the copper also have to be at the right temperature before the "wetting" can take place.

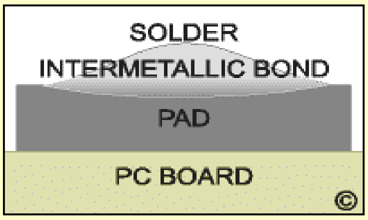

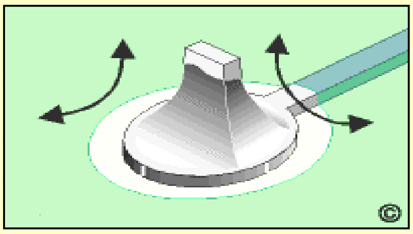

6. Soldering with a soldering iron:

The following 4 factors determine the quality of the soldering connection:

- clean surfaces
- temperature
- time
- quantity of solder

a) Clean surfaces:

"dirt is the enemy of a good soldering connection!"

It is enormously important that there are no impurities on all of the components (including the soldering bit). Soldering on "dirty" components is not possible!
Old PCB's or components sometimes can not be soldered due to a large layer of oxide on the islands and the legs of the components.
In such cases the solder is rejected and goes to everywhere, except for where we want it to go.
In a desperate attempt to solder in such a situation, a common reaction is to raise the soldering temperature. However, this is certainly not a solution. On the contrary!
The only correct solution is thorough cleaning. (e.g. with a brush of glass fibre)

Before the soldering iron can be used, the soldering bit first has to be coated with tin. Some solder has to be melted onto the bit. Afterwards the bit has to be cleaned. For cleaning the soldering bit, it is best is to use a slightly moistened sponge (with demineralized water). The use of ordinary tap water is not reccomended, because the silver layer on the bit gets affected by calcium in the tap water, which results in less thermal contact. This diminished thermal contact runs the risk of cold soldering, etc.
To improve the thermal contact between the bit and the soldering surface, it is advisable to coat the soldering bit with a little bit of solder before every soldering action.

b) Temperature & time :

The temperature of the soldering bit is not the most determining factor for successful soldering: the heating cycle (combination of temperature of the bit and the time the heating takes) is!
How fast the soldering components warm up, how warm they get and how long they stay warm, are the most important factors that have to be controlled to make a reliable soldering connection.

The time soldering takes is very important. A soldering task may only take a couple of seconds: when heated too long, the components and the soldering lands can be damaged.

It is also totally wrong to use the soldering bit to rub melted solder across the soldering surface.

Although the temperature of the bit is not the most determining factor, it is advisable to to do soldering at the lowest possible temperature.

The first that has to be examined before the soldering starts in the 'thermal mass" of the components.

Every soldering task has its own specific "thermal mass" which is determined by the heat absorption of the components. For soldering e.g. the component pins and solder lands.
For components on a heat sink there will be a large thermal mass. In other words: the heat will be "absorbed " by the heat sink, as a result of which more warmth will have to be added than e.g. for soldering the leg of an IC.
Secondly it is very important that the temperature of all the components that will be soldered have more or less the same temperature.
An example: it is necessary that the leg and the soldering land have had at the same time the necessary warming up before the solder is added.
We have to take care that, while warming up, the soldering bit gets in contact with both the leg and the soldering land.
"Thermal contact" is therefore very important.

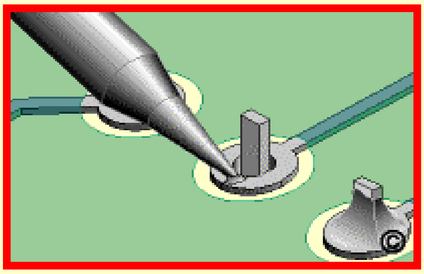

The thermal contact is low
The warmth is transmitted by means of the bit.
to the leg and the soldering land

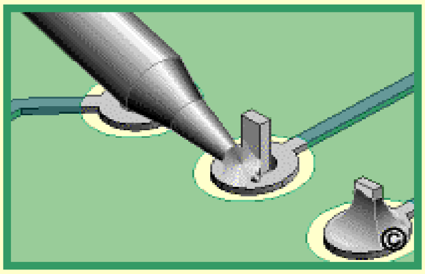

The thermal contact is higher
by adding a bit of solder
to the point of contact

The size and the shape of the soldering bit, as well as the position of the bit, will determine whether or not it makes a good thermal contact. For being able to make a good thermal contact in all circumstances, different types of soldering bits exist (different shapes).
Normally the soldering bit has to be placed on the "largest mass point" of the soldering connection that needs to be made.
This will make fast heating possible.

c) Adding solder & quantity

The last factor necessary for a high quality soldering connection, is adding the correct quantity of solder.
Too much solder is an unnecessary waste and can aid the formation of "bridges" across legs
If not enough solder is added, this can cause an "insufficient soldering" or "dry soldering"
Adding the correct quantity of solder takes a lot of exercise and experience

Melted solder always flows from a "colder" area to a "warmer" area.
Before the solder is added, the surface temperature of the components to be soldered has to be higher than the melting temperature of the solder.
Never let the soldering wire melt by holding it against the soldering bit. This may cause the solder to fall on a surface where the temperature is lower than the melting temperature
This can result in a cold soldering, with almost no or insufficient "wetting".
If the solder is added to a clean, fluxed and sufficiently heated surface, then the solder will melt and flow nicely without direct contact with the heat source. (soldering bit)
Therefore it is advisable to put the soldering wire on the sufficiently heated surface so that the solder melts as a consequence of the contact with the heated surface

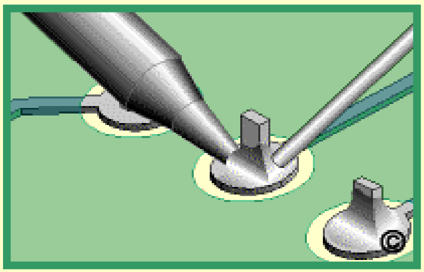

The solder is melted as a consequence of the contact with
the heated surface. No contact with the soldering bit is necessary.

d) Soldering : 7 steps

1. Cleaning of the soldering bit:

For soldering to be effective, the soldering bit has to be clean:
no dirt and no oxides are allowed on the soldering bit.
The soldering bit needs to be cleaned before soldering.

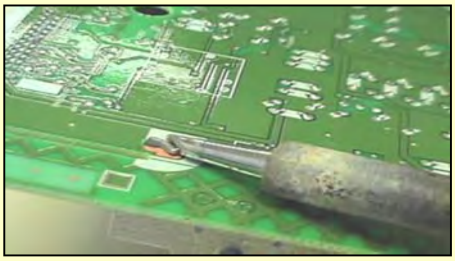

2. Heating of the components that will be soldered.

For soldering to be effective, the components which have to be connected, have to be heated (in this case the pin and the solder lands) to a temperature that at least equals the temperature of the melting point of the soldering alloy. (183 °C)
Bear in mind the "thermal mass", the type of bit and the position of the bit!

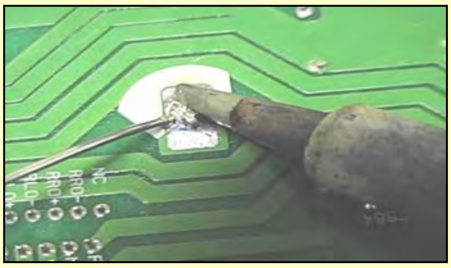

3. Adding the solder:

After both components are heated sufficiently, solder can be added.
Let the solder melt as a consequence of the contact with the heated surface. In other words: don't make any direct contact with the soldering bit.

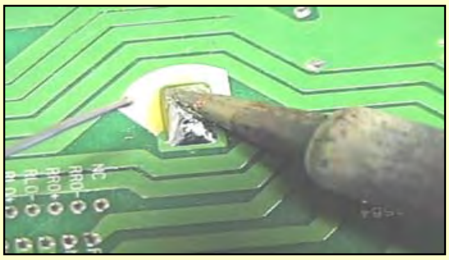

4. Quantity of the solder to be added and letting the solder flow.

Let the solder flow nicely between both components and add solder until the right quantity is reached.

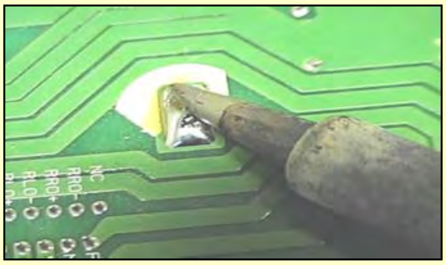

5. Remove the soldering wire

When sufficient solder is added, the soldering wire may be removed.

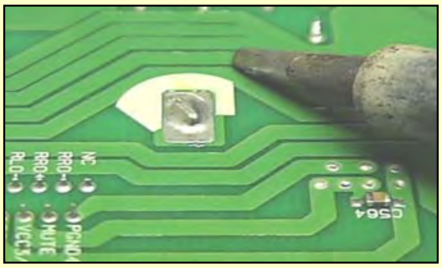

6. Remove the soldering bit.

When the solder has nicely flowed open, and there is "wetting", the soldering bit may be removed so that the solder can solidify.
Pay attention for possible sudden movements during this operation. (risk of "disturbed" soldering)

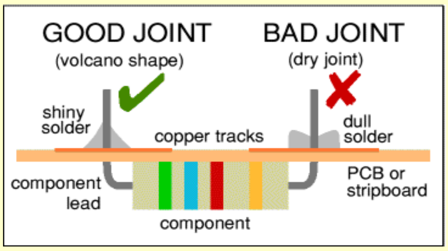

7. Check the soldering connection

Check the soldering connection.
Nicely formed soldering is shining and even. Also pay attention to the shape:
a spherical shape is not good!

e) Soldering of a flatpackage IC

For soldering a "Flat-package IC" (gull wing type) one can use, beside the ordinary point soldering (= leg by leg), a special soldering technique: the DRAGGING TECHNIQUE.
For this technique, one must use a special soldering bit that will, after filling-up, act as a kind of "mini-soldering-wave", and with which one can solder a whole row of legs in one ongoing movement

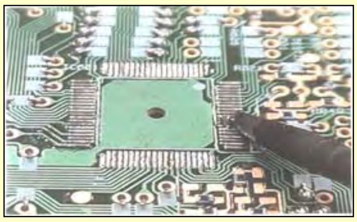

1. Cleaning of the soldering surface:

Remove all possible soldering residues (caused by desoldering) and make all the paths nicely even by going over those paths with the soldering bit. (without adding solder)

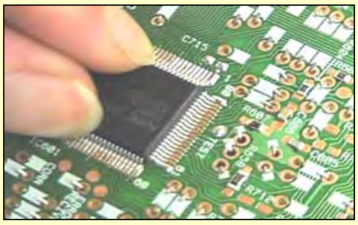

2. Positioning of the IC:

Position the IC on the patterns.
This has to be very precise.
A little movement can cause serious problems.

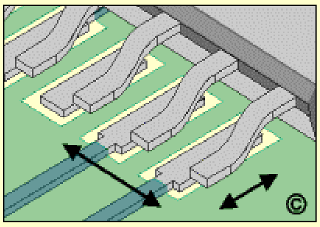

3. Fixing of the IC:

Fix the IC on the PCB, by "temporarily" soldering some legs on both sides.

4. Add sufficient flux on all sides.

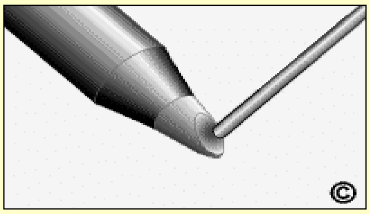

5. Filling of the bit and start of the soldering:

Fill the special (concave) soldering bit with solder.

Now start a slow dragging movement from one leg to another. Pay attention that the solder nicely moves up on the legs of the IC.

If the PCB is equipped with a "soldering thief" (a soldering land that is positioned next to the last leg), one has to solder in the direction of the soldering thief, so that at the end the possible surplus of solder flows on this soldering thief and no short circuit is created on the last two legs of the row.

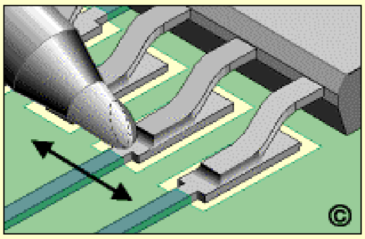

f) Incorrect soldering :

For certain soldering problems or soldering mistakes, specific terms are used:

SOLDER BRIDGE

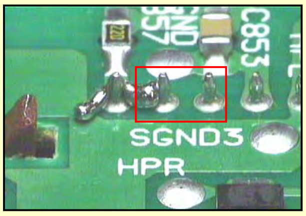
C853
SGND3
HPR

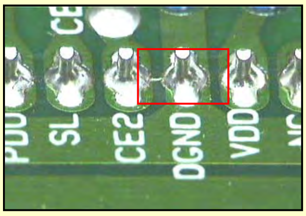
SL
CE2
OGND
VDD

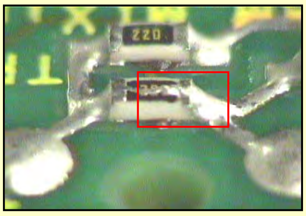

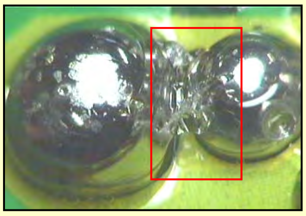

COLD SOLDERING

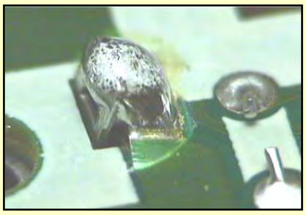

HALF SOLDERING OR DRY SOLDERING

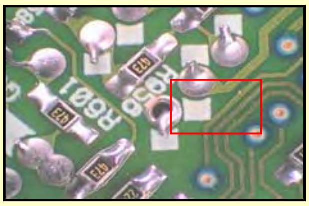

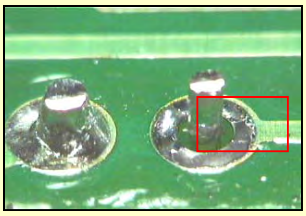

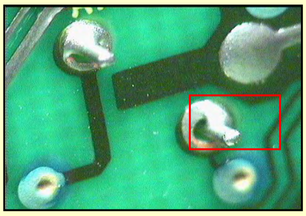

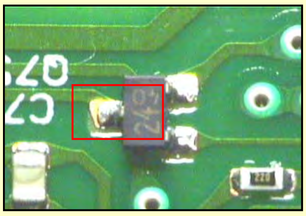

NO SOLDERING

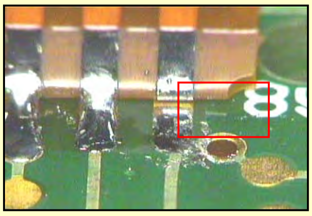

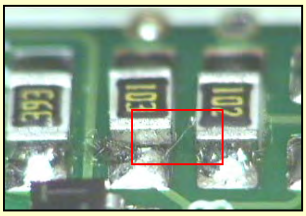

SOLDERING HORN

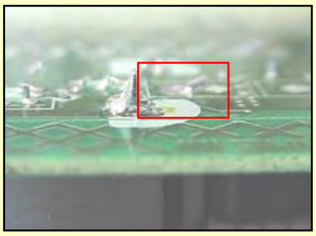

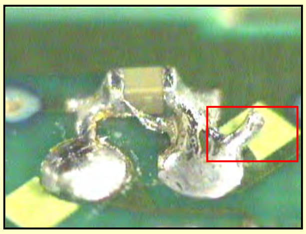

SOLDERING BALL OR SOLDERING SPLASH

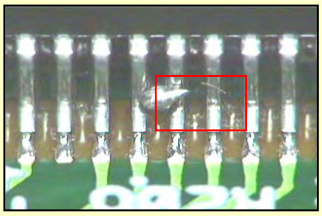

SOLDERING WIRE SHORT CIRCUIT

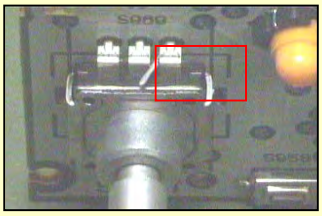

7) (De)soldering with warm air and soldering paste:

In some cases it is impossible to solder or desolder certain components with a classical soldering iron or a normal desoldering station.

This is often the case for SMT components, with contact legs (partly) underneath the component.
In such a case one has to use a (de)soldering appliance that uses warm air or infra-red heat.

SMT component with contact surfaces beneath the component

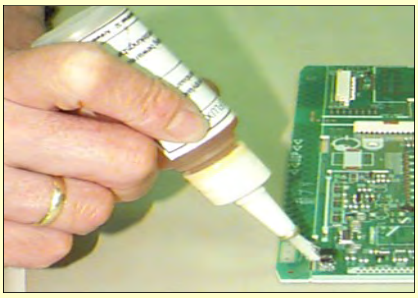

1. Fluxing of the component:
To have a good temperature conductance during warming up, the soldering surfaces need to be moistened with flux.
If these surfaces are parly underneath the component, flux can be put on the borders of the component.

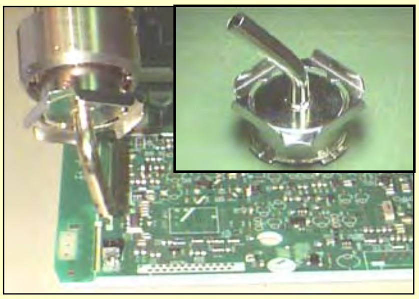

2. Warming-up of the soldered surfaces:
Use a "pointing nozzle" and point it at the component you want to remove. Keep a distance of 1 cm between the nozzle and the surface you want to warm up. Choose the temperature of the warm air e.g. +/- 340°C. (dependant on the temperature profile of the compents.)
Pay attention : if the contact surfaces are partly under the component, it is necessary to warm up the body of the component. Often the component is not useable afterwards.

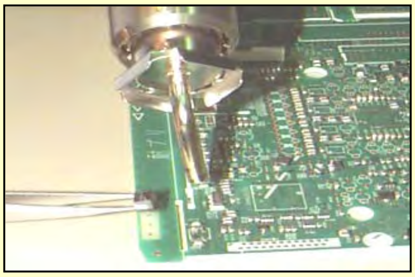

3. Removal of the component

When the solder (paste) has melted, the component can be removed with a pair of tweezers.

4. Cleaning of the soldering surfaces:
Remove all redundant soldering paste with a classical desoldering station.

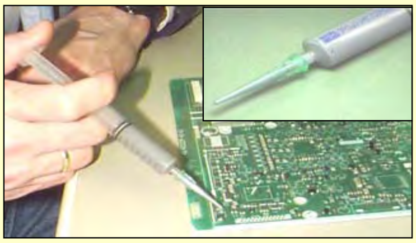

5. Adding soldering paste:
Add with a measured squirt the necessary soldering paste on the soldering lands.
Pay attention: soldering paste needs to be stored in strict circumstances. (see the instructions of the supplier)

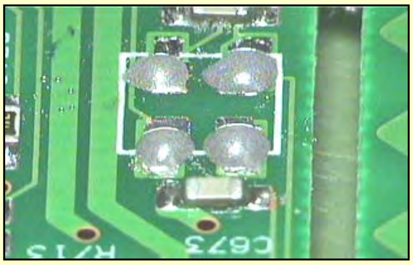

6. Checking the quantity of soldering paste:
Check the quantity of the added soldering paste.
Only the contact surfaces should be covered with soldering paste.
Too much soldering paste can cause a short circuit during "re-flowing".

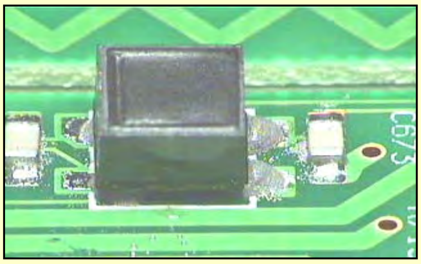

7. Positioning of the component

Position the new component carefully with the contact surfaces in the soldering paste with a pair of tweezers.
Position the component by the guide lines on the PCB.

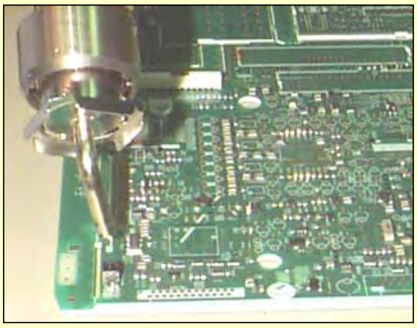

8. Reflowing of the component:

Use a "pointing nozzle" and point it on the component you want to remove. Keep a distance of 1 cm between the nozzle and the surface you want to warm up. Choose the temperature of the warm air e.g. +/- 340°C. (dependant on the temperature profile of the component.) Warm up until one sees that the soldering paste is active or flowing. The result should be a good soldered joint.
Pay attention: For most of the warm air soldering appliances one can choose the air pressure. The speed should be minimum, otherwise there is a possibility of blowing the component away.

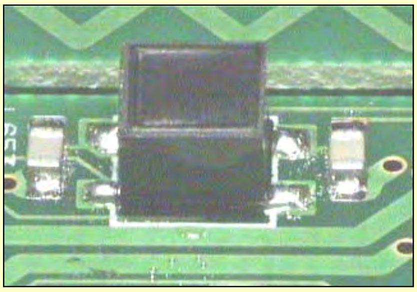

9. Checking of the soldering:

If possible, check the soldered joints.
Good soldering shines and has flowed evenly.

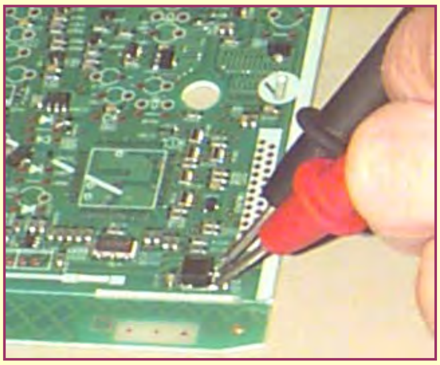

10. Checking of the electrical connection

If possible check the electrical connection with a meter.

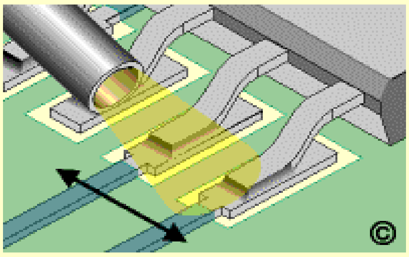

Example of a "warm air soldering appliance": PACE.

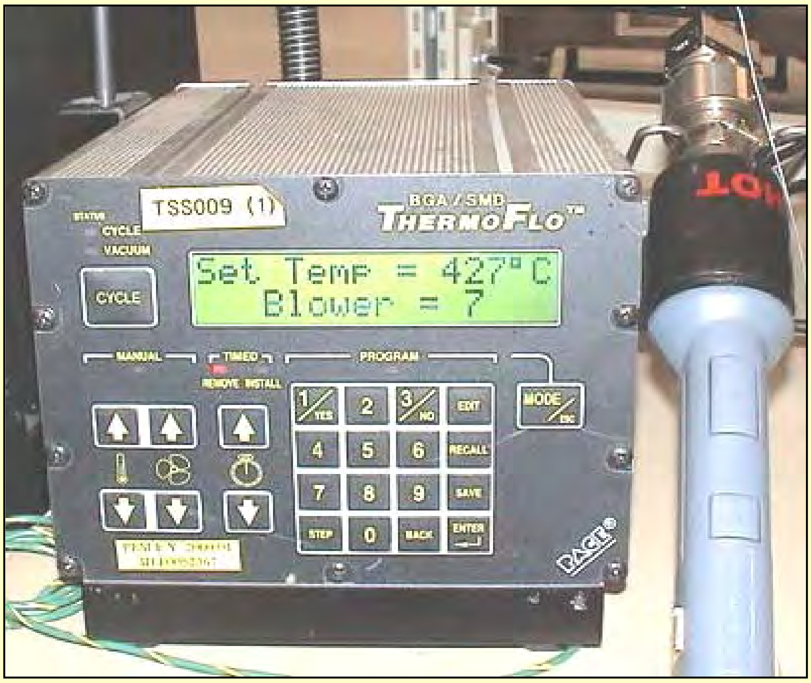
TSS009 (1)
BGA / SMD
THERMOFLO™
STATUS
CYCLE
VACUUM
CYCLE
MANUAL
TIMED
REMOVE INSTALL
PROGRAM
MODE/ESC
EDIT
RECALL
SAVE
ENTER
STEP
BACK
PACE
HOT

Set Temp = 427°C
Blower = 7

8) Lead free soldering

Introduction.

Normally all solderings (mechanical - and hand soldering) were executed with soldering material that contained lead. (40 % lead / 60 % tin)

But lead is toxic and therefore damaging for humans and for nature.

Hence European directives were made in which companies will be obliged to use soldering techniques without lead after July 2006.

Soldering without lead is relatively new in the world of electronics.

Lead-free solder:

If no lead is present in the soldering wire, this has to be replaced. The soldering wire needs to contain another metal, like SILVER & COPPER. (96,5 % tin / 3% silver / 0,5 % copper)

This new composition has the advantage that it is no longer toxic and that the soldering connections are stronger. (lead is relatively weak). But this new composition also has some negative consequences:

1) The cost is higher (silver and copper alloys are more expensive than lead)
2) The joints look different (more granular instead of shining)
3) The melting temperature is higher

One of the main points of interest is the higher melting temperature. One would think that as a consequence of the higher melting temperature of the solder, also the soldering temperature needs to be raised. This however is not always allowed! Some components are only able to handle a certain maximum temperature for a certain time.

In practice.

- You can use your original soldering station if you can reach the correct temperature at the soldertip
- The soldering bit has to be a bit larger than a soldering bit for soldering with lead.

Pay attention: **Solder with lead and solder without lead can never be mixed.**

9) Temperature specifications of the soldering bit

Solder with lead

Process	**Bit Type**	**Solder**	**Min**	**Center**	**Max**
Pins	ETB / ETCC/ 1121-0519	UOS-014	340°C	**380°C**	400°C
LCD	ETF / ETCC	UOS-005	320°C	**340°C**	360°C
OEL Display	ETCC	UOS-004	280°C	**290°C**	300°C
Connector	ETL / ETF	UOS-005	320°C	**340°C**	360°C
Flexible PCB	ETF	UOS-004	280°C	**300°C**	320°C
IC's & QFP IC's	1121-0490	UOS-005	280°C	**300°C**	320°C
Removal bridge	ETB / ETCC	UOS-005	320°C	**340°C**	360°C
Lights	ETF / ETB /1121-0519	UOS-005	280°C	**300°C**	320°C
SMT/SMD repair	ETL / ETF	UOS-005	280°C	**300°C**	320°C

Temp. specifications Pioneer standard ZES-A015 & ZES-B016

Solder without lead

Process	**Bit Type**	**Solder**	**Min**	**Center**	**Max**
Pins	LTB / LTCC	UOS-026	400°C	**410°C**	420°C
LCD	LTF / LTCC	UOS-027	320°C	**340°C**	360°C
OEL Display	LTCC	UOS-028	280°C	**300°C**	320°C
Connector	LTL / LTF	UOS-027	320°C	**340°C**	360°C
Flexible PCB	LTF	UOS-028	280°C	**300°C**	320°C
IC's & QFP IC's	1121-0490	UOS-027	280°C	**300°C**	320°C
Removal bridge	LTB / LTCC	UOS-027	320°C	**340°C**	360°C
Lights	LTF / LTB	UOS-027	280°C	**300°C**	320°C
SMT/SMD repair	LTL / LTF	UOS-027	280°C	**300°C**	320°C

Temp. specifications Pioneer standard ZES-A043

Types of soldering wire:

With lead	UOS-014 (1 mm) + UOS-005 (0.8 mm) + UOS-004 (0.5mm)
Without lead	UOS-026 (1 mm) + UOS-027 (0.8 mm) + UOS-028 (0.5 mm)

10) Types of soldering bits:

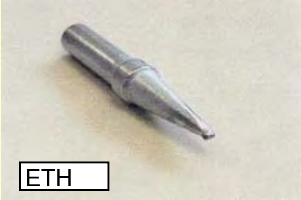
ETH

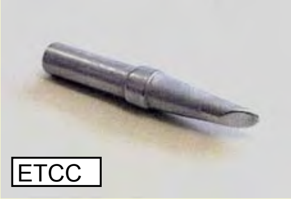
ETCC

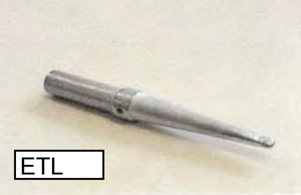
ETL

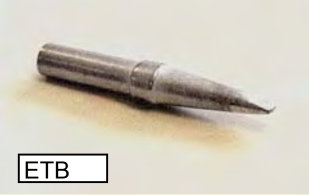
ETB

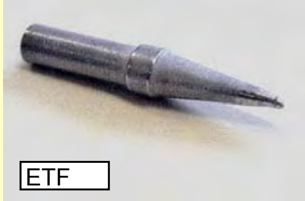
ETF

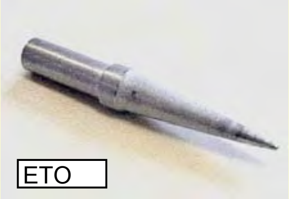
ETO

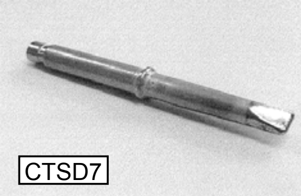
CTSD7

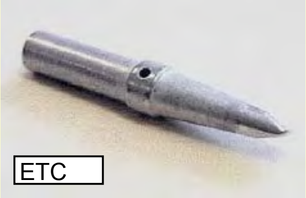
ETC

11) Soldering devices:

"Clean-o-point": for automatic cleaning of the soldering bit with 2 rotating sponges

SOLDERING FLU
Bevat isopropyla
Licht ontvlambaa

"Flux dosing bottle": for easy adding of flux (with brush)

"Bit cleaner": for drying and cleaning of the soldering bit

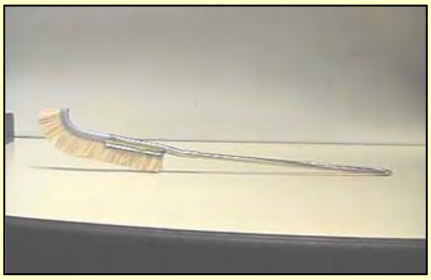

" Brush" : for removal of possible soldering residues after soldering

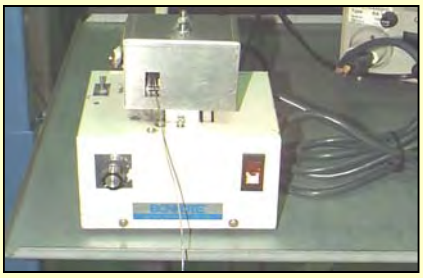

"V-cutter" : for making a cut in the soldering wire, in order to reduce the formation of little soldering balls

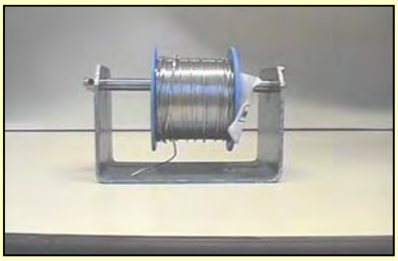

"Solder dispenser" : for a simple supply of soldering wire

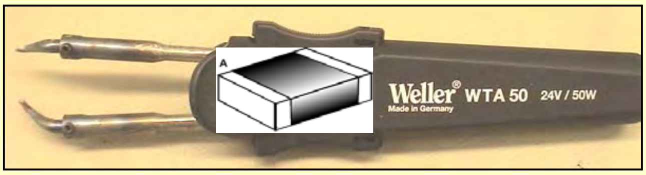
Weller WTA 50 24V / 50W
Made in Germany

"Tweezer-copper bit": handy tool for the removal of SMD/SMT components.

THERMOFLO
Set Temp = 427°C
Blower = 7

"Warm air desoldering appliance": for easy desoldering of flat-package IC's

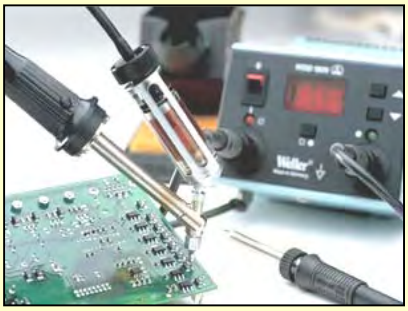

"Desoldering appliance": for easy desoldering of point solderings

"Soldering magnifying glass": for a good view during soldering

The Basic Soldering Guide

http://www.epemag.wimborne.co.uk/solderfaq.htm (for text in this document)

basics for wire (Good intro) : http://www.youtube.com/watch?v=BLfXXRfRIzY
The Happy Soldering Iron: http://www.youtube.com/watch?v=AL-_RGbyf1s
basics: http://www.youtube.com/watch?v=I_NU2ruzyc4
Watt/temp comparison http://www.youtube.com/watch?v=Vh9pWu6K6tc&NR=1
tip sizes: http://www.youtube.com/watch?v=Sfb1Ve52ztY&NR=1
types of solder: http://www.youtube.com/watch?v=COqGkYMOA44
maintenance: http://www.youtube.com/watch?v=krxTfZCFptk&NR=1

Good overview with some odd bits: http://www.youtube.com/watch?v=AOdnGUMi7lQ

Desoldering overview: http://www.youtube.com/watch?v=j-_pnc-Qqm8
2nd good overview: http://www.youtube.com/watch?v=8UN3D2-f64A
Hot air pencil/gun: http://www.youtube.com/watch?v=AxYhF6Ab2CU
Desoldering gun (short): http://www.youtube.com/watch?v=8Z6MvZz_uNc

Crazy Guy from Make Electronics: http://www.youtube.com/watch?v=3N3ApzmyjzE
Not Soldering!! – Bread board: http://www.youtube.com/watch?v=oiqNaSPTI7w

This written guide will help beginners and novices to obtain effective results when soldering **electronic components**. If you have little or no experience of using a soldering iron, then EPE *(Everyday Practical Electronics* magazine) recommends that you practice your soldering technique on some fresh surplus components and clean stripboard (protoboard), before experimenting with a proper constructional project. This will help you to avoid the risk of disappointment when you start to assemble your first prototypes. If you've never soldered before, then read on!

Soldering irons

The most fundamental skill needed to assemble any electronic project is that of ***soldering***. It takes some practice to make the perfect joint, but, like riding a bicycle, once learned is never forgotten! The idea is simple: to join electrical parts together to form an electrical connection, using a molten mixture of lead and tin (solder*) with a soldering iron. A large range of soldering irons is available - which one is suitable for you depends on your budget and how serious your interest in electronics is.

[*Note: the use of lead in solder is now increasingly prohibited in many countries. "Lead free" solder is now statutory instead.]

Electronics catalogues often include a selection of well-known brands of soldering iron. Excellent British-made ones include the universally popular Antex, Adcola and Litesold makes. Other popular brands include those made by Weller and Ungar. A very basic mains electric soldering iron can cost from under £5 (US$ 8), but expect a reasonable model to be approximately £10-£12 (US$ 16 - 20) - though it's possible to spend into three figures on a soldering iron "station" if you're really serious! Check some suppliers' catalogues for some typical types. Certain factors you need to bear in mind include:-

Voltage: most irons run from the mains at 240V. However, low voltage types (e.g. 12V or 24V) generally form part of a "soldering station" and are designed to be used with a special controller made by the same manufacturer.

Wattage: Typically, they may have a power rating of between 15-25 watts or so, which is fine for most work. A higher wattage does ***not*** mean that the iron runs hotter - it simply means that there is more power in reserve for coping with larger joints. This also depends partly on the design of the "bit" (the tip of the iron). Consider a higher wattage iron simply as being more "unstoppable" when it comes to heavier-duty work, because it won't cool down so quickly.

Temperature Control: the simplest and cheapest types don't have any form of temperature regulation. Simply plug them in and switch them on! Thermal regulation is "designed in" (by physics, not electronics!): they may be described as "thermally balanced" so that they have some degree of temperature "matching" but their output will otherwise not be controlled. Unregulated irons form an ideal general purpose iron for most users, and they generally cope well with printed circuit board soldering and general interwiring. Most of these "miniature" types of iron will be of little use when attempting to solder large joints (e.g. very large terminals or very thick wires) because the component being soldered will "sink" heat away from the tip of the iron, cooling it down too much. (This is where a higher wattage comes in useful.)

A proper temperature-controlled iron will be quite a lot more expensive - retailing at say £40 (US$ 60) or more - and will have some form of built-in thermostatic control, to ensure that the temperature of the bit (the tip of the iron) is maintained at a fixed level (within limits). This is desirable especially during more frequent use, since it helps to ensure that the temperature does not "overshoot" in between times, and also guarantees that the output will be relatively stable. Some irons have a bimetallic strip thermostat built into the handle which gives an audible "click" in use: other types use all-electronic controllers, and some may be adjustable using a screwdriver.

Yet more expensive still, ***soldering stations*** cost from £70 (US$ 115) upwards (the iron may be sold separately, so you can pick the type you prefer), and consist of a complete bench-top control unit into which a special *low-voltage* soldering iron is plugged. Some versions might have a built-in digital temperature readout, and will have a control knob to enable you to vary the setting. The temperature could be boosted for soldering larger joints, for example, or for using higher melting-point solders (e.g. silver solder). These are designed for the most discerning users, or for continuous production line/ professional use. The best stations have irons which are well balanced, with comfort-grip handles which remain cool all day. A thermocouple will be built into the tip or shaft, which monitors temperature.

Anti-static protection: if you're interested in soldering a lot of static-sensitive parts (e.g. CMOS chips or MOSFET transistors), more advanced and expensive soldering iron stations use static-dissipative materials in their construction to ensure that static does not build up on the iron itself. You may see these listed as "ESD safe" *(electrostatic discharge proof)*. The cheapest irons won't necessarily be ESD-safe but never the less will still probably perform perfectly well in most hobby or educational applications, if you take the usual anti-static precautions when handling the components. The tip would need to be well earthed (grounded) in these circumstances.

Bits: it's useful to have a small selection of manufacturer's bits (soldering iron tips) available with different diameters or shapes, which can be changed depending on the type of work in hand. You'll probably find that you become accustomed to, and work best with, a particular shape of tip. Often, tips are *iron-coated* to preserve their life, or they may be bright-plated instead. Copper tips are seldom seen these days.

Spare parts: it's nice to know that spare parts may be available, so if the element blows, you don't need to replace the entire iron. This is especially so with expensive irons. Check through some of the larger mail-order catalogues.

A ***solder gun*** is a pistol-shaped iron, typically running at 100W or more, and is completely unsuitable for soldering modern electronic components: they're too hot, heavy and unwieldy for micro-electronics use. Plumbing, maybe..!

Soldering irons are best used along with a heat-resistant ***bench-type holder***, so that the hot iron can be safely parked in between use. Soldering stations already have this feature, otherwise a separate soldering iron stand is essential, preferably one with a holder for tip-cleaning sponges. Now let's look at how to use soldering irons properly, and how to put things right when a joint goes wrong.

The Basic Soldering Guide Photo Gallery

Soldering is a delicate manual skill which only comes with practice. Remember that your ability to solder effectively will determine directly how well the prototype or product functions during its lifespan. Poor soldering can be an expensive business - causing product failure and downtime, engineer's maintenance time and customer dissatisfaction. At hobbyist level, bad soldering technique can be a cause of major disappointment which damages your confidence. It needn't be like that: soldering is really easy to learn, and like learning to ride a bike, once mastered is never forgotten!

These photos illustrate the basic steps in making a perfect solder joint on a p.c.b. If you're a beginner, our advice is that it's best to practice your soldering technique using some clean, new parts with perhaps some new stripboard (protoboard). Be sure to avoid using old, dirty parts; these can be difficult if not impossible to solder.

(Left) Printed circuit board copper tracks must be clean to begin with, especially if they're not previously "tinned" with solder. Clean any raw p.c.b. copper tracks gently with e.g. an abrasive rubber block available from electronics suppliers. **(Right)** Clean the iron "bit" (soldering iron tip) using a damp sponge. The soldering iron featured is an Ungar Concept 2100 Soldering Station. Other popular brands of soldering equipment include Weller and Antex.

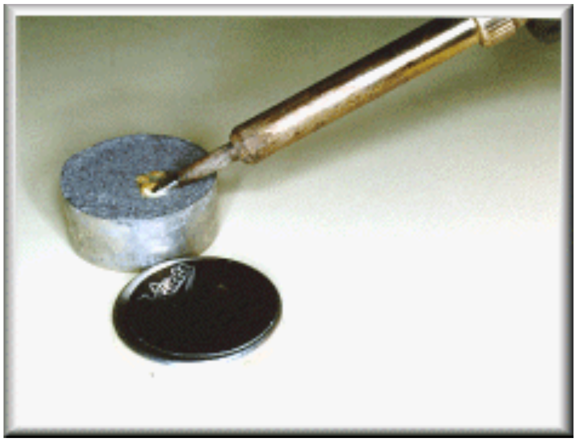

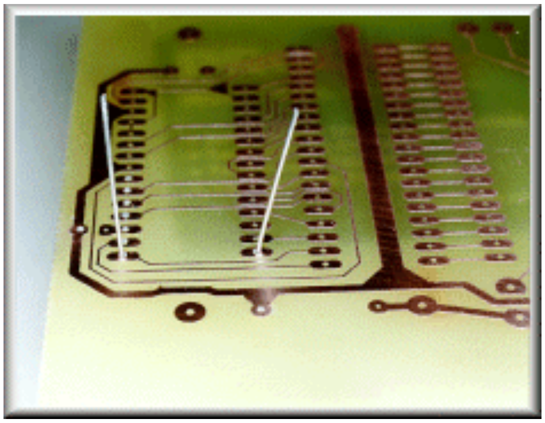

(Left) A useful product is Multicore's Tip Tinner Cleaner (TTC) - a 15 gramme tin of special paste which cleans and "tins" the soldering iron iron, in one go. New tips must be tinned **immediately** when used for the first time. **(Right)** Insert components and splay the leads so that the part is held in place.

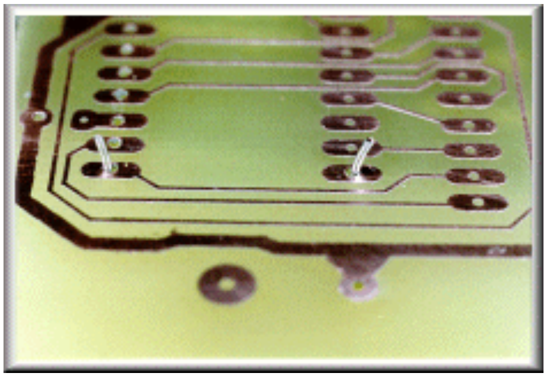

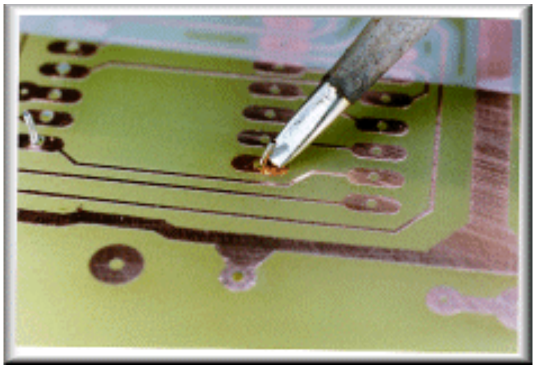

(Left) It's usually best to snip the electronic component wires to length prior to soldering. This helps prevent transmitting mechanical shocks to the copper foil.**(Right)** Apply a clean soldering iron tip to the copper solder pad and the component lead, in order to heat *both* items at the same time.

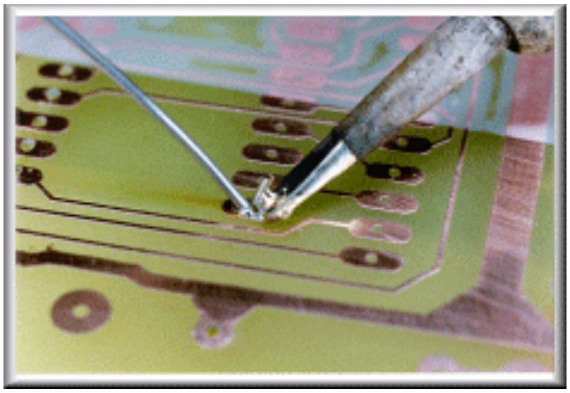

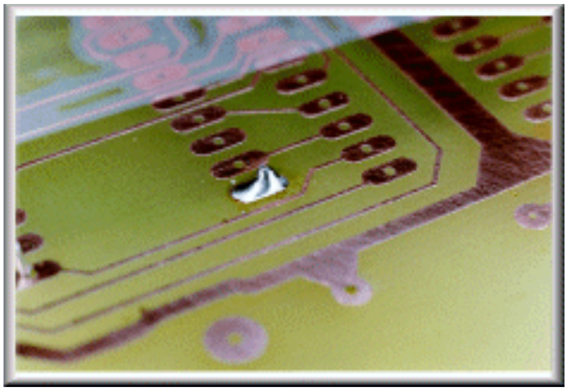

(Left) Continue heating and apply a few millimetres of solder. Remove the iron and allow the solder joint to cool naturally. **(Right)** It only takes a second or two, to make the perfect joint, which should be nice and shiny. Check the Guide for troubleshooting help.

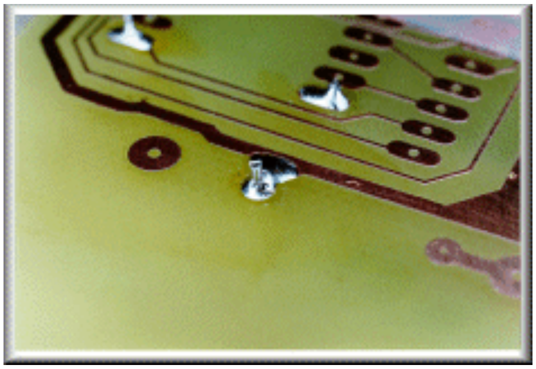

An example of a "dry" or "gray" soller joint - the solder failed to flow, and instead beaded to form globules around the wire.

How to solder

Turning to the actual techniques of soldering, firstly it's best to *secure the work* somehow so that it doesn't move during soldering and affect your accuracy. In the case of a printed circuit board, various holding frames are fairly popular especially with densely populated boards: the idea is to insert all the parts on one side ("stuffing the board"), hold them in place with a special foam pad to prevent them falling out, turn the board over and then snip off the wires with cutters before making the joints. The frame saves an awful lot of turning the board over and over, especially with large boards. Other parts could be held firm in a modeller's small vice, for example.

Solder joints may need to possess some degree of mechanical strength in some cases, especially with wires soldered to, say, potentiometer or switch tags, and this means that the wire should be looped through the tag and secured before solder is applied. The down side is that it is more difficult to *de-solder* the joint (see later) and remove the wire afterwards, if required. Otherwise, in the case of an ordinary circuit board, components' wires are bent to fit through the board, inserted flush against the board's surface, splayed outwards a little so that the part grips the board, and then soldered.

In my view - opinions vary - it's generally better to snip the surplus wires leads off *first*, to make the joint more accessible and avoid applying a mechanical shock to the p.c.b. joint. However, in the case of semiconductors, I often tend to leave the snipping until *after* the joint has been made, since the excess wire will help to sink away some of the heat from the semiconductor junction. Integrated circuits can either be soldered directly into place if you are confident enough, or better, use a dual-in-line socket to prevent heat damage. The chip can then be swapped out if needed.

Parts which become hot in operation (e.g. some resistors), are raised above the board slightly to allow air to circulate. Some components, especially large electrolytic capacitors, may require a mounting clip to be screwed down to the board first, otherwise the part may eventually break off due to vibration.

The perfectly soldered joint will be nice and shiny looking, and will prove reliable in service. I would say that:

- cleanliness
- temperature
- time
- adequate solder coverage

are the key factors affecting the quality of the joint. A little effort spent now in soldering the perfect joint may save you - or somebody else - a considerable amount of time in troubleshooting a defective joint in the future. The basic principles are as follows.

Really Clean

First, and without exception, all parts - including the iron tip itself - must be clean and free from contamination. Solder just will not "take" to dirty parts! Old components or copper board can be notoriously difficult to solder, because of the layer of oxidation which builds up on the surface of the leads. This repels the molten solder and this will soon be evident because the solder will "bead" into globules, going everywhere except where you need it. Dirt is the enemy of a good quality soldered joint!

Hence, it is an absolute necessity to ensure that parts are free from grease, oxidation and other contamination. In the case of old resistors or capacitors, for example, where the leads have started to oxidise, use a small hand-held file or perhaps scrape a knife blade or rub a fine emery cloth over them to reveal fresh metal underneath. Stripboard and copper printed circuit board will generally oxidise after a few months, especially if it has been fingerprinted, and the copper strips can be cleaned using an abrasive rubber block, like an aggressive eraser, to reveal fresh shiny copper underneath.

Also available is a fibre-glass filament brush, which is used propelling-pencil-like to remove any surface contamination. These tend to produce tiny particles which are highly irritating to skin, so avoid accidental contact with any debris. Afterwards, a wipe with a rag soaked in cleaning solvent will remove most grease marks and fingerprints. After preparing the surfaces, avoid touching the parts afterwards if at all possible.

Another side effect of having dirty surfaces is the tendency for people to want to apply *more heat* in an attempt to "force the solder to take". This will often do more harm than good because it may not be possible to burn off any contaminants anyway, and the component may be overheated. In the case of semiconductors, temperature is quite critical and they may be harmed by applying such excessive heat.

Before using the iron to make a joint, it should be "tinned" (coated with solder) by applying a few millimetres of solder, then wiped on a damp sponge preparing it for use: you should always do this immediately with a new bit, anyway. Personally, I always re-apply a very small amount of solder again, mainly to improve the thermal contact between the iron and the joint, so that the solder will flow more quickly and easily. It's sometimes better to tin larger parts as well before making the joint itself, but it isn't generally necessary with p.c.b. work. (All *EPE* printed circuit boards are "roller-tinned" to preserve their quality and to help with soldering.) A worthwhile product is Weller's *Tip Tinner & Cleaner*, a small 15 gram tinlet of paste onto which you dab a hot iron - the product cleans and tins the iron ready for use. An equivalent is Adcola *Tip-Save*.

Normal electronics grade solder is now "lead free" and typically contains Sn 97 Ag 2.5 Cu 0.5 (i.e. 97% tin, 2.5% silver and 0.5% copper). **It already contains cores of "flux"** which helps the molten solder to flow more easily over the joint. Flux removes oxides which arise during heating, and is seen as a brown fluid bubbling away on the joint. The use of separate acid flux paste (e.g. as used by plumbers) should NEVER be necessary in normal electronics applications because electronics-grade solder already contains the correct grade of flux! Other solders are available for specialist work, including aluminium and silver-solder. Different solder diameters are produced, too; 20-22 SWG (19-21 AWG) is 0.91-0.71mm diameter and is fine for most work. Choose 18 SWG (16 AWG) for larger joints requiring more solder.

Temperature

Another step to successful soldering is to ensure that the **temperature** of *all* the parts is raised to roughly the same level before applying solder. Imagine, for instance, trying to solder a resistor into place on a printed circuit board: it's far better to heat *both* the copper p.c.b. *and* the resistor lead at the same time before applying solder, so that the solder will flow much more readily over the joint. Heating one part but not the other is far less satisfactory joint, so strive to ensure that the iron is in contact with *all* the components first, before touching the solder to it. The melting point of most solder is in the region of 188°C (370°F) and the iron tip temperature is typically 330-350°C (626°-662°F). The latest lead-free solders typically require a higher temperature.

Now is the time

Next, the joint should be heated with the bit for just the right amount of time - during which a short length of solder is applied to the joint. Do not use the iron to carry molten solder over to the joint! Excessive time will damage the component and perhaps the circuit board copper foil too! Heat the joint with the tip of the iron, then continue heating whilst applying solder, then remove the iron and allow the joint to cool. This should take only a few seconds, with experience. The heating period depends on the temperature of your iron and size of the joint - and larger parts need more heat than smaller ones - but some parts (semiconductor diodes, transistors and i.c.s), are sensitive to heat and should not be heated for more than a few seconds. Novices sometimes buy a small clip-on heat-shunt, which resembles a pair of aluminium tweezers. In the case of, say, a transistor, the shunt is attached to one of the leads near to the transistor's body. Any excess heat then diverts up the heat shunt instead of into the transistor junction, thereby saving the device from over-heating. Beginners find them reassuring until they've gained more experience.

Solder Coverage

The final key to a successful solder joint is to apply an appropriate amount of solder. Too much solder is an unnecessary waste and may cause short circuits with adjacent joints. Too little and it may not support the component properly, or may not fully form a working joint. How much to apply, only really comes with practice. A few millimetres only, is enough for an "average" p.c.b. joint, (if there is such a thing).

Here's a summary of how to make the perfect solder joint.

1. All parts must be clean and free from dirt and grease.
2. Try to secure the work firmly.
3. "Tin" the iron tip with a small amount of solder. Do this immediately, with new tips being used for the first time.
4. Clean the tip of the hot soldering iron on a damp sponge.
5. Many people then add a tiny amount of fresh solder to the cleansed tip.

6. Heat all parts of the joint with the iron for under a second or so.
7. Continue heating, then apply sufficient solder only, to form an adequate joint.
8. Remove and return the iron safely to its stand.
9. It only takes two or three seconds at most, to solder the average p.c.b. joint.
10. Do not move parts until the solder has cooled.

Troubleshooting Guide

- Solder won't "take" - grease or dirt present - desolder and clean up the parts. Or, material may not be suitable for soldering with lead/tin solder (eg aluminium).
- Joint is crystalline or grainy-looking - has been moved before being allowed to cool, or joint was not heated adequately - too small an iron/ too large a joint.
- Solder joint forms a "spike" - probably overheated, burning away the flux.

Desoldering methods

A soldered joint which is improperly made will be electrically "noisy", unreliable and is likely to get worse in time. It may even not have made any electrical connection at all, or could work initially and then cause the equipment to fail at a later date! It can be hard to judge the quality of a solder joint purely by appearances, because you cannot say how the joint actually formed on the *inside*, but by following the guidelines there is no reason why you should not obtain perfect results.

A joint which is poorly formed is often called a "dry joint". Usually it results from dirt or grease preventing the solder from melting onto the parts properly, and is often noticeable because of the tendency of the solder not to "spread" but to form beads or globules instead, perhaps partially. Alternatively, if it seems to take an inordinately long time for the solder to spread, this is another sign of possible dirt and that the joint may potentially be a dry one.

There will undoubtedly come a time when you need to *remove* the solder from a joint: possibly to replace a faulty component or fix a dry joint. The usual way is to use a ***desoldering pump*** or vacuum pump which works like a small spring-loaded bicycle pump, only in reverse! (More demanding users using CMOS devices might need a pump which is ESD safe.) A spring-loaded plunger is released at the push of a button and the molten solder is then drawn up into the pump. It may take one or two attempts to clean up a joint this way, but a small desoldering pump is an invaluable tool especially for p.c.b. work.

Sometimes, it's effective to actually *add more* solder and then desolder the whole lot with a pump, if the solder is particularly awkward to remove. Care is needed, though, to ensure that the boards and parts are not damaged by excessive heat; the pumps themselves have a P.T.F.E. nozzle which is heat proof but may need replacing occasionally.

An excellent alternative to a pump is to use *desoldering braid*, including the famous American "Soder-Wick" (sic) or Adcola "TISA-Wick" which are packaged in small dispenser reels. This product is a specially treated fine copper braid which draws molten solder up into the braid where it solidifies. The best way is to use the tip of the hot iron to press a short length of braid down onto the joint to be de-soldered. The iron will subsequently melt the solder, which will be drawn up into the braid. Take extreme care to ensure that you don't allow the solder to cool with the braid adhering to the work, or you run the risk of damaging p.c.b. copper tracks when you attempt to pull the braid off the joint. See my photo gallery for more details.

I recommend buying a small reel of de-soldering braid, especially for larger or difficult joints which would take several attempts with a pump. It is surprisingly effective, especially on difficult joints where a desoldering pump may prove a struggle.

The Basic De-soldering Guide Photo Gallery (and Black Museum of Bad Soldering)

De-soldering is required when electronic components need to be removed from a circuit, usually because they are faulty. It may sometimes be necessary during testing or assembly, if a wrong part has been fitted or a modification has to be made. In the field, it's not uncommon for faulty electronic components to be swapped out, or poor joints (perhaps "dry" or gray joints) to need re-making properly, months or years after manufacture. Experienced engineers can often diagnose a particular faulty joint immediately, because they may have seen the same problem on similar electronic equipment before, especially if the equipment has a "reputation". A proper desoldering technique can soon be acquired with practice - all you need to do is buy some scrap boards to have a go with, and desolder to your heart's content!

The next photo sequence illustrates the basic steps for desoldering a printed circuit board, in order to remove a faulty part. Both the technique for using a desoldering suction pump as well as desolder braid are illustrated. Some real-life examples of poor soldering are shown too, in my *Black Museum of Bad Soldering*!

Remember - it costs just as much to get it right as it does to get it wrong! Practice makes perfect.

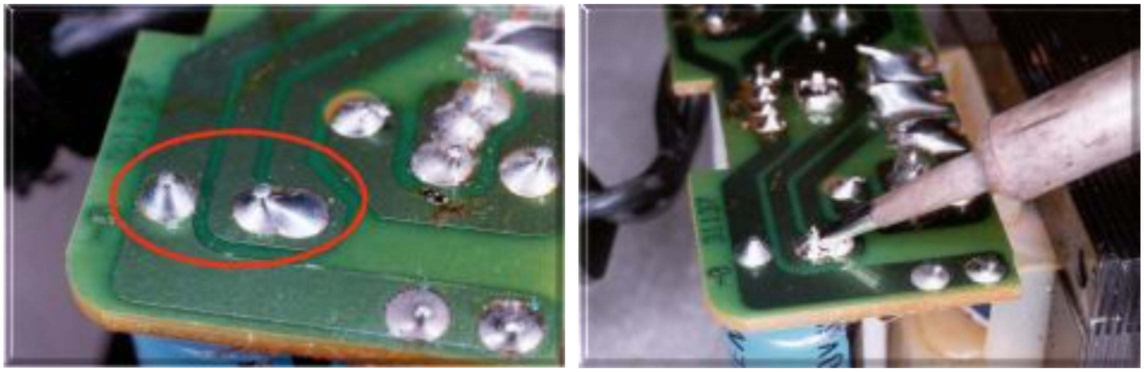

(Left) The two solder joints to be desoldered, to enable a faulty electrolytic capacitor to be removed from the printed circuit board.

(Right) If using a suction-type desoldering pump, apply the soldering iron tip first to melt the solder joint (say for 1-2 seconds). Ensure the spring-loaded desoldering pump is 'primed' and ready to go...

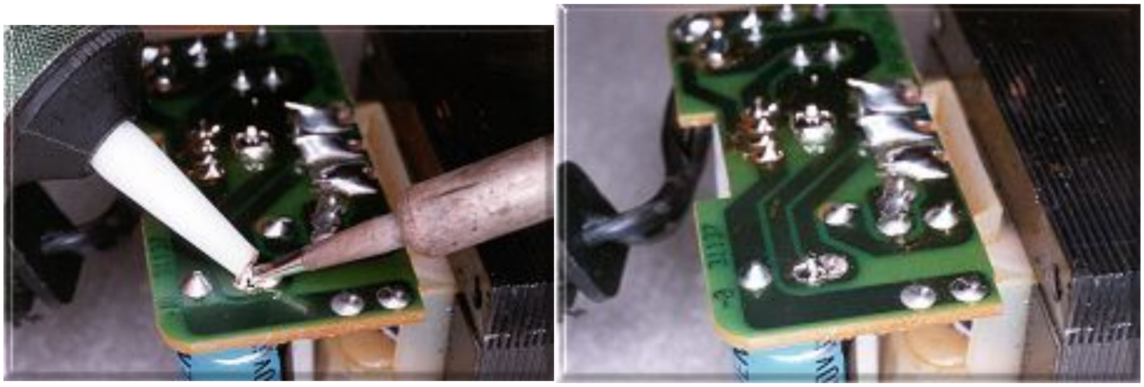

(Left) The PTFE nozzle of the desoldering pump is applied to the molten solder and the spring-loaded plunger is then immediately released, drawing the molten solder up into the pump. Remove the soldering iron tip. Repeat the process if needed. *Handy tip*: sometimes it helps to add some fresh solder and then desolder the whole joint.

(Right) The first p.c.b. joint, now desoldered. The second joint will be desoldered using traditional desoldering braid.

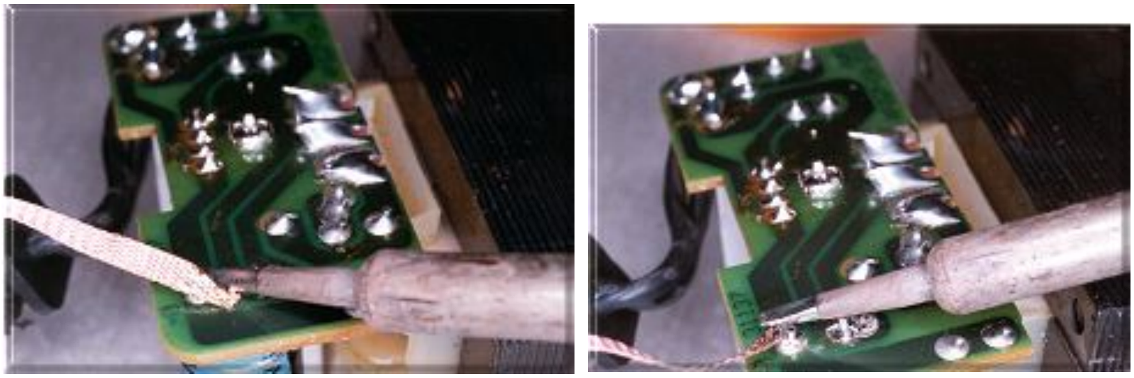

(Left) Select a suitable width of desoldering braid, and **press it down** onto the **COLD** joint using the hot tip of the iron. A flat soldering iron bit is preferable.

(Right) The molten solder is drawn up by capillary action into the desoldering braid. Take care not to overheat the board (the p.c.b. copper track may lift off), nor 'drag whiskers' of solder over the board, nor let the braid solidify onto the joint! Remove the braid while the joint is still molten.

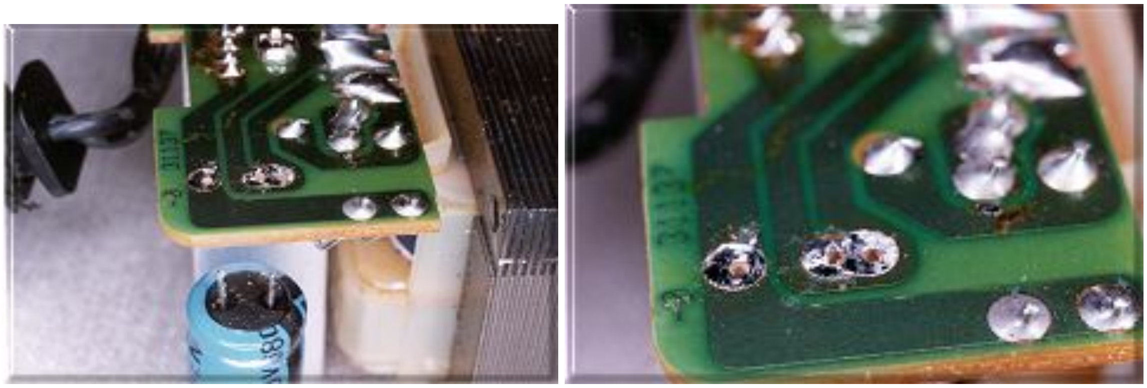

(Left) The faulty electrolytic capacitor dropped out of the board after desoldering. Sometimes, it may need persuading with pliers.... but don't overdo this or you risk damaging the copper tracks on the p.c.b.

(Right) Close-up photograph of both joints, now desoldered and ready for the replacement component to be fitted.

The Black Museum of Bad Soldering

These are all genuine examples which have not been retouched or reworked in any way.

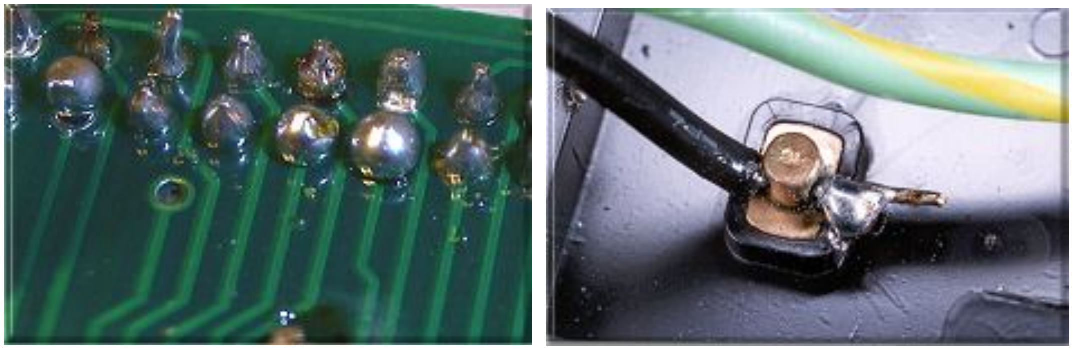

(Left) A tenfold excess of solder on this hand-soldered printed circuit board, and (extreme left) an incomplete solder joint with poor coverage. There is no need to add more solder "for luck".

(Right) An example of a dry (or gray) solder joint found inside a commercial PSU for a computer peripheral. The wire had been fed through the hole in the brass terminal, and merely tacked on with a blob of solder. This is a fire hazard (risk of arcing and overheating).

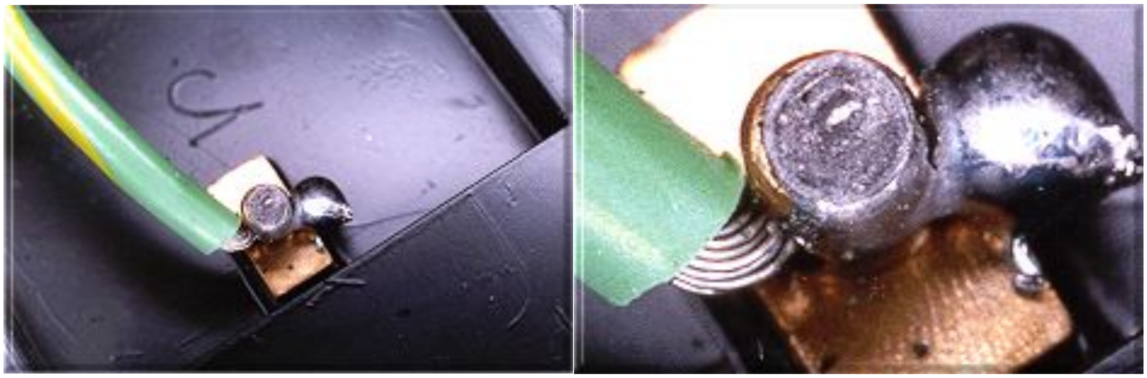

(Left) Hmmmm... this joint looks somewhat suspect as well... it's the earth (ground) wire in the same PSU

(Right) A close-up reveals the terrible standard of soldering (and quality control), with a fracture visible on this ground/ earth joint.

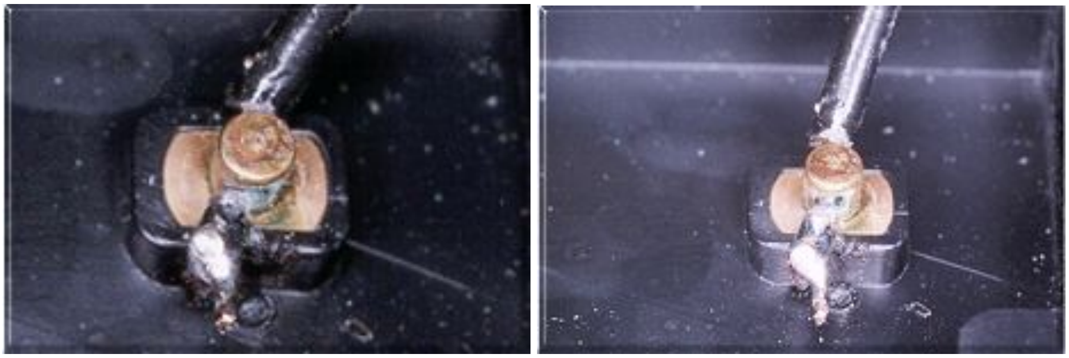

(Left) How **not** to make a mains voltage soldered joint. This solder joint went "dry" and starting arcing, nearly destroying the attached equipment. It is also a fire hazard.

(Right) The same mains connection, the wire merely being 'tacked on' with a blob of solder.

First Aid

If you are unlucky enough to receive burns which require treatment, here's what to do :-

1. Immediately cool the affected area with cold running water for several minutes.
2. Remove any rings etc. before swelling starts.
3. Apply a sterile dressing to protect against infection.
4. Do not apply lotions, ointments etc., nor prick any blisters which form later.
5. Seek professional medical advice where necessary.

A Few Hints about soldering on a Ground Plane

Before starting

- Always remove any grease and water from your hands. This will reduce the extent to which you oxidise and contaminate the board.
- Avoid touching the copper surface with your hands, always hold a board by its edges.
- Ensure all power and signal inputs are turned off (the soldering iron will short anything you touch to ground!!!)

Soldering a component

- Consider the circuit pathway from input to output in order to minimise the distance. This is especially important between ICs.
- Use the cleaning rubbers or solutions to clean away the oxidised surface where you are going to mount the component. This will help improve the electrical characteristic and improve solder flow
- Where you intend to place solder, VERY LIGHTLY scratch a cross-hatch pattern on the copper plate with a sharp tool. This will increase the friction thereby preventing your soldering iron skating across the board surface
- Apply heat to the copper surface first then add the solder so it smoothly flows on to the copper.
- Use pliers to bend the wire/chip legs to form small feet as this will increase the stability and make them much easier to solder. All wire bends should be done with pliers!
- Tin the legs before trying to mount them on the board
- When using wire, devise a colour coding system and stick to it i.e. red +ve rail, black GND rail, yellow –ve rail, blue input etc. This will make debugging so much easier
- If you are soldering an IC, two good techniques are:
    - Bend its legs with pliers so they are all horizontal and the chip can sit flat on the GND plane surface. Then place a bend just after where the pin leg narrows and bend again slightly further down so the leg it touching the ground plane (see image below). This way the IC is firmly held in place.

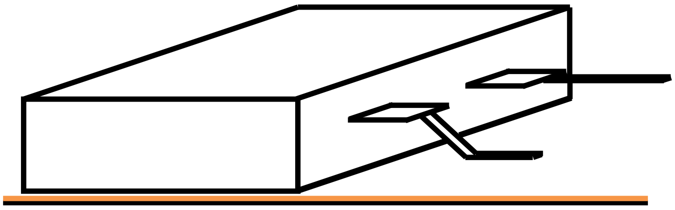

    - Bend legs touching the copper plane at the end to make small feet, then bend all other legs so that they are horizontal. This will result in the chip hovering just off the GND plane reducing the stability, but increasing the space to connect further components.
- Keep component leads as short as possible. This will reduce parasitics, improve mechanical robustness and reduce the chance of accidental shorts

These rules should be followed in conjunction with Dr Clarke's Soldering Guidelines found on the EE2 lab website.

Here is a good copper board layout. Notice the neat leads with none left loose above the circuit, and the way that the layout looks like a schematic. See the web copy (on the AMP handout page) of this picture for more detail.

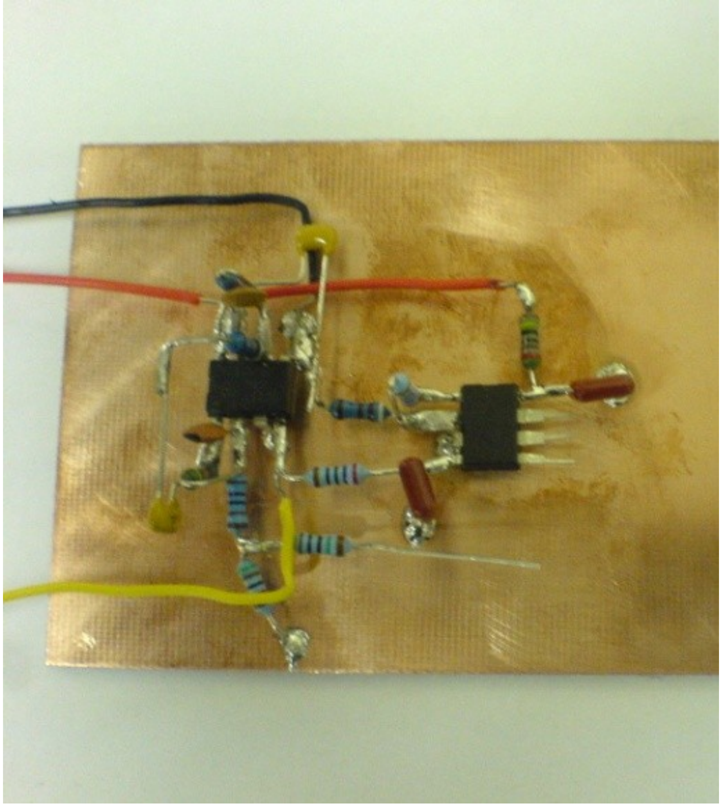

Soldering and Desoldering Instruction

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Soldering is more like gluing with molten metal than anything else. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice.

Remember that when soldering, the rosin in the solder releases fumes. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area. Hot solder is also dangerous. Be sure not to let is splash around because it will burn you almost instantly. Eye protection is also advised.

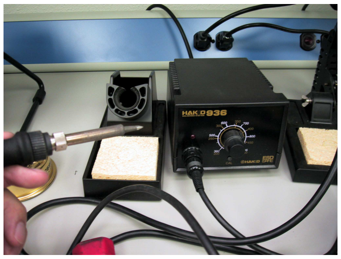

HAKKO 850
HAKKO 850

The Tools

Soldering Iron – Even the cheapest of them will do the job. They may not last as long, but they do get hot enough to melt solder and that is the goal. You only need to make sure that the one you buy has a suitable tip on it. The most typical tip is the one that is about the size of a 1/8” stereo mini-plug.

Solder – Get Rosin Core solder. Rosin will help the solder to flow onto the wires. Solder comes in different thickness. Buy what is appropriate for your job.

Desoldering Gun – This tool will make life a lot easier when you need to rework a previous solder job.
Solder Wick – Braided copper with rosin coating used to draw solder off of circuit boards.

SMD Rework station— Basically this is a high temp hair dryer with special nozzle attached. It heats up entire pins of the SMD device so you can remove them.

The Methods

Soldering – There is not art to soldering, it takes patience and practice to get it right. If you are doing it right, it will be easy and very fast. First, make sure that your soldering iron tip is clean. If it is dirtier than just a light gray color, you need to clean it. You can do this with sandpaper or a Scotch Brite pad. Next, turn your iron on and give it plenty of time to heat up. To test the heat, use a piece of solder touched to the tip. If it melts immediately, it is ready.

The most common way to mess up a solder job is to let the soldering iron stay on the parts to be soldered too long. What this does is oxidize the metal, making it too dirty to attract solder. If you are working with a circuit board, too much heat can damage the board, rendering it useless and in need of repair. You should only hold the soldering iron to the parts to be joined for no more than 10-12 seconds. Any longer and you will melt insulation on wire or damage a circuit board.

It has been said that soldering is a two-person job. You need two hands to hold the parts together, one hand to hold the iron, and another to feed the solder. The correct method for applying solder is to hold the joint perfectly still while heating with the soldering iron. After a few seconds, introduce the end of the solder at the point where the iron meets the parts

If the solder does not melt immediately and flow onto the joint, pull the solder away and try again after a couple seconds longer. If you exceed 10 seconds, pull the iron off and try again after it all cools down. You probably didn't have the iron touching enough of the parts to be soldered. Sometimes, the parts to be soldered are so big that they conduct the heat away very quickly and make it difficult to solder. In this case, it is OK to put some solder on each part individually and then put them together later by melting the solder on each one while they are touching.

Joining Two Wires - Strip off about 3/8" of insulation from the two wire ends to be joined. Place a ½" length of heat shrink tubing over one wire and push it back so that the heat from soldering won't shrink it prematurely. Fold each bare wire end back on itself so that the tip of the wire now comes back to the end of the insulation. Link both wires together using the bends like hooks. Now, twist the ends of the wires around themselves tightly. This type of connection is called a Western Union and it is the strongest method of joining two wires together. Now, heat the connection with the iron and allow solder to flow over the entire joint. Wait a few seconds for it to cool off and then slide the heat shrink tubing down over the connection and heat it with a match or heat gun.

Soldering a Wire to a Circuit Board – This is delicate work that you will do while soldering. Too much heat and your board will be damaged. If you are attaching a wire to an existing hole on the board, make sure that the hole is clean and free of any excess solder (see de-soldering for more detail). The best advice when soldering to a circuit board is to angle your soldering iron tip so that it makes good contact with the pad on the circuit board and the wire or part to be soldered at the same time. Heat the parts up and touch the solder at the point where the tip meets the part and the pad. If the solder does not flow immediately, pull everything off and wait for it to cool down and try again later.

Step 1

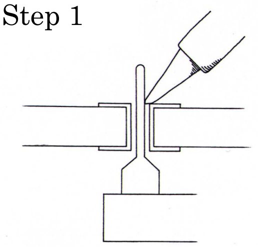

Apply heat to both pin and pad on the circuit board. Make sure tip of the soldering iron is touching both pin and pad.

Step 2

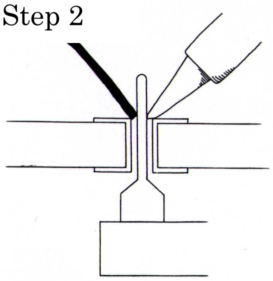

When the both pin and pad are heated enough, apply solder to the pin and pad. Solder should starts to flow.

Step 3

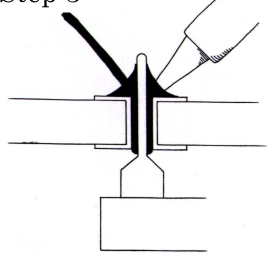

Apply enough but not too much solder. While applying the solder, do not move the tip of iron.

Step 4

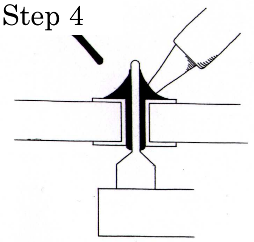

Remove solder away.

Step 5

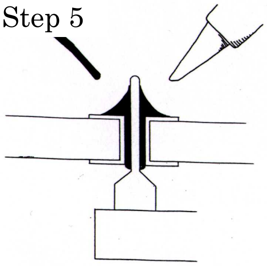

Remove tip of soldering iron also.

Step 6

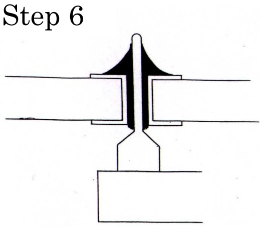

Done! Check the joint to make sure the connection.

Good soldering and Bad soldering Joints

Here are some example of Good soldering and Bad soldering joint.
For Japanese, we often say, “Good soldering joint should look like Mt. Fuji.”

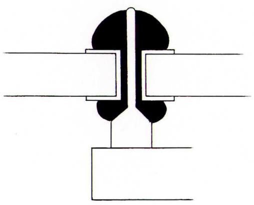

Too much soldering

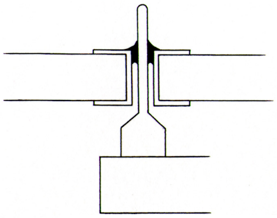

Too little soldering

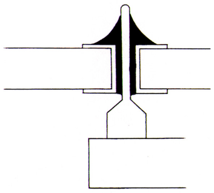

Good soldering

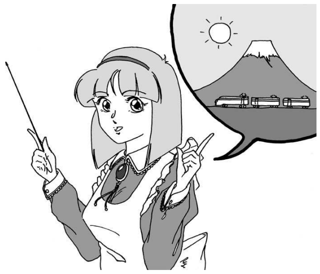

Cold Solder Joints
A cold joint is a joint in which the solder does not make good contact with the component lead or printed circuit board pad. Cold joints occur when the component lead or solder pad moves before the solder is completely cooled. Cold joints make a really bad electrical connection and can prevent your circuit from working.
Cold joints can be recognized by a characteristic grainy, dull gray color, and can be easily fixed. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can re-solder the joint, making sure to keep it still as it cools.

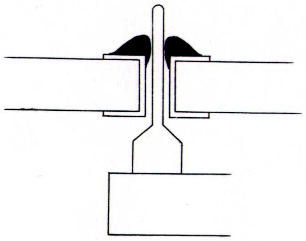

Cold soldering joints

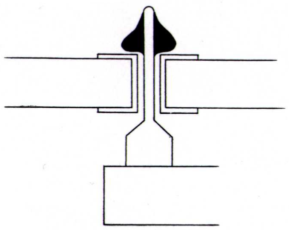

Tips and Tricks

Soldering is something that needs to be practiced. These tips should help you become successful so you can stop practicing and get down to some serious building.

1. Use heat sinks. Heat sinks are a must for the leads of sensitive components such as ICs and transistors. If you don't have a clip on heat sink, then a pair of pliers is a good substitute.

2. Keep the iron tip clean. A clean iron tip means better heat conduction and a better joint. Use a wet sponge to clean the tip between joints.

3. Double-check joints. It is a good idea to check all solder joints with an ohmmeter after they are cooled. If the joint measures any more than a few tenths of an ohm, then it may be a good idea to re-solder it.

4. Use the proper iron or temperature. Remember that bigger joints will take longer to heat up with an 30W iron than with a 150W iron. While 30W is good for printed circuit boards and the like, higher wattages are great when soldering to a heavy metal chassis.

Desoldering

Desoldering is extremely difficult compare to soldering. In the process of RomBoard installation, the parts and circuit board must be in the good shape to re-use them. The tool we use is Desoldering Gun. This device has vacuum pump built in with heater tip. Process of desoldering it self is very simple, but there are some tricks to do clean and safe desoldering job.

Step 1

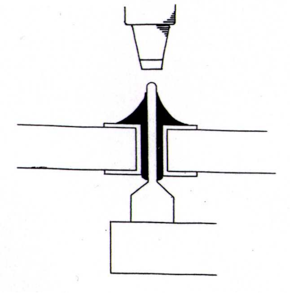

Put the desoldering gun's tip over to the soldering joint. Make sure nozzle is 90° angle from circuit board.

Step 2

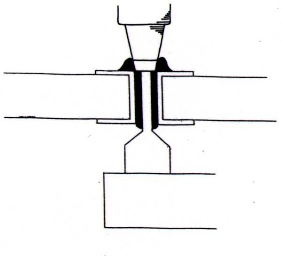

Apply the enough heat to melt entire solder. Multi-layer board or ground pin may take while to transfer heat.

Step 3

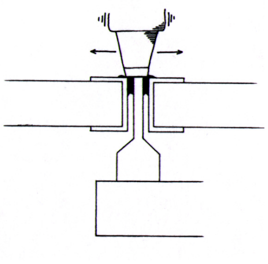

While applying vacuum, move the tip around the pin to get all the solder of around the pin out. At this time, you can also feel pin moves freely, so you know pin is free from circuit board.

Step 4

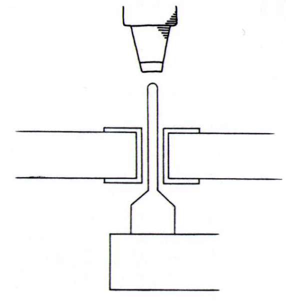

When you can hear air flow sound from hole, you should be done. Stop applying vacuum and remove the gun from the pin. Inspect hole and pin.

When you done with desoldering, the parts that you are trying to remove should move freely. If it doesn't, find which pin is still has solder left, and re-apply fresh solder to it and try desoldering process again. The multi-layer circuit board require more heat to get solder to melt. Make sure pin start to move freely by moving the tip of soldering gun before you apply vacuum to it.

SMD device soldering and removal

In the process of RomBoard installation, chance of handling SMD devices is becoming higher and higher due to ECU configuration change.
Soldering of SMD devices are not much different from regular through hole soldering.
Important thing is positioning of the device is very critical.
Start with applying little solder to two of the each corner of the pads on the circuit board.
Then, place the SMD device and re-heat a one of the solder you just applied to connect a SMD device pin to the pad. Check the position of the device, if position is right re-heat the other end of solder to secure the device completely. If you didn't get position right, re-heat the pin that you just soldered and while heating a pin, move the device to the right position.
Once you positioned the device in the right position, apply Rosin Flux to both pins and pads. This will help your soldering job by keeping solder separate from each pin.
Here is how the joint on SMD device pin should look like.

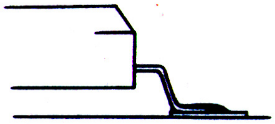

QFP and SOP chip

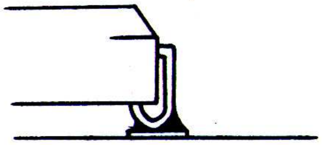

PLCC chip

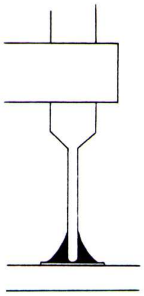

RomBoard adapter socket

Removing SMD device will be the probably the most difficult process. SMD rework station is used to do this. Make sure all the pins are completely heated otherwise you will be removing a pad from the circuit board along with device. Also, you should note, when you are applying heat to the device, some parts around the device is also heated and moves around when you touch. It is good idea to take a note of location of the devices near the target device.

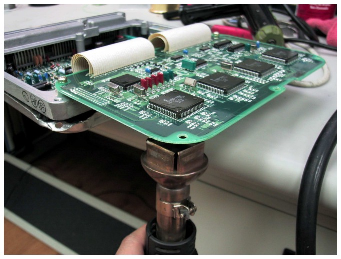

Step 1
Apply pre-heat from back side of the target device. Keep applying heat until you can't touch the chip. By this time, circuit board should have enough heat.

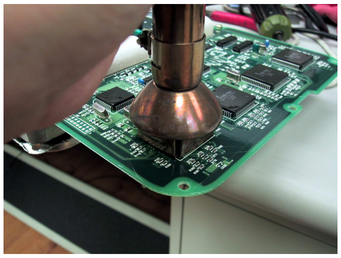

Step 2
Apply heat from top of the target device. Nozzle should cover entire chip. You can check if chip is free from solder by sliding the nozzle left and right. Make sure you don't move other parts. If the chip moves freely, it is ready to be removed.

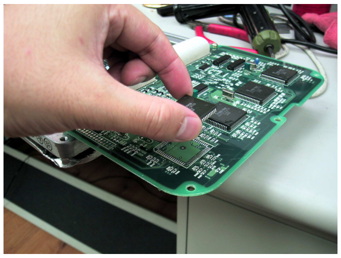

Step 3
Remove the chip straight up to avoid making solder bridge between pins. Chip is extremely hot! Unless you have skin of steel, avoid imitating the picture.

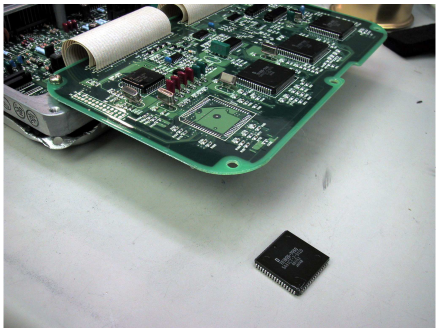

Step 4
After remove the chip, cool down both circuit board and chip.

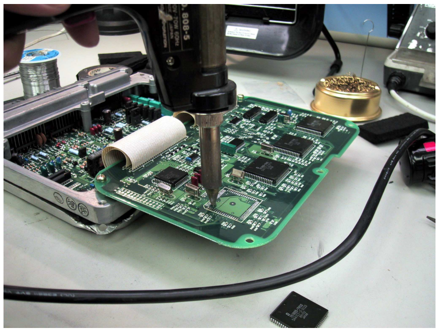

Step 5
Clean the pads on the circuit board with desoldering gun.
Do not apply too much heat to the pad. Pad might peel off from circuit board. Clean the chip by desoldering gun also.
Check for any bridge between pins. If the chip has too much protective coating on it, use solvent to clean it up.

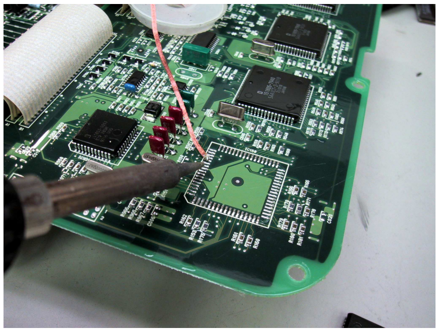

Step 6
If the pads on the circuit board is too small to use desolding gun, try cleaning by solder wick with rosin flux. This may take time but it is much better than peeling pads off.
Once again, do not apply too much heat.

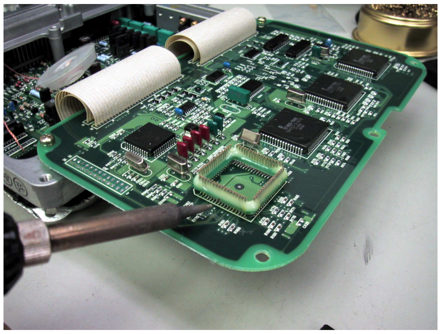

Step 7
Solder the RomBoard adapter where the chip was.
Refer SMD device soldering section for how to solder adapter.

The Final Word

Soldering, desoldering, and working with SMD device will take some effort to learn. You should be relaxed and concentrated when you work.
Keeping every thing clean (tip of iron, gun, work surface, etc) will help you also.
Once you get comfortable to work with soldering you may find your own tips and tricks.
Experience is all bout soldering.

EOF

Strain Gages and Instruments

Tech Tip TT-609

Strain Gage Soldering Techniques

Introduction

The most common method of making electrical connections in strain gage circuits is by means of soft solders, in wire form. Other methods, such as spot welding, brazing, compression bonding, paste solders, and conductive epoxies, are also available, but find only limited application. Solders have many advantages for strain gage use — they are low in cost, readily available in various alloy compositions to provide a range of melting temperatures, and are easily obtained in the form of either solid wire or wire with a core of flux. They are convenient to use, and offer an excellent combination of electrical and mechanical properties.

Although soldering is basically a simple procedure, it must be done with appropriate tools, supplies, and techniques to assure accurate strain measurement. This is particularly true when test requirements are severe in the sense of approaching the limits of the strain gage circuit capabilities; e.g., long-term stability, high-elongation measurements, fatigue endurance, etc. Use of improper materials or techniques can significantly degrade strain gage performance.

The purpose of this Tech Tip is to outline recommended procedures and materials for attaching leadwires to strain gage solder tabs or to bonded printed-circuit terminals. These reliable, experience-proven methods are based on the use of a professional quality soldering station, in conjunction with Micro-Measurements solders and installation accessories.

Soldering Station and Pencil

For precision soldering of strain gages, it is always necessary to use a temperature- or power-controlled soldering station that provides low voltage and adjustable temperature to the soldering iron tip. An unregulated soldering iron, connected directly to the power line, is not ordinarily suitable for strain gage use because the tip temperature is apt to be far too high. This tends to oxidize the tip, and to instantly vaporize the flux, making soldering much more difficult. In addition, the unnecessarily high temperature may damage the strain gage, the bonding adhesive, or even the test specimen. For these reasons, the soldering station should incorporate provision for adjusting the soldering temperature to suit varying installation conditions and requirements. The temperature must be adjusted, of course, to accommodate the melting points of the different solders commonly used for strain gage connections, but also to allow for environmental conditions such as drafts or outdoor soldering in cold weather. Moreover, the temperature controller should be carefully designed to ensure that it does not generate electrical noise that could adversely affect nearby measuring instruments when both are in use.

Design of the soldering pencil also requires special consideration. It should be light in weight, with a very flexible power cord, and with the gripping area thermally insulated from the heating element. These characteristics contribute to the comfort, ease, and precision of soldering, and minimize operator fatigue during long periods of use. The soldering tip itself should be of the flat, chisel, or screwdriver type. Pointed tips should not be used, because they tend to draw solder away from the work area, and thus make it more difficult to achieve a proper joint. In contrast, flat tips act to confine the solder, while offering greater surface area for better heat transfer and more effective soldering, generally.

Micro-Measurements soldering units incorporate all of the above features and a number of others, designed to help the user easily make consistent, reliable solder joints. These soldering units are widely used by professional strain gage installers everywhere, in both stress analysis laboratories and in transducer manufacture.

Solder Selection

The Micro-Measurements Division stocks a broad range of solder types to meet various installation and test requirements. While solders are sometimes selected to provide specific electrical or mechanical properties, the most common basis for selection is simply the melting-temperature range. Low-melting-point solders, for example, are generally used for strain gage installations on nonmetallic test parts to avoid damaging the gage, bonding adhesive, or test material due to overheating. In contrast, high-temperature solders are normally selected only when necessary to satisfy elevated-temperature testing requirements. These solders are somewhat more difficult to handle because the higher working temperature

Strain Gage Soldering Techniques

rapidly vaporizes the flux, and oxidizes the soldering tip, both of which tend to impede the soldering process. Specially designed soldering tips are recommended for high-temperature use.

For routine applications, where test conditions do not dictate the use of either a low- or high-temperature solder, an alloy with an intermediate melting temperature is the normal selection. The 63/37 tin-lead alloy (Type 361A-20R) is an excellent choice for general-purpose strain gage soldering. As an eutectic alloy, it has a sharply defined melting temperature — a characteristic that largely eliminates “cold” solder joints. The addition of a trace of antimony provides superior performance when the soldered connections will be exposed to very low (cryogenic) temperatures for long periods of time.

The general-purpose solders are supplied with a core of activated rosin flux. This makes soldering much more convenient, and is particularly useful in field applications where accessory liquid rosin flux (M-Flux AR) may not be available. Solid-wire solder, with externally applied acid flux (M-Flux SS), is recommended for making soldered connections to Micro-Measurements K- and D-alloy (modified Karma and isoelastic) strain gages. Rosin-core solders should not be used in conjunction with acid flux.

Silver solder (Type 1240-FPA) is available for applications where leadwire connections will be exposed to temperatures above about +550°F (*+290°C*). This solder, in paste form, is not suitable for attaching wires directly to strain gage solder tabs or to bondable terminals, but is intended for connecting instrument leads to preattached strain gage leads, as with WK-Series gages using a special resistance soldering unit. Techniques for making leadwire connections with silver solder are described in Micro-Measurements Tech Tip TT-602, *Silver Soldering Technique for Attachment of Leads to Strain Gages.*

Soldering Flux

The function of a soldering flux is to remove oxidation from the members being joined (solder tabs, terminals, leadwires), and to prevent further oxidation during soldering. For making leadwire splices, or soldering directly to constantan foil or copper terminals, the flux contained in a rosin-core solder is usually sufficient. With higher temperature solders, however, it may be necessary to supply additional flux. A liquid activated-rosin flux such as M-Flux AR is recommended for this purpose.

Acid fluxes should never be used on constantan strain gages or copper terminals, or for splicing copper leadwires; and paste fluxes, containing chlorides, should not be used under any circumstances for strain gage soldering. When tinning bare (without soldering options) solder tabs of Micro-Measurements K- and D-alloy strain gages, a liquid acid flux (M-Flux SS) is recommended. After the tinning operation, the residual flux must be completely neutralized within one to two minutes; and then the leadwire joint can be completed using the same solder and M-Flux AR rosin flux or a rosin-cored solder.

Preparation of the Soldering Tip

New soldering tips should always be tinned with solder prior to initial use. This is easily accomplished by wrapping one to two in (*25 to 50 mm*) of solder wire around the working portion of the tip while the soldering iron is cold, before applying power to the soldering station. If rosin-core solder is used, no external flux is required. With solid-wire solder, however, the wrapped tip should be dipped into liquid rosin flux (M-Flux AR) to provide sufficient flux for initial tinning. Set the control on the soldering station to the appropriate temperature range for the solder, and apply power to the unit. Allow the soldering pencil to heat until the solder wrapped around the tip melts completely. Remove excess melted solder from the tip with a dry gauze sponge. Never knock the heated soldering pencil against any object to remove excess solder, since this may result in personal injury or damage to the soldering pencil.

Note: Cross-alloying of solders can change the electrical, chemical, thermal and mechanical properties of the solder being used. To prevent cross-alloying, it is recommended that only one type of solder be used with each soldering tip. Of course, if one type of solder is incorporated in a gage with solder dots and another type is added, a mixture is produced. This mixture cannot be expected to have melting and strength properties any better than those of the lower temperature component.

Oxidation of the soldering tip seriously hinders the soldering operation. The tendency for oxidation can be minimized by ensuring that excess melted solder remains on the tip at all times when it is not actually in use. Negligent maintenance practices, or wiping the hot tip with materials that char on the surface, will produce a buildup of oxide that prevents proper soldering. If the tip does become oxidized, the following procedure is effective for cleaning and re-tinning:

1. Set the soldering station to the appropriate temperature range for the solder in use.
2. Place several drops of M-Flux SS on a glass plate. Re-tin the soldering surface by holding the heated tip in the SS flux while feeding solder onto the tip. A generous amount of solder is essential for proper tinning.

Strain Gage Soldering Techniques

3. Wipe the excess solder from the tinned tip with a dry gauze sponge. For severely oxidized tips, it may be necessary to repeat this operation several times to obtain a properly tinned surface. The soldering tip should never be filed or sanded, since this may remove the plating on the tip, accelerating the oxidation and leading to the early deterioration of the tip. After the cleaning operation, remove excess solder, re-tin and clean the tip several times, using rosin-core solder, or solid-wire solder withM-Flux AR.

Tinning Solder Tabs and Bondable Terminals

All strain gage solder tabs, terminals, and leadwires must be properly tinned before making soldered connections. This helps ensure active surface wetting and good heat transfer during the soldering operation. Tinning stranded leadwires to produce a formable solid conductor will also greatly simplify the leadwire attachment procedure.

Before tinning the solder tabs on open-face (unencapsulated) strain gages, the measuring grid should be protected with PDT drafting tape. The drafting tape is positioned to cover the entire grid and the upper portion of the solder tabs, as shown in Figure 1. This not only shields the grid from soldering flux and inadvertent solder splash, but also restricts the flow of solder on the tabs. The tinned area on the solder tabs should be only large enough to easily accommodate the leadwire size in use. The latter consideration is particularly important when making installations for dynamic applications or large-strain measurement.

The tinning procedure for strain gage tabs and terminals consists of first cleaning and reapplying a small amount of solder to the hot soldering iron tip. Next, apply a drop of M-Flux AR to the tab or terminal (this step can be omitted if a rosin-core solder is used). When soldering directly to bare Karma or isoelastic foil, use M-Flux SS on the gage tabs only. Hold the soldering pencil in a nearly horizontal position (<30°), with the flat surface of the tip parallel to the solder tab or terminal. Place the solder wire flat on the gage tab, and press firmly with the tinned hot soldering tip for about one to two seconds, while adding approximately 1/8 in (*3 mm*) of fresh solder at the edge of the tip. This procedure assures that there is sufficient solder and flux for effective tinning. Simultaneously lift both the soldering pencil and solder wire from the tab area.

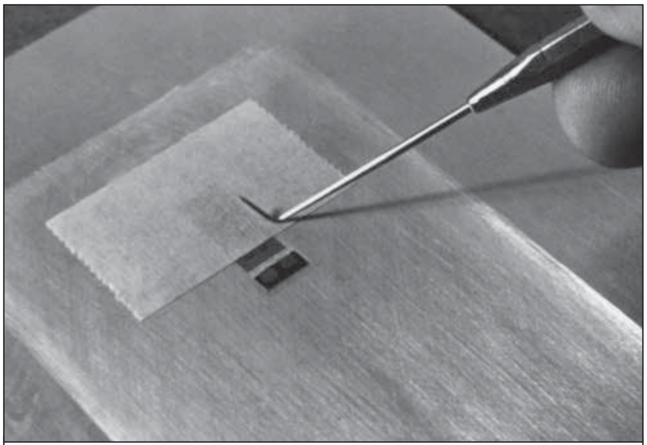

Figure 1 – Gage grid and upper portion of solder tabs masked with drafting tape.

NOTE: Lifting the soldering iron before lifting the solder may result in the end of the solder wire becoming attached to the tab; lifting them in the reverse order can leave a jagged (spike) solder deposit on the tab. When the operation is performed properly, it will produce a small, smoothly tinned area on the tab or terminal.

If M-Flux AR or a rosin-core solder is used in the tinning, it is not necessary to remove the residual soldering flux at this time. However, when M-Flux SS is employed to tin the bare solder tabs of K- or D-alloy gages, the acidic flux residue must be removed immediately following the tinning operation. To remove the residue, apply M-Prep Conditioner A liberally, and wash the area with a soft brush; then blot dry with a clean gauze sponge. Next, wash again with freely applied M-Prep Neutralizer 5A, and blot dry with a clean gauze sponge.

NOTE: Special procedures for tinning and wiring strain gages supplied with preattached solder dots are described in Micro-Measurements Tech Tip TT-606, *Soldering Techniques for Lead Attachment to Strain Gages with Solder Dots.*

Tinning and Attaching Leadwires

Of course, leadwire ends must be stripped of insulation before tinning, and this should be done with a thermal wire stripper to avoid the damage to the wire that often occurs when mechanical wire strippers are used. After the wires are stripped, the ends of stranded conductors should be twisted tightly together before tinning. The bare leadwire ends can then be tinned easily with the following procedure:

1. Remove excess solder from the soldering tip, using a dry gauze sponge. Then melt fresh solder on the hot tip to form a hemisphere of molten solder about twice the diameter of the wire to be tinned.
2. If rosin-core solder is used, slowly draw the bare wire through the molten solder while continuously adding fresh solder to the interface of the wire and soldering tip. With solid-wire solder, apply M-Flux AR to the

Strain Gage Soldering Techniques

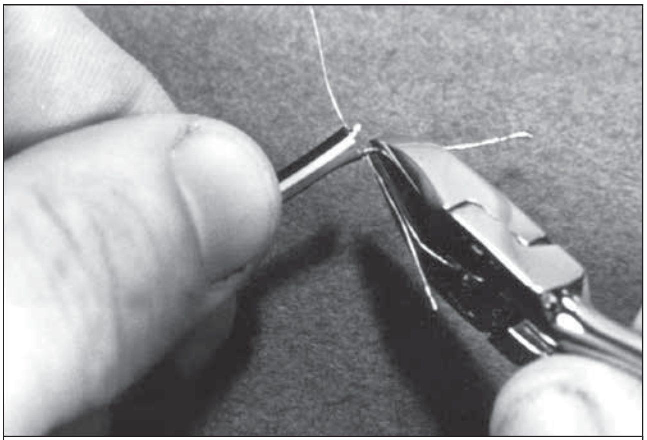

Figure 2 – Trimming leadwire ends before taping in place.

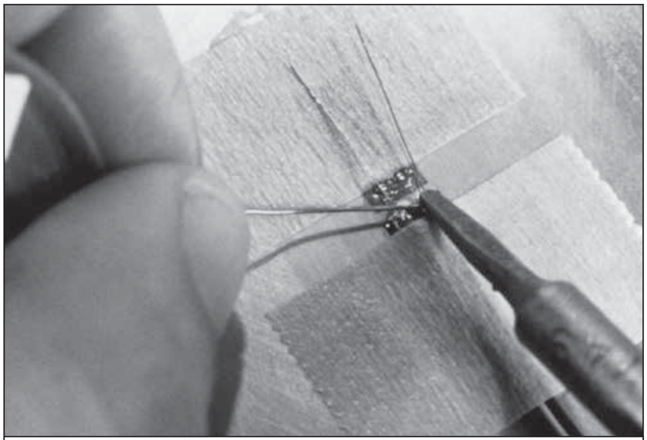

Figure 3 – Leadwire end taped to surface in preparation for soldering.

wire end before starting to tin, and proceed in the same manner. This will produce a smooth, shiny coating of solder over the bare wire.

For applications employing bondable terminal strips and stranded instrumentation wire, it may be convenient to use a single strand of the wire as a jumper between the terminal and the strain gage solder tab. In such cases, the single wire strand should be separated out before twisting and tinning the remaining strands (see Micro-Measurements Tech Tip TT-603, *The Proper Use of Bondable Terminals in Strain Gage Applications*).

Leadwires should be formed and routed to the strain gage or terminal strip, then firmly anchored to the test-part surface with drafting tape before making the soldered connection. Attempting to route the leadwires after completing the solder joint will often result in damage to the gage or terminals. Routing into the connection area should be along a minimum strain direction (such as the "Poisson" direction in a uniaxial stress field) particularly for high elongation or dynamic tests. The tinned leadwire end should be trimmed short enough so that it will not protrude through the connection area, and cannot inadvertently make electrical contact with the test-part surface or adjacent solder connections. Figure 2 illustrates this stage in the procedure. In the final preparatory step, bend the leadwire end slightly to form a spring-like loop, and tape the wire firmly in place over the connection area, using PDT drafting tape. The tape should be within about 1/8 in (*3 mm*) of the connection area, as shown in Figure 3.

Clean and re-tin the soldering iron tip with fresh solder. The temperature of the iron should be adjusted so that the solder is easily melted, without rapidly vaporizing the flux. If the iron temperature is either too low or too high, it may cause poor solder connections, or it may damage the strain gage, terminal, or bonding adhesive. Apply a small amount of M-Flux AR to the joint area and, holding the soldering pencil nearly horizontal, firmly press the flat surface of the tip on the junction for about one second; then lift the tip from the soldered joint. If needed, additional flux can be provided during the joining operation by feeding a little fresh solder into the joint from a spool of rosin-core solder. This procedure should result in a smooth, hemispherical solder joint, without any peaks or jagged areas. If the solder joints are not smooth and uniform in size, repeat the soldering procedure, using additional flux and/or solder as necessary.

Cleanup and Inspection of Soldered Joints

After completing the soldering operation, it is imperative that all traces of residual flux be completely removed with RSK Rosin Solvent. The same solvent is used to soften the mastic of the drafting tape, permitting its easy removal. Do not try to pull away the tape with tweezers or other tools, because this may result in damage to the soldered connections or the strain gage grid. Thoroughly clean the entire installation area with generously applied rosin solvent and a soft-bristled brush. Clean the solder connection area until no visible signs of residual flux remain, and blot the area dry with a clean gauze sponge. Any traces of residual flux can cause gage instability and drift, and will inhibit bonding of the installation's protective coating. Incompletely removed soldering flux is the most common cause of degraded performance in strain gage installations. Residual flux mixed with a protective coating application can completely destroy the coating objective.

Visually inspect the soldered joints for any gritty or jagged joint surfaces, and for traces of flux. Solder connections

Strain Gage Soldering Techniques

should be smooth, shiny, and uniform in appearance. Any soldered joints that look questionable should be re-soldered, and flux removed. Check the resistance-to-ground of the completed gage installation, using the Model 1300 Gage Installation Tester. Low or marginal resistance readings suggest a leakage path between the soldered connections and the test-part surface. This condition usually results from residual soldering flux, or from bare leadwire conductors partially shorting the gage tabs or terminals to the test part. Soldered joints should not be tested by pulling on the leadwire, or by probing at the joint area. These practices frequently cause lifting or tearing of the solder tab from the gage backing material.

Summary

The ability to make consistently good soldered joints is essential for precision strain gage measurements. The techniques described here are straightforward and easily mastered, but they are most effective when used with professional soldering equipment which is specially designed for making soldered connections in strain gage circuits. The soldering pencil should be lightweight, with a flat chisel or screwdriver tip, and it should be connected to the soldering station with a very flexible power cord. Requirements for the soldering station include low-voltage operation of the soldering pencil, and provision for temperature adjustment to suit the type of solder and the application conditions. The equipment should not generate electrical interference that could affect sensitive measuring instrumentation. Solder selection is based primarily on the expected operating temperature range of the strain gage installation; and all solder tabs, bondable terminals, and leadwire ends should be tinned before soldering the joints. Soldered joints should always be smooth and shiny, with no jagged or irregular edges, and all traces of residual flux must be thoroughly removed prior to the application of protective coating. Use of the recommended materials and techniques, with careful attention to detail, will result in consistently proper and reliable soldered connections.

PROJECT LEAD THE WAY
PLTW

Soldering & De-soldering

Digital Electronics

Soldering & De-soldering

This presentation will…

- Review the tools needed to solder and de-solder electronic components.
- Demonstrate how to *tin* a soldering iron tip.
- Demonstrate the soldering process.
- Show the characteristics of a good solder connection.
- Review classic soldering mistakes.
- Demonstrate the de-soldering process.

Soldering Tools

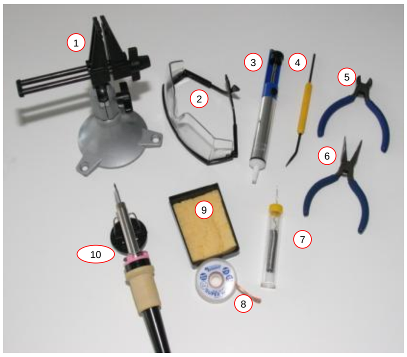

1) Vise
2) Safety glasses
3) Solder sucker
4) Solder tool
5) Diagonal cutters
6) Needle nose pliers
7) Solder
8) Solder wick
9) Damp sponge
10) Soldering iron

Soldering Iron

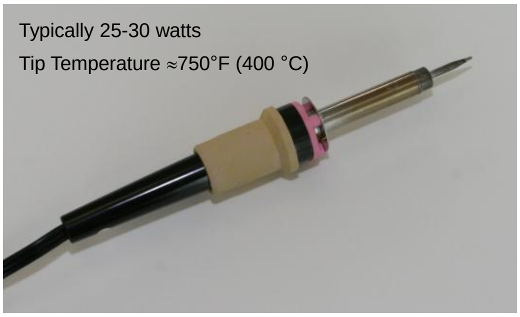

Typically 25-30 watts

Tip Temperature ≈750°F (400 °C)

Solder

- Solder is an alloy of tin and lead.
- The solder used for electronics is frequently called 60/40 solder because it is made of 63% tin and 37% lead.
- 60/40 solder melts at 361° F.
- Lead-free solder**:** As of July 1st, 2006, European laws mandated that new electronics be entirely lead-free. As of yet, no such laws exist in the United States.

Soldering Iron Care & Maintenance

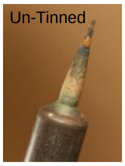

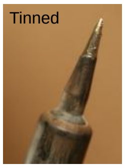

- A soldering iron must be coated with a thin coat of solder. This will allow for the transfer of heat to the work piece.
- This procedure is called tinning.
- The tip must be kept coated with a shiny layer of solder by occasional wiping and applying solder directly to the tip.

Tinning Process

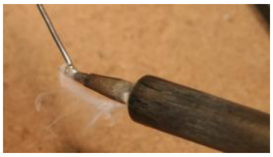
Apply Solder to Soldering Iron Tip

Roll Tip on Damp Sponge

Properly Tinned Soldering Iron Tip

Solder Process

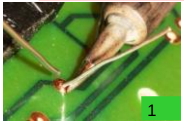

Heat both items at the same time by applying the soldering iron to the copper pad and the component lead.

Continue heating and apply a few millimeters of solder. Remove the iron and allow the solder joint to cool naturally.

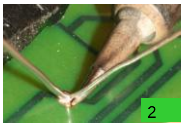

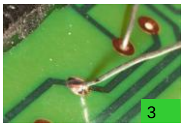

It only takes a second or two to make the perfect joint, which should appear shiny.

A Good Solder Joint

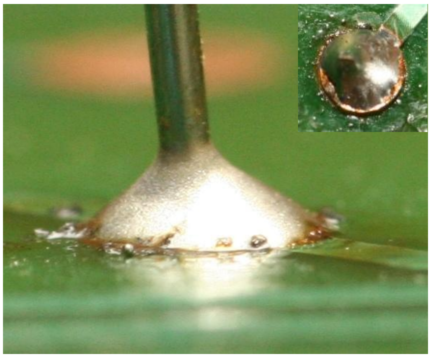

- Smooth
- Bright
- Shiny
- Clean
- Concave fillet

Bad Solder Connections

Too Much Solder

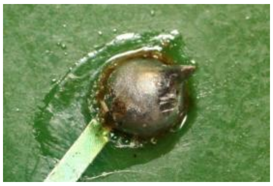

Too Little Solder

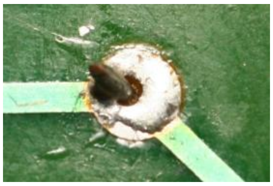

Bad Solder Connections

Cold Solder Joint

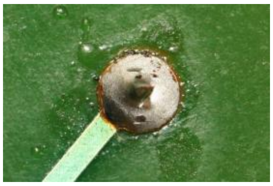

Not Soldered

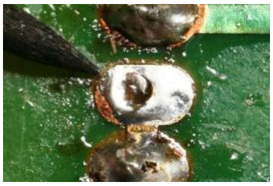

Bad Solder Connections

Solder Bridge

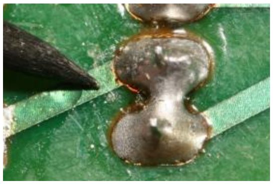

Lifted Trace/Pad

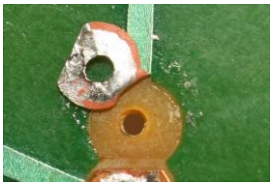

De-Solder Process: *Solder Sucker*

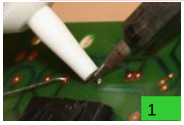

Apply heat to the connection to be de-soldered. When the solder melts, trigger the solder sucker.

Repeat de-soldering as needed until all solder is removed. Remove soldering iron & solder sucker from area.

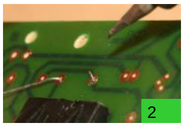

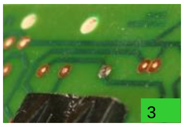

Remove component lead.

De-Solder Process: *Solder Wick*

Solder wick is finely braided copper that is used to *wick away* excess solder from a de-soldered connection.

Apply the solder wick and soldering iron to the de-soldered connection. The solder wick will draw the excess solder off of the PCB pad.

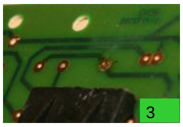

De-soldered PCB pad

Soldering Safety

- Wear safety glasses when soldering. This includes all individuals in the vicinity of someone who is soldering.
- Place soldering iron in an approved holder when not in use. The iron is hot and can cause burns.
- Place the soldering iron so that the cord does not get caught up in your arms or on others.
- Ensure access to proper ventilation.
- Verify that the type of solder is safe to use in your working environment.
- Secure the components to be soldered before beginning the soldering process.

Soldering Safety

- Provide plenty of space to work.
- Use a properly-sized point for the soldering job to be completed.
- Verify that the tip on the soldering iron has a sharp point and has not been damaged in any way.
- Check the power cord for burned or melted sections that show bare wires. Label those cords DO NOT USE and ask the instructor to repair or replace.
- Do not to touch molten solder - it is hot!
- Make sure that the solder strand is long enough to keep fingers away from the hot iron.

Soldering Safety

- Tie back long hair and remove or tuck loose clothing.
- Use heat sinks for heat-sensitive parts. Provide sufficient cooling time before removing parts.
- Do not flick solder off of the iron. Flicking can cause solder to spray and hit skin or eyes.
- Hold the scrap end when cutting excess leads so that the scrap lead is not thrown into the air.
- Cut leads evenly with wire cutters.
- Make sure that leads do not short across other traces or leads.
- Thoroughly wash your hands after handling solder.

Project Lead The Way, Inc.
Copyright 2009

How to Solder Like a Pro

Detailed instruction, helpful hints, safety tips, and a comprehensive troubleshooting guide

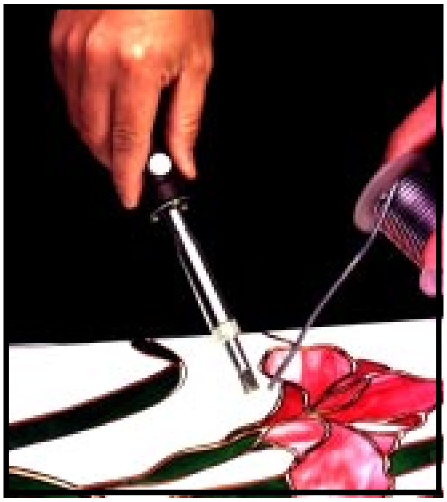

inland affordable products you can trust! inland

Table of Contents

Thank you for your interest in Inland's How to Solder Like a Pro manual. Whether you are a beginner or an experienced stained glass artist, you will find useful information about selecting and maintaining your soldering tools and improving your soldering skills.

Safety

The Importance of Safety 3
Fumes 3
Lead 3
Heat 3

Soldering Irons

Choosing a Soldering Iron 4
Types of Temperature Control 4
Irons with Temperature Controlled Tips 4
Ceramic Heating Element Irons 4
Temperature Controlling Devices 5
Tip Sizes, Styles and Uses 5

Iron and Tip Maintenance

The Importance of Proper Care 6
The Cord 6
The Heating Element 6
The Tips 6

Solder

Why Do We Use Solder? 7
The Advantages of Solder 7
Types of Solder 7
Characteristics of Solder Types 8
Solder Composition Reference Table 8

Flux

Why Do We Use Flux? 9
Types of Flux 9
Characteristics of Fluxes 9
Getting the Most From Your Flux 9

Soldering Copper Foil

Before You Start 10
Cleaning Before Soldering 10
Getting Started 10
Running The Solder Bead 11
Fixing Imperfections 11
Other Things To Consider 11
Ending A Bead 12
Cleanup After Soldering 13
Finishing 13

Soldering Metal Cames

Step by Step 14

Soldering Other Metals

Vase Caps and Spiders 15
Metal Outer Channels 15
Hanging Hardware 16

Patina

How to apply patina 16

Trouble Shooting

Tips and hints to help you fix problems 17-19

Glossary

Definition of terms 20-21

Safety

The Importance of Safety

One of the most important aspect of soldering any stained glass project is to take the proper measures to safeguard your health. Soldering involves working with potentially hazardous materials like lead, flux and harmful fumes. By observing specific safety precautions and using informed common sense, working in stained glass can be a safe and enjoyable past time. The following is intended to serve as a general rule of thumb. Since each person and situation is unique, you should use this information as a starting point to help you make informed safety decisions concerning your soldering habits.

Fumes

Solder only in a well ventilated area, and use an exhaust device that moves solder fumes away from your face. It's preferable to exhaust outside (check local/state building codes and restrictions on venting to outside air). If you don't exhaust to the outside, use a bench top fan or intake device with a replaceable smoke/fume absorber made for stained glass artists. It should draw solder fumes into the replaceable filter. The filter should be activated charcoal and designed to remove particles smaller than 1/2 micron from the air. There are also several OSHA approved respirators available for fumes. Consider wearing one in conjunction with a venting system, especially if you plan on soldering for several hours every day.

Lead

Most popular solders used in stained glass are lead based. When you are using them, follow these precautions:

- 👎 Never eat, drink, or smoke in any area where soldering takes place.
- 👍 Always thoroughly wash your hands after soldering.
- 👍 Make sure your soldering equipment and supplies are kept out of the reach of children.
- 👎 Do not discard lead or solder scraps into the trash. Find a means or place to recycle them.
- 👎 Never use lead based solders or cames on items that will come in contact with food or children, or will be frequently handled. If you are making kaleidoscopes, jewelry or napkin holders, use lead free solder.
- 👍 Consider having your blood lead level checked by your physician on a regular basis to help you monitor your handling practices.
- 👍 If you are pregnant, or considering it, you should check with your doctor before using lead or solder.

Heat

Soldering tools operate at high temperatures, so these safeguards are important:

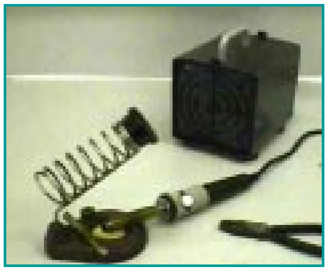

- 👍 Wear safety glasses! Solder and flux can "pop" and "spit."
- 👍 Solder on a fire resistant surface. Homosote, or dry wall are good .
- 👎 Never leave your iron plugged in and unattended.
- 👎 Do not overload a wall outlet with too many electric appliances.
- 👍 Never set your hot iron down on anything other than an iron stand.
- 👍 Replace the cord of your iron if it becomes worn or gets burnt.
- 👍 To prevent burning your fingers, use needle nose pliers or heat resistant gloves to hold small pieces.
- 👎 Never cut off a grounding prong on an iron plug to make it fit an ungrounded receptacle.

Soldering Irons

Choosing a Soldering Iron

There is a lot to consider when you choose an iron for stained glass. The handle should be heat resistant and comfortable for you to hold. You may be holding it for several hours at a time, so consider the weight and balance of the iron. Carefully check the wattage of the iron that you are considering. It must be able to continuously generate enough heat to melt all of the types of solder you plan to work with and be compatible with the type of construction you use (i.e.-copper, foil, lead came, rebar, etc.). You will need an iron that is at least 60 watts, with chisel type tips, and a way to control the temperature.

Types of Temperature Controls

A broad range of soldering irons are available. Most irons, which we will call "conventional irons," are made with wound wire heaters in a barrel with mica insulation. They are often inexpensive, but they do not offer the ability to control operating temperature. A separate temperature controlling device must be used with these irons to achieve the best results. There are three types of soldering irons that are easier to use for stained glass projects, because they allow you to control temperatures more effectively:

Irons with temperature controlled tips.

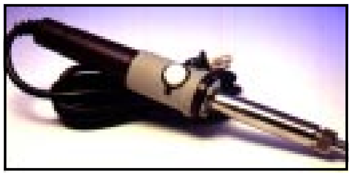

Irons used with temperature controlling devices.

Irons with ceramic heaters.

Irons with Temperature Controlled Tips

These irons are supplied with an internal regulator in the tip that does not allow the iron to exceed a predetermined temperature. An example would be a 600°F tip. The iron heats to that temperature then "shuts off." When heat is required, the iron "turns on" again. Tips are available in predetermined temperatures up to 800°F. These irons are easy for beginners to use, because the temperature is automatically maintained for you, however, as your skills increase, you may prefer to control the amount of heat yourself for different soldering situations. The limited amount of control may become frustrating as your soldering skills increase. With this type of iron you will likely encounter "cold spots" where the iron will not melt your solder. These cold spots occur when heat is being drawn out of the tip faster than it is being replenished. You will need an iron with a minimum of 100 watts.

Ceramic Heating Element Irons

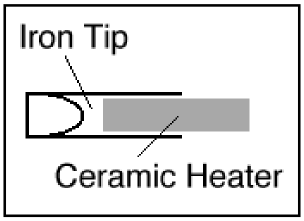

These irons are relatively new to the stained glass trade. They are made with highly efficient ceramic heating elements. Like a ceramic room heater, they produce a consistent temperature using less electrical wattage during operation. When initially heating and when reheating during "recovery" periods, they can draw a "burst" of power exceeding 100 watts and then efficiently reduce electrical consumption, often below 60 watts, during the soldering process. The result is efficiency and economy.

A remarkable feature of ceramic heater irons is they generally reach operating temperature in less than 60 seconds. Of the three types of noncoventional irons, the ceramic heater type best maintains consistent tip temperature over extended periods of time. Problems with slow heat recovery generally don't exist. A ceramic heater iron offers excellent flexibility and can be used with a temperature controller if you like.

Temperature Controlling Devices

A temperature controller is a device that operates similarly to a light dimmer switch. By dialing the control to a higher or lower setting, more or less electricity is fed to the iron. This increases or decreases the iron temperature, allowing the tip temperature to be controlled. You can establish the maximum temperature without changing tips.

A temperature control/iron combination offers you greater control and flexibility than a temperature controlled tip iron when working with different metals and solder mixes. Look for one that is at least 85 watts. Since the flow of electricity to the tip is consistent and never completely stops, encounters with cold spots are nearly eliminated. A temperature control/iron combination is suitable for all types of stained glass construction.

Temperature controllers are generally separate units that the iron plugs into. They are small, easy to use and relatively inexpensive. There are also irons that have controls built into the handle of the iron, which generally cost less than buying an iron and a separate control. Never plug an iron into a temperature control unless the manufacturer specifically states it is suitable for use with it.

Tip Sizes, Styles and Uses

Stained glass irons are generally sold with a "chisel" style tip. There are a variety of tip sizes and styles available for nearly every iron used in the stained glass industry. Different tips can expand the versatility of your soldering iron, so select an iron that offers more than one size replacement tip.

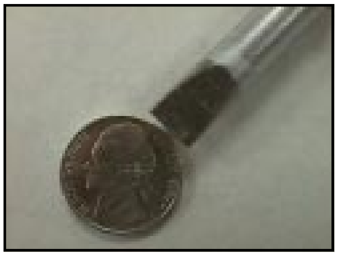

The standard tip size that comes with an iron is usually from 1/4" to 3/8" wide. This size works well for soldering either copper foil or lead came projects.

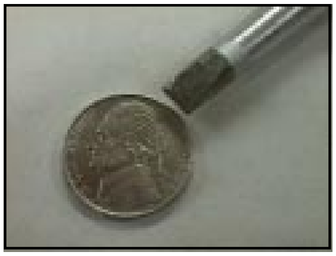

A smaller chisel tip, 3/16" wide, can be useful when soldering small pieces or when soldering in a tight area, such as a narrow inside part of a kaleidoscope. It is also very effective for decorative soldering.

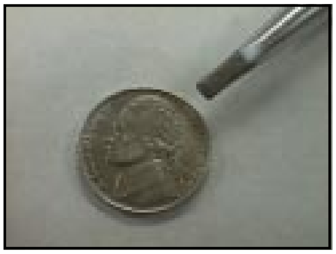

Very small tips, 1/8" or narrower, are usually used for decorative soldering. They allow you to create very fine details and designs with the solder.

Iron and Tip Maintenance

The Importance of Proper Care

Your soldering iron may very well be one of your biggest tool investments, so you will want to do everything possible to ensure that it gives you many years of service. Well maintained soldering irons and tips perform better and make the job of soldering much easier. Always place your soldering iron in a stable iron stand whether it is being used or not.

The Cord

- Make sure you plug the iron into the correct type of outlet.
- Try not to use an extension cord. If you must, use a heavy duty one.
- Don't drop or bang the iron. Ceramic heaters are especially easy to crack or break.
- Regularly check the cord for burns or cracks and have a professional electrician replace worn cords before using the iron.
- Make sure that the cord is not hanging in such a way that it can be pulled off of the table.

The Heating Element

- Do not allow the iron to idle at operating temperatures for extended periods. This could burn out the heater element. Unplug the iron or, if you are using a rheostat, turn it down to a low "idle" setting.
- Occasionally, remove the tip and **lightly** tap the barrel of wire wound heater irons to remove debris.
- If you will not be using your iron for an extended period of time, you may want to store it (after it has fully cooled) in a zipper type bag to protect it from corrosion and humidity.

The Tips

- When you are finished using your iron, remove the tip from the barrel. Removing the tip is essential to preventing "seizing" which can occur if it is left in the iron for extended periods. (If your tip seizes, you can easily damage the heating element trying to remove it. It is best to return you iron to the manufacturer for removal.)

- When reinserting tips, make sure they are properly seated in the barrel.
- Never dip your tip into flux in order to clean it. Instead, use a clean damp sponge to wipe all sides of the tip periodically as you solder. Doing this removes impurities the tip has accumulated from the solder and the environment. It will ensure that you are receiving the maximum heat at the tip surface.

- Properly cleaned tips are bright and shiny. If your tip becomes "blackened," you can usually remove the buildup with a wet sponge, a tinning block or by gently using a brass brush. A “tinning block” (sal-amoniac) is used by placing a small amount of flux on the block and rubbing the tip of your hot iron in it. Wipe the tip on a damp sponge to remove debris. You may need to repeat this several times if your tip is very dirty.)
- Never use sandpaper or any abrasive material to clean a tip.
- The best way to minimize your tip maintenance is to find a good quality solder. Use one that has a high tin content and high metal purity.

Solder

Why Do We Use Solder?

Solder is a combination of tin and lead used to create a strong bond between other metals. Since solder won’t stick to glass, we apply a copper foil tape (our metal) to the edges of the glass. This is refered to as the copper foil method of stained glass construction. Solder is melted over the copper foil, creating a structure that holds the pieces of glass together. The other option is to set the glass into channels of lead or metal, and solder the channels together. This is refered to as the lead came method.

The Advantages of Solder

Some of the advantages of soldering versus other bonding methods are:

- Solders are easy to use and relatively inexpensive.
- Low energy is required to solder.
- Properly soldered joints are highly reliable.
- Solder joints are easy to rework or repair.
- Experienced craftpersons can exercise a high degree of control over the soldering process.
- Solder joints age very well. They can last for years, decades and centuries.

Types of Solder

Stained glass solders are usually a mixture of tin and lead, designated by two numbers representing the percentages of each metal in that specific mix. The first number always refers to the percentage of tin, the second is the percentage of lead. The most commonly used solders in stained glass are 60/40, (60% tin/40% lead) 50/50 and 63/37. "Lead-free" solders have no number designation and are a mixture of tin and small amounts of other metals. Avoid solders containing antimony, a very toxic element. Instead look for lead-free solders containing silver, or copper. They are safer and easy to use. (Note they are, however, more difficult to patina.)

SILVERGLEEM
Quik-Set

In North America, you will find solder is generally sold in solid core wire form on a spool. The common spool size is one pound. In Europe, you will find solders primarily sold in a bar form. Never use acid-core or rosin-core solders for stained glass work. Look for solders that are sold as "free of impurities" in the component metals. Impurities cause a "scum" on your solder bead, degrade soldering iron tips, and interfere with the proper reaction of patina chemicals resulting in undesired finishes.

An important term for solders is the "working range" or "pasty range." This is the range of temperature between which solder transitions from liquid back to solid.

Characteristics of Solder Types

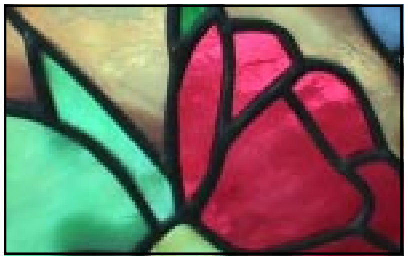

60/40 Solder: Composed of 60% tin and 40% lead, this solder melts at 374°F, but doesn't become completely solid until it cools to 361°F. This means it has a "pasty range" or "working range" of 13 degrees. This solder is your best choice for copper foil work. The liquid temperature and narrow "pasty range" make it easy to form and maintain consistent high, rounded, beaded seams. Because of its relatively low melting point, "60/40" solder is easy to rework to maintain a smooth finish solder bead.

50/50 Solder: This is composed of 50% tin and 50% lead. It is liquid at 421°F, solid at 361°F and has a pasty range of 60 degrees. This solder will produce a much "flatter" bead than 60/40. Because of its higher melting point, 50/50 solder is often used on the back (or inside) of a stained glass project to protect against "melt through" when soldering the front. Because it spreads and flattens out, 50/50 solder is often used when soldering lead came joints.

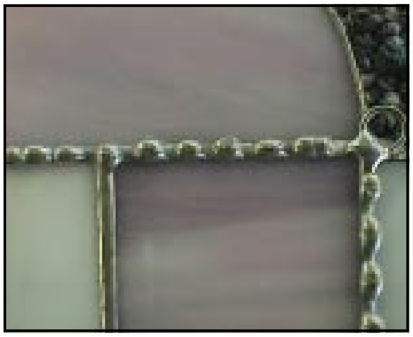

63/37 Solder: This solder is 63% tin and 37% lead. It becomes liquid at 361°F, and solid at 361°F, with a pasty or working range of 0 degrees. This solder is called a eutectic alloy which means at 361°F, you can go instantly from solid to liquid to solid just by applying or removing the heat source. You will often find "63/37" solder referred to as decorative or quick set solder. It is primarily used to create dimensional effects in the solder itself and can be "pulled" and manipulated to produce a variety of textures and designs. 63/37 solder also makes an excellent solder to bead up the outside rim of copper foiled pieces.

Lead-Free Solder: Depending on the specific mix of metals, lead free will produce differing liquid, solid, and pasty range temperatures. Check with the solder manufacturers for these specifics. Lead-free solders will perform similar to a 50/50 mix. Lead-free solders require more practice in order to obtain a smooth bead on copper foiled pieces and should not be used on lead came projects. Lead-free solder is the most expensive solder, but is the solder of choice when constructing pieces that will contact food, that will be handled frequently, or that will be used in a child's play area or room.

Solder Composition Reference Table

Alloy	Tin%	Lead%	Solid to	Liquid at	Pasty Range
50/50	50	50	361^0	421^0	60^0
60/40	60	40	361^0	374^0	13^0
63/37	63	37	361^0	361^0	0^0

Flux

Why Do We Use Flux?

Flux is a chemical compound that is used to promote the bonding of metals by removing the oxide residue simultaneously with the soldering process. Most metals left exposed to the air around us react with the air to form residue on the surface of the metal. The process is oxidization and the residues are oxides. Each mix of metals being joined has a specific flux that best promotes this bonding process. In stained glass, the metals being joined are primarily copper to tin/lead solder and lead, brass or zinc to tin/lead solder.

Types of Flux

Selecting the correct flux for your application is as critical as any other step of the soldering process. The proper flux will assure less soldering problems and a satisfactory solder bond. The best fluxes do three things:

- They remove all the residue that has formed on the surface of the metals you are going to solder.
- They prevent oxides from forming while you are soldering.
- Any post-soldering residue they leave is noncorrosive and easily cleaned off.

Fluxes are available in organic and inorganic forms. If you touch the flux and your skin seems to have a sensitivity to one type of flux, an inorganic type for example, try an organic variety. Often you will find that you are less sensitive to irritation by using the opposite type flux. Organic fluxes are generally some form of oleic (fatty) acids, while inorganic fluxes are most often zinc chloride based.

Characteristics of Fluxes

Liquid flux is the most widely used. It may or may not be water soluble. Water soluble fluxes clean up very easily, but are thinner and some have a tendency to evaporate quickly and require repeat applications. Some liquid fluxes are thinner than others and have a tendency to run or spread out from their point of application.

Gel Flux is generally water soluble and "adheres" well to the surface being soldered. Gels tend not to evaporate as readily as liquid flux, but if you apply too much, they produce more "spitting" and "popping" as you solder.

Paste Flux is very thick, stays where it's applied, and doesn't evaporate easily. It's effective when soldering 3-D forms such as lamps. Paste fluxes can be difficult to clean. Be very careful not to over apply a paste flux as it will also spit and pop, leaving pits on your solder seams.

Getting the Most From Your Flux

It is important to keep your flux clean and free of impurities. This will ensure the purest possible solder joint. One way to keep your flux clean is to avoid dipping in and out of the original container. Pour out the amount you will need for your current project into a separate, smaller container. Don't leave the flux bottle sitting around with the cover off. Never pour leftover flux back into the original container. Fluxes often will produce fumes as you solder. Make sure you are aware of and follow the precautions suggested in the health and safety section of this guide.

Soldering Basics for Copper Foil

The quality of the solder job can make or break your project, yet it is the step that takes the most practice to become good at. Soldering is more difficult than the other steps in stained glass because there are so many things that can effect it: the amount of heat, the amount of solder, the type of flux, the rate at which you move, gaps in the project, etc. The following step-by-step instructions will give you the basics of soldering. If you are just starting out, you will learn a lot. If you are experienced, you may just find the trick that makes it all come together. Great soldering takes a little bit of knowledge and a lot of practice.

Before You Start

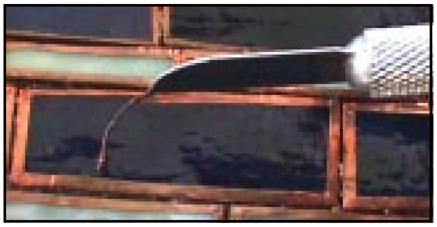

Preparation is the key to producing a good solder bead. Here are some suggestions to help you get of to a great start!

- Make sure that the foil is seated properly over the glass.
- Trim all overlaps in the foil using a sharp craft blade.
- Assemble all the tools you will need so they are at hand.
- Wear safety glasses.
- Have a proper ventilation set up.
- Make sure the foil is well burnished and sealed smoothly to the glass surface.
- Secure the glass pieces using pins, or squaring blocks made for this purpose before soldering.

Set aside enough time to solder your entire project all at once. If you don't, dried flux and solder residues on the unfinished parts will make further soldering difficult. Flux allowed to remain on a project can compromise the strength of the solder joints.

Cleaning Before Soldering

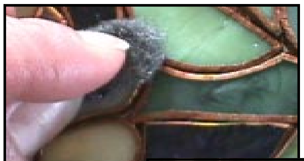

Sometimes the surface you are going to solder needs cleaning to remove visible grime, dirt, residue or oxidization. In this case, use 000 or 0000 steel wool to clean the surface of copper foil. For lead came or excessive oxidation carefully use a soft brass brush. If you won’t be able to solder your project immediately after it is copper foiled, store it in a plastic bag to help reduce oxidation. If you partially solder the project and have to stop before it is complete, be sure to clean off all of the flux and store the project in a plastic bag until you are ready to continue. You will probably still need to use steel wool on the joints when you are ready to solder, but the clean up should be minor.

Getting Started

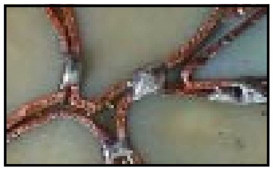

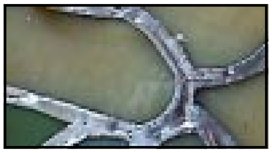

Begin the soldering process by fluxing all intersections of the glass project. Then "flat tack" solder these intersections together by using a very small amount of solder and the flat face of the iron tip. The solder should lay flat on the intersection and you shouldn't have any beads or bumps of solder on the piece. Hence the name "flat-tack" soldering.

Now, fill any gaps between the glass pieces by melting solder into them until the solder is level with the surface of the glass. If you have large gaps, you can ball up some copper foil, with the sticky sides together, and fill the gaps. This will help keep the solder from seeping out the other side.

"Tin" the entire perimeter making sure to work some solder into seams that intersect the perimeter. (Tinning is the process of coating a surface with a thin layer of solder. Apply just enough solder to turn the foil to the silver color of the solder.) This creates a "dam" that prevents solder from spilling over the edge when you solder the interior seams.

Next, flux an area you can comfortably finish soldering before the flux can evaporate. Start at the top of the piece and work down. This prevents you from dragging your sleeves and hands through the flux.

Running The Solder Bead

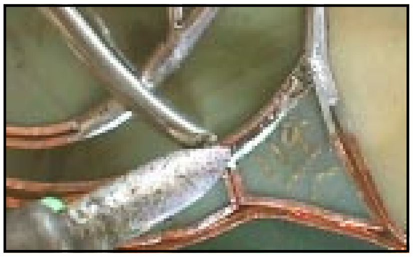

Hold your soldering iron in your writing hand. Position the iron comfortably so the face of the iron tip is perpendicular to the seam. If you hold your iron like a carving knife you will automatically put the tip in the best position for soldering. It will also keep you from "painting" with the iron. You will hold the solder in the opposite hand. Start soldering by placing the iron tip down on the foil seam with the wider face perpendicular to the seam and facing you. Touch the solder to the tip of the iron, just above the point where the iron tip touches the copper foil. Allow the solder to flow down the tip face onto the seam as you move the iron along the seam. Feed the solder at a continuous, even rate to create a consistent, rounded bead of solder over the seam. This takes practice!

If you apply too little solder or move too quickly, your seam will be flat. To correct this problem, allow the solder to cool, then resolder the joint applying more solder. On the other hand, be careful not to feed the solder too fast or move the iron too slowly, as this will cause the solder seam to bulge and flow out over the glass.

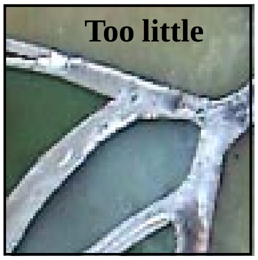

Too little

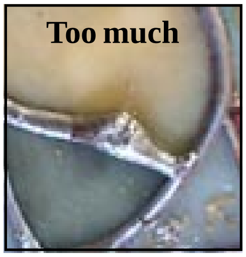

Too much

Fixing Imperfections

If you have too little solder, just add more, being sure to remelt the previous solder line as you apply another coat. If you applied too much solder the first time, it's easy to remove the excess with the following procedure. Clean your iron tip thoroughly on a sponge. With the iron tip positioned with the flat face down, move it across the seam, "pulling" the excess solder onto the glass. Clean your tip and repeat the process until all excess solder is removed. Another method for removing excess solder is to wait until all the seams have been soldered. Then place your project up on edge and run the tip through the seam allowing the excess to run off. Lay the piece flat, reflux, and rework the seam, adding additional solder as needed.

When you solder, work slowly enough to produce a good bead, but not so slowly that solder melts through seams to the back side of the project. Here's a hint: one way to prevent this is to put damp paper towel under your panel before you start soldering. Moving too slow also increases the likelihood of causing heat fractures on your glass.

Other Things To Consider

Always keep your tip clean by frequently wiping all sides of the tip on your damp sponge. Wait momentarily for the tip to reheat before continuing to solder. A clean tip maintains proper heat and removes impurities.

To achieve a smooth solder line, solder the longest continuous seam possible. Don't start or stop a solder seam at an intersection with another seam. As you meet an intersection, allow the solder to break or "y" out over the intersecting seam about 1/2" and then return to the original path you were soldering in one continuous motion. As you continue to solder seams, you will connect these branches, thus preventing pulling out at intersections of seams (which is a noticeable sign of a beginner). When you encounter a "v" type joint. Solder in a direction that moves up into the "v" point, instead of coming down onto the point. This keeps a nice clean "v" joint without excess solder obscuring it. Do this by soldering up one leg of the "v" and as you approach the apex, drop your iron tip down flatter on the glass, so that it crosses over both seams. Then continue soldering into and out of the point of the "v."

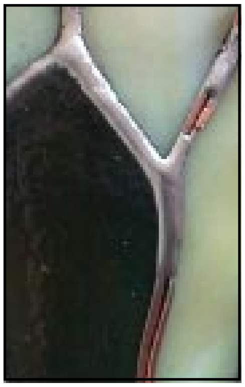

Avoid a "painting" motion when soldering. Visualize your iron as a magnet pulling along a metal strip. Maintain a small puddle of solder at the base of the tip as you solder. When two adjacent pieces of glass vary in thickness, ride the tip on the higher piece of glass as you solder.

Effects of using the iron like a paint brush

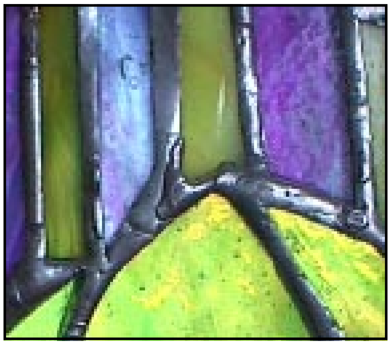

A big advantage of using copper foil is that not-so-perfect solder seams can be reworked. Allow the seam to cool, reflux it, then resolder. Be careful because reworking a seam too many times (more than 2-3) can cause the adhesive on the foil to loosen and bubble.

Solder one side of your project completely, then turn it over and repeat the process on the other side. Some of the flux from the first side you soldered will have seeped to the other side. You may encounter more "spitting," so use flux very sparingly. To achieve a smooth, beaded solder seam: practice! practice!

Ending A Bead

When you are ready to end your seam, pull off across the face of the glass making sure you are in a liquid section of the solder seam and do not lift straight up. This technique leaves a smooth exit point that is virtually unnoticeable. You may notice that this is the same photo used for removing extra solder. It’s not a mistake. The action is very much the same.

Practice!

Creating perfect solder seams takes practice. One good way is to make a practice board. Make sure to include all types of intersections and lines (curvy and straight). Foil the piece and flat solder it on the front and back. Practice each different soldering technique. When you are finished, you can hold the piece vertical and melt the solder off. Now you are ready to start again. Many people use the practice board before they start soldering each project, just to get the feel of soldering again.

Cleanup After Soldering

After soldering, you will need to remove all flux and solder residues from your solder lines and glass. The best cleaners to use are ones that neutralize the acid property of fluxes. To properly clean your project, you will need a good flux cleaner, very warm water and a soft scrub brush. Thoroughly brush the cleaner into every corner of your project, front and back. Follow this with a long rinse in clean, warm water and then towel the project dry. This step is very important because any flux that remains will continue to react with the solder, causing oxidation. If your project looks good after the initial cleaning, but the solder lines have a white or green growth on them after a few weeks, it generally means that all of the flux was not removed from the project. You can use 000 steel wool to remove the patina and finishing compound and thoroughly clean the project. Then reapply the patina and finishing compound.

A small stiff scrub brush is the best way to clean the entire project. After the initial cleaning, you may find some areas that need additional attention.

You may find small pieces of solder stuck to the glass. You can easily remove these using your fingernail, or 000 steel wool. Be careful not to scratch the glass.

An old toothbrush can be helpful for cleaning in tight intersections and right up against the solder line.

Occasionally you may find that after applying patina, certain types of glass acquire a “rainbow” effect. This is most likely to occur on hot colors, black, white and iridized glass. It is one of those things that is unpredictable and difficult or impossible to remove. You can try to remove it by using a paste made from baking soda, but it doesn’t always work. The only way to prevent this occasional problem is to be very careful not to get patina on your glass. Use a cotton swab to apply the patina only to the solder lines of the project. If you have this problem, it is a good idea to make note of it on any remaining glass, so you can be more careful the next time you use that particular glass.

Finishing

After cleaning your project you should use a finishing wax or compound to keep it looking it’s best. A good quality carnauba wax or pre-mixed wax and cleaner should be applied to all of the solder lines. A nice side effect is that the wax also enhances the color of the patina making it either rich black or shinny copper. You may find it helpful to reapply the wax after a few months. (If you would like to change the color of the solder, see the section about patina on page 16.)

Soldering Basics for Metal Cames

Lead came construction requires a different soldering technique than that used for copper foil. You can use the same iron and flux, but the preparation and actual soldering will differ. You will be soldering only the joints, instead of running an entire bead. With a little knowledge and some practice, you will be able to make perfect solder joints on lead or other metals.

Getting Started

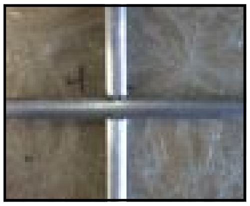

First, make sure that all joints are in line with each other and butt against each other with no gaps. Use a triangle and/or square to check the squareness of the panel and make sure that your lines run true across the panel in all directions. Make any adjustments that are needed. Thoroughly clean surface oxidation from intersections with a soft brass brush.

Soldering

The most desirable solder to use when constructing a lead project is 50/50. It flows out from the point of contact, leaving a nice joint. Do a "lead" test to check the heat of your iron. Allow the iron tip to heat for several minutes. Then, place the flat face of the tip on a scrap of lead. If the lead melts, the iron is too hot. Lower the temperature or change to a lower temperature tip. (This is where a rheostat iron is great.) Now test the tip with 50/50 solder to make sure it can melt it. If it does, your tip temperature is adjusted correctly. If not, raise tip temperature just until you melt the solder. Recheck the tip temperature and test it on the lead again. Now flux all the joints on your panel.

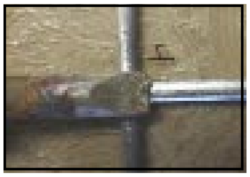

Begin by holding the solder on the joint. Move in to the joint across the glass, not from across the came. The wide face of the tip should be down toward the joint. Coming into the joint over the glass prevents leaving a solder trail on the lead came at one end of the solder joint. Techniques like these are the signature of an accomplished craftsperson.

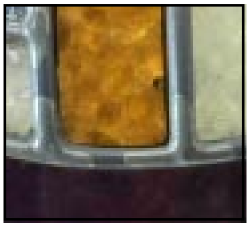

Melt only a small amount of solder and move the tip in a slight circular motion allowing it to flow out over the joint a distance equal to the width of the came being used. Pull the iron tip straight up off the center of the joint. The desired end result is that intersections of the lead cames are not visible through the solder and the solder on them is smooth, not beaded. You should be able to run a fingernail over the came and into the soldered joint without it catching or clicking. Solder the joints on the front of the project. Then turn the project over and repeat the same process on the back.

Finishing

Lead projects need to be finished by applying a cement. The cement will weatherproof the project and secure the glass in the channels. This step is necessary on any came project to ensure a strong, secure panel. Cement is easily applied to the panel by using a small scrub brush to work the cement under the channels both on the front and back of the project. After the cementing process is complete, whiting sprinkled over the entire surface and allowed to absorb excess moisture for a few minutes. Using a stiff brush, vigerously brush the lead channels, removing excess cement from the glass. The brushing process allows a natural patina to form on the lead, so it is not necessary to apply patina to the project. Allow the project to dry thoroughly, usually for a few days, before installing or hanging the project.

Soldering Basics for Other Metals

Vase Caps

Vase caps are generally made from spun brass. The cap must be very clean of all dirt and oils for the solder to adhere. Use 0000 steel wool to clean the areas that you will be soldering. The cap also needs to be quite hot in order to achieve a good bond between the solder and the brass. An easy way to get the cap hot is to heat it in a hot oven for about 15 minutes. Use pliers to carefully remove the cap from the oven. Center the vase cap on the lamp. Flux all areas you will be soldering. (Paste flux is easy to use for this purpose, because it doesn't drip.) Get a small amount of solder on your iron and hold it on the vase cap for a minute or so at the area you want to make a connection to the lamp. When the cap is hot enough, the solder will spread out instead of lumping up. Add more solder to make a connection to the lamp. (Some vase caps have a lacquer finish on them. This must be removed at the areas you need to solder. Use a craft knife to gently scrape the lacquer in these areas.

Spiders

Spiders are made from brass as well. You can either cut the legs to fit, or bend them. Like a vase cap, you need to get the brass quite hot in order to create a good bond with the solder. Unlike vase caps, you can get the smaller area of the spider hot enough using just your soldering iron. Simply hold your iron tip on the area that you need to solder for a minute or so. Apply flux and a little solder, and continue to heat the spider with your iron. When it is hot enough, the solder will smooth out around the spider leg and adhere to the lamp. If the solder is globed up, the spider leg isn't hot enough. Continue in this manner for all of the legs.

Metal Outer Channels

When attaching a zinc, copper or brass edge to a project, use this solder technique: Cut the zinc, copper or brass came vertical members to overlap the horizontal members in a "butt" type joint. A butt joint provides structural stability and distributes weight. Before you solder the butt joints, you can use masking tape to contain the size of the solder joint.

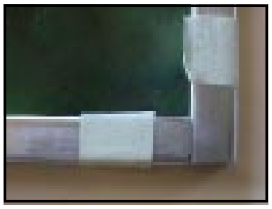

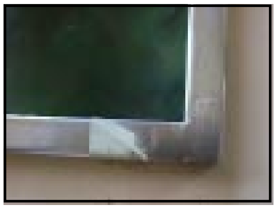

Clean and flux these joints and solder them to each other using the flat tip of the iron on the metal. The metal channel will take a little longer to get hot than copper foil. If the metal is not hot enough the solder will not smooth out. The metal frame should be attached to the panel by soldering it to any lead lines that it touches.

On leaded projects, solder from the outer metal edging onto the lead came, pulling the iron tip straight up over the point of intersection when completing the solder joint. This simple technique will level the solder smoothly over the joint.

For copper foil pieces, stop your solder bead about 1/4-1/2" from the outer edge to allow room for the metal edging to fit onto the perimeter of the project. If you forget to do this initially, flatten the solder bead so the outer edging fits onto the project. After fitting the outer edging material, flux and solder the corners as outlined above. Connect the copper foil beads that intersect with the outer edging metal by continuing the solder bead into the edging making sure that the solder adheres securely to the metal edge.

As you work more with your iron, different solders, lead came, and copper foiled projects, your skills will increase. You will quickly develop the techniques that will make you more confident in tackling increasingly complex projects.

Hanging Hardware

There are many ways to hang your project. You will want to make sure that the method you choose will support the weight of the project. You can purchase pre-made hardware from your stained glass store, or make your own using any type of solderable wire. If you are using wire look for 16 to 18 gauge.

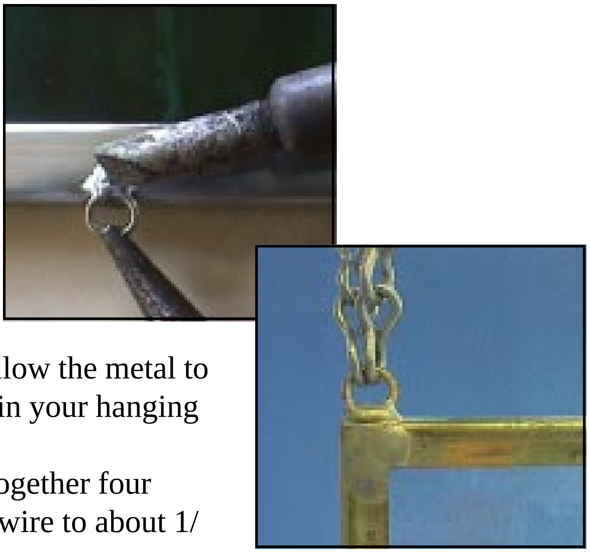

Rings can be added to the metal border of the project if it is not too heavy. A general rule of thumb is about one square foot. Hold the ring with pliers and solder rings to the metal border of the piece wherever they allow it to hang properly. Just like soldering the channel together, if you don't allow the metal to get hot enough, the solder will not adhere well. This will result in your hanging rings failing.

A stronger type of hardware is made using wire. Twist together four strands of 16 -18 gauge wire, leaving a loop at the end. Cut the wire to about 1/2". Insert the wire into the opening of the metal channel, bending the loop the direction you need it to be. Solder the hanger to the channel and fill the opening with solder.

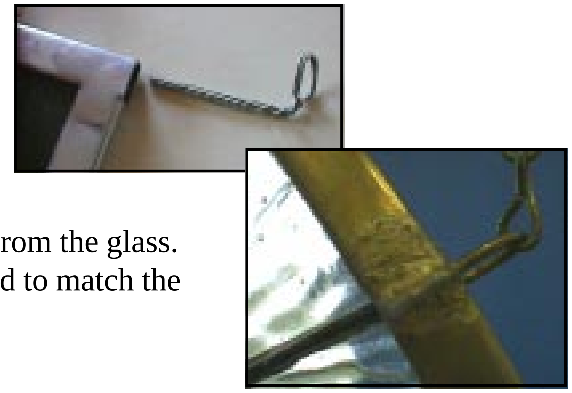

Another option is to make a loop from twisted wire and solder it to the front and back of the metal channel, with the loop extending above the project. Be sure to solder the hanger into an existing solder joint, not just to the channel. This will prevent the weight of the project from pulling the channel away from the glass.

By the way, the joint on the hanger at the right was painted to match the brass channel using a gold paint pen.

Patina

Solder and lead change color naturally over time, but you can change their color instantly by using patinas. There are several different formulas producing a variety of colors. Patinas are easy to use. Make sure that your project is thoroughly clean and free of solder, flux and cutting residues. Use a brush or soft rag to apply the patina to the solder until you reach the desired color. Wash the panel again to remove any patina from the glass.

👍When using antique brass patina, first scrub your solder seams with fine steel wool. Apply the patina sparingly with a clean brush, using only a few strokes. Over use will produce a coppery finish.

👎Never mix patinas and always use proper ventilation and skin protection.

👍To get a nice black finish on brass, add a dash of table salt to a little black patina in a glass or plastic jar. Mix well, apply and allow to dry. Clean and finish a usual.

Trouble Shooting

Problem: The solder won't adhere to the foil or came.

Solutions/Explanations:

1) Did you apply any flux?
2) Did you apply enough flux?
3) Was your iron too cool?
4) Is the foil or lead came too oxidized for flux to activate the surface?
    A) Buff copper foil with 000 or 0000 steel wool.
    B) Clean lead came with brass brush.

Problem: Solder is melting through to the backside of my project.

Solutions/Explanations:

1) Your iron is too hot. Depending on what type of iron you are using, either turn the temperature controller down or change to a lower temperature tip.
2) You may be soldering too long in each area. Move on and let the solder cool. Return to those spots later.
3) The gaps between your glass are too wide. Try to fill the gaps with solder and let it cool. Then go back and solder a bead on top of the filled gaps.
4) Try placing a wet paper towel under the project to provide cooling.

Problem: My solder is "stiff." I am unable to get a smooth bead.

Solutions/Explanations:

1) Your iron is too cool. Depending on what type of iron you are using, either turn the temperature controller up or change to a higher temperature tip.
2) You may not be using enough flux.
3) Is the tip of the iron getting hot? Check to see that the tip is seated correctly down in the barrel of the iron.
4) Are you "painting" with solder, instead of running a long bead with a steady, even, one directional movement? This is one of the most common mistakes that beginners make. In order for the solder to create a bead, it has to get molten. If you are using the iron as a “paint brush”, the solder is not getting hot enough to melt thoroughly.
5) You may be soldering with just a corner of the iron tip. Check for proper positioning of your iron.
6) Try a different solder with a lower melting point.

Problem: My iron seems to be hot, but the solder isn't melting as fast as normal.

Solutions/Explanations:

1) Check your tip. Is it loose?
2) Is your tip too dirty?
    A) Clean it on a wet sponge.
    B) Clean it on a tinning block.
3) Your tip is too corroded or defective. Replace it.
4) Your rheostat may be malfunctioning.
    A) Plug the iron into the outlet you are using without the temperature controller. If the iron works, the controller is malfunctioning. If the iron doesn’t work, the problem may be with the iron itself.
    B) With some types you can check the temperature controller by plugging a lamp into it. Turn on the lamp and see if the controller dims and brightens the lamp.

Problem: My solder seams are flat.

Solutions/Explanations:

1) You are not using enough solder. Reflux your project and add more solder.
2) You are soldering too fast. Move at a speed that lets the solder bead.
3) If the bead was there, but disappeared, you may be working too long in one area causing the solder to melt through to the other side. Allow the area to cool down before trying again. Another option is to place a damp paper towel under the seam you are working on. This will help keep the glass cool, allowing you a little more time to work.
4) Check the type of solder mix you are using. For copper foil work, 60/40 will help produce a higher, more rounded bead. Don't make the mistake of buying 40/60 solder. It has an even longer pasty range than 50/50, and is not recommended for stained glass work.

Problem: My solder seam is too wide.

Solutions/Explanations:

1) Do you have large gaps between the pieces you are soldering?
    A) You may have to recut some pieces.
2) You may be using too much solder causing the seam to bulge over onto the glass. "Pull" or "bleed" the excess solder from the seam.
3) Is the copper foil too wide? This is another common mistake for beginners. Remember that the width of the solder seam is determined by the width of the foil. An attractive solder seam is accomplished by using a foil that shows about 1/16" on both sides of the glass.

Problem: My solder seams are irregular in width.

Solutions/Explanations:

1) You may have poorly fitting pieces that create different size gaps in your project. Recut these pieces.
2) Your foil may be applied unevenly on each side of the glass in some places. You may try to correct this by trimming the foil with a craft knife. If that does not work, remove the foil and start again.

Problem: My glass fractured while I was soldering.

Solutions/Explanations:

1) Your iron is too hot. Depending on what type of iron you are using, either turn the rheostat down or change to a lower temperature tip.
2) You "worked" too long in one location with the iron.
3) There could have been a small chip or crack in the glass which was expanded when it was heated by the iron.

Problem: My solder bead is not bright and shiny. It appears dull and splotchy.

Solutions/Explanations:

1) The solder was applied too cold, so it never fully reached its "liquid state." Turn up the temperature if you are using a rheostat or change to a higher temperature tip if you are using a temperature controlled tip.
2) You may not be using enough flux to "wet" your foil or lead and create a proper solder condition.
3) If you're using 50/50 solder, try 60/40. Because solder crystallizes as it cools through the pasty range, 50/50 is more prone to having a textured look on its surface because of its considerably larger pasty range.
4) The solder you are using may be of inferior quality.

Problem: My solder is "spitting" as I work. What's wrong?

Solutions/Explanations:

1) You are likely using too much flux. It is literally boiling when you apply the solder. Wipe some of the flux off with a paper towel and try soldering again. You may find areas that now need a little more flux, because you removed too much.
2) If the spitting only occurs on the back side of your project (or the second side you are soldering), the problem is still too much flux, but the cure is different. When you use too much flux on the front side of the project, the excess to flows through the panel, so the spitting isn't as bad as it could be. When you use too much flux on the back side of the project, it can't flow through to the front, because the front is already sealed. The obvious cure for this problem is to use less flux the next time. But what about now, when you already have too much and can't remove it from between the glass? The only good solution is to apply your solder, and allow the flux to boil up and out of the seam. (Keep your face as far away from the project as you can.) Once it stops spitting, you can remelt the solder and make it look as good as new.

Problem: I have gooey, glue residue along my solder seams that won't wash away.

Solutions/Explanations:

1) Adhesive from foil may be working up and out onto your glass. This can be caused by poor foiling or burnishing of the copper foil, which allowed flux to seep underneath the foil. To correct this use a cuticle stick or soft toothbrush to clean adhesive away.
2) Your iron is too hot, causing the adhesive on the back of the foil to melt and seep out onto the glass. Depending on what type of iron you are using, either turn the rheostat down or change to a lower temperature tip.
3) If you reworked a seam too often, it may have loosened the foil. Allow the seam to cool to the touch and gently press the foil back down to the glass.

Problem: Attached hooks and rings are pulling away from my project.

Solution/Explanations:

1) Hooks and rings should never be attached to just a foiled edge. Attach hooks/rings to vertical seams in the piece or at a juncture between a vertical seam and the perimeter.
2) If there is no vertical seam or intersection, then attach the ring/hook to a horizontal seam in the piece.

Problem: My solder has a white chalky growth on it. What did I do wrong?

Solution/Explanations:

1) All of the flux was not washed off of the project after it was soldered.
2)The project was not sealed with a finishing compound or wax, or it has worn off. Use 000 steel wool to remove the patina and any remaining wax. Clean the project thoroughly paying special attention to the corners and edges. Reapply the patina and wax.
3) The project has been exposed to the outside elements. Most waxes are intended for indoor use only. Others can be used outside, but must be reapplied regularly. Projects that will be exposed to the elements are better constructed using the lead came method.

Glossary of Soldering Terms

Antimony

An element used in the production of some solders. It should be avoided for use in stained glass.

Ceramic Heater

A type of heating element comprised primarily of ceramic, noted as extremely fast heating and efficient.

Chisel Tip

A soldering iron tip shaped like a chisel tool. This is the most common shape used for soldering stained glass. Chisel tips are made in a variety of sizes, the most common being 1/8”, 1/4” and 3/8”.

Heat Sink

A device used to draw or absorb heat being generated by another source. For example, an object being soldered acts as a heat sink to the soldering iron.

Decorative Soldering

Any decorative effects created in solder. These effects are created with a soldering iron, usually with a very narrow tip. Special solder, like 63/37 or QuickSet, make it easier to create special effects because it has a “zero” pasty range.

Eutectic Point

An exact single temperature point at which an alloy goes from solidus to liquidus with no pasty range. For example, the eutectic point of lead and tin is 361^0 F. This point is obtained only by 63/37 tin/lead alloy.

Flux

A chemical agent used to remove compounds from the surface of metals during the soldering process. (See organic and inorganic flux.)

Idle Temperature

A very low temperature at which the iron is on, maintaining the capability of a more rapid heat up than if the iron was off or “cold”. This is usually between 200^0F and 300^0F.

Inorganic Flux

A flux comprised of one or more inorganic salt such as zinc chloride or ammonium chloride. Inorganic fluxes are more corrosive and conductive than organic fluxes,. They are effective on all common metals.

Leaded Solder

A material used to join metals comprised of tin and lead.

Liquidus

The temperature at which a pure metal becomes completely molten or liquid.

Mica

A mineral based material used as a construction component in wire wound type heaters of soldering irons.

Organic Flux

Organic fluxes are not as corrosive as inorganic fluxes. They are often used when the surface of the glass may be effected by the flux, such as painted glass.

Oxides

Debris on the surface of a metal which is the result of the reaction with chemicals in its environment. Oxides must be removed mechanically (with steel wool or a brass brush) and chemically for proper wetting to occur.

Pasty Range of Solder

The temperature range which is the difference between the solidus and the liquidus temperatures. This is sometimes referred to a the “working range”.

QuickSet™

Another name for 63/37 solder. See decorative solder.

Recovery Time

The time required for a soldering iron to reach soldering temperature after it has hit a “cold spot”.

Rheostat

The term applied to a soldering iron control which is used to vary the temperature of the heater in the iron.

Solder

An alloy of two or more metals with a liquidus temperature of less than 800^0 F.

Solder Bead

The term used to describe the look of solder when it has been properly applied to a copper foil seam.

Solidus

The temperature at which a pure metal or alloy goes from liquid to solid (or “freezes”).

Temperature Control

An electrical or electronic device into which a soldering iron is plugged, or which is within a soldering iron. The device is used to vary the temperature of the soldering iron heater.

Tinning

Applying a thin layer of solder to a metal surface. Most commonly used in reference to the copper foil technique.

Tinning Block

A block of sal-ammoniac on which a soldering tip is cleaned and resurfaced with a layer of solder. It is not actually used for soldering.

Tip Tinning

The process of renewing a soldering iron tip using a sal-ammoniac (tinning) block.

Wetting

The term used to describe the proper flow of liquid solder, promoted by flux, on the surface of another metal. “Wetting” is necessary to form a proper soldering joint.

Wire Wound Heaters

Soldering iron heaters constructed with heating wires wound in a coil around Mica.

Soldering & Desoldering

Soldering/Desoldering Tools & Accessories

Soldering Station with LCD Display, ESD Safe 400 Series

17400

- High-quality sensor and heat exchange system guarantee precise temperature control at the soldering tip
- Soldering iron tool heats rapidly from room temperature to 350C in 30 seconds
- Suitable for general purpose soldering as well as specialized lead-free soldering applications
- Sponge and coiled brass tip cleaner ensure a clean soldering tip free from debris and corrosive material
- Strong, deluxe iron holder provides safe secure place to store heated iron between uses

Product ID	Description
17400	Soldering Station with LCD Display, ESD Safe 400 Series

Soldering Station with LCD Display, ESD Safe 405 Series

17405

- LCD temperature control display with blue backlight
- Long lasting, durable high - quality ceramic heater
- Sponge and recessed area provides a convenient method for cleaning the tip
- Suitable for general purpose soldering as well as specialized lead-free soldering applications
- Safety guard iron holder

Product ID	Description
17405	Soldering Station with LCD Display, ESD Safe 405 Series

Soldering Kit - 6pc

17501

- 40 Watt Soldering Iron heats up to 900° F
- Ideal for use on electrical/electronic components, audio equipment, etc.
- Includes desoldering pump for removing liquid solder
- Provides basic tools needed for many soldering projects
- 6 Pc kit comes complete with replaceable fine tip

Product ID	Description
17501	Soldering Kit - 6pc

Soldering Kit - 7pc

17502

- 40 Watt Soldering Iron heats up to 900° F
- Ideal for use on electrical/electronic components, audio equipment, etc.
- Includes desoldering pump for removing liquid solder
- Provides basic tools needed for many soldering projects
- 6 Pc kit comes complete with replaceable fine tip

Product ID	Description
17502	Soldering Kit - 7pc

Soldering Iron 80W with Fine Tip

17510

- Ideal for general purpose soldering applications including metal board, tube or large components
- Round, soft grip, non-slip handle
- Durable cord with a 5-foot reach
- Fitted with long life replaceable fine tip

Product ID	Description
17521	Soldering Iron 40W with Fine Tip

Soldering Iron 40W with Fine Tip

17521

- Ideal for general purpose soldering
- Round, soft grip, non-slip handle
- Durable cord with a 5-foot reach
- Fitted with long life replaceable fine tip

Product ID	Description
17521	Soldering Iron 40W with Fine Tip

N-Series Soldering Iron Tips

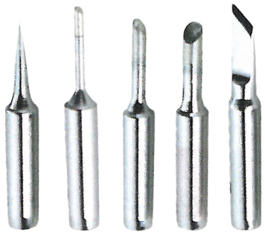

- Long life, lead-free soldering iron tips
- Replacement tips for soldering the different components of PCBs
- Five different sizes
- Includes two tips
- For use with 17400 & 17405 ESD Safe Soldering Stations with LCD Display

Product ID	Description
17400-N9-1	N-Series Soldering Iron Tips (N9-1)
17400-N9-2	N-Series Soldering Iron Tips (N9-2)
17400-N9-3	N-Series Soldering Iron Tips (N9-3)
17400-N9-4	N-Series Soldering Iron Tips (N9-4)
17400-N9-5	N-Series Soldering Iron Tips (N9-5)

C-Series Soldering Iron Tips

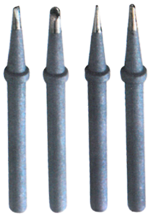

- Long life, lead-free soldering iron tips
- Replacement tips for soldering the different components of PCBs
- Four different sizes
- Includes two tips
- For use with 17521 Soldering Iron 40W

Product ID	Description
17521-C1-1	C-Series Soldering Iron Tips (C1-1)
17521-C1-2	C-Series Soldering Iron Tips (C1-2)
17521-C1-3	C-Series Soldering Iron Tips (C1-3)
17521-C1-4	C-Series Soldering Iron Tips (C1-4)

B-Series Soldering Iron Tips

- Long life, lead-free soldering iron tips
- Includes two tips
- For use with 17510 Soldering Iron 80W

Product ID	Description
17510-B5-1	B-Series Soldering Iron Tips (B5-1)

SMD Soldering

Your reference guide to soldering with surface mount devices

Edited by Lim Siong Boon, last dated 06-May-09.

email: mail@siongboon.com
website: http://www.siongboon.com

Short cut to your reference guides and charts

1. Introduction to SMD soldering
2. SMD soldering (prototyping board)
3. SMD soldering (prototyping PCB socket 2.54mm pitch)

1. Introduction to SMD soldering

SMD IC stands for surface mount device integrated circuit, or commonly known as SMT surface mount technology. Why go into SMD? Small in size, light in weight. This is the advantage of smd. Being small, engineer is able to design small electronics gadget that people can carry around. It is quite obvious that people prefer small mobile phone.

The ultimate advantage will be lower cost to consumer. Being small in size, the printed circuit board (pcb) can be small, meaning low cost. Transportation cost can be reduce because it is now lighter and more products can be packed into the same box. Space required is less, meaning cost saving on the warehouse storage. PCB board is easy to route without the through hole. Better signal integrity. Easily assembled by machine. There is too much advantages. Going SMT is certainly the way forward.

Technology evolution, from big to small.

A phone small enough to carry around.

Smaller mobile phone

Even smaller phone

Mobile phone as thin as biscuit.

I used to build circuits using dual in-line package (DIP or DIL) IC only. I hardly thought of using surface mount device/component (SMD) because I have great doubt that I can solder the fine legs of the IC chip. The pitch for the older DIP IC package are 2.54mm. Pitch is the distance between adjacent pins as illustrated in the picture. Most IC comes with a variety of packages. The IC that I know of, offer both the SMD and DIP packages. I

Through hole packages

	DIP - 8 pins

thought that I will never ever need to use smd packages until one day, I have no more choice. I managed to pick up new circuit designs, and discover more and more new IC chips. Most new IC chips design do not have DIP packages. You can hardly find one. I realised that in order to implement better circuits, I need to learn to use new IC chip. In order to use the new chip, I have to find ways to solder smd components. That's where I venture into the world of smd.

Nowadays people prefer to use smd, because they are small in size, which turns out to be cost efficient fabricating small PCB. With smaller PCB, space & weight is saved, resulting in lower cost for the transportation/distribution and storage.

Research and research, I got to know from other electronics guru that smd is in fact easier to work with than through hole. I try it out and from that day onwards, I am in love with smd components. Small to solder, but it save me the effort to cut the lead for through hole component. I managed to store my component using minimum box and space. There are many many advantage to work with smd.

Many people might think that you need special tools like a rework station or fine tip iron to solder the small smd components with small pitches. With the correct technique, you can use your soldering iron to do the soldering. In fact my 60W goot TQ-95 soldering iron has quite blunt tip. So thick that most people think that it is not possible to solder smd IC with fine pitches. Fine tip iron is easy to reach the fine lead, but I find that it is not as hot as the same iron with blunt tip. If you prefer fine tip, I will recommend hotter soldering iron or those which can allow you to adjust the temperature, they can be hot. I managed to solder smd package TQFP, SSOP QSOP with pitch as small as 0.5mm. Width of the lead of about 0.3mm.

Dealing with SMD components do not necessary requires you to fabricate a professional pcb board. You can also mount smd components on low cost prototyping board. To start off implementing smd components, you might like to try using the soic package. Quite common at this point in time, but may just phase out as what has happened to the dip packaging. SOIC has a pitch of 1.27mm, which is exactly half of DIP packages. This size is great because I can solder the IC to the same old 2.54mm pitch prototyping donut padded board. What I usually do is to cut the donut pad into half. Soldered onto each pad is two pins 1.27mm apart.

You can refer to the following article, for further illustration.

"SMD soldering (prototyping board)"

With so much advantage there is indeed a disavantage. SMD IC comes with many type of packages with different pitches. Unlike DIP IC, the pitch is typical 2.54mm. I can easy purchase a prototype board with 2.54mm, and almost all the DIP will fit to the board. SOIC package can still be mounted onto the 2.54mm prototype board. The rest of the SMD IC chip has quite a wide range of completely different packaging. This make them difficult to start with, without having to fabricate a PCB.

There are actually solution to this. Most IC company usually introduce their IC chip together with a prototype board for the IC.

Another solution will be to use a pre-fabricated prototyping board that allows mounting of various smd footprint. You can search for the various names such as

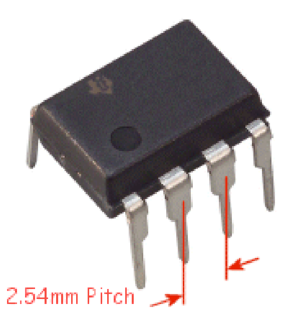	
	resistor
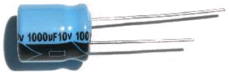	capacitor
	inductor
	diode

SMD packages

1.27mm Pitch	SOIC - 8 pins
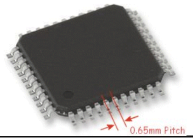	TQFP - 44 pins
	resistor
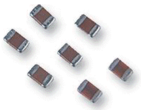	capacitor
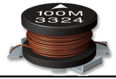	inductor
	diode

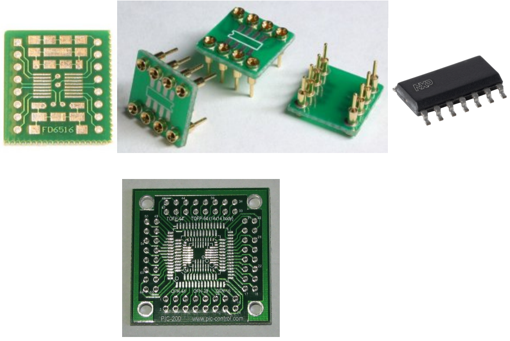

Various distributor for the prototyping adaptor

- prototype adaptor
- chip adaptor
- smd prototyping board
- smd to dip adaptor
- smd socket
- smd adaptor
- smd to 2.54mm converter
- SMD to DIP converter
- chip carrier

You can refer to the following article, for further illustration.

"SMD soldering (prototyping PCB socket 2.54mm pitch)"

In prototyping with DIL IC, pcb mount IC socket is usually used, so that the IC can be removed easily if damaged. Some sockets are designed for programming used where the chip can be inserted and remove easily for programming purpose. To remove the component, you need not have to do de-soldering. Like DIL package, smd also comes with their own IC socket. There are many variety of smd packaging, and getting the socket for your smd components is not going to come cheap & easy compare to DIL. Therefore I usually built prototype without any sockets for smd design. After numerous attempts designing and building of circuits, you will definitely attain a certain level of confident on your design without considering IC socket. Without the socket, cost and space are saved significantly.

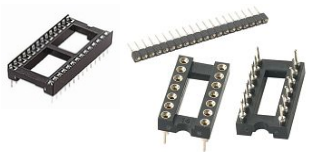

DIL IC sockets/ holder. Turn pins designs is of better quality.

Various distributor for SMD socket

SMD IC sockets:

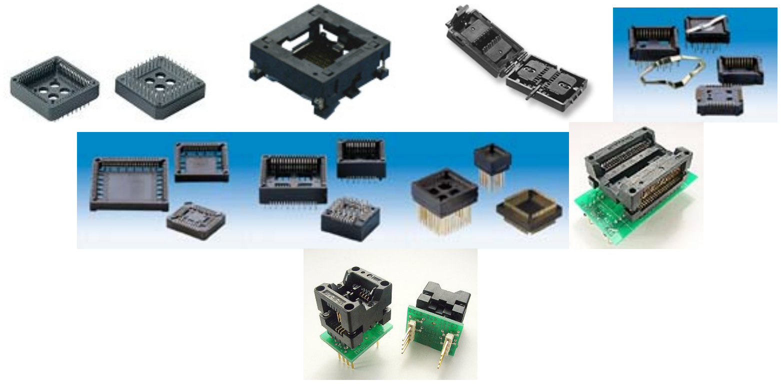

The headache will come when you really need to removed the smd IC. How do you remove the IC from the PCB? For de-soldering of smd component more than 3 leads, you can add in more solder to connect up the IC pins heat pins on all the sides of the IC and eventually pull out the component. It is easier to use the rework station or IR heater to de-solder smd components. The rework station uses hot sir to melt down the solder on the board. You will need to aim the hot air at the solder joint.

Rework Station

If you do not have the lurxy of getting the rework station, another possible method might be to use a special solder that has lower melting point. LowMelt® DeSolder Wire. A lower melting point means that the solder will takes a longer time to cool down to a solid. This allows you more time to remove the IC after heating up the solder. The solder is flooded onto the pins of the IC. The pins are heated up together using a soldering iron and the smd IC can then be removed easily with a tweezer.

Various type of smd packaging

* SOT, SC

* SOIC, ExposedPad™

* SSOP, ExposedPad™

* TQFP, ExposedPad™

* TSOP

* TSSOP, ExposedPad™

Low melt solder wire

The Science Of™ Zephyrtronics A Division of JTI, Inc.

* PLCC

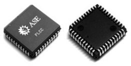

* LQFP, ExposedPad™

* LGA

* PBGA

* PSOP

* Dual Row MicroLeadFrame®

* LQFP PowerQuad®

* MicroLeadFrame® (MLF®)

* MQFP PowerQuad®

* PSvfBGA

* SuperBGA®

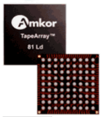

* TapeArray® BGA

* tsCSP

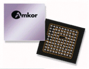

* Ultra CSP™

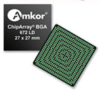

* CABGA, CTBGA, CVBGA

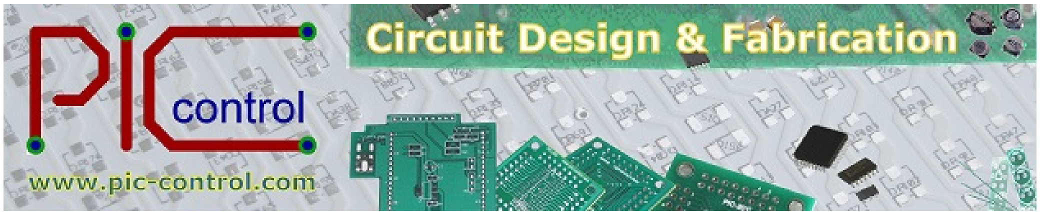
PIC control
Circuit Design & Fabrication
www.pic-control.com

Singapore Customized, custom made Electronics Circuits & Kits

2. SMD soldering (prototyping board)

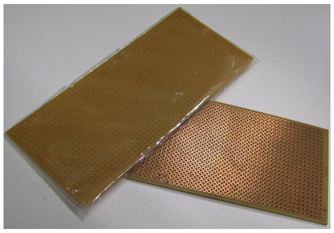

The prototype board above has pad with 2.54mm pitch and is designed for DIL and through hole components. The board is easily available from local store. Dealing with smd components on this 2.54mm pitch board is not difficult as well. I have been using this board to solder quite a number of surface mount components.

- SOIC
- DPAK
- SOT23-6, SOT-3
- TO-263
- PSOP

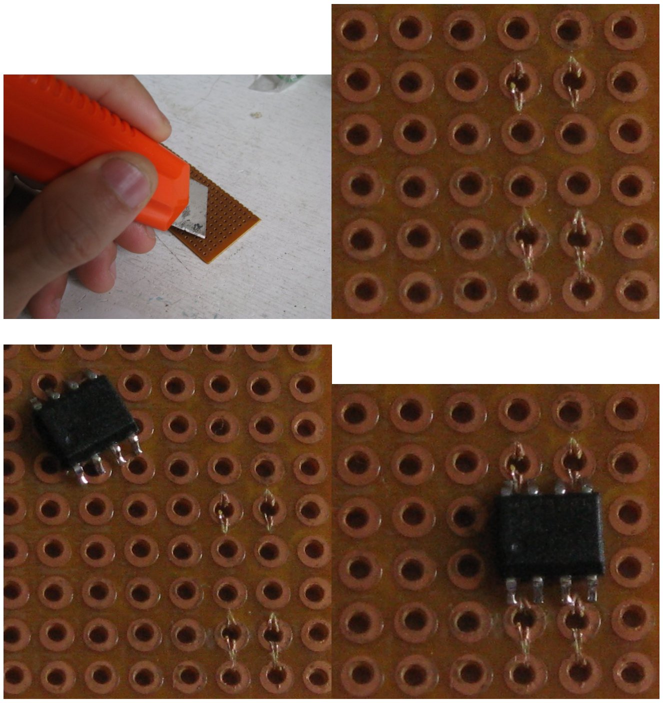

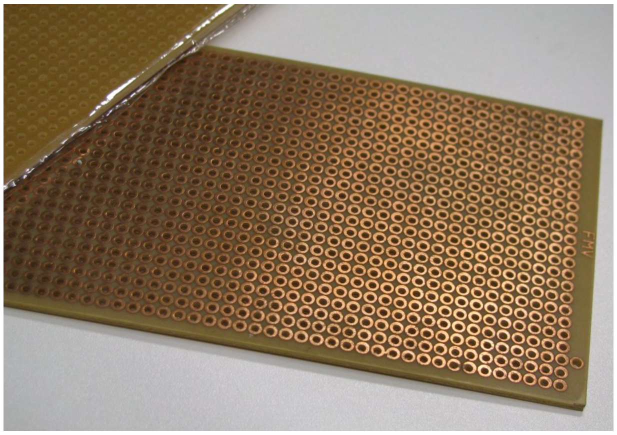

High pin count component with pitch less than 1.27mm will not be suitable to mount onto this board. For passive smd component like resistor, capacitor or inductor which only have 2 pin, mounting is simpler. Size is not an issue soldering onto the board.

Align the SMD IC to the position on the pad. Mark out the pad for the cutting to be done.

Use a penknife to cut the donut pad into half. Ensure that the half pad does not short circuit using your multi-meter. Sometimes the cut may not be deep enough or conductive reside between the gap resulting in a short circuit. Remember to ensure that each is proper cut before you start to solder.

Align the smd with your finger or tweezer and begin soldering one of the pin. Check if the IC is still properly align after soldering the first pin. If not, heat up the soldered pin and realign again. After the ICis align, solder all the rest of the pin. If the solder bridge across to the adjacent pin, use a soldering sucker to remove the excess solder.

The cut between the pad usually make the solder harder to bridge across the adjacent pin. Bridging is still possible, and it is quite easy to have it removed.

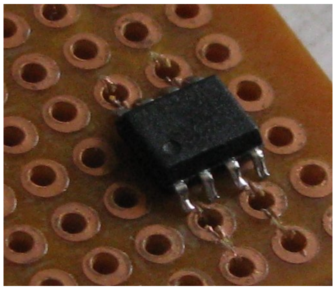

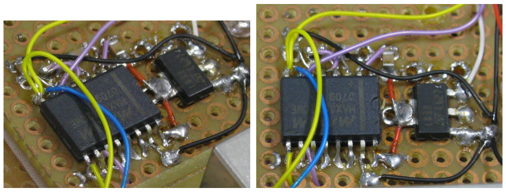

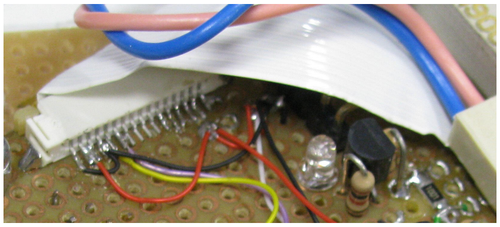

SMD ribbon cable connector being soldered onto the 2.54mm pitch prototype board.

Example of building the circuit using surface mount component on the prototyping board with 2.54mm pitch pads. As shown above, the pins are being wired by fine wire wrapping wire. The single core wire is very flexible and comes in a variety of color, making it easy to identify the type of signal being carried by the wire.

3. SMD soldering (prototyping PCB socket 2.54mm pitch)

What you need...

List of items:
- Surface IC chip (TQFP package)
- Surface mount IC adaptor or converter
- Soldering Iron
- Tweezer or Pincer
- Solder with flux core
- Soldering Flux
- Stick for coating flux
- Wiping paper
- Tinner

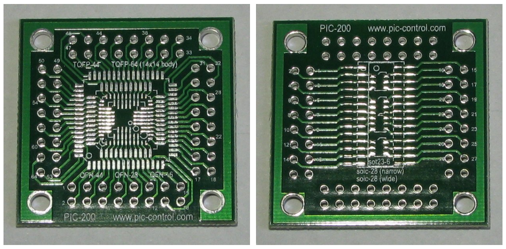

PIC-200 smd to 2.54mm pitch adaptor

from PIC control.

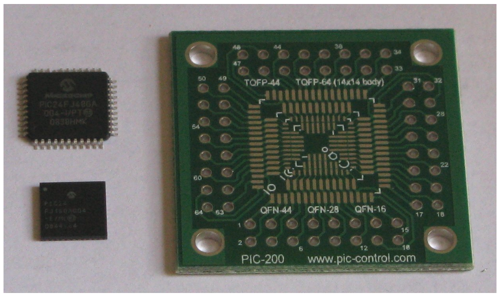

INTRODUCTION

In this setup, I am going to solder a surface mount microcontroller (TQFP 44 pins) onto a smd adaptor. The adaptor will then be used on the commonly available prototyping board, so that I can try out my new design for circuit control.

TQFP, QFN and SOIC are quite some common footprint for prototyping smd microcontroller from 18 to 64pins. This prototyping adaptor pic-200 (on the left) that I have is an ideal adaptor for my microcontroller. Converting the smd device into 2.54mm, I can easily mount the smd unit onto my 2.54mm pitch prototyping board. It is available locally.

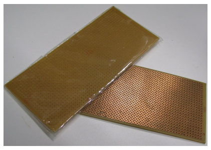

Prototyping board with 2.54mm pitch donut pads.

The pic-200 smd adaptor originate from PIC-CONTROL and is well documented. Commonly used smd footprint can be found on either the top or bottom side of the pcb adaptor. These are some of the smd packages that can be mounted onto pic-200 adaptor.

IC	**Front**	**Back**
TQFP-44		

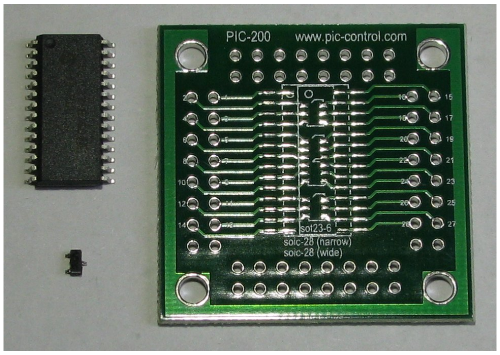

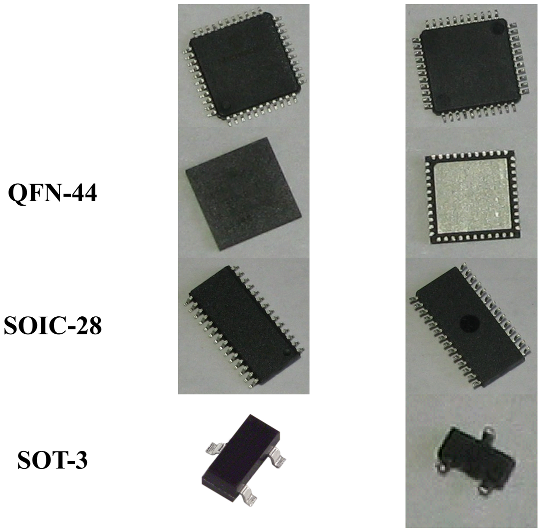

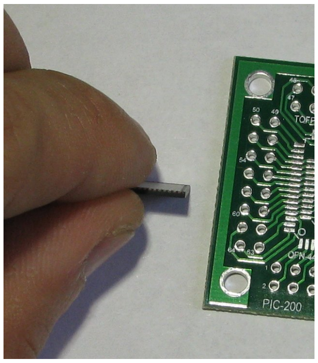

Holding onto the side view of the QFN package IC.

With a wide range of smd components that can be mounted, it is easier to stock up this pcb adaptor for future prototyping use.

The following article is about getting my smd microcontroller PIC24FJ64G004 soldered onto this board. After which I will be on my way to do prototyping for my new control circuit.

This smd adaptor origined from PIC-CONTROL. For other type of adaptors, you can also refer to other manufacturer as mention above.

click here to

Buy this SMD Adaptor Now at the PIC-store

PIC-store

<Step 1>

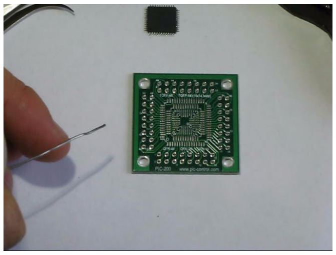

Before soldering, ensure that the soldering pad on the PCB is free of oxidation. PCB is usually tinned to protect the copper surface from oxidation. In cases where the surface is oxidized, you can try cleaning the surface with a contact cleaner. Apply the solution with tissue paper or toothbrush to clean up the contact. This solution make the PCB board rather oily, which I don't quite appreciate.

LPS
LVC
CONTACT CLEANER

Contact cleaner.

Let us start our smd soldering here. Soldering surface mount component. The first important thing to introduce would be the solder.

There are many type of solder. Choosing the correct solder takes a bit of some knowledge too. Leaded or lead free solder? The industries is moving towards Pb free components & PCB, calling for unleaded product. Lead is toxin and not environmental friendly. I would advise ou to wash your hand after soldering or handling pcb board. A soldering fume extractor to vacuum the smoky fume would be quite comfortable for your soldering. The smoke is quite choking and a health hazard if you breath that in.

soldering fume extractor.

Solder also comes with a variety of metal alloy combination. Nowadays it is common in solder core to contain flux. If the solder do not have the flux in the core, the flux have to be manually applied onto the metal surface before the soldering beginning. The flux cleans up the soldering surface, and also make the solder flow rather smooth like liquid. One of the important ingredient to make a good solder joint lays in the soldering flux.

Solder wire also comes in a wide range of diameter to choose from. Generally I use 0.8mm soldering wire roll with flux at the core. Smaller diameter solder wire allows you a better control of the amount of solder to apply. For soldering larger pins or connector component thicker diameter wire will be preferred. Wire too thin is rather troublesome, because you need to feed in a longer length of solder in order to solder the large surface/lead. For general purpose PCB board soldering 0.8mm wire fits somewhere in the middle. Not too thick nor too thin. For smd components, diameter smaller than 0.8mm is preferred. Using 0.8mm for smd soldering is not of any issue.

Solder wire 2mm

1.2mm

1.0mm

0.8mm

0.5mm

<Step 2>

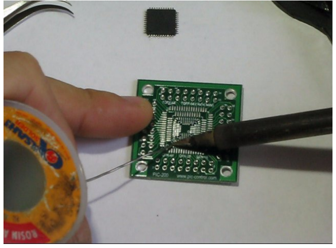

One of the pad being tinned.

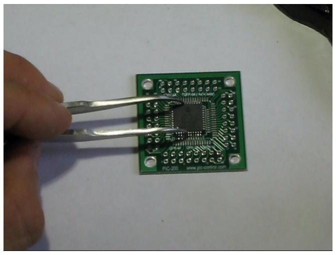

Align smd IC to the respective pad.

My microcontroller TQFP package has 44pins around the square package. To secure the package to the adaptor, I will need to solder one of the pin to the board first. It is important to solder only one pin. In case of mis-alignment, we can correct it easily. If more pins are soldered, it will be more difficult to correct any alignment.

Pin 1 of the adaptor pad is first tinned with a layer of solder. A tweezer is then used to align the IC chip on the pad. The IC is secure to the board with a slight touch from my soldering iron, soldering the IC pin to the pad.

The important factor to consider would be the heat generated by the iron. I prefer to use a 60W iron, which still works well in air-con room. A lower wattage or fine tip iron will not heat up that well in colder environment.

Goot TQ95 60W

If you prefer a fine tip, perhaps the soldering station will be more suitable. The iron temperature can be adjusted to a higher or precise temperature for the component that you want to solder. Component with thick metal contact like connector, heat sink, for thick cables should be solder with a higher temperature. The iron should be held on longer, in order to have the component fully heated up before any solder is applied.

Solder Station

For soldering the IC pin to the pad, you can start off by heating up the pad. Then while heating up the pad, touch the IC pin and apply the solder wire directly to the pad or IC pin, allowing the solder to flow like liquid. The IC pin is heated last to minimize the chance of over heating the pin. If it doesn't flow very well, you can manually some flux to the joint. The solder wire should not touch the iron directly, because this will vaporize the flux, and solder flow will be restricted.

Heat up the lead and pad longer allows the solder to flow and form a good joint. Not for too long as it can also damaged your IC. You should be able to see the solder flow eventually onto the lead & pad. With a bit of practice, you will see the difference between a quick solder touch and another one with lead & pad heat up a little bit longer. It takes a number of practice to get a good joint with minimum amount of heating time.

The soldering iron that I use has rather thick tip, but that does not matter when I do soldering for the smd microcontroller. All we do is to flood the pins with solder. They can be be suck up later.

There is a technology known as ultrasonic soldering. The soldering iron tip has this micro vibration which helps the solder to flow more easy without much use of flux. It can also solderthose difficult to solder material. The following is one article that I found that describe about it. [- Soldering the unsolderable].

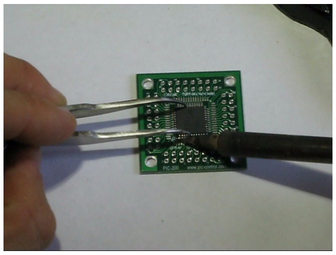
IC being fixed in position

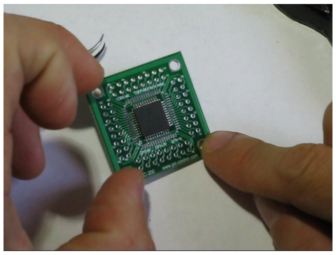
Checking if alignment is correct.

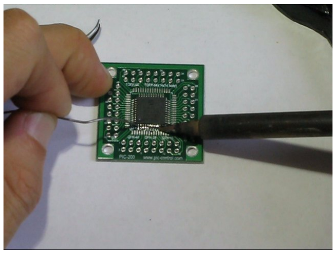
Flooding the pins with the rest of the solder.

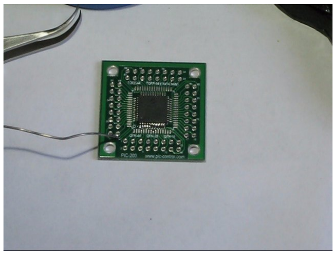
Pins flooded with solder

<Step 3>

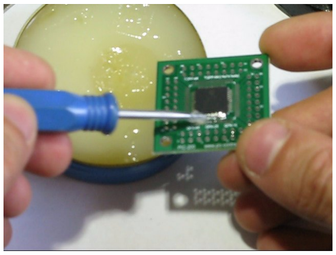

I have applied more flux at this stage. It should have been applied before flooding the pin, but it doesn't matter too much. The flux will make the solder rather liquid so that it is easier to suck up the excess solder.

ACID PASTE FLUX
SP-30
solder flux paste

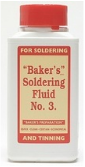

Baker Soldering Fluid

<Step 4>

The flooded solder is being heated up, and the excess solder are being sucked up.

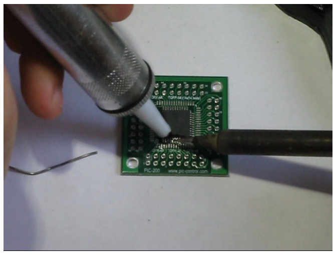

<Step 5>

Each pins is being touch up, ensuring that the pin is solder to pad and that they are not bridged across to the adjacent pins.

Now we have got our smd component soldered onto the board.

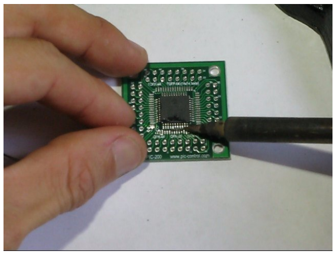

<Step 6>

The flux helps a lot in doing a good solder but it often leaves the board with tiny dots of transparent strain, a bit sticky, also known as grease or flux residue. The board looks untidy with these tiny bits of grease.

These are the various recommended cleaning solution for pcb.
- 50% Alcohol + 50% Water
- Thinner
- Flux remover
- Multicore Prozone MCF800
- Solvent-> Bromopropane; propane, 1-bromo-; propyl bromide (Chemical Formula: C_3H_7Br)

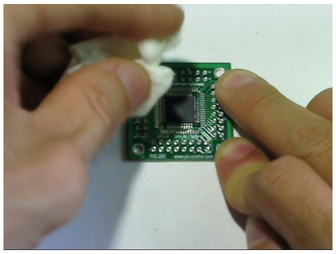

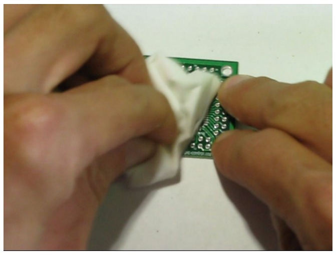

Alcohol MSDS (Material Safety Data Sheet)

Thinner MSDS (Material Safety Data Sheet)

ELECTROLUBE
FLU

Electrolube, FLU Fluxclene Cleaning Solvent, pdf file

Industrial Size
HEAVY DUTY FLUX REMOVER
HEAVY DUTY FLUX REMOVER

MG Chemicals 413B Heavy Duty Flux Remover

MGC-850 Stainless Steel Cleaning Brush

Brush for the crush

Flux Remover

Multicore Prozone MCF800

Besides using chemical, or after using chemical, there may still be some white blur stain. Those are flux spread dry up on the surface. Using a hot air blower, or heater on the surface can melt the dry up flux, and recover the smooth shiny PCB board surface again. Clean away the melt flux immediate after the heat.

Using alcohol is strongly recommended. Thinner is not suitable for certain plastic/material. Care is required when using thinner for cleaner. Using thinner on the PCB, copper pads and soldering joint isn't a problem. When use on plastic, you are advise to do a trial to ensure that the material is able to with stand thinner solution. Flux remover is commercially ready mixed solution. From what I learn in the manufacturer website, they are non-flammable chemical solution.

CAUTION

Alcohol & thinner has a lower flash point, and catch fire easily. You have to be extremely in handling and storing these flammable chemical. Do not store or work on the chemical near electrical appliances, hot area or things that can cause spark. Keep your windows open to allow air to circulate while working with these chemical. They may cause headache, dizziness and uncomfortable when inhaled.

Chemical comes in different level of grade. A slight different in the chemical concentration or purity can results in a large difference in cost. Alcohol can be expensive. For pcb cleaning purpose, request for technical grade alcohol. Higher grade alcohol is typically used for consumption or lab experiment purpose. Since we are using it for cleaning, there is not much advantage in using high quality chemical which is expensive. Dickson chemical is specialized in high graded chemical for laboratory use. They do sell lower cost technical grade alcohol. They are not suitable for consumption but is ok for cleaning application.

Dickson Chemicals

Dickson Chemical selling Technical grade alcohol

You can easily buy thinner from your local hardware. Typically used for cleaning your brushes after all the wall painting works.

The Science Of Zephyrtronics A Division of JTI, Inc.

Non-flammable Flux remover

click here to

Buy Cleaning Solution Now at the PIC-store

PIC-store

These are other recommendation in the web. After researching further, I would advise not use them. These chemical are of health hazard. Exposure will have severe adverse health effects.

- 50% isopropanol + 50% water
- Trichloroethylene
- Carbon tetrachloride

Isopropanol (true chemical name) is also known as Isopropyl alcohol or 2-propanol. 2- is refer as Iso. Other chemical may starts with 1- refer as N. These are some jargon used in the chemical industrial. I managed to learn a few of the terms from the vendor.
Trichloroethylene (true chemical name) is typically used for degreasing the metal. It is an extraction solvent for oil. Another name for Trichloroethylene is Trielin. Based on some read up, they can cause cancer.
http://www.answers.com/topic/trichloroethylene

Carbon tetrachloride can harm our environment, producing CFC which deplete the earth's ozone layer. http://en.wikipedia.org/wiki/Carbon_tetrachloride

Initially I tried using contact cleaner to clean the stain off. I use cloth or tissue but the fabric is often tear by the component sharp edges. Most corners are not easy to reach. Later on I tried using a toothbrush, the result is better.

For mass cleaning job, you can try using the ultrasonic pcb cleaner. It is a deep metal container for the solution, using micro vibration to shake off the dirt. I have never tried on it before, and I believe it is the same machine people used for cleaning their jewelry. Basically, you will need to dip the pcb board into the solution, and the ultrasonic will be activated to do it's job. The solution used is the same as manual cleaning, requiring 3-5 minutes of cleaning.

Ultrasonic cleaner

Ultrasonic jewelry cleaner

There is also a flux that does not leave residue on the PCB. I found it on the internet, and have not try it before.

INTERFLUX, IF 2005M No-Residue™ flux, pdf file

Article relating to removing flux residue:

- pcb cleaning - Aqueous Cleaning Process.pdf
- pcb cleaning - Flux Removal, nuxx.pdf
- pcb cleaning - pcb cleaning article.pdf
- pcb cleaning - PCB Cleaning, Printed Circ.pdf
- pcb cleaning - pcb washing chemical composite.txt
- pcb cleaning - remove solder flux.pdf
- pcb cleaning - Solvent washing PCB boards.pdf
- pcb cleaning - washing the board using alcohol.pdf
- pcb cleaning - PCB cleaning article Inventec_S24_03.pdf

<Step 7>

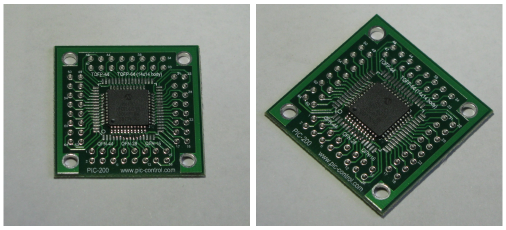

A close up view of the clean and neatly soldered surface mount IC on the prototyping adaptor. Soldering smd components is simple and fast with a bit of practice. Small and easy to clean up. There are too many advantage in working with surface mount, compared to through hole components.

This is the end of the short demonstration on smd soldering. I hope these article can provide you some insight and confident to start experimenting circuits using surface mount components.

Video available. Please click here,

smd soldering.mpg (25MB)

smd cleaning pcb.mpg (5.5MB)

Article on smd soldering.
- 2010-07 How to Solder Surface-Mount Devices.pdf

click here to
Buy this SMD Adaptor Now at the PIC-store

PIC-store

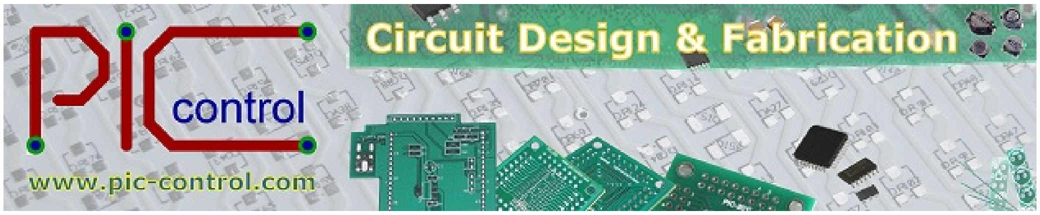

Singapore Customized, custom made Electronics Circuits & Kits

Various Chemical Reference:

Alcohol 70% + Distill Water 30%
MDX4-4159
7MHFE-71DE

Cyclohexanone
Methylethylketone
ethylacetate,
Silicon Wacker
Wacker M4503, General purpose mould making silicon (Using Two Part Silicon Mold Putty to Make Molds)
http://www.kirkside.com.au/, Mould Material Specialist

Propanol, used to clean PCB. (close to white electric oil)
Acetone, remove away ink printed on PVC stickers
n-Hexane (white electric oil)

smd glue for surface mount component mounting / soldering use
Seal-glo NE series, NE3000S, NE8800T, more Seal-glo

email: mail@siongboon.com
website: http://www.siongboon.com

Keyword: surface mount soldering, smd adaptor, smd to 2.54mm convertor

Surface Mount Soldering Procedure

We believe that the following procedures describe the simplest way to reliably assemble surface mount components without special tools. The only tools required are a pair of fine-tip tweezers, a rosin based flux pen, rosin core solder (.025” dia. preferred), and a soldering iron with a small, clean tip. A spool of fine (0.075”) solderwick will come in handy if you need to rework connections. You will need isopropyl alcohol, cotton balls, and cotton swabs for cleanup. If you have access to hot air rework tools and the skill to use them, feel free to use them instead.

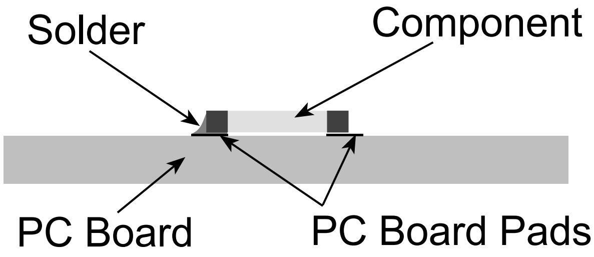

A proper solder joint with fillet

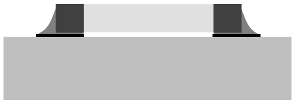

Proper joints

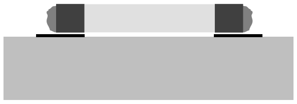

Poor joints

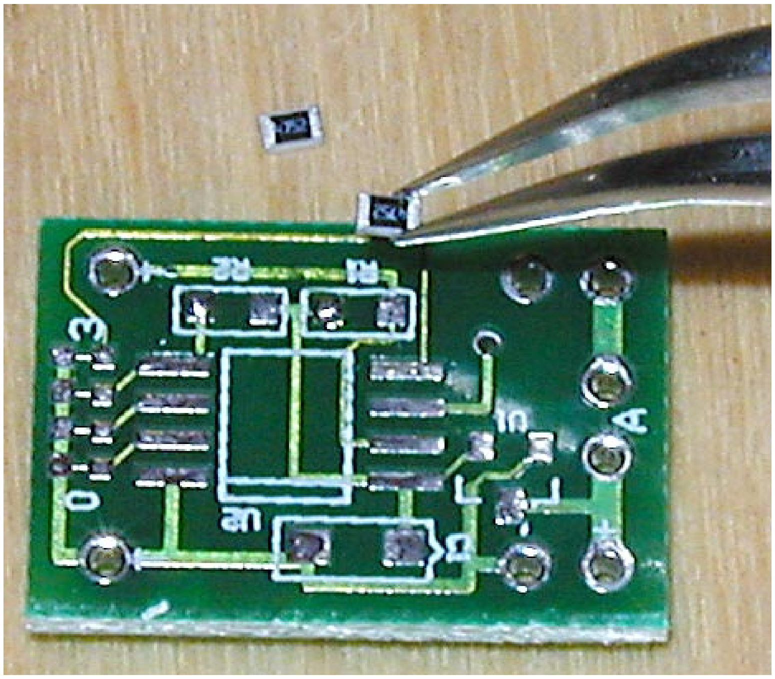

Placing the component

Soldering surface mount components is not difficult, but it does require good eyes, a steady hand, and a soldering iron with a small, clean tip. The illustrations at the left show a proper and a poor solder joint. In the proper joint, the flux has “wetted” both mating surfaces, allowing the molten solder to adhere to both the printed circuit board pad and the component lead. In the illustration of the poor connection, insufficient heat or flux has forced the solder to ball up on the end of the component, without making it to the PCB pad. The component is not attached properly, and the electrical connection will be intermittent at best. This connection can be salvaged by applying flux and heat to the PCB pad to get the solder to flow to the area between the pad and the component.

The most important characteristic of your soldering iron is that it must be equipped with a *small, clean* tip. If the tip is too large or is covered with oxidation, creating a well-flowed solder joint will be nearly impossible. Wipe your tip on a sponge before each joint - it must be shiny and well-tinned in order to transfer its heat properly.

You will find it easiest to work on your board if you clamp it in a moveable vise or stick it to a small block of wood with double-stick tape. This will allow you to rotate the board to whatever angle is most comfortable for access to the component being installed. If you try to hold the tiny parts at awkward angles, you will not have good results!

The first step is to apply a small bead of solder to one of the PCB pads for the component to be installed (the left-hand pads of R1 and R2 in the photo at right). The bead should be approximately 0.020” to 0.030” high (you may want to reevaluate your bead size after you have completed several joints). Apply some liquid *rosin based* flux from a flux pen to the solder bead and to the other pad(s) for the component.

Next, using a pair of tweezers, pick up the component to be installed and place it over the appropriate pads.

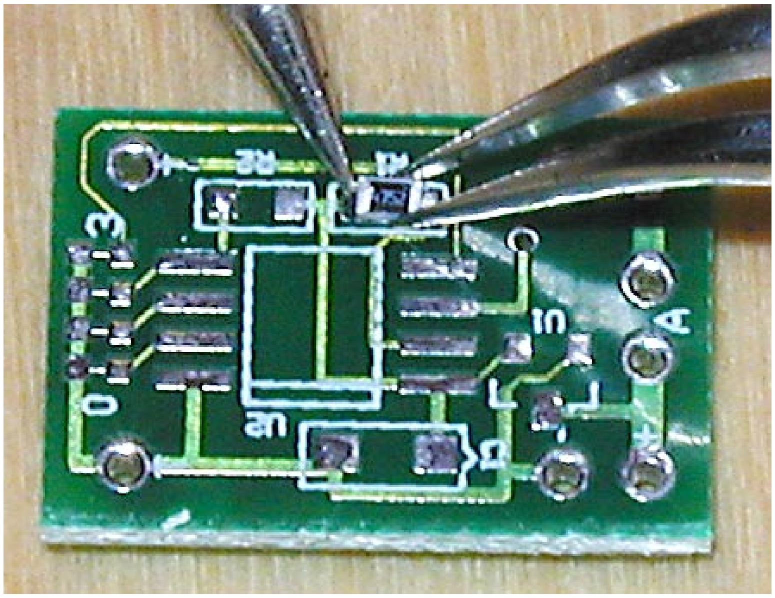

Affix the part by reflowing the first solder bead

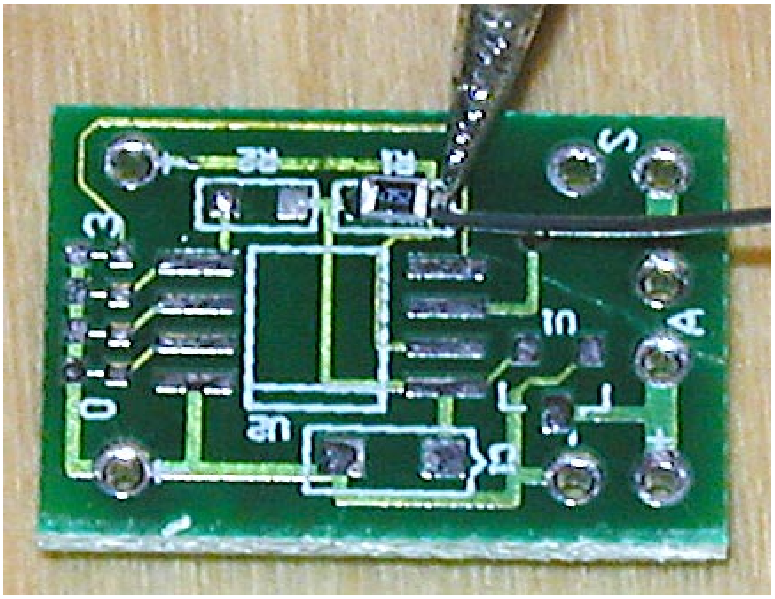

With the part in position, solder the other end

With the part in position, move the soldering iron to the solder bead on the PCB pad. Apply a small amount of heat from the iron to flow the bead, simultaneously lowering the part against the board and correcting for any rotational misalignment. Remove the iron and allow the solder to cool. Inspect the joint -- at this point you are not concerned about the quality of the actual solder joint, just the positioning of the component. The part should be flush against the PCB, with both ends properly contacting the pads. The part should be straight, and centered between the two pads.

Apply a liberal amount of liquid flux (again, a flux pen is the ideal applicator) to both ends of the component. Heat the currently unsoldered end of the component and the adjacent pad with the soldering iron, and carefully wipe on a small amount of solder. You want a small fillet as shown in the earlier illustration, not a large glob of solder. Make sure that the solder has flowed onto the pad as well as the component lead - don't be fooled by a lump of solder on the end of the component that doesn't flow under the component and onto the PCB pad.

After the second end is soldered, go back to the original solder joint and reheat the solder. The flux will allow the solder to flow freely, and a good fillet should be observed. If there seems to be insufficient solder, add a little more. If either of the ends appears to have an excessive amount of solder (a large blob sticking up above the part) it can be removed with solder wick. Apply flux to the blob and the wick, position the wick over the blob, and press lightly on the wick with your iron. As the heat is conducted through the wick the solder blob will melt and be drawn off by the wick. You may end up removing too much of the solder, in which case you can reapply a *small* amount of fresh solder to the joint.

After the joints have cooled, inspect them carefully to make sure that they are solid and make contact with the board. If you have any doubt about the quality of the connections, apply more flux and reflow the joints until you are satisfied.

The same procedure is also used for installing components with three or more leads (transistors, ICs, etc.). Start with a bead of solder on one pad, position and "tack" the part, then add more flux and solder the remaining leads. Don't forget to reflow/improve the first "tacked" connection if necessary.

Home Duane Linda BMW articles Czech Students Search my web Ji-young's travel

Electrical Wire soldering For Beginners

by Randy Glass

__DISCLAIMER-__ First, soldering can be dangerous. This lesson is not meant to be everything you need to know. Your safety is up to you, and you take full responsibility for any damage or injury caused by these instructions, whether as an act of omission or commission on your part or mine. Be aware that an improperly soldered wire can lead to fire, loss of vehicles, injury, and even death. By reading and/or using these instructions you take full responsibility and release me from all liability. Second, I took these photos as I did the work, pressing the remote control for my camera with one finger as I held the soldering pencil in one hand and the solder in the other. I apologize for any images that might be out of focus or poorly composed.

If Yogi Berra ever wrote a book on motorcycles he would probably say, "We seem to spend around forty percent of our time taking things apart, and the other seventy-five percent of the time putting them back together." Some things are easy to put together. We use specialized fasteners like nuts, bolts and screws to hold things together that, some day, we might want to take apart again. But there are some things that, once put together, we would like to know that, to some degree, our fastening system is permanent. One of those places is in electrical wiring. When a wire breaks or needs to be fastened to a connector or another wire in such a way as to be dependable, then the very best way is to solder.

soldering is a method to attach one metal part to another. It is part of a family that uses molten metal to do this job. At the top of the ladder is welding. This is a very high temperature method that actually melts the parts a bit where they are attached, along with some other metal, usually in the form of welding rods. Next is brazing. This melts some hard metal, usually brass, into the gap between two pieces of metal. Normally, the surface of the material being joined is not melted in this process, but it gets very, very hot. Finally there is soldering. soldering flows some molten metal between metal parts at relatively low temperatures.

For our purposes I am going to speak of the type of soldering that attaches two wires together or one wire to a connector- the sort of soldering that any backyard mechanic might need to do when working on any electrical component or the wires that connect them together in almost any type of vehicle.

WHY solder?

Although there are methods that can be used to make these connections without soldering, they are usually inferior. soldering does a lot of things:

1) It Is Strong- You are connecting two metal parts with a liquid metal that becomes solid after it fills the joint and cools. When properly done it leaves no voids so the joint is very strong and the wire is well supported.

2) It Slows or Stops Oxidation - Because solder fills the joint it excludes air and moisture so the metal in the joint will not oxidize. This maximizes electrical flow over time.

3) It Makes An Excellent Electrical Connection- because the solder is a conductor it gives the wire more area from which it can flow electrons to whatever it is connected.

4) Non-permanent - Although soldering is very strong, it can be removed and the joint taken apart in the future if necessary.

5) Connectors do Not Need to be Crimped - Those little metal connectors that are designed to be crimped have a terrible habit of coming loose as well as breaking the little copper wires in the conductor making a weak joint as well as one that does not conduct as well as it should. They easily let air and moisture in which causes corrosion and that leads to poor and broken connections.

MATERIAL LIST

It doesn't take much to set yourself up as a soldering wizard. Here is a list (numbers refer to image to the left):

*1) **__soldering Pencil -__** Something around 25 to 40 watts would probably be fine for most electrical work. You don't need a huge soldering gun to do electrical work. A little smaller is actually better than too big, as we shall see. Get a brand-named pencil so that down the line you will still be able to

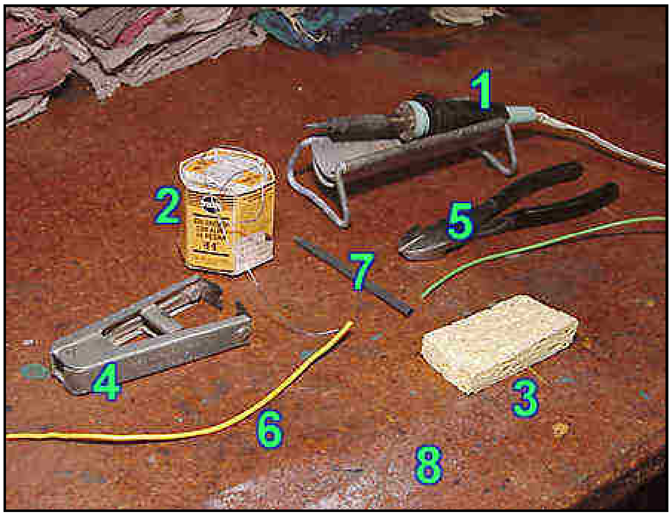

get replacement tips for it. A chisel tip as seen in the pictures that follow is best for most work but you might also need a pointed tip for more delicate work. Those old soldering guns are nice for making battery cables, but are generally too large for what we are doing. it would be like using a sledge hammer to make a picture frame- it would have to be a REALLY big picture frame!

There are also cordless, rechargeable soldering irons available. They generally run about $35-75 or so. Although these cost substantially more than a corded iron, and the replacement tips are more expensive as well, there is no cord to get in the way when working and as they can be carried along on trips. These generally have a built-in light as well making repairs at night easier. I normally loosen the lamp so that it does not light, so the battery lasts longer in use.

There are soldering tools that are actually little, refillable, butane-powered torches with a soldering tip at the end which is heated by the flame. As they are about the size of a fat ball point pen, they are another good way to get a portable soldering tool. These generally are available for less cost than the rechargeable irons mentioned above. I have used electric irons to solder around a gas tank on an automobile- something that I would never consider with one of these.

And a **soldering Pencil Holder** - Most soldering pencils come with a holder or a rest as seen here, but if not, get one. These metal stands are very handy and will help avoid melted parts and burned hands and work surfaces.

*2) **solder** - Resin core solder is what you want. NEVER use acid core solder on electrical components. The acid will stick around and continue to eat into components and wiring well after the soldering job is over.

Resin core solder can be had in various quantities, but I buy the one pound roll for around $12-15 which will be a lifetime supply for most folks. I use Kester brand "44" Resin Core solder in the .031 diameter size. This is about the size of mechanical pencil lead and works quite well for what you will be doing.

It use to be that a standard part of a soldering kit was a little tub of flux. This paste would be spread on the joint ahead of time to help clean the joint chemically as it was heated by the iron. It was necessary as flux-core solder was unavailable. If your wires are clean and you are using a quality, resin core solder, additional flux is not necessary.

*3) A**n Old kitchen sponge** or two

* **Tools** to work on wires such as a (4) wire stripper and a (5) wire cutter, and possibly needle nosed pliers and a small vise (not shown).

* 6) A few bits of **insulated wire** and an old spade connector or two on which to practice. Two or three feet of wire will do.

*7) **Heat Shrink Tubing** - I rarely use electrical tape any more. Heat shrink is really great! I usually buy it in three foot lengths in two or three sizes. This is a sort of memory tubing that, I believe, is made by heating a special tubing and stretching it and allowing it to cool while still stretched. it maintains that size. When heat is applied once again the tubing will shrink back down to its original size.

*8) Protection for the work surface like an old piece of 1/4" plywood about 12" square or such. An old piece of counter-top lamination works also.

* Wicking material (not shown)- You can purchase some special copper wick that is used to unsolder joints. This is usually not necessary, but handy at times, and inexpensive.

* **Eye protection**

* A Ventilated workspace

SOME WARNINGS

HEAT - A soldering pencil as well as molten solder are both really hot. They will burn you and it will hurt. It can hurt a lot. It can make your skin give off smoke and that smells bad. If left unattended a hot soldering pencil can cause fires, melt carpeting, burn workbenches, and make the cat dance like never before. An unsuspecting person can pick up a hot soldering iron and get a nasty burn (trust me on this one!). Never leave a soldering tool warming or hot and unattended.

Additionally, a hot soldering iron can almost instantly melt plastic parts and upholstery in a hurry. Use care when soldering as well as when moving the iron to and from the solder joint. If the cord gets caught when moving the iron it can easily cause you to drop the hot tool onto something expensive or painful. NEVER place a soldering iron on saddles, luggage, body work, painted surfaces, or on or near any heat-sensitive surface.

LEAD and VAPORS- solder contains a good amount of lead and handling it over a long period of time may have negative health consequences. The vapors caused by soldering can, over time, cause health problems as well. Work in a well ventilated area. I usually work out doors and use a fan to blow the vapors away from me. The smoke that is given off by the flux in the solder is also not a good thing to breathe.

MOLTEN METAL - When soldering (and particularly when un-soldering) it is very possible to flick bits of melted solder about. As this stuff is quite hot, even a small drop in the eye can cause permanent damage. Wear eye protection and do not wear your good glasses, or wear them under goggles. The dripped solder can melt plastic parts and permanently damage auto carpeting as well. Be aware of other folks or critters who may be in the area as well.

ELECTRICITY - You will be soldering on wires that, if shorted out, could cause a blown fuse, damaged electrical components, or melting wires and a fire. Always disconnect the battery when soldering wires on a vehicle and remove connectors from heat-sensitive components before soldering on them.

LETS GO

I could spend a lot of time talking about how it all works, but jumping right in is the best way to learn how to solder and I will explain as I go

Begin by creating a safe work area. One free of clutter, with sufficient light, and protected from the soldering iron and melting solder. Set the soldering pencil in its stand and plug it in. be sure that the cord is placed in a way that it will not be accidently pulled or tripped on, upsetting the soldering pencil.

While that is heating up, prepare your sponge. Wet it, squeeze out the excess water, then fold it in

half and rubber band it about half-way down.

Pull out about a foot or two of solder from the roll and touch it to the tip of the soldering iron. Always apply the solder to the flat parts of the tip and not to the sides. When the solder melts fairly easily, melt an amount all over the tip. There will be some smoke from the resin in the solder.

The solder you bought is actually a bit hollow. Inside of the metal there is a channel (or in some cases, channels) that make it sort of like a little garden hose. The channel is filled with a resin. When heated it melts and some of it burns off. The action of the resin when it is heated helps clean the surface to be soldered which helps the solder stick a lot better. Unfortunately, the burned resins can also leave deposits on the soldering pencil, and over time will make it difficult to solder. That's where the sponge comes in. After you have melted some solder onto the tip (about a quarter inch or so of the thin stuff) wipe the soldering pencil's tip off on the damp sponge a few times.

If you examine the tip of the soldering pencil carefully you will see that it is now covered with a thing coat of shiny, silver metal- the solder. You have successfully "tinned" the tip of the pencil. This keeps the metal tip from oxidizing and will assist you in soldering from now on. Always do this before starting a soldering job and occasionally during a larger job.

Place the soldering pencil back into its holder and let's prepare the wires. We are going to solder two wires together, making believe they have broken or need extending.

We start by "stripping" some of the plastic insulation away from where we want to connect the wires. The thicker the wire the more insulation we remove. For most common wires on a motorcycle or in a car, removing about 3/8 to � inch works fine. You will get a

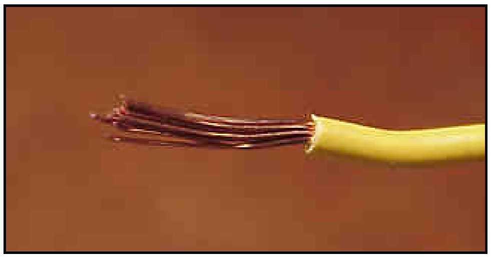

feel for how much in a short period of time. When stripping wire be very careful not to cut the strands of the conductor.

There are special tools available for stripping wire that can save a lot of time. These can be as simple as a carefully wielded razor blade to fancy and somewhat expensive automatic strippers.

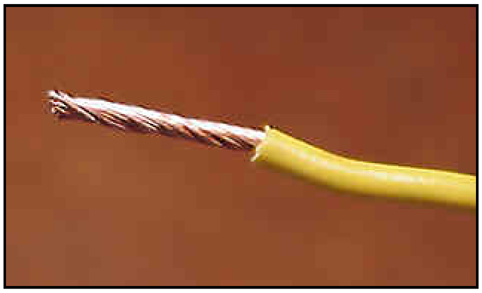

Once stripped, twist the strands together into a neat little bundle.

The exposed wires should look fairly clean and in most cases, shiny, but certainly not green or cruddy. If corroded, then the wires should be cut back and re-stripped until clean wire is found.

Now cut a length of heat shrink tubing off about an inch long. It should be of a large enough diameter that at least two of the wires you are soldering can easily fit into it at the same time. Slide the length of the tubing over one of the stripped wires- it doesn't matter which here, but in real life we would slide it over the lower wire or the one which slopes most downward away from the joint so it will tend to slide out of the way. make sure it is at least three or four inches away from the joint.

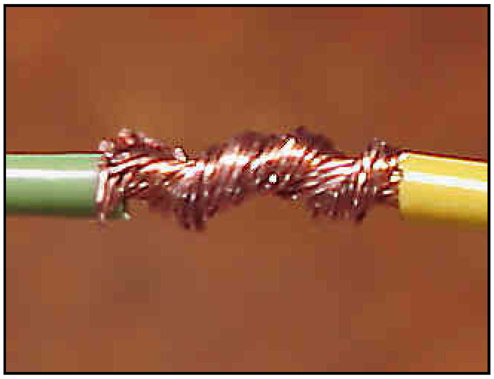

Now, twist the two wires together into a tight, neat pair. This is important because it helps create a good electrical connection. You also do not want any little wires sticking out. Start by crossing the wires about mid-strip, crossing them at about 45 degrees or so, and twisting in opposite directions. When done, the joint should be fairly strong on its own, holding the two wires together tightly. The connected pair of wires should be in line as if they are a single wire and not two joined together.

It's time to solder. Place the roll of solder on your left (if you are right handed) with a good length of solder hanging out of the roll. With the solder in your left hand and the soldering iron in your right hand, check to see that the tip is still tinned and shiny. If not, repeat the process with a bit of solder melted to the tip, then wipe the tip off on the sponge.

Melt a little ball of solder onto the iron....

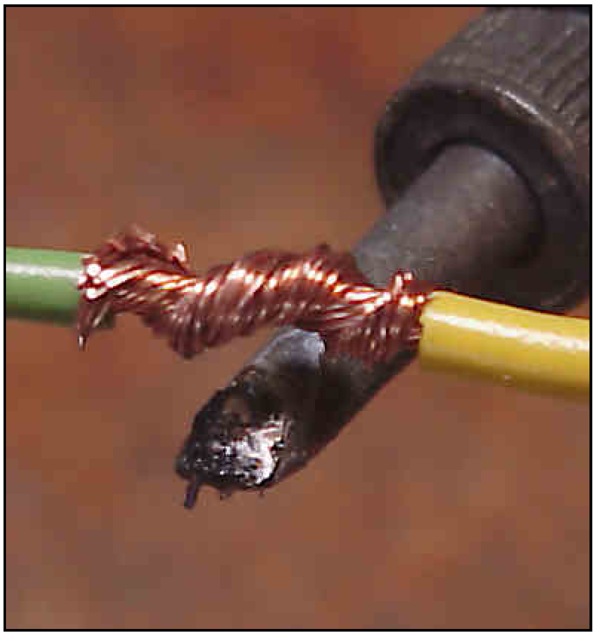

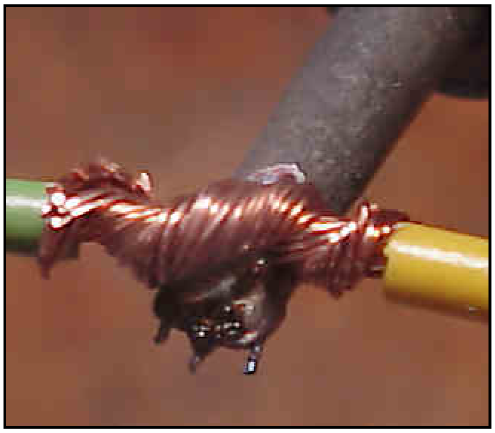

...and immediately place the soldering iron under the twisted wires (if possible) so that the little ball of solder fills much of the gap between the soldering pencil and the wire. Since heat rises, placing the iron under the joint will heat it faster. The little ball of solder works as a heat conductor and will very rapidly conduct the heat from the soldering pencil's tip into the wires. If the ball of solder does not seem like it is contacting the wire very well, without removing the pencil from the joint, melt just a bit more solder in there by feeding the end of the solder wire into the tip. This will not only increase the contact area of the solder to the wire but it will also add a bit of flux that helps the solder adhere to the wire at this time.

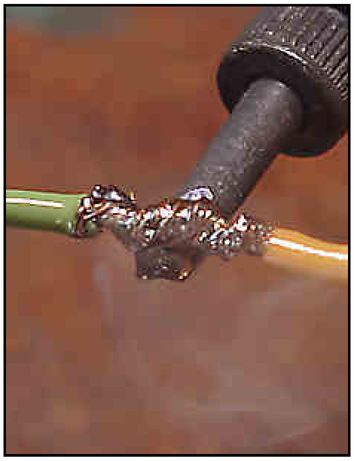

After about two or three seconds, start to feed solder, slowly at first, into the void between the pencil and the wire. You should soon see the solder is flowing into the twisted wires. WHen that happens, increase the feed rate of the solder into the joint, and you can now also feed a little into the top of the joint. Don't move the iron's placement, jut take the end of the solder wire and touch it to the tip of the joint. Continue to feed solder until the individual wires are almost filled with solder and the entire joint has turned silver. When that happens, stop feeding the solder in and immediately but gently remove the soldering iron.

How long do you wait until feeding in solder? One way to test how long to wait is to touch the solder on top of the joint and wait until that area is hot enough to melt the solder. The amount of time it takes varies with the size of the wires to be soldered, the size of the solder, the wattage of the iron, the size of the irons's tip, the ambient air temperature, and other factors. Experience will teach you how long this takes.

The problem comes from waiting too long and putting too much heat into the joint. When you do that it begins to melt the insulation on the wires around the joint (you can see that beginning to happen above at the end of the insulation of the green wire next to the soldered joint). That is why I like tho flow the solder in from where the iron's tip touches the wire. This is the place that will come up to soldering temperature first, and then the flowing solder will more easily flow solder into the joint as it carries heat with it. The larger the wires, the longer it takes and the more important it becomes to feed in some solder from the top of the joint as well. Again, experience will teach you when that is necessary.

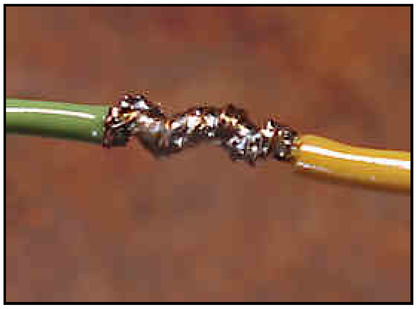

How much solder is enough? Too much? The correct amount of solder is used when the joint is fairly all silver colored but still shows the "outlines" of the individual strands of wire in the joint. If there is very much in the way of copper colored wire showing then you probably didn't use enough solder or the wires weren't clean and would not accept the solder. If there are big blobs and drips of hardened solder on the joint then you probably had either too cool of a joint or used too much solder.

After soldering, let the joint cool for about thirty seconds to a minute or so before handling. If you were successful you should have:

* A pair of wires tightly joined together

* The joint between them is very shiny, looking like polished sterling silver.

* No hardened dribbles or drops of solder hanging off the joint

* No little wires sticking out of the joint.

* A union only slightly larger than each of the wires themselves.

If you were NOT successful you might see:

* Large blobs or possibly a long, pointy finger of hardened solder hanging off the joint. Sometimes, after cleaning the iron, these can be removed by simply heating the joint until the solder melts, then removing the iron which will pull away some of the excess solder with it. This may need to be repeated a few times to remedy the situation.

* Wires that are not soldered together and has movement within it. Re-heat the joint and add some more solder.

* Melted insulation. This can be so bad that the wires are revealed with the insulation dripping off the wires. At this point you will probably need to cut the wires and begin again, or unsolder the joint and add a long piece of heat-shrink tubing to cover the exposed wires.

* Little ends of the copper wires sticking out. These can sometimes be cut off with a wire cutter and/or pressed down with a pair of smooth-jaw pliers, possibly needle nose pliers.

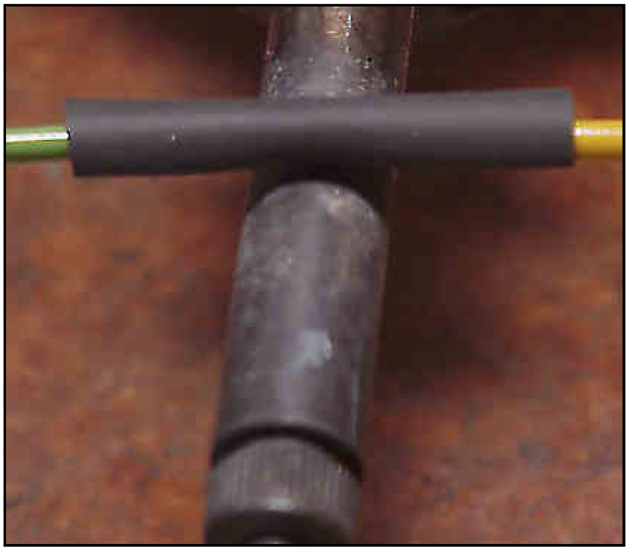

Now it is time to cover the joint with the heat shrink tubing. You DID remember to slide on the heat shrink tubing before beginning, didn't you? Once the joint is cool enough to touch, slide the tubing so it is centered over the joint and gently touch the meaty part of the soldering iron's heating element to the center of the tubing.

Continue this around the center of the tubing until it tightly grips the wire joint, then work outwards until the entire piece of tubing has shrunk tightly around the joint.

Attaching An Electrical Connector

Besides connecting two wires together, soldering is the best way top attach some sort of connector to the end of a wire. The connection is strong as well as fairly well sealed against the elements. It works pretty much the same way as above...

Remember when we "tinned" the tip of the soldering pencil? We did this to apply a film of solder to the tip so it wouldn't oxidize and thus be able to more easily transfer heat. Now I am going to do the same with a piece of wire.

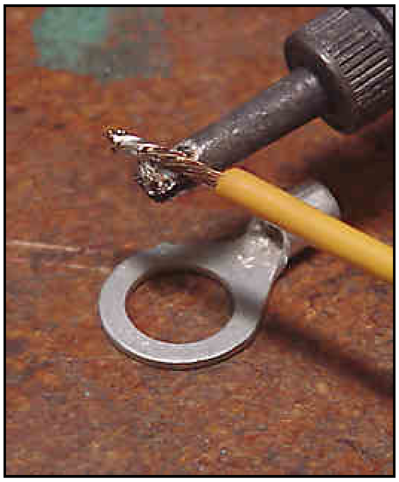

Here I am applying a good amount of solder ahead of time to this wire because I am planning on soldering it to a fairly heavy connector. This way, the wire will be "pre-soldered" and easier to attach, and will have less of a chance of melting the insulation.

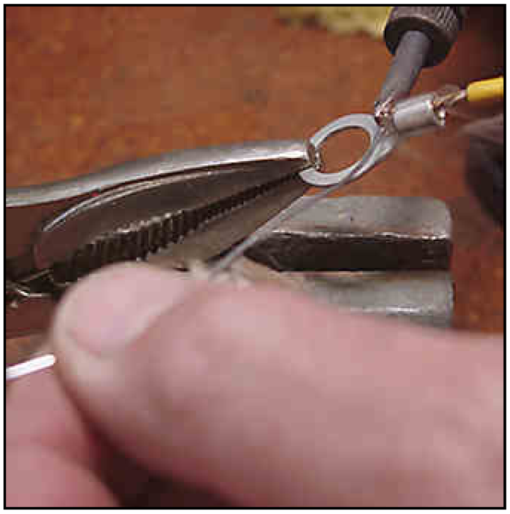

When soldering a connector to the end of a piece of wire it is very helpful to use some sort of vise to hold the connector. A small pair of locking pliers or a pair of needle nose pliers with a rubber band around the handles to hold the jaws shut works well. Here I have locked a pair of locking pliers into a small vise to work as a "hands-free" holder. This not only allows me to use both hands to work but isolates me from the heat of the soldering process.

Remember that the heavy metal pliers will work as a heat sink, pulling heat away from the soldering joint, so it might take longer to achieve the proper heat level in the joint. That is why I have gripped the connector as far away from the area to be soldered as possible,

Be careful as the pliers may get hot near the tips and you might overheat the wire as well. Be sure to slide the heat shrink tubing over then wire BEFORE beginning to solder! If you leave it off and finish soldering, you will be left with the soldering job in one hand, the heat shrink tubing in the other hand, and a stupid look on your face. We call this, "Heat Shrink Tubing Syndrome."

Additionally, if you apply too much solder, particularly with a small connector, it is very possible to fill the connector with solder and thus make it difficult or impossible for the connector to function properly. There are ways to get the solder out, but they are generally more difficult (or dangerous) than just chucking it and starting over. To keep this from happening:

* Never touch the area of the connector where you don't want solder to flow
* Orient the connector with the part where solder shouldn't flow, upwards (or Higher than the end where solder will flow).
* Apply solder sparingly and slowly
* Use a lower wattage soldering pencil
* Use a smaller tip on the soldering pencil

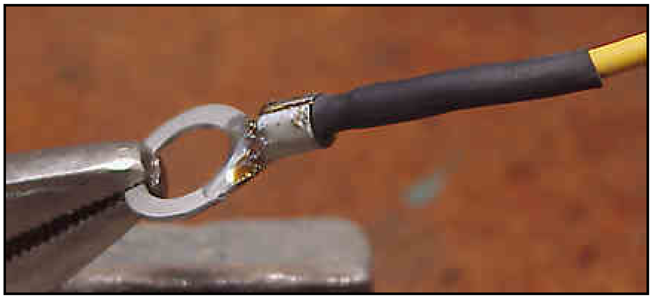

I sometimes use two pieces of heat shrink, one over the other, shrunk one at a time, in areas where extra protection is needed. This is particularly useful when soldering a connector to the end of a wire. I first shrink a small diameter piece about an inch long over the wire up to where it joins the connector as shown here. This acts as a strain relief as well as insulation.

Notice the small amount of flux on the connector ring left over from the soldering process. If you use resin core solder, this stuff is harmless to the electrical device. It can be brushed off with a small wire brush if you wish.

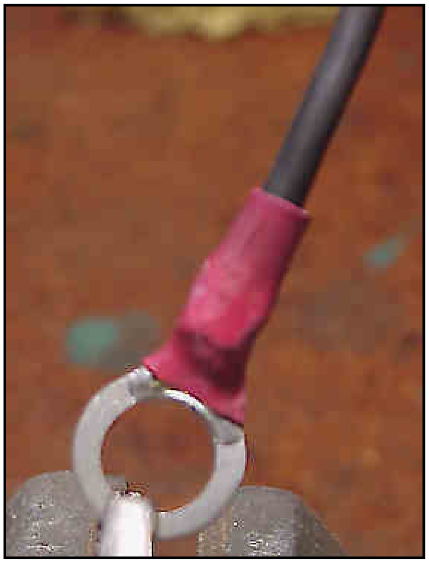

I then shrink another larger diameter piece about � an inch long over the connector to act as an insulating cover. In the case of something like a female spade connector, the entire connector except for the opening can be covered and protected in this way. If the heat shrink interferes with the connection, simply cut the offending part away with a razor blade being careful not to damage anything else including yourself!

In areas where the joint needs to be protected from corrosion, slather the terminal with petroleum jelly, or the special grease for this application.

Following these directions you will become a "master solderer" in a short period of time. It takes some practice to get a feel of how hot for how long with how much solder it takes to create to make a good solder joint, so get some scrap wire and practice before attempting to solder on anything important.

If there are any questions or comments, please feel free to E-mail me at frcn@cncnet.com.

Note by Duane; If you appreciate articles like this, I suggest that you thank Randy at the email address above.

More BMW info

Home page

This page was last edited: 02/03/2006 - copyright Duane Ausherman
Web hosting provided by hostmeister.com

Soldering and Desoldering Instruction

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Soldering is more like gluing with molten metal than anything else. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice.

Remember that when soldering, the rosin in the solder releases fumes. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area. Hot solder is also dangerous. Be sure not to let is splash around because it will burn you almost instantly. Eye protection is also advised.

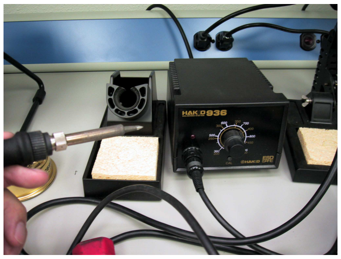

The Tools

Soldering Iron – Even the cheapest of them will do the job. They may not last as long, but they do get hot enough to melt solder and that is the goal. You only need to make sure that the one you buy has a suitable tip on it. The most typical tip is the one that is about the size of a 1/8” stereo mini-plug.

Solder – Get Rosin Core solder. Rosin will help the solder to flow onto the wires. Solder comes in different thickness. Buy what is appropriate for your job.

Desoldering Gun – This tool will make life a lot easier when you need to rework a previous solder job.
Solder Wick – Braided copper with rosin coating used to draw solder off of circuit boards.

SMD Rework station— Basically this is a high temp hair dryer with special nozzle attached. It heats up entire pins of the SMD device so you can remove them.

The Methods

Soldering – There is not art to soldering, it takes patience and practice to get it right. If you are doing it right, it will be easy and very fast. First, make sure that your soldering iron tip is clean. If it is dirtier than just a light gray color, you need to clean it. You can do this with sandpaper or a Scotch Brite pad. Next, turn your iron on and give it plenty of time to heat up. To test the heat, use a piece of solder touched to the tip. If it melts immediately, it is ready.

The most common way to mess up a solder job is to let the soldering iron stay on the parts to be soldered too long. What this does is oxidize the metal, making it too dirty to attract solder. If you are working with a circuit board, too much heat can damage the board, rendering it useless and in need of repair. You should only hold the soldering iron to the parts to be joined for no more than 10-12 seconds. Any longer and you will melt insulation on wire or damage a circuit board.

It has been said that soldering is a two-person job. You need two hands to hold the parts together, one hand to hold the iron, and another to feed the solder. The correct method for applying solder is to hold the joint perfectly still while heating with the soldering iron. After a few seconds, introduce the end of the solder at the point where the iron meets the parts

If the solder does not melt immediately and flow onto the joint, pull the solder away and try again after a couple seconds longer. If you exceed 10 seconds, pull the iron off and try again after it all cools down. You probably didn't have the iron touching enough of the parts to be soldered. Sometimes, the parts to be soldered are so big that they conduct the heat away very quickly and make it difficult to solder. In this case, it is OK to put some solder on each part individually and then put them together later by melting the solder on each one while they are touching.

Joining Two Wires - Strip off about 3/8" of insulation from the two wire ends to be joined. Place a ½" length of heat shrink tubing over one wire and push it back so that the heat from soldering won't shrink it prematurely. Fold each bare wire end back on itself so that the tip of the wire now comes back to the end of the insulation. Link both wires together using the bends like hooks. Now, twist the ends of the wires around themselves tightly. This type of connection is called a Western Union and it is the strongest method of joining two wires together. Now, heat the connection with the iron and allow solder to flow over the entire joint. Wait a few seconds for it to cool off and then slide the heat shrink tubing down over the connection and heat it with a match or heat gun.

Soldering a Wire to a Circuit Board – This is delicate work that you will do while soldering. Too much heat and your board will be damaged. If you are attaching a wire to an existing hole on the board, make sure that the hole is clean and free of any excess solder (see de-soldering for more detail). The best advice when soldering to a circuit board is to angle your soldering iron tip so that it makes good contact with the pad on the circuit board and the wire or part to be soldered at the same time. Heat the parts up and touch the solder at the point where the tip meets the part and the pad. If the solder does not flow immediately, pull everything off and wait for it to cool down and try again later.

Step 1

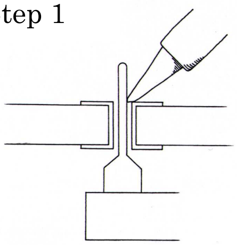

Apply heat to both pin and pad on the circuit board. Make sure tip of the soldering iron is touching both pin and pad.

Step 2

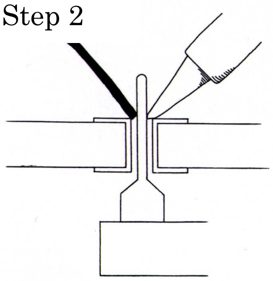

When the both pin and pad are heated enough, apply solder to the pin and pad. Solder should starts to flow.

Step 3

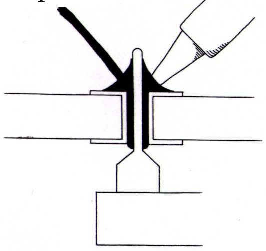

Apply enough but not too much solder. While applying the solder, do not move the tip of iron.

Step 4

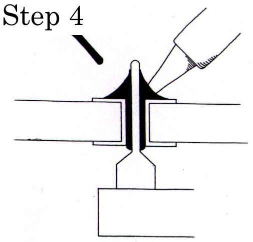

Remove solder away.

Step 5

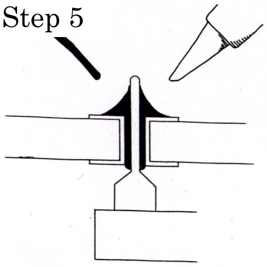

Remove tip of soldering iron also.

Step 6

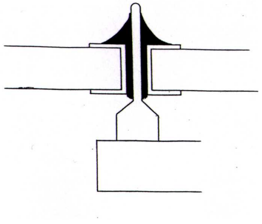

Done! Check the joint to make sure the connection.

Good soldering and Bad soldering Joints

Here are some example of Good soldering and Bad soldering joint.
For Japanese, we often say, “Good soldering joint should look like Mt. Fuji.”

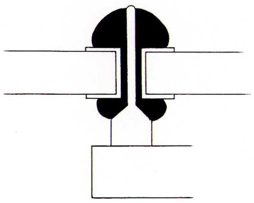

Too much soldering

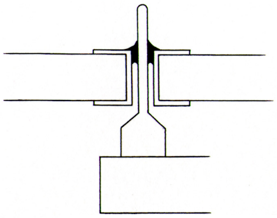

Too little soldering

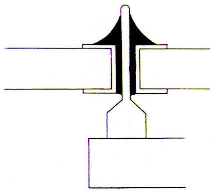

Good soldering

Cold Solder Joints

A cold joint is a joint in which the solder does not make good contact with the component lead or printed circuit board pad. Cold joints occur when the component lead or solder pad moves before the solder is completely cooled. Cold joints make a really bad electrical connection and can prevent your circuit from working.
Cold joints can be recognized by a characteristic grainy, dull gray color, and can be easily fixed. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can re-solder the joint, making sure to keep it still as it cools.

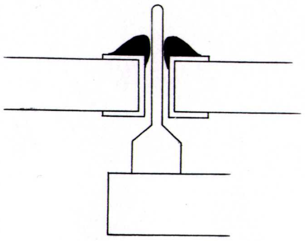

Cold soldering joints

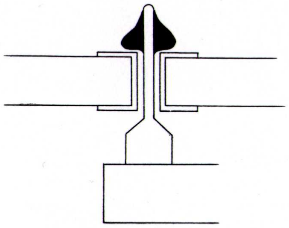

Tips and Tricks

Soldering is something that needs to be practiced. These tips should help you become successful so you can stop practicing and get down to some serious building.

1. Use heat sinks. Heat sinks are a must for the leads of sensitive components such as ICs and transistors. If you don't have a clip on heat sink, then a pair of pliers is a good substitute.

2. Keep the iron tip clean. A clean iron tip means better heat conduction and a better joint. Use a wet sponge to clean the tip between joints.

3. Double-check joints. It is a good idea to check all solder joints with an ohmmeter after they are cooled. If the joint measures any more than a few tenths of an ohm, then it may be a good idea to re-solder it.

4. Use the proper iron or temperature. Remember that bigger joints will take longer to heat up with an 30W iron than with a 150W iron. While 30W is good for printed circuit boards and the like, higher wattages are great when soldering to a heavy metal chassis.

Desoldering

Desoldering is extremely difficult compare to soldering. In the process of RomBoard installation, the parts and circuit board must be in the good shape to re-use them. The tool we use is Desoldering Gun. This device has vacuum pump built in with heater tip. Process of desoldering it self is very simple, but there are some tricks to do clean and safe desoldering job.

Step 1

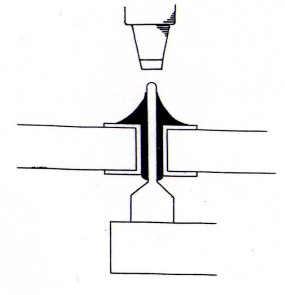

Put the desoldering gun's tip over to the soldering joint. Make sure nozzle is 90° angle from circuit board.

Step 2

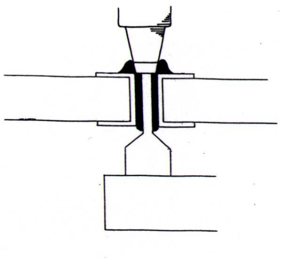

Apply the enough heat to melt entire solder. Multi-layer board or ground pin may take while to transfer heat.

Step 3

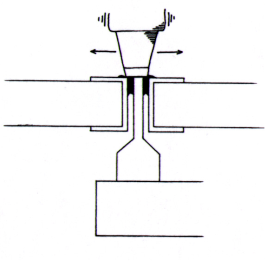

While applying vacuum, move the tip around the pin to get all the solder of around the pin out. At this time, you can also feel pin moves freely, so you know pin is free from circuit board.

Step 4

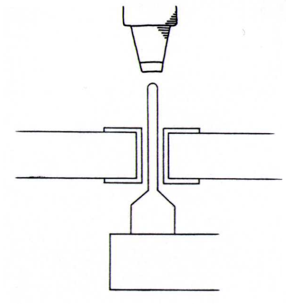

When you can hear air flow sound from hole, you should be done. Stop applying vacuum and remove the gun from the pin. Inspect hole and pin.

When you done with desoldering, the parts that you are trying to remove should move freely. If it doesn't, find which pin is still has solder left, and re-apply fresh solder to it and try desoldering process again. The multi-layer circuit board require more heat to get solder to melt. Make sure pin start to move freely by moving the tip of soldering gun before you apply vacuum to it.

SMD device soldering and removal

In the process of RomBoard installation, chance of handling SMD devices is becoming higher and higher due to ECU configuration change.
Soldering of SMD devices are not much different from regular through hole soldering.
Important thing is positioning of the device is very critical.
Start with applying little solder to two of the each corner of the pads on the circuit board.
Then, place the SMD device and re-heat a one of the solder you just applied to connect a SMD device pin to the pad. Check the position of the device, if position is right re-heat the other end of solder to secure the device completely. If you didn't get position right, re-heat the pin that you just soldered and while heating a pin, move the device to the right position.
Once you positioned the device in the right position, apply Rosin Flux to both pins and pads. This will help your soldering job by keeping solder separate from each pin.
Here is how the joint on SMD device pin should look like.

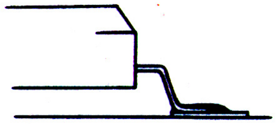

QFP and SOP chip

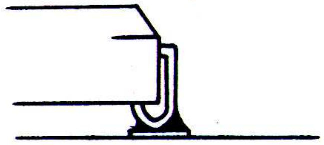

PLCC chip

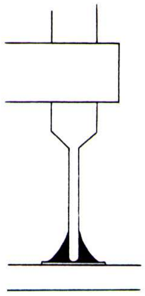

RomBoard adapter socket

Removing SMD device will be the probably the most difficult process. SMD rework station is used to do this. Make sure all the pins are completely heated otherwise you will be removing a pad from the circuit board along with device. Also, you should note, when you are applying heat to the device, some parts around the device is also heated and moves around when you touch. It is good idea to take a note of location of the devices near the target device.

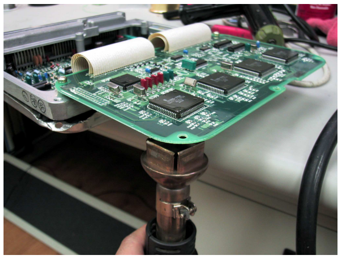

Step 1
Apply pre-heat from back side of the target device. Keep applying heat until you can't touch the chip. By this time, circuit board should have enough heat.

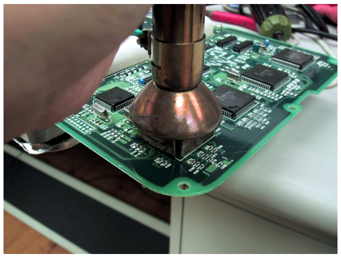

Step 2
Apply heat from top of the target device. Nozzle should cover entire chip. You can check if chip is free from solder by sliding the nozzle left and right. Make sure you don't move other parts. If the chip moves freely, it is ready to be removed.

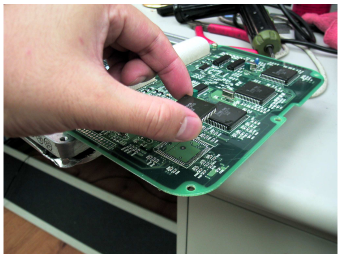

Step 3
Remove the chip straight up to avoid making solder bridge between pins. Chip is extremely hot! Unless you have skin of steel, avoid imitating the picture.

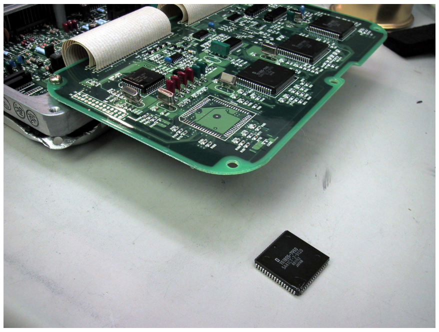

Step 4
After remove the chip, cool down both circuit board and chip.

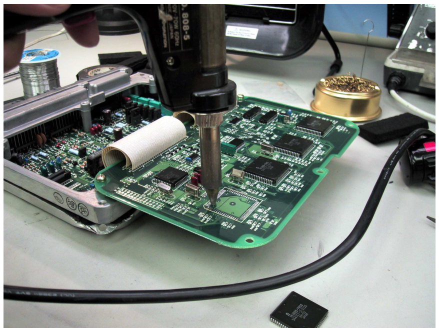

Step 5
Clean the pads on the circuit board with desoldering gun.
Do not apply too much heat to the pad. Pad might peel off from circuit board. Clean the chip by desoldering gun also.
Check for any bridge between pins. If the chip has too much protective coating on it, use solvent to clean it up.

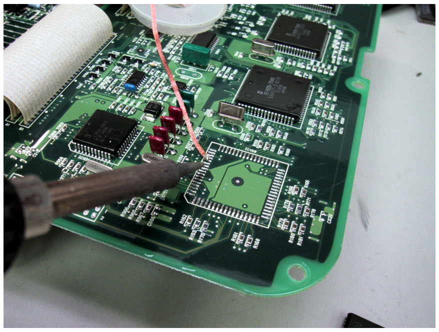

Step 6
If the pads on the circuit board is too small to use desolding gun, try cleaning by solder wick with rosin flux. This may take time but it is much better than peeling pads off.
Once again, do not apply too much heat.

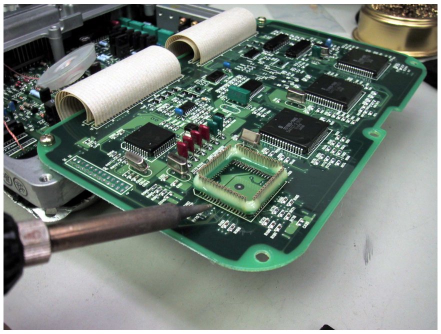

Step 7
Solder the RomBoard adapter where the chip was.
Refer SMD device soldering section for how to solder adapter.

The Final Word

Soldering, desoldering, and working with SMD device will take some effort to learn. You should be relaxed and concentrated when you work.
Keeping every thing clean (tip of iron, gun, work surface, etc) will help you also.
Once you get comfortable to work with soldering you may find your own tips and tricks.
Experience is all bout soldering.

EOF

How To Solder - Soldering Tutorial

Soldering is defined as "the joining of metals by a fusion of alloys which have relatively low melting points". In other words, you use a metal that has a low melting point to adhere the surfaces to be soldered together. Consider that soldering is more like gluing with molten metal, unlike welding where the base metals are actually melted and combined. Soldering is also a must have skill for all sorts of electrical and electronics work. It is also a skill that must be taught correctly and developed with practice. This tutorial will cover the most common types of soldering required for electronics work. This includes soldering components to printed circuit boards and soldering a spliced wire joint. This presentation came from http://www.aaroncake.net/electronics/solder.htm

Soldering Equipment

The Soldering Iron/Gun

The first thing you will need is a soldering iron, which is the heat source used to melt solder. Irons of the 15W to 30W range are good for most electronics/printed circuit board work. Anything higher in wattage and you risk damaging either the component or the board. If you intend to solder heavy components and thick wire, then you will want to invest in an iron of higher wattage (40W and above) or one of the large soldering guns. The main difference between an iron and a gun is that an iron is pencil shaped and designed with a pinpoint heat source for precise work, while a gun is in a familiar gun shape with a large high wattage tip heated by flowing electrical current directly through it.

A 30W Watt Soldering Iron

A 300W Soldering Gun

For **hobbyist electronics** use, a soldering iron is generally the tool of choice as its small tip and low heat capacity is suited for printed circuit board work (such as assembling kits). A soldering gun is generally used in heavy duty soldering such as joining heavy gauge wires, soldering brackets to a chassis or stained glass work.

You should choose a soldering iron with a 3-pronged grounding plug. The ground will help prevent stray voltage from collecting at the soldering tip and potentially damaging sensitive (such as CMOS) components. By their nature, soldering guns are quite "dirty" in this respect as the heat is generated by shorting a current (often AC) through the tip made of formed wire. Guns will have much less use in hobbyist electronics so if you have only one tool choice, an iron is what you want. For a beginner, a 15W to 30W range is the best but be aware that at the 15W end of that range, you may not have enough power to join wires or larger components. As your skill increases, a 40W iron is an excellent choice as it has the capacity for slightly larger jobs and makes joints very quickly. Be aware that it is often best to use a more powerful iron so that you don't need to spend a lot of time heating the joint, which can damage components.

A variation of the basic gun or iron is the soldering station, where the soldering instrument is attached to a variable power supply. A soldering station can precisely control the temperature of the soldering tip unlike a standard gun or iron where the tip temperature will increase when idle and decrease when applying heat to a joint. However, the price of a soldering station is often ten to one hundred times the cost of a basic iron and thus really isn't an option for the hobby market. But if you plan to do very precise work, such as surface mount, or spend 8 hours a day behind a soldering iron, then you should consider a soldering station.

The rest of this document will assume that you are using a soldering iron as that is what the majority of electronics work requires. The techniques for using a soldering gun are basically the same with the only difference being that heat is only generated when the trigger is pressed.

Solder

The choice of solder is also important. There several kinds of solder available but only a few are suitable for electronics work. Most importantly, you will only use *rosin core solder*. *Acid* core solder is common in hardware stores and home improvement stores, but meant for soldering copper plumbing pipes and not **electronic circuits**. If acid core solder is used on electronics, the acid will destroy the traces on the printed circuit board and erode the component leads. It can also form a conductive layer leading to shorts.

For most printed circuit board work, a solder with a diameter of 0.75MM to 1.0MM is desirable. Thicker solder may be used and will allow you to solder larger joints more quickly, but will make soldering small joints difficult and increase the likelihood of creating solder bridges between closely spaced PCB pads. An alloy of 60/40 (60% tin, 40% lead) is used for most electronics work. These days, several lead-free solders are available as well. **Kester "44" Rosin Core** solder has been a staple of electronics for many years and continues to be available. It is available in several diameters and has a non-corrosive flux.

Large joints, such as soldering a bracket to a chassis using a high wattage soldering gun, will require a separate application of brush on flux and a thick diameter solder of several millimeters. Remember that when soldering, the flux in the solder will release fumes as it is heated. These fumes are harmful to your eyes and lungs. Therefore, always work in a well-ventilated area and avoid breathing the smoke created. Hot solder is also dangerous. It is surprisingly easy to splash hot solder onto yourself, which is a thoroughly unpleasant experience. Eye protection is also advised.

Preparing To Solder

Tinning the Soldering Tip

Before use, a new soldering tip, or one that is very dirty, must be tinned. "Tinning" is the process of coating a soldering tip with a thin coat of solder. This aids in heat transfer between the tip and the component you are soldering, and also gives the solder a base from which to flow from.

Step 1: Warm Up the Iron

Warm up the soldering iron or gun thoroughly. Make sure that it has fully come to temperature because you are about to melt a lot of solder on it. This is especially important if the iron is new because it may have been packed with some kind of coating to prevent corrosion.

Step 2: Prepare A Little Space

While the soldering iron is warming up, prepare a little space to work. Moisten a little sponge and place it in the base of your soldering iron stand or in a dish close by. Lay down a piece of cardboard in case you drip solder (you probably will) and make sure you have room to work comfortably.

Step 3: Thoroughly Coat the Tip in Solder

Thoroughly coat the soldering tip in solder. It is very important to cover the entire tip. You will use a considerable amount of solder during this process and it will drip, so be ready. If you leave any part of the tip uncovered it will tend to collect flux residue and will not conduct heat very well, so run the solder up and down the tip and completely around it to totally cover it in molten solder.

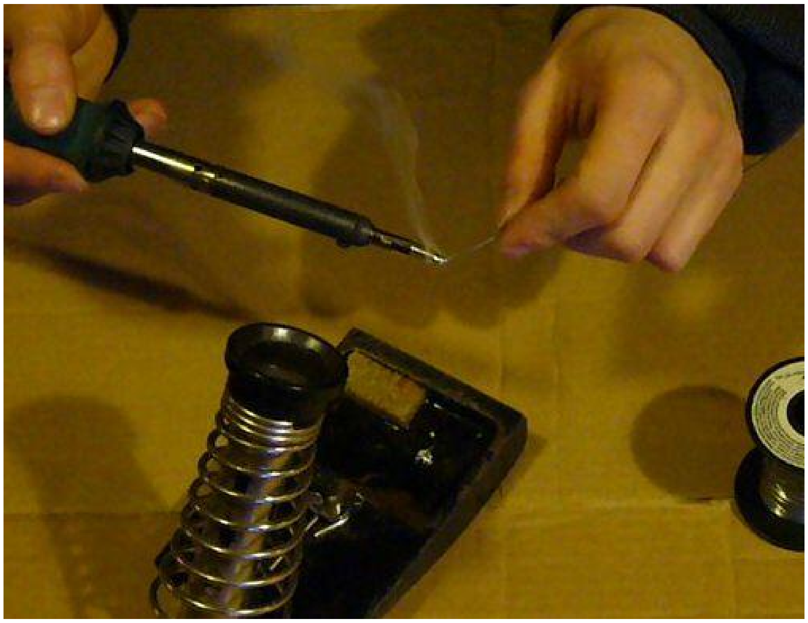

.

Step 4: Clean the Soldering Tip

After you are certain that the tip is totally coated in solder, wipe the tip off on the wet sponge to remove all the flux residue. Do this immediately so there is no time for the flux to dry out and solidify.

Step 5: You're Done!

You have just tinned your soldering tip. This must be done anytime you replace the tip or clean it so that the iron maintains good heat transfer.

Soldering a Printed Circuit Board (PCB)

Soldering a PCB is probably the most common soldering task an electronics hobbyist will perform. The basic techniques are fairly easy to grasp but it is a skill that will take a little practice to master. The best way to practice is to buy a simple electronics kit or assemble a simple circuit (such as an **LED chaser**) on a perf-board. Don't buy that expensive kit or dive into a huge project after only soldering a few joints.

Soldering components onto a PCB involves preparing the surface, placing the components, and then soldering the joint.

Step 1: Surface Preparation:

A clean surface is very important if you want a strong, low resistance solder joint. All surfaces to be soldered should be cleaned well. **3M Scotch Brite pads** purchased from the home improvement, industrial supply store or automotive body shop are a good choice as they will quickly remove surface tarnish but will not abrade the PCB material. Note that you will want *industrial* pads and not the kitchen cleaning pads impregnated with cleaner/soap. If you have particularly tough deposits on your board, then a fine grade of steel wool is acceptable but be very cautious on boards with tight tolerances as the fine steel shavings can lodge between pads and in holes.

Once you have cleaned the board down to shiny copper you can use a solvent such as acetone to clean any bits of the cleaning pad that may remain and to remove chemical contamination from the surface of the board. Methyl hydrate is another good solvent and a bit less stinky then acetone. Be aware that both these solvents can remove ink, so if your board is silk screened, test the chemicals first before hosing down the entire board.

A few blasts with compressed air will dry out the board and remove any junk that may have built up in the holes.

Step 2: Component Placement

After the component and board have been cleaned, you are ready to place the components onto the board. Unless your circuit is simple and only contains a few components, you will probably not be placing all the components onto the board and soldering them at once. Most likely you will be soldering a few components at a time before turning the board over and placing more. In general it is best to start with the smallest and flattest components (resistors, ICs, signal diodes, etc.) and then work up to the larger components (capacitors, power transistors, transformers) after the small parts are done. This keeps the board relatively flat, making it more stable during soldering. It is also best to save sensitive components (MOSFETs, non-socketed ICs) until the end to lessen the chance of damaging them during assembly of the rest of the circuit.

Bend the leads as necessary and insert the component through the proper holes on the board. To hold the part in place while you are soldering, you may want to bend the leads on the bottom of the board at a 45 degree angle. This works well for parts with long leads such as resistors. Components with short leads such as IC sockets can be held in place with a little masking tape or you can bend the leads down to clamp onto the PC board pads.

In the image below, a resistor is ready to solder and is held in place by slightly bent leads.

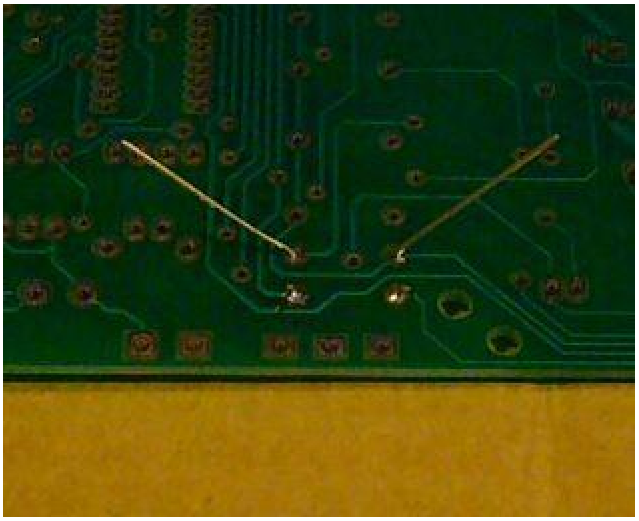

Step 3: Apply Heat

Apply a very small amount of solder to the tip of the iron. This helps conduct the heat to the component and board, but it is **not** the solder that will make up the joint. To heat the joint you will lay the tip of the iron so that it rests against both the *component lead* and the *board*. It is critical that you heat the lead and the board, otherwise the solder will simply pool and refuse to stick to the unheated item. The small amount of solder you applied to the tip before heating the joint will help make contact between the board and the lead. It normally takes a second or two to get the joint hot enough to solder, but larger components and thicker pads/traces will absorb more heat and can increase this time.

If you see the area under the pad starting to bubble, **stop** heating and remove the soldering iron because you are overheating the pad and it is in danger of lifting. Let it cool, then carefully heat it again for much less time.

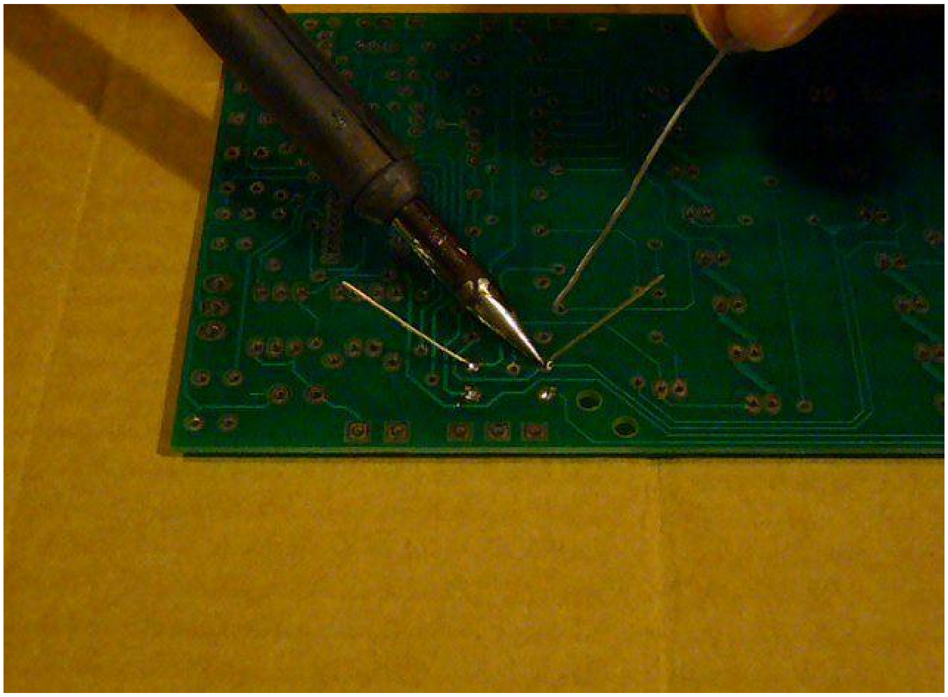

Step 4: Apply Solder to the Joint

Once the component lead and solder pad has heated up, you are ready to apply solder. Touch the tip of the strand of solder to the component lead and solder pad, but **not** the tip of the iron. If everything is hot enough, the solder should flow freely around the lead and pad. You will see the flux melt liquefy as well, bubble around the joint (this is part of its cleaning action), flow out and release smoke. Continue to add solder to the joint until the pad is completely coated and the solder forms a small mound with slightly concave sides. If it starts to ball up, you have used too much solder or the pad on the board is not hot enough.

Once the surface of the pad is completely coated, you can stop adding solder and remove the soldering iron (in that order). Don't move the joint for a few seconds as the solder needs time to cool and solidify. If you do move the joint, you will get what's called a "cold joint". This is recognized by it's characteristic dull and grainy appearance. Many cold joints can be fixed by reheating and applying a small amount of solder, then being allowed to cool without being disturbed.

Step 5: Inspect the Joint and Cleanup

Once the joint is made you should inspect it. Check for cold joints (described a little above and at length below), shorts with adjacent pads or poor flow. If the joint checks out, move on to the next. To trim the lead, use a small set of side cutters and cut at the top of the solder joint.

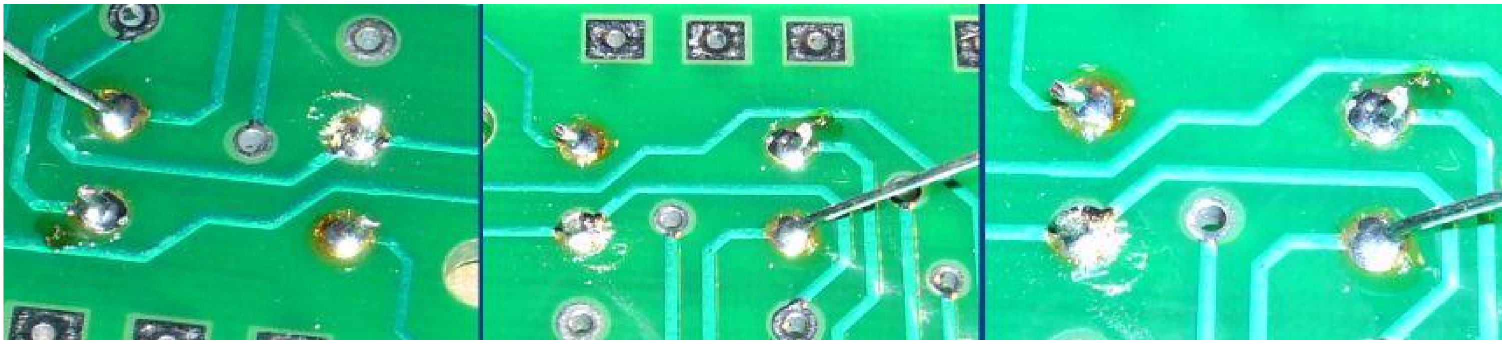

After you have made all the solder joints, it is good practice to clean all the excess flux residue from the board. Some fluxes are hydroscopic (they absorb water) and can slowly absorb enough water to become slightly conductive. This can be a significant issue in a hostile environment such as an automotive application. Most fluxes will clean up easily using methyl hydrate and a rag but some will require a stronger solvent. Use the appropriate solvent to remove the flux, then blow the board dry with compressed air.

Cold Solder Joints

A "cold solder joint" can occur when not enough heat is applied to the component, board, or both. Another common cause is a component moving before the solder has completely cooled and solidified. A cold joint is brittle and prone to physical failure. It is also generally a very high resistance connection which can affect the operation of the circuit or cause it to fail completely.

Cold joints can often be recognized by a characteristic grainy, dull gray color, but this is not always the case. A cold joint can often appear as a ball of solder sitting on the pad and surrounding the component lead. Additionally you may notice cracks in the solder and the joint may even move. Below is the shocking image of every example of a bad solder joint you will ever see. It appears that this **FM transmitter kit** was assembled using the technique of "apply solder to iron then drip onto joint". If your joints are looking like this, then **stop** and practice after rereading this page. Note that not a single of of these joints is acceptable, but amazingly, the circuit worked.

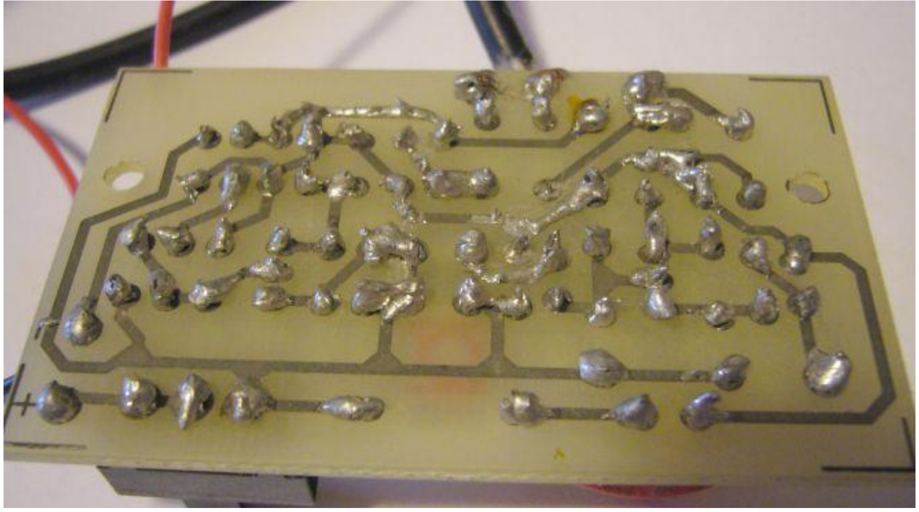

Most cold solder joints can be easily fixed. Generally all that is required is to reheat the joint and apply a little more solder. If there is already too much solder on the joint, then the joint will have to be **desoldered** and then soldered again. This is done by first removing the old solder with a desoldering tool or simply by heating it up and flicking it off with the iron. Once the old solder is off, you can resolder the joint, making sure to heat it thoroughly and keep it still as it cools.

advertisement

HOW TO SOLDER - THE QUICK, THOROUGH GUIDE

Technology (/technology/) > Soldering | by IoTalabs (/member/IoTalabs/) Follow

4,064 ♥ 38 4

Posted Jun. 9, 2016 | (cc) BY-NC-SA

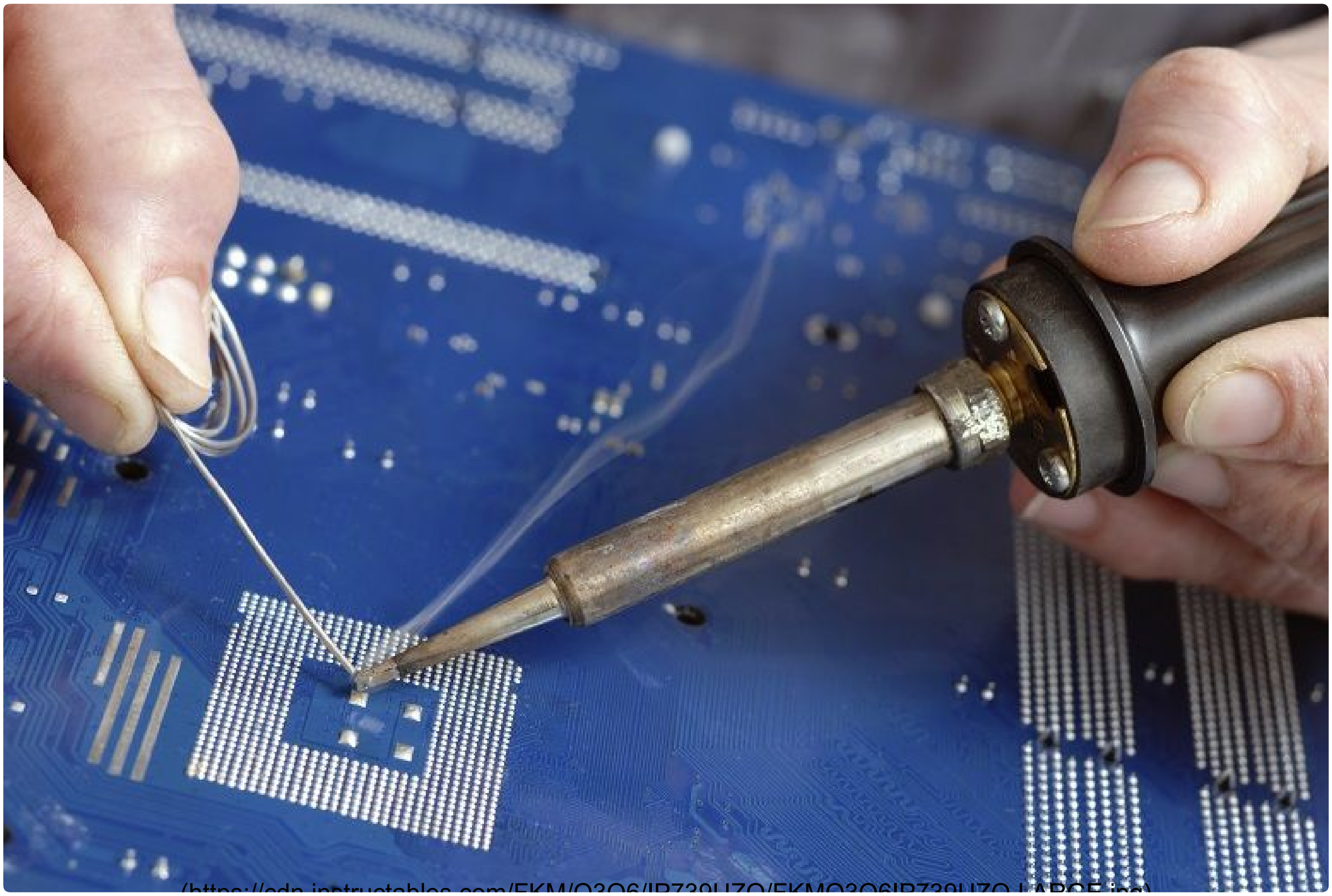

(https://cdn.instructables.com/FKM/O3O6/IP739UZO/FKMO3O6IP739UZO.LARGE.jpg)

We are IoTalabs and we are a group of Internet of Things enthusiasts that love hacking together different devices. Over the past few months we have immersed ourself in the wonderful world of hardware! In our journey, we learned the best soldering practices around.

We've been working on many projects (check out our current project http://doteverything.co/) and we wanted to share our experience on safely and correctly soldering connections!

Add Tip | Ask Question

Step 1: What Is Soldering?

Soldering accomplishes a strong bond between two pieces of metal by joining them together. In this procedure, a material called solder, an alloy mixture of tin and lead, flows over two pre-heated pieces of metal and holds them together. The process is similar to welding but differs because when you weld you are fusing and melting two pieces together to make one. When you solder you are essentially ‘gluing’ two parts together with molten metal. Most metals with the exception of aluminum, white metal and porous cast iron can be soldered. Below, you will find instructions and illustrations that show you how to use a soldering iron.

Add Tip | Ask Question

Step 2: Materials for Soldering

SOLDERING FLUX
ZINC CHLORIDE
Internally!
Eyes and Skin
Contact: Flush
Call Physician
PLENTY of WATER
PHYSICIAN!
REACH OF CHILDREN!
SOLDER
ROSIN CORE
16 OZ.
(1 LB.)
NET WT.
MADE IN
U.S.A.
NET WT. 2 OZ.
alpha
SOLDERING
PASTE FLUX
METAL CLEANER FOR SOLDERING
Ask for ALPHA METALS top quality
RELIABLE
UNGAR

(https://cdn.instructables.com/FZG/59XY/IP74OGXI/FZG59XYIP74OGXI.LARGE.jpg)

Soldering Iron

Modeler’s Vise or Frame (optional)

Solder

Damp Sponge

Flux to remove oxides

Add Tip Ask Question

Step 3: Prepare Your Workspace

Make sure your table or bench top is clear and free of obstructions. You will want as much freedom as possible to move your hand around and make adjustments.

Add Tip Ask Question

Step 4: Turn on the Soldering Iron

The soldering iron needs to be warmed up properly before it can be used to ensure clean application of the solder. Some soldering irons heat up in seconds while others take minutes. Turn it on and leave it in the stand for 2-3 minutes to be safe.

Add Tip Ask Question

Step 5: (Optional) Secure the Items You Are Soldering

Because we only have two hands, it would be nice to stabilize the item we want to solder. This is where a vice or clamp that can hold your item sturdily would come in handy. This is not required, but can greatly increase the ease and quality of your solder joints.

Add Tip Ask Question

Step 6: Cleaning the Soldering Iron

(https://cdn.instructables.com/FVM/SZ4M/IRZ4OGX5/FVMSZ4MIRZ4OGX5.LARGE.jpg)

Because soldering irons get so hot, they oxidize and become dirty quickly. They key to reliable connections is clean components so make sure that your soldering tip and parts you are joining are clean. To accomplish this, pass the tip of your soldering iron on a wet sponge until it shines.

Add Tip Ask Question

Step 7: Apply Flux

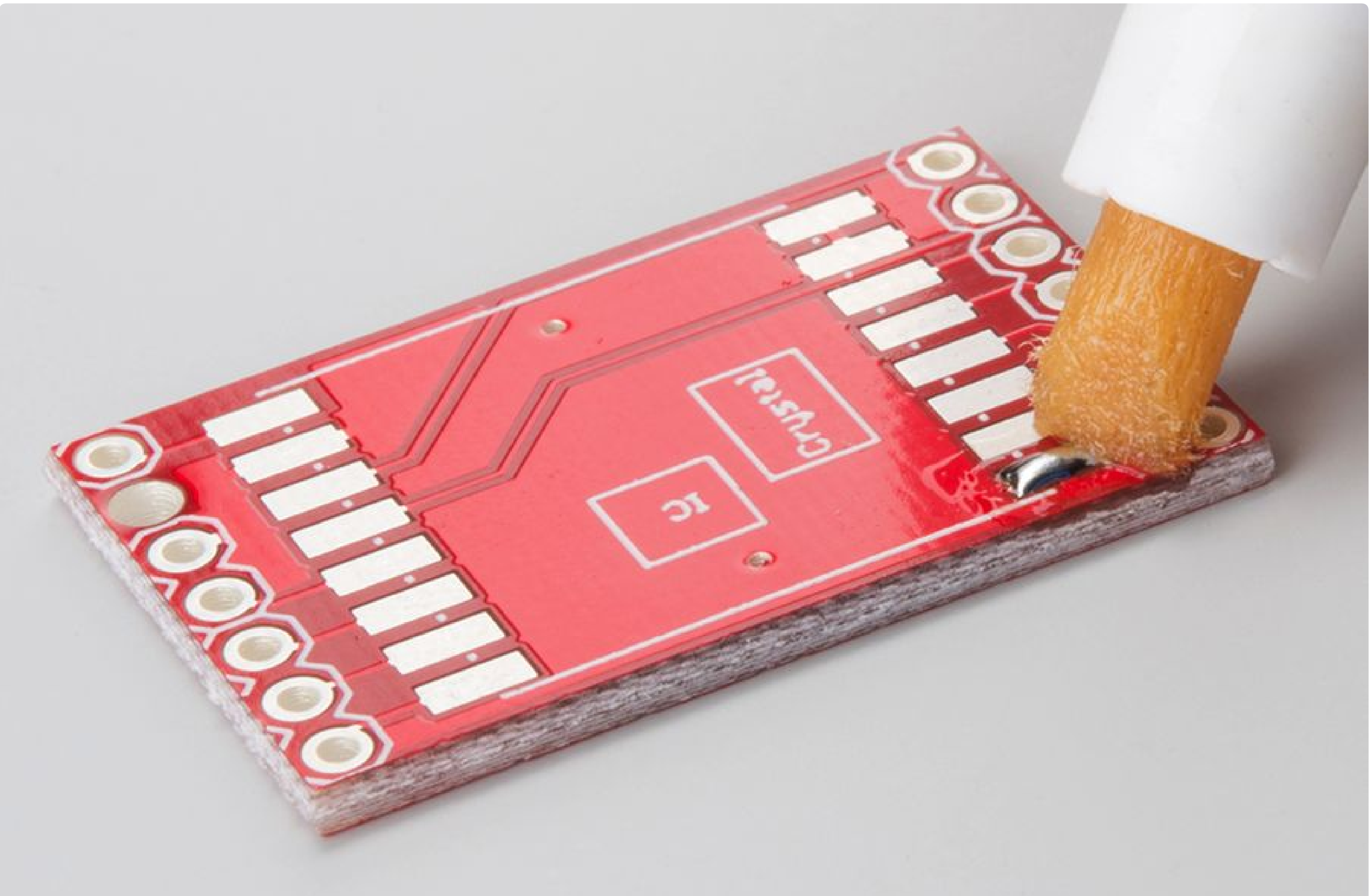

(https://cdn.instructables.com/FTD/PDPV/IBZ4OGXZ/FTDPDPVIBZ4OGXZ.LARGE.jpg)

In soldering it often becomes necessary to use materials called fluxes to help remove oxides and keep them absent while you solder. Flux needs to melt at a temperature lower than solder so that it can do its job prior to the soldering action. There are different methods to apply flux. The method you choose will be dependent on the items you are soldering.

Add Tip Ask Question

Step 8: Tin the Soldering Iron

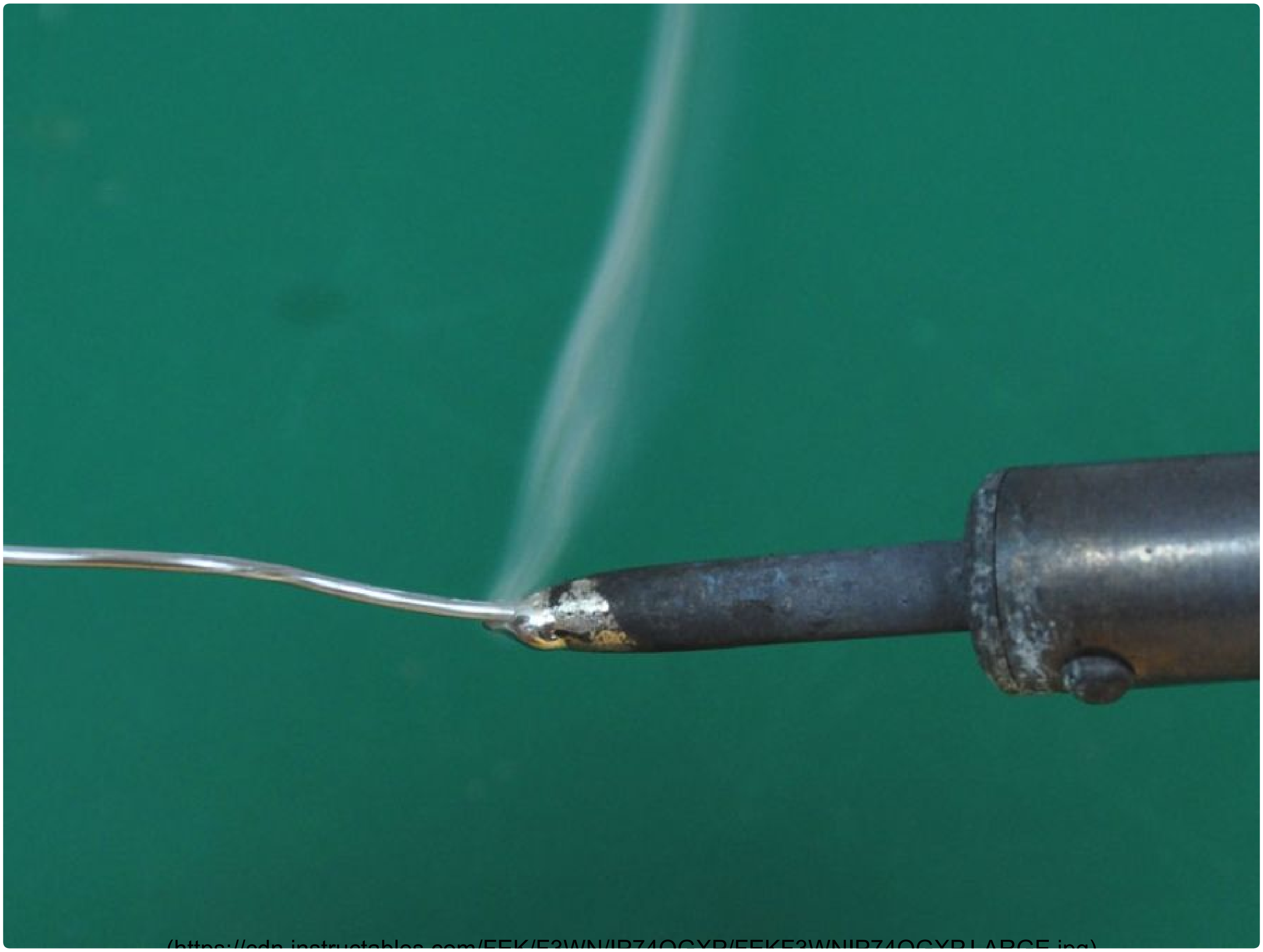

(https://cdn.instructables.com/FFK/F3WN/IRZ4OGXB/FFKF3WNIRZ4OGXB.LARGE.jpg)

If you want to know everything there is to know about how to use a soldering iron, you'll need to know how to tin. Tinning is the process of coating a soldering tip with a thin coat of solder. Melt a thin layer of solder on your iron's tip. This aids in heat transfer between the tip and the component you are soldering, and also gives the solder a base from which to flow from. This process may need to be repeated as you solder. You will only touch the tip of the soldering iron to the solder when you tin. Do not touch the tip of the iron to the solder while you are actually soldering.

Add Tip Ask Question

Step 9: Start Soldering

Hold the soldering iron like you would a pen in the hand you write with and the solder in the other. Then place the tip of the soldering iron. The tip needs to touch both the wire lead and the surface so they achieve the same temperature.

Add Tip Ask Question

Step 10: Feeding the Solder

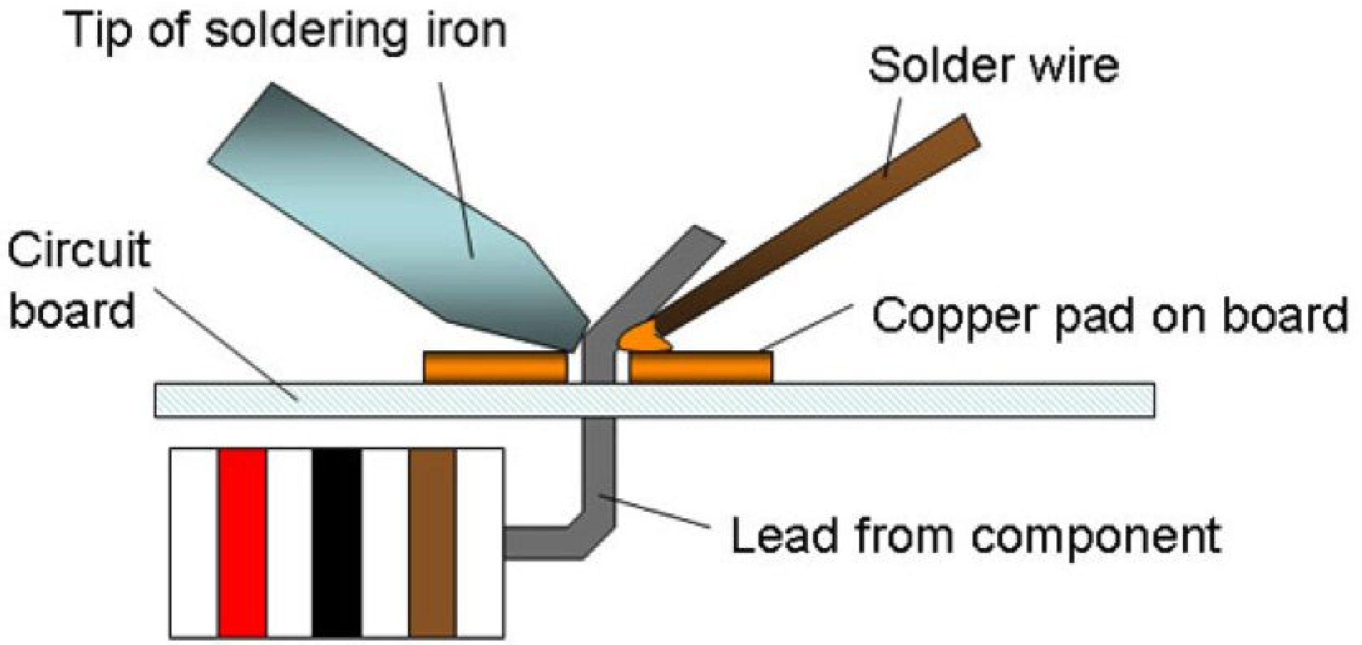
Tip of soldering iron
Solder wire
Circuit board
Copper pad on board
Lead from component

(https://cdn.instructables.com/FIW/HS90/IPZ4OGXV/FIWHS90IPZ4OGXV.LARGE.jpg)

Touch the solder to the side of the connection opposite the soldering iron after heating the area for 2-3 seconds. Then, let the solder flow only until the connection is covered.

Add Tip Ask Question

Step 11: Removing the Solder

Remove the solder first, then the iron otherwise your solder will get stuck to your connection point without the appropriate heat. Make sure the joint remains stationary while it cools.

Add Tip Ask Question

Step 12: Check to See If the Connection Is Solid

A smooth, shiny and volcano shaped joint is what you are looking for. If this isn't what you see, you'll need to reheat and feed in more solder.

Add Tip Ask Question

Step 13: Warnings

- DO NOT lay a soldering iron down on any surface. A soldering iron should either be placed on a stand or sealed with a heat resistant cap after every use.
- Soldering should be completed in a well ventilated area.
- Lead is present in most solders. Be sure to wash your hands after your project, or better yet wear gloves.
- Try not to inhale any of the solder smoke that comes off the solder. This is lead and can be dangerous.
- The tip of a soldering iron is very hot. Contact with the tip of a soldering iron would result in a nasty burn.
- Your soldering iron will perform better if kept clean. A damp sponge can be used to clean residue caused by flux material. A very small skim of flux should be applied to the iron after the cleaning.

THE BASICS OF HOW TO SOLDER

Soldering Tutorial for Beginners: Five Easy Steps

Table of Contents:

- Before you begin
- Step 1: Heat your iron to the appropriate temperature
- Step 2: Make your connection mechanically stable
- Step 3: Clean the tip of your iron
- Step 4: Apply heat and solder
- Step 5: Inspect the joint
- Common mistakes

Before you begin

You will need the following equipment:

A temperature-controlled soldering iron of 25W or more
A tip cleaner such as a brass sponge or a wet sponge

Solder (leaded is easier to work with than lead-free)
(Optional) Helping hands
Safety glasses

Note: molten solder is extremely hot! Irons heat up to around twice the boiling point of water, so be extremely careful. Wear safety glasses when soldering.

Step 1: Heat your iron to the appropriate temperature

Weller®
WES51
°Fx10

For through-hole components in a circuit board, you will want 600-700 degrees fahrenheit.

If you’re soldering a larger connections, with more copper that’s dissipating heat, then you may want a slightly higher temperature. With a temperature-adjustable iron you can see what works and adjust accordingly.

Step 2: Make your connection mechanically stable

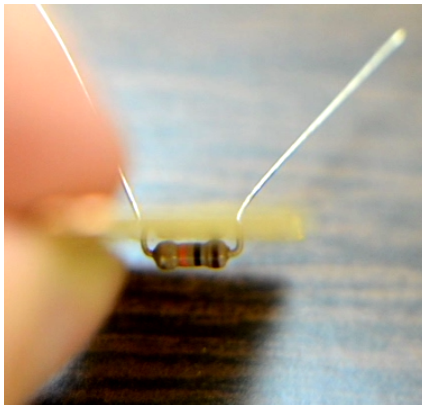

While your iron is heating up, you can work on making your components stay in place without your help. This is where your helping hands may come in handy. Use them to help you arrange your components to be stable without you needing to hold anything. If you’re soldering a through-hole component in a circuit board, you can bend the leads of the component into a V-shape so that it stays in place flush against the board. For components without long leads, you can use a piece of tape to hold them in place.

Step 3: Clean the tip of your iron

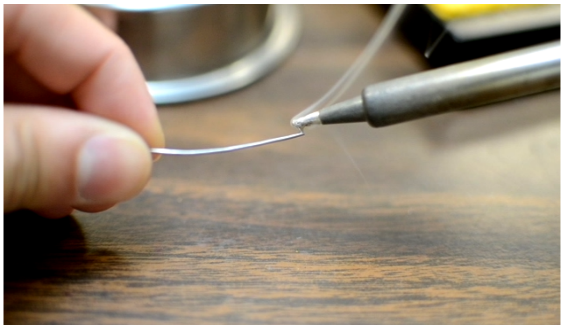

You can check whether your iron is close to its final temperature by touching your solder to the end of the iron and seeing if it melts. When you do this, you'll see some smoke come out. This is not the metal vaporizing, but rather the flux inside of it is burning. Flux is a mild acid that eats away at the oxidation layer that forms on the surface of hot metals. Without flux in the solder to clean your surfaces, you would have a very hard time getting the solder to wet.

Clean your iron using a brass sponge or a wet sponge to remove any oxidation or excess solder from the tip. Remember to always keep your tip clean while you're soldering. A dirty tip is covered in metal oxides that transfer heat badly and can result in poorly soldered joints

Step 4: Apply heat and solder

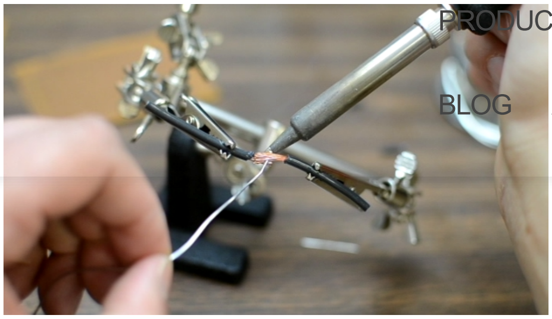

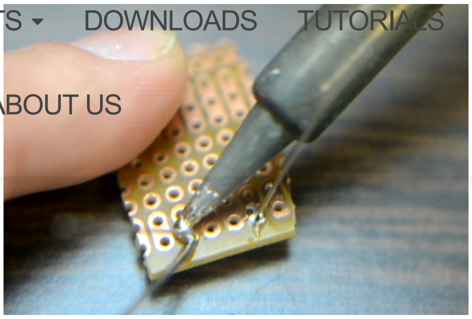

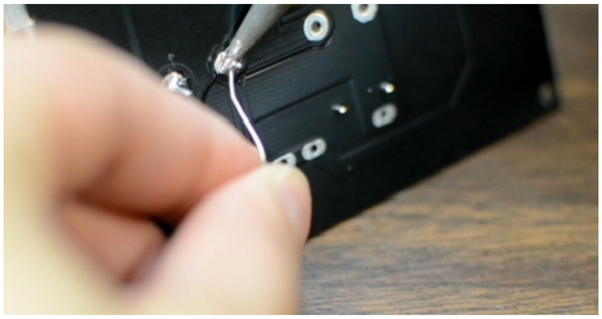

Touch the flat part of your iron's tip to one side of the joint, while feeding solder from the other side of the joint. The idea is to heat up the entire joint to the melting temperature of solder, so when you touch the solder to the surfaces it melts right on. The only way you can ensure a good electrical connection is by letting the solder flow over the components you're joining. Don't apply the solder directly to the tip of the iron. For through-hole components, hold the iron against both the lead and the pad.

Small joints can heat up within a few seconds, but larger joints can take a minute or two. Keep the joint steady for a few seconds after removing the iron, as the solder needs time to cool. Remember that the joint is hot now! don't touch it with your fingers right away.

Step 5: Inspect the joint

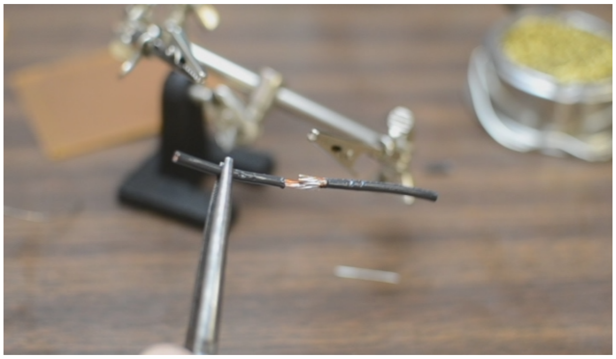

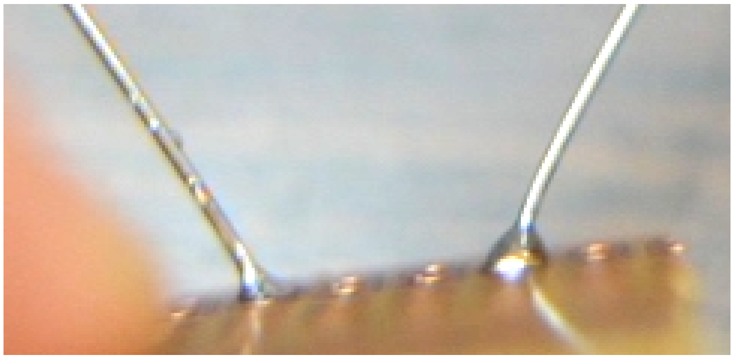

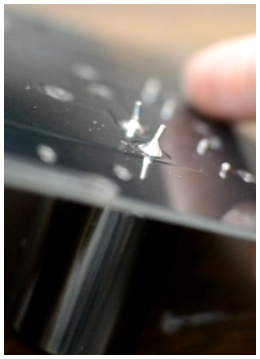

The solder should have wetted to both surfaces. On pads of a printed circuit board, the solder should be covering the entire pad, and it should make a volcano shape between the pad and the lead, indicating good adhesion. Once the joint is perfect, clip the leads of through-hole components.

Common mistakes, or what NOT to do

Now that we’ve gone over what to do, let me tell you what NOT to do.

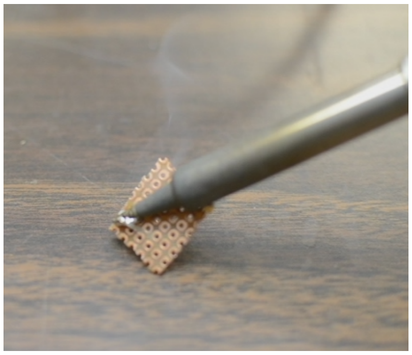

Don’t put a blob of solder on the iron and then try to transfer it to your joint. This burns away all of the flux in the solder, which prevents the flux from cleaning the metal surfaces. Furthermore, it’s easy to transfer the solder blob away from the iron without actually making it wet to the joint. I guarantee you that this will result in an awful electrical connection, so don’t do it.

Basic Tutorials

Soldering Basics
Desoldering Tutorial
Tesla Coil FAQ

tinyTesla Assembly Tutorials

tinyTesla Manual
Heat Sink Hardware Installation
Main Board's Logic Component Installation
Low Voltage Driver Board Test
Main Board's Power Component Installation
IGBT Installation

oneTeslaTS Assembly Tutorials

oneTeslaTS Manual
Endcap and Secondary Assembly Tutorial

Advanced Tutorials

How a Tesla Coil Works
(DR)SSTC Nonidealities and Losses
Thermal Analysis
DRSSTC Driver Guide
DRSSTC Bridge Design
Gate Drive Design
Resonator Design
Tesla Coil Tuning

SIGN UP FOR UPDATES

Enter your email for exclusive updates from oneTesla about new products, promotions, and other announcements

SUBSCRIBE

Limit of Liability

At oneTesla, we have a passion for making learning about hands-on electronics fun. Use our kits for educational purposes only. The operational reliability of the kits is dependent upon how well the kit is assembled, something that only the kit builder has control over. Build oneTesla kits at your own risk.

Site Links

- Contact Us
- Order FAQ
- oneTesla In the Press

One Tesla: DIY Sin…
4,519 likes

Like Page
Shop Now

ACCREDITED BUSINESS

About Us | Privacy Policy | Terms and Conditions | Kickstarter Terms and Conditions | Site Map | Search Terms | Advanced Search |

Makerspaces .com

How To Solder: A Beginner's Guide

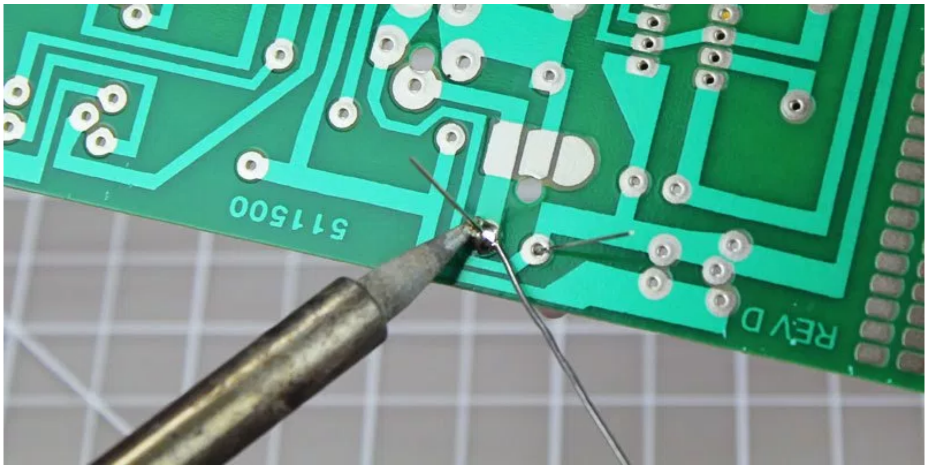

Learning how to solder w/ proper soldering techniques is a fundamental skill every maker should master. In this tutorial, we outline the basics of soldering irons, soldering stations, types of solder, desoldering and safety tips. Whether you're building a robot or working with Arduino, knowing how to solder will come in handy.

FREE EBOOK (PDF) – Learn To Solder Guide (17 pages)

What Is Soldering?

If you were to take apart any electronic device that contains a circuit board, you'll see the components are attached using soldering techniques. Soldering is the process of joining two or more electronic parts together by melting solder around the connection. Solder is a using a desoldering tool as described below.

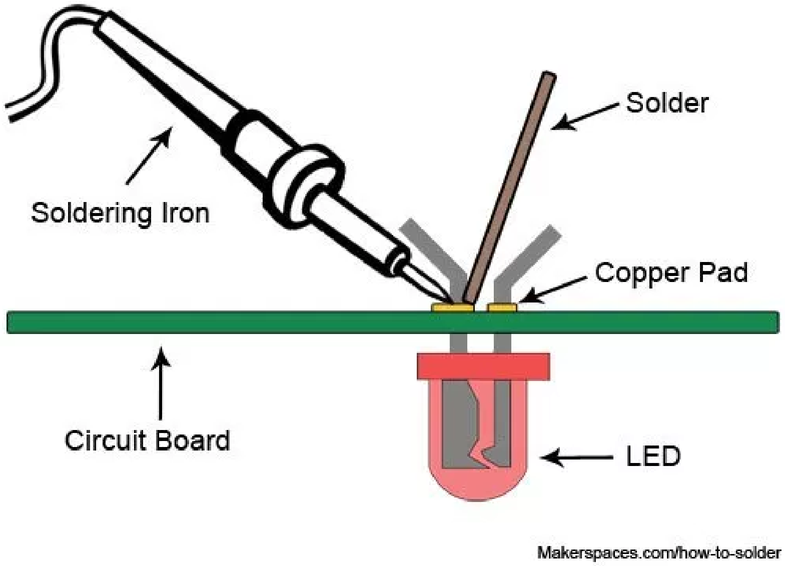
Solder
Soldering Iron
Copper Pad
Circuit Board
LED
Makerspaces.com/how-to-solder

Soldering Tools

The good thing about learning how to solder is the fact that you don't need a lot to get started. Below we'll outline the basic tools and materials you will need for most of your soldering projects.

Soldering Iron

A soldering iron is a hand tool that plugs into a standard 120v AC outlet and heats up in order to melt solder around electrical connections. This is one of the most important tools used in soldering and it can come in a few variations such as pen or gun form. For beginners, it's recommended that you use the pen style soldering iron in type of soldering iron because it can heat up to 896′ F which is extremely hot.

Soldering Station

A soldering station is a more advanced version of the basic standalone soldering pen. If you are going to be doing a lot of soldering, these are great to have as they offer more flexibility and control. The main benefit of a soldering station is the ability to precisely adjust the temperature of the soldering iron which is great for a range of projects. These stations can also create a safer workspace as some include advanced temperature sensors, alert settings and even password protection for safety.

Soldering Iron Tips

At the end of most soldering irons is an interchangeable part known as a soldering tip. There are many variations of this tip and they come in a wide variety of shapes and sizes. Each tip is used for a specific purpose and offers a distinct advantage over another. The most common tips you will use in electronics projects are the conical tip and the chisel tip.

Conical Tip – Used in precision electronics soldering because of the fine tip. Because of its pointed end, it's able to deliver heat to smaller areas without affecting its surroundings.

Chisel Tip – This tip is well-suited to soldering wires or other larger components because of its broad flat tip.

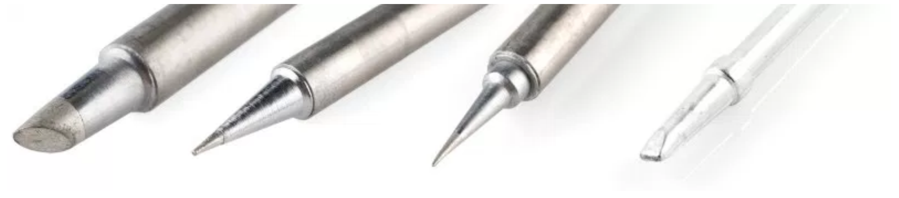

Image Credit – Sparkfun.com

Brass or Conventional Sponge

it did when it was new. You could use a conventional wet sponge but this tends to shorten the lifespan of the tip due to expansion and contraction. Also, a wet sponge will drop the temperature of the tip temporarily when wiped. A better alternative is to use a brass sponge as shown on the left.

Soldering Iron Stand

A soldering iron stand is very basic but very useful and handy to have. This stand helps prevent the hot iron tip from coming in contact with flammable materials or causing accidental injury to your hand. Most soldering stations come with this built in and also include a sponge or brass sponge for cleaning the tip.

Solder

Solder is a metal alloy material that is melted to create a permanent bond between electrical parts. It comes in both lead and lead-free variations with diameters of .032″ and .062″ being the most common. Inside the solder core is a material known as flux which helps improve electrical contact and its mechanical strength.

For electronics soldering, the most commonly used type is lead-free rosin core solder. This type of solder is usually made up of a Tin/Copper alloy. You can also use leaded 60/40 (60% tin, 40% lead) rosin core solder but it’s becoming less popular due to health concerns. If you do use lead solder, make sure you have proper ventilation and that you wash your hands after use.

Net Wt.: 8 oz. (226.7g)
64-009
60/40 ROSIN-CORE SOLDER
LF-99 Lead-Free Solder
99.3% tin 0.7% copper, 4.84g/
5 ft., 0.031" Dia. Rosin Core
WARNING: This product contains or produces a chemical known to the state of

When buying solder, make sure NOT to use acid core solder as this will damage your circuits and components. Acid core solder is sold at home improvement stores and is mainly used for plumbing and metal working.

As mentioned earlier, solder does come in a few different diameters. The thicker diameter solder (.062″) is good for soldering larger joints more quickly but it can make soldering smaller joints difficult. For this reason, it’s always a good idea to have both sizes on hand for your different projects.

Helping Hand (Third Hand)

A helping hand is a device that has 2 or more alligator clips and sometimes a magnifying glass/light attached. This clips will assist you by holding the items you are trying to solder while you use the soldering iron and solder. A very helpful tool to have in your makerspace.

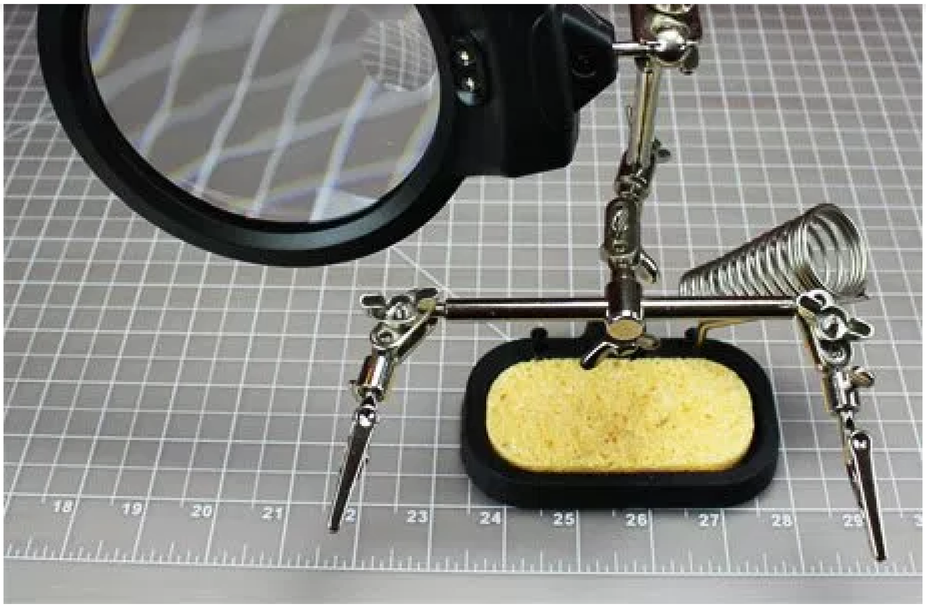

Soldering Safety

Now that you know what tools and materials are required, it’s time to briefly discuss ways of staying safe while soldering.

Soldering irons can reach temperatures of 800′ F so it’s very important to know where your iron is at all times. We always recommend you use a soldering iron stand to help prevent accidental burns or damage.

Make sure you are soldering in a well ventilated area. When solder is heated, there are fumes released that are harmful to your eyes and lungs. It's recommended to use a fume extractor which is a fan with a charcoal filter that absorbs the harmful solder smoke.

It's always a good idea to wear protective eye wear in case of accidental splashes of hot solder. Lastly, make sure to wash your hands when done soldering especially if using lead solder.

Tinning The Tip

Before you can start soldering, you need to prep your soldering iron by tinning the tip with solder. This process will help improve the heat transfer from the iron to the item you're soldering. Tinning will also help to protect the tip and reduce wear.

Step 1: Begin by making sure the tip is attached to the iron and screwed tightly in place.

Step 3: Wipe the tip of the soldering iron on a damp wet sponge to clean it. Wait a few seconds to let the tip heat up again before proceeding to step 4.

Step 4: Hold the soldering iron in one hand and solder in the other. Touch the solder to the tip of the iron and make sure the solder flows evenly around the tip.

You should tin the tip of your iron before and after each soldering session to extend its life. Eventually, every tip will wear out and will need replacing when it becomes rough or pitted.

How To Solder

To better explain how to solder, we’re going to demonstrate it with a real world application. In this example, we’re going to solder an LED to a circuit board.

Step 1: **Mount The Component** – Begin by inserting the leads of the LED into the holes of the circuit board. Flip the board over and bend the leads outward at a 45′ angle. This will help the component make a better connection with the copper pad and prevent it from falling out while soldering.

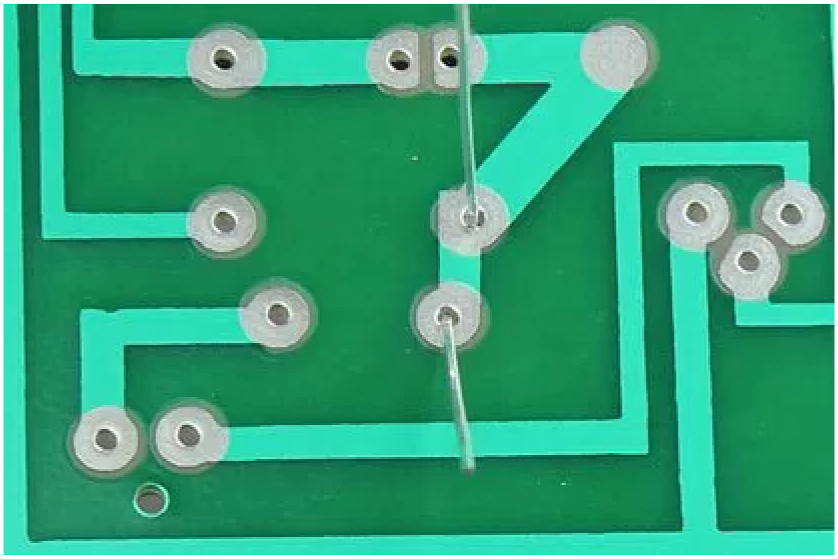

resistor lead at the same time. You need to hold the soldering iron in place for 3-4 seconds in order to heat the pad and the lead.

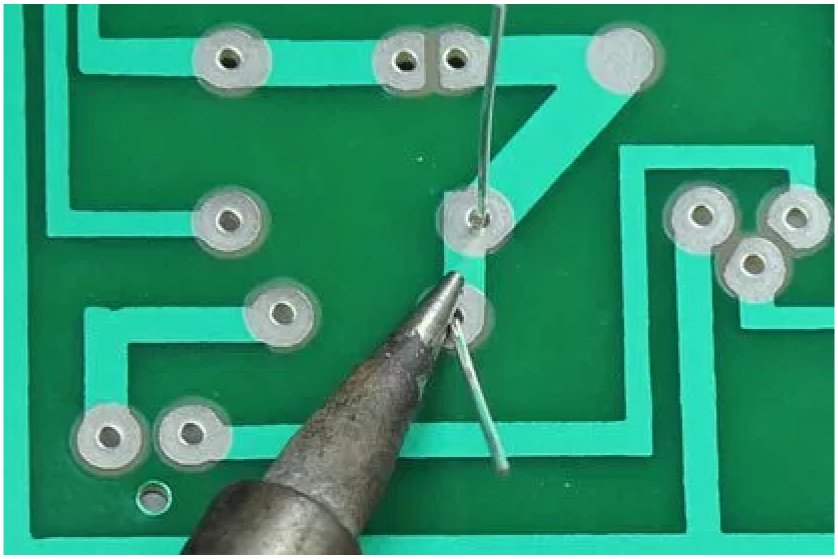

Step 3: **Apply Solder To Joint** – Continue holding the soldering iron on the copper pad and the lead and touch your solder to the joint. **IMPORTANT** – Don't touch the solder directly to the tip of the iron. You want the joint to be hot enough to melt the solder when it's touched. If the joint is too cold, it will form a bad connection.

Step 4: **Snip The Leads** – Remove the soldering iron and let the solder cool down naturally. Don't blow on the solder as this will cause a bad joint. Once cool, you can snip the extra wire from leads.

A proper solder joint is smooth, shiny and looks like a volcano or cone shape. You want just enough solder to cover the entire joint but not too much so it becomes a ball or spills to a nearby lead or joint.

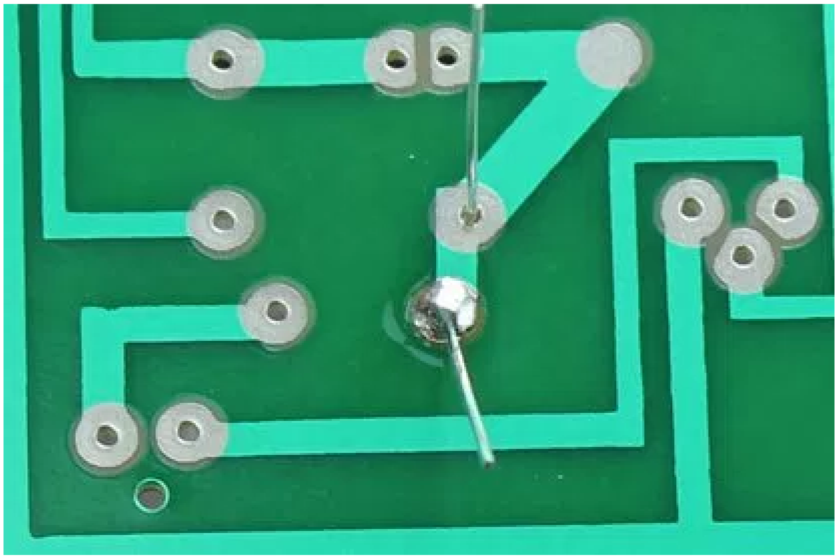

How To Solder Wires

Begin by removing the insulation from the ends of both wires you are soldering together. If the wire is stranded, twist the strands together with your fingers.

Make sure your soldering iron is fully heated and touch the tip to the end of one of the wires. Hold it on the wire for 3-4 seconds.

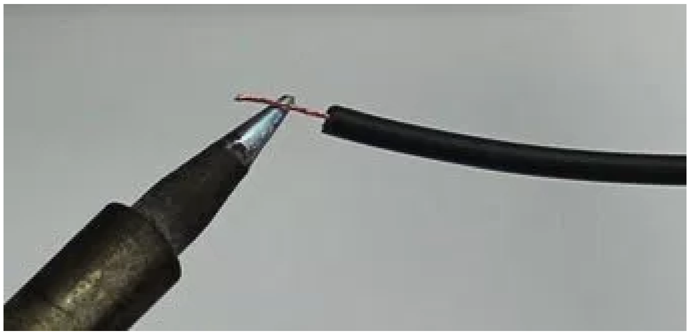

Keep the iron in place and touch the solder to the wire until it's fully coated. Repeat this process on the other wire.

Hold the two tinned wires on top of each other and touch the soldering iron to both wires. This process should melt the solder and coat both wires evenly.

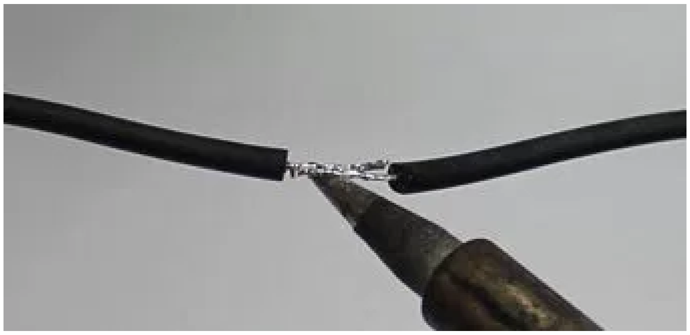

Remove the soldering iron and wait a few seconds to let the soldered connection cool and harden. Use heat shrink to cover the connection.

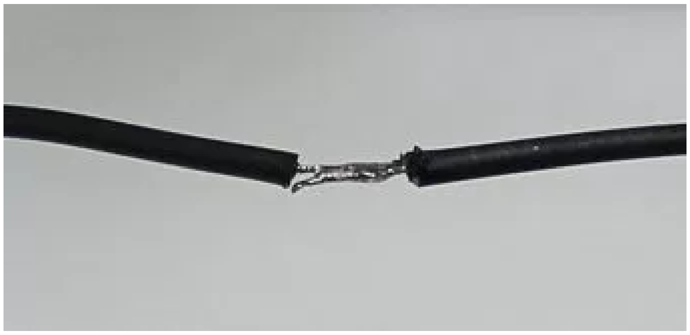

Desoldering

The good thing about using solder is the fact that it can be removed easily in a technique known as desoldering.

To desolder a joint, you will need solder wick which is also known as desoldering braid.

Desoldering Braid
RoHS
6402090
RadioShack

Step 1 – Place a piece of the desoldering braid on top of the joint/solder you want removed.

Step 2 – Heat your soldering iron and touch the tip to the top of the braid. This will heat the solder below which will then be absorbed into the desoldering braid. You can now remove the braid to see the solder has been extracted and removed. Be careful touching the braid when you are heating it because it will get hot.

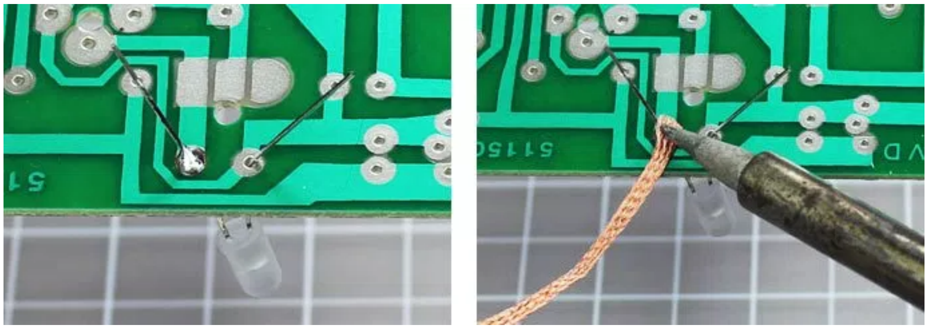

solder with a press of a button.

To use, press the plunger down at the end of the solder sucker. Heat the joint with your soldering iron and place the tip of the solder sucker over the hot solder. Press the release button to suck up the liquid solder. In order to empty the solder sucker, press down on the plunger.

Soldering & Electronics

- Digikey
- Sparkfun
- Adafruit
- Jameco
- Velleman
- Mouser

FREE – Learn To Solder Guide (17 pages)

Makerspaces.com

Makerspaces.com was built to help schools and libraries start and run

Makerspaces
.com

acquire the skills needed for the 21st century.

adafruit learning system

Adafruit Guide To Excellent Soldering

Created by Bill Earl

Last updated on 2014-06-28 08:30:54 PM EDT

Guide Contents

Guide Contents 2

Tools 4

Building a Soldering Toolkit 4

Choosing a Soldering Iron 4

- Basic Irons 5
- Better Irons 5
- Best Irons 5

Irons to avoid 6

- For emergencies only: 6
- Not for circuit board use: 6

Essential Tools and Supplies: 6

- Stand 7
- Solder 7
- Diagonal Cutters 7

Other Handy Tools and Supplies 7

- Vise 8
- Third Hand 8
- Solder Sucker 9
- Solder Wick 9

Preparation 10

- Heat the Iron 10
- Clean the Iron 10
- Tin the Tip 11
- Make sure that the joint is clean 11
- Immobilize the Joint 11
- Steady the Board 12

Making a good solder joint 13

- Heat the joint 13
- Apply the solder 14
- Let It Flow 14

Let It Cool 14
Trim the Lead 14
Congratulations! 15
Problems? 16
Surface Mount Components 17
Immobilize the Joint 17
Heat the Joint 17
Apply the Solder 18
Let it Flow 18
Let it Cool 18
Problems? 18
Common Soldering Problems 20
The Ideal Solder Joint 20
Disturbed Joint 21
Cold Joint 22
Overheated Joint 24
Insufficient Wetting (Pad) 25
Insufficient Wetting (Pin) 27
Insufficient Wetting 28
(Surface Mount) 28
Solder Starved 29
Too Much Solder 30
Untrimmed Leads 31
Solder Bridge 32
Lifted Pad 33
Repairing a Lifted Pad 34
Stray Solder Spatters 34
All of the Above! 35

Tools

Building a Soldering Toolkit

If you are just getting started in Electronics, Ladyada's Electronics Toolkit (http://adafru.it/136) (pictured above) is a great kit full of quality tools - including everything you need to make great solder joints. If you would rather build your toolkit piece-by-piece, read on:

Choosing a Soldering Iron

There are many types of soldering irons. For most Adafruit kits and projects, you will want a pencil-style soldering iron with 25 watts or more.

An under-powered iron is a poor investment. It will end up costing you more in ruined kits and damaged components.

- It will take longer to heat the joint, allowing heat to spread to the component being soldered - potentially overheating and damaging the component.
- Longer heating times will also give more time for oxides to form on the surfaces being soldered. This will prevent the solder from flowing and result in a poor joint.
- Longer recovery times between joints can result in frustration, 'cold joints' or both.

You don't need to spend a fortune to get a good iron. Advanced features such as temperature control and interchangeable tips are nice to have, but not essential for hobbiest-level work.

Basic Irons

There are many basic pencil style irons that are suitable for hobbiest use. But you will need one that is capable of heating the joints quickly enough. Choose an iron with 25 watts at a minimum.

Better Irons

An adjustable temperature iron with a little more power will give you a bit more control and allow you to work faster. The Adjustable 30W 110v Soldering Iron (http://adafru.it/180) in the store is an excellent choice.

This iron is also available as part of Ladyada's Electronics Toolkit (http://adafru.it/136), which contains many other essential soldering tools.

Best Irons

A professional-style temperature-controlled iron with interchangeable tips and 50 watts or more of power is a joy to work with. Feedback control keeps the tip temperature at precisely the level you set. The extra watts speed recovery time so that you can work faster. Interchangeable tips let you select the ideal tip shape for specialized work.

The 65 watt Hakko FX-888 (http://adafru.it/303) is an excellent professional quality soldering iron. The Weller WES51 or WESD51 are also excellent choices for serious electronics work.

Irons to avoid

In addition to underpowered irons, there are several types of irons to avoid for general circuit-board work.

For emergencies only:

These irons are handy for occasions when you have no place to plug in a regular soldering iron. But they are not the best choice for a primary soldering tool:

- **Butane Powered Irons** have plenty of power but are difficult to control.
- **Battery Powered Irons** are generally underpowered for most work.

BERNZOMATIC
CRAFTSMAN

Not for circuit board use:

These tools are not suitable for circuit board work:

- **Torches** of any kind are not suitable for electronics work and will damage your circuit boards.
- **Soldering Guns** are OK for working with heavy gauge wires, but don't have the precision necessary for soldering delicate electronics components.
- ***Cold-Heat™*** Irons inject current into the joint to heat the tip. This current can be damaging to sensitive electonic components. Avoid these irons for electronics work.

Essential Tools and Supplies:

These tools are the bare-minimum essentials required for soldering:

Stand

If your soldering iron does not have a built-in stand, you will need a safe place to rest the hot iron between uses. A Soldering Iron Stand (http://adafru.it/150) will keep your iron from rolling around and protect both you and your work surface from burns.

Most stand holders come with a sponge and tray for cleaning your soldering iron.

ROSIN CORE
GUARANTEED VIRGIN METALS
GUARANTEED ALLOY
60/40
4 OZ.
.031 DIA

Solder

Standard 60/40 lead/tin Rosin Core Solder (http://adafru.it/145) is the easiest type to work with.

170M Xcelite® U.S.A.

Diagonal Cutters

You will also need a pair of Diagonal Cutters (http://adafru.it/152) for trimming component leads after soldering.

Other Handy Tools and Supplies

These are some other tools and supplies you might find useful when working on soldering projects.

Vise

A vise holds your work steady as you solder. This is important for both safety and sound joints. The Panavise Jr (http://adafru.it/151) is an ideal size for most Adafruit kits and projects.

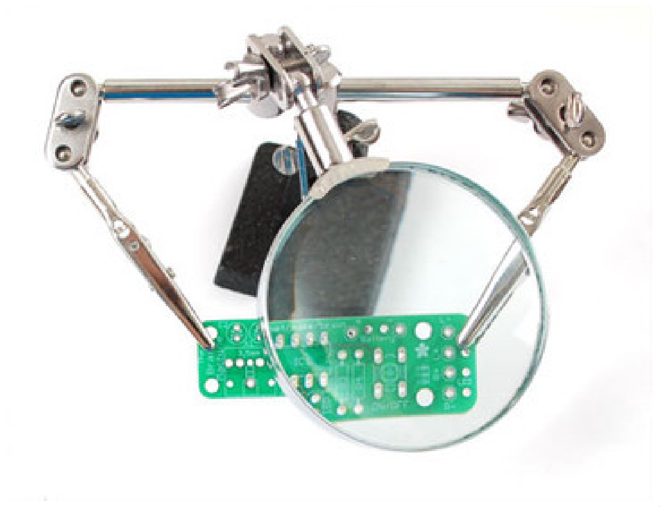

Third Hand

A Helping Third Hand (http://adafru.it/291) Tool is a good for smaller boards, or to hold things in place while terminating or splicing wires.

Solder Sucker

A Solder Sucker (http://adafru.it/148) is a very helpful tools for removing excess solder or when you need to de-solder a joint. As the name implies, this device literally sucks the solder out of the joint.

ROHS
35+YEARS SPIRIG PATENTED VACUUMISATION WICK MANUFACTURE
no clean
size AB (3)
1.5 m / 5 ft
DESOLDERING BRAID
3S-WICK®
part: LS/AB-1.5 m
TRESSE à DESSOUDER
LOTSAUGLITZE®
antistatic
swiss

Solder Wick

Solder Wick (http://adafru.it/149) is another way to clean excess solder from a joint. Unlike the solder sucker, the wick soaks up the molten solder.

Preparation

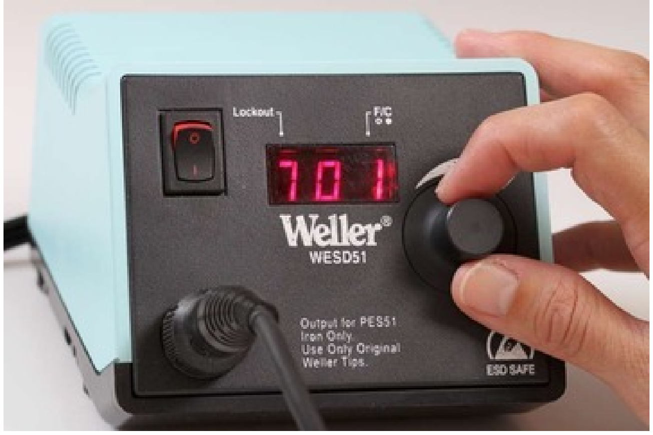
Lockout
F/C
701
Weller®
WESD51
Output for PES51
Iron Only.
Use Only Original
Weller Tips.
ESD SAFE

Heat the Iron

Plug an and/or turn on your soldering iron to warm up. If you are using a temperature controlled iron, set it to 700F/370C for 60/40 or 750F/400C for lead-free solder.

While the iron is heating dampen the sponge with a little bit of water.

Clean the Iron

Wipe the tip of the hot iron on the damp sponge to clean off any oxidation.

Do not use files or abrasives to clean the tip. It will damage the plating and ruin the tip.

Tin the Tip

Apply a small amount of solder to the tip and wipe again to tin the tip. You should have a thin, shiny layer of molten solder on the tip of your iron.

If the tip is badly oxidized and difficult to tin, it can usually be reconditioned with some tip-tinning paste.

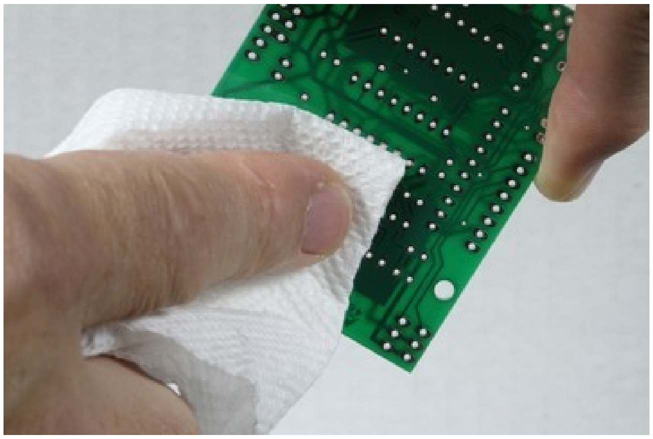

Make sure that the joint is clean

Dirt, oxidation and oily fingerprints can prevent the solder from wetting the solder-pad to create a solid joint. All Adafruit boards are plated to prevent oxidation, but if your board appears dirty from storage or handling, wipe it down with a little isopropyl alcohol.

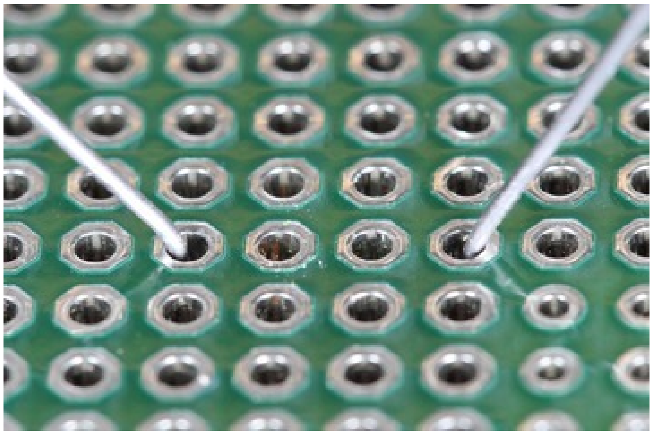

Immobilize the Joint

This is very important! The parts being joined must not move during the soldering process. If there is any movement as the molten solder is solidifying, you will end up with an unreliable 'cold joint'.

Most through-hole components can be immobilized by simply bending the leads on the solder-side of the hole.

Steady the Board

A vise is a good way to keep the board from moving around while you try to solder it.

Once the joint is clean and immobilized, you are ready to apply the solder.

Making a good solder joint

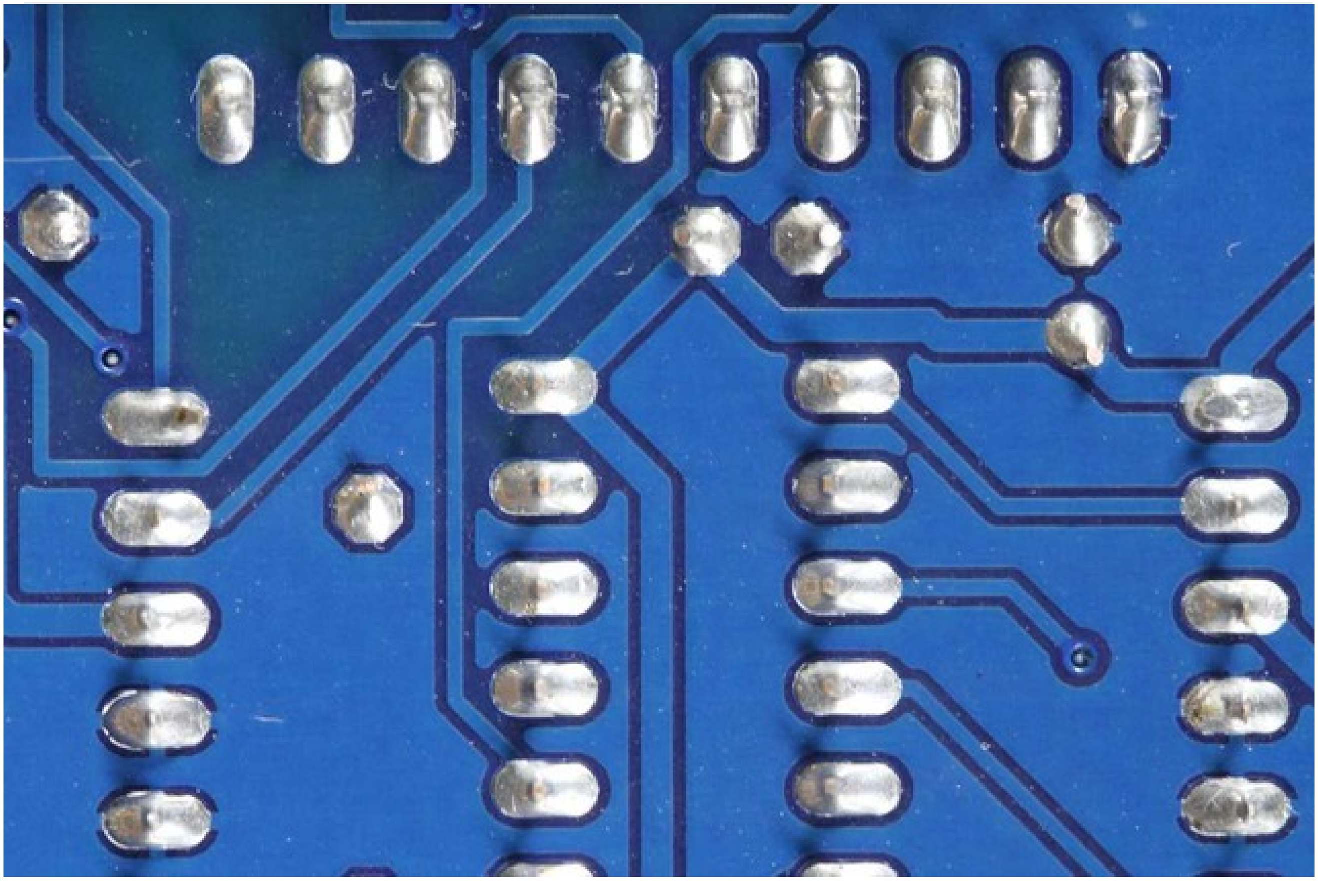

Once you have prepared the your tools and the joint to be soldered, making a good solder joint requires just a few simple steps.

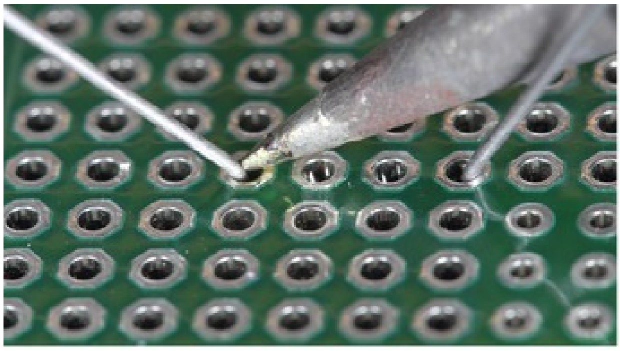

Heat the joint

Heat the joint with the tip of the iron. Be sure to heat both the solder pad and the component lead or pin. A small drop of solder on the tip will help to transfer the heat to the joint quickly.

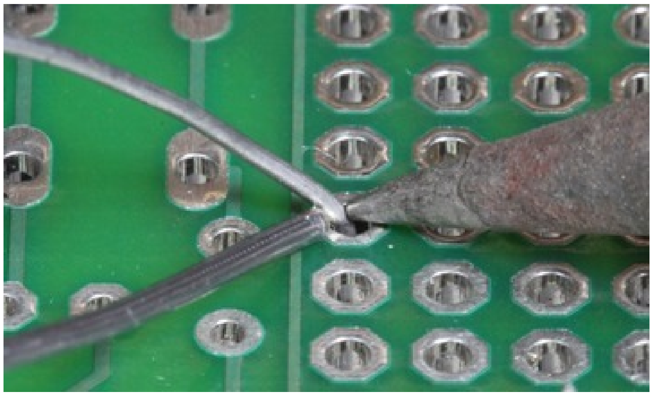

Apply the solder

Touch the end of the solder to the joint so that it contacts both the solder pad and the component lead or pin. It should melt and flow smoothly onto both the pin and the pad. If the solder does not flow, heat the joint for another second or two and try again.

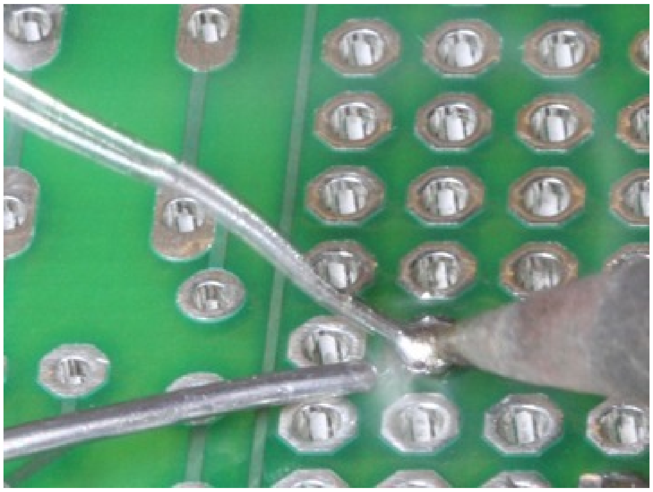

Let It Flow

Keep heating the solder and allow it to flow into the joint. It should fill the hole and flow smoothly onto both the solder pad and the pin or component lead.

Let It Cool

Once enough solder has been added to the joint and it has flowed well onto both the component lead and the solder pad, remove the iron from the joint and allow it to cool undisturbed.

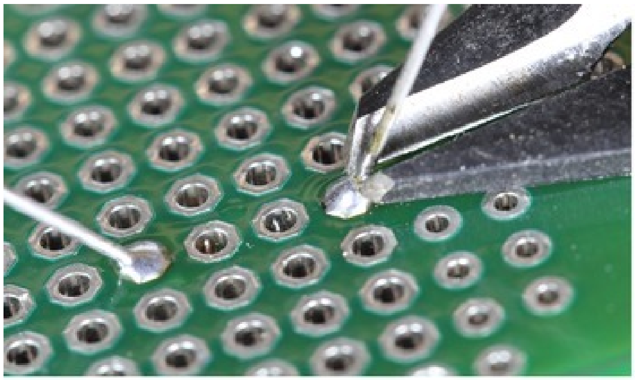

Trim the Lead

Use your diagonal cutters to trim the lead close to the board.

Note: This step applies only to components with wire leads. It is not necessary to trim the pins on Integrated circuit chips or sockets.

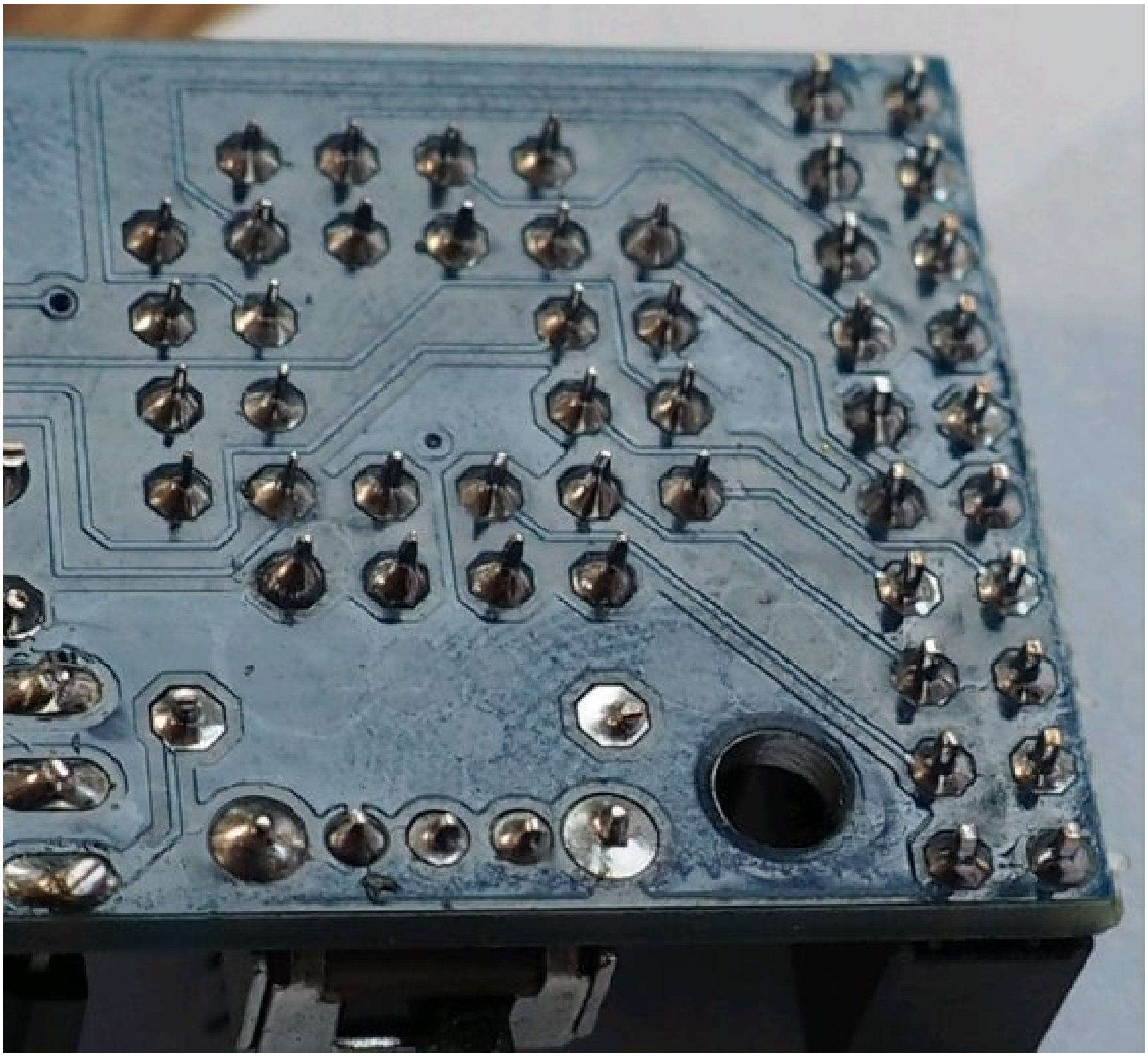

Congratulations!

Reward yourself with a Soldering Badge (http://adafru.it/465).

Problems?

The last page of this guide illustrates a number of common soldering problems with advice on prevention and repair.

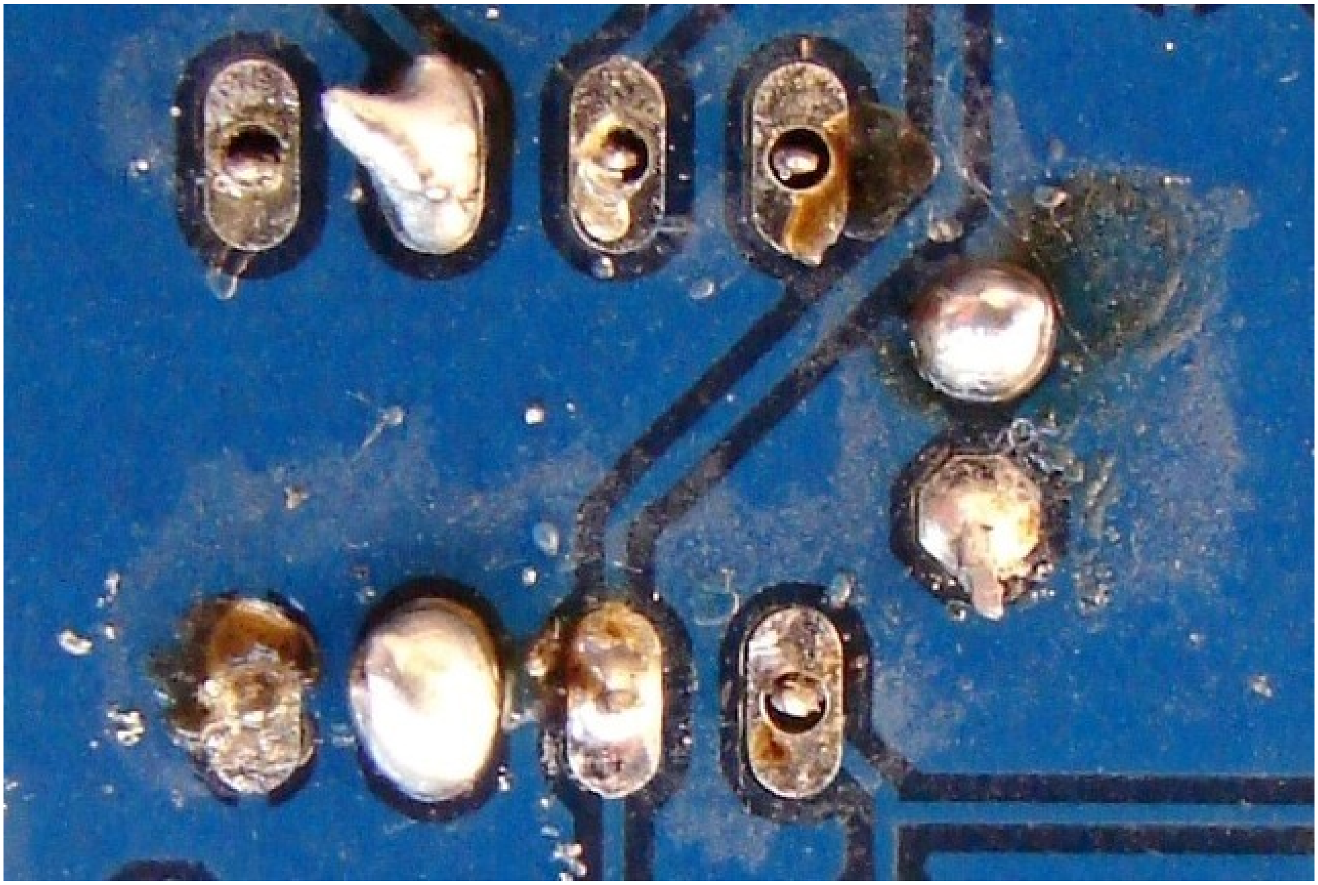

Surface Mount Components

The previous page showed how to make a good through-hole joint. But more and more components are only available in surface mount form these days. Not all surface mount packages are easily worked by hand, but there are plenty that can be managed with the same basic tools used for through-hole soldering.

Let's start with a surface-mount part common to several Adafruit kits: The SD Card Holder:

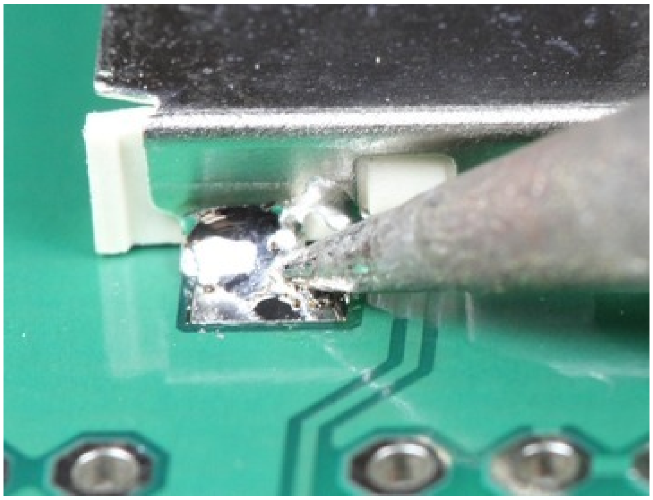

Immobilize the Joint

Unlike many surface mount components, immobilizing the SD card holder is relatively easy. There are small pegs on the back that fit into positioning holes in the board. Once it is in place, solder the four small corner tabs to make it permanent.

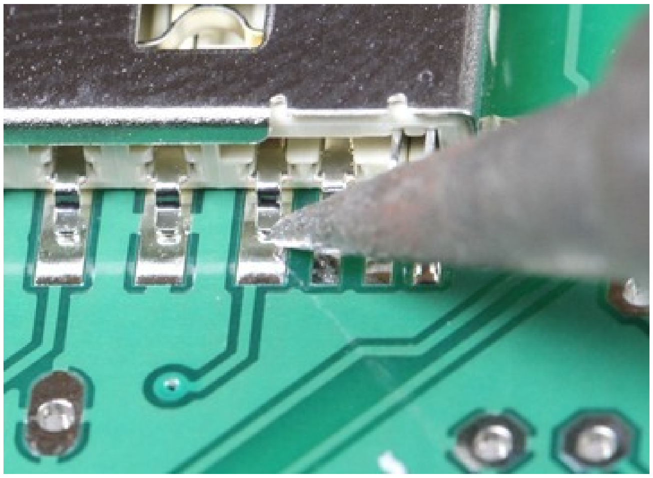

Heat the Joint

Start by putting the tip of the hot iron on the solder pad adjacent to the pin. The pad will take longer to heat, so we apply most of the heat to the pad to start.

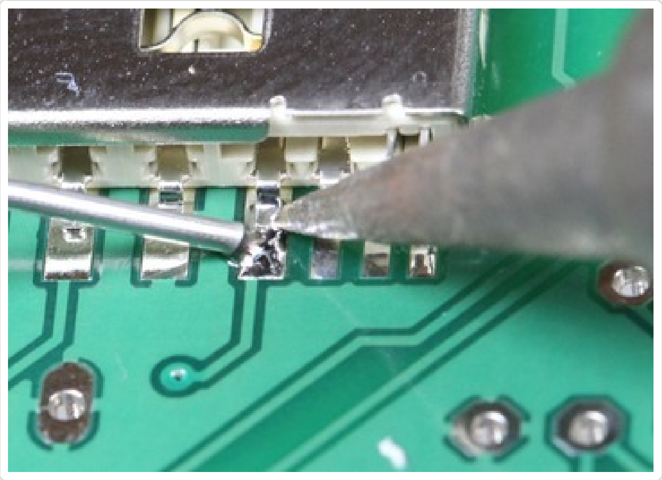

Apply the Solder

When the joint is hot, apply solder to the side opposite the iron. The solder should melt and start to flow into the joint.

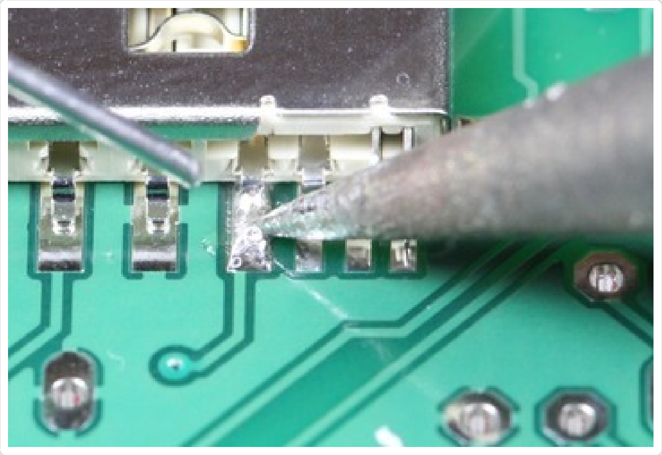

Let it Flow

Apply just enough solder to ensure a good joint, then keep the heat on while the solder wicks up between the pin and the pad to make a good electrical bond.

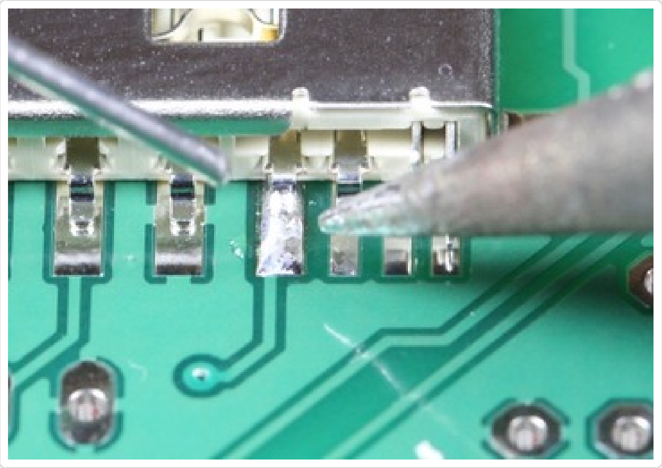

Let it Cool

Remove the iron and allow the joint to cool undisturbed.

Problems?

The last page of this guide illustrates a number of common soldering problems with advice on prevention and repair.

Common Soldering Problems

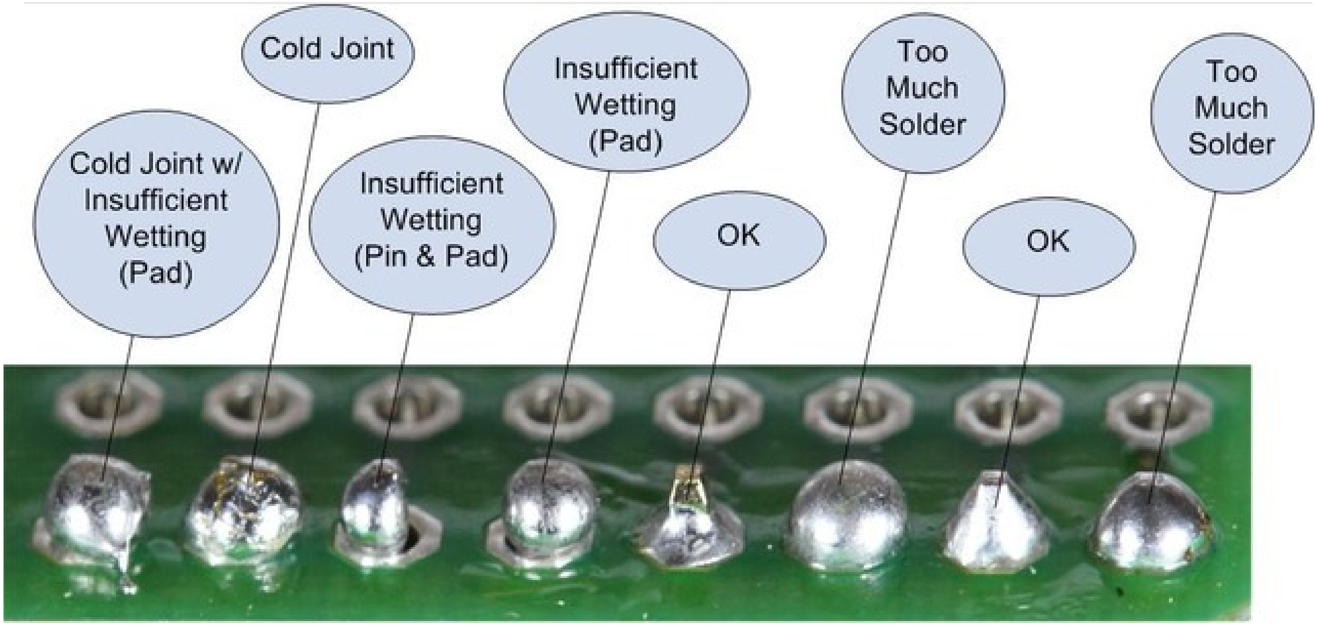
Cold Joint w/ Insufficient Wetting (Pad)
Cold Joint
Insufficient Wetting (Pin & Pad)
Insufficient Wetting (Pad)
OK
Too Much Solder
OK
Too Much Solder

The Ideal Solder Joint

The ideal solder joint for through-hole components should resemble the diagram below.

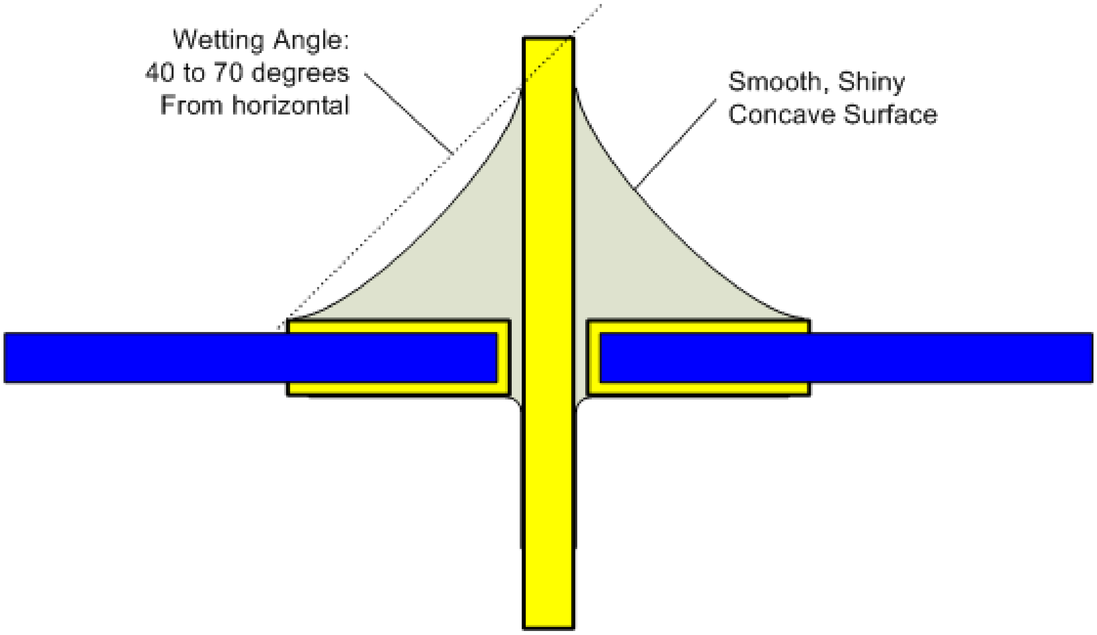
Wetting Angle:
40 to 70 degrees
From horizontal
Smooth, Shiny
Concave Surface

The photos that follow show some common soldering problems, with suggestions for repair and prevention:

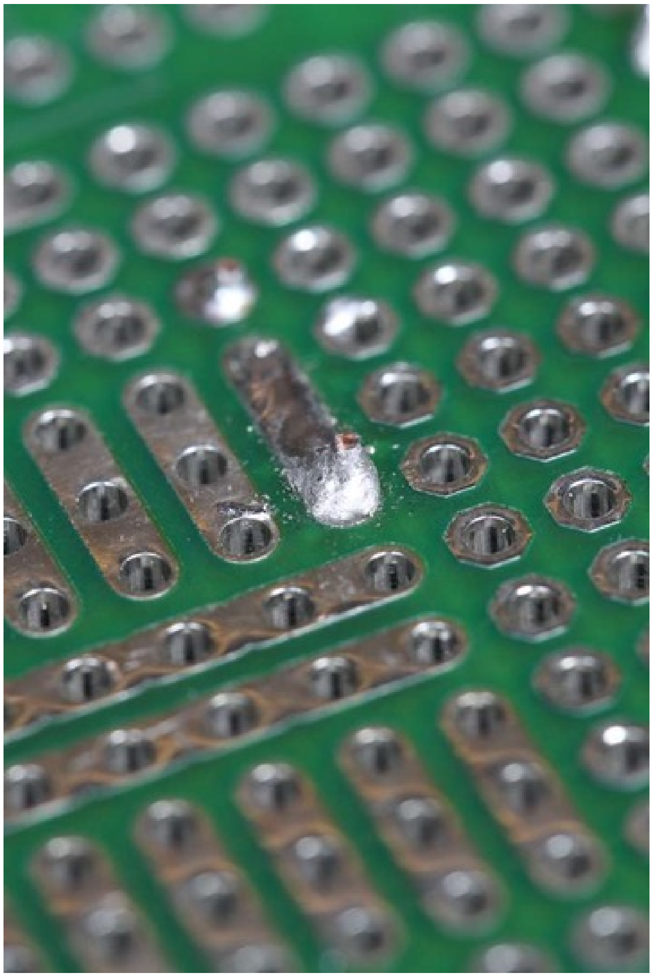

Disturbed Joint

A Disturbed joint is one that has been subjected to movement as the solder was solidifying. The surface of the joint may appear frosted, crystalline or rough.

Often called a 'Cold Joint'. They can look similar to a true cold joint, but the cause is different.

Repair: This joint can be repaired by reheating and allowing it to cool undisturbed.

Prevention: Proper preparation, including immobilizing the joint and stabilizing the work in a vise can prevent disturbed joints.

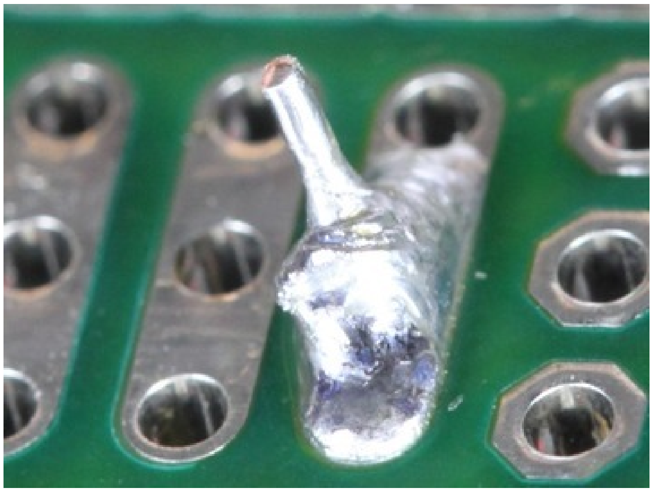

Cold Joint

A 'Cold Joint' is one where the solder did not melt completely. It is often characterized by a rough or lumpy surface. Cold joints are unreliable. The solder bond will be poor and the cracks may develop in the joint over time.

Repair: Cold joints can usually be repaired by simply re-heating the joint with a hot iron until the solder flows. Many cold joints (such as the one pictured) also suffer from too much solder. The excess solder can usually be drawn-off with the tip of the iron.

Prevention: A properly pre-heated soldering iron with sufficient power will help prevent cold joints.

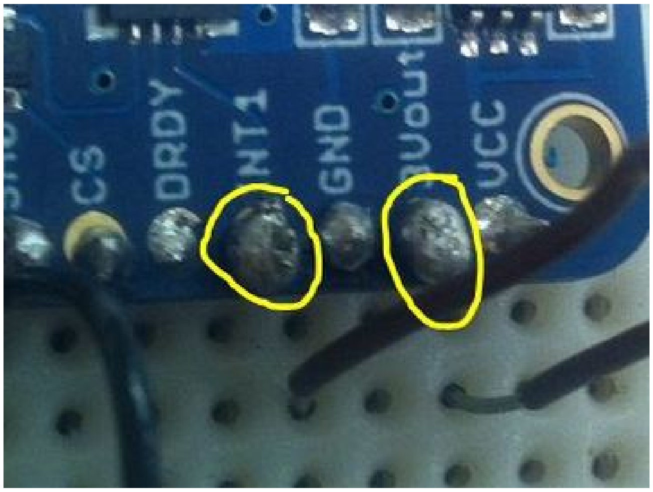

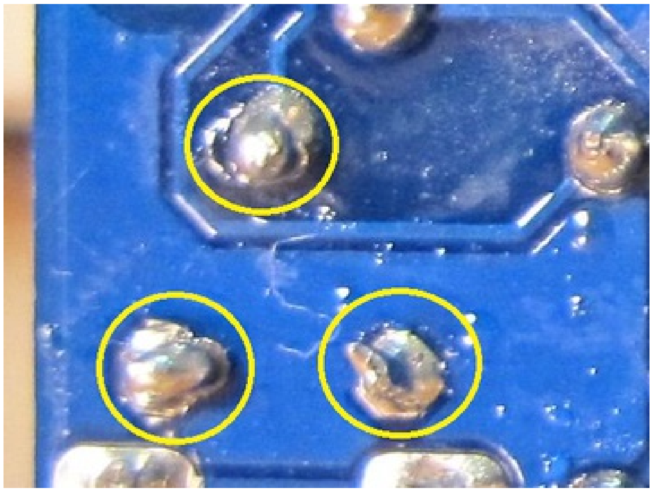

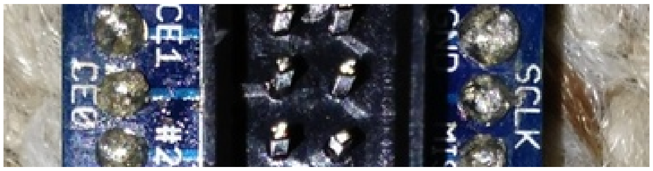

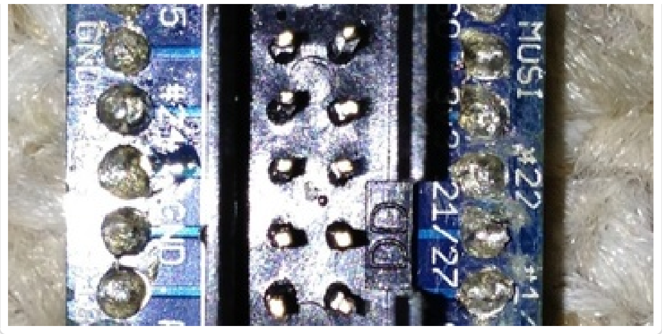
GND
#24
GND
MOSI
#22
21/27
#17
DD

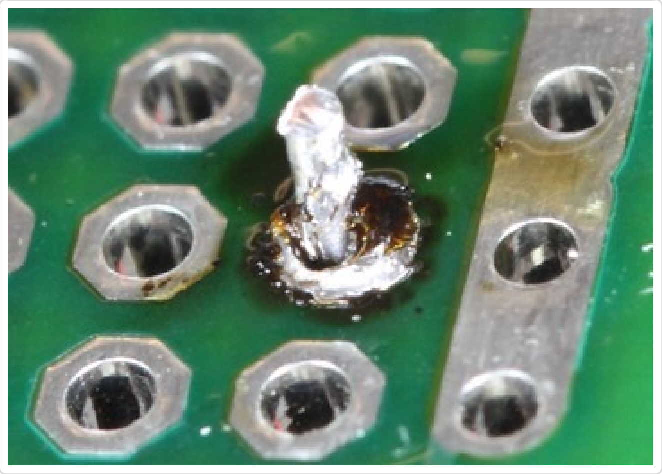

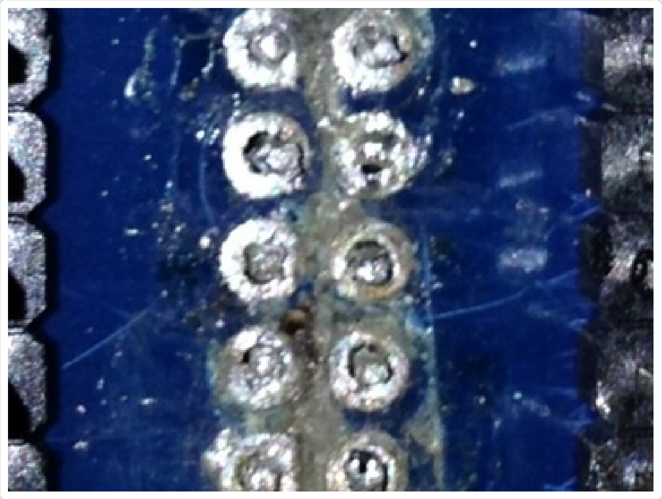

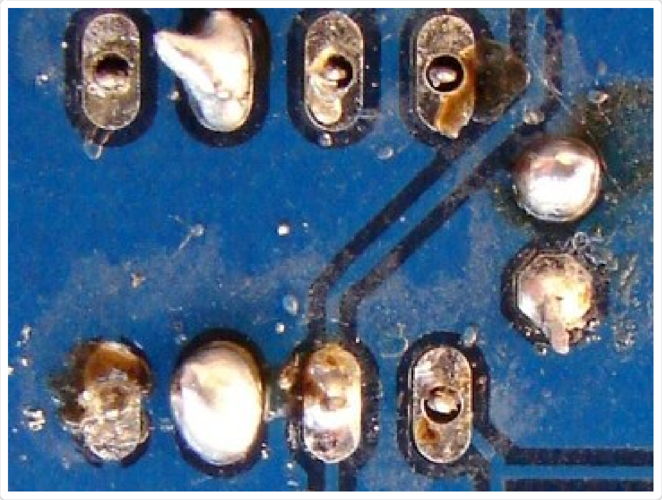

Overheated Joint

At the other extreme, we have the overheated joint. The solder has not yet flowed well and the residue of burnt flux will make fixing this joint difficult.

Repair: An overheated joint can usually be repaired after cleaning. Careful scraping with the tip of a knife, or little isopropyl alcohol & a toothbrush will remove the burnt flux.

Prevention: A clean, hot soldering iron, proper preparation and cleaning of the joint will help prevent overheated joints.

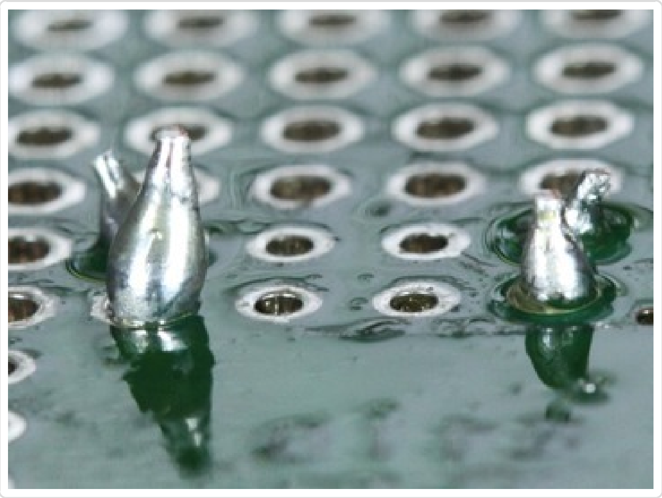

Insufficient Wetting (Pad)

These two joints both show signs of insufficient wetting of the solder pad. The solder has wetted the leads nicely, but it has not formed a good bond with the pad. This can be caused by a dirty circuit board, or by failing to apply heat to the pad as well as the pin.

Repair: This condition can usually be repaired by placing the tip of the hot iron at the base of the joint until the solder flows to cover the pad.

Prevention: Cleaning the board and even heating of both the pad and the pin will prevent this problem.

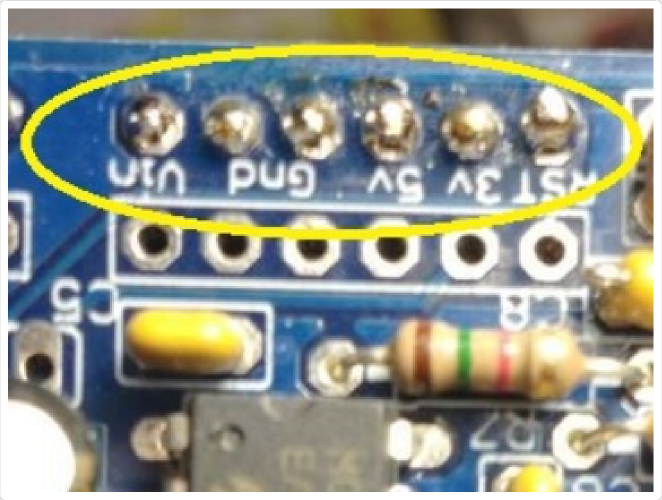

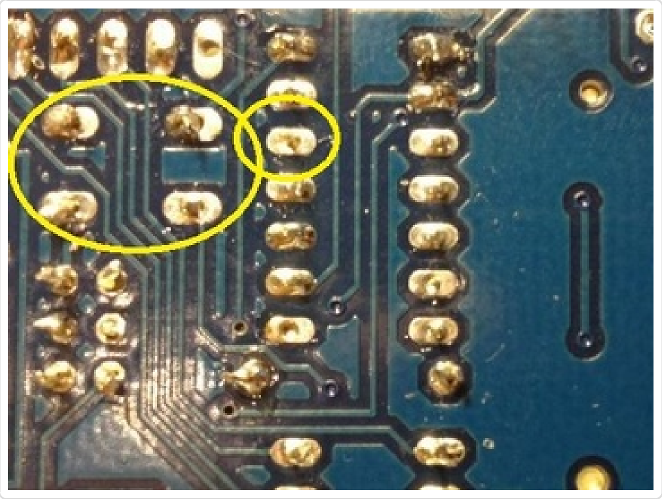

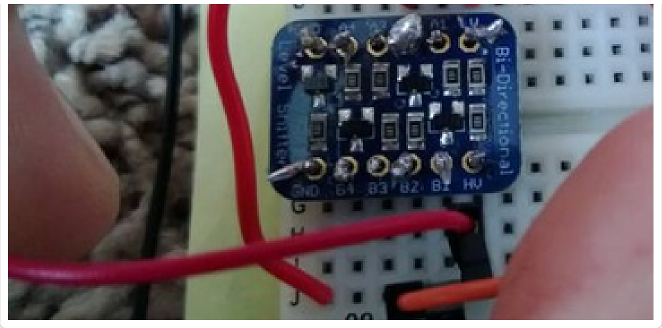
Bi-Directional
Level Shifter
GND B4 B3 B2 B1 HV

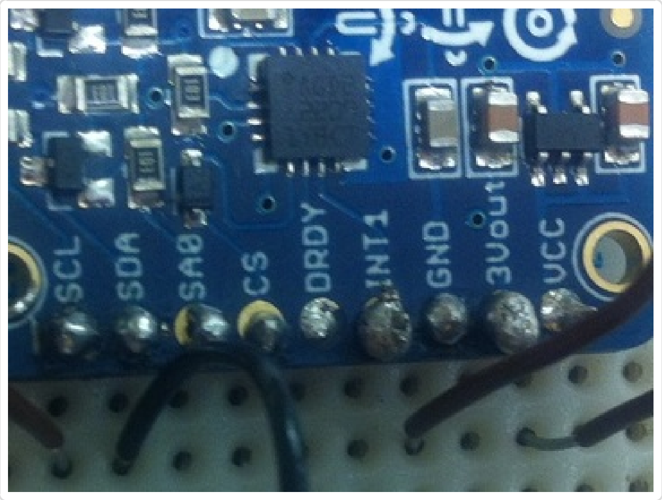
SCL
SDA
SA0
CS
DRDY
INT1
GND
3Vout
VCC

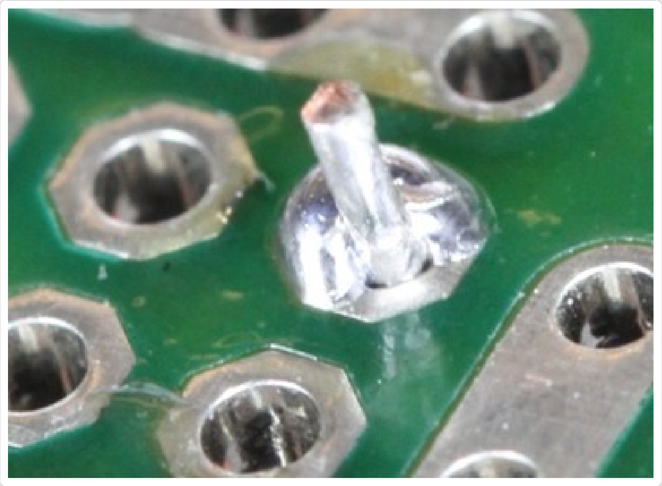

Insufficient Wetting (Pin)

This solder in this joint has not wetted the pin at all and has only partially wetted the pad. In this case, heat was not applied to the pin and the solder was not given adequate time to flow.

Repair: This joint can be repaired by re-heating and applying more solder. Be sure that the tip of the hot iron is touching both the pin and the pad.

Prevention: Even heating of both the pin and the pad will prevent this problem.

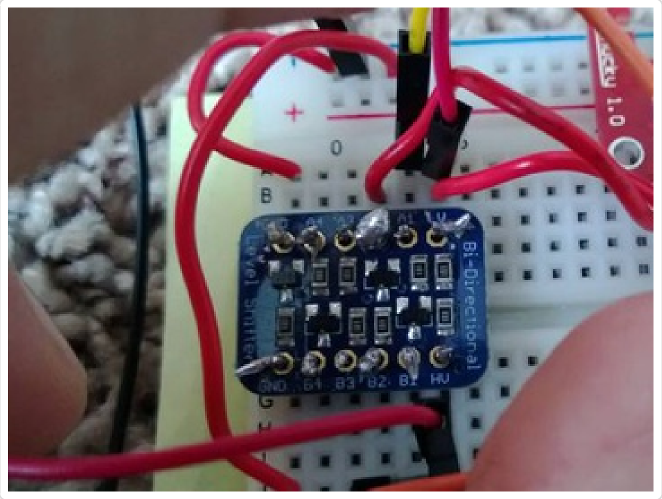

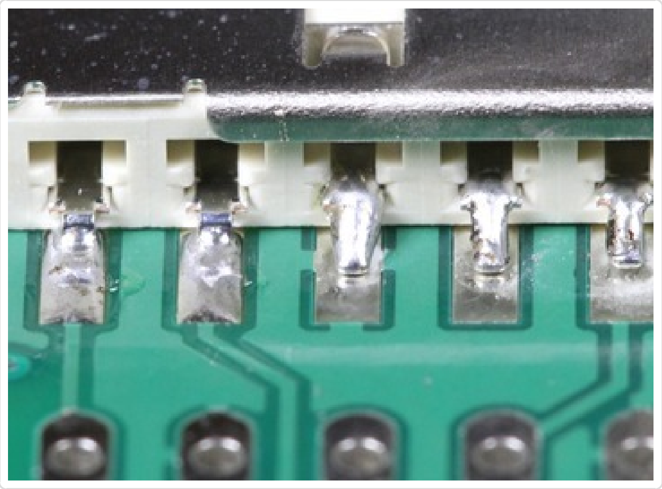

Insufficient Wetting (Surface Mount)

Here we have three pins of a surface mount component where the solder has not flowed onto the solder pad. This is caused by heating the pin instead of the pad.

Repair: This is easily repaired by heating the solder pad with the tip of the iron, then applying solder until it flows and melts together with the solder already on the pin.

Prevention: Heat the pad first.

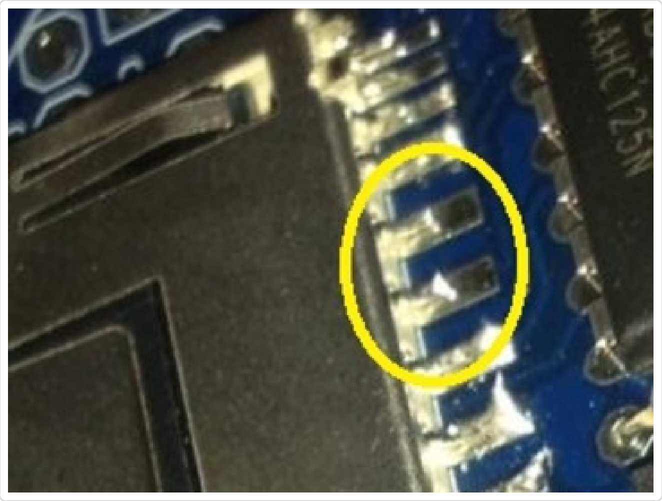

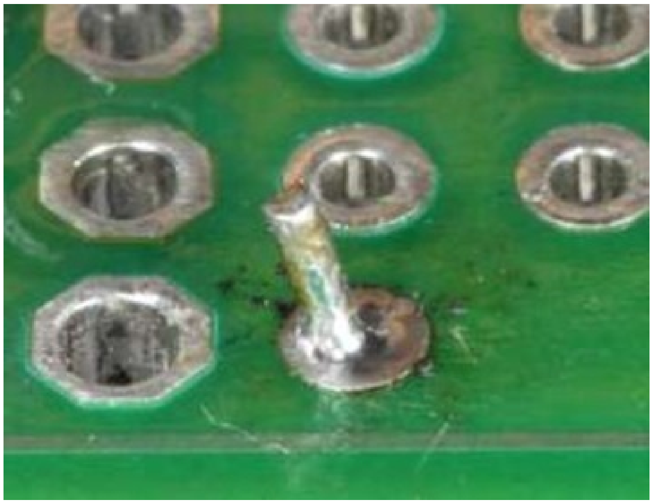

Solder Starved

A solder starved joint simply does not have enough solder. It may make good electrical contact, but it is hard to verify by inspection. In any case, it is not a strong joint and may develop stress cracks and fail over time.

Repair: Re-heat the joint and add more solder to make a good strong joint.

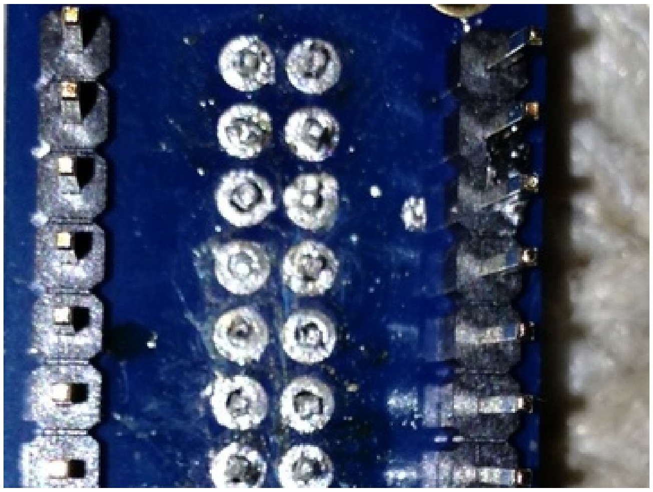

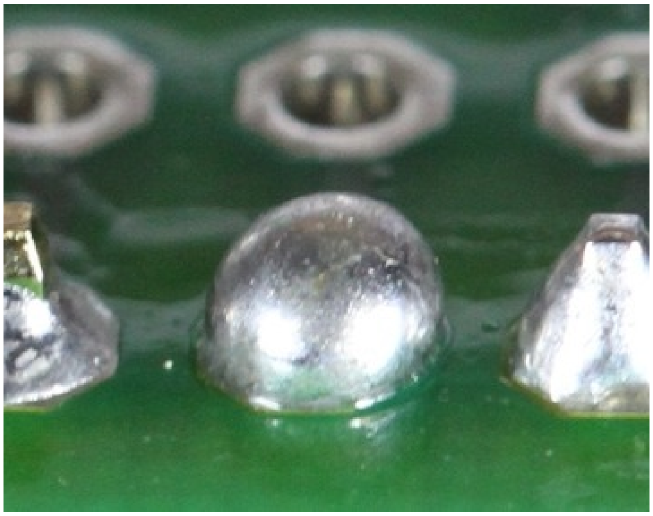

Too Much Solder

This might be a perfectly good joint, but we can't tell for sure. It is entirely possible that this blob of solder wets neither the pin nor the pad and is not a reliable electrical connection. The best evidence of proper wetting (and good electrical contact) is a nice concave surface as on the joint on the far left.

Repair: It is usually possible to draw off some of the excess solder with the tip of a hot iron. In extreme cases, a solder-sucker or some solder wick can be helpful as well.

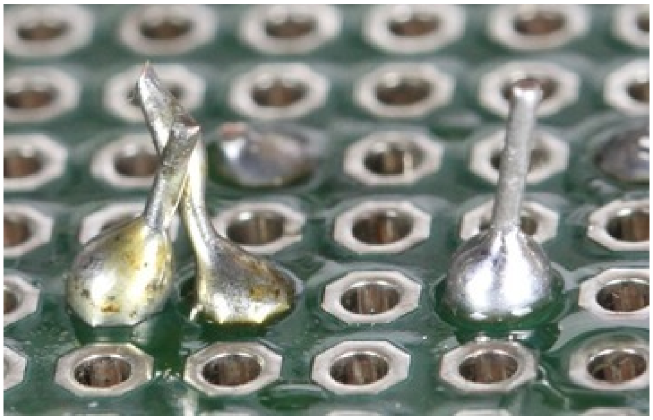

Untrimmed Leads

Leads that are too long are potential short circuits. The two joints on the left are an obvious danger of touching. But the one on the right is long enough to be dangerous as well. It would not take much force to bend that lead over to touch an adjacent trace.

Repair: Trim all leads just at the top of the solder joint.

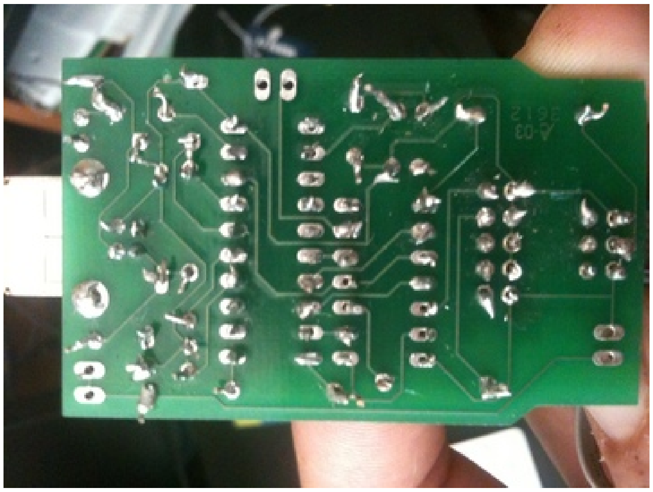

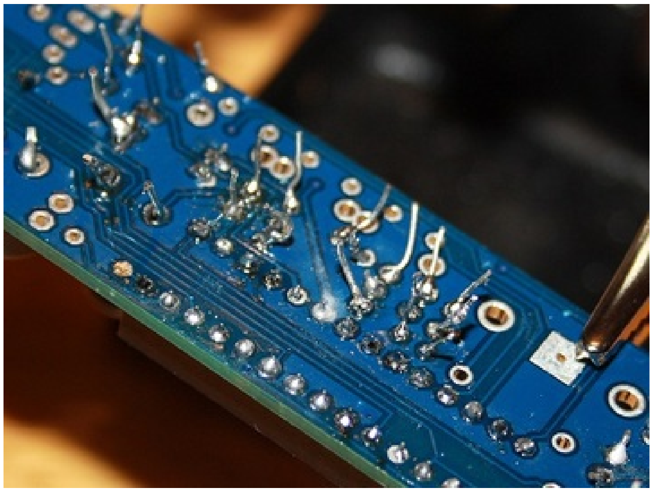

Solder Bridge

The left two solder joints have melted together, forming an unintended connection between the two.

Repair: Sometimes the excess solder can be drawn off by dragging the tip of a hot iron between the two solder joints. If there is too much solder, a solder sucker or solder wick can help get rid of the excess.

Prevention: Solder bridges most often happen between joints with too much solder to begin with. Use only enough solder to make a good joint.

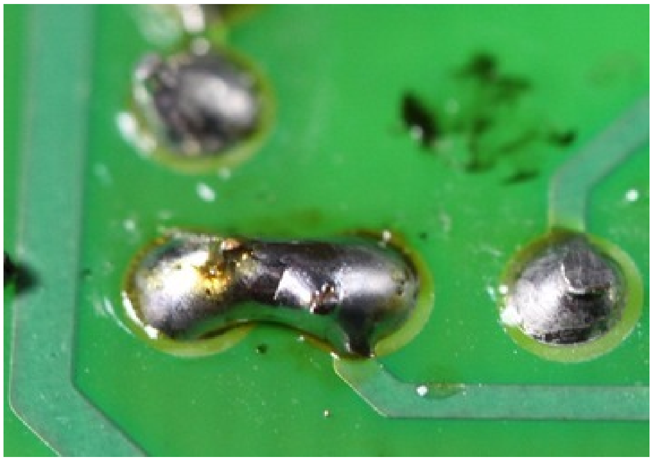

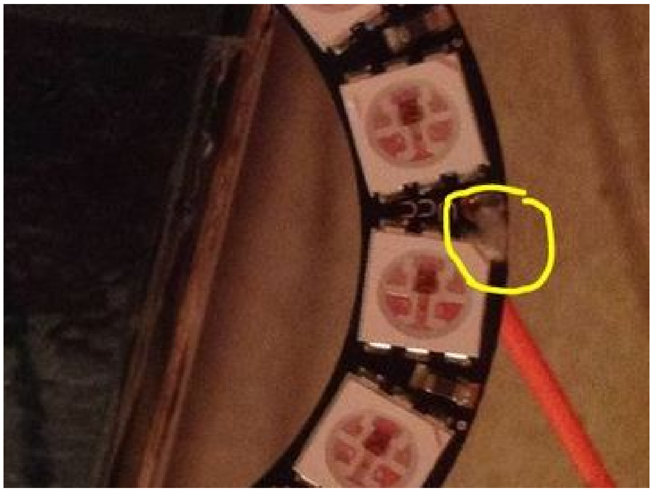

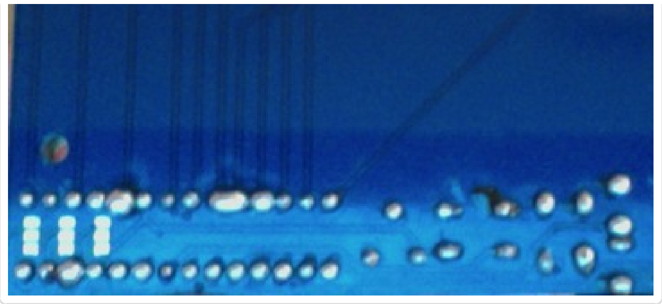

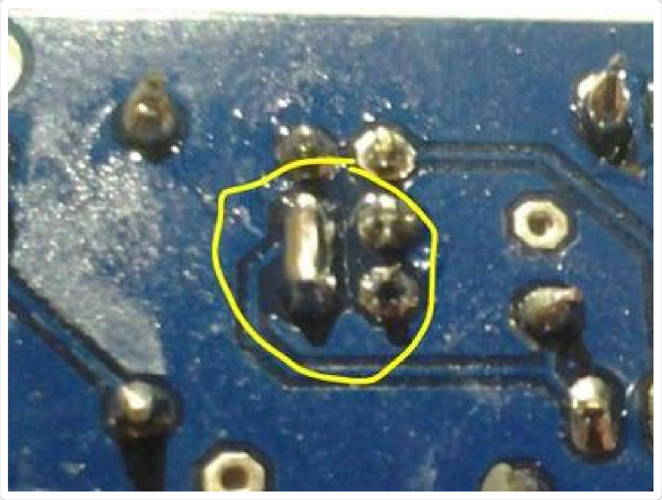

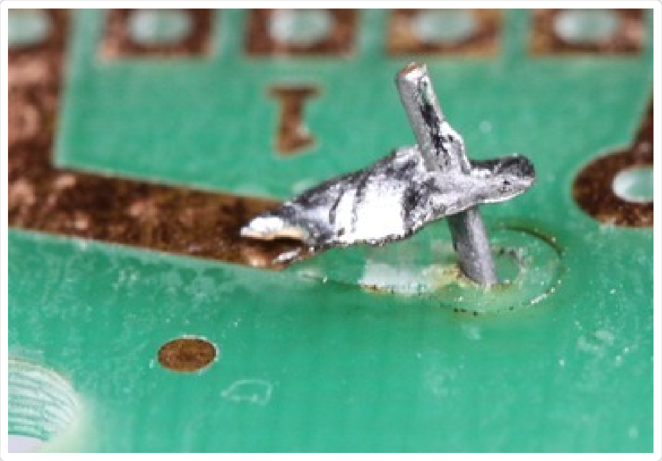

Lifted Pad

This photo shows a solder pad that has become detached from the surface of the circuit board. This most often occurs when trying to de-solder components from the board. But it can result simply from overworking the joint to the point where the adhesive bond between copper and the board is destroyed.

Lifted pads are especially common on boards with thin copper layers and/or no through-plating on the holes.

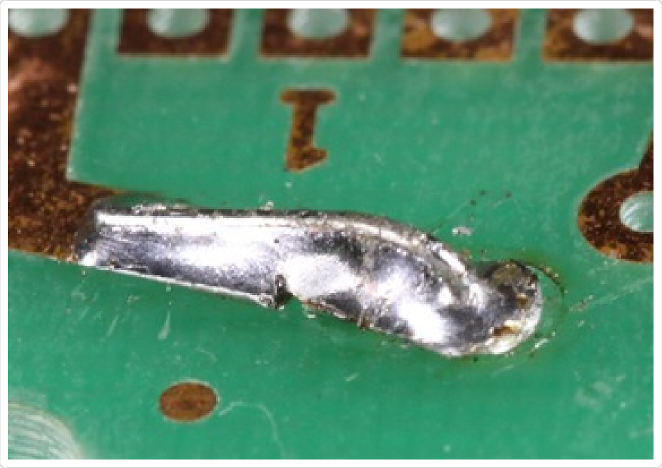

Repairing a Lifted Pad

It may not be pretty, but a lifted pad can usually be repaired. The simplest repair is to fold the lead over to a still-attached copper trace and solder it as shown to the left. If your board has a solder-mask, you will need to carefully scrape off enough to expose the bare copper.

Other alternatives are to follow the trace to the next via and run a jumper to there. Or, in the worst case, follow the trace to the nearest component and solder your jumper to the leg of that. Not exactly pretty, but functional.

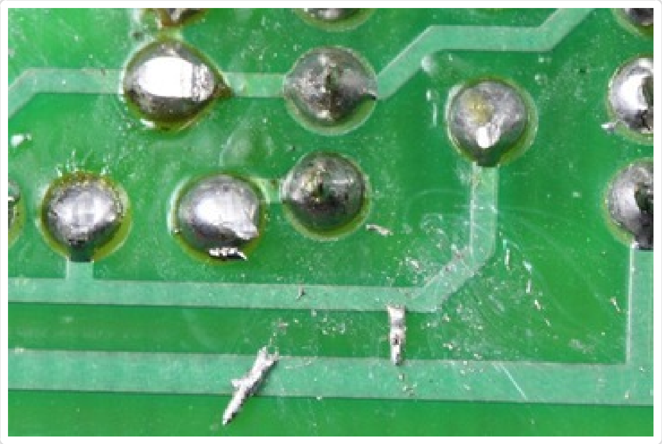

Stray Solder Spatters

These bits of solder are held to the board only by sticky flux residue. If they work loose, they can easily cause a short circuit on the board.

Repair: These are easy to remove with the tip of a knife or tweezers.

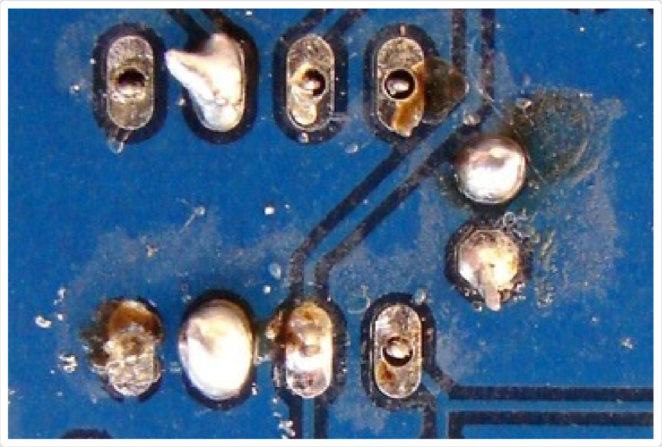

All of the Above!

Don't panic. Take your time. Most joints can be repaired with patience. If the solder refuses to flow the way you want it to:

1. Stop and let the joint cool.
2. Clean and tin your iron.
3. Clean off any burnt flux from the joint.
4. Let the iron come back up to temperature.
5. Then reheat the joint and try again.

AN114

HAND SOLDERING TUTORIAL FOR FINE PITCH QFP DEVICES

Scope

This document is intended to help designers create their initial prototype systems using Silicon Lab's TQFP and LQFP devices where surface mount assembly equipment is not readily available. This application note assumes that the reader has at least basic hand soldering skills for through-hole soldering. The example presented will be the removal, cleanup and replacement of a TQFP with 48 leads and 0.5 mm lead pitch.

Safety

Work should be done in a well-ventilated area. Prolonged exposure to solder fumes and solvents can be hazardous. There should be no presence of sparks or flames when solvents are in use.

Materials

The right materials are key to a good solder job. The list below are the recommendations from Silicon Labs. Other materials may work, so the user should feel free to substitute and experiment. The use of organic solder is highly recommended.

Required

1. Wire wrap wire (30 gauge) *
2. Wire strippers for wire wrap wire *
3. Soldering station - variable temperature, ESD-safe. Should support temperatures 800°F (425°C). This example uses a Weller model EC1201A. The soldering wand should have a fine tip no more than 1 mm wide.
4. Solder - 10/18 organic core; 0.2" (0.5 mm) diameter
5. Solder flux - liquid type in dispenser
6. Solder wick - size C 0.075" (1.9 mm)
7. Magnifier - 4X minimum. An inexpensive headset OptiVISOR by Donegan Optical Co. is used for this example.
8. ESD mat or tabletop and ESD wrist strap - both grounded
9. Tweezers with pointed (not flat) tips
10. Isopropyl Alcohol
11. Small stiff bristle brush for cleaning (nylon or other non-metallic material). Cut off hair to approximately 0.25" (6 mm)

* Required for device removal only.

Optional

1. Board vise to hold printed circuit board
2. Dental pick (90 degree bend)
3. Compressed dry air or nitrogen to dry boards

Copyright © 2003 by Silicon Laboratories

4. Optical inspection stereo microscope 30-40X

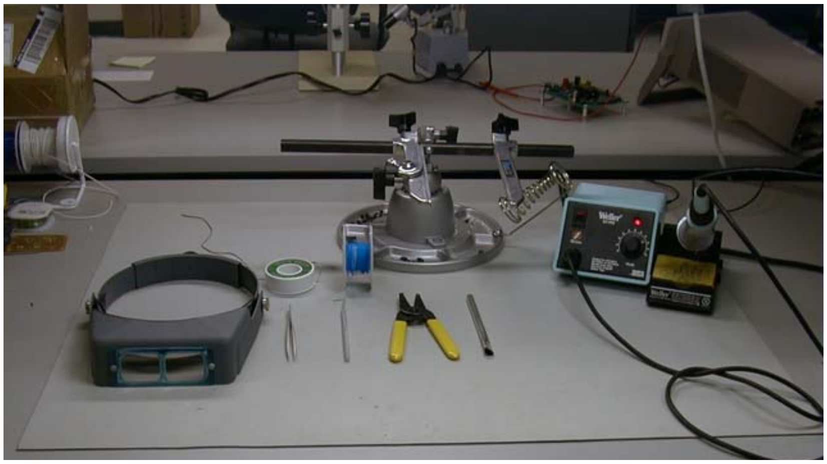

Figure 1. Some of the materials to get started...

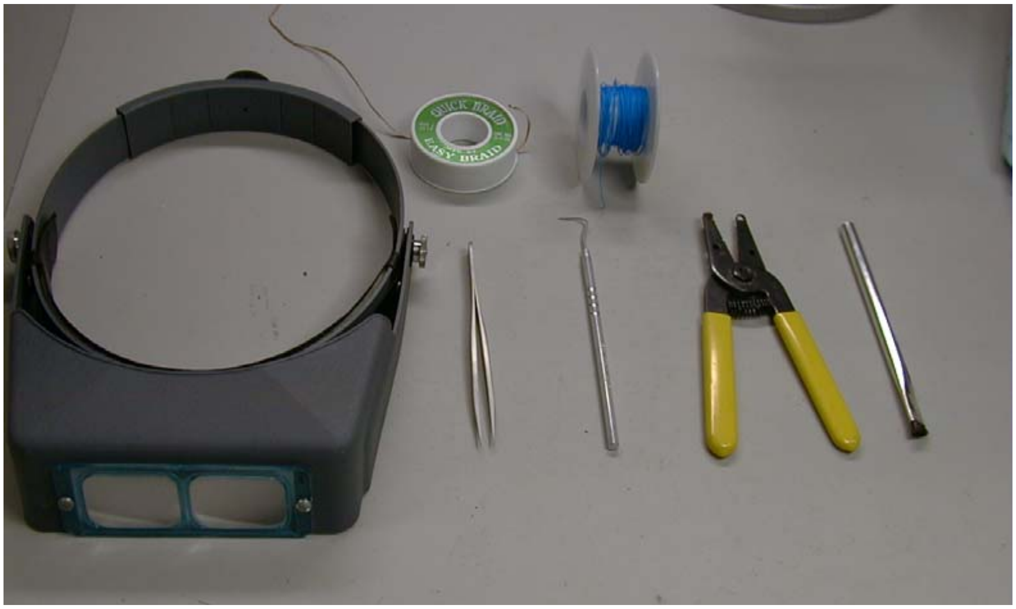

Figure 2. clock-wise from left: 4X magnifying headset, solder wick, wire wrap wire, stiff cleaning brush, wire strippers, and pointed tweezers

QUICK BRAID
EASY BRAID
Wire Wrapping
WIRE WRAPPING WIRE
30
100
R30B-0100

Figure 3. Solder wick and wire wrap wire

HEALTH ★ CARE
AMERICA
ISOPROPYL
RUBBING ALCOHOL
91% BY VOLUME

Figure 3. Isopropyl Alcohol

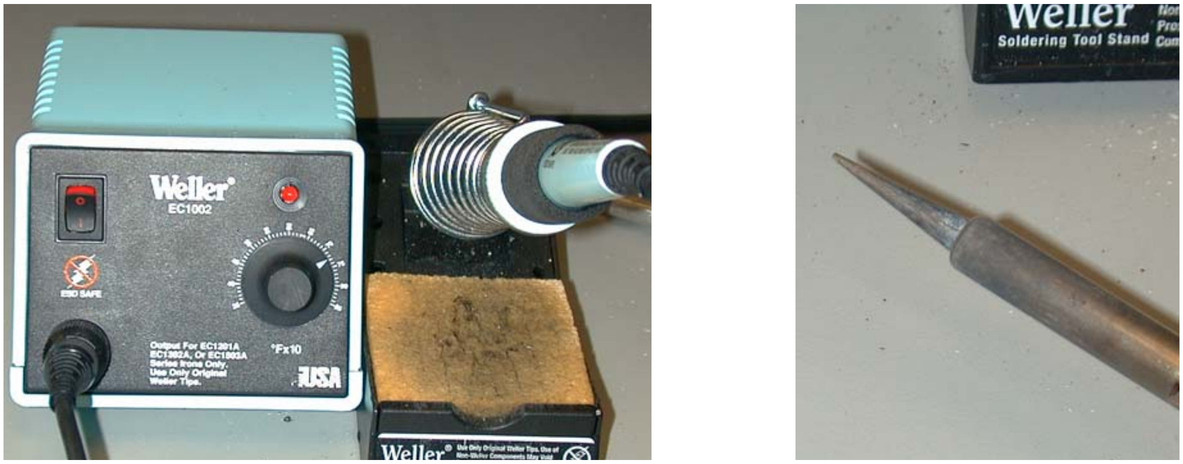
Weller
EC1002
Weller
Soldering Tool Stand

Figure 4. ESD safe soldering station with fine tip wand. This one is Weller model EC1201A

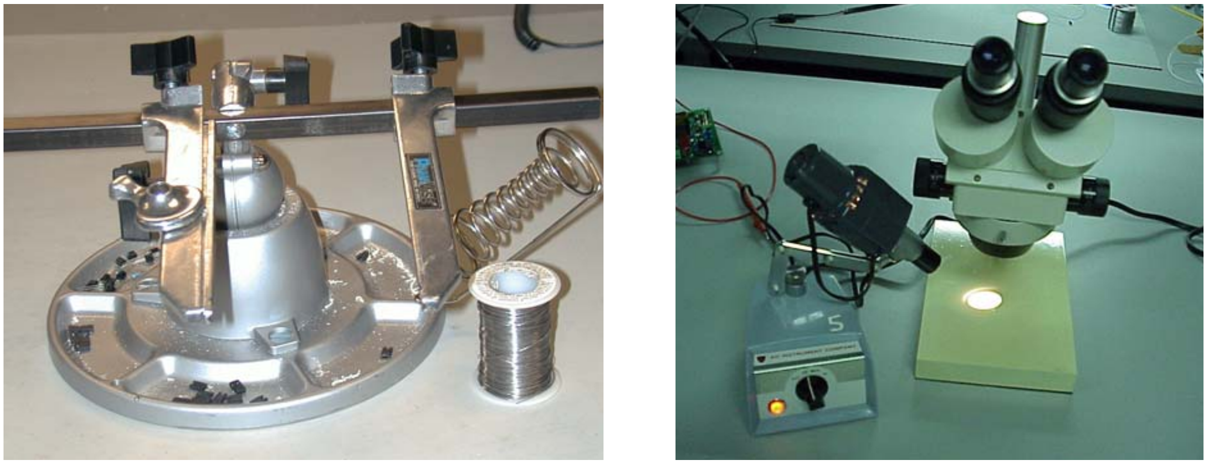

Figure 5. Optional equipment includes a PCB vise and an inspection microscope 7-40X

Procedure

The following procedures cover the replacement of a TQFP 48 pin device with 0.5 mm lead pitch. The lead shape is the standard gullwing associated with JEDEC standard QFPs. This procedure section is divided into three parts:

A. Part Removal

B. Board Cleanup

C. Soldering a new device.

If you are soldering parts to a new printed circuit board, skip part A and refer to the new board cleanup section in part B.

A. Part Removal

Preparation:

- Board with IC to be removed is mounted in a holder or vise. A PCB holder/vise is optional but it is required that the PCB is held steady for the part removal.
- The soldering station is warmed up to 800°F (422°C) and the solder tip is clean.
- ESD precautions have been taken.

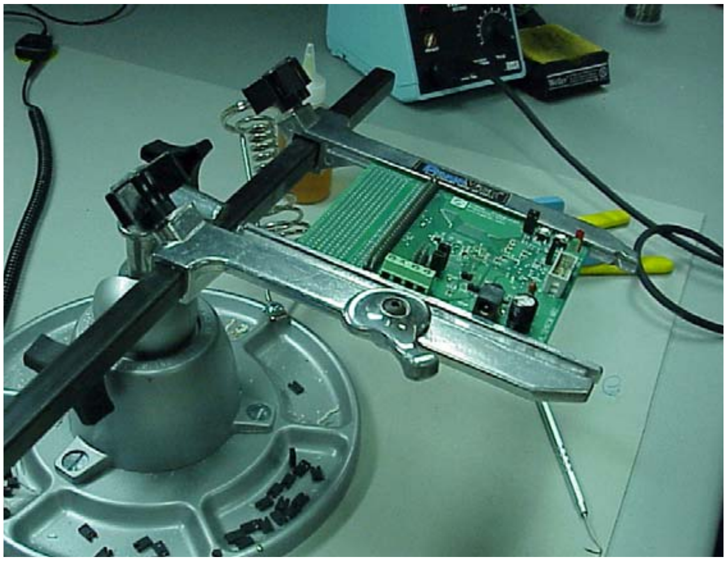

Figure 6. Ready to get started

Begin by wetting all the leads with flux to enhance the initial solder wicking cleanup. Wick up solder as much as possible from the QFP leads. Be careful not to scorch the PCB board with prolonged solder heat.

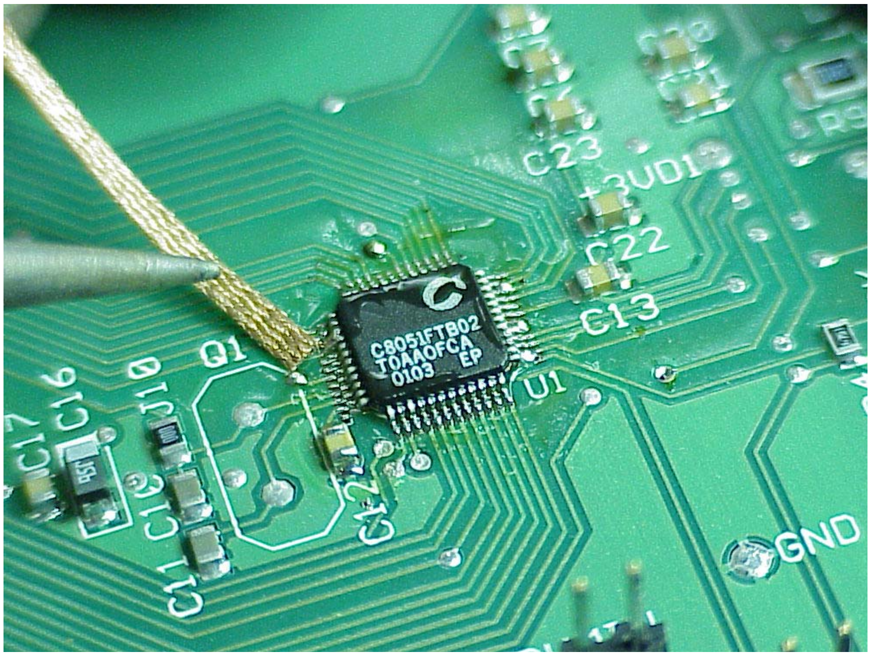
C23
3UD1
C22
C13
C8051FT802
T0AA0FCA
0103 EP
U1
Q1
C16
C17
J10
C12
C11
GND

Figure 7. Apply flux; wick excess solder from pins

Next, strip off approximately 3 inches of insulation from a piece of the 30 gauge wire wrap wire. Cut the wire at a comfortable 12 inch length or so.

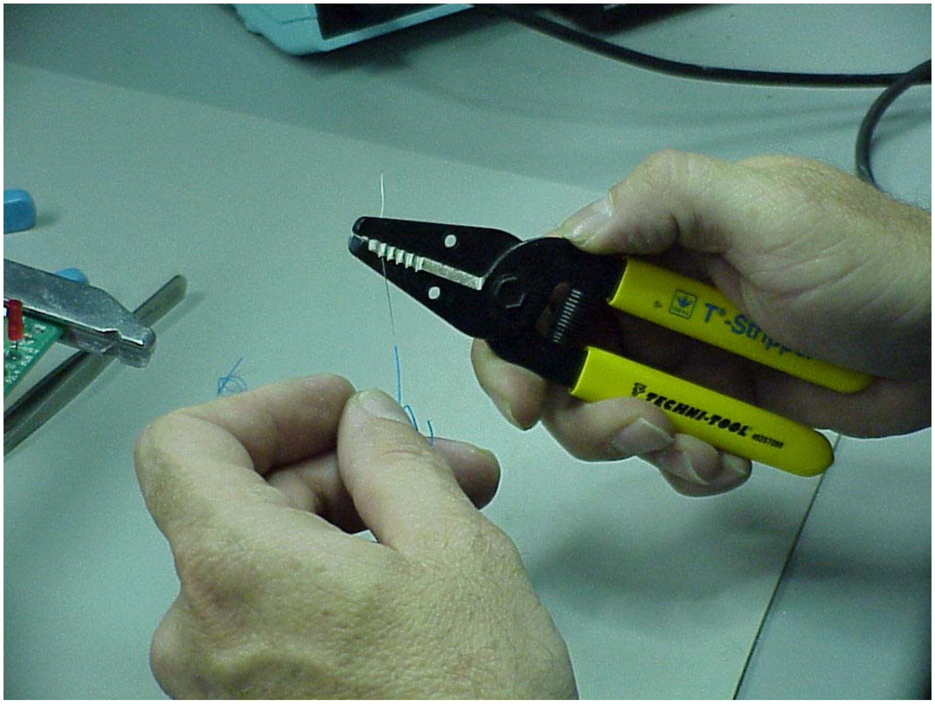
T-Stripper
TECHNI-TOOL

Figure 8. Wire stripping

Feed the wire behind and under the leads on one side of the IC as shown in Figure 9.

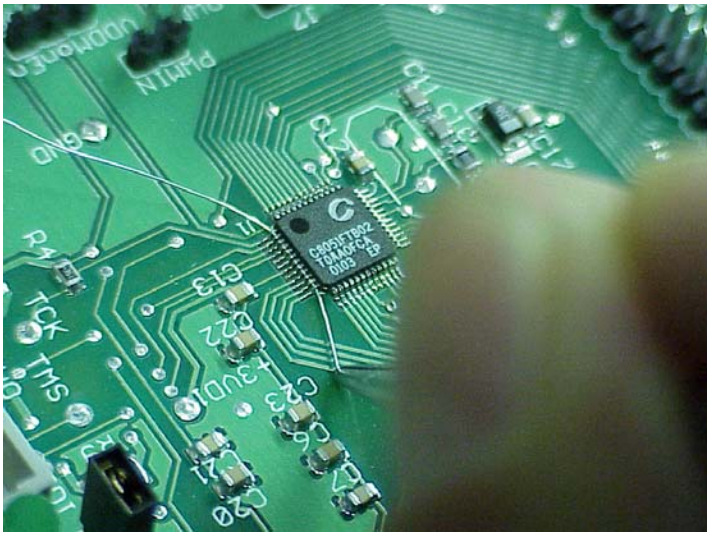

Figure 9. Feeding wire under QFP leads

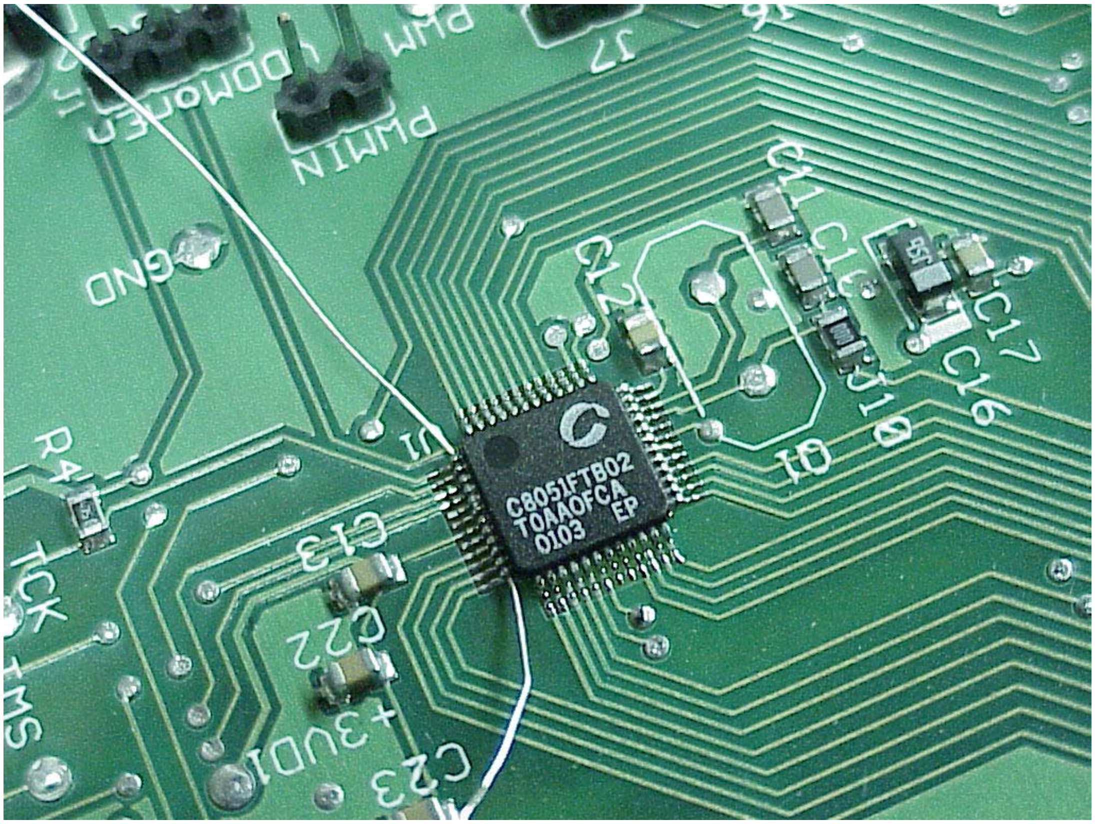

Figure 9. Wire with one side anchored to nearby component

Solder tack (anchor) one end of the 3 inch wire to a nearby via or component on the PCB. The anchor point should be in a location similar to that should in Figure 10.

Dispense a small amount of liquid flux across the leads.

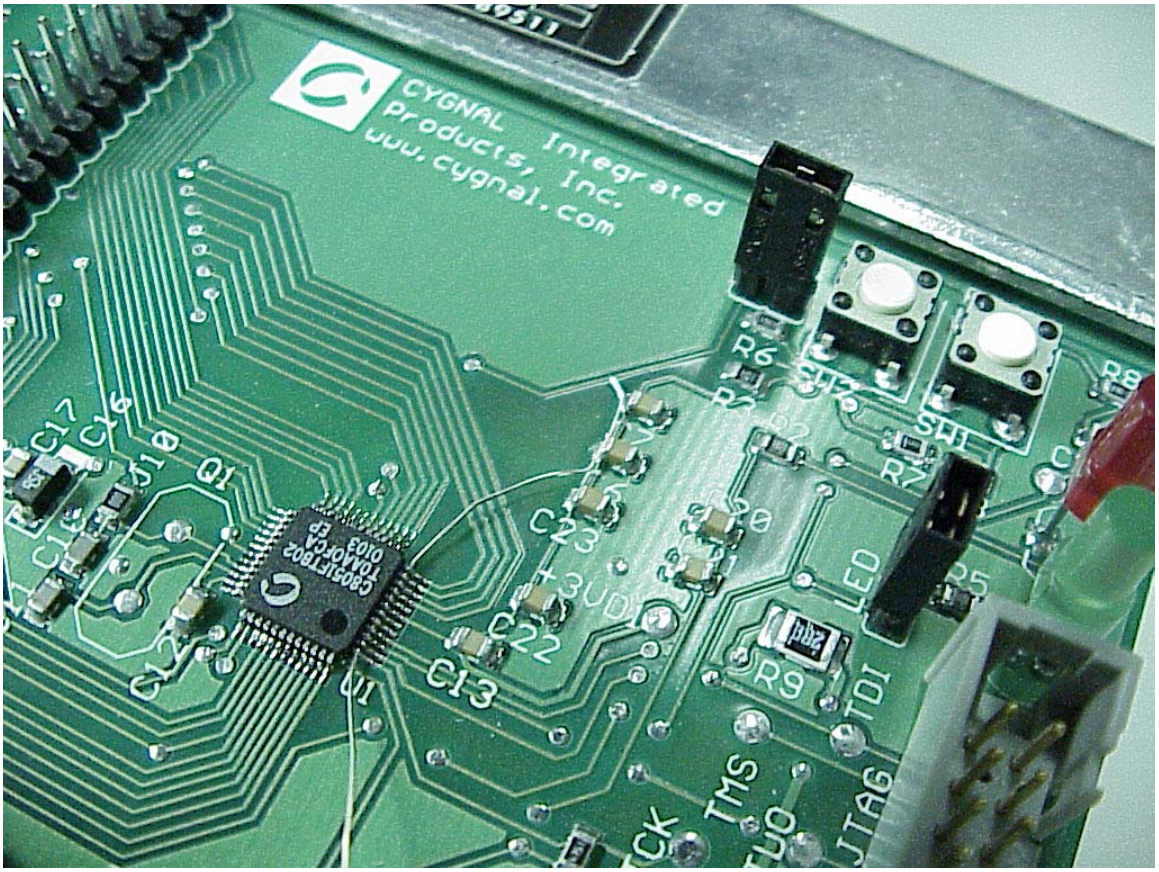
CYGNAL Integrated
Products, Inc.
www.cygnal.com
C8051F302
C17
C16
J10
Q1
U1
C13
C22
+3VD
C23
R6
R9
SW2
SW1
LED
TDI
TMS
TCK
JTAG

Figure 10. Bus wire is anchored on C6

Hold the loose (non-anchored) end of the wire with tweezers in close proximity to the device as shown in Figure 11.

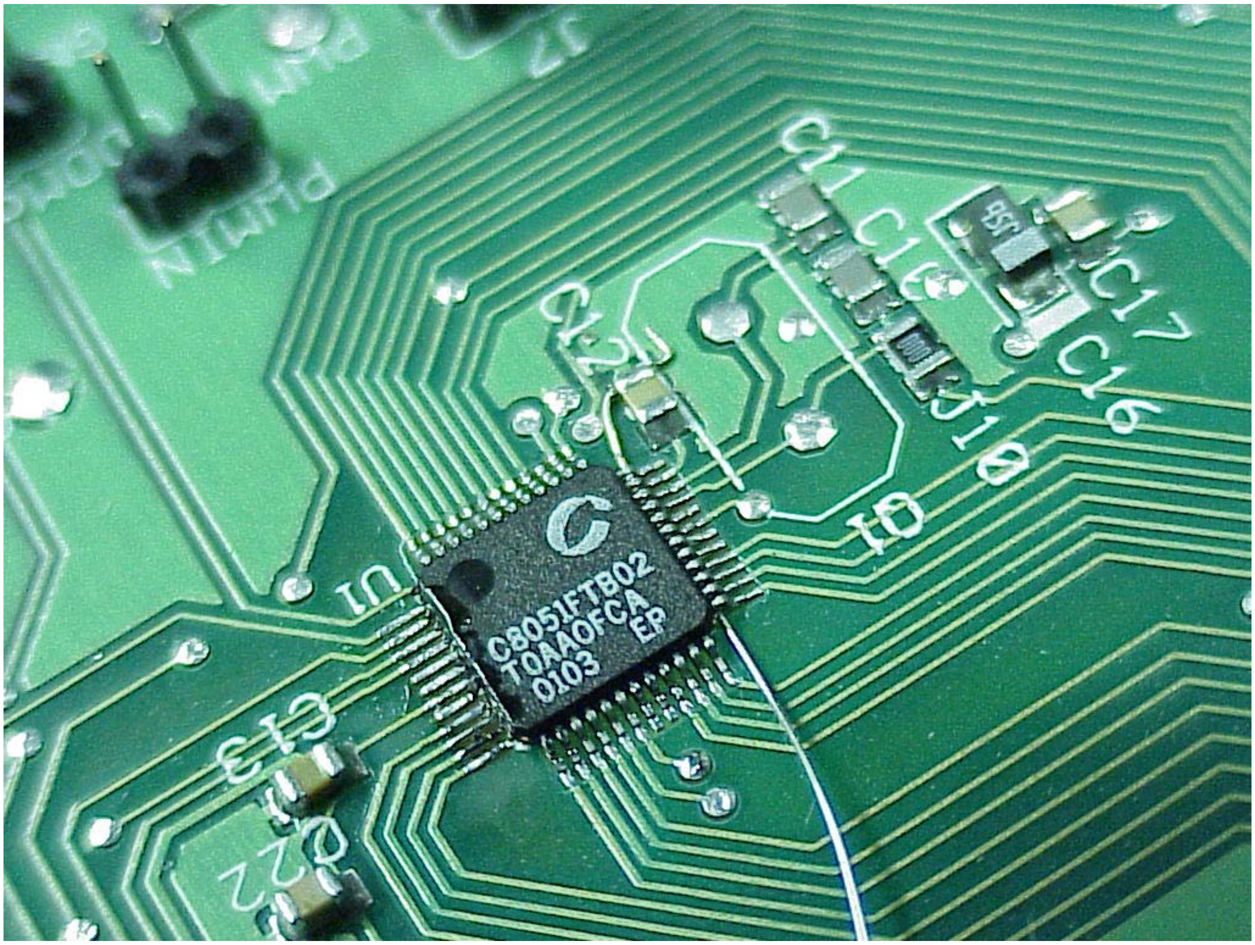

Figure 11. Side 2 anchored and ready for solder heat

You will now need to simultaneously apply solder heat and pull the wire away from the QFP, pulling at a slight upward angle from the board surface. Apply solder heat beginning at the lead closest to your tweezers. As the solder melts, gently pull the wire away from the QFP while continuing to move the solder heat from pin to pin to the right. You should not pull very hard. Pull as the solder melts. Do not leave the solder heat on any lead for more than necessary. The first lead will take the longest to heat, and after the wire gets hot, solder on the other leads will melt quickly. Excessive heat will damage the IC device and the PCB pad. The removal of 12 pins from a 48TQFP should take about 5 seconds total. Signs of excessive heat are:

- Melted plastic on the IC device
- PCB pads that lift off
- Brown scorch marks on the PCB

With one side of the QFP completed, repeat the same procedure on the other three sides of the QFP. Cut off the dirty part of the wire wrap wire or use a new piece for each side. Dispense flux again for each side.

Note that in the following pictures, the old IC device is not being saved. There is slightly more heat being applied here than necessary to speed up the process. This result is some melted plastic and missing gull-wing leads. These are visible in the pictures that follow. If you are trying to save the IC being removed, then you must be very careful to apply as little heat as possible during the removal process such that the QFP leads remain intact in the plastic QFP body. This will require some experimentation with solder heat settings and timing.

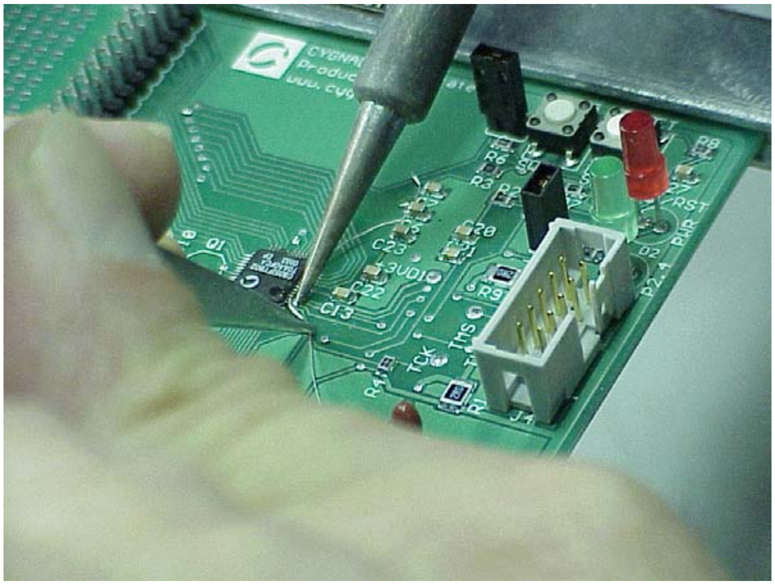

Figure 12. Hold tweezers close to device

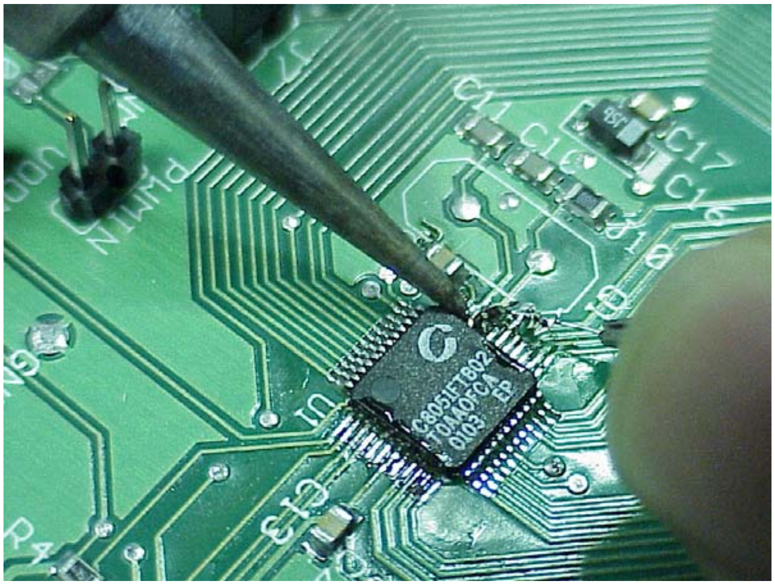

Figure 14. Side 2 almost complete

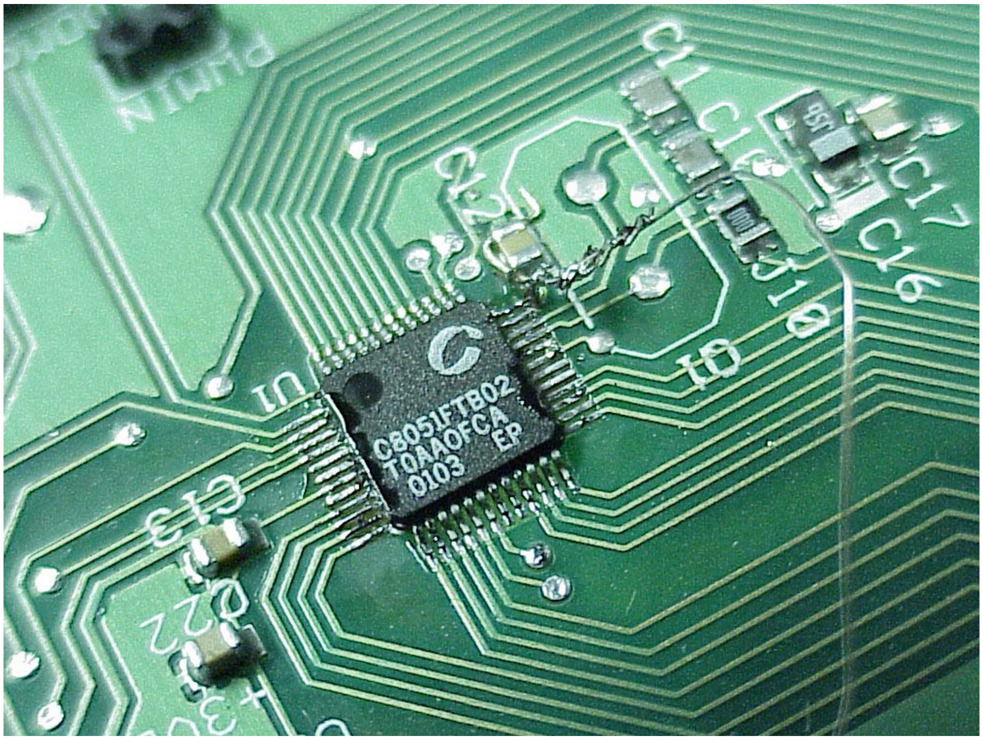

Figure 15. Side 2 is done

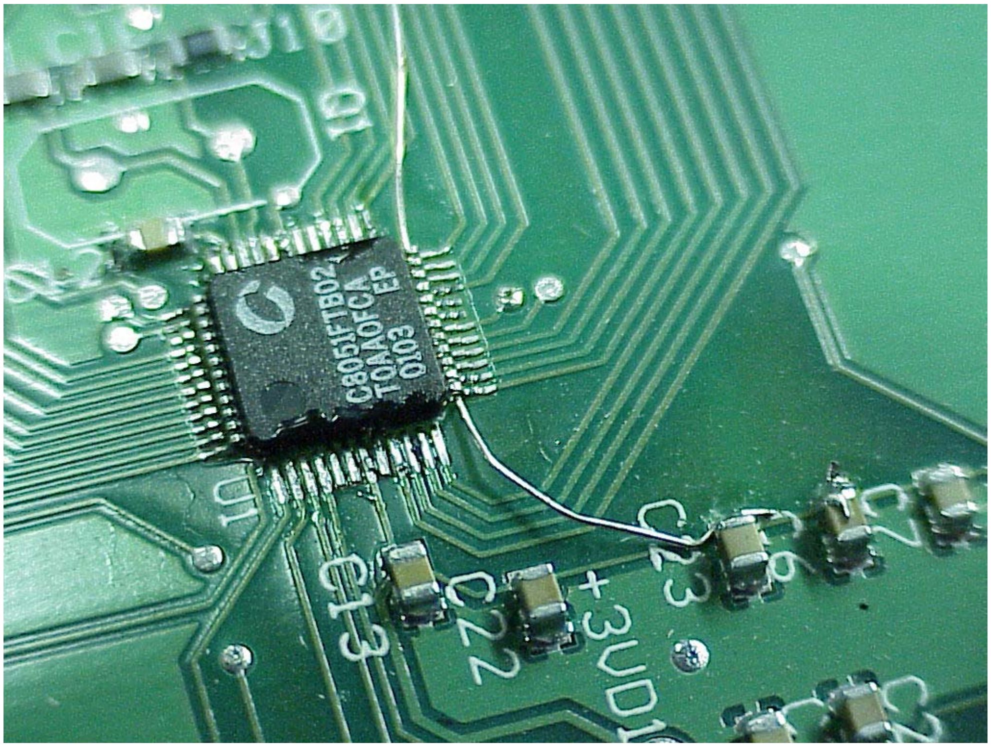

Figure 16. Side 3 anchored and ready

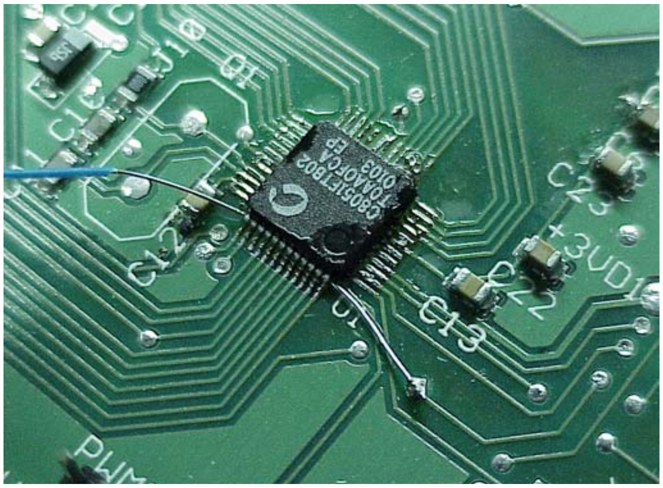

Figure 17. Side 4 anchored to a via

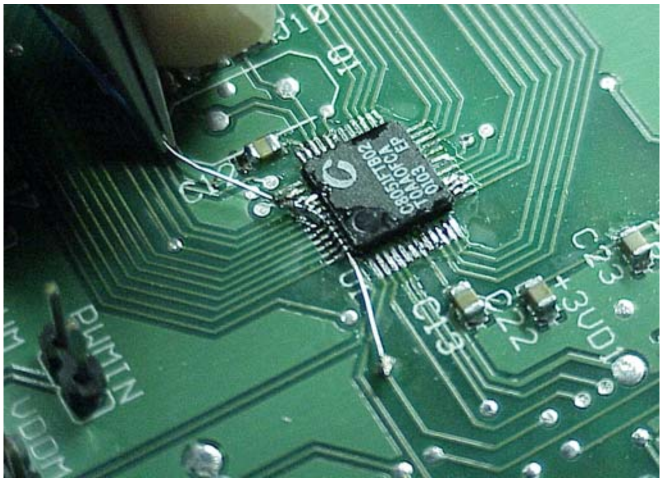

Figure 18. Side 4 removal started

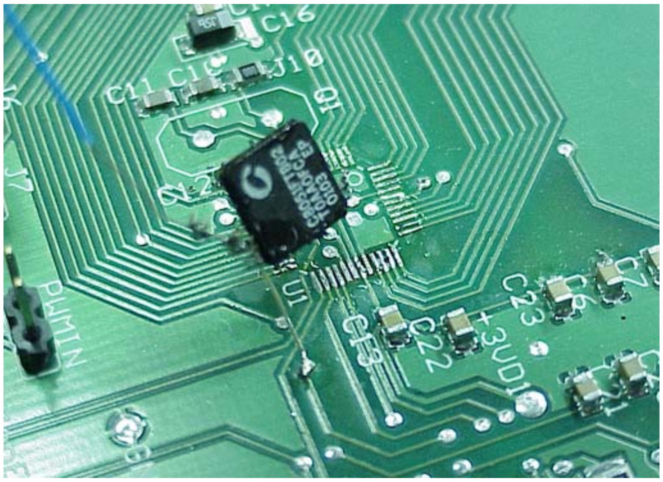

Figure 19. QFP removal a second before completion

B. Board Cleanup

New PCBs

For mounting a device to a new PCB, the amount of cleanup should be minimal. On a new PCB, there should be no solder on the pads. Brushing the pads with isopropyl alcohol (Figure 33) and drying the board should be enough preparation to begin the mounting procedure.

Reworked PCBs

The following section is the cleanup sequence that follows the QFP removal in the previous section. After removing the device, the solder pads will need cleanup. The idea is to clean the pads so that they are flat and free or solder and flux. Solder wick the pads until they are flat and dull. A clean pad appearance should be a dull silver color.

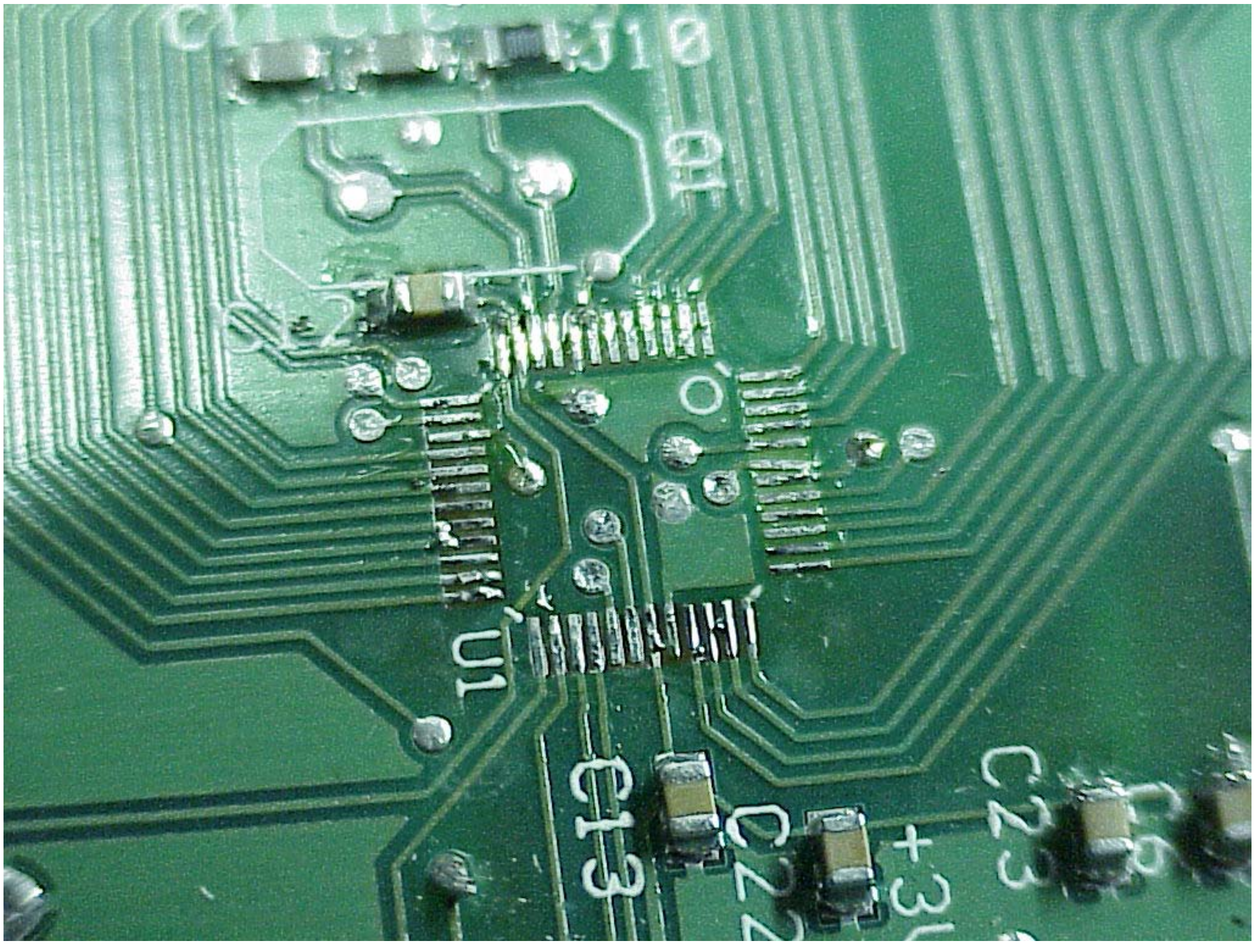

Figure 20. Pads after QFP removal procedure

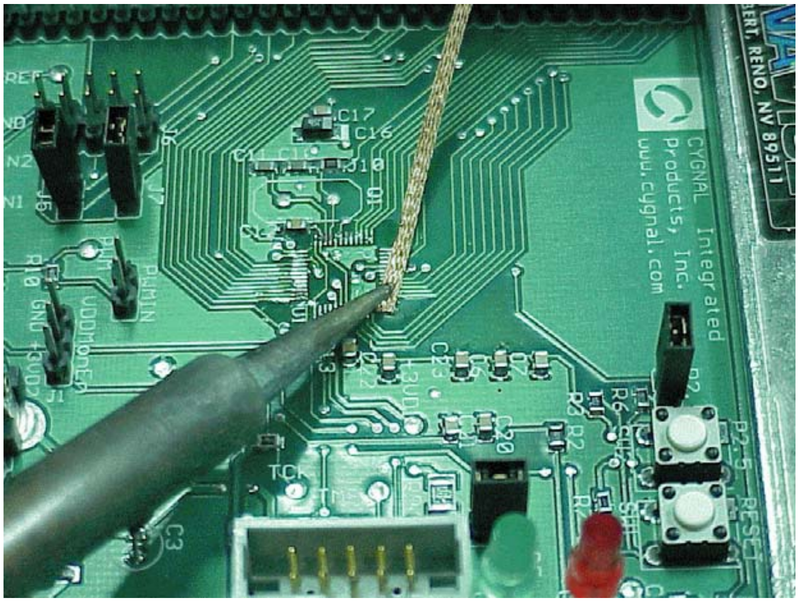

Figure 21. Solder wick to remove solder on the pads

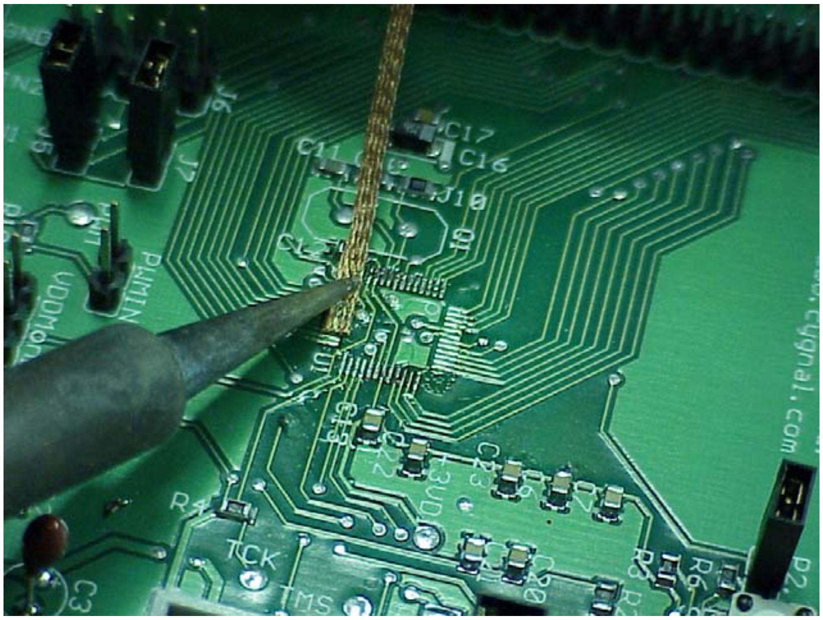

Figure 22. Repeat for all pads

If any pads become loose from the PCB, use a dental pick or other pointed object to re-align the pad (Figure 23 and Figure 24).

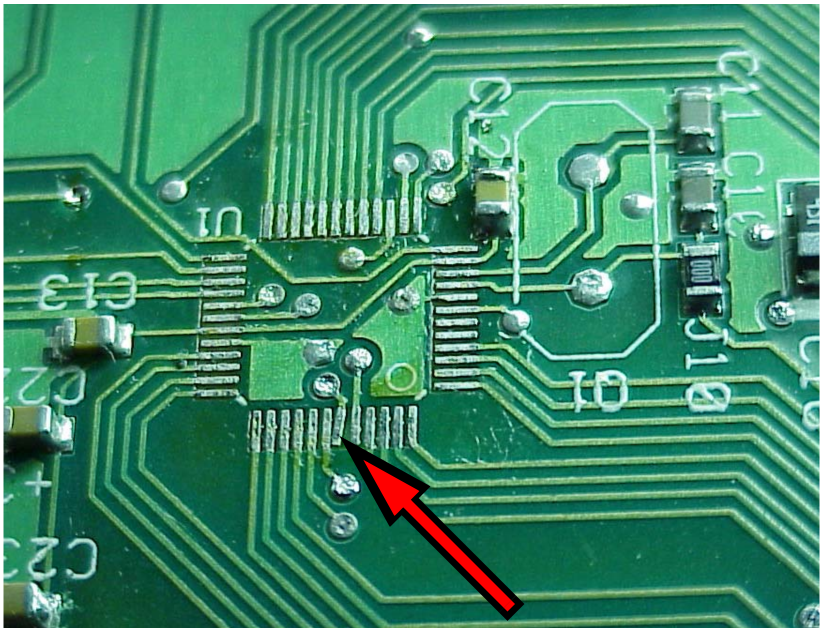

Figure 23. Clean pads but one pad is slightly crooked

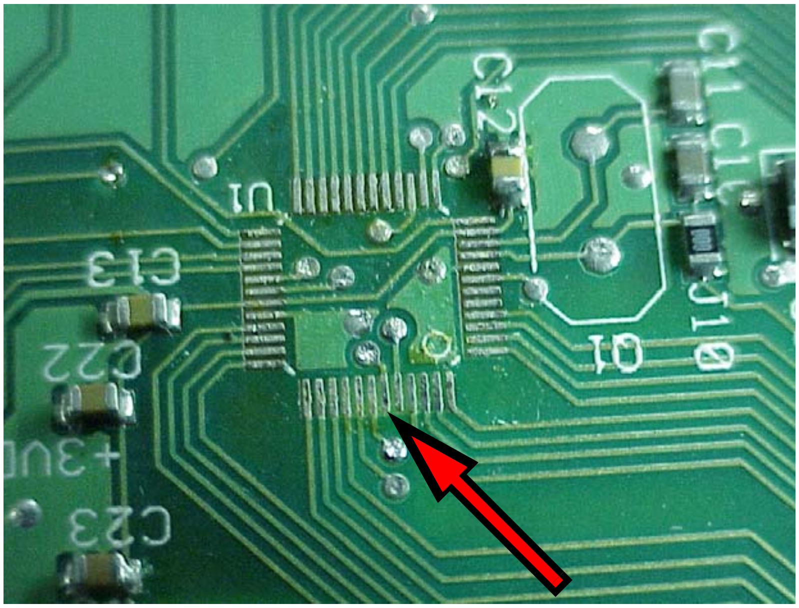

Figure 24. Pad straightened out

B. Soldering a new QFP

The pads on the PCB should be clean and free of any solder.

Carefully place the new QFP device on the PCB using tweezers or another safe method. Make sure the part is not dropped as the leads can be easily damaged.

Align the part over the pads using a small pick or similar tool to push the part. Get the alignment as accurate as you can. Also, make sure that the part is oriented correctly (pin 1 orientation).

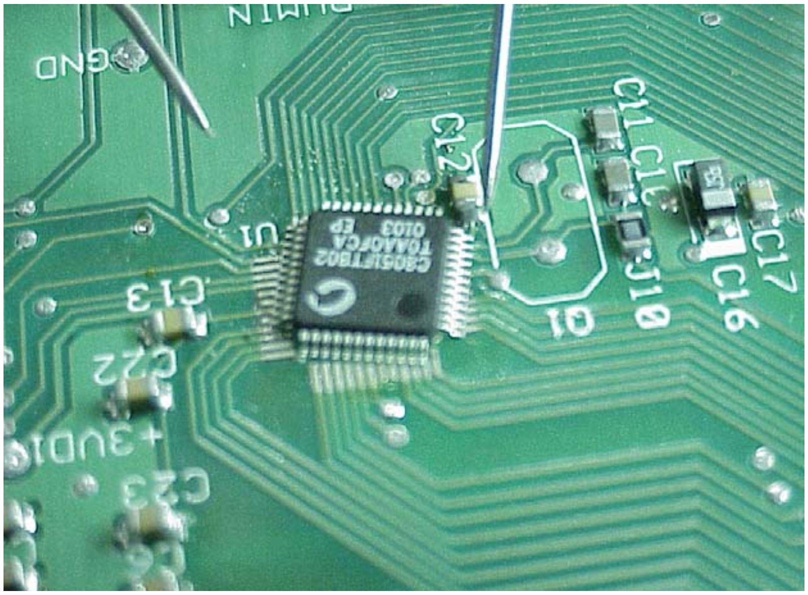

Figure 25. New QFP near pads, ready for alignment

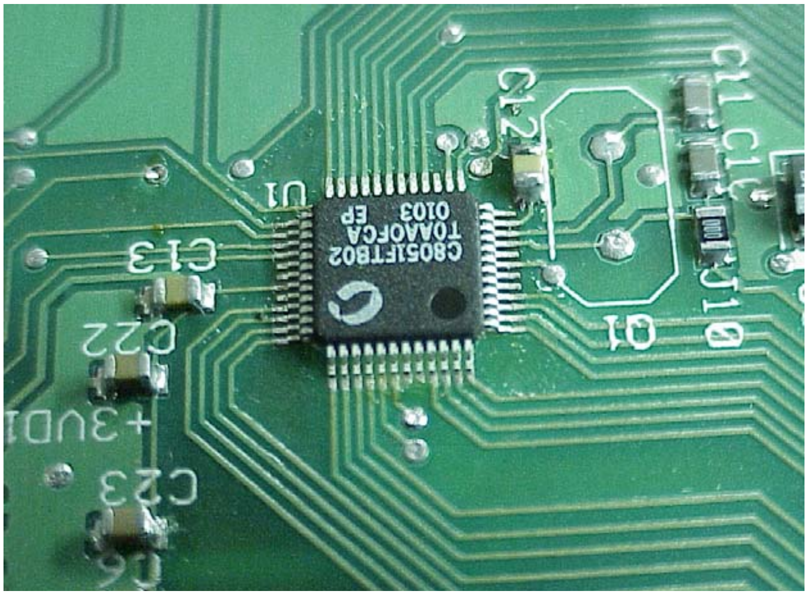

Figure 26. QFP aligned

Adjust the soldering station temperature to 725°F (385°C). Put a small amount of solder on the tip of the solder iron. While holding down the aligned QFP with a pick or other pointed tool, add a small amount of solder flux to the corner leads in two opposite corners. While still holding the part down with the pick, solder down two opposite corner leads on the QFP. Do not worry about excess solder or shorts between adjacent leads at this time. The idea is to anchor down the aligned QFP with solder so it does not move.

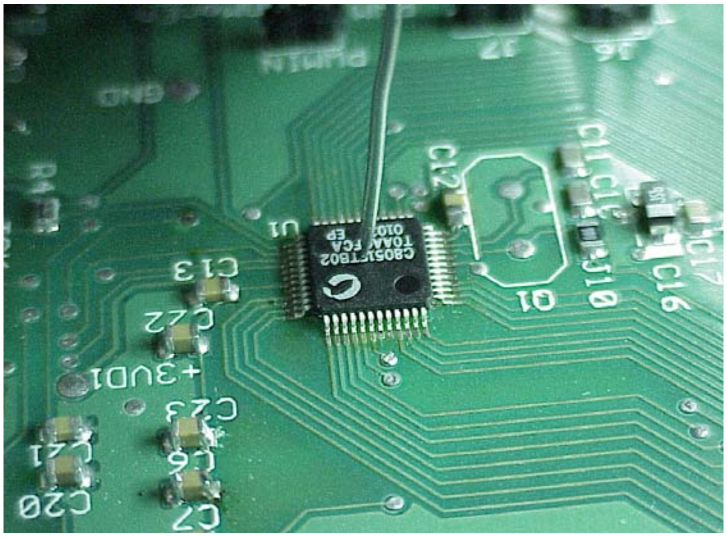

Figure 27. Aligned QFP ready for solder anchoring

Re-check the QFP alignment after soldering the corners. If necessary, make adjustments or remove and start over to get good QFP to PCB alignment.

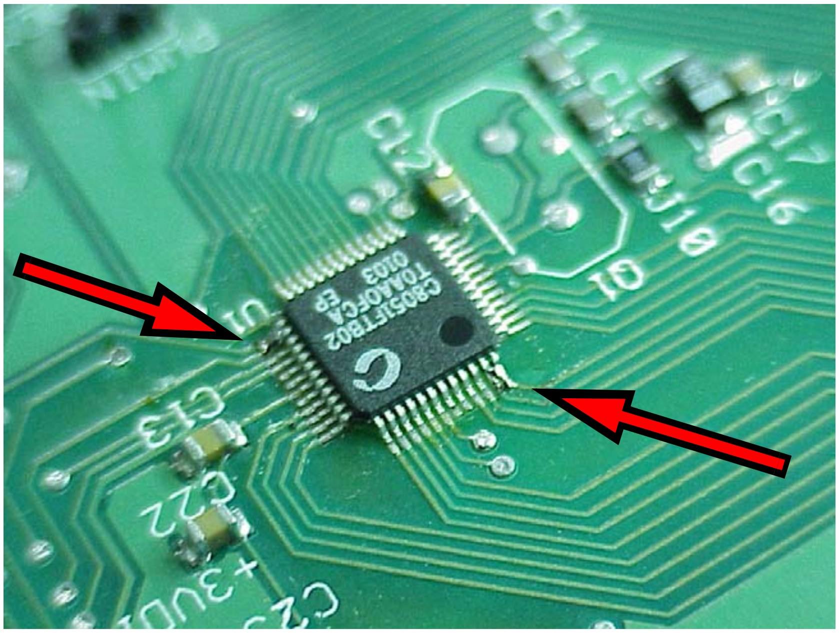

Figure 28. Aligned QFP with corners soldered down

Now you are ready to solder all the leads. Add solder to the tip of the soldering iron. Dispense flux over all the leads to keep them wet.

Touch the solder iron tip to the end of each QFP lead until the solder is seen running up the lead. Repeat for all the leads. Add small amounts of solder as needed to the soldering iron tip. Again, do not worry if you see some solder bridging as you will clean that up in the next step.

When soldering, keep the soldering iron tip parallel with the pins being soldered to prevent excessive solder shorts.

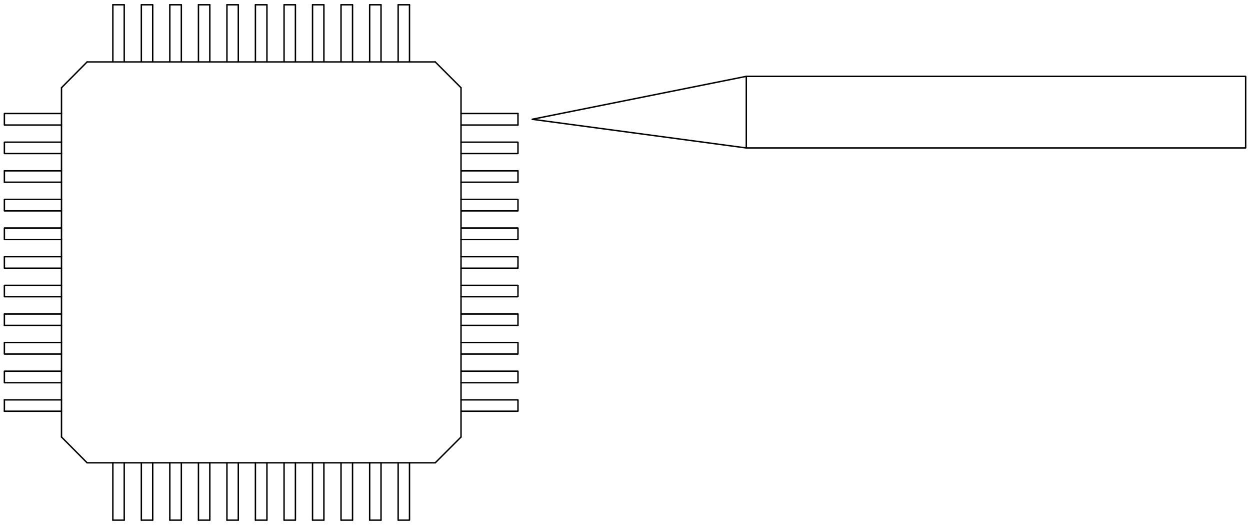

Figure 29. Keep iron tip parallel to pins being soldered

After soldering all the leads, wet all the leads with flux to enhance the solder wicking cleanup. Wick up solder where needed to eliminate any shorts/bridging.

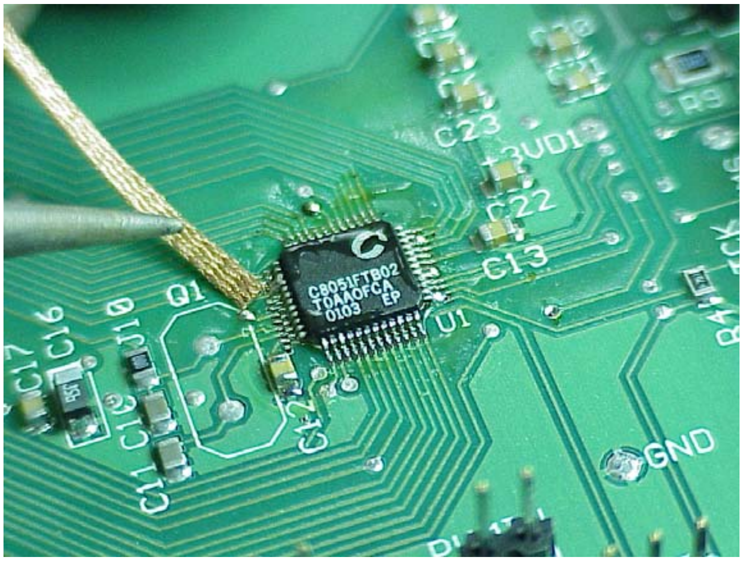

Figure 30. Wicking Solder #1

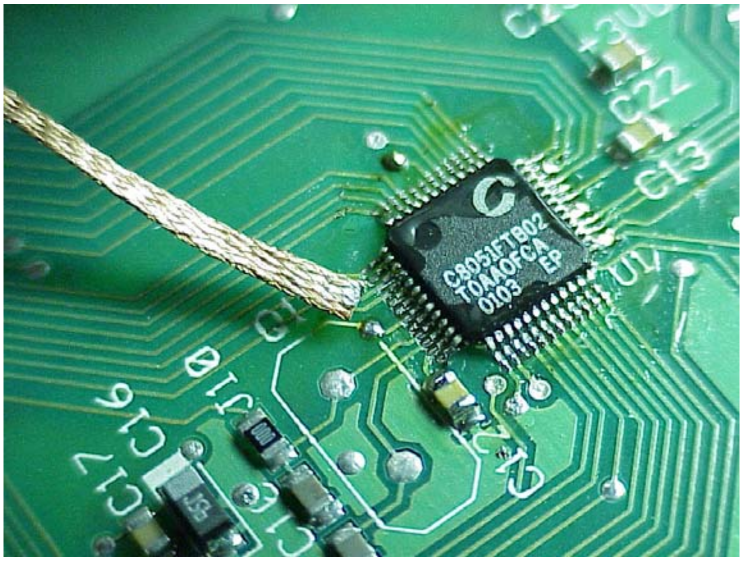

Figure 31. Wicking Solder #2

Hand inspect the board using 4X magnification (or higher) for shorts or marginal solder joints. Solder joints should have a smooth melt transition between each device pin and the PCB. Rework any pins as needed

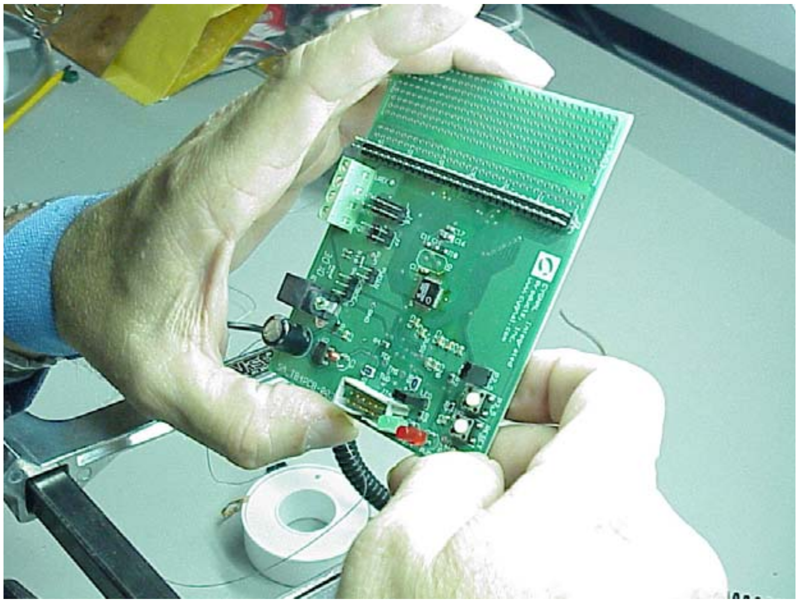

Figure 32. Visual Inspection

After the inspection passes, it is time to clean the flux off the board. Dip the stiff bristle brush into alcohol and wipe in the direction of the leads. Use moderate, but not excessive pressure. Use liberal amounts of alcohol and brush well between the QFP leads until the flux disappears.

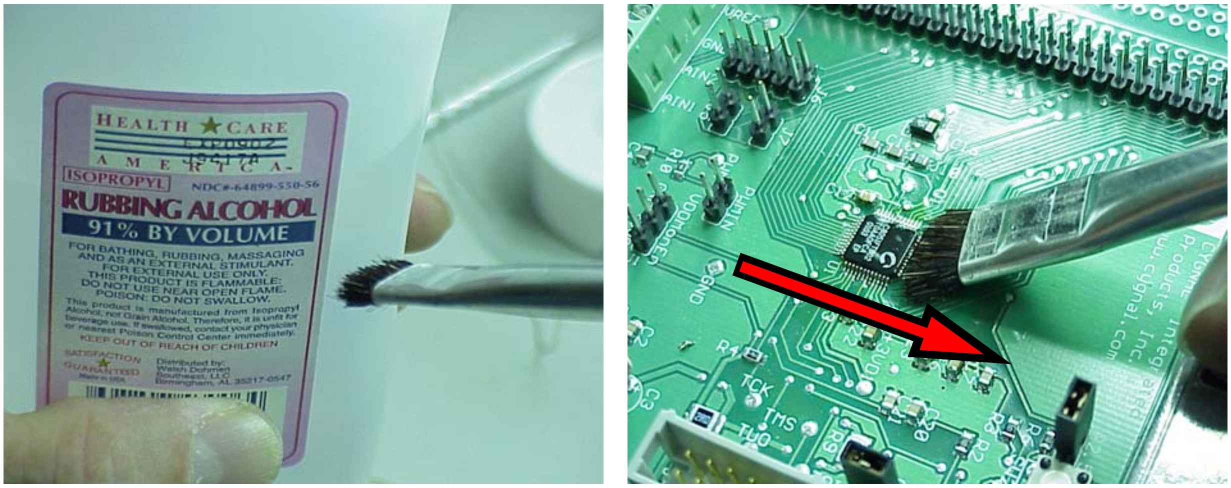

Figure 33. Isopropyl Alcohol and stiff brush used for cleanup. Brush only in direction of leads

Dry the board with compressed dry air or nitrogen. If this is not available, let the board dry for 30 minutes or more to let the alcohol evaporate under the QFP. The QFP leads should look bright and there should be no flux residue.

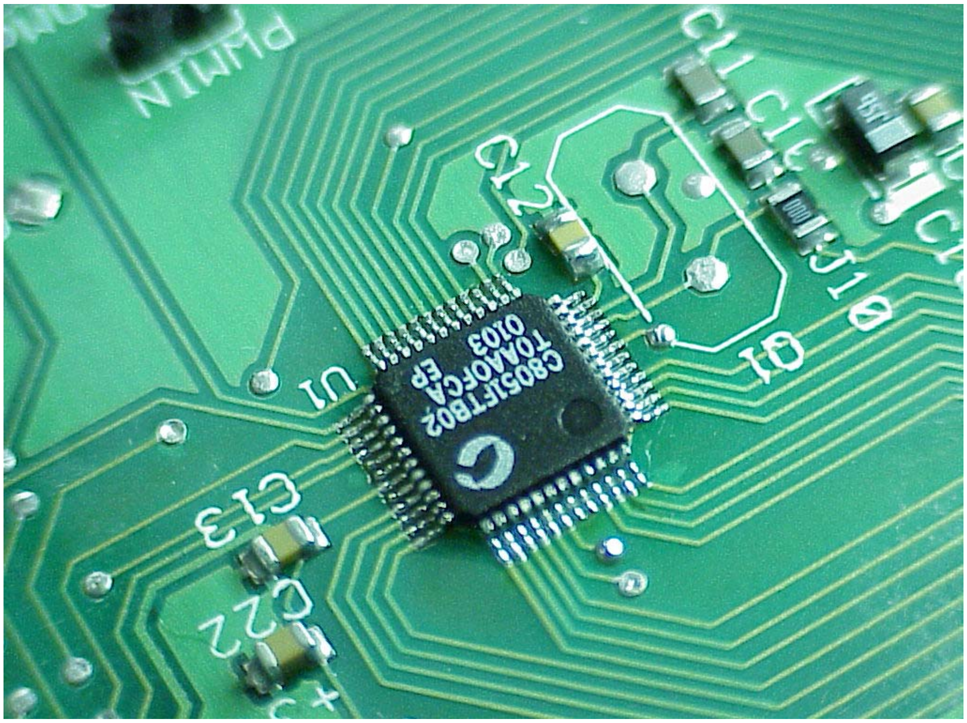
PWMIN
C11
C12
C1
J10
Q1
U1
C8051F002
TOAAOFCA
EP
0103
C13
C22

Figure 34. Clean and Pristine

Re-inspect the board for workmanship. Rework any leads if needed.

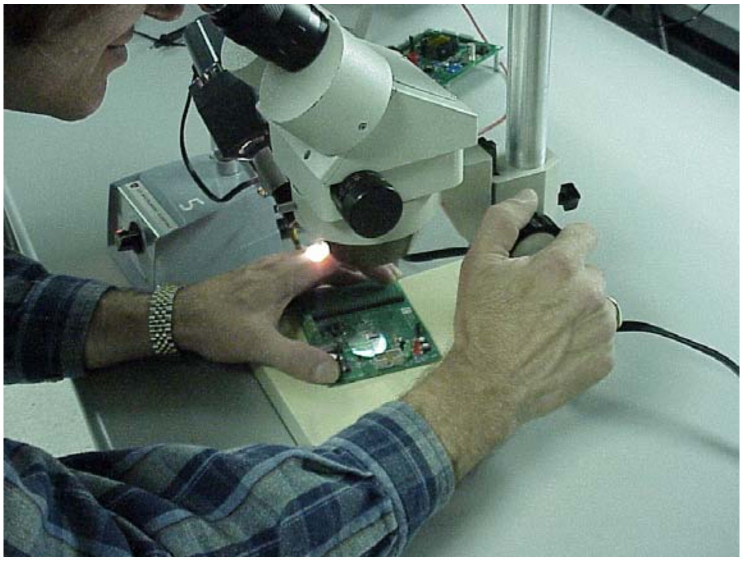

Figure 35. Stereo zoom inspection station (7X to 40X magnification) helps to inspect solder joint workmanship

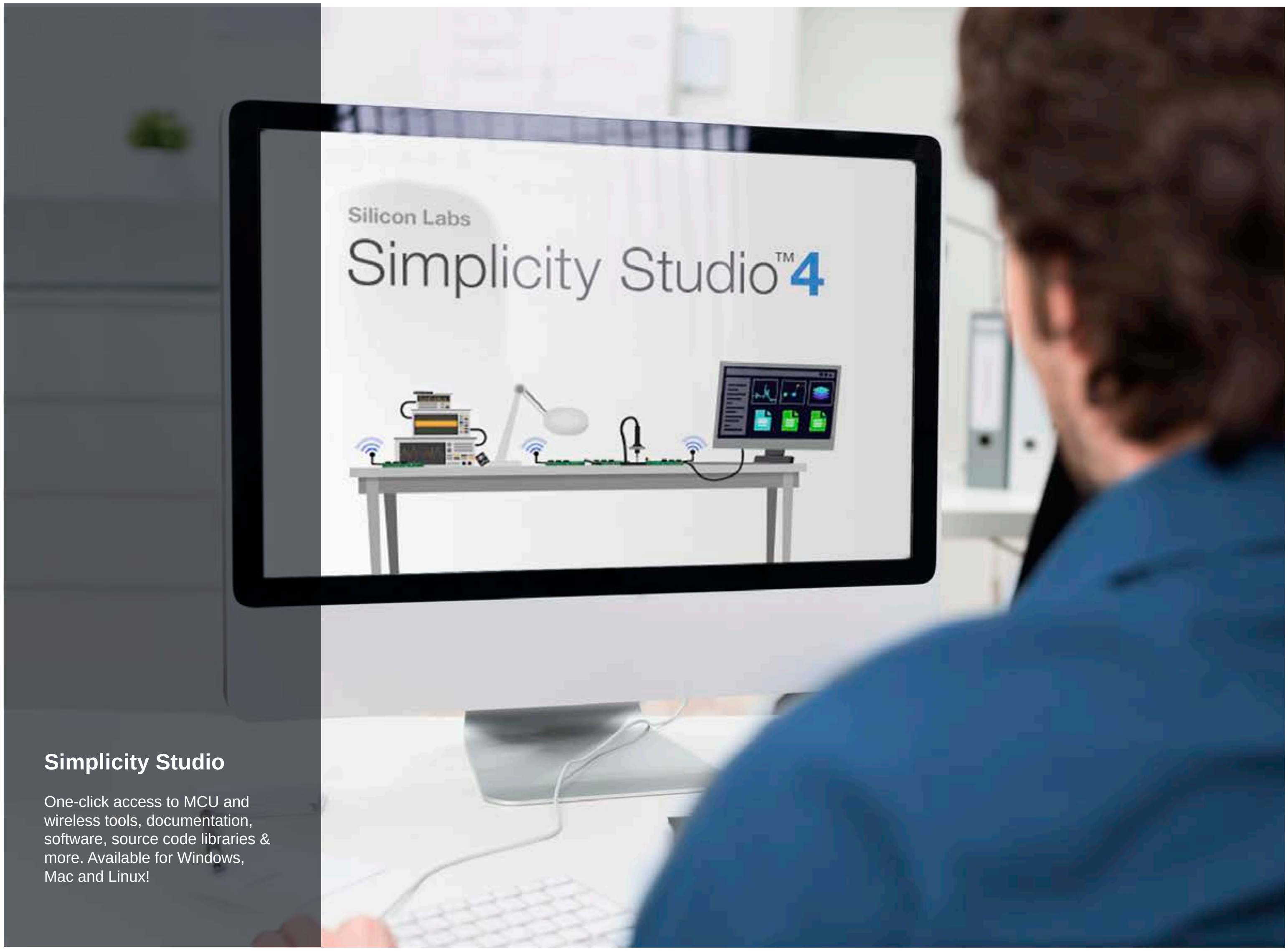

Simplicity Studio

One-click access to MCU and wireless tools, documentation, software, source code libraries & more. Available for Windows, Mac and Linux!

IoT Portfolio
www.silabs.com/IoT

SW/HW
www.silabs.com/simplicity

Quality
www.silabs.com/quality

Support and Community
community.silabs.com

Disclaimer
Silicon Labs intends to provide customers with the latest, accurate, and in-depth documentation of all peripherals and modules available for system and software implementers using or intending to use the Silicon Labs products. Characterization data, available modules and peripherals, memory sizes and memory addresses refer to each specific device, and "Typical" parameters provided can and do vary in different applications. Application examples described herein are for illustrative purposes only. Silicon Labs reserves the right to make changes without further notice and limitation to product information, specifications, and descriptions herein, and does not give warranties as to the accuracy or completeness of the included information. Silicon Labs shall have no liability for the consequences of use of the information supplied herein. This document does not imply or express copyright licenses granted hereunder to design or fabricate any integrated circuits. The products are not designed or authorized to be used within any Life Support System without the specific written consent of Silicon Labs. A "Life Support System" is any product or system intended to support or sustain life and/or health, which, if it fails, can be reasonably expected to result in significant personal injury or death. Silicon Labs products are not designed or authorized for military applications. Silicon Labs products shall under no circumstances be used in weapons of mass destruction including (but not limited to) nuclear, biological or chemical weapons, or missiles capable of delivering such weapons.

Trademark Information
Silicon Laboratories Inc.® , Silicon Laboratories®, Silicon Labs®, SiLabs® and the Silicon Labs logo®, Bluegiga®, Bluegiga Logo®, Clockbuilder®, CMEMS®, DSPLL®, EFM®, EFM32®, EFR, Ember®, Energy Micro, Energy Micro logo and combinations thereof, "the world's most energy friendly microcontrollers", Ember®, EZLink®, EZRadio®, EZRadioPRO®, Gecko®, ISOmodem®, Precision32®, ProSLIC®, Simplicity Studio®, SiPHY®, Telegesis, the Telegesis Logo®, USBXpress® and others are trademarks or registered trademarks of Silicon Labs. ARM, CORTEX, Cortex-M3 and THUMB are trademarks or registered trademarks of ARM Holdings. Keil is a registered trademark of ARM Limited. All other products or brand names mentioned herein are trademarks of their respective holders.

SILICON LABS

Silicon Laboratories Inc.
400 West Cesar Chavez
Austin, TX 78701
USA

http://www.silabs.com

Soldering Guidelines

Solder0 R0.1

EE400/1
Design

Revision History

Rev.	Date	Contributor	
0.1	September 2005	L. Wyard-Scott	Updates to reflect changed Capstone Design Course format and new de/soldering equipment.
0.0	January 2004	V.J. Sieben, L. Wyard-Scott, E. Tiong	Creation.

Contents

1 Soldering Technique **1**
1.1 Required Tools . . . 1
1.2 Important Soldering Tips . . . 2
1.3 Precautions . . . 3
1.4 How to Solder Through-Hole Components . . . 4
1.5 How to Solder Surface-Mount Components . . . 8
1.6 Wire Connectors and Headers . . . 9
1.6.1 Crimp Connectors and Pins . . . 9
1.6.2 DB Connectors (Solder Cup Type) . . . 10
1.6.3 Joining Two Wires . . . 10
1.7 Testing Connections . . . 11
1.8 Milled Board Soldering . . . 12
1.9 Post-Soldering Cleanup . . . 12
1.9.1 PCB Cleanup . . . 12
1.9.2 Work Area Cleanup . . . 12

2 Desoldering Technique **13**
2.1 Solder Pump/Sucker . . . 13
2.2 Desoldering Braid/Wick . . . 14
2.3 Desoldering Iron . . . 15
2.4 Removing Components Effectively . . . 16

3 Glossary **17**

Bibliography **18**

A Pictures of Tools **19**

Chapter 1

Soldering Technique

This guide is intended to give an introduction to basic soldering technique. If the reader desires to pursue advanced soldering techniques, there are several books as well as technical institutions that offer detailed courses in the area of soldering.

1.1 Required Tools

Before discussing soldering techniques, one should must set up a “soldering work station”. The following is a list of tools for a typical soldering station. See Appendix A for pictures of the various parts listed below.

1. **Variable temperature soldering iron:** used for applying heat to joints during the soldering process.
2. **Damp sponge:** for cleaning soldering iron tip.
3. **Rosin-core solder:** to electrically and mechanically bond a component to the PCB.
4. **Wire cutters or side cutter:** for trimming component leads and stripping insulation from wires.
5. **Needle nose pliers:** for holding, placing and shaping components.
6. **Desoldering pump and/or desoldering braid:** for removing solder.
7. **Scotch tape and/or a “Third Hand”:** for securing components.

8. **Safety glasses:** for eye protection. These are mandatory in the lab.
9. **Magnifier:** to provide more detail during intricate work. A magnifying glass is convenient, but an illuminated magnifier is better.
10. **Light source:** to prevent eye-strain.
11. **Ventilation:** to extract and dispel fumes generated during the soldering process.
12. **Flux:** to clean components and PCB pads.
13. **Acid brush:** to assist in the removal of flux residue.

Prior to soldering it is a good idea to have all components organized as it will make populating the PCB more efficient. Make a BOM (bill of materials) for the PCB, and ensure before soldering that the components have been collected. The schematic and PCB layout will also be referred to when populating the board.

1.2 Important Soldering Tips

The following tips provide a quick guideline on how to make proper joints.

Cleanliness: All parts, including the soldering iron tip, must be clean and free from grease, oxidation and contamination. Solder does not flow over contaminated areas; moreover, solder is repelled by dirt. Severe contamination is evident when solder begins to “bead”. A common source of contamination is oxidation. Old components and copper boards will often have an oxide layer that prevents a good solder joint. Ensure all components have shiny leads and the PCB has clean traces. An abrasive such as a blue or pink eraser, emery paper, or stell wool can be used to remove the oxidized layer from the PCB board and components.

Tinning: In addition to being clean, the soldering iron tip must also be tinned (coated with solder). Tinning the tip allows solder to flow on the components more quickly rather than the soldering iron tip itself. Tinning involves adding a few millimetres of solder to the tip and then wiping and rotating the tip on the damp sponge to reveal a shiny surface on the tip of the soldering iron: a thin layer of solder will coat or “tin” the tip of the soldering iron. **When done soldering,**

tinning the iron is required to protect the tip from oxidation thereby dramatically increasing its life.

Temperature: Ensure that both the component leads and the PCB's copper layer are heated at the same time. The soldering iron tip should contact both the component and the PCB pad. This will ensure that each surface is relatively close in temperature resulting in a good joint. If there is a temperature difference between the two surfaces, the solder will form a "dry" joint. Soldering irons are typically set around 650 Fahrenheit, depending on the lead-tin ratio of the solder being used. Too much heat causes excessive "sputtering" of flux, and too little doesn't melt the solder in a timely manner.

Duration: The duration that the iron is in contact with the component and PCB is dependent on the size of the joint and your soldering iron temperature. For the typical PCB through-hole joint, it should take a few seconds to heat the joint and apply the solder. This will require practice, so don't expect to be fast if you are a beginner. Excessive heat (several seconds in duration) will damage sensitive semiconductors. If this is a concern, use a heat sink attached to the component leads: sometimes as simple as an alligator clip. These concerns can sometimes be avoided by soldering sockets instead of the semiconductor itself.

Adequate solder coverage: If too little solder is applied, the joint will not make a secure connection and will cause erratic behaviour. However, if too much solder is applied, the joint may bridge with adjacent joints resulting in electrical shorts. How much solder to apply comes with experience. Figs. 1.4 through 1.7 show good and bad solder joints.

Handling: Most modern electronics systems contain static-sensitive devices. Use proper handling procedures to minimize the likelihood of damage: grounding wrist-straps, grounded soldering irons, grounding mats, etc.

1.3 Precautions

Soldering Irons get very hot (600-800^{0}F, 315-425^{0}C), please ensure you follow precautions during use. Basic safety precautions are listed below.

- Never leave your iron turned on while unattended.

- Turn the soldering iron off when it is not being used. If the iron is left on for long periods of idle time, the soldering iron tip will be destroyed through oxidation.
- Eye protection must always be worn when soldering. Hot flux can spit up and into an unprotected eye. **In the Capstone Design Lab, use of eye protection is mandatory.**
- If the cord of the soldering iron is damaged, inform the lab staff who will ensure it is replaced.
- Never set the soldering iron down on anything other than an iron stand.
- To prevent burning your fingers, use needle nose pliers, heat resistant gloves, or a third hand tool to hold small pieces.
- Familiarize yourself with the safe handling of all materials used during the soldering process. This includes solder, flux, alcohol, and desoldering braid. Each has a Material Safety Data Sheet (MSDS) and can be found in the lab or online. "Safe Operating Procedures" are found posted on the wall close to the soldering facilities.

1.4 How to Solder Through-Hole Components

Most of the soldering done in the Capstone Design Lab is through-hole. A through-hole joint is a type of soldering joint in which the component joins with the PCB pad through a physical hole in the board. The following steps will illustrate how to make a proper through hole solder joint on a PCB.

1. Ensure that the printed circuit board and all components are clean. Cleaning can be achieved with a mild abrasive and/or the application of flux.
2. Plug in the soldering iron, turn it on, and let it warm up for 2–3 minutes.
3. Wet the soldering station sponge with the water provided in the lab. Do not wet the sponge in the bathroom or the water fountain.
4. Clean the tip of the soldering iron and tin it with solder.

5. Insert the component into the holes. Ensure that the component is secure by taping the component or by using a third hand. Optionally, the component leads can be clinched as shown in Fig. 1.1. This technique, however, is not recommended for two-sided boards as the flow of solder to the component side is restricted.

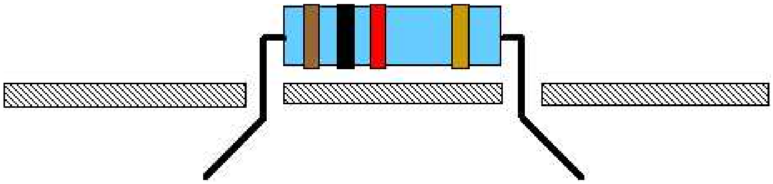

Figure 1.1: Clinched Component Leads

6. Apply the soldering iron tip to one side of joint making contact with the component lead and the board copper foil, ensuring that both are heated up to the same temperature as shown in Fig. 1.2 and Fig 1.3. Notice the tinned tip in Fig. 1.3.

7. Slowly add a few millimetres of solder to the other side of the joint. DO NOT apply solder to the soldering iron tip. If enough heat was applied to the PCB pad and component wire, the solder will flow freely onto the joint.

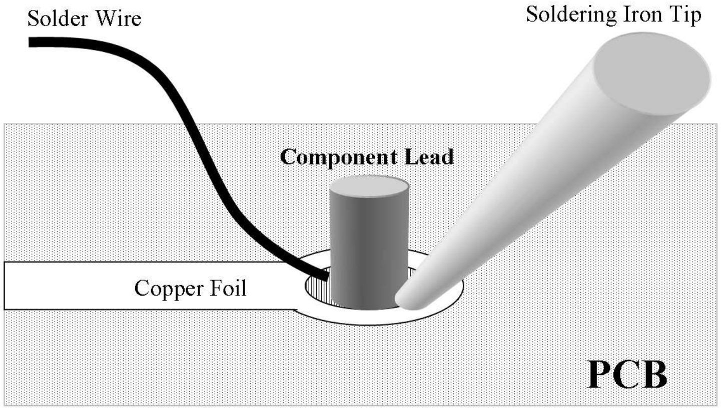

Figure 1.2: Method of Soldering Through Hole Joints

8. Remove the solder when the joint is suitably covered as shown in

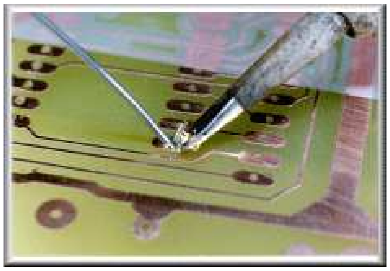

Figure 1.3: Method of Soldering Through Hole Joints, [8]

Fig. 1.4. The goal is to get the joint to be a “fillet”: a curve as shown in Fig. 1.4.

9. If the PCB is double-sided, the solder should flow through the hole around the component lead and make a bond on the component side of the board (opposite to the side that the solder was applied). If this “wicking” does not occur, the hole may be undersized, clinching could be blocking the solder’s path, or the component lead is not clean.

10. Remove the soldering iron and allow the joint to cool naturally.

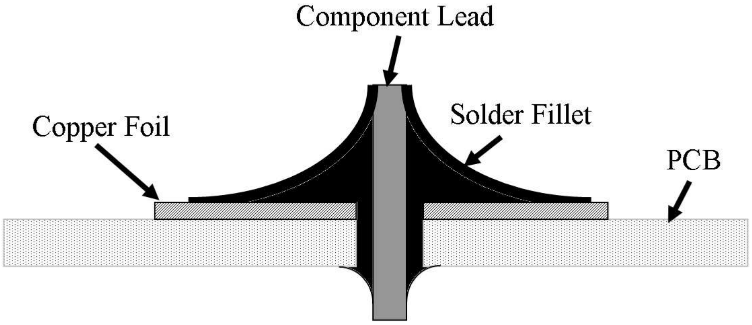

Figure 1.4: Good Solder Joint - Solder Fillet

11. Cut the lead of the component, if necessary.

Fig. 1.6 and Fig. 1.7 below are pictures showing good and bad solder joints.

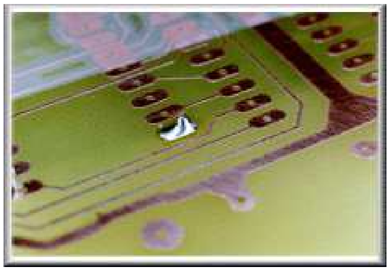

Figure 1.5: Good Solder Joint, [8]

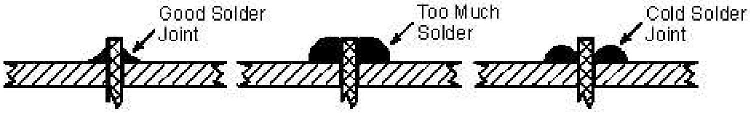
Good Solder
Joint
Too Much
Solder
Cold Solder
Joint

Figure 1.6: Joint Examples, [3]

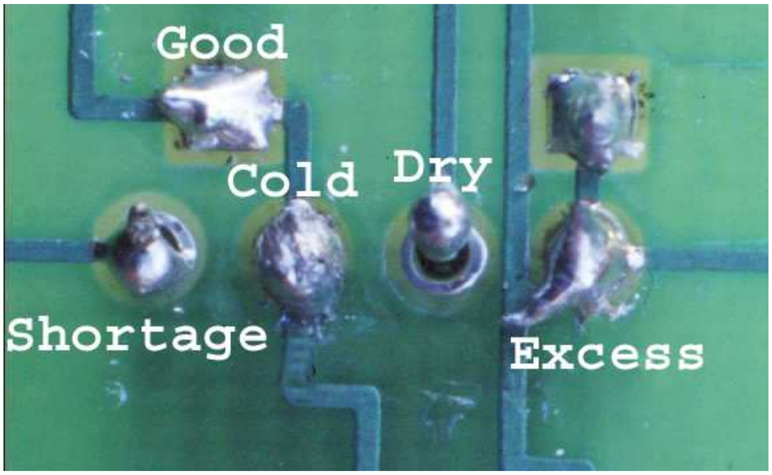
Good
Cold
Dry
Shortage
Excess

Figure 1.7: Joint Examples, [2]

1.5 How to Solder Surface-Mount Components

Surface mount soldering requires more experience and skill than through hole. It is recommended that one practices with through-hole prior to attempting any surface mount soldering. As the name suggests, surface mount involves soldering a component to either the top or bottom surface of a PCB. Depending on the footprint, the pads are usually a spaced closer together (finer pitch), making the soldering more susceptible to solder bridges, etc.

Figs. 1.8 and 1.9 show examples of good surface-mount solder joints.

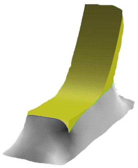

Figure 1.8: Surface Mount Good Solder Joint, [6]

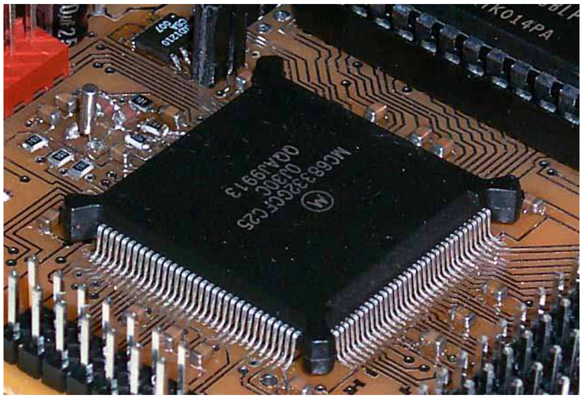

Figure 1.9: Surface Mount IC with Good Solder Joints, [5]

The actual soldering of the joints is similiar to the through-hole method. One difficulty, however, is maintaining the part's alignment on the PCB pads. A good technique is outlined here:

1. Align the component on the PCB pads. This can be aided with the use of tweezers and dental picks.
2. Secure the component to the PCB by applying a small amount of pressure onto the top of the component using a small slot screwdriver.

An index finger resting on the end of the screwdriver provides enough force to secure the device.

3. Solder one of the corner component leads to the PCB pad.
4. Align the remaining pads and solder the opposite corner PCB pad.
5. Solder the remaining pads in a pattern that does not build-up too much heat in the device.

1.6 Wire Connectors and Headers

When PCBs are manufactured they often have connectors to peripheral devices. These connectors are like other components in how they are soldered onto the PCB. However, the plug that matches the connector usually also requires some soldering. **Please note that with very few exceptions, the wire used for cabling is stranded (rather than solid-core) due to its higher strength and flexibility.** Below are procedures for a few of the more common plug types.

1.6.1 Crimp Connectors and Pins

1. Strip off about 1cm of the wire insulation.
2. Place the exposed wire into the crimp-style connector or pin. The wire should just barely show coming out the other side. Some crimp pins have two crimping areas – one for the stripped wire, and the other for the wire *with* the insulation.
3. Crimp the connector with the appropriate crimping tool. Note that the connectors are colour coded to settings on the crimping tool, try and match the correct colour.
4. As added insurance, apply solder to the joint to solidify the connection.

Before crimping all pins destined for a connector housing, start with one and ensure that it fits successfully into the housing. If the crimp is made too tightly, or otherwise mis-shaped, the pin may no longer fit.

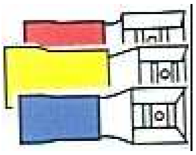

Figure 1.10: Colour Coded Crimp Connectors

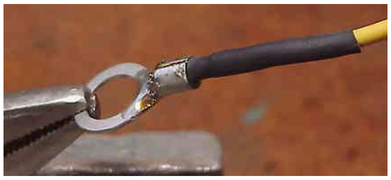

Figure 1.11: Crimp connector without shrink wrap, [1]

1.6.2 DB Connectors (Solder Cup Type)

1. Strip off about 5mm of the wire insulation.
2. Tin the wire.
3. Slide shrink wrap on the wire.
4. Place the wire into the DB - connector solder end.
5. Apply the soldering iron tip to the connector and wire.
6. Apply a few millimetres of solder until the joint is adequately covered.
7. Let the joint cool.
8. Slide the shrink wrap over the joint and heat the shrink wrap with a heat gun (or other heat source) until the wrap shrinks around the joint.

1.6.3 Joining Two Wires

1. Slide shrink wrap on one of the wires.
2. Tin or coat each wire with some solder.
3. Join the wires in a way that provides a good mechanical connection (such as twisting together) and apply the soldering iron tip and some solder.

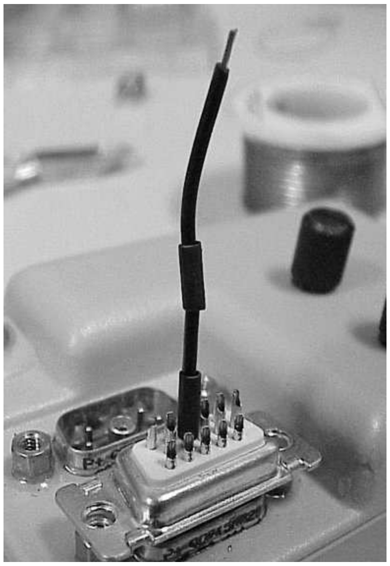

Figure 1.12: DB Connector with shrink wrap, [7]

4. Let the joint cool.
5. Slide shrink wrap over the joint and heat the shrink wrap with a heating gun or a heat source until it the the wrap shrinks around the joint.

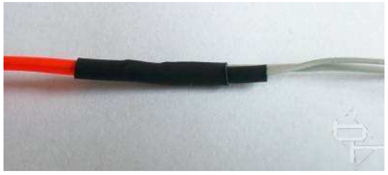

Figure 1.13: Two wire joint insulated by shrink wrap, [4]

1.7 Testing Connections

After completely soldering a component to a PCB, it is good practice to ensure connectivity between the component wire leads and the PCB pads they are soldered to. A DMM (Digital Multi-Meter) is sufficient to determine connectivity; many DMMs include an audible connectivity setting, but failing this, measure the joint resistance.

1.8 Milled Board Soldering

In-lab manufacture of PCBs uses a technique known as "milling". The milling technique involves cutting out the PCB tracks and pads from the copper-clad board.

Due to its construction, a milled PCB is susceptible to solder-bridging across the milled grooves, particularly when an excess of solder is used. Finding a short caused by a bridge is a difficult task, particularly when many solder joints exist. To this end, make a limited number of solder connections and then test for bridges using a multimeter.

1.9 Post-Soldering Cleanup

1.9.1 PCB Cleanup

The flux left behind by rosin-core solders, or perhaps as part of the cleaning process, needs to be removed from the PCB. Due to the flux's sticky nature, dirt gathers and contributes to short-circuiting problems. The most frustrating part of this situation is that a short does not necessarily occur immediately. Weeks, or even years later, a short can develop.

To remove flux, alcohol is used. Apply the alcohol liberally and then brush away with an acid brush, starting at the center of the PCB and working out toward the edges. This is a time-consuming task, but a clean board is well-worth the effort.

1.9.2 Work Area Cleanup

The importance of keeping your work area clean cannot be emphasized enough. When clear of obstructions and garbage, handling a hot soldering iron is safer. The soldering process itself involves chemicals and substances which are known to have ill-effects in humans. Wiping-down the work-area surfaces with a moist paper towel will help reduce some contamination. When you are done soldering, **wash your hands with soap and water** to get rid of contamination. The primary concern here is accidental ingestion of the chemicals. For more information, please refer to the Safe Operating Procedures, posted in the lab's soldering area.

Chapter 2

Desoldering Technique

Desoldering may be required for several reasons:

- a component may have failed
- a wrong part was installed;
- a design modification necessitates a change; or
- if a board contains expensive components that can be salvaged.

Whatever the reason, there are three common techniques to remove solder from a joint: using a desoldering pump, a desoldering wick, or desoldering iron.

Regardless of the method used, if it is permissible to destroy the part during removal, then a lot of time and effort can be saved. For instance, using wire cutters to trim off all the pins of a through-hole IC so they can be removed individually makes the removal process much easier. In most cases, avoiding damage to the PCB is of paramount importance.

2.1 Solder Pump/Sucker

A desoldering pump is exactly what it sounds like: a pump that sucks up solder. Usually the pump is spring-loaded and provides a recoil when released. The four steps below outline how to desolder a joint using a desoldering pump.

1. Prime the desoldering pump. This involves depressing the desoldering pump spring. The pump will click when correctly depressed.

2. Heat the joint from one side with the soldering iron tip. Wait 1-2 seconds until the solder begins to melt from the soldering iron heat.
3. Put the pump tip on the other side of the joint. Don't be afraid to actually touch the joint.

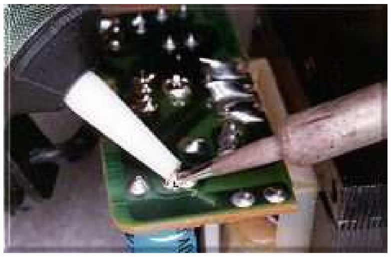

Figure 2.1: Desoldering with a Solder Pump, [8]

4. Press the desoldering pump spring-release button to suck up the solder.

This procedure should be repeated if the joint has a significant amount of solder. If done correctly the joint should eventually look as shown in Fig. 2.2 shown below.

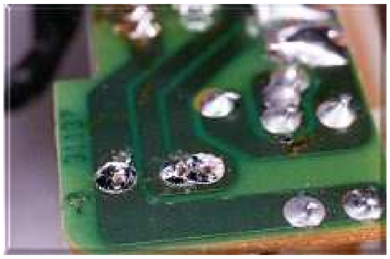

Figure 2.2: Clean Joint from Desoldering with a Solder Pump, [8]

2.2 Desoldering Braid/Wick

A desoldering braid removes solder from a joint using a technique known as "wicking". Desolder braid is a piece of material that sponges up molten solder by capillary action that draws solder away from the joint the braid. The following procedure outlines the basic steps involved in desoldering a joint using desoldering braid.

1. Place the desoldering braid over the joint.
2. Press the desoldering braid onto the joint with the soldering iron tip. This will apply heat to the desoldering braid and the joint allowing the molten solder to flow.
3. Wait for the solder to melt. The solder should flow onto the braid and away from the joint.

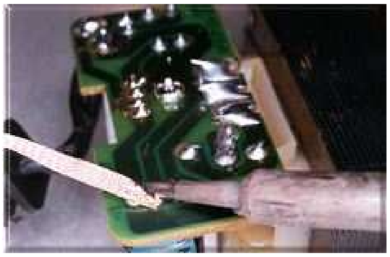

Figure 2.3: Desoldering with a Solder Wick, [8]

4. Cut off the solder coated portion of the desoldering braid. **There should be no copper visible in the portion that is removed: braid is very expensive.**

2.3 Desoldering Iron

Using a desoldering iron, available with higher-end soldering stations, is similar to using a solder sucker. The desoldering iron is essentially a soldering iron with a built-in vacuum.

1. Make contact between the iron and the joint to be desoldered, ensuring the vacuum opening is not blocked.
2. Once the solder on the joint has become molten, depress the button on the iron to activate the vacuum.

It is very important for the life of the desoldering iron that the vacuum assembly be cleaned after every session of use. Please ask the lab personnel for a description of the cleaning procedure.

As with a normal soldering iron, the desoldering iron tip should be tinned before and after every use.

2.4 Removing Components Effectively

To remove a component with little or no damage to the PCB or component takes practice and **patience**. Often when removing a component, PCB pads and tracks may be damaged by "lifting" off the board. Below are some tips that may be useful when removing components from a PCB.

- Desolder all joints pertaining to a component prior to removal. It may be necessary to go over the joints several times before all the solder is removed. Practice patience.
- Use pliers and *gently* pull on components while applying heat to the joints. The assistance of a friend can help with this.
- Do not pull with force, because the PCB pads and tracks will be damaged.
- Do not try and remove components by prying them.
- Do not try to push the component out of the holes with the soldering tip. This will certainly lift the copper pad off of the PCB and ruin the iron tip.
- In order to get better thermal conductivity between the iron and the solder joint, it is sometimes effective to *add* a small amount of solder.

Chapter 3

Glossary

BOM: Bill of Materials, which contains a list of all components and values contained on a particular PCB. Used as a “shopping list”.

Desoldering Braid: A material used to remove solder with capillary action.

Desoldering Pump: A device used to remove solder with a swift vacuum action.

Flux: Cleans the surfaces that are being heated by bringing contaminants to the surface. Most solders include flux in their core. For excessive contamination use a flux pen.

Footprint: The spacing pattern or layout of pads for a particular component or integrated chip, as used on a PCB.

PCB (Printed Circuit Board): A fibreglass board upon which copper traces are laminated to make connections between various components. The copper traces serve as flat wires connecting various components.

Shrink Wrap: A tube that fits over a wire, that when heated shrinks to provide insulation and support for the joints.

Bibliography

[1] R. Glass. Electrical wire soldering for beginners. Technical report. http://www.airheads.org/contrib/solder.html.

[2] D. Lauder. How to solder. Technical report, Oct 2001. http://dragon.herts.ac.uk/ eleqdml/teaching/general/soldering/.

[3] W. M. Leach. The leach amp. Technical report, 2000. See section on Assembly of Circuit Boards, http://users.ece.gatech.edu/ mleach/lowtim/part2.html.

[4] Macroman. Cutting neon string. Technical report. http://www.bit-tech.net/article/59/.

[5] K. Maxon. Have you seen my new soldering iron? Technical report. http://www.seattlerobotics.org/encoder/200006/oven-art.htm.

[6] NIST. Archive of surface evolver code. Technical report. http://www.ctcms.nist.gov/solder/archive.html.

[7] Serial Tester. Serial tester. Technical report. http://army-gps.robins.af.mil/tech/Serial-tester/proc06.jpg.

[8] A. Winstanley. The basic electronics soldering and desoldering guide. Technical report, Feb 1999. http://www.epemag.wimborne.co.uk/solderfaq.htm.

Appendix A

Pictures of Tools

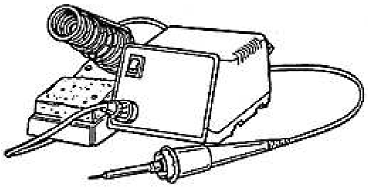

Figure A.1: Soldering Iron

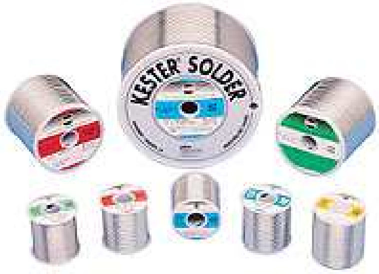
KESTER SOLDER

Figure A.2: Solder Wire

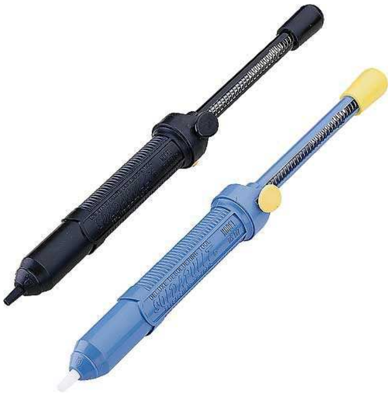

Figure A.3: Desolder Pump

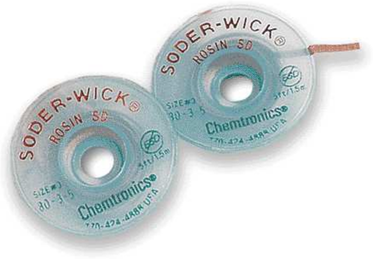
SODER-WICK®
ROSIN SD
Chemtronics®
SODER-WICK®
ROSIN SD
Chemtronics®

Figure A.4: Desolder Braid

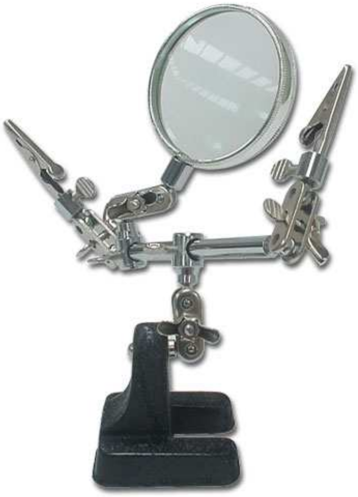

Figure A.5: Third Man

1

Soldering

Contents

Introduction 2
1 Soldering Printed Circuit Boards.. 3
2 Types of Solder.. 4
3 Types of PCB 5
  3.1 Through-Hole Components 5
  3.2 Surface Mount Components 5
4 Flux.. 6
5 Soldering Irons 7
  5.1 Tips 8
  5.2 Tip Contamination and Cleaning.. 9
6 Ovens and Hot Plates.. 10
7 Desoldering... 11
8 Hazards involved in soldering 12
  8.1 Heat... 12
  8.2 Toxic materials 12
  8.3 Spattering.. 13
  8.4 Electrical Safety.. 13

2

Introduction

Soldering is a process used for joining metal parts to form a mechanical or electrical bond. It typically uses a low melting point metal alloy (solder) which is melted and applied to the metal parts to be joined and this bonds to the metal parts and forms a connection when the solder solidifies. It is different to welding in that the parts being joined are not melted and are usually not the same material as the solder.

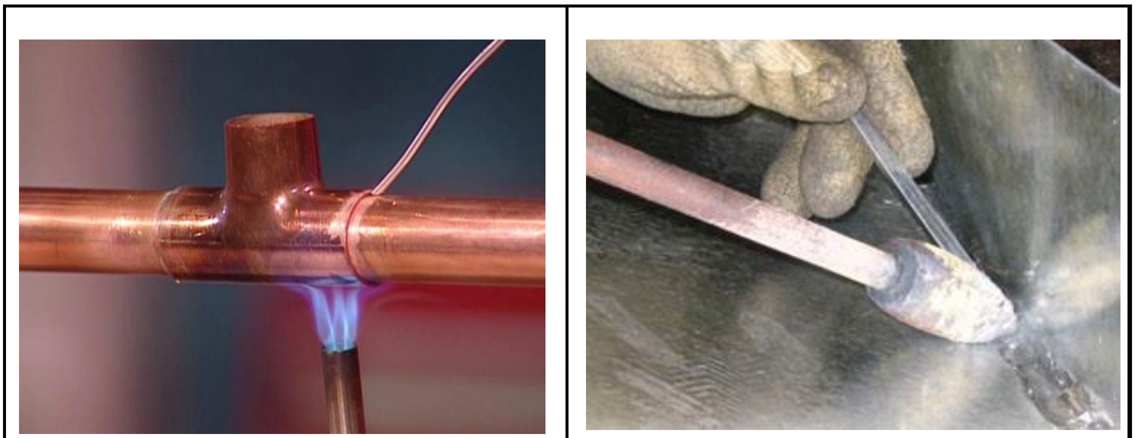

Soldering copper pipe

Sheet metal fabrication

Figure 1 – Different Types of Soldering

Soldering is a common practice for assembling electrical components and wiring. Although it can be used for plumbing, sheet metal fabrication or automotive radiator repair the techniques and materials used are different to those used for electrical work. This document is intended to provide guidance on the safe working methods and proper tools and techniques for soldering of electrical components.

1 Soldering Printed Circuit Boards

Soldering may be used to join wires or attached components to a printed circuit board (PCB). Wires, component leads and tracks on circuit boards are mostly made of copper. The copper is usually covered with a thin layer of tin to prevent oxidization and to promote better bonding to other parts with solder. When soldering bare copper wires they are often “tinned” by applying molten solder before making a joint.

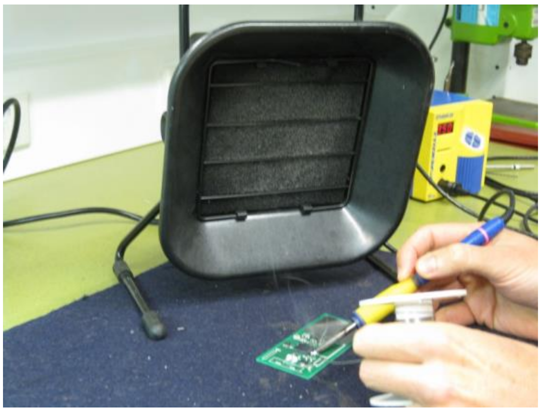	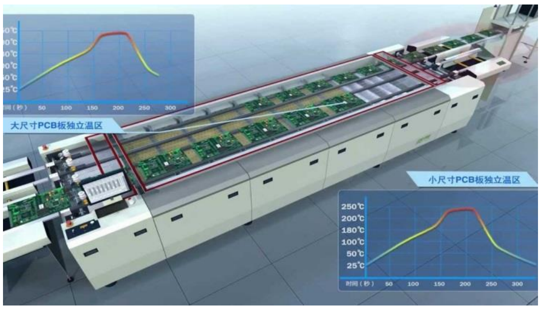 大尺寸PCB板独立温区 小尺寸PCB板独立温区 250℃ 200℃ 180℃ 100℃ 50℃ 25℃ 时间(秒) 50 100 150 200 250 300
Hand soldering components on PCB	Industrial oven for surface mount

Figure 2 – Different Types of PCB Soldering

2 Types of Solder

There are different types of solder used for electrical work. They are broadly classified as tin/lead solders or lead free solders. Tin/lead solders have been used for many years because of their ease of use however they have been phased out of commercial use due to the harmful effects on humans and the environment. Tin/lead solder is still available and is used by "hobbyists" and other non-commercial users as it is still easier to use than lead free types. When using tin/lead (or leaded) solder there are additional safety precautions that must be observed.

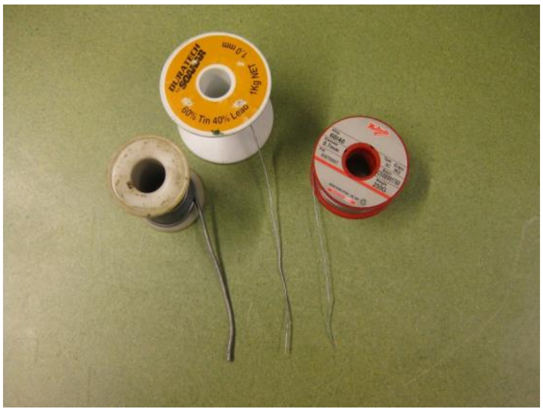	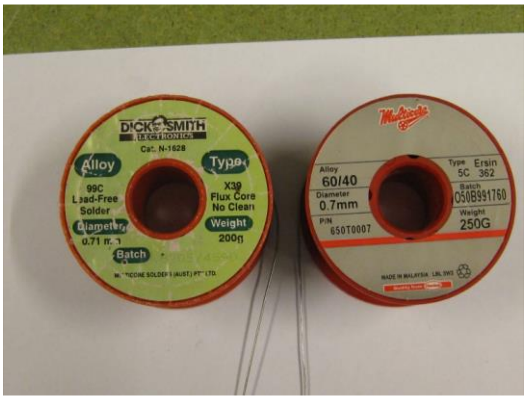
Different gauges of solder wire	Solder composition is labelled (Lead free on left)

Figure 3 – Different Types of Solder

3 Types of PCB

Printed circuit boards (PCBs) are populated by electronic components and these may be “surface mount” or “through-hole” types.

3.1 Through-Hole Components

As the description “through-hole” suggests, the leads of the component are passed through holes in the PCB and then soldered to a “pad” on the reverse side of the PCB. Soldering is accomplished by heating the component lead and PCB pad with a soldering iron and melting solder wire into the joint. This type of construction was common from the 1960’s until early 2000’s and is still used by hobbyists and in small scale production where manual assembly is preferred.

3.2 Surface Mount Components

Commercial circuits are mostly of the surface mount type as these are cheaper to make, more compact and easier to automate assembly. For surface mount construction the component’s pads are on the same side of the PCB as the component and the component connections sit onto these pads. Soldering is accomplished by applying solder paste onto component pads on the PCB, placing the component onto the paste and then heating the entire assembly to melt the solder. Commercial assembly uses ovens to heat the boards. Hobbyists can also use surface mount components and soldering can be accomplished by applying solder paste and melting with a hot plate, small oven or soldering iron. Some surface mount joints can be soldered using a soldering iron and solder wire.

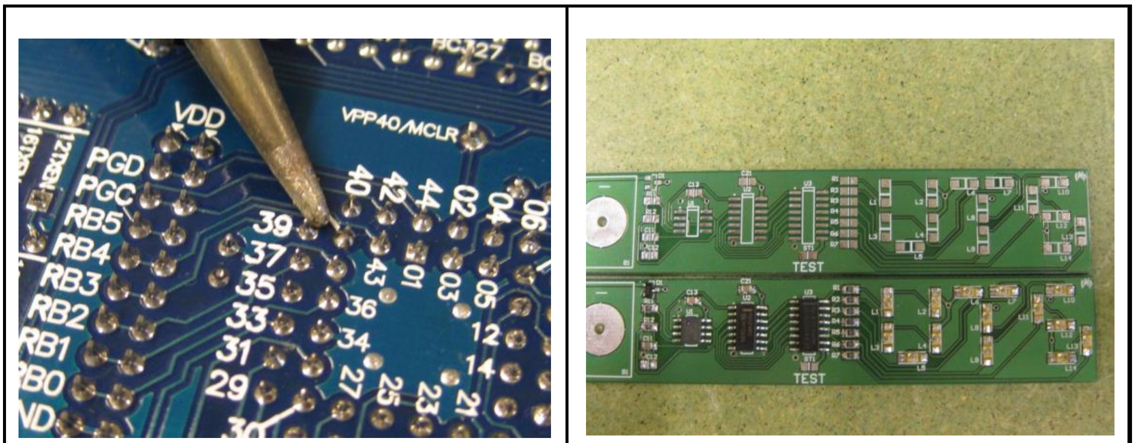

Through-hole PCB
(non-component side)

Surface mount PCB
(component side)

Figure 4 – Different Types of PCB

4 Flux

For electrical soldering both solder wire and solder paste contain flux. This helps to clean the surfaces being soldered and prevent oxidization of the hot solder. The composition of the flux will vary depending on whether it is in a paste or wire, leaded or unleaded solder. Solder wire usually contains a flux called “rosin”. Most fluxes will produce fumes when the solder is heated and these fumes are likely harmful to your health. For occasional soldering it may be sufficient to have a well-ventilated workspace but for longer or repeated exposure a fume extractor should be used. Solder flux can also cause solder to spatter and eye protection should be worn when soldering.

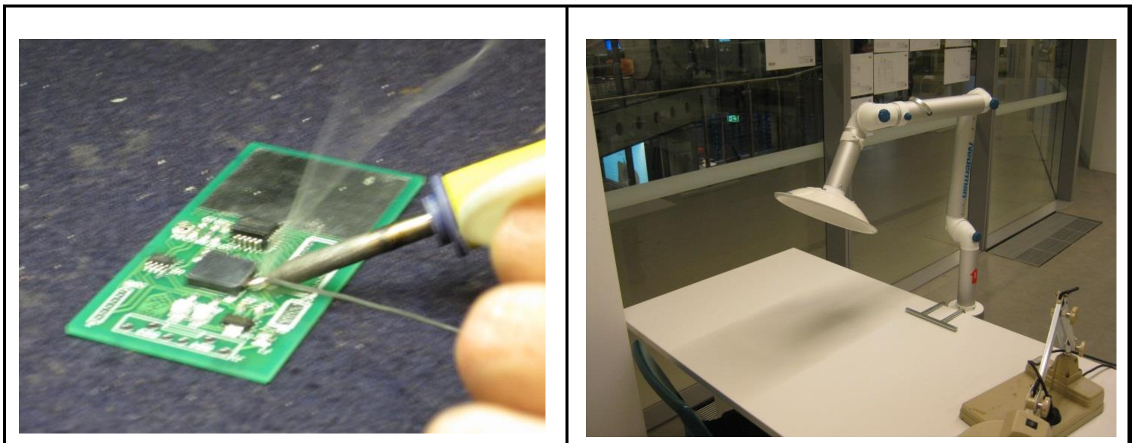

Fumes produced when soldering

Workstation with fume extraction system

Figure 5 – Fumes and Fume Extraction Systems

5 Soldering Irons

Soldering irons come in many varieties and sizes. Soldering irons may be electric, gas powered or externally heated. Most common types are electric. Simple electric soldering irons have no controls and you simply plug them in and wait for them to heat up. Their temperature is regulated by the power of the heating element and heat loss to the environment. Some soldering irons have temperature controls which allow the user to set a desired operating temperature for the soldering iron. This is useful if the soldering iron is being used for different types of solders which have different melting points or if the soldering iron is being used for other purposes such as heating heatshrink. It also introduces a problem if the user does not set an appropriate temperature for the work, solder can be overheated and decompose. Hotter is not better! A temperature of around 320 °C works well for 60/40 leaded solder. Some temperature controlled soldering irons use interchangeable tips to change the temperature at which they operate.

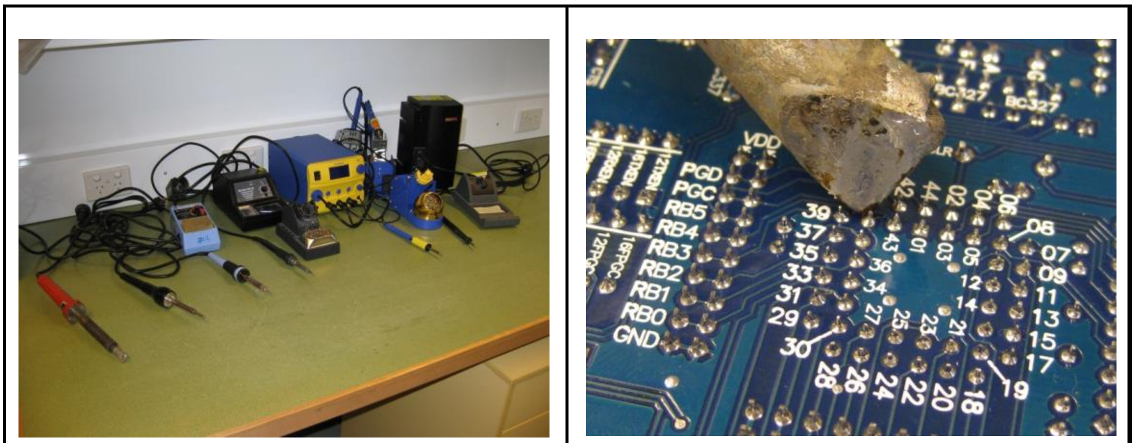

A range of electric soldering irons

Choose an iron appropriate for your task!

Figure 6 – Types of soldering irons

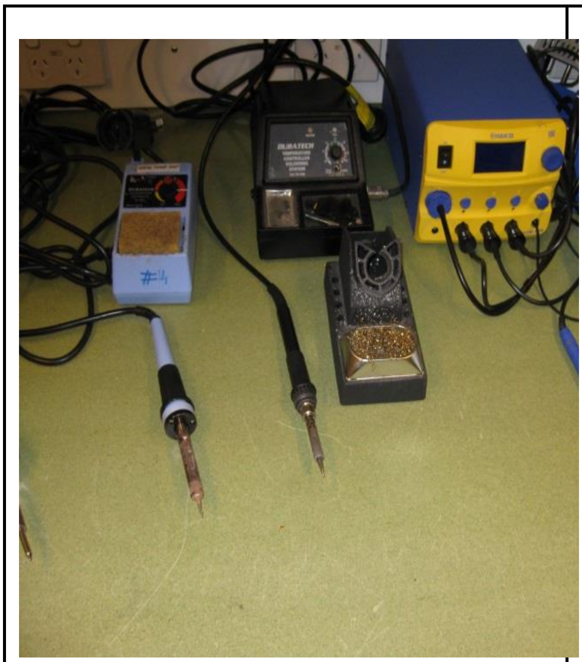	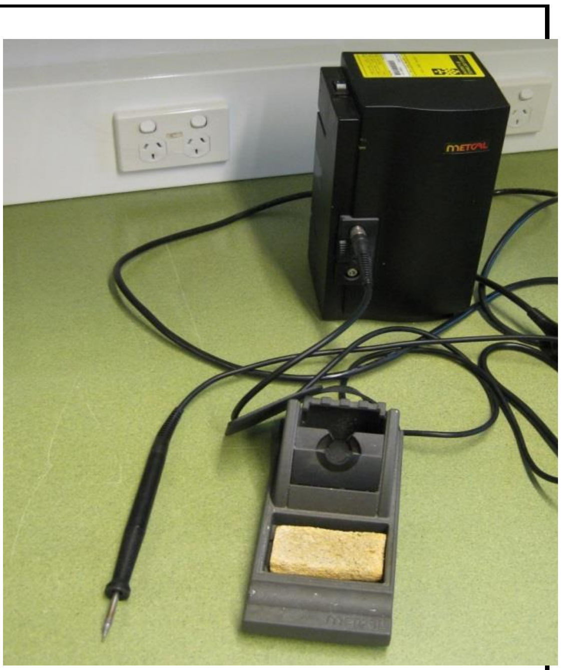
Adjustable temperature soldering stations	Different tips set the temperature for this station

Figure 7 – Types of soldering stations

5.1 Tips

Heat is transferred from the tip of the soldering iron to the joint by thermal conduction enabled by metal to metal contact between the tip and joint. The tips of soldering irons come in various shapes and sizes to enable the best contact to be made. Most tips are either conical or chisel shapes. The shape is largely a personal preference and you can use whichever type works best for you. The size of the tip should be selected to allow the tip to be placed against the joint being soldered without interfering with adjacent parts. The tip should be large enough to conduct sufficient heat into the joint to allow the solder to melt and flow properly. The choice of tip size is not a precise calculation and a “normal” size tip will work for most joints on a PCB.

5.2 Tip Contamination and Cleaning

The thermal conduction from the tip to the joint may be inhibited by contamination on the tip. This contamination can be formed by burnt solder flux or oxidized solder. To make best thermal contact the tip should be cleaned using a tip cleaner(!). Two types are a wet sponge or a brass wire wool. The wet sponge removes the contaminated material when the tip is wiped across it, the water in the sponge cools the solder and the mechanical abrasion removes the contamination leaving a thin coating of clean solder on the tip. This method can cause the tip temperature to dip momentarily. The brass wire wool type removes the contamination by mechanical abrasion and bonding contaminated solder to the brass. The tip is pushed into the brass wool and when it is withdrawn the tip is clean with a thin coating of solder. You must not “wipe” the tip on the brass wool type because the springiness of the brass wool may flick molten solder which may cause burns to people or objects. You should never “flick” excess solder from the soldering iron as this may also cause burns or damage.

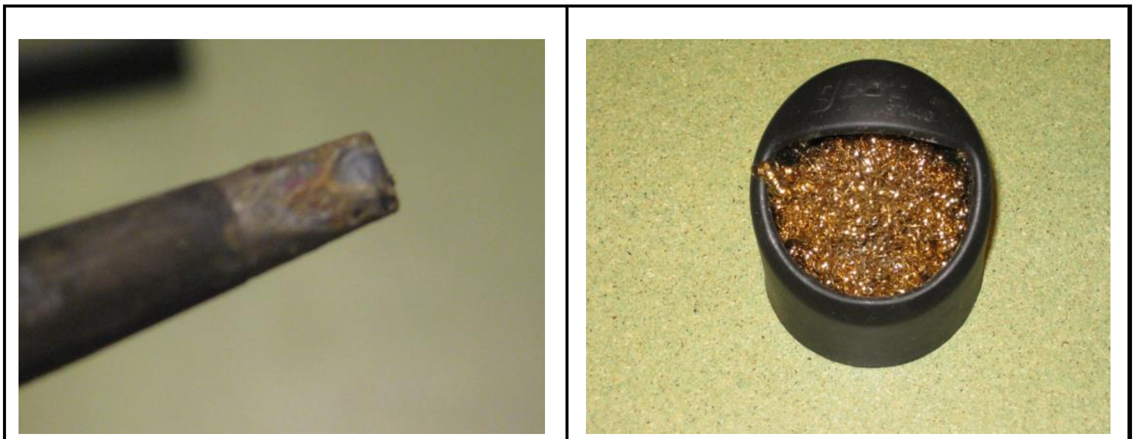

Tip contaminated with old solder and burnt flux

Brass wool type tip cleaner

Figure 8 – Dirty tip and brass wool cleaner

6 Ovens and Hot Plates

For surface mount soldering the heat is usually applied to the whole PCB and all components soldered at the same time. For commercial work this is done in large ovens, often with conveyors to move the boards through the oven. For small scale work simple infrared ovens are available. Another technique uses a hotplate. The same considerations for temperature apply, although because the heat source does not come into contact with the solder or flux, contamination is less likely. Time and temperature are considerations with these methods as the components are exposed to the high temperature for the period required for the solder paste to melt and flow.

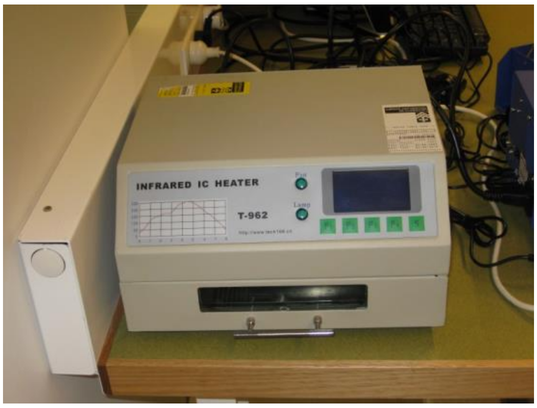	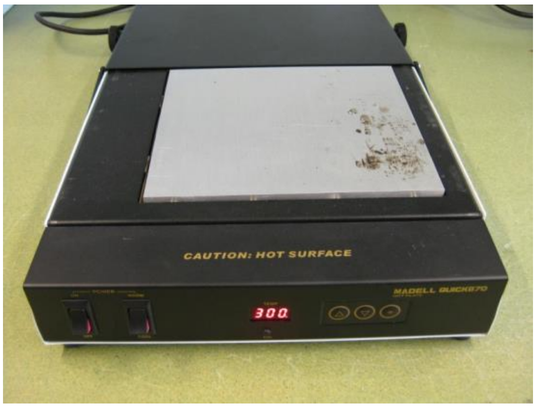
Infrared oven for surface mount soldering	Hotplate for surface mount soldering

Figure 9 – Oven and hot plate used for surface mount soldering

7 Desoldering

If a part that has been soldered needs to be replaced it needs to be “de-soldered”. Depending on the part and type of joint it may be possible to simply re-melt the solder and remove the part, or it may be necessary to remove the solder from the joint so the part can be freed. Some methods for removing solder are solder wick, solder sucker or de-soldering tool. Solder wick is a copper braid which is applied to the joint and heated with a soldering iron. As the solder in the joint is melted it is drawn into the solder wick like a sponge and is removed from the joint. A solder sucker is a spring loaded syringe or rubber bulb. The tip of the solder sucker is placed near the joint as the joint is melted by a soldering iron. When the sucker is operated a vacuum is created which draws the molten solder from the joint into the body of the sucker. A de-soldering tool is a type of soldering iron with a hollow tip and is connected to a pump or vacuum source. The tip of the de-soldering tool is placed onto the joint, typically over a component lead, and once the solder has melted the pump is operated to draw the molten solder away.

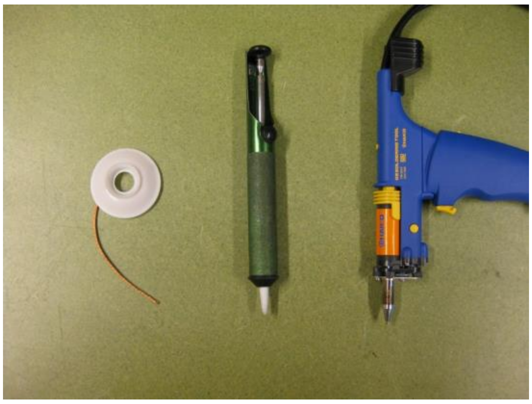	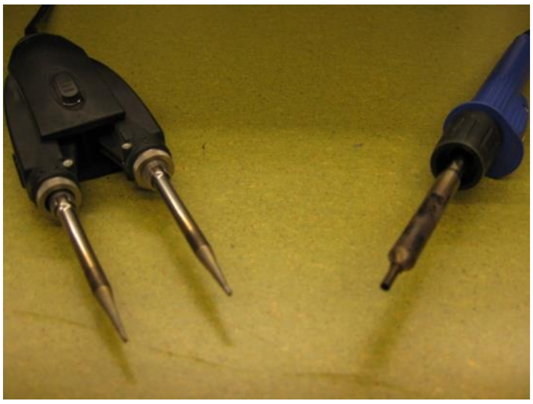
Solder wick, syringe type sucker, vacuum de-soldering tool	Hot tweezers or Hot air nozzle are options for SMDs

Figure 10 – Desoldering tools and tweezers for surface mount devices

8 Hazards involved in soldering

8.1 Heat

Although solder has a relatively low melting point this temperature is more than high enough to cause serious burns to people or objects. It is important to hold the soldering iron only by the insulated handle, never touch the heating element or tip when the soldering iron is on. The soldering iron will remain hot for some time after it is turned off so always check that it has cooled down before touching it, e.g. if changing the tip. When you are not soldering always keep the soldering iron in a proper holder so that you don’t touch it accidently and it doesn’t heat or burn other objects such as the benchtop. Don’t hold parts being soldered with your hands as these will also be heated when being soldered. Don’t flick molten solder from the soldering iron or wipe the tip on brass wool type tip cleaners.

If using a hot plate for surface mount soldering do not touch the hot plate. Use utensils such as pliers to place and remove PCB’s from the hotplate. If using hot air tools for soldering, de-soldering or rework, do not direct the hot air stream onto yourself or other people. If using an oven allow the PCB to cool before handling or use utensils. Don’t place hot PCB’s on temperature sensitive surfaces.

If burns occur they should be treated by holding under cold running water for several minutes and assistance sought if burns are severe. Incidents should be reported.

8.2 Toxic materials

Leaded solder contains lead which is a harmful material. Use of this type of solder will probably involve handling it and your skin may become contaminated by it. Although it is unlikely that the lead can be absorbed directly through your skin it may be ingested indirectly if it is transferred by handling food whilst your skin is contaminated. Always wash your hands thoroughly before eating or handling food.

Solder flux creates fumes when heated during soldering which may be harmful if inhaled. Use a fume extractor to avoid inhaling fumes.

8.3 Spattering

Solder and flux can spit or spatter when heated. Always wear eye protection (safety glasses) when soldering.

8.4 Electrical Safety

Electric soldering irons are plugin appliances and must have a current safety test tag. The test will confirm that the soldering iron conforms to electrical safety standards and has not been damaged at the time of the test. Before use you should visually check that the soldering iron does not have damage such as melted insulation on the lead, broken or cracked handle or exposed conductors. Don't use damaged equipment and report the damage.

For electrical safety the exposed metal parts such as the tip and heating element are earthed. Don't solder on any live equipment as contact with the earthed tip may cause damage to the equipment or soldering iron.

SMD Soldering

Information & Instructions

This is a walkthrough of a SMD Soldering Workshop. The following information and instructions can be used as a simple method for soldering SMD components. The recipient of this information should acknowledge the danger of soldering *(written on the back page of this pamphlet)* before attempting to solder.

Recommended Tools:

- Soldering Iron
- Solder
- Tweezers
- Solder Wick
- Flux Pen

① Add solder to one pad.

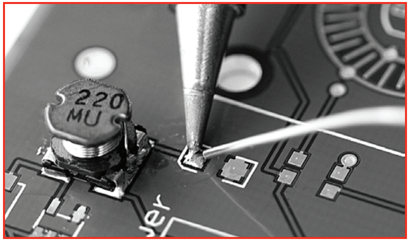

② While that pad is molten, slide the component into place. Do not push down from the top - slide the component into the blob of solder horizontally.

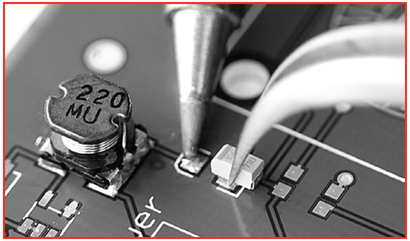

③ Align the component while connection is still molten.

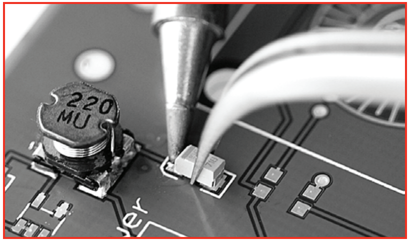

④ Once you have good alignment, continue to hold the component in place, and remove your iron. Continue to hold component for 1-2 seconds while the solder joint solidifies.

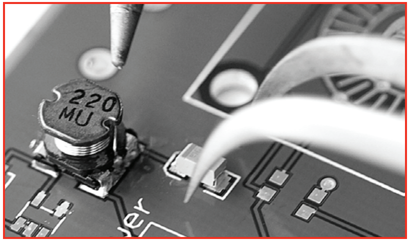

⑤ From above, the alignment looked good. From the side, you can see the rear pad is hovering slightly above the PCB. This can lead to problems on multi-pin components (open connections). Be sure the component is flush up against the PCB before soldering more connections. Re-grip the component, re-heat pad 1 and push the component flush against the PCB.

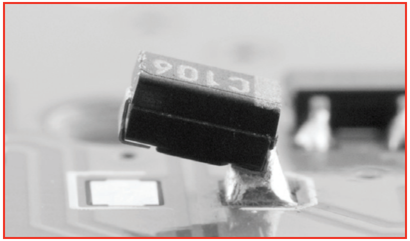

⑥ This is how a tantalum capacitor should look after making both solder connections.

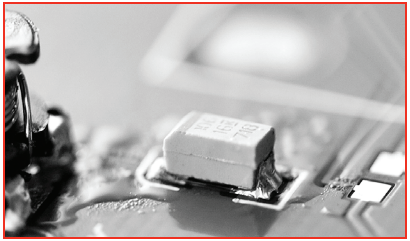

⑦ If alignment is not good, do not solder more than 1 pad! Re-heat the joint, re-adjust component until aligned correctly, then move on to soldering other connections.

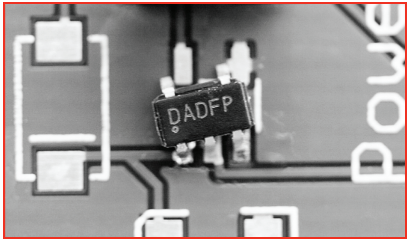

⑧ This is bad. It would be nearly impossible to finish the connections on this part. Make sure you have the component flush against the PCB.

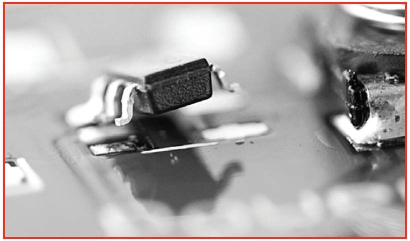

⑨ If you solder multiple pins together, don't worry about it! It can be easily fixed. Do not worry about jumpers! There are actually three pins under that blob.

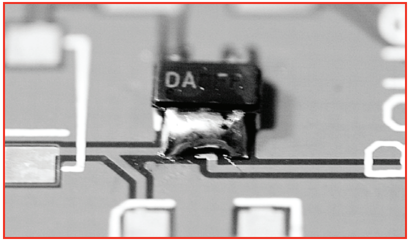

⑩ Pull out some solder wick. Put a small amount of solder on the end of your iron (this will transfer heat from iron to wick to the jumper). Sandwich the wick in between the iron and the solder jumper.

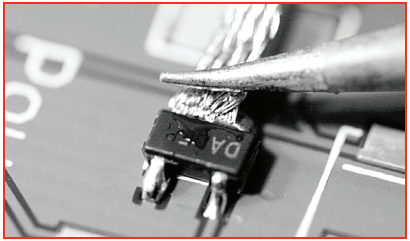

⑪ Hold still for 2-3 seconds. You will see solder start to flow up the wick. Once the excess solder has flowed into the wick, carefully lift up the wick and your iron in one fluid motion.

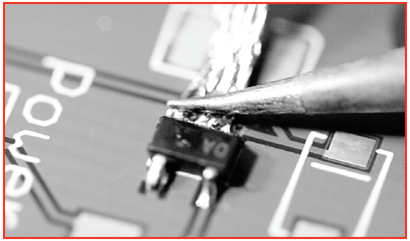

⑫ Nice and clean!

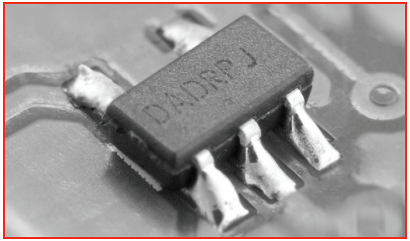

⑬ Bad bad bad. There was not enough solder for the connection on the left. Middle pin is lacking solder and should have been heated for longer. Right pin has had solder applied by an iron rather than applied to two metal contacts (the board and the pin).

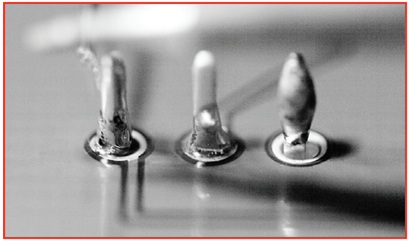

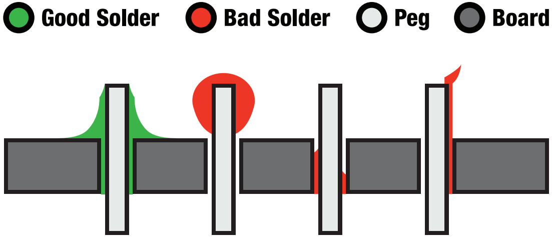

TIPS & TRICKS

A Solder flows smooth all the way to the bottom.

B Error: Solder balls up on top of pad, not connecting pin to pad.
Solution: Flux then wick. If you wick too much add solder.

C Error: Too little solder makes for weak connection.
Solution: Flux then add solder.

D Error: Bad Connection... and ugly... oh so ugly.
Solution: Flux then add solder.

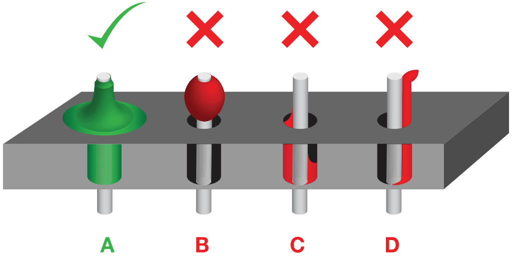

Safety Issues

Soldering

The tip of the iron is normally 700 °F, hot enough to melt metal. It is normal for the handle of the soldering iron to heat up a bit. Hold it like a pencil and move your hand further away from the tip if the heat is uncomfortable. The solder smokes because the rosin inside the solder is burning off - it's not harmful.

sparkfun.com

RoHS COMPLIANT PB

© SparkFun Electronics, Inc. All Rights Reserved. All other trademarks contained herein are the property of their respective owners.

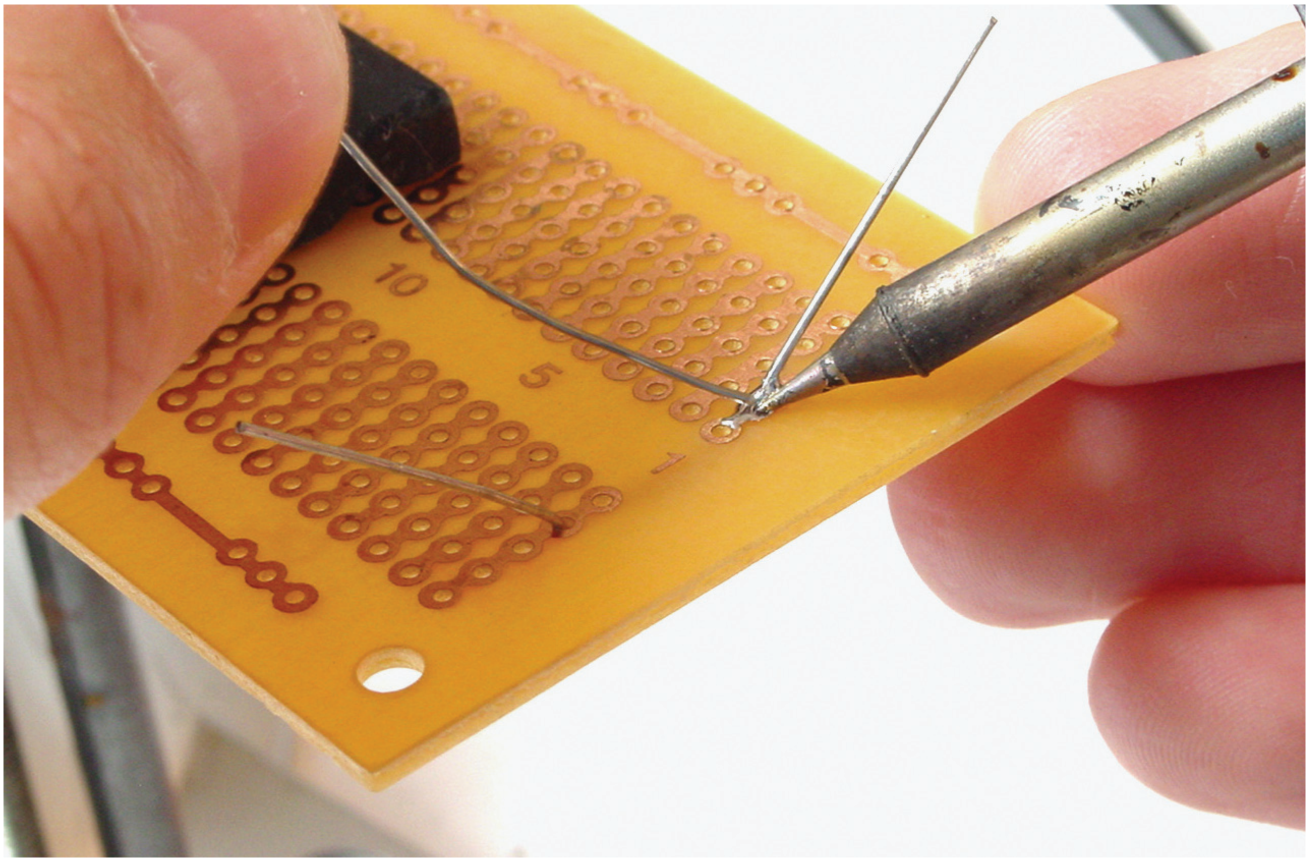

Photography by Joe Grand

Soldering and Desoldering

Step-by-step instructions for making (and unmaking) the perfect solder joint.

By Joe Grand

The two key parts of soldering are good heat distribution and cleanliness of the soldering surface and component. With practice, you'll become comfortable and experienced with the process.

In this primer, I'll explain how to solder a component onto a printed circuit board (PCB). I'll also provide desoldering tips and show you how to remove a surface-mount component from a printed circuit board using a Chip Quik kit. And I'll show you how to remove a component by removing the solder in a way that won't damage the components or the circuit board.

Reprinted with permission from *Hardware Hacking*, copyright 2004, Syngress Publishing, ISBN: 1-932266-83-6, pp. 34–40.

OVERVIEW

TOOLS AND TIPS

Tools of the Trade

Soldering iron You could pay as little as $10 or as much as $1,000 for a soldering iron. I recommend a fine-tip, 700°F, 50W soldering stick iron. A good general-purpose iron for hardware hacking is the Weller W60P Controlled-Output Soldering Iron, which sells for under $70.

Solder Should be thin gauge (0.032" or 0.025" diameter) 60/40 rosin core.

Desoldering tool (aka solder sucker) A manual vacuum device that pulls up hot solder, useful for removing components from circuit boards. I like the one RadioShack sells (#64-2098, $10).

IC extraction tool Helps lift integrated circuits from the board during removal/desoldering.

Chip Quik SMD Removal Kit Allows you to remove surface mount components quickly and easily. Chipquik.com offers the kit for $16.

Sandpaper A very fine-grit sandpaper is useful for removing oxidation from component and circuit board surfaces.

Desoldering braid Woven metal material used to wick up melted solder.

Small, flat-tip screwdriver Comes in handy for removing some types of components.

Needlenose pliers, wire cutters, and vise These common tools will make your job easier.

The Chip Quik SMD Removal Kit

The Chip Quik SMD Removal Kit allows you to quickly and easily remove surface-mount components such as PLCC, SOIC, TSOP, QFP, and discrete packages. The main component of the kit is a low-melting-temperature solder (requiring less than 300°F) that reduces the overall melting temperature of the solder on the SMD pads. Essentially, this enables you to just lift the part right off the PCB.

CHIPQUIK
CURITY
Alcohol Prep
2 Ply - Medium
CHIPQUIK
No-Clean Paste Flux

INCLUDES

- Alcohol pads for cleaning the circuit board after device removal
- A special low-melting-temperature alloy
- Standard no-clean flux
- Application syringe

Desoldering Tips

FOR STANDARD THROUGH-HOLE COMPONENTS

- First grasp the component with a pair of needlenose pliers. Heat the pad beneath the lead you intend to extract and pull gently. The lead should come out. Repeat for the other lead.
- If solder fills in behind the lead as you extract it, use a spring-loaded solder sucker to remove the excess solder.

FOR THROUGH-HOLE ICs OR MULTI-PIN PARTS

- Use a solder sucker or desoldering braid to remove excess from the hole before attempting to extract the part.
- You can use a small, flat-tip screwdriver or IC extraction tool to help loosen the device from the holes.
- Be careful to not overheat components, since they can become damaged and may fail during operation.

BEFORE YOU START

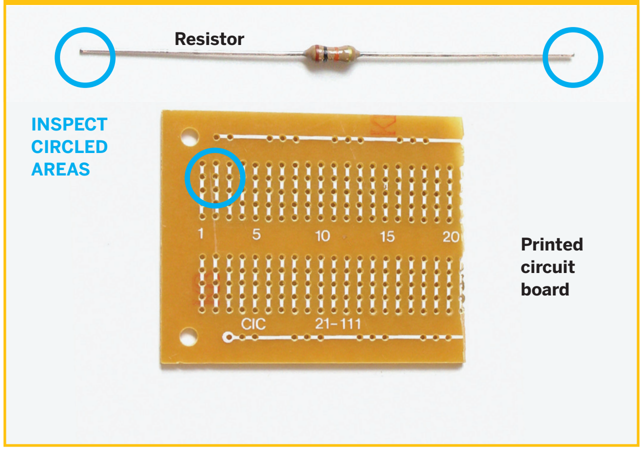

Joe Grand is the president of Grand Idea Studio, Inc. (joe@grandideastudio.com), a product-development and intellectual-property licensing firm. He specializes in embedded system design, computer security research, and inventing new concepts and technologies. He is also a host on The Discovery Channel's *Prototype This*.

SOLDERING A RESISTOR TO A CIRCUIT BOARD

BEFORE YOU START

Inspect the leads or pins for oxidation. If the metal surface is dull, sand with fine sandpaper until shiny. In addition, use the sandpaper to clean the oxidation and excess solder from the soldering iron tip to ensure maximum heat transfer.

This simple example shows the step-by-step process to solder a through-hole component to a printed circuit board (PCB). I used a piece of prototype PCB and a single resistor.

⚠ DANGER: It's important to consider safety precautions. Improper handling of the soldering iron can lead to burns or other physical injuries. Wear safety goggles and other protective clothing when working with solder tools. With temperatures hovering around 700°F, the tip of the soldering iron, molten solder, and flux can quickly sear through clothing and skin. Keep all soldering equipment away from flammable materials and objects. Be sure to turn off the iron when it's not in use and store it properly in its stand.

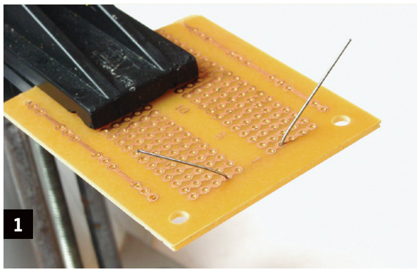
1

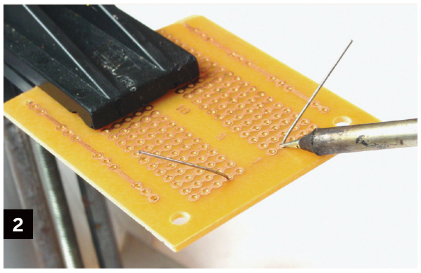
2

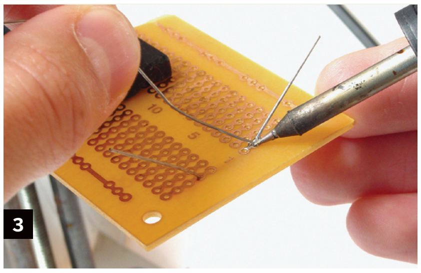
3

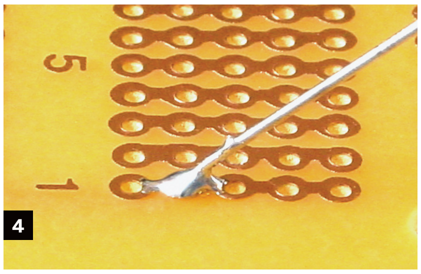
4

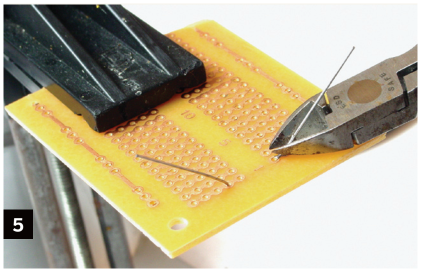
5

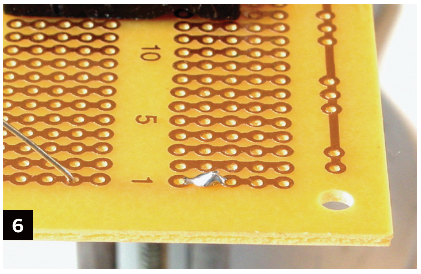
6

1. Bend and insert the component leads into the desired holes on the PCB. Flip the board to the other side. Slightly bend the lead you'll be soldering to prevent the component from falling out when the board is turned upside down.

2. To begin the actual soldering process, allow the tip of your iron to contact both the component lead and the pad on the circuit board for about 1 second before feeding solder to the connection. This will allow the surface to become hot enough for solder to flow smoothly.

3. Next, apply solder sparingly and hold the iron in place until solder has evenly coated the surface. Ensure that the solder flows all around the 2 pieces (component lead and PCB pad) that you're fastening together.

Don't put solder directly onto the hot iron tip before it has made contact with the lead or pad; doing so can cause a cold-solder joint (a common mistake that can prevent your hack from working properly). Soldering is a function of heat, and if the pieces aren't heated uniformly, solder may not spread as desired. A cold-solder joint will loosen over time and can build up corrosion.

4. When it appears that the solder has flowed properly, remove the iron from the area and wait a few seconds for the solder to cool and harden. Do not attempt to move the component during this time.

The solder joint should appear smooth and shiny, resembling the image above. If your solder joint has a dull finish, reheat the connection and add more solder.

5. Once the solder joint is in place, snip the lead to the desired length. Usually, you'll simply cut the remaining portion of the lead that isn't part of the actual solder joint. This prevents any risk of short circuits between leftover component leads on the board.

6. Here's a completed soldering example.

BEFORE YOU START

Use a rubbing alcohol pad to remove any residue from the solder pads. Verify that the solder pads are clean and free of cuts or solder jumps before proceeding. Desoldering, or removing a soldered component from a circuit board, is typically trickier than soldering because you can easily damage the device, the circuit board, or surrounding components. For surface mount devices (SMDs) with more than a few pins, the easiest method to remove the part is the Chip Quik SMD Removal Kit, as shown in the following step-by-step example.

PLEASE READ through this example completely before attempting SMD removal on an actual device. When removing the device, be careful not to scratch or damage any of the surrounding components or pull up any PCB traces.

THREE PRIMARY FUNCTIONS OF FLUX

- **Cleans metal surfaces to assist the flow of *filler metals* (solder) over *base metals* (device pins).**
- **Assists with heat transfer from heat source (soldering iron) to metal surface (device pins).**
- **Helps in the removal of *surface metal oxides* (created by oxygen in the air when the metal reaches high temperatures).**

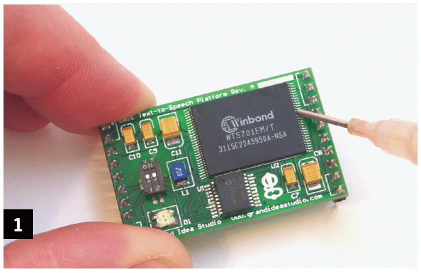
1

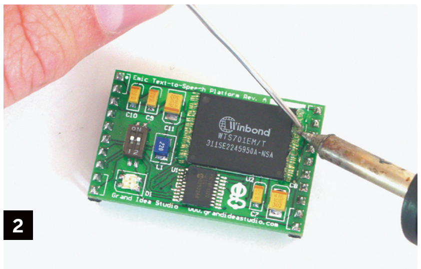
2

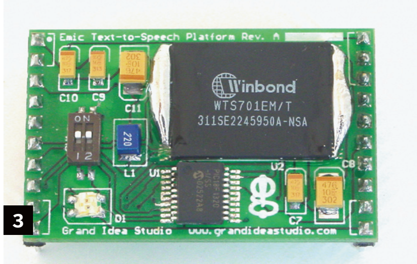
3

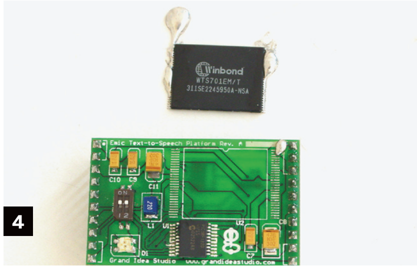
4

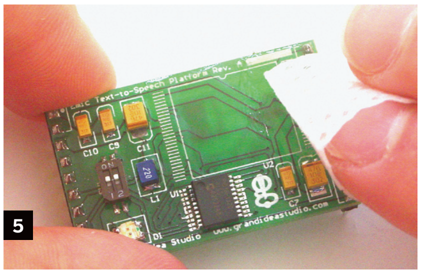
5

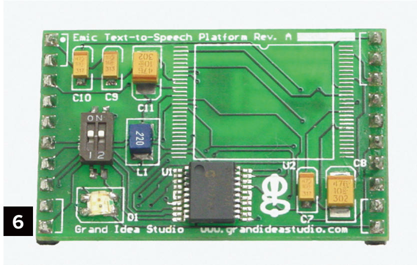
6

1. The first step is to assemble the syringe, which contains the no-clean flux. Simply insert the plunger into the syringe and push down to dispense the compound. The flux should be applied evenly across all the pins on the package you'll be removing. (Flux is a chemical compound used to assist in the soldering or removal of electronic components or other metals.)

2. Once the flux is evenly spread over the pins of the target device, the next step is to apply the special Chip Quik alloy to the device. This step is just like soldering: apply heat to the pins of the device and the alloy at the same time. The alloy has a melting point of approximately 300°F, which is quite low. You shouldn't have to heat the alloy with the soldering iron for very long before it begins to melt. The molten alloy should flow around and under the device pins. Starting at one end of the device, simply heat and apply the alloy. Repeat for the other side(s) of the device.

3. Flux will help ensure a nice flow of the alloy onto the device pins. Make sure the alloy has come in contact with every single pin by gently moving the soldering iron around the edges of the device. Avoid touching nearby components on the PCB with the soldering iron.

4. Now that the alloy has been properly applied to all pins of the device, it's time to remove the device from the board. After making sure that the alloy is still molten by reheating all of it with the soldering iron, gently slide the component off the board. You can use a small, jeweler's flat-tip screwdriver to help with the task. If the device is stuck, reheat the alloy and wiggle the part back and forth to help the alloy flow underneath the pads of the device and loosen the connections.

5. The final step in the desoldering process is to clean the circuit board. This step is important because it will remove any impurities left behind from the Chip Quik kit and get you ready for the next step.

First, use the soldering iron to remove any stray alloy left on the device pads or anywhere else on the circuit board. Next, apply a thin, even layer of flux to all of the pads that the device was just soldered to. Use the included alcohol swab or a flux-remover spray to remove the flux and clean the area.

6. The desoldering process is now complete. The surface-mount device has been removed and the circuit board cleaned.

TECHNOLOGY IN ACTION™

Building Your Own Electronics Lab

A Guide to Setting Up Your Own Gadget Workshop

Dale Wheat
FLUKE

Dale Wheat

For your convenience Apress has placed some of the front matter material after the index. Please use the Bookmarks and Contents at a Glance links to access them.

friendsof™

Apress®

Contents at a Glance

- About the Author xi
- About the Technical Reviewer xii
- Acknowledgments xiii
- Introduction xiv
- Chapter 1: Planning Your Electronics Workshop 1
- Chapter 2: Building Your Tool Chest 19
- Chapter 3: Components 57
- Chapter 4: A Portable Mini-Lab 87
- Chapter 5: The Cozy Corner Lab 105
- Chapter 6: The Small Group Lab and Classroom 135
- Appendix A: Getting Started with Tool Building 145
- Index 161

Introduction

We're surrounded by electronics these days. Did you ever wonder how it all worked? Do you know just enough about electricity to be dangerous? Would you like to learn more? Then this book is for you.

There are many excellent textbooks on how electricity works. This is not one of them. This book helps you get started learning about electricity and electronics by showing you how to set up your own electronics laboratory. You might have guessed that from the title.

Along the way, several simple exercises are presented to show you some basic concepts about how electricity works in the real world—not just in a lecture hall. By the time you get done reading this short book, you'll have a good idea of where to begin and what kinds of tools and components you might need, and you'll have some excellent advice about how to keep it all organized.

Once you've got your own space set up to perform your electronic experiments, the sky is the limit.

Here's what you have to look forward to in this book, one chapter at a time.

Chapter 1: Planning Your Electronics Workshop

Your very own electronics laboratory is a great place to build, repair, invent, and learn more about electronics. We seem to be surrounded by electronics more and more every day. We can easily become dependent upon them, without even understanding how they work. Having your own electronics lab can help you gain some control over your electronic minions. Here you can learn about their inner mysteries, including how to repair them when possible or, better yet, improve them with your own custom modifications. You can also turn your own ideas into reality by building electronic circuits from scratch.

What will you need to plan your lab? What if you already have a basic understanding of electronics and the beginnings of a workshop at your disposal? This chapter helps get you going in the right direction.

Chapter 2: Building Your Tool Chest

You're going to need some tools in your lab, as well as the skills to use them effectively. This chapter will get you started.

If you're starting from scratch, don't worry. There's not a whole lot you absolutely *must* have to get started. Some basic hand tools and a place to keep them organized is all you will need. You don't even have to spend a lot of money at first. As you progress in your hobby, you will most likely want to add to your tool chest and upgrade some of those tools. Electronics can be the perfect hobby because it can take up (1) all your spare time and (2) all your discretionary income. What more could you ask for?

Chapter 3: Components

You're going to need some components to play with, as well as the knowledge to use them effectively. This chapter will introduce you to some of the bits and pieces that make up modern electric and electronic circuits. You'll also learn a little bit about how to identify components from their appearance and markings, when available.

Once you've got an idea about what these parts do in a circuit, you'll learn a little more about how to measure their electrical properties and put them to use. You'll also be shown what *not* to do, in some select examples.

Chapter 4: A Portable Mini-Lab

Having a portable (or at least (*trans*portable) electronics lab comes in handy in several circumstances. Maybe you don't have a place (yet) for a permanent home for all that equipment. Maybe you need to bounce at a moment's notice, heading off on electronic adventures at the drop of a hat. Or maybe you like to keep everything where you can find it in a hurry, without having to rummage through shelves and boxes, looking for just the right tool or component.

Whatever your motivation for wanting a portable lab at your disposal, this chapter should help you get started.

Chapter 5: The Cozy Corner Lab

You don't need an entire garage or extra office for a functional laboratory for your electronic endeavors. It's certainly nice if these areas are available to you, but you might be surprised at how much you can do in just a small space, if you set it up properly and maintain it with determination.

Chapter 6: The Small Group Lab and Classroom

The first part of this book deals with how to get along with electrons and make them do your bidding. In this chapter, you'll explore a completely different topic: other people, and how to get along with them.

You might not ever figure out how to get *them* to do your bidding, but at least you can keep them from stealing your tools. Maybe.

Appendix A: Getting Started with Tool Building

Once you've become comfortable in your lab and had time to play with a few circuits, you might start to notice that some of the "tools" you've been using are just simple electronic circuits themselves. Good examples are power supplies and meters. These are great tool-building projects because you can see useful results early.

This appendix takes a look at building a couple of simple electronic tools that might be of use in your lab. Hopefully you will develop a better understanding of how these tools work. Ideally, you will progress from the basic question of "Does it work?" to the more involved questions of "How well does it work and what can I do to improve it?"

CHAPTER 1

Planning Your Electronics Workshop

Are you interested in electronics? Would you like to set up your own "electro lab" to conduct amazing experiments, build crazy gizmos, and repair or modify your existing electronics? Then this book is for you!

Your very own electronics laboratory is a great place to build, repair, invent, and learn more about electronics. We seem to be surrounded by electronics more and more every day. We can easily become dependent upon them, without even understanding how they work. Having your own electronics lab can help you gain some control over your electronic minions. Here you can learn about their inner mysteries, including how to repair them when possible or, better yet, improve them with your own custom modifications. You can also turn your own ideas into reality by building electronic circuits from scratch.

What will you need to plan your lab? What if you already have a basic understanding of electronics and the beginnings of a workshop at your disposal? This chapter helps get you going in the right direction.

What to Expect

This book can't teach you everything about electronics, and it doesn't try. That would take several lifetimes. What it can do is introduce you to the tools and the skills you will need to set up your own electronics lab. This will include a very basic introduction to electricity, a little bit of theory, some safety tips, and a whole lot of example projects. It's a great starting place. Where it will lead you is mostly determined by you and what you're wanting to do with electronics.

If you're already familiar with electricity and electronic concepts, there is still plenty of fun and interesting stuff waiting for you to explore. The field of electronics itself is constantly expanding. From the time this book was written to the time that you've finished reading this sentence, many advances will have occurred in both our understanding of electronics as well as the development of new applications and electronic devices. This is one of the many things that makes the study of electronics so interesting and exciting.

Every lab is different. Your lab will reflect not only your immediate electronics goals but also your personality. Feel free to pick and choose from all the ideas presented in this book and add in some of your own. The repair shop and the design studio are necessarily going to be arranged differently. There really is no right or wrong way to go.

Even the best laid plans, so the saying goes, end up somewhere unexpected. When planning your electronics lab for the first time, or for the tenth time, keep in mind that your interests, resources, and reasons for wanting to work on electronics are going to change over time. Please feel free to reinvent yourself and your lab as conditions permit. You might also be forced to reconsider your priorities when other factors in your life exert themselves. That being said, stay tuned for "A Cautionary Tale" at the end of this chapter.

Some configurations are more appropriate for certain endeavors, and nothing replaces the right tool for the job. This book should at least give you an idea of what is involved and get you pointed down the right road.

It Starts with a Plan

Most of the decisions you'll make when planning your laboratory will depend on what *you're* interested in doing there. Having a good plan at the beginning is like having a clear map when taking a long journey. It helps you from getting lost and also is handy for figuring out when you've arrived.

■ **Note** If you fail to plan, you plan to fail.

Remember, at this point in the game, you don't have to make any final decisions about what you want to be able to do in your lab. This book is all about illustrating possibilities and giving you the information and advice necessary to branch out into the areas that hold the most interest for *you*.

A Broad Outline

Here's what you need to get started: a little time, a small amount of space, some basic tools, and a few components. Those are the boring parts. The fun stuff includes your ideas, goals, and inspiration. Combine all that with some fundamental information about how electricity and electronics work, and you're in business.

A Little Time

Like any good hobby, setting up your electronics lab and conducting experiments there is going to take up some of your time. For some, this is the *main reason* for getting started in electronics and building a lab: a place to escape, unwind, and tinker with ideas, prototypes, and complex systems. Maybe it's your job. Maybe it's your passion. Perhaps it's a bit of both.

You're going to need time to plan and build your lab. This book will spell out several specific arrangements for various-sized labs, but they are just examples and not hard-and-fast rules. More time will be required for obtaining the proper tools and materials that you will use there. Time will be spent not only *doing* things in your lab, but also in just *thinking* about doing things in your lab. For example, you might build a clever little gizmo that ought to do a simple task, only to find out that it refuses to work. You might need to spend a little bit of time thinking of ways of testing it to find out where the problem lies. This is sometimes called *troubleshooting* or *debugging*.

This is the basis of the scientific method. To help understand how something works, you first form a theory that would explain its behavior. Next you prepare one or more experiments that will demonstrate the accuracy of your theory. The experiments are conducted and their results are analyzed. If the results are in agreement with your theory's predictions, then your theory is more than likely correct. If not, you may need to modify or extend your theory to accommodate the behavior observed. All of these things take time, which helps explain why we don't have jet packs or time machines yet.

A Little Space

You don't have to start out with a cavernous underground bunker for your electronics lab. If you already have one, that's great. You might have an extra classroom available, some unused warehouse space, or an empty bench in a garage.

Then again, you might only have a corner of the dining-room table, and then only between mealtimes. That's plenty of room when you're just getting started! Once you learn more about electronics, and especially when you learn more about what it is that *you* want to do with electronics, then you can start looking for a more permanent home for your toys, tools, and spare parts.

The main points that will be stressed throughout this book when planning and working in your space are safety, good lighting, and organization, in that order.

- Your lab needs to be relatively safe for you, and *very* safe for your visitors and neighbors. The "Go Away!" sign on the door does not guarantee that you will be free from curious visitors, pesky and otherwise. People are naturally attracted to the creative and mysterious activities going on in your lab.
- Be prepared to buy, build, or borrow more lighting, because you're going to need it. Proper lighting improves the quality of your doodlings and tinkering enormously. You might think you already have enough light, and unless you were a photographer in a previous life, you don't. (Photographers understand that you can never have too much light.)
- An organized lab is a productive lab. Don't waste time looking for parts or tools when it would have only taken mere moments to put them in their proper places in the first place. A place for everything, and everything in its place. You get to decide where all these tools and bits get to live, so spend some time and do it wisely.

These three tenets will be repeated throughout the book, so you might as well get used to hearing about them.

Basic Tools

You really don't need a lot of tools to get started. There is a tendency to want to be a bit overprepared before tackling any new project, but it really isn't necessary when it comes to planning and building a new electronics lab. This is also true when sorting out a lab that's gone badly out of control.

You're going to be working with wire quite often. Claude Debussy has been quoted as saying that music is the space between the notes. It's the connections between the various electronic components that make the magic happen in an electronic circuit, in the same way that it's the relationships between people that make their lives so interesting.

The connections between all the various imaginable electronic components almost always start out as bits of wire. These connections can then be optimized into copper traces etched on a printed circuit board or even as metallization links in an integrated circuit. You're going to be working with a lot of wires in this hobby, and that usually boils down to two basic hand tools: *wire cutters* and *wire strippers.*

A wire cutter does what it says and says what it does: it cuts wire. A wire stripper is a more specialized tool that helps remove the outer layer of insulation from a wire. Ever more specialized versions of both tools abound. Some tools aim to perform both tasks at once, while never perfecting the art of either.

Resist the temptation to use your teeth for cutting and stripping wire. If you're caught in a situation without proper tools, use your fingernails. They grow back faster than teeth.

■ **Tip** Use the right tool for the job. Knowing which tool is the right tool is half the battle.

If you're performing detailed work on smaller circuits, you might benefit from having some small needle-nosed pliers, a small vise, clip, or helping-hands jig, and possibly a magnifying glass.

A popular and handy skill in the electronic arena is *soldering*. Soldering is the process of making electrical and mechanical connections. This is done by heating the items to be joined together and adding solder. The solder melts and fills in all the gaps between the items being joined. Once cooled, the solder forms a solid, conductive link.

■ **Note** Soldering is similar to welding, with the exception that welding actually melts the parts being joined.

A suitably sized soldering iron and the right kind of solder are indispensible for many projects, including repair, prototyping, and production. Picking up the skill to make quality connections is obtained in the same way as all worthwhile skills: with lots of practice.

Besides these basic fabrication tools, you will also want some *test and measurement equipment*. These are devices used to measure various aspects of electrical circuit behavior. These range from the very simple, such as voltage meters, current meters, and continuity testers, to the very exotic (and sometimes expensive), which include oscilloscopes and various logic, network, and signal analyzers. See the Appendix for information about building your own simple test equipment for your lab.

This is just a quick summary of the kinds of tools you might want in your electronics lab. Chapter 2 goes into much more detail about the tools you will probably need or want in your lab, as well as the skills needed to operate them safely and effectively.

Fundamental Components

As mentioned in the previous section, you're more than likely going to be wiring up various contraptions and rewiring others. This is probably going to involve a lot of wire.

For small voltages and currents, small-gauge insulated copper wire is often used. It's easy to handle, cut, and splice. This can be done with mechanical connectors such as wire-nuts or screw terminals, as well as with soldering.

Back in the day when computers were all connected using fat cables, lots of tiny wires were bundled up into cables and snaked all over a typical desktop computer installation. With the increasing popularity of wireless devices, those fat, juicy cables are lying around, discarded, unwanted, and just waiting to be harvested for their valuable conductors. You can also buy various kinds of wire in all colors and sizes from most hardware stores and from online suppliers.

Now consider the types of projects you'll be working on:

- If you're most interested in building your own lighting, you're going to want some lightbulbs, sockets, and LEDs, or whatever lighting technology appeals most to you. Don't forget that you're also going to need switches, knobs, dimmers, housings, plugs, and *more wire*.

- If you like to work with things that move, you're going to need motors, relays, solenoids, servos, and fans. You'll probably be needing some gears, pulleys, chains, belts, shafts, and couplings, along with connectors, switches, controllers, position sensors, and *more wire*.
- If you're into making lots and lots of sound, you'll be needing speakers, headphones, microphones, pickups, transducers, transformers, amplifiers, connectors, jacks, plugs, and, you guessed it, *more wire*.
- Are computers and digital devices more to your liking? Be thinking about getting yourself a wide variety of chips, microcontrollers, software, cables, sockets, connectors, and still *more wire*.

Now you've got an idea of some of the different components you might want to have on hand, depending on what kind of inspiration happens to strike. You're going to need a place to keep all these different parts sorted. This will play a big part in the overall planning phase of your lab. Maybe your old fishing tackle box has enough room for your basic hand tools and enough secret compartments to hold an interesting variety of components. On the other end of the spectrum, maybe you need to start thinking about some industrial-grade shelving units or pallet racks. Now is the time to think about all these factors and start making the right kind of plans to help make your lab a successful one.

Again, this is just a quick summary of some of the types of components you might want to work on in your lab. Chapter 3 is going to take you much farther into understanding many of the more common electrical and electronic components that you are likely to encounter in your lab.

Ideas

Admit it: you've already got some ideas for cool projects that you'd like to be working on in your own electronics lab. That's great! You're well on your way to making those dreams a reality.

The best possible thing you could do with your ideas is to write them down somewhere. It can be as simple as a sketch on a piece of scrap paper or as formal as a complete set of project documentation. The important thing is that you actually get the idea across using either words or pictures in a form that you will be able to revisit in the future. This will be of immense benefit to you, as well as anyone else that might be interested in your ideas.

It's also a great idea to keep a log as you work on your projects. That's the perfect place to jot down those ideas as they come popping out of your brain. Be as detailed as you can be. You will thank yourself later.

Even if you don't already have a backlog of ideas jostling for your attention in the lab, it's still a good idea to keep a log and explore the things that interest you. Leonardo da Vinci kept detailed notebooks on almost every possible subject. Anything that interested him got written down, sketched, or diagrammed in his notes. Even today, people are still getting inspiration from his doodles (see Figure 1-1).

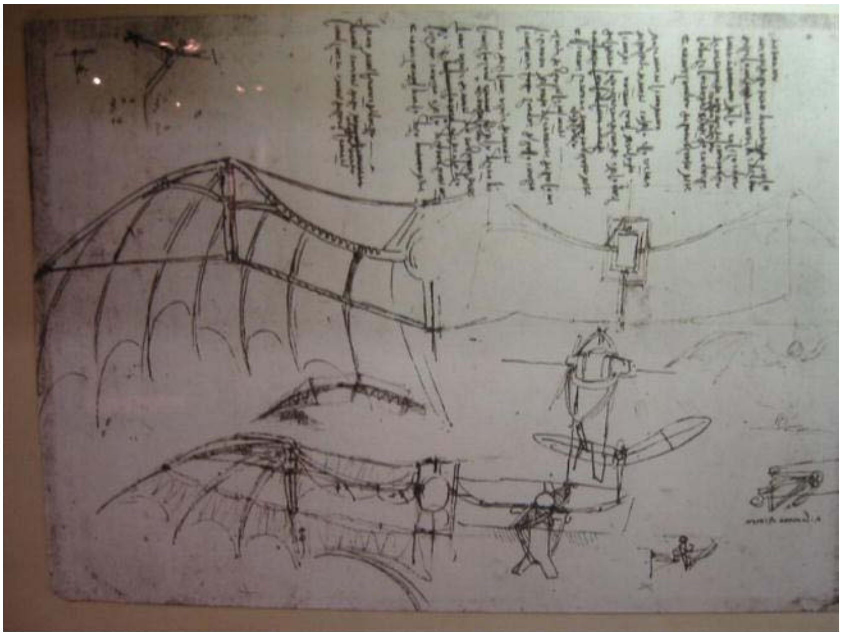

Figure 1-1. *Sketches and notes for a "flying machine" by Leonardo da Vinci. Photograph by Arnaud 25 (public domain), via Wikimedia Commons.*

Writing down your ideas also helps you pick which projects you want to work on first. It just isn't possible to work on all of them at the same time! Narrowing down the list helps get you focused and lets you concentrate all of your talents and energies on the most important projects.

When you're just beginning, the "most important projects" are the ones that provide the most positive reinforcement in the shortest amount of time. Plan a few simple, straightforward projects with a high probability of success at first. You can become more ambitious and take greater risks when you've built up a good reserve of self-confidence and techniques. Remember, the only experiments that fail are the ones from which you don't learn anything. Sometimes learning how *not* to do something is just as important as learning how to do it correctly.

Goals

The best way to turn great ideas into spectacular projects is to have a plan. Set some achievable goals for yourself. When you're first starting, don't be afraid to take those baby steps. You have to crawl before you can teleport.

Project planning is its own art form, and a dark and mysterious art it can be sometimes. It really helps to have clearly defined goals and expectations early in a project. Otherwise, how will you know when you're finished?

Any large, complex problem or project can be broken down into smaller, more manageable subprojects. Unfortunately, this can be carried too far. You don't need to reinvent the wheel every time you start a new project. There's a saying that suggests, "Don't build anything you can buy." The counterpoint to this wisdom is the equally ancient saying, "Don't buy what you can build." You'll have to be the judge as to which one applies best to you and your situation.

Do you have lots of time and very little money? This is good news! This helps simplify the decision-making process enormously. Design and build everything yourself, from scratch. This can include your own tools as well as custom-made components. The trick to this approach is not get lost in all the inevitable details that will arise.

No time to spare, but there's a bit of a budget available? This is likewise good news. With few exceptions, there's someone out there that's willing to part with just what you need for a reasonable price. The trick here is to find them and make that connection.

Inspiration!

Remind yourself what your inspiration is. What initially prompted you to start thinking about building your own electronics lab? Why did you want to get into electronics as a hobby or a career in the first place? Do you even remember?

It really helps to have a firm grasp on what your underlying motives are (or were) when faced with the challenges that regularly present themselves along the way. While it's true that some things just need doing, does it really have to be *you* that does them? Could you farm out some of the work and still feel a sense of satisfaction in the completion of the project? Or do you feel the need to take the do-it-yourself mind frame to its utmost limit?

Whatever your inspiration is, or wherever it came from, do your best to keep it in your sights while pursuing your dreams and goals. It makes it that much more satisfying when you eventually attain them.

How Electricity Works

There are two ways to go about learning how electricity actually works. The first involves just a whole bunch of *fascinating* theoretical information about atoms, electrons, fairies, and dragons. The second way is to just assume that it does indeed work, and spend some time actually putting it to work.

You don't need to know everything about electricity to have fun with it. A passing understanding is sufficient to get you started. As your technical requirements grow more elaborate, you will find yourself going back to revisit some of the basic concepts, until your comprehension is sufficient to carry you forward.

The Theoretical Approach

The prevailing theories about how electricity works center around the orbits of negatively charged subatomic particles called *electrons*. Electrons are quite small, even when compared to individual atoms. Electrons normally whiz around the *nucleus* of the atom, which is usually made up of other subatomic particles called *protons* and *neutrons*. The exception is hydrogen, whose most common isotope has only a single proton and no neutrons. Protons have a positive electrical charge and neutrons have no electrical charge.

When an atom has a balance of protons and electrons, it is said to be electrically neutral. If an atom has more electrons than protons, it has a negative electrical charge. Conversely, if there are more protons than electrons, the atom has a positive charge. An atom that has either a positive or negative electrical charge is called an *ion*.

While flying round and round the nucleus, electrons tend to cluster into well-defined orbits, or *shells*. The innermost shell of an atom can hold as many as 2 electrons. Once this orbit is "full," up to 8 more electrons can be accommodated in the next larger shell. The next shell holds up to 18 electrons, and the one after that can contain up to 32 electrons. After that, the shells get smaller again, stepping back down symmetrically from 18 to 8 to 2.

Each unique combination of electrons and protons represents one of the chemically distinct *elements*. Each element has its own particular chemical characteristics, and cannot be further subdivided by conventional chemical or mechanical means. Toward the end of the *periodic table*, which arranges all the known elements into meaningful rows and columns, you'll start to find some exceptions to the shell symmetry. These heavier elements, however, don't usually turn up in most electronics labs, so we'll not be spending a lot of time there.

Individual atoms can combine into more complex structures called *molecules*. One popular molecule is water, which is made up of two hydrogen atoms and one oxygen atom. Each of the hydrogen atoms has a single electron. The oxygen atom has eight electrons. The first two electrons in the oxygen atom occupy the first shell, leaving six electrons in the next outer shell. That leaves two "open" places in the second shell, which is where the two electrons from the two hydrogen atoms come into play. The electrons are shared by the three atoms, linking them all together and forming a *covalent bond*. Water has vastly different physical properties than either hydrogen or oxygen.

It's also possible to bump an electron out of orbit around a nucleus. When this happens, the atom becomes an ion and is electrically positive. This makes it especially attractive to any loose electrons that happen to be in the vicinity. This is because electrons are negatively charged, and in the case of electrical charges, opposites attract. This movement of electrons is how an electric charge flows through a substance.

Some materials are very good at slinging those electrons around, and as such are called good *conductors* of electrical energy. Examples of good conductors are copper and silver. Some materials resist the flow of electrical charge, and can be used as *insulators*.

An important concept to understand about the flow of electricity is that it is based on the movements of tiny, *negatively charged particles*. This is referred to as the flow of *electron current*, but this term is rarely used. For some reason, the more popular, or conventional, idea of electric current flow is from the more positive point to the more negative point in a circuit. This is referred to as *conventional current* and is the standard used throughout this book.

The fairies and the dragons come into the picture when you start to look any deeper into how the building blocks of matter are actually (or allegedly) constituted. The previous quick summary of how electricity works covers less than 1 percent of the physics needed to understand even how basic atomic principles operate. Even then there are various schools of thought as to how, *exactly*, all that tiny stuff works together to produce rainbows and cousins and moon rocks. A proton, once thought to be an indivisible *elementary particle*, is now thought to be built out of a "down" quark and two "up" quarks. There are also "strange" and "charmed" quarks flitting about, along with an entire pantheon of mythological particles. Go figure.

Now all of that *fascinating* information only scratched the surface of how *electricity* works. Electricity does a lot of work for us. We use it for lighting, heating, motion; just all kinds of things. *Electronics* is where we start to get into some really bizarre physics involving semiconductors, transistors, computer chips, and other miracles of the modern age. Explaining even the basic ideas about how *electronics* works is way beyond the scope of this little book. It's certainly possible to learn and understand it, but for the purposes of setting up a lab and building some example projects, we can just assume that it does indeed work. We can start from there and have some fun with it.

If your curious mind just can't stand *not knowing* how electronics works, the interwebs are your friends, and so are the helpful folks at your local library.

A Practical Approach

An easier way to think about electricity is to skip over *why* it flows and concentrate on *how* it flows. That's what we'll be doing in the remainder of this book, for the most part.

■ **Note** Electricity flows in a circle.

Any continuous flow of electrical current travels in a circle. There are exceptions to this broad generalization, but most of the example circuits encountered in this book will have a definite circular arrangement, if you know how to look for it.

Here are some technical terms that will be used throughout this book. You don't have to memorize any of them right now (there will *not* be a test), but you should at least take a glance at them so you know where to come back when you want to know more.

> *Circuit*: A circuit is anything that forms a continuous loop or circle. The term *circuit* is formed from the same Latin word that means *circle* or *wheel*.
>
> *Amps*: The *intensity* of an electrical current is measured in units called *amperes*, named after André-Marie Ampère. This is usually shortened to *amp* in popular usage. The symbol for amperage in a circuit is *I* (for intensity), while the unit is abbreviated *A*. For example, you could say that the intensity of the electrical current (I) in a circuit was measured to be 3.75A (amperes). Electrical current is measured with an *ammeter*.
>
> *Volts*: The voltage in a circuit is the measure of the difference in electrical *potential*, or how much work it *could* do. The unit of voltage is the volt (V), named after Alessandro Volta. This is sometimes referred to as pressure or tension. For example, a "high-tension wire" means a high-voltage line, not necessarily one that has been pulled really, really tight. Voltage is, not surprisingly, measured with a *voltmeter*.
>
> *Watts*: Electrical power is measured in watts (W), named after James Watt. The easy way to calculate watts, or the amount of energy that is converted into work (either useful or not) is to multiply the voltage across the circuit by the current flowing through the circuit. For example, if a 12V battery is connected to a lightbulb that draws 2A, the bulb converts the electrical energy into both light and heat (mostly heat) to the tune of 24W (12V × 2A = 24W).
>
> *Ohms*: The resistance to the flow of electrical current is measured in *ohms*, named after Georg Ohm. A *resistor* is a simple electronic component that exhibits a fixed resistance to the flow of electrical current. The symbol for resistance is *R* and the unit of measure is the capital Greek letter omega, Ω.

Safety

Make no mistake: electricity can kill you. It bears repeating, as well as emphasis.

■ **Warning** Electricity can kill you.

Working safely with electricity is like working safely with a sharp knife, for example. A good knife is a useful tool. Life would be cumbersome and crude (not to mention short) without good tools. However, when using *any* tools it's important to understand their proper use and operation. The same goes for electricity.

The point here is not to be overdramatic about the dangers of electricity. Billions of people are safely and peacefully coexisting with electricity, largely due to wise safety rules and well-designed electrical appliances. However, there are three major ways that you can be harmed when working around electricity, and it is important that you be aware of them:

- Your body uses tiny amounts of electricity to send nerve impulses to and from your brain. Sending relatively giant shocks of electricity through your body is a good way to stop your heart from beating or your lungs from breathing. You need to keep doing both of those things, so be careful.
- Since your body is not a very good conductor of electricity, it converts the electrical current into heat, which can produce very damaging burns, both on the inside and the outside of your body.
- Additionally, many electrically related injuries are sustained while trying to *get away* (i.e., your body's involuntary reflex when exposed to dangerous voltages).

Yet working safely with electricity is easy. Observe the following rules and you'll be relatively safe, both in your lab and outside of it:

- Always assume a circuit is "live." Electrical devices can retain lethal voltages even when completely disconnected from power.
- Don't work with high voltage (more than 36 volts) unless you're specifically trained. Reading this book does *not* mean you're "specifically trained," even if you read it *twice*.
- Don't short-circuit batteries (i.e., deliberately connect the positive and negative terminals together). Even very small batteries can hold dangerous amounts of power and can produce enough heat to burn you. Even worse, it can start a fire, which is yet another unpleasant way to die.
- Be careful when using metal (conductive) tools around battery terminals so that you don't accidentally short the terminals together or to other circuitry.
- Don't assume other people are aware of these dangers.

Lighten Up!

Now it's time to have a little fun with electricity, don't you think? While it's important to keep safety in mind, it's just as important to be curious and willing to try new things. Don't let the (important) safety rules immobilize your creativity. Use these rules, like any tool, to leverage your enjoyment of this fun and exciting hobby.

Let's start with a simple circuit that is almost guaranteed to lighten things up. Let's take apart a flashlight (or *torch* for you metric types) and see how it works.

First, find a typical household flashlight. The flashlight in Figure 1-2 was purchased for $3, and it included two new D-cell batteries. Put the batteries in the flashlight and make sure it lights up when you turn on the switch. If the bulb is burned out or the batteries are dead, then this experiment is not going to be much fun at all.

DORCY
Frost Brite
ClearView Design
TrueSpot Reflector
Modèle transparent
Réflecteur TrueSpot
DORCY
HEAVY DUTY

Figure 1-2. *A typical household flashlight will be the ~~victim~~ subject of our first electrical experiment.*

Once we've established that the flashlight works as intended, we can begin the experiment.

DISSECTING A FLASHLIGHT

We don't need most of the pieces of the flashlight body. Let's take it apart and have a look inside.

1. Open the flashlight be unscrewing the end with the lightbulb in it.
2. Remove the batteries. Set them aside for the moment. We'll use them in a bit.
3. Remove the plastic retainer that holds the bulb in the reflector.
4. Take out the bulb.
5. We'll need to replace the wiring that was contained within the flashlight's body. Find a short piece of wire 8" to 12" long. A specialized jumper wire with alligator clips on the end was used in the photos for this exercise, but almost any kind of wire will do. See Figure 1-3.

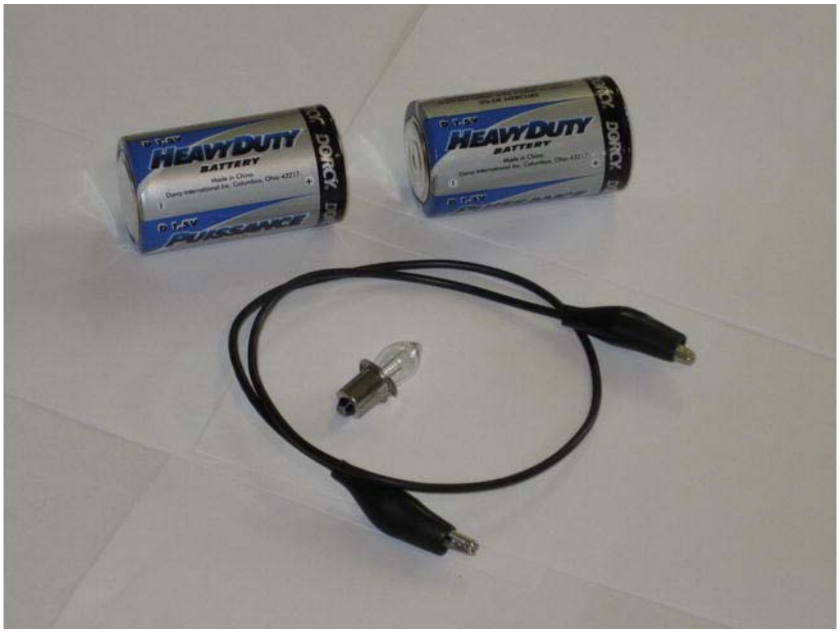

Figure 1-3. *All the pieces you'll need to light up the bulb.*

6. If you're using plain, insulated wire, you'll need to remove a small bit of it from both ends of the wire to expose the *conductor*, or metal wire part inside.
7. This step can be tricky. Try to hold both batteries in one hand, along with the bulb. The batteries need to be pointed in the same direction. If this is too hard to do, try it with just one battery.

8. The positive end of the battery has a small cap on it. Touch the bottom point of the bulb to the top of the battery.
9. Now, using your other hand, connect the wire from the bottom of the battery to the metal base of the bulb.
10. The lightbulb should start to glow. See Figure 1-4.

■ **Caution** If you have a super-bright bulb, it can get hot. Don't burn yourself!

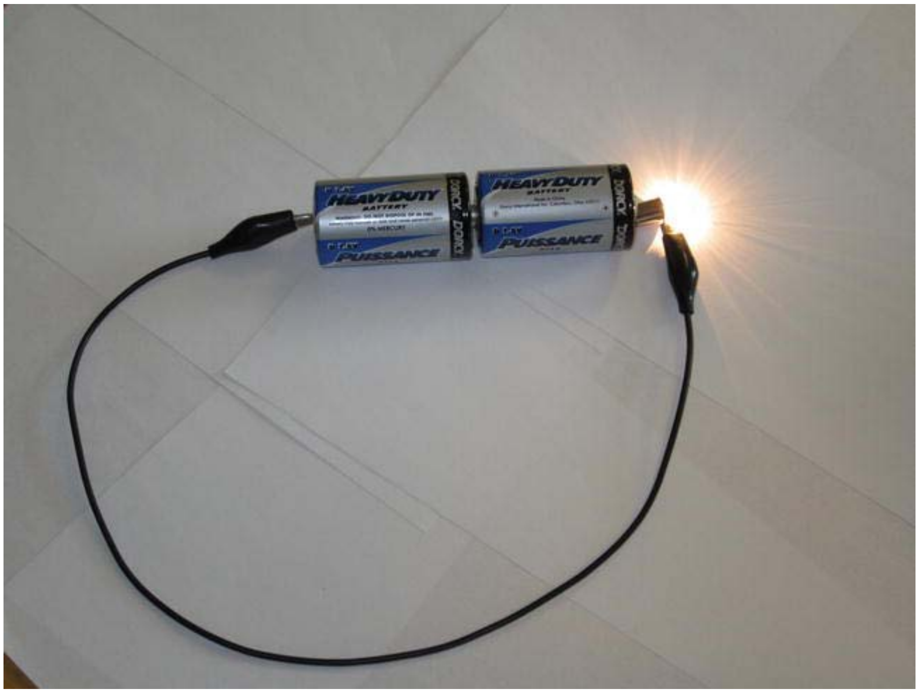

Figure 1-4. *The completed circuit, showing the glowing bulb. Electricity is flowing!*

Several important things are going on in this very simple circuit. Electrical current is flowing, producing useful work (in this case, light). Can you see the "circle"? When you *break* the circuit, the bulb stops glowing. This is exactly how the switch in the flashlight body works. Peek inside the flashlight and see if you can tell how it works. By connecting and disconnecting the wire, you can make the bulb flash. If you only use one battery, the bulb turns on, but is not as bright. Why do you think that happens?

Environmental Impact

You affect your environment every day, and your environment affects you, as well. As you learn more about electronics and all the interesting and wonderful things it can teach you, be mindful of what kind of trail you're leaving behind for others to clean up.

A good way to get a feel for your environmental impact is to take a look at your garbage. It doesn't sound that interesting or fun, but it *will* teach you a lot about yourself and your habits. What are you throwing away? Could you have found either another use for it, or perhaps another home? Is it going to be a nuisance to someone else, or even worse, a hazard?

A good example is the use of "disposable" batteries. We used batteries (technically *cells*, but that distinction will be spelled out in more detail in Chapter 3) in our very first experiment. We buy new batteries and plug them into our devices, and when they run out of power we take them out and throw them away. Where do they go? How do they then interact with the environment? They don't just "go away."

Disposable batteries are very convenient. They are generally inexpensive, compact, self-contained, and reliable. This *convenience*, however, comes with a price all its own. Consider using *rechargeable* batteries whenever possible in your lab. They cost a little more than their disposable counterparts, but can easily and cheaply be refilled with electrical goodness over and over again.

You should also be aware of any hazardous materials that you use or store in your lab. For example, lead is a very common material used in many electrical devices, even though its use is being reduced worldwide. It is found in older-technology rechargeable batteries, solder, and printed circuit boards.

Some of the chemicals used to etch printed circuit boards can be dangerous if not used, stored, and disposed of properly. Take the time to understand what, exactly, you're working with when you're conducting your experiments in your laboratory.

Just like the captain of a seagoing vessel, you're in charge of what happens in *your* lab. With this authority come important responsibilities that rest squarely upon your shoulders. Please take these responsibilities seriously.

Budgeting

How much is all this stuff going to cost? A lot of it depends on how creative you want to be in stocking your lab with tools and parts.

If your budget is small, then start small. If your budget is nonexistent, then you'll just have to start even smaller. You should generally approach acquisitions for your lab under one of these three headings:

- Buying new
- Finding used
- Harvesting or recycling

Buying New

Buying everything new is obviously the most expensive approach, at least from the *money* side of it. What a lot of people fail to take into account is that their *time* is also a very valuable commodity. Try to balance the one with the other, as befits your particular situation.

Except for the folks in the most *dreadful* of hurries, it always makes sense to shop around for the tools, equipment, and components that you are going to need in your lab. You'll find that a lot of "professional-grade" equipment carries an enormous price premium compared with "consumer-" or "hobby-grade" alternatives. This is generally to be blamed on the fact that there is a correspondingly enormous amount of money to be made in the field of electronics.

New components are almost always more desirable to use, especially when you're looking at a production environment and you need to make sure everything produced is of the highest possible quality and reliability. New, unused components are also more appropriate when conducting complex experiments, as this helps to reduce, if not eliminate, many variables in the testing and analysis phases. You don't have to worry that someone (probably you) might have "borrowed" some parts for a quick fix or some other temporary usage and then surreptitiously replaced them back in inventory, having possibly compromised them in some unseen way.

Finding Used

A prudent way to fill up a new lab, especially when first starting out, is to look for bargains on used tools and components. This is especially true of the fixtures, shelves, cabinets, benches, and other furniture that belongs in a lab.

Tools wear out. Tools are replaced. This is a fact of life when you're in the business of working with tools. The good news is that "one man's trash is another man's treasure." If we consider the possibility of women participating, we get three more potential combinations, all of which can end up in a win-win situation.

Many used tools can be retrofitted and placed back into service in your lab. On the other hand, there are a lot of tools that are simply used up and can not feasibly be brought back into useful service. Either their repair is cost prohibitive or they are no longer supported by their original manufacturers.

Another factor working in favor of the patient and observant collector of discarded tools is the inevitable march of progress. What was shiny and top-of-the-line a year ago might be intolerable or insufficient next year, from the standpoint of the original purchaser. Many bargain opportunities arise from this simple understanding.

This can sometimes be applied to the purchase of components that have never been used, but have been sitting somewhere waiting for their chance. A commonly used phrase to describe these parts, which are often sold at a discount, is *new old stock (NOS)*. Sometimes these parts are sold this way because of overstock or because a manufacturing run was cancelled or cut short.

Harvesting or Recycling

There's an absurdly large amount of perfectly good equipment, tooling, fixtures, and components sitting out there, looking for a new home. A lot of companies, as well as individuals, either outgrow their present fittings or turn toward other interests and endeavors. This can be an excellent opportunity to acquire the furnishings you need to populate your lab.

On the other hand, please don't go crazy. You don't have to pick up every discarded appliance or chair you see sitting by the roadside. Why pay rent for something you can easily and cheaply obtain when you actually need it? Are you really wanting to pay the heating and air-conditioning bill for that box full of odd-sized capacitors that are probably past their prime? Don't fall into the trap of the *false economy* when it comes to bargain-hunting for your lab. The most important ingredient in your lab is *you*. Leave a little room to move around a bit.

Another factor to keep in mind is *shelf life*. Not all parts (or tools) last forever, especially when not being used. Batteries are the worst. You can prolong the shelf life of nonrechargeable batteries by storing them in your refrigerator, as this slows down the electrochemical processes within them that produce the electricity. Even solar panels will wear out with either constant use or constant disuse, although this can take years to happen.

Some Example Workshops

Here's a little peek into some actual electronic laboratories, each with its own personality and history. Each one grew from something small into something, well, not so small.

A Fresh Start

Harold Timmis is an electronic engineer, software developer, and author. His curiosity about how things worked started as a small child, which led him from taking apart everything in his parents' house to studying engineering at the Florida Institute of Technology.

Harold keeps a lab in his home where he can work on gadgets in his spare time. Harold recently moved, and used the opportunity to "clean house" and set up everything in his lab exactly the way he wanted things.

He took this photo of his new lab after working on a couple of projects (see Figure 1-5). Note that everything is within easy reach when sitting at the desk. It's still quite tidy, as you can see.

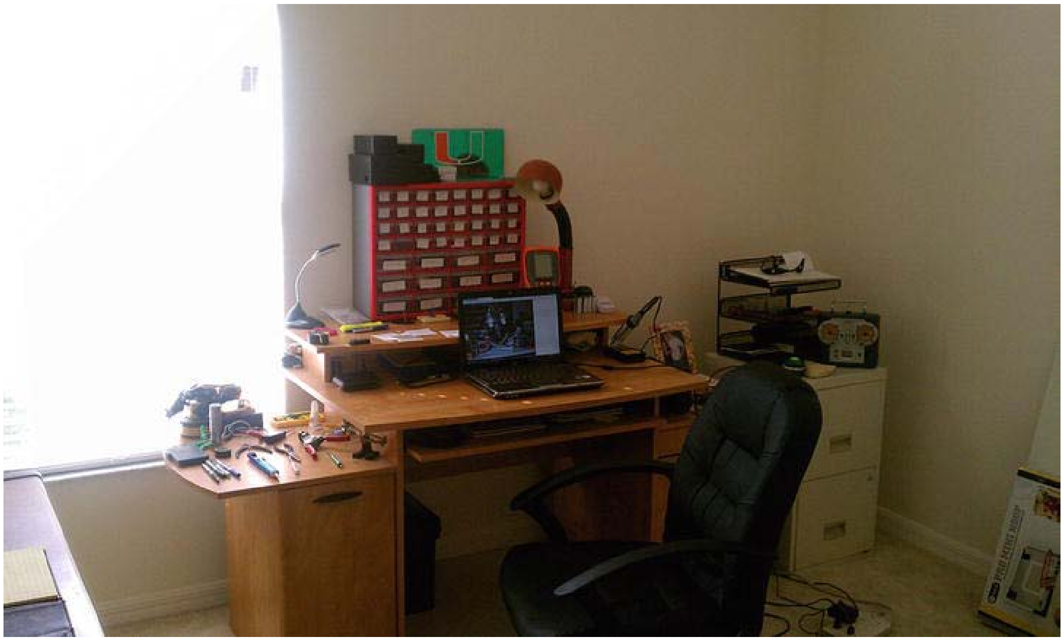

Figure 1-5. *The reorganized home lab of Harold Timmis (photo by Harold Timmis)*

The Robot Clubhouse

The Dallas Personal Robotics Group (`http://dprg.org`) is the world's oldest personal robotics group known to man. From 2002 to 2009, Mike Dodson of Modern Assemblies donated the use of a portion of one of his warehouses to DPRG. This gave DPRG a permanent home for meetings, contests, and social gatherings.

Robot builders enjoyed a spacious and comfortable area for working on robot projects, among other pursuits. Over the years, members contributed time, equipment, and lots of hard work to improving the space. The photo in Figure 1-6 shows one end of the available space, after an extensive remodeling. Several workstations along the wall are in the process of being set up. That good-looking fellow seated at the table is the author.

When Mike retired in 2009, DPRG began a search for another home.

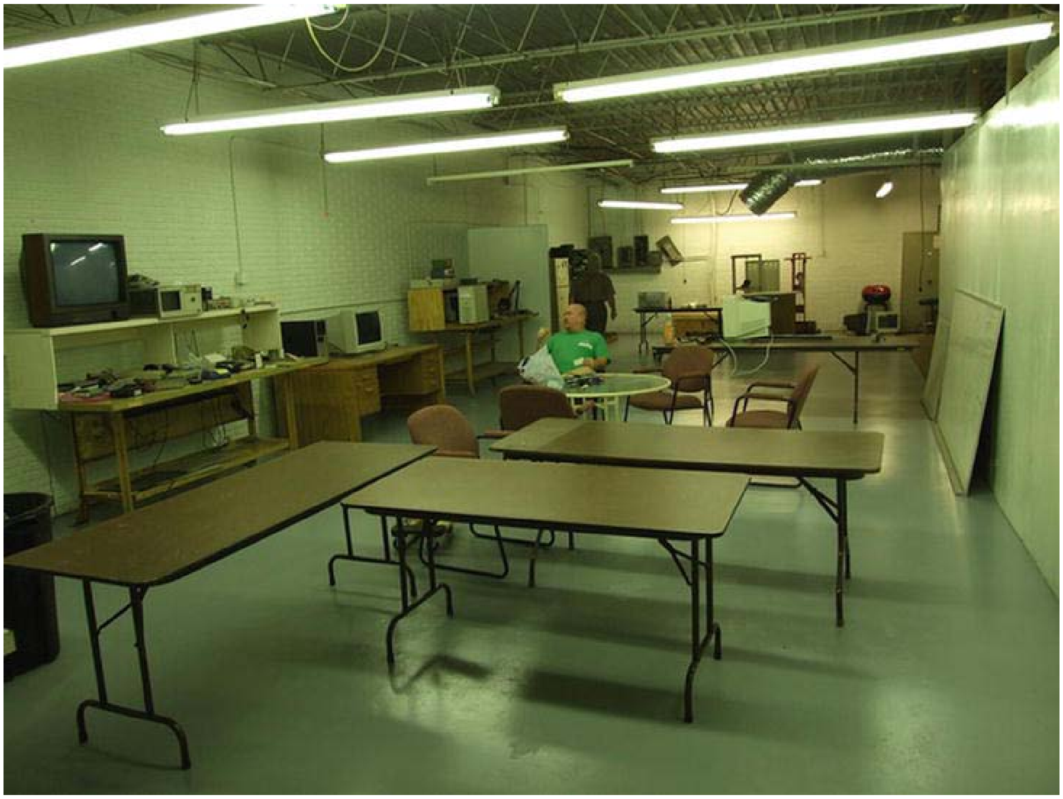

Figure 1-6. *DPRG's world headquarters from 2002 to 2009 (photo by R. Steve Rainwater)*

A Cautionary Tale

It's relatively easy to go from a well-designed and effective lab to something, let's say, less desirable. This doesn't happen overnight, but it can happen. Let this be a warning to you about how even the best of intentions can still produce embarrassing and unproductive confusion (see Figure 1-7).

Unfinished projects and leftover parts and tools compete for the limited space available in this photo. Not only is a lot of time wasted in searching for both components and tools, but the limited space restricts the scale of projects that can be entertained.

Don't let this happen to you.

Figure 1-7. *It's not safe. It's not pretty. It's not even properly "eccentric." You have been warned.*

Summary

So now you know a little something about electricity and electronics. Hopefully this chapter has covered your very basic questions about how all this magic stuff actually works.

Now you're ready to start poking around some circuits. You're going to need the right tools for the job. Chapter 2 will get you started. You'll learn about some basic hand tools that are very useful in the lab, as well as some fairly high-tech machines that you might require in the future.

CHAPTER 2

Building Your Tool Chest

You're going to need some tools in your lab, as well as the skills to use them effectively. This chapter will get you started. If you're starting from scratch, don't worry. There's not a whole lot you absolutely *must* have to get started. Some basic hand tools and a place to keep them organized is all you will need. You don't even have to spend a lot of money at first. As you progress in your hobby, you will most likely want to add to your tool chest and upgrade some of those tools. Electronics can be the perfect hobby because it can take up (1) all your spare time and (2) all of your discretionary income. What more could you ask?

The very basic hand tools, such as screwdrivers and tweezers, are mostly self-evident in their usage. Try not to poke yourself in the eye. Where some helpful advice is appropriate, it will be given.

Here's an excellent example of a simple safety rule that will absolutely improve and extend your life: when using any kind of cutting tool, always direct the cutting motion away from yourself. For example, you might receive a package in the mail, and get all excited about opening it. You then discover it's sealed up with indestructible packing tape and can't simply be ripped open with your bare hands. Use a small knife or box cutter and make a single cut at a time, starting from the point closest to you on the package and cutting outward. Using this method, should the knife slip for any reason, the blade will travel away from you. This allows your blood and other bits to stay inside you, where they can do the most good. Never cut toward yourself.

If you've already got some tools (or even a *lot* of tools), it never hurts to review their proper usage. Also, if you're suffering from an *overabundance* of tools, to the point where you can no longer keep them organized, this chapter might help you to prioritize them, at least as far as permanent residency in your lab is concerned.

■ **Tip** Every tool in your shop must *earn* its place there. It must *deserve* to be there. *No exceptions*.

The Most Important Tool Ever

The most important tool you'll ever use is your mind. Every other tool you will ever work with depends on the correct and proper functioning of your mind. If you're careful with it, you won't misplace it, neglect it, damage it, or use it for the wrong job. Your mind is a wonderful asset when cared for properly, and a terrible liability when ill-treated.

Your mind is in many ways like a muscle. The more you use it, the stronger it gets. Learning new things does not "push out" the old things you've already learned. Your brain is capable of forging new neural pathways when challenged, and there's plenty of extra capacity in that old noggin of yours to handle whatever may arise.

■ **Tip** Try to learn something new every day.

Take care of that brain of yours. Just like you wouldn't use a dull drill bit or a rusty saw, don't try to work on electrical or electronic projects when your mind is not in the right place. Ideally, you should never try to get any work done in your lab when you are sleepy, distracted, angry, upset, medicated, or overly tired. Deal with whatever problems you're facing in a responsible manner and then get back to work in the lab.

- Are you sleepy? Take a nap or head off to bed for the night. Really bad mistakes are made when the brain does not get enough sleep. Sleep deprivation is notorious for being obvious (and obnoxious!) to everyone else except you. "Really bad mistakes" made while working with electrical or electronic projects can produce very hazardous conditions for you and everyone around you, such as fire, damaged equipment, injuries, and even death.
- Are you angry or upset? Then your mind is not going to be adequately focused on what you need to do in the lab. "Fools! I'll destroy them all!" are famous last words *for a reason*. Give yourself some time to calm down from whatever has angered or upset you. Also, it's OK for you to give yourself permission to admit that you are angry or upset about something in the first place. Ignoring it probably won't make it go away, and probably won't make your feelings about it go away, either. Walk away, count to ten, or whatever works for you.
- Feeling distracted? Then now is *not* the time to be conducting electronic experiments in the lab. You need to be *present* in the lab, both bodily and mentally. There are just too many ways to make simple blunders that have far-reaching and sometimes overly expensive consequences.
- Have your cognitive functions been compromised? This can happen from taking certain medicines or even recreational intoxicants. If the little bottle says, "Don't operate machinery or drive an automobile," then you probably shouldn't be making important decisions or working with dangerous circuits in your lab, either. Wait until you're clear-headed enough to proceed with your plan for world domination, adding blinking eyes to a plush doll, or what have you.

The worst thing about cognitive impairment is that the first thing that gets impaired is the ability to judge the level of your cognitive impairment!

Care and Feeding of Your Brain

Here are a few tips to keep your brain (and the rest of you) in good working order:

- Stay hydrated. Drink more water than you think you need. This might result in more trips to the bathroom, but this is much more desirable than the alternative. A good flow of water through your systems helps your body feed and repair itself, as well as aiding in the elimination of toxins. Drink more water.
- Avoid excess consumption of coffee, tea, sodas, and especially "energy" drinks. These contain ingredients that stress both your body and your mental capacity. They do not contain energy, but trick your body into going into overdrive by triggering a stress reaction.
- Take frequent breaks. Get up and walk around, if possible. Raise your arms over your head and hold them there as you count to ten. This redirects blood flow and helps restore a balance within your peripherals.

All Those Other Bits, Too

Just as your mind is the most important tool you'll ever use, the rest of your bodily bits and personal pieces are vital and important in your work in the lab. For the most part, there are no spare parts laying around for things like your eyes, your ears, your hands, or your skin.

Should you wear eye protection in the lab? Yes, you should. Are you always going to do so? Probably not. It's true, it can be inconvenient or perhaps a tad uncomfortable. Consider the cost, however. Even the tiniest scratch on your eyeball can take weeks to heal and leave you susceptible to infection or vision loss. It's just not worth it to be lazy. Ask the next one-eyed person you see. If you're a bit uncomfortable asking one-eyed people, "Hey, what's with the one eye?" then perhaps you can just try to imagine them advocating or encouraging you along the same path they took, whether intentionally or accidentally. They won't do it.

Invest in quality eye protection that is comfortable and properly sized for you. Provide eye protection for visitors to your lab as well, if they will be present when soldering, welding, machine tooling, or wire clipping is being performed. Then invest in the habit of wearing your eye protection whenever it is appropriate.

■ **Note** Don't pretend that your prescription glasses with polycarbonate lenses are "safety glasses." They aren't. They do not protect either your eyes or your face from impact or objects entering from the side, as proper safety glasses should. If you usually wear glasses, use safety glasses that comfortably accommodate your glasses when worn.

Will you be working with really loud noises? These could be from powerful audio amplifiers ("it goes to 11") or power tools. Hearing loss from exposure to excessively loud noises is cumulative. This means it slowly (or suddenly) chips away at your hearing, bit by bit. Why do you think old people turn up the TV so loud?

Hearing protection is both affordable and comfortable. For temporary protection, you can use disposable earplugs. If you need protection all day, consider earmuffs designed for hearing protection. Don't let "Hey, y'all, check this out" be the last thing you ever hear.

It's hard to work with hand tools with no hands. It's possible; just really hard. Take care of your hands. There are a hundred ways to burn, tear, cut, or otherwise mangle those mitten-fittings of yours. Have a variety of work gloves on hand to handle the different needs of your lab. The choice will be ultimately determined by the particular hazards present in your lab. Are you working with chemicals as well as electronics? How about sharp edges or glassware? Normally, you won't need to wear gloves for soldering small electronics, but if you start working with really large-scale projects, such as sculptures or plumbing, you might.

Make sure you are still "handy" when wearing the proper gloves for the job. Make sure the gloves fit you properly and allow you the adequate dexterity you need.

■ **Tip** Be safe. Don't be lazy.

Basic Hand Tools

Assuming that you're going to start small and build toward bigger things, you should also start small with some small hand tools. These tools will make it much easier to work on small parts and wires.

Good hand tools increase your effective strength and improve the accuracy and precision of your motions. They allow you to focus a great deal of your strength on exactly the part that needs it, without wasting a lot of effort. Good hand tools also insulate you from sharp, hot, or otherwise unpleasant objects when you work on them.

Take care of your hand tools. Put them up when you're done with them. Make sure they are clean and ready to be used again the next time you need them. Few things are more frustrating than attempting to start a quick repair job and finding out that you need to spend twice as much time preparing your tools for use.

Have some sort of methodical plan or reasoning involved with the storage of your tools. They need a safe and comfortable home, just like you do. You need to be able to find them quickly when you want them. Try out different organizational strategies until you find one that works for you and your lab. This is an easy task when you have just a few tools to organize. Once you've accumulated a large collection of hand tools, this is no longer optional. You *must* know what tools are available and where they are located.

Frequently used tools should be within easy reach of your primary workstation. If your tools need any sort of adjustment or setup before use, make sure you've got all the proper bits handy. For example, a drill press will often have a chuck key used to tighten and loosen the chuck that holds a drill bit in place. That key needs to be in the immediate vicinity of the drill press, as well as easy to locate. Don't spend a lot of time needlessly hunting for tools. Decide for yourself on some safe and practical policies for where you keep your tools.

■ **Tip** Don't waste time guessing where your tools are located. *Decide* where they are located.

Just like everything else, hand tools have a limited useful lifetime. While some tools can easily outlast you, it's up to you to keep track of exactly how much utility you can continue to expect from your tools. This doesn't mean that you have to trash your hand tools once they develop some tiny flaw. If it's beyond your capabilities to repair the tool, then consider replacing it and reassigning it to the second string or backup tool chest. Having a spare tool, even a slightly imperfect one, on hand when you need it can make the difference between success and failure in a project.

Bear in mind, however, that some tools are simply too dangerous to keep around once they've outlived their useful life span. A dull cutting tool that cannot be reasonably resharpened is a good example of a tool that has lost its honored place among your other tools. Thank it for its service, wipe away that little tear, and then figure out a good way to recycle any of its bits that are still functional.

Wire Cutters and Wire Strippers

As you might have picked up from Chapter 1, you can spend a lot of time working with wire in this hobby. Understanding a little bit about how wire is made and how it functions in a circuit will aid you in your experiments.

How Wire is Made—The Very Short Version

Most wire is made of copper or a copper alloy. Some wire used for electrical wiring in houses is made of aluminum. The wire itself starts out as a fairly thick strand and is stretched and pulled—a process called *drawing*—until the desired diameter is obtained. This strand is then wrapped up on spools or bobbins.

- Some wire is made of a single strand of copper (or aluminum). This is called *solid-core* or *single-strand* wire. It is a single strand of wire, usually circular in cross-section.
- Multiple, thinner strands of wire are often bundled or twisted together to make larger-diameter wire. Stranded wire is much more flexible than solid-core wire, but is more expensive to produce because of the extra manufacturing steps required. Because stranded wire has a larger surface area per unit of length, it is more susceptible to environmental hazards such as corrosion.

Most electrical wire is covered with an insulating jacket, usually made of plastic or rubber. This sheathing must be removed to make electrical connections. There is sometimes more than one layer of insulation on a wire, depending on what its final application is destined to be.

More information about wire as a *component* will be presented in Chapter 3. For now, the emphasis is on the characteristics of wire that help determine the best tools to use when working with wire.

Selecting the Correct Tool

Both copper and the aluminum alloys used for electrical wiring are fairly soft metals and can be easily cut or formed with simple hand tools. Wire *cutters* are specialized tools made to make a clean cut through the wire's strand or stands. Wire *strippers* are another kind of specialized tool used to remove the insulation from a wire.

Figure 2-1 shows two rows of commonly available wire cutters.

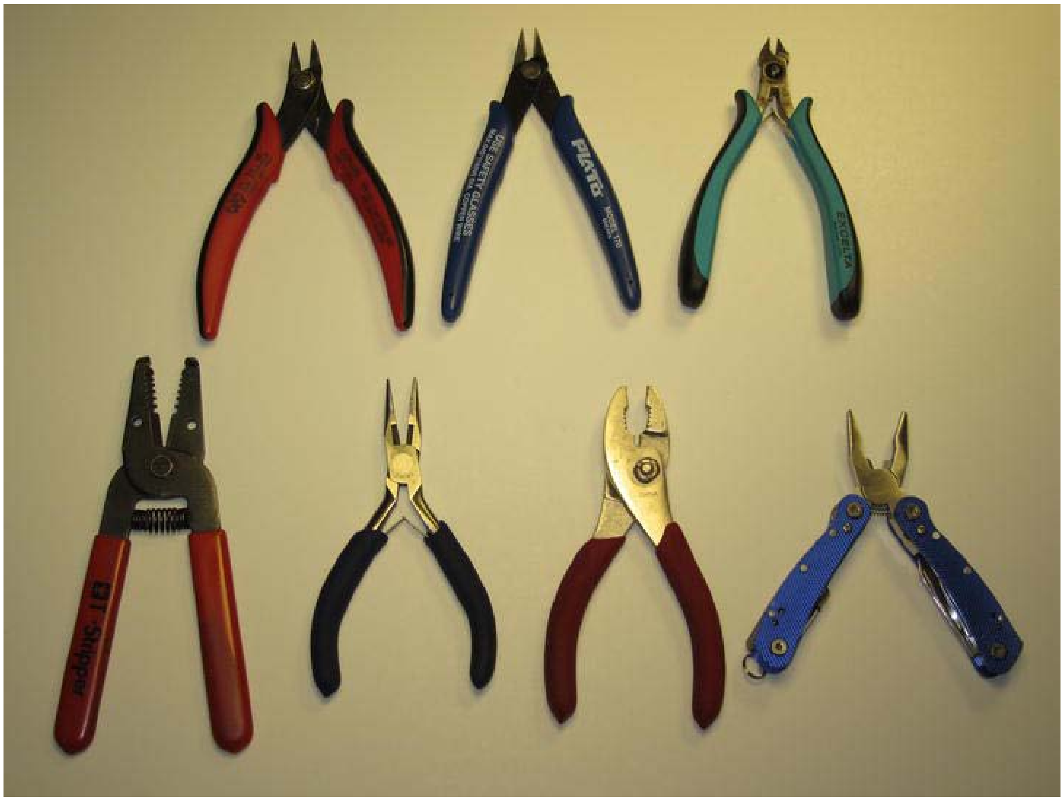

Figure 2-1. *Commonly available wire-cutting tools. The top row shows three styles of diagonal cutters, which are the preferred tools for cutting small-gauge copper wire. The bottom row shows some tools that offer wire cutting as a secondary facility, and should only be used when proper wire cutters are not available.*

The top row of cutting implements in Figure 2-1 are various forms of what are known as diagonal cutters, side cutters, or flush cutters. These cutters have beveled cutting edges that meet in the middle of the wire being cut. One side of the cutters is flat, or *flush*, allowing wires to be cut with some precision. These are the preferred cutters for small-gauge copper wire, as well as for clipping excess component leads from a printed circuit board (PCB). See Figure 2-2.

Compare the cutting action of the side cutters to the other, less optimized wire cutters in the second row of cutters in Figure 2-1. These cutters either use a shearing action to cut the wire or include a flush-cutting area as a secondary feature.

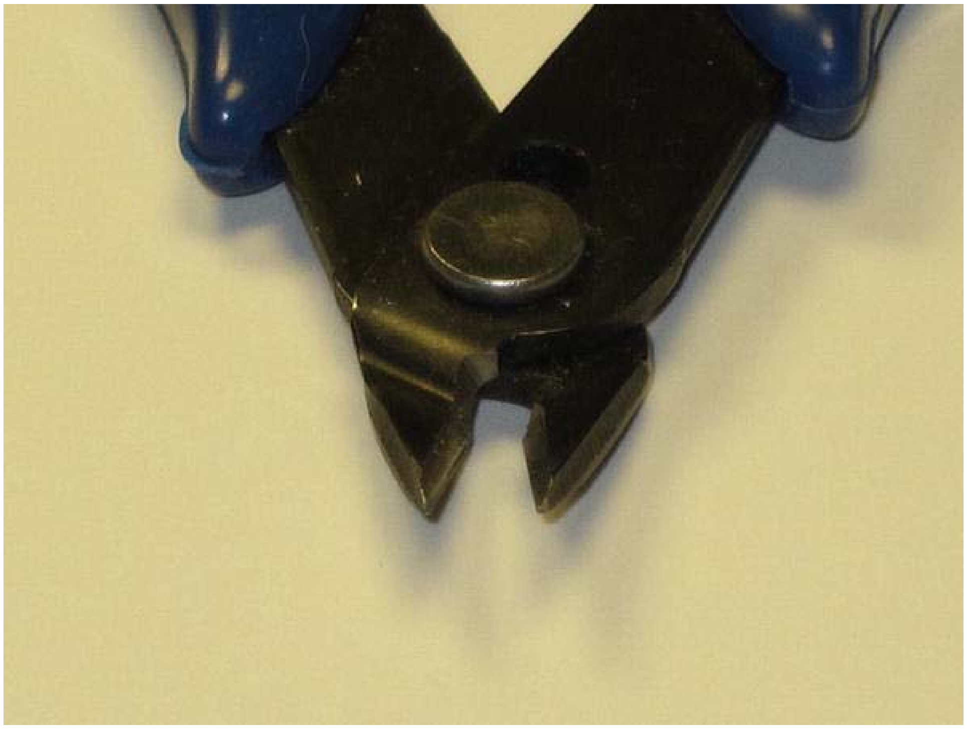

Figure 2-2. *A close-up look at the business end of a pair of side cutters. When the two cutting edges meet in the middle, the bottom side is flat, or flush, allowing the user to get quite close to where a wire or component lead emerges from a solder joint.*

Use caution when clipping component leads or small wires with any wire cutters. The cut lead can easily fly away at sufficient velocity to poke you right in the eye or other sensitive area. If possible, hold *both* ends of a wire being cut or direct the cut leads *away* from your face.

■ **Caution** Use eye protection when clipping leads. Don't say you weren't warned about this hazard.

Wire Strippers

Wire strippers are tools that allow you to easily strip the nonconductive insulation from wires. You can try chewing it off with your remaining teeth, but this is not recommended.

Wire strippers usually consist of a pair of hinged blades with some sort of machined notch. The notch allows the blades to cut the insulation around the wire without cutting or nicking the underlying conductor. A good example is shown in Figure 2-3.

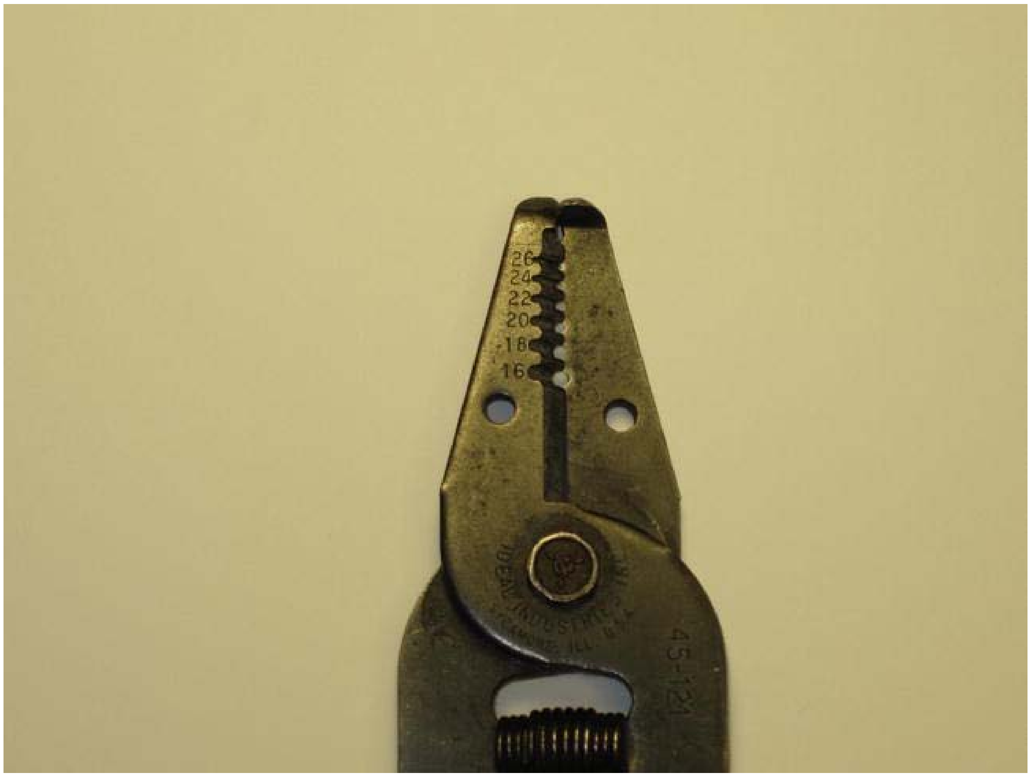

Figure 2-3. *The popular T-Stripper combination wire cutter and stripper tool from Ideal Industries has been around for over half a century. Note the wire gauge sizes stamped into the face of the cutter blade. New models feature laser-engraved text.*

The wire strippers in Figure 2-3 can easily strip the insulation from wires with a conductor diameter, or *gauge*, between 16 (0.0508" or 1.291 mm) and 26 (0.0159" or 0.405 mm). The higher-numbered wire gauges represent *smaller*-diameter wires. Once upon a time, the gauge number represented the number of *drawing* operations that were performed on a wire when being manufactured. The more times it was drawn through the processing equipment, the thinner the wire became. Today, the wire-manufacturing industry in the United States and Canada has adopted the American Wire Gauge standard (AWG). See Table 2-1.

Table 2-1. *The AWG. A more complete listing is available at http://en.wikipedia.org/wiki/American_wire_gauge, from which this partial list was obtained.*

AWG	Diameter Area			Copper Resistance	
	Inch mm		mm^2	Ω/km mΩ/m	Ω/1000 ft mΩ/ft
0000 (4/0) 0.46	00	11.684 107		0.1608	0.04901
000 (3/0)	0.4096	10.404	85.0 0.20	28	0.06180
00 (2/0)	0.3648	9.266	67.4 0.25	57	0.07793
0 (1/0)	0.3249	8.252	53.5 0.32	24	0.09827
1 0.28	93	7.348	42.4 0.40	66	0.1239
2 0.25	76	6.544	33.6 0.51	27	0.1563
3 0.22	94	5.827	26.7 0.64	65	0.1970
4 0.20	43	5.189	21.2 0.81	52	0.2485
5 0.18	19	4.621	16.8 1.02	8	0.3133
6 0.16	20	4.115	13.3 1.29	6	0.3951
7 0.14	43	3.665	10.5 1.63	4	0.4982
8 0.12	85	3.264	8.37 2.06	1	0.6282
10 0.10	19	2.588	5.26 3.27	7	0.9989
12 0.08	08	2.053	3.31 5.21	1	1.588
14 0.06	41	1.628	2.08 8.28	6	2.525
16 0.05	08	1.291	1.31 13.1	7	4.016
18 0.04	03	1.024	0.823 20.9	5	6.385
20 0.03	20	0.812	0.518 33.3	1	10.15

22 0.02	53	0.644	0.326 52	.9 6	16.14
24 0.02	01	0.511	0.205 84	.2 2	25.67
26 0.01	59	0.405	0.129	133.9	40.81
28	0.0126 0	.32 1	0.0810 2	12. 9	64.90
30	0.0100 0	.25 5	0.0509 3	38. 6	103.2

We'll return to this table again in Chapter 3 when the electrical properties of wire are discussed. For now, note that as the wire gauge number goes *up*, the wire's physical diameter goes *down*. Similarly, as a conductor's physical size goes down, its ability to carry current also goes down, and its electrical *resistance* goes up.

■ **Note** Other standards exist throughout the world. Steel-wire manufacturers use a different scale, as their interest is more on the tensile strength of the wire, not its electrical properties.

Note that the AWG is for single-strand copper wire. Equivalent gauges in stranded wire are physically a bit larger, although they have the same effective cross-sectional area and can carry the same amount of current.

When using wire strippers, keep in mind that your goal is to cut and remove only the nonconductive insulation from the wire, leaving the internal conductor intact. Even a small nick can result in a weak point in the wire, both electrically and mechanically.

Practice Makes Perfect

The best way to attain proficiency with a new tool is with practice—lots and lots of practice. This is true for working with your hands as well as with that big brain of yours. Ideally, any task worth doing will involve at least a little bit of both manual dexterity and focused brain power.

Cutting and stripping wires is an excellent example. For your first practice assignment, spend some time turning long, beautiful lengths of wire into little bits of wire with stripped ends.

CUTTING AND STRIPPING WIRE

For this exercise, you're going to need the following:

- Wire cutters
- Wire strippers
- Much wire
- A comfortable, well-lit location

Cut the Wire into Segments

Don't watch TV or be otherwise distracted. Pay attention to what you are doing.

Begin by cutting the wire into various-sized segments from 2" to 6" in length. Accuracy in length is not critical at this point. You need to get a good feel for your wire cutters and how much force is needed to cut the wire, without cutting yourself or others.

Collect the cut wire segments into a bowl or small container for the next step.

Strip the Wire Segments

Using your wire strippers, remove about 1/4" (6 mm) of insulation from both ends of each wire. Try not to cut or nick the wire core. The insulation that you remove can either be discarded or collected for use in some sort of art or craft project.

Once you've completed the cutting and stripping portion of this exercise, sort the wires according to their length, or their color if you used different colors of wire. Compare the quality and consistency of your work by looking at the first few wires and the last ones you did. Do you see any improvements in your work?

If you spend about an hour working on your wire-cutting and wire-stripping technique, you will perhaps master a skill that will prove useful for the rest of your life, as well as prepare a big pile of wires that you can use in many of the sample projects illustrated in this book.

Pliers and Tweezers

Every day, the size of the "typical" electronic component gets smaller and smaller. This miniaturization process has brought about a world where you can carry a phenomenally powerful computer in your pocket, and sometimes even make phone calls with it. The individual transistors, resistors, capacitors, and other components inside such modern marvels are literally microscopic in size. You need a fairly powerful microscope to be able to see them at all.

In your typical home electronics lab, and even in many professional labs, you will not be dealing with these tiny things, except as large clumps all tied together inside a chip or other framework. Still, the typical components you will use are often a bit too small for all but the tiniest of fingers. That's where pliers and tweezers come into play. They are an extension of your own hands and fingers, increasing your precision and allowing you to perform delicate tasks with tiny parts.

There are several types of pliers and tweezers that you may use on a daily basis in your lab. See Figure 2-4.

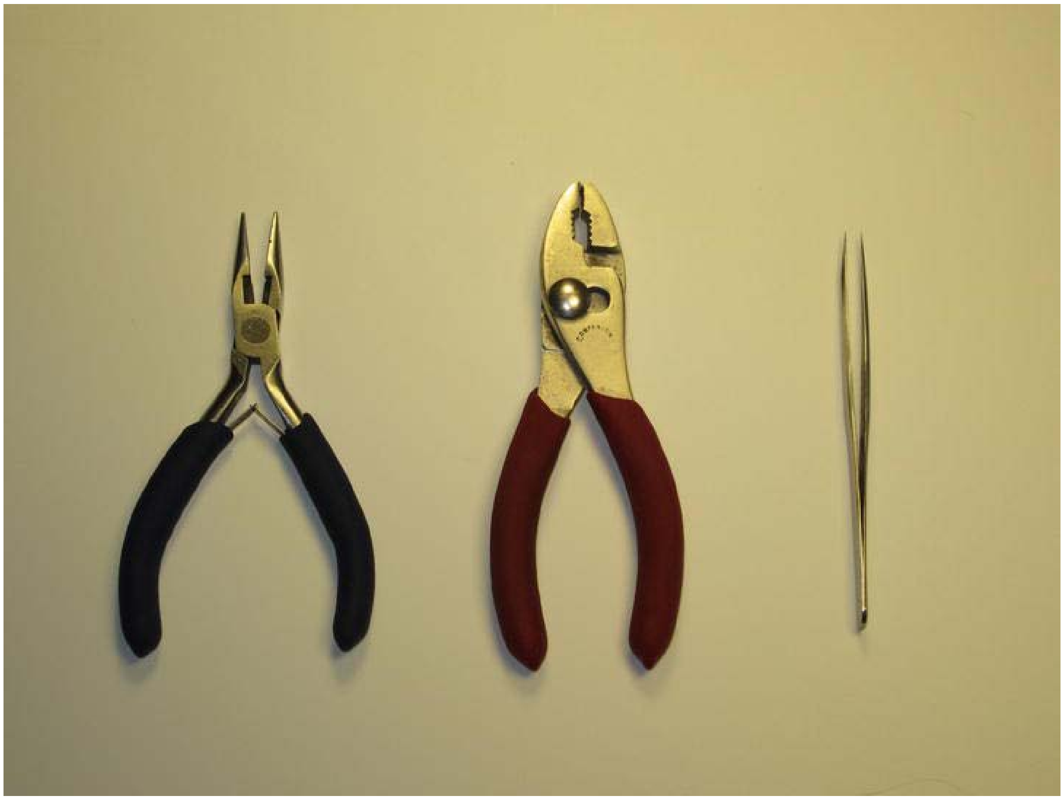

Figure 2-4. *Pliers and tweezers aid in working with small parts. On the left is a pair of needle-nosed pliers with cushioned grips and leaf springs. In the center is a pair of slip-joint pliers that can be adjusted to accommodate two different ranges of gripping action. On the right is a pair of needle-sharp tweezers, which can be used to pick up and place even the tiniest of electronic components—if your hand is steady enough!*

Screwdrivers, Nut Drivers, Wrenches, and Ratchets

A lot of electrical connections can be made using only a screwdriver. Figure 2-5 shows an electrical connection using a *barrier strip*. A specialized connector is crimped onto the end of the wire and held in place with a screw. A variety of crimp-on connectors are available and should be installed using the appropriate tool, as indicated by the connector's manufacturer. Just smashing the connector onto the wire with some pliers is not recommended.

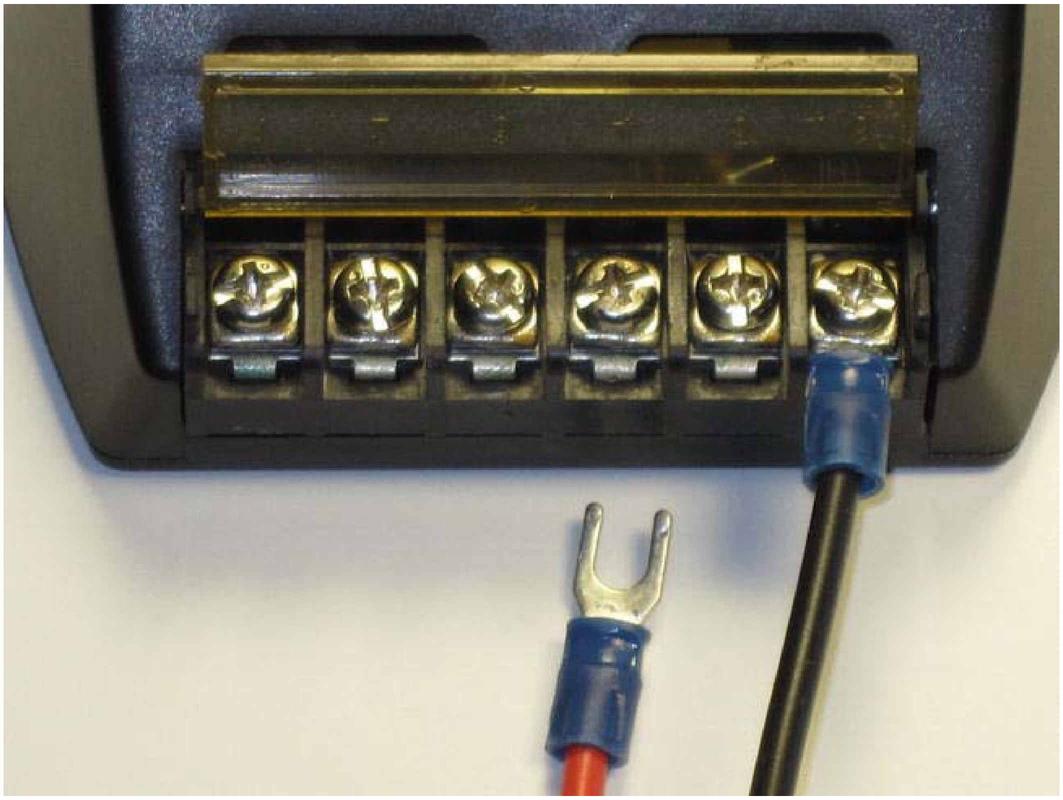

Figure 2-5. *A barrier strip on a low-cost solar-charge controller. No soldering required! The specialized connector is held in place with a screw. The connector is affixed to the end of the wire by deformation using a specially designed crimping tool. You can see the indentations in the insulating collar around the connector.*

Another type of connector that is tightened with a screw is called a *terminal block*. Unlike a barrier strip, no specialized connector is required on the wire. You can simply insert a stripped wire into the terminal block and tighten the connector using a screwdriver. Try to strip the wire only enough to allow the bare conductor to enter the terminal block, without leaving any noninsulated wire exposed that could accidentally short out against another wire. See Figure 2-6.

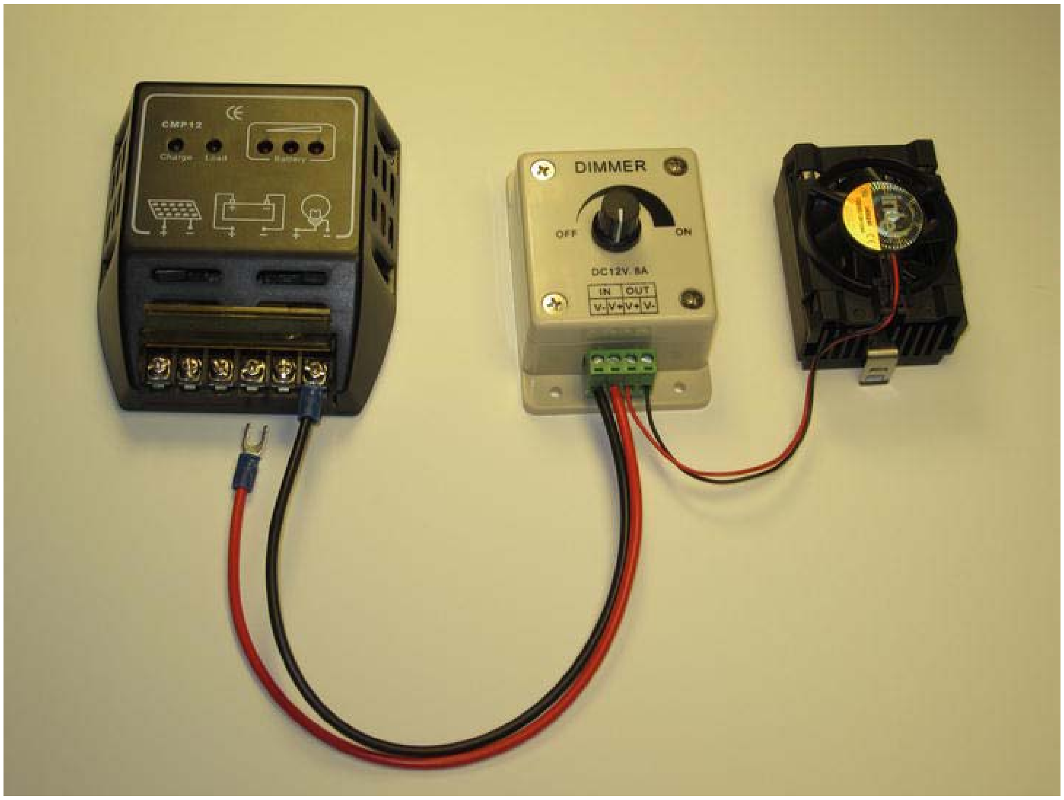
CMP12
Charge
Load
Battery
DIMMER
OFF
ON
DC12V. 8A
IN
OUT
V-
V+
V+
V-

Figure 2-6. *A terminal block uses a screw for tightening the connection with a bare wire.*

Besides being useful for making electrical connections using barrier strips and terminal blocks, screwdrivers are also quite handy in their original function, which is to tighten and loosen mechanical fastening devices such as screws and bolts. See Figure 2-7.

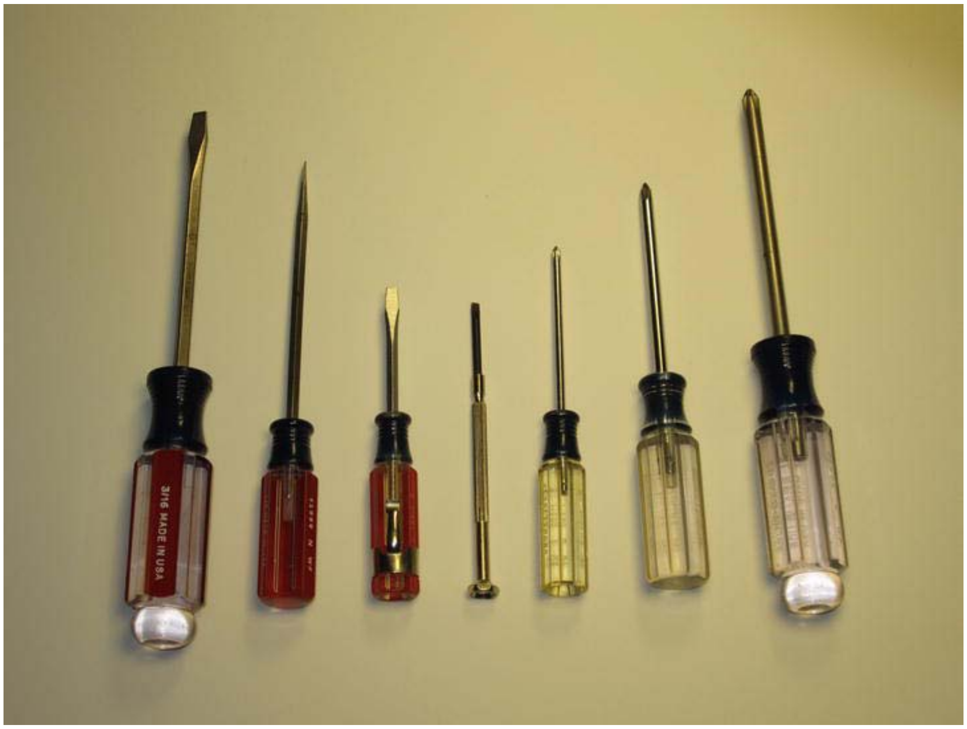

Figure 2-7. *A variety of screwdrivers. The four on the left are standard, or straight-bladed, screwdrivers. The three on the right are Phillips-style screwdrivers. Other specialty bits are also available.*

You should only use screwdrivers to drive screws and bolts. Don't try to use them as levers, pry bars, chisels, punches, or other tools. When you need a chisel, use a chisel. Keep your screwdrivers clean and dry. Put them back where they belong when you're done with them. You get to decide where they belong.

Screws, Bolts, and Nuts

There are a *very* large number of different kinds of screws and bolts. Some screws, such as wood screws, are made to be driven directly into wood and create their own threads in the material as they are installed. Other screws, largely referred to as machine screws, mate with a premachined or tapped hole, or a machine nut.

Nuts are often shaped with four or six sides that allow them to be held in place while the screw is turned. Alternately, the nuts themselves are turned while the screw is held in place. Some screws and bolts have a similarly shaped head that allows them to be turned with a wrench or nut driver instead of a traditional screwdriver.

Please resist the urge to overtighten screws and bolts. The mechanical advantage gained by the inclined plane of the screw thread multiplies the force used to turn the screw into a terrific amount of pressure on the screw itself. You can easily damage either the screw or the threads in the material to be fastened by overtightening a screw.

More Specialized Hand Tools

A good set of wire tools, pliers, and screwdrivers will go a long way when working with electronics. However, sometimes you need a more specialized tool for more specialized work.

Here are some of the more common tools that are a little more focused on a particular task than they are for general-purpose usage.

Wire Crimpers

As mentioned previously in this chapter, you can make solid electrical and mechanical connections using barrier strips and specially terminated wires. The optimized connectors used for this purpose require optimized tools for best performance.

Figure 2-8 depicts some examples of specialized crimp tools for electrical connections.

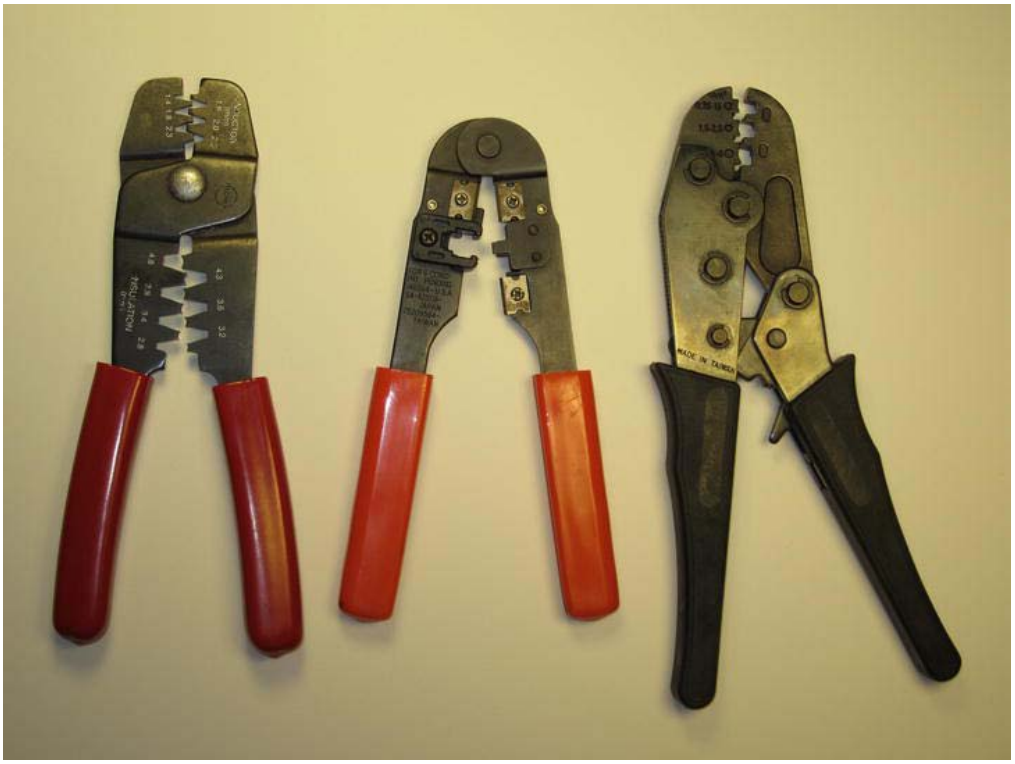

Figure 2-8. *Some common crimping tools for affixing electrical connectors to wire*

Referring again to Figure 2-8, the crimpers on the left are for crimping tiny pins to the ends of tiny wires that are then inserted into plastic connector bodies. The center crimp tool is for crimping modular jacks onto telephone cables. It also includes a wire-cutter blade (bottom) and a pair of controlled-depth stripper blades for removing only the outside layer of insulation from silver-satin flat telephone cord, without disturbing the individually insulated wires contained within. The crimp tool on the right features a ratcheting action that increases the crimping force applied to the connectors. This is the tool used to crimp the connectors show previously in Figure 2-5.

Computers

You might not think about computers as being "power tools" in your toolbox. As the size of computers, laptops, tablets, and smart phones decreases, and their computing power, networking ability, and storage capacities increase, they become much better candidates for use in the lab.

Traditional "desktop" personal computers (PCs), laptops, netbooks, and tablets can perform multiple roles in your laboratory setting. If you are lucky enough to have Internet access on your computer, it can become a conduit for all types of useful electronics information, in addition to its more accepted role as a source of funny cat videos. See Figure 2-9.

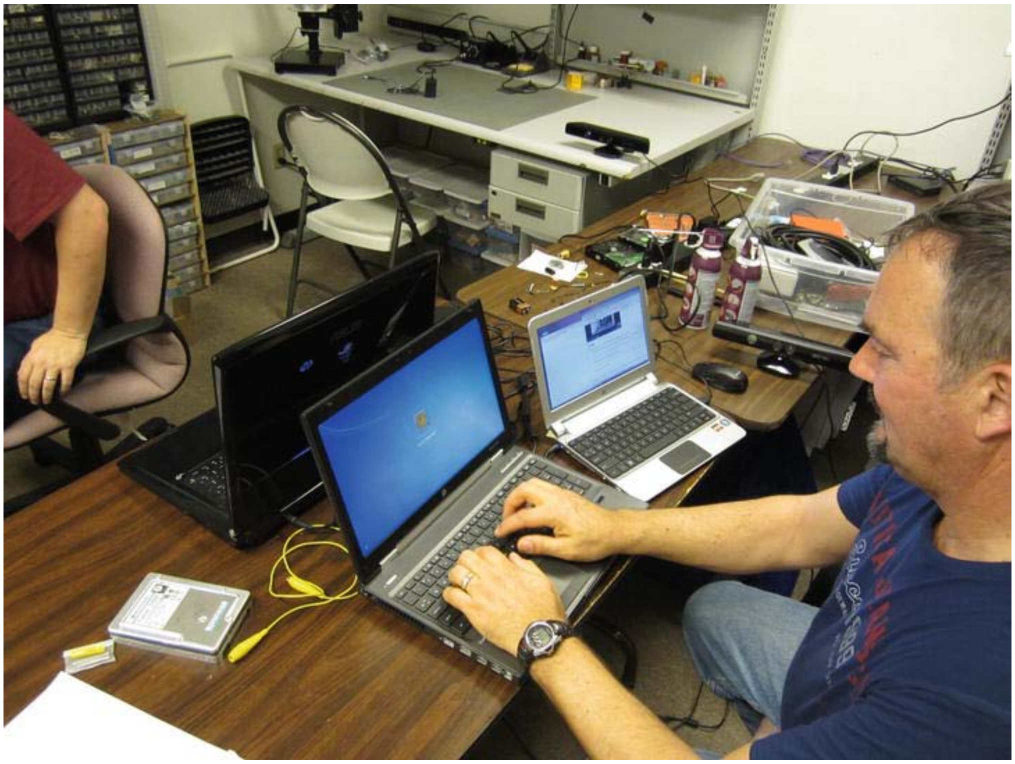

Figure 2-9. *Computers being used in the lab. Paul Bouchier of DPRG uses two laptops at once when testing new robot software with the Microsoft Kinect (just to the right of his mouse).*

Here are just a few of the many possible areas of practical knowledge, learning opportunities, and bits of interesting information that you can access using your computer:

> *Data sheets*: Almost all electronics manufacturers provide online access to a library of technical data sheets and application notes for their products. These should be your first choice for information about how and when to use their products. These sources also contain important product *errata*, which contains corrections and updates for products that have been made since their first publication.
>
> *Online tutorials*: The Web abounds with detailed and thought-provoking tutorials and background information on every topic imaginable, including, not surprisingly, mountains of information on computers and electronics.
>
> *User forums*: People tend to cluster around others with similar interests. Once upon a time when people where geographically limited and unable to communicate or interact with folks around the world, this severely restricted a lot of the more specialized areas of knowledge. Today, you can find user groups centered on every conceivable human endeavor (and some inconceivable ones, too) that are filled with helpful, knowledgeable members that are more than happy to answer questions and share experiences with you. Do try to be courteous, respectful, and fair when asking for help with your projects, and always try to repay in kind by helping those with similar questions that you've had to master.
>
> *Sharing/collaboration web sites*: As the cost and complexity of publishing your work on the Internet becomes more manageable, it becomes faster and easier to let others know about your progress, as well as your problems. You can easily set up your own web page for little or no cash, or join many of the available user groups that allow their members to post, browse, and otherwise benefit from the works of other similarly minded individuals.
>
> *Design tools*: You can find many free or reasonably priced electronic design and simulation tools online. Some manufacturers also provide free tools specifically for optimizing designs for their products.
>
> *Online suppliers*: Where you once had to walk (or drive) down to the store and see what was available in your area as far as electronic components and tools went, now you can browse through the catalogs of suppliers from all over the world, from the comfort of your own lab. Not only does this provide a much larger field from which to choose, it also levels the playing field, so to speak, forcing worldwide competition for your hard-earned cash. Bonus: Some manufacturers offer a limited numbers of free samples. Yip!

This list only scratches the surface of the kinds of information you can access using a computer. You can also use a computer as part of your electronic projects, and several examples of doing this are included later in the book.

Tools for Soldering

You don't have to learn how to solder to work with electronics, but it helps. It really does. It opens a whole new world of possible projects, repairs, and custom modifications to existing equipment. Just remember, as always, to keep your work area clean and safe. See Figure 2-10.

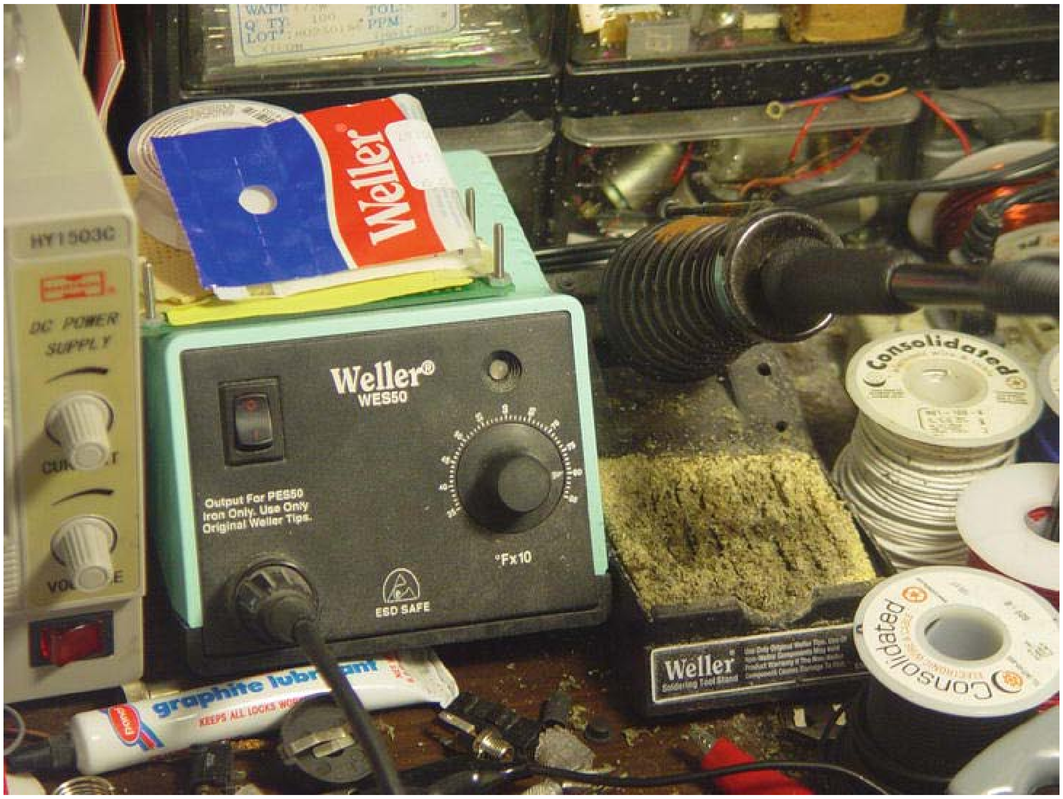
Weller
Weller
WES50
Output For PES50
Iron Only. Use Only
Original Weller Tips.
°Fx10
ESD SAFE
DC POWER
SUPPLY
HY1503C
Consolidated
graphite lubricant

Figure 2-10. *A Weller controlled-temperature soldering station. What a mess!*

Like any worthwhile skill, soldering is best mastered with lots and lots of practice. The best way to learn this useful skill is to have someone show you how it's done. It's difficult, but not impossible, to learn proper soldering techniques from a book. That being said, have a look at Figure 2-11.

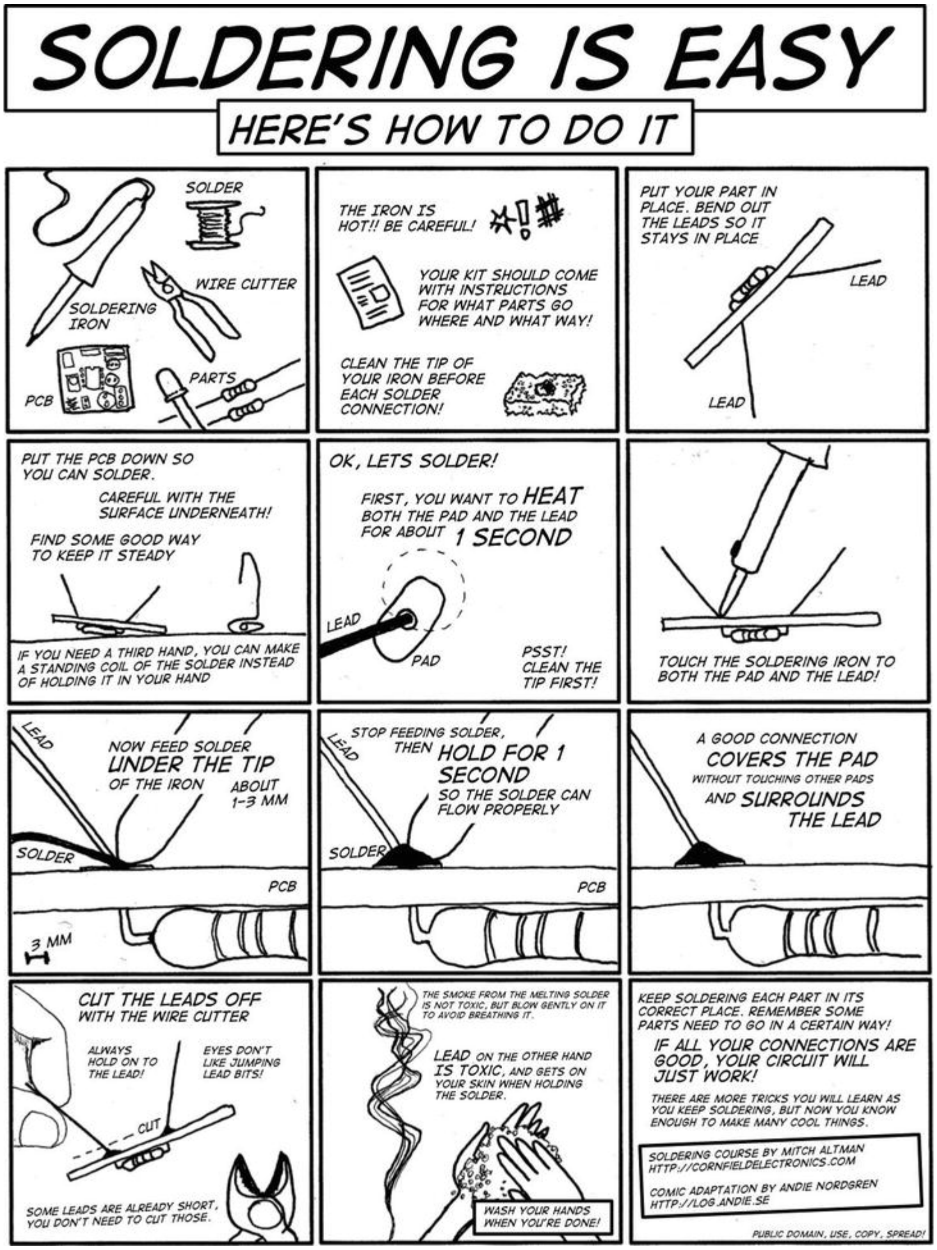
SOLDERING IS EASY
HERE'S HOW TO DO IT
SOLDER
WIRE CUTTER
SOLDERING IRON
PARTS
PCB
THE IRON IS HOT!! BE CAREFUL! *!#
YOUR KIT SHOULD COME WITH INSTRUCTIONS FOR WHAT PARTS GO WHERE AND WHAT WAY!
CLEAN THE TIP OF YOUR IRON BEFORE EACH SOLDER CONNECTION!
PUT YOUR PART IN PLACE. BEND OUT THE LEADS SO IT STAYS IN PLACE
LEAD
LEAD
PUT THE PCB DOWN SO YOU CAN SOLDER.
CAREFUL WITH THE SURFACE UNDERNEATH!
FIND SOME GOOD WAY TO KEEP IT STEADY
IF YOU NEED A THIRD HAND, YOU CAN MAKE A STANDING COIL OF THE SOLDER INSTEAD OF HOLDING IT IN YOUR HAND
OK, LETS SOLDER!
FIRST, YOU WANT TO HEAT BOTH THE PAD AND THE LEAD FOR ABOUT 1 SECOND
LEAD
PAD
PSST! CLEAN THE TIP FIRST!
TOUCH THE SOLDERING IRON TO BOTH THE PAD AND THE LEAD!
LEAD
NOW FEED SOLDER UNDER THE TIP OF THE IRON
ABOUT 1-3 MM
SOLDER
PCB
3 MM
LEAD
STOP FEEDING SOLDER, THEN HOLD FOR 1 SECOND SO THE SOLDER CAN FLOW PROPERLY
SOLDER
PCB
A GOOD CONNECTION COVERS THE PAD WITHOUT TOUCHING OTHER PADS AND SURROUNDS THE LEAD
CUT THE LEADS OFF WITH THE WIRE CUTTER
ALWAYS HOLD ON TO THE LEAD!
EYES DON'T LIKE JUMPING LEAD BITS!
CUT
SOME LEADS ARE ALREADY SHORT, YOU DON'T NEED TO CUT THOSE.
THE SMOKE FROM THE MELTING SOLDER IS NOT TOXIC, BUT BLOW GENTLY ON IT TO AVOID BREATHING IT.
LEAD ON THE OTHER HAND IS TOXIC, AND GETS ON YOUR SKIN WHEN HOLDING THE SOLDER.
WASH YOUR HANDS WHEN YOU'RE DONE!
KEEP SOLDERING EACH PART IN ITS CORRECT PLACE. REMEMBER SOME PARTS NEED TO GO IN A CERTAIN WAY!
IF ALL YOUR CONNECTIONS ARE GOOD, YOUR CIRCUIT WILL JUST WORK!
THERE ARE MORE TRICKS YOU WILL LEARN AS YOU KEEP SOLDERING, BUT NOW YOU KNOW ENOUGH TO MAKE MANY COOL THINGS.
SOLDERING COURSE BY MITCH ALTMAN HTTP://CORNFIELDELECTRONICS.COM
COMIC ADAPTATION BY ANDIE NORDGREN HTTP://LOG.ANDIE.SE
PUBLIC DOMAIN, USE, COPY, SPREAD!

***Figure 2-11.** Instructions for soldering. This delightful illustration by Andie Nordgren captures the basics of soldering as taught by soldering guru Mitch Altman (http://log.andie.se/post/397677855/soldering-is-easy).*

To solder, you will need a soldering iron. You will also need some solder and probably some wire cutters. That's just the tip of the iceberg, however. Like any good tool, your soldering station will soon accumulate many *accessories* to enhance your soldering prowess.

To begin your soldering adventure, you will need a *small* soldering iron, some solder, and a damp sponge. The big, pistol-grip style of soldering irons are for much larger projects, such as plumbing. Ideally you want something in the 15W to 50W range. See Figure 2-12.

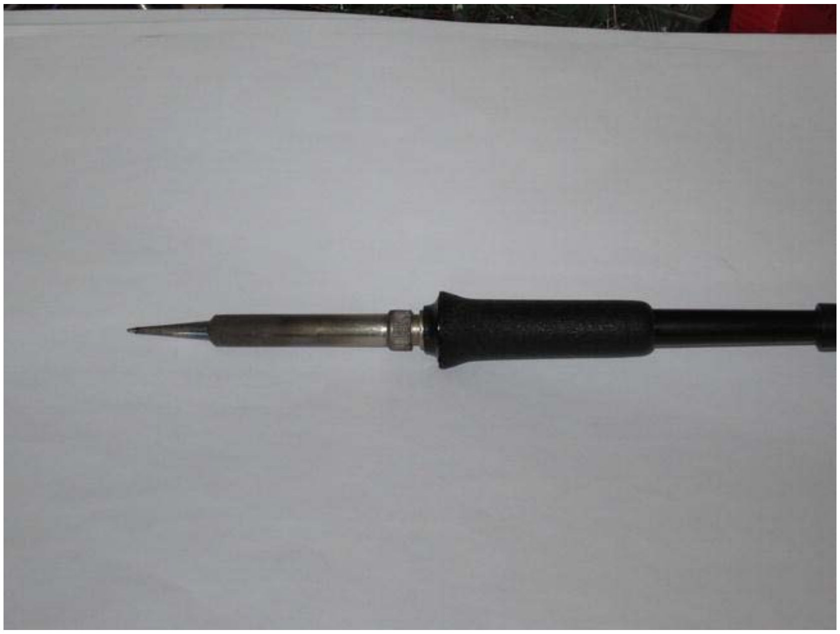

Figure 2-12. *The pencil-grip style of soldering irons is better suited for working on small electronics projects.*

The soldering iron in Figure 2-12 has a replaceable, conical tip. These tips are generally interchangeable for a particular model of soldering iron, but unfortunately not interchangeable between different models or manufacturers. Different tip shapes are also available, depending on what types of things you're wanting to solder.

The tip is usually made of copper, which is an excellent conductor of heat. Since copper really likes to combine with other metals at high temperatures (which is exactly what happens when soldering), most tips are coated in a thin layer of steel to give them much longer useful lifetimes. Tips still wear out, however, and should be replaced.

In this particular soldering iron, the next part of the iron is the removable barrel that holds the tip in place. This barrel is where the heating element is housed, so remember that it is just as hot as the tip! Try not to touch it. You will probably not want to make that mistake more than once.

The next section is the padded and insulated handle of the soldering iron, which is made of dense plastic foam. A quality soldering iron will be comfortable in your hand and won't produce excessive fatigue, even after hours of use. Lower-quality and lower-priced models will more rapidly take their toll on your hands.

Not shown in Figure 2-12 is the flexible power cord running back to the soldering iron's base. This cord supplies the electrical power that is converted to heat in the barrel, as well as some wires for a temperature sensor. This allows the unit to control the temperature of the tip with great accuracy. Note the temperature dial in Figure 2-10. Lower-cost units omit the temperature control features and run at a constant power output.

It will take some amount of time for the soldering iron tip to reach a working temperature. Likewise, it will also take some time to cool off once you've turned off the soldering iron.

■ **Caution** Soldering irons are very hot in use and can cause painful burns. Always treat them with respect!

The solder you use will be determined by what kind of soldering you want to do. For most small, electronics-related soldering projects, a small-gauge tin-lead or lead-free solder can be used. Bear in mind that lead is a poisonous material that slowly accumulates in your system, should you come into contact with it, as your body has no good way of eliminating it. Wash your hands before and after soldering, just to be safe.

Soldering works best when everything being soldered is sparklingly clean. To help facilitate this cleanliness, chemical compounds known as soldering rosin are used. These compounds chemically clean the surfaces being soldered when exposed to high levels of heat. Some types of solder have rosin incorporated into them. You can also obtain rosin as a paste into which you can dip wires and component leads before soldering. Other types can be applied with a dispensing pen.

■ **Caution** Don't breathe the smoke emitted during the soldering process. If possible, use a fan or fume extractor to redirect the soldering fumes away from you.

Once your soldering iron is up to its proper operating temperature, clean the tip by wiping it gently on a damp sponge. Your tips will last longer if you resist the urge to clean them before replacing the soldering iron in its holder. Even though it seems counterintuitive to "put it up dirty," this will more than double the life of your soldering iron tips. Just be sure to clean the tip *before* you solder anything.

Power Tools

Everybody loves power tools, don't they? Sometimes it's enough to just *imagine* that you might need one to justify a trip to the store and subsequent purchase. While this may be human nature (or perhaps only the nature of one particular gender), it doesn't fully justify having one of each in the lab, *just in case*.

Drills and Drill Bits

You will, however, find a few particular tools to be especially useful in your lab. A good hand drill or drill press is one of those tools. There seems to be no end of things electronical that need a few more holes drilled into them. Electronic enclosures and control panels need good, clean holes drilled into them. If you make your own PCBs at home, you will find that they need *lots* of holes drilled into them.

Like any power tool, an electric drill needs and deserves your respect. You'd be surprised how much unintentional damage you can cause with such a simple device. Observe these simple safety rules when operating a power drill:

- Keep your work area clean and well illuminated.
- Remove any watches or jewelry you might be wearing.
- If you have long hair, keep it pulled back and tied out of the way.
- Always clamp down any work to be drilled on a drill press.
- Always remove drill bits from the drill chuck after use.
- Discard dull or rusted drill bits that cannot be resharpened.
- Always wear safety glasses when operating power tools.

Hand drills come in all shapes, sizes, and price ranges. Some are powered by hand cranks and some are electric. Rechargeable batteries make for very portable drills that you can carry almost anywhere. See Figure 2-13. The only drawback is their limited charge time. That's why you should always have at least one spare battery charged up and ready to go. Electric drills with wires attached don't have this problem. Just don't trip over the cord!

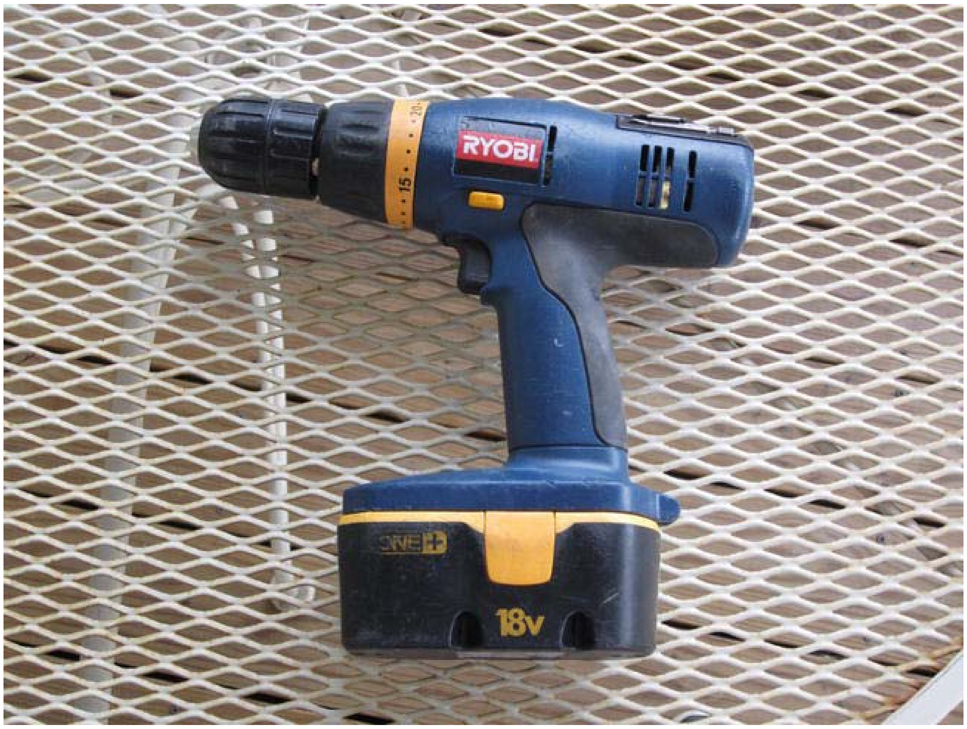

Figure 2-13. *A rechargeable hand drill. This model is both variable speed and reversible, which are two very desirable features in a hand drill.*

■ **Note** Power drills are also very handy when used with screwdriver bits. They can save you time and many blisters!

Drill bits likewise come in all sizes and shapes. You can buy sets with a variety of sizes, and these will usually come with a handy storage container that conveniently labels all the sizes for you. This only works, however, if you take the time to return the drill bits to their proper location once you've finished working with them.

A drill press has a large frame that facilitates drilling a precisely aligned hole, over and over again. You can make spindle-speed adjustments to drill presses (even small ones, such as the one pictured in Figure 2-14) either through an electronic speed control or by changing up the belts and pulleys that drive the spindle. There are even small frames for mounting handheld drills or rotary tools, such as the Dremel brand of tools. You can often, however, find actual, dedicated drill presses on sale for about the same price as the daintier Dremel tools.

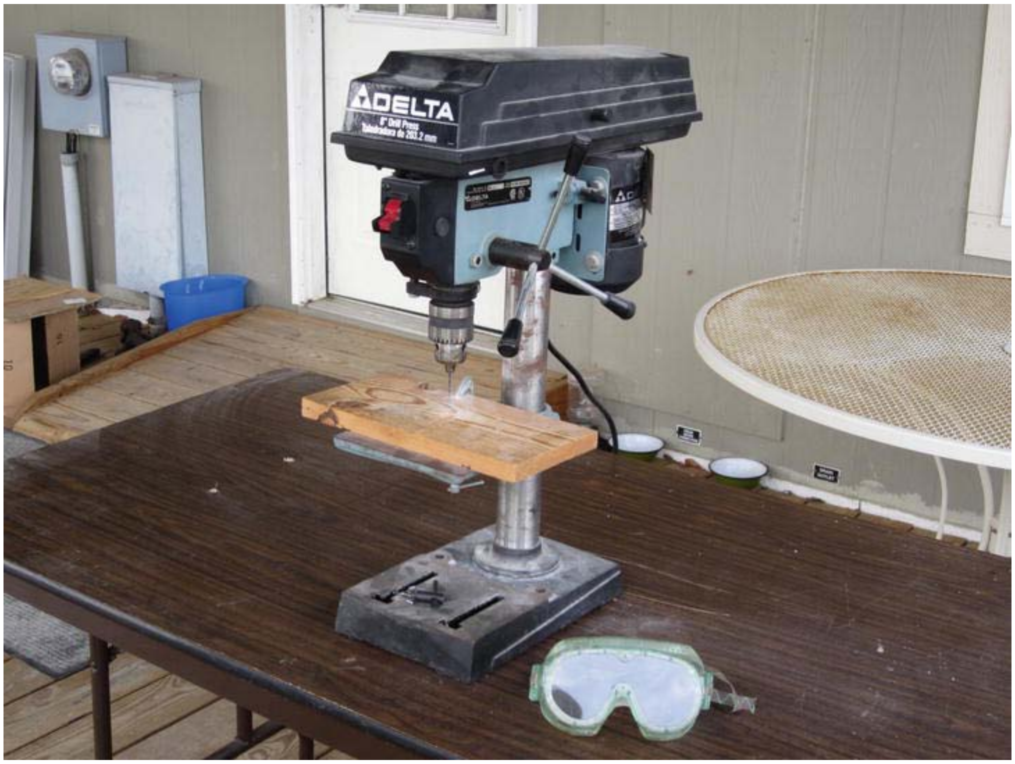

Figure 2-14. *A small tabletop drill press*

■ **Caution** When working with drills and presses, always keep your work area clear. Always wear safety glasses when using power tools. Try to keep the chuck key handy. Don't leave drill bits mounted in the chuck when not in use. Many chuck keys have built-in springs that prevent them from remaining in the chuck. Those that don't risk being accidentally left in the chuck when the drill its turned on, which can either fling the chuck across the workshop or make a big mess of your favorite hands.

Laser Cutters and 3D Printers

While not everyone can afford some of the more expensive and esoteric tools that are available, some of these tools are getting more popular than others; excellent examples are laser cutters and 3D printers. See Figures 2-15 and 2-16.

Laser cutters and engravers use a powerful carbon-dioxide laser to cut through wood and non-chlorine-based plastics. The laser beam is focused onto the work surface using mirrors that are mounted in an XY gantry, allowing the cutting area to be controlled by a computer. Typical installations require venting to the outside because of the various aromas produced in the process, as well as a water-cooling system for the laser tube itself.

There are also some newer laser cutters being built by hobbyists from scratch.

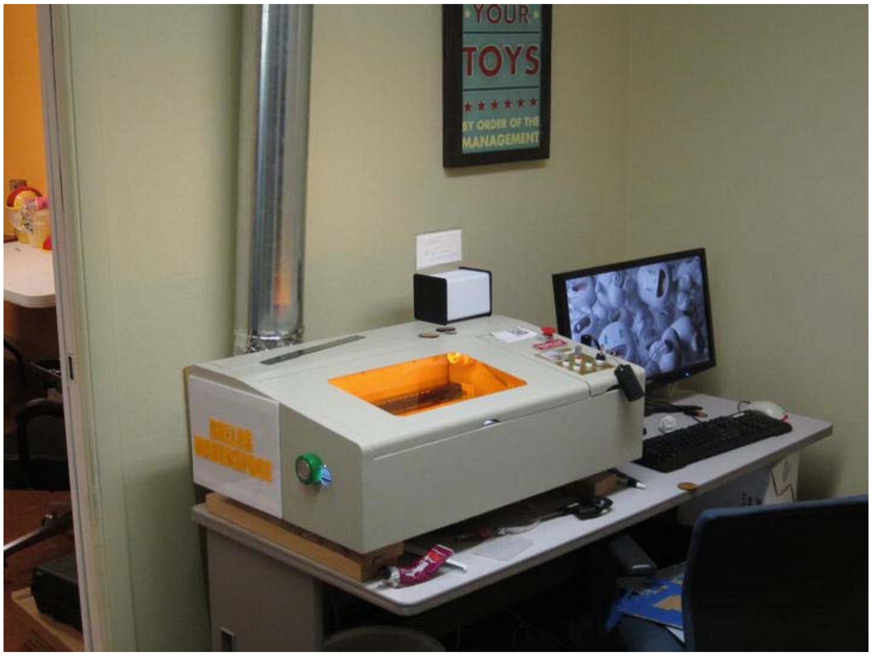

Figure 2-15. *A laser cutter. A powerful 40W carbon-dioxide laser is used to cut through wood and non-chlorine-based plastic, all under computer control. It can also be used as an engraver. This model is manufactured by Full Spectrum Engineering.*

3D printers seem to be all the rage these days. We're all pretty accustomed to the idea of printing on paper with a computer, although it's still a pretty recent innovation, as far as the history of printing goes. The idea of printing a solid object, on the other hand, is considered to be something of a game-changer.

The simplest 3D printers use a plastic extrusion and a motorized gantry to operate as a kind of a robotic glue gun. A plastic filament is fed into a heated fixture that is moved around a print bed, leaving a trail of cooling, gooey plastic. Either the extruder or the print bed is moved using motors under computer control to build up a layer at a time of the desired object.

Higher-technology 3D printers use this and other techniques to produce tangible, solid objects. Some use layers of fine powder that is glued together with a binder dispensed from a print head, similar to an inkjet printer. As each layer is printed, the object is lowered and another layer is deposited. This can even be done with sugar, as in the case of the CandyFab 4000, designed by the clever folks at Evil Mad Scientist Laboratories (`www.evilmadscientist.com/article.php/candyfab`). A stream of heated air is used to fuse a layer of sugar into a single solid structure, and an object is built up of successive layers.

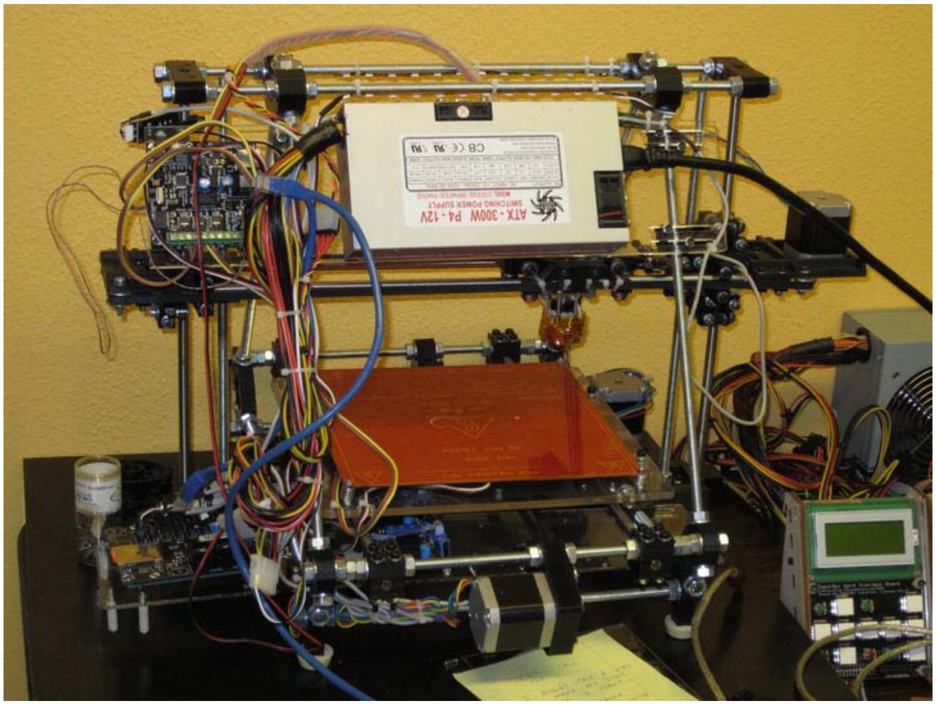

***Figure 2-16.** A 3D printer at Dallas Makerspace. This machine uses a heated extruder to build 3D objects out of plastic.*

Basic Test and Measurement Equipment

When it is behaving itself, electricity is generally invisible. To "see" what it is doing, or not doing, you will need some specialized tools. These fall into the two general categories of *test* and *measurement* equipment.

The line between testing and measuring is sometimes blurry. One possible distinction might be that testing verifies proper operation within acceptable parameters, while measurement actually assigns values to those parameters. An example of testing is, Does that flashlight work? The answer is yes or no, arrived at by the simple expedient of turning it on and observing the results. An example of measurement is, How much current does that flashlight draw? The answer might be 0.78A, and this would be determined using the appropriate measurement device, which in this case could be an ammeter.

Inspection Equipment

Sometimes, however, measuring things isn't hard at all, especially when the things being measured are actual, tangible items that you can hold in your hand. What you will need are some basic tools, primarily to be able to *see* what it is that you are measuring, and secondarily to take accurate measurements of them, at least in regard to their physical dimension. This comes in very handy when trying to identify components during repair or equipment upgrades, when the original technical documentation is not at hand.

Even the youngest and finest of eyeballs will need help with the teeny-tiny components that are available nowadays. You should try to find at least a small magnifying glass to examine fine details. If your budget allows, you might want to consider a microscope of modest power to look at your handiwork. Careful shopping will result in finding some older microscopes at bargain prices. Expect to pay a premium price for the shiny, new stuff. See Figure 2-17.

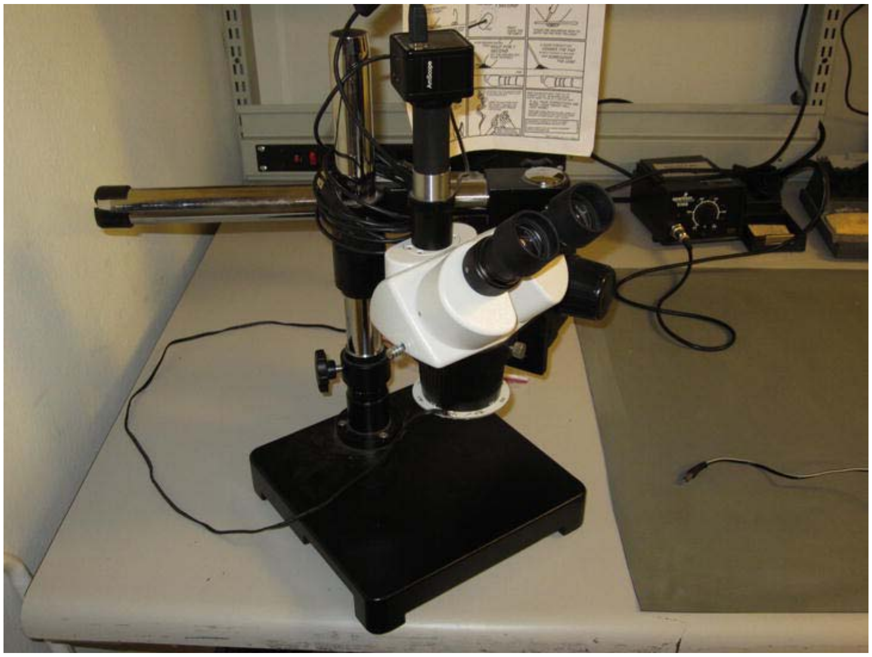

Figure 2-17. *A very nice stereo microscope. The power cord running to the white ring (partially hidden) supplies power to a ring of white LEDs that help illuminate the field of view. The base is very heavy to help stabilize the microscope.*

Even a humble magnifying lens can be very helpful in the lab, especially if it has a built-in light. See Figure 2-18.

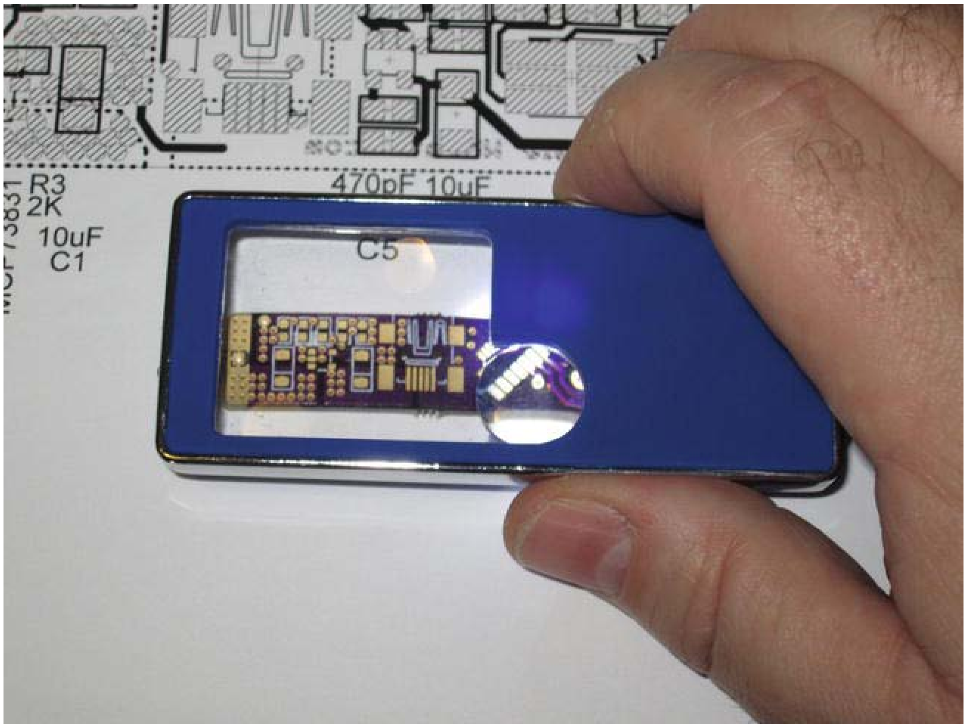

Figure 2-18. *A small, handheld magnifier with a built-in LED light is very handy for examining small features and details, such as are found on many PCBs.*

Physical Measurement Devices

Once you know what something looks like, at extreme close-ups, can you describe in absolute terms how big it is?

Take a look at Figure 2-19. Here we have, from the top to the bottom, a traditional tape measure, marked in both English and metric units (fractional inches and decimal centimeters), a digital caliper that can render measurements in either system at the touch of a button, and an architect's scale. While it may not be evident in the top-down view, the architect's scale is triangular in cross-section, resulting in not two but *six* different measurement options. This is mostly due to architects liking to draw things to scale, instead of full-sized.

The digital caliper is interesting in that it also provides a computer interface so that its readings can be pulled in by a computing device of some sort. These devices used to be quite expensive, but mass production has brought their prices down significantly.

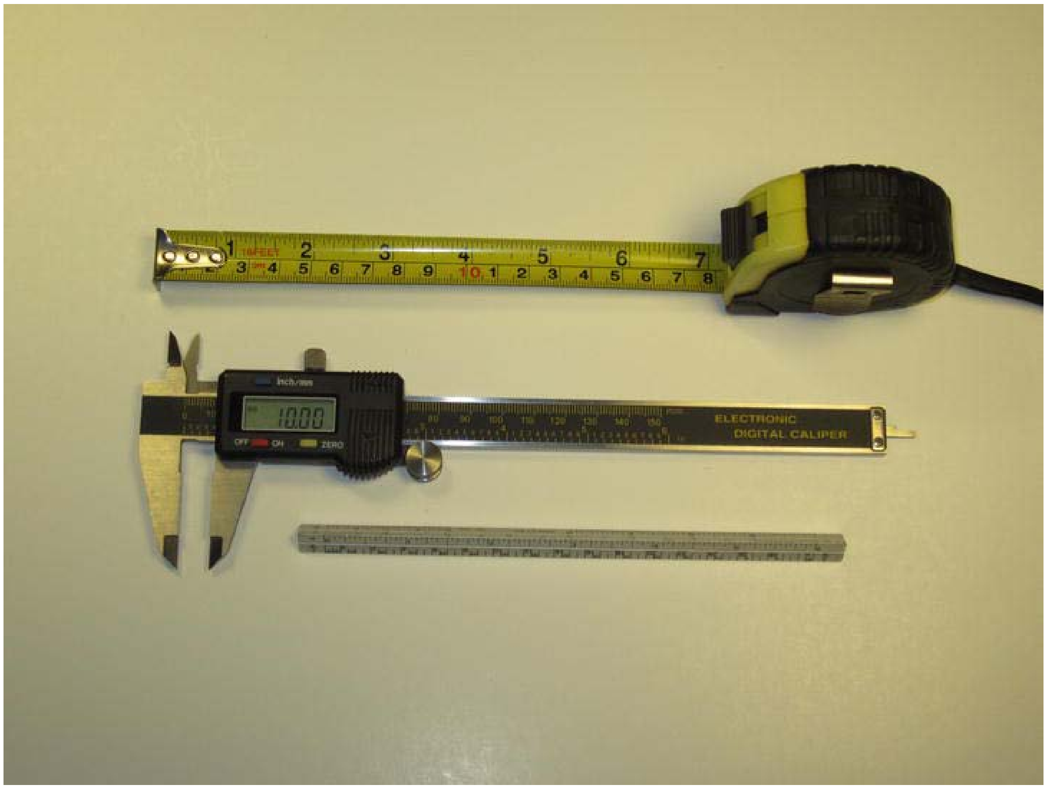

Figure 2-19. *From top to bottom, a tape measure, a set of digital calipers, and an architect's scale.*

Simple Electronic Measurement Tools

Now that you've got the tools to examine what these things look like and how big or small they are, let's move on to the more mysterious art of measuring their electrical properties.

Voltmeters

Voltage is a measure of electromotive force, and the standard unit of measure is the *volt*, named after Alessandro Volta, an Italian physicist. Voltage is sometimes compared to the pressure of water flowing through a garden hose, while amperage is compared to the amount of water flowing through the hose at any given time. It's not particularly useful to push this water analogy too far, but it's helpful in the beginning.

Voltage can be accurately measured with, you guessed it, a voltmeter. See Figure 2-20.

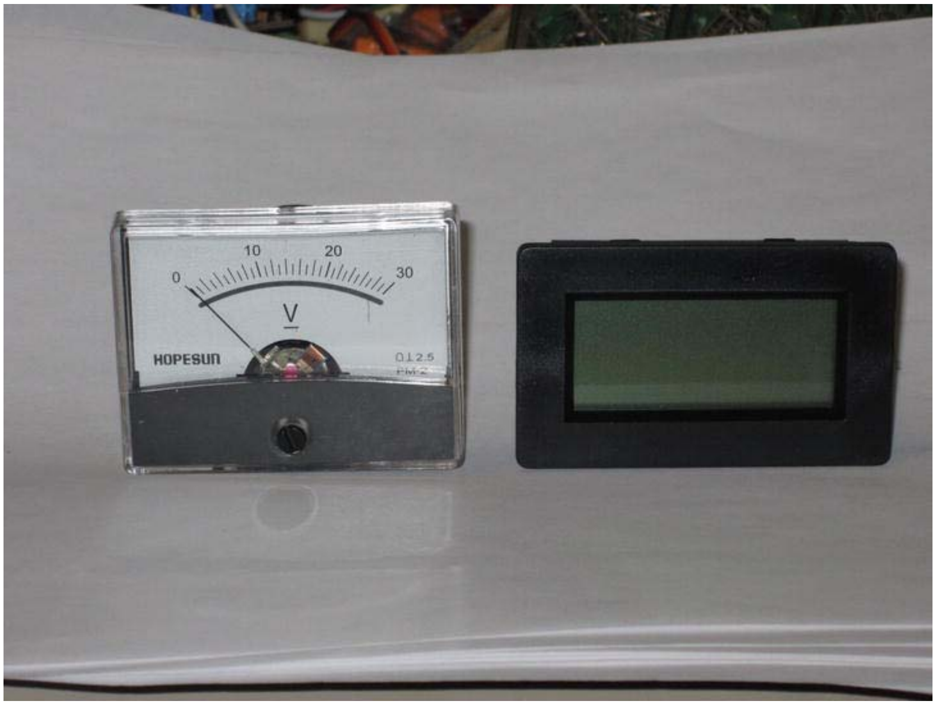

Figure 2-20. *A panel-mount analog voltmeter (left) and a digital version (right)*

A voltmeter measures the voltage, or *potential difference*, in a circuit across two points. Sometimes one of the points is a fixed reference point, such as the electrical *ground* mentioned in many circuits. This is often just a convenience for the purpose of measuring things, as the ground in an electrical circuit may or may not be connected to the ground under your feet.

A voltmeter will have two terminals, labeled positive (+) and negative (–). Looking at the analog voltmeter in Figure 2-20 (the one on the left), we see a needle that can swing from 0 on the left to 30 on the right. The prominent capital *V* on the dial indicates that the unit of measure is the volt. This indicates that you can use this meter to directly read from 0 to 30 volts, just by placing its terminals across a voltage. The meter itself is electromechanical in design, and creates a magnetic field that works against a spring and a fixed magnet. This action causes the needle to move to a point generally corresponding to the labeled voltage on the dial. These needles are not especially accurate, but can be corrected using the screwdriver adjustment at the bottom of the meter.

An analog panel meter such as this one does not require a separate power supply for proper operation. It draws all the power it needs from the signal being measured. This makes it unsuitable for measuring low-power circuits, where the current draw from the meter would unduly influence the reading.

The panel meter on the right is a digital circuit that measures the voltage and displays it on the built-in liquid crystal display (LCD). It requires a separate power supply for normal operation.

Ammeters

An *ammeter* looks and operates just like a voltmeter, except that it measures electrical current flow instead of voltage. The flow of electrical current is measured in units called *amperes* (A). The basic mechanism is almost identical to a voltmeter. To convert a current flow into a voltage, an electrical device known as a *shunt* is used. The shunt is simply a fixed-value resistor that produces a voltage across its terminals in relation to the amount of current that is flowing through it and its inherent resistance to the flow of current.

This relationship between voltage, current, and resistance is elegantly expressed in *Ohm's Law*. If you know two of the measurements in a circuit, you can calculate the third, based on their fixed and proportional relationship to each other. You will find this law to be exceedingly helpful to you when designing, repairing, and upgrading electronics.

To avoid using fractions, which seem to terrify so many people for no good reason, you can use Ohm's Law thusly:

- Voltage (volts) = current (amps) × resistance (ohms)
- Current (amps) = voltage (volts) ÷ resistance (ohms)
- Resistance (ohms) = voltage (volts) ÷ current (amps)

That's all there is to it, if you're just working with volts, amps, and ohms. Things can get weird, however, when you start to work with alternating current (AC), which is the kind of electricity you find in wall outlets and audio circuits.

Ohmmeters

An excellent application of Ohm's Law is found in building a resistance meter, also called an ohmmeter, because it measures resistance. Resistance is measured in units called ohms, often abbreviated with the capital Greek letter omega (Ω).

There are two ways to do measure resistance. One is to pass a known voltage though the unknown resistance and measure the current that is flowing in the circuit. The other is to pass a fixed current though an unknown resistance and measure the voltage across it. This second method is the one most commonly used.

You're most likely going to use a multimeter to measure resistance, as this is one of the more common functions of these handy devices. Mutlimeters are covered in the very next section.

Advanced Test and Measurement Equipment

It would seem that for simple circuits, having just a voltmeter, an ammeter, and an ohmmeter would be all you need. For the most part, this is correct. However, have you considered what would happen if you applied 100V to the 30V meter shown in Figure 2-20? What if you simply reversed the polarity of the signal by accident? Something bad for the meter, that's for sure. It might not actually catch on fire *right away*, but it would certainly damage the delicate, spring-loaded mechanism, possibly rendering it useless as a reliable measuring instrument.

The same can be said for an unprotected ammeter or ohmmeter. Luckily, the nice people that make good test equipment stay up late at night, thinking of the stupid things we might do with their products, and protect them accordingly. Let's look at some meters that measure multiple electrical characteristics, which strangely enough are called *multimeters*.

Multimeters

Multimeters will almost always allow you to measure various ranges of DC (direct current) volts, AC volts, current, and resistance. More advanced meters also measure things like frequency and duty cycle or even temperature. Some meters will even directly measure capacitors, inductors, and transistors using specialized connectors.

As with most tools, you generally get what you pay for with multimeters. The meter on the left in Figure 2-21 was purchased for under $3, including shipping from China. It arrived DOA (dead on arrival) and was not deemed important enough to return. It will make an interesting clock or fake lie detector someday. The unit on the right was purchased new for nearly $400 and has provided over a decade of useful service. The unit in the center is modeled after the unit on the right, it would seem, but can be bought new for between $10 and $20. The unit in the photo was bought wholesale for $4, because sometimes you get lucky. It's often handy to have a second meter in the lab—for example, when you need to measure the voltage *and* the current flowing in a circuit simultaneously, which you just can't do with a single meter.

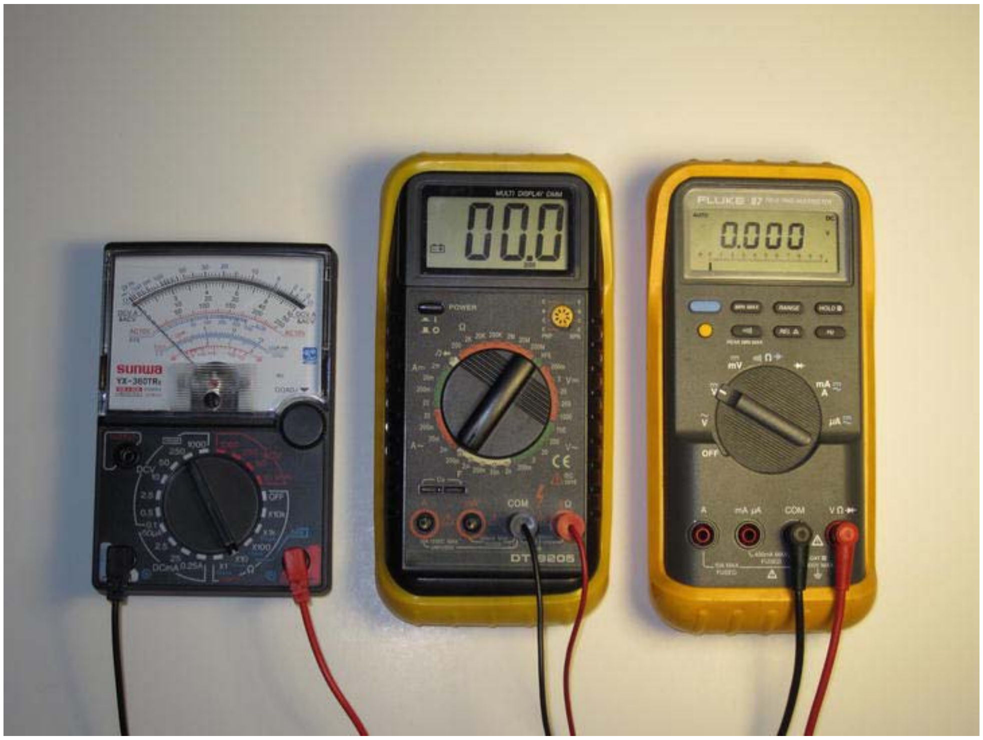

Figure 2-21. *Examples of some multimeters. The meter on the left features an analog display that is easy to read and acommodates several readout scales simultaneously. The remaining two meters have digital readouts and offer more accuracy. All the meters use a large rotary knob for selecting the proper function. The higher-cost unit on the right features auto-ranging, intellignetly deciding what range to use. The right two units also have high-impact boots to help protect them in the event of a fall.*

Your present budget will help determine what level of multimeter you might want to obtain. If you're wanting to sink less cash into a multimeter right now, consider looking for used equipment in good shape. As with any good tool, you can spend as much as you like on these. The sky really is the limit here.

Most all-in-one meters protect their inputs using fuses. You will have to disassemble your meter to replace these fuses, so think twice and measure once (to paraphrase the old carpenter's saying, "Measure twice and cut once").

■ **Tip** Be sure to set up your meter for the proper function and range *before* taking measurements.

Oscilloscopes

Once you've gone beyond measuring simple DC and AC voltages and move on to more complex signals and waveforms, you'll have to lay aside your trusty meter and fire up your favorite oscilloscope. This instrument is more complex than your traditional multimeter, and offers a different perspective on what's going on in your circuit.

While many modern oscilloscopes have many modes of operation, their basic function is to draw a picture of your signal on a small screen. In the olden days, this was done with a moving dot that was swept across a phosphorescent screen, leaving a residual, fading line of light. If this pattern was repeated over time, a visual representation of the signal could be seen on the screen. See Figure 2-22.

Modern oscilloscopes use a computer monitor or built-in screen display to draw the signal. This saves space, weight, and power requirements.

When purchasing an oscilloscope, the features you pay most for are the number of available channels and the bandwidth. The least expensive oscilloscopes will have a single channel. Having multiple channels allows you to "see" multiple signals simultaneously, which is often helpful in understanding what is going on in a circuit.

Bandwidth refers to the range of frequencies that can accurately be measured by the oscilloscope. Lower bandwidth scopes require much simpler electronics, and are thus cheaper to produce.

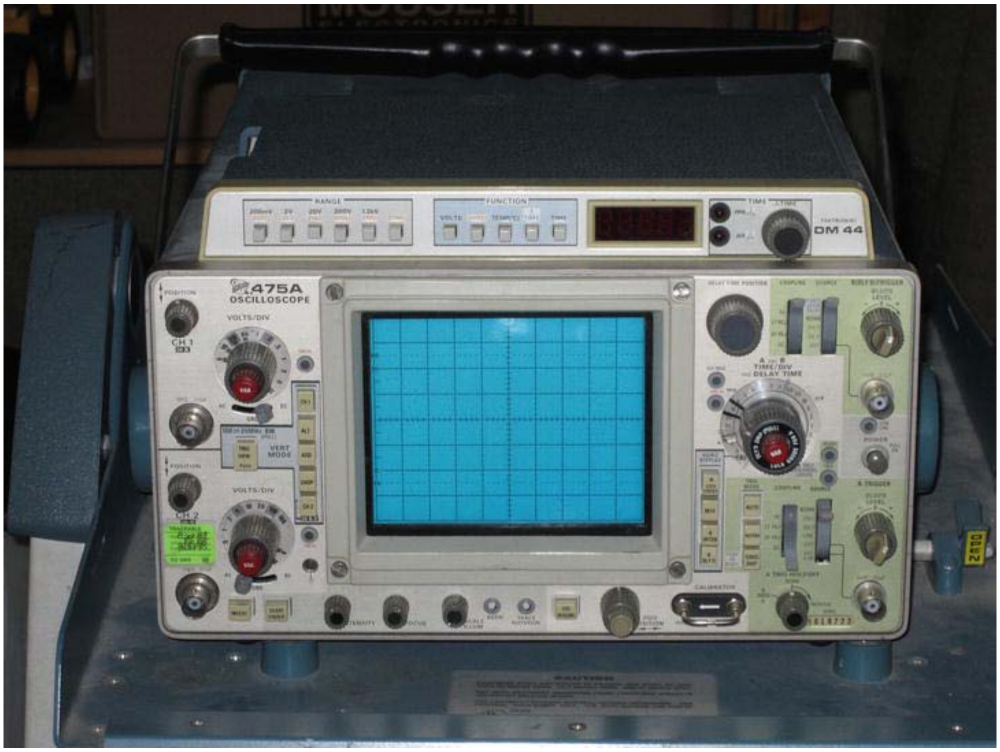

Figure 2-22. *A Tektronix 475A oscilloscope. This piece of vintage test equipment is considered by some to be the best analog oscilloscope ever made. It also has a digital multimeter built in along the top edge.*

The Golden Age

Once upon a time, not so long ago, a quality oscilloscope weighed 40 pounds and had a folding handle to make it, if not portable, at least *transportable*. It also had its own line of furniture to place it on, so that you could wheel it around the lab as you needed. Oh, and it also cost about three months' salary, and that's if you had a *good* job.

The good news? That same scope still works today. It can be repaired or calibrated to the exact same standards as when it was brand new, *plus* you can buy it on eBay for a few hundred dollars, sometimes less. That's because when these devices were designed back in the 1970s, they were made to be the best in the world, and cost was not a factor.

Modern Oscilloscopes

Today you can buy a quality oscilloscope for a few hundred to a few thousand dollars. These scopes not only display your signal on a large, colorful screen, but they can also analyze the signal and tell you all kinds of interesting things about it. For an example of such an oscilloscope, see Figure 2-23.

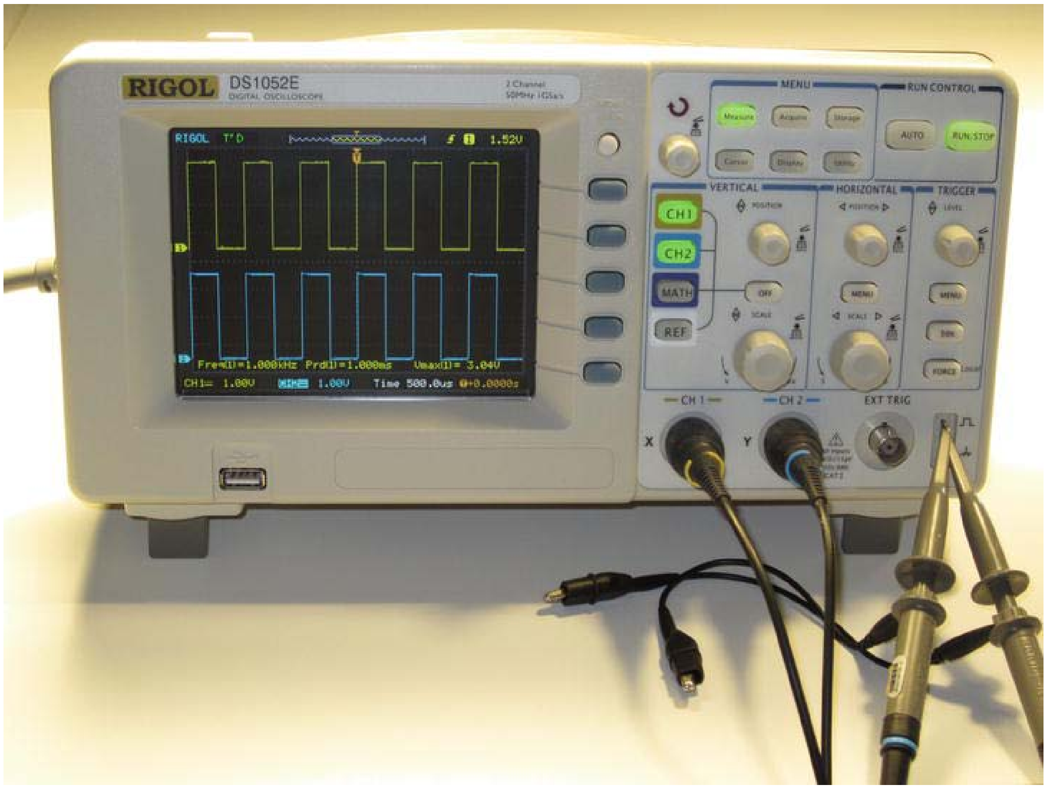
RIGOL
DS1052E
DIGITAL OSCILLOSCOPE
MENU
RUN CONTROL
AUTO
RUN/STOP
VERTICAL
HORIZONTAL
TRIGGER
CH1
CH2
MATH
REF
EXT TRIG
X
Y

Figure 2-23. *A modern digital storage oscilloscope. The RIGOL DS1052E is an affordable scope that will handle most troubleshooting and design challenges that you might encounter in your lab.*

This new breed of digital storage oscilloscope (DSO) can even connect to your PC to be remotely operated and allow you to capture screen shots that can then be pasted into books about how to build your own electronics lab. It’s true. See Figure 2-24.

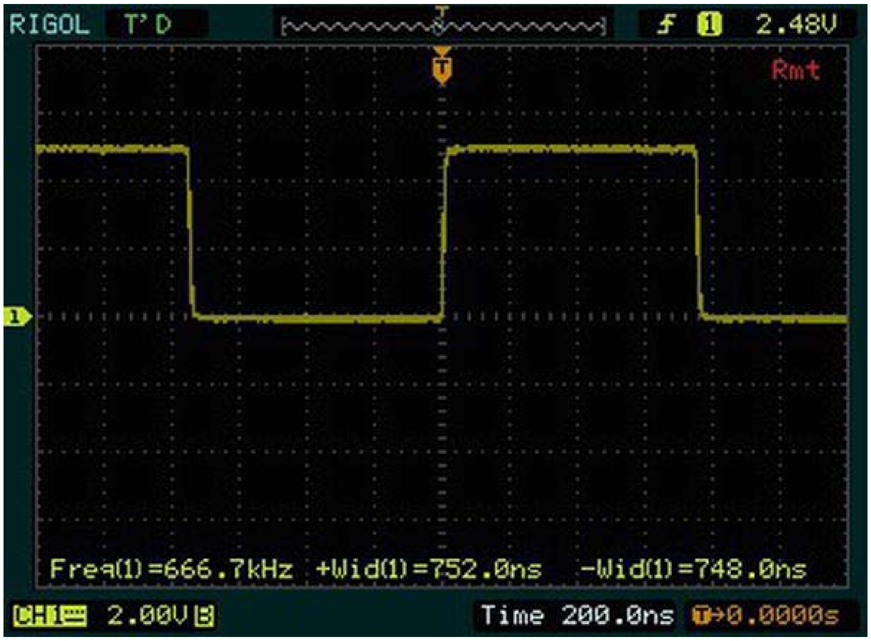

Figure 2-24. *A screen capture made with the RIGOL DS1052E oscilloscope.*

Tiny Oscilloscopes

The march of electronic miniaturization never stops, and so it is with oscilloscopes. If both your needs and your budget are modest enough, perhaps you can get by with something like the pocket-sized wonder shown in Figure 2-25.

While little scopes like this lack the accuracy and bandwidth of their larger relatives, they are definitely handy to have when you need to look at a signal. If you can keep your expectations in check when it comes to the user interface, you can get a lot of utility from such a small package.

Additionally, one of the benefits of a device such as this is that the hardware and internal software (yes, there's a little computer in there, somewhere) is all open source. This means that you have access to all the schematics and source code. And what does that mean? It means that if you have the skills and the right equipment, you can redesign the entire apparatus to fit your own requirements, assuming that you understand the basic limitations of the electronics. The sky is the limit. You are empowered.

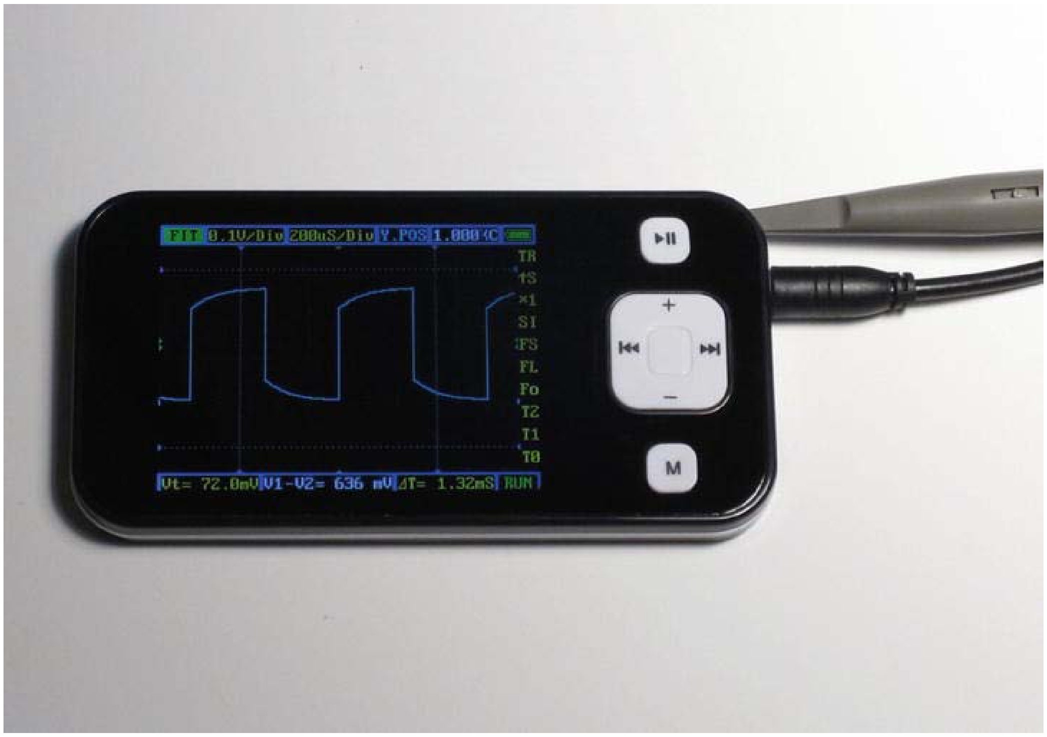
FIT 0.1V/Div 200uS/Div Y.POS 1.000
TR
↑S
×1
SI
FS
FL
Fo
T2
T1
T0
Vt= 72.0mV V1-V2= 636 mV ΔT= 1.32mS RUN
+
−
M

Figure 2-25. *A handheld oscilloscope, which can also fit in your pocket. This is the DSO201, which has been discontinued by the manufacturer, but can still be snapped up for a song from various online distributors, as well as eBay. This unit was purchased new for $60 and features a single channel with a bandwidth of 1 MHz.*

Power Supplies

One of the most important pieces of test equipment in your lab is a reliable power supply. Building a good bench power supply is an excellent first or second electronics project (see Chapter 5 for more on adding power supply to your lab and Appendix A for more on building your own). This is because the basic design is very simple and easy to build, yet allows you to add any number of additional features later.

Once you get tired of throwing away spent 9V batteries and grow weary of waiting for the red light on the battery charger to go out, it's time to invest (either time or money or both) in a decent bench power supply.

It doesn't have to be fancy. In fact, the simpler the better at first. What you're aiming for primarily at this point is reliability. You want to know that when you turn on your 5V supply, it's putting out 5 volts, not 2 or 11, no matter what you do to it. If it's overloaded, it needs to deal with it and not be catching on fire. If it's underloaded, or not loaded at all, it still needs to maintain a nice, clean output.

Want a cheap or free power supply? Do you have any dead or discarded electrical gizmos that came with a charger? These make excellent starter supplies for the new lab. Do you have any extra USB chargers? These are even better, because they supply up to 2.5W (and sometimes more) of *regulated* 5V. We'll use some of these in experiments in Chapter 3.

Organizing Your Tools

So now you have an idea about some of the tools you want (or already have) for your lab. Where are they all going to live? This is a big question, and deserves a big answer.

One way to look at this book is that the entire latter half is dedicated to answering this question. There are, obviously, many possible arrangements that will work for most people. But we're talking about *your* lab, with *your* tools, arranged and ordered to meet, specifically, *your* needs.

As you progress through Chapter 3 (the very next chapter in this book), you'll get a chance to play with both your tools and some interesting electronic components, all in the interest of getting you familiar with some basic concepts. These concepts and ideas are not new, but they may be new to you. The good news is that time invested today and in the very near future should pay off big dividends to you later in life. Who knows what you'll devise in that electronics lab of yours?

Summary

Starting with your brain and working outward, this chapter has covered some of the tools and devices you might use in your electronics lab. These include basic hand tools, computers, tools for soldering, some common power tools you might find handy, as well as some fancy-schmancy tools you might only find in a hackerspace, school, or electronics clubs, such as laser cutters and 3D printers.

You also learned about some of the basic test and measurement tools you'll be needing when working on electronic circuits in your lab. These include magnifying glasses, measuring tapes, and meters (oh my!), as well as oscilloscopes and power supplies.

Additionally, *lots* of advice on safety was presented. These topics were learned the hard way. Do try to learn from the mistakes of others, won't you?

Just a hint was given about how you might want to organize your tools within your lab. Take care of your tools and they will most certainly take care of you. The details about possible organizational scenarios is left for the second half of the book, where you will learn about setting up labs big and small, as befits your needs, budget, and interests.

Chapter 3 ties in much that you have learned here in Chapter 2 by introducing you to not only some of the basic electronics components and their uses and characteristics, but also how to use a good number of the tools described in this chapter. Hopefully this chapter has whetted your appetite and readied you to dive into some circuits and see what kind of mischief you can wreak on the world.

CHAPTER 3

Components

You're going to need some components to play with, as well as the knowledge to use them effectively. This chapter will introduce you to some of the bits and pieces that make up modern electric and electronic circuits. It will also teach you a little bit about how to identify components from their appearance and markings, when available.

Once you've got an idea about what these parts do in a circuit, you'll learn a little more about how to measure their electrical properties and put them to use. You'll also be shown what *not* to do, in some select examples.

There are an enormous number of both manufacturers and vendors of electronic components these days. Trying to list even a few of them here in print doesn't really make much sense, as a good percentage of that information will be incorrect or obsolete as soon as the ink dries on the page.

Having a variety of basic components readily available is convenient, until your collection grows beyond your ability to house or organize it. This happens far too often, in reality. Most common electronic parts can be obtained within a day or two from local sources, unless you live in a really secluded part of the planet. Also, most of the more common parts will typically cost only a few pennies apiece, so there aren't a lot of reasons to stockpile random parts. Keep a few of each on hand for experimentation, but hold off on pointless hoarding . . . if you can.

Conductors and Interconnects

In any electrical or electronic circuit, there are just as many connections between components as there are components. These include wires, cables, connectors, plugs, sockets, and all those other pieces that help the little electrons zip around the circuit.

First, let's take a look at wires and interconnects in a very abstract way. We can think of wires as connections between other components, without having to get into a lot of detail about the wire itself. We just want a way to indicate that there is an electrical connection among all the bits, and using an electrical schematic diagram is an easy and effective way to do this.

Electrical Schematics

A *schematic* diagram is any kind of stylized drawing that represents a simplified view of something. Usually some sorts of symbols are used to represent the parts that make up the whole, as well as the connections or relationships that exist between them and possibly the outside world. You can have schematics for plumbing, hiking trails, or tribbles, a well as electrical and electronic circuits. Let's talk about electronic schematics for a little bit.

The electrical symbol for a wire is almost always just a plain, solid line. If there's something really special about the wire that needs to be used, it is usually indicated in a note somewhere on the schematic. This can be used to indicate what color the wire needs to be, what size is appropriate, whether there are any special requirements for the insulation, and so forth. Most of the time it's just a plain, solid line. For an example of just how simple it can be to draw a wire in a schematic, see Figure 3-1.

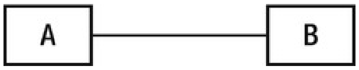

Figure 3-1. *A very simple electrical schematic diagram, showing how to get from point A to point B with a single wire. Drawing wires in schematics is easy!*

If a wire starts in just one place and ends in just one place, everything stays nice and easy. If multiple wires need to converge on a single point, or one signal wire needs to split up to go to more than one destination, it gets just a little more complicated, but not too much. See Figure 3-2.

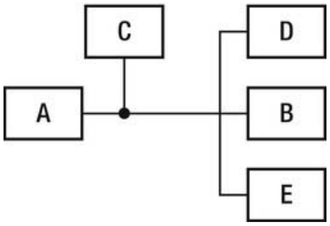

Figure 3-2. *Wires that are connected to each other usually have a dot at their intersection. If there is no dot, there is no connection between overlapping wires.*

In the figure, points A, B, and C are connected together. Points D and E are connected together, but are not connected to points A, B or C, even though the wires look like they are crossing. In a professional schematic, a careful draftsperson would route the wire from point D to point E on the right side of the components to avoid any possible confusion.

Some folks like to draw a little bump, or bridge, when two unconnected wires cross in a schematic, just to make sure that everyone knows that the wires aren't connected. This gets to be a bit tedious once you've got more than two or three components in your schematic, and it can look a bit unsightly. Feel free, however, to draw your schematics however you see fit. Just remember that others might not share your particular drafting sensibilities.

Wire and Cables

Chapter 2 took a quick look at wire and how it is made to help you pick out the best wire-cutting and wire-stripping tools. Now it is time to consider wire as an electrical component within your circuits. Knowing what a wire can and cannot do will help you pick the right kind of wire to use. Yes, wire has limits! This is important both for building new circuits from scratch as well as repairing, replacing, or upgrading wiring in existing circuits.

Wire can be used for all sorts of artsy and craftsy purposes. Copper wire is the most common wire used in electronics. Sometimes the copper conductor is *tinned* with another conductive metal with a low melting point to make it easier to solder. When you cut this wire, you can see the copper on the inside.

Pure copper wire is getting very expensive these days. Unscrupulous people try to pass off copper alloy with inferior electrical properties as the real deal. A quick way to test is with a magnet. Copper is not attracted to magnets.

As mentioned in Chapter 2, wire comes in two major categories: solid core and stranded. Solid core wire is just a single piece of copper wire, while stranded wire is made up of many smaller strands of copper wire, usually twisted together. Stranded wire is generally more flexible than solid core wire.

Copper wire is measured using a gauge system. Different systems exist around the world. A common system used in the United States and Canada is the American Wire Gauge (AWG) standard. It uses a series of single numbers to denote the relative size of the wire, with larger numbers representing smaller-diameter wires. These numbers go all the way down to zero (0 gauge), but they don't stop there. Larger wires than 0 gauge are called 00 gauge, double-ought, or sometimes 2/0. There are also three- and four-digit versions, but after that, those giant wires are specified with their actual cross-sectional diameter.

Using Wire: The Very Short Version

Fatter wires carry more current with less heat being generated due to resistance. The wire's insulation is rated for a maximum voltage and temperature. If this temperature is exceeded, the insulation will soften, melt, or catch on fire. Avoid this by using the right sized wire for the job. Exceeding the maximum voltage can cause more problems if the current arcs through the insulation. This can cause a short circuit and may even cause a fire. It is also likely to give off noxious fumes. Choose wisely!

For low voltages (under 36V) and small current loads (under 1A), you can use whatever you want. It doesn't matter. Pick something pretty. A thing of beauty is a joy forever.

For higher-voltage applications, make sure the wire's insulation is rated *far in excess* of what you anticipate. The rating should be visibly printed on the insulation itself. See Figure 3-3.

Figure 3-3. *The wire's insulation only works up to a point. That point is usually printed on the insulation itself, but you may need to put on your good reading glasses to see it. This wire states, "LL33911 CSA TEW 105°C 600V FTT 14AWG," which translates to the manufacturer's part number, Canadian Standards Association certified thermoplastic equipment wire, maximum temperature of 105°C, 600V insulation dialectric strength, 14 gauge.*

For higher-current applications, you really need to use the highest-gauge wire that is necessary to carry the required current. Hot wires cause fires.

Using Wire: The Longer Version

If you will spend some time learning about the properties of wire, you will be able to make *much* better decisions regarding its use and application. Assuming that the overall shape of the cross section of the wire is round, the gauge of the wire determines how much resistance to the flow of electrical current the wire will exhibit over a given distance. This resistance, in turn, helps determine the amount of electrical current that the wire can safely carry.

When electric current meets resistance, the electrical power is converted into some other form of energy. Usually this energy is given off as heat, or thermal energy. It can also produce magnetic fields, sound, light, and radio waves.

A good example of this is demonstrated by the heater coils on an electric stovetop. When the stove's switch is turned on, current flows through the coil. Most of the current is converted directly into heat. Electric heaters are very efficient this way. A small amount of the electrical power is converted into light. You can see this as a red glow coming from the heater coil when it is hot. The coil also creates a magnetic field.

If you know the gauge of a wire, you can look up the wire's characteristic resistance. These numbers will be close but perhaps not exact, as there will always be small variations in wire diameter, conductor purity, and other factors. Generally, however, the numbers are pretty close.

Let's take another look at the AWG table (see Table 3-1), and look at a few simple examples to help understand what these wires can do for us. (A more complete listing is available at `http://en.wikipedia.org/wiki/American wire gauge`, from which this partial list was obtained.)

Table 3-1. *The AWG*

AWG	Diameter Area			Copper Resistance	
	Inch mm		mm^2	Ω/km mΩ/m	Ω/1000 ft mΩ/ft
0000 (4/0)	0.46 00	11.6 84	107	0.1608	0.04901
000 (3/0)	0.4096	10.404	85.0 0.20	28	0.06180
00 (2/0)	0.3648	9.266	67.4 0.25	57	0.07793
0 (1/0)	0.3249	8.252	53.5 0.32	24	0.09827
1 0.28	93	7.348	42.4 0.40	66	0.1239
2 0.25	76	6.544	33.6 0.51	27	0.1563
3 0.22	94	5.827	26.7 0.64	65	0.1970
4 0.20	43	5.189	21.2 0.81	52	0.2485
5 0.18	19	4.621	16.8 1.02	8	0.3133
6 0.16	20	4.115	13.3 1.29	6	0.3951
7 0.14	43	3.665	10.5 1.63	4	0.4982

8 0.12	85	3.264	8.37 2.	06 1	0.6282
10 0.10	19	2.588	5.26 3.	27 7	0.9989
12 0.08	08	2.053	3.31 5.	21 1	1.588
14 0.06	41	1.628	2.08 8.	28 6	2.525
16 0.05	08	1.291	1.31 13	.1 7	4.016
18 0.04	03	1.024	0.823 2	0.9 5	6.385
20 0.03	20	0.812	0.518 3	3.3 1	10.15
22 0.02	53	0.644	0.326 5	2.9 6	16.14
24 0.02	01	0.511	0.205 8	4.2 2	25.67
26 0.01	59	0.405	0.129	133.9	40.81
28	0.0126 0	.32 1 0	08 10	212.9	64.90
30	0.0100 0	.25 5 0	05 09	338.6	103.2

Conductivity

The first thing you need to know is that copper wire really, really wants to conduct electricity. It wants to conduct electricity more than it wants to *resist* the flow of electrical current. Even so, copper is not a perfect conductor, and it has a certain amount of inherent resistance. This resistance causes the energy carried by the electrical current to be sidetracked, or converted into heat and other forms of energy. The higher the resistance in the wire, the more heat will be produced.

Normally this heat is a very small thing and is easily dissipated into the surrounding air. You usually can't even tell by touching that a wire is a fraction of a degree warmer than its surroundings—but this is only true when the amount of energy is relatively small, like in our rewired flashlight example. When you start to push a *lot* of power through a wire, the heat buildup can outpace the cooling-off provided by the surrounding air, and the temperature of the wire increases.

Again, this is no big deal as long as that temperature stays within a comfortable range. But what happens when the temperature gets really high?

First, the insulation will get soft, melt, or even catch on fire. After the insulation, if any, has burned off, the wire will glow red hot and eventually melt. This is known as *catastrophic wire failure*, which can result in what electrical engineers refer to as an unauthorized thermal event, which we call a fire. Yikes!

But look around you. You're probably sitting right next to some wires right now, very few of which (hopefully) are on fire, even though they are carrying variously large loads. How do they keep from burning up? The answer is in the proper selection of the right *size* of wire for the job. Bigger wires have less resistance to the flow of electricity, so they are able to carry more power with less heat being produced.

Now, you don't want to have to wire up all of your electronic circuits using jumper cables from your car's trunk, as that gets to be cumbersome after a while. Let's look at some good guidelines for what size wire to use for your projects.

In Chapter 2 you learned about (or were reminded of) Ohm's Law, which illustrates the relationships between voltage, current, and resistance in any electrical circuit. In summary, V = I × R, where V is volts, I is current in amps, and R is resistance in ohms. It's not just a good idea, it's the law!

So how much electricity will we lose going through a foot or so of 16-gauge wire? If we know that one foot of 16-gauge wire has a resistance of 4mΩ (and we know this from looking it up in the chart), and we build a circuit that pushes 1A of current through this wire, we can figure out a couple of interesting things using Ohm's Law and a calculator.

Applying Ohm's Law

If you know two of the three variables used in Ohm's Law, you can calculate the third variable. Since we know R and I, we can calculate V. Since V = I × R, we can substitute our values of 1A and 4mΩ (0.004Ω) into the equation, and get 0.004V, or 4 millivolts.

That's not very many volts at all! In this case, that's a good thing. The voltage we just calculated is the *voltage drop* across the length of the wire. Those millivolts never make it from one end of the wire to the other—they're lost. Poor millivolts, they never had a chance.

OK, let's not let their heroic sacrifice be in vain. If we know volts, and we know amps, we can also calculate power, which is the rate at which energy is converted to work. The unit of measure for electrical power is *watts*, and is also the same as Joules per second (J/s), but that's another story. We usually use the capital letter *W* to abbreviate wattage, but we use the capital letter *P* to represent the *value* in equations.

Calculating power is simple if you know the volts and the amps. Just multiply them together: P = V × I. Easy.

So in our example so far, we have 4mV, 1A, and 4mΩ. How much power is being dissipated (wasted) in this circuit? P = 0.004V × 1A = 0.004W, or 4 milliwatts, which is just four-thousandths of a watt. That's just about enough power to light up a dim LED, and not much else. So you're not going to be starting any fires with these numbers. You most likely would not even be able to measure the rise in temperature that is occurring in the wire (oh yes, it's occurring, but it's very, very small).

Let's put 10A through a foot of 30-gauge wire. On second thought, let's not, and just do the math instead. The voltage drop will be V = I × R, or 10A × 0.1032Ω = 1.032V. The power dissipated would be P = V × I, or 1.032V × 10A = 10.32W. That wire would be a crispy critter in no time. Ten watts doesn't sound like much when you're talking about a home stereo or a lightbulb or a hair dryer, but that skinny little wire has almost no surface area with which to dissipate the generated heat. This means that the temperature goes up, quickly. Bye-bye, little piece of wire.

As a side note, you just learned how most current-limiting fuses operate. The conductor within the fuse literally *fuses*, or evaporates, when the internally generated heat produced from an "overcurrent" situation rises beyond its melting point. Poof, the circuit is opened, and property and lives are saved.

■ **Warning** Never circumvent the protection afforded by a fuse.

Ideally what you want is the smallest wire that will carry the required amount of current with an acceptable rise in temperature. Remember that any rise in temperature is in addition to the *ambient* temperature. If you're sitting in a room that is "room temperature," which is assumed to be 25°C or 77°F (the same thing, if you want to check it), and you're expecting a 50°C rise in temperature due to waste heat, your wire is going to be 75°C, which is far too hot to touch. Remember that 100°C is the boiling point of (pure) water (at sea level). Hot!

One more thing that complicates the thermal calculations is that copper, as a conductor, has a *positive thermal coefficient*, meaning that the characteristic resistance of a piece of wire (which we can look up in the AWG table) goes *up* by a known amount as the temperature increases. You need to consider this when you build circuits that are *supposed* to get hot, such as electric heaters. In most other cases, it's best to keep the heat to a minimum, as it represents wasted power, plain and simple.

Testing Your Wires

Can something as simple as a wire go bad? Yes, it most certainly can. How can you tell? Usually, it's because something doesn't work the way you think it should, or *worse*, it works sometimes and not others. A great deal of time, the problem is right at the end of the wire, where it is trying to make a connection with something else. However, it's possible for a wire to develop a problem right in the middle. This can be caused by a tiny nick in the insulation that lets in nasty elements such as rain and dirt, allowing the inner conductor to corrode. This type of problem takes time to make itself apparent, but it can certainly happen, especially in equipment that is exposed to the outside world for any length of time.

A quick way to test a wire is with a piece of test equipment called a *continuity tester*. This simple device usually is battery powered and has two probes. When you connect the probes together, the tester makes a noise, lights up a light, or maybe does both. Make sure your continuity tester's batteries are fresh and that the probes, if they are removable, are firmly seated in their sockets. Even test equipment needs testing from time to time.

■ **Note** A continuity tester is one of the easiest pieces of test equipment to cobble together yourself. See Appendix A for more information about building your own tools.

As mentioned previously, the problem is almost always at the very end of the wire. This is a good thing to know, so you can start looking in the most likely places first, saving valuable troubleshooting time.

A much better way to test a wire is with an ohmmeter. Using an ohmmeter, you can determine the *exact* amount of resistance in a wire, which under ideal circumstances should be very low. One ohm or less is typical for most normal-sized wires that you are likely to encounter. Super-long wires can have higher resistance, and that's where you can consult the handy AWG table to find out what the resistance is *supposed* to be.

When testing wires, especially wires with attached connectors, it's a good idea to wiggle the wire about vigorously, as this will reveal some intermittent connections for you.

Storing Your Wire

If you're buying new wire, it will probably come on a reel. It's often handy to keep common sizes and colors on your bench, if you have room. A paper towel–holder makes an excellent wire dispenser rack, although you can also spend more time or money buying a prebuilt one or designing and fabricating your own.

A custom red-and-black wire dispenser is shown in Figure 3-4. You might use a fixture like this if many pairs of identical-length wires are needed for a project. Not having to chase the wire spools all over the bench really saves a lot of time. This structure was built using a MicroRax miniature slotted aluminum extrusion (*http://www.microrax.com*), which is perhaps a bit of overkill, but was a lot of fun to plan and build. A coat hanger, artfully bent, would have also done the trick.

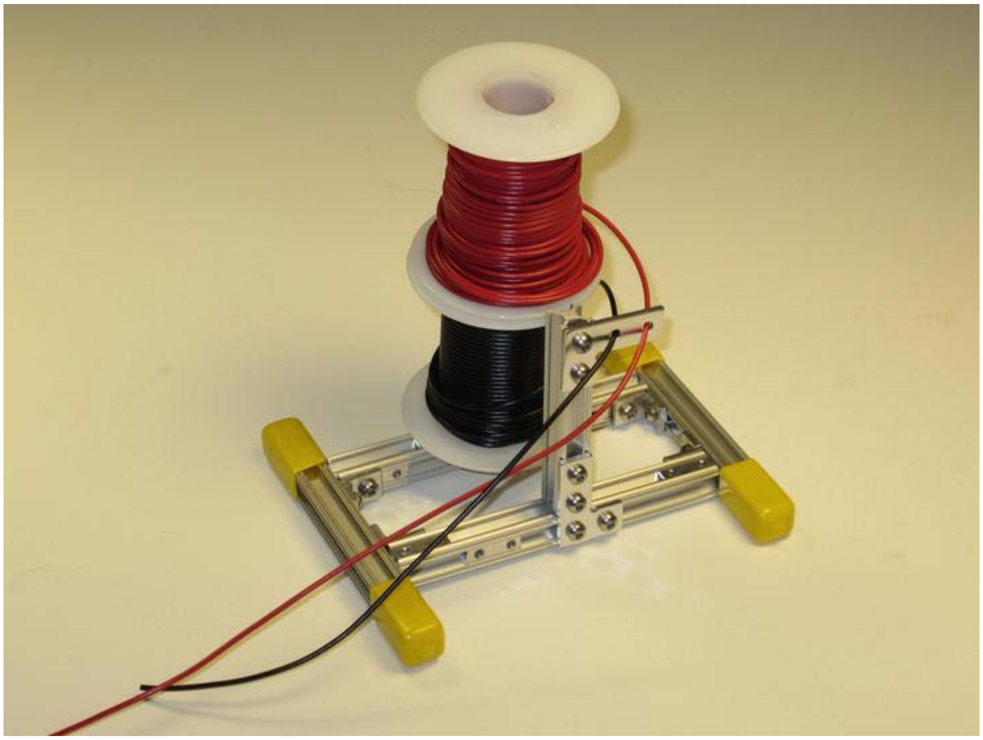

Figure 3-4. *A custom wire-dispensing frame made from modular aluminum extrusion. This kind of fixture makes it easier to cut both a red and a black wire to almost exactly the same length, which happens a lot when you use red and black wire for positive and negative power connections, respectively.*

Connectors

Connectors make life easier for the casual wiring enthusiast. The hardcore types solder everything together, but that is not always practical or even the right thing to do in every circumstance.

Keep in mind that every connector, just like every wire, has some sort of electrical limits. These limits must be respected.

Clip Leads

Also called *alligator clips* or *test leads*, you can buy ready-made wires or build your own from individual connectors. The flexible rubber or plastic boot extends the insulation of the wire. A spring-loaded jaw, often with serrated teeth, is used to clamp down on the wire, lead, or electrical terminal desired (see Figure 3-5).

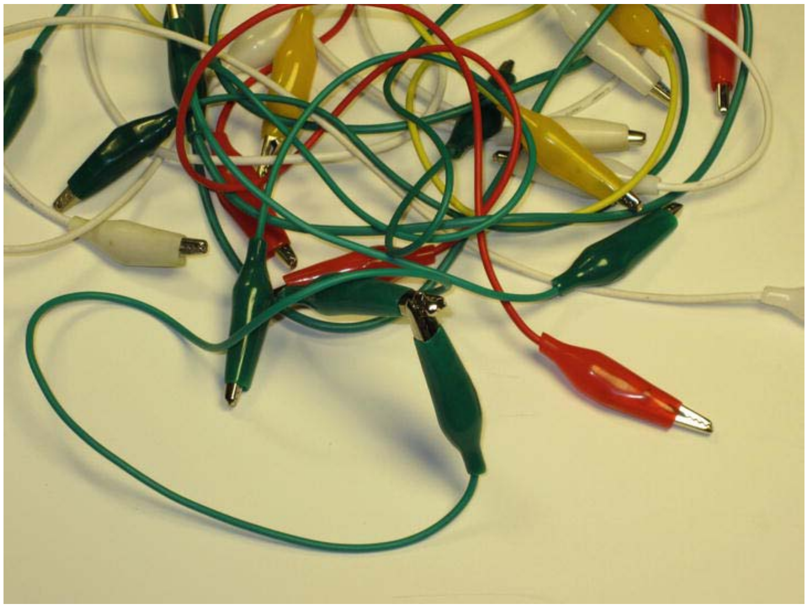

Figure 3-5. *Clip leads come in all sizes, gauges, colors, and terminal types. They are handy for making quick electrical connections.*

Clip leads are quite handy to have around your bench, if you've got the room for them. They tend to get tangled up, so some people hang them from a specially slotted rack, making them easier to pluck at a moment's notice.

Keep in mind that clip leads are generally for *temporary* electrical connections, such as when you're testing something or trying out a prototype circuit. They don't make very good long-term solutions to your electrical connection needs for a variety of reasons.

The first problem is that even when the spring-loaded jaws make a solid, low-impedance connection when first attached, this electrical connection will deteriorate over time as oxidation inevitably forms on the contacts.

Another problem is that many mass-produced clip leads are inexpertly smashed onto the ends of wires, without taking advantage of the alligator clips' built-in features, such as strain reliefs and soldering points. This leads to additional resistance in your circuit at both ends of the clip lead, where you would probably never think to look. The next time you buy some brand-new clip leads, hook them up to your ohmmeter and see just how conductive they really are. A good practice is to carefully remove the insulating boot and make a proper electrical connection yourself, so that you've got one less thing to worry about.

Solderless Breadboards

A very versatile prototyping platform is the solderless breadboard. Once upon a time, people would tack together simple electrical circuits using wires, small nails, and a hammer, utilizing a wooden plank or small board as the substrate. The kitchen breadboard was often just right in size and weight and would mysteriously disappear into the workshop. You can still cobble together simple circuits using this technique.

The solderless breadboard is a miracle of modern manufacturing techniques. A body made of precision-molded plastic has an array of small openings built in, and within each opening is a row of small spring clips that are just the right size to grasp and hold a small-gauge (22–26 AWG) wire. Adjacent points within the array are electrically connected together, which lets you easily connect small components into a circuit using your bare hands, with no soldering required. About all you would normally need would be some solid-core wire (preferably), some wire cutters, and some wire strippers. You can even buy precut jumper wires specifically made for use with solderless breadboards. They're terrifically handy! See Figure 3-6.

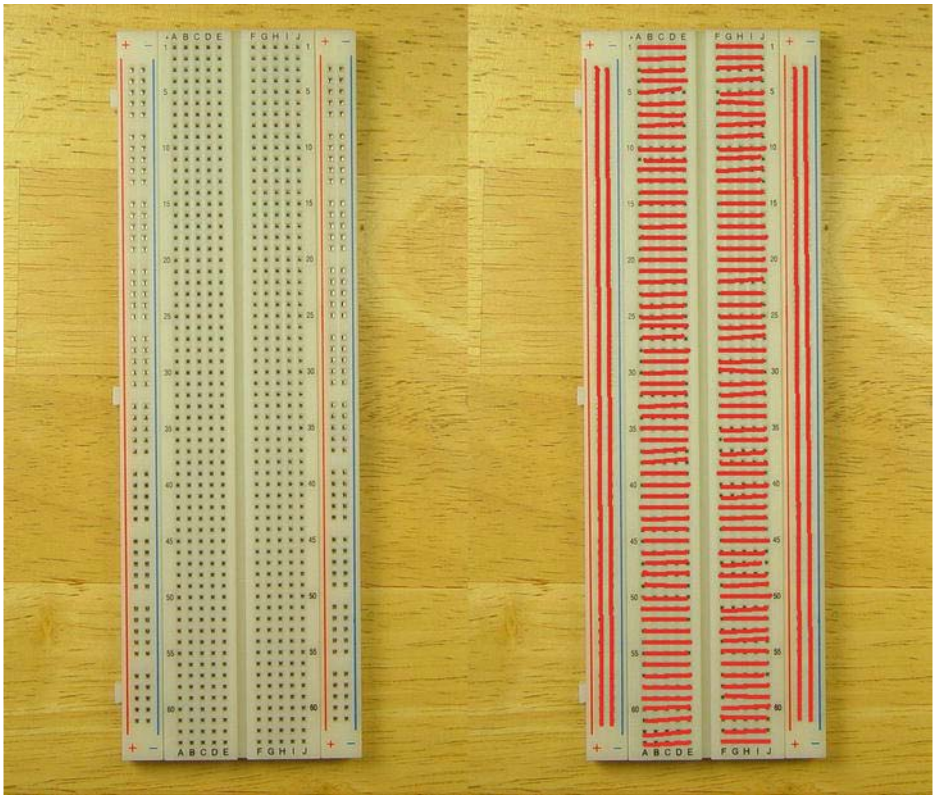

Figure 3-6. *On the left is a solderless breadboard with 830 tie points. Each hole has a spring clip inside that can hold a small wire or component lead. On the right, the underlying electrical connections between the tie points are illustrated.*

The spacing between adjacent tie points is exactly 0.1", which corresponds to many common "footprints" for electronic components. You can easily build simple circuits in just a few minutes. You can even build fairly complex circuits using integrated circuits and other components. Figure 3-7 shows a complete Arduino-compatible microcomputer that you can build with your bare hands and then program using your PC and the free Arduino software.

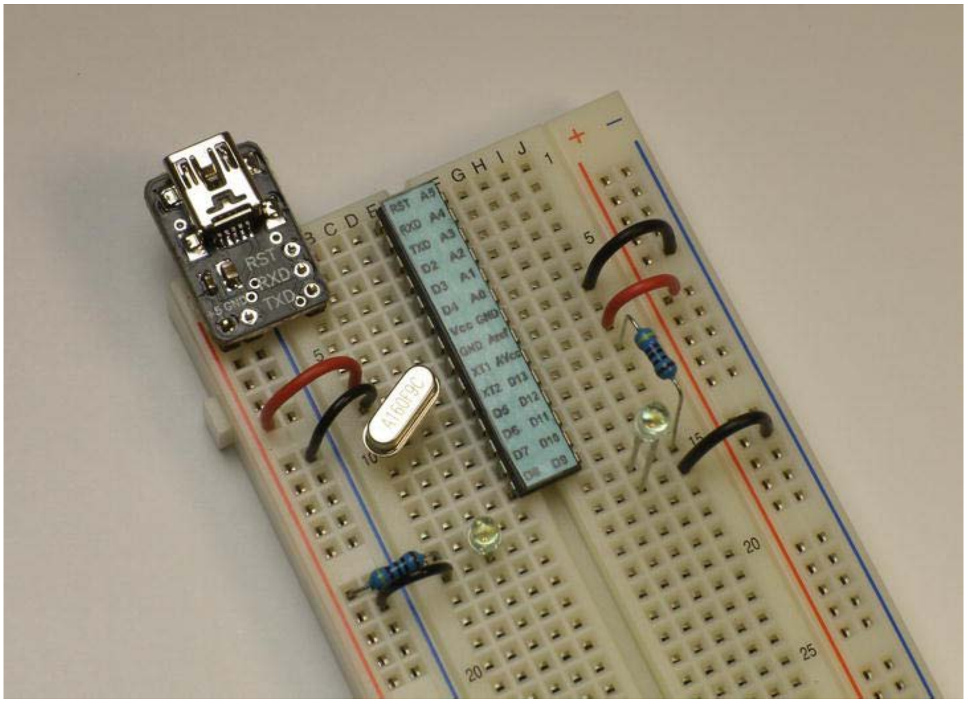
RST
RXD
TXD
GND
RST A5
RXD A4
TXD A3
D2 A2
D3 A1
D4 A0
Vcc GND
GND Aref
XT1 AVcc
XT2 D13
D5 D12
D6 D11
D7 D10
D8 D9
B C D E
G H I J
+
−
1
5
10
15
20
25

Figure 3-7. *A breadboard Arduino, which is compatible with the popular Arduino Duemilanove microcontroller, can be built using just a handful of components on a solderless breadboard, leaving plenty of room for more experimentation. The black, rectangular component in the upper-left corner is a miniature USB adapter that supplies power, and provides communication and control signals for the microcontroller.*

You'll find that projects tend to expand to fill the available area (and budget). See Figure 3-8.

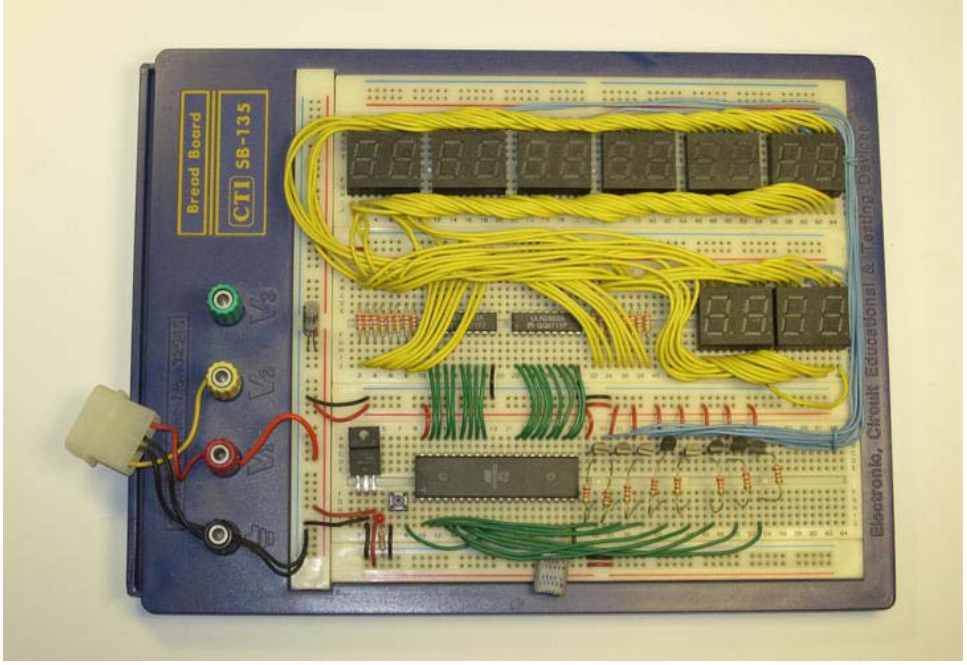

Figure 3-8. *A prototype LED display with computer control, built on a solderless breadboard. That's a lot of wires!*

Passive Components

Most electronic components fall into one of two different categories: *passive* and *active*. There are a few different ideas floating around as to what, exactly, is the difference between passive and active components. For the purposes of this book, we will use the simple definition that an *active* component requires a power source to work properly, while a *passive* component does not.

Resistors

A good example of a passive component is a *resistor*. You've already learned a lot about resistance in an electronic circuit. Resistance is measured in ohms, and it converts electricity into heat. When you're working with wires, cables, and other conductors, resistance is usually a *bad thing*. You want the least amount of resistance possible. However, there are times when you both want and need resistance in a circuit.

One example of when you would want resistance in a circuit is when you want to light up an LED. Unlike an incandescent lightbulb, an LED has no built-in current-limiting ability. Without a fixed amount of external resistance added to the circuit, an LED would self-destruct in a very short time, because it would overheat from excessive current flowing through it. A resistor would limit the amount of current flowing and control the brightness of the LED.

The exception to this dire prediction is when the power source that is lighting up the LED has insufficient gumption to blow up anything, such as with a small coin-cell battery. Anything bigger than that normally has no trouble convincing the LED to release its magic smoke, without which it can no longer function.

A typical resistor has a fixed resistance and a maximum power rating. The resistance is determined by the material from which the resistor is made. The value of a resistor, as you already know, is measured in ohms, and can range from just a fraction of an ohm (often expressed as milliohms) up to millions of ohms.

Through-Hole and Surface-Mount Resistors

Resistors come in all shapes and sizes. Generally speaking, the larger the surface area of the resistor, the more power it can safely dissipate. A typical resistor that you might work with on a solderless breadboard or on an older PCB is shown in Figure 3-9. These are called *through-hole* components because their extended leads are meant to be formed and threaded through holes that are drilled into PCBs. Also in Figure 3-9, for comparison, are some teensy-weensy *surface-mount* resistors. Surface-mount resistors have no leads and are meant to be soldered directly to pads laid out on the surface of the PCB.

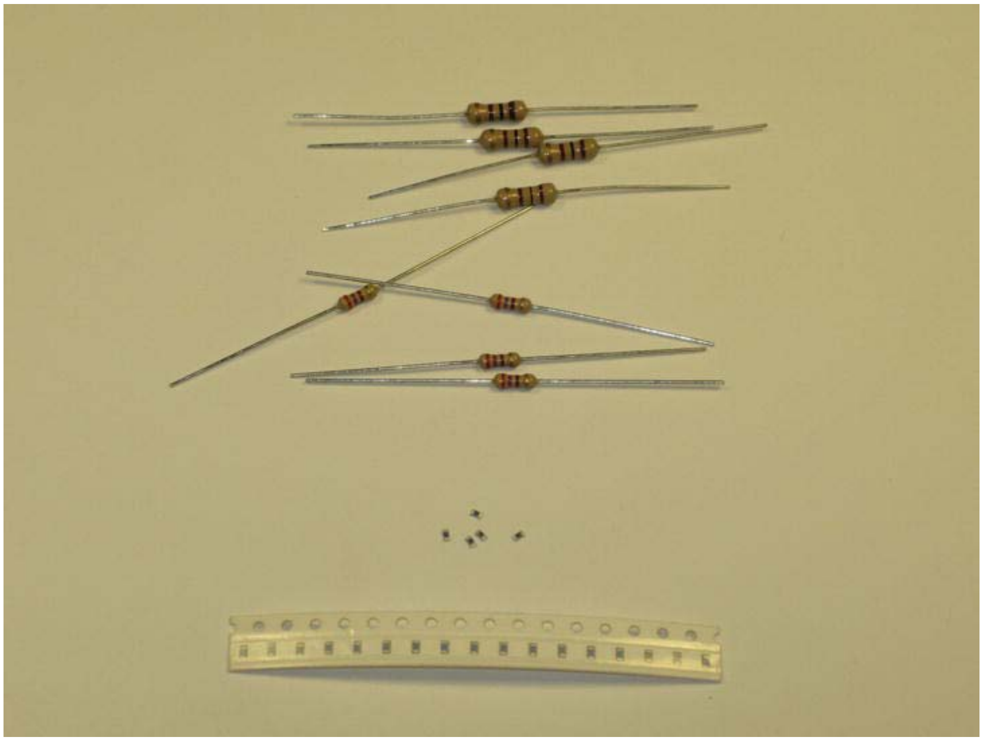

Figure 3-9. *Some through-hole resistors (top) and surface mount resistors (bottom). The surface mount resistors are so small that they are shipped packaged in tape, which is usually wound on reels for automated placement in PCB assembly.*

You can easily form the leads of through-hole resistors by hand or by using a *lead-bending jig*, a small tool with accurately spaced notches. Traditionally, these resistors have their leads formed so as to be threaded through holes in a PCB, and then soldered to the PCB. The extra lead length on the bottom of the PCB is then clipped off. See Figure 3-10.

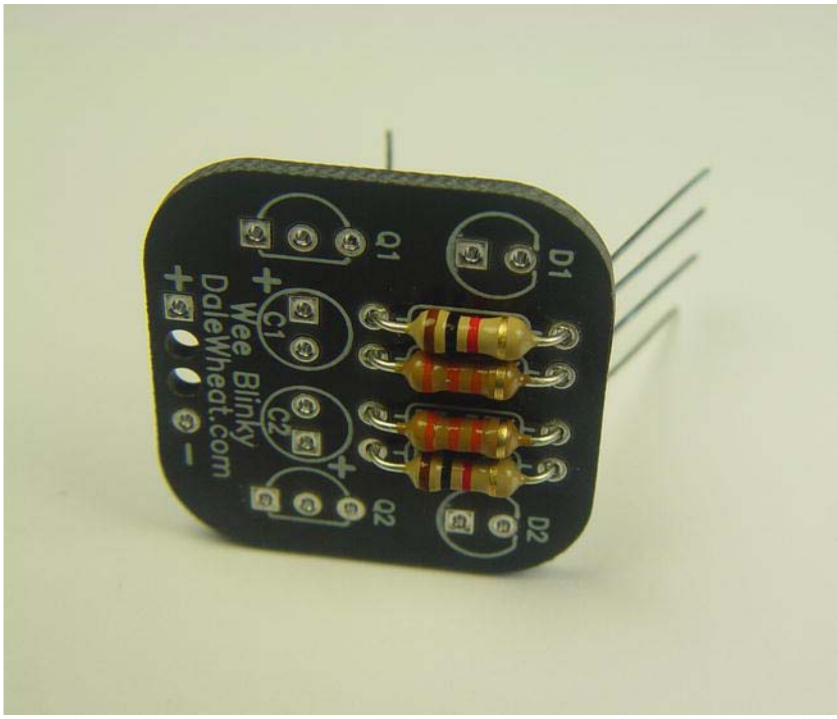

Figure 3-10. *Resistor leads are bent at right angles and inserted into holes in the PCB to be soldered.*

Schematic Symbols and Reference Designators

The schematic symbol for a fixed-value resistor is a zigzag line. See Figure 3-11. Every component in your circuit should have a name, which is also called a *reference designator*. Typically, most reference designators will have a short letter abbreviation to describe what kind of part it is (*R* is for *resistor*) and a sequential number assigned as more parts are added to the schematic. R1 would be the first resistor, R2 would be the second resistor, and so on.

If there is only one resistor in your circuit, then you don't have to number them, do you? You can just called it *R*, or *Galactic Emperor R the Mighty*, or whatever strikes your fancy. Most people stick to simple numbers.

Your numbering scheme depends a lot on the complexity of your schematic. If you only have a few resistors in your circuit, you can start at one and go up as needed. If you have a complex circuit with many subcircuits, you might allocate R101, R102, R103, and so on, to the first subcircuit (e.g., a power supply), and then assigned R201, R202, and so on, to a different section, such as a display. It's entirely up to you. Use a system that helps *you* keep track of things.

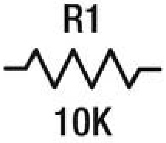

Figure 3-11. *The schematic symbol for a fixed-value resistor is a zigzag. The reference designator for this resistor is R1 and its value is 10,000 ohms, or 10KΩ. A simple rectangle symbol is also used to represent resistors, although many components are available in this physical form factor. Use the symbol you like best.*

Standard Resistor Values and Tolerances

Remember how wires have their own peculiar numbering scheme to describe their basic characteristics (i.e., the AWG, which uses *smaller* numbers to represent *larger* wires)? Well, resistors have their own peculiar numbering system as well to describe their electrical properties, or *value*. Even though the *value* of a resistor is measured in ohms and succinctly describes its characteristic resistance, not every possible value is readily available.

Some smart cookies at the Electronics Industries Alliance (EIA; see `http://www.ecaus.org/eia/site/index.html#`) have come up with some standard values that fill in most of the commonly used values for resistors and other components. Their scale is not linear but logarithmic, meaning that it takes larger steps as the values go up. This makes a little more sense once you understand how component value *tolerances* work.

You can buy, for example, a 1Ω resistor and expect it to be pretty close to being 1Ω at room temperature, but it's not going to be *exactly* 1Ω. It's going to be pretty close. How close? Well, that mostly depends on how much money you're willing to spend on this single resistor. If the answer is "not that much," then "pretty close" is going to mean within 5 percent of the stated value. That means your 1Ω resistor could actually measure as little as 0.95Ω or as much as 1.05Ω and still be "within tolerance." Five-percent resistors cost only a few cents even in small quantities.

Given that there's a little wiggle room for each resistor value, people who can tolerate their resistor values being off by as much as 5 percent don't need to be ordering 50 resistors of 0.99Ω and another 50 of 1.01Ω when they could just as easily (and certainly less expensively) order 100 resistors of 1Ω with a 5 percent tolerance.

This overlap in the values of the resistors allows us to skip from 1.0Ω to 1.1Ω without having to keep track of all the little resistors in between. This is because a 1.1Ω resistor could measure as little as 1.05Ω or as much as 1.15Ω and still be in tolerance.

With a standard resistor tolerance of 5 percent, we can use just 24 values to cover the range between 1Ω and 10Ω. The same multiples can then be used to cover the range between 10Ω and 100Ω, and so on. Aren't logarithms fun? See Table 3-2.

Table 3-2. *The E24 Standard Values for 5 Percent Resistor Value Tolerances (Higher and Lower Decades Are Also Available)*

1.0 10		100	1.0K	10K	100K	1.0M
1.1 11		110	1.1K	11K	110K	1.1M
1.2 12		120	1.2K	12K	120K	1.2M
1.3 13		130	1.3K	13K	130K	1.3M
1.5 15		150	1.5K	15K	150K	1.5M

1.6	16	160	1.6K	16K	160K	1.6M
1.8	18	180	1.8K	18K	180K	1.8M
2.0	20	200	2.0K	20K	200K	2.0M
2.2	22	220	2.2K	22K	220K	2.2M
2.4	24	240	2.4K	24K	240K	2.4M
2.7	27	270	2.7K	27K	270K	2.7M
3.0	30	300	3.0K	30K	300K	3.0M
3.3	33	330	3.3K	33K	330K	3.3M
3.6	36	360	3.6K	36K	360K	3.6M
3.9	39	390	3.9K	39K	390K	3.9M
4.3	43	430	4.3K	43K	430K	4.3M
4.7	47	470	4.7K	47K	470K	4.7M
5.1	51	510	5.1K	51K	510K	5.1M
5.6	56	560	5.6K	56K	560K	5.6M
6.2	62	620	6.2K	62K	620K	6.2M
6.8	68	680	6.8K	68K	680K	6.8M
7.5	75	750	7.5K	75K	750K	7.5M
8.2	82	820	8.2K	82K	820K	8.2M
9.1	91	910	9.1K	91K	910K	9.1M

Color Codes

Even the old-fashioned through-hole parts are a bit too small to legibly label with a numeric value and tolerance factor. A system of color codes was developed to designate resistor values, using colors to represent both single digits, a power-of-ten multiple, and the resistor's tolerance. These colors are painted on the resistors in bands. A little memorization and careful practice will have you reading resistor color codes in no time.

For 5 percent tolerance resistors, only four bands are needed to completely specify the resistor's value and tolerance. The first two bands represent the first two significant digits of the value. The next value is the multiplier. The final value is a special case to tell the resistor's tolerance.

The first two bands, as well as the multiplier value, use the color codes shown in Table 3-3 to represent the digits 0 through 9. (Unfortunately, these colors were chosen before the subtleties of colorblindness were widely understood.)

Table 3-3. *The Resistor Color Codes for the Significant Digits and Multiplier Digit of the Resistor's Value*

Digit	Color
0	Black
1	Brown
2	Red
3	Orange
4	Yellow
5	Green
6	Blue
7	Violet
8	Gray
9	White

The last band indicates the tolerance of the resistor using a different set of colors, as shown in Table 3-4. The most common are 5 percent (gold) and 1 percent (brown). There is supposed to be a slightly larger gap between the digits and the tolerance bands, but this is sometimes hard to see.

Table 3-4. *Resistor Color Codes for the Final Color Band*

Tolerance	Color
0.05%	Orange
0.25%	Blue
0.5%	Green
1%	Brown
2%	Red
5%	Gold
10%	Silver
20%	None

The multiplier digit can be thought of as either the number of 0s after the specified digits or as the exponent to which the number 10 is raised, as in scientific notation. Some examples make this easy to understand and remember.

A resistor marked with brown, black, red, and gold stripes has a resistance of 1KΩ with a tolerance of 5 percent. An easy way to decipher this code is to do a simple substitution for the first two digits (brown is 1 and black is 0, so we have 10 so far). Next decipher the multiplier band and write down that many 0s after the first two digits. Red is 2, so write down two 0s. That gives us 1000 so far. The final band is gold, and that means a 5 percent tolerance; 1,000 ohms is usually abbreviated as 1.0KΩ or simply 1KΩ.

■ **Note** *K* stands for *kilo*, the Greek prefix for 1,000. Likewise, *M* stands for *mega*, the prefix for 1 million.

This method works for most values from 10Ω to 10MΩ. A problem occurs, however, when we need to specify a value less than 10, as a negative multiplier value must be used to get an exponent of less than 0. For the value 1 and those less than 10Ω, a gold band is used for the multiplier. For values less than 1Ω, a silver band is used.

To save even more precious space on our schematics, another convention is to replace the decimal point with the multiplier. For example, 4.7KΩ would be written as just 4K7. 1.8MΩ would become 1M8 and 1.5Ω would be 1E5. 10KΩ would still be 10K using this method. If it is clearly understood that the unit of measure is the ohm, then even the Ω can be omitted.

Resistor Power Ratings

Any resistor has a maximum power-handling capability. This rating is usually expressed in watts or fractions of a watt. The larger through-hole resistors shown previously in Figure 3-9, for example, have a power rating of 1/2W (or 0.5W, if you prefer). This means that if you attempt to dissipate more than half a watt of power through the resistor, it is most likely going to quickly overheat and fail.

Thermal failure in a resistor starts with a thin curl of smoke, and then results in the discoloration of the resistor itself and sometimes actual flames. In any case, it gets dangerously hot, so do try to avoid this situation.

If you know the voltage across the resistor, as well as the current, you can easily calculate the power being dissipated using the simple formula P = V × I. Remember that P is the power value in watts, V is the voltage in volts, and I is the current in amps. One volt at one amp is one watt, which is way more than enough to smoke our little resistors. Be careful! Any time you calculate the correct value for a resistor in ohms, be sure to also calculate the required power-handling capability.

Variable Resistors

So far we've only looked at *fixed*-value resistors. They usually exhibit the same resistance day in and day out. That's what we love about them. But what if you need to vary the resistance in a circuit? We have resistors for that, too! A good example is a volume control on an amplifier. Turn it to the left (down), and the volume goes down. Turn it to the right (up), and the volume goes up. How does that work?

One kind of variable resistor is called a *potentiometer*. Potentiometers are sometimes abbreviated as *pots*. Pots usually have three terminals instead of two. Two of the terminals connect to the fixed resistor that is inside the potentiometer. Instead of being all sealed up like the fixed resistors you've been studying, the internal, resistive conductor is exposed along a track or groove. The third terminal, which is often physically in the center, between the two other terminals, is called the *wiper*, and can be moved from one end of the fixed resistor to the other end, making electrical connection to the inner part of the fixed resistor as it goes. As the wiper moves along the length of the resistive element, it taps into the resistor at a different point, varying the resistance. This is how a potentiometer can be used as a variable resistor.

Figure 3-12 shows a typical three-terminal potentiometer with a round body. The shaft can be turned to adjust the resistance. There are many variations on this basic design.

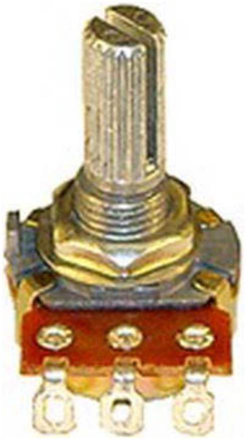

Figure 3-12. *A typical potentiometer, or variable resistor. Image by BG Micro (http://bgmicro.com).*

The schematic symbol for a potentiometer looks suspiciously like a resistor with an extra terminal, which happens to be exactly what it is. The extra terminal is the wiper, shown as the upward-pointing arrow. See Figure 3-13.

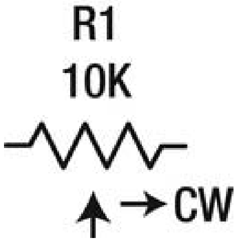

Figure 3-13. *The schematic symbol for a potentiometer is a fixed resistor with a wiper terminal. The "CW" arrow indicates which way the wiper goes when the shaft is rotated clockwise.*

Potentiometers generally come with two distinctive *response curves*, or *tapers*:

- A *linear* curve means that every equal step of rotation about the axis changes the resistance by the same amount. For example, if you were to center the shaft in the middle of its rotation, the resistance from the wiper contact to either of the end contacts would be nearly exactly half of the total resistance between the two end contacts. This is quite useful for when you need to adjust a voltage up or down in a relatively straight line. This is sometimes indicated by adding a capital *B* to the end of the resistance value (e.g., 10KΩB).
- In many audio circuits, however, it has been discovered that the human ear is anything *but* linear. A special response curve that more accurately describes this sensitivity is called an *audio* curve or taper. An *A* is used as a suffix to indicate that an audio taper is used for a potentiometer.

Note that linear and audio tapers are *not* interchangeable. If you find some random potentiometers in a parts bin, you can get a quick idea of what you've got by centering the shafts and seeing if the resistance is "symmetrical" or not, using an ohmmeter. Equal resistance to either end terminal indicates a linear taper, and anything else is mostly probably an audio taper.

Just like any other resistor, a potentiometer has a maximum power rating as well as a tolerance. These are usually either printed on the body of the potentiometer, if there is room, or hidden away in a product data sheet somewhere.

Specialty Resistors

There are many different kinds of resistors made for many special kinds of applications. High-power resistors are built out of sturdier materials that can take the thermal stress required to dissipate large amounts of heat. See Figure 13-14.

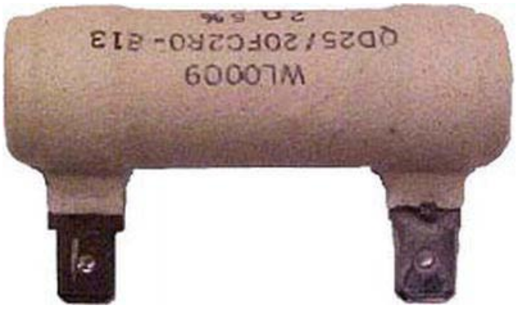

Figure 13-14. *A power resistor can handle much more heat dissipation than the smaller though-hole or surface-mount varieties. They are often so big that they need special mounting hardware. Image by BG Micro (http://bgmicro.com).*

Another popular option is the *multiturn* potentiometer. These potentiometers have an internal gear system that allows their shafts or wipers to be moved slowly across their internal resistive tracks by multiple turns of the control element. Increasing the number of turns allows the user much finer control of the resistance adjustment. It also takes the user longer to get from one point to another along the scale.

Capacitors

The basic definition of a *capacitor* is two conductors separated by an insulator. By itself, it doesn't sound like that would be a very interesting configuration at all. It turns out, however, that there's a lot more to it.

When a voltage is applied across a capacitor, no steady current flows through it. Remember, there is an *insulator* in the middle of that capacitor. Any insulator, just like the insulation on a piece of wire, *does not conduct electricity*.

What happens is that an *electric field* is created between the two conductors, or *plates*, as they are called when they are part of a capacitor. Now we get back to the fairies and dragons and molecules and so forth, way down deep at the heart of the problem. This electric field is a form of energy. Once it is built up by the difference in voltage, or *potential*, between the two conductors, it has nowhere to go, so there it stays . . . at least until something else in the circuit changes.

For this reason, capacitors are sometimes used as *reservoirs* of electrical energy in a circuit, holding a bit of spark here and there as needed. Just as it takes electrical energy to build up this electric field, a certain amount of electrical energy will be released when the field collapses. This will happen when the voltage between the two conductors drops.

Using capacitors as tiny electric batteries in a circuit is only one of many applications for this versatile part. While a capacitor will not conduct a *steady* flow of electric current between its terminals, it will often *appear* to conduct *alternating current* (as opposed to *direct current*), which is why you see them so often in audio circuits. As a charge builds up on one plate within the capacitor, it then attracts electrons from the other side, pulling them in from the other side. Even though they don't "jump the gap" of the central insulator, it looks like a tug-of-war is going on from both sides.

This behavior can also be used to send the higher-frequency components of a signal one way in a circuit and the lower-frequency components of a signal in another direction. This is one way that analog filters are built, and it has all kinds of useful and interesting applications.

Capacitors come in all shapes and sizes. They are made of many different materials, depending on what particular characteristics are needed. See Figure 3-15.

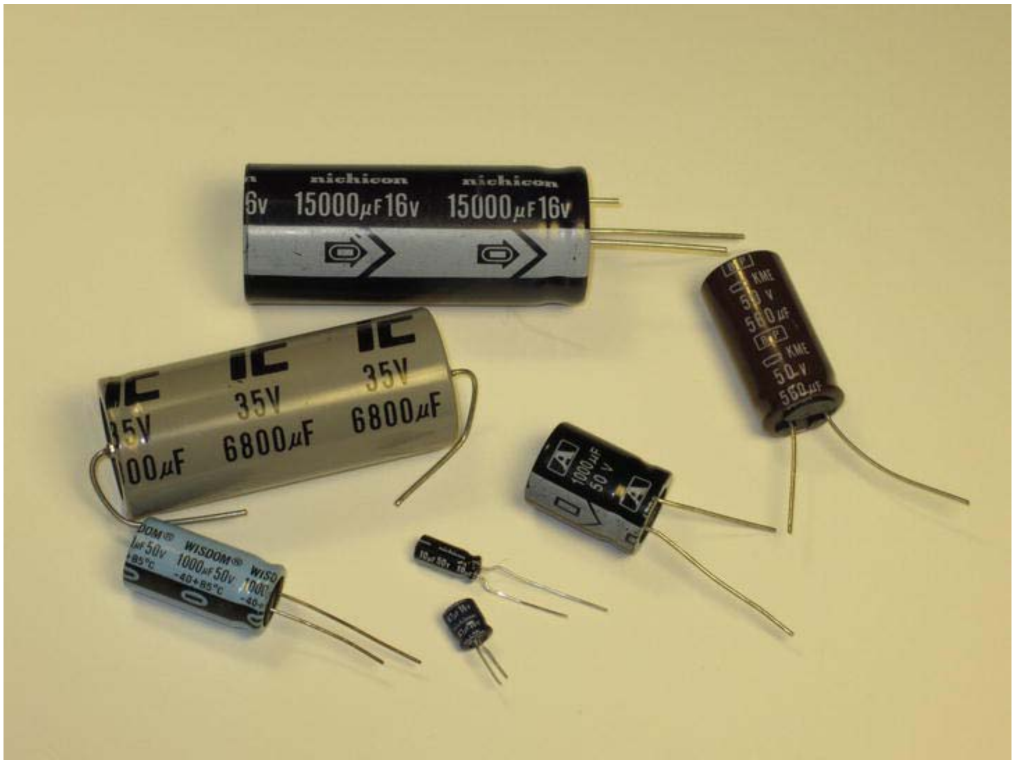
nichicon
15000µF 16V
35V
6800µF
KME
50 V
560µF
1000µF
50 V
WISDOM
1000µF 50V
-40+85°C

Figure 3-15. *Some electrolytic capacitors of various sizes and capacities. Both much larger and much smaller capacitors exist and are very common in electronics.*

Schematic Symbols

While there are a large number of special-purpose capacitors that are custom made for certain applications, a very large number of capacitors fall into two broad categories: polarized and nonpolarized. The schematic symbols for both types of capacitors display the fundamental construction of two conducting plates separated by an insulator. See Figure 3-16.

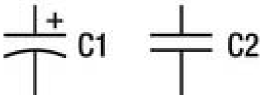

Figure 3-16. *The schematic symbols for polarized capacitors (left) and nonpolarized capacitors (right)*

C1 is a *polarized* capacitor, meaning one terminal is more positive than the other. In practice, this is indicated on the part itself in a variety of ways. One convention is that the *anode*, or more positive lead, is physically a little bit longer than the *cathode*, or negative lead. There is generally some sort of mark made on the capacitor body itself. Usually this is a stripe on the *negative* side of the capacitor.

What's really odd and potentially confusing is that one type of capacitor, specifically that made primarily of tantalum, has a stripe that indicates the *positive* lead of the polarized capacitor. Know your capacitors!

A polarized capacitor that is installed backward will experience terrible physical and thermal stresses and can sometimes explode, sending bits of itself all over the place.

C2 in Figure 3-16 illustrates a *nonpolarized* capacitor. It doesn't matter which way it is installed in a circuit. It is electrically symmetrical and works just as well in either orientation.

Inductors

An inductor is almost always made out of coils of wire. Sometimes inductors are quite large, as with transformers, while other times they are as small as a speck. An inductor would appear, electrically, as a simple short-circuit if it weren't configured to take advantage of certain principles of *electromagnetism*, or the tendency for the flow of electric current in a conductor to form a *magnetic field* around the conductor. The shape and orientation of the loops of wire that compose the inductor determine its electrical characteristics.

In many ways, an inductor is the opposite or complement of the capacitor. While a capacitor does not conduct direct current from one terminal to another, an inductor will happily do so, at least for a time. While a capacitor will resist a change in voltage across its terminal as it builds up its electric field, an inductor will resist a change in current flow across its terminal once it has established a magnetic field.

Because of their complementary natures, the two components can be used together to form a *tuned circuit*—one that resonates at a particular frequency, or allows the passage of only a certain range of frequencies, with the capacitor shuffling the high-frequency portions and blocking the low-frequency portions, while the inductor does the opposite, blocking the high-frequency components and passing the low-frequency ones. This is how many older radios and televisions worked before the advent of low-cost solid-state electronics.

Once an inductor has built up a magnetic field, any drop in the current flow across its terminals will result in the collapse of the field and the generation of a very-high-voltage spike. This is the basic electrical circuit that was used in automobile engines to create the high-voltage spark needed to ignite the fuel-and-air mixture in the combustion chamber.

Inductors can also be used to efficiently pass alternating current via a magnetic field from one inductor to another. This is how the primary windings of a power transformer *induce* a current in the neighboring secondary coil, which, depending on the ratio of windings and a few other factors, determines if the transformer *steps up* the voltage or *steps down* the voltage. It can even keep the voltage the same, for when electrical isolation is needed between two different systems.

Most electric motors employ some sort of inductor that alternately attracts and repels another electromagnet or permanent magnet. The most basic direct-current (DC) motors require a fairly complex mechanical arrangement to perform the necessary switch-swap to alternate the fields. This process is called *commutation*. In other types of electric motors, such as brushless motors and stepper motors, this commutation process is done electronically, externally to the motor itself, and requires more complex circuitry to control the speed and direction of the motors. See Figure 3-17.

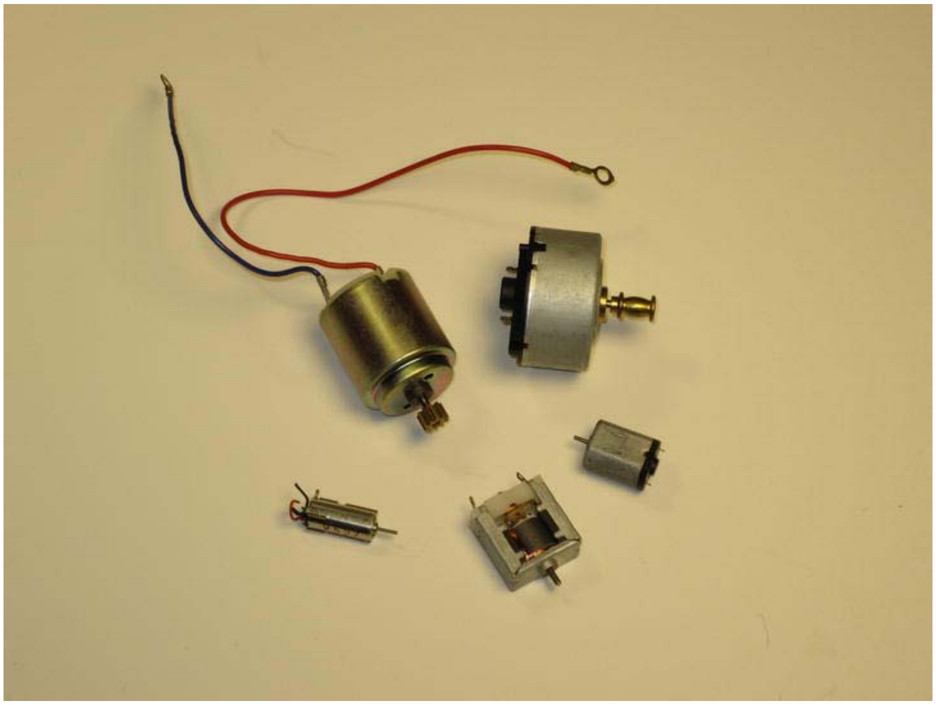

Figure 3-17. *Some simple DC motors. When direct current is applied to their terminals, they spin. The magic happens with an electrical and mechanical arrangement inside that performs the necessary commutation of the magnetic fields.*

Switches

The most basic way to control the flow of electric current in a circuit it to literally *break* the circuit. This means that instead of forming a conductive loop from power source to components and back again, a gap is introduced somewhere in the loop.

This gap can come anywhere in the loop. It can be right next to the battery, for example, or somewhere off in the middle—as long as the circle is broken, no current will flow.

If you perform the lightbulb experiment in Chapter 1, you can see how this works. When everything is lined up properly and the final connection is made, the lightbulb lights up. When any of the connections are broken, the light goes out. It's really as simple as that.

Most of the electrical switches that you operate every day are really mechanical in nature. Their main function is to physically move two conductors either together or apart.

In the lightbulb experiment, you could use the wire as a switch, by physically touching either the battery terminal or the lightbulb (depending on how you were holding everything together). With a little practice, you could probably send Morse code, or at least a variation on "one if by land, two if by sea."

There are many kinds of switches for many kinds of switching tasks. The most basic are *momentary action* switches. These are usually push buttons, and only conduct electricity when pressed. Alternately, they can be built so that they conduct electricity all the time, *except* when pressed:

- The first option is called *normally open*. When a circuit is open, it is not conducting electricity. When you *close* the circuit, by completing the loop, electricity can flow. A normally open switch would be welcome in a doorbell circuit. When no one is pressing the button, no bells are ringing. It would be annoying the other way around.
- A *normally closed* switch acts like a solid piece of wire (being a conductor of electricity) until it is activated, when it ceases to provide a path for the electricity to follow. This type of switch can be used as a *kill switch* on a gasoline-powered engine, such as a lawn mower, for example. By interrupting the flow of electricity generated by the magneto and powering the spark plug, you interrupt the internal combustion cycle, which will eventually cause the engine to slow down and then stop.

This type of switch is called a *single-throw* switch. It's either on or it's off—conducting electricity or not conducting electricity.

■ **Note** For almost every possible configuration of switches, there is a corresponding *opposite* configuration that is almost as popular.

Another useful configuration for a switch is to switch from one circuit to another circuit, and not just on and off. This is referred to as a *double-throw* circuit. A switch like this normally has three terminals. A *common* terminal is usually, but not always, in the center. At any rate, the common terminal is electrically connected to the moving part of the switch that is going to make contact with one or the other two remaining contacts. The *normally closed* contact is the one that will be making contact with the common terminal when the switch is not activated, or at rest. The *normally open* contact is simply the other terminal. This assumes that the switch has some sort of bias or mechanical spring action that returns it to a certain orientation when no one is activating it.

Some double-throw switches can remain in either state indefinitely. A typical light switch used in indoor lighting is a good example. It's just as happy to stay on as off.

Switches can have more than one circuit built into them. These switches are called *double pole*, *triple pole*, and so forth. They're like two or more single-pole switches physically taped together, so that when one is switched, they all are switched. Usually these poles are electrically isolated from each other.

We'll play with switches again in Chapter 5, where you'll learn some interesting ways to wire them up.

Active Components

While it could truthfully be said that *all* electronic components require some electricity to do anything, the distinction made in this book between *passive* and *active* components is that *active* components need a little more power to be really interesting.

First we'll look at sources of this magical electricity, and then we'll move on to some of the fascinating and complex components that will happily use it all up for you.

Power Supplies

You're going to need some sort of power source for any electric or electronic circuit you run across. If it doesn't use electricity in some structured form, it's not really electronics, now is it?

Batteries

You can use regular batteries to power a lot of your circuits, but they eventually wear out, and then you have no real recourse other than to throw them away or recycle them, if possible. If you're going to be using batteries, at least get some good ones. Many of the bargain brands are terribly inferior in their manufacture and are only providing you with a false economy. Spend a little more and get higher-quality parts, and you will be rewarded with longer life and more reliable operation.

A great way to work with batteries is to use the appropriate battery connectors or battery holders. Sometimes you can remove these from obsolete electronic devices that have outlived their useful lifetime. Always observe the proper orientation and polarity of batteries. Even smaller batteries contain an annoying amount of power, especially when it all decides to come out at once, such as when you accidentally plug something in backward or short something out. This can result in damaged parts, burns, or even fires. Be careful with batteries.

Having a fuse holder wired inline with your battery-powered supply is an excellent and cheap way to prevent such "unauthorized thermal events." Keep several spare fuses on hand as well. Fuses cost little and save much.

Rechargeable batteries are becoming more common, and they keep coming down in price while improving in longevity and power capacity, and decreasing in size and weight. Any rechargeable battery technology beats any kind of primary, or nonrechargeable, battery in the long run.

■ **Caution** The same warnings about battery polarity and fuses apply to rechargeable batteries as well.

Since there's no easy way to "turn off" a battery, it would be a good idea to include a power switch in your supply, if possible. Just make sure that whatever switch you select can stand up to the maximum current you will want to be extracting from your power supply. You should also make your switch large and easy to see, so that you can find it quickly when you need it. An obvious power-on indicator is also a great plus for a homemade power supply. This can be as simple as a green LED (with appropriate current-limiting resistor—see below) or as complex as a voltage meter indicating how much juice is left.

Transformers

Another great power supply that you probably already own is a power adapter or transformer from another electronic appliance. You should use caution, of course, with anything that plugs into the wall for power, especially if you're planning on monkeying around with the other end of it.

Almost all commercially produced power adapters have a slew of certifications plastered on them. Somewhere in there should be the basic data concerning what type of power input it expects, what it is going to try to produce, and what the maximum limits are on the output. Make sure that whatever it is that you're going to plug into your wall is specifically designed to plug into *your* wall, as electrical power standards vary from one location to another.

Make sure that any repurposed power supply you use does not get overly warm when in use. This is an indication that it is presently or has been overloaded in the past. Such a device is a fire-starter waiting for its magic moment.

You need to know how much voltage is coming out of your power supply. Even if it is clearly printed on the device itself, if you have access to a voltmeter, you need to verify the output voltage before connecting it to anything. Many lower-cost power adapters are *unregulated* and will produce an output voltage that will vary wildly depending on the load applied to it. A *regulated* voltage output, on the other hand, will provide consistent voltage levels on the output, even with varying loads.

USB adapters and chargers are becoming quite common and can provide *regulated* +5V at up to 1/2A and sometimes more. These are great devices to use for small electronic experiments. They are generally small, lightweight, and reliable.

A dedicated bench power supply is a very handy piece of test and measurement gear to have available in your lab. A good supply will have a variable voltage and a readout to show you what's going on (or coming out). A step up and you get to set the maximum current that it will supply (to keep things from getting out of control too fast). Additional desirable features include multiple outputs and remote control from your PC.

No matter what kind of power supply you use, use it responsibly. Don't leave equipment powered up if you're not using it. Don't attempt to bypass safety measures that are in place for *your* protection.

Diodes

A diode is a semiconductor device that will conduct current in one direction only. That's the simple story. Of course, the reality is much more complex. A typical diode is composed of two pieces of a semiconductor material, usually silicon, but sometimes germanium or some other exotic substance. The diode action occurs in the *junction* between the two types of semiconductors, which have been specially manufactured to contain the exact right balance of impurities required. This is generally not the kind of thing you can build in your lab, unless you're Jeri Ellsworth.

These two bits of semiconductor, called the *N region* and the *P region* (for negative and positive, respectively), are sealed up in either a glass or plastic enclosure. A standard diode has two terminals—an *anode* and a *cathode*—which represent the positive and negative leads, again respectively. Conventional current (i.e., current flowing from the more positive voltage to the more negative voltage) can flow through the diode from the anode to the cathode, but not the other way around. It's magic! Well, technically it's semiconductor physics, but either answer will serve for our purposes.

Diodes come in all sizes. Generally speaking, the more power a diode is expected to handle, the larger it is going to be. *Small-signal* diodes are the same size or smaller than typical resistors. Power diodes, often called *rectifiers*, depending on their application, can be so large they have to be mounted with large bolts.

The two most important electrical characteristics for the happy care and feeding of diodes are their maximum reverse voltage and forward current capability. You don't want to exceed either one.

Since a diode is, at best, only ever going to have a cryptic part number written on it, you'll need to consult the manufacturer's data sheet to find out all the necessary information. Many semiconductor producers offer "commodity" parts that conform to standard part numbering schemes. That way, you know that a 2N4148 diode from one manufacturer is going to be pretty close to a 2N4148 diode from any other manufacturer, but you should always check to make sure.

Diodes typically have a *forward voltage* that must be exceeded before current will flow. This voltage is essentially lost and is dissipated as waste heat from the device. Silicon diodes will have an average forward voltage around 0.7V and up. Germanium diodes will have a smaller forward voltage, in the area of 0.3V. Schottky diodes use a metal-to-semiconductor interface region and achieve substantially lower voltage drops, making them much more suitable for higher-power applications.

LEDs

One special kind of diode is the popular *light-emitting diode*. You see these little guys everywhere these days. In fact, some folks go around with black tape, just trying to cover them all up so they can get their darkness back at night.

The reason that they are everywhere is that they are cheap, reliable, and easy to use. They don't take up much room, don't take up much power, and don't generate that much heat. The exceptions to these admirable qualities, of course, are in the newer generation of LED lighting products that emit tremendous amounts of light and produce tremendous amounts of heat. Every month it seems that a new record is set for efficiency. Let's hope that trend continues.

Being essentially diodes, LEDs have both an anode and a cathode. LEDs only conduct current in a single direction. You'll need to make sure you limit the amount of current flowing through your LEDs, because the LEDs themselves sure aren't going to do it.

LEDs designed to be indicator lights can be powered with as little as 1mA. High-efficiency LEDs can be used for illumination purposes and still only draw 20–30mA each. These very modest power requirements make LEDs a favorite for microcontroller projects, as these output power levels are easily handled by most popular chips.

The forward voltage drop across LEDs is typically much higher than for a rectifier or signal diode. In fact, the voltage goes up as the emitted color of the LED goes from the infrared and red end of the spectrum (1.2–2V) up to the blue and ultraviolet end of the spectrum (3–3.6V).

Transistors

A true modern miracle, the semiconductor transistor ushered in the age of solid-state electronics. More than a million individual transistors can be used to build modern computer processors, all on a single chip.

The first types of transistors to be built and used were called *bipolar junction transistors (BJTs)*. This type of transistor is similar in construction to the diode, except that there are two semiconductor junction areas inside.

This arrangement of magic beans allows a small amount of electrical current to be applied to one of the three terminals of the transistor, called the *base*, which in turn allows a much greater amount of electrical current to flow between the remaining two terminals, known as the *collector* and *emitter*.

Two complementary types of BJTs are made, depending on the way that the positive and negative semiconductors are sandwiched together. NPN transistors will conduct conventional current from the collector to the emitter when current flows from the base to the emitter. The opposite variety, called the PNP transistor, does precisely the opposite, conducting conventional current from the emitter to the collector when the base is negatively biased.

Transistors can be used to make amplifiers, switches, current limiters, voltage regulators, oscillators, and any number of other useful circuits. They are generally very low cost, lightweight, and reliable.

As with any other electrical components, you need to be aware of the electrical limitations of your transistors. Even when a transistor is fully saturated and conducting very efficiently, it's still dissipating a certain amount of its energy as heat. Too much heat and say goodbye to your transistor.

A newer type of transistor is becoming quite popular in power circuit applications, and uses a different type of semiconductor property to accomplish its mission. The *MOSFET (metal-oxide–semiconductor field-effect transistor)* has three terminals, just like the BJT. Each terminal performs a similar job, but does it in a different way. The *gate* of a MOSFET is electrically isolated from the conduction *channel* between the *drain* and *source* terminals. However, when a proper voltage is applied to the gate, it creates an electrical field that influences the conductivity of the channel between the drain and the source.

The important differences between BJTs and MOSFETs lie in the type of signal needed to drive each of the transistors and how much power is wasted by being converted to heat. A BJT is a current-controlled device, where more current into (or out of) the base results in more current flow through the device. The MOSFET is a voltage-controlled device, where more voltage applied to the gate increases the current flowing through the channel.

Additionally, MOSFETs typically have much lower on-state resistance than BJTs, meaning they can more efficiently switch higher currents. Finally, MOSFETs have a *positive temperature coefficient*, meaning that they conduct less power as they heat up—which is a good thing. BJTs are subject to *thermal runaway*—conducting more current as they heat up, which in turn generates more heat, and so on, until something snaps, crackles, or pops.

Integrated Circuits

As parts get more specialized, they become less fun. There are a hundred ways to wire up some transistors, resistors, and capacitors, but there's generally only one way to hook up a thermocouple interface chip.

The good news is that it has become much easier for companies to design and build vastly complex chips that can do things for us that, honestly, most people cannot comprehend. And that's OK. We need airplanes that can, mostly, fly themselves. If we had to depend on humans to get everything done, you can imagine what kind of world we'd live in. Humans are notoriously sleepy, hungry, distracted, and forgetful. A voltage comparator chip does nothing but compare two voltages, all day, every day, perfectly, all the time.

Integrated circuits come in an unbelievably wide range of functions, packages, and capabilities. The first integrated circuits were amplifiers. Integrated circuits allowed manufacturers to make chips that have dozens or hundreds of transistors and other parts *all at once*. Just plop that little silicon chip into a plastic package with legs on it, and you're done.

What kind of chips do you need in your lab? Well, that's going to depend heavily on what kind of circuits you're wanting to design, build, or repair.

Many older chips had a standard 0.1" pin spacing that made them easy to plug into a solderless breadboard for prototyping purposes. Most newer chips have been considerably shrunk, because phones and tablet computers are all the rage these days. It's still possible to work with these newer chips, but it almost always requires mounting them to some sort of breakout board or expander that brings out their tiny leads to a more reasonable, and reachable, size.

Acquiring Components

Where do components come from? The same place as everything else—the store—just like food, just like money.

The electronic component manufacturing industry spans the globe. If someone's not putting something electronic together, they're taking something electronic apart.

Once upon a time it took some hunting to find sources for electronic parts. Swap meets and "ham fests" were great places to find components, assemblies, and entire appliances just ready to be hacked, studied, improved, or rebuilt. Now you have to spend time looking for places to *get away* from all the electronic hustle and bustle.

One of the great things about today's Internet is that it takes almost no effort to find both electronic parts and the information needed to use them. The real trick these days is finding the best deal.

If you're strictly a hobbyist, looking to spend some quality time alone or with your kids, learning about electronics and building fun projects, then it just about doesn't matter if you get the absolute, rock-bottom price on every component and tool. Even if you're a budding entrepreneur, you're still in a much better position today to leverage those brilliant ideas floating around in that noggin of yours than you would have been even ten years ago, all because of the ubiquitous nature of electronics and technology saturating our lives. Not only can you get the right tools and the best parts shipped right to your door for a song, you can easily network with knowledgeable folk from all over the globe, contributing ideas and brainstorming across time zones and borders.

Buying New Components

It would be more than a full-time job just keeping up with all of the new tools, parts, software, and hardware that are available today. Even if you could, it would take many lifetimes to make and grow the connections and relationships you need to successfully compete in the global economy.

That being said, a good place to start is in your own backyard. If at all possible, you should deal locally for the things you need. When what you want or need cannot be obtained locally, expand outward until you find it.

While it's certainly nice to be able to find all your needs met close to home, it's also good to know that the world is shrinking every day. The sort of deals and relationships that couldn't possibly have existed in your parents' time are happening right now.

If you are in a position to spend some money on quality tools and components, then by all means do so. You will not regret it. Buying questionable items is like buying your groceries out of some guy's van in a parking lot. Maybe he'll be there next week and maybe he won't.

Even the small fry can get treated like an honored guest when dealing with the right people. The trick is finding the right people. The best way to find them is to ask your friends and others with similar interest where they trade. Ask them how they feel about their relationships with their vendors. Are they just a number? Or are they a partner?

Recycling Used Components

It's really a bad idea to waste anything. In a more perfect world, there'd be no "other" category at the recyclers. Everything would have an exit strategy, especially the complex and fragile toys we're so addicted to these days.

Until that day, try to look at your trash with new eyes. Is it conceivable that someone else might be able to use this equipment, or is it truly past its useful life? Or are you just being lazy?

You can do your part, and you should believe that others are doing their part as well. Take the time to find them, find out what they can offer you, and find out what you have that they might be able to use. You won't know until you ask.

Be careful when harvesting components with which you are unfamiliar. Novice recyclers should give yesterday's cathode ray tube (CRT) TVs and monitors a wide berth. They can retain *lethal* voltages for considerable periods of time after being retired from service. Just leave them on the side of the road. The same goes for microwave ovens, or any equipment that has refrigerant in it. The stuff they use nowadays is quite toxic.

There is a growing surplus of both components and “last year’s models” available to the student of electronics these days. Sometimes it’s worth your time and money just to take something apart to see how it works. Then sometimes you find some useful parts inside that you can use in projects of your own.

Component Data Sheets

Every component ever manufactured has a data sheet. This data sheet contains everything you ever wanted or needed to know about this particular component. It’s probably available right now on the manufacturer’s web site, along with a dozen application notes. If not, then it’s probably available on a university’s computer system in a project folder.

This isn’t always the case, sadly. Past a certain point, it doesn’t make sense to allocate resources to information that no one wants or needs anymore. That point varies with each company, with some diligently archiving their information and others tossing it out the window at the end of the quarter.

When possible, collect and maintain your own private set of data sheets. When you acquire a new piece of equipment, get everything you can find in electronic form and stash it away somewhere where you can find it when you need it. If you’re experimenting with some new components, squirrel away those data sheets, just in case you strike gold.

Have you built something cool? Write up your own data sheet for it! Pretend that you’re still going to be interested in this project next year. Better yet, imagine that *someone else* could take an interest in your work. You’ll be pleasantly surprised at how useful you’ll find this information in the future, after you’ve slept a bit.

Summary

This chapter has covered a lot of different types of electronic components, but it has only scratched the surface of what is presently available.

It’s perfectly possible to spend a lifetime pleasantly tinkering with electronics and never have a clue how any of the underlying principles work. Sometimes all you need is a *practical* understanding of what the different components do in a circuit. Hopefully this chapter has given you a rough outline of what is involved. Ideally, your interest has been sparked enough to pursue more detailed study elsewhere. Good luck and happy discovering!

Now that you’re armed with some basic tools and some ideas about some of the available components, it’s time to get busy building your own lab, one experiment at a time.

The remainder of this book will take you on a journey of trial, error, and experimentation that will help you build the best possible lab for your needs, both for now and in the future.

CHAPTER 4

A Portable Mini-Lab

Having a portable (or at least (*trans*portable) electronics lab comes in handy in several circumstances. Maybe you don't have a place (yet) for a permanent home for all that equipment. Maybe you need to bounce at a moment's notice, heading off on electronic adventures at the drop of a hat. Or maybe you like to keep everything where you can find it in a hurry, without having to rummage through shelves and boxes, looking for just the right tool or component.

Whatever your motivation for wanting a portable lab at your disposal, this chapter should help you get started.

Have Lab, Will Travel

First, a note or two is in order concerning *travelling* with your electronics lab. Nothing makes security screeners at an airport or customs inspectors at border crossings more nervous than a bunch of mysterious devices, especially devices that are tangled up in a rat's nest of wires.

Do what you can to keep your collection of tools and components tidy and in working order. You will most likely be asked to at least turn on any questionable items and demonstrate that they are not hollowed-out shells for transporting questionable items. It also doesn't hurt to clearly label equipment as to its function and ownership, which makes sense in any case, not just travel scenarios.

The Compact Executive Model

"A full mongoose is a slow mongoose." This is an important lesson learned from Rudyard Kipling's *Rikki-Tikki-Tavi*. Rikki, a mongoose, has to stay fast on his feet to be able to combat Nag the cobra and hope to win. Therefore, Rikki learns to only take a few bites at a time when eating and never gorge himself.

Although hopefully not in a life-or-death situation like Rikki's, you can learn a thing or two from this simple philosophy. Only take what you absolutely need when travelling, and never take something you can obtain at your destination. These things will only slow you down. Don't be a slow mongoose.

■ **Tip** Travel light.

A slim and lightweight briefcase or attaché can hold many of the tools and components you might need for your mobile electronic experimentations. The main advantage is that you will only need one hand to carry it (leaving one hand free to open doors, bribe guards, or fend off ninja—assuming you have two hands to begin with). See Figure 4-1.

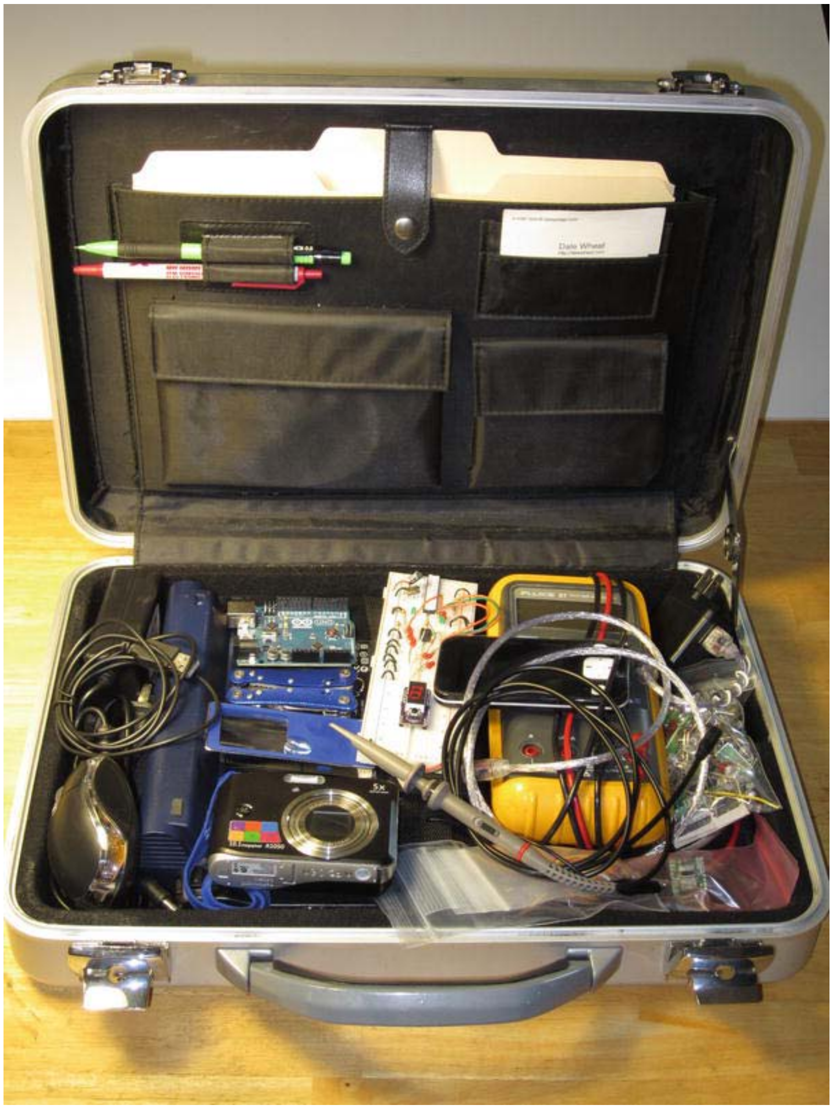

Figure 4-1. *The compact executive mobile electronics laboratory. There's plenty of room for lots of tiny things, but not much else. Also handy for transporting big stacks of cash, should the situation arise.*

Let's take a look at what all you can take with you without a lot of heavy lifting. Starting on the left and moving to the right, we have

- Netbook PC with AC charger and USB mouse (it might be hard to see in Figure 4-1 because it's upside down and covered by other stuff)
- Arduino Uno programmable microcontroller (or *physical computing platform*)
- Folding multitool
- Lighted magnifier
- Digital camera
- Solderless breadboard with lots of precut jumper wires
- Digital multimeter and probes
- Compact single-channel oscilloscope with probe
- AC-powered USB charger—can also be used as a stand-alone regulated 5V power supply
- USB cable
- A variety of LEDs in all shapes and sizes
- Additional passive components
- A variety of batteries, both rechargeable and otherwise, and various battery holders (not shown)
- Static-sensitive integrated circuits and modules stored in antistatic bags
- Notebooks, writing implements, blank media, and documentation
- Your contact information in case of loss

One of the interesting features of this lab is that it is completely self-contained, power-wise. While it's nice to have access to AC power, it's not strictly required, at least for a while. The Arduino and the attached breadboard are powered by the battery in the netbook, via the USB cable. The multimeter has its own 9V battery, while the mini-oscilloscope can be recharged from a spare USB connector. With an extended-life battery, you can expect four to five hours of on-location electronic experimentation.

If your netbook or PC has Wi-Fi or other connectivity to the Internet, you can browse tutorials, data sheets, user forums, and even this book, unless you've already downloaded these items to the internal hard drive.

Have a look at this particular lab deployed and ready for action in Figure 4-2.

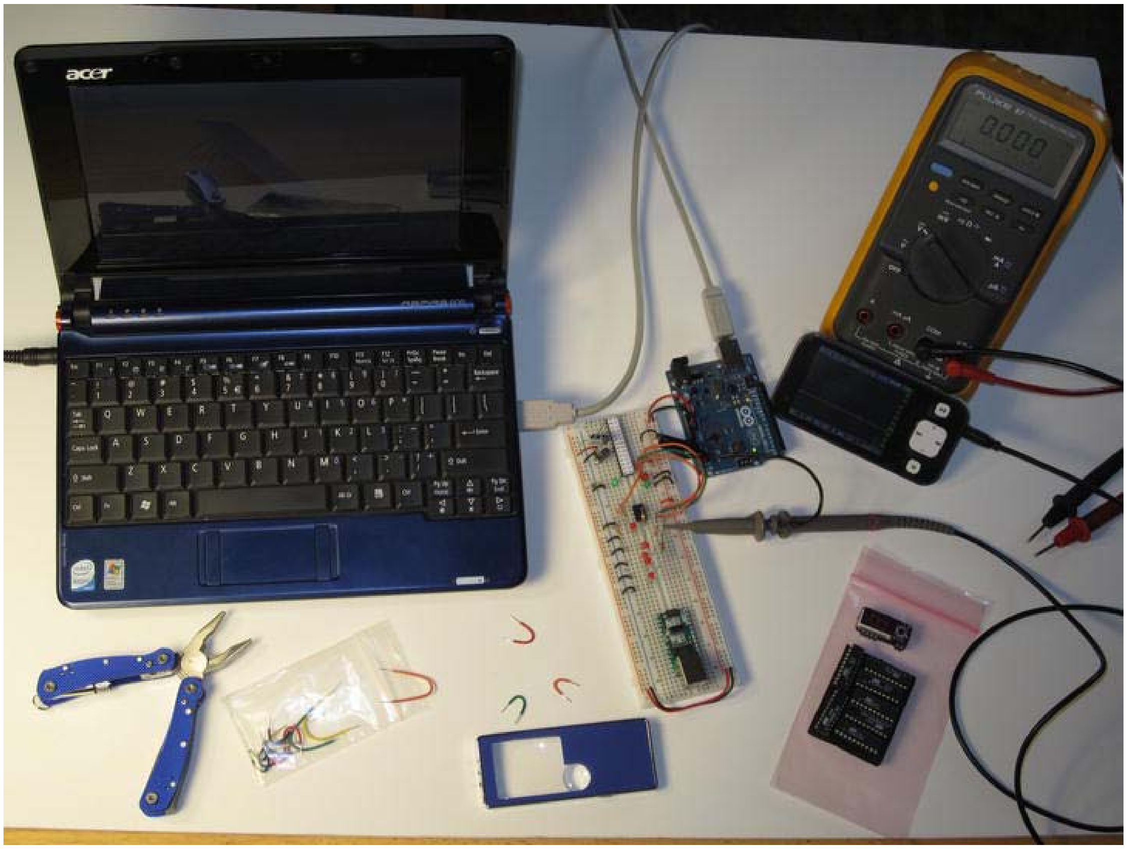

Figure 4-2. *The compact executive lab is deployed on a desktop. You have access to lots of electronic building and debugging capabilities in a small area. Using a solderless breadboard helps reduce the number of fabrication tools that you need to carry with you on your adventures.*

The Arduino was included for its small size, popularity, and ease of use. There's a whole world waiting for you inside that little board, waiting to be explored, but that's a story for another day.

You can use the USB cable to extract a regulated +5V supply from your PC for your low-voltage-experimenting pleasure. You can also use an inexpensive USB charger to do the same thing, if you don't feel like hauling a whole PC around with you. Just don't try to draw more than half an amp of current, which is the "official" USB limit for supplying power to USB devices. Technically, *unenumerated* devices (i.e., just a USB cable with nothing intelligent enough to identify itself to your computer) can only draw up to 100 mA (0.1A) of current. A USB charger, on the other hand, is not so picky.

No Soldering Required

One notable omission from this particular configuration is any sort of soldering tool. You could certainly add a small soldering iron, some solder, and a few more hand tools. You'll either need to boot out some of the bulkier items, such as the PC or the full-sized multimeter, or get a slightly bigger carrying case.

The Ever-Shrinking Computing Device

One option that will be gaining momentum in the very near future is the use of tablet computers or smart phones instead of the physically larger laptops or netbooks. They already have a lot of the computing capabilities of reasonably powered portable computers. Also, you can look forward to innovative hardware and software interfaces that will make them more useful in the lab as programmable test and measurement devices. While name-brand products still command a premium price, many no-name or OEM (original equipment manufacturer) products are being sold for a fraction of the cost.

Electrical Experiments You Can Perform

Here are some simple electronic experiments that you can perform using only the tools and components found in the compact executive mobile electronics lab. Start out with something simple and incrementally build on your success. Don't try the most complex problems first. What you want is to build up both your knowledge *and* your confidence at the same time.

Lighten Up, 20th Century Style

Here's an easy experiment that will teach you several interesting things about electronics, believe it or not. You can light up a typical LED using only a lithium coin cell. It needs to be a lithium coin cell that produces at least three volts (marked "3V"), not one of the smaller batteries that only supplies 1.5V, such as a watch or hearing-aid battery, as these will not produce enough voltage to light up a typical LED.

Slide the coin cell between the leads of your LED. There are four possible outcomes, three of which are "nothing interesting happens." Let's start with the least likely possibility, in which the LED lights up. Sometimes you just get lucky.

First Possible Outcome: The LED Lights Up When Attached to the Coin Cell

Figure 4-3 shows the results of this outcome. Notice that the longer lead of the sLED (the positive, or anode, lead) is touching the side of the coin cell marked with a +. No current-limiting resistor is required in this circuit because the battery's internal resistance is sufficient to keep from blowing up the LED. It just happens to work out that way, this time. Most other times, you will need to add a current-limiting resistor to the circuit.

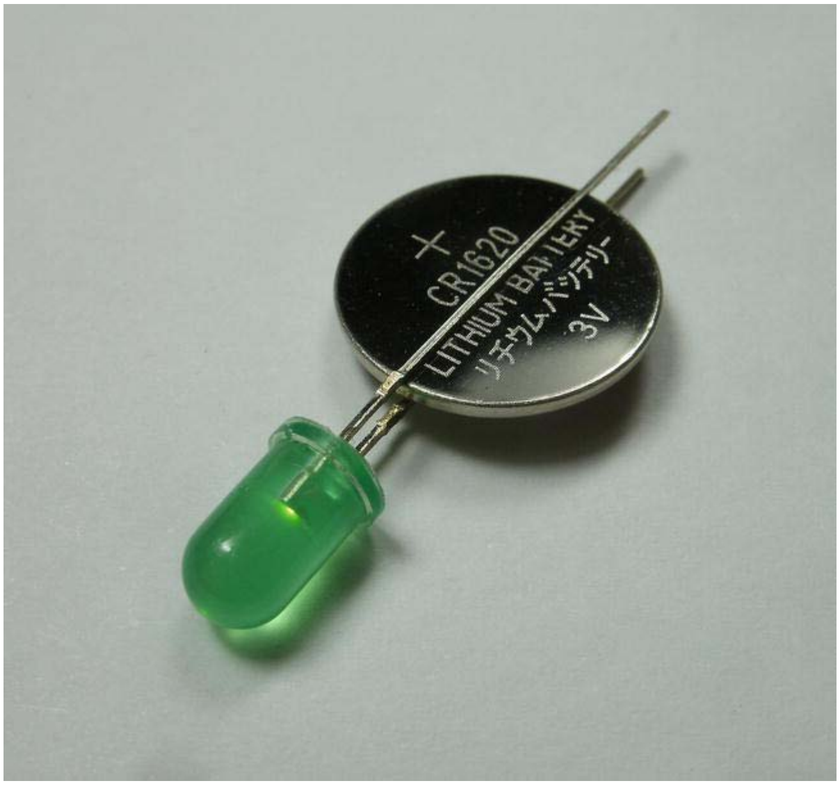

Figure 4-3. *A typical LED lights up when attached to a lithium coin cell. No extra wires needed!*

The other three scenarios? Let's look at them one at a time.

Second Possible Outcome: Nothing Interesting Happens

Why? Because the LED has been installed backward. Alternatively, the coin cell has been installed backward. Remember from Chapter 2 that LEDs are light-emitting diodes, and diodes only carry current in one direction. Therefore, when connected backward, they block the flow of electric current, preventing anything interesting from happening in the rest of the circuit.

Typical lithium coin cells have one broad, flat side that usually has something printed or etched on it, almost always with a + to indicate the positive polarity of the cell. The other side is often textured or rippled in some way to help make better electrical contact when installed in a device. This other side is the negative terminal of the battery.

The LED, being a diode, is *polarized*. It will only work when installed in the proper orientation inside the circuit. Most LEDs will offer one or more ways of easily identifying which lead is which (see Figure 4-4).

- Brand-new LEDs will have extended leads about an inch or so long. One lead will be slightly longer than the other one. The longer lead is the *anode*, or positive lead. The shorter lead is the *cathode*, or negative lead. This information is lost when the leads of the LED are trimmed—for example, after being installed in a PCB.
- Another industry standard for identifying the polarity of an LED is to look at the small ridge or shoulder at the base of the LED body. There will be a flat portion molded into the otherwise circular outline. This flat spot marks the cathode, or negative terminal of the LED.

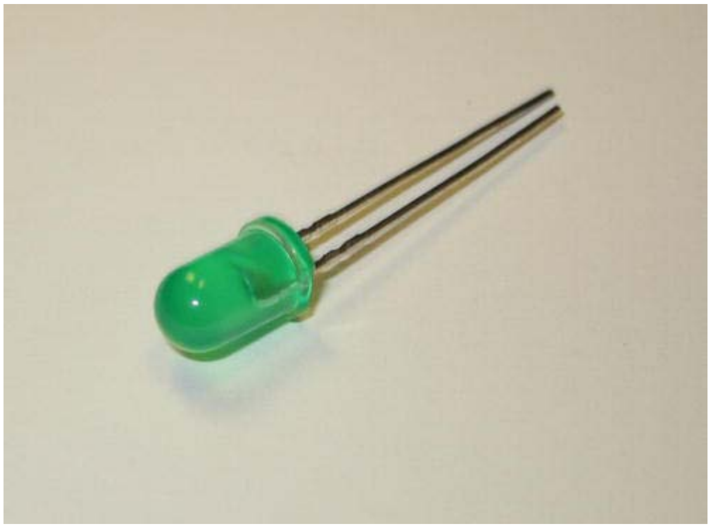

Figure 4-4. *Typical LEDs in the 3mm, 5mm, 8mm, and 10mm sizes almost always have two polarity indicators. The first is that the longer lead is the positive lead, or anode (top). The second is that the flat side on the base indicates the negative lead, or cathode (bottom).*

On really tiny LEDs, there may be an even tinier mark printed or etched onto the LED to indicate the cathode. You will sometimes need a magnifying glass just to see it!

The quick and easy fix for this problem is to swap the direction of either the LED or the battery. If this still produces no good results, we have to consider some more possibilities.

Third Possible Outcome: Nothing Interesting Happens

This time, after establishing that the LED has been properly oriented with respect to the polarity of the battery, we must entertain the possibility that the battery has been previously depleted. Coin cells have a limited amount of energy stored within them, and this can be easily drained in normal use over time.

A quick way to measure a battery's electrical goodness is to use a voltmeter. See Figure 4-5.

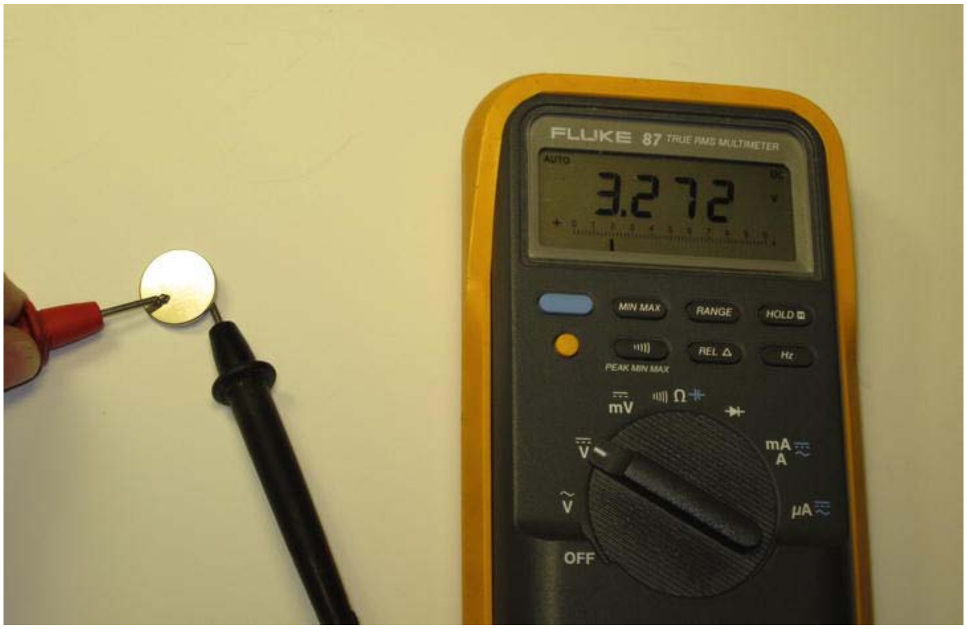

Figure 4-5. *Measuring a coin cell's electrical goodness using a digital multimeter. This one is chock-full of goodness!*

If you have a multimeter, be sure to set it to read DC voltage, in the appropriate voltage range (unless you have a fancy "autoranging" meter). Touch one of the meter's probes to one side of the battery and the other probe to the other side of the battery, and see how much voltage is measured. Anything below 3V should be considered depleted and of no further use, as far as electronic-type projects are concerned. The batteries are shiny and round, so they may have a second career in some sort of art project, but otherwise they should be disposed of in a responsible manner.

Also, perhaps because of their very shininess and roundness, they like to be eaten by small children. Don't let this happen! Take whatever steps are necessary to prevent small children from handling or even touching your electronic tools and components. Use extra care when your lab is mobilized.

■ **Caution** No unsupervised children, no matter how adorable, are allowed in the lab!

Fourth Possible Outcome: Nothing Interesting Happens Because the LED Is Defective

This can happen when an LED is subjected to too much current and the tiny, tiny wires within the plastic body overheat and fail. It can also occur when the LED has been exposed to the elements and moisture has penetrated the casing, allowing the internal connections to oxidize. This scenario, while within the realm of possibility, is certainly the least likely of the four, unless you are habitually cruel to your LEDs. It's also easy to test, if you have a known-good lithium coin cell handy.

Where's the Resistor?

You might be wondering where the current-limiting resistor can be found in this circuit. It's in there, all right, but it's hidden within the battery, so to speak. Coin cells are designed to provide tiny amounts of current over a long period of time. You're not going to arc-weld or start your car with one. As such, they have a much higher *internal resistance* than most any other kind of battery. We use this "feature" to our advantage when testing LEDs using just a coin cell, as the maximum amount of current that will normally flow out of the battery is below that necessary to destroy the LED.

A Slightly More Permanent Circuit

You can build a more permanent circuit for the coin cell and LED using a solderless breadboard. If you recall from the "Solderless Breadboards" section of Chapter 3 (see Figure 3-6), a solderless breadboard is an array of spring-loaded tie points, which serve to both physically hold electronic components and provide electrical connections between certain points. Let's build an LED circuit on a solderless breadboard using an LED, a coin cell, a resistor, a jumper wire, and a custom battery holder (see Figure 4-6). Later on, in Figure 4-7, you can see details of the custom battery holder.

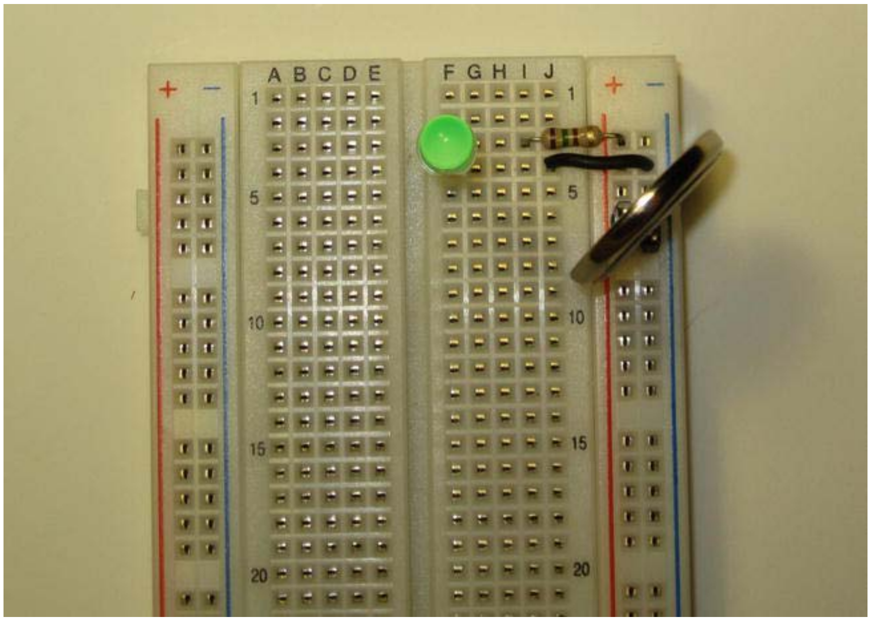

Figure 4-6. *A slightly more permanent LED circuit, built on a solderless breadboard. The tie points on the breadboard are labeled left to right with the letters A–E and F–J for the columns. The rows are numbered starting at 1 at the top. The LED's anode, for example, is installed in tie point F3. The power rails on either side are simply marked with + for positive and – for negative.*

Referring again to Figure 4-6, you can see the overall layout of the LEDcircuit. The power rails along both sides of the breadboard are just longer versions of the electrical connections that connect each of the groups of five tie points together. In this example, the right-hand power rails are used, but the left-hand side is not.

The schematic diagram of this circuit shows how simple it is, at least in principle (see Figure 4-7). Can you see the circle?

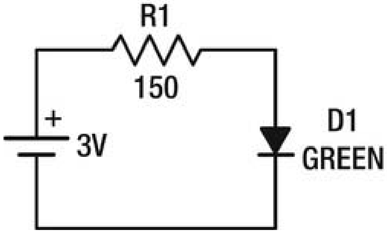

Figure 4-7. *The schematic diagram of the LED circuit as built on the solderless breadboard. Schematics are intentionally simpler than reality (most of the time) in order to help convey the very basic idea of how things are to be wired and the relationships between the components, without bogging you down in all the inevitable details that are involved.*

The Battery

The symbol for the battery actually looks like the battery in this case. The same symbol is used for most single-cell batteries, even when they are of completely different shapes or proportions. Multicell batteries are usually represented with repeated stacks of long and short lines.

The battery symbol's long line represents the positive terminal of the battery, and the short line represents the negative terminal. Again, it's just a coincidence that the wide side of the lithium coin cell is the positive terminal. The addition of the + sign in the schematic is redundant, but you'd be surprised how many people accidentally reverse this symbol. Better safe than sorry when it comes to electricity!

The Resistor

The battery's positive terminal is connected via a wire (drawn as a simple, solid line) to one side of the resistor. The schematic symbol for the resistor is a zigzag line.

The resistor is labeled "R1" (resistor 1) in this schematic, even though it's the only resistor and could just as well have been labeled "RESISTOR #1," or just plain "resistor." The numeric value underneath the resistor is the component's primary electrical characteristic, at least as far as we're concerned when looking at the schematic. In this case, it indicates a resistor whose *value* is 150Ω. In this simple example, the value of the resistor is quite flexible. A larger value, such as 1KΩ (1,000 ohms) or more, would also work, except that the LED would not shine as brightly. This is because *more* resistance results in *less* current flow, and it's the current flow that makes the LED shine. More current, more shiny.

We haven't specified any of the other important characteristics of the resistor, such as the tolerance or the power-handling capacity. Because this is such a low-power circuit, almost anything will do. When you start to work with really big LEDs that require correspondingly really big power supplies, this becomes a much more critical issue. For now, we start simply.

The LED

The other side of the resistor is connected, again, with a simple, solid line, to the anode of the LED, whose *reference designator* in this schematic is D1 (for diode 1). You could have called it "LED," and future generations would probably have understood your intentions, which should be your overriding goal. Its characteristic "value" in this case is "green." Of course, any color will do, as long as you have the eyes to see it.

The LED's schematic diagram effectively communicates its polarized nature. The arrow indicates the flow of *conventional current*. The flat bar indicates the cathode, or negative terminal of the LED. This corresponds to the flat side of the LED body in this example. The cathode is connected via a wire back to the negative terminal of the battery, completing the circuit.

Can you see the circle yet? Hint: It's disguised as a rectangle. The important point is that it forms a complete loop.

Building the Actual Circuit

Comparing the schematic to the photograph of the completed circuit might not immediately click in your head. Take a look at Figure 4-8 for a slightly different angle, where you can see some more details, such as how the battery is being mounted.

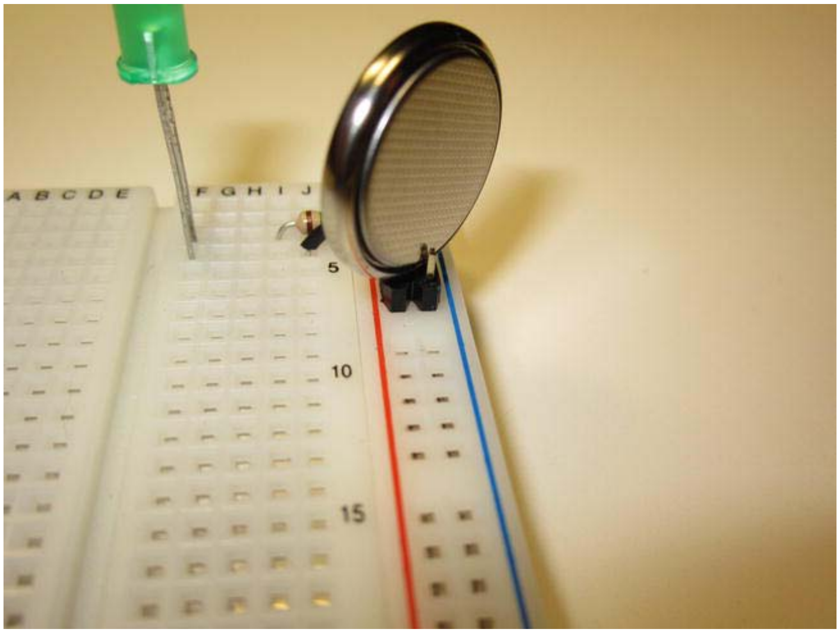

Figure 4-8. *Another view of the LED circuit as built on a solderless breadboard. The battery holder is a modified pin block that has had a couple of its pins removed, allowing the coin cell to fit in on the diagonal. A thinner coin cell could have fit without your having to remove the corner pins. A proper coin cell holder is inexpensive and would provide a much more robust solution to the problem.*

Looking at Figure 4-8, you can see how the positive side of the battery is making physical and electrical contact with the gold-plated pin of the pin block. The remainder of the pin emerges underneath the black plastic molding to make contact with the spring-loaded tie point within the body of the solderless breadboard, which in this case is the negative power rail. The positive terminal, in a similar fashion, makes contact with the positive power rail.

By attaching the battery directly to the right-side power rails, you ensure that your circuits are only a short jumper away from power, no matter where along the breadboard you decide to build them. Handy!

Note that the left-side and right-side power rails are not connected at this time. If you would like them to be, you can easily attach them using jumper wires. See Figure 4-9.

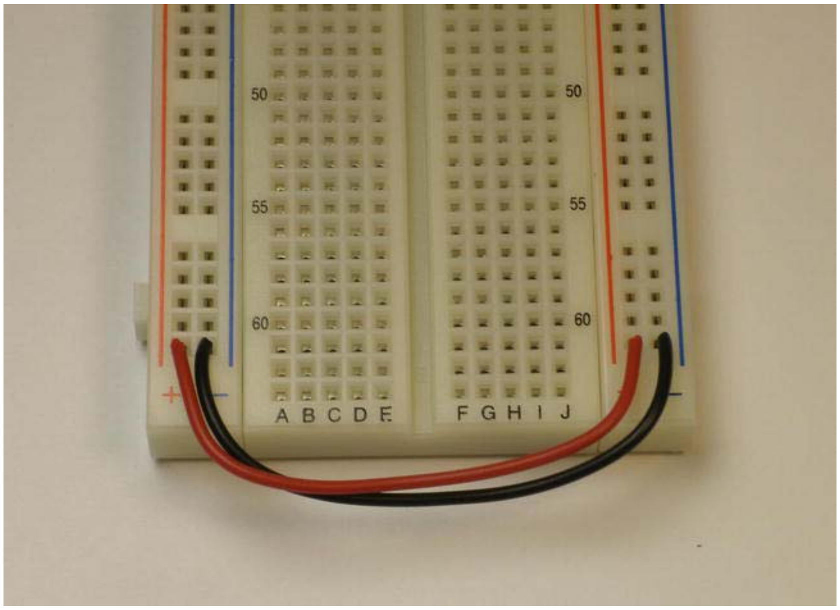

Figure 4-9. *Attaching the left- and right-side power rails using jumpers. By default, the power rails are not connected. There are some situations where you need them to not be connected, such as when using multiple supply voltages within a single circuit. These things happen.*

Now, referring *all the way back* to Figure 4-6 (sorry about that), you should be able to see how the positive terminal of the battery is indeed connected to one side of the resistor. The other side of the resistor, then, makes an electrical connection with the anode of the LED (the non-flat side, when seen from above). The cathode is connected to the negative power rail using a short jumper wire. The negative power rail then completes the circuit back to the battery. The circle is complete.

Note that the resistor could have been placed either before the LED in the circuit (as shown in the photographs) or after the LED. The only important points are that it is in the same circuit, or circle, and that the current must flow through *both* the LED and the resistor. It really doesn't matter which one comes first, especially in this simple circuit. It will only make a difference when you start to do fancy stuff, like adding switches and such to complicate things.

If you're using a different kind of solderless breadboard (and there are a few distinct varieties), you may or may not have power rails available. If so, fret not. You can still make point-to-point connections using more jumper wires. Now you're starting to understand why you were assigned to cut and strip so many jumper wires back in Chapter 2! They should start to be coming in handy just about now.

Multiple Uses for the LED

Now we've moved from the 19th century incandescent bulb to the 20th century LED. What kind of lighting devices will we encounter in the centuries ahead?

One of the nice features of LEDs is that they don't typically require a great deal of power to operate. This is certainly true of indicator-style LEDs, whose only purpose is to be seen as either on or off. We can leave this LED installed right where it is and use it as a handy power-on indicator for future projects. You might be surprised how many heads get scratched in puzzlement, wondering why a newly wired prototype isn't working, only to find out after exhaustive investigation that the power wasn't turned on! Then again, you might not be surprised by this.

Now let's take this perfectly working circuit and break it—on purpose. If we interrupt the circuit, for example, by removing any of the components or their connections, the circuit stops working. To that end, remove the short jumper wire and replace it with two longer jumper wires. Plug in each of the new, longer jumper wires where the two ends of the short jumper were installed. See Figure 4-10.

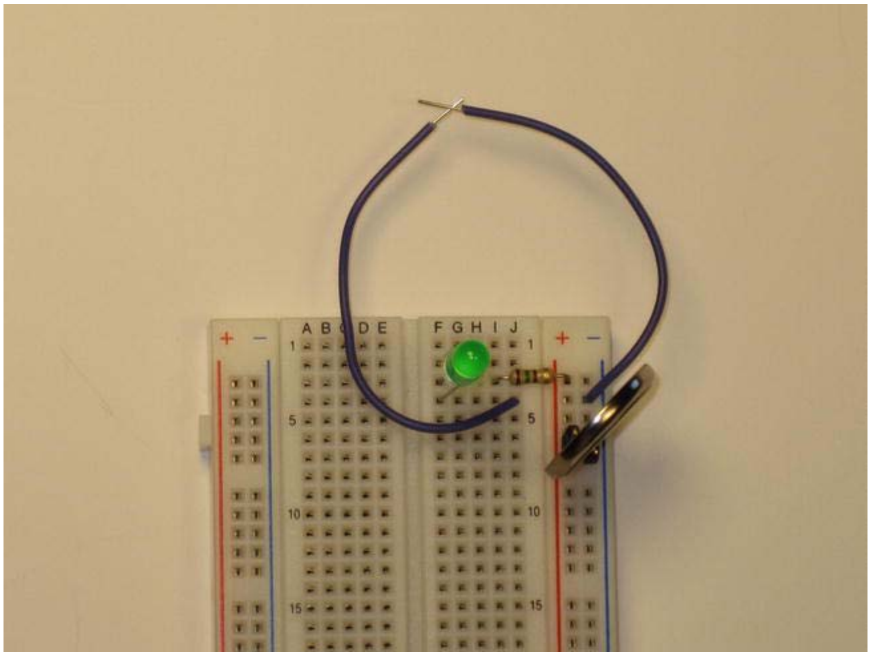

Figure 4-10. *By intentionally breaking the circuit, we stop the flow of electricity through the components, extinguishing the LED. By reconnecting the longer jumpers (or just touching them together), we reestablish the circuit, and the LED should light back up again. We can use this mechanism to test whether there is continuity in another circuit.*

Now, by touching the two probe wires together, you recomplete the circuit, and the LED should glow in happiness. This is exactly how a switch works. You can use this basic idea to build a continuity tester that lights up the LED when the current flows in the circuit. This is very handy for testing cables and wiring in a more complex assembly.

More electronic experiments, including how to both calculate and measure what's happening in a circuit, can be found in Chapter 5.

The Special Project Portable Laboratory

When you move past the simple joys of lighting LEDs and wiring up tutorial circuits, where do you go and what do you build? How about a robot?

Doug Paradis is an award-winning writer and robot builder. When Doug wants to work on one of his robots, he packs up a small kit of tools and heads over the local robot club, which in Doug's case is the Dallas Personal Robotics Group (DPRG). They meet every Tuesday at the Dallas Makerspace for an informal evening of robot building and socializing. See Figure 4-11.

Figure 4-11. *Doug Paradis of DPRG, arriving at the Dallas Makerspace for an evening of robot building. Doug uses the club's laser cutter and other specialized tools to give his robots that winning edge. He brings his own set of hand-picked tools along with him.*

Figure 4-12 shows what Doug brings with him to work on his robots.

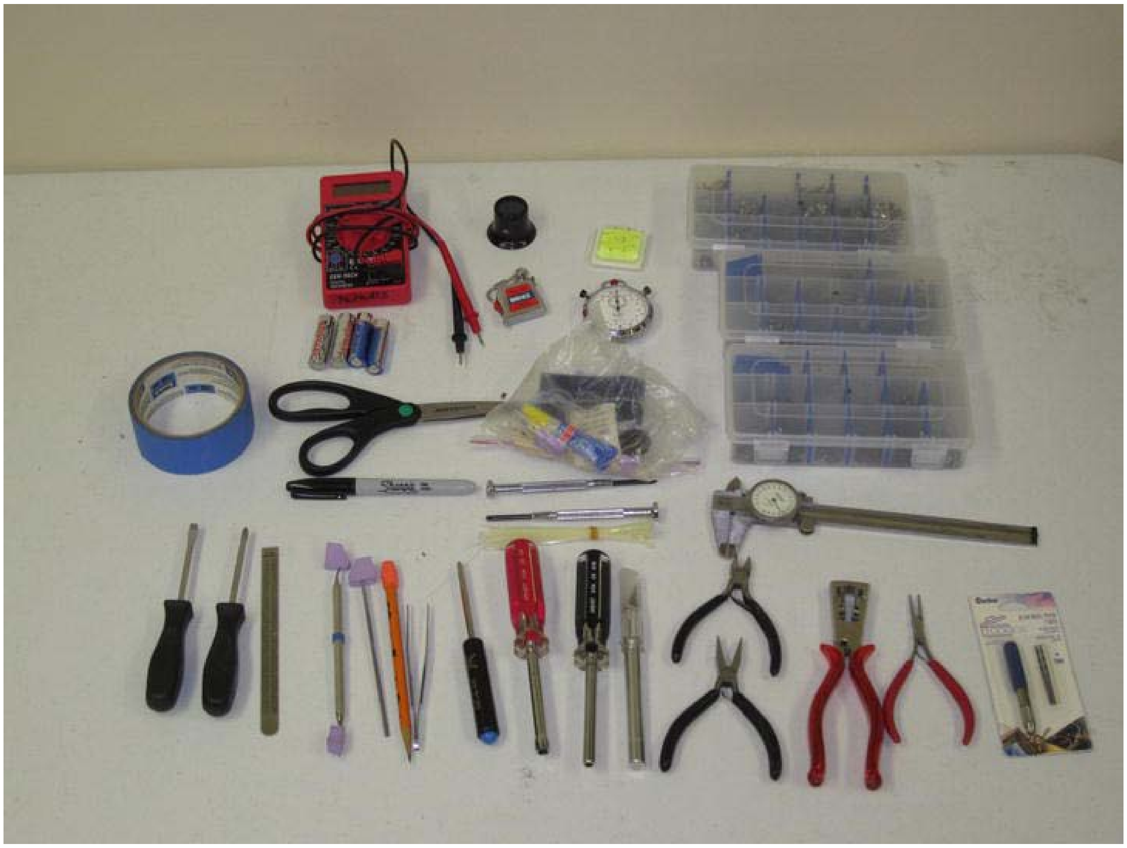

Figure 4-12. *The contents of Doug's portable workshop. A variety of hand tools, precison measurement devices, and fasteners (including tape, tie wraps, and super glue) are easily stored in a small tool bag for easy transport. The comparmentalized storage containers at the upper right contain a large assortment of screws, washers, and nuts.*

Doug has learned the lesson of travelling light. The electronic portions of his robotic creations evolve slowly over time, but the mechanical aspects get a lot of tweaking along the way.

Not everyone has perfected the art of light travelling. See Figure 4-13.

DRAIN OUTLET

Figure 4-13. *A sort-of-portable robotics lab, with lots of tools and spare parts. One way of looking at this bounteous splendor is "three trips to the car."*

This caravan of tools, parts, and miscellaneous supplies forms the entourage of Kyle the Robot. Kyle is very much a work-in-progress. The basic framework is made of hand-tooled aluminum. The large wheels use O-rings for tires and are driven by 24V gear motors. The 24V supply is made up of two 12V, 5Ah rechargeable lead-acid batteries. Ah stands for amp hours, and represents the capacity of the battery. In this case, it can deliver an amp of current for five hours, or five amps for one hour. The custom PCB holds the electronics for the drive motors and will be controlled by a separate microcontroller that has yet to be added. See Figure 4-14.

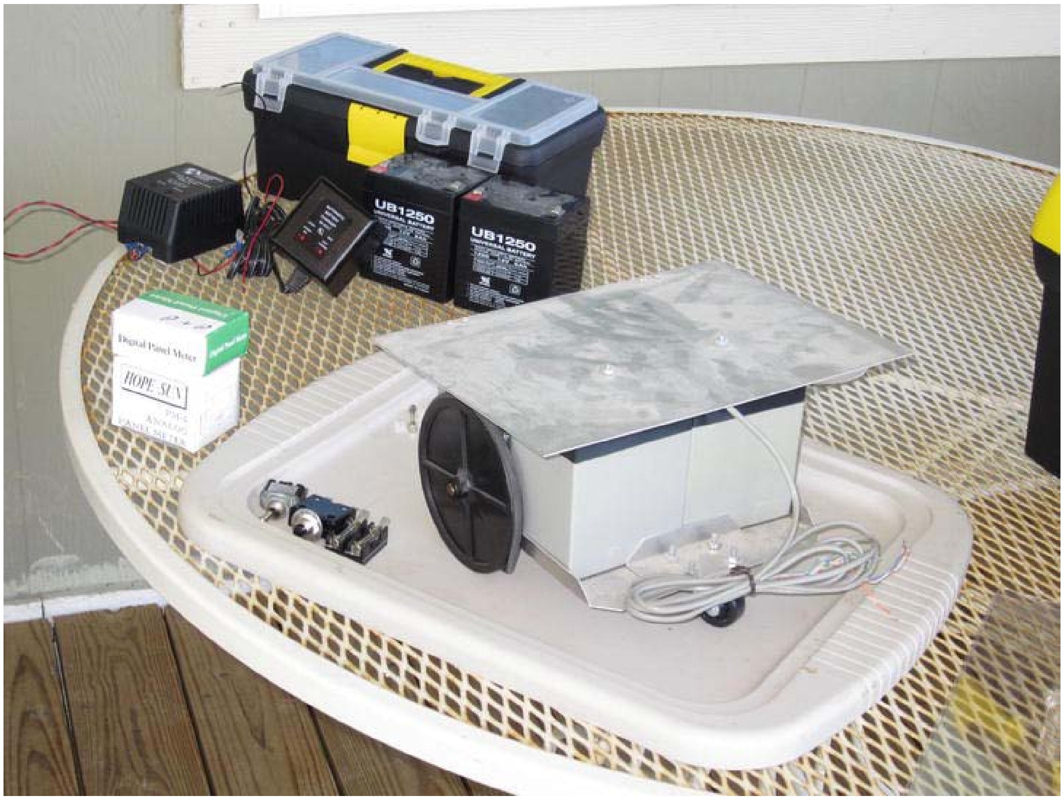

Figure 4-14. *Kyle the Robot is a work-in-progress who requires a wide range of specialized tools and components. Some new components are being fitted, including dedicated battery voltmeters, fuses, a circuit breaker, and a master power switch. Once the sharp corners are filed off, Kyle will be able to roll around the house, looking for chores to do.*

Summary

As we come to the end of this chapter, you should have some good ideas bouncing around in your head about how you want to pack your mobile electronics lab, what tools and components you *absolutely* must have with you, and how you want them organized. The best way to find out for sure what works for you is to make a few field trips, paying careful attention to what was actually used compared to what you anticipated needing. This will help you whittle down your cargo to the bare essentials, which is what you want. Remember, travel light. Excess baggage is a purely optional curse.

In the next chapter, we'll look at some suggestions for organizing and working in a more permanent setting. This could be a domain as small as the corner of a desk or as large as a warehouse. You will soon discover, if you haven't already, that this hobby, like many other projects, will quickly expand to fill the space available. Try to keep in mind that *you* are the one that ultimately makes the decisions. Effort spent planning ahead of time will help keep the sprawl to a minimum, while helping you stay focused on what you wanted to be able to do in the first place.

CHAPTER 5

The Cozy Corner Lab

Chapter 4 focused on the needs of the travelling electronic adept. Here in Chapter 5 you can come home to a lab that is always ready and waiting for some electrical fun and doodling. Nice!

You don't need an entire garage or extra office for a functional laboratory for your electronic endeavors. It's certainly nice if these areas are available to you, but you might be surprised at how much you can do in just a small space, if you set it up properly and maintain it with determination.

The first part of this chapter will deal with the task of rehabilitating a small, dedicated work area that has succumbed to years of poor organizational habits. These things happen. The second half of the chapter explores the possibilities of starting afresh with a nice, clean space—either something brand new or just newly cleaned. Some more projects help break in your new work area, as well as familiarize you with more interesting information about some of the basic electronic components reviewed in Chapter 3.

The Rehabilitation of a Small but Useful Area

Let's perform a little magic, in the form of a complete makeover for a small workstation. Have a look at Figure 5-1. Don't let this happen to you!

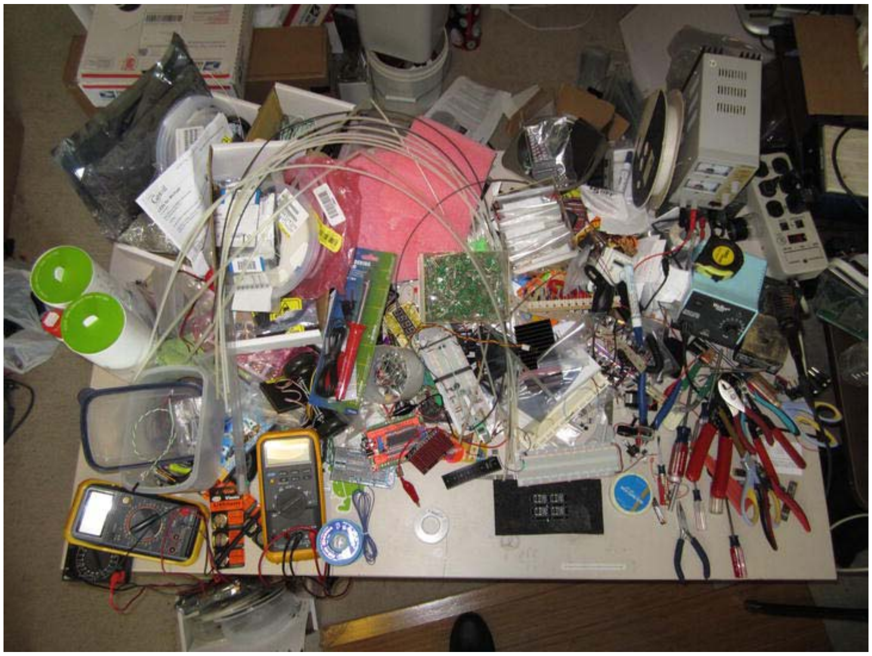

Figure 5-1. *A workstation that has been in continuous use for several years. Not much planning went into the original organization of this work area. Such a space can often accumulate flotsam and jetsam that might have at one time seemed appropriate, but now just competes for precious space. The effective work area is down to only a few dozen square inches! This limits the scope of projects that can be comfortably undertaken. It also causes unnecessary delays in rooting around for tools and parts. It's not pretty and it's not safe.*

Where to begin? Like any other project, this big project is made up of many smaller projects, none of which are especially difficult, if approached in the right order and with the right attitude.

If you're starting from scratch with your electronics lab and don't want to wade through this mess, jump on over to the "Adding a Power Source" section of this chapter.

Science to the Rescue!

Let's use the science of *taxonomy* to help break down this task into smaller, more manageable jobs. Taxonomy comes from the Greek words for "arrangement" (*taxis*) and "method" (*nomia*). That sound like *exactly* what is needed here! An "arrangement method" would assist in building up a plan for attacking this project.

■ **Tip** Method plus rigor equals success! Have a plan, pursue it with determination, and you will succeed.

Let's break down the overall category of "big mess on table" into two major *taxa* (singular: *taxon*), or categories. Ideally, what we will find here are *tools* and *components*. There will be other bits and pieces that don't easily fit into one category or the other (or will maybe fit into both), but it gives us a place to start.

Organization

Divide and *conquer*! These are the two words that will help us through the ordeal of rehabilitating this workspace. We'll get to do a lot of dividing here, and end up as conquerors. The spoils will be a useful (and tidy!) work area that will encourage you to begin new electronic adventures, instead of scaring small children.

The Great Divide

It would be possible to reorganize this workspace *in situ* (i.e., in place), but it would require a great deal of inner peace and lots of patience. Lacking these, a complete reboot is in order. Let's clear off the entire table and start afresh. This will provide the excellent opportunity to arrange things in exactly the best possible order.

First, get two big containers. These can be cardboard boxes or plastic storage bins—it really doesn't matter. We'll just be using them as temporary staging areas as we make like an embryo (and divide and divide again). See Figure 5-2.

Figure 5-2. *Two large cardboard boxes are enlisted to help with the initial organization of the contents of the tabletop. One will be for tools and the other one for components. Further subdivisions will be made later.*

Now, the first thing you will want to do is remove any items that really don't belong in the lab at all, such as coffee cups, personal effects, and any obvious trash. Everything that is about to go into one of the two boxes should eventually find its way back to the work area, so now is the time to omit the things just don't belong here.

Is it a tool? Is it a component? How do we tell? Ask five people and you will get five different answers. For the purposes of this little endeavor, a *component* is something that will become a part of a project, while a *tool* is something that is used to modify or manipulate a component. Another way of looking at it is that components get used up quickly, while tools get used up slowly. If you build enough projects, you will be lucky enough to see exceptions to these guidelines.

Here are some examples. Wire is a component. Wire cutters are tools. That was easy. A flashlight is a "visual inspection aid," so it goes under the category of tools. Batteries for the flashlight are consumable, so they are components. There's no real need to be ultra-precise at this point. Use your best judgment and make up policies that make sense to *you*.

It might help, from a physical logistics standpoint, to use smaller subcontainers to hold the really small bits as you're lumping them into the two broad categories. Folding cardboard bin boxes are handy both on the tabletop as well as on shelves, if you have them. The cardboard bin boxes shown in Figure 5-3 are from Uline (`http://uline.com`), part numbers S-16268 (4" wide) and S-16269 (6" wide).

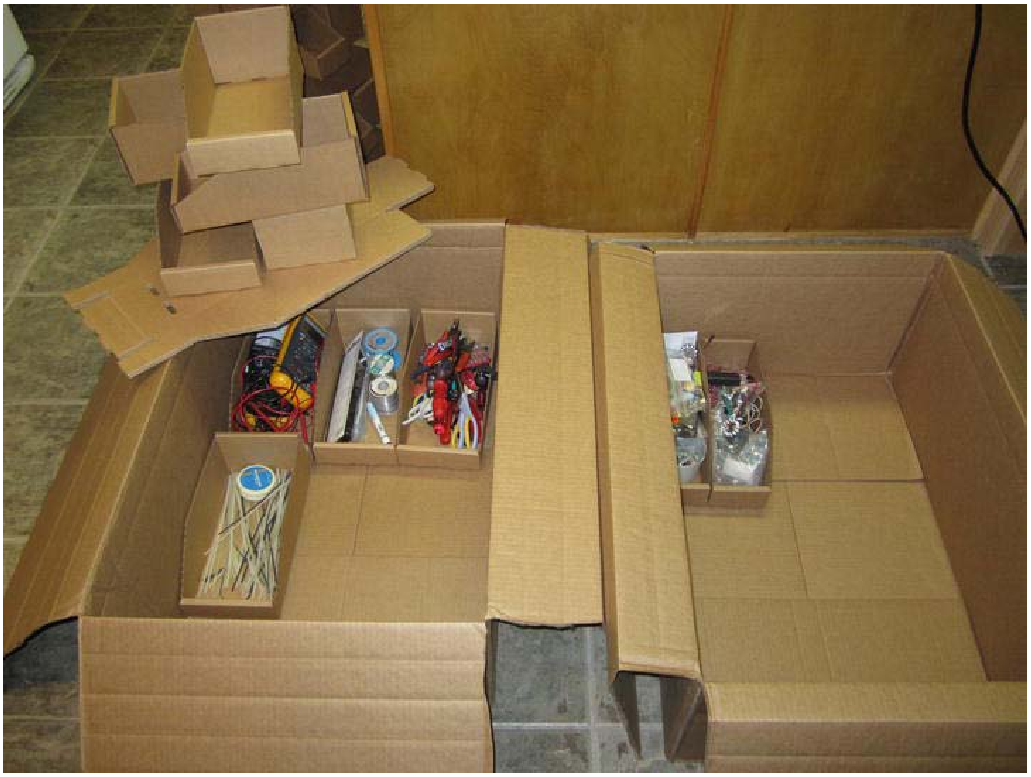

Figure 5-3. *Folding cardboard bin boxes help contain some of the smaller bits and pieces during the Big Sort. You can use whatever happens to be available. Some preliminary subsubdivision is going on here. For the most part, the tools are on the left and the components are on the right.*

After everything is removed from the table and the cardboard boxes are nearly full, the table itself should get a good washing. Behold the *tabula rasa* (blank slate) depicted in Figure 5-4.

Figure 5-4. *The tabletop has been cleared of tools, components, and various bits of debris. A good washing with soap and water completes the cleansing. A burned spot (the result of a poorly supervised thermal test) is covered with a small sticker (see inset).*

Adding a Power Source

Here's where a bit of planning will save you a lot of headaches in the future. What you never seem to have enough of in any laboratory, especially electronics labs, are lots of conveniently located power outlets. Let's take care of that issue right away, before building anything else onto the tabletop. See Figure 5-5.

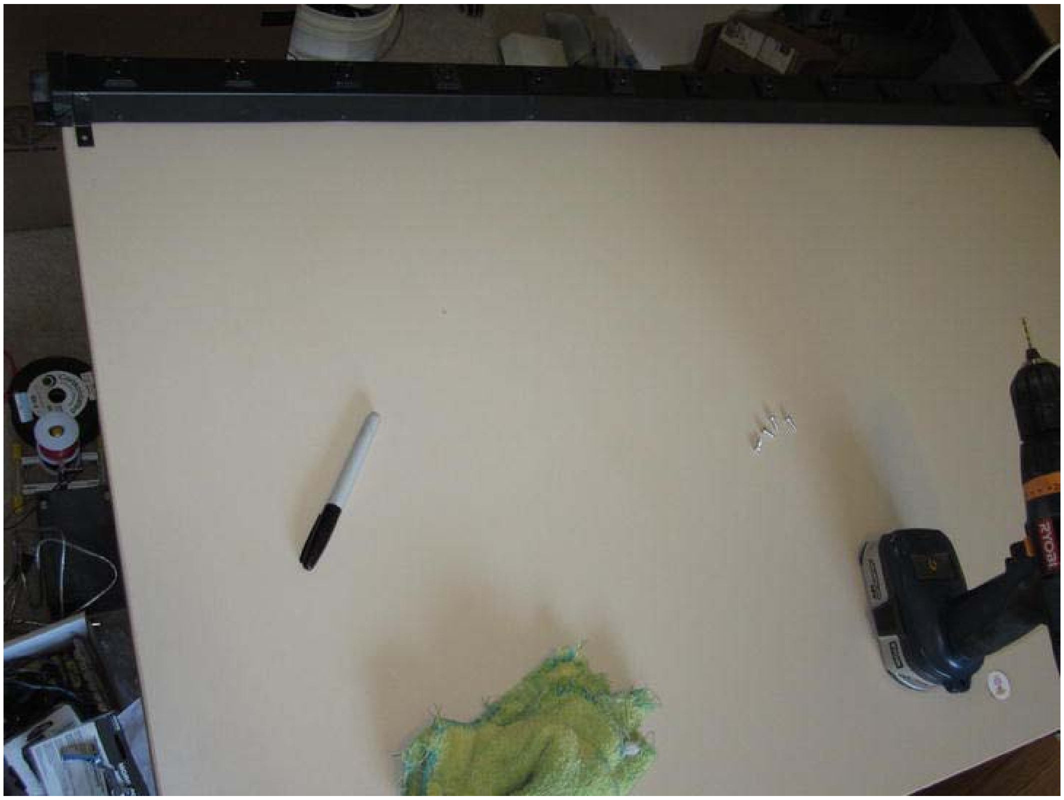

Figure 5-5. *A 12-outlet power strip is attached to the back edge of the tabletop. This will provide a safe and convenient source of power for all those gadgets you'll eventually want on your bench. Tools needed for the job include a drill motor to drill out some small pilot holes for the brackets, a small drill bit, and a felt-tipped marker. Before drilling the pilot holes, align the brackets where you want them to go and use the marker to put a dot where the pilot hole should be drilled. Make sure the power strip is not plugged in when you're drilling right next to it!*

Having lots of extra power outlets is a nice bonus, but is not absolutely required. What you want to avoid, however, is excessive daisy-chaining of power strips and extension cords. This is a fire hazard. Remember from Chapter 3 that every conductor has a certain amount of resistance. This resistance turns electricity into heat. Enough heat makes things catch on fire.

If you do have the luxury of being able to mount a permanent power strip (which, technically, remains a temporary power tap as far as the manufacturer is concerned), then be sure to install it securely using the hardware provided. See Figure 5-6.

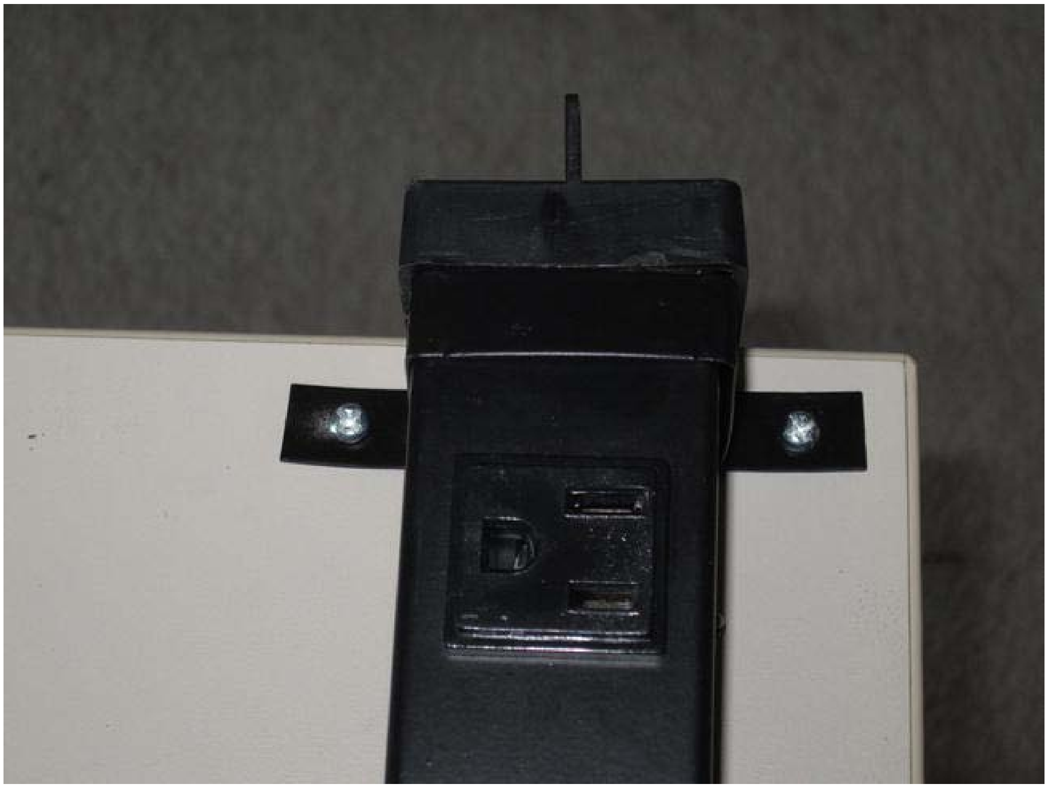

Figure 5-6. *The power strip is fastened securely to the table using the provided hardware. You do not want your power strip wiggling around, or worse yet, falling off the table, dragging all of your power tools with it!*

■ **Tip** Test every outlet of your power strip before permanently mounting it to your workbench.

A Clean, Well-Lighted Space

You can never have enough light, it seems. What passes for normal room illumination might be sufficient for such adventurous undertakings as finding your way to the door or looking for the remote control for the television. For anything more demanding, you're going to have to add some more lighting, and probably lots of it.

At a minimum, you are going to want some *task lighting*, which helps illuminate the immediate area of your focus. In this case, where a small table is being used as a workbench, an articulated fluorescent lamp with a built-in magnifier will do a great job, at least until more lighting can be strung up. See Figure 5-7.

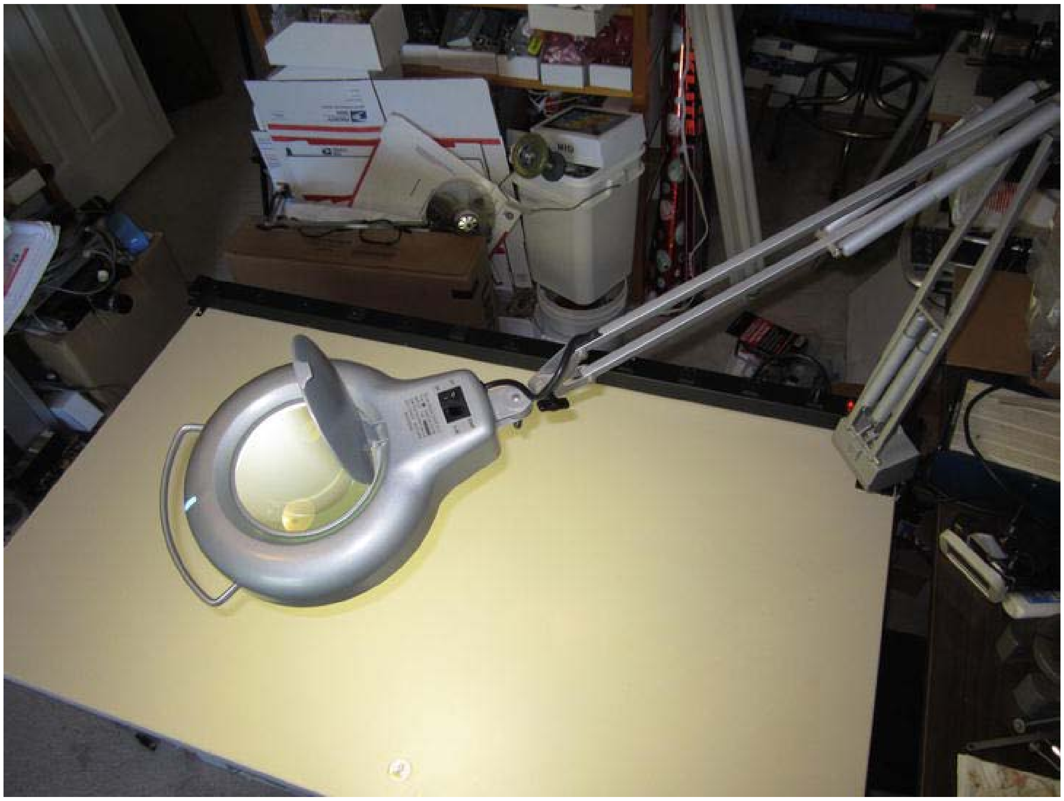

Figure 5-7. *You will need some task lighting for your bench. This desk lamp also includes a handy magnifier for really close work.*

Tools

Now we're ready to bring some of the tools back in to the lab. Here's where you get to decide which tools get a *permanent* home on the bench and which ones get put up and stored when not in use.

The permanently deployed tools on your workbench should be the ones that you find that you are using the most. The ones that get put up after every use are the ones that either only get used infrequently or pose some sort of hazard, such as sharp edges or hot surfaces. Again, you get to decide. It's your lab.

Tools for Soldering

An activity that is practiced frequently in many electronics labs is soldering. If you haven't learned how to solder yet, you are strongly encouraged to do so. Here's a handy setup for soldering, sitting in one corner of the newly refurbished workbench. See Figure 5-8.

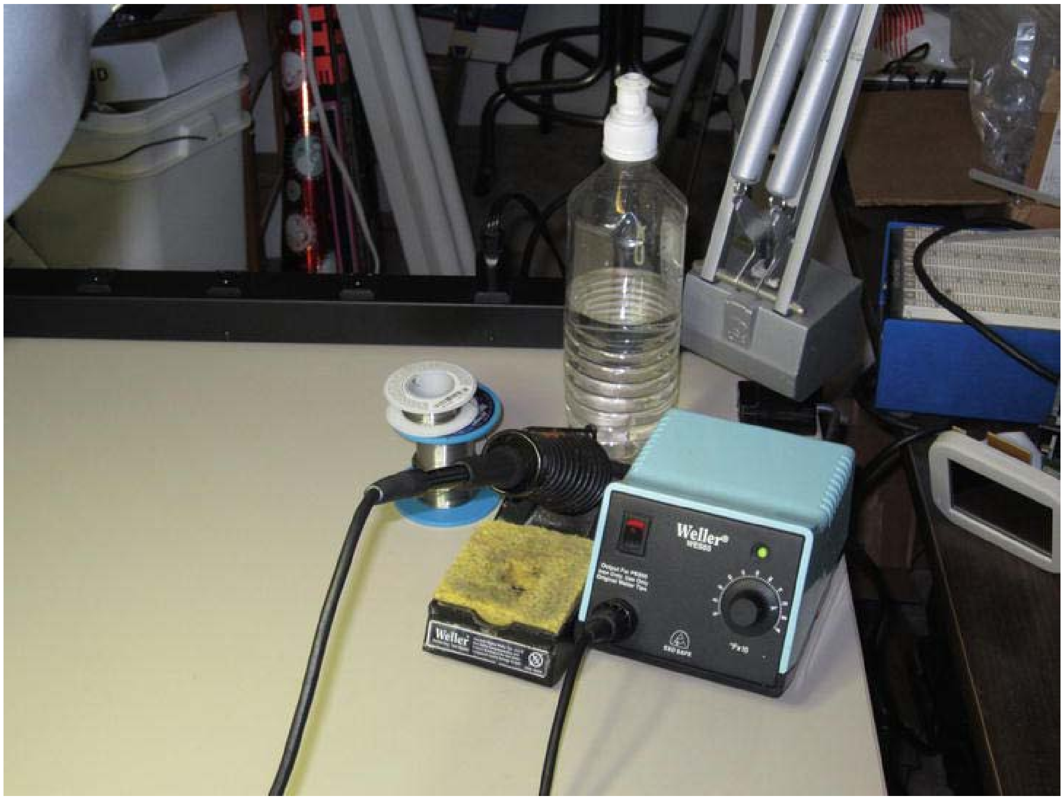

Figure 5-8. *A temperature-controlled soldering station with a stand for the soldering iron, a water bottle to wet the tip-wiping sponge, and some different kinds of solder. Some people prefer to have the soldering station nearer to the front edge of the table. Use whatever configuation makes the most sense in your lab.*

■ **Caution** Soldering irons are hot and can cause serious burns. Do not place a soldering iron where you might accidentally knock it off the work table or brush against it.

Hand Tools and Some Test Equipment

No matter what you work on at the bench, you will most likely discover the need for some common hand tools. When working with electronics, especially, you will often need wire cutters, wire strippers, and small pliers. These tools have been laid out conveniently (for a right-handed person) at the lower-right-hand side of the bench, while the less-often-used hand tools can reside in a box.

A regulated, adjustable bench power supply is also very handy when working with electronics of all kinds. A good, reliable multimeter can also claim a permanent home on the bench, as these will get a lot of use. See Figure 5-9.

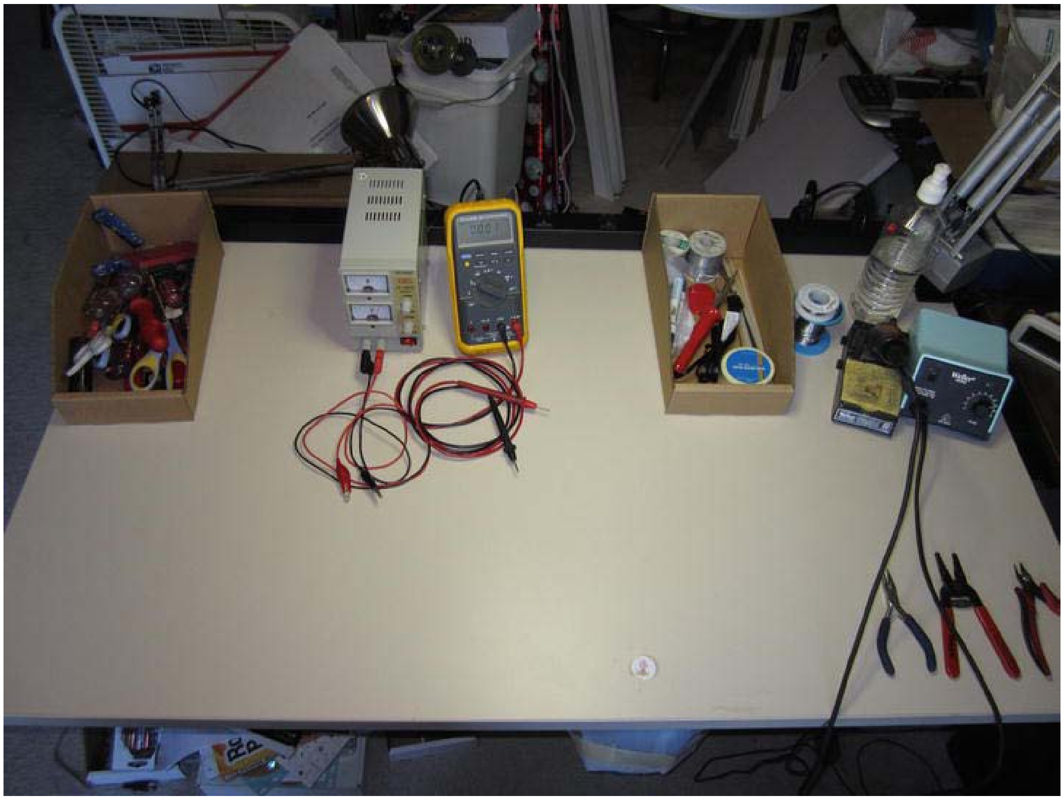

Figure 5-9. *A clean bench with just enough tools to have some fun. Extra soldering supplies, including more kinds of solder , extra soldering iron tips, soldering flux, and an emergency backup soldering iron, sit in a box adjacent to the soldering station. Hand tools and basic test-and-measurement equipment complete the set of "most-useful tools." There's still room for a few more, when you need them.*

Under the Desk

If your workbench has the luxury of drawers, by all means use them! Try to keep your under-desk environment as free of clutter as possible. It's not really a very convenient place to store unused tools or components. You're going to drop the occasional item from time to time, so having to dig out a bunch of junk just to find it is just going to be more frustrating for you. Keep it clean!

You *should* keep a medium-sized trash can under or near your work area, as you will inevitably generate some waste products in this hobby. By minimizing the amount of food-related garbage deposited in your laboratory trash receptacle, you can postpone the invasion of little critters looking for a free meal. Also, bits of wire and old components don't tend to stink after a few days, but a half-eaten sandwich might.

■ **Tip** Copper easily recycles at a premium price. Collect your copper wire snippings and cash them in someday.

A nice addition to the under-the-table accessories you might consider is some low-wattage strip lighting. LEDs are especially good for this. You will appreciate this more once you conduct your first search-and-rescue mission for that teensy-tiny component that dropped off the edge of your table and into the carpet.

Components

Most of our tools are now ready to go. The most frequently used tools are permanently set up, ready for action. The less frequently used tools are carefully stored where we can get to them when we need them, but are not in the way the rest of the time. This is a big improvement over the situation you saw back in Figure 5-1.

Now we need some components with which to work, using our scientifically arrayed tools. Have a look at Figure 5-10 to see one example of how to keep your parts organized yet within easy reach.

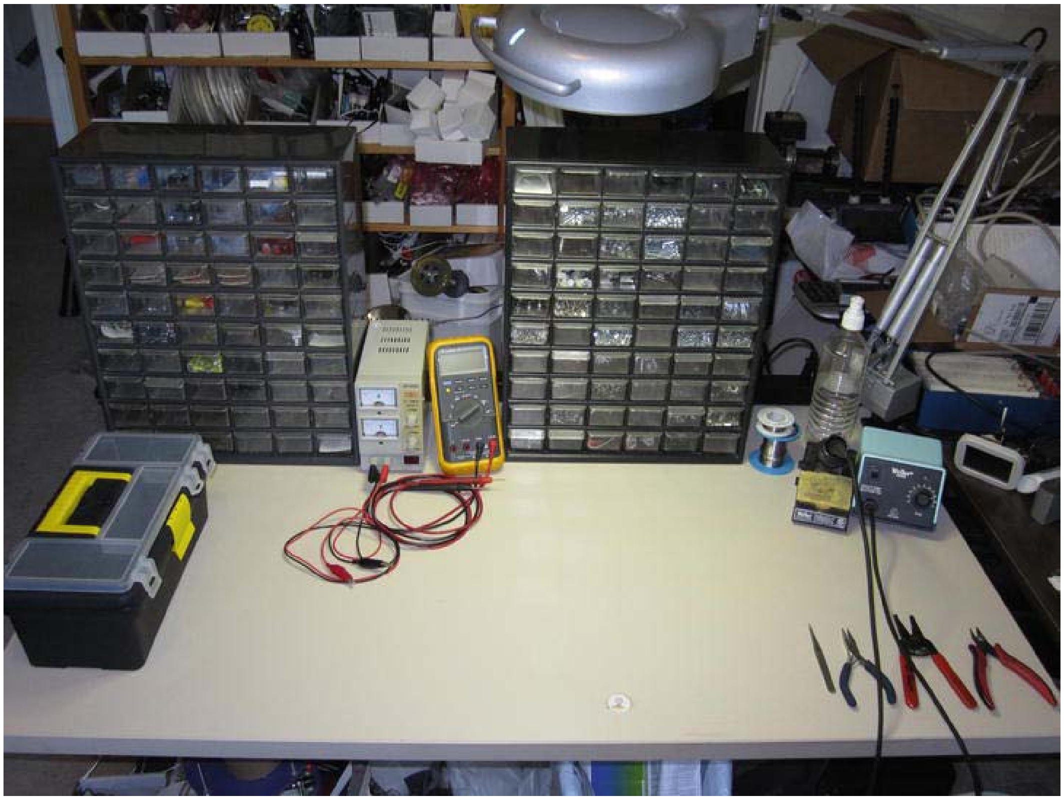

Figure 5-10. *Parts bins are used to organize components. This style of container is available in many sizes with different numbers and sizes of drawers. Even more parts are stored on a shelf behind the work area. The most commonly used hand tools are laid out on the right. The other hand tools are collected in a small toolbox on the left.*

Where to Go from Here

This work space could be further augmented with shelving, both above the work surface and underneath. More lighting wouldn't hurt, either. However, this is plenty to get started. Let's build some electronics projects on this squeaky-clean table and see if we can dirty it up a bit.

Projects

Let's look at a few smaller circuits you can easily build on your tidy, new work area using a solderless breadboard. Expanding on the LED-and-coin-cell experiments in Chapter 4, we'll add a few more components and hopefully start to see how they all interact.

There are any number of different types of solderless breadboards available today. A range of them is shown in Figure 5-11.

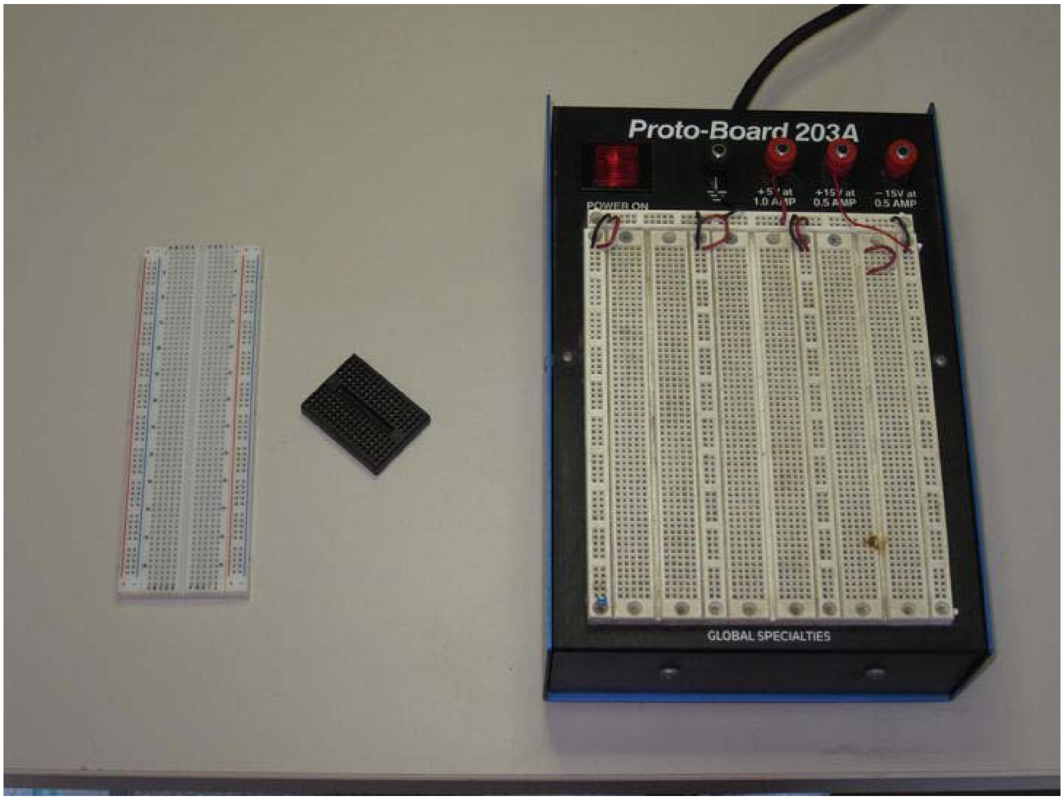

Figure 5-11. *Different sizes of solderless breadboards, all with the same lead pitch and internal connections. The small black unit in the center is perfect for little circuits with only a few small components. The white unit on the left has more tie points and offers power rails along each side. The deluxe, powered unit on the right adds even more space, as well as multiple regulated power supplies for all your tinkering requirements.*

Parallel and Series Circuits

The circuits presented so far have all illustrated the “loop” characteristic of electrical circuits. Electricity generally flows in a circle. It’s possible to have more than one path through a circuit, however. What happens then?

Like most good questions, the answer is, “It depends.” Let’s hook up *two* LEDs at once and try to figure out exactly what’s going on. Figure 5-12 shows how it’s put together. See Figure 5-13 for the schematic diagram.

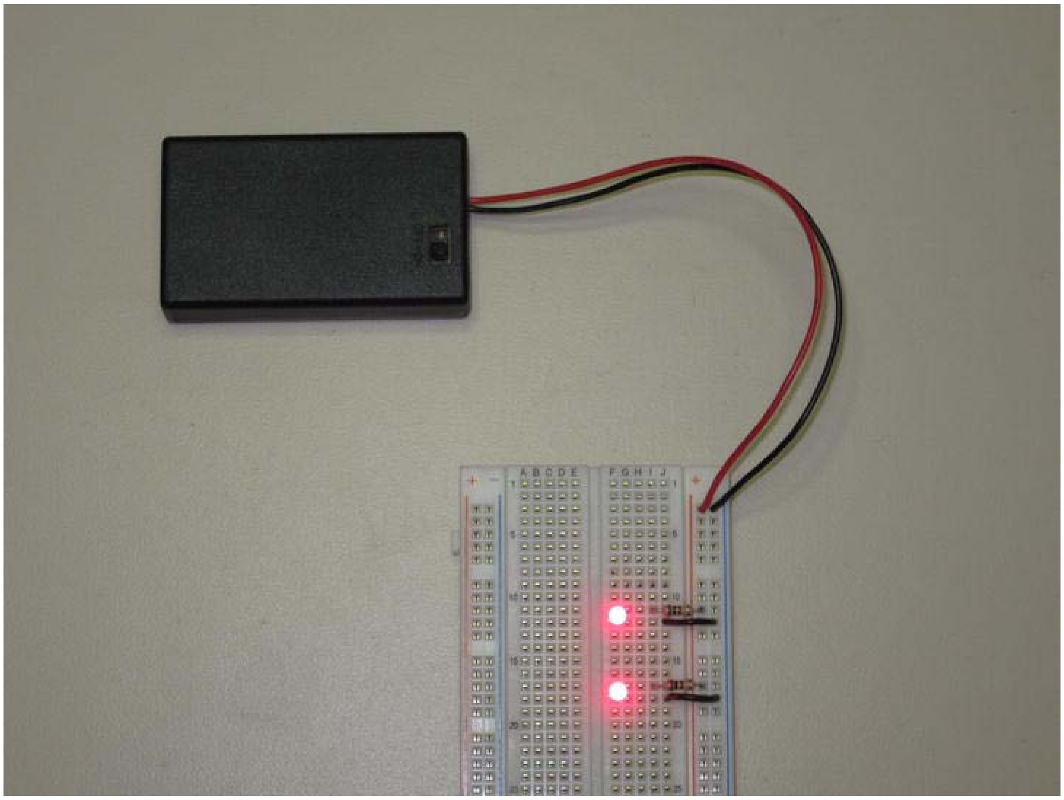

Figure 5-12. *Two LEDs are illuminated at the same time. This simple circuit demonstrates both parallel and series electrical circuits. A 3×AAA battery holder with an integrated power switch is used to provide power to the circuit.*

To build this circuit as shown, you will need the following components:

- A solderless breadboard with power rails
- A 3×AAA battery holder (which holds three AAA cells) and three fresh AAA cells—rechargeable cells are better
- Two LEDs (any visible color will do
- Two 100Ω 1/4W resistors
- Two short jumper wires

Using the letters printed on the breadboard to find the columns across the top and the numbers to find the rows, we can describe each of the tie points using a single letter and a number, such as A-1 or B-2. The following exercise shows how to put it all together.

ASSEMBLY INSTRUCTIONS

1. Install the top LED with its anode (longer lead) in tie point F-11 and the cathode in tie point F-12.
2. Insert one of the resistors from the red power rail on the right at row 11 across to tie point I-11.
3. Connect the right-hand blue power rail at row 12 to tie point J-12 using a short jumper wire.
4. Install another LED just below the top LED, with its anode in tie point F-17 and the cathode in F-18.
5. Install the other resistor from the right-hand power rail at row 17 across to tie point I-17.
6. Use another short jumper wire to connect the blue power rail on the right at row 18 to tie point J-18.
7. Install three fresh AAA cells into the battery holder.
8. Insert the red wire (positive) coming from the battery holder to the red power rail on the right at row 3 (the top).
9. Insert the black wire (negative) from the battery holder to the blue power rail on the right at row 3.
10. If your battery holder has a power switch, turn it on now.

The circuit is now complete. Please verify that both LEDs are illuminated, as shown in Figure 5-12.

Troubleshooting

If both LEDs are not shining at this point, you're going to have to do some troubleshooting.

- The first thing to do is double-check your wiring. Does your breadboard look like the one in Figure 5-12?
- Next, make sure you're using good batteries. If you have a voltmeter, you can test the batteries by measuring their voltage. New alkaline or carbon zinc cells should read 1.5V or more. Rechargeable nickel-cadmium (NiCd) or nickel–metal hydride (NiMH) batteries should read around 1.2V each when completely charged.
- Are the cells properly installed in the holder? Are they in the proper orientation and firmly seated?

- Is the power switch (if available) not in the On position? This happens more frequently than anyone would like to admit. Wiggle the switch back and forth and see if it's only making intermittent connection.
- Check the orientation of the LEDs. They will only work if installed in the correct orientation. It only takes a second to unplug them, swap them around, and try them the other way. At these low voltages, it won't hurt them to be installed backward.
- Are your resistors the right value? The 100Ω value is not critical. Anything from 100Ω to 1KΩ will do nicely. Don't use smaller values, as this will allow too much current to flow through the circuit, perhaps damaging the LEDs. How much current? We'll figure that out in just a bit. Larger-value resistors will decrease the amount of light emitted by the LEDs, down to the point where you can't see them anymore.
- Are you sure your LEDs emit visible light? Millions of infrared LEDs are manufactured every month for use in remote controls and night-vision gear. This light is not visible with the human eye, although some digital cameras can see it. Just because it looks like an LED doesn't mean it *is* an LED.

A Different View of the Same Thing

Your version of the circuit in Figure 5-12 might look a little different, and that's just fine. The important point is that both LEDs are lit up at the same time when you apply power.

Let's look at the traditional schematic for this circuit to get an idea of what's going on here. See Figure 5-13.

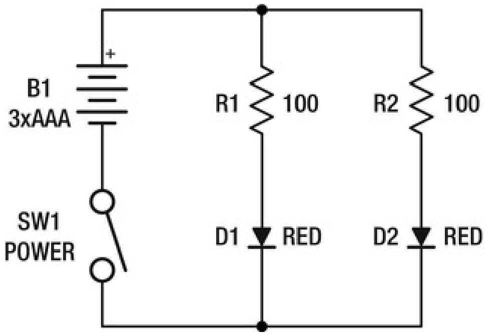

Figure 5-13. *The traditional schematic diagram of the two-LED circuit. This diagram should convey the very basic information needed to build the circuit, as well as understand the important aspects of how the circuit is supposed to work, without going into much detail at all about the actual implementation or assembly of the components.*

By now you should be able to look at the schematic and see the loop formed by the circuit. The only new twist is that there are *two* paths along which the electricity can flow, either through R1 and D1 (the top LED on the breadboard) or through R2 and D2 (the bottom LED, or the initials of a certain adorable robot from the movies).

The three AAA cells in the battery holder are shown as B1 in the diagram. The power switch, labeled SW1, is shown as a separate component to remind us that the circuit can easily be broken, halting the flow of electricity, which is something we want to happen from time to time. The connections between the two resistors and the two LEDs should be somewhat obvious. I'll discuss the details for the parallel and series connections momentarily.

Remember that an electronic schematic diagram is an abstract representation, boiled down to only the essential items involved and the relationships among them. As such, it doesn't need to resemble the final product, but it could. See Figure 5-14.

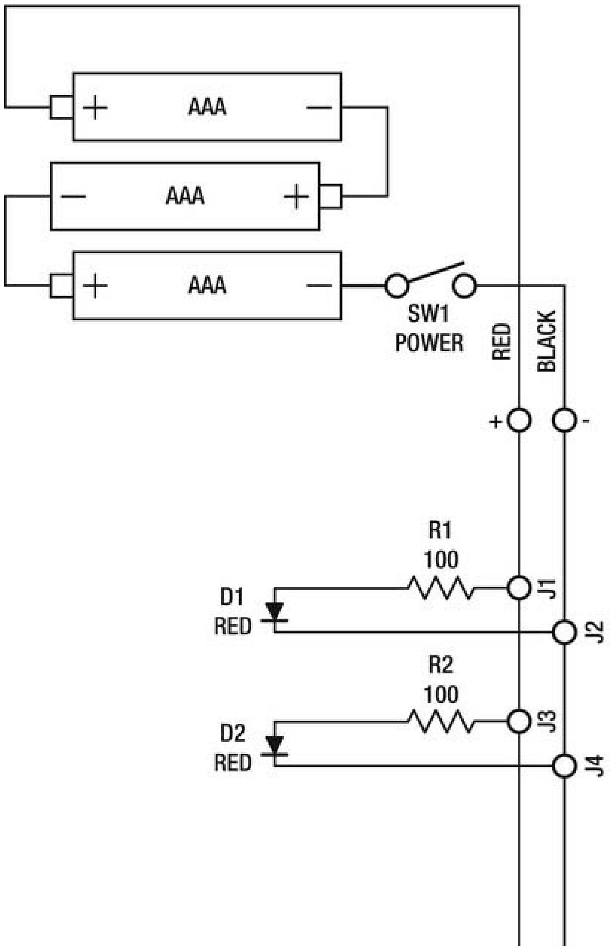

Figure 5-14. *Another schematic diagram of the two-LED circuit, showing the approximate physical layout of the components on the solderless breadboard. This style of schematic is often used when describing how to actually build a circuit, showing the constraints of the actual, physical components and the connections between them. It conveys more specific information than the previous example in Figure 5-13.*

Batteries in Series and Parallel

Let's look at the batteries first. Technically, as mentioned in Chapter 1, there is one *battery* in this circuit, and it happens to be composed of three *cells*. The term *battery* is very often used interchangeably with *cells*, and the distinction is minor. It's better to avoid confusion than it is to be exactly, precisely *correct*—usually.

We see three cells sitting side by side in both the schematic in Figure 5-14 and the actual battery holder shown in Figure 5-12. The cells are connected *in series*, meaning that one connects to another that connects to another, and so forth. They are like beads on a single wire. You should be able to trace the wiring (the solid lines) that connects the individual cells in the schematic.

Note the polarity of each of the cells within the battery. The positive terminal of the bottom-most cell is wired to the negative terminal of the center cell. The positive terminal of the center cell is connected to the negative terminal of the topmost cell.

The negative terminal of the bottom-most cell is connected to one side of an *SPST (single-pole, single-throw)* switch. The other side of the switch emerges from the battery holder's body as the black, or negative, lead from the battery pack. The positive terminal of the topmost cell connects directly to the red, or positive, lead coming out of the battery pack.

The switch could have just as easily been installed in the positive lead, and it would have worked exactly the same. The battery holder used in the photograph happens to switch the negative lead, as was revealed by examining the internal wiring using a multimeter as a continuity tester.

When the switch is *open* (i.e., in the Off position), the circle is broken and no electricity is going to flow. When the switch is *closed*, the circuit is complete and power is made available to the remainder of the LED circuit, assuming that it is wired up correctly.

The amount of *voltage* being supplied by the battery holder depends on the electrical characteristics of the individual cells. If alkaline batteries are used, each cell will provide 1.5V, assuming that the cells are relatively fresh. Rechargeable NiCd or NiMH cells will provide around 1.2V when fully charged.

When batteries like these are connected in series, the total voltage available is the sum of each of the cells. Just add all the voltages together. For example, using fresh alkaline batteries, the total voltage would be 3 × 1.5V = 4.5V. Rechargeable cells would provide 3 × 1.2V = 3.6V when fully charged.

If the cells were wired *in parallel*, with all of the positive terminals being connected together and all of the negative terminals wired together, the total output voltage would only be 1.5V for alkaline cells. The total output *current*, however, would be triple the capacity of each individual cell. See Figure 5-15 for a comparison of series, parallel, and combination circuits using batteries.

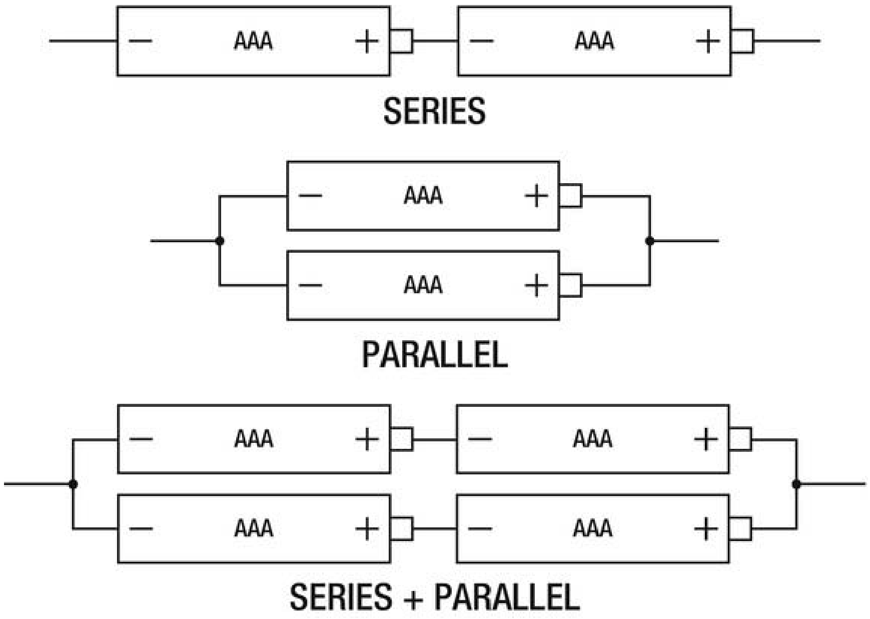

Figure 5-15. *Different arrangements of the wiring of the batteries produce different results. The top two cells are arranged in series. Their respective voltages are added together. The center circuit consists of two cells arranged in parallel. This configuration doubles the current that can be delivered. The bottom drawing shows a combination of both series and parallel circuits, doubling both the voltage and the current.*

When connected in series, the total current output is the same as the current capacity of each individual cell.

In summary, if you need more voltage, stack the cells end to end. If you need more current, stack them side by side. Even more complex arrangements are possible, using both parallel and series circuits to achieve the desired power output characteristics. This practice is very common in hobby model aircraft and battery-powered vehicles. A designation such as 3S2P means that there are two parallel circuits (2P) of three cells in series (3S).

Looking back at the circuit in Figures 5-13 and 5-14, we see that the "battery" used in this circuit consists of three cells in series. This means that if alkaline batteries of 1.5V each are used, the total voltage available to the circuit is 4.5V. If typical rechargeable cells (NiCd or NiMH) are used, we can expect 3.6V to flow through the circuit.

Switches in Series

The power switch, labeled SW1 in the schematic, is *in series* with the rest of the circuit. As mentioned in the previous section, when this switch is open, no current flows through the circuit at all. The switch is intentionally placed at this location in this circuit to act as a bottleneck. When the power switch is turned off, no power flows through the circuit. That seems like a good idea, considering the desired operational goal.

The schematic symbol for SW1 shows a pretty good likeness to the internal wiring of any simple SPST switch. There are two terminals, and a conductor that can be moved to either electrically connect them or electrically isolate them. As drawn, the switch is in the open position.

If we added another identical switch in series with the existing switch, what would happen? Take a look at Figure 5-16.

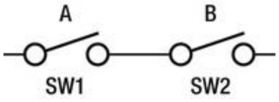

Figure 5-16. *Two switches in series. Both switches must be closed for electricity to flow through this part of the circuit. If either switch is open, the circuit is broken and no current will flow.*

Both switch A (SW1) and switch B (SW2) must be moved to the closed position for any electricity to flow through the circuit. If either switch is opened, then no current will flow. This is sometimes referred to as an *AND* circuit. Both A *and* B must be on for the circuit to operate.

A circuit just like this is often used in security systems. Switches are attached to doors and windows. All of the switches are wired in series, forming a large loop. The switches are all configured to be closed when the doors and windows are closed, and to open when doors or windows are opened.

If all the doors and windows are closed, then the loop of switches will allow the flow of electricity. Although it sounds exactly *backward*, when electricity is flowing in the circuit, this keeps the security alarm *turned off*. When any one of the switches is opened, the circuit is broken. The alarm then sounds.

Most real security systems are more complex than this, of course. The main advantage of this particular configuration is that it is easy to install and requires the least amount of wiring. The chief disadvantage is that when the alarm is triggered, you have no idea exactly *which switch* caused it.

Switches in Parallel

Just like with batteries, switches behave differently when configured in a different manner. Let's connect two switches *in parallel* this time and see what happens. See Figure 5-17.

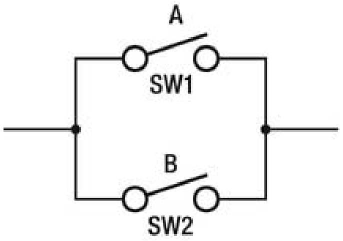

Figure 5-17. *Two SPST switches in parallel. If either one of the switches is closed, then power can flow through this part of the circuit. Both switches must be opened to stop the flow of current.*

The circuit in Figure 5-17 will conduct electricity if switch A (SW1) *or* switch B (SW2) are closed. To stop the flow of current, you have to open both switches. This can be used as a simple *OR* circuit. If switch A *or* switch B is closed, then power can flow through the circuit.

Looking back at the schematics in Figures 5-13 and 5-14, we see that SW1 is in series with the rest of the circuit. This makes it a very effective way to turn power on and off from a single location. Handy!

The Three-Way Switch

You might think that the circuit shown in Figure 5-17 could be used to turn lights on and off from two different locations. You would be partly correct. You could turn lights *on* from either switch location, since flipping either switch to On will cause current to flow and lights (or other appliances) to operate.

Unfortunately, you would not be able to turn *off* the light from either switch if the other switch were on. So this configuration is good for turning lights on but not off. A slightly more complex arrangement is required to do both.

We will need two switches, as before, but they will need to have a little more functionality to accomplish what we've set out to do. Instead of a switch that's either open or closed, we need a switch that alternates between two different routes, depending on how it is positioned. This type of switch is called a *double-throw* switch, as it connects a *common* terminal to either one of two different terminals. It can be *thrown* in two different directions. This should be much easier to understand when looking at a schematic symbol. See Figure 5-18.

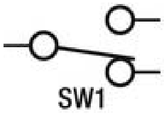

Figure 5-18. *A single-pole, double-throw (SPDT) switch. The common terminal on the left side is connected to either the top or bottom terminal on the right. Variations of this type of switch have a neutral position in the middle. Push-button (i.e., momentary contact) versions of this switch will label the two right-side terminals as either "normally open" or "normally closed," depending on the physical configuration of the switch.*

Now comes the clever bit. We'll connect a pair of these switches together by tying the two non-common terminals on one switch to their counterparts on the other switch. This results in four possible combinations of switch positions (both down, one up, the other up, and both up). Two possibilities turn the light on, and the other two turn the light off. The important point is that the "on" or "off" state of the entire circuit can be changed from either location, independently of the state of the other switch. Problem solved! Again, this should be easier to see when presented as a diagram. See Figure 5-19.

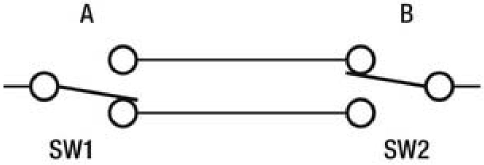

Figure 5-19. *A three-way switch circuit built using two SPDT switches. Power is either carried by the upper wire or the lower wire, as long as both switches are in the same position. If not, then no power flows across this part of the circuit. As shown, the three-way switch is effectively off. Flip either switch to turn it on.*

If you ask a real electrician how to wire a three-way switch, you'll get several answers, or more likely, you'll get, "It depends." That's because *it really does* depend on the actual wiring situation. Sometimes the power is routed through one or the other of the two switches, and sometimes the power is routed through the load. Each of these scenarios requires a different routing of wires and connections.

You are strongly advised to consult a real electrician concerning wiring in your house, where your family, pets, and all your stuff live. Less-than-perfect wiring causes fires or electrocution, which are both very unpleasant ways to die.

■ **Caution** This book does *not* teach you how to wire electricity in your home. Consult a professional.

Resistors in Series

Just like batteries and switches, resistors perform different functions in a circuit depending on their configuration. Let's look at using resistors *in series* first, and then consider using them in parallel later. Both combinations have their uses, as you will see.

Let's look at just a small part of the schematic that we've been working with so far in this chapter. See Figure 5-20.

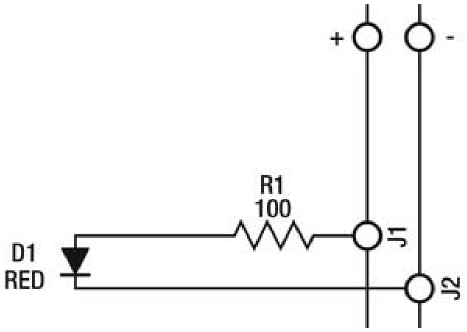

Figure 5-20. *One part of the circuit showing an LED, a resistor, and the power rails from our solderless breadboard. The resistor and LED are connected in series. The battery and power switch are not shown.*

LED Curiosities

D1 is an LED. It is a polarized component and will only work when installed in the proper orientation. The schematic indicates a red LED should be used, but feel free to use whatever color LED you like or happen to have available. The triangle shape of the schematic symbol helps remind us of the direction of *conventional current* through the device. The actual flow of electrons is the other way, but that's only because electrons are negatively charged particles and do everything backward. Apparently this was not well understood when electricity was first being tamed for our use.

LEDs are fascinating devices, emitting light in a wide range of colors, coming in an astounding array of shapes and sizes, and generally requiring very little power to operate. They are almost as easy to use as incandescent bulbs and are certainly more power efficient.

Being semiconductors, however (which leads to their being semimagical in their inner workings), LEDs generally only operate correctly within tightly bounded parameters. For example, a typical LED will only need between 2 and 3V to shine quite brightly. Incandescent lightbulbs, on the other hand, can operate from less than 1V to many hundreds of volts, depending on their construction and intended purpose.

While LEDs have a minimum *forward voltage* requirement that must be met for them to work at all, the brightness of an LED is determined solely by the amount of current flowing through it.

LEDs Have No Self-Control

In addition, LEDs have no internal, inherent mechanism to limit the amount of current flowing through them. Incandescent bulbs, when cold, have a low internal resistance, allowing much current to flow through them when connected to the right power supply. Remember that less resistance means more current can flow, and vice versa. This resistance almost instantly goes way up as the filament of the bulb glows white hot. The filament itself then becomes the limiting factor in how much current flows through the bulb, usually stabilizing at the proper power level.

LEDs have no such sense of self-control. They will conduct as much current though themselves as you are willing to provide, up to and beyond the point where the internal heat generated by the small amount of resistance they do possess causes them to literally burn up. In general, it's usually the tiny, almost microscopic internal bond wires that fizzle out, not the actual LED chip itself. However, dead is dead, and LEDs are generally considered to be "unrepairable" subassemblies. (Believe it or not, that's the correct technical term, even though it's not the correct English word—*irreparable* is the correct term when speaking of non-electrical things).

Care and Feeding of LEDs

The simplest possible solution to this ~~problem~~ opportunity for excellence is the use of a *current-limiting resistor* in series with the LED. Recall that a resistor will resist the flow of current in a circuit. The resistor used in this circuit, R1, is indicated to exhibit a resistance of 100Ω.

Let's say that D1, the red LED in our circuit, requires 2V in normal operation. This is pretty common for red LEDs. Generally speaking, the forward voltage requirement for LEDs goes up as you get farther up the color spectrum. Blue LEDs, for example, often require 3.0V to 3.6V to operate.

If we're using rechargeable cells in our prototype (and we really should be, you know), we can expect 3.6V to be made available to the resistor-plus-LED loop within our circuit. If 2V is needed for the LED, where are the rest of those volts going?

The answer is that they are being used to heat up the current-limiting resistor R1. Resistors turn current into heat. That's what they do. They do it, however, in a very precise and predictable way, which makes it easy for us to calculate how much current, voltage, and heat will be churning around in our circuits.

■ **Caution** Third-grade arithmetic ahead: Addition, subtraction, multiplication, and (shudder) division. You have been warned.

If we started with 3.6V, and two of those volts are being used by the LED, then our good friend Subtraction tells us that we've got 1.6V flowing through R1. For a thorough explanation of why this happens the way it does, feel free to look up Kirchhoff's Voltage Law. It basically says that all the voltages in a circuit end up being equal. The voltage across the resistor and LED will equal the battery voltage. The voltages within the LED-in-series-with-resistor circuit add up, just like battery voltages add up in series.

If we know the voltage across R1 (1.6V) and the resistance of R1 (100Ω), then we can use Ohm's Law to calculate the current flowing through R1. Reviewing what you learned in Chapter 3 about the predictable and inviolable relationships between voltage, resistance, and current, you can see that I = E / R (i.e., current equals voltage divided by resistance), where I is the current in amps, E is the voltage in volts. and R is the resistance in ohms.

It's not that hard! It sounds much worse than it actually is. The answer is 0.016A, or 16mA (1.6V/100Ω). That's the teeny-tiny amount of current needed to light up a typical red LED. Most small LEDs can handle up to 25mA safely, and some can handle much more than that, especially if they are intended for illumination purposes.

One quick side trip to multiplication land and we'll know everything we need to know about R1 in this circuit. While the schematic specified the resistance of R1 as 100Ω (the primary characteristic or component *value*), nothing was said about the power-handling capability of this component.

Power is calculated as voltage multiplied by current. That's all there is to it. We know the voltage: 1.6V. We know the current: 16mA. We multiply them together to get 0.0256W (watts), or almost 26mW. Remember that you have to enter the current as 0.016 because the formula calls for the number of *amps*, not milliamps.

Not quite 26mA is not quite much of anything. You will not be able to tell that the resistor is shedding all those extra volts as heat with your fingertip. It might not even register with a sensitive thermometer. It's just a really tiny amount of power. A typical small resistor (a 1/4W resistor) will handle 250mW of power, which is almost ten times what is needed in this application.

■ **Tip** It's always a good idea to calculate the power requirements of your components, just so you're not *too* surprised when the magic smoke leaves the circuit, rendering it crispy and nonfunctional.

OK, we're back from the frightening land of multiplication. It wasn't *that* bad was it? The remainder of understanding resistors in series depends only upon the science of addition. Counting on your fingers will be good enough.

We can change the brightness of the LED by changing the amount of current flowing through it. We can change the amount of current flowing through the LED by either raising the supply voltage or by changing the resistance of R1. Since we can easily substitute any other value of resistor for the 100Ω resistor with which we began, this is easy to do. Just don't go *too low* in resistor value, as this will allow *too much* current to flow through the LED, possibly damaging it permanently.

Adding resistors in series adds their resistances together. For example, if you placed two 100Ω resistors end to end, the total resistance across them would be 200Ω. That's really all there is to it: 100Ω + 100Ω = 200Ω.

If you perform this experiment with the LED circuit we built on the breadboard, you can see that the LED will not shine as brightly as it did before. This is because the resistance has doubled, which means, all other factors remaining the same, that the current through the LED has been cut in half. Half the current results in half the brightness, since LEDs are controlled by the amount of current flowing through them.

What's interesting about this arrangement is that the power-handling capacity of the two resistors is *also* double what it was when only one resistor was in the circuit. This is because each resistor is now dissipating half the power when compared with the previous scenario. If two 1/4W resistors are used in series, they can together safely dissipate up to 1/2W of power, although pushing any component to 100 percent of its capacity is just asking for trouble, especially when heat is concerned.

By connecting resistors end to end, you add their resistances together, as well as their power-handling capabilities. By keeping a small variety of basic values on hand (e.g., 1Ω, 10Ω, 100Ω, 1KΩ, 10KΩ, etc.), you can easily stack up the right combination needed for almost any electronic application. This is the method used in the very handy device known as a *decade resistor* or *decade box*, which has dials for each digit in the desired resistance, allowing you to dial up any value of resistor needed.

Extra credit: If you've followed along this far, consider researching voltage dividers built using only resistors.

Resistors in Parallel

What happens when resistors are connected side by side instead of end to end? That's where things get interesting, at least from a mathematical point of view.

An easy way to look at parallel resistors is to think about the amount of current flowing through each one, and then add them all up. See Figure 5-21.

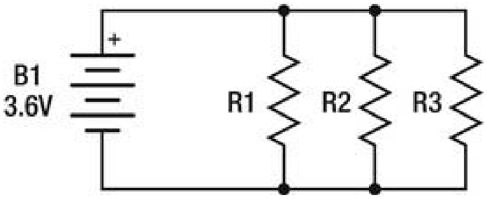

Figure 5-21. *Parallel resistors each conduct a portion of the total current flow through the circuit. Each resistor allows a certain amount of current to flow. The sum of their currents can be calculated if their individual resistances are known.*

Let's assume that R1, R2, and R3 have resistances of 100Ω, 200Ω, and 300Ω, respectively. The total resistance of the circuit across the battery terminals is *not* 600Ω, as would be the case if they were in series instead of parallel. The total resistance is less than 100Ω, which is the least of the resistors. How can this be?

First, let's think in terms of how much current is flowing in each of the three parallel pathways within this circuit. If B1 is putting out 3.6V, and R1 is 100Ω, then we know to use Ohm's Law to calculate the current, with I = E / R. In this case, I = 3.6V / 100Ω = 0.036A, or 36mA. *In addition*, the path through R2 is drawing half that (because R2 is double the value of R1—get it?), or 18 mA. *In addition*, the final pathway is drawing one-third as much current as the first one, or 12 mA. Just to verify: I = 3.6 / 300Ω = 0.012A, or 12mA. Everything checks out.

Now add all those currents together: 36mA + 18mA + 12mA = 66mA. That's the total current, using this method. Again, since we know both the voltage and the current, we can find the *effective* or *equivalent resistance* of all three resistors in parallel, as if they were one single resistor. Rearranging the formula for Ohm's Law, we find that R = E / I. In our case, R = 3.6 / 0.066A = ~54.55Ω. See? Smaller than 100Ω, somehow.

Thinking about the individual *currents* flowing through each resistor is the same as adding the current-supplying capacity of battery cells together when arrayed in parallel. Just add them together. However, we had to calculate each current separately and then add them together. There is a more direct way, depending upon how you look at things, but beware: it involves not only *fractions*, but *fractions upon fractions*. Oh, the horror.

■ **Note** Some people don't dig math. Weird, but true.

Here is the equation for calculating the total resistance (R_T) of three resistors in parallel (R_1, R_2, R_3):

$$R_T = \frac{1}{\left(\frac{1}{R_1}\right) + \left(\frac{1}{R_2}\right) + \left(\frac{1}{R_3}\right)}$$

The good thing about this equation is that you don't need to know what the supply voltage is to get the correct answer. All you need to know are the values of the resistors. If you need to use more resistors, just add more terms to the equation.

Calculating the value of parallel resistors occurs quite frequently when designing electronics. Some special cases are simple enough to do in your head. For example, if two identical-value resistors are placed in parallel, their combined resistance is half the value of each individual resistor. That is, if you placed two 100Ω ohm resistors in parallel, their effective resistance would be 50Ω. Similarly, if you have three identical resistors, the net resistance is one-third the value of each constituent resistor. Make sense? These are special cases of the general equation just shown.

This happens so much that an alternate notation method has been developed: $R_T = R_1|R_2|R_3$. This is not often used for hobby electronics, however.

The power-handling ratings of resistors are added together, so if you have two 10W resistors in parallel, the net power-handling capacity is now 20W. At least that part was easy.

Capacitors in Series and Parallel

Capacitors are completely different from resistors. When calculating their values in series and in parallel, you have to use *exactly the opposite* methods. Isn't electronics interesting?

The values of capacitors *in parallel* are added together. This makes them much more like batteries in parallel than resistors. However, you still have to use the fractions-upon-fractions method to calculate the value of capacitors in series.

Again, some shortcuts are available. Two identical capacitors in series have *half* the capacitance of each individual capacitor, but *double* the working voltage, which is the maximum voltage they were designed to withstand before failing.

Inductors in Series and Parallel

It gets more interesting, believe it or not. Inductors act like resistors as far as adding their values together in series and taking the reciprocal of the sum of the reciprocals of their values in parallel, but *only* when the magnetic fields of all of the inductors involved do not interfere with each other. If the magnetic fields generated when current flows though the inductors overlap, then you have to factor in their *mutual inductance*, which starts to get complicated.

Extra credit: Figure out how all that works and let me know.

LEDs in Series and Parallel

You learned (or were reminded of) a few of the basic characteristics of LEDs in the previous section, although it was really to help illustrate the operation of resistors in series and parallel. A couple of major oversimplifications were introduced to help concentrate on the resistor. We'll address a couple of these issues here.

Lighting up a single LED using a battery and a fixed current-limiting resistor is fun, but it's pretty basic stuff. Lighting up a *lot* of LEDs is where it gets to be a *lot* of fun. To do so, you need to learn a little more about the weirdness that is *semiconductor*.

LEDs are diodes (*LED* stands for *light-emitting diode*, after all). Diodes are semiconductors and are made by elves in a hollow tree. Semiconductors do not have nice linear voltage-to-current curves.

For the purposes of simplicity, in the previous section, it was suggested that LEDs have a certain minimum voltage that they need to operate. This is called their *forward voltage*, and is sometimes referred to as their *voltage drop* within a circuit. So far, this is all true. You can find an LED's forward voltage specification in the data sheet published by the manufacturer.

What you'll find, when you look, is that every manufacturer only publishes a *range* of voltages, such as 1.8V–2.2V, giving a minimum and a maximum voltage. This is what they are contractually obligated to deliver in their product. Anything outside of this range is considered *defective*, or at least out of tolerance. You might get a good deal on "rejects" that fall outside these parameters but still light up quite nicely. Such is the provenance of many of the penny LEDs available from overseas. Yes, they exist. And yes, you get what you pay for.

There are two reasons why LED manufacturers are being so deliberately vague in their specifications. The first is that process variations are going to produce a certain randomness to these values. One batch might be identical to the next batch, but then the third batch might be all over the place. Such things happen.

The second reason is where the real weirdness begins. Forward voltage is not a fixed value in any given device, but instead increases as the current increases. Again, this increase is *nonlinear* and cannot be calculated with a simple formula.

The best way to demonstrate this to yourself is to connect an LED to a power source via a variable resistor, such as a potentiometer. For the protection of your LED, add a fixed resistor in series with the potentiometer so that you establish a maximum current level, even when the potentiometer is turned all the way up (i.e., has almost zero resistance).

Now apply power and adjust the potentiometer until the LED just barely lights up. Now measure the voltage across the terminals of the LED, using a voltmeter. It should be in the neighborhood of 2V–3V. For extra credit, measure the current flowing *through* the LED, using an ammeter. This measurement will most likely be less than 1mA.

Compare these readings to what you find when the LED is at full brightness. The forward voltage across the LED goes up as the current goes up. To map the nonlinearity of this function, you would need to record many readings of both voltage and current, and then plot them in a graph.

Just as you can wire resistors, capacitors, and inductors in series and in parallel, you can do the same with LEDs. As with resistors, the forward voltage of the LEDs are added together when arranged in a series circuit. This raises the minimum voltage required to light up a string of LEDs in series. This means that you can't light up ten LEDs in a single string using a 9V battery. You have to first overcome the minimum forward voltage just to get any current at all flowing. Then you have to up the current to get the desired brightness, which simultaneously raises the total forward voltage through the circuit. You still need a current-limiting resistor in this circuit—but only one is needed. You don't have to have one for each LED.

The amount of current flowing through the string of LEDs is determined by the supply voltage, the cumulative forward voltage of all the LEDs, and the resistance of the current-limiting resistor. The same amount of current is flowing in every part of the circuit. If 25mA is being drawn from the power supply to light up your single string of LEDs, that means that each and every LED has 25mA of juice flowing through it—more than enough to light them up quite nicely.

If you have a 12V power supply (or a car battery), you could string up to five or maybe six red LEDs in series, with a small current-limiting resistor. That's because each red LED will typically have a forward voltage between 1.8V and 2.2V. Other colors higher up the rainbow might need more voltage per LED, meaning fewer can be lit at once using a single string of LEDs in series.

But don't fret. You can light up as many LEDs as you want by placing multiple strings of LEDs in parallel with each other. Each additional string of LEDs adds to the total current draw, so you will eventually find a limit along that dimension as well. For an example of a small LED array using both parallel and series circuits, see Figure 5-22. The schematic is shown in Figure 5-23.

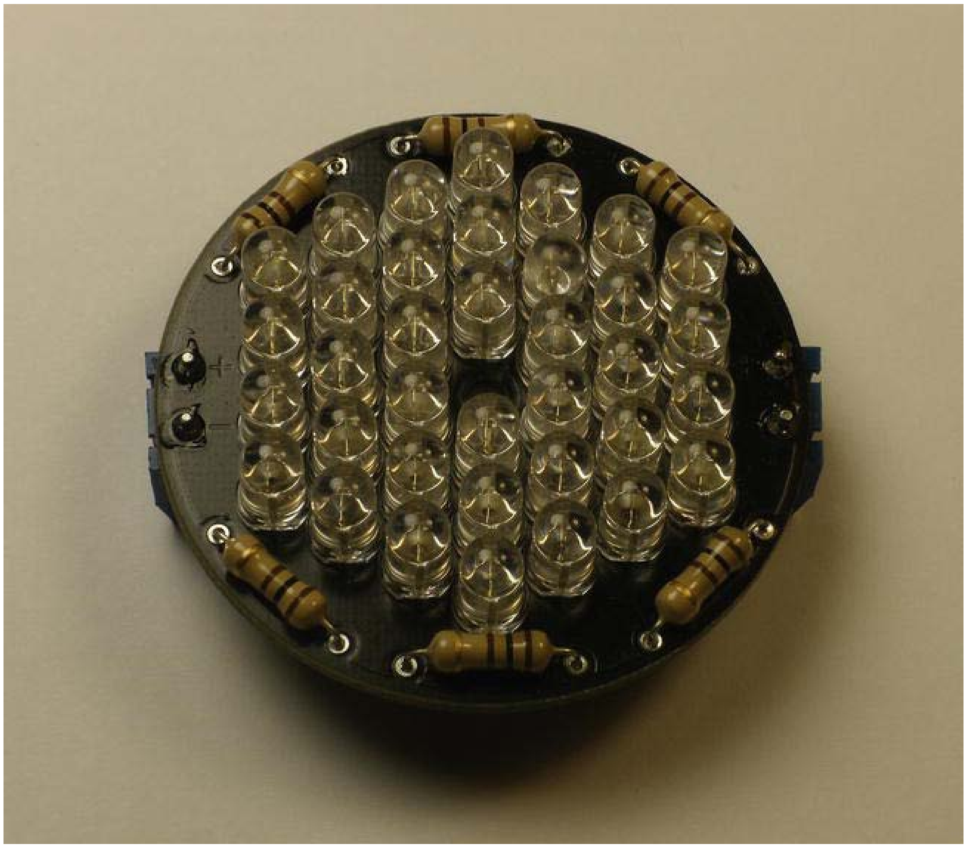

Figure 5-22. *An array of six parallel strings of six infrared LEDs in series. These infrared spotlights are used for covert surveillance and security applications.*

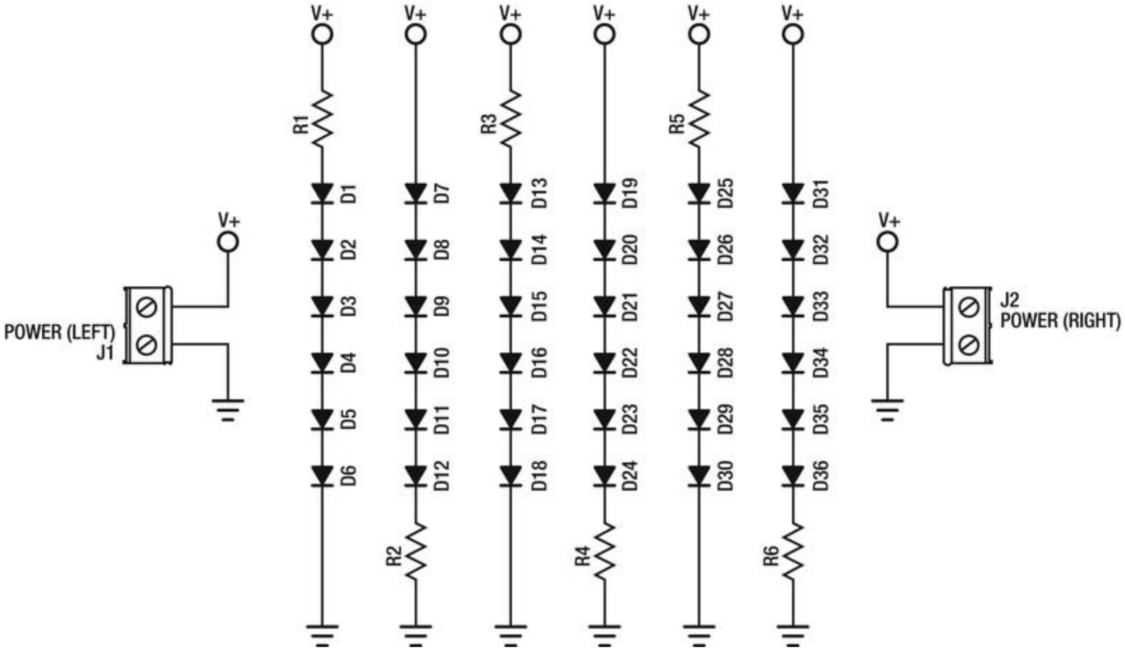
V+
V+
V+
V+
V+
V+
V+
V+
R1
R2
R3
R4
R5
R6
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
D16
D17
D18
D19
D20
D21
D22
D23
D24
D25
D26
D27
D28
D29
D30
D31
D32
D33
D34
D35
D36
POWER (LEFT)
J1
J2
POWER (RIGHT)

Figure 5-23. *The schematic of the LED array, showing both parallel and series circuits. The circuit offers two power connectors to make it easy to wire multiple LED arrays using the same power supply.*

Summary

Even if your workbench doesn't look as bad as the one presented at the beginning of this chapter (or even if it looks *worse*), hopefully you've learned a thing or two about the importance of they saying, "A place for everything and everything in its place." Since *you* get to decide where everything goes, you can either make it easy or hard on yourself as the years go by.

The workspace in this chapter extended to a total of 10 square feet. If this is what you have to work with, you can certainly set up a comfortable, safe, and productive area for your electronic projects. If you have even more space available (adding a shelf either above or below the desk is a great way to do this), then even better.

Hopefully some of the more detailed descriptions of how electronic components can be wired up have given you some ideas for projects of your own. The very best possible way to learn more about electronics is to think up some simple experiments and try to build them in your lab. Have a goal, take lots of notes, and have some fun—even your failures can be used as stepping stones to more ambitious and creative projects in the future.

Have a *lot* more room? Then perhaps you should be thinking about setting up a group lab or classroom. Chapter 6 will give you some helpful advice about what you are going to need to get started. See you there!

CHAPTER 6

The Small Group Lab and Classroom

The first part of this book covered what you need to know to build your own personal electronics lab. Now it's time to share the love with your fellow electronics enthusiasts. Where to begin?

Like most important questions, the answer remains, "It depends." What resources are available to you and your group? How many folks will likely want to be working together, and for how long? How well do you really know these people? And these are just a few of the questions you're going to encounter along the way.

Take the time to understand exactly who "you" (plural; in Texas, *y'all*) are and what your requirements and goals could be. By doing so, you map out the future of your successful facility. Likewise, *not* planning ahead or making reasonable assumptions about your needs is simply an invitation to problems, problems, and more problems. Who wants *more* problems?

■ **Tip** If you fail to plan, you plan to fail.

Know Yourself

The ancient Greek maxim "know thyself" has as many definitions as it has authors, it seems. Plato indicated in several of his dialogues that Socrates would use this phrase to admonish his students to learn more about themselves and their own characters and capabilities before attempting to figure out the rest of the universe. Wise advice then and now.

As far as knowing yourself in relation to setting up a shared electronics lab, you should first think about what role *you* are going to play in this endeavor. Yes, *you*. Here are some of the major ones, and you might end up being all of them at some point.

Facilitator or Coordinator

In the role of the facilitator, you plan on providing an environment for others to use for their projects and explorations in electronics. You may or may not partake in these activities yourself, but have been tasked with (or volunteered for) the organization and coordination of the lab.

It's critically important to understand the needs of your intended ~~lab monkeys~~ users. See the "Know Your Group" section later in the chapter for more information on this topic.

As the lab facilitator, lab workers will come to you with questions, requests, comments, and complaints. Be prepared to address their needs in a responsive and professional manner—even if they're just your cousin.

Your responsibilities will include

- Communicating the scope of the lab (i.e., deciding what *is* and what *is not* acceptable or expected)
- Publishing any important information or updates regarding location, available times, features, rules, and possibly ongoing projects
- Crafting policies involving safety, proper tool usage and training, and participant conduct
- Maintaining supplies of commonly used materials
- Inspecting and repairing tools as the need arises

To balance out your responsibilities, you should be granted adequate authority in the following areas:

- Characterizing the scope of the lab, based on present and anticipated future needs
- Enforcing the policies of the lab
- Budgetary discretion in acquiring, provisioning, and maintaining lab assets
- Granting or revoking access to the lab based on established parameters

If you're going to be the Keeper of the Lab, then you should have both a personal interest in keeping it in good shape as well as the wherewithal to decide how things should be done.

Good luck finding the right balance there.

Teacher or Presenter

Let's assume that your lab has already been set up (or can be set up in a jiffy), and your main goal is to present information or teach classes there. Again, it should be apparent that a good knowledge of your intended audience will play a vital part in your success. See the "Know Your Group" section for more information.

Once you understand the needs of your class and the goals of your presentation, the next step is to leverage the existing facilities to your best advantage. Will you need specialized presentation equipment or materials? If so, see what is already available, and then make a shopping list of everything else you will need. Examples include lighting, a projector, specialized power distribution, and test equipment.

Are you making an interactive presentation? If so, what materials will you need to provide for your students to be able to fully participate in your presentation?

What will be your optimum teacher-to-student ratio? If you have a simple message that's easily communicated and reinforced, a single speaker can effectively address a rather large crowd. Once you start adding layers of complexity to the subject matter, you're going to have to be prepared to backtrack and explain things in more detail from time to time. If you're leading a group through a technical exercise, it's often profitable to have assistants available to help just those individuals that need a little more help, without having to stop the class in its tracks. These assistants need to be up to speed on the subject matter as well as have good communication and troubleshooting skills.

Remember that students look to the teacher for leadership, not only as a subject-matter expert. Take the lead, set a good example, and share your enthusiasm for your topic with your students.

A successful teacher leads a class on a journey of discovery. Make sure you don't lose too many along the way.

Participant

Your goal as primarily a group lab participant is to make effective use of the facilities at hand, making the most of your time and energy. Having access to a well-stocked lab with a good array of tools can be a decided advantage for both your immediate project goals, and it can be an excellent opportunity to accelerate your knowledge of electronics and other mysterious arts.

Imagining yourself as an active participant is the best perspective when laying out the flow of your lab. Set up workstations that have almost everything you need, leaving room for the complex or specialized stuff elsewhere. There are more suggestions about specific workstation configurations in the "Workstations" section of this chapter.

While the active-participant model helps you squeeze out the maximum utilization of your lab, be sure to consider the sort-of-working-but-also-sort-of-goofing-off model. Not everyone wants to be working full-tilt all of the time. Leave room for a certain amount of idleness, contemplation, and whimsy. You might be surprised at the results.

Observer

You want to be in on the action, but not necessarily working on anything in particular. That's great! Don't underestimate the importance of the purely social aspects of technical gatherings.

Some people just like to hang out with other like-minded individuals. This is a good thing—when resources and circumstances allow it. Watching creative, talented, and driven folks working on projects with passion is very often the best kind of inspiration. Let your kids see you having fun and making cool stuff in your lab, and the next thing you know, they are going to want to be there whenever you are. Pretty soon, they are going to want to start working on projects of their own—and that is its own special reward.

Know Your Group

Now you know a little bit about who you are and what your expectations of your lab may be. Now it's time to get to know the group of people who will most benefit from access to the lab.

Identify the needs and requirements of your group by first identifying *precisely* who your intended audience is going to be. Knowing this one item, up front, before any big plans are laid, is the best way to ensure success down the road.

There are several broad categories for which you might need to prepare. The more you know about your particular group, the better. Let's start fairly close to home and expand outward.

Your Family

Once you've established your own electronics lab at home, it's great to be able to share some of the fun with those that are closest to you. This might be your children, parents, siblings, or extended family. A well-planned and well-maintained electronics lab is going to generate some interest, or at least some curiosity. Here's how you can carve out some room for them, while still being able to get things done yourself.

It's very natural to want to share your hobbies and interest with others, and a safe and well-stocked electronics lab is a great place to do so. This is an especially ripe opportunity for parents to pass on their knowledge, skills, values, and work ethic to the next generation.

The family lab can be a logical extension of the family home. Everything is shared, including responsibilities and privileges. In more agrarian societies, this is already well understood at an early age, where chores around the farm and in the fields and a respect for the tools being used help enhance a sense of group belonging and worth. Everyone has a stake in the prosperity of the family endeavor.

Do you want your kids to appreciate what you give them? Then give them chores before you give them presents. Let them clean up the lab, put away tools, inventory parts, and take out the trash before spoiling them with all the fun stuff you get to do in the lab. This will help foster a sense of ownership in the long-term success of the lab, whether you're there to supervise or not.

It bears repeating (from Chapter 4) that no unsupervised children, no matter how adorable, should be allowed in the lab. Would you allow unsupervised cooking in the kitchen? It's the same basic scenario. When everything goes well, everyone has a good time. It only takes a split-second distraction to go from a good time to a trip to the emergency room in either case. Have rules and enforce them, even if it's your elders that you are protecting by doing so.

Sometimes having a resource like a shared electronic laboratory is the first opportunity people have to assert some sort of control over their relationships. Use the opportunity to make friends with your relatives, while you still can.

Your Friends

Unlike your family, as the saying goes, you get to *pick* your friends. Now you have the opportunity to provide a cool place to hang out as well as build all manner of awesome gizmos. It's only natural to want to share some of the fun with your like-minded friends.

The basic framework of rules and accountability apply to the friendly lab as well as any other group endeavor. Let your friends know up front what your expectations of them are in regard to use of the lab. Can they use your tools? Can they store artwork-in-progress at your place? When is it cool to come over to tinker, and when is it not? These kinds of questions should be dealt with early in the game; otherwise, these issues can flare up and become trouble spots. There's no substitute for good communication.

Once you extend your laboratory's hospitality to those outside your immediate family, it's a good idea to clearly designate what is your personal property and what belongs to the lab and all its participants. This can be as simple as putting name tags on all your tools and marking shared resources appropriately.

This is where *trust* really factors into the success or failure of a group lab. Just like in math, you want to eliminate as many variables as possible to ensure that a good time is had by all, and to ensure the long-term safety and viability of the facility itself. Which of your friends do you *trust* in your lab? Think about these things at length before opening the doors to just anybody.

Coworkers and Fellow Students

So you've been tasked with organizing or supervising a lab for a school class or university department. You may or may not have any choice in who will be granted access, or even when they can avail themselves to the facilities.

The balance of responsibility and authority is crucial in this scenario, just like in most other situations in life. Hopefully you will be in a position to enforce whatever rules are already in place. Ideally, you will have some say in crafting new rules as they become necessary, or adjust existing rules when they need it. In either case, it's just not fair for someone to expect you to do a good job, yet not provide you the tools to get it done.

Positions like this become available from time to time in most schools. Sometimes there's an opportunity for a small paycheck, and other times all you are going to get are points with the administration. Hopefully, at least you will also enjoy lab privileges. That's got to be worth something, yes?

If you are helping to run an existing lab, there are probably others you can consult who have had this particular honor in the past. Try to find out who these folks might be and see if you can get any helpful advice from them about your lab and its care and feeding.

If you're starting from scratch, you will enjoy the double-edged sword of setting things up the way you want, as well as setting things up the way everyone else wants it. You *get* to establish whatever level of functionality you want, with the caveat that you then must maintain that level. It's a freedom, like most freedoms, that comes with a price.

Special-Interest Groups

Are you a member of club that has a special interest in electronics? Several popular interests, such as ham radio operators, remote-control model flyers or drivers, robot builders, and computer users, have a love-hate relationship with electronics. Having a dedicated electronics lab available to the membership is a great way to build up the love as well as work out the hate.

Given a highly specialized crowd, such as the members of a university or trade school, you might even find people organized into an electronics club, especially these days when almost every aspect of our lives is touched by this technology in one way or another.

The specialized nature of your group helps focus what areas of electronics are of most practical interest, at least from a lab point of view. This cascades into easier choices when it comes to stocking the lab with workstations, tools, and components.

The Dallas Personal Robotics Group (DPRG; see `http://dprg.org`), for example, had the use of an 1,800-square-foot warehouse for over six years, rent free, due to the generosity of one of the members, Mike Dodson of Modern Assemblies. Having access to an excellent industrial setting allowed the club members to dream of and then realize much more ambitious robot projects than could be attempted at home. After Mike's retirement, the DPRG went into a nomadic phase and finally spun off a dedicated workshop and meeting area, now called the Dallas Makerspace.

The General Public

A lot of people see the popularity of do-it-yourself (DIY) and *maker-movement* projects and envision themselves as entrepreneurs, seeing a need and being able to meet it. Unfortunately, most of these people are wrong.

Building a lab dedicated to public use sounds like a great and noble aim. The realities of such an undertaking, including the financial, logistical, and business related, are not to be underestimated. While it seems to work well in certain areas and at certain times, it's almost always the result of the right combination of people coming together at the right time.

Hackerspaces (a terribly unfortunately chosen label) are springing up all over the world. The term *hacker,* while understood within some technical communities to mean a person of respectable skills, has a decidedly *unrespectable* taint to it when viewed from the outside. To the vast majority of the English-speaking world, a hacker is a rogue computer jockey, a criminal intent on breaking and entering into computer systems for nefarious purposes.

The term *maker,* on the other hand, leaves a more positive first impression. The Dallas Makerspace members deliberately made this decision when naming themselves, while they were still just a subcommittee within the board of directors of the nomadic DPRG, looking for a new home. It was important to them to maintain the very positive image that the DPRG had established and built over its long history, while still looking to the future.

If you're going to set up a lab for use by the general public, whether you decide to call it a hackerspace, a makerspace, or something entirely different, you get to make almost *no* assumptions about the good intentions of your membership. It's a sad state of affairs, to be sure, but it seems to be the prevailing wisdom of the day.

Some popular and successful organizations—such as Noisebridge (`www.noisebridge.net`), in San Francisco, California, which was incorporated in the state as a "non-profit educational corporation for public benefit"—operate almost entirely on a "consensus basis," sometimes known as "anarchic chaos." Noisebridge's overriding motto is "Be excellent to each other," very often shortened to just "Be excellent." Its very real success and sense of community among members is based on what seems to be a very local tolerance that is not to be found in many parts of the globe. If you ever get a chance, do stop by and prepare to be impressed. It's an amazing place.

A good parallel would be to visit your local public library. In the areas that still maintain libraries (shockingly, they are on the decline), you will find a finely tuned machine that meshes well with the surrounding locality. Are there bars on the windows? If so, you're going to need some bars on your windows. Are there metal detectors at the entrance? Oh, dear. Do you see any uniformed security forces present? These are not good signs.

Don't be completely discouraged, however. Sometimes a community in decline needs a little boost of energy and an injection of esprit de corps. This can often be accomplished by introducing some new alternatives for people's spare time. An ancient Muslim saying reads, "Trust in Allah, but tie your camel."

A Shared Resource

Do you have a dedicated space available? If so, you are way ahead of the game. Finding a good location is half the battle. If your space is truly dedicated to your pursuits and activities (i.e., a private home, club, or makerspace), you have the advantage of being able to leave tools, components, and projects in place, returning to them (and finding them still there!) when the opportunity arises. "Truly dedicated" in this sense can be equated with having your own key to the door and reasonably free access. While a 24-hour facility sounds like a great idea, there are a lot of logistical problems associated, such as security, utility costs, scheduling, and parking, just to name a few.

If your space is not-so-dedicated, such as a school lab or friend's garage, you can still make it a home away from home if you are able to arrange for secure storage of all your shiny toys. This could take the form of individual lockers or a storage closet.

No matter what the logistical arrangements, the main challenge to providing a versatile and effective setting for electronic and other kinds of tinkers is basically a social one. Once it's more than just you, it can start to get complicated.

Luckily, there are some time-tested methods for making these types of things run smoothly.

Be a Hospitable Host

If you are taking the lead in arranging a shared work space, please remember that you are responsible for extending the appropriate hospitality to your present and future guests. This single factor will greatly determine the vibe, feeling, or ambiance of the facility, even more than having the latest and greatest tools and gizmos on hand.

This task may or may not fall entirely on your shoulders. Many hands make light work. If you are able to delegate some of the duties of a hospitable host to other members of your group, consider doing so.

Think of your task in the same light as planning a successful party, even if it's one that you anticipate will last for days on end. These basic ideas are as old as civilized living. Think about fabulous parties or memorable events that you have been privileged to attend in the past, and then plan on taking it up a notch.

Try to anticipate the human needs of your attendees as well as their technological expectations. People have what seems like an endless list of requirements, some of which you can address, some not.

Make your location known. Make this information easy to find to those interested, and easy to update when the venue changes. A web site is trivial to set up these days, and even a post to a popular forum will tend to be accessible via Internet search engines, in time. There's a new social media outlet popping up every third Tuesday, so avail yourself to what makes sense for your audience. Word of mouth only goes so far these days.

Once you've gone to the trouble of publishing your location, go the extra mile and put out some sort of actual sign or indicator that your arriving guests have found the right spot. If you have an identifiable logo or symbol associated with your group, be sure to proudly wave that flag. People don't like feeling lost, or unsure whether they're entering a welcoming space or about to trespass on private property.

Let everyone know when they are expected as well as when they are *not* expected. Post your regular meeting times, if you have them. Alternatively, make known the appropriate hours of operation. It's disappointing to be turned away at the door just because it's 1:00 a.m. on a school night.

When possible, welcome your guests. This makes a big impression on both newcomers as well as the regulars. Being part of a group is supposed to *enhance* a feeling of inclusiveness and belonging. We are social animals at heart, and even a simple hello is often the prelude to an interesting and productive time together.

Have rules. Make these rules known. A sign on the front door is not out of place, as long as it's not too negative or off-putting. Rules should *prevent* problems from occurring, and not be enacted as a spontaneous reaction when problems inevitably occur.

Ensure your guests' comfort when attending. Is there enough room for the largest (and the smallest) participant to safely enter and exit the facilities? Is there adequate lighting in the common areas, if they exist? Is there a place for everyone to sit? Who has dominion over the thermostat?

Do you plan on making refreshments available? Will this duty rotate among the membership? Keep track of such things, if only to keep from overextending yourself in all the myriad details of hosting a successful meeting. While massive banqueting and consumption of yummies make for a festive and enjoyable association, these activities are not always aligned with maintaining an effective laboratory. Separate the feeding trough from the work areas. Stay focused on what you're trying to accomplish here.

Will everyone have access to restrooms, and whose job is it to keep these areas tidy?

And the most important question of all: Who takes out the trash?

Be a Courteous Guest

Just as it's important to play the role of the hospitable host, it is likewise critical to be a courteous guest when visiting a facility, whether you are a paid-up member in good standing or an invited guest. Again, these guidelines will vary according to the intents and purposes of your lab, but the basics will always remain the same.

If possible, and when appropriate, let your host know if you are planning on attending. This won't apply in a causal, drop-in type of laboratory. Even so, by making your intentions of attending known ahead of time, you might influence (one way or the other) the attendance of other potential participants. Nothing succeeds like success. A well-attended party has a better potential for awesomeness than a sparsely peopled event.

Arrive on time or within the established time frames set forth by your host. Similarly, unless you have been granted special access, plan on leaving at a reasonable time as well. Benjamin Franklin wrote in his *Poor Richard's Almanack* (1736) that "Fish & Visitors stink in 3 days." Don't be a stinker.

Try to contribute something. Don't show up empty-handed, expecting to be fed, entertained, and enlightened.

Workstations

Now it's time to talk about your workstations. Even if it's just you in the lab, having a separate workstation for each type of activity is a great idea. For example, you might have one workstation set up for programming a computer, another for soldering electronics, and another for working with power tools, such as a drill press or a saw.

Workstations dedicated to a particular task should have everything within easy reach, including tooling, supplies, power, and lighting. Don't waste time by forcing people to wander around the lab looking for the right drill bit or magnifying glass. Keep everything close so you can find what you need, when you need it.

Having a place for everything ahead of time is the best way to follow the wise adage, "A place for everything and everything in its place." You get to decide where everything lives. It doesn't hurt to go ahead and label everything conspicuously so that others can find their way around the lab.

If you have several people wanting to work on the same kinds of projects at the same time, you're going to need multiple, nearly identical workstations. The exact number depends on your available space and the number of people you expect to want to participate simultaneously, as well as your budget.

Most indoor-oriented work can often be performed at a work table. Exceptions include giant-scale mechatronics, pyrotechnics, and underwater robotics. Some of their constituent bits and pieces, however, can be crafted on an adequately prepared work table.

Make sure each work area has plenty of power outlets available. You're almost guaranteed to run short at some point. Take care not to overload your circuits, however. If you anticipate needing a lot more power than is presently provisioned, you're going to have to consider the real possibility of having your facility expanded by qualified, *licensed* professionals. These requirements will vary from one location to the next. Check with your local building codes to be sure.

As repeated often in previous chapters, you can never have *too much* lighting available, especially for the more detail-oriented chores. Typical indoor room lighting is just about adequate for finding your way out of the room, and that's about all. You'll need area lighting for each workstation, and possible additional task lighting, depending on your particular requirements.

Providing network access to each workstation can be challenging. Wireless networking is a good way to accommodate a relatively small number of users at once. More than a handful of computers trying to access the Internet at the same time can bring many wireless routers to a slow crawl. Dedicated high-speed Ethernet connectivity is more expensive and takes more effort to deploy, but offers much greater bandwidth to your lab-folk.

If you're especially cramped (or *dimensionally constrained*) in your lab, you'll need to get inventive. For example, a table placed in the center of the room can be approached from multiple sides. There's no reason all the workbenches have to be pushed up against a wall, allowing access from only a single side. Get creative!

Safety Planning

While it's tempting to fill up all the available space in a lab with workstations, especially when the budget permits it, do try to leave a bit of common area here and there. You always want to be able to walk *away* from a workstation with the same (or sometimes even more) rapidity you approached it with. Fire is an excellent example of such a motivator.

■ **Tip** Keep a properly selected and maintained fire extinguisher at each workstation, if possible.

Post emergency procedures prominently, where everyone, including visitors, can see. Have an evacuation plan, complete with map and "You are here" indicator posted at each workstation. Hopefully you will never need it.

Presentations

Are you excited about your latest project? Have you just learned a great new skill? Setting up your lab for group presentations is an excellent way to communicate these ideas to your friends, family, and coworkers.

Make your ~~victims~~ guests comfortable. Have adequate seating for everyone, with plenty of room between the rows (if you have more than one row of chairs). Ensure that everyone has a good view of what you intend to present.

Will the attendees be expected to take notes? Then perhaps you should set up tables where they can comfortably jot down some of the more spectacular aspects of your presentation. If you will be providing a handout or other documentation packages, make sure you have enough for the anticipated demand. Will you be entertaining questions or comments either during the presentation or afterward? Then make sure that not only can your audience see *you*, but that *you* can see *them*.

If you're going to be using some sort of projector for your presentation, by all means set aside some time to get to know the equipment ahead of time. Nothing ever seems to work just exactly as you imagined that it should. This seems to be an invariant law of presentation physics. No one is really sure why.

Your audience's time is just as valuable as your own. In fact, if there are only two people attending, then their collective time is worth approximately *twice* as much as yours. Think about it. Have all your ducks in a row, so to speak. Wasting time figuring out which cable goes where is not a good use of your time together.

Summary

Wow! You can really go crazy setting up a shared electronics lab, can't you?

Before you even flip on the light in your shiny new group lab, you should already have an insightful idea of who is going to be using it and why. You should also have some honest expectations of what your own expectations are going to be, and how you would like to interact with both the lab itself and its population.

The details, of course, are always going to be specific to your particular situation. Be on guard, however, for the future. It's going to sneak up on you, someday. Not too long ago, for example, it took a fairly hefty desk to keep a modern personal computer up off the floor. That kind of computing power now fits in a pocket. Think of what better uses you could dream up for that space these days!

That kind of forward thinking is exactly what turns the dreams of today into the reality of tomorrow. It's forward-thinking people like you, dear reader, that help map out those realities. Take the opportunity you have right now to spend some quality time daydreaming about how things *could* be. Then take those steps necessary to make it happen. Your family, your society, and your world expect great things from you. You can do it!

APPENDIX A

Getting Started with Tool Building

Once you've become comfortable in your lab and have had time to play with a few circuits, you might start to notice that some of the "tools" you've been using are just simple electronic circuits themselves. Good examples are power supplies and meters. These are great tool-building projects because you can see useful results early.

Let's take a look at building a couple of simple electronic tools that might be of use in your lab. Hopefully you will develop a better understanding of how these tools work. Ideally, you will progress from the basic question of "Does it work?" to the more involved questions of "How well does it work and what can I do to improve it?"

Please bear in mind that these exercises, while hopefully useful, are to illustrate what needs to go into the building of practical, reliable tools. At every stage you should be thinking about what design choices you have and which choices will enhance the usefulness of your tool-building process.

A Very Simple Power Supply

Almost all electronic projects need electricity to work. Having a good, dependable power supply at your disposal on your bench will get you started faster on new projects. You won't have to worry about dead batteries and flaky power adapters. See Figure A-1.

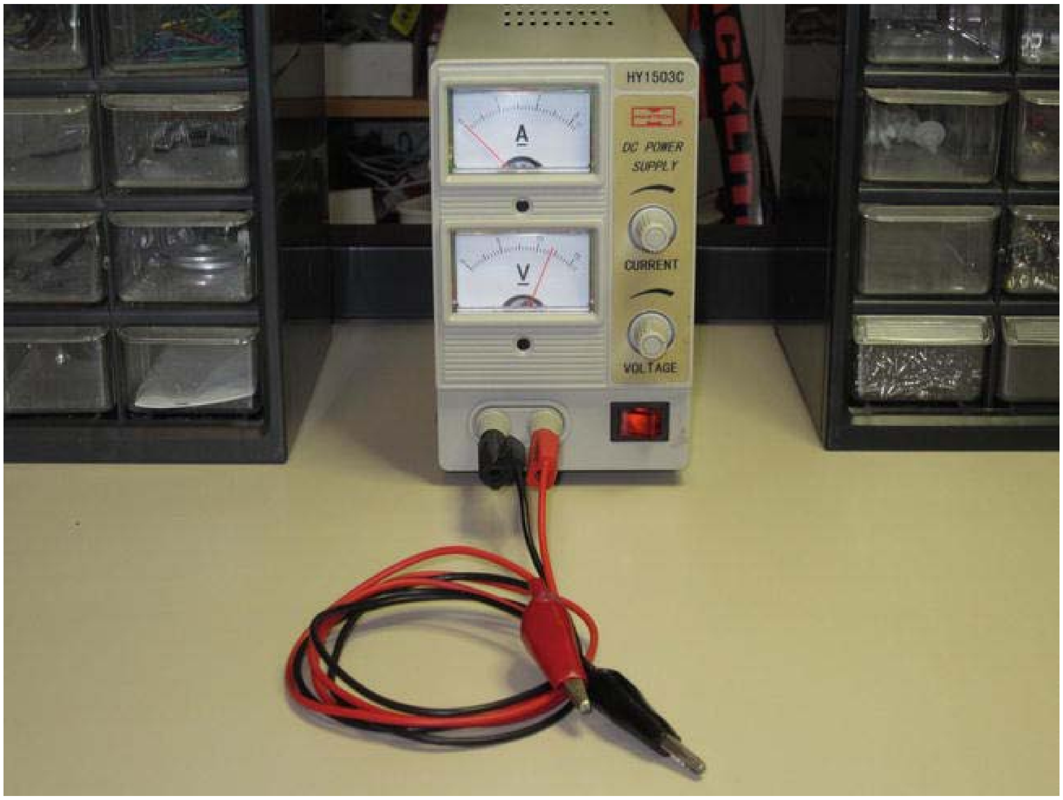

Figure A-1. *A professional bench power supply. This device allows both the voltage and the maximum current to be adjusted. It provides two dedicated meters to show how much power is flowing: one for voltage and the other for current. It also provides a modular set of plugs (banana plugs) for easily connecting and disconnecting power leads.*

A bench power supply such as the one shown in Figure A-1 has many features and includes extra circuitry to make it safe and easy to use. It will also set you back a good chunk of change (i.e., cost money). We're going to build something a bit simpler to start. Feel free to add more bells and whistles as your needs arise.

■ **Caution** Anything plugged into the wall has the potential to shock, burn, or kill you. Do not defeat their safety features!

An Easy 2.5 Watts

You probably have a spare USB charger lying around somewhere. If you don't already, you more than likely will at some point in the near future, as these are becoming more common with smaller electronic devices. If not, you can obtain one relatively cheaply if you do some smart shopping. The USB charger shown in Figure A-2 was obtained from overseas via eBay for less than $1. You will also need a USB cable about 3 or 4 feet long. Again, don't spend a lot of money on a new cable. If you don't already have one floating around in a drawer somewhere, you probably know someone that does.

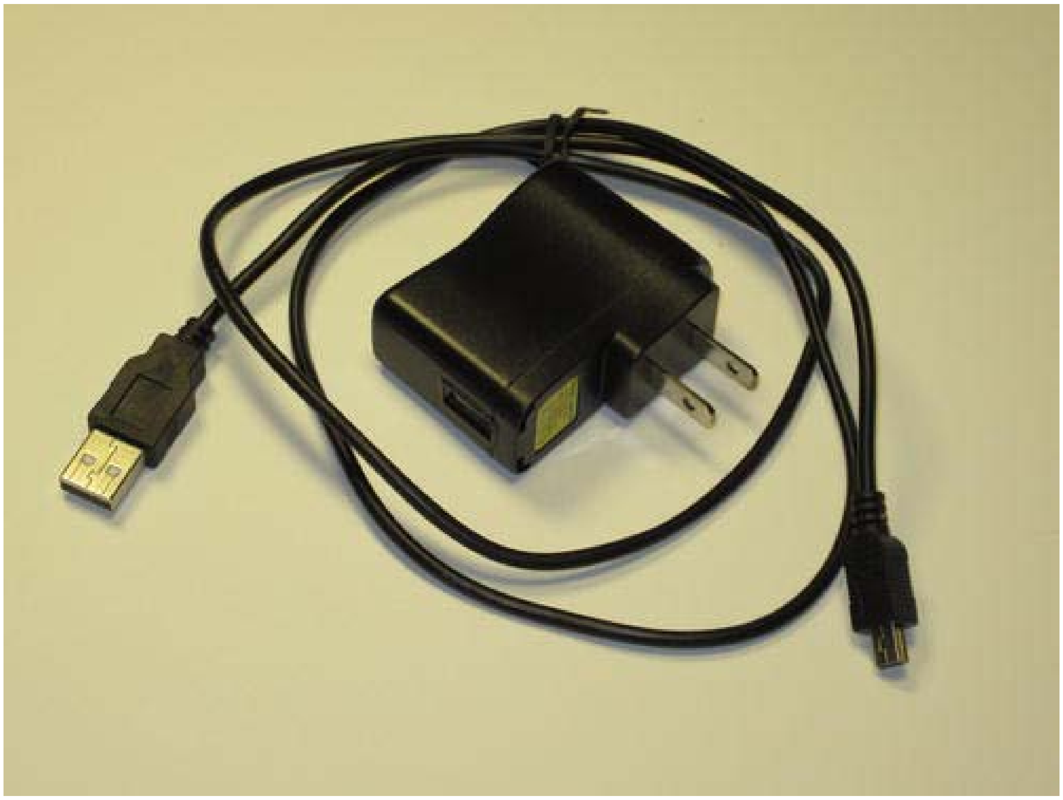

Figure A-2. *A USB charger that plugs into the wall and provides a regulated +5V supply. This plug is intended for US wiring, but will operate from 110AC to 240VAC at 50 or 60Hz, which pretty much covers the whole world, if you have the right adapter plug.*

It doesn't matter what type of USB cable you find, because we're going to cut off the smaller end and use the bare wires inside the cable. It just has to have the USB A–style plug shown on the left side of Figure A-2. The plug will fit into the socket of the charger. You do not need to open or modify the USB charger at all, and you are seriously advised not to do so for any reason.

Let's take a closer look at the label on the charger to find out exactly what it's supposed to be doing for us. See Figure A-3.

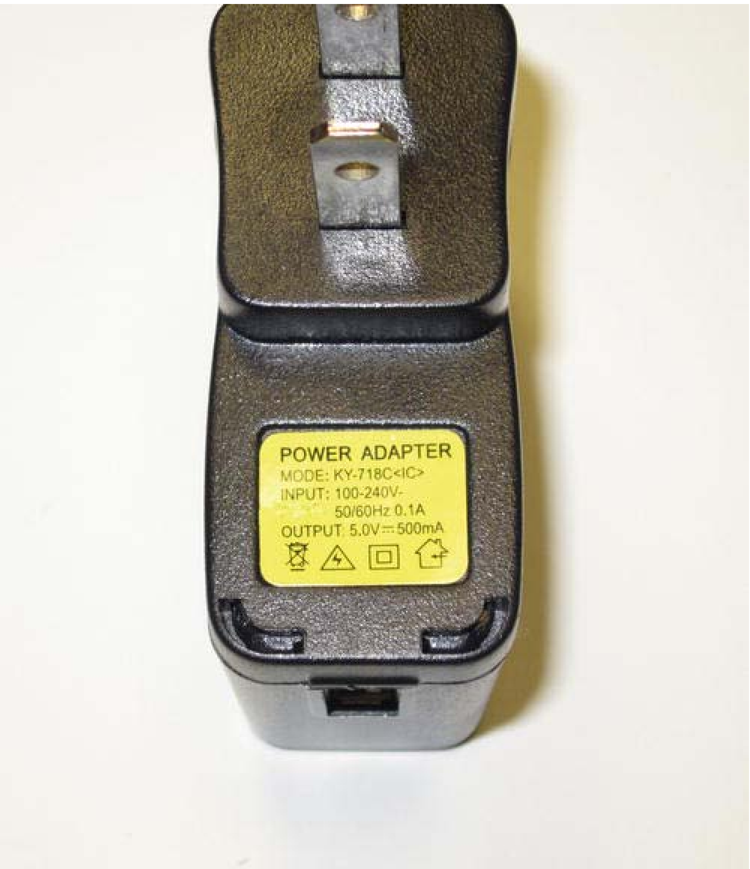

Figure A-3. *The label of the USB charger tells us that it accepts anything from 110VAC to 240VAC at either 50Hz or 60Hz. This wide range of input power allows the same internal electronic design to be used worldwide. Only the prongs that plug into the wall vary from location to location. The label also tells us that the USB charger will output +5V at up to 500mA. That's 2.5W!*

The "OUTPUT: 5.0V ⎓ 500mA" marking tells us what we need to know about the flavor of the power coming out of the adapter. The "5.0V" is the measure of the voltage. This is a very useful voltage level for small projects. Many digital chips require a 5V supply, although 3.3V is also becoming quite popular.

The ⎓ symbol represents *direct current (DC)* voltage. The symbol for *AC (alternating current)* voltage is the tilde, or wavy line symbol ~. The "500mA" is the maximum current that can be supplied by this charger (500 milliamps, or 0.5A), although you are cautioned not to expect a low-cost item such as this one to produce 100 percent of its rated power 100 percent of the time.

While this USB charger will most likely work for our intended purposes, it's best to keep those expectations in check. Note that "Model:" is misspelled ("Mode:"); and also note the absence of the magic letters *USB* or the USB logo. Also, there is a complete lack of any safety certifications present on this humble device, as befits its ~~cheap~~ economically optimized provenance. Compare this with the veritable constellation of safety certifications on the name-brand laptop charger shown in Figure A-4.

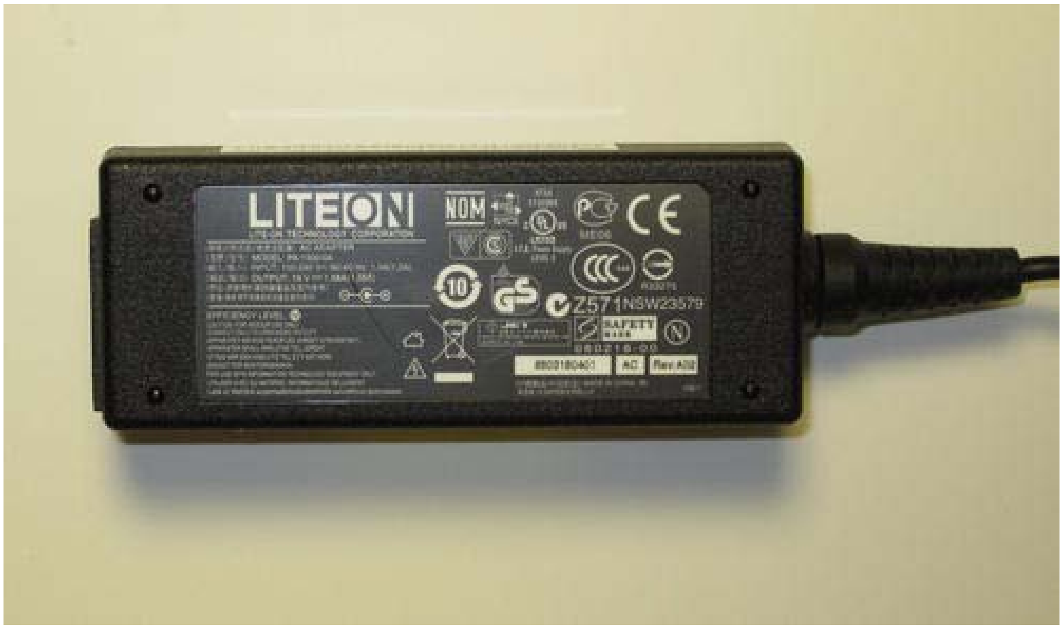

Figure A-4. *A name-brand laptop charger with several international safety certifications. Every one of these certifications costs money to obtain, and shows that the device was demonstrated to be safe for use in a typical home or business application.*

Next we adapt the USB cable to our needs. In the words of the Red Queen, “Off with his head!”

■ **Caution** Make sure the USB charger is not plugged into the wall or a power strip when making modifications to the cable.

First, plug the USB cable into the USB charger. Now cut off the *other* end of the cable. See Figure A-5.

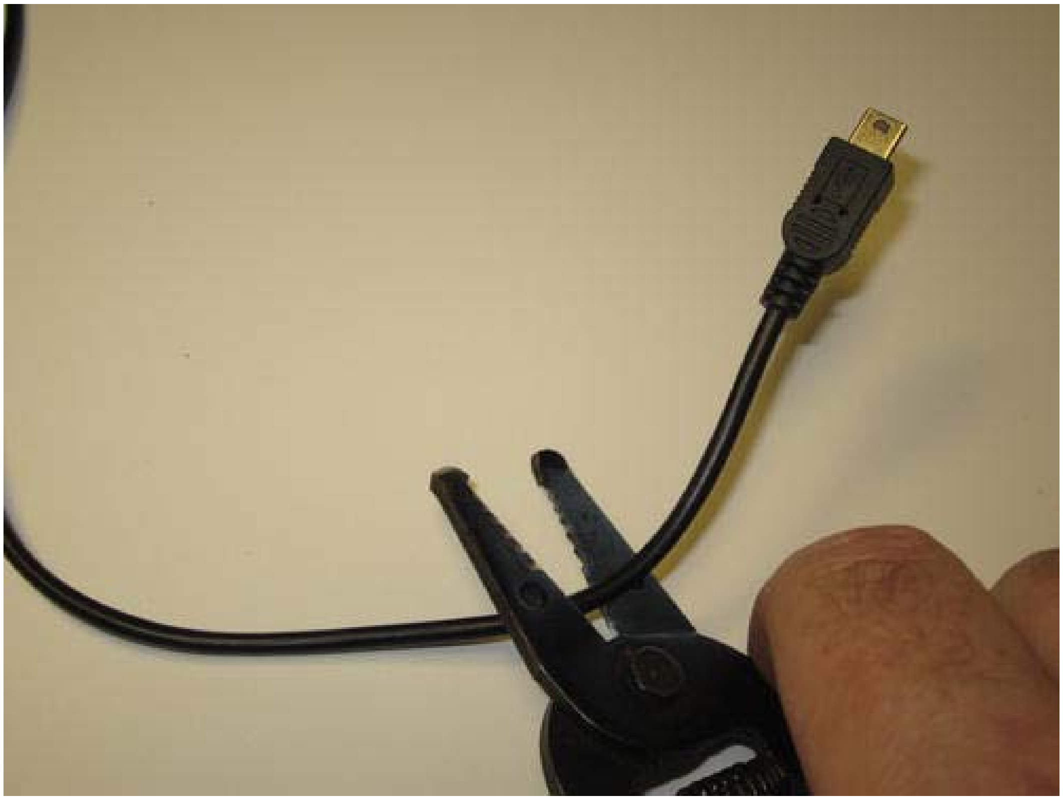

Figure A-5. *Cut off the device end of the USB cable. If you cut off the wrong end, you're not going to be able to plug it back into the charger, at least not without a considerable amount of effort. To ensure you cut the correct end, plug the host end of the cable into the USB charger before snipping, and then cut off the other end.*

Now carefully strip off about an inch of the cable's insulating jacket. Do your best to not nick the insulation on the wires inside. Within the jacket you should find four wires. Two of these wires carry the +5V and ground from the power supply. The other two wires carry the data on the USB connection. Since we're only interested in the power, we can ignore the other two wires for now. See Figure A-6.

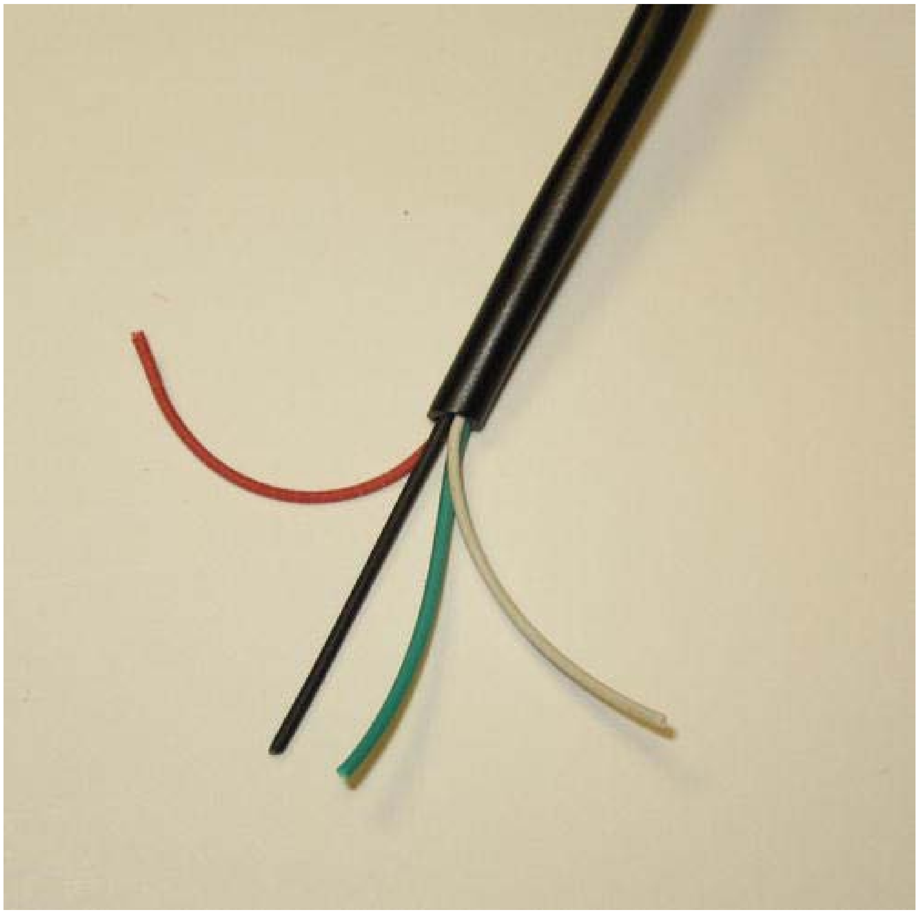

Figure A-6. *The four internal wires within the USB cable. From left to right they are red (VBUS or +5V), black (ground), green (USB D+), and white (USB D–). Use the red and black wires to pull a regulated +5V from the charger.*

The USB standard specifies every little detail about how to build and mark USB cables, even down to the colors of the internal wires in the cable that you're not ever supposed to see! That's dedication.

■ **Note** Apple violates the USB standard by not having a raised USB symbol molded into its cables.

However, sometimes cable manufacturers fudge a bit and use whatever four-conductor cable they happen to have lying around. A very common substitute cable uses red, blue, yellow, and brown wires internally. See Table A-1 for a list of which wires go where.

Table A-1. *Standard USB Wire Colors and Their Functions, Along with Popular Alternate Colors*

Function	Pin Number	Standard USB Color	Alternate Color
VBUS (+5V)	1	Red	Red
D– 2		White	Blue
D+ 3		Green	Yellow
GND 4		Black	Brown

Now is the time to verify the wiring before we get much further. Even if your cable has the official wire colors within the insulating jacket, you are strongly advised to double-check both the voltage and the polarity coming out of the charger before connecting it to your tender, sensitive circuits. Minimize your assumptions about the compliance of others. Testing proves testing works!

Strip about 1/4 inch of the insulation from the ends of the red and black wires. Get out your meter and set it up to read voltage at or above the 5V range. If your fancy meter has autoranging, then just set it to read DC voltage. Plug the USB cable into the charger, if it’s not already installed. Make sure the red and black wires are not touching, and then plug the charger into the wall or a convenient power strip.

Now measure the voltage between the red and black wires. Hopefully you should see a reading at or near 5.0V. The USB standard specifies a tolerance of ±10 percent, so anything between 4.5V and 5.5V is considered OK. See Figure A-7.

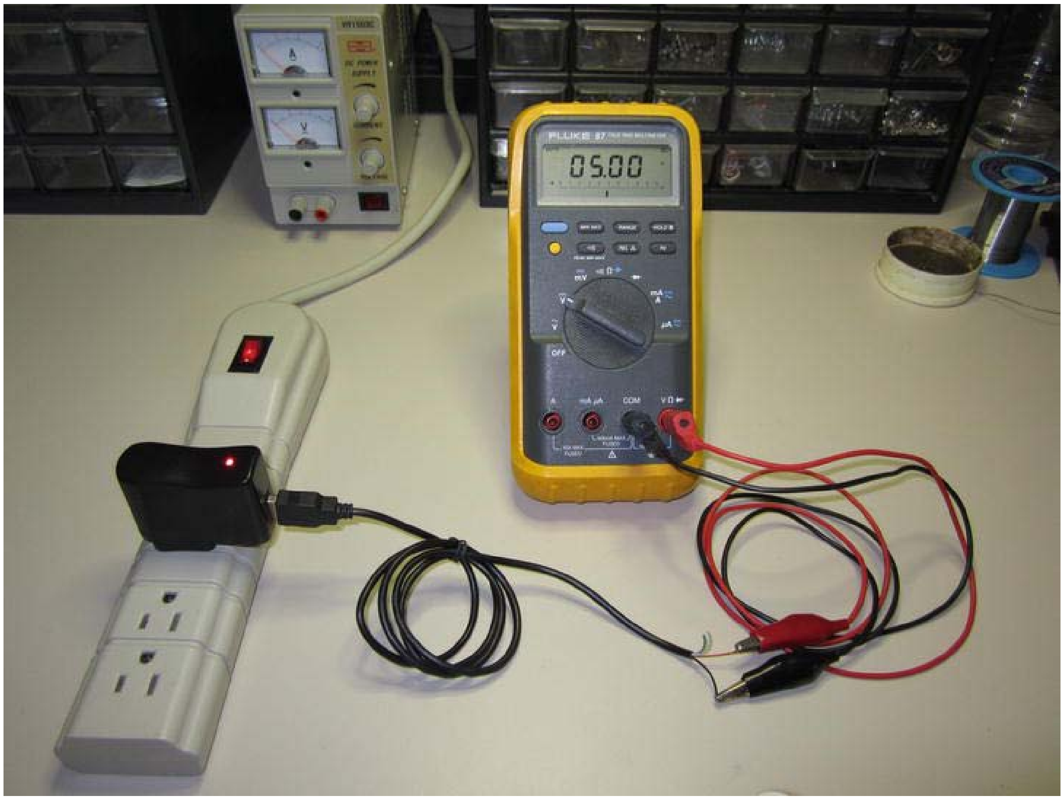

__Figure A-7.__ Measure the voltage coming from the USB charger before connecting it to any of your circuits. This one reads 5V on the dot! This is a much safer method than trial by fire (i.e., plugging it into your circuit and looking for a fire).

Now that you've confirmed that the red and black wires do indeed carry the regulated +5V from the USB charger, you can clip off the unused green and white wires. At this point you can either use the bare, stripped wires of the USB cable or solder on a more robust connector. Since this power supply is destined to be used with a solderless breadboard, a small, two-pin connector with the appropriate spacing will be attached. See Figure A-8.

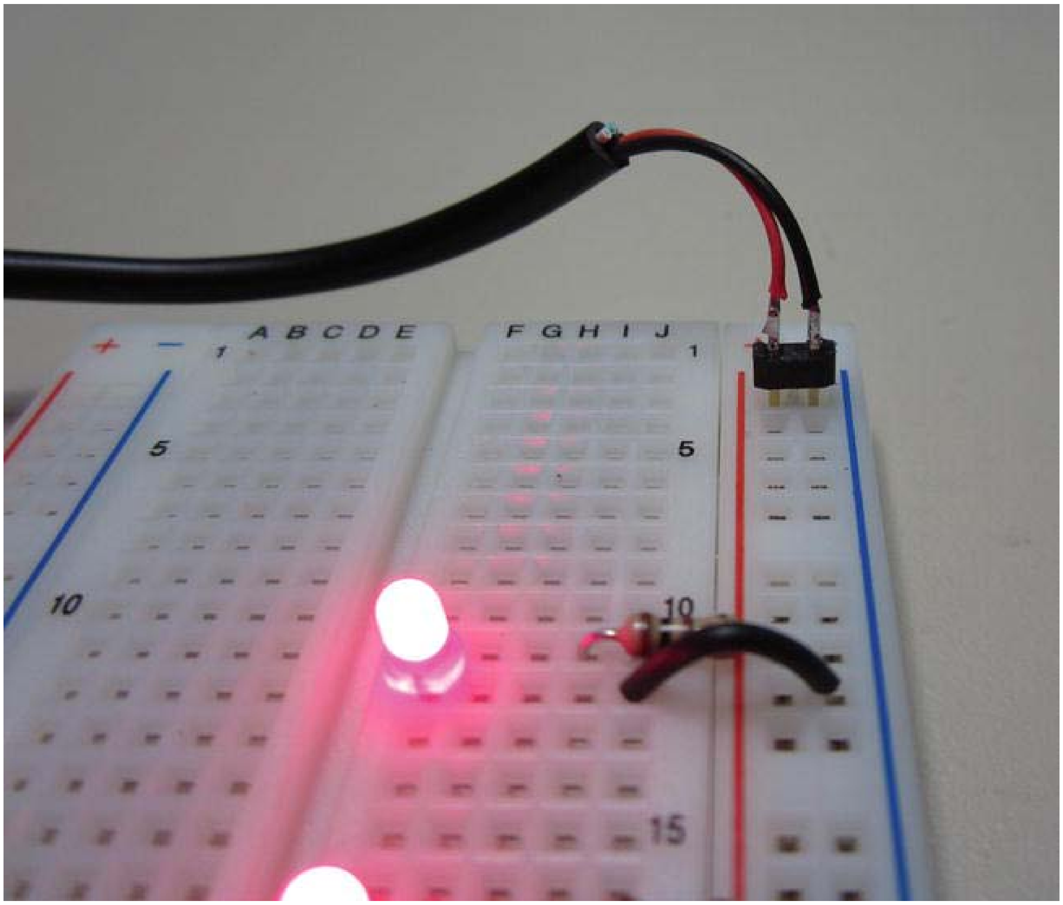

Figure A-8. *A gold-plated pin connector is soldered to the modified USB cable. The red and black insulation is left visible to be a polarity indicator when installing the cable to the solderless breadboard. Another happy LED!*

Adding Features

This power supply works just fine for most projects. If your requirements are, "Translate AC power from the wall into regulated DC power I can use for small projects, without spending a lot of money," then this contraption fits the bill precisely.

You can see for yourself in Figure A-8 that it does indeed work as advertised. Looking back at Figure A-7, you can see that the charger itself has a power-on indicator LED built into the housing. What it doesn't have is a power control or a switch. The power to the circuit is controlled by the switch on the power strip. This works just fine as long as you don't mind turning on and off everything else that happens to be plugged into the same power strip.

Various Indicators

Not all USB chargers have a power indicator, however. Some have both a power indicator LED as well as a *charge status indicator*, which can tell you at a glance if your USB device is fully charged or still charging. A charge status indicator is usually implemented by measuring the amount of current being drawn from the supply. If more than a preset current is flowing, the power supply assumes a battery is still being charged. When that current drops below a certain level, it assumes that the battery has been completely (or maybe just mostly) charged. Some chargers use a red or orange LED to indicate charging, and a green or blue LED to show that the battery is finished charging. Again, this type of charge status display is provided to give the user a quick status update with a single glance. You don't need to consult the product manual to interpret the status code, if you can remember that red means "charging" and green means "charged."

Let's rearrange some of the parts on the solderless breadboard to make room for some new features. See Figure A-9.

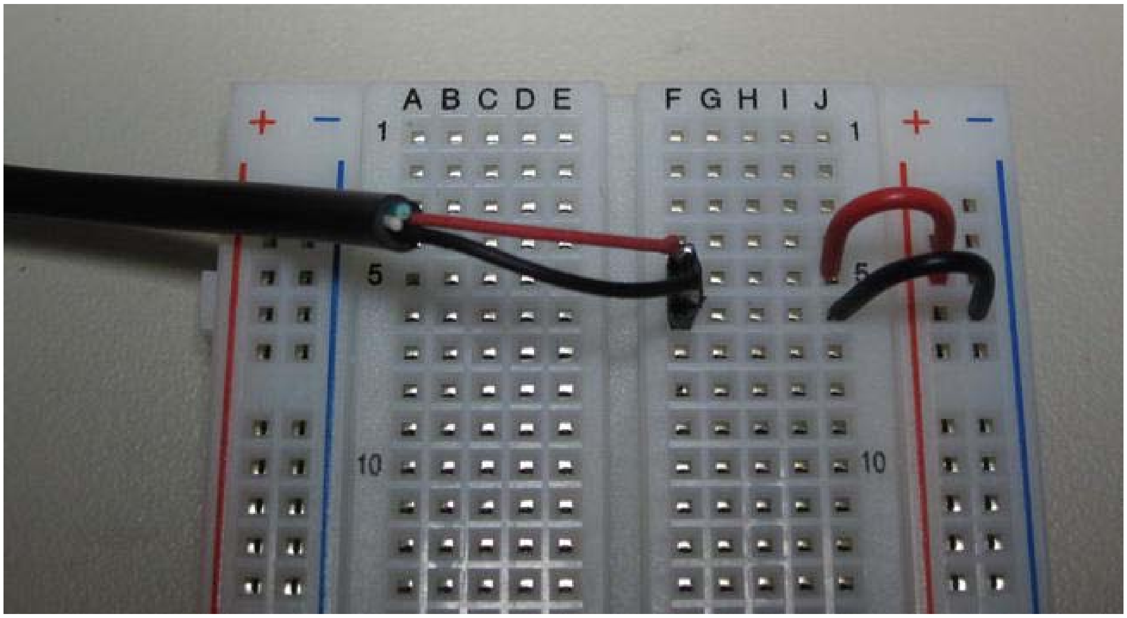

Figure A-9. *The regulated power coming from the USB charger is no longer directly connected to the breadboard's power rails. Instead, it is routed to some internal tie points and then connected to the power rails via two short jumpers.*

Power-On Indicator

First, we add a local power-on indicator by reinstalling the LED and current-limiting resistor used in the previous experiments. While this duplicates the function of any LEDs that might be present on the USB charger itself, it brings the indicator closer to the area of interest, which in this case is the circuitry onboard the solderless breadboard.

The anode of the LED goes into tie point F-3, with the cathode being installed at tie point F-4. The current-limiting resistor spans the gap between tie point I-3 and the right-hand positive (red) power rail. A short jumper connects tie point J-4 and the right-hand negative power rail. See Figure A-10.

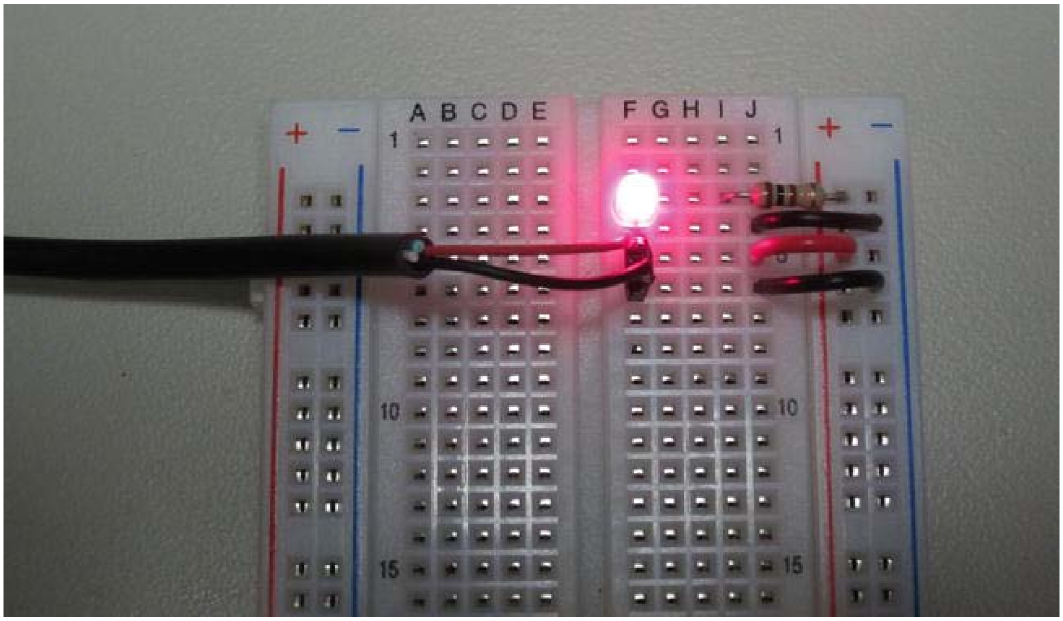

Figure A-10. *Add an LED and a current-limiting resistor across the power rails of the solderless breadboard to give an at-a-glance power indicator, right where the action is taking place.*

Now we have a good power-on indicator, right where we need it. It may or may not be redundant, depending on the features of your USB charger. This type of indicator is digital in the sense that it tells us that the circuit is either on or off, without giving us a lot of precise information about anything in between. For a lot of applications, however, this is more than enough information.

Power Supply Voltage

What if you *did* either want or need more precise information about the power supply? One method would be to permanently connect your multimeter to the circuit so that you can see how many volts are being delivered. That method assumes that you're not going to need your multimeter for anything else, ever.

A practical variation of that method, however, it to provide convenient *test points* in your circuit so that you can quickly and easily take a voltage reading when the need arises. Be sure to position the test points so that they are easy to access. Having a nice, big, readable label wouldn't hurt, either. Also, try to arrange things so that it's not likely that contact with a metal object, such as a tool or your wristwatch, will accidentally short out the power supply by touching both test points at the same time. Even simple "features" get complicated quickly, it seems!

If you want an accurate (and permanent) voltage reading on your circuit, consider adding a voltmeter. Small panel meters are relatively inexpensive and reasonably accurate. A DC voltmeter has two leads, positive and negative, that are attached across the voltage to be measured. See Figure A-11.

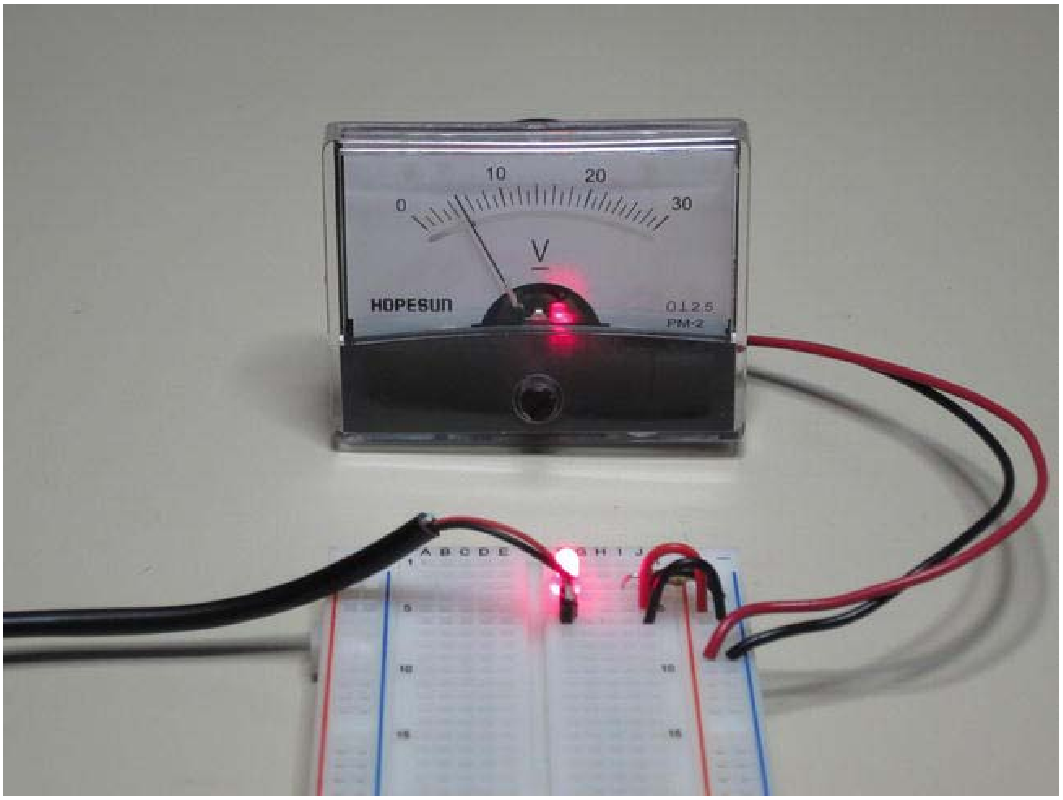

Figure A-11. *A DC voltmeter is attached to the circuit to accurately measure the supply voltage. The V on the meter's legend stands for voltage. The underline beneath the V indicates that it is designed to measure DC voltages, as opposed to AC voltages.*

The panel meter shown in Figure A-11 has a range of 0–30VDC. That's a bit more than we need, since our little power supply is only supposed to produce 5.0VDC ±10 percent. Since the meter's range does not accurately match the expected range of supply voltages, we lose a bit of precision when attempting to read the display. All the action is crowded together on the left side of the meter.

It might seem that a 0–5VDC meter would be preferable, but that would not allow for any overvoltage conditions at all. Remember, the USB standard allows for up to 5.5VDC to be delivered over the cable. A 0–6VDC meter, if available, would be more suitable. The meter introduced in Figure A-11, however, would be perfect when used with the laptop supply shown in Figure A-4, because its output voltage is 19.0V. That leaves a bit of room for overvoltage conditions and gives a wide enough swing of the needle to produce fine detail.

Power Supply Current

Measuring the *voltage* at any point in a circuit is easy. Just put a voltmeter between two points and voilà! In our previous example, we measured the VBUS power supply voltage with reference to GND, the ground potential of the circuit. Many voltage measurements are made with respect to ground.

To measure the current flowing through a circuit, we have two options. The first option is the most direct route. Install an ammeter (a meter that measures *amps* of current) inline with the circuit you want to measure. This requires that you *break* the circuit to make room for the meter. Once installed, the current flowing in the circuit actually flows through the ammeter as well.

The second option involves using a bit of trickery and some math. Instead of installing an ammeter directly inline with the current flow to be measured, we install a resistor of known resistance and measure the *voltage* that develops across the resistor.

Since we know that the values of volts, ohms, and amps are always related in any circuit, we can use Ohm's Law to determine how much current is flowing through a circuit by measuring the voltage across a known resistance.

To make things very easy, let's use a 1Ω resistor in series with the power supply, inserted in the circuit between where the positive lead of the power supply arrives at the circuit and where it is connected to the rest of the circuitry.

Unlike our previous use of Ohm's Law to calculate the missing value based on two known values, we only use *one* known value (the value of the resistor, called a *shunt resistor* when used in this manner). This one known value (resistance) then establishes the relationship between the two other values, voltage and current. As the current increases in the circuit, so does the voltage across the resistor. The two values will mirror each other—as one increases or decreases, the other will follow.

Since we used a 1Ω resistor in the circuit, this sets the relationship between current through the circuit and voltage across the resistor at a 1:1 ratio. For every ampere of current through the circuit, for example, 1V will be measured across the resistor.

Our power supply is only supposed to be able to deliver 0.5A of current, maximum, so we'll be dealing with voltages of less than 0.5V, unless something terribly unexpected happens.

Again, we could permanently attach a multimeter across the leads of a 1Ω resistor inserted into the circuit, or we could attach a voltmeter that is sensitive enough to give accurately readable information in the range we need. The 0–30VDC meter used before is much too large in range to do us any good here. We might see a tiny wiggle in the needle, but that's all.

Instead, we'll use a different kind of meter entirely. A digital voltmeter contains its own dedicated measurement circuitry and a numeric display to indicate the voltage reading. It also requires its own separate power supply. See Figure A-12.

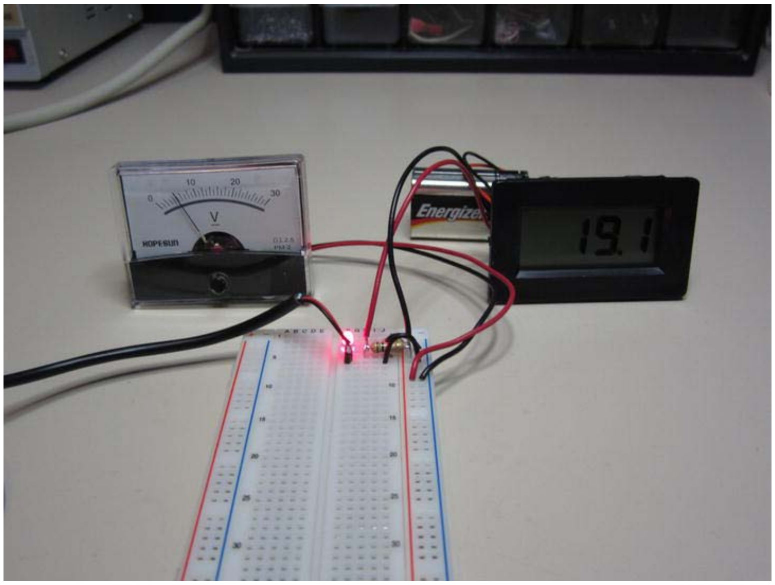

Figure A-12. *A digital voltage meter is used to read the voltage across the shunt resistor installed in the circuit. Because the shunt resistor has a 1.0Ω value, Ohm's Law dictates that the reading in millivolts on the digital meter correspond exactly to the number of milliamps flowing through the circuit. Math is useful, sometimes.*

The jumper wire connecting the incoming power supply directly to the power rail has been replaced with a 1Ω resistor. Leads from both ends of the resistor have been connected to the input terminals of the digital voltage meter on the right. The digital meter, unlike the analog meter on the left, requires an independent power supply to operate, in this case a 9V battery.

So now we've added some features to our simple power supply, including a power-on indicator, a supply voltage reading, and a supply current reading.

These are very useful features to have on any power supply. However, at this point in the prototyping stage, we've just barely covered what the added circuitry is supposed to do. Packaging all this up into an easy-to-use and reliable device takes as much if not more thought and effort.

First, you have to rebuild this entire circuit in a more permanent fashion. Next, you will have to wedge all that clever gadgetry into some sort of enclosure to protect the delicate bits while still permitting access to the inputs, outputs, and controls, if any. Then you get to label everything so that you can remember how to use it in the future.

Summary

So now you just might be developing a certain respect for how much work goes into building even the simplest tools—tools that are useful, accurate, reliable, and an asset to your toolbox. Remember, every tool on your workbench has to earn its place there. This applies to tools you design and build yourself, as well.

Good luck with your tool-building projects. Sometimes it's just as much fun to work on your tools as it is to work *with* your tools.

Index

A

Active electrical and electronic components
  batteries, 81
  diodes, 82–83
  integrated circuits, 84
  LEDs, 83
  power supplies, 81
  transformers, 82
  transistors, 83–84
Active-participant model, 137
American wire gauge (AWG), 26–28, 60–61
Ammeters, 49
Anarchic chaos, 140
AND circuit, 124
Arduino-compatible microcomputer, 67
AWG. *See* American wire gauge (AWG)

B

Barrier strip, 30–31
Bipolar junction transistors (BJTs), 83, 84
Bolts, 33
Budgeting, 14–15

C

Capacitors
  definition, 76
  electrolytic capacitors, 77
  schematic symbols, 78
  in series and parallel, 130
Charge status indicator, 155
Coin cell
  electrical goodness measurement, 94
  lithium
    installed backward, 92
    LED lights up, 91–92
Components, electric and electronic circuits
  acquiring components, 84
    new components purchase, 85
    recycle, used components, 85–86
  active components, 81
    batteries, 81
    diodes, 82–83
    integrated circuits, 84
    LEDs, 83
    power supplies, 81
    transformers, 82
    transistors, 83–84
  conductors and interconnects
    connectors, 64–68
    description, 57
    electrical schematics, 57–58
    wire and cables, 58–64
  data sheets, 86
  passive components
    capacitors, 76–78
    inductors, 78–79
    resistors, 68–76
    switches, 79–80
Computers, 35–36
Conductors and interconnects
  connectors
    clip leads, 65
    description, 64
    solderless breadboards, 66–68
  description, 57
  electrical schematics, 57–58
  wire and cables
    conductivity, 61–62
    description, 58–59
    longer version, 60–61
    Ohm's law, 62–63
    short version, 59–60
    storing, 63–64
    testing, 63
Connectors
  clip leads, 65
  description, 64
  solderless breadboards

Connectors, solderless breadboards (*cont.*)
        Arduino-compatible microcomputer, 67
        description, 66
        prototype LED display, 68
        tie points, 66
Cozy corner lab
    capacitors in series and parallel, 130
    inductors in series and parallel, 130
    LEDs, in series and parallel, 131–133
    parallel and series circuits
        assembly instructions, 119
        batteries in, 122–123
        components, 118
        in different view, 120–121
        schematic diagram, 118
        troubleshooting, 119–120
    rehabilitation, work station, 105–106
        components, 115–116
        divide, 107–109
        lighting, 112
        power source, 110–111
        taxonomy, 106–107
        tools, 112–114
        under-desk environment, 115
    resistors
        in parallel, 129
        in series, 126
    switches
        in parallel, 124
        in series, 123
        three-way, 125–126

D

Dallas Personal Robotics Group (DPRG), 17, 101–102, 139
DC voltmeter, 156–157
Digital caliper, 46, 47
Digital storage oscilloscope (DSO), 53–54
Digital voltage meter, 159
Diodes, 82–83
Double-throw switches, 80
DPRG. *See* Dallas Personal Robotics Group(DPRG)
Drawing, 23
Drills, 40–43
DSO. *See* Digital storage oscilloscope (DSO)

E, F, G

Electrical experiment
    LED
        defective, 95
        installed backward, 92–93
        lights up when attached to lithium coin cell, 91–92
        on solderless breadboard, 95–98
    parallel and series circuits, 117–118
Electrolytic capacitors, 77
Electronic measurement tools
    ammeters, 49
    ohmmeters, 49
    voltmeters, 47–48
Electronics lab
    and classroom (*see* Lab and classroom)
    budgeting, 14–15
    cautionary tale, 17
    electricity,working principles
        practical approach, 9
        theoretical approach, 7–8
    environmental impact, 14
    flashlight, 11–13
    fundamental components, 4–5
    goals, 6
    ideas, 5–6
    of Harold Timmis, 16
    planning, 2
    portable (*see* Portable mini-lab)
    Robot Clubhouse, 17
    safety, 10
    space for, 3
    tools, 3–4 (*see also* Tools)
Environmental impact, 14

H

Hackerspaces, 139
Hand tools, 114
    computers, 35–36
    pliers and tweezers, 29–30
    screwdriver, 30–33
    screws, bolts and nuts, 33
    wire crimpers, 34–35
    wire cutters, 23–25, 29
    wire strippers, 25–29
Harold Timmis' lab, 16

I, J, K

Indicators, 155
Inductors
  DC motors, 79
  description, 78
  usage, 79
Infrared LEDs, 132
Inspection equipment, 45–46
Insulators, 8
Integrated circuits, 84

L

Lab and classroom
  group
    coworkers and fellow students, 138
    family, 137
    friends, 138
    general public, 139–140
    special-interest groups, 139
  presentations, 143
  role
    facilitator/coordinator, 135
    observer, 137
    participant, 137
    teacher/presenter, 136
  shared resource
    courteous guest, 141
    hospitable host, 140–141
  workstations
    description, 142
    safety planning, 142
Laboratory set up. *See* Electronics lab
Laptop charger, 149
Laser cutters, 43
Light-emitting diodes (LEDs), 83
  care and feeding, 127–128
  circuit on solderless breadboard, 95–98
  curiosities, 126–127
  defective, 95
  installed backward, 92–93
  in series and parallel, 131–33
  lights up when attached to lithium coin cell, 91–92
  self-control, lack of, 127
  uses, 100

M

Magnifying lens, 46
Metal-oxide–semiconductor field-effect transistor (MOSFET), 84
Modern digital storage oscilloscope, 52–53
Mobile electronics lab. *See* Portable mini-lab
MOSFET. *See* Metal-oxide–semiconductor field-effect transistor (MOSFET)
Multimeters, 50–51, 94

N

Noisebridge organization, 140
Nuts, 33

O

Ohmmeters, 49
Ohm's law
  description, 49
  application, 49, 62–63
OR circuit, 124
Oscilloscopes
  functions of, 51
  modern, 52–54
  tiny, 54–55

P, Q

Parallel and series circuits
  assembly instructions, 119
  batteries in, 122–123
  in different view, 120–121
  schematic diagram, 118
  troubleshooting, 119–120
Passive electrical and electronic components
  capacitors, 76–78
  inductors, 78–79
  resistors, 68–76
  switches, 79–80
Pencil-grip style, soldering irons, 39
Permanent circuit, 95–97
Physical measurement devices, 46–47
Pliers and tweezers, 29–30
Portable mini-lab
  advantages, 88
  compact executive model

Portable mini-lab, compact executive model (*cont.*)
        electrical experiments (*see* Electrical experiments)
        lab deployed on desktop, 89–90
    components, 89
    DPRG, 101–102
    features, 89
    robotics lab, 103
    travel, 87
Potentiometer, 74–75
Power-on indicator, 155
Power rail attachment, 99
Power supplies, 55–56, 81–82, 145–146
Power tools
    3D printers, 44
    drills, 40–43
    laser cutters, 43
Professional bench power supply, 146

R

Resistors
    color codes, 72–73
    description, 68
    location, 95
    in parallel, 129
    in permanent circuit, 97
    power ratings, 73
    schematic symbols and reference designators, 70–71
    in series, 126
    specialty resistors, 75–76
    standard values and tolerances, 71–72
    through-hole and surface-mount resistors, 69–70
    variable resistors, 74–75
RIGOL DS1052E oscilloscope, 54
Robot Clubhouse, 17

S

Shunt resistor, 158
Single-pole double-throw (SPDT) switch, 125
*Si*ngle-throw switches, 80
Soldering
    definition, 4
    instructions for, 37–38
    tools for, 113
Solderless breadboards, 66
    Arduino-compatible microcomputer, 67
    description, 66
    prototype LED display, 68
    tie points, 66
    types of, 117
SPDT switch. *See* Single-pole double-throw (SPDT) switch
Stereo microscope, 45
Surface-mount resistors, 69–70
Switches
    description, 79
    double-throw switch, 80
    in parallel, 124
    in series, 123
    single-throw switch, 80
    three-way, 125–126

T

Tablet computers, 91
Tabletop drill press, 42
Tektronix 475A oscilloscope, 52
Temperature-controlled soldering station, 113
Terminal block, 31–32
Test and measurement equipment
    multimeters, 50–51
    oscilloscopes
        functions of, 51
        modern, 52–54
        Tektronix 475A, 52
        tiny, 54–55
    power supplies, 55
3D printers, 44
Three-way switch, 125–126
Through-hole resistors, 69–70
Tools
    earmuffs, 22
    electronic measurement
        ammeters, 49
        ohmeters, 49
        voltmeters, 47–48
    eye protection, 21
    for soldering, 37–40
    hand
        computers, 35–36
        pliers and tweezers, 29–30
        protection, 22
        screwdriver, 30–33
        screws, bolts and nuts, 33
        wire crimpers, 34–35

        wire cutters, 23–25, 29
        wire strippers, 25–29
    hearing protection, 22
    indicators, 155
    inspection equipment, 45–46
    mind, 19–21
    physical measurement devices, 46–47
    power
        3D printers, 44
        drills, 40–43
        laser cutters, 43
    power-on indicator, 155
    power supply, 145–46
    USB charger
        description, 147–148
        voltage measurement, 149–153
Transformers, 82
Transistors
    BJTs, 83
    description, 83
    MOSFET, 84
Troubleshooting, 2

U

Unrepairable subassemblies, 127
USB charger
    cable
        colors and functions, 152
        internal wires, 151
    description, 147–148
    regulated power from, 155
    voltage measurement, 149–153

V

Variable resistors, 74–75
Voltmeters, 47–48

W, X, Y, Z

Weller controlled-temperature soldering station, 37
Wire and cables
    conductivity, 61–62
    description, 58–59
    longer version, 60–61
    Ohm's law, 62–63
    short version, 59–60
    storing, 63–64
    testing, 63
Wire crimpers, 34–35
Wire cutters, 23–25, 29
Wire strippers, 25–29
Wood screws, 33
Workshop. *See* Electronics lab

Building Your Own Electronics Lab

■ ■ ■

Dale Wheat

Apress®

Building Your Own Electronics Lab

Copyright © 2012 by Dale Wheat

All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval system, without the prior written permission of the copyright owner and the publisher.

ISBN-13 (pbk): 978-1-4302-4386-1

ISBN-13 (electronic): 978-1-4302-4387-8

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

President and Publisher: Paul Manning
Lead Editor: Matt Moodie
Technical Reviewer: Cliff Wooton
Editorial Board: Steve Anglin, Mark Beckner, Ewan Buckingham, Gary Cornell, Morgan Ertel, Jonathan Gennick, Jonathan Hassell, Robert Hutchinson, Michelle Lowman, James Markham, Matthew Moodie, Jeff Olson, Jeffrey Pepper, Douglas Pundick, Ben Renow-Clarke, Dominic Shakeshaft, Gwenan Spearing, Matt Wade, Tom Welsh
Coordinating Editor: Brigid Duffy
Copy Editors: Damon Larson
Compositor: Bytheway Publishing Services
Indexer: SPi Global
Cover Designer: Anna Ishchenko

Distributed to the book trade worldwide by Springer Science+Business Media, LLC., 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail `orders-ny@springer-sbm.com`, or visit `www.springeronline.com`.

For information on translations, please e-mail `rights@apress.com`, or visit `www.apress.com`.

Apress and friends of ED books may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Special Bulk Sales–eBook Licensing web page at `http://www.apress.com/bulk-sales`.

The information in this book is distributed on an "as is" basis, without warranty. Although every precaution has been taken in the preparation of this work, neither the author(s) nor Apress shall have any liability to any person or entity with respect to any loss or damage caused or alleged to be caused directly or indirectly by the information contained in this work.

Any source code or other supplementary materials referenced by the author in this text is available to readers at `www.apress.com`. For detailed information about how to locate your book's source code, go to `http://www.apress.com/source-code/`.

To my father, John H. Wheat, who patiently taught me about electricity and electronics over 40 years ago. To this day, I remain fascinated by it. My life has been enriched in so many ways by the many things that I learned from my father.

Contents

- **About the Author** xi
- **About the Technical Reviewer** xii
- **Acknowledgments** xiii
- **Introduction** xiv
- **Chapter 1: Planning Your Electronics Workshop** 1
  - What to Expect 1
  - It Starts with a Plan 2
  - A Broad Outline 2
    - A Little Time 2
    - A Little Space 3
    - Basic Tools 3
    - Fundamental Components 4
    - Ideas 5
    - Goals 6
    - Inspiration! 7
  - How Electricity Works 7
    - The Theoretical Approach 7
    - A Practical Approach 9
  - Safety 10
  - Lighten Up! 11
  - Environmental Impact 14

Budgeting 14
Buying New 14
Finding Used 15
Harvesting or Recycling 15
Some Example Workshops 16
A Fresh Start 16
The Robot Clubhouse 17
A Cautionary Tale 17
Summary 18
Chapter 2: Building Your Tool Chest 19
The Most Important Tool Ever 19
Care and Feeding of Your Brain 21
All Those Other Bits, Too 21
Basic Hand Tools 22
Wire Cutters and Wire Strippers 23
Pliers and Tweezers 29
Screwdrivers, Nut Drivers, Wrenches, and Ratchets 30
Screws, Bolts, and Nuts 33
More Specialized Hand Tools 34
Wire Crimpers 34
Computers 35
Tools for Soldering 37
Power Tools 40
Drills and Drill Bits 40
Laser Cutters and 3D Printers 43
Basic Test and Measurement Equipment 44

Physical Measurement Devices ... 46
Simple Electronic Measurement Tools ... 47
Advanced Test and Measurement Equipment ... 49
Multimeters ... 50
Oscilloscopes ... 51
Power Supplies ... 55
Organizing Your Tools ... 56
Summary ... 56
■ **Chapter 3: Components ... 57**
Conductors and Interconnects ... 57
Electrical Schematics ... 57
Wire and Cables ... 58
Connectors ... 64
Passive Components ... 68
Resistors ... 68
Capacitors ... 76
Inductors ... 78
Switches ... 79
Active Components ... 81
Power Supplies ... 81
Diodes ... 82
LEDs ... 83
Transistors ... 83
Integrated Circuits ... 84
Acquiring Components ... 84
Buying New Components ... 85
Recycling Used Components ... 85

Component Data Sheets 86
Summary 86

■ **Chapter 4: A Portable Mini-Lab 87**

Have Lab, Will Travel 87
The Compact Executive Model 87
No Soldering Required 90
The Ever-Shrinking Computing Device 91
Electrical Experiments You Can Perform 91
Lighten Up, 20th Century Style 91
A Slightly More Permanent Circuit 95
Multiple Uses for the LED 100
The Special Project Portable Laboratory 101
Summary 104

■ **Chapter 5: The Cozy Corner Lab 105**

The Rehabilitation of a Small but Useful Area 105
Science to the Rescue! 106
Organization 107
The Great Divide 107
Adding a Power Source 110
A Clean, Well-Lighted Space 111
Tools 112
Under the Desk 115
Components 115
Where to Go from Here 116
Projects 116
Parallel and Series Circuits 117
Batteries in Series and Parallel 122

Switches in Series 123
Switches in Parallel 124
The Three-Way Switch 125
Resistors in Series 126
Resistors in Parallel 129
Capacitors in Series and Parallel 130
Inductors in Series and Parallel 130
LEDs in Series and Parallel 131
Summary 133
Chapter 6: The Small Group Lab and Classroom 135
Know Yourself 135
Facilitator or Coordinator 135
Teacher or Presenter 136
Participant 137
Observer 137
Know Your Group 137
Your Family 137
Your Friends 138
Coworkers and Fellow Students 138
Special-Interest Groups 139
The General Public 139
A Shared Resource 140
Be a Hospitable Host 140
Be a Courteous Guest 141
Workstations 142
Safety Planning 142
Presentations 143
Summary 143

Appendix A: Getting Started with Tool Building 145
A Very Simple Power Supply 145
An Easy 2.5 Watts 147
Adding Features 154
Various Indicators 155
Power-On Indicator 155
Power Supply Voltage 156
Power Supply Current 158
Summary 159
Index 161

About the Author

Dale Wheat is a full-time freelance writer, specializing in electronics and embedded systems. He has written several articles for technical and hobbyist magazines such as *Circuit Cellar*, O'Reilly's *MAKE* magazine, and *Elektor*. He teaches classes on electronics, microcontrollers, and soldering skills. He designs and sells DIY electronics kits from his web site, `http://dalewheat.com`. Before becoming a full-time writer, Dale consulted as a computer programmer and systems analyst for several companies, including IBM, MCI, and GTE (now Verizon). Dale is a two-term past president of the Dallas Personal Robotics Group, the world's oldest personal robotics club. He is a member of the National Honor Society, Phi Theta Kappa, and intends to continue his education as long as they keep the doors open. He lives with his wife, Anne, near Dallas, Texas.

About the Technical Reviewer

■ Cliff Wootton is a former Interactive TV systems architect at BBC News. The "News Loops" service developed there was nominated for a BAFTA and won a Royal Television Society Award for Technical Innovation. An invited speaker on pre-processing for video compression at the Apple WWDC conference. Taught post graduate MA students about real-world computing, multimedia, video compression, metadata and researching the deployment of next generation interactive TV systems based on open standards.

Currently working on R&D projects investigating new Interactive TV technologies, involved with MPEG standards working groups, writing more books on the topic and speaking at conferences when not lecturing on Multimedia at the University of the Arts in London.

Acknowledgments

This book was Michelle Lowman’s idea. I just wrote down some of the words.

Corbin Collins and Brigid Duffy, coordinating editors at Apress, helped keep me on track with this project, despite all my best efforts. Cliff Wootton and Alan Smith provided excellent insight and suggestions to help the chapters along. The entire Apress editorial and production teams make what could be a hard job into a thing of joy.

My wife, Anne, continues to buy into the fantasy that I am a “professional writer,” giving me the time and space I need to do this. Her patience, love, support, and trust make all the difference in the world.

My good neighbor and friend Jerry Johnson’s fervent prayers and constant encouragement and enthusiasm really helped pull this project through the difficult times that were encountered. While I was complaining about “narrative pacing and clarity,” he underwent brain surgery (twice) and chemotherapy for cancer, always with a smiling laugh and a twinkle in his eye.

VOL. 1

Charles Platt

Encyclopedia of Electronic Components

Power Sources & Conversion

Resistors ▪ Capacitors ▪ Inductors
Switches ▪ Encoders ▪ Relays ▪ Transistors

O'REILLY®

Make:
makezine.com

Encyclopedia of Electronic Components
Volume 1

Charles Platt

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Encyclopedia of Electronic Components Volume 1

by Charles Platt

Copyright © 2013 Helpful Corporation. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Brian Jepson
Production Editor: Melanie Yarbrough
Proofreader: Melanie Yarbrough
Indexer: Judy McConville

Cover Designer: Mark Paglietti
Interior Designer: Edie Freedman and Nellie McKesson
Illustrator: Charles Platt
Photographer: Charles Platt
Cover Production: Randy Comer

October 2012: First Edition

Revision History for the First Edition:

2012-10-03 First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449333898* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Encyclopedia of Electronic Components Volume 1*, the cover images, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-33389-8

[TI]

To Mark Frauenfelder, who reacquainted me with the pleasures of Making.

Table of Contents

Preface . . . . **xix**

1. How to Use This Book **1**
- Reference vs. Tutorial 1
- Theory and Practice 1
- Organization 1
- Subject Paths 2
- Inclusions and Exclusions 2
- Typographical Conventions 3
- Volume Contents 3
- Safari® Books Online 3
- How to Contact Us 4

> POWER

> > SOURCE

2. Battery **5**
- What It Does 5
- How It Works 6
  - Electrode Terminology 7
- Variants 7
  - Disposable Batteries 8
  - Rechargeable Batteries 9
- Values 11
  - Amperage 11
  - Capacity 11
  - Voltage 13
- How To Use It 14

What Can Go Wrong 15
  Short Circuits: Overheating And Fire 15
  Diminished Performance Caused By Improper Recharging 15
  Complete Discharge Of Lead-Acid Battery 15
  Inadequate Current 15
  Incorrect Polarity 15
  Reverse Charging 16
  Sulfurization 16
  High Current Flow Between Parallel Batteries 16

> > CONNECTION

3. Jumper 17
What It Does 17
How It Works 17
Variants 18
Values 18
How To Use It 19
What Can Go Wrong 19

4. Fuse 21
What It Does 21
How It Works 21
Values 22
Variants 22
  Small Cartridge Fuses 23
  Automotive Fuses 23
  Strip Fuses 24
  Through-Hole Fuses 24
  Resettable Fuses 24
  Surface Mount Fuses 26
How To Use It 26
What Can Go Wrong 27
  Repeated Failure 27
  Soldering Damage 27
  Placement 28

5. Pushbutton 29
What It Does 29
How It Works 29
Variants 30
  Poles And Throws 30
  On-Off Behavior 30
  Slider 31
  Styles 31
  Termination And Contact Plating 32
  Mounting Style 32
  Sealed Or Unsealed 32

    - Latching . . . 33
    - Foot Pedal . . . 33
    - Keypad . . . 33
    - Tactile Switch . . . 34
    - Membrane Pad . . . 34
    - Radio Buttons . . . 35
    - Snap-Action Switches . . . 35
    - Emergency Switch . . . 35
  - Values . . . 35
  - How To Use It . . . 35
  - What Can Go Wrong . . . 35
    - No Button . . . 35
    - Mounting Problems . . . 35
    - LED Issues . . . 36
    - Other Problems . . . 36

8. Switch . . . 37

  - What It Does . . . 37
  - How It Works . . . 37
  - Variants . . . 38
    - Terminology . . . 38
    - Poles And Throws . . . 38
    - On-Off Behavior . . . 39
    - Snap-Action . . . 39
    - Rocker . . . 40
    - Slider . . . 40
    - Toggle . . . 41
    - DIP . . . 43
    - SIP . . . 44
    - Paddle . . . 44
    - Vandal Resistant Switch . . . 45
    - Tactile Switch . . . 45
    - Mounting Options . . . 45
    - Termination . . . 45
    - Contact Plating Options . . . 45
  - Values . . . 45
  - How To Use It . . . 46
    - Power Switches . . . 46
    - Limit Switches . . . 46
    - Logic Circuits . . . 47
    - Alternatives . . . 47
  - What Can Go Wrong . . . 47
    - Arcing . . . 47
    - Dry Joints . . . 48
    - Short Circuits . . . 48
    - Contact Contamination . . . 48
    - Wrong Terminal Type . . . 48
    - Contact Bounce . . . 48

    - Mechanical Wear 48
    - Mounting Problems 48
    - Cryptic Schematics 49

7. Rotary Switch 51

- What It Does 51
- How It Works 52
- Variants 52
    - Conventional 52
    - Rotary DIP 53
    - Gray Code 54
    - PC Board Rotary Switch 55
    - Mechanical Encoder 55
    - Pushwheel And Thumbwheel 55
    - Keylock 55
- Values 56
- How To Use It 56
- What Can Go Wrong 57
    - Vulnerable Contacts 57
    - Contact Overload 57
    - Misalignment 57
    - Misidentified Shorting Switch 57
    - User Abuse 57
    - Wrong Shaft, Wrong Knobs, Nuts That Get Lost, Too Big To Fit 57

8. Rotational Encoder 59

- What It Does 59
- How It Works 59
- Variants 60
    - Pulses And Detents 61
    - Format 61
    - Output 61
    - Rotational Resistance 61
- Values 61
    - Contact Bounce 61
    - Sliding Noise 62
- How To Use It 62
- What Can Go Wrong 62
    - Switch Bounce 62
    - Contact Burnout 63

9. Relay 65

- What It Does 65
- How It Works 66
- Variants 67
    - Latching 67
    - Polarity 67
    - Pinout Variations 67

Reed Relay 68
Small Signal Relay 68
Automotive Relays 69
General Purpose/Industrial 69
Time Delay Relay 69
Contactor 70
Values 70
How To Use It 71
What Can Go Wrong 72
Wrong Pinouts 72
Wrong Orientation 72
Wrong Type 72
Wrong Polarity 72
AC And DC 72
Chatter 72
Relay Coil Voltage Spike 72
Arcing 72
Magnetic Fields 72
Environmental Hazards 73

> > MODERATION

10. **Resistor** 75
What It Does 75
How It Works 76
Variants 76
Resistor Array 77
Values 79
Tolerance 79
Value Coding 81
Stability 82
Materials 82
How To Use It 84
In Series With LED 84
Current Limiting With A Transistor 84
Pullup And Pulldown Resistors 85
Audio Tone Control 85
RC Network 85
Voltage Divider 86
Resistors In Series 86
Resistors In Parallel 86
What Can Go Wrong 87
Heat 87
Noise 87
Inductance 87
Inaccuracy 87

Wrong Values 88

11. Potentiometer 89

What It Does 89
How It Works 90
Variants 90
Linear And Log Taper 90
Classic-Style Potentiometer 91
Multiple-Turn Potentiometer 92
Ganged Potentiometer 93
Switched Potentiometer 93
Slider Potentiometer 93
Trimmer Potentiometer 93
How To Use It 94
What Can Go Wrong 95
Wear And Tear 95
Knobs That Don't Fit 95
Nuts That Get Lost 95
A Shaft That Isn't Long Enough 96
Sliders With No Finger Grip 96
Too Big To Fit 96
Overheating 96
The Wrong Taper 96

12. Capacitor 97

What It Does 97
How It Works 97
Variants 99
Format 99
Principal Types 101
Dielectrics 103
Values 104
Farads 104
Commonly Used Values 104
Dielectric Constant 105
The Time Constant 105
Multiple Capacitors 106
Alternating Current And Capacitive Reactance 106
Equivalent Series Resistance 106
How To Use It 107
Bypass Capacitor 107
Coupling Capacitor 107
High-Pass Filter 107
Low-Pass Filter 107
Smoothing Capacitor 108
Snubber 108
Capacitor As A Battery Substitute 109
What Can Go Wrong 109

Wrong Polarity 110
Voltage Overload 110
Leakage 110
Dielectric Memory 110
Specific Electrolytic Issues 110
Heat 110
Vibration 110
Misleading Nomenclature 111

13. **Variable Capacitor** **113**
What It Does 113
How It Works 113
Variants 114
Values 115
Formats 115
How To Use It 115
What Can Go Wrong 117
Failure To Ground Trimmer Capacitor While Adjusting It . . . 117
Application Of Overcoat Material Or “Lock Paint” 117
Lack Of Shielding 117

14. **Inductor** **119**
What It Does 119
How It Works 120
DC Through A Coil 121
Magnetic Core 122
EMF And Back-EMF 122
Electrical And Magnetic Polarity 123
Variants 124
Magnetic Cores 124
Nonmagnetic Cores 125
Variable Inductors 125
Ferrite Beads 126
Toroidal Cores 126
Gyrator 127
Values 128
Calculating Inductance 128
Calculating Reactance 128
Calculating Reluctance 129
Datasheet Terminology 129
Series And Parallel Configurations 129
Time Constant 129
How To Use It 130
Core Choices 132
Miniaturization 132
What Can Go Wrong 132
Real-World Defects 132
Saturation 132

RF Problems 133

> > CONVERSION

15. AC-AC Transformer 135
What It Does 135
How It Works 136
The Core 137
Taps 137
Variants 138
Core Shapes 138
Power Transformer 138
Plug-In Transformer 139
Isolation Transformer 139
Autotransformer 140
Variable Transformer 140
Audio Transformer 140
Split-Bobbin Transformer 141
Surface-Mount Transformer 141
Values 141
How To Use It 142
What Can Go Wrong 142
Reversal Of Input And Output 142
Shock Hazard From Common Ground 142
Accidental DC Input 142
Overload 142
Incorrect AC Frequency 142

16. AC-DC Power Supply 143
What It Does 143
Variants 143
Linear Regulated Power Supply 143
Switching Power Supply 144
Unregulated Power Supply 146
Adjustable Power Supply 146
Voltage Multiplier 146
Formats 146
How To Use It 147
What Can Go Wrong 147
High Voltage Shock 147
Capacitor Failure 147
Electrical Noise 147
Peak Inrush 147

17. DC-DC Converter 149
What It Does 149
How It Works 149
Variants 150

    - Buck Converter 150
    - Boost Converter 151
    - Flyback Converter With Inductor 151
    - Flyback Converter With Transformer 151
    - Formats 151
- Values 152
    - Nominal Input Voltage And Frequency 152
    - Output Voltage 153
    - Input Current And Output Current 153
    - Load Regulation 153
    - Efficiency 153
    - Ripple And Noise 154
    - Isolated Or Non-Isolated 154
- How To Use It 154
- What Can Go Wrong 155
    - Electrical Noise In Output 155
    - Excess Heat With No Load 155
    - Inaccurate Voltage Output With Low Load 155

18. DC-AC Inverter 157

- What It Does 157
- How It Works 157
- Variants 158
- Values 158
- How To Use It 159
- What Can Go Wrong 160

> > REGULATION

19. Voltage Regulator 161

- What It Does 161
- How It Works 161
- Variants 163
    - Packaging 163
    - Popular Varieties 163
    - Adjustable Regulators 163
    - Negative And Positive Regulators 164
    - Low-Dropout Linear Regulators 164
    - Quasi-Low-Dropout Linear Regulators 165
    - Additional Pin Functions 165
- Values 165
- How To Use It 165
- What Can Go Wrong 166
    - Inadequate Heat Management 166
    - Transient Response 166
    - Misidentified Parts 166
    - Misidentified Pins 167
    - Dropout Caused By Low Battery 167

Inaccurate Delivered Voltage 167

> ELECTROMAGNETISM

> > LINEAR

20. **Electromagnet 169**
- What It Does 169
- How It Works 169
- Variants 170
- Values 171
- How To Use It 171
- What Can Go Wrong 172

21. **Solenoid 173**
- What It Does 173
- How It Works 174
- Variants 176
  - Low Profile 176
  - Latching 176
  - Rotary 176
  - Hinged Clapper 176
- Values 176
  - Coil Size Vs. Power 177
- How To Use It 177
- What Can Go Wrong 177
  - Heat 177
  - AC Inrush 177
  - Unwanted EMF 177
  - Loose Plunger 177

> > ROTATIONAL

22. **DC Motor 179**
- What It Does 179
- How It Works 179
- Variants 181
  - Coil Configurations 181
  - Gearhead Motor 181
  - Brushless DC Motor 183
  - Linear Actuator 184
  - Values 184
- How To Use It 185
  - Speed Control 186
  - Direction Control 186
  - Limit Switches 187
- What Can Go Wrong 187

Brushes And Commutator 187
Electrical Noise 187
Heat Effects 188
Ambient Conditions 188
Wrong Shaft Type Or Diameter 188
Incompatible Motor Mounts 188
Backlash 188
Bearings 188
Audible Noise 189

23. AC Motor 191
What It Does 191
How It Works 191
  Stator Design 191
  Rotor Design 192
Variants 195
  Single-Phase Induction Motor 195
  Three-Phase Induction Motor 196
  Synchronous Motor 196
  Reluctance Motor 197
  Variable Frequency Drive 198
  Wound-Rotor AC Induction Motor 198
  Universal Motor 198
  Inverted AC Motors 199
Values 199
How To Use It 199
What Can Go Wrong 200
  Premature Restart 200
  Frequent Restart 200
  Undervoltage Or Voltage Imbalance 200
  Stalled Motor 200
  Protective Relays 200
  Excess Torque 200
  Internal Breakage 200

24. Servo Motor 201
What It Does 201
How It Works 201
Variants 203
Values 204
How To Use It 205
  Modification For Continuous Rotation 206
What Can Go Wrong 206
  Incorrect Wiring 206
  Shaft/Horn Mismatch 206
  Unrealistically Rapid Software Commands 207
  Jitter 207
  Motor Overload 207

Unrealistic Duty Cycle 207
Electrical Noise 207

25. Stepper Motor 209
What It Does 209
How It Works 209
Reluctance Stepper Motors 210
Permanent Magnet Stepper Motors 211
Bipolar Stepper Motors 213
Unipolar Motors 213
Variants 214
High Phase Count 214
Hybrid 216
Bifilar 216
Multiphase 216
Microstepping 217
Sensing And Feedback 217
Voltage Control 217
Values 218
How To Use It 218
Protection Diodes 218
Positional Control 219
What Can Go Wrong 219
Incorrect Wiring 219
Step Loss 219
Excessive Torque 219
Hysteresis 220
Resonance 220
Hunting 220
Saturation 220
Rotor Demagnetization 220

> DISCRETE SEMICONDUCTOR

> > SINGLE JUNCTION

26. Diode 221
What It Does 221
How It Works 223
Variants 224
Packaging 224
Signal Diodes 224
Rectifier Diodes 224
Zener Diode 224
Transient Voltage Suppressor (TVS) 225
Schottky Diode 225
Varactor Diode 225

        - Tunnel Diode, Gunn Diode, PIN Diode 226
        - Diode Array 226
        - Bridge Rectifier 226
    - Values 226
    - How To Use It 227
        - Rectification 227
        - Back-EMF Suppression 228
        - Voltage Selection 229
        - Voltage Clamping 230
        - Logic Gate 230
        - DC Voltage Regulation And Noise Suppression 230
        - AC Voltage Control And Signal Clipping 231
        - Voltage Sensing 231
    - What Can Go Wrong 232
        - Overload 232
        - Reversed Polarity 233
        - Wrong Type Of Diode 233

27. Unijunction Transistor 235
    - What It Does 235
    - How It Works 236
    - Variants 238
    - Values 238
        - How To Use It 239
    - What Can Go Wrong 239
        - Name Confusion 239
        - Incorrect Bias 239
        - Overload 240

> > MULTI-JUNCTION

28. Bipolar Transistor 241
    - What It Does 241
    - How It Works 241
        - Current Gain 244
        - Terminology 245
    - Variants 245
        - Packaging 245
        - Connections 246
    - How To Use It 246
        - Darlington Pairs 248
        - Amplifiers 250
    - What Can Go Wrong 251
        - Wrong Connections On A Bipolar Transistor 251
        - Wrong Connections On A Darlington Pair Chip 251
        - Soldering Damage 252
        - Excessive Current Or Voltage 252

Excessive Leakage 252

29. **Field Effect Transistor** **253**
What It Does 253
How It Works 253
JFETs 253
JFET Behavior 255
MOSFETs 256
The Substrate Connection 261
Variants 262
MESFET 262
V-Channel MOSFET 262
Trench MOS 262
Values 262
How To Use It 263
P-Channel Disadvantage 263
Bipolar Substitution 263
Amplifier Front Ends 263
Voltage-Controlled Resistor 263
Compatibility With Digital Devices 263
What Can Go Wrong 263
Static Electricity 263
Heat 263
Wrong Bias 264

Appendix A. **Schematic Symbols** **265**

Index **269**

Preface

At a time when information is widely and freely available in greater quantities than ever before, the reader may wonder whether *The Encyclopedia of Electronic Components* is really necessary. Surely, anything you want to know can be found online?

Well, yes and no. Let's consider the available resources.

1. Datasheets

Datasheets are indispensable, but they have limitations. Some are detailed; others are skimpy. Some show you sample schematics as a guide to using a component; many don't. None of them tells you much about how a component works, because that's not their purpose. Often they don't mention other components that must be added. Some datasheets for DC-DC converters, for instance, say nothing at all about bypass capacitors, even though the capacitors may be essential. A datasheet for an optocoupler says nothing about the pullup resistor required by the open-collector output.

Datasheets don't facilitate comparison shopping. A datasheet from one manufacturer will not compare its products with those from another manufacturer, and may not even provide much guidance about alternatives that are available from the same manufacturer. For example, a datasheet for a linear voltage regulator won't suggest that you might do better to use a DC-DC converter in an application where high efficiency is important.

Most of all, datasheets don't tell you how to avoid common mistakes. What actually happens if you connect that tantalum capacitor the wrong way around? A datasheet gives you the customary list of absolute maximum values, and after that, you are on your own, burning things out, encountering mysterious electronic behavior, and discovering limitations that are so well known, the datasheet didn't bother to mention them. In my experience, relying on datasheets creates a significant risk of reinventing the wheel.

2. Wikipedia

Wikipedia's coverage of electronics is impressive but inconsistent. Some entries are elementary, while others are extremely technical. Some are shallow, while others are deep. Some are well organized, while others run off into obscure topics that may have interested one of the contributors but are of little practical value to most readers. Many topics are distributed over multiple entries, forcing you to hunt through several URLs. Overall, Wikipedia tends to be good if you want theory, but not-so-good if you want hands-on practicality.

3. Manufacturers' Tutorials

A few helpful and enlightened manufacturers have compiled highly authoritative, instructional overviews of the components that they sell. Littelfuse, for instance, publishes an excellent series of documents telling you everything you could possibly want to know about fuses. But now you encounter a different problem: There is so much information, you'll need a couple of hours to dig through it all. Also, because the tutorials tend not to receive high page rankings on Google, they can be hard to find. And if a manufacturer has gaps in its product line, its tutorial is unlikely to mention them. Consequently, you won't know what's missing.

4. Personal Guides

It is a well-known attribute of the Web that many individuals feel the impulse to share everything they know (or think they know) about a particular topic. These personal guides can present surprisingly thorough online coverage of relatively obscure issues, such as the types of capacitors most suitable for loudspeaker crossover circuits, or the correct derivation of amp-hour numbers for lead-acid batteries. Unfortunately, on some sites you can also find errors, unsubstantiated opinions, plagiarism, and eccentricity. My general rule is that three or more guides generally have to agree with each other before their statements can be trusted—and even then, I have a small residue of doubt. The search-inspect-and-verify process can take a while.

So—yes, the information that you want usually does exist somewhere online, but no, it may not be easy to find. The vastness of the Web is not organized like an encyclopedia.

What about books? Generally speaking, they tend to be entry-level, or they specialize in narrow areas. A few broad-ranging books are truly excellent, but they are primarily educational, organized in an instructional sequence. They are not reference books.

The Encyclopedic Solution

Scarcity or inaccessibility of information ceased to be a problem many years ago. Its vast quantity, inconsistency, and dispersal have become the new barriers to acquiring knowledge. If you have to go hunting among datasheets, Wikipedia, manufacturers' tutorials (which may or may not exist), personal guides (which may have unrevealed bias), and multiple educational books, the process will be inconvenient and time-consuming. If you plan to revisit the topic in the future, you'll have to remember which URLs were useful and which ones weren't—and you may find that many of them are not even there anymore.

When I considered these issues during my own work as an electronics columnist for *Make* magazine, I saw a real need for a fact-checked, cross-referenced encyclopedia that would compile the basic information about components concisely, in an organized, consistent format, with informative photographs, schematics, and diagrams. It might save many people a lot of search time if it could summarize how components work, how to use them, what the alternatives are, and what the common errors and problems may be.

That is the modest ambition of *The Encyclopedia of Electronic Components*.

The Audience

Like any reference work, this one hopes to serve two categories of readers: The informed and the not-yet-informed.

Perhaps you are learning electronics, and you see a part listed in a catalog. It looks interesting, but the catalog doesn't tell you exactly what the part does or how it is commonly used. You need to look it up either by function or by name, but you're not sure where to start. An encyclopedic reference can simplify the fact-finding process, can save you from ordering a part that may be inappropriate, and can tell you how it should be used.

Perhaps, instead, you are an electronics engineer or hobbyist, thinking about a new circuit. You remember using a component three or four years ago, but your recollection may not be reliable. You need to refresh your memory with a quick summary—and so, you open the encyclopedia, just to make sure.

Completeness

Obviously, this book cannot include every component that exists. Mouser Electronics claims to have more than 2 million products listed in its online database. *The Encyclopedia of Electronic Components* only has room for a fraction of that number—but still, it can refer you to the primary types. The electronic edition of this book should allow easy insertions and updates. My hope is that it can become an ever-expanding resource.

Acknowledgments

Any reference work draws inspiration from many sources, and this one is no exception. Three were of special importance:

Practical Electronics for Inventors by Paul Scherz (second edition) McGraw-Hill, 2007

Electronic Devices and Circuit Theory by Robert L. Boylestad and Louis Nashelsky (ninth edition) Pearson Education Inc., 2006

The Art of Electronics by Paul Horowitz and Winfield Hill (second edition) Cambridge University Press, 2006

I also made extensive use of information gleaned through Mouser Electronics and Jameco Electronics. And where would any of us be without *Getting Started in Electronics* by Forrest M. Mims III, or *The TTL Cookbook* by Don Lancaster?

In addition, there were individuals who provided special assistance. My editor, Brian Jepson, was immensely helpful in the development of the project. Michael Butler contributed greatly to the early concept and its structure. Josh Gates did resourceful research. My publishers, O'Reilly Media, demonstrated their faith in my work. Kevin Kelly unwittingly influenced me with his legendary interest in "access to tools."

Primary fact checkers were Eric Moberg, Chris Lirakis, Jason George, Roy Rabey, Emre Tuncer, and Patrick Fagg. I am indebted to them for their help. Any remaining errors are, of course, my responsibility.

Lastly I should mention my school friends from decades ago: Hugh Levinson, Patrick Fagg, Graham Rogers, William Edmondson, and John Witty, who helped me to feel that it was okay to be a nerdy kid building my own audio equipment, long before the word "nerd" existed.

—Charles Platt, 2012

How to Use This Book

1

To avoid misunderstandings regarding the purpose and method of this book, here is a quick guide regarding the way in which it has been conceived and organized.

Reference vs. Tutorial

As its title suggests, this is a reference book, not a tutorial. In other words, it does not begin with elementary concepts and build sequentially toward concepts that are more advanced.

You should be able to dip into the text at any point, locate the topic that interests you, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will not find concepts being introduced in a sequential, cumulative manner.

My book *Make:Electronics* follows the tutorial approach. Its range, however, is more circumscribed than that of this encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I am assuming that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently I have not included any proofs of formulae, any definitions rooted in electrical theory, or any historical background. Units are defined only to the extent that is necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Organization

The encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

1. A component merits its own entry if it is (a) widely used or (b) not-so-widely used but has a unique identity and maybe some historical status. A widely used component would be a **bipolar transistor**, while a not-so-widely-used component with a unique identity would be a **unijunction transistor**.
2. A component does not merit its own entry if it is (a) seldom used or (b) very similar in function to another component that is more widely used. For example, the *rheostat* is sub-

sumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls that in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. Instead they are grouped by subject, in much the same way that books in the nonfiction section of a library are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in Figure 1-1. It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **capacitor** entry, for instance, is headed with this path:

power > moderation > capacitor

Any classification scheme tends to run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but should it really be included with solid-state relays and comparators? A decision was made to put it in the **resistor** section, because this seemed more useful.

Some components have hybrid functions. In Volume 2, in the *integrated circuit* subcategory, we will distinguish between those that are *analog* and those that are *digital*. So where should an **analog-digital converter** be listed? It will be found under *analog*, because that category seems better associated with its primary function, and people may be more likely to look for it there.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electromagnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semiconductor	single junction	diode
		unijunction transistor
	multijunction	bipolar transistor
		field-effect transistor

Figure 1-1. *The subject-oriented organization of categories and entries in this encyclopedia.*

Inclusions and Exclusions

There is also the question of what is, and what is not, a component. Is wire a component? Not for the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they have been included as components.

Many similar decisions had to be made on a case-by-case basis. Undoubtedly, some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. Speaking personally, the best I could do was create a book that is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Throughout this encyclopedia, the names of components that have their own entries are presented in **bold type.** Other important electronics terms or component names are presented in *italics* where they first appear in any one section.

The names of components, and the categories to which they belong, are all set in lower-case type, except where a term is normally capitalized because it is an acronym or a trademark. *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

Where formulae are used, they are expressed in a format that will be familiar to computer programmers but may be unfamiliar to others. The * (asterisk) symbol is used in place of a multiplication sign, while the / (slash symbol) is used to indicate division. Where pairs of parentheses are nested, the most deeply nested pair identifies the operations that should be performed first.

Volume Contents

Practical considerations relating to book length influenced the decision to divide *The Encyclopedia of Electronic Components* into three volumes. Each volume deals with broad subject areas as follows.

Volume 1

Power, electromagnetism, and discrete semiconductors.

The *power* category includes sources of power and methods to distribute, store, interrupt, and modify power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the main types of diodes and transistors.

Volume 2

Integrated circuits, light sources, sound sources, heat sources, and high-frequency sources.

Integrated circuits are divided into analog and digital components. *Light sources* range from incandescent bulbs to LEDs and small display screens; some reflective components, such as liquid-crystal displays and e-ink, are also included. *Sound sources* are primarily electromagnetic.

Volume 3

Sensing devices.

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, volumes 2 and 3 are still in preparation, but their contents are expected to be as described above.

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors.

Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from O'Reilly and other publishers, sign up for free at *http://my.safaribooksonline.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

MAKE
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

MAKE unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. MAKE celebrates your right to tweak, hack, and bend any technology to your will. The MAKE audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make is leading—we call it the Maker Movement.

For more information about MAKE, visit us online:

MAKE magazine: *http://makezine.com/magazine/*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com/*

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

http://oreil.ly/encyc_electronic_comp_v1

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

battery

2

This entry covers electrochemical power sources. Electricity is most often generated electromagnetically, but since these sources cannot be classified as components, they are outside the scope of the encyclopedia. Electrostatic sources are excluded for similar reasons.

A battery is sometimes referred to as a *cell* or *power cell*, but can actually contain multiple cells, as defined in this entry. It used to be called an *accumulator* or a *pile*, but those terms are now archaic.

OTHER RELATED COMPONENTS

- **capacitor** (see Chapter 12)

What It Does

A battery contains one or more *electrochemical cells* in which chemical reactions create an electrical potential between two immersed terminals. This potential can be discharged as *current* passing through a *load*.

An electrochemical cell should not be confused with an *electrolytic cell*, which is powered by an external source of electricity to promote *electrolysis*, whereby chemical compounds are broken down to their constituent elements. An electrolytic cell thus consumes electricity, while an electrochemical cell produces electricity.

Batteries range in size from *button cells* to large *lead-acid* units that store power generated by solar panels or windmills in locations that can be off the grid. Arrays of large batteries can provide bridging power for businesses or even small communities where conventional power is unreliable. Figure 2-1 shows a 60KW, 480VDC self-watering battery array installed in a corporate data center, supplementing wind and solar sources and providing time-of-day peak shaving of energy usage. Each lead-acid battery in this array measures approximately 28″ × 24″ × 12″ and weighs about 1,000 lb.

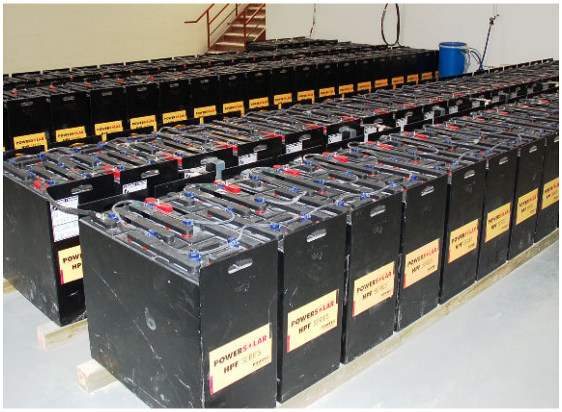

Figure 2-1. *A battery array providing 60KW at 480VDC as backup for a corporate data center. (Photo by permission of Hybridyne Power Systems, Canada, Inc., and the Hybridyne group of companies. Copyright by Hybridyne, an internationally registered trademark of Hybridyne Power Systems Canada Inc. No right of further reproduction unless specifically granted by Hybridyne.)*

Schematic symbols for a battery are shown in Figure 2-2. The longer of the two lines represents the positive side of the battery, in each case. One way to remember this is by imagining that the longer line can be snipped in half so that the two segments can combine to form a + sign. Traditionally, multiple connected battery symbols indicate multiple cells inside a battery; thus the center symbols in the figure could indicate a 3V battery, while those on the right would indicate a voltage greater than 3V. In practice, this convention is not followed conscientiously.

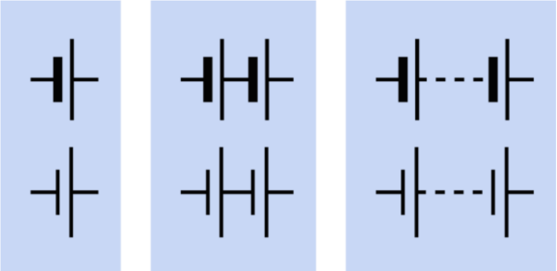

Figure 2-2. *Schematic symbols for a battery. Each pair of symbols within a blue rectangle is functionally identical.*

How It Works

In a basic battery design often used for demonstration purposes, a piece of copper serves as an *electrode*, partially immersed in a solution of copper sulfate, while a piece of zinc forms a second electrode, partially immersed in a solution of zinc sulfate. Each sulfate solution is known as an *electrolyte*, the complete battery may be referred to as a *cell*, and each half of it may be termed a *half-cell*.

A simplified cross-section view is shown in Figure 2-3. Blue arrows show the movement of electrons from the zinc terminal (the *anode*), through an external load, and into a copper terminal (the *cathode*). A *membrane separator* allows the electrons to circulate back through the battery, while preventing electrolyte mixing.

Orange arrows represent positive copper *ions*. White arrows represent positive zinc ions. (An ion is an atom with an excess or deficit of electrons.) The zinc ions are attracted into the zinc sulfate electrolyte, resulting in a net loss of mass from the zinc electrode.

Meanwhile, electrons passing into the copper electrode tend to attract positive copper ions, shown as orange arrows in the diagram. The copper ions are drawn out of the copper sulfate electrolyte, and result in a net accumulation of copper atoms on the copper electrode.

This process is energized partially by the fact that zinc tends to lose electrons more easily than copper.

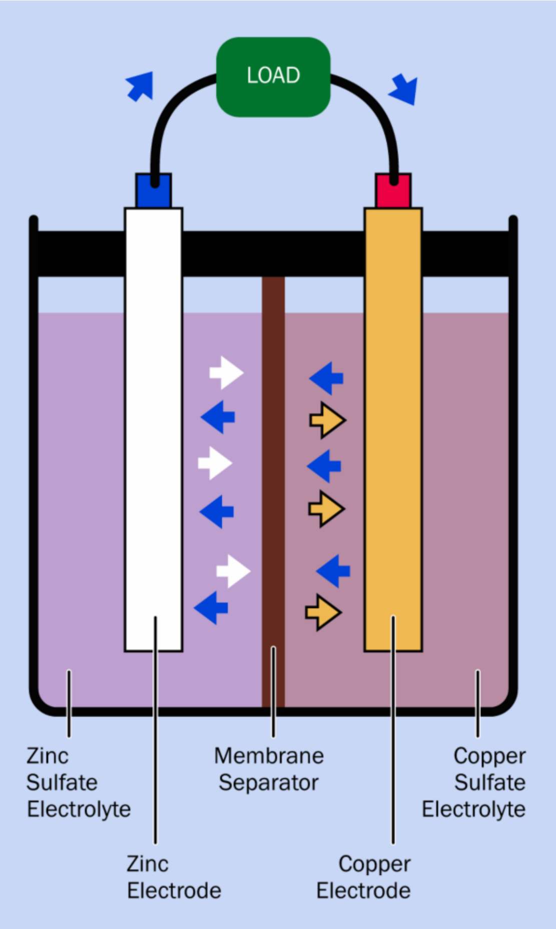

Figure 2-3. *A classically simple electrochemical cell. See text for additional details.*

Batteries for use in consumer electronics typically use a paste instead of a liquid as an electrolyte, and have been referred to as *dry cells*, although this term is becoming obsolete. The two half-cells may be combined concentrically, as in a typical 1.5-volt C, D, AA, or AAA alkaline battery (see Figure 2-4).

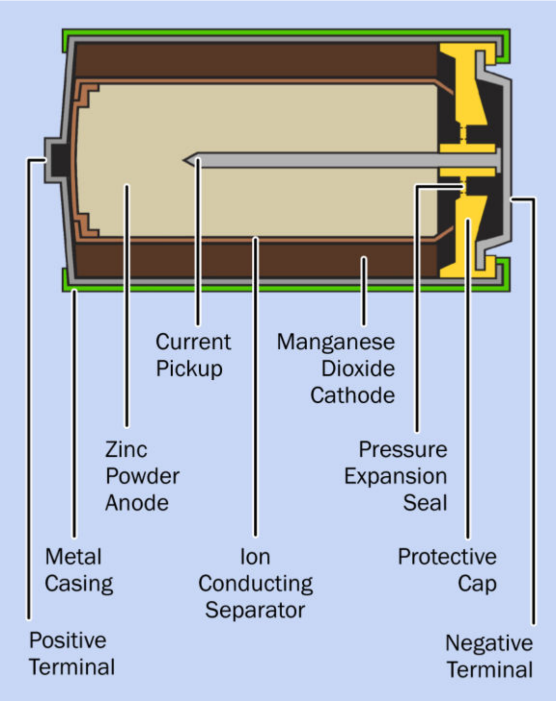

Figure 2-4. *Cross-section view of a typical 1.5-volt alkaline battery.*

A 1.5V battery contains one cell, while a 6V or 9V battery will contain multiple cells connected in series. The total voltage of the battery is the sum of the voltages of its cells.

Electrode Terminology

The electrodes of a cell are often referred to as the *anode* and the *cathode*. These terms are confusing because the electrons enter the anode inside the cell and leave it outside the cell, while electrons enter the cathode from outside the cell and leave it inside the cell. Thus, the anode is an electron emitter if you look at it externally, but the cathode is an electron emitter if you look at it internally.

Conventional current is imagined to flow in the opposite direction to electrons, and therefore, outside the cell, this current flows from the cathode to the anode, and from this perspective, the cathode can be thought of as being “more positive” than the anode. To remember this, think of the letter t in “cathode” as being a + sign, thus: ca+hode. In larger batteries, the cathode is often painted or tagged red, while the anode may be painted or tagged black or blue.

When a reusable battery is recharged, the flow of electrons reverses and the anode and the cathode effectively trade places. Recognizing this, the manufacturers of rechargeable batteries may refer to the more-positive terminal as the anode. This creates additional confusion, exacerbated further still by electronics manufacturers using the term “cathode” to identify the end of a **diode** which must be “more negative”(i.e., at a lower potential) than the opposite end.

To minimize the risk of errors, it is easiest to avoid the terms “anode” and “cathode” when referring to batteries, and speak instead of the negative and positive terminals. This encyclopedia uses the common convention of reserving the term “cathode” to identify the “more negative” end of any type of diode.

Variants

Three types of batteries exist.

1. *Disposable batteries*, properly (but infrequently) referred to as *primary cells*. They are not reliably rechargeable because their chemical reactions are not easily reversible.
2. *Rechargeable batteries*, properly (but infrequently) known as *secondary cells*. They can be recharged by applying a voltage between

the terminals from an external source such as a *battery charger*. The materials used in the battery, and the care with which the battery is maintained, will affect the rate at which chemical degradation of the electrodes gradually occurs as it is recharged repeatedly. Either way, the number of charge/discharge cycles is limited.

3. *Fuel Cells* require an inflow of a reactive gas such as hydrogen to maintain an electrochemical reaction over a long period. They are beyond the scope of this encyclopedia.

A large **capacitor** may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture than a battery of equivalent power storage. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle. See Figure 2-5.

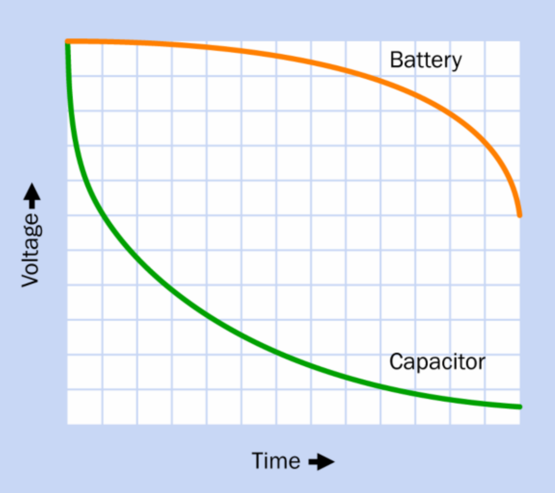

Figure 2-5. *The voltage drop of a discharging capacitor is much steeper initially than that of a battery, making capacitors unsuitable as a battery substitute in many applications. However, the ability of a capacitor to discharge very rapidly at high amperage can sometimes be a significant advantage.*

Capacitors that can store a very large amount of energy are often referred to as *supercapacitors.*

Disposable Batteries

The energy density of any disposable battery is higher than that of any type of rechargeable battery, and it will have a much longer shelf life because it loses its charge more slowly during storage (this is known as the *self-discharge rate*). Disposable batteries may have a useful life of five years or more, making them ideal for applications such as smoke detectors, handheld remotes for consumer electronics, or emergency flashlights.

Disposable batteries are not well suited to delivering high currents through loads below 75Ω. Rechargeable batteries are preferable for higher-current applications. The bar chart in Figure 2-6 shows the rated and actual capabilities of an alkaline battery relative to the three most commonly used rechargeable types, when the battery is connected with a resistance that is low enough to assure complete discharge in 1 hour.

The manufacturer's rating of watt hours per kilo is typically established by testing a battery with a relatively high-resistance load and slow rate of discharge. This rating will not apply in practice if a battery is discharged with a C-rate of 1, meaning complete discharge during 1 hour.

Common types of disposable batteries are *zinc-carbon cells* and *alkaline cells*. In a zinc-carbon cell, the negative electrode is made of zinc while the positive electrode is made of carbon. The limited power capacity of this type of battery has reduced its popularity, but because it is the cheapest to manufacture, it may still be found where a company sells a product with "batteries included." The electrolyte is usually ammonium chloride or zinc chloride. The 9V battery in Figure 2-7 is actually a zinc-carbon battery according to its supplier, while the smaller one beside it is a 12V alkaline battery designed for use in burglar alarms. These examples show that batteries cannot always be identified correctly by a casual assessment of their appearance.

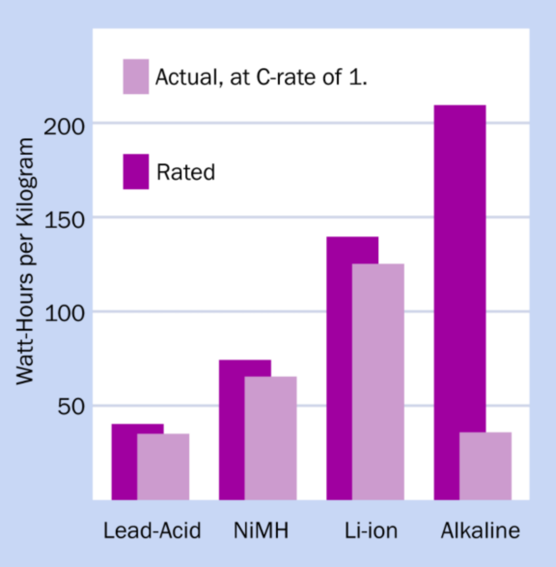

Figure 2-6. *Because of their relatively high internal resistance, alkaline batteries are especially unsuited to high discharge rates, and should be reserved for applications where a small current is required over a long period. (Chart derived from http://batteryuniversity.com.)*

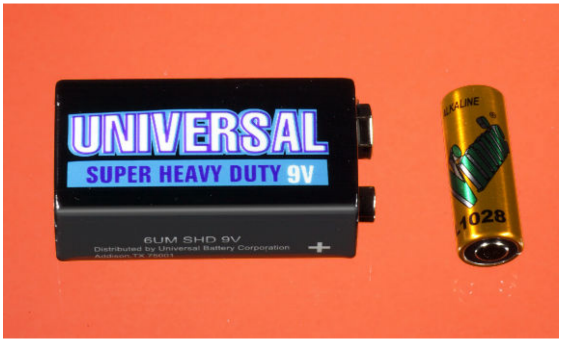

Figure 2-7. *At left, a cheap carbon-zinc battery; at right, a 12V alkaline burglar-alarm battery. See text for additional details.*

In an alkaline cell, the negative electrode is made of zinc powder, the positive electrode is manganese dioxide, and the electrolyte is potassium hydroxide. An alkaline cell may provide between three to five times the power capacity of an equal size of zinc-carbon cell and is less susceptible to voltage drop during the discharge cycle.

Extremely long shelf life is necessary in some military applications. This may be achieved by using a *reserve battery*, in which the internal chemical compounds are separated from each other but can be recombined prior to use.

Rechargeable Batteries

Commonly used types are *lead-acid*, *nickel cadmium* (abbreviated *NiCad* or *NiCd*), *nickel-metal hydride* (abbreviated *NiMH*), *lithium-ion* (abbreviated *Li-ion*), and *lithium-ion polymer*.

Lead-acid batteries have existed for more than a century and are still widely used in vehicles, burglar alarms, emergency lighting, and large power backup systems. The early design was described as *flooded*; it used a solution of sulfuric acid (generically referred to as *battery acid*) as its electrolyte, required the addition of distilled water periodically, and was vented to allow gas to escape. The venting also allowed acid to spill if the battery was tipped over.

The *valve-regulated lead-acid* battery (*VRLA*) has become widely used, requiring no addition of water to the cells. A pressure relief valve is included, but will not leak electrolyte, regardless of the position of the battery. VRLA batteries are preferred for *uninterruptible power supplies* for data-processing equipment, and are found in automobiles and in electric wheelchairs, as their low gas output and security from spillage increases their safety factor.

VRLA batteries can be divided into two types: absorbed glass mat (AGM) and gel batteries. The electrolyte in an AGM is absorbed in a fiber-glass mat separator. In a gel cell, the electrolyte is mixed with silica dust to form an immobilized gel.

The term *deep cycle battery* may be applied to a lead-acid battery and indicates that it should be more tolerant of discharge to a low level—perhaps 20 percent of its full charge (although manufacturers may claim a lower number). The plates in a standard lead-acid battery are composed of a lead *sponge*, which maximizes the surface area available to acid in the battery but can be physically abraded by deep discharge. In a deep cycle battery, the plates are solid. This means they are more robust, but are less able to supply high amperage. If a deep-discharge battery is used to start an internal combustion engine, the battery should be larger than a regular lead-acid battery used for this purpose.

A sealed lead-acid battery intended to power an external light activated by a motion detector is shown in Figure 2-8. This unit weighs several pounds and is trickle-charged during the daytime by a 6″ × 6″ solar panel.

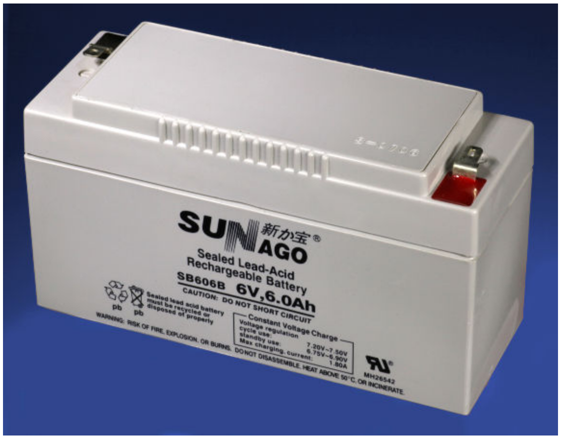

Figure 2-8. *A lead-acid battery from an external light activated by a motion sensor.*

Nickel-cadmium (*NiCad*) batteries can withstand extremely high currents, but have been banned in Europe because of the toxicity of metallic cadmium. They are being replaced in the United States by *nickel-metal hydride* (*NiMH*) types, which are free from the *memory effect* that can prevent a NiCad cell from fully recharging if it has been left for weeks or months in a partially discharged state.

Lithium-ion and lithium-ion polymer batteries have a better energy-to-mass ratio than NiMH batteries, and are widely used with electronic devices such as laptop computers, media players, digital cameras, and cellular phones. Large arrays of lithium batteries have also been used in some electric vehicles.

Various small rechargeable batteries are shown in Figure 2-9. The NiCad pack at top-left was manufactured for a cordless phone and is rapidly becoming obsolete. The 3V lithium battery at top-right was intended for a digital camera. The three batteries in the lower half of the photograph are all rechargeable NiMH substitutes for 9V, AA, and AAA batteries. The NiMH chemistry results in the AA and AAA single-cell batteries being rated for 1.2V rather than 1.5V, but the manufacturer claims they can be substituted for 1.5V alkaline cells because NiMH units sustain their rated voltage more consistently over time. Thus, the output from a fresh NiMH battery may be comparable to that of an alkaline battery that is part-way through its discharge cycle.

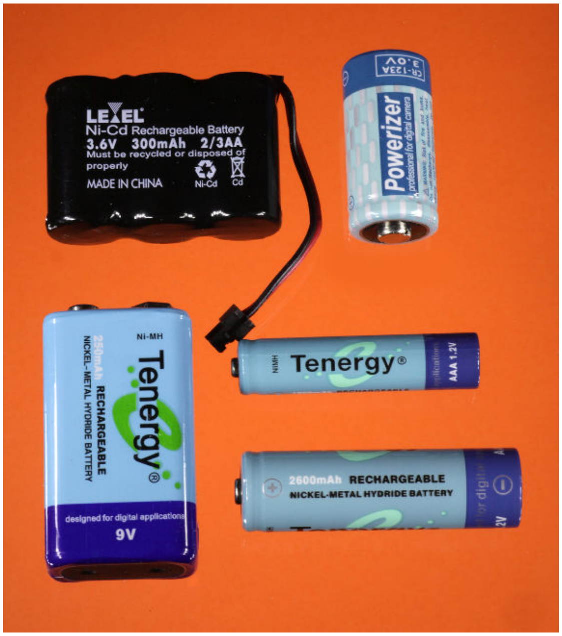

Figure 2-9. *Top left: NiCad battery pack for a cordless phone. Top right: Lithium battery for a digital camera. The other batteries are rechargeable NiMH substitutes for everyday alkaline cells.*

NiMH battery packs are available to deliver substantial power while being smaller and lighter than lead-acid equivalents. The NiMH package in Figure 2-10 is rated for 10Ah, and consists of ten

D-size NiMH batteries wired in series to deliver 12VDC. This type of battery pack is useful in robotics and other applications where a small motor-driven device must have free mobility.

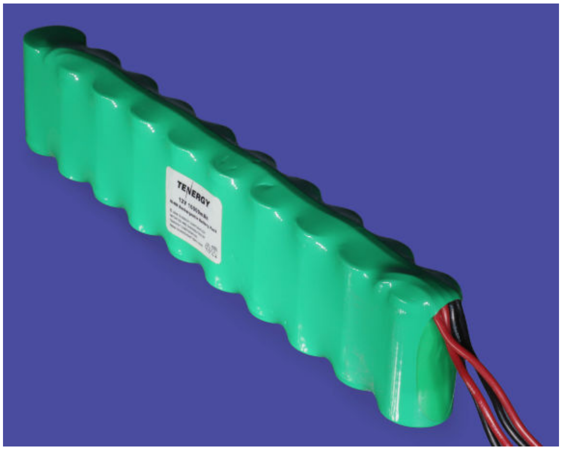

Figure 2-10. *This NiMH battery pack is rated at 10Ah and delivers 12 volts from ten D-size cells wired in series.*

Values

Amperage

The current delivered by a battery will be largely determined by the resistance of the external load placed between its terminals. However, because ion transfer must occur inside the battery to complete the circuit, the current will also be limited by the *internal resistance* of the battery. This should be thought of as an active part of the circuit.

Since a battery will deliver no current if there is no load, current must be measured while a load is attached, and cannot be measured by a meter alone. The meter will be immediately overloaded, with destructive results, if it is connected directly between the terminals of a battery, or in parallel with the load. Current must always be measured with the meter in series with the load, and the polarity of the meter must correspond with the polarity of the battery. See Figure 2-11.

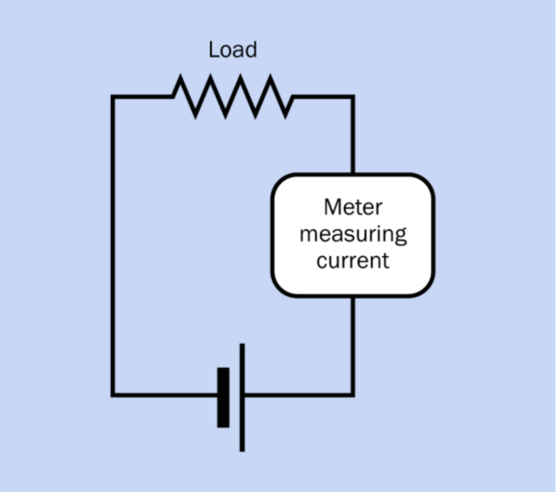

Figure 2-11. *When measuring current using an ammeter (or a multimeter configured to measure amps), the meter must be placed in series with the battery and a load. To avoid damaging the meter, it must never be applied directly across the terminals of the battery, or in parallel with a load. Be careful to observe the polarity of the meter.*

Capacity

The *electrical capacity* of a battery is measured in *amp-hours*, abbreviated *Ah*, *AH*, or (rarely) *A/H*. Smaller values are measured in *milliamp-hours*, usually abbreviated *mAh*. If I is the current being drawn from a battery (in amps) and T is the time for which the battery can deliver that current (in hours), the amp-hour capacity is given by the formula:

```
Ah = I * T
```

By turning the formula around, if we know the amp-hour rating that a manufacturer has determined for a battery, we can calculate the time in hours for which a battery can deliver a particular current:

```
T = Ah / I
```

Theoretically, Ah is a constant value for any given battery. Thus a battery rated for 4Ah should provide 1 amp for 4 hours, 4 amps for 1 hour, 5 amps for 0.8 hours (48 minutes), and so on.

In reality, this conveniently linear relationship does not exist. It quickly breaks down as the current rises, especially when using lead-acid batteries, which do not perform well when required to deliver high current. Some of the current is lost as heat, and the battery may be electrochemically incapable of keeping up with demand.

The *Peukert number* (named after its German originator in 1897) is a fudge factor to obtain a more realistic value for T at higher currents. If n is the Peukert number for a particular battery, then the previous formula can be modified thus:

$$T = Ah / I^{n}$$

Manufacturers usually (but not always) supply Peukert's number in their specification for a battery. So, if a battery has been rated at 4Ah, and its Peukert number is 1.2 (which is typical for lead-acid batteries), and I=5 (in other words, we want to know for how long a time, T, the battery can deliver 5 amps):

$$T = 4 / 5^{1.2} = \text{approximately } 4 / 6.9$$

This is about 0.58 hours, or 35 minutes—much less than the 48 minutes that the original formula suggested.

Unfortunately, there is a major problem with this calculation. In Peukert's era, the amp-hour rating for a battery was established by a manufacturer by drawing 1A and measuring the time during which the battery was capable of delivering that current. If it took 4 hours, the battery was rated at 4Ah.

Today, this measurement process is reversed. Instead of specifying the current to be drawn from the battery, a manufacturer specifies the time for which the test will run, then finds the maximum current the battery can deliver for that time. Often, the time period is 20 hours. Therefore, if a battery has a modern 4Ah rating, testing has probably determined that it delivered 0.2A for 20 hours, not 1A for 4 hours, which would have been the case in Peukert's era.

This is a significant distinction, because the same battery that can deliver 0.2A for 20 hours will not be able to satisfy the greater demand of 1A for 4 hours. Therefore the old amp-hour rating and the modern amp-hour rating mean different things and are incompatible. If the modern Ah rating is inserted into the old Peukert formula (as it was above), the answer will be misleadingly optimistic. Unfortunately, this fact is widely disregarded. Peukert's formula is still being used, and the performance of many batteries is being evaluated incorrectly.

The formula has been revised (initially by Chris Gibson of SmartGauge Electronics) to take into account the way in which Ah ratings are established today. Suppose that AhM is the modern rating for the battery's capacity in amp-hours, H is the duration in hours for which the battery was tested when the manufacturer calibrated it, n is Peukert's number (supplied by the manufacturer) as before, and I is the current you hope to draw from the battery. This is the revised formula to determine T:

$$T = H * (AhM / (I * H)^{n})$$

How do we know the value for H? Most (not all) manufacturers will supply this number in their battery specification. Alternatively, and confusingly, they may use the term *C-rate*, which can be defined as 1/H. This means you can easily get the value for H if you know the C-rate:

$$H = 1 / \text{C-rate}$$

We can now use the revised formula to rework the original calculation. Going back to the example, if the battery was rated for 4Ah using the modern system, in a discharge test that lasted 20 hours (which is the same as a C-rate of 0.05), and the manufacturer still states that it has a Peukert number of 1.2, and we want to know for how long we can draw 5A from it:

$$T = 20 * (4/(5 * 20)^{1.2}) = \text{approximately } 20 * 0.021$$

This is about 0.42 hours, or 25 minutes—quite different from the 35 minutes obtained with the old version of the formula, which should never be used when calculating the probable discharge time based on a modern Ah rating. These issues may seem arcane, but they are of great importance when assessing the likely performance of battery-powered equipment such as electric vehicles.

Figure 2-12 shows the probable actual performance of batteries with Peukert numbers of 1.1, 1.2, and 1.3. The curves were derived from the revised version of Peukert's formula and show how the number of amp-hours that you can expect diminishes for each battery as the current increases. For example, if a battery that the manufacturer has assigned a Peukert number of 1.2 is rated at 100Ah using the modern 20-hour test, but we draw 30A from it, the battery can actually deliver only 70Ah.

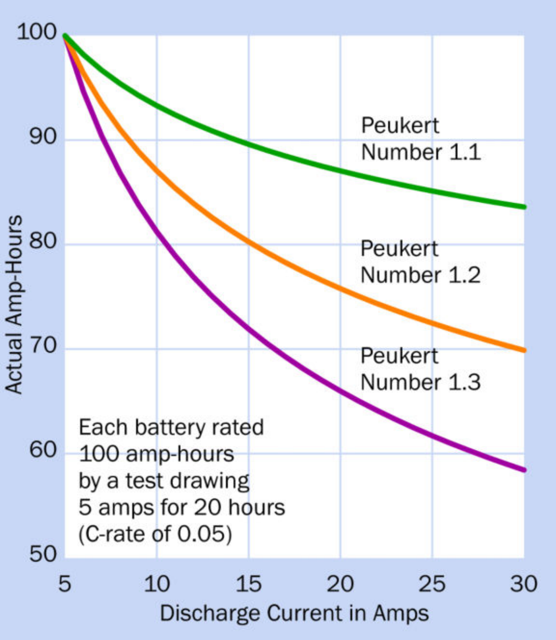

Figure 2-12. *Actual amp-hour performance that should be expected from three batteries of Peukert numbers 1.1, 1.2, and 1.3 when they discharge currents ranging from 5 to 30 amps, assuming that the manufacturer has rated each battery at 100Ah using the modern system, which usually entails a 20-hour test (a C-rate of 0.05).*

One additional factor: For any rechargeable battery, the Peukert number gradually increases with age, as the battery deteriorates chemically.

Voltage

The rated voltage of a fully charged battery is known as the *open circuit voltage* (abbreviated *OCV* or V_{oc}), defined as the potential that exists when no load is imposed between the terminals. Because the internal resistance of a volt meter (or a multimeter, when it is used to measure DC volts) is very high, it can be connected directly between the battery terminals with no other load present, and will show the OCV quite accurately, without risk of damage to the meter. A fully charged 12-volt car battery may have an OCV of about 12.6 volts, while a fresh 9-volt alkaline battery typically has an OCV of about 9.5 volts. Be extremely careful to set a multimeter to measure DC volts before connecting it across the battery. Usually this entails plugging the wire from the red probe into a socket separately reserved for measuring voltage, not amperage.

The voltage delivered by a battery will be pulled down significantly when a load is applied to it, and will decrease further as time passes during a discharge cycle. For these reasons, a **voltage regulator** is required when a battery powers components such as digital integrated circuit chips, which do not tolerate a wide variation in voltage.

To measure voltage while a load is applied to the battery, the meter must be connected in parallel with the load. See Figure 2-13. This type of measurement will give a reasonably accurate reading for the potential applied to the load, so long as the resistance of the load is relatively low compared with the internal resistance of the meter.

Figure 2-14 shows the performance of five commonly used sizes of alkaline batteries. The ratings in this chart were derived for alkaline batteries under favorable conditions, passing a small current through a relatively high-ohm load for long periods (40 to 400 hours, depending on battery type). The test continued until the final voltage for each 1.5V battery was 0.8V, and the final voltage for the 9V battery was a mere 4.8V. These voltages were considered acceptable when the

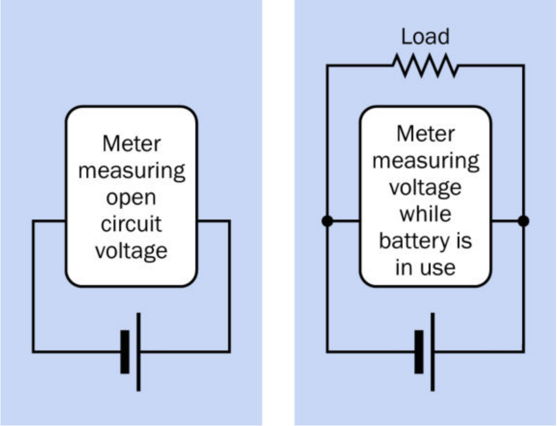
Meter measuring open circuit voltage
Load
Meter measuring voltage while battery is in use

Figure 2-13. *When using a volt meter (or a multimeter configured to measure voltage), the meter can be applied directly between the battery terminals to determine the open-circuit voltage (OCV), or in parallel with a load to determine the voltage actually supplied during use. A multimeter must be set to measure DC volts before connecting it across a battery. Any other setting may damage the meter.*

Ah ratings for the batteries were calculated by the manufacturer, but in real-world situations, a final voltage of 4.8V from a 9V battery is likely to be unacceptable in many electronics applications.

Battery type	Rating (Ah)	Final voltage	Load (ohms)	Current (mA)
AAA	1.15	0.8	75	20
AA	2.87	0.8	75	20
C	7.8	0.8	39	40
D	17	0.8	39	40
9V	0.57	4.8	620	14

Figure 2-14. *The voltage delivered by a battery may drop to a low level while a manufacturer is establishing an amp-hour rating. Values for current, shown in the chart, were calculated subsequently as estimated averages, and should be considered approximate. (Derived from a chart published by Panasonic.)*

As a general rule of thumb, if an application does not tolerate a significant voltage drop, the manufacturer's amp-hour rating for a small battery may be divided by 2 to obtain a realistic number.

How to Use it

When choosing a battery to power a circuit, considerations will include the intended shelf life, maximum and typical current drain, and battery weight. The amp-hour rating of a battery can be used as a very approximate guide to determine its suitability. For 5V circuits that impose a drain of 100mA or less, it is common to use a 9V battery, or six 1.5V batteries in series, passing current through a **voltage regulator** such as the LM7805. Note that the voltage regulator requires energy to function, and thus it imposes a voltage drop that will be dissipated as heat. The minimum drop will vary depending on the type of regulator used.

Batteries or cells may be used in series or in parallel. In series, the total voltage of the chain of cells is found by summing their individual voltages, while their amp-hour rating remains the same as for a single cell, assuming that all the cells are identical. Wired in parallel, the total voltage of the cells remains the same as for a single cell, while the combined amp-hour value is found by summing their individual amp-hour ratings, assuming that all the batteries are identical. See Figure 2-15.

In addition to their obvious advantage of portability, batteries have an additional advantage of being generally free from power spikes and noise that can cause sensitive components to misbehave. Consequently, the need for smoothing will depend only on possible noise created by other components in the circuit.

Motors or other inductive loads draw an initial surge that can be many times the current that they use after they start running. A battery must be chosen that will tolerate this surge without damage.

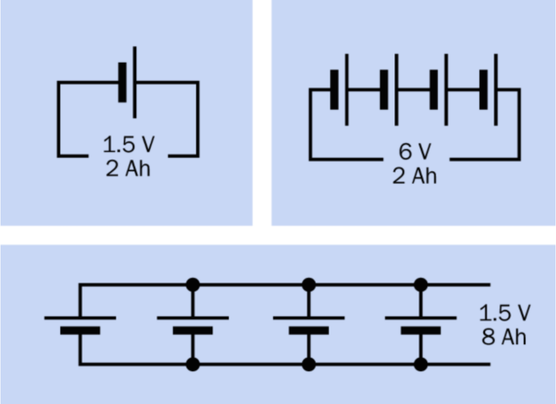

Figure 2-15. *Theoretical results of using 1.5V cells in series or in parallel, assuming a 2Ah rating for one cell.*

Because of the risk of fire, United States airline regulations limit the amp-hour capacity of lithium-ion batteries in any electronic device in carry-on or checked passenger baggage. If a device may be carried frequently as passenger baggage (for example, emergency medical equipment), NiMH batteries are preferred.

What Can Go Wrong

Short Circuits: Overheating and Fire

A battery capable of delivering significant current can overheat, catch fire, or even explode if it is short-circuited. Dropping a wrench across the terminals of a car battery will result in a bright flash, a loud noise, and some molten metal. Even a 1.5-volt alkaline AA battery can become too hot to touch if its terminals are shorted together. (Never try this with a rechargeable battery, which has a much lower internal resistance, allowing much higher flow of current.) Lithium-ion batteries are particularly dangerous, and almost always are packaged with a current-limiting component that should not be disabled. A short-circuited lithium battery can explode.

If a battery pack is used as a cheap and simple workbench DC power supply, a **fuse** or circuit breaker should be included. Any device that uses significant battery power should be fused.

Diminished Performance Caused by Improper Recharging

Many types of batteries require a precisely measured charging voltage and a cycle that ends automatically when the battery is fully charged. Failure to observe this protocol can result in chemical damage that may not be reversible. A charger should be used that is specifically intended for the type of battery. A detailed comparison of chargers and batteries is outside the scope of this encyclopedia.

Complete Discharge of Lead-Acid Battery

Complete or near-complete discharge of a lead-acid battery will significantly shorten its life (unless it is specifically designed for deep-cycle use—although even then, more than an 80% discharge is not generally recommended).

Inadequate Current

Chemical reactions inside a battery occur more slowly at low temperatures. Consequently, a cold battery cannot deliver as much current as a warm battery. For this reason, in winter weather, a car battery is less able to deliver high current. At the same time, because engine oil becomes more viscous as the temperature falls, the starter motor will demand more current to turn the engine. This combination of factors explains the tendency of car batteries to fail on cold winter mornings.

Incorrect Polarity

If a battery charger or generator is connected with a battery with incorrect polarity, the battery may experience permanent damage. The **fuse** or *circuit breaker* in a charger may prevent this from occurring and may also prevent damage to the charger, but this cannot be guaranteed.

If two high-capacity batteries are connected with opposite polarity (as may happen when a clumsy attempt is made to start a stalled car with jumper cables), the results may be explosive. Never lean over a car battery when attaching cables to it, and ideally, wear eye protection.

Reverse Charging

Reverse charging can occur when a battery becomes completely discharged while it is wired (correctly) in series with other batteries that are still delivering current. In the upper section of the schematic at Figure 2-16 two healthy 6V batteries, in series, are powering a resistive load. The battery on the left applies a potential of 6 volts to the battery on the right, which adds its own 6 volts to create a full 12 volts across the load. The red and blue lines indicate volt meter leads, and the numbers show the reading that should be observed on the meter.

In the second schematic, the battery on the left has become exhausted and is now a "dead weight" in the circuit, indicated by its gray color. The battery on the right still sustains a 6-volt potential. If the internal resistance of the dead battery is approximately 1 ohm and the resistance of the load is approximately 20 ohms, the potential across the dead battery will be about 0.3 volts, in the opposite direction to its normal charged voltage. Reverse charging will result and can damage the battery. To avoid this problem, a battery pack containing multiple cells should never be fully discharged.

Sulfurization

When a lead-acid battery is partially or completely discharged and is allowed to remain in that state, sulfur tends to build up on its metal plates. The sulfur gradually tends to harden, forming a barrier against the electrochemical reactions that are necessary to recharge the battery. For this reason, lead-acid batteries should not be allowed to sit for long periods in a discharged condition. Anecdotal evidence suggests that even a very small trickle-charging current can prevent sulfurization, which is why some people recommend attaching a small solar panel to a battery that is seldom used—for example, on a sail boat, where the sole function of the battery is to start an auxiliary engine when there is insufficient wind.

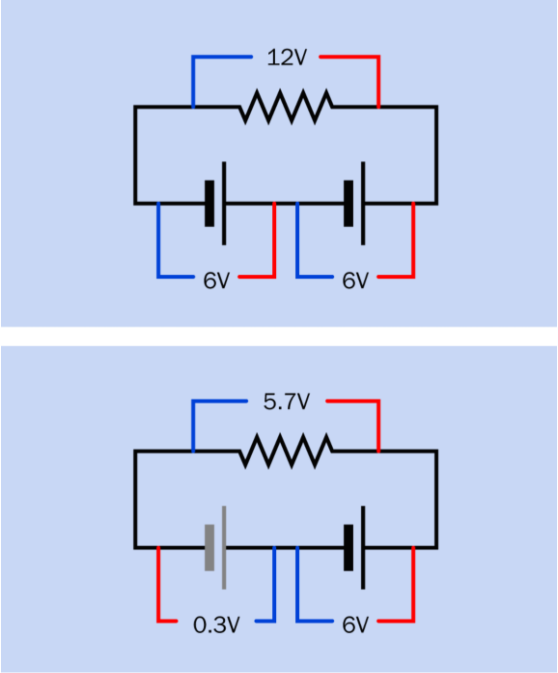

Figure 2-16. *When a pair of 6V batteries is placed in series to power a resistive load, if one of the batteries discharges completely, it becomes a load instead of a power source, and will be subjected to reverse charging, which may cause permanent damage.*

High Current Flow Between Parallel Batteries

If two batteries are connected in parallel, with correct polarity, but one of them is fully charged while the other is not, the charged battery will attempt to recharge its neighbor. Because the batteries are wired directly together, the current will be limited only by their internal resistance and the resistance of the cables connecting them. This may lead to overheating and possible damage. The risk becomes more significant when linking batteries that have high Ah ratings. Ideally they should be protected from one another by high-current **fuses.**

jumper

3

A jumper may also be referred to as a *jumper socket* or a *shunt*. A jumper should not be confused with *jumper wires*, which are not considered components for the purposes of this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)

What It Does

A jumper is a low-cost substitute for a **switch,** where a connection has to be made (or unmade) only a few times during the lifetime of a product. Typically it allows a function or feature on a circuit board to be set on a semipermanent basis, often at the time of manufacture. A *DIP switch* performs the same function more conveniently. See "DIP" (page 43).

There is no standardized schematic symbol to represent a jumper.

How It Works

A jumper is a very small rectangular plastic tab containing two (or sometimes more) metal sockets usually spaced either 0.1" or 2mm apart. The sockets are connected electrically inside the tab, so that when they are pushed over two (or more) pins that have been installed on a circuit board for this purpose, the jumper shorts the pins together. The pins are usually 0.025" square and are often part of a *header* that is soldered into the board. In a parts catalogue, jumpers may be found in a section titled "Headers and Wire Housings" or similar.

Three jumpers are shown in Figure 3-1. The blue one contains two sockets spaced 0.1" and is deep enough to enclose the pins completely. The red one contains two sockets spaced 2mm and may allow the tips of the pins to emerge from its opposite end. The black one contains four sockets, each pair spaced 0.1" apart.

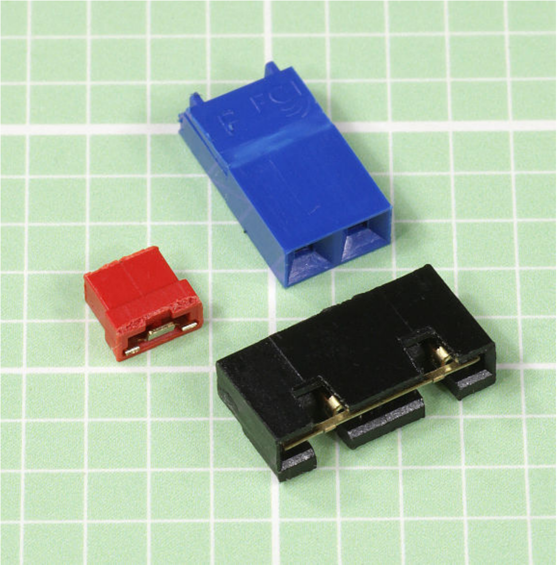

Figure 3-1. *Three jumpers containing two sockets spaced 2mm (left), two sockets spaced 0.1" (top right), and four sockets, each pair spaced 0.1" (bottom right).*

The set of pins with which a jumper is used is often referred to as a *header*. Headers are available with pins in single or dual rows. Some headers are designed to be snapped off to provide the desired number of pins. A dual 28-pin header is shown in Figure 3-2 with a black jumper pushed onto a pair of pins near the midpoint.

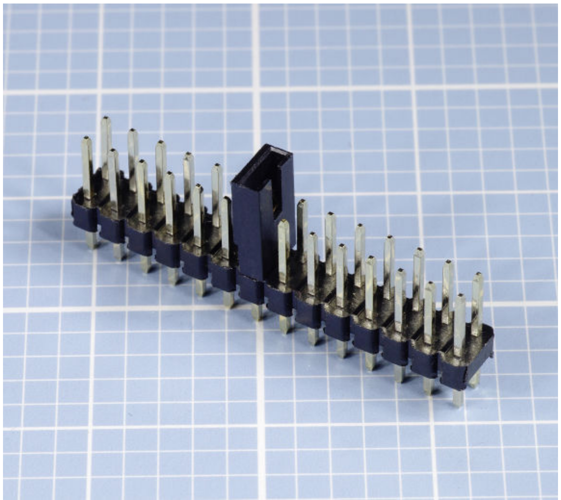

Figure 3-2. *A jumper pushed onto a pair of pins midway along a dual 28-pin header.*

Variants

A *jumper assembly* may be a kit containing not only the jumper but also the array of pins with which it is intended to be used. Check the manufacturer's datasheet to find out exactly what is included.

The most common types of jumpers have two sockets only, but variants are available with as many as 12 sockets, which may be arranged in one or two rows. *Header sockets* may be used as a substitute for purpose-made jumpers, with the advantage that they are often sold in long strips that can be snapped off to provide as many sockets as needed. However, the pins attached to header sockets must be manually connected by soldering small lengths of wire between them.

In some jumpers, the plastic tab extends upward for about half an inch and functions as a finger grip, making the jumper much easier to hold during insertion and removal. This is a desirable feature if there is room to accommodate it.

The sockets inside a jumper are often made from phosphor-bronze, copper-nickel alloy, tin alloy, or brass alloy. They are usually gold-plated, but in some instances are tin-plated.

Rarely, a jumper may consist of a metal strip with U-shaped connections suitable for being used in conjunction with screw terminals. Two jumpers of this type are shown in Figure 3-3. They should not be confused with high-amperage fuses that look superficially similar.

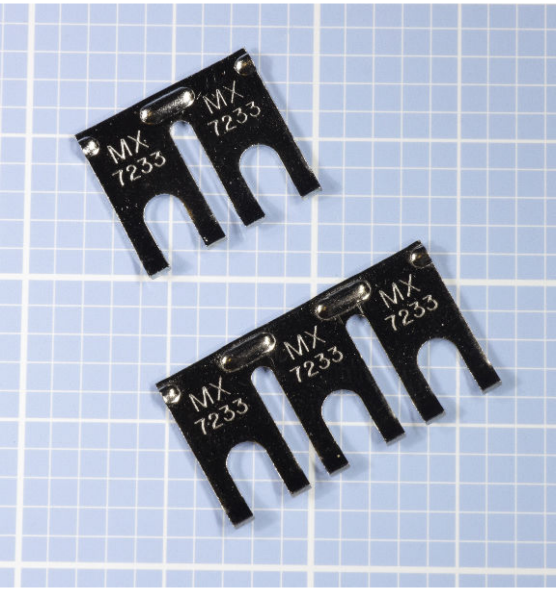

Figure 3-3. *These jumpers are designed to short together two or three screw terminals.*

Values

The spacing between the sockets in a jumper is referred to as its *pitch*. As previously noted, 0.1" and 2mm are the most popular values.

A typical maximum rating for a jumper of 0.1" pitch is 2A or 2.5A at 250V.

How to Use it

A jumper may activate a “set it and forget it” circuit function. An example would be the factory configuration of a product to work with 115VAC or 230VAC power input. End users were expected to set jumpers in some computer equipment sold during the 1980s, but this is no longer the case.

What Can Go Wrong

Jumpers are easily dropped, easily lost, and easily placed incorrectly. When purchasing jumpers, buy extras to compensate for their fragility and the ease of losing them.

Any location where a jumper may be used should be clearly labelled to define the function of each setting.

Cheap, poorly made jumpers may self-destruct from mechanical stresses when removed from their pins. The plastic casing can come away, leaving the sockets clinging naked to the pins protruding from the circuit board. This is another reason why it is a good idea to have a small stock of spare jumpers for emergencies.

Oxidation in jumpers where the contacts are not gold- or silver-plated can create electrical resistance or unreliable connections.

fuse

4

The alternate spelling *fuze* is seldom used.

OTHER RELATED COMPONENTS

- None

What It Does

A fuse protects an electrical circuit or device from excessive current when a metal element inside it melts to create an open circuit. With the exception of *resettable fuses* (discussed separately in "Resettable Fuses" (page 24)), a fuse must be discarded and replaced after it has fulfilled its function.

When high current melts a fuse, it is said to *blow* or *trip* the fuse. (In the case of a resettable fuse, only the word *trip* is used.)

A fuse can work with either AC or DC voltage, and can be designed for almost any current. In residential and commercial buildings, *circuit breakers* have become common, but a large cartridge fuse may still be used to protect the whole system from short-circuits or from overcurrent caused by lightning strikes on exposed power lines.

In electronic devices, the **power supply** is almost always fused.

Schematic symbols for a fuse are shown in Figure 4-1. Those at the right and second from right are most frequently used. The one in the center is approved by ANSI, IEC, and IEEE but is seldom seen. To the left of that is the fuse symbol understood by electrical contractors in architectural plans. The symbol at far left used to be common but has fallen into disuse.

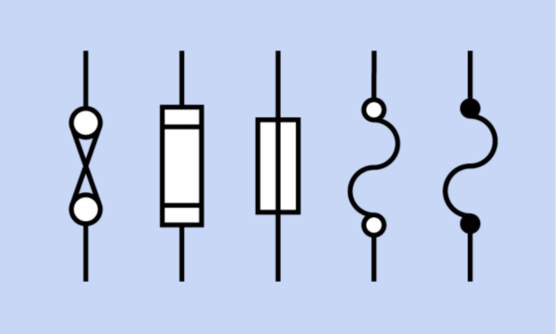

Figure 4-1. *Alternate schematic symbols for a fuse. See text for explanation.*

How It Works

The *element* in a fuse is usually a wire or thin metal strip mounted between two terminals. In a cartridge fuse, it is enclosed in a glass or ceramic cylinder with a contact at each end, or in a small metallic can. (Old-style, large, high-amperage fuses may be packaged in a paper or cardboard tube.) The traditional glass cartridge allows visual inspection to confirm that the fuse has blown.

A fuse responds only to current, not to voltage. When choosing a fuse that will be reliable in conditions of steady current consumption, a safe rule is to figure the maximum amperage when all components are functioning and add 50%. However, if current surges or spikes are likely, their duration will be relevant. If I is the current surge in amps and t is its duration in seconds, the surge sensitivity of a fuse—which is often referred to verbally or in printed format as I2t—is given by the formula:

```
I2t = I² * t
```

Some semiconductors also have an I2t rating, and should be protected with a similarly rated fuse.

Any fuse will present some resistance to the current flowing through it. Otherwise, the current would not generate the heat that blows the fuse. Manufacturer datasheets list the voltage drop that the internal resistance of a fuse is likely to introduce into a circuit.

Values

The *current rating* or *rated current* of a fuse is usually printed or stamped on its casing, and is the maximum flow that it should withstand on a continuous basis, at the ambient temperature specified by the manufacturer (usually 25 degrees Centigrade). The ambient temperature refers to the immediate environment of the fuse, not the larger area in which it may be located. Note that in an enclosure containing other components, the temperature is usually significantly higher than outside the enclosure.

Ideally a fuse should function reliably and indefinitely at its rated maximum amperage, but should blow just as reliably if the current rises by approximately 20% beyond the maximum. In reality, manufacturers recommend that continuous loading of a fuse should not exceed 75% of its rating at 25 degrees Centigrade.

The *voltage rating* or *rated voltage* of a fuse is the maximum voltage at which its element can be counted on to melt in a safe and predictable manner when it is overloaded by excess current. This is sometimes known as the *breaking capacity*. Above that rating, the remaining pieces of the fuse element may form an arc that sustains some electrical conduction.

A fuse can always be used at a lower voltage than its rating. If it has a breaking capacity of 250V, it will still provide the same protection if it is used at 5V.

Four differently rated glass cartridge fuses are shown in Figure 4-2. The one at the top is a slow-blowing type, rated at 15A. Its element is designed to absorb heat before melting. Below it is a 0.5A fuse with a correspondingly thinner element. The two smaller fuses are rated at 5A each. The center two fuses have a maximum voltage rating of 250V, while the one at the top is rated at 32V and the one at the bottom is rated at 350V. Clearly, the size of a fuse should never be used as a guide to its ratings.

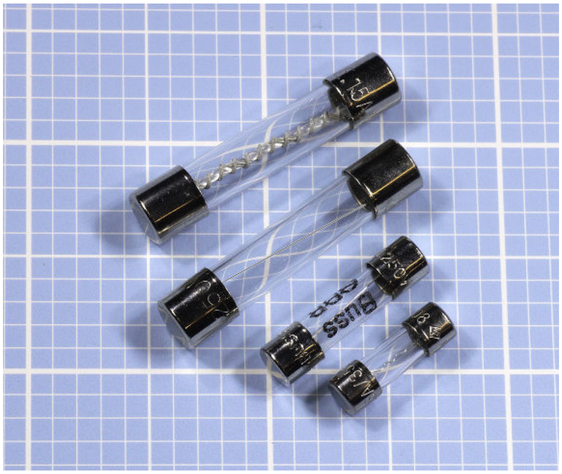

Figure 4-2. *Four glass cartridge fuses. See text for details.*

Variants

Early power fuses in residential buildings consisted of bare nichrome wire wrapped around a porcelain holder. In the 1890s, Edison developed *plug fuses* in which the fuse was contained in a porcelain module with a screw thread, compatible with the base of an incandescent bulb. This design persisted in some U.S. urban areas for more than 70 years, is still found in old buildings, and is still being manufactured.

Small Cartridge Fuses

Small cartridge fuses for appliances and electronics equipment—such as those shown in Figure 4-2--are available in sizes tabulated in Figure 4-3. With the exception of the 4.5mm diameter fuse (a European addition), these sizes were originally measured in inches; today, they are often described only with the equivalent metric measurement. Any cartridge fuse is usually available with the option of a lead attached to it at each end, so that it can be used as a through-hole component.

Fuse type	Diameter (inches)	Diameter (metric)	Length (inches)	Length (metric)
1AG	1/4"	6mm	5/8"	16mm
2AG	0.177"	4.5mm	0.588"	15mm
3AG	1/4"	6mm	1-1/4"	32mm
4AG	9/32"	7mm	1-1/4"	32mm
5AG	13/32"	10mm	1-1/2"	38mm
7AG	1/4"	6mm	7/8"	22mm
8AG	1/4"	6mm	1"	25mm

Figure 4-3. *The approximate physical sizes of commonly used small glass or ceramic cartridge fuses are shown here with the codes that are often used to identify them.*

Fuses may be fast acting, medium acting, or slow-blowing, the last of which may alternatively be referred to as *delay fuses*. Extra-fast–acting fuses are available from some manufacturers. The term *Slo-Blo* is often used but is actually a trademark of Littelfuse. None of the terms describing the speed of action of a fuse has been standardized with a specific time or time range.

Some cartridge fuses are available in a ceramic format as an alternative to the more common glass cylinder. If accidental application of extremely high current is possible (for example, in a multimeter that can be set to measure amps, and may be accidentally connected across a powerful battery), a ceramic cartridge is preferable because it contains a filler that will help to stop an arc from forming. Also, if a fuse is physically destroyed by application of very high current, ceramic fragments may be preferable to glass fragments.

Automotive Fuses

Automotive fuses are identifiable by their use of blades designed for insertion in flat sockets where the fuse is unlikely to loosen as a result of vibration or temperature changes. The fuses come in various sizes, and are uniformly color-coded for easy identification.

A selection of automotive fuses is shown in Figure 4-4. The type at the top is typically described as a “maxi-fuse” while the type at bottom-left is a “mini-fuse.” Here again, size is irrelevant to function, as all three of those pictured are rated 30A at 32V.

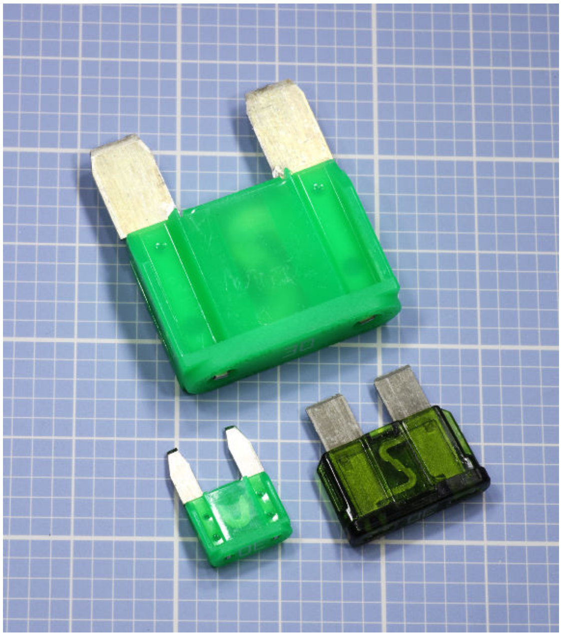

Figure 4-4. *Three automotive fuses. All have the same rating: 30A at 32V.*

In Figure 4-5, the largest of the fuses from Figure 4-4 has been cut open to reveal its element.

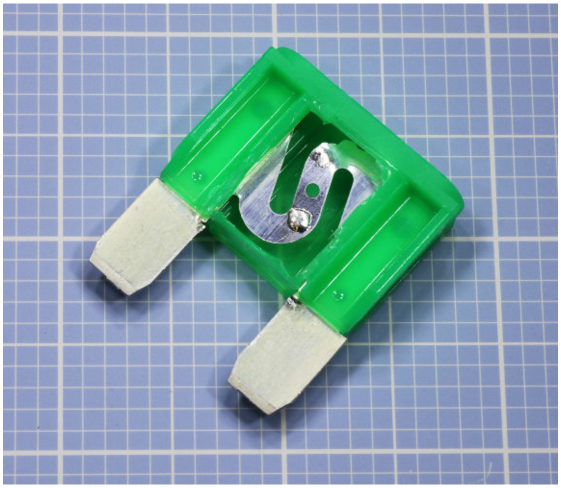

Figure 4-5. *The largest fuse from the previous figure, cut open to reveal its element.*

Usually automotive fuses are mounted together in a block, but if aftermarket accessory equipment is added, it may be protected by an inline fuse in a holder that terminates in two wires. This is shown with two sample fuses in Figure 4-6. Similar inline fuse holders are manufactured for other types of fuses.

Strip Fuses

High-amperage fuses for vehicles may be sold in "strip fuse" format, also known as a *fusible link*, designed to be clamped between two screw-down terminals. Since some **jumpers** may look very similar, it is important to keep them separate. A strip fuse is shown in Figure 4-7.

Through-Hole Fuses

Small fuses with radial leads, which seem appropriate for through-hole insertion in printed circuit boards, are actually often used in conjunction with appropriate sockets, so that they can be easily replaced. They are described in catalogues as "subminiature fuses" and are typically found in laptop computers and their power supplies, also televisions, battery chargers, and air conditioners. Three examples are shown in Figure 4-8. All have slow-blowing characteristics.

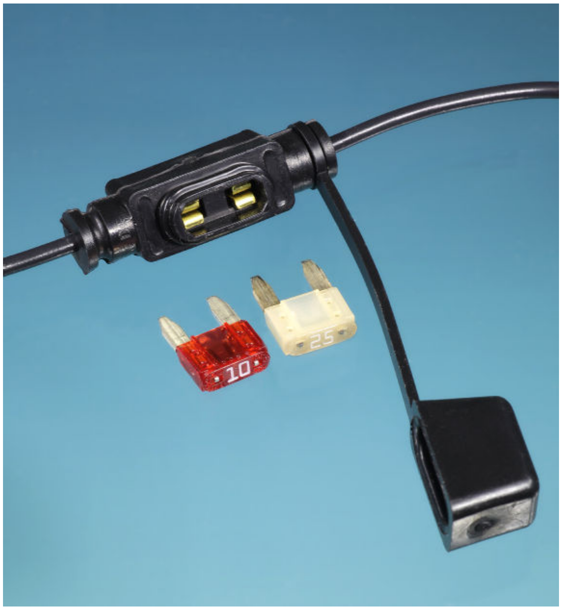

Figure 4-6. *Two blade-type fuses, commonly used for automotive applications, shown with an inline fuse holder. The plastic cap, at right, is closed over the holder when a fuse has been installed.*

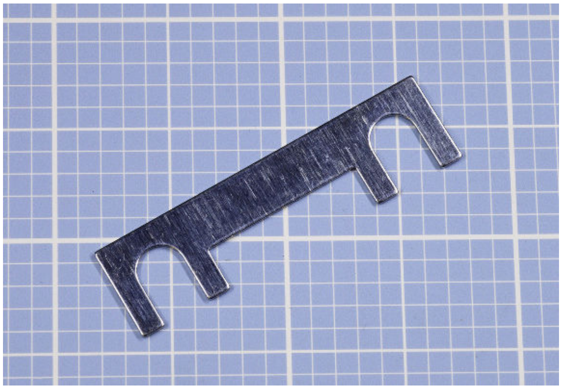

Figure 4-7. *This strip fuse is intended for use in diesel vehicles. The example shown is rated 100A at 36V.*

Resettable Fuses

Properly known as a *polymeric positive temperature coefficient* fuse (often abbreviated *PTC* or *PPTC*), a *resettable fuse* is a solid-state, encapsulated component that greatly increases its resistance in response to a current overload, but gradually returns to its original condition when the flow of current is discontinued. It can be thought

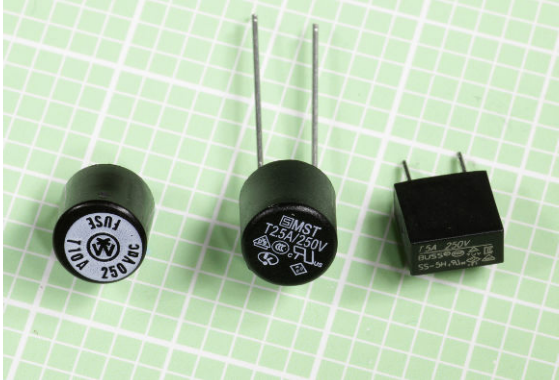

Figure 4-8. *Three subminiature fuses terminating in wire leads. From left to right: 10A at 250V, 2.5A at 250V, and 5A at 250V.*

of as a **thermistor** that has a nonlinear response. Three through-hole examples are shown in Figure 4-9. While different sizes of cartridge fuse may share the same ratings, differently rated resettable fuses may be identical in size. The one on the left is rated 40A at 30V, while the one on the right is rated 2.5A at 30V. (Note that the codes printed on the fuses are not the same as their manufacturer part numbers.) The fuse at the top is rated 1A at 135V.

When more than the maximum current passes through the fuse, its internal resistance increases suddenly from a few ohms to hundreds of thousands of ohms. This is known as *tripping* the fuse. This inevitably entails a small delay, but is comparable to the time taken for a slow-blowing fuse to respond.

A resettable fuse contains a polymer whose crystalline structure is loaded with graphite particles that conduct electricity. As current flowing through the fuse induces heat, the polymer transitions to an amorphous state, separating the graphite particles and interrupting the conductive pathways. A small current still passes through the component, sufficient to maintain its amorphous state until power is disconnected.

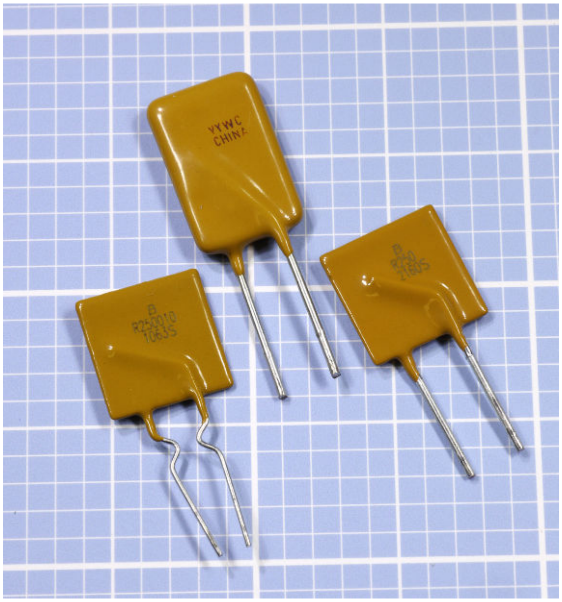

Figure 4-9. *Some through-hole resettable fuses. See text for details.*

After the resettable fuse cools, it gradually recrystallizes, although its resistance does not fall back completely to its original value for more than an hour.

The maximum safe level of current for a resettable fuse is known as the *hold current*, while the current that triggers its response is termed the *trip current*. Resettable fuses are available with trip-current ratings from 20mA to 100A. While conventional appliance and electronics fuses may be rated as high as 600V, resettable fuses are seldom rated above 100V.

Typical cartridge fuses are affected only to a minor extent by temperature, but the current rating of a resettable fuse may diminish to 75% of its normal value at 50 degrees Centigrade and may drop to 50% of its normal value at 80 degrees Centigrade. In other words, a fuse that is rated for 4A at 25 degrees may tolerate a maximum of only 3A when it operates at twice that temperature. See Figure 4-10.

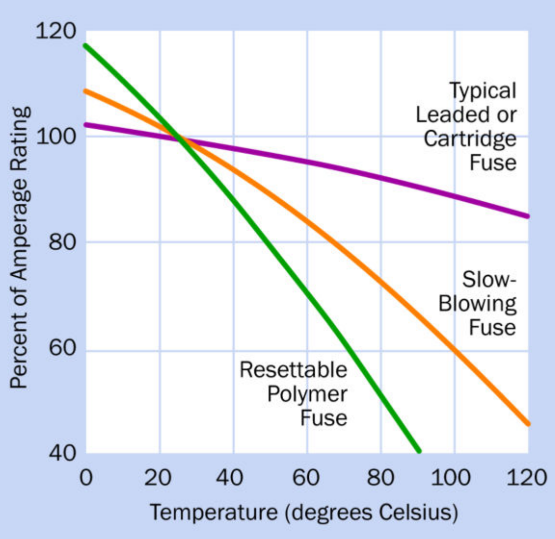

Figure 4-10. *The three curves give a very approximate idea of the temperature sensitivity of three types of fuses commonly used to protect electronic equipment. The left-hand scale provides an approximate value for the amperage which will trip the fuse.*

Conventional slow-blowing fuses are temperature-sensitive, but to a lesser degree than resettable fuses.

Resettable fuses are used in computer power supplies, USB power sources, and loudspeaker enclosures, where they protect the speaker coils from being overdriven. They are appropriate in situations where a fuse may be tripped relatively often, or where technically unsophisticated users may feel unable to replace a fuse or reset a circuit breaker.

Brand names for resettable fuses include PolySwitch, OptiReset, Everfuse, Polyfuse, and Multifuse. They are available in *surface-mount* (SMT) packages or as *through-hole* components, but not in cartridge format.

Surface Mount Fuses

Because surface-mount fuses are difficult or impossible to replace after they have been soldered onto the board, they are often resettable.

A surface-mount resettable fuse approximately 0.3″ square is shown in Figure 4-11. It is rated for 230V and has an internal resistance of 50 ohms. Its hold current is 0.09A and its trip current is 0.19A.

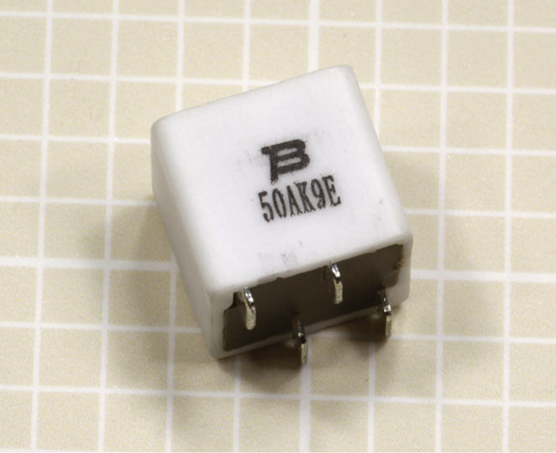

Figure 4-11. *A surface-mount resettable fuse. See text for details.*

How to Use it

Any equipment that may be plugged into a wall outlet should be fused, not only to protect its components but also to protect users who may open the box and start investigating with a screwdriver.

Equipment that contains powerful motors, pumps, or other inductive loads should be protected with slow-blowing fuses, as the initial surge of current when the equipment is switched on is likely to rise well above the rating of the fuse. A slow-blowing fuse will tolerate a surge for a couple of seconds. Other fuses will not.

Conversely, fast-acting fuses should be used with electronic equipment, especially integrated circuits that are quickly and easily damaged.

Any device using substantial **battery** power should be fused because of the unpredictable and generally bad behavior of batteries when they are short-circuited. Parallel connections between multiple large batteries should be fused to avoid the possibility that a highly charged battery may attempt to recharge its neighbor(s). Large "J size" fuses rated from 125A to 450A have become common in the solar power community, where banks of lead-acid batteries are often used. These fuses have a thick brass tab at each end, drilled so that they can be bolted into place. Alternatively, they will push-fit into an appropriate *fuseholder*.

For cartridge fuses up to 1/4" in diameter that don't have leads attached, appropriately sized fuseholders are available in several formats:

Panel mounted fuse enclosure is probably the most common, consisting of a plastic tube with a spring-contact at the bottom, and a plastic cap with a second contact inside. The cap either screws onto the tube of the fuse, or is pushed down and turned to hold it in place. A nut is provided to secure the fuseholder after it has been inserted into a hole drilled in the panel. The fuse is dropped into the tube, and the cap is applied. This type of holder is available in full-length or shorter, "low profile" formats. A low-profile holder is shown in Figure 4-12. It is shown assembled at right, with its component parts disassembled alongside.

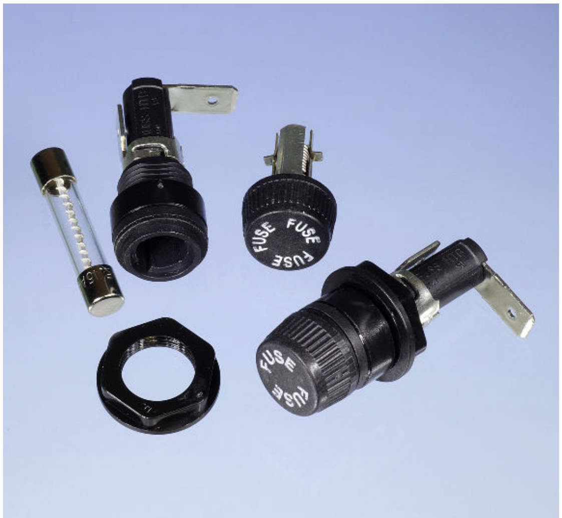

Figure 4-12. *A low-profile panel-mounted fuse holder shown disassembled (left) and assembled (right).*

Circuit board mounted fuse enclosure is basically the same as the panel-mounted version, but with through-hole solder pins attached.

Fuse block is a small plastic block with two clips on its upper surface for insertion of a cartridge fuse.

Fuse clips can be bought individually, with solder pins for through-hole mounting.

Inline fuse holder is designed to be inserted in a length of wire. Usually made of plastic, it will either terminate it, wires or will have metal contacts to crimp or solder at each end. See Figure 4-6.

Through-hole fuse holders are available for subminiature fuses.

What Can Go Wrong

Repeated Failure

When a fuse in a circuit blows frequently, this is known as *nuisance opening*. Often it can result from failure to take into account all the aspects of the circuit, such as a large filtering capacitor in a power supply that draws a major surge of current when the power supply is switched on. The formally correct procedure to address this problem is to measure the power surge, properly known as *peak inrush current*, with an oscilloscope, calculate the $I^2 * t$ of the wave form, and select a fuse with a rating at least 5 times that value.

A fuse should never be replaced with an equivalent length of wire or any other conductor.

Soldering Damage

When a through-hole or surface-mount fuse is soldered into place, heat from the soldering process can cause the soft metal element inside the fuse to melt partially and reflow. This is likely to change the rating of the fuse. Generally, fuses should be treated with the same caution as semiconductors when they are fixed in place with solder.

Placement

A fuse should be placed close to the power source or power input point in a circuit, so that it protects as much of the circuit as possible.

pushbutton

5

Often referred to as a *pushbutton switch* and sometimes as a *momentary switch*. In this encyclopedia, a pushbutton is considered separately from a **switch**, which generally uses a lever-shaped actuator rather than a button, and has at least one *pole* contact where a pushbutton generally has contacts that are not distinguishable from each other.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)
- **rotary switch** (See Chapter 7)

What It Does

A pushbutton contains at least two contacts, which close or open when the button is pressed. Usually a spring restores the button to its original position when external pressure is released. Figure 5-1 shows schematic symbols for pushbuttons. The symbols that share each blue rectangle are functionally identical. At top is a normally-open single-throw pushbutton. At center is a normally-closed single-throw pushbutton. At bottom is a double-throw pushbutton.

Unlike a **switch**, a basic pushbutton does not have a primary contact that can be identified as the *pole*. However, a single pushbutton may close or open two separate pairs of contacts, in which case it can be referred to, a little misleadingly, as a double-pole pushbutton. See Figure 5-2. Different symbols are used for slider pushbuttons with multiple contact pairs; see "Slider" (page 31).

A generic full-size, two-contact pushbutton is shown in Figure 5-3.

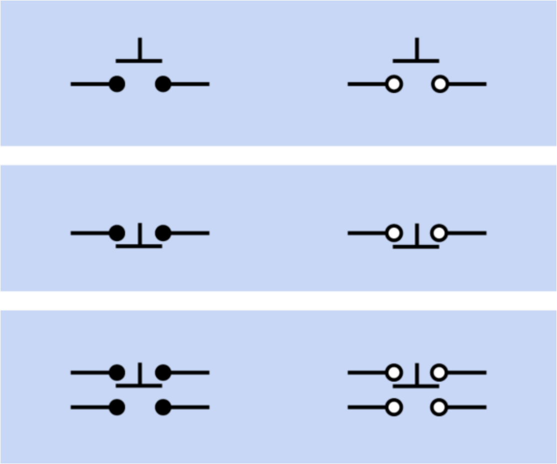

Figure 5-1. *Commonly used schematic symbols to represent a simple pushbutton. See text for details.*

How It Works

Figure 5-4 shows a cross-section of a pushbutton that has a single steel return spring, to create resistance to downward force on the button, and a pair of springs above a pair of contacts, to hold each contact in place and make a firm connection when the button is pressed. The two upper contacts are electrically linked, although this feature is not shown.

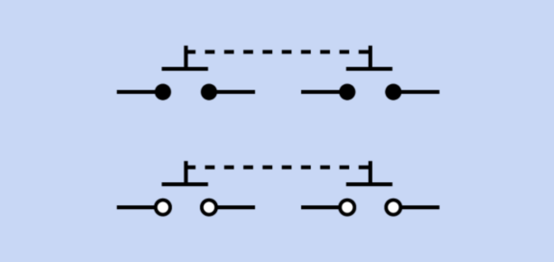

Figure 5-2. *Commonly used schematic symbols to represent a double-pole pushbutton.*

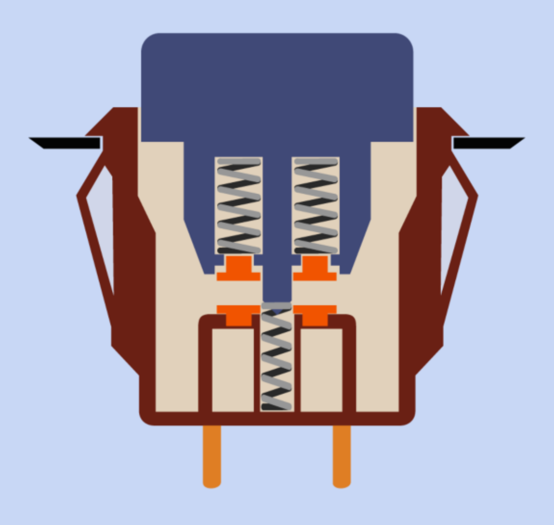

Figure 5-4. *Cross-section of a pushbutton showing two spring-loaded contacts and a single return spring.*

Variants

Poles and Throws

Abbreviations that identify the number of poles and contacts inside a pushbutton are the same as the abbreviations that identify those attributes in a **switch**. A few examples will make this clear:

SPST, also known as 1P1T
Single pole, single throw

DPST also known as 2P1T
Double pole, single throw

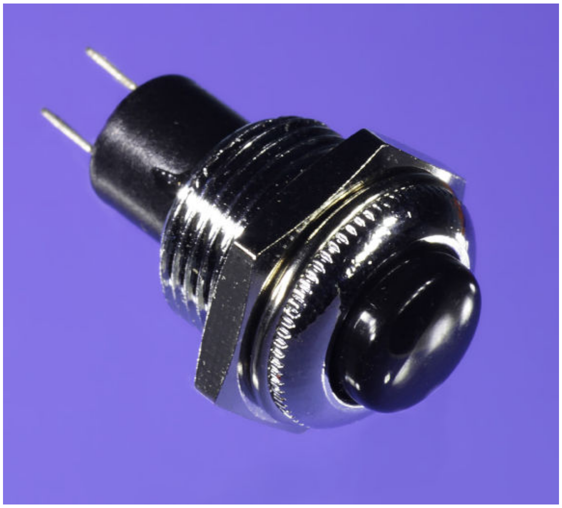

Figure 5-3. *The simplest, traditional form of pushbutton, in which pressing the button creates a connection between two contacts.*

SPDT also known as 1P2T
Single pole, double throw

3PST also known as 3P1T
Three pole, single throw

While a switch may have an additional center position, pushbuttons generally do not.

On-Off Behavior

Parentheses are used to indicate the momentary state of the pushbutton while it is pressed. It will return to the other state by default.

OFF-(ON) or (ON)-OFF
Contacts are normally open by default, and are closed only while the button is pressed. This is sometimes described as a *make-to-make* connection, or as a *Form A* pushbutton.

ON-(OFF) or (OFF)-ON
Contacts are normally closed by default, and are open only while the button is pressed. This is sometimes described as a *make-to-break* connection, or as a *Form B* pushbutton.

ON-(ON) or (ON)-ON
This is a double-throw pushbutton in which one set of contacts is normally closed. When the button is pressed, the first set of contacts

is opened and the other set of contacts is closed, until the button is released. This is sometimes described as a *Form C* pushbutton.

For a single-throw pushbutton, the terms NC or NO may be used to describe it as *normally closed* or *normally open*.

Slider

This type, also known as a *slide* pushbutton, contains a thin bar or rod that slides in and out of a long, narrow enclosure. Contacts on the rod rub across secondary contacts inside the enclosure. Closely resembling a slider switch, it is cheap, compact, and well adapted for multiple connections (up to 8 separate poles in some models). However, it can only tolerate low currents, has limited durability, and is vulnerable to contamination.

A four-pole, double-throw pushbutton is shown in Figure 5-5. A variety of plastic caps can be obtained to press-fit onto the end of the white nylon actuator.

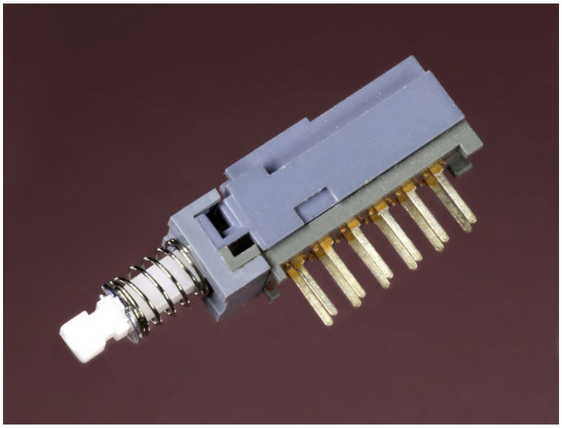

Figure 5-5. *A 4PDT slider pushbutton, shown without the cap that can be snapped onto the end of the actuator.*

Figure 5-6 shows schematic symbols for two possible slide pushbuttons, with a black rectangle indicating each sliding contact. The lead that functions as a pole is marked with a P in each case. Standardization for slide pushbutton schematic symbols does not really exist, but these examples are fairly typical. An insulating section that connects the sliding contacts internally is shown here as a gray rectangle, but in some datasheets may appear as a line or an open rectangle.

Since the symbols for a slide pushbutton may be identical to the symbols for a slide switch, care must be taken when examining a schematic, to determine which type of component is intended.

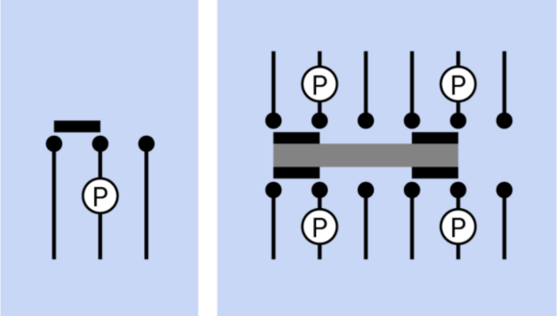

Figure 5-6. *Left: schematic symbol for a simple SPDT slide pushbutton, where a movable contact shorts together either the left pair or right pair of fixed contacts. Right: A 4PDT pushbutton in which the same principle has been extended. The movable contacts are attached to each other mechanically by an insulator. Each pole terminal is marked with a P.*

Styles

Many pushbutton switches are sold without caps attached. This allows the user to choose from a selection of styles and colors. Typically the cap is a push-fit onto the end of the rod or bar that activates the internal contacts. Some sample caps are shown in Figure 5-7, alongside a DPDT pushbutton. Any of the caps will snap-fit onto its actuator.

An *illuminated* pushbutton contains a small *incandescent bulb*, *neon bulb*, or **LED** (*light-emitting diode*). The light source almost always has its own two terminals, which are isolated from the other terminals on the button housing and can be wired to activate the light when the button is pressed, when it is released, or on some other basis. Pushbuttons containing LEDs usually

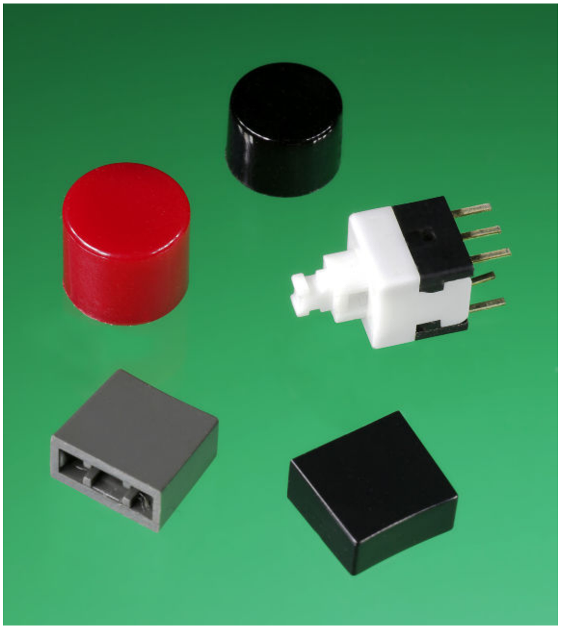

Figure 5-7. *Caps (buttons or knobs) that may be sold as separate accessories for some pushbuttons, shown here alongside a compatible pushbutton switch.*

require external series resistors, which should be chosen according to the voltage that will be used. See the **LED** entry in Volume 2 for additional commentary on appropriate series resistors. An example of an illuminated pushbutton is shown in Figure 5-8. This is a DPDT component, designed to be mounted on a printed circuit board, with an additional lead at each end connecting with an internal LED underneath the translucent white button.

Termination and Contact Plating

These options are the same as for a **switch** and are described in that entry.

Mounting Style

The traditional panel-mounted button is usually secured through a hole in the panel by tightening a nut that engages with a thread on the bushing of the pushbutton. Alternatively, a pushbutton housing can have flexible plastic protrusions on either side, allowing it to be snapped into place in an appropriate-sized panel cutout. This style is shown in Figure 5-4.

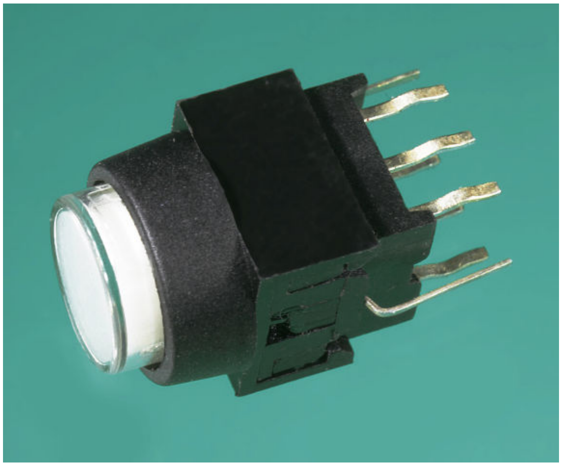

Figure 5-8. *This pushbutton contains an LED underneath the white translucent button.*

PC pushbuttons (pushbuttons mounted in a printed circuit board, or PCB) are a common variant. After the component has been installed in the circuit board, either the button must align with a cutout in the front panel and poke through it when the device is assembled, or an external (non-electrical) button that is part of the product enclosure must press on the actuator of the pushbutton after assembly.

Surface-mount pushbuttons that allow direct fingertip access are uncommon. However, about one-quarter of tactile switches are designed for surface mount at the time of writing. They are typically found beneath membranes that the user presses to activate the switch beneath—for example, in remotes that are used to operate electronic devices.

Sealed or Unsealed

A sealed pushbutton will include protection against water, dust, dirt, and other environmental hazards, at some additional cost.

Latching

This variant, also known as a *press-twice* pushbutton, contains a mechanical ratchet, which is rotated each time the button is pressed. The first press causes contacts to latch in the closed state. The second press returns the contacts to the open state, after which, the process repeats. This press-twice design is typically found on flashlights, audio equipment, and in automotive applications. While *latching* is the most commonly used term, it is also known as *push-push*, *locking*, *push-lock push-release*, *push-on push-off*, and *alternate*.

In a latching pushbutton with *lockdown*, the button is visibly lower in the latched state than in the unlatched state. However, buttons that behave this way are not always identified as doing so on their datasheets.

A six-pole double-throw pushbutton that latches and then unlatches each time it is pressed is shown in Figure 5-9.

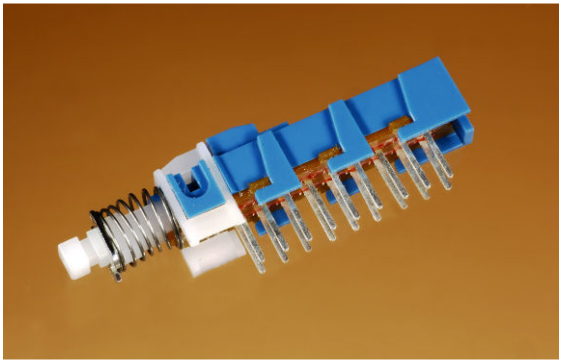

Figure 5-9. *This 6PDT pushbutton latches and then unlatches, each time it is pressed.*

Two more variants are shown in Figure 5-10. On the right is a simple DPDT latching pushbutton with lockdown. On the left is a latching pushbutton that cycles through four states, beginning with one "off" state, the remaining three connecting a different pair of its wires in turn.

A simple OFF-(ON) button may appear to have a latching output if it sends a pulse to a **microcontroller** in which software inside the microcontroller toggles an output between two states. The microcontroller can step through an unlimited number of options in response to each button press. Examples are found on cellular phones or portable media players.

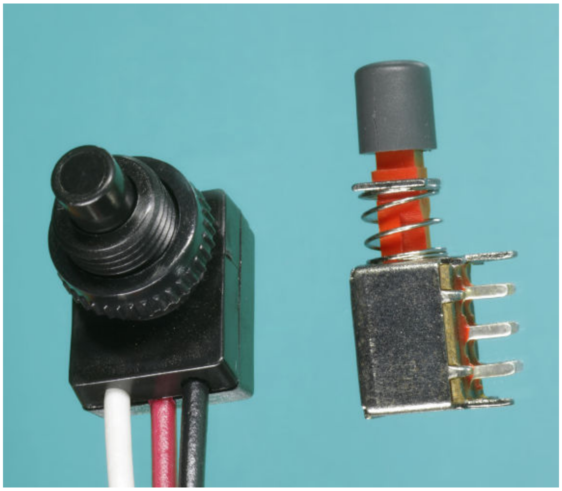

Figure 5-10. *At right, a simple DPDT latching pushbutton with lockdown. At left, this pushbutton cycles through four states, one of them an "off" state, the others connecting a different pair of its wires in turn.*

A mechanically latching pushbutton has a higher failure rate than a simple OFF-(ON) button, as a result of its internal mechanism, but has the advantage of requiring no additional microcontroller to create its output. Microcontrollers are discussed in Volume 2.

Foot Pedal

Foot pedal pushbuttons generally require more actuation force than those intended for manual use. They are ruggedly built and are commonly found in vacuum cleaners, audio-transcription foot pedals, and "stomp boxes" used by musicians.

Keypad

A keypad is a rectangular array of usually 12 or 16 OFF-(ON) buttons. Their contacts are accessed via a *header* suitable for connection with a ribbon cable or insertion into a printed circuit board. In some keypads, each button connects with a separate contact in the header, while all the buttons share a common ground. More often, the buttons are *matrix encoded*, meaning that each of them bridges a unique pair of conductors in a matrix. A 16-button matrix is shown in Figure 5-11. This configuration is suitable for polling by a **microcontroller**, which can be programmed to send an output pulse to each of the four horizontal wires in turn. During each pulse, it checks the remaining four vertical wires in sequence, to determine which one, if any, is carrying a signal. Pullup or pulldown resistors should be added to the input wires to prevent the inputs of the microcontroller from behaving unpredictably when no signal is present. The external appearance of two keypads is shown in Figure 5-12.

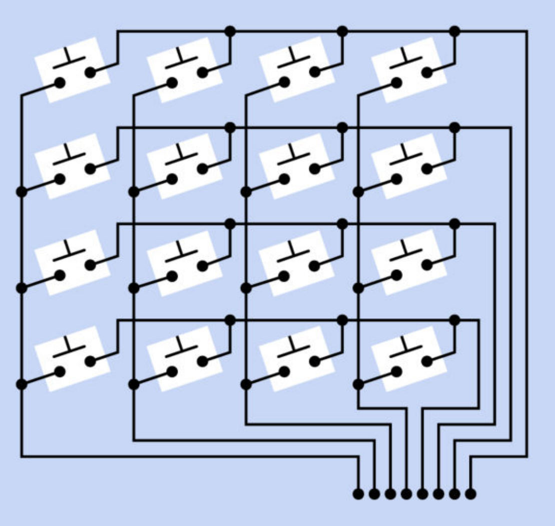

Figure 5-11. *Buttons in a numeric keypad are usually wired as a matrix, where each button makes a connection between a unique pair of wires. This system is suitable for being polled by a microcontroller.*

Tactile Switch

Despite being called a switch, this is a miniature pushbutton, less than 0.4" square, designed for insertion in a printed-circuit board or in a solderless breadboard. It is almost always a SPST device but may have four pins, one pair connected to each contact. Tactile switches may be PC-mounted behind membrane pads. An example is shown in Figure 5-13.

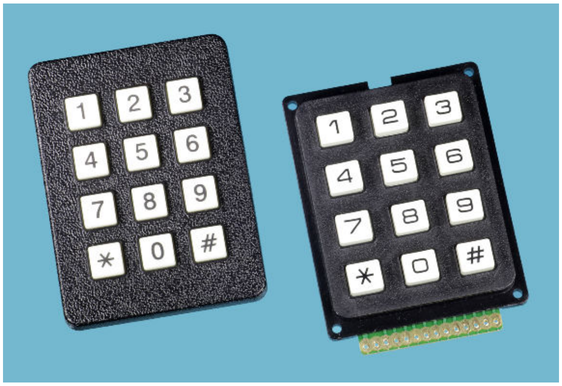

Figure 5-12. *The keypad on the left is matrix-encoded, and is polled via seven through-hole pins that protrude behind it. The keypad on the right assigns each button to a separate contact in its header. See the text for details about matrix encoding.*

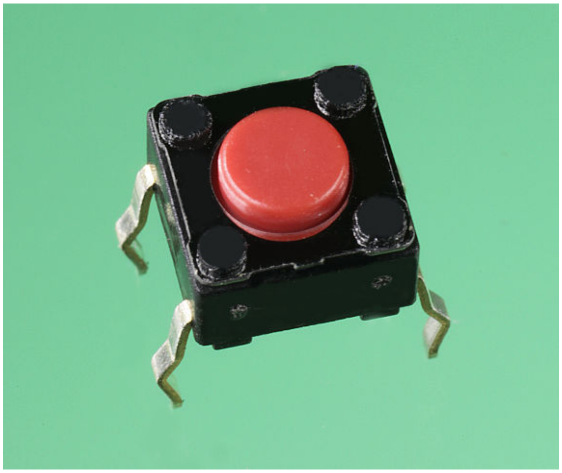

Figure 5-13. *A typical tactile switch.*

Membrane Pad

Typically found on devices such as microwave ovens where contacts must be sealed against particles and liquids. Finger pressure on a membrane pad closes hidden or internal pushbuttons. They are usually custom-designed for specific product applications and are not generally available as generic off-the-shelf components. Some surplus pads may be found for sale on auction websites.

Radio Buttons

The term *radio buttons* is sometimes used to identify a set of pushbuttons that are mechanically interlinked so that only one of them can make an electrical connection at a time. If one button is pressed, it latches. If a second button is pressed, it latches while unlatching the first button. The buttons can be pressed in any sequence. This system is useful for applications such as component selection in a stereo system, where only one input can be permitted at a time. However, its use is becoming less common.

Snap-Action Switches

A snap-action switch (described in detail in the **switch** section of this encyclopedia) can be fitted with a pushbutton, as shown in Figure 5-14. This provides a pleasingly precise action, high reliability, and capability of switching currents of around 5A. However, snap-action switches are almost always single-pole devices.

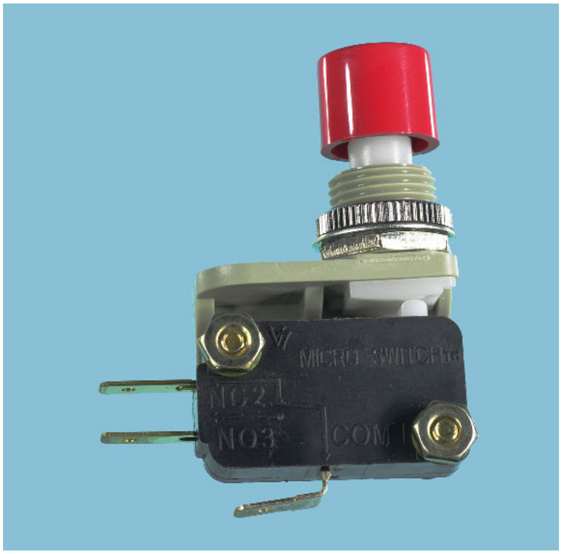

Figure 5-14. *A pushbutton mounted on top of a SPDT snap-action switch.*

Emergency Switch

An emergency switch is a normally-closed device, usually consisting of a large pushbutton that clicks firmly into its "off" position when pressed, and does not spring back. A flange around the button allows it to be grasped and pulled outward to restore it to its "on" position.

Values

Pushbutton current ratings range from a few mA to 20A or more. Many pushbuttons have their current ratings printed on them but some do not. Current ratings are usually specified for a particular voltage, and may differ for AC versus DC.

How to Use it

Issues such as appearance, tactile feel, physical size, and ease of product assembly tend to dictate the choice of a pushbutton, after the fundamental requirements of voltage, current, and durability have been satisfied. Like any electromechanical component, a pushbutton is vulnerable to dirt and moisture. The ways in which a device may be used or abused should be taken into account when deciding whether the extra expense of a sealed component is justified.

When a pushbutton controls a device that has a high inductive load, a *snubber* can be added to minimize arcing. See "Arcing" (page 47) in the **switch** entry of this encyclopedia, for additional information.

What Can Go Wrong

No Button

When ordering a pushbutton switch, read datasheets carefully to determine whether a cap is included. Caps are often sold separately and may not be interchangeable between switches from different manufacturers.

Mounting Problems

In a panel-mount pushbutton that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure when the button is pressed. Conversely, overtightening the nut may strip the threads on the pushbutton bushing, especially in cheaper components where the threads are molded into plastic. Consider applying a drop of Loc-Tite or similar adhesive before completely tightening the nut. Nut sizes vary widely, and finding a replacement may be time-consuming.

LED Issues

When using a pushbutton containing an LED, be careful to distinguish the LED power terminals from the switched terminals. The manufacturer's datasheet should clarify this distinction, but the polarity of the LED terminals may not be clearly indicated. If a diode-testing meter function is unavailable, a sample of the switch should be tested with a source of 3 to 5VDC and a 2K series resistor. Briefly touching the power to the LED terminals, through the resistor, should cause the LED to flash dimly if the polarity is correct, but should not be sufficient to burn out the LED if the polarity is incorrect.

Other Problems

Problems such as arcing, overload, short circuits, wrong terminal type, and contact bounce are generally the same as those associated with a **switch**, and are summarized in that entry in this encyclopedia.

switch

The term **switch** refers here to a physically operated mechanical switch, controlled by flipping a lever or sliding a knob. Although there is some overlap of function, **rotary switches** and **pushbuttons** have their own separate entries. Solid-state switching components are described in entries for **bipolar transistor**, **unijunction transistor**, and **field-effect transistor**. Integrated-circuit switching devices will be found in Volume 2. *Coaxial switches* are used for high-frequency signals, and are not included in this encyclopedia. *Multidirectional switches* differentiate up, down, left, right, diagonal, rotational, and other finger inputs, and are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **pushbutton** (See Chapter 5)
- **rotary switch** (See Chapter 7)

What It Does

A switch contains at least two contacts, which close or open when an external lever or knob is flipped or moved. Schematic symbols for the most basic type of on-off switch are shown in Figure 6-1.

The most fundamental type of switch is a *knife switch*, illustrated in Figure 6-2. Although it was common in the earliest days of electrical discovery, today it is restricted to educational purposes in schools, and (in a more robust format) to AC electrical supply panels, where the large contact area makes it appropriate for conducting high amperages, and it can be used for "hot switching" a substantial load.

How It Works

The *pole* of a switch is generally connected with a movable contact that makes or breaks a connection with a secondary contact. If there is only one pole, this is a *single-pole* switch. If there is an additional pole, electrically isolated from the first, with its own contact or set of contacts, this is a *two-pole* switch, also known as a *double-pole* switch. Switches with more than 4 poles are uncommon.

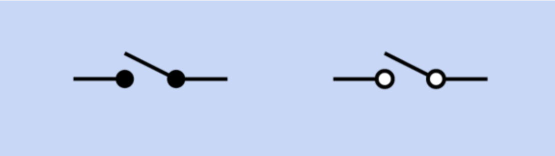

Figure 6-1. *The two most common schematic symbols for a SPST switch, also known as an on-off switch. The symbols are functionally identical.*

If there is only one secondary contact per pole, this is a *single-throw* or *ST* switch, which may also be described as an *on-off* or *off-on* switch. If there is an additional secondary contact per pole, and the pole of the switch connects with the second contact while disconnecting from the first, this is a *double-throw* or *DT* switch, also known as a *two-way* switch.

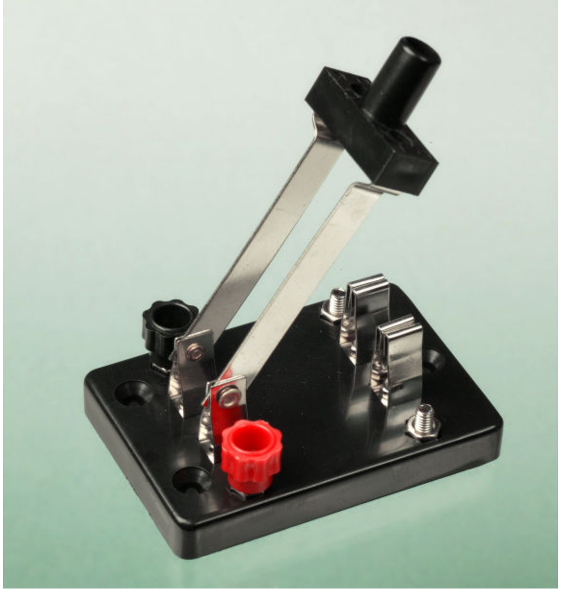

Figure 6-2. *A DPST knife switch intended for the educational environment.*

A double-throw switch may have an additional center position. This position may have no connection (it is an "off" position) or in some cases it connects with a third contact.

Where a switch is spring-loaded to return to one of its positions when manual pressure is released, it functions like a **pushbutton** even though its physical appearance may be indistinguishable from a switch.

Variants

Float switch, *mercury switch*, *reed switch*, *pressure switch*, and *Hall-effect switch* are considered as sensing devices, and will be found in Volume 3.

Terminology

Many different types of switches contain parts that serve the same common functions. The *actuator* is the lever, knob, or toggle that the user turns or pushes. A *bushing* surrounds the actuator on a toggle-type switch. The *common contact* inside a switch is connected with the *pole* of the switch. Usually a *movable contact* is attached to it internally, to touch the secondary contact, also known as a *stationary contact* when the movable contact is flipped to and fro.

Poles and Throws

Abbreviations identify the number of poles and contacts inside a switch. A few examples will make this clear:

SPST also known as 1P1T
Single pole, single throw

DPST also known as 2P1T
Double pole, single throw

SPDT also known as 1P2T
Single pole, double throw

3PST also known as 3P1T
Three pole, single throw

Other combinations are possible.

In Figure 6-3, schematic symbols are shown for double-throw switches with 1, 2, and 3 poles. The dashed lines indicate a mechanical connection, so that all sections of the switch move together when the switch is turned. No electrical connection exists between the poles.

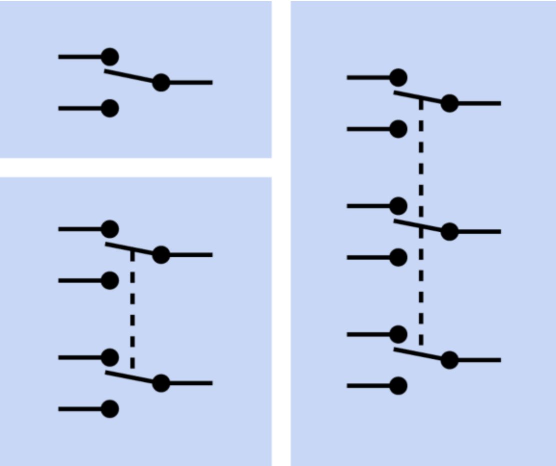

Figure 6-3. *Schematic symbols to represent three types of double-throw switch. Top left: Single-pole. Bottom left: Double-pole. Right: Triple-pole, more commonly known as 3-pole.*

On-Off Behavior

The words ON and OFF are used to indicate the possible states of a switch. The additional word NONE is used by some manufacturers to indicate that a switch does not have a center position. Some manufacturers don't bother with the word NONE, assuming that if the word is omitted, a center position does not exist.

ON-OFF or ON-NONE-OFF

A basic on-off SPST switch with no center position.

ON-ON or ON-NONE-ON

A basic SPDT switch with no center position.

ON-OFF-ON

A double-throw switch with center-off position (no connection when the switch is centered).

ON-ON-ON

A triple-throw switch where the center position connects with its own set of terminals.

Parentheses are used in descriptions of spring-loaded switches to indicate a momentary state that lasts only as long as pressure is applied to the actuator.

(ON)-OFF or OFF-(ON)

A spring-loaded switch that is normally off and returns to that position when pressure is released. Also known as *NO* (normally open), and sometimes described as *FORM A*. Its performance is similar to that of a push-button and is sometimes described as a *make-to-make* connection.

ON-(OFF) or (OFF)-ON

A spring-loaded switch that is normally on and returns to that position when pressure is released. This is sometimes described as a *make-to-break* connection. Also known as *NC* (normally closed), and sometimes described as *FORM B*.

(ON)-OFF-(ON)

A spring-loaded double-throw switch with a no-connection center position to which it returns when pressure on its actuator is released.

Other combinations of these terms are possible.

Most double-throw switches break the connection with one contact (or set of contacts) before making the connection with the second contact (or set of contacts). This is known as a *break before make* switch. Much less common is a *make before break* switch, also known as a *shorting switch*, which establishes the second connection a moment before the first connection is broken. Use of a shorting switch may cause unforeseen consequences in electronic components attached to it, as both sides of the switch will be briefly connected when the switch is turned.

Snap-Action

Also known as a *limit switch* and sometimes as a *microswitch* or *basic switch*. This utilitarian design is often intended to be triggered mechanically rather than with finger pressure, for example in 3D printers. It is generally cheap but reliable.

Two snap-action switches are shown in Figure 6-4. A sectional view of a snap-action ON-(ON) limit switch is shown in Figure 6-5. The pole contacts are mounted on a flexible strip which can move up and down in the center of the switch. The strip has a cutout which allows an inverted U-shaped spring to flip to and fro. It keeps the contacts pressed together in either of the switch states.

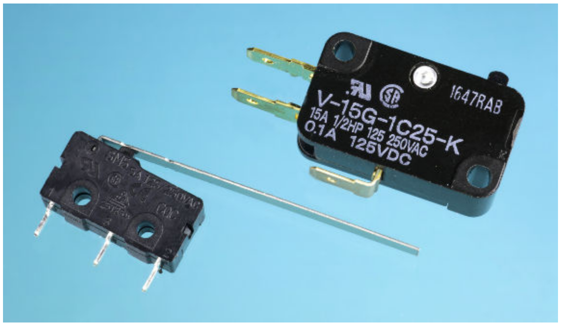

Figure 6-4. *Two SPDT snap-action switches, also known as limit switches. The one on the right is full-size. The one on the left is miniature, with an actuator arm to provide additional leverage. The arm may be trimmed to the required length.*

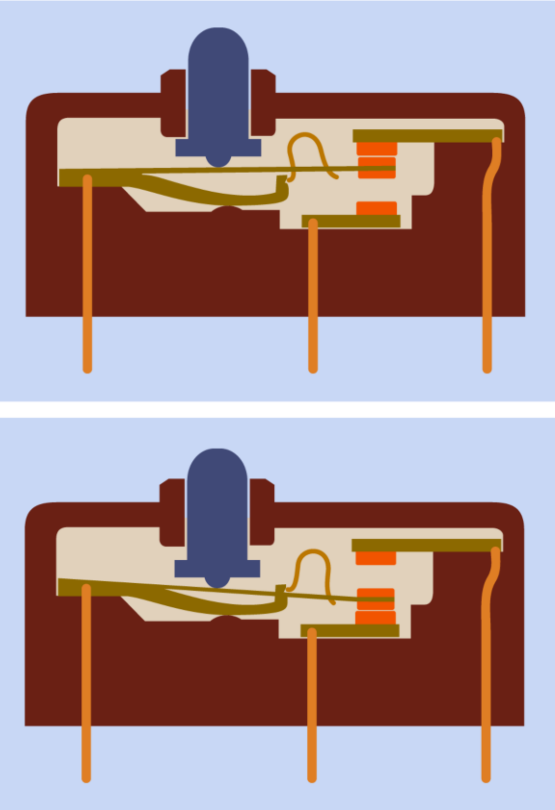

Figure 6-5. *Top: Two contacts inside this limit switch are touching by default. Bottom: When the external button is pressed, it pushes a flexible metal strip downward until it connects with the lower contact. The inverted-U–shaped component is a spring that rests inside a cutout in the flexible strip and resists motion through the central part of its travel.*

The term *snap action* refers to a spring-loaded internal mechanism which snaps to and fro between its two positions. This type of switch is usually SPDT and has a momentary action; in other words, it functions in ON-(ON) mode, although OFF-(ON) and (less often) ON-(OFF) versions are available. The body of the switch is sealed, with a small button protruding through a hole. A thin metal arm may provide additional leverage to press the button. A roller may be mounted at the end of the arm so that the switch can be activated as it slides against a moving mechanical component such as a cam or a wheel. The switch is commonly used to limit the travel or rotation of such a component. Literally thousands of variants are available, in different sizes, requiring different amounts of force for activation. Subminiature snap-action switches can often be actuated by a pressure of only a few grams.

Rocker

Three rocker switches are shown in Figure 6-6. A sectional view of a rocker switch is shown in Figure 6-7. In this design, a spring-loaded ball bearing rolls to either end of a central rocker arm when the switch is turned. Rocker switches are often used as power on-off switches.

Slider

Many types of slider switch (also known as *slide switch*) are widely used as a low-cost but versatile way to control small electronic devices, from clock-radios to stereos. The switch is usually mounted on a circuit board, and its knob or cap protrudes through a slot in the panel. This design is more vulnerable to dirt and moisture than other types of switch. It is usually cheaper than a toggle switch but is seldom designed for use with a high current.

Most slide switches have two positions, and function as SPDT or DPDT switches, but other configurations are less commonly available with more poles and/or positions. A subminiature slide switch is shown in Figure 6-8, while some schematic representations are shown in Figure 6-9,

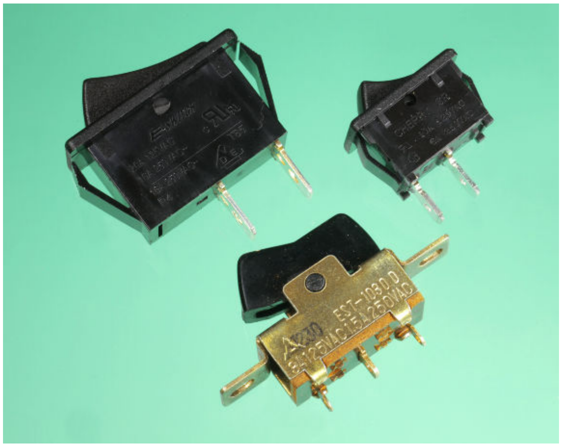

Figure 6-6. *Three rocker switches, the upper two designed for push-insertion into a suitably sized rectangular hole in a panel. The switch at front-center is intended to be screwed in place, and is more than 20 years old, showing that while the choice of materials has changed, the basic design has not.*

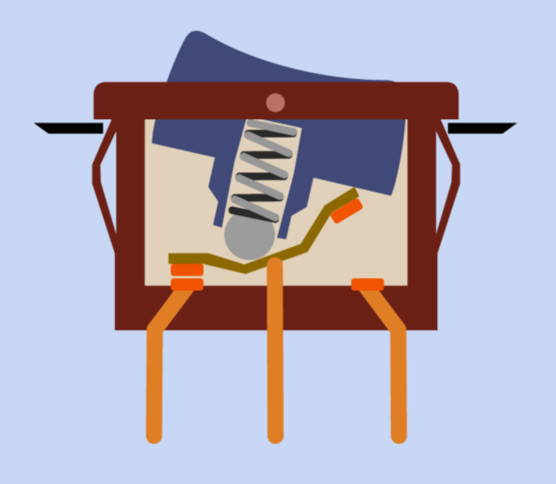

Figure 6-7. *This sectional view of a rocker switch shows a spring-loaded ball-bearing that rolls to and fro along a rocker arm, connecting either pair of contacts when the switch is turned.*

where a black rectangle indicates a sliding internal contact, and a terminal that functions as a pole is identified with letter P in each case. Top left: A SPDT switch using a two-position slider. Top right: A 4PDT slide switch. Bottom left: There are no poles in this switch, as such. The slider can short together any of four pairs of contacts. Bottom right: The slider shorts together three possible pairs of contacts out of four. Here again, there is no pole.

Note that the schematic representation of a slide switch may be identical to that of a slide **push-button**. A schematic should be inspected carefully to determine which type is intended.

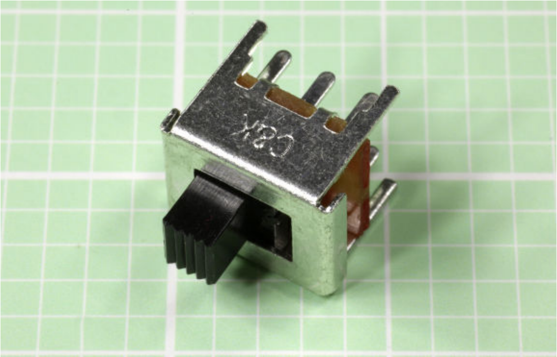

Figure 6-8. *This subminiature slide switch is less than half an inch long, rated 0.3A at 30VDC. Larger versions look almost identical, but can handle only slightly more current.*

The representation of sliders in schematics has not been standardized, but the samples shown are common.

Toggle

A *toggle switch* provides a firm and precise action via a lever (the *toggle*) that is usually tear-drop shaped and nickel plated, although plastic toggles are common in cheaper variants. Formerly used to control almost all electronic components (including early computers), the toggle has declined in popularity but is still used in applications such as automobile accessory kits, motorboat instrument panels, and industrial controls.

Three miniature DPDT toggle switches are shown in Figure 6-10. Two full-size, heavy-duty toggle switches are shown in Figure 6-11. A full-size, four-pole, double-throw heavy-duty toggle switch is shown in Figure 6-12. Toggle switches with more poles are extremely rare.

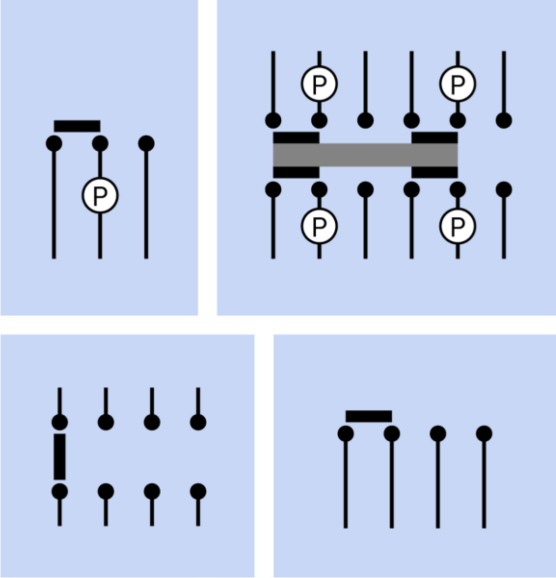

Figure 6-9. *Slide switch schematics. Each black rectangle represents a movable contact that connects two pairs of fixed contacts at a time. Detailed commentary on these variants will be found in the body of the text. Manufacturers may use variants of these symbols in their datasheets (for example, the gray rectangle indicating an insulating contact carrier, at top right, may be represented as a single line, or a black outline with a white center).*

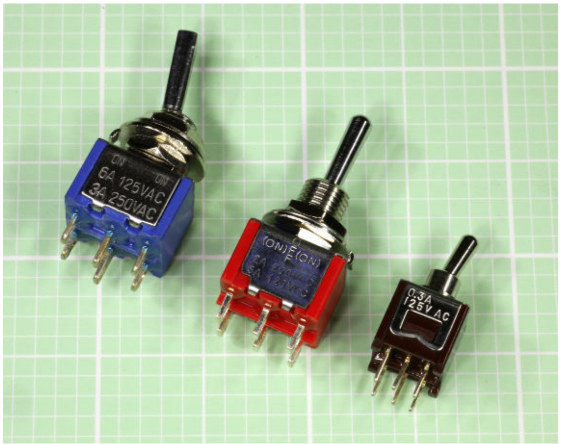

Figure 6-10. *Three miniature toggle switches with current ratings ranging from 0.3A to 6A at 125VAC. Each small square in the background grid measures 0.1" x 0.1".*

An automotive toggle switch is shown in Figure 6-13. Its plastic toggle is extended to minimize operating error.

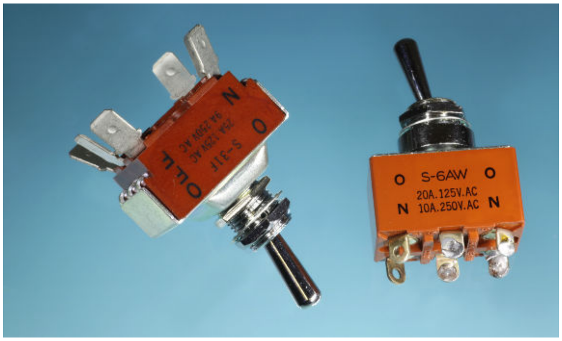

Figure 6-11. *Two full-size toggle switches capable of handling significant current. At left, the switch terminates in quick-connect terminals. At right, the switch has solder terminals (some of them containing residual traces of solder).*

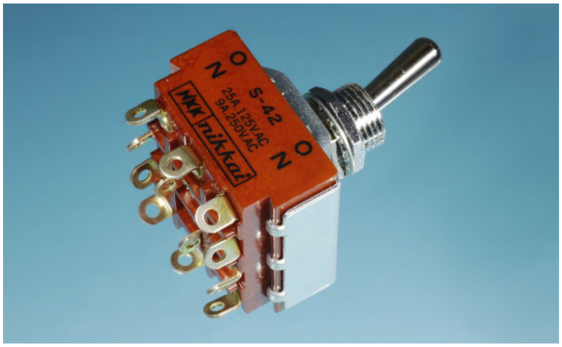

Figure 6-12. *A 4PDT full-size toggle switch with solder terminals, capable of switching 25A at 125VAC. Four-pole switches are relatively unusual.*

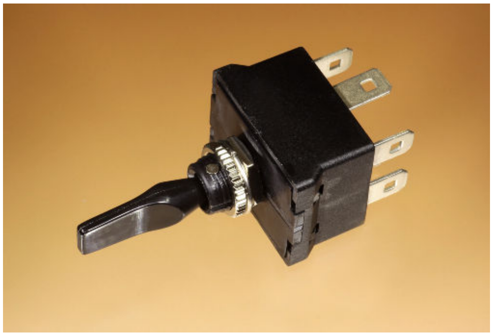

Figure 6-13. *A toggle switch intended for control of automotive accessories.*

High-end toggle switches are extremely durable and can be sealed from environmental contamination with a thin *boot* made from molded rubber or vinyl, which screws in place over the toggle, using the thread on the switch bushing. See Figure 6-14.

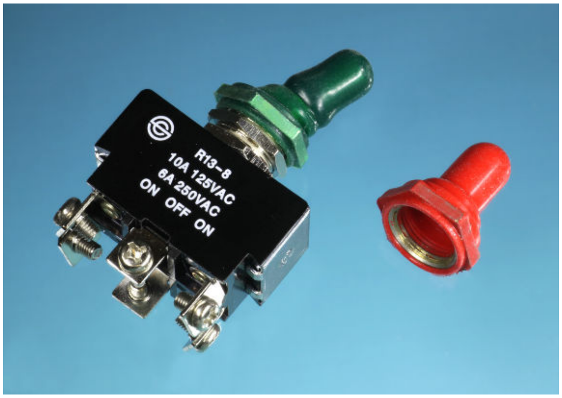

Figure 6-14. *A rubber or vinyl boot can be used to protect a toggle switch from contamination with dirt or water. Each boot contains a nut that screws onto the threads of a toggle switch, as shown at left.*

A *locking toggle switch* has a toggle that must be pulled out against the force of a retaining spring, before the toggle can be moved from one position to another. The toggle then snaps back into place, usually engaging in a small slot in the bushing of the switch.

DIP

A DIP switch is an array of very small, separate switches, designed for mounting directly on a circuit board, either in through-hole or surface-mount format. Through-hole DIP switches have two rows of pins with a 0.1″ pitch, the rows being spaced 0.3″ apart to fit a standard DIP (dual-inline package) socket or comparable configuration of holes in the board. Surface-mount DIP switches may have 0.1″ or 0.05″ pitch.

Most DIP arrays consist of SPST switches, each of which can close or open a connection between two pins on opposite sides of the switch body. The switch positions are usually labelled ON and OFF. Figure 6-15 shows a selection of DIP switches. Figure 6-16 shows the internal connections in a DIP switch.

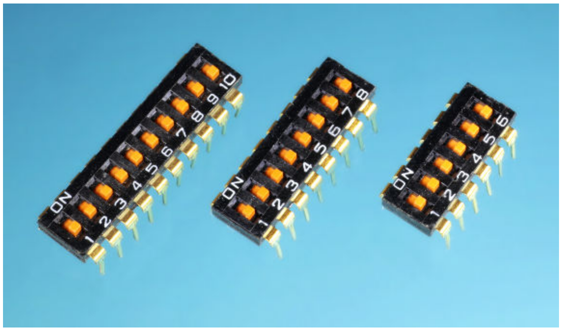

Figure 6-15. *As shown here, DIP switches are available with a variety of "positions," meaning the number of switches, not the number of switch states.*

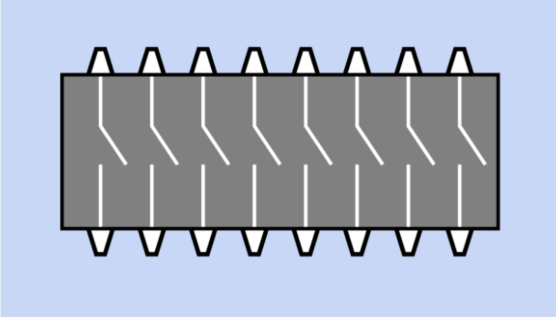

Figure 6-16. *The interior connections of a 16-pin DIP switch.*

The number of switches in a DIP array is usually referred to as its number of "positions." This should not be confused with the two positions of each physical switch lever. SPST DIP switches are made with 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, and 16 positions.

Early IBM-compatible desktop computers often required the user to set the position of an internal DIP switch when making routine upgrades such as installing an additional disk drive. While this feature is now obsolete, DIP switches are still used in scientific equipment where the user is expected to be sufficiently competent to open a cabinet and poke around inside it. Because of the 0.1″ spacing, a small screwdriver or the tip of a pen is more appropriate than a finger to flip individual levers to and fro.

DIP switches may also be used during prototype development, as they allow a convenient way to test a circuit in numerous different modes of operation.

Most DIP switches have wire terminals which are just long enough for insertion into a standard breaboard.

DIP switch package options include standard, low-profile, right-angle (standing at 90 degrees relative to the circuit board), and piano (with switch levers designed to be pressed, like tiny rocker switches, instead of being flipped to and fro).

Some SPDT, DPST, DPDT, 3PST, and 4PST variants exist, but are uncommon. Multiple external pins connect with the additional internal switch contacts, and a manufacturer's datasheet should be consulted to confirm the pattern of internal connections. A surface-mount, 0.1" pitch, DPST DIP switch is shown in Figure 6-17, with a plastic cover to protect the switches from contamination during wave soldering (at left), and with the cover peeled off (at right).

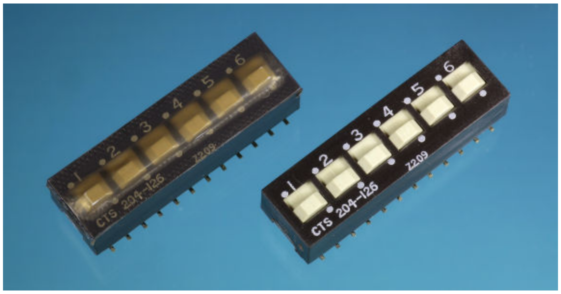

Figure 6-17. *A SPDT surface-mount double-throw DIP switch, sold with a plastic cover (shown at left) to protect it during wave soldering. The cover has been removed at right.*

SIP

A SIP switch is an array of small, separate switches, identical in concept to a DIP switch, but using only one row of pins instead of a double row. The applications for SIP switches are the same as DIP switches; the primary difference is simply that the SIP switch occupies a little less space, while being perhaps slightly less convenient to use.

One terminal of each switch usually shares a common bus. The internal connections in a typical 8-pin SIP array are shown in Figure 6-18. Pin spacing is 0.1", as in a typical DIP switch.

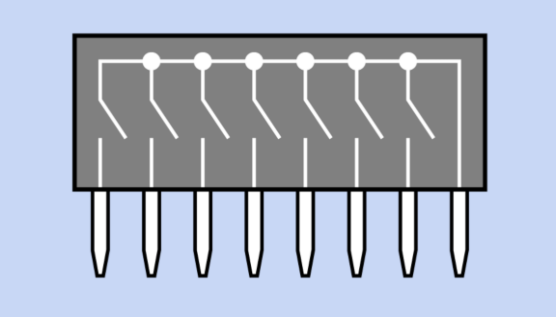

Figure 6-18. *The interior connections of an 8-pin SIP switch incorporating a common bus.*

Paddle

A paddle switch has a flat-sided tab-shaped plastic actuator, relatively large to allow a firm, error-free grip. Internally it is often comparable with a rocker switch, and is generally used with AC power. Some toggle-switch bodies are also sold with paddle-shaped actuators. A subminiature paddle switch is shown in Figure 6-19.

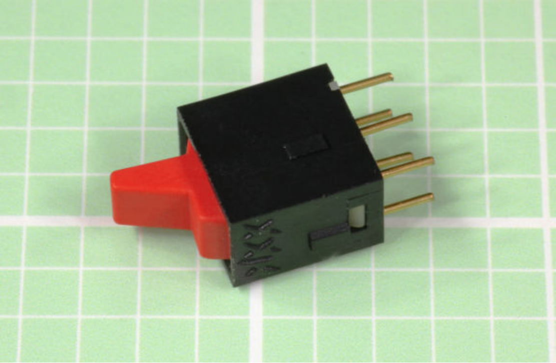

Figure 6-19. *A subminiature paddle switch. Full-size versions are often used as power switches.*

Vandal Resistant Switch

Typically fabricated from stainless steel, this is designed to withstand most types of abuse and is also weather-proof. The pushbuttons that allow pedestrians to trigger a traffic signal are a form of vandal-resistant switch.

Tactile Switch

This is considered to be a **pushbutton**, and is described in that entry. See "Tactile Switch" (page 34).

Mounting Options

A *panel mount* switch generally has a threaded bushing that is inserted from behind the front panel of a product, through a hole of appropriate size. It is supplied with a lockwasher and a nut (often, two nuts) that fit the thread on the switch bushing.

Front panel mount usually means that screws visible from the front of the panel are attached to a bracket on the switch behind the panel. The actuator of the switch is accessible through a cutout in the panel. This mounting style is mostly used for rocker switches and sometimes for slide switches.

Subpanel mount means that the switch is attached to a separate plate or chassis behind the control panel. The actuator of the switch is accessible through a cutout.

Snap-in mount requires a switch with flexible plastic or metal tabs each side, designed to push through a cutout in the panel, at which point the tabs spring out and retain the switch.

PC mount switches have pins that are soldered into a printed circuit board. They may have additional solderable lugs to provide mechanical support.

Surface mount switches are attached to a board in the same manner as other surface-mount components.

Termination

Switches (and pushbuttons) are available with a variety of terminals.

Solder lugs are small tabs, each usually perforated with a hole through which the end of a wire can be inserted prior to soldering.

PC terminals are pins that protrude from the bottom of the switch, suitable for insertion in a printed circuit board. This style is also known as a *through-hole*. The terminals may have a right-angle bend to allow the component to be mounted flat against the board, with the switch actuator sticking out at the side. This termination style is known as *right-angle PC*. Many manufacturers offer a choice of straight or bent pin terminals, but the component may be listed in a catalog under either of those options, with no indication that other options exist. Check manufacturer datasheets carefully.

Quick connect terminals are spade-shaped to accept push-on connectors, commonly used in automotive applications. Hybrid quick-connect terminals that can also function as solder lugs are sometimes offered as an option.

Screw terminals have screws premounted in flat terminals, for solderless attachment of wires.

Wire leads are flexible insulated wires, often with stripped and tinned ends, protruding at least an inch from the body of the component. This option is becoming uncommon.

Contact Plating Options

The internal electrical contacts of a switch are usually plated with silver or gold. Nickel, tin, and silver alloys are cheaper but less common. Other types are relatively rare.

Values

Switches designed for electronic devices vary widely in power capability, depending on their purpose. Rocker switches, paddle switches, and toggle switches are often used to turn power on and off, and are typically rated for 10A at 125VAC, although some toggle switches go as high as 30A. Snap-action or limit switches may be similarly rated, although miniature versions will have reduced capability. Slide switches cannot handle significant power, and are often rated around 0.5A (or less) at 30VDC. DIP and SIP switches have a typical maximum rating of 100mA at 50V and are not designed for frequent use. Generally they are used only when the power to the device is off.

How to Use it

Power Switches

When a simple SPST switch is used to turn DC power on and off, it conventionally switches the positive side of the power supply, also sometimes known as the *high side*. The primary reason for following this convention is that it is widely used; thus, following it will reduce confusion.

More importantly, an on-off switch that controls AC power must be used on the "live" side of the supply, not the "neutral" side. If you have any doubts about these concepts (which go beyond the scope of this book), consult a reference guide on this subject. Using a DPST component to switch both sides of an AC supply may be a worthwhile additional precaution in some applications. The ground wire of an AC supply should never be switched, because the device should always be grounded when it is plugged into an electric outlet.

Limit Switches

An application for two limit switches with a DC motor and two rectifier **diodes** is shown in Figure 6-20. This diagram assumes that the motor turns clockwise when its lower terminal is positive, and counter-clockwise when its upper terminal is positive. Only two terminals are used (and shown) in each limit switch; they are chosen to be normally-closed. Other terminals inside a switch may exist, may be normally-open, and can be ignored.

The motor is driven through a dual-coil, DPDT latching **relay**, which will remain in either position indefinitely without drawing power. When the upper coil of the relay receives a pulse from a pushbutton or some other source, the relay flips to its upper position, which conducts positive current through the lower limit switch, to the lower terminal of the motor. The motor turns clockwise until the arm attached to its shaft hits the lower limit switch and opens it. Positive current is blocked by the lower diode, so the motor stops.

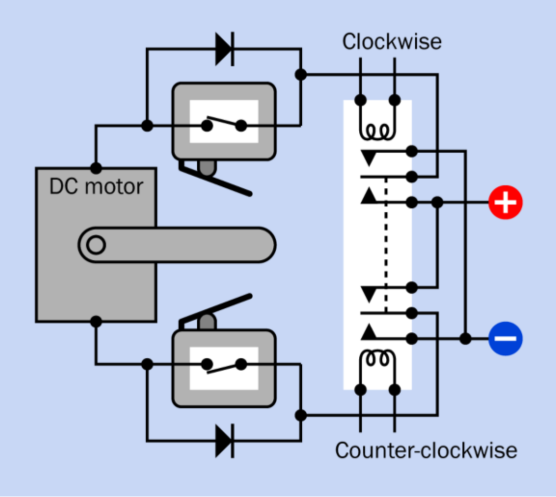

Figure 6-20. *In this schematic, normally-closed limit switches are opened by pressure from an arm attached to a motor, thus switching off its power at each end of its permitted travel and preventing overload and burnout. A two-coil latching relay activates the motor. Rectifier diodes allow power to reach the motor to reverse its rotation when a limit switch is open.*

When the lower coil of the relay is activated, the relay flips to its lower position. Positive current can now reach the upper side of the motor through the upper limit switch. The motor runs counter-clockwise until its arm opens the upper limit switch, at which point the motor stops again. This simple system allows a DC motor to be run in either direction by a button-press of any duration, without risk of burnout when the motor reaches the end of its travel. It has been used for applications such as raising and lowering powered windows in an automobile.

A DPDT pushbutton could be substituted for the latching relay if manual control, only, is acceptable. However, in this scenario, sustained pressure on the pushbutton would be necessary to move the motor arm all the way to the opposite end of its travel. A DPDT switch might be more appropriate than a pushbutton.

Logic Circuits

Logic circuits that depend purely on switches can be constructed (for example, to add binary numbers) but are rare and have no practical applications. The most familiar and simplest example of manually switched logic is a pair of SPDT switches in house wiring, one positioned at the top of a flight of stairs and the other at the bottom, as shown in Figure 6-21. Either switch will turn the light on if it is currently off, or off if it is currently on. To extend this circuit by incorporating a third switch that has the same function as the other two, a DPDT switch must be inserted. See Figure 6-22.

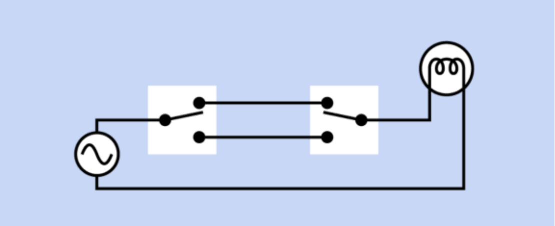

Figure 6-21. *SPDT switches are commonly used in house wiring so that either of them will turn a shared light on if it is off, or off it is on.*

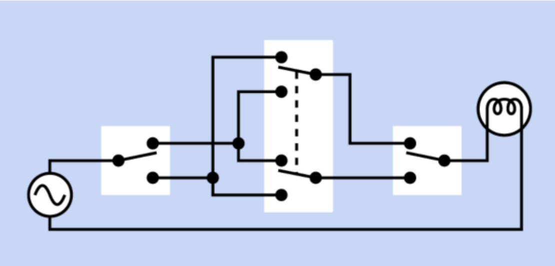

Figure 6-22. *A DPDT switch must be inserted if three switches must have identical function to control the on-off state of a single light bulb.*

Alternatives

As **microcontrollers** have become cheaper and more ubiquitous, they have taken over many functions in electronic products that used to be served by switches. A menuing system driven by a microcontroller can use one **rotational encoder** with a SPST pushbutton built into it to select and adjust numerous parameters in a device such as a car stereo, where functions were once selected and adjusted by individual switches and potentiometers. The rotational-encoder option takes up less space, is cheaper to build (assuming a microcontroller is going to be used in the device for other purposes anyway), and can be more reliable, as it reduces the number of electromechanical parts. Whether it is easier to use is a matter of taste. Cost and ergonomics may be the primary factors to consider when choosing where and how to use switches.

What Can Go Wrong

Arcing

The contacts inside a switch will be rapidly eroded if *arcing* (pronounced "arking") occurs. An electric arc is a spark that tends to form when a switch is opened while conducting a high current or high voltage (typically 10A or more and 100V or more). The most common cause is an inductive load that generates back-EMF when it is switched on and forward-EMF when it is switched off. The surge can be many times the amperage that the load draws during continuous operation. In DC circuits, arcing can be reduced by using a rectifier **diode** in parallel with the load (with its polarity blocking normal current flow). This is often referred to as a *flyback diode* or *freewheeling diode*. In AC circuits, where a diode cannot be used in this way, a *snubber* (a simple combination of capacitor and resistor) may be placed around the load. A snubber can also be used around the switch itself, in DC circuits. See "Snubber" (page 108).

When switching an inductive load, it is generally prudent to use switches rated for twice as much current as the circuit will normally draw.

Dry Joints

Switches that control significant current will have substantial terminals, and these terminals will be attached to heavy-gauge wire. When using solder to make this type of connection, the combined heat capacity of the wire and the terminal will sink a lot more heat than a small component on a circuit board. At least a 30W soldering iron should be used. Lower-wattage irons may be incapable of melting the solder completely (even though they seem to), and a "dry joint" will result, which can have a relatively high electrical resistance and will be mechanically weak, liable to break later. Any good solder joint should withstand flexing of the wire attached to it.

Short Circuits

Because many switches are still wired in with solder tabs, screw terminals, or quick-connect terminals, wires that become accidentally detached can be a significant hazard. Heat-shrink tubing should be applied to enclose wires and terminals at the rear of a power switch, as an additional precaution. Power switches should always be used in conjunction with appropriate **fuses**.

Contact Contamination

Sealed switches should be used in any environment where dirt or water may be present. Slide switches are especially vulnerable to contamination, and are difficult to seal. Switches used in audio components will create "scratchy" sounds if their contacts deteriorate.

Wrong Terminal Type

Because switches are available with a wide variety of terminal types, it's easy to order the wrong type. Switches may be supplied with pins for through-hole insertion in circuit boards; screw terminals; quick-disconnect terminals; or solder lugs. Variants may also be available for surface mount. If a project requires, for example, the insertion of pins in a printed circuit board, and a switch is supplied with solder lugs, it will be unusable.

Part numbers generally include codes to identify each terminal variant, and should be studied carefully.

Contact Bounce

Also known as *switch bounce*. When two contacts snap together, extremely rapid, microscopic vibrations occur that cause brief interruptions before the contacts settle. While this phenomenon is not perceptible to human senses, it can be perceived as a series of multiple pulses by a **logic chip**. For this reason, various strategies are used to *debounce* a switch that drives a logic input. This issue is explored in detail in the entry on logic chips in Volume 2 of the encyclopedia.

Mechanical Wear

Any toggle or rocker switch contains a mechanical pivot, which tend to deteriorate in harsh environments. Friction is also an issue inside these switches, as the design often entails the rounded tip of a lever rubbing to and fro across the center of a movable contact.

The spring inside a snap switch or limit switch may fail as a result of metal fatigue, although this is rare. A slide switch is far less durable, as its contacts rub across each other every time the switch changes position.

In any application that entails frequent switching, or where switch failure is a critical issue, the most sensible practice is to avoid using cheap switches.

Mounting Problems

In a panel-mount switch that is secured by turning a nut, the nut may loosen with use, allowing the component to fall inside its enclosure. Conversely, overtightening the nut may strip the threads on the switch body, especially in cheaper components where the threads are molded into plastic. Consider applying a drop of Loc-Tite or similar adhesive after moderately tightening the nut. Note that nut sizes vary widely, and finding a replacement may be time-consuming.

Cryptic Schematics

In some circuit schematics, the poles of a multipole switch may be visually separated from each other, even at opposite sides of the page, for convenience in drawing the schematic. Dotted lines usually, but not always, link the poles. In the absence of dotted lines, switch segments are often coded to indicate their commonality. For example, SW1(a) and SW1(b) are almost certainly different parts of the same switch, with linked poles.

rotary switch

7

Not to be confused with **rotational encoder**, which has its own entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **switch** (See Chapter 6)
- **rotational encoder** (See Chapter 8)

What It Does

A rotary switch makes an electrical connection between a *rotor*, mounted on a shaft that is turned by a knob, and one of two or more stationary contacts. Traditionally, it was the component of choice to select wavebands on a radio receiver, broadcast channels on a television or inputs on a stereo preamplifier. Since the 1990s, it has been substantially superceded by the **rotational encoder**. However it still has applications in military equipment, field equipment, industrial control systems, and other applications requiring a rugged component that will withstand heavy use and a possibly harsh environment. Also, while the output from a rotational encoder must be decoded and interpreted by a device such as a microcontroller, a rotary switch is an entirely passive component that does not require any additional electronics for its functionality.

Two typical schematic symbols for a rotary switch are shown in Figure 7-1. They are functionally identical. A simplified rendering of the interior of a traditional-style rotary switch is shown in Figure 7-2. A separate contact (not shown) connects with the rotor, which connects with each of the stationary contacts in turn. The colors were chosen to differentiate the parts more clearly, and do not correspond with colors in an actual switch.

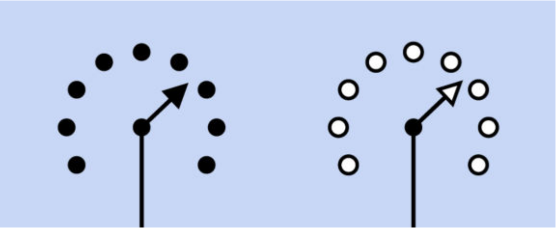

Figure 7-1. *Typical schematic symbols for a rotary switch. The two symbols are functionally identical. The number of contacts will vary depending on the switch.*

A selection of rotary switches is shown in Figure 7-3. At top-left is an *open frame* switch, providing no protection to its contacts from contaminants. This type of component is now rare. At top-right is a twelve-position, single-pole switch rated 2.5A at 125VAC. At front-left is a four-position, single-pole switch rated 0.3A at 16VDC or 100VAC. At front right is a two-position, two-pole switch with the same rating as the one beside it. All the sealed switches allow a choice of panel mounting or through-hole printed circuit board mounting.

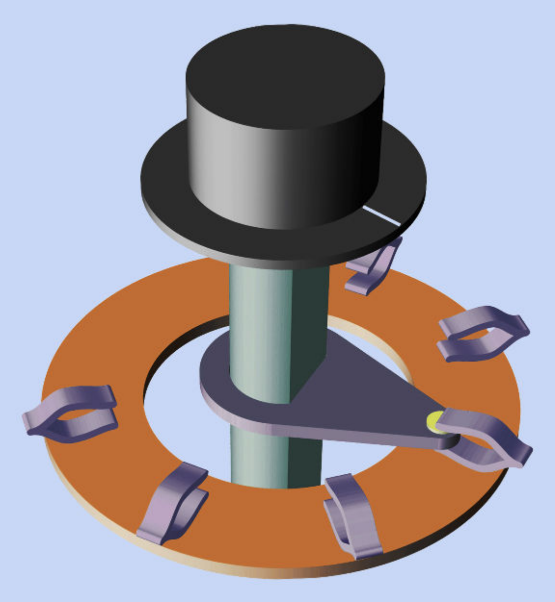

Figure 7-2. *A simplified rendering of interior parts in a basic SP6T rotary switch. Arbitrary colors have been added for clarity.*

How It Works

A switch may have multiple poles, each connecting with its own rotor. The rotors are likely to be on separate *decks* of the switch, but two, three, or four rotors, pointing in different directions, may be combined on a single deck if the switch has only a small number of positions.

Rotary switches are usually made with a maximum of twelve positions, but include provision for limiting the number of positions with a *stop*. This is typically a pin, which may be attached to a washer that fits around the bushing of the switch. The pin is inserted into a choice of holes to prevent the switch from turning past that point. For example, an eight-position rotary switch can be configured so that it has only seven (or as few as two) available positions.

A specification for a rotary switch usually includes the angle through which the switch turns between one position and the next. A twelve-position switch usually has a 30-degree turn angle.

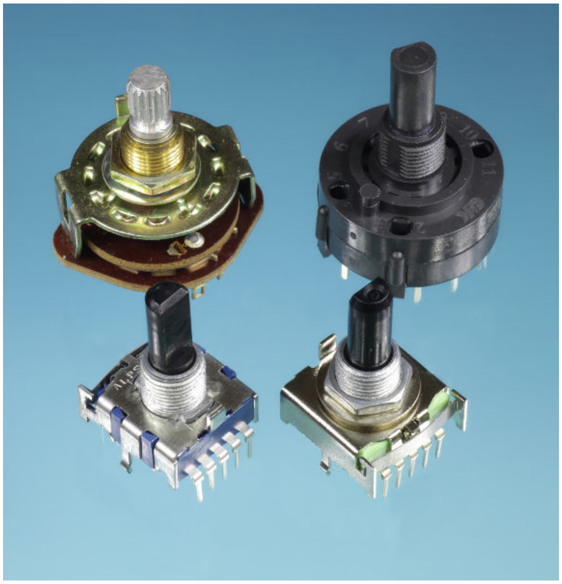

Figure 7-3. *A selection of rotary switches. See text for details.*

Variants

Conventional

The traditional style of rotary switch is designed to be panel-mounted, with a body that ranges from 1″ to 1.5″ in diameter. If there is more than one deck, they are spaced from each other by about 0.5″. The switch makes an audible and tactile “click” as it is turned from one position to the next.

A rugged sealed five-deck rotary switch is shown in Figure 7-4. It has five poles (one per deck), and a maximum of 12 positions. The contacts are rated 0.5A at 28VDC. This type of heavy-duty component is becoming relatively rare.

If the rotor in a switch establishes a connection with the next contact a moment before breaking the connection with the previous contact, this is known as a *shorting* switch, which may also be described as a *make-before-break* switch. In a *nonshorting* or *break-before-make* switch, a tiny

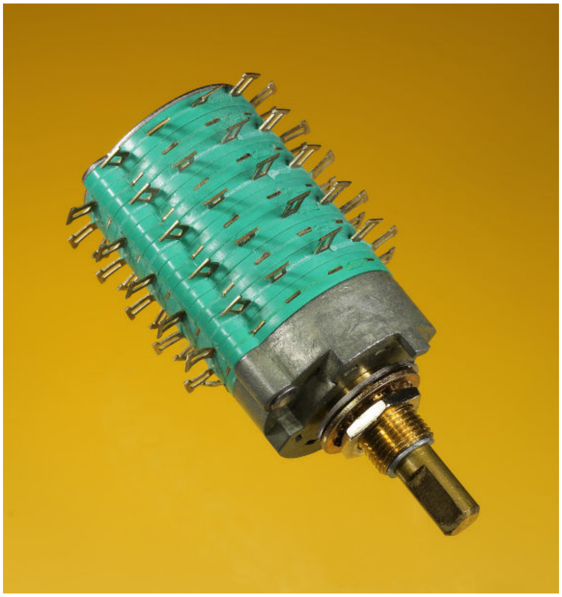

Figure 7-4. *A five-pole, twelve-position rotary switch.*

interval separates one connection from the next. This can be of significant importance, depending on the components that are connected with the switch.

The shaft may be round, splined, or D-shaped in section. A knob is seldom supplied with a switch and must be chosen to match the shaft. Some shaft dimensions are metric, while others are measured in inches, with 1/4" diameter being the oldest standard. Some switches with a splined shaft are supplied with an adapter for a knob of D-shaped internal section; the adapter can be slipped onto the shaft in any of 12 or more positions, to minimize the inconvenience of positioning the body of the switch itself so that the knob is correctly oriented in relation to positions printed on the face of the panel.

Miniature rotary switches may be as small as 0.5" diameter, and usually terminate in pins for through-hole mounting on a PC board. Miniature switches usually have lower current ratings than full-size switches.

Rotary switches must be securely anchored to resist the high turning forces that can be inflicted upon them by users. In a panel-mount design, a nut is tightened around a thread on the bushing of the switch. Through-hole versions can be secured to the PC board with the shaft protruding loosely through a cutout in the panel. To minimize mechanical stress on the circuit board, the detents in a PC-board switch are usually weaker than in a full-size switch, and the knob is usually smaller, allowing less leverage.

Rotary DIP

A conventional *DIP switch* is a linear array of miniature SPST switches designed to fit a standard DIP (dual-inline package) layout of holes in a circuit board. It is described in the **switch** entry of this encyclopedia. A *rotary DIP* switch (also known as an *encoded output rotary switch* or a *coded rotary switch*) does not conform with a DIP layout, despite its name. It is approximately 0.4" square and usually has five pins, one of which can be considered the input or common pin while the other four can function as outputs. The pins are spaced at 0.1" pitch from one another. Pin function and layout are not standardized.

A dial on top of the switch has either 10 positions (numbered 0 through 9) or 16 positions (0 through 9 followed by letters A through F). One switch of each type is shown in Figure 7-5.

Each position of the dial closes pairs of contacts inside the component to create a unique binary-coded decimal pattern (in a 10-position switch) or binary-coded hexadecimal pattern (in the 16-position switch) on the four output pins. The pin states are shown in Figure 7-6. A rotary DIP switch is a relatively flimsy device, and is not designed for frequent or heavy use. It is more likely to be a "set it and forget it" device whose state is established when it is installed in a circuit board.

Because each position of the switch is identified with a unique binary pattern, this is an example of *absolute encoding*. By contrast, a typical **rotational encoder** uses *relative encoding*, as it merely generates a series of undifferentiated pulses when the shaft is turned.

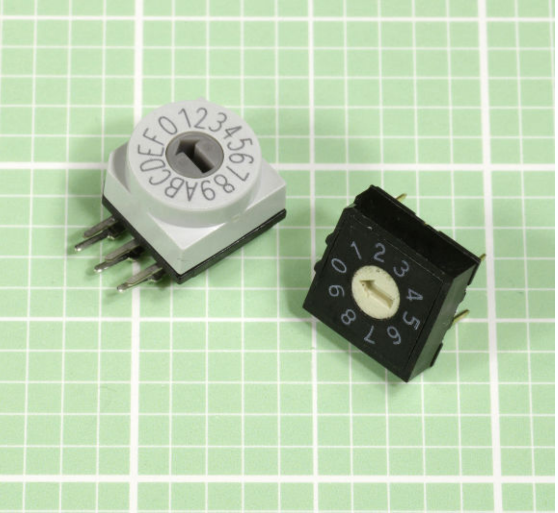

Figure 7-5. *A rotary DIP switch, also known as an encoded output rotary switch, may be used as a substitute for a DIP switch in some applications.*

A *real-coded* rotary DIP makes a connection between input and output pins wherever a binary 1 would exist. In the *complement-coded* version, the output is inverted. The switch is primarily intended for use with a **microcontroller**, enabling only four binary input pins on the microcontroller to sense up to sixteen different switch positions.

A six-pin rotary DIP variant is available from some manufacturers, with two rows of three pins, the two center pins in each row being tied together internally, and serving as the pole of the switch.

Rotary DIPs are available with a screw slot, small knurled knob, or larger knob. The screw-slot version minimizes the height of the component, which can be relevant where circuit boards will be stacked close together. A *right-angle PC* variant stands at 90 degrees to the circuit board, with pins occupying a narrower footprint. The switch on the left in Figure 7-5 is of this type.

While most rotary DIPs are through-hole components, surface-mount versions are available.

Most rotary DIPs are sealed to protect their internal components during wave-soldering of circuit boards.

Switch Position	Pin 1	Pin 2	Pin 3	Pin 4
0	⚫	⚫	⚫	⚫
1	🔴	⚫	⚫	⚫
2	⚫	🔴	⚫	⚫
3	🔴	🔴	⚫	⚫
4	⚫	⚫	🔴	⚫
5	🔴	⚫	🔴	⚫
6	⚫	🔴	🔴	⚫
7	🔴	🔴	🔴	⚫
8	⚫	⚫	⚫	🔴
9	🔴	⚫	⚫	🔴
A	⚫	🔴	⚫	🔴
B	🔴	🔴	⚫	🔴
C	⚫	⚫	🔴	🔴
D	🔴	⚫	🔴	🔴
E	⚫	🔴	🔴	🔴
F	🔴	🔴	🔴	🔴

Figure 7-6. *Positive and negative states of the four output pins of a real-coded 16-position rotary DIP switch, assuming that the common pin of the switch is connected with a positive supply voltage. A ten-position rotary DIP switch would use only the states from 0 through 9. In a complement-coded switch, the positive and negative states would be reversed.*

Gray Code

A Gray code (named after its originator, Frank Gray) is a system of absolute encoding of a switch output, using a series of nonsequential binary numbers that are chosen in such a way that each number differs by only one digit from the preceding number. Such a series is useful because it eliminates the risk that when a switch turns, some bits in the output will change before others, creating the risk of erroneous interpretation. A minority of rotary switches or rotational encoders are available with Gray-coded outputs. Typically, a microcontroller must use a lookup table to convert each binary output to an angular switch position.

PC Board Rotary Switch

Miniature switches with a conventional, non-encoded output are available for printed-circuit board mounting, sometimes requiring a screwdriver or hex wrench to select a position. A single-pole eight-position switch of this type is shown in Figure 7-7. Its contacts are rated to carry 0.5A at 30VDC, but it is not designed to switch this current actively.

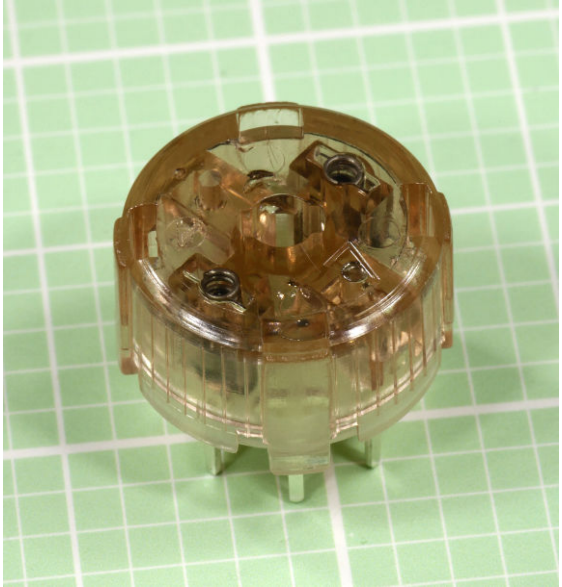

Figure 7-7. *This miniature switch is designed for insertion on a printed circuit board. It can be used to make a setting before a device is shipped to the end user.*

Mechanical Encoder

A *mechanical encoder* functions similarly to a rotary DIP switch but is intended for much heavier use. It outputs a binary-coded-decimal value corresponding with its shaft position, is typically the size of a miniature rotary switch, and is designed for panel mounting. The Grayhill Series 51 allows 12 positions, each generating a code among four terminals. The Bourns EAW provides 128 positions, each generating a code among 8 terminals.

Pushwheel and Thumbwheel

A *pushwheel switch* is a simple electromechanical device that enables an operator to provide a code number as input to data processing equipment, often in industrial process control. The decimal version contains a wheel on which numbers are printed, usually in white on black, from 0 through 9, visible one at a time through a window in the face of the switch. A button above the wheel, marked with a minus sign, rotates it to the next lower number, while a button below the wheel, marked with a plus sign, rotates it to the next higher number. A connector at the rear of the unit includes a common (input) pin and four output pins with values 1, 2, 4, and 8. An additional set of pins with values 1, 2, 3, and 4 is often provided. The states of the output pins sum to the value that is currently being displayed by the wheel. Often two, three, or four pushwheels (each with an independent set of connector pins) are combined in one unit, although individual pushwheels are available and can be stacked in a row.

A *thumbwheel switch* operates like a pushwheel switch, except that it uses a thumbwheel instead of two buttons. Miniaturized thumbwheel switches are available for through-hole mounting on PC boards.

Hexadecimal versions are also available, with numbers from 0 through 9 followed by letters A through F, although they are less common than decimal versions.

Keylock

A *keylock switch* is generally a two-position rotary switch that can be turned only after insertion of a key in a lock attached to the top of the shaft. This type of switch almost always has an OFF-(ON) configuration and is used to control power.

Keylock switches are found in locations such as elevators, for fire-department access; in cash registers; or on data-processing equipment where switching power on or off is reserved for a system administrator.

Values

A full-size rotary switch may be rated from 0.5A at 30VDC to 5A at 125VAC, depending on its purpose. A very few switches are rated 30A at 125VAC; these are high-quality, durable, expensive items.

A typical rotary DIP switch is rated 30mA at 30VDC and has a carrying current rating (continuous current when no switching occurs) of no more than 100mA at 50VDC.

How to Use it

In addition to its traditional purpose as a mode or option selector, a rotary switch provides a user-friendly way to input data values. Three ten-position switches, for instance, can allow user input of a decimal number ranging from 000 to 999.

When used with a **microcontroller**, a rotary switch can have a *resistor ladder* mounted around its contacts, like a multi-point *voltage divider*, so that each position of the rotor provides a unique potential ranging between the positive supply voltage and negative ground. This concept is illustrated in Figure 7-8, where all the resistors have the same value. The voltage can be used as an input to the microcontroller, so long as the microcontroller shares a common ground with the switch. An analog-digital converter inside the microcontroller translates the voltage into a digital value. The advantage of this scheme is that it allows very rapid control by the user, while requiring only one pin on the microcontroller to sense as many as twelve input states.

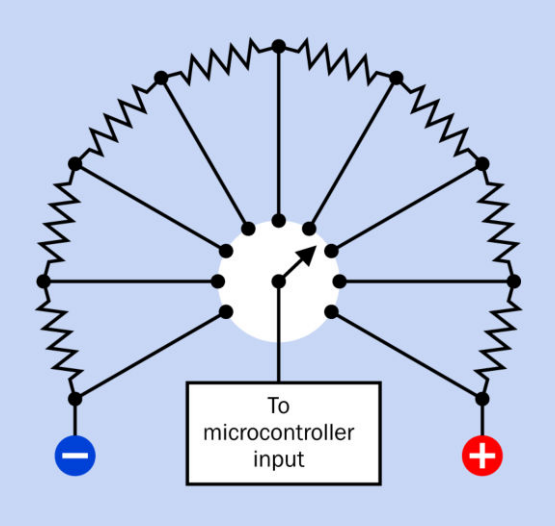

Figure 7-8. *A resistor ladder can be formed around the contacts of a rotary switch, with the pole of the switch connected to a microcontroller that has an analog-digital converter built in. The microcontroller converts the voltage input to an internal digital value. Thus, one pin can sense as many as twelve input states.*

For a ladder consisting of 8 resistors, as shown, each resistor could have a value of 250Ω. (The specifications for a particular microcontroller might require other values.) To avoid ambiguous inputs, a nonshorting rotary switch should be used in this scheme. A *pullup resistor* of perhaps 10K should be added to the microcontroller input, so that there is no risk of it “floating” when the switch rotor is moving from one contact to the next. The code that controls the microcontroller can also include a blanking interval during which the microcontroller is instructed to ignore the switch.

Because the rotary switch is an electromechanical device, it has typical vulnerabilities to dirt and moisture, in addition to being bulkier, heavier, and more expensive than a **rotational encoder**. Rotary switches have also been partially replaced by pushbuttons wired to a microcontroller. This option is found on devices ranging from digital alarm clocks to cellular phones. In addition to being cheaper, the pushbutton alternative is preferable where space on a control panel, and behind it, is limited.

What Can Go Wrong

Vulnerable Contacts

Most modern rotary switches are sealed, but some are not. Any switch with exposed contacts will be especially vulnerable to dirt and moisture, leading to unreliable connections. This was an issue in old-fashioned TV sets, where periodic contact cleaning of the channel selector switch was needed.

Exposed contacts are also more vulnerable to side-effects from temperature cycling (when a device warms up and then cools down).

Contact Overload

The contacts on a cheap rotary switch are especially vulnerable to arcing, as the user may turn the switch slowly, causing gradual engagement and disengagement of contacts instead of the snap-action that is characteristic of a well-made toggle switch. If a rotary switch may control significant currents or current surges, it must be appropriately rated, regardless of the extra expense. For more information on arc suppression in switches, see "Arcing" (page 47).

Misalignment

Most knobs for rotary switches consist of a pointer, or have a white line engraved to provide clear visual indication of the position of the switch. If this does not align precisely with indications printed on the panel, confusion will result. For hand-built equipment, the switch can be installed first, after which the control-panel indications can be glued or riveted in place on a separate piece of laminated card, plastic, or metal for precise alignment. If the switch is not secured tightly, its body may turn slightly under repeated stress, leading to erroneous interpretation of the knob position.

Misidentified Shorting Switch

If a shorting switch is used where a nonshorting switch was intended, the results can be disconcerting or even destructive, as one terminal will be briefly connected with the adjacent terminal while the switch is being turned. Multiple functions of a circuit may be activated simultaneously, and in a worst case scenario, adjacent terminals may be connected to opposite sides of the same power supply.

User Abuse

The turning force that must be applied to a full-size conventional rotary switch is significantly greater than the force that is applied to most other types of panel-mounted switches. This encourages aggressive treatment, and the turning motion is especially likely to loosen a nut holding the switch in place. The lighter action characteristic of miniature rotary switches does not necessarily solve this problem, as users who are accustomed to older-style switches may still apply the same force anyway.

Rotary switches should be mounted in expectation of rough use. It is prudent to use Loc-Tite or a similar compound to prevent nuts from loosening, and a switch should not be mounted in a thin or flimsy panel. When using a miniature rotary switch that has through-hole mounting in a circuit board, the board must be sufficiently robust and properly secured.

Wrong Shaft, Wrong Knobs, Nuts That Get Lost, Too Big to Fit

These problems are identical to those that can be encountered with a **potentiometer**, which are discussed in that entry in this encyclopedia.

rotational encoder

8

The term *rotational encoder* used to be reserved for high-quality components, often using optical methods to measure rotation with precision (more than 100 intervals in 360 degrees). Cheaper, simpler, electromechanical devices were properly referred to as *control shaft encoders*. However, the term *rotational encoder* is now applied to almost any device capable of converting rotational position to a digital output via opening and closing internal mechanical contacts; this is the sense in which the term is used here. It is sometimes distinguished from other types of encoder with the term *mechanical rotary encoder*. Magnetic and optical rotary encoders do not contain mechanical switches, are classified as *sensors* by this encyclopedia, and will appear in Volume 3. They are found in a device such as an *optical mouse*.

OTHER RELATED COMPONENTS

- **rotary switch** (See Chapter 7)

What It Does

A rotational encoder has a knob that a user can turn to display a series of prompts on an LCD screen, or to adjust the input or output on a product such as a stereo receiver. The component is almost always connected to inputs on a **microcontroller** and is usually fitted with *detents* that provide tactile feedback suggesting many closely spaced positions. The encoder often allows the user to make a selection by pushing the knob in, which closes an internal momentary switch. Thus, this type of encoder functions as a pushbutton as well as a switch.

A rotational encoder is an *incremental* or *relative* device, meaning that it merely creates and breaks internal switch connections when rotation occurs, without providing a unique code to identify each absolute rotational position. An *absolute* encoder is discussed in the **rotary switch** entry of this encyclopedia.

No schematic symbol exists to represent a rotational encoder.

How It Works

An encoder contains two pairs of contacts, which open and close out of phase with each other when the shaft rotates. In a clockwise direction, the A pair of contacts may be activated momentarily before the B pair; in a counter-clockwise direction, the B pair may be activated before the A pair. (Some encoders reverse this phase difference.) Thus if one contact from each pair is connected with two inputs of an appropriately programmed microcontroller, and if the other contact of each pair is connected with negative ground, the microcontroller can deduce which way the knob is turning by sensing which pair of contacts closes first. The microcontroller can then count the number of pulses from the contacts and interpret this to adjust an output or update a display.

A simplified schematic is shown in Figure 8-1. The two buttons inside the dashed line represent the two pairs of contacts inside the encoder, while the chip is a microcontroller. The knob and shaft that activate the internal switches are not shown. The schematic assumes that when a contact closes, it pulls the chip input to a low state. A pullup resistor is added to each input of the chip to prevent the pins from “floating” when either pair of contacts is open.

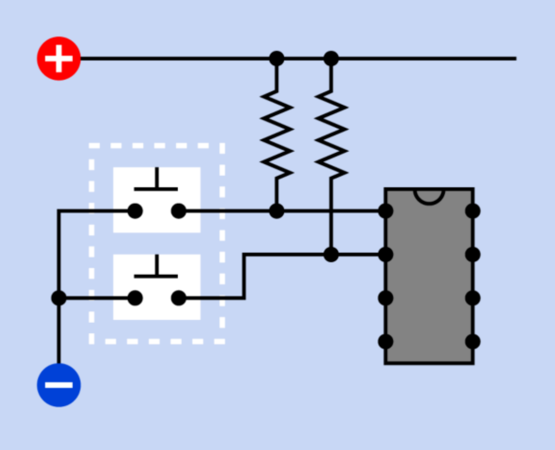

Figure 8-1. *Simplified schematic showing the typical setup for a rotational encoder. The pushbuttons inside the dashed line represent the contacts inside the encoder. The chip is a microcontroller.*

Figure 8-2 gives a conceptual view of the outputs of an encoder that is turned clockwise (top) and then counter-clockwise (bottom). Some encoders may reverse this phase sequence. Red and black colors have been assigned to the pin states on the assumption that the terminals that are common to both pairs of contacts are connected with negative ground. Thus a “high” pulse in the graphical representation actually indicates that the encoder is grounding its output.

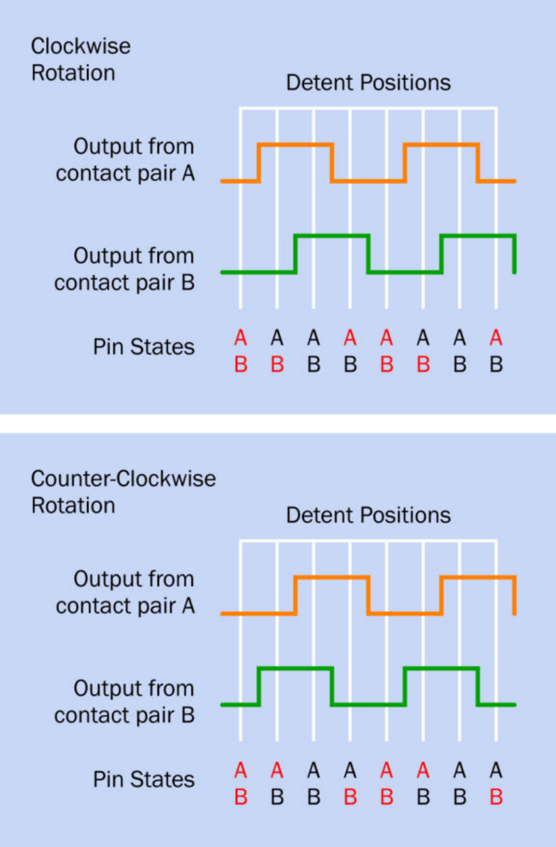
Clockwise Rotation
Detent Positions
Output from contact pair A
Output from contact pair B
Pin States
A A A A A A A A
B B B B B B B B
Counter-Clockwise Rotation
Detent Positions
Output from contact pair A
Output from contact pair B
Pin States
A A A A A A A A
B B B B B B B B

Figure 8-2. *Hypothetical outputs from a rotational encoder, assuming that the common terminals of the contact pairs are connected to negative ground. A high pulse in the graphical representation therefore indicates that the contact pair is grounded. The number of detents relative to the number of pulses per rotation varies from one type of encoder to another.*

Microcontrollers have become so ubiquitous, and rotational encoders are so cheap, they have displaced rotary switches in many applications where a low current is being switched. The combination of a rotational encoder and a microcontroller is very versatile, allowing display and control of an almost unlimited number of menus and options.

Variants

There are two types of rotational encoders containing mechanical contacts: absolute and relative. An absolute encoder generates a code corresponding with each specific rotational position. The code is usually a binary output among four or more pins. It is discussed under *mechanical encoder* in the **rotary switch** section of this encyclopedia. The variants listed here are all relative encoders.

Pulses and Detents

Rotational encoders from different manufacturers may have as few as 4 or as many as 24 pulses per rotation (PPR), with 12 to 36 detents (or no detents at all, in a few models.) The relationship between pulses and detents shown in Figure 8-2 is typical but is far from being universal. The number of detents may be equal to, greater than, or less than the number of pulses per rotation.

Format

Rotational encoders are generally panel-mounted or through-hole devices. In the latter category, most are horizontally mounted, with a minority being at 90 degrees to the board.

Output

In an encoder containing two switches, four switch-state combinations are possible: OFF-OFF, ON-OFF, OFF-ON, and ON-ON. This is known as a *quadrature* output. All of the rotational encoders discussed here conform with that system.

Rotational Resistance

Rotational encoders vary widely in the resistance that they offer when the user turns the knob. This is largely a function of the detents, if they are included. Still, all rotational encoders generally offer less rotational resistance than a rotary switch, and do not have the kind of heavy-duty knobs that are typically used with rotary switches. Since an encoder creates only a stream of pulses without any absolute positional information, a knob with any kind of pointer on it is inappropriate.

Values

Virtually all rotational encoders are designed to work with a low-voltage supply, 12VDC or less. All of them are intended for low currents, reflecting their purpose to drive microcontroller inputs. Some sample rotational encoders are pictured in Figure 8-3. At rear: nine pulses per rotation (PPR), 36 detents, 10mA at 10VDC. Far left: 20PPR, 20 detents, with switch. Far right: 24PPR, no detents, 1mA at 5VDC. Center (blue): 16PPR, no detents, 1mA at 5VDC. Front: 12PPR, 24 detents, 1mA at 10VDC, requires Allen wrench or similar hexagonal shaft to engage with the rotor.

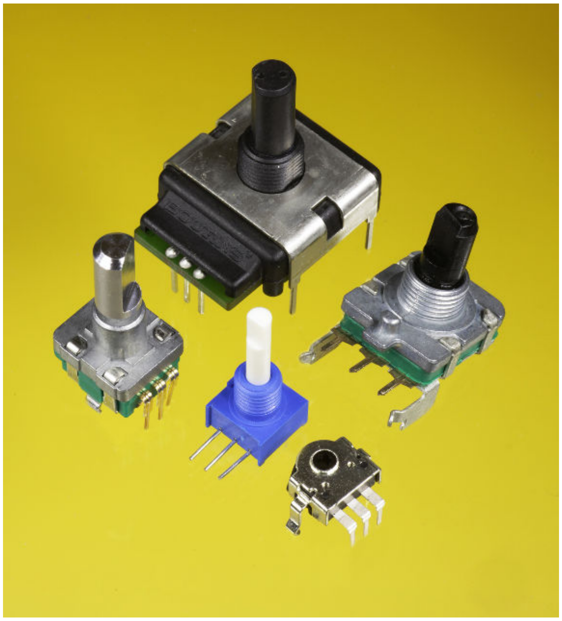

Figure 8-3. *Rotational encoders with a variety of specifications. See text for details.*

Contact Bounce

Any mechanical switch will suffer some degree of *contact bounce* when its contacts close. Datasheets for rotational encoders may include a specification for bounce duration ranging from around 2ms to 5ms, which is sometimes known as the *settling time*. Naturally, a lower value is preferred. The microcontroller that interprets the positional information from the encoder can include a debouncing routine that simply disregards any signals during the bounce period following switch closure.

Sliding Noise

Sliding noise is the opposite of contact bounce. When two contacts have made a connection and then rub across each other (as occurs inside a rotational encoder while the knob is being turned), the connection may suffer momentary lapses. Datasheets for rotational encoders generally do not supply ratings for this.

How to Use it

As noted above, a rotational encoder can only be used in conjunction with a microcontroller or similar device that is capable of interpreting the phase difference between the pairs of contacts, and is capable of counting the number of opening/closing events while the knob is being turned. (Some dedicated chips are designed for this specific purpose.)

It can be adapted to be driven by a **stepper motor**, to provide feedback regarding the rotation of the motor shaft, and its output can also be interpreted to calculate angular acceleration.

Programming the microcontroller is the most significant obstacle. Generally the program should follow a sequence suggested by this pseudocode:

Check:

- If the encoder contains a pushbutton switch, check it. If the pushbutton is being pressed, go to an appropriate subroutine.
- The status of contacts A.
- The status of contacts B.

Compare their status with previously saved states for A and B. If the status has not changed, repeat from Check.

Debounce:

- Recheck the contacts status rapidly and repeatedly for 50ms, and count the states for contacts A and B. (The 50ms duration may be adjusted for different encoders, as an encoder with a higher number of pulses per rotation will tend to create shorter pulses.)
- Compare the total number of changed states with unchanged states.

If the changed states are in a small minority, probably the signal was erroneous, caused by bounce or sliding noise. Go back to Check and start over.

Interpret:

- Deduce the rotational direction from these four possibilities:
  - — Contacts A were open and have closed.
  - — Contacts A were closed and have opened.
  - — Contacts B were open and have closed.
  - — Contacts B were closed and have opened. (The specific type of encoder will determine how these transitions are interpreted.)
- Revise the variable storing the direction of rotation if necessary.
- Depending on the direction of rotation, increment or decrement a variable that counts pulses.
- Take action that is appropriate to the direction of rotation and the cumulative number of pulses.
- Go back to Check again.

What Can Go Wrong

Switch Bounce

In addition to a debouncing algorithm in the microcontroller, a 0.1µF bypass capacitor can be used with each of the output terminals from the encoder, to help reduce the problem of switch bounce.

Contact Burnout

Rotational encoders are TTL-compatible. They are not generally designed to drive even a small output device, such as an LED. The contacts are extremely delicate and will be easily damaged by any attempt to switch a significant current.

relay

9

Properly known as an *electromagnetic armature relay* to distinguish it from a *solid-state relay*. However, the full term is very rarely used. It may also be described as an *electromechanical relay*, but the term **relay** is normally understood to mean a device that is not solid state.

OTHER RELATED COMPONENTS

- **solid state relay** (Volume 2)
- **switch** (See Chapter 6)

What It Does

A relay enables a signal or pulse of electricity to switch on (or switch off) a separate flow of electricity. Often, a relay uses a low voltage or low current to control a higher voltage and/or higher current. The low voltage/low current signal can be initiated by a relatively small, economical switch, and can be carried to the relay by relatively cheap, small-gauge wire, at which point the relay controls a larger current near to the load. In a car, for example, turning the ignition switch sends a signal to a relay positioned close to the starter motor.

While solid-state switching devices are faster and more reliable, relays retain some advantages. They can handle double-throw and/or multiple-pole switching and can be cheaper when high voltages or currents are involved. A comparison of their advantages relative to **solid state relays** and **transistors** is tabulated in the entry on **bipolar transistor** in Figure 28-15.

Common schematic symbols for single-throw relays are shown in Figure 9-1 and for double-throw relays in Figure 9-2. The appearance and orientation of the coil and contacts in the symbols may vary significantly, but the functionality remains the same.

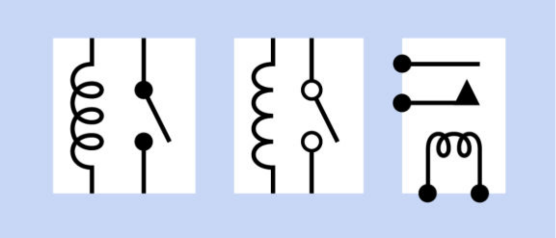

Figure 9-1. *Commonly used schematic symbols for a SPST relay. The symbols are functionally identical.*

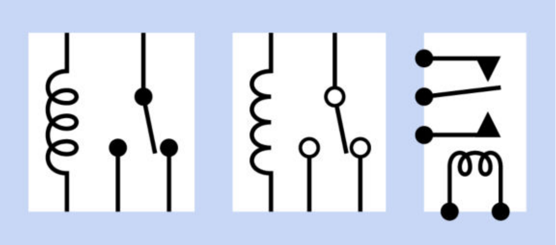

Figure 9-2. *Commonly used schematic symbols for a SPDT relay. The symbols are functionally identical.*

How It Works

A relay contains a *coil*, an *armature*, and at least one pair of *contacts*. Current flows through the coil, which functions as an **electromagnet** and generates a magnetic field. This pulls the armature, which is often shaped as a pivoting bracket that closes (or opens) the contacts. These parts are visible in the simplified rendering of a DPST relay in Figure 9-3. For purposes of identification, the armature is colored green, while the coil is red and the contacts are orange. The two blue blocks are made of an insulating material, the one on the left supporting the contact strips, the one on the right pressing the contacts together when the armature pivots in response to a magnetic field from the coil. Electrical connections to the contacts and the coil have been omitted for simplicity.

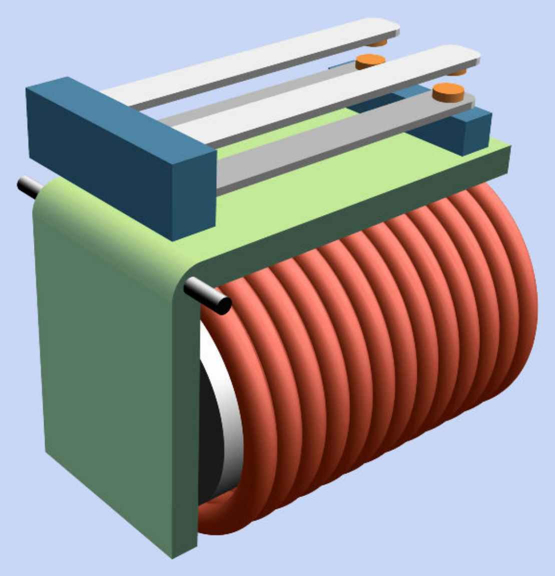

Figure 9-3. *This simplified rendering shows the primary parts of a DPST relay. See text for details.*

Various small relays, capable of handling a variety of voltages and currents, are pictured in Figure 9-4. At top-left is a 12VDC automotive relay, which plugs into a suitable socket shown immediately below it. At top-right is a 24VDC SPDT relay with exposed coil and contacts, making it suitable only for use in a very clean, dry environment. Continuing downward, the four sealed relays in colored plastic cases are designed to switch currents of 5A at 250VAC, 10A at 120VAC, 0.6A at 125VAC, and 2A at 30VDC, respectively. The two blue relays have 12VDC coils, while the red and yellow relays have 5V coils. All are nonlatching, except for the yellow relay, which is a latching type with two coils. At bottom-left is a 12VDC relay in a transparent case, rated to switch up to 5A at 240VAC or 30VDC.

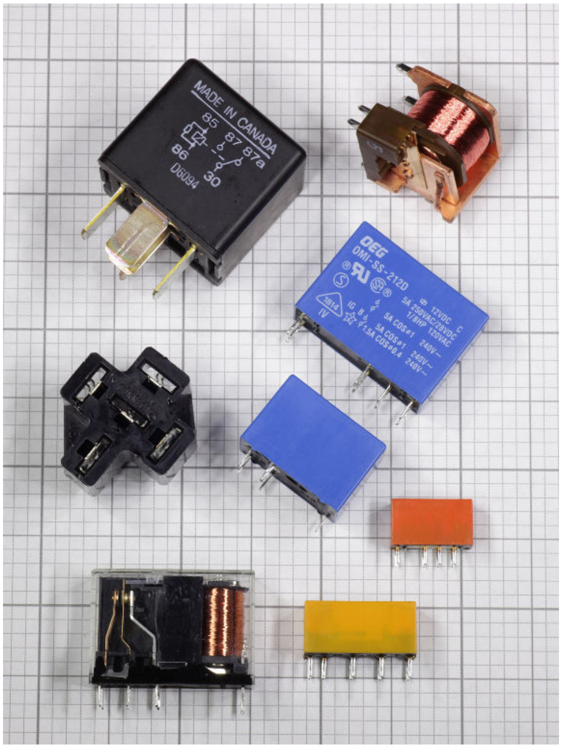

Figure 9-4. *An assortment of small DC-powered relays. See text for details.*

The configuration of a relay is specified using the same abbreviations that apply to a **switch**. SP, DP, 3P, and 4P indicate 1, 2, 3, or 4 poles (relays with more than 4 poles are rare). ST and DT indicate single-throw or double-throw switching. These abbreviations are usually concatenated, as in 3PST or SPDT. In addition, the terminology Form

A (meaning normally open), Form B (normally closed), and Form C (double-throw) may be used, preceded by a number that indicates the number of poles. Thus "2 Form C" means a DPDT relay.

Variants

Latching

There are two basic types of relay: *latching* and *nonlatching*. A nonlatching relay, also known as a *single side stable* type, is the most common, and resembles a *momentary switch* or **pushbutton** in that its contacts spring back to their default state when power to the relay is interrupted. This can be important in an application where the relay should return to a known state if power is lost. By contrast, a latching relay has no default state. Latching relays almost always have double-throw contacts, which remain in either position without drawing power. The relay only requires a short pulse to change its status. In semiconductor terms, its behavior is similar to that of a *flip-flop*.

In a *single-coil* latching relay, the polarity of voltage applied to the coil determines which pair of contacts will close. In a *dual-coil* latching relay, a second coil moves the armature between each of its two states.

Schematic symbols for a dual-coil latching relay are shown in Figure 9-5. Some symbol styles do not make it clear which switch position each coil induces. It may be necessary to read the manufacturer's datasheet or test the relay by applying its rated voltage to randomly selected terminal pairs while testing for continuity between other terminal pairs.

Polarity

There are three types of DC relay. In a *neutral relay*, polarity of DC current through the coil is irrelevant. The relay functions equally well either way. A *polarized relay* contains a diode in series with the coil to block current in one direction. A *biased relay* contains a permanent magnet near the armature, which boosts performance when current flows through the coil in one direction, but blocks a response when the current flows through the coil in the opposite direction. Manufacturers' datasheets may not use this terminology, but will state whether the relay coil is sensitive to the polarity of a DC voltage.

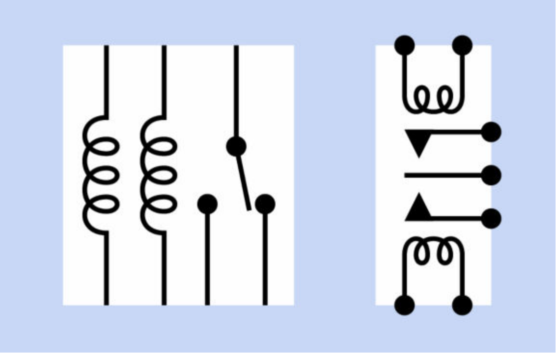

Figure 9-5. *Schematic symbols for a two-coil latching relay. The symbols are functionally identical.*

All relays can switch AC current, but only an *AC relay* is designed to use AC as its coil current.

Pinout Variations

The layout and function of relay pins or quick connects is not standardized among manufacturers. Often the component will have some indication of pin functions printed on it, but should always be checked against the manufacturer's datasheet and/or tested for continuity with a meter.

Figure 9-6 shows four sample pin configurations, adapted from a manufacturer's datasheet. These configurations are functionally quite different, although all of them happen to be for DPDT relays. In each schematic, the coil of the relay is shown as a rectangle, while the pins are circles, black indicating an energized state and white indicating a non-energized state. The bent lines show the possible connections between the poles and other contacts inside the relay. The contacts are shown as arrows. Thus, pole 4 can connect with either contact 3 or contact 5, while pole 9 can connect with either contact 8 or contact 10.

Top-left: Polarized nonlatching relay in its resting condition, with no power applied. Top right: Single-coil latching relay showing energized contacts (black circles) when the coil is powered with the polarity indicated. If the polarity is reversed, the relay flips to its opposite state. Some manufacturers indicate the option to reverse polarity by placing a minus sign alongside a plus sign, and a plus sign alongside a minus sign. Bottom-left and bottom-right: Polarized latching relays with two coils, with different pinouts.

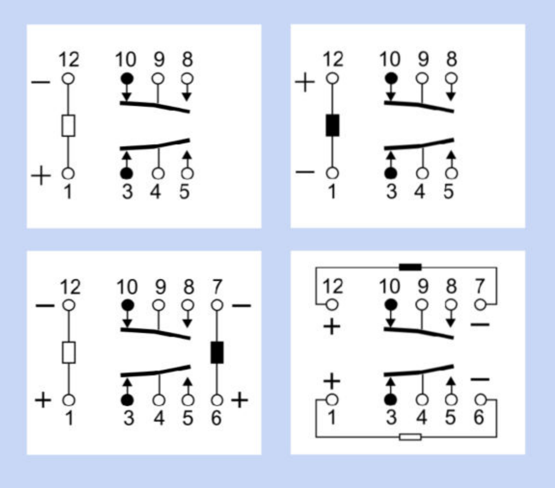

Figure 9-6. *Relay pinouts depicted in the style commonly found in manufacturers' datasheets, showing different relay types. Top-left: Single coil, nonlatching. Top-right: Single coil, latching. Bottom left: Two-coil, latching. Bottom right: Two-coil, latching, alternate pinouts. (Adapted from a Panasonic datasheet.)*

In these diagrams, the relay is seen from above. Some datasheets show the relay seen from below, and some show both views. Some manufacturers use slightly different symbols to indicate interior functions and features. When in doubt, use a meter for verification.

Reed Relay

A *reed relay* is the smallest type of electromechanical relay with applications primarily in test equipment and telecommunications. With a coil resistance ranging from 500 to 2000 ohms, these relays consume very little power. The design consists of a *reed switch* with a coil wrapped around it. Figure 9-7 shows a simplified rendering. The two black contacts are enclosed in a glass or plastic envelope and magnetized in such a way that a magnetic field from the surrounding coil bends them together, creating a connection. When power to the coil is disconnected, the magnetic field collapses and the contacts spring apart.

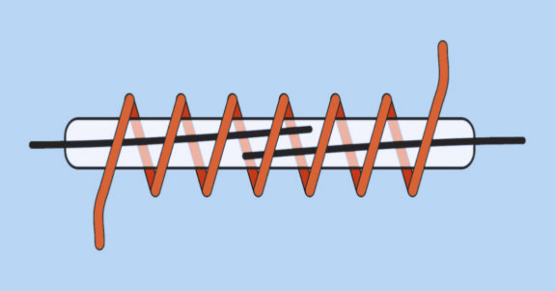

Figure 9-7. *This simplified rendering shows a reed relay, consisting of a magnetized reed switch inside a glass or plastic pod, activated by a coil wrapped around it.*

In Figure 9-8, two reed relays are shown, at top-left and center-right. At bottom-left, the type of relay on the right has been opened by a belt sander to reveal its copper coil and inside that, a capsule in which the relay contacts are visible.

Surface-mount reed relays can be smaller than 0.5″ × 0.2″. Through-hole versions are often around 0.7″ × 0.3″ with pins in two rows, though some are available in SIP packages.

Reed relays have limited current switching capacity and are not suitable to switch inductive loads.

Small Signal Relay

A *small signal relay* is also known as a *low signal relay*. This type may have a footprint as small as a reed relay but generally stands slightly taller, requires slightly more coil current, and is available in versions that can switch slightly higher

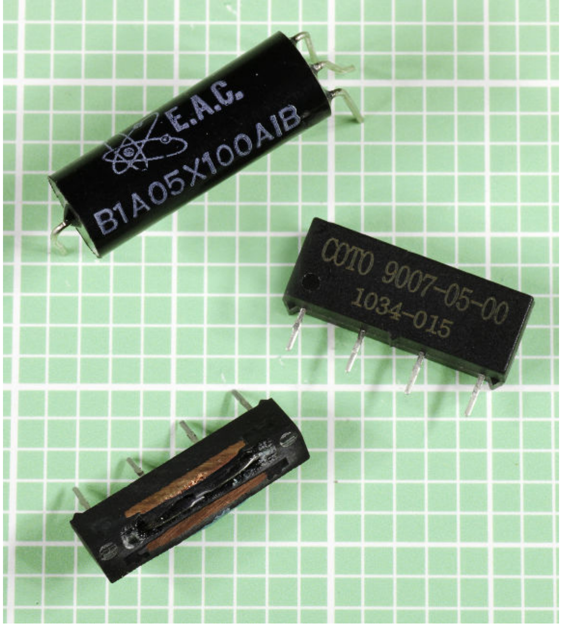

Figure 9-8. *Three reed relays, one of which has had its packaging partially removed by a belt sander to reveal its copper coil and internal contacts.*

voltages and currents. There are usually two rows of pins, spaced either 0.2″ or 0.3″ apart. The red and orange relays in Figure 9-4 are small signal relays.

Automotive Relays

An *automotive relay* is typically packaged in a cube-shaped black plastic case with quick-connect terminals at the bottom, typically plugged into a socket. Naturally they are designed to switch, and be switched by, a 12VDC supply.

General Purpose/Industrial

These relays cover a very wide range and are usually built without significant concern for size. They may be capable of switching high currents at high voltages. Typically they are designed to plug into a socket such as an *octal base* of the type that was once used for **vacuum tubes**. The base, in turn, terminates in solder tabs, screws, or quick connects and is designed to be screwed to a chassis. It allows the relay to be unplugged and swapped without resoldering.

Two industrial relays are shown in Figure 9-9. Both are DPDT type with 12VDC coils and rated to switch up to 10A at 240VAC. The one on the left has an octal base. An octal socket that fits an octal base is shown in Figure 9-10.

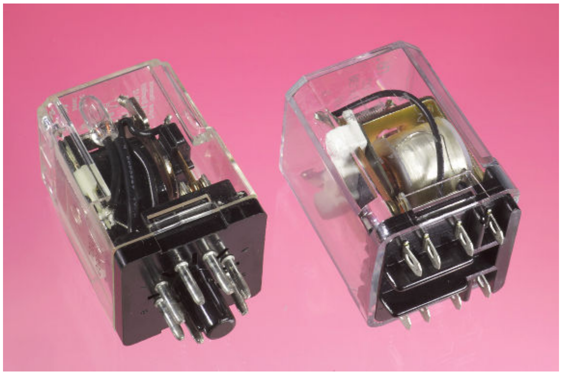

Figure 9-9. *Two relays powered by 12VDC, capable of switching up to 10A at 240VAC.*

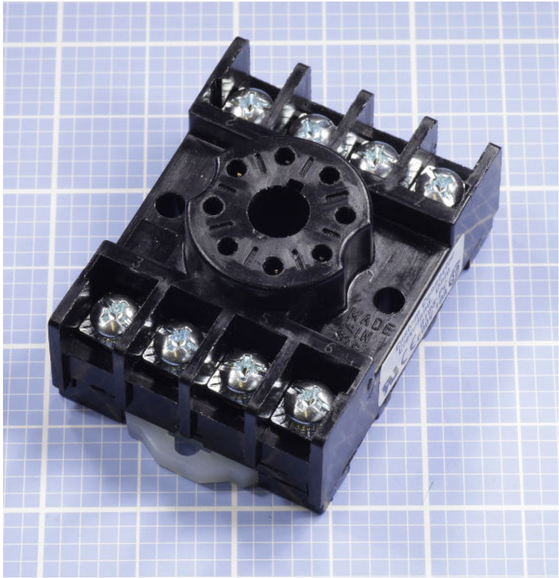

Figure 9-10. *An octal socket with screw terminals, designed to accept a relay with an octal base.*

Time Delay Relay

Generally used to control industrial processes, a *time delay relay* switches an output on and off at preset time intervals that can be programmed to repeat. The example in Figure 9-11 has a 12VDC coil and is rated to switch up at 10A at 240VAC. It has an octal base.

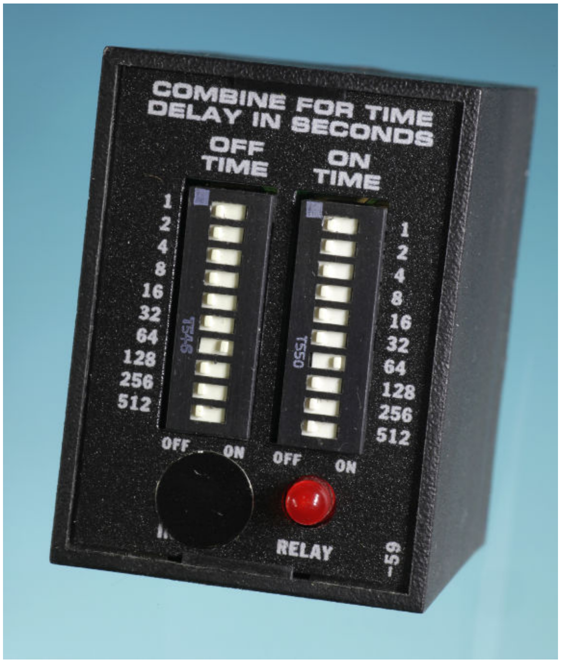

Figure 9-11. *The control switches on a time-delay relay, allowing separately configured "on" and "off" intervals.*

Contactor

A contactor functions just like a relay but is designed to switch higher currents (up to thousands of amperes) at higher voltages (up to many kilovolts). It may range from being palm-sized to measuring more than one foot in diameter, and may be used to control heavy loads such as very large motors, banks of high-wattage lights, and heavy-duty power supplies.

Values

Datasheets usually specify maximum voltage and current for the contacts, and nominal voltage and current for the coil, although in some cases the coil resistance is stated instead of nominal coil current. The approximate current consumption can be estimated, if necessary, by using Ohm's Law. The minimum voltage that the relay needs for activation is sometimes described as the *Must Operate By* voltage, while the *Must Release By* voltage is the maximum coil voltage that the relay will ignore. Relays are rated on the assumption that the coil may remain energized for long periods, unless otherwise stated.

While the contact rating may suggest that a relay can switch a large load, this is not necessarily true if the load has significant inductance.

Reed relays

Usually use a coil voltage of 5VDC and have a contact rating of up to 0.25A at 100V. Through-hole (PCB) versions may have a coil voltage of 5VDC, 6VDC, 12VDC, or 24VDC and in some cases claim to switch 0.5A to 1A at up to 100V, although this rating is strictly for a noninductive load.

Small signal/low signal relays

Usually use a coil voltage ranging from 5VDC to 24VDC, drawing about 20mA. Maximum switching current for noninductive loads ranges from 1A to 3A.

Industrial/general purpose relays

A very wide range of possible values, with coil voltages ranging up to 48VDC or 125VAC to 250VAC. Contact rating is typically 5A to 30A.

Automotive relays

Coil voltage of 12VDC, and contact rating often 5A at up to 24VDC.

Timer relays

Usually these specify a coil voltage of 12VDC, 24VDC, 24VAC, 125VAC, or 230VAC. The timed interval can range from 0.1 sec to 9999 hours in some cases. Common values for contact ratings are 5A up to 20A, with a voltage of 125V to 250V, AC or DC.

How to Use it

Relays are found in home appliances such as dishwashers, washing machines, refrigerators, air conditioners, photocopy machines, and other products where a substantial load (such as a motor or compressor) has to be switched on and off by a control switch, a thermostat, or an electronic circuit.

Figure 9-12 shows a common small-scale application in which a signal from a microcontroller (a few mA at 5VDC) is applied to the base of a transistor, which controls the relay. In this way, a logic output can switch 10A at 125VAC. Note the rectifier diode wired in parallel with the relay coil.

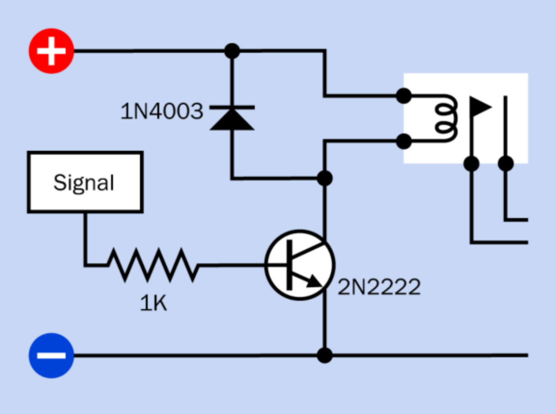

Figure 9-12. *A signal from a digital source such as a microcontroller can switch substantial voltage and current if it is applied to the base of a transistor that activates a relay.*

A latching relay is useful wherever a connection should persist when power is switched off or interrupted, or if power consumption must be minimized. Security devices are one common application. However, the circuit may require a "power reset" function to restore known default settings of latched relays.

A circuit including every possible protection against voltage spikes is shown in Figure 9-13, including a snubber to protect the relay contacts, a rectifier diode to suppress back-EMF generated by the relay coil, and another rectifier diode to protect the relay from EMF generated by a motor when the relay switches it on and off. The snubber can be omitted if the motor draws a relatively low current (below 5A) or if the relay is switching a noninductive load. The diode around the relay coil can be omitted if there are no semiconductors or other components in the circuit that are vulnerable to voltage spikes. However, a spike can affect components in adjacent circuits that appear to be electrically isolated. A severe spike can even be transmitted back into 125VAC house wiring. For information on using a resistor-capacitor combination to form a snubber, see "Snubber" (page 108).

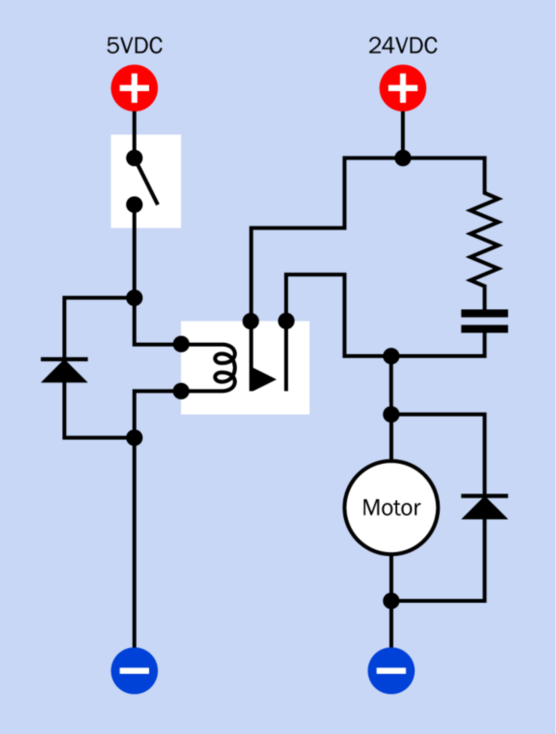

Figure 9-13. *This hypothetical schematic shows three types of protection against voltage spikes induced by an inductive load (a motor, in this instance) and the coil of the relay.*

What Can Go Wrong

Wrong Pinouts

The lack of standardization of relay pinouts can cause errors if one relay is replaced with another that appears to be the same, but isn't. In particular, the pins that connect with normally-closed contacts may trade places with pins connected with normally-open contacts, in relays from different manufacturers.

Pinouts are also confusing in that some datasheets depict them from above, some from below, and some from both perspectives.

Wrong Orientation

Small relays of through-hole type usually have pins spaced in multiples of 0.1". This allows them to be inserted the wrong way around in a perforated board. Almost all relays have an identification mark molded into one end or one corner of the plastic shell. Manufacturers do not standardize the position and meaning of these marks, but they are usually replicated in datasheets. When using a relay of a type that you have not used before, it is a sensible precaution to test it with a meter to verify the functions of its terminals before installing it.

Wrong Type

A latching relay may have exactly the same appearance as a nonlatching relay from the same manufacturer, and the same two pins may energize the coil. However, in a latching relay, the contacts won't spring back to their non-energized position, causing functional errors that may be difficult to diagnose. The part numbers printed on latching and nonlatching versions of the same relay may differ by only one letter or numeral and should be checked carefully.

Wrong Polarity

A relay with a DC-energized coil may require power to be applied with correct polarity and may malfunction otherwise.

AC and DC

A relay coil designed to be powered by DC will not work from AC and vice-versa. The contact rating of a relay is likely to be different depending whether it is switching AC or DC.

Chatter

This is the noise created by relay contacts when they make rapid intermittent connection. Chatter is potentially damaging to relay contacts and should be avoided. It can also create electrical noise that interferes with other components. Likely correctible causes of chatter include insufficient voltage or power fluctuations.

Relay Coil Voltage Spike

A relay coil is an inductive device. Merely switching a large relay on and off can create voltage spikes. To address this problem, a rectifier diode should be placed across the coil terminals with polarity opposing the energizing voltage.

Arcing

This problem is discussed in the **switch** entry of this encyclopedia. See "Arcing" (page 47). Note that because the contacts inside a *reed relay* are so tiny, they are especially susceptible to arcing and may actually melt and weld themselves together if they are used to control excessive current or an inductive load.

Magnetic Fields

Relays generate magnetic fields during operation and should not be placed near components that are susceptible.

The reed switch inside a reed relay can be unexpectedly activated by an external magnetic field. This type of relay may be enclosed in a metal shell to provide some protection. The adequacy of this protection should be verified by testing the relay under real-world conditions.

Environmental Hazards

Dirt, oxidation, or moisture on relay contacts is a significant problem. Most relays are sealed and should remain sealed.

Relays are susceptible to vibration, which can affect the contacts and can accelerate wear on moving parts. Severe vibration can even damage a relay permanently. **Solid-state relays** (discussed in Volume 2) should be used in harsh environments.

resistor

10

OTHER RELATED COMPONENTS

- **potentiometer** (See Chapter 11)

What It Does

A resistor is one of the most fundamental components in electronics. Its purpose is to impede a flow of current and impose a voltage reduction. It consists of two wires or conductors attached at opposite ends or sides of a relatively poor electrical conductor, the resistance of which is measured in ohms, universally represented by the Greek omega symbol, Ω.

Schematic symbols that represent a resistor are shown in Figure 10-1 (Left: The traditional schematic symbol. Right: The more recent European equivalent). The US symbol is still sometimes used in European schematics, and the European symbol is sometimes used in US schematics. Letters K or M indicate that the value shown for the resistor is in thousands of ohms or millions of ohms, respectively. Where these letters are used in Europe, and sometimes in the US, they are substituted for a decimal point. Thus, a 4.7K resistor may be identified as 4K7, a 3.3M resistor may be identified as 3M3, and so on. (The numeric value in Figure 10-1 was chosen arbitrarily.)

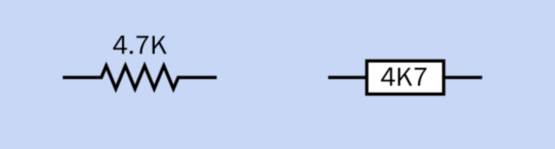

Figure 10-1. *Resistor symbols. The left one is more common in the United States, while the right one is widely used in Europe. The 4.7K value was chosen arbitrarily.*

A resistor is commonly used for purposes such as limiting the charging rate of a **capacitor**; providing appropriate control voltage to semiconductors such as **bipolar transistors**; protecting **LED**s or other semiconductors from excessive current; adjusting or limiting the frequency response in an audio circuit (in conjunction with other components); pulling up or pulling down the voltage at the input pin of a digital logic chip; or controlling a voltage at a point in a circuit. In this last application, two resistors may be placed in series to create a *voltage divider*.

A **potentiometer** may be used instead of a resistor where variable resistance is required.

Sample resistors of various values are shown in Figure 10-2. From top to bottom, their power dissipation ratings are 3W, 1W, 1/2W, 1/4W, 1/4W, 1/4W, and 1/8W. The accuracy (tolerance) of each resistor, from top to bottom, is plus-or-minus 5%, 5%, 5%, 1%, 1%, 5%, and 1%. The beige-colored body of a resistor is often an indication that its tolerance is 5%, while a blue-colored body often indicates a tolerance of 1% or 2%. The blue-bodied resistors and the dark brown resistor contain metal-oxide film elements, while the beige-bodied resistors and the green resistor contain carbon film. For more information on resistor values, see the upcoming *Values* section.

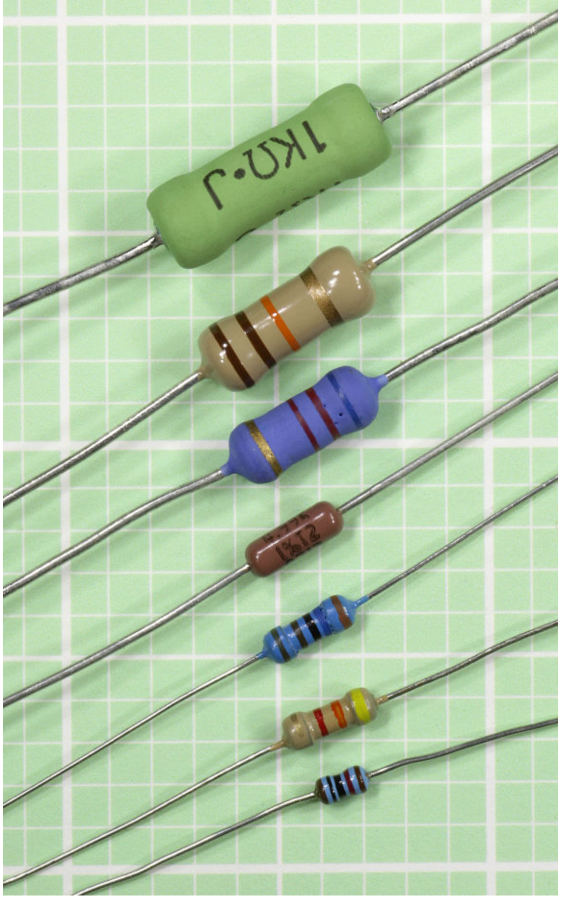

Figure 10-2. *A range of typical resistors. See text for details.*

How It Works

In the process of impeding the flow of current and reducing voltage, a resistor absorbs electrical energy, which it must dissipate as heat. In most modern electronic circuits, the heat dissipation is typically a fraction of a watt.

If R is the resistance in ohms, I is the current flowing through the resistor in amperes, and V is the voltage drop imposed by the resistor (the difference in electrical potential between the two contacts that are attached to it), Ohm's law states:

```
V = I / R
```

This is another way of saying that a resistor of 1Ω will allow a current of 1 amp when the potential difference between the ends of the resistor is 1 volt.

If W is the power in watts dissipated by the resistor, in a DC circuit:

```
W = V * I
```

By substitution in Ohm's law, we can express watts in terms of current and resistance:

$$W = I^2 / R$$

We can also express watts in terms of voltage and resistance:

$$W = V^2 * R$$

These alternates may be useful in situations where you do not know the voltage drop or the current, respectively.

Approximately similar relationships exist when using alternating current, although the power will be a more complex function.

Variants

Axial resistors have two leads that emerge from opposite ends of a usually cylindrical body. *Radial resistors* have parallel leads emerging from one side of the body and are unusual.

Precision resistors are generally defined as having a tolerance of no more than plus-or-minus 1%.

General-purpose resistors are less stable, and their value is less precise.

Power resistors are generally defined as dissipating 1 or 2 watts or more, particularly in power supplies or power amplifiers. They are physically larger and may require *heat sinks* or fan cooling.

Wire-wound resistors are used where the component must withstand substantial heat. A wire-wound resistor often consists of an insulating tube or core that is flat or cylindrical, with multiple turns of resistive wire wrapped around it. The wire is usually a nickel-chromium alloy known as *nichrome* (sometimes written as *Ni-chrome*) and is dipped in a protecting coating.

The heat created by current passing through resistive wire is a potential problem in electronic circuits where temperature must be limited. However, in household appliances such as hair dryers, toaster ovens, and fan heaters, a nichrome element is used specifically to generate heat. Wire-wound resistors are also used in 3D printers to melt plastic (or some other compound) that forms the solid output of the device.

Thick film resistors are sometimes manufactured in a flat, square format. A sample is shown in Figure 10-3, rated to dissipate 10W from its flat surface. The resistance of this component is 1K.

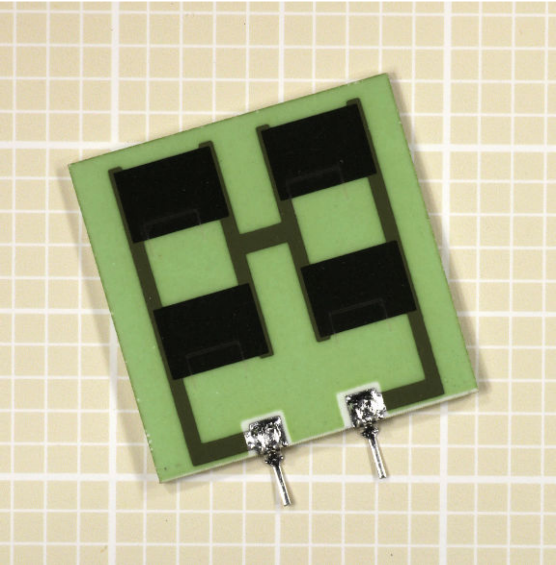

Figure 10-3. *A thick-film resistor measuring about 1" square and 0.03" thick.*

Surface-mount resistors generally consist of a resistive ink film printed on top of a tablet of aluminum oxide ceramic compound, often approximately 6mm long, known as a 2512 form factor. Each surface-mount resistor has two nickel-plated terminations coated in solder, which melts when the resistor is attached to the circuit board. The upper surface is coated, usually with black epoxy, to protect the resistive element.

Resistor Array

This is also known as a *resistor network* or *resistor ladder*, and consists of a chip containing multiple equal-valued resistors.

A resistor array in a *single-inline package* (or *SIP*) may have three possible internal configurations: isolated, common bus, and dual terminator. These options are shown at top, center, and bottom, respectively, in Figure 10-4. The isolated variant is commonly available in SIPs with 6, 8, or 10 pins. The common-bus and dual-terminator configurations generally have 8, 9, 10, or 11 pins.

In the isolated configuration, each resistor is electrically independent of the others and is accessed via its own pair of pins. On a common bus, one end of each resistor shares a bus accessed by a single pin, while the other ends of the resistors are accessed by their own separate pins. A dual-terminator configuration is more complex, consisting of pairs of resistors connected between ground and an internal bus, with the midpoint of each resistor pair accessible via a separate pin. The resistor pairs this function as voltage dividers and are commonly used in emitter-coupled logic circuits that require termination with -2 volts.

A *dual-inline package* (*DIP*) allows a similar range of internal configurations, as shown in Figure 10-5. At top, isolated resistors are commonly available in DIPs with 4, 7, 8, 9, or 10 pins. At center, the common bus configuration is available in DIPs with 8, 14, 16, 18, or 20 pins. At bottom, the dual-terminator configuration usually has 8, 14, 16, 18, or 20 pins.

The external appearance of SIP and DIP resistor arrays is shown in Figure 10-6. From left to right, the packages contain seven 120Ω resistors in

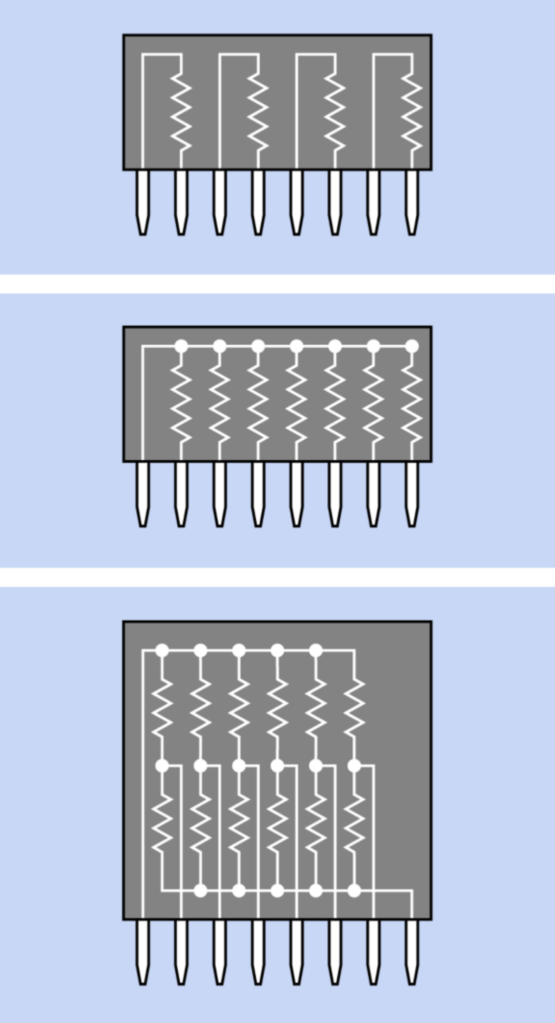

Figure 10-4. *Multiple resistors can be embedded in a single-inline package (SIP) in a variety of formats. See text for additional details.*

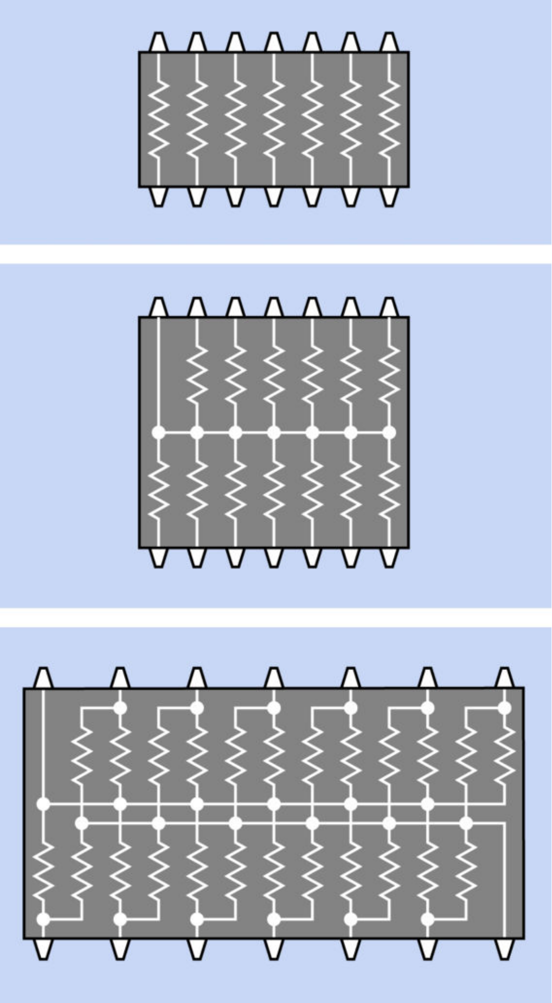

Figure 10-5. *Multiple resistors can be obtained embedded in a dual-inline package (DIP). See text for additional details.*

isolated configuration; thirteen 120Ω resistors in bussed configuration; seven 5.6K resistors in bussed configuration; and six 1K resistors in bussed configuration.

Resistor arrays with isolated or common-bus configurations are a convenient way to reduce the component count in circuits where pullup, pulldown, or terminating resistors are required for multiple chips. The common-bus configuration is also useful in conjunction with a 7-segment **LED display**, where each segment must be terminated by a series resistor and all the resistors share a common ground or common voltage source.

Surface-mount chips are available containing a pair of resistors configured as a single voltage divider.

Chips containing multiple *RC circuits* (each consisting of a capacitor and a resistor in series) are available, although uncommon. A package containing a single RC circuit may be sold as a

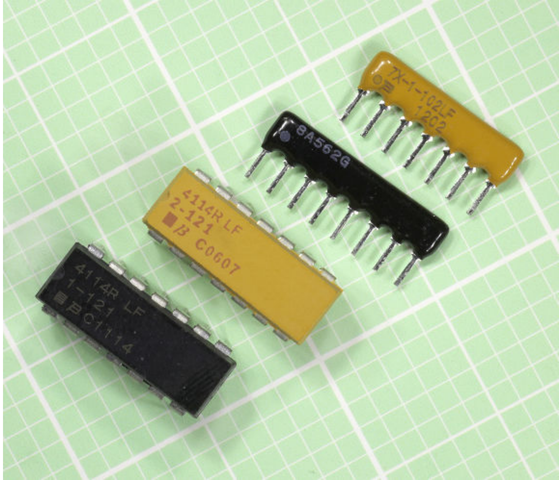

Figure 10-6. *Resistor arrays in DIP and SIP packages. See text for values.*

snubber to protect contacts in a **switch** or **relay** that switches a large inductive load. More information on snubber circuits is in the **capacitor** entry of this encyclopedia; see "Snubber" (page 108).

Values

1 *kilohm,* usually written as 1K, is 1,000Ω. 1 *megohm,* usually written as 1M or 1 meg, is 1,000K. 1 *gigaohm* is 1,000 megs, although the unit is rarely used. Resistances of less than 1Ω are uncommon and are usually expressed as a decimal number followed by the Ω symbol. The term *milliohms* (thousandths of an ohm) is used in special applications. Equivalent resistor values are shown in Figure 10-7.

A resistance value remains unchanged in DC and AC circuits, except where the AC reaches an extremely high frequency.

In common electronics applications, resistances usually range from 100Ω to 10M. Power ratings may vary from 1/16 watt to 1000 watts, but usually range from 1/8 watt to 1/2 watt in most electronic circuits (less in surface-mount applications).

Ohms	Kilohms	Megohms
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 10-7. *Equivalent values in ohms, kilohms, and megohms.*

Tolerance

The tolerance, or precision, of a resistor may range from plus-or-minus 0.001% up to plus-or-minus 20%, but is most commonly plus-or-minus 1%, 2%, 5%, or 10%.

The traditional range of resistor values was established when a tolerance of 20% was the norm. The values were spaced to allow minimum risk of a resistor at one end of its tolerance range having the same value as another resistor at the opposite end of its tolerance range. The values were rounded to 10, 15, 22, 33, 47, 68, and 100, as illustrated in Figure 10-8 where each blue diamond represents the possible range of actual values of a 20% resistor with a theoretical value shown by the white horizontal line at the center of the diamond.

Resistor factors repeat themselves in multiples of 10. Thus, for example, beginning with a resistor of 100Ω, subsequent increasing values will be 150, 220, 330, 470, 680, and 1K, whereas the range of resistors beginning with 1Ω will be 1.5, 2.2, 3.3, 4.7, 6.8, and 10Ω.

Resistance multiplication factors are now expressed as a list of preferred values by the International Electrotechnical Commission (IEC) in their 60063 standard. Intermediate factors have been added to the basic sequence to

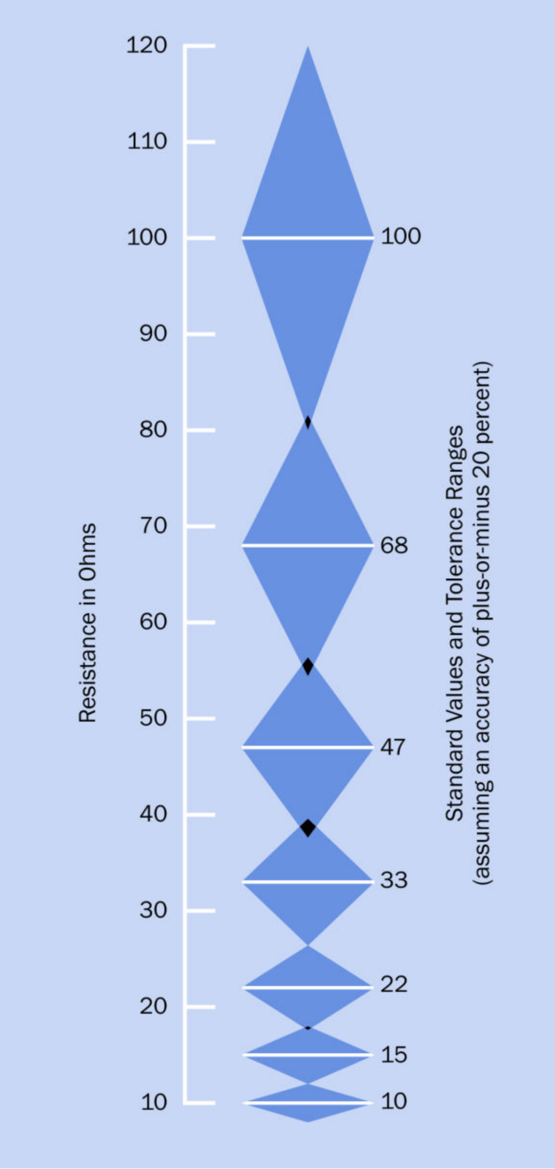

Figure 10-8. *Graphical representation of standard resistor values (white lines) established by the International Electrotechnical Commission, showing the acceptable range of actual values (dark blue areas) assuming precision of plus-or-minus 20%. The overlap, if any, between each range and the next is shown in black.*

accommodate better tolerances. A table showing resistor values for tolerances of plus-or-minus 20%, 10%, and 5% appears in Figure 10-9. Resistors with a tolerance of 5% have become increasingly common.

	Resistor Tolerances (plus or minus)		
	20%	10%	5%
Standard Resistor Values	100	100	100
			91
		82	82
			75
	68	68	68
			62
		56	56
			51
	47	47	47
			43
		39	39
			36
	33	33	33
			30
		27	27
			24
	22	22	22
			20
		18	18
			16
	15	15	15
			13
		12	12
			11
	10	10	10

Figure 10-9. *Standard values for resistors of different precisions. For resistors outside the range shown, values can be found by multiplying or dividing (repeatedly, if necessary) by a factor of 10.*

The IEC has established 3-digit preferred values for resistors with values accurate to plus-or-minus 0.5%.

Because many capacitors still have a tolerance no better than 20%, their values also conform with the old original set of resistance values, although the units are expressed in farads or fractions of a farad. See the **capacitors** entry in this encyclopedia for additional information.

Value Coding

Through-hole axial resistors are traditionally printed with a sequence of three colored bands to express the value of the component, each of the first two bands representing a digit from 0 through 9, while the third band indicates the decimal multiplier (the number of zeroes, from 0 to 9, which should be appended to the digits). A fourth band of silver or gold indicates 10% or 5% tolerance respectively. No fourth band would indicate 20% tolerance, although this has become very rare.

Many resistors now have five color bands, to enable the representation of intermediate or fractional values. In this scheme, the first three bands have numeric values (using the same color system as before) while the fourth band is the multiplier. A fifth band, at the opposite end of the resistor, indicates its tolerance.

In Figure 10-10 the numeric or multiplier value of each color is shown as a "spectrum" at the top of the figure. The tolerance, or precision of a resistor, expressed as a plus-or-minus percentage, is shown using silver, gold, and various colors, at the bottom of the figure.

Two sample resistors are shown. The upper one has a value of 1K, indicated by the brown and black bands on the left (representing numeral 1 followed by a numeral 0) and the third red band (indicating two additional zeroes). The gold band at right indicates a precision of 5%. The lower one has a value of 1.05K, indicated by the brown, black, and green bands on the left (representing numeral 1 followed by numeral 0 followed by a numeral 5) and the fourth band brown (indicating one additional zero). The brown band at right indicates a precision of 1%.

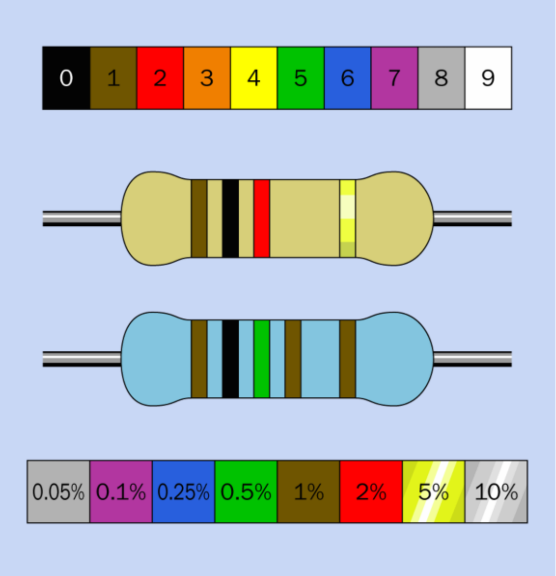

Figure 10-10. *Color coding of through-hole resistors. See text for details.*

In extremely old equipment, resistors may be coded with the body-tip-dot scheme, in which the body color represents the initial digit, the end color represents the second digit, and a dot represents the multiplier. The numeric identities of the colors is the same as in the current color scheme.

In all modern schemes, the three or four bands that show the resistance value are spaced close together, while a larger gap separates them from the band that shows the tolerance. The resistor value should be read while holding the resistor so that the group of closely-spaced numeric bands is on the left.

Confusingly, some resistors may be found where the first three bands define the value, using the old three-band convention; the fourth band indicates tolerance; and a fifth band at the opposite end of the component indicates reliability. However, this color scheme is uncommon.

Other color-coding conventions may be found in special applications, such as military equipment.

It is common for through-hole carbon-film resistors to have a beige body color, while through-hole metal-film resistors often have a blue body color. However, in relatively rare instances, a blue body color may also indicate a *fusible* resistor (designed to burn out harmlessly like a fuse, if it is overloaded) while a white body may indicate a *non-flammable* resistor. Use caution when replacing these special types.

Some modern resistors may have their values printed on them numerically. Surface-mount resistors also have digits printed on them, but they are a code, not a direct representation of resistance. The last digit indicates the number of zeroes in the resistor value, while the preceding two or three numbers define the value itself. Letter R is used to indicate a decimal point. Thus a 3R3 surface-mount resistor has a value of 3.3Ω, while 330 would indicate 33Ω, and 332 indicates 3,300Ω. A 2152 surface-mount resistor would have a value of 21,500Ω.

A surface-mount resistor with a single zero printed on it is a *zero ohm* component that has the same function as a **jumper** wire. It is used for convenience, as it is easily inserted by automated production-line equipment. It functions merely as a bridge between traces on the circuit board.

When resistor values are printed on paper in schematics, poor reproduction may result in omission of decimal points or introduction of specks that look like decimal points. Europeans have addressed this issue by using the letter as a substitute for a decimal point so that a 5.6K resistor will be shown as 5K6, or a 3.3M resistor will be shown as 3M3. This practice is followed infrequently in the United States.

Stability

This term describes the ability of a resistor to maintain an accurate value despite factors such as temperature, humidity, vibration, load cycling, current, and voltage. The *temperature coefficient* of a resistor (often referred to as T_{cr} or T_c, not to be confused with the time constant of a charging capacitor) is expressed in parts per million change in resistance for each degree centigrade deviation from room temperature (usually assumed to be 25 degrees Centigrade). T_c may be a positive or a negative value.

The *voltage coefficient* of resistance—often expressed as V_c—describes the change of a resistor's value that may occur as a function of changes in voltage. This is usually significant only where the resistive element is carbon-based. If V1 is the rated voltage of the resistor, R1 is its rated resistance at that voltage, V2 is 10% of the rated voltage, and R2 is the actual resistance at that voltage, the voltage coefficient, V_c, is given by this formula:

$$V_c = (100 * (R1 - R2)) / (R2 * (V1 - V2))$$

Materials

Resistors are formed from a variety of materials.

Carbon composite. Particles of carbon are mixed with a binder. The density of the carbon determines the end-to-end resistance, which typically ranges from 5Ω to 10M. The disadvantages of this system are low precision (a 10% tolerance is common), relatively high voltage coefficient of resistance, and introduction of noise in sensitive circuits. However, carbon-composite resistors have low inductance and are relatively tolerant of overload conditions.

Carbon film. A cheap and popular type, made by coating a ceramic substrate with a film of carbon compound. They are available in both through-hole and surface-mount formats. The range of resistor values is comparable with carbon-composite types, but the precision is increased, typically to 5%, by cutting a spiral groove in the carbon-compound coating during the manufacturing process. The carbon film suffers the same disadvantages of carbon composite resistors, but to a lesser extent. Carbon film resistors generally should not be substituted for metal film resistors in applications where accuracy is important.

Metal film. A metallic film is deposited on a ceramic substrate, and has generally superior characteristics to carbon-film resistors. During manufacture, a groove may be cut in the metal film to adjust the end-to-end resistance. This may cause the resistor to have higher inductance than carbon-composite types, though it has lower noise. Tolerances of 5%, 2%, and 1% are available. This type of resistor was originally more expensive than carbon-film equivalents, but the difference is now fractional. They are available in both through-hole and surface-mount formats. They are available in lower-wattage variants (1/8 watt is common).

Thick-film resistors are spray-coated, whereas thin-film resistors are sputtered nichrome. Thin-films enjoy a flatter temperature coefficient and are typically used in environments that have a huge operational temperature range, such as satellites.

Bulk metal foil. The type of foil used in metal film resistors is applied to a ceramic wafer and etched to achieve the desired overall resistance. Typically these resistors have axial leads. They can be extremely accurate and stable, but have a limited maximum resistance.

Precision wire-wound. Formerly used in applications requiring great accuracy, but now largely replaced by precision metal foil.

Power wire-wound. Generally used when 1 or 2 watts or more power dissipation is required. Resistive wire is wrapped around a core that is often ceramic. This can cause the resistor to be referred to, inaccurately, as "ceramic." The core may alternatively be fiberglass or some other electrically insulating compound that actively sinks heat. The component is either dipped (typically in vitreous enamel or cement) or is mounted in an aluminum shell that can be clamped to a heat sink. It almost always has the ohm value printed on it in plain numerals (not codes).

Two typical wire-wound resistors are shown in Figure 10-11. The upper resistor is rated at 12W and 180Ω while the lower resistor is rated at 13W and 15K.

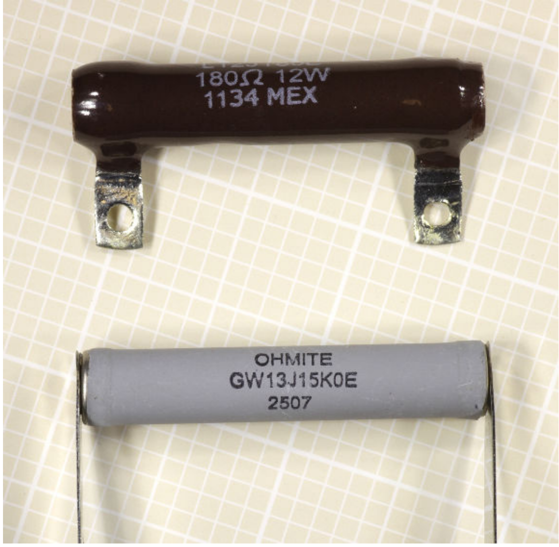

Figure 10-11. *Two wire-wound resistors of greatly differing resistance but similar power dissipation capability.*

A larger wire-wound resistor is shown in Figure 10-12, rated for 25W and 10Ω.

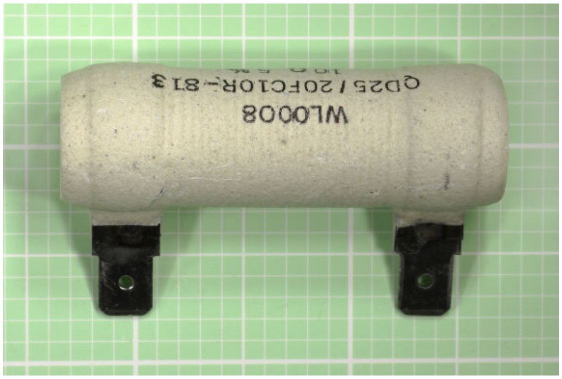

Figure 10-12. *A large wirewound resistor rated to dissipate 25W.*

In Figure 10-13, two resistors encapsulated in cement coatings are shown with the coatings removed to expose the elements. At left is a 1.5Ω 5W resistor, which uses a wire-wound element. At right is a very low-value 0.03Ω 10W resistor.

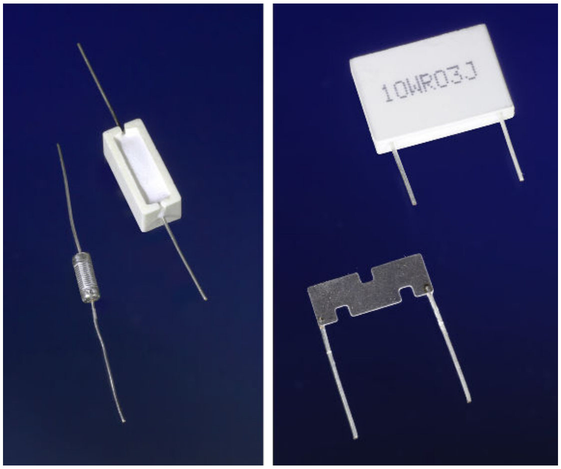

Figure 10-13. *Two low-value resistors with their cement coatings removed to show the resistive elements.*

In Figure 10-14, the resistor at right has an exposed 30Ω element while the resistor at left is rated 10W and 6.5Ω, enclosed in an anodized aluminum shell to promote heat dissipation.

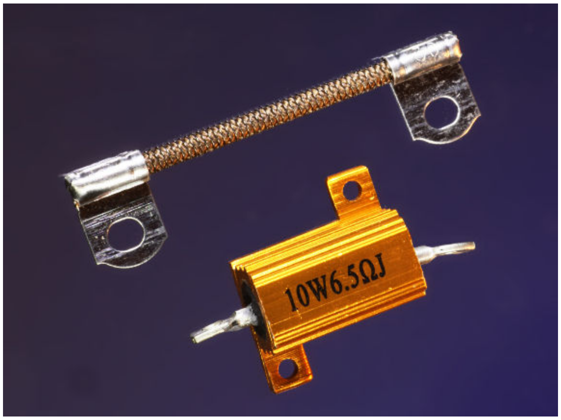

Figure 10-14. *A 30Ω resistor (right) and 6.5Ω resistor (left).*

In power resistors, heat dissipation becomes an important consideration. If other factors (such as voltage) remain the same, a lower-value resistor will tend to pass more current than a higher-value resistor, and heat dissipation is proportional to the square of the current. Therefore power wire-wound resistors are more likely to be needed where low resistance values are required. Their coiled-wire format creates significant inductance, making them unsuitable to pass high frequencies or pulses.

How to Use it

Some of the most common applications for a resistor are listed here.

In Series with LED

To protect an LED from damage caused by excessive current, a *series resistor* is chosen to allow a current that does not exceed the manufacturer's specification. In the case of a single through-hole LED (often referred to as an *indicator*), the forward current is often limited to around 20mA, and the value of the resistor will depend on the voltage being used. (See Figure 10-15.)

When using high-output LEDs (which may contain multiple elements in a single 5mm or 10mm package), or LED arrays that are now being used for domestic lighting, the acceptable current may be much greater, and the LED unit may contain its own current-limiting electronics. A datasheet should be consulted for details.

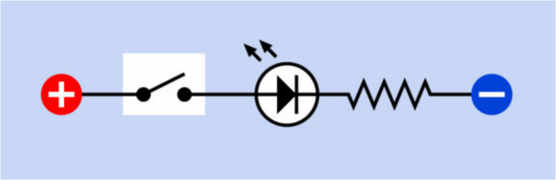

Figure 10-15. *A series resistor is necessary to limit the current that passes through an LED.*

Current Limiting with a Transistor

In Figure 10-16, a transistor is switching or amplifying current flowing from B to C. A resistor is used to protect the base of the transistor from excessive current flowing from point A. Resistors are also commonly used to prevent excessive current from flowing between B and C.

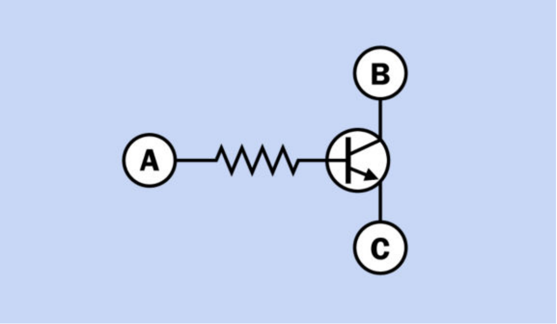

Figure 10-16. *A resistor is typically necessary to protect the base of a transistor from excessive current.*

Pullup and Pulldown Resistors

When a mechanical switch or pushbutton is attached to the input of a logic chip or **microcontroller**, a *pullup* or *pulldown* resistor is used, applying positive voltage or grounding the pin, respectively, to prevent it from "floating" in an indeterminate state when the switch is open. In Figure 10-17, the upper schematic shows a pulldown resistor, whereas the lower schematic shows a pullup resistor. A common value for either of them is 10K. When the pushbutton is pressed, its direct connection to positive voltage or to ground easily overwhelms the effect of the resistor. The choice of pullup or pulldown resistor may depend on the type of chip being used.

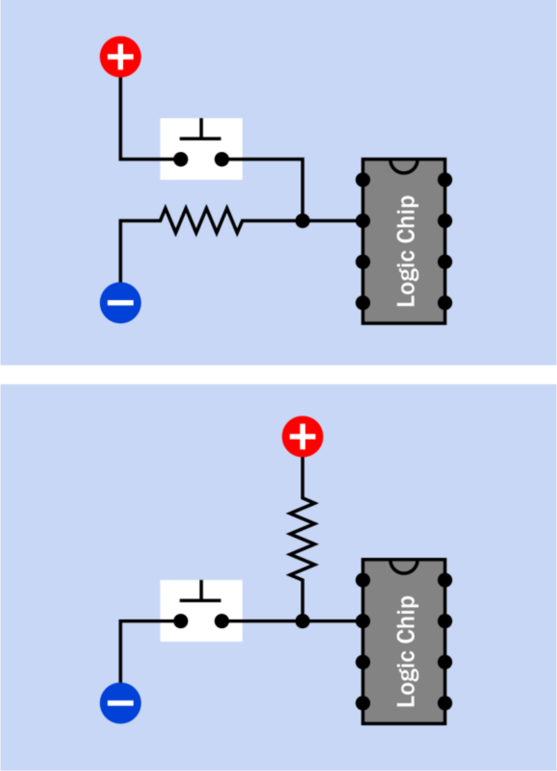

Figure 10-17. *A pulldown resistor (top) or pullup resistor (bottom) prevents an input pin on a logic chip or microcontroller from "floating" in an indeterminate state when the button is not being pressed.*

Audio Tone Control

A resistor-capacitor combination can limit the high-frequency in a simple audio tone-control circuit, as shown in Figure 10-18. Beneath a signal travelling from A to B, a resistor is placed in series with a capacitor that passes high frequencies to ground. This is known as a low-pass filter.

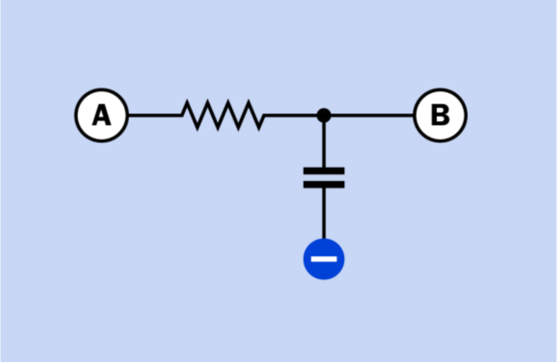

Figure 10-18. *This configuration may be used to remove high frequencies from an audio signal. It is known as a low-pass filter because low frequencies are passed from A to B.*

RC Network

A resistor will adjust the charge/discharge time when placed in series with a capacitor, as in Figure 10-19. When the switch closes, the resistor limits the rate at which the capacitor will charge itself from the power supply. Because a capacitor has an ideally infinite resistance to DC current, the voltage measured at point A will rise until it is close to the supply voltage. This is often referred to as an *RC* (*resistor-capacitor*) network and is discussed in greater detail in the **capacitor** section of this encyclopedia.

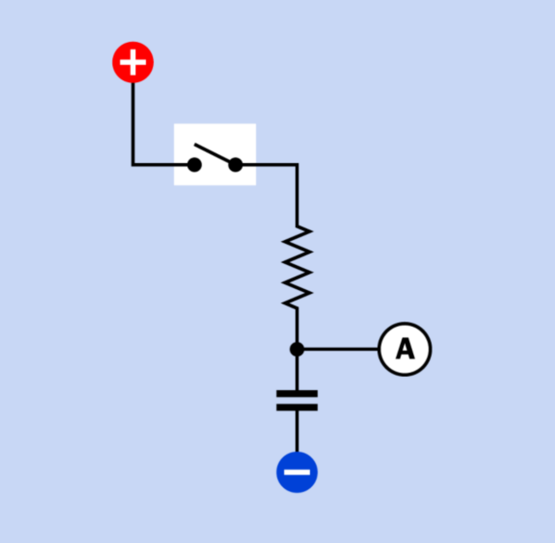

Figure 10-19. *In an RC (resistor-capacitor) network, a resistor limits the rate of increase in potential of the capacitor, measured at A, when the switch is closed.*

Voltage Divider

Two resistors may be used to create a voltage divider (see Figure 10-20). If V_{in} is the supply voltage, the output voltage, V_{out}, measured at point A, is found by the formula:

$$V_{out} = V_{in} * (R2 / (R1 + R2))$$

In reality, the actual value of V_{out} is likely to be affected by how heavily the output is loaded.

If the output node has a high impedance, such as the input to a logic chip or comparator, it will be more susceptible to electrical noise, and lower-value resistors may be needed in the voltage divider to maintain a higher current flow and maintain stability in the attached device.

Resistors in Series

If resistors in series have values R1, R2, R3 . . . the total resistance, R, is found by summing the individual resistances:

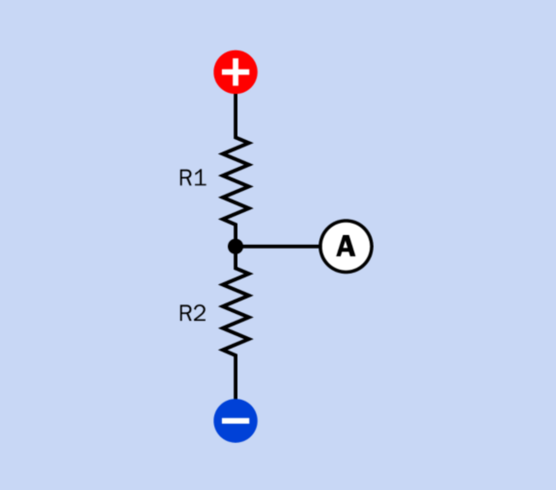

Figure 10-20. *In a DC circuit, a pair of resistors may be placed in series to function as a voltage divider. The voltage measured at A will be lower than the supply voltage, but above ground potential.*

```
R = R1 + R2 + R3. . .
```

The current through each of the resistors will be the same, whereas the voltage across each of them will vary proportionately with its resistance. If the supply voltage across the series of resistors is VS, and the total of all the resistor values is RT, and the resistance of one resistor is R1, the voltage across that resistor, V1, will be given by the formula:

```
V1 = VS * (R1 / RT)
```

Resistors in Parallel

Where two or more resistors (R1, R2, R3 . . .) are wired in parallel, their total resistance, R, is found from the formula:

```
1/R = ( 1/R1 ) + ( 1/R2 ) + ( 1/R3 ). . .
```

Suppose that R1, R2, R3 . . . all have the same individual resistance, represented by RI, and the number of resistors is N. Their total resistance, RT, when wired in parallel, will be:

```
RT = RI / N
```

If each resistor has an equal resistance and also has an equal individual rating in watts (represented by WI), the total wattage (WT) that they can handle when wired in parallel to share the power will be:

```
WT = WI * N
```

Therefore, if an application requires high-wattage resistors, multiple lower-wattage, higher-value resistors may be substituted if they are wired in parallel—and may even be cheaper than a single high-wattage wire-wound resistor. For example, if a 5W, 50Ω resistor is specified, 10 resistors can be substituted, each rated at 0.5W and 500Ω. Bear in mind that if they are tightly bundled, this will interfere with heat dissipation.

What Can Go Wrong

Heat

Resistors are probably the most robust of all electronic components, with high reliability and a long life. It is difficult to damage a resistor by overheating it with a soldering iron.

The wattage rating of a resistor does not necessarily mean that it should be used to dissipate that amount of power on a constant basis. Small resistors (1/4 watt or less) can overheat just as easily as big ones. Generally speaking, it is safe practice not to exceed 75% of a resistor's power rating on a constant basis.

Overheating is predictably more of a problem for power resistors, where provision must be made for heat dissipation. Issues such as component crowding should be considered when deciding how big a heat sink to use and how much ventilation. Some power resistors may function reliably at temperatures as high as 250 degrees Centigrade, but components near them are likely to be less tolerant and plastic enclosures may soften or melt.

Noise

The electrical *noise* introduced by a resistor in a circuit will vary according to the composition of the resistor, but for any given component, it will be proportional to voltage and current. Low-noise circuits (such as those at the input stage of a high-gain amplifier) should use low-wattage resistors at a low voltage where possible.

Inductance

The coiled wire of a wire-wound resistor will be significantly inductive at low frequencies. This is known as *parasitic inductance*. It will also have a resonant frequency. This type of resistor is unsuitable for applications where frequency exceeds 50KHz.

Inaccuracy

When using resistors with 10% tolerance, imprecise values may cause greater problems in some applications than in others. In a voltage divider, for instance, if one resistor happens to be at the high end of its tolerance range while the other happens to be at the low end, the voltage obtained at the intersection of the resistors will vary from its expected value. Using the schematic shown in Figure 10-20, if R1 is rated for 1K and R2 is rated for 5K, and the power supply is rated at 12VDC, the voltage at point A should be:

```
V = 12 * (( 5 / (5 + 1)) = 10
```

However, if R1 has an actual value of 1.1K and R2 has an actual value of 4.5K, the actual voltage obtained at point A will be:

```
V = 12 * (( 4.5 / (4.5 + 1.1)) = 9.6
```

If the resistors are at opposite ends of their respective tolerance ranges, so that R1 has an actual value of 900Ω while the lower resistor has an actual value of 5.5K, the actual voltage obtained will be:

```
V = 12 * (( 5.5 / (5.5 + 0.9)) = 10.3
```

The situation becomes worse if the two resistors are chosen to be of equal value, to provide half of the supply voltage (6 volts, in this example) at their intersection. If two 5K resistors are used, and the upper one is actually 4.5K while the lower one is 5.5K, the actual voltage will be:

```
V = 12 * (( 5.5 / (4.5 + 5.5)) = 6.6
```

Whether this variation is significant will depend on the particular circuit in which the voltage divider is being used.

Common through-hole resistors may occasionally turn out to have values that are outside their specified tolerance range, as a result of poor manufacturing processes. Checking each resistor with a meter before placing it in a circuit should be a standard procedure.

When measuring the voltage drop introduced by a resistor in an active circuit, the meter has its own internal resistance that will take a proportion of the current. This is known as *meter loading* and will result in an artificially low reading for the potential difference between the ends of a resistor. This problem becomes significant only when dealing with resistors that have a high value (such as 1M), comparable with the internal resistance of the meter (likely to be 10M or more).

Wrong Values

When resistors are sorted into small bins by the user, errors may be made, and different values may be mixed together. This is another reason for checking the values of components before using them. Identification errors may be nontrivial and easily overlooked: the visible difference between a 1 megohm resistor and a 100Ω resistor is just one thin color band.

potentiometer

11

Also known as a *variable resistor*; may be substituted for a *rheostat*.

OTHER RELATED COMPONENTS

- **rotational encoder** (see Chapter 8)
- **resistor** (see Chapter 10)

What It Does

When a voltage is applied across a potentiometer, it can deliver a variable fraction of that voltage. It is often used to adjust sensitivity, balance, input, or output, especially in audio equipment and sensors such as motion detectors.

A potentiometer can also be used to insert a variable resistance in a circuit, in which case it should really be referred to as a *variable resistor,* although most people will still call it a potentiometer.

It can be used to adjust the power supplied to a circuit, in which case it is properly known as a *rheostat*, although this term is becoming obsolete. Massive rheostats were once used for purposes such as dimming theatrical lighting, but solid-state components have taken their place in most high-wattage applications.

A full-size, classic-style potentiometer is shown in Figure 11-1.

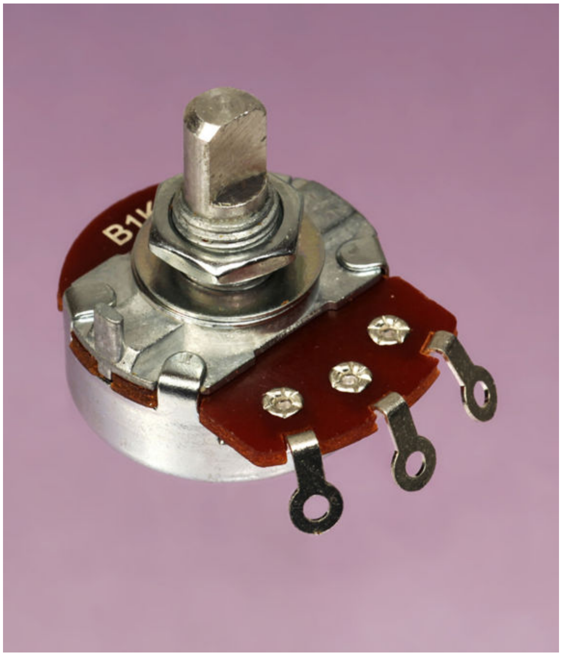

Figure 11-1. *A generic or classic-style potentiometer, approximately one inch in diameter.*

Schematic symbols for a potentiometer and other associated components are shown in Figure 11-2, with American versions on the left and European versions on the right in each case. The symbols for a potentiometer are at the top. The correct symbols for a variable resistor or rheostat are shown at center, although a potentiometer symbol may often be used instead. A preset variable resistor is shown at the bottom, often referred to as a *trimmer* or *Trimpot*. In these examples, each has an arbitrary rated resistance of 4,700Ω. Note the European substitution of K for a decimal point.

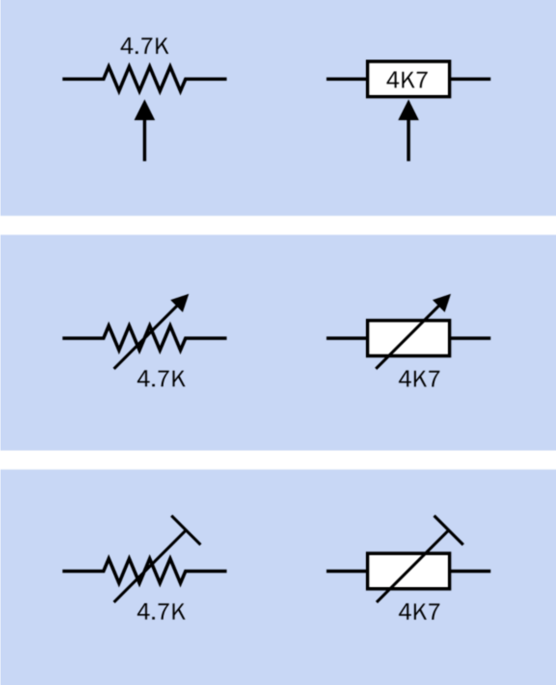

Figure 11-2. *American (left) and European (right) symbols for a potentiometer, a rheostat, and a trimmer potentiometer, reading from top to bottom. The 4.7K value was chosen arbitrarily.*

How It Works

A potentiometer has three terminals. The outer pair connect with the opposite ends of an internal resistive element, such as a strip of conductive plastic, sometimes known as the *track*. The third center terminal connects internally with a contact known as the *wiper* (or rarely, the *pick-off*), which touches the strip and can be moved from one end of it to the other by turning a shaft or screw, or by moving a slider.

If an electrical potential is applied between opposite ends of the resistive element, the voltage "picked off" by the wiper will vary as it moves. In this mode, the potentiometer works as a resistive voltage divider. For example, in a potentiometer with a linear taper (see "Variants," coming up), if you attach the negative side of a 12V battery to the right-hand end terminal and the positive side to the left-hand end terminal, you will find an 8V potential at the center terminal when the potentiometer has rotated clockwise through one-third of its range. In Figure 11-3, the base of the shaft (shown in black) is attached to an arm (shown in green) that moves a wiper (orange) along a resistive element (brown). The voltages shown assume that the resistive element has a linear taper and will vary slightly depending on wire resistance and other factors.

Because a potentiometer imposes a voltage reduction, it also reduces current flowing through it, and therefore creates waste heat which must be dissipated. In an application such as an audio circuit, small currents and low voltages generate negligible heat. If a potentiometer is used for heavier applications, it must be appropriately rated to handle the wattage and must be vented to allow heat to disperse.

To use a potentiometer as a *variable resistor* or *rheostat*, one of its end terminals may be tied to the center terminal. If the unused end terminal is left unconnected, this raises the risk of picking up stray voltages or "noise" in sensitive circuits. In Figure 11-4, a potentiometer is shown adjusting a series resistance for an *LED* for demonstration purposes. More typically, a trimmer would be used in this kind of application, since a user is unlikely to need to reset it.

Variants

Linear and Log Taper

If the resistive element in a potentiometer is of constant width and thickness, the electrical potential at the wiper will change in ratio with the rotation of the wiper and shaft (or with movement of a slider). This type of potentiometer is said to have a linear taper even though its element does not actually taper.

For audio applications, because human hearing responds nonlinearly to sound pressure, a potentiometer that has a linear taper may seem to have a very slow action at one end of its scale and an abrupt effect at the other. This problem used to be solved with a non-uniform or tapered re-

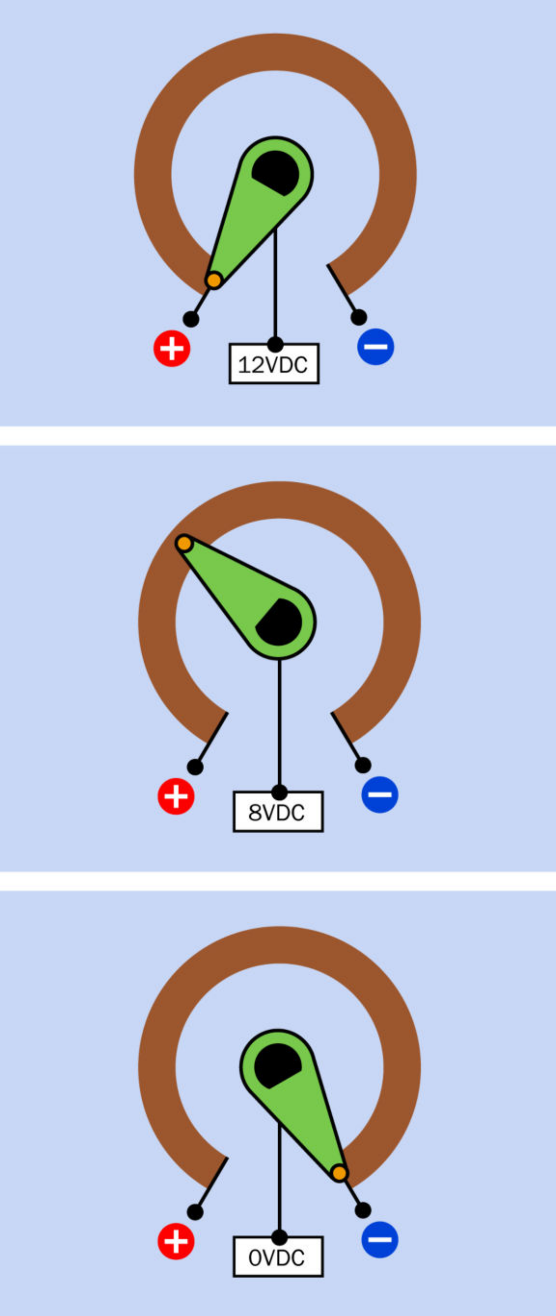

Figure 11-3. *Inside a potentiometer. See text for details.*

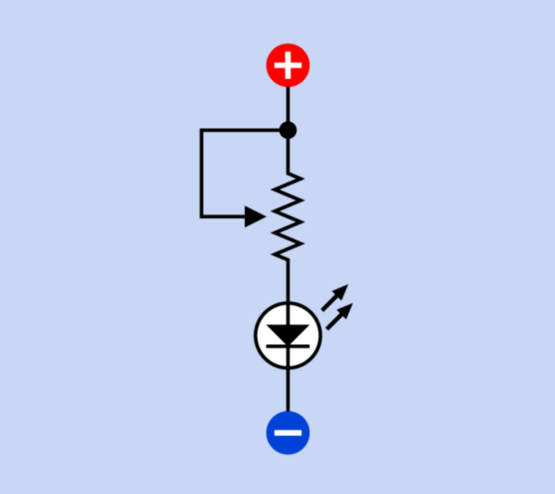

Figure 11-4. *A potentiometer can be used to adjust a series resistance, as shown in this schematic. Tying the wiper to one of the end terminals reduces the risk of picking up electrical noise.*

sistive element. More recently, a combination of resistive elements has been used as a cheaper option. Such a potentiometer is said to have an audio taper or a log taper (since the resistance may vary as a logarithm of the angle of rotation). A reverse audio taper or antilog taper varies in the opposite direction, but this type has become very uncommon.

Classic-style Potentiometer

This consists of a sealed circular can, usually between 0.5″ and 1″ in diameter, containing a resistive strip that is shaped as a segment of a circle. A typical example is shown in Figure 11-1, although miniaturized versions have become more common. A shaft mounted on the can turns the internal wiper that presses against the strip. For panel-mount applications, a threaded bushing at the base of the shaft is inserted through a hole in the front panel of the electronics enclosure, and a nut is tightened on the bushing to hold the potentiometer in place. Often there is also a small offset index pin that, when paired with a corresponding front panel hole, will keep the pot from spinning freely.

Many modern potentiometers are miniaturized, and may be packaged in a box-shaped plastic enclosure rather than a circular can. Their power ratings are likely to be lower, but their principle of operation is unchanged. Two variants are shown in Figure 11-5.

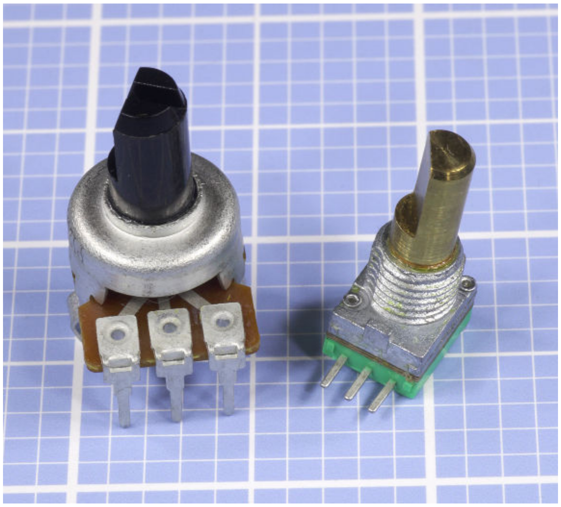

Figure 11-5. *Two modern miniaturized potentiometers. At left: 5K. At right: 10K. Both are rated to dissipate up to 50mW.*

The three terminals on the outside of a potentiometer may be solder lugs, screw terminals, or pins for direct mounting on a circuit board. The pins may be straight or angled at 90 degrees.

The resistive element may use carbon film, plastic, cermet (a ceramic-metal mixture), or resistive wire wound around an insulator. Carbon-film potentiometers are generally the cheapest, whereas wire-wound potentiometers are generally the most expensive.

Wire-wound potentiometers may handle more power than the other variants, but as the wiper makes a transition from one turn of the internal wire element to the next, the output will tend to change in discrete steps instead of varying more smoothly.

In a potentiometer with detents, typically a spring-loaded lever in contact with notched internal wheel causes the shaft to turn in discrete steps that create a stepped output even if the resistive element is continuous.

The shaft may be made of metal or plastic, with its length and width varying from one component to another. A control knob can be fitted to the end of the shaft. Some control knobs are push-on, others have a set screw to secure them. Shafts may be splined and split, or round and smooth, or round with a flat surface that matches the shape of a socket in a control knob and reduces the risk of a knob becoming loose and turning freely. Some shafts have a slotted tip to enable screwdriver adjustment.

Some shaft options for full-size potentiometers are shown in Figure 11-6.

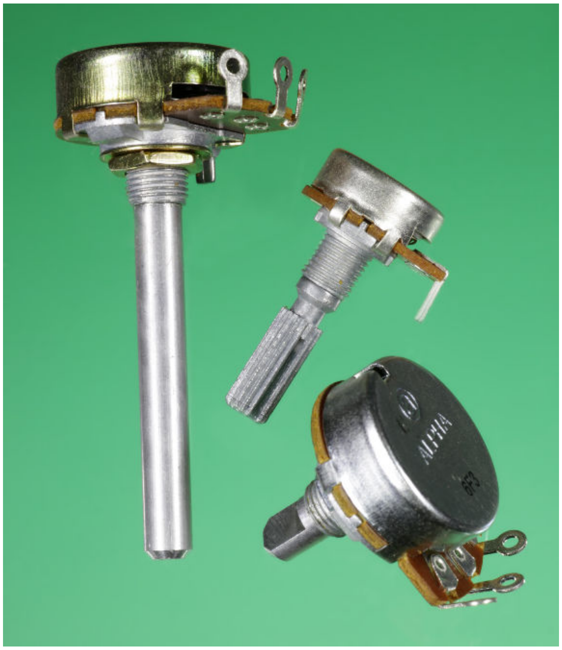

Figure 11-6. *Three shaft options for potentiometers.*

Multiple-Turn Potentiometer

To achieve greater precision, a track inside a potentiometer may be manufactured in the form of a helix, allowing the wiper to make multiple turns on its journey from one end of the track to the other. Such multiple-turn potentiometers typically allow 3, 5, or 10 turns to move the wiper from end to end. Other multiple-turn potentiometers may use a screw thread that advances a wiper along a linear or circular track. The latter is comparable with a trimmer where multiple turns of a screwdriver are used to rotate a worm gear that rotates a wiper between opposite ends of a circular track.

Ganged Potentiometer

Two (or rarely, more) potentiometers can be stacked or combined so that their resistive elements and wipers share the same shaft but can use different voltages or have different taper. Each resistance-wiper assembly is known as a *cup*, and the potentiometers are said to be *ganged*.

Flat ganged potentiometers combine two resistive elements in one enclosure. Some *dual ganged* potentiometers are concentric, meaning that the pots are controlled separately by two shafts, one inside the other. Suitable concentric knobs must be used. You are unlikely to find these potentiometers sold as components in limited quantities.

Switched Potentiometer

In this variant, when the shaft is turned clockwise from an initial position that is fully counterclockwise, it flips an internal switch connected to external terminals. This can be used to power-up associated components (for example, an audio amplifier). Alternatively, a switch inside a potentiometer may be configured so that it is activated by pulling or pushing the shaft.

Slider Potentiometer

Also known as a *slide potentiometer*. This uses a straight resistive strip and a wiper that is moved to and fro linearly by a tab or lug fitted with a plastic knob or finger-grip. Sliders are still found on some audio equipment. The principle of operation, and the number of terminals, are identical to the classic-style potentiometer. Sliders typically have solder tabs or PC pins. In Figure 11-7, the large one is about 3.5″ long, designed for mounting behind a panel that has a slot to allow the sliding lug to poke through. Threaded holes at either end will accept screws to fix the slider behind the panel. A removable plastic finger-grip (sold separately, in a variety of styles) has been pushed into place. Solder tabs underneath the slider are hidden in this photo. The smaller slider is designed for through-hole mounting on a circuit board.

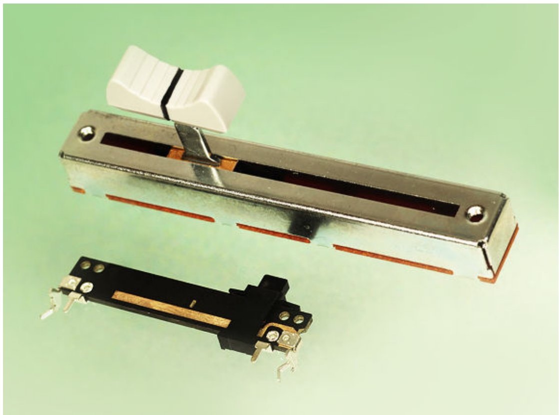

Figure 11-7. *Slider potentiometers.*

Trimmer Potentiometer

Often referred to as Trimpots, this is actually a proprietary brand name of Bourns. They are usually mounted directly on circuit boards to allow fine adjustment or trimming during manufacturing and testing to compensate for variations in other components. Trimmers may be single-turn or multi-turn, the latter containing a worm gear that engages with another gear to which the wiper is attached. Trimmers always have linear taper. They may be designed for screwdriver adjustment or may have a small knurled shaft, a thumb wheel, or a knob. They are not usually accessible by the end user of the equipment, and their setting may be sealed or fixed when the equipment is assembled. In Figure 11-8, the beige Spectrol trimmer is a single-turn design, whereas the blue trimmer is multi-turn. A worm gear inside the package, beneath the screw head, engages with an interior gear wheel that rotates the wiper.

In Figure 11-9, a 2K trimmer potentiometer has a knurled dial attached to allow easy finger adjustment, although the dial also contains a slot for a flat-blade screwdriver.

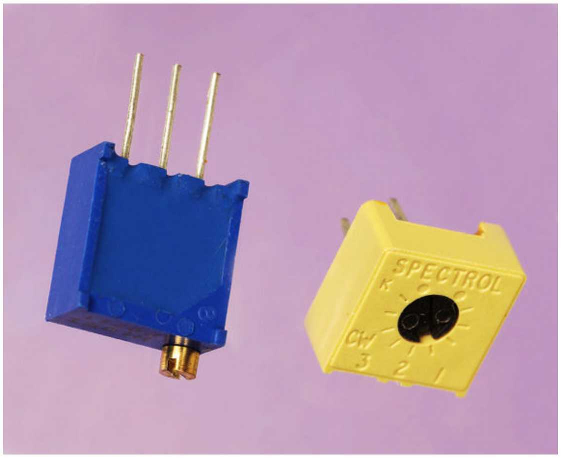

Figure 11-8. *Like most trimmers, these are designed for through-hole mounting on a circuit board.*

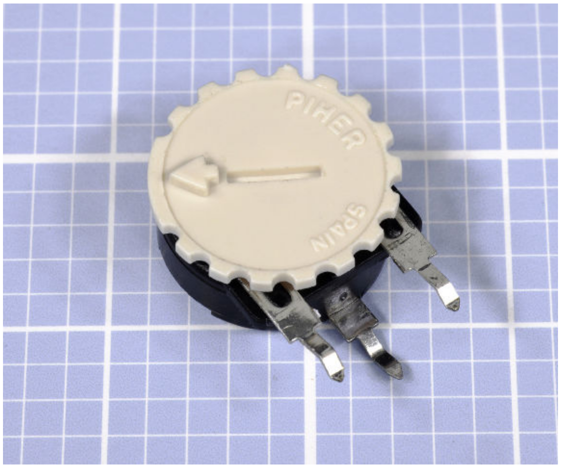

Figure 11-9. *A trimmer potentiometer with a knurled dial to facilitate finger adjustment.*

How to Use it

The classic-style potentiometer was once used universally to control volume, bass, and treble on audio equipment but has been replaced increasingly by digital input devices such as **tactile switches** (see "Tactile Switch" (page 34)) or **rotational encoders** (see Chapter 8), which are more reliable and may be cheaper, especially when assembly costs are considered.

Potentiometers are widely used in lamp dimmers and on cooking stoves (see Figure 11-10). In these applications, a solid-state switching device such as a **triac** (described in Volume 2) does the actual work of moderating the power to the lamp or the stove by interrupting it very rapidly. The potentiometer adjusts the duty cycle of the power interruptions. This system wastes far less power than if the potentiometer controlled the lighting or heating element directly as a rheostat. Since less power is involved, the potentiometer can be small and cheap, and will not generate significant heat.

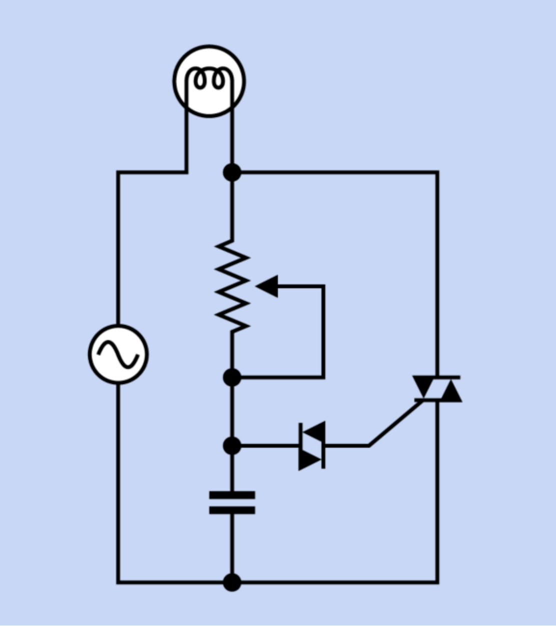

Figure 11-10. *Typical usage of a potentiometer in conjunction with a diac, triac, and capacitor to control the brightness of an incandescent bulb, using an AC power supply. Diacs and triacs are discussed in Volume 2.*

Because true logarithmic potentiometers have become decreasingly common, a linear potentiometer in conjunction with a fixed resistor can be used as a substitute, to control audio input. See Figure 11-11.

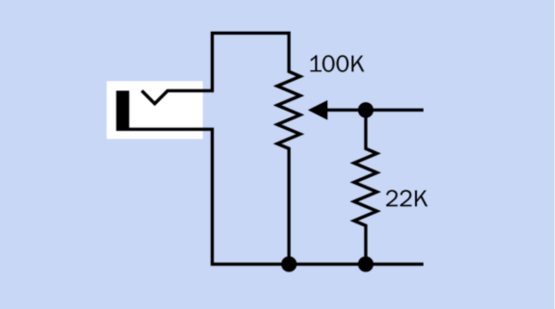

Figure 11-11. *In this circuit, a 100K linear potentiometer is used in conjunction with a 22K resistor to create an approximately logarithmic volume control for an audio system with input coming from a mono jack socket at left.*

A potentiometer may be used to match a sensor or analog input device to an analog-digital converter, or it can calibrate a device such as a temperature or motion sensor.

What Can Go Wrong

Wear and Tear

Since classic-style potentiometers are electromechanical devices, their performance will deteriorate as one part rubs against another. The long open slot of a slider potentiometer makes it especially vulnerable to contamination with dirt, water, or grease. Contact-cleaner solvent, lubricant-carrying sprays, or pressurized "duster" gas may be squirted into a potentiometer to try to extend its life. Carbon-film potentiometers are the least durable and in audio applications will eventually create a "scratchy" sound when they are turned, as the resistive element deteriorates.

If the wiper deteriorates to the point where it no longer makes electrical contact with the track, and if the potentiometer is being used as a variable resistor, two failure modes are possible, shown in Figure 11-12. Clearly the right-hand schematic is a better outcome. This is an argument for always tying the wiper to the "unused" end of the track.

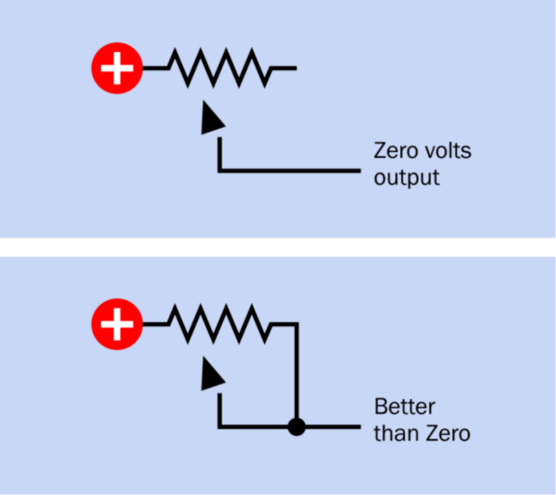

Figure 11-12. *If the wiper of a potentiometer breaks (indicated by the loose arrow head) as a result of wear and tear, and the potentiometer is being used as a variable resistor, the voltage from it will drop to zero (top schematic) unless the wiper has been tied to one end of the track (bottom schematic).*

If you are designing a circuit board that will go through a production process, temperature variations during wave soldering, and subsequent washing to remove flux residues, create hostile conditions for potentiometers, especially sliders where the internal parts are easily contaminated. It will be safer to hand-mount potentiometers after the automated process.

Knobs that Don't Fit

Control knobs are almost always sold separately from potentiometers. Make sure the shaft of the potentiometer (which may be round, round-with-flat, or knurled) matches the knob of your choice. Note that some shaft diameters are expressed in inches, while others are metric.

Nuts that Get Lost

For panel-mounted potentiometers, a nut that fits the thread on the bushing is almost always included with the potentiometer; an additional nut and lock washer may also be supplied. Because there is no standardization of threads on potentiometers, if you lose a nut, you may have some difficulty finding an exact replacement.

A Shaft that Isn't Long Enough

When choosing a shaft length, if in doubt, buy a potentiometer with a long shaft that you can cut to the desired length.

Sliders with No Finger Grip

Slider potentiometers are often sold without a knob or plastic finger-grip, which must be ordered separately and may be available in different styles. The finger-grip usually push-fits onto the metal or plastic tab or lug that moves the slider to and fro.

Too Big to Fit

Check the manufacturer's datasheet if you need to know the physical size of the potentiometer. Photographs may be misleading, as a traditional-style potentiometer that is 0.5" in diameter looks much the same as one that is 1" in diameter. High-wattage potentiometers will be more costly and physically large (2 to 3 inches in diameter). See Figure 11-13.

Overheating

Be sure to leave sufficient air space around a high-wattage potentiometer. Carefully calculate the maximum voltage drop and current that you may be using, and choose a component that is appropriately rated. Note that if you use the potentiometer as a rheostat, it will have to handle more current when its wiper moves to reduce its resistance. For example, if 12VDC are applied through a 10-ohm rheostat to a component that has a resistance of 20 ohms, current in the circuit will vary from 0.4 amps to 0.6 amps depending on the position of the rheostat. At its maximum setting, the rheostat will impose a 4V voltage drop and will therefore dissipate 1.6 watts from the full length of its resistive element. If the rheostat is reset to impose only a 4-ohm resistance, the voltage drop that it imposes will be 2V, the current in the circuit will be 0.5 amps, and the rheostat will therefore dissipate 1 watt from 4/10ths of the length of its resistive element. A wire-wound potentiometer will be better able to handle high dissipation from a short segment of its element than other types of rheostat. Add a fixed resistor in series with a rheostat if necessary to impose a limit on the current.

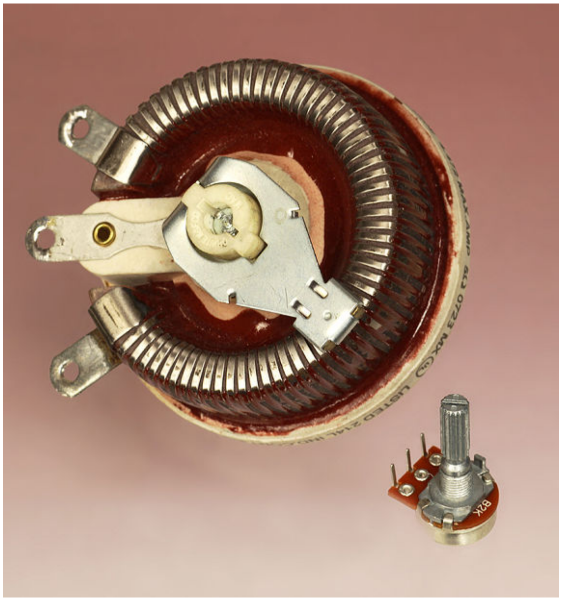

Figure 11-13. *The large potentiometer is approximately 3" in diameter, rated at 5 ohms, and able to handle more than 4 amps. The small potentiometer is 5/8" diameter, rated at 2K and 1/4 watt, with pins designed for through-hole insertion in a circuit board, and a grooved shaft that accepts a push-on knob. Despite the disparity in size, the principle of operation and the basic features are identical.*

When using a trimmer potentiometer, limit the current through the wiper to 100mA as an absolute maximum value.

The Wrong Taper

When buying a potentiometer, remember to check the specification to find out whether it has linear or audio/log taper. If necessary, attach a meter, with the potentiometer set to its center position, to verify which kind of taper you have. While holding the meter probes in place, rotate the potentiometer shaft to determine which way an audio/log taper is oriented.

capacitor

12

Quite often referred to as a *cap*. Formerly known (primarily in the United Kingdom) as a *condenser*, but that term has become obsolete.

OTHER RELATED COMPONENTS

- **variable capacitor** (See Chapter 13)
- **battery** (See Chapter 2)

What It Does

A capacitor connected across a DC power source will accumulate a charge, which then persists after the source is disconnected. In this way, the capacitor stores (and can then discharge) energy like a small rechargeable battery. The charge/discharge rate is extremely fast but can be limited by a series resistor, which enables the capacitor to be used as a timing component in many electronic circuits.

A capacitor can also be used to block DC current while it passes pulses, or electrical "noise," or alternating current, or audio signals, or other wave forms. This capability enables it to smooth the output voltage provided by **power supplies**; to remove spikes from signals that would otherwise tend to cause spurious triggering of components in digital circuits; to adjust the frequency response of an audio circuit; or to couple separate components or circuit elements that must be protected from transmission of DC current.

Schematic symbols for capacitors are shown in Figure 12-1. At top-left is a nonpolarized capacitor, while the other two indicate that a polarized capacitor must be used, and must be oriented as shown. The variant at the bottom is most commonly used in Europe. Confusingly, the nonpolarized symbol may also be used to identify a polarized capacitor, if a + sign is added. The polarized symbols are sometimes printed without + signs, but the symbols still indicate that polarity must be observed.

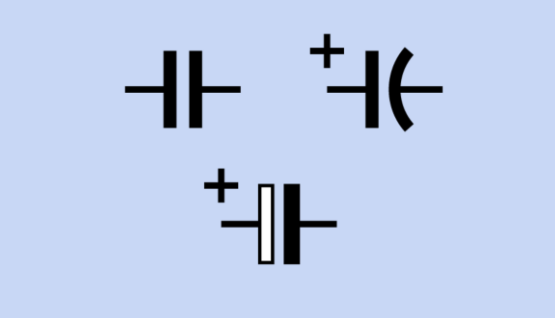

Figure 12-1. *Schematic symbols for polarized and nonpolarized capacitors. See text for details.*

How It Works

In its simplest form, a capacitor consists of two plates, each with a lead attached to it for connection with a DC power source. The plates are separated by a thin, insulating layer known as the *dielectric*, which is usually a solid or a paste but may be liquid, gel, gaseous, or vacuum.

The plates in most capacitors are made from thin metal film or metallized plastic film. To minimize the size of the component, the film may be rolled up to form a compact cylindrical package, or multiple flat sections may be interleaved.

Electrons from the power source will migrate onto the plate attached to the negative side of the source, and will tend to repel electrons from the other plate. This may be thought of as creating *electron holes* in the other plate or as attracting *positive charges*, as shown in Figure 12-2. When the capacitor is disconnected from the power source, the opposite charges on its plates will persist in equilibrium as a result of their mutual attraction, although the voltage will gradually dissipate as a result of *leakage,* either through the dielectric or via other pathways.

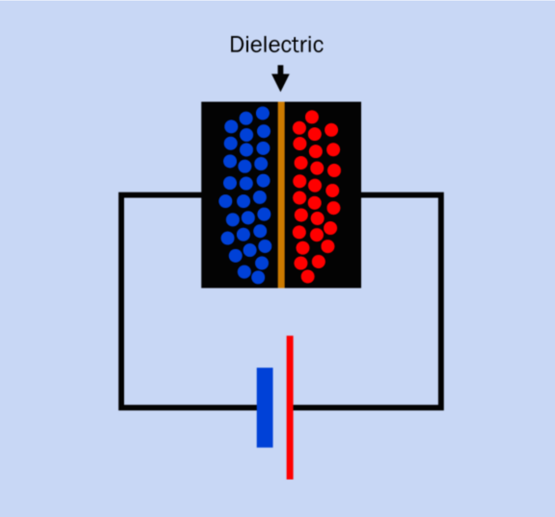

Figure 12-2. *Because the plates of a capacitor are electrically conductive, they will become populated with positive and negative charges when connected with a DC power source. As opposite charges attract each other, they will tend to congregate on either side of the dielectric, which is an insulating layer. The battery symbol is shown here colored for clarity.*

When a **resistor** is placed across the leads of a charged capacitor, the capacitor will discharge itself through the resistor at a rate limited by the resistance. Conversely, if a capacitor is charged through a resistor, the resistor will limit the charging rate.

A resistor in series with a capacitor is known as an *RC network* (Resistor-Capacitor network). In Figure 12-3, an RC circuit is shown with a SPDT switch that charges or discharges the capacitor via a series resistor. The voltage at point A increases nonlinearly (relative to the negative side of the power supply) while the capacitor is charging, and decreases nonlinearly while the capacitor is discharging, as suggested by the graphs. At any moment, the time that the capacitor takes to acquire 63% of the difference between its current charge and the voltage being supplied to it is known as the *time constant* for the circuit. See "The Time Constant" (page 105) for additional information.

When a capacitor is connected across an AC voltage source, each surge of electrons to one plate induces an equal and opposite positive surge to the other plate, and when polarity of the power supply reverses, the charges on the plates switch places. These surges may make it seem that the capacitor is conducting AC current, even though the dielectric that separates the plates is an insulator. See Figure 12-4. Often a capacitor is said to "pass" AC, even though this is not really happening. For convenience, and because the concept is widely established, this encyclopedia refers to capacitors as "passing" AC.

Depending on the size of the capacitor, it will block some AC frequencies while passing others. Generally speaking, a smaller capacitor will pass high frequencies relatively efficiently, as each little surge of current fills each plate. However, the situation is complicated by the *inductive reactance* (which creates the *effective series resistance*) of a capacitor, as discussed below. See "Alternating Current and Capacitive Reactance" (page 106).

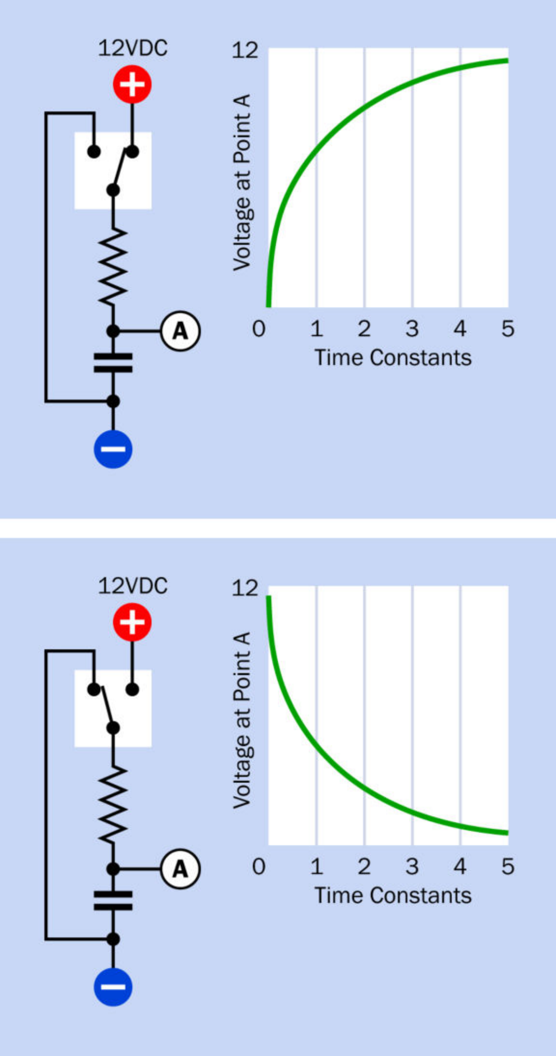

Figure 12-3. *An RC (Resistor-Capacitor) network with a switch to control charge and discharge of a capacitor. At top, the curve gives an approximate idea of the charging behavior of the capacitor. At bottom, the curve illustrates its discharging behavior.*

Variants

Format

The three most common packages for capacitors are cylindrical, disc, and rectangular tablet.

A *cylindrical capacitor* may have *axial leads* (a wire attached to each end) or *radial leads* (both wires emerging from one end). Radial capacitors are more widely used as they allow easy insertion into a circuit board. The capacitor is usually packaged in a small aluminum can, closed at one end, capped with an insulating disc at the other end, and wrapped in a thin layer of insulating plastic. Some samples are shown in Figure 12-5 and Figure 12-6.

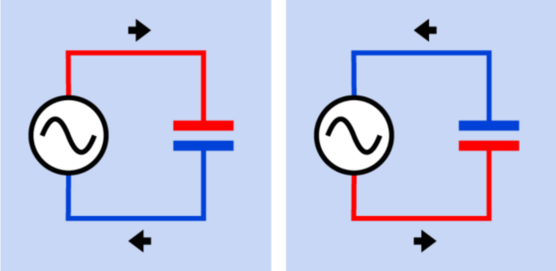

Figure 12-4. *In the left diagram, a source of alternating current charges the upper plate of a capacitor positively and the lower plate negatively. This process entails a flow of conventional current shown by the arrows. A moment later, when the AC current flow reverses, the flow also reverses, creating the impression that the capacitor "passes" AC current.*

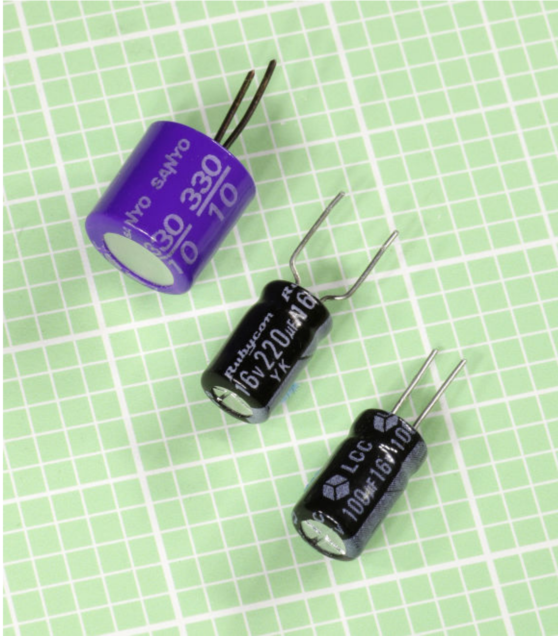

Figure 12-5. *Cylindrical capacitors with radial leads. All are electrolytic.*

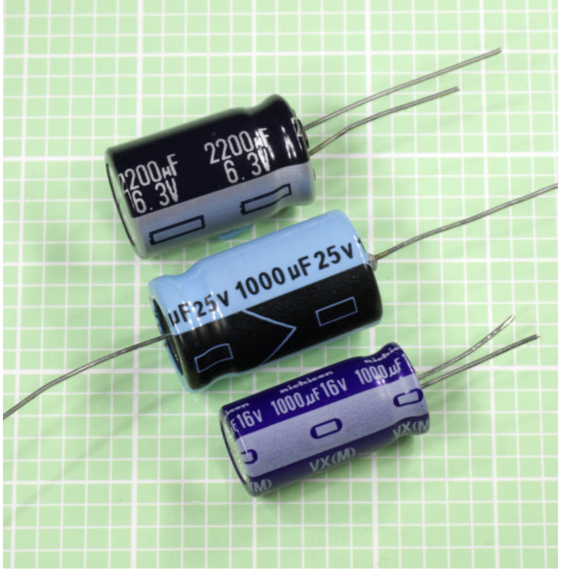

Figure 12-6. *Cylindrical capacitors with radial leads (top and bottom) and axial leads (center). All are electrolytic.*

A *disc capacitor* (sometimes referred to as a *button capacitor*) is usually encased in an insulating ceramic compound, and has radial leads. Modern small-value ceramic capacitors are more likely to be dipped in epoxy, or to be square tablets. Some samples are shown in Figure 12-7.

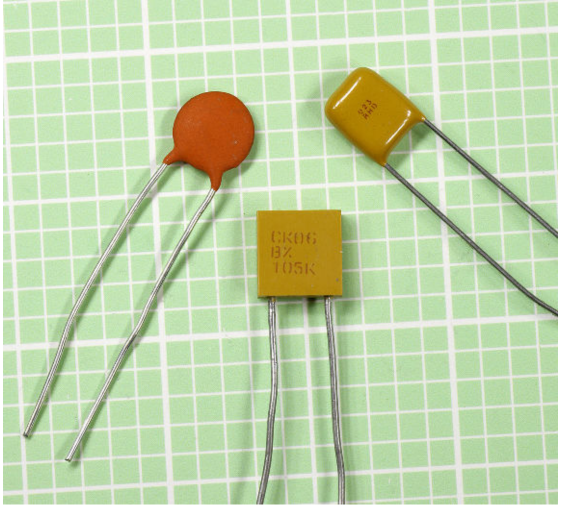

Figure 12-7. *Generic ceramic capacitors. Left: rated for 0.1μF at 50V. Center: 1μF at 50V. Right: 1μF at 50V.*

A *surface-mount capacitor* is square or rectangular, usually a few millimeters in each dimension, with two conductive pads or contacts at opposite ends. It appears almost identical to a surface-mount resistor. Larger-value capacitors are inevitably bigger but can still be designed for surface-mount applications. See Figure 12-8.

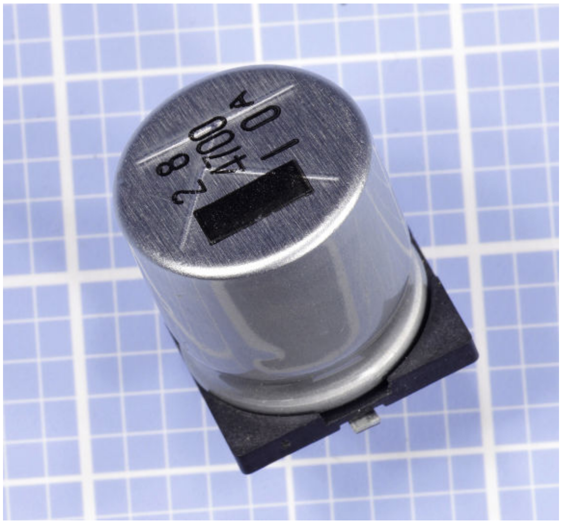

Figure 12-8. *Most surface-mount capacitors are as tiny as other surface-mount components, but this 4,700μF electrolytic (at 10V) has a base approximately 0.6" square. A solder tab is visible at the center of the nearest edge.*

Many capacitors are nonpolarized, meaning that they are insensitive to polarity. However, electrolytic and tantalum capacitors must be connected "the right way around" to any DC voltage source. If one lead is longer than the other, it must be the "more positive" lead. A mark or band at one end of the capacitor indicates the "more negative" end. Tantalum capacitors are likely to indicate the positive lead by using a + sign on the body of the component.

An arrow printed on the side of a capacitor usually points to the "more negative" terminal. In an aluminum can with axial leads, the lead at one end will have an insulating disc around it while the other lead will be integral with the rounded end of the can. The wire at the insulated end must be "more positive" than the wire at the other end.

A *capacitor array* contains two or more capacitors that are isolated from each other internally and accessed by external contacts. They are sold in surface-mount format and also in *through-hole* chips of DIP (dual-inline package) or SIP (single-inline package) format. The internal components may be connected in one of three configurations: isolated, common-bus, or dual-ended common bus. Technically the isolated configuration should be referred to as a capacitor array, but in practice, all three configurations are usually referred to as *capacitor networks*. See Figure 12-9 and Figure 12-10.

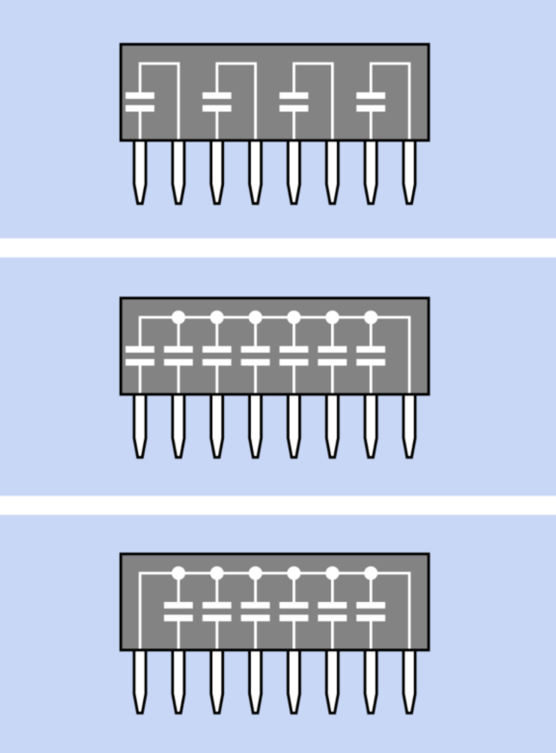

Figure 12-9. *A capacitor network most often consists of a single-inline package (SIP) chip containing multiple capacitors in one of three configurations shown here. Top: Isolated. Center: Common bus. Bottom: Dual-ended common bus. Individual capacitor values ranging from 0.001μF to 0.1μF are common.*

Capacitor networks can reduce the component count in circuits where digital logic chips require bypass capacitors. They are comparable in concept to *resistor arrays*.

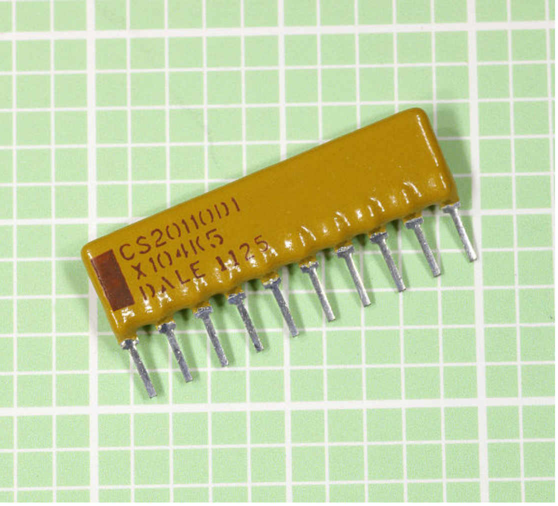

Figure 12-10. *A capacitor array in through-hole, SIP format.*

Chips containing *RC circuits* (multiple resistor-capacitor pairs) are available, although uncommon.

Principal Types

Electrolytic capacitors are relatively cheap, compact, and available in large values. These attributes have made them a popular choice in consumer electronics, especially for power supplies. The capacitive capability of an electrolytic is refreshed by periodic application of voltage. A moist paste inside the capacitor is intended to improve the dielectric performance when voltage is applied, but can dry out during a period of years. If an electrolytic is stored for 10 years or so, it may allow a short circuit between its leads when power is applied to it. The capacitors in Figure 12-5 and Figure 12-6 are all electrolytic. The capacitor in Figure 12-11 is at the high end of the scale.

A *bipolar electrolytic* is a single package containing two electrolytic capacitors in series, end-to-end, with opposed polarities, so that the combination can be used where the voltage of a signal fluctuates above and below 0VDC. See Figure 12-12 and Figure 12-13. This type of component is likely to have “BP” (bipolar) or “NP”

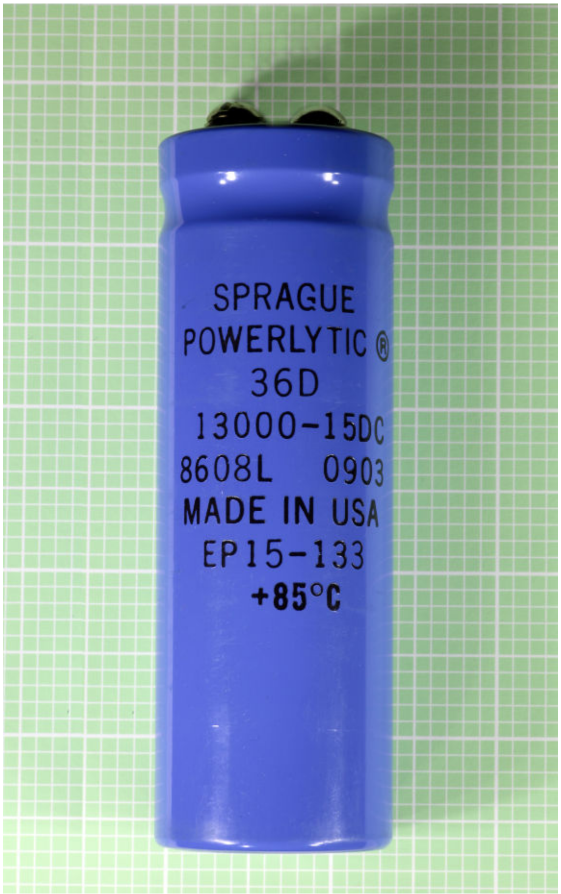

Figure 12-11. *This 13,000μF electrolytic capacitor is larger than would be required in most everyday applications.*

(nonpolarized) printed on its shell. It may be used in audio circuits where polarized capacitors are normally unsuitable, and is likely to be cheaper than non-electrolytic alternatives. However, it suffers from the same weaknesses as all electrolytics.

Tantalum capacitors are compact but relatively expensive, and can be vulnerable to voltage spikes. They are sensitive to application of the wrong polarity. Typically they are epoxy-dipped rather than mounted inside a small aluminum can like electrolytics, and consequently the electrolyte may be less likely to evaporate and dry out. In Figure 12-14, two tantalum capacitors (rated 330μF at 6.3V, left, and 100μF at 20V, right) are shown above a polyester film capacitor (rated 10μF at 100V). Surface-mount tantalum capacitors are decreasing in popularity as large-value ceramic capacitors are becoming available, with smaller dimensions and lower equivalent series resistance.

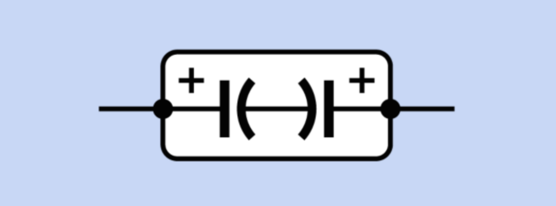

Figure 12-12. *Schematic view of the internal configuration of a bipolar electrolytic capacitor, also known as nonpolarized electrolytic capacitor. It consists of two electrolytics in series, with opposing polarities.*

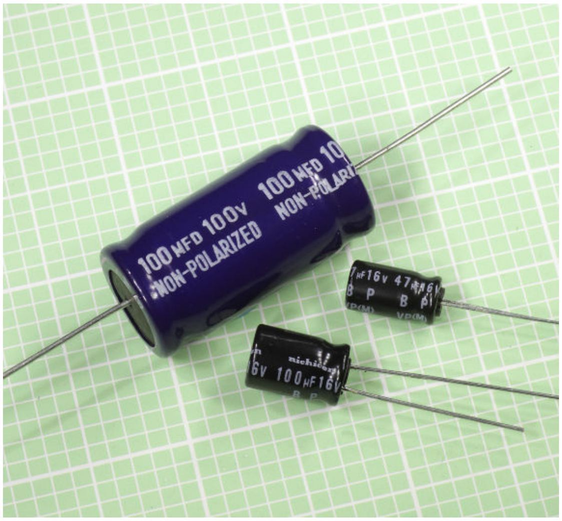

Figure 12-13. *Bipolar electrolytic capacitors. The larger size of the one at top-left is a consequence of its higher voltage rating. "BP" on the other two capacitors is an acronym for "bipolar," meaning that they have no polarity, even though one lead may be shorter than the other.*

Plastic-film capacitors are discussed in the following section.

Single-layer *ceramic* capacitors are often used for bypass, and are suitable for high-frequency or audio applications. Their value is not very stable with temperature, although the "NPO" variants are more stable. Multilayer ceramic capacitors are more compact than single-layer ceramic, and

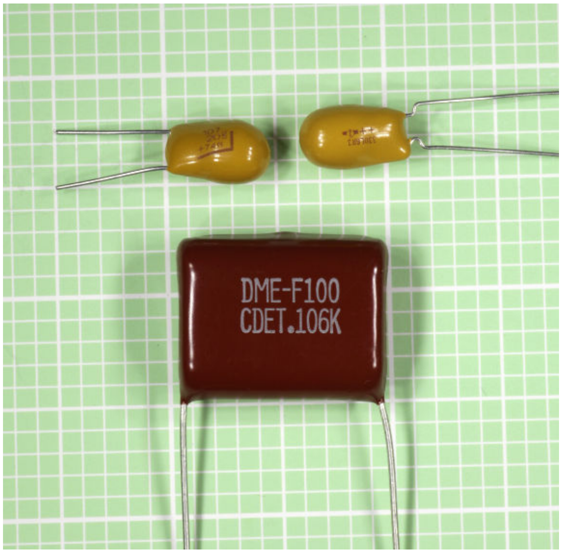

Figure 12-14. *Two tantalum capacitors are shown above a polyester film capacitor. The polarity of the tantalum capacitors is indicated by the plus signs adjacent to the longer lead, in each case. The polyester capacitor is nonpolarized.*

consequently are becoming increasingly popular. Three multilayer ceramic capacitors are shown in Figure 12-15. At bottom-right, even the largest (rated at 47μF at 16V) is only 0.2″ square.

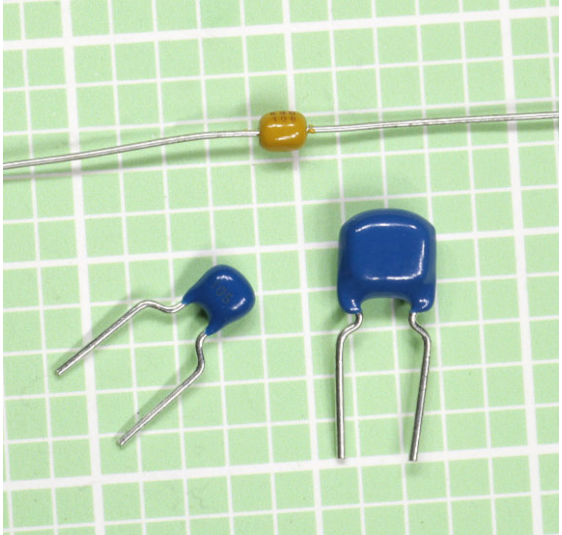

Figure 12-15. *Multilayer ceramic capacitors are extremely compact, and are nonpolarized. Top: 1,000pF (i.e. 1nF) at 100V. Bottom left: 1μF at 25V. Bottom right: 47μF at 16V.*

Dielectrics

The dielectric used in a capacitor most often consists of an electrolytic layer, a ceramic compound, a plastic film (polycarbonate, polypropylene, or polystyrene), or paper.

An *electrolytic layer* in an electrolytic capacitor traditionally consists of paper soaked in an electrolyte. It is interleaved with a thin film of aluminum on which is deposited a layer of aluminum oxide. The layers are rolled up to create a cylindrical component. The functioning dielectric is created when voltage is applied.

Polyester

This is the most common type of plastic film, with the highest dielectric constant, enabling highest capacitance per unit volume. Widely used in DC applications, but the rolled layers create parasitic inductance. Often used in decoupling, coupling, and bypass, but not so suitable for situations requiring stability and low leakage. May not be suitable for high current.

Polycarbonate

Thermally very stable, often specified for filters and timing circuits that require a fixed frequency. An excellent type of capacitor, compatible for mil-spec applications, but expensive.

Mylar, Polyester, and other plastic-film types are often used in audio circuits, where their voltage limitation (typically less than 100VDC) is not a problem, and their nonpolarized attribute is an advantage.

Polypropylene

Vulnerable to heat (a maximum of 85 degrees Centigrade is common), and less thermally stable than polycarbonate. A very low power dissipation factor allows it to handle higher power at higher frequencies. Available with tolerances down to 1%. These capacitors are a popular choice in crossover networks for loudspeaker combinations, and are used in switching power supplies. They tend to be physically larger than other capacitors using film dielectric.

Values

Farads

The electrical storage capacity of a capacitor is measured in *farads,* universally represented by the letter F. A capacitor that can be charged with a potential difference between its plates of 1 volt, in a time of 1 second, during which it draws 1 amp, has a capacitance of 1 farad.

Because the farad is a large unit, capacitors in electronic circuits almost always have fractional values: *microfarads* (μF), *nanofarads* (nF), and *picofarads* (pF). The Greek letter μ (mu) should be used in the μF abbreviation, but a lowercase letter u is often substituted. Thus, for example, 10uF means the same as 10μF.

1F = 1,000,000μF, and 1μF = 1,000,000pF. Therefore, 1 farad is equivalent to 1 trillion picofarads —a very wide range of possible values. See Figure 12-16 and Figure 12-17 for charts showing equivalent values in different units.

pF	nF	μF
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 12-16. *Equivalent values for picofarads, nanofarads, and microfarads. The nF unit is used primarily in Europe.*

μF	F
1	0.000001
10	0.00001
100	0.0001
1,000	0.001
10,000	0.01
100,000	0.1
1,000,000	1

Figure 12-17. *Equivalent values for microfarads and farads. Because the farad is such a large unit, electronic circuits almost always use fractional values.*

The nF unit is more common in Europe than in the United States. A 1nF capacitance is often expressed in the US as 0.001μF or 1,000pF. Similarly, a 10nF capacitance is almost always expressed as a 0.01μF, and a 0.1nF capacitance is more likely to be expressed as 100pF.

European schematics may use value-symbols as a substitute for decimal points. For example, a 4.7pF capacitor may be shown as 4p7, a 6.8nF capacitor may be shown as 6n8, and a 3.3μF capacitor may be shown as 3μ3.

Commonly Used Values

The traditional range of capacitor values was established on the same basis as the traditional range of resistor values, by assuming an accuracy of plus-or-minus 20% and choosing factors that would minimize the possible overlap between adjacent tolerance ranges. The factors 1.0, 1.5, 2.2, 3.3, 4.7, 6.8, and 10 satisfy this requirement. See Chapter 10 for a more detailed explanation, including a graphical representation of values and overlaps in Figure 10-8. While many resistors are now manufactured with high precision, 20%

tolerance is still common for electrolytic capacitors. Other types of capacitors are available with an accuracy of 10% or 5%, but are more expensive.

While large-value capacitors are likely to have their actual value printed on them, smaller capacitors are identified by a variety of different codes. These codes are not standardized among manufacturers, and exist in various colors and abbreviations. A multimeter that can measure capacitance is a quicker, easier, and more reliable method of determining the value of a component than trying to interpret the codes.

In addition to capacitance, a large capacitor is likely to have its working voltage printed on it. Exceeding this value increases the risk of damaging the dielectric. In the case of electrolytic capacitors, a voltage that is much lower than the rated value should also be avoided, because these capacitors require an electrical potential to maintain their performance.

In common electronics applications, values larger than 4,700µF or smaller than 10pF are unusual.

Electrolytics are available at a moderate price in a wider range of values than other commonly used capacitors. They range from 1µF to 4,700µF and sometimes beyond. Working voltages typically range from 6.3VDC to 100VDC, but can be as high as 450VDC.

Tantalum capacitors are usually unavailable in sizes above 150µF or for voltages above 35VDC.

Single-layer ceramic capacitors have small values ranging from 0.01µF to 0.22µF, with working voltages usually not exceeding 50VDC, although very small-value capacitors may be rated much higher for special applications. Poor tolerances of +80% to -20% are common.

Some variants of multi-layer ceramic capacitors are capable of storing up to 47µF, although 10µF is a more common upper limit. They are seldom rated above 100VDC. Some are accurate to plus-or-minus 5%.

Dielectric Constant

If A is the area of each plate in a capacitor (measured in square centimeters), and T is the thickness of the dielectric (measured in centimeters), and K is the *dielectric constant* of the capacitor, the capacitance, C (measured in farads) will be obtained from the formula:

```
C = (0.0885 * K * A) / T
```

The dielectric constant of air is 1. Other dielectrics have different standard values. Polyethylene, for instance, has a constant of approximately 2.3. Thus a capacitor of 1 square centimeter plate area and polyethylene dielectric 0.01 centimeters thick would have a capacitance of about 20pF. A tantalum capacitor of equal plate area and dielectric thickness would have capacitance closer to 100pF, since the dielectric constant of tantalum oxide is much higher than that of polyethylene.

The Time Constant

When a capacitor is charged in series through a resistor (it is used in an RC network), and it begins with no charge on its plates, the *time constant* is the time, in seconds, required to charge the capacitor to 63% of the supply voltage. After an additional, identical interval of time, the capacitor will acquire 63% of the remaining difference between itself and the power supply. In theory the capacitor gets closer and closer to a full charge, but never quite reaches 100%. However, five time constants are sufficient for the capacitor to reach 99%, which is regarded as close enough to being fully charged for all practical purposes.

Refer to Figure 12-3 for a schematic of an RC network.

The time constant is a simple function of the resistance and the capacitance. If R is the value of the resistor (in ohms), and C is the value of the capacitor (in farads), the time constant, TC, will be obtained by the formula:

```
TC = R * C
```

If we multiply the R value by 1,000 while dividing the C value by 1,000, the time constant remains the same, and we can use the more convenient values of kilohms for the resistance and µF for the capacitance. In other words, the formula tells us that a 1K resistor in series with a 1,000µF capacitor has a time constant of 1 second.

The formula suggests that if the value of R diminishes to zero, the capacitor will charge instantly. In reality, the charging time will be rapid but finite, limited by factors such as the electrical resistance of the materials used.

Multiple Capacitors

When two or more capacitors are wired in parallel, their total capacitance is the sum of their separate capacitances. When two or more capacitors are wired in series, the relationship between their total capacitance © and their individual capacitances (C1, C2, C3 . . .) is given by this formula:

$$1 / C = (1/C1) + (1/C2) + (1/C3). \; . \; .$$

The formula to calculate the total capacitance of capacitors connected in series resembles the one used to calculate the total resistance of resistors connected in parallel. See Chapter 10.

Alternating Current and Capacitive Reactance

The apparent resistance of a capacitor to AC is properly known as *capacitive reactance*. In the following formula, capacitive reactance (X_C, in ohms) is derived as a function of capacitance (C, in farads) and AC frequency (f, measured in hertz):

$$X_C = 1 / (2 * \pi * f * C)$$

The formula shows that when frequency becomes zero, capacitive reactance becomes infinite; in other words, a capacitor has theoretically infinite resistance when DC current tries to flow through it. In reality, a dielectric has a finite resistance, and thus always allows some leakage.

The formula also shows that capacitive reactance diminishes when the size of the capacitor increases and/or the frequency being applied to it increases. From this it appears that an AC signal will be attenuated less at higher frequencies, especially if we use a small capacitor. However, a real-world capacitor also exhibits some degree of *inductive reactance*. This value will depend on its configuration (cylindrical vs. multiple flat plates), its physical length, the materials from which it is fabricated, the lengths of its leads, and other factors. Inductive reactance tends to *increase* with frequency, and since capacitive reactance tends to *decrease* with frequency, at some point the curves for the two functions intersect. This point represents the capacitor's *self-resonant frequency*, which is often referred to simply as its *resonant frequency*. See Figure 12-18.

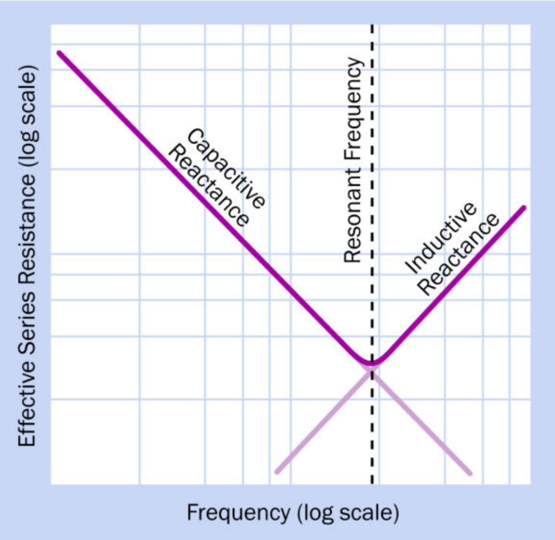

Figure 12-18. *As an AC current applied to a capacitor increases in frequency, the capacitive reactance of the component decreases, while its inductive reactance increases. The resonant frequency of the capacitor is found where the two functions intersect.*

Equivalent Series Resistance

A theoretically ideal capacitor would be purely reactive, without any resistance. In reality, capacitors are not ideal, and they have *equivalent series resistance,* or *ESR*. This is defined as the resistor that you would have to place in series with an ideal version of the capacitor, so that the combination would behave like the real version of the capacitor on its own.

If X_c is the reactance of the capacitor, then its Q factor (which means its *quality factor*) is given by the simple formula:

$$Q = X_C / ESR$$

Thus, the quality factor is higher if the ESR is relatively low. However, the reactance of the capacitor will vary significantly with frequency, and this simple formula is only an approximate guide.

The Q-factor for capacitors should not be confused with the Q-factor for inductors, which is calculated quite differently.

How to Use it

The figures illustrate some simplified schematics for common applications.

Bypass Capacitor

In Figure 12-19, a low-value capacitor (often 0.1µF) is placed near the power input pin of a sensitive digital chip to divert high-frequency spikes or noise to negative ground. This *bypass capacitor* may also be described as a *decoupling capacitor*.

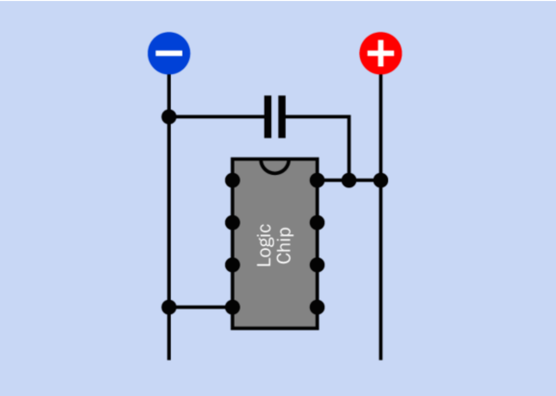

Figure 12-19. *A bypass capacitor (typically 0.1µF) configured to protect an integrated circuit logic chip from voltage spikes and noise in the power supply.*

Coupling Capacitor

In Figure 12-20, a 1µF *coupling capacitor* transmits a pulse from one section of a circuit to another, while blocking the DC voltage. Some reshaping of the waveform may occur.

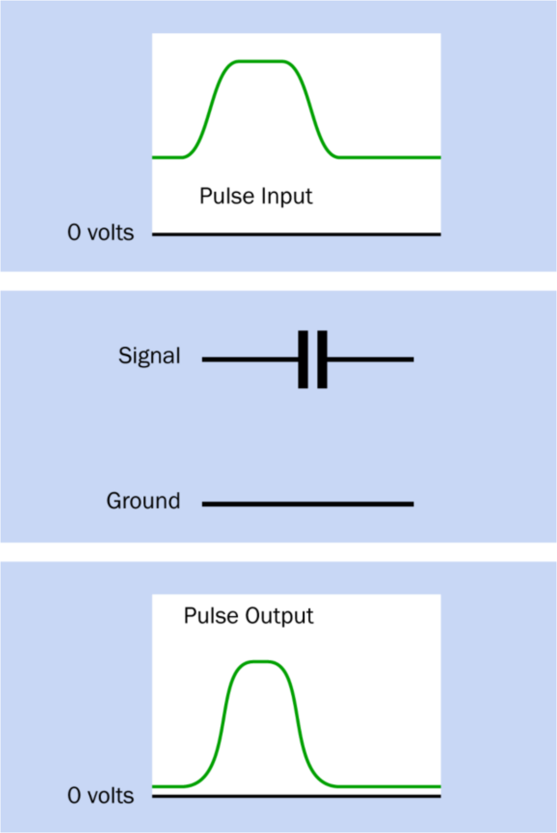

Figure 12-20. *A coupling capacitor (typically around 1µF) preserves DC isolation of one section of a circuit from another, while allowing a pulse to be transmitted.*

High-Pass Filter

In Figure 12-21, a 0.1µF capacitor blocks the low-frequency component of a complex waveform and transmits only the higher frequency that was superimposed on the low frequency.

Low-Pass Filter

In Figure 12-22, a 0.1µF decoupling capacitor diverts the higher frequency component of a complex waveform to negative ground, allowing only

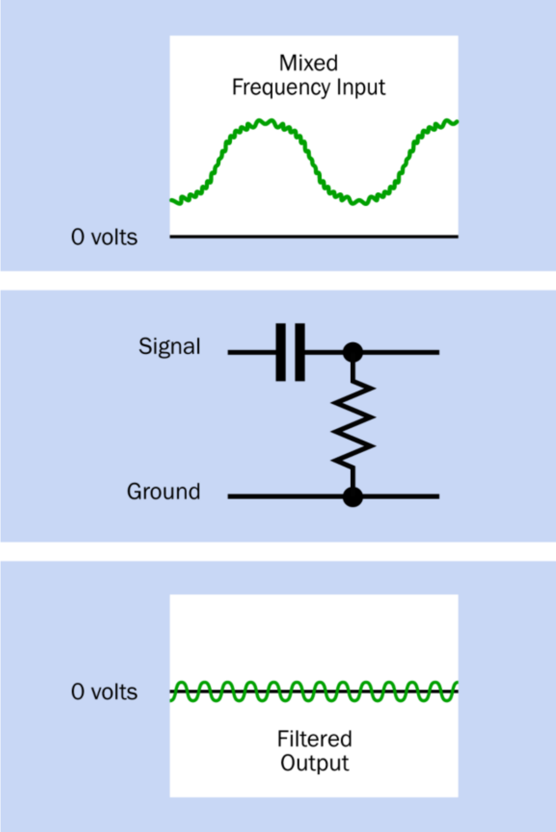

Figure 12-21. *A small capacitor (typically 0.1µF) can be used to create a high-pass filter, passing high frequencies while blocking low frequencies.*

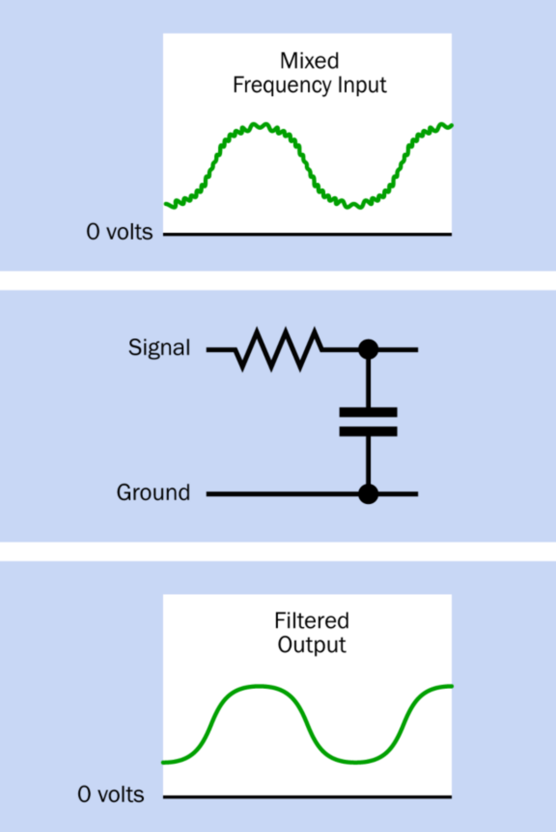

Figure 12-22. *A small capacitor (typically 0.1µF) in this configuration routes high frequencies to negative ground, filtering them out of an analog signal.*

the lower frequency to be preserved. A lower-value capacitor (such as 0.001µF) will bleed away high-frequency noise from an AM radio source without affecting audio frequencies.

Smoothing Capacitor

In Figure 12-23, a 100µF capacitor charges and discharges to smooth an AC signal after a diode has removed the negative portion.

Snubber

In Figure 12-24, an RC network (inside a white dashed line) is known as a *snubber* when used to protect a switch from the problem of *arcing* (pronounced "arking")—that is, a sustained spark that can quickly erode the switch contacts. Arcing may occur in switches, pushbuttons, or relays that control an inductive load, such as a large motor. This problem can become significant at high DC currents (10A or more) or relatively high AC or DC voltages (100V or more).

When the switch is opened, the magnetic field that has been sustained by the inductive load collapses, causing a surge of current, or *forward EMF*. The capacitor in the snubber absorbs this surge, thus protecting the switch contacts. When the switch is closed again, the capacitor discharges itself, but the resistor limits the outrush of current—again, protecting the switch.

A snubber placed around the switch in a DC circuit could typically use a 0.1µF capacitor (polypropylene or polyester) rated for 125VAC/200VDC, and a 100-ohm carbon resistor rated 0.5

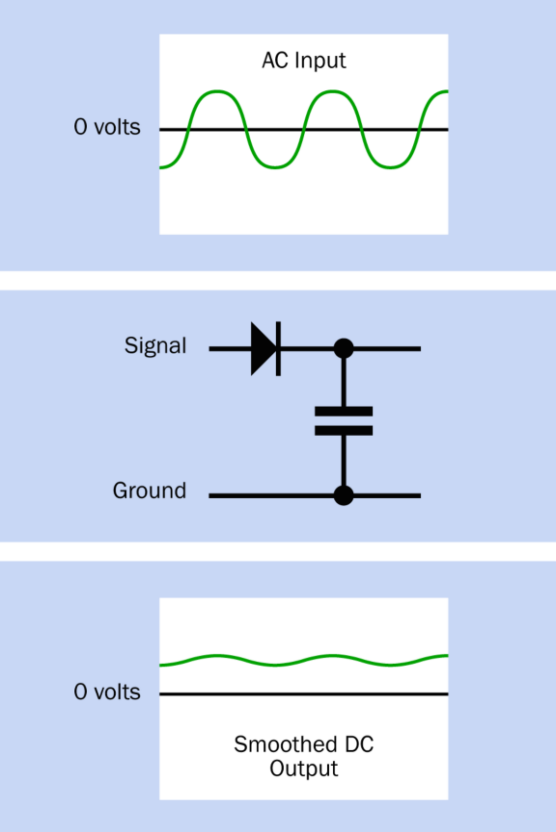

Figure 12-23. *A capacitor of 100µF or more smooths the upper half of an AC signal that has passed through a diode. The capacitor charges during each positive pulse and discharges to "fill the gaps" between them.*

watt or higher. Prepackaged snubbers containing appropriate capacitor-resistor pairs are available from some parts suppliers, primarily for industrial use.

In an AC circuit, a snubber can be placed around the inductive load itself. Although a **diode** is often used this way in a DC circuit, it cannot be used with AC.

Although solid-state switching devices such as a **solid state relay** contain no mechanical contacts, they may still be damaged by substantial pulses of back-EMF, and can be protected by a snubber where they are controlling inductive loads that take 10A or more at 100V or more.

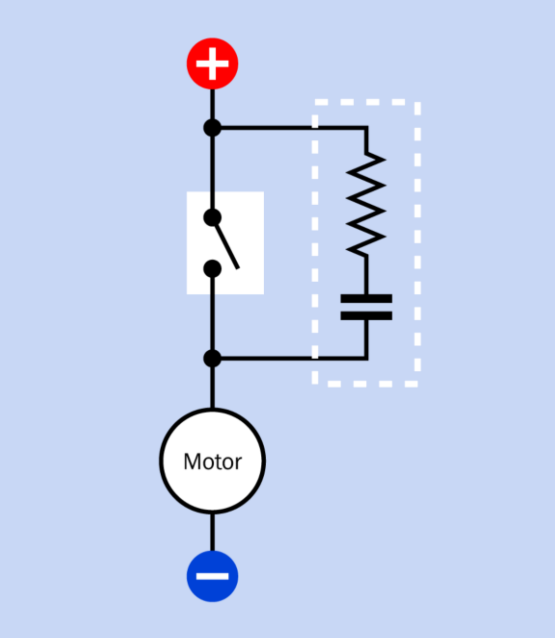

Figure 12-24. *An RC network (outlined with a white dashed line) protects a switch that controls a high inductive load. Used in this way, the RC network is known as a snubber.*

Capacitor as a Battery Substitute

A capacitor may be substituted for a battery for some applications, although it has a lower energy density and will be more expensive to manufacture. A capacitor charges and discharges much more rapidly than a battery because no chemical reactions are involved, but a battery sustains its voltage much more successfully during the discharge cycle.

Capacitors that can store a very large amount of energy are often referred to as *supercapacitors.*

What Can Go Wrong

Common problems associated with capacitors are age-related deterioration (especially in electrolytics), inductive reactance (especially in cylindrical formats), nonlinear response, resistivity, excessive current leakage, and dielectric memory. Some of these problems are discussed below. A manufacturer's datasheet should be consulted carefully in conjunction with the notes regarding compositions in the preceding Variants section before making a commitment to a particular type of capacitor.

Wrong Polarity

A polarized capacitor may offer virtually no resistance if it is connected the wrong way around to a DC power source. A very high current can result, damaging the capacitor and probably other components in the circuit. Failing to observe the polarity of a tantalum capacitor can have destructive or even explosive consequences, depending on the amperage.

Voltage Overload

If the DC working voltage of a capacitor is exceeded, there is a risk of breaking down the dielectric and allowing a spark, or arc, that will form a short circuit. Note that the DC rating of a capacitor does not mean that it can be used safely with an equivalent AC voltage. The maximum AC voltage should be no greater than approximately 0.7 times the DC rated voltage. If a DC-rated capacitor is used directly across an AC power line, it will create an effective short circuit.

If capacitors are connected in series or in parallel, ideally the voltage rating for each capacitor should be the same, and certainly no less than the supply voltage.

Tantalum capacitors are easily damaged by current spikes that exceed their maximum working voltage, and are unsuitable for high-frequency coupling because of their inductance.

Leakage

Charge leakage is a problem especially associated with electrolytic capacitors, which are not suitable for storing a charge over a significant interval. Polypropalene or polystyrene film capacitors are a better choice.

Dielectric Memory

Also known as *dielectric absorption*, this is a phenomenon in which a capacitor's electrolyte displays some percentage of its former voltage after the capacitor has been discharged and then disconnected from the circuit. Single-layer ceramic capacitors especially tend to suffer from this problem.

Specific Electrolytic Issues

Electrolytic capacitors have high inductive reactance, are not manufactured to close tolerances, and deteriorate significantly with age. While other components may be stockpiled and used over a period of years, this is not a sensible policy with electrolytics.

The "capacitor plague" affecting many of these capacitors manufactured from 1999 onward provided a salutary lesson regarding their potential weaknesses. Faulty composition of the dielectric allowed it to deteriorate, liberating hydrogen gas, which eventually caused the aluminum shells of the capacitors to bulge and burst. Circuit boards from major manufacturers were affected. Because the problem took two years to become apparent, literally millions of boards with faulty capacitors had been sold before the fault was diagnosed and eventually corrected.

Unfortunately electrolytics cannot be easily replaced with other types of capacitors in applications such as power supplies, because substitutes will be considerably larger and more expensive.

Heat

The equivalent series resistance (ESR) of a large capacitor inevitably means that it must dissipate some power as heat during use. Ripple current can also create heat. Capacitor performance will change as the temperature increases. A common maximum component temperature for electrolytic capacitors is 85 degrees Centigrade.

Vibration

In a high-vibration environment, electrolytics should be protected by clamping them mechanically in place, using a *capacitor clamp*, also known as a *c-clamp*.

Misleading Nomenclature

Rarely, in the United States, the term “mF” may be used as a probable alternative to μF. This can be a source of confusion and risk because mF is properly (but very rarely) used to mean “millifarads.” The term should always be avoided.

variable capacitor

13

Formerly known (primarily in the United Kingdom) as a *variable condenser*. The term is now obsolete.

OTHER RELATED COMPONENTS

- **capacitor** (See Chapter 12)

What It Does

A variable capacitor allows adjustment of capacitance in much the same way that a **potentiometer** allows adjustment of resistance.

Large variable capacitors were developed primarily to tune radio receivers, in which they were known as *tuning capacitors*. Cheaper, simpler, and more reliable substitutes gradually displaced them, beginning in the 1970s. Today, they are still used in semiconductor fabrication, in RF plastic welding equipment, in surgical and dental tools, and in ham radio equipment.

Small *trimmer capacitors* are widely available and are mostly used to adjust high-frequency circuits. Many of them look almost indistinguishable from *trimmer potentiometers*.

The schematic symbols commonly used to represent a variable capacitor and a trimmer capacitor are shown in Figure 13-1.

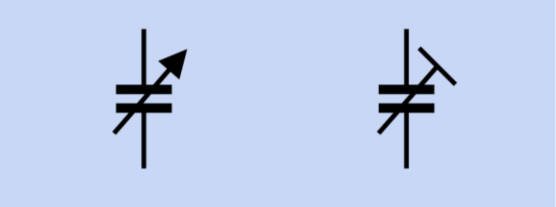

Figure 13-1. *Typical schematic symbols for variable capacitor (left) and trimmer capacitor (right).*

A *varactor* is a form of diode with variable capacitance, controlled by reverse voltage. See "Varactor Diode" (page 225) for this component.

How It Works

The traditional form of variable capacitor consists of two rigid semicircular plates separated by an air gap of 1mm to 2mm. To create more capacitance, additional interleaved plates are added to form a stack. One set of plates is known as the *rotor*, and is mounted on a shaft that can be turned, usually by an externally accessible knob. The other set of plates, known as the *stator*, is mounted on the frame of the unit with ceramic insulators. When the sets of plates completely overlap, the capacitance between them is maximized. As the rotor is turned, the sets of plates gradually disengage, and the capacitance diminishes to near zero. See Figure 13-2.

The air gaps between the sets of plates are the *dielectric*. Air has a dielectric constant of approximately 1, which does not vary significantly with temperature.

The most common shape of plate is a semicircle, which provides a linear relationship between capacitance and the angle of rotation. Other shapes have been used to create a nonlinear response.

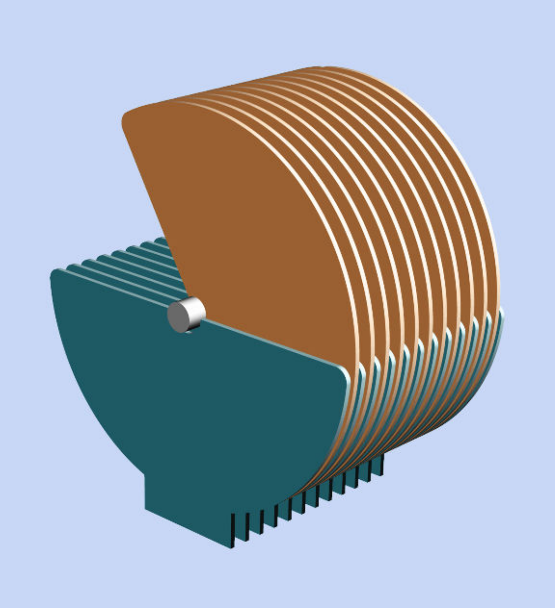

Figure 13-2. *In this simplified view of a variable capacitor, the brown plates constitute the rotor, attached to a central shaft, while the blue plates are the stator. The colors have no electrical significance and are added merely for clarity. The area of overlap between rotor and stator determines the capacitance.*

Reduction gears may be used to enable fine tuning of a variable capacitor, which means multiple turns of a knob can produce very small adjustments of the capacitor. At the peak of variable capacitor design, units were manufactured with high mechanical precision and included *anti-backlash gears*. These consisted of a pair of equal-sized gears mounted flat against each other with a spring between them that attempted to turn the gears in opposite directions from each other. The pair of gears meshed with a single pinion, eliminating the looseness, or backlash, that normally exists when gear teeth interlock. A vintage capacitor with a spring creating anti-backlash gearing (circled) is shown in Figure 13-3. This is a two-gang capacitor—it is divided into two sections, one rated 0 to 35pF, the other rated 0 to 160pF.

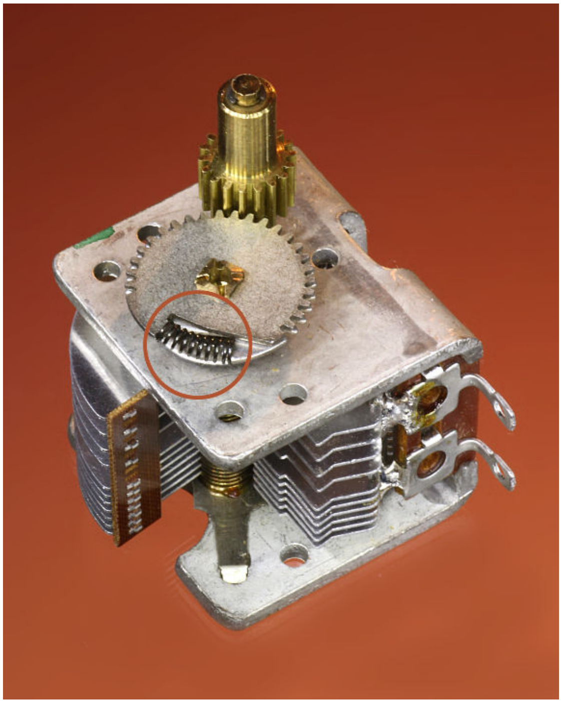

Figure 13-3. *A "traditional style" variable capacitor of the type designed to tune radio frequencies. The spring, circled, enables anti-backlash gearing.*

Variants

The traditional variable capacitor, with exposed, air-spaced, rigid, rotating vanes, is becoming hard to find. Small, modern variable capacitors are entirely enclosed, and their plates, or vanes, are not visible. Some capacitors use a pair of concentric cylinders instead of plates or vanes, with an external thumb screw that moves one cylinder up or down to adjust its overlap with the other. The overlap determines the capacitance.

Trimmer capacitors are available with a variety of dielectrics such as mica, thin slices of ceramic, or plastic.

Values

A large traditional capacitor can be adjusted down to a near-zero value; its maximum will be no greater than 500pF, limited by mechanical factors. (See Chapter 12 for an explanation of capacitance units.)

A maximum value for a trimmer capacitor is seldom greater than 150pF. Trimmers may have their values printed on them or may be color-coded, but there is no universal set of codes. Brown, for example, may indicate either a maximum value around 2pF or 40pF, depending on the manufacturer. Check datasheets for details.

The upper limit of a trimmer's rated capacitance is usually no less than the rated value, but can often be 50% higher.

Formats

All trimmer capacitors are designed for mounting on circuit boards. Many are surface-mount, with a minority being through-hole. Surface-mount units may be 4mm × 4mm or smaller. Through-hole are typically 5mm × 5mm or larger. Superficially, trimmer capacitors resemble single-turn trimmer potentiometers with a screw head in the center of a square package. A through-hole example is shown in Figure 13-4.

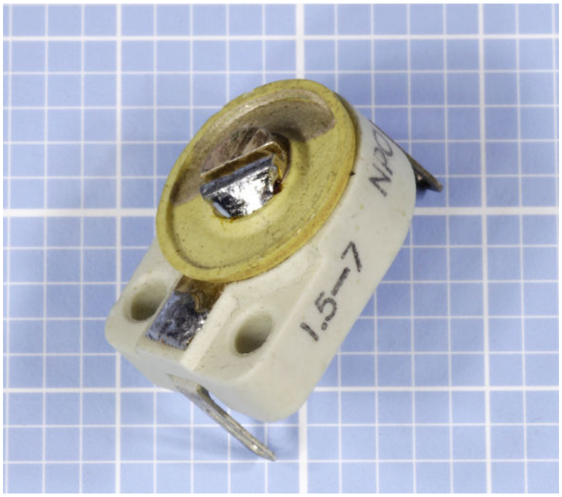

Figure 13-4. *A trimmer capacitor rated 1.5pF to 7.0pF.*

How to Use it

A variable capacitor is often used to tune an *LC circuit,* so called because a coil (with reactance customarily represented by letter L) is wired in parallel with a variable capacitor (represented by letter C). The schematic in Figure 13-5 shows an imaginary circuit to illustrate the principle. When the switch is flipped upward, it causes a large fixed-value capacitor to be charged from a DC power source. When the switch is flipped down, the capacitor tries to pass current through the coil—but the coil's *reactance* blocks the current and converts the energy into a magnetic field. After the capacitor discharges, the magnetic field collapses and converts its energy back into electricity. This flows back to the capacitor, but with inverted polarity. The cycle now repeats with current flowing in the opposite direction. A low-current **LED** across the circuit would flash as the voltage oscillates, until the energy is exhausted.

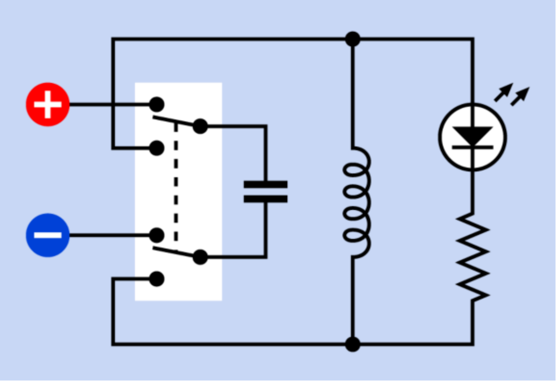

Figure 13-5. *In this imaginary circuit, the capacitor is charged through the double-pole switch in its upper position. When the switch is turned, the capacitor forms an LC (inductance-capacitance) circuit with the coil, and resonates at a frequency determined by their values. In reality, extremely high values would be needed to obtain a visible result from the LED.*

Because the oscillation resembles water sloshing from side to side in a tank, an LC circuit is sometimes referred to as a *tank circuit.*

In reality, unrealistically large values would be required to make the circuit function as described. This can be deduced from the following formula, where f is the frequency in Hz, L is inductance in Henrys, and C is capacitance in Farads:

```
f = 1 / (2π * √(L * C) )
```

For a frequency of 1 Hz, a massive coil opposite a very large capacitor of at least 0.1F would be needed.

However, an LC circuit is well-suited to very high frequencies (up to 1,000MHz) by using a very small coil and variable capacitor. The schematic in Figure 13-6 shows a high-impedance earphone and a diode (right) substituted for the LED and the resistor in the imaginary circuit, while a variable capacitor takes the place of the fixed capacitor. With the addition of an antenna at the top and a ground wire at the bottom, this LC circuit is now capable of receiving a radio signal, using the signal itself as the source of power. The resonant frequency of the circuit is tuned by the variable capacitor. The impedance peaks at the resonant frequency, causing other frequences to be rejected by passing them to ground. With suitable refinement and amplification, the basic principle of an LC circuit is used in AM radios and transmitters.

Because variable capacitors are so limited in size, they are unsuitable for most timing circuits.

Trimmer capacitors are typically found in high-power transmitters, cable-TV transponders, cellular base stations, and similar industrial applications.

They can be used to fine-tune the resonant frequency of an oscillator circuit, as shown in Figure 13-7.

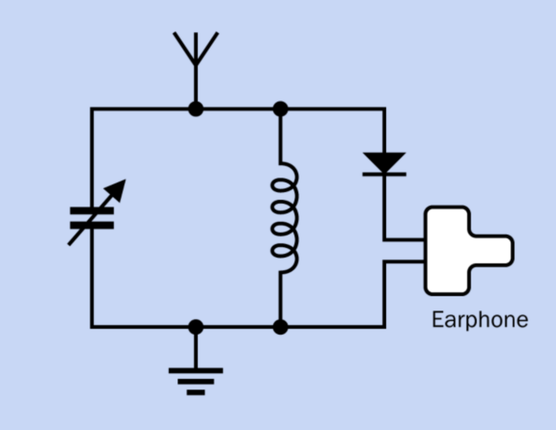

Figure 13-6. *The principle of an LC circuit is used here in a basic circuit that can tune in to a radio station and create barely audible sound through the earphone at right, using only the broadcast signal for power. The variable capacitor adjusts the frequency of the circuit to resonate with the carrier wave of the radio signal.*

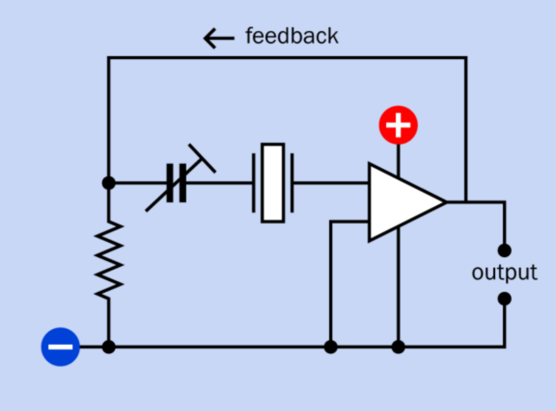

Figure 13-7. *A trimmer capacitor in series with a crystal fine-tunes the frequency of this basic circuit using an op-amp.*

In addition to tuning a circuit frequency, a trimmer capacitor can be used to compensate for changes in capacitance or inductance in a circuit that are caused by the relocation of wires or rerouting of traces during the development process. Readjusting a trimmer is easier than swapping fixed-value capacitors. A trimmer may also be used to compensate for capacitance in a circuit that gradually drifts with age.

What Can Go Wrong

Failure to Ground Trimmer Capacitor While Adjusting it

Although trimmer capacitors are not polarized, the manufacturer may mark one terminal with a plus sign and/or the other with a minus sign. If the capacitor is adjusted while its negative terminal is floating or ungrounded, a metal screwdriver blade will create erroneous readings. Always ground the appropriate side of a trimmer capacitor before fine-tuning it, and preferably use a plastic-bladed screwdriver.

Application of Overcoat Material or "Lock Paint"

Overcoat material is a rubbery adhesive that may be spread over assembled components to immunize them against moisture or vibration. *Lock paint* is a dab of paint that prevents a screw adjustment from turning after it has been set. Most manufacturers advise against applying these materials to a trimmer capacitor, because if penetration occurs, the capacitor can fail.

Lack of Shielding

Variable capacitors should be shielded during use, to protect them from external capacitive effects. Merely holding one's hand close to a variable capacitor will change its value.

inductor

14

The term **inductor** is used here to describe a coil that has the purpose of creating *self-inductance* in an electronic circuit, often while passing alternating current in combination with resistors and/or capacitors. A *choke* is a form of inductor. By comparison, the **electromagnet** entry in this encyclopedia describes a coil containing a center component of ferromagnetic material that does not move relative to the coil, and has the purpose of attracting or repelling other parts that respond to a magnetic field. A coil containing a center component of ferromagnetic material that moves as a result of current passing through the coil is considered to be a **solenoid** in this encyclopedia, even though that term is sometimes more broadly applied.

OTHER RELATED COMPONENTS

- **solenoid** (See Chapter 21)
- **electromagnet** (See Chapter 20)

What It Does

An inductor is a coil that induces a magnetic field in itself or in a *core* as a result of current passing through the coil. It may be used in circuits to block or reshape AC current or a range of AC frequencies, and in this role can "tune" a simple radio receiver or various types of oscillators. It can also protect sensitive equipment from destructive voltage spikes.

The schematic symbol for an inductor includes a coil that can be drawn in two basic styles, shown at the top and at the bottom of Figure 14-1. The style at the bottom has become more common. In each vertical section of the diagram, the functionality of the symbols is identical.

One or two parallel lines alongside the coil indicate that it is wound around a solid core of material that can be magnetized, while one or two dotted lines indicate that it is wound around a core containing metal particles, such as iron filings. Where no core is shown, this indicates an air core.

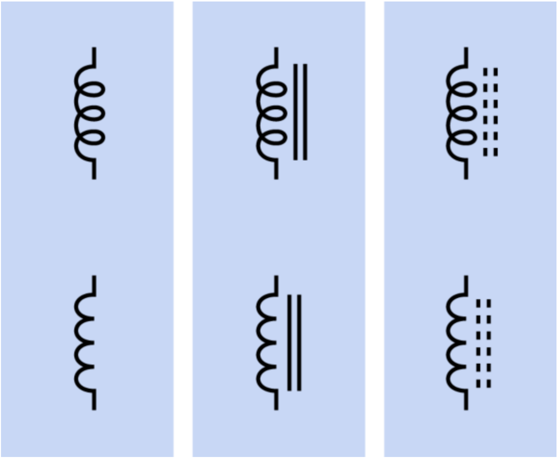

Figure 14-1. *The coil symbol for an inductor may be drawn in two styles which are functionally identical. Line(s) beside the coil indicate a solid core. Dotted line(s) indicate a core containing metal particles.*

A selection of inductors designed for through-hole mounting is shown in Figure 14-2.

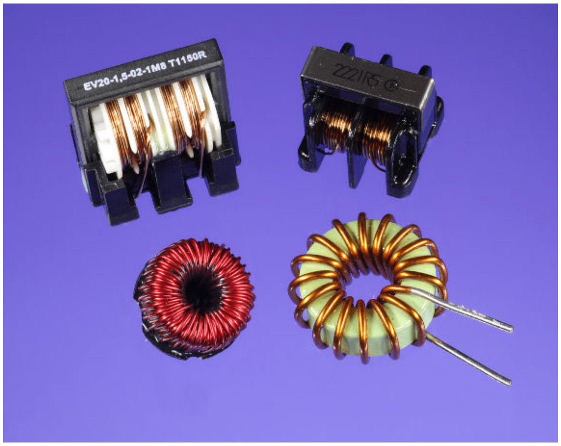

Figure 14-2. *Four inductors designed for through-hole insertion into printed circuit boards.*

How It Works

Direct current passing through an electrical conductor, such as a wire, creates a magnetic field around the conductor. In Figure 14-3, *conventional current* (flowing from positive to negative) is passing through a straight wire from left to right, as indicated by the red/blue arrow. The resulting magnetic field is indicated by the green arrows. If the wire is now bent into a curve, as shown in Figure 14-4, the magnetic field exerts an aggregate force downward through the curve. This magnetic force is conventionally said to flow from *south* to *north*.

If direct current could be induced to circulate through an unbroken circle of wire, the resulting magnetic field could exert a force through the circle as shown in Figure 14-5, assuming clockwise circulation of conventional current as suggested by the red/blue arrows.

Conversely, if a magnet was pushed through the center of the circle, it would induce a pulse of electric current in the circle. Thus, electricity passing through a wire can induce a magnetic field around the wire, and conversely, a magnet moving near a wire can induce an electric current in the wire. This principle is used in an electrical *generator*, and also in a **transformer**, where alternating current in the primary coil induces a fluctuating magnetic field in the core, and the field in the core is turned back into alternating current in the secondary coil.

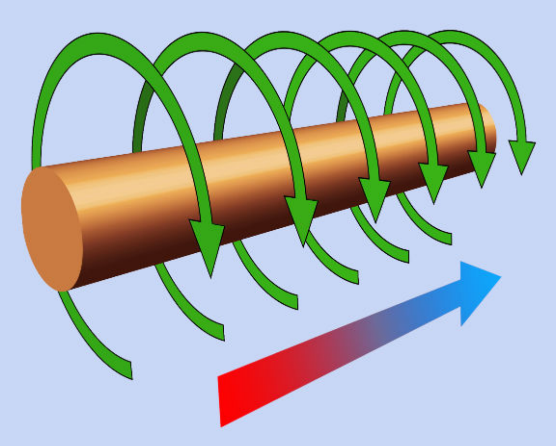

Figure 14-3. *Conventional current passing through a wire from left to right (as indicated by the red/blue arrow) induces a magnetic field around the wire (shown by the green arrows).*

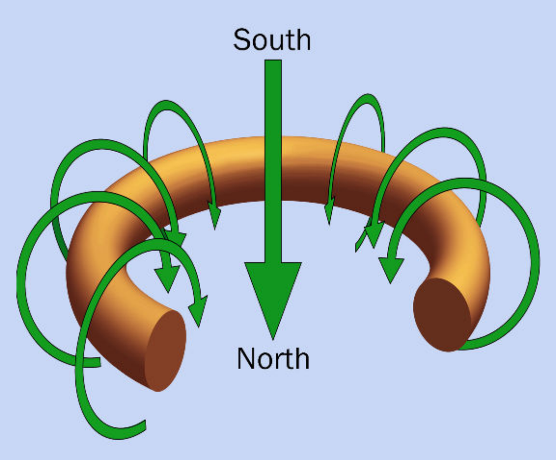

Figure 14-4. *If the wire is bent into a curve, the magnetic fields can create a net force shown by the large green arrow.*

Note that a static or unchanging magnetic field will not induce a flow of electricity.

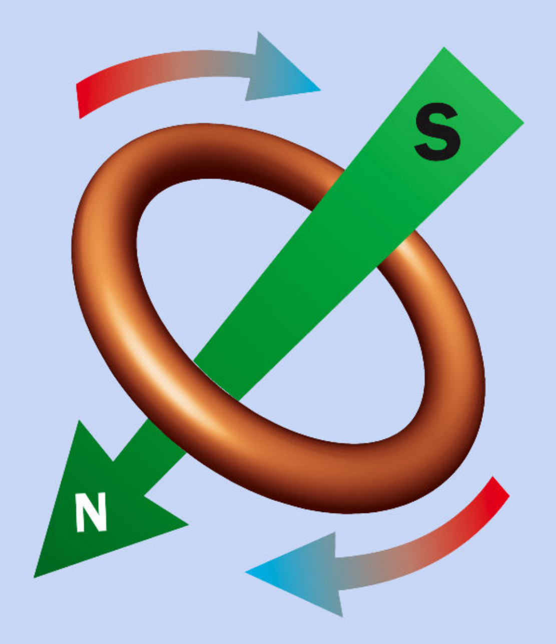

Figure 14-5. *Hypothetically, if conventional current flows around a circular conductor (as suggested by the red/blue arrows), it will create a magnetic field that can create a force as shown by the green arrow.*

DC Through a Coil

If the wire is formed into a helix (a series of approximate circles) as shown in Figure 14-6, and if DC current is passed through the wire, the aggregate of the magnetic fields can create a force in the direction of the green arrow in each example, depending whether the wire is wound clockwise or counter-clockwise, and depending on the direction of the current. The helix is usually referred to as a *coil* or a *winding*.

In actuality, a magnetic field is not open-ended, and its lines of force are completed by circling around outside the inductor, to complete a *magnetic circuit*. This completion of the field can be demonstrated by the traditional high-school experiment of positioning a compass or scattering iron filings on a sheet of paper above a magnet. A simplified depiction of lines of force completing a magnetic circuit is shown in Figure 14-7, where a coil is inducing the magnetic field. Note that throughout this encyclopedia, the color green is used to indicate the presence of magnetic force.

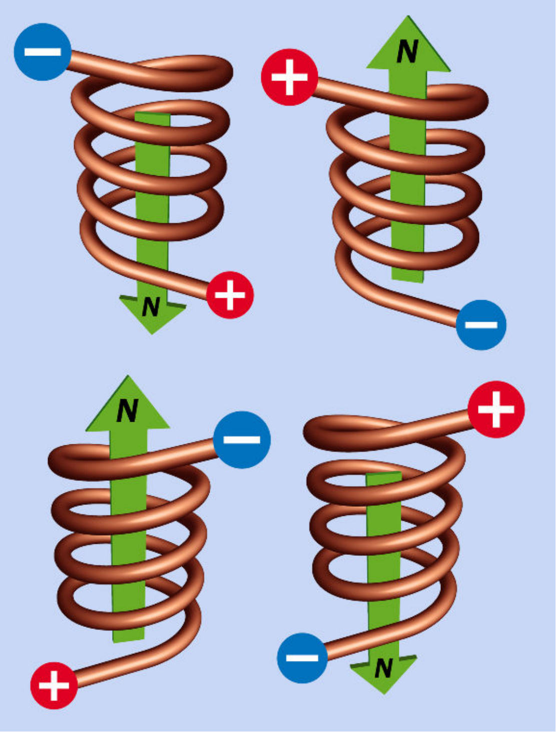

Figure 14-6. *When DC current flows through a coil, it creates magnetic fields that will exert a force whose direction depends on the direction of the current and on whether the coil is wound clockwise or counterclockwise. The force is shown by the green arrow in each case.*

The completion of a magnetic field is not relevant to the primary function of the inductor. In fact the external part of the magnetic field is mostly a source of trouble in electronics applications, since it can interact with other components, and may necessitate the use of magnetic shielding. In addition, the field is weakened by completing itself through air, as air presents much greater *reluctance* (the magnetic equivalent of resistance) than the core of an electromagnet.

The polarity of a magnetic field created by a coil can be demonstrated by moving a small permanent magnet toward the coil, as shown in

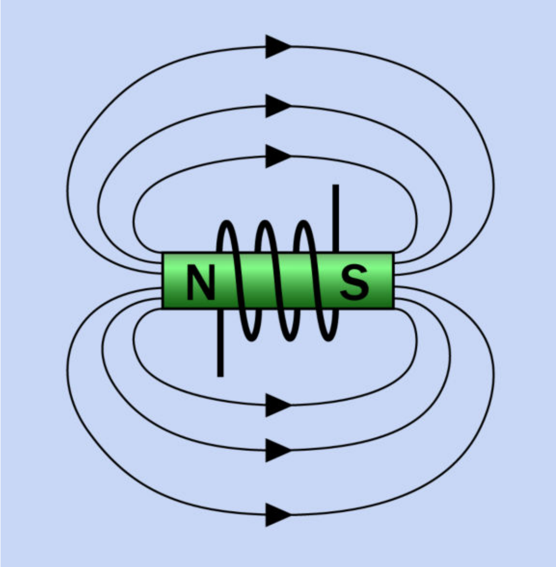

Figure 14-7. *A magnetic field in reality is not open-ended, and each line of force traveling through a rod-shaped magnet or electromagnet is completed outside of the magnet. The completion of magnetic fields has been omitted from other diagrams here for clarity.*

Figure 14-8. If the magnet has opposite polarity to the coil, it will tend to be repelled, as like poles repel. If it has the same polarity, it will tend to be attracted, because opposite poles attract. This principle may be used in **solenoids**.

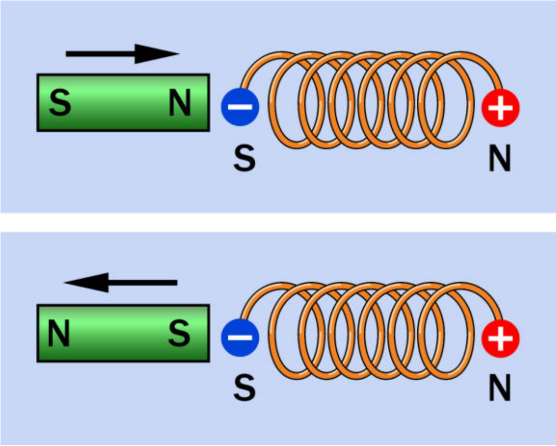

Figure 14-8. *A permanent magnet (left) will either be drawn toward a DC-energized coil or repelled from it, depending on the polarity of the two magnetic fields.*

Magnetic Core

The inductive power of a coil will be enhanced, and the saturation point will be reduced, by using a *magnetic core*. The term "magnetic" in this context does not mean that the core is a permanent magnet; it means that the core can be magnetized briefly by a transient pulse of electricity through the surrounding coil.

A core enhances the effectiveness of an inductor because it has a lower reluctance than that of air. In other words, magnetic flux will flow much more readily through the core than through air.

Roughly speaking, the *permeability* of a magnetic circuit is the opposite of reluctance; it is a measure of how easily a magnetic field can be induced, and is usually expressed relative to the permeability of air, which is approximately 1. The permeability of different core types is discussed in the following "Values" section.

The core of the coil contains *magnetic domains* that behave as tiny magnets, with north and south poles. In the absence of a polarizing magnetic field, the domains are randomly aligned. As a magnetic field is introduced around them and grows stronger, the domains align themselves with it, increasing the total magnetic force. When the domains are almost all uniformly aligned, the core approaches *magnetic saturation* and ceases adding to the net magnetic field. At this point the current in the inductor is said to be *continuous*.

When power to the coil is disconnected, the domains revert partially to their previous random orientation. Thus the core remains a weak permanent magnet. This effect is known as *hysteresis*, while the weak residual field is known as *remanent magnetism*.

EMF and Back-EMF

When DC current is connected through an inductor, the creation of a magnetic field takes a brief but measurable period of time. The field induces an *EMF* (*electro-motive force*) in the wire. Since this force opposes the supplied current, it is referred to as *back-EMF*. It lasts only so long as the field is increasing to its full strength. After the field reaches a steady state, current flows through the coil normally.

This transient resistive effect is caused by the *self-inductance* of the coil, and is opposite to the behavior of a capacitor, which encourages an initial inrush of direct current until it is fully charged, at which point it blocks subsequent current flow.

When high-frequency alternating current attempts to flow through an inductor, if each pulse is too brief to overcome the back-EMF, the coil will block the current. A coil can thus be designed to block some frequencies but not others.

Even a simple electrical circuit that does not contain a coil will still have some self-inductance, simply because the circuit consists of wires, and even a straight length of wire induces a magnetic field when the power is switched on. However, these inductive effects are so small, they can generally be ignored in practical applications.

The transient electrical resistance to alternating current caused by either an inductor or a capacitor is known as *reactance*, although it occurs under opposite electrical conditions, as the coil impedes an initial pulse of DC current and then gradually allows it to pass, while a capacitor allows an initial pulse of DC current and then impedes it.

When a flow of DC current through a coil is switched off, the magnetic field that was created by the coil collapses and releases its stored energy. This can cause a pulse of forward EMF, and like back-EMF, it can interfere with other components in a circuit. Devices such as motors and large relays that contain substantial coils can create problematic spikes of back-EMF and forward-EMF. The forward-EMF that occurs when power to the coil is interrupted is typically dealt with by putting a diode in parallel with the coil, allowing current to circulate through it. This is known as *clamping* the voltage transient. A diode-capacitor combination known as a *snubber* is also commonly used. For a schematic and additional information on this topic, see "Snubber" (page 108).

A schematic to demonstrate EMF and back-EMF is shown in Figure 14-9. The coil can be a 100-foot spool of 26-gauge (or smaller) hookup wire, or magnet wire. It will function more effectively if a piece of iron or steel, such as half-inch galvanized pipe, is inserted through its center. When the button is pressed, current is briefly impeded by the back-EMF created by the coil, and is diverted through D1, making it flash briefly. Then the coil's reactance diminishes, allowing the current to flow through the coil and bypass the LED. When the pushbutton is released, the coil's magnetic field collapses, and the consequent forward-EMF circulates through D2, causing it to flash briefly. Note that the polarity of back-EMF and forward-EMF are opposite, which is why the LEDs in the circuit are oriented with opposite polarities.

The 220Ω resistor should be rated at 1/4 watt minimum, and the button should not be held down for long, as the electrical resistance of the coil is relatively low. The LEDs ideally should be rated for a minimal forward current of no more than 5mA.

Electrical and Magnetic Polarity

Various mnemonics and images have been created to assist in memorizing the polarity or direction of the magnetic field that will be created by a flow of electricity. The *right-hand rule* suggests that if the fingers of the right hand are curled around a coil in the same direction in which the turns of the coil were wound, and if conventional DC current also flows in this direction, the extended thumb will point in the direction of the principal force that can be created by the magnetic field.

By convention, the magnetic field is oriented from *south* to *north*, which can be remembered since the north end of the magnetic field will be

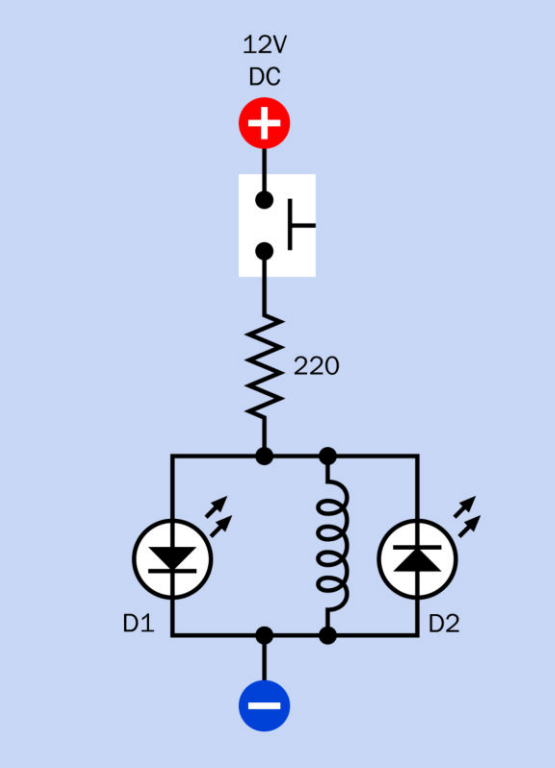

Figure 14-9. *A test circuit to demonstrate the EMF and back-EMF created when DC current starts and stops passing through a coil. See text for details.*

the negative end of the coil (north and negative both beginning with letter N). This mnemonic only works if conventional (positive) current flows through a coil that is wound clockwise.

Another model is the "corkscrew rule" in which we imagine conventional DC current flowing from the handle of a corkscrew, down through its metal section, toward the pointed end. If the corkscrew is turned clockwise, in the same direction as the electricity, the corkscrew will sink into the cork in the same as the direction as the resulting magnetic force.

Variants

Variants include core materials, core shapes, termination style (for through-hole mounting in perforated board, or for surface-mount), and external finish (some inductors are dipped in insulating material, while others allow their copper magnet wire to be exposed).

In addition there are two functional variants: variable inductors and ferrite beads. Their schematic symbols are shown in Figure 14-10.

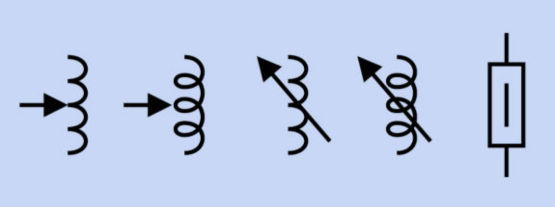

Figure 14-10. *Schematic symbols for a ferrite bead (farthest right) and variable inductors (all other symbols, which are functionally identical).*

Magnetic Cores

A magnetic core may be made from solid iron, plates of iron or steel separated by thin insulating material, powdered iron mixed with a binder, or a ferrite compound derived from nickel, zinc, manganese, or a combination. An iron core has at least 1,000 times the permeability of air, while some ferrites are 10,000 times as permeable.

One major disadvantage of a magnetic core is *hysteresis*, which in this context refers to the tendency of the core to retain some magnetic "memory" as a cycle of alternating current changes from positive to negative. This residual magnetism must be overcome by the next positive pulse of AC. The tendency of the core to retain magnetic polarity is known as its *retentivity*. Iron cores are especially retentive.

Another disadvantage of some magnetic cores is that they may host eddy currents induced by the magnetic field of the coil. These electrical currents tend to circulate through the core, reducing efficiency by generating waste heat, especially if coil currents are high. Forming a core from iron or steel plates, separated by thin layers of insulation, will inhibit these currents. Powdered iron inhibits eddy currents because the particles have limited contact. Ferrites are nonconductive, and are therefore immune to eddy currents. They are widely used.

Hysteresis and eddy currents both incur energy losses with each AC cycle. Therefore, the losses increase linearly as the AC frequency increases. Consequently, inductor cores that suffer either of these problems are not well-suited to high frequencies.

Nonmagnetic Cores

The problems associated with magnetic cores may be avoided by winding the coil around a nonmagnetic core that may be hollow, ceramic, or plastic. A hollow core is referred to as an *air core*. The permeability of ceramic and plastic cores is close to that of air.

An inductor with a nonmagnetic core will be immune to eddy currents and retentivity, but will have to be significantly larger than a magnetic-cored coil with comparable inductance. In the case of a very primitive radio receiver, such as a crystal set, the air-cored coil that selects a radio frequency may be several inches in diameter. A basic circuit diagram for a crystal set (so-called because it uses a diode containing a germanium crystal) is shown in Figure 14-11. The antenna, at top, receives signals broadcast from radio stations. The coil can be tapped (as indicated by the black dots) as a simple way to select different inductance values, blocking all but a narrow range of frequencies. The T-shaped white component at right is a high-impedance earphone. The diode blocks the lower half of the alternating current in a radio signal, and since the signal is *amplitude-modulated*, the earphone responds to variations in intensity in the signal and reproduces the sound encoded in it.

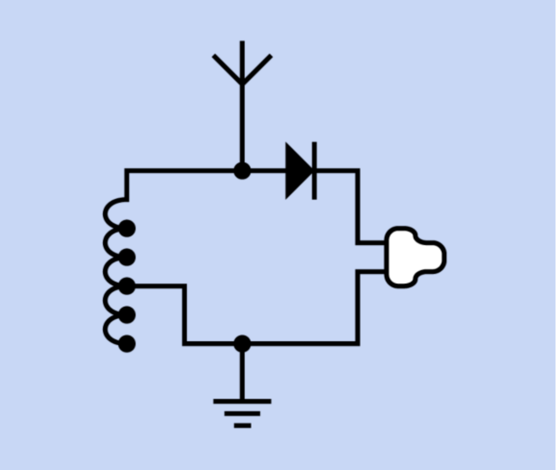

Figure 14-11. *An early and basic application for an inductor is to select radio-station frequencies, as in this schematic for a crystal set. See text for details.*

Variable Inductors

A *variable inductor*, also known as an *adjustable inductor*, is relatively uncommon but can be fabricated by using a magnetic core that penetrates the center of the inductor on an adjustable screw thread. The inductance of the assembly will increase as a larger proportion of the magnetic core penetrates into the open center of the coil. A photograph of a variable inductor is at Figure 14-12.

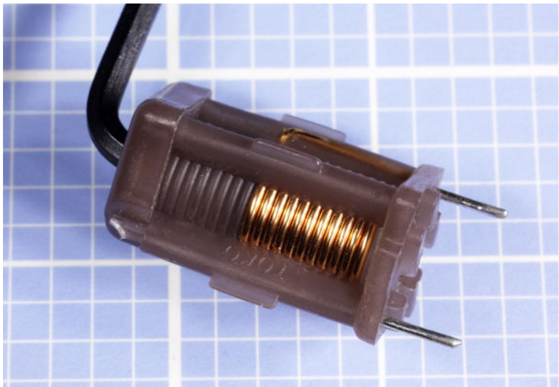

Figure 14-12. *A variable inductor. Its inductance is adjusted via a screw thread that varies the insertion of the core in the coil. In this component the core is turned by inserting a hex wrench, as shown. It is rated from 0.09µH to 0.12µH.*

Ferrite Beads

A *ferrite bead* inverts the design of a typical inductor by running a wire through a hole in the center of the bead, instead of coiling the wire around the core. Two ferrite beads are shown in Figure 14-13. At top, the bead is divided into two sections, each mounted in one-half of a plastic clam shell, which can be closed around a wire. At bottom, the bead must be threaded onto a wire. The purpose is either to limit radio-frequency radiation from a wire by absorbing it into the bead (where it is transformed into heat), or to protect a wire from external sources of radio-frequency radiation. Computer cabling to external devices; lamp dimmers; and some types of motors can be sources of radio frequency.

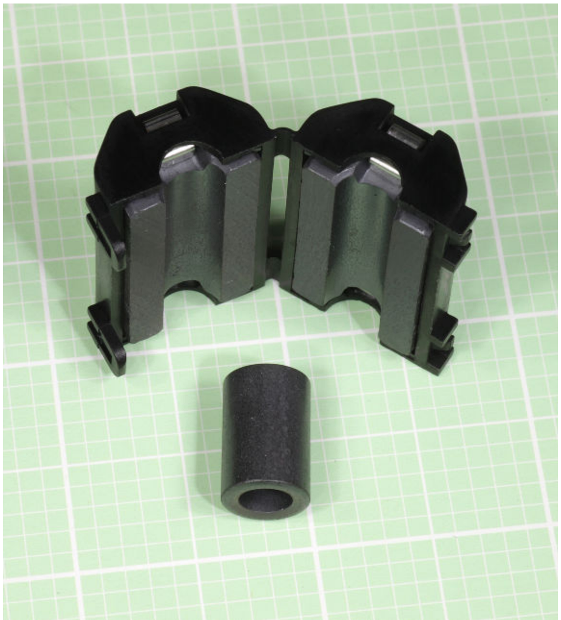

Figure 14-13. *Two examples of ferrite beads. They can inhibit radio-frequency radiation from a wire, or protect the wire from interference.*

Toroidal Cores

The magnetic circuit created by a rod-shaped core must be completed by the lines of force traveling back around from one end of the rod to the other, through the surrounding air. Since air has low permeability, this is a major source of inefficiency. By comparison, a torus (a geometrical shape resembling a donut) completes the entire magnetic circuit inside its core. This significantly increases its efficiency. Also, because its field is better contained, a toroidal inductor needs little or no shielding to protect other components from stray magnetic effects.

Two through-hole toroidal inductors are shown in Figure 14-2. Bottom left: Rated at 345μH. Bottom right: Rated at 15μH. The one at bottom-left has pins beneath it for insertion into a printed circuit board.

Surface-mount inductors often are toroidal to maximize the efficiency of a component that has to function on a very small scale. Examples are shown in Figure 14-14, Figure 14-15, and Figure 14-16.

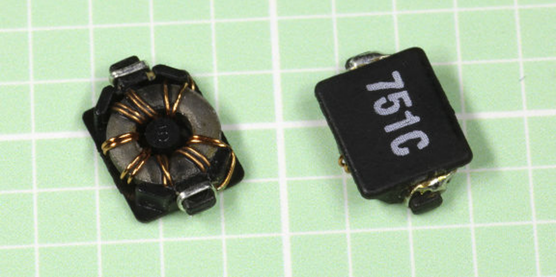

Figure 14-14. *In a typical toroidal inductor, the coil is wrapped around a magnetic core shaped as a torus. This surface-mount component (viewed from the bottom, at left, and from the top, at right) is at the low end of the range of component sizes. It is rated at 750nH.*

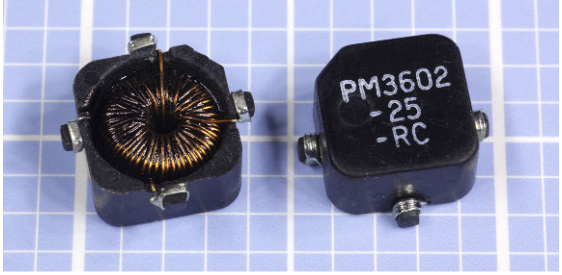

Figure 14-15. *A medium-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at 25μH.*

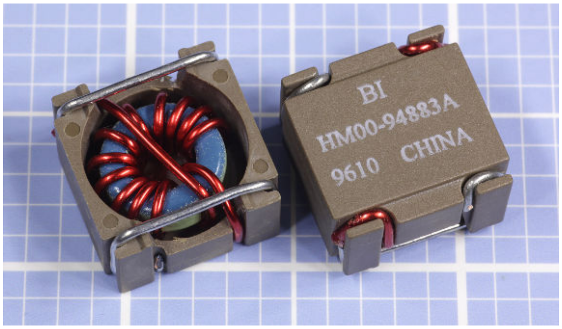

Figure 14-16. *A larger-sized surface-mount toroidal inductor (viewed from the bottom, at left, and from the top, at right). It is rated at 3.8μH.*

A chart showing some types of inductor cores, their commonly available inductances, and their maximum frequencies is shown in Figure 14-17.

Gyrator

A *gyrator* is a small network, sometimes encapsulated in a silicon chip, using **resistors**, a semiconductor, and a **capacitor** to simulate some but not all of the behavior of a coil-based inductor. The semiconductor may be a transistor or a capacitor, depending on the specific circuit. A sample schematic is shown in Figure 14-18. Because no magnetic effects are induced, the gyrator is completely free from the problems of saturation and hysteresis, which affect coils with cores, and also produces no back-EMF. It simply attenuates a signal initially, and then gradually lowers its reactance, thus imitating this aspect of an inductor.

A gyrator may be used where a coil may be unacceptably large (as in a cellular phone) or where signal quality is of paramount importance—for example, in a graphic equalizer or other audio components that perform signal processing at input stages, such as preamplifiers.

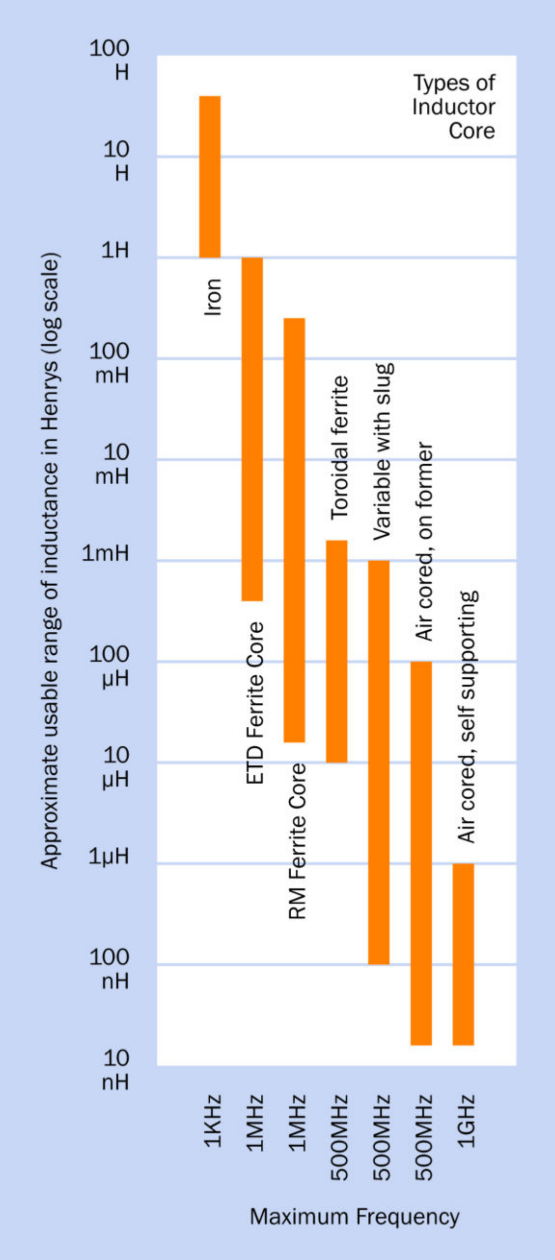

Figure 14-17. *Some commonly used inductor cores and their characteristics. Adapted from "Producing wound components" by R.Clark@surrey.ac.uk.*

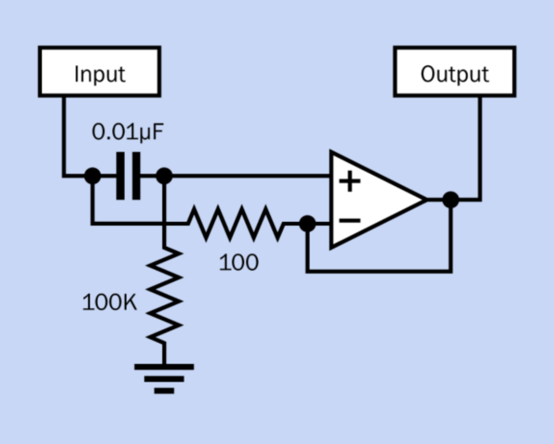

Figure 14-18. *A possible schematic for a coil substitute known as a gyrator, which may be used where a conventional coil would be unacceptably bulky.*

A gyrator does impose some limits on circuit design. While neither side of a real inductor needs to be at ground potential, a gyrator does require a ground connection. However, the performance advantages of gyrators are significant, as they can emulate high inductance without parasitic effects, can be more accurately calibrated (leading to more predictable performance), and do not create magnetic fields that can interfere with other components.

Values

Calculating Inductance

The magnetic inductance of a coil is measured with a unit known as the Henry, named after Joseph Henry, a pioneer in electromagnetism. It is defined by imagining a coil in which current is fluctuating, causing the creation of EMF. If the rate of fluctuation is 1 amp per second and the induced EMF is 1 volt, the inductance of the coil is 1 Henry.

The letter L is commonly used to represent inductance. To derive a useful formula, L will be expressed in microhenrys. If D is the diameter of a coil, N is the number of turns of wire, and W is the width of coil (when the windings are viewed from the side, as shown in Figure 14-19), the precise relationship of the variables is complex but can be reduced to an approximate formula:

$$L = \text{approx } (D^2 * N^2) / 18 * D) + (40 * W$$

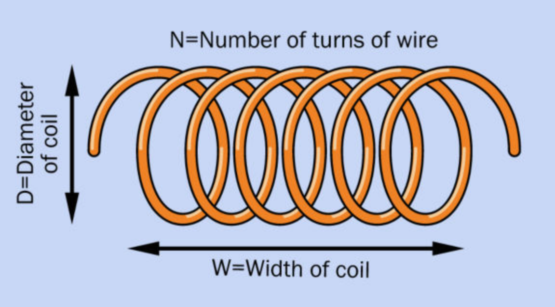

Figure 14-19. *Dimensions of a coil, referenced by a formula to calculate its approximate inductance. See text for details.*

From this, it is clear that inductance tends to increase with coil diameter, and also increases (more significantly) with the square of the number of turns. If the number of turns remains constant, inductance will be higher for a coil that is short and fat than for a coil that is narrow and long.

Because the Henry is a large unit, inductors in electronics circuits typically have their inductances measured in millihenrys (mH), microhenrys (µH), and nanohenrys (nH), where 1 H = 1,000mH, 1 mH = 1,000µH, and 1µH = 1,000 nH. This relationship is shown in Figure 14-20.

Calculating Reactance

The reactance of an inductor (that is, its dynamic resistance to alternating current) varies with the frequency of the current. If f is the AC frequency (in Hertz), and L is the inductance (in Henrys), the reactance, X_L in ohms, is given by the formula:

$$X_L = 2 * \pi * f * L$$

From this equation, it's apparent that as the frequency tends toward zero (DC current), or if the inductance tends toward zero (a short piece of

nH	µH	mH
1	0.001	0.000001
10	0.01	0.00001
100	0.1	0.0001
1,000	1	0.001
10,000	10	0.01
100,000	100	0.1
1,000,000	1,000	1

Figure 14-20. *Inductance is typically measured in nanohenrys (nH), microhenrys (µH), and millihenrys (mH). Equivalent values in these units are shown here.*

straight wire), the reactance will tend toward zero. Conversely, the inductor will impede current increasingly as the frequency and/or the inductance increases.

Calculating Reluctance

The letter S is often used to represent reluctance, while Greek letter µ customarily represents permeability (not to be confused with the use of µ as a multiplication factor of 1/1,000,000, as in µF, meaning "microfarad"). If A is the area of cross-section of the magnetic circuit and L is its length:

```
S = L / µ * A
```

Datasheet Terminology

A typical manufacturer's datasheet should include an *inductance index* for an inductor, expressed in µH per 100 turns of wire (assuming the wire is in a single layer) for inductors with a powdered iron core, and mH per 1,000 turns of wire for inductors with ferrite cores.

The *DCR* is the DC resistance of an inductor, derived purely from the wire diameter and its length.

The *SRF* is the *self-resonant frequency*. An inductor should be chosen so that AC current passing through it will never get close to that frequency.

ISAT (or I_{sat}) is the saturation current, which results in a magnetic core losing its function as a result of magnetic saturation. When this occurs, inductance drops and the charge current rate increases drastically.

Series and Parallel Configurations

Because the inductance of a coil conducting DC current is proportional to the current, the calculations to derive the total inductance of coils in series or in parallel are identical to the calculations used for resistors.

In series, all the coils inevitably pass the same current, and the total inductance is therefore found by summing the individual inductances. When coils are wired in parallel, the current distributes itself according to the inductances; therefore, if L1 is the reluctance of the first coil, L2 is the reluctance of the second coil, and so on, the total reluctance L of the network is found from the formula:

```
1/L = 1/L1 + 1/L2 + 1/L3. . .
```

This is shown in Figure 14-21. In reality, differences between the coils (such as their electrical resistance), and magnetic interaction between the coils, will complicate this simple relationship.

Time Constant

Just as the *time constant* of a capacitor defines the rate at which it accumulates voltage when power is applied through a resistor, the time constant of an inductor defines the rate at which it gradually allows amperage to pass through it, overcoming the EMF generated by the coil. In both cases, the time constant is the number of seconds that the component requires to acquire approximately 63% of the difference between its current value and its maximum value. In the case of an inductor, suppose we assume zero internal resistance in the power source, zero resistance in

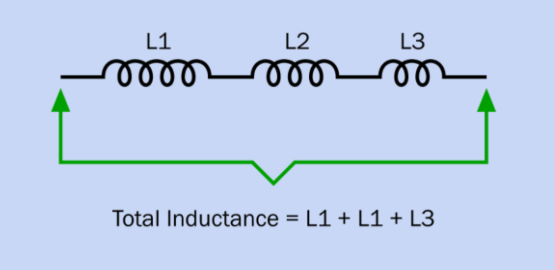

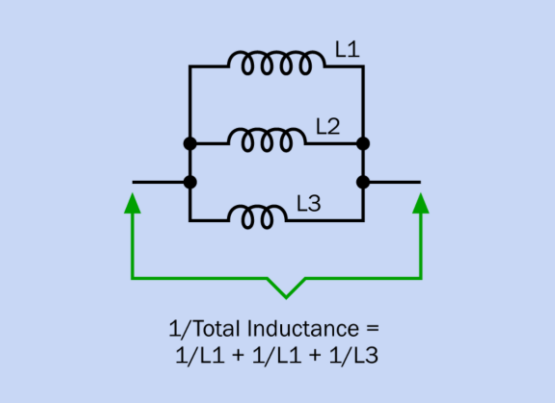

Figure 14-21. *Calculating the total inductance of inductors in parallel (top) and series (bottom).*

the coil windings, and an initial current of zero. If L is the inductance of the coil and R is the value of the series resistor, then the time constant—TC—is given in seconds by the formula

```
TC = L / R
```

Therefore a coil of 10 millihenrys (0.01 Henry) in series with a 100-ohm resistor will pass 63% of the full current in 0.0001 seconds, or 1/10 of a millisecond; it will take an equal additional amount of time for the current to rise by another 63% of the remaining difference between its charge and the maximum amperage of the circuit. In theory, the reactance of a coil can never diminish to zero, but in practice, five time constants are considered adequate to allow maximum current flow.

How to Use it

Because the inductance of an inductor peaks as current increases, and then gradually diminishes, an inductor can be used to block or attenuate high frequencies. A circuit that does this is often referred to as a *low-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-22. A basic application could be the *crossover network* in a loudspeaker system, where high-frequency signals are blocked from a low-frequency driver and are diverted to a high-frequency driver.

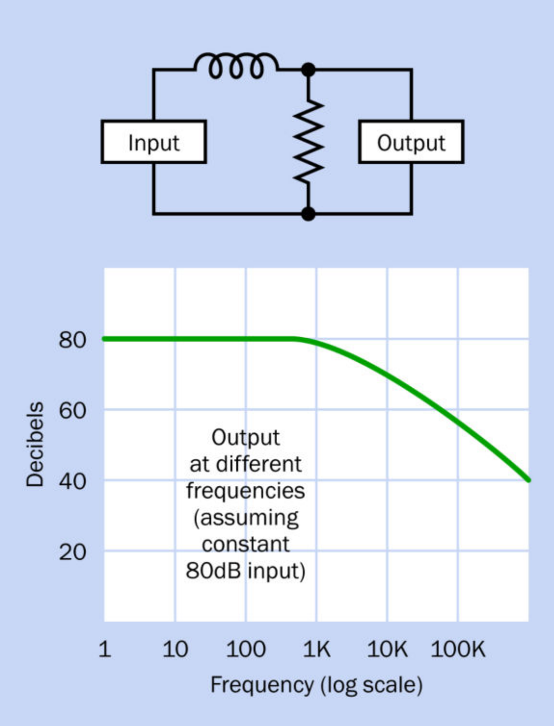

Figure 14-22. *By using the ability of an inductor to block a range of frequencies, a low-pass filter blocks higher frequencies.*

If the location of the inductor is shifted so that it shunts the signal away from the output, the results are reversed, and the circuit becomes a *high-pass filter*. The schematic and a graph suggesting its performance are shown in Figure 14-23.

Note that **capacitors** may also be used to create frequency filters, but because their function is roughly inverse to that of inductors, the place-

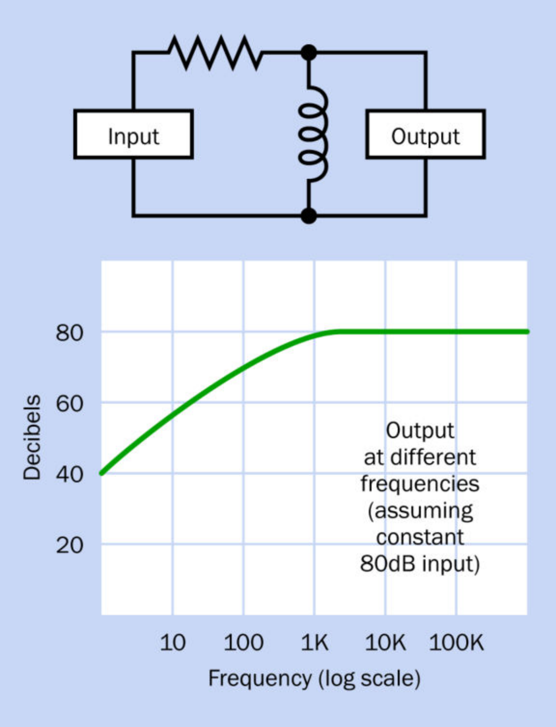

Figure 14-23. *Here the inductor diverts low frequencies away from the output, allowing high frequencies to pass through.*

ment of a capacitor in a circuit would be opposite to the placement of the inductor. Examples of filter circuits using capacitors are found in the entry for that component in this encyclopedia.

An inductor can be combined with a capacitor to form a *bandpass filter*, as shown in Figure 14-24. In this configuration, the inductor blocks the high frequencies while the capacitor blocks the low frequencies, allowing only a limited band of frequencies to get through.

Once again if the location of the components is shifted to shunt the signal away from the output, the results are reversed, as shown in Figure 14-25. This is known as a *notch filter*.

The performance of these filters will depend on the component values, and in most applications, additional components will be necessary. Sophisticated filter circuits are outside the scope of this encyclopedia.

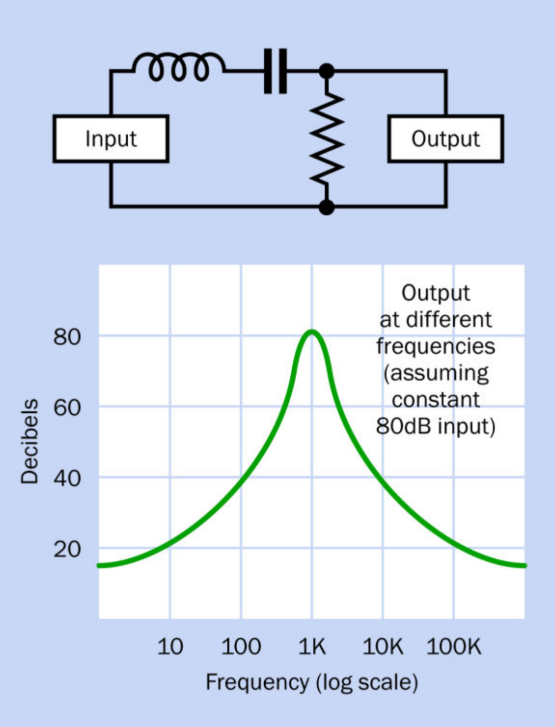

Figure 14-24. *If the values of a capacitor and an inductor are correctly chosen, and the components are placed in series, the inductor blocks high frequencies while the capacitor blocks low frequencies, creating a bandpass filter, in which only a narrow band of frequencies can get through.*

Inductors are of great importance in **DC-DC converters** and **AC-DC power supplies** where voltage changes are enabled by rapid switching. See the relevant entries of this encyclopedia for additional details.

Generally, as electronic equipment has become increasingly miniaturized, the unavoidable bulk of inductors has limited their application. However they may still be used to tune oscillators, to block sudden spikes in power supplies, and to protect equipment from sudden voltage spikes (they are used, for example, in surge suppressors for computing equipment).

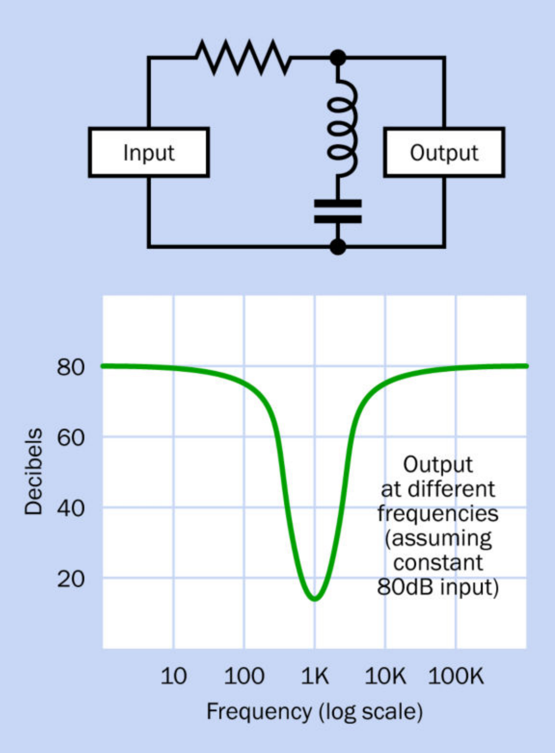

Figure 14-25. *Here the capacitor and inductor block all frequencies except a narrow band, which they divert from the output. The result is a notch filter.*

Core Choices

Air-cored inductors have relatively low inductance, because of their low permeability. However, they can be operated at very high frequencies up to the gigahertz range, and can tolerate higher peak currents.

Inductors with an iron core suffer increasing power losses due to hysteresis and eddy currents as the AC frequency passing through the inductor increases. Consequently, iron-cored inductors are not suitable for frequencies much above 10KHz.

Miniaturization

A low-value inductor can be formed by etching a spiral onto a circuit board, in applications where size must be minimized. They may also be incorporated in integrated circuit chips. However, in small devices such as cellular phones, it is more common to use a coil substitute such as a *gyrator*, as described previously.

What Can Go Wrong

Real-World Defects

The theoretically ideal inductor has no resistance or capacitance and suffers no energy losses. In reality, an inductor possesses both resistance and capacitance, also creates electrical noise, and may pick up electrical noise. It tends to create stray magnetic fields, and generally is more troublesome to deal with than its two cousins, the **resistor** and the **capacitor**.

Parasitic capacitance occurs between adjacent turns of wire. This capacitance becomes more significant at higher frequencies, leading ultimately to a situation where the coil becomes *self resonant*.

The workarounds for these problems involve coil geometries and choices of core material that go beyond the scope of this encyclopedia.

A *gyrator* should be considered as a possible substitute where inductors are troublesome or excessively expensive.

Saturation

Inductance increases as the current passing through a coil increases, but if a magnetic core is used, its contribution to inductance will stop abruptly when the core becomes magnetically *saturated*. In other words, when all of the randomly distributed *magnetic domains* in the core have been induced to align themselves with the pervasive magnetic field, the core cannot become more highly magnetized, and ceases to contribute to the inductance. Note that as a core approaches saturation levels, its hysteresis increases because reversing its magnetization requires greater energy. Antidotes to saturation would include a larger core, a lower current, a smaller number of turns in the coil, and using a core with lower permeability (such as air).

RF Problems

Radio frequencies (*RF*) introduce various problems affecting the efficiency of inductors. The *skin effect* is the tendency of high-frequency AC current to flow primarily on the surface of a strand of wire. The *proximity effect* refers to the tendency of the magnetic fields caused by adjacent wires to introduce eddy currents in the coil. Both of these effects increase the effective resistance of the coil. Various coil geometries have been developed to minimize these effects, but are outside the scope of this encyclopedia. The fundamental lesson is that coils specifically designed for RF are the only ones that should be used with RF.

AC-AC transformer

15

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **DC-DC converter** (See Chapter 17)
- **DC-AC inverter** (See Chapter 18)

What It Does

A transformer requires an input of *alternating current* (AC). It transforms the input voltage to one or more output voltages that can be higher or lower.

Transformers range in size from tiny impedance-matching units in audio equipment such as microphones, to multi-ton behemoths that supply high voltage through the national power grids. Almost all electronic equipment that is designed to be powered by municipal AC in homes or businesses requires the inclusion of a transformer.

Two small power transformers are shown in Figure 15-1. The one at the rear is rated to provide 36VAC at 0.8A when connected with a source of 125VAC. At front, the miniature transformer is a Radio Shack product designed to provide approximately 12VAC at 300mA, although its voltage will be more than 16VAC when it is not passing current through a load.

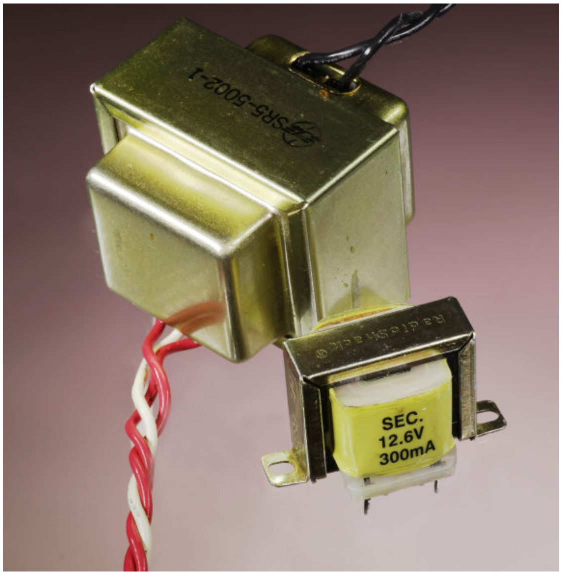

Figure 15-1. *Two small power transformers. The one at the rear measures approximately 1" × 2" × 2" and is rated to provide 36VAC at 0.8A. The term "sec" on the smaller unit is an abbreviation for "secondary," referring to the rating for its secondary winding.*

Transformer schematic symbols are shown in Figure 15-2. The different coil styles at left and right are functionally identical. Top: A transformer with a magnetic core—a core that can be magnetized. Bottom: A transformer with an air core. (This type of transformer is rare, as it tends to be less efficient.) The input for the transformer is almost always assumed to be on the left, through the *primary* coil, while the output is on the right, through the *secondary* coil. Often the two coils will show differing numbers of turns to indicate whether the transformer is delivering a reduced voltage (in which case there will be fewer turns in the secondary coil) or an increased voltage (in which case there will be fewer turns in the primary coil).

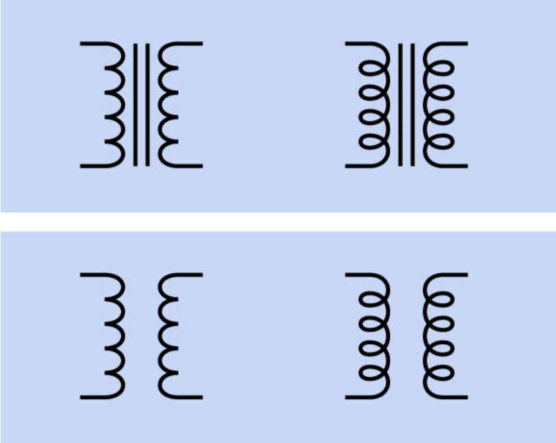

Figure 15-2. *Alternate symbols for a transformer with a ferromagnetic core (top) and air core (bottom). The differing coil symbols at left and right are functionally identical.*

How It Works

A simplified view of a transformer is shown in Figure 15-3. Alternating current flowing through the primary winding (orange) induces magnetic flux in a laminated core formed from multiple steel plates. The changing flux induces current in the secondary winding (green), which provides the output from the transformer. (In reality, the windings usually consist of thousands of turns of thin magnet wire, also known as enameled wire; and various different core configurations are used.)

The process is known as *mutual induction*. If a load is applied across the secondary winding, it will draw current from the primary winding, even though there is no electrical connection between them.

In an ideal, lossless transformer, the ratio of turns between the two windings determines whether the output voltage is higher, lower, or the same as the input voltage. If V_p and V_s are the voltages across the primary and secondary windings respectively, and N_p and N_s are the number of turns of wire in the primary and secondary windings, their relationship is given by this formula:

$$V_p \;/\; V_s = N_p \;/\; N_s$$

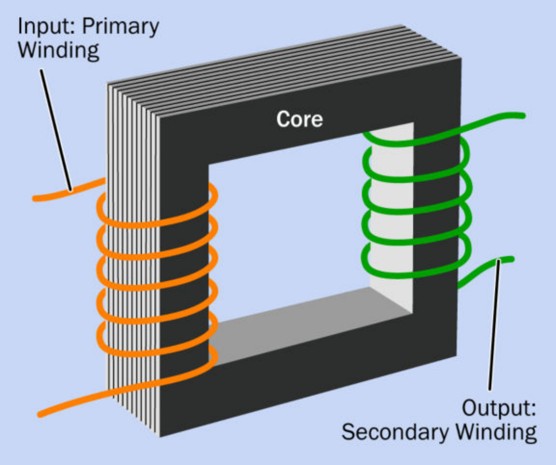

Figure 15-3. *Three basic parts of a transformer, shown in simplified form.*

A simple rule to remember is that fewer turns = lower voltage while more turns = higher voltage.

A *step-up transformer* has a higher voltage at its output than at its input, while a *step-down transformer* has a higher voltage at its input than at its output. See Figure 15-4.

In an ideal, lossless transformer, the power input would be equal to the power output. If V_{in} and V_{out} are the input and output voltages, and I_{in} and I_{out} are the input and output currents, their relationship is given by this formula:

$$V_{in} * I_{in} = V_{out} * I_{out}$$

Therefore, if the transformer doubles the voltage, it allows only half as much current to be drawn from the secondary winding; and if the voltage is cut in half, the available current will double.

Transformers are not 100% efficient, but they can be more than 98% efficient, and relationships between voltage, current, and the number of turns in the windings are reasonably realistic.

When the transformer is not loaded, the primary winding behaves like a simple inductor with reactance that inhibits the flow of current. Therefore a power transformer will consume relatively

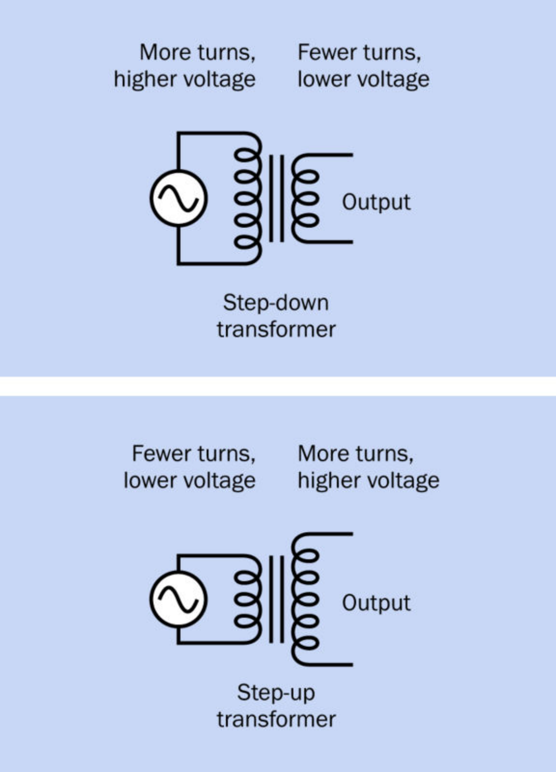
More turns, higher voltage
Fewer turns, lower voltage
Output
Step-down transformer
Fewer turns, lower voltage
More turns, higher voltage
Output
Step-up transformer

Figure 15-4. *The ratio of input voltage to output voltage is equal to the ratio of primary turns to secondary turns in the transformer windings, assuming a transformer of 100% efficiency.*

little electricity if it is left plugged in to an electrical outlet without any load connected to its output side. The power that it does consume will be wasted as heat.

The Core

The ferromagnetic core is often described as being made of iron, but in reality is more often fabricated from high permeability silicon steel. To reduce losses caused by eddy currents, the core is usually laminated—assembled from a stack of plates separated from each other by thin layers of varnish or a similar insulator. Eddy currents tend to be constrained within the thickness of each plate.

Because a DC voltage would cause magnetic saturation of the core, all transformers must operate with alternating current or pulses of current. The windings and geometry of a transformer are optimized for the frequency range, voltage, and current at which it is designed to operate. Deviating significantly from these values can damage the transformer.

Taps

A *tap* on a transformer is a connection part-way through the primary or (more often) the secondary coil. On the primary side, applying an input between the start of a coil and a tap part-way through the coil will reduce the number of turns to which the voltage is applied, therefore increasing the ratio of output turns to input turns, and increasing the output voltage. On the secondary side, taking an output between the start of a coil and a tap part-way through the coil will reduce the number of turns from which the voltage is taken, therefore decreasing the ratio of output turns to input turns, and decreasing the output voltage. This can be summarized:

- A tap on the primary side can increase output voltage.
- A tap on the secondary side can provide a decreased output voltage.

In international power adapters, a choice of input voltages may be allowed by using a double-throw switch to select either the whole primary winding, or a tapped subsection of the winding. See Figure 15-5. Modern electronics equipment often does not require a voltage adapter, because a **voltage regulator** or **DC-DC converter** inside the equipment will tolerate a wide range of input voltages while providing a relatively constant output voltage.

A transformer's secondary winding is often tapped to provide a choice of output voltages. In fact, most power transformers have at least two outputs, since the cost of adding taps to the secondary winding is relatively small. As an

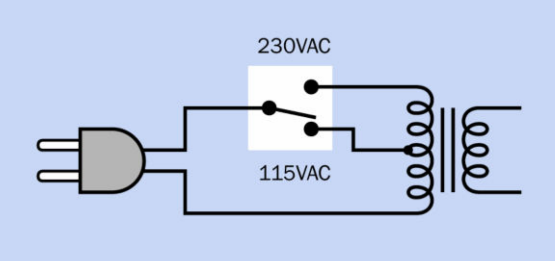

Figure 15-5. *An international power adapter can provide a fixed output voltage by using a double-throw switch to apply 230VAC voltage across a transformer's primary winding, or 115VAC to a tapped midpoint of the primary winding.*

alternative to tapped outputs, two or more separate secondary windings may be used, allowing the outputs to be electrically isolated from each other. See Figure 15-6.

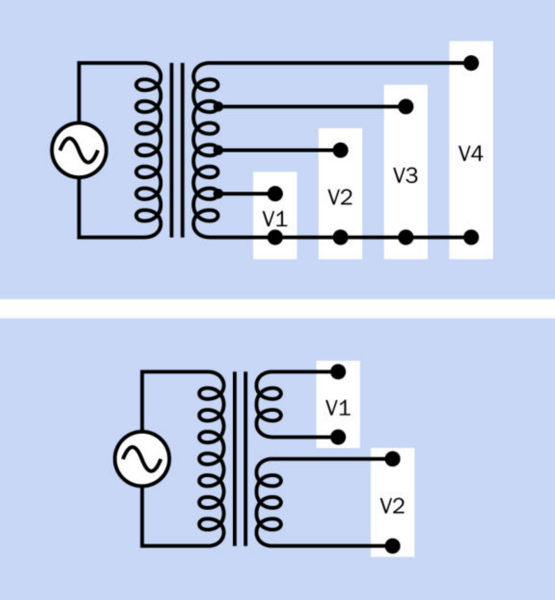

Figure 15-6. *Multiple output voltages may be obtained from a transformer by tapping into the secondary winding (top) or using two or more separate secondary windings (bottom), in which case the outputs will be electrically isolated from each other.*

If the winding on the primary side of a transformer is coiled in the same direction as the winding on the secondary side, the output voltage will be 180 degrees out of phase with the input voltage. In schematics, a dot is often placed at one end of a transformer coil to indicate where the coil begins. If the dots on the primary and secondary sides are at the same ends of the coils, there will be a 180 degree phase difference between input and output. For many applications (especially where the output from a power transformer is going to be converted to DC), this is immaterial.

If there is a center tap on the secondary winding, and it will be referenced as ground, the voltages relative to it, at opposite ends of the secondary winding, will be out of phase. See Figure 15-7.

Variants

Core Shapes

The *shell core* is a closed rectangle, as shown in Figure 15-3. This is the most efficient but most costly to manufacture. A C-shaped core is another option (three sides of the rectangle) and an E-I core is popular, consisting of a stack of E-shaped plates with two coils wound around the top and bottom legs of the E, or wound concentrically around the center leg of the E. An additional stack of straight plates is added to close the gaps in the E and form a magnetic circuit.

In Figure 15-8, the small transformer from Figure 15-1 has been sliced open with a band saw and a belt sander to reveal a cross-section of its windings. This clearly shows that its primary and secondary windings are concentric. It also reveals the configuration of its core, which is in the E-I format. In Figure 15-9, the E-I configuration is highlighted to show it more clearly.

Power Transformer

Typically designed to be bolted onto a chassis or secured inside the case or cabinet housing a piece of electrical equipment with solder tabs or connectors allowing wires to connect the transformer to the power cord, on one side, and a cir-

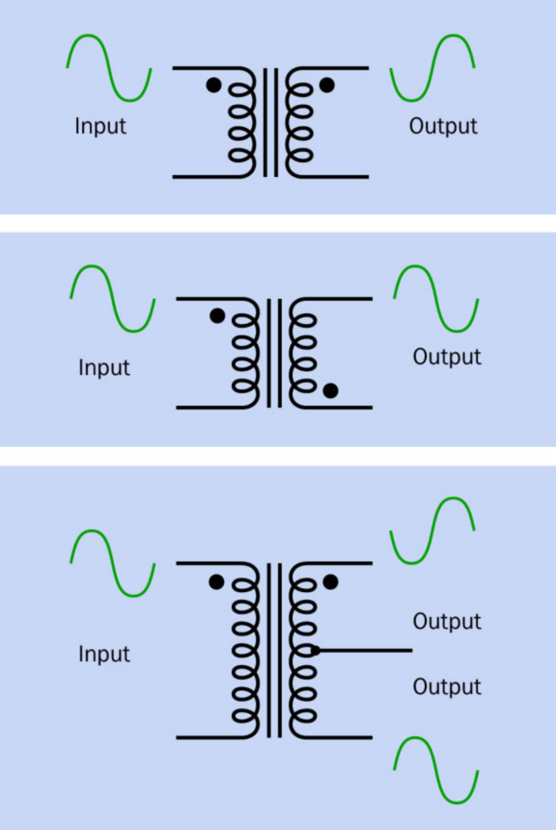

Figure 15-7. *A dot indicates the start of each winding. Where primary and secondary windings are in the same direction, the voltage output will be 180 degrees out of phase with the input. Where the dots indicate windings in opposite directions, the voltage output will have the same phase as the input. Where a center tap on the secondary winding serves as a common ground, the voltages at opposite ends of the secondary winding will be opposite to each other in phase.*

cuit board, on the other side. Smaller power transformers such as the one in Figure 15-1 have "through-hole" design with pins allowing them to be inserted directly onto circuit boards.

Plug-in Transformer

Usually sealed in a plastic housing that can be plugged directly into a wall power outlet. They are visually identical to **AC adapters** but have an AC output instead of a DC output.

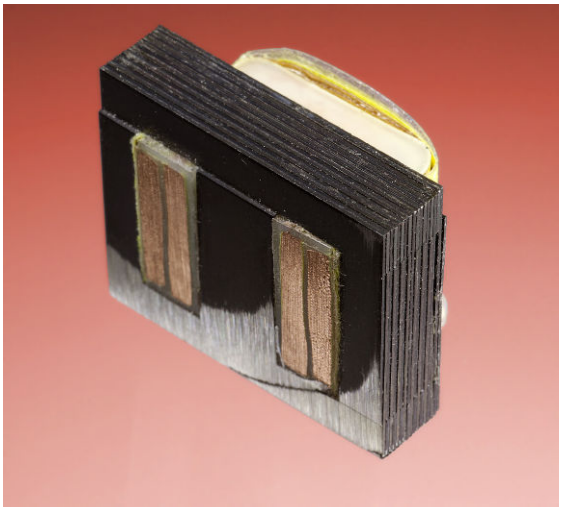

Figure 15-8. *The small transformer from the first figure in this entry is shown sliced open to reveal its internal configuration.*

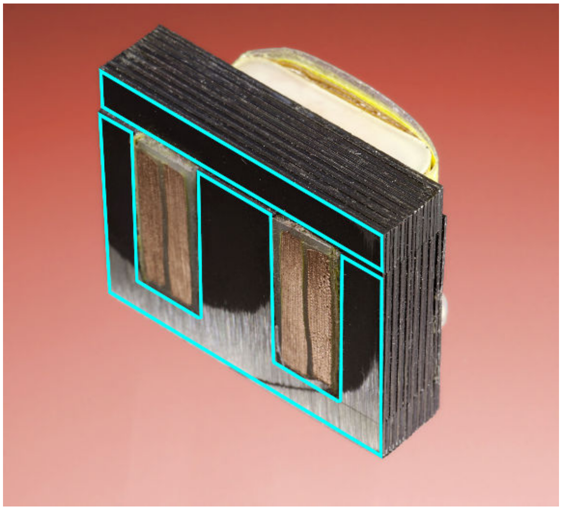

Figure 15-9. *The "EI" shaped plates that form the core of the transformer are outlined to show their edges.*

Isolation Transformer

Also known as a *1:1 transformer* because it has a 1:1 ratio between primary and secondary windings, so that the output voltage will be the same as the input voltage. When electrical equipment is plugged into the isolation transformer, it is separated from the electrical ground of AC power wiring. This reduces risk when working on "live" equipment, as there will be negligible electrical potential between itself and ground. Consequently, touching a grounded object while also touching a live wire in the equipment should not result in potentially lethal current passing through the body.

Autotransformer

This variant uses only one coil that is tapped to provide output voltage. Mutual induction occurs between the sections of the coil. An autotransformer entails a common connection between its input and output, unlike a two-coil transformer, which allows the output to be electrically isolated from the input. See Figure 15-10. Autotransformers are often used for impedance matching in audio circuits, and to provide output voltages that differ only slightly from input voltages.

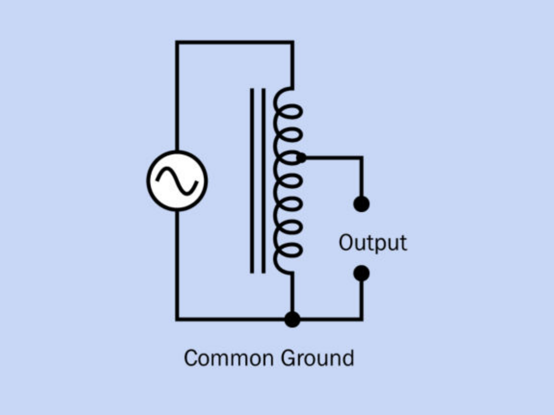

Figure 15-10. *An autotransformer contains only one coil and core. A reduced output voltage can be obtained by tapping into the coil. A common connection prevents the output from being electrically isolated from the input.*

Variable Transformer

A variable transformer, also known as a *variac*, resembles a wire-wound **potentiometer**. Only one winding is used. A *wiper* can be turned to contact the winding at any point, and serves as a movable tap. Like an autotransformer, a variable transformer entails a common connection between input and output.

Audio Transformer

When a signal is transmitted between two stages of a circuit that have different impedance, the signal may be partially reflected or attenuated. (Impedance is measured in ohms but is different from DC electrical resistance because it takes into account reactance and capacitance. It therefore varies with frequency.)

A device of low input impedance will try to draw significant current from a source, and if the source has high output impedance, its voltage will drop significantly as a result. Generally, the input impedance of a device should be at least 10 times the output impedance of the device that is trying to drive it. Passive components (resistors, and/or capacitors, and/or coils) can be used for impedance matching, but in some situations a small transformer is preferable.

If N_p and N_s are the number of turns of wire in the transformer primary and secondary windings, and Z_p is the impedance of a device (such as an audio amplifier) driving the transformer on its primary side, and Z_s is the impedance of a device (such as a loudspeaker) receiving power from the secondary side:

$$N_p \;/\; N_s = \sqrt{(Z_p \;/\; Z_s)}$$

Suppose that an audio amplifier with rated output impedance of 640Ω is driving a loudspeaker with 8Ω impedance. A matching transformer would be chosen with a ratio of primary turns to secondary turns give by:

$$\sqrt{(640/8)} = \sqrt{80} = \text{approximately } 9:1$$

The two transformers in Figure 15-11 are through-hole components designed for telecommunications purposes, but are capable of passing audio frequencies and can be used for impedance matching in applications such as a preamplifier.

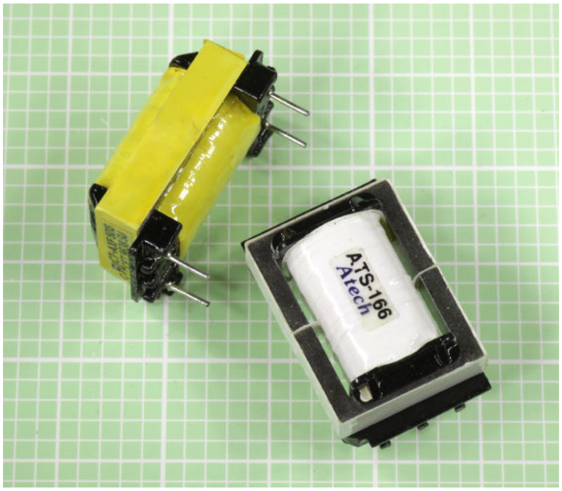

Figure 15-11. *Through-hole transformers. See text for details.*

In Figure 15-12, the transformers are designed for audio coupling. The one on the right has impedances of 500 ohms (primary) and 8 ohms (secondary). On the left is a fully encapsulated line matching transformer with a 1:1 turns ratio.

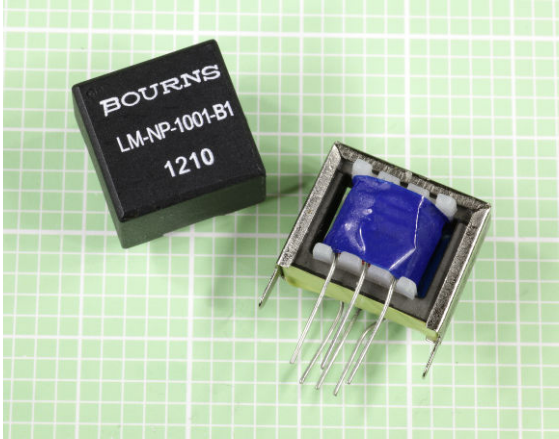

Figure 15-12. *Through-hole transformers. See text for details.*

Split-Bobbin Transformer

This variant has primary and secondary coils mounted side by side to minimize capacitive coupling.

Surface-Mount Transformer

May be less than 0.2" square and is used for impedance matching, line coupling, and filtering. Two surface-mount transformers are shown in Figure 15-13.

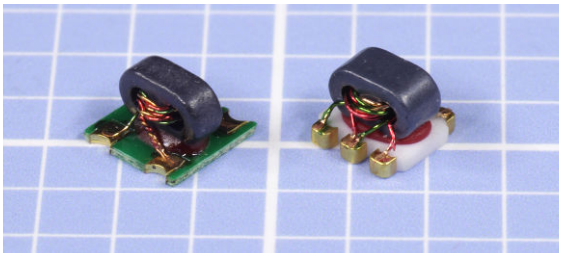

Figure 15-13. *Two surface-mount transformers, each measuring less than 0.2" square, typically used in communications equipment and suited for frequencies higher than 5 MHz.*

Values

When selecting a power transformer, its power handling capability is the value of primary interest. It is properly expressed by the term VA, derived from "volts times amps." VA should not be confused with watts because watts are measured instantaneously in a DC circuit, whereas in an AC circuit, voltage and current are fluctuating constantly. VA is actually the *apparent power*, taking reactance into account.

The relationship between VA and watts will vary depending on the device under consideration. In a worst-case scenario:

```
W = 0.65 VA (approximately)
```

In other words, the averaged power you can draw from a transformer should be no less than two-thirds of its VA value.

Transformer specifications often include input voltage, output voltage, and weight of the component, all of which are self-explanatory. Coupling transformers may also specify input and output impedances.

How to Use it

For most electronic circuits, a power transformer will be followed by a *rectifier* to convert AC to DC, and **capacitors** to smooth fluctuations in the supply. Using a prepackaged **power supply** or *AC adapter* that already contains all the necessary components will be more time-effective and probably more cost-effective than building a power supply from the ground up. See Chapter 16.

What Can Go Wrong

Reversal of Input and Output

Suppose a transformer is designed to provide an output voltage of 10 volts from domestic AC power of 115 volts. If the wrong side of the transformer is connected with 115VAC by mistake, the output will now be more than 1,000 volts—easily enough to cause death, quite apart from destroying components that are connected with it. Reversing the transformer in this way may also destroy it. Extreme caution is advisable when making connections with power transformers. A meter should be used to check output voltage. All devices containing transformers should be fused on the live side and grounded.

Shock Hazard from Common Ground

When working on equipment that uses an autotransformer, the chassis will be connected through the transformer to one side of 115VAC power. So long as a plug is used that prevents reversed polarity, the chassis should be "neutral." However, if an inappropriate power cord is used, or if the power outlet has been wired incorrectly, the chassis can become live. For protection, before working on any device that uses 115VAC power with an autotransformer, plug the device into an isolation transformer, and plug the isolation transformer into the wall outlet.

Accidental DC Input

If DC current is applied to the input side of a transformer, the relatively low resistance of the primary coil will allow high current that can destroy the component. Transformers should only be used with alternating current.

Overload

If a transformer is overloaded, heat will be generated that may be sufficient to destroy the thin layers of insulation between coil windings. Consequently, input voltage can appear unexpectedly on the output side. Transformers with a toroidal (circular) core are especially hazardous in this respect, as their primary and secondary windings usually overlap.

Some (not all) power transformers contain a thermal fuse that melts when it exceeds a temperature threshold. If the fuse is destroyed, the transformer must be discarded.

The consequences of moderate overloading may not be obvious, and can be cumulative over time. Ventilation or heat sinkage should be taken into account when designing equipment around a power transformer.

Incorrect AC Frequency

Single-phase AC power in the United States fluctuates at 60Hz, but Great Britain and some other countries use AC power at 50Hz. Many power transformers are rated to be compatible with either frequency, but if a transformer is specifically designed for 60Hz, it may eventually fail by overheating if it is used with a 50Hz supply. (A 50Hz transformer can be used safely with 60Hz AC.)

AC-DC power supply

16

Also known as an *AC adapter*. When packaged as a palm-sized plastic package that plugs directly into a power outlet, it is occasionally known colloquially as a *wall-wart*.

OTHER RELATED COMPONENTS

- **transformer** (See Chapter 15)
- **DC-DC converter** (See Chapter 17)
- **DC-AC inverter** (See Chapter 18)

What It Does

An AC-DC power supply converts *alternating current* (*AC*) into the *direct current* (*DC*) that most electronic devices require, usually at a lower voltage. Thus, despite its name, a power supply actually requires an external supply of power to operate.

Larger products, such as computers or stereo equipment, generally have a power supply contained within the device, enabling it to plug directly into a wall outlet. Smaller battery-powered devices, such as cellular phones or media players, generally use an external power supply in the form of a small plastic pod or box that plugs into a wall outlet and delivers DC via a wire terminating in a miniature connector. The external type of power supply is often, but not always, referred to as an *AC adapter*.

Although an AC-DC power supply is not a single component, it is often sold as a preassembled modular unit from component suppliers.

Variants

The two primary variants are a *linear regulated power supply* and *switching power supply*.

Linear Regulated Power Supply

A linear regulated power supply converts AC to DC in three stages:

1. A power **transformer** reduces the AC input to lower-voltage AC.
2. A *rectifier* converts the AC to unsmoothed DC. Rectifiers are discussed in the entry on **diodes** in this encyclopedia.
3. A **voltage regulator**, in conjunction with one or more **capacitors**, controls the DC voltage, smooths it, and removes *transients*. The regulator is properly known as a *linear voltage regulator* because it contains one or more **transistors**, which are functioning in linear mode—that is, responding linearly to fluctuations in base current, at less than their saturation level. The linear voltage regulator gives the linear regulated power supply its name.

A simplified schematic of a linear regulated power supply is shown in Figure 16-1.

This type of power supply may be described as *transformer-based*, since its first stage consists of a transformer to drop the AC input voltage before it is rectified.

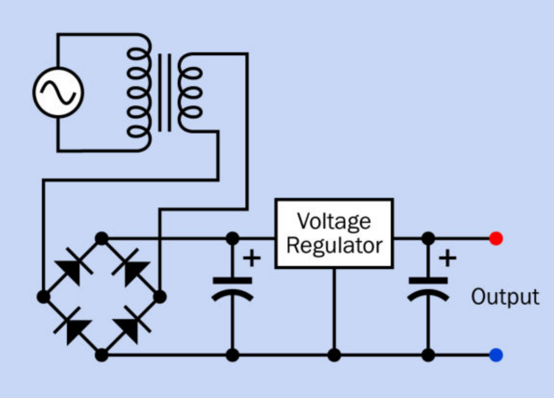

Figure 16-1. *A basic linear regulated power supply.*

Because the rectifier in a power supply generally passes each pulse of AC through a pair of silicon diodes, it will impose a voltage drop of about 1.2V at peak current. A smoothing capacitor will drop the voltage by about 3V as it removes ripple from the current, whereas a voltage regulator typically requires a difference of at least 2V between its input and its output. Bearing in mind also that the AC input voltage may fluctuate below its rated level, the output from the power transformer should be at least 8VAC higher than the ultimate desired DC output. This excess power will be dissipated as heat.

The basic principle of the linear regulated power supply originated in the early days of electronic devices such as radio receivers. A transistorized version of this type of power supply remained in widespread use through the 1990s. Switching power supplies then became an increasingly attractive option as the cost of semiconductors and their assembly decreased, and high-voltage transistors became available, allowing the circuit to run directly from rectified line voltage with no step-down power transformer required.

Some external AC adapters are still transformer-based, but are becoming a minority, easily identified by their relatively greater bulk and weight. An example is shown in Figure 16-2.

Figure 16-2. *A simple transformer-based power supply can be encapsulated in a plastic shell, ready to plug into a power outlet. However, today this format more typically contains a switching power supply, which is usually lighter, smaller, and cheaper.*

Figure 16-3 shows the handful of components inside a cheap, relatively old AC adapter. The output from a power transformer is connected directly to four diodes (the small black cylinders), which are wired as a full-wave rectifier. A single electrolytic capacitor provides some smoothing, but because there is no voltage regulator, the output will vary widely depending on the load. This type of AC adapter is not suitable for powering any sensitive electronic equipment.

Switching Power Supply

Also known as a *switched-mode power supply*, an *SMPS,* or *switcher*, it converts AC to DC in two stages.

1. A *rectifier* changes the AC input to unsmoothed DC, without a power transformer.
2. A DC-DC **converter** switches the DC on and off at a very high frequency using *pulse-width modulation* to reduce its average effective voltage. Often the converter will be the *flyback* type, containing a transformer, but the high-frequency switching allows the transformer to be much smaller than the power transformer required in a linear regulated power supply. See the DC-DC **converter** entry in this encyclopedia for an explanation of the working principles.

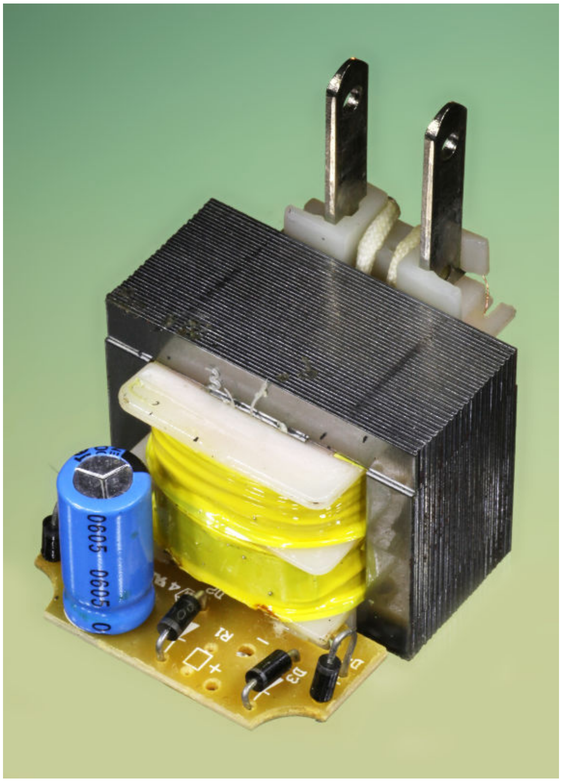

Figure 16-3. *A relatively old, cheap AC adapter contains only the most rudimentary set of components, and does not supply the kind of properly regulated DC power required by electronic equipment.*

A simplified schematic of a switching power supply is shown in Figure 16-4.

The interior of a relatively early switching power supply designed to deliver 12VDC at up to 4A is shown in Figure 16-5. This supply generated considerable waste heat, necessitating well-spaced components and a ventilated enclosure.

The type of small switching power supply that is now almost universally used to power laptop computers is shown in Figure 16-6. Note the smaller enclosure and the higher component count than in the older power supply shown in Figure 16-5. The modern unit also delivers considerably more power, and generates less waste heat. Although this example is rated at 5A, the transformer (hidden under the yellow wrapper at the center of the unit) is smaller than the power transformer that would have been found in an old-style AC adapter delivering just 500mA.

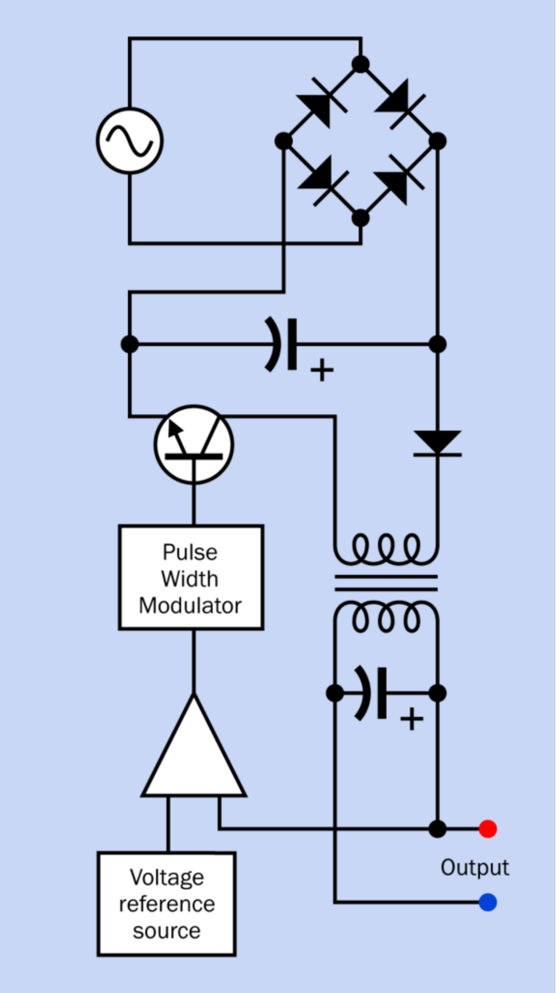

Figure 16-4. *Greatly simplified schematic showing the principal components of a switching power supply. Note the absence of a 115VAC power transformer. The transformer that is inserted subsequently in the circuit functions in conjunction with the high switching frequency, which allows it to be very much smaller, cheaper, and lighter.*

The modern power supply is completely sealed, where earlier versions required ventilation. On

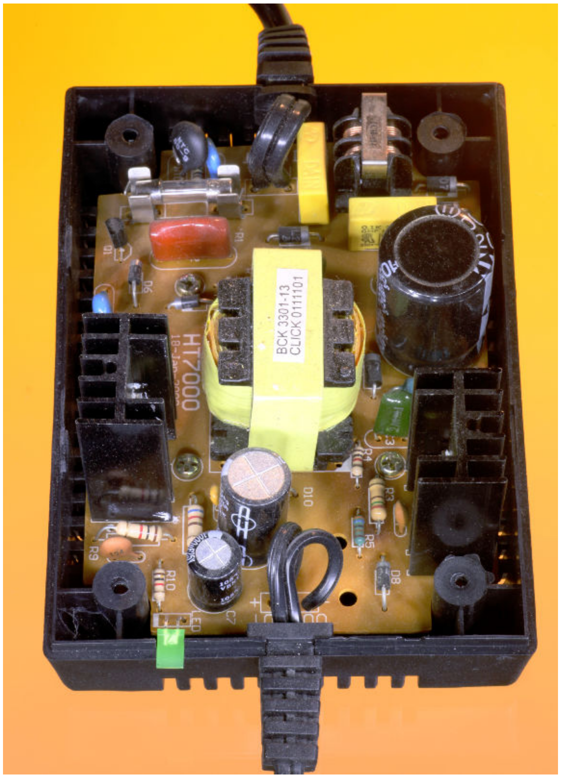

Figure 16-5. *The interior of an early switching power supply.*

the downside, the plastic case of the switching supply requires a metal liner (removed for this photograph) to contain high-frequency electromagnetic radiation.

Unregulated Power Supply

Typically this consists of a transformer and rectifying diodes with little or no smoothing or voltage control of the output.

Adjustable Power Supply

This is usually a linear power supply incorporating an adjustable **voltage regulator**. This type of supply has laboratory applications and is found as a benchtop item to power electronics design projects during their development.

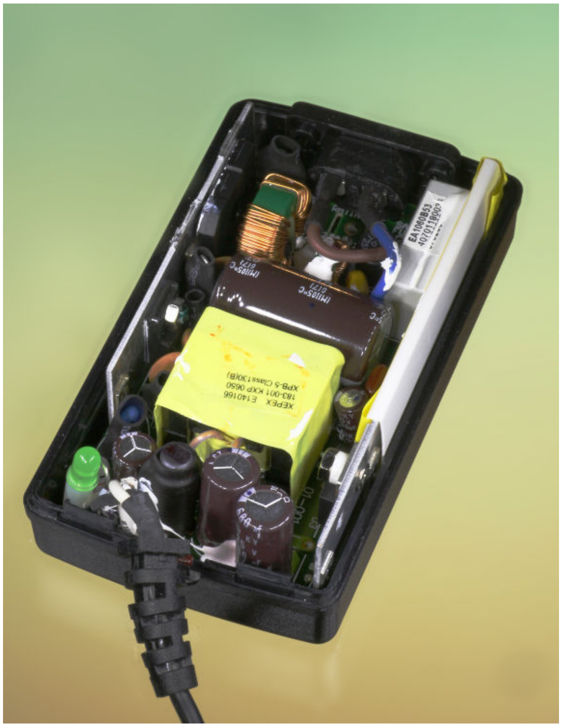

Figure 16-6. *The interior of the type of switching power supply that powers a laptop computer.*

Voltage Multiplier

Devices such as photocopiers and laser printers, televisions, cathode-ray tubes, and microwave ovens require voltages significantly higher than those supplied by domestic AC power outlets. A voltage multiplier usually contains a step-up transformer followed by DC conversion components, but detailed consideration is outside the scope of this encyclopedia.

Formats

An *open frame* power supply consists of components on a circuit board, usually mounted on a metal chassis, with no enclosure or fan cooling.

A *covered* power supply is enclosed in a protective perforated metal box with a cooling fan if needed. Power supplies sold for desktop computers are usually in this format.

Power supplies are also available in rack-mount and DIN-rail formats.

How to Use it

Because a switching power supply contains no power transformer, it is lighter and smaller, and may be cheaper than a linear power supply. It is also more efficient and generates less waste heat. These advantages have made switching power supplies the most popular option to provide DC power for electronics devices. However, the high-frequency switching tends to create *electromagnetic interference* (*EMI*), which must be filtered to protect the output of the device and also to minimize the risk of this interference feeding back into AC power wiring. The high-frequency switched power may also generate harmonics, which must be suppressed.

High-quality linear regulated power supplies still find application in laboratory equipment, low-noise signal processing, and other niches where excellent regulation and low-ripple output are necessary. They are relatively heavy, bulky, and inefficient.

See Figure 16-7 for a chart comparing the advantages and disadvantages of linear and switching power supplies.

What Can Go Wrong

High Voltage Shock

One or more capacitors in a power supply may retain a relatively high voltage for some time after the unit has been unplugged. If the power supply is opened for inspection or repairs, caution is necessary when touching components.

Capacitor Failure

If *electrolytic capacitors* fail in a switching power supply (as a result of manufacturing defects, disuse, or age), allowing straight-through conduction of alternating current, the high-frequency switching semiconductor can also fail, allowing input voltage to be coupled unexpectedly to the output. Capacitor failure is also a potential problem in linear power supplies. For additional information on capacitor failure modes, see Chapter 12.

	Switching power supply	Linear power supply
Component count	High	Medium
Load regulation	0.05% to 0.5%	0.005% to 0.2%
Line regulation	0.05% to 0.2%	0.005% to 0.05%
Ripple (RMS)	10mV to 25mV	0.25mV to 1.5mV
Efficiency	70% to 85%	40% to 60%
EMI	High	Very low
Leakage	High	Low
Physical size	Small	Large
Weight	Light	Heavy
Heat management	Usually fan-cooled	Usually cooled by convection

Figure 16-7. *Comparison of attributes of linear regulated power supplies and switching power supplies. (Adapted from Acopian Technical Company.)*

Electrical Noise

If *electrolytic capacitors* are used, their gradual deterioration over time will permit more electrical noise associated with high-frequency switching in a switching power supply.

Peak Inrush

A switching power supply allows an initial inrush or surge of current as its capacitors accumulate their charge. This can affect other components in the circuit, and requires fusing that tolerates brief but large deviations from normal power consumption.

DC-DC converter

17

Often referred to as a *switching regulator*, and sometimes as a *switcher*, not to be confused with a *switching power supply*.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **voltage regulator** (See Chapter 19)
- **DC-AC inverter** (See Chapter 18)

What It Does

A DC-DC converter, often referred to simply as a *converter*, receives a DC voltage as its input and converts it to a regulated DC voltage as its output. The output voltage may be higher or lower than the input voltage, may be user-adjustable by adding an external resistor, and may be completely electrically isolated from the input, depending on the type of converter that is used. The overall efficiency is not greatly affected by the difference between input and output voltage, and can exceed 90%, minimizing waste heat and enabling the unit to be extremely compact.

A DC-DC converter is an integrated circuit package that includes a high-speed switching device (almost always, a *MOSFET*) in conjunction with an oscillator circuit, an inductor, and a diode. By comparison, a **linear regulator** is usually based around bipolar transistors. Its input must always be at a higher voltage than its output, and its efficiency will be inversely proportional with the voltage drop that it imposes. See the **voltage regulator** entry in this encyclopedia for additional information.

There is no single symbol to represent a DC-DC converter. Some simplified schematics showing the principles of operation of commonly used converters are referenced under the following Variants section.

A DC-DC converter is also typically found in the output stage of a switching **AC-DC power supply**.

How It Works

An internal oscillator controls a *MOSFET* semiconductor that switches the DC input on and off at a high frequency, usually from 50KHz to 1MHz. Output voltage is adjusted by varying the *duty cycle* of the oscillator—the length of each "on" pulse relative to each "off" interval. This is known as *pulse-width modulation*, or *PWM*. The duty cycle is controlled by sampling the output of the converter and using a comparator to subtract the output voltage from a reference voltage, to establish an error value. This is passed to another comparator, which subtracts the error voltage from an oscillator ramp signal. If the error increases, the oscillator signal is more heavily clipped, thus changing the effective ratio of on/off pulse lengths. A simplified schematic of the PWM circuit is shown in Figure 17-1, which omits other components for clarity. The system of subtracting an error voltage from a ramp oscillator voltage to obtain a pulse-width modulated signal is illustrated in Figure 17-2.

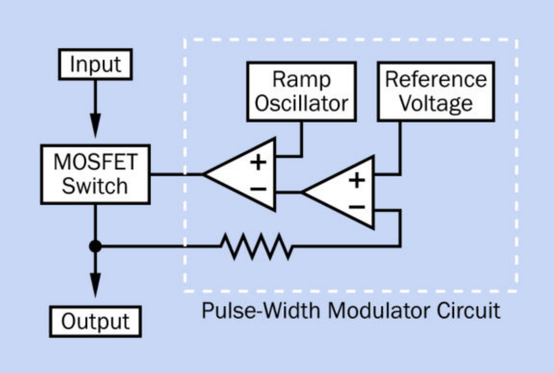

Figure 17-1. *The heart of a DC-DC converter is a MOSFET switch, which operates at a high frequency with pulse-width modulation used to create an adjustable DC output.*

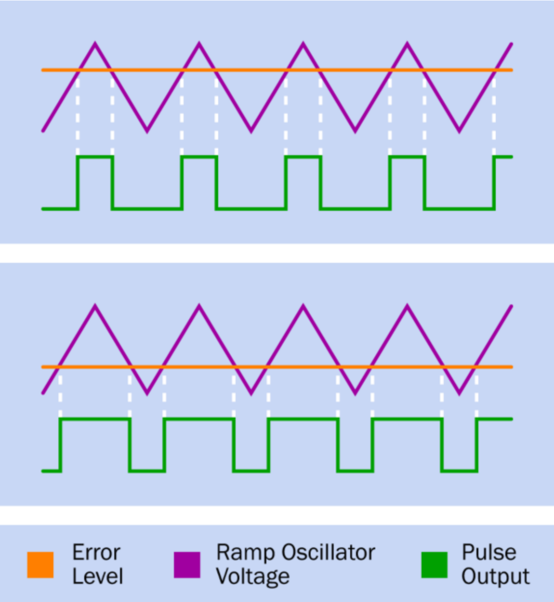

Figure 17-2. *To achieve pulse-width modulation, an error-level voltage is established by comparing the output from the converter with a reference voltage. The error level, shown as an orange line, is then subtracted from the output from a ramp oscillator. The pulse width varies accordingly.*

The key to the efficiency of a DC-DC converter is an inductor, which stores energy in its magnetic field during "on" pulse and releases it in the discharge phase. Thus, the inductor is used as a temporary reservoir and minimizes the ripple current. All converter variants use a coil for this purpose, although its placement varies in relation to the diode and capacitor that complete the basic circuit.

Variants

Four basic switching circuits are used in DC-DC converters and are defined in the coming sections, with a formula to determine the ratio between input voltage (V_{in}) and output voltage (V_{out}) in each case. In these formulae, variable D is the duty cycle in the pulse train generated through an internal MOSFET switch. The duty cycle is the fraction of the total on-off cycle that is occupied by each "on" pulse. In other words, if T_{on} is the duration of an "on" pulse and T_{off} is the "off" time:

$$D = T_{on} / (T_{on} + T_{off})$$

Buck Converter

See Figure 17-3. The output voltage is lower than the input voltage. The input and output share a common ground. For this circuit:

$$V_{out} = V_{in} * D$$

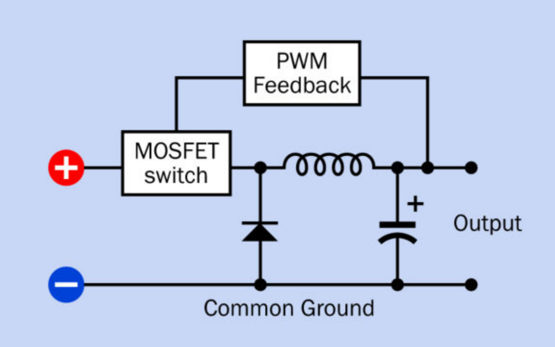

Figure 17-3. *Basic topology of a buck-type DC-DC converter.*

Boost Converter

See Figure 17-4. The output voltage is greater than the input voltage. The input and output share a common ground. For this circuit:

$$V_{out} = V_{in} / (1-D)$$

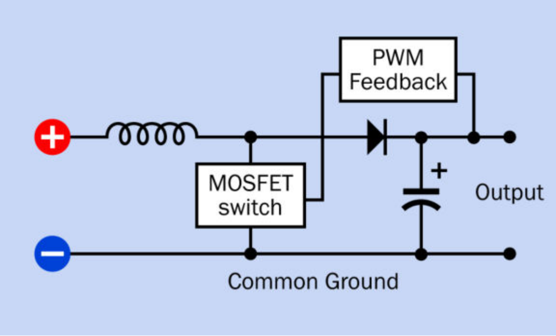

Figure 17-4. *Basic topology of a boost-type DC-DC converter.*

Flyback Converter with Inductor

Commonly known as a *buck-boost converter*. See Figure 17-5. The output voltage can be less than or greater than the input voltage. The input and output share a common ground. For this circuit:

$$V_{out} = V_{in} * (D / (1-D))$$

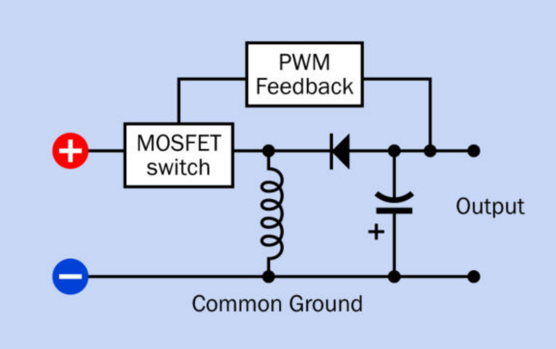

Figure 17-5. *Basic topology of a flyback-type DC-DC converter.*

Flyback Converter with Transformer

See Figure 17-6. The output voltage can be less than or greater than the input voltage. The input and output are isolated from one another. For this circuit:

$$V_{out} = V_{in} * (D / (1-D))$$

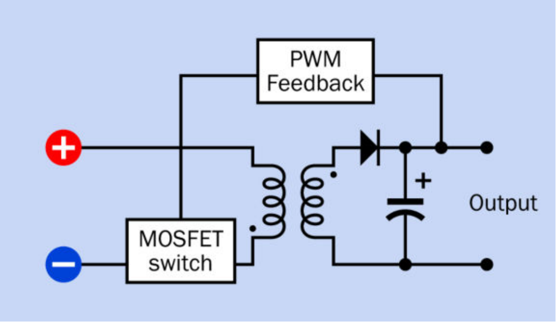

Figure 17-6. *Basic topology of a flyback-type DC-DC converter. (Buck, boost, and flyback topologies adapted from Maxim Integrated Products.)*

Using a transformer in the converter allows multiple outputs with different voltages, supplied through multiple transformer windings.

Formats

A converter may be packaged in a flat rectangular box that requires no additional heat sink and has pins for through-hole insertion into a PC board. Sizes usually range up to to 2″ × 2″. Power handling can range from 5 to 30 watts. Converters of this type are shown in Figure 17-7. (Top: Input range of 9 to 18VDC, fixed output of 5VDC at 3A completely isolated from the input. Typical efficiency of approximately 80%. The case is made of copper, providing good heat dissipation with electrical shielding. Center: Input range of 9 to 18VDC, fixed output of 5VDC at 500mA completely isolated from the input. Typical efficiency of approximately 75%. The manufacturer claims that external capacitors are only needed in critical applications. Bottom: SIP format, fixed input of 12VDC, fixed output of 5VDC at 600mA completely isolated from the input. Typical efficiency of approximately 75%. Requires external capacitors for ripple rejection.)

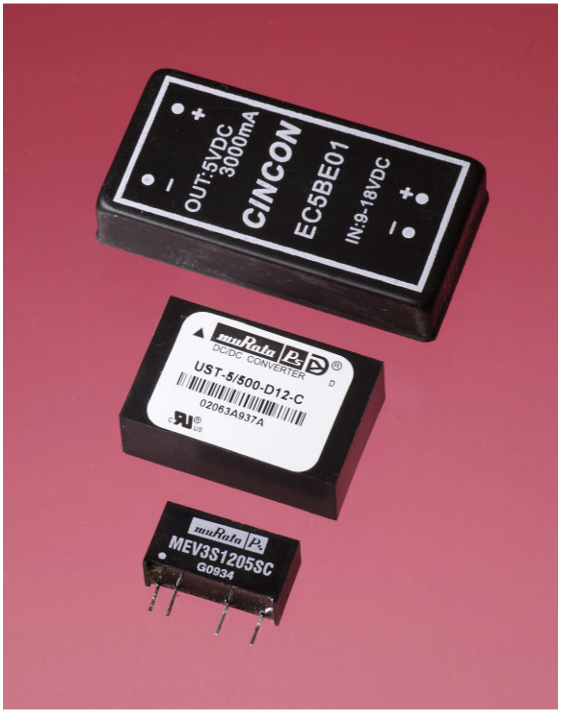

Figure 17-7. *A selection of sealed DC-DC converters.*

Lower-power converters are also available as surface-mount devices.

Some adjustable-output converters are supplied as multiple surface-mounted components preinstalled on a mini-board that has pins for through-hole insertion in a printed circuit board. Their high efficiency enables them to handle a lot of power for their size. In Figure 17-8, the converter accepts a 4.5 to 14VDC input range and has an adjustable output of 0.6 to 6VDC. It is rated at a surprising 10A or 50W and is more than 90% efficient. However, it draws 80mA in a no-load state, causing it to become quite hot. A thermal cutout or automatic shutdown may be used if the converter will not be driving a consistent load.

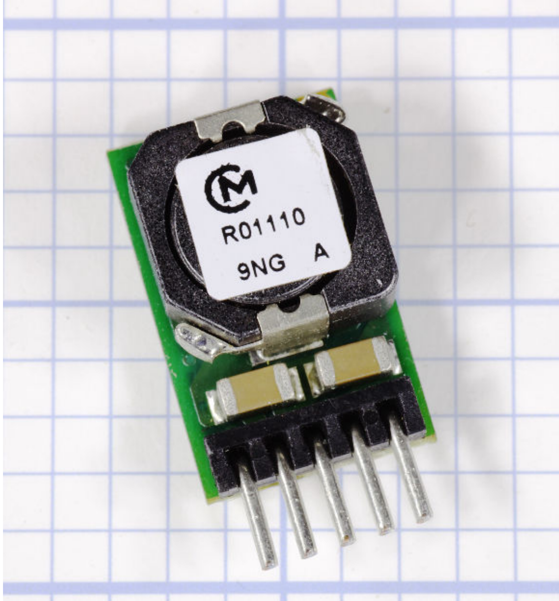

Figure 17-8. *An adjustable DC-DC converter rated for 10A or 50W. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

The miniboard in Figure 17-9 accepts an input voltage from 7 to 36VDC and has an adjustable output ranging from 2.5 to 12.6VDC, at up to 6A. It is non-isolated (has a common negative bus) and claims to be more than 95% efficient at full load.

The miniboard in Figure 17-10 accepts an input voltage from 4.5 to 14VDC and has an adjustable output ranging from 0.6 to 6VDC at up to 20A. It is non-isolated (has a common negative bus) and claims to be more than 90% efficient at full load.

Values

Relevant values include:

Nominal Input Voltage and Frequency

A wide range of input voltages is often acceptable, as the PWM can vary accordingly. Converters

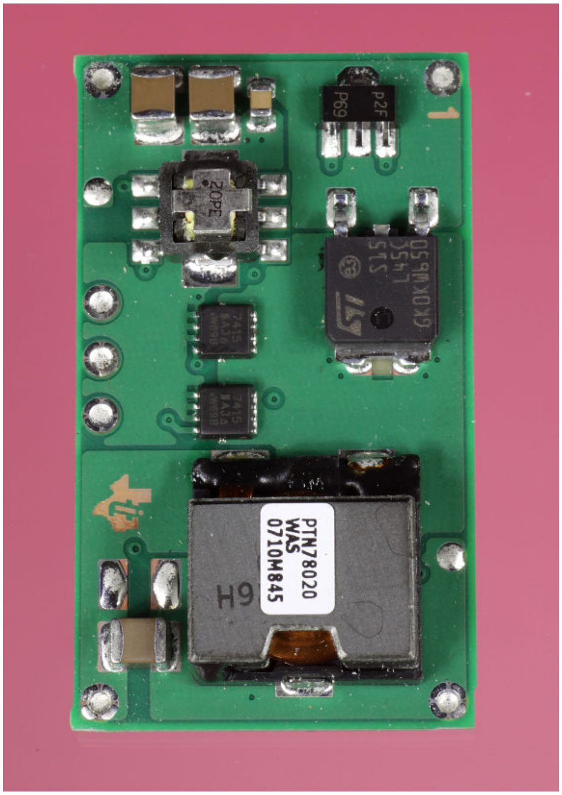

Figure 17-9. *Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

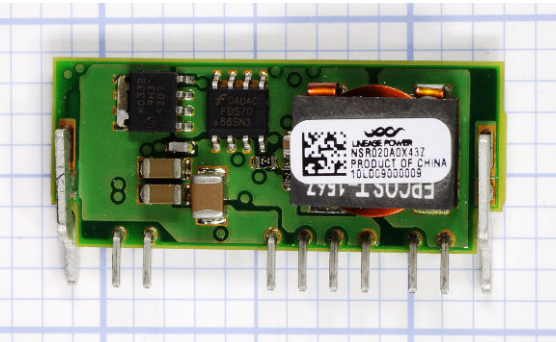

Figure 17-10. *Another adjustable DC-DC converter. The output voltage is determined by adding an external resistor or trimmer potentiometer. External smoothing capacitors are required, as shown in the component's datasheet.*

often allow equipment to be usable internationally, on any voltage ranging from 100VAC to 250VAC, at a fequency of 50Hz or 60Hz, without any adaptation.

Output Voltage

As previously noted, many converters allow the output voltage to be adjusted by adding an external resistor or potentiometer. Alternatively, there may be multiple fixed output voltages, accessible via different pins on the package. They may also provide a positive voltage and equally opposite negative voltage relative to a ground pin.

Input Current and Output Current

Because input voltage and output voltage are likely to be different, the current alone is not a reliable guide to power handling.

A datasheet should specify input current with no load (open circuit on the output side). This current will have to be entirely dissipated as heat.

Load Regulation

This is usually expressed as a percentage and suggests the extent to which output voltage may be pulled down when the load on a DC-DC converter increases. If V_{nil} is the measured output voltage with no load, and V_{max} is the measured output voltage with the maximum rated load:

$$\text{Load regulation} = 100 * (V_{nil} - V_{max})/V_{max}$$

However, note that some converters are designed with the expectation that they will never be used with zero load across the output. In these cases, V_{nil} will be the voltage at minimum rated load.

Efficiency

This is a measure of how much input current must be dissipated as heat. A converter with a 12-volt input, drawing a maximum 300mA input current, will consume 3.6 watts (3,600mW). If it is 80% efficient, it will have to dissipate roughly 20% of its power as heat, or 720mW.

Ripple and Noise

Sometimes abbreviated R/N, this may be measured in mV or as a percentage. Check the specification carefully to determine whether the ripple-and-noise values require use of external smoothing capacitors. Often, this is the case.

Isolated or Non-Isolated

This crucial piece of information is often found near the top of a datasheet, not in the detailed specifications.

How to Use it

Because a converter creates electrical noise, it should be prevented from affecting other components by adding substantial *bypass capacitors* as close as possible to its input and output pins. For most converters, external capacitors are mandatory, and their *effective series resistance* (ESR) should be as low as possible (see the **capacitor** entry in this encyclopedia for an explanation of ESR). Tantalum capacitors are preferable to electrolytics for this reason, and are also more durable. Some manufacturers recommend placing a tantalum capacitor in parallel with an electrolytic. A small ceramic capacitor, typically 0.1 µF, is often recommended in an addition to larger-value capacitors on the output side.

The voltage rating of each capacitor should be twice the voltage at the point in the circuit where it is used. The capacitance value will usually be higher for higher-current converters. Values of 100µF are common, but for high amperage, a value may be as high as 1,000µF.

While datasheets are often inadequate for some types of components, datasheets for DC-DC converters usually include detailed instructions regarding bypass capacitors. Following these instructions is essential. In the relatively rare instances that a datasheet makes no mention of bypass capacitors for a converter, this does not necessarily mean that the capacitors are unnecessary. The manufacturer may assume that they will be used as a matter of course.

Converters are used in a very wide range of devices, supplying power ranging from a few milliamps to tens of amps. At the lower end of the scale, devices such as cellular telephones, portable computers, and tablets contain subcircuits that require different voltages, some of which may be higher than the voltage of a single battery or battery pack that powers the device. A converter can satisfy this requirement. Because a converter can be designed to maintain a fixed output in response to a range of input voltages, it can also compensate for the gradual decline in voltage that occurs during battery usage.

A boost-type converter can be used to double the voltage from a single 1.5V battery in an LED flashlight where 3 volts are required to power the LED. Similarly, a boost-type converter can provide the necessary voltage to run a cold-cathode fluorescent tube that provides backlighting in an LCD computer display.

On a circuit board that is primarily populated with 5VDC components and is fed by a single 5VDC power supply, a converter can be used to supply 12VDC for one special purpose, such as an analog-digital converter or a serial data connection.

If electromechanical **relays** or other inductive loads share a common ground with components ,such as *logic chips* or **microcontrollers**, it may be difficult to protect the sensitive components from voltage spikes. A A flyback converter with a transformer separating the output from the input can allow the "noisy" section of the circuit to be segregated, so long as the converter itself does not introduce noise. Since the electromagnetic interference (EMI) introduced by converters varies widely from one model to another, specifications should be checked carefully.

Very low-power components can pick up EMI from the wires or traces leading into and out of a converter. In this type of circuit, adequate noise suppression may be impossible, and a converter may not be appropriate.

What Can Go Wrong

Electrical Noise in Output

Electrolytic capacitors may be inadequate to smooth the high frequencies used. Multilayer ceramic capacitors or tantalum capacitors may be necessary. Check the manufacturer's datasheet for minimum and maximum values. Also check the datasheet for advice regarding placement of capacitors on the input side as well as across the output.

Excess Heat with No Load

Some converters generate substantial heat while they are powered without a load. The manufacturer's datasheet may not discuss this potential problem very prominently or in any detail. Check the input rating, usually expressed in mA, specified for a no-load condition. All of this current will have to be dissipated as heat, and the very small size of many converters can result in high localized temperatures, especially since many of them allow no provision for a heat sink.

Inaccurate Voltage Output with Low Load

Some converters are designed to operate with at least 10% of full rated load across their output at all times. Below this threshold, output voltage can be grossly inaccurate. Read datasheets carefully for statements such as this: "Lower than 10% loading will result in an increase in output voltage, which may rise to typically double the specified output voltage if the output load falls to less than 5%." Always use a meter to verify the output voltage from a converter at a variety of different loads, and perform this test before installing the converter in a circuit.

DC-AC inverter

18

A power inverter must not be confused with a *logic inverter*, which functions as a digital component in *logic circuits* to invert the state of a low-voltage DC input from high to low or low to high. Logic inverters are discussed in Volume 2.

OTHER RELATED COMPONENTS

- **AC-DC power supply** (See Chapter 16)
- **DC-DC converter** (See Chapter 17)

What It Does

A power inverter is included here as counterpoint to a **power supply** or *AC adapter*, since it has the opposite function. The inverter receives an input of *direct current* (typically 12VDC from a car battery) and delivers an output of *alternating current* (AC) in the range 110VAC-120VAC or 220VAC-240VAC, suitable to power many low-wattage appliances and devices. The interior of a low-cost inverter is shown in Figure 18-1.

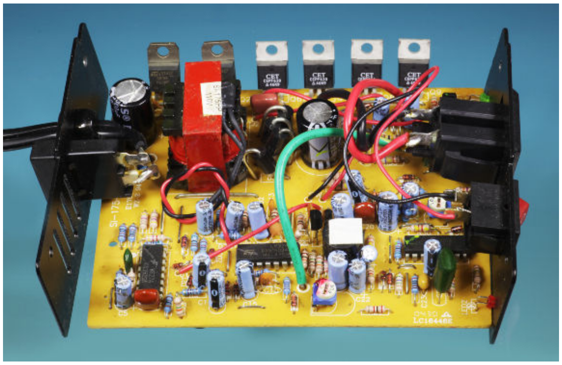

Figure 18-1. *The interior components in a 175-watt inverter.*

How It Works

The first stage of an inverter typically raises a 12VDC input to a higher DC voltage via an internal **DC-DC converter**, then uses a switching circuit to create an approximation of the sinusoidal profile that is characteristic of AC voltage.

Digital switching components naturally tend to create square waves, whose simple appearance conceals the presence of higher frequencies, or *harmonics*, that are ignored by some devices (especially those that convert electricity into heat) but can be troublesome in consumer electronics equipment. A primary objective of inverter design is to adapt or combine square waves to emulate a classic AC sine wave with reasonable fidelity. Generally speaking, the more accurately an inverter emulates a sine wave, the more expensive it tends to be.

The most primitive inverter would create a plain square wave such as that shown in red in Figure 18-2, superimposed on a comparable sine wave (in green). Note that alternating current rated at 115 volts actually peaks at around 163 volts because the number 115 is the approximate *root mean square* (*RMS*) of all the voltage values during a single positive cycle. In other words, if the voltage is sampled x times during a cycle, an

RMS value can be derived by squaring each sample, adding all the samples, dividing by x, and then taking the square root of the result. The RMS value is important as a means to calculate actual power delivered because it can be multiplied by the current to obtain an approximate value in watts.

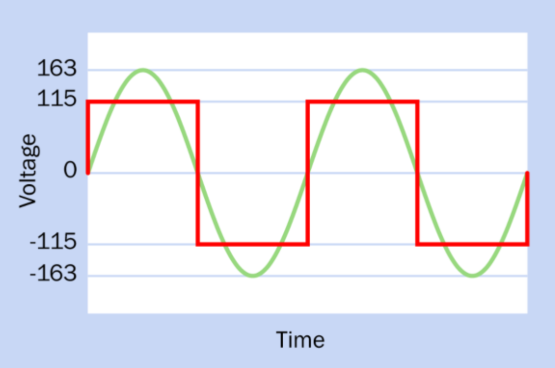

Figure 18-2. *Comparison of an AC voltage sine wave (green) and a square wave of the same frequency (red), both delivering a roughly similar amount of power.*

Variants

As a first step toward a better approximation of a sine wave, gaps of zero voltage can be inserted between square-wave pulses. This "gapped" square wave is shown in Figure 18-3.

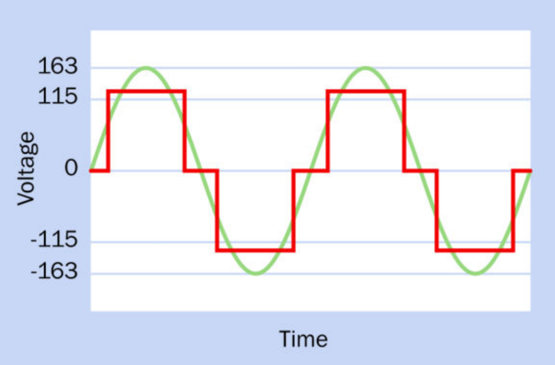

Figure 18-3. *Introducing pauses or gaps of zero voltage between square-wave pulses can produce slightly improved resemblance to a sine wave.*

A further improvement can be achieved if an additional, shorter pulse of higher voltage is added to each primary pulse, as shown in Figure 18-4. Outputs of this kind are referred to as *modified sine wave*, although they are actually square waves modified to emulate a sine wave. Their inaccuracy is expressed as *total harmonic distortion* (*THD*). Some authorities estimate that the THD of gapped square-wave output is around 25%, whereas the addition of shorter square waves reduces this to around 6.5%. This is a topic on which few people agree, but there is no doubt that a "stacked" sequence of square waves provides a closer emulation of a sine wave.

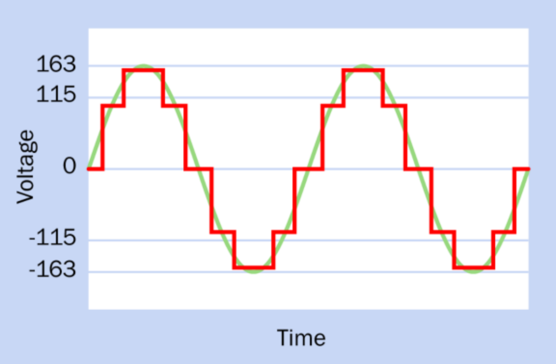

Figure 18-4. *A secondary stream of narrower square-wave pulses can improve the fidelity of an inverter's output.*

A *true sinewave inverter* typically uses *pulse-width modulation* (*PWM*) to achieve THD of less than 1%. It generates a stream of pulses much higher in frequency than that of the AC output, and varies their widths in such a way that their averaged voltage closely approximates the voltage variations in a sine wave. A simplified representation of this principle is shown in Figure 18-5.

Values

Small inverters are typically rated to deliver up to 100 watts and may be fitted with a 12VDC plug for insertion in a vehicle's cigarette lighter. Since a cheap inverter may be only 80% efficient, 100 watts at 135VAC will entail drawing as much as

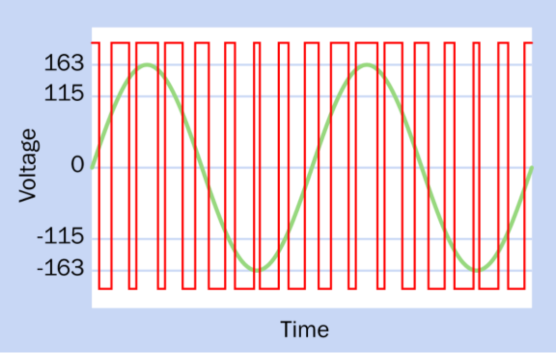

Figure 18-5. *Pulse-width modulation adjusts the widths of pulses delivered at a high frequency. The pulse widths can be averaged to generate voltage that follows a close approximation of a sine wave.*

10 amps at 12VDC. Cigarette lighters are usually fused at 15 or 20 amps, so 100 watts is a reasonable value. Inverters that are rated above 150 watts usually have cables terminating in oversize alligator clips for direct connection to the terminals of a 12V battery.

While the *cold cranking* rating of a car battery may be 100 amps or more, the battery is only designed to deliver that power for up to 30 seconds at a time. Inverters rated for as much as 500 watts will exceed the normal capacity of a single car battery, although if the battery is mounted in a vehicle, it can be supplemented by running the engine so that the alternator shares some of the load. A 500-watt inverter is better supplied by two or more 12-volt car batteries wired in parallel.

How to Use it

Small inverters are typically used in vehicles to run cellphone chargers, music players, or laptop computers. Large inverters are an integral part of off-the-grid solar and wind-powered systems, where battery power must be converted to AC house current. Uninterruptible power supplies contain batteries and inverters capable of running computer equipment for a brief period. Battery-driven electric vehicles with AC motors use inverters with an exceptionally high current rating.

There is a lack of consensus regarding possible harmful effects of powering electronics equipment with a low-cost modified sine wave inverter. Anecdotal evidence suggests that where the equipment uses its own *switching power supply* or uses an *AC adapter* (either mounted internally or as an external plug-in package), the filtering built into the power supply will block harmonics from the inverter.

Other evidence suggests that cheap inverters may have adverse effects on devices containing synchronous motors that run direct from AC. There are reports that fluorescent lighting and photographic electronic flash systems may be unsuitable for use even with modified sine wave inverters. However, differences in product design and component quality make it impossible to generalize. A cheaply made inverter may generate a wave form that is not even a close approximation of a square wave. See Figure 18-6.

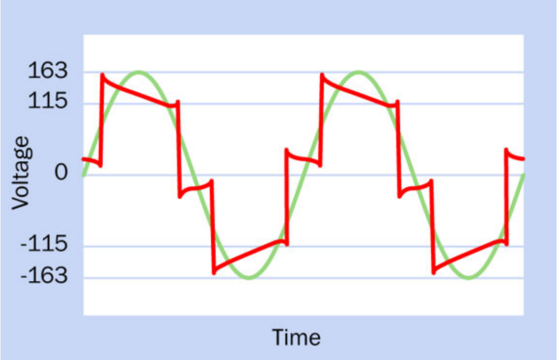

Figure 18-6. *A cheaply made inverter can generate a distorted wave form that is even higher in noise than a pure square wave. This sample is adapted from an actual oscilloscope trace.*

What Can Go Wrong

If multiple batteries are connected in parallel, using suitably heavy-gauge wire to power a large inverter, the batteries must be identical in specification and age, and must all be equally charged to prevent high and potentially dangerous flows of current among the batteries as they attempt to reach an equilibrium among themselves. Interconnections must be firmly clamped to clean battery terminals. For additional information, see the **battery** entry in this encyclopedia.

Problems associated with inverters are likely to be mundane. A 12V wiring to the inverter can overheat if items such as clothes or bedding are left on top of it; a high-wattage fan-cooled inverter can overheat if the fan is obstructed by poor placement or impaired by accumulated dirt; alligator clips may become dislodged from battery terminals; and power surges drawn by inductive loads such as motors may trigger the inverter's *breaker*, especially if they are used in conjunction with other equipment.

As always, high amperage should be treated with caution and respect, regardless of it being delivered at "only 12 volts."

voltage regulator

19

Correctly known as a *linear voltage regulator* to distinguish it from a *switching regulator* or **DC-DC converter**. However, the full term is not generally used, and "voltage regulator" is normally understood to mean a linear voltage regulator.

OTHER RELATED COMPONENTS

- **DC-DC converter** (See Chapter 17)
- **AC-DC power supply** (See Chapter 16)

What It Does

A linear voltage regulator provides a tightly controlled DC output, which it derives from an unregulated or poorly regulated DC input. The DC output remains constant regardless of the load on the regulator (within specified limits). It is a cheap, simple, and extremely robust component.

There is no single schematic symbol for a linear voltage regulator.

The general physical appearance of a commonly used type of regulator, rated for an output of around 1A DC, is shown at Figure 19-1. The LM7805, LM7806, LM7812, and similar regulators in the LM78xx series are encapsulated in this type of package, with pins that are spaced at 0.1″ and have functions as shown. Other types of regulator may differ in appearance, or may look identical to this one but have different pin functions. Always check datasheets to be sure.

How It Works

All linear regulators function by taking some feedback from the output, deriving an error value by comparing the output with a reference voltage (most simply provided by a zener diode), and using the error value to control the base of a *pass transistor* that is placed between the input and the output of the regulator. Because the transistor operates below saturation level, its output current varies linearly with the current applied to its base, and this behavior gives the linear regulator is name. Figure 19-2 shows the relationship of these functions in simplified form; Figure 19-3 shows a little more detail, with a *Darlington pair* being used as the pass transistor. The base of the pair is controlled by two other transistors and a comparator that delivers the error voltage. This version of a voltage regulator is known as the *standard type*.

The voltage difference required between the base and emitter of an NPN transistor is a minimum of 0.6V. Because multiple transistors are used inside a standard-type voltage regulator, it requires a minimum total voltage difference, between its input and its output, of 2VDC. This voltage difference is known as the *dropout voltage*. If the voltage difference falls below this minimum, the regulator ceases to deliver a reliable output voltage until the input voltage rises again. *Low dropout regulators* allow a lower voltage difference, but are more expensive and less commonly used. They are described under the following Variants section.

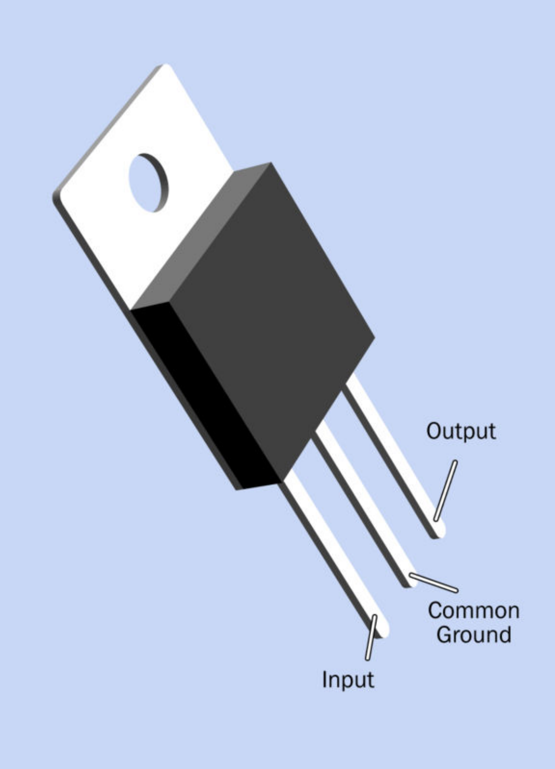

Figure 19-1. *The package design of a commonly used voltage regulator. Others may be significantly different, and the pin functions may vary. Check manufacturer datasheets for details.*

Discrete components could in theory be used to build a voltage regulator, but this ceased to be cost-effective several decades ago. The term is now understood to mean one small integrated package containing the basic circuit augmented with additional, desirable features, such as automatic protection against overload and excessive heat. Instead of burning out if it is overloaded, the component simply shuts down. Most voltage regulators also tolerate accidentally reversed power connection (as when batteries are inserted the wrong way around) and accidentally reversed insertion of the regulator in a circuit board.

Other components can satisfy the requirement to deliver power at a reduced voltage. Most simply, if two resistors in series are placed across a

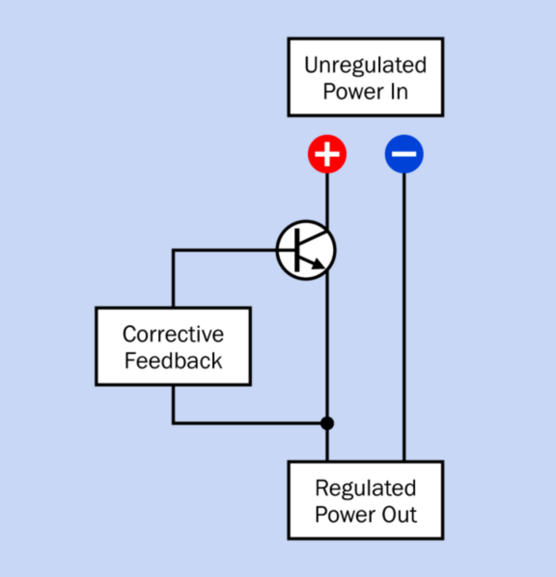

Figure 19-2. *A linear voltage regulator basically consists of a transistor whose base is controlled by corrective feedback derived from the output.*

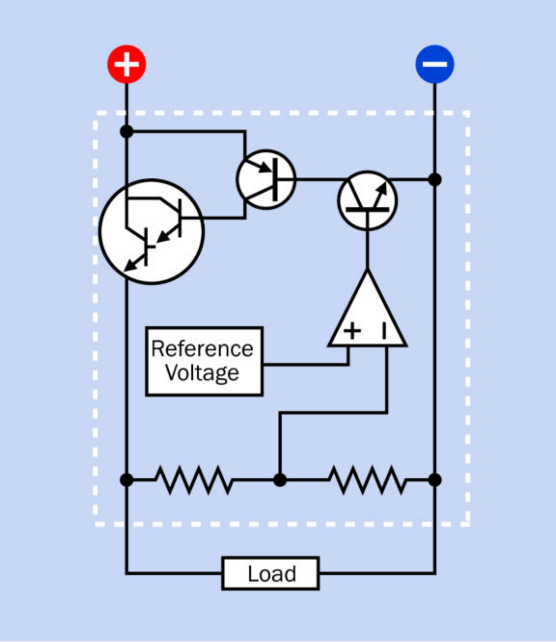

Figure 19-3. *The fundamental internal features of a standard-type voltage regulator, including a Darlington pair, two transistors, a voltage divider, comparator, and reference voltage source, shown inside the dashed white line.*

power source, they form a *voltage divider*, which provides an intermediate voltage at the connection between them. However, this voltage will vary depending on fluctuations in the input voltage and/or load impedance. A voltage regulator is the simplest way to supply a voltage that remains stable regardless of excursions in the input or fluctuations in power consumed by the load.

The disadvantage of a standard-type voltage regulator is that it is inefficient, especially when a relatively high input voltage is used to deliver a relatively low output voltage. If V_{in} is the input voltage, V_{out} is the output voltage, and I_{out} is the output current, the average power loss, P, is given by the formula:

$$P = I_{out} * (V_{in} - V_{out})$$

For example, if the output current is 1A, the input voltage is 9VDC, and the output is 5VDC, 44% of the input power will be wasted, and the component will be only 56% efficient. The wasted power (about 4 watts, in this case) will be dissipated as heat. Even when a standard-type regulator runs at its minimum 2VDC dropout voltage, it must dissipate 1W when delivering 0.5A.

Variants

Packaging

The package for the LM78xx series of regulators, shown in Figure 19-1, incorporates an aluminum plate drilled with a hole so that it can be bolted to a heat sink. Voltage regulators with a lower rated maximum output current (typically, 100mA) do not have the same need for a heat sink, and are available in a package that resembles a small transistor.

Some integrated circuits are available containing two voltage regulators, electrically isolated from each other.

Popular Varieties

In the LM78xx series, the last two digits in the part number specify the output voltage, which is fixed. Thus the LM7805 delivers 5VDC, the LM7806 delivers 6VDC, and so on. For regulators with a fractional voltage output (3.3VDC being common), an additional letter may be inserted in the part number, as in the 78M33.

Many copies of the LM78xx series are made by different manufacturers, the copies being functionally identical, regardless of additional letters that are added to the part number to identify its source or other attributes.

The LM78xx regulators are mostly rated to be accurate within 4%, although actual samples almost always deliver voltages that are more precise than this range suggests.

Adjustable Regulators

While the majority of regulators have a fixed output, some allow the user to set the output by adding one or more resistors. The LM317 is a popular example. Its output voltage can range from 1.25VDC to 37VDC and is set via a resistor and a trimmer potentiometer, as illustrated in Figure 19-4. If R1 is the fixed-value resistor and R2 is the trimmer, as shown in the schematic, the output voltage, V_{out}, is given by the formula

$$V_{out} = 1.25 * (1 + (R2 / R1))$$

Typical values for R1 and R2 would be 240Ω and 5K, respectively. With the trimmer at the middle of its range, V_{out} would be 1.25 * (1 + (2500 / 240)) = approximately 15VDC, requiring an input of at least 17VDC. However, if the trimmer is reduced to 720Ω, the output would be 5VDC. In practice, the value of a trimmer should be chosen so that a mid-range setting provides approximately the desired output. This will enable fine adjustment of the output voltage.

While the versatility of an adjustable regulator is desirable, its overall power dissipation is still proportional to the difference between the input voltage and the output voltage. To minimize heat loss, this difference should not exceed the dropout voltage by a larger amount than is absolutely necessary.

An adjustable regulator may require larger bypass capacitors than a regulator with a fixed output. A manufacturer's recommendations for the LM317 are shown in Figure 19-4.

Negative and Positive Regulators

While most linear voltage regulators are designed for "positive input" (conventional current flow from input to output), some are intended for "negative input." In this variant, the common terminal is positive, and the input and output are negative in relation to it.

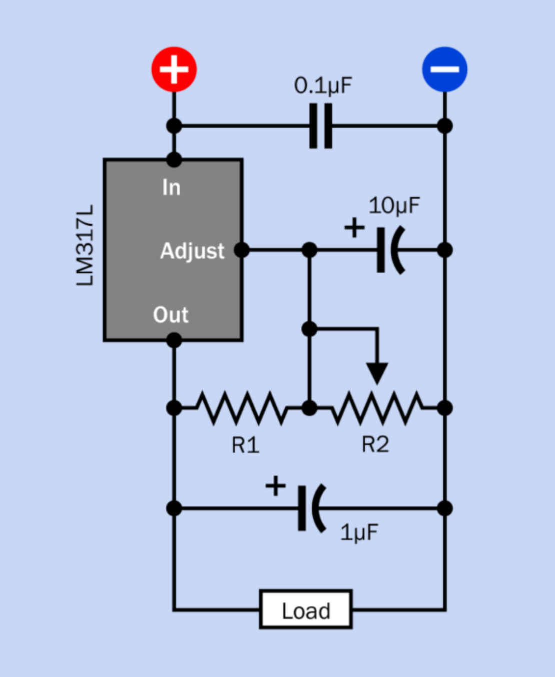

Figure 19-4. *Schematic for the LM317L adjustable voltage regulator, based on a circuit recommended by National Semiconductor, with bypass capacitors added for ripple rejection.*

Low-Dropout Linear Regulators

Low-dropout regulators (sometimes referred to as *LDO* regulators) allow a much lower dropout voltage by using a single PNP or MOSFET transistor. LDO regulators are popularly used in battery-powered devices where efficiency should be maximized and heat dissipation should be minimized. For example, the LM330 is a regulator with a 5VDC output, tolerating a dropout voltage of 0.6V, allowing it to be used with four AAA cells. In an LDO regulator the dropout voltage actually varies with load current and may diminish to as little as one-tenth of its rated value when the output current is minimal.

The majority of low-dropout regulators are sold in surface-mount packages, and are designed for maximum output of 100mA to 500mA. Only a few exceptions exist. They tend to be slightly more expensive than regulators with the typical 2V dropout rating.

Three voltage regulators are shown in Figure 19-5. From left to right, they are rated 5VDC at 1A, 12VDC at 1A, and 5VDC at 7.5A. The two smaller regulators are of the LM78xx series. The larger regulator claims a low maximum dropout voltage of 1.5VDC, and its output voltage can be adjusted with an external potentiometer and resistor.

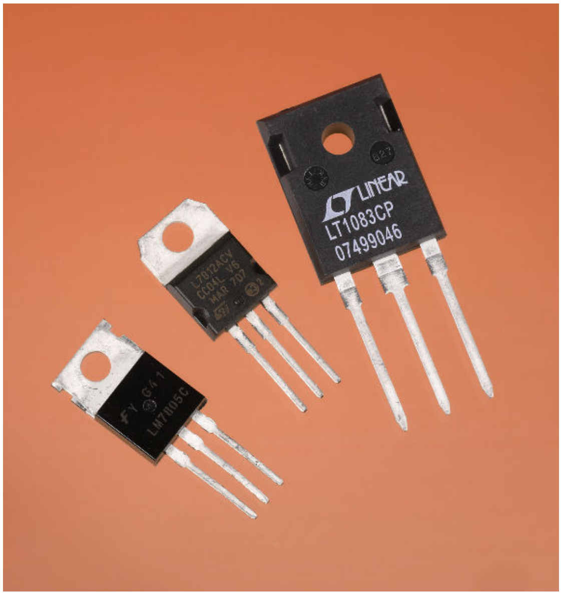

Figure 19-5. *Two voltage regulators from the LM78xx series, and a third high-current, low-dropout, adjustable regulator rated 5VDC (adjustable upward) at 7.5A.*

Quasi-Low-Dropout Linear Regulators

Where a standard regulator uses a Darlington pair as the pass transistor and an LDO uses a single PNP transistor, the so-called Quasi-LDO uses a combination of NPN and PNP transistors and has an intermediate dropout voltage, typically a maximum of 1.5VDC. However, the terms LDO and Quasi-LDO are not used uniformly in the industry. One manufacturer markets Quasi-LDO regulators as LDO regulators, and describes its LDO regulators as *Very Low Dropout* regulators. Consult datasheets to determine the actual specification of the product, regardless of its classification.

Additional Pin Functions

Some voltage regulators include an extra pin, typically known as an *enable pin*, which switches off the device in response to a signal from a microcontroller or logic gate.

Some regulators offer another option, an additional *status pin* that can signal a microcontroller that an error mode exists if the regulator output falls significantly below its rated value.

In battery-powered devices, a low-battery sensor is a desirable feature, since a regulator may simply shut down without warning if the input voltage is insufficient. A few regulators, such as the LP2953, provide a low-battery warning output via an extra pin.

Values

Linear voltage regulators with a single, fixed output are commonly available to supply DC outputs of 3.3, 5, 6, 8, 9, 10, 12, 15, 18, and 24 volts, with a few variants offering fractional values in between. The most commonly used values are 5, 6, 9, 12, and 15 volts. The input voltage may be as high as 35VDC.

Maximum output current is typically 1A or 1.5A, in the traditional three-pin, through-hole, TO-220 format. A surface-mount version is available. Other surface-mount formats have lower power limits.

Accuracy may be expressed as a percentage or as a figure for load regulation in mV. A typical load regulation value would be 50mV, while voltage regulation accuracy ranges from 1% to 4%, depending on the manufacturer and the component. While low-dropout regulators are generally more efficient, they do require more ground-pin current. This is not usually a significant factor.

How to Use it

Some components, such as many old-design CMOS chips or the traditional TTL version of the 555 timer, allow a wide range of acceptable input voltages, but most modern logic chips and microcontrollers must have a properly controlled power supply. Regulators such as the LM7805 are traditionally used to provide this, especially in small and relatively simple devices that draw a moderate amount of current, have a low component count, and are powered via a battery or an **AC adapter**. A fully fledged switching power supply is overkill in this kind of application.

A linear voltage regulator cannot respond instantly to changes in input voltage. Therefore, if the input supply contains voltage spikes, these spikes may pass through the regulator. Bypass capacitors should be applied preventively. A sample schematic showing an LM7805 regulator with bypass capacitors recommended by a manufacturer is shown in Figure 19-6.

In a battery-powered device where standby power is required for long periods and full power is only needed intermittently, the *quiescent current* drawn by a minimally loaded voltage regulator is important. Modern LDO regulators may draw as little as 100µA when they are very lightly loaded. Other types may consume significantly more. Check datasheets to find the most appropriate component for a particular application.

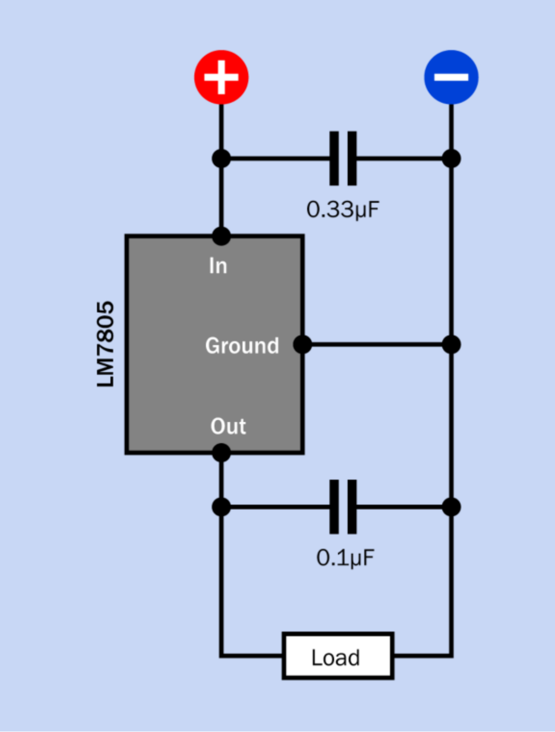

Figure 19-6. *Typical schematic for use of an LM7805 regulator, with capacitor values based on recommendations from Fairchild Semiconductor.*

Note that DC-DC power **converters** may draw a lot of current when they are lightly loaded, and will dissipate large amounts of heat as a result. An LDO is therefore preferable in this situation.

What Can Go Wrong

Inadequate Heat Management

The ability to "dial up" a wide range of voltages from an adjustable regulator such as the LM317 can be a temptation to use it on a "one size fits all" basis, to deliver any output ranging from 5VDC to 18VDC from a uniform 24VDC input. Assuming 1A output current, the worst-case power dissipation in this scenario would be almost 20W. To achieve reasonable efficiency and maintain waste heat at a manageable level, the input voltage should not exceed the output voltage by much more than the dropout voltage.

Even when a voltage regulator is used correctly, it can generate more heat than was expected if the requirements of a circuit are altered during development. An initial handful of components may draw only 100mA, but as more capabilities are requested and more parts are added (especially relays or LED displays) the power consumption can quickly add up, generating an unexpected amount of waste heat and raising the possibility of a sudden (and mysterious) shutdown if the regulator does not have an adequate heat sink.

Transient Response

When there is a major fluctuation in the demand by the load (for example, if an inductive device is switched on elsewhere in the circuit), the voltage regulator requires a finite time to adjust itself and maintain its specified output voltage. This time lag is known as its *transient response*. If a momentary fluctuation is likely, and other components may be sensitive to it, a larger capacitor should be used between the output of the voltage regulator and ground.

The transient response time may also be insufficient to block sudden, brief spikes in input voltage. This may occur, for example, when a low-cost AC adapter that does not have a properly smoothed output is used as the power source. Additional 1µF bypass capacitors may be added at the input and output of a regulator to provide better protection from power fluctuations.

Misidentified Parts

Many types of linear voltage regulators appear physically identical. Care is needed to distinguish those which have fixed output from those that allow a variable output. When using the LM78xx series, double-check the last pair of digits in the part number, which provide the only guide regarding the output. Using an LM7808 instead of an LM7805 may be sufficient to destroy all the 5VDC chips in a logic circuit. It is advisable to use a meter to check the output of any power supply before connecting it with a circuit.

Misidentified Pins

The LM78xx series of voltage regulators uses a very intuitively obvious and consistent scheme for the functions of its pins: input on the left, ground in the center, and output on the right, when looking at the regulator from the front, with its pins facing downward. Unfortunately the consistency of this scheme can encourage an unthinking habit for making connections. The LM79xx series of negative voltage regulators swaps the identity of the input and ground pins, whereas adjustable regulators use yet another different scheme. Good practice suggests checking a component against the manufacturer's datasheet before connecting it.

Dropout Caused by Low Battery

If a regulator rated to deliver 6VDC has a 2VDC dropout voltage and is powered from a 9V battery, the battery can easily drop below the minimum acceptable 8VDC if it becomes old or depleted. When this happens, the output from the regulator will tend to fall, or may oscillate.

Inaccurate Delivered Voltage

A voltage regulator maintains its output voltage between its output pin and ground pin. Thin traces on a circuit board, or a long run of very small-gauge wiring, can impose some electrical resistance, reducing the actual voltage delivered to a component. Ohm's Law tells us that the voltage drop imposed by a trace (or thin wire) will be proportional to the current flowing through it. For example, if the resistance between the output pin of a voltage regulator and a component is 0.5Ω and the current is 0.1A, the voltage drop will be only 0.05V. But if the current increases to 1A, the voltage drop is now 0.5V. Bearing this in mind, a linear voltage regulator should be positioned close to voltage-sensitive components. In printed circuit designs, the traces that deliver power should not have significant resistance.

When using linear voltage regulators with adjustable output, there may be a temptation to connect adjustment resistor R1 to the positive end of the load, to obtain a "more accurate" delivered voltage. This configuration will not produce the desired result. R1 should always be connected as closely as possible between the output pin and the adjustment pin of the voltage regulator, while R2 should connect between the adjustment pin and the negative end of the load. This is illustrated in Figure 19-7, where the gray wire in each schematic indicates that it possesses significant resistance.

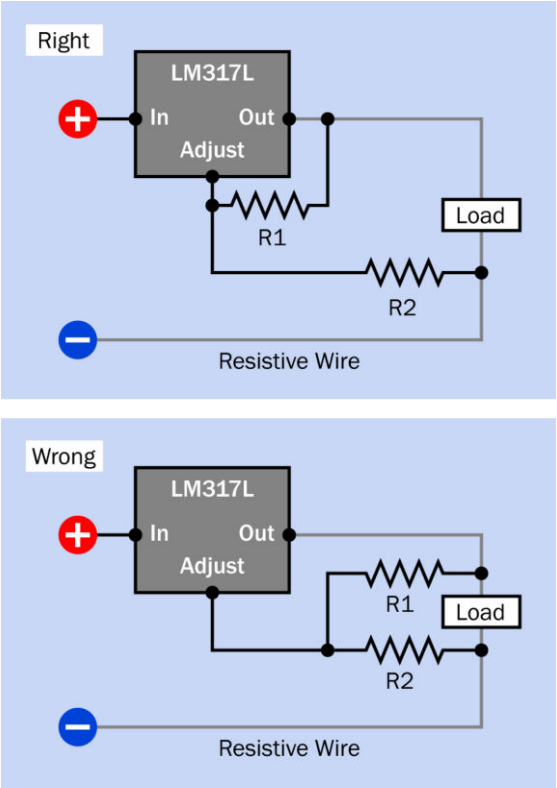

Figure 19-7. *When the connection between an adjustable-output voltage regulator and load components has a significant resistance (shown here as a gray "resistive wire"), R1 should always be connected as closely as possible to the pins of the regulator, as shown in the upper schematic. (Derived from schematics prepared by National Semiconductor.)*

electromagnet

20

The term **electromagnet** is used here to mean a coil containing a core of ferromagnetic material that does not move relative to the coil. The core is used solely to create a magnetic field that attracts or repels other parts that have appropriate inherent magnetic properties. Where a center component moves in response to the magnetic force created by current through a coil, this is discussed in the **solenoid** entry. By comparison, the **inductor** entry describes a coil that may or may not have a ferromagnetic core, and is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **solenoid** (See Chapter 21)
- **electromagnet** (See Chapter 20)
- **DC motor** (See Chapter 22)
- **AC motor** (See Chapter 23)

What It Does

An electromagnet consists of a coil that creates a magnetic field in response to an electric current. The field is channeled and reinforced by a core of magnetic material (that is, material that can be magnetized). Electromagnets are incorporated in motors, generators, loudspeakers, microphones, and industrial-sized applications such as mag-lev trains. On their own, they provide a means for electric current to hold, lift, or move objects in which a magnetic field can be induced.

A very small, basic electromagnet about 1 inch in diameter is shown in Figure 20-1. No specific schematic symbol for an electromagnet exists, and the symbol for an induction coil with a solid core is often used instead, as shown in Figure 14-1 (the center variant of each of the three) in the **inductor** entry of this encyclopedia.

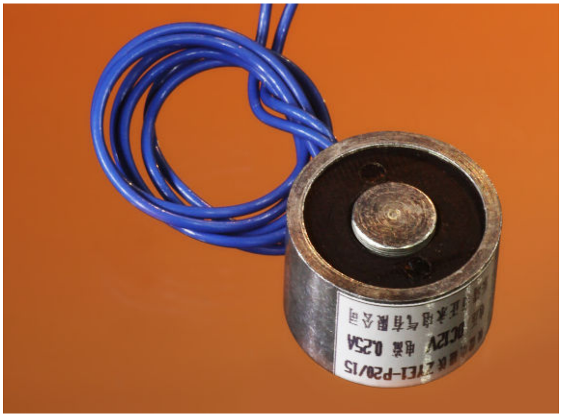

Figure 20-1. *An electromagnet approximately 1 inch in diameter, rated to draw 0.25A at 12VDC.*

How It Works

Electric current flowing through a circle of wire (or a series of connected loops that form a helix or *coil*) will induce a magnetic field through the center. This is illustrated in the **inductor** entry of this encyclopedia, specifically in diagrams Figure 14-3, Figure 14-4, Figure 14-5, and Figure 14-6.

If a stationary piece of ferromagnetic material is placed in the center of the circle or coil, it enhances the magnetic force because the *reluctance* (magnetic resistance) of the material is much lower than the reluctance of air. The combination of the coil and the core is an electromagnet. This is illustrated in Figure 20-2. For a lengthier discussion of this effect, see "Magnetic Core" (page 122).

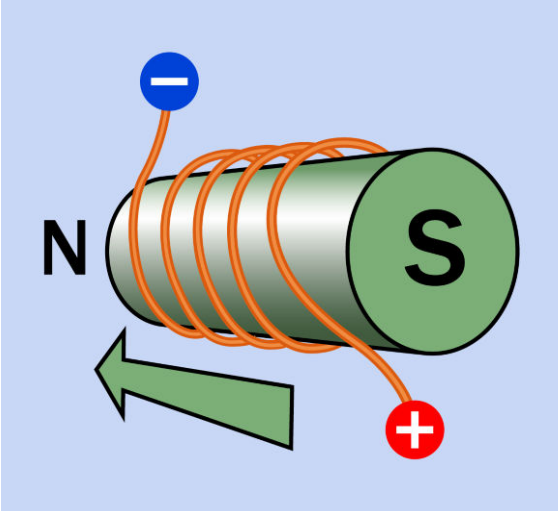

Figure 20-2. *Direct conventional current flowing through a wire coiled around a ferromagnetic rod induces a magnetic force in the rod, conventionally considered to flow from south to north.*

The magnitude of the electromagnetic *flux density* will be proportional to the current flowing through the coil, assuming a DC power source.

Variants

Electromagnet designs vary according to their application. The simplest design consists of a single coil wound around a rod which may terminate in a plate for applications such as lifting scrap metal. This design is relatively inefficient because the magnetic circuit is completed through air surrounding the electromagnet.

A more efficient, traditional design consists of a U-shaped core around which are wound one or two coils. If the U-shaped core is smoothly curved, it resembles a *horse-shoe magnet*, as shown in Figure 20-3. This design has become relatively uncommon, as it is cheaper to make windings across two separate, straight vertical cores and bridge them. However, the horseshoe configuration is extremely efficient, as the coils induce north and south magnetic polarities in the open ends of the U-shaped core, and the magnetic circuit is completed through any object that is attracted toward the open ends and links them. The attracted object is shown as a rectangular plate in Figure 20-3. Because a magnetic circuit will naturally attempt to limit its extent, and because this goal will be achieved when the circuit is completed, the attractive force of the U-shaped magnet is maximized.

An electromagnet powered by direct current naturally produces a consistently polarized, stable magnetic field. When AC current is applied, an electromagnet may still be used to exert an attractive force on a passive object that is not magnetized but is capable of being magnetized. The electromagnet will change its polarity at almost the same frequency as the AC, and will induce equal and opposite fluctuating polarity in the target, causing mutual attraction. The core of the magnet will be composed of plates separated by thin layers of insulation to inhibit the eddy currents induced by the AC, but still an AC-powered electromagnet will be less efficient than a comparable DC-powered electromagnet because it will also suffer from *hysteresis* as power is consumed by repeatedly reversing the polarity of the *magnetic domains* in the core.

Some electromagnets that are described as suitable for AC power actually contain *rectifiers* that convert the AC to DC.

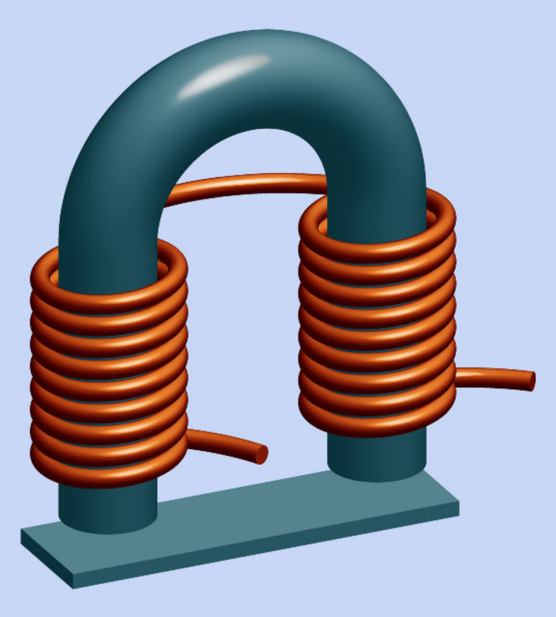

Figure 20-3. *This traditional design for an electromagnet has a pedigree stretching back for more than a century. It maximizes efficiency by completing a magnetic circuit through any object that the magnet attracts.*

Values

Electromagnets are typically calibrated in terms of their power consumption and retaining force (the weight of an iron target that they can support). The retaining force is usually measured in grams or kilograms.

How to Use it

Electromagnets are used mostly as subassemblies in other components, such as motors and generators, relays, loudspeakers, and disk drives. They have also been used in audio (and video) tape recorders to magnetize ferric oxide on tape, using a magnetic field of varying strength to record an audio signal. In this application, a form of horseshoe magnet with an extremely narrow gap is used, the width of the gap determining the highest frequency that the electromagnet can record, in conjunction with the speed of the tape moving past the head.

The tape recording process can be reversed when the electromagnet "reads" the tape and turns the signal back into a weak alternating current that can be amplified and reproduced through a loudspeaker.

A simple application for an electromagnet is in a traditional-style doorbell, where one or two coils attract a spring-loaded lever, at the top of which is a knob that hits a bell. When the lever is pulled toward the bell, it breaks a contact that supplies power to the electromagnet. This allows the lever to spring back to its original position, which reestablishes the circuit, repeating the process for as long as power is applied to the bell. The bulk and weight of the component parts in this type of doorbell are making it obsolete, as electronic versions containing small loudspeakers become relatively cheaper. However, a **solenoid** may still be used in the type of bell that creates a single chime or pair of chimes.

In any device using a *cathode-ray tube*, electromagnetic coils are used to form a *yoke* around the neck of the tube, to deflect the beam of electrons on its way to the screen. A similar principle is used in electron microscopes. In some cases, electrostatically charged plates are used to achieve the same purpose.

An electromagnet may be used to activate a *reed switch* (the diagram in Figure 9-7 shows such a switch). In this application, the combination of the electromagnet and the switch are functioning as a **relay**.

When an electromagnet is energized by alternating current, it can be used to *degauss* (in other words, to demagnetize) other objects. The AC is either applied with diminishing current, so that the alternating magnetic polarities gradually subside to zero, or the electromagnet is gradually moved away from the target, again reducing the magnetic influence to (virtually) zero. This latter procedure may be used periodically to demagnetize record and replay heads on tape recorders, which otherwise tend to acquire residual magnetism, inducing background hiss on the tape.

Traditional large-scale applications for electromagnets tend to involve lifting and moving heavy objects or scrap metal, such as junked cars. A more modern application is in magnetic resonance imaging (MRI), which has revolutionized some areas of medicine.

Very large-scale applications for electromagnets include particle accelerators, in which multiple magnetic coils are energized sequentially, and fusion-power generators, where high-temperature plasma is contained by a magnetic field.

What Can Go Wrong

Because an electromagnet requires constant power to maintain its magnetic force, yet it is not doing any actual work so long as its target remains stationary (in contact with the core of the magnet), the current running through the coil of the magnet must be dissipated entirely as heat. Further discussion of this issue will be found at "Heat" (page 177) in the **solenoid** section of this encyclopedia.

solenoid

21

The term **solenoid** was historically used to describe any coil without a magnetic core. More recently and more commonly it describes a coil inside of which a cylindrical plunger moves in response to the magnetic field generated by the coil. In this encyclopedia, the term **electromagnet** has its own entry, and describes a coil with a center component of ferromagnetic material that does not move relative to the coil. It is used solely to attract or repel other parts that have inherent magnetic properties. By comparison, the **inductor** entry describes a coil that is used for the specific purpose of creating reactance, or self-inductance, in an electronic circuit, often in association with alternating current and in combination with resistors and/or capacitors. The **inductor** entry contains a basic discussion and explanation of magnetic force created by electricity.

OTHER RELATED COMPONENTS

- **inductor** (See Chapter 14)
- **solenoid** (See Chapter 21)

What It Does

A typical solenoid consists of a hollow coil inside a *frame*, which may be a sealed cylinder or box-shaped with open sides. In the case of a cylinder, its opposite ends may be referred to as *pole faces*.

At least one of the pole faces has a hole through which a *plunger* (also known as an *armature*) is pulled or pushed by the solenoid. Thus, the solenoid is a device for applying a linear mechanical force in response to current passing through it. In most solenoids, current must be maintained in order to maintain the mechanical force.

A small open-frame solenoid is pictured in Figure 21-1. The upper section of the figure shows the three basic parts: frame, compression spring, and plunger. The lower part of the figure shows the parts assembled.

A larger, closed, cynlindrical solenoid is shown in Figure 21-2, with the plunger and spring removed.

A 3D rendering showing a simplified, imaginary, cylindrical solenoid cut in half appears in Figure 21-3. The diagram includes a gray cylindrical shell, often described as the frame; the coil, shown in orange; the plunger, which is pulled into the coil by its magnetic field; and the triangular stop, which limits the plunger's upward travel. The frame of the solenoid exists not merely to protect the coil, but to provide a magnetic circuit, which is completed through the plunger.

The lower end of the plunger is often fitted with a nonmagnetic yoke or perforated plate for connection with other components. Stainless steel can be used for this purpose. The stop may be fitted with a thrust rod (also fabricated from stainless steel) if the solenoid is intended to "push" as well as "pull." Springs to adjust the force of the plunger, or to return it to its initial position when the current through the coil is interrupted, are not shown in the rendering.

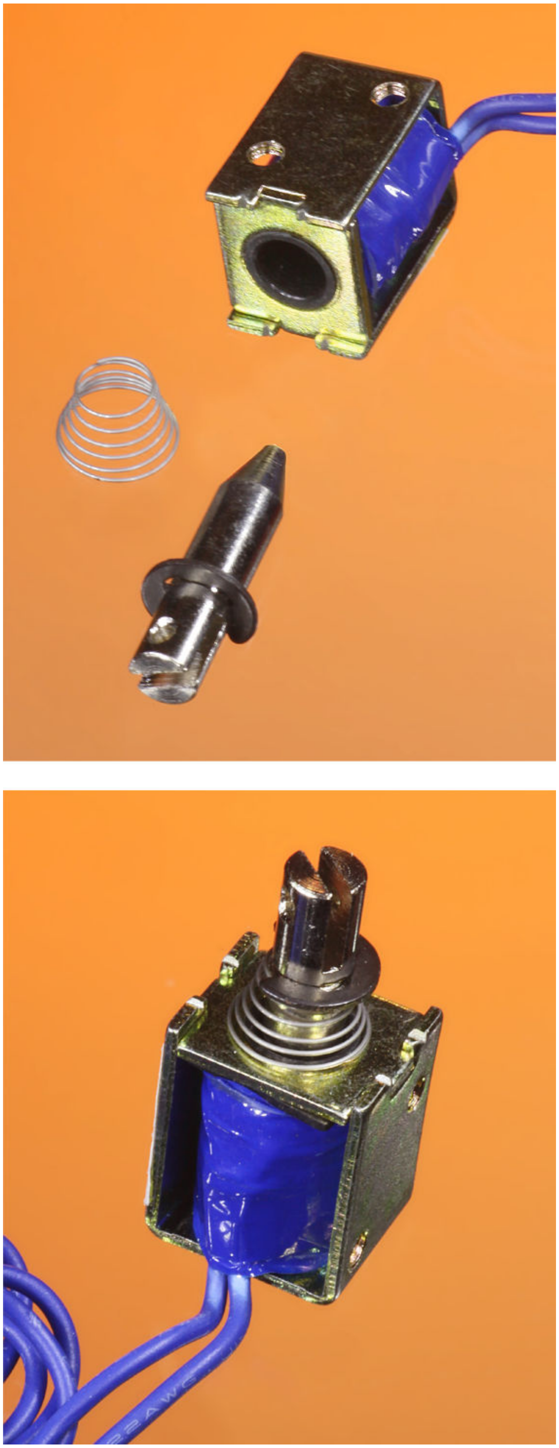

Figure 21-1. *A small 12VDC solenoid.*

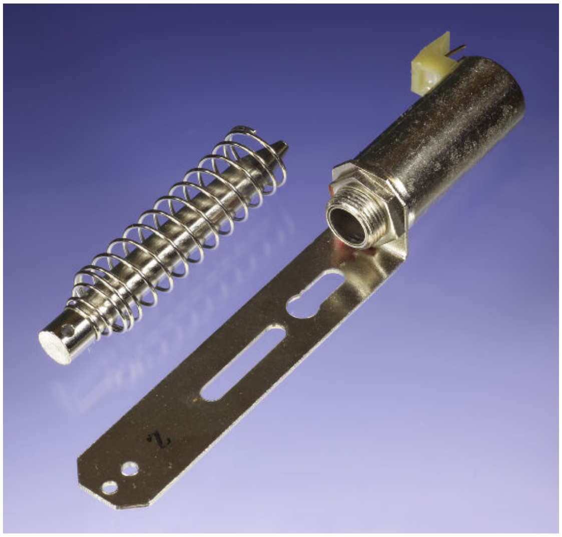

Figure 21-2. *A larger solenoid rated for 24VDC.*

Because there is no standardized schematic symbol for a solenoid, and because this type of component is so widely used in conjunction with valves, any diagram involving solenoids is more likely to emphasize fluid or gas flow with symbols that have been developed for that purpose. In such circuits, a solenoid may be represented simply by a rectangle. However, the symbols shown in Figure 21-4 may occasionally be found.

How It Works

Current flowing through the coil creates a magnetic force. This is explained in the **inductor** entry of this encyclopedia, using diagrams in Figure 14-3, Figure 14-4, Figure 14-5, and Figure 14-6.

If the plunger is fabricated from a material such as soft iron, the coil will induce an equal and opposite magnetic polarity in the plunger. Consequently the plunger will attempt to occupy a position inside the coil where the ends of the plunger are equal distances from the ends of the coil. If a collar is added to the free end of the plunger, this can increase the pulling force on the plunger when it is near the end of its throw because of the additional magnetic pull distributed between the collar and the frame of the solenoid.

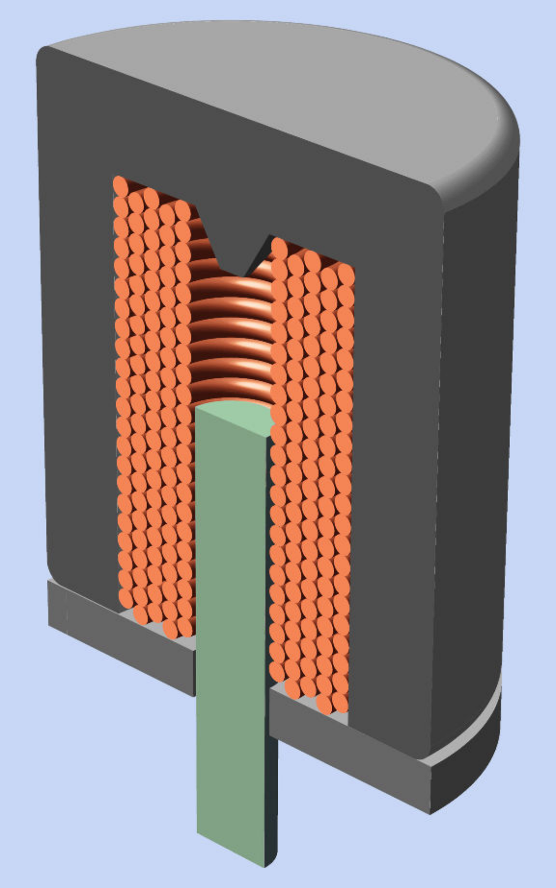

Figure 21-3. *A simplified view of a solenoid cut in half, showing the primary parts.*

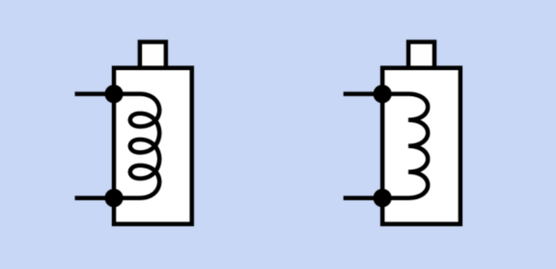

Figure 21-4. *Although no standard schematic symbol for a solenoid exists, these symbols may sometimes be found.*

A spring can be inserted to apply some resistive force to compensate for the increase in pulling force that occurs as a larger proportion of the plunger enters the coil. A spring may also be used to eject the plunger, partially at least, when current to the coil is interrupted.

If the plunger is a permanent magnet, reversing DC current to the coil will reverse the action of the plunger.

A solenoid with a nonmagnetized plunger may be energized by AC current, since polarity reversals in the magnetic field generated by the coil will induce equal and opposite reversals in the polarity of the plunger. However, the force curve of an AC-powered solenoid will be different from the force curve of a DC-powered solenoid. See Figure 21-5. The alternating current is likely to induce humming, buzzing, and vibration.

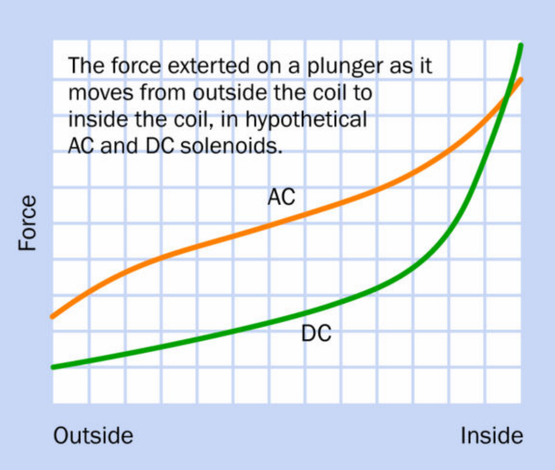

Figure 21-5. *A comparison of the force exerted on a plunger, relative to its position as it enters the coil, in hypothetical AC and DC solenoids.*

The frame of the solenoid increases the magnetic power that the coil can exert by providing a magnetic circuit of much lower *reluctance* than that of air (reluctance being the magnetic equivalent of electrical resistance). For a lengthier discussion of this effect, see "Magnetic Core" (page 122) in the **inductor** entry of this encyclopedia. If current flowing through the coil increases to the point where the frame becomes magnetically *saturated*, the pulling power of the solenoid will level off abruptly.

The heat generated by a solenoid when it is maintained in its energized state may be reduced if the manufacturer includes a series resistor and a switch that functions as a *bypass switch*. The switch is normally closed, but is opened mechanically when the plunger reaches the end of its throw, thus diverting electricity through the series resistor. This itself will generate some heat as a result of the current flowing through it, but by increasing the total resistance of the system, the total heat output will be reduced. The resistor value is chosen to provide the minimum power needed to retain the plunger at the end of its throw.

Variants

The most common variant is tubular, with open-frame as a secondary option. A tubular solenoid has been shown in Figure 21-2.

Additional variants include:

Low Profile

A shorter, fatter solenoid which may be used if a short throw is acceptable.

Latching

A permanent magnet holds the plunger when it reaches the end of its travel, and continues to hold it after power to the solenoid is disconnected. The plunger itself is also a permanent magnet, and is released by running current of reverse polarity through the coil.

Rotary

This variant is similar in principle to a brushless DC motor and causes the armature to rotate through a fixed angle (typically ranging from 25 to 90 degrees) instead of moving linearly. It is used as a mechanical indicator in control panels, although it is being displaced by purely electronic indicators.

Hinged Clapper

Instead of a plunger, a small hinged panel (the "clapper") moves in when the solenoid is active, and springs back when the power is interrupted.

Values

The stroke length, duty cycle, and holding force are the most significant values found in solenoid datasheets.

Holding forces for DC solenoids can range from a few grams to hundreds of kilograms. The holding force will be inversely proportional to the length of the solenoid, if all other variables are equal. The force that the solenoid can exert on its plunger also varies depending on the position of the plunger in the length of its throw.

Duty cycle is of special importance because the solenoid continues to draw power and create heat so long as it is holding the plunger at the end of its throw (assuming the solenoid is not the latching type). The initial current surge in an AC solenoid generates additional heat.

The duty cycle is simply calculated. If T1 is the time for which the solenoid is on and T2 is the time for which the solenoid is off, the duty cycle, D, is derived as a percentage from the formula

```
D = 100 * (T1 / (T1 + T2))
```

Some solenoids are designed to withstand a 100% duty cycle, but many are not, and in those cases, there is a maximum value not only for D but for the peak "on" time, regardless of the duty cycle. Suppose a solenoid is rated for a 25% duty cycle. If the solenoid is appropriately switched on for one second and off for three seconds, the heat will be allowed to dissipate before it has time to reach overload levels. If the solenoid is switched on for one minute and off for three minutes, the duty cycle is still 25%, but the heat that may accumulate during a one-minute "on" cycle may overload the component before the "off" cycle can allow it to dissipate.

Coil Size vs. Power

Because additional windings in a coil will induce a greater magnetic force, a larger solenoid tends to be more powerful than a smaller solenoid. However this means that if a larger and a smaller solenoid are both designed to generate the same force over the same distance, the smaller solenoid will probably draw more current (and will therefore generate more heat) because of its fewer coil windings.

How to Use it

Solenoids are primarily used to operate valves in fluid and gas circuits. Such circuits are found in laboratory and industrial process control, fuel injectors, aircraft systems, military applications, medical devices, and space vehicles. Solenoids may also be used in some electronic locks, in pinball machines, and in robotics.

What Can Go Wrong

Heat

Overheating is the principal concern when using solenoids, especially if the maximum "on" time is exceeded, or the duty cycle is exceeded. If the plunger is prevented from reaching the end of its throw, this can be another cause of overheating.

Because coil resistance increases with heat, a hot solenoid passes less current and therefore develops less power. This effect is more pronounced in a DC solenoid than an AC solenoid. A manufacturer's force curve should show the solenoid performance at its maximum rated temperature, which is typically around 75 degrees Centigrade, in a hypothetical ambient temperature of 25 degrees Centigrade. Exceeding these values may result in the solenoid failing to perform. As in all coils using magnet wire, there is the risk of excessive heat melting the insulation separating the coil windings, effectively shortening the coil, which will then pass more current, generating more heat.

AC Inrush

When an AC solenoid reaches the end of its travel, the sudden stop of the plunger results in forward *EMF* that generates additional heat. Generally speaking, a longer stroke creates a greater surge. Rapid cycling will therefore exacerbate coil heating.

Unwanted EMF

Like any device containing a coil, a solenoid creates back EMF when power is connected, and forward EMF when the power is disconnected. A protection diode may be necessary to suppress power spikes that can affect other components.

Loose Plunger

The plunger in many solenoids is not anchored or retained inside the frame and may fall out if the solenoid is tilted or subjected to extreme vibration.

DC motor

22

In this section, the term "traditional DC motor" is used to describe the oldest and simplest design, which consists of two *brushes* delivering power via a rotating, sectioned *commutator* to two or more electromagnetic coils mounted on the motor shaft. *Brushless* DC motors (in which DC is actually converted to a pulse train) are also described here because "brushless DC" has become a commonly used phrase, and the motor is powered by direct current, even though this is modified internally via *pulse-width modulation*.

OTHER RELATED COMPONENTS

- **AC motor** (See Chapter 23)
- **stepper motor** (See Chapter 25)
- **servo motor** (See Chapter 24)

What It Does

A traditional DC motor uses direct current to create magnetic force, which turns an output shaft. When the polarity of the DC voltage is reversed, the motor reverses its direction of rotation. Usually, the force created by the motor is equal in either direction.

How It Works

Current passes through two or more coils that are mounted on the motor shaft and rotate with it. This assembly is referred to as the *rotor*. The magnetic force produced by the current is concentrated via cores or poles of soft iron or high-silicon steel, and interacts with fields created by permanent magnets arrayed around the rotor in a fixed assembly known as the *stator*.

Power to the coils is delivered through a pair of *brushes*, often made from a graphite compound. Springs press the brushes against a sleeve that rotates with the shaft and is divided into sections, connected with the coils. The sleeve assembly is known as the *commutator*. As the commutator rotates, its sections apply power from the brushes to the motor coils sequentially, in a simple mechanical switching action.

The most elementary configuration for a traditional DC motor is shown in Figure 22-1.

In reality, small DC motors typically have three or more coils in the rotor, to provide smoother operation. The operation of a three-coil motor is shown in Figure 22-2. The three panels in this figure should be seen as successive snapshots of one motor in which the rotor turns progressively counter-clockwise. The brushes are colored red and blue to indicate positive and negative voltage supply, respectively. The coils are wired in series, with power being applied through the commutator to points between each pair of coils. The direction of current through each coil determines its magnetic polarity, shown as N for north or S for south. When two coils are energized in series without any power applied to their midpoint, each develops a smaller magnetic field than an individually energized coil. This is

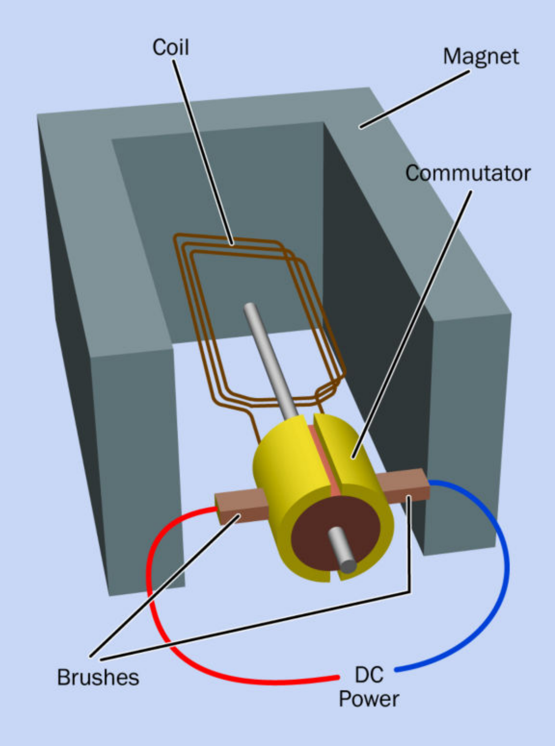

Figure 22-1. *The simplest traditional DC motor contains these parts. The combination of coil, shaft, and commutator is the rotor. The fixed magnetic structure in which it rotates is the stator.*

indicated in the diagram with a smaller white lowercase n and s. When two ends of a coil are at equal potential, the coil produces no magnetic field at all.

The stator consists of a cylindrical permanent magnet, which has two poles—shown in the figure as two black semicircles separated by a vertical gap for clarity—although in practice the magnet may be made in one piece. Opposite magnetic poles on the rotor and stator attract each other, whereas the same magnetic poles repel each other.

DC motors may be quite compact, as shown in Figure 22-3, where the frame of the motor measures about 0.7″ square. They can also be very powerful for their size; the motor that is shown disassembled in Figure 22-4 is from a 12VDC bilge pump rated at 500 gallons per hour. Its output was delivered by the small impeller attached to the rotor at right, and was achieved by using

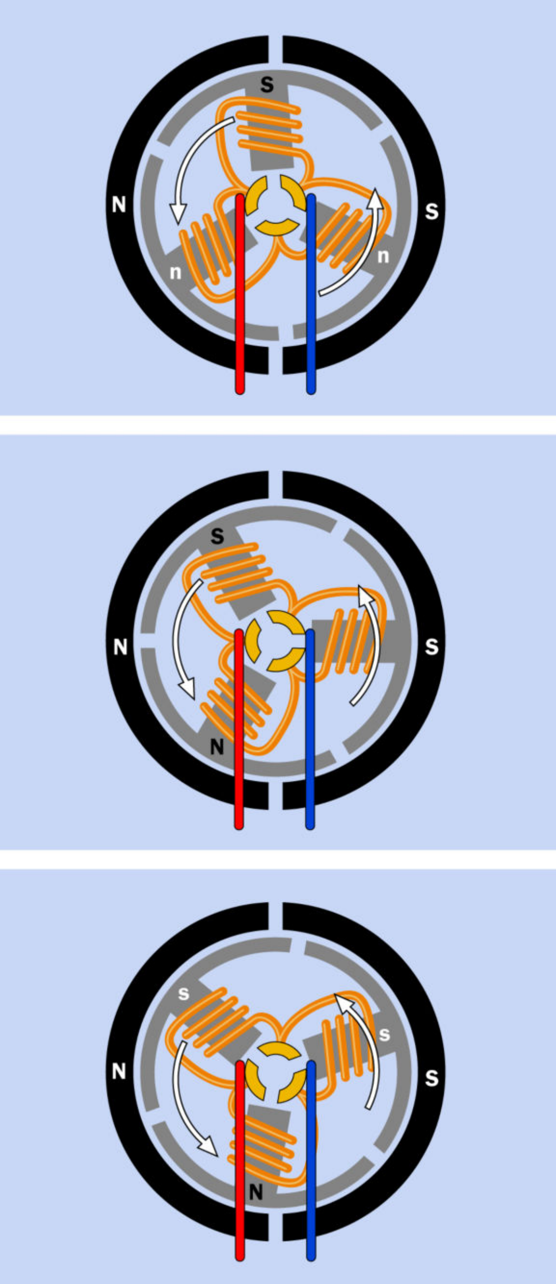

Figure 22-2. *Three sequential views of a typical three-coil DC motor viewed from the end of its shaft (the shaft itself is not shown). Magnetic effects cause the rotor to turn, which switches the current to the coils via the commutator at the center.*

two extremely powerful neodymium magnets, just visible on the inside of the motor's casing (at top-left) in conjunction with five coils on the rotor.

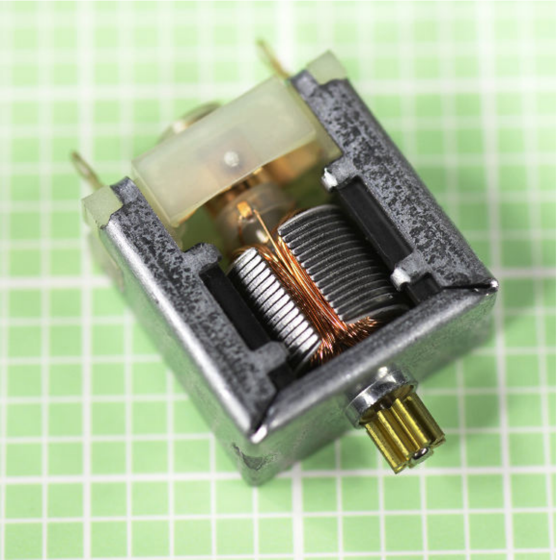

Figure 22-3. *A miniature 1.5VDC motor measuring about 0.7" square.*

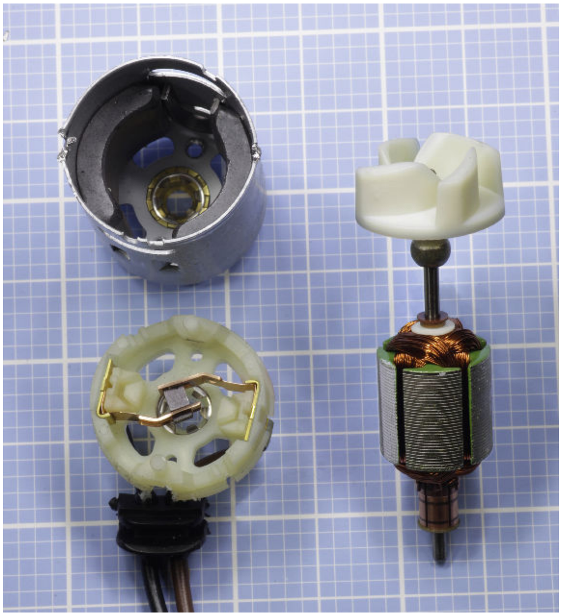

Figure 22-4. *A traditional DC motor removed from its cylindrical casing. The brushes of the motor are attached to the white plastic end piece at bottom-left. Large squares on the graph paper in the background are 1" × 1", divided in 0.1" increments. The motor was used in a small bilge pump.*

Variants

Coil Configurations

The series connection of coils used in Figure 22-2 is known as the *delta configuration*. The alternative is the *wye configuration* (or *Y configuration*, or *star configuration*). Simplified schematics are shown in Figure 22-5. Generally speaking, the delta configuration is best suited to high-speed applications, but provides relatively low torque at low speed. The wye configuration provides higher torque at low speed, but its top speed is limited.

Gearhead Motor

A *gearhead motor* (also often known as a *gear motor*) incorporates a set of reduction gears that increase the torque available from the output shaft while reducing its speed of rotation. This is often desirable as an efficient speed for a traditional DC motor may range from 3,000 to 8,000 RPM, which is too fast for most applications. The gears and the motor are often contained in a single sealed cylindrical package. Two examples are shown in Figure 22-6. A disassembled motor, revealing half of its gear train under the cap and the other half still attached to a separate circular plate, appears in Figure 22-7. When the motor is assembled, the gears engage. As in the case of the bilge-pump motor, the stator magnets are mounted inside the cylindrical casing. Note that the brushes, inside the circular plate of white plastic, have a resistor and capacitor wired to suppress voltage spikes.

Spur gears are widely used for speed reduction. *Planetary gears* (also known as *epicyclic gears*) are a slightly more expensive option. Spur gears such as those in Figure 22-8 may require three or more pairs in series. The total speed reduction is

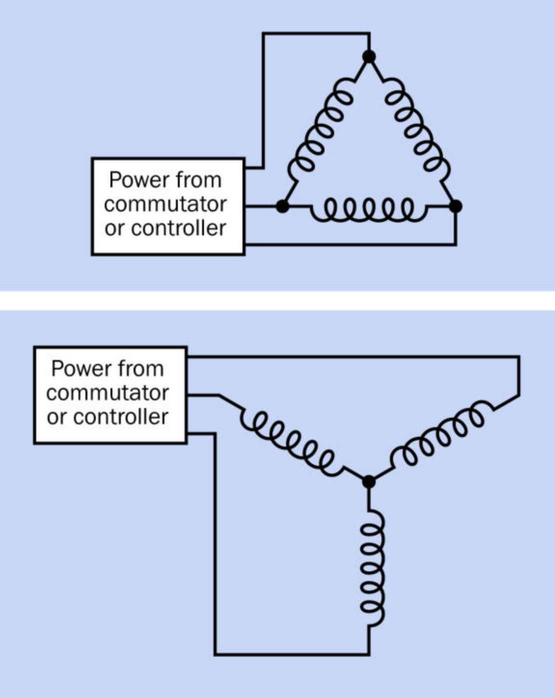
Power from commutator or controller
Power from commutator or controller

Figure 22-5. *Coils on the rotor of a traditional DC motor may be connected in delta configuration (top) or wye configuration (bottom).*

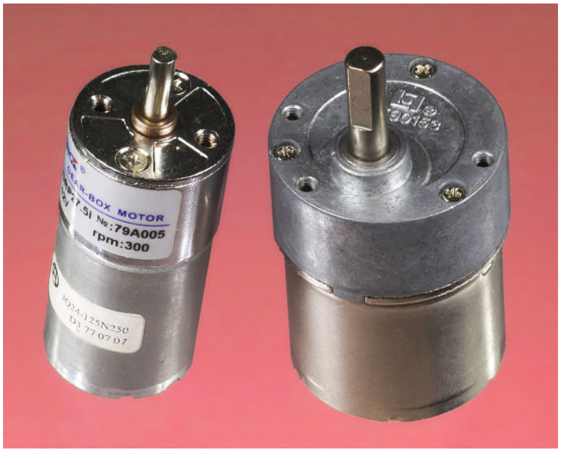

Figure 22-6. *Two typical small gearhead motors.*

found by multiplying the individual ratios. Thus, if three pairs of gears have ratios of 37 : 13, 31 : 15, and 39 : 17, the total speed reduction ® is obtained by:

```
R = ( 37 * 31 * 39 ) / ( 13 * 15 * 17 )
```

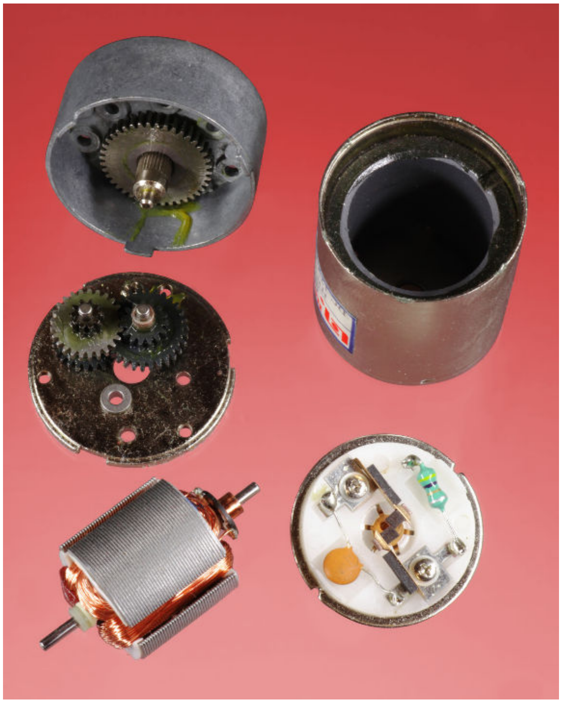

Figure 22-7. *Spur gears from a gearhead motor provide speed reduction and increased torque.*

Therefore:

```
R = 44733 / 3315 = about 13.5 : 1
```

Datasheets almost always express R as an integer. For example, the gear train shown in Figure 22-7 is rated by the manufacturer as having an overall reduction of 50:1. In reality, the reduction can be expected to have a fractional component. This is because if two gears have an integer ratio, their operating life will be shortened, as a manufacturing defect in a tooth in the smaller gear will hit the same spots in the larger gear each time it rotates. For this reason, the numbers of teeth in two spur gears usually do not have any common factors (as in the example above), and if a motor rotates at 500 RPM, a gear ratio stated as 50:1 is very unlikely to produce an output of exactly 10

RPM. Since traditional DC motors are seldom used for applications requiring high precision, this is not usually a significant issue, but it should be kept in mind.

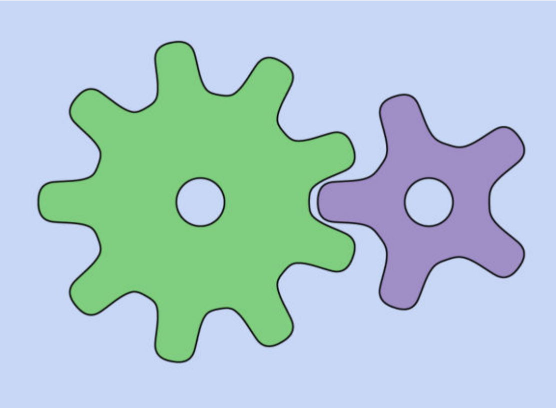

Figure 22-8. *A pair of spur gears.*

Figure 22-9 shows *planetary gears*, also known as *epicyclic gears*. The outer ring gear is properly referred to as the *annulus*, while the *sun gear* is at the center, and the intermediate *planet gears* may be mounted on a *carrier*. The greatest speed reduction will be achieved by driving the sun gear while the annulus is kept in a stationary position and the output is taken from the carrier of the planet gears. If A is the number of teeth in the annulus and S is the number of teeth in the sun gear, the total speed reduction, R, is given by the following formula:

```
R = (S + A) / S
```

Note that in this drive configuration, the number of teeth in each planet gear is irrelevant to the speed reduction. In Figure 22-9, the sun gear has 27 teeth whereas the annulus has 45 teeth. Therefore, the reduction is found by:

```
R = (27 + 45) / 27 = about 2.7 : 1
```

Successive reductions can be achieved by stacking planetary gear sets, using the carrier of one set to drive the sun gear in the next set.

Planetary gears are used primarily if a motor drives a heavy load, as the force is divided among more gear pairs, reducing wear and tear on gear teeth and minimizing the breakdown of lubrication. A planetary gear train may also be more compact than a train of spur gears. These advantages must be evaluated against the higher price and slightly increased friction resulting from the larger number of gears interacting with each other.

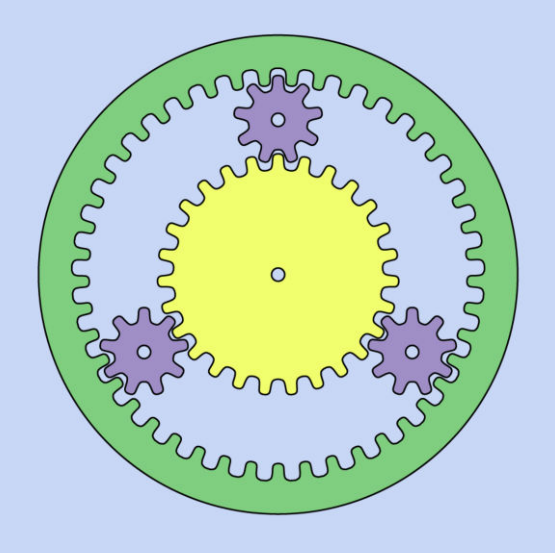

Figure 22-9. *Planetary gears, also known as epicyclic gears, share the torque from a motor among more teeth than simple spur gears.*

Brushless DC Motor

In a *brushless* DC motor, sometimes referred to as a *BLDC motor*, the coils are located in the stator and the permanent magnets are relocated in the rotor. The great advantage of this design is that power can be applied directly to the coils, eliminating the need for brushes, which are the primary source of failure in DC motors as a result of wear and tear. However, since there is no rotating commutator to switch the DC current to the coils, the current must be switched by electronic components, which add to the cost of the motor.

In the *inrunner* configuration the stator surrounds the rotor, whereas in the *outrunner* configuration the stator is located in the center of the motor while the rotor takes the form of a ring or cup that spins around the stator. This is a common design for small cooling fans, where the blades are attached to the outer circumference of a cup that is lined with permanent magnets. An example is shown in Figure 22-10. In this picture, the stator coils are normally hidden from view, being fixed to the fan housing (shown at the top of the picture). Power is controlled by the surface-mount components on the green circular circuit board. The cup attached to the fan blades contains permanent magnets.

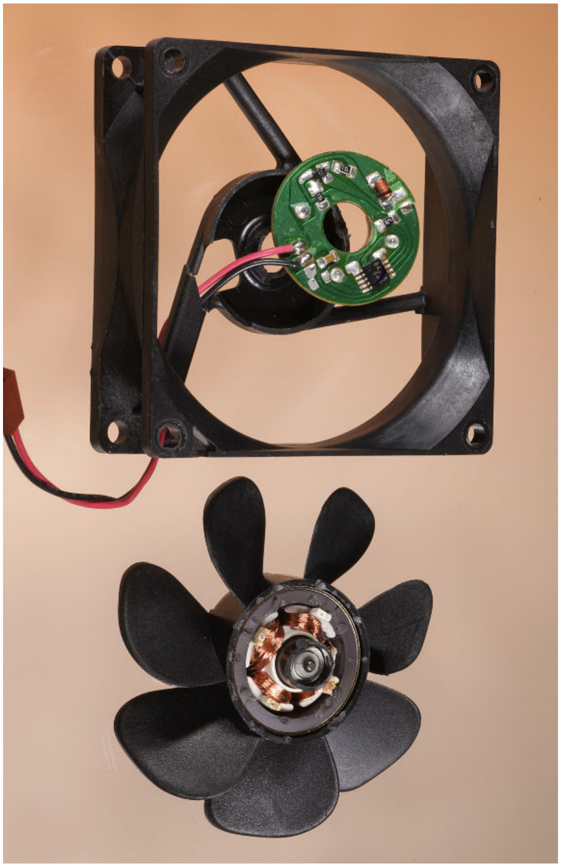

Figure 22-10. *A typical brushless DC cooling fan uses stationary coils, with permanent magnets rotating around them.*

The use of a solid-state switching system to energize the coils sequentially is known as *electronic commutation*. *Hall effect* sensors may be used to detect the position of the rotor and feed this information back to the frequency control circuit, so that it stays "one step ahead" of the rotor (when bringing it up to speed) or is synchronized with the rotor (for a constant running speed). The system is comparable to a *reluctance motor* or *synchronous motor.* These variants are described in the **AC motor** section of this encyclopedia.

While traditional DC motors have been commercially available since the late 1800s, brushless DC motors were not introduced until the 1960s, when the availability of solid-state control electronics began to make the motor design economically viable.

Linear Actuator

Linear actuator is a generic term for any device that can exert a pushing or pulling force in a straight line. In industrial applications, actuators may be powered pneumatically or hydraulically, but smaller-scale units are usually driven by a traditional DC motor. These are more properly (but not often) referred to as *electromechanical linear actuators*.

The rotational force of the motor is typically converted to linear motion by using a threaded motor shaft in conjunction with a nut or collar. The unit is often mounted in an enclosure containing *limit switches* that stop the motor automatically at the limits of travel. For an explanation of limit switches, see "Limit Switches" (page 46) in the **switch** entry in this encyclopedia.

Values

A manufacturer's datasheet should list the maximum operating voltage and typical current consumption when a motor is moderately loaded, along with the *stall current* that a motor draws when it is so heavily loaded that it stops turning. If stall current is not listed, it can be determined empirically by inserting an ammeter (or multimeter set to measure amperes) in series with the motor and applying a braking force until the motor stops. Motors should generally be protected with slow-blowing **fuses** to allow for the power fluctuations that occur when the motor starts running or experiences a change in load.

In addition, the *torque* that a motor can deliver should be specified. In the United States, torque is often expressed in *pound-feet* (or *ounce-inches* for smaller motors). Torque can be visualized by imagining an arm pivoted at one end with a weight hung on the other end. The torque exerted at the pivot is found by multiplying the weight by the length of the arm.

In the metric system, torque can be expressed as gram-centimeters, Newton-meters, or dyne-meters. A Newton is 100,000 dynes. A dyne is defined as the force required to accelerate a mass of 1 gram, increasing its velocity by 1 centimeter per second each second. 1 Newton-meter is equivalent to approximately 0.738 pound-feet.

The speed of a traditional DC motor can be adjusted by varying the voltage to it. However, if the voltage drops below 50% of the rated value, the motor may simply stop.

The power delivered by a motor is defined as its speed multiplied by its torque at that speed. The greatest power will be delivered when the motor is running at half its *unloaded speed* while delivering half the *stall torque*. However, running a motor under these conditions will usually create unacceptable amounts of heat, and will shorten its life.

Small DC motors should be run at 70% to 90% of their unloaded speed, and at 10% to 30% of the stall torque. This is also the range at which the motor is most efficient.

Ideally, DC motors that are used with reduction gearing should be driven with less than their rated voltage. This will prolong the life of the motor.

When choosing a motor, it is also important to consider the *axial loading* (the weight or force that will be imposed along the axis or shaft of the motor) and *radial loading* (the weight or force that will be imposed perpendicularly to the axis). Maximum values should be found in motor datasheets.

In the hobby field, motors for model aircraft are typically rated in watts-per-pound of motor weight (abbreviated w/lb). Values range from 50 to 250 w/lb, with higher values enabling better performance.

Relationships between torque, speed, voltage, and amperage in a traditional DC motor can be described easily, assuming a hypothetical motor that is 100% efficient:

If the amperage is constant, the torque will also be constant, regardless of the motor speed.

If the load applied to the motor remains constant (thus forcing the motor to apply a constant torque), the speed of the motor will be determined by the voltage applied to it.

If the voltage to the motor remains constant, the torque will be inversely proportional with the speed.

How to Use it

A traditional DC motor has the advantages of cheapness and simplicity, but is only suitable for intermittent use, as its brushes and commutator will tend to limit its lifetime. Its running speed will be approximate, making it unsuitable for precise applications.

As the cost of control electronics has diminished, brushless DC motors have replaced traditional DC motors. Their longevity and controllability provide obvious advantages in applications such as hard disk drives, variable-speed computer fans, CD players, and some workshop tools. Their wide variety of available sizes, and good power-to-weight ratio, have encouraged their adoption in toys and small vehicles, ranging from remote-controlled model cars, airplanes, and helicopters to personal transportation devices such as the Segway. They are also used in direct-drive audio turntables.

Where an application requires the rotation of a motor shaft to be converted to linear motion, a prepackaged *linear actuator* is usually more reliable and simpler than building a crank and connecting rod, or cam follower, from scratch. Large linear actuators are used in industrial automation, while smaller units are popular with robotics hobbyists and can also be used to control small systems in the home, such as a remote-controlled access door to a home entertainment center.

Speed Control

A *rheostat* or **potentiometer** may be placed in series with a traditional DC motor to adjust its speed, but will be inefficient, as it will achieve a voltage drop by generating heat. Any rheostat must be rated appropriately, and should probably be wire-wound. The voltage drop between the wiper and the input terminal of the rheostat should be measured under a variety of operating conditions, along with the amperage in the circuit, to verify that the wattage rating is appropriate.

Pulse-width modulation (*PWM*) is preferable as a means of speed control for a traditional DC motor. A circuit that serves this purpose is sometimes referred to as a *chopper*, as it chops a steady flow of current into discrete pulses. Usually the pulses have constant frequency while varying in duration. The pulse width determines the average delivered power, and the frequency is sufficiently high that it does not affect smoothness of operation of the motor.

A **programmable unijunction transistor** or PUT can be used to generate a train of pulses, adjustable with a potentiometer attached to its emitter. Output from the transistor goes to a *silicon-controlled rectifier* (*SCR*), which is placed in series with the motor, or can be connected directly to the motor if the motor is small. See Figure 27-7.

Alternatively, a *555 timer* can be used to create the pulse train, controlling a *MOSFET* in series with the motor.

A **microcontroller** can also be used as a pulse source. Many microcontrollers have PWM capability built in. The microcontroller will require its own regulated power supply (typically 5VDC, 3.3VDC, or sometimes less) and a switching component such as an *insulated-gate bipolar transistor* (*IGBT*) to deliver sufficient power to the motor and to handle the flyback voltage. These components will all add to the cost of the system, but many modern devices incorporate microcontrollers anyway, merely to process user input. Another advantage of using a microcontroller is that its output can be varied by rewriting software, for example if a motor is replaced with a new version that has different characteristics, or if requirements change for other reasons. Additionally, a microcontroller enables sophisticated features such as pre-programmed speed sequences, stored memory of user preferences, and/or responses to conditions such as excessive current consumption or heat in the motor.

A PWM schematic using a microcontroller and IGBT is shown in Figure 22-11.

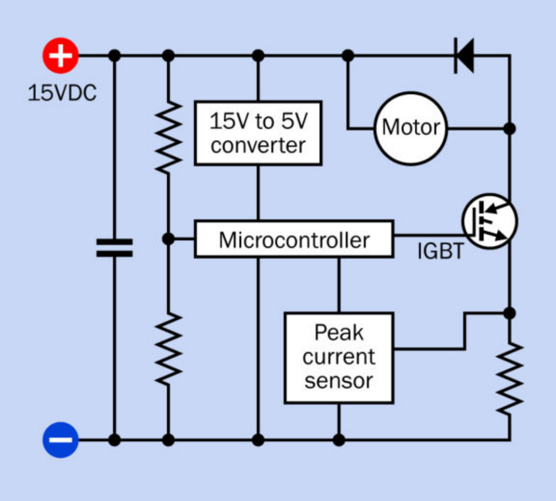

Figure 22-11. *A sample schematic for control of a DC motor via pulse-width modulation, using a microcontroller and an insulated-gate bipolar transistor.*

Direction Control

The *H bridge* is a very early system for reversing the direction of a DC motor simply by swapping the polarity of its power supply. This is shown in Figure 22-12. The switches diagonally opposite each other are closed, leaving the other two switches open; and then to reverse the motor, the switch states are reversed. This is obviously a primitive scheme, but the term "H bridge" is still used when prepackaged in a single chip such as the LMD18200 H bridge motor controller from National Semiconductor.

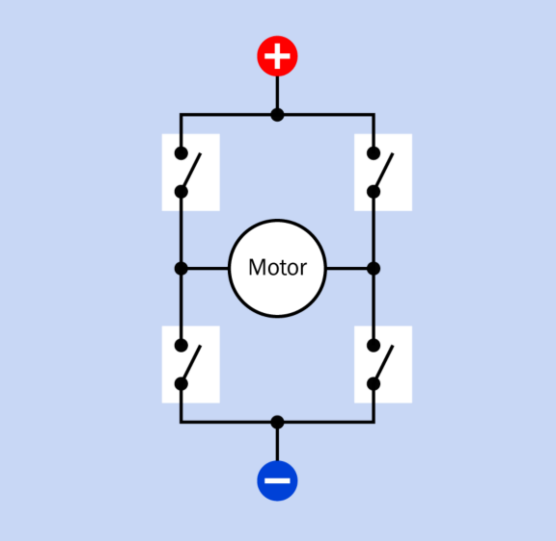

Figure 22-12. *A DC motor can be reversed by this very basic circuit, known as an H bridge, by opening and closing pairs of switches that are diagonally opposite each other.*

A double-throw, double-pole **switch** or **relay** can achieve the same purpose, as shown in Figure 22-13.

Limit Switches

When a traditional DC motor is used reversibly within a restricted range of motion, it can be fitted with *limit switches* to prevent the motor from stalling and burning out at either end of its permitted travel. Limit switches are explained in "Limit Switches" (page 46) in the **switch** entry in this encyclopedia.

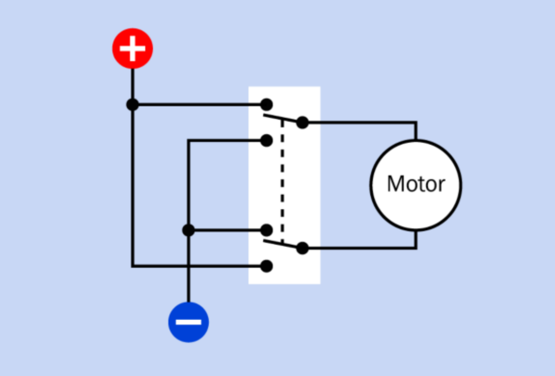

Figure 22-13. *A DPDT switch or relay can reverse the direction of a traditional DC motor simply by swapping the polarity of the power supply.*

What Can Go Wrong

Brushes and Commutator

The primary cause of failure in DC motors is abrasion of the brushes and wear and tear, oxidation, and/or accumulation of dirt on the commutator. Some motors are designed to allow replacement of the brushes; sealed motors and gearhead motors generally are not. High current and high speed will both tend to accelerate wear in the areas where the brushes meet the commutator.

Electrical Noise

The intermittent contact between the brushes and sections of the commutator of a traditional DC motor can induce voltage spikes that may travel back into the power supply for the motor and cause seemingly random effects in other components. Sparking in the commutator can be a significant source of *electromagnetic interference* (*EMI*), especially where cheap or poorly fitted brushes are used. Even if the commutator is running cleanly, the rapid creation of a magnetic field in a motor winding, following by collapse of the field, can create spikes that feed back into the power supply.

Wires that power a motor should be in twisted-pair configuration, so that their radiated EMI tends to cancel itself out. They should be routed away from data lines or encoder outputs, and may be shielded if necessary. Data lines from sensors in brushless motors may also be shielded.

Installing a capacitor across the motor terminals can significantly reduce EMI. Some motors have capacitors preinstalled. If the motor is in a sealed casing, it may have to be disassembled to reveal whether a capacitor is present.

Heat effects

Since all motors in the real world are less than 100% efficient, some power is lost by the motor during normal operation, and will be dissipated as heat. The resistance of motor windings, and consequently the magnetic force that they generate, will decrease as the temperature rises. The motor becomes less efficient, and will try to draw more current, worsening the situation. A manufacturer's rating for maximum temperature should be taken seriously.

The insulation of the coil windings is usually the most vulnerable part of a motor if excess heat persists. Short circuits between adjacent coils as a result of insulation breakdown will degrade the performance of the motor while increasing its power consumption, which will create even more heat.

Where motor casings have protruding ridges, these cooling fins should be exposed to ambient air.

Frequent starting, stopping, and reversing will tend to generate heat as a result of power surges, and will reduce the lifetime of the motor.

Ambient Conditions

A warm, dry environment will tend to dry out bearing lubricants and graphite brushes. Conversely, a very cold environment will tend to thicken the bearing lubricants. If a motor will be used in unusual environmental conditions, the manufacturer should be consulted.

Wrong Shaft Type or Diameter

Motors have a variety of possible output shaft diameters, some measured in inches and others in millimeters, and shafts may be long or short, or may have a D-shaped cross section or splines to mate with appropriate accessories such as gears, pulleys, or couplings. Careful examination of datasheets and part numbers is necessary to determine compatibility. In the hobby-electronics world, retailers may offer purpose-built discs or arms for specific motor shafts.

Incompatible Motor Mounts

Mounting lugs or flanges may or may not be provided, and may be incompatible with the application for which the motor is intended. The same motor may be available with a variety of mount options, differentiated only by one letter or digit in the motor's part number. A mount option that was available in the past may become obsolete or may simply be out of stock. Again, examination of datasheets is necessary.

Backlash

Backlash is the looseness or "slack" in a gear train that results from small gaps between meshing gear teeth. Because backlash is cumulative when gears are assembled in series, it can become significant in a slow-output gearhead motor. When measured at the output shaft, it is generally in the range of 1 to 7 degrees, tending to increase as the load increases. If a geared motor is used as a positioning device, and is fitted with an encoder to count rotations of the motor shaft, control electronics may cause the motor to hunt to and fro in an attempt to overcome the hysteresis allowed by the backlash. A **stepper motor** or **servo motor** is probably better suited to this kind of application.

Bearings

When using a motor that is not rated for significant axial loading, the bearings may be damaged by applying excessive force to push-fit an output gear or pulley onto the motor shaft. Even minor damage to bearings can cause significant noise (see the following section) and a reduced lifetime for the component.

In brushless DC motors, the most common cause of failure is the deterioration of bearings. Attempting to revive the bearings by unsealing them and adding lubricant is usually not worth the trouble.

Audible Noise

While electric motors are not generally thought of as being noisy devices, an enclosure can act as a sounding board, and bearing noise is likely to increase over time. Ball bearings become noisy over time, and gears are inherently noisy.

If a device will contain multiple motors, or will be used in close proximity to people who are likely to be sensitive to noise (for example, in a medical environment), care should be taken to insure that motor shafts are properly balanced, while the motors may be mounted on rubber bushings or in sleeves that will absorb vibration.

AC motor

23

The distinction between AC and DC motors has become blurred as controllers for DC motors make increasing use of *pulse-width modulation*, which can be viewed as a form of alternating current. All motors that consume DC power are referenced in the **DC motor** section of this encyclopedia, regardless of whether they modulate the power internally. **Stepper motors** and **servo motors** are considered as special cases, each with its own entry. AC motors, described here, are those that consume alternating current, usually in the form of a sine wave with a fixed frequency.

OTHER RELATED COMPONENTS

- **DC motor** (see Chapter 22)
- **stepper motor** (see Chapter 25)
- **servo motor** (see Chapter 24)

What It Does

An AC motor uses a power supply of alternating current to generate a fluctuating magnetic field that turns a shaft.

How It Works

The motor consists primarily of two parts: the *stator*, which remains stationary, and the *rotor*, which rotates inside the stator. Alternating current energizes one or more coils in the stator, creating fluctuating magnetic fields that interact with the rotor. A simplified representation is shown in Figure 23-1, where the coils create magnetic forces indicated by the green arrows, N representing North and S representing South.

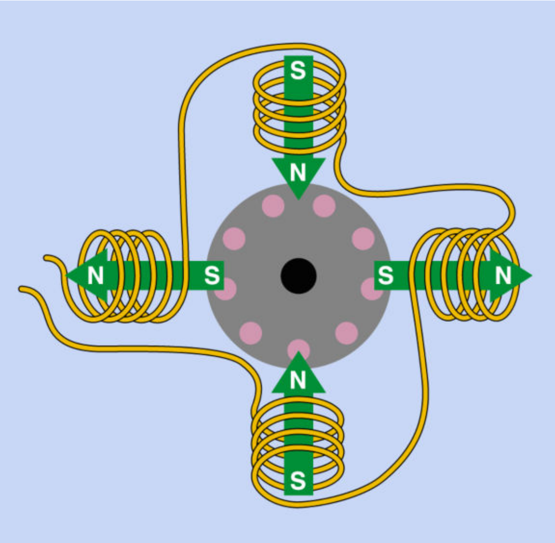

Figure 23-1. *A simplified representation of a basic AC motor. The green arrows indicate magnetic force.*

Stator Design

Plug-in electric fans typically use AC motors. The stator from a large electric fan is shown in Figure 23-2, where the large diameter of each coil maximizes its magnetic effect. The stator from a smaller electric fan is shown in Figure 23-3, in which only one coil is used. (The coil is wrapped in black electrical tape.)

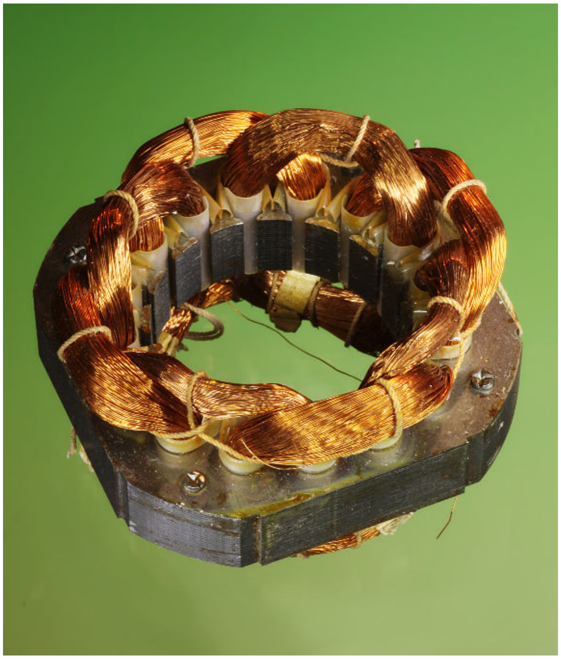

Figure 23-2. *The stator from a large electric fan. Each coil of copper wire is centered on a lug pointing inward to the hole at the center, where the rotor would normally be mounted. The coils overlap because their diameter is maximized to increase their magnetic effect. Each coil is tapped to allow speed selection in steps via an external rotary switch.*

The core of a stator resembles the core of a **transformer** in that it usually consists of a stack of wafers of high-silicon steel (or sometimes aluminum or cast iron). The layers are insulated from one another by thin layers of shellac (or a similar compound) to prevent eddy currents that would otherwise circulate through the entire thickness of the stator, reducing its efficiency.

The coil(s) wound around the stator are often referred to as *field windings*, as they create the magnetic field that runs the motor.

Rotor Design

In most AC motors, the rotor does not contain any coils and does not make any electrical connection with the rest of the motor. It is powered entirely by induced magnetic effects, causing this type of motor to be known generally as an *induction motor*.

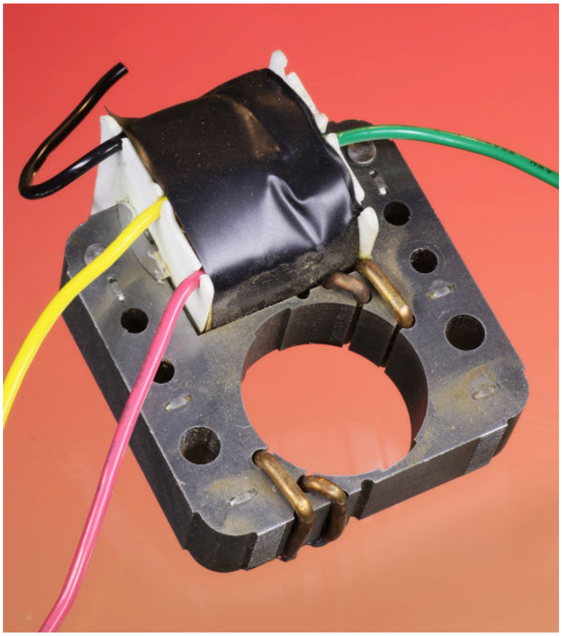

Figure 23-3. *This stator from a smaller electric fan uses only a single coil, wrapped in black tape. It is sufficient to induce a magnetic field but is generally less efficient than a motor using multiple coils.*

As the AC voltage changes from positive to negative, the magnetic force induced in the stator collapses and a new field of opposite polarity is created. Because the stator is designed to create an asymmetrical field, it induces a rotating magnetic field in the rotor. The concept of a rotating magnetic field is fundamental in AC motors.

Like the stator, the rotor is fabricated from wafers of high-silicon steel; embedded in the wafers are nonmagnetic rods, usually fabricated from aluminum but sometimes from copper, oriented approximately parallel to the axis of rotation. The rods are shorted together by a ring at each end of the rotor, forming a conductive "cage," which explains why this device is often referred to colloquially as a *squirrel cage* motor.

Figure 23-4 shows the configuration of a rotor cage with the surrounding steel wafers removed for clarity. In reality, the rods in the cage are almost always angled slightly, as shown in Figure 23-5, to promote smooth running and reduce *cogging*, or fluctuations in torque, which would otherwise occur.

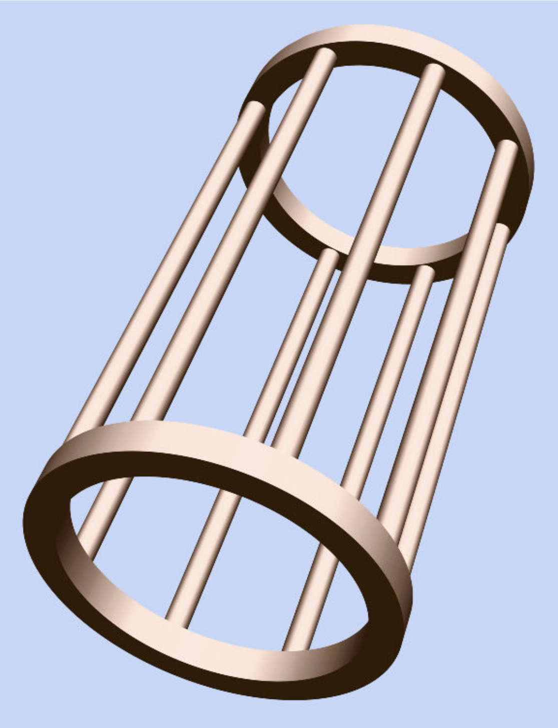

Figure 23-4. *The rotor of a typical AC motor contains a cage of aluminum (or sometimes copper) in which eddy currents occur, as a result of the rotating magnetic field inside the steel body of the rotor (which is omitted here for clarity). These currents cause their own magnetic fields, which interact with the fields generated by coils in the stator.*

In Figure 23-6, the steel wafers of a rotor are shown, with channels to accommodate an angled aluminum cage. Figure 23-7 shows a cross-section of a rotor with the cage elements in pale red and the steel wafer in gray.

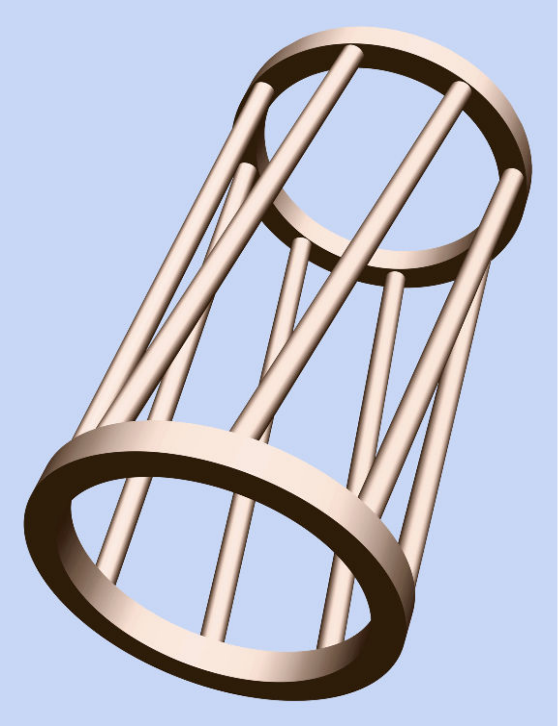

Figure 23-5. *To promote smooth running of the motor, the longitudinal elements of the cage are typically angled, as suggested in this rendering.*

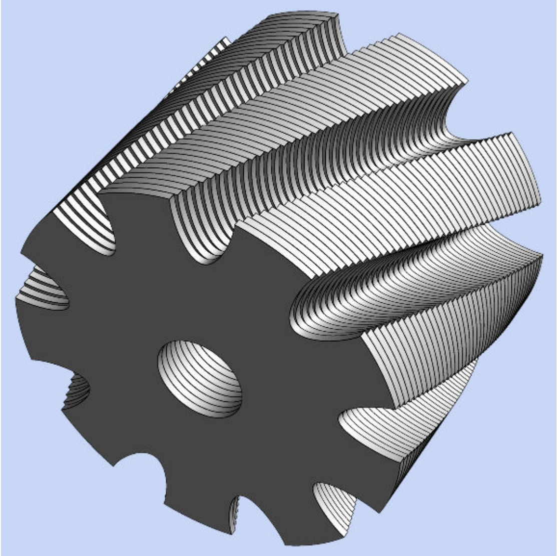

Figure 23-6. *The steel wafers in the rotor of an AC motor are typically offset as shown here. The channels are to accommodate a cage of aluminum or copper conductors.*

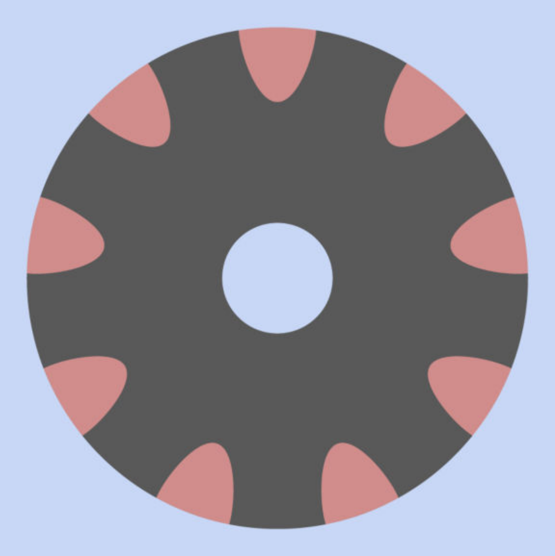

Figure 23-7. *Cross-section of a rotor with steel shown in gray and embedded elements of an aluminum cage shown in pale red.*

The actual rotor from an induction motor is shown in Figure 23-8. This rotor was removed from the stator shown in Figure 23-3. The bearings at either end of the rotor were bolted to the stator until disassembly.

Although the cage is nonmagnetic, it is electrically conductive. Therefore the rotating magnetic field that is induced in the steel part of the rotor generates substantial secondary electric current in the cage, so long as the magnetic field inside the rotor is turning faster than the rotor itself. The current in the longitudinal elements of the cage creates its own magnetic field, which interacts with the fields created by coils in the stator. Attraction and repulsion between these fields causes the rotor to turn.

Note that if the turning speed of the rotor rises to match the frequency of the alternating current powering the coils in the stator, the cage in the rotor is no longer turning through magnetic lines of force, and ceases to derive any power from them. In an ideal, frictionless motor, its unloaded operating speed would be in equilibrium with the AC frequency. In reality, an induction motor never quite attains that speed.

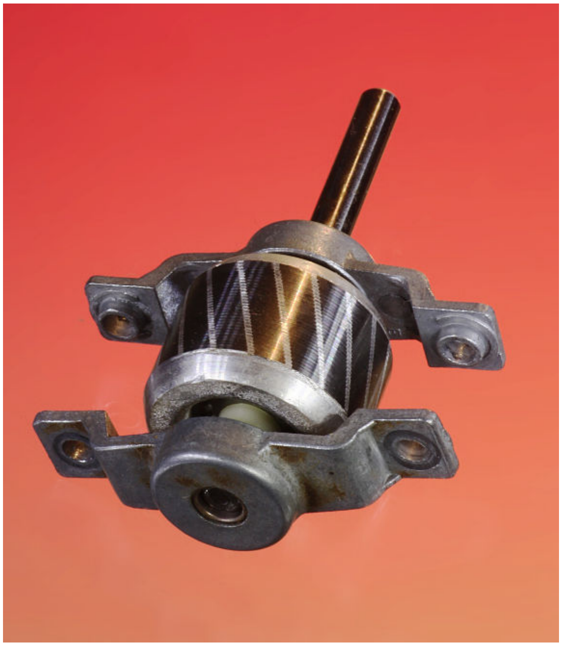

Figure 23-8. *The rotor from a small fan motor. The aluminum cage and its end pieces are the pale gray sections, steel plates are the darker sections.*

When power is applied while the rotor is at rest, the induction motor draws a heavy surge of current, much like a short-circuited **transformer**. Electrically, the coils in the stator are comparable to the primary winding of a transformer, while the cage in the rotor resembles the secondary winding. The turning force induced in the stationary rotor is known as *locked-rotor torque*. As the motor picks up speed, its power consumption diminishes. See Figure 23-9.

When the motor is running and a mechanical load is applied to it, the motor speed will drop. As the speed diminishes, the cage of conductors embedded in the rotor will derive more power, as they are turning more slowly than the rotating magnetic field. The speed of rotation of the field is determined by the frequency of the AC power, and is therefore constant. The difference in rota-

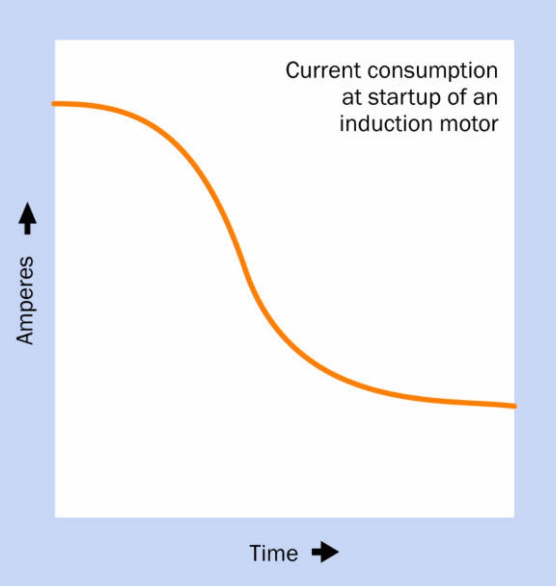

Figure 23-9. *An approximated graph showing the typical current consumption of an AC induction motor as it starts from rest and gains speed over a period of time.*

tional speed between the magnetic field and the rotor is known as *slip*. Higher levels of slip induce greater power, and therefore the induction motor will automatically find an equilibrium for any load within its designed range.

When running under full load, a small induction motor may have a slip value from 4 to 6 percent. In larger motors, this value will be lower.

Variants

Variants of the generic induction motor described above are generally designed to take advantage of either *single-phase* or *three-phase* alternating current.

A *synchronous motor* is a variant in which the rotor maintains a constant speed of rotation regardless of small fluctuations in load.

Some AC motors incorporate a *commutator*, which allows an external connection to coils mounted on the rotor, and can enable variable speed control.

A *linear motor* may consist of two rows of coils, energized by a sequence of pulses that can move a permanent magnet or electromagnet between the coils. Alternatively, the linear motor's coils may move as a result of magnetic interaction with a segmented fixed rail. Detailed description of linear motors is outside the scope of this encyclopedia.

Single-Phase Induction Motor

The majority of induction motors run on single-phase alternating current (typically, from domestic wall outlets). This type of motor is not innately self-starting because the stator coils and rotor are symmetrical. This tends to result in vibration rather than rotation.

To initiate rotation, the stator design is modified so that it induces an asymmetrical magnetic field, which is more powerful in one direction than the other. The simplest way to achieve this is by adding one or more *shorting coils* to the stator. Each shorting coil is often just a circle of heavy-gauge copper wire. This ploy reduces the efficiency of the motor and impairs its starting torque, and is generally used in small devices such as electric fans, where low-end torque is unimportant. Because the shorting coil obstructs some of the magnetic field, this configuration is often known as a *shaded pole* motor.

Copper shorting coils are visible in the fan motor shown in Figure 23-3.

A **capacitor** is a higher-cost but more efficient alternative to a shorting coil. If power is supplied to one or more of the stator coils through a capacitor, it will create a phase difference between these coils and the others in the motor, inducing an asymmetrical magnetic field. When the motor reaches approximately 80% of its designed running speed, a centrifugal switch may be included to take the capacitor out of the circuit, since it is no longer necessary. Switching out the capacitor and substituting a direct connection to the stator coils will improve the efficiency of the motor.

A third option to initiate rotation is to add a second winding in the stator, using fewer turns of smaller-gauge wire, which have a higher resistance than the main winding. Consequently the magnetic field will be angled to encourage the motor to start turning. This configuration is known as a *split-phase* induction motor, in which the starter winding is often referred to as the *auxilliary winding* and consists of about 30% of the total stator windings in the motor. Here again, a centrifugal switch can be incorporated, to eliminate the secondary winding from the circuit when the motor has reached 75 to 80 percent of its designed running speed.

The relationship between motor speed and torque of the three types of motors described above is shown in Figure 23-10. These curves are simplified and do not show the effect that would be produced by introducing a centrifugal switch.

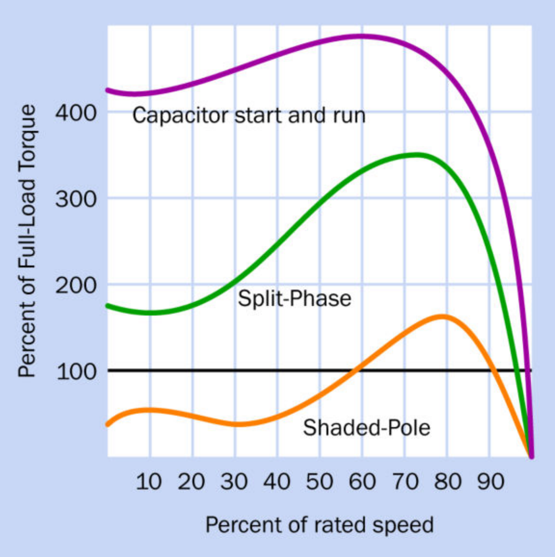

Figure 23-10. *Approximate curves showing the relationship between speed and torque for three types of single-phase induction motor. (Graph derived from AC Induction Motor Fundamentals published by Microchip Technology Inc.)*

Three-Phase Induction Motor

Larger induction motors are often *three-phase* devices. Three-phase AC (which is by far the most common form of *polyphase* AC) is delivered by a power utility company or generator via three wires, each of which carries alternating current with a phase difference of 120 degrees relative to the other two, usually for industrial applications. A common configuration of stator coils for a three-phase motor is shown in Figure 23-11. Since the three wires take it in turns to deliver their peak voltage, they are ideally suited to turn the stator of a motor via induction, and no shorting coil or capacitor is needed for startup. Heavy-duty 3-phase induction motors are extremely reliable, being brushless and generally maintenance-free.

Synchronous Motor

A synchronous motor is a form of induction motor that is designed to reach and maintain equilibrium when the rotor is turning in perfect synchronization with the AC power supply. The speed of the motor will depend on the number of poles (magnetic coils) in the stator, and the number of phases in the power supply. If R is the RPM of a synchronous motor, f is the frequency of the AC current in Hz, and p is the number of poles per phase:

```
R = (120 * f) / p
```

This formula assumes 60Hz AC current. In nations where 50Hz AC is used, the number 120 should be replaced with the number 100.

Two basic types of synchronous motors exist: *direct current excited*, which require external power to start turning, and *non-excited*, which are self-starting. Since non-excited synchronous motors are more common in electronic applications, this encyclopedia will not deal with direct current excited variants.

A *hysteresis motor* is a synchronous motor containing a solid rotor cast from cobalt steel, which has high coercivity, meaning that once it is magnetized, a substantial field is required to reverse the magnetic polarity. Consequently the polarity of the rotor lags behind the constantly changing

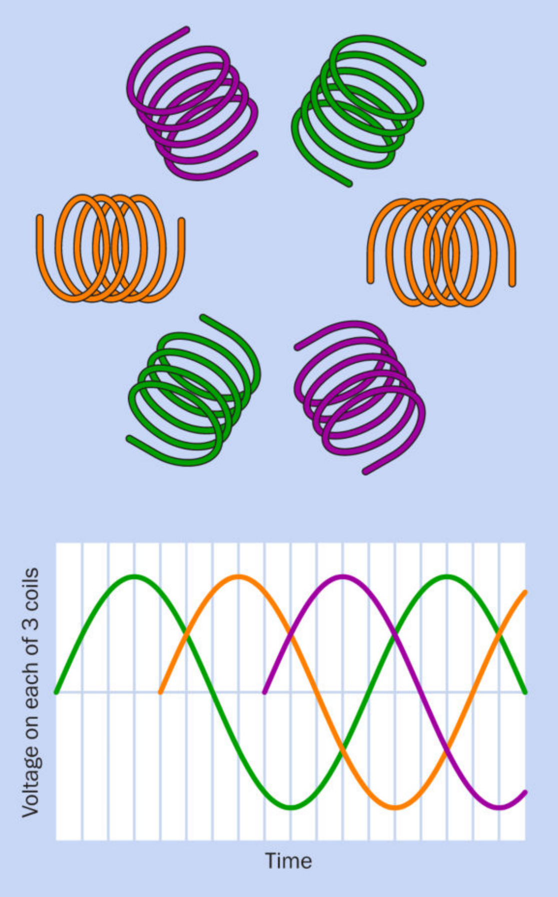

Figure 23-11. *The graph shows voltage delivered via three wires constituting a three-phase power supply. (The curve colors are arbitrary.) A three-phase motor contains a multiple of three coils—often six, as shown here diagrammatically. The three wires of the power supply are connected directly to the coils, which induce a rotating magnetic field.*

polarity of the stator, creating an attracting force that turns the rotor. Because the lag angle is independent of motor speed, this motor delivers constant torque from startup.

Reluctance Motor

Reluctance is the magnetic equivalent to electrical resistance. If a piece of iron is free to move in a magnetic field, it will tend to align itself with the field to reduce the reluctance of the magnetic circuit. This principle was used in very early reluctance motors designed to work from AC and has been revived as electronics to control variable frequency drives have become cheaper.

The simplest reluctance motor consists of a soft iron rotor with projecting lugs, rotating within a stator that is magnetically energized with its own set of inwardly projecting poles. The rotor tends to turn until its lugs are aligned with the poles of the stator, thus minimizing the reluctance.

A basic reluctance motor design is shown in Figure 25-2. It is located in the **stepper motor** section of this encyclopedia, as stepper motors are a primary application of the reluctance principle.

Although a reluctance motor can be used with polyphase fixed-frequency AC power, a variable frequency drive greatly enhances its usefulness. The timing of the frequency is adjusted by the speed of the motor, which is detected by a sensor. Thus the energizing pulses can remain “one step ahead” of the rotor. Since the rotor is not a magnet, it generates no back-EMF, allowing it to reach very high speeds.

The simplicity of the motor itself is a compensating factor for the cost of the electronics, as it requires no commutator, brushes, permanent magnets, or rotor windings. Characteristics of reluctance motors include:

- Cheap parts, easily manufactured, and high reliability.
- Compact size and low weight.
- Efficiencies greater than 90% possible.
- Capable of high start-up torque and high speed operation.

Disadvantages include noise, cogging, and tight manufacturing tolerances, as the air gap between the rotor and stator must be minimized.

A reluctance motor can function synchronously, if it is designed for that purpose.

Variable Frequency Drive

A basic induction motor suffers from significant problems. The surge of power that it draws when starting from rest can pull down the supply voltage enough to affect other devices that share the AC power supply. (Hence, the brief dimming of lights that may occur when the compressor in an air conditioner or refrigerator starts running.) While the motor is turning, it can introduce electrical noise, which feeds back into the power supply, once again causing potential problems for other devices. In addition, the narrow range of speed of an AC induction motor is a great disadvantage.

The advent of cheap solid-state technology encouraged the development of *variable-frequency power supplies* for induction motors. Because the impedance of the motor will diminish as the frequency diminishes, the current drawn by the motor will tend to increase. To prevent this, a variable frequency supply also varies the voltage that it delivers.

Wound-Rotor AC Induction Motor

The stator of this variant is basically the same as that of a single-phase induction motor, but the rotor contains its own set of coils. These are electrically accessible via a *commutator* and *brushes*, as in a **DC motor**. Because the maximum torque (also known as *pull-out* torque) will be proportional to the electrical resistance of the coils in the rotor, the characteristics of the motor can be adjusted by adding or removing resistance externally, via the commutator. A higher resistance will enable greater torque at low speed when the slip between the rotor speed and rotation of the magnetic field induced by the stator is greatest. This is especially useful in corded power tools such as electric drills, where high torque at low speed is desirable, yet the motor can accelerate to full speed quickly when the external resistance is reduced. Typically, the resistance is adjusted via the trigger of the drill.

Figure 23-12 shows a wound-rotor AC induction motor. The disadvantage of this configuration is the brushes that supply power to the rotor will eventually require maintenance. Much larger wound-rotor motors are also used in industrial applications such as printing presses and elevators, where the need for variable speed makes a simple three-phase motor unsuitable.

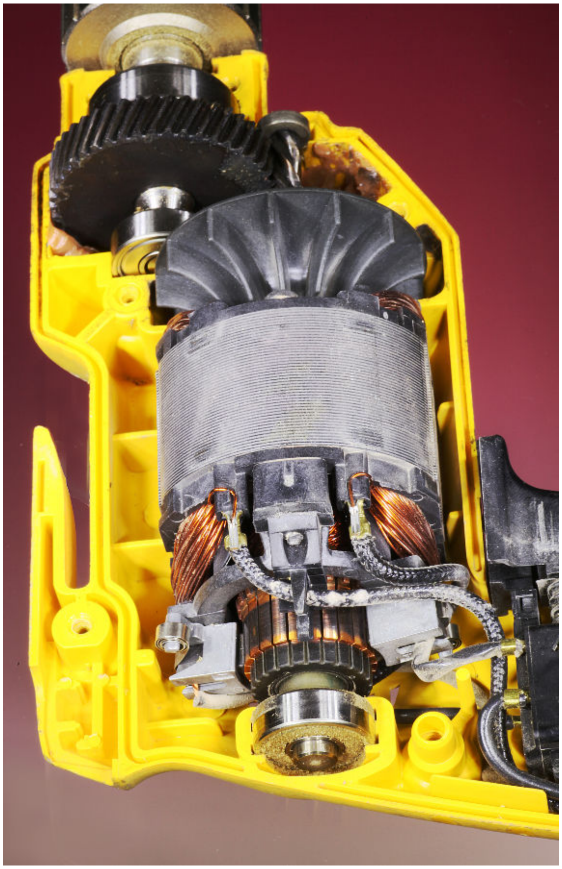

Figure 23-12. *A motor in a corded electric drill uses coils in a brushed rotor to enable variable speed output. In most AC motors, the speed is not adjustable and the rotor does not make any electrical connection with the rest of the motor.*

Universal Motor

A wound-rotor motor may also be described as a *universal motor* if its rotor and stator coils are connected in series. This configuration allows it to be powered by either AC or DC.

DC supplied to the rotor and the stator will cause mutual magnetic repulsion. When the rotor turns, the brushes touching the split commutator reverse the polarity of voltage in the rotor coils, and the process repeats. This configuration is very similar to that of a conventional DC motor, except that the stator in a universal motor uses electromagnets instead of the permanent magnets that are characteristic of a **DC motor**.

When powered by AC, the series connection between stator and rotor coils insures that each pulse to the stator will be duplicated in the rotor, causing mutual repulsion. The addition of a shorting coil in the stator provides the necessary asymmetry in the magnetic field to make the motor start turning.

Universal motors are not limited by AC frequency, and are capable of extremely high-speed operation. They have high starting torque, are compact, and are cheap to manufacture. Applications include food blenders, vacuum cleaners, and hair dryers. In a workshop, they are found in routers and miniature power tools such as the Dremel series.

Because a universal motor requires commutator and brushes, it is only suitable for intermittent use.

Inverted AC Motors

Some modern domestic appliances may seem to contain an AC motor, but in fact the AC current is *rectified* to DC and is then processed with pulse-width modulation to allow variable speed control. The motor is really a **DC motor**; see the entry on this type of motor for additional information.

Values

Because a basic AC induction motor is governed by the frequency of the power supply, the speed of a typical four-pole motor is limited to less than 1,800 RPM (1,500 RPM in nations where 50Hz AC is the norm).

Variable-frequency, universal, and wound-rotor motors overcome this limitation, and can reach speeds of 10,000 to 30,000 RPM. Synchronous motors typically run at 1,800 or 1,200 RPM, depending on the number of poles in the motor. (They run at 1,500 or 1,000 RPM in locations where the frequency of AC is 50Hz rather than 60Hz).

For a discussion of the torque that can be created by a motor, see “Values” (page 184) in the **DC motor** entry in this encyclopedia.

How to Use It

Old-fashioned record players (where a turntable supports a vinyl disc that must rotate at a fixed speed) and electric clocks (of the analogue type) were major applications for synchronous motors, which used the frequency of the AC power supply to control motor speed. These applications have been superceded by CD players (usually powered by brushless DC motors) and digital clocks (which use crystal oscillators).

Many home appliances continue to use AC-powered induction motors. Small cooling fans for use in electronic equipment are sometimes AC-powered, reducing the current that must be provided by the DC power supply. An induction motor generally tends to be heavier and less efficient than other types, and its speed limit imposed by the frequency of the AC power supply is a significant disadvantage.

A simple induction motor cannot provide the sophisticated control that is necessary in modern devices such as CD or DVD players, ink-jet printers, and scanners. A **stepper motor**, **servo motor**, and **DC motors** controlled with *pulse-width modulation* are preferable in these applications.

A *reluctance motor* may find applications in high-speed, high-end equipment including vacuum cleaners, fans, and pumps. Large variable reluctance motors, with high amperage ratings, may be used to power vehicles. Smaller variants are being used for power steering systems and windshield wipers in some automobiles.

What Can Go Wrong

Compared with other devices that have moving parts, the brushless induction motor is one of the most reliable and efficient devices ever invented. However, there are many ways it can be damaged. General problems affecting all types of motors are listed at "Heat effects" (page 188). Issues relating specifically to AC motors are listed below.

Premature Restart

Large industrial three-phase induction motors can be damaged if power is reapplied before the motor has stopped rotating.

Frequent Restart

If a motor is stopped and started repeatedly, the heat that is generated during the initial surge of current is likely to be cumulative.

Undervoltage or Voltage Imbalance

A voltage drop can cause the motor to draw more current than it is rated to handle. If this situation persists, overheating will result. Problems also are caused in a three-phase motor where one phase is not voltage-balanced with the others. The most common cause of this problem is an open circuit-breaker, wiring fault, or blown fuse affecting just one of the three conductors. The motor will try to run using the two conductors that are still providing power, but the result is likely to be destructive.

Stalled Motor

When power is applied to an induction motor, if the motor is prevented from turning, the conductors in the rotor will carry a large current that is entirely dissipated as heat. This current surge will either burn out the motor or blow a fuse or circuit breaker. Care should be taken, in equipment design, to minimize the risk that an induction motor may stall or jam.

Protective Relays

Sophisticated protective relays are available for industrial 3-phase motors, and can guard against all of the faults itemized above. Their details are outside the scope of this encyclopedia.

Excess Torque

As has been previously noted, the torque of an induction motor increases with the slip (speed difference) between the rotation of the magnetic field and the rotation of the rotor. Consequently, if the motor is overloaded and forced to run more slowly, it can deliver more rotational force. This can destroy other parts attached to the motor, such as drive belts.

Internal Breakage

An overloaded induction motor may suffer some cracking or breakage of its rotor. This may be obvious because of reduced power output or vibration, but can also be detected if the motor's power consumption changes significantly.

servo motor

24

Should be referred to as an *RC servo* if it is intended for use in small devices that are remote-controlled and battery powered. However, in practice, the RC acronym is often omitted.

OTHER RELATED COMPONENTS

- **AC motor** (See Chapter 23)
- **DC motor** (See Chapter 22)
- **stepper motor** (See Chapter 25)

What It Does

A servo motor is actually a combination of a motor, reduction gearing, and miniaturized control electronics, usually packaged together inside a very compact sealed plastic case. The motor itself may be AC or DC, and if DC, it may be brushed or brushless. What distinguishes a servo from other types of motor is that it is not designed for continuous rotation. It is a position-seeking device. Its rotational range may be more than 180 degrees but will be significantly less than 360 degrees. Two typical RC servos are shown in Figure 24-1. A side view of a motor is shown in Figure 24-2.

The electronics inside the motor enclosure interpret commands from an external controller. The command code specifies the desired turn angle measured as an offset either side of the center position of the motor's range. The motor turns quickly to the specified position and stops there. So long as the command signal continues and power to the motor is sustained, the motor will hold its position and "push back" against any external turning force. In the absence of such a force, while the motor is stationary, it will use very little current.

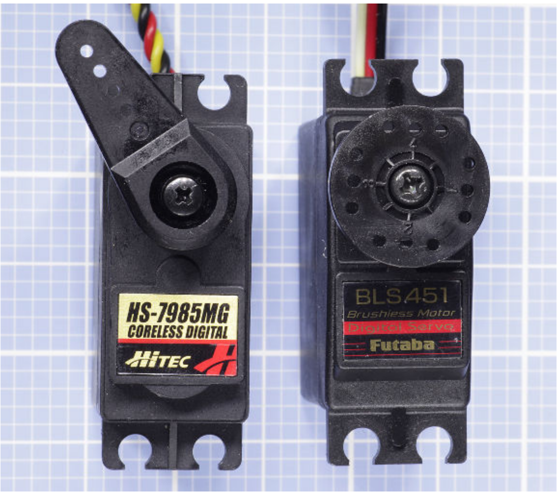

Figure 24-1. *A typical RC servo motor is capable of more than 50 inch-ounces of torque yet can be driven by three or four AA alkaline cells in series, and weighs under 2 ounces.*

The electronics inside a typical RC servo motor are shown in Figure 24-3.

How It Works

Servo motors are generally controlled via *pulse-width modulation* (*PWM*).

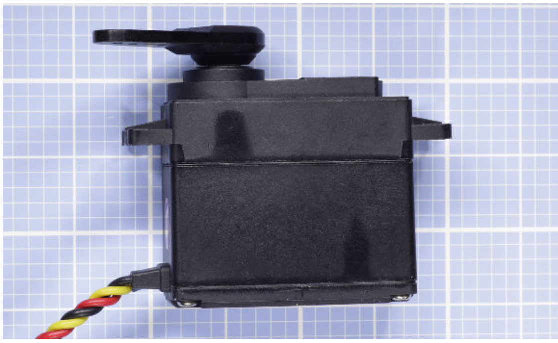

Figure 24-2. *RC servo motors are mostly similar in size. This is a typical side view.*

An industrial servo typically requires a controller that is an off-the-shelf item sold by the manufacturer of the motor. The encoding scheme of the control signals may be proprietary. A heavy-duty servo may be designed to run from three-phase power at a relatively high voltage, and may be used in applications such as production-line automation.

The remainder of this encyclopedia entry will focus primarily on small RC servos rather than industrial servos.

For small RC servos, the stream of control pulses is at a constant frequency of 20ms, with the positive durations of each pulse being interpreted as a positioning command to the motor, and the gaps between the pulses being disregarded. A typical range of pulse widths for a small motor is 1ms to 2ms, specifying a range of -90 to +90 degrees either side of a center location. Many modern motors are capable of excursions beyond these limits, and can be calibrated to establish the precise relationship between pulse width and turn angle. The motor can then be controlled by a lookup table in microcontroller software, or by using a conversion factor between degree-angle and pulse width.

Figure 24-4 shows the typical range of pulse widths within the fixed 20ms period (a frequency of 50Hz) between the start of one pulse and the start of the next, and the meaning of each pulse width to the servo motor. Intermediate pulse widths are interpreted as instructions to rotate to intermediate positions.

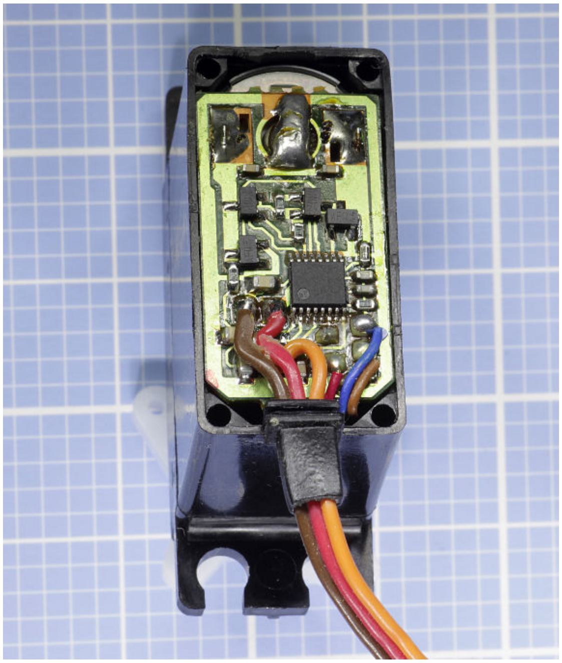

Figure 24-3. *The electronics inside a servo motor decode a stream of pulses that specify the turn angle of the motor.*

Small servo motors require the user to provide a controller that will conform with the above specification. This is often achieved by programming a *microcontroller*, and some microcontroller chips make this especially easy by providing a PWM output specifically tailored to the requirements of an RC servo. Either way, the microcontroller can be directly connected to the servo, enabling an extremely simple and flexible way to manage a positioning device.

Alternatively, a simple pulse generator such as a *555 timer chip* can be used, or controller boards are available from hobbyist supply sources. Some controller boards have USB connections enabling a servo to be governed by computer software.

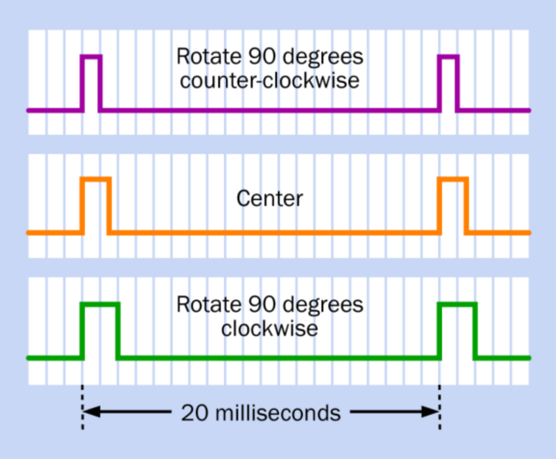

Figure 24-4. *The turn angle of a small RC servo motor is determined by a pulse width from a controller ranging from 1ms to 2ms in duration. The frequency of the pulses is constant at 50Hz.*

In Figure 24-5, a schematic illustrates the connection of a 555 timer with an RC servo, with component values to create a constant frequency of about 48Hz (slightly more than 20ms from peak to peak). The 1μF capacitor in the circuit charges through the 2.2K resistor in series with the diode, which bypasses the 28K resistor. This charging time represents the "on" cycle of the timer chip. The capacitor discharges through the 28K resistor, representing the "off" cycle. The 1K potentiometer, in series with the 5K resistors, acts as a voltage divider applied to the control pin of the timer, adjusting the timer's charge and discharge thresholds. Turning the potentiometer will lengthen or shorten the "on" time of each cycle, without changing the frequency. In practice, because capacitors are manufactured with wide tolerances, the frequency of the timer output cannot be guaranteed. Fortunately most servos will tolerate some inaccuracy.

Since the motor shares the power supply of the timer in this circuit, a protection diode and capacitor have been added between the power supply to the motor and negative ground, to suppress noise and back-EMF.

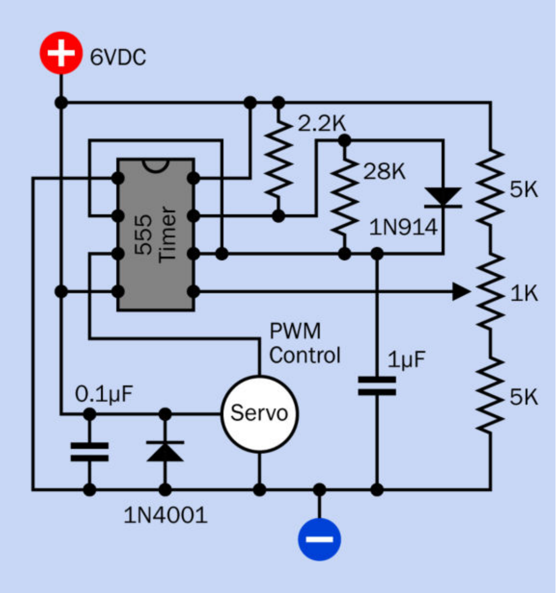

Figure 24-5. *An RC servo can be controlled via a 555 timer with appropriate component values. The potentiometer determines the angular position of the servo.*

Inside a servo motor's casing, the electronics include a **potentiometer** that turns with the output shaft, to provide feedback confirming the motor's position. The limited turning range of the potentiometer determines the turn limits of the motor output shaft.

Variants

Small servos may contain brushed or brushless DC motors. Naturally the brushless motors have greater longevity and create less electrical noise. See the **DC motor** entry in this encyclopedia for a discussion of brushed versus brushless motors.

Servos may use nylon, "Karbonite," or metal reduction gearing. The nylon gears inside a cheaper RC servo are shown in Figure 24-6.

Brushless motors and metal gears add slightly to the price of the motor. Metal gears are stronger than nylon (which can crack under load) but may wear faster, leading to *backlash* and inaccuracy in the gear train. The friction between nylon-to-nylon surfaces is very low, and nylon is certainly

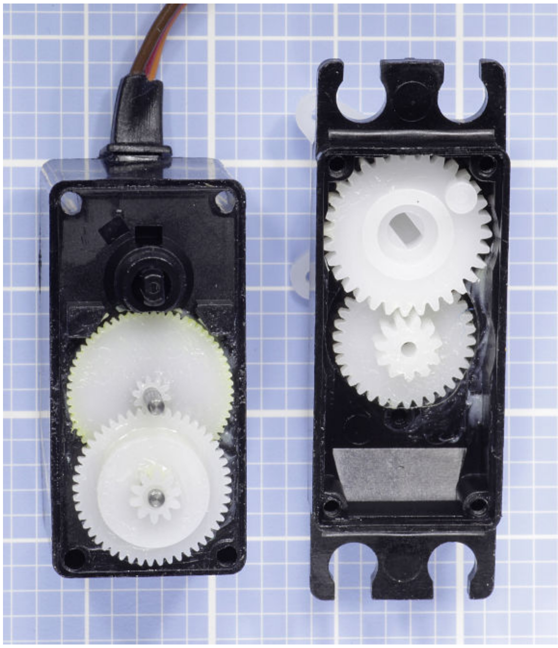

Figure 24-6. *Nylon gearing inside a servo motor.*

adequate and may be preferable if a servo will not be heavily loaded. "Karbonite" is claimed to be five times stronger than nylon and may be a satisfactory compromise. If a gear set experiences a failure (for example, teeth can be stripped as a result of excessive load), manufacturers usually will sell a replacement set to be installed by the user. Installation requires manual dexterity and patience, and some skill.

Servos may have roller bearings or plain sintered bearings, the latter being cheaper but much less durable under side loading.

So-called *digital servos* use faster internal electronics than the older, so-called *analog servos*, and because they sample the incoming pulse stream at a higher frequency, they are more responsive to small, rapid commands from the controller. For this reason they are preferred by hobbyists using servos to control the flight of model airplanes. Externally, the control protocol for digital and analog servos is the same, although a digital servo can be reprogrammed with new code values establishing the limits to its range. A standalone programming unit must be purchased to achieve this.

The most popular manufacturers of small servo motors are Futaba and Hitec. While their control protocols are virtually identical, the motor output shafts differ. The shaft is typically known as the *spline*, and is grooved to fit push-on attachments. The spline of a Futaba motor has 25 grooves, while Hitec uses 24 grooves. Attachments must be appropriate for the brand of motor that has been chosen.

Values

A small servo typically weighs 1 to 2 ounces, has a rotation time of 1 to 2 seconds from one end of its travel to the other, and can exert a surprisingly robust torque of 50 ounce-inches or greater.

Voltage

Small servos were originally designed to run from 4.8V rechargeable batteries in model aircraft. They can be driven with 5VDC to 6VDC on a routine basis. A few servos are designed for higher voltages.

Amperage

The datasheets provided by most manufacturers often fail to specify the power that a servo will draw when it is exerting maximum torque (or indeed, any torque at all). Since small servos are often driven by three or four AA alkaline batteries in series, the maximum current draw is unlikely to be much greater than 1 amp. When the motor is energized but not turning, and is not resisting a turning force, its power consumption is negligible. This feature makes servos especially desirable for remote-controlled battery-powered devices.

Some motors that have a turning range exceeding 180 degrees will respond to pulses of less than 1ms or greater than 2ms. A newly acquired motor should be tested with a microcontroller that steps through a wide range of pulse durations, to determine the limits empirically. Pulses that are outside the motor's designed range will generally be ignored and will not cause damage.

The *turn rate* or *transit time* specified in a datasheet is the time a servo takes to rotate through 60 degrees, with no load on the output shaft. A high-torque servo generally achieves its greater turning force by using a higher reduction gear ratio, which tends to result in a longer transit time.

How to Use it

Typical applications for a small servo include rotating the flaps or rudder of a model aircraft, steering a model boat, model car, or wheeled robot, and turning robotic arms.

A servo generally has three wires, colored red (power supply), black or brown (ground), and orange, yellow, or white (for the pulse train from the controller). The ground wire to the motor must be common with the ground of the controller, and consequently a ceramic bypass **capacitor** of 0.1µF or 0.01µF should be placed between the (red) power wire to the motor and ground. A protection **diode** should also be used. Neither a diode nor a capacitor should be attached to the wire carrying control signals, as it will interfere with the pulse train.

When powering the motor, an AC adapter should only be used with some caution, as its power output may be inadequately smoothed. A voltage regulator is not necessary, but bypass **capacitors** are mandatory. Figure 24-7 shows two hypothetical schematics. The upper section of the figure shows a battery-driven system, possibly using four 1.2V NiMH rechargeable batteries. Since batteries do not generally create voltage spikes, no capacitors are used, but a diode is included to protect the microcontroller from EMF when the servo stops and starts. The lower section of the figure shows the additional precautions that may be necessary when using DC power from an **AC adapter**. The DC-DC **converter**, which derives 6VDC for the motor requires smoothing capacitors (this should be specified in its datasheet), and so does the **voltage regulator**, which delivers regulated 5VDC power to the microcontroller. Once again, the protection diode is included. In both diagrams, the orange wire represents the control wire transmitting pulses to the servo motor.

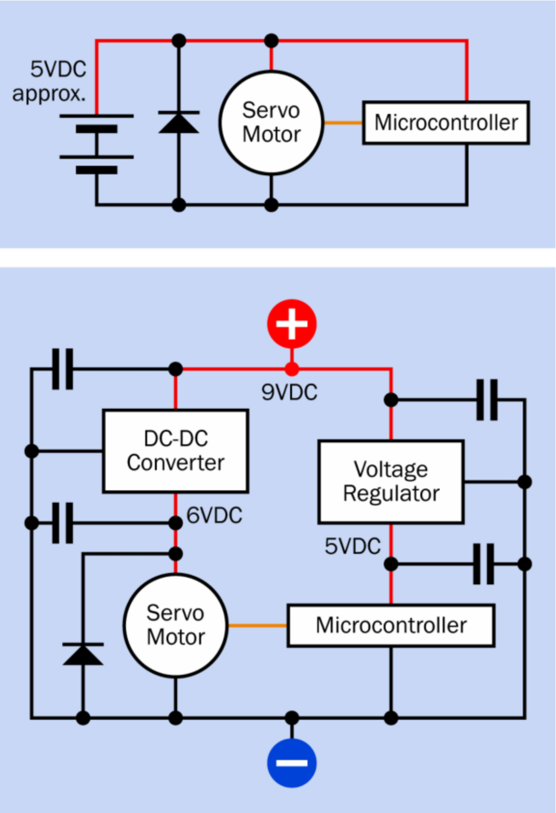

Figure 24-7. *Two possible schematics to run a small servo motor, the upper example using battery power (for example, from four 1.2V NiMH cells) and the lower example using a 9VDC AC adapter. See text for additional explanation.*

Various shaft attachments are available from the same online hobby-electronics suppliers that sell servos. The attachments include discs, single arms, double arms, and four arms in a cross-shaped configuration. A single-arm attachment is often known as a *horn*, and this term may be applied loosely to any kind of attachment. The horn is usually perforated so that other components can be fixed to it by using small screws or nuts and bolts. Figure 24-8 shows a variety of horns.

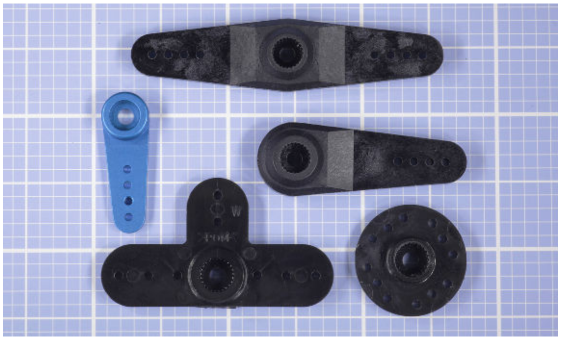

Figure 24-8. *Various shaft attachments, known as horns, are available from motor manufacturers. The blue one is metallic; the others are plastic.*

After the horn is pushed onto the spline (the motor shaft), it is held in place with one central screw. As previously noted, the two major manufacturers of small servos, Futaba and Hitec, have incompatible splines.

Modification for Continuous Rotation

It is possible to modify a small servo motor so that it will rotate continuously.

First the motor case must be opened, and the potentiometer must be centered by using a controller to send some 1.5ms pulses. The potentiometer must then be glued or otherwise secured with its wiper in this precise center position, after which the potentiometer is disconnected from the drive train.

Mechanical stops that would limit the rotation of the motor shaft must be cut away, after which the motor is reassembled. Because the potentiometer has been immobilized, the motor's internal electronics will now "see" the shaft as being in its center position at all times. If the controller sends a pulse instructing the motor to seek a position clockwise or counter-clockwise from center, the motor will rotate in an effort to reach that position. Because the potentiometer will not provide feedback to signal that the motor has achieved its goal, the shaft will continue to rotate indefinitely.

In this mode, the primary distinguishing characteristic of the servo has been disabled, in that it can no longer turn to a specific angle. Also, stopping the servo may be problematic, as it must receive a command that precisely matches the fixed position of the potentiometer. Since the potentiometer may have moved fractionally during the process in which it was immobilized, some trial and error may be needed to determine the pulse width that corresponds with the potentiometer position.

The purpose of modifying a servo for continuous rotation is to take advantage of its high torque, small size, light weight, and the ease of controlling it with a microcontroller.

In response to the interest shown by hobbyists in modifying servos for continuous rotation, some manufacturers now market servos with continuous rotation as a built-in feature. Typically they include a trimmer potentiometer to calibrate the motor, to establish its center-off position.

What Can Go Wrong

Incorrect Wiring

The manufacturer's datasheet should be checked to confirm the color coding of the wires. While a simple DC motor can be reversed by inverting the polarity of its power supply, this is totally inappropriate for a servo motor.

Shaft/Horn Mismatch

Attachments for the spline of one brand of motor may not fit the spline of another brand, and cannot be forced to fit.

Unrealistically Rapid Software Commands

Microcontroller software that positions a servo must allow sufficient time for the servo to respond before the software specifies a new position. It may be necessary to insert delay loops or other wait times in the software.

Jitter

A servo arm that twitches unpredictably usually indicates that the pulse train is being corrupted by external electrical noise. The control wire to the servo should be as short as possible, and should not run closely adjacent to conductors carrying AC or high frequency current switching, or control wires for other servo motors.

Motor Overload

A servo capable of delivering 2 lbs of force 1 inch from its shaft can easily generate enough torque, when it stalls, to break itself free from its mounts, or bend or break any arm or linkage attached to its shaft. Ideally, a relatively "weak link" should be included so that if breakage occurs, it will be predictable and will be relatively easy and cheap to repair.

Unrealistic Duty Cycle

Small servos are designed for intermittent use. Constant cycling will cause wear and tear, especially if the motor has a brushed commutator or metal reduction gears.

Electrical Noise

Brushed motors are always a source of electrical interference, and any servo will also tend to create a voltage dip or surge when it starts and stops. A protection diode may be insufficient to isolate sensitive microcontrollers and other integrated circuit chips. To minimize problems, the servo can be driven by a source of positive voltage that is separate from the regulated power supply used by the chips, and larger filter capacitors may be added to the voltage supply of the microcontroller. A common ground between the motor and the chips is unfortunately unavoidable.

stepper motor

25

Also often referred to as a *stepping motor*, and sometimes known as a *step motor*. It is a type of *induction motor* but merits its own entry in this encyclopedia as it has acquired significant and unique importance in electronics equipment where precise positioning of a moving part is needed and digital control is available.

OTHER RELATED COMPONENTS

- **DC motor** (See Chapter 22)
- **AC motor** (See Chapter 23)
- **servo motor** (See Chapter 24)

What It Does

A stepper motor rotates its drive shaft in precise steps in response to a timed sequence of pulses (usually one step per pulse). The pulses are delivered to a series of coils or *windings* in the *stator*, which is the stationary section of the motor, usually forming a ring around the *rotor*, which is the part of the motor that rotates. Steps may also be referred to as *phases*, and a motor that rotates in small steps may be referred to as having a high *phase count*.

A stepper motor theoretically draws power for its stator coils at a constant level that does not vary with speed. Consequently the torque tends to decrease as the speed increases, and conversely, it is greatest when the motor is stationary or locked.

The motor requires a suitable control system to provide the sequence of pulses. The control system may consist of a small dedicated circuit, or a microcontroller or computer with the addition of suitable driver transistors capable of handling the necessary current. The torque curve of a motor can be extended by using a controller that increases the voltage as the speed of the control pulses increases.

Because the behavior of the motor is controlled by external electronics, and its interior is usually symmetrical, a stepper motor can be driven backward and forward with equal torque, and can also be held in a stationary position, although the stator coils will continue to consume power in this mode.

How It Works

The stator has multiple poles made from soft iron or other magnetic material. Each pole is either energized by its own coil, or more commonly, several poles share a single, large coil. In all types of stepper motor, sets of stator poles are magnetized sequentially to turn the rotor and can remain energized in one configuration to hold the rotor stationary.

The rotor may contain one or more permanent magnets, which interact with the magnetic fields generated in the stator. Note that this is different from a *squirrel-cage* **AC motor** in which a "cage" is embedded in the rotor and interacts with a rotating magnetic field, but does not consist of permanent magnets.

Three small stepper motors are shown in Figure 25-1. Clockwise from the top-left, they are four-wire, five-wire, and six-wire types (this distinction is explained in the following section). The motor at top-left has a threaded shaft that can engage with a collar, so that as the motor shaft rotates counter-clockwise and clockwise, the collar will be moved down and up.

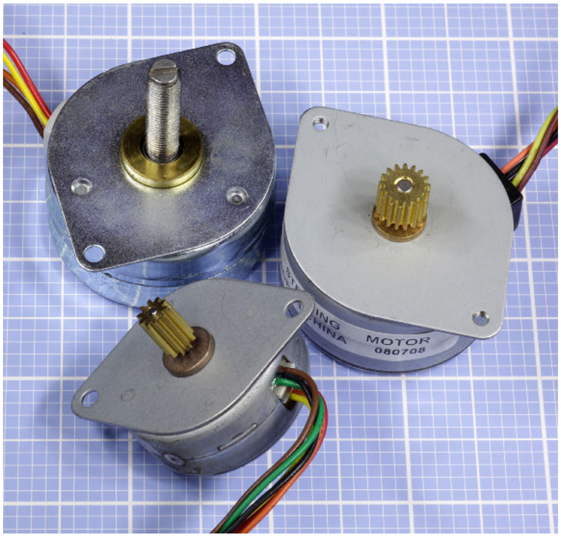

Figure 25-1. *Three small stepper motors.*

Reluctance Stepper Motors

The simplest form of stepper motor uses a rotor that does not contain permanent magnets. It relies on the principle of *variable reluctance*, reluctance being the magnetic equivalent of electrical *resistance*. The rotor will tend to align its protruding parts with the exterior source(s) of the magnetic field, as this will reduce the reluctance in the system. Additional information about variable reluctance is included in "Reluctance Motor" (page 197) in the section of this encyclopedia dealing with the **AC motor**.

A variable reluctance motor requires an external controller that simply energizes the stator coils sequentially. This is shown in Figure 25-2, where six poles (energized in pairs) are arrayed symmetrically around a rotor with four protrusions, usually referred to as *teeth*. Six stator poles and four teeth are the minimum numbers for reliable performance of a reluctance stepper motor.

In the diagram, the core of each pole is tinted green when it is magnetized, and is gray when it is not magnetized. In each section of this diagram, the stator coils are shown when they have just been energized, and the rotor has not yet had time to respond. External switching to energize the coils has been omitted for simplicity. In a real motor, the rotor would have numerous ridges, and the clearance between them and the stator would be extremely narrow to maximize the magnetic effect.

In a 6-pole reluctance motor where the rotor has four teeth, each time the controller energizes a new pair of poles, the rotor turns by 30 degrees counter-clockwise. This is known as the *step angle*, and means that the motor makes 12 steps in each full 360-degree rotation of its shaft. This configuration is very similar to that of a 3-phase AC *induction motor*, as shown in Figure 23-11 in the **AC motor** section of this encyclopedia. However, the AC motor is designed to be plugged into a power source with a constant frequency, and is intended to run smoothly and continuously, not in discrete steps.

Generally, reluctance motors tend to be larger than those with magnetized rotors, and often require feedback from a sensor that monitors shaft angle and provides this information to control electronics. This is known as a *closed loop* system. Most smaller stepper motors operate in an *open loop* system, where positional feedback is considered unnecessary if the number of pulses to the motor is counted as a means of tracking its position.

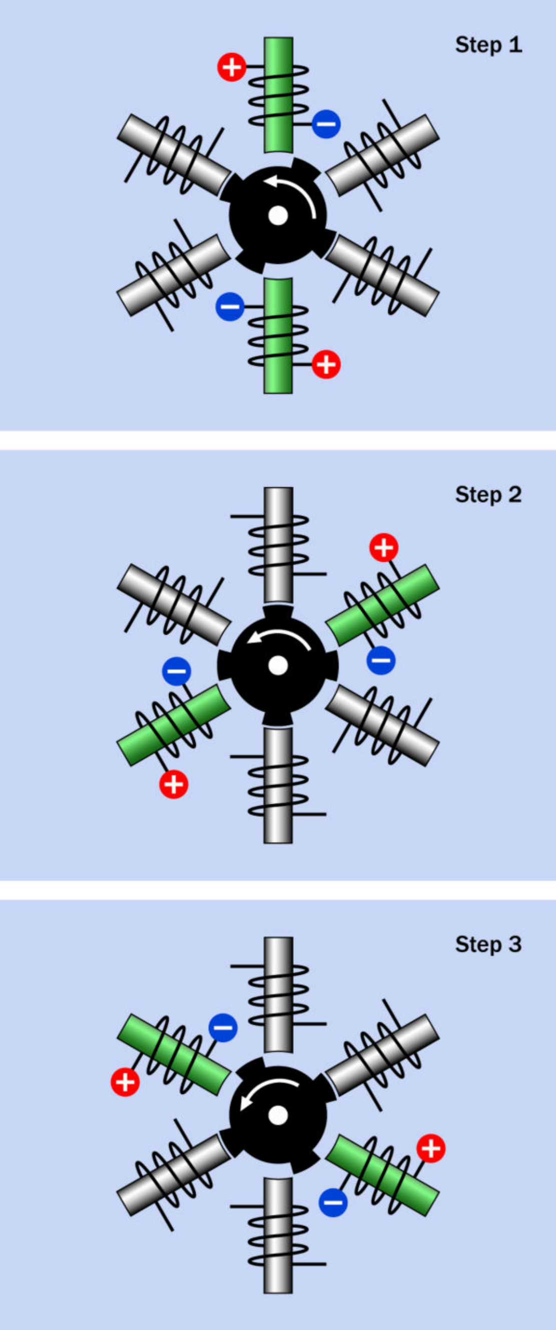

Figure 25-2. *In a variable reluctance stepper motor, the rotor moves to minimize magnetic reluctance each time the next pair of coils is energized. At each step, the coils have been energized a moment before the rotor has had time to respond.*

Permanent Magnet Stepper Motors

More commonly, the rotor of a stepper motor contains permanent magnets, which require the controller to be capable of reversing the magnetic field created by each of the stator coils, so that they alternately attract and repel the rotor magnets.

In a *bipolar* motor, the magnetic field generated by a coil is reversed simply by reversing the current through it. This is shown diagrammatically in Figure 25-3. In a *unipolar* motor, the magnetic field is reversed by applying positive voltage to the center tap of a coil, and grounding one end or the other. This is shown diagrammatically in Figure 25-4.

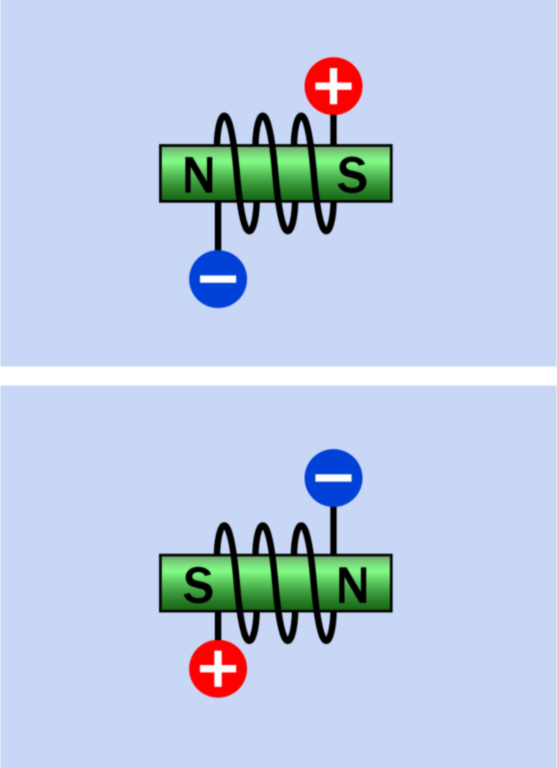

Figure 25-3. *In a bipolar motor, the magnetic field generated by each stator coil is reversed simply by reversing the current through the coil.*

Either type of motor is often designed with an upper and lower deck surrounding a single rotor,

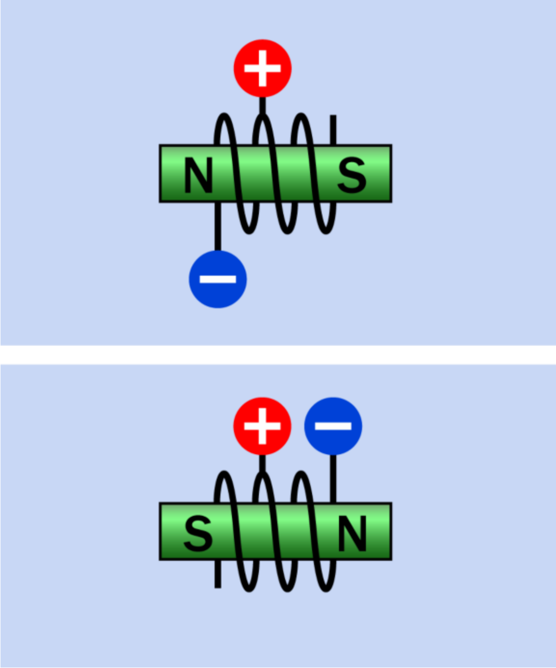

Figure 25-4. *The magnetic field of this coil is reversed by applying positive voltage constantly to a center tap and grounding one end of the coil or the other.*

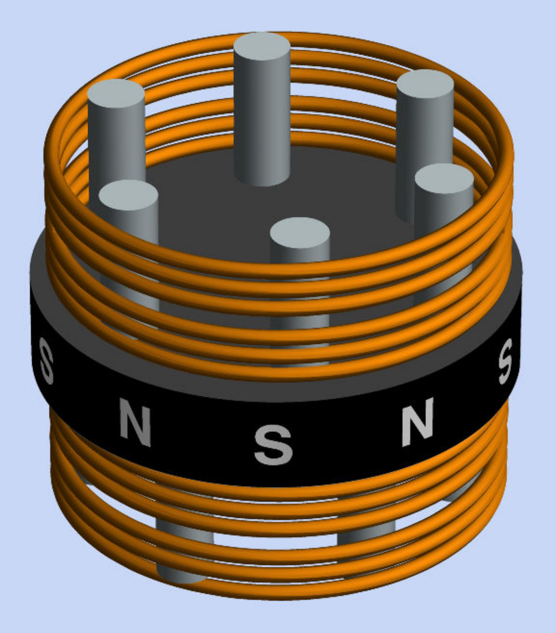

Figure 25-5. *A simplified rendering of the common "two deck" type of motor. See text for details.*

as suggested in Figure 25-5. A large single coil, or center-tapped coil, induces a magnetic field in multiple poles in the top deck, out of phase by one step with a second set of poles, energized by their own coil, in the bottom deck. (All three motors shown in Figure 25-1 are of this type.) The rotor of the motor is tall enough to span both decks, and is rotated by each of them in turn.

In Figure 25-6, the decks of a two-deck four-wire motor have been split apart. The rotor remains in the left-hand section. It is enclosed within a black cylinder that is a permanent magnet divided into multiple poles. In the right-hand section, a coil is visible surrounding metal "teeth" that function as stator poles when the coil is energized.

In Figure 25-7, the same motor has been further disassembled. The coil was secured with a length of tape around its periphery, which has been removed to make the coil visible. The remaining half of the motor, at top-right, contains a second, concealed but identical coil with its own set of poles, one step out of phase with those in the first deck.

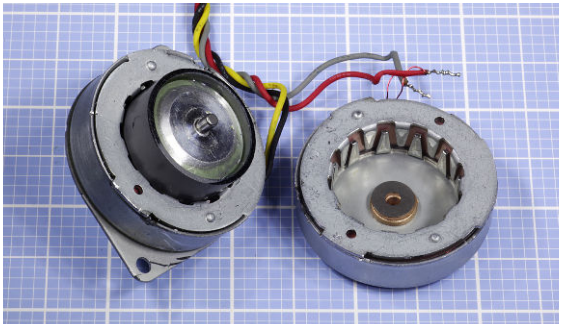

Figure 25-6. *A two-deck stepper motor split open to reveal its rotor (left) and one of the stators (right) encircled by a coil.*

Because the field effects in a two-deck stepper motor are difficult to visualize, the remaining diagrams show simplified configurations with a minimum number of stator poles, each with its own coil.

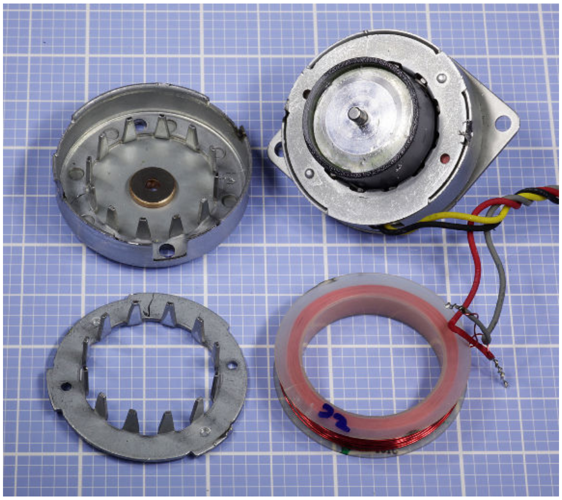

Figure 25-7. *The stepper motor from the previous figure, further disassembled.*

Bipolar Stepper Motors

The most basic way to reverse the current in a coil is by using an *H-bridge* configuration of switches, as shown in Figure 25-8, where the green arrow indicates the direction of the magnetic field. In actual applications, the switches are solid-state. Integrated circuits are available containing all the necessary components to control a bipolar stepper motor.

Four sequential steps of a bipolar motor are shown in Figure 25-9, Figure 25-10, Figure 25-11, and Figure 25-12. The H-bridge control electronics for each coil are omitted for clarity. As before, energized coils are shown with the pole inside the coil tinted green, while non-energized coils are gray, and the rotor is shown before it has had time to respond to the magnetic field in each step.

Unipolar Motors

The control electronics for a unipolar motor can be simpler than those for a bipolar motor, as off-the-shelf switching transistors can ground one end of the coil or the other. The classic five-wire unipolar stepper motor, often sold to hobbyists and used in robotics projects and similar applications, can be driven by nothing more elaborate

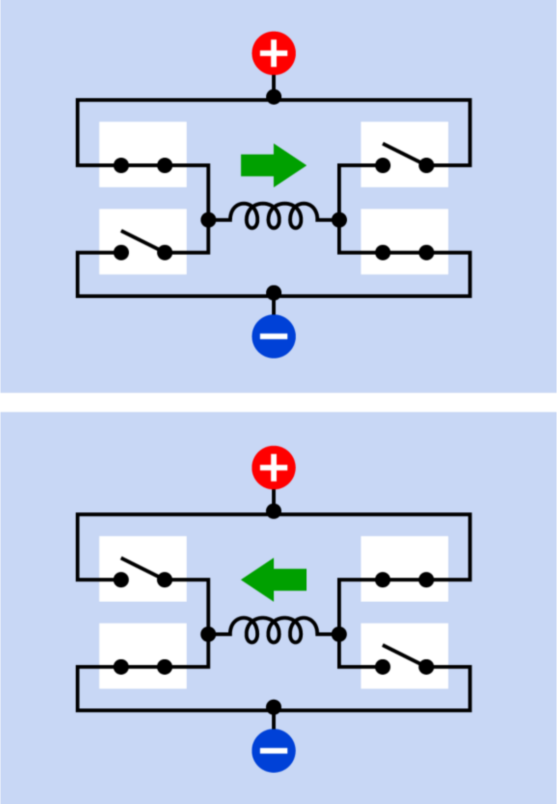

Figure 25-8. *The simplest and most basic way to reverse the current through a coil is via an H-bridge circuit. In practice, the switches are replaced by solid-state components.*

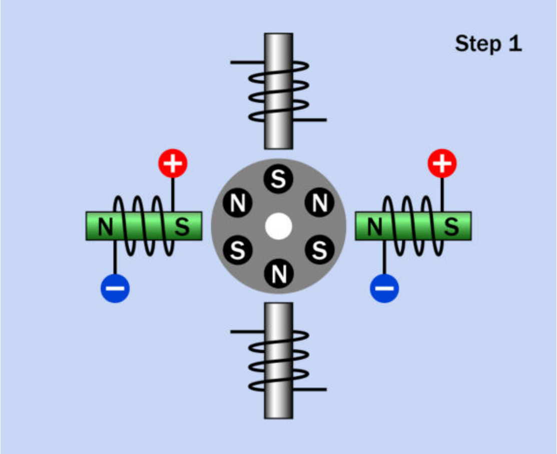

Figure 25-9. *A bipolar stepper motor depicted a moment before the rotor has had time to make its first step in response to magnetic fields created by the stator coils.*

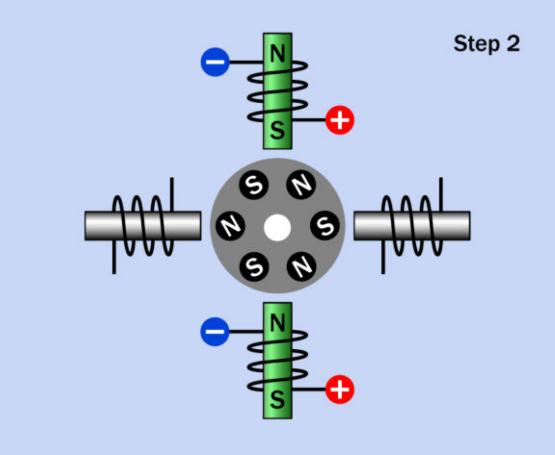

Figure 25-10. *The bipolar stepper motor from the previous figure is shown with its rotor having advanced by one step, and coil polarity changed to induce it to make a second step.*

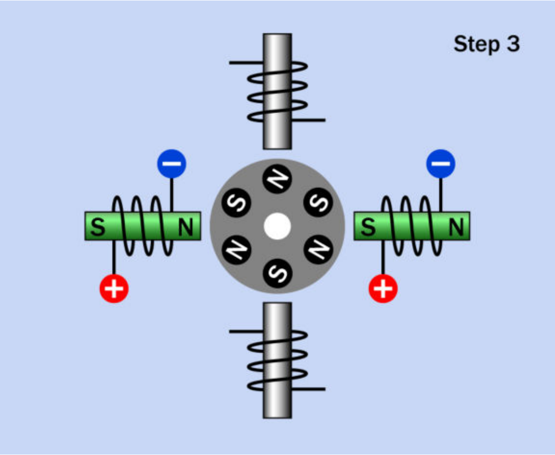

Figure 25-11. *The bipolar stepper motor after taking its second step, immediately before making its third step.*

than a set of *555 timer* chips. However, this type of motor is less powerful for its size and weight because only half of each coil is energized at a time.

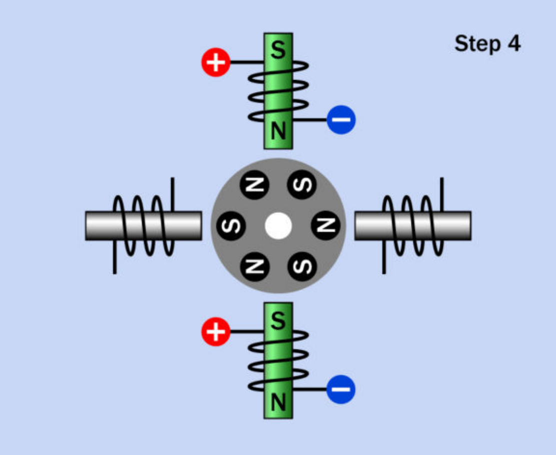

Figure 25-12. *The bipolar stepper motor after taking its third step. When the rotor responds to the new pattern of magnetic fields, its orientation will be functionally identical with that shown in the first step.*

In Figure 25-13, Figure 25-14, Figure 25-15, and Figure 25-16, the simplest configuration of a unipolar system is shown in diagrammatic form using four stator coils and a rotor containing six magnetic poles. Each figure shows the stator coils when they have just been energized, a moment before the rotor has had time to move in response to them. Coils that are energized are shown with the metal cores tinted green. Wires that are not conducting current are shown in gray. The open and closed positions of switches a, b, c, and d suggest the path that current is taking along the wires that are colored black.

Note that coils on opposite sides of the motor are energized simultaneously, while the other pair of coils is de-energized. Adjusting the controller so that it overlaps the “on” cycles of the coils can generate more torque, while consuming more power.

A motor containing more stator poles can advance in smaller steps, if the poles are separately energized. However, if the coils have individual windings, this will increase the cost of the motor.

Variants

In addition to bipolar and unipolar variants, previously described, three others are available.

High Phase Count

This term describes any type of stepper motor in which additional poles reduce the step size. The advantages of a high phase count include

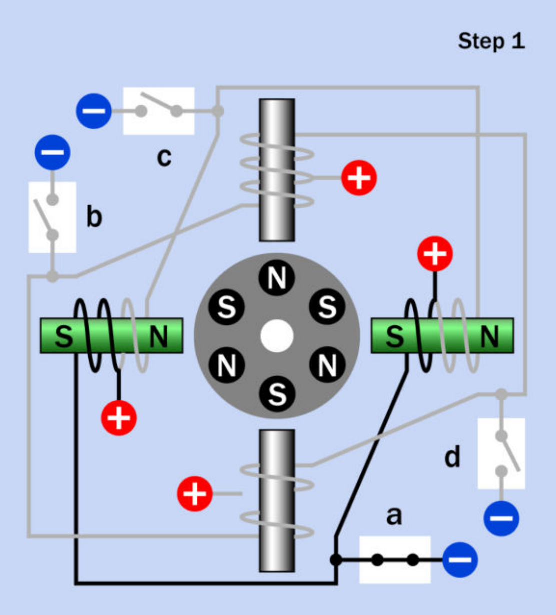

Figure 25-13. *The coils of this unipolar stepper motor are shown an instant after they have been energized, before the rotor has had time to respond by making its first step.*

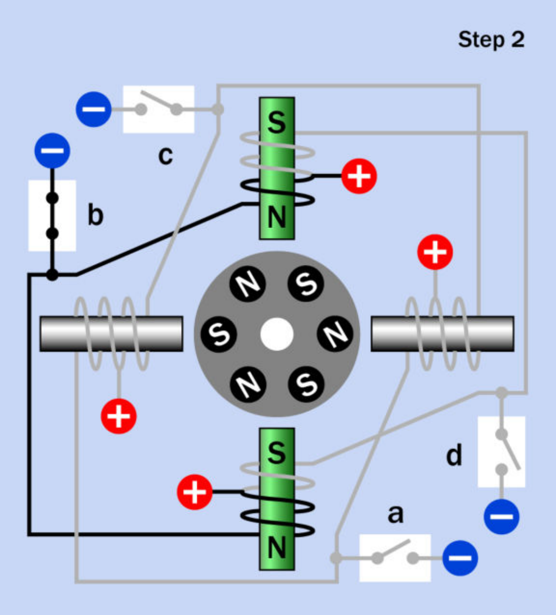

Figure 25-14. *The same motor from the previous figure is shown with coils energized to induce the rotor to make its second step.*

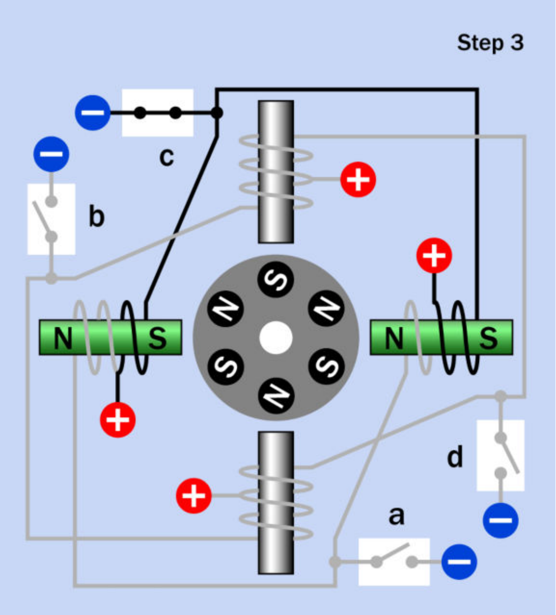

Figure 25-15. *The same motor from the previous figure is shown with coils energized to induce the rotor to make its third step.*

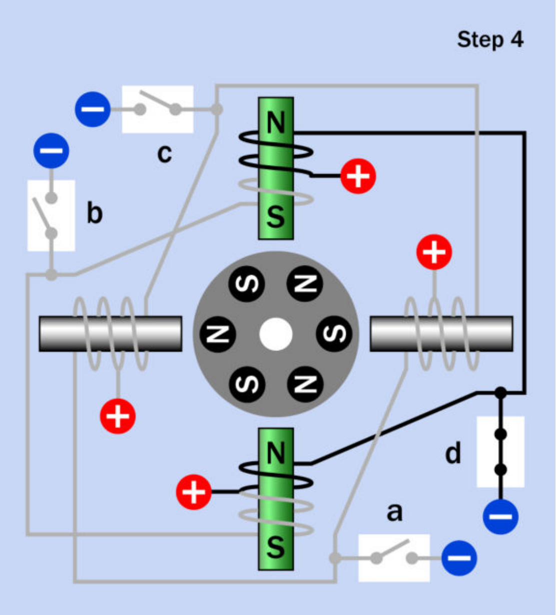

Figure 25-16. *When the rotor makes its fourth step, it will be back in an orientation that is functionally identical with the first figure in this series.*

smoother running at high speed and greater precision when selecting a desired motor position. The additional coils also enable higher power density, but naturally tend to add to the cost of the motor.

Hybrid

This type of motor uses a toothed rotor that provides variable reluctance while also containing permanent magnets. It has become relatively common, as the addition of teeth to the rotor enables greater precision and efficiency. From a control point of view, the motor behaves like a regular permanent-magnet stepper motor.

Bifilar

In this type of motor, also sometimes known as a *universal* stepper motor, two coils are wound in parallel for each stator pole. If there are two poles or sets of poles, and both ends of each winding are accessible via wires that are run out of the motor, there will be eight wires in total. Consequently this type is often referred to as an *8-wire motor*.

The advantage of this scheme is that it allows three possible configurations for the internal coils. By shorting together the wires selectively, the motor can be made to function either in unipolar or bipolar mode.

In Figure 25-17, the upper pair of simplified diagrams depicts one end of one coil connected to the beginning of the other, while positive voltage is applied at the midpoint, as in a unipolar motor. The magnetic polarity of the coil is determined by grounding either end of the coil. The section of each coil that is not conducting current is shown in gray.

The center pair of diagrams shows the adjacent ends of the coils tied together, so that they are now energized in parallel, with the magnetic polarity being determined by the polarity of the voltage, as in a bipolar motor.

The coils may also be connected in series, as shown in the lower pair of diagrams. This will provide greater torque at low speed and lower torque at high speed, while enabling higher-voltage, lower-current operation.

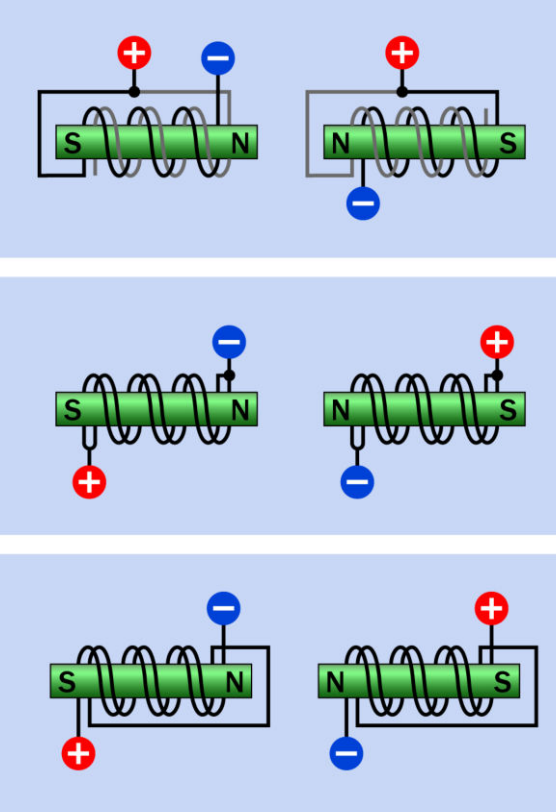

Figure 25-17. *In a bifilar motor, two coils are wound in parallel around each stator pole and can be connected with a center tap to emulate a unipolar motor (upper diagrams), or can be energized in parallel (middle diagrams) or series (lower diagrams) to emulate a bipolar motor.*

Multiphase

In a multiphase motor, multiple stator coils are usually connected in series, with a center tap applied between each pair. A possible configuration is shown in Figure 25-18, where the two diagrams show two consecutive steps in rotation, although the step angle could be halved by changing the voltage polarity in only one location at a time. The way in which the motor is wired enables only one stator coil to be unpowered during any step, because its two ends are at equal potential. Therefore this type of motor is capable of high torque in a relatively small format.

In some multiphase motors, additional wires allow access to both ends of each coil, and the coils are not connected internally. This allows control of the motor to be customized.

Microstepping

An appropriately designed stepper motor can be induced to make very small, intermediate steps if the control voltage is modulated to intermediate levels. Step angles as low as 0.007 degrees are claimed by some manufacturers. However, a motor running in this mode is less able to generate torque.

The simplest form of microstepping is half-stepping. To achieve this in a unipolar motor, each coil passes through an "off" state before its magnetic polarity is reversed.

Sensing and Feedback

So long as the series of pulses to the motor allows the rotor ample time to respond, no feedback mechanism from the rotor is necessary to confirm its position, and an open-loop system is sufficient. If sudden acceleration, deceleration, load fluctuations, and/or rotation reversal will occur, or if high speeds are involved, a *closed loop system*, in which a sensor provides positional feedback, may be necessary.

Voltage Control

Rapid stepping of a motor requires rapid creation and collapse of magnetic fields in the stator windings. Therefore, self-inductance of the windings can limit the motor speed. One way to overcome this is to use a higher voltage. A more sophisticated solution is to use a controller that provides a high initial voltage, which is reduced or briefly interrupted when a sensor indicates that coil current has increased sufficiently to overcome the self-inductance of the windings and has reached its imposed limit. This type of controller may be referred to as a *chopper drive* as the voltage is "chopped," usually by power transistors. It is a form of *pulse width modulation*.

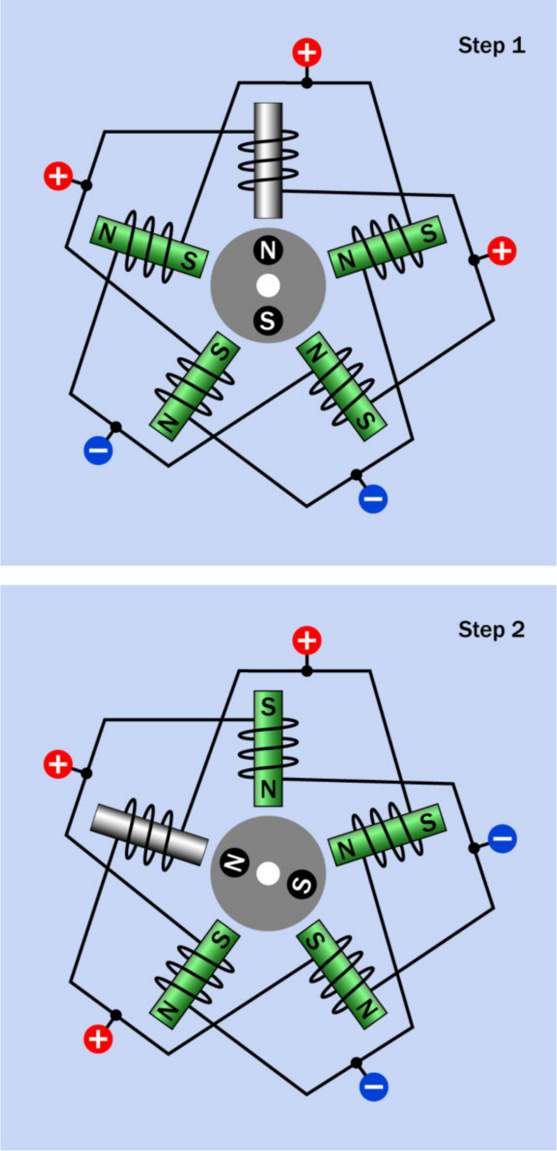

Figure 25-18. *A multiphase stepper motor. By applying voltage in the pattern shown, only one coil is not energized during each step. This enables high torque compared with the size of the motor.*

Values

The *step angle* of a stepper motor is the angular rotation of its shaft, in degrees, for each full step. This will be determined by the physical construction of the motor. The coarsest step angle is 90 degrees, while sophisticated motors may be capable of 1.8 degrees (without microstepping).

The maximum torque that a motor can deliver is discussed in "Values" (page 184) in the **DC motor** entry of this encyclopedia.

Motor weight and size, shaft length, and shaft diameter are the principal passive values of a stepper motor, which should be checked before it is selected for use.

How to Use it

Stepper motors are used to control the seek action in disk drives, the print-head movement and paper advance in computer printers, and the scanning motion in document scanners and copiers.

Industrial and laboratory applications include the adjustment of optical devices (modern telescopes are often oriented with stepper motors), and valve control in fluid systems.

A stepper motor may be used to power a *linear actuator*, usually via a screw thread (properly known as a lead screw) or worm gear. For more on linear actuators, see "Linear Actuator" (page 184). While the stepper motor will enable greater accuracy than a traditional DC motor, the gearing inevitably will introduce some imprecision.

Advantages of stepper motors include:

- Precise positioning, typically within 3 percent to 5 percent per step. The percentage step error does not accumulate as the motor rotates
- Able to run at a wide range of speeds, including very slow speeds without reduction gearing
- Trouble-free start, stop, and reverse action
- Cheap controller hardware where open-loop applications are acceptable
- High reliability, since no brushes or commutator are involved

Disadvantages include:

- Noise and vibration
- Resonance at low speeds
- Progressive loss of torque at high speeds

Protection Diodes

While a small stepper motor may be driven directly from power transistors, *darlington pairs*, or even *555 timers*, larger motors will create *back-EMF* when the magnetic field of each stator coil is induced or forward EMF when the field is allowed to collapse, and bipolar motors will also induce voltage spikes when the current reverses. In a unipolar motor, while only one-half of the coil is actually energized via its center tap, the other half will have an induced voltage, as the coil acts like a *linear transformer*.

A simplified schematic illustrating diode placement for a bipolar motor is shown in Figure 25-19.

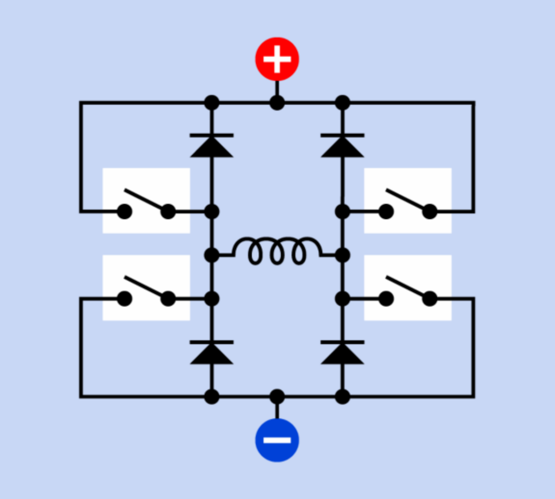

Figure 25-19. *The H-bridge circuit must be augmented with protection diodes to guard against back-EMF created by fluctuating current in the stator coil.*

Integrated circuit chips are available taht incorporate protection diodes, in addition to the necessary power transistors. Stepper motors may also have protection diodes built in. Consult the manufacturer's datasheet for details before attaching a motor to a power source.

Positional Control

The built-in control electronics of a **servo motor** typically turn the shaft to a precisely known position in response to pulse-width modulation from an exterior source such as a **microcontroller**, whereas the angle of rotation of a stepper motor in an open-loop system must be calculated by counting the number of steps from an initial, home position. This limitation of a stepper motor can be overcome by using a closed-loop system, but that will require monitoring the motor, adding complexity to the external controller. The choice between stepper and servo motors should be evaluated on a case-by-case basis.

What Can Go Wrong

General problems affecting all types of motors are listed in "Heat effects" (page 188). Issues relating more specifically to stepper motors are listed in the following sections.

Incorrect Wiring

Because a stepper motor is driven via multiple conductors, there is a significant risk of wiring errors, especially since many motors are not identified with part numbers. The first challenge, then, may be to determine what type of motor it is. When the motor is disconnected from any power, and the shaft is rotated with finger and thumb, a magnetized-rotor motor will not spin as freely as a reluctance motor, because the magnets in the rotor will provide intermittent turning resistance.

If a unipolar motor is relatively small and is fitted with five wires, almost certainly the motor contains two coils, each with a center tap, and their function can be determined by applying positive voltage to the red wire and grounding each of the other wires in turn. Attaching a small piece of tape to the motor shaft will assist in viewing its orientation.

A multimeter set to measure ohms can also be useful in deducing the internal coil connections of the motor, since the end-to-end resistance of a coil should be approximately twice the resistance between the center tap and either end of the coil.

A multiphase motor may have five wires, but in this case, the resistance between any two nonadjacent wires will be 1.5 times the resistance between any two adjacent wires.

Step Loss

In an open-loop system, if the motor skips or misses pulses from the controller, the controller no longer has an accurate assessment of the shaft angle. This is known as *step loss*. Since this can be caused by sudden changes in control frequency, the frequency should be increased (or decreased) gradually. This is known as *ramping* the motor speed. Stepper motors cannot respond instantly to changes in speed, because of inertia in the rotor or in the device that the motor is driving.

Where the motor turns one or more steps beyond its commanded stopping point, this is known as *overshoot*.

Step loss may also occur if the motor continues turning after power has been interrupted (either intentionally or because of an external fault). In an open-loop system, the controller should be designed to reset the motor position when power is initiated.

Excessive Torque

When the motor is stationary and not powered, *detent torque* is the maximum turning force that can be applied without causing the shaft to turn. When the motor is stationary and the controller does deliver power to it, *holding torque* is the maximum turning force that can be applied without causing the shaft to turn, and *pull-in torque* is the maximum torque which the motor can apply to overcome resistance and reach full speed. When the motor is running, *pull-out torque* is the maximum torque the motor can deliver without suffering step loss (pulling it out of sync with its controller). Some or all of these values should be specified on the motor's datasheet. Exceeding any of them will result in step loss.

Hysteresis

When a controller directs a stepper motor to seek a specified position, the term *hysteresis* is often used to mean the total error between the actual position it reaches when turning clockwise, and the actual position it reaches when turning counter-clockwise. This difference may occur because a stepper motor tends to stop a fraction short of its intended position, especially under significant load. Any design that requires precision should be tested under real-world conditions to assess the hysteresis of the motor.

Resonance

A motor has a natural resonant frequency. If it is stepped near that frequency, vibration will tend to be amplified, which can cause positional errors, gear wear (if gears are attached), bearing wear, noise, and other issues. A good datasheet should specify the resonant frequency of the motor, and the motor should run above that frequency if possible. The problem can be addressed by rubber motor mounts or by using a resilient component, such as a drive belt, in conjunction with the drive shaft. *Damping* the vibration may be attempted by adding weight to the motor mount.

Note that if the motor has any significant weight attached directly to its shaft, this will lower its resonant frequency, and should be taken into account.

Resonance may also cause *step loss* (see preceding sections).

Hunting

In a closed-loop system, a sensor on the motor reports its rotational position to the controller, and if necessary, the controller responds by adjusting the position of the motor. Like any feedback system, this entails some lag time, and at certain speeds the motor may start *hunting* or *oscillating* as the controller over-corrects and must then correct its correction. Some closed-loop controllers avoid this issue by running mostly in open-loop mode, using correction only when the motor experiences conditions (such as sudden speed changes), which are likely to cause step loss.

Saturation

While it may be tempting to increase the torque from a stepper motor by upping the voltage (which will increase the current through the stator coils), in practice motors are usually designed so that the cores of the coils will be close to saturation at the rated voltage. Therefore, increasing the voltage may achieve very little increase in power, while causing a significant increase in heat.

Rotor Demagnetization

The permanent magnets in a rotor can be partially demagnetized by excessive heat. Demagnetization can also occur if the magnets are exposed to high-frequency alternating current when the rotor is stationary. Therefore, attempting to run a stepper motor at high speed when the rotor is stalled can cause irrevocable loss of performance.

diode

26

The term **diode** almost always means a semiconductor device, properly known as a *PN junction diode*, although the full term is not often used. It was formerly known as a *crystal diode*. Before that, **diode** usually meant a type of *vacuum tube*, which is now rarely used outside of high-wattage RF transmitters and some high-end audio equipment.

OTHER RELATED COMPONENTS

- **rectifier** (See "Rectification" (page 227))
- **unijunction transistor** (See Chapter 27)
- **LED** (light-emitting diode) (Volume 2)

What It Does

A diode is a two-terminal device that allows current to flow in one direction, known as the *forward direction,* when the *anode* of the diode has a higher positive potential than the *cathode*. In this state, the diode is said to be *forward biased*. If the polarity of the voltage is reversed, the diode is now *reverse biased*, and it will attempt to block current flow, within its rated limits.

Diodes are often used as *rectifiers* to convert alternating current into direct current. They may also be used to suppress voltage spikes or protect components that would be vulnerable to reversed voltage, and they have specialized applications in high-frequency circuits.

A *Zener diode* can regulate voltage, a *varactor diode* can control a high-frequency oscillator, and *tunnel diodes*, *Gunn diodes*, and *PIN diodes* have high-frequency applications appropriate to their rapid switching capability. An **LED** (*light-emitting diode*) is a highly efficient light source, which is discussed in Volume 2 of this encyclopedia. A **photosensitive diode** will adjust its ability to pass current depending on the light that falls upon it, and is included as a *sensor* in Volume 3.

See Figure 26-1 for schematic symbols representing a generic diode.

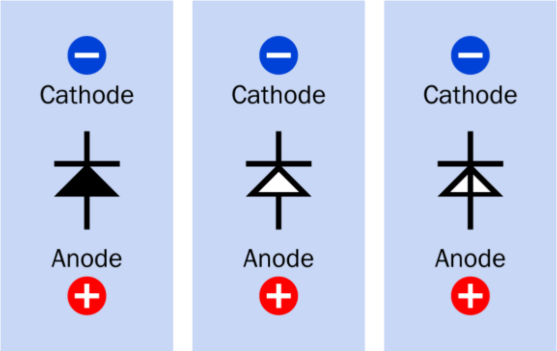

Figure 26-1. *Commonly used schematic symbols for a generic diode. All the symbols are functionally identical. The direction of the arrow formed by the triangle indicates the direction of conventional current (from positive to negative) when the diode is forward-biased.*

The basic diode symbol is modified in various ways to represent variants, as shown in Figure 26-2.

At top
: Each symbol in the group of six indicates a Zener diode. All are functionally identical.

Bottom-left
Tunnel diode.

Bottom-center
Schottky diode.

Bottom-right
Varactor.

A triangle with an open center does not indicate any different function from a triangle with a solid center. The direction of the arrow always indicates the direction of conventional current, from positive to negative, when the diode is forward-biased, although the functionality of Zener diodes and varactors depends on them being reverse-biased, and thus they are used with current flowing opposite to the arrow symbol. The bent line used in the Zener symbol can be thought of as an opened letter Z, while the curled line used in the Schottky diode symbol can be thought of as a letter S, although these lines are sometimes drawn flipped left-to-right.

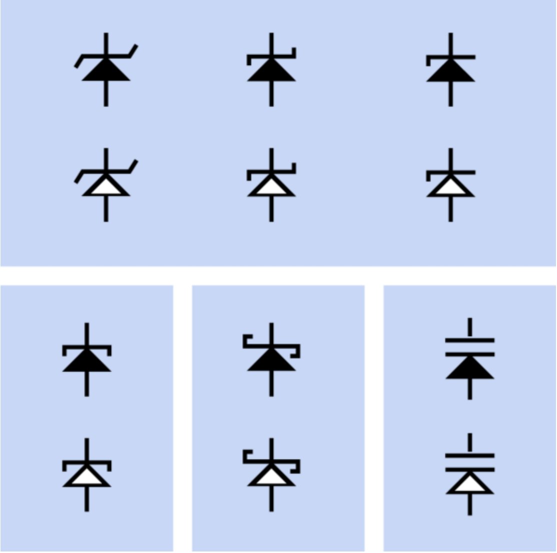

Figure 26-2. *Commonly used schematic symbols for specialized types of diodes. See text for details.*

A range of rectifier and signal diodes is shown in Figure 26-3. (Top: Rectifier diode rated 7.5A at 35VDC. Second from top: Rectifier diode rated 5A at 35VDC. Center: Rectifier diode rated 3A at 35VDC. Second from bottom: 1N4001 Rectifier diode rated 1A at 35VDC. Bottom: 1N4148 signal switching diode rated at 300mA.) All values are for forward continuous current and RMS voltage. Each cylindrical diode is marked with a silver stripe (a black stripe on the 1N4148) to identify its *cathode*, or the end of the diode that should be "more negative" when the component is forward biased. Peak current can greatly exceed continuous current without damaging the component. Datasheets will provide additional information.

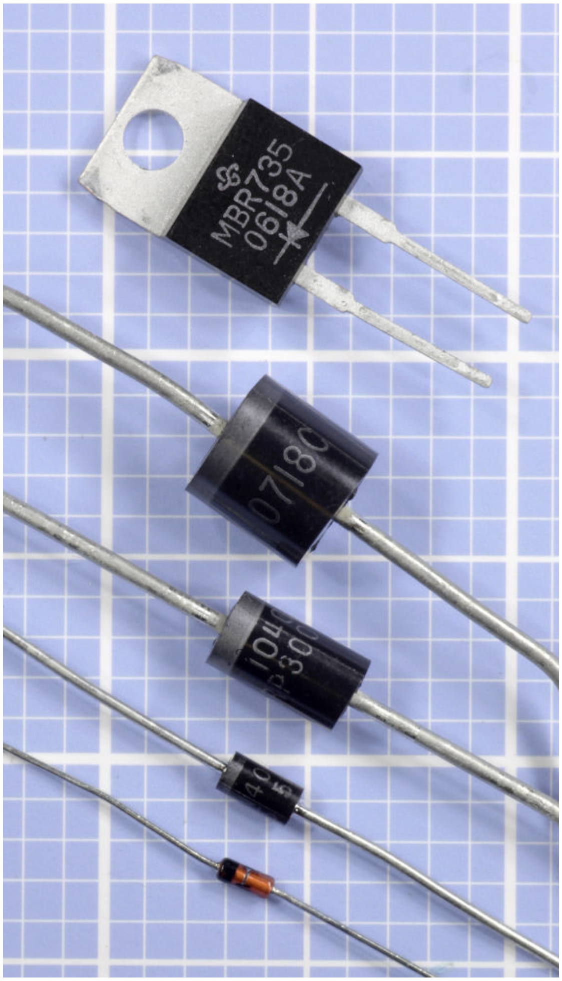

Figure 26-3. *Diodes ranging in continuous forward-current capability from 7.5A (top) to 300mA (bottom). See text for additional details.*

How It Works

A PN diode is a two-layer *semiconductor*, usually fabricated from silicon, sometimes from germanium, and rarely from other materials. The layers are doped with impurities to adjust their electrical characteristics (this concept is explained in more detail in Chapter 28). The *N layer* (on the negative, cathode side) has a surplus of electrons, creating a net negative charge. The *P layer* (on the positive, anode side) has a deficit of electrons, creating a net positive charge. The deficit of electrons can also be thought of as a surplus of "positive charges," or more accurately, a surplus of *electron holes*, which can be considered as spaces that electrons can fill.

When the negative side of an external voltage source is connected with the cathode of a diode, and the positive side is connected with the anode, the diode is forward-biased, and electrons and electron holes are forced by mutual repulsion toward the junction between the n and p layers (see Figure 26-4). In a silicon diode, if the potential difference is greater than approximately 0.6 volts, this is known as the *junction threshold voltage*, and the charges start to pass through the junction. The threshold is only about 0.2 volts in a germanium diode, while in a Schottky diode it is about 0.4 volts.

If the negative side of an external voltage source is connected with the anode of a diode and positive side is connected with the cathode, the diode is now reverse-biased, and electrons and electron holes are attracted away from the junction between the n and p layers. The junction is now a *depletion region*, which blocks current.

Like any electronic component, a diode is not 100% efficient. When it is forward-biased and is passing current, it imposes a small voltage drop of around 0.7V for a silicon-based diode (Schottky diodes can impose a drop of as little as 0.2V, germanium diodes 0.3V, and some LEDs between 1.4V and 4V). This energy is dissipated as heat. When the diode is reverse-biased, it is still not 100% efficient, this time in its task of blocking current. The very small amount of current that manages to get through is known as *leakage*. This is almost always less than 1mA and may be just a few µA, depending on the type of diode.

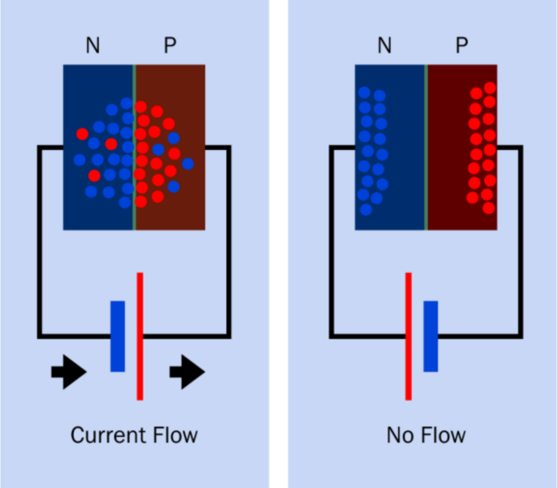

Figure 26-4. *Inside a PN junction diode. Left: in forward-biased mode, voltage from a battery (bottom, with plates colored for clarity) forces charges in the N and P layers toward the central junction of the diode. Current begins to flow. Right: in reverse-biased mode, charges in the N and P layers are attracted away from the central junction, which becomes a depletion region, unable to pass significant current.*

The performance of a theoretical generic PN diode is illustrated in Figure 26-5. The right-hand side of the graph shows that if a diode is forward-biased with a gradually increasing potential, no current passes until the diode reaches its junction threshold voltage, after which the current rises very steeply, as the *dynamic resistance* of the diode diminishes to near zero. The left-hand side of the graph shows that when the diode is reverse-biased with a gradually increasing potential, initially a very small amount of current passes as leakage (the graph exaggerates this for clarity). Eventually, if the potential is high enough, the diode reaches its intrinsic *breakdown voltage*, and once again its effective resistance diminishes to near zero. At either end of the curve, the diode will be easily and permanently damaged by excessive current. With the exception of Zener diodes and varactors, reverse bias on a diode should not be allowed to reach the breakdown voltage level.

The graph in Figure 26-5 does not have a consistent scale on its Y axis, and in many diodes the magnitude of the (reverse-biased) breakdown voltage will be as much as 100 times the magnitude of the (forward-biased) threshold voltage. The graph has been simplified for clarity.

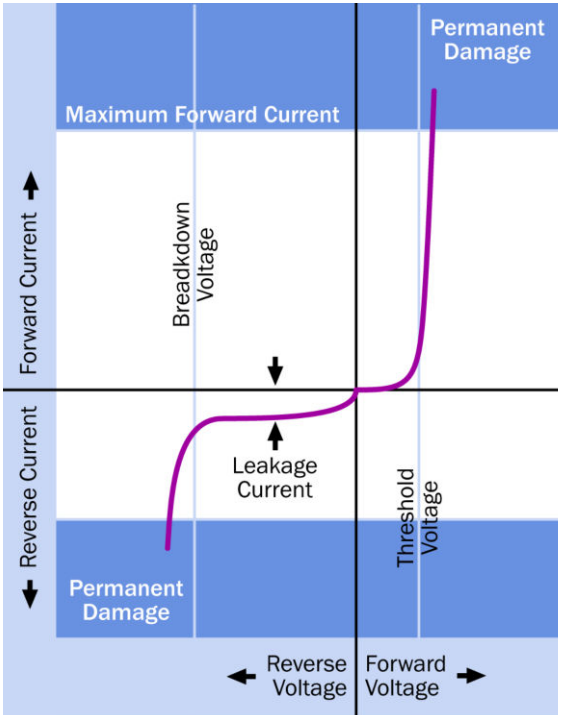

Figure 26-5. *As the forward voltage across a diode reaches the junction threshold, the diode begins passing current. If the voltage across the diode is reversed, initially a small amount of current leakeage occurs. Excessive forward or reverse voltage will create sufficient current to destroy the component.*

Variants

Packaging

Some diodes have no information at all printed on them, while others may have a part number. Any additional information is rare. No convention exists for indicating the electrical characteristics of the component by colors or abbreviations. If one terminal is marked in any way, almost certainly it is the cathode. One way to remember the meaning of a stripe on the cathode end of a rectifier diode or signal diode is by thinking of it as resembling the line in the diode schematic symbol.

Signal Diodes

Also known as *switching diodes* and *high-speed diodes*, their small size provides a low junction capacitance, enabling fast response times. They are not designed to withstand high currents. Signal diodes traditionally were packaged with axial leads for through-hole installation (like traditional-style resistors). Although this format still exists, signal diodes are now more commonly available in surface-mount formats.

Rectifier Diodes

Physically larger than signal diodes, and capable of handling higher currents. Their higher junction capacitance makes them unsuitable for fast switching. Rectifier diodes often have axial leads, although different package formats are used where higher currents are involved, and may include a heat sink, or may have provision for being attached to a heat sink.

There are no generally agreed maximum or minimum ratings to distinguish signal diodes from rectifier diodes.

Zener Diode

A Zener diode generally behaves very similarly to a signal or rectifier diode, except that its breakdown voltage is lower.

The Zener is intended to be reverse-biased; that is, conventional current is applied through it "in the wrong direction" compared with conventional diodes. As the current increases, the *dynamic resistance* of the Zener diode decreases. This relationship is shown in Figure 26-6, where the two colored curves represent the performance of different possible Zener diodes. (The curves are adapted from a manufacturer's datasheet.) This behavior allows the Zener to be used in simple voltage-regulator circuits, as it can allow a reverse current to flow at a voltage limited by the diode's breakdown voltage. Other applications for Zener diodes are described in "DC Voltage Regulation and Noise Suppression" (page 230). A typical Zener diode is shown in Figure 26-7.

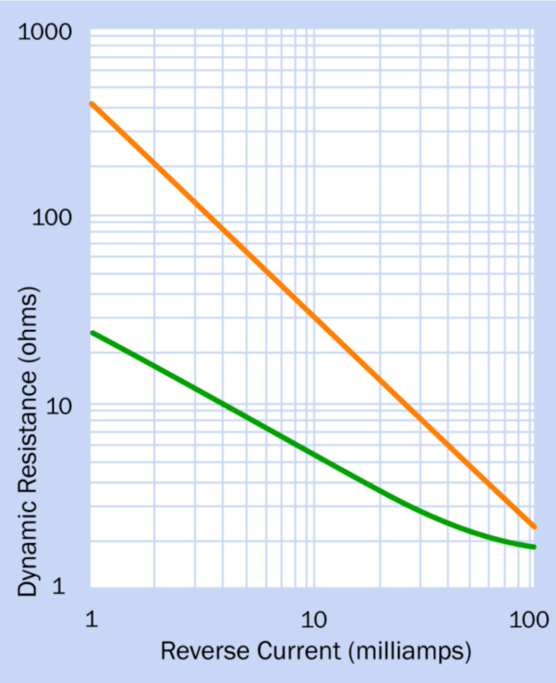

Figure 26-6. *A manufacturer's datasheet may include graphs of this kind, showing the variation in dynamic resistance of two reverse-biased Zener diodes in response to changes in current.*

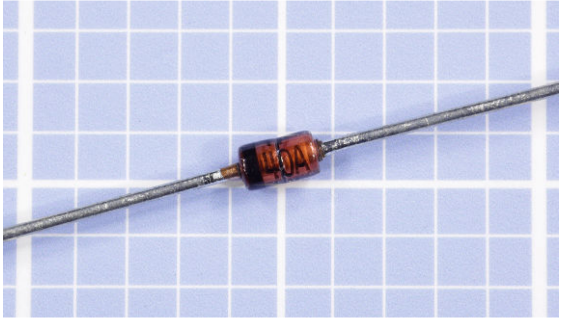

Figure 26-7. *A 1N4740 Zener diode.*

Transient Voltage Suppressor (TVS)

A form of Zener diode designed to protect sensitive devices from transient voltage spikes by clamping them—in other words, diverting the energy to ground. A TVS can absorb as much as 30,000 volts from a lightning strike or static discharge. Typically the Zener diode is incorporated in a network of other diodes in a surface-mount integrated circuit chip.

Zener diodes can also be used in circuits to handle electrostatic discharge (ESD), which can occur when a person unknowingly accumulates an electrostatic potential and then grounds it by touching an electronic device.

Schottky Diode

This type has a low junction capacitance, enabling faster switching than comparable generic silicon diodes. It also imposes a lower forward voltage drop, which can be desirable in low-voltage applications, and allows less power dissipation when a diode is necessary to control current flow. The Schottky diode is fabricated with a semiconductor-to-metal junction, and tends to be slightly more expensive than generic silicon diodes with similar voltage and current specifications.

Varactor Diode

Also known as a *varicap*, this type of diode has variable capacitance controlled by reverse voltage. While other diodes may exhibit this same phenomenon, the varactor is specifically designed to exploit it at very high frequencies. The voltage expands or contracts the depletion region in the junction between the P and N regions, which can be thought of as analogous to moving the plates of a capacitor nearer together or farther apart.

Because the capacitance of a varactor has a low maximum of about 100pF, its uses are limited. It is used extensively in RF applications where its voltage-controlled variable capacitance provides a uniquely useful way to control the frequency of an oscillator circuit. In almost all radio, cellular, and wireless receivers, a varactor controls a phase-locked loop oscillator. In ham radio receivers, it can be used to adjust the tuning of a filter that tracks an incoming radio frequency.

A varactor is always reverse-biased below its breakdown voltage, so that there is no direct conduction. The voltage that controls a varactor must be absolutely free from random fluctuations that would affect its resonant frequency.

Tunnel Diode, Gunn Diode, PIN Diode

Mostly used in very high frequency or microwave applications, where ordinary diodes are unacceptable because they have insufficiently high switching speeds.

Diode Array

Two or more diodes may be encapsulated in a single DIP or (more commonly) surface-mount integrated circuit chip. The internal configuration and the pinouts of the chip will vary from one device to another. Diode arrays may be used for termination of data lines to reduce reflection noise.

Bridge Rectifier

Although this is a diode array, it is commonly indexed in parts catalogues under the term *bridge rectifier*. Numerous through-hole versions are available with ratings as high as 25A, some designed for single-phase input while others process three-phase AC. Screw-terminal components can rectify more than 1,000 volts at 1,000 amps. The package does not usually include any provision for smoothing or filtering the output. See "Rectification" (page 227) for more information on the behavior of a bridge rectifier.

Values

A manufacturer's datasheet for a typical generic diode should define the following values, using abbreviations that may include those in the following list.

- Maximum sustained forward current: I_f or I_o or I_{Omax}
- Forward voltage (the voltage drop imposed by the diode): V_f
- Peak inverse DC voltage (may be referred to as maximum blocking voltage or breakdown voltage): P_{iv} or V_{dc} or V_{br}
- Maximum reverse current (also referred to as leakage): I_r

Datasheets may include additional values when the diode is used with alternating current, and will also include information on peak forward surge current and acceptable operating temperatures.

A typical signal diode is the 1N4148 (included at the bottom of Figure 26-3), which is limited to about 300mA forward current while imposing a voltage drop of about 1V. The component can tolerate a 75V peak inverse voltage. These values may vary slightly among different manufacturers.

Rectifier diodes in the 1N4001/1N4002/1N4003 series have a maximum forward current of 1A and will impose a voltage drop of slightly more than 1V. They can withstand 50V to 1,000V of inverse voltage, depending on the component. Here again, the values may vary slightly among different manufacturers.

Zener diodes have a different specification, as they are used with reverse bias as voltage-regulating devices rather than rectification devices. Manufacturers' data sheets are likely to contain the following terminology:

- Zener voltage (the potential at which the diode begins to allow reverse current flow when it is reverse-biased, similar to breakdown voltage): V_z
- Zener impedance or dynamic resistance (the effective resistance of the diode, specified when it is reverse-biased at the Zener voltage): Z_z

- Maximum or admissible Zener current (or reverse current): I_z or I_{zm}
- Maximum or total power dissipation: P_d or P_{tot}

Zener voltage may be defined within a minimum and maximum range, or as a simple maximum value.

Limits on forward current are often not specified, as the component is not intended to be forward-biased.

How to Use it

Rectification

A *rectifier diode*, as its name implies, is commonly used to rectify alternating current—that is, to turn AC into DC. A half-wave rectifier uses a single diode to block one-half of the AC sinewave. The basic circuit for a half-wave rectifier is shown in Figure 26-8. At top, the diode allows current to circulate counter-clockwise through the load. At bottom, the diode blocks current that attempts to circulate clockwise. Although the output has "gaps" between the pulses, it is usable for simple tasks such as lighting an LED, and with the addition of a smoothing capacitor, can power the coil of a DC relay.

A full-wave bridge rectifier employs four diodes to provide a more efficient output, usually filtered and smoothed with appropriate **capacitors**. The basic circuit is shown in Figure 26-9. A comparison of input and output waveforms for half-wave and full-wave rectifiers appears in Figure 26-10.

Discrete components are seldom used for this purpose, as off-the-shelf bridge rectifiers are available in a single integrated package. Rectifier diodes as discrete components are more likely to be used to suppress *back-EMF* pulses, as described below.

An old but widely used design for a full-wave bridge rectifier is shown in Figure 26-11. This unit measured approximately 2" × 2" × 1.5" and was divided into four sections (as indicated by the solder terminals on the right-hand side), each

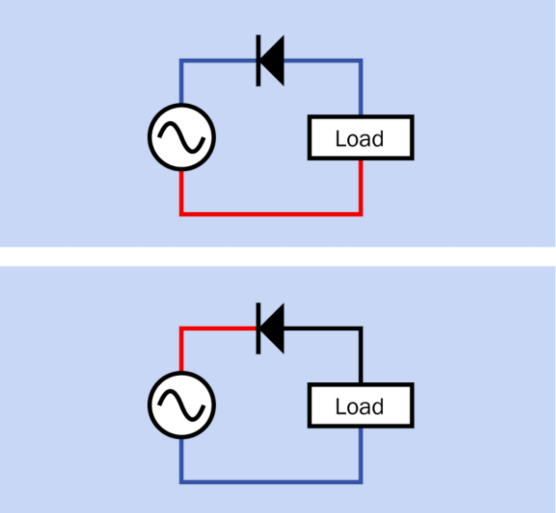

Figure 26-8. *A half-wave rectifier. In this configuration the diode allows AC current to circulate counter-clockwise but blocks it clockwise.*

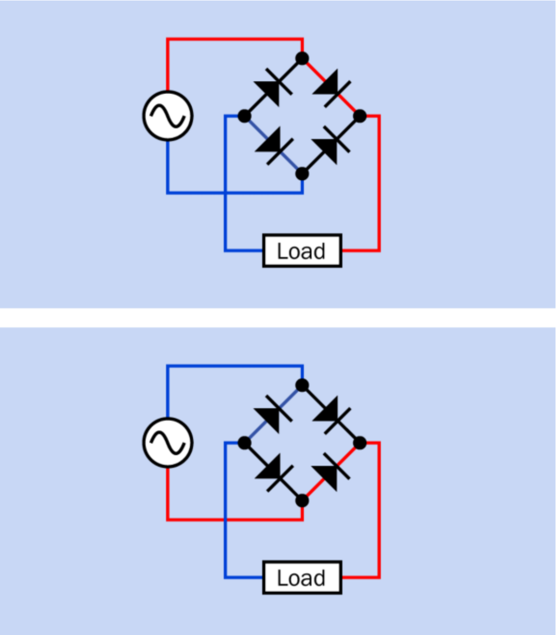

Figure 26-9. *The basic circuit commonly used to form a bridge rectifier, with color added to indicate polarity. Wires shown in black are not passing current because diodes are blocking it. Note that the polarity at the load remains constant.*

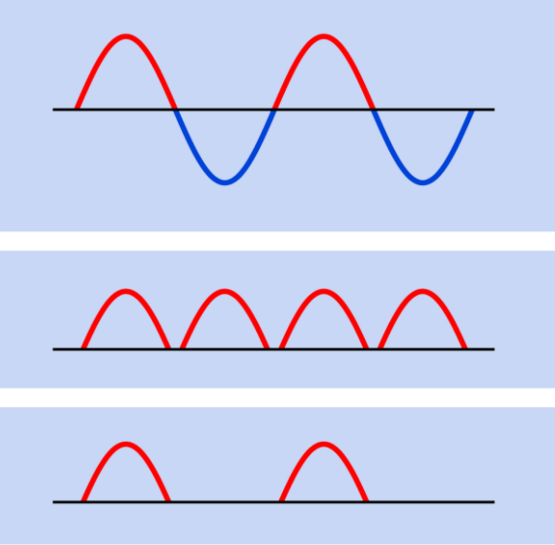

Figure 26-10. *Top: The voltage-amplitude sinewave of an alternating current source that fluctuates between positive voltage (shown red) and negative voltage (shown blue) relative to a neutral (black) baseline. Center: AC current converted by a full-wave rectifier. Because the diodes do not conduct below their threshold voltage, small gaps appear between pulses. Bottom: Output from a half-wave rectifier.*

section corresponding with the functionality of one modern diode. Figure 26-12 shows relatively modern rectifier packages, the one on the left rated at 20A continuous at 800V RMS, the one on the right rated 4A continuous at 200V RMS. In Figure 26-13, the one on the left is rated 4A continuous at 50V RMS, whereas the one on the right is rated 1.5A at 200V RMS.

DC output from rectifier packages is usually supplied via the outermost pins, while the two pins near the center receive AC current. The positive DC pin may be longer than the other three, and is usually marked with a + symbol.

Full-wave bridge rectifiers are also available in surface-mount format. The one in Figure 26-14 is rated for half an amp continuous current.

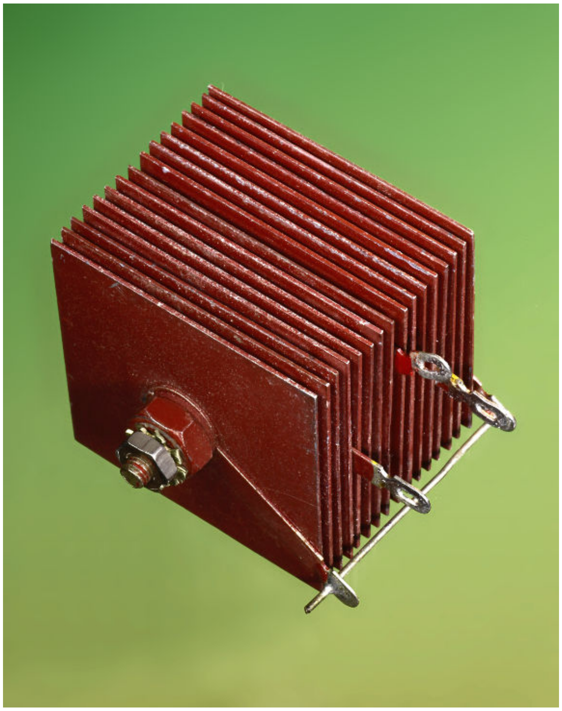

Figure 26-11. *Prior to the perfection of chip fabrication in the late 1960s, it was common to find silicon rectifiers of this type, measuring about 2" square.*

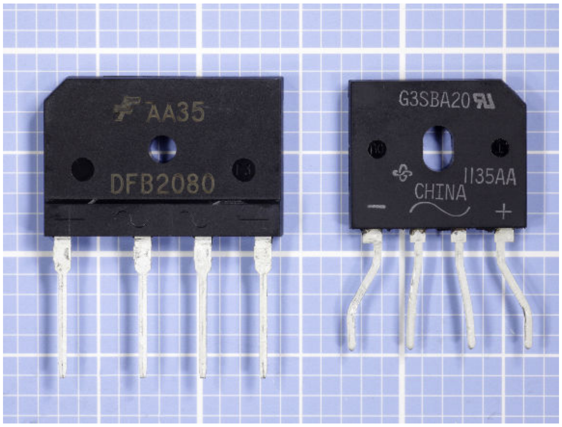

Figure 26-12. *Full-wave bridge rectifies are commonly available in packages such as these. See text for details.*

Back-EMF Suppression

A relay coil, motor, or other device with significant inductance will typically create a spike of voltage when it is turned on or off. This *EMF* can be shunted through a rectifier diode to safeguard other components in the circuit. A diode in this configuration may be referred to as a *protection diode*, a *clamp diode*, or *transient suppressor*. See Figure 26-15.

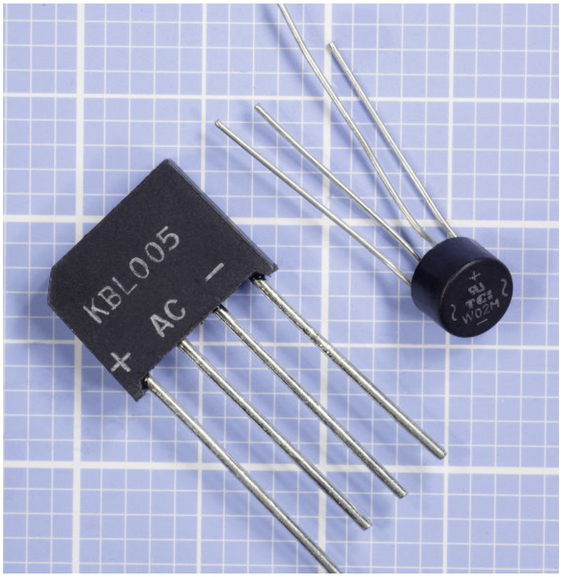

Figure 26-13. *Smaller full-wave bridge rectifiers capable of 1.5A to 4A continuous current.*

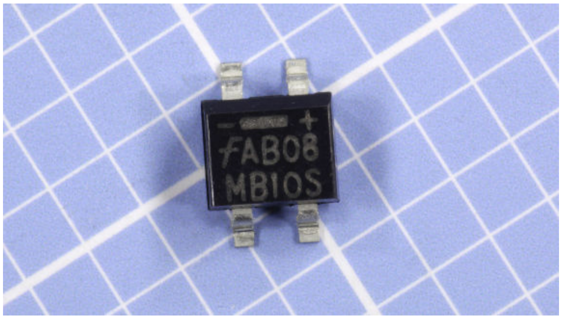

Figure 26-14. *This surface-mount component contains four diodes forming a full-wave bridge rectifier circuit, and can pass 0.5A continuous current. It measures approximately 0.2" square.*

Voltage Selection

A diode is sensitive to the relative voltage between its anode and cathode terminals. In other words, if the cathode is at 9V relative to the ground in the circuit, and the anode is at 12V, the 3V difference will easily exceed the threshold voltage, and the diode will pass current. (Actual tolerable values will depend on the forward voltage capability of the diode.) If the voltages are reversed, the diode will block the current.

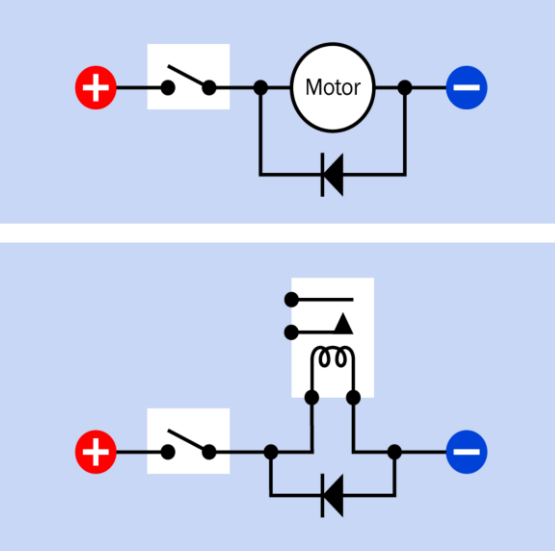

Figure 26-15. *A rectifier diode is very often placed across a motor (top), relay (bottom), or other device with significant inductance that creates a spike of reverse voltage when switched on or off. The surge is shunted through the diode, protecting other components in the circuit.*

This attribute can be used to make a device choose automatically between an AC adapter and a 9V battery. The schematic is shown in Figure 26-16. When an AC adapter that delivers 12VDC is plugged into a wall outlet, the adapter competes with the battery to provide power to a voltage regulator. The battery delivers 9VDC through the lower diode to the cathode side of the upper diode, but the AC adapter trumps it with 12VDC through the upper diode. Consequently, the battery ceases to power the circuit until the AC adapter is unplugged, at which point the battery takes over, and the upper diode now prevents the battery from trying to pass any current back through the AC adapter.

The voltage regulator in this schematic accepts either 12VDC or 9VDC and converts it to 5VDC. (In the case of 12VDC, the regulator will waste more power, which will be dissipated as waste heat.)

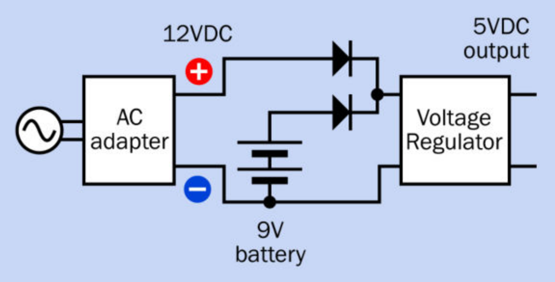

Figure 26-16. *Two diodes with their cathodes tied together will choose automatically between an AC adapter that delivers 12VDC and an internal 9V battery.*

Voltage Clamping

A diode can be used to *clamp* a voltage to a desired value. If an input to a 5V CMOS semiconductor or similarly sensitive device must be prevented from rising out of range, the anode of a diode can be connected to the input and the cathode to a 5V voltage source. If the input rises much above 5.6V, the potential difference exceeds the diode's junction threshold, and the diode diverts the excess energy. See Figure 26-17.

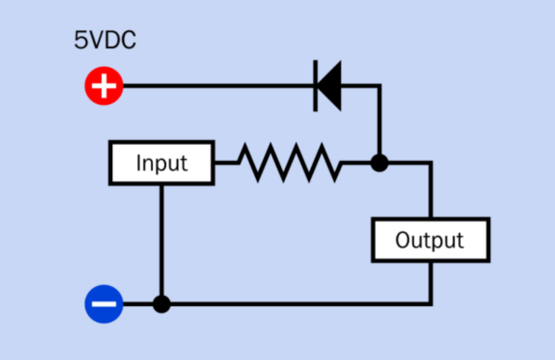

Figure 26-17. *A clamping diode can limit output voltage—in this example, to about 5.6V. If the input rises above that value relative to the common ground, the potential difference across the diode feeds the excess voltage back through it to the 5V source.*

Logic Gate

A signal diode is less than ideal as a logic gate, because it imposes a typical 0.6V voltage reduction, which can be significant in a 5V circuit and is probably unacceptable in a 3.3V circuit. Still, it can be useful on the output side—for example, if two or more outputs from a logic chip or microcontroller are intended to drive, or share, another device such as a single **LED**, as shown in Figure 26-18. In this role, the diodes wired in parallel behave similarly to an OR gate, while preventing either output from the chip from feeding current back into the other output.

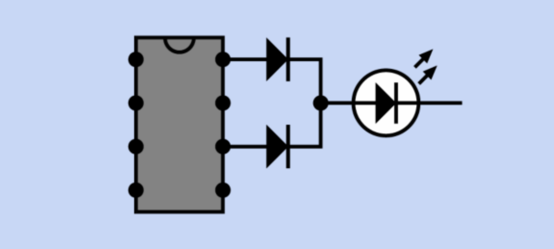

Figure 26-18. *Two or more outputs from a logic chip or microcontroller may be coupled with diodes to power another device, such as an LED, while protecting the chip from backflow of current. The diodes form a logical OR gate.*

DC Voltage Regulation and Noise Suppression

As previously noted, the dynamic resistance of a reverse-biased Zener diode will diminish as the current increases. This relationship begins at the point where breakdown in the diode begins—at its *Zener voltage*--and is approximately linear over a limited range.

The unique behavior of the Zener makes it usable as a very simple voltage controller when placed in series with a resistor as shown in Figure 26-19. It is helpful to imagine the diode and the resistor as forming a kind of voltage divider, with power being taken out at point A in the schematic. If a supply fluctuation increases the input voltage, this will tend to increase the current flowing through the Zener, and its dynamic resistance will diminish accordingly. A lower resistance in its position in the voltage divider will reduce the output voltage at point A, thus tending to compensate for the surge in input voltage.

Conversely, if the load in the circuit increases, and tends to pull down the input voltage, the current

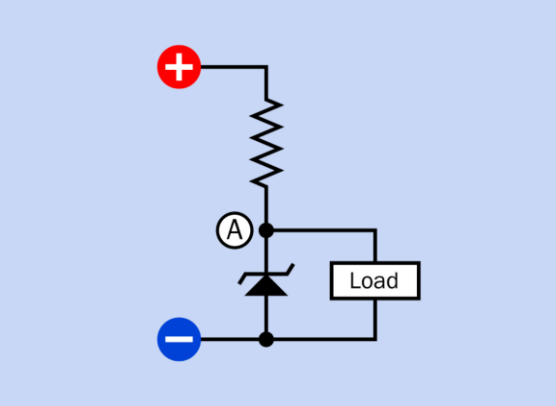

Figure 26-19. *A simplified, basic circuit illustrating the ability of a Zener diode to compensate for variations in the power supply or load in a circuit, creating an approximately constant voltage at point A.*

flowing through the Zener will diminish, and the voltage at point A will tend to increase, once again compensating for the fluctuation in the circuit.

As the series resistor would be a source of heat, a transistor could be added to drive the load, as shown in Figure 26-20.

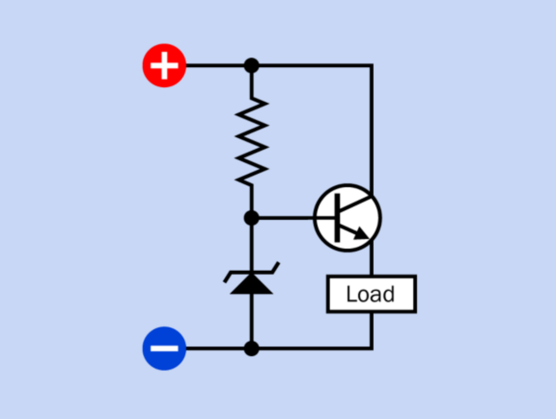

Figure 26-20. *A transistor could be added to the circuit in the previous figure to reduce power waste through the resistor.*

A manufacturer's datasheet may provide guidance regarding the dynamic resistance of a Zener diode in response to current, as previously shown in Figure 26-6. In practice, a packaged **voltage regulator** such as the LM7805 would most likely be used instead of discrete components, since it includes self-calibrating features, requires no series resistor, and is relatively unaffected by temperature. However, the LM7805 contains its own Zener diode, and the principle of operation is still the same.

AC Voltage Control and Signal Clipping

A more practical Zener application would be to limit AC voltage and/or impose clipping on an AC sinewave, using two diodes wired in series with opposed polarities. The basic schematic is shown in Figure 26-21, while clipping of the AC sinewave is illustrated in Figure 26-22. In this application, when one diode is reverse-biased, the other is forward-biased. A forward-biased Zener diode works like any other diode: it allows current to pass relatively freely, so long as the voltage exceeds its threshold. When the AC current reverses, the Zeners trade their functions, so that the first one merely passes current while the second one limits the voltage. Thus, the diodes divert peak voltage away from the load. The Zener voltage of each diode would be chosen to be a small margin above the AC voltage for voltage control, and below the AC voltage for signal clipping.

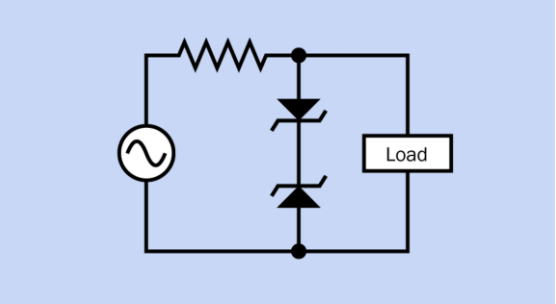

Figure 26-21. *Two Zener diodes placed in series, with opposite polarities, can clip or limit the voltage sinewave of an AC signal.*

Voltage Sensing

A Zener diode can be used to sense a small shift in voltage and provide a switched output in response.

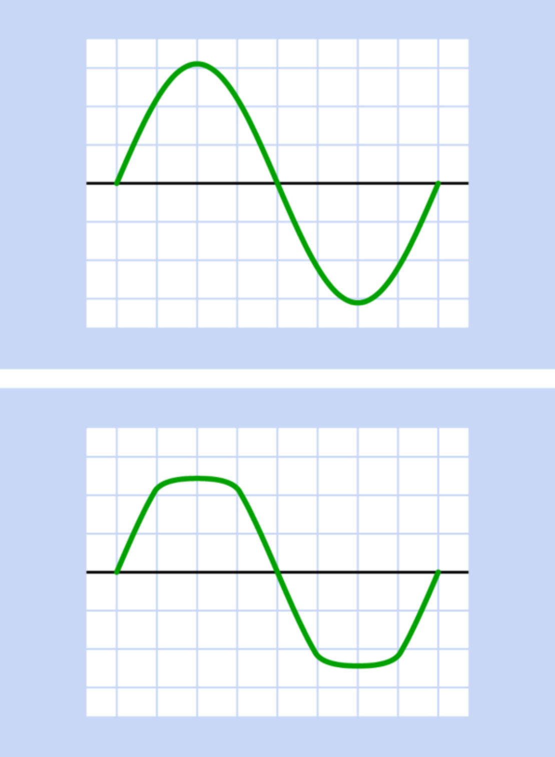

Figure 26-22. *AC input showing a pure sinewave (left) and a clipped version (right) created by Zener diodes wired in series, as in the previous figure.*

In Figure 26-23, the upper schematic shows a Zener diode preventing voltage from reaching the emitter of a PNP transistor while the divided input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, very little current flows through it, and the output is now at near-zero voltage. As soon as the input signal rises above the Zener voltage, the transistor switches on and power is supplied to the output. The input is thus replicated in the output, as shown in the upper portion of Figure 26-24.

In Figure 26-23, the lower schematic shows a Zener diode preventing voltage from reaching the base of an NPN transistor while the input signal is below the Zener (breakdown) voltage of the diode. In this mode, the transistor is relatively non-conductive, and power is supplied to the output. As soon as the input signal rises above the Zener voltage, the transistor is activated, diverting the current to ground and bypassing the output, which is now at near-zero voltage. The input is thus inverted, as shown in the lower portion of Figure 26-24 (provided there is enough current to drive the transistor into saturation).

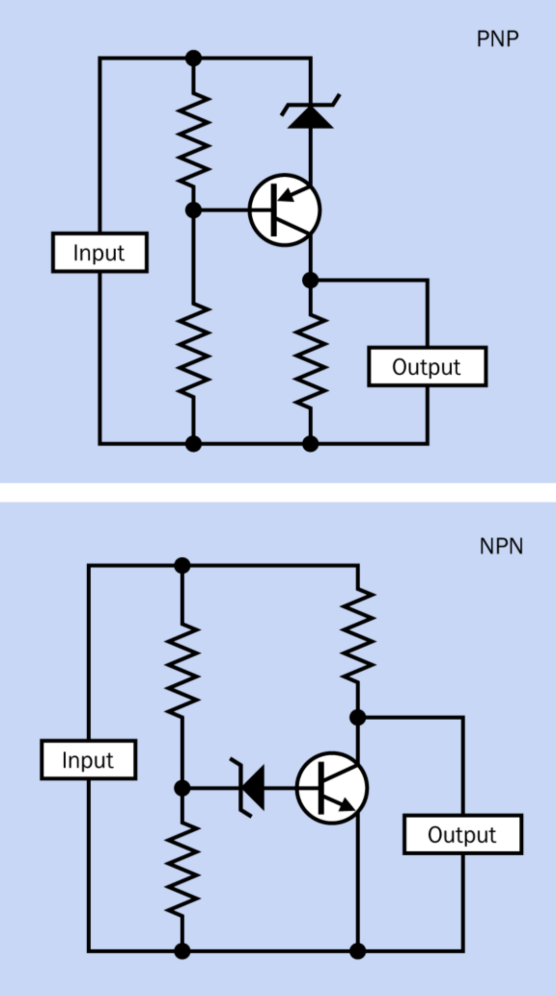

Figure 26-23. *A Zener diode can be used in conjunction with a PNP transistor. See text for details.*

What Can Go Wrong

Overload

If maximum forward current is exceeded, the heat generated is likely to destroy the diode. If the diode is reverse-biased beyond its peak in-

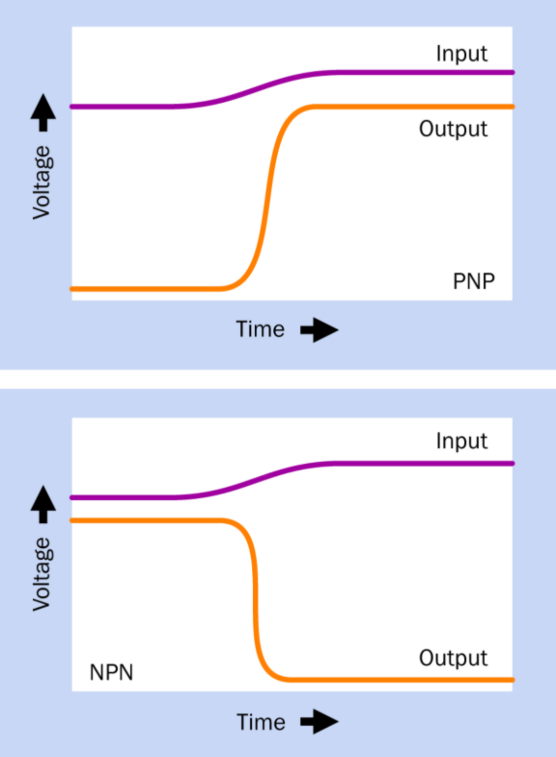

Figure 26-24. *Theoretical output from the transistors in the two previous schematics.*

verse voltage limit, the current will overwhelm the diode's ability to block it, and an *avalanche breakdown* will occur, once again probably destroying the component. The graph in Figure 26-5 illustrates the performance range of a hypothetical generic diode.

Reversed Polarity

Zener diodes look almost identical to other types, and all diodes share the same convention of marking the cathode for identification. Yet Zeners must be reverse-biased while others are forward-biased. This creates a significant risk of installing a diode "the wrong way around," with potentially destructive or at least confusing results, especially when used in a power supply. The very low resistance of a diode to forward current makes it especially vulnerable to burnout if installed incorrectly.

Wrong Type of Diode

If a Zener diode is used accidentally where a signal or rectifier diode is appropriate, the circuit will malfunction, as the Zener will probably have a much lower breakdown voltage, and therefore will not block reverse current. Conversely, if a signal or rectifier diode is used where the circuit calls for a Zener diode, reverse voltage will be clamped (or regulated at the diode's forward voltage value). Since diodes are often poorly marked, a sensible precaution is to store Zener diodes separately from all other types.

unijunction transistor

27

The unijunction transistor (UJT) and programmable unijunction transistor (PUT) are different internally, but are sufficiently similar in function to be combined in this entry.

OTHER RELATED COMPONENTS

- **diode** (See Chapter 26)
- **bipolar transistor** (See Chapter 28)
- **field-effect transistor** (See Chapter 29)

What It Does

Despite their names, the unijunction transistor (UJT) and programmable unijunction transistor (PUT) are not current-amplification devices like bipolar transistors. They are switching components that are more similar to diodes than to transistors.

The UJT can be used to build low- to mid-frequency oscillator circuits, while the PUT provides similar capability with the addition of more sophisticated control, and is capable of functioning at lower currents. The UJT declined in popularity during the 1980s after introduction of components such as the *555 timer*, which offered more flexibility and a more stable output frequency, eventually at a competitive price. UJTs are now uncommon, but PUTs are still available in quantity as through-hole discrete components. Whereas an integrated circuit such as a 555 timer generates a square wave, unijunction transistors in oscillator circuits generate a series of voltage spikes.

The PUT is often used to trigger a **thyristor** (described in Volume 2) and has applications in low-power circuits, where it can draw as little as a few microamps.

Schematic symbols for the two components are shown in Figure 27-1 and Figure 27-2. Although the symbol for the UJT is very similar to the symbol for a field-effect transistor (FET), its behavior is quite different. The bent arrow identifies the UJT, while a straight arrow identifies the FET. This difference is of significant importance.

The schematic symbol for a PUT indicates its function, as it resembles a diode with the addition of a gate connection.

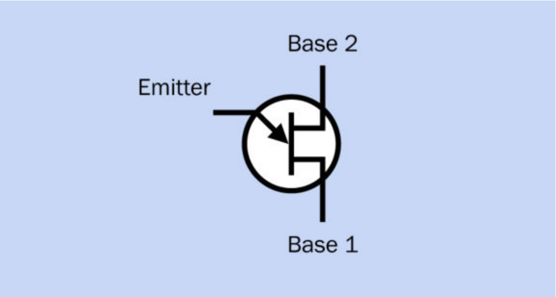

Figure 27-1. *Schematic symbol for a unijunction transistor (UJT). Note the bent arrow. The symbol for a field-effect transistor looks similar, but has a straight arrow. The functionality of the two components is very different.*

In Figure 27-3, the transistors at left and center are old-original unijunction transistors, while the one at right is a programmable unijunction transistor. (Left: Maximum 300mW, 35V interbase voltage. Center: 450mW, 35V interbase voltage. Right: 300mW, 40V gate-cathode forward voltage, 40V anode-cathode voltage.)

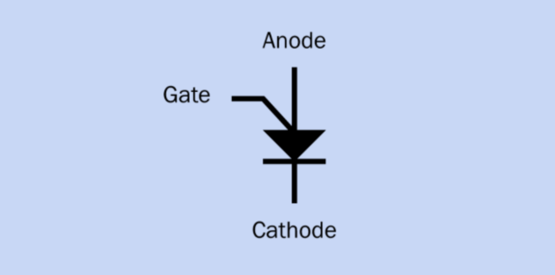

Figure 27-2. *Schematic symbol for a programmable unijunction transistor (PUT). The symbol accurately suggests the similarity in function to a diode, with the addition of a gate to adjust the threshold voltage.*

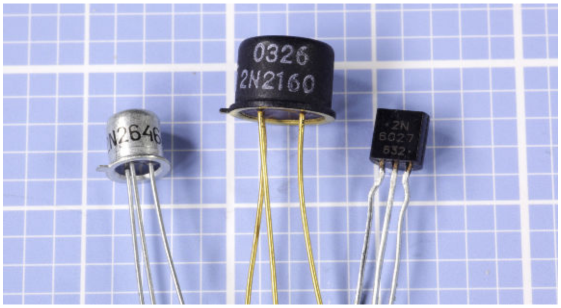

Figure 27-3. *The unijunction transistors at left and center are becoming obsolete; the one at the right is a programmable unijunction transistor (PUT), still readily available and widely used as a thyristor trigger.*

How It Works

The UJT is a three-terminal semiconductor device, but contains only two sections sharing a single junction—hence its name. Leads attached to opposite ends of a single channel of N-type semiconductor are referred to as base 1 and base 2, with base 2 requiring a slightly higher potential than base 1. A smaller P-type insert, midway between base 1 and base 2, is known as the emitter.

The diagram in Figure 27-4 gives an approximate idea of internal function.

When no voltage is applied to the emitter, a relatively high resistance (usually more than 5K) prevents significant current flow from base 2 to base 1. When the positive potential at the emitter increases to a *triggering voltage* (similar to the *junction threshold voltage* of a forward-biased diode), the internal resistance of the UJT drops very rapidly, allowing current to enter the component via both the emitter and base 2, exiting at base 1. (The term "current" refers, here, to conventional current; electron flow is opposite.) Current flowing from base 2 to base 1 is significantly greater than current flowing from the emitter to base 1.

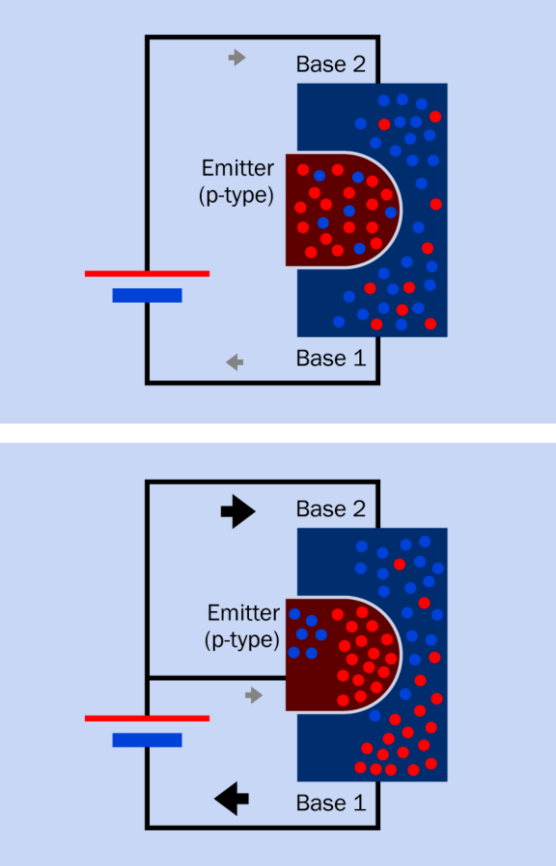

Figure 27-4. *Internal workings of a unijunction transistor.*

The graph in Figure 27-5 outlines the behavior of a UJT. As the voltage applied to the emitter increases, current flowing into the component from the emitter increases slightly, until the triggering voltage is reached. The component's internal resistance now drops rapidly. This pulls down the voltage at the emitter, while the current continues to increase significantly. Because of the drop in resistance, this is referred to as a *negative resistance* region. The resistance actually cannot fall below zero, but its change is negative. After emitter voltage drops to a minimum known as the *valley voltage*, the current continues to increase with a small increase in voltage. On datasheets, the peak current is often referred to as I_p while valley current is I_v.

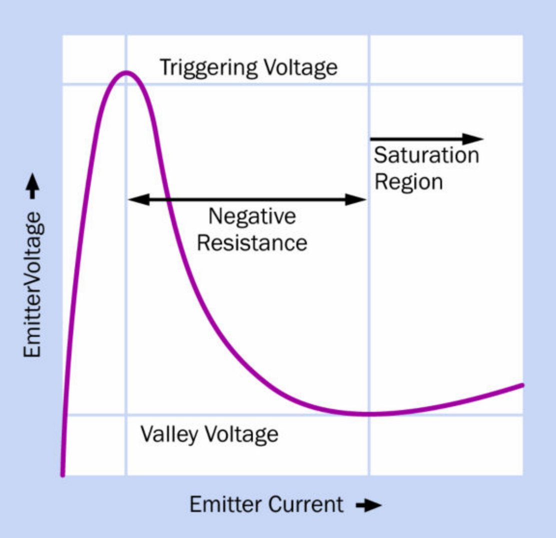

Figure 27-5. *Response curve of a unijunction transistor (UJT). When positive potential at the emitter reaches the triggering voltage, internal resistance drops radically and the component goes through a phase known as "negative resistance" as current increases.*

Figure 27-6 shows a test circuit to demonstrate the function of a UJT, with a volt meter indicating its status. A typical supply voltage would range from 9VDC to 20VDC.

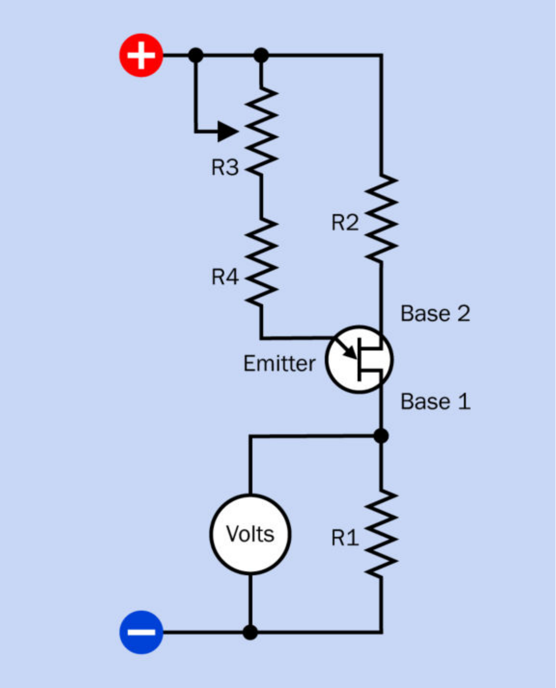

Figure 27-6. *A test circuit for a unijunction transistor (UJT) using a volt meter to show its response as a potentiometer increases the voltage applied to its emitter.*

A PUT behaves similarly in many ways to a UJT but is internally quite different, consisting of four semiconducting layers and functioning similarly to a thyristor.

The PUT is triggered by increasing the voltage on the anode. Figure 27-7 shows a test circuit for a PUT. This component is triggered when the voltage at its anode exceeds a threshold level, while the gate sets the threshold where this occurs. When the PUT is triggered, its internal resistance drops, and current flows from anode to cathode, with a smaller amount of current entering through the gate. This behavior is almost identical to that of a forward-biased diode, except that the threshold level can be controlled, or "programmed," according to the value of the positive potential applied at the gate, with R1 and R2 establishing that potential by functioning as a voltage divider.

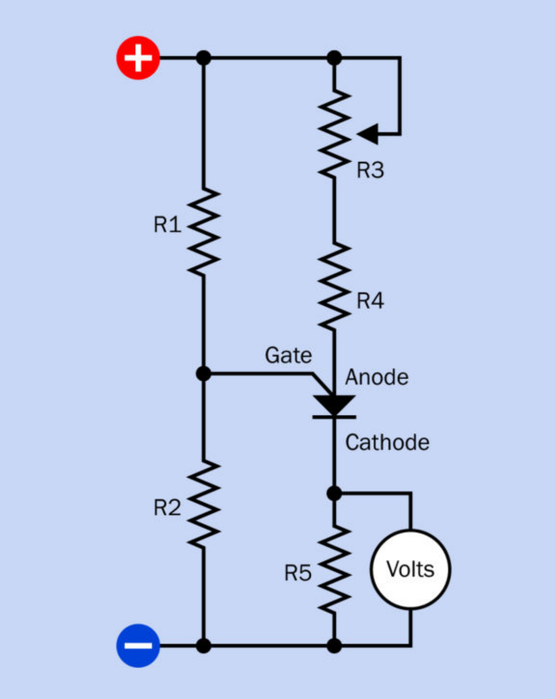

Figure 27-7. *A test circuit for a programmable unijunction transistor (PUT) using a volt meter to show its response as a potentiometer increases the voltage applied to its anode.*

The voltage output of a PUT follows a curve that is very similar to that shown in Figure 27-5, although current and voltage would be measured at the cathode.

Variants

PUTs and UJTs are not made as surface-mount components.

UJTs are usually packaged in black plastic, although older variants were manufactured in cans. PUTs are almost all packaged in black plastic. With the leads pointing downward and the flat side facing toward the viewer, the lead functions of a PUT are usually anode, gate, and cathode, reading from left to right.

Values

The triggering voltage of a UJT can be calculated from the values of R1 and R2 in Figure 27-7 and the voltage at base 1. The term R_{bb} is often used to represent the sum of R1 + R2, with V_{bb} representing the total voltage across the two resistors (this is the same as the supply voltage in Figure 27-6). V_t, the triggering voltage, is given by:

$$V_t = V_{bb} * (R1 / R_{bb})$$

The term ($R1/R_{bb}$) is known as the *standoff ratio*, often represented by the Greek letter ϱ.

Typically the standoff ratio in a UJT is at least 0.7, as R1 is chosen to be larger than R2. Typical values for R1 and R2 could be 180Ω and 100Ω, respectively. If R4 is 50K and a 100K linear potentiometer is used for R3, the PUT should be triggered when the potentiometer is near the center of its range. The emitter saturation voltage is typically from 2V to 4V.

If using a PUT, typical values in the test circuit could be supply voltage ranging from 9VDC to 20VDC, with resistances 28K for R1 and 16K for R2, 20Ω for R5, 280K for R4, and a 500K linear potentiometer for R3. The PUT should be triggered when the potentiometer is near the center of its range.

Sustained forward current from anode to cathode is usually a maximum of 150mA, while from gate to cathode the maximum is usually 50mA. Power dissipation should not exceed 300mW. These values should be lower at temperatures above 25 degrees Centigrade.

Depending on the PUT being used, power consumption can be radically decreased by upping the resistor values by a multiple of 100, while supply voltage can be decreased to 5V. The cathode output from the PUT would then be connected with the base of an NPN transistor for amplification.

How to Use it

Figure 27-8 shows a simple oscillator circuit built around a UJT, Figure 27-9 shows a comparable circuit for a PUT. Initially the supply voltage charges the capacitor, until the potential at the emitter of the UJT or the gate of the PUT reaches the threshold voltage, at which point the capacitor discharges through the emitter and the cycle repeats. Resistor values would be similar to those used in the test circuits previously described, while a capacitor value of 2.2µF would provide a visible pulse of the LED. Smaller capacitor values would enable faster oscillation. In the PUT circuit, adjusting the values of R1 and R2 would allow fine control of triggering the semiconductor.

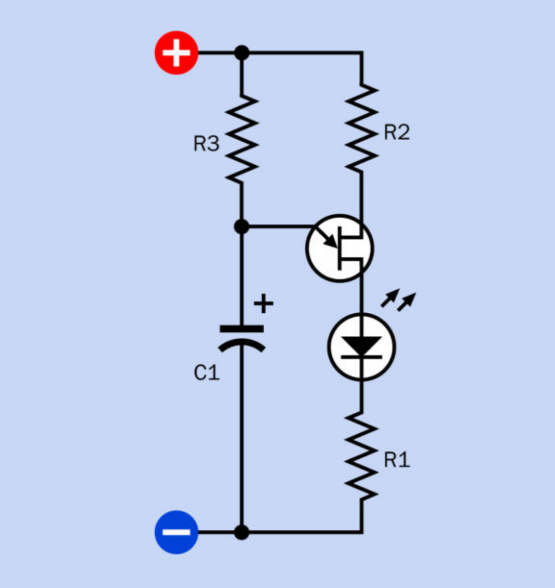

Figure 27-8. *A basic oscillator circuit using a unijunction transistor (UJT). As the capacitor accumulates charge, the voltage on the emitter increases until it triggers the UJT, at which point the capacitor discharges through the emitter.*

Probably the most common use for a PUT at this time is to trigger a **thyristor**.

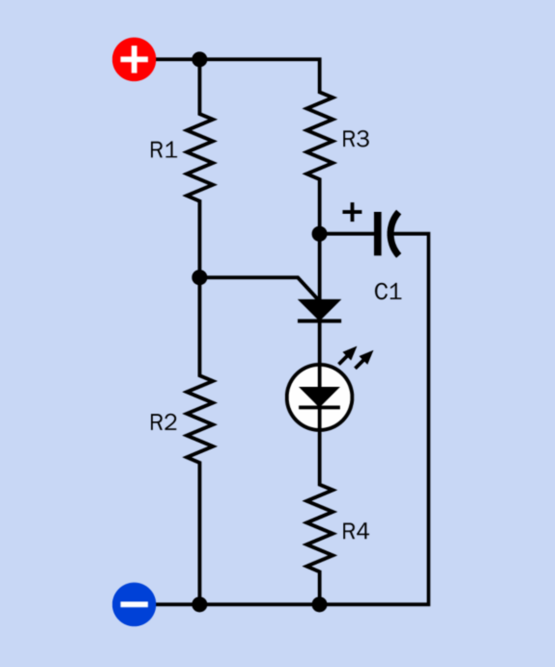

Figure 27-9. *A basic oscillator circuit using a programmable unijunction transistor (PUT). As the capacitor accumulates charge, the voltage on the anode increases until it triggers the PUT, at which point the capacitor discharges through the anode. The gate voltage is preset by R1 and R2 to adjust the triggering voltage.*

What Can Go Wrong

Name Confusion

A programmable unijunction transistor (PUT) is sometimes referred to simply as a “unijunction transistor” (UJT). Bearing in mind the totally different modes of operation of UJT and PUT, the PUT should always be identified by its acronym or by its full name. A circuit will not function if a UJT is substituted for a PUT, or a PUT is substituted for a UJT.

Incorrect Bias

Neither the UJT nor the PUT is designed to operate with reverse bias. In the UJT, a small forward bias should be applied from base 2 to base 1 (that is, base 2 should be at a higher potential relative to base 1) regardless of the voltage on the emitter. The emitter voltage may vary from 0 volts upward. The PUT must be forward biased between its anode and cathode (the anode must have a higher potential relative to the cathode), with an intermediate positive voltage at the gate established by resistors R1 and R2 functioning as a voltage divider (see Figure 27-7). Failure to observe correct biasing will result in unpredictable behavior and possible damage to the component.

Overload

Like any semiconductor, the UJT and the PUT must be protected from excessive current, which can burn out the component. Never connect either of these components directly across a power source without appropriate resistances to limit current flow. Maximum continuous power dissipation for UJTs and PUTs is usually 300mW.

bipolar transistor

28

The word *transistor*, on its own, is often used to mean *bipolar transistor*, as this was the type that became most widely used in the field of discrete semiconductors. However, *bipolar transistor* is the correct term. It is sometimes referred to as a *bipolar junction transistor* or *BJT*.

OTHER RELATED COMPONENTS

- **unijunction transistor** (See Chapter 27)
- **field-effect transistor** (See Chapter 29)
- **diac** (Volume 2)
- **triac** (Volume 2)
- **relay** (Volume 2)
- **solid-state relay** (Volume 2)

What It Does

A bipolar transistor amplifies fluctuations in current or can be used to switch current on and off. In its amplifying mode, it replaced the *vacuum tubes* that were formerly used in the amplification of audio signals and many other applications. In its switching mode it resembles a **relay**, although in its "off" state the transistor still allows a very small amount of current flow, known as *leakage*.

A bipolar transistor is described as a *discrete* semiconductor device when it is individually packaged, with three leads or contacts. A package containing multiple transistors is an *integrated circuit*. A Darlington pair actually contains two transistors, but is included here as a discrete component because it is packaged similarly and functions like a single transistor. Most integrated circuits will be found in Volume 2 of this encyclopedia.

How It Works

Although the earliest transistors were fabricated from germanium, silicon has become the most commonly used material. Silicon behaves like an insulator, in its pure state at room temperature, but can be "doped" (carefully contaminated) with impurities that introduce a surplus of electrons unbonded from individual atoms. The result is an *N-type semiconductor* that can be induced to allow the movement of electrons through it, if it is *biased* with an external voltage. *Forward bias* means the application of a positive voltage, while *reverse bias* means reversing that voltage.

Other dopants can create a deficit of electrons, which can be thought of as a surplus of "holes" that can be filled by electrons. The result is a *P-type semiconductor*.

A bipolar NPN transistor consists of a thin central P-type layer sandwiched between two thicker N-type layers. The three layers are referred to as *collector*, *base*, and *emitter*, with a wire or contact attached to each of them. When a negative charge is applied to the emitter, electrons are forced by mutual repulsion toward the central base layer. If a forward bias (positive potential) is applied to the base, electrons will tend to be attracted out through the base. However, because the base layer is so thin, the electrons are now close to the collector. If the base voltage increases, the additional energy encourages the electrons to jump into the collector, from which they will make their way to the positive current source, which can be thought of as having an even greater deficit of electrons.

Thus, the emitter of an NPN bipolar transistor emits electrons into the transistor, while the collector collects them from the base and moves them out of the transistor. It is important to remember that since electrons carry a negative charge, the flow of electrons moves from negative to positive. The concept of positive-to-negative current is a fiction that exists only for historical reasons. Nevertheless, the arrow in a transistor schematic symbol points in the direction of conventional (positive-to-negative) current.

In a PNP transistor, a thin N-type layer is sandwiched between two thicker P-type layers, the base is negatively biased relative to the emitter, and the function of an NPN transistor is reversed, as the terms "emitter" and "collector" now refer to the movement of electron-holes rather than electrons. The collector is negative relative to the base, and the resulting positive-to-negative current flow moves from emitter to base to collector. The arrow in the schematic symbol for a PNP transistor still indicates the direction of positive current flow.

Symbols for NPN and PNP transistors are shown in Figure 28-1. The most common symbol for an NPN transistor is shown at top-left, with letters C, B, and E identifying collector, base, and emitter. In some schematics the circle in the symbols is omitted, as at top-right.

A PNP transistor is shown in the center. This is the most common orientation of the symbol, since its collector must be at a lower potential than its emitter, and ground (negative) is usually at the bottom of a schematic. At bottom, the PNP symbol is inverted, allowing the positions of emitter and collector to remain the same as in the symbol for the NPN transistor at the top. Other orientations of transistor symbols are often found, merely to facilitate simpler schematics with fewer conductor crossovers. The direction of the arrow in the symbol (pointing out or pointing in) always differentiates NPN from PNP transistors, respectively, and indicates current flowing from positive to negative.

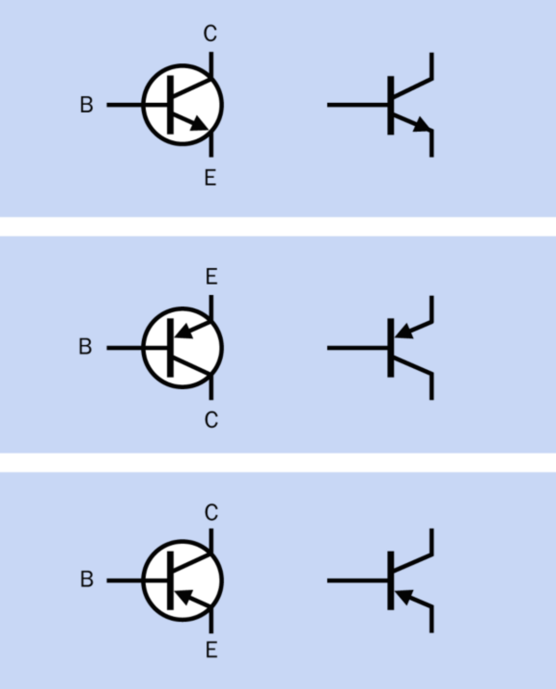

Figure 28-1. *Symbols for an NPN transistor (top) and a PNP transistor (center and bottom). Depending on the schematic in which the symbol appears, it may be rotated or inverted. The circle may be omitted, but the function of the component remains the same.*

NPN transistors are much more commonly used than PNP transistors. The PNP type was more difficult and expensive to manufacture initially, and circuit design evolved around the NPN type. In addition, NPN transistors enable faster switching, because electrons have greater mobility than electron-holes.

To remember the functions of the collector and the emitter in an NPN transistor, you may prefer to think in terms of the collector collecting positive current *into* the transistor, and the emitter emitting positive current *out of* the transistor. To remember that the emitter is always the terminal with an arrow attached to it (both in NPN and PNP schematic symbols), consider that "emitter" and "arrow" both begin with a vowel, while "base" and "collector" begin with consonants. To remember that an NPN transistor symbol has its arrow pointing outward, you can use the mnemonic "N/ever P/ointing i/N."

Current flow for NPN and PNP transistors is illustrated in Figure 28-2. At top-left, an NPN transistor passes no current (other than a small amount of leakage) from its collector to its emitter so long as its base is held at, or near, the potential of its emitter, which in this case is tied to negative or ground. At bottom-left, the purple positive symbol indicates that the base is now being held at a *relatively* positive voltage, at least 0.6 volts higher than the emitter (for a silicon-based transistor). This enables electrons to move from the emitter to the collector, in the direction of the blue arrows, while the red arrows indicate the conventional concept of current flowing from positive to negative. The smaller arrows indicate a smaller flow of current. A resistor is included to protect the transistor from excessive current, and can be thought of as the load in these circuits.

At top-right, a PNP transistor passes no current (other than a small amount of leakage) from its emitter to its collector so long as its base is held at, or near, the potential of the emitter, which in this case is tied to the positive power supply. At bottom-right, the purple negative symbol indicates that the base is now being held at a *relatively* negative voltage, at least 0.6 volts lower than the emitter. This enables electrons and current to flow as shown. Note that current flows into the base in the NPN transistor, but out from the base in the PNP transistor, to enable conductivity. In both diagrams, the resistor that would normally be included to protect the base has been omitted for the sake of simplicity.

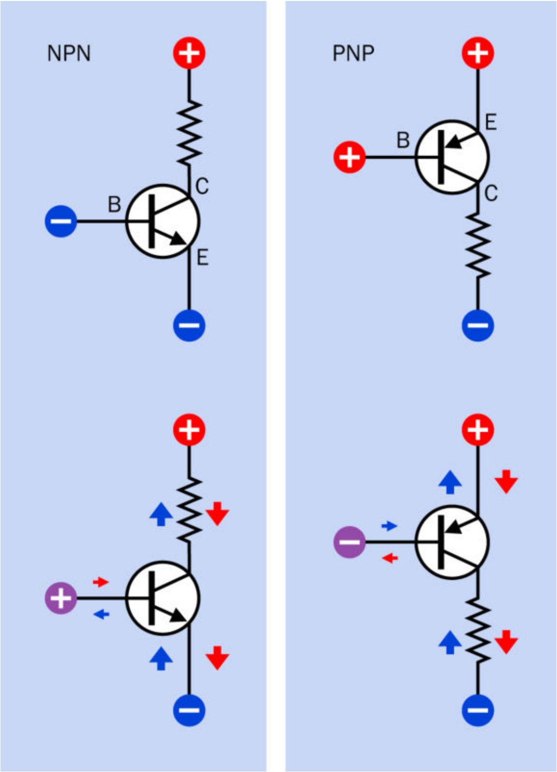

Figure 28-2. *Current flow through NPN and PNP transistors. See text for details.*

An NPN transistor amplifies its base current only so long as the positive potential applied to the collector is greater than the potential applied to the base, and the potential at the base must be greater than the potential at the emitter by at least 0.6 volts. So long as the transistor is biased in this way, and so long as the current values remain within the manufacturer's specified limits, a small fluctuation in current applied to the base will induce a much larger fluctuation in current between the collector and the emitter. This is why a transistor may be described as a *current amplifier*.

A *voltage divider* is often used to control the base potential and ensure that it remains less than the potential on the collector and greater than the potential at the emitter (in an NPN transistor). See Figure 28-3.

See Chapter 10 for additional information about the function of a voltage divider.

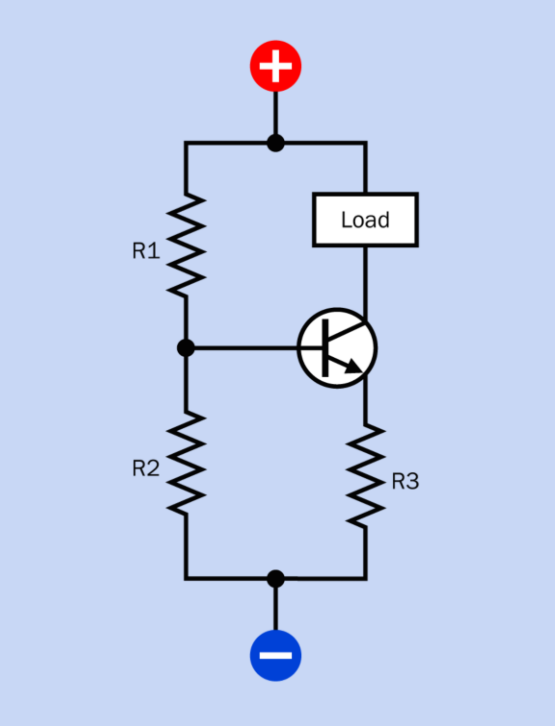

Figure 28-3. *Resistors R1 and R2 establish a voltage divider to apply acceptable bias to the base of an NPN transistor.*

Current Gain

The amplification of current by a transistor is known as its *current gain* or *beta value*, which can be expressed as the ratio of an increase in collector current divided by the increase in base current that enables it. Greek letter β is customarily used to represent this ratio. The formula looks like this:

$$\beta = \Delta I_c \;/\; \Delta I_b$$

where I_c is the collector current and I_b is the base current, and the Δ symbol represents a small change in the value of the variable that follows it.

Current gain is also represented by the term h_{FE}, where E is for the common Emitter, F is for Forward current, and lowercase letter h refers to the transistor as a "hybrid" device.

The beta value will always be greater than 1 and is often around 100, although it will vary from one type of transistor to another. It will also be affected by temperature, voltage applied to the transistor, collector current, and manufacturing inaccuracies. When the transistor is used outside of its design parameters, the formula to determine the beta value no longer directly applies.

There are only two connections at which current can enter an NPN transistor and one connection where it can leave. Therefore, if I_e is the current from the emitter, I_c is the current entering the collector, and I_b is the current entering the base:

$$I_e = I_c + I_b$$

If the potential applied to the base of an NPN transistor diminishes to the point where it is less than 0.6V above the potential at the emitter, the transistor will not conduct, and is in an "off" state, although a very small amount of leakage from collector to emitter will still occur.

When the current flowing into the base of the transistor rises to the point where the transistor cannot amplify the current any further, it becomes *saturated*, at which point its internal impedance has fallen to a minimal value. Theoretically this will allow a large flow of current; in practice, the transistor should be protected by resistors from being damaged by high current resulting from saturation.

Any transistor has maximum values for the collector current, base current, and the potential difference between collector and emitter. These values should be provided in a datasheet. Exceeding them is likely to damage the component.

Terminology

In its *saturated mode*, a transistor's base is saturated with electrons (with no room for more) and the internal impedance between collector and emitter drops as low as it can go.

The *cutoff mode* of an NPN transistor is the state where a low base voltage eliminates all current flow from collector to emitter other than a small amount of leakage.

The *active mode*, or *linear mode*, is the intermediate condition between cutoff and saturated, where the beta value or h_{FE} (ratio of collector current to base current) remains approximately constant. That is, the collector current is almost linearly proportional to the base current. This linear relationship breaks down when the transistor nears its saturation point.

Variants

Small signal transistors are defined as having a maximum collector current of 500 mA and maximum collector power dissipation of 1 watt. They can be used for audio amplification of low-level inputs and for switching of small currents. When determining whether a small-signal transistor can control an inductive load such as a motor or relay coil, bear in mind that the initial current surge will be greater than the rated current draw during sustained operation.

Small switching transistors have some overlap in specification with small signal transistors, but generally have a faster response time, lower beta value, and may be more limited in their tolerance for collector current. Check the manufacturer's datasheet for details.

High frequency transistors are primarily used in video amplifiers and oscillators, are physically small, and have a maximum frequency rating as high as 2,000 MHz.

Power transistors are defined as being capable of handling at least 1 watt, with upper limits that can be as high as 500 watts and 150 amps. They are physically larger than the other types, and may be used in the output stages of audio amplifiers, and in switching **power supplies** (see Chapter 16). Typically they have a much lower current gain than smaller transistors (20 or 30 as opposed to 100 or more).

Sample transistors are shown in Figure 28-4. Top: A 2N3055 NPN power transistor. This type was originally introduced in the late 1960s, and versions are still being manufactured. It is often found in power supplies and in push-pull power amplifiers, and has a total power dissipation rating of 115W. Second row, far left: general purpose switching-amplification PNP power transistor rated for up to 50W power dissipation. Second row, far right: A high-frequency switching transistor for use in lighting ballast, converters, inverters, switching regulators, and motor control systems. It tolerates relatively high voltages (up to 700V collector-emitter peak) and is rated for up to 80W total power dissipation. Second row, center-left and center-right: Two variants of the 2N2222 NPN small signal switching transistor, first introduced in the 1960s, and still very widely used. The metal can is the TO-19 package, capable of slightly higher power dissipation than the cheaper plastic TO-92 package (1.8W vs. 1.5W with a collector temperature no greater than 25 degrees Centigrade).

Packaging

Traditionally, small-signal transistors were packaged in small aluminum "cans" about 1/4" in diameter, and are still sometimes found in this form. More commonly they are embedded in buds of black plastic. Power transistors are packaged either in a rectangular module of black

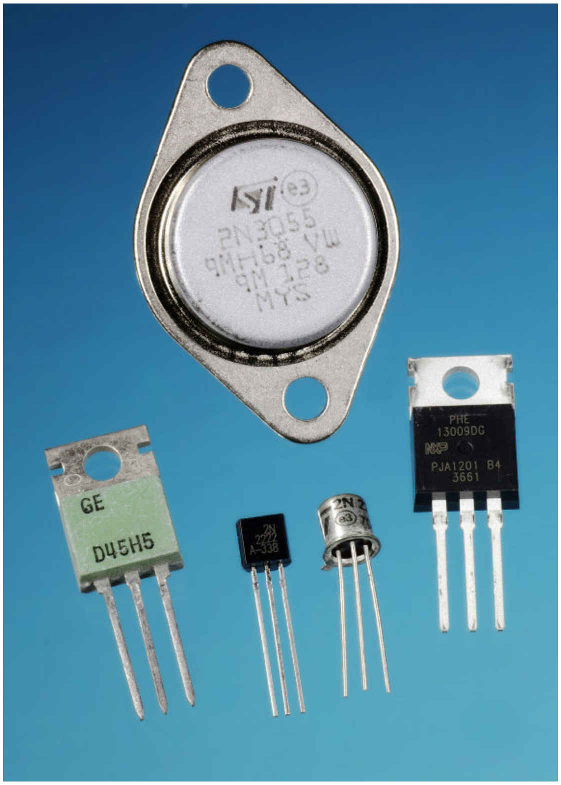

Figure 28-4. *Samples of commonly used transistors. See text for details.*

plastic with a metal back, or in a round metal "button." Both of these forms are designed to dissipate heat by being screw-clamped to a *heat sink*.

Connections

Often a transistor package provides no clue as to which lead is the emitter, which lead is the base, and which lead is the collector. Old can-style packaging includes a protruding tab that usually points toward the emitter, but not always. Where power transistors are packaged in a metal enclosure, it is typically connected internally with the collector. In the case of surface-mount transistors, look for a dot or marker that should identify the base of a bipolar transistor or the gate of a field-effect transistor.

A through-hole transistor usually has its part number printed or engraved on its package, although a magnifying glass may be necessary to see this. The component's datasheet may then be checked online. If a datasheet is unavailable, meter-testing will be necessary to confirm the functions of the three transistor leads. Some multimeters include a transistor-test function, which may validate the functionality of a transistor while also displaying its beta value. Otherwise, a meter can be put in diode-testing mode, and an unpowered NPN transistor should behave as if diodes are connected between its leads as shown in Figure 28-5. Where the identities of the transistor's leads are unknown, this test will be sufficient to identify the base, after which the collector and emitter may be determined empirically by testing the transistor in a simple low-voltage circuit such as that shown in Figure 28-6.

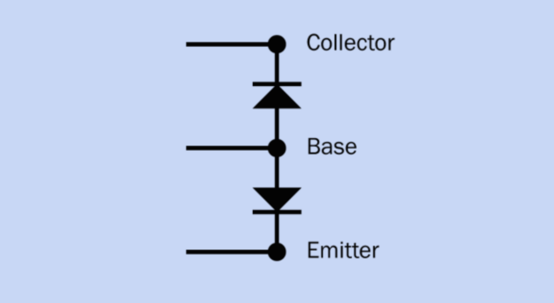

Figure 28-5. *An NPN transistor can behave as if it contains two diodes connected as shown. Where the functions of the leads of the transistor are unknown, the base can be identified by testing for conductivity.*

How to Use it

The following abbreviations and acronyms are common in transistor datasheets. Some or all of the letters following the initial letter are usually, but not always, formatted as subscripts:

h_{FE}
Forward current gain

β
Same as hFE

V_{CEO}
Voltage between collector and emitter (no connection at base)

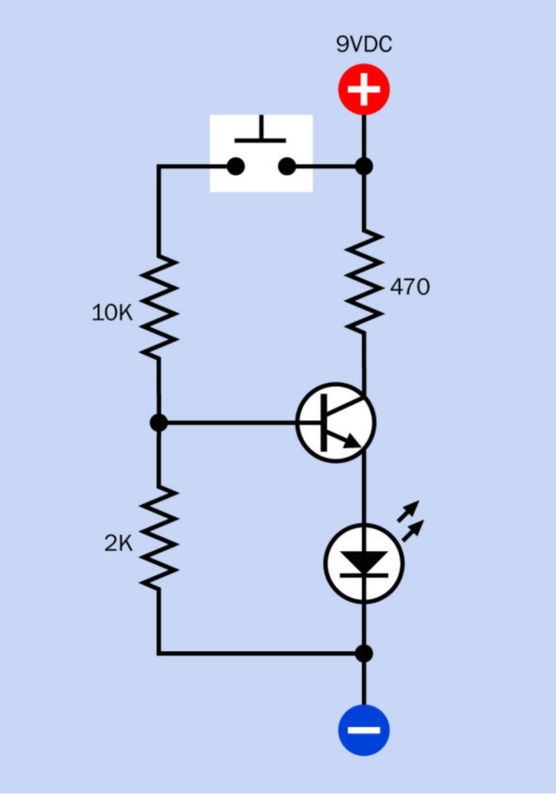

Figure 28-6. *This simple schematic can be used to breadboard-test a transistor empirically, determining its functionality and the identities of its collector and emitter leads.*

V_{CBO}

Voltage between collector and base (no connection at emitter)

V_{EBO}

Voltage between emitter and base (no connection at collector)

V_{CEsat}

Saturation voltage between collector and emitter

V_{BEsat}

Saturation voltage between base and emitter

I_c

Current measured at collector

I_{CM}

Maximum peak current at collector

I_{BM}

Maximum peak current at base

P_{TOT}

Total maximum power dissipation at room temperature

T_J

Maximum junction temperature to avoid damage

Often these terms are used to define “absolute maximum values” for a component. If these maximums are exceeded, damage may occur.

A manufacturer’s datasheet may include a graph showing the *safe operating area* (SOA) for a transistor. This is more common where power transistors are involved, as heat becomes more of an issue. The graph in Figure 28-7 has been adapted from a datasheet for a silicon diffused power transistor manufactured by Philips. The safe operating area is bounded at the top by a horizontal segment representing the maximum safe current, and at the right by a vertical segment representing the maximum safe voltage. However, the rectangular area enclosed by these lines is reduced by two diagonal segments representing the *total power limit* and the *second breakdown limit*. The latter refers to the tendency of a transistor to develop internal localized “hot spots” that tend to conduct more current, which makes them hotter, and able to conduct better—ultimately melting the silicon and causing a short circuit. The total power limit and the second breakdown limit reduce the safe operating area, which would otherwise be defined purely by maximum safe current and maximum safe voltage.

Uses for discrete transistors began to diminish when integrated circuits became cheaper and started to subsume multi-transistor circuits. For instance, an entire 5-watt audio amplifier, which used to be constructed from multiple components can now be bought on a chip, requiring just

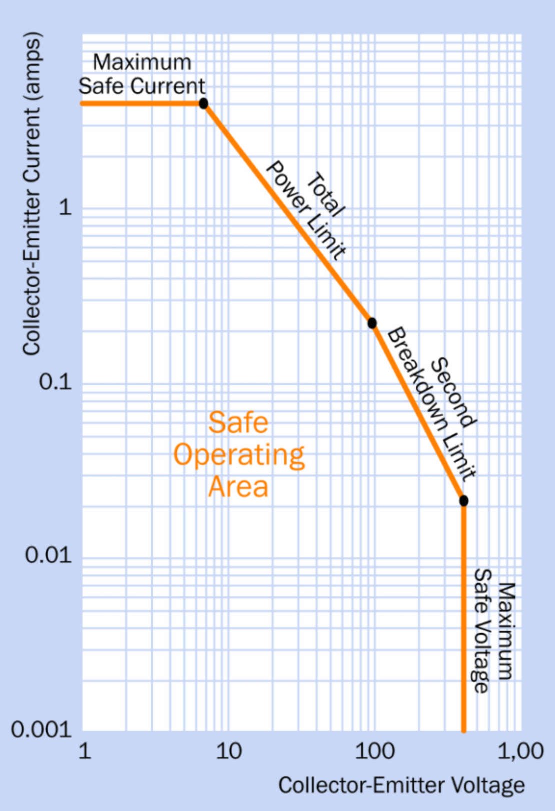

Figure 28-7. *Adapted from a Philips datasheet for a power transistor, this graph defines a safe operating area (SOA) for the component. See text for details.*

a few external capacitors. More powerful audio equipment typically uses integrated circuits to process inputs, but will use individual power transistors to handle high-wattage output.

Darlington Pairs

Discrete transistors are useful in situations where current amplification or switching is required at just one location in a circuit. An example would be where one output pin from a **microcontroller** must switch a small motor on and off. The motor may run on the same voltage as the microcontroller, but requires considerably more current than the typical 20mA maximum available from a microcontroller output. A *Darlington pair* of transistors may be used in this application. The overall gain of the pair can be 100,000 or more. See Figure 28-8. If a power source feeding through a potentiometer is substituted for the microcontroller chip, the circuit can function as a motor speed control (assuming that a generic DC motor is being used).

In the application shown here, the microcontroller chip must share a common ground (not shown) with the transistors. The optional resistor may be necessary to prevent leakage from the first transistor (when in its "off" state) from triggering the second. The diode protects the transistors from voltage transients that are likely when the motor stops and starts.

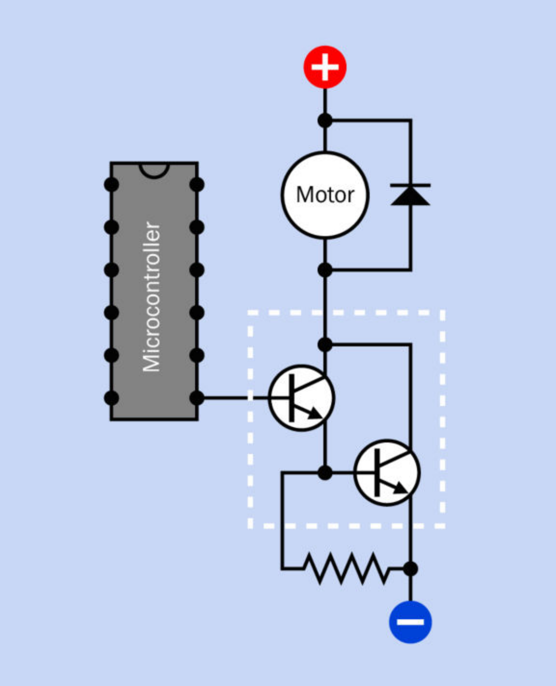

Figure 28-8. *Where the emitter of one NPN transistor is coupled to the base of another, they form a Darlington pair (identified by the dashed rectangle in this schematic). Multiplying the gain of the first transistor by the gain of the second gives the total gain of the pair.*

A Darlington pair can be obtained in a single transistor-like package, and may be represented by the schematic symbol shown in Figure 28-9.

Various through-hole Darlington packages are shown in Figure 28-10.

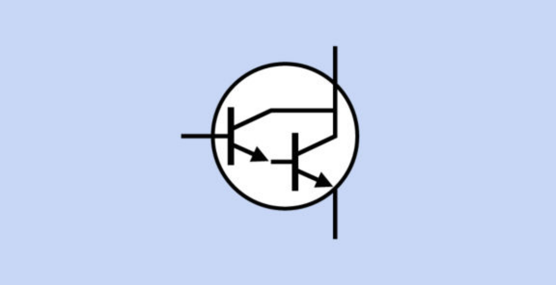

Figure 28-9. *When a Darlington pair is embedded in a single transistor-like package, it may be represented by this schematic symbol. The leads attached to the package can be used as if they are the emitter, base, and collector of a single NPN transistor.*

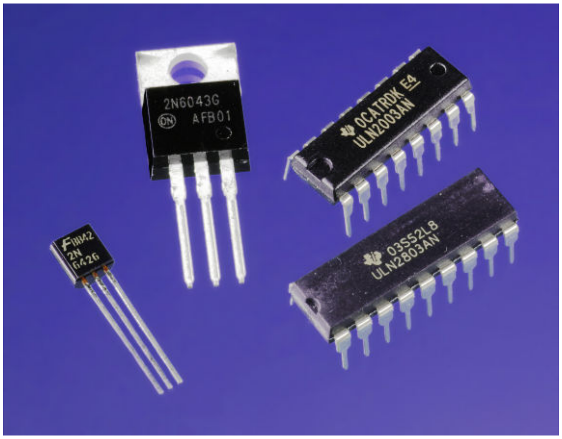

Figure 28-10. *Various packaging options for Darlington pairs. From left to right: The 2N6426 contains a Darlington pair rated to pass up to 500mA continuous collector current. The 2N6043 is rated for 8A continuous. The ULN2003 and ULN2083 chips contain seven and eight Darlington pairs, respectively.*

Seven or eight Darlington pairs can be obtained in a single integrated chip. Each transistor pair in these chips is typically rated at 500mA, but they can be connected in parallel to allow higher currents. The chip usually includes protection diodes to allow it to drive inductive loads directly.

A typical schematic is shown in Figure 28-11. In this figure, the microcontroller connections are hypothetical and do not correspond with any actual chip. The Darlington chip is a ULN2003 or similar, containing seven transistor pairs, each with an “input” pin on the left and an “output” pin opposite it on the right. Any of pins 1 through 7 down the left side of the chip can be used to control a device connected to a pin on the opposite side.

A high input can be thought of as creating a negative output, although in reality the transistors inside the chip are sinking current via an external device—a motor, in this example. The device can have its own positive supply voltage, shown here as 12VDC, but must share a common ground with the microcontroller, or with any other component which is being used on the input side. The lower-right pin of the chip shares the 12VDC supply because this pin is attached internally to clamp diodes (one for each Darlington pair), which protect against surges caused by inductive loads. For this reason, the motor does not have a clamp diode around it in the schematic.

The Darlington chip does not have a separate pin for connection with positive supply voltage, because the transistors inside it are sinking power from the devices attached to it.

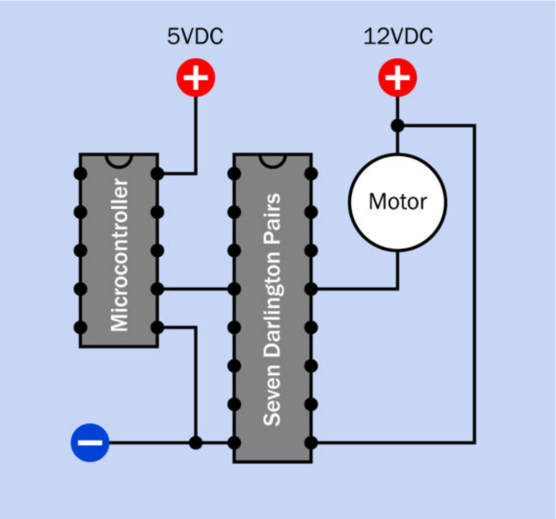

Figure 28-11. *A chip such as the ULN2003 contains seven Darlington pairs. It will sink current from the device it is driving. See text for details.*

A surface-mount Darlington pair is shown in Figure 28-12. This measures just slightly more than 0.1" long but is still rated for up to 500mA collector current or 250mW total power dissipation (at a component temperature no higher than 25 degrees Centigrade).

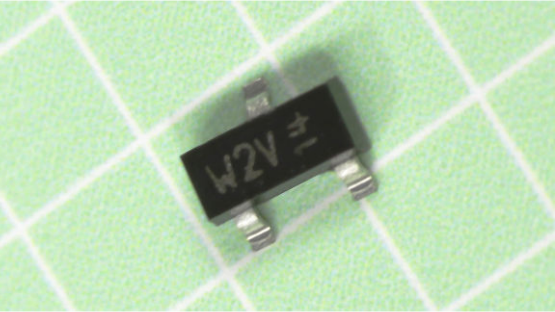

Figure 28-12. *A surface-mount package for a Darlington pair. Each square in the background grid measures 0.1". See text for additional details.*

Amplifiers

Two basic types of transistor amplifiers are shown in Figure 28-13 and Figure 28-14. The *common-collector* configuration has current gain but no voltage gain. The capacitor on the input side blocks DC current from entering the amplifier circuit, and the two resistors forming a voltage divider on the base of the transistor establish a voltage midpoint (known as the *quiescent point* or *operating point*) from which the signal to be amplified may deviate above and below.

The *common-emitter* amplifier provides voltage gain instead of current gain, but inverts the phase of the input signal. Additional discussion of amplifier design is outside the scope of this encyclopedia.

In switching applications, modern transistors have been developed to handle a lot of current compared with earlier versions, but still have some limitations. Few power transistors can handle more than 50A flowing from collector to emitter, and 1,000V is typically a maximum value. Electromechanical relays continue to exist because they retain some advantages, as shown in the table in Figure 28-15, which compares switching capabilities of transistors, **solid-state relays**, and electromechanical **relays**.

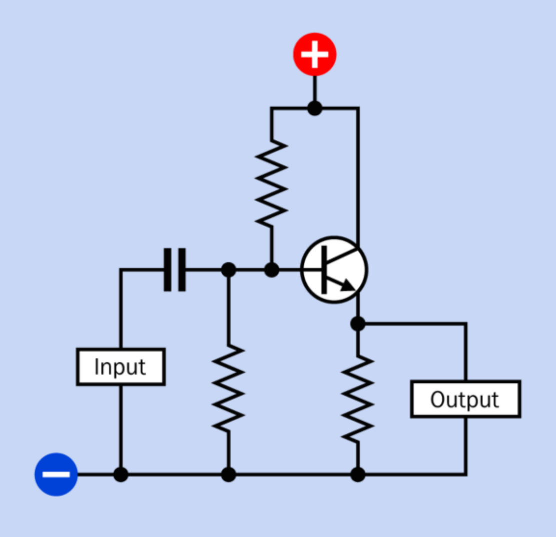

Figure 28-13. *The basic schematic for a common-collector amplifier. See text for details.*

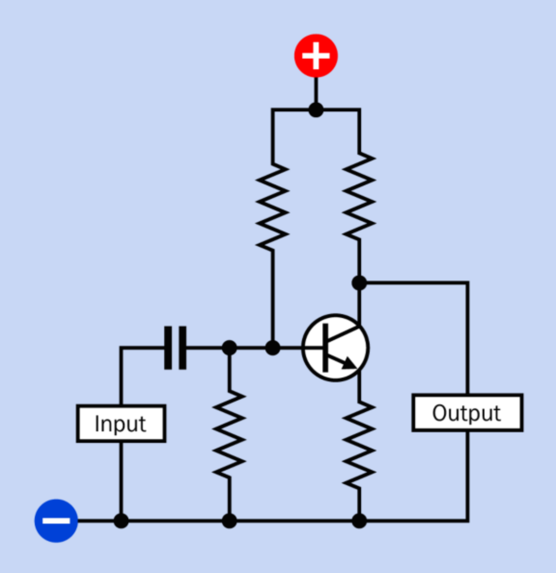

Figure 28-14. *The basic schematic for a common-emitter amplifier. See text for details.*

	Transistor	Solid-State Relay	Electro-Mechanical Relay
Long-term reliability	Excellent	Excellent	Limited
NC contacts?	No	Yes	Yes
DT contacts?	No	No	Yes
Ability to switch large current	Limited	Some	Good
Ability to switch AC	No	Yes	Yes
Can be triggered by AC voltage?	No	Yes	Yes
Suitability for miniaturization	Excellent	Poor	Poor
Vulnerable to heat	Yes	Yes	Not very
Vulnerable to corrosion	No	No	Yes
OK at high speed	Excellent	Good	Poor
Price advantage for low-voltage low-current	Yes	No	No
Price advantage for high-voltage high-current	No	No	Yes
Current leakage when "off"	Yes	Yes	No
Trigger circuit isolated from switched circuit	No	Yes	Yes

Figure 28-15. *A comparison of characteristics of switching devices.*

What Can Go Wrong

Wrong Connections on a Bipolar Transistor

Failing to identify a transistor's leads or contacts correctly can obviously be a potential source of damage, but swapping the collector and emitter accidentally will not necessarily destroy the transistor. Because of the inherent symmetry of the device, it will in fact function with collector and emitter connections reversed. Rohm, a large semiconductor manufacturer, has included this scenario in its general information pages and concludes that the primary indicator of transposed connections is that the β value, or h_{FE} , drops to about 1/10th of specification. If you are using a transistor that works but provides much less amplification than you expect, check that the emitter and collector leads are not transposed.

Wrong Connections on a Darlington Pair Chip

While a single-component package for a Darlington pair functions almost indistinguishably from a single transistor, multiple Darlington pairs in a DIP package may create confusion because the component behaves differently from most other chips, such as logic chips.

A frequent error is to ground the output device instead of applying positive power to it. See Figure 28-11 and imagine an erroneous connection of negative power instead of the 12VDC positive power.

Additional confusion may be caused by reading a manufacturer's datasheet for a Darlington pair chip such as the ULN2003. The datasheet depicts the internal function of the chip as if it contains logic inverters. While the chip can be imagined as behaving this way, in fact it contains bipolar transistors that amplify the current applied to the base of each pair. The datasheet also typically will not show the positive connection that should be made to the common-diode pin (usually at bottom-right), to provide protection from surges caused by inductive loads. This pin must be distinguished carefully from the common-ground pin (usually at bottom-left). The positive connection to the common-diode pin is optional; the common-ground connection is mandatory.

Soldering Damage

Like any semiconductor, transistors are vulnerable to heat and can be damaged while soldering, although this seldom happens if a low-wattage iron is used. A copper alligator clip can be applied as a heat sink to each lead before it is soldered.

Excessive Current or Voltage

During use, a transistor will be damaged if it is subjected to current or voltage outside of its rated range. Passing current through a transistor without any series resistance to protect it will almost certainly burn it out, and the same thing can happen if incorrect resistor values are used.

The maximum wattage that a transistor can dissipate will be shown in its datasheet. Suppose, for example, this figure is 200mW, and you are using a 12VDC supply. Ignoring the base current, the maximum collector current will be 200 / 12 = approximately 15mA. If the transistor's emitter is connected to ground, and the load applied to the transistor output has a high impedance, and if we ignore the transresistance, Ohm's Law suggests that a resistor that you place between the collector and the supply voltage should have a resistance of at least 12 / 0.015 = 800 ohms.

When transistors are used in switching applications, it is customary for the base current to be 1/5th of the collector current. In the example discussed here, a 4.7K resistor might be appropriate. A meter should be used to verify actual current and voltage values.

Excessive Leakage

In a Darlington pair, or any other configuration where the output from one transistor is connected with the base of another, leakage from the first transistor while in its "off" state can be amplified by the second transistor. If this is unacceptable, a bypass resistor can be used to divert some of the leakage from the base of the second transistor to ground. Of course the resistor will also steal some of the base current when the first transistor is active, but the resistor value is typically chosen so that it takes no more than 10% of the active current. See Figure 28-8 for an example of a bypass resistor added to a Darlington pair.

field effect transistor

29

The term **field-effect transistor** encompasses a family primarily consisting of the *junction field-effect transistor* (or *JFET*, which is the simplest generic type) and the *metal-oxide semiconductor field-effect transistor* (or *MOSFET*, also sometimes known as an *insulated-gate field-effect transistor*, or *IGFET*). Because the principles of operation overlap considerably, the entire -FET family is grouped in this entry.

OTHER RELATED COMPONENTS

- **bipolar transitor** (See Chapter 28)
- **unijunction transistor** (See Chapter 27)
- **diode** (See Chapter 26)

What It Does

A field-effect transistor creates an electric field to control current flowing through a channel in a semiconductor. MOSFETs of microscopic size form the basis of complementary metal oxide semiconductor (CMOS) *integrated circuit chips*, while large discrete MOSFETs are capable of switching substantial currents, in lamp dimmers, audio amplifiers, and motor controllers. FETs have become indispensable in computer electronics.

A **bipolar transistor** is generally thought of as a *current amplifier* because the current passing through it is controlled by a smaller amount of current passing through the base. By contrast, all FETs are considered to be *voltage amplifiers*, as the control voltage establishes field intensity, which requires little or no current. The negligible leakage through the gate of an FET makes it ideal for use in low-power applications such as portable hand-held devices.

How It Works

This section is divided into two subsections, describing the most widely used FETs: JFETs and MOSFETs.

JFETs

A *junction field-effect transistor* (or *JFET*) is the simplest form of FET. Just as a bipolar transistor can be of NPN or PNP type, a JFET can have an *N-channel* or *P-channel*, depending whether the channel that transmits current through the device is negatively or positively *doped*. A detailed explanation of semiconductor doping will be found in the **bipolar transistor** entry.

Because negative charges have greater mobility, the N-channel type allows faster switching and is more commonly used than the P-channel type. A schematic symbol for it is shown in Figure 29-1 alongside the schematic for an NPN transistor. These symbols suggest the similarity of the devices as amplifiers or switches, but it is important to remember that the FET is a primarily a voltage amplifier while the bipolar transistor is a current amplifier.

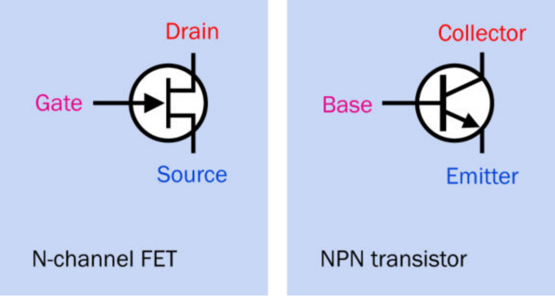

Figure 29-1. *A comparison between schematic symbols for N-channel JFET (left) and NPN bipolar transistor (right) suggests their functional similarity as switches or amplifiers, although their behavior is markedly different.*

Three JFETs are shown in Figure 29-2. The N-channel J112 type is supplied by several manufacturers, the figure showing two samples, one from Fairchild Semiconductor (left) and the other from On Semiconductor (right). Although the full part numbers are different, the specifications are almost identical, including a drain-gate voltage of 35V, a drain-source voltage of 35V, and a gate current of 50mA. The metal-clad 2N4392 in the center has similar values but is three times the price, with a much higher power dissipation of 1.8W, compared with 300mW and 350mW for the other two transistors respectively.

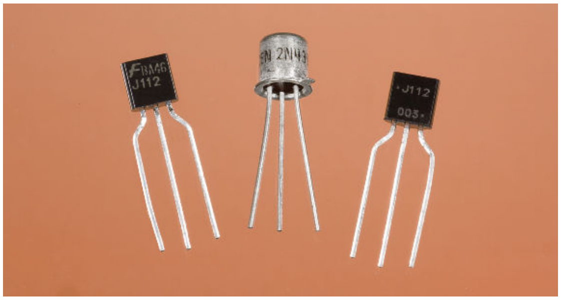

Figure 29-2. *Junction Field Effect Transistors (JFETs). See text for details.*

Schematic symbols for N-channel and P-channel JFETs are shown in Figure 29-3, N-channel being on the left while P-channel is on the right. The upper-left and lower-left symbol variants are both widely used and are functionally identical. The upper-right and lower-right variants likewise mean the same thing. Because the upper variants are symmetrical, an S should be added to clarify which terminal is the source. In practice, the S is often omitted, allowing some ambiguity. While the source and drain of some JFETs are in fact interchangeable, this does not apply to all types.

The circle around each symbol is occasionally omitted when representing discrete components, and is almost always omitted when multiple FETs are shown connected to form an integrated circuit.

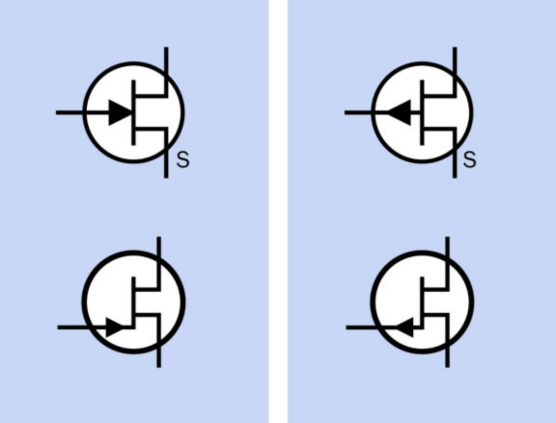

Figure 29-3. *Schematic symbols for junction field-effect transistors (JFETs). Left: N-channel. Right: P-channel. The symbols at top and bottom on each side are functionally identical. Circles may be omitted. Letter S may be omitted from the symmetrical symbol variants, even though this creates some ambiguity.*

The internal function of an N-channel JFET is shown diagrammatically in Figure 29-4. In this component, the source terminal is a source of electrons, flowing relatively freely through the N-doped channel and emerging through the drain. Thus, conventional current flows nonintuitively from the drain terminal, to the source terminal, which will be of lower potential.

The JFET is like a normally-closed switch. It has a low resistance so long as the gate is at the same potential as the source. However, if the potential of the gate is reduced below the potential of the source—that is, the gate acquires a more relatively negative voltage than the source—the current flow is *pinched off* as a result of the field created by the gate. This is suggested by the lower diagram in Figure 29-4.

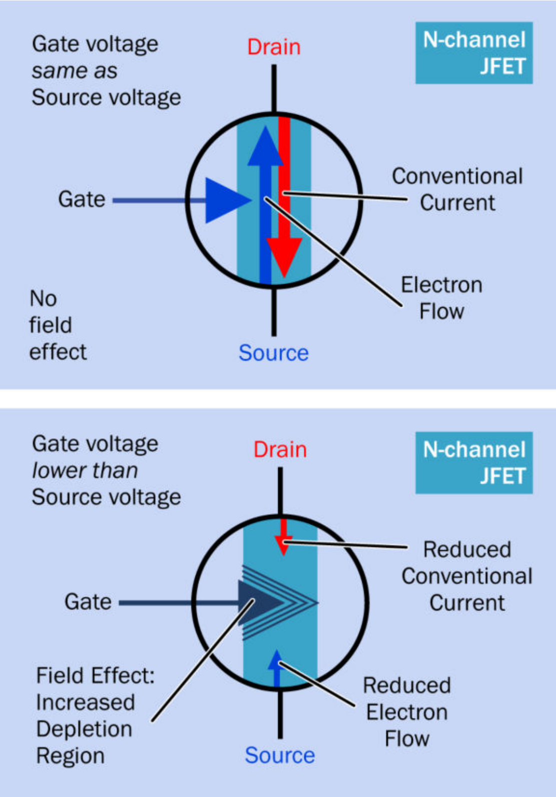
Gate voltage
same as
Source voltage
Drain
N-channel JFET
Gate
Conventional Current
Electron Flow
No field effect
Source
Gate voltage
lower than
Source voltage
Drain
N-channel JFET
Reduced Conventional Current
Gate
Field Effect: Increased Depletion Region
Reduced Electron Flow
Source

Figure 29-4. *At top, conventional current flows freely from drain to source through the channel of an N-doped JFET. At bottom, the lowered voltage of the gate relative to the source creates a field effect that pinches off the flow of current.*

The situation for a P-channel JFET is reversed, as shown in Figure 29-5. The source is now positive (but is still referred to as the source), while the drain can be grounded. Conventional current now flows freely from source to drain, so long as the gate is at the same positive potential as the source. If the gate voltage rises above the source voltage, the flow is pinched off.

A bipolar transistor tends to block current flow by default, but becomes less resistive when its base is forward-biased. Therefore it can be referred to as an *enhancement device*. By contrast, an N-channel JFET allows current to flow by default, and becomes more resistive when its base is reverse-biased, which widens the *depletion layer* at the base junction. Consequently it can be referred to as a *depletion device*.

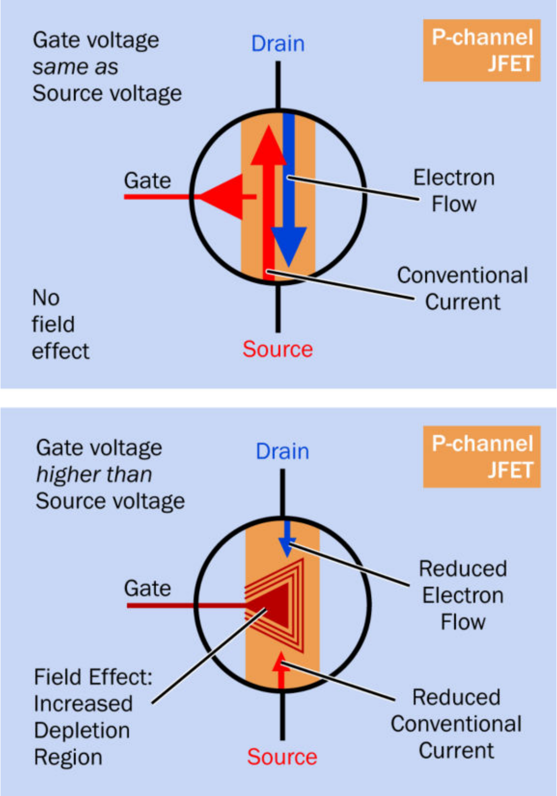
Gate voltage
same as
Source voltage
Drain
P-channel JFET
Gate
Electron Flow
Conventional Current
No field effect
Source
Gate voltage
higher than
Source voltage
Drain
P-channel JFET
Gate
Reduced Electron Flow
Field Effect: Increased Depletion Region
Reduced Conventional Current
Source

Figure 29-5. *At top, conventional current flows freely from source to drain through the channel of a P-doped JFET. At bottom, the higher voltage of the gate relative to the source creates a field effect that pinches off the flow of current.*

The primary characteristics of a junction field-effect transistor relative to an NPN bipolar transistor are summarized in the table in Figure 29-6.

JFET Behavior

The voltage difference between gate and source of a JFET is usually referred to as V_{gs} while the voltage difference between drain and source is referred to as V_{ds}.

	NPN Bipolar Transistor	N-Channel JFET
Type of amplifier	Current	Voltage
Active bias	Positive	Negative
Unbiased state	Nonconductive	Conductive
Biased state	More conductive	Less conductive

Figure 29-6. *This table contrasts the characteristics of an N-channel JFET with those of an NPN bipolar transistor.*

Suppose the gate of an N-channel JFET is connected with the source, so that $V_{gs}=0$. Now if V_{ds} increases, the current flowing through the channel of the JFET also increases, approximately linearly with V_{ds}. In other words, initially the JFET behaves like a low-value resistor in which the voltage across it, divided by the amperage flowing through it, is approximately constant. This phase of the JFET's behavior is known as its *ohmic region*. While the unbiased resistance of the channel in a JFET depends on the component type, it is generally somewhere between 10Ω and 1K.

If V_{ds} increases still further, eventually no additional flow of current occurs. At this point the channel has become *saturated*, and this plateau zone is referred to as the *saturation region*, often abbreviated I_{dss}, meaning "the saturated drain current at zero bias." Although this is a nearly constant value for any particular JFET, it may vary somewhat from one sample of a component to another, as a result of manufacturing variations.

If V_{ds} continues to increase, the component finally enters a *breakdown* state, sometimes referred to by its full formal terminology as *drain-source breakdown*. The current passing through the JFET will now be limited only by capabilities of the external power source. This breakdown state can be destructive to the component, and is comparable to the breakdown state of a typical **diode**.

What if the voltage at the gate is reduced below the voltage at the source—such as V_{gs} becomes negative? In its ohmic region, the component now behaves as if it has a higher resistance, and it will reach its saturation region at a lower current value (although around the same value for V_{ds}). Therefore, by reducing the voltage on the gate relative to the voltage at the source, the effective resistance of the component increases, and in fact it can behave as a *voltage-controlled resistor*.

The upper diagram in Figure 29-7 shows this graphically. Below it, the corresponding graph for a P-channel JFET looks almost identical, except that the current flow is reversed and is pinched off as the gate voltage rises above the source voltage. Also, the breakdown region is reached more quickly with a P-channel JFET than with an N-channel JFET.

MOSFETs

MOSFETs have become one of the most widely used components in electronics, everywhere from computer memory to high-amperage switching power supplies. The name is an acronym for *metal-oxide semiconductor field-effect transistor*. A simplified cross-section of an N-channel MOSFET is shown in Figure 29-8.

Two MOSFETs are shown in Figure 29-9.

Like a JFET, a MOSFET has three terminals, identified as drain, gate, and source, and it functions by creating a field effect that controls current flowing through a channel. (Some MOSFETS have a fourth terminal, described later). However, it has a metal source and drain making contact with each end of the channel (hence the term "metal" in its acronym) and also has a thin layer of silicon dioxide (hence the term "oxide" in its acronym) separating the gate from the channel, thus raising the impedance at the gate to at least

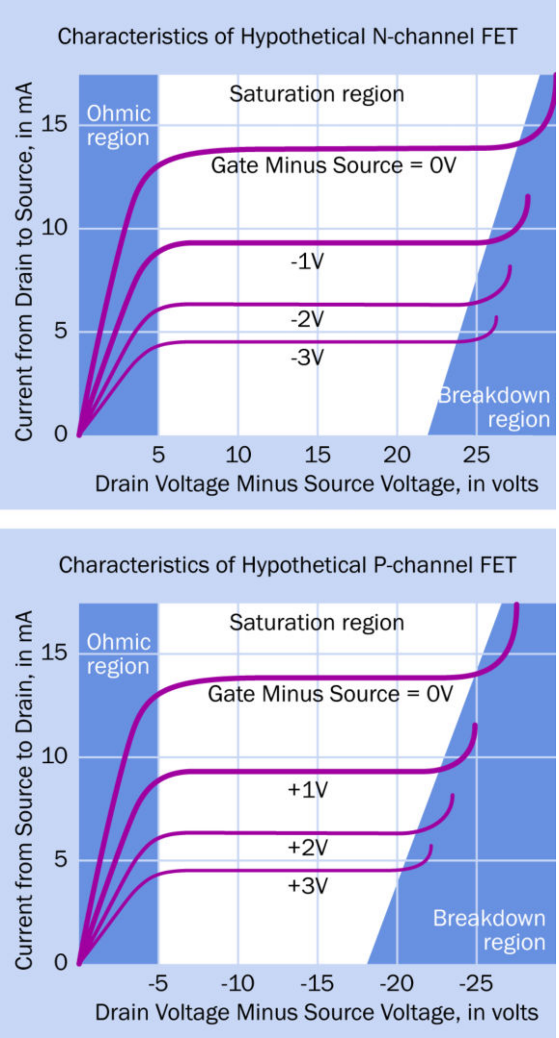

Figure 29-7. *The top graph shows current passing through the channel of an N-channel JFET, depending on gate voltage and source voltage. The lower graph is for a P-channel JFET.*

100,000 gigaohms and reducing gate current essentially to zero. The high gate impedance of a MOSFET allows it to be connected directly to the output of a digital integrated circuit. The layer of silicon dioxide is a *dielectric*, meaning that a field appled to one side creates an opposite field on the other side. The gate attached to the surface of the layer functions in the same way as one plate of a capacitor.

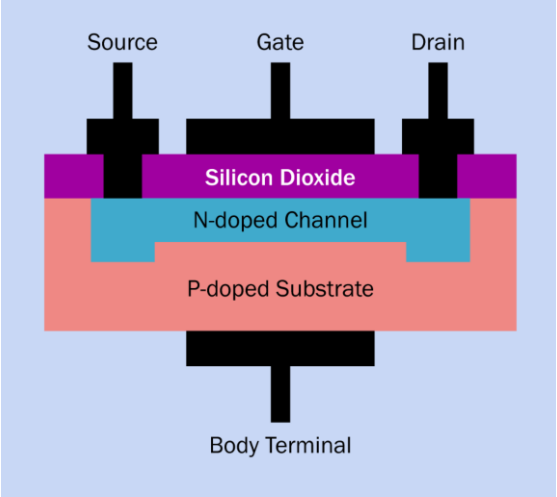

Figure 29-8. *Simplified diagram of an N-channel MOSFET. The thickness of the silicon dioxide layer has been greatly exaggerated for clarity. The black terminals are metallic.*

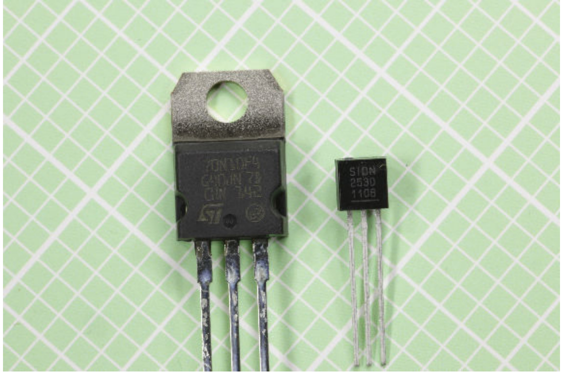

Figure 29-9. *Two MOSFETs. At left, the TO-220 package claims a drain current of up to 65A continuous, and a drain-to-source breakdown voltage 100V. At right, the smaller package offers a drain current of 175mA continuous, and a drain-to-source breakdown voltage of 300V.*

The silicon dioxide also has the highly desirable property of insulating the gate from the channel, thus preventing unwanted reverse current. In a JFET, which lacks a dielectric layer, if source voltage is allowed to rise more than about 0.6V higher than gate voltage, the direct internal connection between gate and channel allows negative charges to flow freely from source to gate, and as the internal resistance will be very low, the resulting current can be destructive. This is why the JFET must always be reverse-biased.

A MOSFET is freed from these restrictions, and the gate voltage can be higher or lower than the source voltage. This property enables an N-channel MOSFET to be designed not only as a depletion device, but alternatively as an *enhancement device*, which is "normally off" and can be switched on by being forward-biased. The primary difference is the extent to which the channel in the MOSFET is N-doped with charge carriers, and therefore will or will not conduct without some help from the gate bias.

In a depletion device, the channel conducts, but applying negative voltage to the gate can pinch off the current.

In an enhancement device, the channel does not conduct, but applying positive voltage to the gate can make it start to do so.

In either case, a shift of bias from negative to positive encourages channel conduction; the depletion and enhancement versions simply start from different points.

This is clarified in Figure 29-10. The vertical (logarithmic) scale suggests the current being conducted through the channel of the MOSFET, while the green curve describes the behavior of a depletion version of the device. Where this curve crosses the center line representing 0 volts bias, the channel is naturally conductive, like a JFET. Moving left down the curve, as reverse bias is applied (shown on the horizontal axis), the component becomes less conductive until finally its conductivity reaches zero.

Meanwhile on the same graph, the orange curve represents an enhancement MOSFET, which is nonconductive at 0 volts bias. As forward bias increases, the current also increases—similar to a bipolar transistor.

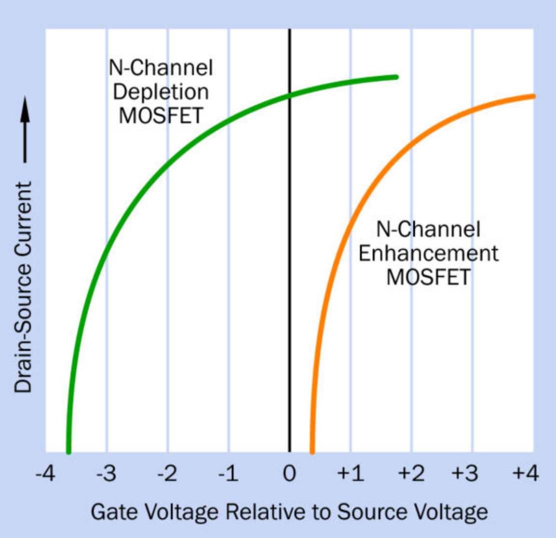

Figure 29-10. *The current conduction of depletion and enhancement N-channel MOSFETs. See text for details. (Influenced by The Art of Electronics by Horowitz and Hill.)*

To make things more confusing, a MOSFET, like a JFET, can have a P-doped channel; and once again it can function in depletion or enhancement mode. The behavior of this variant is shown in Figure 29-11. As before, the green curve shows the behavior of a depletion MOSFET, while the orange curve refers to the enhancement version. The horizontal axis now shows the voltage difference between the gate and the drain terminal. The depletion component is naturally conductive at zero bias, until the gate voltage increases above the drain voltage, pinching off the current flow. The enhancement component is not conductive until reverse bias is applied.

Figure 29-12 shows schematic symbols that represent depletion MOSFETs. The two symbols on the left are functionally identical, representing N-channel versions, while the two symbols on the right represent P-channel versions. As in the case of JFETs, the letter "S" should be (but often is not) added to the symmetrical versions of the symbols, to clarify which is the source terminal. The left-pointing arrow identifies the components as N-channel, while in the symbols on the right, the right-pointing arrows indicate P-channel MOS-

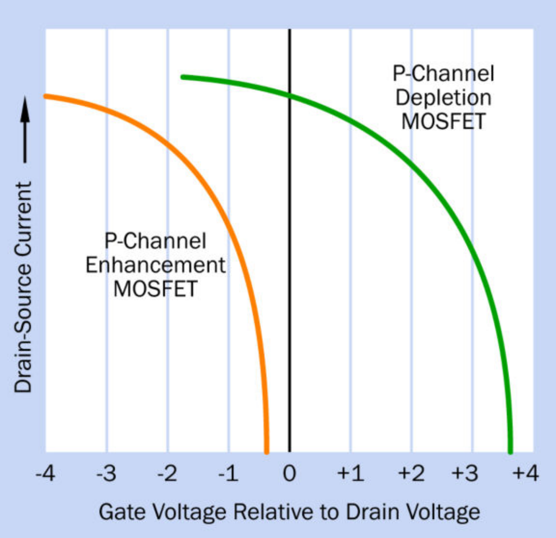

Figure 29-11. *The current conduction of depletion and enhancement P-channel MOSFETs. See text for details.*

FETs. The gap between the two vertical lines in each symbol suggests the silicon dioxide dielectric. The right-hand vertical line represents the channel.

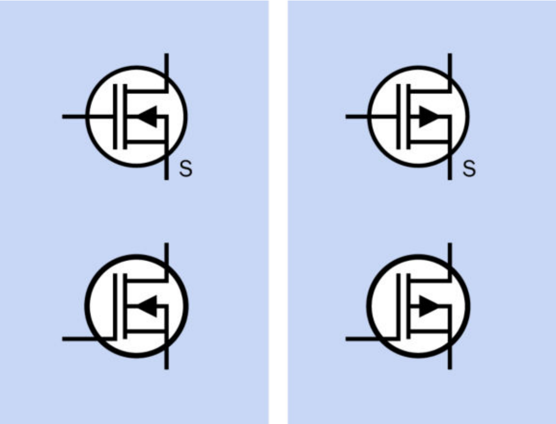

Figure 29-12. *Schematic symbols for depletion MOSFETs. These function similarly to JFETs. The two symbols on the left are functionally identical, and represent N-channel depletion MOSFETs. The two symbols on the right are both widely used to represent P-channel depletion MOSFETs.*

For enhancement MOSFETs, a slightly different symbol uses a broken line between the source and drain (as shown in Figure 29-13) to remind us that these components are "normally off" when zero-biased, instead of "normally on." Here again a left-pointing arrow represents an N-channel MOSFET, while a right-pointing arrow represents a P-channel MOSFET.

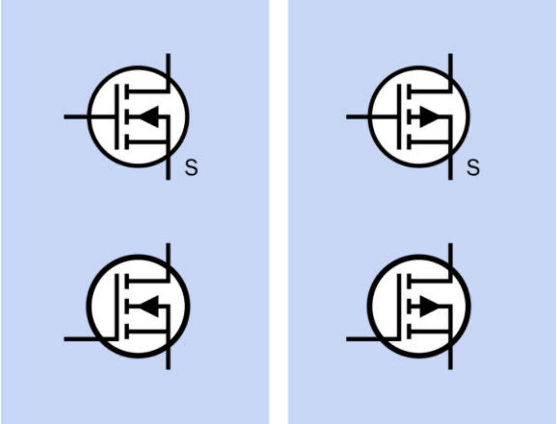

Figure 29-13. *Schematic symbols for enhancement MOSFETs. The two on the left are functionally identical, and represent N-channel enhancement MOSFETs. The two on the right represent P-channel enhancement MOSFETs.*

Because there is so much room for confusion regarding MOSFETs, a summary is presented in Figure 29-14 and Figure 29-15. In these figures, the relevant parts of each schematic symbol are shown disassembled alongside text explaining their meaning. Either of the symbols in Figure 29-14 can be superimposed on either of the symbols in Figure 29-15, to combine their functions. So, for instance, if the upper symbol in Figure 29-14 is superimposed on the lower symbol in Figure 29-15, we get an N-channel MOSFET of the enhancement type.

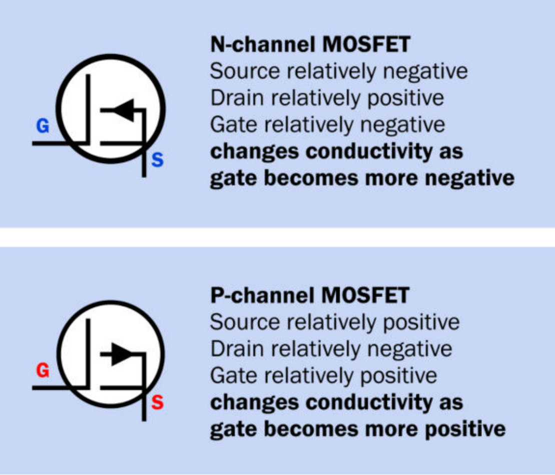
G
S
N-channel MOSFET
Source relatively negative
Drain relatively positive
Gate relatively negative
changes conductivity as gate becomes more negative
G
S
P-channel MOSFET
Source relatively positive
Drain relatively negative
Gate relatively positive
changes conductivity as gate becomes more positive

Figure 29-14. *Either of the two symbols can be combined with either of the two symbols in the next figure, to create one of the four symbols for a MOSFET. See text for details.*

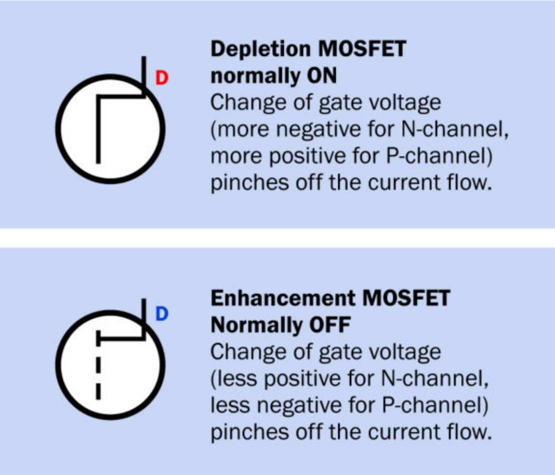
D
Depletion MOSFET normally ON
Change of gate voltage (more negative for N-channel, more positive for P-channel) pinches off the current flow.
D
Enhancement MOSFET Normally OFF
Change of gate voltage (less positive for N-channel, less negative for P-channel) pinches off the current flow.

Figure 29-15. *Either of the two symbols can be combined with either of the two symbols from the previous figure, to create one of the four symbols for a MOSFET. See text for details.*

In an additional attempt to clarify MOSFET behavior, four graphs are provided in Figure 29-16, Figure 29-17, Figure 29-18, and Figure 29-19. Like JFETs, MOSFETs have an initial ohmic region, followed by a saturation region where current flows relatively freely through the device. The gate-to-source voltage will determine how much flow is permitted. However, it is important to pay close attention to the graph scales, which differ for each of the four types of MOSFET.

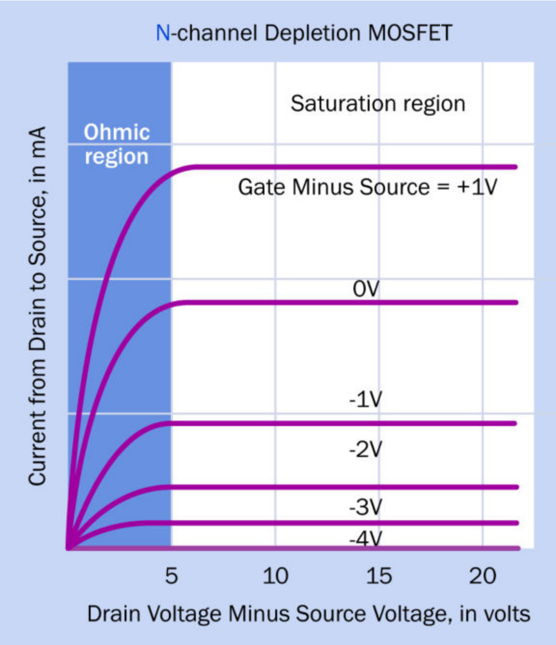

Figure 29-16. *Current flow through a depletion-type, N-channel MOSFET.*

In all of these graphs, a bias voltage exists, which allows zero current to flow (represented by the graph line superimposed on the horizontal axis). In other words, the MOSFET can operate as a switch. The actual voltages where this occurs will vary with the particular component under consideration.

The N-channel, enhancement-type MOSFET is especially useful as a switch because in its normally-off state (with zero bias) it presents a very high resistance to current flow. It requires a relatively low positive voltage at the gate, and effectively no gate current, to begin conducting conventional current from its drain terminal to its source terminal. Thus it can be driven directly by typical 5-volt logic chips.

Depletion-type MOSFETs are now less commonly used than the enhancement-type.

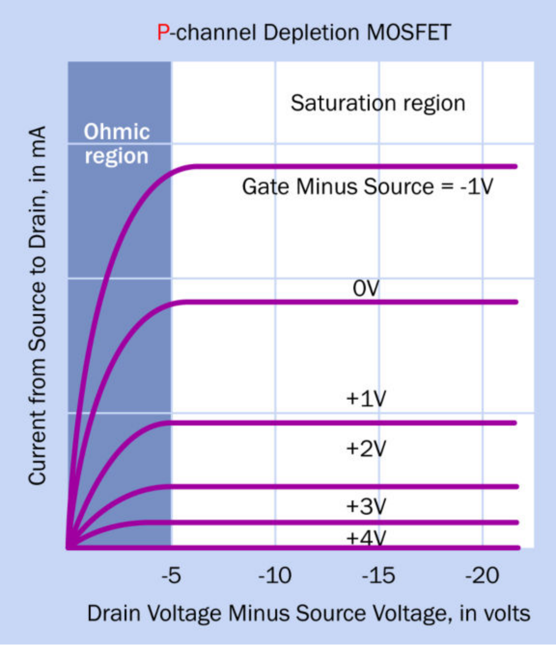

Figure 29-17. *Current flow through a depletion-type, P-channel MOSFET.*

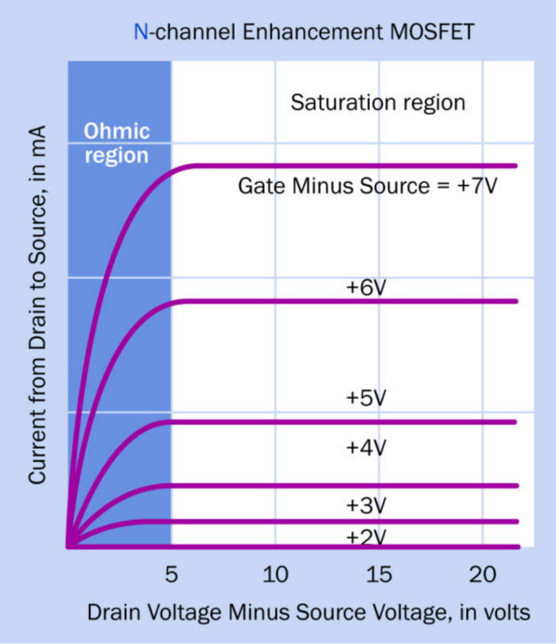

Figure 29-18. *Current flow through an enhancement-type, N-channel MOSFET.*

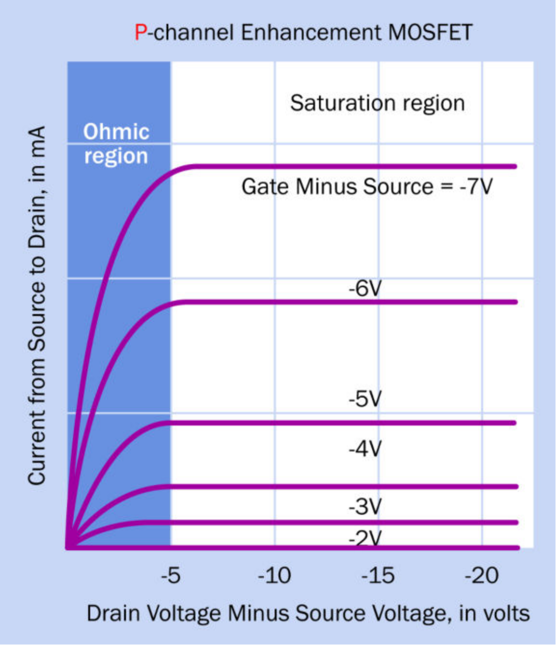

Figure 29-19. *Current flow through an enhancement-type, P-channel MOSFET.*

The Substrate Connection

Up to this point, nothing has been said about a fourth connection available on many MOSFETs, known as the *body terminal*. This is connected to the substrate on which the rest of the component is mounted, and acts as a diode junction with the channel. It is typically shorted to the source terminal, and in fact this is indicated by the schematic symbols that have been used so far. It is possible, however, to use the body terminal to shift the threshold gate voltage of the MOSFET, either by making the body terminal more negative than the source terminal (in an N-channel MOSFET) or more positive (in a P-channel MOSFET). Variants of the MOSFET schematic symbols showing the body terminal are shown in Figure 29-20 (for depletion MOSFETS) and Figure 29-21 (for enhancement MOSFETS).

A detailed discussion of the use of the body terminal to adjust characteristics of the gate is beyond the scope of this encyclopedia.

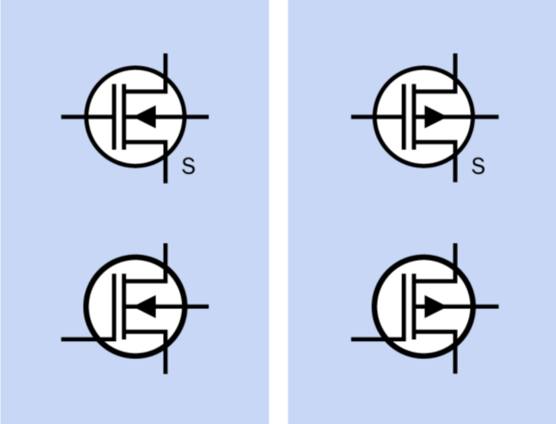

Figure 29-20. *Schematic symbol variants for depletion MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.*

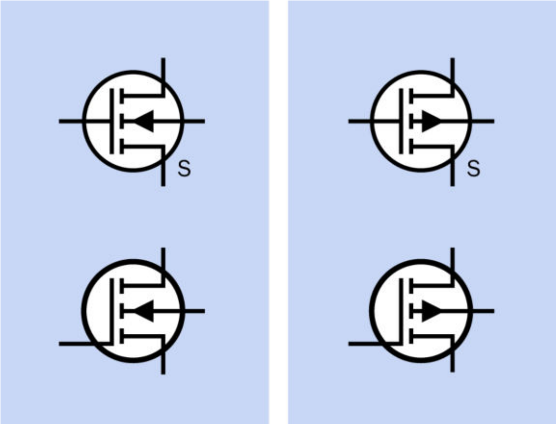

Figure 29-21. *Schematic symbol variants for enhancement MOSFETs, showing the body terminal separately accessible instead of being tied to the source terminal.*

Variants

A few FET variants exist in addition to the two previously discussed.

MESFET

The acronym stands for *MEtal-Semiconductor Field Effect Transistor*. This FET variant is fabricated from gallium arsenide and is used primarily in radio frequency amplification, which is outside the scope of this encyclopedia.

V-Channel MOSFET

Whereas most FET devices are capable of handling only small currents, the *V-channel MOSFET* (which is often abbreviated as a *VMOS FET* and has a V-shaped channel as its name implies) is capable of sustained currents of at least 50A and voltages as high as 1,000V. It is able to pass the high current because its channel resistance is well under 1Ω. These devices, commonly referred to as *power MOSFETs*, are available from all primary semiconductor manufacturers and are commonly used in switching power supplies.

Trench MOS

The TrenchMOS or Trenchgate MOS is a MOSFET variant that encourages current to flow vertically rather than horizontally, and includes other innovations that enable an even lower channel resistance, allowing high currents with minimal heat generation. This device is finding applications in the automobile industry as a replacement for electromechanical **relays**.

Values

The maximum values for JFETs, commonly found listed in datasheets, will specify V_{ds} (the drain-source voltage, meaning the potential difference between drain and source); V_{dg} (the drain-gate voltage, meaning the potential difference between drain and gate); V_{gsr} (the reverse gate-source voltage); gate current; and total device dissipation in mW. Note that the voltage differences are relative, not absolute. Thus a voltage of 50V on the drain and 25V on the source might be acceptable in a component with a V_{ds} of 25V. Similarly, while a JFET's "pinch-off" effect begins as the gate becomes "more negative" than the source, this can be achieved if, for example, the source has a potential of 6V and the gate has a potential of 3V.

JFETs and MOSFETs designed for low-current switching applications have a typical channel resistance of just a few ohms, and a maximum switching speed around 10Mhz.

The datasheet for a MOSFET will typically include values such as gate threshold voltage, which may be abbreviated V_{gs} (or V_{th}) and establishes the relative voltage at which the gate starts to play an active role; and the maximum on-state drain current, which may be abbreviated $I_{d(on)}$ and establishes the maximum limiting current (usually at 25 degree Centigrade) between source and gate.

How to Use it

The combination of a very high gate impedance, very low noise, very low quiescent power consumption in its off state, and very fast switching capability makes the MOSFET suitable for many applications.

P-Channel Disadvantage

P-channel MOSFETs are generally less popular than N-channel MOSFETS because of the higher resistivity of P-type silicon, resulting from its lower carrier mobility, putting it at a relative disadvantage.

Bipolar Substitution

In many instances, an appropriate enhancement-type MOSFET can be substituted for a bipolar transistor with better results (lower noise, faster action, much higher impedance, and probably less power dissipation).

Amplifier Front Ends

While MOSFETs are well-suited for use in the front end of an audio amplifier, chips containing MOSFETs are now available for this specific purpose.

Voltage-Controlled Resistor

A simple voltage-controlled resistor can be built around a JFET or MOSFET, so long as its performance remains limited to the linear or ohmic region.

Compatibility with Digital Devices

A JFET may commonly use power supplies in the range of 25VDC. However, it can accept the high/low output from a 5V digital device to control its gate. A 4.7K pullup resistor is an appropriate value to be used if the FET is to be used in conjunction with a TTL digital chip that may have a voltage swing of only approximately 2.5V between its low and high thresholds.

What Can Go Wrong

Static Electricity

Because the gate of a MOSFET is insulated from the rest of the component, and functions much like a plate of a capacitor, it is especially likely to accumulate static electricity. This static charge may then discharge itself into the body of the component, destroying it. A MOSFET is particularly vulnerable to electrostatic discharge because its oxide layer is so thin. Special care should be taken either when handling the component, or when it is in use. Always touch a grounded object or wear a grounded wrist band when handling MOSFETs, and be sure that any circuit using MOSFETs includes appropriate protection from static and voltage spikes.

A MOSFET should not be inserted or removed while the circuit in which it performs is switched on or contains residual voltage from undischarged capacitors.

Heat

Failure because of overheating is of special concern when using power MOSFETs. A Vishay Application Note (“Current Power Rating of Power Semiconductors”) suggests that this kind of component is unlikely to operate at less than 90 degrees Centigrade in real-world conditions, yet the power handling capability listed in a datasheet usually assumes an industry standard of 25 degrees Centigrade.

On the other hand, ratings for continuous power are of little relevance to switching devices that have duty cycles well below 100%. Other factors also play a part, such as the possibility of power surges, the switching frequency, and the integrity of the connection between the component and its heat sink. The heat sink itself creates uncertainty by tending to average the temperature of the component, and of course there is no simple way to know the actual junction temperature, moment by moment, inside a MOSFET.

Bearing in mind the accumulation of unknown factors, power MOSFETs should be chosen on an extremely conservative basis. According to a tutorial in the *EE Times*, actual current switched by a MOSFET should be less than half of its rated current at 25 degrees, while one-fourth to one-third are common. Figure 29-22 shows the real-world recommended maximum drain current at various temperatures. Exceeding this recommendation can create additional heat, which cannot be dissipated, leading to further accumulation of heat, and a thermal runaway condition, causing eventual failure of the component.

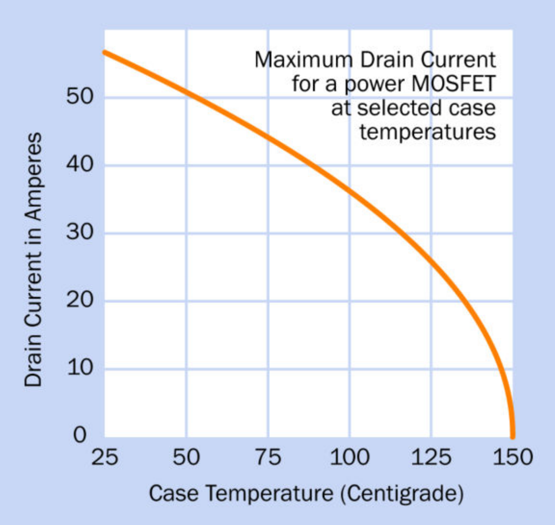

Figure 29-22. *Maximum advised drain current through a power MOSFET, related to case temperature of the component. Derived from EE Times Power MOSFET Tutorial.*

Wrong Bias

As previously noted, applying forward bias to a JFET can result in the junction between the gate and the source starting to behave like a forward-biased diode, when the voltage at the gate is greater than the voltage at the source by approximately 0.6V or more (in an N-channel JFET). The junction will present relatively little resistance, encouraging excessive current and destructive consequences. It is important to design devices that allow user input in such a way that user error can never result in this eventuality.

Schematic Symbols

A

This section contains a compilation of schematic symbols for components that have been described in this volume. They are sequenced primarily in alphabetical order, as this section is intended for use as an index. However, symbols that have a strong similarity are grouped together; thus **potentiometer** is found adjacent to **resistor**, and all types of transistors are in the same group.

The symbol variants shown in each blue rectangle are functionally identical.

Where a component has mandatory polarity or is commonly used with a certain polarity, a red plus sign (+) has been added for guidance. This sign is not part of the symbol. In the case of polarized capacitors, where a plus sign is normally shown (or should be shown) with the symbol, the plus sign is a part of the symbol and appears in black.

This is not intended to be an exhaustive compilation of symbol variants. Some uncommon ones may not be here. However, the list should be sufficient to enable identification of components in this volume.

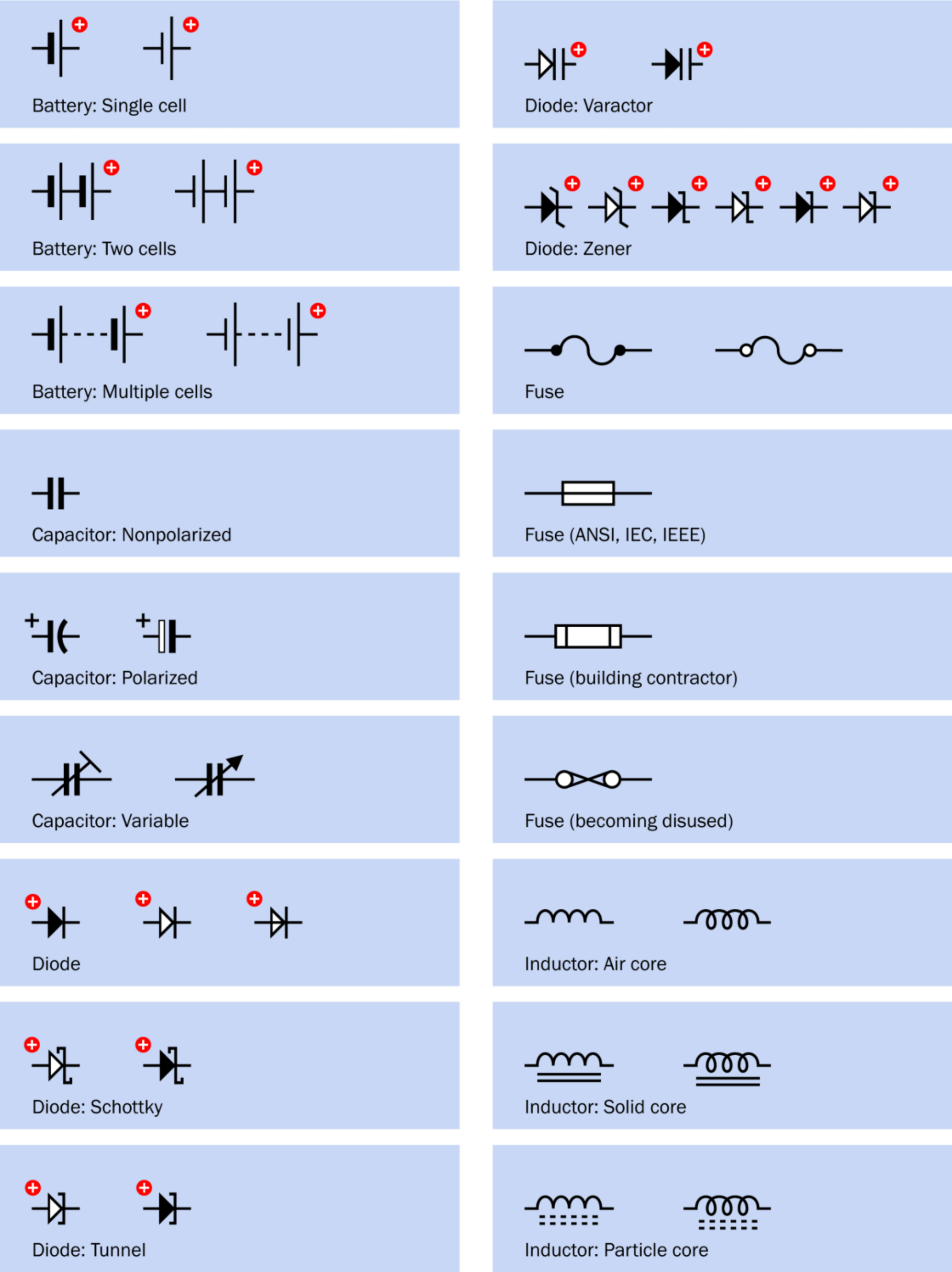
Battery: Single cell
Diode: Varactor
Battery: Two cells
Diode: Zener
Battery: Multiple cells
Fuse
Capacitor: Nonpolarized
Fuse (ANSI, IEC, IEEE)
+
Capacitor: Polarized
Fuse (building contractor)
Capacitor: Variable
Fuse (becoming disused)
Diode
Inductor: Air core
Diode: Schottky
Inductor: Solid core
Diode: Tunnel
Inductor: Particle core

Figure A-1. *Schematic symbols*

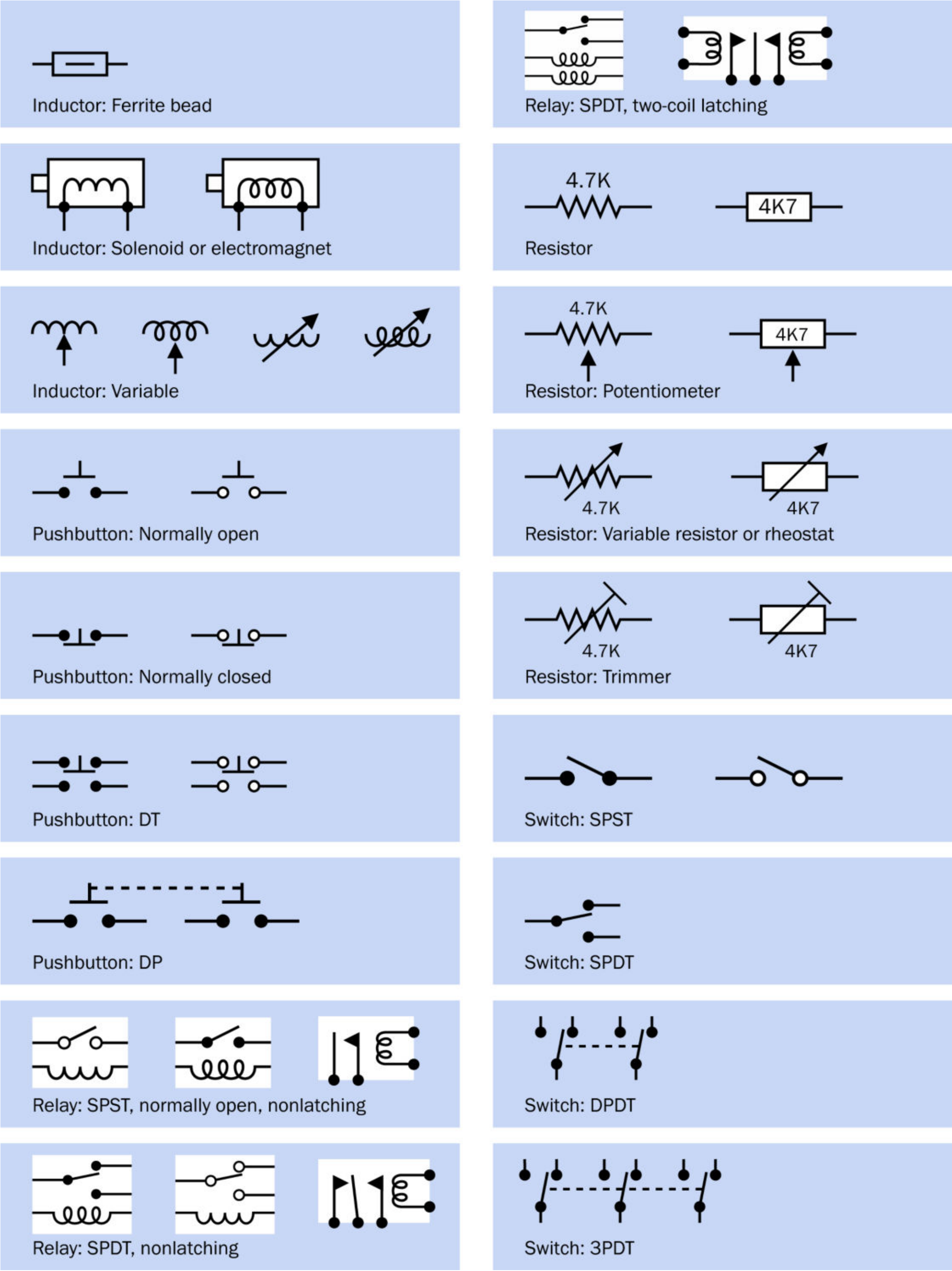

Inductor: Ferrite bead

Inductor: Solenoid or electromagnet

Inductor: Variable

Pushbutton: Normally open

Pushbutton: Normally closed

Pushbutton: DT

Pushbutton: DP

Relay: SPST, normally open, nonlatching

Relay: SPDT, nonlatching

Relay: SPDT, two-coil latching

Resistor

Resistor: Potentiometer

Resistor: Variable resistor or rheostat

Resistor: Trimmer

Switch: SPST

Switch: SPDT

Switch: DPDT

Switch: 3PDT

Figure A-2. *Schematic symbols, continued*

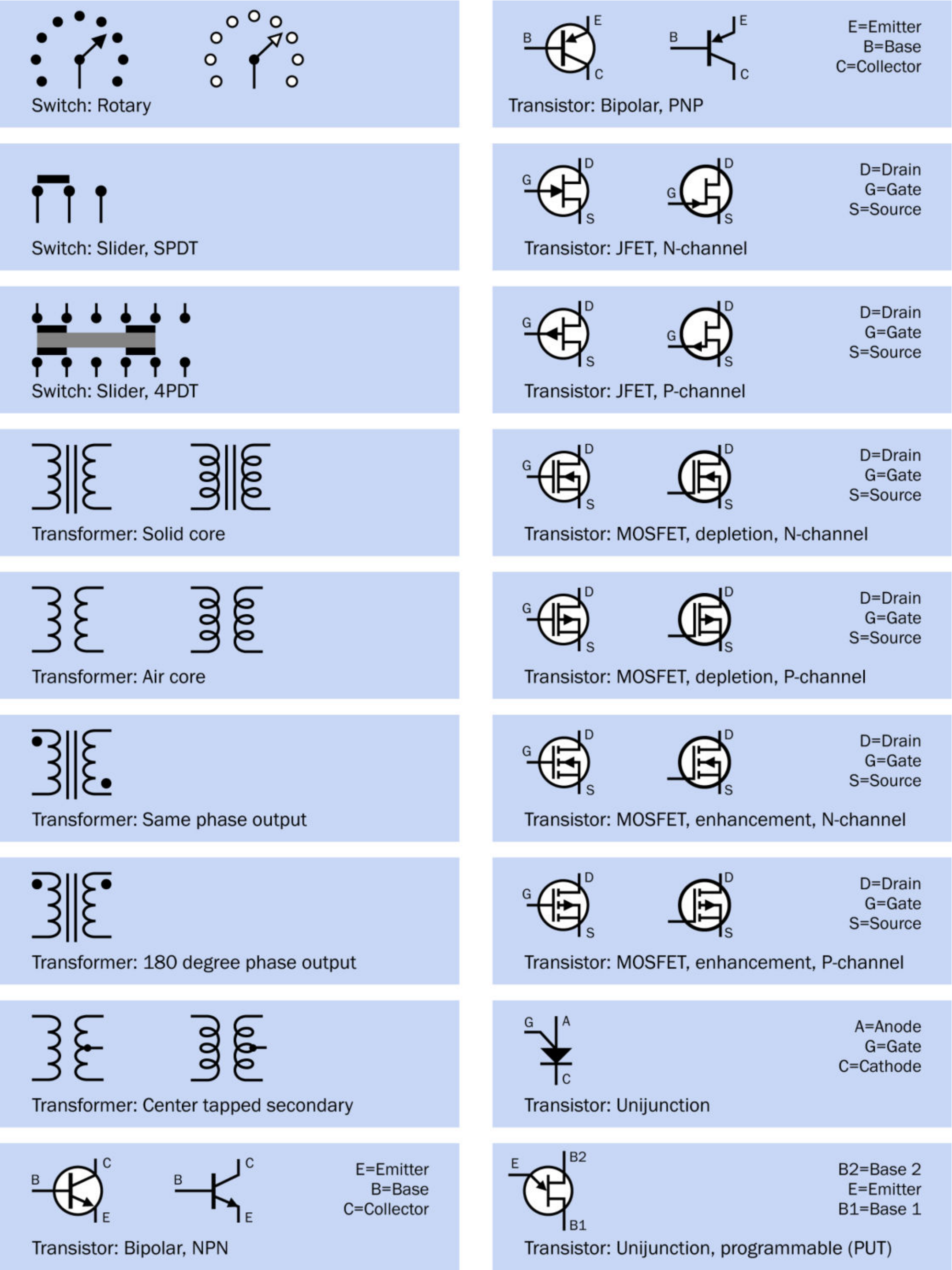
Switch: Rotary

Switch: Slider, SPDT

Switch: Slider, 4PDT

Transformer: Solid core

Transformer: Air core

Transformer: Same phase output

Transformer: 180 degree phase output

Transformer: Center tapped secondary

Transistor: Bipolar, NPN

Transistor: Bipolar, PNP

Transistor: JFET, N-channel

Transistor: JFET, P-channel

Transistor: MOSFET, depletion, N-channel

Transistor: MOSFET, depletion, P-channel

Transistor: MOSFET, enhancement, N-channel

Transistor: MOSFET, enhancement, P-channel

Transistor: Unijunction

Transistor: Unijunction, programmable (PUT)

B
E
C
E=Emitter
B=Base
C=Collector
D
G
S
D=Drain
G=Gate
S=Source
A
A=Anode
G=Gate
C=Cathode
B2
B1
B2=Base 2
E=Emitter
B1=Base 1

Figure A-3. *Schematic symbols, continued*

Index

Symbols

18 * D) + (40 * W, 128
1:1 transformers (see isolation transformers)

A

absolute encoding, 53
AC adapters, 142, 143, 157, 159
AC inrush, 177
AC motors, 191–200
  design of, 191
  function of, 191
  potential problems, 200
  types of, 195
  use of, 199
  values for, 199
AC-AC transformers, 135–142
  design of, 136
  function of, 135
  potential problems, 142
  schematic symbol for, 135
  types of, 138
  use of, 142
  values for, 141
AC-DC power supplies, 143–147
  function of, 143
  potential problems, 147
  types of, 143
  use of, 147
accumulators, 5
active mode, 245
actuators, 38
adjustable inductors, 125
adjustable output converters, 152
adjustable voltage regulators, 163
Ah/AH/A/H (see amp-hours)
air cores, 125
alkaline cells, 8, 13
alternate pushbuttons, 33
alternating current (AC), 135, 143, 157, 196, 221
amp-hour capacity, 11, 13,
amp-hours (Ah), 11
amplifiers, 250
amplitude-modulated, 125
analog servo motors, 204
annulus, 183
anodes, 6, 221
anti-backlash gears, 114
apparent power, 141
arcing, 47, 72, 108
armatures, 66
audio tone control, 85
audio transformers, 140
audio-taper potentiometers, 91, 96
automotive fuses, 23
automotive relays, 69
autotransformers, 140
auxilliary winding, 196
avalanche breakdown, 233
axial leads, 99
axial loading, 185
axial resistors, 76

B

back EMF, 108, 218, 227, 228
backlash, 188, 204
bandpass filters, 131
basic switches, 39
batteries, 5–16
  capacitors as, 109
  capacity, 11
  DC-AC inverters and, 160

We'd like to hear your suggestions for improving our indexes. Send email to index@oreilly.com.

  design of, 6
  disposable, 7, 8
  fuel cells, 8
  function of, 5
  potential problems, 15
  rechargeable, 7, 9
  schematic symbols for, 6
  use of, 14
  values for, 11
  voltage, 13
battery acid, 9
battery chargers, 8
bearings
  failure of, 189
beta value (β), 244
biased relays, 67
biased semiconductors, 241
bifilar motors, 216
bipolar electrolytic capacitors, 102
bipolar junction transistor (BJT) (see bipolar transistors)
bipolar motors, 211
bipolar stepper motors, 213
bipolar transistors, 241–252
  design of, 241
  function of, 241
  potential problems, 251
  schematic symbols for, 242
  types of, 245
  use of, 246
BLDC motors (see brushless DC motors)
blown fuses, 21, 27
body terminals, 261
boost-type converters, 151, 154
boots, 43
break-before-make switches, 39, 53
breakdown state, 256
breakdown voltage, 224
breakers, 160
breaking capacity, 22
bridge rectifiers, 226
brushes, 179, 187, 198
brushless DC motors, 183
buck converters, 150
buck-boost converters (see flyback converters)
bulk metal foil, 83
bushings, 38
button capacitors (see disk capacitors)
button cells, 5
bypass capacitors, 107, 154
bypass switches, 176

C

c-clamps, 110
C-rate, 12
capacitance, 105, 115
capacitive reactance, 106
capacitor arrays, 101
capacitor clamps, 110
capacitor networks, 101
capacitors, 8, 97–111, 97
  (see also variable capacitors)
  as alternative to shorting coils, 195
  design of, 97
  function of, 97
  potential problems, 109
  schematic symbols, 97
  types of, 99
  use of, 107
  values for, 104
capacity, 11–14
caps (see capacitors)
carbon-composite resistors, 82
carbon-film potentiometers, 95
carbon-film resistors, 82
carriers, 183
cartridge fuses, 23
cathode-ray tubes, 171
cathodes, 6, 221
cells, 5, 6
ceramic capacitors, 103
charging
  incorrect voltage and, 15
  reverse charging, 16
chatter, 72
chokes, 119
chopper drives, 217
circuit breakers, 15, 21
circuits
  RC circuits, 79
clamp diodes, 230
closed-loop systems, 210, 217
CMOS (complementary metal oxide semiconductor), 253
code samples
  digital access to, 3
coded rotary switches, 53
cogging, 193
coils, 66, 121, 135, 170
cold cranking rating, 159
collectors, 242, 246
common contacts, 38
common-bus configurations, 78
common-collector amplifiers, 250
common-emitter amplifier, 250
commutators, 179, 195, 198
complement-coded rotary switches, 54,
complementary metal oxide semiconductor (see CMOS)
components
  definition of, 2
condensers (see capacitors)
contact bounce, 61
contacts
  burnout of, 63
  contamination of, 48, 73
  overloading of, 57
  relays and, 66
  vulnerability of, 57
continuous inductors, 122
contractors, 70
control shaft encoders, 59
conventional current, 7
  inductors and, 120
converters (see DC-DC converters)
cores, 119, 132, 137, 138
coupling capacitors, 107
covered power supplies, 146
crossover networks, 130
crystal diodes (see diodes)
crystal sets, 125
cup (potentiometer), 93
current, 5
  inadequate, 15
  stall current, 184
current amplifiers, 243, 253
current gain (hFE), 244

current ratings
  DC motors, 184
  fuses, 22
  pushbuttons, 35
  relays, 70
  rotary switches, 56
  rotational encoders, 61
cutoff mode, 245
cylindrical capacitors, 99

D

damping, 220
Darlington chips, 249
Darlington pair chips, 251
Darlington pairs, 161, 218, 241, 248
DC current
  AC-AC tranformers and, 142
  inductors and, 121
DC motors, 179–189
  current rating, 184
  design of, 179
  function of, 179
  potential problems, 187
  types of, 181
  use of, 185
DC-AC inverters, 157–160
  design of, 157
  function of, 157
  potential problems, 160
  types of, 158
  use of, 159
  values for, 158
DC-DC converters, 149–155
  design of, 149
  function of, 149
  potential problems, 155
  types of, 150
  use of, 154
  values for, 152
DCR (DC resistance), 129
debounce, 48
decks, 52
decoupling capacitors, 107
deep cycle batteries, 9
degauss, 171
delay fuses, 23
delta configuration, 181
demagnetization
  of rotors, 220
demagnetizing objects, 171
depletion device, 255
depletion layer, 255
depletion region, 223
detent torque, 220
detents, 59
dielectric absorption, 110
dielectric constant, 105
dielectric memory, 110
dielectrics, 97, 103, 113, 257
digital access
  code samples, 3
digital servo motors, 204
diode arrays, 226
diodes, 221–233
  DC circuits and, 47
  design of, 223
  function of, 221
  potential problems, 232
  schematic symbol for, 221
  types of, 224
  use of, 227
  values for, 226
DIP (see dual-inline package)
DIP switches, 43, 53
  (see also rotary DIP switches)
  vs. jumpers, 17
direct current (DC), 143, 157
direct current-excited motors, 196
discrete semiconductors, 241
disk capacitors, 100
disposable batteries, 7, 8–9
doping, 253
double-pole pushbuttons, 29
double-pole switches, 37
double-throw (DT) switches, 37
drain-source breakdown, 256
dropout voltage, 161
dry cells, 7
dry joints, 48
dual-coil latching relays, 67
dual-ganged potentiometers, 93
dual-inline package (DIP), 77
dynamic resistance, 224, 225
dyne-meters, 185

E

effective series resistance (ESR), 98, 154
efficiency
  converters and, 153
8-wire motors, 216
electrical capacity (see capacity)
electrical noise, 87, 97, 132, 147, 155, 207, 230
electrical polarity, 123
electrochemical cells, 5
electrochemical power sources, 5
electrodes, 6, 7
electrolysis, 5
electrolytes, 6
electrolytic capacitors, 101, 110, 147
electrolytic cells, 5
electrolytic layer dielectrics, 103
electromagnetic armature relays, 65
electromagnetic interference (EMI), 147, 154, 187
electromagnets, 169–172
  function of, 169, 169
  potential problems, 172
  types of, 170
  use of, 171
  values for, 171
electromechanical linear actuators, 184
electromechanical relays, 65
electromotive force (EMF), 177, 218, 227, 228
electron holes, 98, 223
electronic commutation, 184
electrostatic discharge (ESD), 225, 263
element, 21
emergency switches, 35
EMI (see electromagnetic interference)
emitters, 242, 246
enable pins, 165
encoded output rotary switches, 53
enhancement devices, 255, 258

epicyclic gears, 182
equivalent series resistance (ESR), 110
errata, web address for, 4
ESD (see electrostatic discharge)

F

farads (F), 104
ferrite beads, 126
ferromagnetic cores, 137
field effect transistor (FET)
  design of, 253
  function of, 253
  potential problems, 263
  types of, 262
  use of, 263
  values for, 262
field windings, 192
field-effect transistor (FET), 253–264
555 timer, 186, 202, 214, 218, 235
flat-ganged potentiometers, 93
flip-flop behavior, 67
flooded design, 9
flux density, 170
flyback converters, 151
flyback diodes, 47
flyback-type converters, 144
foot pedal pushbuttons, 33
FORM A/B switches, 39
Form A/B/C pushbuttons, 30
formulae
  presentation of, 3
forward bias, 241
forward direction (diodes), 221
forward EMF, 177
frames, 173
freewheeling diodes, 47
front panel mount switches, 45
fuel cells, 8
fuseholders, 27
fuses, 15, 21–28
  (see also circuit breakers)
  current rating, 22
  design of, 21
  function of, 21
  potential problems, 27
  schematic symbols for, 21
  types of, 22
  use of, 26, 28
  voltage rating, 22
fusible links, 24
fusible resistors, 82

G

ganged potentiometers, 93
gear motors (see gearhead motors)
gearhead motors, 181, 188
gears
  nylon vs. metal, 204
general-purpose resistors, 76
generators, 120
gigaohms, 79
Gray-coded outputs, 55
grounding
  common ground, 142
  variable capacitors and, 117
Gunn diodes, 221, 226
gyrators, 127, 132

H

H bridge system, 187
half-cells, 6
half-stepping mode, 217
Hall effect sensors, 184
harmonics, 157
header sockets, 18
headers, 18, 34
heat
  AC motors and, 200
  capacitors and, 110
  converters and, 155
  diodes and, 233
  effect on motors, 188
  power resistors and, 84
  resistors and, 87
  solenoids and, 177
  voltage regulators and, 166
heat sinks, 76, 246
Henry units, 128
high phase count motors, 216
high side, 46
high-frequency transistors, 245
high-pass filters, 107, 130
high-speed diodes, 224
hold current, 25
holding torque, 220
horse-shoe magnet design, 170
hot switching, 37
hunting, 220
hybrid motors, 216
hysteresis, 122, 124, 170, 220
hysteresis motors, 197

I

IGBT (see insulated-gate bipolar transistor)
illuminated pushbuttons, 31
impedance, 140
in-line fuse holders, 24,
inaccuracy
  resistors and, 87
incandescent bulbs, 31
incremental devices
  rotational encoders, 59
indicators, 84
inductance, 87, 128
inductance index, 129
inductance-capacitance circuits (see LC circuits)
induction motors, 210
induction rotors, 192
inductive loads
  switch choices for, 48
  voltage spikes and, 71
inductive reactance, 98, 106
inductors, 119–133
  design of, 120
  function of, 119
  potential problems, 132
  schematic symbol for, 119
  types of, 124
  use of, 130
  values for, 128
inrunner configuration, 184
insulated-gate bipolar transistor (IGBT), 186

insulated-gate field-effect transistor (IGFET), 253
integrated circuits, 241
internal resistance, 11
inverted AC motors, 199
ions, 6
ISAT (saturation current, 129
isolated vs. non-isolated, 154
isolation transformers, 140

J

jumper assemblies, 18
jumper sockets (see jumpers)
jumper wires, 17
jumpers, 17–19
junction field-effect transistor (JFET), 253, 253–256
junction threshold voltage, 223, 236

K

Karbonite gears, 204
keylock switches, 56
keypad pushbuttons, 34
kilohms, 79
knife switches, 37

L

latching pushbuttons, 33
latching relays, 46, 67
LC (inductance-capacitance) circuits, 115
lead-acid batteries, 5, 9, 15
leakage, 98, 223, 241, 252
LED (light-emitting diode), 36, 221
  pushbuttons and, 31
  resistors and, 84
limit switches, 39, 184, 187
linear actuators, 184, 186, 218
linear mode, 245
linear motors, 195
linear regulated power supplies, 143
linear taper potentiometers, 90
linear voltage regulators (see voltage regulators)
lithium-ion batteries (Li-ion), 9, 15
lithium-ion polymer batteries, 9
load regulation, 153
loads, 5
lock paint, 117
locked-rotor torque, 194
locking pushbuttons, 33
locking toggle switches, 43
log-taper potentiometers, 91, 96
logarithmic volume control, 95,
logic chips, 48, 154
logic circuits, 47, 157
logic gates, 230
logic inverters, 157
low-dropout voltage regulators, 161, 164
low-pass filters, 108, 130
low-signal relays, 69
lubricants
  environmental effects on, 188

M

magnetic circuits, 121
magnetic cores, 122, 124
magnetic domains, 122, 132, 170
magnetic fields, 72
magnetic polarity, 123
magnetic resistance (see reluctance)
magnetic saturation, 122
mAh (see milliamp-hours)
MAKE magazine
  contact information, 4
make-before-break switches, 39, 53
make-to-break connections, 30, 39
make-to-make connections, 30
matrix encoded pushbuttons, 34
mechanical encoders, 55
mechanical rotary encoders, 59
mechanical wear
  of switches, 48
  potentiometers, 95
megohms, 79
membrane pad pushbuttons, 34
membrane separators, 6
memory effect, 10
MESFET (MEtal-Semiconductor Field Effect Transistor), 262
metal-film resistors, 83
metal-oxide semiconductor field-effect transistor (see MOSFET)
meter loading, 88
microcontrollers
  as alternative to switches, 47
  as pulse sources, 186
  increasing output of, 248
  keypads and, 34
  pushbuttons and, 33
  rotary switches and, 56
  rotational encoders and, 59, 62
  servo motors and, 202
microfarads (µF), 104, 111
microhenrys (µH), 128
microstepping mode, 217
microswitches, 39
milliamp-hours (mAh), 11
millihenrys (mH), 128
milliohms, 79
miniaturized potentiometers, 92
modified sine waves, 158
monetary switches, 29, 67
MOSFET (metal-oxide semiconductor field-effect transistor), 149, 186, 253, 256–261, 263
movable contacts, 38
multiphase motors, 217
multiple-turn potentiometers, 93
Must Operate By voltage, 70
Must Release By voltage, 70
mutual induction, 136

N

N layer, 223
N-channel, 253
N-type semiconductors, 241
nanofarads (nF), 104

nanohenrys (nH), 128
negative resistance region, 237
negative voltage regulators, 164
neon bulbs, 31
neutral relays, 67
Newton-meters, 185
nichrome (Ni-chrome) elements, 77
nickel-cadmium batteries (nicad/NiCd), 9, 10
nickel-metal hydride batteries (NiMH), 9, 10
noise
  electrical, 87, 97, 132, 147, 155, 207
  reduction of, 189, 230
non-excited synchronous motors, 196
non-flammable resistors, 82
nonlatching relays, 67
nonmagnetic cores, 125
nonshorting switches, 53, 57
normally closed (NC), 31, 39
normally open (NO), 31, 39
notch filters, 131
nuisance opening, 27
nylon vs. metal gears, 204

O

octal base, 69
OCV (see open circuit voltage)
ohmic region, 256
on-off switches, 37
open circuit voltage (OCV/Voc), 13
open-frame power supplies, 146
open-frame switches, 51
open-loop systems, 210, 217
operating point, 250
optical mouse, 59
oscillating, 220
ounce-inches, 185
output
  voltage regulators and, 166
outrunner configuration, 184
overcoat material, 117
overheating
  DC-AC inverters and, 160
  MOSFETs, 263
  potentiometers, 96
  transformers and, 142
overshoot, 219

P

P layer, 223
P-channel, 253
P-type semiconductors, 241
paddle switches, 44
panel-mount switches, 45, 49
panel-mounted pushbuttons, 32
parallel batteries
  high current flow, 16
parasitic capacitance, 132
parasitic inductance, 87, 103
pass transistors, 161
PC board rotary switches, 55
PC mount switches, 45
PC pushbuttons, 32
PC terminals, 45
peak inrush, 147, 177
peak inrush current, 27
permanent magnet stepper motors, 211
permeability, 122
Peukert number, 12
phase count, 209
phases (steps), 209
pick-offs, 90
picofarads (pF), 104
piles, 5
pin configurations
  correct identification of, 167
  relays, 67, 72
PIN diodes, 221, 226
pitch, 18
planetary gears, 182
plastic-film dielectrics, 103
plug fuses, 22
plug-in transformers, 139
plungers, 173
PN junction diodes (see diodes)
polarity
  danger of incorrect, 15, 110
  danger of reversed, 233
  DC relays and, 67
  mnemonics for, 123
  preventing reversed, 142
polarized relays, 67
pole faces, 173
poles, 29, 30, 37, 38, 38
polymeric positive temperature coefficient fuse, 24
polyphase AC, 196
positive charges, 98
positive voltage regulators, 164
potentiometers, 75, 89–96, 186
  design of, 90
  function of, 89
  potential problems, 95
  schematic symbols for, 89
  types of, 90
  use of, 94
pound-feet, 185
power cells, 5
power cords
  correct choice of, 142
power MOSFETs, 262
power resistors, 76
power transformers, 139
power transistors, 245
power wire-wound resistors, 83
PPR (see pulses per rotation)
precision resistors, 76
precision wire-wound resistors, 83
press-twice pushbuttons, 33
primary cells, 7
primary coils, 135
programmable unijunction transistor (PUT), 235, 239
protection diodes, 218, 228
proximity effect, 133
PTC/PPTC (see polymeric positive temperature coefficient fuse)
pull-in torque, 220
pull-out torque, 220
pulldown resistors, 85
pullup resistors, 56, 85
pulse-width modulation (PWM), 144, 150, 158, 186, 199, 201, 217
pulses per rotation (PPR), 61

push-lock push-release pushbuttons, 33
push-on push-off pushbuttons, 33
push-push pushbuttons, 33
pushbutton switches, 29
pushbuttons, 29–36
    current ratings, 35
    design of, 29
    function of, 29
    potential problems, 35
    schematic symbol for, 29
    types of, 30
    use of, 35
pushwheel switches, 55
PUT (programmable unijunction transistor), 235, 239
PWM (see pulse-width modulation)

Q

quadrature output, 61
quality factor, 107
quasi-low-dropout voltage regulators, 165
quick connect terminals, 45
quiescent current, 166
quiescent point, 250

R

radial leads, 99
radial loading, 185
radial resistors, 76
radio buttons, 35
radio crystal set, 125
radio frequencies (RF), 133
ramping, 219
rated current
    fuses, 22
rated voltage, 22
RC (resistor-capacitor) network, 86, 98
RC circuits, 79
RC servo motors, 201
reactance, 115, 123, 128
real-coded rotary switches, 54,
rechargeable batteries, 7, 9–11
rectified current, 199
rectifier diodes, 224, 227, 233
rectifiers, 142, 143, 144, 170, 221
reed relays, 68, 72
reed switches, 68, 171
relative devices
    rotational encoders, 59
relative encoding, 53
relays, 65–73
    current rating, 70
    design of, 66
    function of, 65
    possible problems, 72
    schematic symbols for, 65, 67
    types of, 67
    use of, 71
reluctance, 121, 129, 170, 176
reluctance motors, 184, 197, 199
reluctance stepper motors, 210
remanent magnetism, 122
reserve batteries, 9
resettable fuses, 21, 24
resistance, 210
    rotational encoders, 61
resistor arrays, 77
resistor ladders, 56
resistor networks/ladders (see resistor arrays)
resistor-capacitor (RC) network, 86, 98
resistors, 75–88
    design of, 76
    function of, 75
    materials, 82
    possible problems, 87
    schematic symbols for, 75
    types of, 76
    use of, 84
    values for, 79
resistors in parallel, 86
resistors in series, 86
resonant frequency, 106, 220
response curves, 237
retentivity, 124
reverse charging, 16
rheostats, 89, 186
right-angle PC rotary switch, 54
right-angle PC terminals, 45
right-hand rule, 123
ripple-and-noise values, 154
RMS (see root mean square)
rocker switches, 40
root mean square (RMS), 158
rotary DIP switches, 53
rotary switches, 51–57
    current rating, 56
    design of, 52
    function of, 51
    potential problems, 57
    schematic symbol for, 51
    types of, 52
    use of, 56
rotational encoders, 47, 59–63
    current rating, 61
    design of, 59
    function of, 59
    potential problems, 62
    types of, 60
    use of, 62
rotors, 51, 113, 179, 191, 192, 209
    demagnetization of, 220

S

Safari Books Online, 3
safe operating area (SOA), 247
saturated mode, 245
saturation, 132, 220, 244
saturation current, 129
saturation region (Idss), 256
saturation, magnetic, 176
schematic symbols, 265
    AC-AC transformers, 135, 138
    batteries, 6
    bipolar transistors, 242
    capacitors, 97
    diodes, 221
    fuses, 21
    inductors, 119
    MOFSET, 259
    potentiometers, 89
    pushbuttons, 29
    relays, 65, 67
    resistors, 75, 82
    rotary switches, 51

  slide switches, 41
  switches, 49
  unijunction transistors, 235
Schottky diodes, 225
SCR (see silicon-controlled rectifier)
screw terminals, 45
secondary cells, 7
secondary coils, 135
self-discharge rate, 8
self-inductance, 119, 217
self-resonance, 132
self-resonant frequency, 106, 129
semiconductors, 223
series resistors, 84
servo motors, 201–207
  design of, 201
  function of, 201
  potential problems, 206
  types of, 203
  use of, 205
  values for, 204
  vs. stepper motors, 219
settling time, 61
shaded pole motors, 195
shell cores, 138
shielding
  variable capacitors and, 117
shock hazards
  AC-AC transformers and, 142
  AC-DC power supplies, 147
short-circuited batteries, 15, 26
short-circuited switches, 48
shorting coils, 195
shorting switches, 39, 53, 57
shunts (see jumpers)
signal clipping, 231
signal diodes, 224, 233
silicon-controlled rectifier (SCR), 186
single side stable relays, 67
single-coil latching relays, 67
single-inline package (SIP), 77
single-layer ceramic capicitors, 110
single-phase induction motors, 195
single-pole switches, 37
single-throw (ST) switches, 37
SIP (see single-inline package)
SIP switches, 44
skin effect, 133
slide pushbuttons, 41
slider potentiometers, 93
slider pushbuttons, 31
slider switches, 40, 48
sliding noise, 62
slip, 195
Slo-Blo fuses, 23
slow-blowing fuses, 26
small cartridge fuses, 23
small-signal relays, 69
small-signal transistors, 245
small-switching transistors, 245
SMPS (see switching power supply)
SMT (see surface-mount fuses)
snap-action switches, 35, 39
snap-in mount switches, 45
snubbers, 35, 47, 79, 108
SOA (see safe operating area)
solder lug terminals, 45
soldering damage, 27, 252
solenoids, 173–177
  design of, 174
  function of, 173
  potential problems, 177
  types of, 176
  use of, 177
  values for, 176
solid-state relays, 65
splines, 204
split-bobbin transformers, 141
split-phase induction motors, 196
sponge (lead), 9
spur gears, 182
squirre-cage motors, 192
squirrel-cage motors, 210
SRF (self-resonant frequency, 129
stability
  resistors and, 82
stall current, 184
stall torque, 185
standard type regulators, 161
star configuration, 181
stationary contacts, 38
stators, 113, 179, 191, 209
status pins, 165
step angle, 210
step loss, 219
step motors (see stepper motors)
step-down transformers, 136
step-up transformers, 136
stepper motors, 209–220
  design of, 209
  function of, 209
  potential problems, 219
  types of, 214
  use of, 218
  values for, 218
stepping motors (see stepper motors)
stops, 52
strip fuses, 24
subject-oriented organization, 2
subminiature fuses, 24
subminiature paddle switches, 44
subminiature slide switches, 41
subminiature switches, 40
subpanel mount switches, 45
substrate connections, 261
sulfurization, 16
sun gear, 183
supercapacitors, 8, 109
surface mount switches, 45
surface-mount (SMT) fuses, 26
surface-mount capacitors, 100
surface-mount resistors, 77
surface-mount transformers, 141
surface-mounted pushbuttons, 32
surge sensitivity, 22
switch bounce, 48, 62
switched potentiometers, 93
switched-mode power supply (see switching power supply)
switchers (see DC-DC converters) (see switching power supply)
switches, 17, 37–49, 51
  (see also rotary switches)
  design of, 37
  function of, 37
  potential problems, 47
  terminology, 38
  types of, 38
  use of, 46
  values for, 45

switching diodes, 224
switching power supplies, 143, 144, 159
switching regulators (see DC-DC converters)
synchronous motors, 184, 196

T

tactile swithes, 34
tank circuits (see LC circuits)
tantalum capacitors, 102, 110
taps, 137
Tcr/Tc (see temperature coefficient)
teeth (rotor), 210
temperature coefficient (Tcr/Tc), 82
terminals, choosing correct, 48
terminals, for switches, 45
THD (see total harmonic distortion)
thick-film resistors, 77, 83
thin-film resistors, 83
three-phase induction motors, 196
through-hole chips (capacitors), 101
through-hole fuses, 24, 26
through-hole terminals, 45
throws, 30, 38
thumbwheel switches, 55
time constant, 105, 130
time-delay relay switches, 70
toggle switches, 41
tolerance
    resistors, 79
toroidal core inductors, 126
torque
    expression of, 185
    fluctuations in, 193
    motor overloading and, 200
    stepper motors and, 220
total harmonic distortion (THD), 158
tracks, 90
transformer-based power supplies, 143
transient response, 166
transient suppressors, 228
transient voltage suppressor (TVS), 225
transients, 143
transistors, 241
    (see also bipolar transistors)
    resistors and, 84
transit time, 205
TrenchMOS, 262
triggering voltage, 236
trimmer capacitors, 113, 117
trimmer potentiometers, 89, 93, 113
Trimpots (see trimmer potentiometers)
trip current, 25
tripped fuses, 21
true sinewave inverters, 158
tuning capacitors, 113
tunnel diodes, 221, 226
turn rate, 205
TVS (see transient voltage suppressor)
two-contact pushbuttons, 29
two-pole switches, 37
two-way switches, 37

U

UJT (see unijunction transistor)
unijunction transistor (UJT), 235–240
    design of, 236
    function of, 235
    potential problems, 239
    schematic symbol for, 235
    types of, 238
    use of, 239, 240
    values for, 238
uninterruptible power supplies, 9
unipolar motors, 211, 214
universal motors, 198
universal stepper motors (see bifilar motors)
unloaded speed, 185

V

V-channel MOSFET (VMOS FET), 262
VA (see volts times amps)
vacuum tubes, 221, 241
valley voltage, 237
value coding, 88
    resistors, 81
valve-regulated lead-acid batteries (VRLA), 9
vandal resistant switches, 45
varactor diodes, 221, 225
variable capacitors, 113–117, 113
    (see also capacitors)
    design of, 113
    function of, 113
    possible problems, 117
    types of, 114
    use of, 115
    values for, 115
variable condensers (see variable capacitors)
variable frequency power supplies, 198
variable inductors, 125
variable reluctance, 210
variable resistors, 89
variable transformers, 140
variacs (see variable transformers)
varicaps (see varactor diodes)
Vc (see voltage coefficient)
Very Low Dropout voltage regulators, 165
vibration
    capacitors and, 110
    damping of, 220
    relays and, 73
Voc (see open circuit voltage)
voltage, 13
    automatic selection of, 229
    imbalance in AC motors, 200
    inaccurate in converters, 155
    inaccurate in voltage regulators, 167
    multiple outputs, 138,
    regulation with diodes, 230

  sensing with diodes, 232
voltage amplifiers, 253
voltage coefficient (Vc), 82
voltage dividers, 56, 56, 75, 86, 163, 244
voltage multipliers, 146
voltage overload
  bipolar transistors and, 252
  capacitors and, 110
voltage rating, 22
voltage regulators, 161–167
  design of, 161
  function of, 161
  potential problems, 166
  types of, 163
  use of, 165, 167
  values for, 165
voltage spikes
  protection against, 71, 221
voltage-controlled resistors, 256
volts times amps, 141
VRLA (see valve-regulated lead-acid batteries)

W

wall-warts, 143
windings, 121, 209
wipers, 90, 140
wire lead terminals, 45
wire-wound resistors, 77
wound-rotor AC induction motors, 198
wye configuration, 181

Y

Y configuration, 181
yokes, 171

Z

Zener diodes, 221, 225, 230, 231, 233
Zener voltage, 230
zero ohm components
  resistors, 82
zinc-carbon cells, 8

Science / Electronics

VOL. 2

Encyclopedia of Electronic Components

Signal Processing

Want to know how to use an electronic component? This second book of a three-volume set includes key information on electronics parts for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips—whether you're refreshing your memory or exploring a component for the first time. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** the first and only encyclopedia set on electronic components, distilled into three separate volumes
- **Incredibly detailed:** includes information distilled from hundreds of sources
- **Easy to browse:** parts are clearly organized by component type
- **Authoritative:** fact-checked by expert advisors to ensure that the information is both current and accurate
- **Reliable:** a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials
- **Instructive:** each component description provides details about substitutions, common problems, and workarounds
- **Comprehensive:** Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers a range of sensing devices.

Charles Platt

Charles Platt's lifelong love of electronics began when he built a telephone answering machine at age 15. A contributing editor to *Make Magazine*, he wrote the widely acclaimed *Make: Electronics*. He's also a science-fiction writer (author of *The Silicon Man*), and a former senior writer at *Wired* magazine.

US $29.99 CAN $31.99
ISBN: 978-1-4493-3418-5

Make:
makezine.com

2
Make: Encyclopedia of Electronic Components
Platt
MAKERMEDIA

VOL. 2

Charles Platt with Fredrik Jansson

Encyclopedia of Electronic Components

Signal Processing

LEDs ▪ LCDs ▪ Audio ▪ Thyristors
Digital Logic ▪ Amplification

Make:
makezine.com

Science / Electronics

VOL. 2

Encyclopedia of Electronic Components

Signal Processing

Want to know how to use an electronic component? This second book of a three-volume set includes key information on electronics parts for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips—whether you're refreshing your memory or exploring a component for the first time. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** the first and only encyclopedia set on electronic components, distilled into three separate volumes
- **Incredibly detailed:** includes information distilled from hundreds of sources
- **Easy to browse:** parts are clearly organized by component type
- **Authoritative:** fact-checked by expert advisors to ensure that the information is both current and accurate
- **Reliable:** a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials
- **Instructive:** each component description provides details about substitutions, common problems, and workarounds
- **Comprehensive:** Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers a range of sensing devices.

Charles Platt

Charles Platt's lifelong love of electronics began when he built a telephone answering machine at age 15. A contributing editor to *Make:* magazine, he wrote the widely acclaimed *Make: Electronics*. He's also a science-fiction writer (author of *The Silicon Man*), and a former senior writer at *Wired* magazine.

US $29.99 CAN $31.99
ISBN: 978-1-4493-3418-5

Make:
makezine.com

2
Make: Encyclopedia of Electronic Components Platt
MAKER MEDIA

VOL. 2

Charles Platt with Fredrik Jansson

Encyclopedia of Electronic Components

Signal Processing

LEDs ▪ LCDs ▪ Audio ▪ Thyristors
Digital Logic ▪ Amplification

Make:
makezine.com

Encyclopedia of Electronic Components Volume 2

Charles Platt
with Fredrik Jansson

MAKER MEDIA™
SEBASTOPOL, CA

Encyclopedia of Electronic Components Volume 2
by Charles Platt
with Fredrik Jansson

Copyright © 2015 Charles Platt. All rights reserved.

Printed in the United States of America.

Published by Maker Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Maker Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Brian Jepson
Production Editor: Melanie Yarbrough
Proofreader: Jasmine Kwityn
Indexer: Last Look Editorial
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator and Photographer: Charles Platt

November 2014: First Edition

Revision History for the First Edition:

2014-11-10: First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449334185* for release details.

Make:, Maker Shed, and Maker Faire are registered trademarks of Maker Media, Inc. The Maker Media logo is a trademark of Maker Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Maker Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

ISBN: 978-1-449-33418-5

[TI]

In fond memory of my father, Maurice Platt

Table of Contents

How to Use This Book . . . . **xxi**

> DISCRETE SEMICONDUCTOR

> > THYRISTOR

1. SCR **1**
What It Does 1
How It Works 1
    Switching Behavior 2
Internal Configuration 3
    Breakdown and Breakover Voltage 4
    SCR Concept Demo 4
    AC Current Applications 5
Variants 5
Values 5
    Commonly Used Abbreviations 6
How to Use It 6
    Phase Control 7
    Overvoltage Protection 7
What Can Go Wrong 8
    Unexpected Triggering Caused by Heat 8
    Unexpected Triggering Caused by Voltage 8
    Confusion of AC and DC Ratings 8
    Maximum Current versus Conduction Angle 9
    Confusing Symbols 9

2. diac **11**
What It Does 11

Symbol Variants 11
How It Works 12
Switching AC 13
Variants 13
Values 14
What Can Go Wrong 14
Unexpected Triggering Caused by Heat 14
Low-Temperature Effects 14
Manufacturing Tolerances 14

3. triac 15
What It Does 15
Symbol Variants 15
How It Works 17
Quadrants 17
Threshold, Latching, and Holding Current 18
Triac Testing 19
Breakover Voltage 20
Switching AC 20
Triac Triggered by a Diac 22
Other Triac Drivers 22
Charge Storage 23
Variants 23
Values 23
What Can Go Wrong 24
Unexpected Triggering Caused by Heat 24
Low-Temperature Effects 24
Wrong Type of Load 24
Wrongly Identified Terminals 24
Failure to Switch Off 24

> INTEGRATED CIRCUIT

> > ANALOG

4. solid-state relay 25
What It Does 25
Advantages 25
Disadvantages 26
How It Works 26
Variants 27
Instantaneous versus Zero Crossing 27
NC and NO Modes 28
Packaging 28
Solid-State Analog Switch 28
Values 29
How to Use It 29
What Can Go Wrong 30

Overheating Caused by Overloading 30
Overheating Caused by Bad Terminal Contact 30
Overheating Caused by Changing Duty Cycle 30
Overheating Caused by Component Crowding 30
Overheating in Dual Packaging 30
Reverse-Voltage Burnout 30
Low Voltage Output Current May Not Work 30
Inability to Measure AC Output 30
Relay Turns On but Won't Turn Off 31
Relays in Parallel Won't Work 31
Output Device Doesn't Run at Full Power 31
Solid-State Relays and Safety Disconnects 31

5. optocoupler 33
What It Does 33
How It Works 34
Variants 35
  Internal Sensors 35
  Basic Optocoupler Types 36
Values 36
How to Use It 36
What Can Go Wrong 37
  Age 37
  LED Burnout 37
  Transistor Burnout 37

6. comparator 39
What It Does 39
  Hysteresis 39
How It Works 39
  Differences from an Op-Amp 42
Variants 42
Values 43
How to Use It 44
  AND gate 45
  Bistable Multivibrator 46
  Relaxation Oscillator 46
  Level Shifter 46
  Window Comparator 46
  Other Applications 47
What Can Go Wrong 47
  Oscillating Output 47
  Confused Inputs 48
  Wrong Chip Type 48
  Omitted Pullup Resistor 48
  CMOS Issues 48
  Erratic Output 48
  Swapped Voltages 48
  Heat-Dependent Hysteresis 48

7. op-amp 49
- What It Does 49
- How It Works 49
  - Dual Inputs 50
  - Negative Feedback 51
  - Op-Amps and Comparators 52
- Variants 52
- Values 52
- How to Use It 53
  - Controlling the Gain 53
  - Calculating Amplification 53
  - Unintentional DC Voltage Amplification 54
  - Low-Pass Filter 54
  - High-Pass Filter 55
  - Relaxation Oscillator 55
  - Single Power Source 56
  - Offset Null Adjustment 56
- What Can Go Wrong 57
  - Power Supply Problems 57
  - Bad Connection of Unused Sections 57
  - Oscillating Output 57
  - Confused Inputs 58

8. digital potentiometer 59
- What It Does 59
  - Advantages 59
- How It Works 60
- Variants 61
  - Volatile and Nonvolatile Memory 61
  - Taper 62
  - Data Transfer 62
  - SPI 62
  - I2C Protocol 63
  - Up/Down Protocol 63
  - Other Control Systems 64
  - Connections and Modes 64
- Values 65
- How to Use It 66
  - Achieving Higher Resolution 66
- What Can Go Wrong 67
  - Noise and Bad Inputs 67
  - Wrong Chip 67
  - Controller and Chip Out of Sync 67
  - Nonlinear Effects 67
  - Data Transfer Too Fast 67

9. timer 69
- What It Does 69
  - Monostable Mode 69

Astable Mode 70
How It Works 70
Variants 70
The 555 Timer 70
555 Monostable Operation 71
555 Astable Operation 72
556 Timer 73
558 Timer 73
CMOS 555 Timer 74
5555 Timer 74
7555 Timer 74
7556 Timer 74
4047B Timer 75
Dual Monostable Timers 75
Values 76
555 Timer Values 76
Time Calculation in Monostable Mode 77
Time Calculation in Astable Mode 77
Dual Monostable Timers 77
How to Use It 79
555 Monostable Mode 79
555 Astable Mode 80
Separate Control of High and Low Output Times 80
555 Fifty Percent Astable Duty Cycle: 1 80
555 Fifty Percent Astable Duty Cycle: 2 81
Use of the 555 Control Pin 81
555 Flip-Flop Emulation 82
555 Hysteresis 83
555 and Coupling Capacitors 84
555 Loudspeaker Connection 84
Burst Mode 84
"You Lose" Game Sound 85
What Can Go Wrong 85
Dead Timer 85
CMOS Confused with Bipolar 86
The Pulse that Never Ends 86
Erratic Chip Behavior 86
Interference with Other Components 86
Erratic Behavior of Output Devices 86
Fatal Damage Caused by Inductive Loads 87

> > DIGITAL

10. **logic gate** **89**
What It Does 89
Origins 89
How It Works 89
Inversion 90
Single-Input Gates 90

Gates with More than Two Inputs 91
Boolean Notation 91
Arithmetical Operations 91
Other Operations 92
Variants 93
Part Numbers 94
Families 95
Family Interoperability 96
Gates per Chip 96
Two Inputs, Single Gate 96
Three Inputs, Single Gate 97
Single Gate, Selectable Function 97
Two Inputs, Dual Gate 98
Original 74xx 14-Pin Format 98
Quad Two-Input 74xx Pinouts 98
Triple Three-Input 74xx Pinouts 99
Dual Four-Input 74xx Pinouts 100
Single Eight-Input 74xx Pinouts 100
74xx Inverters 101
Additional Variations 102
Pinouts in the Original 4000 Series 102
4000 Series Inverters 103
How to Use It 103
Which Family 103
Applications 104
What Can Go Wrong 105
Static 105
Floating Pins 105
Family Incompatibilities 105
Overloaded Outputs 105
Output Pulled Down 105
Incorrect Polarity and Voltages 105
Bent Pins 105
Unclean Input 106
Analog Input 106

11. flip-flop 107
What It Does 107
How It Works 107
NAND-Based SR Flip-Flop 108
NOR-Based SR Flip-Flop 109
Forbidden States 110
The JK Flip-Flop 112
Master-Slave Flip-Flop 113
D-Type Flip-Flops 114
Summary 116
Variants 116
Packaging 117
Values 117

How to Use It . . . 118
What Can Go Wrong . . . 119
    Ambiguous Documentation . . . 119
    Faulty Triggering . . . 119
    Metastability . . . 119
    Other Issues . . . 119

12. **shift register . . . 121**
What It Does . . . 121
    Schematic Representation . . . 122
How It Works . . . 122
    Abbreviations and Acronyms . . . 123
    Parallel Outputs and Inputs . . . 123
Variants . . . 124
    Serial In, Serial Out . . . 124
    Serial In, Parallel Out . . . 124
    Parallel In, Serial Out . . . 124
    Parallel In, Parallel Out . . . 125
    Universal . . . 125
Values . . . 125
    Power Considerations . . . 126
    Three-State Output . . . 126
How to Use It . . . 126
    Dual Inputs . . . 127
    Preloading the Shift Register . . . 127
    Polling a Keyboard . . . 127
    Arithmetical Operations . . . 127
    Buffering . . . 128
What Can Go Wrong . . . 128
    Confusing Classification . . . 128
    Inadequate Setup Time . . . 128
    Unconnected Input . . . 129
    Output Enable Issues . . . 129
    Floating Output Bus . . . 129

13. **counter . . . 131**
What It Does . . . 131
    Schematic Representation . . . 131
How It Works . . . 132
    Modulus and Modulo . . . 132
    Pin Identifiers . . . 133
Variants . . . 134
    Ripple versus Synchronous . . . 134
    Ring, Binary, and BCD . . . 134
    Clock Sources . . . 135
    Rising Edge and Falling Edge . . . 136
    Multiple Stages . . . 136
    Single and Dual . . . 136
    High-State, Low-State, and Three-State . . . 136

    - Descending Output . . . 136
    - Programmable Counters . . . 137
    - Examples . . . 137
  - Values . . . 137
  - What Can Go Wrong . . . 137
    - Lock-Out . . . 137
    - Asynchronous Artifacts . . . 137
    - Noise . . . 138

14. **encoder** . . . **139**
  - What It Does . . . 139
    - Schematic Symbol . . . 139
    - Similar Devices . . . 140
  - How It Works . . . 140
  - Variants . . . 141
  - Values . . . 142
  - How to Use It . . . 142
    - Cascaded Encoders . . . 142
  - What Can Go Wrong . . . 143

15. **decoder** . . . **145**
  - What it Does . . . 145
    - Input Devices . . . 145
    - LED Driver . . . 146
    - Schematic Symbol . . . 146
    - Similar Devices . . . 146
  - How It Works . . . 148
  - Variants . . . 148
  - Values . . . 148
  - How to Use It . . . 149
  - What Can Go Wrong . . . 149
    - Glitches . . . 149
    - Unhelpful Classification . . . 149
    - Active-Low and Active-High . . . 149

16. **multiplexer** . . . **151**
  - What It Does . . . 151
    - Differential Multiplexer . . . 152
    - Similar Devices . . . 152
  - How It Works . . . 153
    - Schematic Symbol . . . 153
    - Pin Identifiers . . . 154
  - Variants . . . 155
  - Values . . . 155
  - How to Use It . . . 156
    - Other Application Notes . . . 157
  - What Can Go Wrong . . . 157
    - Pullup Resistors . . . 157
    - Break Before Make . . . 157

Signal Distortion 157
Limits of CMOS Switching 157
Transients 157

> LIGHT SOURCE, INDICATOR, OR DISPLAY

> > REFLECTIVE

17. LCD 159
What It Does 159
How It Works 159
Variants 160
  Active and Passive Types 161
  Crystal Types 161
  Seven-Segment Displays 161
  Additional Segments 162
  Dot-Matrix Displays 163
  Color 166
  Backlighting Options 167
  Zero-Power Displays 167
How to Use It 167
  Numeric Display Modules 167
  Alphanumeric Display Module 168
What Can Go Wrong 169
  Temperature Sensitivity 169
  Excessive Multiplexing 170
  DC Damage 170
  Bad Communications Protocol 170
  Wiring Errors 170

> > SINGLE SOURCE

18. incandescent lamp 171
What It Does 171
History 172
How It Works 172
  Spectrum 173
  Non-Incandescent Sources 174
  Power Consumption 175
Variants 175
  Miniature Lamps 175
  Panel-Mount Indicator Lamps 176
  Halogen or Quartz-Halogen 176
  Oven Lamps 176
  Base Variants 177
Values 177
  Power 177
  Illuminance 178

    - Intensity 178
    - MSCP 178
    - Efficacy 179
    - Efficiency 179
  - How to Use It 179
    - Relative Advantages 179
    - Derating 180
  - What Can Go Wrong 180
    - High Temperature Environment 180
    - Fire Risk 181
    - Current Inrush 181
    - Replacement Problems 181

19. neon bulb 183

- What It Does 183
- How It Works 184
  - Construction 184
  - Ionization 184
  - Negative Resistance 185
- How to Use It 186
  - Limited Light Output 187
  - Efficiency 187
  - Ruggedness 187
  - Power-Supply Testing 188
  - Life Expectancy 188
- Variants 189
  - Nixie Tubes 189
- What Can Go Wrong 189
  - False Indication 189
  - Failure in a Dark Environment 189
  - Premature Failure with DC 190
  - Premature Failure through Voltage Fluctuations 190
  - Replacement 190

20. fluorescent light 191

- What It Does 191
- How It Works 191
  - Ballast and Starter 192
  - Flicker 193
- Variants 193
  - CCFLs 194
  - Sizes 194
  - Comparisons 194
- Values 195
  - Brightness 195
  - Spectrum 195
- What Can Go Wrong 195
  - Unreliable Starting 195
  - Terminal Flicker 195

Cannot Dim 196
Burned Out Electrodes 196
Ultraviolet Hazard 196

21. laser

21. laser 197
What It Does 197
How It Works 198
  Laser Diode 198
  Coherent Light 200
Variants 201
  CO2 Lasers 201
  Fiber Lasers 201
  Crystal Lasers 201
Values 201
How to Use It 202
  Common Applications 202
What Can Go Wrong 202
  Risk of Injury 202
  Inadequate Heat Sink 202
  Uncontrolled Power Supply 203
  Polarity 203

22. LED indicator 205
What It Does 205
  Schematic Symbols 206
  Common Usage 206
How It Works 207
  Multicolor LEDs and Color Mixing 207
Variants 208
  Size and Shape 208
  Intensity 208
  Efficacy 208
  Diffusion 209
  Wavelength and Color Temperature 209
  Internal Resistor 210
  Multicolored 210
  Infrared 211
  Ultraviolet 211
Values 211
  Forward Current 211
  Low-Current LEDs 211
  Forward Voltage 212
  Color Rendering Index 212
  Life Expectancy 212
  Light Output and Heat 212
  View Angle 213
How to Use It 213
  Polarity 213
  Series Resistor Value 214

- LEDs in Parallel 214
- Multiple Series LEDs 214
- Comparisons with Other Light Emitters 214
- Other Applications 215

What Can Go Wrong 215
- Excessive Forward Voltage 215
- Excessive Current and Heat 215
- Storage Issues 215
- Polarity 215
- Internal Resistors 215

23. LED area lighting 217

What It Does 217
- Trends in Cost and Efficiency 218
- Schematic Symbol 218

How It Works 218
- Visible Differences 220
- Side-by-Side Comparison 220
- Heat Dissipation 222
- Efficacy 222
- Dimming 222
- Ultraviolet Output 222
- Color Variation 222

Variants 223
- Comparisons 223

Values 225

What Can Go Wrong 225
- Wrong Voltage 225
- Overheating 225
- Fluorescent Ballast Issues 225
- Misleading Color Representation 226

> > MULTI-SOURCE OR PANEL

24. LED display 227

What It Does 227

How It Works 228

Variants 228
- LCD comparisons 228
- Seven-Segment Displays 228
- Multiple Numerals 229
- Additional Segments 229
- Dot-Matrix Displays 230
- Pixel Arrays 231
- Multiple Bar Display 232
- Single Light Bar 232

Values 232

How to Use It 232
- Seven-Segment Basics 232

Driver Chips and Multiplexing 233
Sixteen-Segment Driver Chip 234
Dot-Matrix LED Display Modules 234
Pixel Arrays 235
Multiple Bar Display Driver 236
One-Digit Hexadecimal Dot Matrix 236
What Can Go Wrong 237
Common Anode versus Common Cathode 237
Incorrect Series Resistance 237
Multiplexing Issues 237

25. **vacuum-fluorescent display** 239
What It Does 239
How It Works 239
Anode, Cathode, and Grid 240
How to Use It 240
Modern Application 241
Variants 241
Color 241
Character Sets and Pictorial Design 242
Comparisons 242
What Can Go Wrong 242
Fading 242

26. **electroluminescence** 243
What It Does 243
How It Works 243
Phosphors 244
Derivation 244
Variants 244
Panels 244
Flexible Ribbons 245
Rope Light 245
OLED 246

> SOUND SOURCE

> > AUDIO ALERT

27. **transducer** 249
What It Does 249
How It Works 249
Variants 250
Electromagnetic 250
Piezoelectric 250
Ultrasonic Transducer 250
Formats 251
Values 251

    - Frequency Range 251
    - Sound Pressure 251
    - Weighted Sound Values 252
    - Unweighted Values 253
    - Measurement Location 253
    - Limitations 253
    - Voltage 254
    - Current 254
- How to Use It 254
    - Appropriate Sound Intensity 254
    - Volume Control 254
    - AC Supply 254
    - Self-Drive Transducer Circuit 254
- What Can Go Wrong 254
    - Overvoltage 254
    - Leakage 255
    - Component Mounting Problems 255
    - Moisture 255
    - Transducer-Indicator Confusion 255
    - Connection with a Microcontroller 255

28. audio indicator 257

- What It Does 257
- How It Works 257
    - Audio Frequency 258
    - History 258
- Variants 258
    - Sound Patterns 258
    - Formats 258
- Values 259
    - Voltage 259
    - Current 260
    - Frequency 260
    - Duty Cycle 260
- How to Use It 260
    - Appropriate Sound Intensity 260
    - Volume Control 260
    - Wiring 260
- What Can Go Wrong 260

> > REPRODUCER

29. headphone 261

- What It Does 261
- How It Works 261
    - Audio Basics 261
- Variants 262
    - Moving Coil 262
    - Other Types 263

Mechanical Design 264
Values 265
Intensity 265
Frequency Response 265
Distortion 266
Impedance 266
What Can Go Wrong 266
Overdriving 266
Hearing Damage 266
Mismatched Impedance 266
Incorrect Wiring 266

30. **speaker** **267**
What It Does 267
How It Works 267
Construction 268
Multiple Drivers 269
Venting 270
Resonance 270
Miniature Speakers 270
Variants 271
Electrostatic Speaker 271
Powered Speakers 271
Wireless Speakers 271
Innovative Designs 271
Values 271
What Can Go Wrong 272
Damage 272
Magnetic Field 272
Vibration 272

Index **273**

How to Use This Book

This is the second of three volumes. Its purpose is to provide an overview of the most commonly used electronic components, for reference by students, engineers, hobbyists, and instructors. While you can find much of this information dispersed among datasheets, introductory books, websites, and technical resources maintained by manufacturers, the *Encyclopedia of Electronic Components* gathers all the relevant facts in one place, properly organized and verified, including details that may be hard to find elsewhere. Each entry includes typical applications, possible substitutions, cross-references to similar devices, sample schematics, and a list of common problems and errors.

You can find a more detailed rationale for this encyclopedia in the Preface to Volume 1.

Volume Contents

Practical considerations influenced the decision to divide this encyclopedia into three volumes. Each deals with broad subject areas as follows.

Volume 1

Power; electromagnetic devices; discrete semiconductors

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnetic devices* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. A contents listing for Volume 1 appears in Figure P-1.

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in Figure P-2.

Volume 3

Sensing devices

The field of sensors has become so extensive, they easily merit a volume to themselves. *Sensing devices* include those that detect light, sound, heat, motion, pressure, gas, humidity, orientation, electricity, proximity, force, and radiation.

At the time of writing, Volume 3 is still in preparation, while Volume 1 is complete and is available in a variety of formats.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electromagnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semiconductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1. *The subject-oriented organization of categories and entries in Volume 1.*

Primary Category	Secondary Category	Component Type
discrete semiconductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2. *The subject-oriented organization of categories and entries in Volume 2.*

Organization

Reference versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

My books *Make: Electronics* and *Make: More Electronics* follow a tutorial approach. They don't go into as much depth as this Encyclopedia, because a tutorial inevitably allocates a lot of space to step-by-step explanations and instructions.

Theory and Practice

This book is oriented toward practicality rather than theory. I assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently, I have not included proofs of formulae or definitions rooted in electrical theory. Units are defined only to the extent necessary to avoid confusion.

Many books on electronics theory already exist, if theory is of interest to you.

Entries

This encyclopedia is divided into entries, each entry being devoted to one broad type of component. Two rules determine whether a component has an entry all to itself, or is subsumed into another entry:

Rule 1

> A component merits its own entry if it is (a) widely used, or (b) not so widely used but has a unique identity and maybe some historical status. The **bipolar transistor** entry is an example of a widely used component, whereas the **unijunction transistor** entry is an example of a not so widely used component with a unique identity.

Rule 2

> A component does not merit its own entry if it is (a) seldom used, or (b) very similar in function to another component that is more widely used. For example, a *rheostat* is subsumed into the **potentiometer** section, while *silicon diode*, *Zener diode*, and *germanium diode* are combined together in the **diode** entry.

Inevitably, these guidelines required judgment calls which in some cases may seem arbitrary. My ultimate decision was based on where I would expect to find a component if I was looking for it myself.

Subject Paths

Entries are not organized alphabetically. They are grouped by subject, in much the same way that books in the nonfiction section of some libraries are organized by the Dewey Decimal System. This is convenient if you don't know exactly what you are looking for, or if you don't know all the options that may be available to perform a task that you have in mind.

Each primary category is divided into subcategories, and the subcategories are divided into component types. This hierarchy is shown in Figure P-2. It is also apparent when you look at the top of the first page of each entry, where you will find the path that leads to it. The **diac** entry, for instance, is headed with this path:

discrete semiconductor > thyristor > diac

Any classification scheme will run into exceptions. You can buy a chip containing a *resistor array*, for instance. Technically, this is an *analog integrated circuit*, but a decision was made to put it in the **resistor** section of Volume 1, because it can be directly substituted for a group of resistors.

Some components have hybrid functions. A **multiplexer**, for instance, may pass analog signals and may have "analog" in its name. However, it is digitally controlled and is mostly used in conjunction with other digital integrated circuits. This seemed to justify placing it in the digital category.

Inclusions and Exclusions

There is also the question of what is, and is not, a component. Is wire a component? Not for the purposes of this encyclopedia. How about a **DC-DC converter**? Because converters are now sold in small packages by component suppliers, they are included in Volume 1 as components.

Many similar decisions had to be made on a case-by-case basis. Some readers will disagree with the outcome, but reconciling all the disagreements would have been impossible. The best I could do was to create a book which is organized in the way that would suit me best if I were using it myself.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark. The term *Trimpot*, for instance, is trademarked by Bourns, but *trimmer* is not. **LED** is an acronym, but *cap* (abbreviation for **capacitor**) is not.

The European convention for representing fractional component values eliminates decimal points. Thus, values such as 3.3K and 4.7K are expressed as 3K3 and 4K7. This style has not been adopted to a significant degree in the United States, and is not used in this encyclopedia.

In mathematical formulae, I have used the style that is common in programming languages. The * (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. So, in this example:

```
A = 30 / (7 + (4 * 2) )
```

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-3 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top right is avoided, and the configuration at top center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

Photographic Backgrounds

All photographs of components include a background grid that is divided into squares measuring 0.1". Although the grid is virtual, it is equivalent in scale to physical graph paper placed immediately behind the component. If the component is photographed at an angle, the grid may be reproduced at a similar angle, creating perspective on the squares.

Background colors in photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

Because there is no way of knowing if a component may have a long production run, this encyclopedia is cautious about listing specific part numbers. To find a specific part that has a narrow function, searching the websites maintained by suppliers will be necessary. The following suppliers were checked frequently during the preparation of the book:

- Mouser Electronics
- Jameco Electronics

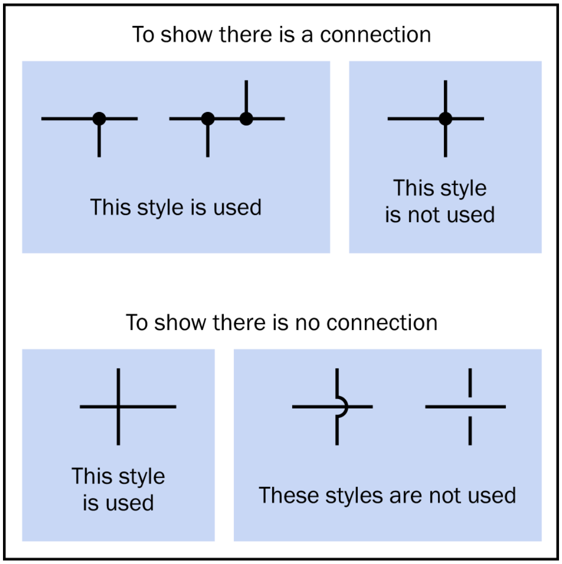
To show there is a connection
This style is used
This style is not used
To show there is no connection
This style is used
These styles are not used

Figure P-3. *Visual conventions that are used in the schematics in this book.*

When seeking obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful.

Issues and Errata

If you believe you have found an error in this book, you will find guidance on how to report it here: *http://bit.ly/eec_v2_errata*.

Before posting your own erratum, please check those that have been submitted previously, to see if someone else already reported it.

I value and encourage reader feedback. However, before you post feedback publicly to a site such as Amazon, I have a request. Please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews. If you feel you have not received a prompt or adequate response from the O'Reilly errata site mentioned here, you can email me personally at:

make.electronics@gmail.com

I check that address irregularly—sometimes only once in a couple of weeks. But I do answer all messages.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like Maker Media, O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

Make:
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make: is leading—we call it the Maker Movement.

For more information about Make:, visit us online:

Make: magazine: *http://makezine.com/magazine/*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com/*

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at: *http://bit.ly/encyclopedia_of_electronic_components_v2*.

Acknowledgments

Any reference work draws inspiration from many sources. Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, Op Amp, & Optelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

I also made extensive use of information on vendor sites, especially:

- Mouser Electronics
- Jameco Electronics
- All Electronics
- sparkfun
- Electronic Goldmine
- Adafruit
- Parallax, Inc.

In addition, some individuals provided special assistance. My editor, Brian Jepson, was immensely helpful in the development of this book. Philipp Marek and Steve Conklin reviewed the text for errors. My publisher demonstrated faith in my work. Kevin Kelly unwittingly influenced me with his legendary interest in "access to tools." It was Mark Frauenfelder who originally brought me back to the pleasures of building things, and Gareth Branwyn who revived my interest in electronics.

Lastly, I should mention my school friends from decades ago: Patrick Fagg, Hugh Levinson, Graham Rogers, William Edmondson, and John Witty, who helped me to feel that it was OK to be a nerd building my own audio equipment, long before the word "nerd" actually existed.

—Charles Platt, 2014

SCR

1

The acronym **SCR** is derived from *silicon-controlled rectifier*, which is a gate-triggered type of *thyristor*. A thyristor is defined here as a semiconductor having four or more alternating layers of p-type and n-type silicon. Because it predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, a thyristor is considered to be a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**), it is considered to be an integrated circuit.

Other types of thyristor are the **diac** and **triac**, each of which has its own entry.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **diac** (see Chapter 2)
- **triac** (see Chapter 3)

What It Does

In the 1920s, the *thyratron* was a gas-filled tube that functioned as a switch and a rectifier. In 1956, General Electric introduced a solid-state version of it under the name *thyristor*. In both cases, the names were derived from the thyroid gland in the human body, which controls the rate of consumption of energy. The thyratron and, subsequently, the thyristor enabled control of large flows of current.

The **SCR** (silicon-controlled rectifier) is a type of thyristor, although the two terms are often used as if they are synonymous. Text that refers loosely to a thyristor may actually be discussing an SCR, and vice versa. In this encyclopedia, the **SCR**, **diac**, and **triac** are all considered to be variant types of thyristor.

An SCR is a solid-state switch that in many instances can pass high currents at high voltages. Like a **bipolar transistor**, it is triggered by voltage applied to a gate. Unlike the transistor, it allows the flow of current to continue even when the gate voltage diminishes to zero.

How It Works

This component is designed to pass current in one direction only. It can be forced to conduct in the opposite direction if the reversed potential exceeds its *breakdown voltage*, but this mistreatment is likely to cause damage.

By comparison, the diac and triac are designed to be bidirectional.

The SCR has three leads, identified as anode, cathode, and gate. Two functionally identical versions of the schematic symbol are shown in Figure 1-1. Early versions sometimes included a circle drawn around them, but this style has become obsolete. Care must be taken to distinguish between the SCR symbol and the symbol that represents a **programmable unijunction transistor** (PUT), shown in Figure 1-2.

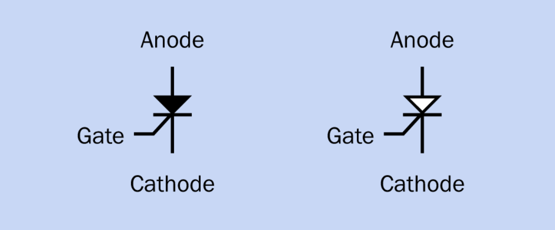

Figure 1-1. *Two functionally identical schematic symbols for an SCR (silicon-controlled rectifier). The symbol on the left is more common.*

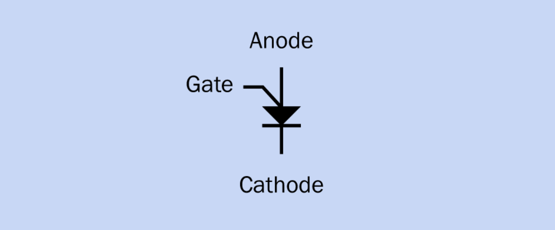

Figure 1-2. *The symbol shown here is for a programmable unijunction transistor (PUT). Care must be taken to distinguish it from the symbol for an SCR.*

Switching Behavior

When the SCR is in its passive or nonconductive state, it will block current in either direction between anode and cathode, although a very small amount of *leakage* typically occurs. When the SCR is activated by a positive voltage at the gate, current can now flow from anode to cathode, although it is still blocked from cathode to anode. When the flow reaches a level known as the *latching current*, the flow will continue even after the triggering voltage drops to zero. This behavior causes it to be known as a *regenerative* device.

If the current between anode and cathode starts to diminish while the gate voltage remains zero, the current flow will continue below the latching level until it falls below the value known as the *holding current*. The flow now ceases. Thus, the only way to end a flow of current that has been initiated through an SCR is by reducing the flow or attempting to reverse it.

Note that the self-sustaining flow is a function of current rather than voltage.

Unlike a transistor, an SCR is either "on" or "off" and does not function as a *current amplifier*. Like a diode, it is designed to conduct current in one direction; hence the term *rectifier* in its full name. When it has been triggered, the impedance between its anode and cathode is sufficiently low that heat dissipation can be managed even at high power levels.

The ability of SCRs to pass relatively large amounts of current makes them suitable for controlling the power supplied to motors and resistive heating elements. The fast switching response also enables an SCR to interrupt and abbreviate each positive phase of an AC waveform, to reduce the average power supplied. This is known as *phase control*.

SCRs are also used to provide *overvoltage protection*.

SCR packages reflect their design for a wide range of voltages and currents. Figure 1-3 shows an SCR designed for on-state current of 4A RMS (i.e., measured as the root mean square of the alternating current). Among its applications are small-engine ignition and *crowbar* overvoltage protection, so named because it shorts a power supply directly to ground, much like a crowbar being dropped across the terminals of a car battery (but hopefully with a less dramatic outcome). See Figure 1-15.

In Figure 1-4, the SCR can handle up to 800V repetitive peak off-state voltage and 55A RMS. Possible applications include AC rectification, crowbar protection, welding, and battery charging. The component in Figure 1-5 is rated for 25A and 50V repetitive peak off-state voltage. To assess the component sizes, bear in mind that the graph line spacing is 0.1".

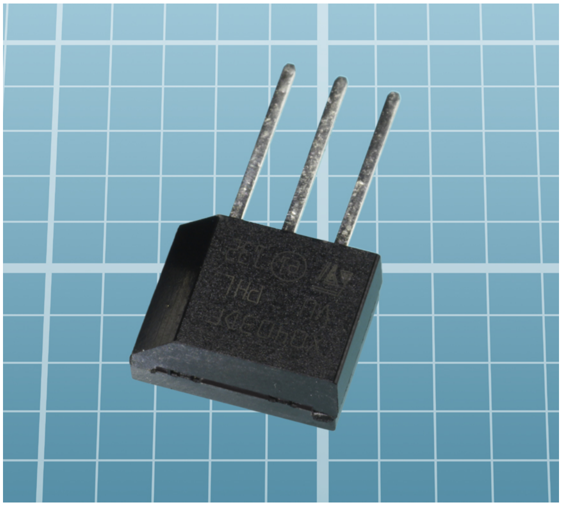

Figure 1-3. *SCR rated for 400V repetitive off-state voltage, no greater than 4A RMS.*

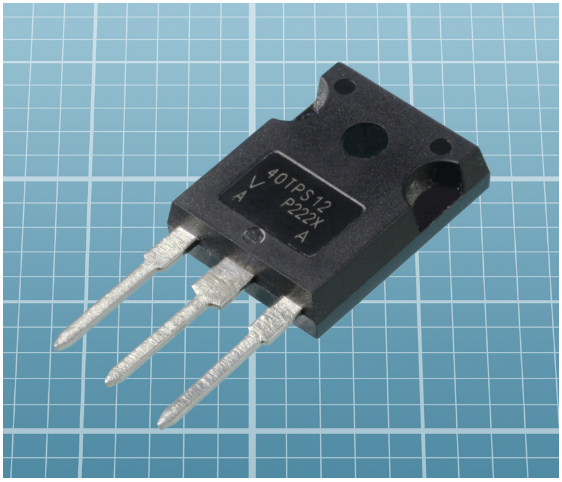

Figure 1-4. *SCR rated for 800V repetitive off-state voltage, no greater than 55A RMS.*

Internal Configuration

The function of an SCR can be imagined as being similar to that of a PNP transistor paired with an NPN transistor, as shown in Figure 1-6. In this simplified schematic, so long as zero voltage is applied to the "gate" wire, the lower (NPN) transistor remains nonconductive. Consequently, the upper (PNP) transistor cannot sink current, and this transistor also remains nonconductive.

When voltage is applied to the "gate," the lower transistor starts to sink current from the upper transistor. This switches it on. The two transistors now continue to conduct even if power to the "gate" is disconnected, because they have created a positive feedback loop.

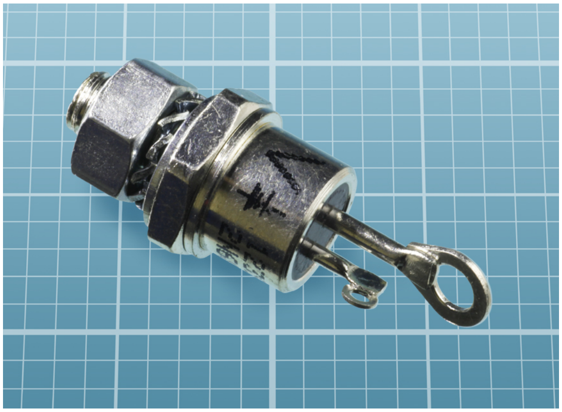

Figure 1-5. *Stud-packaged SCR rated for 50V repetitive off-state voltage, no greater than 25A RMS.*

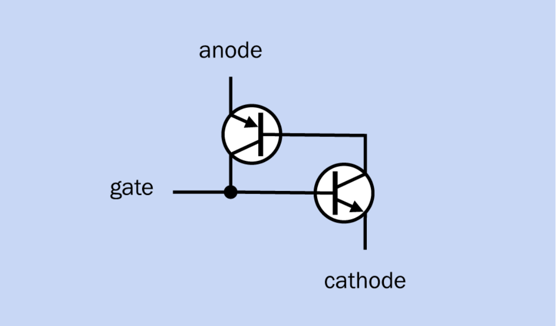

Figure 1-6. *An SCR behaves similarly to an NPN and a PNP transistor coupled together.*

Figure 1-7 shows the same two transistors in simplified form as sandwiches of p-type and n-type silicon layers (on the left), and their combination in an SCR (on the right). Although the actual configuration of silicon segments is not as simple or as linear as this diagram suggests, the SCR can be described correctly as a *PNPN device*.

An SCR is comparable with an electromagnetic **latching relay**, except that it works much faster and more reliably.

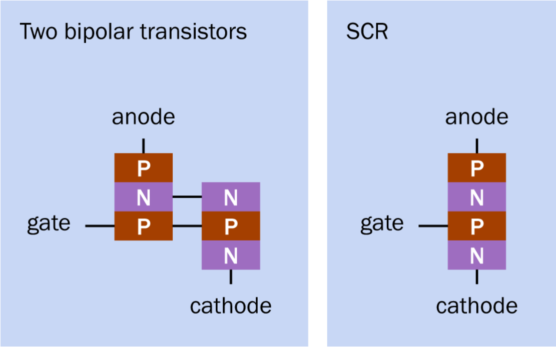

Figure 1-7. *The two transistors from the previous figure are shown here in simplified form as two stacks of p-type and n-type silicon layers. These layers are combined in an SCR, on the right.*

Breakdown and Breakover Voltage

The curves in Figure 1-8 illustrate the behavior of a hypothetical SCR, and can be compared with the curves shown for a diac in Figure 2-5 and a triac in Figure 3-10. Beginning with zero voltage applied between anode and cathode, and zero current flowing (i.e., at the center origin of the graph), if we apply a voltage at the anode that is increasingly negative relative to the cathode (i.e., we attempt to force the SCR to allow negative current flow), we see a small amount of leakage, indicated by the darker blue area (which is not drawn to scale). Finally the *breakdown voltage* is reached, at which point the negative potential overcomes the SCR and its impedance drops rapidly, allowing a surge of current to flow, probably damaging it.

Alternatively, starting once again from the center, if we apply a voltage at the anode that is increasingly positive relative to the cathode, two consequences are possible. The dashed curve assumes that there is zero voltage at the gate, and shows that some leakage occurs until the applied potential at the anode reaches the *breakover voltage*, at which point the SCR allows a large current flow, which continues even when the voltage decreases.

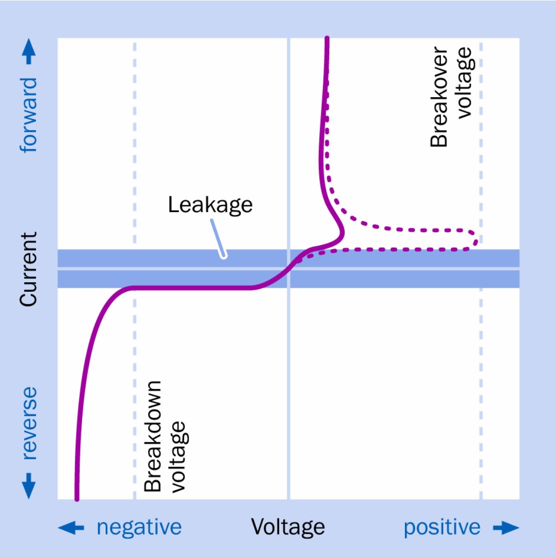

Figure 1-8. *The solid curve shows current passing between the anode and cathode of a hypothetical SCR for varying voltages, while a triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate.*

In practice, the SCR is intended to respond to a positive gate voltage. Under these circumstances, its behavior is shown by the solid curve in the top-right quadrant in Figure 1-8. The SCR begins to conduct current without having to reach the breakover voltage at the anode.

- When used as it is intended, the SCR should not reach breakdown or breakover voltage levels.

SCR Concept Demo

In Figure 1-9, pushbutton S1 applies voltage to the gate of the SCR, which puts the SCR in self-sustaining conductive mode. When S1 is released, the meter will show that current continues to pass between the anode and the cathode. The X0403DF SCR suggested for this circuit has a holding current of 5mA, which a 5VDC supply should be able to provide with the 1K resistor in the circuit. If necessary, this resistor can be reduced to 680Ω.

Now if pushbutton S2 is pressed, the flow is interrupted. When S2 is released, the flow will not resume. Alternatively, if pushbutton S3 is pressed while the SCR is conducting current, the flow is diverted around the SCR, and when the pushbutton is released, the flow through the SCR will not resume. Thus, the SCR can be shut down either by a normally closed pushbutton in series with it (which will interrupt the current), or a normally open pushbutton in parallel with it (which will divert the current).

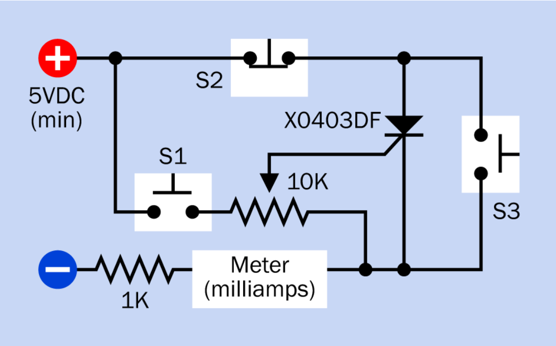

Figure 1-9. *In this test circuit, S1 triggers the SCR, while S2 or S3 will stop it. See text for additional details.*

The test circuit is shown installed on a breadboard in Figure 1-10. In this photograph, the red and blue wires supply a minimum of 5VDC. The two red buttons are tactile switches, the one at top left being S1 in the schematic while the one at bottom right is S3. The large switch with a rectangular button is S2; this is normally closed, and opens when pressed. The X0403DF SCR is just below it and to the right. The square blue trimmer is set to the midpoint of its range.

AC Current Applications

If the SCR is used with alternating current, it stops conducting during each negative cycle, and is retriggered in each positive cycle. This suggests one of its primary applications, as a controllable rectifier that can switch rapidly enough to limit the amount of current that passes through it during each cycle.

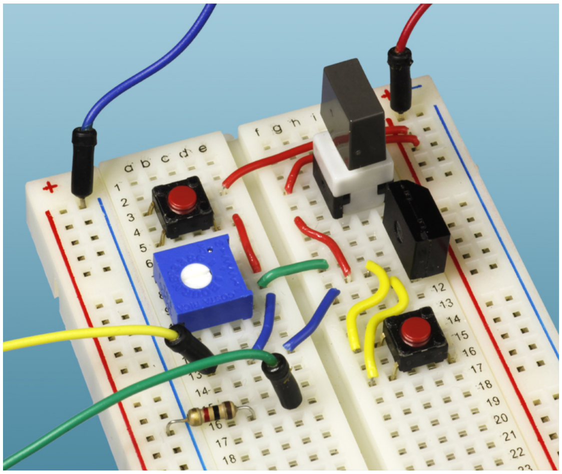

Figure 1-10. *A breadboarded version of the SCR test circuit. The two red buttons correspond with S1 and S3 in the schematic, while the large rectangular button at top right opens S2. See text for details.*

Variants

SCRs are available in surface-mount, through-hole, and stud packages, to handle increasing currents and voltages. Some special-purpose SCRs can control currents of hundreds of amps, while high-power SCRs are used to switch thousands of amps at more than 10,000V in power distribution systems. They are too specialized for inclusion in this encyclopedia.

Typical power ratings for SCRs in general use are summarized in the next section.

Values

Any SCR will impose a forward voltage drop, which typically ranges from around 1V to 2V, depending on the component.

Because SCRs are often used to modify AC waveforms, the current that the component can pass is usually expressed as the root mean square (RMS) of its peak value.

Commonly Used Abbreviations

- V_{DRM} Maximum repetitive forward voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{RRM} Maximum repetitive reverse voltage that may be applied to the anode while no voltage is applied to the gate (i.e., when the SCR is not in conductive mode).
- V_{TM} Maximum on-state voltage while the SCR is in conductive mode. T indicates that this value changes with temperature.
- V_{GM} Forward maximum gate voltage.
- V_{GT} Minimum gate voltage required to trigger.
- V_{GD} Maximum gate voltage that will not trigger.
- I_{DRM} Peak repetitive forward blocking current (i.e., maximum leakage).
- I_{RRM} Peak repetitive reverse blocking current (i.e., leakage in the off state).
- I_{GM} Maximum forward gate current.
- $I_{T(RMS)}$ Maximum RMS current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- $I_{T(AV)}$ Maximum average current between anode and cathode while the SCR is in conductive mode. T indicates that this value changes with temperature.
- I_{GT} Maximum gate current required to trigger.
- I_H Typical holding current.
- I_L Maximum latching current.
- T_C Case temperature, usually expressed as an acceptable range.
- T_J Operating junction temperature, usually expressed as an acceptable range.

Surface-mount variants may tolerate maximum anode-cathode currents that typically range from 1A to 10A. Maximum voltages as high as 500V are allowed in some cases. Leakage in the "off" state may be as high as 0.5mA or as low as 5μA. Gate trigger voltage is likely to range from 0.8V to 1.5V, and trigger current of 0.2mA to 15mA is typical.

Through-hole variants may be packaged in TO-92 format (like discrete transistors) or, more commonly, in TO-220 format (like a typical 1A voltage regulator). They may be rated for a maximum of 5A up to 50A, depending on the component, with maximum voltages ranging from 50V to 500V. Leakage is similar to surface-mount variants. The gate trigger voltage is typically around 1.5V, and trigger current ranges from 25–50mA.

A stud-type SCR may have a maximum 50A to 500A current rating, although some components are capable of tolerating even higher values. Maximum voltages of 50V to 500V are possible. Leakage is likely to be higher than in other formats, with 5mA to 30mA being common. The gate trigger voltage is typically 1.5 to 3V, and trigger current may range from around 50mA to 200mA.

How to Use It

Although other applications are possible, in practice SCRs have two main applications:

- Phase control, which interrupts each positive phase of an AC power supply. It can moderate the speed of a motor or the heat generated by a resistive load.
- Overvoltage protection. This can safeguard sensitive components in a circuit where there is a DC power supply.

SCRs are often incorporated in ground-fault circuit interruptors (although not usually as discrete components) and in automotive ignition systems.

Phase Control

Phase control is a convenient way to control or limit the AC power delivered to a load by abbreviating each pulse in the AC waveform. This is done by adjusting the gate voltage so that the SCR blocks the first part of each positive phase, then conducts the remainder, and then stops conducting below its holding level. The SCR will then block the reversed flow in the negative phase of the AC waveform, but an additional SCR with opposite polarity can be added.

This is a form of *pulse-width modulation*. It is highly efficient, as the effective internal resistance of the SCR is either very high or very low, and the component does not waste significant energy in the form of heat.

On a graph showing the fluctuating voltage of an AC waveform, a single cycle is customarily divided into four stages: (1) zero voltage, (2) maximum positive voltage, (3) zero voltage, (4) minimum negative voltage, all measurements being made between the live side of the supply and the neutral side of the supply.

The cycle then repeats. Its transitions are often referred to as *phase angles* of 0 degrees, 90 degrees, 180 degrees, and 270 degrees, as shown in Figure 1-11.

The fluctuating voltage in an AC power supply is proportional with the sine of the phase angle. This concept is illustrated in Figure 1-12. If an imaginary point (shown as a purple dot) is moving in a circular path counterclockwise at a constant speed, its vertical distance (shown in green) above or below the X axis (horizontal centerline) can represent an AC voltage corresponding with the angle (shown as purple arcs) of the circle radius to the point, each angle being measured from at the center relative to a start position at right on the X axis.

When an SCR is used for phase control, the point at which it starts to conduct may be anywhere from 0 to almost 180 degrees. This is achieved by diverting a small amount of the AC power into an RC network attached to the gate of the SCR, as shown in Figure 1-13. The capacitor in this schematic introduces a delay that can be varied by the potentiometer. This enables the SCR to be triggered even after the peak of the AC power signal. In Figure 1-14, the AC power is shown by the center (green) curve, and the slightly delayed, reduced voltage at the gate is shown by the upper, purple curve. When the gate voltage rises to the trigger level, it causes the SCR to begin conducting current, creating an abbreviated output shown in the bottom curve. In this way, triggering from an AC phase angle of 0 degrees to almost 180 degrees is possible. The phase angle where the SCR begins to allow conduction is known as the *conduction angle*.

If two SCRs with opposite polarity are placed in parallel with each other, they can be used to provide phase control on both the positive-going and negative-going sections of an AC cycle. This configuration is used in high-powered devices. A **triac** is used for the same purpose with lower current.

Six SCRs may be used to control three-phase power.

Overvoltage Protection

The tolerance of an SCR for high current makes it suitable for use in a crowbar voltage limiting circuit.

In Figure 1-15, the SCR does not conduct current (other than a small amount of leakage) until the Zener diode senses a voltage above the maximum level considered safe. The diode then allows power to reach the gate of the SCR. Its impedance drops immediately, and the resulting surge of current trips the fuse. After the cause of the overvoltage condition is corrected, the fuse can be replaced and the circuit may resume functioning.

A capacitor is included so brief spikes in the power supply will be passed to ground without triggering the SCR. A resistor of around 100Ω ensures that the gate voltage of the SCR remains near zero during normal operation. When the Zener diode starts to conduct current, the resistor acts as a voltage divider with the diode, so that sufficient voltage reaches the SCR to activate it.

This circuit may be unsuitable for low-voltage power supplies, because the Zener diode has to be chosen with a high enough rating to prevent small power fluctuations from tripping it. Bearing in mind that the real triggering voltage of the diode may be at least plus-or-minus 5% of its rated voltage, the diode may have to be chosen with at least a 6V rating in a 5V circuit, and it may not be activated until the voltage is actually 6.5V. This may be insufficient to protect the components being used with the power supply.

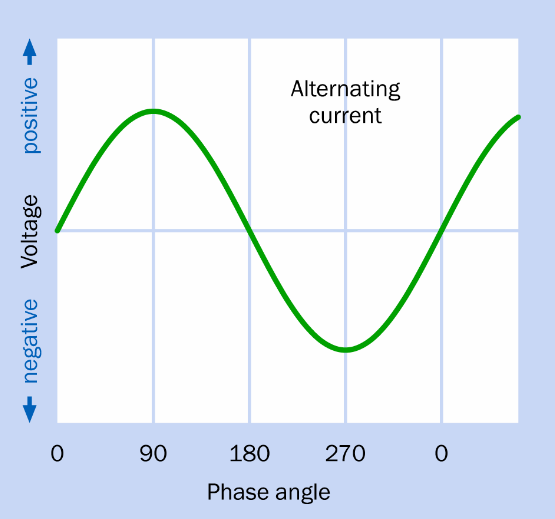

Figure 1-11. *An AC waveform is customarily measured in degrees of phase angle.*

What Can Go Wrong

Like other semiconductors, an SCR can be adversely affected by excessive heat. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, the values for triggering current and holding current are valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Unexpected Triggering Caused by Voltage

A very rapid increase in forward voltage at the anode can induce a triggering voltage in the gate by capacitive coupling. As a result, the SCR can trigger itself without any external application of gate voltage. This is sometimes known as *dv/dt triggering*. If necessary, a snubber circuit can be added across the anode input to prevent sudden voltage transitions.

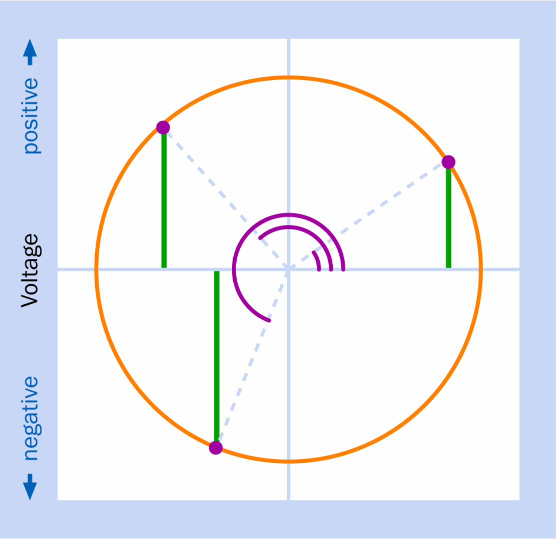

Figure 1-12. *The fluctuating voltage of an AC power supply (shown as vertical green lines) is proportional with the sines of the angles (purple arcs) in this diagram. The angles are referred to as phase angles.*

Confusion of AC and DC Ratings

The on-state current for an SCR is averaged only over the width of each pulse that the SCR actually conducts. It is not time-averaged over an entire AC cycle, and it will be different again from a DC rating. Care must be taken to match the current rating with the way in which the component will actually be used.

Maximum Current versus Conduction Angle

Current-carrying capability will be very significantly affected by the length of the duty cycle when the SCR is being used to abbreviate each positive AC pulse. When the SCR imposes a 120-degree conduction angle, it may be able to handle twice the average on-state current as when it is imposing a 30-degree conduction angle. The manufacturer's datasheet should include a graphical illustration of this relationship. If an SCR is chosen for a high conduction angle, and the angle is later reduced, overheating will result, and damage is likely.

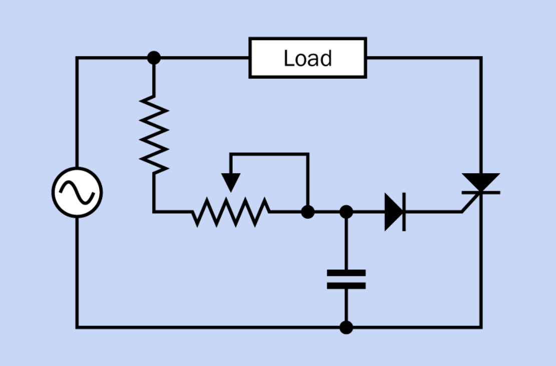

Figure 1-13. *In this schematic, an SCR is used to apply phase control, adjusting the power that passes through a load.*

Confusing Symbols

When reading a schematic, unfortunate errors can result from failure to distinguish between the symbol for a **programmable unijunction transistor (PUT)** and the symbol for an SCR. The characteristics of a PUT are described in Volume 1 of this encyclopedia.

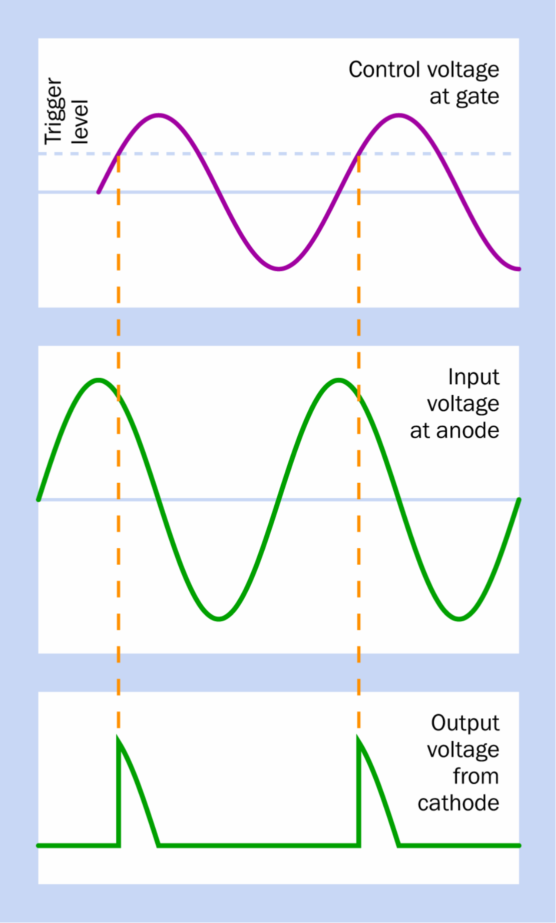

Figure 1-14. *If the AC power applied to the anode of an SCR (center) is reduced in voltage and delayed slightly by an RC network, it can trigger the SCR (top), causing it to pass only an abbreviated segment of each positive AC pulse (bottom).*

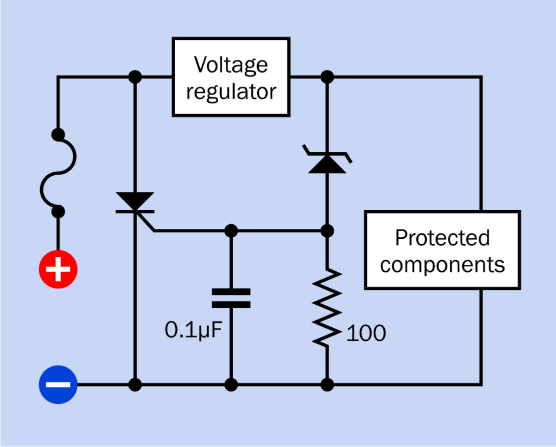
Voltage regulator
Protected components
0.1µF
100
+
−

Figure 1-15. *In this schematic, an SCR is used to provide crowbar overvoltage protection for sensitive components.*

diac

2

A **diac** is a self-triggering type of *thyristor*. Its name is said to be derived from the phrase "diode for AC," and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **triac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see Chapter 1)
- **triac** (see Chapter 3)

What It Does

The **diac** is a bidirectional thyristor with only two terminals. It blocks current until it is subjected to sufficient voltage, at which point its impedance drops very rapidly. It is primarily used to trigger a **triac** for purposes of moderating AC power to an **incandescent lamp**, a resistive heating element, or an AC **motor**. The two leads on a diac have identical function and are interchangeable.

By comparison, a **triac** and an **SCR** are thyristors with three leads, one of them being referred to as the gate, which determines whether the component becomes conductive. A triac and a diac allow current to flow in either direction, while an SCR always blocks current in one direction.

Symbol Variants

The schematic symbol for a diac, shown in Figure 2-1, resembles two diodes joined together, one of them inverted relative to the other. Functionally, the diac is comparable with a pair of Zener diodes, as it is intended to be driven beyond the point where it becomes saturated. Because its two leads are functionally identical, they do not require names to differentiate them. They are sometimes referred to as A1 and A2, in recognition that either of them may function as an anode; or they may be identified as MT1 and MT2, MT being an acronym for "main terminal."

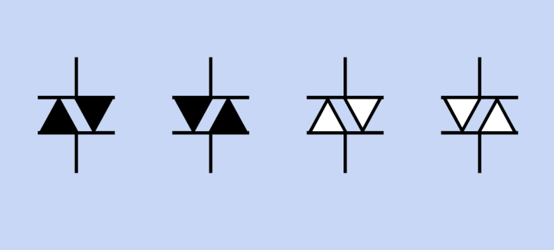

Figure 2-1. *Symbol variants to represent a diac. All four are functionally identical.*

The symbol may be reflected left to right, and the black triangles may have open centers. All of these variants mean the same thing. Occasionally the symbol has a circle around it, but this style is now rare.

When only a moderate voltage is applied (usually less than 30V) the diac remains in a passive state and will block current in either direction, although a very small amount of *leakage* typically occurs. When the voltage exceeds a threshold known as its *breakover level*, current can flow, and the diac will continue to conduct until the current falls below its *holding level*.

A sample diac is shown in Figure 2-2.

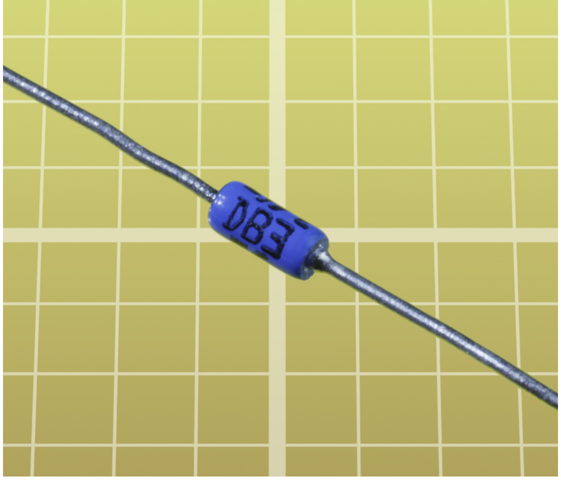

Figure 2-2. *Because a diac is not intended to pass significant current, it is typically packaged in a small format. The graph squares in the photograph each measure 0.1".*

How It Works

Figure 2-3 shows a circuit that demonstrates the conductive behavior of a diac.

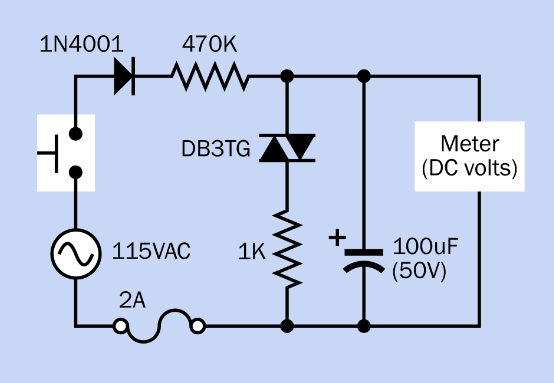

Figure 2-3. *A test circuit to demonstrate the behavior of a diac. See text for details.*

When the pushbutton is held down, current from the positive side of the AC supply flows through the diode and the 470K resistor to the capacitor. The diac is not yet conductive, so the capacitor accumulates a potential that can be monitored with the volt meter. After about 30 seconds, the charge on the capacitor reaches 32V. This is the breakover voltage for this particular diac, so it becomes conductive. The positive side of the capacitor can now discharge through the diac and the 1K series resistor to ground.

If the pushbutton is released at this moment, the meter will show that the capacitor discharges to a potential below the holding level of the diac. The capacitor now stops discharging because the diac has ceased being conductive.

If the pushbutton is held down constantly, the meter will show the capacitor charging and then discharging through the diac repeatedly, so that the circuit behaves as a *relaxation oscillator*. The 1K series resistor is included to protect the diac from excessive current. If a standard quarter-watt resistor is used, it should not become unduly warm because current passes through it only intermittently.

- Because this circuit uses 115VAC, basic precautions should be taken. The fuse should not be omitted, the capacitor should be rated for at least 50V, and the circuit should not be touched while it is connected to the power source. Breadboarding a circuit using this voltage requires caution and experience, as wires can easily come loose, and components can be touched accidentally while they are live.

Figure 2-4 shows the test circuit on a breadboard. The red and blue leads at the top of the photograph are from a fused 115VAC power supply. The live side of the supply passes through a diode to a pushbutton switch that has a rectangular black cap. A 470K resistor connects the other side of the switch to the positive side of a 100µF electrolytic capacitor, and also to the diac (small blue component). A 1K resistor connects the other end of the diac back to the negative side of the capacitor, which is grounded. The yellow and blue wires leaving the photograph at the left are connected with a volt meter, which is not shown.

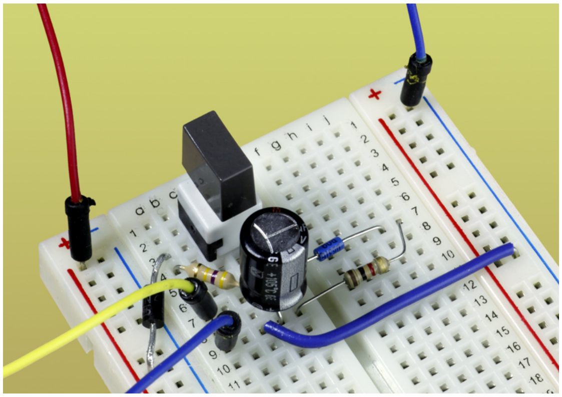

Figure 2-4. *A breadboarded version of the diac test circuit. See text for details.*

The behavior of a diac is also illustrated in Figure 2-5, which can be compared with the curves in Figures 3-10 and 1-8, depicting the behavior of a triac and an SCR respectively.

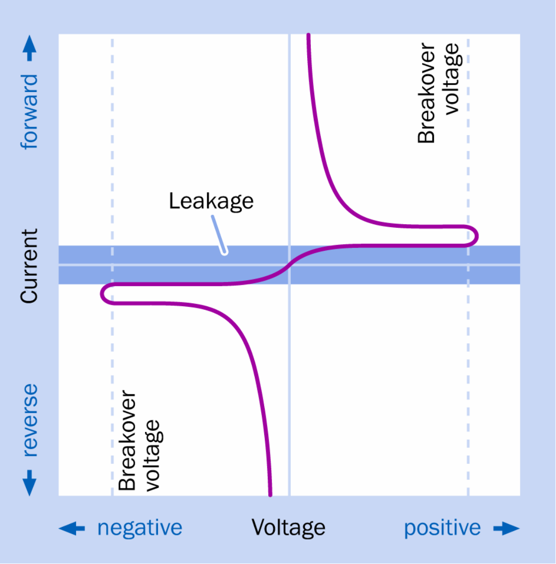

Figure 2-5. *The curve shows current passing through a diac when various voltages are applied.*

Switching AC

The diac cannot function as a switch, because it lacks the third terminal which is found in a triac, an SCR, or a bipolar transistor. However, it is well suited to drive the gate of a triac, because the behavior of a diac is symmetrical in response to opposite voltages, while the triac is not. If an AC voltage applied to a diac is adjusted with a potentiometer in an RC circuit, the diac will pass along a portion of each positive or negative pulse, and will delay it by a brief amount of time determined by the value of the capacitor in the RC circuit and the setting of the potentiometer. This is known as *phase control*, as it controls the *phase angle* at which the diac allows current to flow.

See Figure 3-13 for a schematic showing a diac driving a triac. See Figures 1-14 and 3-11 for graphs illustrating phase control. See "Phase Control" for a discussion of phase in AC waveforms.

Variants

Diacs are available in through-hole and surface-mount formats. Because they are not intended to handle significant current, no heat sink is included.

A *sidac* behaves very similarly to a diac, its name being derived from “silicon diode for alternating current.” Its primary difference from generic diacs is that it is designed to reach its breakover voltage at a higher value, typically 120VAC or 240VAC.

Values

When performing its function to trigger a triac, a diac is unlikely to pass more than 100mA.

The breakover voltage of a diac is usually between 30V and 40V, with a few versions designed for up to 70V. When the diac starts to conduct, its on-state impedance is sufficient to reduce the voltage significantly, with 5V being a typical minimum output voltage.

Although the rise time when a diac responds is very brief (around 1µs), the component is not expected to run at a high frequency. It will normally be used with 50Hz or 60Hz AC to trigger a triac. For this reason, its repetitive peak on-state current is usually specified at no more than 120Hz.

Abbreviations in datasheets are likely to include:

- V_{BO} Breakover voltage (sometimes may be specified as latching voltage, which for a diac is the same thing).
- V_{BO1} - V_{BO2} Breakover voltage symmetry. The hyphen is intended as a minus sign, so that this value is the maximum difference between breakover voltage in each direction.
- V_O Minimum output voltage.
- I_{TRM} Repetitive peak on-state current.
- I_{BO} Breakover current, usually the maximum required, and less than 20µA.
- I_R Maximum leakage current, usually less than 20µA.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a diac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for breakover current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

A higher breakover voltage will be required by a diac operating at low temperatures, although the variation is unlikely to be greater than plus-or-minus 2% within a normal operating range. Temperature has a much more significant effect on a triac.

Manufacturing Tolerances

The breakover voltage for a diac is not adjustable, and may vary significantly between samples of the component that are supposed to be identical. The diac is not intended to be used as a precision component. In addition, while its breakover voltage should be the same in either direction, a difference of plus-or-minus 2% is possible (1% in some components).

triac

3

A **triac** is a gate-triggered type of *thyristor*. Its name was probably derived from the phrase "triode for AC," and because it is not an acronym, it is not usually capitalized.

A thyristor is defined here as a semiconductor having four or more layers of p-type and n-type silicon. Because the thyristor predated integrated circuits, and in its basic form consists of a single multilayer semiconductor, it is categorized as a discrete component in this encyclopedia. When a thyristor is combined with other components in one package (as in a **solid-state relay**) it is considered to be an integrated circuit.

Other types of thyristor are the **SCR** (silicon-controlled rectifier) and the **diac**, each of which has its own entry in this encyclopedia.

Thyristor variants that are not so widely used, such as the *gate turn-off thyristor (GTO)* and *silicon-controlled switch (SCS)*, do not have entries here.

OTHER RELATED COMPONENTS

- **SCR** (see Chapter 1)
- **diac** (see Chapter 2)

What It Does

The **triac** is ubiquitous in AC dimmers for **incandescent lamps**. It is also used to control the speed of AC motors and the output of resistive heating elements. It is a type of *thyristor* which contains five segments of p-type and n-type silicon and has three leads, one of them attached to a gate that can switch a bidirectional flow of current between the other two. Its name was originally a trademark, generally thought to be derived from the phrase "triode for AC." A triode was a common type of vacuum tube when thyristors were first introduced in the 1950s.

By comparison, a **diac** is a thyristor with only two leads, allowing current to flow in either direction when the component reaches a *breakover voltage*. Its name was probably derived from the phrase "diode for AC." It is often used in conjunction with a triac.

An **SCR** (silicon-controlled rectifier) is a thyristor that resembles a triac, as it has three leads, one of them a gate. However, it only allows current to flow in one direction.

Symbol Variants

The schematic symbol for a triac, shown in Figure 3-1, resembles two diodes joined together, one of them inverted relative to the other. While a triac does not actually consist of two diodes, it is functionally similar, and can pass current in either direction.

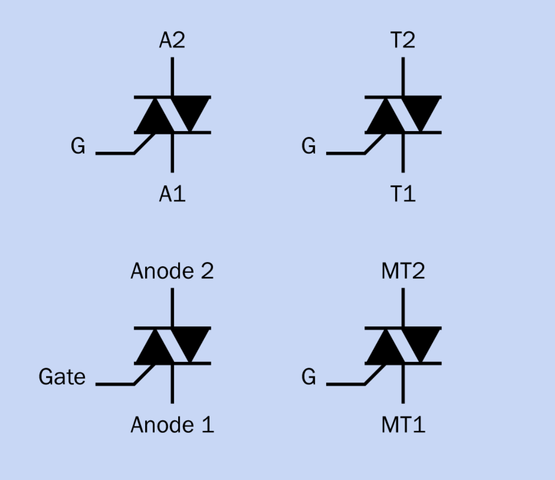

Figure 3-1. *The schematic symbol for a triac, with four naming conventions that are used for its leads. The different conventions do not indicate any functional difference.*

An appended bent line represents the gate. The labels for the other two leads are not standardized, and can be referred to as A1 and A2 (for Anode 1 and Anode 2), or T1 and T2 (for Terminal 1 and Terminal 2), or MT1 and MT2 (for Main Terminal 1 and Main Terminal 2). The choice of terms does not indicate any functional difference. In this encyclopedia entry, A1 and A2 are used.

The A1 terminal (or T1, or MT1) is always shown closer to the gate than A2 (or T2, or MT2). This distinction is important because although the triac can pass current in either direction, its behavior is somewhat asymmetrical.

- Voltages are expressed relative to terminal A1 (or T1, or MT1, if those terms are used).

The schematic symbol may be reflected or rotated, the black triangles may have open centers, and the placement of the bent line representing the gate may vary. However, terminal A1 is always nearer to the gate than terminal A2.

Figure 3-2 shows 12 of the 16 theoretical possibilities. All of these variants are functionally identical. Occasionally the symbol has a circle around it, but this style is now rare.

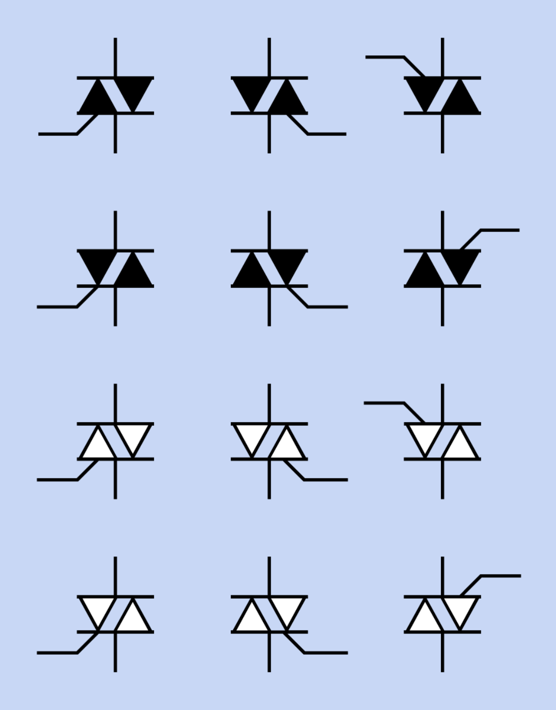

Figure 3-2. *Interchangeable variants of the schematic symbol for a triac.*

Triacs with various characteristics are shown in Figures 3-3, 3-4, and 3-5.

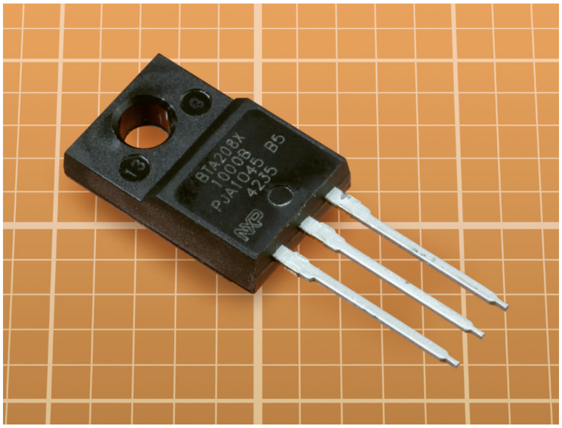

Figure 3-3. *The BTA208X-1000B triac can conduct 8A continuous on-state current RMS, and withstands peak off-state voltage of up to 1,000V. This is a "snubberless" triac.*

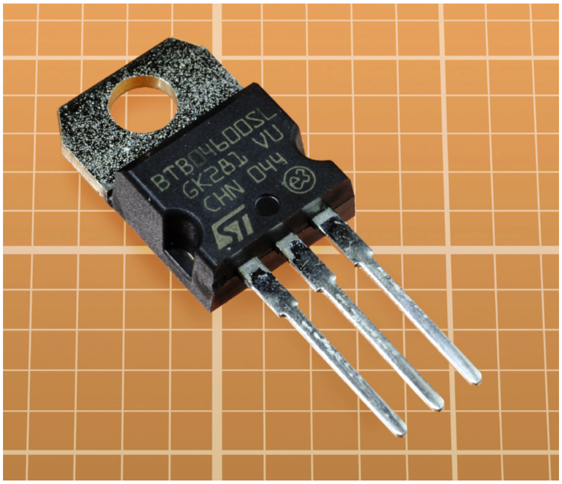

Figure 3-4. *The BTB04-600SL triac can conduct 4A continuous on-state current RMS, and withstands peak off-state voltage of up to 600V.*

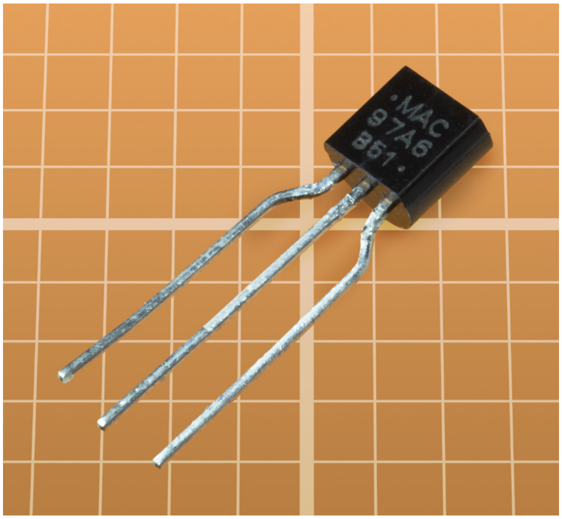

Figure 3-5. *The MAC97A6 triac can conduct 0.8A continuous on-state current RMS, and withstands peak off-state voltage of up to 400V.*

How It Works

When no gate voltage is applied, the triac remains in a passive state and will block current in either direction between A1 and A2, although a very small amount of *leakage* typically occurs. If the gate potential becomes sufficiently positive *or* negative relative to terminal A1, current can begin to flow between A1 and A2 in *either* direction. This makes the triac ideal for controlling AC.

Quadrants

While a gate voltage is applied, four operating modes are possible. In each case, A1 is the reference (which can be thought of as being held at a neutral ground value). Because the triac is conducting AC, voltages above and below ground will occur. The four modes of operation are often referred to as four *quadrants*, and are typically arranged as shown in Figure 3-6.

In some reference sources (especially educational text books), current is shown with an arrow indicating a flow of electrons moving from negative to positive. Because the type of current flow is often undefined, diagrams should be interpreted carefully. In this encyclopedia, current is always shown flowing from a more-positive location to a more-negative location.

Quadrant 1 (upper right)

A2 is more positive than A1, and the gate is more positive than A1. Conventional current (positive to negative) will flow from A2 to A1. (This behavior is very similar to that of an **SCR**.)

Quadrant 2 (upper left)

A2 is more positive than A1, and the gate is more negative than A1. Once again, conventional current (positive to negative) will flow from A2 to A1.

Quadrant 3 (lower left)

A2 is more negative than A1, and the gate is more negative than A1. Conventional current is reversed from A1 to A2.

Quadrant 4 (lower right)

A2 is more negative than A1, but the gate is more positive than A1. Conventional current is reversed from A1 to A2.

- Note that two positive symbols or two negative symbols in Figure 3-6 do not mean that both locations are of equal voltage. They simply mean that these locations are at potentials that are significantly different from A1.

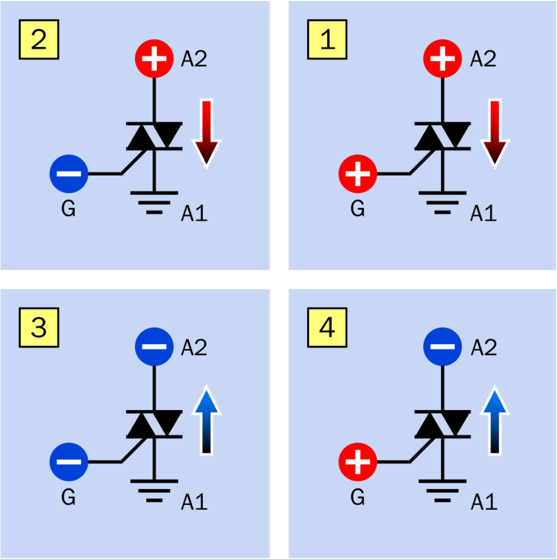

Figure 3-6. *The "quadrants" of triac behavior. Positive and negative symbols indicate which terminal is "more positive" or "more negative" than A1. The ground symbol represents a potential midway between positive and negative. See text for details.*

Suppose that gate current increases gradually. When it reaches the *gate threshold current* of the triac, the component starts conducting between A1 and A2. If the current between A1 and A2 rises above the value known as the *latching current*, it will continue to flow, even if gate current disappears completely.

If the self-sustaining current through the triac gradually diminishes, while there is no voltage applied to the gate, conduction between the main terminals will stop spontaneously when it falls below a level known as the *holding current*. This behavior is similar to that of an SCR. The triac now returns to its original state, blocking current until the gate triggers it again.

The triac is sufficiently sensitive to respond to rapid fluctuations, as in 50Hz or 60Hz AC.

Threshold, Latching, and Holding Current

Figure 3-7 shows the relationship between the gate threshold current, the latching current, and holding current. In the upper half of the figure, gate current is shown fluctuating until it crosses the threshold level. This establishes current flow between the main terminals, shown in the lower half of the figure. Prior to this moment, a very small amount of leakage current occurred (shown in the figure, but not to scale).

In this hypothetical scenario, the triac starts passing current between external components—and the current exceeds the latching level. Consequently, gate current can diminish to zero, and the triac remains conductive. However, when external factors cause the current between the main terminals to diminish below the holding level, the triac abruptly ceases to be conductive, and current falls back to the leakage level.

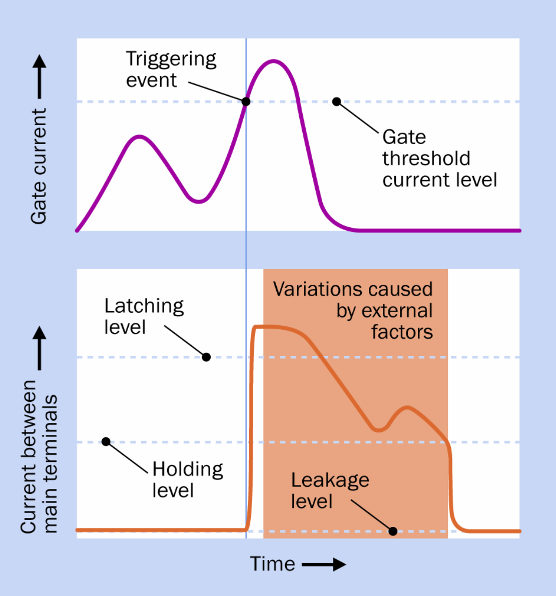

Figure 3-7. *The relationship between gate current of a triac and the current between its main terminals. See text for details.*

Unlike a **bipolar transistor**, a triac is either "on" or "off" and does not function as a *current ampli-*

fier. When it has been triggered, the impedance between A1 and A2 is low enough for heat dissipation to be manageable even at relatively high power levels.

Triac Testing

Figure 3-8 shows a circuit which can demonstrate the conductive behavior of a triac. For simplicity, this circuit is DC powered. In a real application, the triac is almost always used with AC.

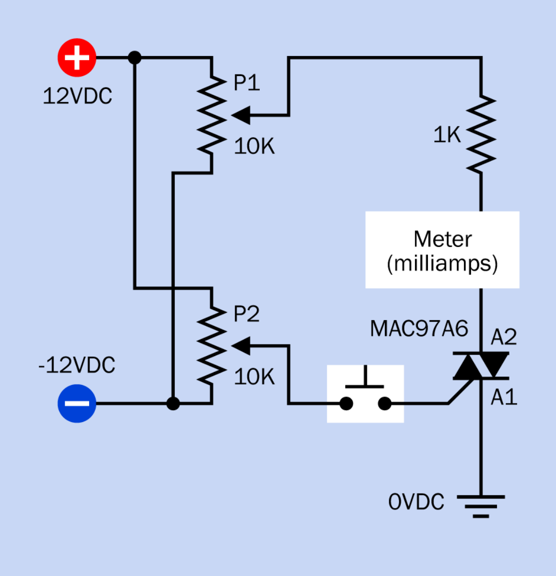

Figure 3-8. *A test circuit to show the behavior of a triac when varying positive and negative potentials are applied to the gate and to the A2 terminal, relative to A1.*

Note that this circuit requires at least a +12VDC and -12VDC power supply (higher values may also be used). The ground symbol represents a midpoint voltage of 0VDC, applied to terminal A1 of the triac, which is an MAC97A6 or similar. If a dual-voltage power supply is unavailable, the gate of the triac can be connected directly to +12VDC, omitting potentiometer P2; but in this case, only two operating modes of the triac can be demonstrated by turning potentiometer P1.

Each potentiometer functions as a voltage divider between the positive and negative sides of the power supply. P1 applies a positive or negative voltage to A2, relative to A1. P2 applies a positive or negative voltage to the gate, relative to A1.

If the test begins with both potentiometers at the top ends of their range, A1 and G both have a positive potential relative to A1, so that the triac is now in quadrant 1 of its operating modes. Pressing the pushbutton should cause it to start conducting current limited by the 1K resistor, and the meter should change from measuring 0mA to around 12mA. If the pushbutton is released, the triac should continue to conduct current, because 12mA is above this triac's latching level. If P1 is slowly moved toward the center of its range, the current diminishes, ceasing when it falls below the holding level. If P1 is now moved back to the top of its range, the current will not resume until the triac is retriggered with the pushbutton.

The test can be repeated with P1 at the top of its range and P2 at the bottom of its range, to operate the triac in quadrant 2; P1 at the bottom of its range and P2 at the bottom of its range, to operate the triac in quadrant 3; and P1 at the bottom of its range and P2 at the top of its range, to operate the triac in quadrant 4. The functionality should be the same in each case. The pushbutton will initiate a flow of current, which will diminish when P1 is turned toward the center of its range.

In any of these quadrants, P2 can be turned slowly toward the center of its range while the pushbutton is pressed repeatedly. This will allow empirical determination of the gate threshold current for this triac. The meter, measuring milliamps, will measure the current if it is inserted between the wiper of the potentiometer and the gate of the triac.

The test circuit is shown installed on a breadboard in Figure 3-9. In this photograph, the red and blue wires at left supply +12VDC and -12VDC relative to the black ground wire at top right. The yellow and green wires connect with a meter set to measure milliamps. The red button is a tactile switch, while the MAC97A6 triac is just above it and to the left. The square blue 10K trimmers are set to opposite ends of their scales, so that the meter will show current flowing when the tactile switch is pressed.

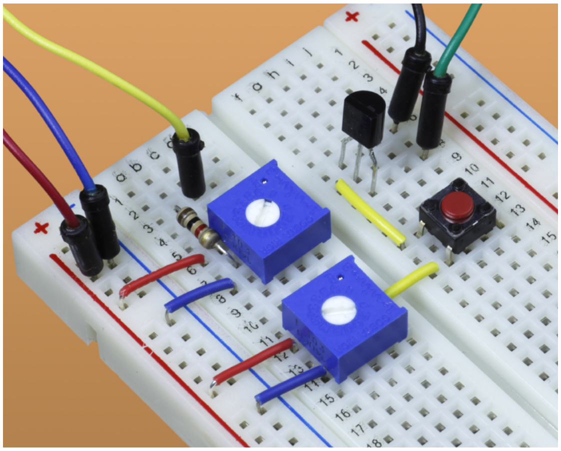

Figure 3-9. *A breadboarded triac test circuit.*

Breakover Voltage

If a much higher voltage is applied to A2, the triac can be forced to conduct current without any triggering voltage being applied to the gate. This occurs when the potential between A1 and A2 reaches the triac's *breakover voltage*, although the component is not designed to be used this way. The concept is illustrated in Figure 3-10, which can be compared with the behavior of an SCR illustrated in Figure 1-8 and the behavior of a diac shown in Figure 2-5. While the term *breakdown voltage* defines the minimum reverse voltage required to force a diode to conduct, *breakover voltage* refers to the minimum forward voltage that has this effect. Because a triac is designed to conduct in both directions, it can be thought of as having a breakover voltage in both directions.

In Figure 3-10, the numbers in yellow squares are the quadrants referred to in Figure 3-6. The solid curve represents current flow if a triggering voltage is applied to the gate while a positive or negative potential is applied to A2, relative to A1. If the gate is not triggered while the voltage between A1 and A2 gradually increases, the dashed section of the curve illustrates the outcome when the component reaches breakover voltage. Although this may not damage the triac, the component becomes uncontrollable.

- In normal usage, the voltage between A1 and A2 should not be allowed to reach breakover level.

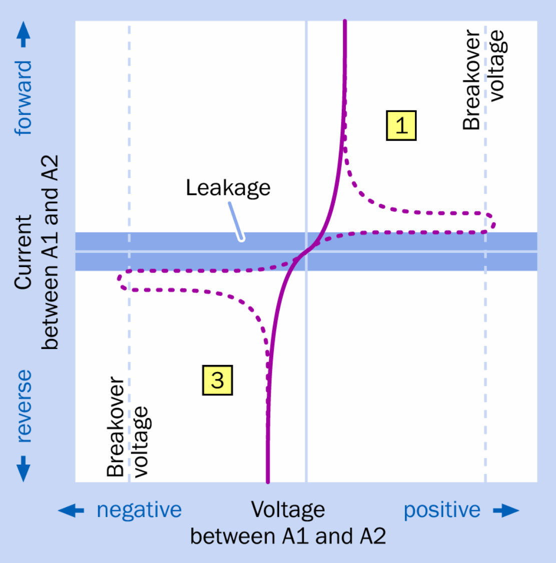

Figure 3-10. *The solid curve shows current passing between A1 and A2 in a hypothetical triac, for varying voltages, while triggering voltage is applied to the gate. The dashed curve assumes that no triggering voltage is applied to the gate. The numbers in yellow squares are the quadrants of triac operation.*

Switching AC

"Switching" AC with a triac means interrupting each pulse of current so that only a portion of it is conducted through to the load. Usually this is done with the triac functioning in quadrants 1 and 3. In quadrant 3, the polarity of the flow between A1 and A2 is opposite to that in quadrant 1, and the gate voltage is also reversed. This enables a relatively simple circuit to control the duration of each half-cycle passing through the triac. The theory of this circuit is shown in Figure 3-11.

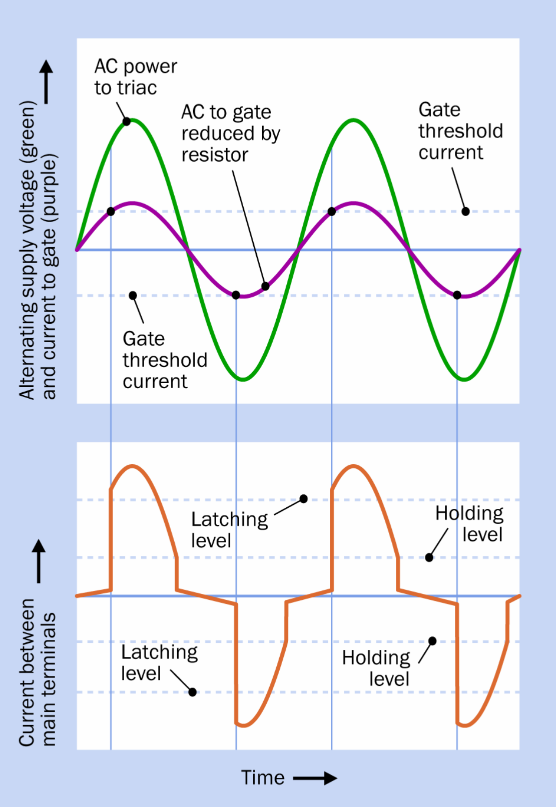

Figure 3-11. *To moderate the power of AC current, a triac can block a section of each AC pulse.*

The upper section of Figure 3-11 shows alternating voltage to the triac in green. The purple curve represents the gate current of the triac, reduced by a variable resistor. (The figure is for conceptual purposes only; the alternating power supply voltage and the fluctuating gate current cannot actually share the same vertical scale of a graph.)

Figure 3-11 can be compared with Figure 3-7, except that the negative threshold level for the gate is now shown as well as the positive threshold level. Remember, either a positive or negative voltage can activate the gate.

In Figure 3-11, initially the triac is nonconductive. As time passes, the gate current reaches the threshold level, and this triggering event enables current to flow between the main terminals of the triac, as shown in the lower part of the figure. This current exceeds the latching level, so it continues to flow, even though the gate current diminishes below its threshold level. Finally the current between the main terminals falls below the holding level, at which point the triac stops conducting. It waits for the next triggering event, which occurs as the power supply swings to negative.

This simple system blocks a section of each AC pulse, which will vary in length depending how much current is allowed to flow through the gate. Because the blocking process occurs rapidly, we notice only the reduced overall power passing through the triac (in terms of the brightness of a light, the heat emitted by a resistive element, or the speed of a motor).

Unfortunately, there is a problem in this scenario: the triac does not quite behave symmetrically. Its gate threshold level for positive current is not exactly equal and opposite to its gate threshold level for negative current. The upper part of Figure 3-11 shows this flaw in the differing vertical offsets of the positive and negative thresholds from the central zero line.

The result is that negative AC pulses through the triac are shorter than positive pulses. This asymmetry produces harmonics and noise that can feed back into power supply wiring, interfering with other electronic equipment. The actual disparities in gate response, in each quadrant of operation for two triacs, are shown in Figure 3-12.

Current capacity of triac	Ratio of gate current required to conduct, relative to quadrant 1		
	Quadrant 2	Quadrant 3	Quadrant 4
4 amp	1.6	2.5	2.7
10 amp	1.5	1.4	3.1

Figure 3-12. *Because the internal structure of a triac is asymmetrical, it requires a different trigger current in each of its operating quadrants. This table, derived from a Littelfuse technical briefing document, shows the ratio of the minimum trigger current in quadrants 2, 3, and 4 relative to quadrant 1.*

See Figure 1-14 for a graph illustrating phase control in the SCR. See "Phase Control" for a discussion of phase in AC waveforms generally.

Triac Triggered by a Diac

The problem of asymmetrical triggering can be overcome if the triac is triggered with a voltage pulse generated by another component that does behave symmetrically. The other component is almost always a diac, which is another type of thyristor. Unlike an SCR or a triac, it has no gate. It is designed to be pushed beyond its breakover voltage, at which point it latches and will continue to conduct until current flowing through it diminishes below its holding level. See Chapter 2 for more information about the diac.

In Figure 3-13, the diac is shown to the right of the triac, and is driven by a simple RC network consisting of a fixed resistor, a potentiometer, and a capacitor. (In an actual application, the RC network may be slightly more complex.) The capacitor takes a small amount of time to charge during each half-cycle of AC. The length of this delay is adjusted by the potentiometer, and determines the point in each AC half-cycle when the voltage to the diac reaches breakover level. Because the delay affects the phase of the AC, this adjustment is known as *phase control*.

As the voltage exceeds breakover level, the diac starts to pass current through to the gate of the triac, and triggers it. The holding level of the diac is lower than its latching level, so it continues to pass current while the capacitor discharges and the voltage diminishes. When the current falls below the holding level, the diac stops conducting, ready for the next cycle. Meanwhile, the triac continues to pass current until the AC voltage dips below its holding level. At this point, the triac becomes nonconductive until it is triggered again.

This chopped waveform will still create some harmonics, which are suppressed by the coil and capacitor at the left side of the circuit in Figure 3-13.

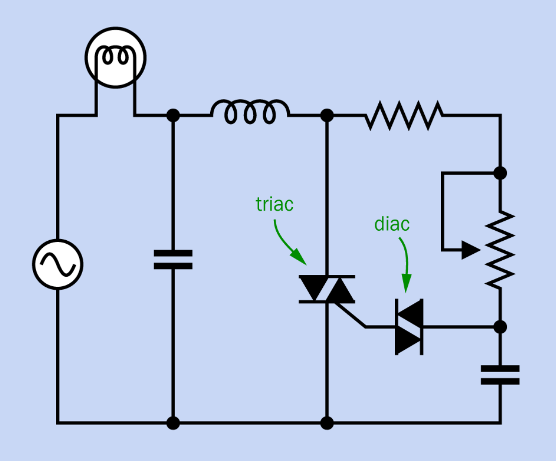

Figure 3-13. *A minimal schematic showing typical operation of a triac, with a diac supplying pulses to the triac gate. The potentiometer adjusts the delay created by the capacitor.*

Other Triac Drivers

It is possible, although unusual, to drive a triac from a source other than a diac.

Simple on-off control can be achieved by using a special optocoupler such as the MOC3162 by Fairchild Semiconductor. This emits a switching signal to a triac only when the AC voltage passes through zero. A *zero cross* circuit is desirable because it creates much less interference. The use of an optocoupler helps to isolate the triac from other components.((("zero cross circuit"))

Phase control can be achieved using an optocoupler such as the H11L1, which can be driven by rectified but unsmoothed AC after it passes through a Zener diode to limit the voltage. The output from the optocoupler is logic-compatible and can be connected with the input to a timer such as the 555, set to one-shot mode. Each pulse from the timer passes through another optocoupler such as the MOC3023, which uses an internal LED to trigger the gate of a triac.

Yet another possibility is to use the programmed output from a microcontroller, through an optocoupler, to control the gate of a triac. An online search for the terms "microcontroller" and "triac" will provide some additional suggestions.

Charge Storage

While switching AC, the internal charge between A1 and A2 inside the triac requires time to dissipate before the reverse voltage is applied; otherwise, *charge storage* occurs, and the component may start to conduct continuously. For this reason, the triac is normally restricted to relatively low frequencies such as domestic 60Hz AC power.

When a triac controls a motor, the phase lag between voltage and current associated with an inductive load can interfere with the triac's need for a transitional moment between a positive and negative voltage cycle. In a datasheet, the term *commutating dv/dt* defines the rate of rise of opposite polarity voltage that the triac can withstand without locking into a continuous-on state.

An RC *snubber network* is often wired in parallel with A1 and A2 to control the rise time of voltage to the triac, as shown within the darker blue rectangle in Figure 3-14, where a resistor and capacitor have been added just to the left of the triac. The highest resistance and lowest capacitance, consistent with trouble-free operation, should be chosen. Typical values are 47Ω to 100Ω for the resistor, and 0.01µF to 0.1µF for the capacitor.

Variants

Triacs are available in through-hole and surface-mount packages.

Some components that are referred to as triacs actually contain two SCR components of opposite polarity. The "alternistor" range from Littelfuse is an example. The SCR will tolerate faster voltage rise times than a conventional triac, and is more suitable for driving inductive loads such as large motors.

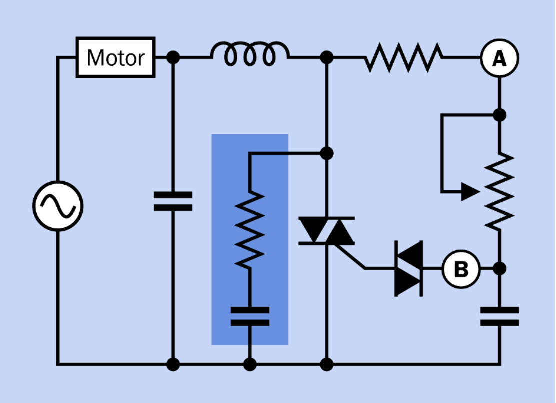

Figure 3-14. *To prevent a triac from locking itself into a continuous-on state while driving an inductive load such as a motor, a snubber circuit can be added (shown here as a resistor and capacitor in the darker blue rectangle to the left of the triac).*

A *snubberless triac*, as its name implies, is designed to drive an inductive load without need for a snubber circuit. An example is the STMicroelectronics BTA24. Datasheets for this type of component impose some limits that may be stricter than for a generic triac.

Values

Surface-mount triacs are typically rated between 2A to 25A of switched AC current (RMS), the higher-current versions being as large as 10mm square. The necessary gate trigger voltage may range from 0.7V to 1.5V. Through-hole packages may be capable of slightly higher currents (up to 40A), with gate trigger voltages of 1V to 2.5V being common.

As noted previously, the majority of triacs are restricted to relatively low frequency switching, 60Hz being very common.

Abbreviations in datasheets are likely to include:

- V_{DRM} or V_{RRM} Peak repetitive reverse off-state voltage. The maximum reverse voltage that the component will withstand in its "off" state without experiencing damage or allowing current to pass.

- V_{TM} The maximum voltage difference between A1 and A2, measured with a short pulse width and low duty cycle.
- V_{GT} Gate trigger voltage necessary to produce the gate trigger current.
- I_{DRM} Peak repetitive blocking current (i.e., maximum leakage).
- I_{GM} Maximum gate current.
- I_{GT} Minimum gate trigger current.
- I_H Holding current.
- I_L Latching current.
- $I_{T(RMS)}$ On-state RMS current. The maximum value passing through the component on a continuous basis.
- I_{TSM} Maximum non-repetitive surge current. Specified at a stated pulse width, usually 60 Hz.
- T_C Case temperature, usually expressed as an acceptable range.
- T_J Operating junction temperature, usually expressed as an acceptable range.

What Can Go Wrong

Like other semiconductors, a triac is heat sensitive. Usual precautions should be taken to allow sufficient ventilation and heat sinking, especially when components are moved from an open prototyping board to an enclosure in which crowding is likely.

Unexpected Triggering Caused by Heat

On a datasheet, a value for triggering current is valid only within a recommended temperature range. A buildup of heat can provoke unexpected triggering.

Low-Temperature Effects

Significantly higher gate current will be required by a triac operating at low temperatures. It is quite possible that the component will need twice as much current at 25° C compared with 100° C, junction temperature. If the triac receives insufficient current, it will not turn on.

Wrong Type of Load

If an **incandescent lamp** is replaced with a **fluorescent light** or **LED area lighting**, a preexisting triac may no longer work as a dimmer. Fluorescent lamps will have some inductance, and may also provide a capacitive load, either of which will interfere with the normal behavior of a triac.

The light output of an LED varies very differently compared with the light output of an incandescent bulb, in response to reduction in power. Therefore an LED should be dimmed using pulse-width modulation that is appropriate for its output characteristics. A triac is generally not suitable.

Wrongly Identified Terminals

A triac is often thought of as a symmetrical device, because it is designed to switch AC current using either positive or negative voltage at the gate. In reality, its behavior is asymmetrical, and if it is installed “the wrong way around” it may function erratically or not at all.

Failure to Switch Off

As already noted (see “Charge Storage” on page 23), a triac will tend to suffer from *charge storage* if there is insufficient time between the end of one half-cycle and the beginning of the next. A component that works with a resistive load may cease to function if it is used, instead, to power an inductive load.

solid-state relay

4

A **solid-state relay** is less-commonly referred to by its acronym, *SSR*. It is sometimes regarded as an **optocoupler**, but in this encyclopedia the two components have separate entries. An optocoupler is a relatively simple device consisting of a light source (usually an **LED**) and a light sensor, in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1A.

A component that works like a solid-state relay but only switches a 5V (or lower) logic signal may be referred to as a *switch*, even though it is entirely solid-state. This type of component is included in this entry because it functions so similarly to a solid-state relay.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **optocoupler** (see Chapter 5)

What It Does

A **solid-state relay** (SSR) is a semiconductor package that emulates an electromagnetic **relay** (see Volume 1). It switches power on or off between its output terminals in response to a smaller current and voltage between its input terminals. Variants can switch AC or DC and may be controlled by AC or DC. An SSR functions as a *SPST switch*, and is available in normally open or normally closed versions. SSRs that function as an SPDT switch are relatively unusual and actually contain more than one SSR.

No single schematic symbol has been adopted to represent a solid-state relay, but some alternatives are shown in Figure 4-1:

Top
: An unusually detailed depiction of an SSR that switches DC current using MOSFETS. Symbols for this device often omit the diodes on the output side and may simplify the MOSFET symbols.

Bottom left
: An SSR that uses an internal **triac** to switch AC. The box labeled 0x indicates that this is a *zero-crossing* relay, meaning that it switches when alternating voltage crosses the 0V level from positive to negative or negative to positive.

Bottom right
: A generic SSR, showing a symbol for a normally open relay, although whether it is designed for AC or DC is unclear.

Advantages

- Great reliability and long life.
- No physical contacts that are vulnerable to arcing and erosion or (under extreme conditions) that could weld themselves together.

- Very fast response, typically 1μs on and 0.5μs off.
- Very low power consumption on the input side, as low as 5mA at 5VDC. Many solid-state relays can be driven directly from logic chips.
- Lack of mechanical noise.
- No contact bounce; a clean output signal.
- No coil that would introduce back EMF into the circuit.
- Safe with flammable vapors, as there is no sparking of contacts.
- Often smaller than a comparable electromagnetic relay.
- Insensitive to vibration.
- Safer for switching high voltages, as there is complete internal separation between input and output.
- Some variants work with input control voltages as low as 1.5VDC. Electromagnetic relays typically require at least 3VDC (or more, where larger relays are required to switch higher currents).

Disadvantages

- Less efficient; its internal impedance introduces a fixed-value voltage drop on the output side (although this may be negligible when switching higher voltages).
- Generates waste heat in its “on” mode, in accordance with the voltage drop.
- Passes some *leakage* current (usually measured in microamps) on the output side when the relay is supposed to be “off.”
- A DC solid-state relay usually requires observation of polarity on the output side. An electromagnetic relay does not.
- Brief voltage spikes on the input side, which would be ignored by a slower electromagnetic relay, may trigger a solid-state relay.
- More vulnerable than an electromagnetic relay to surges and spikes in the current that is switched on the output side.

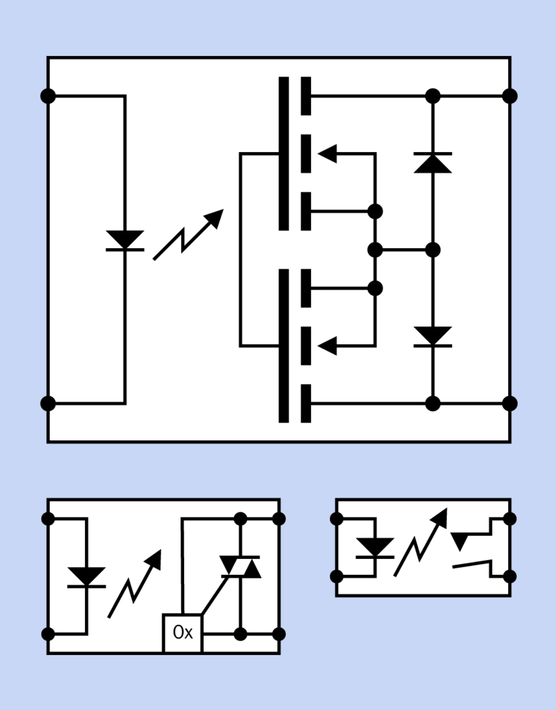

Figure 4-1. *Schematic symbols for solid-state relays have not been standardized. See text for details.*

How It Works

Almost all modern SSRs contain an internal **LED** (light-emitting diode, see Chapter 22) which is switched on by the control input. Infrared light from the LED is detected by a sensor consisting of one or more **phototransistors** or **photodiodes**. In a relay that controls DC current, the sensor usually switches a **MOSFET** (see Volume 1) or an **SCR** (silicon-controlled rectifier—see Chapter 1). In relays that control AC current, a **triac** (see Chapter 3) controls the output. Because the input side and the output side of the SSR are linked only by a light signal, they are electrically isolated from each other.

The MOSFETs require so little power, it can be provided entirely by light falling on an array of 20 or more photodiodes inside the SSR package.

Typical solid-state relays are shown in Figures 4-2 and 4-3.

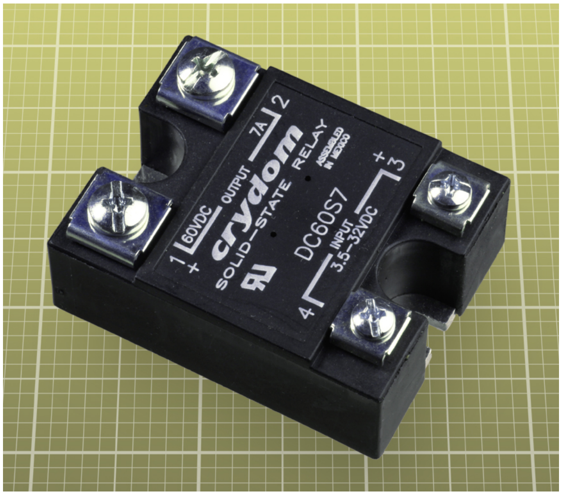

Figure 4-2. *A solid-state relay capable of switching up to 7A DC. See text for a detailed description.*

The Crydom DC60S7 accepts a control voltage ranging from 3.5VDC to 32VDC, with a typical input current of less than 3mA. Maximum turn-on time is 0.1ms and maximum turn-off time is 0.3ms. This relay can switch up to 7A and tolerates a surge of up to twice that current. It imposes a voltage drop of as much as 1.7VDC, which can become a drawback when switching voltages that are significantly lower than its maximum 60VDC. The electronics are sealed in thermally conductive epoxy, mounted on a metal plate approximately 1/8" thick which can be screwed down onto an additional heat sink.

The Crydom CMX60D10 tolerates a more limited range of control voltages (3VDC to 10VDC) and requires a higher input current of 15mA at 5VDC. However, its very low maximum on-state resistance of 0.018Ω imposes a much smaller voltage drop of less than 0.2 volts when passing 10A. This results in less waste heat and enables a *single-inline package* (SIP) without a heat sink. The CMX60D10 weighs 0.4 ounces, as opposed to the 3 ounces of the DC60S7. Relays from other manufacturers use similar packaging and have similar specifications.

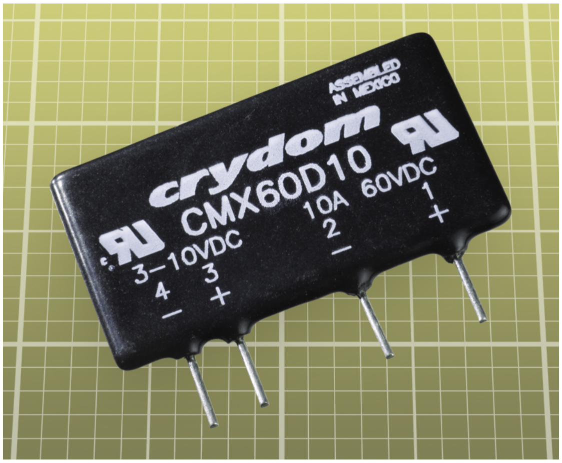

Figure 4-3. *A solid-state relay capable of switching up to 10A. Its lower internal resistance results in less waste heat and enables a smaller package. See text for a detailed description.*

Variants

Many solid-state relays have protective components built into the package, such as a varistor on the output side to absorb transients. Check datasheets carefully to determine how much protection from external components may be necessary when switching an inductive load.

Instantaneous versus Zero Crossing

A *zero crossing* SSR is one that (a) switches AC current and (b) will not switch "on" until the instant when the AC voltage crosses through 0V. The advantages of this type are that it does not have to be built to switch such a high current, and creates minimal voltage spike when the switching occurs.

All SSRs that are designed to switch AC will wait for the next voltage zero crossing before switching to their "off" state.

NC and NO Modes

Solid-state relays are SPST devices, but different models may have a normally closed or normally open output. If you require double-throw operation, two relays can be combined, one normally closed, the other normally open. See Figure 4-4. A few manufacturers combine a normally closed relay and a normally open relay in one package, to emulate a SPDT relay.

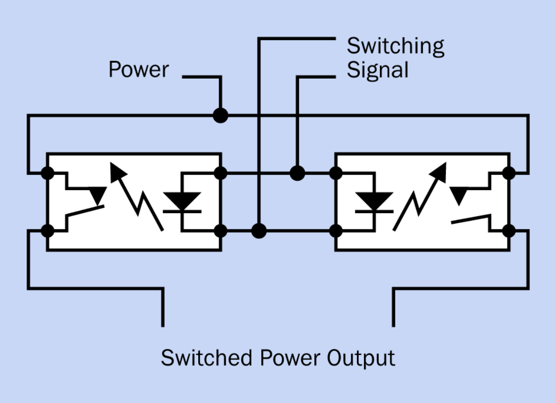

Figure 4-4. *A normally closed solid-state relay can be paired with a normally open solid-state relay to emulate a SPDT switch. This combination is available in a single package from some manufacturers.*

Packaging

High-current solid-state relays are often packaged with screw terminals and a metal base that is appropriate for mating with a *heat sink*. Some are sold with heat sinks integrated. Spade terminals and crimp terminals may be optional. The Crydom DC60S7 shown in Figure 4-2 is an example. This type of package may be referred to as *industrial mount.*

Lower-current solid-state relays (5A or less), and those with a very low output resistance, may be packaged with single-inline pins for through-hole mounting in circuit boards.

Solid-State Analog Switch

DIP packaging may be used for solid-state relays that are designed for compatibility with the low voltages and currents of logic chips. This type of component may be referred to simply as a *switch*. The 74HC4316 is an example, pictured in Figure 4-5.

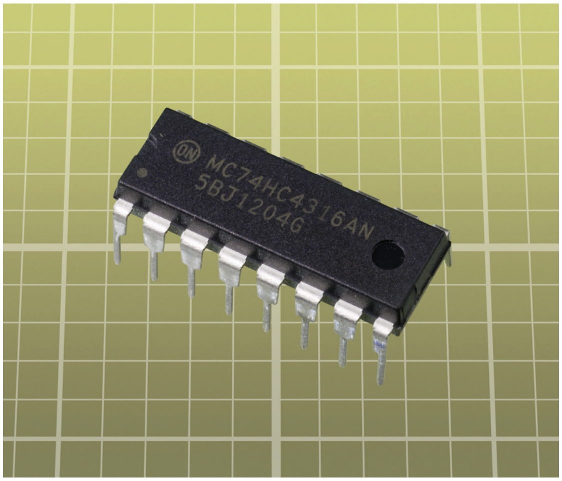

Figure 4-5. *This DIP package contains four "switches" that function as solid-state relays but are restricted to low voltages and currents, compatible with logic chips. See text for details.*

Typically the control voltage and the switched voltage are limited between +7V and −7V, with a maximum output current of 25mA. Each internal switch has its own Control pin, while an additional Enable pin forces all switches into an "off" state if its logic state is high. The simplified functionality of this component is illustrated in Figure 4-6, without showing internal optical isolation.

The "on" resistance of each internal pathway will be approximately 200Ω when the component is powered with +5VDC on the positive side and 0VDC on the negative side. This resistance drops to 100Ω if the negative power supply is -5VDC.

If all of the outputs from the chip are shorted together, it functions as a **multiplexer** (see Chapter 16). In fact, this type of switch component is often listed in catalogs as a multiplexer, even though it has other applications.

Because the component tolerates equal and opposite input voltages, it is capable of switching AC.

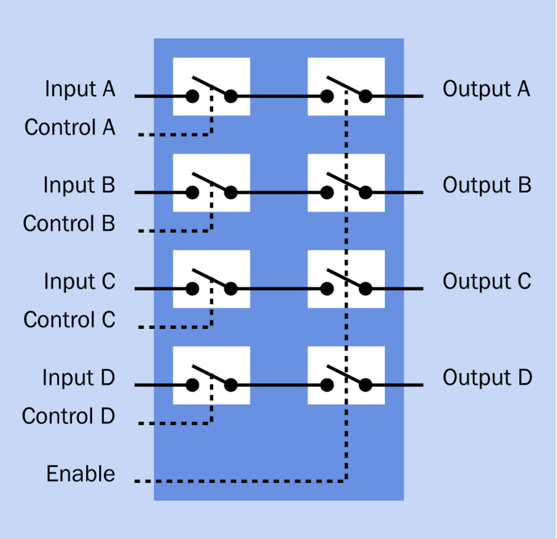

Figure 4-6. *The functionality of a chip containing four solid-state analog switches. A high state on a Control pin closes its associated switch. The Enable pin must be held low for normal operation; a high Enable state forces all the switches into the "off" position. If the outputs are tied together, this component can function as a multiplexer.*

Values

Industrial-mount solid-state relays typically can switch currents ranging from 5A to 500A, with 50A being very common. The higher-current relays mostly require DC control voltage; 4V to 32V are typical, although some versions can go much higher. They contain an SCR or triac to switch AC.

Smaller solid-state relays in SIP, DIP, or surface-mount packages often use MOSFETs on the output side, and are often capable of switching up to 2A or 3A. Some can switch either AC or DC, depending on the way the output is wired. The LED on the input side may require as little as 3mA to 5mA for triggering.

How to Use It

Solid-state relays find their primary uses in telecommunications equipment, industrial control systems and signalling, and security systems.

The component is very simple externally. Power on the input side can come from any source capable of delivering the voltage and current specified by the manufacturer, and any device that doesn't exceed maximum current rating can be connected to the output side, so long as provision is made for suppressing back-EMF from an inductive load, as shown in Figure 4-7. Often a solid-state relay can be substituted directly for an electromagnetic relay, without modifying the circuit.

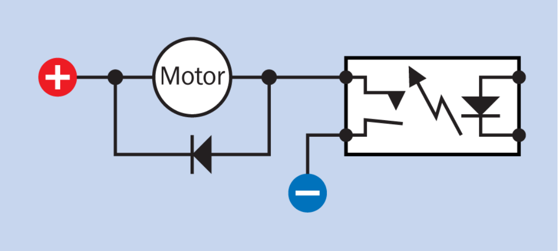

Figure 4-7. *Use of a diode around an inductive load, to protect a solid-state relay from back-EMF.*

Solid-state relays are heat sensitive, and their rating for switching current will diminish as their temperature increases. Manufacturer datasheets will provide specific guidance. Using a heat sink will greatly improve the performance. Bear in mind that the relay generates heat continuously while it is in its "on" mode—about 1 watt per ampere.

Because it requires so little current on the input side (typically no more than 15mA), a solid-state relay can usually be driven directly by chips such as *microcontrollers* that would not be able to activate an equivalent electromagnetic relay.

Applications may take advantage of the solid-state relay's reliability, immunity to vibration, lack of contact sparking, freedom from coil-induced surges on the input side, and lack of contact bounce on the output side. A solid-state relay is ideal within digital equipment that is sensitive to power spikes. It may switch a fuel pump that handles volatile, flammable liquids, or a wastewater pump in a basement subject to flooding (where long-term zero-maintenance reliability is necessary, and contact corrosion could be a risk in electromagnetic relays). Small solid-state relays can switch motors in robots or appliances where vibration is common, and are often used in arcade games.

What Can Go Wrong

Overheating Caused by Overloading

Relays must be *derated* when used at operating temperatures above the typical 20 or 25º C for which their specification applies. In other words, the sustained operating current must be reduced, usually by an amount such as 20% to 30% for each 10-degree increase in ambient temperature. Failure to observe this rule may result in failure of the component. Burnout may also occur if a high-current solid-state relay is used without a heat sink, or the heat sink isn't big enough, or *thermal compound* is not applied between the solid-state relay and the heat sink.

Overheating Caused by Bad Terminal Contact

If the screw terminals on the output side of a high-current solid-state relay are not tightened sufficiently, or if there is a loose spade terminal, or if a crimped connection isn't crimped tightly enough, the poor contact will create electrical resistance, and at high currents, the resistance will create heat, which can cause the solid-state relay to overheat and burn out.

Overheating Caused by Changing Duty Cycle

If a high-current solid-state relay is chosen for an application where it is in its "on" state only half the time, but the application changes during product development so that the solid-state relay is in its "on" state almost all the time, it will have to dissippate almost twice as much heat. Any time the duty cycle is changed, heat should be considered. The possibility of the relay being used in an unconventional or unexpected manner should also be considered.

Overheating Caused by Component Crowding

Overheating increases dramatically when components are tightly crowded. At least 2cm (3/4″) should be allowed between components.

Overheating in Dual Packaging

When a package contains two solid-state relays, the additive effects of the heat created by each of them must be considered.

Reverse-Voltage Burnout

Because a solid-state relay is more sensitive to *back-EMF* than an electromagnetic relay, greater care should be used to protect it from reverse voltage when switching inductive loads. A *protection diode* should be used, and a *snubber* can be added between its output terminals, if it is not included inside the relay package.

Low Voltage Output Current May Not Work

Unlike electromagnetic relays, solid-state relays require some voltage on the output side to enable their internal operation. If there is no voltage, or only a very low voltage, the SSR may not respond to an input. The minimum voltage required on the output side is usually specified in a datasheet.

To test a solid-state relay, apply actual voltages on input and output sides and use a load such as an incandescent light bulb. Merely applying a meter on the output side, set to measure continuity, may not provide sufficient voltage to enable the relay to function, creating the erroneous impression that it has failed.

Inability to Measure AC Output

When a multimeter is used to test continuity across the output of an AC-switching solid-state relay of zero-crossing specification, the meter will generate enough voltage to prevent the solid-state relay from finding zero voltage across its output terminals, and consequently the solid-state relay won't switch its output.

Relay Turns On but Won't Turn Off

When a solid-state relay controls a relatively high-impedance load such as a small **solenoid** (see Volume 1) or a **neon bulb** (see Chapter 19), the relay may switch the device on but will seem unable to switch it off. This is because the leakage current of the solid-state relay, in its "off" state, may be just enough to maintain the load in its "on" state.

If an SSR containing a triac is used erroneously to switch DC, it will not be able to switch off the current.

Relays in Parallel Won't Work

Two solid-state relays usually cannot be used in parallel to switch twice as much current. Because of small manufacturing variances, one relay will switch on a moment before the other. When the first relay is on, it will divert the load current away from the second relay. The second relay needs a small amount of current on its output side, to function. Without any current, it will not switch on. This means the first relay will pass the total current without any help from the second relay, and will probably burn out, while the second relay does nothing.

Output Device Doesn't Run at Full Power

A solid-state relay imposes a voltage reduction on its output side. This will be a fixed amount, not a percentage. When switching 110V, this difference may be negligible; when switching 12V, it may deliver only 10.5V, which represents enough of a drop to cause a motor or a pump to run noticeably more slowly. The internal switching device inside the relay (MOSFET, triac, SSR, or bipolar transistor) will largely determine the voltage drop. Check the manufacturer's datasheet before using the relay.

Solid-State Relays and Safety Disconnects

Because a solid-state relay always allows some leakage in its "off" state, it can still deliver a shock when used to switch high voltages. For this reason, it may not be suitable in a safety disconnect.

optocoupler

5

Sometimes known as an *optoelectronic coupler*, *opto-isolator*, *photocoupler*, or *optical isolator*.

A **solid-state relay** is sometimes referred to as an **optocoupler**, but in this encyclopedia it has a separate entry. An optocoupler is a relatively simple device consisting of a light source (usually an LED) and a light sensor, both embedded in one package. It is used primarily for isolation rather than to switch a high current. A solid-state relay can be thought of as a substitute for an electromagnetic **relay**, usually has additional components in its package, and is intended to switch currents of at least 1A.

OTHER RELATED COMPONENTS

- **electromagnetic relay** (see Volume 1)
- **solid-state relay** (see Chapter 4)

What It Does

An optocoupler allows one section of a circuit to be electrically isolated from another. It protects sensitive components, such as logic chips or a microcontroller, from voltage spikes or incompatible voltages in other sections of a circuit. Optocouplers are also used in medical devices where a patient has to be protected from any risk of electric shock, and are used in devices which conform with the MIDI standard for digital control of music components.

In Figure 5-1, three possible applications for an optocoupler are suggested:

Top

The output from a logic chip passes through an optocoupler to an inductive load such as a relay coil, which may create voltage spikes that would be hazardous to the chip.

Center

The noisy signal from an electromagnetic switch passes through an optocoupler to the input of a logic chip.

Bottom

The low-voltage output from a sensing device on a human patient passes through an optocoupler to some medical equipment, such as an EEG machine, where higher voltages are used.

Internally, an optocoupler works on the same principle as a **solid-state relay**. An LED is embedded on the input side, shining light through an interior channel or transparent window to a sensing component that is embedded on the output side. Because the only internal connection is a light beam, the input and output of the optocoupler are isolated from each other.

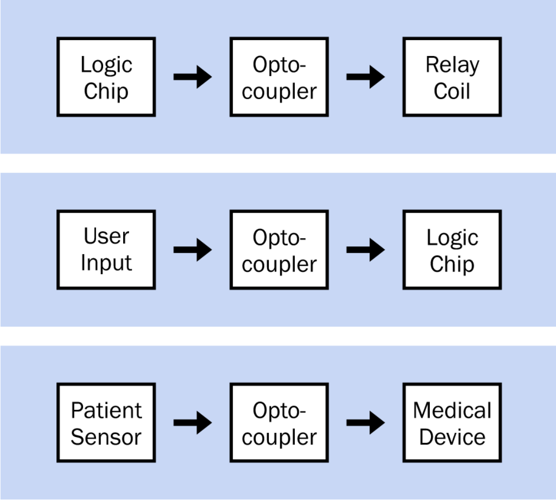

Figure 5-1. *Possible applications for a photocoupler. See text for details.*

Isolation transformers were traditionally used for this purpose prior to the 1970s, when optocouplers became competitive. In addition to being smaller and cheaper, an optocoupler can also pass slow-changing signals or on-off DC states which a transformer would ignore.

More recently, inductive and capacitive coupling components have become available in surface-mount packages that are competitive with optocouplers for high-speed data transfer. They also claim to be more durable. Because of the gradual reduction in output from an LED, the performance of an optocoupler degrades over time, and is typically rated for up to 10 years.

How It Works

The LED in an optocoupler almost always emits light in the near-infrared part of the spectrum, and is matched to the sensitivity of a **phototransistor**, or a **photodiode**, or (less often) a **photoresistor** that provides the output. Photosensitive **triacs** and **SCRs** are also sometimes used.

The most common type of optocoupler uses a bipolar phototransistor with an open-collector output. Schematic symbols for this type are shown in Figure 5-2:

Top left

The most common generic form.

Top right

Two diodes on the input side allow the use of alternating current.

Center left

An additional terminal allows addition of bias to the photosensitive base of the output transistor, to reduce its sensitivity.

Center right

An Enable signal can be used as the input to the NAND, suppressing or enabling the output.

Bottom left

A *photodarlington* allows higher emitter current.

Bottom right

Relatively uncommon, and is also used for a **solid-state relay**.

In each symbol, the diode is an LED, and the zigzag arrow indicates light that is emitted from it. A pair of straight arrows, or wavy arrows, may alternatively be used.

An optocoupler in through-hole DIP format is shown in Figure 5-3.

An *optical switch* can be thought of as a form of optocoupler, as it contains an LED opposite a sensor. However, the LED and the sensor are separated by an open slot, to allow a thin moving object to pass through, interrupting the light beam as a means of detecting the event. It is categorized as a *sensor* in this encyclopedia, and will be found in Volume 3.

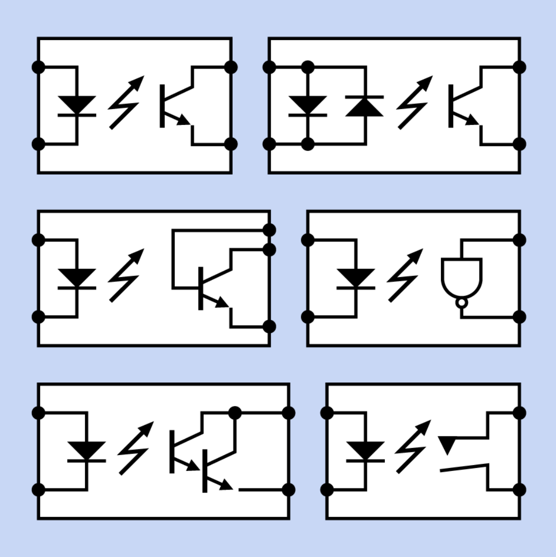

Figure 5-2. *Six variants of schematic symbols that may be used to represent an optocoupler. See text for details.*

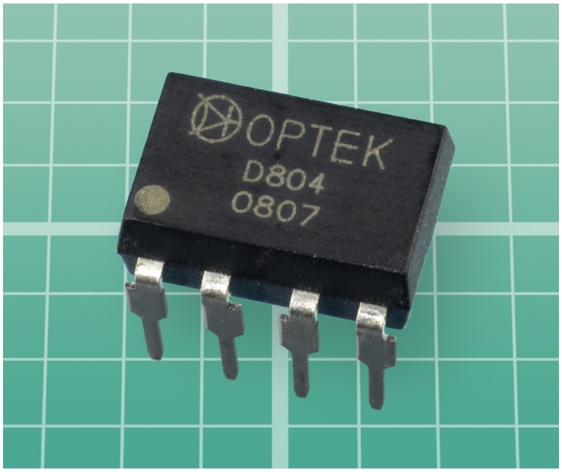

Figure 5-3. *An optocoupler in through-hole 8-pin DIP format.*

Variants

Internal Sensors

Historically, a photoresistor (often referred to as a *photocell*) was the first type of sensor to be used. It has a more linear response than other sensor types, but its response is much slower. It is still found in audio applications. "Stomp box" pedals used by guitarists typically contain an optocoupler that employs a photoresistor, and are valued for their linearity and their immunity from the mechanical wear, contamination, and "scratchiness" that builds up over time in a **potentiometer**. Optocouplers also eliminate *ground loops*, which tend to be induced by small differences in ground potential, introducing hum or buzz in audio applications when two or more power supplies are tied together.

The type of optocoupler that contains a photoresistor and is commonly used by musicians was initially trademarked as a *Vactrol*, and that term is still used generically. Vactrols have also been used to provide audio compression in telephone voice networks, and were used in photocopiers and photographic exposure meters, but these applications are now obsolete.

Photoresistors are becoming uncommon because of their cadmium content, which is unlawful in many countries (especially in Europe) because of its environmental toxicity.

A photodiode provides the fastest response time in an optocoupler, limited primarily by the characteristics of the LED that shines light upon it. A *PIN diode* can respond in less than a nanosecond; its acronym is derived from its fabrication from p-type and n-type semiconductor layers with an *intrinsic* layer connecting them. This additional layer can be responsive to light. When the diode is slightly reverse-biased, a photon entering the intrinsic layer can dislodge an electron, enabling current to flow. The reverse bias enlarges the active area and enhances the effect. In this mode, the PIN acts like a photoresistor, appearing to reduce its resistance in response to light.

When the PIN is used in *photovoltaic* mode, no bias is applied, and the component actually generates a small voltage (less than 1VDC), like a solar cell, in response to incoming light. Where an optocoupler uses a MOSFET on its output side, as many as 30 photodiodes may be connected in series to develop the necessary threshold voltage to trigger the transistor. This arrangement is common in **solid-state relays**.

A bipolar phototransistor is a slower-speed device but is still usually capable of a 5µs response time or better. Its open collector requires external voltage and a *pull-up resistor* to deliver a positive output so long as the phototransistor is nonconductive. When the LED turns on, the phototransistor sinks current, effectively creating a low output. In this way, the optocoupler functions like an *inverter*, although some variants include a noninverting output.

Basic Optocoupler Types

An optocoupler with *high linearity* will respond more proportionally to variations in current to its LED. *High Speed* optocouplers are used for high-frequency data transfer. *Logic-output* optocouplers have a clean high/low output transition, rather than an *analog output*, which varies with fluctuations in the input. Linearity is of importance only where an optocoupler is being used to transmit an analog signal with some fidelity. Some logic-output optocouplers provide the function of a Schmitt trigger on their output side.

While optocouplers are available in various package formats, the DIP style with six or eight pins remains popular, providing sufficient physical space for the LED, the sensor, and a light channel, while providing good electrical isolation.

Variants may have two or four optocouplers combined in one package. A *bidirectional* optocoupler may consist of two optocouplers in parallel, inverted with respect to each other.

Values

In a datasheet, the characteristics of primary importance in an optocoupler are:

- CTR is the Current Transfer Ratio, the ratio of maximum output current to input current, expressed as a percentage. With a bipolar phototransistor output, 20% is a typical minimum CTR. With a photodarlington output, the CTR may be 1,000% but the bandwidth is much lower—the response time may be measured in microseconds rather than nanoseconds. Optocouplers with a photodiode output have a very low CTR, and their output is in microamps. However, they provide the most linear response.
- $V_{CE(MAX)}$ is the maximum collector-emitter voltage difference (in an optocoupler with a bipolar phototransistor output). Values from 20 to 80 volts are common.
- V_{ISO} is the maximum potential difference, in VDC, between the two sides of the optocoupler.
- I_{MAX} is the maximum current the transistor can handle, generally in mA.
- Bandwidth is the maximum transmittable signal frequency, often in the range of 20kHz to 500kHz.

The LED in an optocoupler typically requires 5mA at a forward voltage of 1.5V to 1.6V.

The maximum collector current on the output side of an optocoupler is unlikely to be higher than 200mA. For higher output currents, a solid-state relay should be considered. It provides photo-isolation on the same basis as an optocoupler, but high-current versions tend to be considerably more expensive.

How to Use It

The primary purpose of an optocoupler is to provide protection against excessive voltage—from transients, incompatible power supplies, or equipment with unknown characteristics. If a device is designed to be plugged into a USB port on a computer, for instance, the computer may be isolated via an optocoupler.

A series resistor for the LED is not built into most optocouplers, because the value of the resistor will depend on the input voltage that is used. Care must be taken to determine what the maximum voltage on the input side will be, and a series resistor should be chosen to reduce current appropriately. Allowance should be made for some degradation in the performance of the LED over time.

For an optocoupler with an open-collector output, a pull-up resistor is necessary in most applications. The voltage from the optocoupler must be matched to the input requirements of other components, and the collector current must remain within the specified limits. Some trial and error in resistor selection may be necessary.

In Figure 5-4, a schematic shows typical component values in a test circuit using a pushbutton as input. The separation of the two power supplies is emphasized by the different color shades used for the positive and negative symbols. Although the input side and the output side of an optocoupler may be used with a common ground, this defeats its purpose in providing complete isolation between the sections of the circuit.

The pinouts for an optocoupler must be checked carefully in the manufacturer's datasheet. While the input for an 8-pin DIP chip is usually applied to pins 2 and 3, the output pin functions are not standardized and will vary depending on the internal configuration of the chip. An optocoupler such as the Optek D804, with an enable function using an internal NAND gate, requires its own power supply.

Where an optocoupler allows an external connection to the base of its internal bipolar output phototransistor, reverse bias applied to this pin will decrease the sensitivity of the optocoupler but can increase its immunity to noise on the input side.

What Can Go Wrong

Overload conditions on the input or the output side of an optocoupler will be the most likely cause of failure.

Age

Because optocouplers are typically rated for only 10 years of average use, the age of a component may cause it to fail.

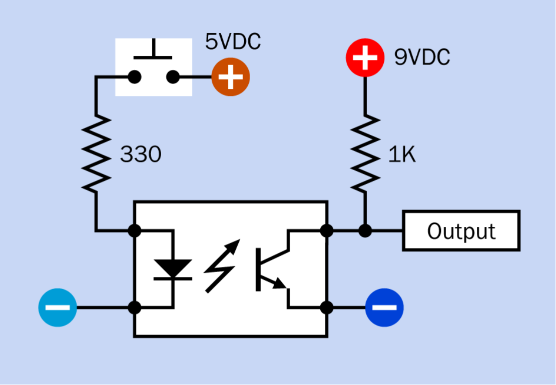

Figure 5-4. *Typical values for a series resistor (to protect the LED) and pull-up resistor (to control current and voltage on the output side) in an optocoupler test circuit.*

LED Burnout

Because the LED is hidden inside the component, there is no immediate indication of its performance. A meter can be inserted into the circuit on the input side to determine if current is passing through the LED. A meter set to measure volts can be used to discover whether the LED is imposing a normal voltage drop. While significant overload will cause immediate burnout, slightly exceeding the current rating of the LED may have more pernicious consequences, as the LED may not fail until days or weeks have passed without any sign of trouble. The failure of the optocoupler will be unexpected and difficult to determine.

Transistor Burnout

Here again the damage caused by excessive current may be progressive, occurring over a prolonged period. The easiest way to test an optocoupler that may have failed is by removing it from the circuit. A socketed DIP package is preferable for this purpose.

comparator

6

Although a **comparator** has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

This entry describes an analog comparator. A *digital comparator* is very different, being a logic chip that compares two binary numbers that can be referred to as A and B. Outputs from the chip indicate whether A>B or A<B or A=B. The digital comparator does not have an entry in this encyclopedia.

OTHER RELATED COMPONENTS

- **op-amp** (see Chapter 7)

What It Does

A **comparator** is an integrated circuit chip that compares a variable voltage on one input pin with a fixed, reference voltage on a second input pin. Depending which voltage is higher, the output from the comparator will be high or low.

The output will make a clean transition between two fixed values, even if the input is infinitely variable. Thus the comparator can function as an *analog-digital converter*, as shown in Figure 6-1.

Because the output voltage range can be adjusted up or down independently of the input range, a comparator can also function as a *voltage converter*.

Hysteresis

If positive feedback is added through external resistors, *hysteresis* can be introduced. We may imagine a *hysteresis zone* extending above and below the reference voltage level. Small input variations that occur within the zone will be ignored. The comparator only reacts when the input signal emerges above or below the hysteresis zone. When the input signal returns into the hysteresis zone, this event also will be ignored. The concept is illustrated graphically in Figure 6-2. A circuit to create hysteresis is shown in Figure 6-10.

How It Works

The schematic symbol for a comparator is shown in Figure 6-3. This seems identical to the symbol for an op-amp, described in Chapter 7, but an op-amp is traditionally a *dual-voltage* device using positive and negative power sources that are equal and opposite, in addition to a zero value midway between the two. Modern comparators mostly use a conventional single voltage, and therefore the negative symbol used in comparator schematics throughout this section of the encyclopedia represents 0 volts. It has the same meaning as the ground symbol found in many schematics elsewhere.

The two inputs to a comparator are described as *inverting* and *noninverting* (for reasons explained later). Confusingly, these are identified with plus and minus symbols inside the triangle that represents the component. These plain black-and-white symbols have nothing to do with the power supply.

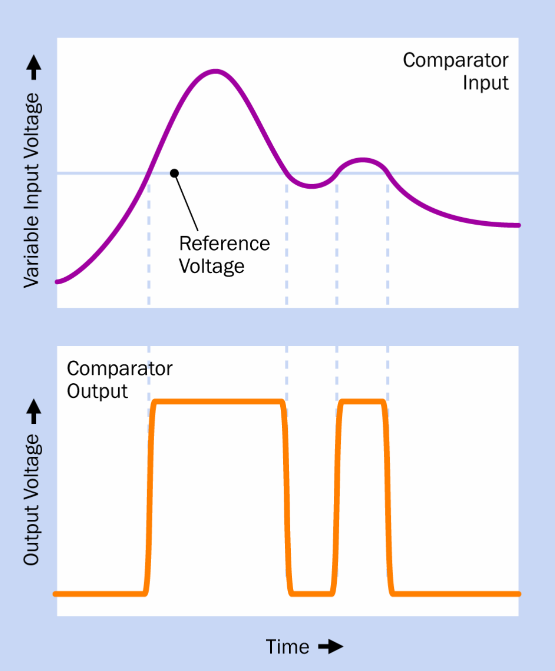

Figure 6-1. *The basic behavior of a simple comparator is shown here.*

Often, in schematics, the power supply is not shown, because it is assumed to be present. However, all comparators require a power supply in order to function.

The basic internal and external connections used in conjunction with a typical comparator are shown in Figure 6-4.

The potentiometer at top left is often a *trimmer*, to fine-tune a reference voltage. The variable input can come from a sensor or any other device capable of delivering a voltage up to the limit set by V1.

The output is often an *open collector* from an internal bipolar transistor, as shown in the figure.

Note that as many as three different voltages can be used, as indicated by the different colors associated with V1, V2, and V3. However, they must share a common ground to enable the comparator to make valid comparisons.

When the noninverting input exceeds the voltage of the inverting input, the output transistor goes into its "off" state, and blocks current from an external *pullup resistor*. Because the current from the resistor now has nowhere else to go, it is available to drive other devices attached to the comparator output, and the output appears to be high.

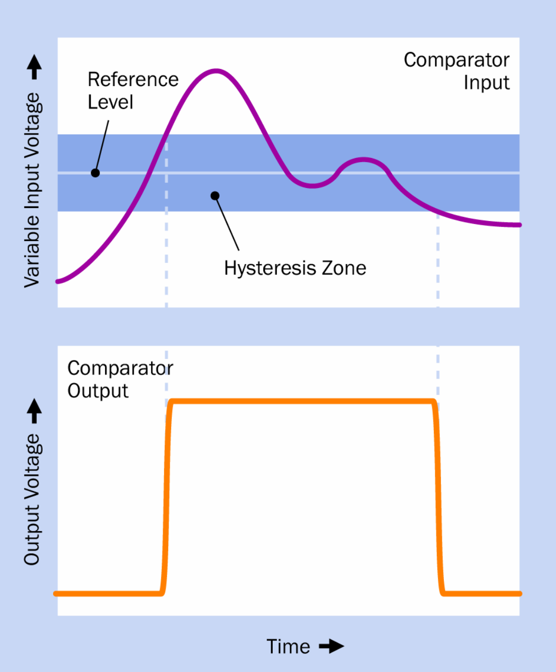

Figure 6-2. *The performance of a comparator shown in the previous figure can be modified by the addition of hysteresis. Small variations that occur within the hysteresis zone are ignored.*

When the noninverting input falls below the voltage of the inverting input, the transistor becomes conductive, and sinks almost all the current from the pullup resistor, assuming other devices attached to the output have a relatively high impedance. The output from the comparator now appears to be low.

This can be summed up as follows:

- When a variable voltage is applied to the noninverting input, and it rises *above* the reference voltage applied to the *inverting input*, the output transistor turns off, and the comparator delivers a *high* output.
- When a variable voltage is applied to the noninverting input, and it falls *below* the reference voltage applied to the *inverting input*, the output transistor turns on, and the comparator delivers a *low* output.

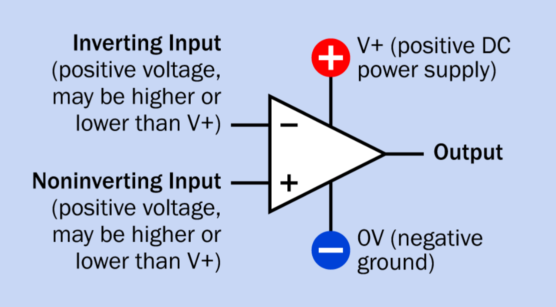
Inverting Input
(positive voltage, may be higher or lower than V+)
Noninverting Input
(positive voltage, may be higher or lower than V+)
V+ (positive DC power supply)
Output
0V (negative ground)
–
+

Figure 6-3. *The symbol for a comparator is the same as the symbol for an op-amp, even though they often require different types of power supply and their functions are significantly different.*

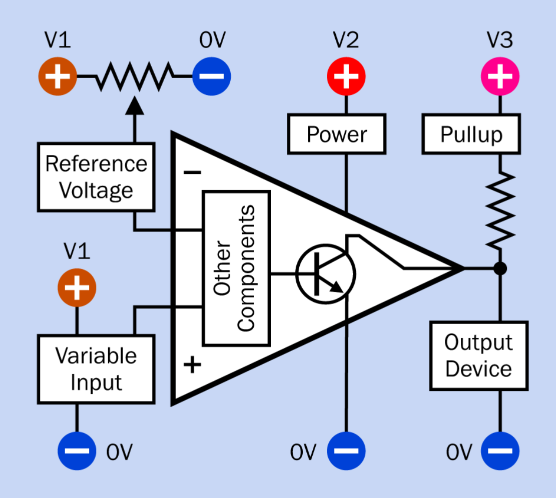
V1
0V
V2
V3
Reference Voltage
Power
Pullup
Other Components
V1
Variable Input
Output Device
0V
0V
0V
–
+

Figure 6-4. *Connections to a comparator, and their functions.*

If the reference voltage and the variable voltage are swapped between the input pins, the behavior of the comparator is reversed. This relationship is illustrated in Figure 6-5. When a voltage transition is applied to the *inverting* input, the transition is *inverted* at the output.

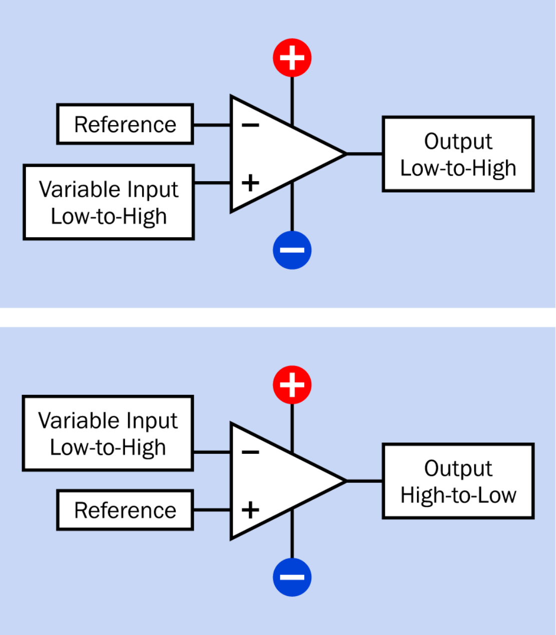
Reference
Variable Input Low-to-High
Output Low-to-High
Variable Input Low-to-High
Reference
Output High-to-Low
–
+

Figure 6-5. *Depending which input pin is used for the reference voltage, and which input carries a variable voltage, the comparator output either follows the variable voltage or inverts it.*

Placement of the plus and minus signs inside the comparator symbol may vary. Most often, the minus sign is above the plus sign, as shown in all the schematics here. Sometimes, however, for convenience in drawing a schematic, the plus sign may be shown above the minus sign. Regardless of their placement, the plus sign always identifies the noninverting input, and the minus sign always identifies the inverting input. To avoid misinterpretations, schematics should be inspected carefully.

Where a power supply for the comparator is shown, the positive side is always attached to the upper edge of the symbol, while 0V ground is always attached to the lower edge.

Differences from an Op-Amp

Saturation versus linearity

The output of a comparator is optimized for saturation (high or low, without intermediate levels, using positive feedback). The output of an op-amp is optimized for linearity (faithful reproduction of nuances in the input, using negative feedback).

Output mode

The majority of comparators have open-collector outputs (or open-drain outputs in CMOS devices) where the voltage is established by a pullup resistor. This can be adjusted for compatibility with other components, especially 5VDC logic. Only a minority have push-pull amplifier outputs that require no pullup resistor. By comparison, among op-amps, a push-pull output that functions as a voltage source is the traditional default.

Faster response

A comparator responds more quickly than an op-amp to changes in input voltage, if the op-amp is used in the role of a comparator. The comparator is primarily a switching device, not an amplifier.

Hysteresis

This is generally desirable in a comparator, for reasons already explained, and some components are designed with hysteresis built in. This feature is undesirable in an op-amp, as it degrades sensitivity.

Open-loop operation

(i.e., without feedback) this can be used with a comparator. An op-amp is intended for use in closed-loop circuits (i.e., with feedback), and manufacturers will not specify its performance in an open loop.

As previously noted, a comparator usually requires a single-voltage power supply, while an op-amp often requires a dual-voltage power supply.

Variants

Where a comparator uses a MOSFET output transistor, it may have an open-drain output, which requires a pullup resistor, as with an open-collector output.

Some comparators have a *push-pull* output, capable of supplying output current (usually a small amount). In these instances, no pullup resistor is necessary or desirable. The output voltage range will be closest to *rail-to-rail* values (i.e., the range of the power supply) where MOSFETs are used for the output, as MOSFETs impose a smaller voltage drop than bipolar transistors.

The advantage of an open collector (or open drain) relative to a push-pull output is that it allows the output voltage to be set independently of the power supply voltage. Another advantage is that multiple outputs can be connected in parallel, as in a *window comparator* circuit (described below).

Some comparators incorporate a reference voltage on the chip, based on the power supply to the chip. In this case, a separate reference voltage does not have to be supplied, and the component will draw less current.

Many chips are available containing two or more comparators. This is often expressed as the number of *channels* in the component. A *dual comparator* typically allows two different voltage sources for the outputs of the comparators. They will share the same 0V ground, however. Chips such as the LM139 and LM339 contain four comparators, and are available in through-hole or surface-mount formats. They have become a generic choice, costing less than $1 apiece.

An LM339 comparator chip is shown in Figure 6-6. This is a quad chip, meaning that it contains four comparators. They share a common power supply. The chip is TTL and CMOS compatible, is typically powered by 5VDC, but can be driven by up to 36VDC. The input differential voltage range also extends up to 36V.

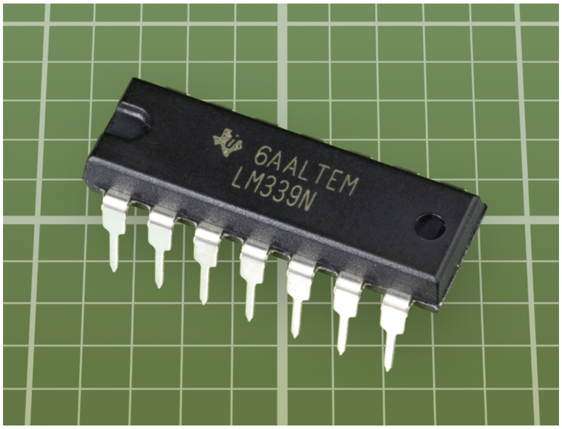

Figure 6-6. *The LM339 quad comparator chip, shown here, was introduced long ago but remains widely used.*

Some comparators have an internal *latch* function that is accessed by a dedicated pin. The latch-enable signal forces the comparator to assess its inputs and hold an appropriate output which can then be checked by other components.

Values

In a datasheet, V_{IO} (also referred to as V_{OS}) is the *input offset voltage*. This is a small voltage, in addition to the reference voltage, which the comparator will require to *toggle* its output in either direction, up or down. Figure 6-7 shows this graphically. V_{IO} sets the limit of *resolution* of the comparator, which will not respond unless the input voltage exceeds the reference voltage by this amount. A smaller value for V_{IO} is better than a larger value. Common values for V_{IO} range from 1mV to 15mV. The actual offset voltage tends to vary between one sample of a component and another. V_{IO} is the maximum allowed value for a component.

Because the comparator will not respond until the reference voltage is exceeded by V_{IO}, the output pulse width will be narrower than if the comparator reacted at the point where the variable voltage input was precisely the same as the reference voltage.

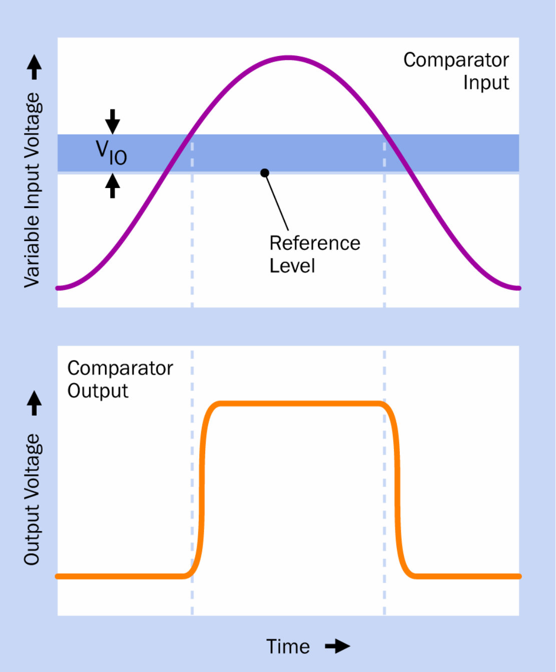

Figure 6-7. *The input offset voltage is the very small voltage that a comparator requires, additional to the reference input voltage, before it will toggle its output from low to high or high to low.*

V_{TRIP+} and V_{TRIP-} are the *rising* and *falling voltages*, respectively, that will trip the comparator output where the comparator exhibits some innate hysteresis without an external feedback loop. They are also referred to as *Lower State Transition Voltage* (LSTV) and *Upper State Transition Voltage* (USTV).

V_{HYST} is the *hysteresis range* defined as V_{TRIP+} minus V_{TRIP-}. The relationship is shown graphically in Figure 6-8.

A_{VD} is the *voltage gain* of a comparator, in which the letter "A" can be thought of as meaning "amplification." The gain is measured as a maximum ratio of output voltage to input voltage. Typically it ranges from 40 to 200.

Supply voltage for modern comparators is often low, as the components are used in surface-mount format for battery-powered devices where low power consumption is a primary concern. Thus, 3VDC is common as a power requirement, and 1.5VDC comparators are available. Still, older chips can use as much as 35VDC.

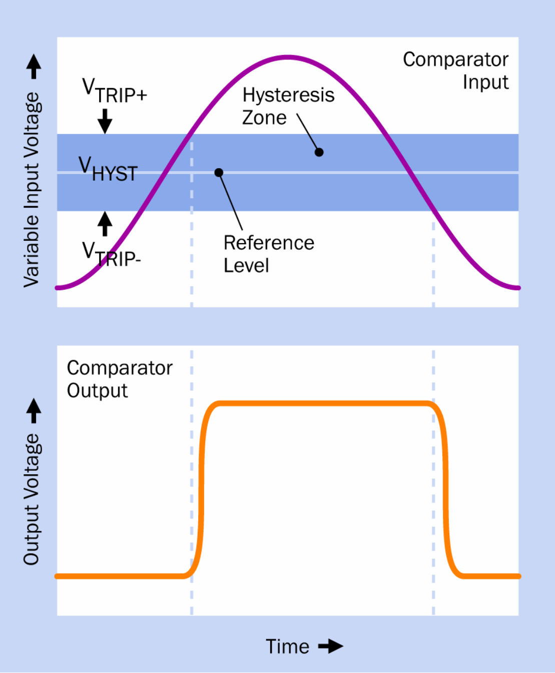

Figure 6-8. *The value of V_{TRIP} shows the hysteresis in a comparator—the range of input voltages, relative to the reference voltage, in which it will not respond.*

Supply current can range from 7mA down to below 1µA.

I_{SINK} is the recommended typical or maximum *sink current* that the component will tolerate, if it has an open-collector output. This value should be considered in relation to the power dissipation, P_D.

The *propagation delay* in a comparator is measured from the moment when an input (usually a square wave) reaches the triggering value, to the time when the consequent output reaches 50% of its final value.

When a comparator is driving CMOS logic using a 5VDC power supply, a typical value for a pullup resistor is 100K. It does not have to be lower, because CMOS has such a high input impedance.

How to Use It

In Figure 6-1, a hypothetical comparator responds immediately when the input voltage equals the reference voltage value. However, this is an idealized scenario. A magnified view, in Figure 6-9, suggests that the comparator is likely to respond with *jitter* when the input signal is very close to the reference voltage, because of tiny variations in heat, current, and other variables. This jitter will cause significant problems if the comparator is driving a device such as a relay, directly or indirectly.

Hysteresis eliminates this uncertainty around the transition level of the input, by telling the comparator to ignore small irregularities in the input voltage. Hysteresis is also useful in many situations where larger variations in a sensor input should be ignored. In Figure 6-2, for instance, suppose that the input voltage comes from a temperature sensor. The small bump in the right-hand section of the curve is probably unimportant; it could be caused by someone opening a door, or a person's body heat in brief proximity to the sensor. There's no point in responding to every little event of this type. In this application, the larger, longer-term temperature trend is what matters, and significant hysteresis is appropriate.

Also, if a comparator is being used as a thermostat, to switch a heating system on and off, we do not want the comparator to respond as soon as the temperature rises just a small amount. The heating system should run for a while before it elevates the temperature beyond the hysteresis zone.

The usual way to create hysteresis is with *positive feedback*. In Figure 6-10, a connection from the output of the comparator runs back through a 1M potentiometer to the variable (noninverting) input. The effect that this has is to reinforce the input voltage with the output voltage, as soon as the comparator input goes high. Now the input can diminish slightly without switching off the comparator. But if the input declines significantly, even the feedback from the output voltage won't be sufficient to maintain the variable input at a higher level than the reference voltage. (Remember, the "high" output voltage from the comparator is a fixed value; it does not change in proportion with the input voltage.) Consequently, the output toggles to low. Now the variable input is deprived of help from the comparator output, so it will be low enough that it has to rise considerably to toggle the comparator back on again. During that period, once again, small variations will be ignored.

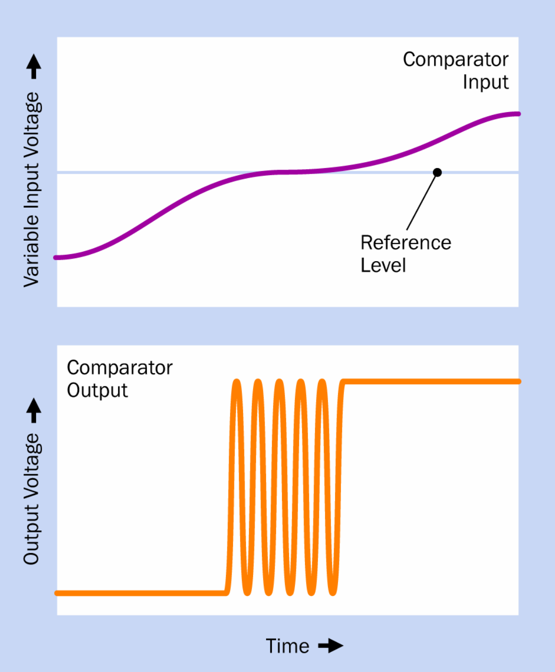

Figure 6-9. *In real-world applications, tiny variations where the variable input voltage crosses the reference voltage can induce jitter in the output from a comparator that has no hysteresis.*

In the schematic, a **phototransistor** (PT1, at left) is in series with a 3.3K resistor to adjust its voltage output to a suitable range. A 1M potentiometer at upper-left is wired as a voltage divider, so that it can establish a reference level that matches the light level that we wish to detect with the phototransistor.

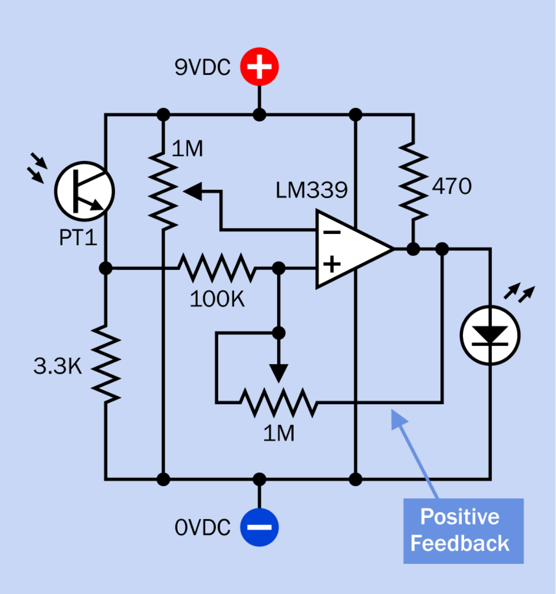

Figure 6-10. *A simple circuit to achieve hysteresis with positive feedback to the variable input of a comparator.*

The 470Ω resistor is the pullup resistor, which protects the LED from excessive current. The lower 1M resistor adjusts the amount of positive feedback, which determines the width of the hysteresis zone.

Values for components may have to be adjusted depending on the supply voltage, the variable input voltage, and other factors. But the principle will remain the same. Note that in the example shown, all the positive voltage sources are identical. In practice, different voltages could be used, so long as they share a common ground.

AND gate

A set of open-collector comparators can function jointly as an AND gate, when their outputs are tied together with one pullup resistor. So long as all the output transistors are nonconductive, the output will be high. If just one comparator toggles into conductive mode, the output will be low. See Figure 6-11.

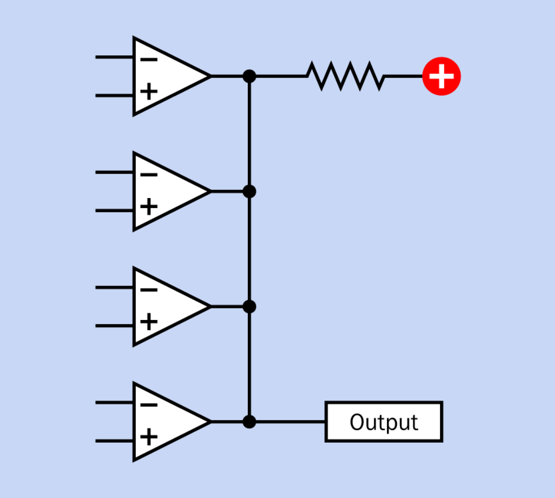

Figure 6-11. *If the outputs of multiple open-collector comparators are tied together with a suitable pullup resistor, they will function as an AND gate.*

Bistable Multivibrator

If positive feedback to the noninverting input of the comparator is sufficiently high, a voltage almost at 0V ground will be required to counter the high output from the comparator—after which, a voltage almost equal to the supply voltage will be needed to turn it back on. In other words, the comparator is behaving like a *bistable multivibrator*, or **flip-flop**.

Relaxation Oscillator

A *relaxation oscillator*, which is a form of *astable multivibrator,* can be created using direct positive feedback in combination with delayed negative feedback. In Figure 6-12, positive feedback goes to the noninverting input, as before, but negative feedback also passes through a 220K resistor to the inverting input of the comparator. A 0.47µF capacitor initially holds the inverting input low, while the capacitor charges. Gradually the capacitor reaches and exceeds the charge on the noninverting input, so the output from the comparator toggles to its low state. This means that its internal transistor is now sinking current, and it discharges the capacitor. Because the noninverting input is being held at a voltage midway between supply and ground by the two 100K resistors forming a voltage divider, eventually the voltage on the inverting input controlled by the capacitor falls below the noninverting voltage, so the cycle begins again.

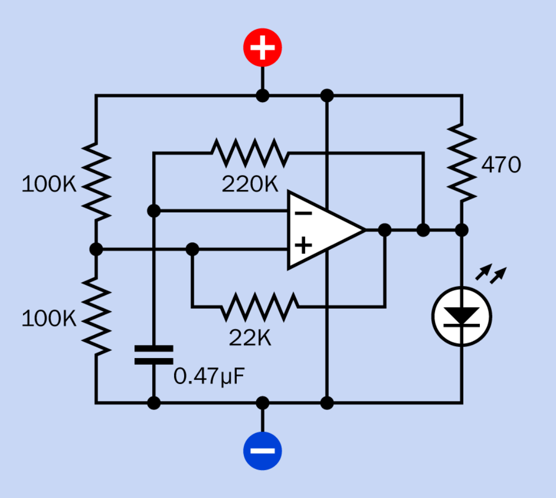

Figure 6-12. *A comparator can be used to create a relaxation oscillator.*

Level Shifter

Where a comparator is used simply to change the level of an input voltage, it can be referred to as a *level shifter*. An example of a level shifter is shown in Figure 6-13, in which a high/low 3VDC logic input is converted to a high/low logic output at 5VDC.

Window Comparator

A *window comparator* is a circuit (not a single component) that will respond to input voltages that deviate outside an acceptable "window" of values. In other words, the circuit responds anytime the variable input is either unacceptably low or unacceptably high.

An example could be an alarm that will sound if a temperature is either too low or too high. In Figure 6-14, two comparators are used to create a window comparator circuit, both sharing a variable voltage input from a sensor. A voltage divider is necessary to establish a higher voltage limit at the noninverting input of the upper comparator, while a separate voltage divider would establish a lower limit at the inverting input of the lower comparator. If an alarm has an appropriate resistance, it can be used instead of a pullup resistor. The alarm will sound when the output from either comparator is low, which happens if the inverting input has a higher voltage than the non-inverting input.

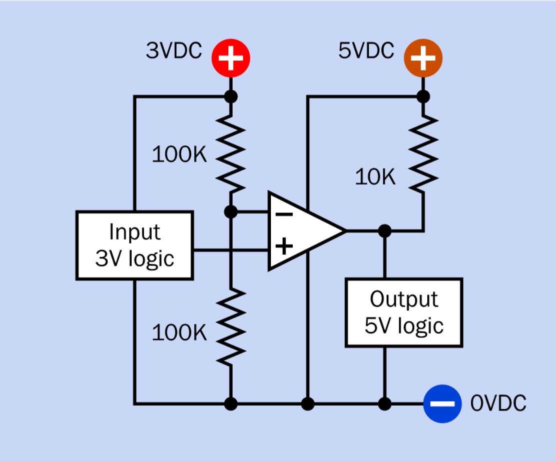

Figure 6-13. *A comparator can be used to convert high and low 3V logic inputs into high and low 5V logic outputs.*

Other Applications

As previously noted, a comparator can be used as a simple *analog-digital converter*. It has "one bit" accuracy (i.e., its output is either high or low).

A comparator can be used as a *zero point finder* when its variable voltage input is attached to an AC signal. The output from the comparator will be toggled whenever the AC signal passes through zero volts. The output will be a square wave (approximately) instead of a sine wave.

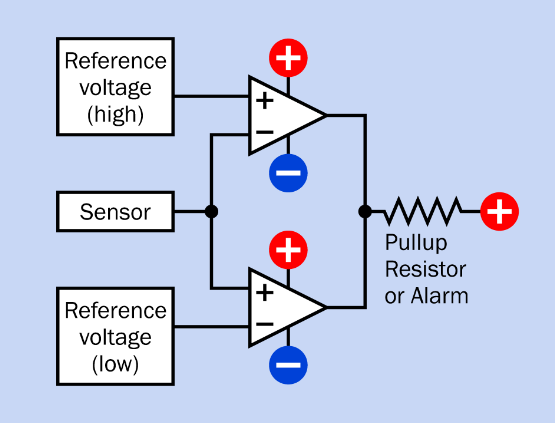

Figure 6-14. *A basic, simplified circuit for a window comparator. See text for details.*

A *continuous converter* changes its output promptly in response to a change in input. This requires continuous current consumption. Because many applications only need to check the output from a comparator at intervals, power can be saved by using a *clocked* or *latched* comparator.

What Can Go Wrong

Oscillating Output

The high input impedance of a comparator is vulnerable to stray electromagnetic fields. If the conductors leading to and from the comparator are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add 1µF bypass capacitors to the power supply on either side of the comparator. However, some manufacturers recommend alternatives such as introducing just a small amount of hysteresis, or reducing the value of input resistors to below 10K.

If a chip contains multiple comparators, and one of them is unused, one of its input pins should be tied to the positive side of the voltage supply while the other should be tied to 0V ground, to eliminate the possibility of an oscillating output.

Confused Inputs

A comparator will function if its two inputs are swapped accidentally, but its high and low output states will be the inverse of what is expected. Also, if positive feedback is used, transposed inputs can create oscillations. Because the comparator symbol may appear in a schematic with the noninverting input either below or above the inverting input, the inputs are easily transposed by accident.

One way to remember which way the inputs should be connected is to use the mnemonic: "plus, high, positive." The *plus* input creates a *high* output when the input becomes *more positive* than the reference voltage on the other input. The opposite is less intuitively obvious: the *minus* input creates a *high* output when the input becomes *more negative* than the reference voltage on the other input.

Wrong Chip Type

Different comparators offer different outputs: open collector, open drain, and push-pull. While open collector and open drain function similarly, the pullup resistor value is likely to be different in each case. If a push-pull output is mistakenly connected as if it is open collector or open drain, it will not work correctly, if at all. Different types of comparators must be sorted and stored in clearly labeled bins.

Omitted Pullup Resistor

It is relatively easy to forget to include the pullup resistor on an open-collector output. In this case, when the transistor inside the comparator is in its nonconductive state, the output pin will be floating, with an indeterminate voltage that will create confusing or random results.

CMOS Issues

As usual when using CMOS chips, it is bad practice to allow unconnected, floating inputs. This is an issue where a chip contains multiple comparators, some of which are not being used. The solution recommended by some manufacturers is to tie one input of an unused comparator to the supply voltage, and the other input of the same comparator to ground.

Erratic Output

If positive feedback is insufficient, the comparator output may show signs of *jitter*. Conversely, if the positive feedback is excessive, the comparator may get stuck in an on state or an off state. Feedback must be chosen carefully.

Swapped Voltages

A comparator is often capable of controlling an output voltage that is much higher than that of its power supply. Because both voltages are applied to different pins on the same chip, mistakes can be made quite easily. The chip is likely to be damaged if the voltages are swapped accidentally between the relevant pins.

Heat-Dependent Hysteresis

Remember that the voltages at which the comparator turns on and off will vary slightly with the temperature of the component. This *drift* should be tested by running the comparator at higher temperatures.

7

op-amp

Although a comparator has the same schematic symbol as an **op-amp**, their applications differ and they are described in separate sections of this encyclopedia.

The unabbreviated name for an op-amp is an *operational amplifier*, but this term is seldom used.

OTHER RELATED COMPONENTS

- **comparator** (see Chapter 6)

What It Does

An op-amp is an *operational amplifier* consisting of multiple transistors packaged in an integrated circuit chip. It senses the fluctuating voltage difference between two inputs, uses power from an external supply to amplify that difference, and uses *negative feedback* to ensure that the output is an accurate replica of the input. Its amplification can be adjusted by changing the values of two external resistors.

Op-amps were developed originally using vacuum tubes, for use in analog computers, before the era of digital computing. Their implementation in integrated circuits dates from the late 1960s, when chips such as the LM741 were introduced (lower-noise versions of it still being widely used today). Multiple op-amps in a single package were introduced in the 1970s.

An LM741 is shown in Figure 7-1. Inside the 8-pin, DIP package is a single op-amp.

How It Works

In alternating current, voltages deviate above and below a zero value, which is sometimes referred to as the *neutral* value. This occurs in domestic power supplies and in audio signals, to name two very common examples. A *voltage amplifier* multiplies the positive and negative voltage excursions, using an external power source to achieve this. Most op-amps are voltage amplifiers.

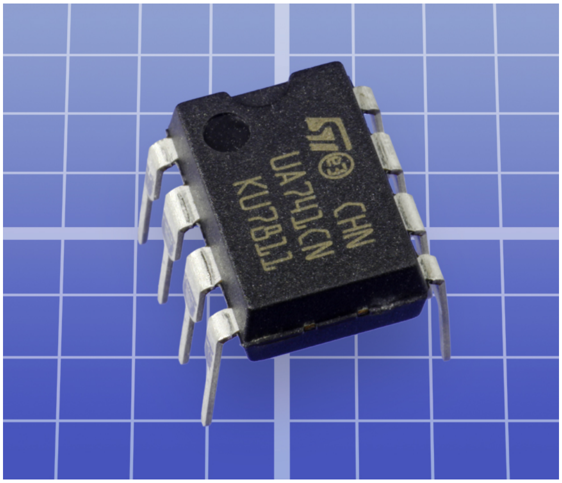

Figure 7-1. *The LM741, shown here, is still one of the most widely used op-amps.*

An ideal amplifier maintains a *linear relationship* between its input and its output, meaning that the output voltage values are a constant multiple of the input voltages over a wide range. This is illustrated in Figure 7-2, where the lower curve is a duplicate of the upper curve, the only difference being that its amplitude is multiplied by a fixed amount (usually much greater than shown here). The ratio is properly known as the *gain* of the amplifier, usually represented with letter A (for amplification).

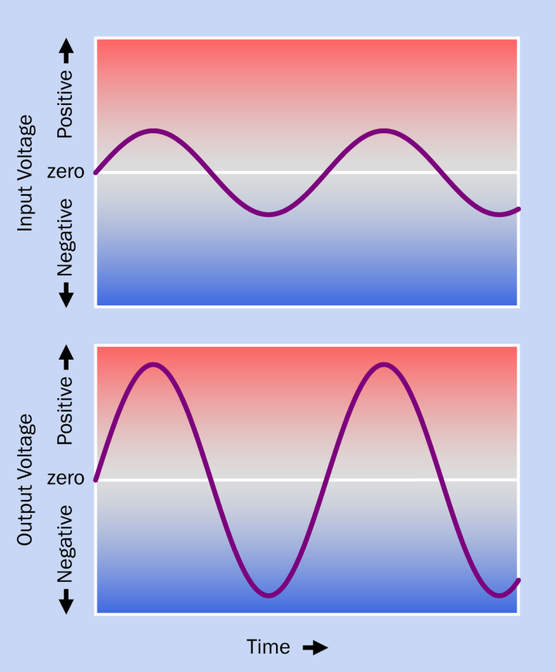

Figure 7-2. *In an ideal voltage amplifier, the output voltage will be a duplicate of the fluctuating input voltage, the only difference being that the amplitude of the output is multiplied by a fixed amount. This ratio is known as the gain of the amplifier.*

Figure 7-3 shows the triangular symbol for a generic single-input amplifier (not an op-amp). It may contain any number of components. The triangle almost always points from left to right, with its input on the left and its output on the right, and power attached above and below. This is often a *dual voltage* power supply, which is convenient for amplifying a signal that fluctuates above and below 0V. In some schematics, the power supply connections may not be shown, as they are assumed to exist.

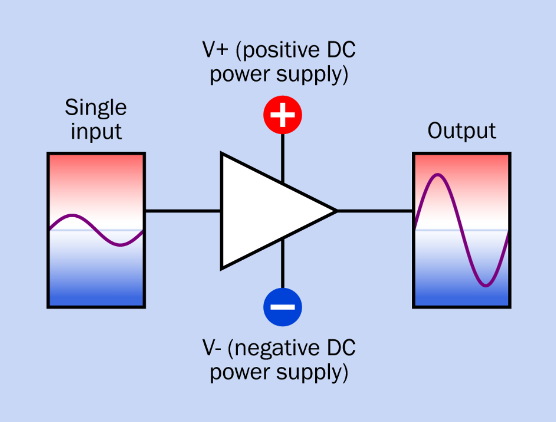

Figure 7-3. *The generic symbol for a single-input amplifier (not an op-amp), with the positive side of its power supply being equal and opposite in value to the negative side, and 0V being at the midpoint between them.*

- While the blue negative symbol is generally used throughout this encyclopedia to indicate 0V ground, it represents a voltage identified as V− in a dual voltage power supply, being equal in value but opposite in polarity to the positive side of the supply, V+. (Sometimes these voltages are indicated as V− and V+.)

The output from this imaginary generic amplifier is shown in the figure as a linear amplification of the input.

Dual Inputs

An op-amp has two inputs instead of one, and amplifies the voltage difference between them. Its symbol is shown in Figure 7-4. The upper input in this figure is held at 0V, midway between V+ and V−. Because the op-amp has so much gain, an accurate reproduction of its input would create an output exceeding the voltage of the power supply. Because this is not possible, the output tends to become saturated and consequently is *clipped* when it reaches its maximum value, as shown in the figure. The thumbnail graphs give only an approximate impression, as they are not drawn to the same scale.

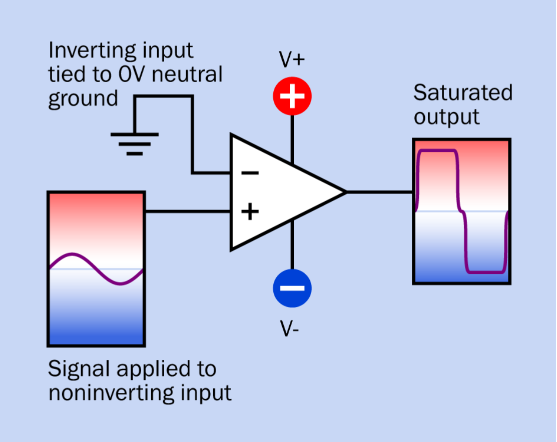

Figure 7-4. *An op-amp has so much gain, its output will tend to saturate, producing a square wave regardless of the shape of the input.*

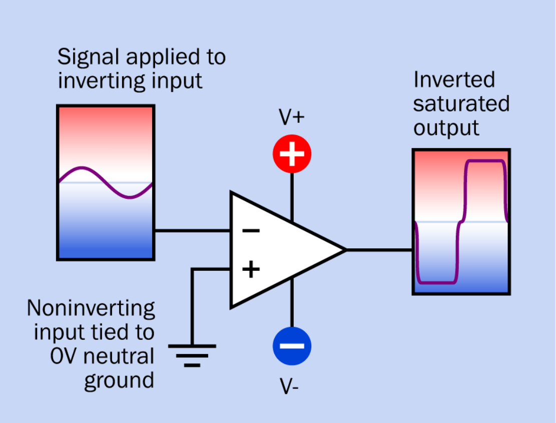

Figure 7-5. *When the incoming signal is applied to the inverting input of an op-amp while its noninverting input is held at 0V ground, the output is inverted.*

The small black plus and minus signs alongside the two inputs to the op-amp have nothing to do with the voltage supplied to the component. The "minus" input is properly referred to as the *inverting input* while the "plus" input is the *noninverting input*, in recognition of their functions.

The inputs are sometimes arranged with the minus above the plus, and sometimes with the plus above the minus. Schematics should be inspected carefully to note which arrangement is being used.

The positive and negative power connections to the op-amp may be omitted, but if shown, they always place V+ at the top, regardless of which way around the inputs are presented.

If a signal is applied to the noninverting input, while the inverting input is held at 0V ground, the op-amp provides an output in which the voltage is not inverted relative to the input.

If the input connections are swapped, so that the inverting input receives the incoming signal while the noninverting input is tied to 0V ground, the output from the op-amp is inverted (the gain remains the same). See Figure 7-5.

- An op-amp that is being used without any other components to moderate its output is functioning in *open loop* mode.

Negative Feedback

To create an output that is an accurate replica of the input, the op-amp must be brought under control with *negative feedback* to the input signal. This is illustrated in Figure 7-6. A resistor connects the output back to the inverting input, so that the input is automatically reduced to the point where the output is no longer saturated. The values of R1 and R2 will determine the gain of the op-amp, as explained in "How to Use It" on page 53. The op-amp is now functioning in its intended *closed loop* mode, meaning that the output is being tapped for feedback.

To obtain a linear output that is noninverted, connections are made as shown in Figure 7-7. The resistors form a voltage divider between the output and 0V ground, effectively increasing the comparison value on the inverting input.

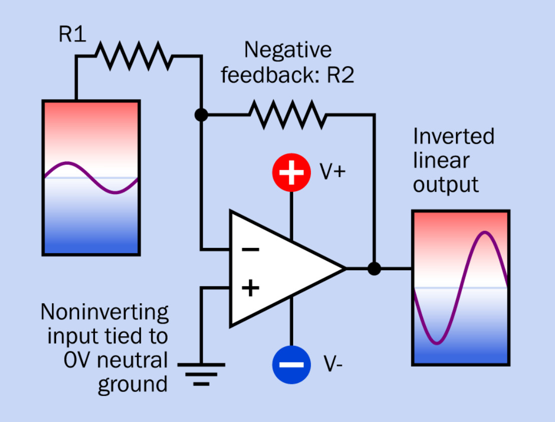

Figure 7-6. *A resistor applies negative feedback to the inverting input of an op-amp, and creates a linear output.*

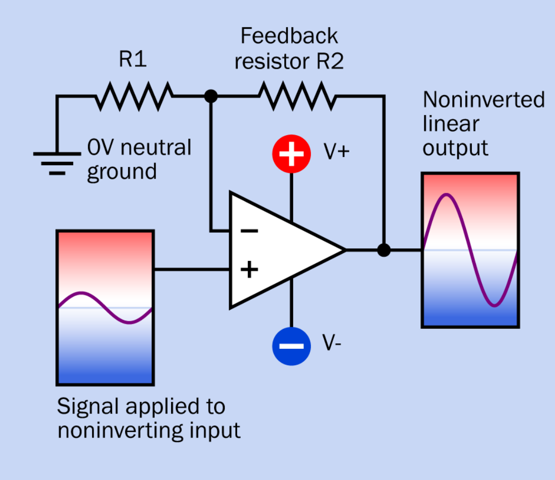
R1
Feedback resistor R2
0V neutral ground
V+
Noninverted linear output
–
+
V-
Signal applied to noninverting input

Figure 7-7. *Where the incoming signal is applied to the noninverting input, negative feedback is created by using a pair of resistors forming a voltage divider between the output and 0V ground.*

- Note that the gain of an op-amp is specific to a particular frequency range of AC signal. This is discussed in "How to Use It" on page 53.

Op-Amps and Comparators

A **comparator** can be regarded as a type of op-amp, and in fact an op-amp can be used as a comparator, comparing a variable DC voltage on one input with a reference voltage on another input. However, the two types of components have diverged in design to the point where they should be considered separately. The distinction is sufficiently important to have prompted Texas Instruments to issue an Application Report in 2001 titled "Op Amp and Comparators—Don't Confuse Them!".

Differences in function are summarized in the previous entry discussing comparators (see "Differences from an Op-Amp" on page 42).

Variants

Because op-amps are mostly low-current devices, they are widely available in very small surface-mount formats, in addition to the through-hole DIP packages which used to be more common.

Many chips are available containing two or more op-amps. This is often expressed as the number of *channels* in the component. A dual chip contains two op-amps, while a quad chip contains four op-amps. Usually all the op-amps in a chip share the same power supply. Bipolar or CMOS transistors may be used.

Because op-amps are widely available in dual and quad packages, it's quite common for a circuit designer to have one op-amp in a chip "left over." The designer may be tempted to use that spare unit as a comparator instead of installing an additional chip. To address this situation, some manufacturers offer hybrid op-amp chips containing an additional comparator. The Texas Instruments TLV2303 and TLV2304 are examples.

Values

The op-amps derived from 1970s designs often tolerate a wide range of power-supply voltages. Plus-or-minus 5VDC to plus-or-minus 15VDC is a common range. Modern op-amps are available that run from as little as 1VDC to as much as 1,000VDC.

Op-amps are available for frequencies ranging from 5KHz all the way up to 1GHz.

A “classic” op-amp such as the LM741, which is still widely used, will operate with a power supply ranging from plus-or-minus 5VDC to plus-or-minus 22VDC. Its output is rated for up to 25mA, and its input impedance is at least 2MΩ. The most current it will draw from an input is around 0.5μA.

V_{IO} is the *input offset voltage*. In an ideal component, the output from an op-amp should be 0V when its inputs have a voltage difference of 0V. In practice, the output will be 0V when the inputs differ by the offset voltage. V_{IO} is likely to be no greater than a couple of mV, and negative feedback can compensate for the offset.

V_{ICR} is the *common mode voltage range*. This is the range of input voltages that the op-amp will tolerate. This can never be more than the positive power supply voltage and will often be less, depending on the types of transistors that are used on the input side. If an input voltage goes outside the common mode voltage range, the op-amp will stop functioning.

V_{IDR} is the *input differential voltage range*—the maximum permissible difference between peak positive and peak negative input voltages. This is often expressed as plus-or-minus the power supply voltage, or slightly less. Exceeding the range can have destructive consequences.

I_B is the *input bias current*, averaged over the two inputs. Most op-amps have extremely high input impedance and consequently use very low input currents.

Slew rate at unity gain is the rate of change of the output voltage caused by an instantaneous change on the input side, when the output of the op-amp is connected directly back to the inverting input (during operation in noninverting mode).

How to Use It

In addition to being an amplifier for AC signals, an op-amp can serve as an oscillator, filter, signal conditioner, actuator driver, current source, and voltage source. Many applications require some understanding of the complexities of mathematics describing alternating current, which are not included this encyclopedia. Almost all the applications have a common starting point, however, which is to establish and control the gain of the feedback circuit.

Controlling the Gain

"A_{VOL} is the *open-loop voltage gain*, defined as the maximum voltage amplification that can be achieved when no feedback is applied from the output to an input. This remains constant until the AC frequency rises to a point known as the *breakover frequency*. If the frequency continues to rise, the maximum gain diminishes quite rapidly, until finally it terminates in 1:1 amplification at the *unity gain frequency*. This transition is shown by the orange line in Figure 7-8. The length of each purple line shows the frequencies which can be tolerated when the op-amp is used in closed-loop mode, and a negative feedback loop limits the gain. For example, where the gain is just 10:1, it can remain constant to just above 10KHz.

Note that both of the scales in this graph are logarithmic.

Calculating Amplification

So long as an op-amp is used within the boundaries of the graph, its voltage amplification can be controlled by choosing appropriate feedback and input resistors. If the op-amp is being used in noninverting mode, and R1 and R2 are placed as shown in Figure 7-7, the amplification ratio, A, is found approximately by the formula:

```
A = (approximately) 1 + (R2 / R1)
```

From this it can be seen that when R1 is very large compared with R2, the gain diminishes to near unity. If R1 becomes infinite and R2 is zero, the gain is exactly 1:1. This can be achieved by replacing R2 with a section of wire (theoretically of zero resistance) and omitting R1 entirely, as in Figure 7-9. In this configuration, the output from the op-amp should be identical with its input.

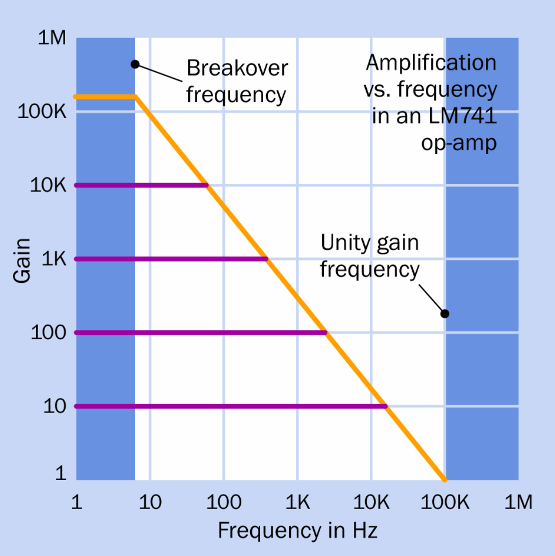

Figure 7-8. *Where each horizontal purple line meets the diagonal orange line, this is the maximum frequency that can be used without reduction in the maximum gain of an op-amp.*

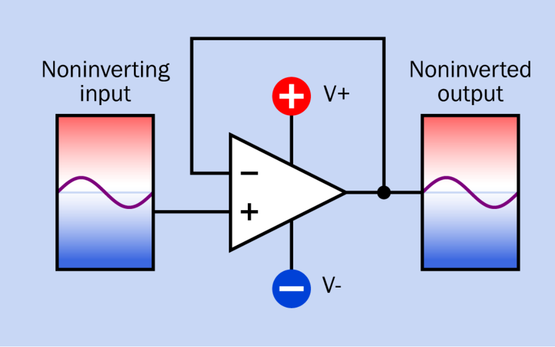

Figure 7-9. *While an op-amp is in noninverting mode, if the feedback resistor is replaced with a section of wire and the 0V ground connection is omitted entirely, the gain of the op-amp diminishes theoretically to 1:1.*

If the op-amp is being used in inverting mode, and R1 and R2 are placed as shown in Figure 7-6, then the voltage amplification ratio, A, is found approximately by the formula:

```
A = (approximately) −(R2 / R1)
```

- Note the minus sign. In inverting mode, gain is expressed as a negative number.
- In a practical circuit, at the expected frequency, the amplification factor established by choice of resistors should be no more than 20.
- An inverting circuit has a relatively low input impedance. For this reason, in most applications, a noninverting circuit is preferred.

Unintentional DC Voltage Amplification

Although the op-amp is intended primarily as an AC signal voltage amplifier, it will also amplify a DC difference between the voltages on its inputs. In the upper section of Figure 7-10, a positive DC offset is inverted and amplified to the point where the output is forced to its negative limit, and the signal is lost, because its fluctuations have been overwhelmed by the positive offset. A coupling capacitor (shown in the lower section of the figure) removes the DC voltage while passing the AC signal. The appropriate capacitor value will depend on the frequency of the signal.

Low-Pass Filter

An op-amp can facilitate a very simple low-pass audio filter, just by adding a capacitor to the basic inverting circuit previously shown in Figure 7-6. The filter schematic is shown in Figure 7-11. Capacitor C1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic inverting circuit is approximately −(R2 / R1), the op-amp functions normally when the impedance of C1 is blocking the low frequencies, forcing them to pass through R2. Higher frequencies, however, are able to bypass R2 through C1, lowering the effective resistance of the feedback section of the circuit, thus reducing its gain. This way, the power of the op-amp is greatly reduced for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

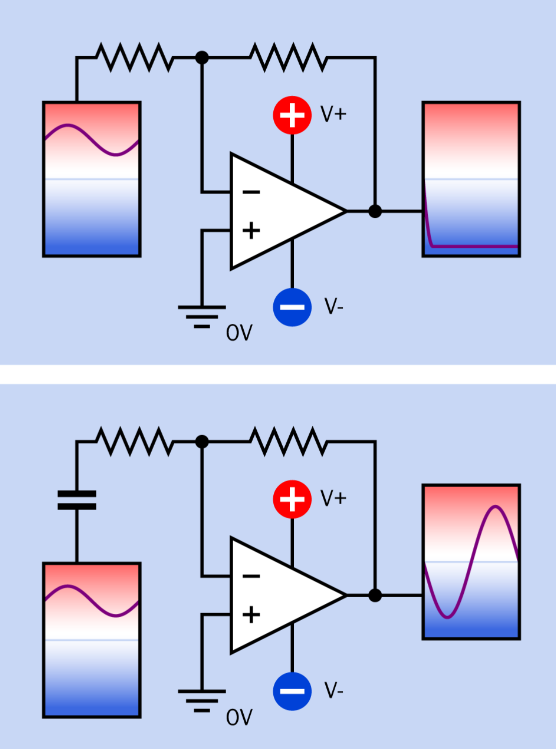

Figure 7-10. *The addition of a capacitor at the input of an op-amp is often necessary to prevent any DC voltage offset being amplified. In the upper section of this figure, a DC offset is large enough to force the inverted output to its negative limit, and the signal is completely lost.*

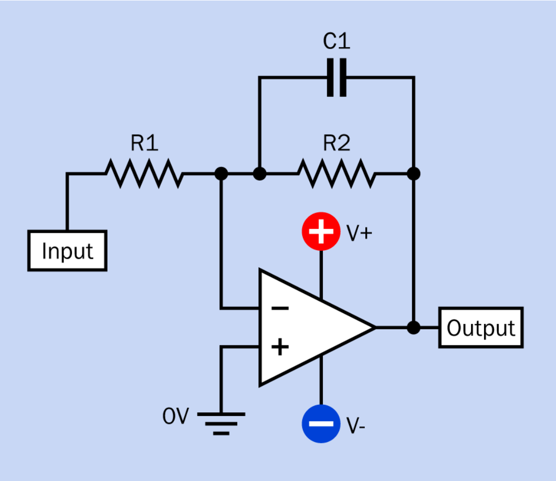

Figure 7-11. *A very basic low-pass filter, which works by allowing capacitor C1 to bypass resistor R2 at higher audio frequencies.*

High-Pass Filter

A simple high-pass filter can be created by adding a capacitor to the basic noninverting circuit previously shown in Figure 7-7. The filter schematic is shown in Figure 7-12. Once again capacitor C1 is chosen with a value that passes higher audio frequencies and blocks lower audio frequencies. Because the gain of the basic noninverting circuit is approximately 1 + (R2 / R1), the op-amp functions normally when the impedance of C1 is blocking the low frequencies, forcing them to pass through R1. Higher frequencies, however, are able to bypass R1 through C1, lowering the effective resistance of that section of the circuit, thus reducing the negative feedback and increasing the gain. This way, the power of the op-amp is increased for higher frequencies compared with lower frequencies. A passive RC circuit could achieve the same effect, but would attenuate the signal, while the op-amp circuit boosts part of it.

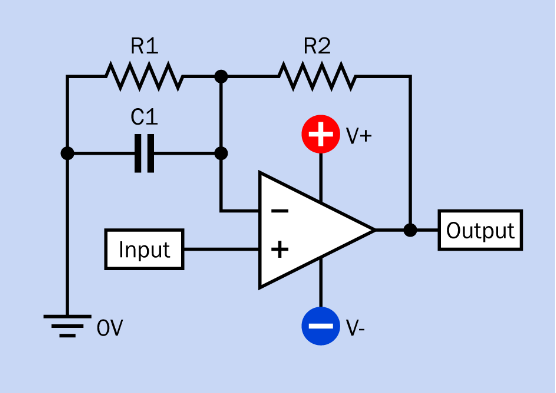

Figure 7-12. *A very basic high-pass filter, which works by allowing capacitor C1 to bypass resistor R1 at higher audio frequencies.*

Relaxation Oscillator

The schematic in Figure 7-13 is similar to the circuit shown in Figure 6-12 using a comparator. It functions as a *relaxation oscillator*, which is a form of *astable multivibrator*. The lower half of the circuit is a positive feedback loop that reinforces the output while the upper half of the circuit is charging the capacitor. Eventually the charge on the capacitor exceeds the voltage on the noninverting input of the op-amp, creating negative feedback that exceeds the positive feedback. The capacitor discharges and the cycle repeats. The component values in the figure should generate an output that runs at around 2Hz. Reducing the value of the capacitor will increase the frequency.

Single Power Source

A few op-amps are designed to work from single voltages, but they are a relatively small minority, and will clip the output signal if the input goes negative. Power supplies are readily available that provide multiple voltages such as +15VDC, 0V, and -15VDC. They are ideal for driving an op-amp—but may not be useful for any other components in the circuit. Can an op-amp that is designed for dual voltages be made to run from a single supply, such as 30VDC?

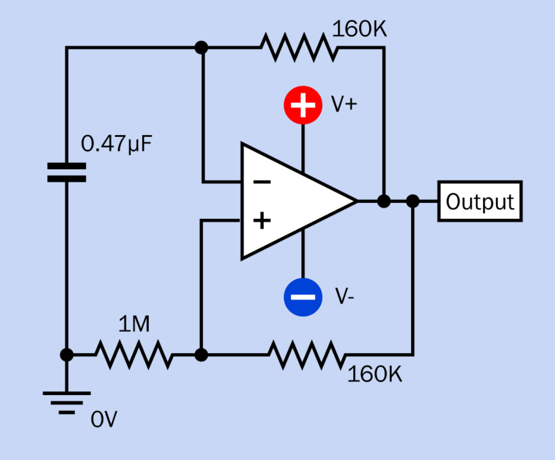

Figure 7-13. *A relaxation oscillator.*

This is relatively easy to do. The op-amp simply needs a potential difference to power its internal transistors, and 30VDC on the V+ pin with 0VDC on the V– pin will work just as well as +15VDC and -15VDC. However, referring back to Figure 7-6, if the op-amp is used in inverting mode, an intermediate voltage must be supplied to the noninverting input. Likewise, in noninverting mode, an intermediate voltage is necessary for one of the inputs, and must be half-way between the extremes of the power supply. If the supply is +15VDC and -15VDC, the midpoint is 0V. If the supply is 30VDC and 0V, the midpoint is 15VDC.

Because the inputs of an op-amp have a very high impedance and draw negligible current, the intermediate voltage can be provided with a simple voltage divider, as shown in Figure 7-14, where R3 and R4 should be no greater than 100K each. Their exact values are not important, so long as they are equal.

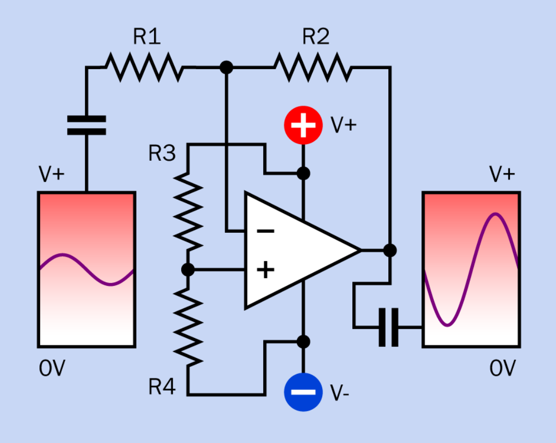

Figure 7-14. *A voltage divider, comprised of R3 and R4 in this schematic, can provide a voltage halfway between V+ and negative ground, enabling the op-amp to use just one power supply instead of two.*

A coupling capacitor should still be used on the input side, as shown, because there is no guarantee that the input signal will be centered precisely on 15V, and any offset will be amplified, potentially causing clipping of the signal. For similar reasons, a coupling capacitor is also added on the output side.

Offset Null Adjustment

Some op-amps provide two pins for *offset null adjustment*, which is a setup process to ensure that identical voltage on the two inputs will produce a null output. This is a way of compensating for any internal inconsistencies introduced during the manufacturing process.

To perform offset null adjustment, both input pins are connected directly to 0V ground, and the ends of a trimmer potentiometer (typically, 10K) are connected with the offset null pins, while the wiper of the potentiometer is centered and then connected with the negative power supply. The probes of a meter that is set to measure DC volts are placed between the output of the op-amp and 0V ground. The potentiometer is then adjusted until the meter shows a reading of 0VDC. A schematic is shown in Figure 7-15.

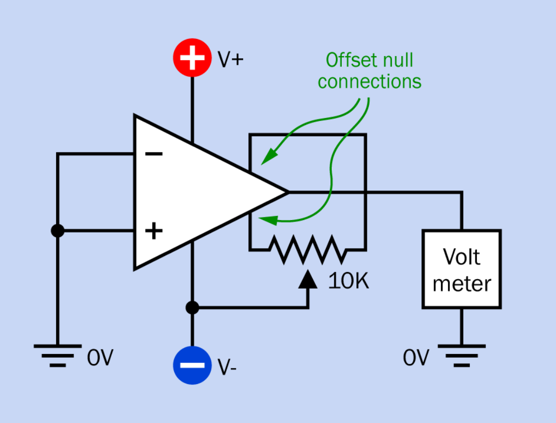

Figure 7-15. *Connections for making an offset null adjustment to an op-amp that allows this procedure.*

What Can Go Wrong

Power Supply Problems

Op-amps are especially vulnerable to reversed polarity in a power supply. If there is even a remote possibility of this occurring, a diode in series with one side of the supply can provide protection.

A more realistic concern is the destructive consequence of an input signal that exceeds the power supply voltage(s) of the op-amp. Even if the input is within the acceptable range, it can still cause permanent damage if it is applied before the op-amp powers up.

Bad Connection of Unused Sections

Multiple op-amps are often combined in a single package. If some of these "sections" remain unused, they will still receive power from the shared supply, and will attempt to function. If the inputs are left unconnected, they will pick up small stray voltages by capacitance or induction, and in the absence of negative feedback, the op-amp will create unpredictable outputs, consuming power and possibly interacting with other sections of the same chip. Figure 7-16 shows three incorrect options for addressing this problem, and one recommended option (derived from Texas Instruments Application Report SLOA067).

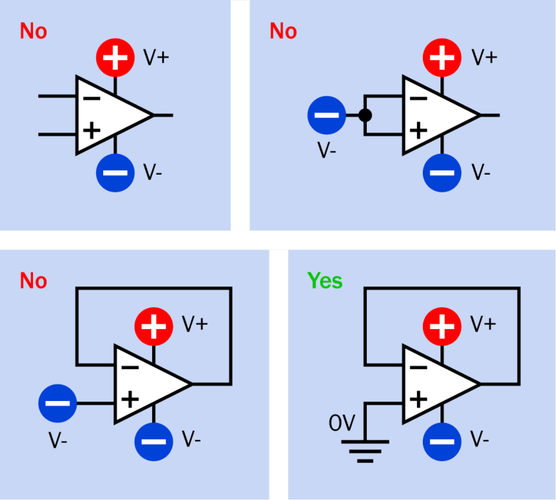

Figure 7-16. *When multiple op-amps share a chip, one that is unused will still receive power from the shared supply. Its inputs must not be allowed to float, and must be connected to minimize activity and power consumption. Three common errors are shown here, with one recommended option. Note the distinction between 0V ground (0V) and negative power (V−).*

Oscillating Output

The inputs of an op-amp are vulnerable to stray electromagnetic fields. If the conductors leading to and from the op-amp are relatively long, the output can couple capacitively with the input during voltage transitions, causing unwanted oscillations.

The commonly recommended solution to this problem is to add a 1μF bypass capacitor between the power supply and 0V ground. However, some manufacturers recommend alternatives such as introducing a very small amount of hysteresis, or reducing the value of input resistors to below 10K.

Confused Inputs

A schematic may show an op-amp with the noninverting input above the inverting input, or vice versa. The only indication of this will be the plus and minus signs inside the chip, which can be extremely small and easily overlooked. For convenience in drawing a diagram, two op-amps in the same circuit may have their inputs shown in opposite configurations. Special care must be taken to verify that the inverting and noninverting input pins on a chip are correctly assigned.

digital potentiometer

A **digital potentiometer** is also known as a *digitally adjustable potentiometer*, a *digitally controlled potentiometer*, a *digitally programmed potentiometer* (with acronym *DPP*), a *digpot*, or a *digipot*. The terms are functionally interchangeable. Because the abbreviation *pot* is often used to describe an analog potentiometer, some people refer to digital potentiometers colloquially as *digital pots*. In printed documentation, the letters in *pot* may be capitalized. Because it is an abbreviation, not an acronym, it is not capitalized here.

Because this component enables digital control of a variable voltage, it is a *mixed signal device*. It is classified here as an analog chip because it primarily emulates the function of an analog device. It may be thought of as a form of *digital-analog converter*, although this encyclopedia does not have a section devoted to that type of component or to *analog-digital converters*, as their application is relatively specialized.

OTHER RELATED COMPONENTS

- **potentiometer** (see Volume 1)

What It Does

This component is an integrated circuit chip that emulates the function of an analog **potentiometer**. It is often described as being *programmable*, meaning that its internal resistance can be changed via a control input.

Digital potentiometers are particularly suited for use in conjunction with a *microcontroller*, which can control the internal resistance of the component. Possible applications include adjustment of the pulse width of an oscillator or multistable multivibrator (e.g., using the Control pin of a 555 **timer** chip); adjustment of the gain in an **op-amp**; specification of voltage delivered by a *voltage regulator*; and adjustment of a *bandpass filter*.

A digital potentiometer in combination with a microcontroller may also be used in conjunction with a pair of external buttons or a rotational encoder, to adjust the gain of an audio amplifier and for similar applications.

Advantages

A digital potentiometer offers significant advantages over an analog potentiometer:

- Reliability. The digital component may be rated for as many as a million cycles (each storing the wiper position in an internal memory location). An analog component may be capable of just a few thousand adjustment cycles.
- Digital interface.
- Elimination of long signal paths or cable runs. The digital potentiometer can be placed close to other chips, whereas an analog potentiometer often has to be some distance away to enable control by the end user. Reduction in the length of signal paths can reduce capacitive effects, while elimina-

tion of cable runs will reduce manufacturing costs.

- Reduction in size and weight compared with a manual potentiometer.

A digital potentiometer also has some disadvantages:

- Its internal resistance is somewhat affected by temperature.
- It is not usually capable of passing significant current. Few chips can sink or source more than 20mA at the output, and 1mA is common. The output is primarily intended for connection with other solid-state components that have high impedance.
- Users may prefer the immediacy and tactile feel of a knob attached to an analog potentiometer, rather than a pair of buttons or a rotational encoder.

How It Works

A digital potentiometer changes the point at which a connection is made along a *ladder* of many fixed resistors connected in series inside the chip. Each end of the ladder, and each intersection between two adjacent resistors, is known as a *tap*. The pin that can connect with any of the taps is referred to as the *wiper*, because it emulates the function of a wiper in an analog potentiometer. In reality, a digital potentiometer does not contain a wiper or any other moving parts.

A fully featured digital potentiometer allows access to each end of the ladder through two pins that are often labeled "high" and "low," even though they are functionally interchangeable (except in the case of a component that simulates a logarithmic taper, as described later). The "low" end of the ladder is sometimes numbered 0. In this case, if there are *n* resistors, the "high" end of the ladder will be numbered *n*. Alternatively, if the "low" end of the ladder is numbered 1, and there are *n* resistors, the "high" end will be numbered *n+1*. This principle is illustrated in Figure 8-1.

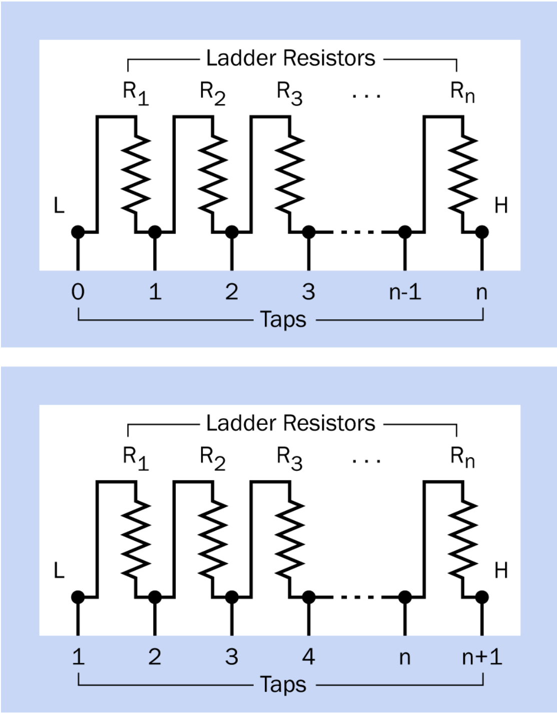

Figure 8-1. *Available wiper connections to a resistor ladder inside a digital potentiometer, showing two numbering systems that may be used.*

The "low" pin on a digital potentiometer may be identified as L, or A, or R_L, or P_A in a datasheet, while the "high" pin may be identified as H, or B, or R_H, or P_B and the pin that accesses the wiper is typically identified as W, or R_W, or P_W. Letters L, H, and W are used below. Although the L and H pins are functionally interchangeable, their labels are useful to identify which direction the W connection will move in response to an external signal.

Digital potentiometers are available with as few as 4 or as many as 1,024 taps, but common values are 32, 64, 128, or 256 taps, with 256 being the most common.

No specific schematic symbol represents a digital potentiometer. Often the component is shown as an analog potentiometer symbol inside a box that has a part number, as suggested in

Figure 8-2. Control pins and the power supply may be omitted if the schematic is just intended to show logical connections. Alternatively, if the digital potentiometer is depicted in a schematic where it is connected with other components such as a microcontroller, multiple pins and functions may be included, as shown in Figure 8-3. The pins additional to L, H, and W are explained below.

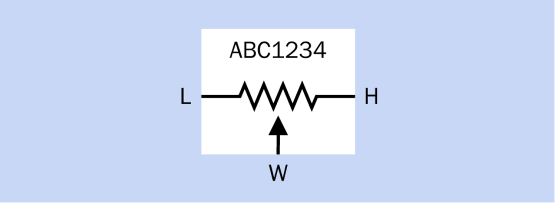

Figure 8-2. *There is no single specific symbol to represent a digital potentiometer. It may be shown using an analog potentiometer symbol in a box with a part number, as suggested here, where power connections and additional pins are omitted for clarity.*

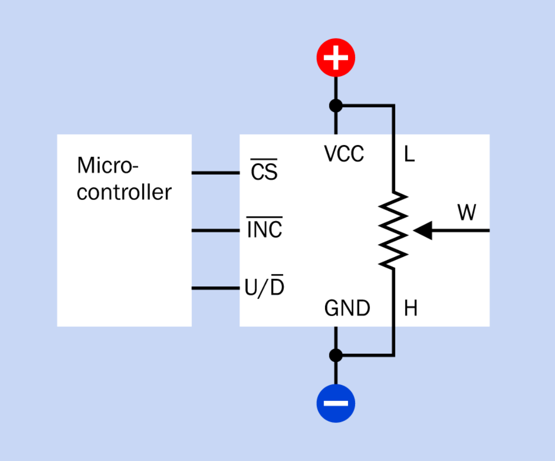

Figure 8-3. *If a digital potentiometer appears in a schematic where it is connected with other components such as a microcontroller, additional pins and functions may be indicated. This generic representation of a digital potentiometer shows some of the functions that can be included.*

Variants

A *dual* digital potentiometer contains two complete units, while a *quad* contains four. Triples exist but are relatively uncommon. A few chips contain six potentiometers. Multiple digital potentiometers in a chip can be used as the digital equivalent of *ganged* analog potentiometers, for simultaneous synchronized adjustment of multiple inputs in an audio system (two channels in a stereo amplifier, or more in a surround-sound system).

The pinouts of a sophisticated quad digital potentiometer chip are shown in Figure 8-4. Other quad chips have different pinouts and capabilities; there is no standardized format as there is with digital logic chips. In this example, the high/low states of Address 0 and Address 1 select one of the four internal resistor ladders, numbered 0 through 3. The Chip Select pin makes the whole chip either active or inactive. The Write Protect pin disables writing to the internal wiper memory. The Serial Clock pin inputs a reference pulse stream to which the serial input data must be synchronized. The Hold pin pauses the chip while data is being transmitted, allowing the data transmission to be resumed subsequently. The NC pins have no connection.

Volatile and Nonvolatile Memory

Any type of digital potentiometer requires memory to store its current wiper position, and this memory may be volatile or nonvolatile. Nonvolatile memory may be indicated in a datasheet by the term *NV*.

A digital potentiometer with volatile memory will typically reset its wiper to a center-tap position if power is disconnected and then restored. A digital potentiometer with nonvolatile memory will usually restore the most recently used wiper position, provided the chip is fully powered down and then fully powered up without glitches in the supply. If a microcontroller is being used to control the digital potentiometer, it can store the most recent resistance value in its own nonvolatile memory, in which case the type of memory in the potentiometer becomes irrelevant.

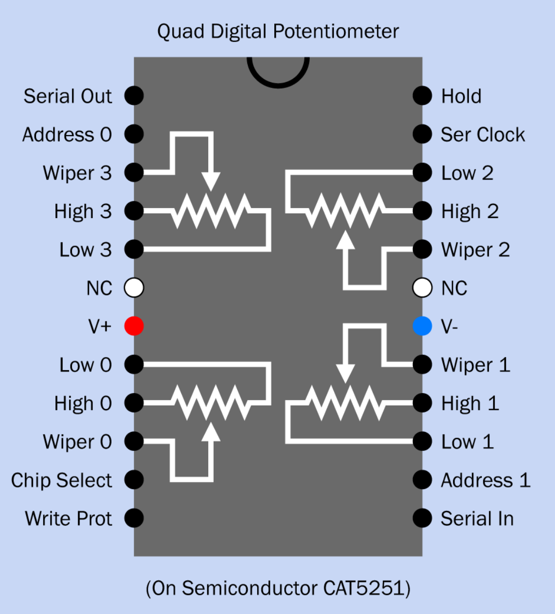

Figure 8-4. *Pinouts of a sophisticated quad digital potentiometer chip. Other chips will have different pinouts and capabilities. This example is available in surface-mount formats only. See text for details.*

Taper

Digital potentiometers are available with *linear taper* or *logarithmic taper*. In the former, each resistor in the ladder has the same value. In the latter, values are chosen so that the cumulative resistance between the wiper and the L end of the ladder increases geometrically as the wiper steps toward the H end of the ladder. This is useful in audio applications where sound intensity that increases exponentially may seem to increase linearly when perceived by the human ear.

A microcontroller can emulate logarithmic steps by skipping some taps in the ladder in a digital potentiometer, but this will result in fewer increments and lower precision.

Data Transfer

Digital potentiometers are mostly designed to use one of three serial protocols:

- *SPI*. This acronym is derived from *serial peripheral interface*, a term trademarked by Motorola but now used generically. The standard is adapted in various radically different ways among digital potentiometers.
- *I2C*. More correctly printed as *I^{2}C* and properly pronounced “I squared C,” this acronym is derived from the term *inter-integrated circuit*. Developed by Philips in the 1990s, it is a relatively slow-speed bus-communication protocol (up to 400kbps or 1Mbps in its basic form). It is built into some microcontrollers. The standard is more uniformly and rigorously defined than SPI.
- *Up/down*, also sometimes known as *pushbutton* or *increment/decrement* protocol.

Both SPI and I2C are supported by many microcontrollers, including the Atmel AVR at the heart of the Arduino.

These three systems for controlling a digital potentiometer are described in more detail in the following sections.

SPI

This is the most widely used protocol, but when reading datasheets, a lot of care must be taken to determine how it varies in each case.

The Microchip 4131-503, shown in Figure 8-5, uses SPI protocol. It contains 128 resistors and can be powered by 1.8VDC to 5.5VDC.

The one feature that all versions of SPI have in common is that a series of high/low pulses is interpreted by the chip as a set of bits whose value defines a tap point in the resistor ladder. In computer terminology, every tap point has an *address*. The incoming bits define the address, after which the status of an additional input pin can tell the chip to move the wiper to that location.

Typically, there will be a chip select pin, identified as CS; a serial data input pin, identified as SDA, SI, DIN, or a similar acronym; and a serial clock pin, identified as SCL, SCLK, or SCK, which must receive a stream of pulses to which the high/low data input pulses must be synchronized. In addition, the SPI protocol allows bidirectional (duplex) serial communication. Only a minority of digital potentiometers make use of this capability, but where it exists, a serial data output pin may be labeled SDO. Alternatively, one pin may be multiplexed to enable both input and output, in which case it may be labeled SDI/SDO.

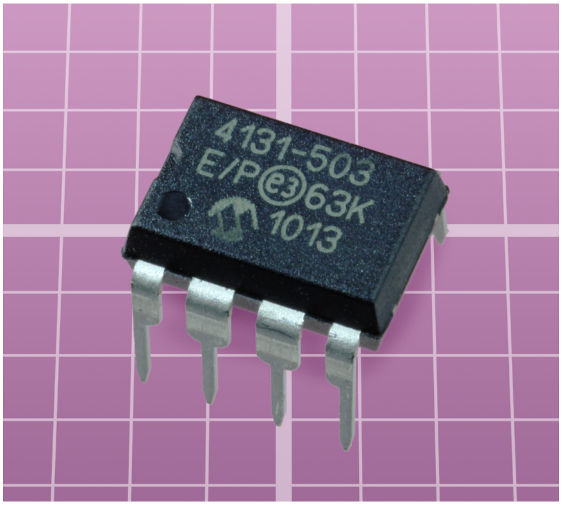

Figure 8-5. *This digital potentiometer uses SPI protocol. See text for details.*

If a pin is active-low, a bar (a horizontal line) will be printed above its acronym.

The most common type of digital potentiometer has 255 resistors and therefore 256 tap points, which have addresses numbered 0 through 255, each of which can be specified by a sequence of eight data bits constituting one byte. However, a different coding system will be applied in chips that have a different number of taps. In a 32-tap component, for instance, data is still sent in groups of eight bits, but only the first five bits define a *tap address*, while the remaining three are interpreted as commands to the chip.

Most 256-tap chips use an SPI protocol in which two eight-bit bytes are sent, the first being interpreted by the chip as a command, while the second specifies a tap address. Each manufacturer may use a different set of command codes, and these will vary among chips even from the same manufacturer.

Most commonly, three wires are used for data transmission and control (causing these chips to be described as *3-wire* programmable potentiometers).

CS is usually, but not always, pulled low to activate the digital potentiometer for input. A series of low or high states is then applied to the data-input pin. Each time the clock input changes state (usually on the rising edge of the clock pulse) the state of the data input is copied to a **shift register** inside the chip. After all the bits have been clocked in, CS can change from low to high, causing the contents of the shift register to be copied into a decoder section of the chip. The first bit received becomes the most significant bit in the decoder. The value of the eight bits is decoded, and the chip connects the W pin directly to the corresponding tap along the ladder of 255 internal resistors.

I2C Protocol

The I2C specification is controlled by NXP Semiconductors (formerly Philips), but can be used in commercial products without paying licensing fees. Only two transmission lines are required: one carrying a clock signal, the other allowing bidirectional data transfer synchronized with the clock (although many digital potentiometers use the I2C connection only to receive data). The pins are likely to be identified by the same acronyms as the pins on a chip that uses SPI protocol.

As in SPI, a command byte is followed by a data byte, although the command set differs from that of SPI and will also differ among various I2C chips. Full implementation of I2C allows multiple devices to share a single bus, but this capability may remain unused.

Up/Down Protocol

This simpler, asynchronous protocol does not require a clock input. The chip will respond to data pulses that are received at any speed (up to its maximum speed), and the pulse widths can be inconsistent.

Each pulse moves the wiper connection one step up or down the ladder. While this has the advantage of simplicity, the taps are not addressable, and consequently the wiper cannot skip to any tap without passing through intervening taps incrementally. This is not an inconvenience when the potentiometer controls audio gain, which is a primary application.

In some chips, an increment pin, usually labeled INC, receives pulses while the high or low state of a second pin, usually labeled U/D, determines whether each pulse will step the wiper up the ladder or down the ladder.

In other chips, pulses to an Up pin will step the wiper up the ladder, while pulses to a Down pin will step the wiper down the ladder.

Either of these chip designs can be referred to as a *two-wire* type. If an additional chip-select pin is included (labeled CS on datasheets), this type of digital potentiometer can be referred to as a *three-wire* type. The chip select pin is likely to be active-low, meaning that so long as it has a high state, the chip will ignore incoming signals.

The CAT5114 shown in Figure 8-6 uses an U/D pin. It contains 31 resistors, is available in 8-pin DIP or surface-mount formats, and can be powered by 2.5VDC to 6VDC. Each of its logic inputs draws only 10µA.

In six-pin chips the INC pin is omitted, one of the H, L, or W pins will be omitted, and the U/D pin will function differently. When CS is pulled low, the chip checks the state of the U/D pin. If it is high, the chip goes into increment mode; if it is low, the chip goes into decrement mode. So long as CS remains low, each transition of the U/D pin from low to high will either increment or decrement the wiper position, depending on the mode that was sensed initially. When CS goes high, further transitions on the U/D pin will be ignored until CS goes low again, at which point the procedure repeats.

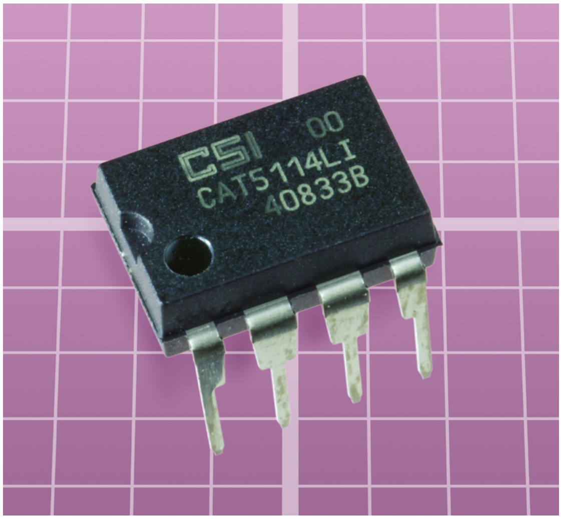

Figure 8-6. *This digital potentiometer contains 31 resistors and uses the simplest up/down protocol to step from one tap to the next.*

The chip does not provide any feedback regarding the position of its wiper, and consequently a control device such as a microcontroller cannot know the current wiper position. If the chip has nonvolatile memory (as is the case in many up/down digital potentiometers), it will resume its previous wiper location at power-up, but here again a control device will have difficulty determining what that position is. Therefore, in its basic form, an up/down chip is only appropriate for simple tasks, especially in response to up/down pushbuttons.

Other Control Systems

A few digital potentiometers use a parallel interface. Because they are relatively uncommon, they are not included here.

Connections and Modes

Some variants of digital potentiometers minimize the chip size and number of connections by limiting accessibility to the internal resistor ladder. In a chip designed to function in *rheostat mode*, the W pin is eliminated and the chip moves an internal connection point to change the resistance between the H and L pins.

In some variants, the low end of the ladder is permanently, internally connected with ground, and the L pin is omitted. In other variants, one end of the ladder is unconnected inside the chip.

A chip designed to function in *voltage divider mode* will include all three pins—H, L, and W—except in some instances where the low end of the ladder is grounded internally.

Variants are shown in Figures 8-7, 8-8, 8-9, and 8-10. Because some pins may be omitted, and there is no standardization of function among the pins that do exist, circuits and chips must be examined carefully prior to use.

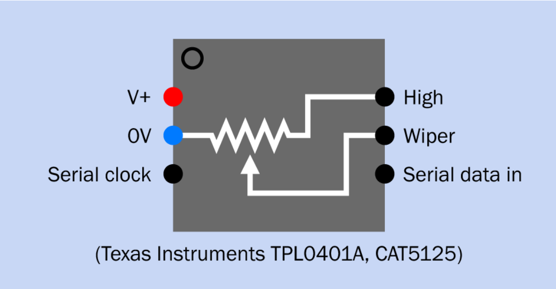
V+
0V
Serial clock
High
Wiper
Serial data in
(Texas Instruments TPL0401A, CAT5125)

Figure 8-7. *Some digital potentiometers minimize chip size and provide specialized functionality by eliminating pins. In the variant shown here, the W pin provides a voltage between H and an internal ground connection. The chip is controlled via I2C serial protocol.*

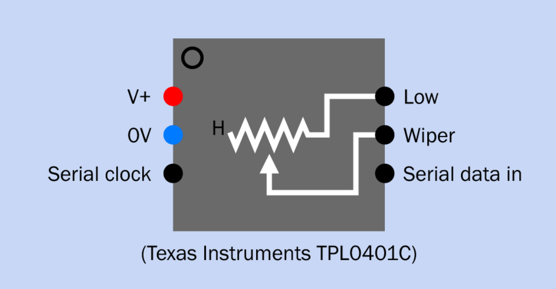
V+
0V
Serial clock
H
Low
Wiper
Serial data in
(Texas Instruments TPL0401C)

Figure 8-8. *In this variant, the H end of the internal resistor ladder is allowed to float inside the chip, and the digital potentiometer functions as a rheostat. The chip is controlled via I2C serial protocol.*

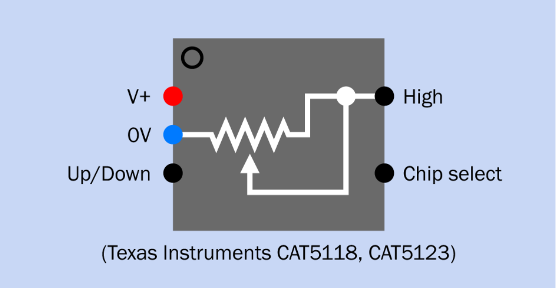
V+
0V
Up/Down
High
Chip select
(Texas Instruments CAT5118, CAT5123)

Figure 8-9. *This variant provides a variable resistance between the H pin and an internal connection with negative ground. Pin 5 is omitted. The chip is controlled by up/down pulses.*

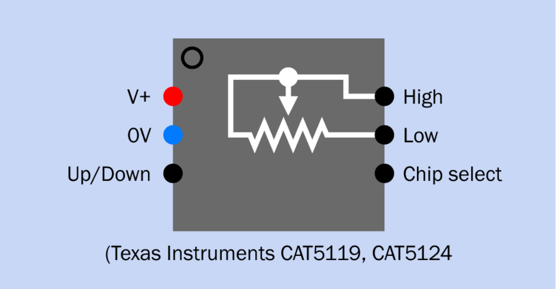
V+
0V
Up/Down
High
Low
Chip select
(Texas Instruments CAT5119, CAT5124

Figure 8-10. *This variant provides a variable resistance between H and L pins, without allowing either end of the resistor ladder to float. The W pin is omitted, as the wiper is tied internally to the H pin. The chips listed are controlled by up/down pulses.*

Values

A primary limitation of digital potentiometers is that they cannot withstand significant current. This may prevent them from being substituted for an analog potentiometer unless changes are made in the circuit. H, L, and W pins are usually unable to source or sink continuous, sustained current exceeding 20mA.

Wiper resistance is the resistance that is added internally by the wiper. This is nontrivial; it is often around 100Ω, and can be as high as 200Ω.

Typical end-to-end resistance of the ladder of internal resistors may range from 1K to 100K. Values of 1K, 10K, and 100K are common.

While the number of taps is likely to be a power of two in chips where the taps are addressable, up/down chips are not so constrained and may contain, for example, 100 taps.

The end-to-end resistance of a whole ladder may vary by as much as 20% from one sample of a chip to the next. Among resistor ladders in digital potentiometers sharing the same chip (i.e., in dual or quad chips) the variation will be much smaller.

Almost all digital potentiometers are designed for a supply voltage of 5V or less. The H and L pins are not sensitive to polarity, but the voltage applied to either of them must not exceed the supply voltage.

How to Use It

While most microcontrollers contain one or more analog-digital converters that change an analog *input* to an internal numeric value, a microcontroller cannot create an analog *output*. A digital potentiometer adds this functionality, although applications will be restricted by its limitation on current.

An up/down digital potentiometer can be controlled directly by a pair of **pushbuttons**, one of which will increase the resistance value while the other will reduce it. The pushbuttons must be *debounced* when used in this way. An alternative to pushbuttons is a **rotational encoder**, which emits a stream of pulses when its shaft is turned. In this case, an intermediate component (probably a microcontroller) will usually interpret the pulse stream and change it to a format that the digital potentiometer can understand.

Where a digital potentiometer is used in audio applications, it should be of the type that moves the wiper connection from one tap to another during a *zero crossing* of the audio signal (i.e., that is, at the moment when the AC input signal passes through 0V on its way from positive to negative or negative to positive). This suppresses the "click" that otherwise tends to occur during switching. Potentiometers with this feature may include phrases such as "glitch free" in their datasheets.

Digital potentiometers that are intended primarily for audio applications often have 32 taps spaced at intervals of 2dB. This will be sufficient to satisfy most listeners.

Achieving Higher Resolution

For sensitive applications where a resolution with more than 1,024 steps is required, multiple digital potentiometers with different step values can be combined. One way of doing this is shown in Figure 8-11. In this circuit, the wipers of P2 and P3 must be moved in identical steps, so that the total resistance between the positive power supply and negative ground remains constant. These two potentiometers could be contained in a dual chip, and would receive identical up/down commands. P1 is at the center of the voltage divider formed by P2 and P3, and is adjusted separately to "fine tune" the output voltage that is sensed at point A.

If all three of the potentiometers in this circuit contain 100 taps, a combined total of 10,000 resistance steps will be possible.

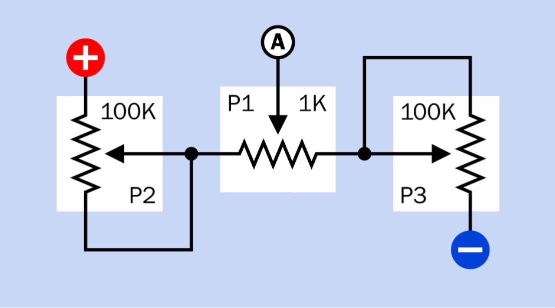

Figure 8-11. *If all three digital potentiometers in this schematic have 100 taps, and the wipers of P2 and P3 are moved in synchronization, the voltage measured at point A can have a high resolution of up to 10,000 steps.*

What Can Go Wrong

Noise and Bad Inputs

Because a digital potentiometer is capable of receiving data at speeds as high as 1MHz, it is sensitive to brief input or power fluctuations, and can misinterpret them as instructions to move the wiper—or can misinterpret them as command codes, in a component using SPI or I2C serial protocol.

To minimize noise in the power supply, some manufacturers recommend installing a 0.1µF capacitor as close as possible to the power supply pin of the component. In addition, it is obviously important to provide clean input signals. This means thorough debouncing of any electromechanical switch or pushbutton inputs.

Wrong Chip

The wide diversity of input protocols and pinouts creates many opportunities for installation error.

Up/down, SPI, and I2C protocol require totally different pulse streams. Many manufacturers offer components that are distinguished from each other by just one or two digits in their part numbers, yet have radically different functionality.

If more than one specific type of digital potentiometer may be used during circuit development, they should be stored carefully to avoid inadvertant substitutions. Using the wrong chip may be particularly confusing in that an inappropriate input protocol will still produce some results, although not those which were intended.

Controller and Chip Out of Sync

As noted in the discussion of data transmission protocols, most digital potentiometers are not capable of providing feedback to confirm the position of the internal wiper. A designer may wish to include a power-up routine which establishes the state of the digital potentiometer by resetting it to a known position, at one end of its scale or the other.

Nonlinear Effects

While the end-to-end resistance of the resistor ladder inside a digital potentiometer is not likely to be affected significantly by changes in temperature, the resistance at the wiper is more heat sensitive.

In an up/down chip, there can be differential errors between incremental and decremental modes. In other words, if a tap is reached by stepping up to it incrementally, the resistance between the W pin and H or L may not be quite the same as if the same tap is reached by stepping down to it decrementally. The difference may not be significant, but may be puzzling for those who are unfamiliar with this phenomenon.

Some differences may be found among resistors in a ladder. That is, in a supposedly linear digital potentiometer, each resistor may differ in value slightly from the next.

Data Transfer Too Fast

When using a microcontroller to send data to a digital potentiometer, a small delay may be necessary between pulses, depending on the microcontroller's clock speed. A digital potentiometer may require a minimum pulse duration of 500ns. Check the manufacturer's datasheet for details.

8

timer

A device that creates a single timed pulse, or a series of timed pulses with timed intervals between them, is properly known as a *multivibrator*, although the generic term **timer** has become much more common and is used here.

Three types of multivibrator exist: *astable*, *monostable*, and *bistable*. The behavior of *astable multivibrators* and *monostable multivibrators* is described in detail in this entry. A timer chip can also be made to function as a bistable multivibrator. This is described briefly below, but it is not a designed function of a timer. The primary discussion of bistable multivibrators will be found in the entry of this encyclopedia dealing with **flip-flops**.

OTHER RELATED COMPONENTS

- **flip-flop** (See Chapter 11)

What It Does

A *monostable* timer emits a single timed pulse of fixed length in response to a triggering input that is usually of shorter duration. Many monostable timers are also capable of running in *astable* mode, in which the timer spontaneously emits an ongoing stream of timed pulses with timed gaps between them. A dual-mode timer can run in either mode, determined either by external components attached to it, or (less commonly) by changing the status of a mode selection pin.

Monostable Mode

In monostable mode, the timer emits a pulse in response to a change from high to low voltage (or, less commonly, from low to high voltage) at a *trigger pin*. Most timers respond to a *voltage level* at the trigger pin, but some are insensitive to any persistent pin state and only respond to a *voltage transition*. This is known as *edge triggering*.

The pulse generated by the timer may consist of a change from low to high (or, less commonly, from high to low) at an output pin. The length of the pulse will be determined by external components, and is independent of the duration of the triggering event, although in some cases, an output pulse may be prolonged by *retriggering* the timer prematurely. This is discussed below.

At the end of the output pulse, the timer reverts to its quiescent state, and remains inactive until it is triggered again.

A monostable timer can control the duration of an event, such as the time for which a light remains on after it has been triggered by a motion sensor. Alternatively, the timer can impose a delay, such as the interval during which a paper towel dispenser refuses to respond after a towel has been dispensed. A timer can also be useful to generate a clean pulse in response to an unstable or noisy input, as from a manually operated pushbutton.

Astable Mode

In astable mode, a timer will generally trigger itself as soon as power is connected, without any need for an external stimulus. However, the output can be suppressed by applying an appropriate voltage to a *reset pin*.

External components will determine the duration of each pulse and the gap between it and the next pulse. The pulse stream can be slow enough to control the flashing of a turn signal in a 1980s automobile, or fast enough to control the bit rate in a data stream from a computer.

Modern timer circuits are often incorporated in chips that have other purposes. The flashing of a turn signal in a modern car, for instance, is now likely to be timed by a microcontroller that manages many other functions. Still, chips that are purely designed as timers remain widely used and are very commonly available in numerous through-hole and surface-mount formats.

How It Works

The duration of a single pulse in monostable mode, or the frequency of pulses in astable mode, is most commonly determined by an external *RC network* consisting of a resistor in series with a capacitor. The charging time of the capacitor is determined by its own size and by the value of the resistor. The discharge time will be determined in the same way. A **comparator** inside the timer is often used to detect when the potential on the capacitor reaches a *reference voltage* that is established by a voltage divider inside the chip.

Variants

The 555 Timer

An eight-pin integrated circuit manufactured by Signetics under part number 555 was the world's first fully functioned timer chip, introduced in 1972. It combined two comparators with a **flip-flop** (see Chapter 11) to enable great versatility while maintaining excellent stability over a wide range of supply voltages and operating temperatures. Subsequent timers have been heavily influenced by this design. A typical 555 timer chip is shown in Figure 9-1.

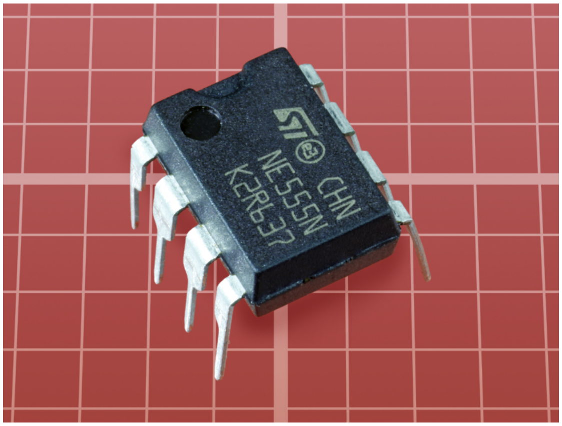

Figure 9-1. *A typical 555 timer chip. Functionally identical versions in which the "555" identifier is preceded or followed by different letter combinations are available from many different manufacturers.*

The 555 was designed by one individual, Hans Camenzind, working as an independent consultant for Signetics. According to a transcript of an interview with Camenzind maintained online at the Transistor Museum, "There was nothing like it at the time. You had to use quite a few discrete components—a comparator, a Zener diode or even two. It was not a simple circuit."

The 555 timer quickly became the most widely used chip in the world, and was still selling an annual estimated 1 billion units three decades after its introduction. It has been used in spacecraft, in intermittent windshield wiper controllers, in the early Apple II (to flash the cursor), and in children's toys. Like many chips of its era, its design was unprotected by patents, allowing it to be copied by numerous manufacturers.

The initial version was built around **bipolar transistors**, and consequently is referred to as the *bipolar version* or (more often) the *TTL version*, this being a reference to *transistor-transistor logic* protocol. Within a few years, CMOS versions based around **MOSFETs** were developed. They reduced the ability of the chip to sink or source current at its output pin, but consumed far less power, making them better suited to battery-operated products. The CMOS versions were and still are pin-compatible with the original bipolar version, both in through-hole and surface-mount formats. Their timing parameters are usually the same.

555 Monostable Operation

The internal functionality of a 555 timer wired to run in monostable mode is illustrated in Figure 9-2 with the chip seen from above. The pins are identified in datasheets by the names shown. To assist in visualizing the behavior of the chip, this figure represents the internal flip-flop as a switch which can be moved by either of two internal comparators, or by an input from the Reset pin.

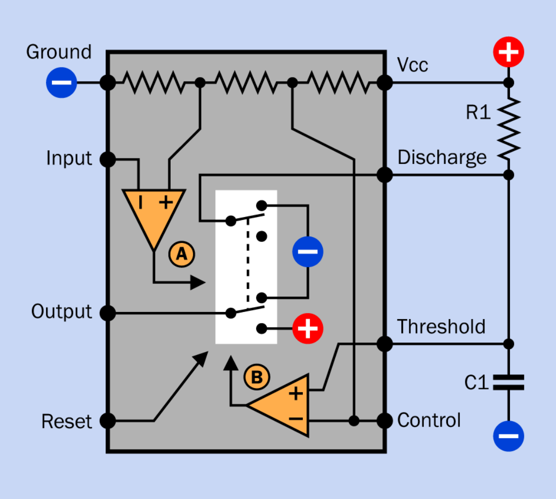

Figure 9-2. *The internal functions of a 555 chip, with its flip-flop represented as a switch that can be moved by either of two comparators, or by a low voltage on the Reset pin. An external resistor and capacitor, shown as R1 and C1, cause the timer to run in monostable (one-shot) mode, generating a single high pulse when the state of the Input pin is pulled from high to low.*

Inside the chip, three resistances of 5K each are connected between V+ (positive supply voltage) and negative ground. It has been suggested that the part number of the 555 chip was derived from these three 5K resistors, but Hans Camenzind has pointed out that Signetics was already using three-digit part numbers beginning with the number 5, and probably chose the 555 part number because the sales department had high expectations for the chip and wanted its number to be easily memorable. (A similar rationale explains the part number of the 2N2222 transistor.)

The resistances inside the timer function as a *voltage divider*, providing a reference of 1/3 of V+ to the noninverting pin of Comparator A and 2/3 of V+ to the inverting pin of Comparator B. (See Chapter 6 for an explanation of the functioning of comparators.)

When power is initially supplied to the timer, if the Input pin is at a high state, Comparator A has a low output, and the flip-flop remains in its “up” position, allowing the Output pin to remain in a low state. The flip-flop also grounds the lower end of R1, which prevents any charge from accumulating on capacitor C1.

If the state of the Input pin is pulled down externally to a voltage less than 1/3 of V+, Comparator A now creates a high output that changes the flip-flop to its “down” position, sending a high signal out through the Output pin. At the same time, C1 is no longer grounded, and begins to charge at a rate determined by its own size and by the value of R1. When the charge on the capacitor exceeds 2/3 of V+, it activates Comparator B, which forces the flip-flop into its “up” position. The Output pin goes low, C1 discharges itself into the Discharge pin, and the timer’s cycle is at an end.

The low voltage on the Input pin of the timer must end before the end of the output cycle. If the voltage on the Input pin remains low, it will re-trigger the timer, prolonging the output pulse.

A pullup resistor may be used on the Input pin to avoid false triggering, especially if an external electromechanical **switch** or **pushbutton** is used to pull down the Input pin voltage.

The Reset pin should normally be held high, either by being connected directly to positive supply voltage (if the reset function will not be needed) or by using a pullup resistor. If the Reset pin is pulled low, this will always interrupt an output pulse regardless of the timer's current status.

If a voltage higher or lower than 2/3 of V+ is applied to the Control pin, this will change the reference voltage on Comparator B, which determines when the charging cycle of C1 ends and the discharge cycle begins. A lower reference voltage will shorten each output pulse by allowing a lower charge limit for C1. If the control voltage drops to 1/3 of V+ (or less), the capacitor will not charge at all, and the pulse length will diminish to zero. If the control voltage rises to become equal to V+, the capacitor will never quite reach that level, and the pulse length will become infinite. A workable range for the control voltage is therefore 40% to 90% of V+.

Because the Control pin is an input to the chip, it should be grounded through a 0.01µF ceramic capacitor if it will not be used. This is especially important in CMOS versions of the timer.

A defect of the bipolar 555 is that it creates a *voltage spike* when its Output pin changes state. If it will be sharing a circuit with sensitive components, a 0.01µF bypass capacitor should be added as closely as possible between the V+ pin and negative ground. The voltage-spike problem was largely resolved by the CMOS 555.

555 Astable Operation

In Figure 9-3, the 555 timer chip is shown with external components and connections to run it in astable mode. The pin names remain the same but have been omitted from this diagram because of limited space. The labeling of the two external resistors and capacitor as R1, R2, and C1 is universal in datasheets and manufacturers' documentation.

When the timer is powered up initially, capacitor C1 has not yet accumulated any charge. Consequently, the state of the Threshold pin is low. But the Threshold pin is connected externally with the Input pin, for astable operation. Consequently, the Input pin is low, which forces the flip-flop into its "down" state, creating a high output. This happens almost instantaneously.

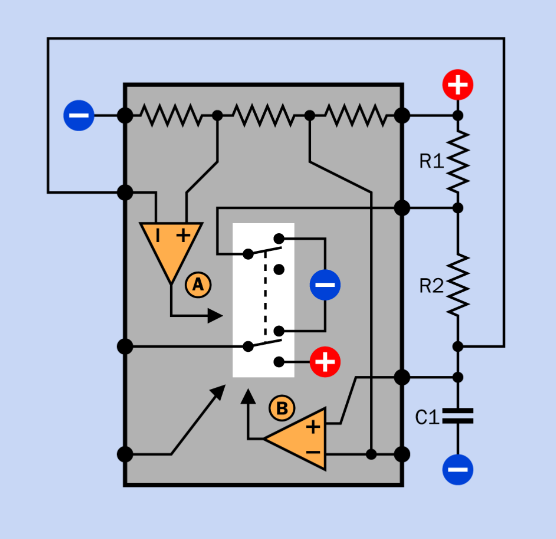

Figure 9-3. *The internal functions of a 555 chip, with two external resistors and a capacitor wired to run the timer in astable (free-running) mode.*

While the flip-flop is "down," the Discharge pin is not grounded, and current flowing through R1 and R2 begins to charge the capacitor. When the charge exceeds 2/3 of positive supply voltage, Comparator B forces the flip-flop into its "up" position. This ends the high pulse on the Output pin, and starts to drain the charge from the capacitor through R2, into the Discharge pin. However, the voltage on the capacitor is still being shared by the Input pin, and when it diminishes to 1/3 of V+, the Input pin reactivates Comparator A, starting the cycle over again.

The functions of the Reset and Control pins are the same as in monostable mode. Because voltage applied to the Control pin changes the length of each pulse and the gaps between pulses, it has the effect of adjusting the frequency of the output in astable mode.

When power is first connected to the timer, C1 must initially charge from an assumed state of zero potential to 2/3 V+. Because subsequent cycles will begin when the capacitor is at 1/3 V+, the first high output pulse from the timer will be slightly longer than subsequent output pulses. This is unimportant in most applications, especially because the rate at which a capacitor accumulates charge is greater when beginning from 0V than when it has reached 1/3 V+. Still, the longer initial pulse can be noticeable when the timer is running slowly.

Because the capacitor charges through R1 and R2 in series, but discharges only through R2, the length of each positive output pulse in astable mode is always greater than the gap between pulses. Two strategies have been used to overcome this limitation. See "Separate Control of High and Low Output Times" on page 80.

556 Timer

The 556 consists of two 555 bipolar-type timers in one package. An example of the chip is shown in Figure 9-4. The pinouts are shown in Figure 9-5. Although 556 timers have become relatively uncommon compared with the 555, they are still being manufactured in through-hole and surface-mount versions by companies such as Texas Instruments and STMicroelectronics, under part numbers such as NA556, NE556, SA556, and SE556 (with various letters or letter pairs appended). Each timer in the chip has its own set of inputs and outputs, but the timers share the same V+ and ground voltages.

558 Timer

This 16-pin chip is now uncommon, and many versions have become obsolete. It has been identified by a part number such as NE558 although different prefix letters may be used. The NTE926, shown in Figure 9-6, is actually a 558 timer.

The chip contains four 555 timers sharing a common power supply, common ground, and common control-pin input. For each internal timer, the Threshold and Discharge functions are connected internally, so that the timers can only be used in one-shot mode. However, one timer can trigger another at the end of its cycle, and the second timer can then retrigger the first, to create an astable effect.

Each timer is edge-triggered by a voltage transition (from high to low), instead of being sensitive to a voltage level, as is the case with a 555 timer. Consequently the timers in the 558 chip are insensitive to a constant (DC) voltage.

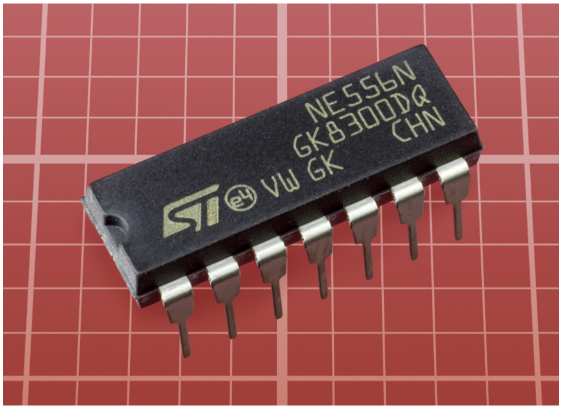

Figure 9-4. *An example of the 556 timer chip.*

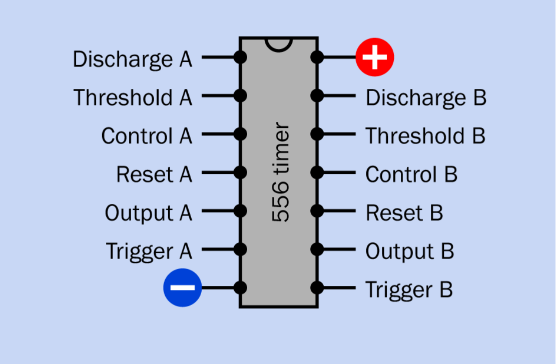

Figure 9-5. *The 556 timer contains two separate 555 timers sharing the same power supply and ground. The pin functions for timer A and timer B are shown here.*

The output from each timer is an open collector, and therefore requires an external pullup resistor. Each output is capable of sinking up to 100mA.

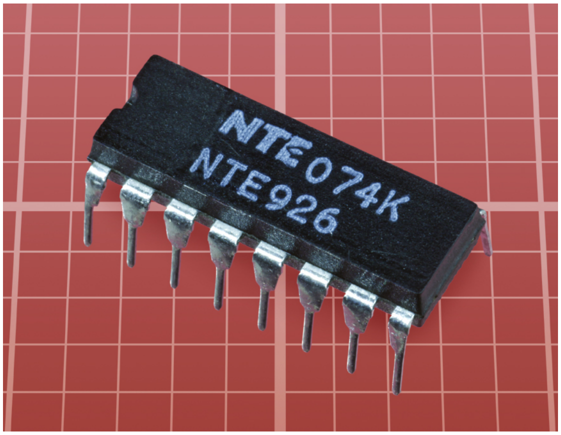

Figure 9-6. *The NTE926 is a 558 timer chip.*

CMOS 555 Timer

While the part numbers of many CMOS versions are significantly different from part numbers of the bipolar versions, in some instances the CMOS numbers are only distinguished by a couple of initial letters. The ST Microelectronics TS555 series and Texas Instruments TLC555 series, for example, use MOSFETs internally. The ST Microelectronics SE555 series and Texas Instruments SA555 series use bipolar transistors internally.

One way to distinguish between the two types, when searching a website maintained by a parts supplier, is to begin by looking generically for a "555 timer" and then add a search filter to show chips either with a minimum power supply of 3VDC (which will be CMOS) or with a minimum power suply of 4.5VDC (which will be bipolar).

CMOS versions of the 555 timer do not create the power spike that is characteristic of the bipolar versions during output transitions. The CMOS chips can also be powered by a lower voltage (3VDC, or 2VDC in some cases), and will draw significantly less current in their quiescent state. They also require very little current for threshold, trigger, and reset functions.

The wiring of external resistors and capacitors to the CMOS version of the chip, and the internal voltage levels as a fraction of V+, are identical to the original 555 timer. Pin functions are likewise identical. The only disadvantages of CMOS versions are their greater vulnerability to static discharge, and their lower output currents. The TLC555, for instance, will source only 15mA (although it can sink 10 times that amount). Other manufacturers have different specifications, and datasheets should be checked carefully.

5555 Timer

The 5555 contains a digital **counter** that enables it to time very long periods. Its full part number is 74HC5555 or 74HCT5555, although these numbers may be preceded or followed by letter combinations identifying the manufacturer. It is not pin-compatible with a 555 timer.

Two input pins are provided, one to trigger the timer on a rising edge, the other on a falling edge, of the input pulse. The inputs are Schmitt-triggered.

The timer is rated for 1Hz to 1MHz (using an external resistor and external capacitor). The counter section can divide the pulse frequency by values ranging from 2 to 256. For longer timed periods, different settings on the digital control pins will divide the frequency by values ranging from 2^{17} through 2^{24} (131,072 through 16,777,216). This enables the timer to achieve a theoretical pulse length lasting for more than 190 days. The timer will accept a clock input from an external oscillator to achieve better accuracy than is available with a resistor-capacitor timing circuit.

7555 Timer

This 8-pin chip is a CMOS version of the 555 timer, manufactured by companies such as Maxim Integrated Products and Advanced Linear Devices. Its characteristics are similar to those of CMOS 555 timers listed above, and the pinouts are the same.

7556 Timer

This 14-pin chip contains two 7555 timers, sharing common power supply and ground connections. Pinouts are the same as for the original 556 timer, as shown in Figure 9-5.

4047B Timer

This 14-pin CMOS chip was introduced in an effort to address some of the quirks of the 555 timer while also providing additional features. It runs in either monostable or astable mode, selectable by holding one input pin high or another input pin low. In astable mode, its duty cycle is fixed at approximately 50%, a single resistor being used for both charging and discharging the timing capacitor. An additional "oscillator" output runs twice as fast as the regular output.

In monostable mode, the 4047B can be triggered by a positive or negative transition (depending on which of two input pins is used). It ignores steady input states and will also ignore additional trigger pulses that occur during the output pulse. However, a retrigger pin is provided to extend the output pulse if desired.

Complementary output pins are provided, one being active-high while the other is active-low.

To time very long periods, the 4047B was designed to facilitate connection with an external counter.

The power supply for the 4047B can be as low as 3VDC. Its maximum source or sink output current is only 1mA when powered at 5VDC, but up to 6.8mA when powered at 15VDC.

The chip is still available from manufacturers such as Texas Instruments (which markets it as the CD4047B) in through-hole and surface-mount formats. However, despite its versatility, the 4047B is less popular than dual monostable timers, described in the next section.

Dual Monostable Timers

Various timers that run only in monostable mode are available in dual format (i.e., two timers in one chip). This format became popular partly because two monostable timers can trigger each other to create an astable output, in which the pulse width, and the gap between pulses, can be set by a separate resistor-capacitor pair on each timer. This allows greater flexibility than is available when using a 555 timer.

Most dual monostable timer chips are edge-triggered by a change in input voltage, and will ignore a steady DC voltage. Consequently, the output from one timer can be connected directly to the input of another, and no coupling capacitors are necessary.

As in the 4047B, the user has a choice of two input pins for each timer, one triggered by a transition from low to high, the other triggered by a transition from high to low. Similarly, each timer has two outputs, one shifting from low to high at the start of the output cycle, the other shifting from high to low.

The values of a single resistor and capacitor determine the pulse duration of each timer.

Dual monostable timers often have the numeric sequence 4528 or 4538 in their part numbers. Examples include the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, and the MC14538B from On Semiconductor. The 74123 numeric sequence identifies chips that have a very similar specification, with chip-family identifiers such as HC or LS inserted, as in the 74HC123 and 74LS123, and additional letters added as prefix or suffix. The pinouts of almost all these chips are identical, as shown in Figure 9-7. However, Texas Instruments uses its own numbering system, and datasheets should always be consulted for verification before any connections are made.

Many chips of this type are described as "retriggerable," meaning that if an additional trigger pulse is applied to the input before an output pulse has ended, the current output pulse will be extended in duration. Check datasheets carefully to determine whether a chip is "retriggerable" or will ignore new inputs during the output pulse.

The 74HC221 dual monostable vibrator (pictured in Figure 9-8) functions very similarly to the components cited above, but has slightly different pinouts.

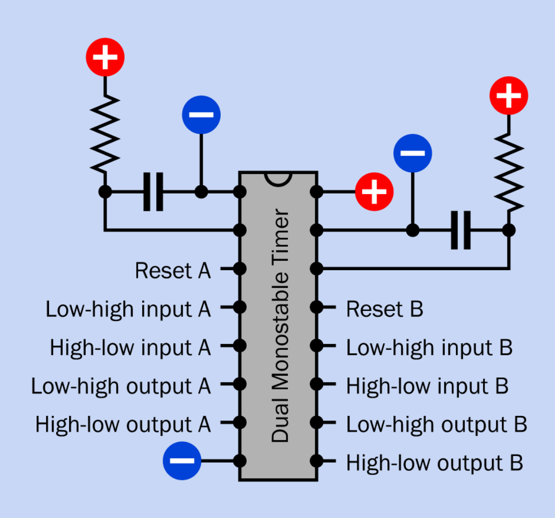

Figure 9-7. *Pin functions for most 4528, 4538, and 74123 series of dual monostable timer chips. An RC network is shown connected for each timer. Note that Texas Instruments uses different pinouts on its versions.*

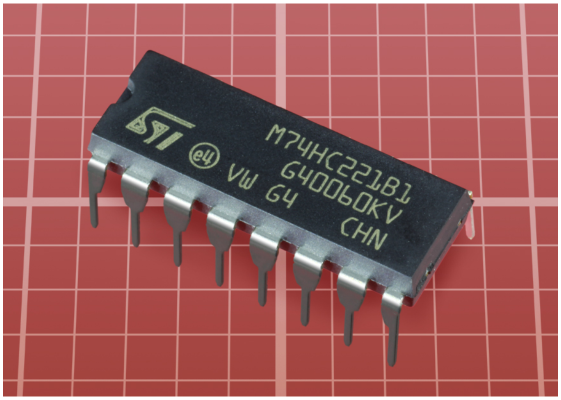

Figure 9-8. *A dual-timer chip containing two monostable multivibrators that can function in astable mode if they are connected externally to trigger each other.*

Values

555 Timer Values

The original bipolar version of the 555 timer was designed to operate using a wide range of positive supply voltages, from 4.5VDC to 16VDC. CMOS versions vary in their recommended V+ values, and datasheets must be consulted for verification.

The output of a bipolar 555 is rated to source or sink up to 200mA. In practice, the maximum current will be lower when the timer is powered at the low end of its range, around 5VDC. Attempting to source more than 50mA will pull down the voltage internally, affecting operation of the timer.

CMOS versions all impose restrictions on output current, allowing higher values for sinking than sourcing. Again, datasheets must be consulted for the values, which vary widely from one component to another.

The voltage measured on the output pin, when it is used for sourcing current, will always be lower than the power supply voltage, and a 1.7V drop is commonly specified for bipolar versions. In practice, the voltage drop that is actually measured may be less, and will vary according to the load on the output.

The voltage drop does not increase significantly with a higher supply voltage, and because it is a relatively constant value, it becomes less significant when a higher value for V+ is used.

CMOS versions of the 555 timer achieve a claimed output source voltage that is only 0.2V less than the power supply.

When choosing values for R1 and R2, a minimum for each resistor is 5K, although 10K is preferred. Lower values will increase power consumption, and may also allow overload of the internal electronics when the chip sinks current from C1. A typical maximum value for each resistor is 10M.

A high-value capacitor may cause the timer to function less accurately and predictably, because large capacitors generally allow more *leakage*. This means that the capacitor will be losing charge at the same time that it is being charged through R1 + R2. If these resistors have high values, and the capacitor has a value of 100µF or more, the rate of charge may be so low that it is comparable with the rate of leakage. For this reason, a 555 timer is not a good choice for timing intervals much greater than a minute. If a large-value capacitor is used, tantalum is preferable to electrolytic.

The minimum practical value for a timing capacitor is around 100pF. Below this, performance may not be reliable.

Although some CMOS versions may enable fast switching, the shortest practical output pulse for a 555 timer is around 10 microseconds. On the input pin, a triggering pulse of at least 1 microsecond should be used.

Time Calculation in Monostable Mode

If R1 is measured in kilohms and C1 is measured in microfarads, the pulse duration, T, in seconds, of a 555 timer running in monostable mode can be found from this simple formula:

```
T = 0.0011 * R1 * C1
```

This relationship is the same in all versions of the 555. Figure 9-9 provides a quick and convenient way to find the pulse value using some common values for R1 and C1. Resistors can be obtained with tolerances below plus-or-minus 1%, but capacitors are often rated with an accuracy of only plus-or-minus 20%. This will limit the accuracy of the pulse values shown in the chart.

Time Calculation in Astable Mode

If R1 and R2 are measured in kilohms and C1 is measured in microfarads, the frequency of pulses, F (measured in Hz) of a 555 timer running in astable mode can be found from this simple formula:

```
F = 1440 / ( ( R1 + (2 * R2) ) * C1)
```

This relationship is found in all versions of the 555. Figure 9-10 shows the frequency for common values of R2 and C1, assuming that the value of R1 is 10K. In Figure 9-11, a value of 100K is assumed for R1.

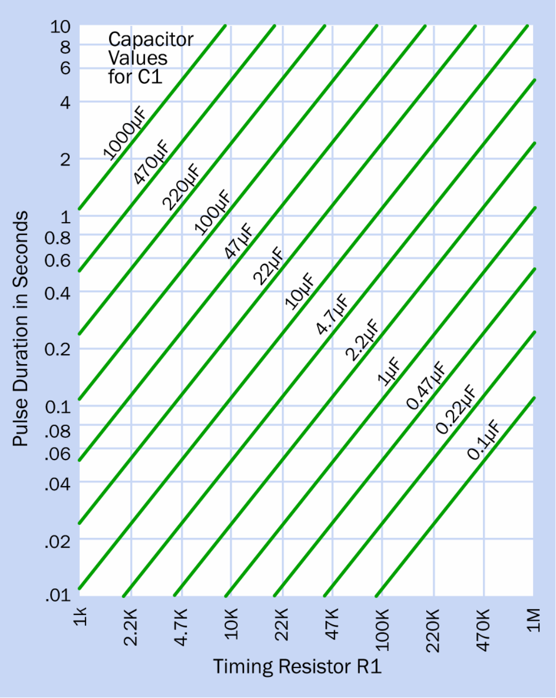

Figure 9-9. *To determine the pulse duration of a 555 timer running in monostable mode: find the value of R1 on the horizontal scale, follow its vertical grid line upward to the intersection with a green line which corresponds with the value of capacitor C1, and read across to the vertical scale providing the duration in seconds. Both axes are logarithmic.*

Dual Monostable Timers

Dual chips such as the HEF4528B from NXP, the M74HC4538 from STMicroelectronics, the MC14538B from On Semiconductor, and the 74HC123 from Texas Instruments have widely varying requirements for power supply. Some accept a limited range from 3VDC to 6VDC, while others tolerate a range of 3VDC to 20VDC. When powered with 5VDC their required input and output states are compatible with those of 5V logic chips.

Output pins of these chips source and sink no more than 25mA (much less in some instances). Because there are so many variants, they cannot be summarized here, and datasheets must be consulted for details.

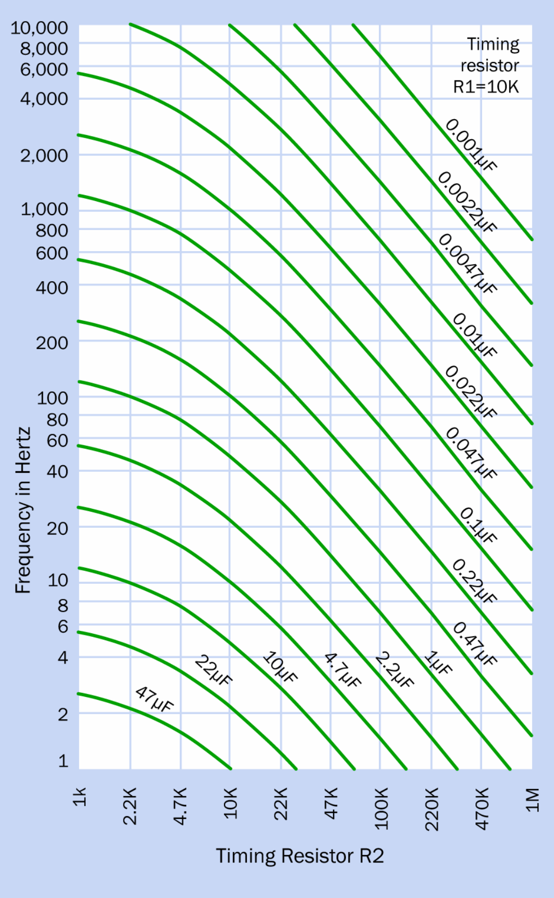

Figure 9-10. *To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 10K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.*

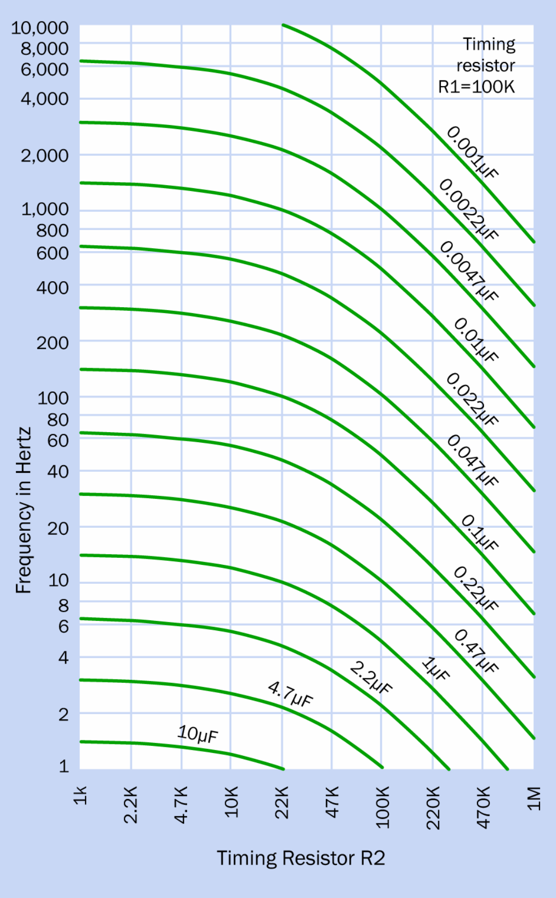

Figure 9-11. *To obtain the frequency of a 555 timer running in astable mode, when R1 has a value of 100K: find the value of R2 on the horizontal scale, follow its vertical grid line upward to the intersection with a green curve which corresponds with the value of capacitor C1, and read across to the vertical scale providing the frequency in Hertz. Both axes are logarithmic.*

As these timers are all monostable, and each timer uses just one resistor and one capacitor, the only formula required is to give the pulse time as a function of these two variables. If R is the resistor value in ohms, and F is the capacitor value in farads, and K is a constant supplied by the manufacturer, the pulse time T, in seconds, is found from the formula:

```
T = R * F * K
```

K ranges between 0.3 and 0.7 depending on the manufacturer and also on the voltage being used. Its value should be found in the manufacturer's datasheet. If R is measured in megohms and F is measured in microfarads, the formula is still valid, as the multipliers cancel each other out.

Generally speaking, these dual monostable CMOS timers are not intended for pulse duration exceeding 1 minute.

The timing capacitor should be no larger than 10µF, as it discharges directly and rapidly through the chip.

How to Use It

Where a timer is required to drive a load such as a relay coil or small motor directly, the original TTL version of the 555 timer will be the only choice. Even in this instance, a protection diode must be used across the inductive device.

For smaller loads and applications in chip-to-chip circuits, CMOS versions of the 555, including the 7555, use less power, cause less electrical interference, and are pin-compatible while using the same formulae to calculate frequency in astable mode or pulse duration in monostable mode. They are of course more vulnerable to static discharge, and care must be taken to make a connection to every pin (the capacitor that grounds the Control pin, if Control is not going to be used, is mandatory).

In dual monostable timers, unused rising-edge trigger inputs should be tied to V+ while unused falling-edge trigger inputs should be tied to ground. A Reset pin that will not be used should be tied to V+, unless that entire timer section of the chip will be unused, in which case the pin should be grounded.

To measure durations longer than a few minutes, a timer which incorporates a programmable counter to divide the clock frequency is the sensible choice. See the description of the 5555 timer that was included earlier in this entry.

The original bipolar version of the 555 remains a robust choice in hobby applications such as robotics, and its design allows some versatile variations which may even be used in logic circuits. A variety of configurations are shown in the schematics below.

555 Monostable Mode

The basic schematic for a 555 timer running in monostable mode is shown in Figure 9-12. In this particular example, a pushbutton that is liable to suffer from *switch bounce* is connected to the Input pin of the timer, which responds to the very first connection made by the pushbutton and ignores the subsequent "bounces," thus producing a "clean" output. To avoid retriggering, which results in a prolonged output pulse, the timer's output should exceed the time for which the button is likely to be pressed. The output should also exceed the duration of any possible switch bounce, which can otherwise create multiple output pulses. In the schematic, an LED is attached to the timer output for demonstration purposes.

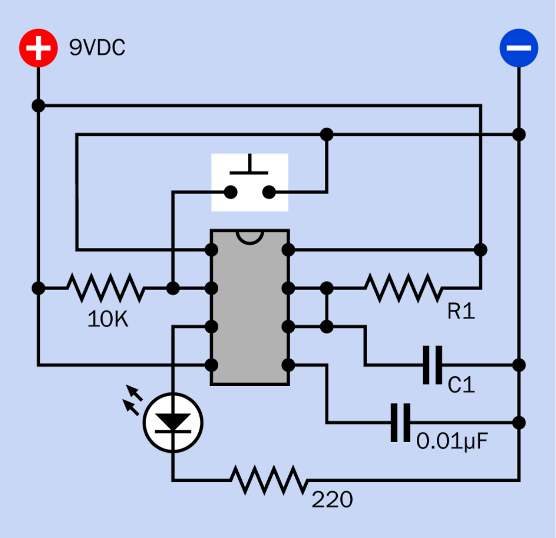

Figure 9-12. *The basic monostable configuration of a 555 timer. This particular circuit debounces an input from a pushbutton switch and converts it to a clean pulse of fixed duration, powering an LED for demonstration purposes.*

This circuit is shown on a breadboard in Figure 9-13. The red and blue wires, at the top of the photograph, supply 9VDC to the board. R1 is 1M, while C1 is 1µF, creating a pulse of just over 1 second. A tactile switch, just above the timer, provides the input.

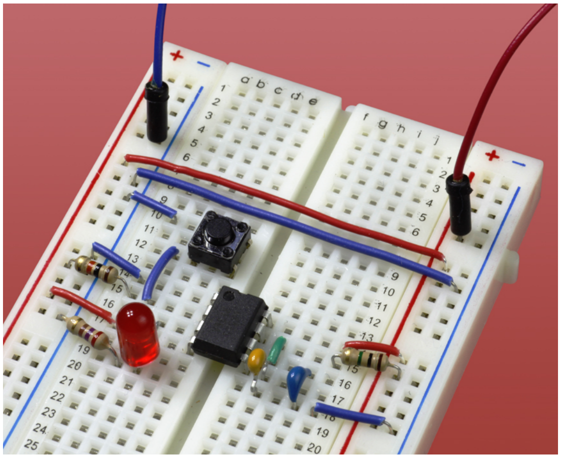

Figure 9-13. *The basic monostable configuration for a 555 timer, mounted on a breadboard.*

555 Astable Mode

A basic schematic for a 555 timer running in astable mode is shown in Figure 9-14. Once again, an LED is attached to the output for demonstration purposes. If the pulse rate exceeds the persistence of vision, a small loudspeaker can be used instead, in series with a 47Ω resistor and a 100µF capacitor.

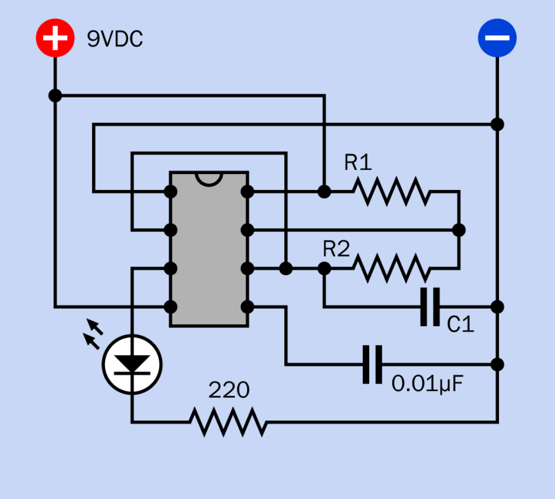

Figure 9-14. *A 555 timer with external connections and components causing it to run in astable (free-running) mode.*

Separate Control of High and Low Output Times

In Figure 9-15, a bypass diode has been added around R2. The capacitor now charges primarily through R1, as the diode has a much lower effective resistance than R2. It discharges only through R2, as the diode blocks current in that direction. Consequently, the length of the high output pulse can be adjusted with the value of R1 only, while the length of the low output pulse can be adjusted with the value of R2 only. The duration of the high output can be lower than, or equal to, the duration of the low output, which is not possible with the basic configuration of components in Figure 9-14.

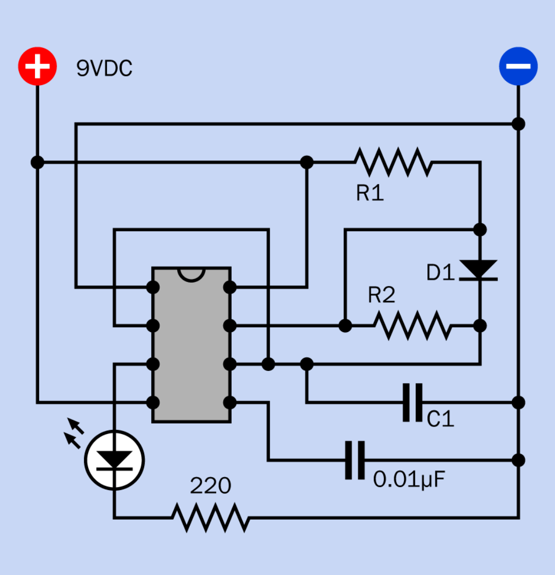

Figure 9-15. *In this circuit, a diode bypasses R2, so that the "on" time and the "off" time of the 555 timer can be set independently of each other, with R1 and R2, respectively.*

555 Fifty Percent Astable Duty Cycle: 1

In Figure 9-16, the circuit enables a fixed astable output duration of approximately 50% high and 50% low. Initially, C1 has no charge, pulling the Input of the timer low, and causing it to begin a cycle with a high pulse from the Output pin, as usual. In this demonstration circuit, the output illuminates an LED. At the same time, resistor R1 is attached to the output and charges C1. When the voltage on C1 reaches 2/3 of V+, this is communicated to the timer Input pin, which ends the "high" cycle and initiates low status on the Output pin. This starts to sink the charge from C1, through R1. When the voltage drops to 1/3 V+, this initiates a new cycle. Because only one resistor is used to charge and discharge the capacitor, we may imagine that the charge and discharge times should be identical. However, a higher load on the output will probably pull down the output voltage to some extent, lengthening the charge time. Conversely, a load on the Output pin that has low resistance will probably sink at least some of the charge from the capacitor, shortening the discharge cycle.

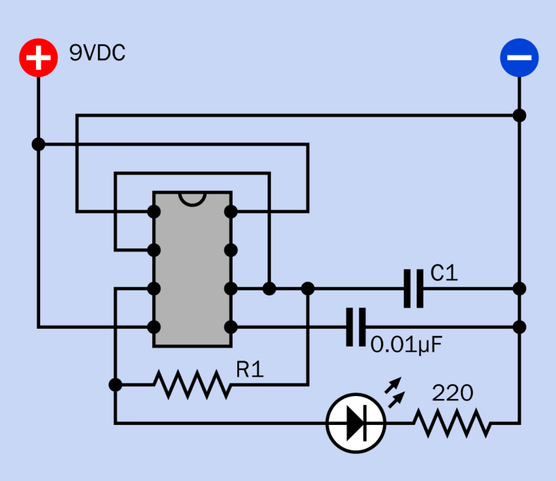

Figure 9-16. *This configuration provides an approximate 50-50 on-off duty cycle at the output pin, although the precise duration will depend on the load.*

555 Fifty Percent Astable Duty Cycle: 2

In Figure 9-17, a small modification of the basic astable circuit shown in Figure 9-14 provides another way to enable a 50% duty cycle. Compare the two schematics, and you will see that just the connection between R1 and R2 has been altered so that C1 now charges only through R1, and discharges only through R2. However, in this configuration the capacitor is discharging into a voltage divider created by the two resistors. Empirical adjustment of the resistor values may be necessary before the duty cycle is precisely 50%.

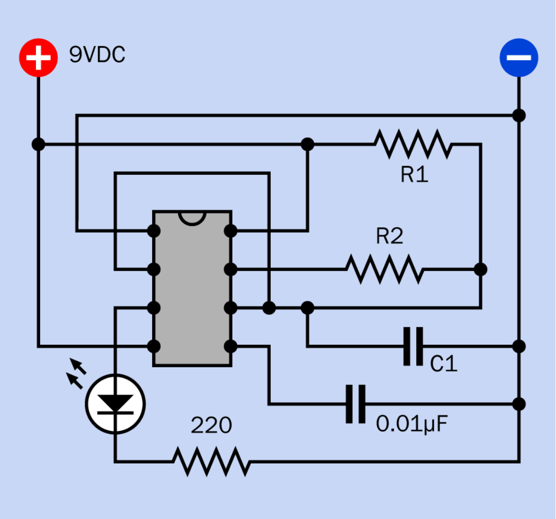

Figure 9-17. *An alternative configuration to provide an approximate 50-50 on-off duty cycle in a 555 timer.*

Use of the 555 Control Pin

In Figure 9-18, a potentiometer and two series resistors allow a varying voltage to be applied to the Control Pin. This will lengthen or reduce both the charge and the discharge times of the timing capacitor. If values for the capacitor and its associated resistors are chosen to create a frequency of approximately 700Hz, a 10K potentiometer should demonstrate more than an octave of audible tones through the loudspeaker. Other components can be substituted for a potentiometer, creating possibilities for producing pulse-width modulation. Alternatively, if a large capacitor is added between the Control pin and ground while a second 555 timer, running slowly in astable mode, applies its output to the Control pin, the charging and discharging of the capacitor will apply a smoothly rising and falling voltage. If the first 555 timer is running at an audio frequency, the output will have a "wailing siren" effect.

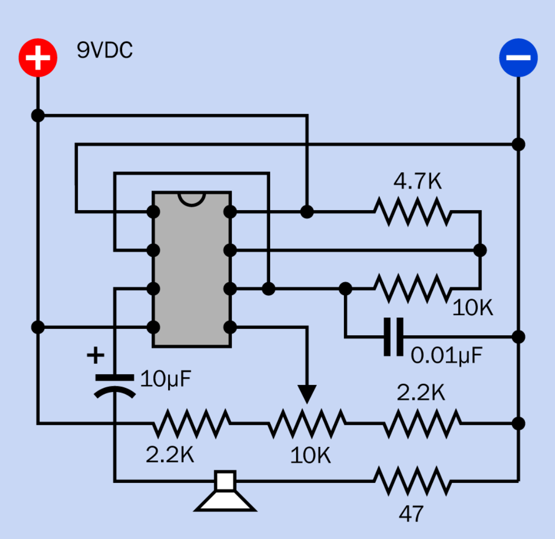

Figure 9-18. *A circuit that allows adjustment of the astable 555 frequency by increasing or lowering the voltage on the Control pin.*

Figure 9-19 shows the components specified in Figure 9-18 mounted on a breadboard.

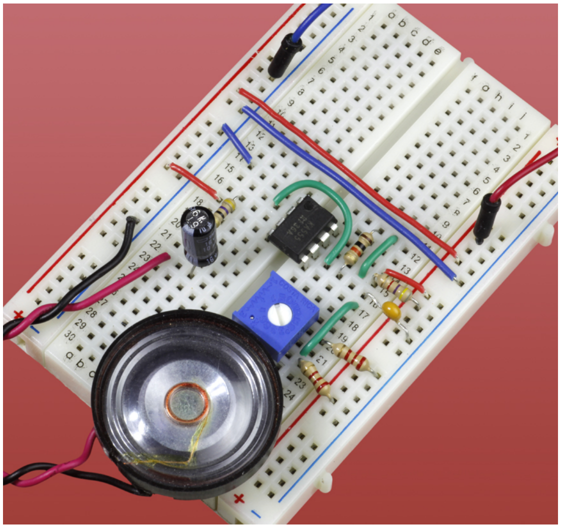

Figure 9-19. *The components in the previous schematic are shown here mounted on a breadboard. They will generate an audio output ranging between approximately 425Hz and 1,075Hz. A lower value for the timing capacitor will shift the audio range higher in frequency.*

555 Flip-Flop Emulation

The flip-flop inside a 555 timer can be accessed to control the timer's outputs. In Figure 9-20, pushbutton switch S1 applies a negative pulse to the Input pin, creating a high output from the timer, which illuminates LED D1. Normally the pulse length would be limited by the charge time of a capacitor attached to the Threshold pin, but in this circuit there is no capacitor, and the Threshold pin is hard-wired to negative ground. So, it never rises to 2/3 of positive power, and the output from the timer remains high indefinitely.

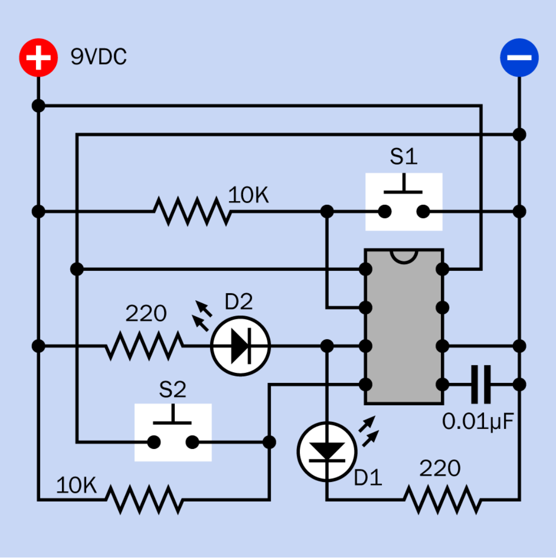

Figure 9-20. *A 555 timer can have its timing features disabled so that it functions as a flip-flop.*

However, if S2 is pressed, it grounds the Reset pin of the timer, which ends the high output and pulls the Output pin down to a low state. D1 goes out and D2 lights up, as the timer is now sinking current through it. When S2 is released, the timer output remains low and D2 remains illuminated, because the Input pin is held high by a pullup resistor. Therefore, the timer now functions in bistable mode, as a flip-flop. While this may be seen as an inappropriate use of the chip, because its full functionality is being disabled, its ability to deliver substantial current and to tolerate a wide range of supply voltages may make it more convenient to use than a digital flip-flop. See Chapter 11 for more information about flip-flops.

A 555 timer emulating a flip-flop is shown on a breadboard in Figure 9-21.

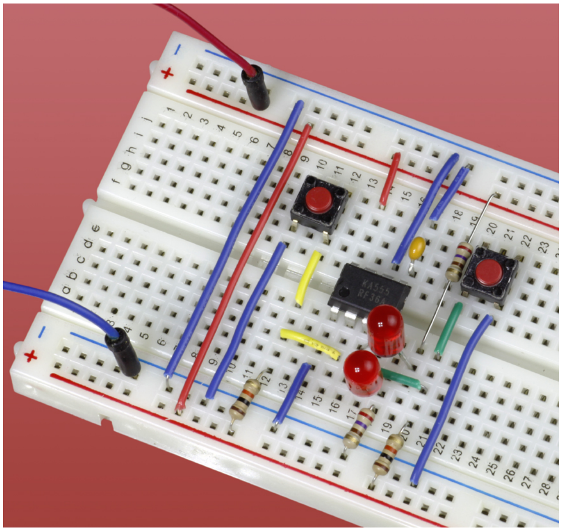

Figure 9-21. *The schematic in which a 555 timer acts as a flip-flop is shown here adapted for a breadboard.*

555 Hysteresis

The comparators inside a 555 timer enable the chip to produce hysteresis. In Figure 9-22, the Input pin and the Threshold pin are shorted together, and C1, the timing capacitor, is omitted. A 10K potentiometer, wired as a voltage divider, delivers a voltage to the Input pin ranging smoothly from V+ to negative ground. As the input dips below 1/3 V+, the Output pin goes high, lighting LED D1. Now if the input voltage gradually rises, the output remains high, even as the input rises above 1/3 V+. The output state is "sticky" because the timer does not end an output pulse until the Threshold pin tells it to, by reaching 2/3 V+. When this finally occurs, the Output pin goes low, D1 goes out, and D2 comes on, sinking current into the Output pin.

Suppose, now, the input voltage starts to go down again. Once again the output state is "sticky" because it remains low until the Input pin drops below the 1/3 level. When that happens, the output finally flips back to a high state, D2 goes out, and D1 comes on.

In the "dead zone" between 1/3 and 2/3 of supply voltage, the timer remains in its current mode, waiting for the input to stray outside of those limits. This behavior is known as *hysteresis*, and is of special importance when processing a varying signal, such as the voltage from a temperature sensor, to control an on/off device such as a thermostat. In fact the 10K potentiometer in this demo could be replaced with a thermistor or a phototransistor, wired in series with a resistor to create a voltage divider which will have an input range compatible with the 555 timer. The hysteresis can then be adjusted by varying the supply voltage that powers the timer, as this will change the values of 1/3 V+ and 2/3 V+. Alternatively, varying the voltage on the Control pin will affect the hysteresis.

A **comparator** can provide much more versatile control of hysteresis by using positive feedback (see Chapter 6 for additional details). But the 555 timer provides a quick-and-simple substitute, and its greater ability to source or sink current enables it to be connected with a wider range of other components.

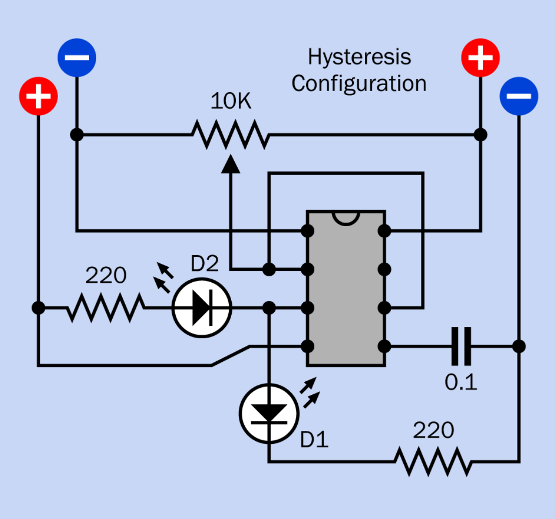

Figure 9-22. *A 555 timer wired so that it creates hysteresis, when supplied by a variable input voltage.*

555 and Coupling Capacitors

As previously noted, when a basic bipolar 555 timer (and some of its variants) is wired in monostable mode, it will retrigger itself indefinitely if its input remains low. One way to avoid this is by using a coupling capacitor. This will pass a transition from high to low, but will then block a steady subsequent voltage. In Figure 9-23, a phototransistor in series with a resistor provides a variable voltage to the noninverting input of a comparator. The reference voltage of the comparator is adjusted with a potentiometer, and resistor R3 provides positive feedback, ensuring that the output from the comparator will be quick and clean. The output from the 555 timer goes through a transistor to the relay, shown at the bottom.

It is important to see the function of the coupling capacitor, C3, with the pullup resistor, R2, which holds the Input pin of the 555 timer high by default. When the output from the comparator drops from high to low, C3 passes this transition to the Input pin of the timer, momentarily overcoming the positive potential, and triggering the timer. After the timer responds, however, C3 blocks any continuation of low voltage from the comparator. Pullup resistor R2 resumes its function of holding the input high, and prevents the timer from being retriggered.

555 Loudspeaker Connection

A small 8-ohm loudspeaker can be driven from the output of a bipolar 555 timer running in astable mode, but should be isolated from it with a 10µF to 100µF capacitor. A series resistor of 47Ω (minimum) should be used. See Figure 9-24.

Burst Mode

It is sometimes useful to create a short beep of fixed length in response to a button-press. The beep should terminate even if the button is held down. This "burst mode" can be achieved with the circuit in Figure 9-25, where the button connects power to a bipolar-type 555 timer running in astable mode, and an RC network applies a decreasing potential to a 47µF capacitor, which is wired to the Reset pin of the timer. The resistor in series with the capacitor will vary the length of the beep. When voltage to the pin drops below approximately 0.3V, output from the timer stops and cannot restart until the button is released.

A resistor of greater than 1.5K may not allow the input value at the reset pin to fall below the voltage, which is necessary to enable a reset. If a lower power supply voltage than 9VDC is used, the resistor value should be higher—for example, a 5VDC power supply works well with a 1.5K to 2K resistor.

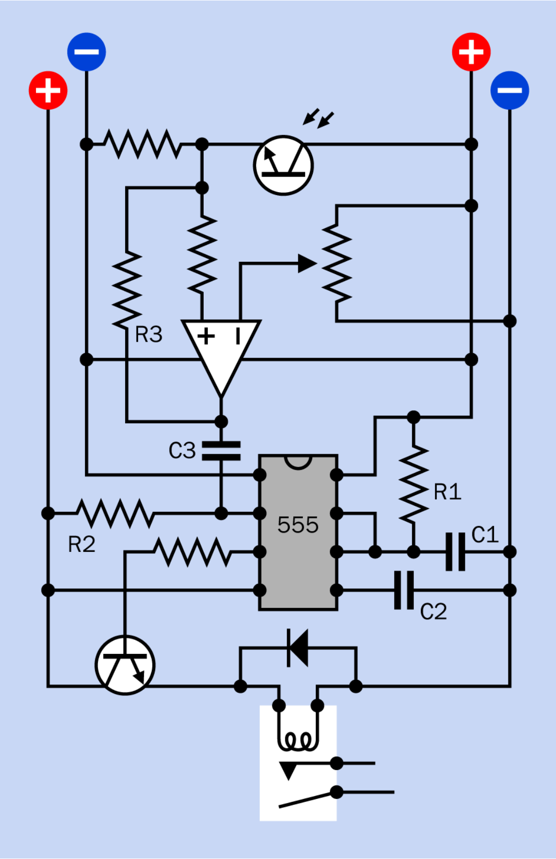

Figure 9-23. *A coupling capacitor (C3) is used in this circuit to isolate the 555 timer from a sustained low input from the comparator. The capacitor only passes a transition from high to low. The rest of the time, the pullup resistor (R3) holds the input high.*

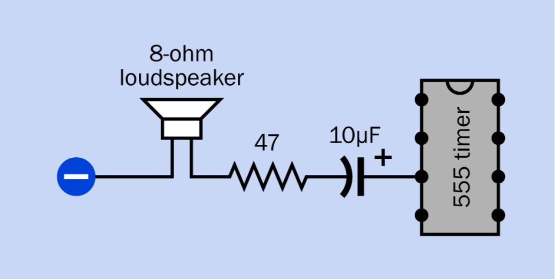

Figure 9-24. *A small 8-ohm loudspeaker can be attached through a capacitor and a resistor to the output of a bipolar 555 timer.*

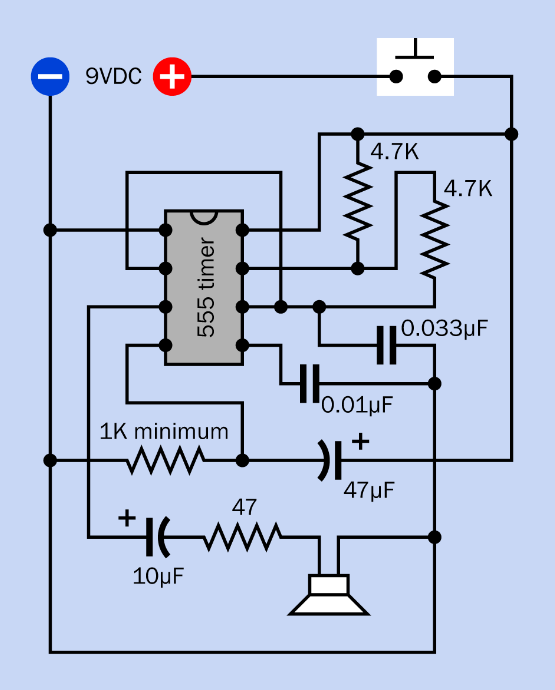

Figure 9-25. *An RC circuit, wired to apply a decreasing voltage on the Reset pin of a bipolar 555 timer, will shut off the timer shortly after it is powered up. This can be used to create a fixed-length beep in response to a button press of any duration.*

Figure 9-26 shows the components installed on a breadboard.

"You Lose" Game Sound

A timer is a simple, cheap way to create a variety of simple game sounds. The schematic in Figure 9-27 makes a groaning sound as the 100µF capacitor wired to the Control pin of a bipolar-type 555 gradually charges through the 1K resistor. Note that if a larger resistor is paired with a smaller capacitor, the effect will differ. The 150K resistor is included to discharge the capacitor reasonably quickly in time for the next cycle.

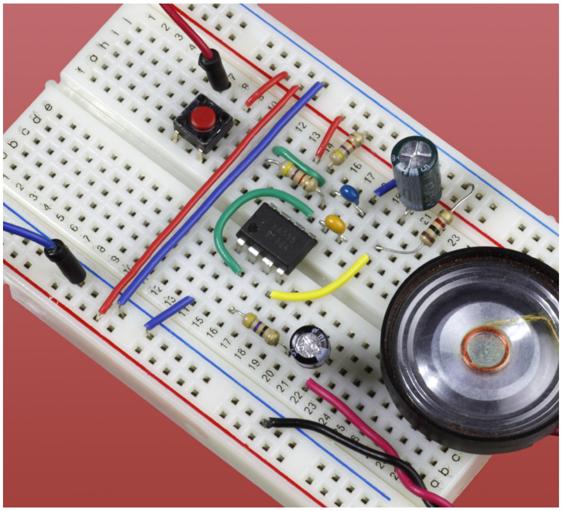

Figure 9-26. *The "burst mode" circuit installed on a breadboard with a miniature loudspeaker.*

What Can Go Wrong

Dead Timer

Like any chip, the 555 can be damaged by overvoltage, excessive source current or sinking current, static electricity, incorrectly applied polarity of power supply, and other forms of abuse. The TTL version of the timer is fairly robust, but the CMOS type much less so. Check for obvious errors such as lack of supply voltage, incorrect or ambiguous input voltages, and unusual current draw (too high, or none at all, at the V+ pin). Use the meter probes on the actual pins of the chip, in case there is a break in the wiring that feeds them. Because timer chips are cheap, a reserve supply of them should be maintained.

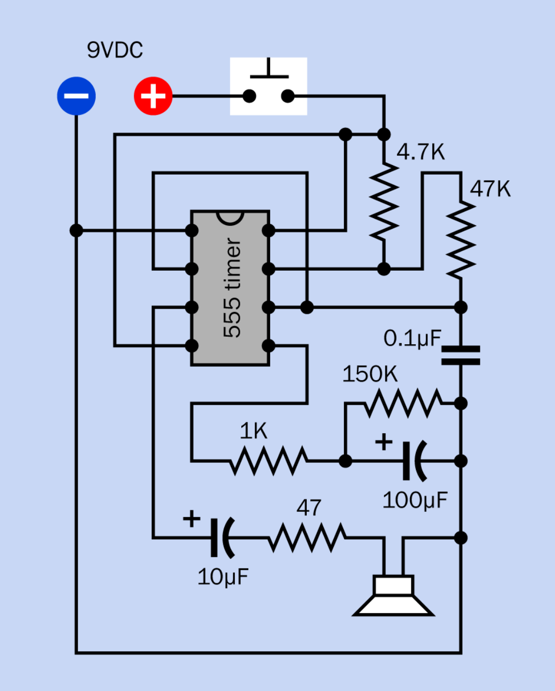

Figure 9-27. *An RC circuit, wired to apply an increasingly positive voltage to the Control pin of a bipolar timer running in astable mode, will gradually pull down the frequency at the Output pin, creating a sound that may be useful in simple game applications.*

CMOS Confused with Bipolar

The part numbers of some bipolar chips are very similar to those of some CMOS versions, and the chips look physically identical. But the CMOS version is easily overloaded, as it may source only 10mA to 20mA maximum while the TTL version is capable of 200mA. Make sure that your chips are carefully labeled when they are stored.

The Pulse that Never Ends

If a 555 timer responds correctly to a high-to-low transition on the input pin, but the output pulse continues indefinitely, check the voltage on pin 6 to see if the timing capacitor is charging above 2/3 of V+. While a 555 can run from 5VDC, a high-current device on the Output pin can pull down the voltage inside the chip to the point where the capacitor never charges sufficiently to end the cycle.

Also check that the input transition from high to low lasts for a shorter time than the pulse. A persistent low input can retrigger the timer.

Erratic Chip Behavior

Possible causes include:

- Floating pins. The Input pin, in particular, should always be connected with a defined voltage (via a 10K pullup resistor, if necessary), and must not be allowed to float at an indeterminate potential.
- Voltage spikes. A timer can be triggered by transients from other components, especially inductive loads. If the input to a monostable timer dips for even a fraction of a second, the timer will initiate a new cycle. A *protection diode* should be used in conjunction with an inductive load.
- Voltage spikes can also introduce variations in the pulse train from an astable timer.
- TTL versions of the 555 timer will tolerate a wide range of supply voltage, but if a voltage regulator is not used, fluctuations in voltage can have unpredictable consequences.

Interference with Other Components

Because the bipolar version of a 555 timer creates a voltage spike when its output changes state, it can interfere with the normal function of other components, especially CMOS chips. A 0.1µF bypass capacitor can be applied between the timer's V+ pin and ground.

Erratic Behavior of Output Devices

If a 555 timer powers an output device such as a relay, and the relay is not opening or closing in a reliable manner, first check that it is receiving sufficient voltage. If the 555 timer is powered with 5VDC, its output will be only around 4VDC.

This problem can be avoided by using the output from the timer to control the voltage on the base of a transistor which switches a separate source of power to the relay coil.

Fatal Damage Caused by Inductive Loads

While it is possible to drive an inductive load such as a small motor or relay directly from a TTL 555 timer, two precautions should be taken. First, the motor or the coil of the relay should have a clamping diode added around it, as is standard practice. Second, because the output of the timer is capable of sinking current as well as sourcing current, it can be protected from sinking back-EMF by inserting a diode in series with the load. This is illustrated in Figure 9-28.

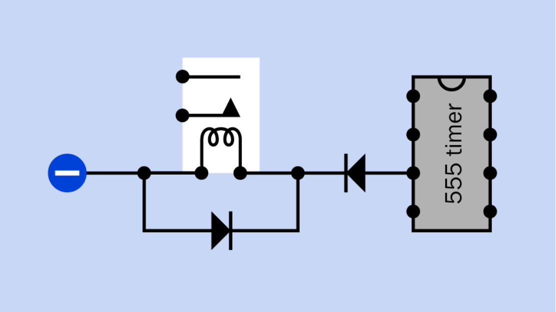

Figure 9-28. *In addition to a standard protection diode clamped around an inductive load such as a relay coil, the 555 timer can be protected from back-EMF by adding a diode in series. The series diode must of course be rated to carry sufficient current through the coil. When choosing a relay, allowance must be made for the voltage drop that will be imposed by the series diode.*

logic gate

10

Only basic **logic gates** are included in this entry—that is, components that perform a Boolean logic operation on two to eight inputs (or one input, in the case of an inverter) to create a single high or low logical output.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)

What It Does

A **logic gate** is a circuit that delivers an output, either high or low, depending on the states of its two inputs, either or both of which can be high or low.

Some gates may have more than two inputs, and an *inverter* only has one input, but the basic gates all conform with the two-input, one-output model. The components that constitute a logic gate are almost always etched into a wafer inside a silicon chip.

In a digital computer, a *high logic state* is traditionally close to 5VDC and represents a value of 1 in *binary arithmetic*, while a *low logic state* is traditionally close to 0VDC and represents a binary 0. Modern devices may use a lower voltage for the high state, but the principle is still the same.

A small network of logic gates can perform binary addition, and all other operations in a digital computer are built upon this foundation.

Origins

The concept of digital logic originated in 1894, when English mathematician George Boole announced his invention of a form of algebra (now referred to as *Boolean algebra*) to analyze combinations of two logical states that could be interpreted as "true" and "false." This concept had few practical applications until the 1930s, when Claude Shannon saw that because a basic switch has two states, Boolean algebra could enable analysis of complex networks of switches that were being used in telephone systems.

Because the state of a switch could also be used to represent the values 0 or 1 in binary arithmetic, and because a transistor could function as a switch, Boolean algebra was implemented in solid-state digital computing equipment.

How It Works

While conventional arithmetic uses *arithmetical operators* to represent procedures such as addition or multiplication, Boolean algebra uses *Boolean operators*. The operators of special interest in digital electronics are named AND, NAND, OR, NOR, XOR, and XNOR.

Although each gate actually contains multiple transistors, it is represented by a single logic symbol, as shown in Figure 10-1. The names of the Boolean operators are customarily printed all in caps. A gate requires a power supply and a connection with negative ground, separate from its inputs, but these connections are omitted from gate schematics because they are assumed to exist.

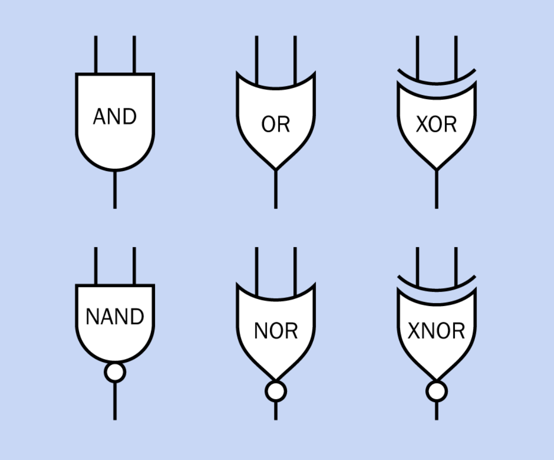

Figure 10-1. *Six types of two-input logic gates are used in digital electronics, although the XNOR gate is rare, as it has few applications. The names are customarily printed in uppercase letters.*

The functions of the gates with two inputs can be defined in electrical terms. In Figure 10-2, the four possible combinations of inputs are tabulated in the left column, with red indicating a high input and black indicating a low input. The corresponding output from each gate is shown beneath its name. This kind of tabulation is known as a *truth table*, as it is derived from Boolean algebra which originally concerned itself with "true" and "false" states.

Input states of a two-input gate	Gate Outputs					
	AND	NAND	OR	NOR	XOR	XNOR
⚫ ⚫	⚫	🔴	⚫	🔴	⚫	🔴
⚫ 🔴	⚫	🔴	🔴	⚫	🔴	⚫
🔴 ⚫	⚫	🔴	🔴	⚫	🔴	⚫
🔴 🔴	🔴	⚫	🔴	⚫	⚫	🔴

Figure 10-2. *The four possible combinations of input states in a 2-input logic gate are shown at left. The corresponding output from each gate is shown beneath its name. Red indicates a high state, while black indicates a low state.*

The truth table assumes that *positive logic* is being used. Negative logic is very uncommon, but if it were used, the red dots in the truth table would correspond with low inputs and outputs, while the black dots would correspond with high inputs and outputs.

Inversion

The small circles appended to the outputs of NAND, NOR, and XNOR gates mean that the output of each gate is inverted compared with the AND, OR, and XOR gates. This can be seen by inspection of the output states shown in Figure 10-2. The circles are known as *bubbles*.

Sometimes logic symbols are shown with a bubble applied to one input, as in Figure 10-3. In these cases, the circle indicates that an input must be inverted. More than one gate may be needed to achieve this logic function in an actual circuit. The style is often used to show the inner workings of an IC, using a minimum number of logic symbols.

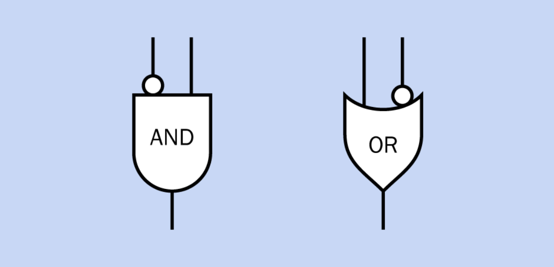

Figure 10-3. *The circle in a logic gate symbol indicates that a signal is being inverted. Circles can be inserted at gate inputs, but in a real circuit a separate inverter is likely to be needed to create this effect.*

Single-Input Gates

Two gates exist that have a single input and a single output, shown in Figure 10-4. The *buffer* should not be confused with the symbol for an **op-amp** or a **comparator**. (Those components always have two inputs.) The output state of a buffer is the same as its input state, but the component may be useful to deliver more current or to isolate one section of a circuit from another.

When a bubble is appended to a buffer, it becomes a NOT gate, more commonly known as an *inverter*. Its function is to create an output state that is opposite to its input state.

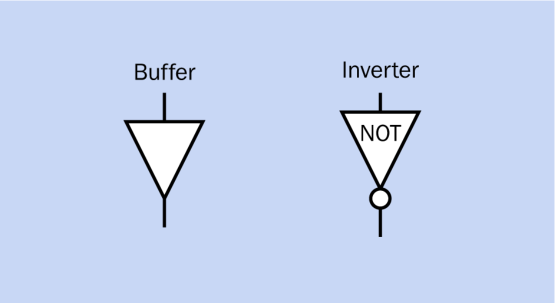

Figure 10-4. *The two logic gates that have only one input and one output. Note that in some schematics showing internal logic of ICs, the bubble on an inverter may be found on the input side instead of the output side.*

Gates with More than Two Inputs

AND, NAND, OR, and NOR gates can have any number of inputs, as suggested in Figure 10-5, although practical factors usually limit the inputs to a maximum of eight.

Input states of a gate with more than two inputs	Gate Outputs			
	AND	NAND	OR	NOR
All low	●	🔴	●	🔴
All high	🔴	●	🔴	●
At least 1 low and at least 1 high	●	🔴	🔴	●

Figure 10-5. *The previous table has been modified to show the outputs from logic gates that have more than two inputs. XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.*

The rules can be summarized like this:

- Output from an AND gate: Low if any of its inputs is low, high if all of its inputs are high.
- Output from a NAND gate: High if any of its inputs is low, low if all of its inputs are high.
- Output from an OR gate: High if any of its inputs is high, low if all of its inputs are low.
- Output from a NOR gate: Low if any of its inputs is high, high if all of its inputs are low.

In the case of XOR and XNOR gates, their logic requires that a unique output state must exist if one input is high while the other input is low.

In fact, so-called three-input XOR gates do exist, an example being the 74LVC1G386 chip, in which the output is high if all three inputs are high, or if one input is high, but not if two inputs are high or no inputs are high. Further discussion of more-than-two-input XORs is outside the scope of this encyclopedia.

Boolean Notation

For reference, the original written notation for Boolean operators is shown in Figure 10-6. Unfortunately, the notation for these operators was never properly standardized, and in more than one instance, multiple symbols acquired the same meaning. The letters P and Q are often, but not always, used to represent two input states that can be true or false.

- The use of a horizontal line above a symbol, to indicate that its state has been reversed, has carried over to datasheets where this notation can show that an output state from any digital chip is inverted. The line is known as a *bar*.

Arithmetical Operations

Suppose we wish to sum two binary numbers, each containing two digits. There are four digits altogether, and depending on their values, there are 16 different possible addition sums, as shown in Figure 10-7.

If A0 and B0 represent the rightmost digits of the two numbers being added, and S0 is the sum of those two digits, inspection of the figure shows that the sum can be derived using just three rules:

1. If A0 = 0 and B0 = 0, then S = 0.
2. If A0 and B0 have opposite states, then S0 = 1.
3. If A0 = 1 and B0 = 1, then S0 = 0, and carry 1 to the next place left.

If A0 and B0 are the two inputs to an XOR logic gate, the output of the gate satisfies all three rules, except the need to carry 1 to the next place left. This last function can be satisfied with an AND gate. The function of two gates is known as a *half adder*, and is shown in the top section of Figure 10-8.

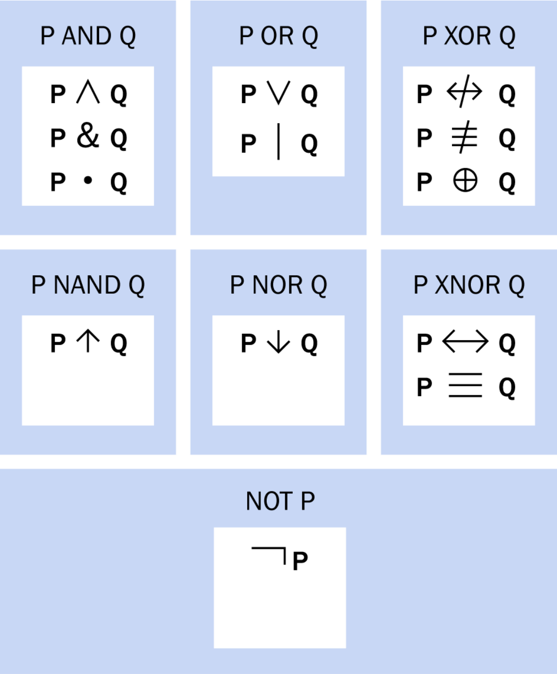

Figure 10-6. *Boolean operators as they have been expressed in written notation. Lack of standardization has resulted in more than one symbol representing some of the operators.*

When we consider the next pair of binary digits to the left, the situation now becomes more complicated, because we may be carrying 1 into this addition sum from the previous stage, and we still need to be able to to carry 1 out (if necessary) to the next stage. An assembly of five logic gates can deal with this, and their combination is known as a *full adder*. This is shown in the bottom section of Figure 10-8.

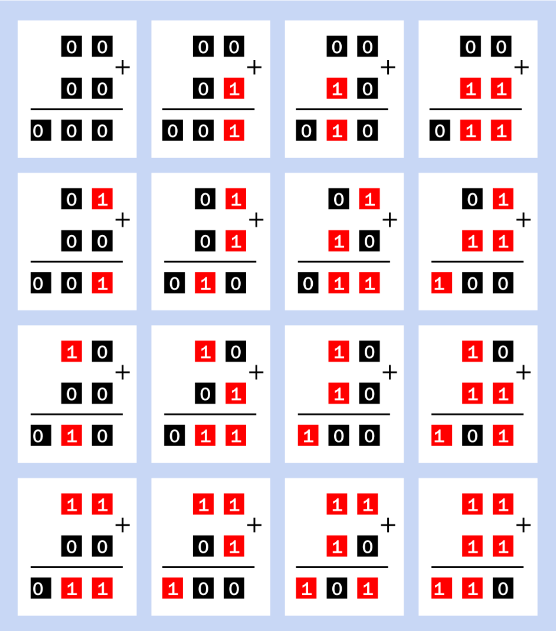

Figure 10-7. *Sixteen different addition sums are possible, when summing two binary numbers of two digits each.*

The combination of XOR and AND gates shown in Figure 10-8 is not the only one that works to add binary numbers. However, it may be the most intuitively obvious.

Other Operations

Binary arithmetic remains the most important application of logic gates, but individually packaged gates are seldom used for that purpose anymore. They were long since subsumed into large multifunction computing chips.

Single gates still have application in small systems, or to modify the inputs and outputs of microcontrollers, or to convert the output from one complex digital chip to make it compatible with the input of another. This last application is often referred to as *glue logic*.

Applications for single gates are discussed in "How to Use It" on page 103.

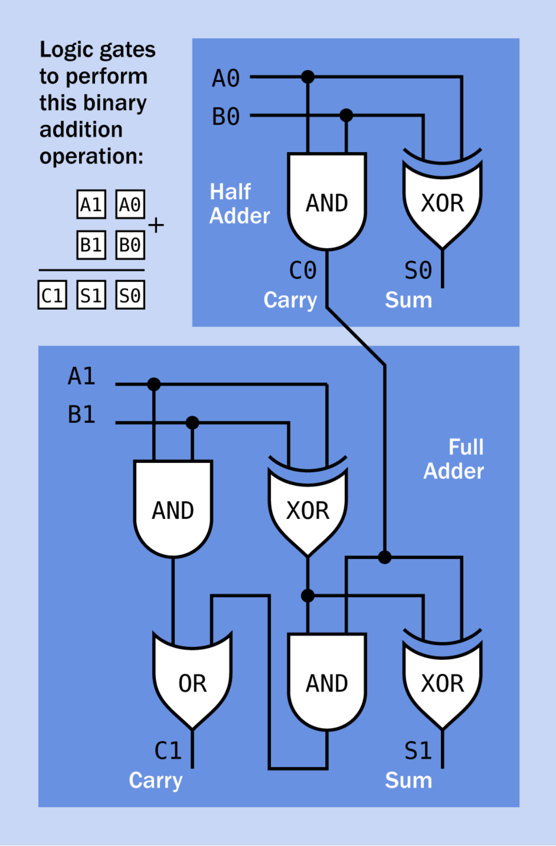

Figure 10-8. *Logic gates can be used to add binary numbers, using a high input or output to represent a binary 1 and a low input or output to represent a binary 0. This schematic shows one possible way for gates to add two two-digit binary numbers.*

Variants

Chips containing logic gates were introduced in the 1960s. The 7400 NAND chip, from Texas Instruments, was the first of a series that became so influential, the same basic part numbers (with letters added before, after, and among the digits) are still used today. An example of a currently available through-hole 7400 chip is shown in Figure 10-9.

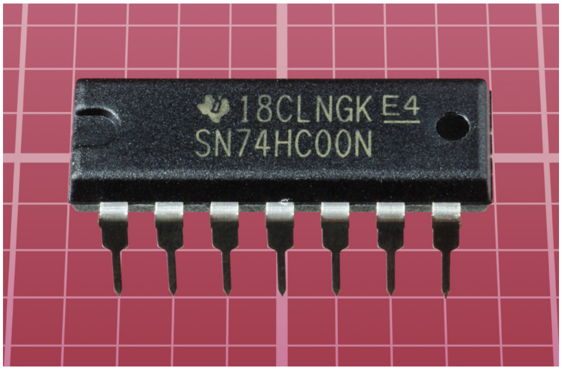

Figure 10-9. *A modern version of a 7400 chip containing four NAND gates.*

Initially, these chips conformed with a *transistor-transistor logic* (*TTL*) standard that had been invented at TRW in 1961 and introduced in commercial products by Sylvania in 1963. It established the now-familiar standard of 5VDC for the power supply. Many logic chips now use other voltages, but the term “high” still means an input or output that is near to the supply voltage, while “low” means an input or output that is near to negative ground. The exact definition of “near” will be found in datasheets for the chips.

The *7400 series* was successful partly because it was engineered for compatibility. The output from one gate could be connected directly to the input of another gate, with a few bypass capacitors added on a circuit board to suppress voltage spikes caused by rapid switching. Earlier components had not been so easy to interface with each other. The new standard dominated the industry to the point where dozens of manufacturers started making chips that conformed with it, and a single board could mix-and-match chips from multiple vendors.

Because many logic chips acquired part numbers that began with 74, they are often referred to as the 74xx series, where other digits (sometimes more than two) can be substituted for xx. This avoids ambiguity, as the very first chip in this format was a NAND gate that had 7400 as its actual part number. In the text below, 7400 will refer to that specific chip, whereas 74xx will refer to the whole series of chips.

RCA introduced a competing family of logic gates in 1968, using CMOS transistors. As each part number began with a 4 and contained four digits, this was referred to as the *4000 series*. The CMOS chips were slower and more expensive, but tolerated a wider range of power supply voltages (3V to 12V, initially). Their biggest advantage was that they used much less current. This was important, as TTL chips created a lot of waste heat. The lower power consumption of CMOS also enabled one chip to control the inputs of many others, which simplified circuit design. This one-to-many relationship is known as *fanout*.

Ultimately, CMOS chips transcended their early limitations. While they were reserved initially for battery-powered devices in which very low power consumption was more important than speed, CMOS is now used almost everywhere, still maintaining its advantage of low current (almost zero, in fact, while a chip is quiescent) while equalling the speed of TTL. However, CMOS logic chips are very often pin-compatible with the old TTL components, and modern CMOS part numbers are often derived from the old 74xx series.

Most CMOS logic chips in the old 4000 series are still available, and may be used in situations where a power supply greater than 5VDC is convenient.

Part Numbers

As the performance of semiconductors gradually improved, successive *families* of logic chips were introduced, identified by one-letter, two-letter, or three-letter acronyms. The acronym was inserted into the part number, so that a 7400 NAND gate in the HC (high-speed CMOS) family became a 74HC00 NAND gate.

Because these chips were available from multiple sources, the part number was also preceded with one or more letters indicating the manufacturer. And because each chip was manufactured in different versions (for example, some complied with military specifications, while others didn't), letters were also appended to the end of the part number. Today, the appended letters may indicate whether the chip is of the old through-hole format, or conforms with a more recent surface-mount format.

Summing up:

- Prefix: manufacturer ID.
- Numerals, omitting any letters in the middle: Chip functionality.
- Middle letters: Chip family.
- Suffix letters: Package format.

Thus, for example, the actual part number for a 74HC00 NAND chip could be SN74HC00N, where the SN prefix indicates that it is manufactured by Texas Instruments and the N suffix means that it is in plastic dual-inline-pin (DIP) format. (The SN prefix was introduced by Texas Instruments in the earliest days of integrated circuits as an acronym for "semiconductor network," meaning that multiple transistors were "networked" on a wafer of silicon. Other manufacturers used their own schemes for part numbering, and so SN became exclusively identified with Texas Instruments.)

The system of augmenting part numbers has been further extended by inserting 1G, 2G, or 3G immediately after the family identifier, to indicate surface-mount chips that contain one, two, or three logic gates. If the "G" identifier is missing, the chip usually has four logic gates, which was the standard used in the original 74xx series. This rule applies even in surface-mount formats, where the surface-mount pads of four-gate chips have the same functions as the pinouts of the original TTL versions (except in the case of square-format surface-mount chips, which are not discussed here).

When searching a catalog to find a chip by its part number, it helps to remember that searching for a 7400 chip may not find any hits, but searching for a 74HC00 (or any other valid number containing a family identifier) is much more likely to be understood.

A key to understanding part numbers is shown in Figure 10-10. The upper part of the figure is a guide to interpreting numbers on a generic basis, while the lower part interprets the specific part number shown.

Families

As of 2013, the HC family in the 74xx series has become so widely used, it can be considered the default in the traditional DIP 14-pin format. Incremental improvements are still being made, and new families are being introduced, primarily in surface-mount formats which use lower power-supply voltages (down to around 1VDC).

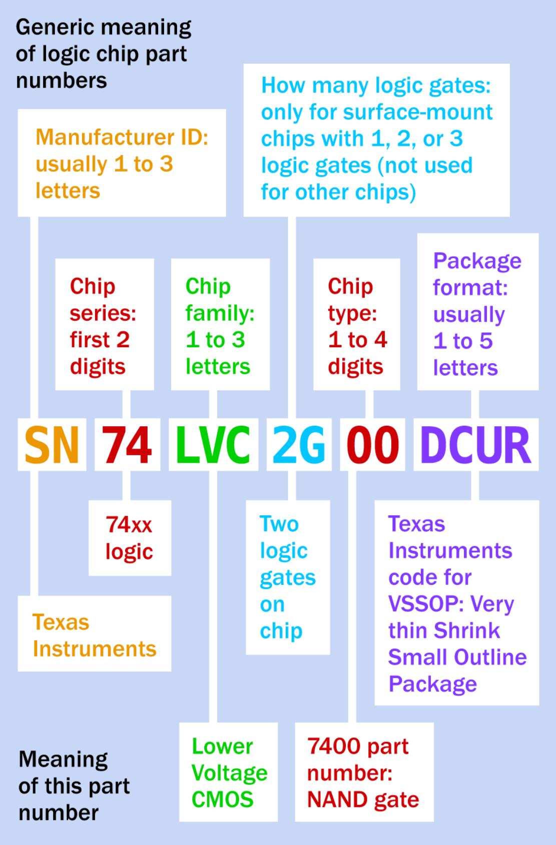
Generic meaning of logic chip part numbers
Manufacturer ID: usually 1 to 3 letters
How many logic gates: only for surface-mount chips with 1, 2, or 3 logic gates (not used for other chips)
Chip series: first 2 digits
Chip family: 1 to 3 letters
Chip type: 1 to 4 digits
Package format: usually 1 to 5 letters
SN 74 LVC 2G 00 DCUR
74xx logic
Two logic gates on chip
Texas Instruments code for VSSOP: Very thin Shrink Small Outline Package
Texas Instruments
Meaning of this part number
Lower Voltage CMOS
7400 part number: NAND gate

Figure 10-10. *How to interpret the segments of a logic chip part number in the 74xx family (in this case, a 7400 NAND gate).*

Here is an historical summary of the most important chip families.

- 74xx: Original series of bipolar TTL chips.
- 74Hxx: Bipolar TTL, high speed, about twice as fast as the original 74xx chips, but twice the power consumption.
- 74Lxx: Bipolar TTL, lower power consumption than the original TTL, but also much lower speed.
- 74LSxx: Bipolar TTL, lower power with Schottky input stages, faster than original TTL. Some LS chips are still being manufactured.
- 74ASxx: Bipolar TTL, Advanced Schottky, intended to supercede the 74Lxx.
- 74ALSxx: Bipolar TTL, Advanced Low Power Schottky, intended to supercede the 74LSxx.
- 74Fxx: Bipolar TTL, Faster.
- 74HCxx: CMOS high-speed emulation of 74LSxx.
- 74HCTxx: CMOS but with similar high-state input voltage threshold to bipolar TTL chips, for compatibility.
- 74ACxx: Advanced CMOS.
- 74ACTxx: Advanced CMOS emulation of TTL with similar high-state input voltage threshold to TTL, for compatibility.
- 74AHCxx: Advanced Higher-Speed CMOS, three times as fast as HC.
- 74VHCxx: Very High Speed CMOS.
- 74AUCxx, 74FCxx, 74LCXxx, 74LVCxx, 74ALVCxx, 74LVQxx, 74LVXxx: Various specifications, many using power supply voltages of 3.3V or below.

In the 4000 series, an early significant improvement was the 4000B family, which allowed a higher power supply limit (18V instead of 12V) and was much less susceptible to damage by static discharge. The 4000B family almost totally replaced the old 4000 family, and most 4000B chips are still available, as they are useful in situations where a power supply delivers more than 5VDC.

- When it is referenced casually, the B at the end of a chip number in the 4000 series may be omitted. When the number is listed in a catalog, the B is included.

Chips with 45 as their first two digits were introduced as a new generation, but were not widely adopted. After that, the 4000 series ceased to evolve, and CMOS chips adopted 74xx part numbers, distinguishing themselves by the insertion of letter groups in the center of the number.

To add to the confusion, some 4000 series part numbers were appended to 74xx part numbers, so that, for example, the 74HCT4060 is designed to be compatible with the old 4060B chip.

Family Interoperability

One of the most important issues relating to chip families is their differing specifications for a low-state voltage and a high-state voltage in inputs and outputs.

The original 74xx TTL series, using a 5VDC power supply, used these approximate specifications:

- Output: 74xx voltage representing a low state (at most 0.4V to 0.5V)
- Input: 74xx input voltage interpreted as a low state (maximum 0.8V)
- Output: 74xx voltage representing a high state (at least 2.4V to 2.7V)
- Input: 74xx input voltage interpreted as a high state (minimum 2V)

This provided a safe margin of error of at least 0.4V when chips were communicating with each other.

In the CMOS 4000 family, however, logic chips required a minimum input of 3V to 3.5V to be interpreted as a high state. The minimum acceptable output from a TTL chip was below this level, creating problems if anyone should try to use the output from a TTL chip to communicate with an input on a CMOS chip.

One solution is to add a 4.7K pullup resistor to the TTL output, guaranteeing that it won't fall too low. But this wastes power, and the need for the resistor is easily forgotten. Another option is to use the HCT or ACT family of CMOS logic. The "T" in these family names indicates that they have been engineered to share the input standards of the old TTL chips. They still deliver the same high output as other types of CMOS, making them seem to be the best possible solution. Unfortunately, it does entail a compromise: the "T" chips are more sensitive to noise, among other factors.

- Ideally, chip families should not be intermingled.

Gates per Chip

Each of the original 74xx chips contained multiple gates within the limits of a uniform 14-pin through-hole format. The gates that were most commonly used had two inputs, and there were four of these gates per chip.

However, the desire for miniaturization, and the use of automatic chip-placement and soldering equipment, made one-gate and two-gate logic chips desirable and practical in surface-mount format. (Three-gate surface-mount chips exist, but are sufficiently unusual that they are not described in this encyclopedia.)

Two Inputs, Single Gate

Where a chip contains just one logic gate, it is almost always a surface-mount component, and the part number has 1G in the middle to indicate "one gate." Pad functions are shown in Figure 10-11. The layout is standardized for all logic gates, with the exception of XNOR gates, which are not manufactured in surface-mount format.

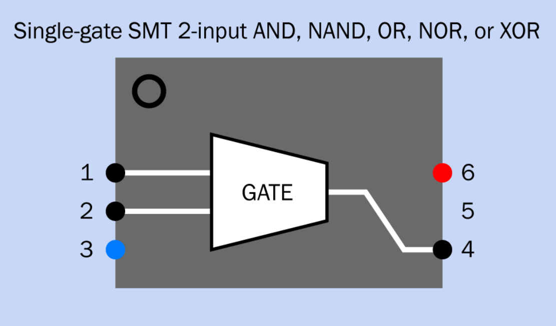

Figure 10-11. *Internal configuration and solder-pad functions for a two-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, NOR, or XOR gate. XNOR gates are not manufactured in this format.*

In the figure, a gate is shown in generic form, indicating that it may be an AND, NAND, OR, NOR, or XOR gate, depending on the part number of the chip. Inputs are on the left of the gate, while its output is on the right. The chip does not have a solder pad in position 5, but the pad at top right is identified as pin 6 for consistency with the numbering pattern in other surface-mount components where six pads are common.

The generic part numbers for single-gate surface-mount two-input logic chips are shown here, with letter x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND gate: x74x1G08x
- OR gate: x74x1G32x
- NAND gate: x74x1G00x
- NOR gate: x74x1G02x
- XOR gate: x74x1G86x

Three Inputs, Single Gate

AND, NAND, OR, and NOR single gates are available with more than two inputs. Their output is determined by rules shown in Figure 10-5. XOR and XNOR gates are not included in the table, because a strict interpretation of their logic requires that a unique output state exists if one input is high while the other is low.

The pad functions for a surface-mount single-gate logic chip with three inputs are shown in Figure 10-12. The generic part numbers for these chips are shown below. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

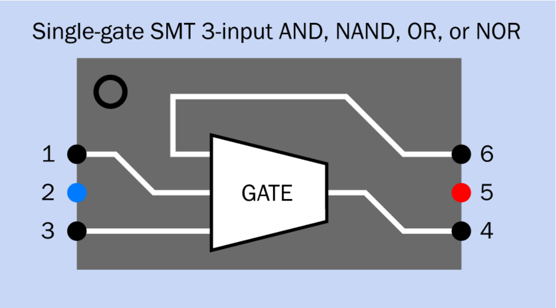

Figure 10-12. *Internal configuration and solder-pad functions for a three-input surface-mount single-gate logic chip that can contain an AND, NAND, OR, or NOR gate.*

The generic part numbers for single-gate surface-mount three-input logic chips are shown here, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x1G11x
- NAND: x74x1G10x
- OR: x74x1G32x
- NOR: x74x1G27x

Single Gate, Selectable Function

A few surface-mount chips can emulate a variety of two-input gates, by using appropriate external connections. The internal logic of one example, with generic part number x74x1G97x (an actual example would be Texas Instruments SN74LVC1G97), is shown in Figure 10-13. Depending which pin is grounded and which other pins are used as inputs, the chip can emulate all five of the most commonly used gates. To achieve this, however, some inputs have to be inverted.

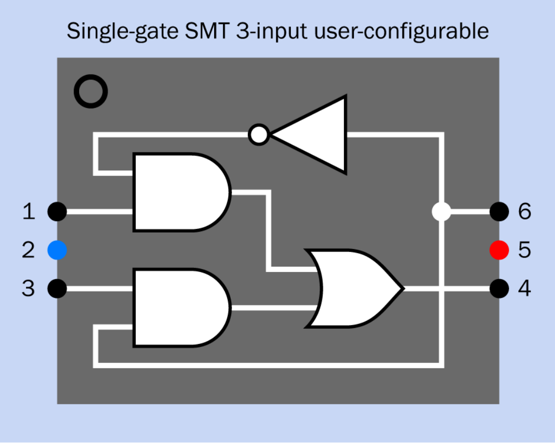

Figure 10-13. *Internal configuration for a configurable surface-mount chip that can emulate various two-input logic gates, depending which inputs are used and which are grounded. Some inputs have to be inverted to emulate some gates.*

Two Inputs, Dual Gate

Two-input surface-mount AND, NAND, OR, NOR, and XOR gates are available in dual layout (two gates per chip). The internal logic and pad functions are shown in Figure 10-14. The generic part numbers for these chips are shown here. Again, each x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

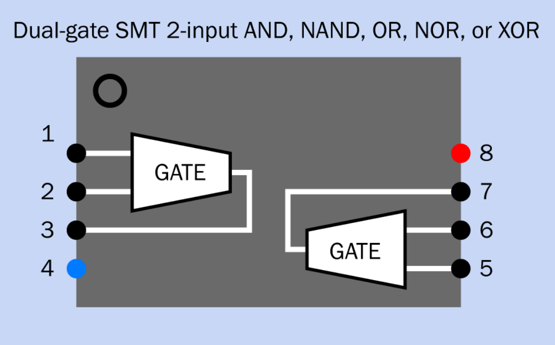

Figure 10-14. *Internal configuration and solder-pad functions for a two-input surface-mount dual-gate logic chip that can contain two AND, NAND, OR, NOR, or XOR gates. XNOR chips are not manufactured in this format.*

The generic part numbers for dual-gate surface-mount two-input logic chips are shown below, with each x indicating that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format:

- AND: x74x2G08x
- NAND: x74x2G00x
- OR: x74x2G32x
- NOR: x74x2G02x
- XOR: x74x2G86x

Original 74xx 14-Pin Format

Each of the original 74xx TTL chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates with two inputs each
- Triple 3-input: Three gates with three inputs each
- Dual 4-input: Two gates with four inputs each
- Dual 5-input: Two gates with five inputs each
- Single 8-input: One gate with eight inputs

The five-input chips have become so uncommon that they are not described in this encyclopedia.

Quad Two-Input 74xx Pinouts

14-pin DIP 74xx quad two-input logic chips are available in AND, NAND, NOR, XOR, or XNOR versions, all of which have an internal layout shown in Figure 10-15. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

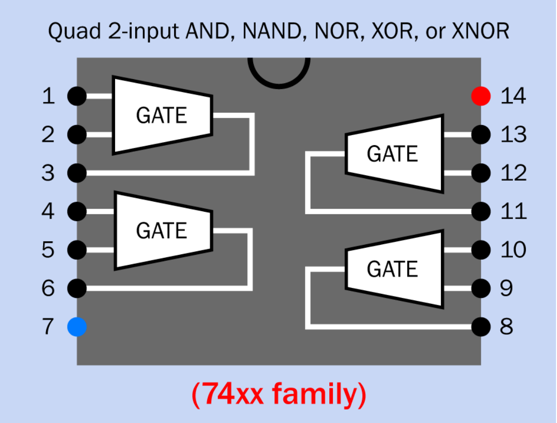
Quad 2-input AND, NAND, NOR, XOR, or XNOR
GATE
GATE
GATE
GATE
1
2
3
4
5
6
7
14
13
12
11
10
9
8
(74xx family)

Figure 10-15. *In a 14-pin quad two-input 74xx logic chip, the AND, NAND, NOR, XOR, and XNOR versions all share this generic layout.*

- The 14-pin quad two-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-16.

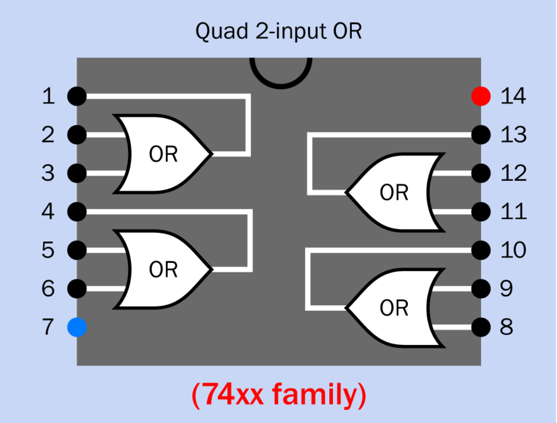
Quad 2-input OR
OR
OR
OR
OR
1
2
3
4
5
6
7
14
13
12
11
10
9
8
(74xx family)

Figure 10-16. *In a quad two-input 74xx OR chip, this layout is used, which is different from that used in all the other quad two-input logic gates.*

Triple Three-Input 74xx Pinouts

The AND, NAND, and NOR versions of a 14-pin DIP 74xx triple three-input logic chip all have an internal layout shown in Figure 10-17. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. Three connections leading to a gate are its inputs, while a single connection from a gate is its output.

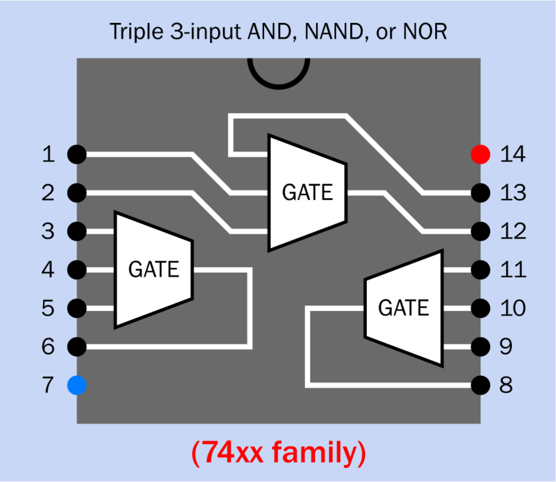
Triple 3-input AND, NAND, or NOR
GATE
GATE
GATE
1
2
3
4
5
6
7
14
13
12
11
10
9
8
(74xx family)

Figure 10-17. *In a 14-pin triple three-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout.*

- The 14-pin triple three-input OR chip has different pinouts from all the other 74xx logic chips. It is shown in Figure 10-18.

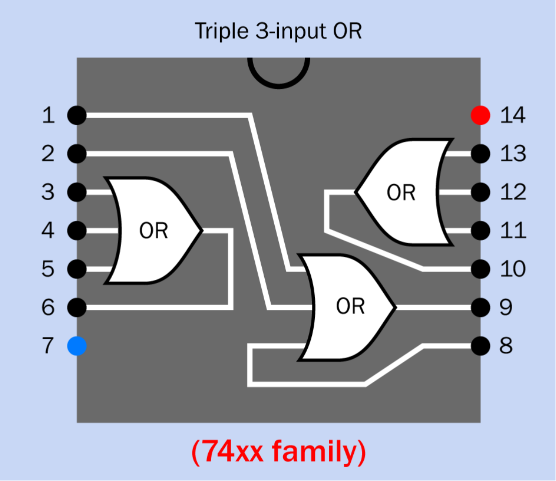
Triple 3-input OR
OR
OR
OR
1
2
3
4
5
6
7
14
13
12
11
10
9
8
(74xx family)

Figure 10-18. *In a triple three-input 74xx OR chip, this layout is used, which is different from that used for all the other triple three-input logic gates.*

Dual Four-Input 74xx Pinouts

A 14-pin DIP 74xx dual four-input logic chip contains two four-input gates. The AND, NAND, and NOR versions all have an internal layout shown in Figure 10-19. The layout is unchanged in surface-mount format. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type.

- There is no OR chip of the 14-pin dual four-input type.

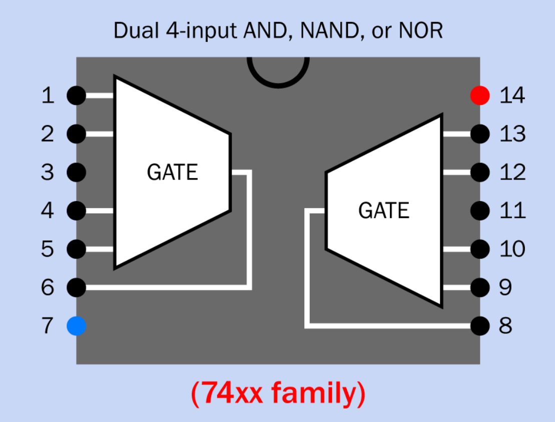

Figure 10-19. *In a 14-pin dual four-input 74xx logic chip, the AND, NAND, and NOR versions all share this generic layout. There is no 74xx OR chip with four inputs per gate.*

Single Eight-Input 74xx Pinouts

A 14-pin DIP 74xx single eight-input NAND chip contains one eight-input gate, as shown in Figure 10-20. The layout is unchanged in surface-mount format.

- There is no AND chip of the 14-pin single eight-input type.

A 14-pin eight-input logic chip in the 74xx series, able to function as both an OR and a NOR, is shown in Figure 10-21. The output from the NOR gate is connected with pin 13, but also passes through an inverter to create an OR output at pin 1. (Because a NOR gate is equivalent to an inverted-OR, when its output is inverted again, it returns to being an OR.)

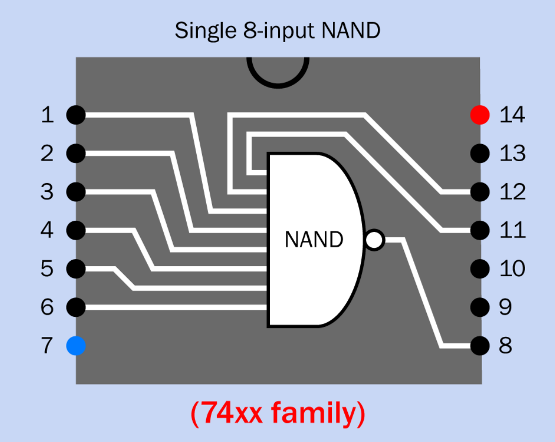

Figure 10-20. *The internal layout of single eight-input NAND chip in the 14-pin 74xx series. There is no 74xx AND chip with eight inputs per gate.*

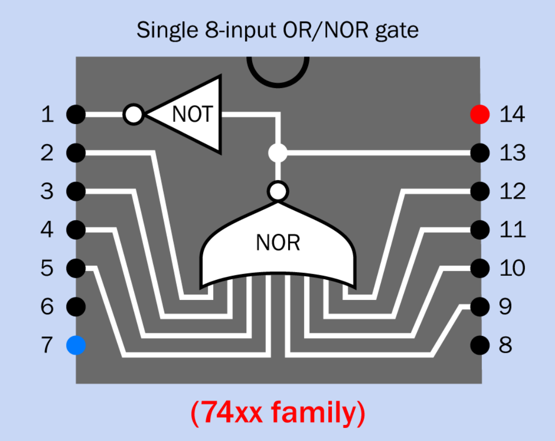

Figure 10-21. *The internal layout of single eight-input OR/NOR chip in the 14-pin 74xx series. Pin 13 has the NOR output, while pin 1 has the OR output.*

The following list shows the generic part numbers for DIP and surface-mount versions of 14-pin logic chips in the 74xx series that have two or more inputs per gate. As before, an x indicates that letter sequences are likely to be inserted to indicate manufacturer, logic family, and chip format.

- Quad 2-input AND: x74x08x
- Quad 2-input NAND: x74x00x
- Quad 2-input OR: x74x32x
- Quad 2-input NOR: x74x02x
- Quad 2-input XOR: x74x86x
- Quad 2-input XNOR: x74x266x
- Triple 3-input AND: x74x11x
- Triple 3-input NAND: x74x10x
- Triple 3-input OR: x74x4075x
- Triple 3-input NOR: x74x27x
- Dual 4-input AND: x74x21x
- Dual 4-input NAND: x74x20x
- Dual 4-input NOR: x74x4002x
- Single 8-input NAND: x74x30x
- Single 8-input OR/NOR: x74x4078x

74xx Inverters

Single, dual, and triple inverter packages in the 74xx series are available in surface-mount format only. Their internal arrangement is shown in Figures 10-22, 10-23, and 10-24.

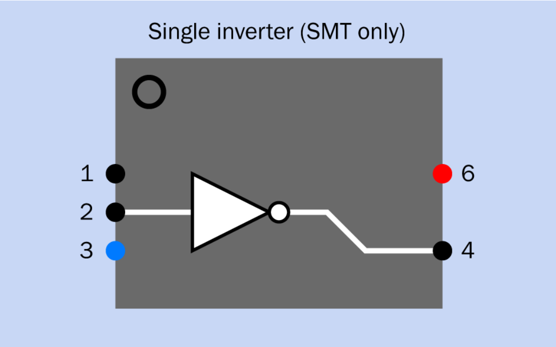

Figure 10-22. *The internal layout of a 74xx series logic chip containing one inverter. This is available in surface-mount format only. Pin 5 is absent. Pin 1 is not connected.*

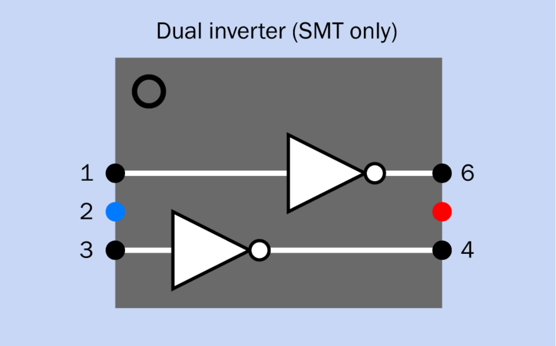

Figure 10-23. *The internal layout of a 74xx series logic chip containing two inverters. This is available in surface-mount format only.*

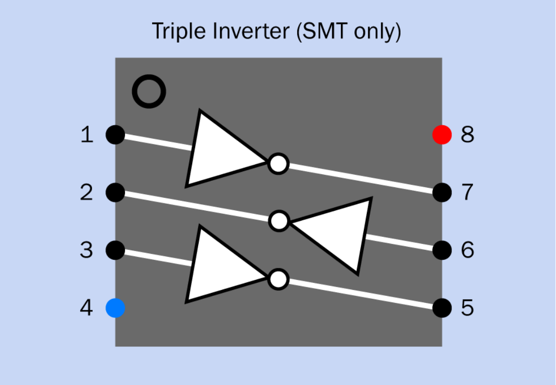

Figure 10-24. *The internal layout of a 74xx series logic chip containing three inverters. This is available in surface-mount format only.*

In the 14-pin format, a hex inverter chip (containing six inverters) is available, as shown in Figure 10-25. The layout is the same for DIP and surface-mount formats.

Generic part numbers for inverter chips are as follows:

- Single inverter: x74x1G04x
- Dual inverter: x74x2G04x
- Triple inverter: x74x3G14x
- Hex inverter: x74x04x

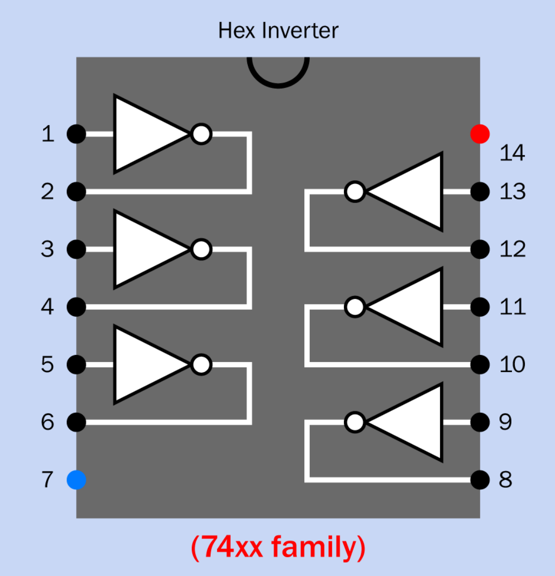

Figure 10-25. *The internal layout of a 14-pin 74xx hex inverter logic chip, containing six inverters. This layout is the same for DIP and surface-mount versions.*

Additional Variations

Some chips in the 74xx series (both DIP and surface mount versions) have variants with open drain or open collector outputs, while others have inputs that are configured as Schmitt triggers. These variants will be found as hits when searching supplier websites for logic chips by gate name and number of inputs.

Pinouts in the Original 4000 Series

Each of the original 4000 CMOS chips contained multiple gates within the limits of a uniform 14-pin chip format. The available options were, and still are:

- Quad 2-input: Four gates of two inputs each
- Triple 3-input: Three gates of three inputs each
- Dual 4-input: Two gates of four inputs each
- Single 8-input: One gate of eight inputs

In the 4000 family, 14-pin quad two-input logic chips are available in AND, OR, NAND, NOR, XOR, or XNOR versions, all of which have an internal layout shown in Figure 10-26. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The four connections leading to a gate are its inputs, while the single connection from a gate is its output.

Unlike the 74xx family, the quad two-input OR chip in the 4000 family has the same pinouts as the other types of quad two-input logic chips.

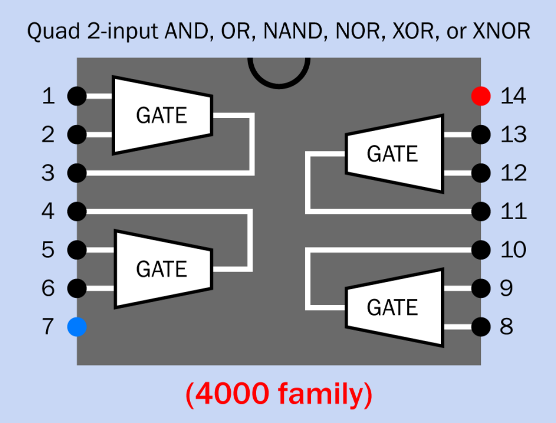

Figure 10-26. *In the 4000 family, the AND, OR, NAND, NOR, XOR, and XNOR versions of a quad two-input logic chip all share this generic layout.*

In the 4000 family, a 14-pin triple three-input logic chip contains three three-input gates. The AND, OR, NAND, and NOR versions all have an internal layout shown in Figure 10-27. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. The three connections leading to a gate are its inputs, while a single connection from a gate is its output.

Unlike the 74xx family, the triple three-input OR chip in the 4000 family has the same pinouts as the other types of triple three-input logic chips.

In the 4000 family, a 14-pin dual four-input logic chip contains two four-input gates. The AND, NAND, OR, and NOR versions all have an internal layout shown in Figure 10-28. The gates are shown in generic form, as the layout remains the same regardless of which type of gate is in the chip. All the gates in any one chip are of the same type. Each pair of connections leading to a gate are its inputs, while the single connection from a gate is its output.

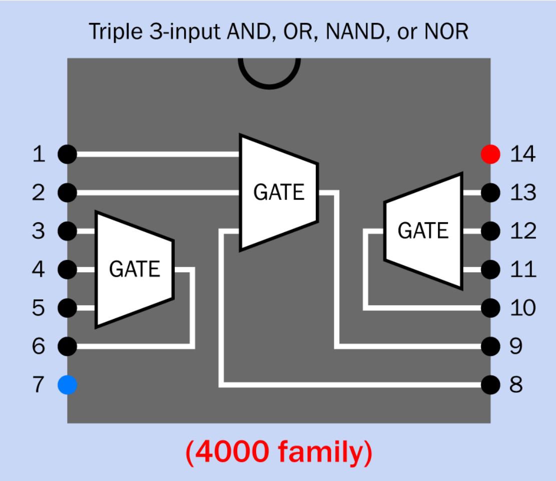

Figure 10-27. *In the 4000 family, the AND, OR, NAND, and NOR versions of a triple three-input logic chip all share this generic layout.*

Note that the 4000 family does have a dual four-input OR chip, whereas the 74xx family does not.

In the 4000 family, a 14-pin eight-input logic chip with AND and NAND outputs is available, as shown in Figure 10-29.

The following list shows the generic part numbers for 14-pin logic chips in the 4000 family that have two or more inputs per gate (in actual part numbers, letters will be substituted where an x appears):

- Quad 2-input AND: x4081x
- Quad 2-input NAND: x4011x
- Quad 2-input OR: x4071x
- Quad 2-input NOR: x4001x
- Quad 2-input XOR: x4070x
- Quad 2-input XNOR: x4077x
- Triple 3-input AND: x4073x
- Triple 3-input NAND: x4023x
- Triple 3-input OR: x4075x
- Triple 3-input NOR: x4025x
- Dual 4-input AND: x4082x
- Dual 4-input NAND: x4012x
- Dual 4-input OR: x4072x
- Dual 4-input NOR: x4002x
- Single 8-input AND/NAND: x4068x

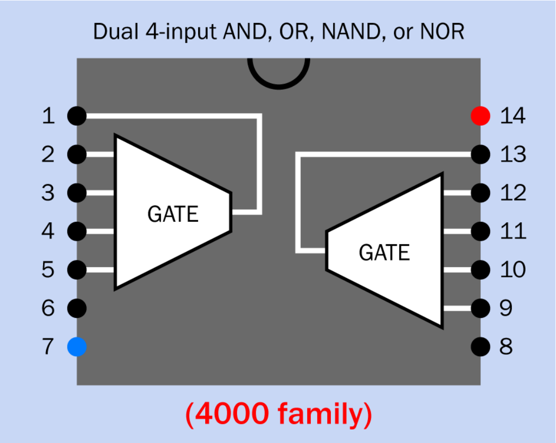

Figure 10-28. *In the 4000 family, the AND, OR, NAND, and NOR versions of a dual four-input logic chip all share this generic layout.*

4000 Series Inverters

In the 4000 family, the 4069B is a 14-pin hex inverter chip (containing six inverters), as shown in Figure 10-30. This has the same pinouts as the x74x04x chip.

How to Use It

Which Family

In DIP format, the HC family has existed for more than 30 years, and has become established as a widely used default choice.

In surface-mount formats, the choice of family will largely be determined by the choice of supply voltage.

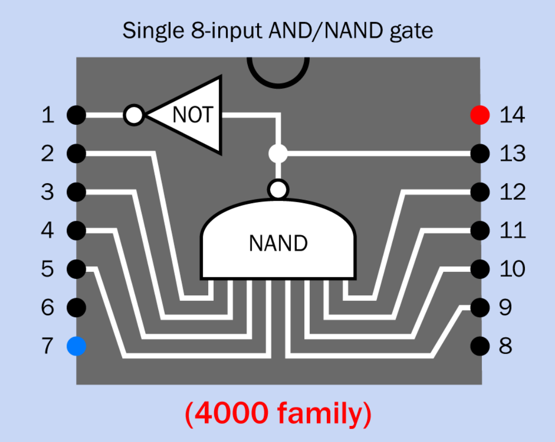

Figure 10-29. *In the 4000 family, a single eight-input AND/NAND chip has this internal layout. The inverted output from the NAND gate becomes an AND output from pin 1 of the chip.*

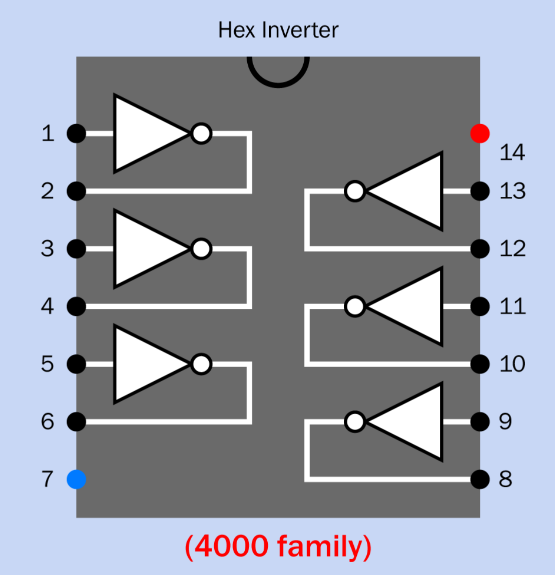

Figure 10-30. *The internal layout of a 14-pin 4069B hex inverter logic chip, containing six inverters. This layout is the same as for the x74x04x chip.*

Although the 4000 series is now more than 40 years old, it may still be useful where a 5VDC power supply is not required for other reasons in a circuit and would be added purely to power a 74xx series logic gate or other digital chip. If a circuit contains a 9VDC or 12VDC relay, for instance, a Darlington pair may be used with that voltage to drive the relay, and an old-school 4000 series logic chip could share the same supply. The relay coil would need a clamping diode to protect the logic chip from transients.

Applications

The output from a logic chip may be used as an input for a microcontroller, to enable multiple inputs to share one pin. An eight-input NAND gate, for instance, could combine the inputs from eight normally on motion sensors. If just one sensor responds to an intrusion, the gate output would change from high to low.

Logic gates may be useful in any simple device that has to respond to a single, specific combination of inputs. A digital combination lock is one example; games of chance are another. Most simple dice simulations use logic gates to convert the output from a counter to drive a dice-pattern of LEDs.

A logic gate may be used as an interface between an electromechanical **switch** and a circuit containing digital chips. A 10K pullup or pulldown resistor prevents the gate input from floating when the switch is open. A buffer can be used for this purpose, or an inverter, or any "spare" gate on a logic chip that is already in the circuit. One input of the chip can be tied to the positive power supply or negative ground, to create an appropriate input from the chip when the switch, attached to the other input, is opened or closed.

A jam-type **flip-flop** can be used to *debounce* a switch input. See Chapter 11 for details. If two NOR or two NAND gates are unused in a circuit, they can form a flip-flop.

In the original CMOS 4000 family, a positive output may be capable of driving an LED if the current does not exceed 5mA with a power supply of 5VDC or 10mA with a power supply of 10VDC. Note that the output voltage will be pulled down significantly by these loads. In the 74HCxx family, chips can source or sink as much as 20mA, but here again the output voltage will be pulled down. Note that the total limit for all outputs from a 74HCxx chip is around 70mA.

The output from a logic chip can be passed through a buffer such as the 7407, which has an open-collector output capable of sinking as much as 200mA. This enables direct drive of modest loads, so long as they are not inductive.

Solid-state relays and **optocouplers** can be driven directly from logic chips, as they draw very little current. A solid-state relay can switch 50A or more.

What Can Go Wrong

Two problems are common when using CMOS digital chips: damage from *static electricity*, and erratic behavior caused by *floating pins*.

Static

The early 4000 series CMOS chips were especially vulnerable, but more recent CMOS designs generally include diodes at the inputs, which reduces the risk. Still, logic chips should be protected by inserting them into anti-static foam or enclosing them in conductive wrappers until they are installed in a board. While handling chips, it is good practice to be grounded, ideally using a wrist-mounted ground wire.

Floating Pins

Any pin which is unconnected in a logic chip is considered to be "floating," and can pick up signals by capacitive coupling, possibly disrupting the behavior of the chip and also causing power consumption, as the ambiguous pin state will tend to prevent that gate in the chip from entering quiescent mode.

Generally speaking, input pins in a TTL logic chip that are not being used for any purpose should be tied to the positive voltage supply, while unused CMOS pins should be tied to negative ground.

Family Incompatibilities

As previously noted, older TTL logic chips may deliver a "high" output voltage that is lower than the minimum expected by newer CMOS logic chips. The best option is not to mix families, but if chips are stored carelessly, some intermingling can occur. Part numbers should be checked if one chip appears to be ignoring output from another.

Overloaded Outputs

If a circuit calls for a logic chip with an open-collector output, and a regular chip is used by mistake, it will almost certainly be damaged.

Output Pulled Down

If the output from one logic chip is connected with the input of another logic chip, and if the output from the first chip is also connected to an **LED**, the LED may pull down the output voltage so that the second chip will not recognize it as a high state. As a general rule, a logic output can drive an LED, or can drive another logic chip, but not both. Very-low-current LEDs, which draw as little as 2mA, may be acceptable.

Incorrect Polarity and Voltages

Logic chips can be knocked out by applying incorrect polarity, or voltage to the wrong pin, or the wrong voltage. Modern logic chips tolerate a very limited voltage range, and a 74xx series chip will be irrevocably damaged if it is used where a 4000 series chip was specified for a power supply higher than 6VDC.

If a chip is inserted upside-down, it will probably be damaged when voltage is applied.

Bent Pins

Like all through-hole chips, DIP logic chips can be inserted accidentally with one or more pins bent underneath the chip. This error is very easy to miss. The bent pins will not make contact with any socket that is used, and the chip will behave unpredictably. Check for proper pin insertion with a magnifying glass if necessary.

Unclean Input

Logic chips expect a clean input without voltage spikes. A 555 timer of TTL type generates spikes in its output which can be misinterpreted as multiple pulses by the input of a logic chip. A CMOS-type 555 timer is more suitable for connection with logic chips.

If a pushbutton, rotational encoder, or electromechanical switch provides a high or low input, the input must be debounced. In hardware, this is traditionally done with a **flip-flop**. It can also be done with code in a microcontroller.

Analog Input

The input of a logic chip can be connected directly with a thermistor, phototransistor, or similar analog component, but only if there is some certainty that the voltage at the input pin will remain within the range that is acceptable to the chip. In the case of a phototransistor, for example, it should be exposed to a limited, known range of light intensity.

In general, it is best to avoid applying intermediate-voltage signals to a digital logic input, as they can create unpredictable output, or output of an intermediate voltage. A **comparator** can be placed between the analog source and the digital logic chip, or a logic chip with a Schmitt-trigger input can be used.

flip-flop

11

The term **flip-flop** is sometimes printed with a space instead of a hyphen, but the hyphenated form seems to predominate in the United States. Therefore, the hyphen is included here. The term *flipflop* (with neither a hyphen nor a space) is sometimes seen, but is unusual. The acronym *FF* is confined mostly to logic diagrams or schematics.

The term *latch* is sometimes used interchangeably with *flip-flop* but is assumed here to describe a minimal asynchronous circuit that responds immediately and transparently to an input. A flip-flop can function as a latch and also as a synchronous device which is opaque, meaning that the input does not flow directly through to the output.

OTHER RELATED COMPONENTS

- **counter** (see Chapter 13)
- **shift register** (see Chapter 12)

What It Does

Transistors enable logic gates; logic gates enable flip-flops; and flip-flops enable many mathematical, storage, and retrieval functions in digital computing. Most flip-flops today are embedded in much larger integrated circuits that have complex functions. However, they are still available as separate components in chip form, and will be discussed on that basis here.

A flip-flop is the smallest possible unit of memory. It can store a single bit of data, represented by either a high or low *logic state*. (A full explanation of logic states is included in the **logic gate** entry. See Chapter 10.) Flip-flops are especially useful in **counters**, **shift registers**, and *serial-to-parallel converters*.

A flip-flop circuit can be classified as a form of *bistable multivibrator*, as each of its outputs is stable in one of two states until an external trigger stimulates it to "flip" from one state and "flop" into the other. (For a comprehensive discussion of monostable and astable multivibrators, see the **timer** entry in Chapter 9.)

An *asynchronous* flip-flop will respond immediately to a change of input, and can be used for applications such as debouncing the signal from an electromechanical **switch** or building a *ripple* **counter**. More often, a flip-flop is *synchronous*, meaning that a change in input state will be unrecognized until it is enabled by a low-to-high or high-to-low transition in a stream of pulses from an external *clock*.

How It Works

Every flip-flop has two outputs, each of which may have a high or low state. When the flip-flop is functioning normally, the outputs will be in opposite logical states, one being high while the other is low. These outputs are typically identified as Q and NOT-Q (the latter term meaning a letter Q with bar printed above it, sometimes referred to verbally as "Q-bar"). In datasheets and other documents where a bar symbol cannot be represented easily above a letter, the NOT-Q output may be represented as letter Q with an apostrophe after it, as in Q'.

Almost always, in a schematic diagram, a flip-flop is represented by a simple rectangle, with inputs and outputs identified by letters and other marks. Because a description of the inner workings is necessary before the different types of flip-flop can be understood, schematic symbols for various flip-flops will not be introduced until "Variants" on page 116.

The simplest flip-flop contains two logic gates whose function can be most easily understood if the inputs are controlled by a SPDT switch. It can be created from two NAND gates or two NOR gates, as described next. This type of component can be described as:

- *asynchronous*: Will accept data on an impromptu basis, as it is not synchronized with a clock.
- *jam-type*: Colloquial equivalent of *asynchronous*. The input is jammed in at any time, forcing an immediate change of output.
- *transparent*: The input state flows straight through to the output.

NAND-Based SR Flip-Flop

Figure 11-1 shows two NAND gates attached to a SPDT switch, with two pullup resistors. When either of the NANDs has a floating input from the switch, the pullup resistor attached to that input will maintain it in a high state. The data inputs for the NAND gates are labeled S and R, meaning Set and Reset, giving this component its name as an *SR* flip-flop:

- In a NAND-based SR flip-flop, a low state is considered an active logic input, as indicated by the bar placed above each letter.
- A high state is considered an inactive logic output.

The schematic style in this figure, with diagonally crossing conductors, is universally used and easily recognizable. The equivalent schematic in Figure 11-2, which might be created by circuit-drawing software, has the same functionality but would not be immediately recognizable as a flip-flop. The "classic" crossed-conductor representation is preferable.

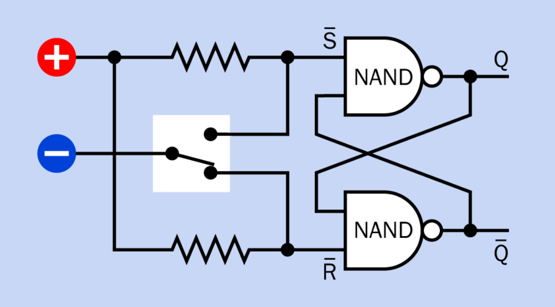

Figure 11-1. *The schematic for a simple NAND-based SR-type flip-flop, with a switch and pullup resistors added to control two inputs.*

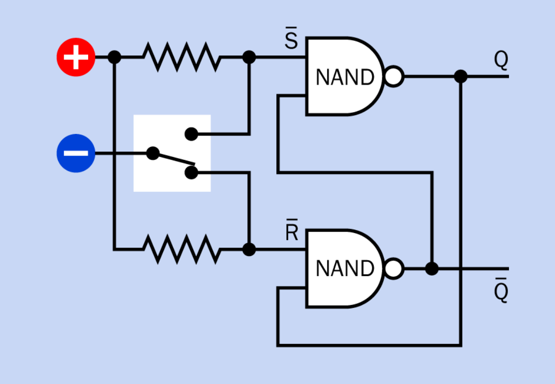

Figure 11-2. *An alternative component layout for an SR flip-flop, functionally identical to the previous schematic, but not so easily recognizable. The layout with a pair of diagonally crossing conductors has become so ubiquitous, it should be considered to be a standard.*

The first step toward understanding the behavior of flip-flops is to recall the relationship between the two inputs and the output of NAND or NOR gates. This is shown in Figure 11-3, where red indicates a high logic state and black indicates a low logic state.

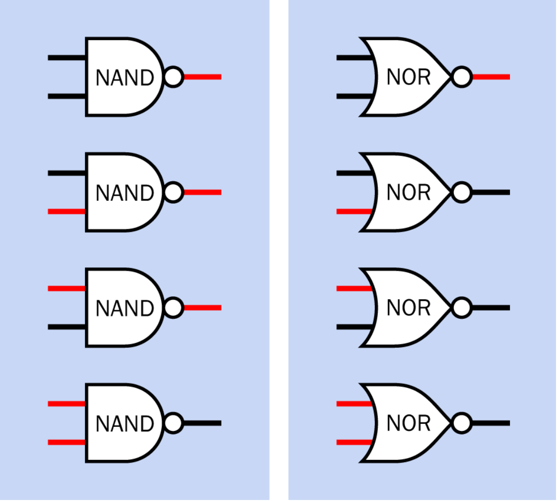

Figure 11-3. *The four possible input combinations for a NAND gate and a NOR gate, with the corresponding logical outputs. A flip-flop can be built from two NANDs or two NORs.*

The behavior of a NAND gate can be summarized:

- Both inputs high: Output low.
- Other input combinations: Output high.

Figure 11-4 shows a series of snapshots of the SR flip-flop circuit as the switch moves from one position, through an intermediate state where it makes no connection, to the other position. Remember that in this circuit, the active logical input state is low, and the active logical output state is high.

In the top panel, the pullup resistor of the lower NAND is overwhelmed by the direct connection to negative ground, which holds the R input in a low state. The other input of this gate is irrelevant, because the output from a NAND will be high if either of its inputs is low. So, the lower NAND has a high output, which feeds back to the secondary input of the upper NAND gate. The S input of this gate is high, because of the pullup resistor. Because both inputs of this gate are now high, its output is low, which feeds back to the lower gate. The lower gate doesn't change its output, because either of its low inputs is enough to keep its output high. So, the circuit is in equilibrium. A high state on the NOT-Q output is known as the *Reset* state for a NAND-based flip-flop.

The second panel shows what happens if the switch now moves up into a neutral, disconnected position. The R input of the lower NAND now becomes high, because of the pullup resistor. But this NAND still has one low input, supplied by the output of the upper gate, so its output remains high, and the circuit is still in equilibrium. This is known as the *Hold* state for the NAND circuit.

Suppose the switch bounces to and fro between the states shown in the first two panels. The output from the circuit won't change. This shows that the circuit provides a method for eliminating *switch bounce*—the very fast, momentary spikes that occur when the mechanical contacts of a switch open and close.

The third panel shows what happens if the switch now moves to its upper position. The top input of the upper NAND gate is now pulled low. Consequently, its output goes high. This feeds back to the lower gate. Its other input is high because of the pullup resistor. With both of its inputs high, its output goes low. The gate outputs have flipped and swapped values. A high state on the Q output is known as the *Set* state for a NAND-based flip-flop.

The circuit still remains in equilibrium even if the switch returns to its central, disconnected position shown in the bottom panel. Therefore, the debouncing capability of the circuit works equally well for both positions of the switch.

NOR-Based SR Flip-Flop

Figure 11-5 shows a similar circuit using two NOR gates attached to an SPDT switch. Because the NOR gates function differently, this circuit uses active-high input logic, and pulldown resistors are needed instead of pullup resistors. The output from the circuit still uses active-high logic, and is identical with the NAND-based circuit in this respect, although the relative positions of the Q and NOT-Q outputs have been swapped.

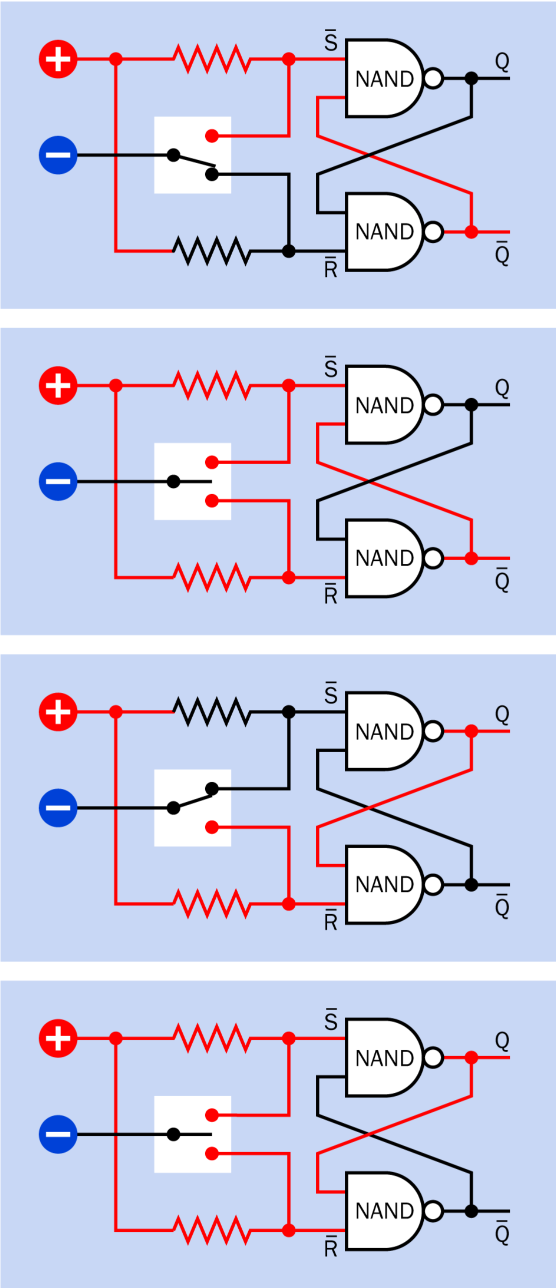

Figure 11-4. *Four snapshots of transitions in a NAND-based SR flip-flop as the switch moves down and up through an intermediate no-connection zone. See text for details.*

- In a NOR-based SR flip-flop, a high state is considered an active logic input, as indicated by the absence of a bar placed above the letters S and R.
- A high state is considered an active logic output.

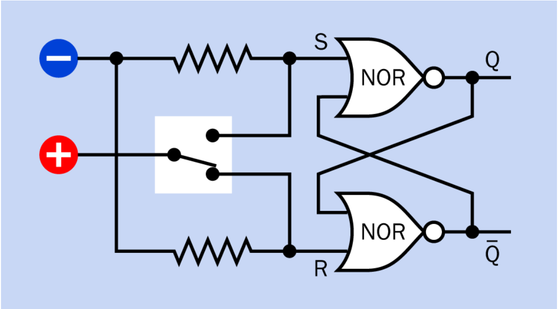

Figure 11-5. *The schematic for a simple SR flip-flop using NOR gates instead of NANDs.*

In the case of the NOR gate:

- Both inputs low: Output high.
- Other input combinations: Output low.

Figure 11-6 shows a series of snapshots as the switch moves from one position to the other, through intermediate states where it makes no connection. Remember that in this circuit, the active logical state is high at both the inputs and the outputs.

In this circuit, as in the previous circuit using NAND gates, it will ignore switch bounce, allowing the gate outputs to remain stable.

Forbidden States

Either of the circuits described so far depicts an SR flip-flop, regardless of whether it is NAND-based or NOR-based. Its input and output states are summarized in Figure 11-7. However, as this table suggests, there are some input states that create problems.

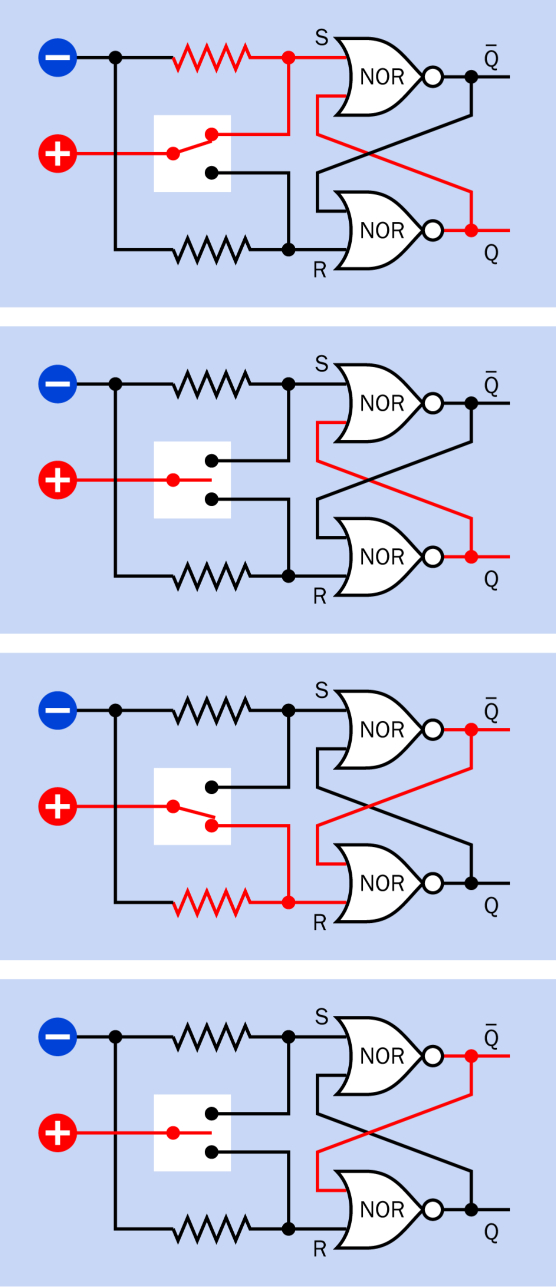

Figure 11-6. *Four snapshots of a NOR-based SR flip-flop, showing the consequence of changing switch positions, comparable with the NAND-based flip-flop.*

Flip-Flop Inputs		Flip-Flop Outputs			
		NAND-based		NOR-based	
S	R	Q	$\bar{Q}$	Q	$\bar{Q}$
●	●	🔴 Problem 🔴		Same as Previous	
●	🔴	🔴	●	●	🔴
🔴	●	●	🔴	🔴	●
🔴	🔴	Same as Previous		● Problem ●	

Figure 11-7. *A table of input states and the consequent output states for NAND-based and NOR-based SR flip-flops.*

In either the NAND-based flip-flop or the NOR-based flip-flop, the output when the switch is in its unconnected center position will remain the same as when the switch was in its previous position. This is the usefulness of the flip-flop: it remembers the previous state. These situations are identified as “Same as Previous” in the table.

The pullup resistors (in a NAND-based flip-flop) and pulldown resistors (in a NOR-based flip-flop) are intended to guarantee that both inputs will be high (NAND) or both inputs will be low (NOR) even when the switch makes no connection. Therefore, it should be impossible for both inputs to be low (NAND) or high (NOR).

But what happens if the circuit is powered up with the switch in the unconnected position? One input of each gate is controlled by the output of the other gate. But what will those outputs be?

In the NAND-based version, the outputs from the NANDs will be low while the chips are powering up. As soon as the NAND chips are functioning, each of them will sense that it has one input high and one input low, so it will change its output to high.

But now that each chip has a high output, it will feed back to the secondary input of the other chip. Now both chips have both inputs high. This will cause them both to change their outputs to low—but this will feed back again, flipping the outputs back to high again. In fact, if the gates are absolutely identical, the circuit will oscillate very rapidly. This is sometimes known as *ringing*.

In real life, the gates will not be absolutely identical, and eventually one of them will respond fractionally ahead of the other, tipping the circuit into the state shown either in the second panel or the fourth panel in Figure 11-4. But which chip will win? There is no way of knowing. This is known as a *race condition*, and the winner is unpredictable.

A similar but opposite situation occurs in the NOR-based flip-flop if it is powered up with the switch in the disconnected position, and the S and R outputs are both low, because of the pulldown resistors. Here again it will be a race condition.

We can address the problem by making a rule that the switch must always be in one position or the other when the flip-flop is powered up. But what if there is a faulty switch? Or what if a power interruption occurs while the switch is changing position?

Another problem occurs if the switch makes one contact a fraction before it breaks the other contact. This would result in both S and R inputs being low, in a NAND flip-flop. The same state could occur if a separate logic circuit is driving the S and R inputs, and an error causes it to make S and R both low. This is shown in Figure 11-8. Because the output from a NAND gate is always high if at least one of its inputs is low, both gates now have a high output, and the circuit is stable.

The problem is, the states of the outputs from a flip-flop should always be opposite to each other. If both of them are high, this can create logic problems in the rest of the circuit attached to the flip-flop.

- In a NAND-based SR flip-flop, if S and R are both low, this is known as a *forbidden state* or a *restricted combination*.

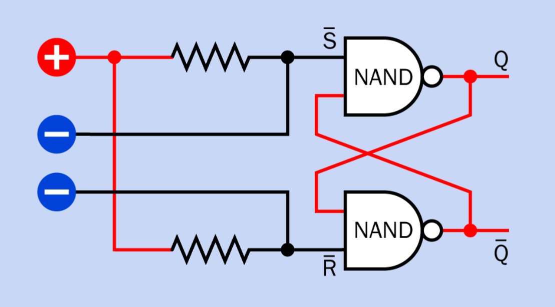

Figure 11-8. *What happens when both S and R inputs to a NAND flip-flop are low as a result of an error in a separate control circuit.*

A similar problem afflicts a NOR-based SR flip-flop, except that the forbidden state will occur when the S and R inputs are both high.

- In a NOR-based SR flip-flop, if S and R are both high, this is a *forbidden state* or a *restricted combination*.

The SR flip-flop is useful as a switch debouncer, but for computing applications, it is vulnerable to errors.

The JK Flip-Flop

Because the name of the JK flip-flop shares the initials of Jack Kilby, who won a Nobel prize for his fabrication of the world's first integrated circuit, some people speculate that this type of flip-flop was named after him. The attribution seems implausible, and may have gained currency simply because a flip-flop was the first device that Kilby happened to build when he was developing an integrated circuit.

Regardless of how it came to be named, the JK design is shown in Figure 11-9. This is commonly referred to as a JK *latch*. The electromechanical switches that were shown driving the SR flip-flop, along with pullup or pulldown resistors, are no longer included, because the inputs at positions J and K are assumed to come from other devices that have properly defined high and low states. Their behavior may be unpredictable, but neither of them will ever have a floating state.

This is a *gated* circuit, meaning that an additional input stage blocks direct access to the output stage, and it is also a *synchronous* circuit, as it uses a train of pulses at a clock input. Two three-input NAND gates are placed in front of a NAND-based SR flip-flop, and they address the problem of simultaneous identical inputs by using crossover feedback from the second stage to the first stage (via the conductors at top and bottom of the schematic).

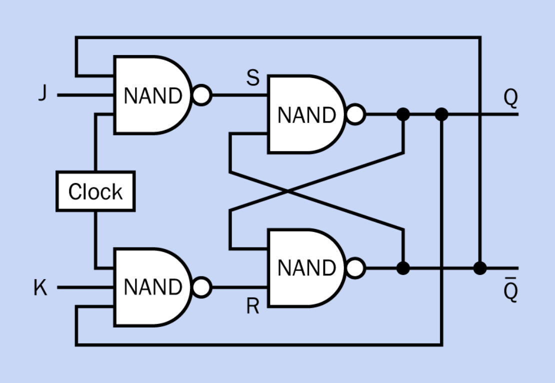

Figure 11-9. *The basic circuit for a clocked JK flip-flop, using two additional NAND gates prepended to an SR flip-flop.*

Versions of a JK flip-flop are possible using NOR gates, but are less common. Only the NAND-based version will be considered here.

In the case of a three-input NAND gate:

- All three inputs high: Output low.
- Other input combinations: Output high.

Because of the additional pair of NANDs, the circuit now recognizes a high input as logically active, instead of the low-active input in the previous SR flip-flop using NAND gates. Consequently, two simultaneous high inputs might be expected to create the type of forbidden state that was caused by two simultaneous low inputs previously. However, in Figure 11-10, the top and bottom panels show that simultaneous high inputs at J and K will support two possible valid outputs, where the state at Q is always opposite to the state at NOT-Q. In fact, when both inputs are high, a positive pulse on the clock input will *toggle* the outputs (i.e., they will switch places). In fact, the toggling will continue so long as the clock input is high. Consequently, this type of flip-flop is intended for use with short clock pulses.

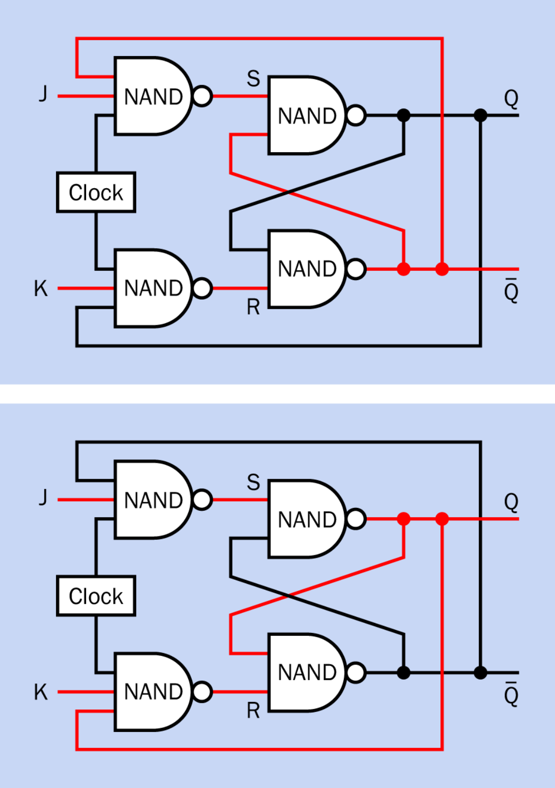

Figure 11-10. *When the J input and the K input are both high, this no longer causes a forbidden state. The combination will toggle the outputs of the flip-flop between the two states shown here.*

Master-Slave Flip-Flop

A more stable form is shown in Figure 11-11 where yet another stage has been added, this one being a "master" to the first. In fact, this configuration is known as a *master-slave* flip-flop, the slave stage being driven by the master stage but remaining inactive until a low clock input at the master stage passes through an *inverter* to become a high clock input at the slave stage. The master and slave stages thus take turns, one being activated by a high clock pulse while the other is activated by the low part of the pulse cycle. The output from the slave stage cannot feed back to the master stage while the clock pulse is still high, and thus the timing issue in the single-stage JK latch is eliminated. Because the master-slave version of the JK configuration is not *transparent*, it is correctly known as a flip-flop rather than a latch.

In addition, Preset and Clear inputs may be added to override the clock to Set or Reset the outputs. These inputs are active-low.

Figure 11-12 summarizes the behavior of a JK master-slave flip-flop that is triggered by the falling edge of each clock pulse (shown as a downward-pointing arrow in the Clock column of the table). Note that the output will be delayed while the slave stage waits for the second part of each clock cycle.

The letter X in the table indicates that the state in that cell is irrelevant.

When J and K are both low, the states of Q and NOT-Q will remain the same as in the previous cycle, and this is still referred to as a *Hold* state. When J and K are both high, the outputs toggle, meaning that their new states will be opposite to the previous states.

D-Type Flip-Flops

A D-type flip-flop places an inverter between two inputs to guarantee that they will always be in opposite states, and uses a clock signal to copy their states to a pair of logic gates.

When an inverter is added between the inputs in this way, either an SR flip-flop or a JK flip-flop can become a D-type flip-flop. Figure 11-13 shows the simplest possible D-type circuit, added to a basic SR flip-flop. Only one data input is now required (customarily labeled D), because it drives the other through the inverter.

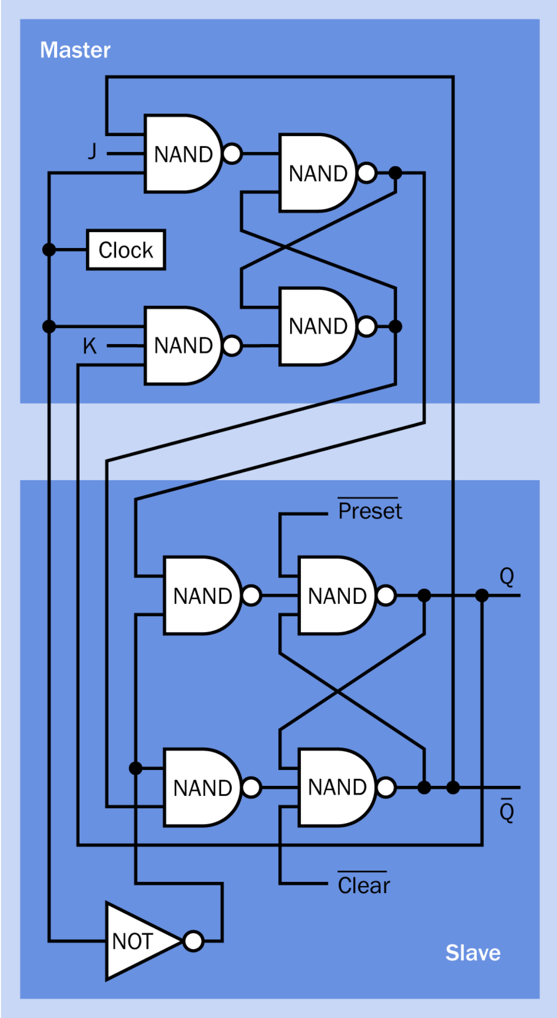

Figure 11-11. *A master-slave circuit that drives one flip-flop with another.*

Flip-Flop Inputs					Outputs	
Clock	Preset	Clear	J	K	Q	$\overline{Q}$
X	●	🔴	X	X	🔴	●
X	🔴	●	X	X	●	🔴
↓	🔴	🔴	●	●	Same as Prev	
↓	🔴	🔴	●	🔴	●	🔴
↓	🔴	🔴	🔴	●	🔴	●
↓	🔴	🔴	🔴	🔴	Toggle	

Figure 11-12. *A table showing inputs and outputs for a JK master-slave flip-flop.*

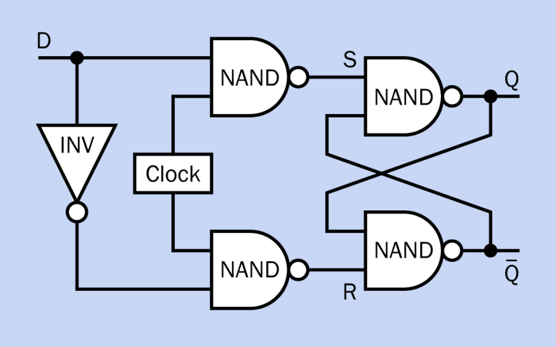

Figure 11-13. *A simple D-type flip-flop. The inverter guarantees that the state of one input will always be opposite to the state of the other.*

Figure 11-14 uses snapshots to show how the circuit responds to changing input and clock conditions. Initially, in the top panel, the data input is high, the clock input is high, and the Q output is high. In the second panel, the clock goes low, causing the output from the upper NAND gate in the input stage to change from low to high. But the upper NAND gate in the output stage still has one low input, so its state remains unchanged. In fact, the S and R inputs of the output NANDs are now both high, which creates the hold condition.

In the third panel, the D input changes from high to low, but this has no effect so long as the clock is low. The D input can fluctuate repeatedly, and nothing will happen until the clock goes high, as shown in the fourth panel. Now the clock copies the new D input state through to the output.

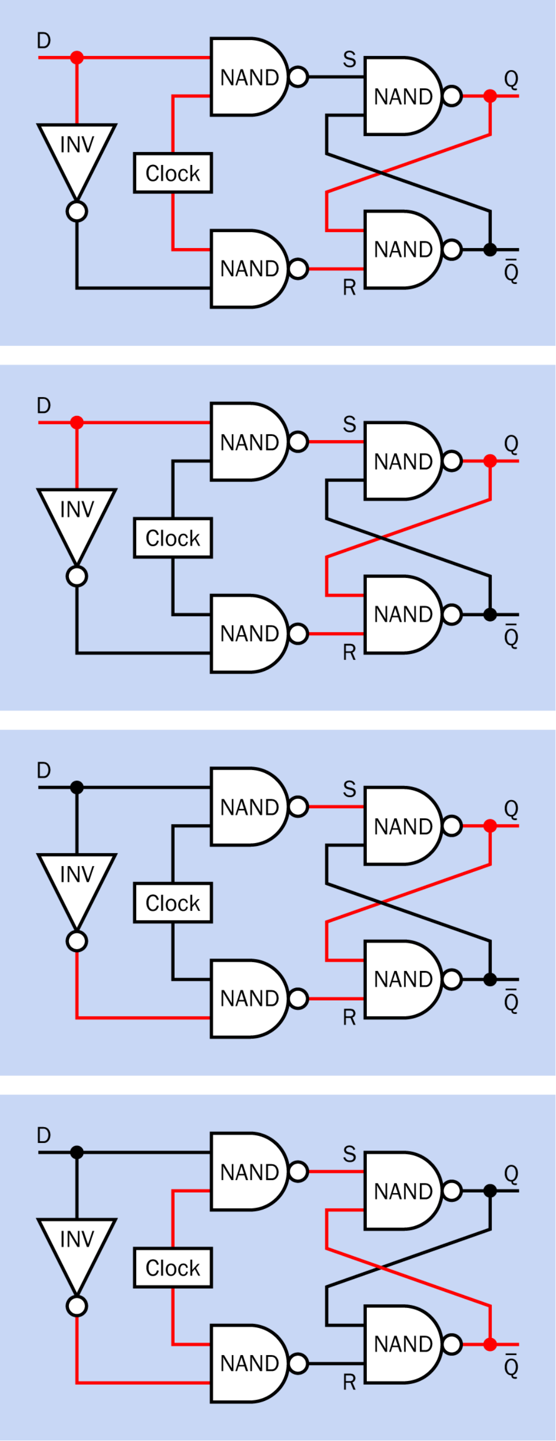

Figure 11-14. *Four snapshots showing the behavior of a D-type flip-flop.*

Summary

- An SR flip-flop can be used for switch debouncing, but in other applications it can enter an unacceptable race condition if its inputs and power supply are not carefully controlled.
- A JK flip-flop is gated, meaning that an SR circuit is preceded with an input stage and a clock input. This eliminates the race condition, adds the ability to toggle the outputs, but requires a very brief clock input. The circuit is edge-triggered.
- A master-slave flip-flop consists of two flip-flops, one driving the other. They can be JK type or SR type. The first flip-flop is activated by a positive clock state, while the second is activated by the subsequent negative clock state. Timing issues are resolved.
- A D-type flip-flop is gated with an inverter between the inputs, so that they cannot be simultaneously high or low. Consequently, only one input, labeled D, is needed. A high state on the D input causes a Set condition, while a low input causes a Reset condition, but only when the clock copies the status of the inputs through to the outputs. The status of the outputs remains stable (the flip-flop enters a hold condition) after the clock goes low.
- The JK circuit used to be widely used, because of its versatility. The D-type circuit now predominates.
- A T-type (toggling) flip-flop exists but is uncommon, and is not included in this encyclopedia.
- Many flip-flop circuits exist in addition to the ones that have been illustrated here. Only the most commonly cited circuits have been included.

A chip containing two positive-edge triggered D-type flip-flops is shown in Figure 11-15. Each flip-flop in this component has its own data, set, and reset input and complementary outputs.

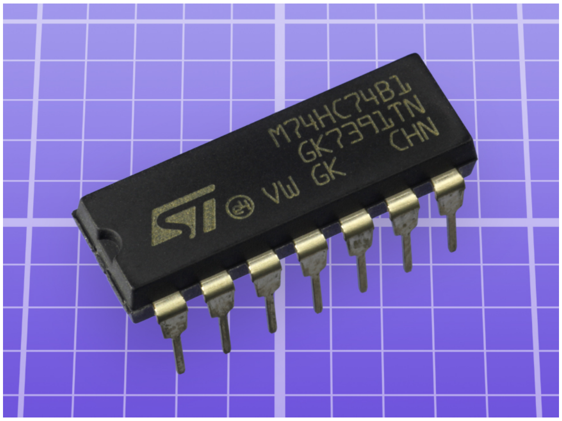

Figure 11-15. *This chip contains two positive-edge triggered D-type flip-flops.*

Variants

A selection of schematic symbols representing flip-flops is shown in Figure 11-16. Letters S, R, J, K, or D define the type of flip-flop. Q and NOT-Q are the outputs. CLK is the clock input but may alternatively be identified with letter E, meaning *Enable*. SRCK or SCLK may also identify it, the abbreviations being intended to mean “serial clock.

A triangle preceding CLK indicates that the flip-flop is positive-edge triggered. A circle, properly termed a *bubble*, preceding the triangle, indicates that the flip-flop is negative-edge triggered. In other locations, the bubble indicates that the input (or output) is inverted; it means the same thing as a bar printed above the text abbreviation, and indicates active-low logic. Synchronous inputs are shown on the left side of the flip-flop with the CLK input, while asynchronous inputs (if any) are shown above and below the flip-flop rectangle.

Using these guidelines, the examples in Figure 11-16 can be decoded:

1. An unclocked SR flip-flop with active-low inputs (probably NAND-based).
2. An unclocked SR flip-flop with active-high inputs (probably NOR-based).

3. An SR flip-flop with active-high inputs, pulse-triggered by an active-high clock input.
4. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input. The bubble on the lower Q output means the same thing as a letter Q with a bar printed above it.
5. A D-type flip-flop pulse-triggered by an active-low clock input.
6. A D-type flip-flop edge-triggered by a falling-edge clock input.
7. A JK flip-flop with active-high inputs, edge-triggered by a rising-edge clock input, with asynchronous active-low Preset and Clear inputs.

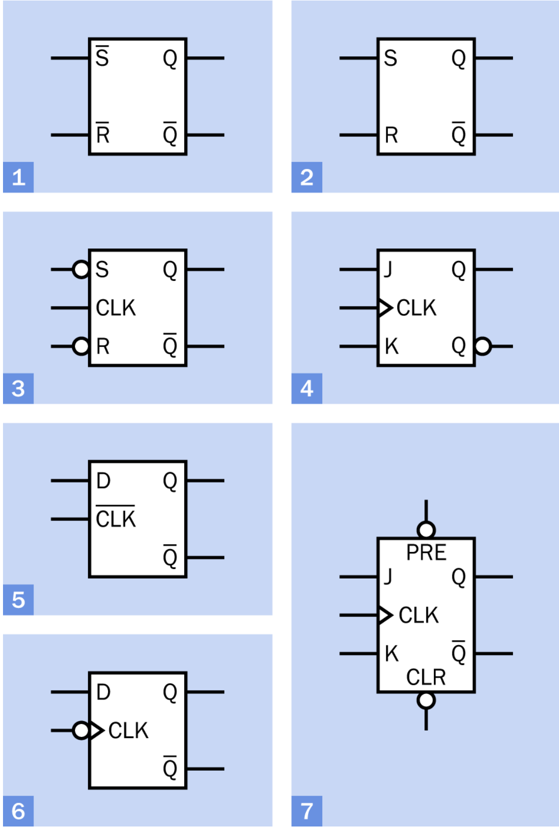

Figure 11-16. *The schematic symbol for a flip-flop is an annotated rectangle. See text for an explanation of the letters and marks.*

Packaging

Only about 10% of the flip-flops listed by a typical parts warehouse are through-hole chips. The rest are now surface-mount. Still, even if the search is narrowed further to through-hole packages in the 74xx and 4000 series, at least 100 types still exist. They provide opportunities in education and prototyping work, even though they are less often used as standalone components.

A package often contains more than one flip-flop. Dual and quad arrays are common. The flip-flops may be independently clocked, or may share a single clock input; datasheets should be checked carefully for details. Octal flip-flops, such as the D-type 74x273, are intended for use as eight-bit registers.

Many of the older flip-flops are numbered in the 74xx series of logic chips. See Chapter 10 for a detailed guide to this numbering system and the various logic families. D-type flip-flops include 74x74, 74x75, 74x174, and 74x175, where an acronym for the logic family is substituted for the x. Old-style CMOS flip-flops include the 4042B D-type latch, the 4043B quad NOR SR flip-flop, and the 4044B quad NAND SR flip-flop. The last two are synchronous, and both allow two Set inputs, labeled S1 and S2 in the datasheet.

Examples of JK flip-flops include the 74x73, 74x76, and 74x109.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See "Variants" on page 93 for a discussion of acceptable high and low logic input voltages. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

If a flip-flop is used for high-speed operation, the following values must be considered:

- t_S Setup time: The minimum time in nanoseconds for an input to be constant before the next clock pulse can process it.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. The interaction between a clock pulse and an input state takes a brief but measurable amount of time; errors may occur if the clock is given less than that amount of time to do its job.
- t_{CO} Clock-to-output: The elapsed time after an active clock edge, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock edge, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL}.
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock edge, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH}.
- f_{MAX} Maximum clock frequency for reliable operation.
- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds.

In a shift register or counter, where multiple flip-flops are cascaded but they share the same clock, the t_{CO} of one flip-flop must be shorter than the hold time of the next flip-flop, to allow the input of data to be complete before the window of opportunity is over.

How to Use It

The asynchronous SR flip-flop is of primary use in debouncing switches. Examples are the single MAX6816, dual MAX6817, and octal MAX6818.

D-type flip-flops are widely incorporated in frequency dividers, which are used to count pulses and display a binary output. If the NOT-Q output is wired back to the D input, the pulse stream to the clock input will have the following effect:

1. Suppose the initial D state is low and the initial state of the NOT-Q output is low.
2. The first high clock pulse propagates the low D state into the flip-flop.
3. The next low clock state forces the NOT-Q output high. This feeds back and creates a high D input.
4. The second clock pulse propagates the high D state into the flip-flop.
5. The next low clock state pulls the NOT-Q output low. This feeds back and creates a low D input.

The sequence then repeats. Only one high output is generated at NOT-Q (or at Q) for every two clock pulses; thus the circuit can become a divide-by-two **counter**. If the Q output is tapped to serve as the clock input for another flip-flop, that circuit now has a divide-by-four output. A series of many flip-flops can be chained together, so long as the propagation of signals along the chain is fast enough to occur before the next clock pulse. This is known as an *asynchronous counter*.

For more information on the use of counters, see Chapter 13.

While flip-flops have tended to be integrated with other components in digital computing, they are still used as registers where 8 or 16 bits of serial data must be assembled at a time, prior to being disseminated as parallel data.

What Can Go Wrong

Ambiguous Documentation

For reasons which are unclear, instructional texts and tutorials can be erratic when describing flip-flops:

- A truth table may fail to clarify whether the circuit uses active-high or active-low logic.
- Truth tables from different sources are often inconsistent in their representation of current and future output states, and may even fail to include the clock status in a clocked flip-flop.
- Tutorials may include logic diagrams for some types of circuit, but not others.
- NOR gates may be used, without any mention that NAND gates can also be used (and may be more common or convenient).
- The active-low or active-high status of inputs in an SR flip-flop may not be shown.

Bearing this in mind, manufacturer datasheets should be consulted whenever possible as the primary source of information.

Faulty Triggering

In many cases, a flip-flop designed for edge triggering can give erroneous results if it is level-triggered, and vice versa. Rising-edge-triggered flip-flops must be distinguished from falling-edge-triggered flip-flops. As always, it is important for similar parts that have similar functions to be stored separately.

Metastability

The behavior of flip-flops has been described in this entry under ideal conditions, where they are operating well within parameters established by the manufacturer. In reality, non-ideal scenarios may occur, especially where inputs such as data and clock, or clock and reset, are almost simultaneous. This may be difficult to avoid if a signal is received from an external source such as a sensor, with no way to control its arrival time. If the input occurs within the setup time or the hold time of a clock pulse, the flip-flop may be unable to determine whether the input precedes or follows the clock.

This may lead to *metastability*, meaning an unpredictable output and/or oscillations that take several clock cycles to settle into a stable state. If the output from a flip-flop may be used by two separate components with slightly different response times, one may interpret the oscillating output as a high state while the other interprets it as low. In a computing circuit, metastability can lead to calculation errors or a system crash. To avoid these issues, limits in datasheets should be observed. Attention should be paid to the manufacturer specifications for minimum setup time and hold time, so that the circuit has sufficient opportunity to recognize a signal and respond.

One solution to metastability is to connect multiple flip-flops in series, all sharing a common clock signal. This will tend to filter out irregularities, at the expense of requiring additional clock cycles if the flip-flops are not transparent.

Metastable-hardened flip-flops minimize metastability but cannot eliminate it completely.

Other Issues

Problems that tend to affect digital chips generally are listed in the section of the entry on logic gates (see "What Can Go Wrong" on page 105).

shift register

12

The term **shift register** is rarely hyphenated. In this encyclopedia, no hyphen is used.

A shift register can function as a *queue*, but this term is more usually applied to software. When the output from the last stage of a shift register is connected back to its input, it can function as a *ring counter*, but that application is described in the **counter** entry of this encyclopedia.

Component catalogs sometimes list shift registers as binary ripple **counters**, instead of giving them their own section. In this encyclopedia, a binary counter is considered to have binary-weighted outputs (with values 1, 2, 4, 8…. in decimal notation) and is described in the **counter** entry. A shift register has outputs that are not necessarily binary-weighted.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)
- **counter** (see Chapter 13)
- **multiplexer** (see Chapter 16)

What It Does

A *register* is a component (or a small section of computer memory) that stores information. The smallest unit of information is one *bit* (i.e., one *binary digit*) with a value 1 or 0 that can be represented by a high or low logic state. A **shift register** most commonly is designed to store eight bits, although some store four.

Each bit is memorized by the status of a **flip-flop** inside the register. For a detailed description of flip-flops, see Chapter 11. When a pulse from an external *clock* is received by the shift register, all of the bits in storage are moved along one step, from each flip-flop to the next. The high or low status of an input pin at that moment is *clocked in* to the first flip-flop, while the bit in the last flip-flop is overwritten by the bit preceding it. A diagram representing the function of a basic four-bit shift register is shown in Figure 12-1.

Note that the status of the input pin is ignored until the moment when a clock pulse copies it into the first flip-flop. In the figure, when the input pin has a brief high state that ends immediately before clock pulse three, the high state is ignored.

A shift-register chip is shown in Figure 12-2.

Because the functionality of a shift register is now often incorporated in much larger logic chips, it is less widely used as a stand-alone component than it used to be. It is still useful for purposes of serial-parallel or parallel-serial conversion, and for small tasks such as scanning a matrix-encoded keyboard or keypad. It also has educational applications and can be used in conjunction with a microcontroller.

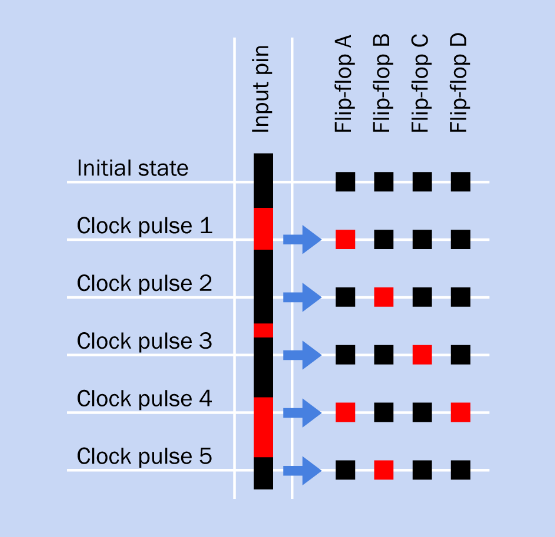
Input pin
Flip-flop A
Flip-flop B
Flip-flop C
Flip-flop D
Initial state
Clock pulse 1
Clock pulse 2
Clock pulse 3
Clock pulse 4
Clock pulse 5

Figure 12-1. *The function of a four-bit shift register in which each flip-flop may be set to a high or a low state, represented here with red or black squares. After a high bit is clocked into the chip, it is moved one space along by each subsequent clock pulse.*

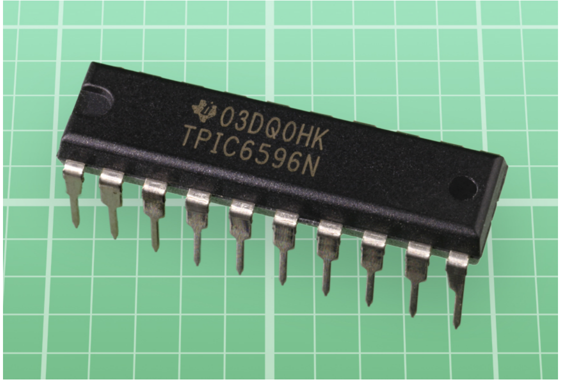
03DQ0HK
TPIC6596N

Figure 12-2. *This 8-bit shift-register chip is unusual in that it uses "power logic," in which open-drain outputs enable it to drive relatively high-current devices. It can sink up to 250mA at each of its output pins, at up to 45VDC.*

Schematic Representation

No specific symbol exists for a shift register. It is represented in a schematic by a simple rectangle, often (but not always) with control inputs on the left, data inputs arrayed along the upper edge, and data outputs along the lower edge. An example is shown in Figure 12-3, along with a diagram showing the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be described in "How It Works" on page 122.

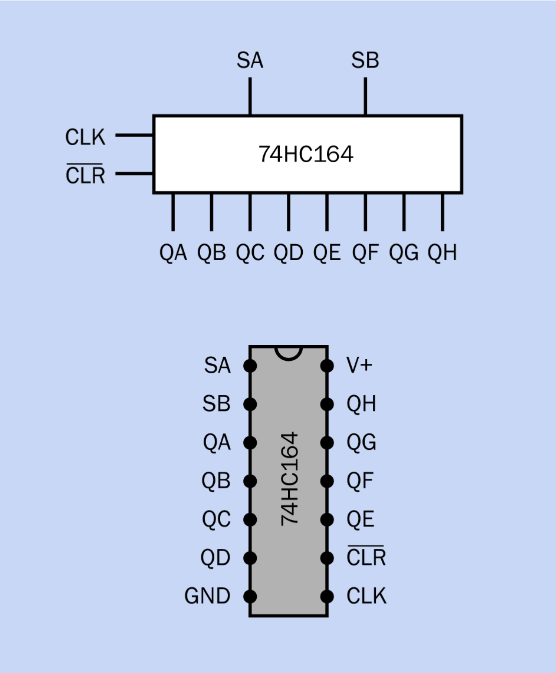
SA
SB
CLK
$\overline{\text{CLR}}$
74HC164
QA QB QC QD QE QF QG QH
SA
SB
QA
QB
QC
QD
GND
V+
QH
QG
QF
QE
$\overline{\text{CLR}}$
CLK
74HC164

Figure 12-3. *Typical schematic representation of a shift register, compared with the pinouts of the actual component.*

The schematic symbol representing a shift register may appear superficially similar to the physical form of the chip which contains it, but the physical layout of the pins is unlikely to be the same.

How It Works

A shift register generally consists of a chain of D-type flip-flops. See the entry describing **flip-flops** in Chapter 11 for a detailed explanation of this component.

The simplest shift register functions as a *serial-in, serial-out* device, abbreviated with the acronym *SISO*. Because the first bit that enters it will be the first to leave at the opposite end, it can also be described as a *first-in, first-out* data storage device, using the acronym *FIFO*.

The basic connections between flip-flops in a four-bit SISO shift-register are shown in Figure 12-4. The D input in each section refers to the fact that it is a D-type flip-flop. The primary output from each flip-flop is identified with letter Q.

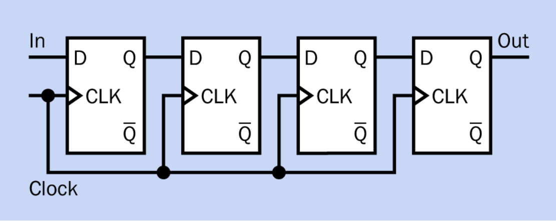

Figure 12-4. *The simplest shift register is a serial-in, serial-out (SISO) device. This example contains four D-type flip-flops.*

Each clock input is labeled CLK. When the output of each flip-flop is coupled to the input of the next flip-flop, and both share the same clock signal, the clock signal will cause the state of the third flip-flop to be sent to the fourth, the output of the second to be copied to the third, the output of the first to the second, and the input state will be copied to the first.

Abbreviations and Acronyms

The shift register will usually have an additional input that forces an immediate "clear" of all the registers, regardless of the clock state at that moment. This input is usually labeled CLR and will have a bar printed above it if it is active-low (which is the usual convention). If there is a pin labeled MR (meaning "Master Reset"), it will have the same function as CLR.

Because its effect is independent of the clock state, the clear signal is described as an *asynchronous* input.

While the abbreviation CLK is frequently used to identify the clock input, SCLK is also used (meaning "serial clock"), and occasionally the abbreviation CP may be found, meaning "clock pulse" input. If the shift register contains two stages, one to clock data in and the second to clock data out, they may be separately clocked, in which case they will be identified with different abbreviations. These are not standardized, but should be explained in the manufacturer's datasheet. No matter which abbreviation is used for a clock input, it will have a bar printed above it if the input is active-low.

Shift registers are generally *edge triggered*, meaning that the rising or falling edge of a clock pulse triggers the bit-shifting operation. If the component responds to a clock transition from low to high, it is *rising-edge triggered*. If it responds to a transition from high to low, it is *falling-edge triggered*, and this may be indicated in the schematic by a small circle, properly known as a *bubble*, preceding the triangle which indicates that this is an edge-triggered device.

Most shift registers are positive-edge triggered.

Parallel Outputs and Inputs

In many shift registers, data may be read out in parallel (from all flip-flops simultaneously), using pins provided for this purpose. In this mode, the shift register can function as a *serial-parallel converter* (*serial in, parallel out*, represented by the acronym *SIPO*). A simplified schematic of the internal connections is shown in Figure 12-5.

Where parallel outputs are provided, they are often identified as QA, QB, QC, and so on (moving from left to right) but may alternatively be described as Q1, Q2, Q3, Q4, and so on.

In a schematic, the input pin is conventionally shown as being at the left end of the component. Often two inputs are provided, connected internally as inputs to a NAND gate. The inputs are likely to be labeled A and B, but may alternatively be named SA and SB, indicating that they are serial inputs. S1 and S2 are alternative classifications. If parallel inputs exist, they may be identified as PA, PB, PC, and so on.

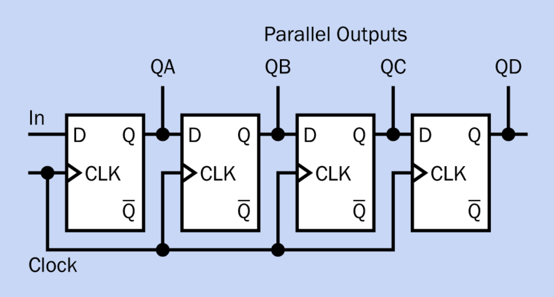

Figure 12-5. *Many shift registers have pins connected to points along the chain of flip-flops. These connections enable data to be read from the shift register in parallel.*

If serial data is supplied asynchronously, as in the illustration in Figure 12-1, it will be ignored until the shift register is triggered by the next clock pulse. The input state at that moment will then be copied into the first flip-flop, while the data that is already being stored in the shift register will be moved along the chain. In datasheets, this is customarily represented by a diagram such as the one in Figure 12-6. This diagram assumes that the shift-register is rising-edge triggered. Note that a brief fluctuation in the input which does not coincide with a clock-trigger event will be ignored.

Variants

Serial In, Serial Out

A basic SISO shift register allows only for serial input (at one end of the chain of flip-flops) and serial output (at the other end of the chain). No pins are available for parallel output of data.

This type of component usually permits 64-bit storage, where parallel output is simply impractical, as too many pins would be required. An example is the 4031B chip. This includes provision for recirculation of bits, so that it will also function as a ring counter (see Chapter 13 for a discussion of this function). As is always the case with logic chips, the part number will be preceded by letter(s) identifying the manufacturer, and a suffix will distinguish variants of the chip.

Another type of SISO shift register is programmable. It will store any number of bits from 1 through 64, determined by a binary number applied in the form of high/low states to five pins reserved for this purpose. An example is the 4557B.

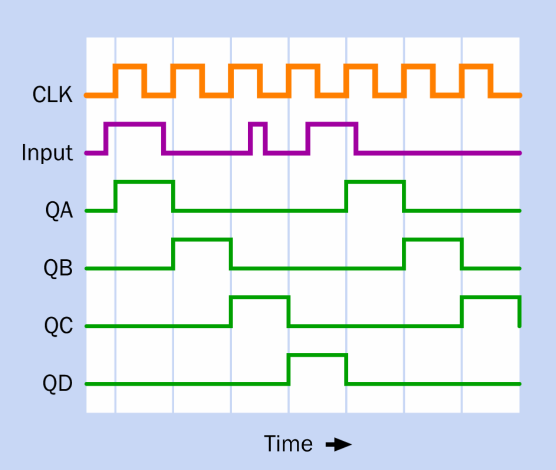

Figure 12-6. *In a rising-edge-triggered shift register, the high or low state of an asynchronous input (purple line) is copied into the first flip-flop of a shift register by each clock pulse (orange line). Brief fluctuations that do not coincide with a rising clock pulse are ignored. Existing data in the register is shifted from one flip-flop to the next.*

Serial In, Parallel Out

The majority of serial-input shift registers allow parallel output from points along the chain in addition to serial output at the end of the chain. These chips almost all are 8-bit registers. Typically two inputs are provided, one of which can be used to receive bits that recirculate from the end of the chain, back to the beginning. Widely used examples are the 4094B and the 74x164, where an acronym identifying the logic family will be substituted for the x.

Parallel In, Serial Out

A minority of shift registers are able to function as *parallel-serial converters* (*parallel in, serial out*, represented by the acronym *PISO*). Typically this type of chip allows *jam-type* parallel data input, meaning that the data is forced into the flip-flops via a separate pin for each of them. Parallel input is enabled by the status of a serial/parallel control pin. When the control pin reverts to its opposite status, each clock pulse will now shift data along the chain of flip-flops, allowing it to be read from the final output one bit at a time. Thus, data can be entered into the chip in parallel and read out of it serially. Examples are the 4014B and 4021B. Both are 8-bit.

Parallel In, Parallel Out

Shift registers that permit parallel output in addition to parallel input are almost all of the *universal* type, described in the next section.

Universal

A *universal* shift register is capable of all four modes of operation: SISO, SIPO, PISO, and PIPO. The four modes of the component are selected by the high or low status of two *mode select* pins. In addition, this component may have the ability to shift the register states either left or right. A *bidirectional* shift register has this same capability, and may also have PISO and PIPO capability, depending on the chip. Examples are the 74x195 and 74x299, where an acronym identifying the logic family will be substituted for the x in the number.

Universal shift registers are almost all 4-bit or 8-bit. They often have relatively complicated features, such as access to internal JK flip-flops, or pins that are multiplexed to provide different functionality depending whether an enable pin is held high or low. Datasheets must be checked carefully to ensure correct use.

Dedicated shift registers of SIPO or PISO type will be easier to use.

Values

As is the case with other logic chips, most flip-flops in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

If a shift register is used for high-speed operation, the following values must be considered (identical notation, and similar values, are found in specifications for flip-flops):

- t_S Setup time: The input state of a shift register must exist for a very brief period before the clock trigger that processes it. This period is known as the *setup time*. In the 4000 series of chips, recommended setup may be as long as 120ns. The value will be much lower in 74xx chips.
- t_H Hold time: The minimum time in nanoseconds for an input to persist after the active edge of a clock pulse that has processed it. In many shift registers, no hold time is necessary, as the chip has already been activated by the rising edge of the clock pulse.
- t_{CO} Clock-to-output: The elapsed time after a clock trigger, before the output changes. This is a function of the internal workings of the chip, and may be broken down into low-to-high and high-to-low output transitions, as follows.
- T_{PLH} Propagation to Low-to-High: The elapsed time after an active clock trigger, before a low-to-high swing occurs at an output. This may not be identical to T_{PHL}.
- T_{PHL} Propagation to High-to-Low: The elapsed time after an active clock trigger, before a high-to-low swing occurs at an output. This may not be identical to T_{PLH}.
- f_{MAX} Maximum clock frequency for reliable operation. In the older design of 4000 series chips, 3MHz may be recommended with a power supply of 5VDC. Higher frequencies are possible with a higher voltage power

supply. Frequencies as high as 20MHz are possible in the 5VDC 74HC00 series.

- $t_{W(H)}$ The minimum high clock pulse width in nanoseconds. In the older design of 4000 series chips, 180ns may be recommended with a power supply of 5VDC. Shorter pulses are possible with higher voltage power supply. Pulses as short as 20ns are possible in the 5VDC 74HC00 series.
- $t_{W(L)}$ The minimum low clock pulse width in nanoseconds. This is likely to be the same as $t_{W(H)}$.

Power Considerations

Shift registers conform with the usual power-supply requirements for logic families. These are described in detail in the **logic gate** entry in Chapter 10. Likewise, the ability of a shift register to source or sink current is usually determined by its logic family. In a few cases, however, shift registers have an additional *open-drain* output stage for each register, capable of sinking currents as high as 250mA. The Texas Instruments TPIC6596 shown in Figure 12-2 is an example. When an open-drain output is connected with a logic device whose input cannot be allowed to float indeterminately, a *pullup resistor* must be added.

Three-State Output

A chip may have a *three-state output* (also known as *tri-state output*, a term which was trademarked but is now used generically). This means it will be capable of changing its outputs so that instead of sourcing or sinking current in a logical high or low state, they can have a high impedance. The chip then becomes “invisible” to others which may be sharing the same output bus. The high-impedance state is usually applied to all outputs from the shift register simultaneously when enabled by a separate pin, often identified as *OE*, meaning *output-enable*. Examples of three-state shift registers are the 74x595 or the 4094B chip.

The high-impedance state can be thought of as being almost equivalent to switching the outputs of the shift register out of the circuit. Consequently, if other components sharing the bus are also in high-impedance output mode, the state of the bus will “float,” with uncertain results. To avoid this, a pullup resistor of 10K to 100K can be used on each bus-line.

Where the internal components of a shift register are shown in a datasheet, a three-state output is usually represented with a buffer symbol or inverter symbol that has an additional control input located on its upper edge, as shown in Figure 12-7. This should not be confused with the similar placement of a positive power supply input to an amplifier or op-amp. (Schematics showing the interior elements of a logic chip almost never include power-supply connections.)

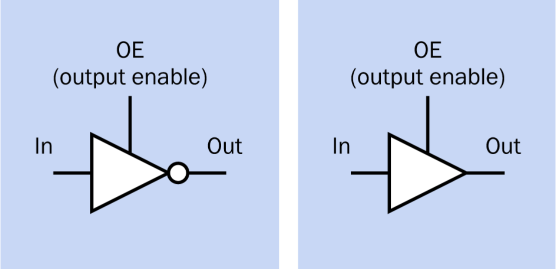

Figure 12-7. *A shift register may be capable of a three-state output, where high impedance is an option additional to the usual high or low logic state. An output enable pin allows this. It is typically shown as an additional input to an inverter (left) or buffer (right) inside the shift-register chip.*

How to Use It

The SISO application of a shift register can be used simply to delay the transmission of data by storing it and moving it from one flip-flop to the next before it is read out of the end of the chain.

The SIPO application of a shift register (serial in, parallel out) may be useful where a microcontroller has insufficient outputs to control multiple devices. Serial data can be sent from a single microcontroller output to the input of a shift register. The chip can then drive a separate device from each of its parallel output pins. If there are eight devices, the microcontroller can send a sequence of eight bits, each of which will control the on/off status of one device when the bits are read out of the shift register in parallel. If more devices are used, an additional shift-register can be daisy-chained to the output from the first.

Clock signals can be supplied from the microcontroller, along with a signal to the clear input of the shift register, if desired. Alternatively, the old bit states in the shift register can simply be "clocked out" and replaced with a new set of serial data. During the process of "bit banging," the parallel outputs of the shift register can remain connected directly with the output devices if the clock speed from the microcontroller is fast enough, as devices such as relays will not respond to extremely brief pulses.

For generic shift registers that do not have open-drain outputs, a buffer will be needed to provide sufficient current for any device using more current than an LED.

If a shift register is configured for PISO mode (parallel in, serial out) it can be placed on the input side of a microcontroller, polling a variety of devices and feeding their states into the microcontroller serially.

Dual Inputs

Where a shift register has two serial inputs (as is often the case), they are almost always linked as inputs to an internal NAND gate. This allows the output from the end of the chain of flip-flops to be connected back to the beginning of the chain, if the shift register is to function as a ring counter. However, if this function is not used and a single input is required, the two inputs to the shift register can be tied together. In this configuration, the input becomes active-low. The two possible configurations are shown in Figure 12-8. It is important never to leave an input *floating*, or unconnected.

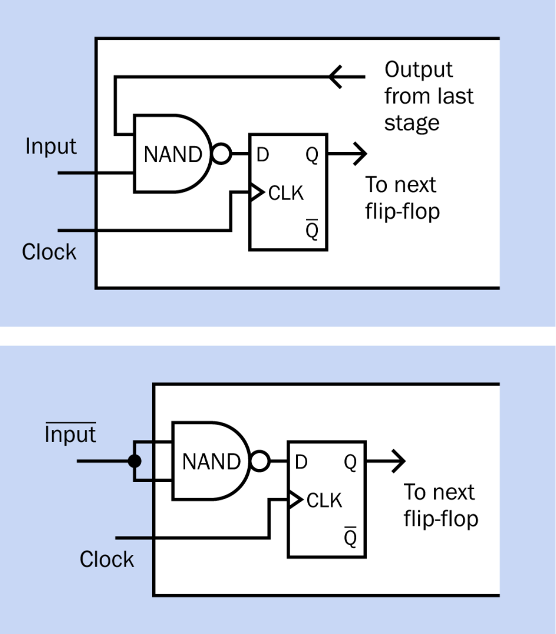

Figure 12-8. *Two possible configurations where a shift register allows two inputs linked with an internal NAND gate.*

Preloading the Shift Register

Where a shift register will be used to output a single recirculating bit (or in similar applications), the component must be preloaded with a high state in its first register. This may require a one-shot (monostable) timer which is activated only when the circuit is powered up.

Polling a Keyboard

Two shift registers can be used to scan the data lines in a matrix-encoded keyboard or keypad. This is often known as *polling* the keyboard. Provided the scan rate is sufficiently fast, the user experiences a seemingly immediate response to a key-press.

While the full schematic is too large and complex to be included here, many examples can be found online.

Arithmetical Operations

Shift registers were traditionally used to perform arithmetical operations on binary numbers. If the number is represented by eight bits (i.e., one byte) with the most significant digit on the left, shifting the bits one space to the right will have the effect of dividing the byte value by 2. If the bits are shifted one space to the left, the byte value will be multiplied by 2 (provided an additional register is available to store the most significant bit after it has been shifted). This concept is illustrated in Figure 12-9.

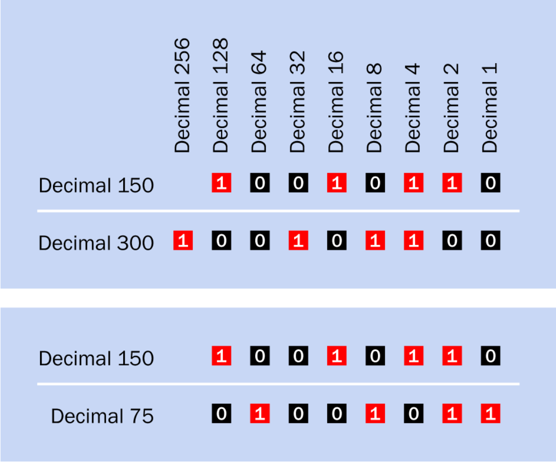
Decimal 256
Decimal 128
Decimal 64
Decimal 32
Decimal 16
Decimal 8
Decimal 4
Decimal 2
Decimal 1
Decimal 150 1 0 0 1 0 1 1 0
Decimal 300 1 0 0 1 0 1 1 0 0
Decimal 150 1 0 0 1 0 1 1 0
Decimal 75 0 1 0 0 1 0 1 1

Figure 12-9. *In the upper section of this diagram, a binary number represented by eight bits in a shift register is multiplied by 2 by shifting all the bits one space to the left. In the lower section, the same binary number is divided by 2 by shifting all the bits one space to the right. The binary values are shown in decimal notation to the left.*

In the upper section of the figure, the binary number 10010110 (chosen arbitrarily) is represented in the eight flip-flops of a shift register. The decimal place value of each digit is indicated. Adding up the place values of the 1s in the number, the total is 128 + 16 + 4 + 2 = 150. Below the white line, the digits have been shifted one place to the left, with the leftmost digit carried over to an additional location, while a zero is inserted in the rightmost location. Assuming that the additional place at the leftmost location has a place value of 256, the total is now 256 + 32 + 8 + 4 = 300.

In the lower section of the figure, the same binary number has been shifted one space to the right, with a 0 introduced in the leftmost location. The decimal value is now 64 + 8 + 2 + 1 = 75.

While this application for shift registers was common during development of digital computing in the 1960s and 1970s, the shift register as a separate component became less common subsequently, as its functionality was acquired by modern CPU chips.

Buffering

A shift register may also be used as a *buffer* between two circuits where the clock speeds are different. Digits are clocked in by the first circuit, then clocked out by the second. Some shift registers allow two clock inputs and can be used for this purpose.

What Can Go Wrong

Problems that tend to affect digital chips generally are listed in the entry on logic gates (see "What Can Go Wrong" on page 105).

Confusing Classification

Because of the functional similarity to a binary ripple **counter**, a shift register is sometimes listed by component suppliers as a counter. In fact, a binary counter will almost always have outputs that have place values 1, 2, 4, 8.... and upward, while the outputs from a shift register will not have place values.

When searching for a shift register, it can be found by specifying a "counting sequence" of serial to parallel, serial to serial, parallel to serial, or parallel to parallel. If the "counting sequence" is simply up or down, the component is a counter, not a shift register.

Inadequate Setup Time

Each flip-flop in a shift register must have a stable input state before the next triggering event shifts the data. If this setup time is reduced below the minimum specified in the datasheet, results will be unpredictable.

Unconnected Input

Because many shift registers have a choice of two inputs to the same chain of internal flip-flops, it is easy to leave one of them unconnected by accident. A floating input will be vulnerable to any stray electromagnetic fields, and is almost certain to create random effects.

Output Enable Issues

The output enable pin on a shift register that has three-state logic outputs is usually active-low. Consequently, if the pin is left unconnected, the logic outputs may go into high-impedance mode, or will fluctuate unpredictably. Where three-state outputs are not required, a safe course of action will be to avoid using three-state chips.

Floating Output Bus

If a pullup resistor is omitted from a bus that is shared by three-state chips, the results will be unpredictable. Even if the circuit design seems to guarantee that at least one chip will have a high or low output on the bus, a pullup resistor should still be included.

counter

13

The term **counter** is used here to mean a digital-logic chip. A counter could be built from discrete transistors, but this approach is obsolete. Counters may also be devised from parts such as multiple relays, or a solenoid advancing a ratchet wheel, but such electromechanical devices are not included in this encyclopedia.

In this encyclopedia, a counter by definition has binary-weighted outputs with values 1, 2, 4, 8…. in decimal notation. The exception to this rule is a *ring counter*, which does not have binary-weighted outputs but is included here because its name identifies it as a counter. A **shift register** may be used as a ring counter, but is more versatile and has many other functions; hence it has a separate entry.

Gray code counters, in which successive outputs differ by only one binary digit, are not described in this encyclopedia.

OTHER RELATED COMPONENTS

- **flip-flop** (see Chapter 11)
- **shift register** (see Chapter 12)

What It Does

A counter can be used to count events, or can measure time in convenient intervals if it is connected with a component such as a *quartz crystal* that operates at a precise and reliable frequency. The counter receives input pulses (usually referred to as a *clock input*) and counts a predetermined number of them before restarting from the beginning. It will repeat in this fashion so long as power is connected, and the clock pulses continue, and a reset signal is not supplied.

Almost all counters create some form of output during the count. Most commonly, the output is a pattern of high and low states expressing the number of clock pulses in binary code. Where a counter counts to a very high number before recycling, some intermediate binary digits may be omitted.

While standalone counter chips are not as widely used now as in the early days of computing, they still find application in industrial processes, small devices, and education, and can be used to control incremental devices such as stepper motors. They can be used in conjunction with microcontrollers.

Schematic Representation

No specific logic symbol exists for a counter. It is most often shown in a schematic as a rectangle with clock input(s) and clear input(s) on the left and outputs on the right. An example appears in Figure 13-1, above a representation of the physical chip and its pinouts. The meaning of the abbreviations identifying the inputs, outputs, and control functions will be found in "How It Works" on page 132. Because the two MR inputs for this particular counter are ANDed inside the chip, the AND symbol is included with the counter symbol.

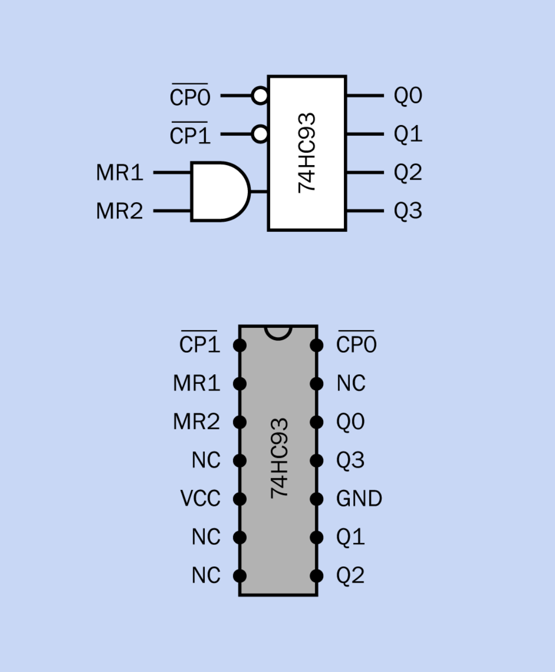

Figure 13-1. *Typical schematic representation of a counter, compared with the pinouts of the actual component.*

A counter chip is shown in Figure 13-2.

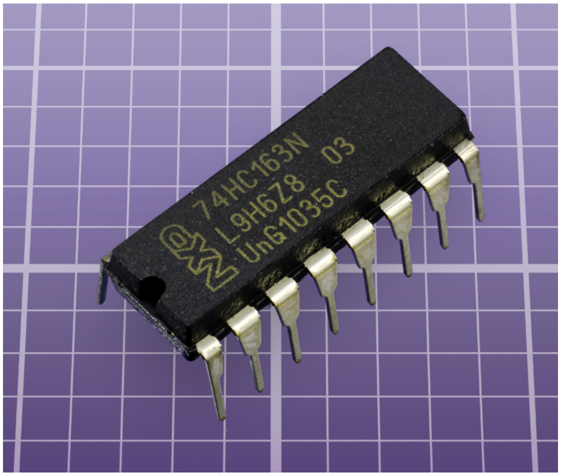

Figure 13-2. *The 74HC163 shown in this photograph is a 4-bit synchronous counter capable of being preloaded with a starting value, and able to do a synchronous reset.*

How It Works

A counter is built from a chain of **flip-flops**, with each one triggering the next. JK, T-type, or D-type flip-flops may be used. For a thorough description of a flip-flop, see Chapter 11. In Figure 13-3, a D-type flip-flop is shown, triggered by each rising clock pulse.

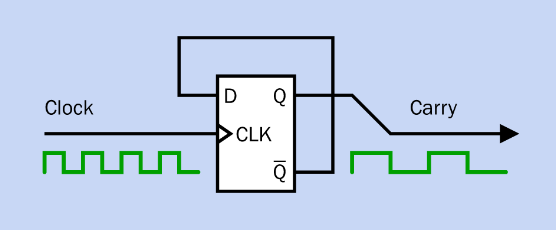

Figure 13-3. *When the complementary (NOT-Q) output from a D-type flip-flop is connected back to its input, the Q output frequency is half of the clock input frequency. See text for additional details.*

Initially the Q output of the flip-flop is low, so its NOT-Q output (identified by Q with a bar above it) is high. This feeds back to the D input, but has no effect until the rising edge of the next clock pulse copies the high D input to the Q output. The Q output is now latched high while the NOT-Q output is latched low and feeds back to the D input. The triggering event has passed, so the low D input does not have any immediate effect. The rising edge of the next clock pulse copies the low state of the D input to the output, and changes the NOT-Q output to high, causing the cycle to repeat. As a result, the output frequency of the flip-flop is one-half of the input frequency from the clock. If the output is carried to the next flip-flop to become its clock input, once again the frequency will be divided by two.

Modulus and Modulo

The *modulus* of a counter is the number it will count up to, before it repeats. This is sometimes written incorrectly as *modulo*.

In fact, *modulo* is the name of an arithmetical operation, often abbreviated as *MOD* (usually capitalized, even though it is not an acronym). This operation consists of a division sum in which the *remainder* from the division is the result of the operation. Thus 100 modulo 5 gives a result of 0, because there is no remainder when 100 is divided by 5. But 100 modulo 7 gives a result of 2, because 2 is the remainder of the division operation.

To exacerbate the confusion, *MOD* is also used as an adjective referring to the modulus of a counter. Thus a MOD-4 counter has a modulus of 4, and a MOD-16 counter has a modulus of 16. As a general rule, when a counter is being described, *modulo* and *MOD* will mean the same thing as *modulus*. This may be confusing to people such as computer programmers who are already familiar with the correct usage of MOD as an arithmetical operator.

In a counter, to obtain a modulus that is not a power of two, logic gates inside the chip can intercept a particular value (such as 1010 binary, which is 10 decimal) and use this as a signal to restart the count at zero. External connections to the chip can achieve the same purpose.

Pin Identifiers

Abbreviations and acronyms are used in datasheets to identify pin functions. These identifiers have not been standardized, and many variants exist.

CLK is the abbreviation most commonly used for the clock input, sometimes alternatively shown as CK or CP. If it is active-low, or if its falling edge will be active, a bar will be printed above it. Where a printed font does not permit an underscore to be placed in this way, CLK' (the abbreviation followed by an apostrophe) may be used instead. Often two or more stages of a counter can be separately clocked, and the input pins will be identified with abbreviations such as CLK1 and CLK2, or 1CLK and 2CLK, or CKA and CKB, or CP1 and CP2, or similar.

Where a clock input is edge-triggered, this is indicated by a small triangle. The triangle can be seen in Figure 13-3.

CLR designates a pin which will clear the count and reset it back to zero. This signal is often active-low, indicated by a bar printed above the abbreviation.

In a schematic, a small circle, properly known as a *bubble*, may be placed at any input which is active-low. On a clock input, the circle indicates that it is falling-edge triggered. See Figure 11-16 for various implementations of symbols with flip-flop schematics.

The CLR operation may be *synchronous* (the pin state will not be recognized until the next clock pulse) or *asynchronous* (in which case the pin state overrides the clock and resets the counter immediately). MR stands for "Master Reset," and has the same function as CLR.

Where two or more counters (or multiple stages within one counter) can be reset separately, more than one clear input will be shown, and may be identified with abbreviations such as CLR1 and CLR2, or MR1 and MR2.

Output pins will almost always be identified as Q0, Q1, Q2 … or QA, QB, QC … up to the maximum necessary to express the modulus. If two or more counters are included in one chip, outputs may be prefixed with an appropriate number. Thus 2Q3 would be the third output in the second counter. Multiple counters in one chip are identified with numbers starting from 1.

Where internal flip-flops are shown, they will have identifiers such as FF1 or FF2. Each flip-flop will have its own clear function, identified as C or CD (the latter meaning "clear data"), and may have inputs labeled D1, D2, D3… in a D-type flip-flop or J and K in a JK flip-flop. See Chapter 11 for an explanation of flip-flop inputs and outputs.

The input to a counter is always imagined to begin from the left, and consequently the leftmost flip-flop shown in an internal schematic will express the least significant bit of the current value of the count, even though a binary number is written with the least significant bit in the rightmost place.

If a counter is capable of receiving parallel data as an input (explained below) it will have a pin labeled PE for parallel enable. It may also have a CE or CET pin, for count enable.

As is generally the case in logic chips, VCC or V+ are typically used to identify the positive power supply pin, while GND or V- will identify the negative-ground pin. NC means that a pin has no internal connection at all, and consequently requires no external connection.

Variants

All counter chips use binary code internally, and the number of *bits* (*binary digits*) in the counter's modulus will be the same as the number of internal flip-flops. A 4-bit counter (the usual minimum) will have a modulus of 2^4 which is 16. A 21-bit counter (the maximum typically available) will have a modulus of 2^{21} which is 2,097,152. For higher moduli, counters can be chained together, each sending a carry signal to the next. This is known as a *cascade*.

Multiple counters, with different moduli, may be chained in a single chip. For example, in a digital clock that displays hours and minutes using the 60Hz frequency of an American domestic power supply as its timebase, the initial counting stage will have a modulus of 60, to count individual seconds. The next counting stages will have moduli of 60, 10, and 6, so that they can count from 00 to 59 minutes. Additional stages in the chip will tally hours.

A counter with a *parallel input* can be preloaded with an initial value (in binary code) from which it may count up or down. A *parallel-enable* pin may put the counter into a mode where the number can be *jam loaded*, meaning that it is jammed into the counter regardless of the clock state. Other types of counter are loaded synchronously.

Ripple versus Synchronous

In a *ripple counter* each internal flip-flop triggers the clock input of the next, so that their states change in a rapid but incremental sequence from the input to the output. This is also known as an *asynchronous* counter. Because the final state will not be valid until the clock pulse has rippled all the way through the counter (and through additional counters if they are cascaded together), a ripple counter will tend to suffer from a *propagation delay* of up to a microsecond. Ripple counters may also create output spikes or momentary transient count values that are invalid. Therefore, they are more suitable for applications such as driving a numeric display than for interfacing at high speed with other logic chips.

In a *synchronous counter*, all the flip-flops are clocked simultaneously. A synchronous counter is better suited to operation at high speed.

Of the counter chips available today, about half are synchronous and half are asynchronous.

Ring, Binary, and BCD

A counter that activates output pins one at a time sequentially is said to have a *decoded output*. It is often referred to as a *ring counter*. It has the same number of output pins as its modulus. An example is the 4017B chip.

A *binary counter* is more common and has an *encoded output*, meaning that it will express the running total of the count in binary code through *weighted outputs* that typically have (decimal) values of 1, 2, 4, 8, and so on. A modulus-8 counter (often referred to as an *octal counter*) will require three outputs which represent the binary numbers 000, 001, 010, 011, 100, 101, 110, and 111 (decimal 0 through 7) before going back to 000 and repeating.

A *modulus-16* counter, also known as a *hexadecimal counter* or a *divide-by-16 counter*, will have a binary output represented by four output pins, counting from 0000 through 1111 (decimal 0 through 15). Four-digit binary counters are very common, and their outputs are compatible with other components such as a **decoder**, which converts a binary-number input into a ring-counter-style output.

A *decade counter* is a modulus-10 binary counter. It is described as having a *binary-coded decimal* output (often expressed with the acronym *BCD*), using four weighted output pins which represent the numbers 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, and 1001 (decimal 0 through 9) before repeating. Because this counter skips binary outputs from 1010 through 1111 (decimal 10 through 15), it is said to have a *shortened modulus*.

Figure 13-4 shows a schematic diagram of JK flip-flops in a decade ripple counter. The J and K inputs are all tied to the positive power supply, as this causes the clock input to toggle the output high and low. Note that because the primary input is always shown at the left end of the component, the least significant output bit (Q0) is in the leftmost position.

To intercept binary 1010 (decimal 10), an internal NAND gate is used. Its output goes low when its two inputs, from Q1 and Q3, go high. The output from the NAND immediately activates the CLR function on all the flip-flops, so that as soon as the decade counter reaches 1010 (decimal 10), it resets itself to 0.

In this particular chip, the preload for each flip-flop is tied to the positive power supply, so that it is always inactive. In some counters, the preload feature of each flip-flop is accessible via pins outside the chip. This creates the potential hazard of preloading the counter with one of the numbers that it normally skips (for instance, 11 decimal in a decade counter). This is referred to as an *invalid number* or *disallowed state*. (This use of the term "state" refers only to the binary number stored in the counter's flip-flops. It has nothing to do with the high-state or low-state voltages used to represent binary 0 or 1.)

The counter's datasheet should include a *state diagram* showing how the counter will deal with this situation. It may reset itself to a valid value after a maximum of two steps, but this can still cause confusion, depending on the application.

The state diagram for a 74HC192 counter is shown in Figure 13-5.

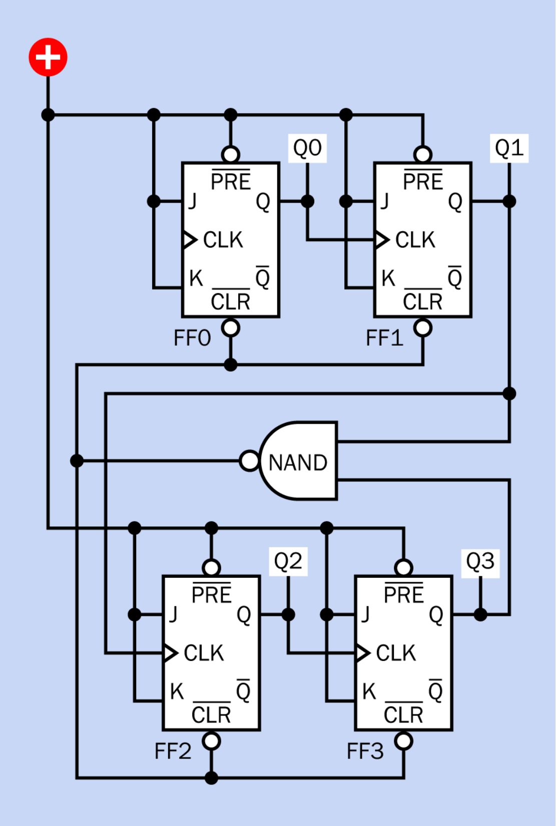

Figure 13-4. *The internal logic of a synchronous decade counter that uses JK flip-flops.*

Clock Sources

The clock input may be provided by a **timer** chip or by an *RC network*, which has the advantage of being able to run at a relatively low speed for purposes where this is desirable. It may alternatively be provided by a *quartz crystal* oscillating at a much higher frequency such as 1MHz. Successive counters may be necessary to reduce this value, depending on the application.

In some counters, the clock is built into the chip. More commonly, a resistor and capacitor may be used externally to establish a clock rate in conjunction with logic gates inside the chip. The datasheet for this type of component will include a formula for calculating the clock frequency from the resistor and capacitor values. The 4060B chip is an example.

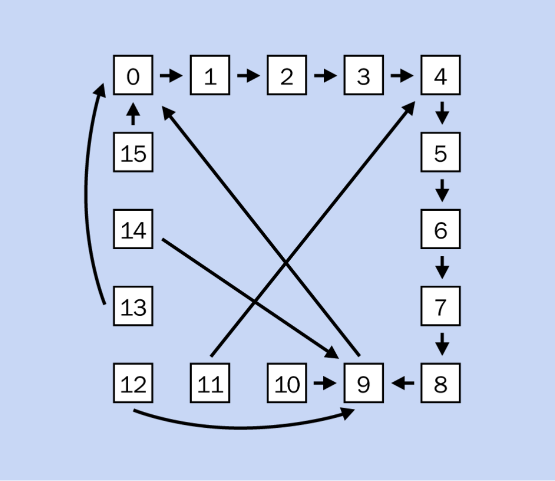

Figure 13-5. *A state diagram shows the transitions that a counter will make from each number to the next (in decimal notation), including the transitions which enable it to exit from disallowed states. This example is for a 74HC192 chip.*

Rising Edge and Falling Edge

A counter may be designed to be triggered either by the rising edge or the falling edge of the clock input, or by its high or low logic state. Generally speaking, ripple counters use the falling edge, so that the final output from one counter can become the clock input of the next. In other words, when the most significant digit of the first counter changes from a high to low logic state, this transition toggles the least significant bit of the second counter.

Synchronous counters generally use the rising edge of the clock input. If multiple synchronous counters are cascaded, they must all share the same clock signal, and will all change their flip-flop states simultaneously.

Multiple Stages

It is common for a counter chip to contain two or more *stages* with differing moduli. To take a common example, a divide-by-2 stage and a divide-by-5 stage that are both present in a single chip can be used to create a decade counter by connecting external pins. The extra stages provide a choice of moduli if they are used individually.

Single and Dual

Counter chips may contain two counters of the same modulus. This is known as a *dual* counter. Dual 4-bit counter chips are common. Each counter can be used separately, or they can be cascaded, in which case the total modulus will be found by multiplying the individual moduli together.

High-State, Low-State, and Three-State

Almost all counters use positive logic where a 1 is represented by a high state and 0 by a low state. Some counters allow an additional output state which has a high impedance and is equivalent to an open circuit. This feature is useful when two or more chips share the same output bus. It is discussed in the entry for **shift registers** in "Three-State Output" on page 126.

Descending Output

Most components only create an ascending count. The output can be converted to a descending count by passing each binary state through an inverter, but this will only work properly if the modulus is equal to the number of states. In a BCD counter, its inverted outputs will count from decimal 15 to decimal 6, not from decimal 9 to decimal 0.

A few counters are available which are designed to create a correct descending count. Other counters are available which allow the user to set the mode to ascending or descending. Examples are the 74x190 or 74x192 (where an acronym for the chip family will be substituted for the letter x).

A descending output is useful in combination with a parallel input, where a user may set an initial value from which the counter will descend to 0. With suitable logic, this can enable a user-specified delay period.

Programmable Counters

A programmable counter can usually allow a modulus ranging from 2 to more than 10,000. The counter counts down by dividing an initial number repeatedly with a value that is preset with binary inputs. An example is the 4059B chip.

Examples

Many counter specifications date back to the 4000 family of logic chips. Versions of them subsequently became available in the 74xx series, often with the old 4000 part number preceded by 74x (where x is replaced by a designation of the logic family). For example, a version of the 4518B dual BCD chip can be obtained as the 74HC4518. As is the case with all logic chips, this part number will be preceded by letter(s) designating the particular manufacturer, with a suffix added to differentiate minor variants of the chip. The 74xx series has the advantage of higher speed and a greater ability to source or sink current at its output pins.

Most of the original CMOS chips, such as the 4518B, are still available, even in surface-mount versions. These offer the possible advantage of being able to use a higher power supply voltage.

Many counters offer multiple options such as different modulus values that can be selected by external pin connections. Some chips are tolerant of slow clock frequencies; others are not. Most are edge-triggered, but a few are level triggered. Some, such as the 4518B mentioned above, allow a choice of a rising-edge clock input and falling-edge clock input on different pins. For a specific application, it is really necessary to read a variety of datasheets to select the chip that is most suitable.

Values

As is the case with other logic chips, most counters in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount 74xx versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low logic-input states. On the output side, the 4000 series chips are able to source or sink less than 1mA at 5VDC, but the 74HCxx series can usually manage around 20mA.

A few counters are capable of delivering more power through additional output stages that can drive LEDs. The 4026B decade counter is still being manufactured, capable of powering modest 7-segment displays. The 4033B has the additional option of blanking any leading zeros in a multi-digit display. Other chips that were designed for direct connection to LED numerals have become obsolete as the need for this application has diminished. The 74C925, 74C926, 74C927, and 74C928 are examples. They may still be found from surplus outlets, but should not be specified in new circuit designs.

What Can Go Wrong

The entry that deals with problems affecting shift registers (see "What Can Go Wrong" on page 128), describes issues which also affect counters. The entry that deals with logic chips (see "What Can Go Wrong" on page 105), describes problems affecting all types of logic chips. In addition, the potential problems listed below are specific to counters.

Lock-Out

This is the condition which occurs if a counter with a shortened modulus is loaded with a binary state that is out of its range. Consult the datasheet and study its state diagram to determine the most likely outcome if this problem occurs.

Asynchronous Artifacts

Because the flip-flops in an asynchronous (ripple) counter do not change simultaneously, they create very brief false outputs while the ripple process is taking place. In a 4-bit counter, the binary number 0111 (decimal 7) should be followed by 1000 (decimal 8). However, the rightmost digit (i.e., the least significant bit) will change to a 0 initially, creating 0110 as a momentary binary output (decimal 6). The carry operation will then change the next digit to a 0, creating 0100 (decimal 4). The carry operation continues, changing the next digit to a 0, creating 0000. Finally the operation completes by creating 1000 as the correct output.

These intermediate states on the output pins are often referred to as *glitches*. Because they are extremely brief, they will be indetectable when a counter is used to drive a display. They can cause significant issues, however, if the outputs of the counter are connected with other logic chips.

Another type of asynchronous problem will occur if the clock speed is sufficiently high that a new pulse is received at the first flip-flop before the ripple of changing states has passed all the way through to the final flip-flop. This will result in a different brief invalid value on the output pins.

Noise

Old TTL-type counters, such as the 74LSxx series, are especially noise-sensitive. Adding a 0.1μF or 0.047μF bypass capacitor as close to the power supply pin as possible is recommended. Breadboarding counters of this type may result in errors if a high-frequency clock is used, because conductors such as patch-cords are liable to pick up noise. Modern 74HCxx counters are preferable.

encoder

14

In this encyclopedia, an **encoder** is a digital chip that converts a decimal-valued input into a binary-coded output.

The term “encoder” may alternatively refer to a **rotational encoder** (also known as a *rotary encoder*) which has a separate entry in Volume 1 of this encyclopedia. The term may also describe a *code hopping encoder*, which is an encryption device used in keyless entry systems for automobiles.

OTHER RELATED COMPONENTS

- **decoder** (see Chapter 15)
- **multiplexer** (see Chapter 16)

What It Does

An encoder is a *logic chip* that receives an input consisting of an active logical state on one of at least four input pins, which have decimal values from 0 upward in increments of 1. The encoder converts the active pin number into a binary value represented by logic states on at least two output pins. This behavior is opposite to that of a **decoder**.

Encoders are identified in terms of their inputs and outputs. For example:

- 4-to-2 encoder (four input pins, two output pins)
- 8-to-3 encoder (eight input pins, three output pins)
- 16-to-4 encoder (sixteen input pins, four output pins)

In the early days of computing, encoders processed interrupts. This application is now rare, and relatively few encoder chips are still being manufactured. However, they are still useful in small devices—for example, if a large number of inputs must be handled by a microcontroller that has insufficient pins to receive data from each individually.

Schematic Symbol

Like other logic-based components, the encoder does not have a specific schematic symbol and can be represented by a plain rectangle as in Figure 14-1, with inputs on the left and outputs on the right. The bars printed above some of the abbreviations indicate that an input or output is active-low. In this chip, the 74LS148, all inputs and outputs are active-low.

Generally speaking, inputs labeled D0, D1, D2… are used for data input, although they may simply be numbered, with no identifying letter. The encoded outputs are typically identified as Q0, Q1, Q2… or A0, A1, A2… with Q0 or A0 designating the least significant bit in the binary number.

Pins labeled E and GS are explained in the following section.

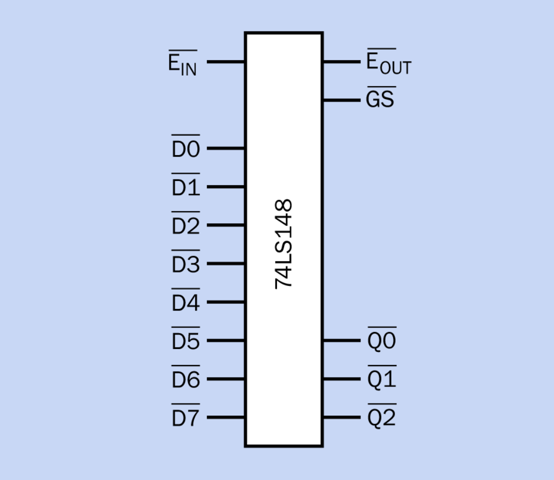

Figure 14-1. *While no specific schematic symbol exists for an encoder chip, this style is commonly used. Shown here is a 16–to–4 encoder with active-low inputs and outputs.*

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An analog multiplexer may allow its inputs and outputs to be swapped, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

How It Works

An encoder contains **logic gates**. The internal logic of an 8-to-3 encoder is shown in Figure 14-2, where the darker blue rectangle represents the chip. The switches in this figure are external and are included only to clarify the concept. An open switch is imagined to provide an inactive logic input, while a single closed switch provides an active logic input. (Multiple active inputs can be handled by a *priority encoder*, described below).

Each input switch has a numeric status from 1 to 7. The switch with value 0 does not make an internal connection, because the output from the OR gates is 000 by default.

The logic state of each OR output represents a binary number, weighted with decimal values 1, 2, and 4, as shown at the bottom of the figure. Thus, if switch 5 is pressed, by tracing the connections it is clear that the outputs of OR gates 4 and 1 become active, while the output from gate 2 remains inactive. The values of the active outputs thus sum to 5 decimal.

Figure 14-3 shows the outputs for all possible input states of a 4-to-2 encoder. Figure 14-4 shows the outputs for all possible input states of an 8-to-3 encoder. These diagrams assume that a high logic state is an active logic state, on input or output. This is usually the case.

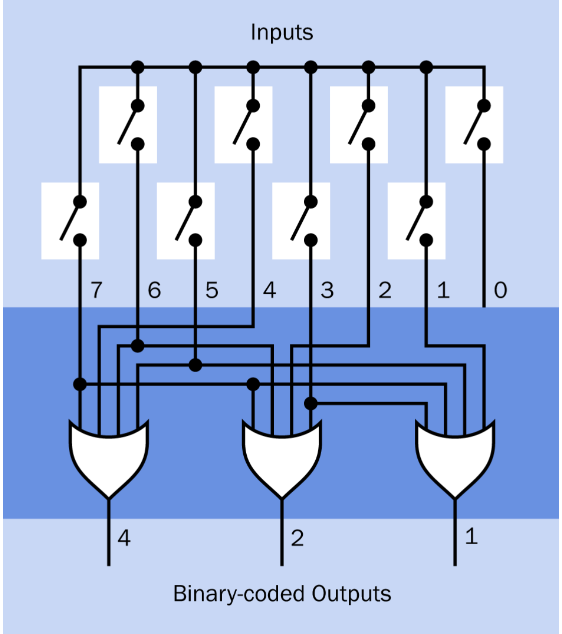
Inputs
7
6
5
4
3
2
1
0
4
2
1
Binary-coded Outputs

Figure 14-2. *A simplified simulation of the internal logic of an 8-to-3 encoder. The dark blue rectangle indicates the space inside the chip. The external switches are included only to clarify the concept. An encoder chip would have an Enable line to create an active output.*

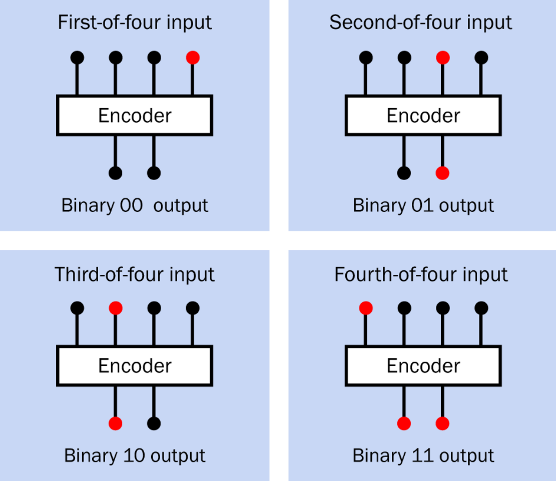
First-of-four input
Encoder
Binary 00 output
Second-of-four input
Encoder
Binary 01 output
Third-of-four input
Encoder
Binary 10 output
Fourth-of-four input
Encoder
Binary 11 output

Figure 14-3. *The four possible inputs of a 4-to-2 encoder (top of each panel) and the encoded outputs (below).*

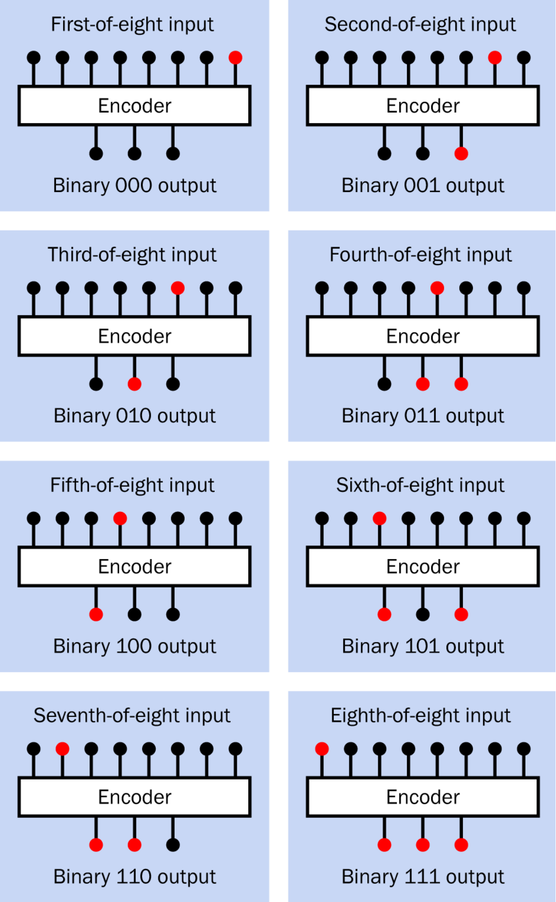
First-of-eight input
Encoder
Binary 000 output
Second-of-eight input
Encoder
Binary 001 output
Third-of-eight input
Encoder
Binary 010 output
Fourth-of-eight input
Encoder
Binary 011 output
Fifth-of-eight input
Encoder
Binary 100 output
Sixth-of-eight input
Encoder
Binary 101 output
Seventh-of-eight input
Encoder
Binary 110 output
Eighth-of-eight input
Encoder
Binary 111 output

Figure 14-4. *The eight possible inputs of an 8-to-3 encoder (at the top of each panel), and the encoded outputs (below). Note that one input of an encoder must always be logic-high. All logic-low inputs are not a valid state.*

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders respond within two or three nanoseconds.

Variants

A *simple encoder* assumes that only one input pin can be logically active at a time. A *priority encoder* assigns priority to the highest-value input pin if more than one happens to receive an active input. It ignores any lower-value inputs. An example is the 74LS148, which is an 8-to-3 chip.

A few encoders feature *three-state* outputs (also known as *tri-state*), in which a high-impedance or "floating" output state is available in addition to the usual high and low logic states. The high-impedance state allows multiple chips to share an output bus, as those that are in high-impedance mode appear to be disconnected. This is useful if two or more encoders are cascaded to handle a larger number of inputs.

Values

As is the case with other logic chips, most encoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low input states. On the output side, the 4000 series chips are able to source less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

How to Use It

Suppose that a microcontroller should respond to an eight-position rotary switch. Because the switch cannot be turned to more than one position at a time, all of its eight contacts can be connected with the inputs on an encoder, which will deliver a 3-bit binary number to three inputs of the microcontroller. Code inside the microcontroller then interprets the pin states.

This is shown in Figure 14-5. Pulldown resistors would be needed on the input pins of the encoder, to prevent them from floating when they are not connected by the rotary switch. They have been omitted from this diagram for simplicity. Debouncing the switch would be handled by the microcontroller.

Other forms of input may be used instead of a rotary switch. For example, the outputs from eight comparators or eight phototransistors could be passed through an encoder.

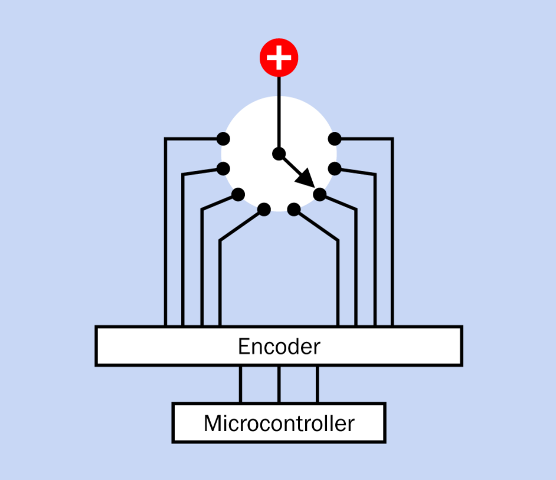

Figure 14-5. *Output from an eight-input rotary switch could be connected through an 8-to-3 encoder to provide input to a microcontroller using a reduced number of pins. Pulldown resistors have been omitted for simplicity.*

Cascaded Encoders

Encoders are often provided with features to facilitate handling additional inputs via multiple chips. Typically, a second Enable pin is provided, as an output that connects with the Enable input of the preceding chip. This preserves the priority function, so that an input on the second chip prevents any additional input to the first chip from affecting the output. In a datasheet, the enable pins may be labeled E_{IN} and E_{OUT}, or E_I and E_O.

In addition, a GS pin will be included, meaning "Group Select." It is logically active only when the encoder is enabled and at least one input is active. The GS pin of the most-significant encoder provides an additional binary digit.

The outputs from two encoders can be linked via OR gates, as shown in Figure 14-6, where the lower chip's GS output provides the most significant bit of a four-bit binary number.

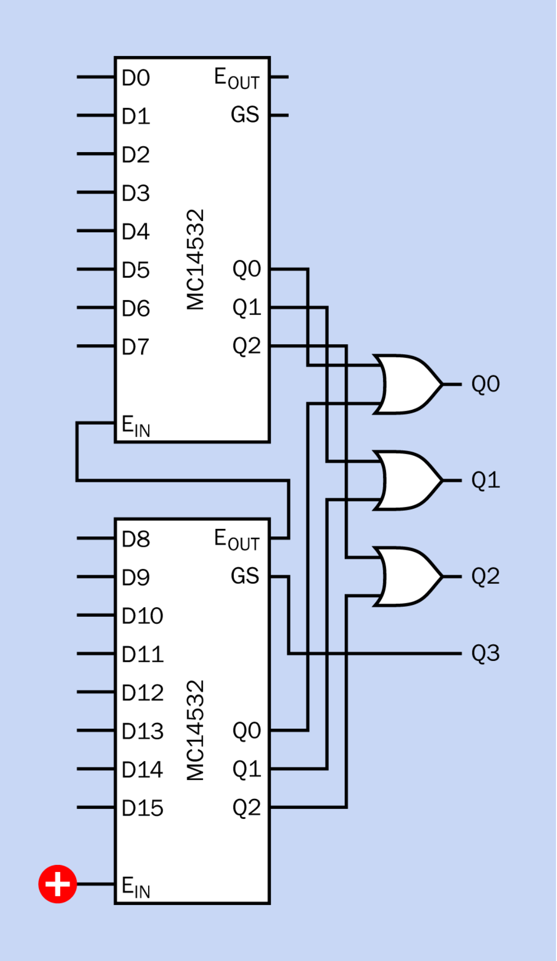
D0
D1
D2
D3
D4
D5
D6
D7
EIN
EOUT
GS
Q0
Q1
Q2
MC14532
Q0
Q1
Q2
Q3
D8
D9
D10
D11
D12
D13
D14
D15
EIN
EOUT
GS
Q0
Q1
Q2
MC14532

Figure 14-6. *Two eight-bit encoders can be cascaded to handle 16 separate inputs. In this example, the encoders use active-high logic.*

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in "What Can Go Wrong" on page 105.

See "What Can Go Wrong" on page 149 in the entry describing **decoders** for a list of more specific problems that also afflict encoders.

decoder

15

In this Encyclopedia, a **decoder** is a digital chip that receives a binary-coded input and converts it to a decimal output by applying a logic state to one of a sequence of pins, each of which is assigned an integer value from 0 upward.

The term "decoder" also refers to components and devices that have other functions, such as decoding audio or video formats. These functions are not included here.

OTHER RELATED COMPONENTS

- **encoder** (see Chapter 14)
- **multiplexer** (see Chapter 16)

What it Does

A decoder receives a binary-coded number on two or more input pins. It *decodes* that number and expresses it by activating one of at least four output pins.

The behavior of a decoder with a two-bit binary input is shown in four sequential snapshots in Figure 15-1, where the least significant bit of the input is on the right in each diagram, and the output moves from right to left.

Figure 15-2 shows a similar sequence in a decoder where various values of a three-bit input are decoded to create an eight-pin output.

One sample state of a four-bit decoder is shown in Figure 15-3.

All of these figures assume that a high state represents an active input or output. In a few chips, a low state is used to represent an active output.

Decoders with 2, 3, or 4 input pins are common. To handle a binary input greater than 1111 (decimal 15), decoders can be chained together, as described below.

Manufacturers' datasheets often describe decoders in terms of their inputs and outputs. Typical examples would include:

- 2-to-4 decoder (two input pins, four output pins)
- 3-to-8 decoder (three input pins, eight output pins)
- 4-to-10 decoder (for converting *binary-coded decimal* to decimal output)
- 4-to-16 decoder (also known as a *hex decoder*).

Input Devices

The input pins of a decoder can be driven by a counter that has a binary-coded output. A decoder can also be driven by a microcontroller, which may have an insufficient number of output pins to control a variety of devices. Two, three, or four of the outputs can be used to represent a binary number which is passed through the decoder to activate the devices one at a time, perhaps with transistors or Darlington arrays introduced to handle the load. This is suggested in Figure 15-4.

A **shift register** can be used for a similar purpose, but often has only one pin for input. This pin must be supplied sequentially with a serial pattern of bits that will match the desired high/low states of the output pins. The relative advantage of this system is that a shift register can generate any pattern of output states. A one-of-many decoder can activate only one output at a time.

LED Driver

A special case is a *seven-segment decoder* designed to drive a seven-segment **LED display** numeral. A binary-coded decimal number on four input pins is converted to a pattern of outputs appropriate for lighting the segments of the display that will form a number from decimal 0 through 9.

Schematic Symbol

Like other logic-based components, the decoder does not have a specific schematic symbol and is represented by a plain rectangle as in Figure 15-5, with inputs on the left and outputs on the right. The bars printed above the E and LE abbreviations (which stand for Enable and Latch Enable, respectively) indicate that they are active-low. In this chip, the 74HC4514, all outputs are active-high, but in a related 4-to-16 decoder, the 74HC4515, all outputs are active-low. In both of these chips, the Enable pin is held low to activate the outputs. The Latch Enable pin freezes the current state of the outputs (i.e., it latches them) when it is held low.

Generally speaking, pins labeled A0, A1, A2… in a datasheet are often the binary inputs (although A, B, C… may be used), with A0 designating the least significant bit. Outputs are usually labeled Y, and are activated in sequence from Y0 when the binary input starts counting upward.

Similar Devices

The similarities and differences between encoder, decoder, multiplexer, and demultiplexer can cause confusion.

- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive logic state.
- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An analog multiplexer may allow its inputs and outputs to be reversed, in which case it becomes a **demultiplexer**. It can connect a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

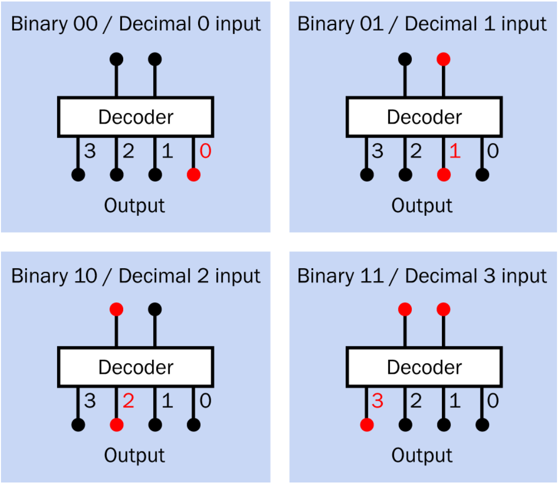
Binary 00 / Decimal 0 input
Decoder
3
2
1
0
Output
Binary 01 / Decimal 1 input
Decoder
3
2
1
0
Output
Binary 10 / Decimal 2 input
Decoder
3
2
1
0
Output
Binary 11 / Decimal 3 input
Decoder
3
2
1
0
Output

Figure 15-1. *A decoder with two input pins can interpret their binary-number representation to create an active logic state on one of four output pins.*

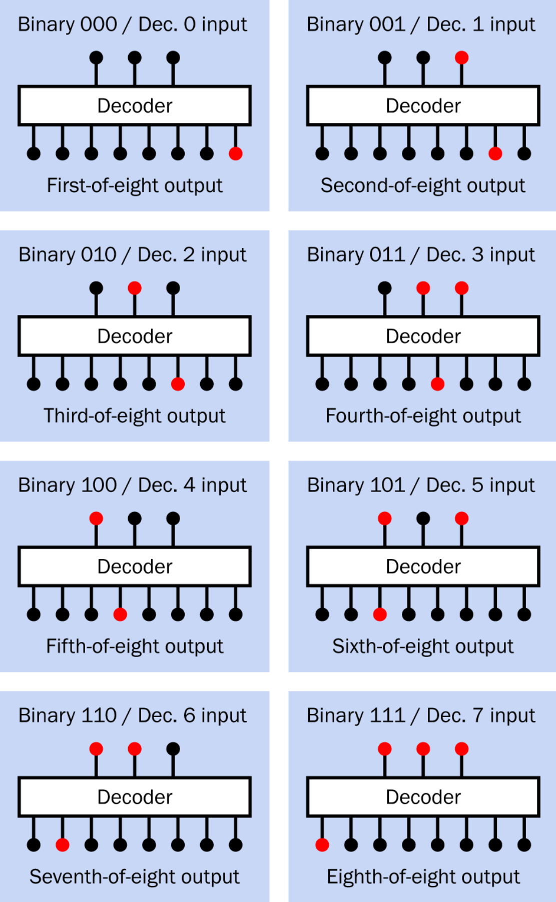
Binary 000 / Dec. 0 input
Decoder
First-of-eight output
Binary 001 / Dec. 1 input
Decoder
Second-of-eight output
Binary 010 / Dec. 2 input
Decoder
Third-of-eight output
Binary 011 / Dec. 3 input
Decoder
Fourth-of-eight output
Binary 100 / Dec. 4 input
Decoder
Fifth-of-eight output
Binary 101 / Dec. 5 input
Decoder
Sixth-of-eight output
Binary 110 / Dec. 6 input
Decoder
Seventh-of-eight output
Binary 111 / Dec. 7 input
Decoder
Eighth-of-eight output

Figure 15-2. *A decoder with three input pins can interpret their binary-number representation to create a high logic state on one of eight output pins.*

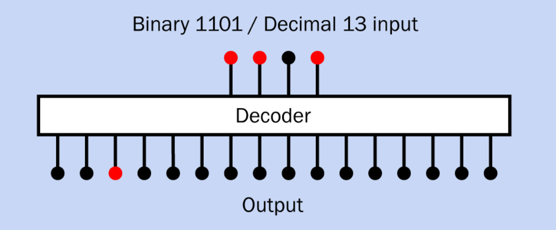
Binary 1101 / Decimal 13 input
Decoder
Output

Figure 15-3. *A decoder with four input pins can interpret their binary-number representation to create a high logic state on one of 16 output pins. Only one of the 16 possible states is shown here.*

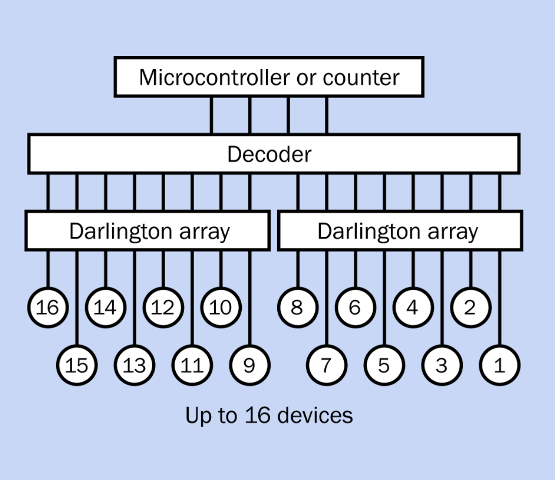
Microcontroller or counter
Decoder
Darlington array
Darlington array
16
14
12
10
8
6
4
2
15
13
11
9
7
5
3
1
Up to 16 devices

Figure 15-4. *Four outputs from a binary counter or microcontroller can be used by a decoder to activate one of up to 16 output devices.*

A photograph of a 74HC4514 decoder chip appears in Figure 15-6.

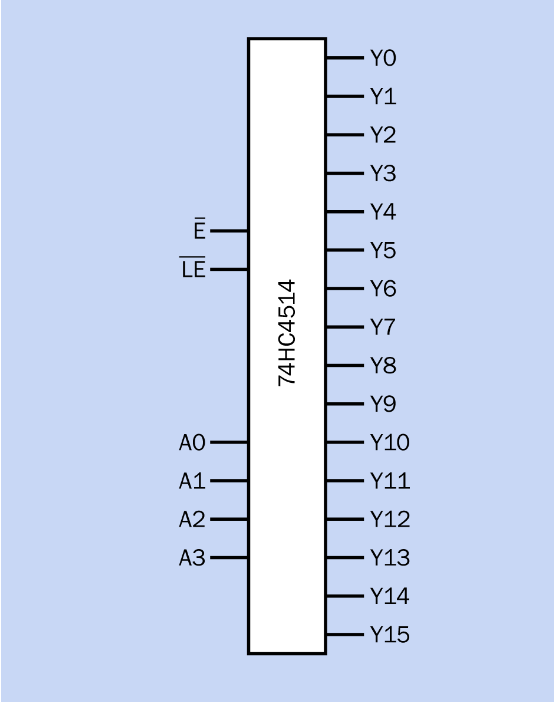
74HC4514
$\overline{E}$
$\overline{LE}$
A0
A1
A2
A3
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
Y10
Y11
Y12
Y13
Y14
Y15

Figure 15-5. *While no specific schematic symbol exists for a decoder chip, this style is commonly used. Shown here is a 4-to-16 decoder.*

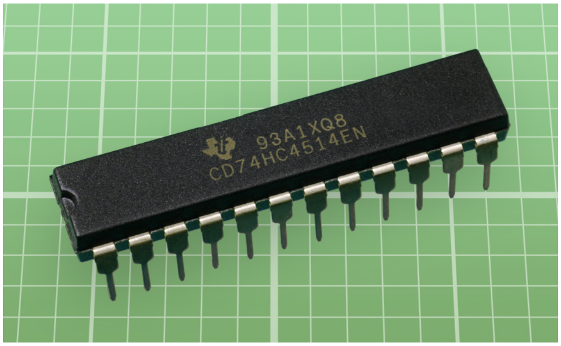

Figure 15-6. *The 24-pin 74HC4514 decoder chip processes a 4-bit input and represents it by making one of its 16 output pins active-high.*

How It Works

A decoder contains logic gates, each of which is wired to respond to a unique binary pattern of inputs. (In the case of a seven-segment decoder, the internal logic is more complicated.) Figure 15-7 shows the logic of a 2-to-4 decoder. The darker blue area contains the components inside the chip. The external switches are included only to clarify the function of the decoder. An open switch is imagined to provide a low logic input, while each closed switch provides a high logic input.

Unlike ripple counters, where propagation delays can reduce the overall response time of the component, decoders function within two or three nanoseconds.

Variants

Decoder variants have not proliferated with time, and relatively few are available. Most are 3-to-8, 4-to-16, and binary-coded-decimal types.

The 7447 and 74LS47 are seven-segment decoders that have an open-collector output capable of driving a 7-segment display directly. The 7448 is similar but also contains built-in resistors and a capability to blank out leading zeros in a display. However, some suppliers now list the 74LS48 as obsolete. It may be still available from old stock, but should not be specified in new circuits.

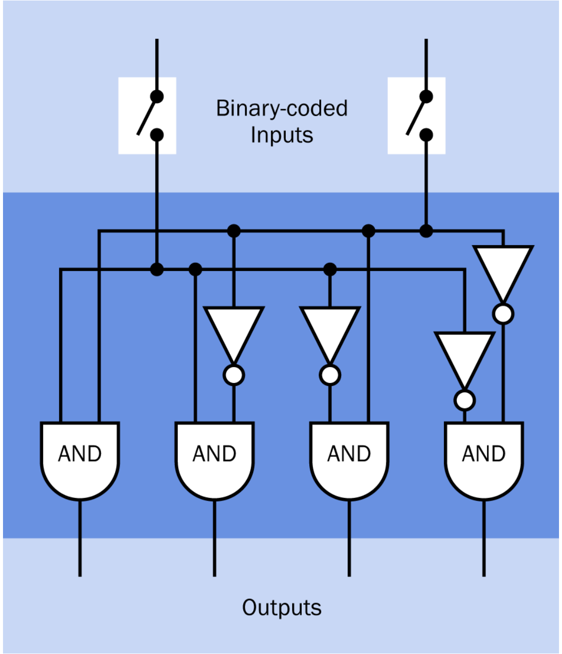

Figure 15-7. *A simplified simulation of the logic in a decoder. An actual chip would have an Enable line to activate the output. The dark blue rectangle indicates the space inside the chip.*

Although 74LS47 is still being manufactured, and is available in surface-mount as well as through-hole format, a version is not available in the widely used HC family of 74xx chips. Care must be taken to satisfy the input voltage requirements of the 74LS47 when driving it with 74HCxx chips.

Values

As is the case with other logic chips, most decoders in the through-hole 74xx series are intended for 5VDC power supply while the older 4000 series may tolerate up to 18VDC. Surface-mount versions may use voltages as low as 2VDC.

See the section on logic gates in Chapter 10 for a discussion of acceptable high and low input states. On the output side, the 4000 series chips are able to source or sink less than 1mA at 5VDC, but the 74HCxx series can manage around 20mA.

How to Use It

The original applications for decoders in computer circuits have become uncommon, but the chips can still be useful in small appliances and gadgets where multiple outputs are controlled by a **counter** or microcontroller.

Although 16 is usually the maximum number of outputs, some chips are designed to allow expansion. The 74x138 (where a chip family identifier such as LS or HC can be substituted for the letter x) is a 3-to-8 decoder with two logic-low Enable pins and one logic-high Enable. If a value-8 binary line is applied to the low-enable of one chip and the high-enable of another, the first chip will be disabled when the line goes high to indicate that the binary number 1000 has been reached, and the second chip can continue upward from there by sharing the same three less-significant-bit inputs. As many as four chips can be chained in this way.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates in "What Can Go Wrong" on page 105.

Glitches

Although a decoder typically functions faster than a ripple counter, it suffers the same tendency to introduce brief *glitches* in its output. These are momentary invalid states which occur while processes inside the chip that are slightly slower are catching up with other processes that reach completion slightly faster. A brief *settling time* is necessary to ensure that the output is stable and valid. This will be irrelevant when powering a device such as an **LED indicator**, which will not display such brief transients. The problem may be more important if the output from the decoder is used as an input to other logic chips.

If the input to a decoder is derived from a ripple counter, the input may also contain glitches, which can cause erroneous outputs from the decoder. It is better to use a synchronous counter on the input side of a decoder.

Unhelpful Classification

Online parts suppliers tend to list decoders under the same category heading as encoders, multiplexers, and demultiplexers, making it difficult to find what you want. Under this broad subject heading (which will include thousands of chips), if you search by selecting the number of inputs relative to the number of outputs that you have in mind, this will narrow the search considerably.

Active-Low and Active-High

Chips with identical appearance and similar part numbers may have outputs that are either active-low or active-high. Some may offer a latch-enable pin, while others have enable pins that must be pulled low or forced high to produce an output. Accidental chip substitution is a common cause of confusion.

multiplexer

May be abbreviated as a *mux* (this term is sometimes printed all in caps), and may be referred to alternatively as a *data selector*. Some sources maintain that a multiplexer has no more than two channels, whereas a data selector has more, but there is no consensus on this, and datasheets continue to use the term "multiplexer" predominantly.

Analog multiplexers are usually bidirectional, and thus will function equally well as *demultiplexers*. Consequently, this encyclopedia does not contain a separate entry for demultiplexers.

OTHER RELATED COMPONENTS

- **encoder** (see Chapter 14)
- **decoder** (see Chapter 15)

What It Does

A multiplexer can select one of two or more input pins, and connect it internally with an output pin. Although it is an entirely solid-state device, it behaves as if it contains a **rotary switch** in series with a SPST switch, as shown in Figure 16-1. A binary code applied to one or more Select pins chooses the input, and an Enable pin establishes the connection with the output. The Select and Enable functions are processed via an internal section referred to as a *decoder*, not to be confused with a **decoder** chip, which has its own entry in this encyclopedia.

All multiplexers are digitally controlled devices, but may be described as either *digital* or *analog* depending how they process the input signal. A digital multiplexer creates an output that is adjusted to logic-high or logic-low within the limits of its logic family. An analog multiplexer does not impose any processing on the voltage, and passes along any fluctuations. Thus, it can be used with alternating current.

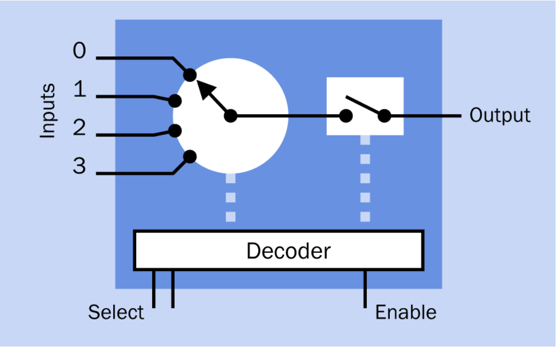

Figure 16-1. *A multiplexer functions as if it contains a rotary switch. The switch position is determined by a binary number applied to external Select pins. The internal connection is completed by applying a signal to an Enable pin.*

Because an analog multiplexer merely switches a flow of current, it can be *bidirectional*; in other words, it can function as a *demultiplexer*, in which case the input is applied to the pole of the (imaginary) internal switch and outputs are taken from the terminals.

Differential Multiplexer

A *differential* multiplexer contains multiple switches that are differentiated from one another (i.e., they are electrically isolated, although they are controlled by the same set of select pins). A differential multiplexer is conceptually similar to a rotary switch with two or more decks controlled by a single shaft. See Figure 16-2.

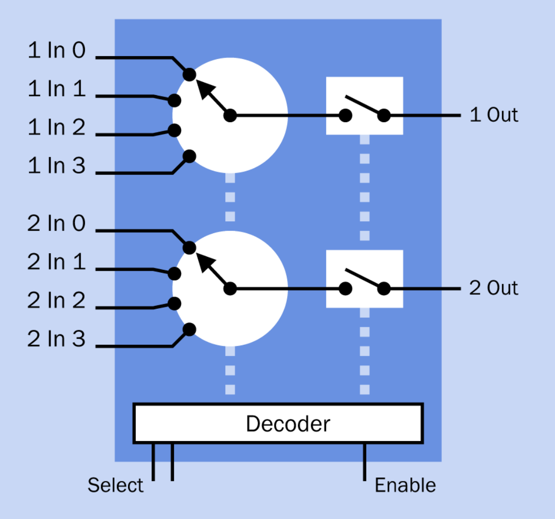

Figure 16-2. *A differential multiplexer contains two or more electronic switches that are differentiated from one another, similarly to the decks on a rotary switch. Although the channels into each switch are typically numbered from 0 upward, the switches are numbered from 1 upward.*

A bidirectional dual 4-channel differential analog multiplexer is shown in Figure 16-3.

Modern multiplexers are often found switching high-frequency data streams in audio, telecommunications, or video applications.

Similar Devices

The similarities and differences between multiplexer, demultiplexer, encoder, and decoder can cause confusion:

- A **multiplexer** can connect a choice of multiple inputs to a single output, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which input should be connected with the output pin. The alternative term *data selector* evokes the function of this device more clearly.
- An *analog* multiplexer may allow its inputs and outputs to be reversed, allowing it to become a **demultiplexer**, connecting a single input to one of multiple outputs, for data transfer. The logic state of an enable pin, or a binary number applied as a pattern of logic states to multiple control pins, chooses which output should be used. The alternative term *data distributor* evokes the function of this device more clearly.

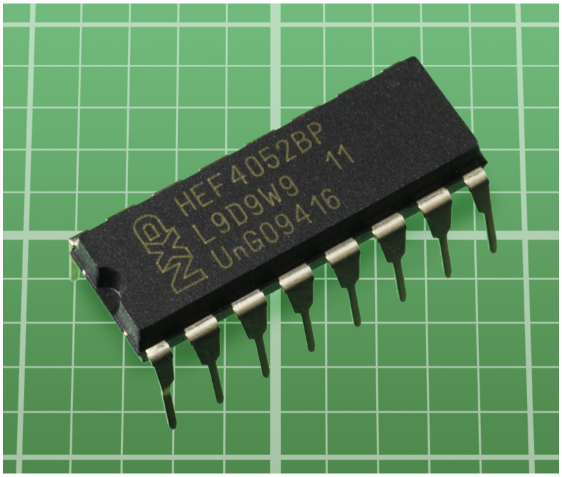

Figure 16-3. *This CMOS chip contains two four-channel differential analog multiplexers.*

- In an **encoder**, an active logic state is applied to one of four or more input pins, while the rest remain in an inactive logic state. The input pin number is converted to a binary code which is expressed as a pattern of logic states on two or more output pins.
- In a **decoder**, a binary number is applied as a pattern of logic states on two or more input pins. This value determines which one of four or more output pins will have an active logic state, while the rest remain in an inactive

logic state. A digital multiplexer does not allow reversal of its inputs and outputs, but a decoder functions as if it were a digital demultiplexer.

How It Works

The multiple inputs to a multiplexer are referred to as *channels*. Almost always, the number of channels is 1, 2, 4, 8, or 16. A 1-channel component is only capable of "on" or "off" modes and functions similarly to a SPST switch.

If there are more than two channels, a binary number will determine which channel is connected internally. The number of channels is usually the maximum that can be identified by the number of select pins, so that 2 pins will control 4 channels, 3 pins will control 8 channels, and 4 pins (the usual maximum) will control 16 channels.

In multiplexers with three or more channels, an enable pin is usually still present to activate or deactivate all the channels simultaneously. The enable feature may be described alternatively as a *strobe*, or may have an inverse function as an *inhibit* pin.

Although a rotary switch is helpful in conceptualizing the function of a multiplexer, a more common representation (sometimes in datasheets) is an array of SPST switches, each of which can be opened or closed by the decoder circuit. A typical example, depicting a dual differential multiplexer, is shown in Figure 16-4. Note that the internal decoder can only close one switch in each channel at a time.

The switch analogy is appropriate in that when an output from a multiplexer is not connected internally (i.e., its switch is "open") it is effectively an open circuit. However, some multiplexers contain *pullup resistors* to give each output a defined state. This can be an important factor in determining whether the multiplexer is suitable for a particular application.

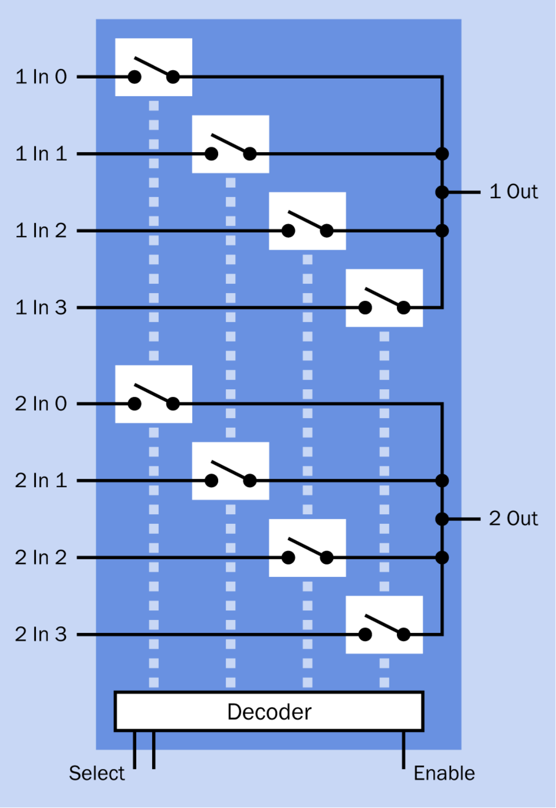

Figure 16-4. *The internal function of a dual multiplexer is commonly represented as a network of SPST switches, each of which is controlled by decoder logic.*

A digital multiplexer actually contains a network of logic gates, shown in simplified form in Figure 16-5.

A demultiplexer has internal logic shown in simplifier form in Figure 16-6.

Schematic Symbol

In a schematic, a multiplexer and demultiplexer may be represented by a trapezoid with its longer vertical side oriented toward the larger number of connections. This is shown in Figure 16-7. However, this symbol is falling into disuse.

More often, as is the case with most logic components, a multiplexer or demultiplexer is represented by a rectangle with inputs on the left and outputs on the right, as shown in Figure 16-8. The distinction between inputs and outputs is problematic, however, in an analog multiplexer which will allow data flow to be reversed.

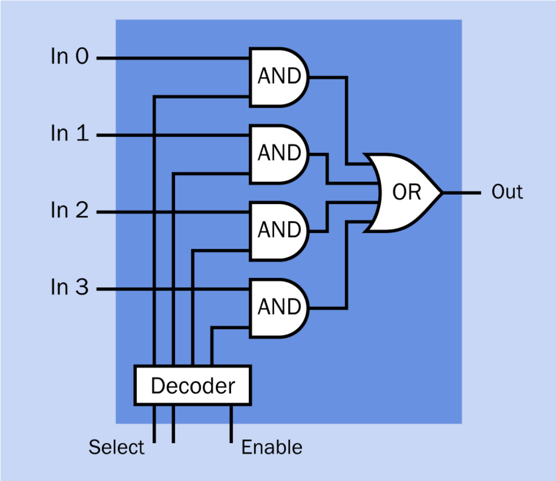
In 0
In 1
In 2
In 3
AND
AND
AND
AND
OR
Out
Decoder
Select
Enable

Figure 16-5. *A simplified representation of the logic gates in a digital multiplexer.*

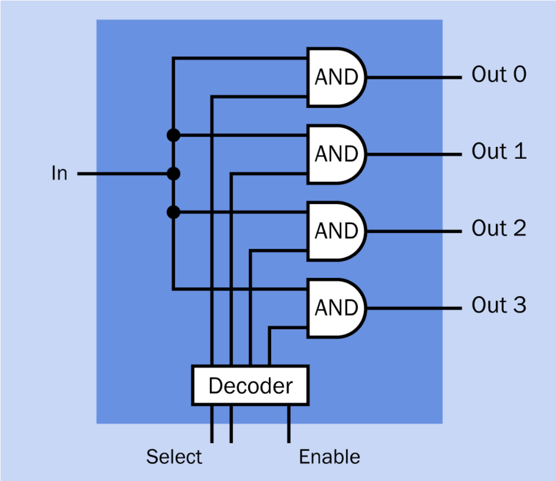
In
AND
AND
AND
AND
Out 0
Out 1
Out 2
Out 3
Decoder
Select
Enable

Figure 16-6. *A simplifier representation of the logic gates in a digital demultiplexer.*

Pin Identifiers

The lack of standardization in the identification of pin functions is perhaps more extreme in the case of multiplexers than for other types of logic chips.

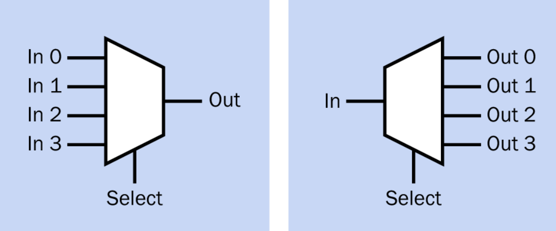
In 0
In 1
In 2
In 3
Out
Select
In
Out 0
Out 1
Out 2
Out 3
Select

Figure 16-7. *The traditional symbol for a multiplexer (left) and demultiplexer (right). The trapezoid is oriented with its longer vertical side facing the larger number of connections. This symbol is falling into disuse.*

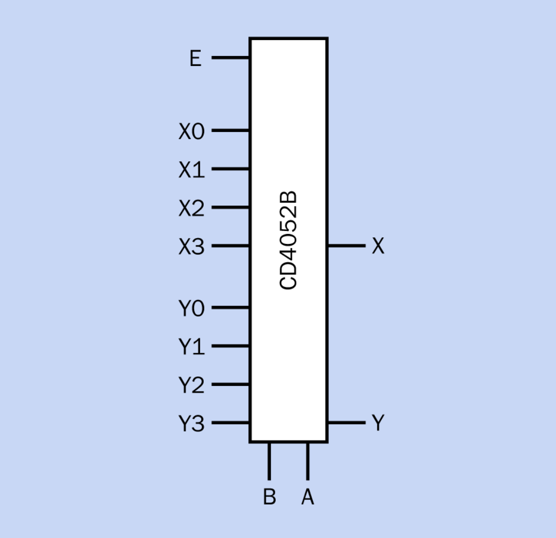
E
X0
X1
X2
X3
Y0
Y1
Y2
Y3
CD4052B
X
Y
B
A

Figure 16-8. *A simple rectangle is most often used as a schematic symbol for a multiplexer, but the abbreviations assigned to pin functions are not standardized. See text for details.*

An output enable pin will be shown as E or EN, or occasionally OE. It may alternatively be described as an inhibit pin, labeled INH, or sometimes will be called a strobe. The function is the same in each case: one of its logic states will enable the internal switches, while its other logic state will prevent any internal switches from closing.

Switch inputs may be labeled S0, S1, S2… or X0, X1, X2… or may simply be numbered, almost always counting up from 0. Where two or more sets of switches coexist in one package, each set of inputs may be distinguished from the others by preceding each identifier with a numeral or letter to designate the switch, as in 1S0, 1S1, 1S2… or 1X0, 1X1, 1X2… (Switches are generally numbered from 1 upward, even though their inputs are numbered from 0 upward.) Outputs may be identified using the same coding scheme as inputs, bearing in mind that the inputs and outputs of an analog multiplexer usually are interchangeable. Some manufacturers, however, prefer to identify each multiplexer output by preceding it with letter Y. Alternatively, Z1, Z2, Z3… may identify the outputs from switches 1, 2, 3… Fortunately, datasheets usually include some kind of key to this grab-bag of abbreviations.

Control pins are often identified as A, B, C… with letter A representing the least significant bit in the binary number that is applied to the pins.

Voltages can be confusing in multiplexers. Components intended for use with digital inputs are straightforward enough, as the supply voltage will be identified as V_{CC} and is typically 5VDC for through-hole packages (often lower for surface-mount), while negative ground is assumed to be 0VDC. However, where a multiplexer may be used with AC inputs in which the voltage varies above and below 0V, supply voltages above and below 0VDC are also possible—such as +7.5VDC and −7.5VDC, to take a random example. Three power-supply pins may be provided for this purpose. The positive supply will usually be identified as V_{DD} (the D refers to the Drain in the internal MOSFETs). A V_{EE} pin may be at 0VDC or at a negative value equal and opposite to V_{DD}. The E in this abbreviation is derived from Emitter voltage, even though the component may not contain a bipolar transistor with an emitter. Customarily, a V_{SS} pin (the S being derived from the Source in the internal MOSFETs) will be at 0VDC, and other voltages will be measured above and below this baseline. This ground pin may alternatively be labeled GND.

As is customary in logic chips, low-active control pins will have a bar printed above their identifiers, or an apostrophe will be placed after an identifier if the font does not permit printing the bar. Alternatively, low-active pins may be represented by showing a small circle, properly referred to as a *bubble*, at the input or output point of the symbol for the multiplexer. Note that analog inputs and outputs are neither high-active nor low-active; they merely pass voltages through.

Variants

Most multiplexers are “break before make” devices, where one input is disconnected before the next input is connected. However, some exceptions exist, and datasheets should be checked for this. It can be a significant issue, because make-before-break switching will briefly connect external devices with each other, through the chip.

Many multiplexers can tolerate control voltages above the usual high value in a logic circuit—as high as 15VDC in some cases. The voltage that is switched by the multiplexer may be the same as the control voltage, or may be higher.

Some analog multiplexers have *overvoltage protection* that allows them to withstand input voltages that are twice or three times the recommended maximum.

Datasheets may mention “internal address decoding,” meaning that the binary number input, specifying a channel to be switched, is decoded inside the chip. In fact, virtually all multiplexers now have on-chip address decoding, and this feature should be assumed to exist, regardless of whether it is mentioned.

Values

The voltage to be switched will usually be referred to as the *input voltage*, V_{IN}.

An analog multiplexer should not be subjected to current exceeding the value that it is designed to switch. This is known as the maximum *channel current*. A typical value would be 10mA, although many modern surface-mount components are designed for currents in the microamp range.

The *on-resistance* is the resistance imposed by the analog multiplexer on the signal flowing through it. While modern, specialized analog multiplexers may have an on-resistance as low as 5Ω, these are relatively unusual. An on-resistance of 100Ω to 200Ω is more common. This value will vary within a component depending on the power supply voltage and the voltage being switched. It will increase slightly as V_{IN} deviates above (or below) 0V, will increase substantially for lower values of supply voltage, and will increase significantly with temperature.

The curves in Figure 16-9 show on-resistance of an analog multiplexer varying with input voltage, with three different power supplies: plus-and-minus 2.5VDC (described in the graph as a "spread" of 5VDC), plus-and-minus 5VDC (a "spread" of 10VDC), and plus-and-minus 7.5VDC (a "spread" of 15VDC). These curves were derived from a datasheet for the MC14067B analog multiplexer; curves for other chips will be different, although the basic principles remain the same.

Switching time is an important consideration in high-speed applications. The "on" and "off" times specified in a datasheet (often as t_{ON} and t_{OFF}) are a function of the propagation delay from the control input to the toggling of the switch, and are generally measured from the halfway point of the rising or falling edge of the control input, to the 90% point of the output signal level.

Leakage current is the small amount of current (often measured in picoamperes) that the solid-state switch will pass when it is in its "off" state. This should be insignificant except when very high-impedance loads are used.

Separate switches inside a multiplexer may have characteristics that differ slightly from one another. Differences in on-resistance between adjacent switches can be important when switching parallel analog signals. A datasheet should mention the extent to which switches have matched characteristics, and may define the maximum deviation from one another using the abbreviation R_{ON} even though this same term may be used, confusingly, to denote the on-resistance of each individual switch.

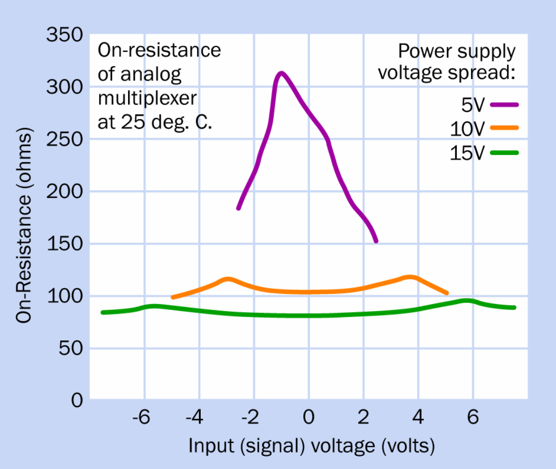

Figure 16-9. *Variations in on-resistance in an analog multiplexer. Each voltage "spread" is the difference between positive supply voltage and an equal-and-opposite negative ground voltage. Thus a "spread" of 10VDC means plus and minus voltages of 5VDC. (Curves derived from On Semiconductor datasheet for MC14067B analog multiplexer.)*

How to Use It

A multiplexer may be used as a simple switch to choose one of multiple inputs, such as a choice of input jacks on a stereo system. A dual differential multiplexer is useful in this application, as it can use a single select signal to switch two signal paths simultaneously.

A multiplexer can also be used as a digital volume control by switching an audio signal among a variety of resistances, similar to a **digital potentiometer**. In this application, the possible presence of pullup resistors inside the multiplexer must be considered.

Where a microcontroller must monitor a large number of inputs (for example, a range of temperature sensors or motion sensors), a multiplexer can reduce the number of input pins required. Its data-select pins will be cycled through all the possible binary states by the microcontroller, to select each data input in turn, while its single-wire output will carry the analog data to a separate pin on the microcontroller which performs an analog-digital conversion.

Conversely, a demultiplexer (i.e., an analog multiplexer such as the 4067B chip which can be used in demultiplexer mode) can be used by a microcontroller to switch multiple components on and off. Four outputs from the microcontroller can connect with the control pins of a 16-channel demultiplexer, counting from binary 0000 through binary 1111 to select output pins 0 through 15. After selecting each pin, the microcontroller can send a high or low pulse through it. The process then repeats. (A **decoder** can be used in the same way.)

Other Application Notes

Multiplexers may be *cascaded* to increase the inputs-to-outputs ratio.

Modern multiplexers are found on computer boards where they choose among video output ports, or as PCI express channel switches.

A multiplexer may be used as a parallel-to-serial converter, as it samples multiple channels and converts them into a serial data stream.

In telecommunications, a multiplexer can sample voice signals from multiple separate inputs and combine them into a digital stream that can be transmitted at a faster bit rate over a single channel. However, this application goes far beyond the simple uses for multiplexers described here.

What Can Go Wrong

Problems that are common to all digital chips are summarized in the section on logic gates (see "What Can Go Wrong" on page 105).

Pullup Resistors

While they are often necessary to prevent connections from floating, pullup resistors built into a multiplexer may have unexpected consequences if the user is unaware of them.

Break Before Make

For most applications, it is desirable for each internal solid-state switch to break one connection before making a new one. This avoids the possibility of separate external components being briefly connected with each other through the multiplexer. Datasheets should be checked to verify that a multiplexer functions in break-before-make mode. If it doesn't, the enable pin can be used momentarily to disable all connections before a new connection is established.

Signal Distortion

Where a multiplexer is passing analog signals, signal distortion can result if the on-resistance of multiple internal switches varies significantly at different voltages. A datasheet for an analog multiplexer should usually include a graph showing on-resistance over the full signal range. The flatter the graph is, the less distortion the component will create. This is often described in datasheets as R_{ON} Flatness.

Limits of CMOS Switching

Although most multiplexers are built around CMOS transistors, their switching speed may be insufficient for video signals, and their on-resistance may vary enough to introduce distortion. Multiplexers are available with complementary bipolar switching for very high-speed applications. They impose some penalties in cost and power consumption.

Transients

Switch capacitance inside a multiplexer can cause transients in the output when the switch changes state. An allowance for *settling time* may be necessary. This will be additional to the switching speed claimed by the datasheet.

LCD

17

The full term *liquid-crystal display* is seldom used. Its acronym, **LCD**, is much more common. Sometimes the redundant combination **LCD display** is found. All three terms refer to the same device. In this encyclopedia, the first two words in *liquid-crystal display* are hyphenated because they are an adjectival phrase. Other sources often omit the hyphen.

The acronym **LED** (for **light-emitting diode**) is easily confused with **LCD**. While both devices display information, their mode of action is completely different.

OTHER RELATED COMPONENTS

- **LED display** (see Chapter 24)

What It Does

An LCD presents information on a small display panel or screen by using one or more segments that change their appearance in response to an AC voltage. The display may contain alphanumeric characters and/or symbols, icons, dots, or pixels in a *bitmap*.

Because of its very low power consumption, a basic monochrome LCD is often used to display numerals in battery-powered devices such as digital watches and calculators. A small liquid-crystal display of this type is shown in Figure 17-1.

Color-enabled, backlit LCDs are now frequently used in almost all forms of video displays, including those in cellular telephones, computer monitors, game-playing devices, TV screens, and aircraft cockpit displays.

How It Works

Light consists of electromagnetic waves that possess an electric field and a magnetic field. The fields are perpendicular to each other and to the direction in which the light is traveling, but the field polarities are randomly mixed in most visible radiation. This type of light is referred to as *incoherent*.

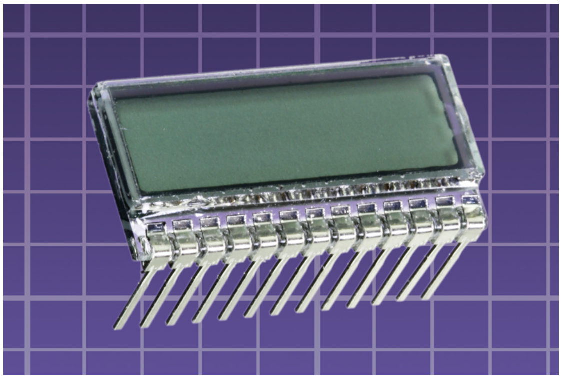

Figure 17-1. *A small, basic monochrome LCD.*

Figure 17-2 shows a simplified view of an LCD that uses a backlight. Incoherent light emerges from the backlight panel (A) and enters a vertical polarizing filter (B) that limits the electric field vector. The polarized light then enters a liquid crystal (C) which is a liquid composed of molecules organized in a regular helical structure that rotates the polarity by 90 degrees when no voltage is applied to it. The light now passes through a horizontal polarizing filter (D) and is visible to the user.

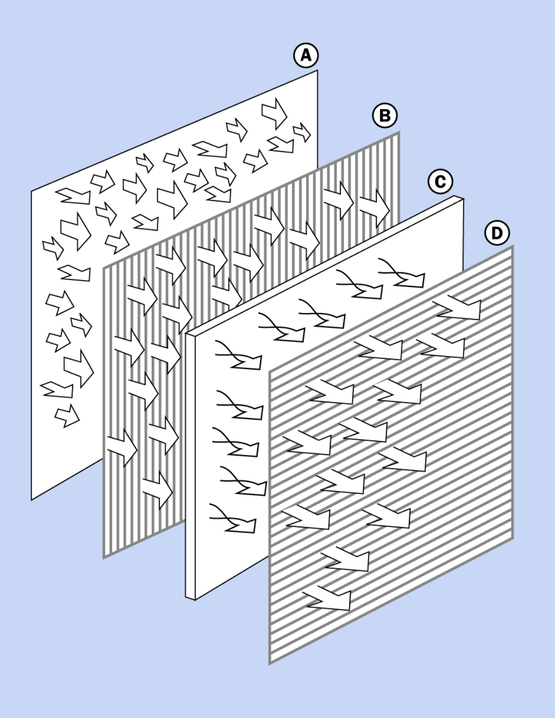

Figure 17-2. *The combination of two polarizers and a liquid crystal appears transparent when voltage is not applied. See text for details.*

- A liquid crystal itself does not emit light. It can only modify light that passes through it.

Figure 17-3 shows what happens when voltage is applied to the liquid crystal via transparent electrodes (not included in the figure). The molecules reorganize themselves in response to the electric potential and allow light to pass without changing its polarity. Consequently, the vertically polarized light is now blocked by the front, horizontally polarized filter, and the display becomes dark.

A liquid crystal contains ionic compounds that will be attracted to the electrodes if a DC voltage is applied for a significant period of time. This can degrade the display permanently. Therefore, AC voltage must be used. An AC frequency of 50Hz to 100Hz is common.

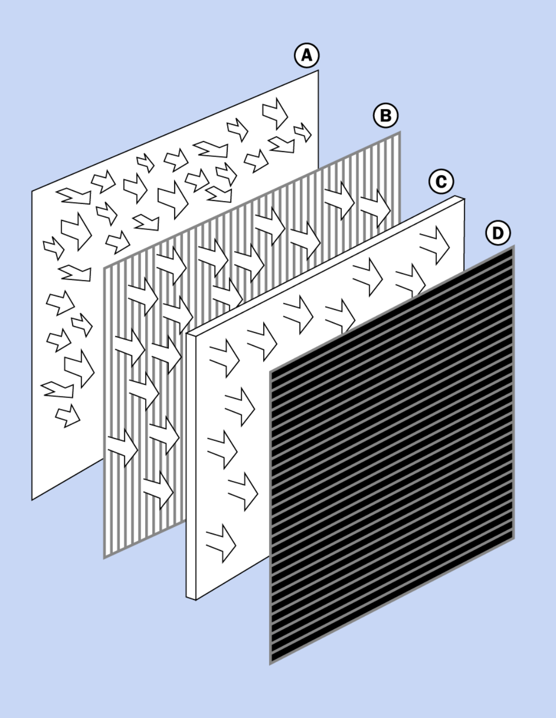

Figure 17-3. *The LCD appears dark when voltage is applied. See text for details.*

Variants

A *transmissive LCD* requires a backlight to be visible, and is the type illustrated in Figure 17-2. In its simplest form, it is a monochrome device, but is often enhanced to display full color by adding red, green, and blue filters. Alternatively, instead of a white backlight, an array of pixel-sized red, green, and blue **LEDs** may be used, in which case filters are unnecessary.

Backlit color LCDs have displaced *cathode-ray tubes*, which used to be the default system in almost all video monitors and TVs. LCDs are not only cheaper but can be fabricated in larger sizes. They do not suffer from *burn in*, where a persistent unchanging image creates a permanent scar in the phosphors on the inside of a tube. However, large LCDs may suffer from dead pixels or stuck pixels as manufacturing defects. Different manufacturers and vendors have varying policies regarding the maximum acceptable number of pixel defects.

In a *reflective LCD*, the structure is basically the same as that shown in Figure 17-2 except that a reflective surface is substituted for the backlight. Ambient light enters from the front of the display, and is either blocked by the liquid crystal in combination with the polarizing filters, or is allowed to reach the reflective surface at the rear, from which it reflects back through the liquid crystal to the eye of the user. This type of display is very easily readable in a bright environment, but will be difficult to see in dim conditions and will be invisible in darkness. Therefore, it may be augmented with a user-activated light source mounted at the side of the display.

A *transreflective LCD* contains a translucent rear polarizer that will reflect some ambient light, and is also transparent to enable a backlight. While this type of LCD is not as bright as a reflective LCD and has less contrast, it is more versatile and can be more energy efficient, as the backlight can be switched off automatically when ambient light is bright enough to make the display visible.

Active and Passive Types

An *active matrix* LCD adds a matrix of thin-film transistors to the basic liquid-crystal array, to store the state of each segment or pixel actively while the energizing AC voltage transitions from positive to negative. This enables a brighter, sharper display as *crosstalk* between adjacent pixels is reduced. Because *thin-film transistors* are used, this is often described as a *TFT* display; but the term is interchangable with *active matrix*.

A *passive matrix* LCD is cheaper to fabricate but responds sluggishly in large displays and is not so well suited to fine gradations in intensity. This type of component is used primarily in simple monochrome displays lacking intermediate shades of gray.

Crystal Types

Twisted Nematic (TN) are the cheapest, simplest type of LCD, allowing only a small viewing angle and average contrast. The appearance is limited to black on gray. The response rate is relatively slow.

Super Twisted Nematic (STN) displays were developed in the 1980s for passive LCDs, enabling better detail, wider view angle, and a faster response. The natural appearance is dark violet or black on green, or dark blue on silver-gray.

Film-compensated Super Twisted Nematic (FSTN) uses an extra coating of film that enables a pure black on white display.

Double Super Twisted Nematic provides further enhancement of contrast and response times, and automatic contrast compensation in response to ambient temperature. The appearance is black on white. This display requires backlighting.

Color Super Twisted Nematic (CSTN) is an STN display with filters added for full color reproduction.

Seven-Segment Displays

The earliest monochrome LCDs in devices such as watches and calculators used seven segments to display each numeral from 0 through 9. This type of LCD is still used in low-cost applications. A separate control line, or electrode, connects to each segment, while a *backplane* is shared by all the segments, connecting with a *common* pin to complete the circuit.

Figure 17-4 shows a typical seven-segment display. The lowercase letters *a* through *g* that identify each segment are universally used in datasheets. The decimal point, customarily referred to as "dp," may be omitted from some displays. The array of segments is slanted forward to enable more acceptable representation of the diagonal stroke in numeral 7.

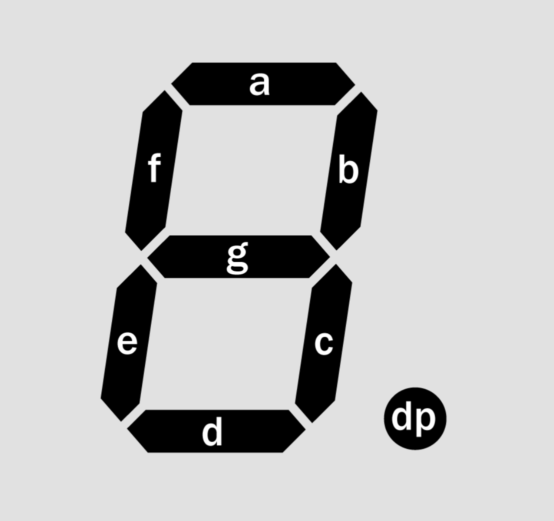

Figure 17-4. *Basic numeric display format for LCD numeric displays (the same layout is used with LEDs). To identify each segment, lowercase letters are universally used.*

Seven-segment displays are not elegant but are functional and are reasonably easy to read, as shown in Figure 17-5. Letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments) may be added to enable display of hexadecimal values.

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of 7-segment displays, as suggested in Figure 17-6.

The advantage of this system is low cost, as 7-segment displays are cheap to fabricate, entail the fewest connections, and require minimal decoding to create each alphanumeric character. However, numbers 0, 1, and 5 cannot be distinguished from letters O, I, and S, while letters containing diagonal strokes, such as K, M, N, W, X, and Z, cannot be displayed at all.

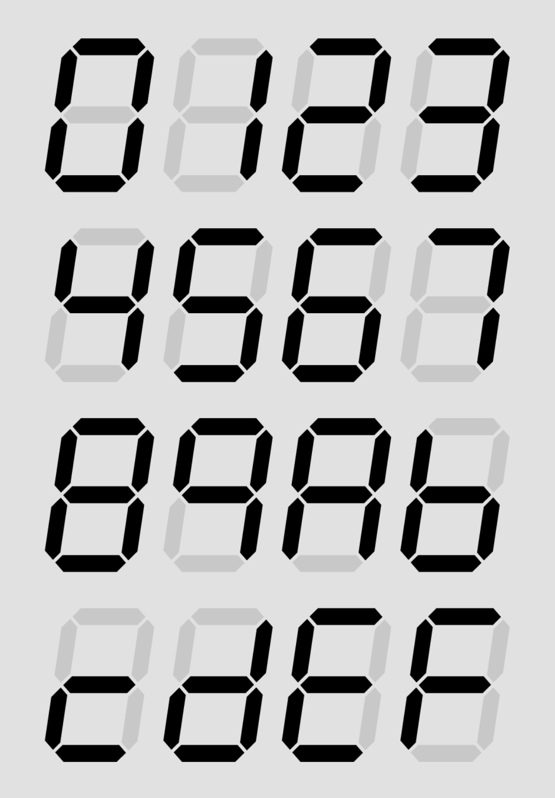

Figure 17-5. *Numerals and the first six letters of the alphabet created with a 7-segment display.*

Additional Segments

Alphanumeric LCDs were developed using 14 or 16 segments to enable better representation of letters of the alphabet. Sometimes these displays were slanted forward, like the 7-segment displays, perhaps because the style had become familiar, even though the addition of diagonal segments made it unnecessary. In other cases, the 14 or 16 segments were arrayed in a rectangle. See Figure 17-7.

The same words represented in Figure 17-6 are shown in Figure 17-8, using 16-segment LCDs. Clearly, the advantage gained by enabling diagonal strokes entailed the disadvantage of larger gaps in the letters, making made them ugly and difficult to read.

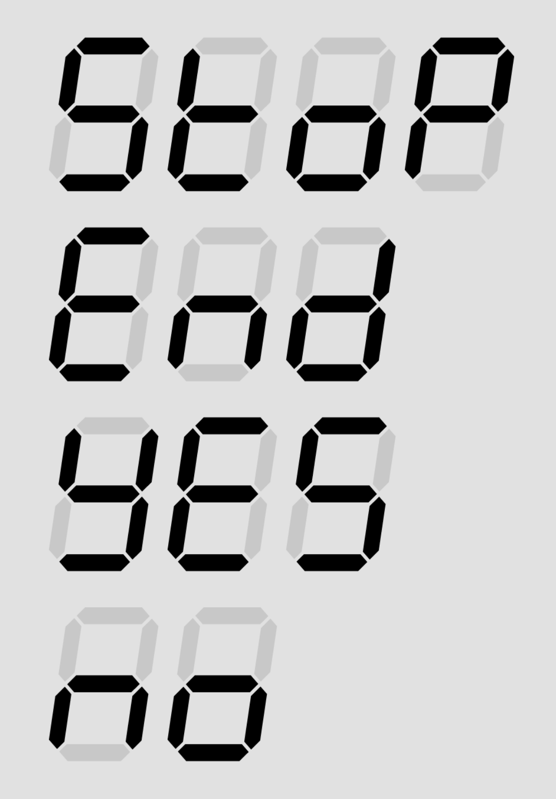

Figure 17-6. *Basic text messages can be generated with 7-segment displays, although they cannot contain alphabetical letters that use diagonal strokes.*

A full character set using 16-segment LCDs is shown in Figure 17-9. This conforms partially with the ASCII coding system, in which each character has an identifying numeric code ranging from 20 hexadecimal for a letter-space to 7A hexadecimal for letter z (although this character set does not attempt to represent lowercase letters differently from uppercase). The ASCII acronym stands for *American Standard Code for Information Interchange*.

Because backlit LCDs had become common by the time 16-segment displays were introduced, the characters were often displayed in light-on-dark or "negative" format, as suggested in this figure. LEDs, of course, have always used the light-on-dark format, as an LED is a light-emitting component.

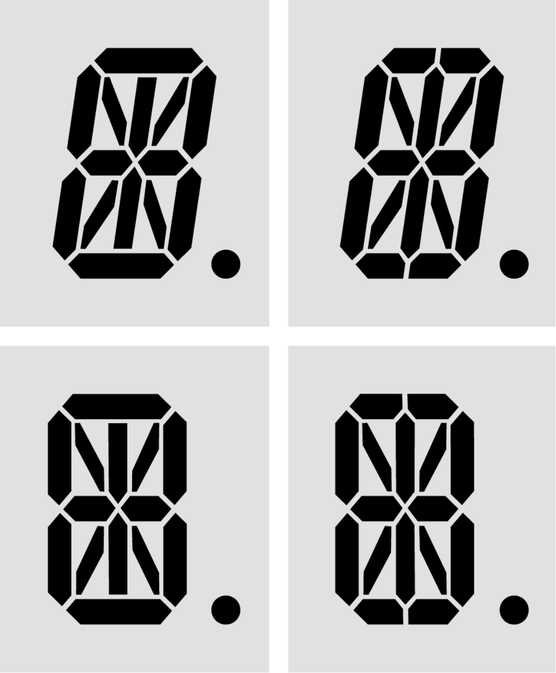

Figure 17-7. *LCDs using 14 segments (left) and 16 segments (right) were introduced to represent a full alphabet in addition to numerals. Sometimes these displays were slanted forward, like the previous 7-segment type, even though this was no longer necessary to represent the number 7.*

Dot-Matrix Displays

The 16-segment displays were never widely popular, and the declining cost of microprocessors, LCD fabrication, and ROM storage made it economic to produce displays using the more easily legible 5x7 dot-matrix alphabet that had been common among early microcomputers. Figure 17-10 shows a dot-matrix character set that is typical of many LCDs.

Because the original ASCII codes were not standardized below 20 hexadecimal or above 7A hexadecimal, manufacturers have represented a variety of foreign-language characters, Greek letters, Japanese characters, accented letters, or symbols using codes 00 through 1F and 7B through FF. The lower codes are often left blank, allowing user installation of custom symbols. Codes 00 through 0F are often reserved for control functions, such as a command to start a new line of text. There is no standardization in this area, and the user must examine a datasheet for guidance.

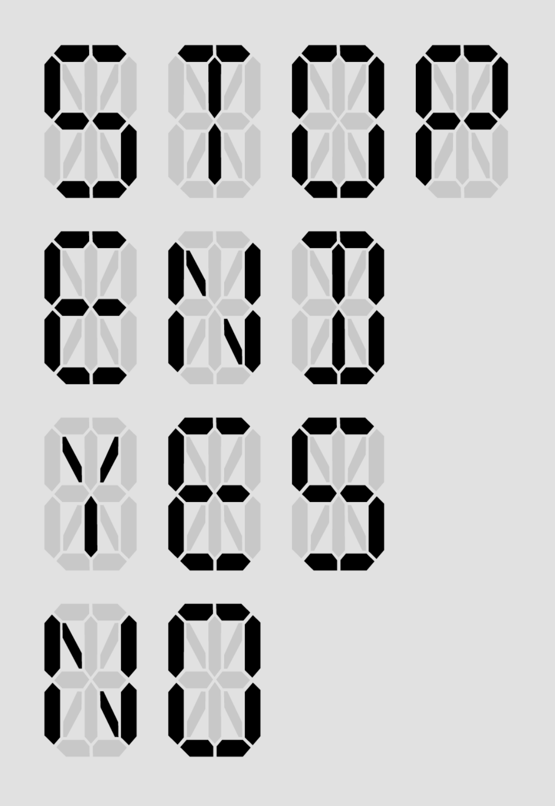

Figure 17-8. *The same text messages shown previously using 7-segment LCDs are shown here using 16-segment displays.*

Dot-matrix LCDs are usually packaged in arrays consisting of eight or more columns and two or more rows of characters. The number of columns is always stated before the number of rows, so that a typical 8 x 2 display contains eight alphanumeric characters in two horizontal rows. An array of characters is properly referred to as a *display module*, but may be described, confusingly, as a *display*, even though a single seven-segment LCD is itself a display. A 16x2 display module is shown from the front in Figure 17-11 and from the rear in Figure 17-12.

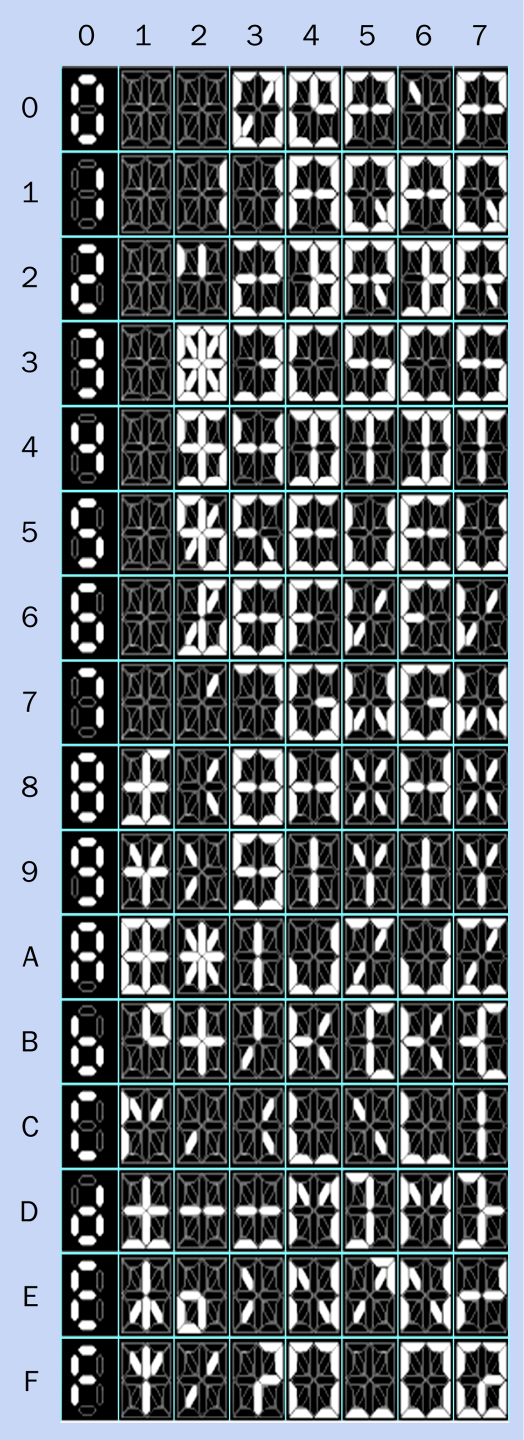

Figure 17-9. *A full character set using 16-segment LCDs.*

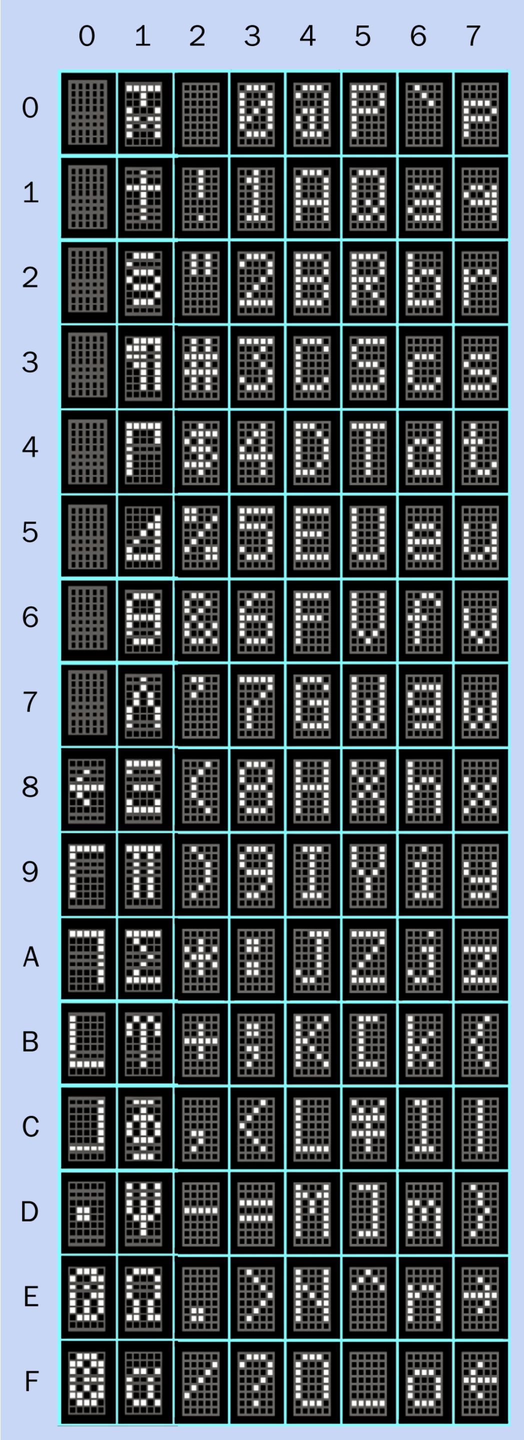

Figure 17-10. *A dot-matrix character set typical of LCDs capable of displaying a matrix of 5×7 dots.*

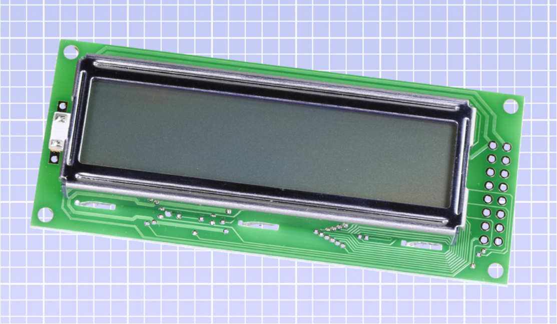

Figure 17-11. *A 16x2 LCD display module seen from the front.*

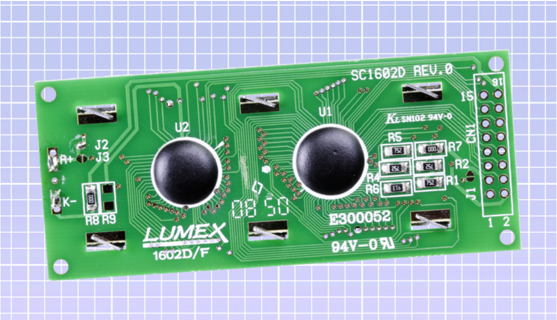

Figure 17-12. *The same 16x2 LCD display module from the previous figure, seen from the rear.*

Multiple-character display modules have been widely used in consumer electronics products such as audio components and automobiles where simple status messages and prompts are necessary—for example, to show the volume setting or broadcast frequency on a stereo receiver. Backlighting is almost always used.

Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, color displays are likely to displace monochrome dot-matrix LCD display modules in many applications. Similarly, touchscreens will tend to displace pushbuttons and tactile switches. Touchscreens are outside the scope of this encyclopedia.

Color

The addition of filters to create a full color display is shown in simplified form in Figure 17-13.

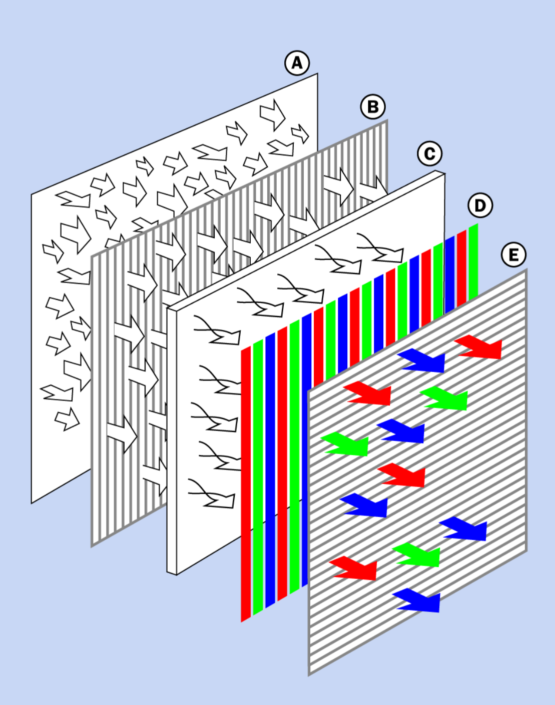

Figure 17-13. *The addition of red, green, and blue color filters, in conjunction with variable density liquid crystal pixels, enables an LCD full-color display.*

Red, green, and blue are almost always used as primary colors for transmitted light, because the combination of different intensities of these *RGB* primaries can create the appearance of many colors throughout the visible spectrum. They are said to be *additive* primaries, as they create brighter colors when they are combined. The principle is illustrated in Figure 17-14.

The use of the word "primaries" to refer to red, green, and blue can cause confusion, as full-color printed materials use a different set of *reflective* primaries, typically cyan, magenta, and yellow, often with the addition of black. In this *CMYK* system, additional layers of pigment will absorb, or subtract, more visible frequences. See Figure 17-15.

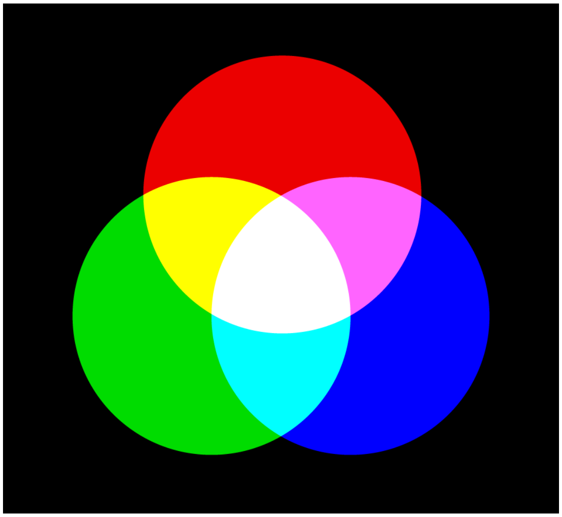

Figure 17-14. *When colors red, green, and blue are transmitted directly to the eye, pairs of these additive primaries create secondary colors cyan, magenta, and yellow. Combining all three additive primaries creates an approximation of white light. This can be verified by viewing a color monitor with a magnifying glass.*

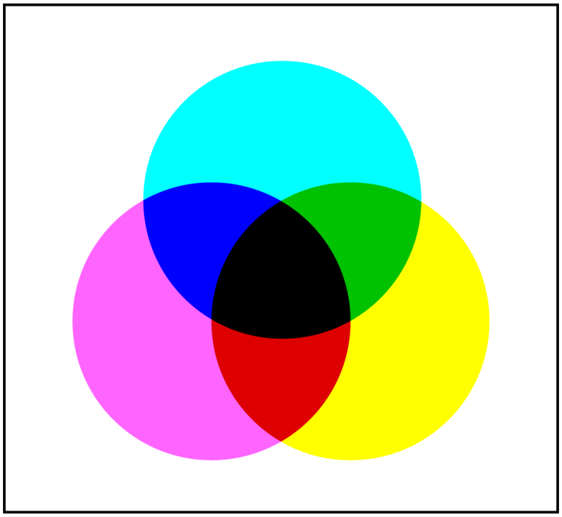

Figure 17-15. *When ink colors cyan, magenta, and yellow are superimposed on white paper and are viewed in white light, pairs of these subtractive primaries create secondary colors red, green, and blue. Overprinting all three subtractive primaries creates an approximation of black, limited by the reflective properties of available pigments. Black ink is usually added to provide additional contrast.*

The complete range of colors that can be created as a combination of primaries is known as the *gamut*. Many different RGB color standards have been developed, the two most widely used being *sRGB* (almost universal in web applications) and *Adobe 1998* (introduced by Adobe Systems for Photoshop, providing a wider gamut). None of the available systems for color reproduction comes close to creating the full gamut that can be perceived by the human eye.

Backlighting Options

For monochrome LCDs, **electroluminescent** backlighting may be used. It requires very low current, generates very little heat, and has a uniform output. However, its brightness is severely limited, and it requires an inverter that adds significantly to the current consumption.

For full-color LCDs, **fluorescent** lights were originally used. They have a long lifetime, generate little heat, and have low power consumption. However, they require a relatively high voltage, and do not work well at low temperatures. Early flat screens for laptop computers and desktop monitors used *cold-cathode fluorescent panels.*

Subsequently, white light-emitting diodes (**LED**s) were refined to the point where they generated a range of frequencies that was considered acceptable. Light from the LEDs passes through a diffuser to provide reasonably consistent illumination across the entire screen. LEDs are cheaper than fluorescent panels, and allow a thinner screen.

High-end video monitors use individual red, green, and blue LEDs instead of a white backlight. This eliminates the need for colored filters and produces a wider gamut. So-called *RGB* LCD monitors are more expensive but are preferred for professional applications in video and print media where accurate color reproduction is essential.

Zero-Power Displays

Some techniques exist to create LCDs that require power only to flip them to and fro between their transparent and opaque states. These are also known as *bistable* displays, but have not become as widely used. They are similar in concept to *e-ink* or *electronic paper* displays, but the principle of operation is different.

How to Use It

So long as an LCD consists of just one numeral, it can be driven by just one decoder chip that translates a binary-coded input into the outputs required to activate the appropriate segments of the LCD. The evolution of multi-digit displays, alphanumeric displays, dot-matrix displays, and graphical displays has complicated this situation.

Numeric Display Modules

An LCD consisting of a single digit is now a rare item, as few circuits require only one numeral for output. More commonly, two to eight numerals are mounted together in a small rectangular panel, three or four numerals being most common. A typical digital alarm clock uses a four-digit numeric display module, incorporating a colon and indicators showing AM/PM and alarm on/off. Other numeric display modules may include a minus sign.

Modules that are described as having 3.5 or 4.5 digits contain three full digits preceded by a numeral 1 composed of two segments. Thus, a 3-digit module can display numbers from 000 through 999, while a 3.5-digit display can display numbers from 000 through 1999, approximately doubling the range.

Numeric display modules of the type described here do not contain any decoder logic or drivers. An external device, such as a microcontroller, must contain a lookup table to translate a numeric value into outputs that will activate the appropriate segments in the numbers in a display, with or without decimal points and a minus sign. To avoid reinventing the wheel, a programmer may download code libraries for microcontrollers to drive commonly used numeric display modules. It is important to remember, though, that segments in monochrome LCDs must be activated by AC, typically a square wave with a frequency of 30Hz to 90Hz.

An alternative is to use a decoder chip such as the 4543B or 4056B, which receives a binary-coded decimal input (i.e., 0000 through 1001 binary, on four input puts) and translates it into an output on seven pins suitable for connection with the seven segments of a 7-segment display. The 4543B requires a square-wave input to its "phase" pin. The square-wave must also be applied simultaneously to the backplane of the LCD, often identified as the "common" pin on datasheets. Pinouts for the 4543B are shown in Figure 17-16.

The 4543B includes provision for "display blanking," which can be used to suppress leading zeros in a multidigit number. However, the lack of outputs to control a minus sign or decimal point limits the decoder to displaying positive integers.

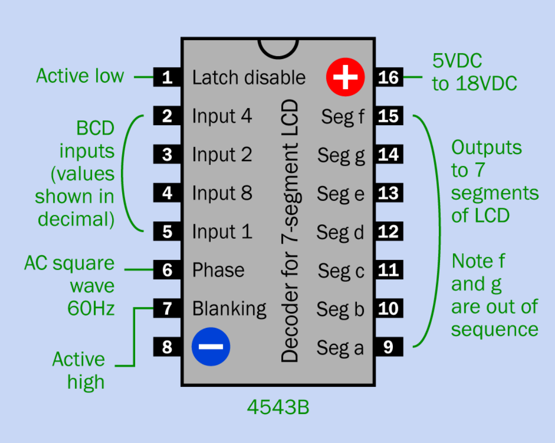

Figure 17-16. *Pinouts for the 4543B decoder chip, which is designed to drive a seven-segment numeric LCD.*

The power supply for a 4543B can range from 5VDC to 18VDC, but because the logic-high output voltage will be almost the same as that of the power supply, it must be chosen to match the power requirements of the LCD (very often 5VAC).(((

To drive a three-digit numeric display module, a separate decoder chip can be used to control each digit. The disadvantage of this system is that each decoder requires three inputs, so that a three-digit display will require nine outputs from the microcontroller.

To deal with this issue, it is common to *multiplex* a multi-digit display. This means that each output from the decoder is shared among the same segments of all the LCD numerals. Each LCD numeral is then activated in sequence by applying AC voltage to its common pin. Simultaneously, the decoder sends the data appropriate to that LCD. This process must be fast enough so that all the digits appear to be active simultaneously, and is best managed with a microcontroller. A simplified schematic is shown in Figure 17-17. It can be compared with a similar circuit to drive LED displays, shown in Figure 24-13.

Alphanumeric Display Module

Arrays of dot-matrix LCDs that can display alphabetical characters as well as numerals require preset character patterns (usually stored in ROM) and a command interpreter to process instructions that are embedded in the data stream. These capabilities are often built into the LCD module itself.

While there is no formal or de facto standard, the command set used by the Hitachi HD44780 controller is installed in many displays, and code libraries for this set are available for download from sites dedicated to the Arduino and other microcontrollers. Writing code from scratch to control all aspects of an alphanumeric display is not a trivial chore. The Hamtronix HDM08216L-3-L30S is a display that incorporates the HD44780.

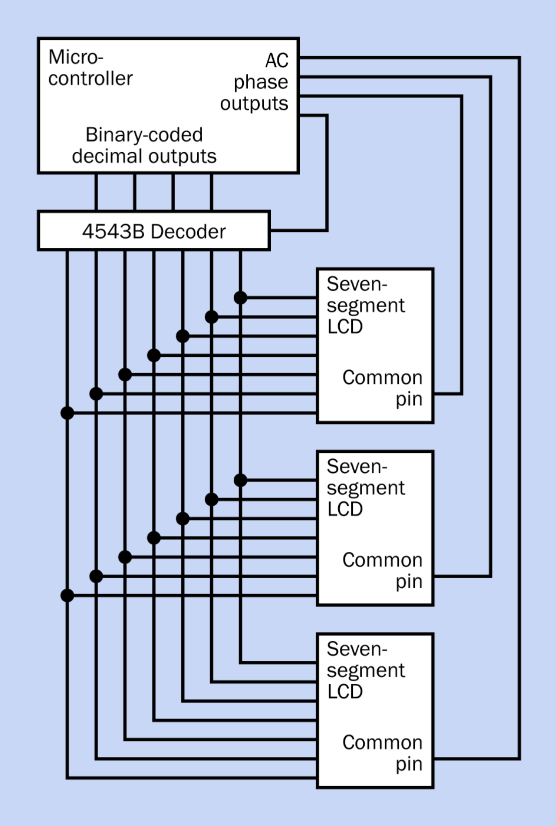

Figure 17-17. *When two or more numeric displays are multiplexed, a control device (typically, a microcontroller) activates each of them in turn via its backplane (common terminal) while sending appropriate data over a shared bus.*

Regardless of which standard is used, some features of alphanumeric display modules are almost universal:

- Register select pin. Tells the display whether the incoming data is an instruction, or a code identifying a displayable character.
- Read/write pin. Tells the display whether to receive characters from a microcontroller or send them to a microcontroller.
- Enable/disable pin.
- Character data input pins. There will be eight pins to receive the 8-bit ASCII code for each displayable character in parallel. Often there is an option to use only four of these pins, to reduce the number of microcontroller outputs necessary to drive the display. Where four pins are used, each 8-bit character is sent in two segments.
- LED backlight pin. Two may be provided, one connected to the anode(s) of the LED backlight, the other to the cathode(s).
- Reset pin.

Embedded instruction codes can be complex, including commands to reposition the cursor at a specific screen location, backspace-and-erase, scroll the display, and erase all characters on the screen. Codes may be included to adjust screen brightness and to switch the display between light-on-dark (negative) and dark-on-light (positive) characters.

Some display modules also have graphics capability, allowing the user to address any individual pixel on the screen.

Because of the lack of standardization in control codes, manufacturer's datasheets must be consulted to learn the usage of a particular alphanumeric display module. In addition to datasheets, online user forums are a valuable source of information regarding quirks and undocumented features.

What Can Go Wrong

Temperature Sensitivity

Liquid crystals vary in their tolerance for low and high temperatures, but generally speaking, a higher voltage may be necessary to create a sufficiently dense image at a low temperature. Conversely, a lower voltage may be necessary to avoid "ghosting" at a high temperature. An absolutely safe operating temperature range is likely to be 0 through 50 degrees Celsius, but check the manufacturer's datasheet for confirmation. Special-purpose LCDs are available for extreme temperatures.

Excessive Multiplexing

A twisted nematic display is likely to perform poorly if its duty cycle is greater than 1:4. In other words, more than four displays should not be multiplexed by the same controller.

DC Damage

An LCD can be damaged quickly and permanently if it is subjected to DC current. This can occur by accident if, for example, a timer chip is being used to generate the AC pulse stream, and the timer is accidentally disconnected, or has an incorrect connection in its RC network. Check timer output with a meter set to measure AC volts before allowing any connection to the common pin of an LCD.

Bad Communications Protocol

Many alphanumeric display modules do not use a formal communications protocol. Duplex serial or I2C connection may not be available. Care must be taken to allow pauses of a few milliseconds after execution of embedded commands, to give the display sufficient time to complete the instruction. This is especially likely where a command to clear all characters from the screen has to be executed. If garbage characters appear on the screen, incorrect data transfer speed or lack of pause times may be to blame.

Wiring Errors

This is often cited by manufacturers as the most common cause of failure to display characters correctly, or lack of any screen image at all.

incandescent lamp

18

The terms *incandescent light*, *incandescent bulb*, and *incandescent light bulb* are often used interchangeably with **incandescent lamp**. Because the term "lamp" seems to be most common, it is used here. A *panel-mounted indicator lamp* is considered to be an assembly containing an incandescent lamp.

A *carbon arc*, which generates light as a self-sustaining spark between two carbon electrodes, can be thought of as a form of incandescent lamp, but is now rare and is not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **LED area lighting** (see Chapter 23)
- **LED indicator** (see Chapter 22)
- **neon bulb** (see Chapter 19)
- **fluorescent light** (see Chapter 20)

What It Does

The term *incandescent* describes an object that emits visible light purely as a consequence of being hot. This principle is used in an **incandescent lamp** where a wire *filament* glows as a result of electric current passing through it and raising it to a high temperature. To prevent oxidation of the filament, it is contained within a sealed bulb or tube containing an inert gas under low pressure or (less often) a vacuum.

Because incandescent lamps are relatively inefficient, they are not considered a wise environmental choice for area lighting and have been prohibited for that purpose in some areas. However, small, low-voltage, panel-mount versions are still widely available. For a summary of advantages of miniature incandescent lamps relative to **light-emitting diodes** (LEDs) see "Relative Advantages" on page 179.

Schematic symbols representing an incandescent lamp are shown in Figure 18-1. The symbols are all functionally identical except that the one at bottom right is more likely to be used to represent small panel-mounted indicators.

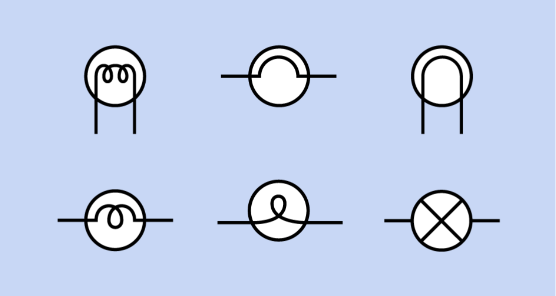

Figure 18-1. *A variety of symbols can represent an incandescent lamp. The one at bottom right may be more commonly used for small panel-mounted indicators.*

The parts of a generic incandescent light bulb are identified in Figure 18-2:

A: Glass bulb.

B: Inert gas at low pressure.

C: Tungsten filament.

D: Contact wires (connecting internally with brass base and center contact, below).

E: Wires to support the filament.

F: Internal glass stem.

G: Brass base or cap.

H: Vitreous insulation.

I: Center contact.

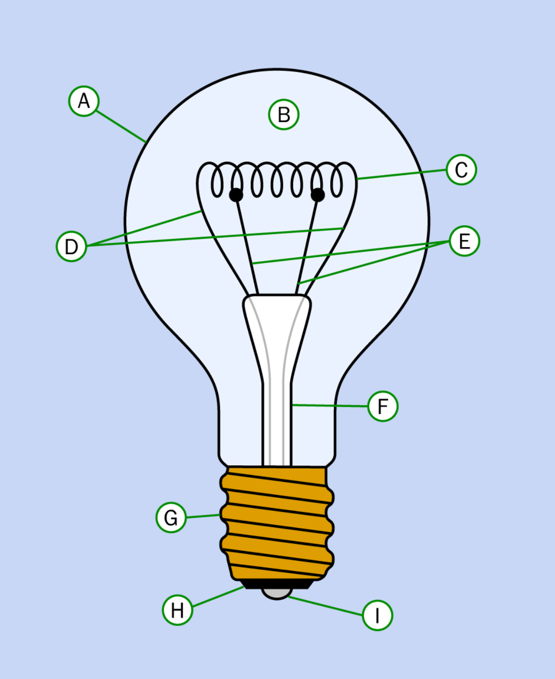

Figure 18-2. *The parts of a typical incandescent lamp (see text for details).*

History

The concept of generating light by using electricity to heat a metal originated with Englishman Humphrey Davy, who demonstrated it with a large battery and a strip of platinum in 1802. Platinum was thought to be suitable because it has a relatively high melting point. The lamp worked but was not practical, being insufficiently bright and having a short lifespan. In addition, the platinum was prohibitively expensive.

The first patent for an incandescent lamp was issued in England in 1841, but it still used platinum. Subsequently, British physicist and chemist Joseph Swan spent many years attempting to develop practical carbon filaments, and obtained a patent in 1880 for *parchmentized thread*. His house was the first in the world to be illuminated by light bulbs.

Thomas Edison began work to refine the electric lamp in 1878, and achieved a successful test with a carbonized filament in October 1879. The bulb lasted slightly more than 13 hours. Lawsuits over patent rights ensued. Carbonized filaments were used until a tungsten filament was patented in 1904 by the German/Hungarian inventor Just Sándor Frigyes and the Croatian inventor Franjo Hanaman. This type of bulb was filled with an inert gas, instead of using a vacuum.

Many other pioneers participated in the effort to develop electric light on a practical basis. Thus it is incorrect to state that "Thomas Edison invented the light bulb." The device went through a very lengthy process of gradual refinement, and one of Edison's most significant achievements was the development of a power distribution system that could run multiple lamps in parallel, using filaments that had a relatively high resistance. His error was insisting on using *direct current* (DC) while his rival Westinghouse pioneered *alternating currrent* (AC), enabling power transmission over longer distances through the use of **transformers**. The use of AC also enabled Tesla's brushless induction motor.

By the mid-1900s, most incandescent bulbs used tungsten filaments.

How It Works

All objects emit electromagnetic radiation as a function of their temperature. This is known as *black body radiation*, based on the concept of an object that absorbs all incoming light, and thus does not reflect any sources from outside itself. As its temperature increases, the intensity of the radiation increases while the wavelength of the radiation tends to decrease.

If the temperature is high enough, the wavelength of the radiation enters the visible spectrum, between 380 and 740 nanometers. (A nanometer is one-billionth of a meter.)

The melting point of tungsten is 3,442 degrees Celsius, but a lamp filament typically operates between 2,000 and 3,000 degrees. At the higher end of this scale, evaporation of metal from the filament tends to cause deposition of a dark residue on the inside of the bulb, and erodes the filament more rapidly, to the point where it eventually breaks. At the lower end of this scale, the light will be yellow and the intensity will be reduced.

Spectrum

The color of black-body radiation is measured using the Kelvin temperature scale. The increment of 1 degree Kelvin is the same as 1 degree Celsius, but the Kelvin scale has a zero value at *absolute zero*. This is the theoretical lowest conceivable temperature, at which there is complete absence of heat. It is approximately –273 degrees Celsius.

From this it is evident that if K is a temperature in degrees Kelvin and C is a temperature in degrees Celsius:

`K = C + 273` (approximately)

Calibration of light sources in degrees Kelvin is common in photography. Many digital cameras allow the user to specify the *color temperature* of lights that are illuminating an indoor scene, and the camera will compensate so that the light source appears to be pure white with all colors in the visible spectrum being represented equally.

Some computer monitors also allow the user to specify a white value in degrees Kelvin.

Color temperature is used in astronomy, because the spectrum of many stars is comparable with that of a theoretical black body.

A color temperature of 1,000 degrees K will have a dark orange hue, while 15,000 degrees K or higher will have a blue hue comparable to that of a pale blue sky. The color temperature of the sun is approximately 5,800 K. Interior lighting is often around 3,000 K, which many people find acceptable because it creates pleasant flesh tones. An incandescent bulb described by the manufacturer as "soft white" or "warm" will have a lower color temperature than one which is sold as "pure white" or "paper white."

Graphs showing the emission of wavelengths at various color temperatures are shown in Figure 18-3. The rainbow section indicates the approximate range of visible wavelengths between ultraviolet, on the left, and infrared, on the right. For purposes of clarity, the peak intensity for each color temperature has been equalized. In reality, increasing the temperature also increases the light output.

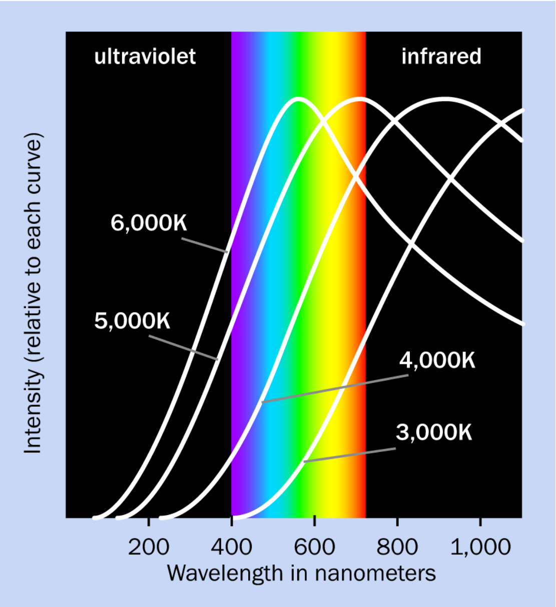

Figure 18-3. *Approximate peak wavelengths for black-body radiation at various color temperatures in degrees Kelvin. The curves have been adjusted so that their peak values are equalized. Adapted from an illustration in the reference book Light Emitting Diodes by E. Fred Schubert.*

Non-Incandescent Sources

So long as light is generated by heating a filament, plotting the intensity against wavelength will result in a smooth curve without irregularities. A higher Kelvin value will simply displace and compress the curve laterally without changing its basic shape to a significant degree.

The introduction of **fluorescent** sources and, subsequently, light-emitting diodes (**LEDs**) has complicated this scenario. Because they are *luminescent* rather than *incandescent*, they do not generate an evenly weighted, continuous range of wavelengths.

LEDs tend to emit monochromatic light, meaning that it is tightly centered around just one color. A "white" LED is really a blue LED in which a phosphor coating on the semiconductor die is excited to create light over a broader range. A fluorescent light tends to create *spectral lines* which show up as sharp peaks at a few wavelengths determined by the mercury inside the bulb. Figure 18-4 illustrates these problems.

The human eye tends to compensate for the yellow emphasis of incandescent lamps and for the irregularities in spectra emitted by other light sources. Also, the eye is often unable to distinguish between "white" light created as a mix of all the visible wavelengths, and light that appears white even though it is dominated by a few isolated wavelengths from a fluorescent source.

However, when the eye views colors that are illuminated by a source that has gaps in its spectrum, some of the colors will appear unnaturally dull or dark. This is true also if an imperfect source is used as a backlight to create colors on a video monitor. Colors rendered by different light sources are shown in Figure 23-7 and subsequent figures.

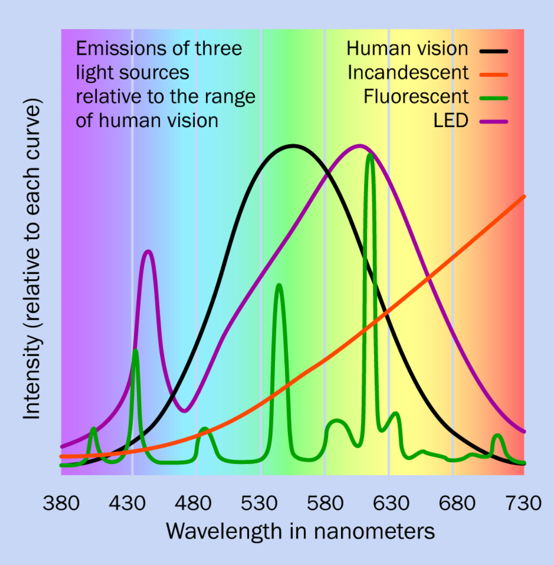

Figure 18-4. *The relative performance of three light sources compared with sensitivity of the human eye to the visible spectrum. Note that the range of wavelengths on the horizontal scale in this figure is not the same as the range in the previous figure. The color assigned to each curve is arbitrary. Adapted from VU1 Corporation.*

Photography is adversely affected by the use of LEDs or fluorescents as a light source. Reds, for example, can seem dark when lit by white LEDs, while blues can be inappropriately intense. Because the source does not have an emission curve comparable to that of an incandescent light, the auto-white balance feature of a digital camera may be unable to address this problem, and it cannot be resolved by entering a different Kelvin number manually.

The fidelity with which a light source is capable of displaying the full visible spectrum is known as the *color rendering index* (CRI), ranging from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected "white" LED may score as low as 80.

Power Consumption

Approximately 95% of the power consumed by an incandescent lamp generates heat instead of visible light. This wastage of power in room lighting is compounded by the power consumption of air conditioning to remove the heat from enclosed spaces in hot climates. While the heat from incandescent lamps does reduce the need for space heating in cold environments, heat is delivered more efficiently by using systems designed for that purpose. Consequently, greater energy efficiency can be achieved with a light source that generates less heat, regardless of ambient air temperature.

Variants

Miniature Lamps

Prior to the development of **LED**s, all light-emitting panel-mounted indicators were either **neon bulbs** or **incandescent lamps**. The use of neon is limited by its need for a relatively high voltage.

Miniature incandescents were the traditional choice for battery-powered light sources, and at the time of writing are still used in cheap flashlights. Variants are available that are as small as a 5mm LED, with a claimed life expectancy that is comparable, although they draw more current to generate an equivalent light intensity, because much of their power is wasted in infrared wavelengths.

The photograph in Figure 18-5 is of a miniature lamp terminating in pins spaced 0.05" apart. The total height of the lamp, including its ceramic base, is less than 0.4," while its diameter is just over 0.1". It draws 60mA at 5V and is rated for 25,000 hours.

The photograph in Figure 18-6 is of a lamp of similar size and power consumption, but terminating in wire leads and rated for 100,000 hours. It emits 0.63 lumens.

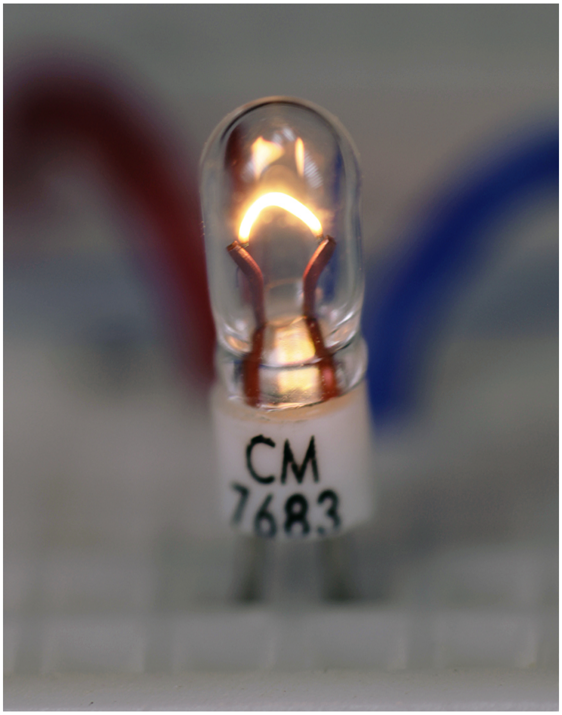

Figure 18-5. *A miniature lamp less than 0.4" high, terminating in pins spaced 0.05" apart.*

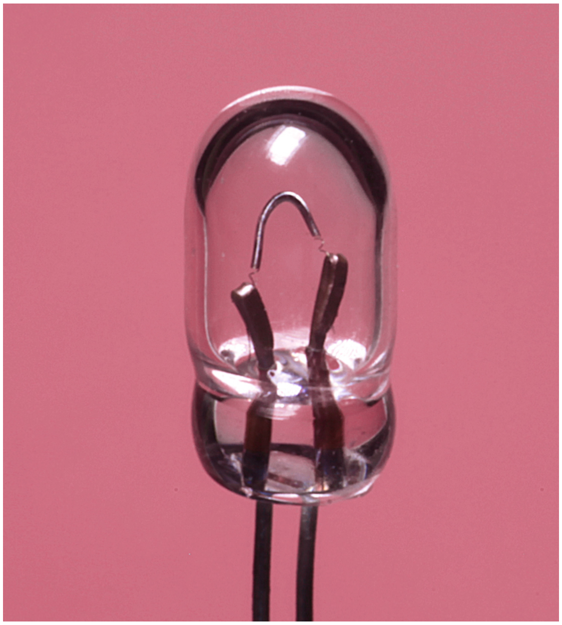

Figure 18-6. *This lamp is 0.25" high and terminates in wire leads.*

The lamp pictured in Figure 18-7 is slightly larger, with a glass envelope about 0.25” diameter. It is rated for less than half the lifetime of the lamp in Figure 18-6 but emits three times as much light —a typical tradeoff. Various base styles are available.

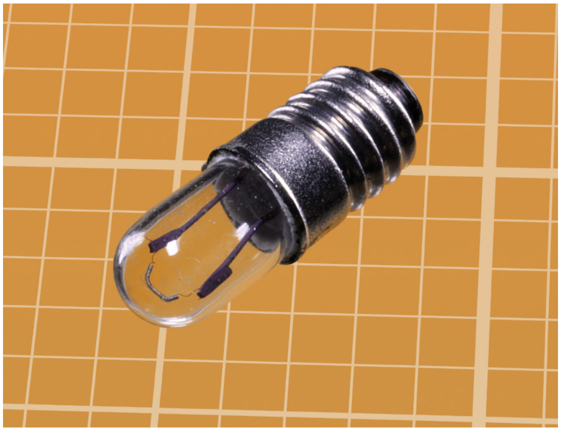

Figure 18-7. *This lamp has a glass envelope about 0.35” high. Its screw-in base makes it easier to replace than an LED.*

In the United States, the light output from miniature incandescent lamps may be measured in lumens, but is more often rated in *mean spherical candlepower* (MSCP). An explanation of light measurement is included in “MSCP” on page 178.

Lamp lenses provide a quick and simple way to add color to a miniature incandescent lamp. Usually the lens is cylindrical with a hemispherical end cap, and is designed to push-fit or snap-fit over a small lamp. Even when the cap is translucent, it may still be referred to as a lens.

Panel-Mount Indicator Lamps

This term often refers to a tubular assembly containing a miniature lamp, ready for installation.

The enclosure is often designed to snap-fit into a hole drilled in the panel. If the incandescent bulb inside the enclosure cannot be replaced, the component is said to be “non-relampable.” Figure 18-8 shows a 12-volt panel-mount indicator lamp.

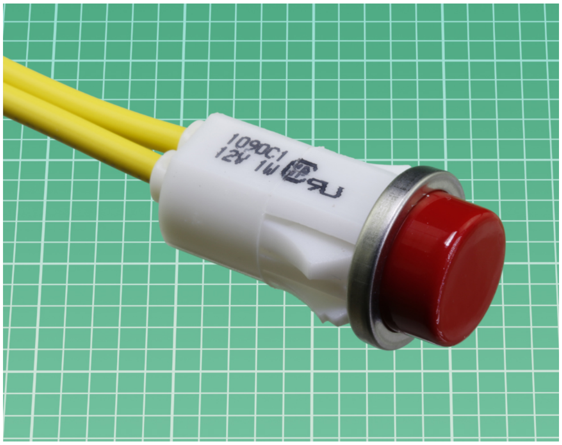

Figure 18-8. *This panel-mount indicator lamp is designed to push-fit into a hole 1/2” in diameter. The bulb inside it is not replacable, causing the assembly to be classified as “non-relampable.”*

Halogen or Quartz-Halogen

This is a type of incandescent lamp containing gases under pressure in which halogens such as iodine or bromine cause evaporated tungsten atoms to be redeposited on the filament. A halogen lamp can therefore operate at a higher temperature, creating a light that is less yellow and brighter than that from a comparable incandescent lamp. It also enables a smaller bulb, but requires an envelope of borosilicate-halide glass (often termed *fused quartz*) instead of regular glass. A halogen lamp will be slightly more efficient than an incandescent bulb of the same wattage, and will last longer.

Halogens are available in a variety of formats. The small bulb pictured in Figure 18-9 consumes 75W, emitting 1,500 lumens at 3,000 degrees Kelvin. The light intensity is claimed to be equivalent to that of a 100W incandescent bulb. It has a mini-candelabra base.

Oven Lamps

Oven lamps are designed to withstand the high temperature in an oven. Typically they are usable with ambient temperatures up to 300 degrees C. A common power rating is 15W.

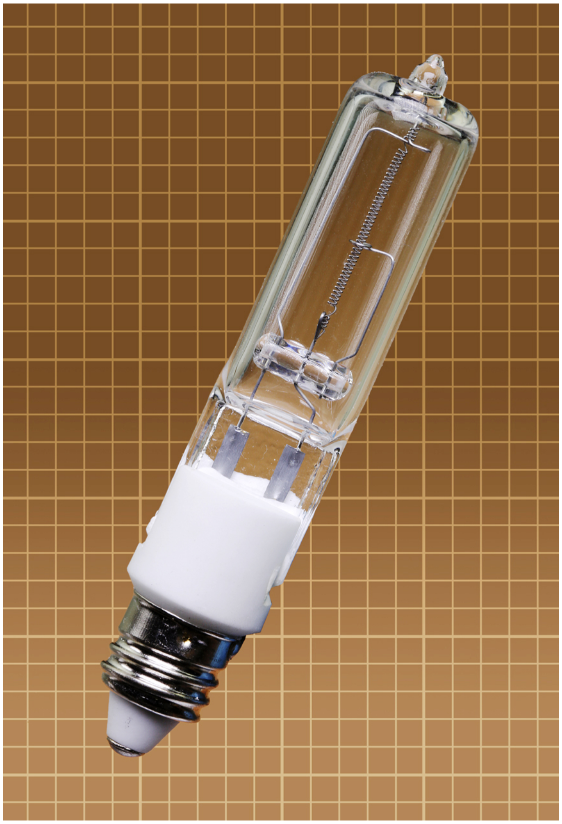

Figure 18-9. *A halogen bulb slightly more than 2" in length, designed for 115VAC.*

Base Variants

Miniature lamps are available with a wide variety of connection options, including wire terminals, single-contact bayonet, double-contact bayonet, miniature screw base, and fuse style. Most of these options require a matching socket.

Screw-in lamps for room illumination are common in household lighting in the United States and many other countries (but not in the UK, where bayonet fittings are used). The US socket size is designated by letter E followed by a number that gives the socket diameter in millimeters. Common sizes are E10, E14 and E27.

A *bayonet base* is fitted with two small lugs protruding on opposite sides. The lamp is secured by pushing it in and twisting it to engage the lugs in slots in the socket. The advantage of a bayonet base is that the bulb is less likely to become loose as a result of vibration.

A *pin base* consists simply of a pair of pins that will push-fit into small holes in a socket.

A *flange base* has a flange that engages in a socket where flexible segments will retain it.

A *wedge base* is forced between two contacts which retain the bulb by friction.

Some indicator lamps terminate simply in long, thin leads that can be soldered.

Values

While the power consumption of full-size incandescent lamps is rated in watts, small indicator lamps are rated in milliamps at the voltage for which they are designed. Miniature lamps may require specific voltages ranging from as low as 2V to 24V. A higher voltage generally necessitates a longer filament, which may entail a larger bulb.

The light that a lamp will emit can be measured in two ways: either as the power of the lamp (not its power consumption, but its radiating power), or as the light delivered to a specific area at a specific distance. These two measurements may differ because a lamp may concentrate its light in a beam, as in the case of a reflector bulb or an LED.

Power

Flux, in watts, is a measurement of energy flow in joules per second. The total radiating power of a lamp, in all wavelengths, in all directions, is known as its *radiant flux*. Because invisible wavelengths are of little interest when assessing the brightness of a lamp, the term *luminous flux* is used to describe the apparent brightness of the lamp in the visible spectrum. The unit for luminous flux is the *lumen*.

The human eye is most responsive to yellow-green hues in the center of the spectrum. Consequently, the measurement of luminous flux is weighted toward green at a wavelength of 555 nanometers. Red and violet are considered to have low luminous flux, while infrared and ultraviolet have a zero value.

When considering a value expressed in lumens, remember:

- Lumens are a measure of the total radiated power output of a light source, in all directions, in the visible spectrum only, weighted toward the characteristics of the human eye.
- The number of lumens of a light source does not define the direction in which the light is shining, or its uniformity.
- The abbreviation for lumen is *lm*.

A conventional incandescent lamp that consumes 100W of electricity is likely to have a light output of about 1,500 lumens. A 40W fluorescent tube can have a light output of about 2,600 lumens.

Illuminance

The *illuminance* of a light source is defined as the luminous flux per unit of area. This can be thought of as the brightness of a surface illuminated by the source.

Illuminance is measured in *lux*, where 1 lux = 1 lumen per square meter. For accurate calibration, the illuminated surface should be spherical in shape, and must be located 1 meter from the light source, with the source at the geometrical center of the sphere.

Illuminance used to be measured in *foot-candles*, where 1 foot-candle was 1 lumen per square foot.

- The number of lumens per square meter (lux) does not define the size of the illuminated area, only the brightness per unit of area.
- A lamp that has a tightly focused beam can achieve a high lux rating. When selecting a lamp for an application, the angle of dispersion of the beam must be considered in conjunction with its lux rating.

Intensity

A *candela* measures the luminous flux within an angle of dispersion. The angle is three-dimensional, and can be imagined as the sharpness of a point of a cone, where the light source is at the point and the cone represents the dispersion of light.

The three-dimensional angle of dispersion is measured in *steradians*. If a light source is at the center of a sphere that has a radius of 1 meter, and is illuminating one square meter of the surface of the sphere, the angle of dispersion is 1 steradian.

- A source of 1 lumen which projects all its light through a dispersion angle of 1 steradian is rated at 1 candela.
- The number of candelas does not define the angle of dispersion, only the intensity within that angle.
- A light source rated for 1,000 candelas could have a power of 10 lumens concentrated within an angle of 0.01 steradians, or could have a power of only 1 lumen concentrated within an angle of 0.001 steradians.
- There are 1,000 millicandelas in 1 candela. The abbreviation for candela is *cd* while the abbreviation for a millicandela is *mcd*.
- LEDs are often rated in mcd. The number describes the intensity of light within its angle of dispersion.

MSCP

Although the term *candlepower* is obsolete, it has been redefined as being equal to 1 candela. *Mean spherical candlepower* (MSCP) is a measurement of all the light emitted from a lamp in all directions. Because the light is assumed to be omnidirectional, it fills 4 * π (about 12.57) steradians. Therefore 1 MSCP = approximately 12.57

lumens. In the United States, MSCP is still the most common method of rating the total light output of a miniature lamp.

Efficacy

The *radiant luminous efficacy* (abbreviated *LER*) assesses how effective a lamp is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. LER is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths.

Thus, if VP is the power emitted in the visible spectrum, and AP is the power emitted in all wavelengths:

```
LER = VP / AP
```

LER is expressed in *lumens per watt*. It can range from a low value of around 12 lm/W for a 40W incandescent bulb to 24 lm/W for a quartz halogen lamp. Fluorescent lamps may average 50 lm/W. LEDs vary, but can achieve 100 lm/W.

Efficiency

The *radiant luminous efficiency* (abbreviated *LFR*) of a lamp measures how good its radiant luminous efficacy is, compared with an imaginary ideal lamp. (Note the difference between the words "efficiency" and "efficacy.") LFR is determined by dividing the radiant luminous efficacy (LER) by the maximum theoretical LER value of 683 lm/W, and multiplying by 100 to express the result as a percentage. Thus:

```
LFR = 100 * ( LER / 683 )
```

The LFR ranges from around 2% for a 40W bulb to 3.5% for a quartz halogen lamp. LEDs may be around 15% while fluorescents are closer to 10%.

How to Use It

When first introduced, LEDs were limited by their higher price, lower maximum light output, and inability to display blue or white. The price difference has disappeared for small indicators, while gaps in the color range have been filled (although the color rendering index of LEDs is still inferior).

Brightness remains an advantage for large incandescents relative to LEDs, as they are more upwardly scalable. However, fluorescents and vapor lamps have an advantage for very high light output, as in the lighting of big-box stores or parking lots. Thus the range of applications for incandescent bulbs is diminishing, especially because common types are now illegal for domestic light fixtures in many parts of the world.

Relative Advantages

When choosing whether to use an incandescent lamp or an LED, these advantages of an incandescent lamp should be considered:

- The intensity can be adjusted with a **triac**-based dimmer. Regular fluorescents cannot be dimmed, while LEDs often require different dimmer circuitry.
- The intensity can also be adjusted with a rheostat. The output from fluorescents cannot.
- Easy white-balance correction. LEDs and fluorescents do not naturally produce a consistent output over the visible spectrum.
- Can be designed to operate directly from a wide range of voltages (down to around 2V and up to around 300V). A higher voltage entails a longer filament wire, which may require a larger bulb. LEDs require additional components and circuitry to use higher voltages.
- Incandescent bulbs are more tolerant of voltage fluctuations than LEDs. With battery operation, the incandescent will still provide some reduced light output when the voltage has diminished radically. LEDs will not perform at all at currents lower than their threshold.
- An incandescent is nonpolarized and may be socketed, which simplifies user replacement.

LEDs are polarized and are usually soldered in.

- Can be powered by AC or DC without any modification or additional circuitry. LEDs require DC, which must be provided through a transformer and rectifier, or similar electronics, if AC power is the primary source.
- Can be equally visible from a wide range of viewing angles. LEDs have restricted viewing angles.
- The heat output from an incandescent bulb may occasionally be useful (for example in a terrarium, or in incubators for poultry).
- Trouble-free switching. Fluorescents tend to hesitate and blink when power is applied, and they require a *ballast* to energize them. The lifespan of fluorescents is reduced by frequent switching.
- No low-temperature problems. Incandescent lamps are not significantly affected by low temperatures. Fluorescents may not start easily in a cold environment, and may flicker or glow dimly for 10 minutes (or more) until they are warm enough to function properly.
- Easy disposal. Fluorescent lights contain small quantities of mercury that are an environmental hazard. They should not be mixed with ordinary trash. Compact fluorescent lamps (CFLs) and LEDs used for room lighting will be packaged with electronics that should ideally be recycled, although this is not very practical. Incandescent bulbs impose the least burden on the environment when they are thrown away.

However, the incandescent lamp has some obvious disadvantages:

- Relatively inefficient.
- More susceptible to vibration.
- More fragile.
- Likely to have a shorter natural life expectancy than LEDs, fluorescents, or neon bulbs, although the lifetime of a small panel indicator can be equal to that of an LED if a low color temperature is acceptable.
- Requires a filter or tinted glass envelope to generate colored light. This further reduces the lamp's efficiency.
- Cannot be miniaturized to the same degree as an LED indicator.

Derating

The lifespan of a lamp can be greatly extended by choosing one with a higher current rating or using it at a lower voltage. The light output will be reduced, and the color temperature will be at a lower Kelvin number, but in some situations this tradeoff may be acceptable.

The graphs in Figure 18-10 suggest that if the voltage of a hypothetical miniature lamp is reduced to 80% of the manufacturer's recommended value, this can make the lamp last 20 times as long. Note, however, that this will cut the light intensity to 50% of its normal value.

Conversely, using 130% normal voltage will give 250% of the normal light output, while shortening the life of the lamp to 1/20 of its normal value. Naturally these figures are approximations that may not apply precisely to a specific lamp.

What Can Go Wrong

High Temperature Environment

If an incandescent lamp is used in an environment hotter than 100 degrees Celsius, the life of the lamp is likely to be reduced by the "water cycle." Any water molecules inside the glass envelope will break down, allowing oxygen to combine with the tungsten filament to form tungsten oxide. The tungsten is deposited on the inside of the glass while the oxygen is liberated and begins a new cycle.

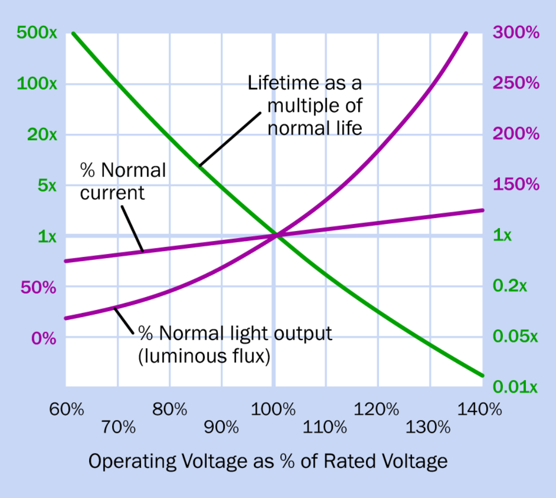

Figure 18-10. *The life expectancy of a hypothetical miniature lamp is very strongly influenced by voltage. Applying only 60% of the rated voltage can make a lamp last 500 times its normal lifespan, although it will greatly reduce light output. Note that the vertical axes apply to curves of the same color. Adapted from "Characteristics of Miniature Lamps" from Toshiba Lighting and Technology Corporation.*

Fire Risk

The partially evacuated bulb of an incandescent lamp provides some separation and protection from the heat in the filament, but if the bulb cannot disperse heat by radiation or convection, its temperature can rise to the point where it ignites flammable materials.

Halogen lamps have an elevated fire risk because they operate at a higher temperature and are smaller, providing less surface area to disperse the heat. They also contain gases under seven to eight atmospheres of pressure. Thermal stress can cause a halogen bulb to shatter, and fingerprints on the glass can increase this risk.

Current Inrush

When an incandescent lamp is first switched on, its filament has one-tenth the resistance that it will exhibit when it becomes hot. Consequently, the lamp will take a large initial surge of current, which stabilizes after about 50 milliseconds. This should be considered if one or more small lamps shares a DC power supply with components such as logic chips that may be sensitive to voltage fluctuations.

Replacement Problems

Because of the limited life of incandescent lamps, they should be installed in such a way that they are easy to replace. This can be an issue with panel indicators, where disassembly of a device may be necessary to reach the lamp.

The range of small incandescent lamps is diminishing, and may continue to diminish in the future. Future availability of replacement lamps should be considered when designing a circuit. When building equipment in small quantities, spare lamps should be purchased for future use.

19

neon bulb

The terms **neon bulb**, *neon indicator*, and *neon lamp* tend to be used interchangeably. In this encyclopedia, a neon *bulb* is defined as a glass capsule containing two electrodes in neon gas (or a combination of gases in which neon is present). A neon *lamp* is an assembly containing a neon bulb, usually using a plastic tube with a tinted transparent cap at one end. A neon *indicator* is a miniature neon lamp that is usually panel-mounted.

Large-scale neon tubes used in signage are not included in this encyclopedia.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see Chapter 18)
- **fluorescent light** (see Chapter 20)
- **LED indicator** (see Chapter 22)

What It Does

When voltage is applied between two electrodes inside a neon bulb, the inert gas inside the bulb emits a soft red or orange glow. This color may be modified by using a tinted transparent plastic cap, known as a *lens*, in a *neon lamp* assembly.

A neon bulb is usually designed for a power supply of 110V or higher. It functions equally well with alternating or direct current.

The schematic symbols in Figure 19-1 are commonly used to represent either a neon bulb or a neon lamp. They are all functionally identical. The black dot that appears inside two of the symbols indicates that the component is gas filled. The position of the dot inside the circle is arbitrary. Even though all neon bulbs are gas filled, the dot is often omitted.

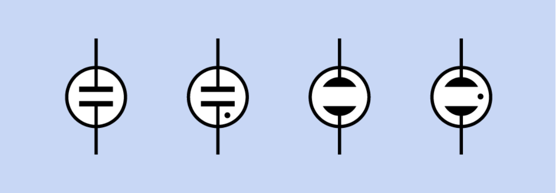

Figure 19-1. *Any of these symbols may represent a neon bulb or a neon lamp. The dot in two of the symbols indicates that the component is gas filled. All neon bulbs are gas filled, but the dot is often omitted.*

The photograph in Figure 19-2 shows a neon bulb with a series resistor preattached to one lead. Many bulbs are sold in this configuration, because a resistor must be used to limit current through the bulb. The bulb has no polarity and can be used on an AC or DC power supply. The same bulb is shown in its energized state in Figure 19-3.

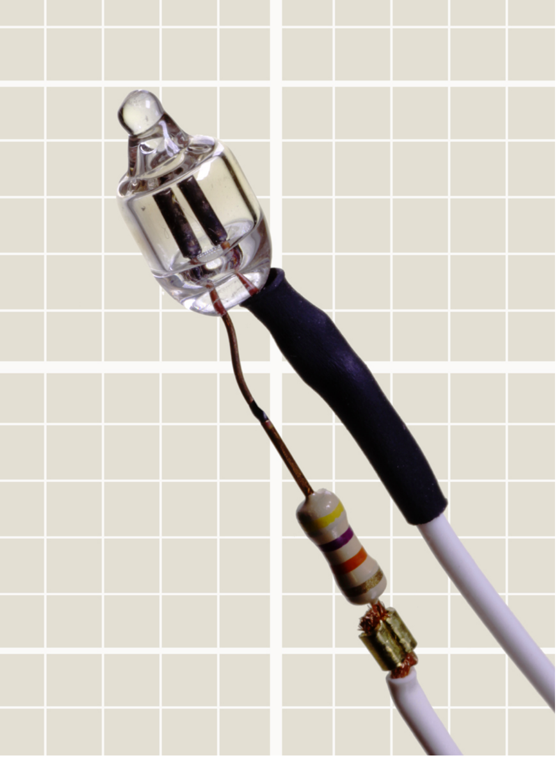

Figure 19-2. *A typical neon bulb with series resistor attached to one lead.*

How It Works

Construction

The parts of a neon bulb are illustrated in Figure 19-4. When the bulb is fabricated, it begins as a glass tube. The leads are made of *dumet*, consisting of a copper sheath around a nickel iron core. This has the same coefficient of expansion as glass, so that when the glass is heated and melted around the leads, it forms a seal that should be unaffected by subsequent temperature fluctuations. This area is known as the *pinch* in the tube.

Nickel electrodes are welded onto the leads before the leads are inserted into the tube. The electrodes have an emissive coating that reduces the minimum operating voltage. The glass tube is filled with a combination of neon and argon gases, or pure neon for higher light output (which reduces the life of the component). The top end of the glass tube is heated until it melts, and is pinched off. This creates a distinctive protrusion known as the *pip*.

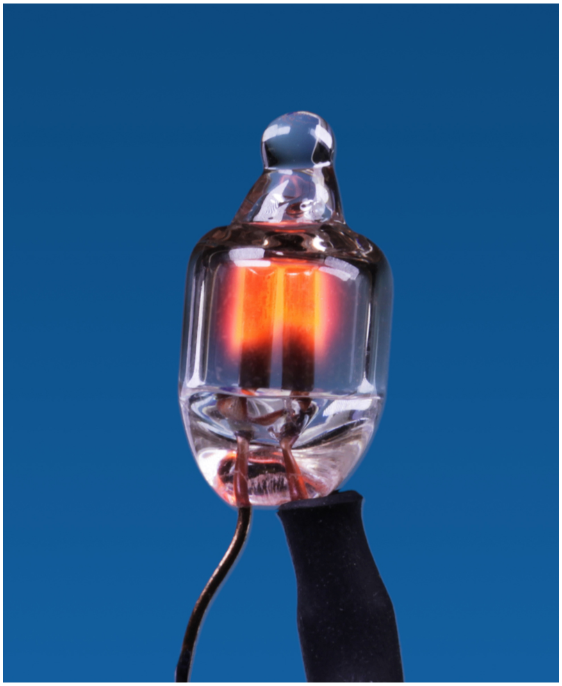

Figure 19-3. *The same bulb from the previous photograph, energized with 115VAC.*

Ionization

When a voltage is applied between the leads to the bulb, the gas becomes ionized, and electrons and ions are accelerated by the electric field. When they hit other atoms, these are ionized as well, maintaining the ionization level. Atoms are excited by collisions, moving their electrons to higher energy levels. When an electron returns from a higher level to a ground state, a photon is emitted.

This process begins at the *starting voltage* (also known as the *striking voltage*, the *ignition voltage*, or the *breakdown voltage*) usually between 45V and 65V for standard types of bulb, or between 70V and 95V for high-brightness types.

When the bulb is operating, it emits a soft radiance known as a *glow discharge* with a wavelength ranging from 600 to 700 nanometers.

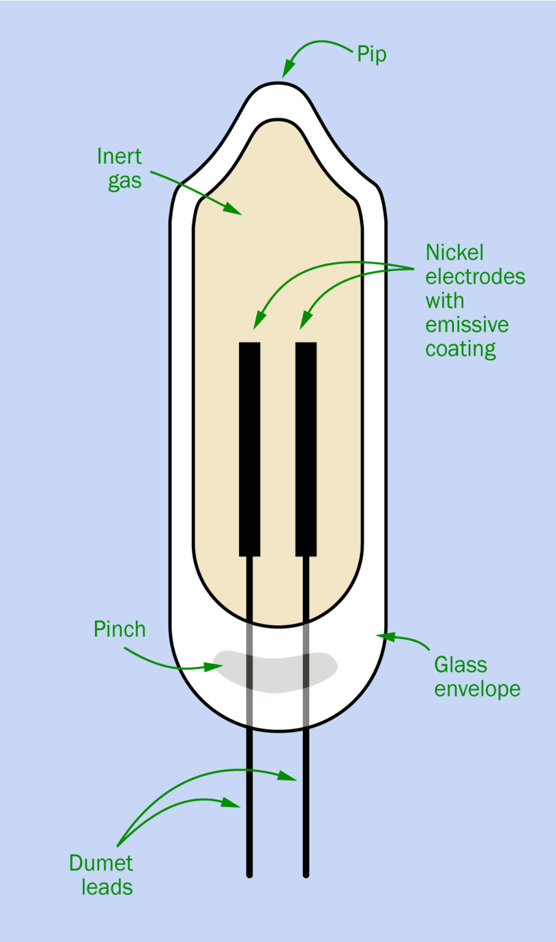

Figure 19-4. *The parts of a neon bulb. See text for details.*

The ionization of the gas allows current to flow through it. This will continue even if the power supply is reduced by 10 to 20 volts to a level known as the *maintaining voltage*.

Negative Resistance

When the glow discharge persists below the starting voltage, this is a form of *hysteresis*, meaning that the neon bulb tends to "stick" in its on state. It remains on while its power supply decreases to the maintaining voltage, but once it switches off, it will "stick" in its off state until the power supply increases again above the maintaining voltage to the starting voltage. The concept of hysteresis is discussed in the entry on **comparators**. See Figure 6-2.

A neon bulb is said to have *negative resistance*. If the current is allowed to increase without restraint, the resistance eventually decreases while the current increases further. If this runaway behavior is not controlled, the bulb will destroy itself.

This behavior is characteristic of gas-discharge tubes generally. A graph showing this appears in Figure 19-5. Note that both scales are logarithmic. Also note that the curve shows how current will be measured in response to voltage. If the voltage is reduced after it has increased, the transitional events shown by the graph will not recur in reverse order. This is especially true if arcing is allowed to begin, as it will almost certainly destroy the component.

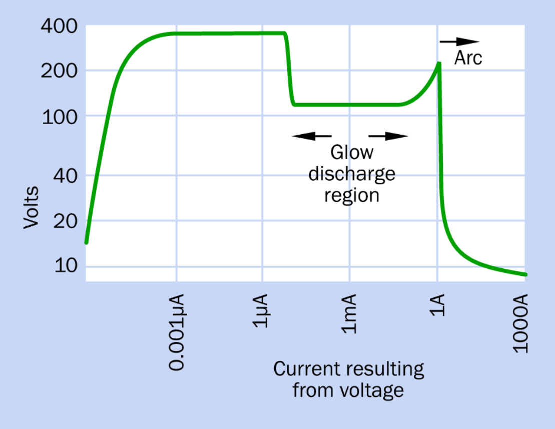

Figure 19-5. *A gas discharge tube, such as a neon bulb, is said to have a negative resistance, as current passing through it tends to increase uncontrollably after the gas is ionized and becomes conductive. (Derived from measurements made by David Knight, on a web page named after his radio ham call sign, G3YNH.)*

A neon bulb can be controlled very simply with a series resistor that maintains it in gas-discharge mode. To understand the operation of the resistor, consider the combination of the lamp and the resistor as a voltage divider, as shown in Figure 19-6. Before the lamp begins to pass current, it has an almost infinite resistance. Therefore, the voltage on both sides of the resistor will be approximately equal, the bulb passes almost no current, and it remains dark.

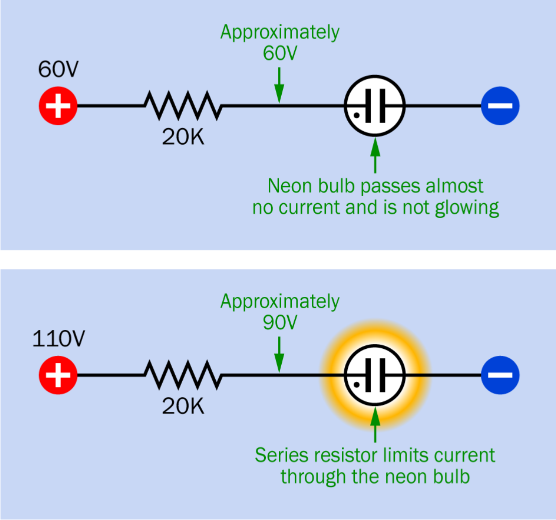

Figure 19-6. *A series resistor is essential to limit the current through a neon bulb.*

After the lamp begins to pass current, the requirement now is for the series resistor to reduce the voltage from the supply level (probably around 110V) to the maintaining level (probably around 90V). This means that the desired voltage drop is 20V, and if the manufacturer's specification tells us that the lamp should pass 1mA (i.e., 0.001 amps), R, the value of the series resistor, is given by Ohm's Law:

```
R = 20 / 0.001
```

Thus, the value for R is 20K. In fact, the value of a resistor supplied with a neon bulb may range from 10K to 220K, depending on the characteristics of the bulb and the supply voltage that will be used.

Now if the bulb's effective internal resistance falls radically, the resistor still limits the currrent. In a hypothetical worst-case scenario, if the bulb's resistance drops all the way to zero, the resistor must now impose the full voltage drop of 110V, and the current, I, will be found by Ohm's law:

```
I = 110 / 20,000
```

That is, about 5mA, or 0.005A.

Neon tubes used in signage require a more sophisticated voltage control circuit which is not included in this encyclopedia.

How to Use It

The use of a neon bulb for an indicator lamp is primarily limited to situations where domestic supply voltage (115VAC or 220VAC) is readily available. "Power on" lights are the obvious application, especially as neon indicators can accept AC. The switch shown in Figure 19-7 is illuminated by an internal neon bulb. The rectangular indicator in Figure 19-8 is designed to run on domestic supply voltage, and its internal bulb and resistor can be clearly seen through the green plastic. The assembly in Figure 19-9 is about 0.5" in diameter, which is the lower limit for neon indicators.

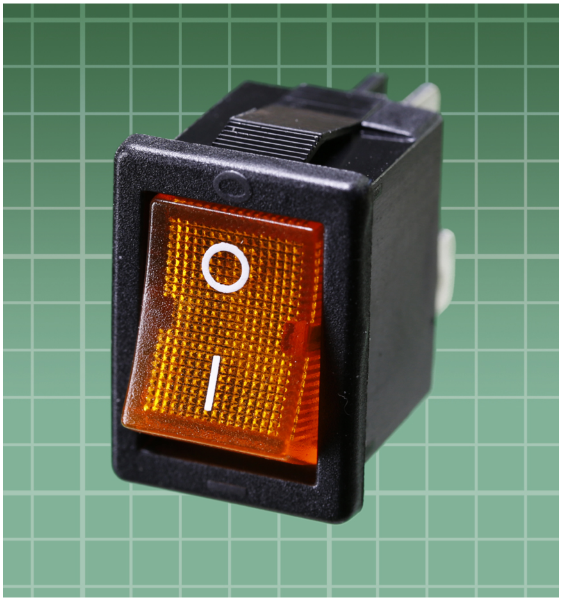

Figure 19-7. *This power switch is illuminated by an internal neon bulb.*

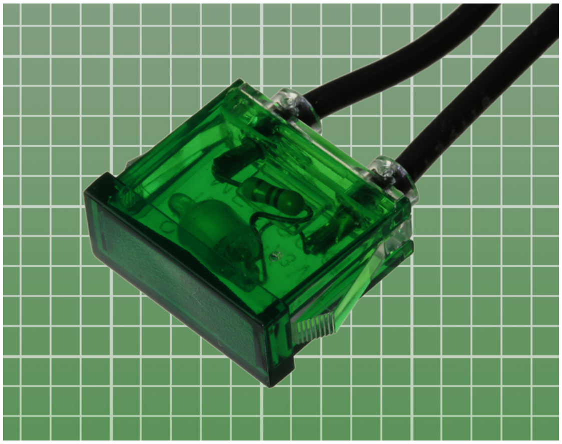

Figure 19-8. *The neon bulb and its series resistor are visible inside this indicator.*

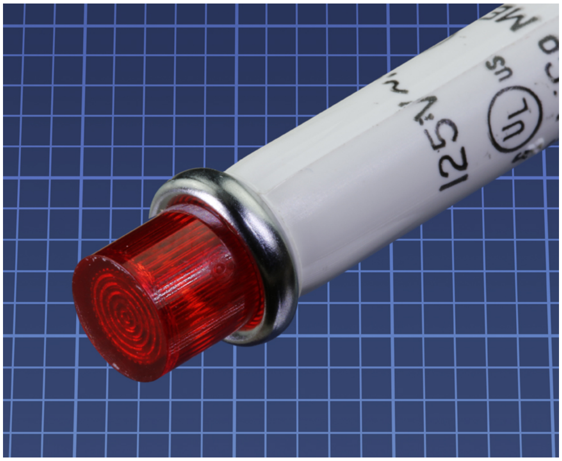

Figure 19-9. *A relatively small neon indicator lamp, designed for insertion in a hole 0.5" diameter.*

Limited Light Output

Neon bulbs have a light output of around 0.06 lumens per milliamp of consumed power (standard brightness type) or 0.15 lumens per milliamp of consumed power (high brightness type).

Comparing this value with the intensity of **LED** indicators is difficult. Their light output is customarily measured in *millicandelas* (mcd), because LED indicators almost always include a lens that focuses the light, and the candela is a measurement of luminous flux within an angle of dispersion. Moreover, because the intensity of neon indicators is not of great interest in most applications, datasheets usually do not supply an intensity value.

One way around the problem of comparisons is to use the standard of *radiant luminous efficacy* (LER), which is defined in the entry on incandescent lamps (see "Efficacy" on page 179). A standard-brightness neon bulb has an LER of about 50 lumens per emitted watt of luminous flux. A light-emitting diode may reach an LER of 100 lm/W. However, a neon bulb operates typically around 1mA while an LED indicator may use 20mA. Therefore, a typical LED indicator may appear to be 30 to 50 times brighter than a typical neon bulb.

Consequently, neon may be an inferior choice in a location where there is a high level of ambient light. Direct sunlight may render the glow of a neon indicator completely invisible.

Efficiency

Because a neon bulb does not use a lot of power and generates negligible heat, it is a good choice where current consumption is a consideration (for example, if an indicator is likely to be on for long periods). The durability and low wattage of neon bulbs, and their convenient compatibility with domestic power-supply voltage, made them a favorite for night-lights and novelty lamps in the past. Figure 19-10 shows an antique bulb containing an ornamental electrode, while Figure 19-11 is a piece of folk art, approximately 1" in diameter, mounted on a plug-in plastic capsule containing a neon bulb.

Ruggedness

Neon bulbs are a good choice in difficult environments, as they are not affected by vibration, sudden mechanical shock, voltage transients, or frequent power cycling. Their operating temperature range is typically from -40 to +150 degrees Celsius, although temperatures above 100 degrees will reduce the life of the lamp.

Figure 19-10. *In bygone decades, ornamental neon bulbs with specially shaped electrodes were popular.*

Figure 19-11. *Neon folk art survives in this hand-painted night-light sold in a Florida tourist shop.*

Power-Supply Testing

When driven by DC current, only the negative electrode (the *cathode*) of a neon bulb will glow. When AC current passes through the bulb, both terminals will glow.

If a bulb (with series resistor) is placed between the "hot" side of a domestic AC power supply and ground, the bulb will glow. If it is placed between the neutral side of the supply and ground, it will not glow.

These features enable a neon bulb to be used for simple power-supply testing.

Life Expectancy

The metal of the electrodes gradually vaporizes during everyday use of a neon bulb. This is known as *sputtering* and can be observed as the glass capsule becomes darkened by deposition of vaporized metal. The electrodes will have a more limited life in a lamp used with DC voltage, where sputtering affects only the cathode. Using AC, the electrodes take turns functioning as the cathode, and vaporization is distributed between both of them.

Failure of a neon lamp can occur as sputtering erodes the electrodes to the point where the maintaining voltage will increase until it almost reaches the level of the power supply. At this point, the bulb will flicker erratically.

Failure can also be defined as a gradual reduction in brightness to 50% of rated light output, caused by accumulated deposition in the glass capsule. Because deposition occurs more heavily on the sides of the bulb, a longer apparent life is possible if the bulb is mounted so that it is viewed from the end.

Typically, neon bulbs are rated for 15,000 to 25,000 hours (two to three years of constant operation). However, the life can be greatly increased by a slight reduction in voltage, which may be achieved by substituting a series resistor with a slightly higher value.

The relationship between operating life and resistor value is shown below. If LA is the normal operating life, LB is the extended operating life, RA is the normal resistor value, and RB is a higher resistor value:

```
LB = LA * ( RB / RA ) ^3.3
```

For example, if a normal resistor value is 20K, and it is increased to 22K, the life of the lamp should increase by a factor of slightly more than 1.4.

Variants

A typical neon bulb terminates in leads, and a lamp assembly often has solder tabs, although it may have a base with a screw thread, flange, or bayonet pins for insertion into a compatible socket. A lamp assembly that does not use a base will either snap-fit into a hole of appropriate size and shape, or may be retained with a nut that engages with a plastic thread on the cylinder of the lamp.

Some neon bulbs or lamp assemblies terminate in pins for direct insertion into a printed circuit board.

Almost all neon bulbs operate either in the 100V to 120V range or in the 220V to 240V range.

Light intensity is expressed either as "standard" or "extra-bright," although datasheets usually do not define those terms.

Nixie Tubes

Nixie tubes, first marketed in 1955, were used to display numerals from 0 through 9 in the days before LEDs took over this capability. They are no longer being manufactured.

Each numeral was physically formed from metal and functioned as an electrode inside a tube filled with a neon-based gas mixture. The typographical elegance of the digits and their aesthetically pleasing glow made Nixies enduringly popular. With a long lifespan, vintage tubes are still usable and can be purchased cheaply from sources such as eBay. Many originate in Russia, where Nixie-type displays were manufactured into the 1980s. The Russian tubes can be identified by their use of a numeral 5 that is a numeral 2 turned upside-down.

Nixie tubes typically require 170VDC. This creates a challenge for a power supply and switching, and can be a safety hazard.

Figure 19-12 shows six Nixie-type tubes repurposed for use as a 24-hour digital clock.

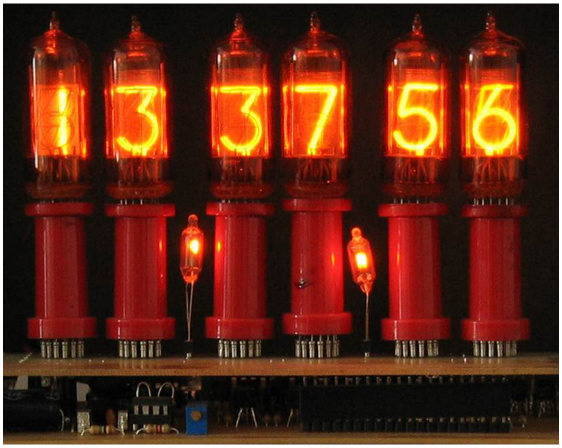

Figure 19-12. *A 24-hour clock using Nixie-type tubes. Source: Wikipedia, public domain.*

What Can Go Wrong

False Indication

Because a neon bulb requires so little power, it may be energized by induced voltages from elsewhere in a circuit, especially if inductive components such as transformers are used. To prevent this, a high-value resistor can be placed in parallel with the bulb, in addition to the series resistor that must always be used.

Failure in a Dark Environment

Because a neon bulb requires a minimal amount of light to initiate its own photon emissions, it may take time to start glowing in a very dim environment, and may not light at all in total darkness. A few bulbs include a small amount of radioactive material that enables them to self-start in complete absence of ambient light.

Premature Failure with DC

The life expectancy quoted in datasheets for neon bulbs usually assumes that they are powered by AC. Because DC results in faster vaporization of the electrodes, the expected lifetime should be reduced by 50% if DC power will be used.

Premature Failure through Voltage Fluctuations

Because the deterioration of a neon bulb accelerates rapidly with current, a sustained voltage that passes slightly more current can radically reduce the expected lifespan.

Replacement

Replacement can be an issue with panel indicators, where disassembly of a device may be necessary to reach the bulb. Bear in mind, however, that an easily removable bulb becomes vulnerable to tampering.

fluorescent light

20

This entry deals primarily with *fluorescent tubes* (infrequently but sometimes described as *fluorescent lamps*), and *compact fluorescent lamps* (CFLs) that are marketed as a substitute for **incandescent lamps**. *Cold-cathode fluorescent lamps* (CCFLs) are also mentioned.

Vacuum fluorescent devices have a separate entry in this encyclopedia. A fluorescent tube or CFL does not contain a vacuum.

Although the diode(s) in a white **LED area lighting** unit are coated with a layer of fluorescent phosphors, they are not categorized here as fluorescent lights, and have their own entry.

A **neon bulb** resembles a fluorescent light in that it is a *gas-discharge* device, but the interior of its glass envelope is usually not coated with fluorescent phosphors, and therefore it has its own entry.

OTHER RELATED COMPONENTS

- **incandescent lamp** (see Chapter 18)
- **LED area lighting** (see Chapter 23)
- **vacuum fluorescent** devices (see Chapter 25)
- **neon bulb** (see Chapter 19)

What It Does

Fluorescent tubes or *compact fluorescent lamps* (CFLs) are primarily used for area lighting. A partially disassembled CFL appears in Figure 20-1, showing the control electronics that are normally hidden inside the base.

There is no standardized schematic symbol to represent a fluorescent light. Figure 20-2 shows three commonly used symbols for a fluorescent tube on the left, and three symbols for a CFL on the right. Note that two of the symbols for a CFL are the same as those for an **incandescent lamp**, shown in Figure 18-1.

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see Chapter 18 for a description of **incandescent lamps**.)

Fluorescence is a form of luminescence. It occurs when electrons in a material are energized and then make a transition back to ground level, at which point they radiate their energy as visible light. The incoming energy can consist of other light at a higher frequency. Some creatures, including species of arachnids and fish, will fluoresce when they are lit with ultraviolet light.

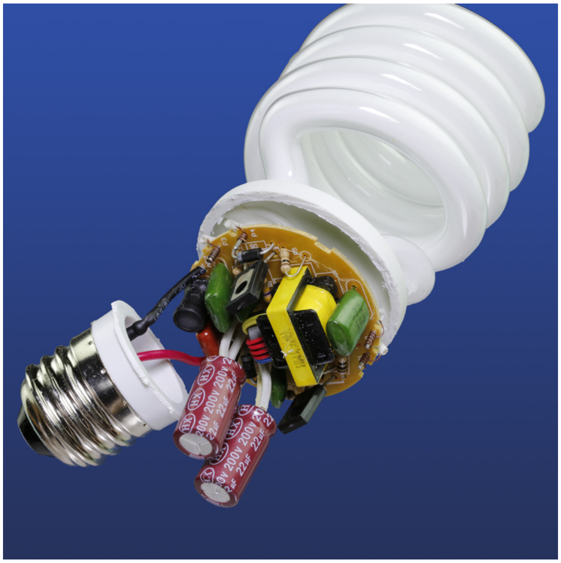

Figure 20-1. *A compact fluorescent lamp with its base cut away to reveal the control electronics.*

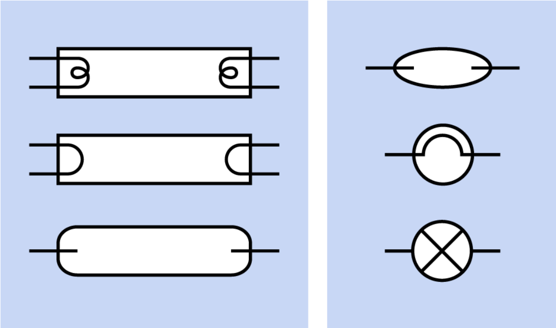

Figure 20-2. *Schematic symbols to represent fluorescent tubes and bulbs are not standardized. See text for details.*

A *fluorescent tube or lamp* contains a very small amount of mercury vapor that can be stimulated to emit ultraviolet light. This encounters a thin layer of *phosphors* coating the inner surface of the glass enclosure. The light causes the phosphors to *fluoresce*, emitting a diffuse radiance in the visible spectrum.

The tube or lamp also contains one or more inert gases such as argon, xenon, neon, or krypton at about 0.3% of normal atmospheric pressure. Two electrodes inside the glass enclosure are made primarily from tungsten, which can be preheated to initiate ionization of the gas. Confusingly, both electrodes are often referred to as *cathodes*.

The function of the gas is not to emit light, but to conduct electric current, so that free electrons may encounter mercury atoms, raising their electrons briefly to a higher energy level. When one of these electrons reverts from its unstable energized state to its previous energy level, it emits a photon at an ultraviolet wavelength.

Figure 20-3 provides a diagram showing the interior of a fluorescent tube.

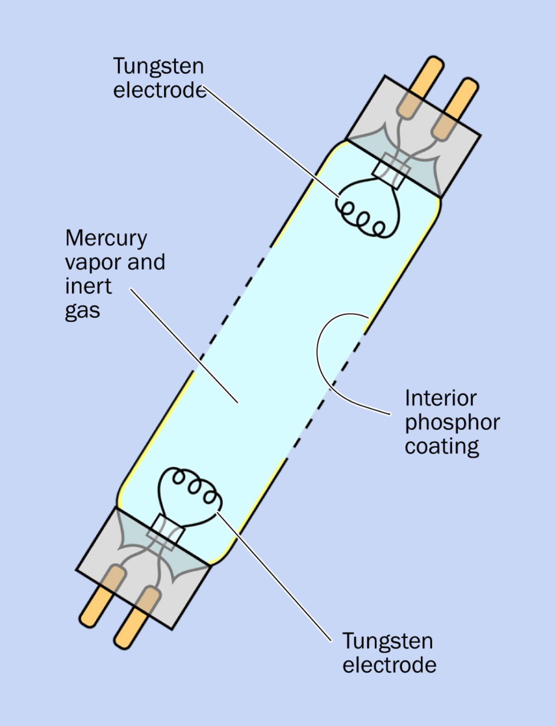

Figure 20-3. *The basic parts of a fluorescent tube.*

Ballast and Starter

Heating the tungsten electrodes is necessary but not sufficient to trigger ionization. A high-voltage pulse is also needed when the light is switched on. In a typical 48″ tube, the pulse may range from 200V to 300V.

After current flow has been established, the gas, which is now a plasma, enters a phase of negative resistance. Current passing through it will tend to increase even if the voltage decreases. This process must be controlled to prevent the formation of an *arc*, which will destroy the electrodes. (A similar process occurs in any gas discharge tube, such as a **neon bulb**, and is described in a graph in Figure 19-5.)

To heat the electrodes, ionize the gas, and then control the current, the fixture for a fluorescent tube contains components that are separate from the tube. In their simplest, traditional form, these components consist of a *starter* and a *ballast*. The starter is a neon bulb that contains a bimetallic strip serving as a normally closed switch. It allows current to flow through the electrodes in series, to heat them. The basic circuit is shown in Figure 20-4.

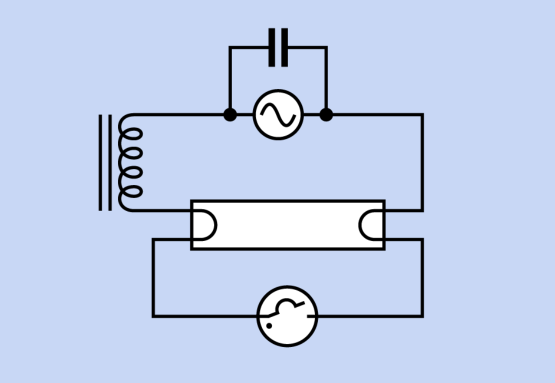

Figure 20-4. *The traditional circuit to trigger ionization of the gas in a fluorescent tube uses a starter (shown at the bottom as a neon tube containing a bimetallic strip, which serves as a switch) and ballast (an inductive load, shown at left).*

The starting process may not be immediately successful, in which case the starter may repeat several times in succession, causing the tube to flicker before its discharge becomes stable. In a cold environment, the tube will have more difficulty starting.

After the tube becomes conductive, current between the electrodes bypasses the starter. At this point, the *ballast* limits the current to prevent an arc from forming. The simplest form of ballast is a coil that functions as an **inductor**.

In a more modern system, an *electronic ballast* replaces the starter-ballast combination. It not only applies the initial surge of high current but also raises the 50Hz or 60Hz frequency of the power supply to 10KHz or more. This increases the efficiency of the tube and eliminates any visible flickering of the light.

All compact fluorescent bulbs (CFLs) contain electronic ballasts. The small components visible in Figure 20-1 are the ballast.

Flicker

When a fluorescent tube uses a conventional ballast and is illuminated with 50Hz or 60Hz AC, the glow discharge stops each time the current flow passes through the zero point in its cycle. In fact, the ionized gas in the tube cannot conduct until it is close to the maximum voltage, and stops conducting when the voltage rolls off. Consequently, the voltage across the tube fluctuates in an approximate square wave, and the light output begins and ends very abruptly. Although this occurs 100 times per second on a 50Hz supply and 120 times per second on a 60Hz supply, some people complain that the flicker is noticeable and can induce headaches.

The rapid on-off discharge is hazardous when it illuminates rotating parts in machinery, as a stroboscopic effect can make the parts seem to be stationary. To mitigate this effect, adjacent tubes in a fixture are powered by separate supplies that are out of phase. This is done either by using a three-phase power supply or by adding an LC circuit to the supply for one of the tubes.

Variants

The traditional type of ballast is also known as a *rapid-start ballast*. By preheating the electrodes, it reduces damage to them that otherwise tends to occur during the starting process. A tube designed for use with a rapid-start ballast has two contacts at each end, and is referred to as a *bi-pin* tube.

An electronic ballast is also known as an *instant-start ballast*. It does not preheat the electrodes, and a tube designed to work with it has only one pin at each end.

CCFLs

A *cold cathode fluorescent lamp* (CCFL) may resemble a miniature fluorescent tube, typically measuring 2mm to 5mm in diameter. The tube may be straight or bent into a variety of shapes. It works on the same principle as a full-size fluorescent tube, containing mercury vapor and one or more inert gases, with an interior coating of phosphors to enable fluorescence. CCFLs are available in many colors and many shades of white.

As its name implies, the electrodes in a CCFL are not heated to establish ionization. Instead, a very high voltage (1,000VAC or more) is applied, dropping to 500VAC to 600VAC after the flow of current has been established. Because CCFLs have been often used to backlight laptop computer screens, inverter circuits are commonly available that create a high-frequency output at a high voltage from an input that can range from 3VDC to 20VDC. The inverter also includes provision to dim the CCFL by using *pulse-width modulation*.

Some CCFLs are designed for illumination of small spaces—for example, the interior of a display case. A few CCFLs look exactly like CFLs and can be used in light fixtures. Some may be compatible with the type of dimmer designed for incandescent lamps.

A CCFL usually has a limited light output compared with that of a conventional fluorescent tube, but has the advantage of working better at low temperatures. Some are designed for signage and exterior lighting in cold-weather locations.

They have a relatively long lifetime of up to 60,000 hours. A hot-cathode fluorescent lamp may fail between 3,000 and 15,000 hours.

Any tube or bulb that uses unheated electrodes to ionize a gas is technically a cold-cathode device, but will not be identified as a CCFL unless it also has an inner layer of phosphors to achieve fluorescence.

It is important to match a tube with the type of ballast installed in a fixture. This is not an issue with CFLs, as they have the appropriate ballast built in.

Sizes

Straight bi-pin tubes are sold in the United States in the following standard sizes:

- T5: 5/8″ diameter. A more modern tube, but still with tungsten electrodes that serve to heat it.
- T8: 1″ diameter. Very often 24″ or 48″ in length, consuming 18W or 36W respectively.
- T12: 1-1/2″ diameter.
- T17: 2-1/8″ diameter.

CFLs are sold in a very wide variety of configurations.

Comparisons

Fluorescent lights have significant advantages and disadvantages. On the plus side:

- After the fixture containing the ballast has been paid for, a tube is relatively cheap. A CFL or an LED light does not have this advantage, as the electronics are built in and will be discarded when the light fails.
- Fluorescent lights have a longer life than incandescent bulbs.
- Fluorescent lights are available in a wide range of shades of white.
- Fluorescent tubes create a diffuse radiance that is ideal for general lighting using ceiling-mounted fixtures. They do not cast harsh shadows.

On the minus side:

- Fluorescents were traditionally more energy-efficient than any other light source, but **LED area lighting** is now more efficient in some designs. LEDs are expected to become more efficient in the future.
- A fluorescent tube with a traditional type of ballast may cause complaints of flickering. By comparison, an LED light uses DC, and an incandescent bulb retains sufficient heat between power cycles so that it does not appear to flicker.
- Fluorescent flicker creates problems when shooting video.
- The fluorescent emission spectrum has sharp peaks that give the lighting an unnatural look.
- In applications that require a defined beam of light, a fluorescent source cannot be used.
- Conventional ballasts can create radio interference, especially in the AM band.
- Because fluorescent tubes and bulbs contain mercury, they require proper disposal, which can incur fees.
- Even an instant-on fluorescent light tends to hesitate briefly when it is switched on.
- The lifespan of a fluorescent light is greatly reduced if it is cycled on and off frequently. An incandescent bulb is less severely affected by cycling, and an LED light is not affected at all.
- Fluorescent lights have difficulty starting at low temperatures.

Values

Brightness

The intensity of a fluorescent light is measured in *lumens per watt*. Because invisible wavelengths are of little interest when assessing brightness, *luminous flux* is used to describe apparent brightness in the visible spectrum. The unit for luminous flux is the *lumen*. Additional information about light measurement is included in the entry describing incandescent lamps (see “Power” on page 177).

Spectrum

The spectrum of photons emitted from mercury vapor in a fluorescent light has wavelengths that peak at 253.7 nanometers and 185 nanometers. (A *nanometer*, customarily abbreviated as nm, is one-billionth of a meter.) These wavelengths are invisible, being in the ultraviolet range, but when the light is transposed into the visible spectrum by the layer of phosphors, “spikes” in the range of wavelengths are still present. For a comparison of output curves for incandescent, fluorescent, and LED lights, see the graph in Figure 18-4.

Various formulations for the phosphors in a tube or CFL attempt to modify the character of the light to suit the human eye, but none of them looks as “natural” as the radiance from an incandescent bulb, probably because the characteristics of incandescent light are very similar to those of sunlight.

What Can Go Wrong

Unreliable Starting

At a low temperature, the mercury inside a fluorescent tube may be slow to vaporize. At very low temperatures, vaporization may not be possible at all. Until the mercury vaporizes, fluorescence will not occur.

Terminal Flicker

As a tube ages, it may start to conduct current only in one direction, causing it to flicker visibly. As it ages more, the gas discharge becomes even less reliable, and the flicker becomes erratic. Eventually, the gas discharge fails completely. In this state, a tube may show only a dim light at each end, in proximity to the tungsten electrodes.

Cannot Dim

Neither the older style of “conventional” ballast nor a modern electronic ballast will respond appropriately to a dimmer of the type designed for incandescent bulbs. This may be an important factor when an incandescent bulb is swapped out for a CFL.

Burned Out Electrodes

Like the tungsten filament in an **incandescent lamp**, the tungsten electrodes in a fluorescent tube suffer progressive erosion. This is evident when a black tungsten deposit forms on the inside of the tube at one or both ends.

Ultraviolet Hazard

Some critics of CFLs maintain that the complex shape of a coiled or zig-zag tube tends to permit small imperfections in the internal phosphor coating, potentially allowing ultraviolet light to escape. If this occurs, and if a CFL is used in a desk fixture in close proximity to the user, ultraviolet light could elevate the risk of skin cancer.

laser

21

The term *maser* was coined in the 1950s to describe a device that used stimulated emission to amplify microwaves. When a device using similar principles amplified visible light in 1960, it was termed an *optical maser*. However, that term is now obsolete, having been replaced with **laser**. This term is always printed in lowercase letters, even though it is an acronym for Light Amplification by Stimulated Emission of Radiation.

The invented verb *to lase* is derived from *laser* and is used to describe the process of generating laser light, with the past participle *lased* and present participle *lasing* sometimes being used.

Thousands of laser variants exist. Because of space limitations, this entry will concentrate primarily on *laser diodes*, which are the smallest, most common, and most affordable type.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)

What It Does

A **laser** generally emits a thin beam of intense light, often in the visible spectrum, and usually in such a narrow range of wavelengths, it can be considered *monochromatic*. The light is also *coherent*, as explained below.

Light output from a laser has three important attributes:

- Intensity. A high-powered laser can deliver energy to a very small, well-defined area, where it may be capable of burning, cutting, welding, or drilling. Large lasers may also be used as weapons, or for power transmission.
- Collimation. This term describes a beam of light that has parallel boundaries, and therefore does not disperse significantly when passing through a transparent medium such as air, glass, or a vacuum. A laser beam can have such excellent collimation, it can be used in precision measuring devices, and has been transmitted over very long distances, even from the Earth to the Moon, where astronauts placed reflectors during the Apollo missions.
- Controllability. Because the beam can be generated with eletrical power, its intensity can be modulated rapidly with relatively simple electronic circuits, enabling applications such as burning microscopic pits in the plastic of a CD-ROM or DVD.

Laser diodes are now more common than all other forms of lasers. They are found in pointers, printers, barcode readers, scanners, computer mice, fiber-optic communications, surveying tools, weapon sights, and directional lighting sources. They are also used as a light source to trigger more powerful lasers.

No generic symbol is used for a laser, but a laser diode is often represented with the same symbol that is used for a light-emitting diode. See Figure 22-2 in the entry for **LED indicators**.

How It Works

A laser is built around a *gain medium*, which is a material that can amplify light. The medium can be a solid, liquid, gas, or plasma, depending on the type of laser.

Initially, an input of energy provides stimulation for some atoms in the gain medium. This is known as *pumping* the laser. The energy input can come from a powerful external light source, or from an electric current.

Stimulation of an atom raises the quantum energy level of an electron associated with the atom. When the electron collapses back to its former energy state, it releases a photon. This is known as *spontaneous emission*.

If one of the photons encounters an atom that has just been excited by the external energy source, the atom may release two photons. This is known as *stimulated emission*. Beyond a threshold level, the number of released photons can increase at an exponential rate.

If two parallel reflectors are mounted at opposite ends of the gain medium, they form a *resonant cavity*. Light bounces to and fro between the reflectors, while pumping and stimulated emission amplifies the light during each pass. If one of the mirrors is partially transparent, some of the light will escape through it in the form of a laser beam. The partially transparent mirror is known as the *output coupler*.

Laser Diode

A laser diode contains an **LED**. (See "How It Works" on page 207 for a more detailed description of the function of an LED.) The p-n junction of the diode functions as the resonant cavity of the laser. Forward bias injects charges into the junction, causing spontaneous emission of photons. The photons, in turn, cause other electrons and electron-holes to combine, creating more photons in the process of stimulated emission. When this process crosses a threshold level, current passing through the diode causes it to *lase*.

The original patent for a laser diode was filed by Robert N. Hall of General Electric in 1962, and the diagram in Figure 21-1 is derived from the drawing in that patent, with color added for clarity.

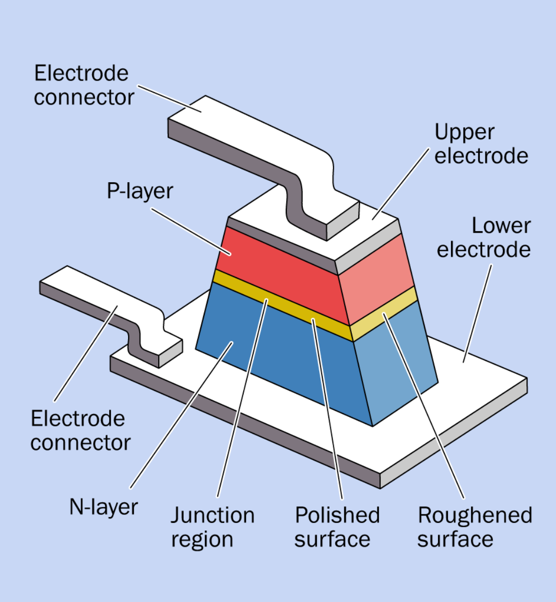

Figure 21-1. *The original design for a laser diode, from the patent filed in 1962.*

In the figure, the junction shown as a yellow layer forms the resonant cavity in which lasing occurs. It measures only 0.1 microns thick (the diagram is not drawn to scale). Its vertical front side is highly polished, and is parallel to the back side, which is also highly polished. Thus, photons reflect between these two vertical sides. The slanted face visible in the figure, and the other slanted face opposite it, are oriented and roughened to minimize internal reflection between them.

Figure 21-2 shows a simplified cross-section of the laser diode.

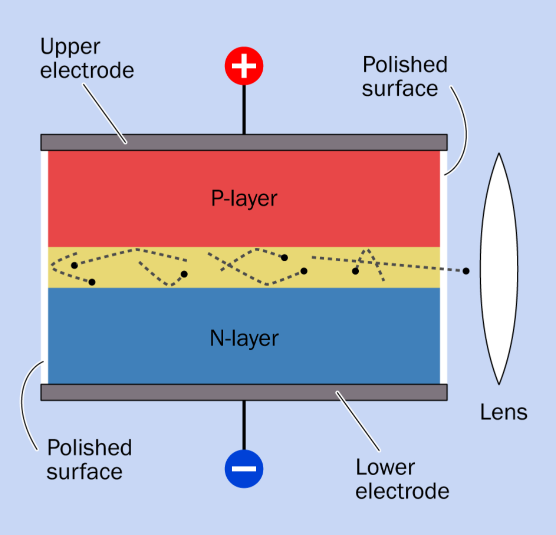

Figure 21-2. *Simplified cross-section of a laser diode.*

Figure 21-3 shows a cross-section of the diode installed in a component sold as a laser. It includes a photodiode to sense the intensity of light emerging through the polished rear end of the laser diode. External electronics are necessary to control the intensity of the laser, using feedback from the photodiode.

The component has three pins (shown pale yellow in the figure), one connecting to the photodiode, another connecting to the p-type layer of the laser diode, and the third being common to the n-type layer of the laser diode and the ground side of the photodiode.

A photograph of a laser diode is shown in Figure 21-4. Note the three pins, comparable to the pins shown in Figure 21-3, indicating that this component requires external control electronics.

In Figure 21-5, a laser is shown with a surface-mount chip adjacent to the solder pad connecting the blue wire. The presence of this chip, with only two wires, indicates that this component has its own control electronics and requires only a DC power supply.

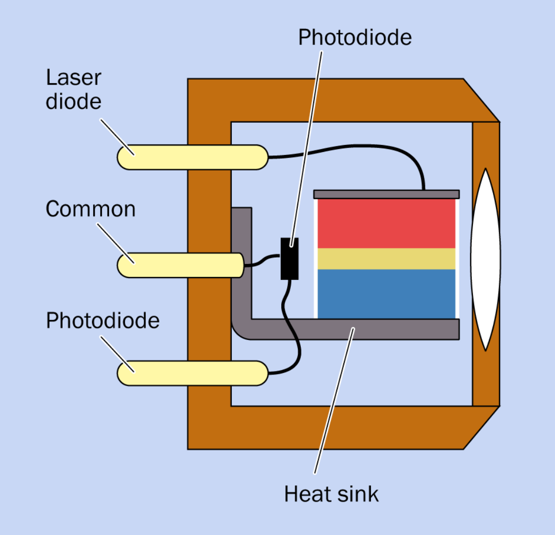

Figure 21-3. *A laser diode is typically mounted with a photodiode to provide feedback for a driver circuit, to control the current consumed by the laser.*

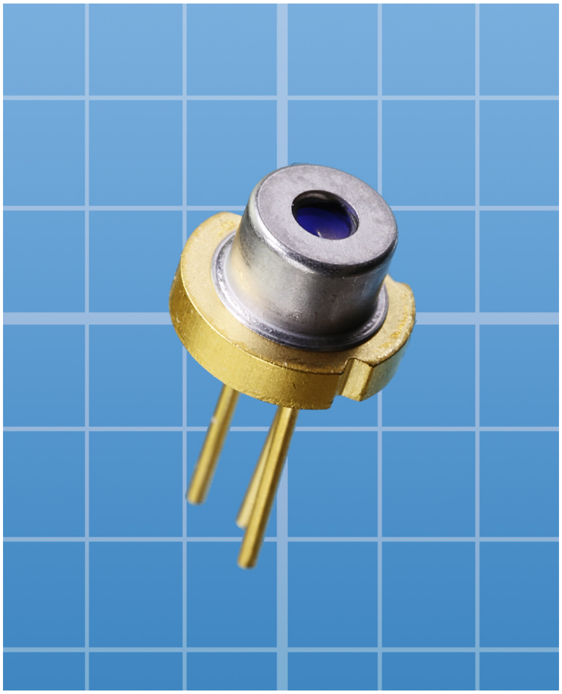

Figure 21-4. *Lite-On 505T laser diode that emits light at 650nm. Power consumption 5mW at 2.6VDC. As indicated by the graph squares, this component is only about 0.2" in diameter.*

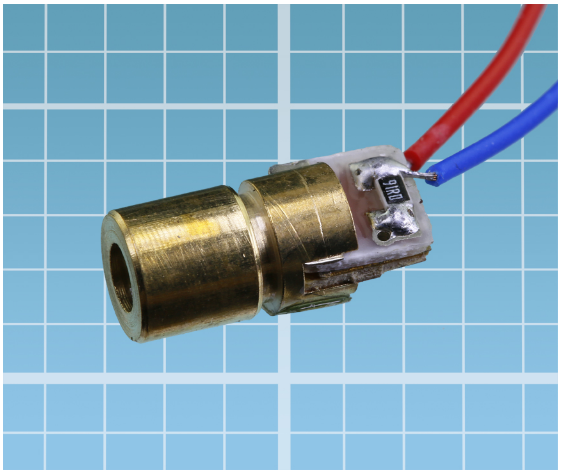

Figure 21-5. *This laser incorporates its own control electronics and requires only a 5VDC power supply. It draws 30mA and generates an output up to 5mW.*

Coherent Light

The emission of *coherent light* by a laser is often explained by suggesting that wavelengths are synchronized with each other. In fact, there are two forms of coherence that can be described approximately as *spatial coherence* and *wavelength coherence*.

If an observer looks up at a cloudy sky, the eye will perceive light radiating chaotically from many distances and directions. Thus, the light is not spatially coherent. The light also consists of many wavelengths, and thus it is not wavelength-coherent.

The filament of an incandescent lamp is a much smaller source of light, but still large enough to generate a profusion of light emissions that are spatially incoherent. The light also includes many different wavelengths.

Suppose a barrier containing a very tiny hole is placed in front of the **incandescent lamp**. If the aperture is very small, an observer on the far side will see the light as a point source. Consequently, the light that emerges from it is now spatially coherent, and will not have chaotically overlapping waves. If the light then passes through a filter, its wavelengths also will become coherent. This is suggested in Figure 21-6 where the light source is an incandescent lamp emitting a wide range of wavelengths.

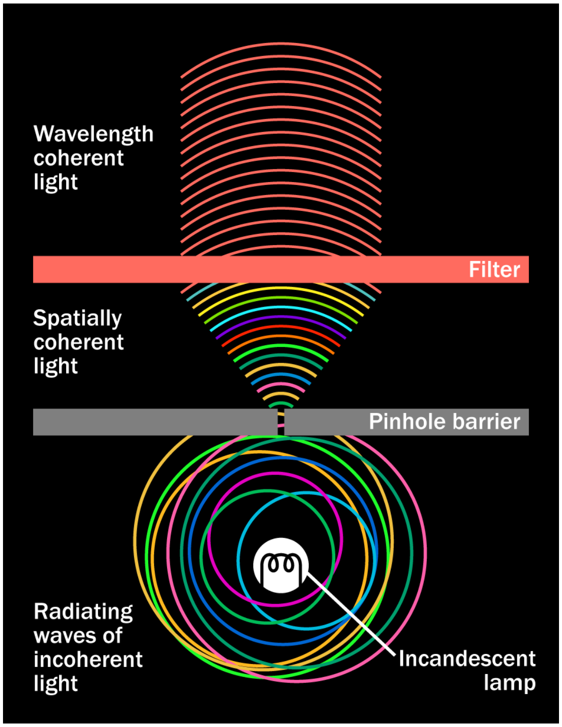

Figure 21-6. *An incandescent lamp, at the bottom of the figure, emits incoherent light at many wavelengths (exaggerated here for clarity). When it passes through a pinhole, it becomes spatially coherent. When it then passes through a colored filter, it becomes wavelength coherent.*

The small amount of light emerging through a pinhole is inevitably much dimmer than light from the original source. A laser, however, amplifies its light output, as well as tending to behave like a point source. The “hall of mirrors” effect of the parallel reflective surfaces in the resonant cavity causes much of the light to shuttle to and fro over a long distance before it emerges through the output coupler. Any light that deviates significantly from the axis of the laser will not escape at all, because the deviations will be cumulative with each reflection. Thus, the light from a laser appears to come from a point source at an almost infinite distance.

Because of the particular geometry of a light-emitting diode, the output from a laser diode is not naturally collimated, and tends to spread by an angle of around 20 degrees. A lens must be used to focus the beam.

Variants

Lasers are sold generally as fully assembled tools for a specific purpose. A very brief summary of CO2, fiber, and crystal lasers is included here.

CO2 Lasers

The gain medium is primarily carbon dioxide but also contains helium and nitrogen, with sometimes hydrogen, water vapor, and/or xenon. The laser is electrically pumped, causing a gas discharge. Nitrogen molecules are excited by the discharge and transfer their energy to the CO2 molecules when colliding with them. Helium helps to return the nitrogen to base energy state and transfer heat from the gas mixture.

CO2 lasers are infrared, and are commonly used in surgical procedures, including ophthalmology. Higher powered versions have industrial applications in cutting a very wide range of materials.

Fiber Lasers

Light is pumped via diodes and amplified in purpose-built glass fibers. The resulting beam has a very small diameter, providing a greater intensity than CO2 lasers. It can be used for metal engraving and annealing, and also for working with plastics.

Crystal Lasers

Like fiber lasers, they are pumped by diodes. These compact lasers are available in a very wide variety of wavelengths, covering the whole visible spectrum, infrared, and ultraviolet. They find applications in holography, biomedicine, interferometry, semiconductor inspection, and material processing.

Values

The output power of a laser is measured in watts (or milliwatts). This should not be confused with the power consumed by the device.

In the United States, any device sold as a laser pointer is limited to a power output of 5mW. However, laser diodes packaged similarly to laser pointers can be mail-ordered with an output of 200mW or more. The legal status of these lasers may be affected by regulations that vary state by state.

In a CD-RW drive that is capable of burning a disc, the diode may have a power of around 30mW. A laser mounted in a CD-ROM assembly is shown in Figure 21-7.

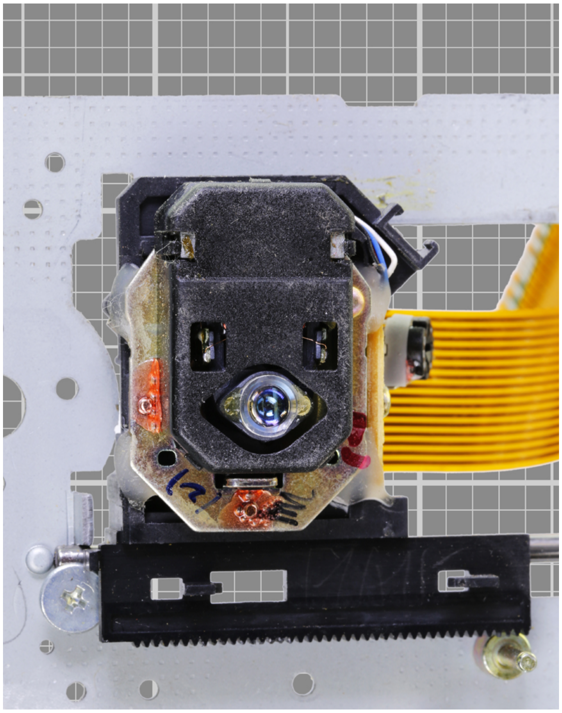

Figure 21-7. *An assembly incorporating a laser for reading a CD-ROM.*

Lasers have such a narrow range of wavelengths, they are given specific output values in nanometers. A laser in an optical mouse may have a wavelength of 848nm; in a CD drive, 785nm; in a bar-code reader, 670nm; in a modern laser pointer, 640nm; in a Blu-ray disc player, 405nm.

How to Use It

While powerful lasers in a laboratory setting have exotic applications, a typical low-power laser diode has become so affordable (costing less than $5 in some instances, at the time of writing) it can be considered merely as a useful source of a clearly defined light beam, ideal for detecting the position of a movable mechanical component or the presence of an intruder.

Generic light-emitting diodes are made with a view angle (i.e., a dispersion angle) as low as 3 degrees, but the beam is soft-edged compared with the precise boundary of a laser beam, and cannot be used reliably in conjunction with sensors at a distance of more than a few inches.

Laser diodes that are sold as components may or may not have current-limiting control electronics built in. Applying power to the laser diode directly will result in thermal runaway and rapid destruction of the component. Drivers for laser diodes are available separately as small, preassembled circuits on breakout boards.

For many applications, it may be simpler and cheaper to buy a laser diode as an off-the-shelf product. A laser pointer provides an easy way to get a source of laser light, and if it would normally be driven by two 1.5V batteries, it can be adapted to run off a 5V supply by using a 3.3V voltage regulator.

Common Applications

In addition to being used with PowerPoint presentations and in conjunction with position sensors, laser pointers have other applications:

- Astronomy. A high-powered laser beam is visible even in clear air as a result of interaction with air molecules. This is known as *Rayleigh scattering*. The phenomenon allows one person to point out a star (or planet) for another person. Because celestial objects are so far away, parallax error is not detectable by two people viewing the beam while standing next to each other. A laser pointer may also be mounted on a telescope to assist in aiming the telescope at an object of interest. This is easier than searching for an object through an eyepiece.
- Target acquisition. Lasers are commonly used on firearms to assist in targeting, especially in low-light conditions. Infrared lasers can be used in conjunction with infrared viewing goggles.
- Survival. A small laser can be included in emergency supplies to signal search teams. A laser can also be used to repel predatory animals.

What Can Go Wrong

Risk of Injury

Lasers are potentially dangerous. Those that have an infrared or ultraviolet output are more dangerous than those with a visible beam, as there is no visual warning that the laser is active. A laser is capable of scarring the retina, although controversy exists regarding the power output that should be considered a high risk.

If a project incorporates a laser, it should be switched off while building or testing the device. It may be advisable to wear protective glasses that block laser light even when an experimenter feels confident that a laser is switched off.

Active lasers should never be pointed at people, vehicles, animals (other than dangerous animals), or oneself.

Inadequate Heat Sink

Lasers may be designed and rated for intermittent use. The burner assembly for a CD-ROM drive, for instance, will be rated for pulsed power, not continuous power. Read datasheets carefully, and provide an adequate heat sink.

Uncontrolled Power Supply

A diode laser that does not have a feedback system in place to control the flow of current can self-destruct.

Polarity

Both the light-emitting diode and the photodiode in a three-pin laser package can be damaged by incorrect polarity of applied power. Pin functions should be checked carefully against datasheets.

LED indicator

22

In this encyclopedia, an **LED indicator** is defined as a component usually 10mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED.*

LED indicators that emit infrared and ultraviolet light are included in this entry. LEDs that are designed to illuminate large interior or exterior areas are discussed in a separate entry as **LED area lighting**. They are sometimes described as *high-brightness LEDs* and almost always emit white light.

The term **light-emitting diode** is becoming less common, as the acronym **LED** has become ubiquitous. The acronym does not usually include periods between the letters.

The words "light emitting" are hyphenated here, as they form an adjectival phrase, but in everyday usage the hyphen is often omitted, and no definitive rule seems to exist.

Originally, a standard LED contained only one diode, but may now include multiple diodes, either to emit additional light or to provide a range of colors. In this encyclopedia, a single epoxy or silicone capsule is still considered to be an **LED indicator** regardless of how many diodes it contains. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- or 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

OTHER RELATED COMPONENTS

- **LED area lighting** (see Chapter 23)
- **LED display** (see Chapter 24)
- **incandescent lamp** (see Chapter 18)
- **neon bulb** (see Chapter 19)
- **laser** (see Chapter 21)

What It Does

An **LED indicator** emits light in response to a small current, typically around 20mA (but sometimes much less), at a voltage lower than 5VDC. It is usually molded from epoxy or silicone that may be colorless and transparent (often referred to as *water clear*), or colorless but translucent, or tinted and transparent, or tinted and translucent.

The color of the light is initially determined by the chemical compounds used internally, and by their dopants; therefore, a water-clear LED may emit colored light.

Ultraviolet LEDs are usually water-clear. Infrared LEDs often appear to be black, because they are opaque to the visible spectrum while being transparent to infrared.

When an LED indicator is described as being *through hole,* it has leads for insertion into holes in a circuit board. The term does not mean that the indicator itself is meant to be pushed through a hole in a panel, although this may also be done. The LED is cylindrical with a hemispherical top that acts as a lens. The leads are relatively thick, to conduct heat away from the component. A dimensioned diagram of a typical LED measuring 5mm in diameter is shown in Figure 22-1.

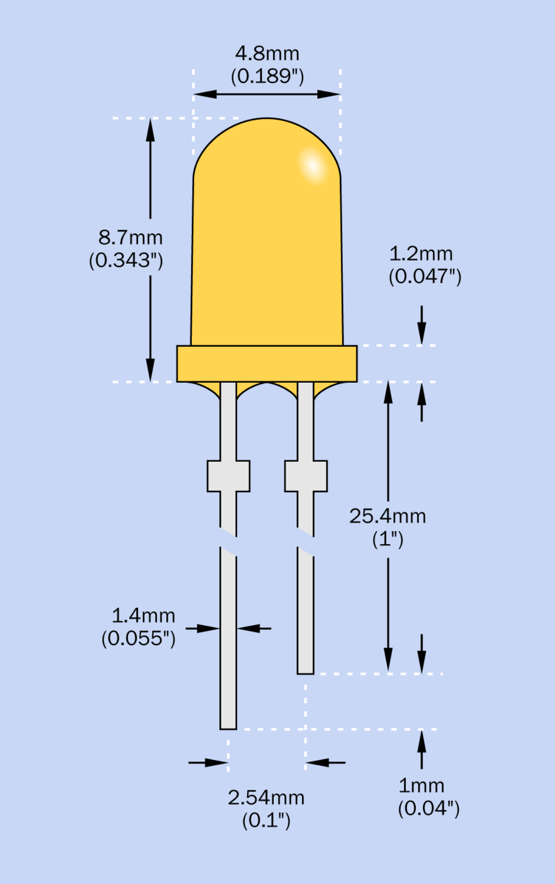

Figure 22-1. *Dimensions of a typical 5mm LED. The longer wire connects with the anode, while the shorter wire connects with the cathode. Adapted from a datasheet published by Lite-On Technology Corporation.*

An LED indicator that is not the through-hole type is usually a surface-mount component. LEDs for surface mounting are mostly rectangular and can be as small as 1mm x 0.5mm. They may require a *heat sink*.

Schematic Symbols

Figure 22-2 shows a variety of symbols that are commonly used to represent an LED. The triangle at the center of each symbol points in the direction of conventional (positive-to-negative) current flow—from the anode to the cathode. Each pair of arrows radiating away from the diode indicates emitted light. Wavy arrows are sometimes used to represent infrared (thermal) radiation. Often, however, an infrared LED is represented in exactly the same style as an LED that emits visible light. With the exception of the wavy arrows, the various styles of schematic symbol are functionally identical and do not identify different attributes of the component such as size or color.

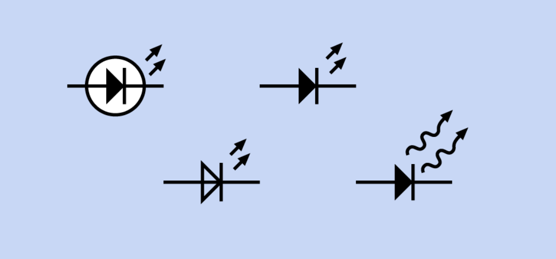

Figure 22-2. *Various symbols may be used to represent an LED. See text for details.*

Common Usage

LED indicators have mostly displaced **neon bulbs** and miniature **incandescent lamps** for the purpose of showing the status of a device. They are found in industrial control panels, home audio systems, battery chargers, washer/dryers, and many other consumer-electronics products. Higher output variants are used in flashlights, traffic signals, taillights on automobiles, and for illumination of subjects that are being photographed. LED indicators may be assembled in large numbers in attention-getting displays such as Christmas lights.

Red, orange, yellow, green, and blue are the basic standard colors. LEDs that appear to generate white light are common, but they do not emit an evenly weighted spectrum of wavelengths. See "Non-Incandescent Sources" on page 174 for a discussion of this topic.

How It Works

An LED, like any **diode**, contains a semiconductor *PN junction* that conducts current only in the forward direction (i.e., from the more-positive side of a power supply to the more-negative side). The diode becomes conductive above a *threshold voltage* sufficient to force electrons in the n-type region and holes in the p-type region to combine with each other. Each time this occurs, energy is released. The energy liberated by one electron-hole combination creates a *photon*, or one quantum of light.

The amount of energy released depends on the *band gap*, which is a property of the semiconductor material. The band gap is the smallest energy that can create an electron-hole pair. The energy determines the light's wavelength, and thus the color.

The band gap also determines the threshold voltage of the LED. For this reason, LEDs of different colors have widely different threshold voltages.

Because an LED will often be used in devices where the DC power supply exceeds the maximum forward voltage, a *series resistor* is customarily used as a simple way to restrict current through the diode.

The light emitted by a colored LED indicator tends to include only a narrow range of wavelengths. However, the addition of a phosphor coating to the diode can broaden the output. This technique is used to make the light from a blue LED appear white, as shown in Figure 22-3. Most white LEDs are actually blue LEDs with a colored phosphor layer added. See the section on **LED area lighting** in Chapter 23 for a more detailed discussion of this topic.

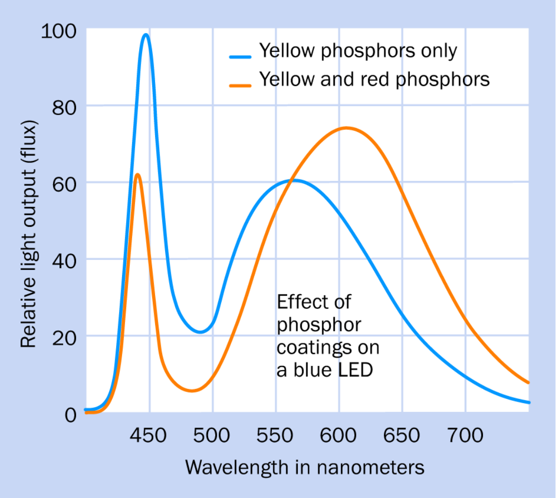

Figure 22-3. *Increasing the range of emitted wavelengths by adding phosphors to a blue LED. Source: Philips Gardco Lighting.*

Multicolor LEDs and Color Mixing

If red, green, and blue light sources are mounted extremely close together, the eye perceives them as a single source, of a color determined by their combined relative intensities. This system of additive color mixing is shown in Figure 17-14 in the entry dealing with LCDs. It is used in LED indicators that contain red, green, and blue light-emitting diodes in a single epoxy or silicone capsule.

While most video monitors use white LEDs or **fluorescent lights** to form a backlight for an **LCD** video screen, some high-end monitors use a matrix of very tiny red, green, and blue LEDs, because the combination of these separate colors generates a wider *gamut* of color wavelengths. The concept of gamut is discussed in "Color" in the **LCD** entry. The tiny LEDs in a backlight cannot be considered as indicators, but indicators are used for this purpose in billboard-sized video displays.

Variants

LED indicators vary widely in size, shape, intensity, view angle, diffusion of light, wavelength of light, minimum and maximum forward voltage, and minimum and maximum forward current.

Size and Shape

The original sizes for round LED indicators were 3mm, 5mm, or (more rarely) 10mm in diameter. Today, through-hole LEDs are sold in many intermediate sizes, although 3mm and 5mm are still most widely used.

The traditional round LED indicator is now augmented with square and rectangular shapes. In a parts catalog, a pair of dimensions such as 1mm × 5mm suggests that the LED is rectangular.

Intensity

The light intensity of an LED is usually expressed in *millicandelas*, abbreviated *mcd*. There are 1,000 mcd in a candela. For more information about units for measurement of light, see "Intensity" on page 178.

The candela measures the *luminous flux*, or visible radiant power, contained within a specified angle of dispersion, usually referred to as the *view angle*. This can be imagined as the rotated angle at the apex of a cone, where the cone defines the "spread" of the light, and the source is at the apex.

If a diode is emitting a fixed amount of luminous flux, the rating in mcd will increase with the inverse square of the view angle. This is because the light delivered to an area in front of the LED will become more intense as the angle becomes smaller. The use of mcd to rate the brightness of an LED can be misleading if it is not considered in comparison with the view angle.

For example, suppose an LED is rated at 1,000 mcd and has a view angle of 20 degrees. Now suppose the same diode is embedded in a different epoxy or silicone capsule with a lens that creates a view angle of only 10 degrees. The LED will now be rated at 4,000 mcd, even though its total power output is unchanged.

- To compare the brightness of two LED indicators meaningfully, they should share the same view angle.

Four through-hole LED indicators with a wide range of specifications are shown in Figure 22-4. From left to right: water-clear white generic, 10mm; Vishay TLCR5800 5mm (emitting red, even though the capsule is water-clear), rated for 35,000mcd with 4 degrees view angle; Everlight HLMPK150 5mm red diffused, rated for 2mcd with 60 degrees view angle; and Chicago 4302F5-5V 3mm green, rated for 8mcd at 60 degrees view angle, containing its own series resistor to allow direct connection with a 5VDC power supply.

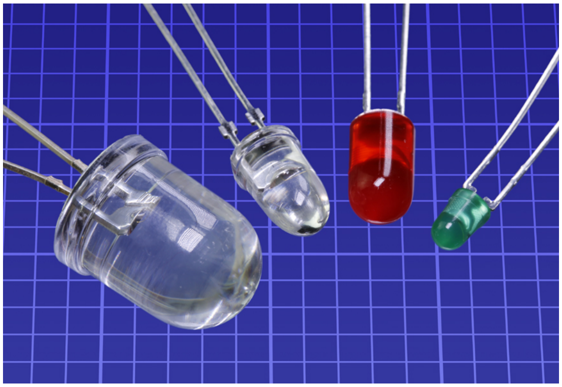

Figure 22-4. *Four assorted LED indicators with very different specifications. See text for details.*

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source compares how effective it is at channeling its output within the visible spectrum, instead of wasting it in other wavelengths, especially infrared. Note that the word *efficacy* has a different meaning from *efficiency*. The LER acronym may help to avoid confusion.

LER is expressed in lumens per watt, and in an incandescent bulb it is calculated by dividing the power emitted in the visible spectrum (the *lumi-*

nous flux) by the power emitted over all wavelengths. This is described in detail on "Efficacy" on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be within the visible spectrum, but some power is wasted by generating heat internally. The efficacy varies depending on the type of LED; thus a red-orange indicator can have an efficacy of 98% while a blue LED will be probably below 40%.

Diffusion

Some LED indicators use epoxy or silicone that is formulated to be translucent or "cloudy" instead of transparent. They diffuse the light so that it is not projected in a defined beam, has a softer look, and has an approximately equal intensity when viewed from a wider range of angles.

"Clear" and "diffused" are options that must be taken into account when choosing LEDs from an online catalog, unless the user is willing to turn a clear LED into a diffuse LED by applying some sandpaper.

Wavelength and Color Temperature

The wavelength of light is measured in *nanometers* (abbreviated *nm*), a nanometer being 1 billionth of a meter. The visible spectrum extends from approximately 380nm to 740nm. Longer wavelengths are at the red end of the spectrum, while shorter wavelengths are at the blue end.

A typical LED emits a very narrow range of wavelengths. For example, Figure 22-5 shows the emission from a standard red LED indicator manufactured by Lite-On. Graphs of this type are typically included in manufacturers' datasheets.

Because a red LED stimulates the cones in the eye that respond to red light, it "looks red" even though the color is not comparable with the natural red that is seen, for instance, in a sunset. That natural color actually contains an additional spread of wavelengths.

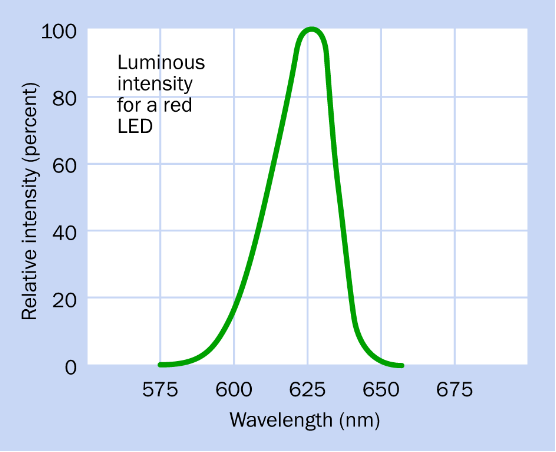

Figure 22-5. *The narrow range of wavelengths emitted by a typical red LED indicator.*

The following list shows the ranges of peak output values, in nanometers, for the most commonly available basic LED indicators (LEDs that emit other wavelengths are available, but they are less common):

- Infrared LED: 850 to 950
- Red LED: 621 to 700
- Orange LED: 605 to 620
- Amber LED: 590 to 591
- Yellow LED: 585 to 590
- Green LED: 527 to 570
- Blue LED: 470 to 475
- Ultraviolet LED: 385 to 405

Figure 22-6 shows this list graphically, omitting infrared and ultraviolet LEDs.

For almost 30 years, blue LEDs were a laboratory curiosity of little practical value, as efficiencies were stuck around 0.03%. An efficiency of more than 10% was finally achieved in 1995. Blue LEDs were marketed soon afterward.

However, when yellow phosphors are added to create the impression of white light by spreading the output over the whole visible spectrum, the wavelengths around 500nm are still not well represented, as suggested in Figure 22-3.

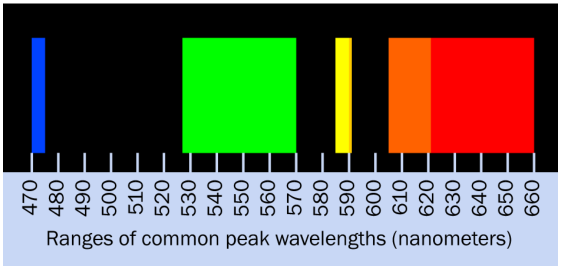

Figure 22-6. *Ranges for peak wavelengths of the most commonly used LEDs. (Source: Survey of approximately 6,000 through-hole LEDs stocked at www.mouser.com.)*

Fluorescent lights perform even more poorly than white LEDs, as can be seen in Figure 18-4 in the entry describing **incandescent lamps**.

Because white LEDs do not emit a single peak of wavelengths, their color is expressed in color temperature rather than nanometers. The concept of color temperature is explained in "Spectrum" on page 173. White LEDs are available rated from 2,800 to 9,000 degrees Kelvin, and are discussed in more detail in the **LED area lighting** entry in this encyclopedia.

Internal Resistor

To eliminate the chore of adding a series resistor to limit current through an LED, some indicators are sold with a series resistor built in. They may be rated for use with 5VDC or 12VDC, but are externally indistinguishable from each other. They are also externally indistinguishable from LEDs that do not contain series resistors. Figure 22-7 shows two 3mm LEDs, the one on the right containing its own series resistor, the one on the left being a generic LED without a series resistor.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an "out of range" error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

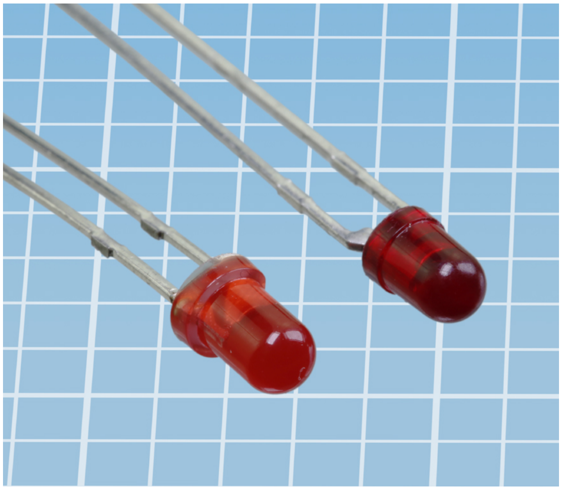

Figure 22-7. *An LED (left) that does not contain its own series resistor is usually indistinguishable from one that does (right).*

One way to determine whether an LED contains an internal series resistor would be to connect it with a variable power supply through a multimeter set to measure mA. Carefully increase the supply voltage from zero until the meter shows a current of 20mA. If the LED does not contain a series resistor, the supply voltage will be close to the recommended forward voltage for that type of LED (no lower than 1.6V for a red LED, and no higher than 3.6V for a white LED). If the LED does contain a series resistor, the supply voltage will be higher. This procedure is time consuming, but may be worthwhile to evaluate multiple LEDs that are known to be identical.

Multicolored

The leads for an LED indicator containing two or more diodes can be configured in several ways:

- Two leads, two colors. Two diodes are mounted internally in parallel, but with opposite polarity.
- Three leads, two colors. Two diodes share a common anode or common cathode.
- Four leads, three colors (RGB). Three diodes share a common anode or common cathode.

- Six leads, three colors. Three diodes, each with its own pair of leads, separate from the others.

Infrared

Most infrared emitters are LEDs that generate wavelengths longer than 800nm. They are found in handheld remotes to control consumer-electronics devices such as televisions and stereo systems, and are also used in some security systems, although *passive infrared motion detectors*, which assess infrared radiation from sources such as people or vehicles, are more commonly used for this purpose.

In conjunction with an infrared emitter, an infrared sensor is necessary, and must be sensitive to the same wavelength. To prevent false positives, the emitter modulates its output, typically with a carrier frequency between 10 and 100kHz. Remotes often use carrier frequencies of 30 to 56kHz. At the receiving end, the signal is processed with a band-pass filter matching the modulation frequency. Many different pulse-coding schemes are used, and no particular standard is dominant.

Ultraviolet

Because ultraviolet radiation can damage the eyes, LED indicators that emit ultraviolet light are potentially dangerous and should be used with caution. A yellow eyeshield can be worn to block the short wavelengths.

Ultraviolet light can be used to cure some adhesives and dental filling material. It can also kill bacteria, and can detect fluorescent print on bank notes, to check for counterfeiting. Ultraviolet flashlights are sold to detect some species of pests, such as scorpions, which fluoresce in response to ultraviolet light.

Values

The specification for an LED will include the wavelength of emitted light, luminous intensity, maximum forward voltage and current, maximum reverse voltage and current, and working values for voltage and current. All these values are important when choosing an indicator for a specific function.

White LEDs for room lighting or external use are calibrated differently. See the entry for **LED area lighting** in Chapter 23.

Forward Current

About half of all the thousands of available types of LED indicators are rated for a typical forward current of 20mA to 25mA. Absolute maximum ratings may be twice as high, but should not generally be applied.

The light intensity of a typical 5mm red LED indicator is plotted against its forward current in Figure 22-8. Note that current and light intensity have an approximately linear relationship up to the typical working current of 20mA. Even above this point, to the absolute maximum of 50mA, the light intensity rolls off only a very small amount.

Although an LED indicator can be dimmed by controlling the current passing through it, the current does not have a linear relationship with the applied voltage, and the indicator will stop functioning completely when the voltage drops below the threshold required by the diode. Consequently, LEDs are commonly dimmed by using *pulse-width modulation*.

Because of the nonlinear response of a diode, LEDs with or without internal resistors cannot be distinguished from each other reliably with a multimeter. If the meter is set to measure ohms, typically it will give an “out of range” error to all types of LED. If it is set to identify a diode, the reading will not tell you if the LED contains a resistor.

Low-Current LEDs

Indicators that require a very low forward current are convenient for direct connection to output pins of logic chips and other integrated circuits. Although a single output from an HC family chip is capable of supplying 20mA without damaging the chip, the current will pull down the output voltage, so that it cannot be used reliably as an input to another chip while also lighting the LED.

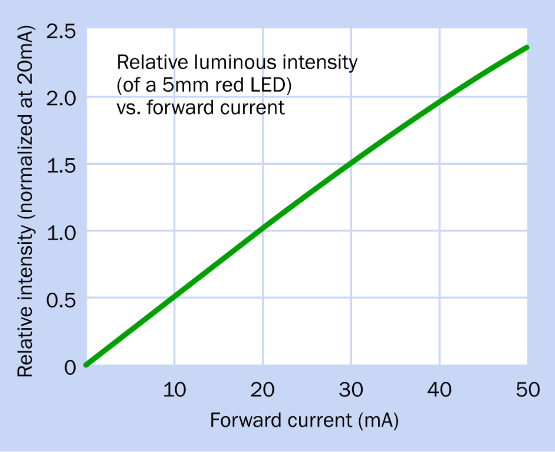

Figure 22-8. *The relationship between forward current and light intensity of a typical 5mm LED indicator is approximately linear up to the 20mA operating current, and almost linear up to the absolute maximum of 50mA.*

Various LED indicators drawing 2mA or 1mA are available, with intensities typically ranging between 1.5mcd and 2.5mcd. This low light output is still bright enough for viewing in a lab-bench environment. Low-current blue LEDs are not available. The only LEDs that draw as little as 1mA are red, as this is the most efficient type.

Using a higher value series resistor with a generic LED will of course reduce its current consumption, and some light will be visible so long as the forward voltage across the LED remains at its minimum level or above.

Forward Voltage

Red is the color that requires not only the least forward current, but the lowest forward voltage. In the range of 1.6VDC to 1.7VDC, all the LEDs are red. Typical forward voltages for various colors are shown here:

- Infrared LED: 1.6V to 2V
- Red LED: 1.6V to 2.1V
- Orange LED: 1.9V to 2.1V
- Amber LED: 2V to 2.1V
- Yellow LED: 2V to 2.4V
- Green LED: 2.4V to 3.4V
- Blue LED: 3.2V to 3.4V
- Ultraviolet LED: 3.3V to 3.7V
- White LED: 3.2V to 3.6V

Color Rendering Index

The *color rendering index* (CRI) evaluates the fidelity with which a light source is capable of displaying the full visible spectrum. It ranges from a perfect score of 100 down to 0 or even lower (sodium-vapor street lighting has a negative value). Computing the index requires standard reference color samples and has been criticized for generating scores that do not correlate well with subjective assessments.

Incandescent bulbs can have a CRI of 100, while an uncorrected white LED may score as low as 80.

Life Expectancy

Because the light output from an LED tends to decrease very gradually with time, the life expectancy is often defined as the number of hours required for the output to diminish to 70% of its output when new. Life expectancy is commonly stated on datasheets for high-brightness white LEDs, but is often omitted from datasheets for LED indicators.

Unlike incandescent lamps and fluorescent lights, LEDs do not have a shorter lifespan if they are frequently cycled on and off.

Light Output and Heat

The light intensity of an LED, measured in mcd, can vary from a few mcd to a maximum of 40,000mcd. Intensities above 30,000mcd generally are achieved by limiting the view angle to 15 degrees or less. Because the candela is weighted toward the central, green segment of the visible spectrum, green LEDs are likely to have a relatively high mcd rating. LEDs rated between 20,000mcd and 30,000mcd, with a view angle of 30 degrees, are almost all green.

Datasheets may often include a *derating curve* showing the lower limit that should be placed on forward current through an LED indicator when its temperature increases. In Figure 22-9, the LED should be operated only within the boundary established by the green line.

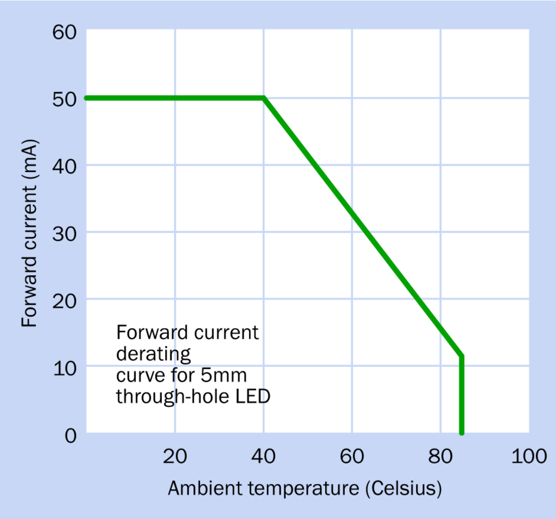

Figure 22-9. *Safe operation of an LED entails limiting the forward current if the temperature increases. The green line shows the boundary for operation of this particular component.*

View Angle

An LED formed from transparent epoxy or silicone (either water-clear or tinted) will create a well-defined beam with a view angle as narrow as 4 degrees or as wide as 160 degrees (in a few instances). The most common view angles for LED indicators are 30 degrees and 60 degrees.

Datasheets for LED indicators often include a *spatial distribution* graph showing the relative intensity of the light when viewed at various angles from the axis of the LED. The spatial distribution graph in Figure 22-10 is for an LED with a view angle defined as 40 degrees. This is the angle at which the relative luminous intensity diminishes to 50%.

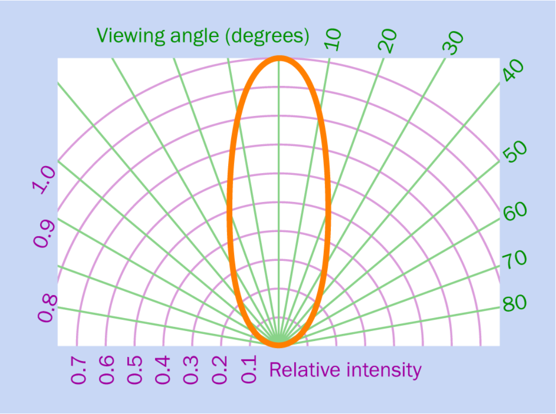

Figure 22-10. *A spatial distribution graph shows the relative intensity of light from an LED at various view angles.*

The view angle is of special concern in devices such as flashlights, where the spread of the beam affects the functionality.

How to Use It

Like all semiconductor devices, LEDs can be impaired by excess forward current and may break down irreversibly if subjected to excess reverse voltage. Their limits for reverse voltage are much lower than those of a rectifier diode. They are also vulnerable to heat, but are not particularly vulnerable to static electricity.

Polarity

A through-hole LED will have two leads of unequal length. The longer lead connects internally with the *anode* of the diode, and should be wired externally to the “more positive” side of a power source. The shorter lead connects internally with the *cathode* of the diode, and should be wired externally to the “more negative” side of a power source.

To remember the functions of the leads, consider that the plus sign would be twice as long as a minus sign if its horizontal and vertical stroke were disassembled and placed end to end.

If a round LED has a flange around its base, a flat spot in the flange will be closest to the cathode side of the component.

Series Resistor Value

Because the effective internal resistance of a diode is not a constant value at different voltages, a trial-and-error approach may be necessary to determine the ideal value for a series resistor with an LED indicator. For this purpose, a trimmer potentiometer can be used with a sample LED while measuring the current through it and the voltage drop across it. A fixed-value resistor can then be substituted. If the choice is between a resistor value that is a little too high and a value that is a little too low, the higher value resistor should be used.

An approximate value can be found using a very simple formula in which R is the resistor value, V_{CC} is the supply voltage, V_F is the forward voltage specified for the LED, and I is the desired current:

$$R = (V_{CC} - V_F) / I$$

Normally a series resistor rated at 1/4 watt will be acceptable, and 1/8 watt may be used in 5VDC circuits. However, care should be taken with a power supply of 9V or higher. Suppose an LED is rated for 1.8V forward voltage at 20mA. In a 5V circuit, the voltage drop across the series resistor will be:

```
V = 5 - 1.8 = 3.2
```

Therefore, the resistor must dissipate 3.2V * 20mA = 64mW. This is comfortably below the 125mW rating of a 1/8 watt resistor. However, with a 9V power supply, the voltage drop across the series resistor will be:

```
V = 9 - 1.8 = 7.2
```

Now the resistor must dissipate 7.2V * 20mA = 144mW. This exceeds the 125mW limit for a 1/8 watt resistor.

LEDs in Parallel

If multiple LEDs are to be driven in parallel, and none of them has to be switched individually, it is naturally tempting to save time by using a single series resistor for all of them. In these circumstances, assess the maximum current carefully and multiply by the voltage drop imposed by each of the LEDs, to determine the wattage of a series resistor.

Linking dissimilar LEDs in parallel is not recommended, because the threshold voltage decreases with increasing temperature. The hottest LED will therefore receive the largest current, and thus become even hotter. Thermal runaway can result.

LEDs containing their own series resistors can safely be wired in parallel.

Multiple Series LEDs

A series resistor wastes current by dissipating it as heat. In an application where two or more LED indicators will be illuminated simultaneously, the LEDs can be connected in series with a lower-value resistor, and three LEDs in series may eliminate the need for a resistor completely, depending on the voltage of the power supply. Here again a trimmer potentiometer should be used to determine an ideal value for any series resistor that may be necessary.

Comparisons with Other Light Emitters

Because LED indicators have largely replaced neon bulbs and miniature incandescent lamps, comparisons are of limited importance at this point. The situation regarding **LED area lighting** is different in that it is still competing actively with fluorescent lights and, in some instances, halogen. A list of advantages and disadvantages for high-intensity white LEDs is given in "Comparisons" on page 223. The advantages of incandescent lamps are listed in "Relative Advantages" on page 179.

Other Applications

LEDs are used in **optocouplers** and in **solid-state relays**. Usually an infrared LED is embedded inside a chip or a plastic module, and emits light through an interior channel to activate a phototransistor. This arrangement provides electrical isolation between the switching signal and the switched current.

Some sensors use an LED paired with a phototransistor at opposite sides of a U-shaped plastic mount. A sensor of this type can monitor industrial processes or may be found inside a photocopy machine, to detect the presence of a sheet of paper.

What Can Go Wrong

Excessive Forward Voltage

Like any diode, the LED has a *threshold voltage* in the forward direction. If this threshold is exceeded, the effective internal resistance of the LED falls very rapidly. Current rises equally rapidly, and quickly damages the component, unless it is protected by an appropriate series resistor.

Excessive Current and Heat

Exceeding the recommended value for forward current, or allowing an LED to overheat, will shorten its lifetime and cause a premature dimming of light output. LEDs generally require some current limiting or regulation (most commonly with a series resistor). They should not be connected directly to a voltage source such as a battery, even if the battery voltage matches the voltage of the diode. The exception to this rule is if the internal resistance of the battery is high enough to limit the current, as in the case of button-cell batteries.

Storage Issues

LEDs of different types are often indistinguishable from each other. They can also be indistinguishable from photodiodes and phototransistors. Careful storage is mandatory, and reusing LEDs that have been breadboarded may cause future problems if they are wrongly identified.

Polarity

If the leads on an LED indicator are trimmed, and if the indicator lacks a flange in which a flat spot will identify the cathode, the component is easily misused with reversed polarity. If it is connected with a component that has limited current sourcing capability (for instance, the output pin of a digital chip), the LED will probably survive this treatment. However, maximum reverse voltage is often as low as 5VDC. To minimize the risk of errors, the anode lead can be left slightly longer than the cathode lead when they are trimmed for insertion in a breadboard or perforated board.

Internal Resistors

As previously noted, it is difficult to distinguish an LED that contains its own series resistor from another LED that does not. The two types should be stored separately, and should be reused circumspectly.

23

LED area lighting

The term **LED area lighting** is used in this encyclopedia to describe a white LED source that is bright enough to illuminate rooms, offices, or outdoor areas. It may also be used in desk lamps or table lamps as *task lighting*. LEDs for these purposes may be categorized as *high-brightness*, *high-power*, *high-output*, or *high-intensity*. A complete fixture containing at least one light source is properly known as a *luminaire*, although the term is not uniformly applied and is sometimes written incorrectly as a *luminary*.

The full term *light-emitting diode* is not normally applied to an LED used for area lighting. For this purpose, the LED acronym has become universal. Periods are not normally placed between the letters.

While an LED area-lighting package may contain more than one diode, it is still categorized here as a single source. By contrast, any component consisting of multiple separately discernible light-emitting diodes, as in a seven-segment numeral, a 14- and 16-segment alphanumeric character, a dot-matrix character, or a display of multiple characters, is listed in a separate entry as an **LED display**.

The term *OLED* is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminiscent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **incandescent lamp** (see Chapter 18)
- **fluorescent light** (see Chapter 20)
- **neon bulb** (see Chapter 19)
- **electroluminescence** (see Chapter 26)

What It Does

High-brightness white LEDs provide a plug-compatible alternative to **incandescent lamps**, *halogen lighting*, and **fluorescent lights** for work spaces and the home.

At the time of writing, products are still evolving rapidly in the field of LED area lighting. A shared goal of manufacturers is to increase efficiency while reducing retail price to the point where high-brightness LEDs will displace fluorescent tubes for most low-cost lighting applications.

A wall-mounted LED reflector-bulb that emulates a halogen fixture is shown in Figure 23-1. A small LED floodlight for exterior use is shown in Figure 23-2. An early attempt to package an LED area light in a traditional-style bulb is shown in Figure 23-3. Within a decade, as LED area lighting continues to evolve, some of these examples may look quaint. Configurations are evolving, with final results that remain to be seen.

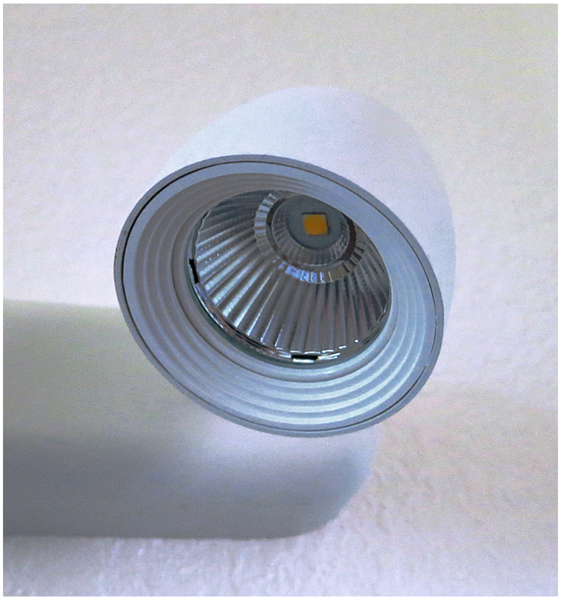

Figure 23-1. *A small LED reflector-light emulating a halogen fixture. Note the square of yellow phosphors mounted on the diode.*

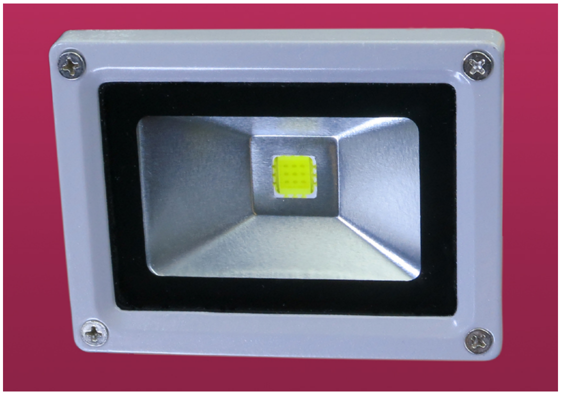

Figure 23-2. *A floodlight for exterior use. Nine LEDs are assembled behind the yellow phosphors. The steel frame measures about 4" by 3".*

Trends in Cost and Efficiency

The *luminous flux* of a source is the total power that it radiates in all directions, within the visible spectrum. The unit that measures luminous flux is the *lumen*. For a more detailed discussion of this topic, see "Power" on page 177. Since 1965, the cost per lumen of light from a given color of LED has decreased by about a factor of 10, while the maximum number of lumens emitted by an LED package has increased by a factor of 20, during each decade. This is known as Haitz's Law, named after Dr. Roland Haitz of Agilent Technologies. Figure 23-4 illustrates it graphically.

Figure 23-3. *An LED light bulb. Unlike an incandescent bulb, it focuses the illumination in one direction, like a reflector-light. Consuming only 6W, it is claimed to be equivalent to a 40W incandescent bulb.*

Schematic Symbol

Schematic symbols that are commonly used to represent an LED are shown in Figure 23-5. The symbol remains the same regardless of the size or power of the component, but architectural plans may represent any type of light using the circle-and-X symbol at bottom right.

How It Works

A high-brighteness LED functions on the same basis explained in the entry describing **LED indicators**. Photons are emitted when electrons are sufficiently energized to cross a PN junction and combine with electron-holes.

An LED that appears white, or off-white, actually emits blue light that is re-radiated over a wide range of wavelengths by adding a layer of yellow phosphors to the chip. A cutaway diagram of an

LED chip (properly known as a *die*), mounted under a silicone lens, is shown in Figure 23-6.

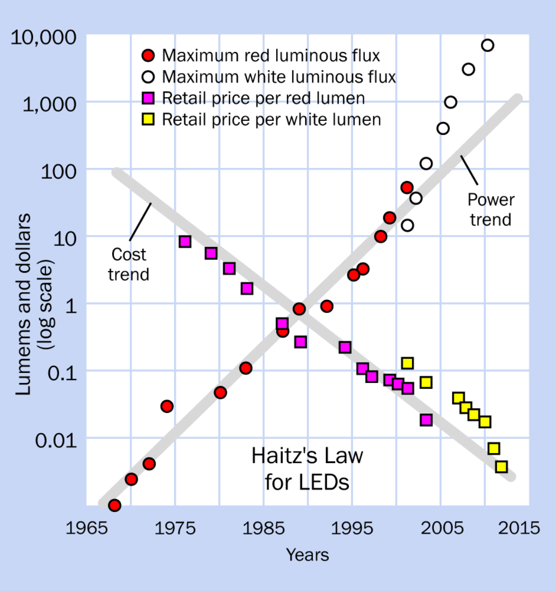

Figure 23-4. *The increase in light output (luminous flux, in lumens) of a single LED compared with the decrease in cost-per-lumen during the years since 1965. The vertical logarithmic scale measures both dollars and lumens. Source: Philips Gardco site-lighting fact sheet with additional data from a "Strategies in Light Report" published by Semiconductor Equipment and Materials International in 2013.*

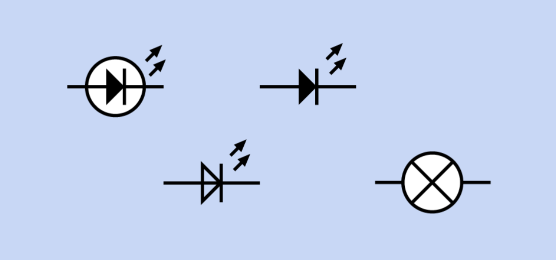

Figure 23-5. *The symbol for an LED remains the same regardless of its size and power, but architectural plans may use the circle-and-X symbol at bottom right for any type of light fixture.*

LEDs are mass-produced by etching them into crystals that are then cut into *wafers* before being subdivided into *dies*, like silicon chips. Most of the blue LEDs that form the basis of white lighting use sapphire crystals as their substrate. The crystal may range in diameter from two inches to six inches. Large sapphire wafers are also finding potential applications in camera lens covers and scratch-resistant cover plates for cellular phones.

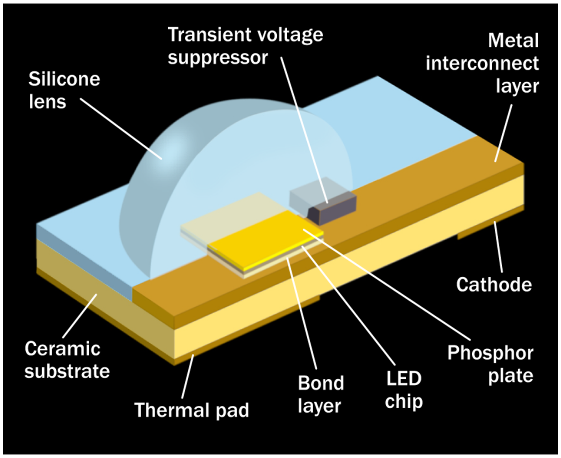

Figure 23-6. *Cutaway diagram of a high-brightness white LED. Adapted from Philips Lumileds Technical Reference document.*

While a die for an LED indicator may be 0.3mm x 0.3mm, a die in a high-brightness LED is often 1mm x 1mm. The size is limited by technical issues involving total internal reflection of the generated light.

The precise color of the light is adjusted by adding red phosphors to the yellow phosphors. This reduces the overall efficiency of the LED by around 10% but creates a "warmer" radiance. A graphical illustration of this principle is shown in Figure 22-3, in the entry on LED indicators.

The *color temperature* of white or offwhite light is measured in degrees Kelvin, typically ranging from 2,500K to 6,500K, where a lower number represents a light with more red in it and a higher number represents a light with more blue in it. This system of measurement was originally used with **incandescent bulbs** to define the temperature of the filament, which determined its color. See "Spectrum" on page 173 for a detailed explanation.

Visible Differences

The effects of different types of illumination are compared in Figure 23-7. To create this figure, first a color chart was prepared in Photoshop and printed on high-gloss photo-grade paper with a Canon Pro9000 Mark II inkjet printer, which has separate colors for red and green in addition to cyan, pale cyan, magenta, pale magenta, yellow, and black.

The color chart was then photographed twice with a Canon 5D Mark II, using a fixed white balance of 4000K. The first exposure was made with "daylight spectrum" LED lighting (claimed color temperature of 6500K) while the second was made with halogen lighting (claimed color temperature of 2900K). The photographs were adjusted in Photoshop for levels only, to fill the available range of 256 values. The two exposures show how the same chart would appear when viewed under the different lights, if the human eye did not adjust itself at all. Note the large area of the LED exposure which is rendered in shades of blue or purple. Also note the dullness of the reds. This confirms the everyday belief that "daylight spectrum" LEDs tend to have a cold, purplish cast while incandescents have a warmer, yellow look.

The same camera was then used to make two more exposures, this time with the white balance set to 6500 for LED lighting and 2900 for halogen lighting, which would be the recommended standard procedure, suggesting the kind of compensation that the human eye also tends to make for different ambient lighting. The result is shown at Figure 23-8. The LED version has improved, but the reds and yellows are still muted. The halogen version also looks better than before, but the magenta end of the spectrum has too much yellow in it. These images show the limits of white-balance correction for indoor photography.

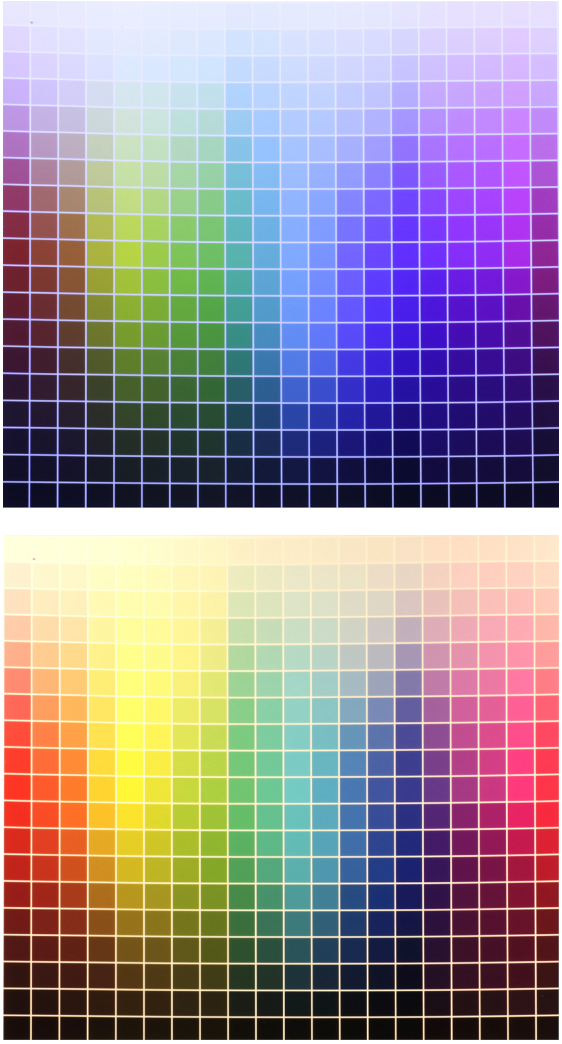

Figure 23-7. *The same printed color chart viewed with "daylight white" LEDs (top) and halogen lighting (bottom), without any compensation to allow for the different light spectra. A fixed white balance of 4000K was used for both pictures.*

Side-by-Side Comparison

Because the human eye is much better at comparing colors when they are adjacent to each other, another color chart was prepared using just six color bars of fully saturated red, yellow, green, cyan, blue, and magenta, with paler and darker versions above and below. The bars were separated with large white gaps. The chart was photographed first with the white balance set to

6500 under "daylight white" LED lighting and then again with the white balance at 2900 for halogen lighting. In Photoshop, the bars from the halogen version were copied and pasted beside the bars from the LED version to facilitate an A-B comparison. The result is shown in Figure 23-9.

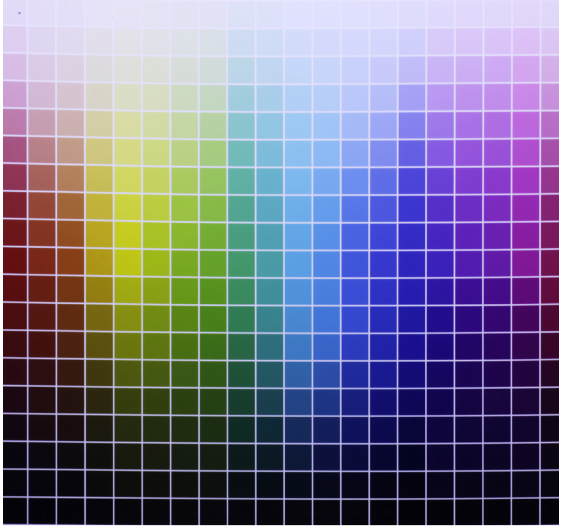

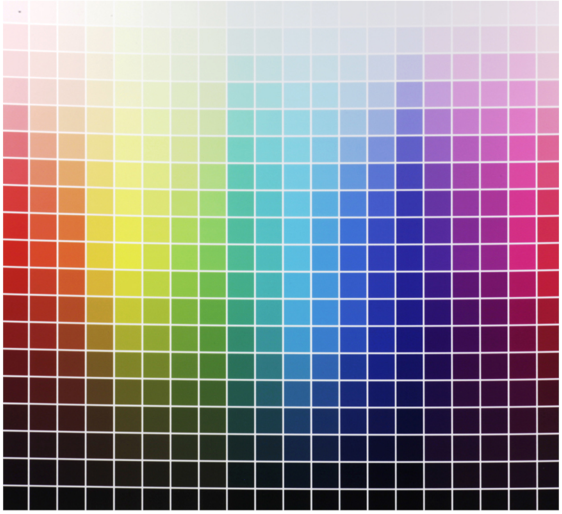

Figure 23-8. *The same color chart as before, photographed with appropriate color temperature settings of 6500K (top, using "daylight-spectrum" LEDs) and 2900K (bottom, using halogen).*

In each pair of colors, LEDs lit the one on the left, and halogen lights lit the one on the right. This figure shows the dramatic difference at the red end of the spectrum, and the poor reproduction of yellows by LED lighting. However, the LED rendering of green is better, and likewise the rendering of magenta, except where its darker version is concerned. Among the pale versions of the colors, the LED lights produce much less density (i.e., they have a brighter look) in the blues, greens, and cyans. The low densities will show up as pale highlights in a photograph of an object, and the picture will tend to have excessive contrast. This will also contribute to the "harsh" look of "daylight white" LED lighting which may be perceived by the eye.

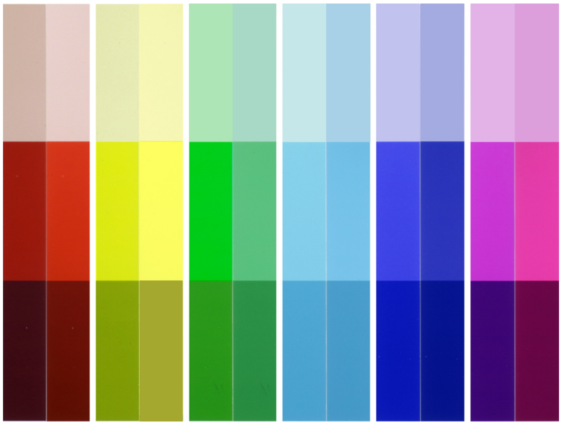

Figure 23-9. *A range of six fully saturated colors, with lighter and darker shades added above and below, photographed first with "daylight spectrum" LEDs and then with halogen lights, after which the two sets of colors were paired for easy comparison. The LED version is on the left in each pair.*

Halogen is deficient at the blue-violet end of the spectrum, even when the camera has an appropriate white-balance setting. Photographers can correct this using image-editing software. LED "daylight spectrum" lights are more difficult to correct. LEDs classified as "warm" should reproduce reds better, but may not do so well with blues.

Diffuse light from a uniformly cloudy sky may be the most ideal form of lighting for photographing objects, but this is of little help for people who work (or take photographs) under artificial lights.

LED lights that contain separate red, green, and blue emitters may perform better, but create a different problem, in that shadows will tend to have color fringes caused by the small offsets between the colored emitters.

Heat Dissipation

An LED is less than 100% efficient because not all electrons mate with electron holes. Some manage to bypass the semiconductor junction; some recombine without generating light; and some transfer their energy to other atoms. In each instance, waste heat is created. While the heat in an incandescent bulb is mostly dissipated by radiation, an LED must get rid of the heat almost entirely by conduction, typically through a *heat sink*. This complicates the design of a fixture, because the integrity of the pathway to dispose of the heat must be retained when the LED bulb or tube is replaced.

Efficacy

The *radiant luminous efficacy* (LER) of an incandescent light source measures how effective it is at channeling its output within the visible spectrum instead of wasting it in infrared radiation. LER is expressed in lumens per watt, and in an incandescent bulb, it is calculated by dividing the power emitted in the visible spectrum (the *luminous flux*) by the power emitted over all wavelengths. This is described in detail in "Efficacy" on page 179 in the entry dealing with incandescent lamps.

In an LED indicator, almost all the radiation can be contained within the visible spectrum, which suggests that its efficacy should be 100%. However, because some waste heat is still created internally, the efficacy is calculated by dividing the light output, in lumens, by the power input, in watts, at the voltage required by the LED. (Lumens can be converted directly to watts, and therefore this division makes a comparison between similar units).

In an LED lighting fixture that contains its own electronics to convert higher voltage AC to lower voltage DC, the power consumption of the fixture is measured not at the diode, but at the input side of the electronics. Therefore, the inefficiency of the electronics reduces the efficacy value of the lighting unit.

Dimming

An incandescent bulb is very sensitive to reduction in power. It becomes radically inefficient, emitting perhaps 1% of its normal light output if the power is reduced to 40%.

LEDs have an almost linear response to the supplied power. Usually a triac-based dimmer will not work well with LED area lighting, and a dimmer designed for LEDs must be substituted, using pulse-width modulation.

Ultraviolet Output

The gas plasma in a fluorescent light generates ultraviolet wavelengths that are shifted to the visible spectrum by the phosphor coating inside the glass envelope. Imperfections in the phosphor coating can potentially allow leakage of ultraviolet light, causing some researchers to claim that the use of CFLs (compact fluorescent lighting) for close-up work with desk lamps can increase the risk of developing some forms of skin cancer. (This claim remains controversial.)

LED manufacturers are quick to point out that white LEDs do not emit any ultraviolet radiation. Figure 23-10 shows spectral power distribution curves derived from measurements of three high-brightness Color Kinetics LEDs manufactured by Philips. The manufacturer states categorically that "The LED-based color and white light products made by Color Kinetics do not emit outside the visible spectrum." Infrared radiation is also negligible.

Color Variation

The *correlated color temperature* (CCT) is determined by finding the conventional incandescent color temperature which looks most similar to the light from a white LED. Unfortunately, because the CCT standard is insufficiently precise, and because small manufacturing inaccuracies can occur, two LED sources with the same CCT number may still appear different when they are side by side. While the human eye adjusts itself to overall color temperature, it is sensitive to differences between adjacent sources. If two or more white LEDs in a lighting fixture do not have identical spectra, the difference will be noticeable.

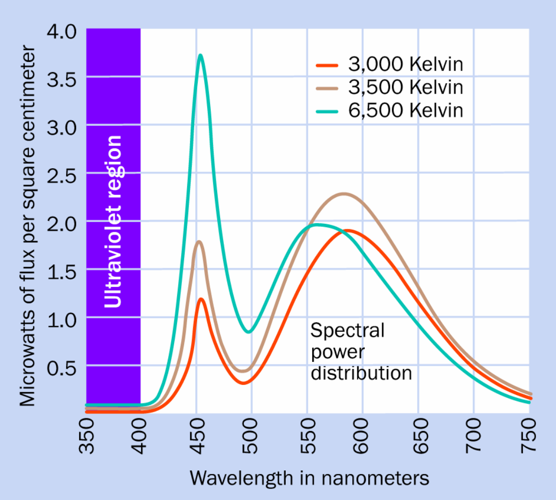

Figure 23-10. *Spectral power distribution curves for three high-brightness white LED lamps showing no ultraviolet emissions. (Adapted from a graph in a white paper published by Color Kinetics Incorporated.)*

To address the issue, manufacturers introduced the concept of "binning," in which lights are subclassified to tighter specifications and are assigned bin numbers based on their measured characteristics. The Philips Optibin system, for instance, assesses the light from various angles, as well as perpendicularly to the source. This is especially important where a large area such as a building lobby is painted white and is lit by high-brightness LEDs that must appear uniform in color temperature.

Variants

LED area lighting products are often designed to emulate the form factors of incandescent bulbs, halogen reflector bulbs, or fluorescent tubes. The standard screw-in base of an LED light bulb, the pin-base of a miniature 12V LED reflector bulb, and the pins on an LED tube enable easy migration to the newer technology.

Strip lights are unique to LED lighting systems. They are thick, flexible plastic ribbons in which are embedded a series of LEDs. For area lighting, the LEDs are white, and the strips can contain necessary control electronics for conversion of AC power. The strips can be placed behind ledges or moldings to provide soft, even illumination of the ceiling above.

Strip lights are also available for 12VDC power, to create lighting effects in customized automobiles and trucks. These strip lights are available in various colors in addition to white. Many have multicolor capability and can be controlled with a handheld remote.

Comparisons

The advantages of an **incandescent lamp** are listed in "Relative Advantages" on page 179, while advantages of **fluorescent lights** are listed in "Comparisons" on page 194. These lists can be compared with the following advantages for LED area lighting:

- While the life of an incandescent lamp for room lighting can be as little as 1,000 hours, LED area lighting typically claims up to 50,000 hours.
- The lifetime of an incandescent bulb is the average time it can emit light before catastrophic failure. The lifetime of an LED is the average time it can emit light before gradually dimming to 70% of its rated output. This is a much gentler, less inconvenient failure mode that does not require immediate replacement.
- Unlike a fluorescent light or incandescent bulb, the LED does not contain hot tungsten that fails as a result of erosion.

- Unlike a fluorescent light, an LED does not contain mercury, and therefore does not require special recycling arrangements that entail associated fees.
- While fluorescents can have difficulty starting in low temperatures, an LED is not sensitive to a cold environment.
- Bright LEDs are available in a wide range of colors that do not require filtering. Filters greatly reduce the efficiency of incandescent bulbs when they are used in applications such as traffic signals or rear lights on automobiles.
- High-brightness LEDs can be dimmable. Fluorescent lights are usually not dimmable, or perform poorly in this role.
- LEDs are inherently directional, because the die radiates light at an angle of 90 degrees to its plane. This makes it ideal for ceiling mounting, where as much light as possible should be directed downward. A fluorescent tube or incandescent bulb often requires a reflector which reduces the overall efficiency.
- LEDs are insensitive to cycling. The life expectancy of an incandescent bulb or (especially) a fluorescent tube is reduced by cycling it on and off.
- No flickering. Fluorescent tubes may start to flicker as they age.
- No electrical interference. Fluorescent tubes can interfere with AM radio reception and some audio devices.
- Safe from breakage. LED area lighting does not necessarily use any glass.

However, high-brightness LEDs still have some barriers to overcome:

- Cost. In the United States, before 60W incandescent bulbs were legislated out of existence, they could be sold profitably for less than $1 each. A T8 fluorescent tube, measuring 1” diameter and 48” long, currently costs between $5 and $6 (retail) but has a life expectancy in the region of 25,000 hours, and uses only 20% of the power of an incandescent bulb to generate two to three times as much light. Clearly the fluorescent tube is a more economical choice, despite the price of the electronics that must be included in the fixture to start the tube. By comparison, currently the purchase price of an LED tube is three times that of a fluorescent tube. It may last twice as long, but is not significantly more efficient, generating perhaps 100 lumens per watt while a fluorescent is typically capable of 90 lumens per watt. Prototype high-brightness LEDs have exceeded 200 lumens per watt, and should be competitive with fluorescents by 2020, but even then, migration will take time.
- Heat sensitivity. Heat reduces the light output and the lifespan of LED fixtures.
- Placement issues. Because LEDs are heat sensitive, they must be installed in locations that do not become excessively hot, their heat sinks must be correctly oriented, and they must have adequate ventilation.
- Color shift. Heat and age may cause the color temperature of an LED to shift slightly, as the color is usually derived from two types of phosphors.
- Nonuniformity. Manufacturing inconsistencies can cause LEDs of the same type to display slightly different color temperatures. Fluorescents and incandescents are more uniform.
- Lower heat output than incandescents. While this is an advantage from the point of view of efficiency, it can be a disadvantage in applications such as traffic signals or airport runway lighting where waste heat can help to keep the lights free from snow or ice.

Values

Although the output from an LED area light is directional, while the output from an incandescent bulb or a fluorescent light is omnidirectional, the intensity is measured the same way in each instance, using *lumens*. This unit expresses the total light emission, without taking directionality into account. (The intensity of **LED indicators** is calibrated in candelas, which measure the power within an angle of dispersion; but candelas are not used for area lighting.)

Typical values for incandescent bulbs are 450 lumens for a power consumption of 40 watts, 800 lumens for a consumption of 60 watts, 1,100 lumens for a consumption of 75 watts, and 1,600 lumens for a consumption of 100 watts. Because much of the output from an incandescent bulb may be wasted by using inefficient reflectors or allowing the light to shine in directions where it is not needed, a high-brightness LED rated at 1,000 lumens may actually appear brighter than a 75-watt incandescent bulb.

A T8 fluorescent tube measuring 48″ long by 1″ in diameter consumes only 32 watts but emits almost 3,000 lumens—when it is new. This output gradually diminishes by as much as 40% over the lifetime of the tube.

Incandescent bulbs deliver between 10 and 15 lumens per watt, approximately. A new fluorescent tube produces around 80 to 90 lumens per watt, and LED area lighting at the time of writing can provide 100 lumens per watt, under real-world conditions.

What Can Go Wrong

Wrong Voltage

Many high-brightness LED lighting units can be used with either 115VAC or 230VAC. There are exceptions, however. Check the specifications to make sure. Also, it is important to avoid applying domestic supply voltage to 12V LED miniature reflector-bulbs that are intended to replace 12VAC halogen bulbs of the same size.

Overheating

If a high-brightness LED fixture is equipped with a heat sink, this must be exposed to freely flowing air. Any vanes on the heat sink should be oriented vertically to encourage convection, and the fixture must not be placed in an enclosure. Overheating will radically shorten LED life.

Fluorescent Ballast Issues

A fluorescent fixture contains a *ballast* to limit the tendency of the tube to draw excessive current. The ballast is contained in a plastic box attached to the back of the frame in which the tube is mounted.

A *magnetic ballast* contains a coil, and is bypassed by an additional *starter* that applies unlimited current for one second when the power is switched on, preheating the tube to initiate plasma discharge.

An *electronic ballast* performs the same function without a separate starter.

Some LED tubes designed as substitutes for fluorescent tubes may allow a magnetic ballast to remain in the circuit, but may not tolerate an electronic ballast. Other LED tubes require any type of ballast to be unwired from the circuit. The unwiring operation will require disconnection of a couple of wires by removing wire nuts (assuming that the fixture has been designed to comply with U.S. building codes). The wires are then reconnected to apply power directly to the tube, and the wire nuts are reapplied to complete the new connection. The ballast can remain passively in the fixture.

Failing to remove the ballast and/or the starter from a fluorescent fixture before installing an LED tube that requires direct connection to the power supply can damage the tube. Connecting the power incorrectly to the LED tube may result in it failing to light up. Documentation supplied with the LED tube should provide guidance for disconnecting the ballast and connecting the tube. Note that the pin functions on LED tubes are not standardized at this time.

Misleading Color Representation

Because the spectrum of a white LED is not evenly weighted across all wavelengths, it will fail to represent some colors accurately, as shown previously. This can be important if LEDs are used to illuminate full-color printing or artwork, or if they are installed in stores selling merchandise such as clothes, furnishings, or food.

LED display

24

In this encyclopedia, a component consisting of multiple separately discernible light-emitting diodes, such as a seven-segment numeral, 14- or 16-segment alphanumeric character, a dot-matrix character, or a display module containing multiple characters, is categorized as an **LED display**. The term *light-emitting diode* is hardly ever used to describe an LED display, as the LED acronym has become ubiquitous. The acronym does not usually include periods between the letters.

An **LED indicator** is defined here as a component usually 5mm or smaller in diameter, made of transparent or translucent epoxy or silicone, most often containing one *light-emitting diode*. It is purposed as a status indicator in a device, rather than as a source of illumination, and is sometimes referred to as a *standard LED*.

LEDs that are designed to illuminate large living or working areas are discussed in a separate entry as **LED area lighting**. They are sometimes referred to as *high-brightness LEDs* and almost always emit white light.

The term *OLED* is an acronym for *Organic Light-Emitting Diode*, a thin panel in which an organic compound is contained between two flat electrodes. Despite its functionality as a form of LED, its design is similar to that of thin-film electroluminiscent light sources. Therefore it is discussed in the entry on **electroluminescence**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LED area lighting** (see Chapter 23)
- **vacuum-fluorescent** (see Chapter 25)
- **electroluminescence** (see Chapter 26)
- **LCD** (see Chapter 17)

What It Does

An LED display presents information on a panel or screen by using multiple segments that emit light in response to a DC current, almost always at a voltage ranging between 2VDC and 5VDC. The display may contain alphanumeric characters and/or symbols; simple geometrical shapes; dots; or pixels that constitute a *bitmap*.

A *liquid-crystal display*, or **LCD**, serves the same purpose as an LED display and may appear very similar, except that a liquid crystal reflects incident light while an LED emits light. The increasing use of backlighting with LCDs has made them appear more similar to LED displays.

There is no schematic symbol to represent an LED display. Where a segmented display is used, often the segments are represented with drawn outlines.

The simplest, most basic, and probably the best-known example of an LED display is the *seven-*

segment numeral, one of which is shown in Figure 24-1. This is a Kingbright HDSP-313E with a character height of 0.4″.

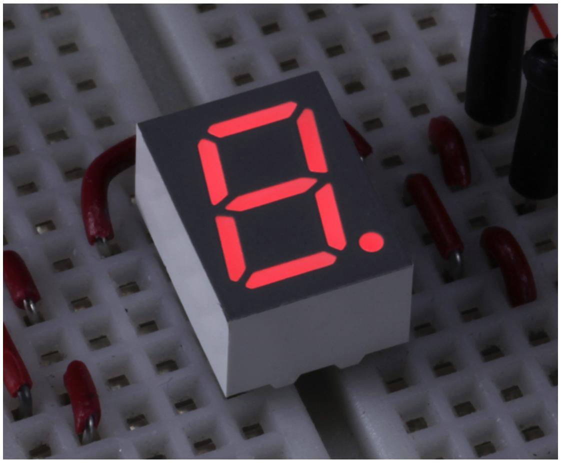

Figure 24-1. *The most basic LED display, able to create numerals from 0 through 9 using seven light-emitting segments that can be illuminated individually. An eighth segment forms the decimal point.*

How It Works

The process by which an LED generates light is explained in "How It Works" on page 207, in the entry dealing with **LED indicators**. Each light-emitting diode in an LED display is functionally the same as the diode in an LED indicator.

LEDs must be driven with DC. This is a primary distinction between an LED display and an LCD, which requires AC.

Variants

LCD comparisons

LCDs and LED displays can look very similar. This raises the obvious question: which is appropriate for a particular application?

LCDs (without backlighting) are more appropriate for applications such as digital watches and solar-powered calculators where power consumption must be minimized. They are capable of running for years from a single button cell.

LCDs are easily visible in bright ambient light, where LED displays are not. LCDs can also be designed to display complex pictographic shapes and symbols, while the segments of an LED display are more constrained to be simple in shape.

An LCD is more likely to be affected by temperature than an LED, and powering it entails some slight inconvenience, because it requires an AC source that is unlikely to be useful elsewhere in a circuit. If the LCD uses LED backlighting, it will also require a low-voltage DC power source for the backlight. An LED display is easier to use in that it can be driven directly from a microcontroller or logic chip, with only some series resistors to limit the current, and the addition of transistors to provide additional power where necessary.

Seven-Segment Displays

Early seven-segment LED displays were used in digital calculators, before LCDs became an affordable, practical alternative that greatly extended battery life. Initially, the size of the diodes was limited, sometimes requiring magnifying lenses to make them legible.

Seven-segment displays are still used in some low-cost applications, although LCDs have become more common.

Figure 24-2 shows how the segments are identified with letters *a* through *g*. This scheme is used universally in datasheets, and is also used for LCDs. The decimal point, customarily referred to as "dp," is omitted from some displays. The segments are slanted forward to enable more acceptable reproduction of the diagonal stroke in numeral 7.

Although seven-segment displays are not elegant in appearance, they are functional and are reasonably easy to read. They also enable the representation of hexadecimal numbers using letters A, B, C, D, E, and F (displayed as A, b, c, d, E, F because of the restrictions imposed by the small number of segments), as shown in Figure 24-3.

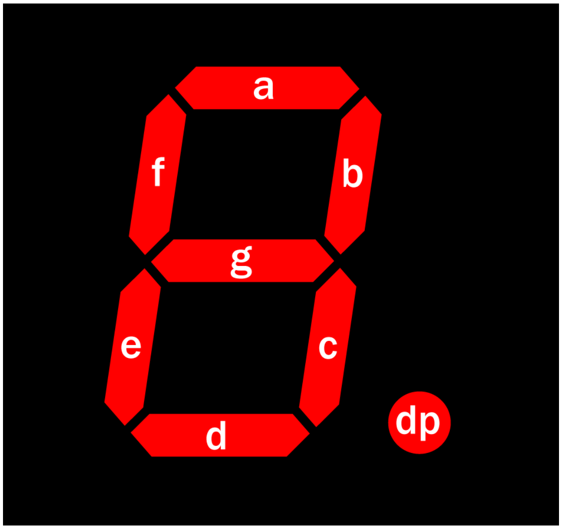

Figure 24-2. *A seven-segment LED display. The lowercase identifying letters are universally used in datasheets.*

In appliances such as microwave ovens, very basic text messages can be displayed to the user within the limitations of seven-segment displays, as suggested in Figure 24-4.

Numbers 0, 1, and 5 cannot be distinguished from letters O, I, and S, while letters containing diagonal strokes, such as K, M, N, V, W, X, and Z, cannot be displayed at all.

Multiple Numerals

Displays consisting of a single numeral are now rare, as few applications require only one digit. Displays of two, three, and four digits are more common, as shown in Figure 24-5.

Additional Segments

Displays with 14 or 16 segments were introduced in an effort to enable the representation of all the letters of the alphabet. The segment layout of these LED displays is identical to that of comparable LCDs. The differences between 14-segment and 16-segment displays are shown in Figure 24-6. Some are angled forward like seven-segment displays, even though the addition of diagonal segments makes this unnecessary for display of characters such as numeral 7.

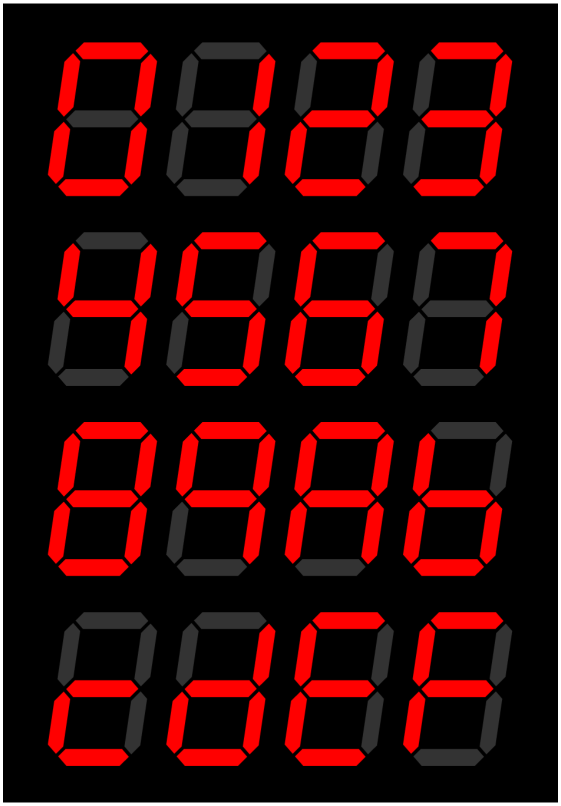

Figure 24-3. *Numerals and the first six letters of the alphabet created with seven-segment displays.*

Figure 24-7 shows the scheme for identifying the segments of a 16-segment display. This naming convention is used in all datasheets. The lowercase letters that were customary with seven-segment displays are usually abandoned in favor of uppercase, perhaps to avoid confusion with the letter L. Note that letter I is omitted from the sequence.

For a complete alphanumeric character set enabled by a 16-segment display, see Figure 17-9 in the entry discussing LCDs.

An example of a 16-segment alphanumeric LED display is shown in Figure 24-8, mounted on a breadboard and wired to show the letter N. This is a Lumex LDS-F8002RI with a character height of 0.8". The component is still available at the time of writing, but in limited quantities.

Generally speaking, 16-segment displays were never very popular, because the gaps between adjacent segments impaired legibility. LED versions remain more readily available than LCD versions, but dot-matrix displays allow a better-looking, more easily legible alphabet, with the added possibility of simple graphics.

Dot-Matrix Displays

In the 1980s, some personal computers used a video character set in which each letter, numeral, punctuation mark, and special character was formed on a video screen from a fixed-size matrix of dots. A similar alphabet is now used in LED dot-matrix displays (and LCDs, as shown in Figure 17-10).

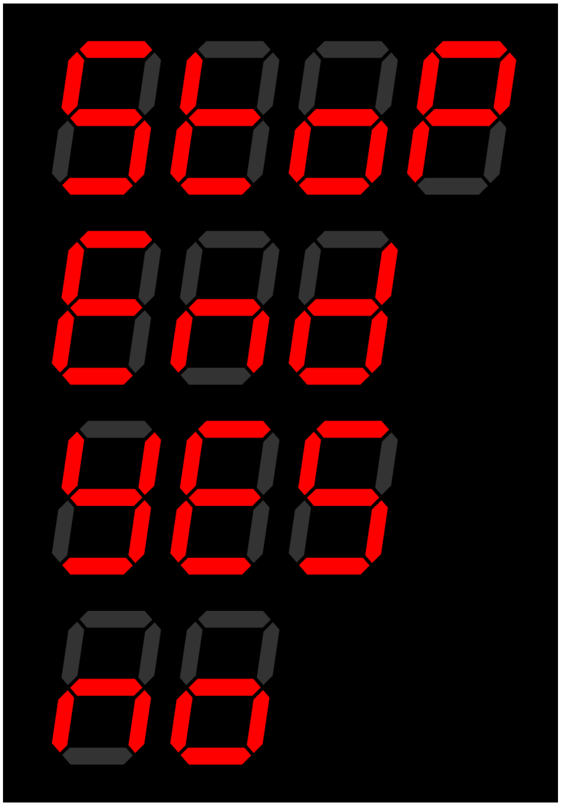

Figure 24-4. *Basic text messages can be generated with seven-segment displays, although they cannot represent alphabetical letters containing diagonal strokes.*

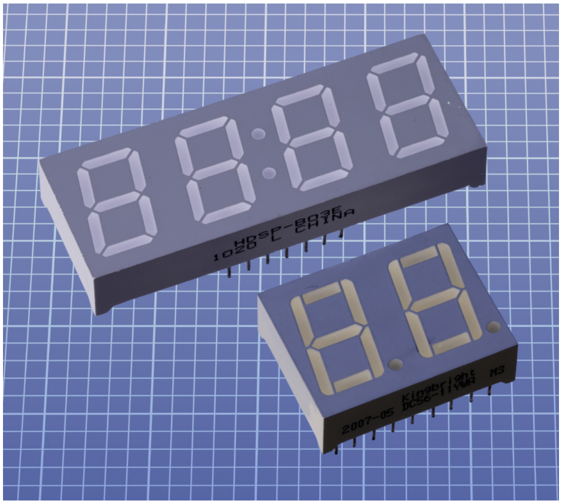

Figure 24-5. *Multiple seven-segment LED displays are often combined in a single component. Top: An Avago 2.05VDC 20mA display designed for a clock. Bottom: A Kingbright two-digit display which draws 20mA at 2.1VDC. The unlit outlines of the numerals would normally be hidden behind panels that are tinted to the same colors emitted by the LED segments when lit.*

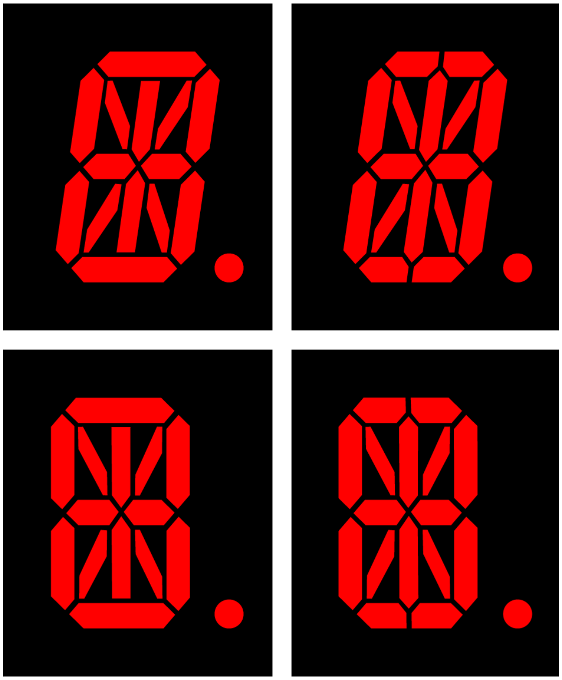

Figure 24-6. *Layouts for 14-segment and 16-segment alphanumeric LEDs are identical to those of LCDs.*

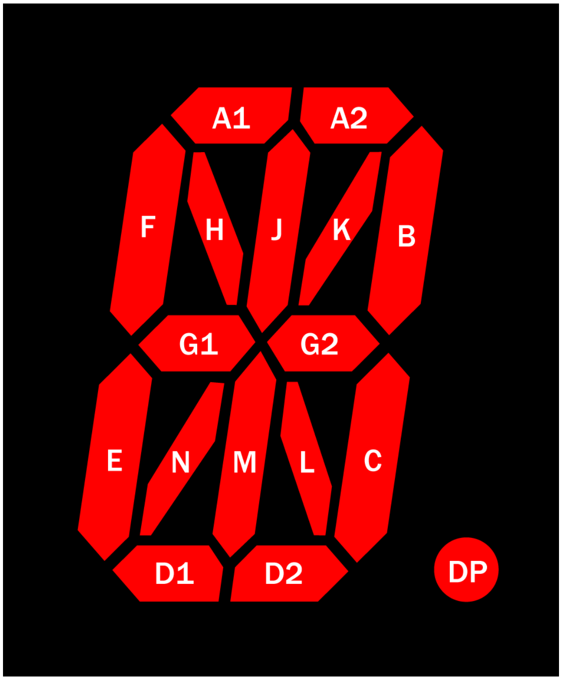

Figure 24-7. *The scheme for identification of segments in a 16-segment alphanumeric LED display.*

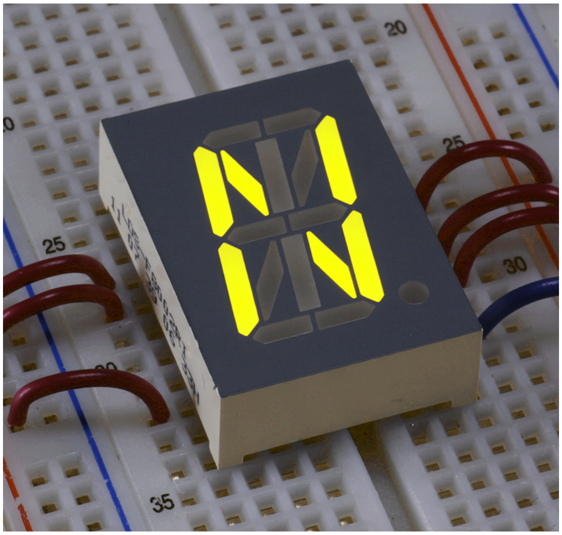

Figure 24-8. *A 16-segment alphanumeric LED display showing the letter N.*

Alphanumeric dot-matrix characters are often grouped in two or more rows with eight or more characters per row. The number of characters is always listed before the number of rows, so that an 8x2 display would contain eight alphanumeric characters in two horizontal rows. This type of component is properly described as a *display module*.

Display modules are used in consumer electronics products such as a stereo receiver where simple status messages and prompts are necessary —for example, to show the tone control settings or the frequency of a radio station. Because the cost of small, full-color, high-resolution LCD screens has been driven down rapidly by the mass production of cellular phones, and because these high-resolution screens are much more versatile, they have already displaced dot-matrix display modules in many automobiles and are likely to follow a similar path in other devices.

Pixel Arrays

The 8x8 pixel array of LED dots shown in Figure 24-9 measures 60mm square (slightly more than 2") and contains 64 LEDs, each approximately 5mm in diameter. Similar arrays are available in other sizes and with different numbers of dots. Displays of the same type may be assembled edge-to-edge to enable scrolling text or simple graphics.

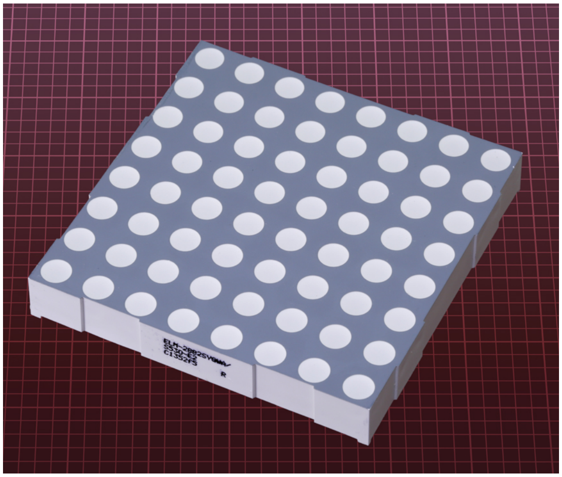

Figure 24-9. *An 8x8 matrix of LED dots measuring more than two inches square.*

Multiple Bar Display

A bar display is a row of small rectangular LEDs in a single component. It may be used for digital representation of an analog signal. The higher the voltage of the signal, the more bars will be illuminated. A typical application would be to show the signal strength of an input to an audio recorder. Ten bars are often used, as in the display shown in Figure 24-10, but multiple components can be combined end to end.

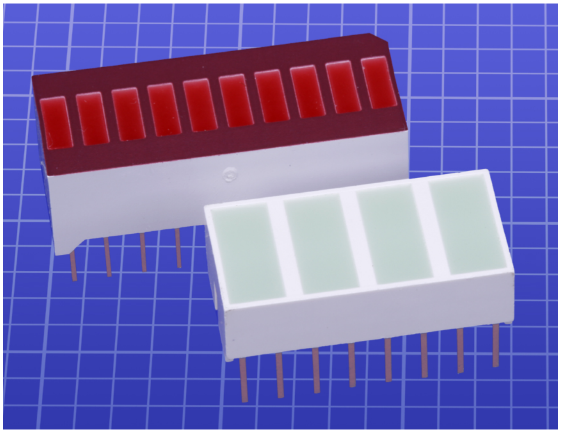

Figure 24-10. *Two LED bar displays in which segments can be lit individually.*

Single Light Bar

A light bar can be thought of as a single-source LED, as it is configured as a single square or rectangle. It is mentioned here, rather than in the entry for single-source LED indicators, because variants may be subdivided into two, three, four, or (sometimes) more discrete sections. These variants are often included in the same datasheet as the monolithic version.

A light bar contains multiple LEDs (often, four in number) behind a translucent panel that provides evenly diffused radiance.

Values

The values for most LED displays are basically the same as for LED indicators, in terms of color, brightness, current consumption, and voltage. See "Values" on page 211 for information.

Multiple-character dot-matrix LED display modules may have different requirements for forward voltage and forward current, depending on drivers that are incorporated in the module. Because there is no standardization for these modules, it will be necessary to consult the manufacturer's datasheet.

How to Use It

Seven-Segment Basics

The diodes in a seven-segment LED display share either a *common anode* or a *common cathode*, the latter being more frequently used. The two types of internal wiring are provided for convenience only. Externally, the displays function identically.

A schematic suggesting the internal wiring and pinouts of a typical ten-pin common-cathode display is shown in Figure 24-11. The pins are numbered as seen from above. Appended to each number is the identity of the segment to which it is connected. Pins 3 and 8 are connected with the cathodes of all the internal LEDs. Both of these pins should be used, to serve as heat sinks for the display.

Note that series resistors are not included inside the display and must be added externally. Their value will be determined by the power supply, to limit the forward current and forward voltage through the LEDs to the extent specified by the manufacturer.

An encapsulated *resistor array* containing either seven or eight resistors in an SIP or DIP chip can be used instead of individual resistors. A seven-segment LED display would require the type of resistor array in which both ends of each resistor are accessible.

Where two or more numerals are combined in a single component, this type of display is likely to have two horizontal lines of pins. In this case, pin 1 will be at the bottom-left corner, seen from above. As always, the pins are numbered counterclockwise, seen from above.

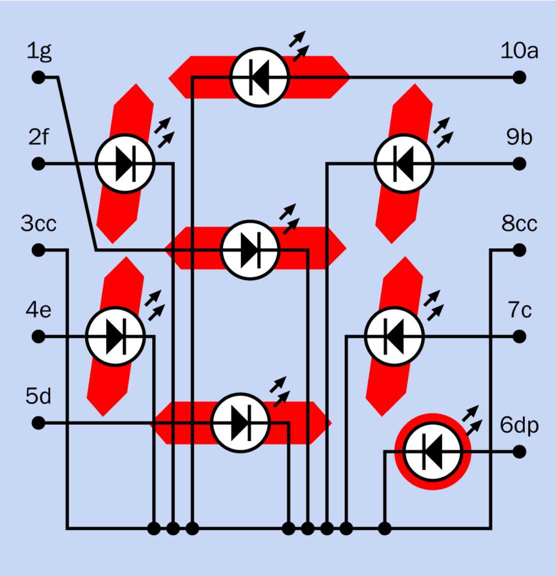
1g
10a
2f
9b
3cc
8cc
4e
7c
5d
6dp

Figure 24-11. *A schematic view of internal connections and pinouts of a seven-segment common-cathode LED display. The numbers refer to the pins of the component, seen from above. The 1 pin may also have a mark beside it on the component, for identification. The orientation of the display can be deduced if there is a decimal point, as this should be at the bottom-right corner.*

Where three or more numerals are combined in a single component, the pinouts may be designed for multiplexing rather than individual access to every segment of each numeral. A four-digit clock display, for instance, may have seven pins that connect in parallel to respective segments in all of the numerals, and four additional pins that can ground each numeral in turn, so that they can be selected sequentially.

Driver Chips and Multiplexing

Illuminating the appropriate segments in a single numeral can be done directly from a microcontroller, or through a driver chip such as the well-known and widely used 4543B that converts a binary-coded decimal input into appropriate segment output patterns. The chip can source sufficient current to drive each segment through a series resistor. Its pinouts are shown at Figure 24-12.

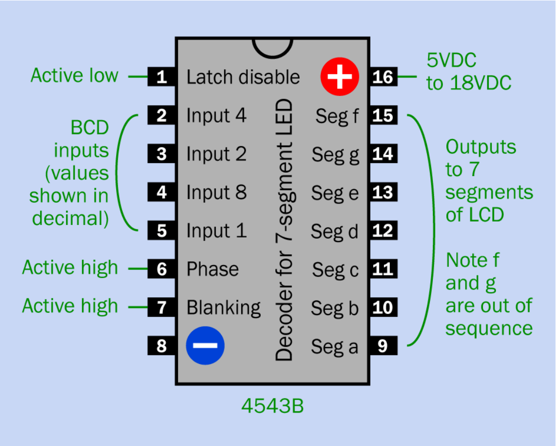
Active low
1
Latch disable
16
5VDC
to 18VDC
BCD
inputs
(values
shown in
decimal)
2
Input 4
3
Input 2
4
Input 8
5
Input 1
Active high
6
Phase
Active high
7
Blanking
8
Decoder for 7-segment LED
Seg f
15
Seg g
14
Seg e
13
Seg d
12
Seg c
11
Seg b
10
Seg a
9
Outputs
to 7
segments
of LCD
Note f
and g
are out of
sequence
4543B

Figure 24-12. *Pinouts of the 4543B seven-segment LED driver chip.*

When used in conjunction with a microcontroller, the 4543B can drive several seven-segment displays by multiplexing them. The basic schematic to achieve this is shown in Figure 24-13, omitting optional features such as leading-zero blanking or connections for a decimal point. The microcontroller sends the binary code for the first numeral and simultaneously grounds the common cathode of that numeral through a transistor, which is needed because as many as seven segments of the numeral may be passing current in parallel. The microcontroller then sends the binary code for the second numeral, and grounds it; then sends the binary code for the third numeral, and grounds it; and the cycle repeats. So long as this process is performed at sufficient speed (at least 50Hz), persistence of vision will create the illusion that all the numerals are active simultaneously. The circuit can be compared with a similar circuit to drive LCDs, shown in Figure 17-17.

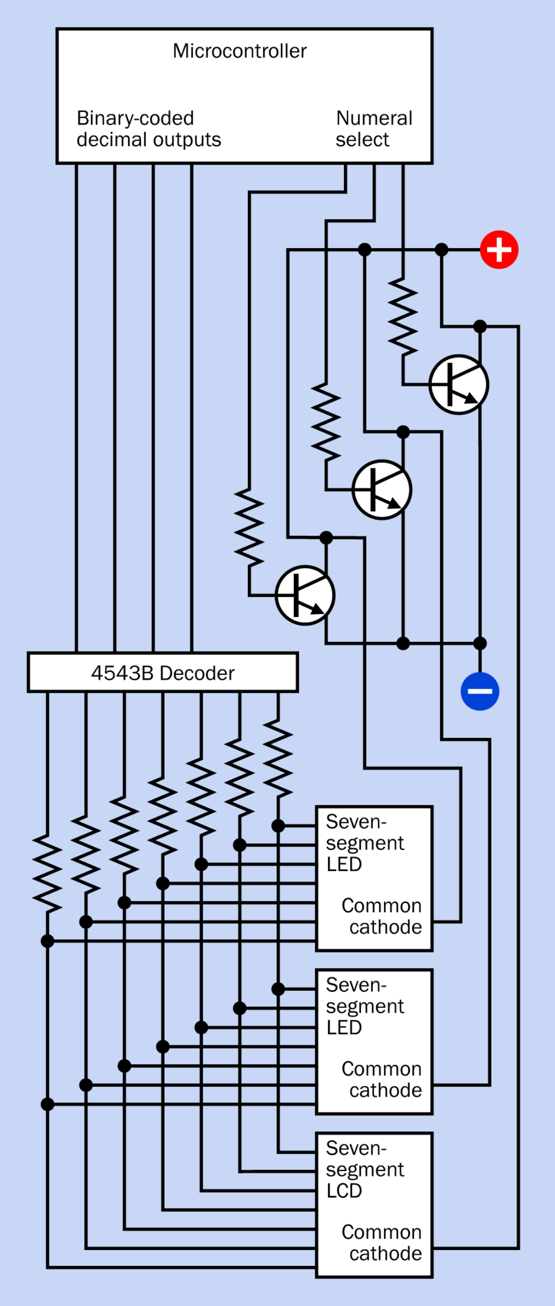

Figure 24-13. *A basic, simplified circuit for driving multiple seven-segment LED displays by multiplexing them.*

The disadvantage of this system is that the microcontroller must update the numerals constantly while performing other duties. To reduce this burden, a "smarter" driver such as the MC14489 can be used, controlling up to five 7-segment digits, or the Intersil ICM7218, which can control up to eight 7-segment digits.

The MC14489 controller receives data serially, using SPI protocol, and handles the details of addressing the LEDs. Because it contains latches to sustain the displayed data, a microcontroller only needs to communicate with the driver when the displayed information needs to be updated.

The ICM7218 is a more sophisticated chip, available in several variants, one of which can receive data on an 8-bit bus and run the seven-segment displays in hexadecimal mode.

Sixteen-Segment Driver Chip

The MAX6954 by Maxim can drive up to eight 16-segment alphanumeric LED displays using a scheme known as Charlieplexing, named after a Maxim employee named Charlie Allen who came up with the concept as a way of reducing the pin count required for multiplexing. Other Maxim controllers use this same protocol, which is transparent to the user.

A microcontroller sends data serially via I2C protocol to the MAX6954, which contains a variety of features. It can drive 14-segment and 7-segment displays as well as 16-segment displays, and contains a 104-character alphabet for each of them. Setting up a microcontroller to send the various necessary command codes to the MAX6954 is not a trivial matter, and bearing in mind the probably impending end-of-life of 16-segment displays, a better option may be to use dot-matrix LED display modules that have controller logic built in.

Dot-Matrix LED Display Modules

A dot-matrix LED display module requires data to define a character set, and a command interpreter to process instructions that will be embedded in a serial data stream. These capabilities are provided either by separate chips or (more often) are incorporated into the LED display module itself.

The SSD1306 is a monochrome graphical controller capable of I2C or SPI serial communication, or parallel communication. When this capability is built into a display module, only one of these types of communication may be activated.

The SSD1331 is a color graphical controller with similar communication capabilities.

The WS0010 is a monochrome controller, compatible with HD44780, which is designed to control LCDs.

Typical controller functions are summarized in "Alphanumeric Display Module" on page 168. Because there is no standardization in this field, precise details must be found in manufacturers' datasheets.

Pixel Arrays

The connections inside an 8x8 pixel array are shown in Figure 24-14, where the schematic symbols for LEDs have been replaced by gray circles for space reasons. To illuminate one LED, power is supplied to the intersection where it resides. In the figure, each vertical conductor (identified as A1, A2 . . . A8) can power the anodes of a column of eight LEDs, while each horizontal conductor (identified as C1, C2 . . . C8) can ground the cathodes of a row of eight LEDs. If only one vertical conductor is connected with positive power while one horizontal conductor is grounded, only one LED will light up, at the intersection of the active conductors.

A problem occurs if we wish to illuminate two LEDs. Suppose they are located at (A3,C2) and (A6,C5). Unfortunately, providing power to them will also result in activating LEDs at (A3,C5) and (A6,C2), as shown in Figure 24-15, where the yellow circles represent LEDs that have been switched on.

The answer to this problem is to rasterize the process. In other words, data is supplied on the array one line at a time, as in the process by which a TV picture is generated. If this is done quickly enough, persistence of vision will create the illusion that the LEDs are illuminated simultaneously.

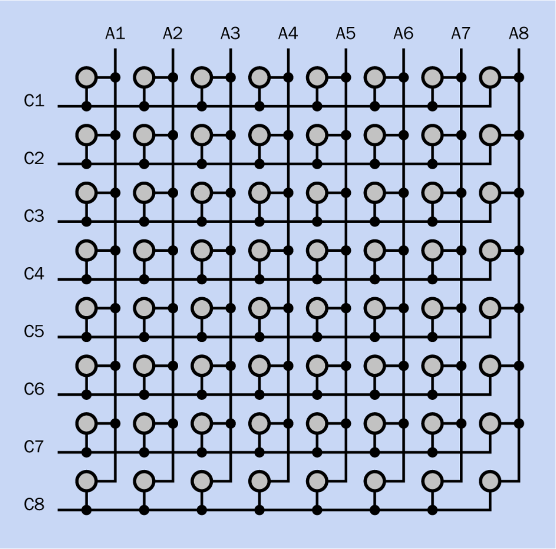

Figure 24-14. *Internal connections in the 8x8 matrix. Each gray circle represents an LED.*

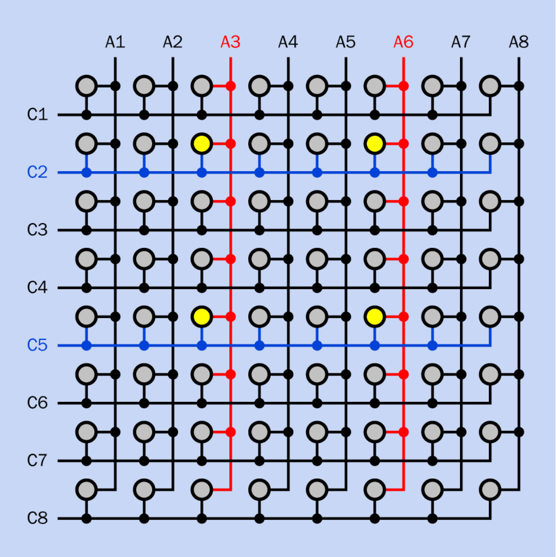

Figure 24-15. *An attempt to illuminate LEDs at (A3,C2) and (A6,C5) will also activate the LEDs at (A3,C5) and (A6,C2).*

A form of multiplexing is used to achieve this. One row of LEDs is connected to negative ground for a brief interval. During this interval, the anodes of selected LEDs are powered momentarily. Then the next row is grounded, and selected LEDs along that row are powered momentarily. This process is repeated for all eight rows before being repeated.

If several 8x8 matrices are assembled edge to edge, their horizontal conductors can be common to all of them. A horizontally scrollable display (sometimes referred to by the archaic term, *electric newspaper*) would then be possible, although the circuit design would be nontrivial.

Multiple Bar Display Driver

The LM3914 is a driver for a bar display that compares an analog input with a reference voltage and provides power to the segments of a multiple bar display, ranging from 2mA to 30mA, adjustable to match the specification of the display that is being used. The chip can generate either a “thermometer” effect, as more outputs are activated when the analog input increases, or a “moving dot” effect, in which only one output is on at a time.

One-Digit Hexadecimal Dot Matrix

While multi-character dot-matrix LED display modules are a versatile way to display prompts and numbers, a simpler component is sometimes sufficient. The Texas Instruments TIL311 is a minimal dot-matrix LED display that receives a binary value from 0000 through 1111 on its four input pins and generates the output in hexadecimal form, using numerals 0 through 9 and letters A through F. The sixteen possible outputs in the dot-matrix display are shown in Figure 24-16. Although this component is no longer being manufactured, it is widely available from many sources, especially in Asia. It eliminates the series resistors and controller chip that are customary for a seven-segment display, and has a better-looking output.

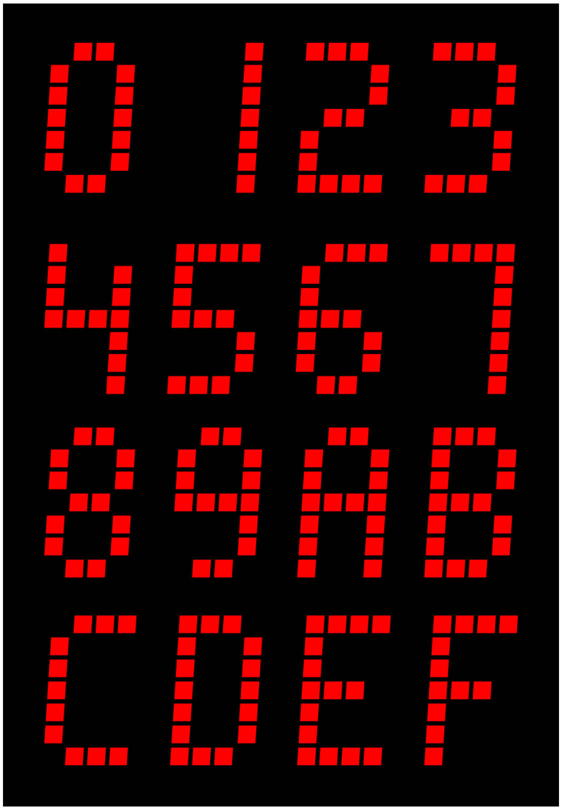

Figure 24-16. *Sixteen possible outputs that can be displayed by the Texas Instruments TIL311 to show a hexadecimal value in response to a four-digit binary input.*

A sample of the TIL311 is shown in Figure 24-17 displaying the number 2.

If two or more of these chips are put together, they can be multiplexed to display multi-digit decimal or hexadecimal integers.

The chip features two decimal points, one to the left of the displayed numeral, and one to the right. If they are activated, they require their own series resistors to limit the current.

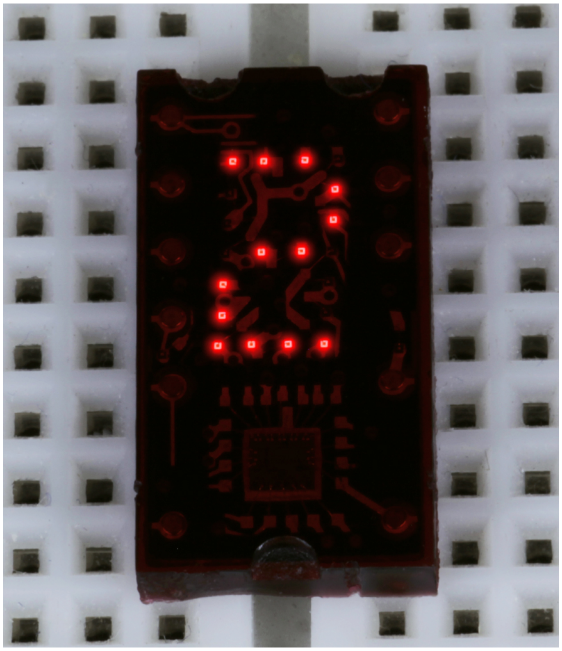

Figure 24-17. *The Texas Instruments TIL311 can be driven directly by a microcontroller or counter chip, with no series resistors necessary. It can generate a hexadecimal output.*

What Can Go Wrong

Common Anode versus Common Cathode

An LED display containing a common cathode is usually identical in appearance to a display containing a common anode, and the two versions will be distinguished by only one digit or letter in their part numbers. Because LED displays have a limited tolerance for reverse voltage, part numbers should be double-checked before applying power.

Incorrect Series Resistance

A common error is to assume that only one series resistor is necessary for a seven-segment LED display, either between the common cathode pin and ground, or, if there is a common anode, between that and the positive power supply. The problem is that if the resistor is suitable for a single LED, its value will be too high when several segments of the display are sinking current or drawing current through it. If its value is reduced, it will be too low when only two segments are using it (as when generating the number 1).

To provide equal illumination of all the segments, each must have its own series resistor.

Multiplexing Issues

When several displays are multiplexed, they naturally appear dimmer, creating a temptation to compensate by upping the current. Because current is only being applied to each display intermittently, a natural assumption is that a higher current can be safely used.

This may or may not be true. When running an LED device with pulsed current, the peak junction temperature, not the average junction temperature, determines the performance. At refresh rates below 1kHz, the peak junction temperature is higher than the average junction temperature, and the average current must therefore be reduced.

Datasheets must be checked to determine whether a device is designed with multiplexing in mind and, if so, what the recommended peak current is. Very often this value will be accompanied by a maximum duration in milliseconds, and a calculation may be necessary to determine the refresh rate, bearing in mind how many other LED displays are being multiplexed in the same circuit at the same time.

Irresponsible multiplexing will shorten the life of an LED display or burn it out.

25

vacuum-fluorescent display

The term **vacuum-fluorescent display** is seldom hyphenated, but the first two words are hyphenated here as they constitute an adjectival phrase. The acronym *VFD* is becoming increasingly popular, although it is ambiguous, being also used to identify a *variable frequency drive*. In both instances, the acronym is printed without periods between the letters.

The entry in this encyclopedia dealing with **fluorescent lights** does not include VFDs, because their purpose and design are very different. A VFD is an informational display, often showing numerals and letters, while a fluorescent light merely illuminates a room or work area. Although a VFD does use fluorescent phosphors, they are printed onto light-emitting segments of the display instead of being applied to the inside surfaces of a glass envelope.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LCD display** (see Chapter 17)
- **electroluminescence** (see Chapter 26)

What It Does

A **vacuum-fluorescent display** or *VFD* superficially resembles a backlit monochrome **LCD** or an **LED display**, as it can represent alphanumeric characters by using *segments* or a *dot matrix*, and can also display simple shapes. It is often brighter than the other information display systems, and can emit an intense green phosphorescent glow that some people find aesthetically pleasing, even though a grid of very fine wires is superimposed internally over the displayed image.

There is no specific schematic symbol to represent a vacuum-fluorescent display.

How It Works

The display is mounted inside a sealed capsule containing a high vacuum. A widely spaced series of very fine wires, primarily made of tungsten, functions as a *cathode*, moderately heated to encourage electron emission. The wires are often referred to as *filaments*.

A **fluorescent light** uses AC, and both of its electrodes are often confusingly referred to as cathodes. A VFD uses DC, and its cathode array has the function that one would expect, being connected with the negative side of the DC power supply.

Opposite the cathode, just a few millimeters away, is an anode that is subdivided into visible alphanumeric segments, symbols, or dots in a matrix. Each segment of the anode is coated with *phosphors*, and individual segments can be separately energized via a substrate. When electrons strike a positively charged anode segment, it emits visible light in a process of *fluorescence*. This behavior can be compared with that of a *cathode-ray tube*. However, the cathodes in a VFD are efficient electron emitters at a relatively low temperature, while the cathodes in a cathode-ray tube require substantial heaters.

Anode, Cathode, and Grid

A *grid* consisting of a mesh of very fine wires is mounted in the thin gap between the filaments of the cathode and the segments of the anode. A simplified view of this arrangement is shown in Figure 25-1.

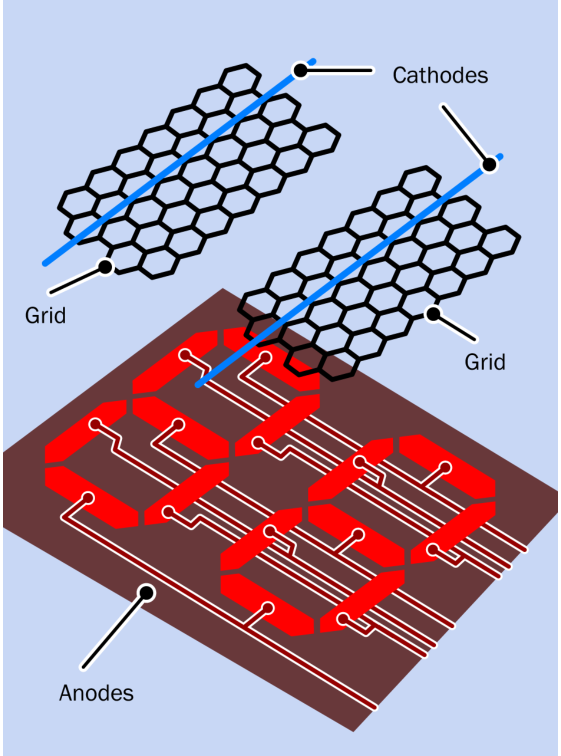

Figure 25-1. *The basic elements of a vacuum-fluorescent display.*

The polarity of the charge on the grid controls and diffuses electrons emitted by the cathode. If a grid section is negatively charged, it repels electrons and prevents them from reaching the sections of the anode beneath it. If the grid section is positively charged, it encourages electrons to reach the anode. Thus, the grid functions in the same way as the grid in a triode vacuum tube, but its conductors are so thin, they are barely visible.

How to Use It

Electronic calculators used vacuum-fluorescent displays during the 1970s, before LED displays and LCDs became more competitive. Purely numeric VFD modules are still available as strings of digits, although they are becoming uncommon and have been replaced by alphanumeric dot-matrix modules where each VFD character is mounted in its own glass module on a separate substrate.

Figure 25-2 shows the interior of a Commodore calculator from the 1970s, with its nine-digit vacuum-fluorescent display enclosed in one glass capsule.

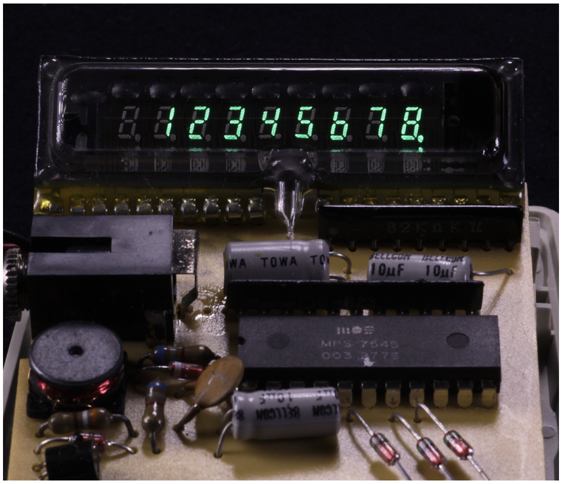

Figure 25-2. *The vacuum-fluorescent display from a 1970s Commodore calculator.*

A closeup of three digits from the previous figure appears in Figure 25-3, showing the grid superimposed above each numeral.

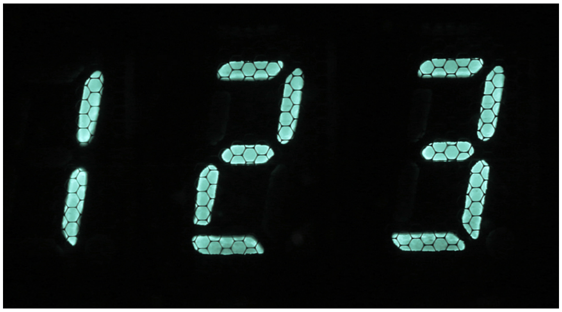

Figure 25-3. *Three digits from the previous figure, showing the grid that controls their illumination.*

A tinted filter of the same color as the display is usually placed in front of it, to conceal some of its workings. Thus, the Commodore calculator used a green filter in front of its green numerals. Figure 25-4 shows two seven-segment numerals from another device, with the filter removed. This reveals not only the grid but also the horizontal wires that function as the cathode. Connections between the segments of the numerals and a backplane are also visible.

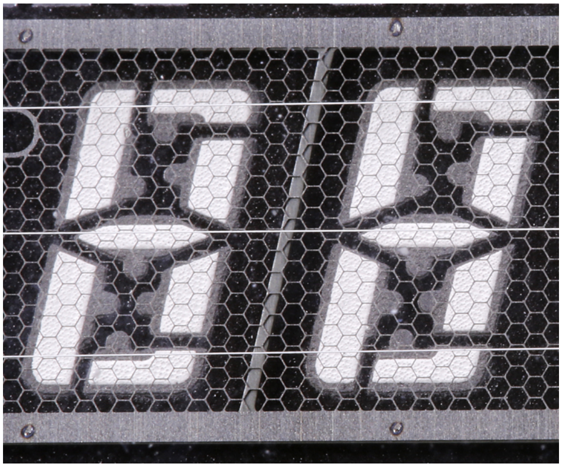

Figure 25-4. *Seven-segment numerals viewed without a colored filter, revealing the cathode (horizontal wires) and the grid (wire mesh).*

Modern Application

A modern VFD module is likely to be mated with a driver that converts 5VDC to the higher voltage (typically 50VDC to 60VDC) required for the display. Built-in logic may offer the option to receive data via an 8-bit parallel bus or with SPI serial protocol, and will contain a character set. A typical display resolution is 128 x 64 pixels.

The combination of a grid and a segmented anode enables a VFD to be controlled by *multiplexing*. For instance, in a display of four seven-segment numerals, the same equivalent segments in all four numerals can be connected in parallel while a separate grid covers each numeral. When each grid is positively energized, it selects the corresponding numeral, and the on-off segment patterns appropriate to that numeral are supplied. This procedure is repeated for each numeral in turn. Persistence of vision makes it appear that they are all active simultaneously.

Variants

Color

Although a VFD cannot provide a full-color display, selected anode segments can be coated with different phosphor colors, which can fluoresce simultaneously. Two or three individual colors are typically used, as in the display for a CD player where color helps to distinguish a variety of different functions. A closeup of a portion of the display from a CD player (with color filter removed) appears in Figure 25-5.

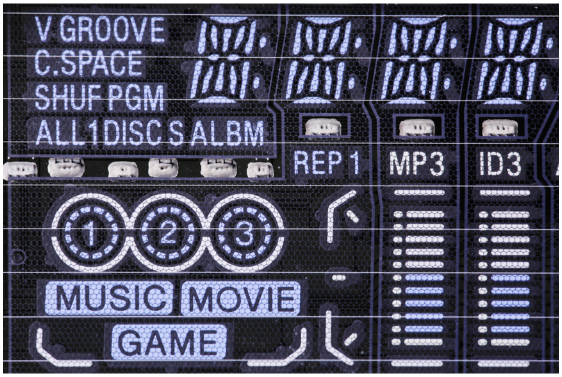

Figure 25-5. *The lefthand section of a vacuum-fluorescent display from a CD player.*

Character Sets and Pictorial Design

In the past, VFDs have combined seven-segment numerals in the same display as custom-shaped anodes. Solid-state gain meters in an audio amplifier, for instance, have used numerals beside a pictorial representation of gain levels resembling analog meters. The look and layout of a display of this type has been unique to a particular product.

Modern VFDs tend to use a generic dot-matrix display in which a character set in firmware dictates how patterns of dots are grouped to form numbers, letters, symbols, or icons.

The appearance of character sets generated with generic segments and dot-matrix arrays is thoroughly discussed and illustrated in the entry describing liquid-crystal displays in Chapter 17. VFD alphanumeric modules are identical in visual design to LCD modules, even though the internal electronics are different.

Comparisons

Two advantages of a VFD are that it functions well at low temperatures (unlike an LCD) and has sufficient brightness and contrast to be usable in sunlight (unlike many LED displays). It can be viewed from almost any angle.

Typical applications have included digital instrumentation in automobiles, information displays in audio and video consumer-electronics equipment, and numerical readouts in vending machines, medical devices, and some digital clocks.

Because a VFD requires a relatively high voltage, has significant power consumption, can show only a limited range of fixed colors, and is more expensive than LED displays or LCDs, its popularity has declined since the end of the 1990s.

What Can Go Wrong

Fading

VFDs gradually fade with age, as a result of reduced electron emission from the electrodes or diminishing performance of the phosphor coatings. Increasing the working voltage can prolong the life of a display.

electroluminescence

26

The field of **electroluminescent** devices is sometimes referred to as *EL*. The same acronym can also be applied as an adjective to an individual electroluminescent device, as in, for example, "an EL panel."

An *organic light-emitting diode*, more commonly known by its acronym *OLED*, is included in this entry because it is technically an electroluminescent device and its design concept is similar to that of an electroluminescent panel. Generic LEDs are also technically electroluminescent, but are not commonly described as such, and have their own entries in this encyclopedia under the subject categories **LED indicator**, **LED area lighting**, and **LED display**.

OTHER RELATED COMPONENTS

- **LED indicator** (see Chapter 22)
- **LCD display** (see Chapter 17)
- **fluorescent light** (see Chapter 20)
- **vacuum-fluorescent** display (see Chapter 25)

What It Does

An electroluminescent device configured as a *panel*, *ribbon*, or *rope-light* contains phosphors that emit light in response to a flow of electricity.

Panels can be used as backlights for LCD displays or, more often, as always-on low-power devices such as exit signs and night lights. Ribbons and rope lights (the latter being also known, more accurately, as *light wires*) are used mainly as recreational novelties. They can be battery powered through a suitable voltage converter. A battery-powered rope light can be wearable.

Thin-film OLED electroluminescent panels are used in small video screens in handheld devices. At the time of writing, OLED TV screens measuring 50" or more have been demonstrated, but are not yet economic for mass production.

No specific schematic symbol exists to represent any electroluminescent device or component.

How It Works

Luminescence is the emission of light as a result of a process that does not require heat. (The opposite phenomenon is *incandescence*, in which heating causes an object to emit light; see Chapter 18 for a description of **incandescent lamps**.)

Electroluminescence is luminescence resulting from stimulation by electricity. This very broad definition really includes devices such as **LED**s, although they are hardly ever described in those terms. Electroluminescence generally refers to panels, films, or wires where electrodes are in direct contact with light emitters such as phosphors.

The exception is an *organic LED*, usually known by its acronym *OLED*, which is frequently described as an electroluminescent device, perhaps because its configuration as a sandwich of thin, flat layers resembles an electroluminescent panel. Two of the layers are semiconductors, and they interact as light-emitting diodes.

Phosphors

A *phosphor* is a compound such as zinc sulfide that will emit light when it receives an energy input from another light source or from electricity. Typically the compound must be mixed with an *activator* such as copper or silver.

For many decades, TV sets and video monitors were built around *cathode-ray tubes* in which the interior of the screen, at the front of the tube, was coated with phosphors. A beam of electrons that fluctuated in intensity generated a picture on the screen by drawing it as a series of lines.

Derivation

The term *phosphor* is derived from *phosphorescence*, which in turn comes from the name of the element *phosphorous*, which will glow when it oxidizes in moist air. (These terms were established before other forms of luminescence were discovered and understood. The behavior of phosphorous is really an example of *chemiluminscence*.)

For our purposes, a phosphor is a compound that is capable of fluorescence or electroluminescence.

Variants

Panels

Electroluminescent panels using phosphor powder, sometimes referred to as *thick phosphor*, are a popular choice where a constant, uniform, low light output is acceptable.

An electric potential is established between two films that act as electrodes, separated by a layer of phosphor crystals. Some manufacturers refer to this configuration as a *light-emitting capacitor* because the structure resembles a capacitor, even though that is not its purpose. The front film is transparent, allowing light to escape.

An electroluminescent panel can be powered by AC or DC but requires at least 75V. Its power consumption is self-limiting, so that no control electronics are required other than a voltage converter if battery power is used.

The phosphors generate a constant, evenly distributed luminescence over the entire area, although the output is not very intense. Applications include night-lights, exit signs, and backlighting for wristwatches.

Panelescent electroluminescent lighting by Sylvania was used for instrument panel displays in some car models such as the Chrysler Saratoga (1960 through 1963) and Dodge Charger (1966 through 1967). It is still used for night-lights. *Indiglo* electroluminiscent displays are still widely used in wristwatches.

The interior components of a disassembled electroluminescent night-light are shown in Figure 26-1. The panel emits a natural pale green glow. A separate blue or green filter passes the glow while blocking other colors of incident light that would otherwise reflect off the panel.

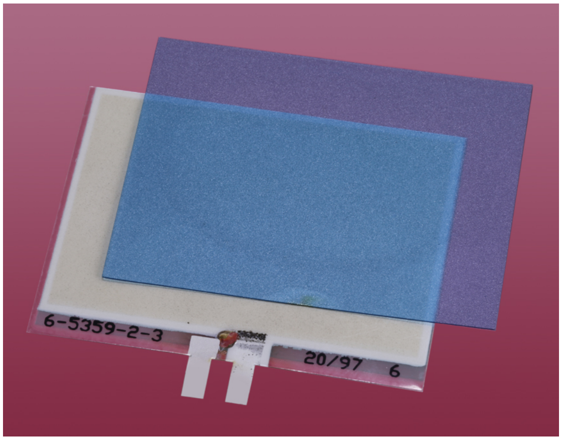

Figure 26-1. *The two interior components of an electroluminiscent night-light: the luminescent panel, and a separate translucent filter.*

Electroluminescent night-lights were popular in the 1970s and 1980s, often featuring cartoon characters to appeal to children. Figures 26-2 and 26-3 show the same night-light in its daytime off-state and its night-time on-state, respectively.

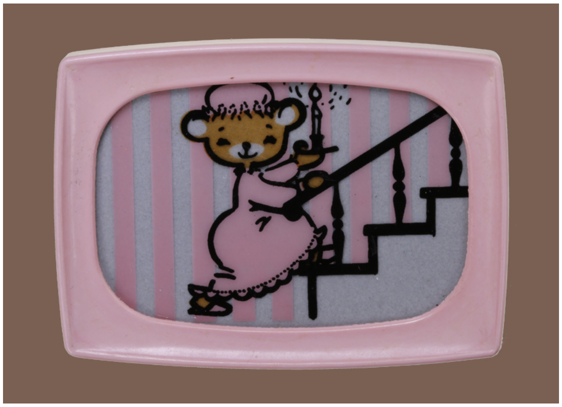

Figure 26-2. *A vintage Panelescent brand night-light, several decades old, in its off-state.*

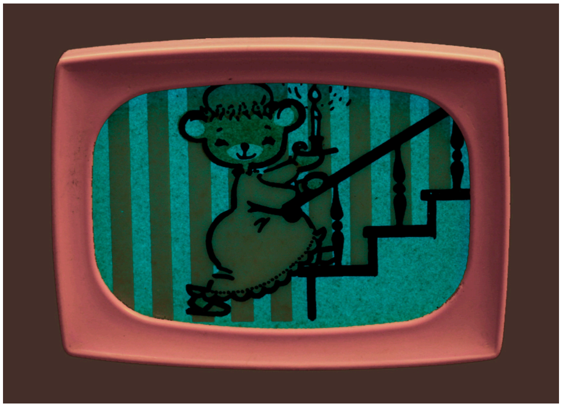

Figure 26-3. *The same night-light, with its green radiance visible under conditions of low ambient light.*

Advantages of electroluminescent panels include the following:

- Low current consumption. One US manufacturer claims that a single exit sign will use electricity costing less than 20 cents per year, while the annual cost of a night-light will be less than 3 cents per year.
- Long life, up to 50,000 hours.
- Self-regulating; no control circuitry required.
- Omnidirectional light output.
- Very wide operational temperature range, between approximately -60 and +90 degrees Celsius.
- Can be plugged directly into a wall outlet.

Disadvantages include:

- Limited light output.
- Very limited choice of colors.
- Not very efficient, 2 to 6 lumens per watt (although the low light output naturally entails low power consumption).
- Gradual reduction in phosphor performance over time.
- High voltage required: 60V to 600V. Ideal for plugging into a wall outlet, but requires a converter when used with battery-powered devices.

Flexible Ribbons

The light-emitting layers inside a night-light are somewhat flexible, and can be made more flexible by reducing their thickness. The result is an electroluminescent ribbon that has some novelty value, and may be used for customizing automobiles. Figure 26-4 shows a ribbon about 1.5″ wide and 12″ long, designed for 12VDC power applied through an inverter.

Rope Light

A rope light or wire light may resemble a *glowstick*. However, a glowstick generates light from chemiluminescence (chemical reactions that release photons), while a rope light uses electricity.

Figure 26-5 shows a rope light powered by two AA batteries connected through an inverter.

At the center of the rope light is a conductor that serves as one electrode. It is coated in phosphors, and the layer of phosphors is protected by a transparent sheath. One or more thin wires is wrapped around the sheath in a spiral, with large gaps between one turn and the next. These wires serve as the second electrode. The wires are enclosed in transparent insulation that forms an outer sheath.

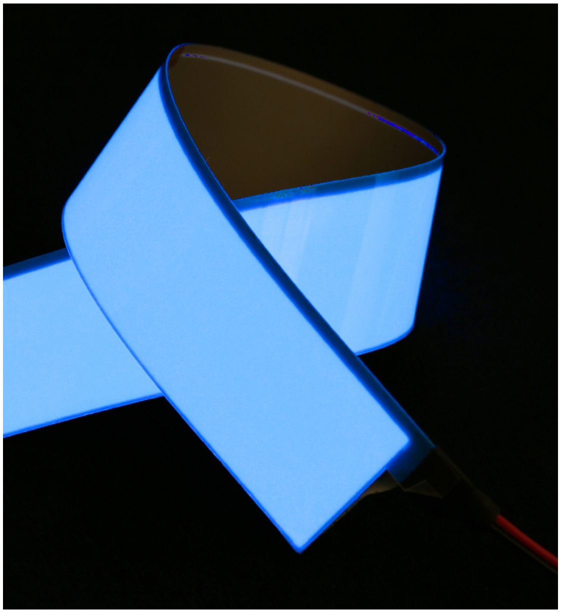

Figure 26-4. *A 12" length of electroluminescent ribbon.*

When AC is applied between the electrodes, the layer of phosphors emits light that radiates out in the gaps between the thin wires. The color of the light can be modified by using tinted outer insulation.

OLED

An OLED uses two thin, flat electrodes, somewhat like a thick-phosphor electroluminescent panel, except that it contains more layers and is capable of generating more light. The layers in an OLED are "organic" in that they consist of chemically organic molecules containing carbon and hydrogen atoms and generally do not contain heavy metals.

While an LCD video monitor or TV screen must have a separate backlight, an OLED generates its own light. This reduces the thickness of the display to a few millimeters and makes it potentially more efficient.

The semiconductor layers are subdivided into pixels, each functioning as a light-emitting diode, while additional layers carry a matrix of conductors for pixel addressing. In an AMOLED, the conductors form an active matrix, while in a PMOLED, they form a passive matrix.

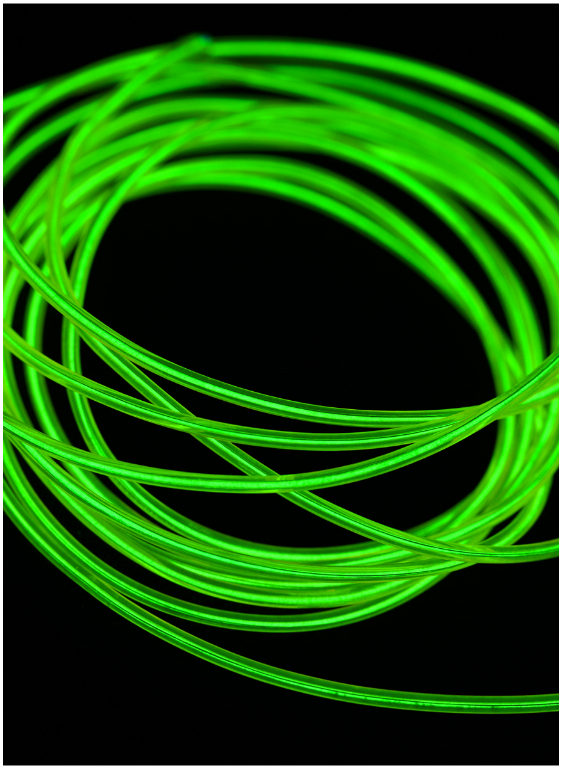

Figure 26-5. *A length of glowing rope light, also known as a light wire.*

In an active matrix, each pixel is backed with a *thin-film transistor* to store its state while the energizing voltage transitions. This is often described as a *TFT* display; but the term is interchangable with "active matrix."

In a passive matrix, each pair of conductors simply supplies current to a pixel. This is cheaper and easier to fabricate but is less responsive.

The terms "active matrix" and "passive matrix" have the same meaning as when used to describe a **liquid-crystal display**.

Monochrome OLED display modules with dot-matrix characters are currently available from China for just a few dollars. Although they appear superficially similar to LCD modules, they generate pure white-on-black characters.

Small full-color OLED screens are used in smartphones and on camera backs, but at the time of writing, large OLED screens are not a mature technology, partly because of production costs. A great variety of chemicals and layer configurations have been tried, and the application of pixels to a substrate has been attempted with vacuum deposition through a shadow mask and with a system similar to inkjet printing. Pixels that emit red, green, and blue light have also been used. Filtered pixels have been used. One dominant process has not yet emerged.

Longevity and brightness have been problems. Where red, green, and blue diodes have been used, the different colors deteriorate at different rates. While the human eye tolerates an overall reduction in brightness, it does not tolerate a slight color shift caused by blue pixels, for instance, losing brightness more rapidly than red pixels.

Because OLED screens promise to be thinner, lighter, and brighter, and may eliminate the need for a fragile glass substrate, there is a strong incentive to develop this technology, which seems likely to gain dominance in the future.

OLED panels may also become a source of diffuse, shadowless room lighting or office lighting when practical problems have been solved and costs have fallen significantly.

transducer

27

The term **transducer** is used here to describe a *noise-creating device* that is driven by external electronics. By comparison, an **audio indicator** (discussed in the next entry) contains its own internal electronics and only requires a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

A **speaker**, more properly termed a *loudspeaker*, is an electromagnetic transducer but is seldom described in those terms. It has a separate entry in this encyclopedia and is defined here as a *sound reproduction device* that is larger and more powerful than a typical transducer and has a more linear frequency response.

While piezoelectric transducers formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

Some transducers convert sound into electricity, but these are categorized as *sensors*, and will be discussed in Volume 3. The only transducers discussed in this entry are those that convert electricity into sound.

OTHER RELATED COMPONENTS

- **audio indicator** (see Chapter 28)
- **headphone** (see Chapter 29)
- **speaker** (see Chapter 30)

What It Does

An audio **transducer** is a device that can create an *alert*. It requires an AC signal that is supplied by external electronics, and in its simplest form may be referred to as a *buzzer* or a *beeper*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often used in conjunction with touch pads, to provide audio confirmation that a tactile switch has been pressed.

The schematic symbols in Figure 27-1 can be used to represent any kind of audio alert, including **indicators**, which contain their own electronics to generate a simple tone or series of tones. Type A is probably the most popular symbol. Types B and C often appear with the word "buzzer" printed beside them for clarification. D and E are really symbols for a **speaker**, but are often used for an alert. F is the symbol for a *crystal*, now sometimes used to indicate a piezoelectric noise maker. G specifically represents an electromagnetic transducer, but is seldom used.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring from around 0.5″ to 1.5″ in diameter. The enclosure is sealed at the bottom but has an opening at the top, so that sound can emerge from the upper side of the diaphragm without being partially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

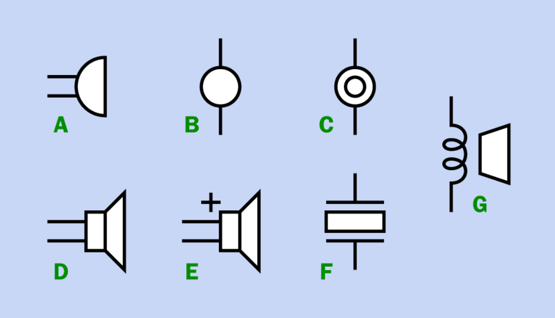

Figure 27-1. *An assortment of symbols which can represent a transducer or an indicator. See text for details.*

The diaphragm is activated either electromagnetically or piezoelectrically, as described next.

Externally, a transducer may be indistinguishable from an audio indicator such as the one pictured at Figure 28-1.

Variants

Electromagnetic

An electromagnetic transducer contains a diaphragm that is usually made of plastic. Mounted on it is a smaller *ferromagnetic disc* that responds to the fluctuating field from AC passing through a coil. When the diaphragm vibrates, it creates pressure waves that are perceived by the human ear as sound.

A car horn is a particularly loud form of electromagnetic transducer.

Piezoelectric

A piezoelectric transducer contains a diaphragm consisting of a thin brass disc on which is mounted a *ceramic wafer*. When an AC signal is applied between the piezoelectric wafer and the disc, the disc flexes at that frequency.

The term *piezo* is derived from the Greek *piezein*, which means "to squeeze or press."

Ultrasonic Transducer

The diaphragm in an *ultrasonic transducer* vibrates at a frequency above the range of human hearing. This component may be electromagnetic, piezoelectric, or crystal-based. Often it is used in conjunction with an ultrasonic receiver as a distance measuring device. The two components can be sold pre-mounted on a breakout board. An output from the board can consist of a pulse train where the pulse duration is proportional to the distance between the transducer and the nearest sound-reflecting object.

An ultrasonic transducer is pictured in Figure 27-2. Its internal components are shown in Figure 27-3.

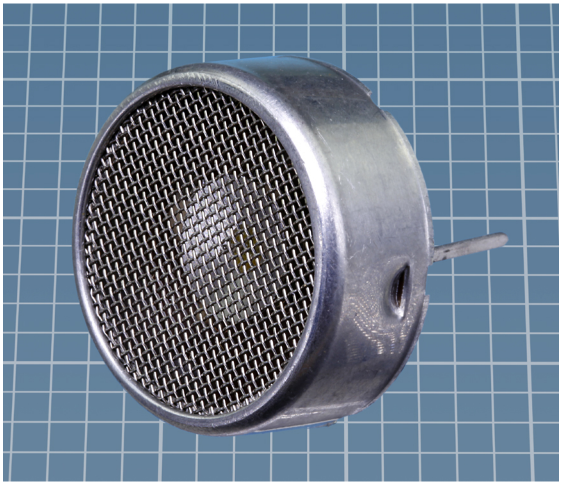

Figure 27-2. *The exterior of an ultrasonic transducer.*

Submersible ultrasonic transducers may be used in cleaning systems, where they agitate a liquid that dislodges dirt or debris. Ultrasonic transducers are also used in echo-sounding and sonar equipment with marine applications.

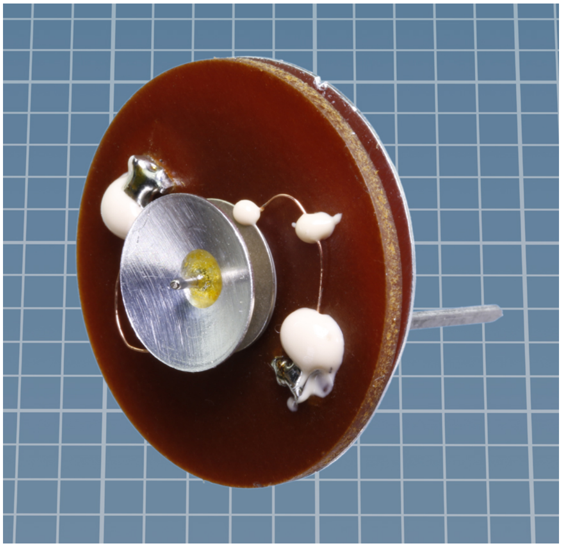

Figure 27-3. *Inside an ultrasonic transducer, a small aluminum cone is the sound radiating element. The white blobs are adhesive to secure the thin wires.*

Formats

Some transducers are available in surface-mount format, measuring about 0.5″ square or less. Because the resonant frequency is related to the size of the component, surface-mount transducers usually generate a high-pitched beep.

Values

Frequency Range

Audio frequency is measured in Hertz, abbreviated Hz, named after Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. The H in Hz is capitalized because it refers to a real name. One thousand Hertz can be written as 1 kiloHertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

The most common frequencies applied to audio transducers range between 3kHz and 3.5kHz. Piezoelectric elements are inefficient for generating sounds below 1kHz, but electromagnetic transducers are better able to generate lower frequencies. Their response curve can be approximately flat to frequencies as low as 100Hz.

Sound Pressure

Sound pressure can be measured in Newtons per square meter, often abbreviated as Pa. Newtons are units of force, while Pa is an abbreviation of Pascals.

The *sound pressure level* (SPL) of a sound is not the same as its sound pressure. SPL is a logarithmic value, to base 10, in units of decibels (dB), derived from the pressure of a sound wave relative to an arbitrary reference value, which is 20 micro-Pascals (20μPa). This is the agreed minimum threshold of human hearing, comparable to a mosquito at a distance of three meters. It is assigned the level of 0dB.

Because the decibel scale is logarithmic, a linear increase in the decibel level of a sound does not correspond with a linear increase in actual sound pressure:

- For each additional 6dB in the SPL, the actual sound pressure approximately doubles.
- For each additional 20dB in the SPL, the actual sound pressure is multiplied by 10.

Bearing in mind that 0dB corresponds with the reference sound pressure of 20μPa, an SPL of 20dB represents a sound pressure of 200μPa (that is 0.0002Pa), and so on.

Many tables show an estimated decibel level for various noise sources. Unfortunately, these tables may contradict each other, or may fail to mention the distance at which a sound is measured. Figure 27-4 shows estimates derived by averaging eight similar tables. It should be viewed as an approximate guide.

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 27-4. *Approximate decibel values for some sound sources (averaged from a selection of eight similar charts).*

Sometimes the claim is made that an increase of +10 on the decibel scale will correspond with a subjective experience that the noise is "twice as loud." Unfortunately, this statement cannot be quantified.

Weighted Sound Values

Subjective assessment of sound is complicated by the nonlinear frequency response of the human ear, which causes some frequencies to seem "louder" than others, even though their sound pressure is the same. The frequency weighting of the ear can be determined by playing a reference tone of 1kHz at 20dB and then doing an A-B comparison with a secondary tone at another frequency, asking the subject to adjust the gain of the secondary tone up or down until the two tones seem equally loud.

This procedure is performed for a range of frequencies. The test is then repeated with a louder 1kHz reference tone, at 30dB. Repetitions continue to a final reference tone of 90dB.

The resulting curves are known as *equal-loudness contours*. An averaged set, from multiple sources, has become an international standard with ISO number 226:2003. The curves shown in Figure 27-5 are derived from that standard. The curves show that the sound pressure of lower frequencies must be boosted by a significant amount to sound as loud as a 1kHz frequency, while a frequency around 3kHz must be reduced slightly, because it tends to sound louder than all others.

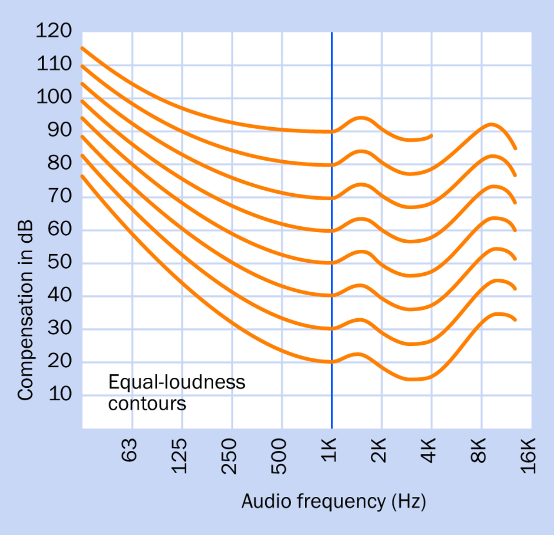

Figure 27-5. *Equal loudness contours derived from ISO 226:2003. See text for details.*

Although the accuracy of equal-loudness contours is controversial, they have been the basis of a widely used weighting system to adjust dB values to represent subjective perceptions of loudness. This *A-weighting* system remains the best-known and most widely applied audio standard in the United States, even though it has been criticized for assigning too little value to sounds that are brief in duration. If a sound level is expressed in dBA, it is A-weighted, meaning that the sounds to which the ear is least sensitive are assigned a value that is lower than their measured value. Thus, a tone of 100Hz has a dBA value about 20dB lower than its dB value, because the human ear is relatively insensitive to low-pitched sounds. dBA values are used in regulations that limit noise in the work place and other environments.

Unweighted Values

If sound intensity is expressed in dBSPL, it is a measurement of the actual Sound Pressure Level and has not been adjusted with the A-weighting system. A graph of unadjusted dbSPL values will display low frequencies as if they are more intense than the ear will perceive. In practical terms, subjective perception of low-end rolloff will be even more severe than the graph makes it appear.

If sound intensity is expressed merely in dB, probably it is unweighted and should be considered as dBSPL.

From a practical point of view, when choosing a tone for a transducer, a 500Hz tone may sound relatively mellow and not subjectively loud. A 3.5kHz tone can be a good attention-getting signal, as the ear is most sensitive in that range.

Transducers generally have a sound pressure rating in dBSPL ranging between 65dBSPL to 95dBSPL, with just a few products that can make more or less noise.

Measurement Location

The sound pressure from an audio alert will naturally diminish if the measurement point moves farther away. Therefore, any rating in decibels should be expressed with reference to the distance at which the measurement is made.

Measurement locations may be expressed in centimeters or inches, and may vary from 10cm to 1 meter, even in datasheets for different devices from the same manufacturer. If the measurement distance doubles, the SPL diminishes by approximately 6dB.

Limitations

A piezoelectric transducer is not intended as a sound reproducer, and does not have a smooth or flat frequency response. The curve for the Mallory PT-2040PQ is not unusual, reproduced in Figure 27-6. This component measures about 3/4" in diameter, is rated for 5VDC, and uses only 1.5mA to generate 90dB (measured at a distance of 10cm). Like many piezoelectric audio devices, its response peaks around 3500kHz and diminishes above and below that value, especially toward the low end. While it is perfectly adequate as a "beeper," it will not reproduce music successfully.

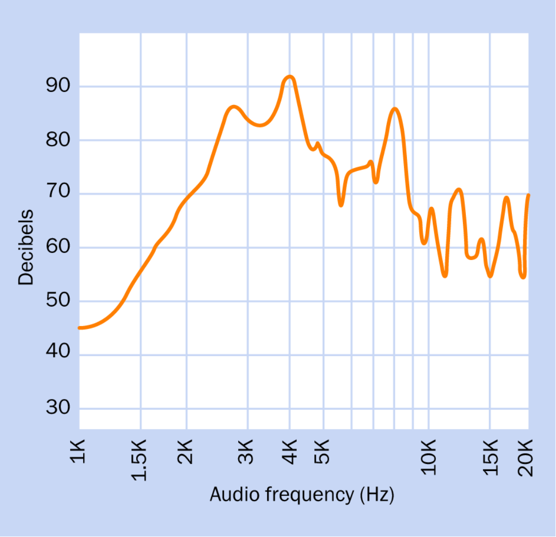

Figure 27-6. *The frequency response from a typical small piezoelectric transducer.*

An electromagnetic transducer is better able to generate low frequencies than a piezoelectric transducer. It has a low impedance that may be suitable in some circuits. However, it is slightly heavier than a comparable piezoelectric transducer, uses much more power, and as an AC device containing a coil, it can create electromagetic interference or may cause fluctuations in the circuit as an inductive load. It is also vulnerable to magnetic interference from elsewhere, while a piezoelectric transducer is not.

While an electromagnetic transducer can be used to reproduce speech or music, and will do a better job than a piezoelectric transducer, its performance will still be dissatisfying. A miniature speaker is more appropriate for the task.

Voltage

Transducers are typically designed to work with voltages ranging from 5VAC to 24VAC. The ceramic wafer in a piezoelectric transducer usually cannot withstand voltages much above 40VAC, and its sound output will not increase significantly above 30VAC.

Current

Typical piezoelectric transducers use less than 10mA and generate negligible heat. An electromagnetic transducer may draw as much as 60mA.

How to Use It

Appropriate Sound Intensity

An alert should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than ambient background noise.

Volume Control

Sound pressure can be lowered by reducing the voltage. Because a transducer does not consume much current, a *trimmer* can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

AC Supply

Although a transducer is an AC device, it is unlikely to be designed for voltage that fluctuates positively and negatively either side of a neutral value. Typically it is intended for voltage that fluctuates between 0V (ground) and the rated positive value of the power supply, and its pins, wires, or terminals are usually marked accordingly. If it has wire leads, the red lead should be connected to the more-positive side of the supply. If it has pins, the longer pin should be more positive.

The alternating signal for a transducer can be supplied by any simple oscillator or astable multivibrator circuit. For a given peak voltage, a square wave will generate a louder signal than a sinusoidal wave. A simple 555 timer circuit can be used, with a second monostable timer to limit the duration of the beep if necessary. An astable 555 can be used to test the transducer and select the audio frequency that sounds best.

Self-Drive Transducer Circuit

If a transducer has three wires or pins, it is probably a *self-drive* type. The datasheet may identify its inputs as M, G, and F, meaning Main, Ground, and Feedback. The Feedback terminal is connected with a section of the diaphragm which vibrates 180 degrees out of phase with the Main terminal. This facilitates a very simple external drive circuit, such as that in Figure 27-7, where the frequency is determined by the transducer's resonant frequency.

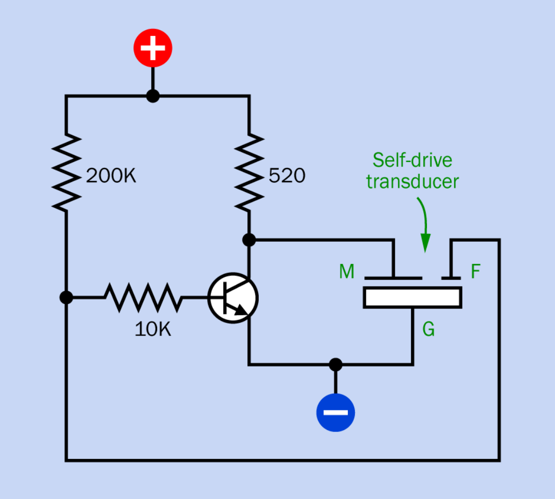

Figure 27-7. *A circuit to control a self-drive type of piezoelectric transducer.*

What Can Go Wrong

Overvoltage

Mallory Sonalerts, one of the largest producer of piezoelectric alerts, states that in the "vast majority" of returned products, the failure was caused by excessive voltage, often in the form of a transient voltage spike.

Leakage

If an alert makes a small, low-volume sound when it is supposed to be off, some current is leaking through it. Less than 1mA may be enough to cause this problem. According to one manufacturer, it can be fixed by placing a 30V Zener transient voltage suppressor diode in series with the alert, or by wiring a small **incandescent lamp** in parallel with the alert.

Note that when the alert is activated, the full supply voltage will be seen at the lamp.

Component Mounting Problems

Some alerts are packaged with mounting holes, but many are not. Those with pins can be soldered into a board, but those without must be glued in place or inserted into a cavity from which they cannot shake loose. Silicone adhesive is recommended, but care must be taken to avoid any of it dripping into the alert before it sets.

Moisture

If an alert will be used in a location where it is vulnerable to moisture, it should be of a type that is sealed against the environment. Even a sealed unit should ideally be oriented so that it faces slightly downward.

Transducer-Indicator Confusion

Externally, a transducer and an indicator often look identical, and some of them are not marked with a manufacturer's part number. Damage can be caused by applying DC to a transducer or AC to an indicator. If both types of parts are kept in inventory, they should be carefully labeled.

Connection with a Microcontroller

A piezoelectric transducer can be driven by a microcontroller, but an electromagnetic transducer is not appropriate in that role, because of its relatively higher current consumption and its behavior as an inductive load.

audio indicator

28

An **audio indicator** is defined here as a noise-creation device that generates a simple tone or series of tones. Unlike a **transducer**, which requires an external source of AC to determine its audio frequency, an indicator contains its own electronics and requires only a DC power supply. Either of these components is often described as a *beeper* or *buzzer*.

While piezoelectric alerts formerly used crystals, only the more modern piezoelectric type that uses a ceramic wafer will be considered here.

OTHER RELATED COMPONENTS

- **transducer** (see Chapter 27)
- **headphone** (see Chapter 29)
- **speaker** (see Chapter 30)

What It Does

When DC power is applied to an **audio indicator**, in its simplest form it creates a continuous or intermittent tone of a fixed frequency. This is called an *alert*.

Audio alerts are used in microwave ovens, washer/dryers, automobiles, gasoline pumps, security devices, toys, phones, and many other consumer products. They are often applied with touch pads, to provide audio confirmation that a tactile switch has been pressed.

A few indicators are programmed to create a two-tone sound, or multiple-tone sequences.

See Figure 27-1 in the previous entry for an assortment of schematic symbols that may be used to represent either an alert or a transducer.

How It Works

A circular diaphragm is glued at its edges inside a cylindrical plastic enclosure, usually measuring from around 0.5″ to 1.5″ in diameter. The enclosure is sealed at the bottom but has a small hole at the top, so that sound can emerge from the upper side of the diaphragm without being partially cancelled by sound of opposite phase that is emitted from the underside of the diaphragm. The enclosure also contains electronics to generate one or more audio tones, and amplifies the sound by resonating with it, in the same way that the body of a guitar or violin amplifies a note being played on the strings.

A PUI XL453 piezoelectric audio indicator is pictured in Figure 28-1, fully assembled on the right, and with its circuit board and diaphragm removed on the left. This indicator creates a pulsed tone at 3.5kHz with a sound pressure of 96dB. It draws 6mA at 12VDC and measures approximately 1″ in diameter.

For more information about the measurement of sound frequency and pressure, see "Frequency Range" on page 251 and "Sound Pressure" on page 251 in the previous entry.

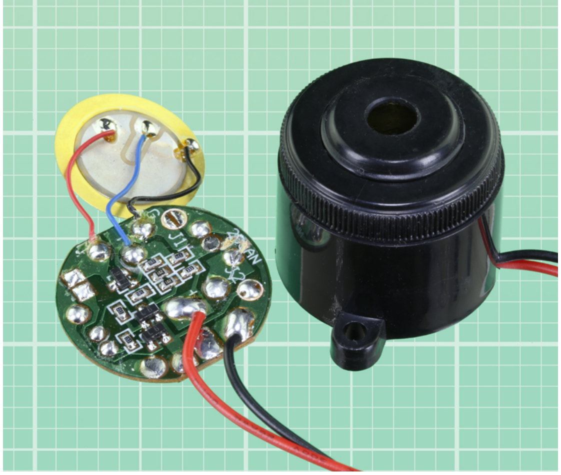

Figure 28-1. *A typical piezoelectric audio indicator.*

Externally, an audio indicator may be indistinguishable from a transducer. However, internally, an indicator is almost always a piezoelectric device, in which a *ceramic wafer* is mounted on a thin brass diaphragm. The term *piezo* is derived from the Greek *piezein*, which means to squeeze or press.

A transducer (described in the previous entry) is a piezoelectric or electromagnetic alert that does not usually contain its own circuitry and must be driven by an external source of AC, which establishes the audio frequency.

The distinction between an indicator and a transducer is often unclear in parts catalogs, where all alerts may be identified as *buzzers*, even though they mostly beep rather than buzz.

Audio Frequency

For a discussion of audio frequency, see "Frequency Range" on page 251 in the previous entry.

History

Probably the earliest form of electrically activated alert was the door bell, in which a 6VDC battery-powered solenoid pulled a spring-loaded lever terminating in a small hammer. The hammer struck the bell, but the motion of the lever also opened a pair of contacts, cutting off power to the solenoid. The lever sprang back to its rest position, which closed the contacts and repeated the cycle so long as power was supplied through an external pushbutton.

Subsequent systems used a small loudspeaker powered by AC house current through a step-down transformer. This created a buzzing sound and may have been the origin of the term "buzzer."

Small components that made a beeping sound only became common when digital equipment required a simple, cheap way to confirm user input or attract attention to the status of a device.

Variants

Sound Patterns

Because an audio indicator contains its own electronics, the manufacturer has the freedom to create various patterns of sound output.

The default is a steady tone. Other common variants include an intermittent tone and a dual tone that fluctuates rapidly between two frequencies. This is sometimes referred to as a *siren*. A few variants can generate an output pattern consisting of several tones in sequence, or effects such as warbling or whooping sounds, which are used mainly in alarm systems.

Formats

Some audio indicators are available in surface-mount format, measuring 1/2" square or less. Because the resonant frequency is related to the size of the component, surface-mount alerts usually make a high-pitched beep.

Panel-mount and board-mount formats range from about 1/2" to 1.5" in diameter. A small audio alert designed to be mounted on a circuit board is shown in Figure 28-2, with its top removed on the right to expose the brass diaphragm glued around the edges. The same component is shown with its plastic enclosure removed completely in Figure 28-3.

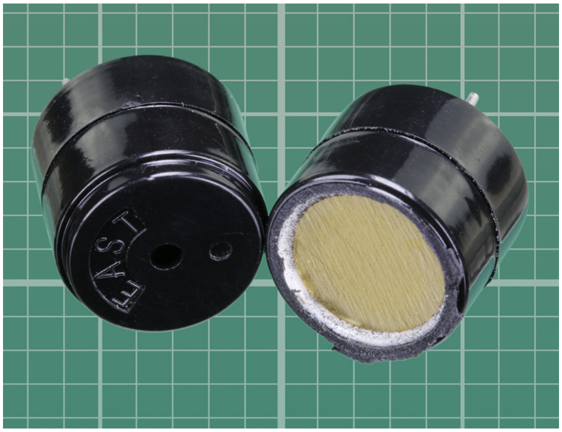

Figure 28-2. *An audio indicator approximately 0.5" in diameter, partially disassembled on the right, revealing its brass diaphragm.*

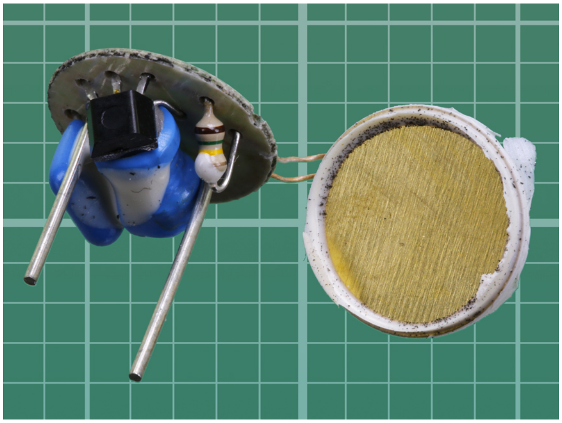

Figure 28-3. *The same indicator from the previous photograph, with its enclosure completely removed.*

Values

For an explanation and discussion of sound pressure and its measurement in decibels, see "Sound Pressure" on page 251 in the previous entry.

Audio indicators generally have a sound pressure rating in dBSPL ranging between 65dBSPL to 95dBSPL, with just a few products that make more or less noise. At 120dB and above, most products are packaged as alarm sirens ready for installation, often with a small horn attached. Their power consumption can be 200mA or more, and they are many times the price of a simple indicator designed for circuit-board mounting.

Voltage

An audio indicator containing its own electronics will almost always be rated somewhere in the range from 5VDC to 24VDC. Sirens intended for use with burglar alarms are often designed for 12VDC or 24VDC, as these are popular values for security systems with battery backup. However, in addition to a rated voltage, a datasheet may specify a wide range of acceptable operating voltages. For example, an indicator with a *rated* voltage of 12VDC may have an *operating* voltage of 3VDC to 24VDC. Naturally, the sound intensity will vary with the voltage, but not as much as one might assume. The graph at Figure 28-4 shows that the sound output from an alarm, measured in decibels, increases by only 8dB when voltage increases by almost a factor of five. Of course, the decibel scale is not linear, but human perception of sound is not linear, either.

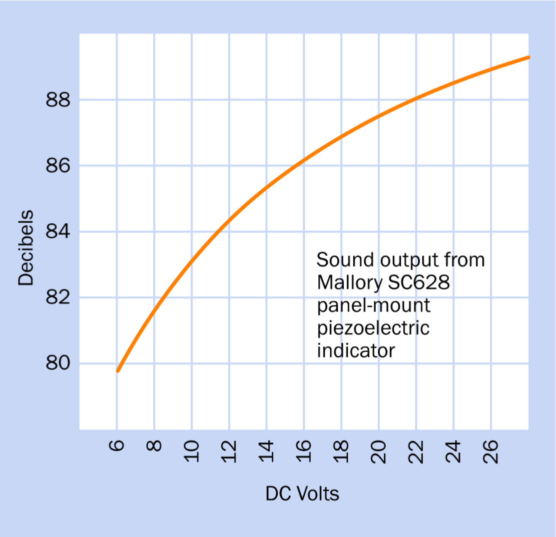

Figure 28-4. *Variation of sound output relative to voltage, in a commonly used piezoelectric indicator.*

Current

Typical piezoelectric indicators use less than 10mA (often as little as 5mA) and generate negligible heat.

Frequency

The most common frequencies for indicators range between 3kHz and 3.5kHz. Piezoelectric elements are inefficient for generating sounds below 1kHz.

Duty Cycle

Piezoelectric alerts generate very little heat and can be run on a 100% duty cycle.

If an alert will be pulsed briefly, the minimum pulse time is 50ms. A shorter duration will merely generate a clicking sound.

How to Use It

Appropriate Sound Intensity

An indicator should be chosen with reference to the environment in which it will be used. To be easily heard, it should be at least 10 dB louder than the ambient background noise.

Volume Control

Sound intensity can be reduced by reducing the voltage. Because an indicator does not consume much current, a trimmer can serve as a volume control. Alternatively, a rotary switch with a set of fixed-value resistors can select preset sound values.

However, in many indicators, variations in voltage may have relatively little effect on sound output, as shown in Figure 28-4.

Wiring

An indicator requires DC voltage. Because the indicator contains a transistor, polarity of the power supply is important. If the indicator has leads attached, the one intended for connection to the positive side of the power supply will be red. If it has pins, the longer pin will be for the positive connection.

What Can Go Wrong

The potential problems in an indicator are the same as those for a transducer. See "What Can Go Wrong" on page 254 in the previous entry.

headphone

29

The term **headphone** is used here to include almost any device that fits into or over the ear for the purpose of sound reproduction. (*Hearing aids* are not included.) Because headphones are used in pairs, the term is usually pluralized.

The term *phones* is fairly common as a colloquial diminution of *headphones* but is not used here.

An *earphone* used to be a single sound reproduction device designed for insertion into the ear, but has become rare. Pairs of *earbuds* are now common.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled headphones, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **transducer** (see Chapter 27)
- **speaker** (see Chapter 30)

What It Does

A headphone converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for reproduction of music for entertainment purposes, or for speech in telecommunications, broadcasting, and audio recording.

Two symbols for headphones are shown in Figure 29-1. The symbol on the left shows a single headphone or earphone; when this symbol is flipped horizontally, it can represent a microphone. The pictographic symbol on the right has been used for many decades, but is still often found in schematics.

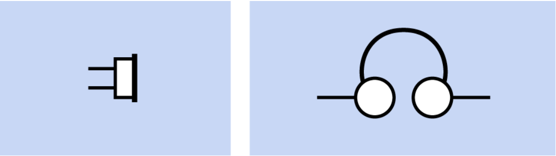

Figure 29-1. *Schematic symbols for a single earphone or headphone (left) and a pair of headphones (right).*

How It Works

Audio Basics

Sound is transmitted as pressure waves through a medium that is usually air but can be a gas, fluid, or solid. The speed of transmission will vary with the density and other attributes of the medium. Small hairs in the inner ears, known as *cilia*, vibrate in sympathy with pressure waves and transmit nerve impulses to the brain, which interprets the impulses as sound.

Three quantities describe the propagation of any type of wave, including a sound wave: its frequency (customarily represented with letter f), its speed of propagation (represented with letter v, for velocity), and its wavelength from peak-to-peak (represented by the Greek letter lambda, which appears as this λ symbol).

The relationship is defined by a very simple equation:

```
v = λ * f
```

Velocity is usually measured in meters per second, wavelength in meters, and frequency in Hertz, abbreviated Hz. One cycle per second is 1Hz. The H is always capitalized, as it refers to the name of Heinrich Rudolf Hertz, the first scientist to prove the existence of electromagnetic waves. One thousand Hertz can be written as 1 kiloHertz, almost always abbreviated as 1kHz (note that the k is lowercase).

The human ear is often described as being able to detect sounds between 20Hz and 20kHz, although the ability to hear sounds above 15kHz is relatively unusual and diminishes naturally with age. Sensitivity to all frequencies can be impaired by long-term exposure to loud noise.

Naturally occurring sounds can be converted to fluctuations in voltage by a **microphone**, which will be found listed as a *sensor* in Volume 3 of this encyclopedia. Artificial sounds can be generated as voltage fluctuations by oscillators and other electronic circuits. In either case, the output fluctuations can range between an upper limit set by a positive supply voltage and a lower limit established by electrical ground (which is assumed to be 0 volts). Alternatively, the fluctuations can range between the positive supply voltage and an equal and opposite negative supply voltage, with 0V lying midway between the two. This option can be less convenient electrically but is a more direct representation of sound, because sound waves fluctuate above and below ambient air pressure, which can be considered analogous to a ground state.

The concept of positive and negative sound waves is illustrated in Figure 29-2 (originally published in the book *Make: More Electronics*).

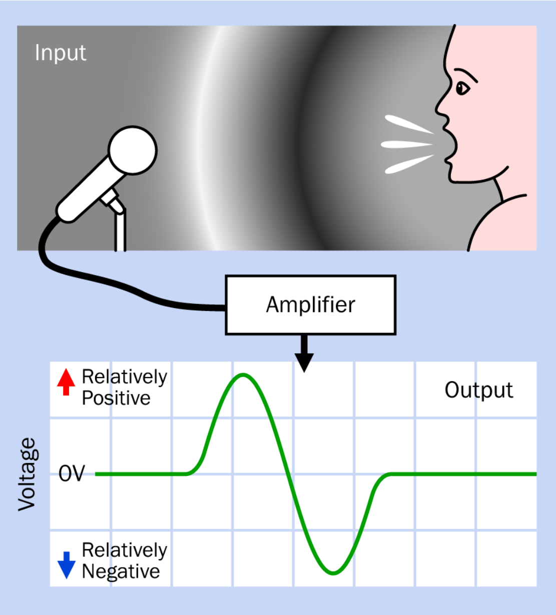

Figure 29-2. *The use of positive and negative voltages to represent a wave of high sound pressure followed by a trough of lower pressure.*

The topic of sound amplification is explored in detail in the entry on *op-amps* in Chapter 7.

A headphone inverts the function of a microphone by converting electricity back into air-pressure waves. This is done electromagnetically (moving a diaphragm in response to an electromagnet) or electrostatically (moving a membrane in response to electrostatic force between two charged electrodes).

Variants

Moving Coil

The most enduringly popular type of headphone uses a coil attached to a *diaphragm*. This is known as a *moving-coil* headphone, as the coil moves with the diaphragm. It can also be referred to as having a *dynamic driver* or *dynamic transducer*, “dynamic” referring to the movement of the coil.

The moving-coil concept is illustrated in Figure 29-3. The coil slides into a deep, narrow, circular slot in a magnet that is attached to the plastic frame of the headphone. The diaphragm is supported at its edges by a flexible rim. Variations in current passing through the coil create a fluctuating magnetic field that interacts with the field of the fixed magnet, causing the diaphragm to move in and out. A very similar configuration is used in many **loudspeakers**. Detail modifications may be made to increase efficiency, reduce production costs, or enhance sound quality, but the principle remains the same.

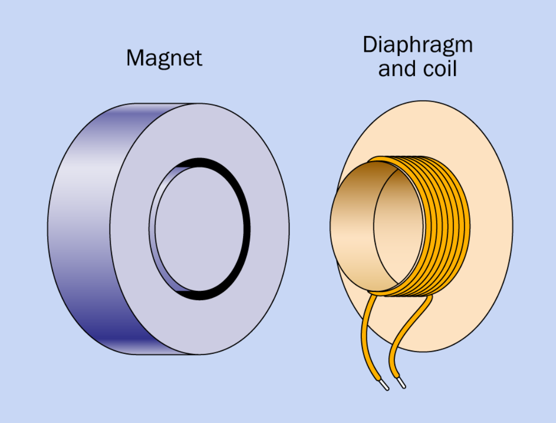

Figure 29-3. *The basic elements of a moving-coil headphone.*

The internal element of a headphone is shown in Figure 29-4. A plastic diaphragm is visible, measuring slightly less than 2" in diameter. The magnet and coil are concealed underneath.

The element in the previous figure is normally enclosed in an assembly such as the one in Figure 29-5, which incorporates a soft padded rim to rest upon the ear.

In an effort to achieve a more balanced frequency response, some designs use two moving-coil drivers in each headphone, optimized for low frequencies and high frequencies, respectively.

Earbuds, described after the next section, often use a miniaturized version of the moving-coil design.

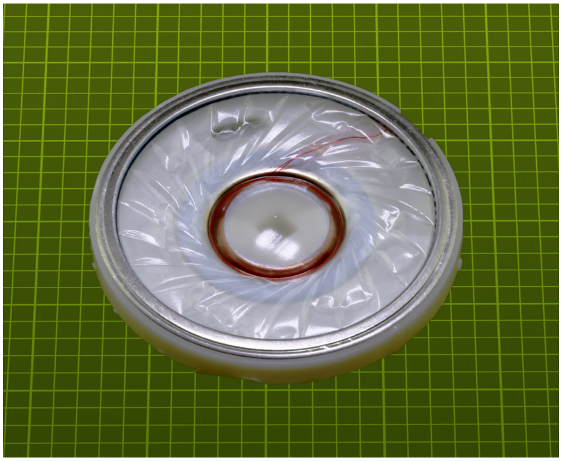

Figure 29-4. *The sound reproducing element removed from a headphone.*

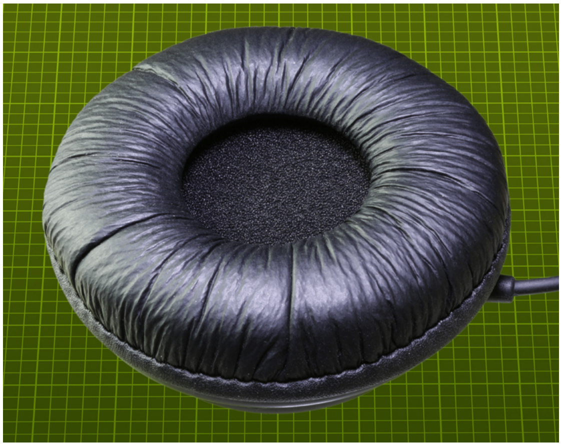

Figure 29-5. *The sound reproducing element from the previous figure is normally packaged inside this enclosure.*

Other Types

Electrostatic headphones use a thin, flat diaphragm suspended between two grids that function as electrodes. A fluctuating potential between the grids, coupled with a reverse-phase voltage on the diaphragm, will vibrate the diaphragm, generating pressure waves. A relatively high voltage is necessary to achieve this, anywhere from 100V to 1,000V, supplied through a conversion unit between the headphones and an amplifier. Electrostatic headphones are known for low distortion and an excellent high frequency response, at some extra cost.

Electret headphones work on a similar principle, except that the membrane is permanently charged, and a high voltage is not required. Electret headphones tend to be small, inexpensive, and not of high sound quality.

A *balanced armature* design, often referred to by the acronym *BA*, uses a pivoting magnet that is claimed to increase efficiency while reducing stress on a diaphragm. BA drivers can be extremely compact, contained within a sealed metal enclosure measuring less than 10mm x 10mm x 5mm. They are commonly used in conjunction with *in-ear* earphones, described in the next section.

Mechanical Design

Circumaural headphones use large soft pads to encircle the ear and block external noise. Their size tends to make them heavy, requiring a well-designed headband to provide comfortable support. *Supra-aural* headphones are smaller and lighter, resting on the ears instead of enclosing them. They cannot exclude ambient noise, and may have inferior bass response compared with the circumaural type.

Open-back headphones, also known as *acoustically transparent*, are favored by some audiophiles because their vented outer surfaces are thought to create a more natural sound, similar to that of a **speaker**. The open backs naturally allow ambient noise to intrude, but also allow the sound generated by the headphones to be heard by others in a room. *Closed-back* headphones contain their sound and provide more insulation against ambient noise.

Earbuds rest just within the outer folds of the ear, facing inward like a pair of tiny **speakers**. They are easily dislodged and provide very little insulation against ambient noise. Their use became common after the introduction of Apple's iPod. A pair of earbuds, one of them with its plastic cover removed, is shown in Figure 29-6.

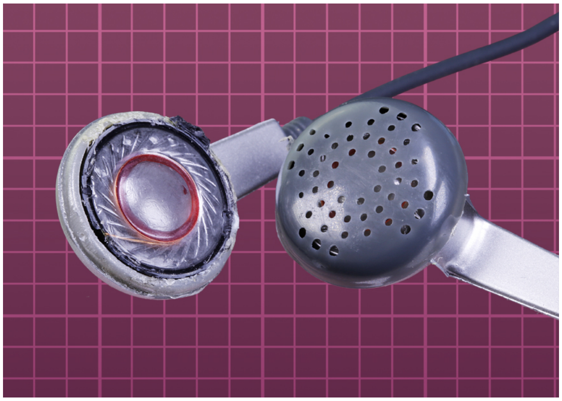

Figure 29-6. *A pair of earbuds, one with its cover removed to show the sound reproducing element, which closely resembles the diaphragm in a full-sized headphone.*

In-ear headphones are designed for insertion into the ear canal, often using a soft sheath that conforms with the ear like an earplug. This sheath is disposable for hygienic reasons, and because it may lose some of its plasticity with use. It excludes most ambient noise, and by minimizing the air gap between the driver of the headphone and the ear drum enables a high quality of sound reproduction.

In-ear headphones are also known as *in-ear monitors*, *IEMs*, *ear canal headphones*, *earphones*, and *canalphones*. A pair of in-ear headphones is shown in Figure 29-7, one of them with its foam sheath removed. The rectangular silver-colored object in the headphone on the left contains a transducer to create sound pressure.

A *headset* consists of one or two headphones plus a flexible microphone that extends to the promiximity of the mouth of the user.

Noise-cancelling headphones, popularized by Bose, monitor external noise with a built-in microphone and generate sound of opposite phase, to provide some cancellation. They are particularly effective on jet aircraft, where background noise tends to be consistent.

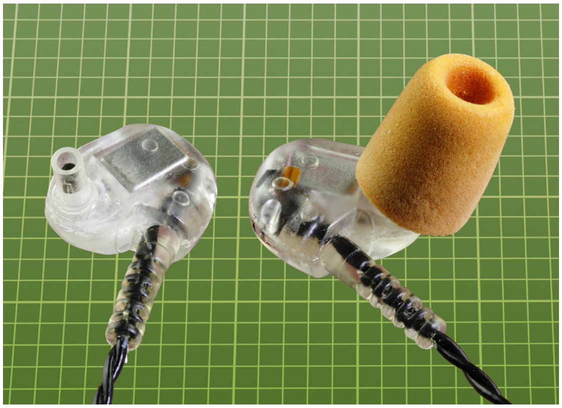

Figure 29-7. *A pair of in-ear headphones, supplied with disposable foam plugs that conform flexibly with the ear canal. The headphone on the left is shown with its plug removed.*

Although an *earphone* is almost obsolete, it is still obtainable from specialty suppliers. It has a high impedance, making it suitable for use with a *crystal-set radio*. An earphone is pictured in Figure 29-8.

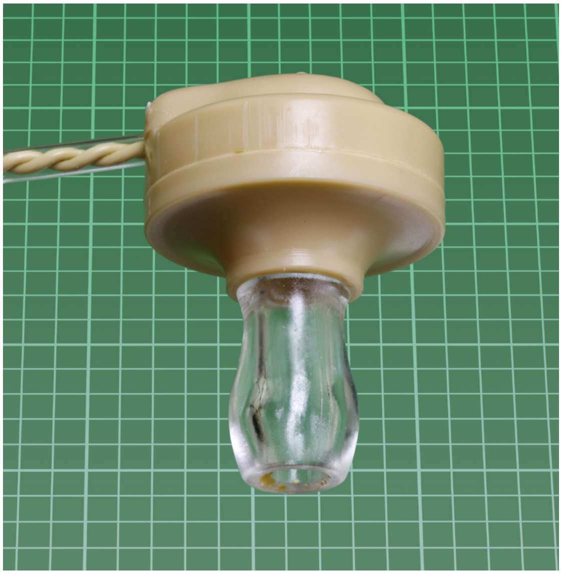

Figure 29-8. *A vintage earphone of the type suitable for use with a crystal-set radio.*

Values

Intensity

Sound pressure is measured in decibels. For a complete explanation and discussion of weighted and unweighted decibel scales, see "Sound Pressure" on page 251 in the **transducer** entry.

Frequency Response

A plot of sound pressure as a function of frequency shows the *frequency response* of a headphone. Measuring the sound pressure meaningfully is a challenge, because the ear canal will add coloration to the sound and can amplify some frequencies while masking others. Ideally, measurement should be done at the ear drum, but this is not feasible. Consequently, high-end headphones are evaluated by making sound measurements inside simulated ear canals in a dummy human head.

A comparison between a high-quality $500 audio product and a transducer that is sold as a component for less than $1 illustrates the difference in frequency response; see Figure 29-9. The Sennheiser headphones have a smooth response that rises toward the low end, compensating for the lack of bass response that tends to be a problem in headphones, and the relative insensitivity of the human ear to low frequencies. The fluctuations at the high end are within about 5dB.

By comparison, the Kobitone emphasizes the range between 3kHz and 4kHz because its primary task is to be heard, and these are the frequencies where human hearing is most sensitive. Its low-frequency response trails off (although is still much better than that of a piezoelectric transducer, where the low response typically diminishes by 40dB to 50dB). The low-frequency output of the Kobitone is actually impressive bearing in mind that the component is only 9mm in diameter. It draws 60mA at 5VAC.

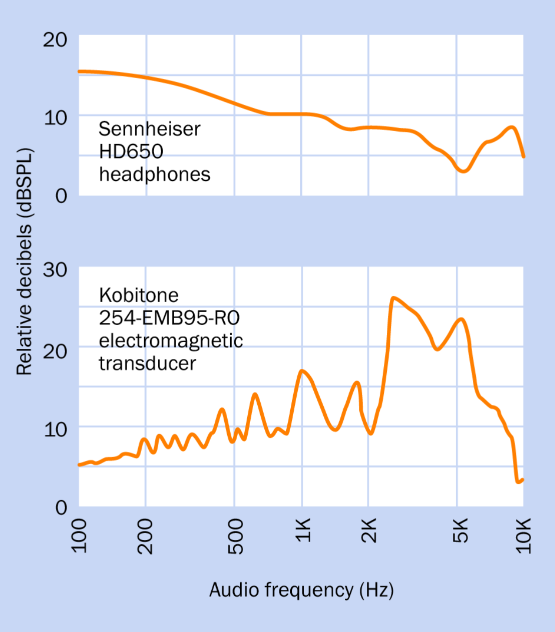

Figure 29-9. *Comparison between frequency responses of a $1 electromagnetic transducer intended as an audio alert and a $500 pair of headphones intended for sound reproduction. The upper graph is derived from a review online at headroom.com; the lower graph is from the manufacturer's datasheet.*

Some manufacturers of audio reproduction equipment prefer not to supply frequency response curves. Instead they may claim, for example, that the frequency response of a product ranges from 100Hz to 20kHz. This claim means very little unless it is accompanied by a range of sound pressure levels. If the frequency response is consistent within a range of, say, plus-or-minus 5dB, this may be acceptable. If the range is plus-or-minus 20dB, it is not acceptable. The ability to reproduce a high note or a low note is not useful if the sound is too faint to hear.

Distortion

The *total harmonic distortion* (THD) of any audio equipment measures its tendency to add spurious *harmonics* of a single frequency. If headphones are required to reproduce a pure 1kHz sinewave, they will also tend to create an additional 3kHz tone that is an artifact. This can be caused by the mechanical behavior of a vibrating diaphragm. The human ear recognizes distortion as a fuzzy or rasping sound. A square wave theoretically contains all the harmonics that are odd multiples of its fundamental frequency, and sounds extremely distorted.

THD should be less than 1% in good-quality audio devices.

Impedance

The electrical impedance of headphones is relevant in that it should match the output specification of the amplifier that drives them.

What Can Go Wrong

Overdriving

Headphones can be damaged by overdriving them. Because a low frequency requires larger excursions of a diaphragm to transmit the same energy as a high frequency, headphones are especially vulnerable to being damaged by bass at high volume.

Hearing Damage

Human hearing can be damaged by prolonged listening to headphones at a high volume. Some controversy remains regarding an acceptable limit for sound pressure.

Mismatched Impedance

If the impedance of headphones does not match the output of the amplifier driving them, distortion or a skewed frequency response can result. This is known as *mismatching*.

Incorrect Wiring

In most consumer products, a pair of headphones will share a common ground. While the connections in a typical three-layer jack plug have been standardized, hand-wired repairs or extensions should be tested carefully. Incorrect wiring will cause unpredictable results.

speaker

30

The term **speaker** is a diminution of *loudspeaker*. The full word is now so rarely used, some catalogs do not recognize it as a search term. This encyclopedia acknowledges contemporary usage by using **speaker** rather than *loudspeaker*.

A fully assembled consumer product can be referred to as a "speaker," but it also contains one or more individual components that are described as "speakers." To resolve this ambiguity, referring to the components as *drivers* is helpful, but this practice can create more ambiguity because other types of components are also called "drivers." The only real guide to the meaning of *speaker* is the context in which it is used.

For the purposes of this entry, a speaker is a *sound reproduction device*, distinguished from a typical electromagnetic **transducer** by being larger and more powerful, with a more linear frequency response. A transducer may be used as a *noise-creating device* to provide an alert, informing the user of the status of a piece of equipment. Because some speakers have been miniaturized for use in handheld products, they may be used as transducers, allowing some overlap between the two categories.

Because this encyclopedia assigns more emphasis to electronic components than to consumer products, this entry provides only a superficial overview of fully assembled speakers, and deals more with the drivers inside them, their principles of operation, and the general topic of sound reproduction.

OTHER RELATED COMPONENTS

- **headphone** (see Chapter 29)
- **transducer** (see Chapter 27)

What It Does

A speaker converts fluctuations of an electric signal into pressure waves that the human ear perceives as sound. It can be used for entertainment purposes or to provide information in the form of spoken words or distinctive sounds (as in the case of a miniature speaker in a cellular phone, playing a ring tone).

The internationally accepted schematic symbol for a speaker is shown in Figure 30-1.

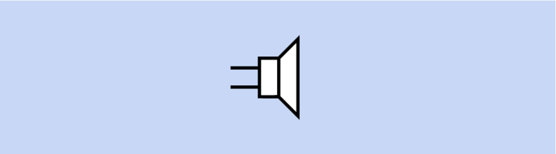

Figure 30-1. *Only one symbol exists to represent a speaker. This is it.*

How It Works

For a summary of basic concepts and terminology relating to sound and its reproduction, see "Audio Basics" on page 261 in the previous entry.

Construction

A speaker contains a *diaphragm* or *cone* with a coil attached to it. Fluctuations of current through the coil interact with a permanent magnet, causing the speaker to emit pressure waves that are proportionate with the current. The design is similar in concept to that of a headphone, shown diagrammatically in Figure 29-3. The primary difference is that a speaker of around 2″ or more will use a cone rather than a flat diaphragm. The cone shape is more rigid and creates a more directional sound.

A 2″ speaker rated for 1/4W with a 63Ω coil is shown in Figure 30-2, undamaged on the left but with its cone cut out on the right. The neck of the cone, which is normally inserted in the circular groove in the speaker magnet, is shown with the inductive coil wrapped around it.

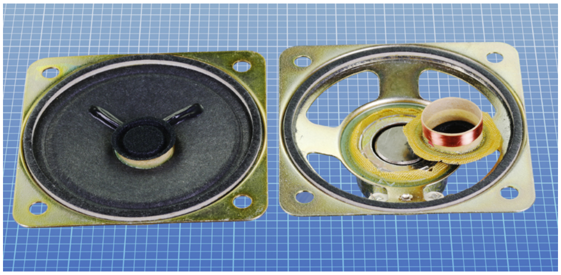

Figure 30-2. *On the left is a 2″ speaker. On the right, its cone has been cut away to reveal the magnet, with a circular groove in it. The neck of the cone, which normally slides into the groove, is shown removed.*

A speaker with a cone 4″ in diameter is shown from the rear in Figure 30-3.

A miniature surface-mount speaker is shown from the front and from the rear in Figures 30-4 and 30-5. It measures just under 0.4″ diameter and was made for Motorola. Its power rating is 50mW.

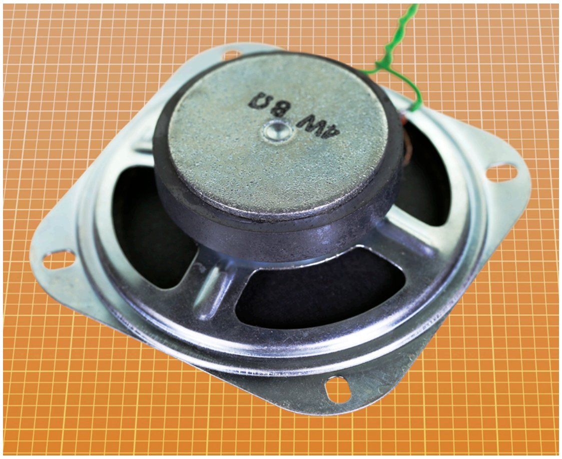

Figure 30-3. *The back side of a speaker with a cone measuring approximately 4″ in diameter. Its magnet is the large round section that is uppermost. This unit is rated for 4W and has an impedance of 8Ω.*

A speaker designed for a cellular phone is shown in Figure 30-6. Note the close resemblance in design to the driver used in an earbud, shown in Figure 29-6 in the previous entry.

In the past, speaker cones were made from tough, fibrous paper. Modern cones are more likely to be plastic, especially in small sizes.

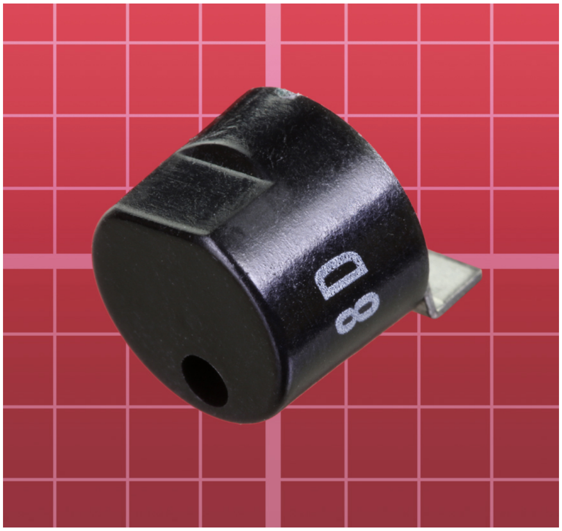

Figure 30-4. *Front view of a surface-mount speaker measuring less than 0.4″ diameter.*

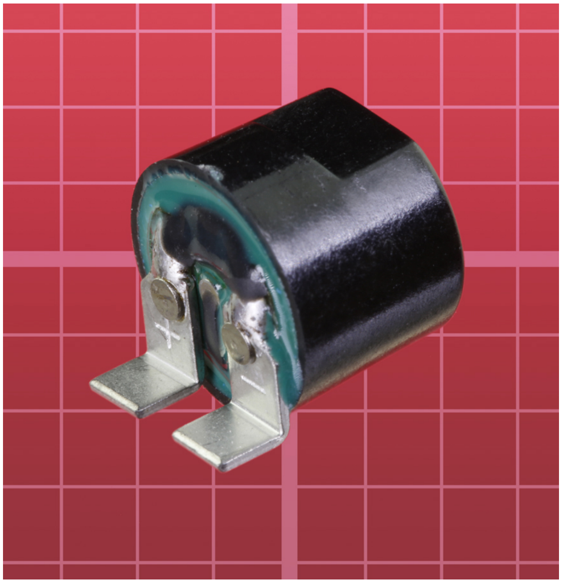

Figure 30-5. *Rear view of the speaker shown in the previous figure.*

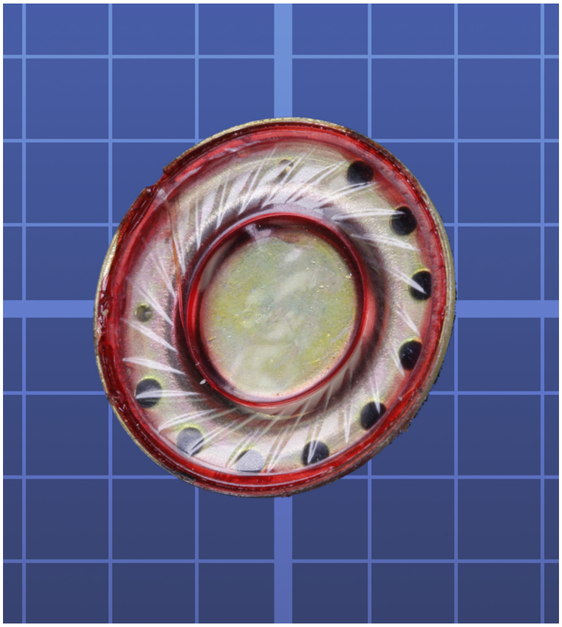

Figure 30-6. *A miniature speaker measuring slightly more than 1/2" diameter and only 0.13" thick, designed for use in a cellular phone. It has an impedance of 150Ω.*

Multiple Drivers

Generally speaking, a large-diameter speaker cone is more effective than a small cone at moving the greater volumes of air associated with reproduction of bass notes. However, the inertia of a large cone impairs its ability to vibrate at high frequencies.

To address this problem, a large speaker and a small speaker often share a single enclosure. A *crossover network* using coils and capacitors prevents low frequencies from reaching the small speaker and high frequencies from reaching the large speaker. The basic principle is shown in the simplified schematic in Figure 30-7.

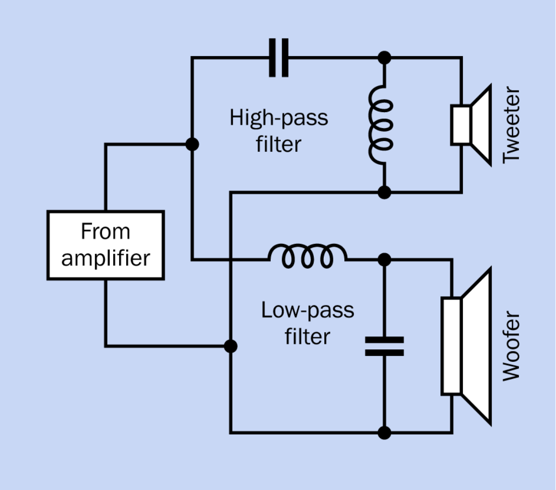

Figure 30-7. *The basic principle of a crossover network.*

Because the crossover network must be "tuned" to match the characteristics of the speakers, and because the combined sound pressure of the speakers must be relatively consistent over a wide range of frequencies, an actual network usually includes additional components.

Because the audio output from an amplifier consists of alternating current, polarized capacitors cannot be used. Polyester capacitors are common.

The small speaker in a pair is known as a *tweeter* while the large speaker is a *woofer*. Although these appellations were whimsically coined, they have endured.

More than two speakers may be combined in an enclosure, in a wide variety of configurations.

Venting

A speaker radiates pressure waves from the back side of its cone as well as from the front, and because the waves from front and back are opposite in phase, they will tend to cancel each other out.

In a miniature speaker, this problem can be addressed simply by sealing the section of the enclosure at the rear. For larger components, a more efficient enclosure can be designed with a *vent* or *reflex port* at the front. Pressure waves from the back of the speaker are diverted over a sufficient distance inside the cabinet so that by the time they emerge through the port, they are approximately in phase with low frequencies from the front of the speaker, although the back wave will lag the front wave by one wavelength.

This design is referred to as a *bass-reflex enclosure*, and was almost universal in high-fidelity components until amplifiers became increasingly powerful during the 1960s. At that point, Acoustic Research, located in Massachusetts, marketed a product line in which speaker enclosures were sealed, the argument being that when an amplifier can deliver 100W per channel, efficiency is no longer an issue, and a sealed enclosure can eliminate compromises associated with a bass-reflex design.

Acoustic Research referred to their concept as "air suspension," as the cushion of air in the sealed cavity helped to protect the speaker by limiting its excursion. This configuration is now often referred to as a *closed-box speaker*. Some audiophiles argue that it must always be inherently superior to a bass-reflex design, partly because of the one-wavelength lag time associated with a reflex port. However, as in many aspects of sound reproduction, the debate is inconclusive.

Resonance

The enclosure for a speaker will tend to have a dominant resonant frequency. This should be lower than the lowest frequency that the speaker will reproduce; otherwise, the resonance will emphasize some frequencies relative to others, creating unwanted peaks in the response.

One reason why high-quality speakers tend to be physically heavy is to reduce their resonant frequency. A modern Thiel speaker assembly, for instance, uses a front panel of particle board that is a full 2″ thick. However, heavy enclosures are expensive to transport and inconvenient to locate or relocate in the home.

To address this problem, a tweeter and a woofer can be mounted in separate boxes. The enclosure for the tweeter can be very small, lightweight, and suitable for placement on a shelf, while the heavy box for the woofer can go on the floor. Human senses have difficulty locating the source of low-frequency sound, so the woofer can be located almost anywhere in a room. In fact, its single speaker can serve both stereo channels.

This configuration has become the default for computer speakers. It is also used in home-theater systems, where the woofer has now become a subwoofer capable of very low frequency reproduction.

Miniature Speakers

If an electronics project has an audio output, and the circuit board will be sharing an enclosure with a small speaker, the size of the box and the material from which it is fabricated will affect the sound quality significantly. A box made from thin hardwood may add resonance that sounds pleasing, if the speaker is being used just for simple electronic tones. By comparison, a metal box may sound "tinny." A box fabricated from a plastic such as ABS will be relatively neutral, provided the plastic is reasonably thick (1/4″ being preferable).

Variants

Electrostatic Speaker

The principle of an electrostatic speaker is the same as that of an electrostatic headphone. A charged membrane is stretched between two grids in front of it and behind it that act as electrodes. Because the membrane is so light, it responds with very little latency, and its large surface area creates a diffused sound that many audiophiles find pleasing. However, a high voltage is required to drive electrostatic speakers, and they are not cheap.

Powered Speakers

A unit containing its own driver electronics is referred to as a *powered speaker*, and is used almost universally with desktop computers, because the computer itself does not contain a power amplifier. Powered speakers may also enable a more versatile crossover network.

A subwoofer may have its own amplifier allowing control of the cutoff frequency above which the speaker will not attempt to reproduce sound. The electronics can include protection for a speaker against being overdriven.

Wireless Speakers

A wireless link between a stereo receiver and its speakers will eliminate the speaker wires that are normally necessary. However, the speakers themselves must be powered, and will have to be wired to electric outlets.

Innovative Designs

The need for small speakers in consumer products such as laptop computers has encouraged innovative designs. The speaker in Figure 30-8 is just 1" square, and its shape is easier to accommodate in a small product than the traditional circular speaker. In Figure 30-9, the interior of the same speaker shows that inductive coils are applied to a square plastic diaphragm.

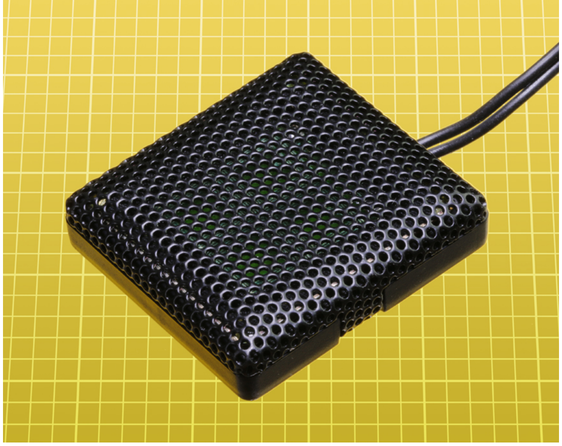

Figure 30-8. *A speaker 1" square, suitable for use with a small electronic device.*

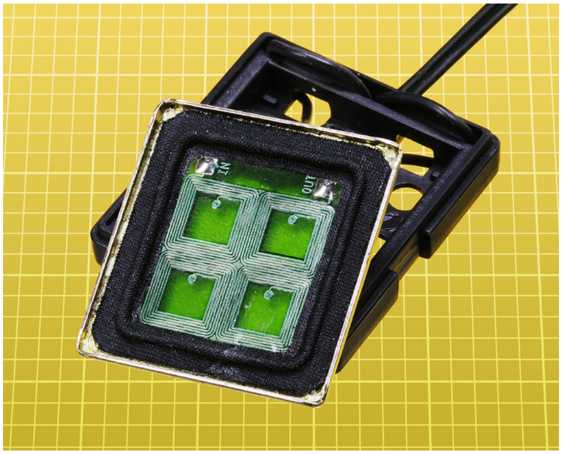

Figure 30-9. *The speaker from the previous figure, opened to show its inductive coils applied to a square plastic diaphragm.*

Values

The typical *impedance* for speakers in audio systems is 8Ω. Small speakers may have a higher impedance, which can be useful when driving them from devices that have limited power, such as a TTL-type 555 timer.

In the United States, the diameter of a circular speaker is usually expressed in inches. Speakers larger than 12" are rare for domestic use. A 4" speaker used to be considered minimal because of its limited low-end frequency response, but much smaller speakers have become common in portable devices.

The low-end frequency response of a miniature loudspeaker designed to be surface-mounted on a circuit board will be very poor. The graph in Figure 30-10 was derived from data supplied by the manufacturer.

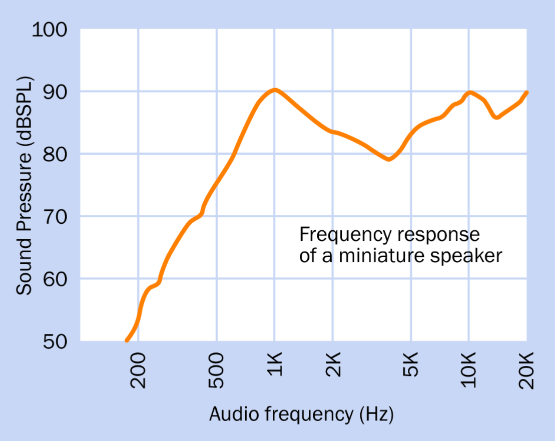

Figure 30-10. *Frequency response for a speaker measuring 15mm x 15mm x 5mm. Small dimensions and lack of an enclosure result in a negligible low-end response.*

Power rating for speakers is in watts, measured on a root-mean-square (RMS) basis.

Sensitivity is measured in decibels, at a distance of one meter, while a speaker is reproducing a single constant tone with a power input of 1W. A speaker assembly designed for undemanding home use may be rated at 85dB to 95dB.

Efficiency is a measure of sound power output divided by electrical power input. A value of 1% is typical.

What Can Go Wrong

Damage

As is the case with headphones, the most common problem affecting a speaker is damage caused by overdriving it. Because a low frequency requires larger excursions of a speaker cone to transmit an amount of energy comparable to that of a high frequency, loud bass notes can be hazardous to the cone. On the other hand, if an amplifier generates distortion (perhaps because it, too, is being overdriven), the harmonics created by the distortion can damage high-frequency speakers.

Magnetic Field

Even a small speaker (measuring 2" or less) contains a magnet that is sufficiently powerful to cause problems if it is in close proximity to other components, especially if sensors such as reed switches or Hall-effect switches are being used. Initial circuit testing should be done with the speaker as far away as possible, to eliminate it as a source of interference.

Vibration

Solder joints will be stressed if they are subjected to low-frequency vibration from a speaker. Loose parts may rattle, and parts that are bolted into place may become unbolted. The speaker itself may become loose. Thread-locker such as Loc-Tite should be applied to nuts before they are tightened.

Index

Symbols

2N682 SCR, 2
4047B timer, 75
40TPS12 SCR, 2
4131-503 digital potentiometer, 62
4543B decoder, 168
4543B decoder chip, 168
555 timer, 70–73, 76, 79–84
5555 timer, 74
556 timer, 73
558 timer, 73
7447 decoder, 148
74HC00 logic gate, 93
74HC123 timer, 75
74HC163 counter, 132
74HC221 timer, 76
74HC4514 decoder, 147
74HC4515 decoder, 139, 147
74HC5555 timer, 74
74HCT5555 timer, 74
74LS123 timer, 75
74LS148 decoder, 139
74LS47 decoder, 148
7555 timer, 74
7556 timer, 74
BTA208X-1000B triac, 16
BTB04-600SL triac, 16
CAT5114 digital potentiometer, 64
CD4047B timer, 75
CMX60D10 relay, 27
D804 optocoupler, 37
DB3TG diac, 12
DC60S7 relay, 27, 28
HEF4528B timer, 75
LM339 comparator, 42
LM3914 driver, 236
LM741 op-amp, 49, 53
M74HC4538 timer, 75
MAC97A6 triac, 16
MAX6954 driver chip, 234
MC14538B timer, 75
NTE926 timer, 73
OPTEKD804 optocoupler, 34
PT-2040PQ transducer, 253
PUI XL453 audio indicator, 257
SSD1306 LED display module, 235
SSD1331 LED display module, 235
TIL311 driver, 236
WS0010 LED display module, 235
X0403DF SCR, 2, 5

A

A-weighting audio standard, 252
absolute zero, 173
acoustically transparent headphones, 264
activator, electroluminescent, 244
active matrix LCD, 161
additive primaries, 166
address of tap, digital potentiometer, 62, 63
Adobe 1998, 167
alert, 249, 257
alternating current, 172
American Standard Code for Information Interchange (ASCII), 163
AMOLED (active-matrix OLED), 246
analog input/output, 66

analog integrated circuits
  comparator, 39–48
    (see also comparator)
  digital potentiometer, 59–67
    (see also digital potentiometer)
  op-amp, 49–58
    (see also op-amp)
  optocoupler, 33–37
    (see also optocoupler)
  solid-state relay, 25–31
    (see also solid-state relay)
  timer, 69–87
    (see also timer)
analog multiplexer, 151, 152
analog output optocoupler, 36
analog-digital converter, 39, 47, 59
AND gate, 45, 90, 91, 92, 97, 98, 100, 102, 103
anode
  LED display, 232
  LED indicator, 213
  vacuum-fluorescent display (VFD), 240
arithmetical operations, binary, 91
arithmetical operators, 89
ASCII (see American Standard Code for Information Interchange)
astable multivibrator, 46, 55, 69
  (see also timer)
astable timer, 70, 72, 77, 80
asynchronous counter, 118, 134
asynchronous flip-flop, 107
asynchronous input, 123
audio alerts
  audio indicator, 257–260
  transducer, 249–255
audio frequency, 54, 251, 252, 253, 254, 257, 258
audio indicator, 249, 257–260
  PUI XL453, 257
  current, 260
  design, 257–258
  duty cycle, 260
  formats, 258
  frequency, 258, 260
  function, 257
  history, 258
  potential problems, 260
  sound intensity, 260
  sound patterns, 258
  use of, 260
  values, 259
  variants, 258
  voltage, 259
  volume control, 260
  wiring, 260
audio transducer (see transducer)

B

back-EMF, 29, 30, 87
backlighting, electroluminescent, 167
backplane, 161
ballast, 180, 192, 194, 225
band gap, 207
bandpass filter, 59
bandwidth, 36
bass-reflex enclosure, 270
bayonet base, 177
BCD output, 135
beeper, 249, 257
bi-pin tube, 194
bidirectional multiplexer, 151
bidirectional optocoupler, 36
bidirectional shift, 125
bidirectional thyristor, 11
binary addition, 91
binary arithmetic, 89, 92
binary counter, 134
binary digit, 121, 134
binary ripple counter, 121
binary-coded decimal (BCD) output, 135
bipolar transistor, 1, 18, 70
bistable displays, 167
bistable multivibrator, 46, 69, 107
  (see also flip-flop)
bit, 121, 134
bitmap, 159, 227
black body radiation, 173
Boolean algebra, 89
Boolean operators, 89, 91
breakdown voltage, 1, 4, 20, 184
breakover frequency, 53
breakover level, 12
breakover voltage, 4, 14, 15, 20
bubble, 90, 116, 123, 133, 155
buffer, 90, 128
burn in, 160
burst mode, 84
buzzer, 249, 257, 258

C

Camenzind, Hans, 70
canalphones, 264
candela, 178
candlepower, 178
capacitor
  coupling, 84
  light-emitting, 244
carbon arc, 171
cascade, 134
cascaded encoders, 142
cascading, 157
cathode, 188, 192, 232, 239
cathode vacuum-fluorescent display (VFD), 240
cathode-ray tube (CRT), 160, 239, 244
CCFL (see cold-cathode fluorescent lamp)
CCT (see correlated color temperature)
cd (candela), 178
ceramic wafer, 250, 258
CFL (see compact fluorescent lamp)
channel current, 155
channels, 42, 52, 153
charge storage, 23, 24
chemiluminescence, 244
chip select pin (CS), 63, 64
cilia, 261
circumaural headphones, 264
clipped output, 50
clock, 121
clock input, 131, 133
clocked comparator, 47
clocked in, 121

closed-back headphones, 264
closed-box speaker, 270
CLR operation, 133
CMOS 555 timer, 74
CMOS chips, 48, 94
CMYK system, 166
CO2 lasers, 201
code hopping encoder, 139
coherent light, 197, 200
cold-cathode fluorescent lamp (CCFL), 191, 194
cold-cathode fluorescent panel, 167
color
  fluorescent lights, 194, 241, 242
  in LCDs, 166–167
  of LEDs, 205, 207, 209–211, 212, 219–222, 222, 226
  primary, 166–167
color rendering index (CRI), 174
color super twisted nematic (CSTN) LCD, 161
color temperature, 173, 209, 219, 222
color-rendering index (CRI), 212
common anode, 232
common cathode, 232
common mode voltage range, 53
common pin, 161
commutating dv/dt, 23
compact fluorescent lamp (CFL), 191
comparator, 39–48, 70, 83, 90
  LM339, 42
  analog-digital converter, 47
  AND gate, 45
  bistable multivibrator, 46
  clocked, 47
  continuous converter, 47
  design, 39–42
  digital, 39
  dual, 42
  function, 39
  hysteresis, 39, 44
  latched, 47
  level shifter, 46
  and op-amps, 42, 52
  potential problems, 47–48
  relaxation oscillator, 46
  use of, 44–47
  values, 43–44
  variants, 42–43
  window, 42
  window comparator, 46
  zero point finder, 47
conduction angle, 7, 9
cone, 268
confused inputs, 48
continuous converter, 47
converter, serial to parallel, 107
correlated color temperature, 222
counter, 107, 121, 131–138
  74HC163, 132
  asynchronous, 118, 134
  binary, 134
  binary ripple, 121
  clock sources, 135
  decade, 135
  decoded output, 134
  descending output, 136
  design, 132–134
  divide-by-16, 134
  function, 131–132
  glitches, 138
  gray code, 131
  hexadecimal, 134
  modulus-16, 134
  modulus/modulo, 132
  multiple stages, 136
  octal, 134
  output states, 136
  pin identifier, 133
  potential problems, 137–138
  programmable, 137
  ring, 131, 134
  ripple, 107, 134, 148
  rising edge and falling edge, 136
  schematic representation, 131–132
  single and dual, 136
  synchronous, 134, 136
  values, 137
  variants, 134–137
coupler, output, 198
coupling capacitor, 84
CRI (see color rendering index)
crossover network, 269
crosstalk, 161
crowbar overvoltage protection, 2
CRT (see cathode-ray tube)
crystal, 249
crystal lasers, 201
crystal-set radio, 265
CS (chip select pin), 63, 64
current
  alternating, 172
  audio-indicator, 260
  channel, 155
  direct, 172
  forward, 211
  gate threshold, 18
  holding, 2, 18
  latching, 2, 18
  leakage, 156
  transducer, 254
current amplifier, 2, 18
Current Transfer Ration (CTR), 36

D

D-type flip-flop, 114, 116
data distributor, 140, 146, 152
data selector, 140, 146, 151, 152 (see also multiplexer)
data transfer, 62
Db (see decibels)
dBA, 253
debounced, 66
decade counter, 135
decibel, 251–252, 259, 265
decoded output counter, 134
decoder, 134, 139, 140, 143, 145–149, 151, 153, 157
  7447, 148
  74HC4514, 147
  74HC4515, 147
  74LS47, 148
  design, 148
  function, 145–146
  input devices, 145–146
  LED driver, 146
  potential problems, 149–149
  schematic symbol, 146–146

- seven-segment, 146
- similar devices, 146
- use of, 149
- values, 148
- variants, 148

decoder chip, 151
demultiplexer, 140, 146, 151
derated relay, 30
derating curve, 213
descending output counter, 136
diac, 1, 11–14, 15, 22
- DB3TG, 12
- abbreviations, 14
- design, 12–13
- function, 11
- potential problems, 14
- switching AC, 13
- symbol variants, 11
- values, 14
- variants, 13

diaphragm, 262, 268
dies, 219
differential multiplexer, 152
digipot (see digital potentiometer)
digital comparator, 39
digital integrated circuits
- counter, 131–138
- decoder, 145–149
- encoder, 139–143
- flip-flop, 107–119
- logic gate, 89–106
- multiplexer, 151–157
- shift register, 121–129

digital multiplexer, 151
digital potentiometer, 59–67, 156
- 4131-503, 62
- CAT5114, 64
- address of tap, 62, 63
- advantages/disadvantages, 59–60
- connections and modes, 64
- data transfer, 62
- design, 60–61
- function, 59–60
- higher resolution, 66
- I2C protocol, 63
- potential problems, 66
- programmable, 59
- SPI protocol, 62
- taper, 62
- up/down protocol, 63
- use of, 66
- values, 65–66
- variants, 61–65
- volatile/nonvolatile memory, 61

digital-analog converter, 39, 47, 59
digitally adjustable potentiometer (see digital potentiometer)
digitally controlled potentiometer (see digital potentiometer)
digitally programmed potentiometer (DPP) (see digital potentiometer)
digpot (see digital potentiometer)
diode, PIN, 35
diode, protection, 30
direct current, 172
disallowed state, 135
discrete semiconductor
- diac, 22

discrete semiconductors, 1
- (see also thyristor)

dispersion, 178, 208, 225
display
- electroluminescent (see electroluminscent (EL) display)
- LED (see LED display)

display module, 164, 231
divide-by-16 counter, 134
dot-matrix displays, 163–165, 230, 234–236, 239
double super twisted nematic LCD, 161
drift, 48
driver, 267
driver chip, 233–234
driver, dynamic, 262
dual comparator, 42
dual counter, 136
dual digital potentiometer, 61
dual inputs, 50–51, 127
dual monostable timer, 75–76, 77
dual voltage device, 39
dual voltage power supply, 50
duty cycle, 260
dv/dt triggering, 8
dynamic driver, 262
dynamic transducer, 262

E

e-ink displays, 167
ear canal headphones, 264
earbuds, 261, 263, 264
earphone, 261, 264
edge pass, 69
edge triggering, 69
edge-triggered shift register, 123
efficiency, 272
EL (see electroluminescent (EL) display)
electret, 264
electric newspaper, 236
electroluminescence, 243
electroluminescent (EL) display, 243–247
- derivation, 244
- design, 243
- flexible ribbons, 245
- function, 243
- OLED, 246–247
- panels, 244–245
- phosphors, 244
- rope light, 245
- variants, 244–247

electroluminescent backlighting, 167
electromagnetic relay, 33
electromagnetic transducer, 250, 253
electronic ballast, 193, 225
electronic paper displays, 167
electrostatic headphones, 263
electrostatic speaker, 271
Enable, 116
encoded output, 134
encoder, 139–143, 152
- 74HC4515 decoder, 139
- 74LS148 decoder, 139
- cascaded, 142
- code hopping, 139
- design, 140–141

  function, 139–140
  potential problems, 143
  priority, 140, 141
  rotary, 139
  rotational, 66
  schematic symbol, 139
  similar devices, 140
  simple, 141
  tri-state, 141
  use of, 142
  values, 142
  variants, 141
equal-loudness contours, 252
erratic output, 48

F

falling voltage, 43
falling-edge triggered shift register, 123
fanout, 94
feedback
  negative, 51–52
  positive, 44
ferromagnetic disc, 250
FF (see flip-flop)
fiber lasers, 201
FIFO (see first-in, first-out (FIFO) shift register)
filament, 171, 239
film-compensated super twisted nematic (FSTN) LCD, 161
first-in, first-out (FIFO) shift register, 123
flange base, 177
flexible ribbon, 245
flicker, 193
flip-flop, 46, 70, 89, 107–119, 121, 132
  asynchrononous/synchronous, 107
  D-type, 114
  design, 107–116
  forbidden states, 110–112
  function, 107
  JK, 112–113
  master-slave, 113
  NAND-based, 108–109
  NOR-based, 109–110
  packaging, 117
  potential problems, 118–119
  summary of types, 116
  use of, 118
  values, 117–118
  variants, 116–117
flip-flop emulation, 82
floating inputs, 127
floating pins, 105
fluoresce, 192
fluorescence, 191, 239
fluorescent lamp, 191
fluorescent light, 24, 167, 174, 191–196, 207, 217, 223, 239
  (see also vacuum-fluorescent display)
  ballast and starter, 192
  brightness, 195
  CCFLs, 194
  comparisons, 194
  design, 191–195
  flicker, 193
  function, 191
  potential problems, 195
  sizes, 194–194
  spectrum, 195
  tubes, 191
  values, 195
  variants, 193
fluorescent tubes, 191
flux, 177
foot-candles, 178
forbidden states, 110–112
frequency
  audio, 54, 251, 252, 253, 254, 257, 258, 260
  breakover of op amp, 53
  unity gain, 53
frequency response, 265
fused quartz, 176

G

gain, 50
gain medium, 198
game sound, 85
gamut, 167, 207
ganged analog potentiometer, 61
gas-discharge device, 191
gate threshold current, 18
gate turn-off thyristor (GTO), 1, 11, 15
gated circuit, 113
glitches, 138, 149
glow discharge, 185
gray code counters, 131
ground loops, 35
GTO (see gate turn-off thyristor)

H

half adder, 92
halogen lighting, 176, 217, 221
harmonics, 266
headphone, 261–266
  acoustically transparent, 264
  balanced armature, 264
  circumaural, 264
  closed-back, 264
  design, 261–262
  distortion, 266
  electret, 264
  electrostatic, 263
  frequency response, 265–266
  function, 261
  impedence, 266
  in-ear, 264
  intensity, 265
  mechanical design, 264–265
  moving coil, 262–263
  noise-cancelling, 264
  open-back, 264
  potential problems, 266
  supra-aural, 264
  values, 265–266
  variants, 262–265
headset, 264
hearing aids, 261
heat sensitivity, 14
heat sink, 28, 206, 222
Hertz, Heinrich Rudolph, 262
hexadecimal counter, 134
high linearity optocouplers, 36
high logic state, 89
high speed optocoupler, 36

high-brightness LED, 205, 217, 227
high-intensity LED, 217
high-output LED, 217
high-pass filter, 55
high-power LEDs, 217
Hold state, 109, 114
holding current, 2, 18
holding level, 12
hysteresis, 42, 44–45, 48, 83, 185
hysteresis range, 43
hysteresis zone, 39

I

I2C protocol, 62, 63
IEM (see in-ear monitor)
ignition voltage, 184
illuminance, 178
impedence, 271
in-ear earphones, 264
in-ear headphones, 264
in-ear monitor, 264
incandescence, 191, 243
incandescent bulbs, 219
incandescent lamp, 11, 15, 24, 171–181, 200, 223
  advantages/disadvantages, 179–180
  base variants, 177
  derating, 180
  design, 172–175
  efficacy, 179
  efficiency, 179
  function, 171–172
  halogen or quartz halogen, 176
  illuminance, 178
  intensity, 178
  mean spherical candlepower (MSCP), 178
  miniature lamps, 175–176
  non-incandescent sources, 174
  oven lamp, 176
  panel-mount indicator lamp, 176
  potential problems, 180
  power, 177–178
  power consumption, 175
  schematic symbols, 171
  spectrum, 173
  use of, 179–180
  values, 177–179
  variants, 175–177
incandescent lamps, 206, 217
incoherent light, 159
increment/decrement protocol, 62
Indiglo electroluminscent displays, 244
inductor, 193
infrared emitters, 211
inhibit pin, 153
input
  analog, 66
  asynchronous, 123
  clock, 131, 133
  confused, 48, 58
  dual, 50–51, 127
  floating, 127
  inverting/noninverting, 39, 51
  parallel, 123, 134
input bias current, 53
input devices, decoder, 145
input differential voltage range, 53
input offset voltage, 43, 53
input voltage, 155
instant-start ballast, 194
integrated circuits (see analog integrated circuits, digital integrated circuits)
inter-integrated circuit, 62
internal sensors, 35
internal series resistor, 210
intrinsic layer, optocoupler, 35
invalid number, 135
inversion, logic gates, 90
inverter, 36, 89, 91, 114
inverting input, 39, 51
ionization, 184
isolation transformers, 34

J

jam loaded, 134
jam-type flip-flop, 108
jam-type parallel data input, 124
jitter, 44
JK flip-flop, 112–113, 116

K

keyboard, polling, 127

L

ladder, 60
lamp
  oven, 176
  panel-mount indicator lamp, 176
lamp lenses, 176
laser, 197–203
  CO2 lasers, 201
  coherent light, 200–201
  common applications, 202
  crystal lasers, 201
  design, 198–201
  fiber lasers, 201
  function, 197
  laser diode, 197, 198–199
  potential problems, 202
  use of, 202
  values, 201
  variants, 201
laser diode, 197, 198–199
latch, 107, 112
latch function, 43
latched comparator, 47
latching current, 2, 18
latching relay, 4
LCD (liquid-crystal display), 159–170, 207, 227, 246
  4543B decoder, 168
  active and passive types, 161
  additional segments, 162–163
  alphanumeric display module, 168–169
  backlighting, 167

  - color, 166–167
  - color super twisted nematic (CSTN), 161
  - crystal types, 161
  - design, 159–160
  - dot-matrix displays, 163–165
  - double super twisted nematic, 161
  - film-compensated super twisted nematic (FSTN), 161
  - function, 159
  - LED comparisons, 228
  - numeric display module, 167–168
  - potential problems, 169–170
  - reflective, 161
  - seven-segment displays, 161–162
  - super twisted nematic (STN), 161
  - transreflective, 161
  - twisted nematic (TN), 161
  - use of, 167–169
  - variants, 160–167
  - zero-power displays, 167
- LCD display (see LCD (liquid crystal display))
- leakage
  - capacitors, 77
  - diac, 12
  - SCR, 2
  - solid-state relay (SSR), 26
  - triac, 17
- leakage current, 156
- LED (generic), 26, 159, 167, 174, 227, 243
- LED area lighting, 24, 191, 195, 217–226, 227
  - color variation, 222
  - comparisons, 220–222, 223–224
  - costs and efficiency, 218
  - design, 218–223
  - dimming, 222
  - efficacy, 222
  - function, 217–218
  - heat dissipation, 222
  - high-brightness, 217
  - high-output, 217
  - high-power, 217
  - potential problems, 225
  - schematic symbol, 218
  - ultraviolet output, 222
  - values, 225
  - variants, 223–224
  - visible differences, 220–220
- LED display, 146, 227–237
  - LM3914, 236
  - MAX6954, 234
  - SSD1306, 235
  - SSD1331, 235
  - TIL311, 236
  - WS0010, 235
  - additional segments, 229–230
  - design, 228
  - dot-matrix displays, 230, 234–236
  - driver chips and multiplexing, 233–234
  - function, 227–228
  - LCD comparisons, 228
  - multiple bar display, 232, 236
  - multiple numerals, 229
  - one-digit hexadecimal dot matrix, 236
  - pixel arrays, 231, 235–236
  - potential problems, 237
  - seven segment numeral, 228
  - seven-segment displays, 228–229, 232–233
  - single light bar, 232
  - sixteen-segment driver chip, 234
  - use of, 232–236
  - values, 232
  - variants, 228–232
- LED indicator, 205–215, 218, 227
  - color rendering index (CRI), 212
  - common usage, 206
  - comparisons to other light emitters, 214
  - design, 207
  - diffusion, 209
  - efficacy, 208
  - forward current, 211
  - forward voltage, 212
  - function, 205–207
  - high-brightness, 205
  - infrared, 211
  - intensity, 208
  - internal resistor, 210
  - LEDs in parallel, 214
  - life expectancy, 212
  - light output and heat, 212
  - low-current LEDs, 211
  - multicolor LEDs and color mixing, 207
  - multicolored, 210
  - multiple series LEDs, 214
  - other applications, 215
  - polarity, 213
  - potential problems, 215
  - schematic symbols, 206
  - series resistor value, 214
  - size and shape, 208
  - ultraviolet, 211
  - use of, 213–215
  - values, 211–213
  - variants, 208–211
  - view angle, 213
  - wavelength and color temperature, 209–210
- lens, 183
- level shifter, 46
- light
  - coherent, 197, 200–201
  - incoherent, 159
  - monochromatic, 197
- light wires, 243
- light-emitting capacitor, 244
- light-emitting diode (see LED)
- linear relationship, 50
- linear taper, 62
- linearity versus saturation, 42
- liquid-crystal display (see LCD)
- logarithmic taper, 62
- logic chip, 139
- logic gate, 89–106, 108, 126, 140
  - 74HC00, 93
  - 4000 series, 94
  - 7400 series, 93
  - binary arithmetical operations, 91
  - Boolean notation, 91
  - chip families, 95–96
  - design, 89

  function, 89
  gates and inputs, 90, 96–103
  inversion, 90
  part numbers, 94–95
  potential problems, 105–106
  uses, 103
  variants, 93
logic state, 89, 107
logic, positive, 90
logic, transistor-transistor, 70, 93
logic-output optocoupler, 36
loudspeaker, 84, 249, 263, 267
low logic state, 89
low-current LEDs, 211
low-pass filter, 54
Lower State Transition Voltage (LSTV), 43
lumen (lm), 177, 195, 218, 225
lumens per watt, 179, 195
luminaire, 217
luminary, 217
luminescence, 191, 243
luminous flux, 177, 195, 208, 218
lux, 178

M

magnetic ballast, 225
maintaining voltage, 185
maser, 197
master-slave flip-flop, 113, 116
mc (millicandela), 178
mean spherical candlepower (MSCP), 176, 178
metal-oxide semiconductor field-effect transistor (see MOSFET)
metastability, in flip-flops, 119
microcontroller, 29, 59
microphone, 262
millicandelas, 187, 208
miniature speakers, 270
mismatching, 266
mixed-signal device, 59
MOD, 133
mode select pins, 125
modulo/modulus, 132
modulus-16 counter, 134
monochromatic light, 197
monostable multivibrator, 69
  (see also timer)
monostable timer, 69, 71, 77, 79
MOSFET (metal-oxide semiconductor field-effect transistor), 26, 70
motion detector, passive infrared, 211
motor, AC, 11
moving coil headphones, 262
MSCP (see mean spherical candlepower)
multicolor LED, 207
multiple bar display, 232, 236
multiple series LED, 214
multiple-stage counter, 136
multiplexer, 28, 140, 146, 151–157
  analog, 151, 152
  bidirectional, 151
  demultiplexer, 140, 146, 151, 152, 157
  design, 153
  differential, 152
  digital versus analog, 151
  function, 151
  pin identifiers, 154–155
  potential problems, 157
  schematic symbol, 153
  similar devices, 152–153
  use of, 156
  values, 155
  variants, 155
multiplexing, 168, 233–234, 241
multisource lighting
  electroluminescent (EL) display, 243–247
  LED display, 227–237
  vacuum-fluorescent display (VFD), 239–242
multivibrator, 69
  (see also timer)
  astable, 46, 55, 69
  bistable, 46, 69, 107
MUX (see multiplexer)

N

NAND gate, 37, 90, 91, 93, 94, 95, 97, 98, 100, 102, 103, 108, 109, 123, 127, 135
  (see also NAND-based SR flip-flop)
NAND-based SR flip-flop, 108–109, 116
nanometer, 195, 209
negative feedback, 49, 51–52
negative resistance, 185
neon bulb, 31, 183–190, 191, 193
  construction, 184
  efficiency, 187
  function, 183
  how it works, 183–186
  ionization, 184
  life expectancy, 188
  light output, 187
  negative resistance, 185–186
  Nixie tubes, 189
  potential problems, 189
  power-supply testing, 188
  ruggedness, 187
  use of, 186–189
  variants, 189
neon bulbs, 206
neon indicator, 183
neon lamp, 183
neon lamp assembly, 183
neutral value, 49
  (see also zero value)
Newtons per square meter (Pa), 251
Nixie tubes, 189
noise-cancelling headphones, 264
noise-creating device, 249, 267
noninverting input, 39, 51
nonvolatile memory, 61
NOR gate, 90, 91, 97, 98, 100, 102, 103, 108, 109
  (see also NOR-based SR flip-flop)
NOR-based SR flip-flop, 109–110
numeric display module, 167–168

NV (see nonvolatile memory)

O

octal counter, 134
OE (output-enable), 126
offset null adjustment, 56
OLED (organic light-emitting diode), 217, 227, 243, 246–247
on-resistance, 156
one-digit hexadecimal dot matrix, 236
op-amp, 39, 49–58, 90
  LM741, 49, 53
  calculating amplification, 53
  and comparators, 52
  confused input, 58
  controlling the gain, 53
  design, 49–52
  differences from comparator, 42
  dual inputs, 50–51
  as high-pass filter, 55
  as low-pass filter, 54
  negative feedback, 51–52
  offset null adjustment, 56
  oscillating output, 57
  potential problems, 57–58
  as relaxation oscillator, 55
  as single power source, 56
  use of, 53–57
  values, 52
  variants, 52
  what it does, 49
open collector, 40
open loop mode, 51
open loop operation, 42
open loop voltage gain, 53
open-back headphones, 264
open-drain outputs, 126
operating voltage, 259
operational amplifier (see op-amp)
optical maser, 197
optical switch, 34
optocoupler, 25, 33–37, 215
  D804, 37
  OPTEKD804, 34
  analog output, 36
  basic types, 36
  bidirectional, 36
  design, 34–34
  function, 33–34
  high speed, 36
  high-linearity, 36
  internal sensors, 35–36
  logic-output, 36
  potential problems, 37
  schematic symbols, 34
  use of, 36–37
  values, 36
  variants, 34–36
OR gate, 90, 91, 97, 98, 99, 100, 102, 103, 140, 142
organic LED (see OLED (organic light-emitting diode))
oscillating output, 47, 57
oscillator, relaxation, 12, 46, 55
output
  analog, 66
  clipped, 50
  descending, 136
  encoded, 134
  erratic, 48
  open-drain, 126
  oscillating, 47, 57
  parallel, 123
  push-pull, 42
  three-state, 142
  types of, 134
  ultraviolet, 222
  weighted, 134
output coupler, 198
output mode, comparator, 42
oven lamp, 176
overvoltage protection, 2, 6, 155

P

Pa (Pascals), 251
panel, 243
Panelescent electroluminescent lighting, 244
parallel input, 134
parallel-enable pin, 134
parallel-in, parallel-out (PIPO) shift converters, 125
parallel-in, serial-out (PISO) shift register, 124
parallel-serial converters, 124
parchmentized thread, 172
passive infrared motion detector, 211
passive matrix LCD, 161
phase angle, 7, 13
phase control, 2, 7–7, 13, 22
phones (see headphone)
phosphor, 192, 239, 244
phosphorescence, 244
photocell, 35
photodarlington, 34
photodiode, 26, 34
photon, 207
photoresistor, 34
phototransistor, 26, 34, 45
photovoltaic PIN diode, 35
piezein, 250
piezo, 258
piezoelectric alerts (see audio indicator)
piezoelectric transducer, 249, 250, 253
pin
  common, 161
  floating, 105
  inhibit, 153
  parallel-enable, 134
  reset, 70
  trigger, 69
pin base, 177
PIN diode, 35
pin identifier, 133
pip, 184
PIPO (see parallel-in, parallel-out (PIPO) shift converters)
pixel arrays, 231, 235–236
PMOLED (passive-matrix OLED), 246
PN junction, 207
PNPN device, 3
polling a keyboard, 127
positive feedback, 44
positive logic, 90

potentiometer, 35
  (see also digital potentiometer)
power rating, 272
powered speaker, 271
primary colors, 166–166
priority encoder, 140, 141
programmable counter, 137
programmable digital potentiometer, 59
programmable unijunction transistor (PUT), 2, 9
propagation delay, 44, 134
protection diode, 30
pullup resistor, 36, 40, 48, 71, 126, 153
pulse-width modulation, 7, 194, 211
pumping a laser, 198
push-pull output, 42
pushbutton, 66, 71
pushbutton protocol, 62

Q

quad digital potentiometer, 61
quadrants, 17–18
quartz crystal, 131, 135

R

race condition, 112
radiant flux, 177
radiant luminous efficacy (LER), 179, 187, 208, 222
radiant luminous efficiency (LFR), 179
radiation, black body, 173
rail-to-rail values, 42
rapid-start ballast, 194
rated voltage, 259
Rayleigh scattering, 202
RC network, 70
rectifier, 2
reference voltage, 70
reflective display (see LCD (liquid crystal display))
reflective LCD, 161
reflective primaries, 166
reflex port, 270
regenerative device, 2
register, 121
relaxation oscillator, 12, 46, 55
relay
  electromagnetic, 33
  latching, 4
  solid-state (see solid-state relay)
  zero-crossing, 25, 27
remainder, 132
reproducers
  headphone, 261–266
  speaker, 267–272
reset pin, 70
reset state, 109
resistor array, 232
resolution, 43
resonant cavity, 198
response time, comparator, 42
restricted combination, 112
retriggering, timer, 69
RGB LCD monitors, 167
RGB primaries, 166
rheostat mode, 64
ring counter, 121, 131, 134
ringing, 112
ripple counter, 107, 134, 148
rising voltage, 43
rising-edge triggered shift register, 123
rope light, 243, 245
rotary encoder, 139
rotary switch, 151
rotational coder, 139
rotational encoder, 66

S

saturation versus linearity, 42
SCR (silicon-controlled rectifier), 1–9, 11, 15, 26, 34
  2N682, 2
  40TPS12, 2
  X0403DF, 2, 5
  abbreviations, 6
  AC current applications, 5
  breakdown and breakover voltage, 4
  concept demo, 4
  design, 1–2
  function, 1
  internal configuration, 3–5
  overvoltage protection, 7
  phase control, 7–7
  potential problems, 8–9
  switching behavior, 2
  use, 6–8
  values, 5
  variants, 5, 6
screw-in lamps, 177
SCS (see silicon-controlled switch)
segments, 239
self-drive transducer circuit, 254
sensitivity, 272
sensor, 34, 35, 249, 262
serial peripheral interface (see SPI)
serial-in, parallel-out (SIPO) shift register, 123, 124
serial-in, serial-out (SISO) shift register, 123, 124
serial-parallel coverter, 123
serial-to-parallel converters, 107
series resistor, 207
series resistor value, 214
settling time, 149, 157
setup time, 125
seven-segment decoder, 146
seven-segment displays, 228–229, 232–233
seven-segment numeral, 228
shift register, 63, 121–129, 131, 146
  abbreviations and acronyms, 123
  arithmetical operations, 127
  buffering, 128
  design, 122
  function, 121–122
  parallel outputs and inputs, 123–124
  polling the keyboard, 127
  potential problems, 128–129

  power considerations, 126
  preloading, 127
  schematic representation, 122
  three-state output, 126
  use of, 126–128
  values, 125–126
  variants, 124–125
shift registers, 107
shortened modulus, 135
sidac, 14
silicon-controlled rectifier (see SCR)
silicon-controlled switch (SCS), 1, 11, 15
simple encoder, 141
single counter, 136
single light bar, 232
single power source, 56
single-inline package (SIP), 27
single-input gates, 90
single-source lighting
  fluorescent light, 191–196
  incandescent lamp, 171–181
  lasers, 197–203
  LED area lighting, 217–226
  LED indicator, 205–215
  neon bulb, 183–190
sink current, 44
SIPO (see serial-in, parallel-out (SIPO) shift register)
siren, 258
SISO (see serial-in, serial-out (SISO) shift register)
sixteen-segment driver chip, 234
slew rate at unity gain, 53
snubber, 30
snubber network, 23
snubberless triac, 23
solid-state analog switch, 28
solid-state relay (SSR), 1, 11, 15, 25–31, 33, 36, 215
  (see also optocoupler)
  CMX60D10, 27
  DC60S7, 27, 28
  advantages/disadvantages, 25
  design, 26–27
  function, 25–26
  instantaneous versus zero crossing, 27
  NC and NO modes, 28
  packaging, 28
  potential problems, 30
  solid-state analog switch, 28
  use of, 29–30
  values, 29
  variants, 27–28
sound pressure level, 251–252, 251
sound reproduction device, 249, 267
sound sources (see audio alerts, reproducers)
spatial coherence, 200
spatial distribution, 213
speaker, 249, 267–272
  closed-box, 270
  construction, 268
  design, 267–270
  electrostatic, 271
  function, 267
  innovative designs, 271
  miniature speakers, 270
  multiple drivers, 269–270
  potential problems, 272
  powered, 271
  resonance, 270
  values, 271
  variants, 270–271
  venting, 270
  wireless, 271
spectral lines, 174
SPI (serial peripheral interface), 62
spontaneous emission, 198
SPST switch, 25
sputtering, 188
SR flip-flop (see NAND-based SR flip-flop)
sRGB, 167
SSR (see solid-state relay)
standard LED, 205, 227
starter, 192, 225
starting voltage, 184
state diagram, 135
steradians, 178
stimulated emission, 198
striking voltage, 184
strip lights, 223
strobe, 153
super twisted nematic (STN) LCD, 161
supra-aural headphones, 264
surface-mount transducers, 251
swapped voltages, 48
switch, 25, 71, 107
  rotary, 151
  silicon-controlled (SCS), 1, 11, 15
  solid-state analog switch, 28
  SPST, 25
switch bounce, 109
synchronous circuit, 113
synchronous counter, 134, 136
synchronous flip-flop, 107

T

T-type flip-flop, 116
tap, 60
tap address, 62, 63
taper, 62
task lighting, 217
TFT (see thin-film transistor)
THD (see total harmonic distortion)
thermal compound, 30
thick phosphor, 244
thin-film OLED, 243
thin-film transistor (TFT), 161, 246
3-wire programmable potentiometers, 63
three-state output, 126, 142
three-wire chips, 64
threshold voltage, 207
through hole, 206
thyratron, 1
thyristor
  diac, 11–14, 15
  SCR, 1–9
  triac, 15–24
timer, 69–87
  4047B, 75
  555, 70–85
  5555, 74
  556, 73
  558, 73

    - 74HC123, 75
    - 74HC221, 76
    - 74HC5555, 74
    - 74HCT5555, 74
    - 7555, 74
    - 7556, 74
    - CD4047B, 75
    - HEF4528B, 75
    - M74HC4538, 75
    - MC14538B, 75
    - NTE926, 73
    - astable mode, 70, 72, 77, 80
    - bipolar version, 70
    - burst mode, 84
    - CMOS 555, 74
    - CMOS version, 71
    - and coupling capacitors, 84
    - design, 70
    - dual monostable, 75–76, 77
    - function, 69–70
    - game sound, 85
    - hysteresis, 83
    - loudspeaker connection, 84
    - monostable mode, 69, 71, 77, 79
    - output time control, 80
    - potential problems, 85–87
    - TTL version, 70
    - use of, 79–85
    - values, 76–79
    - variants, 70–76
- toggle, 43, 113
- total harmonic distortion (THD), 266
- transducer, 249–255, 257, 267
    - PT-2040PQ, 253
    - AC supply, 254
    - current, 254
    - design, 249
    - dynamic, 262
    - electromagnetic, 250, 253
    - formats, 251
    - frequency range, 251
    - function, 249
    - limitations, 253–254
    - measurement location, 253
    - piezoelectric, 249, 250, 253
    - potential problems, 254
    - self-drive transducer circuit, 254
    - sound intensity, 254
    - sound pressure, 251–252
    - surface-mount, 251
    - ultrasonic, 250
    - unweighted values, 253
    - use of, 254
    - values, 251–254
    - variants, 250–251
    - voltage, 254
    - volume control, 254
    - weighted sounds values, 252–253
- transformer, isolation, 34
- transistor, bipolar, 1
- transistor-transistor logic, 93
- transistor-transistor logic protocol, 70
- transparent, 108, 114
- transreflective LCD, 161
- tri-state encoder, 142
- triac, 1, 7, 11, 15–24, 25, 34
    - BTA208X-1000B, 16
    - BTB04-600SL, 16
    - MAC97A6, 16
    - breakover voltage, 20
    - charge storage, 23
    - design, 17–23
    - function, 15–16
    - other drivers, 22
    - potential problems, 24
    - quadrants, 17–18
    - snubberless, 23
    - switching AC, 20–22
    - symbol variants, 15–16
    - testing, 19–20
    - triggered by diac, 22
    - values, 23–24
    - variants, 23
- triac-based dimmer, 179
- trigger pin, 69
- trimmer, 40, 254
- truth table, 90
- TTL (see transistor-transistor logic)
- tweeter, 269
- twisted nematic (TN) LCD, 161
- two-wire chips, 64

U

- ultrasonic transducer, 250
- ultraviolet output, 222
- unity gain frequency, 53
- universal shift register, 125
- unweighted sound values, 253
- up/down protocol, 62, 63
- Upper State Transition Voltage (USTV), 43

V

- Vactrol, 35
- vacuum-fluorescent display, (VFD), 239–242
    - anode in, 240
    - cathode, 240
    - character sets and pictorial design, 242
    - color, 241
    - comparisons, 242
    - design, 239–240
    - function, 239
    - modern application, 241
    - potential problems, 242
    - use of, 240–241
    - variants, 241–242
- variable frequency drive, 239
- velocity, 262
- vent, 270
- VFD (see vacuum-fluorescent display)
- view angle, 208
- volatile memory, 61
- voltage, 259
    - breakdown, 1, 4, 20, 184
    - breakover, 4, 15, 20
    - dual, 39
    - falling, 43
    - forward, 212
    - ignition, 184
    - input, 155
    - input offset, 43, 53
    - maintaining, 185
    - in multiplexers, 155
    - operating, 259

- rated, 259
- reference, 70
- rising, 43
- starting/striking, 184
- threshold, 207
- transducer, 254

voltage amplification, 53
voltage amplifier, 49
voltage converter, 39
voltage divider, 71
voltage divider mode, 65
voltage gain, 43
voltage level, 69
voltage regulator, 59
voltage spike, 72
voltage transition, 69

W

wafers, 219
water clear, 205
wavelength, 262
wavelength coherence, 200
wedge base, 177
weighted output, 134
weighted sound values, 252–253
window comparator, 42, 46
wiper, 60
wiper resistance, 65
wireless speaker, 271
woofer, 269

X

XNOR gate, 90, 91, 98, 100, 102, 103
XOR gate, 90, 91, 92, 97, 98, 100, 102, 103

Z

zero point finder, 47
zero value, 49
zero-crossing relay, 25, 27
zero-crossing signal, 66

About the Authors

Charles Platt is a contributing editor and regular columnist for *Make:* magazine, where he writes about electronics. He is the author of the highly successful introductory hands-on book, *Make: Electronics*, and its sequel, *Make: More Electronics*. His science fiction novels are currently being reissued by Stairway Press.

Platt was a Senior Writer for *Wired* magazine and has written various computer books. As a prototype designer, he created semi-automated rapid cooling devices with medical applications, and air-deployable equipment for first responders. He was the sole author of four mathematical-graphics software packages, and has been fascinated by electronics since he put together a telephone answering machine from a tape recorder and military-surplus relays at age 15. He lives in a Northern Arizona wilderness area, where he has his own workshop for prototype fabrication and projects that he writes about for *Make:* magazine.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in the Netherlands, where he works on swarm robotics and simulates sea animals in the computational science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN. He also fact-checked Charles Platt's previous book, *Make: More Electronics*.

Colophon

The cover and body font is Myriad Pro, the heading font is Benton Sans, and the code font is Ubuntu Mono.

Technology & Engineering / Sensors

VOL. 3

Encyclopedia of Electronic Components

Sensors

Want to know how to use an electronic component? This third book of a three-volume set includes key information on electronic sensors for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** **the first and only multivolume encyclopedia of electronic components, distilled into three separate volumes**
- **Instructive:** **each component description provides details about substitutions, common problems, and workarounds**
- **Incredibly detailed:** **includes information distilled from hundreds of sources**
- **Easy to browse:** **parts are clearly organized by component type**
- **Authoritative:** **fact-checked by experts to ensure current and accurate information**
- **Reliable:** **a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials**
- **Comprehensive:** **Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers all major types of sensing components**

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make:* magazine, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN.

US $29.99 CAN $34.99
ISBN: 978-1-4493-3431-4

Make:
makezine.com

3
Make: Encyclopedia of Electronic Components
Platt & Jansson

Make:

VOL. 3

Charles Platt and Fredrik Jansson

Encyclopedia of Electronic Components

Sensors

Location ▪ Presence ▪ Proximity ▪ Orientation ▪ Oscillation ▪ Force ▪ Load ▪ Human Input ▪ Liquid and Gas Properties ▪ Light ▪ Heat ▪ Sound ▪ Electricity

Technology & Engineering / Sensors

VOL. 3

Encyclopedia of Electronic Components

Sensors

Want to know how to use an electronic component? This third book of a three-volume set includes key information on electronic sensors for your projects—complete with photographs, schematics, and diagrams. You'll learn what each one does, how it works, why it's useful, and what variants exist. No matter how much you know about electronics, you'll find fascinating details you've never come across before.

Convenient, concise, well-organized, and precise

Perfect for teachers, hobbyists, engineers, and students of all ages, this reference puts reliable, fact-checked information right at your fingertips. Beginners will quickly grasp important concepts, and more experienced users will find the specific details their projects require.

- **Unique:** the first and only multivolume encyclopedia of electronic components, distilled into three separate volumes
- **Instructive:** each component description provides details about substitutions, common problems, and workarounds
- **Incredibly detailed:** includes information distilled from hundreds of sources
- **Easy to browse:** parts are clearly organized by component type
- **Authoritative:** fact-checked by experts to ensure current and accurate information
- **Reliable:** a more consistent source of information than online sources, product datasheets, and manufacturer's tutorials
- **Comprehensive:** Volume 1 covers power, electromagnetism, and discrete semiconductors; Volume 2 includes integrated circuits, and light and sound sources; Volume 3 covers all major types of sensing components

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make:* magazine, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN.

US $29.99 CAN $34.99
ISBN: 978-1-4493-3431-4

Make:
makezine.com

3 Make: Encyclopedia of Electronic Components Platt & Jansson

Make:

VOL. 3

Charles Platt and Fredrik Jansson

Encyclopedia of Electronic Components

Sensors

Location ▪ Presence ▪ Proximity ▪ Orientation ▪ Oscillation ▪ Force ▪ Load ▪ Human Input ▪ Liquid and Gas Properties ▪ Light ▪ Heat ▪ Sound ▪ Electricity

Encyclopedia of Electronic Components Volume 3

Charles Platt and Fredrik Jansson

MAKER MEDIA™
SAN FRANCISCO, CA

Encyclopedia of Electronic Components, Volume 3

by Charles Platt

Copyright © 2016 Charles Platt. All rights reserved.

Printed in Canada.

Published by Maker Media, Inc., 1160 Battery Street East, Suite 125, San Francisco, CA 94111.

Maker Media books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact O'Reilly Media's institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Brian Jepson
Production Editor: Melanie Yarbrough
Copyeditor: Christina Edwards
Proofreader: Charles Roumeliotis
Indexer: Charles Platt
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Charles Platt

April 2016: First Edition

Revision History for the First Edition

2016-04-05: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449334314* for release details.

Make:, Maker Shed, and Maker Faire are registered trademarks of Maker Media, Inc. The Maker Media logo is a trademark of Maker Media, Inc. *Encyclopedia of Electronic Components, Volume 3* and related trade dress are trademarks of Maker Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Maker Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

978-1-449-33431-4

[TI]

To Brian Jepson

Table of Contents

Preface .... **xix**

1. **GPS** **1**
   - What It Does 1
     - Schematic Symbol 1
     - GPS Segments 1
   - How It Works 2
   - Variants 2
   - Values 3
   - How to Use It 3
     - Pulse per Second Output 4
   - What Can Go Wrong 4
     - Electrostatic Discharge 4
     - Failure to Ground Properly 4
     - Cold Joints 4
     - Restricted Availability 4
     - Inability to Detect Satellites 4
     - Exceeding Maximum Velocity or Altitude 4

2. **magnetometer** **5**
   - What It Does 5
     - Schematic Symbol 5
     - IMU 5
     - Applications 6
   - How It Works 6
     - Magnetic Fields 6
     - Earth’s Axes 7
     - Coil Magnetometer 8
     - Hall Effect and Magnetoresistance 8
   - Variants 9

How to Use It 9
What Can Go Wrong 9
  Bias 9
  Mounting Errors 10

3. object presence sensor 11

What It Does 11
  Schematic Symbol 12
Variants 12
Optical Detection 12
  Transmissive Optical Sensors 13
  Retroreflective Optical Sensors 15
Magnetic Sensors 16
Reed Switch 17
  Reed Switch Variants 17
  Reed Switch Values 18
  How to Use a Reed Switch 18
Hall-Effect Sensor 18
  How a Hall-Effect Sensor Works 19
  Hall-Effect Sensor Variants 19
  Other Applications 20
Values 20
How to Use a Hall-Effect Sensor 20
Configuration of Object Presence Sensors 21
  Linear Motion 21
  Sensing by Interruption 21
  Angular Motion 21
Sensor Comparisons 22
  Advantages of Optical Presence Sensors 22
  Disadvantages of Optical Presence Sensors 22
  Advantages of a Reed Switch 22
  Disadvantages of a Reed Switch 22
  Advantages of a Hall Effect Sensor 22
What Can Go Wrong 23
  Optical Sensor Issues 23
  Reed Switch Issues 23

4. passive infrared sensor 25

What It Does 25
  Schematic Symbols 25
  Applications 25
How It Works 26
  Pyroelectric Detector 26
  Elements 26
  Lenses 27
Variants 29

What Can Go Wrong . . . 30
  Temperature Sensitivity . . . 30
  Detector Window Vulnerability . . . 30
  Moisture Vulnerability . . . 30

5. **proximity sensor . . . 31**
What It Does . . . 31
  Schematic Symbols . . . 31
  Applications . . . 32
Variants . . . 32
  Ultrasound . . . 32
  Infrared . . . 32
  Relative Advantages . . . 33
Ultrasonic Examples . . . 33
  Imports . . . 33
  Individual Elements . . . 34
Infrared Examples . . . 34
  Trends in Infrared Proximity Sensing . . . 35
Capacitive Displacement Sensor . . . 36
  Applications . . . 37
How It Works . . . 37
  Sources of Error . . . 37
  Values . . . 38
What Can Go Wrong with Optical and Ultrasound Proximity Sensors . . . 38
  Object Too Close . . . 38
  Multiple Signals . . . 38
  Inappropriate Surfaces . . . 38
  Environmental Factors . . . 38
  Deterioration of LEDs . . . 38

6. **linear position sensor . . . 39**
What It Does . . . 39
  Applications . . . 39
  Schematic Symbol . . . 39
How It Works . . . 39
  Linear Potentiometer . . . 40
  Magnetic Linear Encoders . . . 41
  Optical Linear Encoders . . . 41
  Linear Encoder Applications . . . 42
  Linear Variable Differential Transformers . . . 42
What Can Go Wrong . . . 43
  Mechanical Issues . . . 43
  LED Longevity . . . 43

7. **rotary position sensor . . . 45**
What It Does . . . 45

    - Applications 45
    - Schematic Symbol 46
- Potentiometers 46
    - Arc-Segment Rotary Potentiometer 46
    - End Stops 46
    - Multiturn Rotary Potentiometer 46
    - Magnetic Rotary Position Sensor 47
    - Rotary Position Sensing Chips 48
    - Rotary Encoders 48
    - Optical Rotary Encoders 48
    - Optical Products 49
    - Computer Mouse Principles 50
    - Rotational Speed 50
    - Absolute Position 51
    - The Gray Code 51
    - Magnetic Rotary Encoders 52
- How to Use It 53
- What Can Go Wrong 53
    - Wiring Errors 53
    - Coding Errors 53
    - Ambiguous Terminology 54

8. tilt sensor 55
- What It Does 55
    - Schematic Symbol 56
- How It Works 56
    - Simplified Version 57
    - Applications 57
- Variants 58
    - Mercury Switches 58
    - Pendulum Switch 58
    - Magnetization 59
- Tilt Sensors 59
    - Two-Axis Tilt Sensors 59
- Values 60
- How to Use It 61
- What Can Go Wrong 61
    - Contact Erosion 61
    - Random Signals 61
    - Environmental Hazard 61
    - Requirement for Gravity 61
    - Requirement for Stability 62

9. gyroscope 63
- What It Does 63
    - Schematic Symbol 63

  - IMU 63
  - Applications 63
- How It Works 64
  - Vibrating Gyroscope 64
- Variants 66
  - IMUs 66
- Values 67
- How to Use It 67
- What Can Go Wrong 67
  - Temperature Drift 67
  - Mechanical Stress 68
  - Vibration 68
  - Placement 68

10. accelerometer 69
- What It Does 69
  - IMU 69
  - Schematic Symbols 70
  - Applications 70
- How It Works 70
  - Gravity and Free Fall 71
  - Rotation 71
  - Calculation 72
- Variants 72
- Values 73
- What Can Go Wrong 74
  - Mechanical Stress 74
  - Other Problems 74

11. vibration sensor 75
- What It Does 75
  - Schematic Symbols 75
- Variants 75
  - Pin-and-Spring 76
  - Piezoelectric Strip 76
  - Chip-Based Piezoelectric 77
  - “Mousetrap” Type 77
  - Magnetic 77
  - Mercury 78
- Values 78
  - Primary Variables 78
  - Dynamic Attributes 79
- How to Use It 79
- What Can Go Wrong 79
  - Long Cable Runs 79
  - Interference 79

- Correct Grounding 80
- Fatigue Failure 80

12. force sensor 81

- What It Does 81
  - Applications 81
  - Schematic Symbol 82
- How It Works 82
  - Strain Gauge 82
  - Wheatstone Bridge Circuits 83
  - Wheatstone Bridge Errors 84
  - Strain-Gauge Amplification 84
  - Other Strain-Gauge Modules 85
  - Plastic-Film Force Sensors 85
- How to Use It 86
- Values 87
  - Film-Based Force Sensors for User Input 87
  - Specifications for Film-Based Force Sensors 87
  - Strain Gauges 88
- What Can Go Wrong 88
  - Soldering Damage 88
  - Bad Load Distribution 88
  - Water Damage 88
  - Temperature Sensitivity 88
  - Leads Too Long 88

13. single touch sensor 89

- What It Does 89
  - Applications 90
  - Schematic Symbols 90
- How It Works 90
- How to Use It 91
  - Obtaining Touch Pads 91
  - Individual Touch Pad 91
  - Wheels and Strips 92
  - Design Considerations 92
- What Can Go Wrong 93
  - Insensitive to Gloves 93
  - Stylus Issues 93
  - Conductive Ink 93

14. touch screen 95

- What It Does 95
  - Schematic Symbol 95
- Variants 95
  - Resistive Sensing 95

Capacitive Sensing 96
Screens Available as Components 97

15. liquid level sensor 99
What It Does 99
Schematic Symbols 99
Applications 99
How It Works 100
Binary-Output Float Sensor 100
Analog-Output Float Sensor 101
Incremental-Output Float Sensor 101
Displacement Level Sensors 102
Ultrasonic Level Sensors 102
Reservoir Weight 103
Pressure Sensing 103
What Can Go Wrong 104
Turbulence 104
Tilting 104

16. liquid flow rate sensor 105
What It Does 105
Schematic Symbols 105
Paddlewheel Liquid Flow Rate Sensors 105
Turbine Flow Rate Sensors 106
Limitations of Paddlewheels and Turbines 107
Thermal Mass Liquid Flow Rate Sensor 107
Sliding Sleeve Liquid Flow Switch 108
Sliding Plunger Liquid Flow Switch 108
Ultrasonic Liquid Flow Rate Sensor 108
Magnetic Liquid Flow Sensor 108
Differential Pressure Liquid Flow Meter 109
What Can Go Wrong 109
Vulnerability to Dirt and Corrosive Materials 109

17. gas/liquid pressure sensor 111
What It Does 111
Schematic Symbols 111
Applications 111
Design Considerations 112
Units 112
How It Works 112
Basic Sensing Elements 112
Relative Measurement 113
Variants 114
Ambient Air Pressure 114
Altitude 114

  - Gas Pressure 115
- What Can Go Wrong 116
  - Vulnerability to Dirt, Moisture, and Corrosive Materials 116
  - Light Sensitivity 116

18. gas concentration sensor 117
- What It Does 117
  - Schematic Symbol 117
- Semiconductor Gas Sensors 117
- Oxygen Sensors 119
- Humidity Sensors 119
  - Dew-Point Sensor 120
  - Absolute Humidity Sensors 120
  - Relative Humidity Sensors 120
  - Humidity Sensor Output 121
  - Analog Humidity Sensor 121
  - Design Considerations 122
  - Digital Humidity Sensor 122
- What Can Go Wrong 122
  - Contamination 122
  - Recalibration 123
  - Soldering 123

19. gas flow rate sensor 125
- What It Does 125
  - Applications 125
  - Schematic Symbol 125
- How It Works 126
  - Anemometer 126
  - Mass Flow Rate Sensing 127
  - Applications 128
  - Units 128
  - Measuring Higher Volumes 128
  - Output 128
- What Can Go Wrong 129

20. photoresistor 131
- What It Does 131
  - Schematic Symbol 131
- How It Works 132
  - Construction 132
- Variants 132
  - Photoresistors in Optical Isolators 132
- Values 133
  - Comparisons with a Phototransistor 133
- How to Use It 133

Choosing a Series Resistor 134
What Can Go Wrong 134
Overload 134
Excessive Voltage 134
Confusion Among Components 134

21. **photodiode 135**
What It Does 135
Schematic Symbols 135
Applications 135
How It Works 135
Variants 136
PIN Photodiodes 136
Avalanche Diodes 136
Packages 136
Wavelength Range 137
Photodiode Arrays 137
Output Options 137
Specific Variants 137
Values 138
How to Use It 139
What Can Go Wrong 140

22. **phototransistor 141**
What It Does 141
Schematic Symbols 141
Applications 142
How It Works 142
Variants 142
Optional Base Connection 142
Photodarlington 142
PhotoFET 142
Values 143
Behavior Compared to Other Light Sensors 143
Binning 144
How to Use It 144
Output Calculation 145
What Can Go Wrong 145
Visual Classification Errors 145
Output Out of Range 145

23. **NTC thermistor 147**
What It Does 147
Schematic Symbols 147
Applications 148
How an NTC Thermistor Works 148

Output Conversion for Temperature Sensing 149
Choosing a Series Resistor 149
Wheatstone Bridge Circuit 150
Deriving the Temperature Value 150
Inrush Current Limiter 150
Restart 151
Thermistor Values 152
Time and Temperature 152
Resistance and Response 152
Kilohms and Kelvin 152
Reference Temperature 152
Reference Resistance 152
Dissipation Constant 152
Temperature Coefficient 152
Thermal Time Constant 152
Tolerance 153
Temperature Range 153
Switching Current 153
Power Limitations 153
Interchangeability 153
What Can Go Wrong 153
Self-Heating 153
Heat Dissipation 153
Lack of Heat 153
Addendum: Comparison of Temperature Sensors 153
NTC Thermistor 154
PTC Thermistor 154
Thermocouple 154
Resistance Temperature Detector 154
Semiconductor Temperature Sensor 155

24. PTC thermistor 157
What It Does 157
Schematic Symbols 157
PTC Overview 158
Silistor for Temperature Measurement 158
RTDs 159
Nonlinear PTC Thermistors 159
Over-Temperature Protection 159
Over-Current Protection 160
PTC Inrush Current Limiting 161
PTC Thermistor for Lighting Ballast 162
PTC Thermistor as a Heating Element 162
What Can Go Wrong 163
Self-Heating 163
Heating Other Components 163

25. **thermocouple** 165
- What It Does 165
  - Schematic Symbol 166
- Thermocouple Applications 166
- How a Thermocouple Works 167
  - Thermocouple Details 168
- How to Use It 168
  - Types of Thermocouples 169
  - Seebeck Coefficients 169
  - Chips for Output Conversion 170
- Thermopile 171
- What Can Go Wrong 171
  - Polarity 171
  - Electrical Interference 171
  - Metal Fatigue and Oxidation 171
  - Using the Wrong Type 171
  - Heat Damage from Creating a Thermocouple 172

26. **RTD (resistance temperature detector)** 173
- What It Does 173
- RTD Attributes 173
  - Schematic Symbol 174
  - Applications 174
- How It Works 174
- Variants 175
  - Wiring 175
  - RTD Probe 176
  - Signal Conditioning 176
- What Can Go Wrong 176
  - Self-Heating 176
  - Insulation Affected by Heat 176
  - Incompatible Sensing Element 176

27. **semiconductor temperature sensor** 177
- What It Does 177
  - Semiconductor Temperature Sensor Applications 178
  - Schematic Symbol 178
  - Attributes 178
- How It Works 179
  - CMOS Sensors 179
  - Multiple Transistors 179
  - PTAT and the Brokaw Cell 180
- Variants 180
  - Analog Voltage Output 181
  - Analog Current Output 182
  - Digital Output 183

CMOS Semiconductor Temperature Sensors 185
What Can Go Wrong 185
  Different Temperature Scales 185
  Interference in Cable Runs 185
  Latency 186
  Processing Time 186

28. infrared temperature sensor 187

What It Does 187
  Applications 188
  Schematic Symbol 188
How It Works 188
  Thermopile 189
  Temperature Measurement 190
Variants 190
  Surface-Mount Specifications 191
  Sensor Arrays 191
Values 191
  Temperature Range 191
  Field of View 192
What Can Go Wrong 192
  Inappropriate Field of View 192
  Reflective Objects 192
  Glass Obstruction 192
  Multiple Heat Sources 192
  Thermal Gradients 192

29. microphone 193

What It Does 193
  Schematic Symbol 193
How It Works 194
  Carbon Microphone 194
  Moving-Coil Microphone 194
  Condenser Microphone 194
  Electret Microphone 195
  MEMS Microphone 195
  Piezoelectric Microphone 196
Values 196
  Sensitivity 196
  Directionality 197
  Frequency Response 197
  Impedance 198
  Total Harmonic Distortion 198
  Signal-to-Noise Ratio 198
What Can Go Wrong 198
  Cable Sensitivity 198

Noisy Power Supply 198

30. **current sensor** **199**
What It Does 199
Applications 199
Ammeter 199
Schematic Symbol 200
Ammeter Wiring 200
Series Resistor 200
Current-Sense Resistors 201
Voltage Measurement 202
Hall-Effect Current Sensing 202
What Can Go Wrong 203
Confusing AC with DC 203
Magnetic Interference 203
Incorrect Meter Wiring 203
Current Out of Range 203

31. **voltage sensor** **205**
What It Does 205
Applications 205
Volt Meter 205
Schematic Symbol 206
Volt Meter Wiring 206
How It Works 206
Load-Related Inaccuracy 207
Bar Graph 207
What Can Go Wrong 208
Confusing AC with DC 208
High Circuit Impedance 208
Voltage Out of Range 208
Voltage Relative to Ground 208

Appendix A. **Sensor Output** **209**

Glossary **217**

Index **221**

Preface

This third and final volume of the *Encyclopedia of Electronic Components* is devoted entirely to sensors.

Two factors have caused very significant changes in the field of sensors since the 1980s. First, features such as antilock braking, airbags, and emissions controls stimulated the development of low-priced sensors for automotive applications. Many of these sensors were fabricated in silicon as MEMS (microelectromechanical) devices.

The second wave began in 2007 when MEMS sensors were installed in the iPhone. A modern phone may contain almost a dozen different types of sensors, and their size and price have been driven down to a point that would have been unimaginable 20 years previously.

Many MEMS sensors are now as cheap as basic semiconductor components such as a voltage regulator or a logic chip, and they are easy to use in conjunction with microcontrollers. In this Encyclopedia, we have allocated significant space to this segment of the market, hoping that the specific products that we have chosen will remain popular and available for at least the next decade.

In addition, we have devoted space to older components where durability has been proven.

Purpose

While much of the information in this volume can be found dispersed among datasheets, introductory texts, Internet sites, and technical briefings published by manufacturers, we believe there is a real need for a durable resource that assembles all the relevant data in one place, properly organized and verified, including details that may be hard to find elsewhere.

This volume may also serve a useful purpose by attempting to categorize and classify components in a field that is remarkably chaotic. For example, is an *object presence sensor* different from a *proximity sensor*? Some manufacturers seem to think so; others disagree. Understanding the distinctions and the underlying principles can be important if you are trying to decide which sensor to use.

Sensor terminology can also be confusing. To take another example, what is the difference between a *reflective interrupter*, a *reflective object sensor*, a *reflective optical sensor*, a *reflective photointerrupter*, and an *opt-pass sensor*? These terms are used in various datasheets to describe components that are all *retroreflective sensors*. Understanding the proliferating variety of terminology can be essential if you simply want to find something in a product index.

Organization

As in volumes 1 and 2, this volume is organized by subject. For example, if you want to measure temperature, you'll find the entries for a thermistor and a thermocouple next to each other, in an entire section devoted to the sensing of heat. This will help you to compare capabilities and choose the component that best suits your application.

The subject path leading to each sensor is shown at the top of the first page of each entry. For gas flow rate, for instance, you would follow this path:

```
fluid > gas > flow rate
```

Note that the word "fluid" is properly used to include gases as well as liquids.

Exceptions and Conflicts

Unfortunately, some sensors are not easily categorized. There are four problems in this area.

1. What Does a Sensor Really Sense?

A GPS chip is a radio receiver, picking up transmissions from satellites. Does this mean it should be categorized as a sensor of radio waves? No, its purpose is to tell you your location. Therefore, it is categorized as a location sensor. This leads to the first general rule: sensors are categorized by their primary purpose. Secondary purposes may be found in the index.

2. How Many Sensors Are in a Sensor?

Many surface-mount chips perform more than one sensing function. For example, an inertial measurement unit (often identified by its acronym, *IMU*) can contain three gyroscope sensors and three accelerometers—and may contain three magnetometers, too. How should it be categorized?

The answer is that an IMU will be mentioned in more than one entry in the Encyclopedia, because it performs more than one function; but it will not have its own separate entry, because each entry in the Encyclopedia is for a single primary sensing function.

The names of multisensor chips are, of course, included in the index.

3. How Many Stimuli Can One Sensor Sense?

A single sensing element may be used in multiple different types of sensors. The most notable example is the Hall-effect sensor, which can be found in magnetometers, object presence sensors, speed sensors, current sensors, and dozens more. Modern automobiles can contain Hall-effect sensors everywhere from the ignition system to the trunk-locking mechanism. If you are using a hard drive with rotating platters, it probably contains a Hall-effect sensor to monitor the speed of rotation. If you have a generic computer keyboard, each keypress is probably detected with a Hall-effect sensor.

Bearing this in mind, how should a Hall-effect sensor be classified? And where should you expect to find an explanation of how it works?

The answer is that where different types of components contain the same type of sensing element, the entry for each component will include a cross-reference to one location where the sensing element is explained in detail.

This location will be chosen for its relevance. Thus, Hall-effect sensors are explained in the entry for **object presence** sensors, because this is their primary function. While it is true that a Hall-effect sensor works by detecting a magnetic field, that is not its most common application.

4. Too Many Sensors!

Wikipedia lists more than 100 general types of sensors, and even that list is probably not complete. Consequently, we had to pick and choose. Some of the decisions may seem arbitrary, but all of them were made on the grounds of practicality. There were three principles for deciding what to include and what to leave out.

1. Is it a component? We are more interested in board-mounted components than in packaged products that happen to contain sensors. For instance, a thermocouple is often enclosed in a tubular steel probe, and its wire is often plugged in to a specially designed meter that displays temperature. While we do include a photograph of a probe, we are primarily interested in the welded wires of the thermocouple inside it.
2. How much does it cost? An industrial ultrasonic sensor to check items on a factory conveyor belt will be sealed into a module with a waterproof grommet around a shielded cable—which is all very nice, but will not be very affordable. This Encyclopedia is more interested in board-mountable components for one-tenth of the price.
3. How many people are likely to want it? The stock of each type of sensor was checked on component vendor sites. If a sensor wasn't in the inventory, or if only a couple of variants were stocked, we concluded that the limited demand probably didn't justify including it here. For example, a Ferraris acceleration sensor responds to eddy currents in a rotating motor shaft, as a way of measuring vibration in the shaft. This is a really interesting device, but is unlikely to be on most people's shopping lists.

Volume Contents

Having explained the organization of this book and our decisions to include or omit various components, we now present a summary of the contents of all three Encyclopedia volumes:

Volume 1

Power; electromagnetic devices; discrete semiconductors.

The *power* category includes sources of electricity and methods to distribute, store, interrupt, convert, and regulate power. The *electromagnetism* category includes devices that exert force linearly, and others that create a turning force. *Discrete semiconductors* include the primary types of diodes and transistors. See Figure P-1 for a contents listing.

Primary Category	Secondary Category	Component Type
power	source	battery
	connection	jumper
		fuse
		pushbutton
		switch
		rotary switch
		rotational encoder
	moderation	relay
		resistor
		potentiometer
		capacitor
		variable capacitor
	conversion	inductor
		AC-AC transformer
		AC-DC power supply
		DC-DC converter
		DC-AC inverter
	regulation	voltage regulator
electro-magnetism	linear output	electromagnet
		solenoid
	rotational output	DC motor
		AC motor
		servo motor
		stepper motor
discrete semi-conductor	single junction	diode
		unijunction transistor
	multi-junction	bipolar transistor
		field-effect transistor

Figure P-1 *The subject-oriented organization of categories and entries in Volume 1 of this Encyclopedia.*

Volume 2

Thyristors (SCRs, diacs, and triacs); integrated circuits; light sources, indicators, and displays; and sound sources.

Integrated circuits are divided into analog and digital components. *Light sources, indicators, and displays* are divided into reflective displays, single sources of light, and displays that emit light. *Sound sources* are divided into those that create sound, and those that reproduce sound. A contents listing for Volume 2 appears in Figure P-2.

Volume 3

All the most common types of sensing devices, including those that detect location, presence, proximity, orientation, oscillation, force, load, human input, liquid properties, gas types and concentrations, pressure, flow rate, light, heat, sound, and electricity. A contents listing for Volume 3 appears in Figure P-3.

Method

Reference Versus Tutorial

As its title suggests, this is a reference book, not a tutorial. A tutorial such as *Make: Electronics* begins with elementary concepts and builds sequentially toward concepts that are more advanced. A reference book assumes that you may dip into the text at any point, learn what you need to know, and then put the book aside. If you choose to read it straight through from beginning to end, you will find some repetition, as each entry is intended to be self-sufficient, requiring minimal reference to other entries.

Theory and Practice

This book is oriented toward practicality rather than theory. We assume that the reader mostly wants to know how to use electronic components, rather than why they work the way they do. Consequently we do not include detailed proofs of formulae or definitions rooted in electrical theory.

Primary Category	Secondary Category	Component Type
discrete semi-conductor	thyristor	SCR
		diac
		triac
integrated circuit	analog	solid-state relay
		optocoupler
		comparator
		op-amp
		digital potentiometer
		timer
	digital	logic gate
		flip-flop
		shift register
		counter
		encoder
		decoder
		multiplexer
light source, indicator or display	reflective	LCD
	single source	incandescent lamp
		neon bulb
		fluorescent light
		laser
		LED indicator
		LED area lighting
	multi-source or panel	LED display
		vacuum-fluorescent
		electroluminescence
sound source	audio alert	transducer
		audio indicator
	reproducer	headphone
		speaker

Figure P-2 *The subject-oriented organization of categories and entries in Volume 2.*

Primary Category	Attribute to be Sensed	Type of Sensor
spatial	location	GPS
		magnetometer
	presence	object presence
		passive infrared
	distance	object proximity
		linear position
	orientation	rotary position
		tilt
		gyroscope
		accelerometer
mechanical	oscillation	vibration
	force	force
	human input	single touch
		touch screen
fluid	liquid	liquid level
		liquid flow rate
	gas/liquid	pressure
	gas	gas concentration
		gas flow rate
radiation	light	photoresistor
		photodiode
		phototransistor
	heat	NTC thermistor
		PTC thermistor
		thermocouple
		RTD
		semiconductor
		infrared temperature
	sound	microphone
electricity	metering	current
		voltage

Figure P-3 *The subject-oriented organization of categories and entries in Volume 3.*

Sensor Output

In Volumes 1 and 2 of the *Encyclopedia*, each entry included hints on how to use a component. However, many sensors have identical forms of output, which are processed in a similar way. To avoid repetition, general guidance for using nine principal types of sensor outputs has been placed in Appendix A at the back of this volume.

For example, many sensors provide an analog voltage output that varies with the phenomenon that is being sensed. In Appendix A, you will find suggestions on how to adjust the range of the output, if necessary, or how to digitize it with an analog-to-digital converter.

You will also find a comparison between serial protocols such as I2C and SPI, both of which are commonly used when a microcontroller communicates with a digital sensor via a bus.

Glossary

In the world of sensors, many terms tend to recur. *Hysteresis* is one; *MEMS* is another. Rather than define these terms repeatedly, some quick definitions are gathered in a Glossary. Please remember the existence of the glossary if you encounter a term that is unfamiliar. See Glossary.

In many instances, terms that are italicized in the text are defined in the glossary.

Typographical Conventions

Within each entry, **bold type** is used for the first occurrence in each entry of the name of a component that has its own entry elsewhere. Other important electronics terms or component names may be presented in *italics*.

The names of components, and the categories to which they belong, are all set in lowercase type, except where a term is normally capitalized because it is an acronym or a trademark, or contains a proper noun. The term *Hall effect*, for instance, has an initial cap because it is named after a person named Hall. The term *GPS* is all in caps, because it is an acronym; but *psi* (meaning pounds per square inch) remains in lowercase, because even though it is an acronym, the lowercase form is more common.

The situation is different when specifying units that are named after electrical pioneers. All of these units should be lowercased when spelled out. Thus, when referring to the SI unit of force, it is “the newton.” However, where a unit named after a person is abbreviated, the abbreviation is capitalized, as in N for newtons, Hz for hertz, Pa for pascals, and A for amperes.

Mathematical Syntax

In mathematical formulae, we have used the style that is common in programming languages. The * (asterisk) is used as a multiplication symbol, while the / (forward slash) is used as a division symbol. Where some terms are in parentheses, they must be dealt with first. Where parentheses are inside parentheses, the innermost ones must be dealt with first. Consider this example:

```
A = 30 / (7 + (4 * 2) )
```

You would begin by multiplying 4 times 2, to get 8; then add 7, to get 15; then divide that into 30, to get the value for A, which is 2.

Visual Conventions

Figure P-4 shows the conventions that are used in the schematics in this book. A black dot always indicates a connection, except that to minimize ambiguity, the configuration at top-right is avoided, and the configuration at top-center is used instead. Conductors that cross each other without a black dot do not make a connection. The styles at bottom right are sometimes seen elsewhere, but are not used here.

All the schematics are formatted with pale blue backgrounds. This enables components such as switches, transistors, and LEDs to be highlighted in white, drawing attention to them and clarifying the boundary of the component. The white areas have no other meaning.

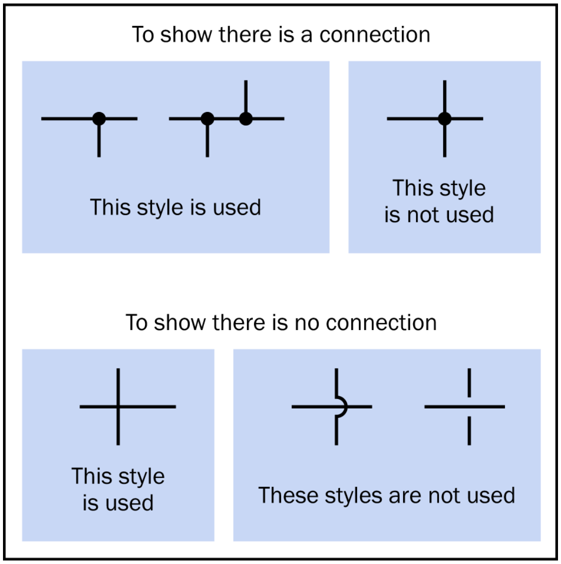
To show there is a connection
This style is used
This style
is not used
To show there is no connection
This style
is used
These styles are not used

Figure P-4 *Visual conventions that are used in the schematics in this book.*

Units and Backgrounds

So long as the United States clings stubbornly to the habit of expressing dimensions in inches, there is a good argument to follow this custom in books intended for an American audience. With this in mind, Volumes 1 and 2 mostly avoided metric units of length. However, as time passed, the books were translated for use in many parts of the world where the inch is regarded as an anachronism.

Recognizing that we now have an international audience, we have used the metric system throughout this volume (with very few exceptions, such as a photograph of an American plumbing fixture that is designed to fit 3/4” pipe). For readers who are metrically impaired, here are some units of length, and their abbreviations:

- 1 nanometer (nm)
- 1 micrometer (µm) = 1,000nm
- 1 millimeter (mm) = 1,000µm
- 1 centimeter (cm) = 10mm
- 1 meter (m) = 100cm = 1,000mm

A micrometer is also known as a *micron*.

The basic conversion factor from meters to inches is 0.0254. Thus:

- 1 inch = 2.54cm = 25.4mm
- 1/1000 inch = 25.4µm

Sometimes 1/1000 inch is called a *mil*.

In many of the component photographs, a graph-paper background is included. Each square in these backgrounds is 1mm.

To avoid confusion, please remember that a few of these same component photographs appeared in books such as *Make: More Electronics*, where the background grid was in tenths of an inch. Captions to photographs in this volume will remind you that millimeters are now used.

Background colors in the photographs were chosen for contrast with the colors of the components, or for visual variety. They have no other significance.

Component Availability

The world of sensors is changing rapidly, and we have no way of knowing if a component will enjoy a long production run. We recommend checking availability at the following suppliers, which we used frequently during the preparation of the book:

- *http://www.mouser.com*
- *http://www.jameco.com*
- *http://www.sparkfun.com*
- *http://www.adafruit.com*

For obsolete parts, or those that are nearing the end of their commercial life, eBay can be very useful. Alternatively, new substitutions for old parts are often listed at *http://www.mouser.com*.

Issues and Errata

There are three situations where the reader and the writer may want to communicate with each other.

- We may want to tell you if the book contains a mistake of some significance. This is *us-informing-you* feedback.
- You may want to tell us if you think you found an error in the book. This is *you-informing-us* feedback.
- You may be having trouble making something work, and you don't know whether we made a mistake or you made a mistake. You would like some help. This is *you-asking-us* feedback.

Here's how you can deal with each of these situations.

Us Informing You

If you already registered your contact information in connection with *Make: Electronics* (second edition) or *Make: More Electronics*, you don't need to register again for updates relating to the Encyclopedia. If you have not already registered, here's how it works.

The only way you can be notified if there's an error in the book is if you supply your contact information. If we have your email address:

- You will be notified of any significant errors that are found in this book, and you will receive a correction.
- You will be notified if there is a completely new edition of this book, or of *Make: Electronics*, or any other books by Charles Platt. These notifications will be very rare.

Your contact information will not be used for any other purpose.

Simply send a blank email (or include some comments in it, if you like) to:

make.electronics@gmail.com

Please put REGISTER in the subject line.

You Informing Us

If you only want to report an error that you have found, it's really better to use the "errata" system maintained by our publisher. The publisher uses the "errata" information to fix the error in updates of the book.

If you feel sure that you found an error, please visit:

http://bit.ly/encyclopedia_electronic_components_v3

The web page will tell you how to submit errata.

You Asking Us

Our time is obviously limited, but if you have a question, a quick answer may be available. You can send email to *make.electronics@gmail.com* for this purpose. Please put the word HELP in the subject line.

Going Public

There are dozens of forums online where you can discuss this book and mention any problems you are having, but please be aware of the power that you have as a reader, and use it fairly. A single negative review can create a bigger effect than you may realize. It can certainly outweigh half-a-dozen positive reviews.

Responses in the past have been generally positive, but in a couple of cases people have been annoyed over small issues such as being unable to find a part online. Help is available on this kind of topic, if you need it. All you have to do is send a request to *make.electronics@gmail.com*.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

You can access the errata page at *http://bit.ly/encyclopedia-electronic-components-v3*.

Make: unites, inspires, informs, and entertains a growing community of resourceful people who undertake amazing projects in their backyards, basements, and garages. Make: celebrates your right to tweak, hack, and bend any technology to your will. The Make: audience continues to be a growing culture and community that believes in bettering ourselves, our environment, our educational system—our entire world. This is much more than an audience, it's a worldwide movement that Make is leading. We call it the Maker Movement.

For more information about Make:, visit us online:

Make: magazine: *http://makezine.com/magazine*
Maker Faire: *http://makerfaire.com*
Makezine.com: *http://makezine.com*
Maker Shed: *http://makershed.com*

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com.

Acknowledgments

Datasheets and tutorials maintained by component manufacturers were considered the most trustworthy sources of information online. In addition, component retailers, college texts, crowd-sourced reference works, and hobbyist sites were used. The following books provided useful information:

Boylestad, Robert L. and Nashelsky, Louis: *Electronic Devices and Circuit Theory*, 9th edition. Pearson Education, 2006.

Braga, Newton C.: *CMOS Sourcebook*. Sams Technical Publishing, 2001.

Hoenig, Stuart A.: *How to Build and Use Electronic Devices Without Frustration, Panic, Mountains of Money, or an Engineering Degree*, 2nd edition. Little, Brown, 1980.

Horn, Delton T.: *Electronic Components*. Tab Books, 1992.

Horn, Delton T.: *Electronics Theory*, 4th edition. Tab Books, 1994.

Horowitz, Paul and Hill, Winfield: *The Art of Electronics*, 2nd edition. Cambridge University Press, 1989.

Ibrahim, Dogan: *Using LEDs, LCDs, and GLCDs in Microcontroller Projects*. John Wiley & Sons, 2012.

Kumar, A. Anand: *Fundamentals of Digital Circuits*, 2nd edition. PHI Learning, 2009.

Lancaster, Don: *TTL Cookbook*. Howard W. Sams & Co, 1974.

Lenk, Ron and Lenk, Carol: *Practical Lighting Design with LEDs*. John Wiley & Sons, 2011.

Lowe, Doug: *Electronics All-in-One for Dummies*. John Wiley & Sons, 2012.

Mims III, Forrest M.: *Getting Started in Electronics*. Master Publishing, 2000.

Mims III, Forrest M.: *Electronic Sensor Circuits & Projects*. Master Publishing, 2007.

Mims III, Forrest M.: *Timer, Op Amp, & Optelectronic Circuits and Projects*. Master Publishing, 2007.

Predko, Mike: *123 Robotics Experiments for the Evil Genius*. McGraw-Hill, 2004.

Scherz, Paul: *Practical Electronics for Inventors*, 2nd edition. McGraw-Hill, 2007.

Williams, Tim: *The Circuit Designer's Companion*, 2nd edition. Newnes, 2005.

In addition, three individuals provided special assistance. Our editor, Brian Jepson, was immensely helpful in the development of this book. Philipp Marek reviewed the text for errors, and Erico Narita collaborated on the Photoshop work.

GPS

1

The acronym **GPS** properly refers to the entire Global Positioning System, including satellites and ground-based control installations. However, a *GPS sensor* consists of a surface-mount chip that processes signals from GPS satellites using a small rectangular antenna, often mounted on top of a *GPS chip*.

A *GPS module* is usually a small board on which a GPS sensor is mounted with additional components. A *GPS receiver* is a device including a data display and other features, such as memory, in addition to a GPS module. In casual colloquial usage, someone who refers to "a GPS" usually means a GPS receiver.

GPS is almost always capitalized without periods.

OTHER RELATED COMPONENTS

- **magnetometer** (see Chapter 2)

What It Does

The Global Positioning System is a navigational aid jointly funded by the U.S. Department of Defense and the U.S. Department of Transportation, while being maintained by the U.S. Air Force. Signals from GPS satellites can be received and processed by modules in a wide variety of equipment ranging from aircraft to wristwatches. The signals provide location data, and may also be used as an accurate time reference.

Schematic Symbol

There is no specific schematic symbol for a GPS chip. It is likely to be shown as a box containing abbreviations that define pin functions, similar to any integrated circuit chip.

GPS Segments

The Global Positioning System consists of three segments:

The space segment

This originally required 24 communications satellites, but was revised in 2011 to require 27, to provide better global coverage. As of August 2015, there were actually 31 satellites in service, with additional "spares" that can be activated if necessary. The satellites occupy orbits 12,500 miles above the Earth, allowing each of them to circle the planet twice in 24 hours. Specifications are maintained online.

The control segment

This includes a master ground-based control station, an alternate master control station, 12 command and control antennas, and 16 monitoring sites, all maintained by the U.S. Air Force.

The user segment
This consists of receiving devices, including those that are government-owned and those that are privately owned.

How It Works

Each satellite carries multiple atomic clocks that maintain precise time, and a pseudo-random number generator in the form of a linear-feedback **shift register** (see Volume 2).

A GPS receiver can distinguish the signals from at least four satellites by comparing their received pseudo-random bit sequences, and can compute the receiver's distance to each satellite by comparing the arrival times of satellite signals.

When a satellite appears above the horizon, it approaches a receiver. After passing overhead, it moves away. This relative motion causes a *Doppler shift* in the received frequency, which the receiver circuit must take into account.

GPS satellites transmit on several frequencies simultaneously. The one for civilian use is 1575.42 MHz, called L1. Another one, 1227.6 MHz, called L2, is reserved for military use.

Variants

A GPS chip generally processes input from an antenna and provides output through solder pads. The antenna is often integrated as a ceramic square or rectangle mounted above the chip, but many chips can also process input from an external antenna. Figure 1-1 shows a GPS chip with a metal shield that is easily mistaken for an antenna. In Figure 1-2, the GPS sensor does incorporate a ceramic antenna.

Some GPS chips contain flash memory for internal data logging, although this is not a standard feature.

Suppliers such as Adafruit and Sparkfun offer GPS modules mounted on breakout boards for easier connection with other components, as shown in Figure 1-2. Some breakout boards also include provision for battery backup with a button cell.

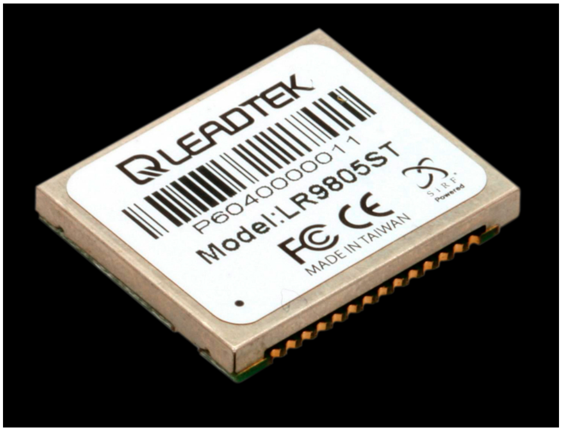

Figure 1-1 *A GPS sensor. This surface-mount chip is hidden beneath a metal shield.*

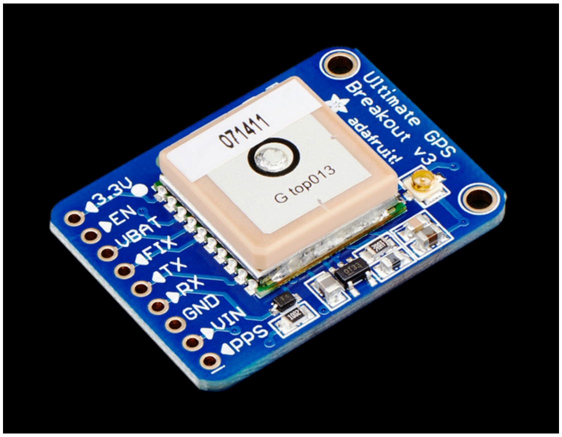

Figure 1-2 *A breakout board offered by Adafruit, incorporating a GPS sensor.*

GPS capability is almost always included in modern smartphones and tablets. It is used in handheld devices for navigation on foot, and in devices designed to be mounted in motor vehicles. Cars offer GPS capability as an option accessed via a built-in screen.

A *GPS tracker* is a device that may lack a display and simply logs its position to internal memory, from which data can be downloaded to a computer later. Many (older) handheld GPS receivers have a connector giving access to a serial or USB port, and provide data in the same NMEA format as the GPS modules described below.

After the Global Positioning System became widely used, other competing systems were introduced. These include the European Galileo, the Russian GLONASS (an acronym for Global Navigation Satellite System), and the Chinese Beidou. As of 2015, GLONASS had become fully operational. Some receivers, including those in cellular phones, compare signals from GPS and GLONASS satellites to achieve higher accuracy.

Values

Sensitivity is expressed usually in dBm, meaning the power ratio in decibels (dB) of the measured power referenced to one milliwatt (mW).

Time to First Fix (TTFF) is the time required to obtain an initial fix from a satellite.

Number of channels is the number of satellites that a GPS receiver can track simultaneously. Early GPS receivers could sense only four channels. Modern units may be able to deal with 22.

Power consumption may be measured in milliwatts. For example, the FGPMMOPA6H GPS standalone module by G.top claims power consumption of 82mW during acquisition of satellite signals and 66mW while tracking subsequently. At a typical voltage of 4VDC, the chipset consumes about 20mA and approximately 17mA, respectively.

Form factor. This is the size of a chip, often determined by the dimensions of the ceramic antenna on top of it. Dimensions may be 15mm x 15mm or larger.

Update rate. The number of position measurements per second. While 1 update per second is often sufficient, some chips generate updates more rapidly. The frequency of updates is expressed in hertz.

Output type. This is often TTL-level serial providing NMEA data. The baud rate can vary and is often selectable.

Supply voltage. Often below 5VDC.

Current consumption. Higher during satellite acquisition.

How to Use It

A GPS module requires only a DC power supply, and will start outputting data as soon as it has identified satellites that are currently within range.

Data provided by a GPS module uses a rather slow and primitive plain-ASCII protocol known as NMEA, developed by the National Marine Electronics Association. Each block of data is known as a *sentence*, and can be parsed independently of previous and subsequent sentences. The default transmission rate is 4,800 bits per second, using 8 bits to identify an ASCII character, no parity, and 1 stop bit. However, some GPS modules use a serial rate of 9,600 bps or faster.

A sentence begins with a two-letter abbreviation defining the type of device employing the sentence. For a GPS device, the abbreviation is GP. The sentence continues with another abbreviation of three characters or more, describing the type of data being transmitted, so that numeric values in the sentence can be interpreted correctly.

The remainder of a sentence consists of letters and numerals in plain ASCII, with values separated by commas. A sentence cannot contain more than 80 characters. A sentence will specify the latitude, longitude, and altitude of the GPS, and a value defining the time when the readings were derived from satellite signals. Some sentence data structures are proprietary, established by the device manufacturer, and will begin with the letter P.

A GPS device may send a variety of different sentence types in succession, to overcome the 80-character sentence length limit. Each sentence will be preceded with an identifier. Sentence types and data contents will be defined in a manufacturer's datasheet.

Output from GPS chips may be compatible with a microcontroller. Output from a GPS breakout board will almost certainly be compatible, and the board is likely to have its own voltage regulator. The microcontroller can receive serial data from the GPS chip, can stop the GPS chip via its enable pin, and can start and stop internal logging of data in flash memory on the GPS chip if this feature is available.

Code libraries are available online for a microcontroller, to enable it to receive and interpret serial data from a GPS device.

Pulse per Second Output

As the GPS positioning depends on calculating distances from the time it takes for a radio signal to travel, accurate timekeeping is needed. When a GPS receiver obtains a fix for its position, it will get a value for the current time as well. This makes GPS receivers usable to supply time and frequency standards. Most receiver modules report the time together with the position. Many also provide a special PPS output, which produces one pulse per second, precisely synchronized with the satellites in view.

The accurate time provided by the GPS receiver can be used to *discipline* a crystal oscillator. This means measuring the crystal oscillator frequency relative to a GPS-provided reference, and continuously adjusting the crystal frequency to keep it stable.

What Can Go Wrong

Problems generally affect chips and modules rather than devices.

Electrostatic Discharge

The patch antenna on a GPS chip connects with the chip via an RF input. If the antenna is subjected to an electrostatic discharge, the chip can be permanently damaged. Likewise, damage will result if a discharge is applied to the RF input, for example through a soldering iron. The chip must be grounded before any work is done involving the RF input.

Failure to Ground Properly

The ground solder pad on a chip, or the ground pin on a breakout board, must make contact before voltage is applied to other pads or pins.

Cold Joints

The patch antenna on a GPS chip can function as a heat sink, increasing the risk of cold solder joints when the chip is being mounted on a board.

Restricted Availability

U.S. regulations limit the export of some GPS devices capable of rapid positional updates that could be usable in military aircraft or missiles. Other restrictions also apply. Suppliers may be inconsistent in their refusal to allow purchase of restricted items outside of the United States.

Inability to Detect Satellites

Any GPS device may fail to detect satellite signals if its view of the sky is obstructed. Reception is usually possible through window glass, but may not be available through walls, a roof, heavy tree cover, and other natural obstructions.

Exceeding Maximum Velocity or Altitude

Security regulations limit the capability of GPS devices to function above 60,000 feet or at greater than 1,200 miles per hour. Outside of these limits, the GPS device will not provide data. This may affect applications in amateur rocketry or high-altitude balloons.

magnetometer

2

This entry deals only with magnetic sensors that respond to the Earth's magnetic field. Small magnetic sensors such as the ubiquitous *Hall sensor* may be used for many other purposes, such as determining the position or rotational speed of mechanical components. These applications involve **object presence** sensing; see Chapter 3. In that entry, Hall sensors are discussed at "Hall-Effect Sensor".

In the past, a **magnetometer** was a bulky measurement device incorporating knobs or other controls and some form of display. While that use of the word is still common, this entry deals only with chip-based magnetic sensors.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **gyroscope** (see Chapter 9)
- **GPS** (see Chapter 1)

What It Does

A traditional compass consists of a thin magnetized strip of metal balanced on a pivot. It aligns itself with the Earth's magnetic field.

A *scalar magnetometer* measures the total strength of a magnetic field. A *vector magnetometer* can measure the strength in a specified direction. In particular, it can provide a numeric output describing the angle between the orientation of the measuring device and the Earth's magnetic poles.

Chip-based magnetometers are usually vector-type, containing three sensors mounted orthogonally—that is, each of them at 90 degrees to the other two. Suitable software can interpret the analog readings from the sensors to calculate magnetic north or south regardless of the angle at which the instrument is being held, relative to the ground.

Schematic Symbol

There is no specific schematic symbol for a magnetometer.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

Magnetometers are found in handheld equipment such as digital compasses, cameras, and cellular phones. They are usually surface-mount chips that are manufactured in large quantities and may be used in conjunction with microcontrollers. For the hobby-electronics community, or for experimental product development, a magnetometer may be mounted on a breakout board for ease of use. A board using a Honeywell HMC5883L is shown in Figure 2-1.

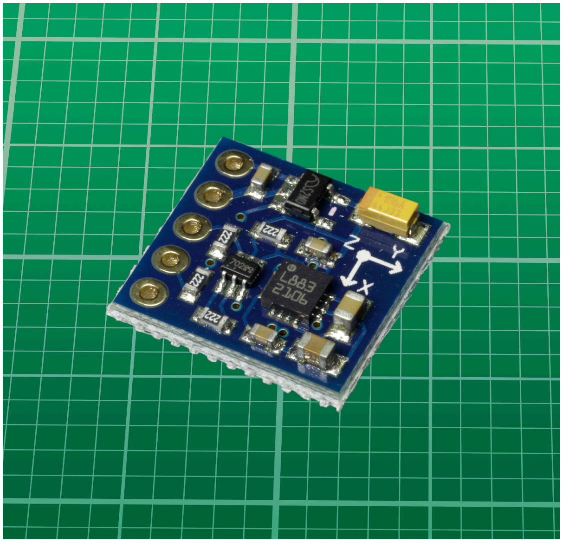

Figure 2-1 *The Honeywell HMC5883L 3-axis magnetometer mounted on a breakout board. The background grid is in millimeters.*

How It Works

An explanation of magnetometers requires an understanding of the fundamentals of magnetism.

Magnetic Fields

A magnetic field is often represented by *field lines* that show the strength and vector of the field. Field lines associated with a simple permanent magnet are shown in Figure 2-2, where the strength, or *flux density*, at any point, is indicated by the spacing the lines, while the angle tangential to a line indicates its vector. (For an extended discussion of magnetism, see **electromagnet** in Volume 1.)

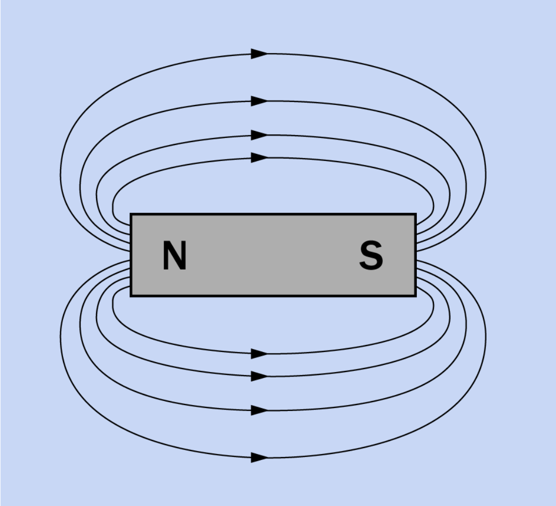

Figure 2-2 *Field lines representing magnetic flux created by a bar magnet. Space between the field lines is inversely proportional to flux density. In reality this is a three-dimensional effect, and a more accurate representation would show the magnet and field lines revolved around the axis of the magnet.*

The flux density of a magnetic field is usually represented by the letter B, and is measured in *newton-meters per ampere*, more commonly referred to as *teslas* (T). An older unit of measurement was gauss (G), 1 tesla being equivalent to 10,000 gauss. Some datasheets still refer to gauss.

The Earth's magnetic field is believed to result from convection currents in the outer liquid of the Earth's core. The strength of this field varies from 25 to 65 microteslas (0.25 to 0.65 gauss) depending on the location where it is measured. To a very rough approximation, the Earth behaves as if a giant bar magnet connects the magnetic north pole and the magnetic south pole. See Figure 2-3.

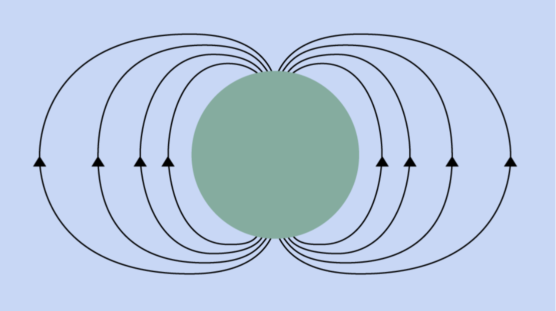

Figure 2-3 *The magnetic field of the Earth approximately resembles the field around a bar magnet.*

Note that in northern and southern latitudes, the field lines angle steeply down toward the surface, while the lines are approximately parallel with the surface near the equator. Consequently a magnetometer held tangentially to the surface of the Earth will tend to measure a stronger horizontal field near the equator than near the poles.

The varying angle of the field lines tangential to the Earth is known as their *inclination*. Variations in field strength can be used to determine an approximate location, although the **GPS** (Global Positioning System, using satellites) enables this much more accurately.

Confusion results from the fact that the nothern magnetic pole of the Earth actually behaves as a south pole, while the southern magnetic pole behaves as a north pole. That is, when a permanent magnet is free to rotate, its north pole will orient itself toward the so-called northern magnetic pole of the Earth, even though opposite poles attract. The northern pole of the Earth should therefore be thought of as the pole that attracts the north end of a compass.

Earth's Axes

Planet Earth revolves around an imaginary line referred to as the *axis of rotation*. This line is close to, but not quite coincident with, the *magnetic axis* connecting the magnetic north pole and magnetic south pole, as shown in Figure 2-4.

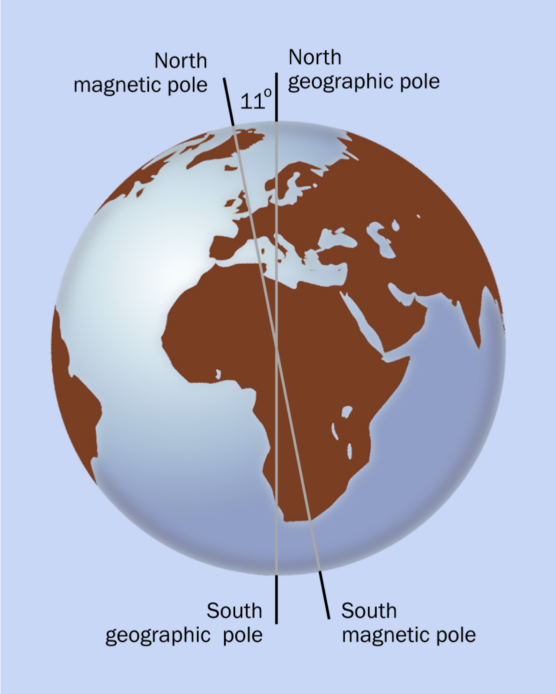

Figure 2-4 *The angle between the magnetic axis and the rotational axis of the Earth is approximately 11 degrees.*

Magnetic declination is the angle between the magnetic north pole and the geographical north pole, as perceived by an observer. This angle will vary depending on the observer's location on the surface.

Because of declination, the direction of magnetic force at points on the surface of the planet will vary with latitude and longitude, as shown in Figure 2-5. In this figure, the *magnetic meridians* are shown in red, superimposed on the *geographical meridians* shown in green. Magnetic meridians show the direction of magnetic force, while geographic meridians are drawn between the ends of the axis of rotation of the Earth. While there is an approximate correlation between the two, in some areas, especially near the north and south poles, the discrepancy can be more than 40 degrees.

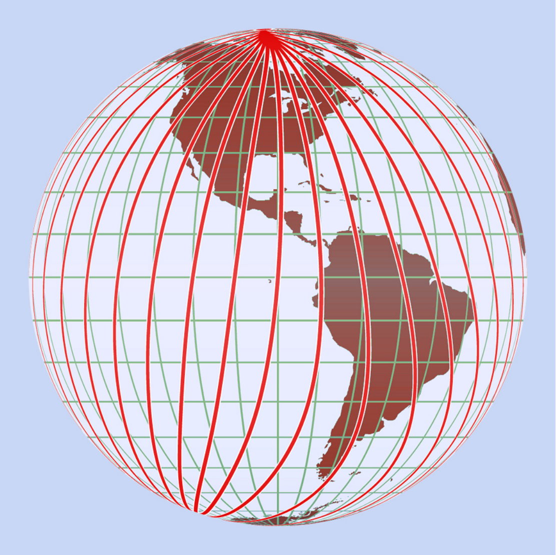

Figure 2-5 *The red lines indicate the direction in which a compass would be likely to point to magnetic north. The green lines connect the geographical poles of the planet. (From Wikimedia Commons.)*

Standard declination tables for locations on the Earth are available, and these values must be added to, or subtracted from, the reading from a compass or magnetometer to determine the direction of geographical north. Navigational systems customarily express the heading of a vehicle or vessel relative to geographical north, as shown in Figure 2-6.

Coil Magnetometer

Current flowing through a wire creates a magnetic field with flux density that is directly proportional to the current in amperes. Conversely, a changing magnetic field will induce current in a wire. This principle is used in a *coil magnetometer*, which can detect buried objects when the coil moves above them. A *rotating coil magnetometer* can determine magnetic field strength while remaining in a stationary position. However, the coil in a coil magnetometer must be relatively large.

Hall Effect and Magnetoresistance

In modern handheld devices, a magnetometer will generally use either the *Hall effect* (see "Hall-Effect Sensor") or *magnetoresistance*, described here.

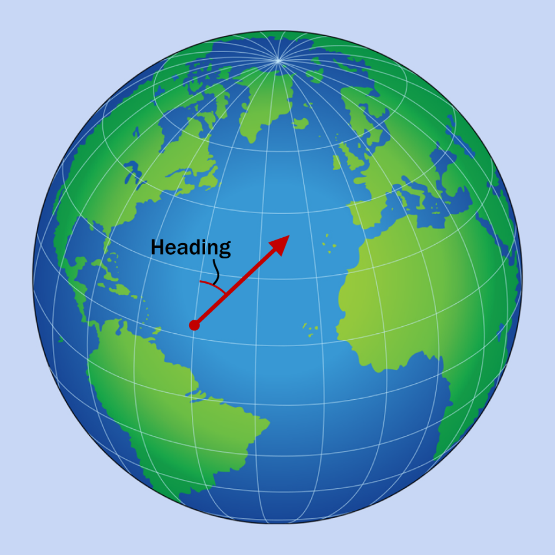

Figure 2-6 *Heading is an angle usually calculated relative to geographical north, sometimes referred to as "true north."*

Magnetoresistance is a phenomenon where the resistance of a material changes fractionally when it is exposed to a magnetic field. This is capable of yielding greater accuracy than Hall-effect sensors, but has tended to be more expensive.

Orthogonally oriented sensors contained in a surface-mount chip are aligned with axes identified with letters X, Y, and Z. These sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in *registers* that are available to other devices, often via the *I2C* protocol, which is widely used by microcontrollers.

Typically there will be two eight-bit registers for each axis, one defining the high byte and the other defining the low byte for the digital value. In reality the ADC is likely to use 10 to 13 bits, with the remaining 6 to 3 bits being unused.

Variants

The Freescale Semiconductor FXMS3110 is a typical low-cost chip containing a 3-axis magnetometer sensor. Many chips now additionally include a 3-axis accelerometer sensor. An example is the LSM303 by STMicroelectronics, which is sold on a breakout board by Adafruit. See Figure 2-7.

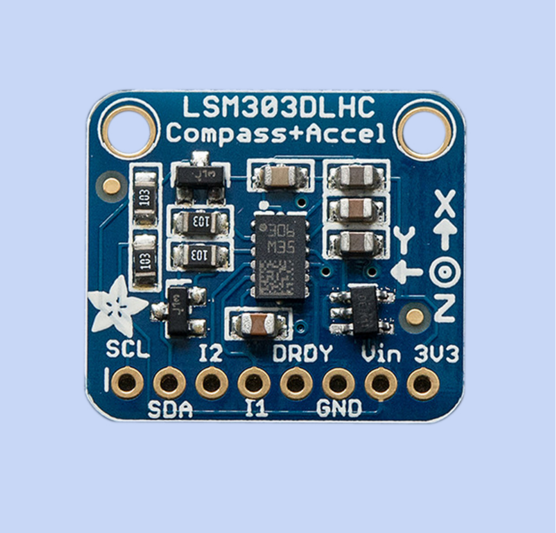

Figure 2-7 *The LSM303 is a chip manufactured by STMicroelectronics. It is shown here on a breakout board from Adafruit.*

See Chapter 10 for an explanation of the function of accelerometers.

The InvenSense MPU-9250 is a highly sophisticated IMU, including a 3-axis **gyroscope** in addition to a 3-axis magnetometer and 3-axis accelerometer.

A processing unit in the MPU-9250 reconciles the nine variables, and the digital output can be accessed via I2C or SPI protocol at speeds up to 1MHz. All functions of the chip are contained within a package measuring less than 3mm square.

See Chapter 9 for additional details about the functioning of a gyroscope.

How to Use It

A basic 3-axis magnetometer sensor such as the HMC5883L can be tested with a microcontroller that can receive data from its registers via I2C protocol. This is relatively easy with an Arduino, which was designed to be I2C-capable.

Several breakout boards are available with the HMC5883L mounted on them. Many of these boards contain a voltage regulator, allowing a 5VDC power supply to be used even though the chip is designed for a typical supply of 2.5VDC.

In addition to the power supply, the breakout board only requires two connections for I2C communication, to its SCL pin (serial clock input) and SDA pin (serial data input/output). If basic I2C software is installed on the microcontroller, it will read digital values from the magnetometer registers. If additional software is used, it will convert the values to magnetic flux densities in microteslas for the X, Y, and Z axes. Code libraries to achieve these goals are widely available online.

A more sophisticated chip such as the InvenSense MPU-9250 can be used similarly, but yields additional data for conversion. Here again, code libraries can be found online. The slightly older MPU-9150 is sold on a breakout board with downloadable code from Sparkfun.

What Can Go Wrong

Bias

Magnetometers are sensitive to their environment, which can induce *magnetometer bias* of two types.

Hard-iron bias is primarily caused by magnetized material inside the device containing the magnetometer. Because this effect is usually unvarying, compensating for it is relatively easy. *Soft-iron bias* is caused by interaction between variations in the Earth's magnetic field and materials inside a magnetometer that can be magnetized.

A common example of soft-iron bias would be power lines, generating a magnetic field that can affect model aircraft and drones using magnetometers for navigation.

Mounting Errors

Placement of a chip-based magnetometer on a circuit board is critical. The field effects of transformers or relays must be taken into account, and even the low voltage and low current in a circuit trace can create a magnetic field sufficient to disturb a chip. No traces on any layer of the board should pass across the footprint of the chip. A magnetometer should not be mounted within a ferromagnetic case.

object presence sensor

3

An **object presence** sensor may also be described as an *object detector* or *detection sensor*.

The term **proximity** sensor may be applied to this component. However, in this Encyclopedia a proximity sensor has the capability of estimating the distance to a target. See Chapter 5. An object presence sensor merely detects whether an object is within a preset range, and does not provide additional information.

Optical and *magnetic* object-presence sensors are described and compared in this entry. *Ultrasonic sensors* are described in the entry dealing with proximity sensors, because they tend to be used for distance measurement rather than just for detection. Other methods of object presence sensing, including *capacitive*, *doppler*, *inductive*, *radar*, and *sonar*, are not included in this Encyclopedia.

A sensor that detects an object by receiving light reflected from it is categorized as a *reflective* sensor, and is included in this entry. (If a module includes a light source as well as a light sensor, it is properly categorized as *retroreflective*, although that term is not always used.)

A sensor that detects an object when it interrupts a beam of light is a *transmissive* sensor, and is included in this entry. It may also be described as a *through-beam* sensor or as an *optical switch*.

A sensor that responds to the motion of an object that emits infrared radiation is a **passive infrared** motion sensor, also described by its acronym, PIR, and often referred to simply as a *motion sensor*. It has its own separate entry. See Chapter 4.

Phototransistors and **photodiodes** may be used as sensing elements in presence sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **proximity** sensor (see Chapter 5)
- **passive infrared** sensor (see Chapter 4)

What It Does

An **object presence** sensor verifies the presence or absence of an object within a predetermined range, without necessarily measuring how far away it is or how fast it is moving. The object may be described as a *target*.

Object presence sensing is often used to verify the correct function of an automated system—for example, the placement of objects on a conveyor belt. It can also be used to count objects as they pass a sensor.

Some types of security systems use presence sensors to sound an alarm if an intruder interrupts a beam of light. They can verify that a door or window is closed. They may also function as a limit switch to control the operation of a motor.

Schematic Symbol

In schematics, an optical presence sensor may be indicated with the symbol for an LED, plus the symbol for a **phototransistor**, with one or two arrows connecting them, as shown at top-left in Figure 3-1. Wavy-line arrows may indicate an infrared connection.

A **photodiode** may be substituted for a phototransistor, as shown at top-right in the figure.

A magnetic sensor may be shown with the symbol for a *Hall-effect sensor*, as in the lower half of Figure 3-1.

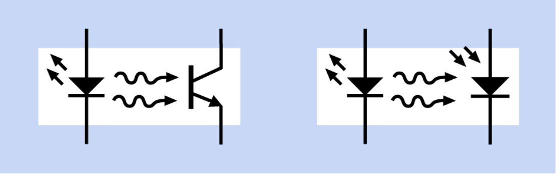

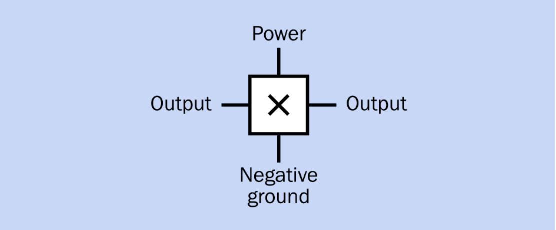
Power
Output
Output
Negative ground

Figure 3-1 *Top: two possible schematic symbols for an optical presence sensor, using a phototransistor (left) or photodiode (right). Other variants are possible. Bottom: the schematic symbol for a Hall-effect sensor, commonly used in a magnetic sensor.*

Variants

To assist the reader in comparing different options for detecting the presence of an object, this entry includes two primary variants: *optical* and *magnetic*.

The optical sensors are divided into *transmissive* and *reflective* (including *retroreflective*). The magnetic sensors are *reed switches* and *Hall-effect sensors*. A chart showing the categories and subcategories is shown in Figure 3-2.

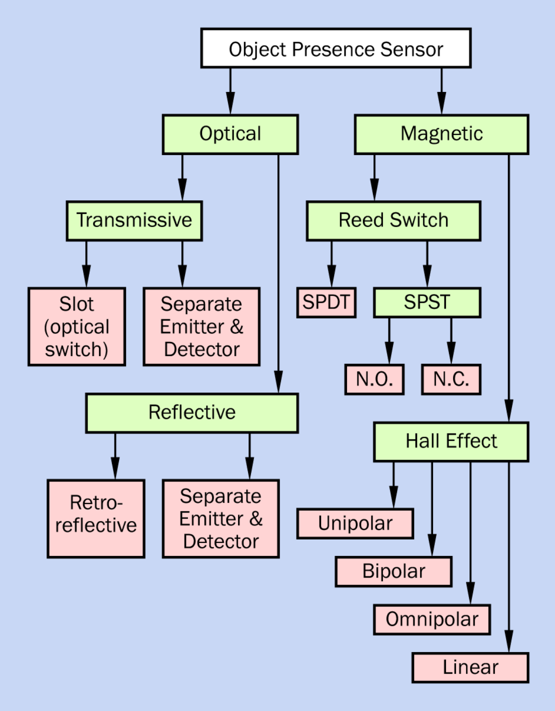
Object Presence Sensor
Optical
Magnetic
Transmissive
Reed Switch
Slot (optical switch)
Separate Emitter & Detector
SPDT
SPST
N.O.
N.C.
Reflective
Hall Effect
Retro-reflective
Separate Emitter & Detector
Unipolar
Bipolar
Omnipolar
Linear

Figure 3-2 *Categories of object presence sensors discussed in this entry. Other types of object presence sensors exist, but are less common, and are not included here.*

Optical Detection

A *transmissive optical sensor*, also known as a *through-beam sensor*, is really a pair of components, one emitting light and the other receiving it. The sensor is triggered if an object interrupts or reflects the light beam, as shown in Figure 3-3.

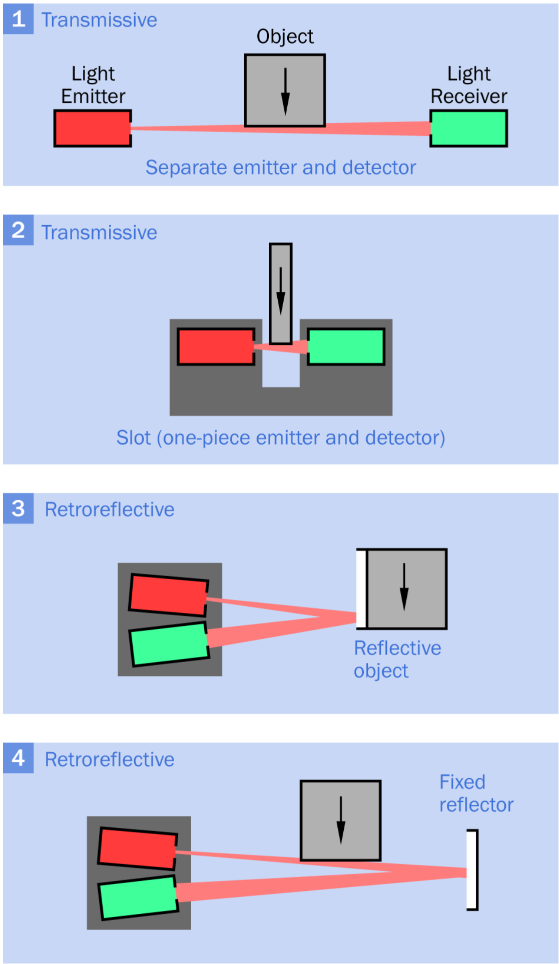

Figure 3-3 *Various configurations of optically activated object presence sensors. See text for details.*

If the light emitter and the light detector face each other across a small gap, they may be contained in a single module (usually with a slot in it) as shown in section 2 of Figure 3-3. This is often referred to as an *optical switch* (not to be confused with the solid-state switching devices used in telecommunications). It is sometimes described as a *photointerruptor* or *photointerrupter*.

A *reflective optical sensor* also consists of a light emitter and a light detector, but they are placed adjacent to each other, facing in the same general direction. When they are mounted in one module, as is often the case, the arrangement is properly known as a *retroreflective sensor*. Either way, the combination may be triggered in one of two ways:

- An object passing in front of the light beam reflects it back to the detector. The object must be naturally reflective (for example, glass containers or white boxes on a conveyor), or must have a reflective patch applied to it, or the light source must be bright enough to reflect from an object that is not highly reflective. This configuration is shown in section 3 of Figure 3-3.
- A stationary reflector can be mounted opposite the light emitter, in which case a detector beside the light emitter is triggered when an object interrupts the reflected light beam. This configuration is shown in section 4 of Figure 3-3.

Transmissive Optical Sensors

A light source and light detector may be sold as separate components in a matched pair. An example is the Vishay TCZT8020, shown in Figure 3-4. These components are small, each measuring no more than 5mm x 3mm. They are designed to be placed just a few millimeters apart. The light source is an infrared LED, while the detector is a **phototransistor** (see Chapter 22 for information about phototransistors).

The source and detector are both designed to use 5VDC. Output from the phototransistor is an open collector. Current through the collector must not exceed 50mA, and must be limited by a pullup resistor of 100 ohms or more. Current through the source must not exceed 60mA, and must be limited by an appropriate series resistor.

Details on using an open-collector output will be found in the Appendix. See Figure A-4.

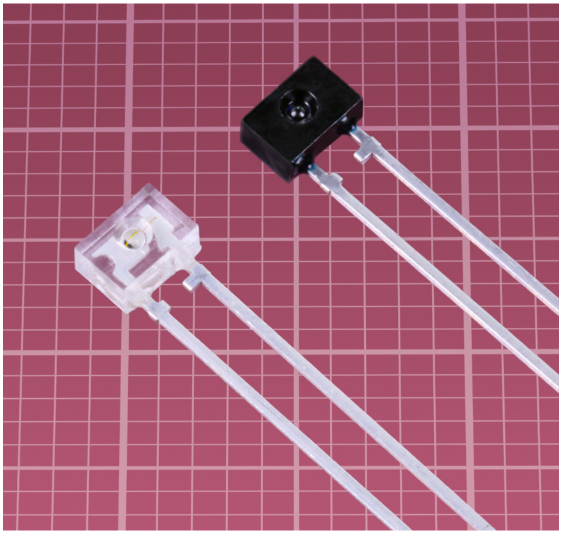

Figure 3-4 *A matched light source and light detector, for use as a transmissive optical sensor. The background grid is in millimeters.*

The Omron EE-SX range includes a variety of configurations of light source and detector separated by a 5mm slot in one module. The source is an infrared LED, and the detector is a phototransistor.

The Omron components tolerate a wide range of supply voltages, from 5VDC to 24VDC, with no series resistor necessary for the LED. Open-collector output from the phototransistor can tolerate 50mA to 100mA, depending on the particular version of the sensor. A red indicator LED is illuminated when an object blocks the slot in the sensor. Some versions have a high output when the slot is open, while others have a high output when the slot is closed. Because of their various features, these sensors are relatively expensive.

A much cheaper optical switch is the Everlight ITR9606 (described by the manufacturer as an "opto interrupter"). It is pictured in Figure 3-5. This is intended as a 5V device and has an open-collector output. It requires a series resistor for the LED and a pullup resistor for the open-collector output. Many similar detectors are available.

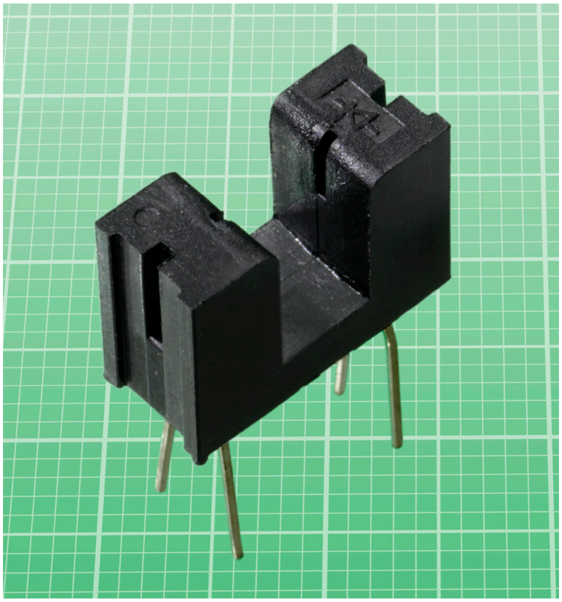

Figure 3-5 *A low-cost transmissive sensor, more commonly known as an optical switch. The background grid is in millimeters. The slot in the sensor is about 5mm wide.*

For longer-distance detection, an infrared receiver can be mounted separately from an infrared LED. The TSSP77038 from Vishay detects infrared light from as far away as 50cm, and pulls its open collector output low in response. The light must be modulated at 38kHz.

Polulo Robotics and Electronics sells a very affordable breakout board containing a TSSP77038 receiver paired with an infrared LED modulated by a 555 timer. It is shown in Figure 3-6. Because this board contains a light source as well as a light detector, it is really a retroreflective sensor.

Where distances exceeding 1 meter are involved, a laser coupled with a phototransistor that is shielded from ambient light may be necessary.

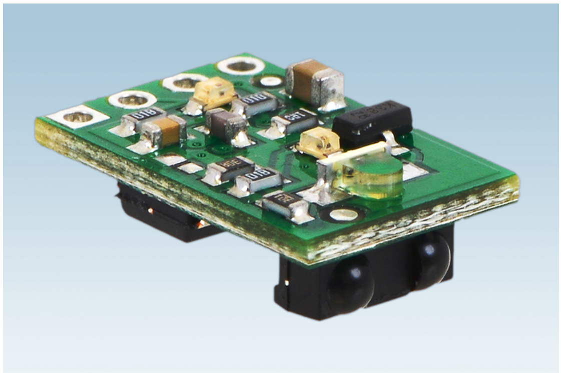

Figure 3-6 *The Vishay TSSP77038 installed with an appropriate light source on a breakout board from Polulo Robotics and Electronics.*

Retroreflective Optical Sensors

As is the case with a transmissive object detector, the retroreflective type may be listed by vendors as an *optical switch*. Other terms used in datasheets include *reflective interrupter*, *reflective object sensor*, *reflective optical sensor*, *reflective photointerrupter*, *opt-pass sensor*, and *photomicrosensor (reflective)*. The remarkable lack of standardization in terminology creates a problem when searching online for these devices. Why so many different names evolved for the exact same component is unclear.

Many retroreflective object-detection sensors are available in packages ranging from 5mm x 5mm to 10mm x 10mm in size. Almost all of these modules are analog devices using an infrared LED as the light source and a phototransistor as the sensor, with an open-collector output. (For more information about phototransistors, see Chapter 22.)

With a suitable pullup resistor, the output voltage will be proportional with the inverse of the distance. If V is the voltage, d is the distance, and k is a conversion factor:

```
V = k * ( 1 / d)
```

While many of the smaller modules are surface-mount, some have leads, as shown in Figure 3-7. A major limitation of these small components is that they have a very limited range, typically less than 5mm. They can only be used in applications such as process control where the position of a target is controlled and predictable.

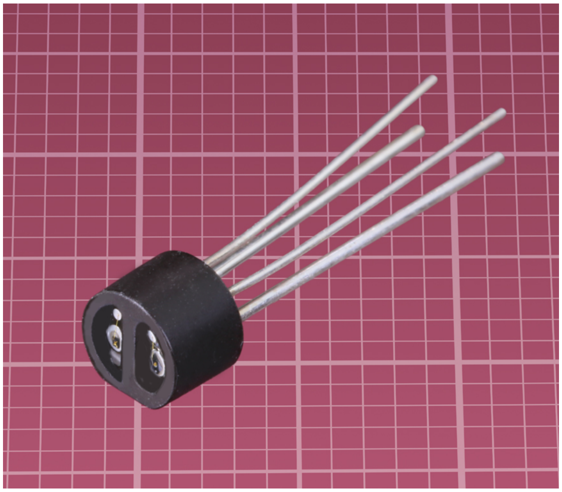

Figure 3-7 *The Rodan RT-530 is a small object-presence sensor with a limited range typical of this type of retroreflective component. The background grid is in millimeters.*

Another example of a retroreflective sensor in a small package is the Optek OPB606A shown in Figure 3-8. The background grid is in millimeters.

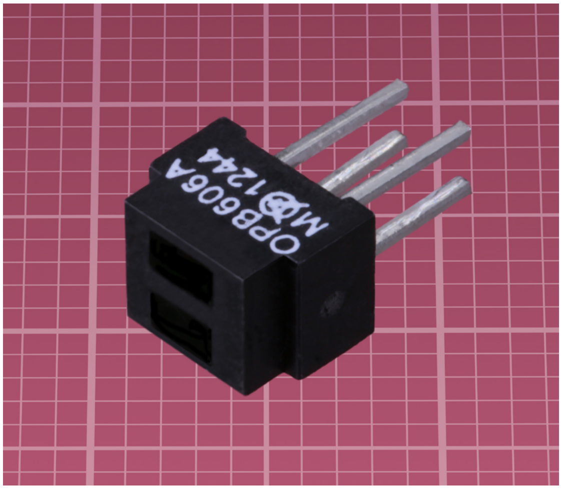

Figure 3-8 *The Optek OPB606A. The background scale is in millimeters.*

A retroreflective module, with a lensed LED and lensed phototransistor to focus the outgoing and reflected beams, is the Vishay TCRT5000 shown in Figure 3-9.

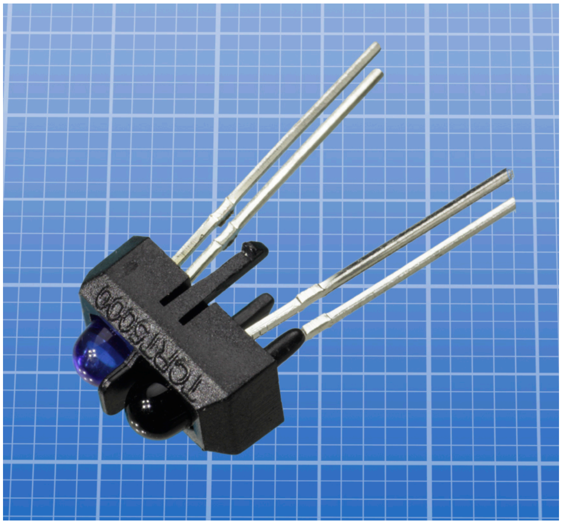

Figure 3-9 *The Vishay TCRT5000 retroreflective sensor. The background grid is in millimeters.*

Retroreflective modules that have a greater range tend to be physically larger, less common, and more expensive. Sharp makes a popular series. Some part numbers and distance sensing limits are: GP2Y0D805Z0F (5mm to 5cm), GP2Y0D810Z0F (2cm to 10cm), and GP2Y0D815Z0F (5mm to 15cm). Figure 3-10 shows the GP2Y0D810Z0F mounted on a small breakout board from Polulo Robotics and Electronics. The board is useful, as the pins on the sensor are spaced only 1.5mm apart. The size of the board is approximately 8mm x 20mm.

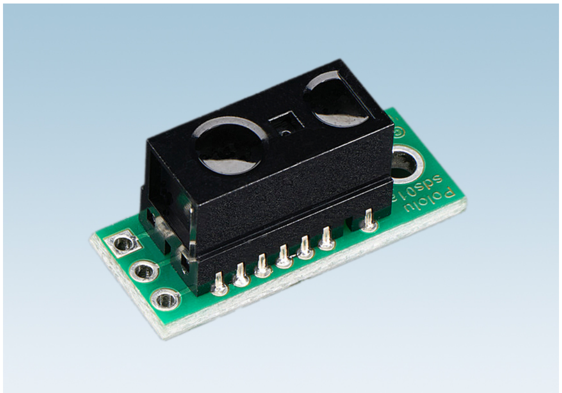

Figure 3-10 *The Sharp GP2Y0D810Z0F detection sensor mounted on a board from Polulo Robotics and Electronics. From a photograph by Adafruit Industries.*

Each of the detection sensors in this series is described in Sharp datasheets as a "Distance Measuring Sensor Unit," but in fact they do not measure distance. A single output is normally logic-high and drops to logic-low when a target is within range. Sharp refers to this as a "digital" output, but in fact it is binary, and should not be confused with the digital buffer on a more sophisticated proximity sensor that contains an analog-to-digital converter and provides a numeric output.

It is important to distinguish the Sharp object-presence sensors listed above from the range of Sharp **proximity** sensors described in the entry in Chapter 5. The proximity sensors are physically larger, and most have an analog output that varies with the distance of a target.

Magnetic Sensors

Prepackaged magnetic sensing units are sold in many configurations for industrial and military applications. Although they may be referred to as "magnetic sensors," they are outside the scope of this Encyclopedia. Here we discuss board-mounted components. Almost always, they use a *reed switch* or a *Hall-effect* sensor as their sensing elements.

Reed Switch

A reed switch is a magnetically activated mechanical switch. It consists of two metallic contacts in a small enclosure that is usually a glass capsule. The contacts are magnetic, and move in response to a magnetic field. A permanent magnet is used to activate the switch. Two reed switches are shown in Figure 3-11.

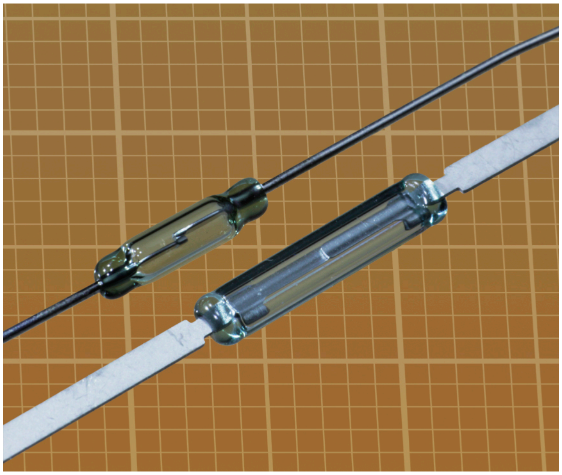

Figure 3-11 *Two SPST reed switches. Although the contacts may seem to be touching, in fact they are separated by a tiny gap, and these switches are the normally open type. The background grid is in millimeters.*

A reed switch exhibits a small amount of *hysteresis*, because the magnetic field strength required to overcome the mechanical resistance of the springy contacts is greater than the field strength required to keep them closed.

Very small electromagnetic relays that only switch a very low current may use a coil-activated reed switch. For purposes of this Encyclopedia, such a component is considered to be a relay, not a sensor. For more information about **relays**, see the entry in Volume 1.

The most common everyday application for a reed switch is in an alarm system that is triggered by an intruder entering a building. A magnet in a sealed plastic enclosure is attached to a door or window, while a reed switch, in another sealed plastic enclosure, is attached to the frame, very close to the magnet. Typical components of this type are shown in Figure 3-12. A diagram illustrating the mode of operation appears in Figure 3-13.

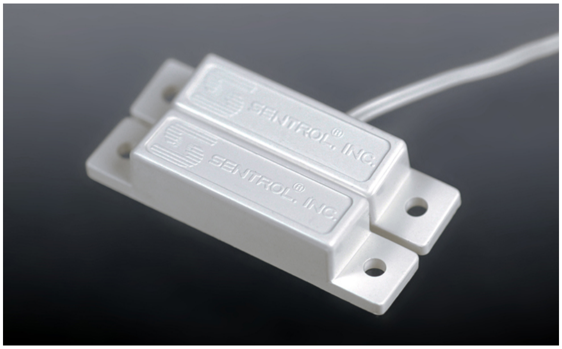

Figure 3-12 *A typical alarm sensor to detect the opening of a door or window. The nearer module contains a magnet, while the module behind it contains a reed switch.*

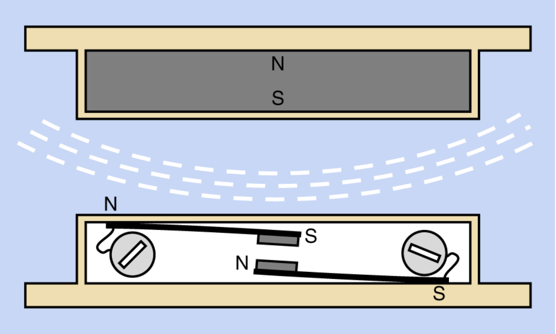

Figure 3-13 *The dashed white lines suggest the magnetic field that closes the contacts of a reed switch.*

So long as the alarmed door or window remains closed, the magnet activates the reed switch. If the door or window is opened, the magnet moves away from the reed switch, and its contacts relax. Usually in this application the reed switch is the normally open type, and is held in its closed state by the magnet. This allows multiple switches to be wired in series, completing a circuit. If a single switch opens, the circuit is broken, and an alarm is triggered.

Reed Switch Variants

Most reed switches are SPST, either normally open or normally closed, although normally open variants are more common. Some switches are SPDT, although this variant is relatively rare. An example of a SPDT reed switch is shown in Figure 3-14.

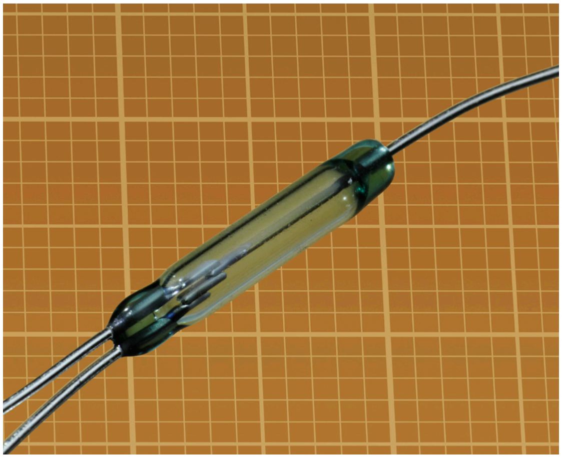

Figure 3-14 *A SPDT reed switch. The background grid is in millimeters.*

The physical size of a reed switch tends to vary roughly in proportion with its current-switching capability.

Reed switches are most commonly supplied with axial leads. A minority are available for surface-mount applications.

Some reed switches are packaged in a plastic pod to provide physical protection for the glass capsule.

Reed Switch Values

The datasheet for a reed switch is likely to contain the following values:

Pull-in: The minimum magnetic field strength required to activate the switch, often measured in ampere-turns.

Drop-out: The maximum magnetic field strength that allows the contacts of the switch to relax, often measured in ampere-turns. Pull-in will be a higher value than drop-out.

Maximum switching current: While a few industrial reed switches can switch as much as 100A, they are uncommon and expensive. A common value for a reed switch about 15mm long is 500mA.

Maximum carry current: If specified, it will be higher than the switching current.

Maximum switched power: Because a reed switch can be used with alternating current, its switching capability may be expressed in watts, or as VA (volts multiplied by amps). 10VA is a common value.

Maximum voltage: While reed switches are most often used with low voltages, some are rated to switch up to 200V.

How to Use a Reed Switch

While an optical object presence sensor may be supplied in a single package containing a light emitter and a light detector, a reed switch always requires an activating magnet that is mounted separately. For reliable operation, the maximum distance between the switch and the magnet is usually restricted to a few millimeters.

The orientation of a magnet that activates a reed switch is not crucial, but will affect the sensitivity of the switch. A manufacturer's datasheet should be consulted for information about optimal magnetic polarity.

As in other mechanical switches, the contacts in reed switches vibrate very briefly when opening or closing. This is known as *contact bounce*, and can be misintepreted as a series of separate signals by digital logic or a microcontroller. *Debouncing* with hardware designed for that purpose, or software (in a microcontroller), may be necessary. See the entry on **switches** in Volume 1.

Hall-Effect Sensor

A Hall-effect sensor responds to a magnetic field by generating a small voltage, usually amplified by a transistor included in the package with the sensor.

When a Hall sensor is in its "off" state (i.e., not triggered by a magnetic field), the sensor creates high resistance between the collector of an internal NPN transistor and negative ground. Consequently, if voltage through a pullup resistor is applied to the collector, the voltage on the collector will be high.

When the sensor is in its "on" state, the resistance drops, the voltage supplied to the collector from the pullup resistor is shunted to ground, and the output voltage appears to go low. As a general rule:

- Activated Hall sensors appear to have a low output.
- Inactive Hall sensors appear to have a high output.

Information about using open-collector outputs is contained in the Appendix. See Figure A-4.

The clean output, reliability, small size, and cheapness of Hall sensors have encouraged their use in devices as different such as hard drives, cameras, keyboards, and automobiles. They are useful in almost any situation where a sensor must detect a mechanical operation at close range. Three through-hole Hall sensors are shown in Figure 3-15. Surface-mount versions are much smaller.

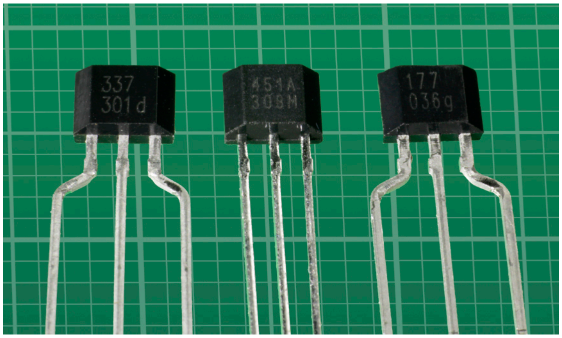

Figure 3-15 *Three through-hole Hall-effect sensors. The background grid is in millimeters.*

How a Hall-Effect Sensor Works

When current is flowing through the length of a conductor, and a magnetic field is applied across the width of the conductor, the field generates a force causing electrons and electron holes to accumulate asymmetrically on opposite sides. This is known as the *Hall effect*.

The voltage difference between the electron-rich and electron-depleted zones is known as the *Hall voltage*. It is proportional to the magnetic field, and inversely proportional to the density of free electrons in the material. For this reason, the Hall effect is easiest to observe in semiconductors, where the density of electrons or electron holes is low.

Hall sensor components contain amplifier circuitry in addition to the sensing element itself. Typically there is an open-collector output, and a comparator or Schmitt trigger to provide some hysteresis.

Hall-Effect Sensor Variants

Four primary variants of Hall sensors are widely used.

Unipolar Hall Sensor

This is activated when an external magnetic field exceeds a threshold value. When the field diminishes, the sensor switches off. Unipolar sensors are available in versions activated by the north magnetic pole or south magnetic pole.

Bipolar Hall Sensor

Proximity to one magnetic pole will switch it on. Proximity to the opposite magnetic pole will switch it off. The sensor remains in its current state (on or off) in the absence of a magnetic field.

Omnipolar Hall Sensor

Proximity to a strong magnetic field of either polarity will switch it on. Removal of the magnetic field will switch it off. An omnipolar sensor can be thought of as a pair of unipolar sensors mounted in opposite directions and with their (open-collector) outputs wired together. This component functions similarly to a reed switch, although of course it still requires a power supply.

Linear Hall Sensor

Also known as an *analog* Hall sensor, its output voltage varies in proportion to an external magnetic field instead of switching cleanly between high and low states. When no magnetic field is detected, the output is half of the sensor's supply voltage. In response to one magnetic polarity, the output can drop almost to zero. The opposite polarity can increase the output almost to supply voltage.

The output from a linear sensor usually is supplied from the emitter of an internal NPN transistor, not the collector. A minimum 2.2K resistor should be connected between output and ground.

The variable output can be interpreted as a measurement of distance between the sensor and a magnet. In this mode, a Hall sensor functions as a **proximity** sensor. However, it is not usually capable of measuring a distance of more than 10mm.

Other Applications

Hall sensors are incorporated in other types of components. A **magnetometer**, for example, may contain Hall sensors.

Additional discussion of Hall-effect sensors, with test circuits, will be found in the book *Make: More Electronics*, from which some of the figures here have been excerpted.

Values

Magnetic field at operating point is the minimum field necessary for the output to switch on. It is measured in tesla or gauss, and the abbreviation B_{OP} is used.

Magnetic field at release point is the maximum field that allows the output to switch off. It is measured in tesla or gauss, and the abbreviation B_{RP} is used.

Magnetic field range may be specified for linear (analog) Hall sensors.

Supply voltage may range as widely as 3VDC to 20VDC, or may be restricted between 3VDC to 5.5VDC. Check datasheets carefully.

Sourcing or sinking capability for the open-collector output is typically 20mA.

How to Use a Hall-Effect Sensor

Hall sensors are often made in 3-pin packages. Through-hole variants are usually made of black plastic and look like TO-92 transistors, but are slightly smaller.

Surface-mounted variants are common.

A typical through-hole Hall sensor has one bevelled face and a flat face on the opposite side. The bevelled face may be referred to as the "front" of the component in a datasheet. The sensor responds when an appropriate magnetic pole is brought close to the front face of the sensor.

The part number printed on the front of the sensor may be abbreviated as three digits. The code below this usually refers to the date of manufacture.

A simple circuit for a Hall sensor resembles a typical circuit for a phototransistor. Positive supply voltage and negative ground are applied to two of the three leads. Positive voltage is also applied, through a pullup resistor, to the third lead, which is the open-collector output (except in the case of a linear Hall sensor, described previously). The output pin is then tapped as the output from the sensor, to be applied to a component that will not draw more than 20mA.

Configuration of Object Presence Sensors

While most of the following suggestions relate specifically to Hall-effect sensors, some general principles may be applied to optical sensors.

Linear Motion

A presence sensor can be activated when the triggering source (such as light or a magnet) approaches it directly. This is sometimes referred to as *head-on mode*. Alternatively, triggering can be arranged when the source moves past the sensor. This is sometimes referred to as *slide-by mode*. The two modes are illustrated in sections 1 and 2 of Figure 3-16.

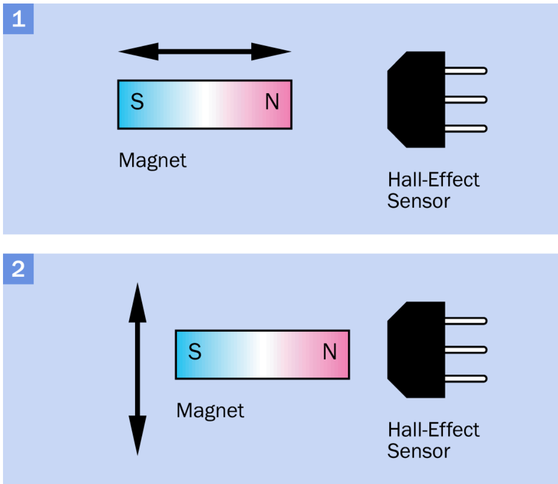

Figure 3-16 *Section 1 of this figure illustrates a presence sensor being used in head-on mode, while section 2 illustrates slide-by mode.*

Slide-by mode may be preferred because it eliminates the risk of damage to the sensor if overshoot occurs in head-on mode.

In slide-by mode, using a bipolar Hall-effect sensor, two magnets can be placed together with opposite polarity, creating a very steep transition in the overall magnetic field. This minimizes the risk of imprecise triggering. Using neodymium magnets, the triggering point can be adjusted with a precision of 0.01mm or better. See Figure 3-17.

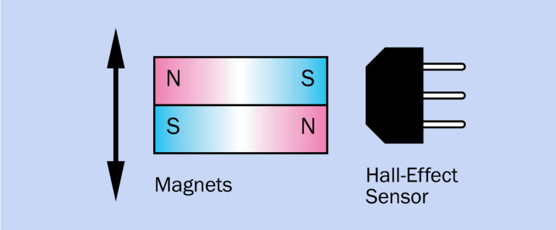

Figure 3-17 *In slide-by mode, two magnets with opposite polarity can be put together to create a very precise transition in a bipolar Hall-effect sensor.*

Sensing by Interruption

An *optointerrupter* is sensitive to an object passing between the light source and the light sensor. A comparable arrangement can be used with a Hall-effect sensor or a reed switch, but only if the interrupting object is thin and ferrous. This configuration is known as a *ferrous vane interruptor*.

Note that the magnet will exert significant force on the ferrous vane. This becomes an issue in sensors where the mechanical force is limited—for example, in paper-path sensors in a photocopy machine.

Additional information about the detection and measurement of moving objects will be found in the entry describing **linear position** sensors. See Chapter 6.

Angular Motion

One or more magnets can be used with a Hall-effect sensor to detect the angular motion, relative angular position, or absolute angular position of a rotating part. This data can be used to determine its speed of rotation. Some techniques to achieve this are described in the entry describing **rotary position** sensors. See Chapter 7.

Sensor Comparisons

Advantages of Optical Presence Sensors

- Not affected significantly by magnetic fields that can interfere with operation of a Hall-effect sensor or a reed switch.
- May be contained in a small, all-in-one package.
- Some optical sensors can operate over a distance of 50cm.
- Very well suited to sensing an object that blocks the light source (the *optointerrupter* configuration).

Disadvantages of Optical Presence Sensors

- Must have clear line-of-sight with the object and/or a reflector.
- Performance will be degraded by accumulation of dust or dirt.
- Limited lifetime of LED light source, if used continuously.
- May be accidentally triggered or impaired by some types of ambient light.
- Often requires a load resistor for the LED as well as a pullup resistor for the open-collector output.
- Range of acceptable voltage is usually narrow.

Advantages of a Reed Switch

- No polarity.
- No additional components required, other than a magnet.
- Able to switch AC or DC.
- Able to switch to voltages as high as 200V in some cases.
- Can be maintained by a magnet in open or closed state indefinitely without any power consumption.
- Many variants are capable of switching 500mA, and higher-current types are available.
- Can be activated through nonmagnetic materials (plastic, paper).
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.

Disadvantages of a Reed Switch

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Cannot be miniaturized to the same extent as a surface-mount chip.
- Glass envelope is easily damaged.
- Arcing may occur between contacts.
- Will not work reliably when the activating magnet is more than a few millimeters from the switch.
- Can be activated accidentally by other magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.
- Must be debounced when connected with a logic chip or controller.

Advantages of a Hall Effect Sensor

- Robust solid-state component.
- Can be miniaturized for surface-mount applications.
- Very low cost.
- Fast response.
- No contact bounce.

- Extremely durable, with an almost unlimited lifetime.
- Not significantly affected by dust or dirt that can degrade performance of an optical switch.
- Low cost.

Disadvantages of a Hall Effect Sensor

- Requires a separate magnet (which must be placed carefully to avoid affecting other components).
- Open-collector output typically limited to 20mA or less.
- May be vulnerable to magnetic fields.
- When sensing an object that comes between the switch and a magnet, only a ferrous object can be used.

What Can Go Wrong

Optical Sensor Issues

Deterioration of LEDs

Most object presence sensors depend on infrared LEDs as a light source. An LED has many virtues (discussed in Volume 2), but does tend to suffer from gradual reduction in light output over a period of years. In a device such as a photocopy machine, which is used intermittently and may have an “economy” or “sleep” mode in which most of its components are powered down, LED-based detectors should last almost indefinitely. In other applications where an LED is powered continuously, its light output will diminish signicantly in 3 to 5 years. Bearing this in mind, an optical sensor should be chosen to operate well below its light-detection limits.

Object Too Close

Some optical and ultrasonic detectors triangulate on an object; that is, the light emitter and light sensor are angled slightly toward each other (as shown in sections 3 and 4 of Figure 3-3). The signal output from the sensor will peak at the distance where the emitter and sensor focus at a point. Consequently the output voltage will diminish when the object comes closer, which may create the misleading impression that the object is moving further away. To avoid spurious readings, detectors should not be used with objects that are closer than the minimum distance specified by the manufacturer.

Reed Switch Issues

Mechanical Damage

Bending the axial leads on a reed switch can easily fracture the glass envelope. Reed switches must be handled with care.

Contact Bounce

If the switch is wired to the input of a logic chip or microcontroller, contact bounce when the switch opens or closes is likely to be misinterpreted as multiple switching events. Debouncing will be necessary, either by additional components or by insertion of a momentary delay in the code embedded in a microcontroller.

Arcing

When switching high voltages or currents, an arc may be created briefly between the switch contacts, most often when they are moving from the closed to the open state. The arc erodes the contacts of the switch. Inductive loads make the arcing problem worse. If the switched voltage is kept below 5V arcing generally does not occur, extending the life of the switch.

passive infrared sensor

4

In coloquial speech, the term *motion sensor* is usually understood to mean a **passive infrared** motion sensor.

The acronym *PIR* is often used for a passive infrared sensor. It is always capitalized, without periods.

Object presence sensors and **proximity** sensors require an active source of a magnetic field, ultrasound, or infrared radiation. A passive infrared sensor does not require any such source, and responds passively to heat radiated from the object that is being detected.

OTHER RELATED COMPONENTS

- **object presence** sensor (see Chapter 3)
- **proximity** sensor (see Chapter 5)

What It Does

A **passive infrared** motion sensor, often described as a *PIR*, detects *black-body radiation*, which all objects emit as a function of their temperature relative to absolute zero. The sensor responds to infrared radiation centered around a wavelength of 10µm (10 microns, or 10,000nm). This is the approximate body temperature of people and animals.

The word “passive” in the term “passive infrared” refers to the behavior of the detector, which receives infrared radiation passively. **Proximity** sensors must generate their own infrared radiation actively, which is interrupted or reflected by nearby objects. See Chapter 5.

Schematic Symbols

Schematic symbols that are sometimes used to represent a passive infrared motion sensor are shown in Figure 4-1.

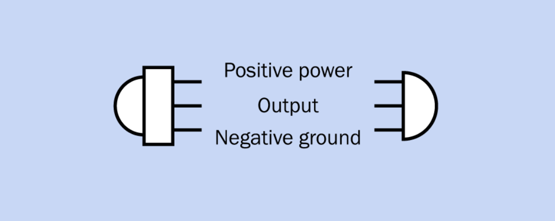

Figure 4-1 *Schematic symbols that may represent a passive infrared motion sensor. The orientation (pointing left or right) is arbitary. The pin sequence may vary.*

Applications

Motion-sensitive outdoor lighting almost always is based around a PIR. Similarly, a security system may sound an alarm or activate a video camera when a PIR indicates human activity.

Wildlife monitoring systems use PIRs to start a video camera that can then run for a preset interval.

Warning systems for automobiles have been developed that use a rear-facing PIR to detect pedestrians.

Industrial indoor lighting may use PIRs that switch the lights on automatically when people enter a room, and then switch the lights off (after a timed delay) when people are no longer detected in the room. The goal is to prevent wastage of energy as a result of employees forgetting to switch the lights off.

How It Works

A PIR module contains multiple components. Most visible is an array of at least 15 small lenses that focus infrared light from zones in the environment onto a *pyroelectric detector*, also known as a *pyroelectric sensor*. The response of the detector is processed by an amplifier, so that the signal can trigger an electromechanical **relay** or **solid-state relay** (see Volume 2). The relay operates an external device such as a light or an alarm.

Additional circuitry may allow the user to control the sensitivity of the PIR module and the length of time that the relay remains closed. The user may also be able to set the time of day when the PIR is active, or an additional **phototransistor** can shut down the PIR during daylight hours. If a phototransistor is included, its sensitivity is adjustable. For more information about phototransistors, see Chapter 22.

Pyroelectric Detector

The pyroelectric detector is actually a type of *piezoelectric* device. It is based around a wafer of lithium tantalate, which generates a small voltage in response to incident thermal radiation. However, like other piezoelectric components, it does not respond to a steady-state input, and must be activated by a transition. This distinguishes it from other types of light sensors, such as an infrared **photodiode**, in which the response is consistently related to a temperature input.

The response of a pyroelectric detector is suggested by the graphs in Figure 4-2.

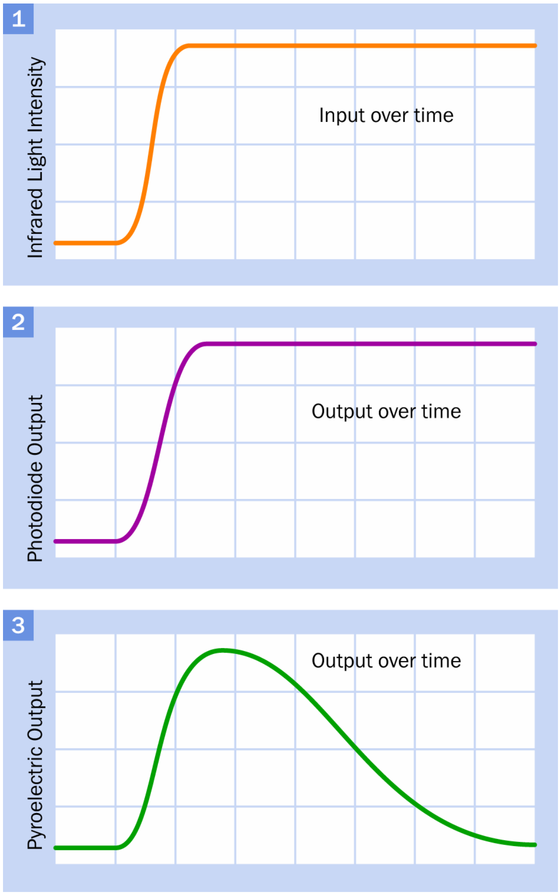

Figure 4-2 *Top: incident infrared light intensity. Center: the voltage output from a hypothetical photodiode. Bottom: the voltage output from a hypothetical pyroelectric detector.*

A pyroelectric detector in a PIR module is mounted in a sealed metallic container, as shown in Figure 4-3. The rectangular window in the detector is usually made of silicon, which is opaque to visible wavelengths but transparent to long-wave infrared radiation.

Elements

The pyroelectric detector used in a PIR contains at least two *elements* with opposite polarities, connected in series. If a sudden change in temperature affects both elements equally, their responses will cancel each other out. Thus, the detector ignores changes in ambient temperature. However, if a source of infrared radiation in the appropriate waveband affects one element before the other, the detector will emit two pulses of opposite polarity. See Figure 4-4.

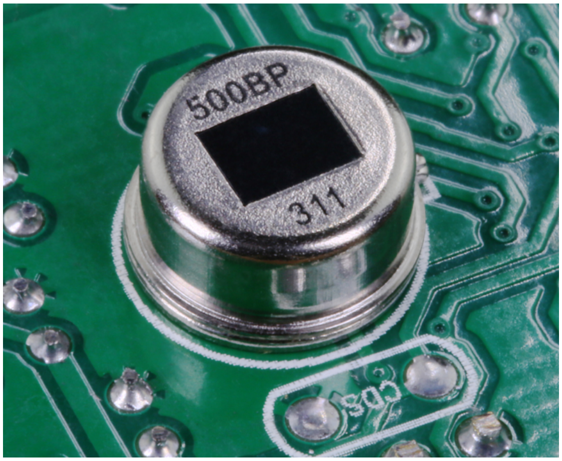

Figure 4-3 *A pyroelectric detector mounted on a small circuit board in a passive infrared radiation sensor.*

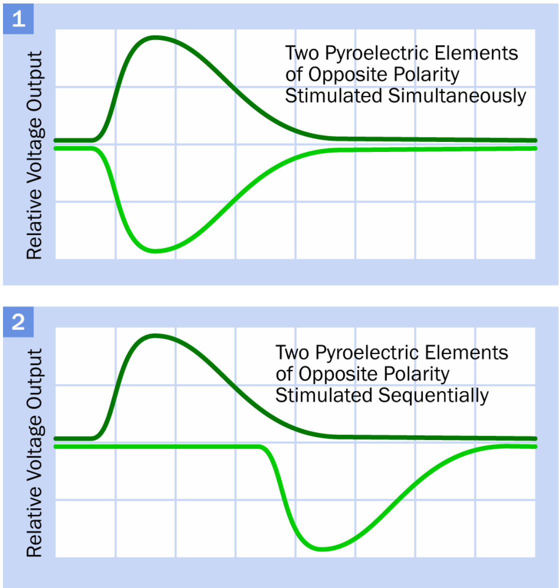

Figure 4-4 *Top: in a pyroelectric detector, if a change in temperature affects two elements of opposite polarity simultaneously, their voltages cancel each other out. Bottom: if one element is triggered before the other, the detector emits a signal.*

Lenses

Lenses are used to trigger the elements sequentially. Each lens faces toward one visible zone in the target area. When a source of infrared radiation moves from one zone to the next, it energizes the elements alternately, creating an output.

In some PIRs, four pyroelectric elements are used instead of two, to provide better coverage. The pairs of sensors may be wired in series or in parallel, but the principle remains the same.

The lenses are molded into a polyethylene dome that is often white and covers the pyroelectric detector. The dome appears smooth on the outside, but fine patterns of concentric ridges are molded on the inside. These are *fresnel lenses*, which are much cheaper, smaller, lighter, and easier to fabricate than conventional optical lenses. A fresnel lens introduces some distortion and aberration, but these defects are unimportant in a PIR.

Figure 4-5 illustrates the principle of a simple fresnel lens. The first section of this figure shows a conventional optical lens with one flat side and one curved side. A distant object emits almost-parallel beams of infrared light that are focused by the lens. Section 2 shows the same lens divided into segments that are stacked with no space between them. They behave in exactly the same way as the original lens. In section 3, each segment has been reduced in width, but because the optical faces still have the same geometry, they will still have the same function, although a small amount of distortion will be added by the reduction in width. This is a fresnel lens, which found an early application in lighthouses, where it greatly reduced the weight of very large glass lenses that focused the beam.

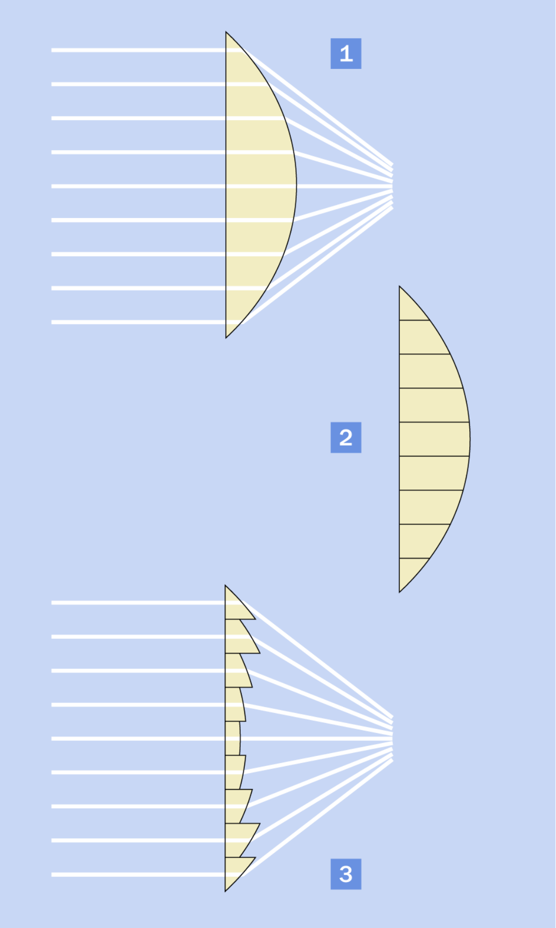

Figure 4-5 *Principle of a fresnel lens. See text for details.*

The same principle can be applied to a lens in which both surfaces are curved, as shown in Figure 4-6. In practice this will introduce more imperfections in the image, although some compensation is possible by adjusting the exact shape of the lens.

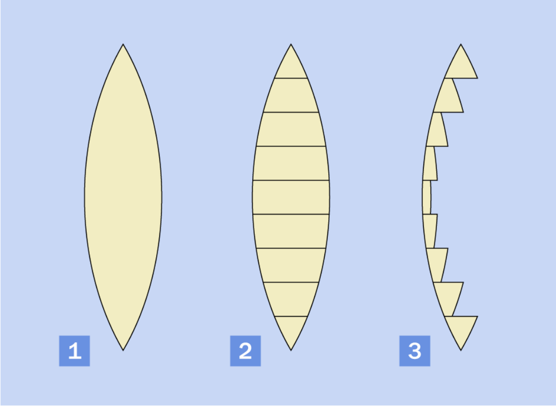

Figure 4-6 *The principle of a fresnel lens applied to a conventional lens in which both surfaces are curved.*

Figure 4-7 shows three curved fresnel lenses placed edge-to-edge, seen from above. In section 1 of this figure, infrared rays from a distant source are focused by the first lens onto the right-hand element of a pyroelectric sensor. In section 2, the external source has moved laterally, and the rays now focus on the left-hand element. In section 3, the source has moved into the zone covered by the centrally located fresnel lens, which focuses on the right-hand element again. The fluctuating inputs will trigger the sensor.

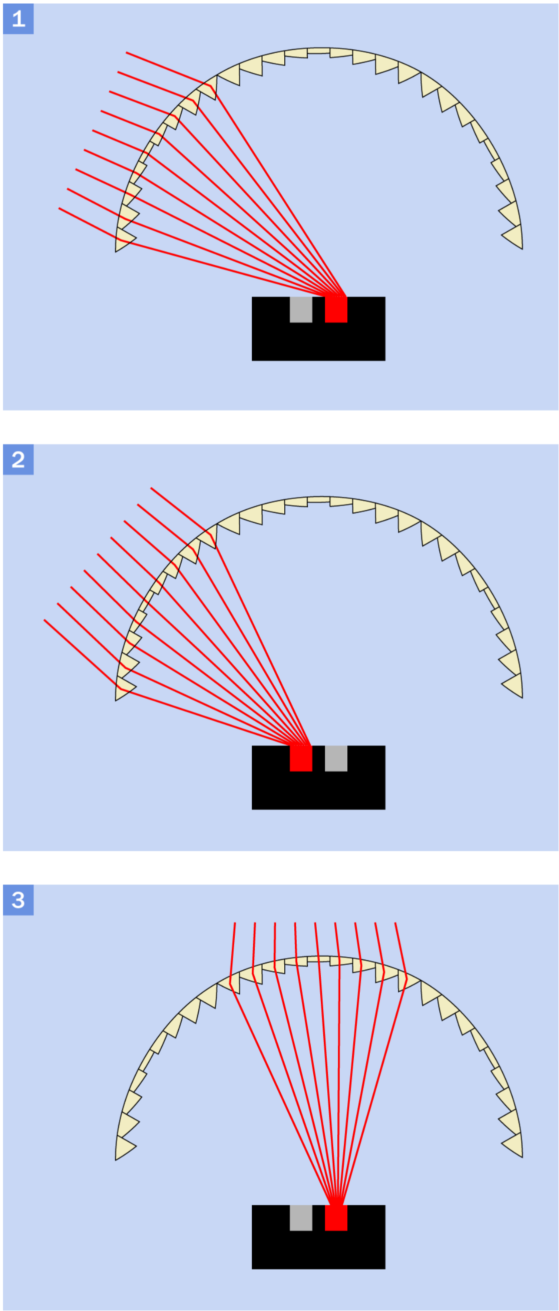

Figure 4-7 *Sections 1, 2, and 3 show fresnel lenses focusing an external source of infrared radiation on individual elements of a pyroelectric sensor.*

PIRs may combine fresnel lenses in a variety of patterns. Figure 4-8 shows an evenly weighted mosaic that would be suitable for a motion detector mounted on a ceiling, facing directly downward. Figure 4-9 shows a pattern weighted toward lateral motion, providing less sensitivity for motion above and below the primary band.

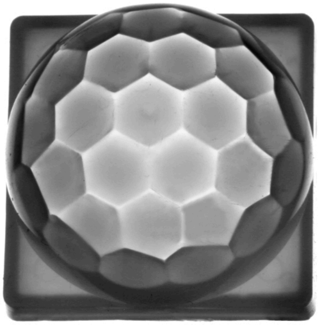

Figure 4-8 *An evenly weighted mosaic of fresnel lenses.*

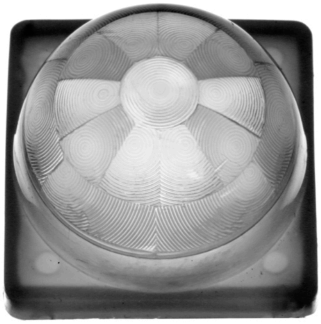

Figure 4-9 *A mosaic weighted more toward sensing lateral motion. The grooves of the fresnel lenses are visible.*

Variants

PIR sensor modules are available mounted on a small board such as the one shown in Figure 4-10 from Parallax, Inc. The detection range is 5 to 10 meters, selected by a jumper on the board. The three pins visible in the photograph are for power supply (3VDC to 6VDC), ground, and output. The output can source up to 23mA with a 5VDC power supply. Power consumption of the module is only 130µA when it is idle, or 3mA when it is active but has no load.

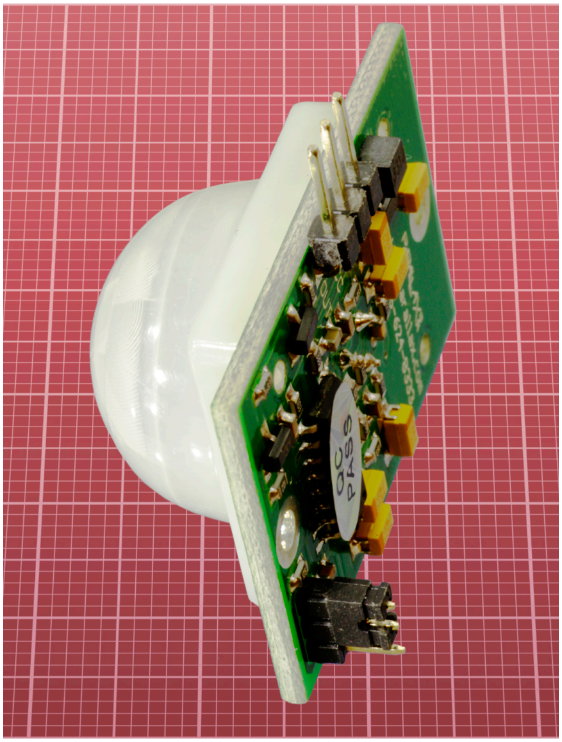

Figure 4-10 *A passive infrared detector mounted with basic necessary components on a small board.*

A board of this type still requires additional components to set the "on" time for a light or alarm, and to deactivate the PIR during daylight hours.

Various lens patterns are available, sold separately.

A PIR can be bought as a single component containing two elements and FET transistors to amplify the signal. Surface-mount and through-hole versions are available, requiring a typical power supply of 3VDC to 15VDC.

However, a PIR bought as a "bare" component requires significant external circuitry, using comparators or op-amps. Circuit design is nontrivial, entailing practical problems such as op-amps being sensitive to voltage spikes caused by activating a relay that shares the same power supply.

An easier alternative is an all-in-one detector, lens, and control circuit such as the Panasonic AMN31111, which is ready for board mounting. Its small output current of 100µA would be capable of activating a solid-state relay. Similar Panasonic PIRs offer a variety of ranges, sensitivities, and supply voltages.

The AMN31111 is in Panasonic's AMN series. There are many type numbers for combinations of analog or digital output, lens shape, and black or white lens. A selection of lens shapes, derived from the manufacturer's datasheet, is shown in Figure 4-11.

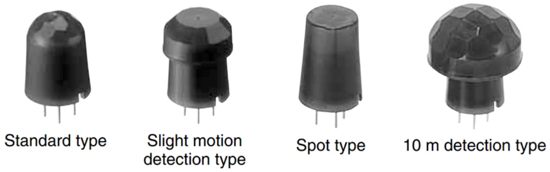

Figure 4-11 *Four lenses from the Panasonic AMN series of passive infrared sensors.*

What Can Go Wrong

Temperature Sensitivity

In warmer weather, objects in the field of view of a PIR will tend to be warmer, and the temperature difference between them and human skin will diminish. This can degrade the performance of a PIR.

Detector Window Vulnerability

The silicon window on a detector is vulnerable to dirt or grease. Avoid touching the component if it is not protected by lenses.

Moisture Vulnerability

Water absorbs far-infrared light. Consequently, condensation on the lens or detector can degrade performance, and a PIR may not function in heavy rain or snow.

5

proximity sensor

Proximity sensors that use infrared, ultrasound, and capacitance are described here. This Encyclopedia does not include proximity sensors that use magnetism, inductance, or other methods of determining distance.

Sometimes proximity sensors are referred to as *distance sensors*. An ultrasonic proximity sensor may be described as a *range finder* or a *ranger*.

High-end ultrasound proximity sensors sold as large, sealed modules with cabling can monitor the status of industrial processes. These commercial units are outside the scope of the Encyclopedia.

A sensor that detects whether an object is present, but does not measure the distance to it, is considered to be an **object presence** sensor and has its own separate entry. See Chapter 3. Many devices that are sold as proximity sensors or distance sensors actually do not provide meaningful distance data, and therefore in this Encyclopedia they are included with presence sensors.

Phototransistors and **photodiodes** may be used as sensing elements in proximity sensors. These components have their own entries as light sensors. See Chapter 22 and Chapter 21.

OTHER RELATED COMPONENTS

- **object presence** sensor (see Chapter 3)
- **passive infrared** sensor (see Chapter 4)

What It Does

A proximity sensor measures the distance between itself and a physical object that is often described as the *target*. The output from the sensor may be analog (voltage), serial data, or pulse-width modulation. It may be transmitted via a serial protocol such as SPI, TTL, or I2C, and may be stored as digital data in a register that is accessed by a microcontroller, using I2C. For additional details about protocol, see Appendix A.

Schematic Symbols

Either of the schematic symbols in Figure 5-1 may represent a proximity sensor, but are not used consistently. The sensor may also be represented as a rectangle containing text describing its function.

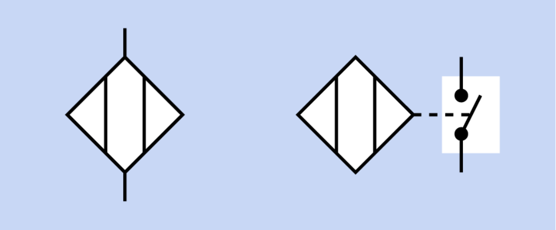

Figure 5-1 *Two ways to represent a proximity sensor in a schematic.*

Applications

In robotics, a proximity sensor can prevent a collision between a robot and an object or barrier in front of it. Some sophisticated proximity sensors can provide sufficient data for software to map the entire environment, but are outside the scope of this entry.

Proximity sensors can also be used in alarm systems, or for liquid-level sensing in storage tanks, or in automobiles to sound a warning if the driver is backing into an obstacle (although these devices are being augmented with rear-view video monitors).

In handheld devices, proximity sensors are used to sense the presence of the user's hand or face —for example, to shut down the display when a person raises a phone to talk into it.

Variants

This entry is subdivided to describe proximity sensors that use ultrasound, infrared light, and capacitance.

Ultrasound

An ultrasonic proximity sensor functions by emitting a short burst of sound and then listening for echoes from objects in front of it.

The sound is created by a **piezoelectric transducer** (see Volume 2) at a frequency between 30kHz and 50kHz, well above the range that can be detected by the human ear. The transducer may do dual duty as a microphone, sending and receiving sound on an alternating basis; or a second transducer, serving as a microphone, may be mounted beside the emitter on a small circuit board. The economically priced HC-SR04 is an ultrasound proximity detector popular in the robotics community, working reliably in the range of 2cm to 5m.

A board on which the sensor is mounted may include a microcontroller to measure the delay between propagation of a pulse and reception of the echo. Distance to the reflecting object is then calculated using the speed of sound in air at sea level, which is approximately 340 meters per second.

Infrared

An infrared proximity sensor requires a beam of infrared light from an LED that may be incorporated in the sensing module or mounted separately. Light reflects from the target and is detected by a **phototransistor** or **photodiode**. From the angle of the reflected light, onboard electronics can calculate the distance to the target by a process known as *triangulation*. See Figure 5-2. (This diagram is simplified. An actual sensor may use a linear array of photodiodes to assess the angle of the returning light beam.)

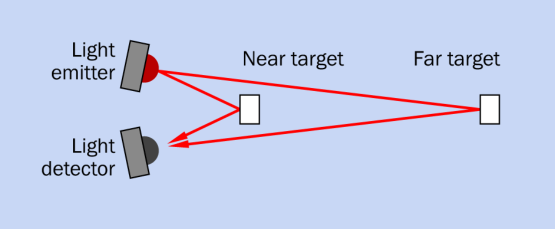

Figure 5-2 *An infrared proximity sensor determines the distance to an object by assessing the angle of reflected light.*

To reduce the risk of false positives, light from the LED contains only a narrow range of infrared wavelengths. Also, it is modulated at a frequency recognized by sensing circuitry in the module.

Relative Advantages

Ultrasound Devices

- Generally more suitable for detecting objects that are more than 1 meter away.
- Unaffected by direct sunlight, fluorescent tubes, and other light sources that can interfere with infrared devices.
- More accurate, often capable of placing objects within 5mm.
- Able to measure the distance to liquids and transparent objects, which may not be easily detected with infrared.

Infrared Devices

- Physically smaller—especially surface-mount versions.
- Able to measure the distance to soft objects, which may not be easily detected by ultrasound.
- More appropriate for targets that are closer than 10mm.
- More affordable.

Ultrasonic Examples

The proximity sensor in Figure 5-3 is a low-priced model, popular in the robotics community. Manufactured by MaxSonar, it uses a single element to send and receive. The manufacturer claims that it can detect a 6mm (quarter-inch) dowel 1.8 meters directly in front of it, and a 9cm dowel at 3.3 meters. This performance is beyond the capability of almost all infrared sensors.

MaxSonar offers a variety of sensors that appear physically very similar but have different range capabilities. Each has three outputs that can be accessed simultaneously: serial data at 9,600bps, analog voltage, and pulse-width modulation.

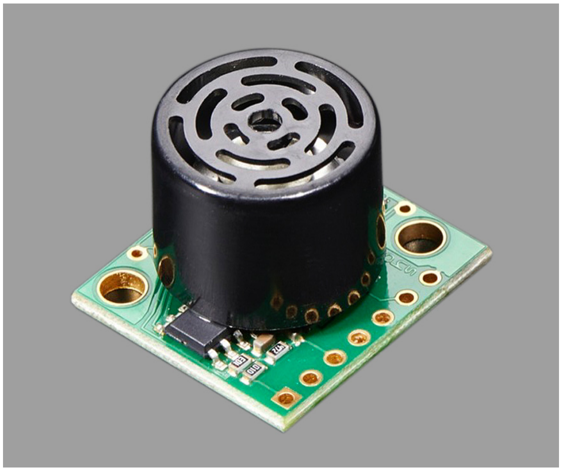

Figure 5-3 *The MaxSonar MB1003 can detect large solid objects as far away as 5 meters. Photo derived from an image by Adafruit.*

The serial output uses the RS232 protocol and consists of the letter R followed by four ASCII-coded numerals representing the measured range in millimeters. Thus, R1000 would indicate an object at 1 meter distance.

The analog voltage ranges linearly from 293mV when sensing an object at 300mm to 4,885mV for an object at 5,000mm.

The pulse-width output sends pulses ranging from 300µs indicating an object 300mm away to 5,000µs indicating an object 5 meters away.

The unit incorporates a temperature sensor that compensates for the lower density of air when its temperature increases. Weatherproofed versions are available. The power supply is 5VDC and must be smoothed.

Imports

Some international sources offer minimal paired ultrasound components at a very low price, such as the HC-SR04 from Cytron Technologies in Malaysia. See Figure 5-4.

Sound dispersion of the transducer is claimed to be plus-or-minus 15 degrees, and the claimed range is up to 4 meters. The module requires a triggering input pulse that must last at least 10µs. This prompts it to emit eight rapid ultrasound signals at 40kHz. The module measures the response time and applies a high state to its echo pin for a duration that is proportional with the distance measured. An external microcontroller must time the duration, then divide by a factor of 58 to obtain distance in centimeters.

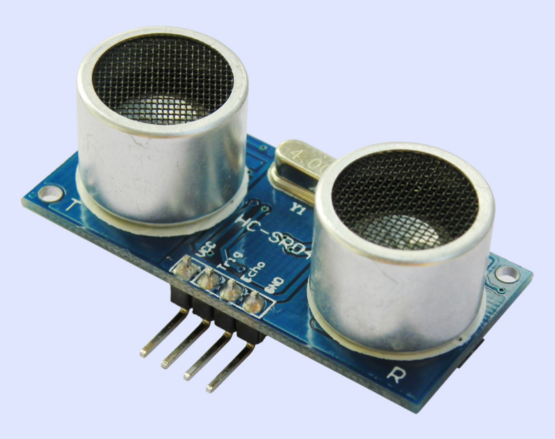

Figure 5-4 *The HC-SR04 is a very low-cost import that can provide acceptable performance when used with an external microcontroller.*

Many online sources offer simple code libraries for Arduino or PICAXE microcontrollers for use in conjunction with the HC-SR04.

Individual Elements

Individual ultrasonic components such as those shown in Figure 5-5 are available from many vendors. The user must add circuitry to generate a high frequency for a short duration, amplify the microphone signal, measure the time difference, and calculate distance.

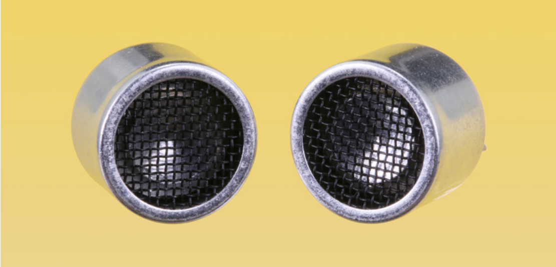

Figure 5-5 *These components, listed as the "40TR12B-R ultrasonic sensor kit" by the online supplier Jameco, could form the basis of a DIY ultrasonic proximity sensor.*

Infrared Examples

Sharp manufactures four infrared proximity sensors that are widely regarded as accurate and easy to use. They are popular in the robotics community. Their part numbers and ranges are:

- GP2Y0A51SK0F (20mm to 150mm)
- GP2Y0A21YK0F (10cm to 80cm)
- GP2Y0A02YK0F (20cm to 150cm)
- GP2Y0A60SZLF (10cm to 150cm).

The GP2Y0A60SZLF is the most recent product, with the most impressive specification. The GP2Y0A21YK0F is shown in Figure 5-6.

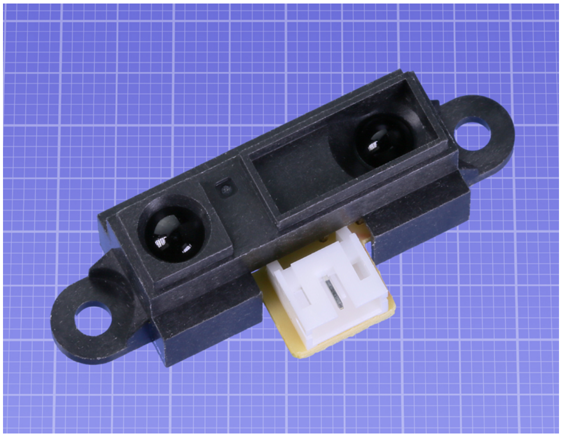

Figure 5-6 *The Sharp GP2Y0A21YK0F infrared proximity sensor. The background grid is in millimeters.*

Sharp describes these sensors as having an *analog output*. The voltage on an output pin varies in inverse proportion with the distance being measured. The relationship is illustrated by the graph in Figure 5-7, derived from the manufacturer's datasheet for model GP2Y0A02YK0F.

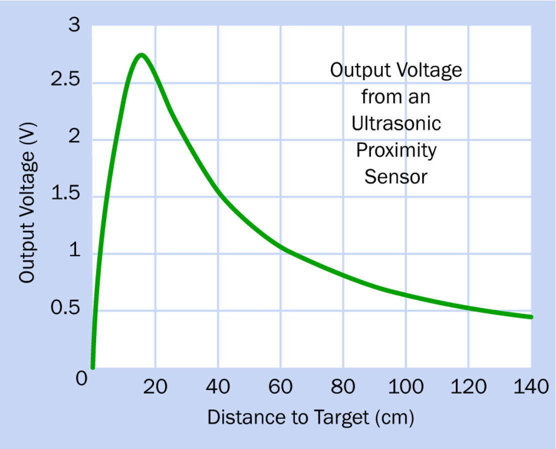

Figure 5-7 *Relationship between output voltage and target distance for a Sharp infrared proximity sensor, derived from the manufacturer's datasheet.*

The Sharp sensors can work with 5VDC. They consume around 30mA, except for the GP2Y0A60SZLF, which uses less current. Because the infrared LED functions in bursts, the manufacturer recommends protecting other components sharing the power supply by placing a 10µF capacitor across the sensor supply pins.

Trends in Infrared Proximity Sensing

Like many other types of sensors, proximity sensors have been affected by the huge market for handheld devices.

Handheld applications have had four consequences:

Miniaturization

> Infrared proximity sensors are now commonly found in surface-mount chips measuring 5mm × 3mm or smaller.

Onboard processing

> The status of a photodiode can be processed by a microcontroller on the same chip, to determine what the sensor is really "seeing." Input from an included ambient light sensor is factored into the evaluation.

Cost reduction

> While chip-based proximity sensors have become increasingly sophisticated, their unit cost has plummeted, so that they are now actually much cheaper than simpler devices such as the Sharp analog sensors described above.

Complexity

> Modern sensors must be programmed with a complex variety of instructions, and their coded output must be interpreted with a separate microcontroller running its own program. Whether this compensates for the low price and added functionality is something for the individual developer or experimenter to decide.

The Silicon Labs Si1145/46/47 series are chips with the kind of sophisticated capabilities required for handheld devices. An external microcontroller communicates with the sensor via an I2C connection, and can instruct it to adjust its distance range (from 1cm to more than 50cm), its analog-to-digital conversion sensitivity, and its current-sinking capability for up to 3 external LEDs. The chip incorporates ultraviolet sensing and ambient-light sensing capability. Its I2C connection can run at up to 3.4Mbps. Because its light output is pulsed for only 25.6µs every 800ms at 180mA, the average power consumption is only 9µA, assuming a supply voltage of 3.3VDC.

For additional details about protocols such as I2C, see Appendix A.

In addition to its use in handheld devices, the manufacturer suggests applications including heart-rate monitoring, pulse oximetry, and display backlighting control. These applications require only a subset of the features built into the sensor, but its price is so low, it may be cost-effective even when many of its capabilities are unused.

Proximity sensors with comparable specifications are available from many manufacturers. Examples are the Vishay VCNL4040 and the Avago HSDL-9100. Figure 5-8 shows the Silicon Labs SI1145 on the left, and the Avago HSDL-9100 on the right.

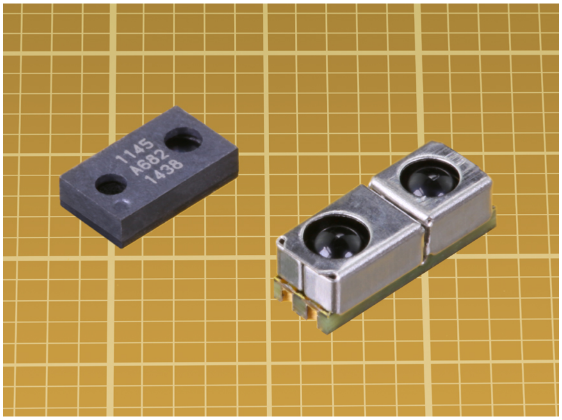

Figure 5-8 *Two sophisticated surface-mount proximity sensors with digital output. Left: Silicon Labs SI1145. Right: Avago HSDL-9100. The background grid is in millimeters.*

Several guidelines must be observed when using this kind of sensor. First and most obviously, if an external LED is used, it must have a peak wavelength compatible with the photodiode in the sensor. The LED must be placed as near as possible to the photodiode, as reducing the separation increases the sensitivity; but crosstalk between the photodiode and LED must be minimized, usually by placing a thin barrier between them that is barely taller than the higher component.

If the light emitter and/or light sensor are protected behind a transparent glass or plastic panel, it must have minimal resistance to infrared wavelengths, and its thickness must be chosen in compliance with guidelines in the sensor datasheet. To prevent light from reaching the sensor by reflection from the rear of the panel, a thin, opaque tube can be installed between the LED and the panel.

A sensor of this type may be configurable with a "detection scheme," meaning high and low sensing threshold levels appropriate to the object that is likely to be detected. Determining this may be a process of trial and error.

Capacitive Displacement Sensor

This is also known as a *capacitive linear displacement sensor*. It should not be confused with a capacitive **single touch** sensor, a human-input device that has its own entry. See Chapter 13.

A **capacitive displacement** sensor measures the distance between itself and a target that must be electrically conductive. Unlike optical or ultrasound position sensors, no additional source of light, sound, or other radiation is required. Unlike a magnetic position sensor, it does not require a separate permanent magnet. It simply measures its electrical capacitance with the target.

High-precision capacitive displacement sensors are used mainly for industrial process management. Lower-precision variants are much less expensive and can be used as **object presence** sensors to determine if an object is anywhere within a specified range.

A typical maximum range would be 10mm. For larger distances, optical and ultrasound sensors are more appropriate.

A selection of cylindrical sensor probes is shown in Figure 5-9.

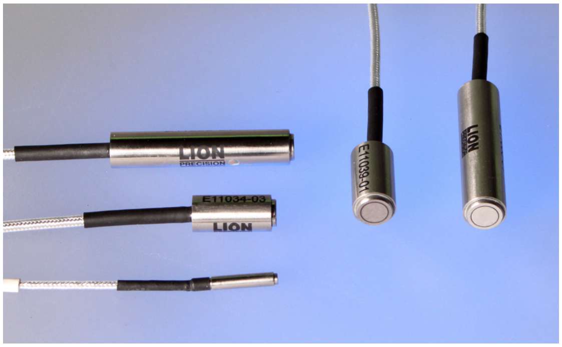

Figure 5-9 *Some high-precision capacitive displacement sensors from Lion, which builds them into cylindrical probes ranging from 3mm to 18mm in diameter.*

Applications

The high-precision variant of this type of sensor is commonly used during the production of small devices, such as disk drives. It can also measure vibration of a rotating metal part, such as a motor shaft, and may maintain automatic focus of a microscope.

Lower-precision variants can be used for applications such as counting objects on a conveyor.

When used to measure materials thickness, the sensor finds applications in checking automobile brake rotor fabrication and the thickness of silicon wafers.

How It Works

When two plates of electrically conductive material are placed opposite each other, they possess capacitance. This means that they have an electrical storage capacity enabled by accumulation of opposite charges on each plate.

The capacity is directly proportional to the plate area, inversely proportional to the distance between the plates, and is affected by the medium separating the plates, known as the *dielectric*.

If the plate area and dielectric remain constant, the distance between the plates will be the only factor affecting the capacitance. Therefore, the distance between the plates can be calculated by measuring the capacitance.

Measurement can be performed by evaluating the *displacement current* that passes through the dielectric from one plate to another when a pulse of voltage is applied. (Hence the term “capacitive displacement” in the name of this type of sensor.)

A detailed explanation of displacement current is included in the second edition of *Make: Electronics*.

The sensor itself functions as one plate of the capacitor while the target functions as the opposite plate. Alternating current is applied as a rapid series of pulses, and the current that passes between the plates is proportional to the distance between them.

Ideally, the target should be grounded to the current source. However, since AC is being used, capacitive coupling of the target to the current source is also possible, so long as the additional capacitor has a value of 0.1µF or higher.

Sources of Error

To obtain a meaningful measurement, the electric field from the sensor is focused on the target. Some dispersion will still occur, and the combination of sensor and target must be compatible. A target typically should be flat and should have a larger surface area than the sensor.

Humidity can affect sensor performance, as it changes the value of the dielectric. Temperature can affect performance, partly because it causes small dimensional variations in the sensor and the target.

The surface of the sensor and the surface of the target should be precisely parallel, because the spot where the field hits the target will be elongated if the target is tilted relative to the sensor. Elongation increases the capacitive area and affects the accuracy of measurement.

This type of sensor can also measure the thickness of a nonconductive material, if the material is a thin sheet that can be interposed between two sensors. In this mode the material functions as a dielectric, and its thickness will affect the AC current passing through it.

While lower-precision capacitive displacement sensors are relatively unusual, they can potentially serve as a relatively affordable and simple **object presence** sensor, so long as the target is conductive and will not be damaged by passing a small alternating current at a relatively low voltage.

Values

A high-precision capacitive displacement sensor can measure distances usually ranging from 0.25mm to 10mm with an accuracy that can be better than 0.05mm. High voltages are not required, with a supply of plus-or-minus 15V being common.

The sensing element, often referred to as a *probe*, is usually plugged into a custom-designed control unit that converts the capacitance measurement into a variable output voltage. The performance is then expressed in millimeters per volt. Thus if the voltage varies by 5V over a distance of 1mm, the sensor is rated as providing 0.2mm per volt.

What Can Go Wrong with Optical and Ultrasound Proximity Sensors

Object Too Close

Because both types of proximity sensors (ultrasonic and infrared) may include emitters that are angled toward the distance range for which they are designed, they may fail to “see” an object that is closer. Consequently the sensor will provide no response, or may detect a different object that is further away. In either case, if the sensor is being used on a moving device, the device may collide with the undetected near object.

Multiple Signals

If two or more sensors and emitters are used concurrently, their combined signals can interfere with each other and create inaccurate readings.

Inappropriate Surfaces

Ultrasonic proximity sensors are intended to identify a single object that is closer to the sensor than other objects, within a narrow beam dispersion angle. Multiple objects, complex surfaces, soft surfaces such as clothing or furnishings, or an unusual configuration of interior walls can create inaccurate readings.

Infrared sensors may be unable to “see” liquids or transparent objects, and may give different assessments of distance depending on the properties of a surface. Human skin, for example, is a poor reflector, as it absorbs some infrared radiation.

Environmental Factors

An ultrasonic transducer uses a very small moving diaphragm to generate sound. Like any system containing moving parts, it will be vulnerable to moisture or excessive humidity, and may need to be protected.

After a device is built and tested indoors, in a controlled environment, it is likely to behave differently if it is moved outside where the temperature is significantly higher or lower.

Deterioration of LEDs

As noted in the section on presence sensors (see Chapter 3), LEDs tend to suffer from gradual reduction in light output over time. The performance of an infrared proximity sensor may deteriorate over a period of years, depending on how intensively it is used.

linear position sensor

A **linear position** sensor may also be described as a *linear displacement sensor* or a *linear position transducer*. It is sometimes categorized as a type of **proximity** sensor, but in this Encyclopedia a proximity sensor emits a signal and receives an echo to measure the distance from itself to an object. By comparison, a linear position sensor measures the position of a sliding object within a stationary enclosure.

An **object presence** sensor can be considered as a form of linear position sensor, but it only responds to the presence of an object, without measuring its position.

OTHER RELATED COMPONENTS

- **proximity** sensor (see Chapter 5)
- **object presence** sensor (see Chapter 3)
- **rotary position** sensor (see Chapter 7)

What It Does

Control of a mechanical device may require accurate and timely information about the position of a movable part in the device. A **linear position** sensor can be used for this purpose.

Three attributes are likely to be of interest:

- Position
- Direction of motion
- Speed of motion

Typically a linear position sensor measures only the first attribute. Additional electronics may calculate the second and third attributes by taking multiple position readings. Thus a *speed sensor* is very likely to be built around a position sensor, and therefore this Encyclopedia does not have a separate entry for speed sensors.

Applications

Robotic arm positioning, wing flap and rudder positions on an aircraft, computer-controlled machine tools, 3D printers, and automobile seat position sensors are some of the many applications.

Schematic Symbol

In a schematic, a linear position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, or others).

How It Works

Linear potentiometers, *magnetic linear encoders*, *optical linear encoders*, and *linear variable differential transformers (LVDTs)* are described here. Other options are available, but they tend to be more specialized, and are not included.

Linear Potentiometer

For a full description of **potentiometers**, see Volume 1.

A linear potentiometer, often referred to as a *slider potentiometer*, contains an electrical resistance in the form of a straight section of *track*. The track may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied across its full length, as shown in Figure 6-1. A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper's position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

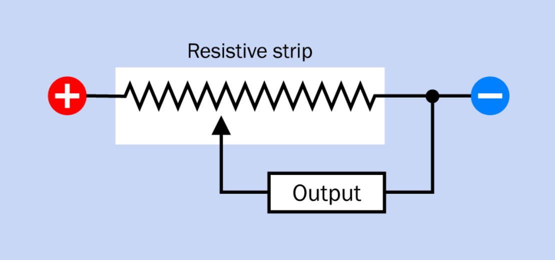

Figure 6-1 *A linear potentiometer can consist of a fixed track with a known voltage applied at each end, and a movable wiper.*

For audio applications, a slider potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

For sensing purposes, the potentiometer is usually protected by a long, narrow box or tube through which a rod slides on sealed bearings. An example is shown in Figure 6-2.

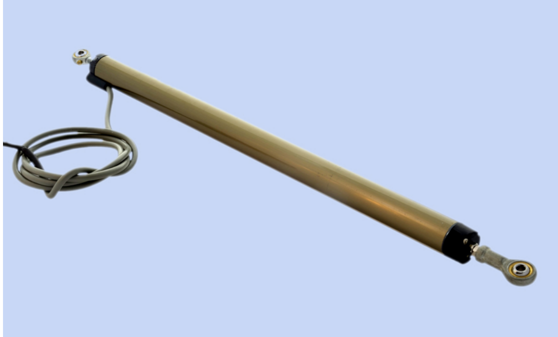

Figure 6-2 *A linear potentiometer contained in a tube with sealed bearings.*

Small linear potentiometers for position sensing are available from companies such as Bourns. The one shown in Figure 6-3 is about 20mm long, and the rod that slides through it has a travel of approximately 10mm. The component is available with resistance values ranging from 1K to 50K, and its power rating is 1/8 watt. The manufacturer claims a life expectancy of 500,000 cycles.

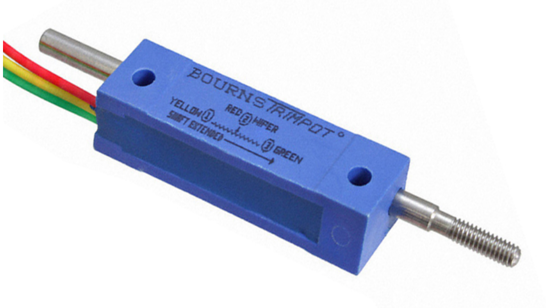

Figure 6-3 *A miniature linear potentiometer. The body is about 20mm long.*

A linear potentiometer is simple, inexpensive, compact, and requires few additional components. The track contains a lubricant, but inevitably some wear occurs as a result of motion of the wiper. Life expectancy will be reduced by vibration or by contamination with dirt or moisture.

A linear potentiometer may rarely be described as a *linear potentiometric sensor*.

Magnetic Linear Encoders

A ferrous rod or strip can be magnetized with alternating north and south poles. When it slides past a single bipolar Hall-effect sensor, it generates a pulse train from the sensor that can be interpreted to provide positional information. The principle of the component is shown in Figure 6-4. (Magnetic rotary encoders also exist; see "Rotary Encoders".)

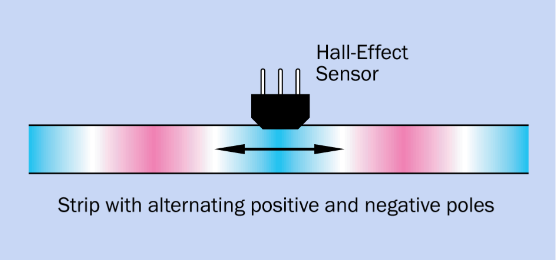

Figure 6-4 *When a strip magnetized with alternating north and south poles slides past a sensor, the pulse train from the sensor can be decoded to indicate the relative position of the strip.*

The sensing element may be described as a *read head*. If two are used, with a spacing equal to half the interval between the north and south poles on the strip, the phase difference between the pulse trains from the sensors will indicate the direction in which the magnetized strip is moving. This is shown in Figure 6-5.

The combination of pulse trains is known as *quadrature*, because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high. The same principle is used in an optical rotary position sensor; see Figure 7-6.

This type of linear position sensor is often described as a *magnetic encoder*, meaning that the position of the sliding part is encoded in a series of pulses. Relatively high resolution is possible, as the north and south poles on a strip of ferrous material can be as close as 2mm. Optical encoders may use the same principle; see "Optical Linear Encoders".

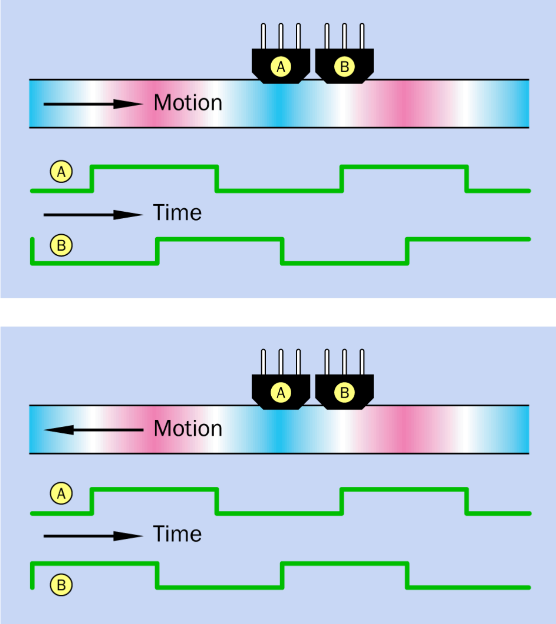

Figure 6-5 *Pulse trains from Hall-effect sensors A and B are shown in green. The phase difference between them can be interpreted to show the direction in which the magnetized strip is moving.*

The sensor can be built into a module containing an analog-to-digital converter that provides numeric output defining the position of the read head.

In an *absolute* magnetic encoder, nonvolatile memory can store the digitized position when the device is switched off. An *incremental* magnetic encoder does not store this information, and will require at least one additional *home sensor* to detect when the encoder is at either end of its travel. At power-up, an initialization routine moves the magnetized strip until the home sensor is triggered.

For details of Hall-effect sensors see "Hall-Effect Sensor".

Optical Linear Encoders

The operation of an optical linear encoder is identical to that of the magnetic linear encoder described immediately above, except that a sliding *optical grating* is used in conjunction with a light source and a detection device such as a **phototransistor** or **photodiode** that functions as the read head. The principle is shown in Figure 6-6. The grating may be described as a *codestrip*.

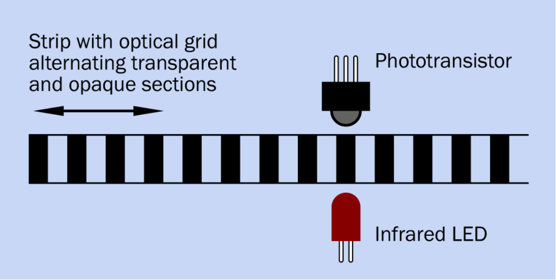
Strip with optical grid alternating transparent and opaque sections
Phototransistor
Infrared LED

Figure 6-6 *An optical linear encoder uses the same general principle as a magnetic linear encoder.*

For details of phototransistors, see Chapter 22. For details of photodiodes, see Chapter 21.

An example of a low-cost optical linear encoder is the Avago HEDS-9 series, which consists of a horseshoe-shaped module with an LED in one arm and an array of photodiodes in the opposite arm. When a codestrip passes between the arms, the module emits two pulse trains from internal comparators. The pulse trains are 90 degrees out of phase, and can be interpreted to show which way the codestrip is moving.

These sensors have a body measuring approximately 10mm wide and are designed to read opaque/transparent intervals ranging from 1.5 to 7.87 cycles per millimeter. The output rate can be as high as 20kHz. No pullup resistor is required on the output, as a 2.5K resistor is integrated.

A *codewheel* may be used, in which case the module senses rotation instead of linear motion. This is described in detail in the entry on **rotary position** sensors. See Chapter 7.

Linear Encoder Applications

Optical or magnetic linear encoders are found in some laboratory equipment, machine tools, and industrial robots. The mean time between failures can range from 100,000 to 1,000,000 hours. Optical encoders must be sealed to provide good protection from dust and dirt.

Linear Variable Differential Transformers

This type of sensor, often referred to by the acronym LVDT, tends to be used in industrial environments where great reliability is required under severe conditions. Examples are high-temperature steam valves and nuclear reactor control mechanisms. However, the robust, frictionless design suggests other possible applications, and custom fabrication is a possibility.

Figure 6-7 illustrates the general principle. Three coils are wound sequentially around a (nonmagnetic) stainless-steel tube, enclosed in a second tube also made of stainless steel. The coils act as transformers with the variable voltage ratio determined by the position of a solid iron armature that slides through them.

The center coil is the primary winding, to which AC is applied between 2kHz to 50kHz, depending on the application. (The frequency must be at least ten times the maximum rate of motion of the armature.) The iron armature is attached to a nonmagnetic rod. The voltages on the secondary windings provide the output from the sensor.

While more than one wiring arrangement is possible, the most common schematic is shown in Figure 6-8. The secondary coils are in series, with one of them reversed, so that the phase of the output is inverted as the armature moves from one end of its travel to the other. The phase detector responds to the phase difference by creating a DC output that varies with movement of the armature. All the functions shown in the schematic are available on a single integrated circuit chip.

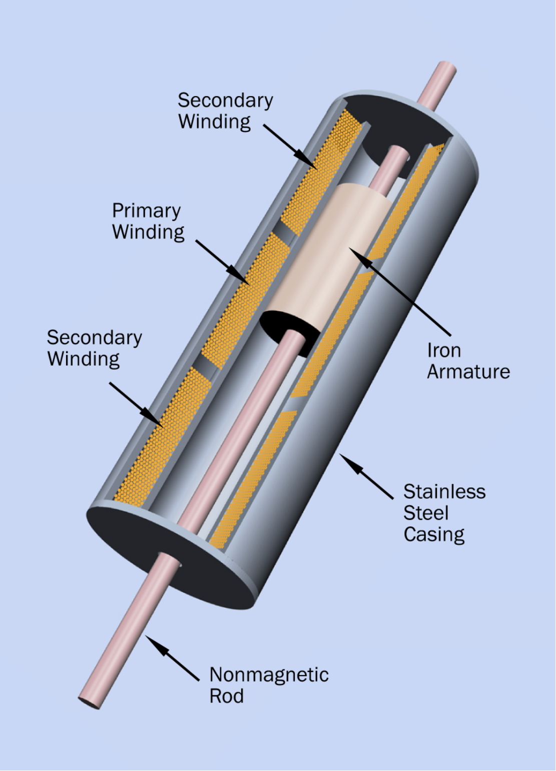
Secondary Winding
Primary Winding
Secondary Winding
Iron Armature
Stainless Steel Casing
Nonmagnetic Rod

Figure 6-7 *Cutaway drawing showing the internal design of a linear variable differential transformer, in which the position of a sliding iron armature determines the voltage induced in the secondary windings.*

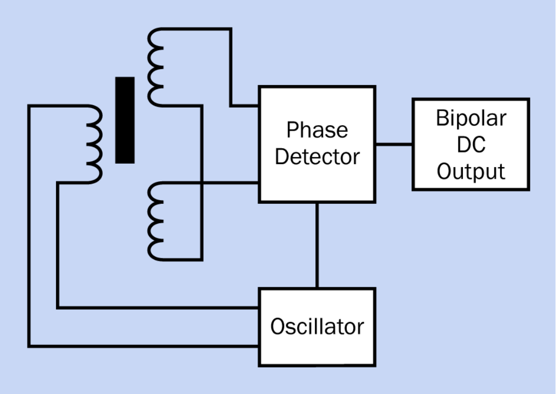
Phase Detector
Bipolar DC Output
Oscillator

Figure 6-8 *Typical schematic for using a linear variable differential transformer.*

An example of a linear variable differential transformer is shown in Figure 6-9.

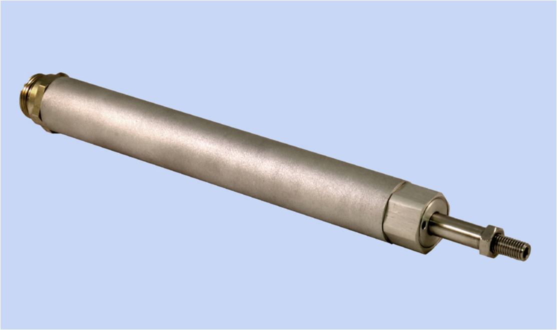

Figure 6-9 *External view of a linear variable differential transformer.*

What Can Go Wrong

Mechanical Issues

Any sliding mechanism will involve friction, and will be vulnerable to wear and tear, resulting in looseness that will degrade its accuracy. In addition, optical systems are vulnerable to dust and dirt.

LED Longevity

Light output of an LED diminishes over a period of years if the LED is “always on.” This will limit the life of the sensor.

rotary position sensor

7

Alternative terms for **rotary position** sensor are *rotary sensor*, *rotational position sensor*, *angular position sensor*, and *angle sensor*.

The terms *rotary* and *rotational* are used interchangeably. In this entry, an attempt has been made to use the term that is most common for each specific application. For example, "rotary position sensor" versus "rotational encoder."

A few sensors are available that specifically measure *rotary speed*, but generally a rotary speed sensor uses information from a rotary position sensor. Therefore, this Encyclopedia does not include a separate entry for rotary speed sensing.

A **rotational encoder** can be used as a rotary position sensor. It is mentioned briefly here but is described in more detail in Volume 1 of this Encyclopedia in the entry describing **switches**.

OTHER RELATED COMPONENTS

- **linear position** sensor (see Chapter 6)

What It Does

Control of a mechanical device may require accurate and timely information about the orientation of a rotating part in the device. A **rotary position** sensor can be used for this purpose.

A sensor may be capable of measuring three attributes:

- Angular orientation
- Direction of rotation
- Speed of rotation

Typically a rotary position sensor measures only the first attribute. Additional electronics are required to calculate the second and third attributes by taking multiple position readings. Thus a *speed sensor* is very likely to be built around a position sensor, and therefore, this Encyclopedia does not have a separate entry for speed sensors.

Applications

In robotics, a rotary position sensor is commonly used to show the orientation of a pivoting arm or strut. It can also be used as a limit switch on a motor.

Specific applications include solar array positioning, remotely piloted vehicles, guidance and navigation, antenna positioning, and wind turbine pitch control.

Pulses from a rotary position sensor are used to measure speed of rotation in vehicles, industrial processes, and aviation. Small rotary speed sensors are built into devices such as cooling fans and computer hard drives.

Schematic Symbol

In a schematic, a rotary position sensor may be represented by symbols for the sensing elements that are inside it (potentiometer, LED, phototransistor, and others).

Potentiometers

Single-turn and multiturn potentiometers may be used as rotary position sensors. For basic information about **potentiometers**, see the entry describing this component in Volume 1.

Arc-Segment Rotary Potentiometer

An *arc-segment rotary potentiometer* is often referred to simply as a "potentiometer," as this type is more common than the multiturn type or the linear type. When used as a sensor, it can measure a turn angle that is less than 360 degrees.

This component contains a resistor in the shape of an arc, referred to as the *track*. It may be a strip of *resistive polymer* or (less often) may consist of an insulator with a coil of nichrome wire wrapped around it.

For sensing purposes, the potentiometer is wired as a voltage divider, and a fixed potential is applied along the full length of the track, as shown in Figure 7-1. A *wiper* slides along the track, sensing a voltage that varies linearly with the wiper's angular position. Output from the wiper can be used directly to control an analog indicator such as a meter, or can be processed by an analog-to-digital converter.

For audio applications, an arc-segment potentiometer may have resistance that varies logarithmically with the position of the wiper. However, this type of component is not generally used as a position sensor.

Low-cost potentiometers were traditionally used as volume or tone controls in stereo systems. When a potentiometer is designed for use as a sensor, it tends to be more ruggedly built and better protected against dust, dirt, and moisture. Its advantages are that it is simple, inexpensive, compact, and requires few additional components.

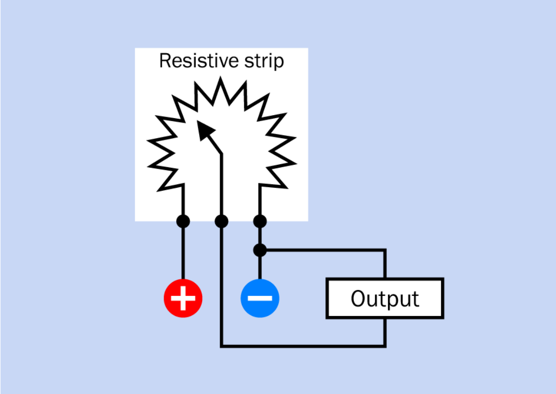

Figure 7-1 *An arc-segment potentiometer consists of a fixed track in the shape of an arc with a known voltage applied at each end, and a movable wiper.*

The major disadvantage is that although the track contains a lubricant, some wear will gradually result from friction with the wiper. Life expectancy will be further reduced if there is vibration or contamination with dirt or moisture.

End Stops

An arc-segment potentiometer usually has *end stops* to prevent the wiper from running off either end of the track. Typically, these stops limit rotation to around 300 degrees.

A few arc-segment potentiometers allow unrestricted rotation. The Bourns 6639 series is an example, although the wiper still passes through a "dead zone" of 20 degrees between the start and end of its track. An application for this type of potentiometer could be as a direction sensor for a weather vane.

Multiturn Rotary Potentiometer

A spiral track, which resembles a coil spring, enables a rotary potentiometer to make multiple turns. The wiper rotates inside the track and follows its contour. A potentiometer of this type will still be calibrated in degrees; thus, a 10-turn component may be listed as allowing an *electrical travel* of approximately 3,600 degrees.

The exterior of a multiturn rotary potentiometer is shown in Figure 7-2.

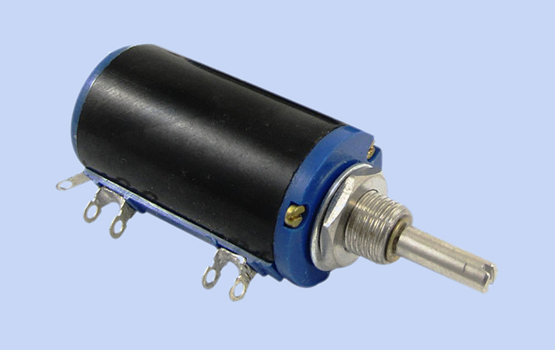

Figure 7-2 *A multiturn rotary potentiometer. Each pair of solder tags is attached internally to one end of the internal coiled track. The smaller tag at the far end connects with the wiper.*

A simpler type of multiturn potentiometer is intended for use as a trimmer—a small potentiometer that can be mounted on a circuit board to allow adjustment or calibration, often during the manufacturing process. This type of trimmer contains a worm gear that engages with a spur gear internally. The wiper is mounted on the spur gear. It has no applications as a sensor, but is mentioned here to avoid ambiguity, as it is probably the component that is most commonly referred to as a "multiturn potentiometer."

Magnetic Rotary Position Sensor

Externally, a modern magnetic rotary position sensor may look very much like an arc-segment rotary potentiometer. Internally, a permanent magnet is attached to the base of the shaft, and one or more **Hall-effect** sensors are mounted on a small circuit board immediately below the magnet, in the bottom of the enclosure. A simplified diagram appears in Figure 7-3.

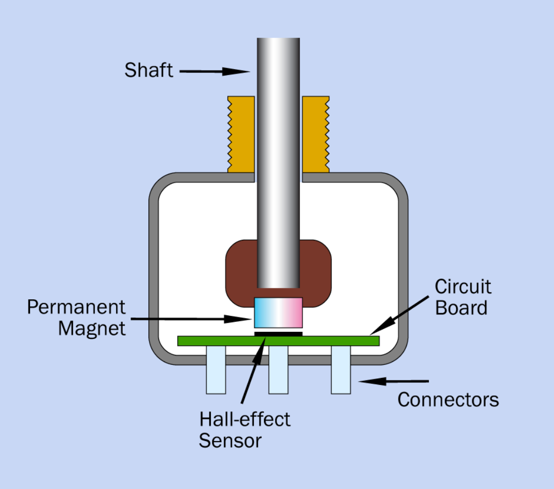

Figure 7-3 *Simplified interior view of a magnetic rotary position sensor.*

For more information about Hall-effect sensors, see "Hall-Effect Sensor".

A magnetic rotary position sensor may be described as a *noncontacting* sensor. Two views of an example are shown in Figure 7-4.

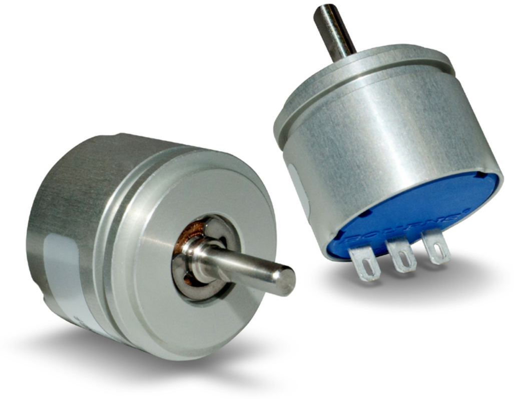

Figure 7-4 *Two views of a Bourns AMS22 magnetic rotary position sensor.*

The AMS22 sensor shown in Figure 7-4 has an analog output ranging from 0.1VDC to 4.9VDC when powered with 5VDC. A noncontacting sensor of this type will cost three to four times as much as a conventional potentiometer, but has the great advantage of extreme durability, with the manufacturer claiming a life of 50 million shaft rotations. A disadvantage is its low-current output, limited to 10mA. The maximum rotation speed of its shaft is 120rpm.

Rotary Position Sensing Chips

Chips of the type used in a magnetic rotary position sensor are available as individual components, many of which have advanced features. For example, the AM8192B angular magnetic encoder by RLS is a 44-pin surface-mount chip containing Hall sensors that detect the orientation of a permanent magnet above or below the chip. Various outputs provide information such as sine or cosine of the turn angle, incremental pulses, and digitized output via an SPI interface.

Rotary Encoders

Many rotating position sensors communicate their angle of rotation with an output that consists of a pulse train or some other coded signal. This type of rotary sensor is known as a *rotary encoder* or *rotational encoder*. (Linear encoders also exist; see "Magnetic Linear Encoders".)

The simplest version of this component is a *mechanical encoder* containing two small electromechanical switches that are activated, out of phase, by a toothed wheel attached to the rotating shaft. This component is described in detail in Volume 1 of the Encyclopedia, where it is categorized as being a form of **switch**. Its low cost and simplicity has made it popular for rotary controls on car radios and small stereo systems, but its switches have a limited life expectancy and create a "noisy" output that must be *debounced* if connected with a logic chip or microcontroller. Typically the microcontroller will include a pause of up to 50ms in its program code to allow time for the switch contacts to settle (although some manufacturers claim only 5ms).

Confusingly, a mechanical encoder is often identified only as a "rotational encoder," even though optical and magnetic rotational encoders also exist, as described immediately below. As a general rule, if a component is described simply as a rotational encoder, it probably contains electromechanical switches.

Optical Rotary Encoders

This type of component works on the same principle as an *optical linear encoder* of the type described in "Optical Linear Encoders". The difference is that a *codewheel* is used instead of a *codestrip*. Typically the codewheel is supplied by the manufacturer of the component that is designed to read it.

A *transmissive* codewheel is shown in Figure 7-5. The distance between the light emitter and the light detector has been exaggerated in this diagram for clarity.

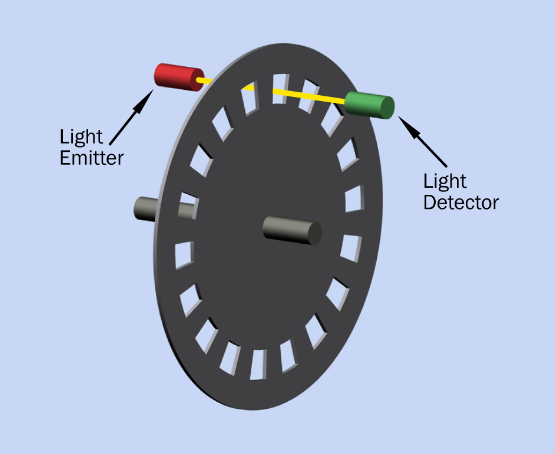

Figure 7-5 *A light-transmissive codewheel for use in an optical rotary encoder.*

Some optical rotary encoders use a *reflective* codewheel, in which case alternating sections of the wheel are light-absorptive and light-reflective, and the emitter and detector are both on the same side of the wheel.

If only one light emitter and light detector are used, the pulse train from the sensor reveals how many increments the wheel has turned relative to its previous position.

A single sensor cannot indicate the direction of rotation, but if a second emitter-sensor pair is added, 90 degrees out of phase with the first, a microcontroller can assess the phase difference between pulse streams to determine which way the wheel is turning. This principle is illustrated in Figure 7-6 where one transmitter-detector pair is located in position A and another is in position B. The resulting pulse trains, for clockwise and counterclockwise rotation of the wheel, are shown in green. The combination of pulse trains is known as *quadrature*, because there are four possible combinations: A and B both high, A and B both low, A high and B low, or A low and B high.

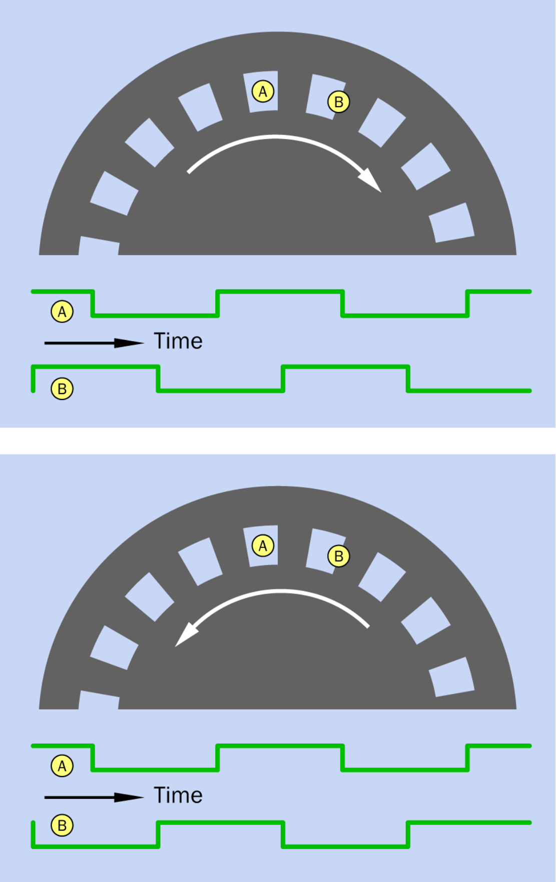

Figure 7-6 *With transmitter-detector pairs mounted at positions A and B, the phase difference between their pulse trains can show the direction of rotation of the wheel.*

This is the same principle as illustrated for a magnetic linear position sensor in Figure 6-5.

An alternative system of quadrature for an optical rotary encoder uses two separate tracks on the disc, each having an equal number of opaque and transparent sections, but half an interval out of phase.

Any system that reveals the relative motion of the wheel, but cannot determine its absolute angular position, is an *incremental sensor*.

Optical Products

High-end optical rotary encoders may use discs with as many as 600 sequential opaque and transparent segments to provide extremely high resolution. They are outside the scope of this Encyclopedia.

Moderately priced optical encoders are available as shaft-driven assemblies very similar in external appearance to potentiometers. Typically they have four terminals, one pair for power and ground connections and another pair for quadrature output from the two internal sensors, usually identified as A and B in datasheets. Some encoders also contain an on-off switch that is activated by pressing the shaft, in which case two additional terminals will be provided.

Bourns is a leading manufacturer of this type of encoder, an example being the EM14, which is mounted in a box-shaped body measuring 14mm square. It uses a 5VDC power supply and provides pulses of 4VDC minimum with intervals of 0.8VDC maximum. Variants are available with 8 to 64 pulses per revolution. Intended for audio applications, this type of encoder has a maximum rotation speed of 120rpm.

An example of an optical rotary encoder from a German manufacturer is shown in Figure 7-7. Its resolution is 25 pulses per rotation. The rectangular package measures approximately 19mm × 25mm. The power supply can be 3.3VDC or 5VDC.

Optical rotary encoders of this type cost about five times as much as mechanical rotary encoders at the time of writing, but their longevity and their clean output signals make them an attractive alternative, and the price difference may diminish over time.

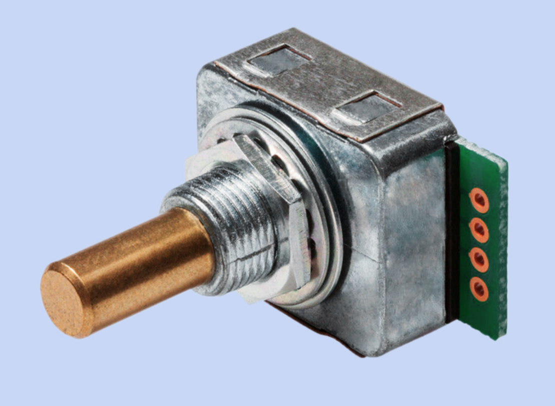

Figure 7-7 *A compact incremental optical rotary encoder in the MRB25 series from Megatron Elektronik AG & Co.*

Encoders such as the Avago HEDS-9 series are not protectively enclosed, and require assembly with a codewheel supplied by the manufacturer. See Figure 6-6 for additional information.

A more basic optical rotary encoder, shown in Figure 7-8, is sold by Cytron Technologies as an *optical switch* mounted on a small board, with a separate codewheel consisting of a slotted disc. This low-cost kit is intended for use in DIY robotics. More information about optical switches will be found in the entry dealing with **optical presence** sensors. See Figure 3-3.

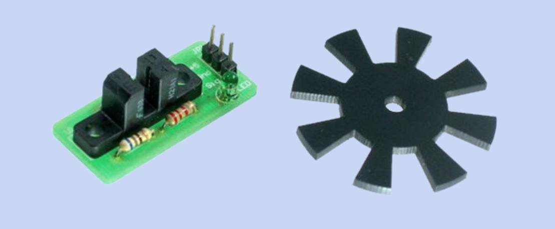

Figure 7-8 *A bare-bones optical encoder for DIY robotics.*

Computer Mouse Principles

The original design of a computer mouse, with a hard rubber ball, contained two optical rotary encoders oriented at right angles to each other. Each of them used a transmissive codewheel. The rolling ball turned the codewheels as the mouse was moved across a desktop, and electronics in the mouse converted the outputs from the encoders into a pulse train that could be interpreted by a computer. Figure 7-9 shows the primary components.

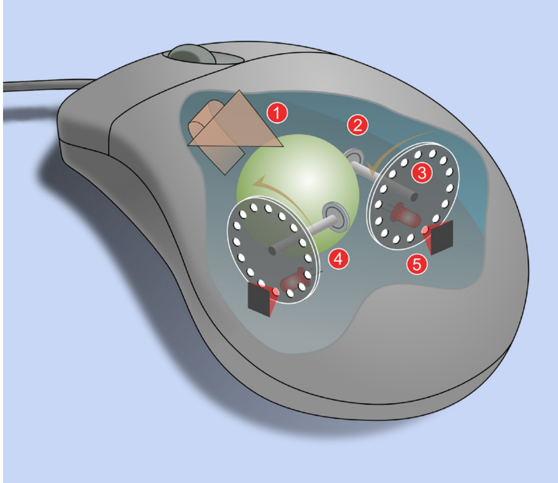

Figure 7-9 *From Wikimedia Commons, this rendering by Jeremy Kemp shows (1) rotation of the ball, (2) a roller touching the ball, (3) a transmissive optical codewheel, (4) an infrared LED that shines through a second codewheel, and (5) a sensor detecting the pulses of light.*

An *optical mouse* works on a different principle, maintaining a monochrome image of the desk surface on an optical array that functions like a very low-resolution camera sensor. Electronics in the mouse detect displacement of the image as the mouse is moved.

Rotational Speed

Incremental rotary encoders, which only supply relative data, are adequate in many applications, especially speed measurement, where a microcontroller can compare a pulse stream from a sensor with a target frequency, and can also provide feedback to control motor speed appropriately. This is convenient when a stepper motor is being used, or a DC motor that is controlled with pulse-width modulation. (See the entry on **motors** in Volume 1.)

Toothed wheels or magnetized wheels are commonly used to measure speed of rotation in applications ranging from automobile transmissions to computer disk drives.

While a second sensor can be added to determine the direction of rotation, this still does not provide information about the absolute orientation of a part.

Absolute Position

If associated electronics are equipped with nonvolatile memory, it may be used to store a sensor-wheel position from one session to the next. This may be sufficient in noncritical applications such as volume control in a car radio or stereo system.

Alternatively, an additional single window in an optical wheel can activate a *home sensor*. When the device is powered up, the wheel is turned until the home sensor is triggered, at which point the orientation of the wheel is known, and subsequent pulses from rotation sensor(s) will add or subtract angular information.

The pulse generated by a home sensor may be described in datasheets as a *reference signal* or *index signal*. In the early days of desktop computers, each 5.25-inch floppy diskette was perforated with an index hole for this purpose.

The Gray Code

For greater reliability in determining absolute position, an optical wheel can be divided into several concentric tracks, each of which contains a different coded sequence and is assigned its own light emitter and light sensor. The sensors are arranged in a radial line to scan the disc as it rotates. Since each detector will provide either a signal or no signal, output from the set of sensors can be combined to create a binary number.

Figure 7-10 shows binary codes from 0000 through 1111, corresponding with decimal numbers 0 through 15, where a white square is equivalent to a numeral 1 and a black square is equivalent to a numeral 0.

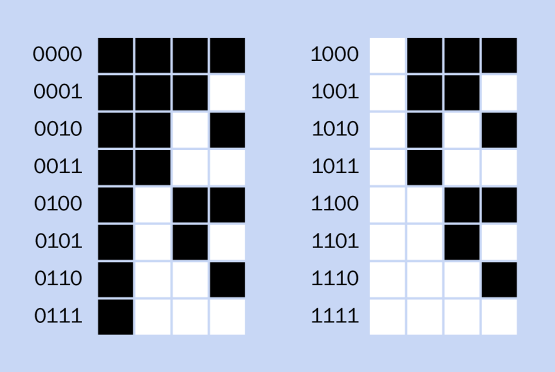

Figure 7-10 *Binary codes from 0000 through 1111, using white to represent 1 and black to represent 0.*

Figure 7-11 shows this system mapped onto a codewheel. The red circles indicate the locations of four stationary light detectors, which begin by providing a reading of 0000, since they coincide with four opaque areas of the wheel. Now if the wheel makes 1/16th of a full rotation in the direction of the arrow, the detectors will register 0001. If the wheel continues to rotate, the detectors will count in binary up to 1111 before the sequence repeats.

The problem with this design is that small manufacturing inaccuracies and other imperfections will result in some light detectors responding fractionally more quickly than others as the wheel rotates. This will occur in transitions where two or more adjacent segments of the wheel change between transparency and opacity. For instance, where 0011 is followed by 0100, transient values of 0010, 0001, 0111, 0110, or 0101 are possible. Although they will be brief, these values may trigger associated electronics.

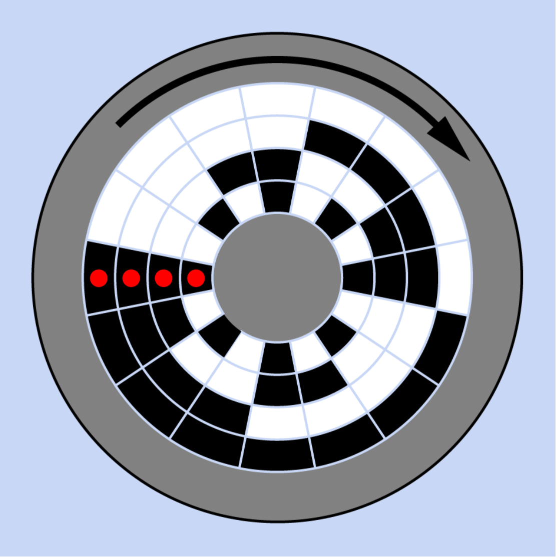

Figure 7-11 *The binary sequence mapped onto a codewheel as areas of opacity and transparency. Red circles indicate light detectors.*

To eliminate this problem, a different code sequence can be used in which only one of the four sensors is allowed to make a transition from each value to the next. This is called a *Gray code*, and eliminates the issue of simultaneous transitions. A commonly used Gray code is shown in Figure 7-12.

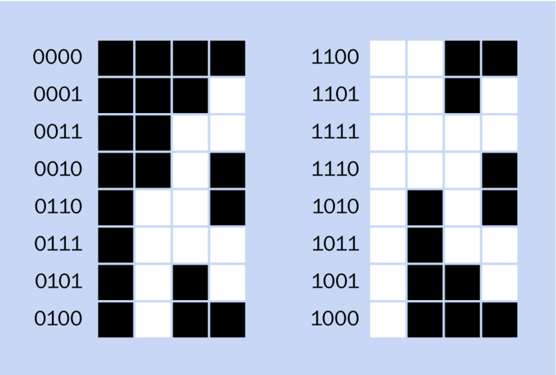

0000	■	■	■	■	1100	□	□	■	■
0001	■	■	■	□	1101	□	□	■	□
0011	■	■	□	□	1111	□	□	□	□
0010	■	■	□	■	1110	□	□	□	■
0110	■	□	□	■	1010	□	■	□	■
0111	■	□	□	□	1011	□	■	□	□
0101	■	□	■	□	1001	□	■	■	□
0100	■	□	■	■	1000	□	■	■	■

Figure 7-12 *A Gray code that allows only one binary digit to change in each transition from one value to the next.*

Magnetic Rotary Encoders

If a ferrous wheel is magnetically polarized in multiple domains, its rotation can be detected by a Hall-effect sensor in much the same way that a wheel divided into transparent and opaque sections can be assessed by light detectors and emitters. This is illustrated in Figure 7-13, where magenta and cyan bands indicate north and south magnetic poles.

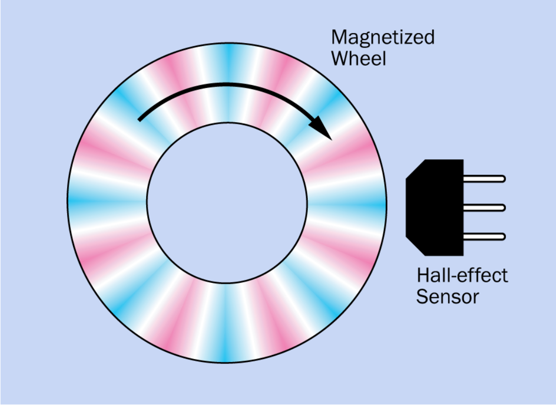

Figure 7-13 *A Hall-effect sensor can detect the rotation of a wheel that is divided into multiple north and south magnetic poles.*

See "Hall-Effect Sensor" for information about Hall-effect sensors.

An additional Hall sensor can be added, offset from the first, as in Figure 7-14. Once again the phase difference between the pulse trains can be used to determine the direction of rotation.

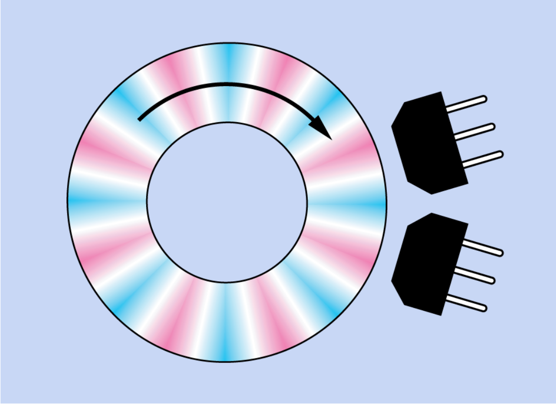

Figure 7-14 *Direction of rotation of the wheel can be deduced from the phase difference between the pulse trains from two sensors.*

Alternatively, a toothed wheel can be used, as suggested in Figure 7-15.

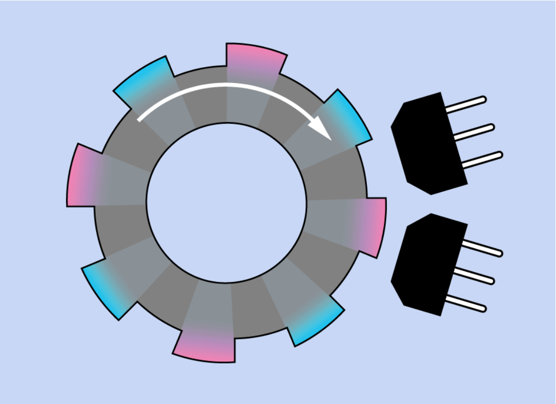

Figure 7-15 *A toothed wheel can trigger a Hall sensor if the teeth are magnetized.*

Another option is to mount a pair of magnets on a nonmagnetic wheel, with a Hall-effect sensor in the center. If it is a sensor with an analog output, the voltage will fluctuate smoothly between positive and negative, relative to the power supply for the sensor. Alternatively, a bipolar Hall-effect sensor can be used to provide a binary output. The concept is illustrated in Figure 7-16. An advantage of this configuration is that it provides approximate information about the absolute position of the wheel.

The two magnets will provide usable linear sensor outputs over a span of plus-or-minus 30 degrees of rotation, approximately. Additional magnets or sensors can produce a more complex output that would be decoded by a microcontroller.

How to Use It

An optical or rotary encoder is well suited for use with a microcontroller program that can count pulses, compare pulse trains, or interpret a Gray code. The microcontroller then takes appropriate action. For example, if a rotational encoder is used to control the gain of an audio amplifier, the microcontroller determines the direction and angle through which the encoder turns, and can respond by changing the value of a **digital potentiometer** (see Volume 1).

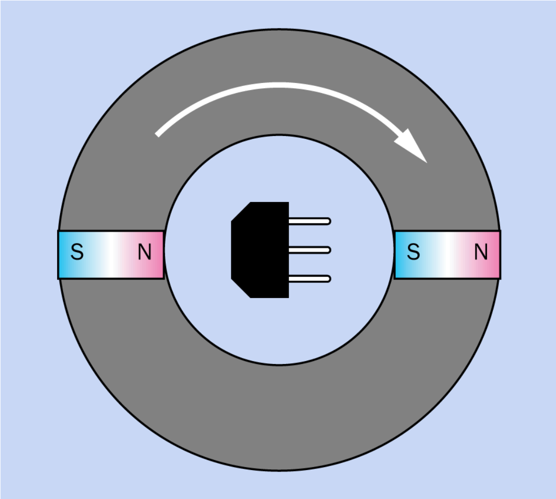

Figure 7-16 *A Hall-effect sensor can detect the angle of rotation of a ring on which two magnets are mounted as shown.*

Integrated circuit chips are available to convert a sequence of quadrature signals into an up-pulse or a down-pulse, thus eliminating the chore of achieving this with a microcontroller program. The LS7183 by LSI Computer Systems is an example.

What Can Go Wrong

Wiring Errors

If two sensors are used to detect the direction of rotation of a rotary encoder, and outputs from the sensors are accidentally swapped, the component will seem to work normally except that the indicated direction will be inverted.

Coding Errors

Microcontroller code that interprets quadrature signals must be fast enough to keep up with the pulse streams. If a microcontroller is performing other tasks, an interrupt may be necessary for processing the rotational data. This should not be a significant problem when interpreting human input from a knob or dial, but for a motor-driven encoder, pulse counting in hardware may be a better alternative.

Ambiguous Terminology

Rotational encoders are often referred to simply as “encoders,” because they are more common than linear encoders. When searching for an encoder, check each datasheet carefully to determine which type you are dealing with.

8

tilt sensor

A *tilt switch* is defined here as an electromechanical switch, while a *tilt sensor* uses electronics. Both types are included in this entry.

A *tipover switch* is very similar to a tilt switch and uses the same principles. Therefore, it is included in this entry.

Some manufacturers' datasheets refer to tilt sensors as *tip sensors*. Some Asian supply catalogues refer to tilt switches as *breakover switches*.

An **accelerometer** can measure the angle at which it is held relative to the downward force of gravity, but it has additional capabilities. Therefore, it has its own separate entry in this Encyclopedia.

An *inclinometer* measures the incline, or positive slope, from an observation point to the top of an object such as a building or tree. The height of the object can be calculated from the angle. A *clinometer* can additionally measure a decline, or negative slope. These measurement devices are fully featured products as opposed to sensors, and are not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **vibration** sensor (see Chapter 11)

What It Does

Three principal types of tilt sensor exist.

1. Single axis, single output. The sensor responds to being tilted around one horizontal axis, relative to the downward force of gravity.
2. Dual axis, dual output. The sensor contains two sensing elements at 90 degrees to each other. Each has an output determined by its angle of tilt from vertical around one axis.
3. Dual axis, single output. A single sensor responds to an angle of tilt from vertical around any horizontal axis.

A *tilt switch* is usually of the third type, and is defined here as containing an electromechanical or electronic switch that opens or closes a connection. Most tilt switches are SPST (normally open) or SPST (normally closed). A minority are DPDT.

A *tipover switch* is a type of high-current tilt switch that cuts power to a device such as an electric heater when it is tipped over.

This Enclopedia defines a *tilt sensor* as being an electronic component, as opposed to an electromechanical component. The distinction is often observed in datasheets, but not always.

Schematic Symbol

No specific schematic symbol is generally used for any variant of a tilt sensor. It can be represented with an annotated switch symbol.

How It Works

Because a tilt switch is a simpler device than a tilt sensor, it will be described first.

The most common type of tilt switch consists of a cylindrical metal or plastic enclosure, often measuring about 5mm by 15mm, containing two spherical steel balls that may be nickel-plated or gold-plated. When the switch is tilted, the balls eventually run downhill, and the lower ball completes an electrical connection between two contacts or between a single contact and the metal enclosure of the switch. The second ball is included to add weight and suppress vibration in the first.

Figure 8-1 shows a switch manufactured by Comus Global, rated for 0.25A at 60VAC or 60VDC, maximum. The body of the switch measures approximately 10mm x 5mm, and the switch is activated when tilting -10 degrees from horizontal. It is deactivated when tilting +10 degrees. A scale drawing of the interior is shown in Figure 8-2.

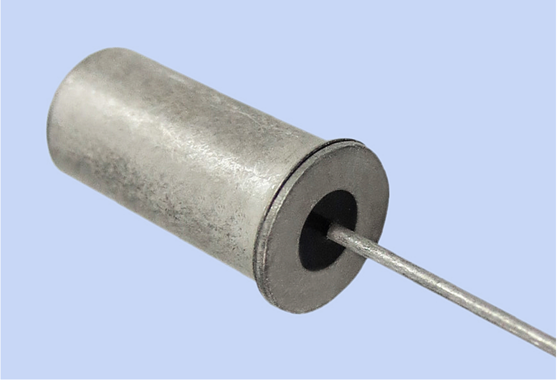

Figure 8-1 *The CW1300 tilt switch manufactured by Comus Global.*

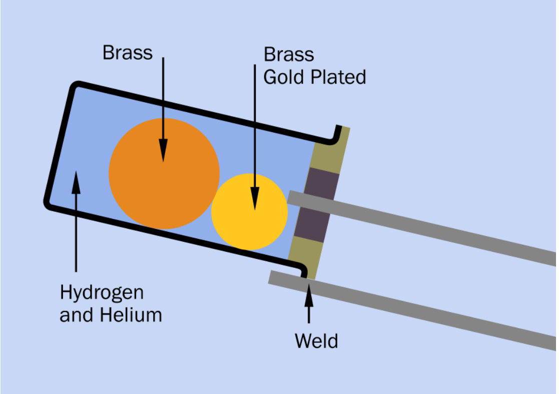

Figure 8-2 *Interior of the CW1300 tilt switch, from a scale drawing supplied by the manufacturer. The lower lead is welded to the shell of the sensor. The leads may be inserted in a circuit board.*

Figure 8-3 shows three common internal configurations of a generic tilt switch. The top version has axial leads and uses the metal shell of the switch to complete the circuit. The center version has radial leads, with a plastic shell. The bottom version has radial leads, one of which is attached to the metal shell to complete the circuit.

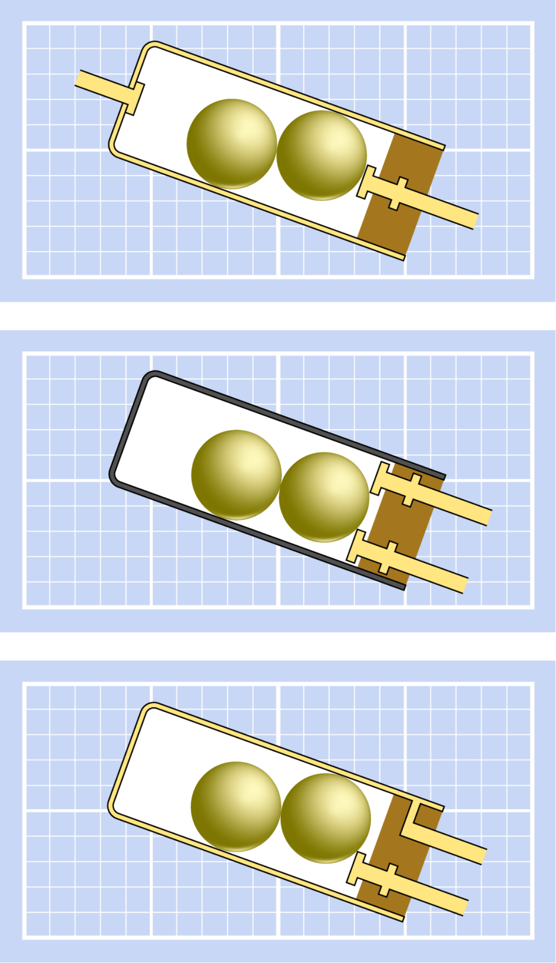

Figure 8-3 *Three variants of a generic ball-operated tilt switch. The graph-paper scale is in millimeters.*

Figure 8-4 shows the parts of a disassembled tilt switch.

Simplified Version

Where size is not a significant consideration, a tilt switch can be created by attaching a pivoting, weighted arm to a small snap-action switch.

Applications

An old-style (nonelectronic) thermostat may contain a tilt switch attached to the end of a bimetallic strip coiled into a spiral. When the strip bends in response to a drop in temperature, the switch closes its contacts, activating a relay that starts a heating unit. If the temperature rises, an additional set of contacts in the same relay may activate an air-conditioning unit. In old thermostats, the tilt switch may contain mercury in a glass tube, which should be handled with caution.

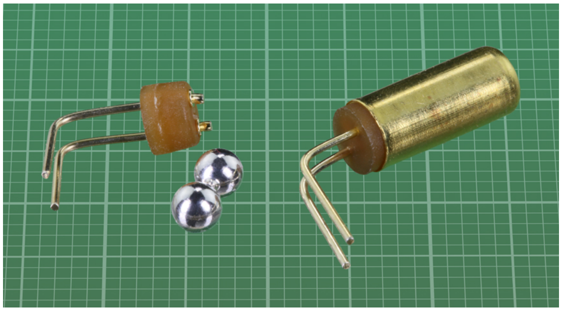

Figure 8-4 *At right, a tilt switch. At left, the cap removed, and the two balls that make internal contact. The background grid is in millimeters.*

A tilt switch may detect the opening of a door or window in a simple alarm system.

Tilt switches have been used in automobiles to switch on the interior light in the trunk when its lid is opened.

A normally closed tilt switch is often used to stop the inflow of granular material to a bin when it is almost full. This is colloquially known as a *bin switch*. In industrial applications of this kind, the switch is activated by a long lever that has a ball mounted at the end. The switch assembly is physically large. See Figure 8-5.

A normally open tilt switch may operate a valve or start a pump when the liquid in a tank drops below a certain point. If the switch uses a float to sense the level of the liquid, it is often known as a *float switch*. This is described in the entry on liquid level sensors. See Chapter 15.

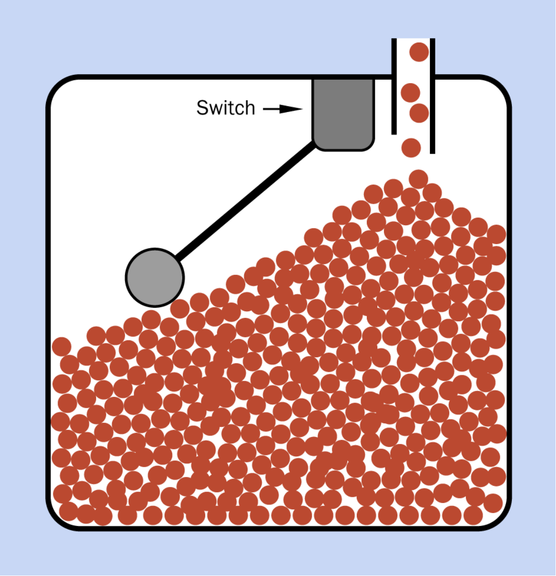

Figure 8-5 *The flow of granular material into a bin can be sensed and stopped by a tilt switch of this type, often known as a bin switch.*

A *tipover switch* may use the simplified system of a weighted arm that activates a snap-action switch. When used in conjunction with a room heater, the switch must handle substantial current.

A motorcycle may contain a tipover switch to stop the electric fuel pump if the motorcycle falls on its side.

Four tilt switches placed in a cross-shaped pattern on a flexible mount can be used as the basis of a very simple game controller, with a joystick mounted in the center.

Variants

The three configurations of ball-type tilt switches shown in Figure 8-3 are functionally interchangeable and can be chosen for convenience of their leads in fitting the circuit.

Mercury Switches

Early tilt switches contained a blob of mercury in a glass bulb. When the bulb was tilted, the mercury rolled to the end and made an electric connection between two metal contacts that penetrated the bulb.

A small mercury switch is shown in Figure 8-6. This type of sensor became less common after many countries classified mercury as an environmental hazard and established regulations restricting its use.

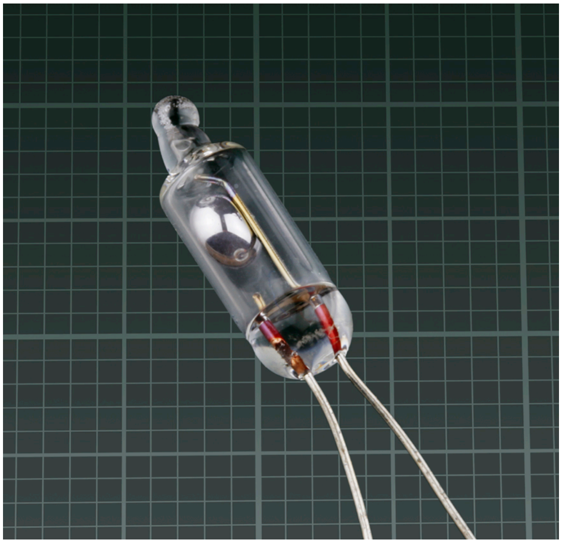

Figure 8-6 *Small mercury switch rated for 0.3A at 24VDC or 24VAC. Larger mercury switches can switch more power; 1A at 230V is common. The background grid is in millimeters.*

Mercury is an excellent electrical conductor. It remains in a liquid state between about -38 degrees Celsius and +356 degrees Celsius, and has very high surface tension, encouraging it to form a single blob instead of breaking up into small droplets. Because free space in the bulb is filled with an inert gas to prevent oxidation of the electrodes, a mercury switch can have a very long operating life. In the United States in the 1970s, some light switches were sold containing mercury switches with a claimed life expectancy of 100 years.

Pendulum Switch

This type of switch, now relatively rare, was found in vintage pinball machines. It consists of a pendulum about 5cm long, suspended inside a steel ring about 1cm in internal diameter. If the machine was rocked sufficiently to bring the pendulum in contact with the ring, the game was cancelled and the word “Tilt” appeared on the display. Consequently this was referred to as a tilt switch, although really it was a form of **vibration** sensor with a long period of oscillation.

Magnetization

Some tilt switches use a steel ball that is weakly magnetized, so that it will seat itself more firmly when it rolls into a circular depression or ring. This type of switch must be tilted back through a larger angle to dislodge the ball. Therefore it will exhibit greater *hysteresis*.

Tilt Sensors

Unlike a tilt switch, a tilt sensor is not built around an electromechanical switch.

The principle of a rolling ball has been miniaturized and encapsulated in a small enclosure (10mm square or smaller), in which the ball rolls to interrupt a beam from an internal LED shining on a phototransistor. Examples are found in the Panasonic AHF series. Internal circuitry ensures a clean on-off signal, free from the switch bounce that is a problem in basic ball-type tilt switches. However, the switch requires a power supply, and the open-collector output must be used with a pullup resistor. By comparison, a simple electromechanical tilt switch can be wired directly to the device that it controls.

Diagrams from the Panasonic datasheet show the three types of AHF sensor available for vertical, horizontal, and reverse mounting. In each case, the ball (dotted circle) rests in a shallow cup (dotted curve) where it obstructs the beam from an internal LED (not shown). See Figure 8-7. An exterior view of the AHF22 is shown in Figure 8-8.

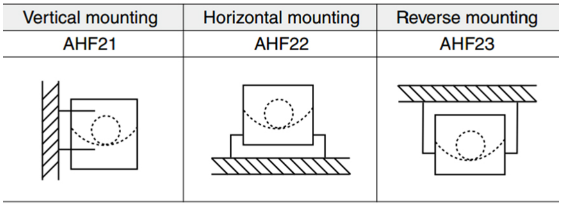

Figure 8-7 *Three types of Panasonic tilt sensor, from the manufacturer’s datasheet. See text for details.*

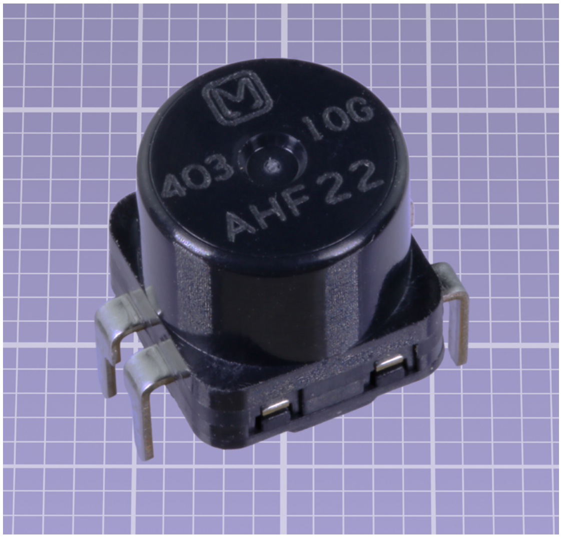

Figure 8-8 *Exterior view of the Panasonic AHF22 tilt sensor. The background grid is in millimeters.*

Two-Axis Tilt Sensors

The Rohm RPI-1035 is a surface-mount tilt sensor about 4mm square, with two phototransistor outputs that indicate which axis the sensor is tilting around. The outputs can be interpreted as a 2-bit binary number, with its four possible states indicating the rotation of the switch around two axes at 90 degrees to each other. Switches of this type were developed to indicate the orientation of consumer-electronic devices such as digital cameras, but more sophisticated sensors containing accelerometers are becoming price-competitive.

Surface-mount 2-axis tilt sensors have been made available on small breakout boards that are easy to use experimentally. An example is the Parallax 28036 shown in Figure 8-9.

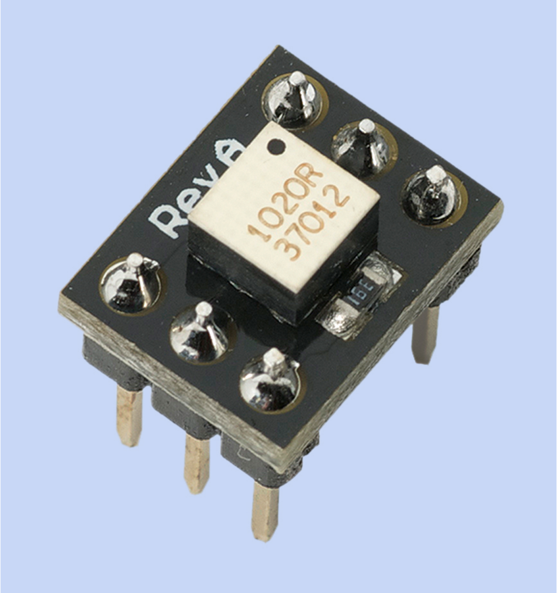

Figure 8-9 *A 2-axis tilt sensor mounted on a breakout board available from Parallax.*

The behavior of the rolling-ball sensor at the heart of this board is shown in Figure 8-10. The sensor contains a square cavity in which the ball is depicted as a blue circle. At one corner of the cavity is a red LED, while two phototransistors, labeled A and B, are at the corners on the left and right. When the sensor is held as in section 1 of the figure, with the LED at the top and the ball resting at the bottom, both phototransistors have a high output as they receive light from the LED.

In section 2 of Figure 8-10, the sensor has been turned through 90 degrees. The ball now prevents light from reaching sensor B, while phototransistor A is still active. In section 3 of the figure, the sensor has been turned through another 90 degrees, so that the ball prevents light from escaping from the LED, and both phototransistors are now dark. In section 4, the ball obstructs phototransistor A but not phototransistor B.

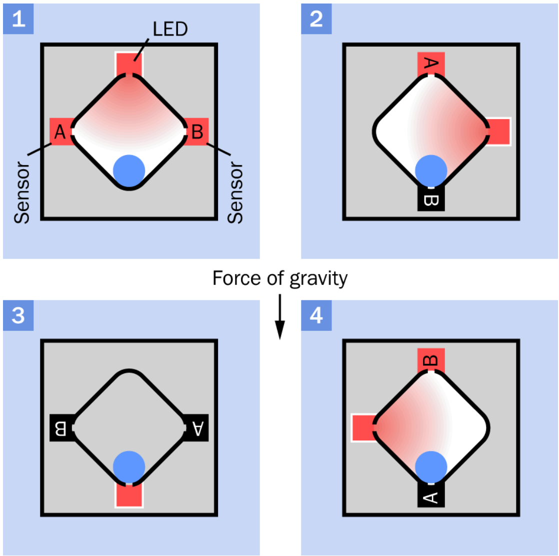

Figure 8-10 *The rolling-ball sensor inside the Parallax 28036. See text for details.*

Suppose that the sensor is placed flat on a horizontal surface. The sensor will now respond to being tipped either way around two horizontal axes. This justifies its description as a 4-directional tilt sensor, although its design suggests that it may have been intended for use as described above, rotating through four positions around one horizontal axis.

Values

A heavy-duty tilt switch can be rated as highly as 10A at 240VAC. More commonly, a tilt switch about 15mm long can be expected to switch about 0.3A at 24VAC or 24VDC.

The *operating angle* is the angle through which the switch must be turned to activate it, relative to its normal rest position.

The *return angle* is the angle to which the switch must be returned to deactivate it. Hysteresis results from the return angle being smaller than the angle that activates the sensor.

Tilt sensors with an open-collector output will specify maximum forward current for the internally mounted LED (usually no greater than

50mA), the maximum collector-emitter voltage at the output (typically 30V), and the maximum collector current (often 30mA). An explanation of open-collector outputs is given in Appendix A (see "3. Analog: Open Collector").

How to Use It

An electromechanical tilt switch can be connected directly between a power supply and a device, so long as the device does not draw more current than the switch is rated to handle. Note that inductive loads such as motors draw an initial surge that can be at least twice the operational rating, while relays may be more likely to create a voltage spike when disconnecting. Switches should be chosen accordingly. For more discussion of this topic, see **switch** in Volume 1.

A small tilt switch can be used in conjunction with a relay or transistor to amplify its signal sufficiently to drive a larger load.

If an electromechanical tilt switch is connected to an electronic device such as a microcontroller or logic chip, output from the switch will have to be *debounced* to prevent a series of brief voltage spikes that can cause false triggering when the switch turns on or off. A debouncing logic circuit or chip can be used, or the program code in a microcontroller can introduce a wait period of up to 50 milliseconds to allow the contacts to settle.

Mercury switches are much less likely to create a noisy output than rolling-ball switches, and may require little or no debouncing.

For an application that must sense rotation around two or three axes, multiple single-axis tilt switches can be combined. A microcontroller or logic gates will be necessary to evaluate signals from the switches, to determine the orientation.

What Can Go Wrong

Contact Erosion

If a ball-type tilt switch is subjected to current that exceeds its specification, arcing may erode its contacts, and they will become less reliable, especially if the contacts are plated with a thin metallic film that is eroded. For additional information on arcing in switches, see the **switch** entry in Volume 1.

Random Signals

During the brief time when a ball-type tilt switch is turning from one position to the other, vibration of the ball(s) inside it is likely to create erratic, random signals. If the output from the switch is being evaluated by a microcontroller, a debouncing routine may be insufficient to prevent the random signals from being sensed, and some programming will be necessary to ignore the signals during this transitional phase. If the switch is connected directly to a relay, the intermittent signals may occur sufficiently rapidly that the relay will ignore them.

Environmental Hazard

A device that incorporates a mercury switch may have to be re-engineered in the future if the availability of mercury switches becomes unreliable as a result of tighter environmental regulations. For the same reason, the end user may have difficulty replacing a mercury switch if it fails. Therefore, a ball-type tilt switch should be used instead of a mercury switch in any newly designed device.

Requirement for Gravity

Because a tilt switch depends on gravity to roll a ball or move a blob of mercury, it will not work in low-gravity, reversed-gravity, or zero-gravity conditions—for example, in a rocket during the unpowered phase of ascent and descent, or in an aircraft that performs aerobatic maneuvers. Performance of a tilt switch in a vehicle that accelerates or decelerates suddenly may also be unreliable. Likewise, it cannot be used on a small boat.

Requirement for Stability

A tilt switch will tend to give erroneous results in a location where there is significant vibration or where the object containing the switch may be turned or repositioned unpredictably by the user.

gyroscope

Historically, a **gyroscope** always contained a spinning disc. Some devices for navigation still depend on rotating elements, but they are outside the scope of this Encyclopedia. This entry deals primarily with *vibrating gyroscopes*, also known as *resonator gyroscopes*, that are MEMS devices contained within silicon chips.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **GPS** (see Chapter 1)
- **magnetometer** (see Chapter 2)

What It Does

A **gyroscope** resists rotation around any axis at right angles to its own axis of rotation or vibration. Consequently, if the gyroscope is allowed to move freely on gimbals in a sealed enclosure, the gyroscope will tend to maintain its orientation while the enclosure can rotate freely around it.

Taking this concept a step further, if the enclosure is mounted in an aircraft, the aircraft's rotation around two axes can be determined by referring to the gyroscope. If additional gyroscopes are added orthogonally to the first, the aircraft's rotation around all three axes can be determined.

A gyroscope does not measure linear motion in any direction, or any static angle of orientation.

Schematic Symbol

A chip-based gyroscope, magnetomer, or accelerometer may be represented in a schematic as a rectangular box containing abbreviations to identify pin functions (as in any integrated circuit chip).

IMU

An **accelerometer** measures variations in linear motion and will also measure its own static orientation relative to the force of gravity. If an accelerometer rotates around its own axis, it will not measure angular velocity.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Applications

The first chip-based gyroscope was used in automobiles in 1998 as a yaw sensor in a skid-control system. Subsequent automotive applications include active suspension control, air bag sensors, rollover detection and prevention, and navigation systems.

Gyroscopes may be installed in military ordnance to provide backup in case an onboard **GPS** system fails, possibly as a result of radio jamming.

Handheld 3D game controllers and headsets may use gyroscopes to control images displayed to the viewer. A digital camera may employ a gyroscope to provide image stabilization. Gyroscopes are usually found in quadcopters or drones, are used to stabilize two-wheeled vehicles such as the Segway, and are used in robotics.

How It Works

The traditional form of gyroscope is a rotating wheel, which will resist turning forces perpendicular to its own axis of rotation. In Figure 9-1, three directions at right angles to each other are defined in the bottom-right corner of the diagram as X, Y, and Z. The wheel is rotating around the X axis, as shown by the green arrow. It will resist any turning force around the Y axis (red arrows) or the Z axis (yellow arrows).

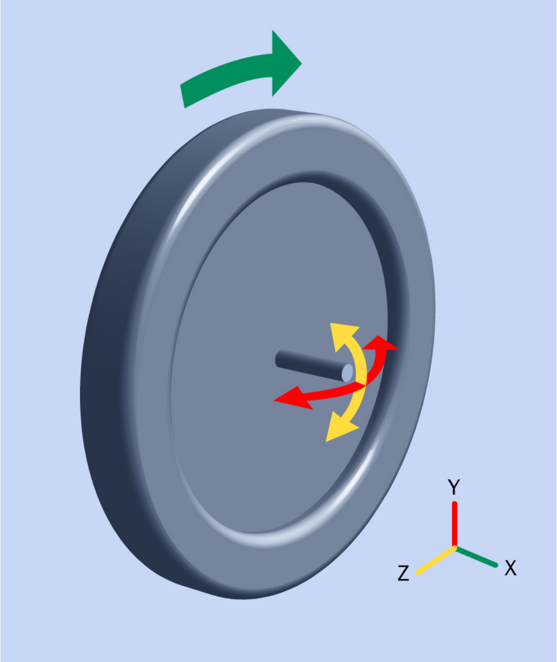

Figure 9-1 *In a traditional gyroscope, a wheel that is rotating (shown by the green arrow) will resist a turning force on either of the axes (shown by red and yellow arrows) perpendicular to its axis of rotation.*

Vibrating Gyroscope

A vibrating fork can be substituted for a wheel. In Figure 9-2 a fork is secured at its base while its tines are induced to vibrate toward each other and away from each other, as suggested by the double-ended arrow. In a chip-based gyroscope, this vibration is induced piezoelectrically or by static electricity.

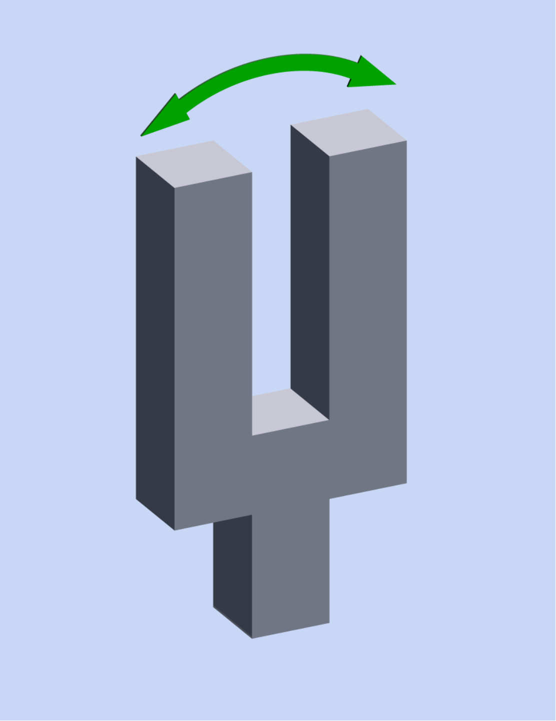

Figure 9-2 *A vibrating fork (green arrows) can be substituted for a rotating wheel in a gyroscope.*

Now suppose a turning force is applied around the vertical axis at the base of the fork, as suggested by the lower arrow in Figure 9-3.

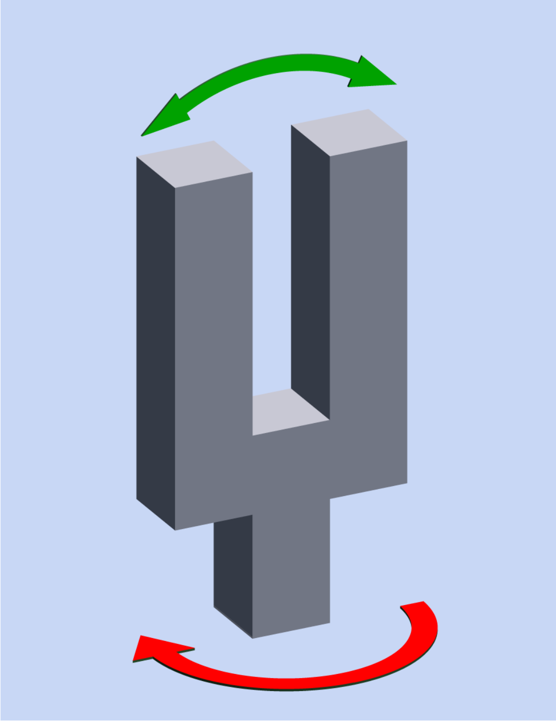

Figure 9-3 *A turning force is applied to the base of the fork around a vertical axis, as shown by the lower arrow.*

The angular momentum of the vibrating tines causes them to resist this turning force, and consequently they will tend to bend, as shown by the yellow arrows in Figure 9-4. The amount of their deflection can be measured capacitively. This system is used in many chip-based gyroscope systems, and may be referred to as a *vibrating gyroscope* or a *resonator gyroscope*.

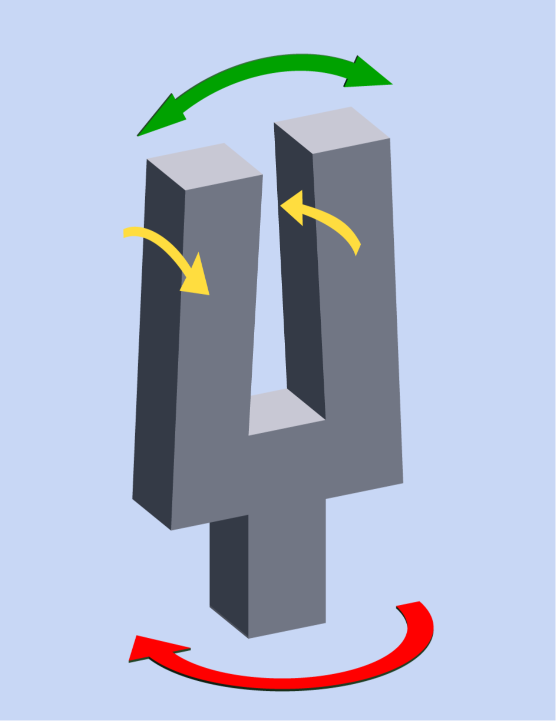

Figure 9-4 *The angular velocity of the rotating fork causes deflection of its vibrating tines, shown by the yellow arrows.*

An assembly of microscopic forks can be etched into a silicon chip. In Figure 9-5, an electron micrograph shows the interior of this type of chip. It contains three rotational sensors, responding to the X, Y, and Z axes of motion. These sensors respond to *pitch* (rotation around the X axis), *roll* (rotation around the Y axis), and *yaw* (rotation around the Z axis).

Fork-based sensors are analog devices whose values are converted to digital values by an onboard *analog-to-digital converter* (ADC). The values are stored in registers that are available to other devices, often via the I2C protocol, which is widely used by microcontrollers.

For additional details about I2C, see Appendix A.

Typically there will be two 8-bit registers for each axis. Each register stores the binary equivalent of a signed integer, where the positive or negative value represents the direction and magnitude of deflection, usually in degrees per second (dps).

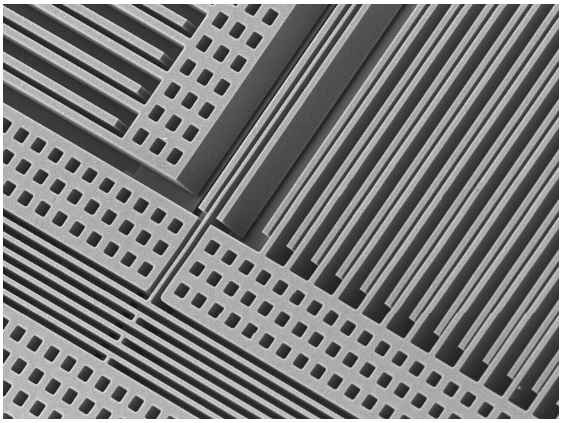

Figure 9-5 *Electron micrograph of the STMicroelectronics LIS331DLH vibrating gyroscope installed in the Apple iPhone 4. The parallel plates at bottom-left function as a spring, while the elements at top-left and at right measure capacitance as their orientation varies according to rotational velocity. Photo courtesy of MEMS Journal published by Chipworks*

Variants

The L3G420D by STMicroelectronics is a 3-axis gyroscope-only chip. It communicates via the SPI or I2C protocol, is approximately 4mm square, and can measure rotational rates up to plus-or-minus 2,000 degrees per second.

The Freescale FXAS21002C has a similar specification. Gyroscope-only chips of this type have fallen in price to the point where they are comparable to the retail cost of everyday components such as a small-signal relay or an audio amplifier on a chip.

IMUs

Chips that only contain gyroscopes are becoming less common as the cost of adding accelerometers decreases.

Gyroscopes and accelerometers are complementary, as gyroscopes are insensitive to linear motion or the Earth's gravity, but accelerometers can measure the rate of change of linear motion and orientation of the chip relative to the Earth. Software can combine this data to calculate the shape of the path being described by a device containing the chip, in addition to the changing velocity of the chip along that path.

The InvenSense MPU-6050 is a common 3-gyroscope, 3-accelerometer chip. It also includes an interface for connecting an external 3-axis magnetometer. The SPI and I2C communications protocols are supported. The MPU-6050 has been a popular choice in the hobby-electronics community, so that Arduino-compatible code to interpret its data is available from many sources. Breakout boards are available with the MPU-6050 installed. An example is the Sparkfun SEN-11028 shown in Figure 9-6.

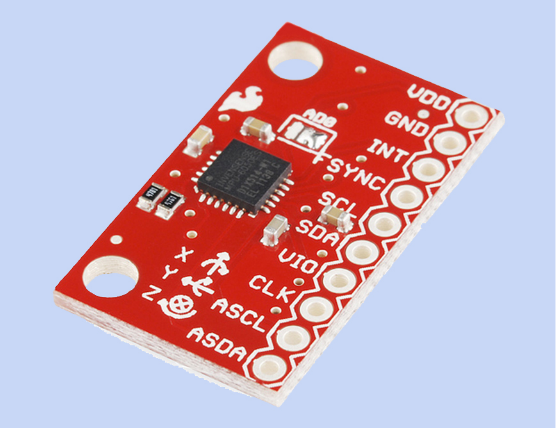

Figure 9-6 *A breakout board from Sparkfun, providing easy access to the InvenSense MPU-6050 chip combining three gyroscopes and three accelerometers.*

Values

The *rotational velocity* of a gyroscope element is usually expressed in degrees of rotation per second (dps), and sometimes in rotations per minute (RPM).

A datasheet will specify the number of sensor axes (usually 3), supply voltage (3.3VDC is common), maximum digital-low and minimum digital-high output voltages, and power consumption in normal mode and sleep mode. Power consumption is usually less than 10mA.

The *dynamic range* is the maximum forward and reverse rotational velocity, which usually will not exceed plus-or-minus 2,000 degrees per second. Lower ranges may be user-selectable. The advantage of selecting a lower maximum rate of change is that it can be converted to a digital value with higher precision.

The *sensor resonant frequency* will be several kilohertz, and must be higher than the frequency of any vibration that is applied to the sensor during use.

The *communications protocol* is usually I2C and SPI may be offered as an option, with a selectable digital output data rate.

Bias temperature coefficient describes the effect of temperature on the gyroscope.

The *resolution* of the gyroscope relates to the number of bits used in the digital output from the onboard ADC. A 16-bit resolution is common.

How to Use It

Using a smart chip such as the MPU-6050, the circuit designer can take advantage of its onboard digital motion processor (DMP). Still, obtaining orientation information from the contents of registers on the MPU-6050 is nontrivial. Online sources and code libraries are necessary. The book *Make: Sensors* contains code listings for the Raspberry Pi as well as the Arduino.

What Can Go Wrong

Temperature Drift

Vibrating materials at the heart of a chip-based gyroscope are likely to change their behavior with temperature. Typically the chip will contain a temperature sensor, the value of which can be used to adjust the output value of the gyroscope.

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The vibrating parts of a chip-based gyroscope may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Vibration

Because a chip-based gyroscope depends on the consistent behavior of internal vibrating parts, external vibration can degrade its accuracy. Sensor design can minimize the effects of vibration, but the datasheet should be consulted for details.

Placement

A gyroscope should be placed on a circuit board near a hard mounting point where deflection of the board will be minimized.

accelerometer

10

OTHER RELATED COMPONENTS

- **GPS** (see Chapter 1)
- **gyroscope** (see Chapter 9)
- **tilt** sensor (see Chapter 8)
- **vibration** sensor (see Chapter 11)

What It Does

Acceleration is the rate at which velocity changes over time. If a car takes 10 seconds to increase its speed from 30kph to 40kph relative to the road on which it is traveling, it is accelerating at an average of 1kph each second. If it then reduces its speed back to 30kph during another 10-second interval, it is decelerating at the same rate—although deceleration is really just negative acceleration.

While a car is accelerating, passengers will feel a lateral force exerted on them. Similarly, astronauts in a rocket that blasts off will feel a downward force. According to Einstein's theory of equivalence, forces resulting from acceleration are indistinguishable from the force of gravity.

Consequently, a sensor that measures acceleration can also measure gravity. This sensor is an **accelerometer**. Its output may be measured in *gravities*, abbreviated with the letter g (not to be confused with the usage of G to measure the strength of a magnetic field in gauss).

If three accelerometers are mounted orthogonally (at 90 degrees to each other), their readings can reveal:

- The direction of acceleration of a moving object.
- If an object has been dropped and is falling freely.
- Which way up it is being held in a stationary position.
- The severity of an impact when a moving object collides with some other object.

IMU

A **gyroscope** measures the rate of rotation of the enclosure in which it is mounted. This is properly known as the *angular velocity*. A gyroscope will also respond to changes in the rate of rotation. It does not measure linear motion or a static angle of orientation.

A **magnetometer** measures the magnetic field surrounding it, and may be sufficiently sensitive to determine its orientation relative to the Earth's magnetic field.

When an accelerometer and a gyroscope are contained in the same package, optionally with a magnetometer, they may be described as an *IMU* (inertial measurement unit), which can provide necessary data to maneuver aircraft, spacecraft, and watercraft, especially when **GPS** signals are unavailable.

Schematic Symbols

A chip-based accelerometer may be represented in a schematic as a rectangular box containing abbreviations to identify pin functions, as in any integrated circuit chip. No specific symbol is used for any of these components.

Applications

In the past, accelerometers were laboratory devices that calibrated the performance of cars, airplanes, and other types of vehicles. Measuring the ability of car tires to withstand cornering forces was an application where an accelerometer was used.

Miniaturization of accelerometer elements, coupled with a radical reduction in their cost, has enabled them to be installed in small electronic devices ranging from smartphones to hard drives.

In a phone or a camera, accelerometers can determine which way up the user is holding the device. The camera can rotate the picture appropriately, and the orientation of the picture can be saved with its image data.

In an external hard drive containing rotating platters, accelerometers can protect the read-write heads by rapidly parking them during the fraction of the second that elapses between someone dropping the hard drive and its impact with the floor.

Accelerometers can be installed in a 3D mouse or virtual-reality headset to determine its orientation and motion. This enables a video image to be updated appropriately. For example, the Nintendo Wii Remote has been marketed with an ADXL330 accelerometer.

In an automobile, an accelerometer can trigger the deployment of an air bag when the deceleration caused by an impact exceeds a threshold level.

How It Works

The simplest conceptual model of an accelerometer consists of a mass attached to one end of a coiled compression spring. The other end of the spring is anchored in an object whose acceleration is being measured. The mass can only move along the same axis as the spring.

Figure 10-1 shows three views of a simplified accelerometer, which is a sealed tube shown in dark red. In the center image, the accelerometer is in its rest state. The top image shows the mass (a dark blue square) responding when the tube accelerates from left to right. The third image shows it decelerating (that is, undergoing negative acceleration, or acceleration from right to left). Using an ideal spring, the displacement of the mass will be proportional to the rate of acceleration, within reasonable limits. The displacement can be measured optically or capacitively.

Note that the rest state will resume when the acceleration stops, regardless of constant motion in any direction. An accelerometer only measures a *change* in velocity. It does not measure a constant velocity.

An accelerometer cannot measure rotation around its own axis of movement. Therefore, it may be used in conjunction with a **gyroscope**, which measures angular velocity.

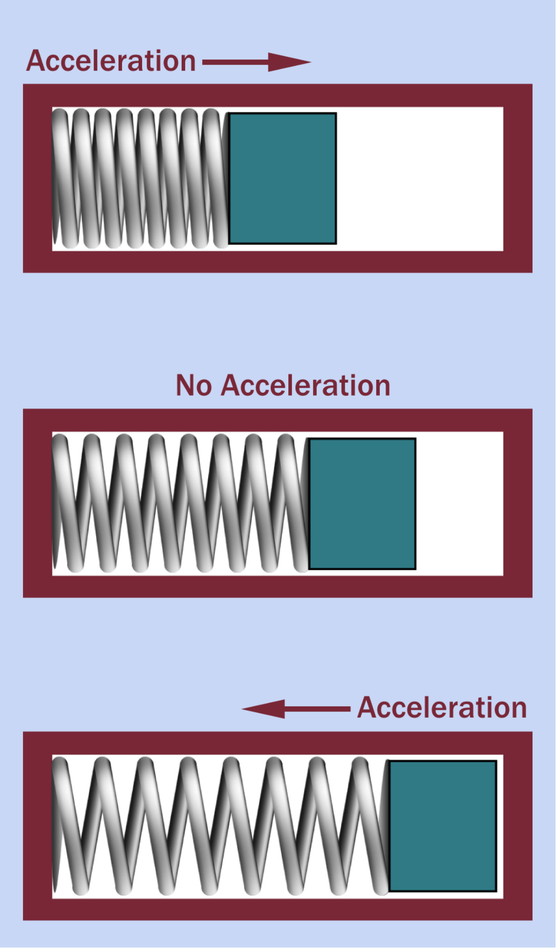

Figure 10-1 *Simplified view of an accelerometer consisting of a sealed tube (dark red outline) with a spring anchored to it at left. A small mass (dark blue) is attached to the right end of the spring. The mass responds to acceleration of the tube.*

Gravity and Free Fall

If a device containing the simplified accelerometer described previously rests on the ground, and is mounted vertically as in the left section of Figure 10-2, gravity acting on the mass will apply a force to one end of the spring, while the other end is restrained. The accelerometer will now measure 1g as a downward force.

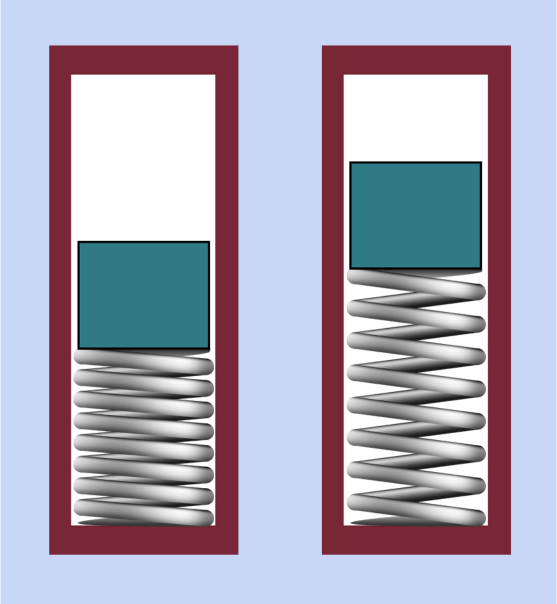

Figure 10-2 *Left: an accelerometer resting on the ground measures 1g, as the force of gravity pulls the mass downward. Right: in free fall, the accelerometer measures 0g.*

If the device is dropped so that it falls freely under the influence of gravity, it is in *free fall* and will accelerate at approximately 9.8 meters per second each second. This is usually written as 9.8m/sec^2, or can be described as 1 gravity, often expressed as 1g.

An accelerometer in free fall will measure 0g, as shown on the right in Figure 10-2, because all parts of the accelerometer are now accelerating equally under the force of gravity.

If three accelerometers are assembled orthogonally, and if one accelerometer is vertical, and the device is held motionless relative to the Earth, the vertical sensor will show 1g while the other two accelerometers will show 0g. If the device is dropped, all sensors will show 0g.

Rotation

If a device containing three accelerometers is turned over, and is not in free fall, accelerometers mounted orthogonally will show values that vary depending on their alignment with the force of gravity.

Calculation

The force acting upon an object will cause a rate of acceleration that can be calculated by Newton's Second Law of Motion, provided that the object can move freely and is not being subjected to any additional forces, such as the force of gravity. If F is the force, m is the mass of the object, and a is the acceleration:

```
F = m * a
```

And therefore:

```
a = F / m
```

If the mass is restrained by a spring whose compression or extension has an approximately linear relationship to the force applied to it, acceleration can be calculated as a function of the linear displacement of the mass.

These statements ignore relativistic effects that are insignificant unless ultra precise time and motion measurement may be involved.

In a real-world accelerometer, movement of the mass will require some form of damping to prevent it from oscillating.

Variants

Accelerometer prices dropped radically after 2010. In an effort to maintain profitability, manufacturers have loaded more features onto chips. While a 2-axis accelerometer such as the Memsic 2125 seemed a good option when first introduced, it is now facing obsolescence as 3-axis accelerometer chips that also contain 3-axis gyroscopes have become ubiquitous—and no more expensive.

Early chip-based accelerometers provided analog outputs where voltage was proportional with acceleration and could be processed by a comparator. On some breakout boards, such as the Dimension Engineering DE-ACCM6G, which used the STMicroelectronics LIS244ALH chip, a comparator was included (see Figure 10-3).

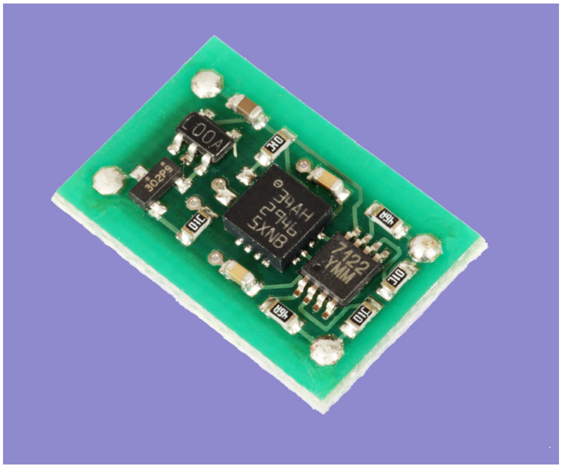

Figure 10-3 *A relatively early, relatively expensive breakout board using the 2-axis accelerometer LIS244ALH chip by STMicroelectronics. It has been superceded by chips that combine accelerometers, gyroscopes, and processors to provide a digital output.*

Because this board allowed a maximum output of only 0.83mA, it was only suitable for high-impedance logic chips or a microcontroller. However, because the output was analog, it could be passed through another comparator for direct connection to a piezo beeper to create a device that would sound an alarm when tilted. This is shown in Figure 10-4.

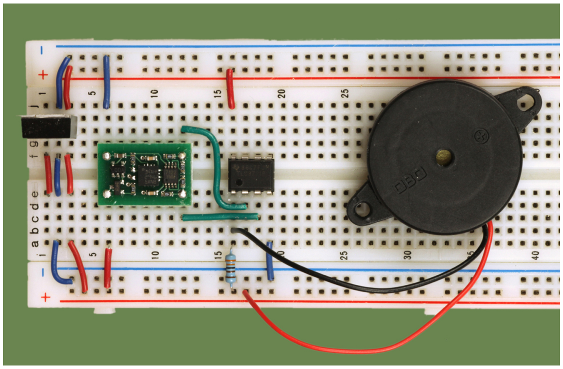

Figure 10-4 *Connecting the analog output from the DE-ACCM6G 2-axis comparator breakout board through a comparator to a piezo beeper.*

Many chips now contain gyroscopes as well as accelerometers. The electron micrograph in Figure 10-5 shows the interior of a chip of this kind, where the zig-zag shapes are "springs" etched into the die, the large areas patterned with dots are masses that can respond to various forms of motion, and the parallel plates are capacitive sensors.

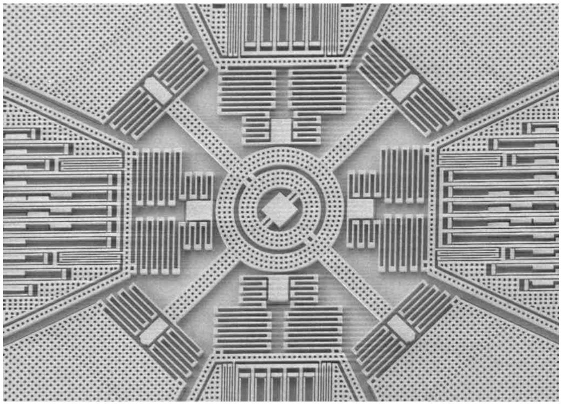

Figure 10-5 *Interior of a chip combining gyroscopes with accelerometers.*

The complexity of a chip combining two types of sensors creates a need for more complicated code to process the six outputs. Consequently most accelerometer chips now have their own onboard ADCs, and digital registers for communication with microcontrollers via the I2C communication protocol. Some chips also have onboard processing to interpret the mix of data. However, microcontrollers still need code to make sense of the data.

The hobby-electronics community has responded. A breakout board such as the LSM9DS0 from Adafruit, using a chip from STMicroelectronics that shares that same part number, tries to make this extremely complex chip accessible to experimenters. Shown in Figure 10-6, the LSM9DS0 contains a 3-axis magnetometer in addition to a 3-axis gyroscope and a 3-axis accelerometer.

The datasheet for the LSM9DS0 runs to more than 70 pages. At the time of writing, Adafruit is still refining code for the Arduino to make the features of this chip accessible.

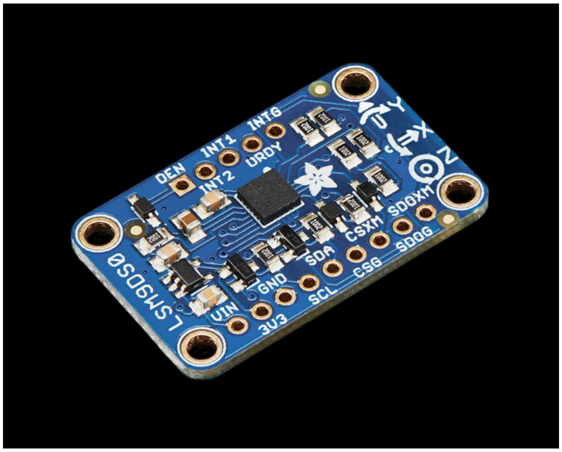

Figure 10-6 *This breakout board from Adafruit is built around an LSM9DS0 chip that combines accelerometers with gyroscopes and magnetometers.*

Despite the complexity of the LSM9DS0 chip, the breakout board sells for approximately the same price, at the time of writing, as the DE-ACCM6G 2-axis analog-output magnetometer four years ago. In the future we may expect IMU prices to fall still further, so that this type of multifunction chip becomes the default, and users simply ignore the functions that they don't need.

Values

Current consumption will vary depending on the activity of the chip, and may be broken down for different types of sensors on multi-sensor chips. A modern accelerometer may typically draw less than 1mA. Gyroscope power consumption will be greater, as a segment of the chip will be maintained in a state of vibration.

Linear acceleration measurable by an accelerometer is customarily expressed in gravities, abbreviated g. Current generations of chips may be able to measure as much as plus-or-minus 16g, but because the value is converted internally to a digital quantity, smaller accelerations will not be expressed so accurately. Therefore the measurement range for acceleration is usually user-selectable by sending an appropriate code to the chip. Ranges may include plus-or-minus 2g, 4g, 6g, 8g, and 16g.

Sensitivity defines the smallest increment of acceleration measurable by the least significant bit (LSB) in an output register, for each of the acceleration ranges. In a range of plus-or-minus 2g, the internal 16-bit ADC may be capable of measuring 0.06 milligravities. When the range is reset to plus-or-minus 16g, the smallest increment may be around 0.7 milligravities.

The measurable range of gravities should not be confused with the maximum acceleration that the chip can tolerate without suffering internal mechanical damage, either while it is powered or unpowered. This will be more than 1,000g provided the duration is brief, and would only be experienced during an impact.

Linear acceleration sensitivity change versus temperature is usually expressed as a percentage, such as plus-or-minus 1.5%.

Output type will be analog or digital. If digital, the data protocol will be I2C or SPI. If I2C, the address of the device should be configurable. The data transfer rate is typically at least 100kHz, and this also may be configurable to different rates.

For additional details about protocols such as I2C, see Appendix A.

What Can Go Wrong

Mechanical Stress

Stress can be induced when a surface-mount chip is soldered to a board. The moving parts in a chip-based accelerometer may be adversely affected. Datasheets will supply information regarding maximum acceptable temperature during the soldering process.

Other Problems

If the accelerometer function is combined with other sensing functions such as magnetometer or gyroscope, see Chapter 2 or Chapter 9 for additional cautions about potential problems affecting these kinds of sensors.

vibration sensor

11

An **accelerometer** can measure some aspects of vibration. However, this entry deals primarily with mechanical devices (often described as *vibration switches*) and piezoelectric devices (often described as *vibration sensors*) that are solely intended to measure vibration.

A *vibrometer* measures vibration using a laser beam aimed at a reflective spot applied to a surface. It is usually a laboratory instrument, beyond the scope of this Encyclopedia.

OTHER RELATED COMPONENTS

- **accelerometer** (see Chapter 10)
- **tilt** sensor (see Chapter 8)
- **force** sensor (see Chapter 12)

What It Does

A vibration sensor responds to repetitive mechanical motion. Most versions contain two switch contacts that are normally open and will close if the sensor vibrates in its designed frequency range. In some sensors, the frequency range and sensitivity are manually adjustable.

Large sensors are used as automatic shutdown switches responding to excessive vibration in machinery, and may be capable of switching substantial currents (10A or higher). Smaller versions can shut down domestic appliances such as a washing machine that is seriously out of balance during a spin cycle.

A vibration sensor can be used as a simple user-input device in toys and games.

A shock sensor can detect abuse of a sensitive device, for example, by including the sensor and a data logger when the device is transported.

Schematic Symbols

Either of the symbols in Figure 11-1 may represent a piezoelectric or piezoresistive vibration sensor, but they also represent other piezo-based devices.

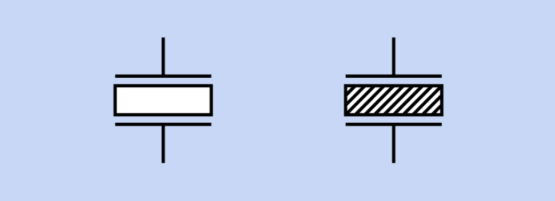

Figure 11-1 *Either of these symbols may represent a piezoelectric or piezoresistive device, including (but not limited to) vibration sensors that operate on this principle. The symbol on the left is more common.*

Variants

Vibration sensors use a wide variety of detection methods.

Pin-and-Spring

Probably the simplest type of sensor consists of a small, thin pin in the center of a miniature coil spring. The spring is anchored at its base, but its other end is free to vibrate. If the vibration reaches a sufficient amplitude, the spring touches the pin, completing a circuit between the two leads of the device.

An example is shown in Figure 11-2, where two identical sensors are shown, one of them cut open to reveal the gold-plated rod and spring inside. This sensor is rated for 10mA at up to 12VDC.

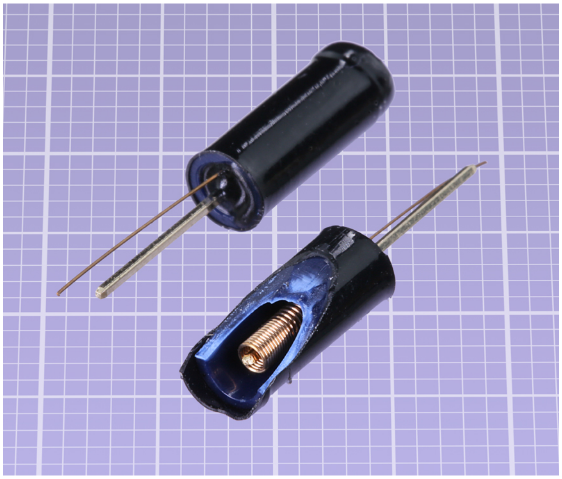

Figure 11-2 *When the spring in this sensor vibrates, it touches the pin that is centered in it. The background grid is in millimeters.*

Advantages of this system are low cost, ability to respond along two out of three axes, no power supply requirement, and ability to switch AC or DC. However, because the internal contact is extremely brief, it must be connected with a latching component of some type. A flip-flop could be used, or a 555 timer. The switch may also be connected with an input pin on a microcontroller, provided a pullup or pulldown resistor is used to prevent the input from floating when the switch is open.

Externally, the packaging of a pin-and-spring vibration sensor is almost indistinguishable from a small **tilt** sensor that contains one or two spherical metal balls. The tilt sensor may respond to vibration, but only of a large amplitude and low frequency.

A miniature board containing a pin-and-spring vibration sensor, comparator, and trimmer potentiometer for sensitivity control is sold cheaply by the Chinese supplier Elecrow as their product SW-18015P, shown in Figure 11-3. Elecrow also offers a wide range of other low-cost sensors.

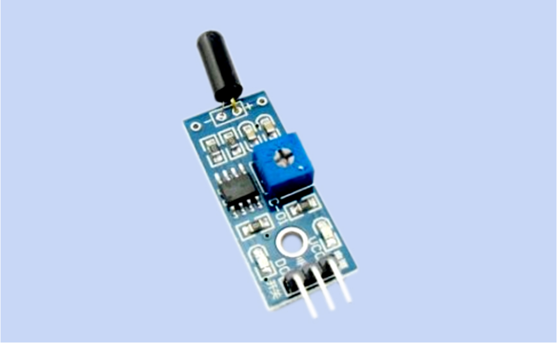

Figure 11-3 *A pin-and-spring vibration sensor mounted on a board with sensitivity control.*

Piezoelectric Strip

The LDT0-028K by Measurement Specialties is a section of piezoelectric polymer film laminated to a polyester substrate. The film is designed to be anchored at one end, allowing the other end to vibrate. An unweighted version and a weighted version are shown in Figure 11-4, each measuring about 13mm x 25mm. Addition of the weight alters the resonant frequency of the sensor.

Deflection of about 2mm is sufficient to generate a surprising 7VDC between the two leads. Larger deflections will generate higher voltages. The manufacturer suggests that direct connection to a CMOS component is possible. An op-amp may be used for signal conditioning.

A piezoelectric device of this type only generates voltage during the process of deflection. If the strip is held in a curved position, its output diminishes to zero.

This sensor has a resonant frequency around 170Hz when there is no weight attached to its free end.

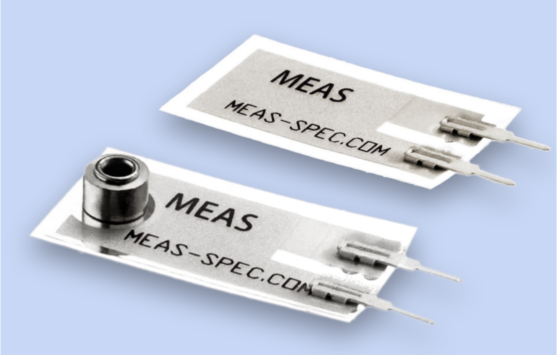

Figure 11-4 *Two versions of the LDT0-028K vibration sensor from Measurement Specialties, one with an added weight to lower its resonant frequency.*

Chip-Based Piezoelectric

The Murata PKGS series is an example of a surface-mount piezoelectric shock sensor. Measuring only about 1mm x 2mm x 4mm it has an analog output designed for connection through an op-amp. The manufacturer suggests application in a hard-disk drive to block read-write operations when vibration occurs. Similarly, these shock sensors may be used in CD-ROM or DVD drives. They may also be installed in cash dispensing machines to sound an alarm if vandalism occurs.

Toshiba's TB6078FUG is a similar product. Note that because these devices contain electronic components, they require a power supply (usually 3.3VDC to 5VDC) for operation.

"Mousetrap" Type

Some vibration switches rely on a simple system of leverage that is comparable to a mousetrap, in that a relatively small stimulus releases a strong spring. In Figure 11-5 the upper part of the figure shows the switch at rest, held in its position by a powerful spring and by the weight of a mass attached to a pivoted arm. In the lower part of the figure, severe vertical vibration has caused the assembly to move up and down with sufficient energy to overcome the tension in the spring while the inertia of the mass has resisted the motion. Consequently the arm has moved past the position where the spring is aligned with the pivot, and the spring now acts to hold the arm against a snap-action switch. A system of this type is used in sensors on some power-station cooling towers, where the loss of a large fan blade can result in major vibration.

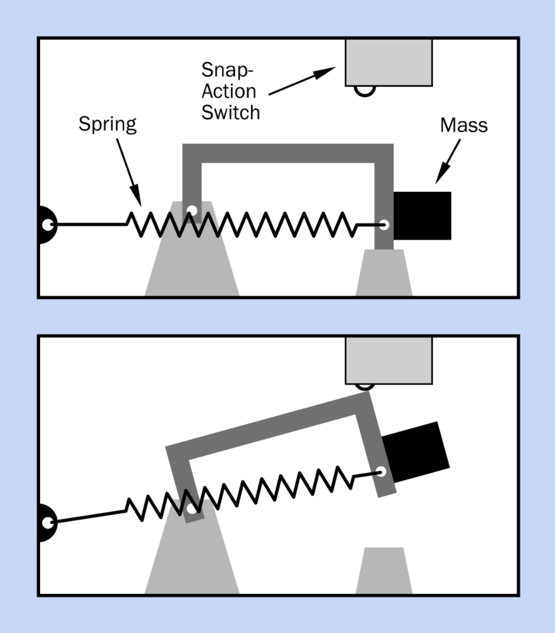

Figure 11-5 *A spring-loaded vibration sensor.*

Magnetic

Typically this method is used to detect excessive vibration in machines or other devices containing heavy rotating mechanical components. The sensor may be physically large, moving switch contacts that are designed to handle currents of 1A or significantly more.

In one system, a steel ball is retained by a permanent magnet that is barely powerful enough to prevent the ball from falling. Excessive low-frequency vibration will dislodge the ball, which falls and completes a circuit between two contacts. This activates a relay that powers down the piece of machinery that is vibrating.

The ball may be spaced a small distance from the magnet by a beveled nonmagnetic seat, and the seat may be movable with an external screw. This will adjust the sensitivity of the switch. After the switch has been triggered, it must be reset, which may entail using an external lever to raise the ball back to its location near the magnet.

Another magnetic system is shown in Figure 11-6, where a magnet on a vertical arm can be displaced by horizontal vibration, and an inertial mass on a spring-loaded horizontal rod can also dislodge the magnet in response to vibration along the other two axes.

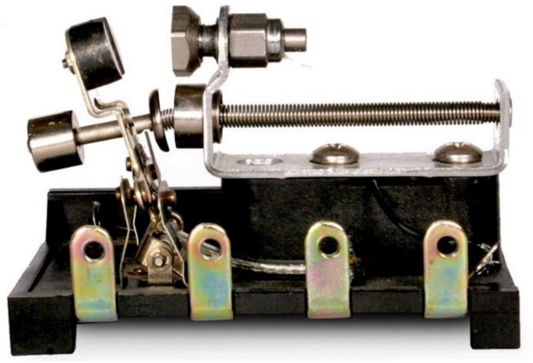

Figure 11-6 *A magnetic vibration switch. See text for details.*

Mercury

A small *mercury switch* may be used as a vibration sensor, although this application is uncommon. See Chapter 8 for more information about mercury switches.

Values

Measurement of vibration is a complex science of special interest in mechanical design, especially in areas such as the powertrain and suspension geometry in an automobile. Only a few fundamentals will be summarized here.

Primary Variables

The four primary variables in vibration are frequency, displacement, velocity, and acceleration. Frequency describes how rapidly the vibration occurs; displacement describes how far the vibrating object moves in each direction; velocity describes how fast it moves during each cycle; and acceleration describes how rapidly the velocity changes during each cycle. Different types of sensors can be chosen for their responsiveness to each attribute.

Figure 11-7 shows the theoretical relationships between displacement, velocity, and acceleration plotted against the frequency of vibration. The y axis of this graph (the vertical axis) is often labeled “amplitude,” but in reality it is being used to measure three different units, as shown. The curves indicate that if the velocity of vibration remains constant while frequency increases, acceleration must increase as a function of the frequency while displacement decreases. The acronym “rms” denotes that the values are measured as the “root mean square” of their fluctuations.

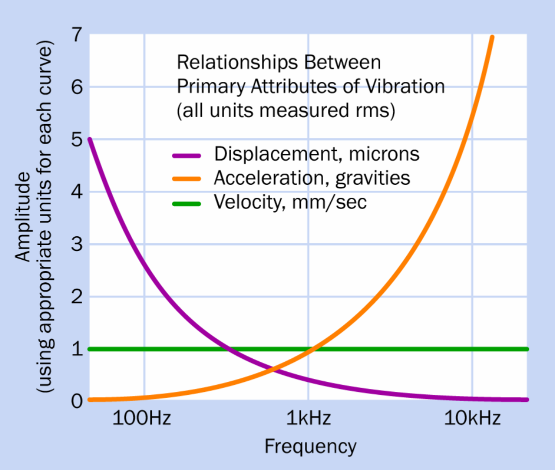

Figure 11-7 *Theoretical relationships among the primary attributes of vibration.*

Mechanical sensors or switches that respond to displacement are best suited to low frequencies, while piezoelectric sensors that are sensitive to acceleration are best suited to higher frequencies. Serious mechanical problems tend to result in low-frequency vibrations, while wear on bearings and in gear trains will be more likely to create high-frequency vibration.

Dynamic Attributes

Datasheets for small piezoelectric sensors may only show basic values such as the voltage range that is likely to be created.

Small mechanical sensors of the pin-and-spring type will be rated for maximum voltage and switching current. Typical values may be around 12V (AC or DC) and 10mA, indicating that the output from this type of sensor should be used with an op-amp, microcontroller, logic chip, solid-state relay, or other semiconductor with a high-impedance input.

Sensors that are manufactured for industrial applications will be rated for attributes such as measurable acceleration range (in gravities), temperature sensitivity, frequency response, resonant frequency, capacitance, and power requirement.

The sensitivity of piezoelectric sensors is usually expressed in mV/g. This type of sensor will require a comparator to process the tiny amount of current that it creates.

How to Use It

If a vibration sensor has an analog output requiring a comparator, the output from the comparator is likely to be of the open-collector type. This will require a pullup resistor of a value that provides an appropriate voltage for the next stage in the circuit. For more information about comparators, see Volume 2. For more information about the use of an open-collector output, see "3. Analog: Open Collector".

A *coupling capacitor* can remove the DC component from the comparator output, allowing only the frequency of vibration to pass through. The choice of capacitor value will depend on the frequency.

When using a piezoelectric sensor with analog output, a 10M resistor may be installed across its two terminals to reduce voltage drift.

The primary challenge in getting a vibration sensor to work successfully will be matching it to the source of vibration. Manufacturers' datasheets for chip-sized sensors often provide very little information about the optimal values for their products. Peak performance will occur when the natural resonant frequency of a sensor is close to the frequency of vibration that it must detect. Trial and error may be necessary.

A sensor must be mounted appropriately. Most sensors are directional, at least to some extent, and many will not respond significantly to vibration at 90 degrees to their primary axis of sensitivity. Their performance will also diminish if they are placed too far from the vibration source, or if they are mounted on a flexible or yielding surface that will tend to absorb vibration.

While industrial vibration switches may be adjusted manually, the response of small devices designed for circuit-board mounting can only be tweaked using external components to filter out unwanted signals from the sensor.

What Can Go Wrong

Long Cable Runs

The output from a piezoelectric vibration sensor is primarily an AC signal, fluctuating at the frequency of the vibration. Long cabling, or inadequately shielded cabling, can potentially introduce capacitive effects that can degrade the sensor signal. This issue will only affect higher frequencies.

Interference

Sensor signals can also be affected by electromagnetic interference from power lines, transformers, and large motors. This is a significant issue, as a primary application for industrial vibration sensors is to measure vibration created by motors.

Correct Grounding

For large sensing equipment, grounding may be important to shield cables that transmit data. In an industrial environment, however, grounding is primarily motivated by safety considerations, and an electrical ground can carry unwanted interference. Ground loops may be created if there are multiple ground points. Ideally, a "ground tree" should be used, where there is only one primary grounding point, and grounds to equipment branch out from it.

Fatigue Failure

In installations where some vibration normally exists, cables should be anchored properly to minimize the risk of fatigue failures.

force sensor

12

A *load cell* or *load sensor* is generally intended to measure a static load, while a *force sensor* can respond dynamically. However, this semantic distinction is not always observed. This entry differentiates between load cells and force sensors but includes them both.

Traditional types of *hydraulic* and *pneumatic* load sensors are not electronic devices. They are outside the scope of this Encyclopedia.

Capacitive, *ultrasonic*, *magnetic*, *optical*, and *electrochemical* force sensors are relatively unusual, and are not included in this Encyclopedia.

A force sensor is occasionally described as a **pressure** sensor, but that term is ambiguous, as it is more often used in conjunction with fluids. This Encyclopedia assumes that a pressure sensor only measures gas or liquid pressure. See Chapter 17.

A **vibration** sensor reacts to rapidly changing forces, but usually cannot measure them accurately, and is simply triggered when vibration exceeds a threshold. See Chapter 11.

Impact sensors that measure the force of a collision are outside the scope of this Encyclopedia.

A sensor designed to respond to a single touch from a fingertip is considered a human-input device and is discussed in the **single touch** sensor entry. See Chapter 13.

OTHER RELATED COMPONENTS

- **vibration** sensor (see Chapter 11)
- **single touch** sensor (see Chapter 13)

What It Does

A **force** sensor measures physical force that is applied to it, either by a person or by an object. Many force sensors respond rapidly and can measure fluctuating forces.

A *load cell* or *load sensor* is usually intended to measure the static weight of an object.

Applications

In robots, force sensors can provide feedback to limit the grip of a mechanical hand. They can also provide haptic feedback for a surgeon performing robotic surgery. In the future, force sensors may find increasing application in agriculture, as mechanized handling of fruit and other foods requires a carefully controlled gripping force.

In medicine, the use of force sensors to evaluate muscle strength in hands or limbs can be important as an indicator of neurological problems or to monitor progress in occupational therapy. Thin force sensors can be installed in shoes to check the weight distribution of each foot. They can also be used for entertainment purposes, to light LEDs in sneakers.

A force sensor may respond to a single-touch user input. See Chapter 13 for more information about **single touch** sensors. Some video game controllers use resistive force sensors to measure the amount of pressure applied to a button. The PlayStation is an example. (Old PlayStation DualShock 2 controllers are a salvageable source of pressure-sensitive buttons.)

Load sensors are used to weigh industrial products, and are also used domestically in kitchen and bathroom scales.

A load sensor can also detect human presence —for example, in the passenger seat of an automobile, where an air bag must not deploy if a young child is present, or in a hospital, to monitor the number of times the patient gets out of bed.

Schematic Symbol

No specific schematic symbol is used for either a force sensor or a load sensor. If a force sensor uses a piezoelectric or piezoresistive element, it may be represented by the symbol shown in Figure 11-1, which is used for many piezo devices.

How It Works

Two methods of force measurement are commonly used: resistive and piezoelectric.

A *piezoelectric* force sensor uses a piezoelectric element, often consisting of a quartz crystal, to convert force to a small voltage that can be amplified. However, this type of sensor only responds to changes in force. If a constant load is applied, the output peaks quickly and then gradually diminishes to zero.

Resistive force sensors change their electrical resistance when force is applied. They include *metallic strain gauges* and *plastic-film* sensors in which two layers of conductive ink are pressed together.

In SI (standard international) units, the force needed to activate a sensor is measured in newtons, abbreviated with a capital letter N. A newton is defined as the force that would accelerate a mass of 1 kilogram at 1 meter per second each second. More practically, in the gravitational field at the surface of the Earth, 1N = about 100 grams of weight. There are about 28 grams in an ounce; thus 1N is slightly less than 4 ounces.

Strain Gauge

A strain gauge is often made from metallic foil applied to an insulating flexible backing. The backing is glued to a shaped piece of metal, usually steel or aluminum, which is designed to flex slightly under pressure and may be referred to as a *spring*, even though it is often one solid object. Its deflection will be related to the force imposed on it.

The maximum deformation of the spring under a strain gauge is usually 500 to 2,000 parts per million (ppm) when subjected to the maximum force that it is designed to measure. A change of 1ppm is referred to as a *microstrain* (abbreviated με).

The strain gauge has no polarity, and functions like a force-controlled **potentiometer** (see Volume 1). The ratio of the change in its resistance to the change in the strain that it experiences is called the *gauge factor*. For metal foil gauges, the gauge factor is usually around 2.0. This is an approximately linear relationship.

The most common type of foil pattern is shown in Figure 12-1. In the figure, if a stretching force is applied horizontally, the multiple thin sections of foil are slightly elongated, and their resistance increases within the limits of elasticity of the foil. This effect is multiplied by the number of sections. If the stretching force is applied vertically, the sections are merely separated slightly, and the result is negligible.

Wheatstone Bridge Circuits

The very small changes in resistance in a strain gauge must be amplified to be usable, and the first step is to use a Wheatstone bridge circuit. The simplest form of this circuit is shown in Figure 12-2.

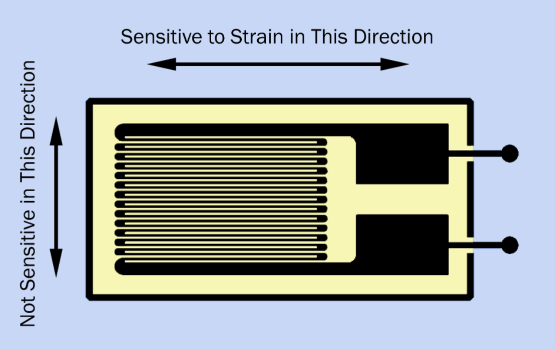
Sensitive to Strain in This Direction
Not Sensitive in This Direction

Figure 12-1 *The pattern of metallic foil used in a typical strain gauge.*

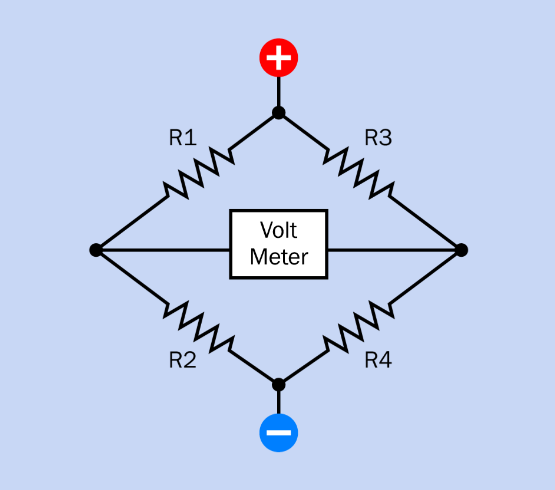

Figure 12-2 *A demonstration version of the basic Wheatstone bridge circuit that is often used to detect small changes in a resistance.*

Each pair of resistors (R1 + R2, and R3 + R4) functions as a voltage divider. If all the resistors have an exactly equal value, the voltage at the midpoint of each pair will be identical, and the volt meter at the center will have a zero reading. However, if the value of one resistor changes slightly, the meter will register the imbalance. This circuit is commonly used because of its sensitivity to small variations.

In Figure 12-3 two strain gauges have been substituted for resistors R3 and R4. The upper strain gauge has been mounted so that it experiences an increase in force at the same time that the lower strain gauge experiences a decrease, usually because one gauge is mounted on the top side of a flexing element while the other gauge is mounted on the underside.

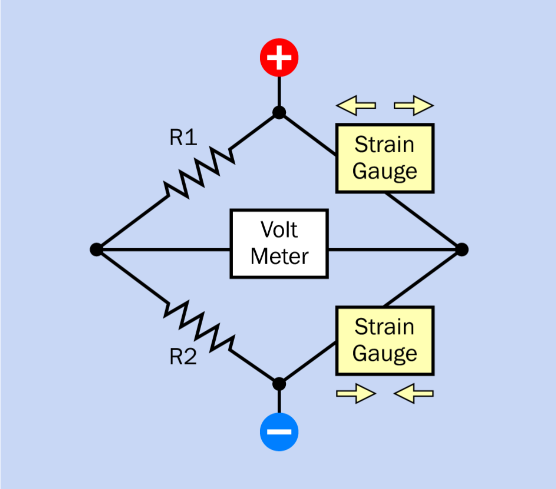

Figure 12-3 *Two strain gauges, oppositely oriented, can be used as resistances in the Wheatstone bridge circuit.*

Using two strain gauges in this way doubles the sensitivity of the Wheatstone bridge circuit. The configuration is known as a "half Wheatstone" force sensor, and will have three connecting wires. One will be black, one will be red, and the third will be a different color. The red and black wires are for connection to the power supply, as shown in the schematic, while the third wire is common and should be considered as an output.

If an additional two strain gauges are inserted in the Wheatstone bridge circuit, this is now a "full Wheatstone" force sensor (see Figure 12-4). Note, however, the diagonally symmetrical orientations of the strain gauges, necessary to multiply their effect.

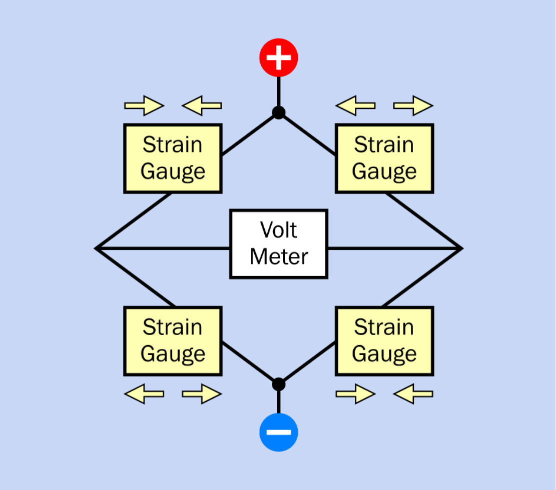

Figure 12-4 *Two additional strain gauges create a "full Wheatstone" circuit.*

A typical digital bathroom scale contains two half-Wheatstone force sensors, wired to create a full-Wheatstone configuration.

The force sensors shown in Figure 12-5 are rated for up to 50kg each, and can thus weigh up to 100kg if they are combined in a scale. In this figure, one sensor has been turned over to show its underside. The strain gauges are hidden in each sensor where the steel sections overlap.

Wheatstone Bridge Errors

Where more than one strain gauge is used in a Wheatstone bridge, they should ideally have identical performance. Since this is impossible as a result of manufacturing tolerances, some error correction is built into devices using multiple strain gauges.

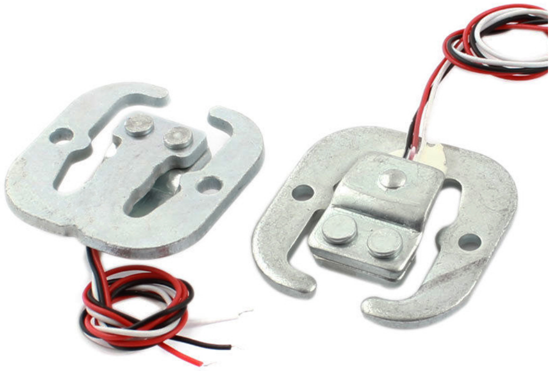

Figure 12-5 *Two force sensors suitable for a bathroom scale, each containing a pair of strain gauges that will flex oppositely under load.*

Strain-Gauge Amplification

The voltage output from a Wheatstone bridge circuit is given by the following formula, where V_{IN} is the supply voltage, V_{OUT} is the output, and R1 through R4 are the resistance values that were used in Figure 12-2.

$$V_{OUT} = [(R3/(R3+R4)) - (R2/(R1+R2))] * V_{IN}$$

The good news is that the output when a strain gauge is used will vary linearly with the load applied. The bad news is that it will be very small.

To amplify it, an op-amp such as the AD620 is often recommended. Using external resistors, its amplification factor can be adjusted from 1:1 to 10,000:1. Alternatively a chip such as the HX711 by Avia Semiconductor contains a 10-bit analog-to-digital converter and is specifically designed for use in weighing scales. Its digital output uses a very simple serial format. Sparkfun sells a breakout board incorporating this chip. See Figure 12-6.

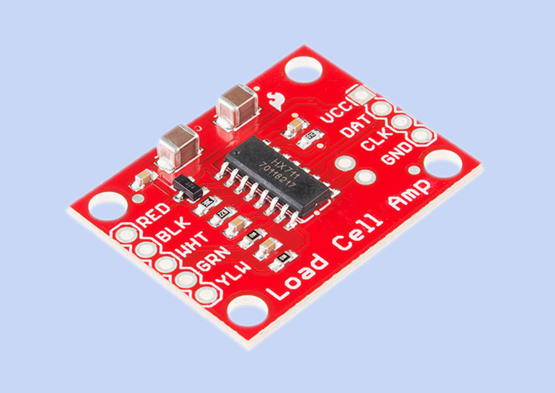

Figure 12-6 *Sparkfun HX711 breakout board for the HX711 amplifier chip, specifically designed for a full-Wheatstone array of strain gauges in a force sensor.*

Other Strain-Gauge Modules

Strain gauges are built into a variety of sensor modules. Figure 12-7 shows a selection, all of which are available from Sparkfun. Many more can be found online.

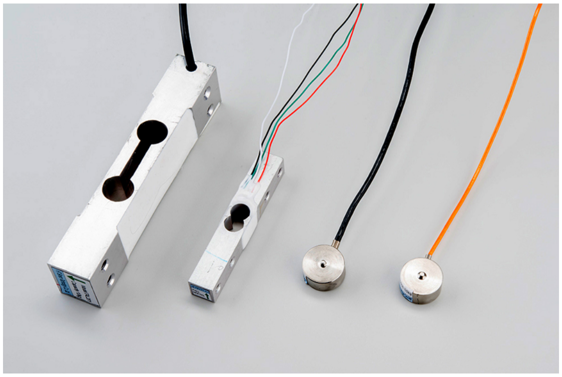

Figure 12-7 *A selection of load cells containing strain gauges, available from Sparkfun.*

Plastic-Film Force Sensors

Plastic-film resistive sensors contain two layers of conductive ink, sealed between two layers of thin, transparent plastic film. The resistance between the ink layers diminishes when they are pressed together. It may vary from as little as 30K when fully loaded to more than 1M when unloaded.

Like a strain gauge, this sensor has no polarity and requires no power supply.

Examples of this type of sensor are shown in Figures 12-8 and 12-9. Other shapes and sizes are available. Manufacturers include Tekscan, whose product is named FlexiForce; Alpha Electronic, in Taiwan; and Interlink Electronics.

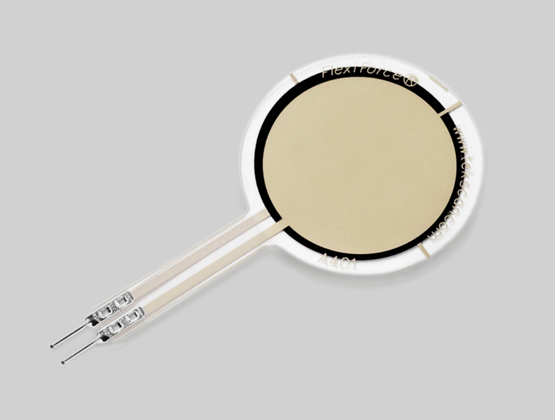

Figure 12-8 *A FlexiForce A401 resistive film sensor made by Tekscan, Inc. Its sensing area is 25.4mm in diameter (1 inch) and is rated to measure up to 111N (25lbs).*

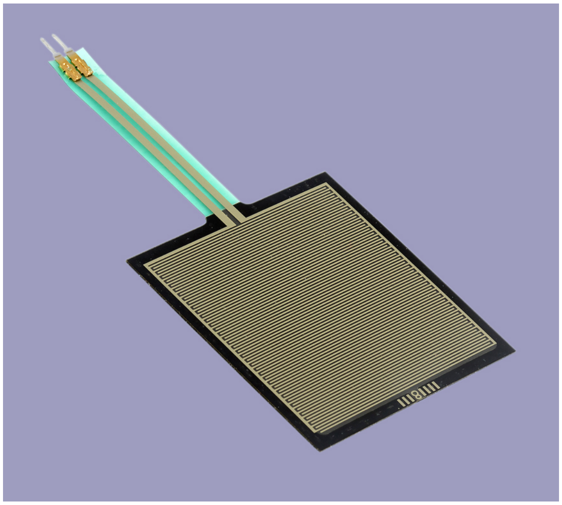

Figure 12-9 *An Interlink FSR406 resistive film sensor. Its sensing area is slightly less than 40mm square, and is rated to measure up to 20N (4.5lbs).*

Plastic-film sensors should not be confused with film-based piezoelectric vibration sensors, which are described in the **vibration** entry (see "Piezoelectric Strip"). Those sensors provide a transient output when they flex rapidly. The resistive sensors provide a stable output in response to a steady load.

Deformative Force Sensors

A sheet of natural or silicone-based rubber can be impregnated with conductive particles. The conductivity of the sheet may not change significantly when it is compressed, but if it is separated from a metal plate by a mesh of thin nylon fibers, compression will result in greater conductivity by pushing the rubber into the gaps between the fibers. See Figure 12-10.

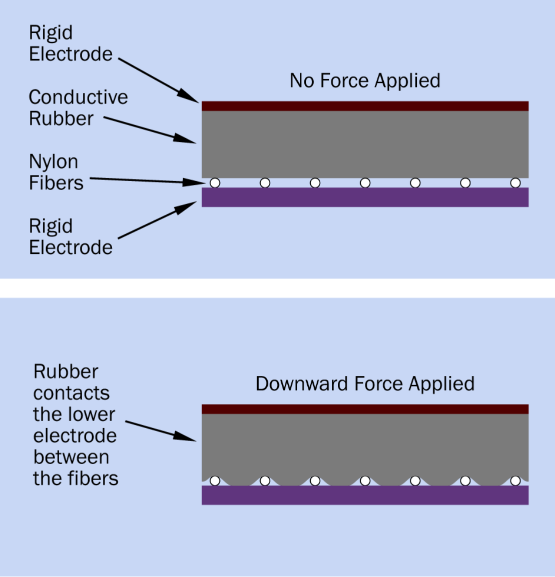

Figure 12-10 *In the top image, a flexible conductive layer is separated from an underlying rigid electrode by nylon mesh. In the lower image, a load applied to the flexible layer has forced more of it into contact with the electrode, reducing the resistance between them.*

Improvised Resistive Sensors

Polyethylene film impregnated with carbon particles is available under the brand name Velostat, owned by 3M. Although it was developed as an antistatic packaging material for semiconductors, it can be used to make a DIY force sensor. When the material is stretched, the embedded particles are more dispersed, and electrical resistance increases. When the material is compressed, its resistance is reduced.

Antistatic foam of the type used to package CMOS components can be used in the same way, although some types behave like memory foam, being slow to recover from pressure. The foam can be sandwiched between a pair of copper-plated circuit boards as electrodes.

How to Use It

Plastic-Film Resistive Force Sensors

The conductivity of a plastic-film sensor has an almost linear relationship with the force applied. In other words, if F is the force and I is the current:

```
I = k * F
```

where k is a constant determined by the characteristics of the materials used.

By Ohm’s Law, R = V / I where V is the voltage drop across a resistance of value R. By substitution, using k * F instead of I:

```
R = V / (k * F)
```

Therefore, if a constant voltage is applied across the force sensor, the resistance of the sensor will be proportional to the reciprocal of the force (i.e., 1 / F). These relationships are shown in Figure 12-11.

For convenient measurement, it will be helpful if the resistance of the sensor can be converted to a voltage output that varies linearly with the force applied. To achieve this, a resistive force sensor of this type is customary amplified with an **op-amp** (see Volume 2). Using the schematic shown in Figure 12-12, the amplification ratio, A, is found from this formula:

```
A = 1 + (R2 / R1)
```

where R2 if the potentiometer, and R1 is the resistive force sensor. Therefore, the output from the op-amp should have an approximately linear relationship with the force applied to the sensor.

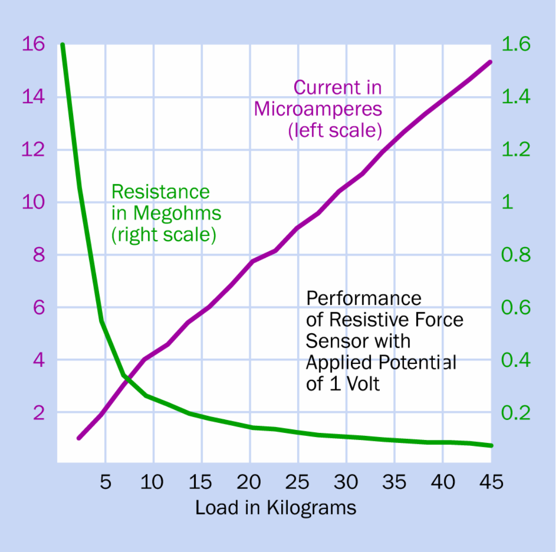

Figure 12-11 *The relationship between force, current, and resistance in a flexible resistive sensor, assuming a constant voltage of 1V is applied across it (derived from FlexiForce Sensors User Manual).*

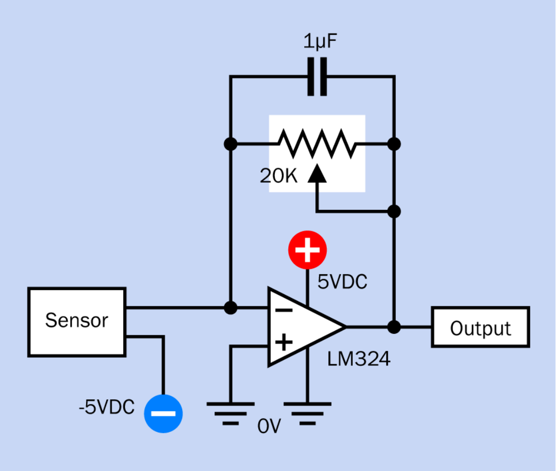

Figure 12-12 *An amplification circuit with an output that is approximately linear with force applied to a resistive film sensor.*

The capacitor in this circuit is added to suppress noise that the circuit might otherwise pick up.

An alternative is to wire a resistive sensor in series with a capacitor, and connect the other side of the capacitor to a comparator that has adjustable feedback. The resistance of the sensor will determine how quickly the capacitor charges. However, because of the mathematics describing the charge rate, the output from the comparator would not be linear with the force on the sensor. Also, provision would be necessary to discharge the capacitor intermittently.

Values

Film-Based Force Sensors for User Input

A very light pressure with a fingertip could be around 50g. A more defined finger-press would be 250g, and a heavy push with a finger would be around 1kg.

The Interlink range of flexible force sensors requires a minimum pressure of 0.2N, or about 20g. Similarly, Alpha products range from a minimum of 10g to 30g.

These specifications suggest that film-based force sensors may be used for one-touch user input, but the no-load resistance will be at least 1M, and more than 10M in some instances. A small amount of pressure is unlikely to reduce this electrical resistance much below 500K. Using an op-amp or comparator to detect that difference and convert it into a reliable on-off output may be vulnerable to noise and power-supply disturbance.

Another consideration is that film-based sensors provide no tactile feedback. For these reasons, and because film-based products are described by their manufacturers as “force sensors” rather than as “touch sensors,” they are included in this entry rather than in Chapter 13. That said, they may be considered as an option for one-touch user input where they are appropriate—for instance, in games where players are likely to slap or hit a sensor vigorously.

Specifications for Film-Based Force Sensors

Durability of film-based sensors is excellent, with manufacturers claiming that performance is not degraded after 1 million applications of a 20kg load. Sensors have no polarity, will function using voltages from 1V to 15V in most cases, and have a response time of less than 5ms. They do not generate, and are not vulnerable to, electromagnetic interference.

Important attributes when evaluating film-type sensors are their limits of tolerance for applied force, and resistance values at each end of the range. Unfortunately these values are poorly documented for many sensors, or may be unspecified.

Maximum force may range from 20N to 440N, depending on the brand and model of sensor.

Electrical resistance is at least 1M when unloaded, and may be as high as 20M.

Accuracy ranges from plus-or-minus 2% to plus-or-minus 5% from one application of force to the next, depending on the model of sensor and the manufacturer. If force is not applied each time in exactly the same area of the sensor, results will vary. If a sensor is swapped with another sensor of the same type, sensor-to-sensor consistency will be uncertain. Therefore, film-based force sensors are not a good choice for applications where accuracy is important.

Sensors may be rated for 5% to 10% hysteresis.

The area of active detection may range from about 4mm wide (FlexiForce sensors) to more than 40mm wide (Interlink FSR-406).

Strain Gauges

Strain gauges are not sold as individual components by most electronics suppliers. They are sometimes available as surplus parts or from sites such as eBay, where specifications range from 100 ohms to 1K as the resistance when no load is applied.

A load sensor on which a strain gauge has been preinstalled is much easier to use, and will be plug-compatible with an appropriate amplifier chip as previously described.

What Can Go Wrong

Soldering Damage

The pins on plastic-film resistive sensors are embedded in thin plastic. Heat from a soldering iron can easily damage this plastic. Heat-sink alligator clips should be used while soldering, or the pins can be socketed instead of soldered.

Bad Load Distribution

Film-based sensors will not provide accurate readings if a load is imposed unevenly or inconsistently, or extends outside the detection area. A *puck* consisting of a small, rigid disc may be interposed between the source of the force and the sensor, to distribute the load within the maximum area. A puck may also be referred to as a *shim*.

Similarly, the sensor must be mounted on a flat, smooth surface, and if this is not available, a rigid plate should be interposed.

Water Damage

Although film-based sensors are enclosed in plastic, they are not waterproof. Immersion may cause the layers to delaminate.

Temperature Sensitivity

Because electrical resistance tends to vary with temperature, readings from resistive force sensors will vary with temperature.

Ambient temperatures of 70 degrees Celsius and above may damage a film-based sensor.

Leads Too Long

Although film-based sensors are supplied with a variety of lead lengths enclosed in the laminated layers of flexible plastic, the leads may be too long for a particular application. If they are trimmed, wires cannot be attached with solder, as it will melt the plastic. Conductive epoxy should be used.

single touch sensor

13

This entry only describes *capacitive* touch sensors. A *conductive* sensor, which uses the fingertip to complete a circuit between two exposed contacts, is not very common, and is not included here.

The type of touch sensor described in this entry requires no physical pressure for activation. It should not be confused with a resistive or piezoelectric **force** sensor that requires pressure. See Chapter 12.

An integrated circuit chip that processes a signal from a *touch pad* is often described as a *touch sensor*, even though it does not contain a sensing element. This entry describes it as a "touch sensor chip" to eliminate ambiguity, and refers to touch input elements as "touch pads."

Touch pads that contain *tactile switches* or *membrane switches* are described in the entry discussing **switches** in Volume 1. All types of switches are described in that volume, with the exception of a *reed switch*, which is magnetically activated and is therefore categorized as a sensor.

Capacitive touch sensors are sometimes referred to as *capacitive proximity sensors*, because they sense the proximity of a human fingertip. In this Encyclopedia, and in most other sources, a **proximity** sensor measures distance, not touch. See Chapter 5.

A *capacitive displacement sensor* employs the same principle as a capacitive touch sensor, but is used to detect the position of an object, not for human input.

OTHER RELATED COMPONENTS

- **force** sensor (see Chapter 12)
- **touch screen** (see Chapter 14)

What It Does

A *touch pad* detects the presence of a human fingertip (or other part of the body) and signals an integrated circuit chip, which is very often termed a *touch sensor*, even though it does not contain a sensing element itself. The chip creates an output to signify that human touch has been recognized.

A keypad of the type found on microwave ovens may appear similar to an array of touch pads, but is more likely to contain *membrane switches* or *tactile switches*, which are described with other forms of switches in Volume 1. The type of touch pad described in this entry requires no physical force and contains no parts that move or flex when pressed.

A modern **touch screen** is usually a capacitive device and can be thought of as an array of touch pads. See Chapter 14.

Applications

Capacitive touch sensors have become common as their cost has fallen relative to simpler components that respond to being pressed.

A touch sensor can be used to start or stop a process, or to power up or power down a device. Multiple sensors may be found wherever user input of a few alphanumeric characters is required. Because touch pads can be completely sealed, they are useful where hygiene is important.

Specific applications include the activation of a backlight in a handheld device, wake-up from standby, ear detection in a cellular telephone, control of medical devices, and activation of interior lighting in some automobiles.

The absence of moving parts or electrical contacts means that a touch pad is more reliable than any type of electromechanical switch. A disadvantage is that it provides no tactile feedback, and therefore will require a visual or audible confirmation when it responds to input. Lack of tactile feedback makes touch pads unsuitable for computer keyboards and other key-entry devices where rapid typing is required.

When the capacitive elements of a touch pad are transparent, it can be mounted in front of a screen.

Schematic Symbols

Either of the schematic symbols in Figure 13-1 may sometimes be used to represent a touch sensor, but not on a consistent basis.

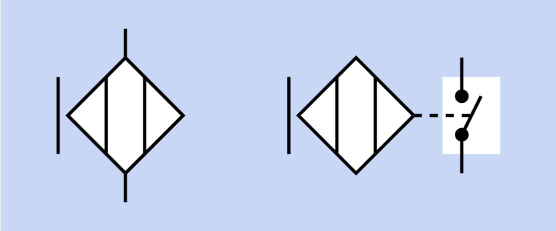

Figure 13-1 *Two possible schematic symbols that may represent a touch sensor.*

How It Works

The two plates in a capacitor are separated by an insulator known as the *dielectric*. Paper, plastic, glass, air, and other insulators can serve this purpose. Although no electrical connection exists between the plates, AC passes through the dielectric as a field effect.

Electrical capacitance exists between any pair of electrical conductors. The human body has high electrical resistance but still is electrically conductive, and therefore has capacitance with other conductive objects.

A touch pad can function as one side of a capacitor, with a fingertip functioning as the other side. In this mode, AC can pass from the touch pad and through the human body to ground. The current is very small, but fluctuations can be detected by an appropriately designed integrated circuit chip, or by a microcontroller.

The precise characteristics of a dielectric will affect the performance of a capacitor to some extent, but will not prevent it from working. Therefore a touch pad can function even when it is shielded behind a protective layer of glass or plastic, as is often the case.

A touch sensor chip generates pulses of low-voltage AC and sends them to a touch pad. The chip detects any variation in the current through the touch pad, indicating the presence of a fingertip. Where an input occurs, the chip changes its output, which usually requires a microcontroller for processing.

How to Use It

Touch sensor chips are available in as many as 40 different formats and configurations. All of them are surface-mount. For breadboarding, an experimenter can use breakout boards that have sensor chips installed. Figure 13-2 shows a product from Adafruit that is capable of addressing 12 touch pads. Its output is accessible from a microcontroller via the I2C protocol. For additional details about protocols such as I2C, see Appendix A.

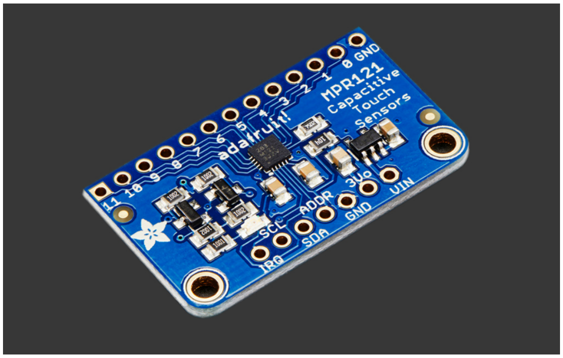

Figure 13-2 *A capacitive touch sensor chip on a breakout board from Adafruit.*

Similar breakout boards are available from Sparkfun, and from large online vendors such as *http://www.mouser.com* (where they are categorized as *development tools*).

While most touch sensor chips require a microcontroller, a few are available with the same number of output pins as touch-pad input pins, and each output pin will transition between logic-high and logic-low when an input on the corresponding pin is detected. Another breakout board from Adafruit, the AT42QT1070, uses this simple system.

A library exists for Arduino that enables two pins to sense touch on a piece of aluminum foil.

It can also work with conductive ink or paint.

Obtaining Touch Pads

Sensor chips are widely available as components, and are very inexpensive. On the other hand, touch pads are not common as components, probably because a touch pad is usually created as a pattern of copper traces etched onto a circuit board by a device manufacturer.

Touch pads from hobby-electronics sources usually include touch sensor chips. Sparkfun offers a 12-key keypad on this basis, and also a 9-key keypad designed as a “touch shield” for use with the Arduino. Both of the Sparkfun products are shown in Figure 13-3. They include the same MPR121 touch sensor chip as the breakout board from Adafruit shown in Figure 13-2, and require an I2C connection with a microcontroller.

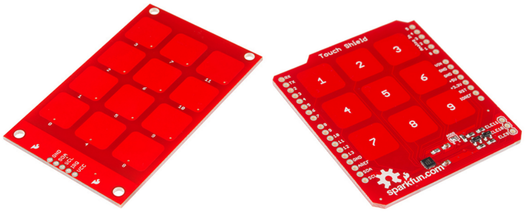

Figure 13-3 *Two capacitive keypads from Sparkfun, the one on the right designed as an Arduino shield.*

Because a capacitive touch pad is usually mounted inside an enclosure, the appearance of the bare touch pad as a component is unimportant. The outside of the enclosure can be printed with a design showing key outlines.

Individual Touch Pad

Adafruit sells the AT42QT1010 touch pad that emulates a momentary switch. Its output transitions from logic-low to logic-high when a finger presence is detected, and transitions back to logic-low when the finger is removed.

An alternative, shown in Figure 13-4, toggles between a logic-high and logic-low output and latches in each state with each single key press.

Both of these keypads contain sensor chips to generate the output.

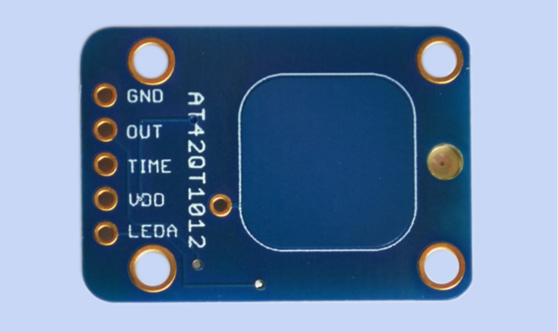

Figure 13-4 *The output from the AT42QT1012 sensor chip on this touch pad from Adafruit toggles between logic-high and logic-low each time the pad is touched.*

Wheels and Strips

A touch wheel uses a circular pattern of conductive traces, often referred to as *electrodes*, to receive finger input. A simple configuration is shown in Figure 13-5. The traces interlock without making contact with each other, so that moving the finger in a circular motion creates a capacitive input that rises and falls sequentially on each element. In the figure, three sections are used, each colored differently here for purposes of clarity. Other touch wheels may contain more sections.

Firmware that is designed to interact with wheel-shaped touch pads will generally assume that two electrodes are receiving input at one time. The firmware attempts to calculate the position and motion of the finger by assessing the relative capacitance of adjacent segments. Ideally the capacitance values should vary linearly and complementarily; that is, from a 50-50 value at a midpoint between two segments, the progression should change to 60-40, 70-30, and so on, as the finger moves around the wheel.

A *touch potentiometer* consists of multiple touch pads, usually arrayed in a straight line. This may be described as a *touch strip*. An example is made by GHI Electronics and sold by Robot Shop as their L12 Capacitive Touch Module, shown in Figure 13-6.

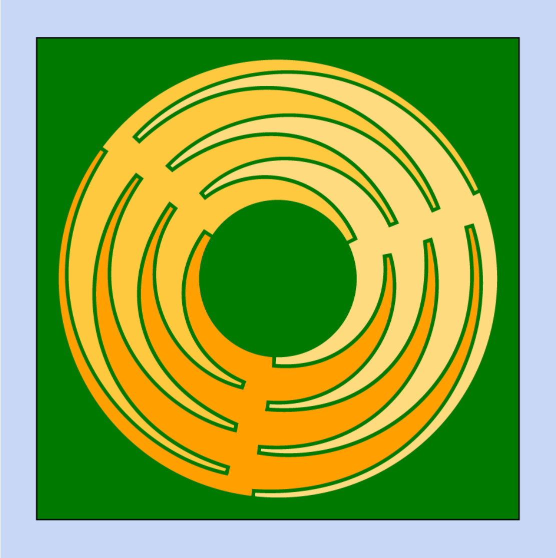

Figure 13-5 *A touch wheel created by copper traces on a circuit board (green). Each section is colored differently, for purposes of clarity.*

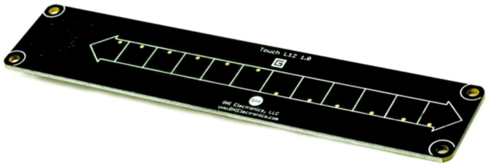

Figure 13-6 *Multiple capacitive touch pads arrayed as a strip, made by GHI Electronics.*

Design Considerations

A simple touch pad is often surrounded by additional copper that is grounded and may be described as a *shield* or a *guard*. The capacitance of a fingertip (spaced above the electrode by a layer of plastic or glass that functions as the dielectric) interferes with the field between the electrode and the shield.

The underside of the circuit board is often plated and grounded to protect the touch pad from electromagnetic interference. This, too, can be referred to as a shield or a guard. The ground plating can be in a hatched pattern to reduce its capacitance with the electrode above.

Layout of a circuit involving a single touch sensor chip can affect its performance significantly. A touch sensing device may have to be "tuned" to detect finger presence reliably.

A long trace between the touch sensor and an electrode will tend to pick up noise and will increase capacitance.

The distance between adjacent traces from multiple touch pads must be maximized to reduce capacitance between them. If output from a sensor chip uses the I2C or SPI digital protocol, any trace carrying that digital signal should be at least 4mm from input traces. If they cross, they should be at 90 degrees to each other.

Electrodes should not be shaped to resemble numbers or other characters printed above them. A single basic electrode should be circular.

What Can Go Wrong

Insensitive to Gloves

Gloves are a challenge for touch-sensor design, as they alter the dielectric and the distance between the electrode and the finger. Capacitive touch sensors may not work at all with some types of gloves. However, special gloves containing metallic threads are available.

Stylus Issues

A nonconductive stylus cannot activate a touch pad.

Conductive Ink

Ink that prints the shapes of touch pads on the exterior of a device should be nonconductive.

touch screen

14

The term **touch screen** is written as two words in this Encyclopedia. In many sources, the words are concatenated as "touchscreen."

The two words are hyphenated here when they function as an adjective, but not otherwise. In manufacturers' datasheets where the two words are used, they are not usually hyphenated.

OTHER RELATED COMPONENTS

- **single touch** sensor (see Chapter 13)
- **force** sensor (see Chapter 12)

What It Does

A touch screen is a video display with embedded touch sensing. The screen reports the position of the touch, and is used as a pointing device as an alternative to a mouse or trackpad. Some touch screens report pressure as well as position.

Touch screens are widely used in smartphones and tablets, and also in some laptop computers. Smaller, simpler touch screens may be found in office equipment such as photocopiers.

Schematic Symbol

No specific schematic symbol is used to represent a touch screen.

Variants

Early designs used infrared LEDs recessed into the edges of a frame around a screen. A matching photodiode picked up the focused beam from each LED. The presence of a fingertip was detected when it interrupted one or more of the beams. This system was not capable of high resolution, but was adequate for detecting user input at predefined locations.

Most touch screens currently are either resistive or capacitive.

Resistive Sensing

A resistive touch screen consists of two transparent layers that can be installed over a separate video display.

Each of the layers has uniform electrical resistance. Pressure from a fingertip on the outer layer (which we can refer to as layer 1) forces it to make contact at a point with the inner layer (referred to here as layer 2).

Two vertical electrodes connect with layer 1 along its left and right edges. Two horizontal electrodes connect with layer 2 along its top and bottom edges. When voltage is applied between the vertical electrodes on layer 1, the layer acts as a horizontal voltage divider. The voltage at the point where the layer is being pressed is applied to layer 2, and can be read from either of the electrodes on layer 2, so long as the metering has a much higher impedance than that of the layer. The voltage is decoded as a value for the horizontal position on layer 1.

An external switching device now repeats the procedure, except that it applies voltage to layer 2, and reads it from layer 1 to supply a value for the vertical position on layer 2. The sequence is illustrated by the top and bottom sections of Figure 14-1.

Because only four connections are necessary, this is referred to as a *four-wire* resistive touch screen. Five-wire variants exist, but are less common, and are not included in this entry.

Advantages of a resistive screen include:

- Simplicity. Only four connections are necessary, and the layers of the screen do not have to be subdivided into separate conductors.
- Low cost, relative to capacitive touch screens.
- Will respond equally well if a user wears gloves or uses a stylus.

Disadvantages of a resistive screen include:

- Some resistive versions require a stylus input instead of finger pressure.
- Resistive screens only respond to a one-location input. Two-finger gestures are not supported.
- Contact bounce occurs when the flexible layer is pressed against the underlying layer, and voltage spikes may be associated with switching power to the screen. To address this issue, firmware in a microcontroller may have to take a median value from several rapid readings.
- The flexible membrane is vulnerable to damage from sharp objects.

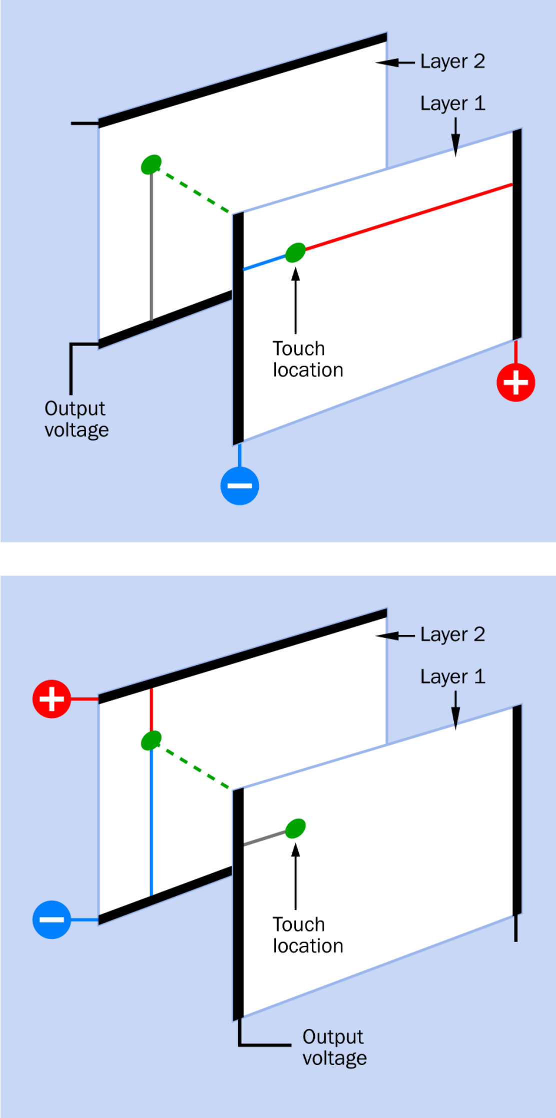

Figure 14-1 *Sections of a resistive touch screen are shown displaced for clarity. In reality they would be separated by a very small gap, allowing one section to make contact with the other in response to finger pressure.*

Capacitive Sensing

A capacitive touch screen can consist of an array of **single touch** sensors printed onto a glass panel as vertical and horizontal lines of transparent conductive ink.

Alternatively, a small capacitive screen can measure the tiny amount of current drained by a fingertip from four sources located at corners of the screen.

For more information about capacitive touch sensing, see the entry on touch sensors in Chapter 13.

Screens Available as Components

A wide variety of screens can be found in diagonal sizes ranging from 2 inches upward. Onboard electronics can be suitable for connection with a microcontroller using the I2C and SPI protocols, or USB. Different screen resolutions are available.

An example of a 3.5-inch touch screen with 320x200 resolution, mounted on a breakout board that can be used with a breadboard and an Arduino, is shown in Figure 14-2.

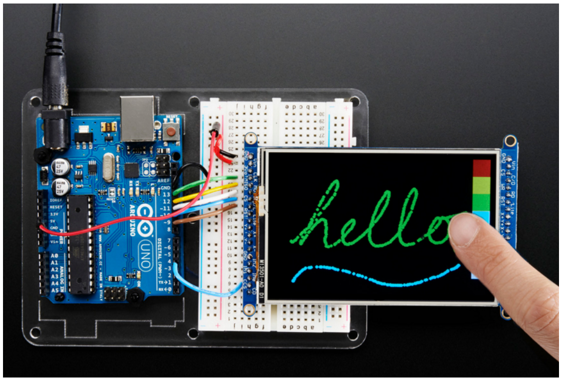

Figure 14-2 *An Arduino-compatible touch screen mounted on a breakout board, available from Adafruit.*

A 7-inch resistive touch screen that can be mounted on a separate video display is shown in Figure 14-3. It can be used with the STMPE610 controller chip, which converts resistive screen values into digital coordinates and can be accessed by a microcontroller over both SPI and I2C. This surface-mount chip is available on a breakout board.

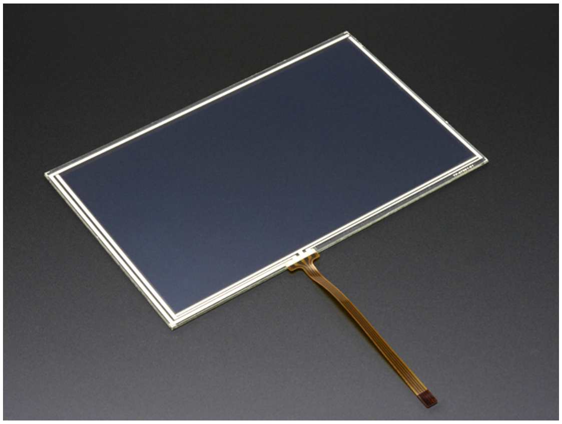

Figure 14-3 *This resistive touch screen is intended for use as a layer applied to a 7-inch video display. The screen is available from Adafruit.*

When choosing a touch screen as a component for a DIY project, the availability of microcontroller code libraries for reading and refreshing the display is an important consideration.

liquid level sensor

15

Level indicators that contain no electronic components are not included in this entry.

Specialized industrial level-sensing equipment is generally outside the scope of this Encyclopedia. This entry discusses small-scale, lower-cost sensors.

Liquid volume can be assessed by measuring liquid pressure at the bottom of a reservoir. Sensors for this purpose are discussed in the entry describing **gas/liquid pressure** sensors. See Chapter 17.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see Chapter 16)
- **gas/liquid pressure** sensor (see Chapter 17)

What It Does

Measuring the volume of liquid in a storage vessel or reservoir is such a fundamental task, countless methods have been devised, of which only the simplest and most common will be discussed here.

A liquid level sensor can have a binary output, meaning that it signals when the volume rises above or falls below a level that can be preset or reset. Often the sensor will be connected to a pump or valve that maintains a relatively constant volume in a container.

Alternatively a sensor can indicate the actual stored volume, either with an analog output or in digital increments.

Schematic Symbols

Three variants of a schematic symbol for a simple liquid level sensor are shown in Figure 15-1. They are not always used, however, and a sensor may be shown simply as an annotated switch.

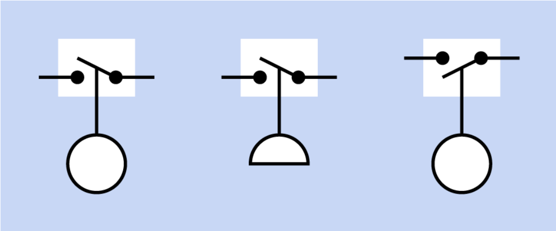

Figure 15-1 *Three variants of a schematic symbol to represent a liquid level sensor. The rightmost symbol indicates that a rising level closes, rather than opens, a switch.*

Applications

The fuel gauge in a vehicle is one of the most commonly encountered applications of a liquid level sensor. The water tank in a recreational vehicle or boat may use similar electronics. In industry, the choice of a sensor will be influenced by the type of liquid that is being stored, the desired accuracy, the temperature range, and whether the storage tank is sealed or open to atmospheric pressure.

How It Works

Desirable attributes of any liquid-level sensor include resistance to vibration, some damping to average out fluctuations caused by turbulence or sloshing in the liquid, resistance to chemical reactions with the liquid, and few moving parts that may require maintenance if the sensor is inside a sealed tank. Desirable attributes of an analog float sensor include a linear response, and some hysteresis if the application requires it.

This entry compares a variety of sensing strategies.

Binary-Output Float Sensor

The term "binary output" is used here to describe an output that only has two states (on and off, or logic-high and logic-low). The simplest type of liquid-level sensor with a binary output consists of a donut-shaped float that contains a permanent magnet and is free to slide vertically up and down a sealed tube containing a reed switch. The tube is supported on a bracket that can be mounted on the wall or the lid of the vessel containing the liquid.

The tube and float must be nonmagnetic, and the float must have a significantly lower specific gravity than the liquid that is being used ("significantly" because the float requires sufficient buoyancy to carry the weight of the magnet and overcome any friction between itself and the tube). A diagram illustrating this configuration is shown in Figure 15-2.

To change the level setting of the sensor, the bracket may be mounted on a screw thread to adjust its vertical position.

The reed switch can be normally open or normally closed, as needed to respond to a rising or falling liquid level. For basic information about reed switches, see "Reed Switch". Additional, detailed information about reed switches is included in the book *Make: More Electronics*.

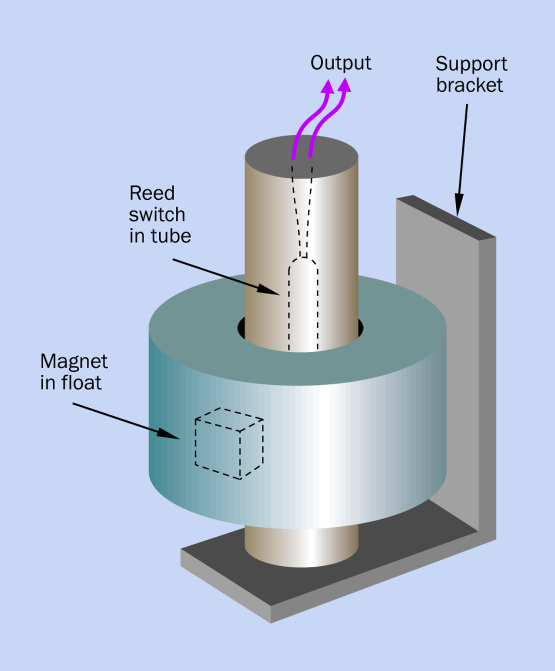

Figure 15-2 *The principal parts of a basic binary-output float sensor.*

For increased reliability, a Hall-effect sensor could be substituted for a reed switch. See "Hall-Effect Sensor" for general information about Hall-effect sensors.

Another binary-output float sensor is shown in Figure 15-3. This is a sealed plastic capsule containing a snap-action switch and a steel ball. The cable is attached to the underside of the top of a tank, and the capsule dangles into liquid in the tank. A separate weight (not shown) has a hole in the middle, and is threaded over the wire. The weight keeps the wire approximately in a vertical position as it dangles into the tank.

Figure 15-3 *An air-filled float that switches external power depending on its orientation.*

Figure 15-4 shows the components inside the float. When the liquid level in the tank falls, the float adopts the position shown on the left in the figure. The ball drops against a lever that closes a snap-action switch, which starts an external pump to replenish the tank. As the liquid level rises, the buoyancy of the air-filled float changes its orientation to that shown on the right in Figure 15-4. The ball drops and the switch opens, stopping the pump.

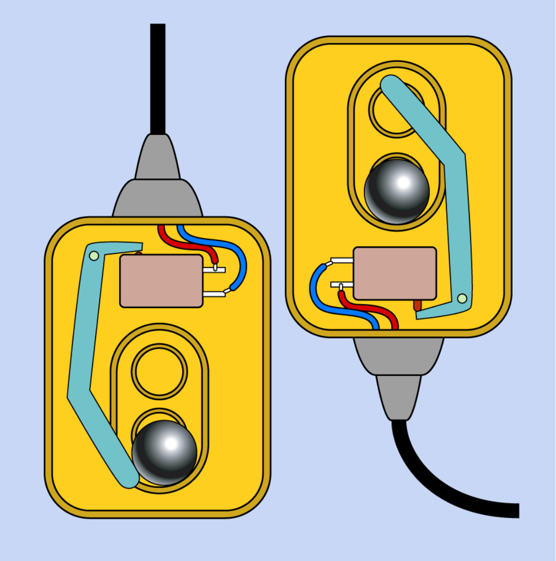

Figure 15-4 *Internal components of the float shown in the previous figure.*

Two circular indents on the inner surface of the plastic capsule prevent the ball from rolling erratically if there is turbulence in the liquid. They also provide some hysteresis.

Analog-Output Float Sensor

The simplest type of liquid-level sensor with an analog output consists of a float on an arm attached to a potentiometer, as shown in Figure 15-5. This very basic design was used in fuel tanks in vehicles for many decades. Disadvantages include a nonlinear response and the limited life expectancy characteristic of potentiometers. Some compensation for the nonlinear response can be made by using an analog fuel gauge with a nonlinear scale.

For more information about potentiometers, see "Arc-Segment Rotary Potentiometer".

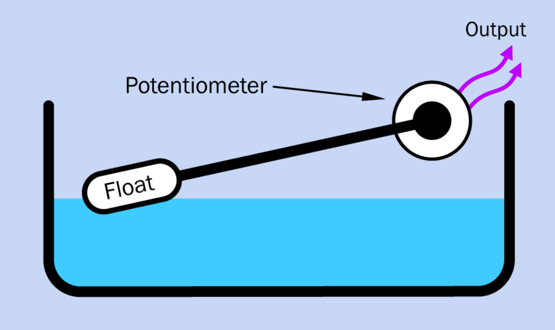

Figure 15-5 *A basic float sensor with an analog output.*

Incremental-Output Float Sensor

A schematic diagram for a simple float sensor with incremental output is shown in Figure 15-6. A magnet embedded in a donut-shaped float, similar to that shown in Figure 15-2, interacts with a sequence of reed switches installed in the central tube. The switches are spaced at equal intervals and apply power between resistors of equal value wired in series. This system has been used in motorcycle and automobile fuel tanks, where the switches may be enclosed in a (nonmagnetic) stainless-steel tube. The accuracy is limited by the number of reed switches.

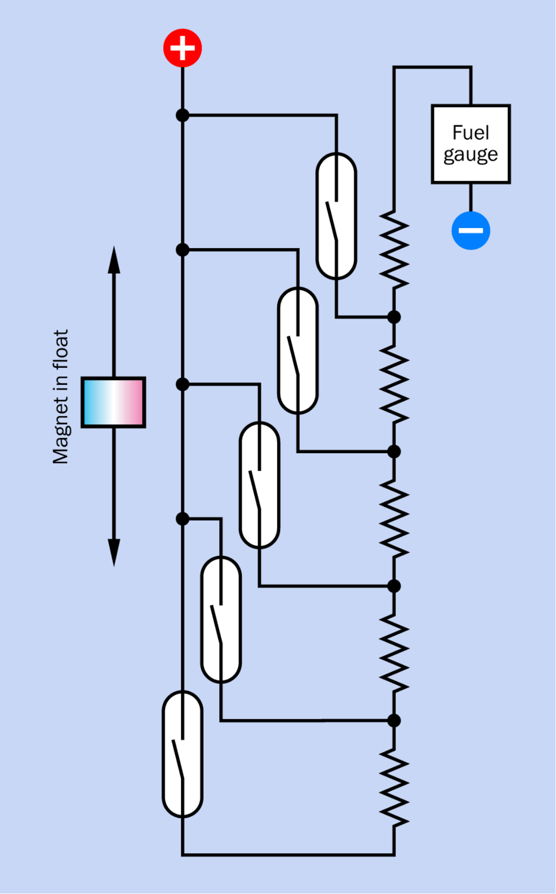

Figure 15-6 *A float sensor with incremental output.*

Displacement Level Sensors

If a heavy object, described as the *displacer*, is suspended in liquid, the effective weight of the object diminishes as the liquid rises up around it. This occurs because according to Archimedes' Principle, the upward buoyant force is equal to the weight of liquid that the object displaces. The displacer is suspended from a load sensor that measures its weight. Analog output from the sensor will be approximately linear with liquid level.

For a simplified diagram showing a displacement level sensor using this concept, see Figure 15-7.

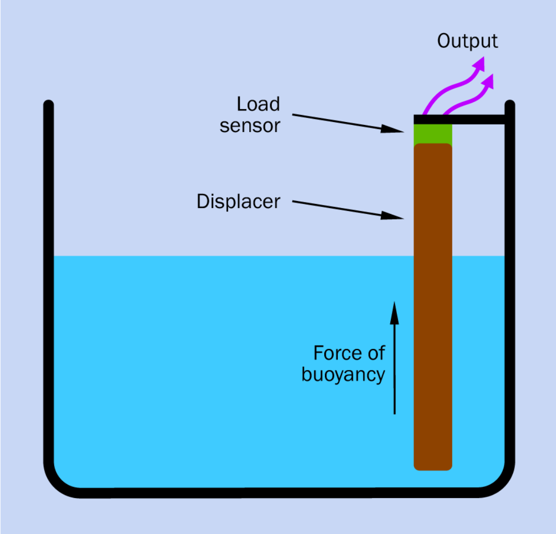

Figure 15-7 *A displacement sensor. The displacer is heavier than the liquid around it, but its effective weight diminishes as the liquid rises.*

For more information on load sensors, see Chapter 12.

Ultrasonic Level Sensors

An ultrasonic sensor can be used to measure the level of liquid in a reservoir, as shown in Figure 15-8. For more information about this type of sensor, see the entry discussing proximity sensors in Chapter 5. A disadvantage of using ultrasound for liquid level sensing is that the speed of sound will be affected by any vapor given off by a volatile liquid.

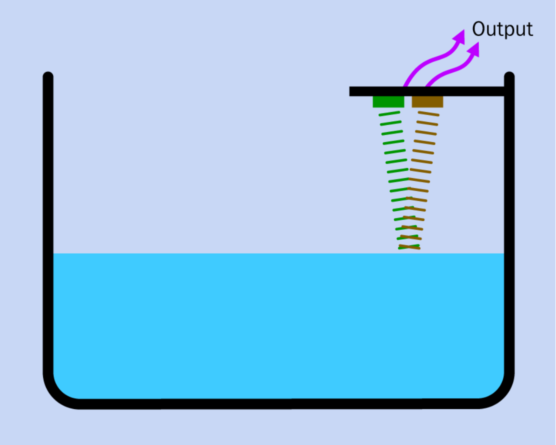

Figure 15-8 *An ultrasonic proximity sensor can measure the level of liquid in a reservoir.*

Reservoir Weight

The weight of a reservoir can be measured to assess the volume of liquid in it. This can be done by mounting the reservoir on load sensors. However, pipes leading to and from the reservoir must be designed so that they do not add or subtract any significant weight. Figure 15-9 suggests an arrangement, although the outflow will still change the weight to some extent depending on the amount of suction that is applied. For more information on load sensors, see Chapter 12.

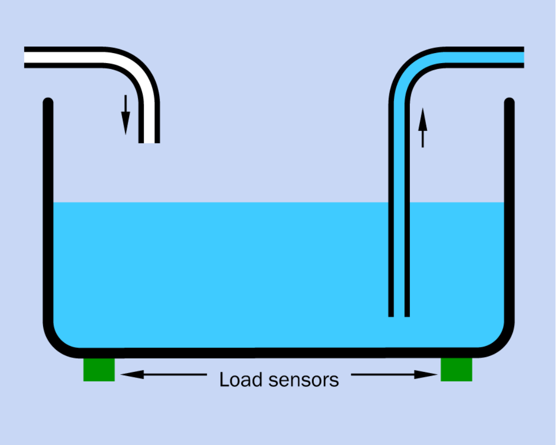

Figure 15-9 *Using load sensors to assess the weight of liquid in a reservoir requires that the weight of plumbing should not be imposed on the structure of the reservoir.*

Pressure Sensing

A differential pressure sensor can be added to a pipe near the bottom of a reservoir. The sensor measures the difference between the liquid pressure and ambient air pressure. See Figure 15-10.

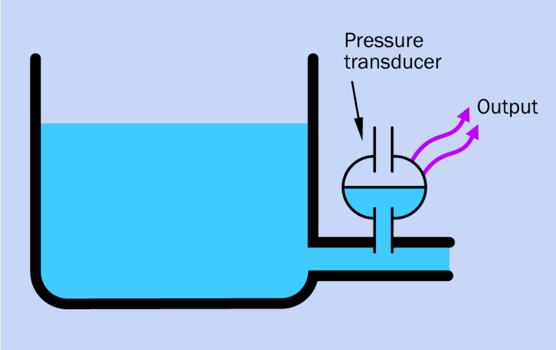

Figure 15-10 *A pressure sensor can assess the volume of liquid in a reservoir.*

This arrangement assumes that the reservoir is vented so that air pressure above the liquid level is equal to air pressure at the sensor. If the reservoir is not vented, a pipe must connect the reference port on the sensor with the space above the liquid.

The reservoir must have straight, vertical sides for the pressure to be directly proportional to the liquid volume.

Liquid volume can also be assessed by measuring the pressure inside a container, near the bottom. A submersible pressure sensor can be used, typically consisting of a watertight capsule fitted with a diaphragm that connects with an internal strain gauge. The sensor is lowered on a cable that also contains an air line. Because pressure in a liquid is affected by atmospheric pressure above the surface of the liquid, the sensor requires an air line so that its measurements are relative to the outside air.

Submersible pressure sensors are useful where access is limited—for example, when measuring fluctuations in an open-air municipal water reservoir.

What Can Go Wrong

Turbulence

Surface turbulence consisting of ripples, waves, or sloshing of liquid will tend to occur when a reservoir is refilled rapidly or is subjected to lateral movements as a result of being mounted in a moving vehicle. To minimize output fluctuations, some *damping* is desirable.

Baffles consisting of perforated plates inside the reservoir are a common strategy, as shown in Figure 15-11. In the upper section of the figure, lateral acceleration causes submersion of a float sensor in the reservoir. In the lower section, perforated baffles minimize the problem.

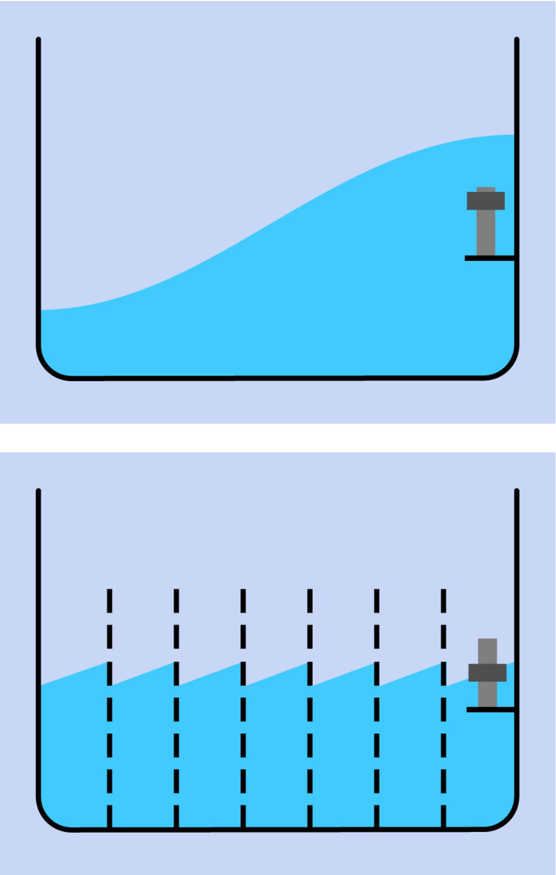

Figure 15-11 *Insertion of perforated baffles in a reservoir minimizes the sloshing that otherwise tends to occur when the reservoir is subject to lateral motion.*

Sensors that measure the weight or pressure of a liquid are less susceptible to turbulence. In a displacement sensor, the weight of the displacer provides a damping effect.

Tilting

All level sensors will tend to give inaccurate readings when a reservoir is tilted. A float sensor will be affected less if it is mounted centrally in a reservoir, because the reservoir will tilt around the sensor, as shown in Figure 15-12.

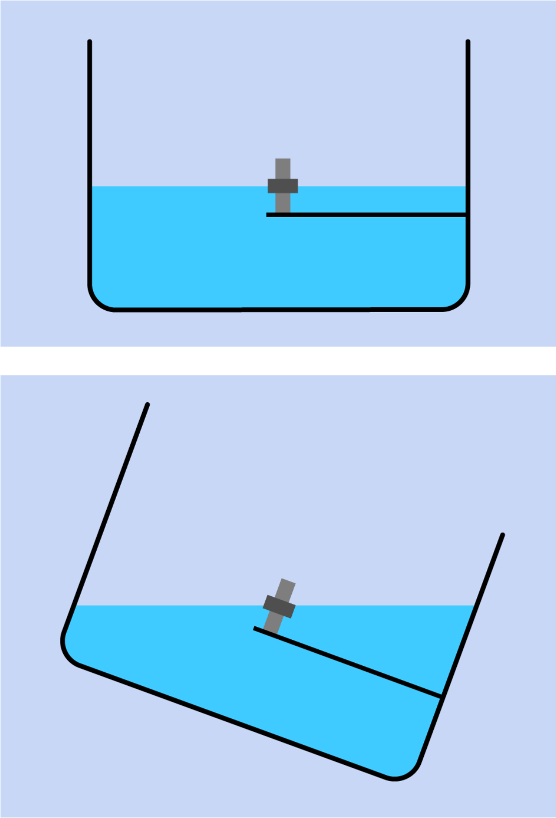

Figure 15-12 *If a float sensor is mounted centrally in a reservoir, it will be significantly less affected if the reservoir tilts.*

liquid flow rate sensor

16

Flow rate sensors that contain no electronic components are not included in this entry.

Some methods of liquid flow sensing can also be applied to gases, but sensors are usually designed for one application or the other. Therefore, gas flow sensors have their own entry. See Chapter 19.

Many liquid flow rate sensors are large devices designed for industrial applications. This entry focuses on lower-cost solid-state sensors.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see Chapter 15)
- **gas/liquid pressure** sensor (see Chapter 17)
- **gas flow rate** sensor (see Chapter 19)

What It Does

A liquid flow rate sensor measures the rate at which liquid flows past or through the device. A water meter is an example of a flow rate sensor.

A sensor may have a binary output, meaning that it signals when flow stops or starts, or if its rate rises above or falls below a level that can be preset or reset. However, most flow meters have an analog output that varies with the volume per unit of time.

Measuring the flow rate of a liquid can be challenging if the viscosity is very high, the liquid is chemically reactive, or the rate is very low. Such factors may require specialized equipment that is outside the scope of this Encyclopedia.

This entry compares the most popular sensing strategies.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic schematics, and therefore they are not included here.

Paddlewheel Liquid Flow Rate Sensors

The simplest and most common liquid flow rate sensor uses a *paddlewheel*, also referred to as a *rotor*, that is mounted with its axis of rotation at 90 degrees to the direction of liquid flow. An example is the Koolance INS-FM16 shown in Figure 16-1. This sensor is intended for use in an aftermarket cooling system for the CPU in an overclocked desktop computer, but can be used in any system where the rate of flow ranges from 0.5 to 15 liters per minute. The paddlewheel has a pair of magnets mounted in it, activating a *reed switch* that is mounted in a sealed enclosure beneath the wheel. (For more information regarding reed switches, see "Reed Switch" in the entry describing **object presence** sensors.)

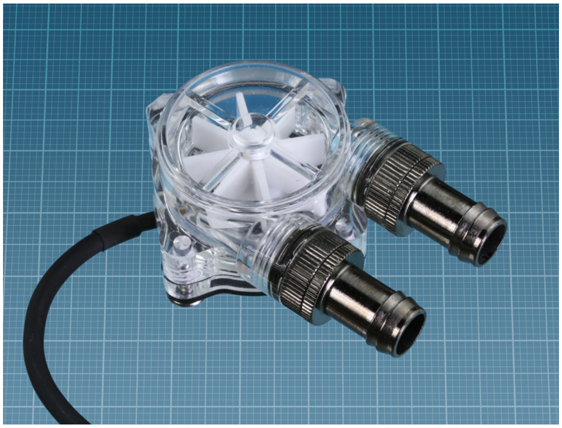

Figure 16-1 *A low-cost, simple paddlewheel sensor designed for rates of 0.5 to 15 liters per minute. The background grid is in millimeters.*

While the reed switch in the Koolance flow sensor inevitably suffers from contact bounce, it has the advantage of simplicity and can be used in conjunction with appropriate hardware or microcontroller code to debounce the pulse stream.

As in any device with a rotating part, friction and wear will afflict a paddlewheel sensor, especially because the bearings are often in the chamber through which the liquid passes. This eliminates the possibility for roller bearings. Typically a plain bearing is used, consisting of a pin that engages in a hole in the casing. Friction wears the bearing surface, creating a larger gap, which allows the rotor to vibrate or bounce instead of spinning smoothly. Flow resistance increases and accuracy is degraded.

In modern designs, the mass of the rotor is minimized to reduce the friction. Also, if the shaft is horizontal (but still at 90 degrees to the direction of the flow), buoyancy of the rotor in the liquid can reduce friction still further, as suggested in Figure 16-2. Ideally, the density of the rotor and the density of the liquid will be the same.

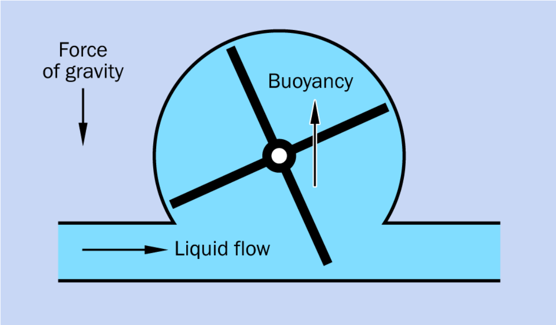

Figure 16-2 *In this configuration, friction on the rotor bearings is mitigated by taking advantage of the buoyancy of a low-density rotor in the liquid that passes through.*

The U-shaped flow path in the Koolance sensor in Figure 16-1 maximizes the responsiveness of the rotor, but an inline path is more common. An example is shown in Figure 16-3.

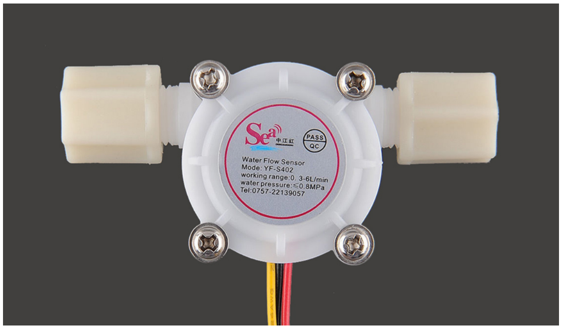

Figure 16-3 *An inline flow sensor rated for 3 to 6 liters per minute.*

Turbine Flow Rate Sensors

In a turbine-type sensor, two or more spiral blades are attached to a hub that rotates around an axis in line with the liquid flow, as shown in Figure 16-4. A magnet in each spiral blade triggers a reed switch or Hall-effect sensor mounted in a bracket that suspends the turbine from the interior walls of the tube.

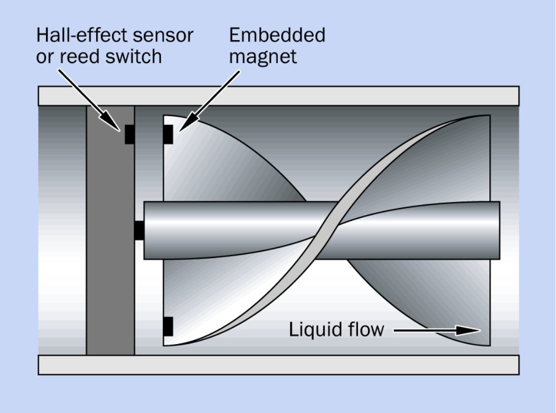

Figure 16-4 *Simplified view of a turbine flow rate sensor mounted inside a tube.*

The support bracket often consists of four struts, which impose resistance to liquid flow. The bearings suffer from the same kinds of problems as the bearings in a paddlewheel sensor, and must withstand additional load resulting from the inline force exerted by flow. Overall, while the turbine type of sensor is popular in laboratory equipment, it has disadvantages that are not shared by the bulkier paddlewheel type.

Limitations of Paddlewheels and Turbines

Both the paddlewheel and turbine types of sensors require a minimum flow to overcome the friction in their bearings. Below this minimum, liquid will find its way around the rotor without turning it. Even when the flow exceeds the minimum, response of the rotor is likely to be nonlinear as a result of turbulence and other factors.

Above a limit stated by the manufacturer, turbulence increases to the point where output from the sensor is no longer meaningful. Wear on the bearing will also increase with flow rate.

A significant problem for these types of sensors is that they do not respond well to sudden variations in flow. The paddlewheel, in particular, has inertia as a function of the diameter of the rotor, and will take some time to spin up in response to an increase in flow. Conversely, when the flow diminishes, the paddlewheel will tend to overrun.

The viscosity of the liquid passing through a paddlewheel or turbine sensor will have a very significant effect on its performance.

Thermal Mass Liquid Flow Rate Sensor

The thermal-mass system is commonly used when volumes are extremely low. The system is illustrated in Figure 16-5. A tube containing liquid is fabricated from a heat-conductive metal such as aluminum. It is enclosed in a larger tube, and the gap between them is filled with thermal insulation. A temperature sensor such as a thermistor measures the temperature of liquid entering the system. A second sensor, combined with a small resistive heater in the form of a coil around the tube, is placed downstream. Liquid passing through the tube will tend to remove heat more effectively at higher flow rates, and the difference in temperature between the two sensors is a logarithmic function of the flow rate.

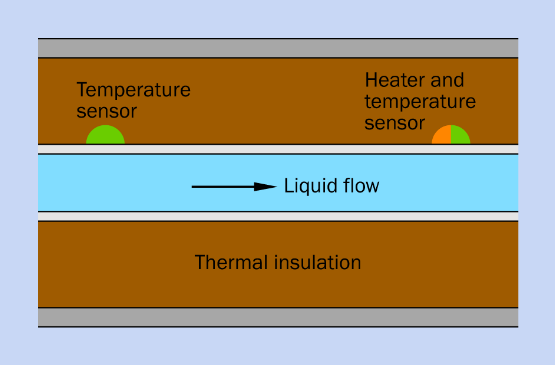

Figure 16-5 *In this type of low-flow sensor, the temperature differential between the two sensors is a logarithmic function of the flow rate.*

Variants of this system use slightly different tube configurations and sensor placement, but the principle is the same. Its advantages include the lack of any moving parts, and the avoidance of any probes intruding into the liquid, which is desirable in biochemical and medical applications.

The same principle is applied in many **gas flow rate** sensors. See "Mass Flow Rate Sensing".

Sliding Sleeve Liquid Flow Switch

This sensor is used in some domestic systems, where flow-activated water heating is required. A vertical section of brass (nonmagnetic) water pipe contains a sliding inner sleeve that incorporates a magnet. An external reed switch is activated when the sleeve is moved by water flowing through it. When the flow stops, the sleeve is returned to its rest position by the force of gravity.

Sliding Plunger Liquid Flow Switch

Figure 16-6 shows an exploded view of a similar device, using a plastic plunger that slides inside a nylon plumbing fixture designed for 3/4-inch pipe. The plunger contains a magnet and is restrained by a compression spring and a circular perforated plate. When water flow is sufficient to overcome the resistance of the spring, the plunger slides far enough for the magnet to activate a reed switch sealed into the external housing.

Ultrasonic Liquid Flow Rate Sensor

This type of sensor passes ultrasound through a liquid in a pipe. The speed of sound through the liquid is affected by the flow rate, and external electronics translate this lag time into a value for volume-per-minute. The system adjusts for variations in temperature that also affect the speed of sound.

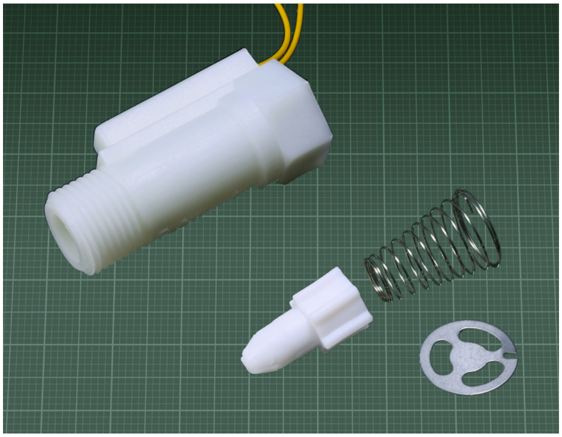

Figure 16-6 *Parts of a flow switch. The small plunger is inserted into the pipe and retained with the perforated plate and compression spring.*

Various configurations are available, some allowing ultrasound sources and detectors to be clamped to the outside of a pipe, as shown in Figure 16-7. To eliminate other variables, one ultrasound pulse is transmitted in the same direction as the flow, followed by another pulse contrary to the flow, and the difference between the two transmission times is used as an indicator of the flow rate.

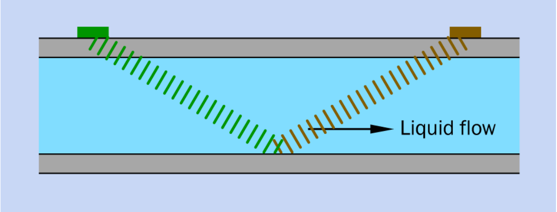

Figure 16-7 *Some ultrasound flow sensors are designed to be clamped externally to a pipe.*

Magnetic Liquid Flow Sensor

A magnetic field is induced in a metal pipe by a coil generating its field perpendicular to flow. The inside of the pipe is lined with nonconductive material in which two electrodes are mounted. Because water containing ions is conductive, the flow of water through the magnetic field induces a small potential difference between the electrodes. This voltage can be used as an indication of the flow rate.

To eliminate external factors that would also affect the potential difference, the polarity of current through the coil around the pipe alternates rapidly. The induced field remains the same regardless of the direction of the current.

A magnetic flow sensor should not be confused with a magnetic flow *switch.* Various types of switches are made, including large, heavy-duty devices where flow moves a magnet that triggers a shutoff valve. This type of industrial device is outside the scope of the Encyclopedia.

Differential Pressure Liquid Flow Meter

In this system, a pipe contains a perforated plate or some similar constrictor that partially obstructs the flow of liquid. Pressure is measured by a pair of pressure transducers placed before and after the constrictor. The pressure difference is an indicator of flow, because it increases as the flow rate increases.

This system was developed originally for large industrial applications but has been miniaturized and etched into silicon to measure very small flow rates. The Omron D6F-PH is an example, measuring less than 3cm square. It contains digital correction to enable a close-to-linear output. Because of its small size, it can be used only for slow flow rates, or as a bypass sensor. The general concept of a bypass sensor is illustrated in Figure 19-7, in the entry discussing gas flow sensors.

What Can Go Wrong

Vulnerability to Dirt and Corrosive Materials

MEMS liquid flow rate sensors containing very delicate, very small sensing elements are very vulnerable to contamination with dirt. Liquids should be filtered to minimize this risk. A manufacturer's datasheet should provide information about the use of corrosive or chemically active liquids.

gas/liquid pressure sensor

17

Pressure measurement devices that do not contain any electronics, such as a nondigital tire-pressure gauge or a mercury manometer, are outside the scope of this Encyclopedia.

Many pressure sensing methods can be used with both gases and liquids. To avoid duplication, gas pressure sensors and liquid pressure sensors do not have separate entries. Both types are described here.

This entry deals almost exclusively with MEMS components. It does not include pressure measurement devices sold as industrial products.

Some manufacturers and vendors use the term *pressure sensor* to describe a component that measures mechanical load or force. In this Encyclopedia, the term only describes components that measure liquid or gas pressure. Mechanical load cells and force sensors will be found in the section on **force** sensors. See Chapter 12.

OTHER RELATED COMPONENTS

- **liquid level** sensor (see Chapter 15)
- **liquid flow rate** sensor (see Chapter 16)
- **gas flow rate** sensor (see Chapter 19)

What It Does

A pressure sensor measures the force exerted by a gas or liquid, often in a container or pipe.

Static pressure is measured under conditions that change slowly or not at all. *Dynamic pressure* is subject to fluctuations. Pressure sensors tend to be designed for one condition or the other.

Schematic Symbols

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, including pressure sensors. Typically they involve a single letter or an X in a circle. These symbols are not generally found in electronic schematics, and therefore they are not included here.

Applications

Barometric sensors are found in barometers and weather stations. Altimeters are really a specialized form of barometric sensor, used in airborne vehicles. Gas pressure sensors have many industrial applications, and are used to monitor tire inflation in vehicles and the output from air compressors. They may also measure liquid pressure indirectly, as in a blood-pressure cuff.

Liquid pressure sensors are widely used to measure oil pressure in automobile engines and hydraulic braking systems. They have medical applications, and monitor water pressure in municipal systems and pumped supplies.

Design Considerations

As a result of random molecular movement, gases will tend to disperse to fill a container. After a gas reaches equilibrium, the pressure will be almost equal in all directions, affected only slightly by gravity. Where a gas is in a sealed, rigid container, pressure will vary linearly with temperature.

A liquid will tend to accumulate at the bottom of any container under the force of gravity. Liquid pressure in a container will be highest at the bottom, because of the weight of liquid above it. However, because almost all liquids are not easily compressible, they transmit force from any point in a container to any other point, including the sides of a container.

Units

Pressure is measured as force per unit area, which can be expressed in a confusing variety of units.

In the United States, gas and liquid pressure are still often expressed in pounds per square inch, abbreviated as PSI, or more often as psi, or sometimes as lb/in^2.

In standard international (SI) units, 1 bar of pressure is approximately equal to atmospheric pressure at sea level. Millibars are popular among meteorologists, 1 millibar being 1/1000 of a bar, equivalent to 100 pascals, where 1 pascal = 1 newton per square meter.

A bar is equivalent to 14.504 psi.

Blood pressure is measured in millimeters of mercury, because mercury manometers were used for this purpose originally. Atmospheric pressure may also be measured in millimeters of mercury, because the earliest barometers used a tube containing mercury. In the United States, some sources still refer to inches of mercury.

How It Works

Pressure sensing usually entails three stages.

1. A *sensing element* translates pressure into the mechanical displacement of a flexible part.
2. A *transducer* converts mechanical displacement into an electrical effect, either modifying resistance or creating a small voltage or current.
3. Electronics are used for *signal conditioning*. This may entail modifying a nonlinear signal, or may convert an analog output to a digital output.

As pressure sensors are increasingly taking the form of MEMS devices, all three stages may be combined in one silicon chip.

Basic Sensing Elements

Figure 17-1 shows four types of sensing elements that are used, or have been used, to convert pressure into mechanical motion. In each case, a green arrow shows where gas or liquid is introduced under pressure. 1: A Bourdon tube flexes under pressure, increasing its radius. The tube is hollow, open at one end and sealed at the other. 2: A coiled Bourdon tube uncoils partially under pressure, causing the top end to rotate. 3: A simple flat diaphragm. 4: A ribbed diaphragm.

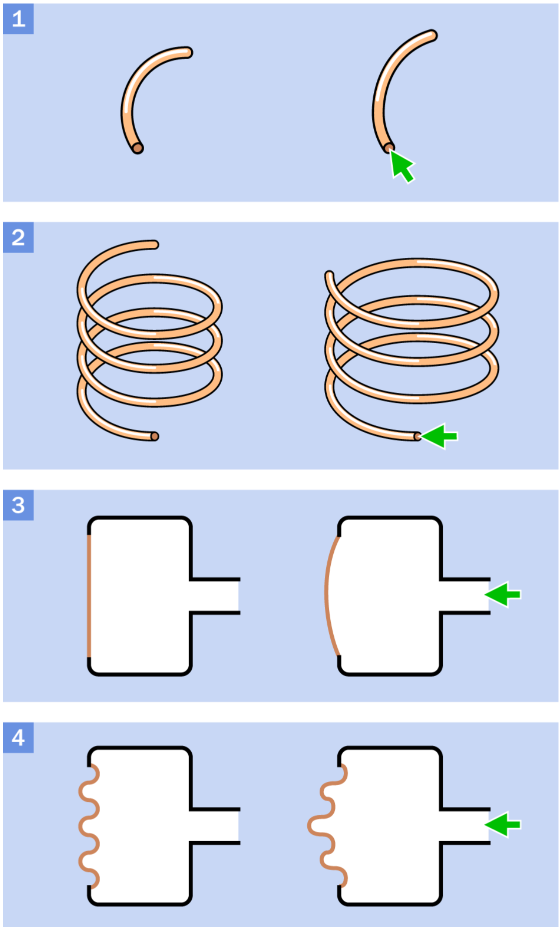

Figure 17-1 *Sensing elements. See text for details.*

The sensing elements numbered 1 and 2 are less likely to be used in modern systems, although a coiled Bourdon tube may be used to turn a potentiometer in low-cost oil-pressure sensing designs. Number 3 is well suited to MEMS devices, as it can be etched into silicon using the principle illustrated in Figure 17-2.

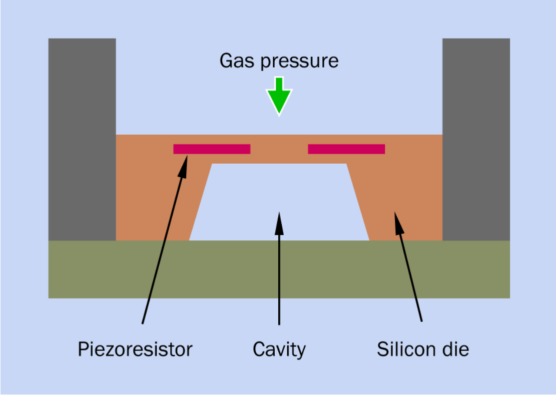

Figure 17-2 *A section of a flat-diaphragm pressure sensor, viewed from the side. An aperture in the top face of the chip allows air to enter. The sensor is etched into silicon, and the deflection of the wafer is measured using embedded piezoresistors.*

Relative Measurement

Pressure is a relative measurement. It is expressed relative to a *reference pressure* of some kind. Three types of measurement are commonly used:

1. *Absolute pressure*, relative to the zero value of a vacuum.
2. *Gauge pressure*, relative to ambient pressure (that is, pressure in the environment around the sensor). Air pressure is the reference source for gauge pressure, and a *vent* is incorporated into the sensing system.
3. *Differential pressure*. In this case, the pressure being measured is relative to some other pressure—for example, the pressure differential between two sealed tanks.

Figure 17-3 illustrates these three measurement types.

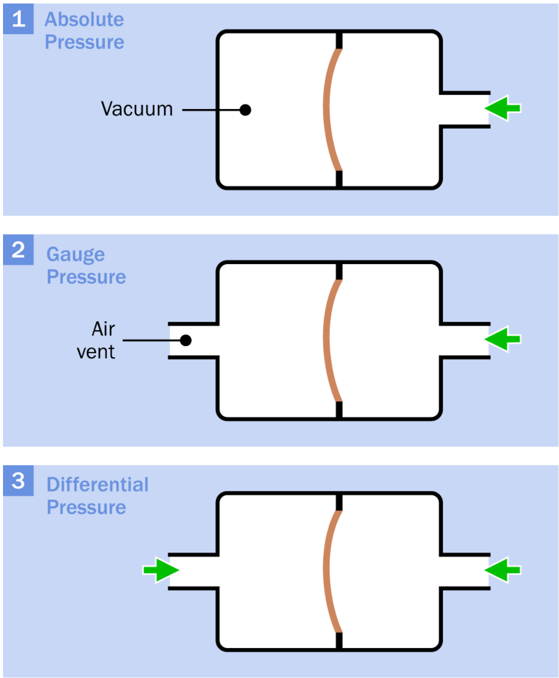

Figure 17-3 *Three ways in which pressure is measured.*

Variants

Ambient Air Pressure

A *barometric* sensor measures the pressure of air around it. This is an absolute value, relative to a vacuum.

The earliest barometers consisted of a tube sealed at one end, containing mercury. No air was allowed to enter the tube. It was inverted so that its open end was immersed in a small reservoir that was open to the air, the pressure of which supported the column of mercury in the tube. Thus, the height of the column was ratiometrically equivalent to atmospheric pressure. Because low air pressure is often an indicator of inclement weather, a barometer was a very simple tool for weather forecasting.

A modern barometric sensor consists of a chip with a vent hole in its upper surface, allowing ambient air pressure to reach a sensor inside the chip. A popular example has been the Bosch BMP085, shown soldered onto a breakout board in Figure 17-4. The BMP180 has a very similar specification, the primary difference being that it is designed for a SPI bus instead of I2C. For additional details about protocols such as SPI and I2C, see Appendix A.

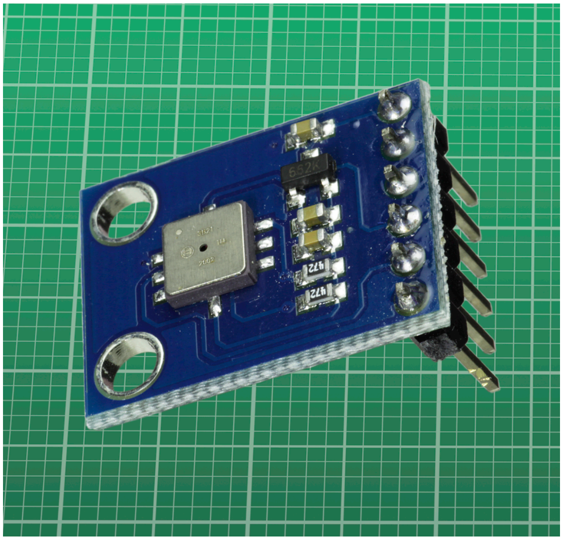

Figure 17-4 *A Bosch barometric sensor on a breakout board. The background grid is in millimeters.*

This sensor uses a supply voltage ranging from approximately 2VDC to 3.5VDC, but a breakout board includes a voltage regulator that tolerates a 5VDC supply. The output is digitized for access by a microcontroller, but the format is quite complex and consists of raw data that must be converted to air pressure by applying a formula. The manufacturer offers free code in the C language for this purpose. A thermometer that is built into the chip enables temperature compensation.

Subsequent products from Bosch include the BME280, which adds a humidity sensor. Breakout boards using this and other barometric chips are available from sources such as Sparkfun and Adafruit, where Arduino code libraries to interpret the data are available for download.

Altitude

A barometric sensor can be used to determine altitude. At sea level, air pressure is approximately 101kPa (kilopascals), or 760 millimeters of mercury. At 5,000 meters altitude, air pressure drops to 56kPa. The transition is nonlinear, changing most rapidly at lower altitudes.

Air pressure will be affected by temperature and weather conditions. Weather is unlikely to affect the value by more than plus-or-minus 1%, and temperature has a smaller effect; but the barometric sensor can be reset to a zero point based on current temperature and weather data.

Gas Pressure

Thousands of board-mountable gas pressure sensors are available from retail component suppliers. The need to make connections via tubing imposes a minimum size limit on these parts. Many have a through-hole format, but even the surface-mount variants can be relatively large (e.g., 10mm × 10mm). Most tend to have barbed tubing ports, the "barbs" being ridges to retain push-on flexible tubing. Output may be analog or digital, the digital protocol being designed for I2C, SPI, or plain TTL serial buses.

These components are designed for pressures ranging up to 500kPa (kilopascals). Some manufacturers use psi in their specifications, while others use bars and millibars, and a few state gas pressures equivalent to inches of water.

Dual-ported gas sensors for differential pressure measurement can be used for gauge measurement (i.e., relative to ambient air) if one port is left open. Absolute pressure sensors have only one port, as they measure pressure relative to a vacuum maintained inside the chip, behind the diaphragm.

- Some gas pressure sensors can do dual duty as liquid pressure sensors, but others cannot, and datasheets must be consulted carefully, as online vendors may not include this information prominently.

The sensor in Figure 17-5 by All Sensors Corporation is designed for flexible tubing with 1/16" internal diameter (1.6mm). Its pin spacing is 0.1" (2.54mm), making it breadboardable. This sensor is intended for gases only, but has a "10-inch" water-equivalent rating, meaning that it can tolerate a maximum pressure equivalent to that of a 10-inch (25.4cm) column of water.

When supplied with up to 16VDC, an internal Wheatstone bridge circuit provides an analog output that varies ratiometrically by a few millivolts over the full range of measurable pressures (the exact specification depending on the power supply). The output may be amplified with an external op-amp. Although this is a differential sensor, it can provide gauge pressure if one of the ports is left open.

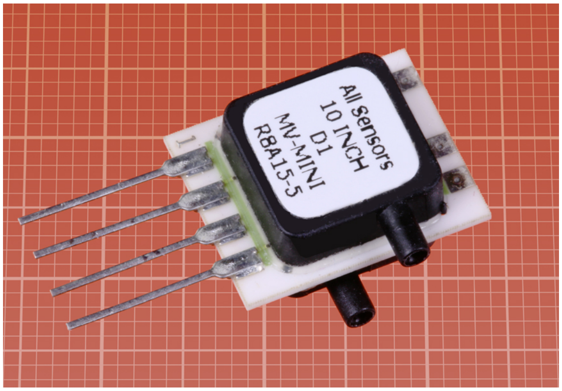

Figure 17-5 *A gas pressure sensor for use with 1/16" (1.6mm) internal diameter flexible tubing. The background grid is in millimeters.*

The larger sensor in Figure 17-6 is in the ADCA range from All Sensors Corporation, designed for flexible tubing with 1/8" internal diameter (3.2mm). This is intended for gases only, and has a "5-inch" water-equivalent rating. It requires a 5VDC power supply and has an internal op-amp that provides an output that varies ratiometrically by 0.2VDC (centered on 4VDC) over the range of measurable differential pressures. Like its smaller cousin, it can provide gauge pressure if one of the ports is left open.

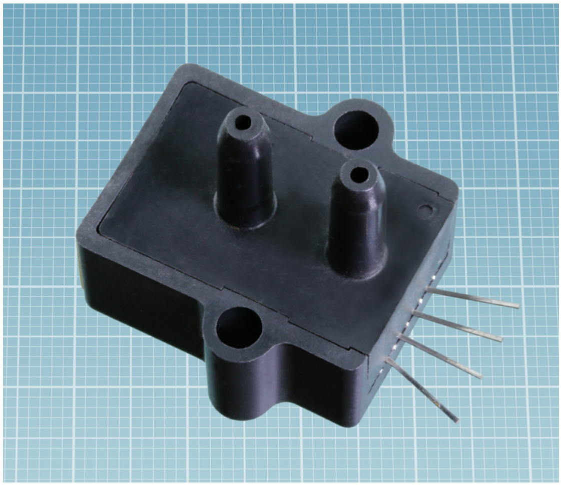

Figure 17-6 *A gas pressure sensor for use with 1/8" (3.2mm) internal diameter flexible tubing. The background grid is in millimeters.*

What Can Go Wrong

Vulnerability to Dirt, Moisture, and Corrosive Materials

MEMS pressure sensors containing very delicate, very small sensing elements must inevitably come into direct contact with gases and, in some cases, liquids. Datasheets will provide warnings regarding humidity and corrosive fluids, but the risk of contamination with dirt remains. Liquids should be filtered to minimize this risk.

Barometric sensors have a hole in the package, exposing the sensor element to the ambient air. This vent must be protected from direct contact with the environment.

Light Sensitivity

If light is allowed into the vent hole in a barometric sensor, it can create photocurrents in the chip. This will affect accuracy.

gas concentration sensor

18

Humidity sensors and *vapor sensors* are included in this entry, as they measure concentrations of a liquid while it is in its gas phase.

Sophisticated *industrial sensors* are available for accurate gas sensing, but this entry deals almost entirely with lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **gas/liquid pressure** sensor (see Chapter 17)

What It Does

Small semiconductor-based gas sensors provide a low-cost method for detecting specific gases in ambient air. The sensors vary their electrical resistance or capacitance in response to the concentration of the gas. They may be used in conjunction with an alarm system that will sound if gases such as *propane* or *carbon monoxide* exceed a preset level, or if the concentration of *oxygen* drops below a preset level.

Because vapor consists of a liquid in its gas phase, gas sensors can respond to vapor, as in the case of *alcohol sensors*.

A *humidity sensor* measures the amount of water vapor in the air, which can be important in refrigeration, HVAC (heating, venting, and air conditioning), medical equipment, meteorology, and storage rooms where art objects, antiques, or paper archives must be preserved in an environment that is neither too dry nor too moist, and is held at a constant temperature. Humidity control can be important also in climates where growth of mold is a concern. Humidity sensors are also used in automobile climate control and windshield defogging, and in the storage of food, fabrics, wood products, and medications.

High humidity coupled with high temperature contributes to the decomposition of many substances, while low humidity can cause desiccation. High humidity can also create damage if it causes materials to expand as a result of taking up moisture.

Many methods have been devised to detect individual gases, but semiconductor sensors are now dominant in applications where accuracy and gas-specificity are not crucial.

Schematic Symbol

No specific schematic symbol exists to represent a semiconductor gas sensor.

Semiconductor Gas Sensors

During the development of transistors in the 1950s, engineers noticed that semiconductor p-n junctions were sensitive to the presence of some gases in the atmosphere. This was regarded as a problem, which was solved by encapsulating transistors to prevent their exposure.

During the 1980s, Japanese law required installation of sensors in every home to detect hazardous concentrations of propane gas. This encouraged the development of cheap, long-lasting components that took advantage of semiconductor sensitivity.

Tin oxide is widely used in a variety of solid-state gas sensors. A sintered layer of the compound is deposited on a ceramic substrate in combination with other compounds such as antimony oxide. The granular layer functions as an n-type semiconductor in which electron transfer increases when certain gases are adsorbed among the grains. When the gas concentration diminishes, oxygen atoms displace the gas molecules, and the sensor returns to its original state. It is unimpaired by being activated, and has a life expectancy of at least 5 years during active (powered) use.

Each sensor includes a tiny resistive heater that is necessary for the chemical reaction to occur. Voltage must be applied to two pins that are connected with the heater. Two other pins connect internally with the sensing element. The resistance between these pins will vary with presence of the target gas; thus, this type of component has a resistive output.

It is a feature of semiconductor gas sensors that they tend to suffer from *cross-sensitivity*. That is, one sensor may respond to more than one gas. Manufacturers control this problem to some extent by adding filtering material around the semiconductor element, or by adjusting the proportions of dopants used in the semiconductor. Datasheets should be consulted carefully to see if a sensor may give false positives as a result of other gases that are likely to be present in the area where it will be used.

Figure 18-1 shows an MQ-5 propane sensor from Hanwei Electronics. The component also responds to methane, hydrogen, alcohol, and carbon monoxide, but with much less sensitivity. The manufacturer recommends calibrating the sensor by using a resistor ranging from 10K to 47K in series with the output resistance.

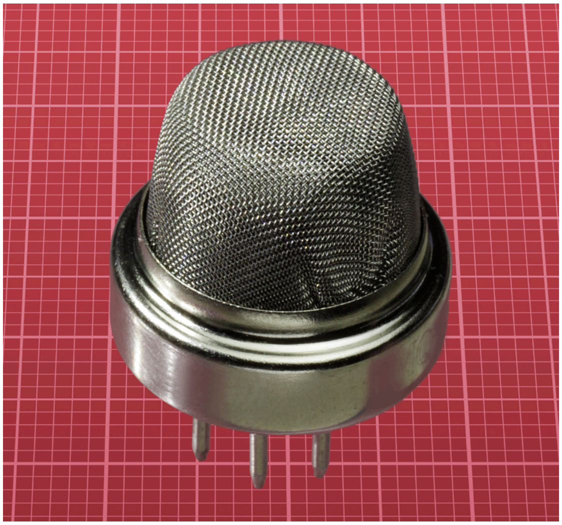

Figure 18-1 *A propane gas sensor. The background grid is in millimeters.*

Figure 18-2 shows an MQ-3 alcohol sensor, also manufactured by Hanwei. It has some sensitivity to benzene, but this is unlikely to be a problem, as benzene is seldom present in significant concentrations in ambient air. However, the response of the sensor to alcohol varies with temperature and humidity. Consequently this component can only be used as a "breathalyzer" if the user can be satisfied with an approximate response.

Semiconductor gas sensors are also available for detecting methane, carbon monoxide, hydrogen, ozone, and other gases.

These components are not low-current devices. Typically the internal heater in the Hanwei range has a resistance of slightly more than 30 ohms, and will draw 150mA to 160mA at 5V (i.e., slightly less than 1W). Because the heater is a simple resistive device, it can be used with AC or DC. The output resistance of the sensor can also be assessed with an AC or DC signal.

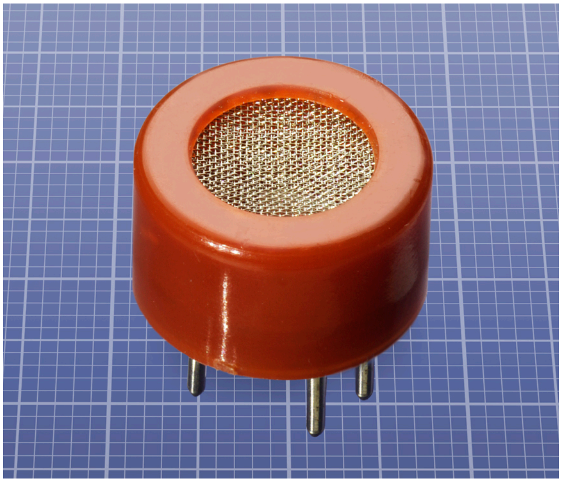

Figure 18-2 *An alcohol sensor. The background grid is in millimeters.*

A breakout board from Parallax simplifies the use of Hanwei gas sensors. The board is shown in Figure 18-3. It is compatible with carbon monoxide, propane, methane, and alcohol sensors, each of which can be plugged into a socket on the board. Two trimmers establish the sensitivity and the trip point for the sensing element, and the TTL output is then logic-high when a gas is detected, and logic-low otherwise.

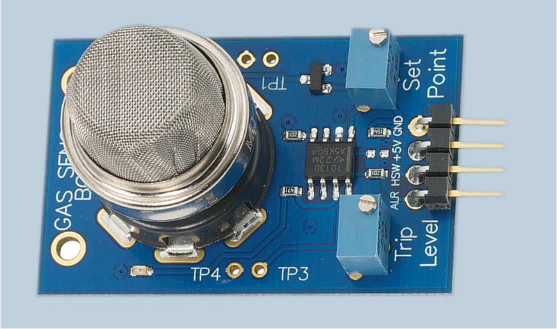

Figure 18-3 *A Parallax breakout board to simplify the use of Hanwei gas sensors.*

Oxygen Sensors

Oxygen sensors are often built using a membrane made of zirconium dioxide. This material has the property that it can transport oxygen ions when heated. One setup is to have a zirconium membrane separating the gas to be measured from ambient air. This is a type of fuel cell, called a concentration cell or Nernst cell. If the oxygen concentration differs between the two sides of the membrane, oxygen ions will flow through it. Only oxygen ions can move—not neutral oxygen atoms or molecules. The ions are negatively charged, so the transport will lead to a potential difference over the cell, which can be measured with platinum electrodes.

To comply with emissions regulations, automobiles sense the oxygen level of exhaust gases. This data controls the fuel-air ratio in the fuel injection system. Too much air results in the formation of nitrogen oxides, while too little air results in excessive carbon monoxide.

Humidity Sensors

Moisture content of the air is expressed in three different ways:

Absolute humidity

This is the weight of water vapor in a fixed volume of air. It is measured in grams per cubic meter in the metric system. An absolute humidity sensor is properly called a *hygrometer*.

Dew point

If a sample of air is cooled without a change of pressure, the dew point is the temperature at which moisture will start to condense. The dew point is a way of describing how humid the ambient air is currently, as water will condense more readily in humid conditions.

Relative humidity

This is often referred to by its acronym, RH. If temperature, pressure, and volume of a sample of air remain constant, relative humidity is the ratio between the current value of absolute humidity and the hypothetical level where the addition of water vapor would result in condensation. The ratio is expressed as a percentage. Thus, if moisture is already condensing in an air sample, relative humidity is 100%. If an air sample contains half the weight of moisture required for condensation to begin, relative humidity is 50%. If there is no moisture at all in the air, relative humidity is 0%.

Colloquial usage of the term “humidity” usually means relative humidity, and sensor output is usually convertible to this value. However, some absolute humidity sensors do exist.

Dew-Point Sensor

Historically, meteorologists used a *chilled mirror hygrometer*, in which a metal mirror was exposed to the atmosphere and cooled until the surface was seen to become misty with condensation. The temperature where this occurred was the dew point.

This system is still used in conjunction with an LED and a **phototransistor**. The LED is positioned so that its light reflects from a mirror, directly onto the phototransistor. The mirror is cooled until mist starts to form, which diffuses the reflected light and causes a well-defined change in output from the phototransistor.

Although the formula linking dew point with relative humidity is complex, a simplified approximation is available that is reasonably accurate so long as the relative humidity is 50% or greater. If RH is the relative humidity, t is the current temperature, and t_D is the dew-point temperature at which mist forms:

$$RH = 100 - (5 * (t - t_D)) \text{ approx.}$$

Although a chilled-mirror dew-point sensor has a reputation for being accurate, it is heavy, expensive, and impractical for most applications outside of meteorology.

Absolute Humidity Sensors

An absolute humidity sensor may use two NTC (negative temperature coefficient) **thermistors** in a Wheatstone bridge circuit. One thermistor is sealed in a compartment containing dry nitrogen, with zero humidity. The other is exposed to the atmosphere. Current passing through the thermistors raises their temperature to at least 200 degrees Celsius. Because heat radiates less efficiently when there is moisture in the air, the exposed thermistor will run hotter, and its resistance will be higher, at higher levels of humidity. This type of sensor is used in clothes dryers and wood kilns, among other applications.

See Chapters 23 and 24 for more information about thermistors.

Relative Humidity Sensors

Two main types of sensing elements are used to measure relative humidity: resistive and capacitive.

In a *resistive* sensing element, a thin layer of polymer, salt, or other hygroscopic substance is deposited on a substrate consisting of ceramic or other unreactive material. When the substance absorbs water, its electrical conductivity increases. Voltage is applied to the sensing element, and AC is used to avoid polarizing it. The current flow is processed externally, with conversion to DC followed by linearization, meaning that the current is processed to establish an almost linear relationship with gas concentration, with temperature compensation factored in. Alternatively, these functions can be performed by hardware built into the sensor, and a digital value for relative humidity can be accessed by an external microcontroller.

In a *capacitive sensor*, once again a thin film of polymer or metal oxide is deposited on a ceramic or glass substrate, but the film functions as a *dielectric* between two metal electrodes that serve as the plates of a capacitor. The dielectric value changes as the film absorbs moisture, and this causes the capacitance to vary, typically by 0.2pF to 0.5pF for each 1% change in relative humidity. This is almost a linear relationship extending over the entire range from 0% to 100% relative humidity.

The actual capacitance value at 50% relative humidity is likely to be between 100pF and 500pF. The sensor may be excited with AC from an external source, or can be incorporated in a chip that derives AC from a DC power supply and provides a digital output.

To determine the dew point or absolute humidity from a value for relative humidity, ambient temperature must also be measured. A chip such as the Si7005 from Silicon Labs includes a temperature sensor with a relative-humidity sensor based around a capacitor in which polyimide film forms the dielectric. If condensation forms, an on-chip heater will cause it to evaporate so that normal operation can resume. Data from the chip is supplied via an I2C interface.

Humidity Sensor Output

When the output is analog, resistance or capacitance of the internal sensing element is available via two pins or solder pads on the sensor. The analog value must be converted to a value for relative humidity by performing a calculation that takes the current temperature into account. The component may or may not include a temperature sensor.

With a digital output, an internal analog-to-digital converter can be accessed by a microcontroller via serial, I2C, or SPI interface. Alternatively, the sensor may communicate values using pulse-width modulation. Either way, the output is a value for relative humidity, calculated on the chip with reference to an onboard temperature sensor.

Analog Humidity Sensor

The Humirel HS1101 is a low-cost analog-output humidity sensor that varies its internal capacitance between approximately 160pF and 200pF as relative humidity increases from 0% to 100%. The response is almost linear, with the curve steepening slightly when humidity exceeds 80%. The component is shown in Figure 18-4.

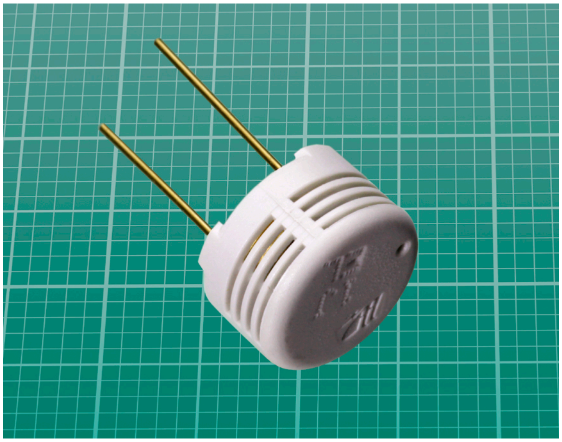

Figure 18-4 *A Humirel HS1101 humidity sensor. The background grid is in millimeters.*

The manufacturer claims a recovery time of 10 seconds after 150 hours of condensation. In other words, the performance of the sensor should be restored to its original specification.

A microcontroller can evaluate the output by measuring the charge time of the internal capacitor in the sensor. If the sensing element is wired in parallel with a 10M resistor, this will allow the capacitor to discharge before the microcontroller charges it again. A 220-ohm series resistor should be used between the microcontroller and the sensor, to limit the charge current. This circuit is illustrated in Figure 18-5.

Alternatively, the sensor manufacturer suggests using the capacitance value to control the output frequency of a 555 timer. A counter or microcontroller may be used to count the number of pulses per unit of time.

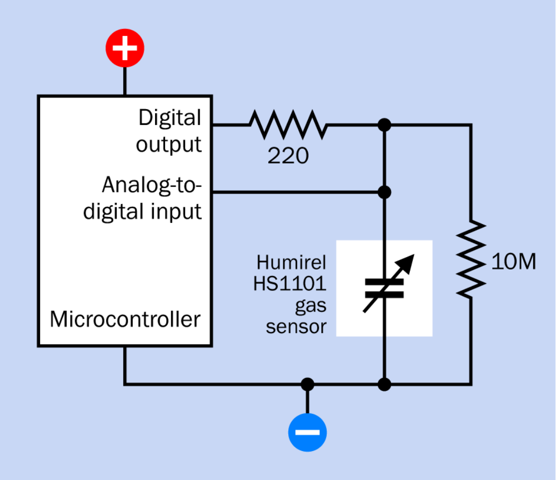

Figure 18-5 *Using a microcontroller to measure the charge time of a gas sensor with capacitive output.*

The Honeywell HIH4030 is a surface-mount humidity sensor with a more convenient analog voltage output that increases almost linearly from approximately 0.8VDC at 0% humidity to 3.8VDC at 100% humidity, assuming a 5VDC power supply. The sensor is available on a miniature breakout board from Sparkfun, shown in Figure 18-6.

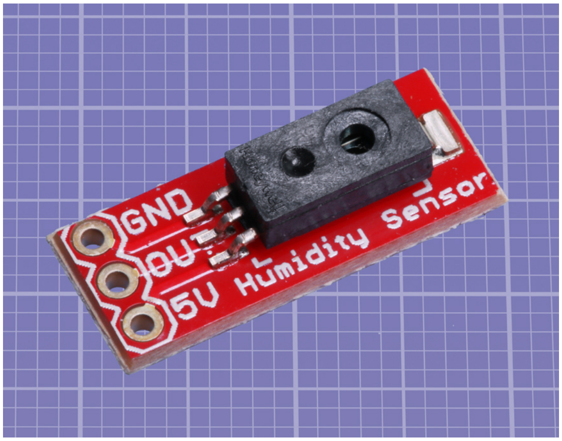

Figure 18-6 *The Honeywell HIH4030 on a breakout board from Sparkfun.*

Design Considerations

As relative humidity is temperature-dependent, a relative-humidity sensor must have the same temperature as the air it measures. Many datasheets recommend that when a sensor is mounted on a printed circuit board, slots should be milled around it to minimize heat transfer to it. The sensor should also be mounted as far away as possible from heat-generating components.

Where a capacitive humidity sensor with an analog output is located some distance from the electronics that will process its output, shielded cables or twisted pair cables should be used to minimize capacitance in the wiring. A decoupling capacitor between the supply voltage and ground close to the sensor can help to keep the supply voltage stable.

Digital Humidity Sensor

The AM2302, available from Adafruit, is a capacitive humidity sensor with a digital output accessible by a microcontroller via the I2C protocol. Onboard electronics calculate relative humidity with reference to an included temperature sensor. This component is pictured in Figure 18-7.

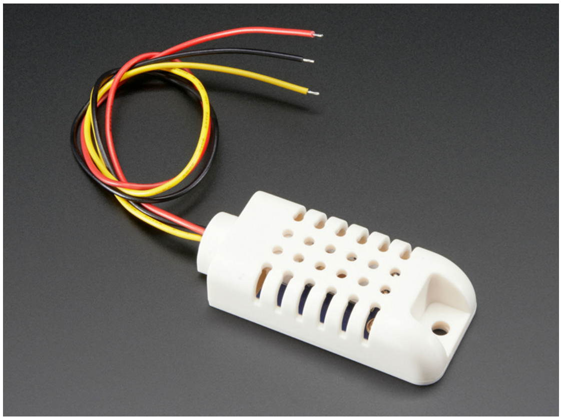

Figure 18-7 *A low-cost humidity sensor with temperature-compensated digital output. From a photograph by Adafruit.*

What Can Go Wrong

Contamination

A semiconductor gas sensor can be damaged by exposure to volatile chemical vapors. This would be unusual in a domestic environment, but is still an important consideration, as there will be no obvious indication that damage has occurred.

Recalibration

If a humidity sensor is exposed to very high humidity where condensation occurs, some datasheets advise a baking procedure, where the sensor is placed in warm, dry air for several hours and then allowed to rehumidify.

Soldering

Semiconductor gas sensors should be soldered quickly and at a controlled temperature, to minimize heat transfer.

gas flow rate sensor

19

Sensors that contain no electronic components are not included in this entry.

A **gas flow rate** sensor may often be described as a *mass flow* sensor or *mass flow rate* sensor. Although it measures volume rather than mass, the mass of gas can be calculated so long as its temperature and pressure are controlled.

Some methods of gas flow sensing can also be applied to liquids, but sensors are usually designed for one application or the other. Therefore, **liquid flow rate** sensors have their own entry. See Chapter 16.

An *anemometer* is a gas flow rate sensor that measures air speed. It is included in this entry.

Many gas flow rate sensors are large devices designed for industrial applications. This entry focuses mostly on lower-cost solid-state sensors that are often classified as *board-mount* components.

OTHER RELATED COMPONENTS

- **liquid flow rate** sensor (see Chapter 16)

What It Does

A **gas flow rate** sensor measures the volume of gas flowing past or through the device, usually inside a pipe. In most applications, users wish to know the mass of the gas that is passing per unit of time. Consequently a gas flow rate sensor is very often identified as a *mass flow rate* sensor. If it functions by heating the gas and measuring the heat dissipation, it is a *thermal mass flow rate sensor*.

A sensor that measures the flow of open air is often described as an *anemometer*. Its output will be expressed as a velocity, not as a volume or mass.

Applications

Mass flow rate sensors are used frequently for laboratory and medical applications, although the reliability and affordability of thermal mass flow rate sensors is now making them attractive for metering municipal gas supplies.

Anemometers are used mostly in meteorology, aviation, and on boats.

Schematic Symbol

Many specialized symbols are used in flow diagrams to represent pumps, valves, and sensors, but they are not schematic symbols of the type generally found in electronic circuit diagrams, and are not included here.

How It Works

Because an anemometer functions so differently from a gas flow rate sensor that is enclosed in a pipe, the two types of sensors will be discussed separately.

Anemometer

The name of this device is derived from the Greek word *anemos*, meaning “wind.” The *cup anemometer* was invented in 1846, using four hemispherical cups attached to horizontal arms rotating on a vertical shaft. The rate of rotation was proportional to wind speed over a substantial range, but the conversion factor between wind speed and RPM varied depending on the size of the cups and their distance from the shaft.

Anemometer design was simplified to three cups in 1926 and was modified in 1991 by adding a tag to one cup. This causes the speed of rotation to fluctuate as the tag rotates through the wind stream, and the direction of the wind can be calculated from the speed fluctuations. Not all anemometers rely on this principle, however, and a separate wind vane can be used to determine wind direction.

The basic design of a modern anemometer is illustrated in Figure 19-1.

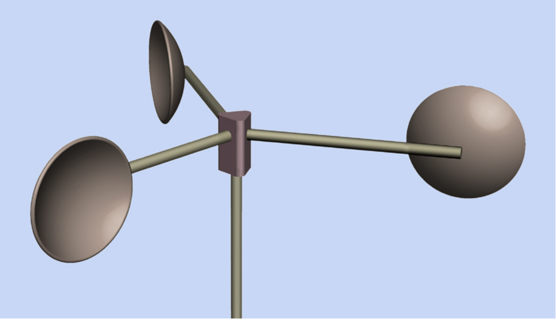

Figure 19-1 *The basic design of a metereological cup anemometer.*

Anemometers traditionally used a mechanical counter to log the number of rotations, which were checked at fixed intervals to derive the wind speed. The output of a modern anemometer may be achieved by generating AC or DC power, or from *Hall-effect sensors* (see “Hall-Effect Sensor” for a full discussion of Hall-effect sensors).

Handheld Anemometer

A digital handheld anemometer for personal use is shown in Figure 19-2. A cup anemometer made by Vaavud for use with a mobile phone (with appropriate software) is shown in Figure 19-3.

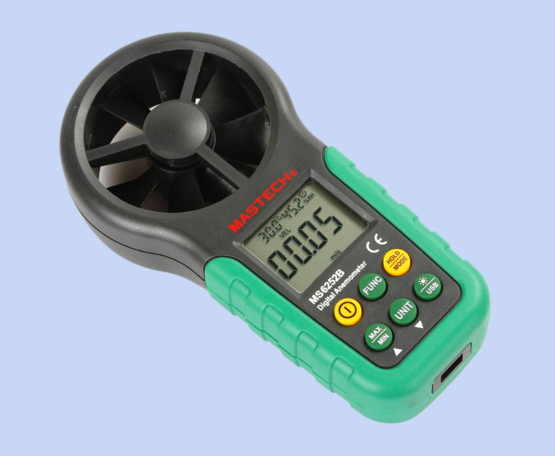

Figure 19-2 *A handheld digital anemometer.*

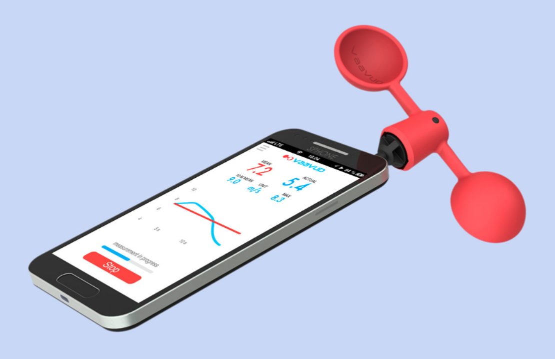

Figure 19-3 *A cup anemometer sold as an accessory for smartphones.*

Ultrasound Anemometer

The movement of air affects the speed of sound, enabling calculation of both wind direction and wind velocity by using an array of ultrasound emitters and receivers. The lack of rotating parts promises greater reliability, and in Figure 19-4 an ultrasound anemometer manufactured by Biral Metereological Sensors also includes heaters so that it will be immune to snow or ice accumulation during freezing conditions.

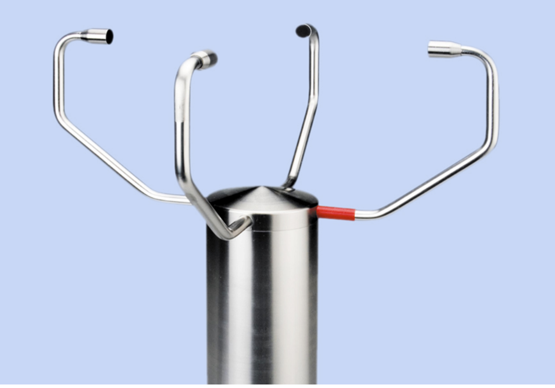

Figure 19-4 *Made by Biral Metereological Sensors, this anemometer uses ultrasound to determine wind speed and direction.*

Ultrasound anemometers have been built on a DIY basis by hobbyists, typically using off-the-shelf ultrasound emitters and an Arduino to decode the signals. Several of these projects are documented online.

Hot Wire Anemometer

A hot wire anemometer measures the cooling effect of the air. It heats a thin wire by passing current through it, and measures the heating power needed to keep the temperature constant.

It is also possible to keep either the voltage over the wire or the current constant, and to assess the wire temperature. The temperature can be measured directly or calculated from the wire's resistance, which increases as the wire gets hotter.

Mass Flow Rate Sensing

A mass flow rate sensor measures the flow speed of a gas. When this is multiplied by the density, the mass flow rate can be calculated.

Most sensors of this type heat the gas and are categorized as *thermal mass flow rate* sensors. The gas passes over a *thermopile* (consisting of several thermocouples wired in series), then a heater, and then another thermopile. These components are miniaturized and can be etched into a chip measuring 2mm square or less.

The temperature difference between the two thermopiles increases as the flow of gas becomes faster and transports more heat to the second thermopile. This is known as the *heat transfer principle*, illustrated in Figure 19-5.

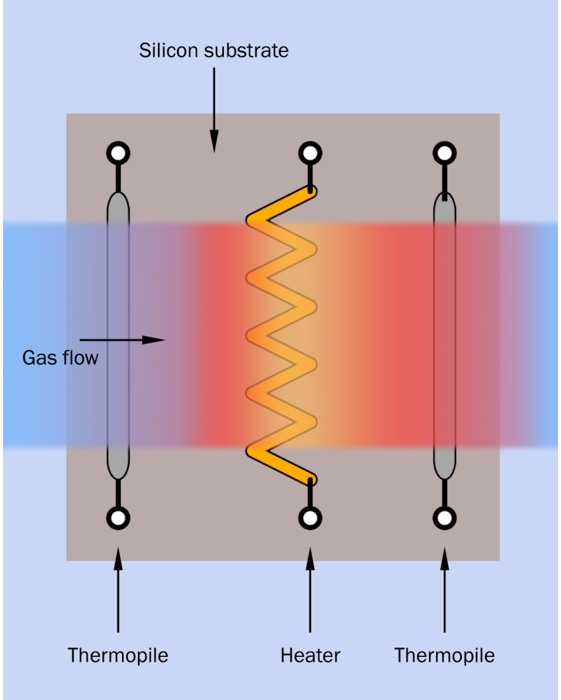

Figure 19-5 *In a thermal mass flow sensor, the flow of gas heats the thermopiles disproportionately, and the temperature difference can be used to evaluate the rate of flow.*

This principle is used also in a liquid thermal mass flow rate sensor, described in "Thermal Mass Liquid Flow Rate Sensor".

An example of a thermal mass flow rate sensor is shown in Figure 19-6.

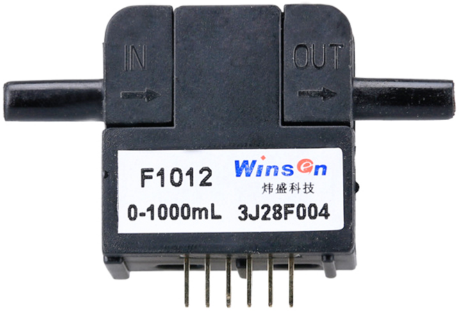

Figure 19-6 *A thermal mass flow rate sensor made by the Chinese manufacturer Zhengzhou Winsen Electronic Technology Co. Ltd.*

Applications

In medicine, mass flow rate sensors are used in anesthesia delivery, respiratory monitoring, sleep apnea machines, ventilators, and other devices. Industrial uses include air-to-fuel ratio measurement, gas leak detection, and gas metering.

While the metering of municipal gas supplies was traditionally done with all-mechanical devices, MEMS-based meters are replacing many of the 400 million mechanical gas meters estimated to exist worldwide.

Units

Mass flow rate sensors are often rated in SLM, meaning standard liters per minute. A "standard liter" has a temperature of 0 degrees Celsius and a pressure of 101.325 kPa (kilopascals). This pressure is equivalent to that of air at sea level. Because temperature and pressure are specified, the mass of a gas in a standard liter can be calculated by knowing the density of the gas. Thus SLM is a way of measuring mass flow, even though it is a measurement of volume.

The acronym SLPM is also used, but it means the same thing as SLM. SLs and SLPs are measurements of standard liters per second, while SCCM refers to standard cubic centimeters per minute.

Measuring Higher Volumes

MEMS sensors are typically equipped with inflow and outflow nozzles suitable for flexible tubing of 3mm or 5mm internal diameter. The nozzles are "barbed," meaning that they have ridges to retain the tubing.

Small tubing can only deal with low flow rates. A few sensors are threaded for standard plumbing pipes, and can accept volumes of up to 10 liters per minute. They are a minority. A low-rate sensor can still be used to measure higher volumes if it is supplied with just a percentage of the primary flow. This principle is illustrated in Figure 19-7, where an adjustable vane in the main pipe creates a pressure differential. A narrower, constricting section of pipe could have a similar effect, but it would not be adjustable.

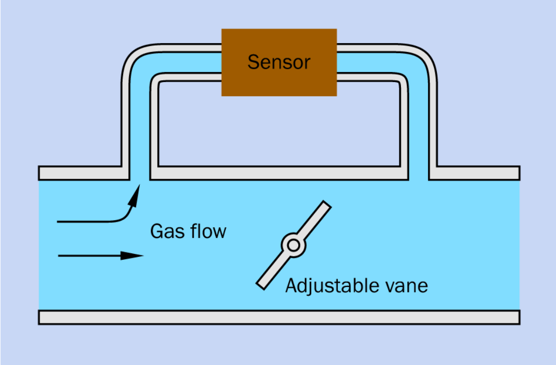

Figure 19-7 *A vane in a primary pipe can divert a percentage of the flow to a sensor.*

Output

Many mass flow sensors have an analog output consisting of a voltage that varies ratiometrically with gas flow. With a typical 5VDC power supply, output voltage may vary from around 1VDC to 4VDC.

Some sensors now incorporate analog-to-digital converters and data processing to provide SLM digital values, accessible from a microcontroller via an I2C interface.

What Can Go Wrong

The primary risk for gas flow rate sensors is damage caused by particles and contaminants in the gas stream. A 5-micron filter is recommended. A *dust segregation system* can also be used, consisting of a small compartment containing semicircular centrifugal chambers. Dust tends to follow the outer edge of the flow path, while the flow sensor is placed on the inner side of the path.

photoresistor

20

A **photoresistor** has a function similar to that of a **phototransistor**, but as its name implies, it is a purely passive component that varies its resistance in response to light.

The term *photocell* was used formerly, but has been displaced by the term **photoresistor**, which describes its function more accurately. The term *photoconductive cell* is sometimes used, or *light-dependent resistor* (or its acronym, *LDR*).

OTHER RELATED COMPONENTS

- **resistor** (see Volume 1)
- **photodiode** (see Chapter 21)
- **phototransistor** (see Chapter 22)

What It Does

Formerly known as a *photocell*, a **photoresistor** is a disc-shaped component with two leads. When light falls on the surface of the disc, resistance between the leads will diminish. Some photoresistors have a resistance in darkness as high as 10 megohms. A few have a resistance in bright light as low as 500 ohms, although several kilohms would be more common.

A photoresistor is less sensitive to light than a **phototransistor** or **photodiode**, and unlike them it is a passive component with no polarity. It presents equal resistance to current in either direction, and may be used with DC or AC.

Because cadmium sulfide is commonly used in this component and is regarded as hazardous to the environment, photoresistors are now unavailable in some regions (notably, Europe). However, at the time of writing, they are still available from many Asian sources, and from some importers in the United States.

Schematic Symbol

Six schematic symbols for a photoresistor are shown in Figure 20-1. They are functionally identical, regardless of whether the single slanting arrow across the resistor symbol is omitted.

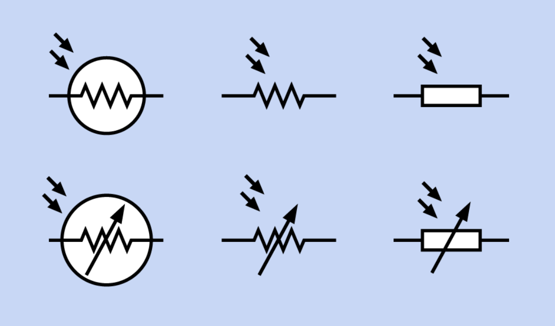

Figure 20-1 *All six symbols for a photoresistor are functionally identical.*

How It Works

Cadmium sulfide is a semiconductor. When exposed to light, more charge carriers are excited into states where they are mobile and can participate in conduction. As a result, electrical resistance decreases.

Construction

A closeup of a photoresistor appears in Figure 20-2. The brown material is a layer of cadmium sulfide deposited onto a ceramic base. The silver material is a conductive compound evaporated onto the cadmium sulfide to form two interlocking electrodes, in a pattern that maximizes the length of the boundary between each of them and the semiconductor. The electrodes connect with leads projecting from the back of the component.

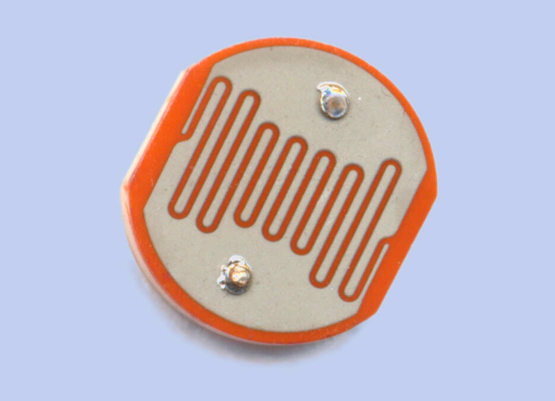

Figure 20-2 *Closeup of a photoresistor. Two interlocking electrodes are mounted on a brown semiconductor layer.*

Variants

Figure 20-3 shows a variety of photoresistors, illustrating the range of sizes available. Small photoresistors may be less than 5mm in diameter; large ones may be 25mm in diameter. The size generally suggests the ability of the component to pass current.

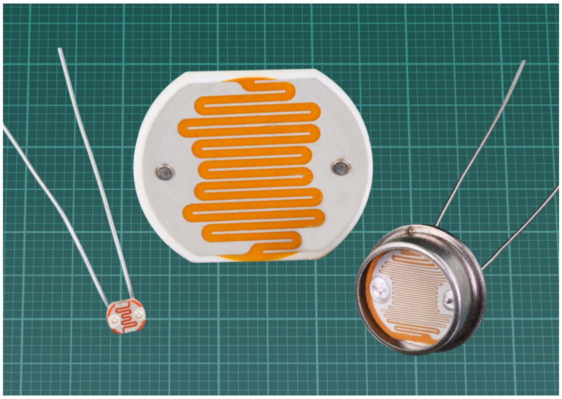

Figure 20-3 *Photoresistors are available in a wide range of sizes. The component in the center is shown at the same scale as the others. Generally speaking, larger components are able to conduct higher currents. The background grid is in millimeters.*

Photoresistors in Optical Isolators

An *optical isolator*, popularly known as an *optocoupler*, contains an LED opposite to a photoresistor, in a sealed package. This is discussed in Volume 2. A *Vactrol* is a similar component, except that the LED is placed opposite to a photoresistor. An example is shown in Figure 20-4.

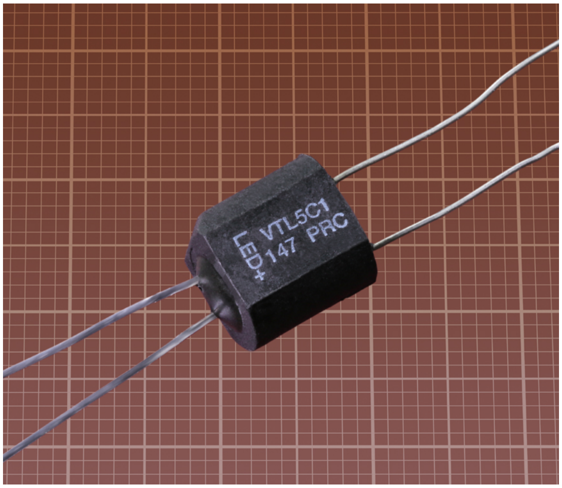

Figure 20-4 *A Vactrol, containing an LED opposite to a photoresistor. The background grid is in millimeters.*

Vactrol is a brand name owned by its initial manufacturer, Vactec. It was developed to control a vacuum tube, hence its name. In the 1950s, Vactrols were used in guitar amplifiers to control tremolo.

Because of the relatively slow response time of a photoresistor, and its sensitivity to temperature, optical isolators based on photoresistors are not used in digital devices. They still retain some utility in audio components and music equipment, where the ability of the photoresistor to pass AC is an advantage and its response time is adequate.

Values

Datasheets for a few photoresistors can still be found from some suppliers, such as Digi-Key, but are mostly unobtainable, as major semiconductor companies have stopped making these components. Vendors may quote basic values, but in the absence of part numbers or a manufacturer name, a buyer cannot verify the information.

The resistance range can be determined by testing a sample component. A typical range would be from 50K in a light intensity of 10 lux, up to 1M in darkness. Maximum power dissipation is likely to range from 100mW for a small photoresistor to 500mW for a large one.

Maximum voltage may be as high as 200V, but photoresistors will work just as well at low voltages.

Comparisons with a Phototransistor

Slower response

A photoresistor typically takes several milliseconds to respond to bright light, and can require more than 1 second to regain its dark resistance. A phototransistor is much more responsive, and a photodiode is faster still.

Narrower range of resistance

The maximum resistance of a photoresistor is almost always significantly lower than the effective maximum resistance of a phototransistor in darkness, and the minimum resistance is likely to be significantly higher than that of a phototransistor in bright light.

Greater current-carrying capacity

Often a photoresistor is rated for twice as much current as the output from a phototransistor.

Not directional

Because a photoresistor is not packaged with a lens, it is sensitive to incident light from anywhere in front of it. If an application requires that light sensitivity should be confined within a narrow angle while the component is insensitive to incident light from other directions, a phototransistor or photodiode should be used.

Temperature dependent

The resistance of a photoresistor varies more with temperature than the effective resistance of a phototransistor.

Cost

At this time, photoresistors are likely to be more expensive than phototransistors.

Lack of information

Photoresistors are often sold without means of checking their specification in a datasheet.

How to Use It

While the effective resistance of phototransistors and photodiodes varies with the applied voltage, photoresistors present the same resistance for a particular light intensity regardless of the voltage applied. This property has made them suitable for use in "stomp box" guitar-effects pedals.

Because the minimum resistance of a photoresistor in bright light tends to be relatively high, and because its response time is quite slow, it is suitable primarily as an analog component rather than as a switch.

In principle, any resistor in a circuit can be replaced by a photoresistor, and its smooth response to light variations would make it suitable, for example, as the resistance in an RC oscillator circuit, where it would determine the charge time of a capacitor. The frequency of the circuit would thus become light-dependent.

Cadmium sulfide photoresistors respond most actively to wavelengths of light ranging from 400nm to 800nm. This is especially important where an **LED indicator** is used as a light source, as LEDs often emit an extremely narrow range of wavelengths. (See the entry for LED indicators in Volume 2.)

Choosing a Series Resistor

To convert light intensity into a voltage, a photoresistor can be connected in series with a regular resistor, to form a voltage divider. There are two ways to do this, as shown in Figure 20-5. On the left side of this figure, light falling on the photoresistor will cause the output voltage to rise. On the right side, light will cause the output to drop.

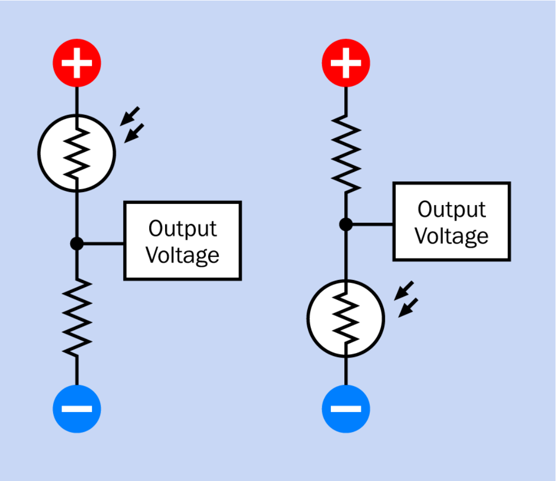

Figure 20-5 *Using a photoresistor to create a variable voltage. See text for details.*

If R_{MIN} is the minimum photoresistor value in the brightest light that will be used, and R_{MAX} is the maximum value in the dimmest light that will be used, a simple formula can be used to find R_S, the optimum value for a series resistor, which will provide the widest range of voltage values at the center of the voltage divider:

$$R_s = \sqrt{R_{min} * R_{max}}$$

What Can Go Wrong

Overload

Because a datasheet may be unavailable, a photoresistor may have to be used on a trial-and-error basis. A test-to-destruction approach may be necessary to determine the limits of the component.

Excessive Voltage

Exceeding the maximum rated voltage for a photoresistor, even for a short time, can cause irreversible damage. Depending on the component, overvoltage can range from 100V to 300V.

Confusion Among Components

Because photoresistors often have no part numbers printed on them and may be sold in miscellaneous assortments, one component may be easily confused with another that has different characteristics. Where two or more photoresistors are used in one device, their characteristics should be measured to determine whether they are functionally identical.

photodiode

21

OTHER RELATED COMPONENTS

- **diode** (see Volume 1)
- **photoresistor** (see Chapter 20)
- **phototransistor** (see Chapter 22)

What It Does

Light falling on a **photodiode** causes it to generate a very small current. This is often called the *photovoltaic effect*. The component functions like a solar panel; in fact, a solar panel can be thought of as being an array of very large photodiodes.

Often, a DC power source is used to apply *reverse bias* to a photodiode. This enables the component to deliver more current. It is now operating in *photoconductive* mode.

Schematic Symbols

Schematic symbols for a photodiode are shown in Figure 21-1. Wavy arrows are often (but not always) used to indicate infrared light.

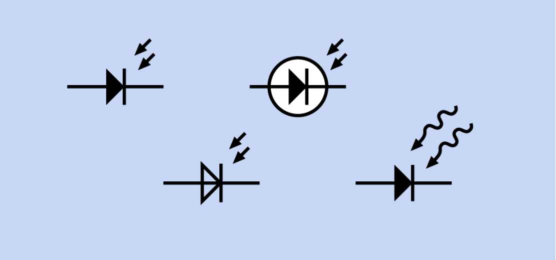

Figure 21-1 *Symbols representing a photodiode.*

Applications

The rapid response of photodiodes makes them suitable for use in optical disc drives, telecommunications, infrared data transfer, digital cameras, and optical switches. Many sensors in this Encyclopedia use a photodiode. **Proximity** sensors, optical encoders, and light meters are examples.

How It Works

When light falls on a semiconductor, it can excite an electron to a higher energy state. The electron then becomes mobile, and leaves behind an *electron hole* (See the entry discussing **diodes** in Volume 1.)

In photovoltaic mode, incident light creates pairs of electrons and electron holes in the semiconductor material. The electrons move to the cathode of the diode while the holes move to the anode, creating voltage between the two. Note that this happens to some extent even in the absence of visible light, as the photodiode may respond to infrared radiation. The tiny amount of current created without visible light is called *dark current*.

In photoconductive mode, light falling on the photodiode creates pairs of electrons and electron holes in the semiconductor material. These will move in opposite directions due to the bias voltage, which means a small current flows through the diode.

Photoconductive mode enables a faster response than photovoltaic mode, because reverse-bias voltage makes the depletion layer wider, which in turn decreases the capacitance. (The same effect is used in *capacitance diodes*.)

Simplified circuits showing the component in photovoltaic mode and photoconductive mode appear in Figure 21-2.

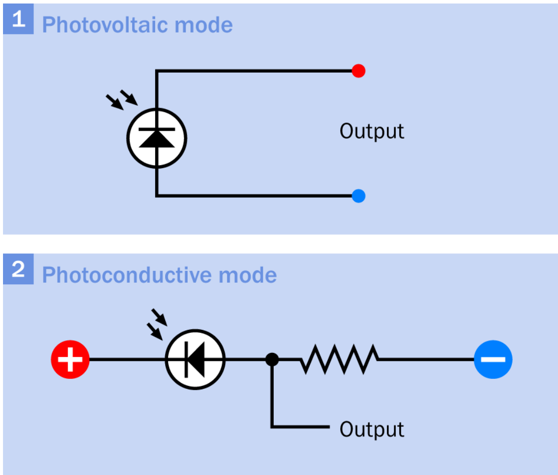

1 Photovoltaic mode

2 Photoconductive mode

Figure 21-2 *Two modes in which a photodiode may be used. In photovoltaic mode, voltage between the two pins must be measured.*

Variants

PIN Photodiodes

Like *PIN diodes*, PIN photodiodes incorporate an undoped (intrinsic) semiconductor layer between the p- and n-doped layers. They are more sensitive and have faster responses than regular PN photodiodes. Many of the photodiodes available are of the PIN type.

Avalanche Diodes

When light enters the undoped region of the avalanche photodiode, it triggers the creation of electron-hole pairs. When electrons migrate toward the *avalanche region* of the diode, the cumulative field strength increases their velocity to the point where collisions with the crystal lattice create further electron-hole pairs.

This behavior causes the avalanche diode to be more sensitive than a PIN photodiode. However, the sensitivity also makes it vulnerable to electrical noise, and it is significantly affected by heat. A guard ring is added around the p-n junction, and a heat sink is often used.

Packages

Many surface-mount and through-hole versions are available. A selection is shown in Figure 21-3 and Figure 21-4. While a through-hole photodiode may look indistinguishable from a through-hole 3mm or 5mm LED, versions are available without lenses, and some are sensitive to incident light coming from the side (they are referred to as *side-looking* or *side-view* variants).

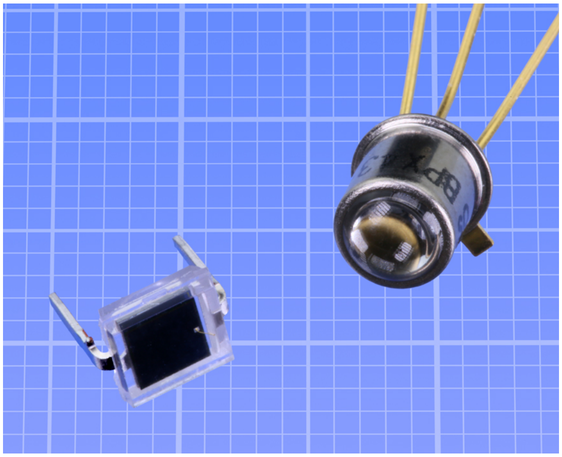

Figure 21-3 *Two sample photodiodes. Left: top-view unfiltered Vishay BPW34. Right: Osram BPX43 in metal can suitable for temperatures up to 125 degrees Celsius.*

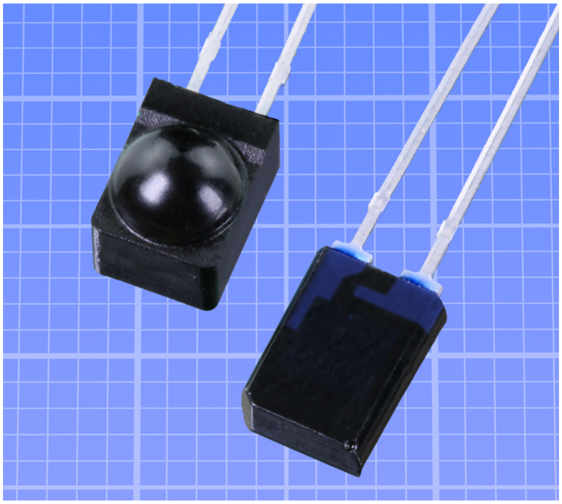

Figure 21-4 *Two side-looking photodiodes. Left: Vishay BPV22NF with lens. Right: Vishay BPW83 without a lens. Both have daylight-blocking filters.*

Wavelength Range

In order for a photon to be detected by the photodiode, it must carry enough energy to be able to create an electron-hole pair. This energy is a property specific to the semiconduct material, and is known as its *band gap*. Additionally, the epoxy packaging of the photodiode may be designed to block some wavelengths of light. Often, an application will require that the component should only respond to infrared light, not visible light.

Photodiode Arrays

A *photodiode array* has several photodiodes mounted in a row or in a grid, for imaging or measurement applications. A row of photodiodes may be used in a flatbed scanner, where it is moved relative to the reflective object being scanned.

In some arrays, photodiodes are available with color filters preinstalled, to facilitate full-color scanning using the transmitted primary colors.

Output Options

Because the output from a photodiode must be processed to be usable, options exist to convert it to a more convenient form, such as a wider range of voltages, a square wave signal where the frequency is proportional with light input, or a binary value accessible by a microcontroller via a serial bus such as I2C. For a discussion of sensor outputs see Appendix A.

Specific Variants

Light to Frequency

The Taos TSL235R is a 3-pin, through-hole chip. It combines a photodiode with logic that creates a square-wave pulse train in which frequency is proportional to the light intensity.

Logarithmic Light Meter

The Sharp GA1A1S202WP light sensor has an output voltage that changes logarithmically with the light level. This gives the sensor a large dynamic range, from 3 lux to 55,000 lux, without requiring a high-resolution analog-to-digital converter. (Human perception of light and sound levels is approximately logarithmic.) This is a surface-mount chip, but is available on a breakout board from Adafruit.

Ultraviolet to Analog

The ML8511 from Lapis Semiconductor combines an ultraviolet-sensitive photodiode with an op-amp that provides an output voltage from approximately 1V to 3V, varying with ultraviolet intensity. A breakout board containing this surface-mount chip is available from Sparkfun and many other sources.

Ultraviolet to Digital

The SI1145 from SiLabs combines ultraviolet sensing with data processing to create a UV index, readable from a microcontroller with I2C. Adafruit offers it on a breakout board.

Color to Digital

The Taos TCS3414FN module contains photodiodes sensitive to red, green, blue, and clear (no filtering). Four 16-bit analog-to-digital converters, one for each channel, provide a digital output accessible over an I2C bus. This module can be used to determine the color of ambient light with some accuracy.

Color to Analog

The Taos TCS3200 also uses red, green, blue, and clear photodiodes but encodes the output from each as a square wave in which the frequency corresponds with the light intensity. The surface-mount chip is available on a breakout board from Robot Shop, shown in Figure 21-5.

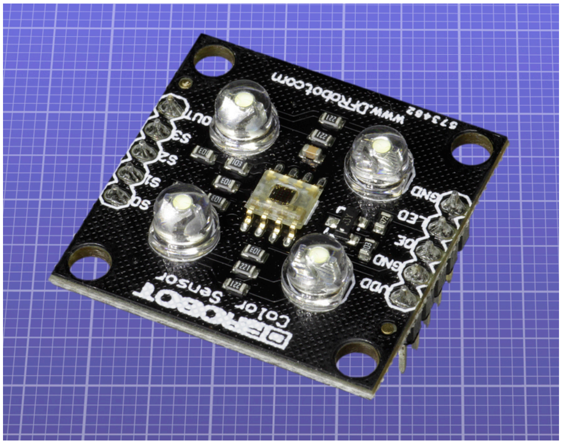

Figure 21-5 *This breakout board from Robot Shop uses the Taos TCS3200 chip to analyze the color of incident light.*

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH229FA infrared photodiode, which resembles a 3mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In Figure 21-6 the SFH229FA is shown beside the SFH229, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to below 400nm, in the green part of the visible spectrum. With the exception of their spectral range, the two photodiodes have identical specifications.

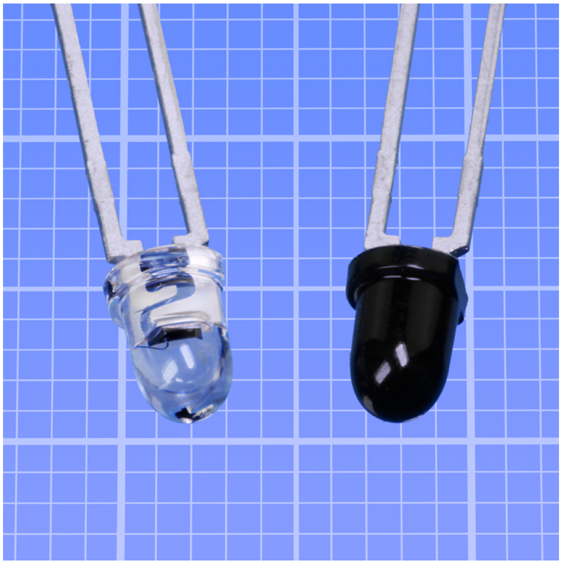

Figure 21-6 *Osram infrared photodiodes SFH229 (left) and SFH229FA (right). The background grid is in millimeters.*

- Typical forward voltage: V_F (1.3V)
- Typical photocurrent: I_P (20µA)
- Maximum power dissipation: P_{TOT} (150mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which the sensitivity has dropped by 50% (plus or minus 17 degrees)
- Dark current: I_R (50pA)
- Wavelength of maximum sensitivity: $\lambda_{S\ MAX}$ (880nm)
- Response speed is the rise and fall time of photocurrent: t_r and t_f (10ns)

Infrared photodiodes exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

The angle of sensitivity depends on the geometry of the package.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical photodiode can be 1,000 times faster than that of a phototransistor. See "Values" in Chapter 22 for a comparison. Note also that the dark current of a photodiode is much lower than that for a phototransistor.

How to Use It

In photoconductive mode, the photodiode can be connected in series with a suitable resistor so that a voltage divider is formed, as shown in Figure 21-2. The voltage at the output will then vary almost linearly with the intensity of the light.

In the photoconductive mode of operation, the output signal is generally measured in millivolts and microamperes. This signal needs to be amplified, usually with an **op-amp** (described in Volume 2).

Figure 21-7 shows simplified schematics for a standard voltage amplifier, in section 1, and a transimpedance amplifier, in section 2.

A transimpedance amplifier measures the current through the photodiode and converts it into a voltage, without the need for a voltage divider. Advantages include less noise, and no need to determine the value of a voltage-divider resistor.

The output voltage of this amplifier is calculated by using this simple formula:

$$V = R * I_P$$

R is the value of the feedback resistor, which determines the gain of the amplifier.

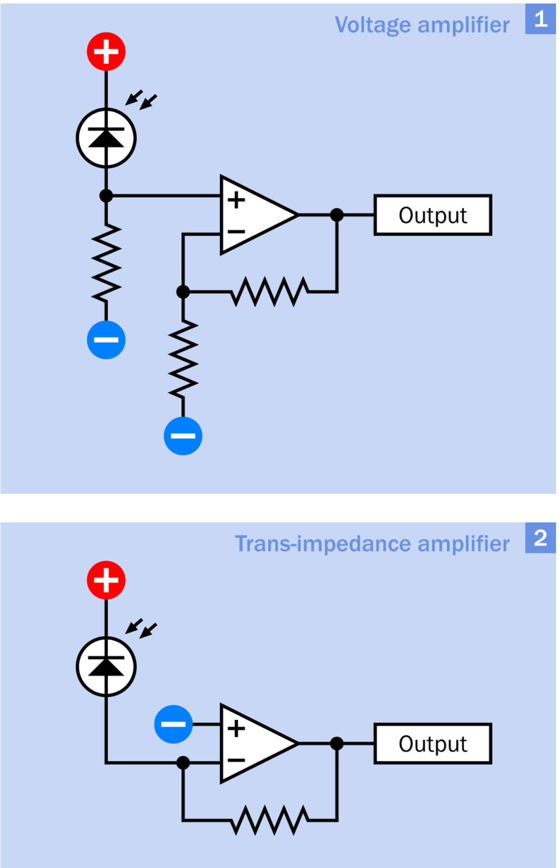

Figure 21-7 *Simplified op-amp circuits for use with photodiodes.*

What Can Go Wrong

Photodiodes can be hard to distinguish from phototransistors and LEDs, in particular infrared ones. They are typically not marked with type numbers. Using a meter may not be helpful, as regular LEDs behave similarly to photodiodes.

Decision procedure:

1. Does the component conduct in either direction while shielded from light? If so, it is a diode and not a phototransistor or photodiode.
2. Pass a weak current (e.g., 4mA) in the forward direction. If it emits visible or infrared light, it's an LED. (Infrared light may be visible when using a digital camera, or can be detected with a known phototransistor or photodiode.) If the package is clear but cloudy so that the die cannot be seen, it is probably a white LED in which the cloudiness is the fluorescent pigment converting blue light into white.

phototransistor

22

OTHER RELATED COMPONENTS

- **photodiode** (see Chapter 21)
- **photoresistor** (see Chapter 20)
- **passive infrared** sensor (see Chapter 4)
- **transistor** (see Volume 1)

What It Does

A **phototransistor** is a transistor controlled by exposure to light. (**Transistors** are described in Volume 1.) It can be either a bipolar transistor or field-effect transistor (FET), and its body is often superficially similar in appearance to a 3mm or 5mm **photodiode** or **LED indicator** encased in resin or plastic. (LED indicators are described in Volume 2.) However, some phototransistors are encased in a metal shell with a window in it.

The window or the plastic body is either transparent to visible light, or may appear black if the component is intended for use only with infrared while blocking visible wavelengths. A selection of phototransistors is shown in Figure 22-1 (left: Optek/TT Electronics OP506A with a broad spectral response centered around 850nm; center: Vishay TEKT5400S with a side-view lens; right: Vishay BPW17N).

Typically a phototransistor has two leads that connect internally with its collector and emitter (or source and drain, in the case of an FET). The base of the transistor (or the gate of an FET) responds to light and controls the flow of current between the leads.

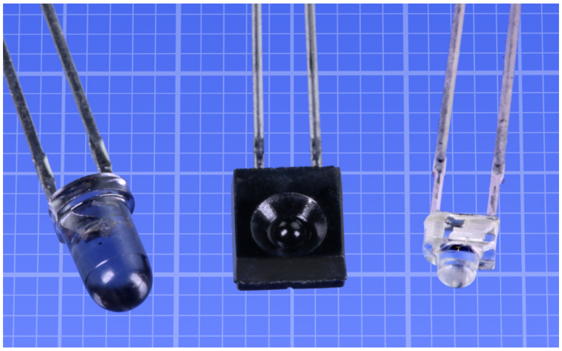

Figure 22-1 *A variety of phototransistors. The background grid is in millimeters. See text for details.*

In the absence of light, a bipolar phototransistor permits leakage between collector and emitter of 100nA or less. When exposed to light, it conducts up to 50mA. This alone differentiates it from a photodiode, which cannot pass much current.

Schematic Symbols

Symbols for a phototransistor are shown in Figure 22-2. They are functionally identical, with the exception of type C, where an additional connection to the base is included. Often (but not always), wavy arrows indicate infrared light.

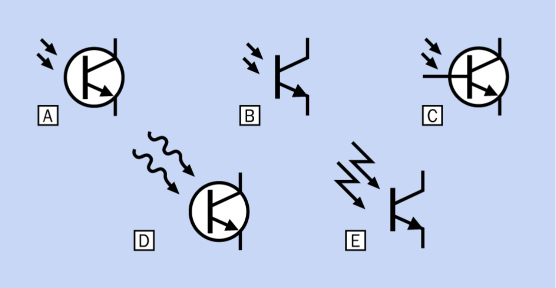

Figure 22-2 *Schematic symbols for a phototransistor.*

Applications

A phototransistor may be used for light measurement or as a light-sensitive switch.

Often the output from a phototransistor is used in conjunction with a *microcontroller* containing an analog-to-digital converter.

Where a clean on-or-off signal is required, a phototransistor can drive the input of a logic chip that contains a *Schmitt trigger*, or it can be processed by a **comparator**.

An *optocoupler* or **solid-state relay** (described in Volume 1) usually contains a phototransistor that is activated by an internal LED. Its purpose is to switch current while electrically isolating one section of a circuit from another.

How It Works

Like a photodiode, the phototransistor responds to light when the light generates electron-hole pairs in the semiconductor material. For a bipolar NPN phototransistor (the most common variant), the important region for pair generation is the reverse-biased collector-base interface. Photocurrent generated here acts as current injected into the base of an ordinary transistor, and permits a larger current to pass from the collector to the emitter.

The behavior of a phototransistor can be visualized as being similar to a photodiode controlling an ordinary bipolar transistor, as shown in Figure 22-3.

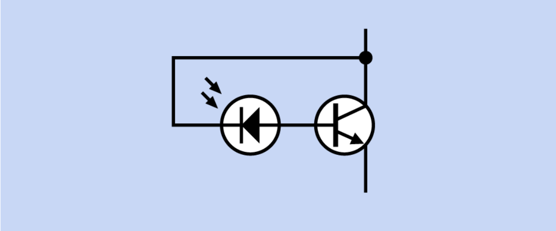

Figure 22-3 *A phototransistor is functionally similar to a photodiode controlling an ordinary transistor.*

Variants

While surface-mount variants are very widely available, through-hole packaging also remains common. When encapsulated like an LED indicator, a phototransistor gathers light from a relatively narrow angle. Variants with a flat surface are sensitive to light from almost any direction in front of the component.

Optional Base Connection

The base of a phototransistor is usually not accessible. However, some variants provide a base connection (or gate connection, in an FET) in addition to the collector and emitter (or the source and drain, in an FET). This third connection allows the application of a bias current, which can prevent low light levels from triggering the transistor.

Photodarlington

A photodarlington is a pair of bipolar transistors, the first being sensitive to light while the second acts as an amplifier for the first. This configuration is very similar to that of a *Darlington transistor* (described in Volume 1). The two-stage design makes it more sensitive to light than a regular phototransistor, but makes the response slower and less linear.

PhotoFET

A field-effect phototransistor is sometimes identified as a *photoFET*. They are relatively uncommon as separate components, but are used in optocouplers, because they have some interesting properties:

- Provided the applied voltage is low enough (less than 0.1V) the photoFET works as a controllable resistance—in contrast to bipolar transistors, where the current is controlled, and is relatively independent of the applied voltage.
- The FET transistor is symmetrical, functioning similarly for signals of either polarity. This makes an FET optocoupler suitable for AC signals.

Values

Abbreviations found in datasheets are included in the list below, with values in parentheses for an Osram SFH300FA infrared photodiode, which resembles a 5mm through-hole LED. It has a peak sensitivity of 880nm and appears black to the human eye, being opaque to wavelengths of light shorter than 700nm, the red end of the visible spectrum.

In Figure 22-4 the SFH300FA is shown beside the SFH300, which has the same peak sensitivity of 880nm but is encapsulated in an untinted module, allowing a sensitivity that tapers gradually to 450nm, in the green part of the visible spectrum. With the exception of their spectral range, the two phototransistors have identical specifications.

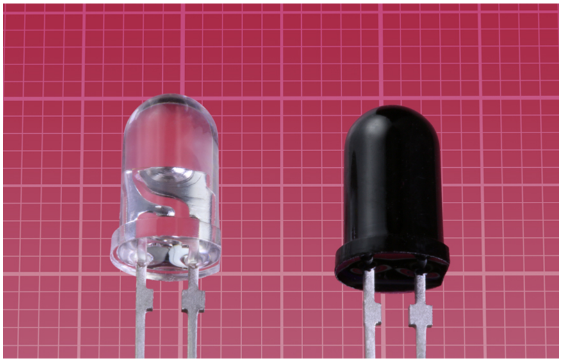

Figure 22-4 *Osram infrared phototransistors SFH300 (left) and SFH300FA (right). The background grid is in millimeters.*

- Maximum collector-emitter voltage: V_{CE} (35V)
- Maximum collector current: I_C (50mA)
- Maximum power dissipation: P_{TOT} (200mW)
- The *half angle* is measured from the axis of the photodiode to the angle at which sensitivity has dropped by 50% (plus or minus 25 degrees)

The angle of sensitivity depends on the geometry of the package. For phototransistors that resemble an LED indicator, with a rounded end that acts as a lens, typical values are plus-or-minus 20 degrees.

- Dark current (when the phototransistor receives no incident light): I_{CE0} (1nA)
- Wavelength of maximum sensitivity: $\lambda_{S\ MAX}$ (880nm)

Infrared phototransistors exist with a variety of peak wavelengths. They are designed to function in conjunction with an LED that has a matching wavelength.

- Response speed is the rise and fall time of photocurrent: t_r and t_f (10 µs)

Behavior Compared to Other Light Sensors

An extended list of comparisons between a photoresistor and a phototransistor will be found in the entry on **photoresistors**. See "Comparisons with a Phototransistor".

A photodiode has a close-to-linear electrical response over a much wider range of intensities of light than a phototransistor. Consequently, photodiodes tend to be the component of choice in applications where measurement of light must be wide-ranging and precise.

Photodiodes can pass less current than a phototransistor, but also tend to draw less current, making them appropriate in battery-powered devices that must draw as little current as possible.

Rise and fall speeds are important for high-speed measurement, signaling, or data transfer. The rise and fall time of a typical phototransistor can be 1,000 times slower than that of a photodiode. See "Values" in Chapter 21 for a comparison. Note also that the dark current of a phototransistor is much higher than that for a photodiode.

The ability of a phototransistor to sink 20mA to 50mA at its output is useful where it will be connected to a component that has relatively low impedance. For instance, a phototransistor can drive a piezoelectric audio transducer directly, or an LED indicator.

Unlike a photodiode, a phototransistor is a solid-state switch. Its saturation voltage (listed in datasheets as $V_{SE(SAT)}$ as described above) is the voltage drop between collector and emitter, and seldom exceeds 0.5V.

Binning

Small variations that occur during the fabrication process can cause inconsistency in the performance of phototransistors that share the same part number. To provide a more consistent response, manufacturers use *binning*, meaning that units sharing the same bin number are likely to share a tighter tolerance. (The same concept is used to minimize variations in **LED area lighting**, described in Volume 1.)

Datasheets will provide information on the availability and meaning of bin numbers, if available.

Bins with higher photocurrents typically have longer response times.

How to Use It

Most phototransistors are bipolar devices with an *open collector output*. That is, the collector of the transistor is accessible from one of the two leads, making it "open" to being used. See Figure A-4 for general instructions on using an open-collector output. A summary relating to phototransistors is included here.

The schematic in Figure 22-5 shows the basic concept. The resistor is referred to as a *pullup resistor*. When the phototransistor is receiving very little light, its effective resistance is high. Consequently almost all the current flowing through the pullup resistor will go to any device attached to the output, and the output voltage will seem to be "high."

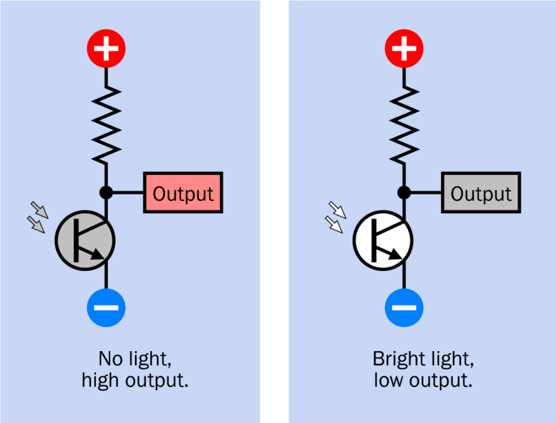

Figure 22-5 *How an open-collector output works.*

If the phototransistor is exposed to a significant light source, the effective resistance between the collector and emitter drops dramatically, and the phototransistor will sink current to ground. Consequently, the output will seem to be "low."

The pullup resistor is necessary between the power source and the collector pin to protect the phototransistor from sinking excessive current when it is exposed to light. The ideal value of the resistor will depend partly on the impedance of any device attached to the output.

In this scenario, exposure to light causes low output whereas darkness causes high output. What if we wish to have it the other way around?

The protective resistor can be moved to the emitter pin, where it becomes a pulldown resistor. It will still protect the phototransistor from passing excessive current, so long as the output is connected to a high impedance. The output is taken from the emitter pin, and will transition from low to high when the component is exposed to light. This is illustrated in Figure 22-6.

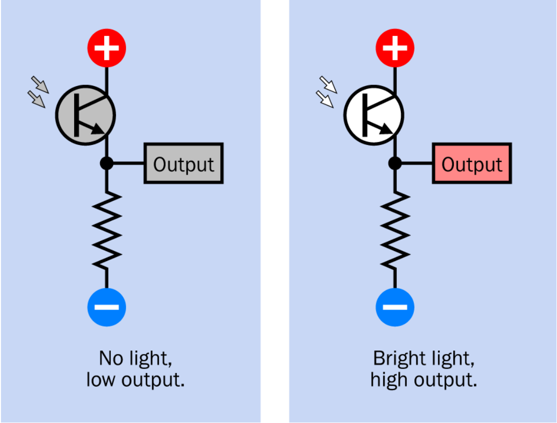

Figure 22-6 *Moving the resistor and taking an output from the emitter pin inverts the behavior of a phototransistor. Any device attached to the output must have a relatively high impedance to protect the phototransistor from excessive current.*

Output Calculation

Using an open-collector output, the photocurrent is almost independent of the applied voltage V_{CE}, provided that the voltage is higher than the saturation voltage $V_{CE(SAT)}$, which is typically between 0.4V and 0.5V.

If the pullup resistor has value R, the voltage across it is

$$U = R * I_P$$

where I_P is the photocurrent passed by the phototransistor.

When choosing R, one should consider the range of currents expected in the light conditions that will be measured, and the voltage range suitable for the next stage of the circuit. 10K is a reasonable starting point (e.g., when measuring light intensity with a microcontroller's analog input). The resistor value can be reduced from there, if necessary.

The value of the pullup resistor must also be chosen to restrict current within the limits specified by the phototransistor datasheet. A value for the resistor guaranteed to be safe is

```
R = V / I~MAX~
```

where V is the supply voltage and I_{MAX} is the maximum allowed current. With this value, the resistor limits the current to the highest allowed value, even if the phototransistor is brightly illuminated and assumed to conduct perfectly.

When V = 5V and I_{MAX} = 15mA, R should be at least 330 ohms.

What Can Go Wrong

Visual Classification Errors

Phototransistors can be visually similar to LEDs and photodiodes. They are easily confused, as neither type of component is usually marked with any identification number. The entry for **photodiodes** describes a system for distinguishing the three types of components. See "What Can Go Wrong" in Chapter 21.

Output Out of Range

The output voltage from a phototransistor will depend on the intensity of incident light, the value of any pullup resistor that is used, and the supply voltage. While a circuit is being developed, the light range may seem predictable, but in actual use the output range may not fall within expectations.

NTC thermistor

23

PTC thermistors, in which the resistance increases as the temperature increases, have a separate entry. See Chapter 24.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Semiconductor temperature sensors and **thermocouples** each have their own entries.

Infrared temperature sensors and **passive infrared** motion sensors have their own entries. They are *noncontact* temperature sensors that respond to infrared radiation.

OTHER RELATED COMPONENTS

- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **passive infrared** motion sensor (see Chapter 4)
- **semiconductor** temperature sensor (see Chapter 27)
- **thermocouple** (see Chapter 25)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

An NTC thermistor is the most common type of discrete-component temperature sensor, and is usually the most affordable. Its resistance diminishes as its temperature increases. This behavior is referred to as a *negative temperature coefficient*, which is the source of the acronym NTC.

This is a simple, passive component that is not polarized. It requires no separate power supply, but an external device must pass a small AC or DC current through it to determine its resistance. This is known as an *excitation current*.

Schematic Symbols

Schematic symbols for a thermistor are shown in Figure 23-1. Those in the top row may still be found in the United States, but are being replaced by the European variants in the second row. The addition of -t° to the symbol indicates an NTC type of thermistor, while +t° indicates that it is the PTC type, with a positive coefficient (see Chapter 24). If no indication is shown, the thermistor is likely to be the NTC type.

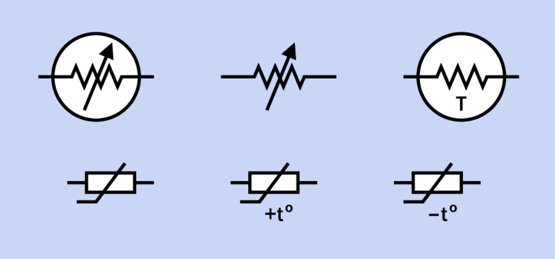

Figure 23-1 *Schematic symbols representing a thermistor. Letter t preceded by a plus or minus sign indicates whether the thermistor is the PTC or NTC type, respectively.*

Applications

Thermistors monitor temperature in air-conditioning systems, clothes washers, refrigerators, pool and spa controls, dishwashers, toasters, and other domestic devices. They are used in laser printers, 3D printers, industrial process controls, and medical equipment.

As many as 20 thermistors may be found in a modern automobile, measuring temperature in locations ranging from the transmission to the ambient air in the passenger compartment.

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors, which measure temperature by making contact with the source, are divided into five main categories, each of which has a separate entry. For convenience, these categories are listed in a comparative summary at the end of this entry. See "Addendum: Comparison of Temperature Sensors".

How an NTC Thermistor Works

Although the term *thermistor* suggests that it is a thermally sensitive resistor, in fact an NTC thermistor is a semiconductor.

Some metal oxides, such as ferric oxide or nickel oxide, become n-type semiconductors when they are treated with dopants. The exact mix is a proprietary secret of each manufacturer. Raising the temperature of this kind of material increases the number of charge carriers in it, promoting electron mobility and thus lowering its effective resistance.

To create a thermistor, the metal oxide mix is heated until it melts and turns into a ceramic. Typically a thin sheet is cut into small pieces for individual sensors. After two leads are connected, the assembly is dipped into epoxy or encapsulated in glass. The most common packages consist of a glass bead, surface-mount chip, or ceramic disc.

Figure 23-2 shows three NTC thermistors. At left is a Murata NXFT15XH103FA2B100 approximately 1mm in diameter, with a reference resistance of 10K and a tolerance of plus-or-minus 1%. At center is a Vishay NTCALUG03A103GC rated 10K at 2%, fitted with a mini-lug. At right is a TDK B57164K153K rated 15K and 3%.

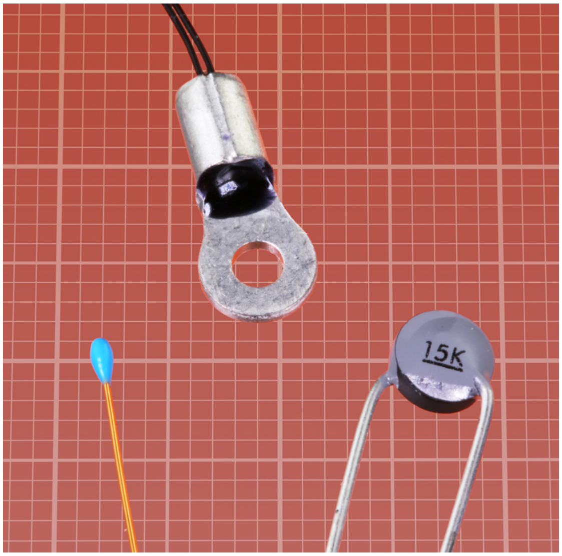

Figure 23-2 *Sample NTC thermistors. See text for details. The background grid is in millimeters.*

Output Conversion for Temperature Sensing

Ideally, the electrical behavior of a temperature sensor should be a linear function of temperature. Thermistors fail in this respect, as their resistance is an approximately inverse exponential function. This is illustrated in Figure 23-3, where the measured resistance of a thermistor rated for 5K at 25 degrees is plotted against temperatures from 0 degrees to 120 degrees Celsius.

- In many datasheets, graphs of this kind may appear flatter because they are customarily plotted against a vertical logarithmic scale.

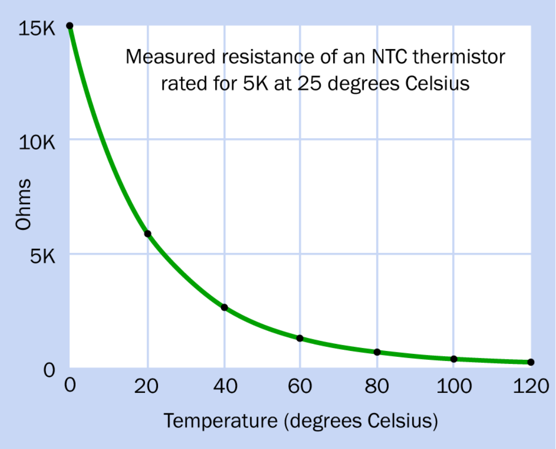

Figure 23-3 *Resistance of a thermistor from 0 to 120 degrees Celsius.*

To monitor the resistance of a thermistor, it can be placed in a simple voltage divider as shown in Figure 23-4, where the fluctuating resistance of the component creates a fluctuating voltage at point A.

- The voltage can be used as an input to a microcontroller that contains an analog-to-digital converter. Alternatively it can be connected directly to a solid-state relay, or amplified with an op-amp, or can be passed through a comparator to create an adjustable switching threshold.

Although this circuit is a voltage divider, it is also known as a *half bridge*, as it is half of a Wheatstone bridge.

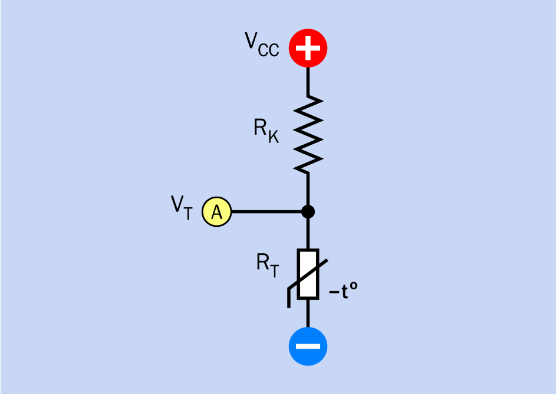

Figure 23-4 *A half-bridge circuit for determining the resistance of a thermistor.*

If V_{CC} is the supply voltage, V_T is the measured voltage at point A, R_T is the resistance of the thermistor, and R_K is the constant value of the series resistor, the basic formula for a voltage divider looks like this:

$$V_T = V_{CC} * (R_T / (R_T + R_K))$$

By transposing terms, a formula can be derived to obtain a value for R_T from the measured voltage and the value of R_K:

$$R_T = (R_K * V_T) / (V_{CC} - V_T)$$

Choosing a Series Resistor

The value for R_K in the formula should be chosen to provide a reasonably wide response over the range of temperatures for which the thermistor is likely to be used. To calculate R_K, another formula must be applied. If R_{MIN} is the resistance of the thermistor at the lowest likely temperature, and R_{MAX} is its resistance at the highest likely temperature:

$$R_k = \sqrt{R_{min} * R_{max}}$$

(This is the same formula as suggested in Figure 20-5 to find the value of a series resistor for use with a photoresistor.)

Wheatstone Bridge Circuit

The half-bridge circuit has the disadvantage that it does not compensate for nonlinearity of a thermistor. Voltage values will change rapidly at the low end of the temperature range, but will change more slowly at the high end, requiring an analog-to-digital converter with a high degree of accuracy to distinguish one voltage value from the next.

A full Wheatstone bridge circuit has a nonlinear output that compensates, somewhat, for the inverse nonlinearity of the thermistor. Referring to the circuit shown in Figure 23-5, the three resistors R_K are chosen using the formula above.

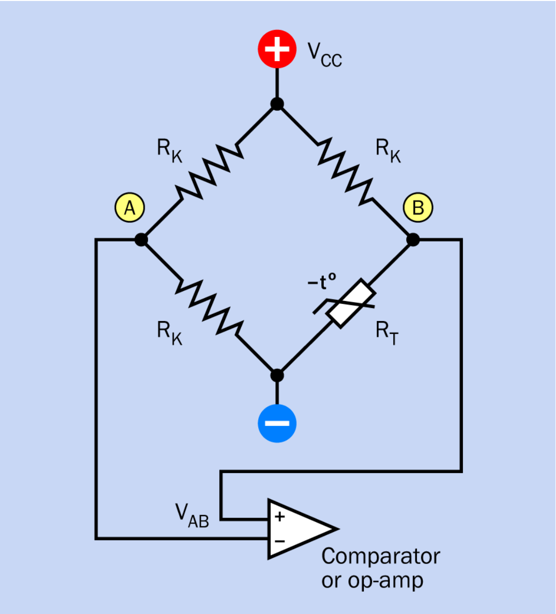

Figure 23-5 *A thermistor may be placed in a full Wheatstone bridge circuit. Outputs A and B are often connected with the two inputs of an op-amp or comparator.*

A standard formula provides the relationship between R_T, the resistance of the thermocouple; V_{CC}, the supply voltage; R_K, the fixed resistances; and V_{AB}, the output voltage measured between points A and B:

$$V_{AB} = (V_{CC} / 2) * (R_T - R_K) / (R_T + R_K)$$

From this formula, a version can be derived to calculate R_T by measuring the output voltage, V_A:

$$R_T = R_K * (V_{CC} + (2*V_{AB})) / (V_{CC} - (2*V_{AB}))$$

- The polarity of V_{AB} is reversible, depending on whether R_T is greater or less than R_K. To accommodate this, A and B can be connected to the two inputs of a comparator or op-amp.

Deriving the Temperature Value

After the resistance of the thermistor has been calculated, it can be converted to a temperature value. The datasheet for a thermistor will usually provide a table showing temperature values tabulated against resistance values, so that a lookup table can be created in a microcontroller program.

Alternatively, a datasheet usually includes constants that can be inserted in a resistance-to-temperature conversion equation, but this is nontrivial and requires natural logarithms, which may not be available in a language implemented on a particular microcontroller.

Inrush Current Limiter

NTC thermistors with appropriate characteristics can be used to limit the inrush of current that tends to occur when a circuit is switched on and large capacitors in the power supply charge very quickly.

An *inrush current limiter* is also known as a *surge limiter*, or may be referred to by its acronym, *ICL*. It is an NTC thermistor whose initial resistance diminishes rapidly as its temperature increases.

While NTC thermistors are the type most often used for inrush limiting, PTC thermistors can serve this purpose if wired differently. See "PTC Inrush Current Limiting". The remainder of this discussion refers only to NTC current limiters.

A suitable NTC thermistor can be placed as shown in the simplified schematic in Figure 23-6, where a rectified AC source is connected with a DC-to-DC converter, and a large smoothing capacitor is used. Initially the thermistor has resistance that is sufficient to limit current and generate heat. But the rise in temperature causes the resistance of the thermistor to fall. Eventually it reaches a steady state where it remains sufficiently warm to maintain a low resistance that imposes a negligible load on the circuit.

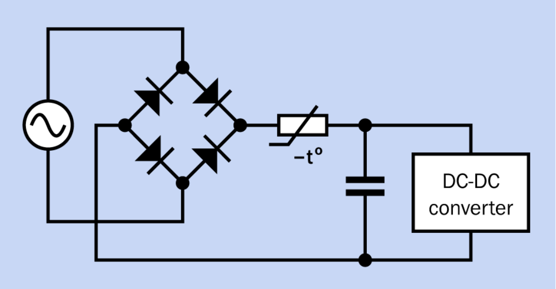

Figure 23-6 *Placement of an NTC thermistor that is designed for inrush current limiting.*

In thermistors that are used for temperature measurement, self-heating is an undesirable attribute. By contrast, an inrush current limiter depends on self-heating to perform its function.

The TDK B57237S509M inrush limiter, shown in Figure 23-7, is rated for 5A and has an initial resistance of 5 ohms at 25 degrees Celsius while not passing current. When tested with a 2,800μF capacitor at 110VAC, its resistance drops to a minimum of 0.125 ohms at 5A. The relationship between current and resistance is shown in Figure 23-8.

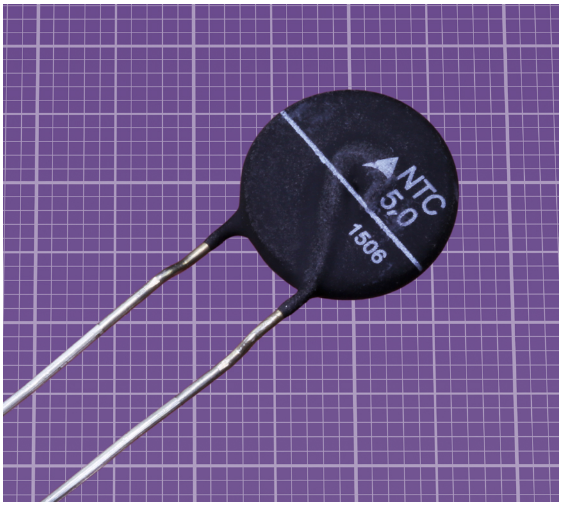

Figure 23-7 *A TDK B57237S509M NTC thermistor designed as an inrush current limiter, rated 5 ohms at 5 amps. The background grid is in millimeters.*

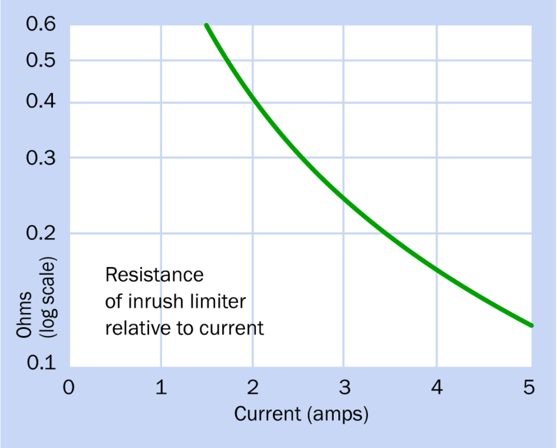

Figure 23-8 *This graph shows the relationship between resistance and current in a TDK B57237S509M inrush limiter.*

Restart

If a protected device is switched off momentarily and is then switched on again, the thermistor cannot provide protection, as it has not had time to cool down and regain its resistance. However, during the 30 seconds to 2 minutes required for heat in the thermistor to dissipate, smoothing capacitors are unlikely to lose much of their charge. Therefore, if the device is restarted, an inrush of current should not occur.

Thermistor Values

Datasheets for thermistors may be more complex and cryptic than for many components.

When examining a datasheet, first check to see if the thermistor is described as being suitable for temperature measurement or inrush current limiting. A component designed for temperature measurement will not survive inrush current, while one designed for inrush current limiting will have a very low resistance making it unsuitable for temperature measurement.

Time and Temperature

In most datasheets, lowercase letter t is used for values relating to time, whereas an uppercase T is used for values relating to temperature. Unfortunately, T may also be used as an abbreviation for "thermistor."

Resistance and Response

Letter R often means resistance, but may indicate response time, depending on the context in which it is used. For example, R_T is the resistance of a thermistor, and t_R is a response time.

Kilohms and Kelvin

Letter K may be used to represent temperature in degrees Kelvin, 0 degrees on the Celsius scale being approximately 273 degrees Kelvin. However, letter K is also used to represent thousands of ohms, sometimes in the same datasheet. In both instances, K is capitalized.

Reference Temperature

This is the temperature at which many attributes of the component are measured, such as its temperature coefficient and resistance. Usually the reference temperature is 25 degrees Celsius, but in some cases it may be 0 degrees, and other values are occasionally used. The term is abbreviated T_{REF}.

Reference Resistance

The reference resistance for a thermistor (sometimes described as its nominal resistance) may be referred to as R_R, and is the resistance at the reference temperature. It may be referred to as the "R value," but in thumbnail product descriptions it can be cited simply as "Resistance."

In datasheets, R25 or R_{25} is the resistance at 25 degrees Celsius. If this is the reference temperature, R_R and R_{25} will be the same.

Dissipation Constant

DC is the *power dissipation constant*, a ratio normally expressed as milliwatts per degree Celsius (written as mW/°C). This is a measurement of how much thermal power the thermistor can transfer to the environment for a 1 degree increase in temperature.

Temperature Coefficient

TC may be used as an acronym for the *temperature coefficient*, which represents the sensitivity of the thermistor. (Sometimes TCR is used instead of TC, the letter R denoting resistance. The two acronyms both mean the same thing.) The value is the percentage change in resistance for each change in temperature of 1 degree Celsius. Thus, if the resistance of a thermistor drops from 800 ohms to 768 ohms when the temperature increases from 28 to 29 degrees Celsius, TC = -4%. For NTC thermistors, which have a resistance that decreases when temperature increases, the temperature coefficient is negative. However, the minus sign may be omitted.

The coefficient may be expressed in parts per million (abbreviated ppm) instead of as a percentage. To convert parts per million to a percentage, divide by 10,000. Thus, a figure of 50,000ppm is equivalent to 5%.

Thermal Time Constant

Unfortunately TC is also used to represent the thermal *time constant*. If T_D is the temperature difference between the thermistor's initial temperature and a new, higher ambient temperature in which it finds itself, TC is the time it takes for the thermistor to add 63.2% of T_D to its current temperature. TC is expressed in seconds, and is defined without power being applied to the thermistor. A low thermal time constant is characteristic of a physically small thermistor that acquires heat rapidly. (TC is very similar to the concept of a time constant for a capacitor acquiring charge. See the entry on **capacitors** in Volume 1.)

Tolerance

The *tolerance* of a thermistor is a measure of its accuracy, usually at 25 degrees Celsius, unless a range of temperatures is stated. A thermistor rated for 5K, with a tolerance of plus-or-minus 1% at 25 degrees, may be found to have an actual resistance ranging from 4,950 to 5,050 ohms at that temperature. Some thermistors have a tolerance of plus-or-minus 20%. A tolerance better than plus-or-minus 1% is relatively rare.

Temperature Range

The *working temperature range* of any thermistor that uses silicon dioxide is usually between about -50 and +150 degrees Celsius (slightly wider for versions encapsulated in glass, and slightly narrower if accuracy is important).

Switching Current

For a thermistor with a nonlinear response, the switching current is the approximate current that forces a sharp transition in resistance. It is represented by I_S.

Power Limitations

Operating current is the maximum current recommended to avoid self-heating. The *power rating* is the maximum allowed power (usually 100mW to 200mW).

Interchangeability

To measure temperature reliably, two thermistors of the same type, from the same manufacturer, should display the same characteristics. This is known as *interchangeability*. A value of plus-or-minus 0.2 degrees Celsius is common for a modern thermistor, but is not often mentioned in a datasheet.

What Can Go Wrong

Self-Heating

Self-heating can affect the accuracy of an NTC thermistor that is used for temperature measurement. To get accurate temperature readings, keep the current as small as possible. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heat Dissipation

Where a thermistor is used for inrush current limiting, it will create some heat during the whole time that a device is switched on. If insufficient air space is allowed between the thermistor and other components, they may be affected.

Lack of Heat

An NTC thermistor will sometimes fail as an inrush current limiter. In very cold climates, it may never become warm enough for its resistance to drop to an acceptable level. Conversely, in a very hot location (such as close proximity to a hot-water pump) it may not get cool enough to provide adequate initial protection.

Addendum: Comparison of Temperature Sensors

A chart illustrating the five main types of contact sensors, and their variants, is shown in Figure 23-9.

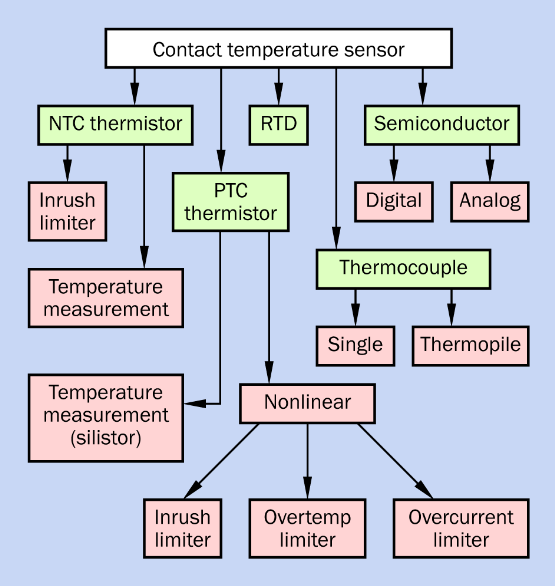

Figure 23-9 *Five types of contact temperature sensors (green boxes) and the variants (red).*

NTC Thermistor

The electrical resistance of an **NTC thermistor** diminishes as its temperature increases. Thus, it has a *negative temperature coefficient*, which is the source of the acronym NTC.

An NTC thermistor is traditionally used where low cost and simplicity are desirable and a relatively limited temperature range is acceptable (often -50 to +150 degrees Celsius). It has the advantage of familiarity, having existed in its present form for many decades. It remains the lowest-cost option among the various types of temperature sensors, and can be connected directly with an external device such as a solid-state relay, in which case no microcontroller is necessary.

PTC Thermistor

The sensing element for a positive-coefficient thermistor is a polycrystalline compound that increases in resistance very rapidly above a threshold temperature. This makes it suitable for blocking a high current to prevent circuit overload.

A *silicon temperature sensor*, sometimes called a *silistor*, can be considered as a PTC thermistor, in that it is a resistive component with a positive temperature coefficient. Its sensing element is etched into silicon.

PTC thermistors are passive, nonpolarized components with two leads or solder pads. For more information about them, see Chapter 24.

Thermocouple

This sensor consists of two wires made from different metals, joined at one end. The differing thermoelectric properties of the wires creates a very small voltage between their free ends. Thermocouples have the widest range of any contact sensor. They are simple, robust, and free from self-heating effects, as they consume no power. Their response is rapid, but very nonlinear, and their sensitivity is limited. They are used in industry and in laboratories, often plugged into a panel meter that combines a digital temperature display with hardware to decode the signal from the type of thermocouple being used.

For more information about thermocuples, see Chapter 25.

Resistance Temperature Detector

Often known by its acronym **RTD**, and sometimes referred to as a *Resistive Temperature Device*, it commonly uses a sensing element fabricated from pure platinum, nickel, or copper. The element may consist of wire wound around a core, or a very thin film deposited on an insulating substrate.

An RTD has a positive temperature coefficient, as its resistance increases while its temperature increases. It is very accurate and stable, providing an almost linear output, especially near the center of its range. However, its sensitivity is often one-tenth of that of an NTC thermistor.

Like a thermistor or a thermocouple, an RTD is a passive device, able to operate at a wide range of voltages and requiring no power supply. It is nonpolarized, with two leads or solder pads.

For more information about resistance temperature detectors, see Chapter 26.

Semiconductor Temperature Sensor

This is a chip-based sensor that requires no additional components to linearize its output, as linearization is performed in the chip.

The temperature range is similar to that of an NTC thermistor, but the output is a variable voltage with a positive temperature coefficient of about 20mV per degree Celsius, supplied by a built-in op-amp. Response time is 4 to 60 seconds.

This type of sensor requires a power supply, typically of 5VDC or less. It does not have to be calibrated before use, as it is trimmed during the production process to achieve accuracy that can be better than that of a thermistor. Manufacturers may claim plus-or-minus 0.15 degrees over the whole temperature range, which is usually -50 degrees Celsius to +150 degrees, but may be less for variants in which accuracy is more important.

The linear analog output is very convenient for use with a microntroller that has an analog-to-digital converter, and the relatively low cost makes this type of sensor increasingly competitive with thermistors.

An analog-to-digital converter may be included on the sensor chip, in which case it is often described as a *digital temperature sensor* or *digital thermometer*, with an output in degrees Celsius (or, sometimes, Fahrenheit) accessible via I2C or SPI bus. For additional details about protocols such as I2C and SPI, see Appendix A.

A *digital thermostat* or *thermostatic switch* is a semiconductor temperature sensor with a binary output that transitions from logic-high to logic-low (or vice-versa) if the temperature goes above a maximum or below a minimum level. The level can be programmed into the chip.

Semiconductor temperature sensors are identified with a variety of other names. For more information, see Chapter 27.

PTC thermistor

24

A *silistor*, or *silicon-based thermistor*, is included in this entry as a form of **PTC thermistor**.

A *resettable fuse* is not quite the same as a PTC thermistor. For more information, see the entry on **fuses** in Volume 1.

NTC thermistors, in which the resistance decreases as the temperature increases, have a separate entry. See Chapter 23.

A *resistance temperature detector* or **RTD** has a resistance that increases as its temperature increases, but it is not usually classified as a thermistor, because its sensing element is fabricated differently. Its entry will be found at Chapter 26.

Infrared temperature sensors, **semiconductor temperature** sensors, and **thermocouples** each have their own entries.

OTHER RELATED COMPONENTS

- **infrared temperature** sensor (see Chapter 28)
- **semiconductor** temperature sensor (see Chapter 27)
- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

The electrical resistance of a PTC thermistor increases as its temperature increases. Variants can measure temperature or can protect circuits by detecting excessive heat or current.

Because a PTC thermistor is a resistive sensor, it has no polarity. Current may flow through it in either direction, or AC may be used.

Schematic Symbols

The schematic symbol for a PTC thermistor is very similar to the symbol for an NTC thermistor. See Figure 23-1.

Comparison of Temperature Sensors

In this Encyclopedia, contact temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for

NTC thermistors. See "Addendum: Comparison of Temperature Sensors". Also see Figure 23-9.

PTC Overview

PTC thermistors can be subdivided into two groups:

- *Linear*, with a chip-sized silicon-based sensing element. They are sometimes referred to as *silistors*. The component has a very linear response and is used for temperature measurement. It may be connected directly to a microcontroller.

- *Nonlinear*, mostly using a sensing element containing barium titanate in a polycrystalline compound that increases in resistance very sharply above a threshold temperature. This type of sensor may be described as a *switching thermistor*, because its nonlinear output can activate a switching device.

The sensing elements in positive-coefficient thermistors are different in principle from the element in an NTC thermistor.

Nonlinear thermistors are used in two different ways:

Externally heated

The thermistor responds to ambient heat or to the temperature of a device to which it is attached. It can be used to protect a circuit or a motor from overheating. Current through the thermistor is minimized to avoid self-heating.

Internally heated

The thermistor responds to its own temperature caused by current passing through it. It can activate a warning signal or shut down equipment in the event of a short circuit. It can also control current for starting a motor or a fluorescent tube, and is sometimes used as a source to create localized heat.

Silistor for Temperature Measurement

A silicon-based PTC thermistor, sometimes known as a *silistor*, provides a highly desirable, almost linear relationship between temperature and resistance. A popular example is the KTY81 series from NXP, a sample of which is shown in Figure 24-1.

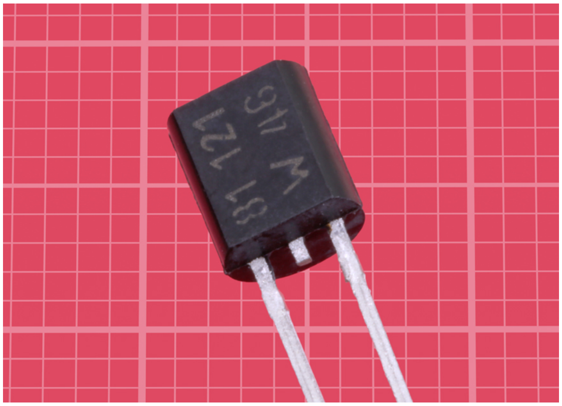

Figure 24-1 *A KTY81 thermistor from NXP. The background grid is in millimeters. Note the amputated center lead.*

The response of this thermistor is shown in Figure 24-2.

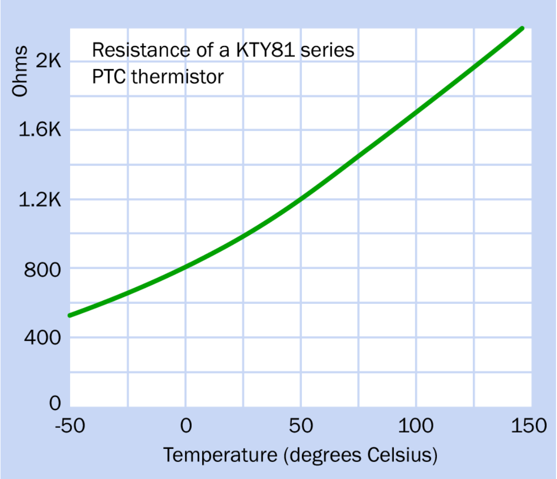

Figure 24-2 *Resistance of the KTY81 thermistor in response to temperature.*

Note that this graph is plotted with a linear vertical scale, unlike the performance curves for many thermistors that are plotted with a log scale. The log scale tends to make a response curve look flatter.

The sensor is a silicon chip designed on the "spreading resistance principle," in which current fans out from a metal contact through a thin layer of silicon to a metallized bottom plane. This effect progresses less actively as the temperature increases. Although the result is partly dependent on polarity, a second metal contact is biased in the opposite direction, and when the two active regions of the chip are wired in series, the result is a component that has no polarity.

- The almost-linear output of this type of sensor makes it easy to use with a microcontroller that has a built in analog-to-digital converter.

Tolerance ranges from plus-or-minus 1% to 5%, depending on the temperature. Variants have a typical reference resistance of 1K or 2K. The temperature coefficient is commonly about 1%, which is considerably lower than that of a typical NTC thermistor, where 4% is common.

- Guidance on reading thermistor datasheets will be found in the entry describing NTC thermistors. See "Thermistor Values".

For correct operation, a typical silistor requires a current ranging from around 0.1mA to 1mA.

- The lower sensitivity and slightly higher price of a PTC temperature-measurement thermistor, compared with an NTC thermistor, may explain why the NTC type seems to remain more popular, with many more variants available. In addition, the NTC type is much more tolerant of variations in current.
- Silistors continue to find some automotive applications, measuring oil temperature, transmission temperature, and climate control, among other parameters.

As a simple strategy to determine its resistance, a series resistor can be used with a PTC sensor to create a voltage divider. The circuit is identical to that used for NTC thermistors. See "Output Conversion for Temperature Sensing".

RTDs

A *resistance temperature detector* or **RTD** is sometimes classified as a PTC thermistor. However, it has a different type of pure-metal sensing element, much lower sensitivity, and is discussed in a separate section of this Encyclopedia. See Chapter 26.

Nonlinear PTC Thermistors

Over-Temperature Protection

This type of nonlinear thermistor is externally heated, but has a switching function. If it is incorporated among other components on a circuit board, its output can be used to activate a warning signal, or can trigger a relay to shut down the circuit until the temperature subsides. This is of special interest for battery chargers where excessive heat can often be a problem, but is also useful in electronic devices generally.

To avoid the possibility of self-heating, current passing through the thermistor must be minimized to a few milliamps.

Some thermistors in the Vishay PTCSL series will make a transition at a temperature as low as 70 degrees Celsius. Others will be triggered by temperatures above 100 degrees. A typical response curve is shown in Figure 24-3, where resistance rises from 100 ohms at 25 degrees to around 1K at the transitional reference temperature of 90 degrees, and reaches at least 4K at 105 degrees.

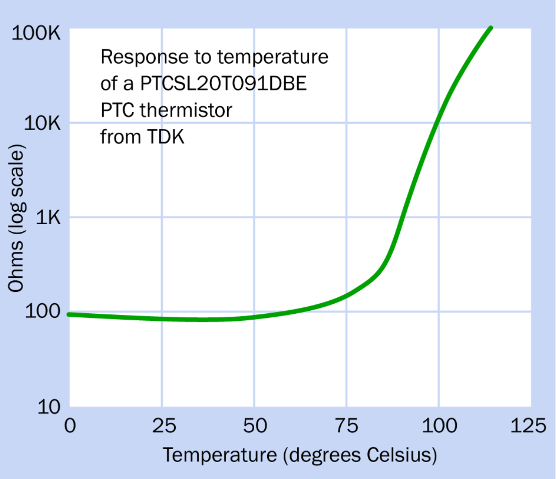

Figure 24-3 *The relationship of resistance to heat in an over-temperature protection thermistor.*

To respond to this transition, the manufacturer recommends a Wheatstone bridge circuit with its outputs connected to a comparator, as suggested for an NTC thermistor in Figure 23-5. The comparator can then activate a signal or a relay.

A picture of the PTCSL20T091DBE thermistor appears in Figure 24-4.

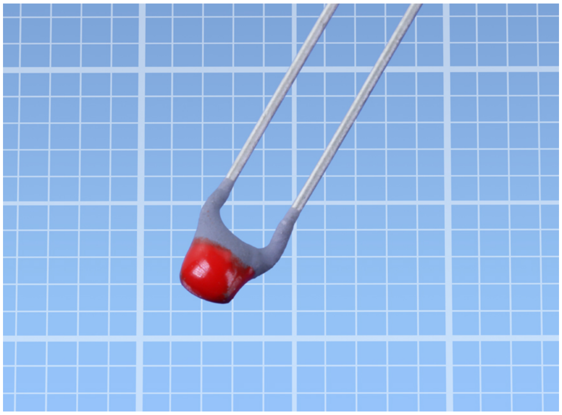

Figure 24-4 *A thermistor in the PTCSL range from TDK. It is color coded using a proprietary scheme by the manufacturer to indicate a reference temperature of 90 degrees Celsius. The background grid is in millimeters.*

This type of thermistor can tolerate a maximum of 30V (AC or DC).

Over-Current Protection

This type of nonlinear thermistor is a substitute for a fuse, as it responds to internal heat created by current passing through it. If the flow of current is excessive, the resistance of the thermistor increases, throttling the flow. When the over-current problem is resolved, the thermistor returns to its normal state. Whereas a fuse must be placed in a location allowing replacement, the thermistor is unharmed by its transition and does not have to be replaced.

Over-current may occur as a result of failure of other components, such as a rectifier diode or a capacitor, or can occur in situations such as a DC motor locking up.

The B598 series from TDK can tolerate voltages over 240V, AC or DC. They typically respond when currents exceed 100mA to 1A, depending on the specific component (a few fall outside that range), and many can withstand 1A to 7A. The B59810C0130A070 pictured in Figure 24-5 is switched by 980mA, can tolerate as much as 7A, and has a reference resistance of 3.5 ohms, rising above 10K when excessive current causes sufficient heat.

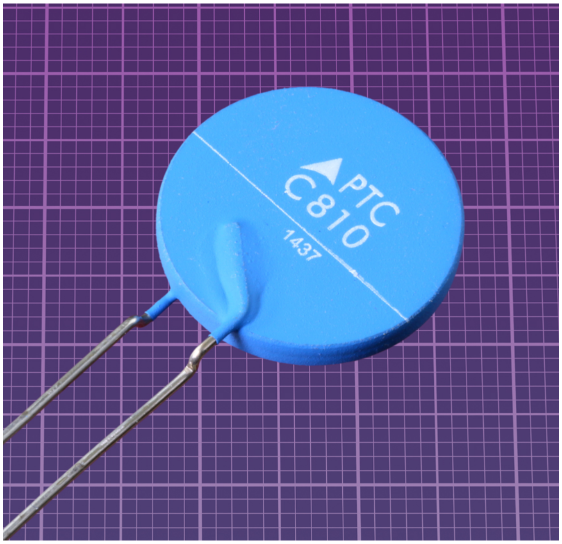

Figure 24-5 *A large over-current protection PTC thermistor. The background grid is in millimeters.*

An over-current thermistor of this type remains wired into the power supply for a device on a permanent basis. Its reference resistance will generate some heat, which is why this type of component is usually restricted to applications where the triggering current is below 1A.

The Murata PTGL07BD220N3B51B0 pictured in Figure 24-6 provides over-current protection with a reference resistance of 22 ohms, has a trip current of 200mA, and tolerates a maximum of 1.5A.

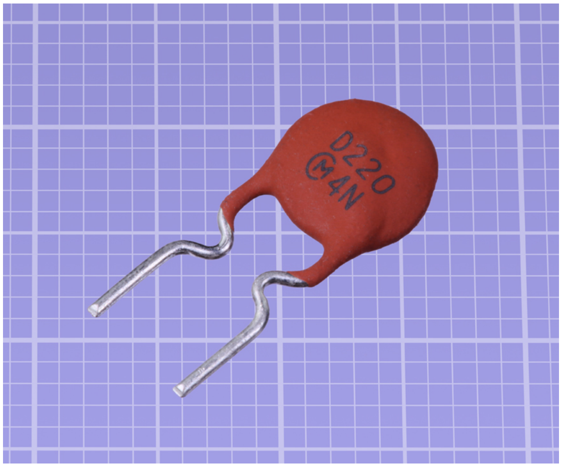

Figure 24-6 *An over-current protection PTC thermistor with a trip current of 200mA. The background grid is in millimeters.*

PTC Inrush Current Limiting

This nonlinear thermistor responds to internal heat caused by an inrush of current when power to a device is switched on. The inrush occurs when current flows rapidly into smoothing capacitors, charging them very rapidly. This can overload a power supply and shorten its life expectancy.

NTC thermistors are traditionally used as inrush limiters. The initially high resistance of this type of component blocks the surge in current, but as heating occurs, the resistance of the NTC thermistor drops rapidly. It remains in the circuit, imposing a relatively small load while the device functions normally. For more details of this application, see "Inrush Current Limiter" in the entry discussing NTC thermistors.

However, an NTC thermistor used in this way will waste some power. Suppose a supply of 120VAC is being used. If the power consumption of a device is 1,000W, the current will be approximately 8A. An NTC thermistor that has a resistance of 0.2 ohms while running warm will introduce a voltage drop of approximately 1.6V, consuming about 13W. This power loss will be greater in applications where the current is even higher—for example, in an electric vehicle recharging station.

To eliminate the loss, a timed bypass relay can be added around the thermistor. The relay closes automatically after a short interval, eliminating the power loss. This is known as *active inrush current limiting*.

However, in this arrangement, an ordinary resistor could be used instead. But in that case, why not use a PTC thermistor that has a reference (cold) resistance of 50 ohms or more? This not only limits the inrush current, but provides additional protection. If a smoothing capacitor in the circuit suffers a short circuit, or if the bypass relay fails to close, excess current passing through the PTC thermistor quickly raises its resistance, protecting the rest of the circuit.

The B5910 series of PTC thermistors from TDK is designed for inrush current limiting. They are packaged in a flame-retardant phenolic resin plastic case, as shown in Figure 24-7. The B59105J0130A020 has a reference resistance of 22 ohms that rises quickly beyond 10K when the temperature exceeds 120 degrees Celsius, as shown in Figure 24-8. This type of component is robust enough to withstand a complete short circuit across a 220-volt supply.

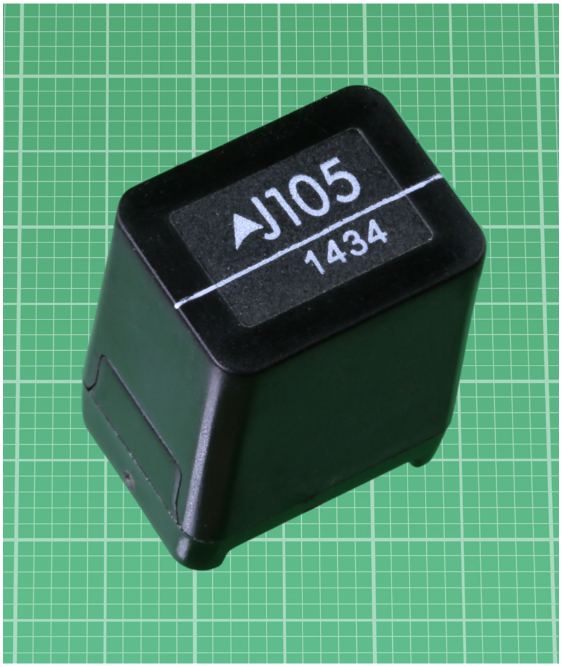

Figure 24-7 *This inrush current-limiting PTC thermistor by TDK is packaged in a flame-retardant case. The background grid is in millimeters.*

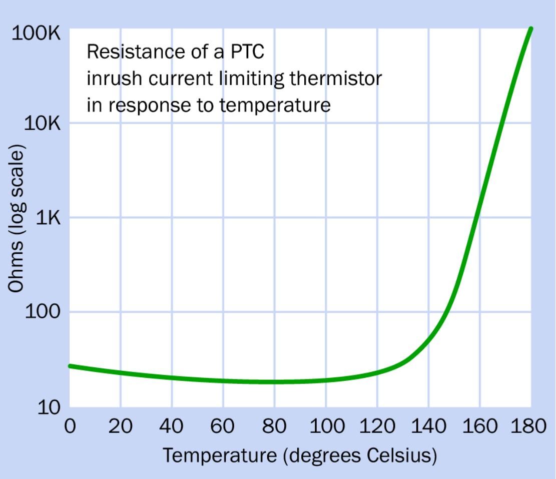

Figure 24-8 *Relationship of resistance to temperature in a PTC inrush current limiting thermistor. Note that the vertical axis has a logarithmic scale.*

PTC Thermistor for Starting Current

In some applications an initial inrush of current is actually necessary and desirable. An air conditioning compressor, for example, requires a surge of high current for "torque assist" when it is starting from a rest state.

High-current PTC thermistors may be used in this kind of situation. The Vishay PTC305C series is an example. These are heavy-duty components that have a switching time of about half a second, a maximum voltage rating of 410VAC or more, and a current rating from 6 to 36 amps.

The PTC thermistor has a relatively high temperature while the motor is running, and must be allowed time to cool before a restart is possible after shutdown. A waiting time of 3 to 5 minutes is imposed by a thermostat or separate time-delay relay.

PTC Thermistor for Lighting Ballast

The starting sequence for a fluorescent lamp requires that current should flow through the cathode heater initially. The thermistor allows this by bypassing a capacitor. Within less than a second, the resistance of the thermistor rises to block current. By this time, the heater has done its job and the lamp runs from high-frequency AC.

PTC Thermistor as a Heating Element

For small applications, a heating element can be made from a PTC thermistor, using its internal resistance to create heat. It has the advantage of being self-limiting, as its resistance rises with temperature. The TDK 5906 series is an example, shown in Figure 24-9. The component is approximately 12mm in diameter, and is designed to be clamped in place, not soldered. It has automotive applications for diesel fuel preheating and spray nozzle defrosting. Residential applications include vaporizers for air fresheners.

The initial resistance is as low as 3 or 4 ohms, rising very quickly at a transition temperature ranging from 70 to 200 degrees Celsius, depending on the specific component.

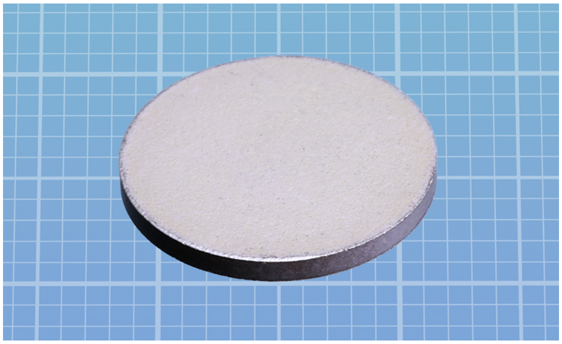

Figure 24-9 *This TDK B59060A0060A010 heating element is a PTC thermistor whose resistance rises rapidly around 80 degrees Celsius. Rated for 12VDC, it is intended for automotive applications. The background grid is in millimeters.*

What Can Go Wrong

Self-Heating

Self-heating may affect the accuracy of a temperature sensor. To get accurate readings, keep the current small. When the resistance of a thermistor is at the high end of its range, brief pulses of current can be used.

Heating Other Components

In cases where the self-heating of thermistors serves a useful purpose, as in surge protectors and when used for delays, the heat can damage nearby components or materials.

thermocouple

25

Because a *thermopile* is an assembly of thermocouples, it is included at the end of this entry. Other types of temperature sensors have their own entries.

OTHER RELATED COMPONENTS

- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **semiconductor** temperature sensor (see Chapter 27)
- **RTD** (resistance temperature detector) (see Chapter 26)
- **infrared temperature** sensor (see Chapter 28)

What It Does

A **thermocouple** measures temperature by using a pair of wires made from dissimilar metals. At one end of each wire, they are joined together, often by welding them. The differing thermoelectric characteristics of the wires generates a very small voltage between their free ends, from which the temperature of the joined ends can be derived.

No power supply is needed for a thermocouple, but the voltage that it generates is extremely small (measured not just in millivolts, but microvolts) and very nonlinear, requiring hardware and/or software to convert it to a temperature value. Laboratory equipment or integrated circuit chips are available for this purpose.

Different types of thermocouples are available to measure different temperature ranges, and each type has its own characteristics, requiring appropriate conversion.

A “raw” thermocouple looks very unimpressive, as it merely consists of two wires welded together at one end. This is illustrated in Figure 25-1. The full length of the photocouple is shown in Figure 25-2.

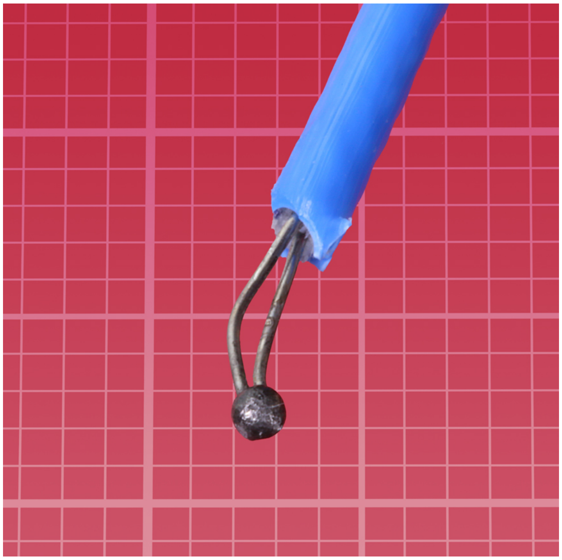

Figure 25-1 *Closeup of the welded wires in a K-type thermocouple. The background grid is in millimeters.*

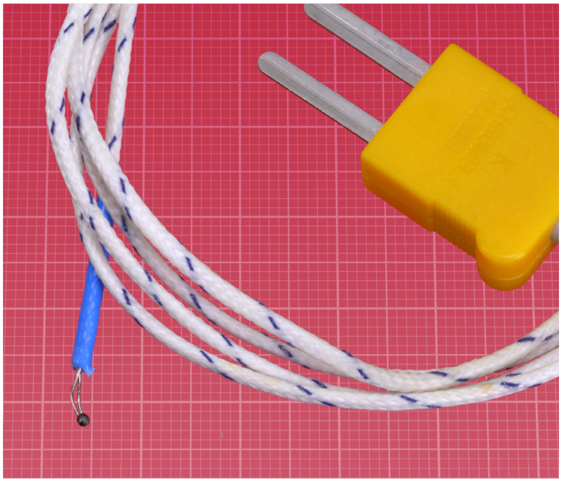

Figure 25-2 *Overview of the thermocouple in the previous photograph.*

A thermocouple sold as a commercial product is likely to be enclosed in a probe, as shown in Figure 25-3.

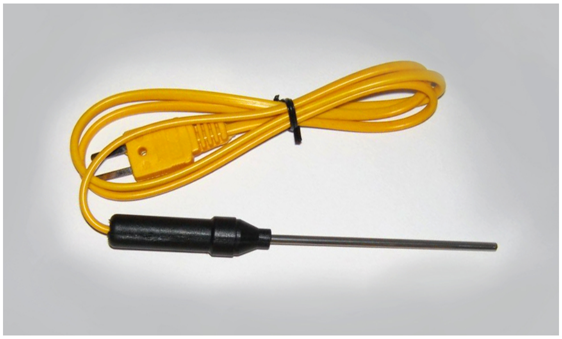

Figure 25-3 *A probe that contains a thermocouple.*

Schematic Symbol

A schematic symbol that is often used to represent a thermocouple is shown in Figure 25-4. Because this component does not consume current, the plus and minus signs do not mean that power should be applied to the wires. The positive sign indicates which wire will generate a higher voltage than the wire with the negative sign.

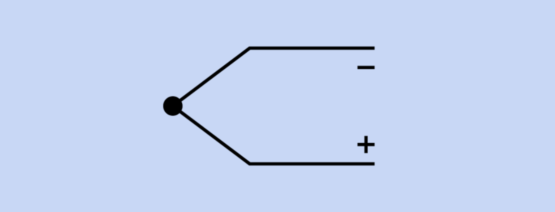

Figure 25-4 *A symbol that is often used for a thermocouple.*

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

Thermocouple Applications

Thermocouples have a wider range than any other form of contact temperature sensor, some types being capable of measuring up to 1,800 degrees Celsius. The main limitation is the ability of the joint between the wires to withstand the heat. Appropriate insulation must be used, but segments of ceramic tube are marketed to serve this purpose if necessary.

The very small thermal mass of a thermocouple enables a rapid response to temperature fluctuations. No self-heating occurs, because the thermocouple consumes no power. It is simple and robust. However, its response is very nonlinear, and the tiny voltages involved are vulnerable to corruption by electrical noise. Accuracy is usually not better than plus-or-minus 0.5 degrees Celsius, and may be less at low temperatures.

Thermocouples are commonly found in laboratories and in some industrial applications, such as monitoring the temperature in a blast furnace or inside an internal combustion engine.

They may also measure temperatures as low as -200 degrees, but at temperatures below -100 degrees the temperature coefficient diminishes to the point where voltage increments are less than 30µV per degree Celsius.

How a Thermocouple Works

When one end of a piece of wire is maintained at a temperature that is different from the other end, the temperature gradient along the wire creates a small electromotive force that manifests itself as a difference in electrical potential between one end of the wire and the other. This is known as the *Seebeck effect*, named after the man who discovered it. The magnitude of the potential will depend on two factors: the temperature difference between the ends of the wire, and the type of wire that is used.

Figure 25-5 illustrates the concept. Part 1 of this figure shows two wires, named A and B. The left ends of the wires are heated to the same temperature, T_X, while the right ends remain at a cooler temperature, T_Y. Because the wires are composed of different metals, the voltage drop across each wire will be different.

To make this model useful, some factors must be eliminated. In part 2 of Figure 25-5, the hot ends of the wires have been welded together. This now guarantees that they share the same temperature and the same voltage, V_X. We do not yet know what these X values are.

In part 3 of the figure, the cold ends of the wires are clamped in an isothermal block, which keeps them at an equal temperature, still represented as T_Y. The block is not electrically conductive, so the cold ends of the wires still have different voltages, V_A and V_B. We cannot measure these voltages directly, because they are relative to V_X, which is unknown. However, a volt meter can measure V_A and V_B relative to each other.

The volt meter will have its own voltage gradient on its wires, and possibly a temperature gradient too, but both of these wires are made of the same metal (probably copper) and share the same temperature gradient. Therefore, their effects will be equal.

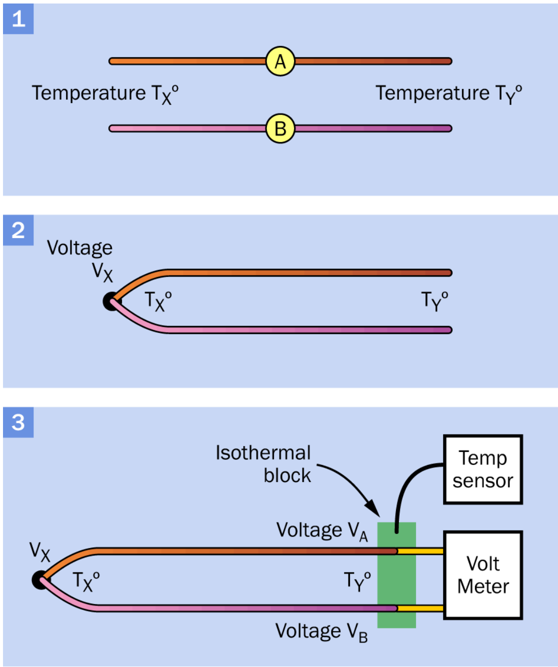

Figure 25-5 *Basic principles of a thermocouple. See text for details.*

A mathematical relationship exists between the temperature gradient and the voltage difference in each thermocouple wire. Suppose K_A is a constant or function that enables the voltage difference in wire A to be determined from its temperature gradient, and K_B serves the same function for wire B. Suppose T_{DIF} is the difference in temperature between T_X and T_Y. We may state:

$$K_A * (T_{DIF}) = V_X - V_A$$

$$K_B * (T_{DIF}) = V_X - V_B$$

By subtracting the second equation from the first and rearranging the terms, we get:

$$T_{DIF} * (K_A - K_B) = V_X - V_A - V_X + V_B$$

The two V_X terms cancel out, leaving $V_B - V_A$ on the right. We can recognize $V_B - V_A$ as the voltage difference measured by the meter. Call it V_M. So:

$$T_{DIF} = V_M / (K_A - K_B)$$

This enables calculation of the temperature difference between the ends of the wires, based on the meter reading and the conversion factor for each wire, which can be found experimentally. Because T_Y is being held at a known, constant value, we can determine the value of T_X:

$$T_X = T_Y + T_{DIF}$$

Thermocouple Details

When the thermocouple was first invented, the cold ends of the wires were placed in a bath of ice and water, forcing them to acquire and maintain a known temperature of 0 degrees Celsius.

The advent of accurately calibrated thermistors made it possible simply to measure the temperature of the cold ends. In this way, a thermistor enables a thermocouple to work. This prompts the question: why not just use the thermistor to measure T_X, and throw away the thermocouple? The reason is that a thermistor has a more limited range, seldom being used for temperatures above 150 degrees Celsius.

Note that the "hot end" of the thermocouple wires does not actually have to be hotter than the "cold end," even though those terms are commonly used. The equation to find T_X works just as well if T_Y is higher than T_X. The temperature difference will simply have a negative value instead of a positive value.

Because "hot" and "cold" are misleading terms, modern documents generally refer to the "measurement junction" and the "reference junction" of the wires. Note, however, that the wires are not actually joined with each other at the reference junction.

A common misconception is that voltage is generated where the wires are joined at the measurement junction. This is not correct. The voltage is a function of the temperature gradient between the measurement junction and the reference junction in each wire. Therefore, the way in which the wires are joined is irrelevant, provided there is an electrical connection between them. They can be welded, soldered, brazed, or crimped together.

How to Use It

Where a thermocouple is used in a laboratory, typically each wire is insulated, and they terminate in a plug that is inserted in a meter. The reference junction is hidden inside the meter, along with some electronics to decode the temperature data. The meter must have a setting that is appropriate to the specific type of thermocouple being used, so that the conversion factors are correct.

Because the type of metal in each wire must be consistent all the way from the measurement junction to the reference junction, other types of wires cannot be used to extend the reach of a thermocouple. Any extension must use wires made from the same metals. Connectors, also, must have pins and sockets that match the types of metals in the wires.

An extension wire for a thermocouple is shown in Figure 25-6.

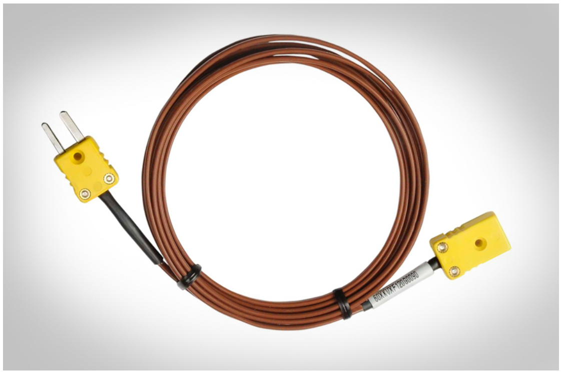

Figure 25-6 *Extension wire for a type K thermocouple. Note the polarized plug.*

Types of Thermocouples

Thermocouples are identified by ANSI standard codes consisting of single letters of the alphabet, shown below. Temperature ranges are approximate, in Celsius, with minimums rounded up and maximums rounded down to the nearest 50 degrees. Some data sources recommend narrower temperature ranges for practical use.

K type

-250 to +1,350 degrees. Most popular type of thermocouple. Positive wire is a nickel-chromium alloy, negative wire is a nickel-aluminum alloy. Commonly used in 3D printers.

J type

-200 to +1,200 degrees. Positive wire is iron, negative wire is a copper-nickel alloy. The iron wire is magnetic and vulnerable to corrosion. This thermocouple is not recommended for low temperatures, even though it is theoretically capable of measuring them.

T type

-250 to +400 degrees. Recommended for cryogenic applications. Positive wire is copper, negative wire is a copper-nickel alloy.

E type

-250 to +1,000 degrees. Most sensitive type, with the highest temperature coefficient. Positive wire is a nickel-chromium alloy, negative wire is a copper-nickel alloy.

N type

-250 to +1,300 degrees. An alternative to the K type, more stable at higher temperatures. Positive wire is a nickel-chromium-silicon alloy, negative wire is a nickel-silicon-magnesium alloy.

R type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

S type

-50 to +1,750 degrees. For high-temperature applications. Positive wire is a platinum-rhodium alloy, negative wire is platinum. Very low temperature coefficient.

Seebeck Coefficients

Datasheets for thermocouples list the Seebeck coefficient, which is the temperature coefficient caused by the Seebeck effect, measured in microvolts per degree. In other words, the value provided by a Seebeck coefficient is the number of additional microvolts that a thermocouple will generate for an increase in 1 degree Celsius.

Each type of thermocouple has a different coefficient, and because thermocouples tend to have a very nonlinear response, the coefficients will vary with temperature. Figure 25-7 shows the variations for six types of thermocouples, over a range from -400 to +1,400 degrees Celsius. It is important to understand that the vertical scale shows the coefficient for each type of thermocouple—that is, the *change* in voltage, not the *actual* voltage, at temperature values along the horizontal axis.

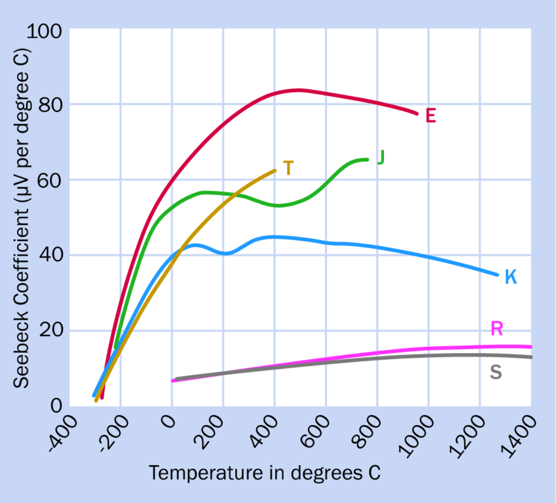

Figure 25-7 *The Seebeck (temperature) coefficient for six types of thermocouples. Partially derived from a datasheet published by Analog Devices.*

R and S type thermocouples have a relatively consistent response, but cannot achieve great accuracy, because the voltage increment is so small for each change in temperature. The K type thermocouple does relatively well between 0 and 1,200 degrees, but the J type only performs adequately between 0 and 800 degrees, while the T and E types are quite inconsistent.

For lower voltage readings, electrical noise becomes an issue. Thermocouple wires are often twisted together, and may also be shielded, to reduce sensitivity to noise. The electronics that decode the thermocouple voltage should include a filter to reject 50Hz or 60Hz interference from wiring in the vicinity.

Chips for Output Conversion

Meters that are designed to interpret the output from a thermocouple and display a temperature tend to be expensive, and may not be convenient for a custom-built application. Fortunately integrated circuit chips are now available to amplify the thermocouple output and apply signal conditioning to create an almost linear response.

The AD8494 and AD8496 from Analog Devices are precalibrated by laser wafer trimming to match the characteristics of J type thermocouples, while the AD8495 and AD8497 match the characteristics of K type. The chips require a power supply of at least 3VDC and have an analog output of 5mV per degree Celsius, enabling measurement over a range of almost 1,000 degrees. They require a very low supply current of 180µA. The manufacturer claims an accuracy of plus-or-minus 2 degrees Celsius.

The chip contains a temperature sensor which should be at the same temperature as the reference junction of the thermocouple. This means that the reference junction (typically, at the socket where a thermocouple plug is inserted) should be as close to the chip as possible, and the chip should be protected from heat created by other components. Any difference in temperature between the reference junction and the chip will create a temperature measurement error.

The MAX31855K chip from Maxim Integrated Products is another thermocouple-to-digital converter. It linearizes the output from the thermocouple and digitizes it as a temperature that is accessible by a microcontroller via a serial SPI bus. Breakout boards with this chip are available. The last letter of the chip number specifies the thermocouple type. Variants for J, K, N, T, S, and R are available.

The AD8495 is mounted on a breakout board from Adafruit, and the MAX31855K is mounted on a breakout board from Sparkfun. These boards are pictured in Figure 25-8.

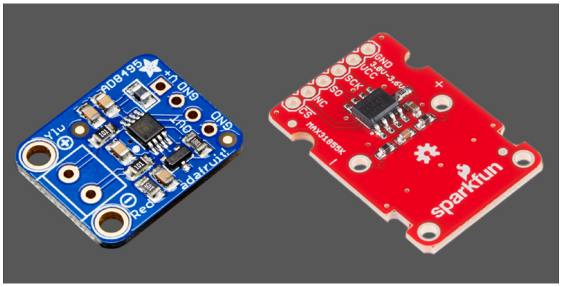

Figure 25-8 *Thermocouple amplifier/conversion chips from Adafruit (left) and Sparkfun (right).*

Thermopile

A thermopile is an assembly of thermocouples connected in series, as suggested in Figure 25-9, where a hot area is shown on the left and a cooler area is shown on the right. The figure assumes that the orange-colored wires have a voltage difference of 5mV from left to right, as a result of the temperature difference, while the purple-colored wires only have a voltage difference of 1mV from left to right. Therefore, the voltage difference between each reference junction and the next is 4mV, as shown on the right side of the figure. So long as the temperature difference exists, the voltage differences will be cumulative, totaling 16mV between top and bottom in this example.

Note that the multiple thermocouple junctions are not electrically connected with each other in each temperature zone.

In reality, more thermocouples than this will be added, and the voltage differences may be lower.

Generally a thermopile is not sold as a separate component from retail suppliers, but is built into other devices. It may be used to generate small amounts of current from a heat difference, as in an infrared thermometer. It can also be used as a safety device to shut down a gas supply if a burner is not lit. See Chapter 28.

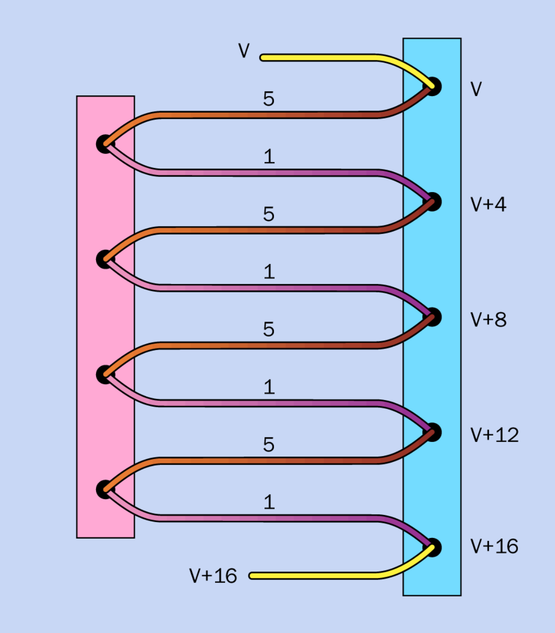

Figure 25-9 *The operating principle of a thermopile. The numbers represent mV, but are chosen arbitrarily as an example.*

What Can Go Wrong

Polarity

The output from a thermocouple has polarity. If this is not observed, an error will result.

Electrical Interference

Thermocouple wires are vulnerable to electrical interference, and should be a twisted pair or, ideally, shielded.

Metal Fatigue and Oxidation

The wire used in some thermocouples tends to be relatively brittle, and cannot withstand much flexing. Also, some metals or alloys are vulnerable to oxidation.

Using the Wrong Type

Different types of thermocouples have totally different characteristics. The electronics to decode the signal from a thermocouple must be matched to the type of thermocouple being used. The plugs on the ends of the thermocouple wires are often retained with screws. A detached plug should be replaced immediately, to avoid the error of attaching it to the wrong type of thermocouple.

Heat Damage from Creating a Thermocouple

If a thermocouple is made from two wires on a DIY basis by welding the tips of the wires together, minimal heat must be used to avoid degrading the alloys in the wires.

RTD (resistance temperature detector)

26

RTD is an acronym, either for *resistance temperature detector* or *resistive temperature device*. No definitive information seems to exist regarding which term is correct, but *resistance temperature detector* is more common.

Occasionally an RTD may be described as a *PTC thermistor*, but its sensing element is different, consisting of pure metal wire or film.

OTHER RELATED COMPONENTS

- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **semiconductor** temperature sensor (see Chapter 27)
- **infrared temperature** sensor (see Chapter 28)

What It Does

A *resistance temperature detector*, also known as a *resistive temperature device*, is usually referred to as an **RTD**. It has a positive temperature coefficient (that is, its resistance increases as its temperature increases) but differs from a **PTC thermistor** in that its sensing element is pure metal instead of a semiconductor.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

RTD Attributes

Positive attributes of RTDs include:

- Accuracy, often plus-or-minus 0.01 degrees Celsius. This very small tolerance allows excellent interchangeability.
- Stability, with a response that drifts by as little as 0.01 degree per year.
- The output is an almost linear function of temperature, making them easily used with a microcontroller.
- Immunity to electrical noise.
- Reasonably rapid response to temperature changes (about 1 to 10 seconds).

Undesirable attributes include:

- Temperature coefficient about one-tenth of that of an NTC thermistor.
- To measure the resistance, some current must pass through the sensor, raising the possibility of self-heating (as in the case of other temperature sensors, with the exception of a thermocouple).
- Relatively high cost, especially of the wire-wound type.

The resistance curves for three generic NTC thermistors are shown in Figure 26-1, plotted against the resistance for a generic platinum RTD that has a reference resistance of 100 ohms at 0 degrees Celsius. Note that this graph is unlike many that illustrate the response of NTC thermistors, in that its vertical scale is not logarithmic.

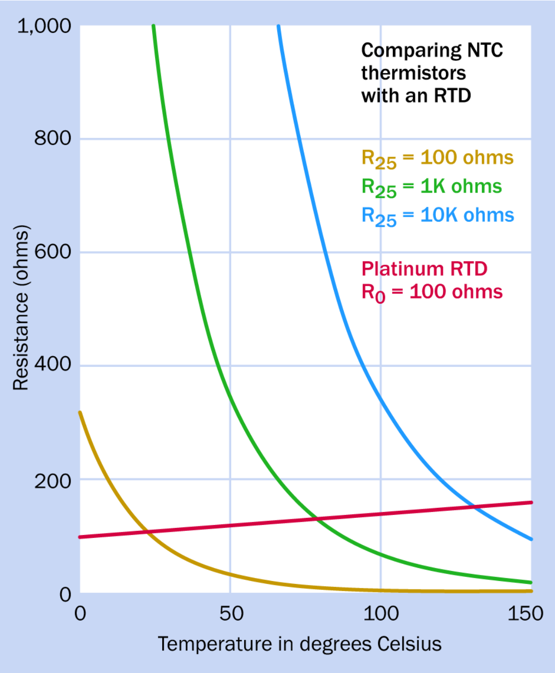

Figure 26-1 *The brown, green, and blue curves show the resistance of three generic NTC thermistors varying with temperature. The red line shows the resistance of a platinum RTD. Derived from a diagram created by Texas Instruments.*

Schematic Symbol

There is no specific schematic symbol for an RTD. Often the symbol for a thermistor may be used. See Figure 23-1.

Applications

Because of its high accuracy, an RTD may be used where precision is important. It can calibrate other temperature sensors, and may measure the temperature of a reference junction of a thermistor. However, it requires sensitive electronics for signal conditioning, because of its low temperature coefficient.

How It Works

An RTD exploits the fractional increase in electrical resistance of a metal film, metal filament, or (in some cases) a carbon film, when the temperature of the metal rises. In its simplest form, an RTD is a 2-wire device with no polarity.

The sensing element is often made from platinum, as this has a linear response to temperature over a wide range. High-quality RTD sensors with a wide temperature range usually consist of platinum wire that is wound around a glass or ceramic core. Smaller sensors may be fabricated from a thin layer of platinum evaporated onto an insulating substrate. Nickel may be substituted for platinum, and has a more sensitive but less linear response.

The wire-wound type is usable at temperatures as high as 500 Celsius (up to 1,000 degrees for some platinum-element types). Some variants are able to measure temperatures as low as -250 degrees.

DIN 60751 is an international standard defining the performance of platinum RTDs. It specifies a reference resistance of 100.00 degrees at 0 degrees, and a temperature coefficient between 0 degrees and 100 degrees. Outside of this range, a formula defines the response.

The response is almost precisely linear, ranging from 100 ohms at 0 degrees to approximately

138 ohms at 100 degrees. The temperature deviation from a straight line, between 0 and 100 degrees, is no more than plus-or-minus 0.8 degrees.

However, current through the RTD must be restricted to avoid self-heating. A range of 0.5mA to 1mA is recommended.

Variants

Some RTDs are potted in glass or resin, as shown in Figure 26-2. This shows an RTD in the TFPTL range from Vishay, containing a nickel thin-film sensing element with a temperature coefficient of about 0.4% and a tolerance of 0.01%. It is available with a very wide range of reference resistances, from 100 ohms to 5K (measured at 25 degrees Celsius). The temperature range is -55 to +70 or -55 to +150 degrees, and the maximum voltage is 30V to 40V, depending on the specific component.

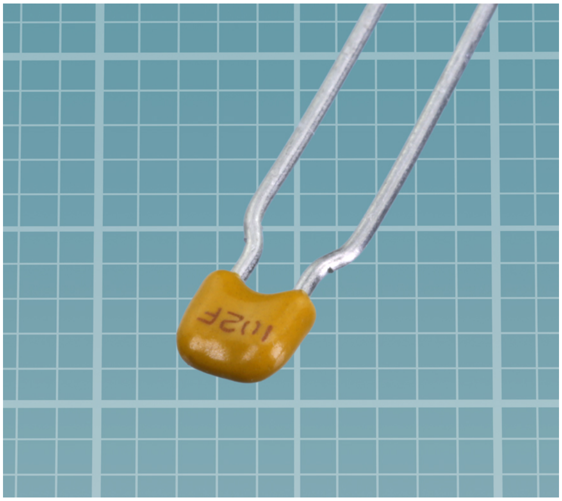

Figure 26-2 *An RTD in the TFPTL series from Vishay. The background grid is in millimeters.*

The flat package shown in Figure 26-3 may be encased in a protective sheath of plastic or silicone rubber, and can be used for surface temperature sensing where the component is glued to the exterior of a flat-sided container. This figure shows an RTD in the L420 range from Heraeus Sensor Technology, containing a platinum thin-film sensing element with a temperature coefficient of 0.385%. It is available in reference resistances of 100, 500, and 1,000 ohms (measured at 25 degrees Celsius). The temperature range is -50 degrees to +400 degrees.

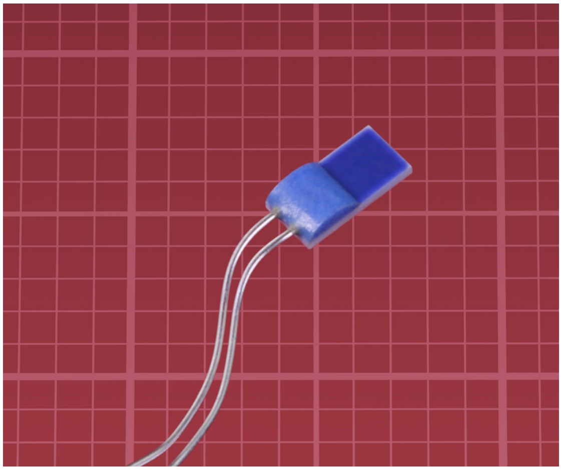

Figure 26-3 *An RTD in the L420 series from Heraeus. The background grid is in millimeters.*

Wiring

The leads to an RTD can be a source of error. If a simple 2-wire configuration is used, the leads will have an unknown resistance that will be affected by temperature, just as the element inside the RTD will have a temperature-sensitive resistance.

To enable temperature compensation, a three-wire design can be used. Figure 26-4 illustrates the principle. In the first section of this figure, resistances R_A and R_B remain unknown. In the second section, the resistances of R_B and R_C can be found by passing a test current through one wire and back through the other, bypassing the component. Assuming that all of the leads are identical in length and composition, the resistance of $R_A + R_B$ will be equal to that of $R_B + R_C$.

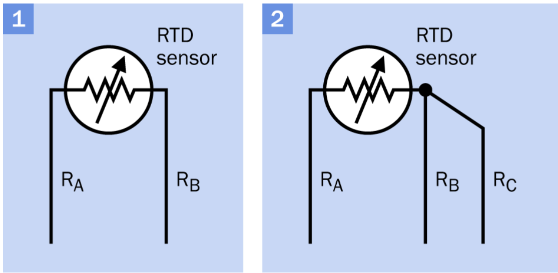

Figure 26-4 *A three-wire configuration enables temperature compensation for the leads to an RTD. See text for details.*

RTD Probe

For practical applications, an RTD sensor is often packaged inside a probe that can be indistinguishable from the type of probe used with a **thermocouple**. However, a thermocouple always uses two wires, as the wires themselves create the voltage. An RTD often uses three wires, as shown in Figure 26-5. This particular sensor is sold for use in a "Brew-Magic" system for brewing craft beer on a commercial basis.

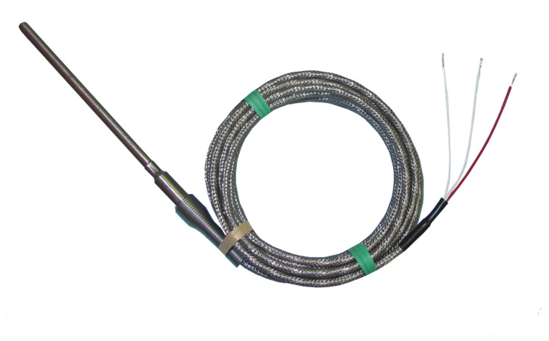

Figure 26-5 *A three-wire RTD is packaged inside this steel probe.*

Signal Conditioning

To process the signal of an RTD, a chip such as the LM75 from National Semiconductor can be used. This is calibrated for connection with a platinum RTD. Internally it converts the resistance of an RTD to a value of 5mV per degree Celsius. This then passes through an analog-to-digital converter on the chip, creating a digital value that can be read via an I2C bus.

What Can Go Wrong

Self-Heating

Self-heating is an issue for RTDs, just as it is for thermistors. Current through an RTD should be limited to 1mA, especially when measuring low temperatures.

Insulation Affected by Heat

The resistance of insulation on the wires leading to a sensor can change with temperature, leading to incorrect resistance readings. This is a more likely source of problems for RTDs than for thermistors, as RTDs are often used at higher temperatures and have a lower temperature coefficient.

Incompatible Sensing Element

If signal conditioning is applied to an RTD that has an incompatible sensing element, temperature readings will be incorrect. For example, an RTD with a nickel element should not be used with signal conditioning designed for a platinum element.

semiconductor temperature sensor

27

This type of sensor may also be referred to as a *bandgap temperature sensor*, a *diode temperature sensor*, a *chip-based temperature sensor*, or an *IC temperature sensor*.

The unfortunate term *integrated silicon-based sensor* is sometimes used, which can be confused with *silicon temperature sensor* (also known as a *silistor*), which is a type of **PTC thermistor**. See Chapter 24.

Some vendors do not divide temperature sensors into clear categories. Semiconductor temperature sensors may be classified as *board-mount* temperature sensors, even though many of them have leads and are not specifically designed to be mounted on circuit boards.

A semiconductor temperature sensor with a digital output is sometimes described as a *digital temperature sensor* or *digital thermometer*. This can be misleading, as the outputs from other types of temperature sensors may be digitized with appropriate components.

OTHER RELATED COMPONENTS

- **thermocouple** (see Chapter 25)
- **NTC thermistor** (see Chapter 23)
- **PTC thermistor** (see Chapter 24)
- **infrared temperature** sensor (see Chapter 28)
- **RTD** (resistance temperature detector) (see Chapter 26)

What It Does

A **semiconductor** temperature sensor is an integrated circuit chip incorporating a sensing element composed of transistor junctions. It has an approximately linear response and is easy to use, in some cases being designed for direct connection with a microcontroller, requiring no additional components.

In analog variants, the output consists either of voltage or current that varies with temperature. These components have a positive temperature coefficient, except for a few CMOS variants where a voltage output diminishes as temperature increases.

Digital variants are becoming more common, creating a numeric output accessible by a microcontroller.

In almost all semiconductor temperature sensors, the characteristics of silicon dioxide limit the temperature range to approximately -50 to +150 degrees Celsius (sometimes less).

This type of component is not (yet) as low in cost as a thermistor, but can include its own amplification, signal processing, and (optionally) analog-to-digital conversion on one chip.

Comparison of Temperature Sensors

In this Encyclopedia, temperature sensors are divided into five main categories, each of which has a separate entry. For convenience, a comparative summary is included in the entry for **NTC thermistors**. See "Addendum: Comparison of Temperature Sensors" for an overview. Also see Figure 23-9.

Semiconductor Temperature Sensor Applications

When a semiconductor temperature sensor is used in surface-mount format, it can measure the temperature of the board on which it is mounted. This enables protection from overheating, often in power supplies.

Because the sensing elements and signal processing circuits are all chip-based, they can be transplanted into other types of sensors. For example, a gas pressure sensor or a proximity sensor can have onboard compensation using a semiconductor temperature sensor. They have also been built into computer CPUs such as the Pentium series from Intel.

Some variants are manufactured in a three-lead TO-92 package, appearing superficially similar to bipolar transistors. They are suitable for remote temperature sensing, and have automotive applications such as measuring the temperature of the transmission, engine oil, or cabin interior. They may also be found in some heating and air-conditioning systems, and some kitchen equipment.

Schematic Symbol

No unique schematic symbol has been developed for a semiconductor temperature sensor. It may be represented by a rectangle containing text abbreviations to represent pin functions, similar to other types of integrated circuit chips.

In the case of a sensor with an output consisting of current that varies with temperature, the sensor may be shown as a *current source*, using the symbol in Figure 27-1. However, this symbol is not specific to temperature sensors; it is used for any component that is a current source.

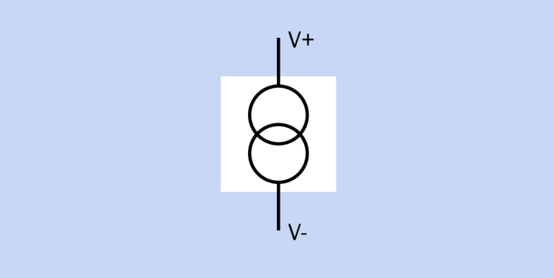

Figure 27-1 *A temperature sensor whose output consists of current varying with temperature may be shown in a schematic as a current source, using this symbol.*

Attributes

Desirable attributes of a semiconductor temperature sensor include:

- Easy to use. Few or no external components may be required, and little or no signal processing.
- Factory-calibrated, with an almost linear response.
- Versions with a digital output are easy to add to any system that already has an I2C bus. For additional details about protocols such as I2C, see Appendix A.

Undesirable attributes of semiconductor temperature sensors include:

- Limited temperature range, the same as thermistors.
- Self-heating issues, especially in versions where signal-processing functions are built into the same chip.

- Not as rugged as some types of temperature sensors.

For an explanation of terminology used for temperature sensors in datasheets, see "Thermistor Values".

How It Works

When a constant current is flowing through a p-n junction in a diode, the voltage across the diode will change by about 2mV for each change in temperature of 1 degree Celsius. This can be demonstrated by the simple circuit shown in section 1 of Figure 27-2.

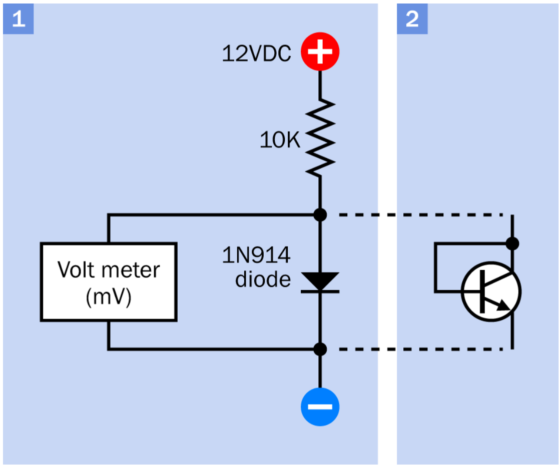

Figure 27-2 *Left, a basic circuit for demonstrating the temperature sensitivity of a diode. Right, an NPN transistor can be substituted to emulate the diode.*

Similarly, the voltage across the p-n junction in an NPN transistor varies with temperature, if the current is constant. A transistor can be substituted for a diode as suggested in section 2 of Figure 27-2. Integrated circuit chips that contain transistors can measure temperature by exploiting this phenomenon.

CMOS Sensors

Some semiconductor temperature sensors use CMOS instead of bipolar transistors. The general concept is similar, but they are described separately, below. See "CMOS Semiconductor Temperature Sensors".

Multiple Transistors

The heat sensitivity of a bipolar transistor can be defined with an equation. If the base-emitter voltage is V_{BE}, q is the charge of an electron, k is a constant (known as Boltzmann's constant), T is the temperature in degrees Kelvin (relative to absolute zero), I_C is the collector current, and I_S is the saturation current (which is less than I_C):

$$V_{BE} = ((k*T) / q) * \log_e (I_C / I_S)$$

The term $\log_e$ means, "the logarithm to base e of the expression in parentheses."

Because k and q have known values, the base-emitter voltage turns out to be proportional to the logarithm of the collector current divided by the saturation current. However, the saturation current depends on the geometry of a transistor, and varies with temperature in a nonlinear way.

To eliminate the factor of saturation current, one transistor can be compared with another transistor that has a larger emitter area. This enables the derivation of a new equation that specifies temperature while getting rid of the troublesome saturation currents, with their nonlinear behavior.

However, it may not be easy to fabricate two transistors, in the same silicon chip, that have the same characteristics except that the emitter area of one is bigger than that of the other. It is much easier to add multiple transistors in parallel, each of them identical to the first. The total emitter area will then be equal to the area in one transistor multiplied by the number of transistors.

In Figure 27-3, assuming all the transistors are at the same temperature in addition to being of identical specification, we can now write two equations. The figure allows room to show only three transistors, but suppose there are N of them. If V_{BE0} is the base-emitter voltage of transistor Q0, on the left, and V_{BEN} is the aggregate base-emitter voltage of the N transistors on the right:

$$V_{BE0} = ((\ k*T\ )\ /\ q)\ *\ \log_e\ (I_C\ /\ I_S)$$

$$V_{BEN} = ((\ k*T\ )\ /\ q)\ *\ \log_e\ (I_C\ /\ N*I_S)$$

From these, an equation can be derived that gets rid of I_C and I_S:

$$V_{BE0}\ -\ V_{BEN} = ((\ k*T\ )\ /\ q\ )\ *\ \log_e\ (N)$$

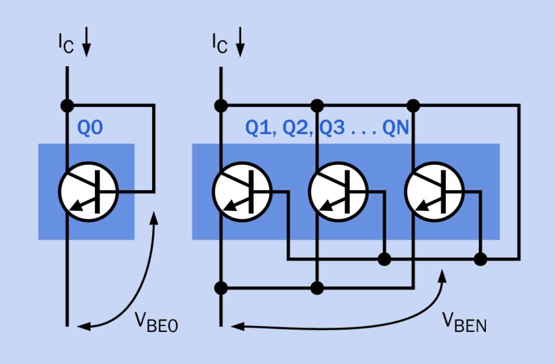

Figure 27-3 *Comparing the base-emitter voltage of one transistor with a set of identical transistors can enable measurement of temperature regardless of the collector current and saturation currents, so long as all the transistors are at the same temperature. See text for details.*

PTAT and the Brokaw Cell

Now if a comparator is added to control the current, a circuit known as the Brokaw Cell is created, shown in Figure 27-4. This is also known generically as a *bandgap temperature sensor*. (A couple of resistors have been omitted for the sake of simplicity.)

Typically, N = 8. That is, there is a set of eight transistors in addition to Q0 (only three being shown here). The voltage difference in the previous equation, V_{BE0} - V_{BEN}, now appears across R2 in the figure, and the voltage across R1 is *proportional to absolute temperature*, often referred to by its acronym, *PTAT*. This voltage can be found from this equation:

$$V_{PTAT} = ((k*T)/q)\ *\ \log_e(N)\ *\ (2*R1/R2)$$

The Brokaw Cell was the basis of the AD580 chip introduced in 1974 by Analog Devices, and the principle is now used very widely in semiconductor temperature sensors.

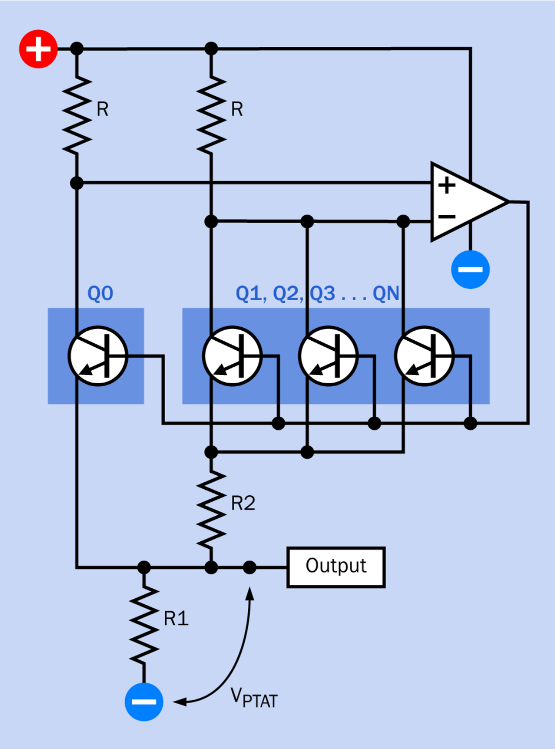

Figure 27-4 *The Brokaw Cell. See text for details.*

Variants

Three output types are used:

- Analog voltage output (voltage varies with temperature).
- Analog current output (current varies with temperature).
- Digital output.

A fourth type creates an output in the form of a square wave, either with a frequency or wavelength proportional to temperature. The Maxim MAX6576 and MAX6577 are examples. However, this type of output is so rare, it is not described in detail here.

Some semiconductor temperature sensors are CMOS-based, and have a voltage output with a negative temperature coefficient. They are described separately. See "CMOS Semiconductor Temperature Sensors".

Analog Voltage Output

LM35 Series

The LM35 is a typical, widely used semiconductor temperature sensor, available from Analog Devices, Texas Instruments, and other manufacturers. Its output voltage changes by 10mV per degree Celsius over a range of approximately -50 degrees to +150 degrees. Accuracy is stated to be plus-or-minus 0.25 degrees at room temperature and plus-or-minus 0.75 degrees over the whole range.

The sensor can be obtained packaged like a transistor, in a TO-92 plastic capsule or metal can. It is also available as a surface-mount component, or in a TO-220 package, like a 5V voltage regulator, as shown in Figure 27-5.

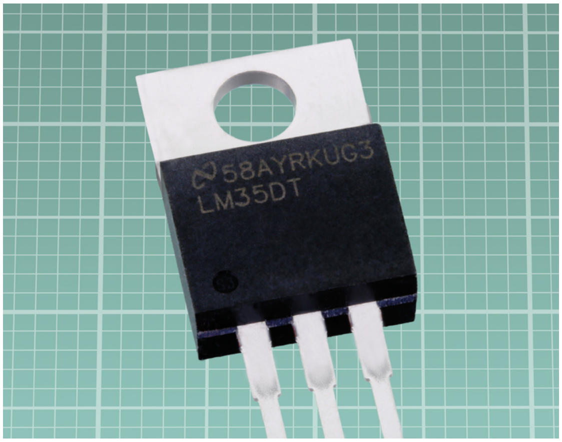

Figure 27-5 *This version of the LM35 can be used to measure surface temperature when secured with a bolt. The background grid is in millimeters.*

This is a three-wire component, two pins or solder pads being used for the power supply while the third serves as the sensor output. The supply voltage typically ranges from 4V to 30V. Necessary current consumption is only 60μA, which minimizes self-heating.

Because this device is specifically designed for the Celsius temperature range, its output is scaled to 0mV at 0 degrees. A pulldown resistor can be added to measure temperatures below zero.

A bypass resistor of 200 ohms between the output and ground is recommended as a precaution against capacitive effects in the cable run.

The LM34 is almost identical to the LM35, except that its output changes by 10mV per degree Fahrenheit instead of 10mV per degree Celsius.

LM135 Series

Although this sensor contains multiple NPN junctions, the manufacturer describes it as behaving like a zener diode in which the breakdown voltage is directly proportional to absolute temperature. The output increases by 10mV per degree over a range from -55 to +150 degrees Celsius.

For the LM135, the manufacturer claims an error of less than plus-or-minus 1 degree Celsius between 0 and 100 degrees. For the LM235 and LM335, in the same product series, the temperature range is narrower, the accuracy is lower, and the price, also, is lower. An LM335 sensor is shown in Figure 27-6.

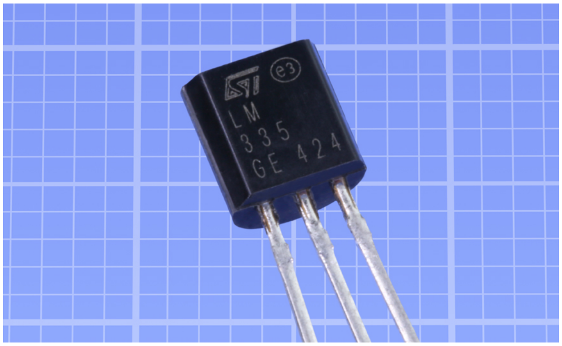

Figure 27-6 *A sample of the LM335 temperature sensor in a TO-92 package. The background grid is in millimeters.*

The sensor is available in a TO-92 package (plastic, like a transistor) or a TO-46 (metal can). It is also manufactured in a surface-mount format. The negative terminal is connected directly to ground, while the positive terminal is connected through a series resistor to the positive side of a power supply that can range from 5V to 40V. The third terminal, labeled "ADJ" on datasheets, allows for the output to be adjusted. Figure 27-7 shows the basic circuit. The value of R1 can be chosen to establish an optimal current of 1mA through the sensor, although a range of 400µA to 5mA is tolerable.

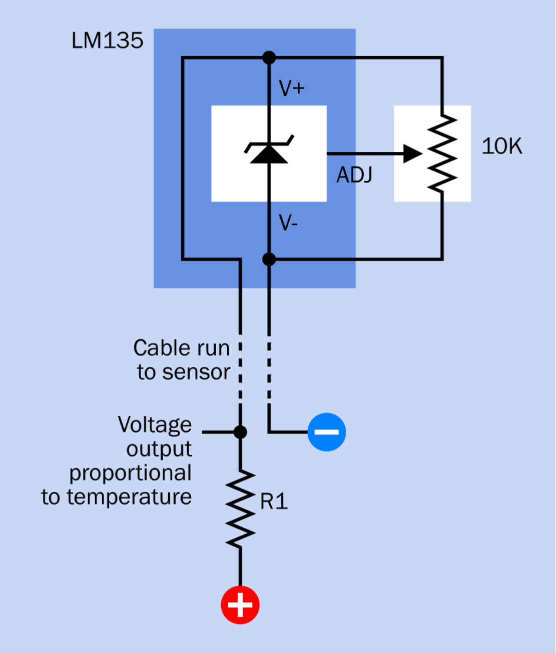

Figure 27-7 *Basic schematic for using an LM135, including output adjustment. Because the sensing element behaves like a zener diode, it is represented with the zener symbol.*

Analog Current Output

Fewer components exist using output current to measure temperature. The output is applied to a grounded resistor, and the voltage across the resistor then changes with current from the sensor.

The useful aspect of a current output is that its accuracy is unaffected by a cable run as long as 200 or 300 feet. Therefore, this type of component is appropriate as a remote sensor.

LM234-3 Series

This is a three-wire sensor, two wires being used for bias power supply and ground (labeled V+ and V- on the datasheet) and a third (labeled R) that delivers current proportional to temperature. The current from pin R passes through an external resistor to ground, and the voltage across this resistor varies by 214µV per degree Kelvin. A bias voltage ranging from 1V to 40V is required.

A sample of the LM234 is shown in Figure 27-8.

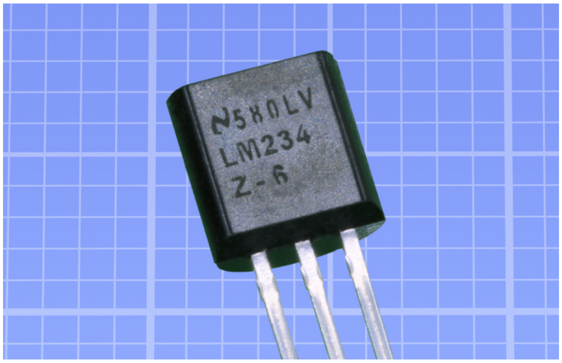

Figure 27-8 *The LM234Z temperature sensor in a TO-92 package. The background grid is in millimeters.*

If the component is used for remote sensing, the resistor should be 230 ohms and can be connected directly between pin R and pin V- of the sensor at the far end of a wire run. At the "home" end of the wire run, temperature output is taken from above a 10K resistor that is placed between the return wire and ground, as shown in Figure 27-9. With these component values, the output voltage will change by 10mV per degree Kelvin.

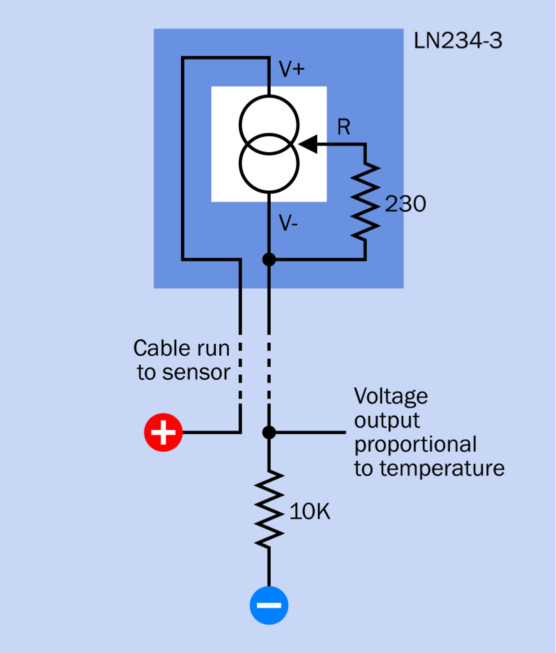

Figure 27-9 *Using an LM243-3 sensor with a current output that varies with temperature.*

The LM234-3 can be encapsulated in a plastic TO-92 package or a TO-46 metal can. A surface-mount version is also available.

The claimed accuracy is plus-or-minus 3 degrees. The temperature range is -25 degrees to +100 degrees.

AD590 Series

The Analog Devices AD590 (successor to the original AD580) is a current-output sensor that uses only two wires. Like the LN234-3 it is available in a TO-46 metal can, but with one lead making no internal connection. It also can be bought in a two-wire "flatpack," or as a surface-mount chip (with eight solder pads, only two of which are connected).

Using a supply voltage of 4V to 30V, the sensor's high-impedance output changes by 1μA per degree Kelvin. Voltage supply variations produce very small errors in the output current; substituting 10V for 5V creates a deviation of only 1μA.

Figure 27-10 shows an application for the AD590 using a resistor and a trimmer to adjust the scale factor. When properly set up, this circuit provides an output that changes by 1mV per degree Kelvin.

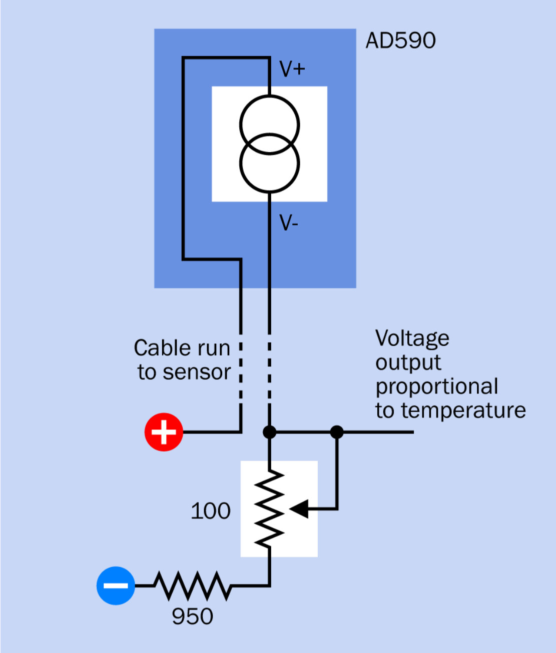

Figure 27-10 *The simplest circuit allowing fine adjustment for use with an AD590 sensor.*

Digital Output

Some of the most popular examples of semiconductor temperature sensors with digital output include the TMP102 series from Texas Instruments, MCP9808 series from Microchip, LM73 series from Texas Instruments and National Semiconductor, and DS18B20 series from Maxim. All of these components can measure a range typical of semiconductor temperature sensors, from approximately -50 to +150 degrees Celsius. Most of them claim an accuracy in the region of plus-or-minus 1 degree over the full range or 0.5 degree in the 0 to 100 range. With the exception of the Maxim DS18B20, which uses its own unique protocol, the components communicate via either the I2C or SMBus protocol.

TMP102 Series

This is available only in a surface-mount format. It has fewer features than the other sensors listed here, and is less accurate, claiming plus-or-minus 3 degrees Celsius over its maximum range of -40 to +125 degrees Celsius. However, it is less expensive. (For greater accuracy, the TMP112 is available.) This is a low-voltage chip requiring 1.4V to 3.6V as its power supply, drawing a quiescent current of 10μA. Temperatures are stored in a 12-bit or 13-bit format that requires some conversion, as a single bit represents 0.0625 degrees Celsius. An alert pin is activated when the measured temperature deviates above or below limits that are preset by the user. No hysteresis adjustment is available for the alerts. The TMP102 is available on a breakout board from Sparkfun, as shown in Figure 27-11.

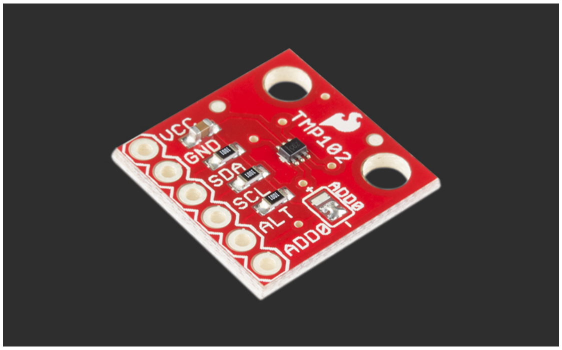

Figure 27-11 *The Texas Instruments TMP102 on a breakout board from Sparkfun.*

MCP9808 Series

This multifunction sensor is available either as a regular surface-mount component, or as a surface-mount with an exposed “thermal pad.” It conforms with the I2C bus standard at up to 400kHz, allowing up to 16 sensors to share the same bus. The chip has a variety of temperature alert features, including high and low limits that can activate a dedicated “alert” pin, and a hysteresis value that can be set for the limits, to ignore brief temperature excursions. The chip can be put into “comparator mode,” where it simply provides logic-high or logic-low output if the temperature is above or below a user-specified value. This feature makes the chip operate as a thermostat. Temperature resolution is user-selectable. The temperature storage format requires some conversion to obtain a Celsius value, to deal with negative values and fractional values. However the chip is available on a breakout board from Adafruit, as shown in Figure 27-12, and an Arduino code library is available.

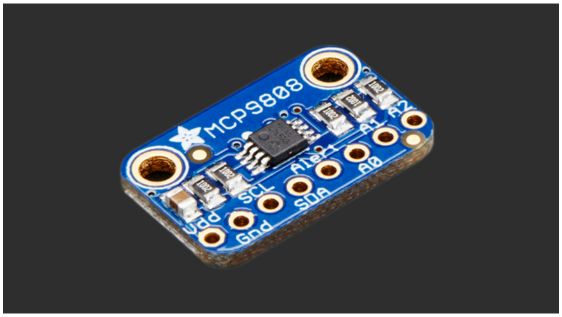

Figure 27-12 *The Microchip MCP9808 on a breakout board from Adafruit.*

LM73 Series

This sensor is only available in surface-mount format. It conforms with the I2C bus standard at up to 400kHz. Its temperature resolution can be set to 11, 12, 13, or 14 bits. An “alert” pin becomes active if the temperature exceeds a preprogrammed limit. An “address” pin can select one of three device addresses by being held in logic-high, ground, or disconnected status. The chip can be put into shutdown mode when power conservation is necessary.

DS18B20 Series

Unlike most digital sensors, this is a three-wire component, because it uses Maxim’s proprietary “1-wire bus” with a unique protocol. The bus allows access to a 2-byte register storing digital output from the temperature sensor, but also allows the user to perform other functions, such as setting the resolution of an onboard analog-to-digital converter (which has a maximum resolution of 12 bits), setting a high-temperature and low-temperature alarm, and allowing the sensor to be identified, as each component has a unique 48-bit serial number in ROM.

The chip can draw sufficient power from the data bus to operate, so long as the bus is held high by a 4.7K pullup resistor. (Maxim describes this as "parasite power.") An internal capacitor sustains the chip briefly while the bus is used for its normal purpose of transferring data, but if the bus has low voltage for more than 480µs the chip will reset itself. The "parasite power" feature also will not work above 100 degrees Celsius. Perhaps recognizing that this system may create more problems than it solves, Maxim has also given the component a normal power input pin.

The DS18B20 is available in the TO-92 package and two sizes of surface-mount chip. Its lack of a standard I2C bus, and its use of complicated proprietary codes, create a steep learning curve. Still, it remains a popular sensor, and an Arduino code library for it is available online.

CMOS Semiconductor Temperature Sensors

CMOS variants of semiconductor temperature sensors have appeared relatively recently compared with the bipolar variants. They draw a very low quiescent current (typically, a few microamps) and can work with a power supply from 5.5VDC down to 2.2VDC, making them suitable for handheld battery-powered devices. An analog output is common. Popular examples are the LM20 and the LMT86 series.

Like bipolar sensors, the LMT86 sensors have a limited temperature range, between approximately -50 and +150 degrees Celsius. Again, like the bipolar sensors, they are available optionally in TO-92 and surface-mount packages. A significant difference is that the output has a negative temperature coefficient, diminishing by 10mV per degree Kelvin, because of the characteristics of CMOS semiconductors. The claimed accuracy is plus-or-minus 0.25 degrees Celsius. The output voltage covers a range of about 2V, diminishing from 0.5V below the supply voltage at -50 degrees Celsius.

A sample of the LMT86 is shown in Figure 27-13.

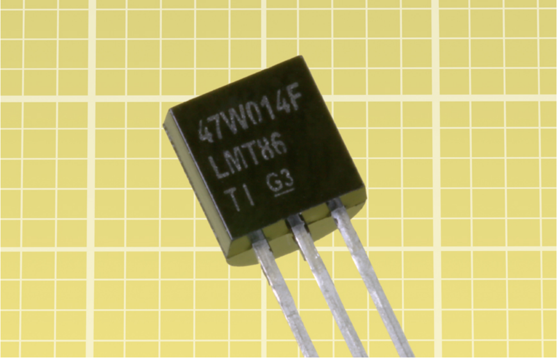

Figure 27-13 *A sample of the LMT86 CMOS temperature sensor. The background grid is in millimeters.*

What Can Go Wrong

Different Temperature Scales

Some voltage-output sensors create an output convertible to degrees Kelvin, while others use degrees Celsius. While the temperature degrees are the same in each scale, a component with an output in millivolts may assign 0mV either to 0 degrees Celsius or 0 degrees Kelvin (equivalent to -273.15 degrees Celsius). The advantage of using a Kelvin scale is that it avoids the problem of negative temperature values.

Rarely, a sensor may use degrees Fahrenheit.

Interference in Cable Runs

Sensors with a voltage output are susceptible to electrical interference. Twisted-pair or shielded cable runs are recommended when sensors are placed remotely.

For the Maxim DS18B20, which uses a 1-wire bus, multiple sensors should be connected along one run of wire (linear topology) instead of each sensor being connected to a central point (star topology). If the cable lengths are longer than a few meters, the topology starts to matter.

Latency

The packaging of semiconductor temperature sensors can create latency in their response time. While a thermocouple consists only of a pair of wires joined by a small dot of melted metal, a TO-92 semiconductor package adds thermal mass that will slow the response considerably. Moreover, copper leads will conduct heat from a circuit board if the board is warmer than its environment.

Surface-mount chips have a very low mass, but must be soldered to some kind of board, even if it is a very small one.

Generally, other types of sensors may be appropriate where rapid response is necessary.

Processing Time

In a sensor with a digital output, the onboard analog-to-digital converter will add a small delay before the data becomes available, and during that delay, the component cannot respond to a new temperature. The output from an analog device may be more suitable for rapid detection of temperature variations.

infrared temperature sensor

28

An **infrared temperature** sensor is sometimes described as a *thermopile*. In reality the sensor module *contains* a thermopile. In this Encyclopedia, a thermopile is considered to be a separate component, described in the entry discussing **thermocouples**. See "Thermopile" in Chapter 25.

Other terms that are sometimes used for an infrared temperature sensor are *contactless thermometer* or *infrared thermometer*. This Encyclopedia classifies a thermometer as a commercially marketed product, not a component.

Devices such as a *radiation pyrometer*, *IR pyrometer*, *optical pyrometer*, or *thermal imager* provide ways of measuring infrared radiation, but are outside the scope of this Encyclopedia.

A **passive infrared** motion sensor (PIR) can detect infrared radiation, but only responds to fluctuations in intensity. An infrared temperature sensor measures the steady-state value of incident radiation.

OTHER RELATED COMPONENTS

- **passive infrared** motion sensor (see Chapter 4)
- **thermocouple** (see Chapter 25)

What It Does

Most temperature sensors discussed in other entries in this Encyclopedia are *contact sensors*, meaning that to measure the temperature of an object, liquid, or gas, they must make contact with it. In situations where contact is not possible or desirable, an **infrared temperature** sensor can be used. It responds to the *black-body radiation* (sometimes known as *characteristic radiation*) that is emitted by all materials above absolute zero (0 degrees Kelvin). This varies with temperature as a result of the movements of molecules.

Situations where noncontact sensors may be preferable to contact sensors include:

- An object is inconveniently located or too far away.
- The temperature of a large area must be measured.
- Contact with a small object would change the temperature of the object. The act of measurement would change the value being measured.
- The object is corrosive, abrasive, or otherwise liable to damage a sensor.
- The object is moving or vibrating.
- The surface of the object must not be contaminated (for example, unprotected foods).

- The temperature of the object is lower than around -50 degrees Celsius or higher than 1,300 degrees Celsius.

However, noncontact sensors have some limitations:

- Normally, only the surface temperature of a target can be measured.
- The optics of the sensor must be protected from dust, dirt, and liquid.
- The target must be clearly visible, in line-of-sight.
- Air pollution will degrade the temperature measurement. Some gases, such as carbon dioxide, will tend to absorb infrared radiation.
- The sensor will be affected by other sources, including reflected, transmitted, and convective heat.
- While an infrared sensor can theoretically respond to a very wide range of temperatures, in practice separate sensors of differing sensitivity are needed to cover a full range.
- Different types of materials emit differing intensities of black-body radiation, even if they are at the same temperature. Some compensation is necessary, or the surface of the object may have to be painted.

Applications

Handheld contactless thermometers were an early application for noncontact sensors.

In astronomy, the thermal radiation from the Sun and other stars is of interest to astronomers.

More recently, the declining cost of an infrared temperature sensor, and the ease of deploying it, have made it appropriate in consumer products. A significant area of adoption is in notebook computers and handheld devices, where processor performance must be balanced against the need to prevent the case from becoming too hot to hold comfortably. In this kind of application, gluing a sensor to the interior of the case would be a manual operation during the production process and would require a wired connection. An infrared temperature sensor mounted on the circuit board, viewing the underside of the case, can achieve the same objective more simply.

A contactless sensor is also very useful for measuring the temperature of rotating objects, such as heating rollers in a laser printer.

Schematic Symbol

No specific schematic symbol exists for an infrared temperature sensor.

How It Works

While nanometers (abbreviated nm) are generally used to measure visible wavelengths, the longer wavelengths of far-infrared are often measured in micrometers (abbreviated μm). The measurable infrared values are defined as ranging from 0.7μm to 14μm, corresponding to peak emissions from a black body ranging in temperature from 200 degrees Kelvin to 6,000 degrees Kelvin (about -70 to +5,700 Celsius).

Unfortunately an object does not emit just one wavelength of black-body radiation for each temperature value. It emits a spread of wavelengths that becomes wider as the temperature increases. However, the peak intensity also increases with temperature, when measured as *spectral radiance*, which is defined as watts per steradian, per micrometer of wavelength. (A steradian is the solid angle at the top of a cone, in this case the cone being of emitted radiation.) Because the intensity increases, it can be used to calculate the temperature.

Figure 28-1 illustrates this concept. Note that both of the axes have logarithmic scales.

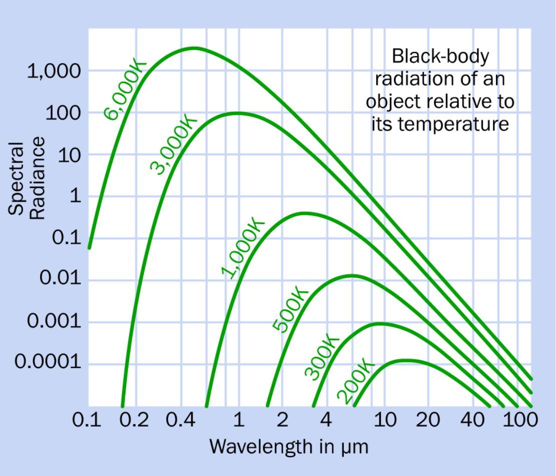

Figure 28-1 *The increase in intensity, and the widening spread of wavelengths, of black-body radiation emitted by one object at six different temperatures (in degrees Kelvin).*

All the curves are for one object. Each curve is specific to one temperature in degrees Kelvin, showing how the intensity of emitted radiation varies with wavelength. Note that radiation at wavelengths shorter than 0.7µm is within the visible spectrum; thus, objects at 1,000 degrees Kelvin, or hotter, may be seen to glow visibly.

Because of the wide variety of intensities and temperatures, an infrared sensor that is ideal for measuring a temperature of 1,000 degrees Kelvin will not provide an accurate result at 200 degrees Kelvin. The peak spectral radiance at 1,000 degrees is more than 10,000 times the peak at 200 degrees. Also, the curves in the figure are for an "ideal" object emitting pure black-body radiation. In reality, glass, plastic, and many other materials have a much lower *emissivity*, meaning that they emit less radiation, and are classified as *gray bodies*. A metallic object with a polished surface may emit one-tenth of true black-body radiation.

These issues cannot be ignored, but may be dealt with by relatively simple strategies. Infrared sensors will be rated according to their suitability for different temperature ranges, and the emissivity of the object being measured can be determined by consulting standard tables. Alternatively, the object can be spray-painted with special black paint (such as "Senotherm" or "3-M Black") that has a known emissivity of about 0.95 of pure black-body radiation. Alternatively, a specially formulated black sticker may be applied to the object that is being measured, so long as its temperature is within reasonable limits.

However, a basic infrared temperature sensor will not function reliably if it is pointed randomly at a variety of objects that vary widely in temperature. Specialized, expensive industrial devices incorporate compensation to deal with these issues, but they are outside the scope of this Encyclopedia.

Thermopile

A typical low-cost, chip-based infrared temperature sensor contains a *thermopile* consisting of multiple thermocouples etched into silicon and connected in series. The concept of a thermopile is illustrated in Figure 25-9, where a brief explanation is included.

The configuration of the thermopile is arranged so that the hot junctions of its thermocouples are all clustered in a small central area, where they receive incoming radiation through a window (often made of silicon) that is transparent to infrared wavelengths. The cold junctions are dispersed around the periphery, where they are shielded from incoming radiation. One way to visualize this is shown in Figure 28-2, although this is not a literal depiction of an actual sensor.

Instead of using alternating types of wire, as in a **thermocouple**, a chip-based thermopile often uses alternating segments of n-type and p-type silicon. The hot junctions are mounted on a thin film that has very little heat capacity, while the cold junctions are mounted on a thicker substrate that acts as a heat sink.

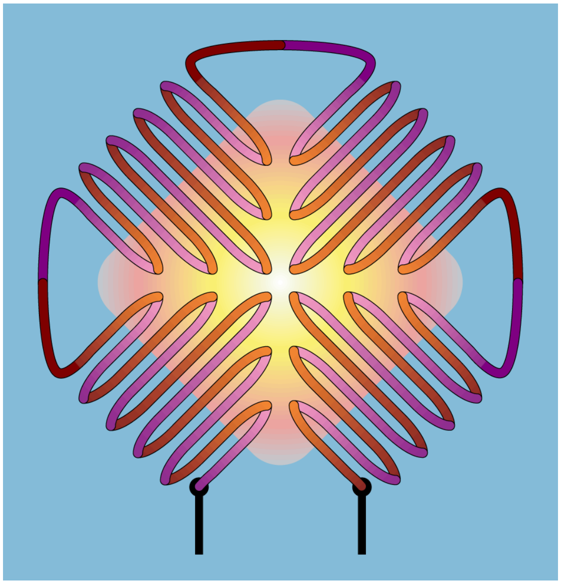

Figure 28-2 *Simplified diagram of the thermopile configuration inside an infrared sensor chip. Radiation arriving through a window in the chip affects the thermocouple junctions in the central region, while junctions around the edges remain at a lower temperature.*

Temperature Measurement

The voltage generated by the thermopile is related to the difference in temperature between the hot and cold thermocouple junctions. Thus there are three interrelated variables: hot temperature, cold temperature, and voltage. To calculate one variable, we must know the other two.

The hot temperature is what we wish to know. Therefore we must establish the voltage (which can easily be measured) and the cold temperature. The cold temperature can be determined by adding a thermistor inside the chip.

Typically an infrared temperature sensor with an analog output will have two pins that provide access to the internal thermistor, so that its temperature can be calculated from its resistance. Another two pins provide the voltage between the ends of the thermopile.

Interpreting and reconciling these values is not a trivial matter, especially bearing in mind that the thermistor has a negative temperature coefficient and a nonlinear output, and the thermopile will also have some nonlinearity. To simplify this situation, some infrared temperature sensors incorporate electronics to perform the necessary calculations and provide a digital output. This output can be converted to degrees of temperature by some fairly simple mathematical operations in an external microprocessor.

Variants

Two types of sensors are popular. One is surface-mounted, such as the TMP006, shown in Figure 28-3. This type generally has a digital output. The other type is a discrete component with four leads, such as the Amphenol ZTP135, shown in Figure 28-4. Discrete components may have either an analog or a digital output.

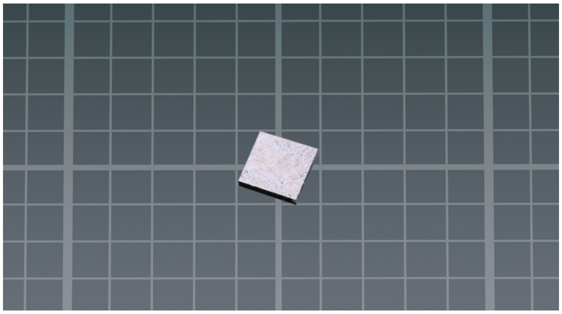

Figure 28-3 *A surface-mount infrared temperature sensor with digital output. Eight tiny solder pads are located on the underside. The background grid is in millimeters.*

Both types of sensors allow infrared light to enter through an area that is opaque to the visible spectrum but transparent to the appropriate range of wavelengths.

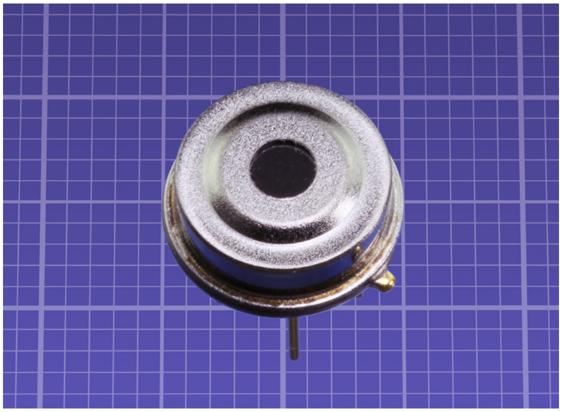

Figure 28-4 *A through-hole infrared temperature sensor with analog output. The background grid is in millimeters.*

The ZTP135 has an analog output shown in Figure 28-5.

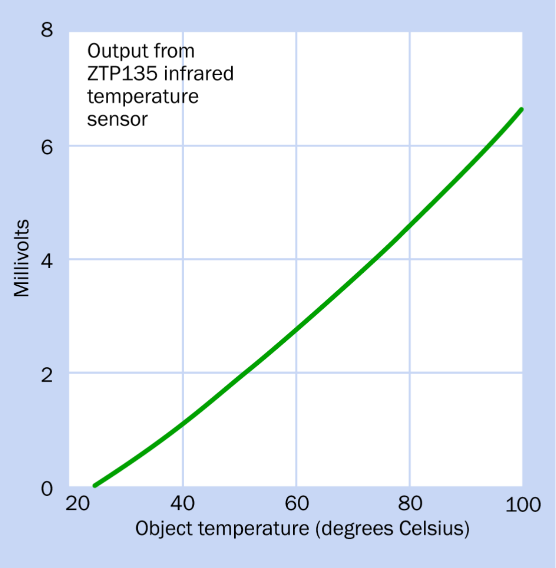

Figure 28-5 *Analog output from an infrared temperature sensor.*

The TMP006 is only about 1.5mm square, but is available on a breakout board from Sparkfun. Its successor, the TMP007, is available on a breakout board from Adafruit, shown in Figure 28-6.

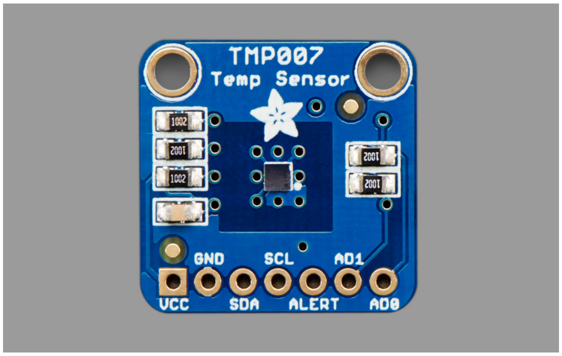

Figure 28-6 *The TMP007 sensor mounted on a breakout board from Adafruit.*

Surface-Mount Specifications

The TMP006 and TMP007 require a supply voltage that can range from 3.3V to 5V. These chips support the SMBus and I2C bus protocols, using a bus address that can be user-selected. An internal analog-to-digital converter uses 1 least-significant bit to represent 1/32nd of a degree Celsius, and data is saved as a 14-bit signed integer. Up to 16 temperature samples can be averaged internally.

Measurable temperature range is -40 to +125 degrees Celsius. Some hysteresis is built in. The TMP007 supports an alert mode if temperature falls above or below a user-specified threshold.

Sensor Arrays

Using multiple thermopile sensors arrayed in a line or a grid, with an array of lenses, it is possible to capture an image of temperature variations over a surface or a scene. This is known as *thermal imaging*. It can detect heat leakage from buildings, indicating poor insulation, or can locate hot spots in electronic circuits. Heimann Sensor has pioneered the miniaturization of a 31 x 31 grid of thermopile sensors in a single TO-8 or TO-39 package.

Values

Temperature Range

Chip-based infrared temperature sensors are typically designed for a range from about -20 degrees Celsius to about +125 degrees Celsius. Their peak sensitivity is to wavelengths between 4µ and 16µm.

Other types of infrared temperature sensors can have a much wider temperature range, but are more costly.

Field of View

Often referred to by its acronym, *FOV*, the field of view is the angle at the apex of an imaginary cone extending outward from the sensor, defining a boundary where the sensitivity diminishes below 50% of the value directly in front of the sensor. Greek letter ϕ may be used to represent the angle between the surface of the cone and the center line, while θ represents the angle between the opposite surfaces of the cone (i.e., 2 * ϕ). This is shown in Figure 28-7. θ is usually the angle defined as field of view.

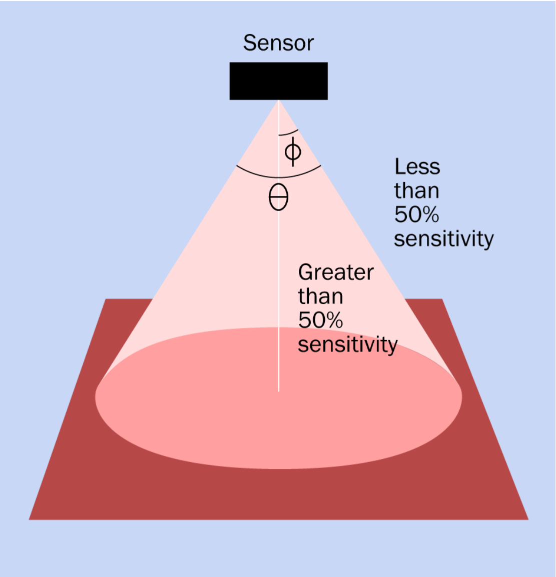

Figure 28-7 *Measurement of the field of view from a sensor, defined as the boundary of an imaginary cone where sensitivity drops below 50%.*

Because most infrared temperature sensing components do not have a lens, they are sensitive over a wide angle. The field of view is typically 90 degrees.

What Can Go Wrong

Inappropriate Field of View

The object being evaluated must fill the field of view of the sensor, to avoid measuring other objects around it.

Reflective Objects

A reflective object has lower infrared emissivity, and also may provide misleading output if the sensor is actually measuring thermal radiation reflected in the surface of the object in front of it. In a permanent installation, such as inside a device, the surface to be measured may have to be painted to reduce its reflectivity.

Glass Obstruction

Because glass is opaque to the spectrum of infrared that is of interest, temperature cannot be measured through a glass window. Silicon is opaque to visible wavelengths but is transparent to wavelengths longer than 2µm.

Multiple Heat Sources

Heat is transferred by convection, conduction, and radiation. While an infrared temperature sensor is designed to be sensitive to radiation, it will also respond to other heat sources. Warm or cold air currents will affect its response, and so will heat conducted through the material on which it is mounted. Careful placement of the sensor is important. A shield around the sensor, with a small hole in the center, can prevent convection, while correct location on a circuit board can minimize conduction.

Thermal Gradients

An infrared temperature sensor should be mounted in a stable environment where it will not be exposed to thermal gradients (one side being hotter than another). This asymmetry can cause inaccurate readings.

microphone

29

OTHER RELATED COMPONENTS

- **speaker** (see Volume 2)
- **headphone** (see Volume 2)

What It Does

The sensation of sound is created by rapid waves of air pressure impinging upon the eardrum. A microphone can convert these pressure waves into an alternating electrical signal that can be amplified, recorded, broadcast, transmitted through wires, and reproduced as sound by a headphone or speaker. The principle is illustrated in Figure 29-1. (For more information about sound reproduction, see the entries on **headphone** and **speaker** in Vol. 2.)

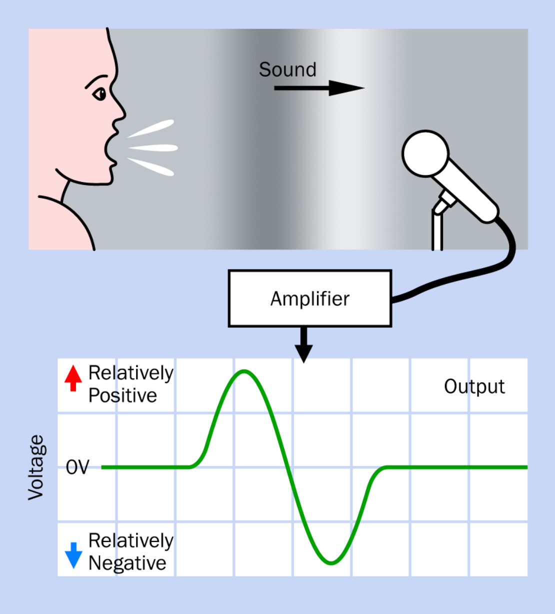

Figure 29-1 *The principle of converting pressure waves into an alternating electrical signal (adapted from an illustration in Make: More Electronics).*

Schematic Symbol

Various schematic symbols for a microphone have been used during the decades since its invention. A selection is shown in Figure 29-2. Each symbol assumes that sound is traveling from left to right. This is important when interpreting the symbol at top right, which can represent an earphone when it points in the opposite direction. Unfortunately, some schematics do not conform with this rule.

The two symbols at the bottom, showing a capacitor inside the microphone, should be reserved for condenser or electret microphones.

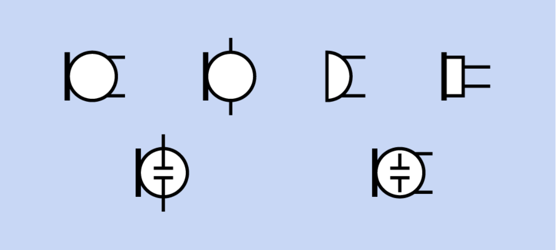

Figure 29-2 *A selection of schematic symbols that represent a microphone.*

How It Works

Some types of microphones generate a small voltage, while others have a fluctuating resistance that modulates a DC current.

Carbon Microphone

This was a very early attempt to reproduce sound. It contained carbon granules whose packing density increased and decreased in response to air pressure waves. When the density increased, the resistance between the granules diminished, and vice-versa. The principle is illustrated in Figure 29-3, and was patented by Thomas Edison in 1877 for use in telephones. As late as the 1950s (and even later in some countries), wired telephone handsets contained carbon microphones. Their bandwidth was extremely limited.

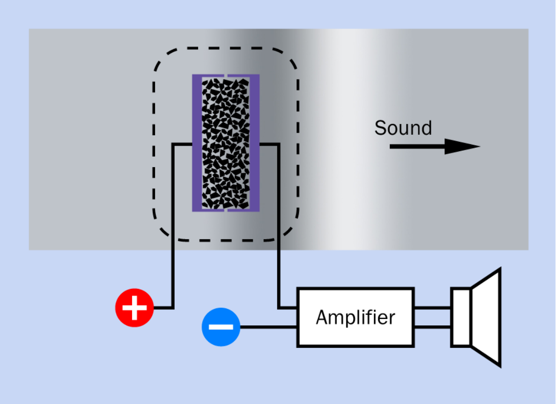

Figure 29-3 *The principle of a carbon microphone.*

Moving-Coil Microphone

Also known as a *dynamic* microphone, this consists of a very small, light coil of thin wire on a cylindrical tube that can vibrate along the axis of a permanent magnet. This principle is illustrated in Figure 29-4. A diaphragm is attached to the front of the tube, and responds to air pressure waves that penetrate the perforated enclosure of the microphone. Movements of the coil around the magnet create small alternating currents in the wire. The inertia of the coil, tube, and diaphragm, and the force needed to overcome the interaction between the coil and the magnet, impose a limit on the high-frequency response of this design.

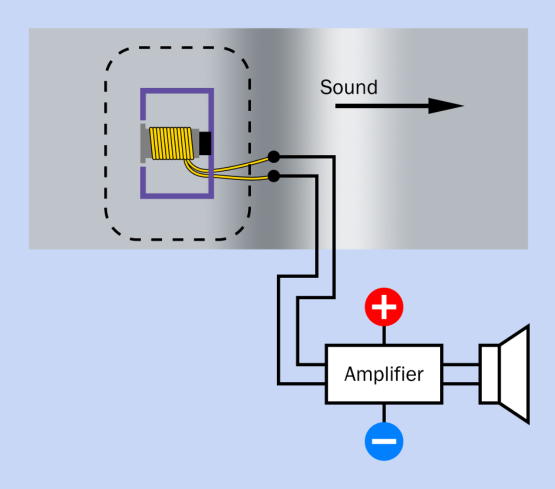

Figure 29-4 *The principle of a moving-coil microphone.*

Condenser Microphone

This type of microphone contains two thin discs or plates that form a capacitor. (In the early days of electricity, a capacitor used to be known as a condenser. The terminology has persisted for microphones.) An equal and opposite charge is applied to the plates. One plate is flexible, and as it responds to pressure waves, the capacitance between it and the other, rigid plate fluctuates. If the charge on the plates is kept approximately constant while the capacitance fluctuates, the voltage across the capacitor fluctuates also. These fluctuations can be amplified, as suggested in Figure 29-5.

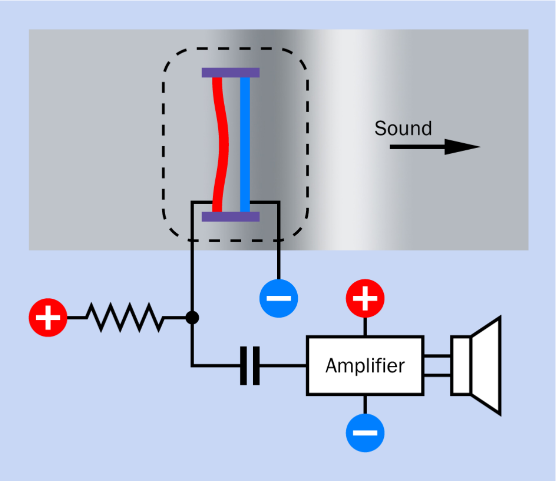

Figure 29-5 *The principle of a condenser microphone.*

Electret Microphone

This works on the same principle as a condenser microphone, except that its plates are made from a ferroelectric material that retains an electrical charge, just as iron will retain a magnetic polarization. The name of the microphone is derived from "electrostatic" and "magnet." While early electret microphones were of poor quality, they have evolved to rival condenser microphones, and are extremely affordable. Because the electret creates very small currents, it usually includes a transistor or op-amp in its package to boost the signal, and has an open-collector output. The basic circuit for an electret is shown in Figure 29-6. For more information about using an open-collector output, see Figure A-4.

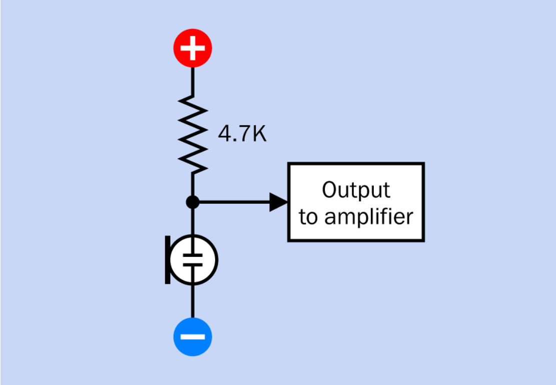

Figure 29-6 *The basic circuit for using an electret microphone.*

An example of a low-cost electret microphone is shown in Figure 29-7. This type of component is sold either with leads attached, or with solder pads.

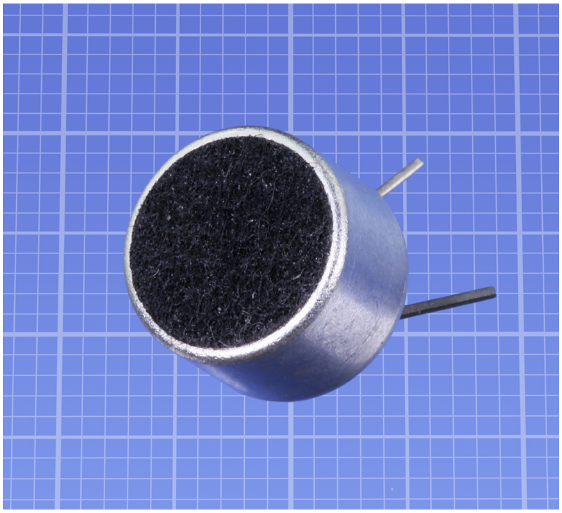

Figure 29-7 *A generic electret microphone. The background grid is in millimeters.*

MEMS Microphone

This type, often used in mobile phones, is a capacitive device that works on the same principle as a condenser microphone, although the component is etched in silicon and has a diaphragm that measures only about 1mm square. Many MEMS microphones have an analog output that is amplified in the same chip. Others have a digital output, using *PDM encoding*. This reduces the analog signal to a very fast bit stream, in which the density of the bits represents the amplitude of each fluctuation in a sound wave. PDM is an acronym for *pulse density modulation*. It requires an external clock signal to time the bit stream.

A breakout board from Sparkfun, on which is mounted an Analog Devices ADMP401 MEMS microphone with a preamplifier, is shown in Figure 29-8.

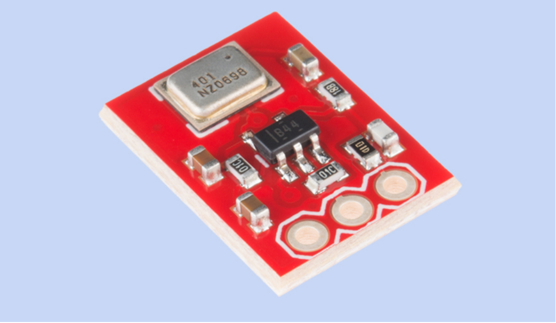

Figure 29-8 *A breakout board for a MEMS microphone (the metal-clad rectangular component at the far end).*

Piezoelectric Microphone

This has also been known as a *crystal* microphone. It contains a diaphragm that functions as a transducer. When it flexes in response to pressure waves, the mechanical energy is transformed into a small amount of electrical energy. Piezoelectric microphones were replaced by the moving-coil type in domestic audio devices when vacuum tubes were replaced with transistors, but may still be used as contact microphones to amplify acoustic musical instruments, or to trigger the playback of digitally sampled musical sounds.

Other variants include *ribbon* microphones (which were common in recording studios in the 1950s and 1960s, but have become rare), *laser* microphones, and *fiber-optic* microphones. They are not sufficiently common to be included in this entry.

Values

Sensitivity

Sound pressure is a complicated topic, explored in detail in the **transducer** entry in Volume 2. It can be measured in pascals, where 1 pascal = 1 newton per square meter.

The *sound pressure level* is a different concept. It measures the *relative* intensity of a sound, in a logarithmic scale calibrated in *decibels* (abbreviated dB). The reference value for this relative scale is 20 micropascals, considered to be the threshold of human hearing, comparable to a mosquito three meters away. This is assigned the value of 0dB.

From this point upward, the actual sound pressure doubles for each additional 6dB. A table of noise sources and their approximate decibel values is shown in Figure 29-9. This is derived from averaging eight similar tables, which are not always consistent in their estimates. It is an approximate guide only.

Decibels	Noise Example
140	Jet engine at 50 meters
130	Threshold of pain
120	Loud rock concert
110	Automobile horn at 1 meter
100	Jackhammer at 1 meter
90	Propeller plane 300 meters above
80	Freight train at 15 meters
70	Vacuum cleaner
60	Business office
50	Conversation
40	Library
30	Quiet bedroom
20	Leaves rustling
10	Calm breathing at 1 meter
0	Auditory threshold

Figure 29-9 *Decibel values for some common sound sources. Reproduced from Volume 2 of this Encyclopedia.*

The decibel unit is important when understanding the specifications of microphones, because it is used to measure their response. Microphone sensitivity is established with a standard input sine wave of 1kHz in frequency and 94dB in intensity (equivalent to 1 pascal in actual sound pressure), measured at the microphone. The sensitivity of an analog microphone is then defined as the number of decibels in an output signal of 1V. Because the output is an AC signal, voltage is measured as a root-mean-square (RMS) value.

For digital microphones, sensitivity is measured as the decibels that can be reproduced by a full-scale digital output. This value is abbreviated as dBFS.

Directionality

A microphone that has a directional response is desirable in many situations. Often, for example, sounds in front of the microphone are more important than sounds from behind the microphone. The directionality of a microphone (sometimes referred to as its *directivity*) is usually represented with a *polar graph* in which the microphone is seen from above, and its sensitivity to sounds from various directions is shown with a curve such as those in Figure 29-10. The circles are drawn at intervals of 5dB, with 0dB at the periphery and -30dB at the center. The precise response for an individual microphone should be shown in its documentation.

Frequency Response

Every microphone tends to be more sensitive to some sound frequencies than others. A manufacturer will provide a graph showing this sensitivity, in decibels, plotted against sound frequency on a logarithmic horizontal axis. Theoretically, human hearing extends from around 20Hz to 20kHz, but few people are actually capable of hearing the high end of that range, and 15kHz may be a more realistic limit for a young person, diminishing to 10kHz with middle age.

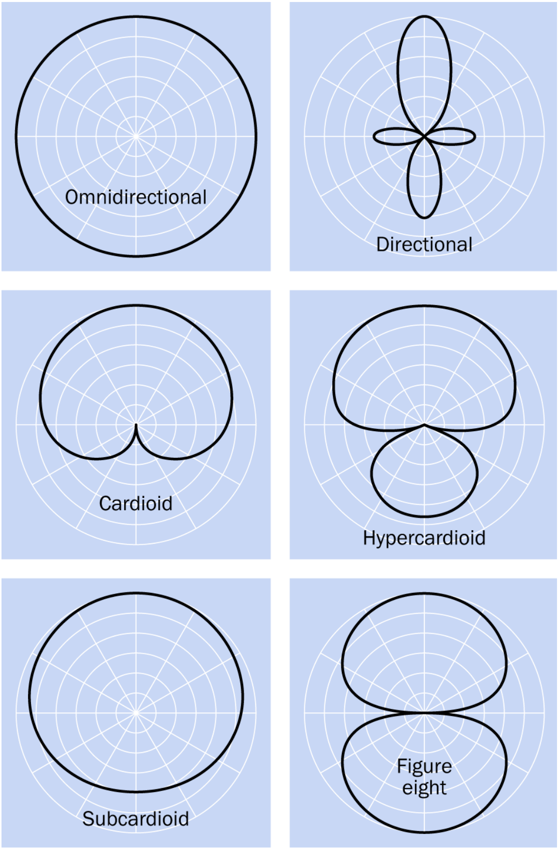

Figure 29-10 *Six generic sensitivity patterns. Individual microphones will deviate somewhat from these generic curves.*

An ideal flat response would show that a microphone is equally sensitive to all frequencies. In reality, *rolloff* usually occurs at low frequencies, and will eventually occur at high frequencies, although it may be preceded by a rise in response. If the central section of a curve is flat within plus-or-minus 1dB, this is a level of performance that was attained only by expensive studio microphones in the past. Electrets and MEMS microphones can now provide equivalent frequency response for $1 or $2 apiece, as opposed to the hundreds or even thousands of dollars that used to be necessary for professional equipment.

The response curve shown in Figure 29-11 is for the eMerging i436, an electret microphone sold in a module as an accessory to enable high-quality recordings on handheld devices. The rise around 15kHz may have been introduced deliberately by the manufacturer to compensate for reduced sensitivity of the human ear in that range.

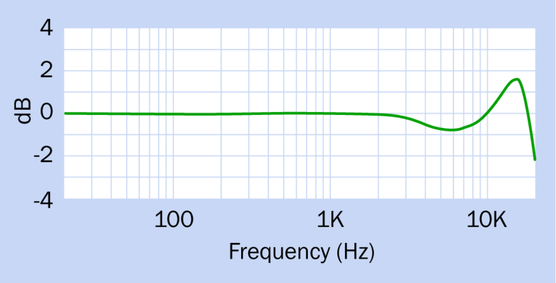

Figure 29-11 *Frequency response for an electret microphone.*

Impedance

The impedance value for a microphone is a function of its resistance, capacitance, and inductance. An amplifier input will also have an impedance rating, and for ideal power transfer between microphone and amplifier, the impedance values should be identical. However, a more important consideration in audio equipment is to avoid voltage loss between the output device (in this case, the microphone) and the input device (the amplifier). To achieve this, the output device should have a low impedance, while the input device should have a high impedance. Most microphones are rated at 150 to 200 ohms, while an amplifier may be rated at 1.5K to 3K.

Total Harmonic Distortion

When an audible sine wave is converted to an electrical output by a microphone (and by its preamplifier, if one is included in the module), the output may become corrupted by some multiples of the basic frequency. These are known as *harmonics*, and they are considered as a distortion of the signal. Total harmonic distortion, measured by a spectrum analyzer over the entire frequency range, should ideally be less than 0.01%.

Signal-to-Noise Ratio

Often abbreviated as S/N or SNR, the signal-to-noise ratio in a microphone is measured in decibels, and should be 60dB or higher.

What Can Go Wrong

Cable Sensitivity

Audio amplification is always vulnerable to electrical noise, which tends to be amplified along with the signal. Small signals from microphones require the use of shielded cables to reduce hum and other types of interference.

Noisy Power Supply

For similar reasons, power supplies must be as free as possible from voltage spikes and other fluctuations.

current sensor

30

The entry describes components that can be installed to monitor current on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

A *current transformer* may be used to measure current, but is not included in this Encyclopedia.

OTHER RELATED COMPONENTS

- **voltage** sensor (see Chapter 31)

What It Does

A current sensor measures the flow of electricity through a wire or a device, and supplies an output that can be interpreted either visually or by a microcontroller to provide a reading in amperes or fractions of an ampere.

Applications

Current sensing is important in industrial applications such as the control of high-powered motors. It can be used to monitor the performance of an inverter, or for everyday purposes such as monitoring the long-term power consumption of an appliance. During product development, a current sensor can indicate the power consumption of a circuit as it changes with modifications.

This entry describes three methods to measure current: an ammeter, series resistor, and Hall sensor. While other methods exist, they are outside the scope of this Encyclopedia.

Ammeter

An ammeter that is sold as a standalone device with leads for circuit testing is often described as a *test meter*. Its functionality is usually built into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

An ammeter designed for permanent installation in a device or prototype is a type of *panel meter*, such as the one shown in Figure 30-1. This traditional-style analog meter may be less expensive than the many digital types that are available. It uses a magnetic field created by current flowing through a coil to pull a needle across a scale, against the force of a spring.

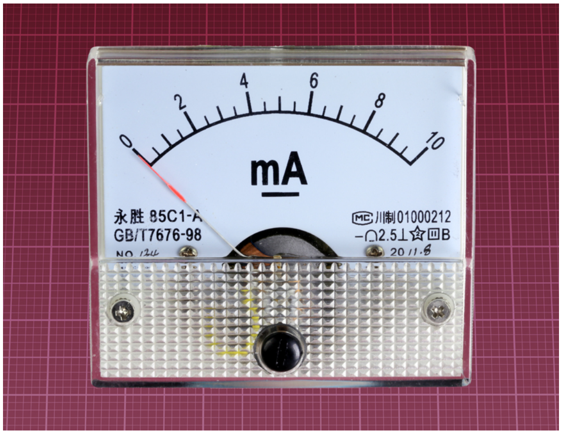

Figure 30-1 *A traditional-style analog ammeter.*

A digital ammeter allows a wider range of values to be viewed more easily. The meter from Adafruit in Figure 30-2 has a range of 0A to 9.99A, at voltages from 4.5VDC to 30VDC. The meter can be powered parasitically from the currents that it is measuring, or can use a separate isolated 5VDC supply.

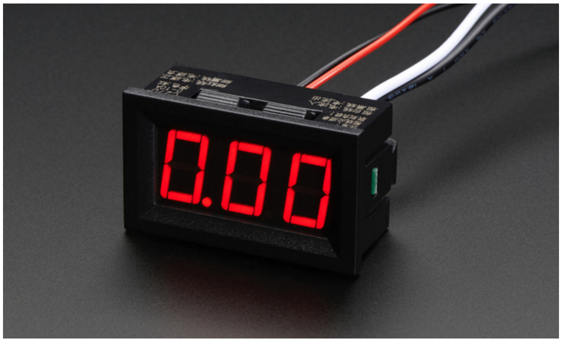

Figure 30-2 *A panel-mount digital ammeter that can measure up to 9.99A.*

Schematic Symbol

An ammeter may be represented in a schematic with the letter A inside a circle, as shown in Figure 30-3.

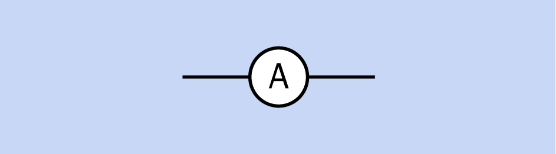

Figure 30-3 *An ammeter may be represented like this in a schematic.*

Ammeter Wiring

Two ways to use an ammeter in a circuit are illustrated in Figure 30-4, where the load may be any equipment, device, or component that provides some electrical resistance. Because current is the same at all points in a simple circuit, the current that the meter measures flowing through itself will be the same as the current flowing through the load, and the sequence of components is immaterial.

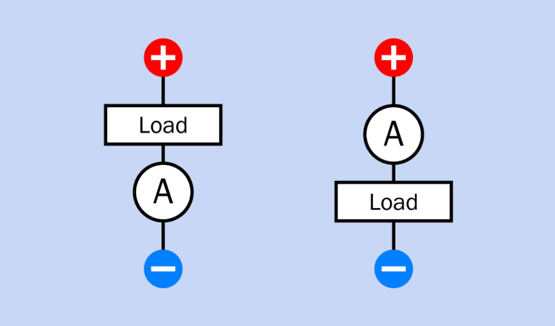

Figure 30-4 *Two options for placement of an ammeter in a circuit.*

However, regardless of the placement of the meter, the process of measuring current will inevitably change the value of the current being measured. This is because the ammeter imposes some internal resistance of its own. The resistance is extremely low, and may be considered negligible for loads of more than a few ohms.

- The low internal resistance of an ammeter means that it must never be connected in parallel with a load, or directly across a power source.

A disadvantage shared by analog and digital meters is that they are not usually interchangeable between AC and DC.

Series Resistor

The current flowing through a load can be calculated by measuring the voltage across a series resistor that is inserted between the load and its ground connection. The concept is illustrated in Figure 30-5.

Using Ohm's law, if U is the voltage drop, I is the current, and R is the value of the resistor:

```
I = U / R
```

This tells us that for a fixed value of R, current is proportional to voltage. Therefore, measuring the voltage enables calculation of the current, so long as the value of R is known.

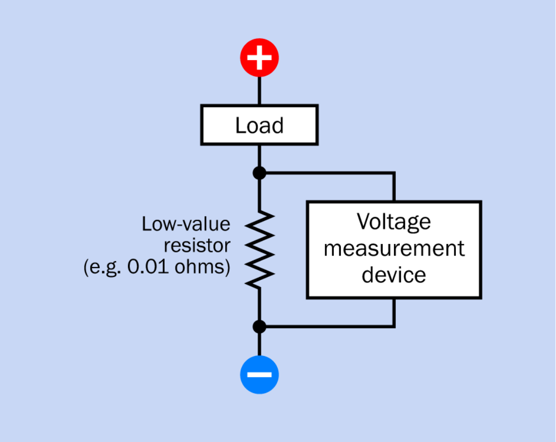

Figure 30-5 *A basic circuit for measuring current as a function of the voltage drop across a small-value resistor. The load in this figure is any circuit or device with a relatively higher resistance. The voltage measurement device could be a microcontroller or analog-to-digital converter.*

Suppose that R has a very small value, which is trivial compared with the resistance of the load. Consequently, the current in Figure 30-5 will be mainly determined by the load, and we may consider the value of the current to be almost the same with or without the addition of R. In that case, the voltage drop across the resistor will be smaller if R is smaller. A smaller voltage drop will not be as easy to measure, but a lower resistance will result in less wastage of power.

If P is the power:

$$P = R * I^2$$

An example may help to make this clear. Suppose the resistor has a value of 0.5 ohms, and the voltage drop across it is measured to be 1V. Ohm's law shows that current flow is 1 / 0.5 = 2A. The power formula shows that P = 0.5 * 4 = 2W.

To waste less power, the value of the resistor should be reduced further. Suppose a resistor of 0.01 ohms is used, and the voltage drop across it is measured as 0.02 volts. The current is 0.02 / 0.01 = 2A, as in the preceding example, but the power dissipation is now only 0.01 * 4 = 0.04W, which is negligible.

But are resistors available that have values measured in fractions of an ohm?

Current-Sense Resistors

In fact, many *current sense* resistors are available, with values of 0.1 ohms, 0.001 ohms, 0.0001 ohms, and many in between. Some resistors have values measured in micro-ohms. Examples are shown in Figures 30-6, 30-7, and 30-8.

Measuring a small voltage drop is easily done by using a microcontroller. However, the connection to the microcontroller must be made as near to the resistor as possible, to eliminate the additional resistance of wires or circuit-board traces. For this reason, precision current-sense resistors may be equipped with four terminals. Two are wider, and are meant for connection to the flow of current. The other two are narrow, for measuring the voltage over the resistor. With this *4-point* configuration, voltage drop over the resistor can be measured as close to it as possible. The 0.001-ohm surface-mount resistor in Figure 30-8 is designed for 4-point measurement.

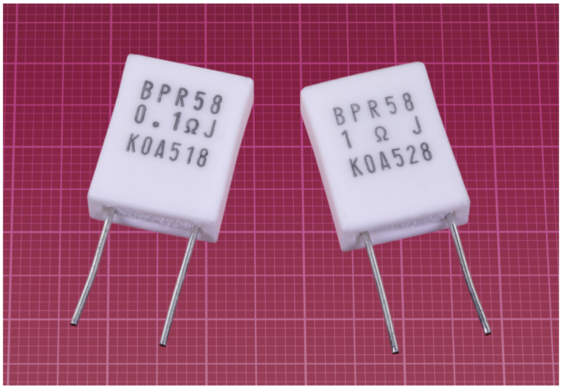

Figure 30-6 *Two current resistors manufactured by KOA Speer. Left: 0.1 ohms, 5W, 5%. Right: 1 ohm, 5W, 5%. The background grid is in millimeters.*

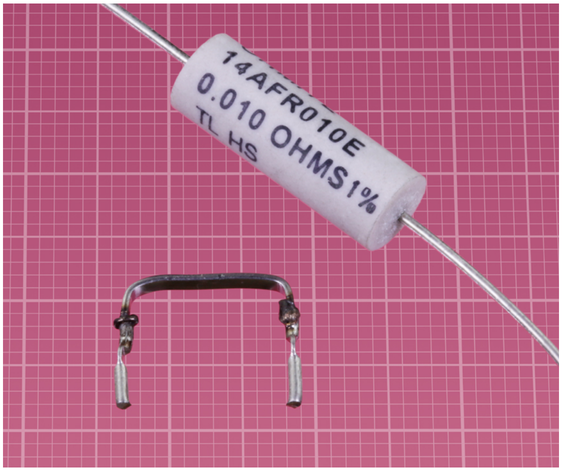

Figure 30-7 *Two styles of current resistors rated for 0.01 ohms. Bottom left: A plug-in version from TT Electronics, rated 1W and 5%. Top right: Ohmite, 4W, 1%. The background grid is in millimeters.*

Resistors that have the lowest values and are intended to tolerate high current may consist of just a metallic strip welded to solderable pins. This type of component is sometimes called an *open-air resistor*. It is commonly used in multimeters, for measuring currents up to 10A or above.

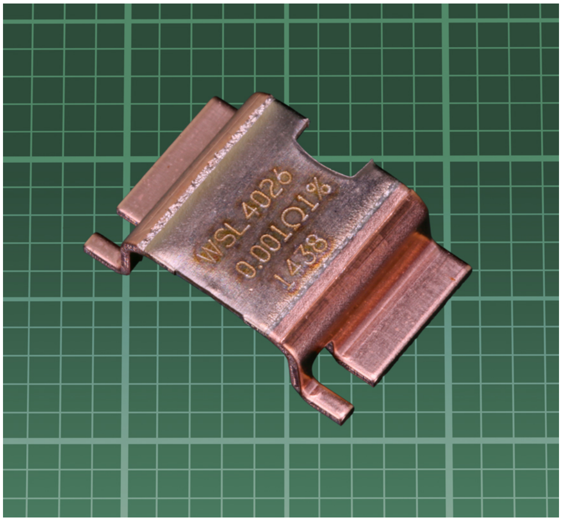

Figure 30-8 *This Vishay 4-terminal surface-mount current resistor is rated for 0.001 ohms, 3W, 1%. The background grid is in millimeters.*

Voltage Measurement

Some chips are designed for amplifying the voltage drop across a current-sensing resistor. An example is the Texas Instruments INA169.

A few chips contain an analog-to-digital converter in addition to the amplifier. The INA219 by Texas Instruments is designed to measure voltage as well as current, on the "high side" of a circuit—that is, between the positive power supply and the power input of the circuit. It makes its digital data available over an I2C bus.

For additional details about protocols such as I2C, see Appendix A.

Measuring current from the voltage drop across a series resistor offers the advantages of simplicity, ability to work with AC or DC, and low cost (although some resistors of extremely low value can be relatively expensive). A possible disadvantage is that the measurement circuit is not isolated from the circuit whose current is being measured.

Hall-Effect Current Sensing

The principle of a Hall-effect sensor is explained in the entry for **object presence** sensors. See "Hall-Effect Sensor". Normally this type of sensor is activated by an external permanent magnet, but it can also react to the magnetic field generated by current flowing through a wire.

Because the field generated around a wire is proportional with the current, the analog output voltage generated by a linear Hall-effect sensor can also be proportional with the current.

Hall-effect sensors for this specific purpose are available in 8-pad surface-mount packaging. The current to be measured passes through a copper conductor that is embedded in the chip. An example is Allegro's ACS712, for AC or DC currents up to 30A. The internal resistance of the current path through this chip is stated as 1.2 milliohms, and the path is isolated from the sensing circuitry.

Three variants of the chip are available, for currents of plus-or-minus 5A, 20A, and 30A. Depending on which variant is used, the output will range from 66mV to 185mV for each increase of 1A in the current path. Because the current path is isolated, the chip requires a separate power supply of 5VDC.

The 5A version of the AVS712 can be bought on breakout boards from Sparkfun. In Figure 30-9, the board at top-left contains only the ACS712, while the board at lower-right adds an op-amp with sensitivity control, to amplify the voltage output when measuring small currents.

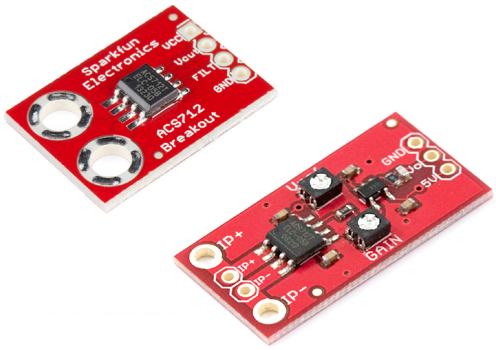

Figure 30-9 *Breakout boards using the AVS712 Hall-effect current sensor. The one at lower-right includes an op-amp to amplify small signals.*

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa. Erroneous readings or damage to the meter may result.

Magnetic Interference

A disadvantage of Hall-effect current sensing is that the sensor can be affected by stray magnetic fields. Because a Hall-effect chip is responding to very small magnetic effects, it is vulnerable to interference. Consult the manufacturer's datasheet carefully regarding correct placement of a chip on a circuit board.

Incorrect Meter Wiring

The correct wiring of an ammeter is in series with a load, not in parallel with a load. This may seem an elementary error, but is easy to make if an ammeter and a volt meter are both being used, and they look similar.

Because some panel meters are not fused, applying an ammeter directly across a power supply, without any series resistance, may result in immediate and hazardous destruction of the meter.

Incorrect wiring can also occur when using a digital meter that has four wires emerging from it: two for current testing and two for a separate power supply. This issue is especially important when measuring substantial currents (above 1A).

Current Out of Range

Attempting to measure a current that exceeds the range of an ammeter may damage the meter or blow its internal fuse, if it has one.

voltage sensor

31

The entry describes the type of component that can be installed to monitor voltage on an indefinite basis. It does not include test equipment, such as test meters or multimeters.

OTHER RELATED COMPONENTS

- **current** sensor (see Chapter 30)

What It Does

A voltage sensor measures the electrical potential between any two points in a circuit, or the voltage supplied by a power source, and provides data in volts or fractions of a volt. It should not be confused with an *analog-to-digital converter*, which is not a sensor in itself, but can process the voltage output from a sensor by digitizing it. More information about analog-to-digital conversion is in the Appendix. See Appendix A.

Applications

Voltage measurement is important in conjunction with all types of power supplies, to verify their performance. A volt meter may also be used to show the output from various types of analog sensors that have voltage output.

A graphical display may be used in audio equipment to indicate signal level, which can be proportional with voltage.

Volt Meter

A volt meter that is sold as a standalone device with leads for circuit testing is often described as a *test meter*. Its functionality is usually built into a *multimeter*. Test meters and multimeters are outside the scope of this Encyclopedia.

A volt meter designed for permanent installation in a device or prototype is a type of *panel meter*, which is described here.

An antique analog panel meter is shown in Figure 31-1.

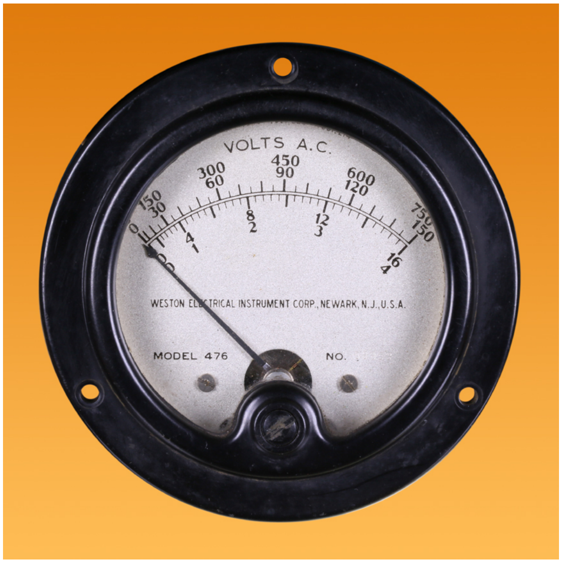

Figure 31-1 *An antique analog volt meter.*

The four scales on the dial correspond with separate input terminals at the rear of the unit. The unequally divided scales are a simple way to compensate for the nonlinear response of the mechanical movement inside the meter.

Modern analog volt meters are still manufactured, and may be cheaper than their digital equivalents. However, a digital volt meter allows a wider range of values to be viewed more easily. The meter in Figure 31-2, sold as a low-cost battery tester, measures voltages from 4VDC to 13VDC to an accuracy of two decimal places. It needs no separate power supply.

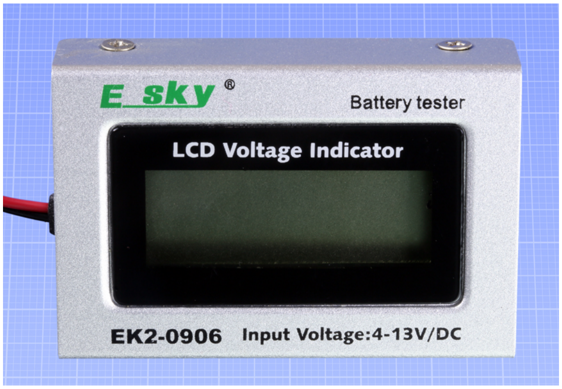

Figure 31-2 *A panel-mount digital volt meter.*

Sometimes the number of digits in a volt meter is specified as ending with one-half, as in 3.5 or 3-1/2 digits. This means that the most significant (leftmost) digit can only be a 1 or a blank. The extra “half digit” may seem inconsequential but doubles the range of displayable values. For example, a 2-digit display can only show 100 values, from 0 to 99. A 2-1/2-digit display can show 200 values, from 0 to 199.

Schematic Symbol

A volt meter may be represented in a schematic with the letter V inside a circle, as shown in Figure 31-3.

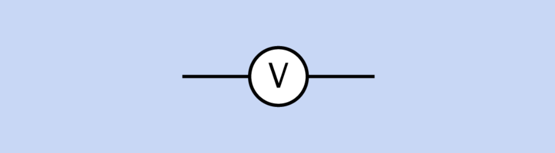

Figure 31-3 *A volt meter may be represented like this in a schematic.*

Volt Meter Wiring

Two ways to use a volt meter in a circuit are illustrated in Figure 31-4, where the load may be any equipment, device, or component that provides some electrical resistance.

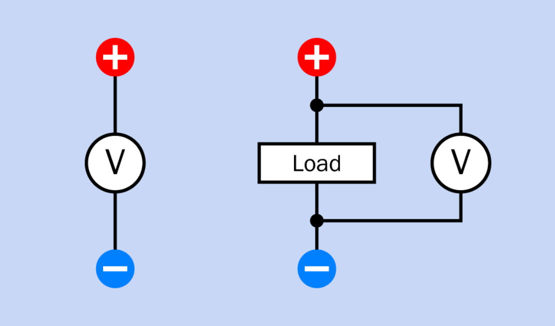

Figure 31-4 *Two options for placement of a volt meter.*

On the left side of the figure, a volt meter may be connected directly across a power source, because the internal resistance of the meter is so high, it will draw very little current. In this configuration, the meter measures the voltage of a source when the source is virtually unloaded.

On the right, the meter can measure the voltage drop across a load, or can measure the voltage across any component, or group of components, that are a subset of the load.

A disadvantage shared by dedicated analog and digital volt meters is that they are not usually interchangeable between AC and DC.

How It Works

An analog volt meter usually contains a built-in, high-value fixed resistance, and an ammeter measuring the current passing through it. The current sensed by the ammeter is then converted to a reading in volts.

Suppose that R represents the fixed resistance, U is the voltage drop across the resistance, and I is the current flowing through it. Ohm's Law tells us:

```
U = I * R
```

The formula shows that if the resistance is fixed, the voltage will vary in proportion with the current, and therefore can be calculated from it.

Load-Related Inaccuracy

When the meter is measuring the voltage drop created by a load that has high resistance comparable to the internal resistance of the meter itself, the meter will not give an accurate reading. Figure 31-5 illustrates this problem.

On the left side of the figure, two resistors 10M each are wired in series between a 9VDC power source and negative ground. Because the resistors are equal, each of them imposes an equal voltage drop of 4.5V.

On the right side of the figure, an additional 10M resistance has been added in parallel with the lower resistor. When two resistors of values R1 and R2 are wired in parallel, their total resistance, R, is given by this formula:

```
1 / R = ( 1 / R1 ) + ( 1 / R2 )
```

Therefore, the total resistance in the bottom half of the circuit is now 5M instead of 10M, and the voltage drop across the upper resistor becomes twice the voltage drop across the two lower resistors.

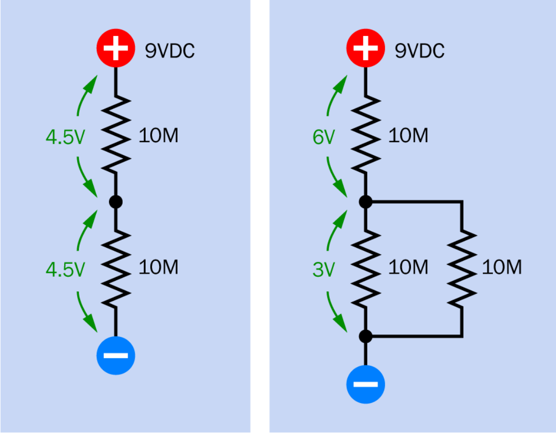

Figure 31-5 *If a meter measuring the voltage drop across a load has an internal resistance comparable to that of a load, the meter will not give an accurate reading. See text for details.*

Now suppose that the additional 10M resistor is actually the fixed resistance inside a volt meter. In fact, many volt meters do have an internal resistance of around 10M. Because the meter has reduced the resistance in the bottom half of the circuit by a factor of 2, it measures the voltage drop as 3V. If the meter had an ideal, infinite resistance, it would give the correct value of 4.5V. In the real world, that is impossible.

Bar Graph

Sometimes it is useful to represent voltage with a graphical display. A *bar graph* component makes this possible, and is often used in audio equipment.

The bar graph can consist of a row of *LEDs*. In a component designed for this purpose, there are usually 10 or more. To represent 0 volts, all the LEDs remain dark. More LEDs are illuminated as the voltage increases.

Because the bar graph itself only contains LEDs, a driver is necessary to convert voltage into the "rising thermometer" effect. Examples are the LM3914 (linear), the LM3915 (logarithmic, 3 dB per step), and the LM3916 (using VU or Volume Units, for audio). Each of these components has 10 LED outputs, and is fabricated using a chain of resistors as a multitap voltage divider, with 10 comparators.

A driver can be set to show only one LED corresponding with the current voltage level, or (more commonly) a cumulative string of LEDs from zero upward. Some bar graphs also contain different colored LEDs, such as green for the first 7, yellow for the next 2, and red for the final one.

A microcontroller containing an analog-to-digital converter can be used to illuminate a bar graph, instead of a driver chip.

An example of a bar graph is the Avago HDSP-4830, shown in Figure 31-6.

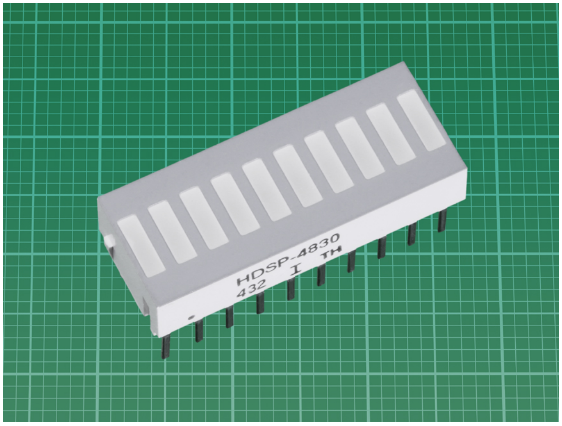

Figure 31-6 *A bar graph LED display, which can be used to represent voltage. The background grid is in millimeters.*

What Can Go Wrong

Confusing AC with DC

A panel meter that is designed only to measure DC should not be used with AC, and vice-versa. Erroneous readings or damage to the meter may result.

High Circuit Impedance

When a meter is measuring the voltage drop across a high-value resistance, it will give an inaccurate reading, as it is diverting a significant proportion of the current.

Voltage Out of Range

Attempting to measure a voltage that exceeds the limited range of a meter can damage the meter or blow its internal fuse, if it has one. Attempting to measure a voltage that is at the very low end of a meter's range will result in an inaccurate value.

Voltage Relative to Ground

If one input to a digital meter is connected with ground, the meter will only be able to measure voltage in a circuit relative to ground.

Sensor Output

This appendix provides some basic information about nine forms of sensor outputs, and the ways in which they can be processed. Other types of encoded output exist, but those examined here are the ones most likely to be found.

Figure A-1 provides an overview. Every sensor initially creates an analog output, which is sometimes connected directly to an output pin. In thermistors and photoresistors, for example, the internal resistance of the component constitutes its output. In many sensors, however, the behavior of the sensing element is processed internally to create voltage, open collector, encoded pulse stream, or current output.

If the sensor is chip-based, it may process the analog sensor response internally to create a binary output or a digital output.

In this Encyclopedia, the term "binary output" means "an output that has two states," usually logic-low or logic-high. The states may be accessible via an output pin, or may be processed internally to create a pulse stream. In that case the stream fluctuates between the two states as a way of encoding an analog value with pulse-width modulation (PWM) or frequency. Other types of encoding are also possible, but are unusual.

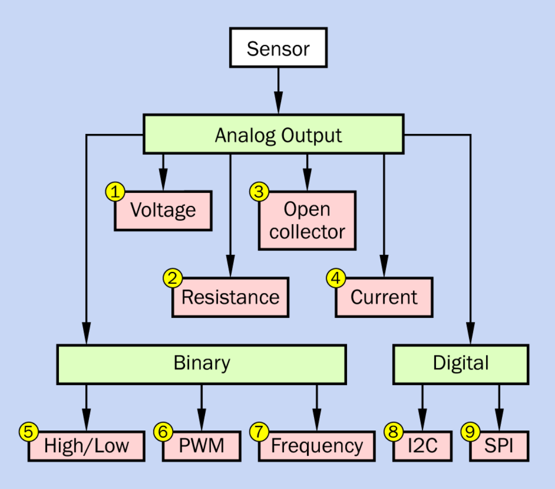

Figure A-1 *Nine possible types of output from sensors. The primary categories are highlighted in green.*

The term *digital output* is used here to mean one or two bytes of data that are stored in a register (memory location) in the sensor chip. While other forms of output are "always on" and can be accessed at any time from an output pin, a digital output is usually not available until an external device, such as a microcontroller, sends an instruction to the sensor chip, telling it to return the data. This two-way communication usually is handled by the I2C or SPI protocols (other protocols exist, but are less common).

Analog Outputs

Voltage is by far the most common form of analog output. Other forms of analog output can easily be converted to a voltage value, using the simple techniques described here.

1. Analog: Voltage

Direct Connection: Analog-to-Analog

An analog voltage output can be connected directly to an analog input, so long as the range is compatible and the sensor can provide sufficient current. Examples of external analog devices would be an analog volt meter, a light source or sound source that changes intensity, or a transistor or op-amp that will amplify the output for other audio/visual purposes.

If the voltage output from the sensor rises above a usable range, it can be converted to a lower value by applying it across two resistors connected in series to form a voltage divider. This is shown in Figure A-2.

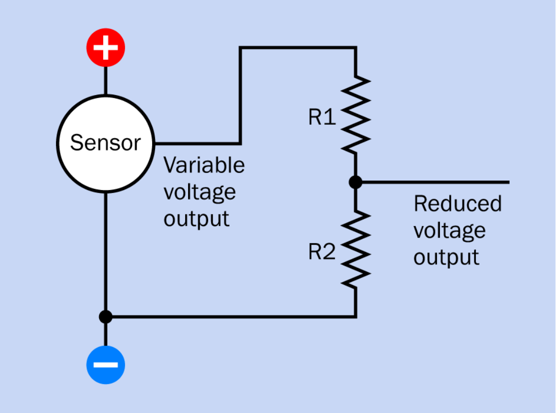

Figure A-2 *Using a voltage divider to reduce the range from a sensor.*

The values of R1 and R2 can be derived from this basic formula, where V_{SEN} is the voltage from the sensor and V_{OUT} is the output from the voltage divider:

Example A-1

$$V_{OUT} = V_{SEN} * (R2 / (R1 + R2))$$

The impedance of the device used for voltage sensing must be high relative to the values of R1 and R2. Note also that if the analog voltage out from the sensor varies linearly with the phenomenon being sensed, this relationship is likely to be disturbed by the voltage divider.

Analog-to-Binary Conversion

The term "binary" is used here to mean an output that can be in one of two states, such as logic-high and logic-low.

A varying analog voltage output can be simplified by passing it through a component that transforms the signal into binary form. This may be done by using a logic chip with a *Schmitt trigger* input, a *zener diode*, or a **comparator**. (For a description of comparators, see the entry in Volume 2.) A comparator provides desirable features such as adjustable positive feedback to create hysteresis. It may be used, for example, to convert a slowly changing signal from a phototransistor while the sun is setting, to a high/low output that can activate a relay to switch on a light.

Analog-to-Digital Conversion

The analog voltage output from a sensor can be digitized by an external analog-to-digital converter (ADC), either inside a microcontroller or by a separate ADC chip.

If a microcontroller is used, a sensor may often be wired directly to an input pin that connects internally with an ADC. A program in the microcontroller can then assess the integer output from the ADC and either execute a conditional statement or convert the value to a format appropriate for another device, such as a digital display.

If an ADC chip is used, there are thousands to choose from. A few basics:

- A *flash* converter contains a row of comparators with different reference voltages generated with a chain of equal resistors. The comparator outputs are fed into a priority **encoder** that outputs a binary number. This system is very fast but has limited resolution.

- A *successive approximation* converter uses a single comparator, comparing the input voltage with the output from a DAC. The binary number that is supplied to the DAC is determined one bit at a time, from the most significant to the least significant bit, using the comparator's result to determine if the bit should be 0 or 1. These bits are stored in a register, called a successive approximation register (SAR). When the process finishes, the SAR contains a binary representation of the input voltage. This type of ADC can achieve high resolution (many bits) at the cost of lower conversion speed.
- In a *dual slope* converter, a capacitor is charged for a fixed time, with a rate proportional to the input voltage, then discharged at a known rate while measuring time by counting clock pulses. The resulting count is the ADC output.

  The name of the converter is derived from the voltage on the capacitor.
- A *voltage-to-frequency* converter uses a voltage-controlled oscillator to produce pulses with a frequency proportional to the input voltage. If the pulses are counted over a fixed time interval, the count is proportional to the signal level.

The number of the bits in the output from an ADC must be sufficient to digitize the input voltage range with desired accuracy. Because the voltage range may contain unexpected peaks, a cautious strategy is to use many more bits than are necessary. However, this means that for most of the time, only a few bits will be used to represent the low end of the voltage range, and accuracy will suffer.

For example, suppose a voltage input normally ranges from 0V to 2V, with occasional brief excursions to 8V. An 8-bit ADC can provide 256 digital values to represent input voltages. If the values are spread uniformly over the full 8V input range, the least significant bit can measure 1/32nd of a volt, or about 31mV. Smaller voltage fluctuations will be ignored. On the other hand, if the 256 values are used to measure a range of just 2 volts, the least significant bit can measure 1/128th of a volt, or slightly less than 8mV—but voltages higher than 2V will be clipped.

An ADC will typically require a reference voltage, and will digitize the range from 0V to that voltage. The reference voltage must be chosen with the issues of accuracy and range in mind.

A microcontroller may provide a feature in its language to perform automatic scaling of an analog input within limits set by a variable in program code. This is done by comparing the input with a selectable voltage level, such as the voltage of the power supply, an externally supplied voltage, or a fixed built in reference. While the ADC in the microcontroller normally digitizes values from 0V to the supply voltage for the chip, the conversion routine can instruct the microcontroller to use its full number of bits (often, 10) to digitize an input range from 0V to 1V.

For a higher sample rate, an ADC chip may be connected with the microcontroller over an I2C or SPI bus.

2. Analog: Resistance

Resistance-to-Voltage Conversion

A sensor that changes its resistance as it responds to its environment can be placed in a *voltage divider* to provide an analog voltage output. This is illustrated in Figure A-3.

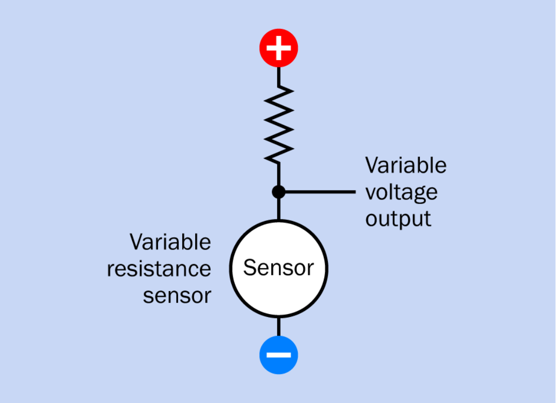
Variable
voltage
output
Variable
resistance
sensor
Sensor

Figure A-3 *The basic principle of putting a variable-resistance sensor in series with a fixed resistance, to create a voltage divider.*

To choose the value for the series resistor, if R_{MIN} and R_{MAX} are the minimum and maximum resistance values for the sensor, the optimum value R_S for the series resistor, to produce the widest variation in voltages, will be found from this formula:

$$R_s = \sqrt{(R_{min} * R_{max})}$$

When the sensor has been set up in this way, the output can now be processed in the same way as any analog voltage output from a sensor.

The datasheet of the sensor should be consulted to make sure that a series resistor value does not allow any chance of the sensor being damaged by excessive current.

3. Analog: Open Collector

Many sensor packages or modules include a bipolar transistor that has an *open-collector* output (or *open drain*, if a CMOS transistor is used). The transistor may or may not be incorporated in an internal **op-amp** (described in Volume 2). Either way, the principle is the same.

Figure A-4 shows a sensor, in darker blue, that has one connection for positive power, another connection for negative ground, and a third connection to the collector of the internal transistor.

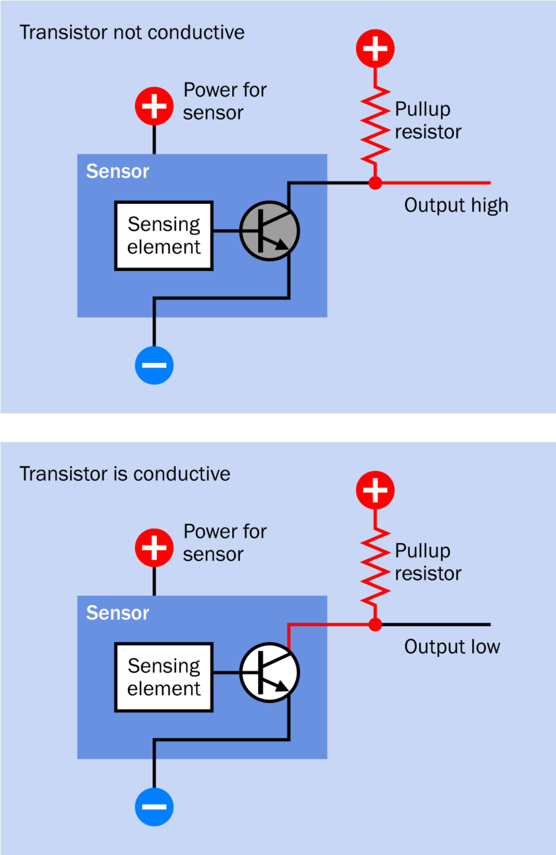
Transistor not conductive
Power for
sensor
Pullup
resistor
Sensor
Output high
Sensing
element
Transistor is conductive
Power for
sensor
Pullup
resistor
Sensor
Output low
Sensing
element

Figure A-4 *How to use an open-collector output from an internal transistor.*

In the upper section of the figure, the sensing element is not applying voltage to the base of the internal transistor, and the transistor only conducts a tiny amount of leakage current. Power applied to it through an external *pullup resistor* cannot reach negative ground in any significant quantity, and therefore it can provide a voltage input to a high-impedance device such as a microcontroller, or can power a component such as an LED, which draws relatively little current (20mA or less).

In the lower section of the figure, the sensing element is now applying voltage to the base of the transistor, drastically lowering its effective resistance. The transistor diverts current from the pullup resistor to ground, and the output appears to go low.

The type of sensing element will determine whether the transistor becomes conductive or nonconductive when the element detects a stimulus.

The value that should be chosen for the pullup resistor will depend on the impedance of any device attached to the open-collector output. A 10K resistor may be appropriate for use with a device such as a microcontroller, which has very high impedance. At the other extreme, if the device is an LED, a 330-ohm resistor may be necessary. The value of the pullup resistor must be sufficiently low to enable reliable operation, but sufficiently high to prevent excessive current from passing through the internal transistor when it becomes conductive (20mA is a common maximum value).

The voltage from the open collector can be processed in the same way as any analog voltage output from a sensor.

An open-collector output may be used when the outputs from multiple devices share the same bus. One device can drive the bus without the problem of other devices attempting to hold the bus voltage high.

4. Analog: Current

Relatively few sensors provide an output consisting of variations in current. Some semiconductor temperature sensors function in this way. The output current can be converted to a voltage output simply by placing a fixed series resistor, as shown in Figure A-5.

The voltage at the point shown, relative to negative ground, will vary linearly with the current. The value of the resistor should be defined in a datasheet for the sensor.

The voltage can now be processed in the same way as any analog voltage output from a sensor.

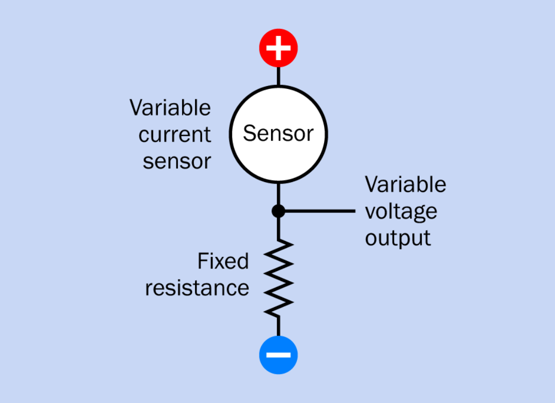

Figure A-5 *How to convert the output from a variable-current sensor.*

5. Binary: High/Low

A sensor that provides a binary output (that is, an output that is either logic-high or logic-low) can be connected directly with a microcontroller, if the voltage range is compatible. Program code in the microcontroller can then test the pin to establish its state. Note that some microcontrollers require a 3.3VDC power supply, while sensor chips may use 5VDC.

A binary output can also be used to control a solid-state relay, or an electromagnetic relay if transistor amplification is used. The output may be sufficient to power an LED indicator.

6. Binary: PWM

PWM is an abbreviation for *pulse-width modulation*. The sensor emits a stream of square-wave pulses with a fixed frequency, but the width of each pulse varies with the stimulus to which the sensor is responding. The width of each high pulse, relative to the wavelength between the start of one pulse and the start of the next, is called the *duty cycle*. A duty cycle of 0% means that there are no pulses at all. With a duty cycle of 100%, there are no gaps between pulses, so the output is high all the time. With a duty cycle of 50%, the duration of each high pulse is the same as the duration of the gap between pulses.

Various microcontrollers offer different ways to decode a PWM pulse stream. The most basic way is for a program to check an input pin repeatedly, as fast as possible, until a high state is detected. The microcontroller copies the value of its internal clock into a variable, then continues to check the input repeatedly until the pulse ends. The pulse duration that has been measured can be converted to a sensor value with a formula or lookup table.

This system is not recommended, as the microcontroller may miss the next pulse while it is converting the value of the previous pulse. To address this problem, a microcontroller language may offer a function that blocks execution of code while waiting for a pulse. The pulseIn() function on the Arduino is an example of this feature. However, the microprocessor must now spend most of its time waiting for a pulse instead of doing useful work.

A better solution is to write a program that is interrupt-driven.

Another option for decoding PWM is to use a low-pass filter that converts the pulse stream into an analog voltage, although some ripples will tend to remain.

Finally PWM can be used directly to power an LED or a DC motor, with transistor amplification as required. The speed of the motor or the brightness of the LED will vary with the duty cycle.

7. Binary: Frequency

Here again, the pulseIn() function on the Arduino may be used, so long as the frequency is a square wave with a known duty cycle.

8. Digital: I2C

In digital electronics, a *bus* is a communal pathway for sharing data among components or devices. The *I2C bus* is an abbreviation for *inter-integrated circuit* bus, developed originally by Philips in 1982. (Philips has since been subsumed into NXP Semiconductors.) The correct notation for I2C is I^{2}C, and it is spoken as "I-squared-C." However, the term is very commonly written as I2C.

The I2C standard defines a data sharing protocol that is limited to 400kHz (with some exceptions) and designed to work on a small scale, almost always within one device, and usually on one circuit board. It is a low-cost, simple design. Data is transmitted serially over two wires, and the devices sharing the bus are connected in parallel.

Typically there is one *master* device on the bus, and a number of *slave* devices. Masters and slaves can both transmit information, but the master normally initiates communication. It also emits a clock signal for synchronization of data.

A sensor is a slave device that can be interrogated by a microcontroller in its role as the master device. Because multiple slave devices can share a bus, the microcontroller needs a way to identify the slave that it wishes to talk to, and each slave is assigned a unique address for this purpose. Often a slave will allow the user to modify the last two bits of the address, so that up to four identical devices can share a bus.

Code libraries to support the I2C protocol are available for most microcontrollers, and communication with a sensor that uses I2C should simply require knowledge of the sensor's I2C address. However, the data registers in a sensor can be quite elaborate, requiring careful study of the manufacturer's datasheet. Multiple procedures may be required to set functions on a device (such as the sensitivity range of an accelerometer, or the threshold for a temperature alarm). Multiple procedures may also be required to read data out of a sensor (such as two bytes to define a temperature, and several bytes to obtain time as well as location readings from a GPS module).

9. Digital: SPI

SPI is an acronym for serial peripherals interface, a standard introduced by Motorola that serves a similar function to the I2C bus, described immediately above. The SPI standard is slightly more sophisticated, enabling duplex communication and higher data transfer speeds. However, SPI requires a minimum of three wires shared by all devices on the bus, and an additional device-selection line for each slave device. The benefit of the extra device lines is that devices are easier to select and address than on an I2C bus, where more program statements are required. As is the case for I2C, code libraries for microcontrollers are widely available to support SPI. However, the requirement for SPI to use three pins on a microcontroller, plus an additional pin for each slave device, is a disadvantage.

More sensors have I2C capability than SPI capability. The SPI protocol is potentially much faster than I2C.

A sensor that is SPI-enabled will almost certainly be available in a similar version that uses I2C. An increasing number of chip-based sensors support both protocols.

Glossary

This glossary is not comprehensive. It contains only the technical terms that have been used most frequently in this book in conjunction with sensor attributes.

ADC Analog-to-digital converter, which accepts a varying signal (usually, a voltage) as an input, and converts it to a digital value in the form of a binary number. This number is likely to have a high limit ranging from 255 decimal to 65535 decimal. Many microcontrollers contain their own ADC, which is multiplexed to assess the inputs on several pins. On an Arduino Uno, the ADC creates a digital value ranging from 0 through 1,023.

analog output When a sensor creates a voltage or varies its resistance without steps or increments, as a function of the phenomenon that it is measuring, this is an analog output.

binary output In this Encyclopedia, the term "binary output" describes a sensor output that only has two states: logic-high and logic-low, or on and off. The term is used in some datasheets, but more often a binary output is described, misleadingly, as an analog output.

breakout board A small printed-circuit board containing one or more integrated circuit chips, usually surface-mounted. The board makes the features of the chips easier to access, because it has pins or connectors with 2.54mm (0.1") spacing for convenient experimental use with a breadboard. Additional features may be included, such as a voltage regulator.

chip-based sensor This term is used in the Encyclopedia to describe a sensor that is etched into a silicon chip and usually has signal conditioning components and circuitry built in.

contact bounce The tiny and rapid vibrations of mechanical switch contacts, when the switch opens or closes. If the switch is connected with a digital device such as a logic chip, debouncing hardware may be necessary, to allow the contacts time to settle. If the switch is connected with a microcontroller, a delay from 5ms to 50ms may be written into program code. Different switches have widely different settling times.

decibel A unit that expresses relative power or intensity, often (but not exclusively) applied to audible sound. The decibel is one-tenth of a bel, and is abbreviated dB, with the B capitalized because it is derived from the name of Alexander Graham Bell. Because dB is a logarithmic unit, the scale has no zero origin. However, 0dB may be assigned arbitrarily to any intensity, in which case lower intensities will have a negative value. An increase of 3dB corresponds to a doubling of sound intensity (acoustic energy). However, when a sound is sensed by the human ear and evaluated by the brain, its subjective loudness doubles when the intensity increases by 10dB.

dielectric The insulating layer separating two plates in a capacitor.

hysteresis The difference between thresholds for switching an output on and off. When a sensor exhibits hysteresis, it may be unresponsive to a stimulus slightly above or below its equivalent current value. This may be useful to eliminate numerous responses to very small stimuli —for example, in a room thermostat.

I2C Interintegrated circuit bus. Sometimes written as I^2C, and often referred to verbally as "I-squared-C." A communications protocol that is often used between a microcontroller and other components on a circuit board. For a description, see "8. Digital: I2C" in the Appendix.

IMU Inertial measurement unit, consisting of three accelerometers and three gyroscopes, sometimes with the addition of three magnetometers. It can be used as a navigational aid. It may also be used in handheld user input devices, such as game controllers.

Kelvin A temperature scale, often abbreviated with the letter K, in which each degree is equivalent to a Celsius degree, but 0 degrees is at absolute zero—the temperature at which materials have no heat energy at all. 273 degrees K is approximately equal to 0 degrees C.

MEMS Microelectromechanical system, i.e., an integrated circuit chip that also contains tiny moving parts. For example, a MEMS accelerometer is built around microscopic springs that respond to accelerative forces.

newton A unit of force named after Isaac Newton, abbreviated with the capital letter N. A force of 1N will accelerate a mass of 1kg at a rate of 1 meter per second each second.

open collector output Many sensors have an open collector output, or contain an op-amp that has an open collector output. The output pin is attached to the collector of an internal transistor, with its emitter connected to negative ground. Positive voltage applied through a pullup resistor to the open collector will be grounded when the internal transistor is conducting current, but will be available for other devices when the transistor is off. See "3. Analog: Open Collector".

orthogonal Angled at 90 degrees. Three orthogonal elements in a sensor will all be angled at 90 degrees to each other.

pascal A unit of pressure equivalent to 1 newton of force per square meter.

PIR Passive infrared sensor. See Chapter 4.

pullup resistor A resistor that pulls up an output or input voltage in the absence of a signal. May be used in conjunction with an **open collector output**.

quadrature An encoding system for output from a pair of sensors. If the sensors are identified as A and B, four output combinations are possible: A high and B low; A high and B high; A low and B high; A low and B low. A common application is to show the direction of movement of a magnetic or optical pattern past the sensor pair.

reference temperature The temperature at which the output signal of a temperature sensor is measured. This is often listed in a datasheet.

register A section of memory that stores a digital value (usually 1 or 2 bytes in a sensor).

target Any object that is being detected by a motion sensor, proximity sensor, or presence sensor.

temperature coefficient The percentage increase or decrease in the value of a sensor as a result of unit change in temperature (usually 1 degree Celsius). Often abbreviated as TC. The value may be resistance, voltage, or current, depending on the sensor. The temperature coefficient should be negative if the value of the sensor diminishes when its temperature increases. If it is expressed in parts per million, abbreviated ppm, it can be converted to a percentage by dividing by 10,000.

Wheatstone bridge A network of four resistors. At least one of the resistors has an unknown value, while the others have precisely known reference values. The network enables calculation of the unknown value. See Figure 12-2.

Index

A

absolute humidity, 119, 120
absolute magnetic encoder, 41
absolute pressure, 113
accelerometer
  and gyroscope, 72
  applications, 70
  free fall, 71
  gravity, 71
  how it works, 70
  IMU, 5, 63, 66, 69
  Newton's Second Law, 72
  rotation, 71
  schematic symbol, 70
  spring, 70
  values, 73
  variants, 72
  what can go wrong, 74
  what it does, 69
ADC, 210, 217
air pressure, 114
alcohol vapor sensor (see gas concentration sensor, alcohol)
altimeter, 114
ammeter (see current sensor)
analog-to-digital conversion, 210, 217
anemometer, 125, 126
angle sensor (see rotary position sensor)
angular motion, 21
angular position sensor (see rotary position sensor)
Arduino-compatible
  gas flow rate sensor, 127
  gyroscope, 67
  IMU, 67
  magnetometer, 9
  pressure sensor, 114
  proximity sensor, 34
  temperature sensor, 184, 185
  touch screen, 97
  touch sensor, 91
avalanche diode, 136
axis of the earth, 7

B

bandgap temperature sensor (see semiconductor temperature sensor)
barometer, 114
barometric pressure, 114
Beidou positioning system, 3
bin switch, 57
black-body radiation, 25
Bourdon tube, 112
breakout board, 217
breakover switch (see tilt sensor)
Brokaw cell, 180

C

capacitive displacement sensor, 36
capacitive proximity sensor (see touch sensor)
capacitive touch screen, 96
capacitive touch sensor, 90
carbon microphone, 194
chip-based temperature sensor (see semiconductor temperature sensor)
clinometer, 55
coil magnetometer, 8
communications satellites, 1
compass, 6, 8
condenser microphone, 194
contact bounce, 18, 23
contents
  volume 1, xxi
  volume 2, xxii
  volume 3, xxii
crystal microphone, 196
current sensor
  ammeter, 199-200
  applications, 199
  Hall effect, 202
  how it works, 199
  microcontroller, 201
  panel meter, 199
  schematic, 200
  schematic symbol, 200

  series resistor, 200
  voltage drop, 200
  what can go wrong, 203
  what it does, 199
current-sense resistor, 201

D

debouncing, 18, 48, 61, 217
decibels, 196, 217
declination, 7
detection sensor (see presence sensor)
dew point, 119
dielectric, 90, 218
differential pressure, 113
diode temperature sensor (see semiconductor temperature sensor)
distance sensor (see proximity sensor)
dual slope converter, 211
dynamic pressure, 111

E

electret microphone, 195
encoder
  linear, 41
  magnetic, 41
  rotary, 48
errata, xxv

F

feedback to and from authors, xxv
flash converter, 211
float sensor, 100
float switch, 57
flux density, 6
flux, magnetic, 6
force sensor
  applications, 81
  deformative, 86
  how it works, 82
  how to use it, 86
  piezoelectric, 82
  plastic film, 82, 85-87
  resistive, 82, 86
  schematic symbol, 82
  strain gauge, 82, 83, 85, 88
  units of measurement, 82
  values, 87
  what can go wrong, 88
  what it does, 81
  wheatstone bridge, 83
four-wire touch screen, 96
frequency response, 197
fresnel lens, 27

G

Galileo positioning system, 3
gas concentration sensor
  alcohol, 118
  breakout board, 119
  cross-sensitivity, 118
  how it works, 117
  humidity, 117, 119
  oxygen, 119
  propane, 118
  schematic symbol, 117
  semiconductor-based, 117
  what can go wrong, 122
  what it does, 117
gas flow rate sensor
  anemometer, 125, 126
  applications, 128
  data output, 128
  schematic symbol, 125
  thermal mass, 127
  thermopile, 127
  units of measurement, 128
  what can go wrong, 129
  what it does, 125
gas pressure sensor (see pressure sensor)
gas sensor (see gas concentration sensor)
gauge pressure, 113
geographical meridians, 7
Global Positioning System (see GPS)
GLONASS positioning system, 3
GPS
  antenna, 2
  definitions, 1
  frequencies, 2
  how to use it, 3
  NMEA protocol, 3
  satellites, 1
  schematic symbol, 1
  time-keeping, 4
  tracker, 2
  values, 3
  what can go wrong, 4
  what it does, 1
gray code, 51
gyroscope
  and accelerometer, 72
  applications, 63
  axes, 64
  how it works, 64
  how to use it, 67
  IMU, 5, 63, 66, 69
  resonator (see gyroscope, vibrating)
  schematic symbol, 63
  values, 67
  variants, 66
  vibrating, 64
  what can go wrong, 67
  what it does, 63

H

Hall effect
  current sensor, 202
Hall effect sensor
  advantages, 22
  applications, xx, 52
  float sensor, 100
  how it works, 18
  how to use, 20
  magnetometer, 8
  rotary position sensor, 48
  variants, 19
hard-iron bias, 9
heading, navigational, 5
humidity sensor (see gas concentration sensor, humidity)
hygrometer, 120
hysteresis

  float sensor, 101, 218
  reed switch, 17, 218

I

I2C, 214, 218
IC temperature sensor (see semiconductor temperature sensor)
IMU, 5, 63, 66, 69, 218
inclination, 7
inclinometer, 55
inertial measurement unit (see IMU)
infrared proximity sensor, 32-34
infrared radiation, 188
infrared temperature sensor
  advantages, 187
  applications, 188
  array, 191
  compared with other temperature sensors, 153
  how it works, 188
  limitations, 188
  schematic symbol, 188
  thermopile, 189
  values, 191
  variants, 190
  what can go wrong, 192
  what it does, 187
inrush current limiter, 150, 161
interintegrated circuit bus (see I2C)

K

Kelvin temperature scale, 218

L

light meter, 137
light-dependent resistor (see photoresistor)
linear displacement sensor (see linear position sensor)
linear feedback shift register, 2
linear position sensor
  absolute, 41
  applications, 39, 42
  how it works, 39
  incremental, 41
  linear potentiometer, 40
  magnetic linear encoder, 41
  optical linear encoder, 41
  quadrature, 41, 218
  schematic symbols, 39
  what can go wrong, 43
  what it does, 39
linear potentiometer, 40
linear variable differential transformer, 42
liquid displacement sensor, 102
liquid flow rate sensor
  differential pressure, 109
  magnetic, 108
  paddlewheel, 105
  reed switch, 105
  rotor, 105
  schematic symbols, 105
  thermal mass, 107
  turbine, 106
  ultrasonic, 108
  what can go wrong, 109
  what it does, 105
liquid flow switch, 108
liquid level sensor
  displacement, 102
  float, 100
  float switch, 57
  how it works, 100
  incremental output, 101
  potentiometer, 101
  pressure of liquid, 103
  schematic symbols, 99
  tilting, 104
  turbulence, 104
  ultrasonic, 102
  weight of reservor, 103
  what can go wrong, 104
  what it does, 99
liquid pressure sensor (see pressure sensor)
load cell (see load sensor)
load sensor, 81
  (see also force sensor)
  applications, 82
  deformative, 86
  liquid level measurement, 103
  schematic symbol, 82
  strain gauge, 83, 85, 88
  values, 87
  what can go wrong, 88
  what it does, 81
LVDT (see linear variable differential transformer)

M

magnetic
  axis, 7
  ball in tilt sensor, 59
  declination, 7
  encoder, 41, 52
  field, 6
  flux, 6
  liquid flow rate sensor, 108
  meridians, 7
  presence sensors, 16
  rotary position sensor, 47
magnetic, linear encoder, 41
magnetometer
  breakout board, 6, 9
  coil, 8
  hard-iron bias, 9
  heading, 5
  how it works, 6
  how to use it, 9
  IMU, 5, 63, 69
  magnetoresistance, 8
  meridians, 7
  placement, 10
  scalar, 5
  schematic symbol, 5
  soft-iron bias, 9
  vector, 5
  what can go wrong, 9
  what it does, 5
magnetoresistance, 8
mass flow rate sensor (see gas flow rate sensor)
mathematical syntax, xxiv
membrane switch, 89
MEMS, xix, 218
MEMS microphone, 195
mercury switch, 58, 61, 78

meridians, 7
metric system, xxiv
microelectromechanical devices (see MEMS)
microphone
  carbon, 194
  condenser, 194
  crystal, 196
  directionality, 197
  dynamic, 194
  electret, 195
  frequency response, 197
  how it works, 194
  impedance, 198
  MEMS, 195
  moving coil, 194
  piezoelectric, 196
  rolloff, 197
  schematic symbols, 193
  sensitivity, 196
  signal-to-noise ratio, 198
  total harmonic distortion, 198
  units of measurement, 196
  what can go wrong, 198
  what it does, 193
motion sensor (see PIR)
mouse, 50
moving coil microphone, 194

N

navigational heading, 5
newton unit of force, 218
Newton's Second Law, 72
NMEA protocol, 3
north pole, 7
NTC thermistor (see thermistor, NTC)

O

object detector (see presence sensor)
object presence sensor (see presence sensor)
open collector output, 218
opt-pass sensor, 15
optical
  linear encoder, 41
optical isolator, 132
optical presence sensor, 12
optical rotary encoder, 48, 53
optical switch, 13, 15, 50
optocoupler, 142
optointerrupter, 21
orthogonal, 218
output
  analog current conversion, 213
  analog voltage conversion, 210, 217
  analog-to-binary conversion, 210
  analog-to-digital conversion, 210
  binary, 209, 210, 213, 217
  digital, 209
  dual slope converter, 211
  flash converter, 211
  frequency conversion, 214
  I2C, 214, 218
  open collector, 212, 218
  options, 209
  pullup resistor, 213, 218
  PWM conversion, 209, 213
  resistance-to-voltage conversion, 211
  scaling, 211
  SPI, 215
  successive approximation converter, 211
  voltage divider, 210
  voltage-to-frequency converter, 211
over-current protection, 160
over-temperature protection, 159
oxygen sensor (see gas concentration sensor, oxygen)

P

paddlewheel flow rate sensor, 105
panel meter, 199, 205
pascal unit of pressure, 218
passive infrared motion sensor (see PIR)
pendulum switch, 58
photocell (see photoresistor)
photoconductive cell (see photoresistor)
photodarlington, 142
photodiode
  applications, 135
  array, 137
  avalanche diode, 136
  band gap, 137
  color sensing, 137
  dark current, 135
  how it works, 135
  how to use it, 139
  light meter, 137
  photoconductive mode, 135
  phototransistor comparison, 143
  photovoltaic mode, 135
  PIN type, 136
  reverse bias, 135
  schematic symbols, 135
  ultraviolet, 137
  values, 138
  variants, 136-137
  wavelength range, 137
  what can go wrong, 140
  what it does, 135
photointerruptor, 13
photomicrosensor, 15
photoresistor
  availability, 131
  how it works, 132
  how to use it, 133
  optical isolator, 132
  phototransistor comparison, 133
  schematic symbols, 131
  sizes, 132
  voltage divider, 134
  wavelengths of sensitivity, 134
  what can go wrong, 134
  what it does, 131
phototransistor
  applications, 142
  base connection, 142
  binning, 144
  FET, 142
  how it works, 142
  how to use it, 144

- optocoupler, 142
- output current calculation, 145
- photodarlington, 142
- photodiode comparison, 143
- photoresistor comparison, 133
- schematic symbols, 141
- values, 143
- what can go wrong, 145
- what it does, 141

photovoltaic effect, 135

piezoelectric
- force sensor, 82
- heat detector, 26
- microphone, 196
- ultrasonic transducer, 32
- vibration sensor, 76

PIN photodiode, 136

pin-and-spring vibration sensor, 76

PIR
- applications, 25
- fresnel lens, 27
- how it works, 26
- schematic symbols, 25
- variants, 29
- what can go wrong, 30
- what it does, 25

potentiometer
- arc segment, 46
- end stops, 46
- float sensor, 101
- logarithmic, 46
- multiturn, 46
- rotary position, 46

presence sensor
- angular motion, 21
- configuration, 21
- linear motion, 21
- magnetic, 16
- optical, 12, 22, 23
- optointerrupter, 21
- reflective, 12
- retroreflective, 12, 15
- schematic symbol, 12
- transmissive, 12-13
- what can go wrong, 23
- what it does, 11

pressure sensor
- absolute pressure, 113
- altitude, 114
- ambient air, 114
- applications, 111
- barometric, 114
- Bourdon tube, 112
- diaphragm, 112
- differential pressure, 113
- dynamic, 111
- gas and liquid behavior, 112
- gases, 115
- gauge pressure, 113
- how it works, 112
- schematic symbols, 111
- sensing elements, 112
- static, 111
- units of measurement, 112
- what can go wrong, 116
- what it does, 111

propane sensor (see gas concentration sensor, propane)

proximity sensor
- applications, 32
- capacitive displacement, 36
- infrared, 32-34
- schematic symbols, 31
- triangulation, 32
- ultrasound, 32-33
- what can go wrong, 38
- what it does, 31

PTC thermistor (see thermistor, PTC)

pullup resistor, 213, 218

pulse-width modulation (see PWM)

PWM, 209, 213

pyroelectric detector, 26

Q

quadrature, 41, 49, 218

R

range finder (see proximity sensor)

reed switch
- advantages, 22
- alarm system, 17
- float sensor, 100
- how it works, 17
- how to use it, 18
- hysteresis, 17
- liquid flow rate sensor, 105
- values, 18
- variants, 17
- what can go wrong, 23

reference sources, xxvii

reference temperature, 218

reflective interrupter, 15

reflective object sensor, 15

reflective optical sensor, 15

reflective photointerrupter, 15

register, 218

relative humidity, 119, 120

resistance temperature detector (see RTD temperature sensor)

resistance-to-voltage conversion, 211

resistive force sensor, 82, 86

resistive temperature device (see RTD temperature sensor)

resistive touch screen, 95

retroreflective presence sensor, 12, 15

rotary encoder, 48, 52-53

rotary position sensor
- absolute, 51
- accelerometer, 71
- applications, 45
- gray code, 51
- Hall Effect sensor, 48
- incremental, 49
- magnetic, 47
- mouse, 50
- optical encoder, 48
- potentiometer, 46
- quadrature, 49, 218
- reflective optical, 48
- rotary encoder, 48
- schematic symbols, 46
- sensing chips, 48
- transmissive optical, 48
- what can go wrong, 53
- what it does, 45

rotary sensor (see rotary position sensor)

rotating coil magnetometer, 8

rotation sensor (see rotary position sensor)
rotational encoder (see rotary encoder)
rotational position sensor (see rotary position sensor)
RTD temperature sensor
  applications, 174
  attributes, 173
  compared with other temperature sensors, 153
  how it works, 174
  output conversion, 176
  probe enclosure, 176
  schematic symbol, 174
  three-wire variant, 175
  variants, 175
  what can go wrong, 176
  what it does, 173

S

Safari Books Online, xxvi
satellites, 1
scalar magnetometer, 5
schematic conventions, xxiv
schematic symbol
  accelerometer, 70
  current sensor, 200
  force sensor, 82
  gas concentration sensor, 117
  gas flow rate sensor, 125
  GPS, 1
  gyroscope, 63
  infrared temperature sensor, 188
  linear position sensor, 39
  liquid flow rate sensor, 105
  liquid level sensor, 99
  load sensor, 82
  magnetometer, 5
  microphone, 193
  photodiode, 135
  photoresistor, 131
  phototransistor, 141
  PIR, 25
  presence sensor, 12
  pressure sensor, 111
  proximity sensor, 31
  rotary position sensor, 46
  RTD temperature sensor, 174
  semiconductor temperature sensor, 178
  thermistor, NTC, 147
  thermistor, PTC, 157
  thermocouple, 166
  tilt sensor, 56
  touch screen, 95
  touch sensor, 90
  vibration sensor, 75
  voltage sensor, 206
Seebeck effect, 167
semiconductor temperature sensor
  analog current output, 182
  analog voltage output, 181
  applications, 178
  attributes, 178
  Brokaw cell, 180
  CMOS type, 185
  compared with other temperature sensors, 153
  digital output, 183
  how it works, 179
  multiple transistors, 179
  schematic symbol, 178
  variants, 180
  what can go wrong, 185
  what it does, 177
sensor
  categories, xx
  chip-based, 217
  multifunction, xx
  output options, 209
  suppliers, xxv
serial peripherals interface (see SPI)
shift register, 2
silistor, 158
slider potentiometer, 40
soft-iron bias, 9
solar panel, 135
sound pressure, 196
sources, xxvii
south pole, 7
speed sensor, 45
SPI, 215
static pressure, 111
stomp box, 133
strain gauge, 82-83, 85, 88
successive approximation converter, 211
suppliers, xxv
surge limiter (see inrush current limiter)

T

tactile switch, 89
target, 218
temperature coefficient, 147, 152, 218
temperature sensor
  comparison of types, 153
  infrared (see infrared temperature sensor)
  noncontact type, 187
  RTD (see RTD temperature sensor)
  semiconductor (see semiconductor temperature sensor)
  thermistor, NTC (see thermistor, NTC)
  thermistor, PTC (see thermistor, PTC)
  thermocouple (see thermocouple)
thermal mass flow rate sensor (see gas flow rate sensor)
thermal mass liquid flow rate sensor, 107
thermistor, NTC
  applications, 148
  compared with other temperature sensors, 153
  datasheet terminology, 152
  how it works, 148
  inrush current limiter, 150
  output conversion, 149
  schematic symbols, 147
  values, 152
  what can go wrong, 153
  what it does, 147
  wheatstone bridge, 150
thermistor, PTC

  compared with other temperature sensors, 153
  externally heated, 159
  heating element, 162
  inrush current limiter, 161
  lighting ballast, 162
  linear, 158
  nonlinear, 158, 159
  over-current protection, 160
  over-temperature protection, 159
  schematic symbols, 157
  silistor, 158
  starting current, 162
  what can go wrong, 163
  what it does, 157
thermocouple
  applications, 166
  compared with other temperature sensors, 153
  how it works, 167
  how to use it, 168
  identifiers, 169
  measurement junction, 168
  output conversion, 170
  reference junction, 168
  schematic symbol, 166
  Seebeck coefficients, 169
  Seebeck effect, 167
  thermopile, 171
  what can go wrong, 171
  what it does, 165
thermopile
  gas flow rate sensor, 127
  infrared temperature sensor, 189
  overview, 171
tilt sensor
  applications, 57
  bin switch, 57
  breakout board, 59
  float switch, 57
  how it works, 56
  how to use it, 61
  magnetic ball, 59
  mercury switch, 58, 61
  miniaturized, 59
  pendulum switch, 58
  schematic symbol, 56
  two-axis, 59
  values, 60
  what can go wrong, 61
  what it does, 55
tilt switch (see tilt sensor)
tip sensor (see tilt sensor)
tipover switch (see tilt sensor)
touch pad (see touch sensor)
touch screen
  Arduino-compatible, 97
  capacitive, 96
  four-wire, 96
  resistive, 95
  schematic symbol, 95
  variants, 95, 97
  what it does, 95
touch sensor
  applications, 90
  breakout boards, 91
  capacitive, 90
  design considerations, 92
  how it works, 90
  how to use it, 91
  schematic symbols, 90
  tactile feedback, 90
  touch pad availability, 91
  what can go wrong, 93
  what it does, 89
  wheels and strips, 92
touch strip, 92
touch wheel, 92
triangulation, 32
turbine flow rate sensor, 106

U

ultrasonic anemometer, 126
ultrasonic flow rate sensor, 108
ultrasonic liquid level sensor, 102
ultrasonic proximity sensor, 32-33
units of measurement, xxiv

V

Vactrol, 132
vapor sensor (see gas concentration sensor)
vector magnetometer, 5
vibrating gyroscope, 64
vibration sensor
  how to use it, 79
  magnetic, 77
  mercury switch, 78
  mousetrap type, 77
  piezoelectric, 76
  pin-and-spring, 76
  schematic symbols, 75
  values, 78
  variables, 78
  variants, 75
  what can go wrong, 79
  what it does, 75
vibration switch (see vibration sensor)
video display, 95
volt meter (see voltage sensor)
voltage divider, 210
voltage sensor
  accuracy, 207
  bar graph, 207
  how it works, 206
  panel meter, 205
  schematic, 206
  schematic symbol, 206
  volt meter, 205
  what can go wrong, 208
  what it does, 205
voltage-to-frequency converter, 211

W

wheatstone bridge, 83, 150, 160, 219

About the Authors

Charles Platt is the author of *Make: Electronics* and *Make: More Electronics*. He is a former senior writer for *Wired* magazine, and is a contributing editor to *Make:* magazine, for which he writes a column on electronics.

Fredrik Jansson is a physicist from Finland, with a PhD from Åbo Akademi University. He is currently living in The Netherlands, where he works on swarm robotics and simulates sea animals in the Computational Science group at the University of Amsterdam. Fredrik has always loved scavenging discarded household electronics for parts, and is a somewhat inactive radio amateur with the call sign OH1HSN. He also fact-checked Charles Platt's previous book, *Make: More Electronics*.

Colophon

The cover and body fonts are Myriad Pro, the heading font is Benton Sans, and the code font is Dalton Maag's Ubuntu Mono.

US 20030197093A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2003/0197093 A1**

St.Clair (43) **Pub. Date: Oct. 23, 2003**

(54) MAGNETIC VORTEX WORMHOLE GENERATOR

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/127,098**

(22) Filed: **Apr. 20, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64C 27/00**; B64C 39/00

(52) **U.S. Cl.** **244/62**

(57) ABSTRACT

This invention relates to a magnetic vortex generator which has the ability to generate negative mass and a negative spring constant which, according to Einstein's General Theory of Relativity, is required in order to create a stable wormhole between our space and hyperspace. Two separate, but electrically connected, toroidal coils of differing radii, carry magnetic flux in opposite directions about their common centerline. According to Maxwell's equation, this produces bucking electric fields along said centerline. Because the two solenoids have different radii, the parallel spring constant of both coils is negative. The negative mass together with the negative spring constant produce a real resonant frequency which can distort the spacetime curvature due to the creation of powerful spikes of negative mass. This phenomenon, similar to the common electrical thunderstorm, opens up a wormhole into hyperspace through which low-density hyperspace energy can enter into our dimension. This energy finds many application in new types of power supplies, inertia-less and mass-less spacecraft, vehicles that can travel light-years by moving out of dimension through hyperspace, surgery-less medical tables, cranes for lifting heavy objects, cold-welded crystals for crystal rotors, folding space waveguides, and electromagnetic field propulsion vehicles using highly relativistic fields.

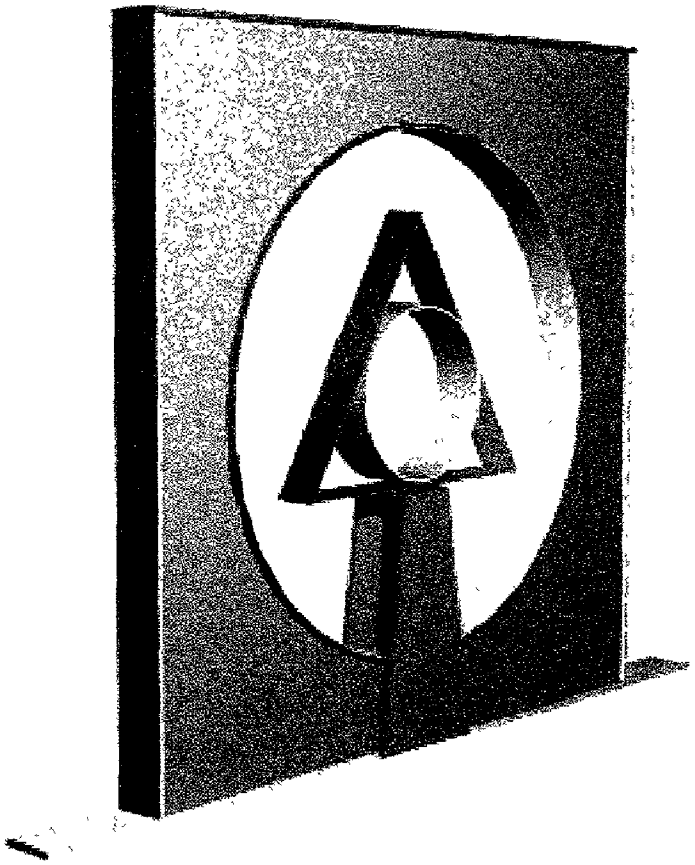

Figure 1

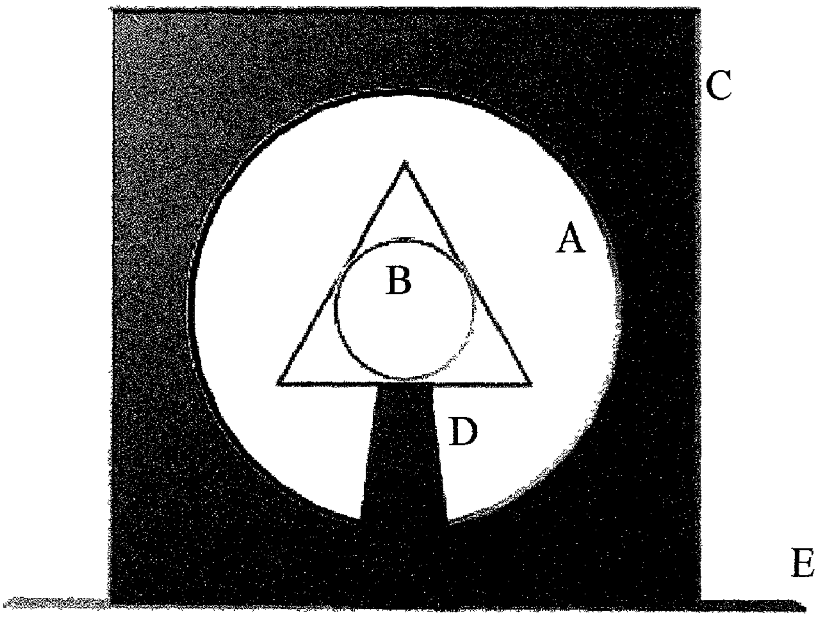

Figure 2

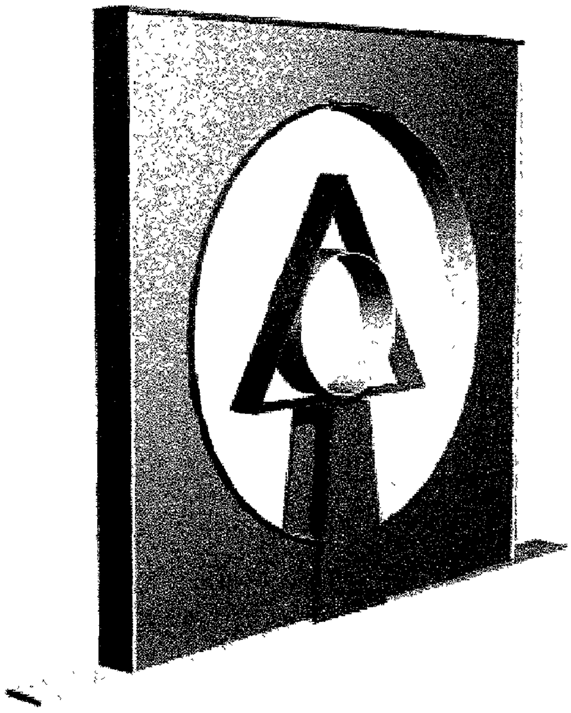

Figure 3

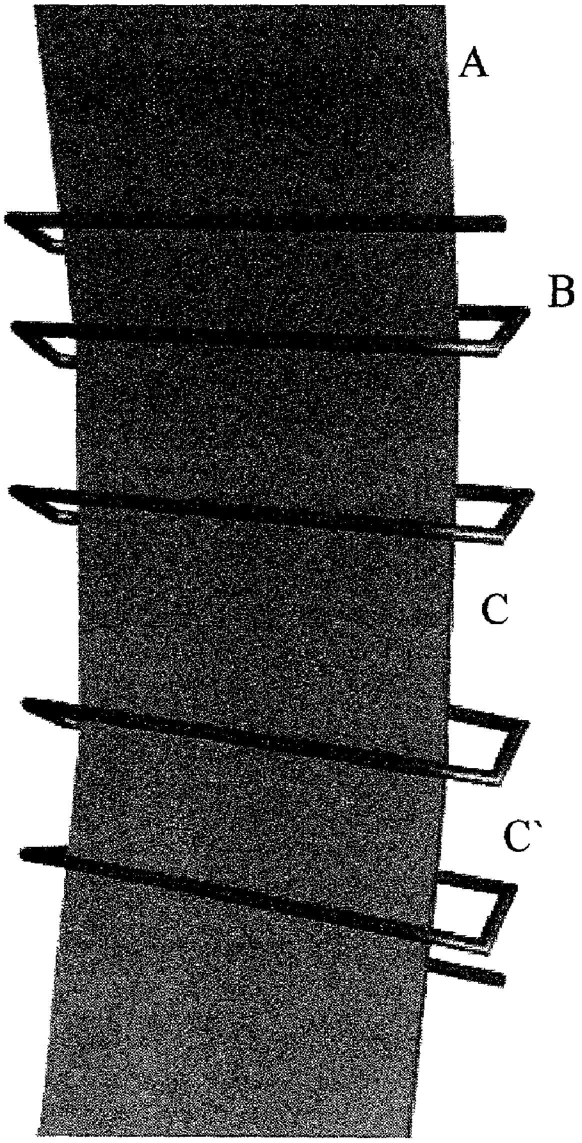

Figure 4

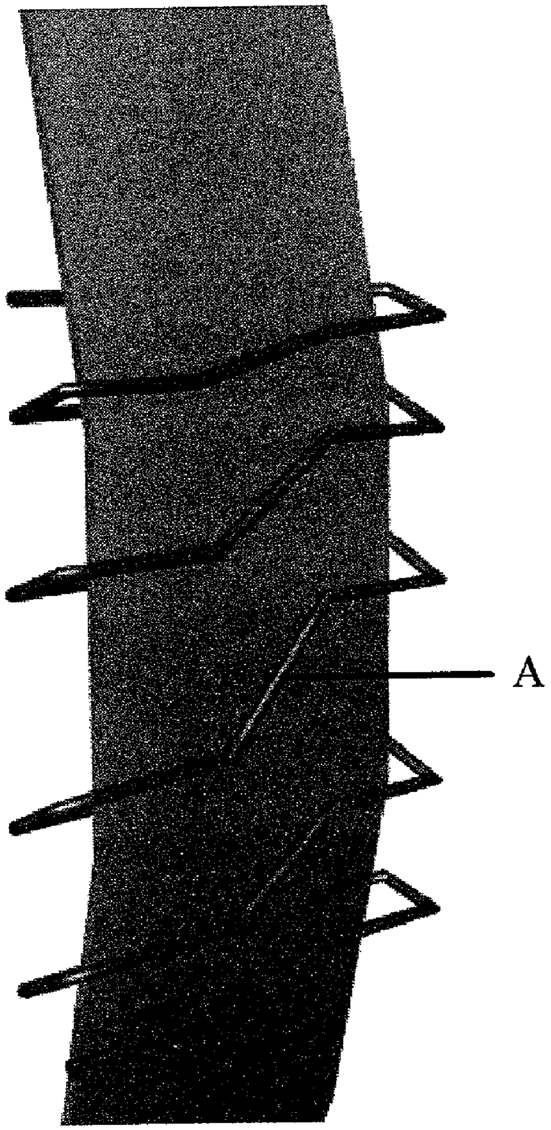

Figure 5

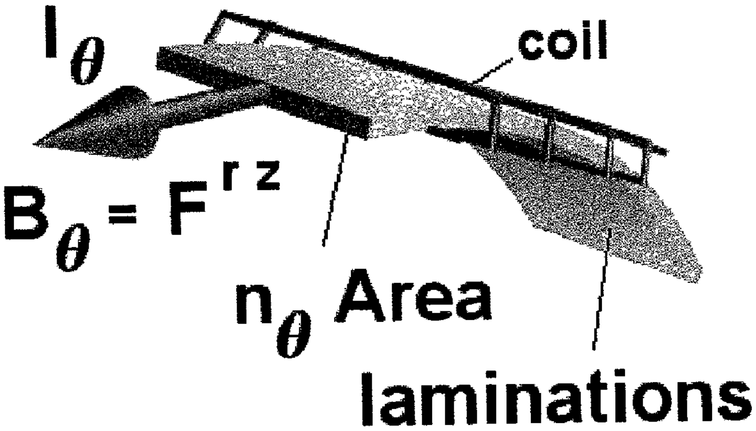
I_θ
coil
$B_\theta = F^{rz}$
n_θ Area
laminations

Figure 6

$$F^{\mu\nu} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & 0 & 0 & 0 & 0 \\ r & 0 & 0 & 0 & B_\theta \\ \theta & 0 & 0 & 0 & 0 \\ z & 0 & -B_\theta & 0 & 0 \end{array}$$

Figure 7

$$k^z = \frac{\text{coul}}{\text{sec}} \cdot \text{m}^2 \cdot \frac{1}{\text{m}^2} \cdot \frac{\text{kg}}{\text{sec} \cdot \text{coul}} = \frac{\text{newton}}{\text{m}} = \frac{\text{force}}{\text{m}}$$

Figure 8

$$k^z = \frac{1}{(r+a)^2} - \frac{1}{r^2} = -a \frac{(2 \cdot r + a)}{\left[(r+a)^2 \cdot r^2\right]}$$

Figure 9

$$g_{\alpha\beta} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & -1 & & B(\cos\theta - \sin\theta) & 0 \\ r & 0 & 1 & 0 & 0 \\ \theta & B(\cos\theta - \sin\theta) & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 10

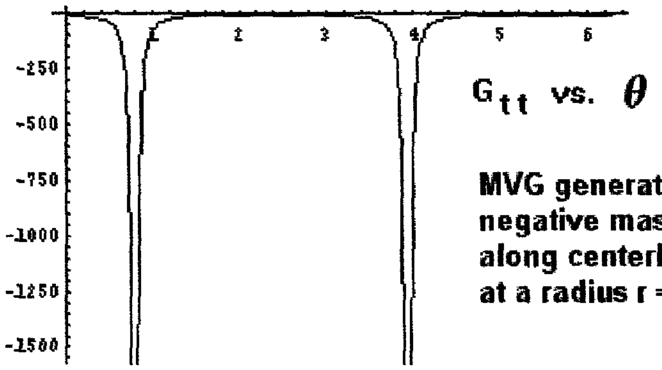

MVG generates two negative mass spikes along centerline of coils at a radius r = 0.

Figure 11

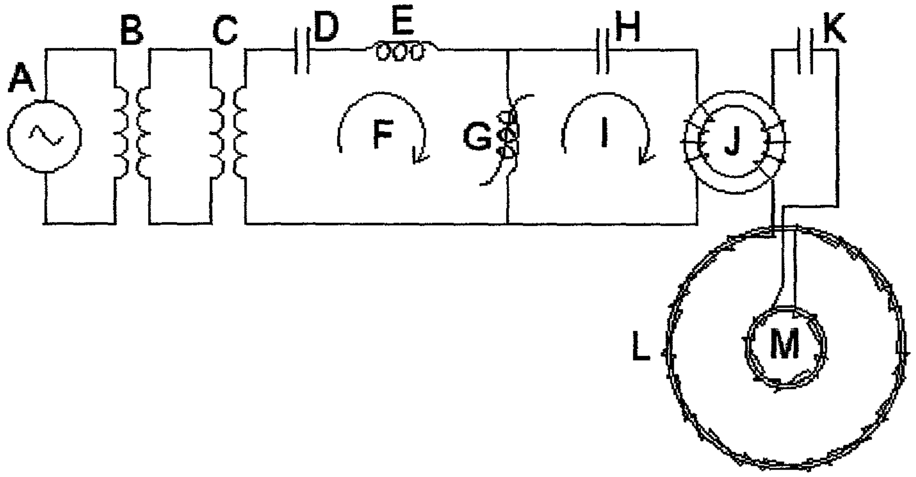
A
B
C
D
E
H
K
F
G
I
J
L
M

Figure 12

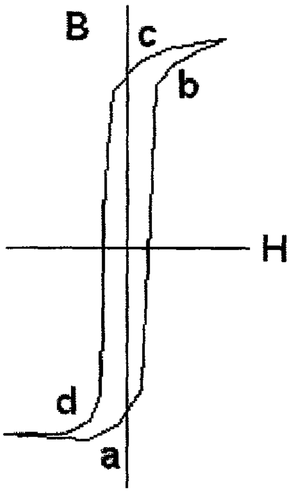

MAGNETIC VORTEX WORMHOLE GENERATOR

BRIEF SUMMARY OF THE INVENTION

[**0001**] This invention, which is the subject of my present application, is comprised of two solenoids wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminates of differing radii. The smaller solenoid is mounted along the centerline of the larger solenoid. This circular magnetic geometry creates linear bucking electric fields along the centerline of the coils. Because the magnetic flux in the laminates travels in opposite directions along arcs of differing radii in the two coils, a negative mass and a negative spring constant are generated by the system. From the theory of gravitational physics, a negative mass is prerequisite to producing a wormhole because it allows the throat of the wormhole to remain open and stable. The creation of the wormhole is facilitated by the appearance of a negative spring constant which allows the spacetime curvature to resonate to such a degree that said wormhole develops between our dimension and another co-dimension of hyperspace. Because the physics constants of hyperspace are different from ours, the wormhole allows hyperspace energy having a low speed of light to enter our dimension. Because electromagnetic fields obey the Lorentz transformation, it is now possible with this lower velocity of light to create huge relativistic fields which can drive the new electromagnetic field propulsion vehicles.

BACKGROUND OF THE INVENTION

[**0002**] The idea for this coil configuration comes from the observation of powerful thunderstorms, as described by physicist Dr. Richard Feynman in his *Lectures on Physics,* a copy of which is enclosed as a reference. Upon reading his explanation, I realized that the thunderstorm is actually a hyperspace physics phenomenon.

[**0003**] After the passage of a large lightning storm, people have observed that a car tire rim has merged with the trunk of a tree growing in the ground. Due to the large branches of the tree, there is no possible way that it could slide down the branches and around the trunk. It was observed also that a straw of wheat became embedded in the hard wood of a telephone pole. It turns out that the thunderstorm offers an explanation as to how this can occur.

[**0004**] After reading Feynman's explanation, it can be seen that the key to this phenomenon is that there is a downward and an upward lightning bolt, sometimes occurring together if the leader branches into two paths. Bolts of lightning also like to strike tall objects such as telephone poles or trees. Now an electrical current moving downward produces a clockwise magnetic B field, as seen from above. On the return stroke, the current is moving upward which produces a counterclockwise magnetic B field. Thus the thunderstorm produces two huge bucking magnetic B fields which is the magnetic geometry of this magnetic vortex generator. Using my tetrahedron diagram, I will then show that the low density hyperspace energy with its low speed of light is able to pull the rim out of dimension so that it can merge with the tree at the moment the lightning strikes.

SUMMARY OF THE INVENTION

[**0005**] It is the object of this invention to produce two toroidal oppositely-directed magnetic flux fields in two separate yet electrically connected solenoids of differing radii. Because the lines of flux are traveling in toroidal, curved-sheet transformer laminates along arcs of different curvature, the fields produce what is known in gravitational physics as negative mass. Negative mass has the property that if you project it onto a hot surface, the surface will cool rather than heat up. The negative mass, together with the resonance of the spacetime curvature involving a negative spring constant, create a wormhole into hyperspace. This interdimensional connection allows low speed of light hyperspace energy into our dimension which can be used to decrease the weight of an object, or produce huge relativistic electromagnetic fields that can be used to drive the new electromagnetic field propulsion spacecraft.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[**0006**] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[**0007**] **FIG. 1**. Front view of magnetic vortex generator.

[**0008**] **FIG. 2**. Perspective view of the magnetic vortex generator.

[**0009**] **FIG. 3**. The non-linear coil winding on the interior face of the laminations.

[**0010**] **FIG. 4**. The non-linear coil winding on the exterior face of the laminations showing that there is an electrical current component in the theta direction.

[**0011**] **FIG. 5**. The coil variables for the tensor calculation of the negative spring constant.

[**0012**] **FIG. 6**. The Faraday electromagnetic tensor showing the position of the magnetic field in the {r,z} slots.

[**0013**] **FIG. 7**. The units of the spring constant.

[**0014**] **FIG. 8**. The parallel spacetime spring constant.

[**0015**] **FIG. 9**. The cylindrical g metric tensor including the magnetic fields.

[**0016**] **FIG. 10**. A plot of the mass term G_n for Einstein's G curvature tensor showing that two negative mass spikes are created along the centerline of the generator. Negative mass is required to keep open the throat of the wormhole.

[**0017**] **FIG. 11**. Circuit diagram for the magnetic vortex generator.

[**0018**] **FIG. 12**. BH curve for non-linear SuperMalloy toroidal core. From the equation, $B=\mu H$, the slope of the line is equal to the permeability $\partial B/\partial H=\mu$.

DETAILED DESCRIPTION OF THE INVENTION

[**0019**] 1. The fact that negative mass is required to keep the throat of a wormhole open has been shown to be valid by physicist Dr. Kip Thorne in the enclosed reference physics paper. The key point of this invention is that if you have a negative mass, you also require a negative spring constant in order to get a real frequency and vice versa. The angular frequency of vibration is equal to the square root of the spring constant K divided by the mass M, or $\omega=\sqrt{K/M}$. If the mass is negative, and the spring constant is positive, the frequency is imaginary. Therefore, in order to get a real frequency, the spring constant has to be negative also.

[0020] 2. Referring to **FIG. 1**, the magnetic vortex generator consists of a large toroidal solenoid (A) and its support structure (C), a smaller toroidal coil (B) with its support structure (D) which holds the coil along the centerline of the larger coil. The entire structure sits on a base (E).

[0021] 3. The two solenoids are wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminations of differing radii. From the right hand rule, a changing circular magnetic field in the direction of the fingers produces a linear electric field in the direction of the thumb. Because there are two coils producing two magnetic fields in opposite directions, there are two bucking electric fields down the centerline of the coils. This duplicates the thunderstorm conditions. A perspective view of the generator is shown in **FIG. 2**.

[0022] 4. A enlarged view of the non-linear coil winding is shown in **FIG. 3**. The wire (B) is wound around the thin transformer laminations (A) with a non-linear coil spacing as shown by the difference in spacing between length (C) and (C'). On this inner side of the coil, the wire is wound straight across where it can be seen that the wire is normal to the edge of the laminations. Not shown are two strips of Velcro which keep the wire aligned and offset from the tape-covered metal laminations.

[0023] 5. Since the wire is straight across on this side, the opposite side has to have the wire run in a diagonal manner as seen in **FIG. 4A**. In terms of cylindrical coordinates, the components of the electrical current in the coil flow in the z-direction across the breadth of the lamination and in the O-direction around the lamination.

[0024] 6. The reason for the non-linear spacing is to preserve the vector potential of the coil. The vector potential is a more important field than the magnetic field because it can extend out past the windings of a long solenoid. If you look at the units, it is the field momentum per charge or kilogram meter per second coulomb. Notice that the derivative of the vector potential with respect to time is an electric field, while a derivative with respect to length is the magnetic field. The inductance of the coil times the current density is equal to the vector potential. Thus the inductance gradient of the coil times the current per meter is the magnetic field. So the non-linear coil picks up an additional magnetic field around the coil besides the one created in the laminations by the winding itself

[0025] 7. Taking a closer look at the coil in **FIG. 5**, it is constructed of thin laminations taped together to form a cylindrical shape with a wire coil wrapped around it. This creates a magnetic field in the theta θ direction within the laminations. Because the coil is wrapped as a helix, there is a current component I_θ in the theta direction. The cross-sectional area Area through which the magnetic flux flows times a normal vector n in the theta direction is the tensor area n_θ Area. If the radius of the coil is r, then the curvature K is $1/r^2$ pointing in the radial direction. Because there are two coils with differing radii, the generator has two curvatures associated with it.

[0026] 8. In the geometry of electromagnetism, the magnetic field is part of the electromagnetic Faraday tensor $F^{\mu v}$ which is a 4 by 4 spacetime matrix having rows and columns of time t, radius r, angle θ and height z in cylindrical coordinates. The first index μ refers to the row, and the second index v refers to the column. The diagonal of the matrix is zero. The first row and column belong to the electric field. All the other slots are filled by the components of the magnetic field. The B_θ component is located in the complementary slots of r and z as shown in **FIG. 6**. Now a tensor product can be written with the available variables I^θ no Area $K_r F^{rz} = k^z$. This says that the current around the loop in the theta direction times the lamination area vector in the theta direction times the curvature K of the coil in the radial direction times the magnetic field in the theta direction is equal to a spring constant in the z-direction, or normal to the plane of the coil. All the tensor components cancel out except for the z-direction. That is, the coil produces a spacetime spring constant through the center of the coil where there are resonant bucking electric fields. So the coil is creating a magnetic spring. The units of the spring constant are shown in **FIG. 7** as force per meter.

[0027] 9. Because there are two coils operating in opposite directions in regions of differing curvature, there are two parallel spring constants generated along the centerline. Two springs in parallel sum, but the inner coil is negative due to the triple product of negative signs of current, field and area. Looking at the front view of the coils, the right hand rule shows the field going counterclockwise with the thumb pointing along the electric field in the positive z-direction. The inner coil with the field going in the clockwise direction has the electric field in the negative z-direction. Therefore the outer coil's positive spring constant sums with the inner coil's negative spring constant to produce an overall negative spring constant as shown in **FIG. 8**. The inner coil has a radius r, and the spacing between the outer and inner coil is a. In this design, the inner radius is 1, and the outer radius is 3 which is the magic ratio in physics of 1/3. That makes length a equal to 2. Substituting r=1 and a=2 into the spring constant equation shows that the ratio is negative 8/9. That is, the spring constant is negative as previously asserted. If the spring constant is negative, it must produce negative mass in order to have a real frequency of resonance. Since it produces negative mass, then it can produce a wormhole as shown by Dr. Kip Thorne.

[0028] 10. The geometry of hyperspace physics is based on the geometry of the tetrahedron which is circumscribed by the sphere. The corners of the tetrahedron touch the sphere at a latitude of 19.47122063° which turns out to be, in terms of planet cosmology, the location where all the large volcanoes and vortices occur on Earth, Mars, Jupiter, Uranus and Neptune. Furthermore, the cosine squared of this angle is 8/9, which is the spring constant ratio for the magnetic vortex generator. That is, the coil is interacting with the geometry of spacetime which is why it is such an effective wormhole generator. As will be demonstrated later, the tetrahedral geometry of hyperspace shows that the electron and proton are one and the same particle. This is a new discovery in science coming out of this research. The diagram also shows that when the speed of light is reduced, due to the low density of energy coming into our dimension through the wormhole, Planck's constant divided by the speed of light puts the electron at the boundary between space and hyperspace. That is, the electron and the proton go out of dimension which is the reason that the car tire rim can become merged with the tree. The ramifications of this magnetic vortex generator are enormous. It means developing new types of crystals through interdimensional merging which will be needed in the development of these spacecraft and their crystal rotors. It means the possibility of teleporting spacecraft through hyperspace over distances of light years using the new folding space waveguides. It means light-weight or inertia-less spacecraft which can be accelerated at hundreds of thousands of g's. It means the ability to lift extraordinary loads in construction work. It means surgery-less medical tables where tumors can be pulled directly out of the patient's body.

[**0029**] 11. Because we have a magnetic spring, then there is some resonant frequency at which the coil can be operated in order to create a large spacetime distortion. It is this distortion that creates the wormhole into hyperspace. From this it follows that the wormhole attaches to a low pressure region with a magnetic monopole. Because pressure is linear mass times the speed of light squared divided by area, a low pressure signifies a low speed of light and less dense matter. What this means is that opposite polarity spacetime curvature at two ends of an electrode will produce a voltage along the electrode, effectively creating a power supply. The reason for this is that spacetime curvature, as shown by Einstein's General Theory of Relativity equation $G_{\alpha\beta}=8\pi T_{\alpha\beta}$, is equal to the square of the electromagnetic fields in the stress-energy tensor T. So the curvature on the ends of the electrode appear as voltages and the system acts as a battery supply to power the spacecraft.

[**0030**] 12. The other characteristic of the magnetic vortex generator is that it can pull objects out of dimension allowing the object to apparently move through solid walls. The object doesn't actually move through the wall because the object is not in the same dimension as the wall; it just appears that way. Moving out of dimension and then back again on the other side of the wall would effectively move the object through the wall for all intents and purposes.

[**0031**] 13. The tetrahedron diagram is a physics diagram which plots the natural logarithm of mass to the natural logarithm of wavelength. The product of mass times wavelength is equal to Planck's constant divided by the speed of light. So the product of the electron mass times its wavelength is equal to the proton mass times its wavelength is equal to the Planck mass times the Planck wavelength. The Planck wavelength is the bottom dimensional limit of the universe. We live in the Planck box which is bounded by the Planck wavelength and Planck mass. Outside this box is hyperspace.

[**0032**] 14. Because logarithms sum, the sum of the log of the mass plus the log of the wavelength is a constant sum. Thus the electron and proton and Planck mass slide on a 45° line known as the base constant which is equal at the axes to the log of Planck's constant divided by the speed of light. Planck's constant is measured in joule-sec so that multiplying it by the frequency 1/second of light gives the energy of the photon particle.

[**0033**] 15. Planck's constant is equal to the Planck wavelength times the Planck mass times the speed of light. The 45° base constant is Planck's constant divided by the speed of light, which means that the speed of light cancels out top and bottom, leaving the area of the Planck box as the value of the base constant. Hyperspace has a low linear mass compared to our dimension. Therefore, Planck's constant is reduced when this energy enters our dimension through the wormhole created by the generator. And the base constant is also so reduced. In terms of logs, this means that the 45° base line becomes more negative and moves to the right on the diagram. As it does so, the base line intersects the electron at the Planck wavelength which is the separation point between space and hyperspace. That is, the electron moves out of dimension. Because the electron and proton are one and the same particle, as shown in reference tetrahedron diagram tet0565, the proton and hence the entire atom is taken out of dimension as well.

[**0034**] 16. All of this can be seen more easily graphically on the tetrahedron diagram itself, referring to reference tetrahedron diagram tet3025. The 45° line which intersects the electron at point (b) is the base constant for our dimension. As you can see, this base line intersects the horizontal axis at a value of −95.91546344 which is the log of Planck's constant h divided by the speed of light. Because Planck's constant is proportional to the linear mass, it is reduced in value by the low density hyperspace energy and, in terms of logs, becomes more negative. This moves the base line to the right at a value of around minus **105**. The new base line intersects the electron at point (a) which is located on the Planck wavelength that is the boundary between space and hyperspace. Thus the electron at point (a) goes out of dimension.

[**0035**] 17. It was inferred previously that a negative spring constant meant a negative mass was produced by the wormhole generator. This can actually be calculated using Einstein's General Theory of Relativity. The calculation starts with the g metric tensor which is a spacetime measurement of distance in terms of time t, radius r, horizontal angle θ and length z. This 4 by 4 matrix is shown in **FIG. 9** where the diagonal line has a signature in cylindrical coordinates of values equal to $\{-1, 1, r^2, 1\}$. All the other terms of the matrix are zero except for the magnetic fields in the two coils. Because the field is changing sinusoidally with time in the theta direction, the field has to go into the {t, θ} and {θ, t} slots of the matrix. Because the field in the inner coil is in the negative direction compared to the outer coil, and including a 90° phase shift between the two fields, a suitable magnetic field function would be B (cos(θ)-sin(θ)).

[**0036**] 18. Using a general relativity software package, Einstein's G curvature tensor can be calculated for this particular metric. The first term G t in the upper left hand slot in the corner is the mass term for the tensor. All the other terms are either electromagnetic fluxes or pressure terms involving the squares of the fields.

[**0037**] 19. Referring to **FIG. 10, a** plot of the mass as a function around a small circle shows that two negative mass spikes occur around the circumference of the circle. Because the radius is almost zero, the two spikes are actually coincident.

[**0038**] 20. This next section shows the electrical system used to drive the magnetic vortex generator. Referring to **FIG. 11**, the system is driven by a sinusoidal voltage source (A) into a 1:1 turns isolation transformer (B). The voltage is stepped up into the range of thousands of volts using a step-up hi-pot transformer (C). The first loop consists of a direct current blocking capacitor (D), a transformer choke (E), a variable inductance (G) and the current in the loop (F). The variable inductance is a coil winding on a toroidal core wound with thousandth-inch thick SuperMalloy tape. This coil acts as a magnetic switch due to its variable permeability.

[0039] 21. Referring to the accompanying drawing in **FIG. 12**, the slope of the BH curve is actually the permeability of the core. At point (a) on the curve, the slope is very low and therefore the permeability is very small. The resistance of the coil is the frequency of the current times the inductance of the coil. If the permeability is small, then the inductance is low, which means that the coil resistance is low initially. With a low resistance in the coil, current (F) flows through the winding rather easily. Then the coil goes from point (a) to point (b) where the permeability and resistance increase. This change in resistance from a low to a high value dumps the magnetic energy into capacitor (H). Then the non-linear coil saturates between points (b) and (c) where again the coil has a small slope and the coil resistance switches to a low value. Capacitor (D) then dumps its charge (I) through coil (G) producing a large voltage spike in the input and output winding of transformer toroidal coil (J). The magnetic flux in coil (J) then produces a voltage spike in coils (L) and (M) of the magnetic vortex generator. The frequency of oscillation of the generator is determined by capacitor (K) and the overall inductance of the two coils. The diagram shows that the winding is non-linear and in opposite directions going from outer coil to the inner coil.

[0040] 22. Frequency of oscillation has to be kept under 20 MHz in order to create a soft wormhole that connects to low pressure regions of hyperspace. The pressure regions of hyperspace are similar to the pressure produced by a dam holding water. The upper surface of the water, where there is no water pressure, is analogous to the black void into which our universe is expanding. In the middle of the dam, there is a region just above us having a lower water pressure which corresponds to the low density hyperspace energy. This analogy is not perfect because the many frequencies of hyperspace are quantized, as we know from quantum physics, rather than being a continuous spectrum of lower and lower frequencies down to the zero frequency of the black void.

What I claim as my invention is

1. A magnetic vortex generator which can generate negative mass according to Einstein's General Theory of Relativity, which is a tested and proven theory. As a result of this theory, it can be shown that negative mass is required to create a stable wormhole between space and hyperspace. Without the negative mass, the throat would close.

2. The generator consists of two solenoids wound with a common wire in opposite directions on two separate toroidal curved-sheet transformer laminates of differing radii. The smaller solenoid is mounted along the centerline of the larger solenoid.

3. According to Maxwell's equations, said geometrical and magnetic arrangement produces linear bucking electric fields along the centerline of said coils. Since curvature and the square of the electromagnetic fields are one and the same phenomenon according to Einstein's spacetime curvature tensor $G=8\pi\ T$, this resonance of the electric fields causes a resonance of the spacetime curvature and the opening of the wormhole.

4. Due to the fact that the magnetic flux travels within the laminations at a curvature equal to the inverse of the radius squared, each coil produces a spring constant which depends on the current, lamination area, magnetic field strength and the individual curvature of each coil.

5. Due to the fact that the flux travels in opposite directions in each solenoid, the spring constant of the outer coil is positive, and the spring constant of the inner coil is negative. These two spring constants, one positive and one negative, add in parallel to create a negative spring constant for the generator. Because the frequency of resonance is equal to the square root of the negative spring constant divided by the negative mass, the resonant frequency is positive real.

6. This process of creating and keeping open the wormhole allows low density hyperspace energy to enter our dimension. Because the linear mass is lower, Planck's constant, equal to the Planck mass times the Planck wavelength times the speed of light, is reduced to such an extent that the electron is moved out of dimension. Because the electron and proton are one and the same particle, when considering a path through space and hyperspace, the proton is also moved out of dimension. Thus this process of moving the atom in and out of dimension using the magnetic vortex generator has the ability to create cold-welded crystals, new types of materials and new ways to cold solder one disparate material to another.

7. As a consequence of said process, highly relativistic electromagnetic fields can be created because the velocity of light has been reduced considerably. These fields than can be used to produce life in the new class of electromagnetic propulsion vehicles.

8. As a consequence of said process, refrigeration systems can be created because negative energy cools rather than heats.

9. As a consequence of said process, power supplies can be made such that a differential spacetime curvature on the ends of a carbon electrode can create a differential voltage similar to a regular chemical battery.

10. As a consequence of this process, surgery-less medical tables can be fabricated whereby tumors can be pulled directly from the patient's body.

11. As a consequence of this process, the negative mass produced by the generator can offset the mass of a spacecraft to create a mass-less, inertia-less vehicle which can accelerate at hundreds of thousands of g's.

12. As a consequence of this process, the folding space waveguide becomes a reality whereby hyperspace energy, which has a low spring constant, can be easily folded and curved using powerful, relativistic electromagnetic fields. Spacecraft will be able to teleport themselves out-of-dimension over huge distances measured in terms of light-years.

13. Because the wormhole opens up an interdimensional connection to hyperspace having a magnetic monopole, a radial magnetic field is created. As a consequence of this process, a changing magnetic monopole field crossed with a changing electric field can produce a toroidal electromagnetic flux around the circular hull of the spacecraft. As a consequence of the merger of these two fields, a spacetime curvature G_{zz} is produced over the hull of the spacecraft which creates a tension or lift force which enables the spacecraft to ascend, hover or descend.

14. An electrical system, comprised of a variable frequency generator and amplifier, an isolation transformer and voltage step-up transformer, drives a non-linear inductance which switches on and off the current in the output circuit in such a manner as to produce large voltage spikes through magnetic vortex generator. The generator responds by producing two large negative mass spikes close to the centerline of the two coils.

15. The winding on each coil of the generator has a non-linear spacing to enhance the magnetic field and to reduce the interwinding coil capacitance.

16. The ratio of the radius of the small coil to that of the larger coil is 1/3, which is the magic ratio in physics. This creates a spring constant that is proportional to 8/9. The square of the cosine of the tetrahedral angle of 19.47° is equal to this ratio. Also the ratio of the area-to-volume ratio of the circumscribing sphere of a tetrahedron to the area-to-volume ratio of the tetrahedron is also 1/3. And the corners of the tetrahedron touch the circumscribing sphere at 19.47°. All the large volcanoes and vortices on Earth, Mars, Jupiter, Uranus and Neptune are located at this latitude. Thus this invention is more effective in developing a wormhole because it is tuned geometrically to the tetrahedral geometry of space.

* * * * *

US 20060071122A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0071122 A1**

St. Clair (43) **Pub. Date: Apr. 6, 2006**

(54) **FULL BODY TELEPORTATION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
4A
52 KINGS COURT
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/953,212**

(22) Filed: **Sep. 29, 2004**

Publication Classification

(51) **Int. Cl.**
B64C 39/00 (2006.01)

(52) **U.S. Cl.** **244/62**

(57) **ABSTRACT**

A pulsed gravitational wave wormhole generator system that teleports a human being through hyperspace from one location to another.

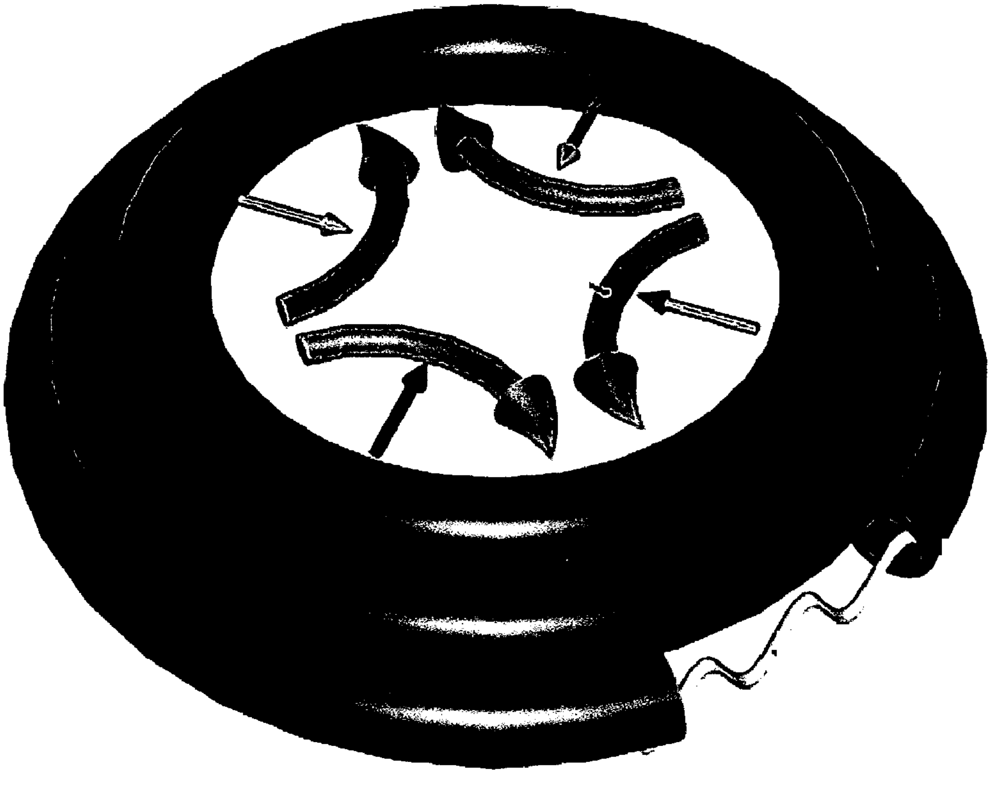

Figure 1

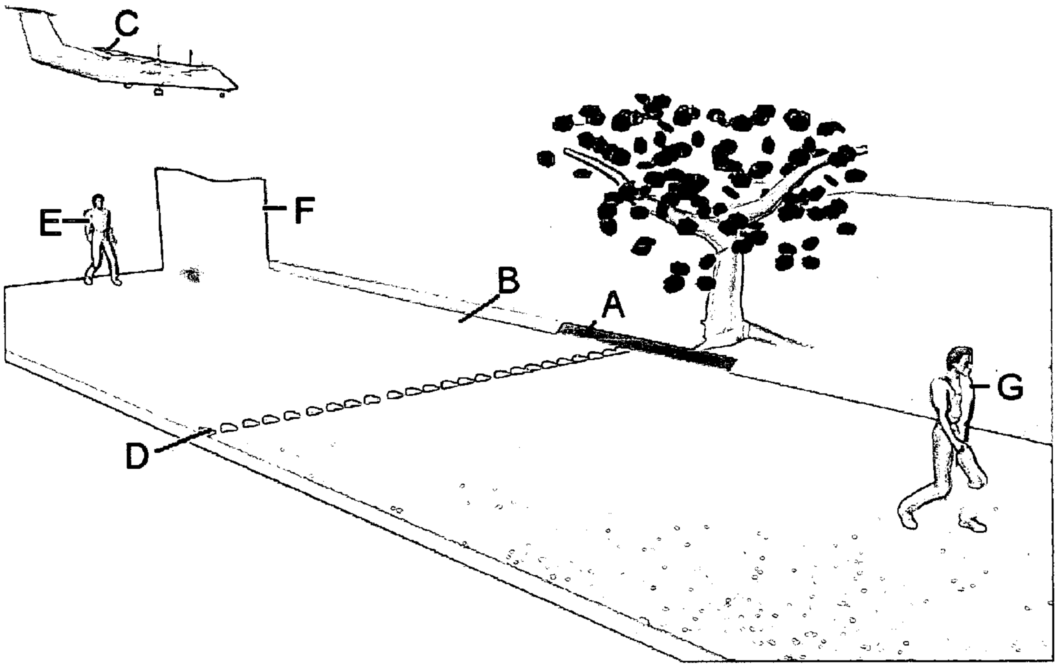

Figure 2

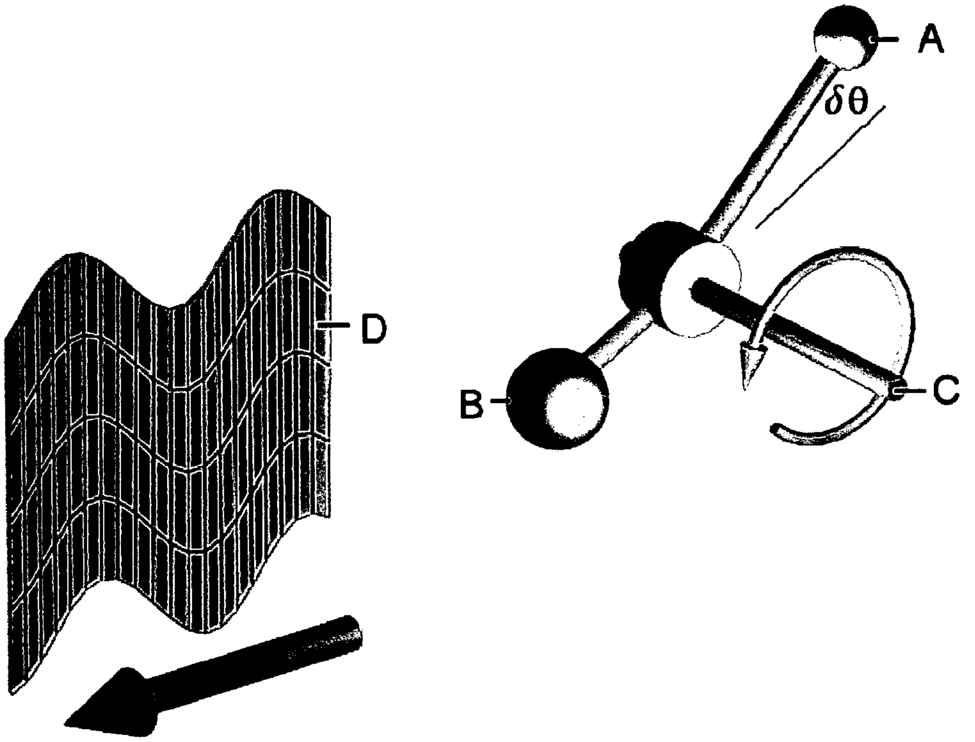

Figure 3

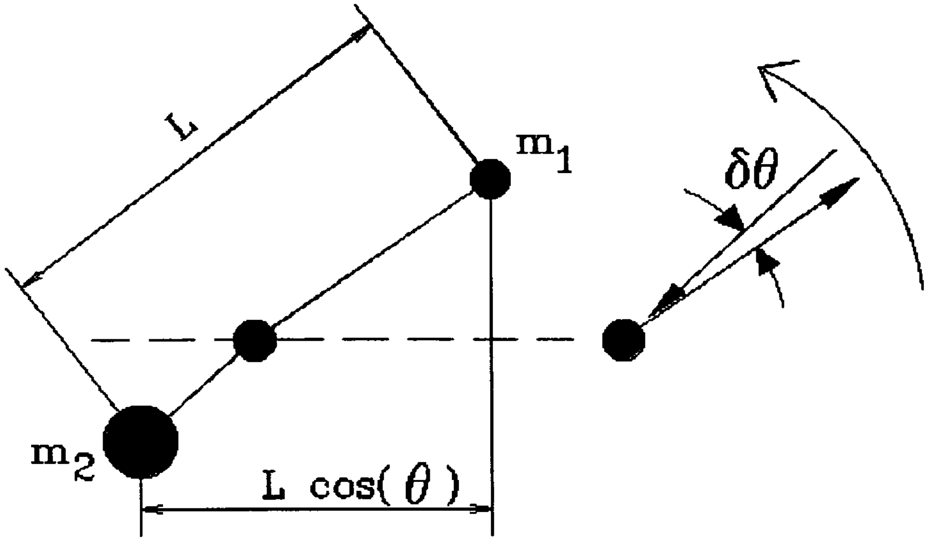
L
m_1
$\delta\theta$
m_2
L cos(θ)

Figure 4

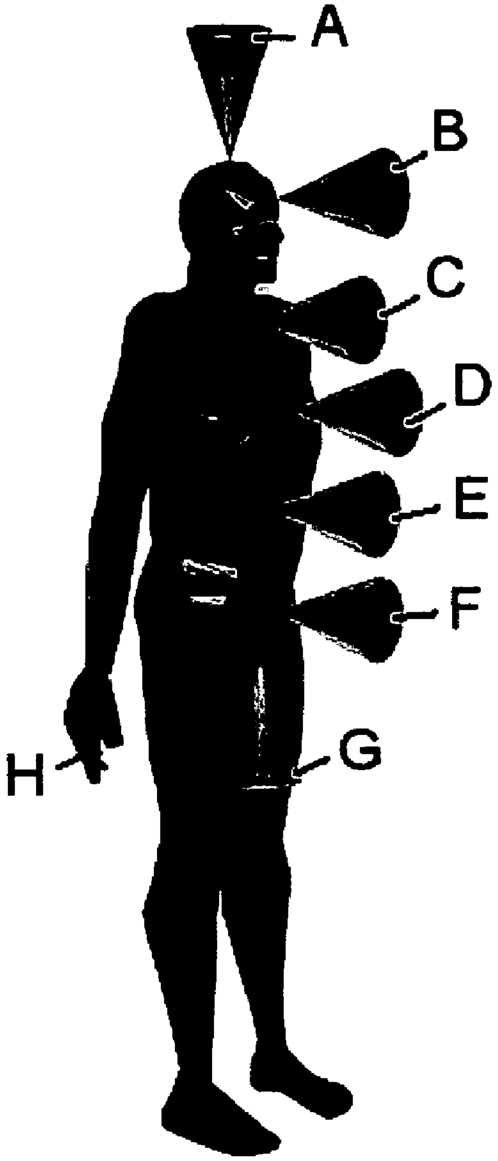

Figure 5

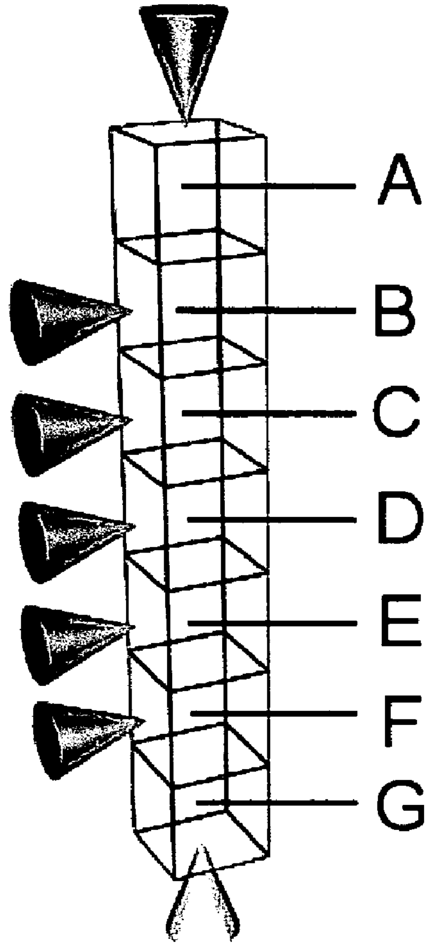

Figure 6

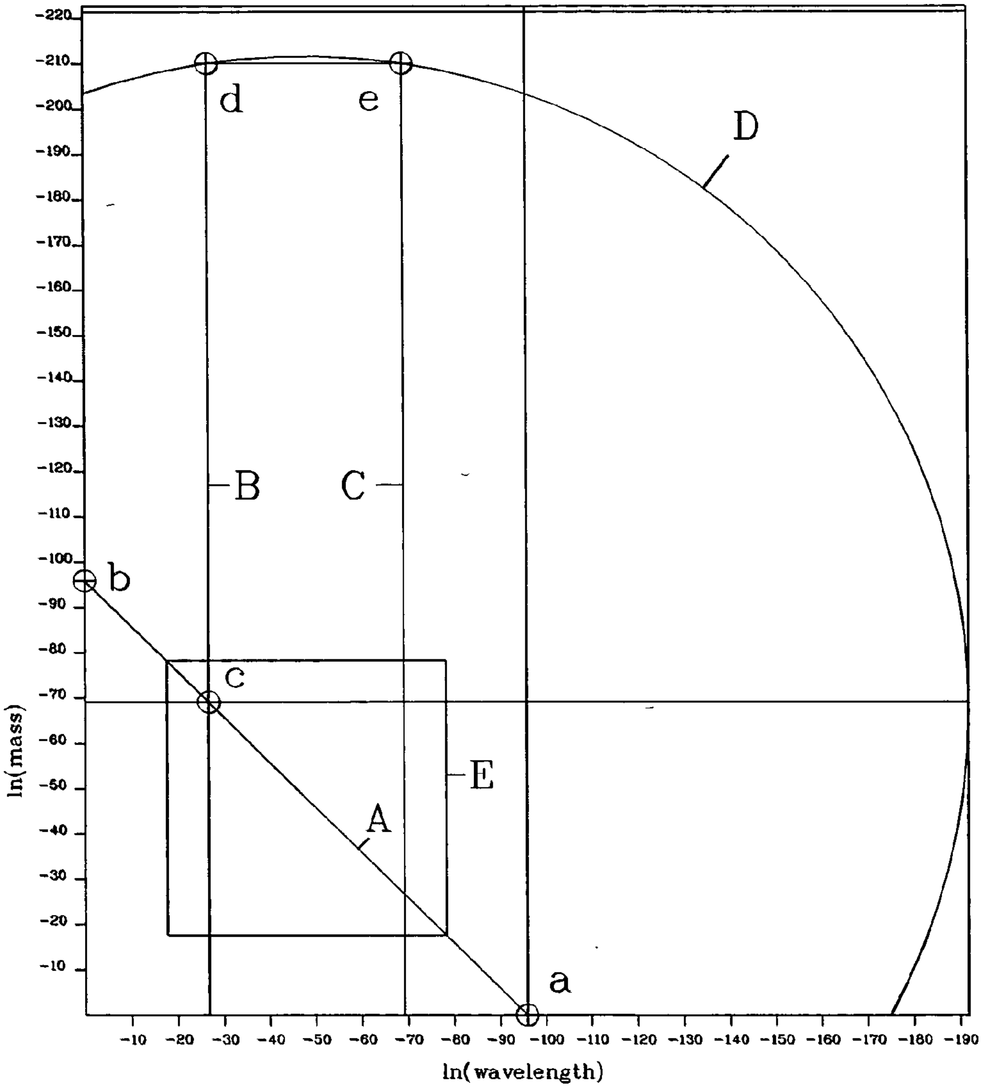
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
d
e
D
B
C
b
c
E
A
a

Figure 7

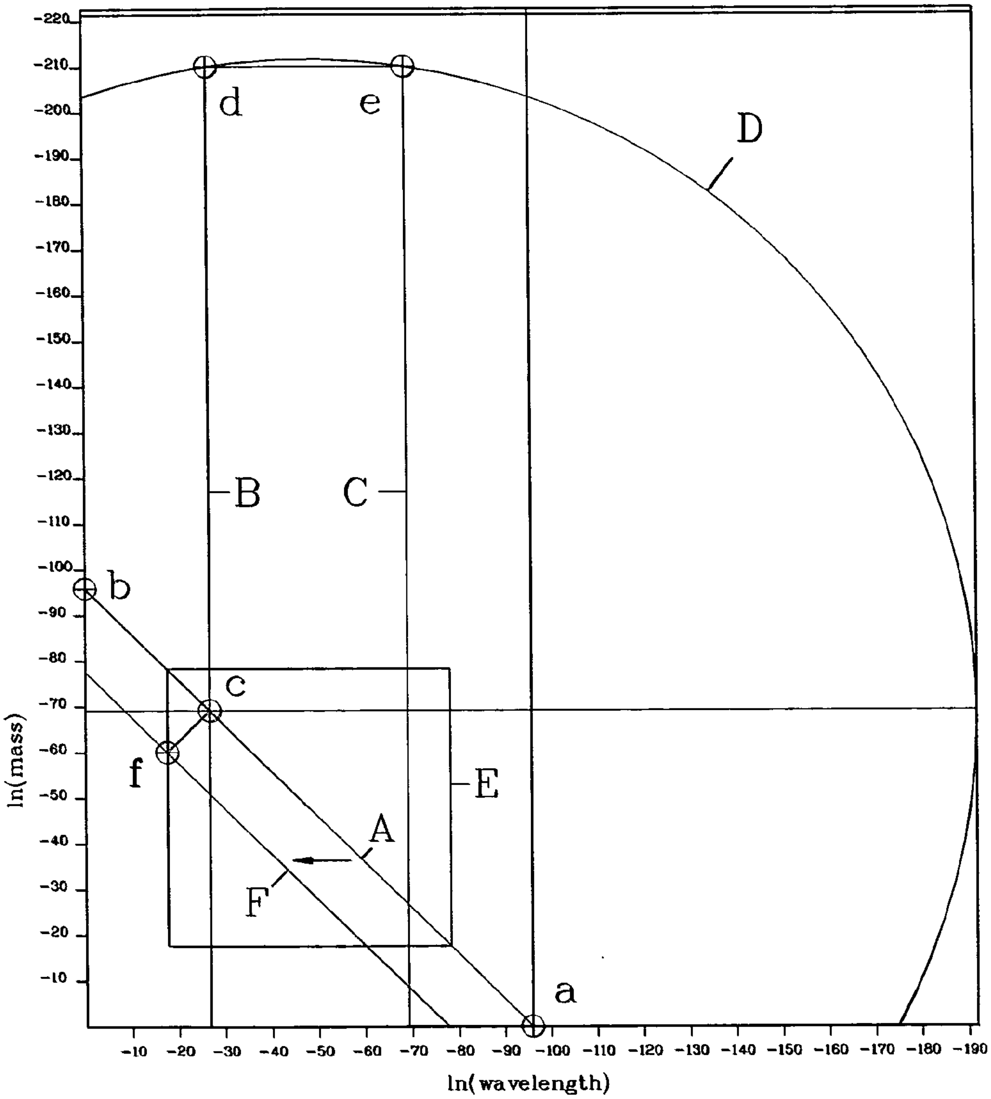
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
a
b
c
d
e
f
A
B
C
D
E
F

Figure 8

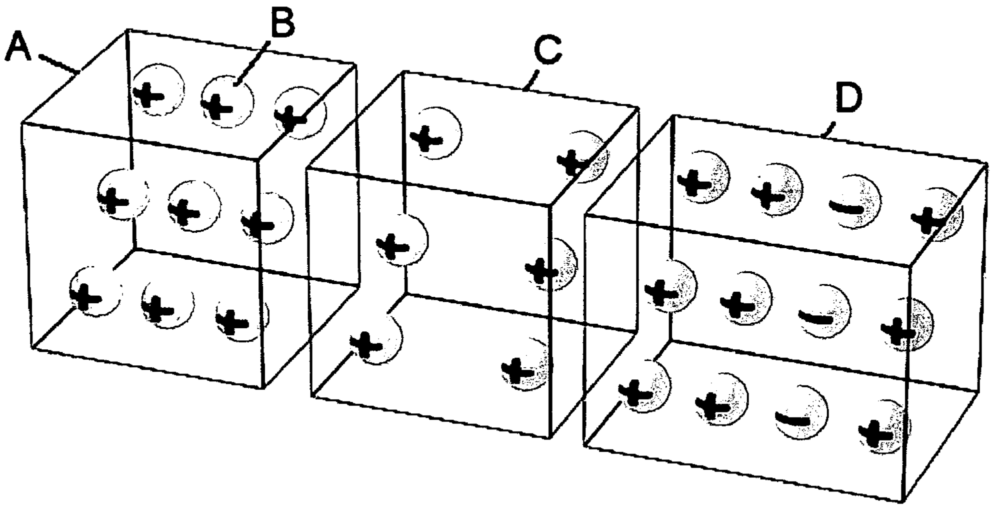

Figure 9

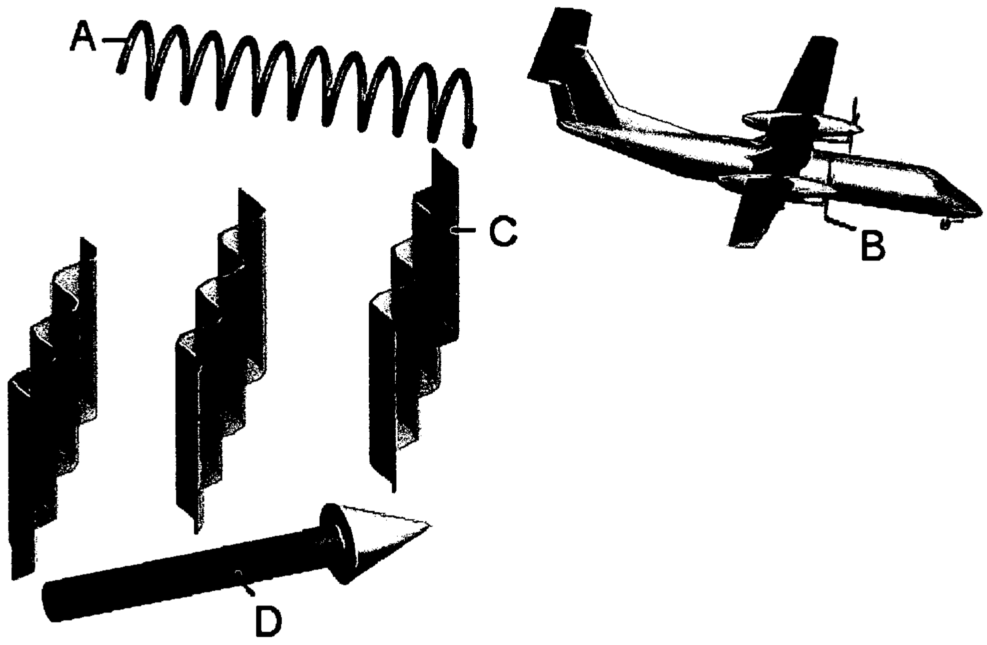

Figure 10

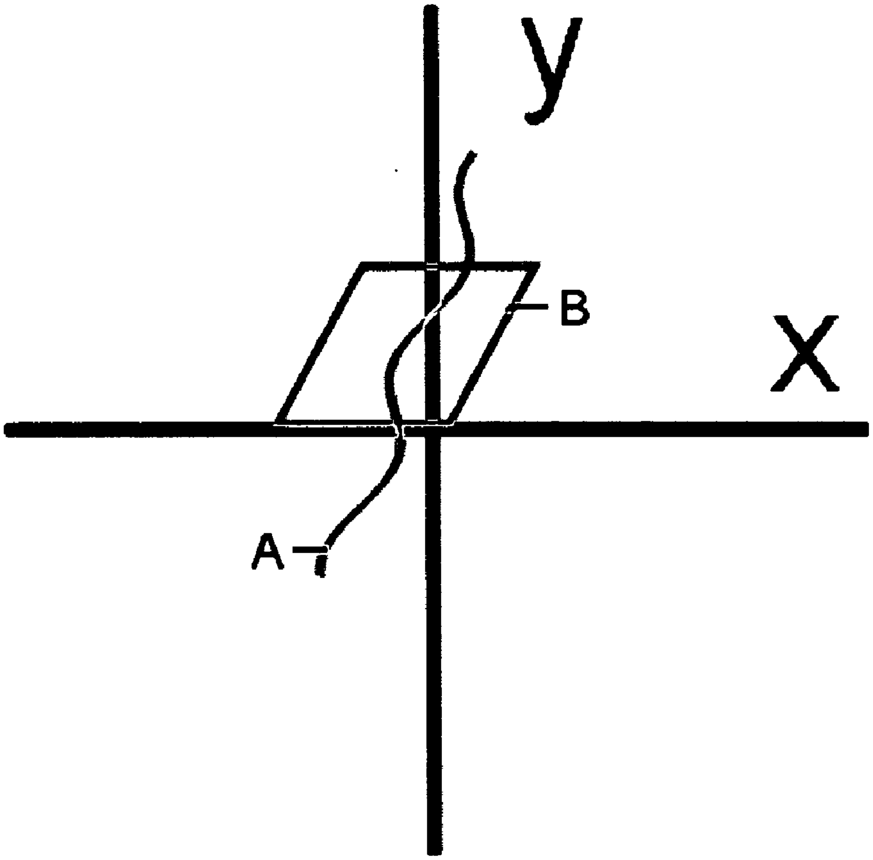

Figure 11

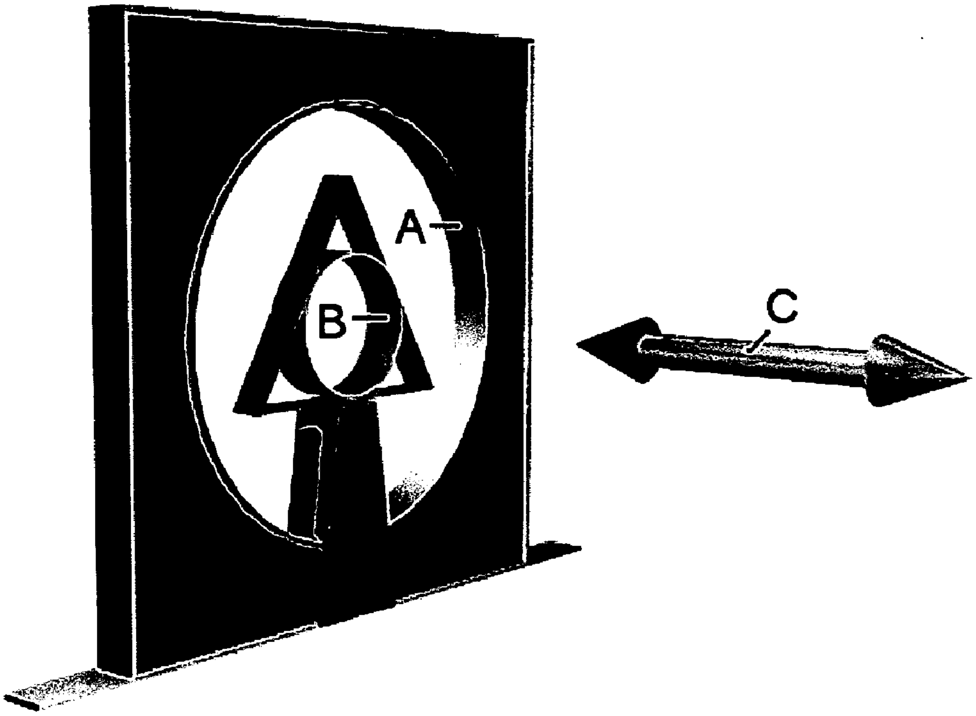

Figure 12

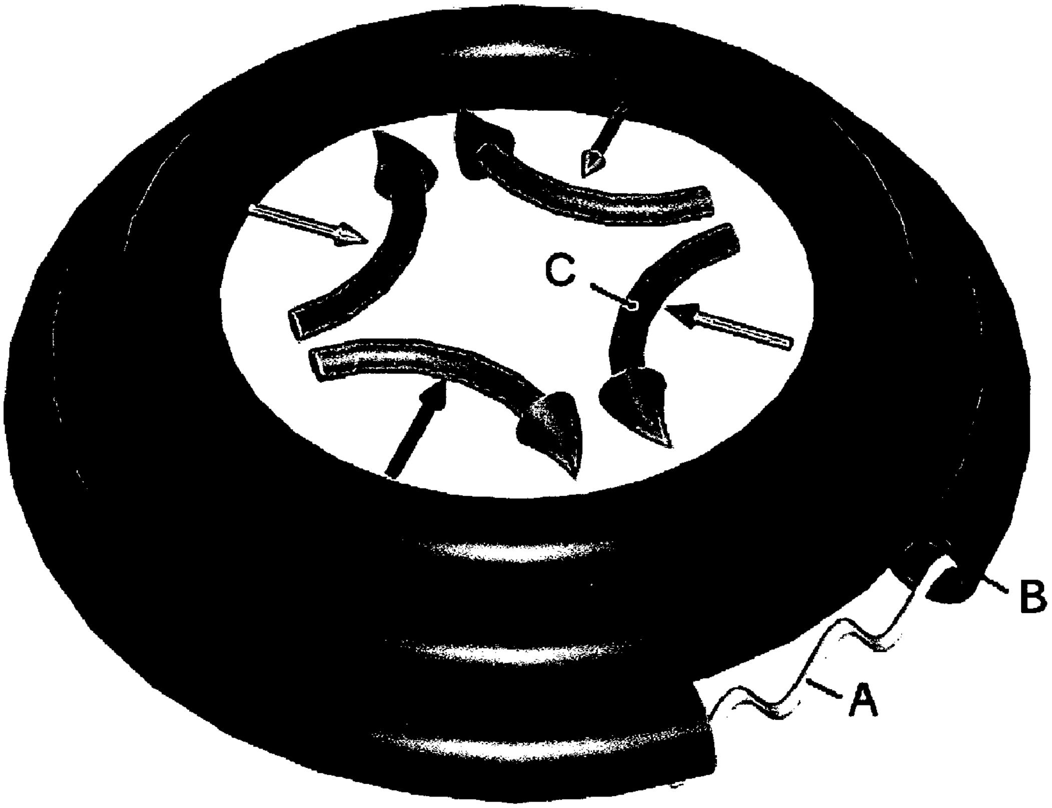

Figure 13

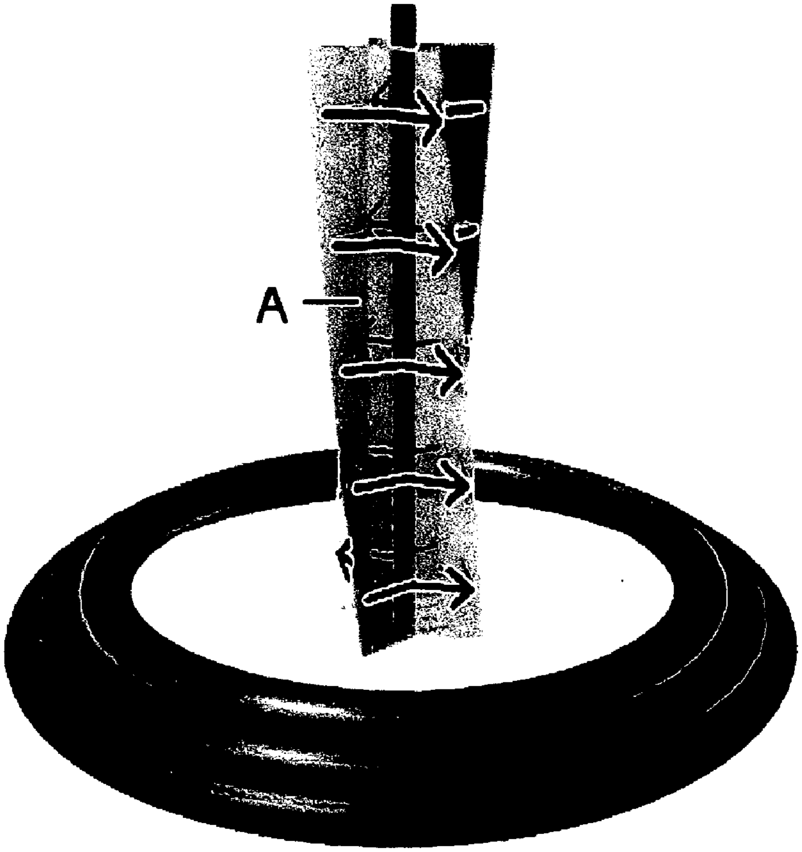

Figure 14

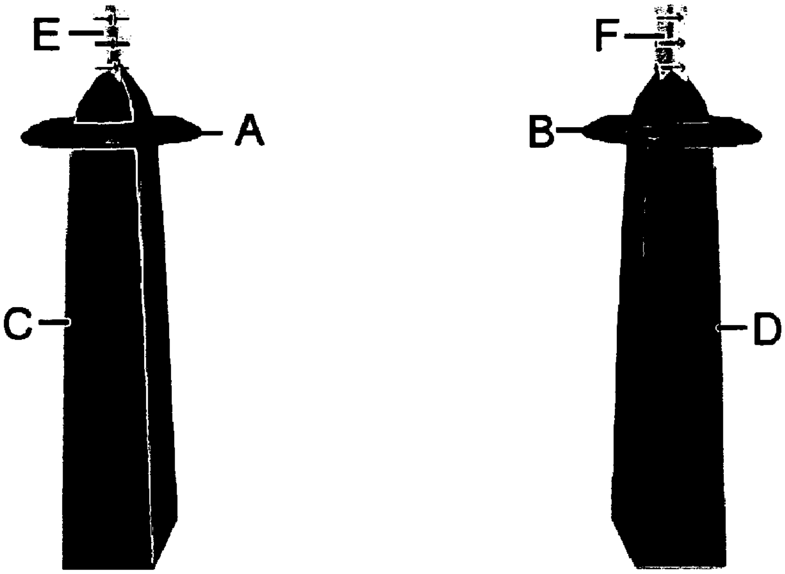

Figure 15

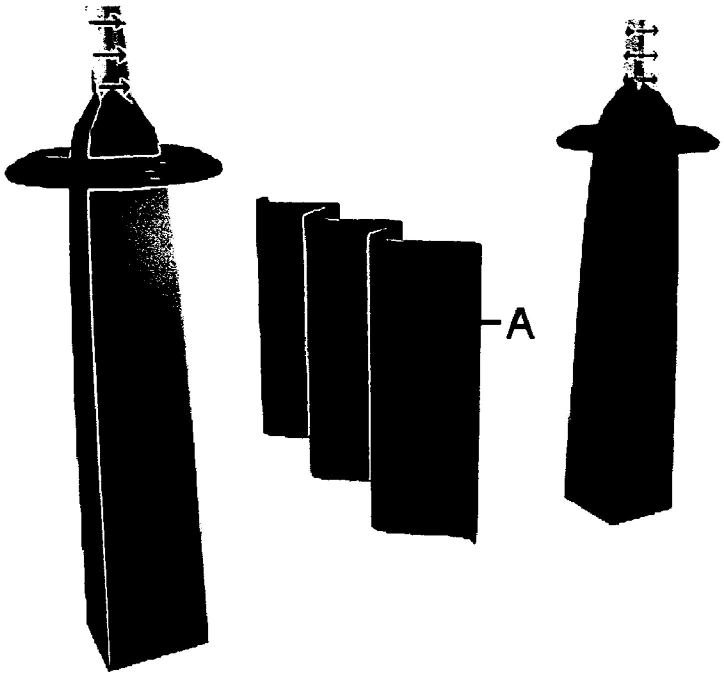

Figure 16

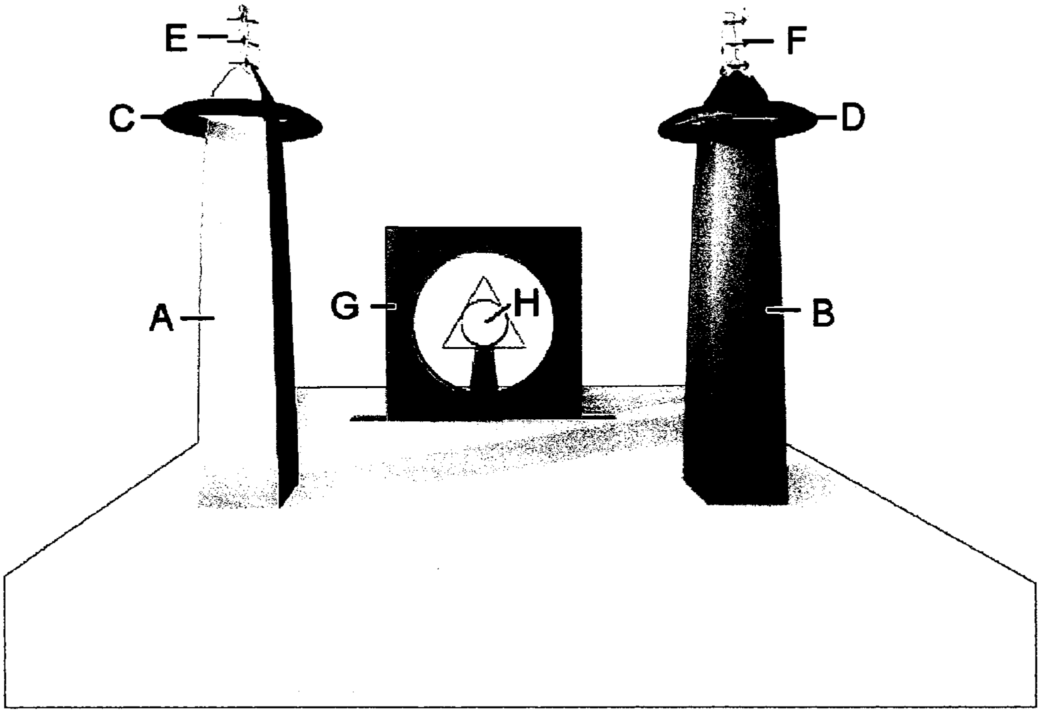

FULL BODY TELEPORTATION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a system that teleports a human being through hyperspace from one location to another using a pulsed gravitational wave traveling through hyperspace.

BACKGROUND OF THE INVENTION

[0002] The basis for this invention is an event, referring to **FIG. 1**, occurring on May 2, 2004, in which the inventor ("he") personally experienced a full-body teleportation while walking to the bus stop (A) along a road (B) that runs perpendicular to the nearby commercial airport runways where planes are landing. There is a wide iron grating (D) for water drainage that crosses the road at the center of the bus stop. The grating width is such that one has to make a concerted effort to jump across it in order to get from one side to the other. Approximately 50 meters from the iron grating, he (E) felt a vertical wave (F), similar to a flag waving in the breeze, traveling down the street toward the bus stop. The wave velocity was about 1 meter per second, which was slightly faster than his walking speed. In the next instance, he (G) found himself down the street near the corner of the next block. Realizing that he had passed the bus stop, he turned around to see the iron grating approximately 50 meters up the street in back of him. Because there was no recollection of having jumped across the iron grating nor of having passed the bus stop's yellow marker line, he realized that he had been teleported a distance of 100 meters while moving along with the traveling wave. It was obvious that the wave was pulsed because the front edge overtook the inventor, moved with him momentarily, and then the back edge of wave left him as it moved on down the street. While contemplating this sequence of events, he then looked up and saw in a span of a few seconds a twin-turboprop airplane (C) in the distance crossing above the road while making a shallow descent in order to land at the airport.

[0003] It took a number of days in order to understand this sequence of events. The explanation involves knowledge of a wide range of subjects such as gravitation physics, hyperspace physics, wormhole electromagnetic theory and experimentation, quantum physics, and the nature of the human energy field.

[0004] It is obvious from the above scenario that the airplane momentarily crossing perpendicular to the road generates the aforementioned pulse. Because the airplane has an engine on each wing, there are two propellers which conceivably are rotating out-of-phase with each other. That is, the blade of one propeller could be pointing up and the equivalent blade on the other engine could be pointing in a slightly different direction. Notice that the tip of the blade traces out a helix as the plane is landing.

[0005] In gravitation physics, referring to **FIG. 2**, it is known that two masses of mass m**1** and m**2** (A,B) attached by lever arms slightly offset by an angle δθ along the radial direction to the rotating shaft (C), will produce a gravitational wave (D) traveling perpendicular to the shaft. The mass and wave are referred to as the source and receptor respectively. Referring to a side view looking along the shaft **FIG. 3**, the product of the mass m times the angular acceleration a is a constant such that m**1**_a_**1** is equal to m**2**_a_**2**. The distance between the masses is length L, which makes an angle θ with the horizontal axis. The difference in time of travel to the receptor gives rise to a difference in phase δθ equal to the angular velocity ω of the rotating shaft times the length L times the cosine of the angle θ

$$\delta\theta = \omega L \cos(\theta)$$

[0006] At the receptor, the amplitude of the wave is equal to the mass times the acceleration times the phase difference divided by the radius r to the receptor

$$A = \frac{m_1 a_1}{r}\delta\theta \approx \left(\frac{m\omega L\sin(\theta)}{r}\right)(\omega L\cos(\theta)) = \frac{m\omega^2 L^2 \sin(2\theta)}{r}$$

Even though the turboprop airplane engines have a high rotational speed and a large separation distance between masses, the gravitational wave which is produced is small and not noticed. The problem is that the gravitational constant G in this dimension has such a small value equal to the speed of light c squared divided by the linear mass Ω of the universe

$$G = \frac{c^2}{\Omega} = \frac{(299792458\ \mathrm{m/s})^2}{1.346812891\cdot 10^{27}\ \mathrm{kg/m}} = 6.673200002\cdot 10^{-11}\ \frac{\mathrm{m}^3}{\mathrm{kgs}^2}$$

[0007] On the other hand, a gravitational wave traveling in hyperspace would be magnified enormously due to the face that the linear mass is so small. The magnitude of the gravitational constant in hyperspace can be estimated in the following manner. At the beginning of the 20th century, a man's parents were dying of tuberculosis. With their permission, he placed them and their beds on weighing scales. When each one passed away, each scale registered a drop in mass equal to 0.071 kilograms. This is the mass of the hyperspace energy being which resides in the physical body. Because hyperspace is co-dimensional with our dimension, the energy being interpenetrates the body and controls its movement.

[0008] Referring to **FIG. 4**, a human being has seven vortices (A through G) which are aligned along the centerline of the body. Each vortex is actually a co-gravitational field K which causes a pendulum placed in the field to spin in circles. For this reason, the K field has units of inverse seconds similar to an angular velocity. The vortex transports energy from our dimension to the energy being located in hyperspace. The gravitational field g and the co-gravitational field K are equivalent gravitationally to the electric E field and the magnetic B field found in electromagnetism. The equivalent gravitational solution to an electromagnetic problem can be obtained by substituting the following gravitational constants for the electromagnetic constants

Electromagnetic	Gravitational
q (charge)	m (mass)
ρ (volume charge density)	ρ (volume mass density)
σ (surface charge density)	σ (surface mass density)
λ (line charge density)	λ (line mass density)
J (convection current density)	J (mass current density)

-continued

Electromagnetic	Gravitational
E (electric field)	g (gravitational field)
B (magnetic field)	K (co-gravitational field)
ϵ_0 (permittivity of space)	$-1/4\pi G$
μ_0 (permeability of space)	$-4\pi G/c^2$
$-1/4\pi\epsilon_0$ or $-\mu_0\ c^2/4\pi$	G (gravitational constant)

[0009] Referring to **FIG. 5**, each vortex is connected through the pineal gland by light cords to a separate hyperspace quantum well having its own frequency and dimension. The reason for this separation is that the conical spiritual eye, attached to vortex B, has to have its own energy structure which is different from vortex (A) which is connected to the quantum energy field in which the mental processes are developed. Due to the high speed of light in our dimension, the quantum wells are the size appropriate to molecules and atoms. In hyperspace, where the speed of light is one meter per second, the quantum wells are huge and can be manipulated. This manipulation has shown that the quantum wells are in the shape of a cube about a meter on a side. This makes the whole structure about seven meters tall for a volume of seven cubic meters. Thus the mass density ρ of hyperspace is about

$$\rho_h = \frac{.071\ \mathrm{kg}}{7\ \mathrm{m}^3} = .01\ \frac{\mathrm{kg}}{\mathrm{m}^3}$$

which per unit area is the same value. Therefore the hyperspace gravitational constant is equal to

$$G_h = \frac{c^2}{\Omega_h} = \frac{\left(1\ \frac{\mathrm{m}}{\mathrm{s}}\right)^2}{.01\ \mathrm{kg/m}} = 100\ \frac{\mathrm{m}^3}{\mathrm{kgs}^2}$$

The enormous magnification of the gravitational constant is therefore of the order of

$$\frac{G_h}{G} = \frac{100}{6.6732 \cdot 10^{-11}} \approx 1.5 \cdot 10^{12}$$

The question is how does this amplified gravitational wave created by the rotating propellers and turbines get into hyperspace from our dimension?

[0010] The answer comes from experiments done using the ancient Chinese form of breathing known as Chi Kung. Using this breathing technique, we have been able to levitate the human body over six feet in the air. The internal temperature of the stomach is around 200 degrees Fahrenheit. By simultaneously squeezing the diaphragm to bring hot air up through the lungs, and breathing through the nose to bring cold air down, rotating vortices are generated in the lung passages when these two air masses meet and twist around each other as depicted in the famous Yin-Yang diagram. Because the lung has variable diameter passages from the large diameter at the throat to the final small air sacs, there is a spectrum of rotating frequencies.

[0011] From quantum physics it is known that if there is a temperature fluctuation occurring among a group of harmonic oscillators in the environment, then Planck's reduced constant $\hbar$ is increased by the cotangent of the constant times the frequency ω of the oscillator divided by twice Boltzmann's constant k times the temperature T

$$\hbar = \hbar\coth\left(\frac{\hbar\omega_n}{2kT}\right)$$

[0012] The effect of increasing Planck's constant, referring to **FIG. 6**, can be seen in the tetrahedron diagram. This diagram, of which there are now over 4000, plots the natural logarithm of mass on the vertical axis versus the natural logarithm of wavelength on the horizontal axis. In terms of mathematics, it is a subspace logarithmic manifold which projects geometrically the physics constants into our **4**D spacetime dimension. That is, it is the geometry of the tetrahedron circumscribed by the sphere that determines the mass of the proton and electron. The mass of the electron times its wavelength is equal to the mass of the proton times its wavelength which in turn is equal to Planck's constant h divided by the speed of light c

$$m_e\lambda_e = m_p\lambda_p = \frac{h}{c}$$

Taking the natural logarithm of the above equation shows that the mass plus the wavelength is equal to what is termed the base constant

$$\ln(m_e) + \ln(\lambda_e) = \ln\left(\frac{h}{c}\right) = -95.91546344$$

which is represented in **FIG. 6** by the 45 degree line (A) from point (a) on the horizontal axis to the vertical axis at point (b). The electron is located at point (c) which is the intersection of the electron wavelength (B) with line (A). The electron wavelength (B) reflects off the sphere (D) at points (d) and (e) and returns along line (C) as the electron mass. As shown in tetrahedron diagram tet0565, stored in the Library of Congress, the clockwise path of the electron transitions into the counter-clockwise path of the proton showing that the electron and proton are one and the same particle. Because the electron and proton travel in opposite directions along the path, they have the same charge but of opposite sign.

[0013] Our dimension is represented by Planck box (E) which is bounded by the Planck mass and the Planck wavelength. The Planck mass is equal to the linear mass of the universe times the Planck length which is the bottom limit of our dimension. The Planck wavelength is 2π times the Planck length. Notice that the electron is located within the Planck box.

[0014] Referring to **FIG. 7**, if there is an increase in Planck's constant due to the temperature fluctuations among the harmonic oscillators, the 45 degree base line (A) moves to the left on the tetrahedron diagram as shown by line (F).

Because of the increase in the base constant, there is a corresponding increase in the electron mass and wavelength. The electron moves from point (c) to point (f) which places it at the edge of the Planck box (F) which is the boundary between space and hyperspace. At point (f), the electron is essentially no longer in our dimension.

[0015] Referring to **FIG. 8**, imagine a box (A) filled with nine electron oscillators (B). If Planck's constant is increased near the three oscillators in the middle, these electrons will leave this dimension. This leaves six oscillators as shown in the box (C). However, box (C) is the equivalent of box (D) in which there are still nine positive mass oscillators together with 3 negative mass oscillators. Thus there is an accumulation of negative energy (−ρ) when information is lost from the environment to another dimension.

[0016] Dr. Kip Thorne, who co-authored the book *Gravitation* with Dr. Archibald Wheeler of Princeton University, has shown in a General Relativity spacetime curvature calculation that negative energy is required to open and stabilize the throat of a wormhole between space and hyperspace. The accumulation of negative energy in the aforementioned example generates wormholes between into hyperspace. Hyperspace has a low energy density because of the reduced speed of light in that dimension. Ordinarily, energy would not flow from hyperspace to space because space has a higher potential than the potential of hyperspace. This, of course, is the reason that the body vortices can flow energy into the energy field of the human being who is located in hyperspace. By creating negative energy, the potential becomes reversed such that low density hyperspace energy flows into our dimension as seen by the positive head

$$PE=\rho_{hyperspace}-(-\rho_{space})=+2\rho$$

The low-density energy fills the body which allows a human being to float upwards like a helium balloon as verified by Chi Kung breathing as well as spinning on a motorized platform known as the Chakra Vortex Accelerator. The latter device resulted in the first mechanical means to produce anti-gravity.

[0017] The process of creating spinning thermal fluctuations is the same as found in the hot air vortices created by the jet airplanes landing at the airport near the road where the full-body teleportation occurred. Large vortices are created over the wing of the airplane at the same time that the turbine engines are spinning hot vortices into relatively cold air. These conditions produce wormholes between space and hyperspace. It takes a twin turboprop airplane landing behind the jet to generate the gravitational wave in the region where the wormholes have formed. The gravitational wave then traverses the wormholes into hyperspace, becoming highly amplified due to the change in linear mass and speed of light. Because the propeller blades are co-linear with the road, the gravitational wave travels in the direction along the road where it was encountered by the inventor.

[0018] From experiments with cavitating bubbles (see patent application Cavitating Oil Hyperspace Energy Generator), it was found that it is possible to produce a wormhole if the surfaces of the bubble collapse asymmetrically. A symmetric collapse of a spherical bubble produces enormous spacetime curvature distortions. An asymmetric collapse, using a magnetic field to distort the collapse, produces, in addition to the same severe spacetime distortions, negative energy as the bubble collapses. Due to some General Relativity considerations, the wormhole that is created starts rotating in a manner similar to the beacon light produced by a lighthouse.

[0019] Referring to **FIG. 9**, due to the forward helical motion (A) of the propellers (B) as the airplane crosses the road, the pulsed gravitational wave (C) is skewed backward at an angle (D). Due to the wormholes created by the presence of thermal vorticity fluctuations generated by the wing and turbines of the airplane, this skewed wave moves into hyperspace where it is highly magnified and detected by the inventor.

[0020] Referring to top view **FIG. 10**, the gravitational wave (A) causes a skewed compression and expansion of the hyperspace quantum wells (B) which constitute the human energy being. Due to this asymmetric distortion in the xy-plane, the quantum wells take the physical body out of dimension as long as the wave pulse is traveling with the human energy field. Once the back edge of the gravitational wave moves on past the quantum wells, the body is then brought back into dimension.

SUMMARY OF THE INVENTION

[0021] It is the object of this invention to teleport a human being from one location to another by creating a pulsed gravitational wave traveling through hyperspace that asymmetrically compresses and expands the quantum wells of the human energy being. This spacetime curvature distortion of the hyperspace quantum wells pulls the physical body out of dimension such that the human being is teleported along with the wave. As the pulsed wave moves on past the quantum wells, the human is brought back into dimension at some distant location. The invention requires (1) a device that will generate a wormhole between space and hyperspace, and (2) a device that will generate a gravitational wave which can be inserted through the wormhole.

[0022] Referring to **FIG. 11**, a magnetic vortex wormhole generator has already been developed which generates a wormhole between space and hyperspace as described in a previous patent application entitled Magnetic Vortex Wormhole Generator. Using this generator, it was found that smoke blown through one side of the coil does not appear on the other side of cylindrical coil. The smoke flows through the wormhole and appears in a hyperspace co-dimension. It was this experiment that resulted in making first contact with the androids of the Grey aliens who told me, in a remote viewing session, that "We saw you blowing smoke into hyperspace."

[0023] The wormhole generator consists of two concentric cylindrical coils (A,B), one of larger radius than the other, made of thin transformer iron laminate wrapped in opposite directions with one continuous wire driven by a sinusoidal current. The solenoidal coil generates a magnetic field through the laminate. Because the electrical current flows in opposite directions at different radii through the two windings, bucking electric fields (C) are created along the centerline of the generator. These radially-offset magnetic fields and bucking electric fields, as shown by a calculation using Einstein's General Theory of Relativity, generate both an enormous spiking spacetime curvature and negative energy at small radius along the centerline where the wormhole is formed. The gravitational wave generator is then coupled to this wormhole generator.

[0024] Referring to **FIG. 12**, it is known from gravitation physics that injecting an electromagnetic wave (A) into a hollow toroidal waveguide (B) produces a hyperbolic spacetime curvature stress (C) in the plane of the waveguide. The tips of the arrows indicate compression and the tail of the arrows indicate expansion or stretching of spacetime. The reason for this spacetime curvature is because the waveguide forces the electromagnetic wave to curve around and travel in a circle. Spacetime has to compensate for this toroidal-generated stress by creating hyperbolic lines of stress in the inner plane of the toroid so that the overall spacetime curvature is zero. For a greater gravitational effect, three toroidal waveguides, phased 120 degrees apart, are used to seal off the curvature.

[0025] Referring to **FIG. 13**, the three toroids create a rotating, twisting, vertical propagating gravitational field (A) through the centerline of the toroids provided that the period of the electromagnetic wave is twice the period of the gravitational wave. This phase relationship is adjusted by selecting the correct radius for the frequency of the monochromatic wave.

[0026] In order to effectively use this gravitational wave, referring to **FIG. 14**, three phased toroidal waveguides (A,B) are mounted at the top of each of two identical square granite obelisks (C,D). The two obelisks are offset by a short distance between them. As the vertical gravitational wave rotates around along the vertical axis inside the obelisk, the edges of the square obelisks are compressed and expanded such as to create two cylindrical asymmetric gravitational waves traveling radially outward.

[0027] Referring to **FIG. 15**, these waves meet to form a plane gravitational wave (A) which travels down the centerline between the two obelisks.

[0028] Referring to **FIG. 16**, the full body teleportation system consists of the twin granite obelisks (A,B) on which are mounted near the top of each the toroidal waveguides (C,D) which produce the pulsed gravitational waves (E,F) that run the length of the obelisks. Because the gravitational wave is rotating inside the obelisk, the granite stone undergoes a very small asymmetrical compression and expansion. A cylindrical gravitational wave propagates out from each obelisk such that along the centerline between the two there is generated a plane gravitational wave. This wave enters the wormhole (H) created by the magnetic vortex generator which is located a short distance from and parallel to the obelisks. The wave is amplified by a factor of almost 10^{13} when it enters the hyperspace co-dimension.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0029] **FIG. 1**. Perspective view of site where full-body teleportation occurred.

[0030] **FIG. 2**. Perspective view of gravitational wave generator.

[0031] **FIG. 3**. Planar view of gravitational wave generator.

[0032] **FIG. 4**. Perspective view of seven vortices of human energy being.

[0033] **FIG. 5**. Perspective view of seven large quantum wells of human energy being.

[0034] **FIG. 6**. Tetrahedron diagram showing Planck's constant and electron.

[0035] **FIG. 7**. Tetrahedron diagram showing electron moving out of dimension.

[0036] **FIG. 8**. Perspective view showing production of negative energy.

[0037] **FIG. 9**. Perspective view of skewed gravitational wave produced by propellers.

[0038] **FIG. 10**. Planar view of skewed quantum wells deformed by gravitational wave.

[0039] **FIG. 11**. Perspective view of magnetic vortex wormhole generator.

[0040] **FIG. 12**. Perspective view of hyperbolic lines of stress generated by toroidal waveguide.

[0041] **FIG. 13**. Perspective view of rotating, twisting, propagating gravitational wave generated by toroidal waveguides.

[0042] **FIG. 14**. Perspective view of toroidal waveguides attached to obelisks.

[0043] **FIG. 15**. Perspective view of gravitational wave generated by obelisks.

[0044] **FIG. 16**. Perspective view of magnetic vortex wormhole generator and obelisk gravitational wave generator.

DETAILED DESCRIPTION OF THE INVENTION

[0045] 1. The obelisks are quarried out of granite stone and cut with a large-diameter diamond saw that is used in highway construction. The beveled piece at the top is cut separately and cemented in place. A tapered aluminum bracket holds the toroids in place.

[0046] 2. The electronics for the magnetic vortex generator are similar to that used in the patent application Magnetic Vortex Wormhole Generator.

[0047] 3. The electronics for the toroidal waveguides is the familiar stub and coaxial cable driven by an amplifier and pulsed variable-frequency generator.

I claim:

1. A full body teleportation system consisting of:

generating a pulsed gravitational wave which propagates through a magnetic vortex wormhole generator; and

generating a wormhole with the magnetic vortex generator whereby the pulsed gravitational wave traverses through the wormhole and enters into hyperspace where the wave is enormously magnified due to the lower speed of light in that dimension.

2. The method of claim 1, wherein the step of generating the pulsed gravitational wave comprises:

using two granite stone obelisks;

mounting monochromatic-wave toroidal waveguides on top of each obelisk to create a rotating, twisting, propagating gravitational wave through the vertical axis of each obelisk; and

creating a cylindrical compression and expansion in each obelisk to produce a plane gravitational wave traveling down the centerline between the two obelisks.

3. The method of claim 1, wherein the step of generating a wormhole into hyperspace comprises:

using two concentric cylindrical solenoidal coils of different radii connected by a single wire wrapped in opposite directions on thin iron transformer laminate;

generating bucking electric fields down the centerline of the vortex generator which creates a spacetime curvature distortion with negative energy in accordance with Einstein's General Theory of Relativity.

4. A teleportation system comprising:

generating a gravitational wave traveling through hyperspace which interacts with the human energy being; and

pulling the human energy being and physical body out of dimension when interacting with the pulsed gravitational wave such that the person is teleported from one location to another through hyperspace and back again into our **4**D spacetime dimension.

* * * * *

US 20060145019A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0145019 A1**

St. Clair (43) **Pub. Date: Jul. 6, 2006**

(54) TRIANGULAR SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/017,093**

(22) Filed: **Dec. 20, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) ABSTRACT

A spacecraft having a triangular hull with vertical electrostatic line charges on each corner that produce a horizontal electric field parallel to the sides of the hull. This field, interacting with a plane wave emitted by antennas on the side of the hull, generates a force per volume combining both lift and propulsion.

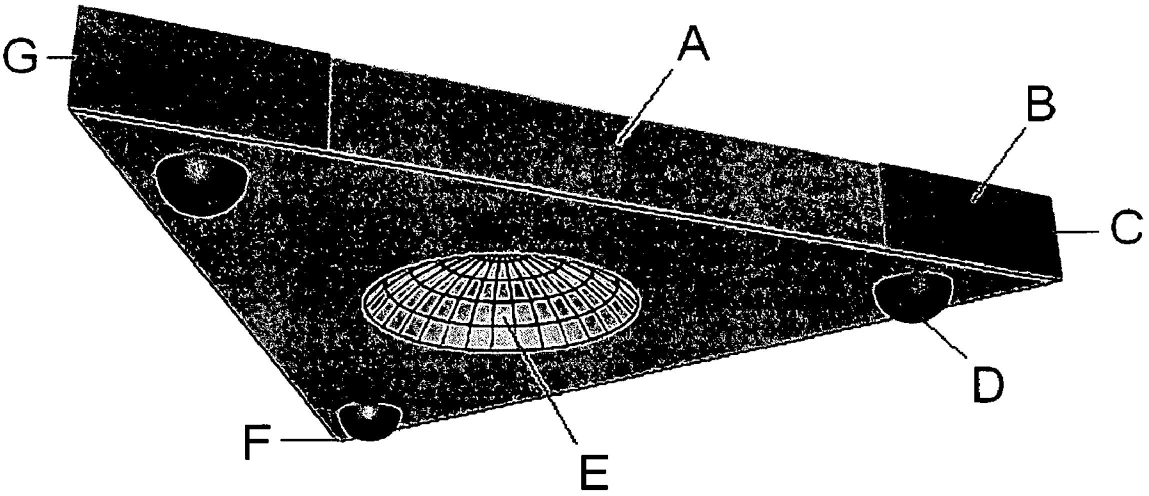

Figure 1

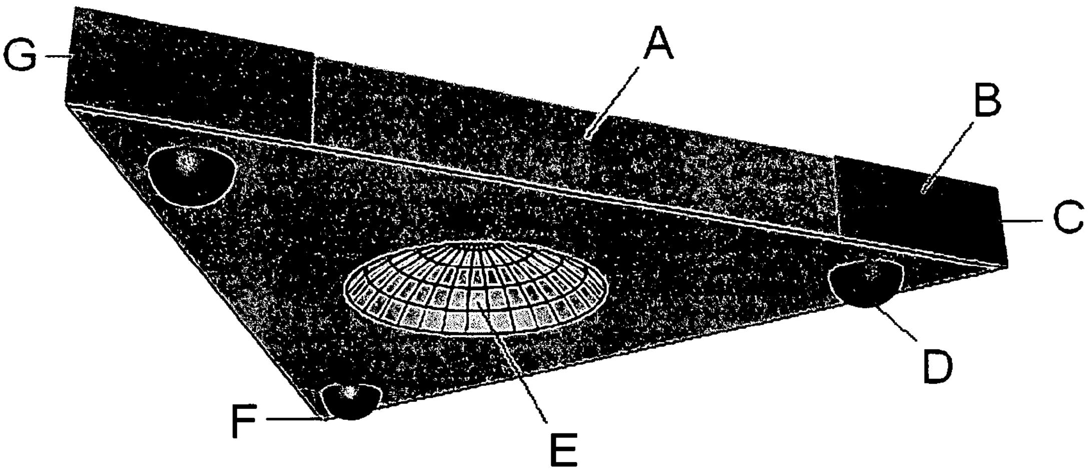

Figure 2

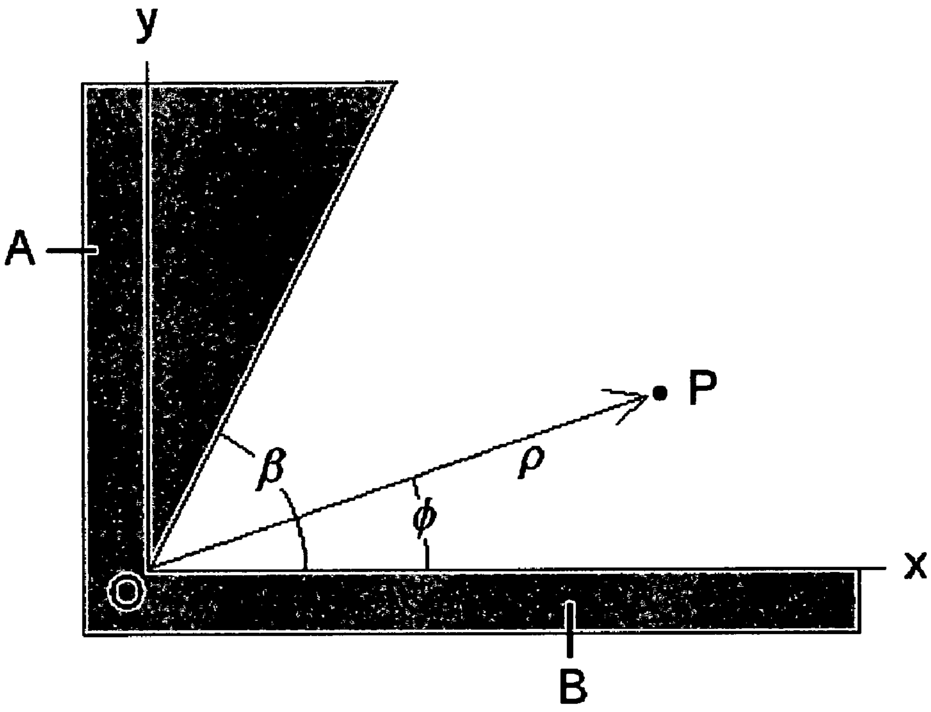

Figure 3

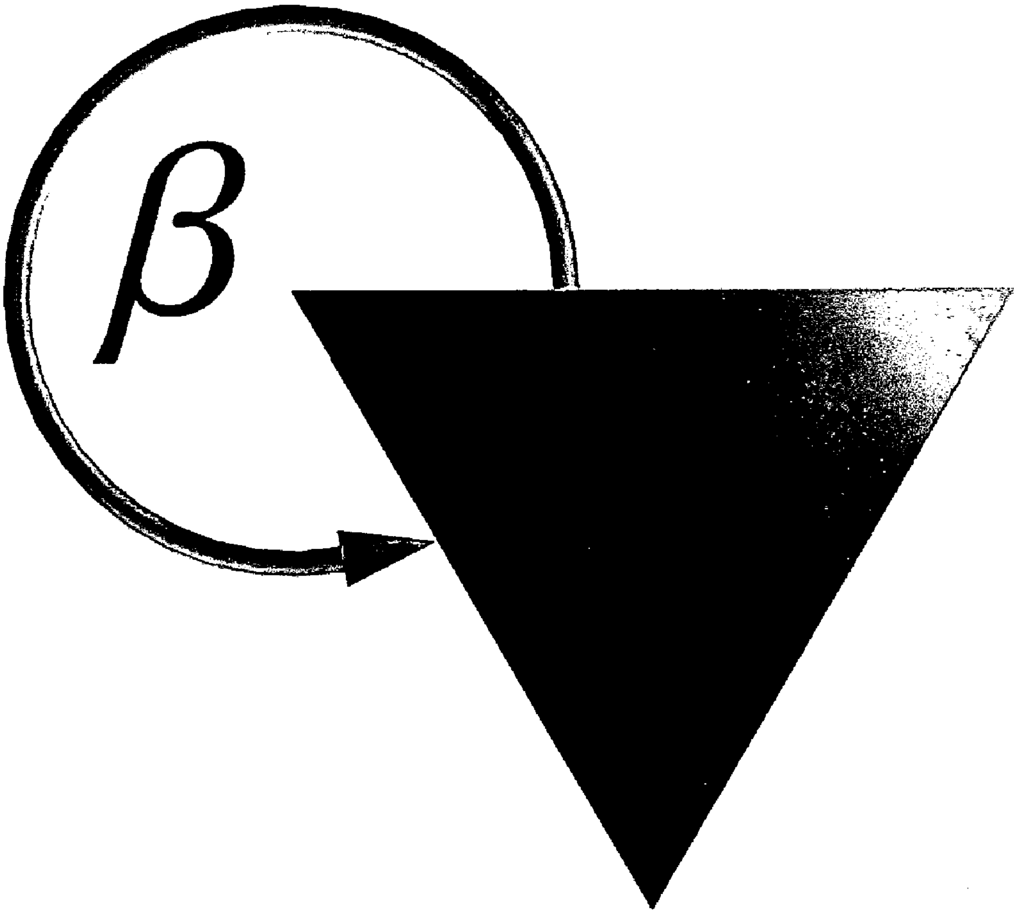

Figure 4

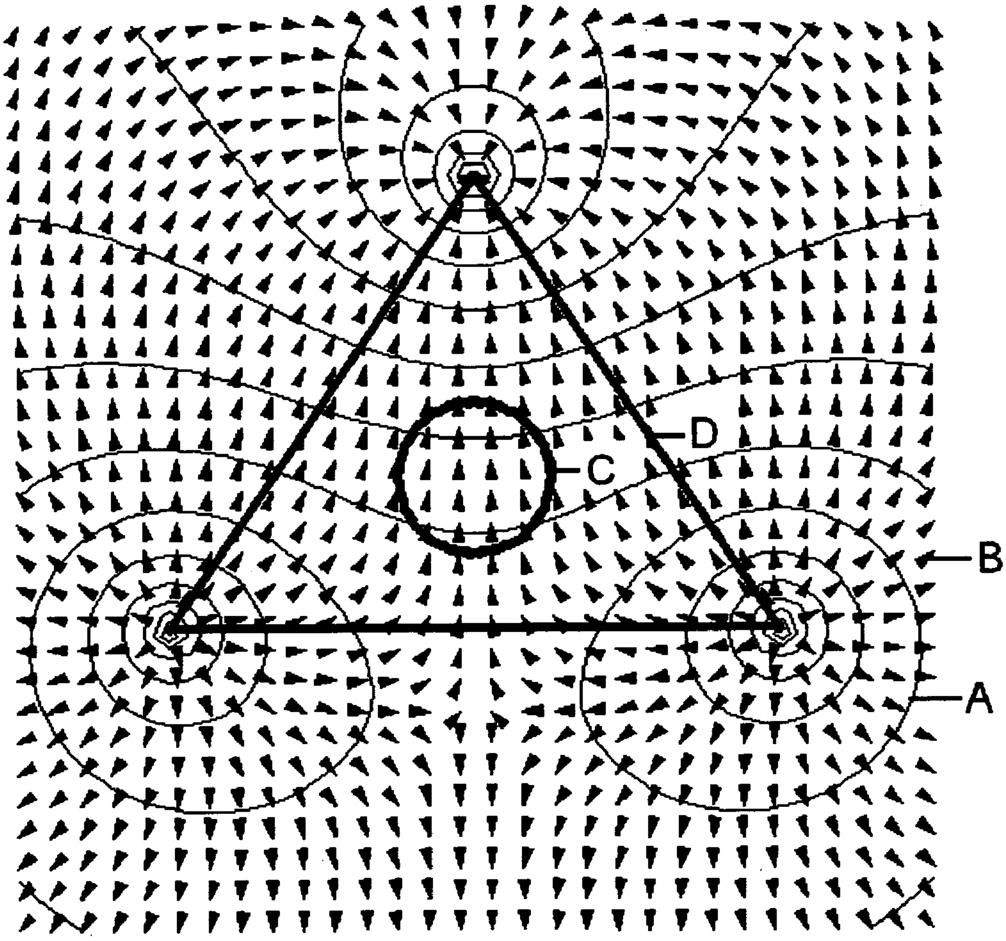

Figure 5

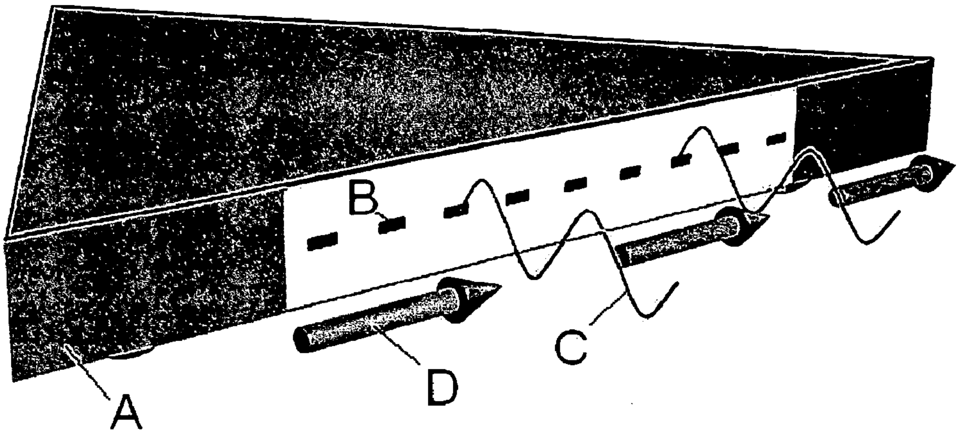

Figure 6

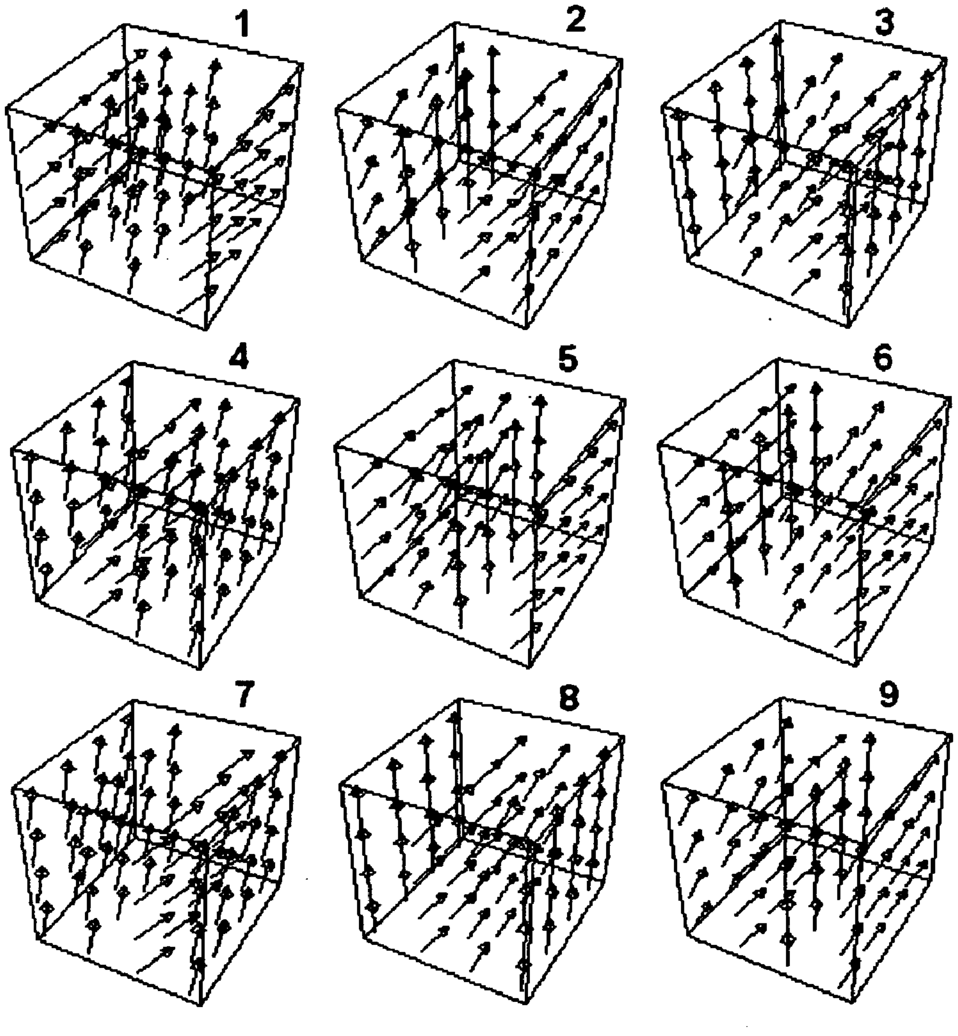

Figure 7

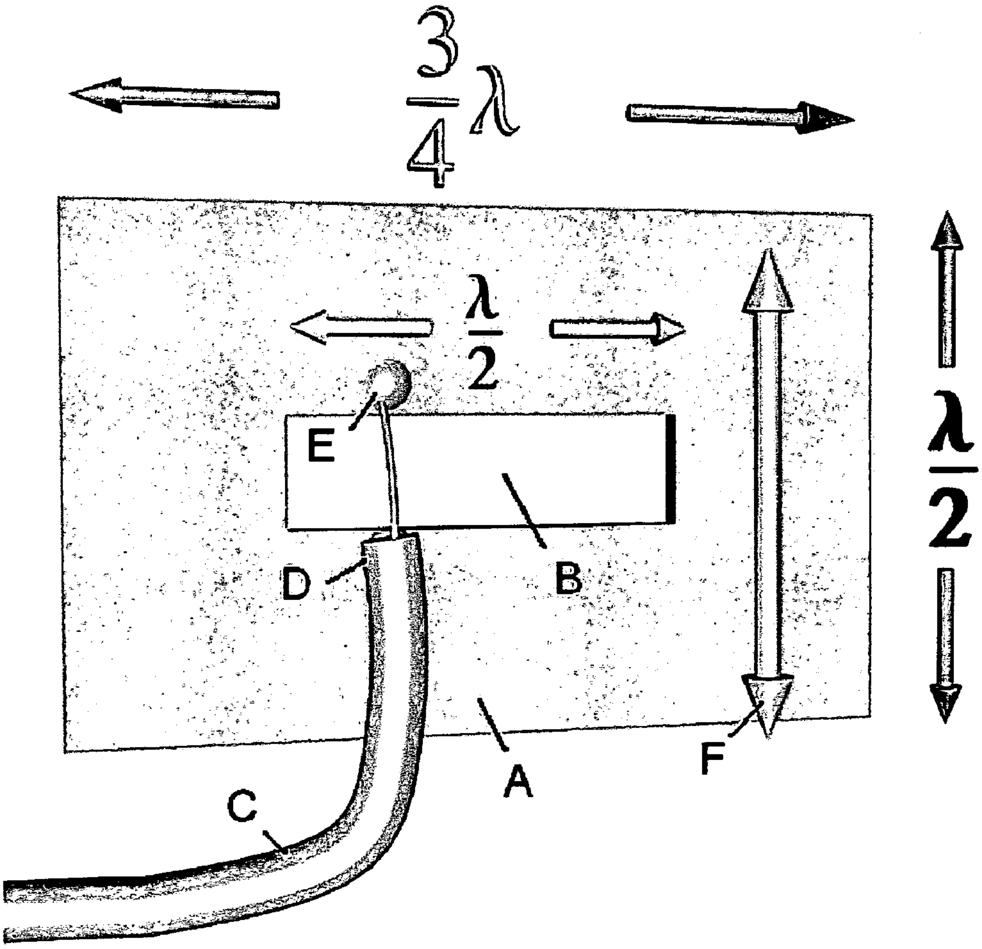
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
E
D
B
A
F
C

TRIANGULAR SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft having a triangular hull with vertical electrostatic line charges on each corner. The line charges create a horizontal electric field that, together with a plane wave emitted by antennas on the side of the hull, generates a force per volume providing a unique combination of both lift and propulsion.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, the spacecraft has a hull in the shape of an equilateral triangle. A parabolic antenna (E) is centrally located in the bottom of the hull. An array of horizontal slot antennas is located along the side of the hull (A). Each back corner (F,G) has a corner conducting plate which is charged to a positive voltage +V. The forward corner (C) has a conducting plate charged to a negative voltage −V. A motion control hemisphere (D) is located on the bottom surface in each of the three corners.

[0003] Referring to **FIG. 2**, two planes (A,B) intersect at the origin O at an opening angle β. Each plane (x,y) is charged to a voltage V. The potential at point P is determined in polar coordinates {ρϕ}. The Laplace equation for the potential Φ in polar coordinates is given by:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\Phi}{\partial\phi^2}=0$$

Using a separation of variables solution, the potential is given as the product of two functions:

$$\Phi(\rho,\phi)=R(\rho)\Psi(\phi)$$

which when substituted into the Laplace equation becomes:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)+\frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=0$$

Since the two terns are separately functions of ρ and ϕ respectively, each one has to be constant with the sum of the constants equal to zero:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)=v^2 \quad \frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=-v^2$$

These two equations have solutions:

$$R(\rho)=a\rho^{v+b\rho-v}$$

$$\psi(\phi)=A\cos(v\phi)+B\sin(v\phi)$$

The azimuthal angle ϕ is restricted to a value in the range 0≦ϕ≦β. The boundary condition is that the potential Φ is equal to V for any radius ρ when ϕ=0 and ϕ=β. This means that v has to be an integer value of π so that the sine function is zero:

$$\sin(v\beta)=\sin\left(\frac{m\pi}{\beta}\beta\right)=\sin(m\pi)=0 \quad m=1,2\ldots$$

which in turn means that the coefficient A of the cosine term has to be zero in the solution above. Choosing b=0 makes the general solution for the potential equal to:

$$\Phi(\rho,\phi)=V+\sum_{m=1}^{\infty}a_m\rho^{m\pi/\beta}\sin(m\pi\phi/\beta)$$

which shows that when the angle is zero, the sine is zero and the potential is V. If the angle is β, then there is a multiple of π such that the sine is zero again.

[0004] Because the series involves positive powers of the radius, for small enough ρ, only the first term *m*=1 in the series is important. Thus around ρ=0, the potential is approximately

$$\phi(\rho,\phi)\approx V+a,\rho^{\pi/\beta}\sin(\pi\phi/\beta)$$

[0005] The electric field component is the negative gradient of the potential:

$$E_\phi(\rho,\phi)=-\frac{1}{\rho}\frac{\partial\Phi}{\partial\phi}=-\frac{\pi a_1}{\beta}\rho^{(\pi/\beta)-1}\cos(\pi\phi/\beta)$$

The surface charge distribution σ at ϕ=0 and ϕ=β is equal to the electric field perpendicular to the surface times the permittivity of space ϵ_0:

$$\sigma(\rho)=\varepsilon_0E_\phi(\rho,0)=-\frac{\varepsilon_0\pi a_1}{\beta}\rho^{\frac{\pi}{\beta}-1}$$

Notice that if angle of intersection β is less than π, then the equation says that there is a very small radius to a positive power which means little charge density accumulation.

[0006] Referring to **FIG. 3**, the value of β, in the case of the triangular hull, is equal to 360° less 60° for a total of 300° or:

$$\beta=\frac{300}{180}\pi=\frac{5}{3}\pi$$

$$\rho^{\frac{\pi}{\frac{5}{3}\pi}-1}=\frac{1}{\rho^{\frac{2}{5}}}$$

which says that there is a charge density singularity to the two fifths power for small radius. Thus, the corner plates on the hull create a huge line charge density along the sharp vertical corner edge. The equation for the potential of a line charge density is given as:

$$\Phi(x,y)=-\frac{\lambda}{2\pi\varepsilon_0}\mathrm{Ln}((x-x_0)^2+(y-y_0)^2)$$

where λ is the charge per unit length in the vertical z-direction, and x_0 and y_0 are the location of the line charge in the xy-plane.

[0007] Referring to **FIG. 4**, the triangular hull (D) is plotted together with the potential contours (A) and the electric field arrows (B) created by the three corner line charges. The line charges are perpendicular to the paper. Notice that the electric field arrows are parallel crossing the center parabolic antenna (C). The electric field is also parallel to the sides (D) of the triangle.

[0008] Referring to **FIG. 5**, along the side of the triangle (A), an array (B) of horizontal slot antennas emit electromagnetic waves that have a vertically polarized electric E field (C). These traveling waves interact with the electric field (D) produced by the line charges on the corners of the triangle.

[0009] Using differential forms mathematics, this combination of fields is represented by the Hodge star of the differential of the wedge product of the two fields. The antenna electromagnetic field is a combination of a traveling magnetic field B_w, and electric field E_w. The stationary field E created by the line charges is perpendicular to the traveling wave.

$$*\,d(E\wedge(B_w+E_w\wedge dt))\frac{\varepsilon}{c}=\frac{\text{force}}{\text{volume}}$$

where ϵ is the linear capacitance of space and c is the speed of light. Thus there is a force per volume around the hull.

[0010] This combination of fields produces a spacetime curvature as determined by Einstein's General Theory of Relativity. The traveling electric field has an amplitude in the vertical z-direction and travels in the x-direction

$$E_{w=Ez}\cos(x-t)$$

The Faraday electromagnetic tensor contains all the electric and magnetic fields in all the {x,y,z} directions. The first row and first column contain the two electric fields

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & E_x & 0 & E_z\cos(x-t) \\ x & E_x & 0 & 0 & 0 \\ y & 0 & 0 & 0 & 0 \\ z & E_z\cos(x-t) & 0 & 0 & 0 \end{array}$$

The stress exerted on spacetime occurs in the xx, yy and zz-direction as calculated from the stress-energy tensor T of gravitational physics

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\mu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor for Cartesian space

$$g_{\alpha\beta}=\begin{array}{c|cccc|} t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

where the diagonal components are the coefficients of the elementary spacetime length ds squared

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

The calculation produces three stresses T^{xx},T^{yy} and T^{zz} in their respective {x,y,z} directions.

[0011] Referring to **FIG. 6**, these three stresses are plotted together as a **3**D vector field animated over time in nine frames. The graphs show that there is a lift force as depicted by the vertical arrows as well as a force of propulsion as shown by the interspersed horizontal arrows. With the passage of time, these vectors exchange places with each other so that the lift becomes the propulsion and vice versa, creating a wavy stress-energy field around the hull.

SUMMARY OF THE INVENTION

[0012] This invention is a spacecraft with a triangular hull having charged flat plates on the vertical corners of the three sides. The two rear corners are charged to a potential V. The forward corner is charged to a potential −V. The 60° angle on the corner creates a line charge density singularity that produces a huge horizontal electric field pointing from the back to the front of the craft which is also parallel to the sides of the triangle. An array of horizontal slot antennas located on the sides of the triangular hull produce an electromagnetic wave with the electric field polarized in the vertical direction. This combination of fields produces a spacetime force in both the vertical and horizontal directions such that the spacecraft receives a lift force and a force of propulsion.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1**. Perspective view of triangular spacecraft.

[0014] **FIG. 2**. Drawing of the intersection of two charged plates in order to calculate the charge density in the corner.

[0015] **FIG. 3**. Perspective view of the corner angle β for the equilateral triangle.

[0016] **FIG. 4**. Planar **2**D graph showing the electric field produced by three line charges on the corners of the triangular hull.

[0017] **FIG. 5**. Perspective view of electric field produced by the linear charge interacting with the traveling electromagnetic wave produced by the slot antenna.

[0018] **FIG. 6**. **3**D vector animation of the lift and thrust force generated by the fields.

[0019] **FIG. 7**. Perspective view of slot antenna.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Referring to **FIG. 7**, the antenna (A) is made out of sheet copper in which a rectangular horizontal slot (B) has been notched out using a die press and sheet metal fixture. A coaxial cable from the amplifier and frequency generator is attached across the slot by soldering the outer cable (D) to one side of the slot and the inner cable (E) to the other side of the slot. This creates the positive and negative charges across the gap which forms the vertical electric field (F) which radiates out perpendicularly to the copper sheet.

[0021] Although the invention has been described with reference to specific embodiments, such as a particular antenna system, those skilled in the art will appreciate that many modifications and variations are possible without departing from the teachings of the invention. All such modifications and variations are intended to be encompassed within the scope of the following claims.

1. A spacecraft comprised of the following components:

(a) a triangular hull in the form of an equilateral triangle;

(b) two copper plates attached on opposite vertical sides at each of the three corners of the hull (**1***a*) such that a sharp vertical edge is formed where they come together;

(c) an electrostatic generator used to charge the back two copper-cladded corners (**1***b*) to a high positive voltage, and the third forward copper-cladded corner to a high negative voltage;

(d) a horizontal slot antenna array mounted-on the sides of the hull; and

(e) a frequency generator, antenna and coaxial cables to drive the antenna array (**1***d*).

2. To create, by claims (**1***a*, **1***b*, **1***c*), an intense vertical line charge at the corners (**1***b*) and a horizontal electric field that that is parallel to the sides of the hull (**1***a*);

3. To create, by claims (**1***d*,**1***e*), an electromagnetic wave with a vertically polarized electric field traveling outward from the side of the hull (**1***a*); and

4. To create, by claims (**2**,**3**), an interaction of the electrostatic field (**2**) with the electromagnetic wave (**3**) such that a combined spacetime curvature pressure is generated on the hull in the upward and forward direction to produce lift and propulsion respectively.

* * * * *

March 25, 1969 W. C. BROWN ET AL 3,434,678

MICROWAVE TO DC CONVERTER

Filed May 5, 1965

Sheet 1 of 3

FIG. 1

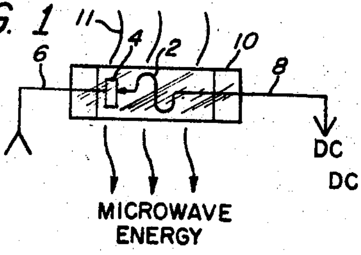
11
4
2
10
6
8
DC
MICROWAVE ENERGY

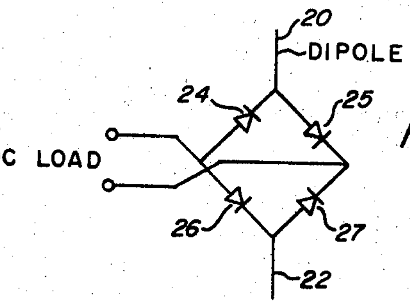
20
DIPOLE
24
25
DC LOAD
26
27
22

FIG. 2

FIG. 3

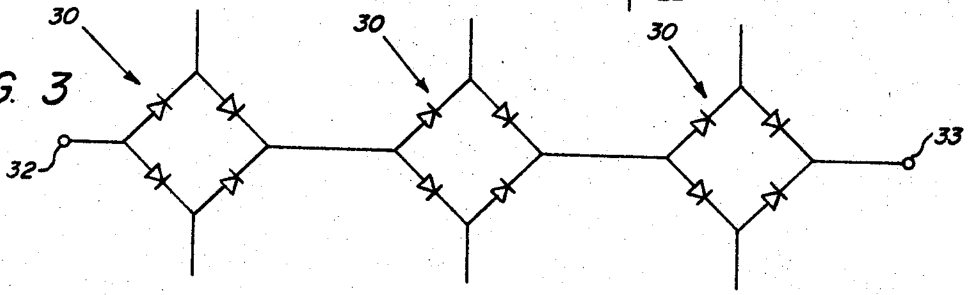
30
30
30
32
33

FIG. 4

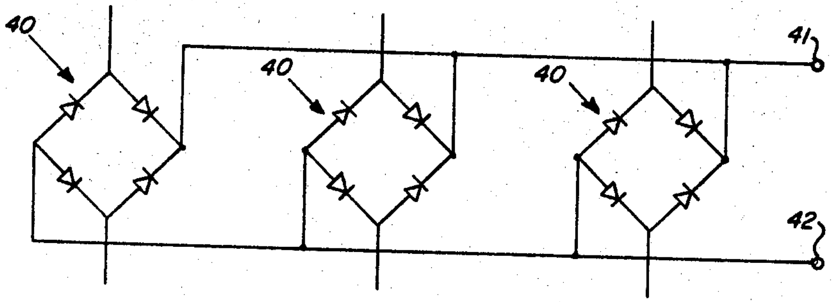
40
40
40
41
42

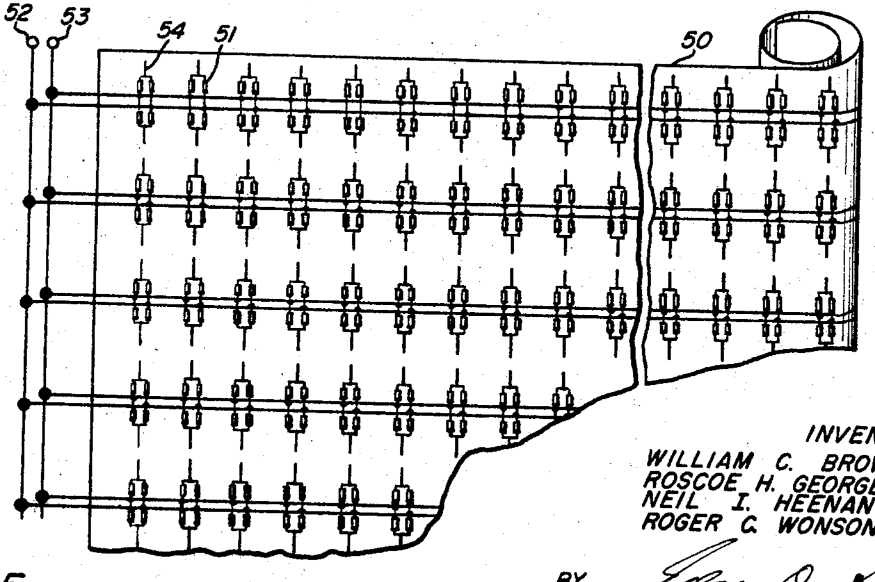
52
53
54
51
50

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

FIG. 5

BY Edgar O. Rost
ATTORNEY

March 25, 1969 W. C. BROWN ET AL 3,434,678

MICROWAVE TO DC CONVERTER

Filed May 5, 1965

Sheet 2 of 3

FIG. 7

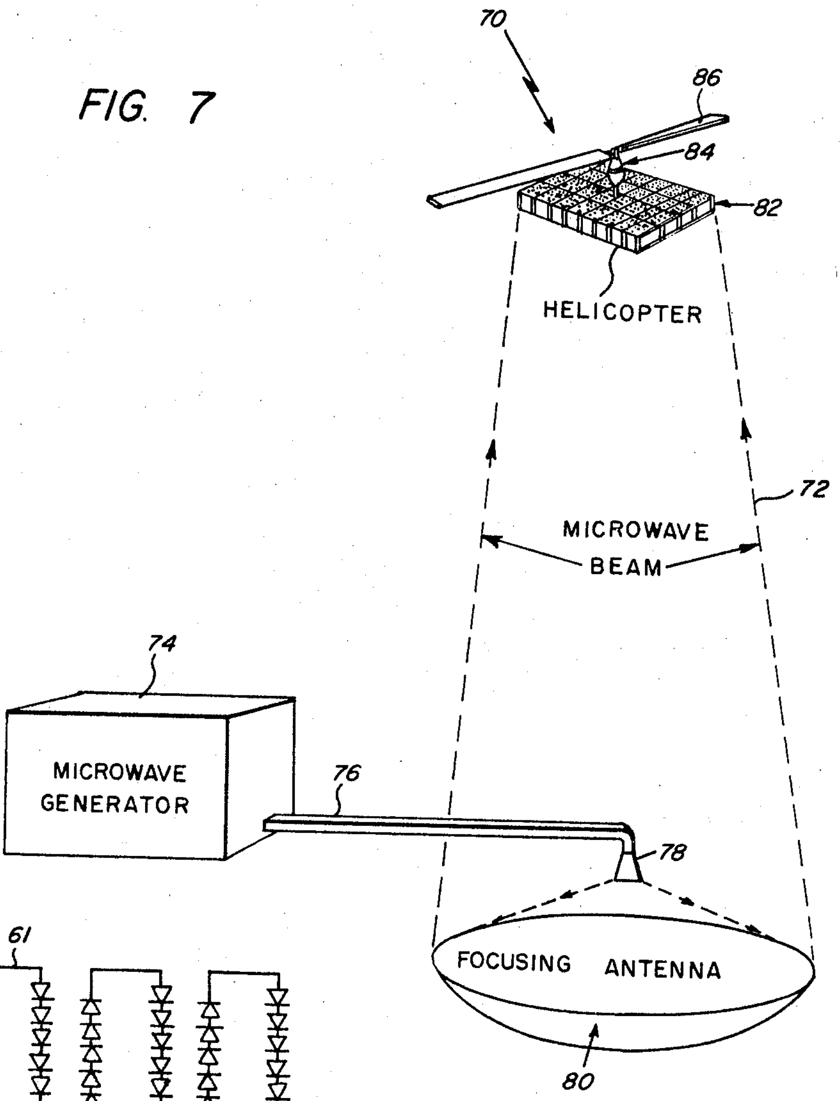
70
86
84
82
HELICOPTER
72
MICROWAVE
BEAM
74
MICROWAVE
GENERATOR
76
78
FOCUSING ANTENNA
80

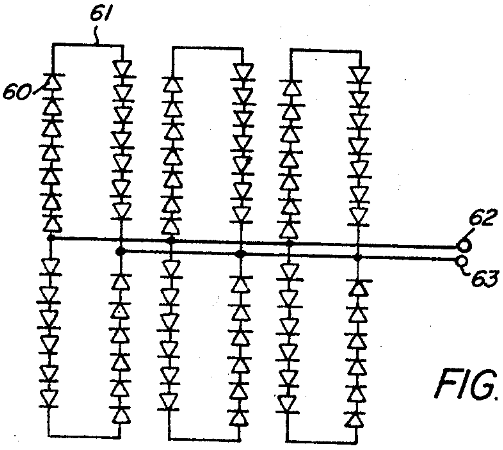

FIG. 6

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

BY Edgar O. Kost
ATTORNEY

March 25, 1969 W. C. BROWN ET AL **3,434,678**

MICROWAVE TO DC CONVERTER

Filed May 5, 1965

Sheet 3 of 3

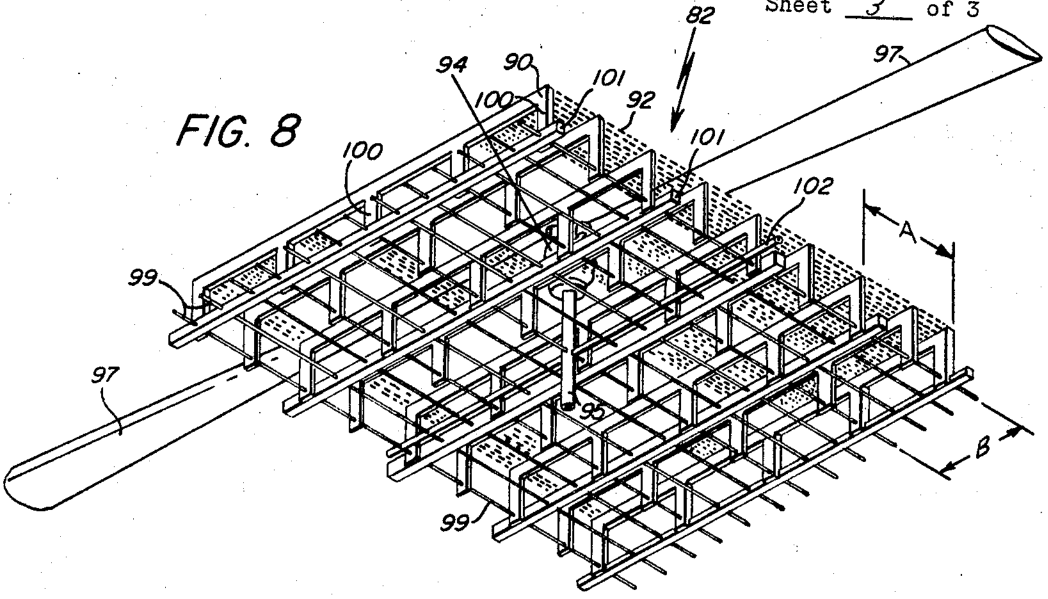
82
94
90
101
92
100
100
101
97
102
A
99
97
95
B
99

FIG. 8

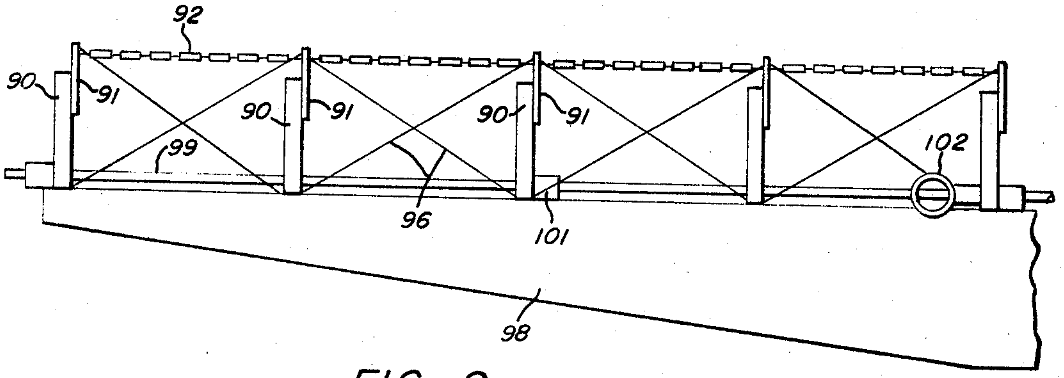
92
90
91
90
91
90
91
99
96
101
102
98

FIG. 9

INVENTORS
WILLIAM C. BROWN
ROSCOE H. GEORGE
NEIL I. HEENAN
ROGER C. WONSON

BY Edgar O. Rost
ATTORNEY

3,434,678
Patented Mar. 25, 1969

3,434,678

MICROWAVE TO DC CONVERTER

William C. Brown, Weston, Mass., Roscoe H. George, West Lafayette, Ind., and Neil I. Heenan, Needham, and Roger C. Wonson, Beverly, Mass., assignors to Raytheon Company, Lexington, Mass., a corporation of Delaware

Filed May 5, 1965, Ser. No. 453,415
Int. Cl. B64g 1/00
U.S. Cl. 244—1 3 Claims

ABSTRACT OF THE DISCLOSURE

A combined antenna and conversion mechanism for reception of beamed high frequency electromagnetic energy in space including a large array of unidirectional current semiconductor rectifier devices. A self-supporting space vehicle utilizing the rectified DC electrical energy for propulsion is disclosed in an illustrative embodiment.

The present invention relates in general to the transfer of energy by means of an electromagnetic wave beam and more particularly to interception and rectification of such energy into low frequency electrical DC energy with a high degree of efficiency.

Improved technology in the field of microwave energy generation at superpower levels has resulted in the realization of electrical energy transmission over considerable distances for remote energization of devices or vehicles without the aid of wires. The transmission of microwave electromagnetic energy into space has been commonly employed in the radar pulse echo systems for the detection and orientation of desired objects within a predetermined scanning range of a transmitting antenna. Beams of a similar nature may now be employed for other useful purposes and the advantages attendant the utilization of electromagnetic energy in the microwave region in contrast with other wavelengths may now be enumerated.

Microwaves have been generally defined as high frequency radio waves whose wavelength is less than 30 centimeters, with a lower wavelength limit on the order of 1 millimeter sometimes being applied to what is commonly referred to as the "microwave region." The superiority of high frequency microwaves is due in part to the fact that it is generally desirable to focus the transmitted energy so as to achieve a high power density at a remote point or area with respect to a given power source. In accordance with the laws of optics, the sharpness of the microwave beam produced by a transmitting antenna varies as the ratio of antenna dimensions to the wavelength of the transmitted energy. Therefore, for a given or desired power density or beam sharpness, a decrease in the wavelength of the transmitted energy permits a corresponding decrease in the dimensions of the antenna. From the standpoint of mechanical considerations, it is desirable to employ small antennas and other components, and it is therefore advantageous to employ high frequency energy of very short wavelength. In addition, the difficulties encountered in long wave transmission as a result of natural and man-made interference or noise do not occur with any appreciable significance at microwave frequencies. Further, in aerospace applications with considerable distances separating the transmitter at an earth or mother planet location and the employment of shorter wavelength beamed energy is preferred since longer wave signals will generallly be reflected at certain altitudes by reflecting layers in the atmosphere.

In view of certain losses due to absorption which may occur in the atmosphere, microwaves in the region having the approximate bounds of 2 and 30 centimeters are readily adaptable to the convenient radiation of power to remote points without the utilization of wires. The preferred wavelengths are of the order of 5 or 10 centimeters to provide efficient focusing with existing transmitting antenna systems which may be maintained at a reasonable size. An illustrative device of the superpower high frequency microwave generators operative in the desired band is the so-called Amplitron which is an amplifier having a broad bandwidth and excellent performance characteristics for the focusing of the beam. Such devices are capable of producing 15 or 20 kilowatts of average continuous wave power in the neighborhood of 10 centimeters in wavelength with capabilities expected in the region of 500 kilowatts or more average power with 50 megawatts peak power. A complete description of such devices may be had by referring to Patent No. 2,933,723 issued Apr. 19, 1960 to William C. Brown and assigned to the assignee of the present invention.

With microwave energy capable of being generated and directed over longer distances conversion of such high frequency electromagnetic energy is of paramount concern. One conversion mechanism in the prior art involves direct conversion of such energy into heat which may then be utilized directly or indirectly for propulsion or generation of flight-producing forces. Examples of such devices for heat energy exchange as well as space vehicles utilizing such energy may be noted in Patent No. 3,174,705, issued Mar. 23, 1965, to D. Schiff et al., as well as U.S. Letters Patent No. 3,083,528, issued Apr. 2, 1963 and No. 3,114,517, issued Dec. 17, 1963, to William C. Brown. The heat exchanger method of conversion of electromagnetic energy into useful power is limited by the overall efficiencies of approximately 25 percent in the conversion of heat into mechanical or electrical work. Desirable, therefore, would be the direct rectification of the high frequency electromagnetic energy into low frequency electrical energy for the operation of many useful aerospace devices as well as systems.

The present invention has for its primary object the conversion of high frequency electromagnetic energy in the microwave region directly into low frequency electrical energy.

A further object of the present invention is the provision of a combined nondirectional receiving antenna and microwave electromagnetic energy to low frequency electrical energy conversion means in a unitary structure.

A still further object of the present invention is a provision of a new and novel combined nondirectional receiving antenna and microwave to DC energy converter for aerospace applications.

Another object of the present invention is the provision of a new and novel nondirectional receiving antenna and microwave to DC energy converter having a high degree of efficiency.

Still another object of the present invention is the provision of a new and novel aerospace vehicle with nondirectional receiving antenna and microwave to DC energy converter means with said vehicle being capable of being supported by its own energy generation means at a distance spaced apart from the power generation means.

In accordance with the teachings of the present invention, the above and other objects are achieved by the employment of efficient unidirectional microwave power rectifiers and dipole antenna means. Such rectifying devices, while being individually limited in power-handling capabilities, normally in the order of fractions of watts, have been found to be highly efficient means for the rectification of microwave power when assembled in large numbers in various arrays. It is interesting to note that the observed collective efficiency was on the order of 40 to 70 percent. In an illustrative embodiment, point-contact semiconductor diodes were arranged in four arm bridge connected networks with the networks interconnected in various configurations such as series, parallel and series-parallel.

In discussing aerospace applications, an additional problem is encountered in the beaming of microwave energy to a remote point and the interception and utilization of such electrical energy. In such applications the advantages of a vehicle which may be maintained in space for indeterminate periods of time without employing a local fuel source are readily apparent. Such devices could readily provide communication networks, surveillance functions using radar techniques along with numerous other functions. The acapture of the beamed high frequency electromagnetic energy raises the need for an efficient antenna means capable of intersecting the beam at high altitudes. Conventional techniques employed in microwave radar usage such as receiving antenna horns are capable of intersecting only a small portion of the beam energy and add considerable weight in applications involving heavier-than-air vehicles. In an exemplary embodiment of the invention a space vehicle, namely a helicopter, is disclosed for either moving flight or a stationary location with self-supporting electrically operative propulsion means. The semi-conductor diode rectifier arrays have been demonstrated to fulfill the receiving antenna functions as well as the electrical energy rectification means in a highly efficient manner. Such combined antenna and rectifier means has also assisted in reduction of the weight problem in airborne devices. Further, it has provided a nondirectional means for the interception of the microwave energy to thereby reduce the problems of focusing inherent in prior art directional horn type receiving antennas.

With the above features, advantages and objects in mind the invention will now be described by reference to the following detailed description together with the accompanying drawings in which:

FIG. 1 is a perspective view of an illustrative diode rectifier;

FIG. 2 is a schematic circuit diagram of a bridge connected diode network with dipole antenna means;

FIG. 3 is a schematic circuit diagram of a plurality of bridge connected networks arranged in series;

FIG. 4 is a schematic circuit diagram of a parallel bridge connected network array;

FIG. 5 is a perspective view of an illustrative embodiment of a combined antenna and rectifier array in a folded or rolled up configuration;

FIG. 6 is a schematic circuit diagram illustrating the bridge connected diode array incorporated in the aerospace vehicle shown in FIG. 7;

FIG. 7 is a schematic representation in elevation illustrative of a heavier-than-air aerospace vehicle incorparating the structure of the present invention;

FIG. 8 is a perspective view of the aerospace vehicle embodiment as viewed from the under portion thereof; and

FIG. 9 is an enlarged partial view in elevation of a portion of the illustrative embodiment shown in FIG. 8.

FIG. 1 illustrates a point-contact semiconductor diode rectifier of the type employed in radar microwave receiver apparatus to rectify returned radar pulses. Any of the high burnout semiconductor diodes having high rectification characteristics are preferred and are commercially available, such as the 1N82 or 1N830. The rectifying junction is formed by whisker element **2** contacting the semiconductor element **4** respectively connected to leads **6** and **8**. Silicon is preferred over germanium for element **4** because of its ability to operate at higher temperatures and thereby handle higher powers. Envelope **10** houses the rectifying elements and may be of a hermetically sealed dielectric material or combination metal and ceramic composition. The inherent characteristic of such diode rectifiers is that the microwave energy is intercepted and rectified in a unidirectional manner and the line **11** indicate pictorially the rays of the beamed electromagnetic microwave energy in a plane normal to the envelope. In FIG. 2 a full-wave bridge connected diode network is illustrated with the forward direction of the rectified DC electric current indicated by the direction of the arrow symbols. The network shown consists of half-wave dipoles **20** and **22** each terminated with a diode rectifier element **24** to **27** in an arm of the bridge connected network. The dipole elements **20** and **22** are of the half-wave configuration and may be spaced apart from each other a one-half wavelength at the frequency of the beamed electromagnetic energy.

Referring now to FIG. 3, an array of bridge connected diode networks each with the half-wave dipoles are shown connected in series. Each network is referred to by the numeral **30** and is similar in the bridge connections to the single element network shown in FIG. 2. The DC output of the collective rectified energy is coupled by means of terminals **32** and **33**. In FIG. 4, a similar number of individual bridge connected diode-dipole networks are shown connected in a parallel array. Each network is indicated by the numeral **40**, and the output terminals are indicated as **41** and **42**.

Any number of diode-dipole networks may be provided and in FIG. 5 such a multi-element array is illustrated by mounting on a flexible material **50** which may be rolled or folded into any desired package or enclosed within a capsule to be launched and released at a predetermined point in space. Any flexible material which is pervious to electromagnetic energy is preferred. The total power desired would be the determining factor in a number of individual diode-dipole elements required. In this embodiment, the bridge connected networks **51** are connected in parallel to the output load indicated by terminals **52** and **53**, and representative measurements of electrical characteristics have shown that approximately five watts of DC electrical energy is realizable for each square foot of area of the combined antenna-rectifier. While the dipole elements **54** have been indicated in a particular array, it is within the scope of the invention to stagger the placement of such dipoles to increase the overall efficiency of the antenna-rectifier.

To further increase the DC powder output, the full-wave bridge connected networks are preferably arranged with a plurality of diodes in series in each arm of the bridge. An illustrative schematic circuit diagram of such a configuration is shown in FIG. 6 wherein seven diodes **60** are shown in each arm of the bridge circuit and are connected in series for a total of twenty-eight diodes in each bridge network. The dipole members will then be the substantially U-shaped end portions **61** at the ends of each brace of seven diodes. In the illustration three such twenty-eight diode bridge networks are shown connected in parallel to terminals **62** and **63**. This closer spacing and compact arrangement has been shown to be a source of improved power output and is capable of a high degree of reliability through the redundant nature of the parallel series connections within each bridge network. If one of the diode rectifiers fails to function the over-all voltage drop across this element would be divided among the six remaining diode rectifiers. If any of the connecting wires between the diode elements should break, the adjacent arms of the other bridge assemblies would take the additional load due to the close proximity of the respective arms to each other. In addition, it is possible to have a number of open connections or inoperative diodes dispersed throughout the array without any serious impairment in performance.

In relation to the array concept to be hereinafter described it may be stated that within a six inch square area ten such individual bridge networks each containing twenty-eight diode rectifiers for a total of 280 diode rectifiers may be deployed in such a manner as to provide maximum exposed area for each diode as well as the connecting leads. Such an arrangement will be hereinafter referred to as a "module" and a DC output in excess of fourteen watts has been measured for such a module. Any number of such modules could be connected provided for a desired power yield and this module concept readily lends itself to use in certain aerospace applications now to be described.

In FIG. 7 a propelled type of space vehicle **70** is shown wholly supported by means of the transfer and rectification of continuous wave electromagnetic energy via a microwave beam **72**. The source of the microwave energy which may be of the Amplitron type device as described in the aforementioned issued Patent No. 2,933,723 is indicated as **74**. This energy is fed by waveguide means **76** to a transmitting horn **78** to illuminate an ellipsoidal beam forming focusing antenna **80** for the transmission of the microwave beam **72**. It will be appreciated by those skilled in the art that the representations of the microwave generation and transmitting antenna means are pictorial representations to illustrate the usage of the invention in diagrammatic form and the present invention is not limited to any particular source of microwave energy or transmitting antenna assembly. It may be stated the reflector of the antenna assembly is considerably larger than most of the reflectors of the prior art in order to focus a large amount of the microwave power at high altitudes for use in the transfer of energy to space vehicles. Such antenna assemblies may be partially supported in a large hollowed area on the earth's surface or other convenient means of support.

The space vehicle or helicopter **70** can be described as a main body member supporting antenna-rectifier means **82** including a large number of the so-called modules connected together and rigidly supported in a planar parallel array. A motor **84** is supported by the combined body member and the receiving antenna-rectifier means and actuates the rotor **86** of conventional design employed in such self-propelled hovering vehicles. The disclosed vehicle provides for the illumination of the planar array of the semiconductor diode dipole elements by the microwave beam and the direct conversion of the microwave power transmitted by the beam into usable electrical energy for the self-propulsion of the device without any local fuel supply being required.

FIGS. 8 and 9 illustrate a space vehicle **82** comprising a plurality of the combined receiving antenna-rectifier module means for interception and rectification of the electromagnetic microwave energy beam emanating from an earth or mother planet source. A planar array of the antenna-rectifier modules is mechanically supported by means of structural members **90** of any lightweight wood or metal. Insulators **91** positioned coextensive with the members **90** support the diode rectifier array and avoid interference with the receiving and electrical performance characteristics by the structural support members. Carrying forward the module concept of 280 diode rectifiers to provide an approximate power output of 14 watts, it was noted that any number of such modules may be coupled together since the individual module outputs are relatively insensitive to a wide range of load resistances connected to the common output terminals. To achieve the desired electrical output of approximately 120 volts and 250 watts of power, subgroups of four modules each were assembled and parallel connected with an approximate 30 volts available for each subgroup. Four such subgroups were series-connected to result in a total of 4,480 diode rectifiers or 16 modules assembled in a two foot square self-supporting planar parallel array structure. The individual diode rectifiers connected in each arm of the bridge network are indicated by the numeral **92**. An exemplary module configuration would extend within the area delineated by the dotted lines and reference letter A on one side and similar dotted lines and reference letter B on the other side.

A motor **94** is connected to the DC side of the overall array and may be additionally supported by tubular member **95**. A shaft and propulsion means consisting of rotor blades **97** provide for the upward lift of the overall vehicle for the self-supporting of same in space applications. Additional structural support such as interlaced rigging **96** of a high tensile strength material such as nylon or steel wire, as well as bracing member **98**, may be employed for strengthening of the body means to withstand the vibrational forces and downwash from the propulsion means.

In accordance with the well known technology of microwave transmission the combined array of diode rectifiers and propulsion means presents a specific load impedance which must be suitably matched to the transmitted microwave energy beam to result in maximum efficiency. In aerospace applications a mismatch of approximately ten to one may be evident. Matching of the load impedance to a value of approximately 377 ohms as the free space value will be provided by a plurality of coplanar parallel metallic rod members **99** disposed in a grating array in front of the diode rectifiers a predetermined distance. Rod members **99** are linearly disposed and extend in a similar direction as the assembled diode rectifiers. A selected frontal spacing of one-quarter of the wavelength of the microwave frequency being transmitted has been experimentally determined to be suitable for impedance matching purposes. An approximate spacing of two inches between the respective members was preferred for a selected microwave frequency of 2,450 megacycles. Each of the members **90** are provided with lateral sections **100** to support the elongated bar members **101** which in turn maintain the rod members **99** in the desired position. A tubular member **102** of a lightweight metal may also be provided to combine with the motor support member **95** for structural support.

The combined antenna-rectifier array provides a source of electrical energy to render any space vehicle self-supporting. The diode rectifier elements when assembled in the antenna array have been found to be nondirectional with respect to interception of the beamed microwave energy. This represents a large step forward in the utilization of high power microwave energy over the prior art horn-type receiving antennas which must be accurately focused and pointed in a particular direction for the reception of any energy. The connections between the respective members of the diode rectifier array and deployment in the parallel configuration serves to provide maximum exposed area. Such connections and in particular the end loop portions adjacent the terminus of each arm of the bridge networks serves as an efficient dipole for the interception of the microwave energy.

Although it is not intended as a full explanation of the high degree of efficiency attained with the disclosed antenna-rectifier array, it is believed that the whisker elements within the semiconductor diodes themselves are a contributing factor and may function as additional dipole elements. The disclosed embodiment functioned efficiently when illuminated by microwave energy generating a vertically polarized beam. Hence, an efficient and light weight energy conversion apparatus is disclosed which may be self-supporting without the requirement of a large local fuel supply payload.

It may be within the purview of the invention to use the available rectified electrical energy for performing many functions in addition to the actuation of the propulsion means. Hence, communications' payloads may be maintained at predetermined positions in space in a hovering attitude utilizing a portion of the electrical energy available. Relay signals to other such vehicles or return signals to ground stations would then be within the realm of possibility. Such available energy may also be employed for servomechanisms, stabilizing and countertorque systems for the navigation of such vehicles.

The electrical efficiencies realized with the combined receiving antenna and rectifier means have also provided certain weight advantages over other energy converters in aerospace applications. Examples of such converters would be heat exchangers or solar cells. In comparison to the present invention where five to eight pounds per kilowatt of energy realized is a normal characteristic, other energy conversion means weigh in the vicinity of 150 pounds per kilowatt of realizable energy. The inherent advantages of the present invention are therefore apparent. While the technology in the diode rectifier art is being continually advanced, new diode power rectifiers as well as integrated circuit techniques are readily available to future configurations of the present invention. The so-called Schottky barrier diodes could be employed to produce combined antenna-rectifier means weighing even less than two pounds per kilowatt of available energy.

Although the foregoing detailed description has referred to DC power rectification it will be evident that with suitable circuit components low frequency AC energy may also be made available. In addition, other propulsion means may be readily substituted using electrical energy. The embodiments disclosed herein are illustrative only and other modifications or alterations will be apparent to those skilled in the art which do not depart from the scope of the broadest aspects of the present invention as defined in the appended claims.

What is claimed is:

1. A space vehicle comprising:

body means;

said body means including spaced structural support members;

combined antenna and DC electrical energy rectification means for the interception and rectification of incident high frequency electromagnetic microwave energy carried by said support members in a planar parallel array;

said rectification means comprising a plurality of four arm full-wave bridge connected rectifier circuit networks each having a plurality of unidirectional semiconductors in each arm;

said networks being electrically interconnected to common output terminals;

electrically operable propulsion means comprising a motor and rotor members carried by said body means and connected to said terminals for the utilization of said rectified DC energy; and

means for matching the load impedance of said combined antenna and electrical energy rectification means to the incident microwave energy.

2. A space vehicle according to claim 1 wherein said load impedance matching means are arranged in a coplanar array coextensive with said antenna and energy rectification means array, and spaced therefrom a distance of approximately one-quarter of a wavelength at the frequency of the microwave energy.

3. A space vehicle according to claim 2 wherein said load impedance matching means comprise a plurality of parallel disposed elongated metallic members.

References Cited

UNITED STATES PATENTS

1,217,149	2/1917	Caldwell	321—27 X
2,165,055	7/1939	Kafka	321—27
2,444,458	7/1948	Master	321—8
2,927,321	3/1960	Harris	343—6.8 X
3,098,971	7/1963	Richardson	325—592 X
3,174,705	3/1965	Schiff et al.	244—1

RODNEY D. BENNETT, *Primary Examiner.*

MALCOMB F. HUBLER, *Assistant Examiner.*

U.S. Cl. X.R.

307—151; 318—16; 321—27; 325—494; 343—100

Teleportation of Entangled Coherent States through Visual Pathways in the Human Brain

[1,2,3]**V. Salari**[*] , [1,2]**M. Rahnama**[†] and [4]**J. A. Tuszynski**[‡]

[1]*Department of Physics, Shahid Bahonar University of Kerman, P. O. Box 76175, Kerman, Iran*
[2]*Afzal Research Institute, Kerman, Iran*
[3]*Neuroscience Research Center, Kerman, Iran*
[4]*Department of Physics, University of Alberta, Edmonton, T6G 2J1, Canada*

Abstract:

In this paper we have proposed a new model for visual consciousness based on the premise there exists a quantum teleportation mechanism between the eye and the brain. In our model, some open questions will be answered.

1) Introduction:

Schrödinger's book "What is life?" has had an enormous influence on the development of molecular biology [1]. His insight has since then inspired many researchers to investigate the molecular basis of a living organism [2,3,4]. Several researchers have noticed the sweeping consequences that would follow from the discovery that living organisms might process information quantum mechanically, either at the bio-molecular level, or the cellular/neuronal level [5,6,7,8,9]. Computational models of the behavior of networks of neurons are still mainly based on the integrate-and-fire model of neural function. Mainstream cognitive neuroscience typically ignores the role of quantum physical effects in the neural processes underlying cognition and consciousness. However, many unsolved problems remain, suggesting the need to consider new approaches. Recent contributions to the investigations of quantum effects in the human brain are due to Jibu and Yasue, Pribram, Lockwood, Mavromatos and Nanopoulous, Hameroff and Penrose, and Stapp [10]. These authors have proposed models in which the operation of consciousness is associated with some sort of explicit wave function collapse. There have been numerous suggestions that consciousness is a macroscopic quantum effect possibly involving superconductivity, superfluidity, electromagnetic fields, Bose-Einstein condensation or some other physical mechanism. Perhaps the most specific model developed thus far is that of Penrose and Hameroff and it asserts that quantum information processing takes place in microtubules (MTs). It has been argued that MTs can process information similarly to a cellular automaton, and hence Hameroff and Penrose suggest that neuronal MTs may operate as a quantum computer.

In this paper, we investigate visual information transfer from a quantum information point of view. Theoretically we consider the conscious observer looking at a quantum system and propose that the state of this system is reported via superposed photons. We address the question whether the observer can receive the exact same state of this system quantum mechanically in his/her brain or this quantum state collapses before reaching the brain. Below we investigate the problem in detail.

2) Evolution of Information from the Eye to the Brain

We assume that a conscious observer directs his/her attention to a quantum system. For simplicity we consider this system to be a manifestation of the celebrated *Schrödinger's cat*. This system can exist in two states: Live$|L\rangle$, or Dead$|D\rangle$.

---

[*] *vahidsalari@graduate.uk.ac.ir*

[†] *Majid.rahnama@mail.uk.ac.ir*

[‡] *jtus@phys.ualberta.ca*

$$\psi_{sys} = \frac{1}{\sqrt{2}}(|L\rangle + |D\rangle) \tag{2-1}$$

The state of this system is then reported via superposed photons. 4% of these photons are reflected from the cornea. 50% of remaining photons are dissipated through ocular media absorption. Other photons enter the 200-250 μm thick retina. There, photons interact with the photoreceptors in the rods and cons layer after 80% loss due to retinal transmission [11],[12]. In this case, we consider just a few remaining photons which are in a superposed quantum state. The key question here is whether this quantum state of photons can be reported to the brain.

When this state interacts with the last layer of retina, according to Thaheld, this superposed photon undergoes a wavefunction collapse. On the other hand, photons can be absorbed and then transformed into classical signals. Here, we use the symbols introduced by Tegmark [13] for the observer. The symbol $|\ddot{\_}\rangle$ denotes the state for which the information on photons is not received by the brain and thus the observer is amphoteric. The symbol $|\ddot{\smile}\rangle$ stands for the state in which the information received in the brain reports that the cat is alive (and the observer is happy). Finally, the symbol $|\ddot{\frown}\rangle$ corresponds to the state in which the information received in the brain indicates that the cat is dead (and the observer will be sad). It means that:

$$U|\ddot{\_}D\rangle = |\ddot{\frown}D\rangle \tag{2-2-a}$$

$$U|\ddot{\_}L\rangle = |\ddot{\smile}L\rangle \tag{2-2-b}$$

Where $U = \exp\left[-\frac{i}{\hbar}\int H_{photon-brain}\, dt\right]$.

Now, we consider another state in which the brain interacts with itself. Penrose and Hameroff have proposed a model of consciousness involving quantum computation with objective reduction in *microtubules* (or MTs) within the brain's neurons (see Figure 1) [14],[15],[16],[17]. MTs are cylindrical polymers comprised of the protein tubulin which organize numerous cellular activities including neuronal motor transport. According to Hameroff and Penrose, switching of tubulin conformational states is governed by quantum mechanical forces within each tubulin interior, and an essential feature of the Orch OR model is that tubulins may exist in quantum superpositions of two or more conformations. Therefore, these states could function as quantum bits, or "qubits" by interacting non-locally (through their entanglement) with other tubulin qubits so that MTs may act as quantum computers. When sufficiently many entangled tubulins are superposed for long enough to reach Penrose's OR threshold given by E=h/T, an objective reduction (OR) "conscious event" occurs as stated in the Orch-OR model.

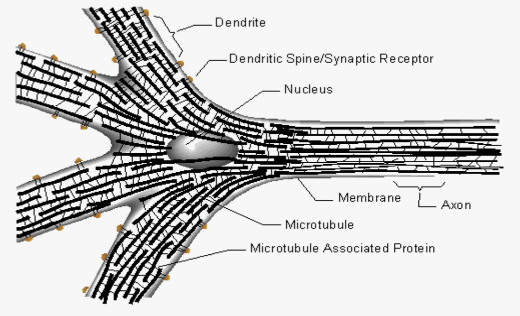
Dendrite
Dendritic Spine/Synaptic Receptor
Nucleus
Membrane
Axon
Microtubule
Microtubule Associated Protein

Figure 1 Representation of MTs inter a brain neuron. The Orch OR model suggests that the main information processing is implemented in these structures.

If previous evolution is described by Penrose's self-collapse in the brain (i.e. Orch-OR), MTs in the neurons of the brain collectively evolve and then collapse (i.e. conscious event) to one of the happy or sad states. It mathematically means that:

$$U|\ddot{\_}\rangle = \frac{1}{\sqrt{2}}(|\ddot{\smile}\rangle + |\ddot{\frown}\rangle \qquad (2\text{-}3)$$

where $U = \exp[-\frac{i}{\hbar}\int H_{brain}\, dt]$.

If we compare this state (2-3) and the state which has the information about the photon (2-2), we can say that there's a great amount of correlation between the retina and the visual cortex, because their result should be identical. We know that in accordance with the Einstein-Podolsky-Rosen (EPR) pair, when two entities originate from a common source they can be entangled with each other. Retina has a similar layered structure as the gray-matter top layers of the cerebral cortex of the brain. In fact, retina is an extension of the central nervous system (the brain and spinal cord) that forms during embryonic development. One reason why scientists are interested in retinal processing is that retina is an accessible part of the brain that can be easily stimulated with light [18]. Thus, we can say that retina and visual cortex are entangled with each other. The other argument for this entanglement is that the image of a system is inverted on the retina but it is perceived in the brain directly. As we know, in two entangled entities when one has a property in some direction the other exhibits that property but in the opposite direction. Thus retina receives an inverted image, while it is received in the visual cortex upright. This property is special to entanglement.

The human body is made up of many organs, which themselves are made up of many millions of cells. How can such a system, with millions, or even billions, of parts function effectively and coherently? This is a no small feat as even small-size human societies often undergo periods of turbulence and trouble due to conflict and poor organization. Now some scientists think that quantum coherence in MTs is a major factor responsible for our bodies, and especially our brains, being so efficiency. Here, we intend to describe visual processing on the basis of quantum information theory.

3) Generation of Entangled Coherent States

Centrioles and cilia, which are microtubular structures, are involved in photoreceptor functions in single cell organisms and primitive visual systems. Cilia are also found in all retinal rod and cone cells. The dimensions of centrioles and cilia are comparable to the wavelengths of visible and infrared light (see Figure 2) [50]. Moreover, cytoskeletons are found mostly among the retina and the visual cortex in the cells of the optic nerve. Cytoskeletal structures of the centrioles can be expected to vibrate like a harmonic oscillator in its ground state. Vibrational dynamics of MT's has been the subject of a recent paper where typical frequency ranges have been discussed [19].

When a photon interacts with a centriole, its electric field can displace the potential of the harmonic oscillator and then releases it, thus generating coherent states [20]. It means that:

$$D(\alpha)|0\rangle = |\alpha\rangle \qquad (3\text{-}1)$$

Where $D(\alpha)$ is the displacement operator; $D(\alpha) = e^{\alpha a^{\dagger} - \alpha^{*} a}$, and $|0\rangle$ is the ground state of harmonic oscillator, and $|\alpha\rangle$ is the coherent state: $|\alpha\rangle = e^{-\frac{|\alpha|^2}{2}} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle$.

Hameroff [21] and Penrose [22] have suggested that MTs inside cells permit long-range quantum coherence, enabling quantum information processing to take place at the sub-cellular level. They use this hypothesis to develop their theory of consciousness. Cells interconnected by gap junctions form networks which fire synchronously, behaving like one giant neuron [23], and possibly accounting for synchronized neural activity such as coherent 40 Hz waves [24]. Marshall [25] suggested that coherent quantum states known as Bose-Einstein condensation occurred among neural proteins [26], [27], [28]. Other issues, such as preconscious-to-conscious transitions were identified by Stapp [29] with the collapse of a quantum wave function in presynaptic axon terminals [7]. The other reason for coherence of these structures is that light is an electromagnetic wave and thus is vibrational, and according to Froehlich's theory [44],[45],[46] it can take these cytoskeletal structures (i.e. nonlinear structures composed of electric dipoles) into one mode of frequency and coherence.

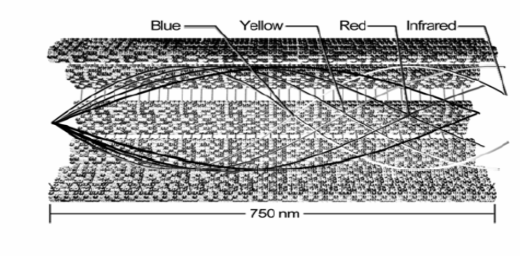

Figure 2 Representation of one centriole, the dimensions of centrioles are comparable to wavelengths of visible and infrared light.

Centrioles are two cylinders composed of MTs which are perpendicular to each other and in accordance with the Hameroff theory of the origin of cancer [30], centrioles are entangled with each other. Because of this entanglement, when a coherent state $|\alpha\rangle$ is generated in one centriole, in the other it will generate state $|-\alpha\rangle$. Now, we can say that after interaction of photons with centrioles, they generate "entangled coherent states" in these structures in the retina, i.e.:

$$|\varphi\rangle_{12} = (A|\alpha\rangle_1|\alpha\rangle_2 - B|-\alpha\rangle_1|-\alpha\rangle_2) \tag{3-2}$$

where $|\varphi\rangle_{12}$ is an entangled coherent state in centrioles with two modes 1 and 2. QED-cavity model of MTs [31] describes that coherent modes of electromagnetic radiation can be sustained in the interior of the MTs. These modes are provided by the interaction of the electric dipole moments of the ordered-water molecules in the interior of MTs with quantized electromagnetic radiation [32],[33]. Jibu, et. al.[49], have proposed that the quantum dynamical system of water molecules and the quantized electromagnetic field confined inside the hollow MT core can manifest a specific collective dynamical effect called superradiance [34] by which the MT can transform any incoherent, thermal and disordered molecular, atomic or electromagnetic energy into coherent photons inside the MT. Furthermore, they have also shown [49] that such coherent photons created by superradiance penetrate perfectly along the internal hollow core of the MT as if the optical medium inside it were made "transparent" by the propagating photons themselves. This is referred to as the quantum phenomenon of self-induced transparency [35]. Superradiance and self-induced transparency in cytoskeletal MTs can lead to "optical" neural holography [36]. Neurons (and maybe also other cells) may contain microscopic coherent optical supercomputers with enormous capacity. Thus Jibu, et. al.[49], suggest that MTs can behave as optical waveguides which result in coherent photons. They estimate that this quantum coherence is capable of superposition of states among MT spatially distributed over hundreds of microns. These in turn are in superposition with other MTs hundreds of microns away in other directions and so on. With the above conclusions $|\varphi\rangle_{12}$ can produce those photons which produced itself, thus if the state $|\varphi\rangle_{12}$ can be restored in the brain, it will reproduce the photons which were absorbed in the retina.

Additional arguments in favor of the feasibility of production of photons in the brain can be found in the conclusions of the paper [37], which also asserts that there exists a neural activity-dependent ultra-weak photon (biophoton) emission in the brain. Thus there is the possibility to restore the initial state of the photon in the brain after absorption in the eye. This process can be implemented through teleportation mechanism between retina and the visual cortex as will be discussed in the following sections.

4) The Decoherence Problem

The important question about the quantum processing in the living systems is: how is it possible for MTs to process information quantum mechanically while the environment surrounding them is relatively hot, wet and noisy?

According to the Orch-OR model, microtubular structures in the neurons of the brain process information quantum mechanically and to avoid decoherence, like lasers maintain quantum coherence against thermal noise. Water within cells is itself not truly liquid, but has been shown to be, to a large extent, ordered [38]. Most of the ordered water in the cell in fact surrounds the cytoskeleton [39]. MTs and other cytoskeletal components are embedded in cytoplasm which exists in alternating phases of (1) “sol” (solution, liquid); and (2) “gel” (gelatinous, solid). Among the most primitive of biological activities, “sol-gel transformations” within neurons and other living cells are caused by assembly and disassembly of cytoskeletal actin (e.g. regulated by calcium ions through the protein calmodulin, in turn regulated by MTs). Sol-gel transformations are essential in basic cellular activities such as (“amoeboid”) movement, growth and synaptic formation and neurotransmitter vesicle release [40], [41]. Transitions can occur rapidly (e.g. 40 sol-gel cycles per second), and some actin gels can be quite solid, and withstand deformation without transmitted response [42]. Cyclical encasement of MTs by actin gels may thus be an ideal quantum isolation mechanism. In the gel phase of cytoplasm, the water ordering surfaces of a MT are within a few nanometers of actin surfaces which also order water. Thus bundles of MTs encased in actin gel may be effectively isolated extending over the radius of the bundle, on the order of hundreds of nanometers. There are many mechanisms which can protect these structures against decohering factors. In general, Quantum states of tubulin/MTs are protected from environmental decoherence by biological mechanisms which include phases of actin gelatin, plasma-like Debye layering, coherent pumping and topological quantum error correction [54]. Moreover, quantum spin transfer between quantum dots connected by benzene rings (the same structures found in aromatic hydrophobic amino acids) is more efficient at warm temperature than absolute zero[43]. It is conjectured that the “flexibility” of the resonant benzene electrons is advantageous to quantum processes by harnessing ambient thermal energy. MTs may possibly utilize nonspecific thermal energy for "laser-like" coherent pumping, for example in the GHz range by a mechanism of "pumped phonons" suggested by Froehlich [44,45,46].

5) What is the mechanism of Teleportation?

According to the definition of teleportation as stated in [47], in the process of quantum teleportation, one can construct an exact replica of the original unknown quantum state at a cost of destroying the original state. Therefore, to call a quantum state transfer operation- quantum teleportation, the process should not only generate output states with better qualities than what can be done classically but also obey the no-cloning theorem [48].The quantum state of a system can be transmitted from a location to a distant one using only classical information provided that a quantum channel exists between the sender and the receiver. Sharing entangled states between the two parties opens the necessary quantum channel. Research in quantum state transfer, especially the quantum teleportation, has emerged as one of the major research areas of theoretical and experimental quantum mechanics. Assume that Alice wants to send Bob an unknown quantum state but when she receives this state she does not know anything about that unless she affects it and collapses it to classical state, or in other hand she destroys that quantum state. She just can send classical signals to Bob through classical channel, but if there is a shared entangled channel between Alice and Bob, Bob can construct initial quantum state with the help of classical signal which is sent by Alice and quantum channel between them. This operation is implemented by operating special unitary operators. For more details see [47]. Here, we want to simulate visual information transfer with teleportation mechanism. We know that when photon penetrates the retina, it changes to action potentials or electrical signals and these classical signals are sent to the brain for interpretation. It means that *retina* (Alice) wants to send the *brain* (Bob) a *photon state* (unknown quantum state), but *retina* (Alice) absorbs it (collapses the quantum state) and changes it to *action potential* (classical state) and send it through *membranes of axons of the brain neurons* (classical channel). *Brain* (Bob) can reconstruct the initial *state of photon* (unknown quantum state) to process it to emerge consciousness. Our arguments for this simulation are below conclusions:

> *Orch OR* [14], [15]: There is quantum processing in the neurons of the brain (there is a quantum channel between retina and brain)
>
> *Tegmark* [13]: Displacement of ions through membranes of brain neurons is a classical phenomenon (action potentials are classical signals and membranes of neurons are classical channels).
>
> *Thaheld* [11],[12]: Superposed photons do collapse in the retina (the quantum state is collapsed by the sender [Alice]).
>
> We assert that they describe different aspects of the teleportation mechanism. We simulate visual consciousness with the teleportation mechanism as shown in Table 5-1.

Table 5- 1 Simulation of the transferring visual information from the eye to the brain to the Teleportation mechanism.

Human Brain	***Quantum Teleportation Mechanism***
Retina	Alice
Membrane of axons in neurons	Classical channel
Cytoskeletal structures	Entangled channel (quantum channel)
Visual cortex	Bob
Action potentials	Classical signals

Now we want to investigate this teleportation mechanism via teleporting of entangled coherent states through visual pathways in more details. We will show how photon states are constructed in the visual cortex.

6) Teleportation of Entangled Coherent States through Visual Pathways

Super radiance and self induced transparency [49] besides Froehlich oscillations can cause the coupling of MT dynamics over wide areas and take them into a superposition and a coherent state. While in superposition, tubulins communicate with entangled tubulins in the same manner, and in MTs in neighboring neurons, and through macroscopic regions of the brain via tunneling through gap junctions and possibly tunneling nanotubes [50]. As we have already said, retina and the visual cortex are entangled with each other. Thus there is a quantum channel between retina and the visual cortex which is composed of microtubular structures. MTs interact with membrane structures mechanically by linking proteins, chemically by ions and second messenger signals, and electrically by voltage fields. The transduction of light into electrical signals takes place in the photoreceptors. Except for the ganglion cells, none of the retinal cells display action potentials [51].

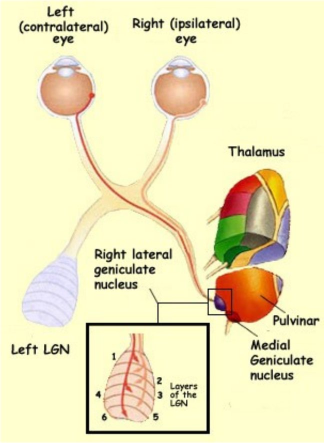

Figure 3 The LGN is a laminated structure, having 6 layers. Contralateral fibers and ipsilateral fibers couple in the LGN. The ipsilateral fibers of the optic nerve terminate in laminae 2,3 and 5 of LGN, while the contralateral fibers terminate in laminae 1, 4 and 6.[52]

Axons leaving the temporal half of the retina traverse the optic nerve to the optic chiasm, where they join the optic tract and project to ipsilateral structures. Axons leaving the nasal half of the retina cross the midline at the chiasm and terminate in contralateral structures. This arrangement means that all the axons in the optic tract carry information about the contralateral visual field. Axons of the optic tract terminate in three areas of the central nervous system, the lateral geniculate nucleus (i.e. LGN), the superior colliculus and the pretectal area. The trajectory through the LGN is the largest most direct and clinically most important pathway by which visual information reaches the cerebral cortex. About 80% of the optic tract axons synapse in the LGN. The LGN is a laminated structure, having 6 layers. Contralateral fibers and ipsilateral fibers couple in the LGN. The ipsilateral fibers of the optic nerve terminate in laminae 2,3 and 5 of LGN, while the contralateral fibers terminate in laminae 1, 4 and 6 of LGN (see Figure 3). There are about 10^6 neurons in each LGN, all of which project to the ipsilateral occipital cortex (area 17) as the optic radiations. The portion of the cerebral cortex that receives LGN axons is called the striate cortex and is usually labeled V1 to designate it as the primary visual cortical area (Figure 3). Virtually all information in the visual system is recognized as being processed by V1 first, and then passed out to higher order systems [53]. The upper visual cortex receives signals from the lower visual field and similarly, lower visual cortex process information from the upper visual field. The right visual cortex processes the left field of view and vice versa (see figure 4).

Now, we investigate the information transfer through visual pathways. The important model among quantum models about the processing in the human brain is the Orch-OR model which is based on the structure of the cytoskeleton. As already said this model asserts that the main processing in the neurons of the brain is performed in the MTs and the nature of the processing is mainly quantum mechanical. The processing unit in this model is *tubulin* which can be in a superposed state. Tubulins act like qubits in quantum computers. *Max Tegmark* believes that there is no quantum processing in the human brain. He has calculated decoherence times for every superposition state which can be possible in the neurons of the brain [13]. In his opinion, superposition states include ions such as Na^+ which are “in” and “out” of the membrane of axon. On the other hand, Na^+ ions are in the superposition of “in” and “out” with a separation distance comparable to the membrane thickness. He has considered three factors which can destroy this superposition state in neurons. *Collisions with the neighboring ions, Collisions with the water molecules* and *Interaction with distant ions* are the factors which Tegmark investigated for decoherence. He has estimated 10^{-19} s and 10^{-20} s for decoherence times. It is clear that above times are very small for the brain processes such as seeing, thinking, speaking and the other cognitive processes. Typically, dynamical timescales for neuron firing and cognitive processes are in the range of 10^{-4} to 1 seconds, whereas decoherence timescales are many orders of magnitude shorter. Thus, action potentials should be regarded as classical signals or the displacement of ions through the membrane of axons should be investigated classically. It is remarkable that Tegmark has also calculated decoherence time for MTs, but he has used wrong assumptions about these structures (for more details see [54]) and we just accept his calculations about action potentials. Thaheld [11], [12] believes that the wave function of any superposed photon state or states is always objectively changed within the complex architecture of the eye, and any incident photons have to run a very daunting gauntlet before they are even converted or transduced to retinal ganglion cell spike trains (To learn more about Thaheld arguments, the reader is referred to refs. [55,56,57]). According to Thaheld, the quantum state of photons does collapse in the retina and it does not reach the brain.

Is Thaheld right? Is not there any mechanism to rebuild the quantum state of photons in the brain? Here we accept that the states of photons collapse in the retina but we believe that they can be restored in the visual cortex via the teleportation mechanism.

Now, the question is how can it be possible to restore the exact state of photons in the brain while its state is collapsed in the retina. The other question which one may ask is: if this state is reported through action potentials how is this information reported to the brain and how can it interpret action potentials to obtain the exact state of the photons? Our solution to the above problems involves the teleportation of entangled coherent states through visual pathways. The state of the photon is teleported from the eye to the brain. On the other hand, the state of the photon is transferred via some "cut-and-paste" mechanism from the eye to the brain. But how is it possible?

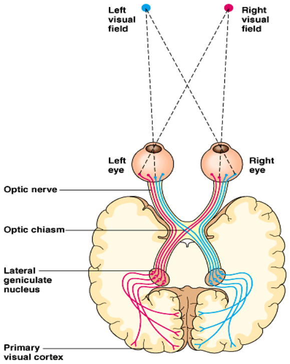

Figure 4 Visual pathways from the eye to the brain. See crossing of pathways.

We concluded before that retina and the visual cortex are entangled. Also we explained how the entangled coherent state is generated in the retina. Now, we want to formulate the process of information transfer from the retina to V1. The state (3-2) with two modes 1 and 2 should be teleported to V1. After the interaction of light with retina, modes 3, 4 and 5, 6 are generated through microtubular structures between retina and V1, and thus they produce entangled coherent channels between retina and V1. It means that the channels are:

$$|\psi\rangle_{35} = \frac{1}{\sqrt{N_\alpha}}(|\alpha\rangle_3|\alpha\rangle_5 - |-\alpha\rangle_3|-\alpha\rangle_5) \qquad (6\text{-}1)$$

$$|\phi\rangle_{46} = \frac{1}{\sqrt{N_\alpha}}(|\alpha\rangle_4|\alpha\rangle_6 - |-\alpha\rangle_4|-\alpha\rangle_6) \qquad (6\text{-}2)$$

Where N_α is the number of tubulins in each channel. Each mode is reported via a special fiber through visual pathways. All of the neurons which are collected in the LGN are divided into two major pathways: ipsilateral fibers and contralateral fibers. Information transfer in the contralateral fibers takes longer than information transfer in ipsilateral fibers because contralateral fibers have crossing relative to ipsilateral fibers and then they have longer lengths than ipsilateral fibers. On the other hand, contralateral fibers have a retarded phase relative to ipsilateral fibers. Now we want to answer the following questions. What is this phase difference? What is the role of this crossing? And how does crossing restore the initial state in the retina?

7) The Role of Phase Shift to Restore Information in LGN

When the information is collapsed in the retina, action potentials are produced. The shape of action potentials is the same for each neuron, but the main problem is which neurons are fired, or in other words which neurons carry action potentials and information. Consider two fibers selected from ipsilateral fibers and two fibers selected from contralateral fibers. The two ipsilateral fibers are called 1 and 2, and the two contralateral fibers are called 3 and 4 while the two fibers from the LGN to V1 are called 5 and 6 which are selected from the group of magnocellular and parvocellular fibers. Now, we start from the retina. The state of centrioles and channels is:

$$
\begin{aligned}
|\psi'\rangle &= |\varphi\rangle_{12} \otimes |\psi\rangle_{35} \otimes |\phi\rangle_{46} \\
&= \frac{1}{N'_{\propto}}\big(A|\alpha\rangle_1|\alpha\rangle_2|\alpha\rangle_3|\alpha\rangle_4|\alpha\rangle_5|\alpha\rangle_6 \\
&\quad -A|\alpha\rangle_1|\alpha\rangle_2|\alpha\rangle_3|-\alpha\rangle_4|\alpha\rangle_5|-\alpha\rangle_6 \\
&\quad -A|\alpha\rangle_1|\alpha\rangle_2|-\alpha\rangle_3|\alpha\rangle_4|-\alpha\rangle_5|\alpha\rangle_6 \\
&\quad +A|\alpha\rangle_1|\alpha\rangle_2|-\alpha\rangle_3|-\alpha\rangle_4|-\alpha\rangle_5|-\alpha\rangle_6 \\
&\quad -B|-\alpha\rangle_1|-\alpha\rangle_2|\alpha\rangle_3|\alpha\rangle_4|\alpha\rangle_5|\alpha\rangle_6 \\
&\quad +B|-\alpha\rangle_1|-\alpha\rangle_2|\alpha\rangle_3|-\alpha\rangle_4|\alpha\rangle_5|-\alpha\rangle_6 \\
&\quad +B|-\alpha\rangle_1|-\alpha\rangle_2|-\alpha\rangle_3|\alpha\rangle_4|-\alpha\rangle_5|\alpha\rangle_6 \\
&\quad -B|-\alpha\rangle_1|-\alpha\rangle_2|-\alpha\rangle_3|-\alpha\rangle_4|-\alpha\rangle_5|-\alpha\rangle_6\big) \qquad (7\text{-}1)
\end{aligned}
$$

All of the above states are collected in the LGN. But here the role of action potentials is very important. They determine which fibers are fired. If fibers 1 and 2 carry action potentials, then it shows that information passes through ipsilateral fibers. Thus to select information from the LGN to send it via fibers 5 and 6 to V1 there is no need for phase difference (or to apply the phase shift operator on the states) and thus the state of (3-2) can be transferred like its first state through fibers 5 and 6. In the formulation (7-1) we see that if fibers x and y are fired, the expression with $|\alpha\rangle_x|\alpha\rangle_y$ should be selected from the terms with coefficients +A besides the expression $|-\alpha\rangle_x|-\alpha\rangle_y$ from the terms with coefficients $-B$. Hence,

$$1,2\ firing \xrightarrow{yields} |\phi'\rangle_{56} = A|\alpha\rangle_5|\alpha\rangle_6 - B|-\alpha\rangle_5|-\alpha\rangle_6 \qquad (7\text{-}2)$$

In another state, if fibers 1 and 4 are fired it means that one fiber is selected from ipsilateral fibers and the other is from contralateral fibers, thus they have phase difference with respect to each other. Hence,

$$1,4\ firing \xrightarrow{yields} |\phi'\rangle_{56} = A|\alpha\rangle_5|-\alpha\rangle_6 - B|-\alpha\rangle_5|\alpha\rangle_6 \qquad (7\text{-}3)$$

To restore initial information, the operator

$$R(\varphi) = e^{-i\pi a^{\dagger}{}_6 a_6}$$

should operate on the state in LGN in which fibers 1 and 4 have conveyed action potentials. This operator changes the ket $|\alpha\rangle_6$ to $|-\alpha\rangle_6$ and vice versa. It means that fiber 4 has a π radian phase difference with respect to fiber 1, and this phase difference can restore the exact state of the photon. If fibers 2 and 3 are fired, this yields:

$$2,3\ firing \xrightarrow{yields} |\phi'\rangle_{56} = A|-\alpha\rangle_5|\alpha\rangle_6 - B|\alpha\rangle_5|-\alpha\rangle_6 \qquad (7\text{-}4)$$

In this case the operator

$$R(\varphi) = e^{-i\pi a^{\dagger}{}_5 a_5}$$

should be involved. For the case of 3 and 4 firing, this yields,

$$3,4\ firing \xrightarrow{yields} |\phi'\rangle_{56} = A|-\alpha\rangle_5|-\alpha\rangle_6 - B|\alpha\rangle_5|\alpha\rangle_6 \qquad (7\text{-}5)$$

in which case the operator

$$R(\varphi) = e^{-i\pi(a^{\dagger}{}_5 a_5 + a^{\dagger}{}_6 a_6)}$$

should be involved. In this case we see that the main path is that of ipsilateral fibers which are direct to each eye and fibers 3 and 4 both have a π radian phase difference with it. We also know that there are two LGNs and the left and right V1. Now, another question emerges. How do these two left and right parts in V1 can instantaneously receive information? To answer this question, we can say that the synaptic β-neurexin/neuroligin-1 adhesive protein complex is claimed to be a device mediating entanglement between the cytoskeletons of the cortical neurons. Thus the macroscopic coherent quantum state can extend through large brain cortical areas [58].

We see that crossing or rotation of neurons in the visual pathways has an important role in restoring information in the brain. Maybe rotations or crossings of neurons throughout the body are there for this very reason.

8) Discussion

In general, we can briefly summarize our approach by listing the following properties:

1- It can combine the Orch-OR model with Tegmark's approaches and the Thaheld conclusion in a compact physical model which is called Teleportation.
2- It investigates visual pathways from atomic scales to macroscopic scales. This approach includes classical descriptions as well as new answers to open questions.
3- It explains why the shape of action potentials stays the same. Classical models state that "sensations" are action potentials that reach the brain via sensory neurons, and "perception" is the awareness and interpretation of the sensation. It is reasonable to assume that the constant shape of action potentials cannot result in different profiles of information. Thus the shape of information should be due to neurons. In this approach MTs are the representatives of information carriers. In our approach action potentials just determine which neurons fire and which do not.
4- It describes why neurons cross at some point. This crossing causes a phase shift relative to a special pathway. In teleportation of entangled coherent states the phase shift operators can rebuild initial information.

5- It explains that how the inverted image on the retina is perceived in the brain as upright.
6- It can describe how different information can be simultaneously perceived as a binding nature of conscious experience. This can be done via quantum parallel processing.
7- It explains how the brain of the observer can receive quantum information from the environment. We can see that there still exists this possibility that the mind can play the main role in the measurement problem, and this is in accord with what London, Bauer, von Neumann and Wigner (initially) asserted.

9) Conclusions

In this paper we have theoretically demonstrated the plausibility of a quantum teleportation mechanism between the eye and the brain which can describe different aspects of visual processing through visual pathways. Our model covers both quantum and classical aspects of neuroscience. This mechanism can combine some features of the Orch-OR model with Tegmark's conclusions and Thaheld's belief in one general model.

Acknowledgement

V. Salari and M. Rahnama thank Dr. Vahid Sheybani, the head of Neuroscience Research Center, for his good instructions and helps.

References:

---

[1] Schrödinger. E, What is life? Cambridge University Press, London, 1944.
[2] Davies. P, Does quantum mechanics play a non-trivial role in life?, Biosystems, 78**,** 69, 2004.
[3] Matsuno. K, *Quantum and biological computation*, Biosystems, 35, 209-212, 1995.
[4] Penrose. R, *Shadows of the mind*, Oxford University Press, 1994.
[5] Matsuno. K and Paton. R. C, *Is there biology of quantum information?,* Biosystems, **55**, 39-46 2000.
[6] Vedral. V, *Entanglement hits the big time*, Nature, **425**, 28-9, 2003.
[7] Beck. F and Eccles. J.C, *Quantum aspects of brain activity and the role of consciousness*, Proc. Nat. Acad. Sci. USA, **89**, 11357-11362, 1992.
[8] Penrose. R, The Emperor's New Mind (Oxford University Press, Oxford, 1989); S.R., *Quantum coherence in MTs: a neural basis for emergent consciousness?*, Journal of Consciousness Studies, **1**, 98–118, 1994.
[9] Davies. P.C.W.s, *The Origin of Life* (Penguin, London, 2003); The Fifth Miracle (Simon & Schuster, New York, 1998)
[10] Tarlaci. S., NeuroQuantology 2003; 3: 292-294

[11] Thaheld. F. 2005 *Preprint* quant-ph/0509042
[12] Thaheld. F. 2006 *Preprint* quant-ph/060418

[13] Tegmark. M., Phys. Rev. E61, 4194-4206 (2000). It's preprint exists in [arXiv:quant-ph/9907009 v2].

[14] Penrose. R. and Hameroff. S.R., J. Conscious. Stud. **2**, 98(1995).
[15] Hameroff. S.R. and Penrose. R., J. Conscious. Stud. **3**, 36(1996).
[16] Hameroff. S.R., Philos. Trans. R. Soc. London Ser. A **356**, 1869(1998).
[17] Hameroff. S.R. ,NeuroQuantology , March 2007 , Vol 5 , Issue 1 , Page 1-8

[18] Dowling. J. E., *The Retina: An Approachable Part of the Brain*. Belknap Press of Hrward University Press, Cambridge. (1987).

[19] Portet. S., Tuszynski. J. A., Hogue. C. M. and Dixon. J. M., *Elastic vibrations in seamless microtubules*, European Biophysics Journal, 34, Number7, October 2005, 912-920.

[20] Gerry, C., Knight, P. "*Introductory Quantum Optics*", Cmbridge University Press,(2005).

[21] Hameroff, S.R., 1994. *Quantum coherence in MTs: a neural basis for emergent consciousness?* J. Conscious. Stud. 1, 98–118.

[22] Penrose, R., 1989. *The Emperor's New Mind*. Oxford UniversityPress, Oxford.

[23] Kandel, E. R., Siegelbaum, S. A. & Schwartz, J. H. 1991 *Synaptic transmission.* In Principles of neural science (ed. E. R. Kandel, J. H. Schwartz & T. M. Jessell), 3rd edn, pp. 121-134. New York: Elsevier. Karplus, M. & McCammon

[24] Jibu, M. *On a heuristic model of the coherent mechanism of the global reaction process of a group of cells. Bussei Kenkyuu (Mater. Phys. Res.)* **53**, 431-436. (In Japanese.) (1990)

[25] Marshall, I. N. *Consciousness and Bose-Einstein condensates*. New Ideas Psychol. **7**, 73, 83. (1989)

[26] Penrose, R. *Newton, quantum theory and reality*. In 300 years of gravity (ed. S. W. Hawking& W. Israel). Cambridge University Press. (1987).

[27] Bohm, D. & Hiley, B. J. 1993 *The undivided universe*. New York: Routledge.

[28] Jibu, M. & Yasue, K. 1995 *Quantum brain dynamics: an introduction*. Amsterdam: John Benjamins.

[29] Stapp, H. P. 1993 *Mind, matter and quantum mechanics*. Berlin: Springer.

[30] Hameroff. S. R., BioSystems **77** 119-136 . (2004).

[31] Mavromatos, N. E., Mershin, A. and Nanopoulos, D. V. (2002), "*QED-Cavity model of MTs implies dissipationless energy transfer and biological quantum teleportation*," http://arxiv.org/abs/quant-ph/0204021

[32] Del Giudice. E., Doglia. S., Milani. M. and Vitiello. G., Nucl. Phys. B275 (FS 17) (1986), 185.

[33] Del Giudice. E., Preparata. G. and Vitiello. G., Phys. Rev. Lett. 61 (1988), 1085.

[34] Dicke. R. H., Phys. Rev. 93 (1954) 99.

[35] McCall. S. L. and Hahn. E. L., Phys. Rev. Lett. 18 (1967) 908

[36] Hameroff. S. R., *Ultimate Computing: Biomolecular Consciousness and Nanotechnology* (North-Holland, Amsterdam, 1987), and Journal of Consciousness Studies, 1 (1994) 91.

[37] Bokkon. I., *Phosphene phenomenon: A new concept*, BioSystems 92(2008) 168-174.

[38] Clegg. J. S. Amer. J. Physiol. **246**, R133 (1984).

[39] Clegg. J. Coll. Phenomena **3**, 289 (1981).

[40] Miyamoto, S. "*Changes in mobility of synaptic vesicles with assembly and disassembly of actin network.*" *Biochim. Biophys. Acta* **1244**, 85-91. (1995).

[41] Muallem, S., Kwiatkowska, K., Xu, X. & Yin, H. L. *Actin filament disassembly is a sufficient final trigger for exocytosis in nonexcitable cells*. *J. Cell Biol.* **128**, 589-598. (1995).

[42] Wachsstock, D. H., Schwarz, W. H. & Pollard, T. D. *Cross-linker dynamics determine the mechanical properties of actin gels*. *Biophys. Jl* **66**, 801-809. (1994).

[43] Ouyang. M., Awschalom. D. D., " *coherent spin transfer between moleculary bridged quantum dots*". Science, 301, 1074-1078, (2003).

[44] Fröhlich. H., Int, J. Quantum Chem. **2**, 641 (1968).

[45] Fröhlich. H., Nature (London) **228**, 1093 (1970).

[46] Fröhlich. H., Proc. Natl. Acad. Sci. U.S.A. **72**, 4211 (1975).

[47] Bennett. C. H., Brassard. G., Crepeau. C., Jozsa. R., Peses. A., and Wootters. W. K., Phys. Rev. Lett. **70**, 1895 (1993).

[48] Wootters. W. K. and Zurek. W. H., Nature (London) **299**, 802 (1982); Dieks .D., Phys. Lett. A **92**, 271 (1982).

[49] Jibu, M., Hagan, S., Hameroff, S.R., Pribram, K.H., Yasue, K. (1994). "*Quantum optical coherenc in cytoskeletal MTs: implications for brain function*", Biosystems 32, 95-209.

[50] Hameroff. S., Tuszynski. J., "*Quantum states in proteins and protein assemblies: The essence of life?*", Proceedings of SPIE Conference on Fluctuationa and Noise, Canary Islands, June 2004

[51] D. D. Georgiev, arXiv: quant-ph/ 0208053

[52] Web address: *http://thebrain.mcgill.ca/.../i_09_cl_dev.html*

[53] Zeki, S. " *The visual image in mind and brain*", Scientific American, 267:42-50 (1993).

[54] Hagan. S., Hameroff. S. R., and Tuszynski. J. A., *Quantum computation in brain MTs: Decoherence and biological feasibility.* Physical Review E, Volume **65**, 061901. (2002).

[55] Thaheld. F. H., *A method to explore the possibility of nonlocal correlations between brain electrical activities of two spatially separated animal subjects.* BioSystems **73**, 205-216.(2004).

[56] Thaheld. F. H., *Can we determine if the linear nature of quantum mechanics is violated by the perceptual process?* BioSystems **71**, 305-309.(2003).

[57] Thaheld. F. H., Comment on *Quantum superpositions and definite perceptions: envisaging new feasible experimental tests.* Phys. Lett. A **273**,232-234.(2000).

[58] Georgiev. D. D., Glazebrook. J. F., Informatica **30,** 221–232 (2006).

DTIC Copy

AFRL-PR-ED-TR-2003-0034 **AFRL-PR-ED-TR-2003-0034**

Teleportation Physics Study

Eric W. Davis
Warp Drive Metrics
4849 San Rafael Ave.
Las Vegas, NV 89120

August 2004

Special Report

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
EDWARDS AIR FORCE BASE CA 93524-7048

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 25-11-2003	**2. REPORT TYPE** Special	**3. DATES COVERED** *(From - To)* 30 Jan 2001 – 28 Jul 2003

4. TITLE AND SUBTITLE

Teleportation Physics Study

5a. CONTRACT NUMBER
F04611-99-C-0025

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62500F

6. AUTHOR(S)

Eric W. Davis

5d. PROJECT NUMBER
4847

5e. TASK NUMBER
0159

5f. WORK UNIT NUMBER
549907

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Warp Drive Metrics
4849 San Rafael Ave.
Las Vegas, NV 89120

8. PERFORMING ORGANIZATION REPORT NO.

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory (AFMC)
AFRL/PRSP
10 E. Saturn Blvd.
Edwards AFB CA 93524-7680

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-PR-ED-TR-2003-0034

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study was tasked with the purpose of collecting information describing the teleportation of material objects, providing a description of teleportation as it occurs in physics, its theoretical and experimental status, and a projection of potential applications. The study also consisted of a search for teleportation phenomena occurring naturally or under laboratory conditions that can be assembled into a model describing the conditions required to accomplish the transfer of objects. This included a review and documentation of quantum teleportation, its theoretical basis, technological development, and its potential applications. The characteristics of teleportation were defined and physical theories were evaluated in terms of their ability to completely describe the phenomena. Contemporary physics, as well as theories that presently challenge the current physics paradigm were investigated. The author identified and proposed two unique physics models for teleportation that are based on the manipulation of either the general relativistic spacetime metric or the spacetime vacuum electromagnetic (zero-point fluctuations) parameters. Naturally occurring anomalous teleportation phenomena that were previously studied by the United States and foreign governments were also documented in the study and are reviewed in the report. The author proposes an additional model for teleportation that is based on a combination of the experimental results from the previous government studies and advanced physics concepts. Numerous recommendations outlining proposals for further theoretical and experimental studies are given in the report. The report also includes an extensive teleportation bibliography.

15. SUBJECT TERMS

teleportation; physics, quantum teleportation; teleportation phenomena; anomalous teleportation; teleportation theories; teleportation experiments; teleportation bibliography

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Franklin B. Mead, Jr.
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	A	88	**19b. TELEPHONE NO** *(include area code)*
Unclassified	**Unclassified**	**Unclassified**			(661) 275-5929

Standard Form 298
(Rev. 8-98)
Prescribed by ANSI Std. 239.18

NOTICE

USING GOVERNMENT DRAWINGS, SPECIFICATIONS, OR OTHER DATA INCLUDED IN THIS DOCUMENT FOR ANY PURPOSE OTHER THAN GOVERNMENT PROCUREMENT DOES NOT IN ANY WAY OBLIGATE THE US GOVERNMENT. THE FACT THAT THE GOVERNMENT FORMULATED OR SUPPLIED THE DRAWINGS, SPECIFICATIONS, OR OTHER DATA DOES NOT LICENSE THE HOLDER OR ANY OTHER PERSON OR CORPORATION; OR CONVEY ANY RIGHTS OR PERMISSION TO MANUFACTURE, USE, OR SELL ANY PATENTED INVENTION THAT MAY RELATE TO THEM.

FOREWORD

This Special Technical Report presents the results of a subcontracted study performed by Warp Drive Metrics, Las Vegas, NV, under Contract No. F04611-99-C-0025, for the Air Force Research Laboratory (AFRL)/Space and Missile Propulsion Division, Propellant Branch (PRSP), Edwards AFB, CA. The Project Manager for AFRL/PRSP was Dr. Franklin B. Mead, Jr.

This report has been reviewed and is approved for release and distribution in accordance with the distribution statement on the cover and on the SF Form 298. This report is published in the interest of scientific and technical information exchange and does not constitute approval or disapproval of its ideas or findings.

//Signed//

FRANKLIN B. MEAD, JR.
Project Manager

//Signed//

RONALD E. CHANNELL
Chief, Propellants Branch

//Signed//

PHILIP A. KESSEL
Technical Advisor, Space and Missile
Propulsion Division

//Signed//

RANNEY G. ADAMS III
Public Affairs Director

AFRL-ERS-PAS-04-155

This Page Intentionally Left Blank

Table of Contents

Section **Page**
1.0 INTRODUCTION……………1
  1.1 Introduction……………1
  1.2 The Definitions of Teleportation……………1
2.0 vm -TELEPORTATION……………3
  2.1 Engineering the Spacetime Metric……………3
    2.1.1 Wormhole Thin Shell Formalism……………3
    2.1.2 "Exotic" Matter-Energy Requirements……………11
  2.2 Engineering the Vacuum……………11
    2.2.1 The Polarizable-Vacuum Representation of General Relativity……………20
  2.3 Conclusion and Recommendations……………26
3.0 q-TELEPORTATION……………30
  3.1 Teleportation Scenario……………30
  3.2 Quantum Teleportation……………32
    3.2.1 Description of the q-Teleportation Process……………34
    3.2.2 Decoherence Fundamentally Limits q-Teleportation……………38
    3.2.3 Recent Developments in Entanglement and q-Teleportation Physics……………38
  3.3 Conclusion and Recommendations……………46
4.0 e-TELEPORTATION……………50
  4.1 Extra Space Dimensions and Parallel Universes/Spaces……………50
  4.2 Vacuum Hole Teleportation……………52
  4.3 Conclusion and Recommendations……………53
5.0 p-TELEPORTATION……………55
  5.1 PK Phenomenon……………55
    5.1.1 Hypothesis Based on Mathematical Geometry……………60
  5.2 Conclusion and Recommendations……………61
6.0 REFERENCES……………63
APPENDIX A – A Few Words About Negative Energy……………73
  A.1 A General Relativistic Definition of Negative or Exotic Energy……………73
  A.2 Squeezed Quantum States and Negative Energy……………73
APPENDIX B – THεμ Methodology……………75

List of Figures

Figure **Page**

Figure 1. Diagram of a Simultaneous View of Two Remote Compact Regions (Ω_1 and Ω_2) of Minkowski Space Used to Create the Wormhole Throat $\partial\Omega$, Where Time is Suppressed in This Representation……………..……………………..……………………………………5

Figure 2. The Same Diagram as in Figure 1 Except as Viewed by an Observer Sitting in Region Ω_1 Who Looks Through the Wormhole Throat $\partial\Omega$ and Sees Remote Region Ω_2 (Dotted Area Inside the Circle) on the Other Side……………………………………………………. 6

Figure 3. A Thin Shell of (Localized) Matter-Energy, or Rather the Two-Dimensional Spacelike Hypersurface $\partial\Omega$ (via (2.3)), Possessing the Two Principal Radii of Curvature ρ_1 and ρ_2……..8

Figure 4. A Schematic of Vacuum Quantum Field Fluctuations (a.k.a. Vacuum Zero Point Field Fluctuations) Involved in the "Light-by-Light" Scattering Process That Affects the Speed of Light………………………………………………………………………………………13

Figure 5. A Schematic of the Casimir Effect Cavity/Waveguide………………………………………...15

Figure 6. Classical Facsimile Transmission (Modified IBM Press Image)………………………………35

Figure 7. Quantum Teleportation (Modified IBM Press Image)…………………………………………36

Figure 8. Quantum Teleportation (From www.aip.org)..43

Approved for public release; distribution unlimited.

List of Tables

Table **Page**
Table 1. Metric Effects in the PV-GR Model When K > 1 (Compared With Reference Frames at Asymptotic Infinity Where K = 1)…………………………………………………………..…21
Table 2. Metric Effects in the PV-GR Model When K < 1 (Compared With Reference Frames at Asymptotic Infinity Where K = 1)…………………………………………………………...22
Table 3. Substantial Gravitational Squeezing Occurs When $\lambda \geq 8\pi r_s$ (For Electromagnetic ZPF)............28

Approved for public release; distribution unlimited.

Glossary

AEC	Average Energy Condition
AFRL	Air Force Research Laboratory
AU	Astronomical Unit
BBO	Beta (β)-Barium Borate
CGS	Centimeter-Gram-Second
CIA	Central Intelligence Agency
DARPA	Defense Advanced Research Projects Agency
DEC	Dominant Energy Condition
DIA	Defense Intelligence Agency
DNA	Deoxyribo Nucleic Acid
DoD	Department of Defense
EPR	Einstein, Podolsky and Rosen
ESP	Extrasensory Perception
eV	Electron Volt
FRW	Friedmann-Robertson-Walker
FTL	Faster-Than-Light
IBM	International Business Machines
INSCOM	Intelligence and Security Command
IR	Infrared
MeV	Mega-Electron Volt
MKS	Meter-Kilogram-Second
NEC	Null Energy Condition
NLP	Neuro-Linguistic Programming
NMR	Nuclear Magnetic Resonance
NSA	National Security Agency
PK	Psychokinesis
PPN	Parameterized Post-Newtonian
PRC	Peoples Republic of China
PV-GR	Polarizable-Vacuum Representation of General Relativity
QED	Quantum Electrodynamics
QISP	Quantum Information Science Program
R&D	Research and Development
SAIC	Science Applications International Corporation
SEC	Strong Energy Condition
SRI	Stanford Research Institute
USSR	Union of Soviet Socialist Republics
UV	Ultraviolet
WEC	Weak Energy Condition
ZPE	Zero-Point Energy
ZPF	Zero-Point Fluctuations

Acknowledgements

This study would not have been possible without the very generous support of Dr. Frank Mead, Senior Scientist at the Advanced Concepts Office of the U.S. Air Force Research Laboratory (AFRL) Propulsion Directorate at Edwards AFB, CA. Dr. Mead's collegial collaboration, ready assistance, and constant encouragement were invaluable to me. Dr. Mead's professionalism and excellent rapport with "out-of-the-box" thinkers excites and motivates serious exploration into advanced concepts that push the envelope of knowledge and discovery. The author owes a very large debt of gratitude and appreciation to both Dr. David Campbell, Program Manager, ERC, Inc. at AFRL, Edwards AFB, CA, and the ERC, Inc. staff, for supporting the project contract and for making all the paperwork fuss totally painless. Dr. Campbell and his staff provided timely assistance when the author needed it, which helped make this contract project run smoothly.

There are two colleagues who provided important contributions to this study that I wish to acknowledge. First, I would like to express my sincere thanks and deepest appreciation to my first longtime mentor and role model, the late Dr. Robert L. Forward. Bob Forward was the first to influence my interests in interstellar flight and advanced breakthrough physics concepts (i.e., "Future Magic") when I first met him at an AIAA Joint Propulsion Conference in Las Vegas while I was in high school (ca. 1978). The direction I took in life from that point forward followed the trail of exploration and discovery that was blazed by Bob. I will miss him, but I will never forget him. Second, I would like to express my sincere thanks and appreciation to my longtime friend, colleague and present mentor, Dr. Hal Puthoff, Institute for Advanced Studies-Austin, for our many discussions on applying his Polarizable Vacuum-General Relativity model to a quasi-classical teleportation concept. Hal taught me to expand my mind, and he encourages me to think outside the box. He also gave me a great deal of valuable insight and personal knowledge about the Remote Viewing Program. Last, I would like to offer my debt of gratitude and thanks to my business manager (and spouse), Lindsay K. Davis, for all the hard work she does to make the business end of Warp Drive Metrics run smoothly.

Eric W. Davis, Ph.D., FBIS
Warp Drive Metrics
Las Vegas, NV

Preface

The Teleportation Physics Study is divided into four phases. Phase I is a review and documentation of quantum teleportation, its theoretical basis, technological development, and its potential application. Phase II developed a textbook description of teleportation as it occurs in classical physics, explored its theoretical and experimental status, and projected its potential applications. Phase III consisted of a search for teleportation phenomena occurring naturally or under laboratory conditions that can be assembled into a model describing the conditions required to accomplish the disembodied conveyance of objects. The characteristics of teleportation were defined, and physical theories were evaluated in terms of their ability to completely describe the phenomenon. Presently accepted physics theories, as well as theories that challenge the current physics paradigm were investigated for completeness. The theories that provide the best chance of explaining teleportation were selected, and experiments with a high chance of accomplishing teleportation were identified. Phase IV is the final report.

The report contains five chapters. Chapter 1 is an overview of the textbook descriptions for the various teleportation phenomena that are found in nature, in theoretical physics concepts, and in experimental laboratory work. Chapter 2 proposes two quasi-classical physics concepts for teleportation: the first is based on engineering the spacetime metric to induce a traversable wormhole; the second is based on the polarizable-vacuum-general relativity approach that treats spacetime metric changes in terms of equivalent changes in the vacuum permittivity and permeability constants. These concepts are theoretically developed and presented. Promising laboratory experiments were identified and recommended for further research. Chapter 3 presents the current state-of-art of quantum teleportation physics, its theoretical basis, technological development, and its applications. Key theoretical, experimental, and applications breakthroughs were identified, and a series of theoretical and experimental research programs are proposed to solve technical problems and advance quantum teleportation physics. Chapter 4 gives an overview of alternative teleportation concepts that challenge the present physics paradigm. These concepts are based on the existence of parallel universes/spaces and/or extra space dimensions. The theoretical and experimental work that has been done to develop these concepts is reviewed, and a recommendation for further research is made. Last, Chapter 5 gives an in-depth overview of unusual teleportation phenomena that occur naturally and under laboratory conditions. The teleportation phenomenon discussed in the chapter is based on psychokinesis (PK), which is a category of psychotronics. The U.S. military-intelligence literature is reviewed, which relates the historical scientific research performed on PK-teleportation in the U.S., China and the former Soviet Union. The material discussed in the chapter largely challenges the current physics paradigm; however, extensive controlled and repeatable laboratory data exists to suggest that PK-teleportation is quite real and that it is controllable. The report ends with a combined list of references.

1.0 INTRODUCTION

1.1 Introduction

The concept of teleportation was originally developed during the Golden Age of 20th century science fiction literature by writers in need of a form of instantaneous disembodied transportation technology to support the plots of their stories. Teleportation has appeared in such SciFi literature classics as Algis Budry's *Rogue Moon* (Gold Medal Books, 1960), A. E. van Vogt's *World of Null-A* (Astounding Science Fiction, August 1945), and George Langelaan's *The Fly* (Playboy Magazine, June 1957). The Playboy Magazine short story led to a cottage industry of popular films decrying the horrors of scientific technology that exceeded mankind's wisdom: *The Fly* (1958), *Return of the Fly* (1959), *Curse of the Fly* (1965), *The Fly* (a 1986 remake), and *The Fly II* (1989). The teleportation concept has also appeared in episodes of popular television SciFi anthology series such as *The Twilight Zone* and *The Outer Limits*. But the most widely recognized pop-culture awareness of the teleportation concept began with the numerous *Star Trek* television and theatrical movie series of the past 39 years (beginning in 1964 with the first TV series pilot episode, *The Cage*), which are now an international entertainment and product franchise that was originally spawned by the late genius television writer-producer Gene Roddenberry. Because of *Star Trek* everyone in the world is familiar with the "transporter" device, which is used to teleport personnel and material from starship to starship or from ship to planet and vice versa at the speed of light. People or inanimate objects would be positioned on the transporter pad and become completely disintegrated by a beam with their atoms being patterned in a computer buffer and later converted into a beam that is directed toward the destination, and then reintegrated back into their original form (all without error!). "*Beam me up, Scotty*" is a familiar automobile bumper sticker or cry of exasperation that were popularly adopted from the series.

However, the late Dr. Robert L. Forward (2001) stated that modern hard-core SciFi literature, with the exception of the ongoing *Star Trek* franchise, has abandoned using the teleportation concept because writers believe that it has more to do with the realms of parapsychology/paranormal (a.k.a. psychic) and imaginative fantasy than with any realm of science. Beginning in the 1980s developments in quantum theory and general relativity physics have succeeded in pushing the envelope in exploring the reality of teleportation. A crescendo of scientific and popular literature appearing in the 1990s and as recently as 2003 has raised public awareness of the new technological possibilities offered by teleportation. As for the psychic aspect of teleportation, it became known to Dr. Forward and myself, along with several colleagues both inside and outside of government, that anomalous teleportation has been scientifically investigated and separately documented by the Department of Defense.

It has been recognized that extending the present research in quantum teleportation and developing alternative forms of teleportation physics would have a high payoff impact on communications and transportation technologies in the civilian and military sectors. It is the purpose of this study to explore the physics of teleportation and delineate its characteristics and performances, and to make recommendations for further studies in support of Air Force Advanced Concepts programs.

1.2 The Definitions of Teleportation

Before proceeding, it is necessary to give a definition for each of the teleportation concepts I have identified during the course of this study:

- *Teleportation* – SciFi: the disembodied transport of persons or inanimate objects across space by advanced (futuristic) technological means (adapted from Vaidman, 2001). We will call this *sf-Teleportation*, which will not be considered further in this study.

- *Teleportation* – psychic: the conveyance of persons or inanimate objects by psychic means. We will call this *p-Teleportation*.

- *Teleportation* – engineering the vacuum or spacetime metric: the conveyance of persons or inanimate objects across space by altering the properties of the spacetime vacuum, or by altering the spacetime metric (geometry). We will call this *vm-Teleportation*.

- *Teleportation* – quantum entanglement: the disembodied transport of the quantum state of a system and its correlations across space to another system, where *system* refers to any single or collective particles of matter or energy such as baryons (protons, neutrons, etc.), leptons (electrons, etc.), photons, atoms, ions, etc. We will call this *q-Teleportation*.

- *Teleportation* – exotic: the conveyance of persons or inanimate objects by transport through extra space dimensions or parallel universes. We will call this *e-Teleportation*.

We will examine each of these in detail in the following chapters and determine whether any of the above teleportation concepts encompass the instantaneous and or disembodied conveyance of objects through space.

2.0 vm-TELEPORTATION

2.1 Engineering the Spacetime Metric

A comprehensive literature search for vm-Teleportation within the genre of spacetime metric engineering yielded no results. No one in the general relativity community has thought to apply the Einstein field equation to determine whether there are solutions compatible with the concept of teleportation. Therefore, I will offer two solutions that I believe will satisfy the definition of vm-Teleportation. The first solution can be found from the class of traversable wormholes giving rise to what I call a true "stargate." A stargate is essentially a wormhole with a flat-face shape for the throat as opposed to the spherical-shaped throat of the Morris and Thorne (1988) traversable wormhole, which was derived from a spherically symmetric Lorentzian spacetime metric that prescribes the wormhole geometry (see also, Visser, 1995 for a complete review of traversable Lorentzian wormholes):

$$ds^2 = -e^{2\phi(r)}c^2dt^2 + [1-b(r)/r]^{-1}dr^2 + r^2d\Omega^2 \qquad (2.1),$$

where by inspection we can write the traversable wormhole metric tensor in the form

$$g_{\alpha\beta} = \begin{pmatrix} -e^{2\phi(r)} & 0 & 0 & 0 \\ 0 & [1-b(r)/r]^{-1} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin^2\theta \end{pmatrix} \qquad (2.2)$$

using standard spherical coordinates, where c is the speed of light, $\alpha,\beta \equiv (0 = t, 1 = r, 2 = \theta, 3 = \varphi)$ are the time and space coordinate indices ($-\infty < t < \infty$; r: $2\pi r$ = circumference; $0 \le \theta \le \pi$; $0 \le \varphi \le 2\pi$), $d\Omega^2 = d\theta^2 + sin^2\theta d\varphi^2$, $\phi(r)$ is the freely specifiable redshift function that defines the proper time lapse through the wormhole throat, and $b(r)$ is the freely specifiable shape function that defines the wormhole throat's spatial (hypersurface) geometry. Such spacetimes are asymptotically flat. The Einstein field equation requires that a localized source of matter-energy be specified in order to determine the geometry that the source induces on the local spacetime. We can also work the Einstein equation backwards by specifying the local geometry in advance and then calculate the matter-energy source required to induce the desired geometry. The Einstein field equation thus relates the spacetime geometry terms comprised of the components of the metric tensor and their derivatives (a.k.a. the Einstein tensor) to the local matter-energy source terms comprised of the energy and stress-tension densities (a.k.a. the stress-energy tensor). The flat-face wormhole or stargate is derived in the following section.

2.1.1 Wormhole Thin Shell Formalism

The flat-face traversable wormhole solution is derived from the thin shell (a.k.a. junction condition or surface layer) formalism of the Einstein equations (Visser, 1989; see also, Misner, Thorne and Wheeler, 1973). We adapt Visser's (1989) development in the following discussion. The procedure is to take two copies of flat Minkowski space and remove from each identical regions of the form $\Omega \times \Re$, where Ω is a three-dimensional compact spacelike hypersurface and $\Re$ is a timelike straight line (time axis). Then identify these two incomplete spacetimes along the timelike boundaries $\partial\Omega \times \Re$. The resulting spacetime is geodesically complete and possesses two asymptotically flat regions connected by a wormhole. The throat of the wormhole is just the junction $\partial\Omega$ (a two-dimensional space-like hypersurface) at which the two original Minkowski spaces are identified (see Figures 1 and 2).

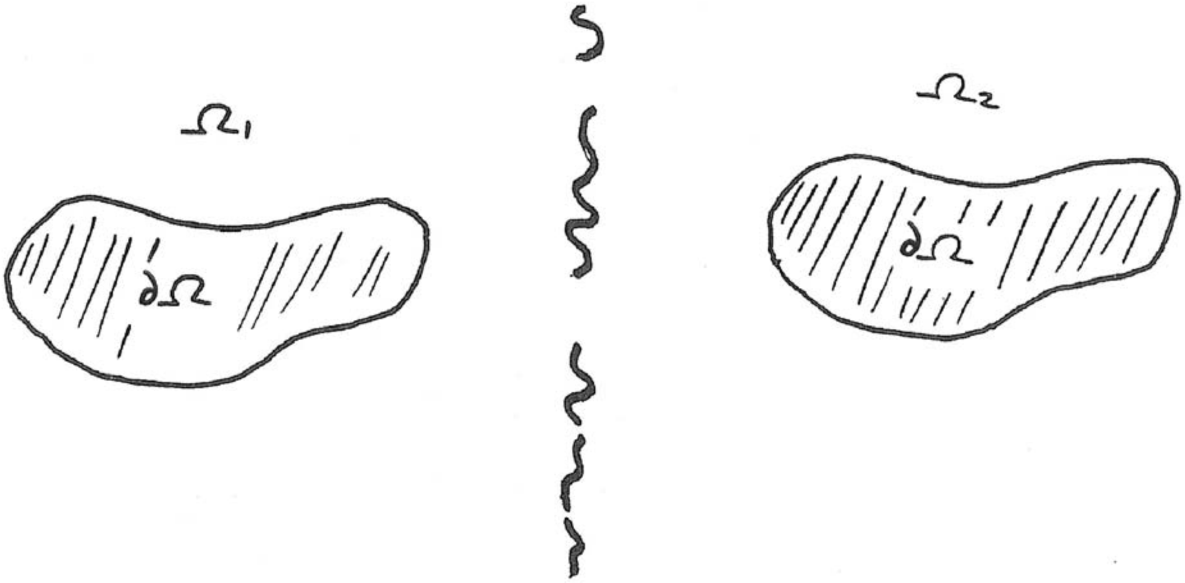
Ω_1
$\partial\Omega$
Ω_2
$\partial\Omega$

Figure 1. Diagram of a Simultaneous View of Two Remote Compact Regions (Ω_1 and Ω_2) of Minkowski Space Used to Create the Wormhole Throat $\partial\Omega$, Where Time is Suppressed in This Representation (adapted from Bennett et al., 1995)

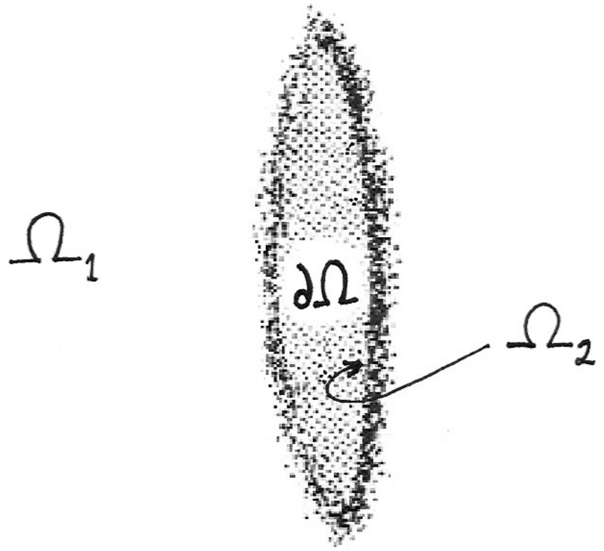

Figure 2. The Same Diagram as in Figure 1 Except as Viewed by an Observer Sitting in Region Ω_1 Who Looks Through the Wormhole Throat $\partial\Omega$ and Sees Remote Region Ω_2 (Dotted Area Inside the Circle) on the Other Side

The resulting spacetime is everywhere Riemann-flat except possibly at the throat. Also, the stress-energy tensor in this spacetime is concentrated at the throat with a δ-function singularity there. This is a consequence of the fact that the spacetime metric at the throat is continuous but not differentiable, while the connection is discontinuous; thus causing the Riemann curvature to possess a δ-function singularity (causing undesirable gravitational tidal forces) there. The magnitude of this δ-function singularity can be calculated in terms of the second fundamental form on both sides of the throat, which we presume to be generated by a localized thin shell of matter-energy. The second fundamental form represents the extrinsic curvature of the ∂Ω hypersurface (i.e., the wormhole throat), telling how it is curved with respect to the enveloping four-dimensional spacetime. The form of the geometry is simple, so the second fundamental form at the throat is calculated to be (McConnell, 1957):

$$\begin{aligned} K^{i}{}_{j}{}^{\pm} &= \pm\begin{pmatrix} \kappa_0 & 0 & 0 \\ 0 & \kappa_1 & 0 \\ 0 & 0 & \kappa_2 \end{pmatrix} \\ &= \pm\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/\rho_1 & 0 \\ 0 & 0 & 1/\rho_2 \end{pmatrix} \end{aligned} \qquad (2.3),$$

where $i,j = 0,1,2$ and $K^{i\,\pm}_{j}$ is the second fundamental form. The full 4×4 matrix $K^{\alpha}{}_{\beta}$ has been reduced to 3×3 form, as above, for computational convenience because the thin shell (or hypersurface) is essentially a two-surface embedded in three-space. The overall ± sign in equation (2.3) comes from the fact that a unit normal points outward from one side of the surface and points inward on the other side. We hereafter drop the ± sign for the sake of brevity in notation. The quantities κ_0, κ_1, and κ_2 measure the extrinsic curvature of the thin shell of local matter-energy (i.e., the stuff that induces the wormhole throat geometry). Since the wormhole throat is a space-like hypersurface, we can exclude time-like hypersurfaces and their components in the calculations. Therefore we set $\kappa_0 = 0$ in equation (2.3) because it is the time-like extrinsic curvature for the time-like hypersurface of the thin shell of matter-energy. As seen in equation (2.3) κ_1 and κ_2 are simply related to the two principal radii of curvature ρ_1 and ρ_2 (defined to be the eigenvalues of $K^{i}{}_{j}$) of the two-dimensional spacelike hypersurface ∂Ω (see Figure 3). It should be noted that a convex surface has positive radii of curvature, while a concave surface has negative radii of curvature.

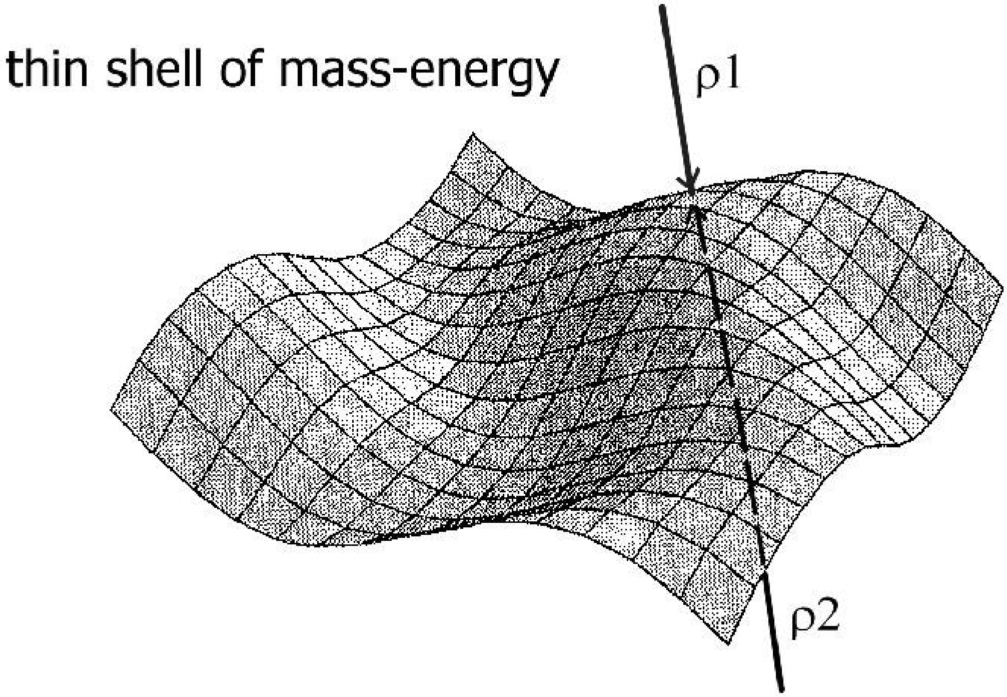
thin shell of mass-energy
ρ1
ρ2

Figure 3. A Thin Shell of (Localized) Matter-Energy, or Rather the Two-Dimensional Spacelike Hypersurface $\partial\Omega$ (via (2.3)), Possessing the Two Principal Radii of Curvature ρ_1 and ρ_2

Approved for public release; distribution unlimited.

It is a standard result of the thin shell or junction condition formalism that the Einstein field equation may be cast in terms of the surface stress-energy tensor S^i_j of the thin matter-energy shell localized in ∂Ω (note: we are exploiting the symmetry of the wormhole with respect to interchange of the two flat regions Ω_1 and Ω_2):

$$S^i{}_j = -\frac{c^4}{4\pi G}\left(K^i{}_j - \delta^i{}_j K^k{}_k\right) \quad (2.4),$$

where G is Newton's gravitational constant and δ^i_j is the (three-dimensional) unit matrix. K^k_k is the trace of equation (2.3):

$$\begin{aligned} K^k{}_k &= Tr\, K^i{}_j \\ &= \frac{1}{\rho_1} + \frac{1}{\rho_2} \end{aligned} \qquad (2.5)$$

and

$$\delta^i{}_j K^k{}_k = \begin{pmatrix} \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 & 0 \\ 0 & \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 \\ 0 & 0 & \frac{1}{\rho_1} + \frac{1}{\rho_2} \end{pmatrix} \quad (2.6).$$

Substituting (2.3) and (2.6) into (2.4) gives (after simplification):

$$S^i{}_j = \frac{c^4}{4\pi G}\begin{pmatrix} \frac{1}{\rho_1} + \frac{1}{\rho_2} & 0 & 0 \\ 0 & 1/\rho_2 & 0 \\ 0 & 0 & 1/\rho_1 \end{pmatrix} \quad (2.7).$$

The thin matter-energy shell's surface stress-energy tensor may be interpreted in terms of the surface energy density σ and principal surface tensions ϑ_1 and ϑ_2:

$$S^i{}_j = \begin{pmatrix} -\sigma & 0 & 0 \\ 0 & -\vartheta_1 & 0 \\ 0 & 0 & -\vartheta_2 \end{pmatrix} \quad (2.8).$$

Thus we arrive at the Einstein field equation by equating (2.8) and (2.7) and multiplying both sides by –1:

$$\begin{pmatrix} \sigma & 0 & 0 \\ 0 & \vartheta_1 & 0 \\ 0 & 0 & \vartheta_2 \end{pmatrix} = -\frac{c^4}{4\pi G} \begin{pmatrix} \frac{1}{\rho_1}+\frac{1}{\rho_2} & 0 & 0 \\ 0 & 1/\rho_2 & 0 \\ 0 & 0 & 1/\rho_1 \end{pmatrix} \qquad (2.9),$$

which gives the final result

$$\sigma = -\frac{c^4}{4\pi G}\left(\frac{1}{\rho_1}+\frac{1}{\rho_2}\right) \qquad (2.10a)$$

$$\vartheta_1 = -\frac{c^4}{4\pi G}\frac{1}{\rho_2} \qquad (2.10b)$$

$$\vartheta_2 = -\frac{c^4}{4\pi G}\frac{1}{\rho_1} \qquad (2.10c).$$

These are the Einstein equations. Equations (2.10a-c) imply that (for $\partial\Omega$ convex) we are dealing with ***negative*** surface energy density and ***negative*** surface tensions. This result is in fact the primary matter-energy requirement for traversable wormholes, as was proved by Morris and Thorne (1988), and later by Visser (1995), within the paradigm of classical Einstein general relativity. The negative surface tension (= *positive outward pressure*, a.k.a. gravitational repulsion or antigravity) is needed to keep the throat open and stable against collapse. The reader should **not** be alarmed at this result. Negative energies and negative stress-tensions are an acceptable result both mathematically and physically, and they manifest gravitational repulsion (antigravity!) in and around the wormhole throat. One only needs to understand what it means for stress-energy to be negative within the proper context. In general relativity the term "*exotic*" is used in place of "*negative*." The effects of negative energy have been produced in the laboratory (the Casimir Effect is one example). In short, negative energy arises from Heisenberg's quantum uncertainty principle, which requires that the energy density of any electromagnetic, magnetic, electric or other fields must fluctuate randomly. Even in a vacuum, where the average energy density is zero, the energy density fluctuates. This means that the quantum vacuum can never remain truly empty in the classical sense of the term. The quantum picture of the vacuum is that of a turbulent plenum of virtual (i.e., energy non-conserving) particle pairs that spontaneously pop in and out of existence. The notion of "zero energy" in quantum theory corresponds to the vacuum being filled with such fluctuations going on. This issue is further elaborated on and clarified in greater detail in Appendix A. We will also revisit this in Section 2.2. Finally, it should be noted that for the analysis in this section we assumed an ultrastatic wormhole [i.e., $g_{00} \equiv 1 \Rightarrow \phi(r) = 0$ in equation (2.1)] with the "exotic" matter-energy confined to a thin layer, and we dispensed with the assumption of spherical symmetry.

We can now build a wormhole-stargate and affect *vm-Teleportation* such that a traveler stepping into the throat encounters **no** exotic matter-energy there. This will require that our wormhole be flat shaped. To make the wormhole flat requires that we choose the throat $\partial\Omega$ to have at least one flat face (picture the thin shell in Figure 3 becoming a flat shell). On that face the two principal radii of curvature become $\rho_1 = \rho_2 = \infty$ as required by standard geometry. Substituting this into equations (2.10a-c) gives

$$\sigma = \vartheta_1 = \vartheta_2 = 0 \qquad (2.11),$$

which is a remarkable result. A further consequence of this is that now $K^i_j = 0$, thus making the Riemann curvature and stress-energy tensors (Riemann: $R^\alpha{}_\beta \sim K^\alpha{}_\beta$; stress-energy: $T^\alpha{}_\beta \sim K^\alpha{}_\beta$) at the throat become zero such that the associated δ-function singularities disappear there. This means that a traveler encountering and going through such a wormhole will feel no tidal gravitational forces and see no exotic matter-energy (that threads the throat). A traveler stepping through the throat will simply be ***teleported*** into the other remote spacetime region or another universe (note: the Einstein equation does not fix the spacetime topology, so it is possible that wormholes are inter-universe as well as intra-universe tunnels). We construct such a *teleportation* stargate by generating a thin shell or surface layer of "exotic" matter-energy much like a thin film of soap stretched across a loop of wire.

2.1.2 "Exotic" Matter-Energy Requirements

Now we have to estimate the amount of negative (or exotic) mass-energy that will be needed to generate and hold open a *vm-Teleportation* wormhole. A simple formula originally due to Visser (1995) for short-throat wormholes using the thin shell formalism gives:

$$\begin{aligned} M_{wh} &= -\frac{r_{throat}c^2}{G} \\ &= -(1.3469\,x10^{27}\,kg)\frac{r_{throat}}{1\,meter} \qquad (2.12), \\ &= -(0.709\,M_{Jupiter})\frac{r_{throat}}{1\,meter} \end{aligned}$$

where M_{wh} is the mass required to build the wormhole, r_{throat} is a suitable measure of the linear dimension (radius) of the throat, and $M_{Jupiter}$ is the mass of the planet Jupiter (1.90×10^{27} kg). Equation (2.12) demonstrates that a mass of $-0.709\ M_{Jupiter}$ (or -1.3469×10^{27} kg) will be required to build a wormhole 1 meter in size. As the wormhole size increases the mass requirement grows negative-large, and vice versa as the wormhole size decreases. After being alarmed by the magnitude of this, one should note that M_{wh} is **not** the total mass of the wormhole as seen by observers at remote distances. The non-linearity of the Einstein field equations dictates that the total mass is zero (actually, the total net mass being positive, negative or zero in the Newtonian approximation depending on the details of the negative energy configuration constituting the wormhole system). And finally, Visser et al. (2003) have demonstrated the existence of spacetime geometries containing traversable wormholes that are supported by ***arbitrarily small quantities*** of exotic matter-energy, and they proved that this was a general result. In Section 2.3 we will discuss how or whether we can create such a wormhole in the laboratory.

2.2 Engineering the Vacuum

Engineering the spacetime vacuum provides a second solution that also satisfies the definition of vm-Teleportation. The concept of "*engineering the vacuum*" was first introduced to the physics community by Lee (1988). Lee stated:

"*The experimental method to alter the properties of the vacuum may be called vacuum engineering…If indeed we are able to alter the vacuum, then we may encounter some new phenomena, totally unexpected.*"

This new concept is based on the now-accepted fact that the vacuum is characterized by physical parameters and structure that constitutes an energetic medium which pervades the entire extent of the universe. We note here the two most important defining properties of the vacuum in this regard (Puthoff et al., 2002):

- ☐ Within the context of quantum field theory the vacuum is the seat of all energetic particle and field fluctuations.

- ☐ Within the context of general relativity theory the vacuum is the seat of a spacetime structure (or metric) that encodes the distribution of matter and energy.

We begin our look into this concept by examining the propagation of light through space. We know from quantum field theory that light propagating through space interacts with the vacuum quantum fields (a.k.a. vacuum quantum field fluctuations). The observable properties of light, including the speed of light, are determined by these interactions. Vacuum quantum interactions with light lead to an effect on the speed of light that is due to the absorption of photons (by the vacuum) to form virtual electron-positron pairs followed by the quick re-emission (from the vacuum) of the photon (see Figure 4). The virtual particle pairs are very short lived because of the large mismatch between the energy of a photon and the rest mass-energy of the particle pair. A key point is that this process makes a contribution to the observed vacuum permittivity ε_0 (and permeability μ_0) constant and, therefore, to the speed of light c [$c = (\varepsilon_0\mu_0)^{-1/2}$].

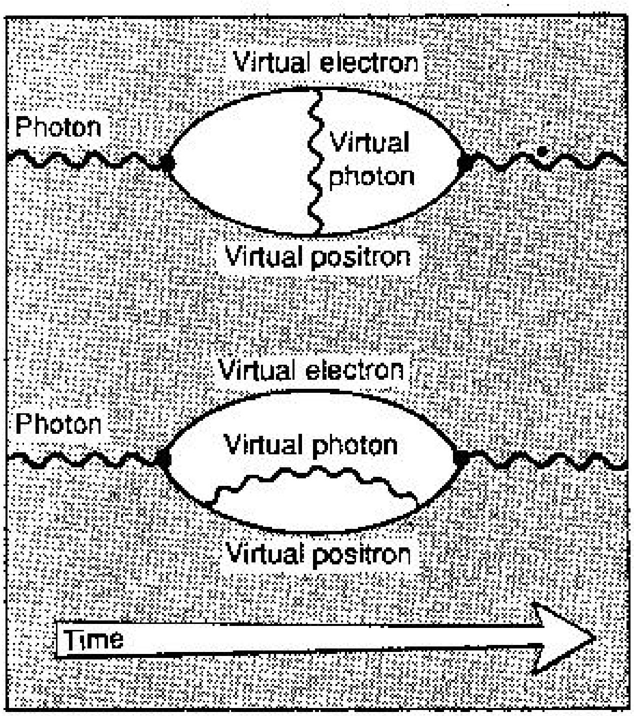
Virtual electron
Photon
Virtual photon
Virtual positron
Virtual electron
Photon
Virtual photon
Virtual positron
Time

Photons can change into a variety of short-lived, or "virtual", particles. as they fly through the vacuum. The processes above, represented by Feynman diagrams, are "two-loop"' processes. They affect light's speed

Figure 4. A Schematic of Vacuum Quantum Field Fluctuations (a.k.a. Vacuum Zero Point Field Fluctuations) Involved in the "Light-by-Light" Scattering Process That Affects the Speed of Light (from Chown, 1990)

The role of virtual particle pairs in determining the ε_0 (μ_0) of the vacuum is analogous to that of atoms/molecules in determining the relative permittivity ε (and μ) of a dielectric material. We know that the absorption/re-emission of photons by atoms/molecules in a transparent medium (note: there are no strongly absorbing resonances, so the atoms/molecules remain in their excited states for a very short time before re-emitting photons) is responsible for the refractive index of the medium, which results in the reduction of the speed of light for photons propagating through the medium. This absorption/re-emission process is also known in physics as a scattering process. We know from experiment that a change in the medium leads to a change in ε (μ), thus resulting in a change of the refractive index. The key point arising from this analogy is that a modification of the vacuum produces a change in ε_0 (μ_0) resulting in a subsequent change in c, and hence, a corresponding change in the vacuum refraction index.

Scharnhorst (1990) and Latorre et al. (1995) have since proved that the suppression of light scattering by virtual particle pairs (a.k.a. coherent light-by-light scattering) in the vacuum causes an ***increase*** in the speed of light accompanied by a decrease in the vacuum refraction index. This very unique effect is accomplished in a Casimir Effect capacitor cavity (or waveguide) whereby the vacuum quantum field fluctuations (a.k.a. zero-point fluctuations or ZPF) inside have been modified (becoming anisotropic and non-translational invariant) to satisfy the electromagnetic boundary conditions imposed by the presence of the capacitor plates (or waveguide walls). The principal result of this modification is the removal of the electromagnetic zero-point energy (ZPE) due to the suppression of vacuum ZPE modes with wavelengths longer than the cavity/waveguide cutoff ($\lambda_0 = 2d$, where d = plate separation; see Figure 5). This removal of free space vacuum ZPE modes suppresses the scattering of light by virtual particle pairs, thus producing the speed of light increase (and corresponding decrease in the vacuum refraction index). We know from standard optical physics and quantum electrodynamics (QED) that the optical phase and group velocities can exceed c under certain physical conditions, but dispersion always ensures that the signal velocity is $\leq c$. But recent QED calculations (see, Scharnhorst, 1990 and Latorre et al., 1995) have proved that in the Casimir Effect system, the dispersive effects are *much weaker* still than those associated with the increase in c so that the phase, group and signal velocities will therefore *all increase* by the same amount. Note that, in general, no dispersion shows up in all of the modified vacuum effects examined by investigators.

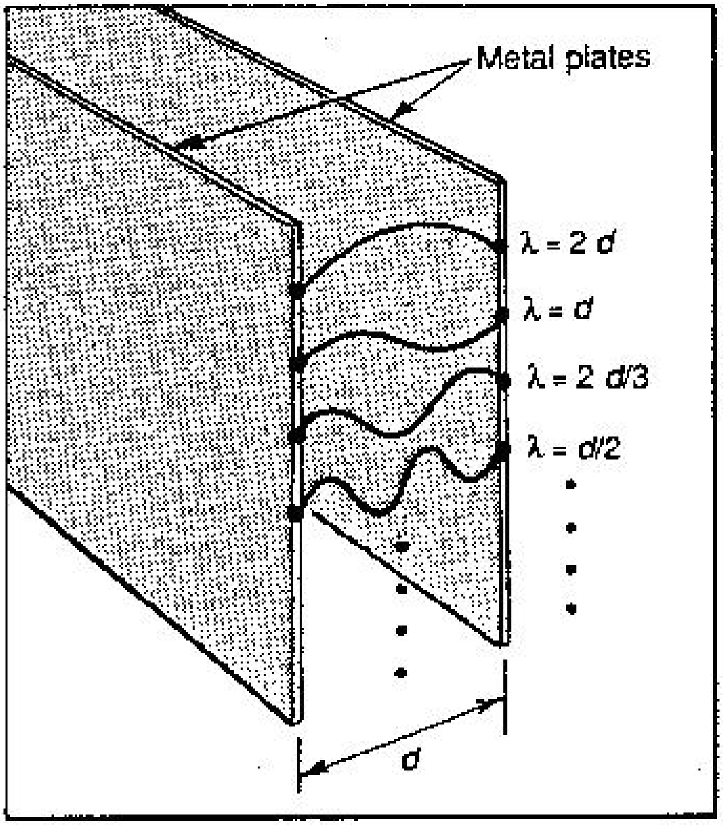
Metal plates
$\lambda = 2d$
$\lambda = d$
$\lambda = 2d/3$
$\lambda = d/2$
d

Casimir effect: the vacuum is full of virtual photons, but photons with wavelengths (λ), more than twice the separation of the plates, are excluded from the space between them. The imbalance pushes the plates together

Figure 5. A Schematic of the Casimir Effect Cavity/Waveguide (from Chown, 1990)

Examples demonstrating the increase in light speed (decrease in vacuum refraction index) via the Casimir Effect vacuum and other modified vacuum effects, as well as those effects producing a decrease in light speed (increase in vacuum refraction index), are described as follows. The vacuum modification effect on the speed of light described in the previous paragraph is (Scharnhorst, 1990):

$$\begin{aligned}\frac{c_\perp{}^*}{c_0} &= \left(1+\frac{11}{2^6\bullet(45)^2}\frac{e^4}{(m_e a)^4}\right) \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \\ &= \left(1+\frac{11\pi^2}{8100}\alpha^2\frac{1}{(m_e a)^4}\right) > 1\end{aligned} \qquad (2.13),$$

where $c_\perp{}^*$ is the (modified) speed of light propagation perpendicular to the Casimir Effect capacitor plates, c_0 is the speed of light in free space (3×10^8 m/s in MKS units), m_e is the electron mass, α is the fine structure constant (≈ 1/137), e is the electron charge ($e^2 = 4\pi\alpha$ in quantum field theory natural units), a is the plate separation, $\hbar$ is Planck's reduced constant, and ε_0 is the vacuum permittivity constant. The condition $\hbar = c_0 = \varepsilon_0 = \mu_0 = 1$ stresses that (2.13), and all the equations that follow, are in quantum field theory natural units. The speed of light and vacuum refraction index measured parallel to the plates is unchanged from their free space values ($c_\| = c_0$, $n_\| = n_0 = 1$). The modified vacuum refraction index measured perpendicular to the plates is (Scharnhorst, 1990):

$$n_\perp = \left(1-\frac{11}{2^6\bullet(45)^2}\frac{e^4}{(m_e a)^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.14).$$

Equations (2.13) and (2.14) show that in general $n_\perp < 1$ and $c_\perp{}^* > c_0$. But $c_\perp{}^* \to c_0$ and $n_\perp \to 1$ when $a \to \infty$ as expected, since we are now allowing all of the vacuum ZPE modes to re-enter the Casimir cavity in this case.

We now survey the additional examples of modified vacuums which increase/decrease light speed (from Latorre et al., 1995):

- ❑ For light (photons) propagating in a Friedmann-Robertson-Walker (FRW) vacuum (i.e., a homogeneous and isotropic Robertson-Walker gravitational background with Friedmann cosmology):

$$\frac{c^*}{c_0} = \left(1+\frac{11}{45}\alpha G\frac{\rho_r + p}{m_e{}^2}\right) > 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.15),$$

where c^* is the modified vacuum speed of light, G is Newton's constant, ρ_r is the energy density and p is the pressure of a radiation-dominated universe ($p = \rho_r/3$). Here the speed of light is increased.

- ❑ For light (photons) propagating in a homogeneous and isotropic thermal vacuum:

$$\frac{c^*}{c_0} = \left(1-\frac{44\pi^2}{2025}\alpha^2\frac{T^4}{m_e{}^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = k_B = 1) \qquad (2.16),$$

where T is the temperature of the vacuum and k_B is the Boltzmann constant. Here the speed of light is decreased.

- ☐ For light (photons) propagating in an anisotropic vacuum given by an external constant uniform magnetic field ***B***:

$$\frac{c_{\|}^{*}}{c_0} = \left(1 - \frac{8}{45}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\sin^2\theta\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1)$$
$$\frac{c_{\perp}^{*}}{c_0} = \left(1 - \frac{14}{45}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\sin^2\theta\right) < 1 \qquad (2.17),$$

where the speed of light is decreased in this vacuum for polarizations coplanar (||) with and perpendicular (⊥) to the plane defined by ***B*** and the direction of propagation, and θ is the angle between ***B*** and the direction of propagation. Latorre et al. (1995) calculated the polarization-average of (2.17) to give the averaged (modified) speed of light in the B-field:

$$\frac{c^{*}}{c_0} = \left(1 - \frac{22}{135}\alpha^2 \frac{\mathbf{B}^2}{m_e^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \quad (2.18).$$

- ☐ For light (photons) propagating in an anisotropic vacuum given by an external constant uniform electric field ***E***, the polarization-averaged modified speed of light is:

$$\frac{c^{*}}{c_0} = \left(1 - \frac{22}{135}\alpha^2 \frac{\mathbf{E}^2}{m_e^4}\right) < 1 \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.19).$$

Here the speed of light is decreased.

Equations (2.16) – (2.19) are the result of vacuum modifications that populate the vacuum with virtual or real particles that induce coherent (light-by-light) scattering, which reduces the speed of massless particles. By examining the form of equations (2.13) and (2.15) – (2.19) Latorre et al. (1995) discovered that the low energy modification of the speed of light is proportional to the ratio of the modified vacuum energy density (as compared to the standard vacuum energy density, $\rho_{vac} = 0$) over m_e^4, with a universal numerical coefficient and the corresponding coupling constants. And a general rule became apparent from their analysis that is applicable to modified vacua for massive and massless quantum field theories, for low energy:

$c^* > c_0$ (vacuum refraction index < 1) when the modified vacuum has a lower energy density
$c^* < c_0$ (vacuum refraction index > 1) when the modified vacuum has a higher energy density
$c^* = c_0$ (vacuum refraction index = 1) when the vacuum is free (or un-modified) with $\rho_{vac} = 0$

The first two rules explain the sign of the change of the speed of light. From this rule and the mathematical commonality between the form of (2.13) and (2.15) – (2.19) Latorre et al. (1995) found a single unifying expression to replace these equations:

$$\frac{c^{*}}{c_0} = 1 - \frac{44}{135}\alpha^2 \frac{\rho}{m_e^4} \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \qquad (2.20),$$

Approved for public release; distribution unlimited.

where ρ is the energy density of the modified vacua under consideration such that $\rho \to \rho_E \sim \boldsymbol{E}^2$ for the electric field vacuum, $\rho \to \rho_B \sim \boldsymbol{B}^2$ for the magnetic field vacuum, and $\rho \to \rho_T \sim \pi^2 T^4$ for the thermal vacuum. If the vacuum is a FRW gravitational vacuum, then one has to substitute one factor of α in (2.20) by $-m_e^2 G$ and $\rho \to \rho_r$. Equation (2.13) for the Casimir Effect vacuum studied earlier is recovered when $\rho \to \rho_{\mathrm{Casimir}} = -(\pi^2/240)a^{-4}$.

Let us recast (2.20) into a more useful form. We subtract one from both sides of (2.20), do some algebra, and thus define the ratio of the change in the speed of light Δc in a modified vacuum to the speed of light in free space c_0:

$$\frac{c^*}{c_0} - 1 = \frac{c^* - c_0}{c_0} \equiv \frac{\Delta c}{c_0}$$

$$\frac{\Delta c}{c_0} = -\frac{44}{135}\alpha^2 \frac{\rho}{m_e^4} \qquad (\hbar = c_0 = \varepsilon_0 = \mu_0 = 1) \quad (2.21).$$

Equations (2.20) and (2.21) are in quantum field theory natural units, which is completely undesirable for estimating physically measurable values of $\Delta c/c_0$. We thus transform or "unwrap" (2.20) and (2.21) back into MKS or CGS units by making the following substitutions (Puthoff, 2003)

$$\rho\ \text{(natural units)} \to \frac{\rho}{\hbar c}\ \text{(MKS or CGS units)}$$

$$m_e\ \text{(natural units)} \to \frac{m_e c}{\hbar}\ \text{(MKS or CGS units)},$$

and after some algebra and rearranging we arrive at the final result:

$$\frac{c^*}{c_0} = 1 - \frac{44}{135}\alpha^2 \frac{\rho}{m_e c_0^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \qquad (2.22)$$

and

$$\frac{\Delta c}{c_0} = -\frac{44}{135}\alpha^2 \frac{\rho}{m_e c_0^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \qquad (2.23),$$

where all quantities are now in MKS or CGS units. We chose the former units so that $c_0 = 3\times10^8$ m/s, $\hbar = 1.055\times10^{-34}$ J-s, $m_e = 9.11\times10^{-31}$ kg, and $\alpha = 1/137$. Note that the ratio of the modified vacuum energy density to the electron rest-mass energy has the dimension of $(volume)^{-1}$ while the quantity in the bracket is the cubed Compton wavelength of the electron having the dimension of (*volume*), and the product of these is dimensionless.

An excellent example for estimating the magnitude of the change in the speed of light (in a modified vacuum) is the Casimir Effect vacuum, since Casimir Effect experiments are common and widespread such that this would be ideal to experimentally test (2.23). We substitute the Casimir vacuum energy density $\rho_{\mathrm{Casimir}} = -(\pi^2\hbar c_0/240)a^{-4}$ (in MKS units) into (2.23), do the algebra, insert the MKS values for the physical constants, and make further simplifications to get:

$$\begin{aligned}\frac{\Delta c}{c_0} &= -\frac{44}{135}\alpha^2\left(-\frac{\pi^2}{240}\frac{\hbar c_0}{a^4}\right)\frac{1}{m_e{c_0}^2}\left(\frac{\hbar}{m_e c_0}\right)^3 \\ &= \frac{11}{8100}\alpha^2\pi^2\left(\frac{\hbar}{m_e c_0 a}\right)^4 \qquad (2.24), \\ &\approx \left(1.59\times10^{-56}\right)a^{-4}\end{aligned}$$

where a (the plate separation) is in meters. Another useful equation is:

$$c^* = \left(1+\frac{\Delta c}{c_0}\right)c_0 \qquad (2.25),$$

where we make the substitution $c^* \to c_\perp^*$ for the present case. H. E. Puthoff and the author (Puthoff, 2003) compared the third line in (2.24) with equation (26) in Scharnhorst (1990) and discovered that the result cited there is in error, because the numerical coefficient is four orders of magnitude too small (Scharnhorst originally pointed out this error to Forward, 1996).

We now set $a = 10^{-6}$ m (1 μm) and we get $\Delta c/c_0 \approx 10^{-32}$ and $c_\perp^* \approx c_0$, which is a horrifically small 1 part in 10^{32} change that we cannot hope to measure at present. But for $a = 10^{-10}$ m (1 Å) we get $\Delta c/c_0 \approx 10^{-16}$ and $c_\perp^* \approx c_0$, which is a 1 part in 10^{16} change that could be measurable at present or in the very near future using high precision laser technology. Last, for $a = 1.1229\times10^{-14}$ m (11.229 fm or ≈ 11 times the nuclear diameter; 1 fm = 10^{-15} m) we find that $\Delta c/c_0 \approx 1$ and $c_\perp^* \approx 2c_0$. We are not able to do technical work at nuclear distances at this time; however, that could change as ultrahigh precision measurement technology continues to evolve. The threshold for the onset of significant changes in light speed occurs when $a < 10^{-12}$ m. This result is generally true for the other modified vacua surveyed in (2.15) – (2.19), since accessible (everyday) values for electric and magnetic field strengths, thermal temperatures and radiation densities are not large enough to overcome the size of the electron mass to create a measurable effect. However, there is a class of ultrahigh intensity tabletop lasers that have achieved such extreme electric and magnetic field strengths and temperatures that it may now be possible to consider using them to explore vacuum modification effects in the lab. We will return to this theme in a later section.

•**Key Point**: As disappointing as the Casimir Effect vacuum (and other modified vacua) results are, it should be strongly pointed out that special relativity theory says that if in one inertial reference frame an object travels only one part in 10^{16} (or even one part in 10^{32}) times faster than c_0, then one can find another reference frame where departure and arrival times of the object are simultaneous, and thus the velocity is ***infinite***. This is what motivates us to look at a teleportation mechanism based on engineering of the vacuum.

•**Technical Notes**:

- Equation (2.15) is interpreted as an increase in the speed of light due to a decrease in the number of vacuum ZPE modes. However, this effect is totally unrelated to light-by-light scattering in the vacuum because the gravitational background "squeezes" (as in squeezed quantum optics states; see Davis, 1999a) the ZPE modes, therefore reducing the vacuum energy density. We further note that the coefficient of 11 is the same for the gravitational vacuum as for the other modified vacua examples based on QED. This factor also appears in the coefficient of the Euler-Poincare characteristic spin-½ contribution to the gravitational trace anomaly (Birrell and Davies, 1982). It is beyond the scope of this study to consider the deep connections between quantum field theory and gravitation.

- We have excluded from our survey the Latorre et al. (1995) results pertaining to all other (high or low energy) modifications of the speed of massless particles. That is because the other examples invoked different QED theories possessing massless ($m_e = 0$), massive and intrinsic mass scales that introduced complex correction terms (beyond the leading low energy terms surveyed above) which are mass-related or running mass-related, and they introduced no new speed modification effects (beyond the low energy electron-positron virtual pair contributions); or no genuine speed modification was possible (especially for the massless Quantum Chromodynamic sector involving pseudo-Goldstone particles).

- There is ongoing (very noisy) controversy within the physics community over the effects of $c^* > c_0$ on causality. As this topic is beyond the scope of this study, I will make three points in this regard: 1) There are no grounds for microcausality violations in accordance with Drummond and Hathrell (1980). 2) A new definition of causality is in order for FTL (faster-than-light) phenomena. 3) Investigators have found that time machines (a.k.a. closed timelike curves) do not affect Gauss's theorem, and thus do not affect the derivation of global conservation laws from differential ones (Friedman et al., 1990). The standard conservation laws remain globally valid while retaining a natural quasi-local interpretation for spacetimes possessing time machines (for example, asymptotically flat wormhole spacetimes). Thorne (1993) states that it may turn out that causality is violated at the macroscopic scale. Even if causality is obeyed macroscopically, then quantum gravity might offer finite probability amplitudes for microscopic spacetime histories possessing time machines. Li and Gott (1998) found a self-consistent vacuum for quantum fields in Misner space (a simple flat space with closed timelike curves), for which the renormalized stress-energy tensor is regular (in fact zero) everywhere. This implies that closed timelike curves could exist at least at the level of semi-classical quantum gravity theory. Therefore, FTL causality paradoxes are just a reflection of our ignorance or inadequate comprehension of the physics of chronology and causality.

In this section we have shown how "vacuum engineering" can modify the speed of light, and how this can, in principle, lead to vm-Teleportation. The vacuum modification concepts summarized above lead us to a formal theory that implements the concept of vacuum engineering within a framework that parallels general relativity theory. This theory is called the Polarizable-Vacuum Representation of General Relativity. In the next section we will introduce and summarize this theory.

2.2.1 The Polarizable-Vacuum Representation of General Relativity

The polarizable-vacuum representation of general relativity (a.k.a. PV-GR) treats the vacuum as a polarizable medium of variable refractive index (Puthoff, 1999a, 2002a, b; Puthoff et al., 2002) exemplifying the concept of the vacuum modification (or vacuum engineering) effects surveyed and discussed in the previous section. The PV-GR approach treats spacetime metric changes in terms of equivalent changes in the vacuum permittivity and permeability constants (ε_0 and μ_0), essentially along the lines of the "$TH\varepsilon\mu$" methodology (see Appendix B for a brief description of this) used in comparative studies of alternative metric theories of gravity (Lightman and Lee, 1973; Will, 1974, 1989, 1993; Haugan and Will, 1977). Such an approach, relying as it does on parameters familiar to engineers, can be considered a "metric engineering" approach. Maxwell's equations in curved space are treated in the isomorphism of a polarizable medium of variable refractive index in flat space (Volkov et al., 1971); the bending of a light ray near a massive body is modeled as due to an induced spatial variation in the refractive index of the vacuum near the body; the reduction in the velocity of light in a gravitational potential is represented by an effective increase in the refractive index of the vacuum, and so forth. This optical-engineering approach has been shown to be quite general (de Felice, 1971; Evans et al., 1996a, b).

As recently elaborated by Puthoff (1999a, 2002a, b; Puthoff et al., 2002) the PV-GR approach, which was first introduced by Wilson (1921) and then developed by Dicke (1957, 1961), can be carried out in a self-consistent way so as to reproduce to appropriate order both the equations of general relativity and the match to the standard astrophysics weak-field experimental (PPN parameters and other) tests of those equations while posing testable modifications for strong-field conditions. It is in application that the PV-GR approach demonstrates its intuitive appeal and provides additional insight into what is meant by a curved spacetime metric.

Specifically, the PV-GR approach treats such measures as the speed of light, the length of rulers (atomic bond lengths), the frequency of clocks, particle masses, and so forth, in terms of a variable vacuum dielectric constant K in which the vacuum permittivity ε_0 transforms as $\varepsilon_0 \rightarrow K\varepsilon_0$ and the vacuum permeability transforms as $\mu_0 \rightarrow K\mu_0$ (see also, Rucker, 1977). In a planetary or solar gravitational potential $K = exp(2GM/rc_0^2) > 1$ (M is a local mass distribution, r is the radial distance from the center of M) while $K = 1$ in "empty" or free asymptotic space (Puthoff, 1999a, 2002a, b; Puthoff et al., 2002). In the former case, the speed of light is reduced, light emitted from an atom is redshifted as compared with a remote static atom (where $K = 1$), clocks run slower, objects/rulers shrink, etc. See Table 1.

Table 1. Metric Effects in the PV-GR Model When *K* > 1 (Compared With Reference Frames at Asymptotic Infinity Where *K* = 1; adapted from Puthoff et al., 2002**)**

Variable	**Determining Equation** (subscript 0 is asymptotic value where $K = 1$)	$K > 1$ *(typical mass distribution, M)*
modified speed of light $c^*(K)$	$c^* = c_0/K$	speed of light < c_0
Modified mass $m(K)$	$m = m_0K^{3/2}$	effective mass increases
modified frequency $\omega(K)$	$\omega = \omega_0K^{-1/2}$	redshift toward lower frequencies
modified time interval $\Delta t(K)$	$\Delta t = \Delta t_0K^{1/2}$	clocks run slower
modified energy $E(K)$	$E = E_0K^{-1/2}$	lower energy states
Modified length $L(K)$	$L = L_0K^{-1/2}$	objects/rulers shrink
dielectric-vacuum "gravitational" forces $F(K)$	$F(K) \propto \nabla K$	attractive gravitational force

When $K = 1$ we have the condition that $c^* = c_0$ (vacuum refraction index = 1), because the vacuum is free (or un-modified, and $\rho_{vac} = 0$) in this case. When $K > 1$, as occurs in a region of space possessing a gravitational potential, then we have the condition that $c^* < c_0$ (vacuum refraction index > 1), because the modified vacuum has a higher energy density in the presence of the local mass distribution that generates the local gravitational field. This fact allows us to make a direct correspondence between the speed of light modification physics discussion in Section 2.2 and the underlying basis for the physics of the PV-GR model. Under certain conditions the spacetime metric can in principle be modified to reduce the value of K to below unity, thus allowing for faster-than-light (FTL) motion to be physically realized. In this case, the local speed of light (as measured by remote static observers) is increased, light emitted from an atom is blueshifted as compared with a remote static atom, objects/rulers expand, clocks run faster, etc. See Table 2. We therefore have the condition that $c^* > c_0$ (vacuum refraction index < 1) because the modified vacuum has a lower energy density. In fact, Puthoff (1999a, 2002a) has analyzed certain special black hole metrics and found $K < 1$ from the model. We will return to this theme later. In what follows we briefly review and summarize the key points and equations from the development of the PV-GR model, and we refer the reader to Puthoff (1999a, 2002a, b) for more extensive discussion and derivations.

Table 2. Metric Effects in the PV-GR Model When $K < 1$ (Compared With Reference Frames at Asymptotic Infinity Where $K = 1$; adapted from Puthoff et al., 2002**)**

Variable	**Determining Equation** (subscript 0 is asymptotic value where $K = 1$)	$K < 1$ *(typical mass distribution, M)*
modified speed of light $c^*(K)$	$c^* = c_0/K$	speed of light > c_0
modified mass $m(K)$	$m = m_0K^{3/2}$	effective mass decreases
modified frequency $\omega(K)$	$\omega = \omega_0K^{-1/2}$	blueshift toward higher frequencies
modified time interval $\Delta t(K)$	$\Delta t = \Delta t_0K^{1/2}$	clocks run faster
modified energy $E(K)$	$E = E_0K^{-1/2}$	higher energy states
modified length $L(K)$	$L = L_0K^{-1/2}$	objects/rulers expand
dielectric-vacuum "gravitational" forces $F(K)$	$F(K) \propto \nabla K$	repulsive gravitational force

We begin by recalling that in flat space electrodynamics, the electric flux vector ***D*** in a linear, homogeneous medium can be written

$$\begin{aligned}\mathbf{D} &= \varepsilon\mathbf{E} \\ &= \varepsilon_0\mathbf{E} + \mathbf{P} \\ &= \varepsilon_0\mathbf{E} + \alpha_V\mathbf{E}\end{aligned} \qquad (2.26),$$

where ε is the permittivity of the medium, the polarization ***P*** corresponds to the induced dipole moment per unit volume in the medium whose polarizability per unit volume is α_V, and ***E*** is the electric field. The identical form of the last two terms naturally leads to the interpretation of ε_0 as the polarizability per unit volume of the vacuum. The quantum picture of the vacuum, where it has been shown that the vacuum acts as a polarizable medium by virtue of induced dipole moments resulting from the excitation of virtual electron-positron particle pairs (Heitler, 1954), completely justifies the interpretation that the vacuum is a medium. Note that there are other virtual particle pairs in the vacuum that also contribute to this picture; however, it is the electron-positron pairs that dominate the others, as shown in Section 2.2. The basic postulate of the PV-GR model for curved space conditions is that the polarizability of the vacuum in the vicinity of localized mass-energy distributions differs from its asymptotic free space value by virtue of vacuum polarization effects induced by the presence of the local mass-energy. Thus the postulate for the vacuum itself is

$$\begin{aligned}\mathbf{D} &= \varepsilon\mathbf{E} \\ &\equiv K\varepsilon_0\mathbf{E}\end{aligned} \qquad (2.27),$$

where K (a function of position) is the modified dielectric constant of the vacuum due to the induced vacuum polarizability changes under consideration. Equation (2.27) defines the transformation $\varepsilon = K\varepsilon_0$.

Table 1 shows the various quantitative effects a polarizable vacuum (in the presence of positive mass-energy distributions) has on the various measurement processes important to general relativity. The effects demonstrated in the middle and right columns demonstrate the basis of the polarizable vacuum approach to general relativity. Table 2 shows what effects are manifested when negative mass-energy distributions induce vacuum polarizability changes that lead to FTL phenomenon. Experimental observations impose constraints on the model causing key physical constants to remain constant even with variable polarizability present in the local space. Puthoff (1999a, 2002a, b) has shown that the fine structure constant is constrained by observational data to remain constant within a variable polarizable vacuum, and this constraint actually defines the transformation $\mu = K\mu_0$. The elementary particle charge e is also taken to be constant in a variable polarizable vacuum because of charge conservation. And $\hbar$ remains a constant by conservation of angular momentum for circularly polarized photons propagating through the (variable polarizability) vacuum. The remaining constant of nature is the speed of light, and although the tables showed how this was modified in variable polarizability vacuums, it is interesting to see how this modification comes about. In a modified (variable polarizability) vacuum the speed of light is defined, as it is in standard electrodynamics, in terms of the permittivity and permeability by:

$$\begin{aligned} c^* &\equiv (\varepsilon\mu)^{-1/2} \\ &= (K\varepsilon_0 \bullet K\mu_0)^{-1/2} \\ &= (K^2\varepsilon_0\mu_0)^{-1/2} \\ &= \frac{1}{K}(\varepsilon_0\mu_0)^{-1/2} \\ &= \frac{c_0}{K} \end{aligned} \qquad (2.28),$$

where the permittivity/permeability transformations and the free space (un-modified vacuum) definition for c_0 were inserted. Note that (2.28) can be re-written as $c^*/c_0 = 1/K$, and this is to be compared with (2.22). Thus we see from (2.28), and by comparison with (2.22), that K plays the role of a variable refractive index under conditions in which the vacuum polarizability is assumed to change in response to general relativistic-type influences. One further note of interest is that the permittivity/permeability transformations also maintains constant the ratio

$$\sqrt{\frac{\mu}{\varepsilon}} = \sqrt{\frac{\mu_0}{\varepsilon_0}} ,$$

which is the impedance of free space. This constant ratio is required to keep electric-to-magnetic energy ratios constant during adiabatic movement of atoms from one position in space to another of differing vacuum polarizability (Dicke, 1957, 1961). And this constant ratio is also a necessary condition in the $TH\varepsilon\mu$ formalism for an electromagnetic test particle to fall in a gravitational field with a composition-independent acceleration (Lightman and Lee, 1973; Will, 1974, 1989, 1993; Haugan and Will, 1977).

Now we make the "crossover connection" to the standard spacetime metric tensor concept that characterizes conventional general relativity theory, as originally shown by Puthoff (1999a, 2002a, b). In flat (un-modified or free) space the standard four-dimensional infinitesimal spacetime interval ds^2 is given (in Cartesian coordinates with subscript 0) by

$$ds^2 = -c_0^{\ 2} dt_0^{\ 2} + \sum_{i=1}^{3} dx_{i0}^{\ 2} \qquad (2.29),$$

where $i \equiv (1 = x,\ 2 = y,\ 3 = z)$. This metric means that measuring rods and clocks are non-varying wherever one goes in spacetime to make measurements. However, this has been shown to be incorrect in general relativity theory, so the length and time transformations (between proper and coordinate values) given in the tables (middle columns) indicate that measuring rods and clocks do vary when placed in regions where $K \neq 1$. Therefore, we replace the time and space differentials in (2.29) with the length and time transformations in the tables into (2.29), and derive the general relativistic spacetime interval

$$ds^2 = -\frac{1}{K} c_0^{\ 2} dt^2 + K\left(\sum_{i=1}^{3} dx_i^{\ 2}\right) \qquad (2.30).$$

Note that observers within a $K \neq 1$ region will always measure the speed of light to be c_0. Equation (2.30) defines an isotropic coordinate system, which is a common and useful way to represent spacetime metrics in general relativity studies. By inspection the metric tensor is written

$$g_{\alpha\beta} = \begin{pmatrix} -1/K & 0 & 0 & 0 \\ 0 & K & 0 & 0 \\ 0 & 0 & K & 0 \\ 0 & 0 & 0 & K \end{pmatrix} \qquad (2.31).$$

The Lagrangian density for matter-field interactions in a vacuum of variable K is given by Puthoff (1999a, 2002a, b) as

$$\begin{aligned} L_d = &-\left[\frac{m_0 c_0^{\ 2}}{\sqrt{K}}\sqrt{1-\left(\frac{v}{c_0/K}\right)^2} + q\Phi - qA_i v^i\right]\delta^3(\mathbf{r}-\mathbf{r}_0) \\ &-\frac{1}{2}\left(\frac{\mathbf{B}^2}{K\mu_0} - K\varepsilon_0\mathbf{E}^2\right) - \frac{c_0^{\ 4}}{32\pi G}\frac{1}{K^2}\left[(\nabla K)^2 - \frac{1}{(c_0/K)^2}\left(\frac{\partial K}{\partial t}\right)^2\right] \end{aligned} \qquad (2.32),$$

where the first term is the Lagrangian density for a free particle of mass m_0, charge q and 3-vector velocity $\boldsymbol{v}$ ($v = |\boldsymbol{v}|$, 3-vector components are labeled by i) interacting with electromagnetic fields via the electromagnetic field 4-vector potential $\boldsymbol{A}_\mu = (\Phi, A_i)$ (note that $\delta^3(\mathbf{r} - \mathbf{r}_0)$ is the delta function that locates the point particle at position $\mathbf{r} = \mathbf{r}_0$); the second term is the Lagrangian density for the electromagnetic fields themselves, and the last term is the Lagrangian density for K (treated here as a scalar variable). This last term emulates the Lagrangian density for the gravitational field. Equation (2.32) does not include any quantum gauge field interaction terms because it is beyond the scope of the present incarnation of the PV-GR approach to include them. We can obtain the equations of particle motion in a variable dielectric vacuum by performing the standard variations of the Lagrangian density $\delta(\int L_d\, dx\, dy\, dz\, dt)$ with respect to the particle variables. However, we are more interested in obtaining the "master equation" for K by varying the Lagrangian density with respect to K, and Puthoff (1999a, 2002a, b) gives the result:

$$\nabla^2\sqrt{K} - \frac{1}{\left(c_0/K\right)^2}\frac{\partial^2\sqrt{K}}{\partial t^2}$$

$$= -\frac{8\pi G}{c_0^{\ 4}}\sqrt{K}\left\{\frac{\left(m_0 c_0^{\ 2}/\sqrt{K}\right)}{\sqrt{1-\left(\dfrac{v}{c_0/K}\right)^2}}\,\frac{1}{2}\left[1+\left(\frac{v}{c_0/K}\right)^2\right]\delta^3(\mathbf{r}-\mathbf{r}_0)\right. \qquad (2.33).$$

$$\left. +\frac{1}{2}\left(\frac{\mathbf{B}^2}{K\mu_0}+K\varepsilon_0\mathbf{E}^2\right) - \frac{c_0^{\ 4}}{32\pi G}\frac{1}{K^2}\left[(\nabla K)^2+\frac{1}{\left(c_0/K\right)^2}\left(\frac{\partial K}{\partial t}\right)^2\right]\right\}$$

This equation describes the generation of general relativistic vacuum polarization effects due to the presence of matter and fields. By inspecting the right-hand side of the equation, we observe that changes in K are driven by the mass density (1st term), electromagnetic energy density (2nd term), and the vacuum polarization energy density itself (3rd term). In fact, the 3rd term emulates the gravitational field self-energy density. Note that the 2nd and 3rd terms in (2.33) appear with opposite signs with the result that electromagnetic field effects can counteract the gravitational field effects. Puthoff found that (2.33) gives the solution $K = exp(2GM/rc_0^{\ 2})$ in the vicinity of a static spherically symmetric (uncharged) mass M (in the low velocity limit $v << c_0$, $\partial K/\partial t = 0$, $\boldsymbol{E} = \boldsymbol{B} = 0$, $q = 0$), which reproduces to appropriate order the standard general relativistic Schwarzschild spacetime metric for the weak gravitational field conditions prevailing in the solar system. This solution guarantees that $K > 1$ near mass concentrations.

Of major importance to the present study are solutions giving $K < 1$ so that teleportation can be realized. Puthoff has found one such solution by studying the case of a static spherically symmetric mass M with charge Q familiar from the study of the Reissner-Nordstrφm spacetime metric. In this case Puthoff found the result

$$K = \left[\cos\left(\frac{\sqrt{b^2-a^2}}{r}\right) + \frac{a}{\sqrt{b^2-a^2}}\sin\left(\frac{\sqrt{b^2-a^2}}{r}\right)\right]^2 \qquad \left(b^2 > a^2\right) \qquad (2.34),$$

where $a^2 = (GM/c_0^{\ 2})^2$, $b^2 = Q^2G/4\pi\varepsilon_0 c_0^{\ 4}$, and r is the radial distance from the center of M. And in this case (2.34) gives $K < 1$, which shows that FTL solutions are available in the PV-GR approach (as they are also in the Einstein theory). (For $a^2 > b^2$ the solution is hyperbolic-trigonometric and describes the standard Reissner-Nordstrφm metric where $K > 1$.)

Generally speaking, in Einstein general relativity the Reissner-Nordstrφm metric can be manipulated along with two shells of electrically charged matter to form a traversable wormhole (Schein and Aichelburg, 1996). But there are two drawbacks to this. The first is that the scheme involves dealing with the collapsed state of the stellar matter that generates the metric (a.k.a. Reissner-Nordstrφm black hole) along with the unpleasant side effects that are encountered, such as the crushing singularities and multiple (unstable) event horizons. Second, the traversable wormhole is an eternal time machine connecting remote regions of the same universe together. Now there are no black hole solutions found in the PV-GR model because in that approach stellar matter collapses smoothly to an ultra-dense state and without the creation of singularities and event horizons (Puthoff, 1999b).

In either case, the Reissner-Nordstrφm metric does not offer a viable mechanism for vm-Teleportation. We are more interested in examining other PV-GR cases (where $K < 1$ or even $K << 1$)

that emulate the effects of traversable wormhole metrics that do obey the vm-Teleportation definition, such as the example presented in Section 2.1. Equation (2.33) suggests that we search for a vacuum engineering concept that exploits electromagnetic fields to alter the vacuum dielectric constant *K* to induce the desired vm-Teleportation effect in the modified vacuum. (However, we can insert other source terms that will lead to the desired result.) We envision this particular teleportation concept to resemble Figure 2. [Note: Before this report went to press H. E. Puthoff, C. Maccone and the author discovered a number of $K < 1$ solutions to equation (2.33) that uniquely meet the definition of vm-Teleportation and FTL motion. We discovered that the generic energy density required to generate $K < 1$ solutions must be negative, and that the total energy density of the system as seen by remote observes is approximately zero. This unique result compares very well with the traversable wormhole mass-energy density requirements discussed in Section 2.1.2. This discovery will be the subject of a forthcoming paper.]

2.3 Conclusion and Recommendations

The concept we envision for vm-Teleportation is that animate or inanimate objects would be placed inside an environmentally enclosed vessel that would simply be moved into the teleportation device. The "teleporter" would be activated, and the vessel would almost immediately disappear and then reappear at the remote destination as if it were briefly moving through a portal or "stargate." The teleportation device might be required to operate in the vacuum of space outside of the Earth's atmosphere. We have shown two practically equivalent ways to implement vm-Teleportation. There is the manipulation of spacetime geometry via exploiting negative (i.e., quantum vacuum zero point) energy as shown by Einstein's general relativity theory, and there is the modification of the vacuum dielectric constant as shown by the PV-GR model. Both have a great deal of theoretical foundation to begin exploring experimentally. The PV-GR model needs additional theoretical work for the present application, but it is now mature enough for experimental exploration.

There already is extensive theoretical, and more importantly, experimental research proving that the vacuum can be engineered (or physically modified) so that the vacuum ZPE can be exploited (via the Casimir Effect, for example) to extract electrical energy or actuate microelectromechanical devices (see for example, Ambjϕrn and Wolfram, 1983; Forward, 1984, 1996, 1998; Puthoff, 1990, 1993; Cole and Puthoff, 1993; Milonni, 1994; Mead and Nachamkin, 1996; Lamoreaux, 1997; Chan et al., 2001, and the references cited therein). But most of this research involves very low energy density regimes, which are much too low for our purposes. The Mead and Nachamkin (1996) device is actually designed to extract electrical energy from the higher frequency/higher energy density ZPE modes. However, new ultrahigh-intensity lasers became available in the 1990s that have achieved extreme physical conditions in the lab that are comparable to the extreme astrophysical conditions expected to be found in stellar cores and on black hole event horizons (Perry, 1996; Mourou et al., 1998; Perry, 2000). The power intensity of these lasers has reached the point to where they actually probe QED vacuum physics and general relativistic physics, and they have even modified the vacuum itself. The lasers were originally called petaWatt lasers (operating range of $10^{14} - 10^{18}$ Watts/cm^2 at femtosecond pulses), but they have now reached power intensity levels in the $10^{25} - 10^{30}$ Watts/cm^2 range. The lasers were made possible by a novel breakthrough called "chirped pulse amplification" whereby the initial low energy/low power intensity laser beam is stretched, amplified and then compressed without experiencing any beam distortions or amplifier damage. This laser system was initially designed as a large-optics beam-line power booster for the NOVA laser fusion experiment at Lawrence Livermore National Laboratory. But researchers found a way to shrink the optics down to tabletop scale, and one can now own and operate a tabletop ultrahigh-intensity laser for ≈ \$500,000. The dimensions of the optical bench used by the University of California-San Diego is ≈ 5 m $\times$ 12 m (or ≈ 60 m^2; see Mourou et al., 1998). In tabletop lab experiments ultrahigh-intensity lasers have generated >> gigagauss magnetic fields, $\geq 10^{16}$ Volt/cm electric field strengths, >> terabar light pressures and $>> 10^{22}$ m/sec^2 subatomic particle accelerations. These ultrahigh-intensity tabletop lasers are thus the ideal instrument with which to explore the fundamental physics underlying the two possible concepts for vm-Teleportation.

There are several ideas on how to generate negative energy in the lab that could potentially be extracted and concentrated in the proper fashion to induce the traversable flat-face wormhole outlined in Section 2.1.1 or induce the $K < 1$ condition (in the PV-GR model) outlined in Section 2.2.1. The schemes for generating negative energy are:

- Casimir Effect (described in Section 2.2): This is the easiest and most well known way to generate negative energy in the lab. The energy density $\rho_{\mathrm{Casimir}} = -(\pi^2 \hbar c_0/240)a^{-4}$ within a Casimir capacitor cavity is negative and manifests itself by producing a force of attraction between the capacitor plates. This has been measured in the lab (see above references). Forward (1998) proposes a mechanism for the endless extraction of energy from the vacuum in a Casimir cavity by cyclic manipulation of the cavity dimensions.

- Moving Mirror: Negative quantum vacuum energy can be created by a single moving reflecting surface (a moving mirror). If a mirror moves with increasing acceleration, then a flux of negative energy emanates from its surface and flows out into the space ahead of the mirror (Birrell and Davies, 1982). However, this effect is known to be exceedingly small, and it is not the most effective way to generate negative energy.

- Optically Squeezed Laser Light: Negative quantum vacuum energy can also be generated by an array of ultrahigh intensity lasers with an ultrafast rotating mirror system. In this scheme a laser beam is passed through an optical cavity resonator made of lithium niobate crystal that is shaped like a cylinder with rounded silvered ends to reflect light. The resonator will act to produce a secondary lower frequency light beam in which the pattern of photons is rearranged into pairs. This is the quantum optical "squeezing" of light effect. (See Section A.2 in Appendix A for a complete definition and description of squeezed quantum states.) Therefore, the squeezed light beam emerging from the resonator will contain pulses of negative energy interspersed with pulses of positive energy. Another way to squeeze light would be to manufacture extremely reliable light pulses containing precisely one, two, three, etc. photons apiece and combine them together to create squeezed states to order. Superimposing many such states could theoretically produce bursts of intense negative energy. For the laser beam resonator example we find that both negative and positive energy pulses are of $\approx 10^{-15}$ second duration. We could arrange a set of rapidly rotating mirrors to separate the positive and negative energy pulses from each other. The light beam is to strike each mirror surface at a very shallow angle while the rotation ensures that the negative energy pulses are reflected at a slightly different angle from the positive energy pulses. A small spatial separation of the two different energy pulses will occur at some distance from the rotating mirror. Another system of mirrors will be needed to redirect the negative energy pulses to an isolated location and concentrate them there.

- Gravitationally Squeezed Vacuum Energy: A natural source of negative quantum vacuum energy comes from the effect that gravitational fields (of astronomical bodies) in space have upon the surrounding vacuum. For example, the gravitational field of the Earth produces a zone of negative energy around it by dragging some of the virtual particle pairs (a.k.a. virtual photons or vacuum ZPF) downward. This concept was initially developed in the 1970s as a byproduct of studies on quantum field theory in curved space (Birrell and Davies, 1982). However, Hochberg and Kephart (1991) derived an important application of this concept to the problem of creating and stabilizing traversable wormholes, and their work was corrected and extended by Davis (1999a). They proved that one can utilize the negative vacuum energy densities, which arise from *distortion of the electromagnetic zero point fluctuations* due to the interaction with a prescribed gravitational background, for providing a violation of the energy conditions (see

Section A.1 in Appendix A). Hochberg and Kephart (1991) showed that the squeezed quantum states of quantum optics provide a natural form of matter having negative energy density. And since the vacuum is defined to have vanishing energy density, anything possessing less energy density than the vacuum must have a negative energy density. The analysis, via quantum optics, shows that gravitation itself provides the mechanism for generating the squeezed vacuum states needed to support stable traversable wormholes. The production of negative energy densities via a squeezed vacuum is a necessary and unavoidable consequence of the interaction or coupling between ordinary matter and gravity, and this defines what is meant by gravitationally squeezed vacuum states. The magnitude of the gravitational squeezing of the vacuum can be estimated from the squeezing condition, which simply states that substantial gravitational squeezing of the vacuum occurs for those quantum electromagnetic field modes with *wavelength* (λ in meters) > *Schwarzschild radius* (r_S in meters) of the mass in question (whose gravitational field is squeezing the vacuum). The Schwarzschild radius is the critical radius, according to general relativity theory, at which a spherically symmetric massive body becomes a black hole; i.e., at which light is unable to escape from the body's surface. We can actually choose any radial distance from the mass in question to perform this analysis, but using the Schwarzschild radius makes equations simpler in form. The general result of the gravitational squeezing effect is that as the gravitational field strength increases the negative energy zone (surrounding the mass) also increases in strength. Table 3 shows when gravitational squeezing becomes important for example masses. The table shows that in the case of the Earth, Jupiter and the Sun, this squeeze effect is extremely feeble because only ZPF mode wavelengths above 0.2 m – 78 km are affected. For a solar mass black hole (radius of 2.95 km), the effect is still feeble because only ZPF mode wavelengths above 78 km are affected. But note from the table that quantum black holes with Planck mass will have enormously strong negative energy surrounding them because all ZPF mode wavelengths above 8.50×10^{-34} meter will be squeezed; in other words, all wavelengths of interest for vacuum fluctuations. Black holes with proton mass will have the strongest negative energy zone in comparison because the squeezing effect includes all ZPF mode wavelengths above 6.50×10^{-53} meter. Furthermore, a black hole smaller than a nuclear diameter ($\approx 10^{-16}$ m) and containing the mass of a mountain ($\approx 10^{11}$ kg) would possess a fairly strong negative energy zone because all ZPF mode wavelengths above 10^{-15} meter will be squeezed.

Table 3. Substantial Gravitational Squeezing Occurs When $\lambda \geq 8\pi r_S$ (For Electromagnetic ZPF; adapted from Davis, 1999a**)**

Mass of body	Schwarzschild radius of body, r_S	ZPF mode wavelength, λ
Sun = 2.0×10^{30} kg	2.95 km	≥ 78 km
Jupiter = 1.9×10^{27} kg	2.82 m	≥ 74 m
Earth = 5.976×10^{24} kg	8.87×10^{-3} m	≥ 0.23 m
Typical mountain ≈ 10^{11} kg	≈ 10^{-16} m	≥ 10^{-15} m
Planck mass = 2.18×10^{-8} kg	3.23×10^{-35} m	≥ 8.50×10^{-34} m
Proton = 1.673×10^{-27} kg	2.48×10^{-54} m	≥ 6.50×10^{-53} m

•**Recommendations**:

- Theoretical Program 1: A one to two year theoretical study (cost ≈ \$80,000) should be initiated to explore the recently discovered $K < 1$ (FTL) solutions to equation (2.33) in order to define, characterize and model the negative energy density source(s) that induce the FTL vacuum modification. The study should also identify potential lab experiments designed to test theoretical predictions.

- Theoretical Program 2: A one to two year study (cost ≈ \$80,000) should be initiated to conduct a detailed review of the negative energy generation schemes summarized above to define their characteristics, performances and requirements. The study should develop technical parameters for each of the schemes in order to identify potential lab experiments.

- Experimental Program 1: An experimental study should be conducted to test Forward's (1998) Casimir energy extraction proposal. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 2: An experimental study using ultrahigh-intensity lasers should be conducted to test the Optically Squeezed Laser Light proposal. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 3: An experimental study using ultrahigh-intensity lasers should be conducted to probe QED vacuum physics and vacuum modification as well as test elements of the PV-GR model. A starting point for this program would be to use such lasers to perform the Ding and Kaplan (1989, 1992, 2000; see also, Forward, 1996) experiment. This is an important fundamental physics experiment to do, because it can distinguish between the rival quantum vacuum electromagnetic ZPE fluctuation and fluctuating charged particle source field theory models, which would settle the acrimonious debate over whether the vacuum really fluctuates or not. R. L. Forward (1999) told the author that a Nobel Prize rides on performing this experiment and settling the issue once and for all. The Ding and Kaplan proposal is already designed to probe QED vacuum physics and vacuum modification. [The essence of the Ding and Kaplan proposal is to demonstrate that a form of photon-photon scattering predicted by QED gives rise to 2nd-harmonic generation of intense laser radiation in a DC magnetic field due to the broken symmetry of interaction (in the Feynman "box" diagram approximation). This effect is possible only when the field system (optical wave + DC field) is inhomogeneous, in particular when a Gaussian laser beam propagates in either a homogeneous or inhomogeneous DC magnetic field. In other words, a vacuum region is filled with a DC magnetic field that polarizes the virtual particle pairs (a.k.a. virtual photons) in the vacuum. This polarized vacuum then scatters incident ultrahigh-intensity laser photons of frequency ν (energy E), thereby generating outgoing photons of frequency 2ν (energy $2E$).] An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

- Experimental Program 4: An experimental study using ultrahigh-intensity lasers should be conducted to establish the extreme physical conditions necessary to test the strong-field limit of general relativity with an emphasis on generating spacetime curvature and negative energy in order to induce a putative micro-wormhole. (Experimental Programs 3 and 4 could be done together to determine whether Puthoff's PV-GR theory or Einstein's general relativity theory is the correct model for nature.) A Nobel Prize is in the offing if this question were to be addressed and settled. An experiment definition study will be required to estimate the experimental method, procedure, equipment needs and costs.

3.0 q-TELEPORTATION

3.1 Teleportation Scenario

Future space explorers and their equipment will need to easily and quickly travel from an orbiting spacecraft to the surface of some remote planet in order to get their work done, or military personnel in the United States need to easily and quickly travel from their military base to another remote location on Earth in order to participate in a military operation, or space colonists will need quick transport to get from Earth to their new home planet. Instead of using conventional transportation to expedite travel the space explorer, military personnel or space colonist and/or their equipment go into the "Teleporter" (a.k.a. "Transporter" in Star Trek lingo) and are "beamed down" or "beamed over" to their destinations at light speed. The mechanism for this teleportation process is hypothetically envisioned to be the following:

1. Animate/inanimate objects placed inside the teleporter are scanned by a computer-generated and -controlled beam.

2. The scan beam encodes the entire quantum information contained within the animate/inanimate object(s) into organized bits of information, thus forming a digital pattern of the object(s).

3. The scan beam then dematerializes the object(s) and stores its pattern in a pattern buffer, thus transforming the atomic constituents of the dematerialized object(s) into a matter stream. Alternative 1: The dematerialization process converts the atoms into a beam of pure energy. Alternative 2: The scan beam does not dematerialize the object(s).

4. The teleporter then transmits the matter/pure energy stream and quantum information signal in the form of an annular confinement beam to its destination. Alternative: Only the quantum information signal is transmitted.

5. At the receiving teleporter the matter/pure energy stream is sent into a pattern buffer whereby it is recombined with its quantum information, and the object(s) is rematerialized back into its original form. Alternative 1: The receiving teleporter recombines the transmitted quantum information with atoms stored inside a reservoir to form a copy of the original. Alternative 2: The quantum information is reorganized in such a way as to display the object on some three-dimensional (holographic) visual display system.

Problem: This generic scenario is modeled after teleportation schemes found in SciFi. There are a lot of important little details that were left out of the teleportation process because we simply do not know what they are. This technology does not yet exist. And we are left with the question of which one of the alternative processes identified in items 3 – 5 one wants to choose from. The above scenario is only an outline, and it is by no means complete since it merely serves to show what speculation exists on the subject. The above scenario describes a speculative form of what we call *q-Teleportation*.

There are questions to be addressed in the above scenario. Does the teleporter transmit the atoms and the quantum bit information signal that comprises the animate/inanimate object or just the quantum bit information signal? There are $\approx 10^{28}$ atoms of matter combined together in a complex pattern to form a human being. How does one transmit this much information and how do we disassemble that many atoms? Computer information gurus would insist that it is not the atoms that matter but only the bits of information representing them when considering the transmission of large "bodies" of information. But are humans simply the sum of all the atoms (and the related excited atom quantum states) that comprise them? We could possibly learn to reconstitute a beam of atoms into a chemically accurate human being. However, would this also include the reconstruction of a person's consciousness (personality, memories, hopes, dreams, etc.) and soul or spirit? This question is beyond the scope of this study to address, but it is nevertheless one of the most important concepts awaiting a complete scientific understanding.

For the teleporter to process and transmit the quantum bit information signal that encodes the animate/inanimate object's pattern will require stupendous digital computer power. For each atom comprising the object we must encode its location in space (three position coordinates), its linear and angular momentum (three vector components for each quantity), and its internal quantum state (electron orbital-energy levels and their excitation/de-excitation and ionization states, binding to other atoms to form molecules, molecular vibrational/rotational states, bound nuclei states, spin states for electrons and nuclei, etc.), etc. If we assume that we can digitally encode all of this information for a single atom with a minimum of one kilobyte (1 byte = 8 bits, 1 bit ≡ 0 or 1) of data, then we will require a minimum of 10^{28} kilobytes to encode and store an entire human being (in three-dimensions). To digitally store and access this much information at present (and for the foreseeable future) is nontrivial. It will take more than 2,400 times the present age of the universe (≈ 13 billion years) to access this amount of data using commercially available computers (operating at ≈ 10 gigabyte/sec). Top-of-the-line supercomputers will not reduce this time significantly. The computer technology needed to handle such a large data storage requirement simply does not exist. The largest commercially available computers can store ≈ 40 gigabytes on a single hard drive. We will need ≈ 10^{20} of these hard drives to store the encoded information of just one human being. Also, wire and coaxial/fiber optic cables do not have the physical capacity to transmit this amount of data between devices. These numbers will not be significantly different for macroscopic inanimate objects. The information processing and transfer technology required for the teleportation system may become possible in 200 – 300 years if improvements in computer storage and speed maintains a factor of 10 – 100 increase for every decade. There is speculation that emergent molecular, bio-molecular (DNA-based systems) and quantum computer technology may achieve the performances required for a teleportation system. In the former case molecular dynamics mimics computer logic processes and the ≈ 10^{25} particles in a macroscopic sample will all act simultaneously, making for far greater digital information processing and transfer speeds. Researchers have given no formal performance estimates for this emergent technology. In the latter case quantum computing would take advantage of entangled quantum states of subatomic matter or photons, whereby digital logic processes would occur at light speed. This technology is in its infancy, and there has been no clear direction on what performance levels will be possible in the future. This topic will be discussed further in Section 3.2.3.

In the above teleportation scenario we might consider dematerializing animate/inanimate objects into a matter stream consisting of only the object's constituent atoms or atomic subcomponents (protons, neutrons and electrons) and transmitting them at the speed of light (or close to it). To push atoms or subatomic particles to near the speed of light will require imparting to them an energy comparable to their rest-mass energy, which will be at a minimum of one order of magnitude larger than the amount of energy required to break protons up into free quarks. The energy required to completely dematerialize (or dissolve) matter into its basic quantum constituents or into pure energy is alone stupendous. At first one will have to impart to every molecule within the object an energy that is equivalent to the binding energy between atoms (atomic binding energy ~ chemical energy ~ *several* eV) in order to break apart the molecules comprising the object's macro-structure. After this an energy equivalent to nuclear binding energies (≈ *several* × 10^6 times atomic binding energy, or ≈ *several* MeV) must be imparted to every free atomic nucleus inside the object in order to break apart the protons and neutrons residing within each nucleus. And last, an energy equivalent to the binding energy that holds together the three quarks residing within each proton and neutron must be imparted to each of the free protons and neutrons within the object. According to the Standard Model and experimental data, the quark binding energy is practically infinite. But all is not lost, because the Standard Model also predicts that if we could heat up the nuclei to ≈ 10^{13} °C (≈ 10^6 times hotter than the core temperature of the Sun, or ≈ 10^3 MeV), then the quarks inside would suddenly lose their binding energies and become massless (along with other elementary matter). This heat is also equivalent to the rest-mass energy of protons and neutrons. Therefore, to heat up and dematerialize one human being would require the annihilation of the rest mass-energy of all 10^{28} protons-neutrons or the energy equivalent of 330 1-megaton thermonuclear bombs. Compare this stupendous explosive energy with the explosive yield of the largest thermonuclear bomb ever detonated on Earth, which was a 50-megaton bomb that was built by Andrei Sakharov in the USSR and detonated on October 30, 1961; it was called "Tzar Bomba." Its first incarnation (ca. early October 1961) comprised a uranium fusion tamper, which gave an estimated explosive yield of ≈ 100 megatons. But the weapon was too heavy (27 metric tons) for a bomber to carry, so the tamper was replaced by one made of lead, which reduced both the weight and the yield. In the end we see that it is not a trivial problem to simply heat up and dematerialize any human or inanimate objects. The technology to do so does not exist unless we invoke new physics to get around the energy requirement.

Finally, we must consider the resolution and aperture of the optics required to scan and transmit the animate/inanimate object's matter (or energy) stream. The Heisenberg quantum uncertainty principle fundamentally constrains the measurement resolution of conjugate observable quantities, such as position and momentum or energy and time. The measurement of any combination of (conjugate) observables with arbitrarily high precision is not possible, because a high precision measurement of one observable leads to imprecise knowledge of the value of the conjugate observable. The quantum uncertainty principle makes it impossible to measure the exact, total quantum state of any object with certainty. The scan resolution of a teleportation system is defined by the wavelength of light used to illuminate the object's atomic/subatomic constituents and record their configurations. To resolve matter at atomic/subatomic distance scales requires that the energy of the scanner light (photons) be extremely large (according to the uncertainty principle); and during the scan this large light energy will be conveyed to the constituents, causing them to drastically change their speed and direction of motion. This means that it is physically impossible to resolve an object's atomic/subatomic particle components and their configurations with the precision necessary to accurately encode and later recreate the object being teleported. To resolve atomic/subatomic particles requires wavelengths smaller than the size of these constituents, which will typically be 1 Å – 1 fm. Such wavelengths are in the gamma ray part of the spectrum, and this becomes a major technical problem for us because at present there is no gamma ray electro-optics with which to work with. Now consider the example of teleporting an object from the surface of a planet back to its spacecraft in orbit some *several* × $10^2 - 10^3$ km away. The optical aperture required to illuminate and scan an object with ≈ 1 Å – 1 fm resolution from orbit will be >> *several* × $10^2 - 10^3$ km. If we are to consider teleporting an object from planet to planet or from star to star then the aperture required will be >> *several* × $10^8 - 10^{13}$ km. These technical problems are truly insurmountable unless totally new physics becomes available.

3.2 Quantum Teleportation

It turns out that there does in fact exist a form of teleportation that occurs in nature despite the numerous technical roadblocks described in the previous section. It is called *quantum teleportation*, which is based on the well-known concept of quantum entanglement. Erwin Schrödinger coined the word "entanglement" in 1935 in a three-part paper (Schrödinger, 1935a, b, c, 1980). These papers were prompted by the Einstein, Podolsky and Rosen (1935; denoted hereafter as EPR) paper that raised fundamental questions about quantum mechanics, whereby Einstein had loudly complained that quantum mechanics allowed physical processes resembling "spooky action at a distance" to occur. EPR recognized that quantum theory allows certain correlations to exist between two physically distant parts of a quantum system. Such correlations make it possible to predict the result of a measurement on one part of a system by looking at the distant part. On this basis, EPR argued that the distant predicted quantity should have a definite value even before being measured, if quantum theory is complete and respects locality (a.k.a. causality). EPR concluded that, from a classical perspective, quantum theory must be incomplete because it disallows such definite values prior to measurement. Schrödinger's perspective on this argument gives the modern view of quantum mechanics, which is to say that the wavefunction (a.k.a. quantum state vector) provides all the information there is about a quantum system. In regards to the nature of entangled quantum states, Schrödinger (1935a, b, c, 1980) stated that, "The whole is in a definite state, the parts taken individually are not." This statement defines the essence of pure-state entanglement. Schrödinger went on to give a description of quantum entanglement by introducing his famous cat experiment.

To better understand the concept of quantum entanglement/teleportation we will focus on the quantum wavefunction (a.k.a. quantum state function). Any quantum system such as a particle that possesses a position in space, energy, angular and linear momentum, and spin is completely described by a wavefunction. This is usually symbolized in a variety of ways, and we choose to represent a generic wavefunction using the traditional "bra-ket" notation of quantum mechanics: $|\varphi\rangle$. Anything that we want to know about the particle is mathematically encoded within $|\varphi\rangle$. As we discussed in the previous section the wavefunction can never be completely known because there is no measurement that can determine it completely. The only exception to this is in the special case that the wavefunction has been prepared in some particular state or some member of a known basis group of states in advance. By measuring one of the properties of a quantum system, we can get a glimpse of the overall quantum state that is encoded within $|\varphi\rangle$. According to the quantum uncertainty principle the act of doing such a measurement will destroy any ability to subsequently determine the other properties of the quantum system. So the act of measuring a particle actually destroys some of the information about its pristine state. This makes it impossible to copy particles and reproduce them elsewhere via quantum teleportation. However, it turns out that one can recreate an unmeasured quantum state in another particle as long as one is prepared to sacrifice the original particle. The trick is to exploit the EPR process to circumvent the quantum uncertainty principle.

As discussed previously, EPR discovered that a pair of spatially separated quantum sub-systems that are parts of an overall quantum system can be "entangled" in a non-local (i.e., non-causal) way. When two particles come into contact with one another, they can become "entangled." In an entangled state, both particles remain part of the same quantum system so that whatever you do to one of them affects the other one in a predictable fashion. More precisely, a measurement on one of the entangled sub-systems puts it into a particular quantum state, while instantaneously putting the sub-system with which it is entangled into a corresponding quantum state, while the two sub-systems are separated by arbitrarily large distances in spacetime (even backwards in time!). A simple example of this phenomenon is to prepare a pair of photons in the same quantum state such that they are entangled, and then allow them to fly apart to remote locations without any form of communication occurring between them along their journey. Measuring the polarization of one of the pair of entangled photons induces the other photon, which may be light-years away, into the same state of polarization as that which was measured for its entangled twin. The basic operation of quantum teleportation can be described as determining the total quantum state of some large quantum system, transmitting this state information from one place to another, and making a perfect reconstruction of the system at the new location. In principle, entangled particles can serve as "transporters" of sorts. By introducing a third "message" particle to one of the entangled particles, one could transfer its properties to the other one, without ever measuring those properties.

Historically, quantum entanglement was never reconciled with the quantum uncertainty principle and the requirement of locality (or causality) in observed physical phenomena, thus it became a paradox in quantum theory. A three-decade debate began following the appearance of the EPR paper over whether quantum entanglement (a.k.a. "spooky action at a distance") was a real quantum phenomenon or not, and this debate came to be called the "EPR dilemma." Einstein's only solution to the dilemma was to suggest that quantum mechanics was incomplete and needed a reformulation to incorporate local hidden-variables that can account for observed physical phenomena without violating causality. Bell (1964) later solved the EPR dilemma by deriving correlation inequalities that can be violated in quantum mechanics but have to be satisfied within every model that is local and complete. Such models are called “local hidden-variable models.” Bell showed that a pair of entangled particles, which were once in contact but later moved too far apart to interact directly (i.e., causally), can exhibit individually random behavior that is too strongly correlated to be explained by classical statistics. Bell’s inequalities make it possible to test whether local hidden-variable models can account for observed physical phenomena in lab experiments. Groundbreaking experimental work by Aspect et al. (1982a, b) along with further theoretical and experimental work done by others (Freedman and Clauser, 1972; Aspect, 1983; Aspect and Grangier, 1985; Hong and Mandel, 1985; Bennett and Wiesner, 1992; Tittel et al., 1998a, b; Tittel and Weihs, 2001) demonstrated violations of the Bell inequalities, which therefore invalidated the local hidden-variable models. The key result of recent theoretical and experimental work is that an observed violation of a Bell inequality demonstrates the presence of entanglement in a quantum system.

3.2.1 Description of the q-Teleportation Process

The experimental work of Bennett et al. (1993) followed by the theoretical and experimental work of others (Vaidman, 1994; Kwiat et al., 1995; Braunstein, 1996; Braunstein and Kimble, 1998; Pan et al., 1998; Stenholm and Bardroff, 1998; Zubairy, 1998; Vaidman and Yoran, 1999; Kwiat et al., 1999) made the breakthrough that was necessary to demonstrate the principle of quantum teleportation in practice. It was a remarkable technical breakthrough that settled, once and for all, the nagging question of whether quantum entanglement could be used to implement a teleportation process to transfer information between remotely distant quantum systems non-causally (i.e., at FTL speed). It is easy to describe how quantum teleportation works in greater detail. Figure 6 compares conventional facsimile transmission with the quantum teleportation process seen in Figure 7. In a conventional facsimile transmission the original document is scanned, extracting partial information about it, but it remains more or less intact after the scanning process. The scanned information is then sent to the receiving station, where it is imprinted on new paper to produce an approximate copy of the original. In quantum teleportation (Figure 7) one scans out part of the information from object A (the original), which one wants to teleport, while causing the remaining, unscanned, part of the information in A to pass, via EPR entanglement, into another object C which has never been in contact with A. Two objects B and C are prepared and brought into contact (i.e., entangled), and then separated. Object B is taken to the sending station, while object C is taken to the receiving station. At the sending station object B is scanned together with the original object A, yielding some information and totally disrupting the states of A and B. This scanned information is sent to the receiving station, where it is used to select one of several treatments to be applied to object C, thereby putting C into an exact replica of the former state of A. Object A itself is no longer in its original initial state, having been completely disrupted by the scanning process. The process just described is teleportation and not replication, and one should not confuse the two. There is a subtle, unscannable kind of information that, unlike ordinary information or material, can be delivered via EPR correlations/entanglement, such that it cannot by itself deliver a meaningful and controllable message. But quantum teleportation delivers exactly that part of the information in an object that is too delicate to be scanned out and delivered by conventional methods.

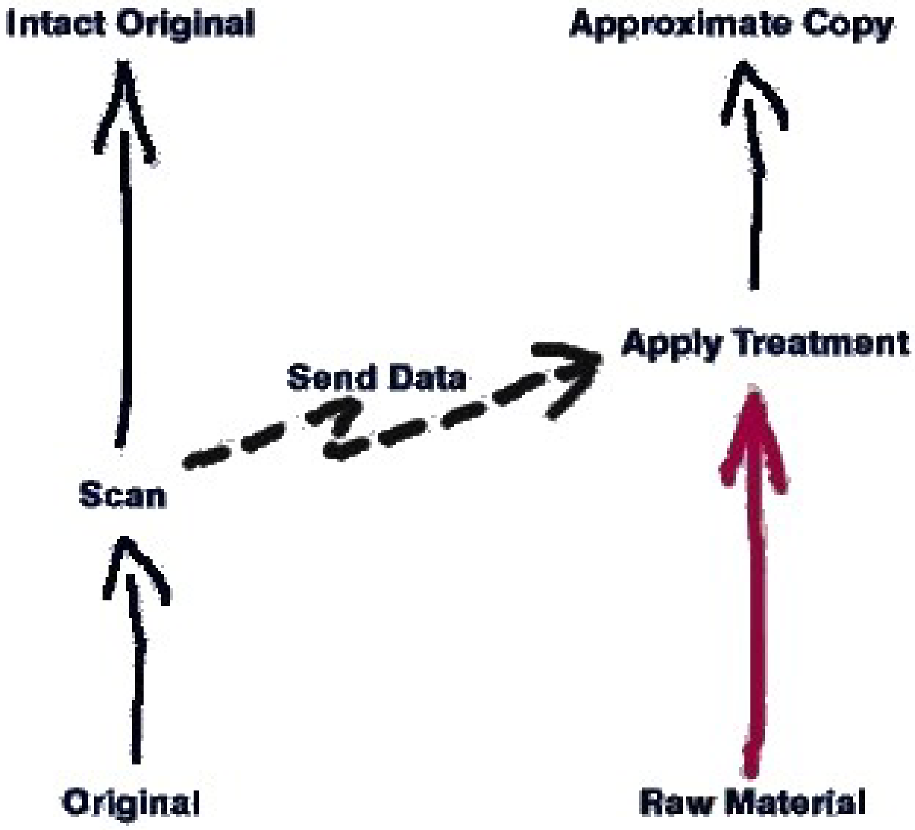
Intact Original
Approximate Copy
Apply Treatment
Send Data
Scan
Original
Raw Material

Figure 6. Classical Facsimile Transmission (Modified IBM Press Image)

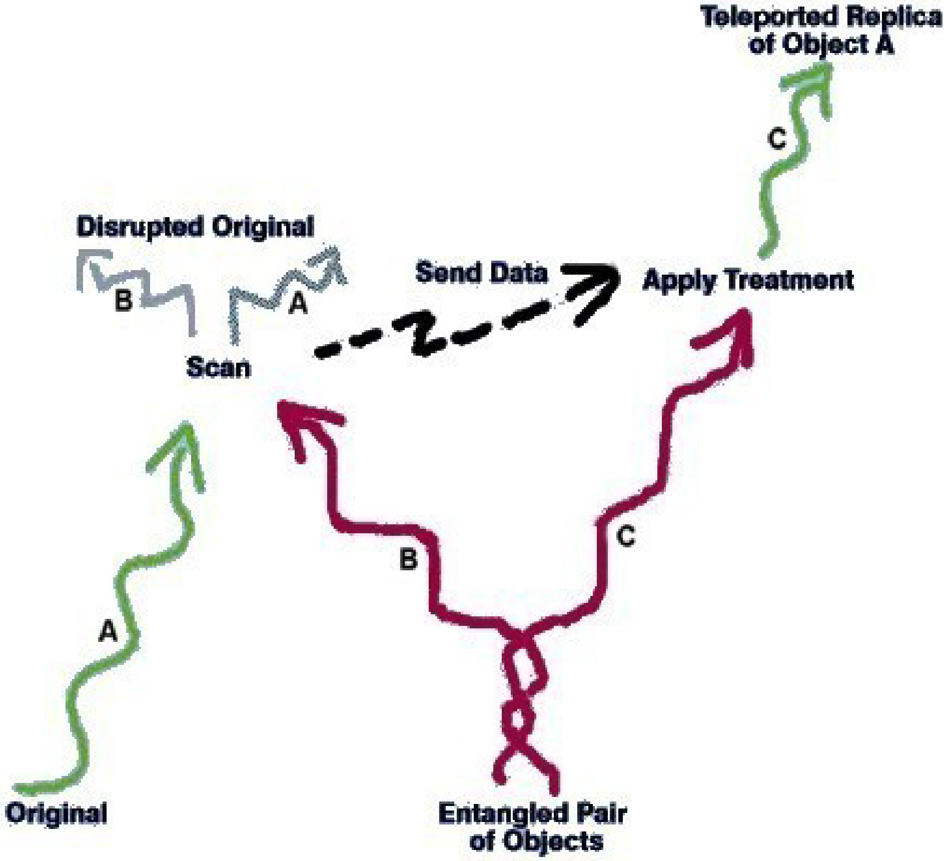
Teleported Replica of Object A
C
Disrupted Original
B
A
Send Data
Apply Treatment
Scan
B
C
A
Original
Entangled Pair of Objects

Figure 7. Quantum Teleportation (Modified IBM Press Image)

Approved for public release; distribution unlimited.

We now go one more final step to give a simplified outline of the actual teleportation process according to Bennett et al. (1993). They propose a multistep procedure by which any quantum state $|\chi\rangle$ of a particle or a photon (that correspond to an N-state system) is to be teleported from one location to another. For example, $|\chi\rangle$ might be a two-level system that could refer to the polarization of a single photon, the nuclear magnetic spin of a hydrogen atom, or the electronic excitation of an effective two-level atom. The following scenario outlines the q-Teleportation process in a very simplified way:

1. Prepare a pair of quantum subsystems $|\varphi\rangle$ and $|\psi\rangle$ in an EPR entangled state so that they are linked together. $|\varphi\rangle$ and $|\psi\rangle$ are maximally entangled and together constitute a definite pure state superposition even though each of them is maximally undetermined or mixed when considered separately.

2. Transport $|\varphi\rangle$ to the location of the teleportation transmitter and transport $|\psi\rangle$ to the location of the teleportation receiver. (In the technical literature the transmitter is called "Alice" and the receiver is called "Bob.") The transmitter and receiver can be many light years apart in space. Note that the two subsystems are non-causally correlated via entanglement, but they contain no information about $|\chi\rangle$ at this point. The two subsystems represent an open quantum channel that is ready to transmit information.

3. Now Alice brings the teleported state $|\chi\rangle$ into contact with the entangled state $|\varphi\rangle$ and performs a quantum measurement on the combined system $|\chi\rangle|\varphi\rangle$. Bob and Alice have previously agreed upon the details of the quantum measurement.

4. Using a conventional classical communication channel, Alice transmits to Bob a complete description of the outcome of the quantum measurement she performed on $|\chi\rangle|\varphi\rangle$.

5. Bob then subjects $|\psi\rangle$ to a set of linear transformations (i.e., suitable unitary rotations) that are dictated by the outcome of Alice's quantum measurement. The quantum subsystem Bob originally first received is no longer in state $|\psi\rangle$ after the linear transformations because it is now in a state identical to the original state $|\chi\rangle$. Therefore, $|\chi\rangle$ has in effect been teleported from Alice to Bob.

Bennett et al. (1993) showed in their experimental work that this scheme requires both a conventional communication channel and a non-causal EPR channel to send the state $|\chi\rangle$ from one location to another. In addition to this, a considerable pre-arrangement of entangled states and quantum measurement procedures is required to make the process work. Bennett et al. (1993) analyzed the information flow implicit in the process and showed that Alice's measurement does not provide any information about the quantum state $|\chi\rangle$. All of the quantum state information is passed by the EPR link between the entangled particle states $|\varphi\rangle$ and $|\psi\rangle$. We can think of the measurement results as providing the "code key" that permits the EPR information to be decoded properly at Bob's end. And because the measurement information must travel on a conventional communications channel, the decoding cannot take place until the code key arrives, insuring that no FTL teleportation is possible.

The q-Teleportation scheme teleports the state of a quantum system without having to completely measure its initial state. The outcome of the process is that the initial quantum state $|\chi\rangle$ is destroyed at Alice's location and recreated at Bob's location. It is very important for the reader to understand that it is the *quantum states* of the particles/photons that are destroyed and recreated in the teleportation process, and not the particles/photons themselves. The quantum state or wavefunction contains the information on the state of a particle, but is not a directly observable physical quantity like mass-energy. The quantum information contained within a state is available in the form of probabilities or expectation values.

Therefore, q-Teleportation *cannot* teleport animate or inanimate matter (or energy) in its physical entirety. However, some experts argue that because a particle's or a photon's quantum state is its defining characteristic, teleporting its quantum state is completely *equivalent* to teleporting the particle/photon even though the original particle's/photon's quantum state (and defining characteristic) was completely destroyed in the process (more on this in Section 3.3). Therefore, no quantum cloning is possible and we are left with a (near-perfect) copy of the now-destroyed original after teleportation (Wootters and Zurek, 1982; Barnum et al., 1996). And finally, classical information itself *cannot* be teleported faster than the speed of light via the non-causal EPR channel; however, quantum information can (more on this in Section 3.2.3).

3.2.2 Decoherence Fundamentally Limits q-Teleportation

Finally, the reader must understand that the q-Teleportation scenario described in the previous section was simplified because we unrealistically assumed that Alice and Bob shared an EPR entangled pair that was free of noise or decoherence. Decoherence is the process, whereby an object's quantum states degrade when information leaks to or from the environment (i.e., environmental noise) through stray interactions with the object. In reality, Alice and Bob have quantum systems that interact directly or through another mediating quantum system like two ions in an ion trap that interact through phonon modes of the trap, or Rydberg atoms in a laser cavity that interact via photons (Sackett, 2001; Raimond et al., 2001). Decoherence degrades the fidelity of the quantum link (i.e., the set of pure EPR entangled pairs) between two quantum systems, thus introducing a certain level of error in the exchange of quantum information between the systems.

In a real-world example of an application of q-Teleportation to quantum computation (discussed in the next section), we can devise an array of interconnected ion traps with each trap holding a small number of ions that are coupled by ions that are moved between the traps or by traveling photons (Wineland et al., 2002). The quantum link (or EPR interaction) between a pair of systems is subject to noise or decoherence through photon loss or heating of the phonons. At present, decoherence imposes a fundamental limit on our ability to perform quantum information processing. Research is continuing on whether decoherence can be reduced, circumvented, or otherwise be (partially or totally) eliminated. Dür and Briegel (2003) have taken the first step towards this goal at rudimentary level by showing that fault-tolerant quantum computation can be achieved in the presence of very high noise levels occurring in the interaction link between small quantum systems, if one assumes that local quantum processing on each end is nearly error free. They showed that the interaction link can have an error rate of two-thirds.

3.2.3 Recent Developments in Entanglement and q-Teleportation Physics

Quantum teleportation physics is still in its infancy. Both theoretical and experimental developments are advancing in many different directions, but are far from maturity at this point in time because the field is still evolving at present. Technical applications of entanglement and q-Teleportation are just becoming conceptualized for the first time, while a small number of basic physics breakthroughs and their related applications are in experimental progress at present. The research community is still in the process of discovering the full nature of entanglement and q-Teleportation, its rules, and what roadblocks nature has in store for its applications and further progression. The literature cited in this study is by no means complete, and only represents a subset of the entire field, because the research is still evolving.

An important application of quantum entanglement and q-Teleportation was the discovery made by Shor (1994, 1997) that computation with quantum states instead of classical bits can result in large savings in computation time. For example, the best algorithms take exponentially more resources to factor ever-larger numbers on a classical computer. A 500-digit number needs 10^8 times as many computational steps to factor as a 250-digit number. The latter classically requires $\approx 5\times10^{24}$ computational steps, or about 150,000 years computing time at terahertz speed, to factor. Shor found a polynomial-time quantum algorithm that solves the problem of finding prime factors of a large integer.

He showed that his algorithm rises only polynomially so that a 500-digit number takes only eight times as many computational steps to factor as a 250-digit number. And by using the quantum factoring algorithm, a 250-digit number requires only $\approx 5\times10^{10}$ steps or < 1 second to factor at terahertz speed, so that a 500-digit number will take ≤ 1 second to factor. No classical polynomial-time algorithm for this problem exists at present. This breakthrough generated a cottage industry of research into quantum computing and quantum information theory.

IBM (2001) constructed a prototype quantum computer that uses the nuclear spins of seven atoms that are part of a large molecule with the iron-based chemical composition $H_5C_{11}O_2F_5Fe$. The computer uses entangled nuclear spins for storage and has a capacity of seven qubits (qubits are defined in the bulleted list in the next two paragraphs below). All of the Fluorine atoms in the large molecule are Fluorine isotope 19 and two of the Carbon atoms are Carbon isotope 13. All the other non-hydrogen atoms have even isotope numbers and no nuclear spins. The objective of the prototype quantum computer was to factor the number 15 into its two prime factors 3 and 5 by using Shor's quantum factoring algorithm. The quantum computation required that a sample of $\approx 10^{18}$ of the large molecules be placed in a magnetic field and manipulated by nuclear magnetic resonance (NMR) techniques. This mechanism allows the spins to function as qubits, whereby Schor's algorithm can be performed via manipulation of the NMR fields. NMR was used to implement quantum computing in this prototype, because the nuclear spins are well isolated from decoherence as a result of the very long decoherence time (the time after which quantum coherence is lost due to environmental noise) in the system.

To factor larger numbers will require a system that uses more than seven qubits. It is estimated that a quantum computer using ≈ 36 qubits could very quickly perform computations that would require a conventional computer ≈ 13 billion years to perform. And such a computer could solve one of the technical problems of human teleportation discussed in Section 3.1. However, a scale-up in the number of qubits is difficult because the IBM prototype has reached the technology limit of NMR quantum computing. The prototype's operation requires that all of the qubits must be in the same molecule. And molecules with more than seven spins that can be used as qubits are not feasible at present. However, there are alternative technologies for quantum computing that show promise for scaling up the number of qubits. The technologies of nuclear spin orientation of single atom impurities in semiconductors, electron spin orientation in quantum dots, and the manipulation of magnetic flux quanta in superconductors all show promise of providing a basis for scalable quantum computers. Finally, the primary technical problem in quantum computing at the present time is decoherence, and this must be eliminated or otherwise mitigated before new quantum technology can become competitive with conventional computer technology.

A byproduct of the recent quantum computing and information research is that a modern theory of entanglement has emerged. Researchers now treat entanglement as a quantifiable physical resource that enables quantum information processing and computation. Entanglement is no longer treated as a paradox of quantum theory. It has been recently discovered that (Nielsen and Chuang, 2000; Nielsen, 2003; Terhal et al., 2003):

- various kinds of pure and mixed entangled states may be prepared in addition to the simple pure-state superpositions that was described in the previous section

- the members of an entangled group of objects do not have their own individual quantum states, only the group as a whole has a well-defined state (i.e., "the whole is greater than the sum of its parts")

- entangled objects behave as if they were physically connected together no matter how far apart they actually are, distance does not attenuate entanglement in the slightest – it has been demonstrated that information can be teleported over 40 km using existing technology (H. Everitt, Army Research Office, 2000)

- if something is entangled with other objects, then a measurement of it simultaneously provides information about its partners

- some quantum systems can have a little entanglement while others will have a lot

- the more entanglement available, the better suited a system is to quantum information processing

- decoherence degrades the fidelity of the quantum link (i.e., the set of pure EPR entangled pairs) between two quantum systems, thus introducing a certain level of error in the exchange of quantum information between the systems; thus limiting our ability to perform quantum information processing (see more on this issue in the next paragraph below)

- mixed entangled states may be measured, distilled, concentrated, diluted, and manipulated

- the basic resource of classical information is the bit (i.e., the two values 0 and 1), while quantum information comes in quantum bits (i.e., qubits) that are described by their quantum state; qubits can exist in superpositions that simultaneously involve 0 and 1, thus giving them an infinite range of values; groups of qubits can be entangled; qubits must be insulated against decoherence, so that the coherent state of the quantum system in a quantum computer is preserved for a time that is long enough to set up a calculation, perform it, and read out the results

- quantum computers processing qubits or entangled qubits can outperform classical computers; functional requirements of quantum computers:

    - they must have the ability to initialize any qubit in a specified state, and to measure the state of a specific qubit
    - they must have universal quantum gates, which are logical elements capable of arranging any desired logical relationship between the states of qubits
    - they must also have a processor capable of interlinking quantum gates to establish rules and boundary conditions for their inter-relationships – in a quantum computation, the arrangement of quantum gates connects the qubits in a logical pattern, according to a program or algorithm, and after an interval the qubits assigned to the result are read out

- quantum error correction codes exist, whereby qubits are passed through a circuit (the quantum analogue of logic gates) that will successfully fix an error in any one of the qubits without actually reading what all the individual qubit states are; no qubit cloning is required

- a completely secure quantum key can be generated and distributed (for communication and decoding of encrypted messages) using entangled photons has been demonstrated (Tittel et al., 2000; Jennewein et al., 2000; Naik et al., 2000); any eavesdropper's attempt to intercept the quantum key will alter the contents in a detectable way, enabling users to discard the compromised parts of the data

- in an experiment which verified that EPR entanglement obeys Special Relativity (Seife, 2000; Scarani et al., 2000; Gisin et al., 2000; Zbinden et al., 2000a, b), and involving a photon detector moving at relativistic speeds (for example, Bob moves away from Alice at close to the speed of light), investigators determined that quantum information via EPR photon pair entanglement must travel > 10^7 times light speed (the photon detectors were 10.6 km apart)

- investigators are still developing quantitative laws of entanglement to provide a set of principles for understanding the behavior of entanglement and how it is used to do information processing

- investigators are working to develop an understanding of the general principles that govern complex quantum systems such as quantum computers

Other developments are equally as interesting or compelling. For example, the quantum state of the object we wish to teleport does not have to describe single microscopic systems like photons, ions, atoms or electrons. Quantum states can describe large collections of atoms like chemical compounds, humans, planets, stars, and galaxies. Hartle and Hawking (1983) even derived the quantum wavefunction of the Universe in closed form, although, it was extremely simplified and excluded the presence of quantum matter-energy. So it has become possible to consider teleporting large quantum systems. We summarize the more recent spectacular developments in the following:

- Generation of entanglement and teleportation by Parametric Down-Conversion (Bouwmeester et al., 1997; Zeilinger, 2003): EPR entangled photon pairs are created when a laser beam passes through a nonlinear β-barium borate or BBO crystal. Inside the crystal (BBO, for example) an ultraviolet photon (λ = 490 nm) may spontaneously split into two lower energy infrared photons (λ = 780 nm), which is called parametric down-conversion. The two "down-conversion" photons emerge as independent beams with orthogonal polarizations (horizontal or vertical). (The orthogonal polarization states represent a classic example of the discrete quantum state variables that can be teleported. Other examples of discrete quantum variables that have been teleported using other schemes include the nuclear magnetic spin of a hydrogen atom, electronic excitations of an effective two-level atom, elementary particle spins, etc.) In the two beams along the intersections of their emission cones, we observe a polarization-entangled two-photon state. For the experimental realization of quantum teleportation, it is necessary to use pulsed down-conversion. Only if the pulse width of the UV light, and thus the time of generating photon pairs is shorter than the coherence time of the down-converted photons, then interferometric Bell-state analysis can be performed. In this type of experiment, the pulses from a mode-locked Ti:Saphire laser have been frequency doubled to give pulses of ≈ 200 fs duration (1 fs = 10^{-15} second). The interfering light is observed after passage through IR filters of 4 nm bandwidth giving a coherence time of ≈ 520 fs. After retroflection during its second passage through the crystal, the UV pulse creates another pair of photons. One of these will be the teleported photon, which can be prepared to have any polarization. Beam splitters and photon detectors are used to perform the Bell-state analysis during the standard teleportation process that ensues. See Figure 8 for a schematic showing the layout of a standard parametric down-conversion entanglement-teleportation experiment.

- Teleportation of squeezed states of light and continuous quantum state variables (Furusawa et al., 1998; Sørensen, 1998; Braunstein and Kimble, 1998; Opatrný et al., 2000; Braunstein et al., 2001; Zhang et al., 2002; Bowen et al., 2002; Bowen et al., 2003; Zeilinger, 2003): Squeezed light (see Section A.2 in Appendix A) is used to generate the EPR entangled beams, which are sent to Alice and Bob. A third beam, the input, is a coherent state of unknown complex amplitude. This state is teleported to Bob with a high fidelity only achievable via the use of quantum entanglement. Entangled EPR beams are generated by combining two beams of squeezed light at a 50/50 beam splitter. EPR beam 1 propagates to Alice's sending station, where it is combined at a 50/50 beam splitter with the unknown input state, in this case a coherent state of unknown complex amplitude. Alice uses two sets of balanced homodyne detectors to make a Bell-state measurement on the amplitudes of the combined state. Because of the entanglement between the EPR beams, Alice's detection collapses Bob's field (EPR beam 2) into a state conditioned on Alice's measurement outcome. After receiving the classical result from Alice, Bob is able to construct the teleported state via a simple phase-space displacement of the EPR field 2. Quantum teleportation in this scheme is theoretically perfect, yielding an output state which equals the input with a fidelity $F = 1$. In practice, fidelities less than one are realized due to imperfections in the EPR pair, Alice's Bell measurement, and Bob's unitary transformation. By contrast, a sender and receiver who share only a classical communication channel cannot hope to transfer an arbitrary quantum state with a fidelity of one. For coherent states, the classical teleportation limit is $F = 0.5$, while for light polarization states it is $F = 0.67$. The quantum nature of the teleportation achieved in this case is demonstrated by the experimentally determined fidelity of $F = 0.58$, greater than the classical limit of 0.5 for coherent states. The fidelity is an average over all input states and so measures the ability to transfer an arbitrary, unknown superposition from Alice to Bob. This technique achieves the teleportation of continuous quantum state variables, as opposed to the discrete quantum state variables used in the Bennett et al. (1993) teleportation protocol and its variants. The teleportation of a squeezed state of light from one beam of light to another demonstrates the teleportation of a continuous feature (of light) that comes from the superpositions of an infinite number of basic states of the electromagnetic field, such as those found in squeezed states. This line of research also involves the experimental demonstration of the mapping of quantum states from photonic to atomic media via entanglement and teleportation. Hald et al. (1999) reported on the experimental observation of a spin-squeezed macroscopic ensemble of 10^7 cold atoms, whereby the ensemble is generated via quantum state entanglement/teleportation from non-classical light to atoms.

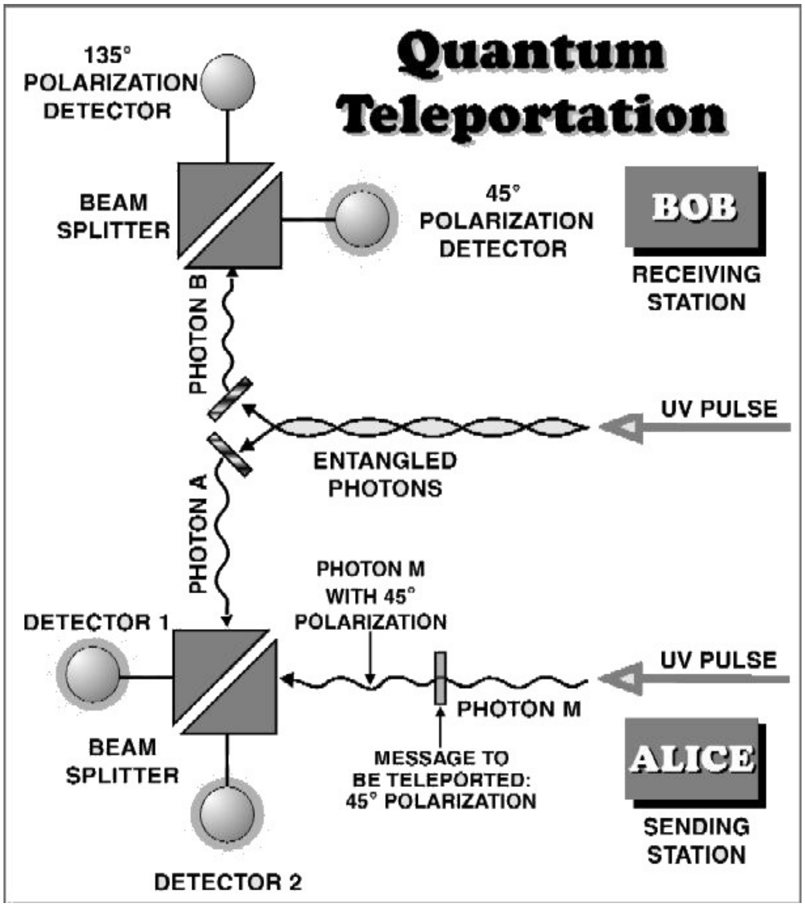
Quantum Teleportation
135° POLARIZATION DETECTOR
BEAM SPLITTER
45° POLARIZATION DETECTOR
BOB
RECEIVING STATION
PHOTON B
UV PULSE
ENTANGLED PHOTONS
PHOTON A
PHOTON M WITH 45° POLARIZATION
DETECTOR 1
UV PULSE
PHOTON M
BEAM SPLITTER
MESSAGE TO BE TELEPORTED: 45° POLARIZATION
ALICE
SENDING STATION
DETECTOR 2

Figure 8. Quantum Teleportation (From www.aip.org)

At the sending station of the quantum teleporter, Alice encodes a "messenger" photon (M) with a specific state: 45 degrees polarization. This travels towards a beam splitter. Meanwhile, two additional entangled photons (A and B) are created. The polarization of each photon is in a fuzzy, undetermined state, yet the two photons have a precisely defined interrelationship. Specifically, they must have complementary polarizations. For example, if photon A is later measured to have horizontal (0 degrees) polarization, then the other photon must collapse into the complementary state of vertical (90 degrees) polarization. Entangled photon A arrives at the beam splitter at the same time as the message photon M. The beam splitter causes each photon to either continue toward detector 1 or change course and travel to detector 2. In 25% of all cases, in which the two photons go off into different detectors, Alice does not know which photon went to which detector. This inability for Alice to distinguish between the two photons causes quantum weirdness to kick in. Just by the very fact that the two photons are now indistinguishable, the M photon loses its original identity and becomes entangled with A. The polarization value for each photon is now indeterminate, but since they travel toward different detectors Alice knows that the two photons must have complementary polarizations. Since message photon M must have complementary polarization to photon A, then the other entangled photon (B) must now attain the same polarization value as M. Therefore, teleportation is successful. Indeed, Bob sees that the polarization value of photon B is 45 degrees: the initial value of the message photon.

Approved for public release; distribution unlimited.

- Entanglement of Atoms (Hagley et al., 1997; Sackett et al., 2000): EPR entanglement at the level of atoms has been experimentally demonstrated using rubidium atoms prepared in circular Rydberg states (i.e., the outer electrons of the atom have been excited to very high energy states and are far from the nucleus in circular orbits). The experimental apparatus produces two entangled atoms, one atom in a ground state and the other atom in an excited state, physically separated so that the entanglement is non-local. And when a measurement is made on one atom, let us say the atom in a ground state, then the other atom instantaneously presents itself in the excited state – the result of the second atom wave function collapse, thus determined by the result of the first atom wave function collapse. This work is now evolving towards the demonstration of entanglement for molecules and larger entities followed by teleportation of their states. Bose and Home (2002) have improved on this concept by proposing a single, simple generic method by which any atoms, ions and macroscopic objects can be entangled and teleported.

- Teleportation of an Atomic State via Cavity Decay (Bose et al., 1999; Sackett et al., 2000): It has been shown how the state of an atom trapped in a cavity can be teleported to a second atom trapped in a distant cavity simply by detecting photon decays from the cavities.

- Biological Quantum Teleportation (Mavromatos et al., 2002): There are several obstacles to teleporting large complicated objects, especially biological entities. Decoherence is the primary obstacle. That is because observable quantum effects in biological matter is thought to be strongly suppressed due to the macroscopic nature of most biological entities and the fact that such systems live at near room temperature, and there is always contact between biological entities and the environment (the source of decoherence). These conditions result in very fast collapse of pertinent quantum wavefunctions to one of the allowed classical states of the biological entity. Mavromatos et al. (2002) propose a daring model that predicts dissipationless energy transfer along shielded macromolecules at near room temperatures as well as quantum teleportation of states across microtubules and perhaps neurons. It is proposed that under certain circumstances it is in principle possible to obtain the necessary isolation against environmental decoherence, so that meso/macroscopic quantum coherence, and entanglement extending over scales that are larger than the atomic scale, may be achieved and maintained for times comparable to the characteristic times for biological and cellular processes. Microtubules are comprised of tubulin that is a common polar protein found in the cytoskeleton of eukariotic cells, which is especially enriched in brain tissue. The model treats microtubules as quantum mechanically isolated high-Q QED cavities, exhibiting properties analogous to those of electromagnetic cavities routinely used in quantum optics. The model builds a microtubule network that achieves quantum teleportation of coherent quantum states, leading to decoherence-resistant bulk quantum information processing and computing within the biological matter. It is speculated that the model can explain how consciousness works, and how the brain processes and computes information.

- Teleportation of a laser beam with embedded radio signal (Bowen et al., 2003): The teleportation of a laser beam from one part of a lab to another has been demonstrated. Investigators embedded a radio signal into a laser beam, then disintegrated the beam and reassembled it a meter away, virtually instantaneously. The laser beam was destroyed in the teleportation process, but the radio signal survived. The laser light at one end of an optical communications system was disassembled and its replica was recreated elsewhere in the lab. Even though the laser beam did not survive teleportation, its encoded message did. This system could be used to transport secure data, such that it could become possible to construct a perfect cryptography system. When two parties want to communicate with one another, one can enable the secrecy of the communication to be absolutely perfect.

- Entanglement and Teleportation of a Macroscopic Ensemble of Atoms (Julsgaard et al., 2001): Expanding upon the earlier work of Hald et al. (1999) and Sackett et al. (2000), investigators experimentally demonstrated the entanglement of two macroscopic objects, each consisting of a cesium gas sample containing $\approx 10^{12}$ atoms. Entanglement is generated via interaction of the samples with a pulse of light, which performs a non-local Bell measurement on the collective spins of the samples. The entangled spin-state can be maintained for 0.5 milliseconds. The teleportation of macro-ensemble atom quantum states is expected to follow this experiment. This work is evolving towards the experimental demonstration of the Bose and Home (2002) proposal, which proved that there is a single generic process that can entangle and teleport any atoms, ions and macroscopic objects.

- Entanglement/teleportation of internal state and external motion information of atoms (Opatrný and Kurizki, 2001): Investigators propose an experiment for transmitting an atom's full information, including its "external" states, such as its energy of motion. This procedure replicates the quantum features of the external motion of a particle. For example, if particle-to-be-teleported C yielded a diffraction pattern after passing through two slits, then the same pattern would be produced by particle B, which receives the teleported information. The researchers propose the following idea: Dissociate a very cold molecule with a laser pulse into two atoms (called A and B). Then manipulate the two atoms so that they become entangled: each one is in a fuzzy state individually, but has a precisely defined relationship with its partner. Then let one of the entangled particles (such as A) collide with particle C, whose unknown state should be teleported. After their collision, the momentum values of the collision partners A and C are measured. With that information, the researchers know how to "kick" and deflect atom B, so that the motion of B precisely emulates that of particle C. The investigators say that state-of-the-art equipment for studying atomic collisions and quantum effects makes this experiment difficult, but feasible, to do. If this proposal proves to be correct, then the implication is that it will become possible to experimentally expand this concept to the teleportation of a large ensemble of atoms, such that the entire physical motion and quantum states of the ensemble can be teleported. This could lead to the future development of a teleportation process similar to what was discussed in Section 3.1.

- Laser-like Amplification of Entangled Particles and Entangled-Photon Lasers (Lamas-Linares et al., 2001): Entangled particles are notoriously difficult to create in bulk. To create entangled photons, for example, researchers use the parametric down-conversion technique to send laser light through a barium borate crystal. Passing through the crystal, a photon sometimes splits into two entangled photons (each with half the energy of the initial photon). However, this only occurs for one in every ten billion incoming photons. To increase the yield, researchers added a step: they put mirrors beyond the crystal so that the laser pulse and entangled pair could reflect, and have the chance to interact. The entangled pair and reflected laser pulse interfere constructively to generate fourfold more two-photon pairs or interfere destructively to create zero pairs. Following these steps, the researchers increased production of two-photon entangled pairs, and also of more rare states such as four-photon entangled quartets. This achievement could represent a step towards an entangled-photon laser, which would repeatedly amplify entangled particles to create greater yields than previously possible, and also towards the creation of new and more complex kinds of entangled states.

This list is by no means complete as new developments in this field continue to arise.

3.3 Conclusion and Recommendations

Given the incredible advancements that have been made in the entanglement and teleportation of macroscopic objects the size of 10^{12} atoms, we are still very far away from being able to entangle and teleport human beings (and even simpler biological entities such as cells, etc.) and bulk inanimate objects (tools, technical equipment, pencils and pens, weapons platforms, communications devices, personal hygiene supplies, etc.). There still remain four essential problems:

- One needs an entangled pair of such bulk objects.
- The bulk objects to be entangled and teleported must be in a pure quantum state (as in a Bose-Einstein condensate, for example). And pure quantum states are very fragile.
- The bulk objects to be entangled and teleported must be extremely isolated from the environment to prevent the onset of decoherence.
- The Bell-state measurement of animate or inanimate objects during entanglement/teleportation will require extracting an amount of information (in bits) that equals or exceeds the number of atoms contained within the object. This infers that the computer storage and processing requirements to entangle and teleport a complete bulk object will be astronomically huge (recall the discussion in Section 3.1).

It is difficult to imagine how we can achieve an extreme level of environmental isolation for an object, let alone a living being that breathes air and radiates heat. Experiments with atoms and larger objects must be done in a high vacuum to avoid collisions with molecules. Thermal radiation from the walls of a teleportation apparatus would easily disturb a tiny amount of matter. At present, decoherence imposes a fundamental limit on quantum entanglement and teleportation. Decoherence is the primary reason why we do not routinely see any quantum effects in our everyday world. Research is continuing on whether decoherence can be reduced, circumvented, or otherwise be eliminated. And some minor progress has been made in that direction.

In q-Teleportation it is the ***quantum states*** of the objects that are destroyed and recreated, and not the objects themselves. Therefore, q-Teleportation ***cannot*** teleport animate or inanimate matter (or energy) in its physical entirety. However, some experts argue that because an object's quantum state is its defining characteristic, teleporting its quantum state is completely ***equivalent*** to teleporting the object, even though the original object's quantum state (and defining characteristic) was completely destroyed in the process. This goes to the heart of what is meant by identity. When an object has all the right properties and features, it will be the same object that one observes whether it was observed now or 24 hours ago. Quantum physics reinforces the point that objects of the same type in the same quantum state are indistinguishable from each other. One should, according to this quantum principle, be able to swap all the atoms in a particular object with the same atoms from a mound of raw materials, and reproduce the original object's quantum states exactly with the end result that the new object is identical to the original. Last, we do not know how to put a human being into a pure quantum state or what doing so would mean for biological functioning (including brain function), but we do know how to put $\leq 10^{12}$ gas atoms/ions and a beam of photons into a pure state in practice. Further research will be required to ascertain whether microbiological and higher-level biological systems, in addition to bulk inanimate matter, can be put into pure quantum states and entangled/teleported.

To perform a Bell-state measurement on (bulk) animate or inanimate objects, during the entanglement/teleportation process, to extract and encode its information will require extracting an amount of information (in bits) that equals or exceeds the number of atoms contained within the object. An object containing a few grams of matter will require the extraction of $> 10^{28}$ bits of data. A simple virus of $\approx 10^7$ atoms would require the extraction of $\geq 10^8$ bits of information during the entanglement/teleportation process, whereas the extraction of a minimum of 10^{28} kilobytes will be required to encode and store an entire human being. This is beyond the capability of present digital electronic computer technology to store and process. It is difficult to see how far computer technology will advance towards meeting this requirement.

It is difficult to fathom what will be in store for the teleportation of human beings given some possible future technology. What about the effects of the q-Teleportation process on the human consciousness, memories and dreams, and the spirit or soul? We know from quantum physics that “the whole is greater than the sum of its parts.” So what happens to the fundamental characteristics of a human being when he/she steps into the teleporter-transmitter, where their quantum states (i.e., their complete identity) are destroyed during the quantum entanglement/teleportation process, and then their copy is created at the teleporter-receiver an instant later? What will things be like during the entanglement process? Will a teleported individual’s consciousness, memories and dreams, and spirit/soul be successfully and accurately teleported or not? This is a major ethical and technical question that will have to be addressed by future research.

•**Recommendations**:

- Broad-spectrum Quantum Computing Technology Development Program: At present, the Quantum Information Science Program (QISP) is coordinated by the U.S. Army Research Office with funding and support from the Army, the National Security Agency, DARPA, and the Office of the Deputy Director of Defense for Research and Engineering. The Naval Research Lab and the CIA are both involved in their own programs. The CIA vets new commercial development of computer technology and computer information processing via its In-Q-Tel company (reference 44). This includes R&D on quantum entanglement and teleportation for computer, information processing and secure communications. QISP was funded for $19 million in 1999. The program involves 34 projects by researchers at 21 universities, three government laboratories and two corporate laboratories. QISP goals include building a quantum computer, developing quantum information processing, and further advances in quantum teleportation. The AFRL should join QISP and provide partnership funding on the order of $1 million per year. An alternative to this would be for AFRL to collaborate with In-Q-Tel and participate in its technology R&D venture capital programs. This R&D investment would allow the Air Force to acquire very advanced quantum physics and related technological applications that can support its mission. The R&D investment benefits would include the development and implementation of quantum computing/information processing and secured quantum communications technology, which can significantly enhance the performance and security of Air Force computing and communication systems infrastructure, and aerospace weapons systems.

- Quantum Cryptography: A dedicated research program should be implemented to develop a mature quantum cryptography technology. Theoretical and experimental work is in progress among a small number of select groups (QISP, In-Q-Tel, universities, etc.), but this field is not advancing fast enough for practical applications to become available to meet increasing adversarial threats against secured military and intelligence communications. The goal of proposed quantum cryptography research is to bring the theoretical and experimental foundation of quantum cryptography and secure quantum information processing to maturity, and to fully develop and implement quantum entanglement/teleportation-based cryptography technology. Recent experimental work has demonstrated that a completely secure quantum key can be generated and distributed for the communication and decoding of encrypted messages using entangled photons. Any eavesdropper’s attempt to intercept the quantum key will alter the contents in a detectable way, enabling users to discard the compromised parts of the data. There is much more work that needs to be done in this area. I recommend that the AFRL implement a

$1 million/year program for five years in order to advance the state-of-art in quantum cryptography technology.

- Quantum Decoherence: Decoherence is the primary reason why we do not routinely see any quantum effects in our everyday world. And it imposes a fundamental limit on quantum entanglement and teleportation via the interaction between entangled/teleported quantum systems and their local environment. In order to advance quantum entanglement/teleportation physics and develop applied technologies, it is necessary that a research program be implemented by the AFRL to explore whether decoherence can be significantly reduced, circumvented, or otherwise be eliminated. An insufficient number of small university groups have slowly made minor progress in this direction. I recommend that a $500,000 - 750,000 per year R&D program be conducted for five years to overcome this technical challenge.

- Pure Quantum States: In order to entangle and teleport quantum particles and bulk objects, they both must be prepared in a pure quantum state. And pure quantum states are very fragile to decoherence. A technical challenge for entanglement/teleportation physics is whether the requirement for pure quantum states can be relaxed and how much decoherence will play a role in this situation, what technical challenges will arise when increasing the size of entangled/teleported matter to larger macroscopic scale ($>> 10^{12}$ atoms), and whether matter of mixed composition (such as a gas or Bose-Einstein condensate of mixed atomic elements) can be entangled/teleported in both pure and mixed quantum states. I recommend that a $250,000 – 500,000 per year research program be conducted for five years to study this problem.

- Entangling Bulk Matter and Bell-State Measurement to Extract Information: Recent experiments demonstrated the entanglement of two macroscopic objects, each consisting of a cesium gas sample containing $\approx 10^{12}$ atoms. Entanglement was generated via interaction of the samples with a pulse of light, which performs a non-local Bell measurement on the collective spins of the samples. In order to push the envelope on this development and take it to higher practical levels, it will be necessary to ascertain the limit on the size and composition of bulk matter entanglement (given the decoherence and pure-state constraints); and to determine what other quantum states can be used for entanglement, what other Bell-state measurement techniques can be used, and whether multiple quantum states can be entangled. The chief technical challenge is the computer technology that will be required to facilitate the huge amount of data that must be extracted, processed and stored from bulk matter quantum states during the Bell-state measurement process. I recommend that a $500,000 – 1 million per year research program be implemented for five years in order to explore these questions and ascertain what solutions may be technically available, and to develop such solutions.

- Biological Quantum Teleportation: The Mavromatos et al. (2002) theoretical model for biological entanglement and teleportation is a remarkable concept that could result in the development of a workable physics theory of consciousness. The model has potential applications to advanced quantum computing/information processing physics and the physics of psi phenomena (see Chapter 5). A research program should be implemented to continue the Mavromatos et al. (2002) work and bring their model to theoretical maturity. It is recommended that this program be funded at $500,000 – 800,000 per year for five years. A parallel or follow-up program should be implemented to experimentally test this model and ascertain any useful technological applications. One application that should be explored in the proposed research program is advanced, ultra-fast, ultra-high-capacity quantum computing and information processing using natural and/or artificial biological systems. The parallel or follow-up experimental research program should be funded at $800,000 – 1.5 million per year for five years.

- FTL Communication: Experiments verifying that EPR entanglement obeys Special Relativity (Seife, 2000; Scarani et al., 2000; Gisin et al., 2000; Zbinden et al., 2000a, b) determined that quantum information via EPR photon pair entanglement must travel > 10^7 times light speed. Can this mechanism be exploited to achieve FTL communication? If so, then the potential military and commercial applications will be revolutionary, and the science and industry of communications will be forever transformed. A comprehensive theoretical and experimental research program should be implemented to answer this question. It is recommended that this program be funded at $700,000 – 1 million per year for five years. A modest experiment definition study should be funded at $80,000 for one year to delineate the most promising experimental approaches to be used for the larger research program. [There is much controversy and debate over FTL (a.k.a. superluminal) signals/communication, and the reader should see the selected superluminal references in the Teleportation References section of this study.]

- New Entanglement/Teleportation Breakthroughs: The most exciting developments in quantum teleportation physics has included the teleportation of a laser beam with an embedded radio signal, the teleportation of squeezed states of light (and hence, continuous quantum state variables), the teleportation of photon states to atoms/ions (from light to matter!), the entanglement of two similar/dissimilar quantum particles that are created by two (independently) different particle sources, the laser-like amplification of entangled particle/photon pairs, parametric down-conversion entanglement and teleportation (of discrete quantum state variables), quantum cryptography with unbreakable keys, the teleportation of quantum information at speeds > 10^7 times light speed, the entanglement and teleportation of macroscopic (10^{12} atoms) matter quantum states, etc. There is also the yet-untested proposal to entangle/teleport the external physical motion and internal quantum state information of atoms. This shows that quantum physics sets no apparent limit on what it is that can be teleported/entangled and how it is to be teleported/entangled, or where it is to be teleported/entangled. At present teleportation technology requires fiber optic and coaxial cables to teleport quantum state information from one location to another. Can we avoid the use of cables and teleport through free space? [Note: Before this report went to press, Aspelmeyer et al. (2003) reported their outdoor experiment that demonstrated the distribution of quantum entanglement (of laser photons) via optical free-space links to independent receivers separated by 600 m across the Danube River (during inclement nighttime weather), with no line of sight between them. This experiment is revolutionary and begins the step toward conducting satellite-based distributed quantum entanglement.] We have not discovered all the possibilities that nature has in store for us. The present breakthrough discoveries will likely introduce novel military and intelligence technology applications in the near and far future. But further R&D must be conducted in order to discover new applications for these recent breakthroughs, to make additional breakthroughs and discoveries, and to advance the state-of-art in quantum teleportation physics to meet future challenges to the Air Force mission. I recommend that a two-track R&D program be implemented over five years. The first track should be funded at $250,000 – 750,000 per year for the purpose of developing new entanglement/teleportation breakthroughs in quantum teleportation physics. The second track should be funded at $750,000 – 1.5 million per year for the purpose of developing applications for any new breakthroughs with the proviso that such applications benefit the Air Force mission and have commercial dual-use capability to leverage advance technology in the private sector.

4.0 e-TELEPORTATION

4.1 Extra Space Dimensions and Parallel Universes/Spaces

A literature search for proposed e-Teleportation concepts based on the conveyance of objects through extra space dimensions and/or parallel universes/spaces has yielded only one result (see Section 4.2). The present state-of-art in research on parallel universes/spaces and extra space dimensions has been strictly limited to the work on developing a grand unified quantum field theory and a quantum theory of gravity, whereby the former necessarily includes the latter. Quantum gravity/unified field theory research has been evolving since the 1920s when Kaluza and Klein published the first papers to describe a model for the unification of gravity with electrodynamics. Many of the more prominent theories today invoke extra spatial dimensions, the existence of parallel universes/spaces, or both in order to quantise gravity and/or to unify gravity with the other forces of nature. It is beyond the scope of this study to provide an in-depth review of all of the research that has been done in this area, so we list below a select few of the historically prominent models that have largely gained a secure foothold in present-day research:

- Kaluza-Klein Electromagnetic-Gravity Unification Theory/Modern Kaluza-Klein Gravity Theories (Kaluza, 1921; Klein, 1926; de Sabbata and Schmutzer, 1983; Lee, 1984; Appelquist et al., 1987; Kaku, 1993, 1994; Overduin and Wesson, 1998): It was originally suggested that Maxwellian electrodynamics and Einstein gravitation could be unified in a theory of five-dimensional Riemannian geometry, where the gravitational and electromagnetic potentials together would determine the structure of spacetime. The fifth space dimension is curled up into a ball of space with a radius slightly larger than 10^{-35} m, and it was originally regarded as having no physical significance because it was simply a mathematical tool used to catalyze unification. At present, the generic name of Kaluza-Klein stands for a wide variety of approaches to quantising and unifying gravitation with other quantum fields using any number of dimensions greater than four.

- Superstring Theories (Green, 1985; Kaku, 1988, 1993, 1994): These theories come in a wide variety of interrelated concepts, and they are a highly evolved form of Kaluza-Klein theories. They are based on the dynamics of string-like fundamental quanta, whereby the observed fundamental particles are manifested by the vibrational ground or excitation states of a quantum string (open or closed loop). The superstrings are $\approx 10^{-35}$ m (i.e., the Planck length) in size. There are different versions of these theories that require ten, eleven or twenty-six extra space dimensions to unify and quantise gravity, whereby the extra dimensions are curled up (i.e., compactified) into balls of space with a radius $< 10^{-35}$ m. These theories later evolved into versions that are now called F- and M-theory. The mathematics behind this class of theories is very ugly, and it is difficult for even the best superstring theorists to make simple or sophisticated calculations and predictions. And so far, this class of quantum gravity theories has escaped experimental verification.

- D-Brane and 3-Brane Theories/Parallel Spaces (Rubakov and Shaposhnikov, 1983a, b; Polchinski, 1995; Antoniadis et al., 1998; Randall and Sundrum, 1999a, b; Weiss, 2000; Pease, 2001; Arkani-Hamed et al., 1998, 2000, 2002): D-brane theory is a recent incarnation of the original superstring theories in which open strings, corresponding to the fundamental particles of the standard model (quarks, leptons, gauge bosons), have their free ends stuck on a (hypersurface) membrane called a D-brane (D = Dirichlet boundary conditions). But the graviton, which corresponds to a closed loop of string, can propagate in all the dimensions. It provides both unification and quantization of gravity by assuming that there are n new spatial dimensions in addition to the three infinite spatial dimensions we know about. And the extra space dimensions are $\approx 10^{-35}$ m in extent. A very recent alternative version of this model is called "3-brane" theory. In this theory, each of the n extra space dimensions is of finite extent $R \approx 2\times10^{(32/n)-17}$ centimeters. The space spanned by the new dimensions is called "the bulk." In this theory, the particles of the standard model live within our familiar realm of three spatial dimensions, which forms a three-dimensional (hypersurface) membrane or "3-brane" within the bulk. The propagation of electroweak and strong nuclear forces is then confined to our 3-brane. However, at distances (r) less than R, gravity (via gravitons) propagates in the full $(3 + n)$-dimensional space, whereby its strength falls as $r^{-(2+n)}$ with increasing separation r. When $r > R$, the gravitational force reverts to its normal Newtonian r^{-2} falloff because there is no longer any extra-dimensional space for it to spread into. If $n = 1$, then the size of the extra-dimension would have to be $R \approx 2\times10^{15}$ cm (or 2×10^{10} km = 133.3 AU; 1 AU = 1.5×10^{8} km is the mean Earth-Sun distance) in order to account for the weakness of gravity, but an extra space dimension this large would have already made itself obvious in the observed dynamics of the solar system. For this reason, investigators have discounted the possibility that $n = 1$. If $n = 2$, then the size of both extra space dimensions would have to be $R \approx 0.2$ cm (or 2 mm). In any case, inconspicuous neighboring 3-branes may be separated from the 3-brane we live on by only a fraction of a millimeter, or even much smaller distances, across the higher-dimensional bulk. Such neighboring 3-branes may be distant folds of our own 3-brane, with the same physics, but able to influence us across shortcuts through the bulk. Or they may be completely separate 3-branes possessing their own fundamental laws and parameters of nature that are completely different from our own. Several tabletop Cavendish-type experiments are now looking for sub-millimeter deviations from Newtonian gravitation as a first step towards verifying 3-brane theory, and other experiments are now being planned or are already underway (Pease, 2001). At present the preliminary experimental results have been negative for the existence of extra space dimensions, and the experimental data suggests that two extra space dimensions are now constrained to length scales << 0.2 – 0.3 millimeters while seven extra space dimensions can be no larger than 2 femtometers (Pease, 2001).

- Parallel Universes/Parallel Spaces (Everett, 1957; Wheeler, 1957, 1962; DeWitt, 1970; DeWitt and Graham, 1973; Jammer, 1974; Davies, 1980; Wolf, 1988; Kaku, 1994; Visser, 1995 and Section 2.1): There are only two other research tracts that are concerned with parallel universes besides 3-brane theory. The first tract is the traversable wormhole research that was discussed in Section 2.1. Traversable wormholes can connect many different universes in the "multiverse" (i.e., a conglomeration of many universes), and these are called inter-universe wormholes. However, traversable wormhole physics (a.k.a. Einstein's General Relativity Theory) does not provide a physical prescription for the existence and nature (i.e., fundamental parameters and physical laws) of other putative universes. The difference between inter-universe and intra-universe (i.e., two distant regions of one universe are connected with each other) wormholes arises only at the level of global geometry and global topology. Local physics near the throat of a traversable wormhole is insensitive to issues of intra-universal or inter-universal travel. An observer in the vicinity of the throat, while making local measurements, would not be able to tell whether he was traveling to another universe or to a remote part of our own universe. And one cannot rely on the topological (as opposed to geometrical) information to determine which is the case, because topological information is not enough to uniquely characterize an inter-universe connection. And General Relativity Theory does not fix the topology of spacetime, so we cannot ascertain the existence of other universes. [Note: Traversable wormholes are also geometrically possible for higher dimensional spaces.] The second tract is the "Many Worlds" interpretation of quantum theory. This version of quantum theory requires the simultaneous existence of an infinite number of equally real worlds, all of which are more-or-less causally disjoint, in order to interpret consistently the relationship between observed phenomena and observers. The theory was proposed in an attempt to overcome a number of deep paradoxes inherent in the interpretation of the theory of measurement and quantum theory. The Many Worlds theory argues that quantum theory requires the existence of a "superspace" of worlds spanning the range of all possible quantum observations (or quantum measurements). Through our acts of measurement we are imagined to trace a path through the mesh of possible outcomes. All the "worlds" are causally disjoint, and the uncertainty of quantum observation can be interpreted as an artifact of our access to such a limited portion of the superspace of possible worlds. The evolution in the superspace as a whole is entirely deterministic.

At present, none of the theoretical concepts outlined above have been brought to a level of technical maturity, where it becomes meaningful to ascertain whether any form of e-Teleportation is theoretically possible between extra space dimensions and different or parallel universes/spaces. However, as mentioned in the item on parallel universes/parallel spaces, there is the exception that traversable wormholes (three- and higher-dimensional) provide a solid physics principle for the implementation of teleportation between parallel universes/spaces. And traversable wormholes can be devised to connect 3-branes together. See Section 2.1 for the discussion on teleportation via traversable wormholes. Also, Kaluza-Klein theories, superstring theories and D-brane theory all have the common feature that their extra space dimensions are $\leq 10^{-35}$ m in extent, which makes it impossible for any useful form of macroscopic-level teleportation to occur between space dimensions. Last, it is not yet possible to do theoretical calculations or even experimentally verify most of these theories. Three-brane theory is the best parallel space theory there is, with the possibility that macroscopic-level teleportation is possible between space dimensions (only if the extra space dimension(s) has length scale(s) >> millimeters). But this theory is still in the stage of maturing theoretically and achieving experimental verification (or falsification). Therefore, we can go no further in this section.

4.2 Vacuum Hole Teleportation

An unusual teleportation concept has been proposed by Leshan (1999, 2002), which describes the teleportation of objects throughout our universe by using the geometrical properties of spacetime. The proposal posits that there is a "zero-space" that exists outside the boundary of our universe, whereby this zero-space is a "point form" space, where the distance between any two points is always equal to zero. Leshan also calls this space a "hole." Further requirements and assumptions of the model are:

- time does not exist as a property in zero-space
- the cosmological principle (i.e., there are no privileged frames relative to another place or point in the universe) requires that the boundary or border of the universe must pass through every point of space
- virtual holes (or zero-space) in spacetime must exist at every point of the universe, which are also called "vacuum holes"
- vacuum holes exist as virtual particles

The last item is interesting because it implicitly says that vacuum holes (a.k.a. zero-space) must also be virtual particles, and in Section 2.2 we showed that virtual particles are a representation of the vacuum ZPF. Therefore, this infers that vacuum holes can be considered to be vacuum zero-point fluctuations in

Leshan's model. Thus, a teleportation mechanism can arise in this model because distances between zero-space and any other point in the universe are zero, so that the vacuum holes can potentially exist at every point in the universe simultaneously. Therefore, if an object is sent "out of the universe" and into a vacuum hole (a.k.a. zero-space), then the object can appear at random at any spacetime point in the universe.

The mechanism for teleportation in this model is:

- to send an object outside of the universe by creating a closed surface (i.e., "hole sphere"), which consists of vacuum holes, around the object;
- while inside the hole sphere, the object then ceases to exist because objects cannot really exist outside of the universe;
- however, the object simultaneously exists at any other remote location in the universe (via the cosmological principle) at the instant it became enclosed by the hole sphere;
- therefore, it has been teleported to some remote location in the universe

Leshan points out that the teleportation device must curve spacetime so that the starting and destination points in the universe coincide, and the curved geometry must be similar to that of a black hole for an instant, so that a channel between the two points can be formed. (This sounds suspiciously like creating a traversable wormhole via an Einstein-Rosen bridge, which can be made traversable by perturbing the Schwarzschild spacetime metric an infinitesimal amount.) There is no space to traverse, so therefore there will be no passage of time during teleportation. The only expenditure of energy in this teleportation scheme is the energy that will be needed to curve spacetime.

This teleportation concept is very convoluted. Leshan does not offer any further explanations that are useful nor does he offer any precise technical description for the vacuum holes, and how they are to be produced and manipulated. There is also no mathematical physics derivation published by Leshan to support this concept. I am totally unable to evaluate this concept in the absence of a rigorous theoretical framework. This concept is too sketchy and full of technical "holes" to seriously consider it any further for this study. The reader should note that it has already been demonstrated that traversable wormholes are the best physical principle available to implement teleportation between universes and extra space dimensions.

4.3 Conclusion and Recommendations

At present, none of the theoretical concepts explored in this chapter have been brought to a level of technical maturity, where it becomes meaningful to ascertain whether any form of e-Teleportation is theoretically possible between extra space dimensions and different or parallel universes/spaces. However, there is the exception that traversable wormholes (three- and higher-dimensional) provide a solid physics principle for the implementation of teleportation between parallel universes/spaces. And traversable wormholes can be devised to connect 3-branes together. Kaluza-Klein, superstring and D-brane theories do not allow for any useful form of macroscopic-level teleportation to occur between space dimensions, because these theories require that the extra space dimensions be $\leq 10^{-35}$ m in extent. Last, it is not yet possible to do theoretical calculations to make predictions or even to experimentally verify most of these theories. Three-brane theory is the best parallel space theory there is with the possibility that macroscopic-level teleportation is possible between space dimensions. But this theory is still in the stage of maturing theoretically and getting experimental verification.

•**Recommendations**:

- The recommendations outlined in Section 2.3 are relevant to the investigation of the possibility for e-Teleportation.

Approved for public release; distribution unlimited.

5.0 p-TELEPORTATION

5.1 PK Phenomenon

P-Teleportation is a form of psychokinesis (or PK) similar to telekinesis but generally used to designate the movement of objects (called apports) through other physical objects or over great distances. Telekinesis is a form of PK, which describes the movement of stationary objects without the use of any known physical force. And PK is essentially the direct influence of mind on matter without any known intermediate physical energy or instrumentation. Rigorously controlled modern scientific laboratory PK, and related psychic (a.k.a. "psi", "paranormal" or parapsychology), research has been performed and/or documented by Rhine (1970), Schmidt (1974), Mitchell (1974a, b, see also the references cited therein), Swann (1974), Puthoff and Targ (1974, 1975), Hasted et al. (1975), Targ and Puthoff (1977), Nash (1978, see also the references cited therein), Shigemi et al. (1978), Hasted (1979), Houck (1984a), Wolman et al. (1986, see also the references cited therein), Schmidt (1987), Alexander et al. (1990), Giroldini (1991), Gissurarson (1992), Radin (1997, see also the references cited therein), Tart et al. (2002), Shoup (2002), and Alexander (2003).

A well-known theoretical/experimental/operational program directed by H. E. Puthoff, R. Targ, E. May and I. Swann was conducted at SRI International and the NSA, and sponsored at various times by the Central Intelligence Agency (CIA), the Defense Intelligence Agency (DIA), and the Army Intelligence and Security Command (INSCOM) over more than two decades; and the program was later carried on by E. May at SAIC (Alexander, 1980; Puthoff, 1996; Targ, 1996; Schnabel, 1997; Tart et al., 2002). This was called the Remote Viewing program, and it was a compartmentalized special access program possessing a variety of codenames during its 22 years of operation. Remote viewing involves precognition and clairvoyance, and it allows a practitioner to acquire information irrespective of intervening distance or time. The Remote Viewing program ended in 1994 and President W. J. Clinton officially declassified it in 1995. The reader should note that the very first U. S. military-intelligence R&D programs on psi, PK and mind control were conducted by H. K. (Andrija) Puharich, M.D., L.L.D during his military service at the Army Chemical and Biological Warfare Center at Fort Detrick, Maryland in the 1940s-50s. Puharich had an interest in clairvoyance and PK, and dabbled in theories for electronically and pharmaceutically enhancing and synthesizing psychic abilities. While in the Army, Puharich took part in a variety of parapsychology experiments, and he lectured Army, Air Force and Navy groups on possibilities for mind warfare. He was a recognized expert in hypnotism and microelectronics.

PK phenomenon was also explored in the Remote Viewing program. Col. J. B. Alexander (USA ret.) credits professional aerospace engineer Jack Houck for "capturing PK phenomenon and transitioning it into an observable form" (Houck, 1982, 1984a, b; Alexander et al., 1990; Alexander, 2003). During the past three decades, Houck (along with Alexander) held a number of PK sessions, whereby attendees are taught the PK induction process, and initiate their own PK events using various metal specimens (forks, spoons, etc.). Individuals were able to completely bend or contort their metal specimens with no physical force being applied whatsoever. Numerous government science advisors and senior military officials took part in and/or witnessed these events, which took place at the Pentagon, at officers' or scientists' homes, and at one quarterly INSCOM retreat attended by the commanding general and a group of colonels and generals commanding INSCOM units around the globe. Spontaneous deformation of the metal specimens was observed at the PK session conducted during the INSCOM retreat, causing a great deal of excitement among those present. Other notable trained observers were also present at this session, and they critically reviewed the events. Psychic Uri Geller (1975) is the original model for demonstrating

PK metal bending. During a talk that he gave at the U.S. Capitol building, Uri caused a spoon to curve upward with no force applied, and then the spoon continued to bend after he put it back down and continued with his talk (Alexander, 1996). Jack Houck continues doing extensive experimental work and data collection on micro- and macro-PK phenomena. Scientifically controlled PK experiments at the Princeton University Engineering Anomalies Research Laboratory were conducted by Robert Jahn (Dean Emeritus of the School of Engineering), who reported that repeatedly consistent results in mentally affecting material substances has been demonstrated in the lab (Jahn and Dunne, 1987). In the 1980s, Jahn attended a meeting on the PK topic at the Naval Research Laboratory, and warned that foreign adversaries could exploit micro- or macro-PK to induce U.S. military fighter pilots to lose control of their aircraft and crash.

Very early investigations of, and experiments on, p-Teleportation occurred during the 19th and early 20th centuries. Many cases that were studied, and the experiments that were performed, were undoubtedly due to fraud, and few experiments have occurred under controlled conditions during that period. However, most of the credible, scientific reports of p-Teleportation phenomenon and related (controlled) experiments occurred in the late 20th century (see for example, Alexander et al., 1990; Radin, 1997). Some of that scientific work involved the investigation of Uri Geller and a variety of other recurrent spontaneous PK phenomena (Hasted et al., 1975; Puthoff and Targ, 1975; Targ and Puthoff, 1977; Nash, 1978; Wolman et al., 1986). Psychics Uri Geller (1975) and Ray Stanford (1974) claimed to have been teleported on several occasions. Most claimed instances of human teleportation of the body from one place to another have been unwitnessed. There are also a small number of credible reports of individuals who reported being teleported to/from UFOs during a UFO close encounter, which were scientifically investigated (Vallee, 1988, 1990, 1997). But there are a larger number of such reports that are anecdotal, whereby the witness data tends to be unreliable. However, we will confine our discussion to the controlled laboratory experiments that have been performed and reported.

One of the more interesting examples of controlled experiments with Uri Geller was one in which he was able to cause a part of a vanadium carbide crystal to vanish (Hasted et al., 1975). The crystal was encapsulated so it could not be touched, and it was placed in such a way that it could not be switched with another crystal by sleight of hand. A more spectacular series of rigorously controlled (and repeatable!) laboratory experiments occurred in the Peoples Republic of China (PRC). In September 1981, an extraordinary paper was published in the PRC in the journal Ziran Zazhi (transl.: Nature Journal), and this paper was entitled, "Some Experiments on the Transfer of Objects Performed by Unusual Abilities of the Human Body" (Shuhuang et al., 1981). The paper reported that gifted children were able to cause the apparent teleportation of small objects (radio micro-transmitters, photosensitive paper, mechanical watches, horseflies, other insects, etc.) from one location to another (that was meters away) without them ever touching the objects beforehand. The experiments were operated under exceptionally well-controlled conditions (both blind and double-blind). The researchers involved included not only observers from various PRC colleges and medical research institutes, but also representatives from the PRC National Defense Science Commission. Because of the involvement of the latter, it was deemed necessary that an unclassified Intelligence Information Report be prepared by the DIA (see Shuhuang et al., 1981), which included a detailed English translation of the article.

Additional research carried out by the Aerospace Medicine Engineering Institute in Beijing was reported in the July 1990 issue of the Chinese Journal of Somatic Science (Kongzhi et al., 1990; Jinggen et al., 1990; Banghui, 1990), which was also translated into English by the DIA. Reported in several articles are experiments involving the videotaping and high-speed photography of the transfer of test specimens (nuts, bundles of matches, pills, nails, thread, photosensitive paper, chemically treated paper, sponges dipped in $FeCl_3$, etc.) through the walls of sealed paper envelopes, double layered KCNS type paper bags, sealed glass bottles and tubes with sealed caps, and sealed plastic film canisters without the walls of any of these containers being breached. All of the Chinese experiments reported using gifted children and young adults, who possessed well-known extraordinary PK ability, to cause the teleportation of the various test specimens. In all the experimental cases that were reported, the test specimens that were teleported were completely unaltered or unchanged from their initial state, even the insects were unaffected by being teleported. The experiments were well controlled, scientifically recorded, and the experimental results were always repeatable.

The Chinese papers are all extremely interesting and very well written, and they show photographs and schematic diagrams of the various experimental setups. The experimental protocols were explained in lengthy detail, and thorough data and statistical analysis were presented in the results. The combined results from the several Chinese experiments showed that:

- different research groups designed different experimental protocols, used different gifted psychics, used different sealed containers, and used different test specimens (live insects, bulk inanimate objects, and even radio micro-transmitters were used to track the location of the specimens) that were to be teleported;

- the time required for the teleportation of test specimens through various barriers was anywhere from a fraction of a second to several minutes, and this was not dependent on the test specimen that was used, the sealed container that was used (or its barrier thickness), which experimental protocol was used, or which psychic was being used

- the high-speed photography/videotaping recorded in one series of experiments that test specimens would physically "meld" or blend with the walls of sealed containers; and recorded in a different series of experiments that test specimens would simply disappear from inside the container only to reappear at another location (after seconds to several minutes of time transpired), such that the test specimen did not actually undergo total material disintegration/reintegration during teleportation – this data is important, because without the aid of electronic monitoring instruments, the average person's sensory organs and usual methods of detection are temporarily unable to perceive the test specimen's (ambiguous) existence during the teleportation process;

- the radio micro-transmitter used as a test specimen in one series of experiments (Shuhuang et al., 1981) transmitted a radio signal to several stationary electronic instruments/receivers, so that the specimen could be tracked and monitored (via signal amplitude and frequency measurements) during the teleportation process; the experimenters discovered that there was large fluctuations in the intensity (in both amplitude and frequency) of the monitored signal to the effect that it would either completely disappear or become extremely weak (to the extent that the monitoring instruments could scarcely detect it) – it was discovered that there was a definite correlation between the change in strength (i.e., radical frequency shifts were observed) of the monitored radio signal and the teleportation of the test specimen, such that the weak or absent signal indicated that the specimen was "nonexistent" (or in an altered physical state) during teleportation (note: the monitored signal amplitude and frequency of the micro-transmitter specimen were stable before and after teleportation);

- before and after "passing through the container wall/barrier", the test specimen and the container's wall/barrier are both complete solid objects;

- the gifted psychics were never allowed to see (they were blindfolded in many experiments) or touch each of the test specimens or the sealed containers before and after experiments were conducted, and only the experimenters touched the specimens and containers (using both blind and double-blind protocols);

- the experimental results were all repeatable

- the conditions for fraud and sleight of hand were totally eliminated, and multiple independent outside witnesses (technical and military-intelligence experts) were present at all times to ensure total fidelity of the experiments

The experimental radio micro-transmitter and high-speed photography/videotaping data offer an important clue on what the teleportation mechanism is, and this will be discussed further in Section 5.1.1. The Chinese were unable to offer any significant physics hypothesis that could explain their results. Some researchers stated that it is necessary to invoke a new physics, which somehow unifies the human consciousness (i.e., physics of consciousness) with quantum and spacetime physics, in order to understand p-Teleportation and related PK phenomena. The researchers were amazed by their repeated results, and were barely able to fathom the altered "state of being" that test specimens underwent during teleportation.

It is also important to point out that during the Cold War the DIA produced three (now declassified) reports on the parapsychology research of the Soviet Union and its Warsaw Pack allies (LaMothe, 1972; Maire and LaMothe, 1975; DIA Report, 1978; other related studies were reported by Groller, 1986, 1987). The purpose of the reports was to collate and summarize collected intelligence, describe in great detail, and assess the Soviet Union and Warsaw Pact R&D on parapsychology and paraphysics. The reports outlined the history of pre-revolutionary (Czarist) Russian, and WWII and post-WWII era Soviet R&D on psychotronics, human mind/behavior control, and the entire spectrum of parapsychology. The Soviet information also mentions the psychotronic/parapsychology R&D materials that Soviet military forces took from various Nazi research centers in and around Germany at the end of WWII. The entire spectrum of parapsychology phenomena was explored by the Soviets, which resulted in the generation of a wealth of experimental data and related scientific research literature. One DIA report noted that there was an East versus West science debate in the Soviet literature over whether paranormal phenomenon and related experimental data was real or even scientifically sound in comparison to western scientific practice and philosophy. Another DIA report lists the names and affiliations of all the researchers, as well as the names of the various Soviet and Warsaw Pact research centers, that were involved. Also, Pratt (1986) reviews and summarizes the history of Soviet psychotronics research.

The U.S. military-intelligence establishment was concerned with the possibility that the Soviets and their Warsaw Pact allies were conducting psychotronics and mind control R&D in order to discover how to exploit and control powerful phenomena that could be used against the U.S. and its allies. LaMothe (1972) chronicled how the Soviets had been researching methods of influencing human behavior for over sixty years. The Soviets and their allies extensively explored an influence technology that they called "controlled offensive behavior", which is defined as "research on human vulnerability as it applies to methods of influencing or altering human behavior" (LaMothe, 1972). Also, LaMothe (1972) describes the revolutionary techniques the Soviets studied to influence human behavior, which included: sound, light, color, odors, sensory deprivation, sleep, electromagnetic fields, biochemicals, autosuggestion, hypnosis, and parapsychology phenomena (such as psychokinesis, telekinesis, extrasensory perception-ESP, astral projection, clairvoyance, precognition, and dream state, etc.). The LaMothe (1972) report became an aid in the development of countermeasures for the protection of U.S. and/or allied personnel.

Psychotronics is the general term that was used in the former Soviet Union/Warsaw Pact countries to categorize many psychic phenomena undergoing scientific research. The conclusions that were reached in the DIA reports are that within the category of psychotronics, the Soviets identified two discrete skills (LaMothe, 1972):

- bioenergetics: those phenomena associated with the production of objectively detectable effects such as psychokinesis, telekinesis, levitation effects, transformations of energy, i.e. the altering or affecting of matter

- bioinformation: those phenomena associated with the obtaining of information through means other than the normal sensory channels (i.e., ESP), such as telepathy, precognition, and clairvoyance, i.e., using the mind to tap into the thoughts of others or to acquire present or future information about objective events in the world

These phenomena involve using the mind and/or some "field" of the body to affect other minds and inanimate objects irrespective of intervening distance or elapsed time, and without engaging any conventional tools. Bioenergetics and bioinformation are two classifications that form a single branch of science the Soviets preferred to call biocommunications. Soviet biocommunications research is primarily concerned with exploring the existence of a definite group of natural phenomena controlled by laws that are not based on any known (energetic) influence. The types of biocommunication (a.k.a. psychotronics) phenomena includes special sensory biophysical activities, brain and mind control, telepathic communications or bioinformation transceiving, bioluminescent and bioenergetic emissions, and the effects of altered states of consciousness on the human psyche. Psychotronics and remote viewing provide capabilities that have obvious intelligence applications. The Soviets and their Warsaw Pact allies invested millions of dollars in psychotronics R&D because they understood this, and saw the potential payoff for military and intelligence applications.

The U.S. response to Soviet psychotronics R&D programs was the Remote Viewing program. In addition, the U.S. Army began the JEDI Project in 1983, which sought to increase human potential using teachable models of behavioral/physical excellent by unconventional means (Alexander et al., 1990). The JEDI Project was essentially a human-performance modeling experiment based on neuro-linguistic programming (NLP) skills, whereby advanced influence technologies to model excellence in human performance was used. The program ran under the auspices of the Army INSCOM and the Organizational Effectiveness School, and was sponsored by a U.S. government interagency task force. Finally, it should be pointed out that the program had successfully trained several hundred people, including members of Congress (such as Al Gore, Jr. and Tom Downey), before being terminated.

There is a wealth of factual scientific research data from around the world attesting to the physical reality of p-Teleportation and related anomalous psi phenomena (Mitchell, 1974b; Targ and Puthoff, 1977; Nash, 1978; Radin, 1997; Tart et al., 2002). The skeptical reader should not be so quick to dismiss the subject matter in this chapter, because one must remain open-minded about this subject and consider p-Teleportation as worthy of further scientific exploration. The psychotronics topic is controversial within the western scientific community. The debate among scientists and scientific philosophers is highly charged at times, and becomes acrimonious to the point where reputable skeptical scientists cease being impartial by refusing to examine the experimental data or theories, and they prefer to bypass rational discourse by engaging in ad hominem attacks and irrational "armchair" arguments.

P-Teleportation and related phenomena are truly anomalous, and they challenge accepted modern scientific paradigm. Lightman and Gingerich (1991) wrote, "Scientists are reluctant to change paradigms for the purely psychological reasons that the familiar is often more comfortable than the unfamiliar and that inconsistencies in belief are uncomfortable." And theories change over time when anomalies enter the picture. Anomalies are particularly helpful for they point to the inadequacies of an old model and point the way to a new one. Anomalous scientific facts are unexpected and difficult to explain within an existing conceptual framework. Kuhn (1970) describes scientific discovery as a complex process, in which an anomalous fact of nature is recognized, and then followed by a change in conceptual framework (i.e., paradigm) that makes the new fact no longer an anomaly. Kuhn stated that, "Discovery commences with the awareness of anomaly, that is, with the recognition that nature has somehow violated the pre-induced expectations that govern normal science." This statement neatly describes exactly what transpired during the historical revolution that took place in physics between the classical mechanics/electrodynamics age in the 19th century and the quantum/atomic/nuclear/relativistic age in the 20th century. And this isn't the only time in human history that scientific paradigms have dramatically changed. The discovery of p-Teleportation already commenced in the 20th century, so let us continue the discovery and create a new physics paradigm for the 21st century.

5.1.1 Hypothesis Based on Mathematical Geometry

The Chinese researchers reported in their teleportation experiments that high-speed photography/videotaping recorded test specimens physically “melding” or blending with the walls of sealed containers, and in a different series of experiments the test specimens would simply disappear from inside the container only to reappear at another location (after seconds to several minutes of time transpired). They also reported in the series of radio micro-transmitter experiments that there were large fluctuations in the intensity (in both amplitude and frequency) of the monitored signal to the effect that it would either completely disappear or become extremely weak (to the extent that the monitoring instruments could scarcely detect it); and they discovered that there was a definite correlation between the change in strength (i.e., radical frequency shifts were observed) of the monitored radio signal and the teleportation of the radio micro-transmitter, such that the weak or absent signal indicated that the specimen was “nonexistent” (or in an altered physical state) during teleportation. This data is important because without the aid of electronic monitoring instruments, the average person’s sensory organs and usual methods of detection are temporarily unable to perceive the test specimen’s (ambiguous) existence during the teleportation process. This data offers an important clue on what the teleportation mechanism is.

It is beyond the scope of this study to propose a complete self-consistent physics theory of consciousness/mind, which explains how the mind can activate p-Teleportation and related psychotronics phenomena. This topic has been under study in recent decades by a legion of medical science, bio- and neuro-physiology, psychology, mathematics, philosophy, and physics experts. Many different theories with varying degree of theoretical maturity and self-consistency have been proposed over the years, and most of them have not yet been experimentally tested for various reasons. However, some first-order experimental work has been done (Mitchell, 1974b; Targ and Puthoff, 1977; Wolman et al., 1986; Radin, 1997; Tart et al., 2002). Ironically, quantum mechanics theory, and the related physics of quantum entanglement and teleportation, has become the primary focus of all of the physics theories of consciousness/psychotronics that have been recently proposed (see for example, Shan, 2003). Wolman et al. (1986) and Radin (1997) provide a review and discussion on recent theories and experiments that are based on quantum physics theory (see also, Walker, 1974; Targ and Puthoff, 1977; Mitchell, 1999, and the references cited therein; Tart et al., 2002). It appears that the physics of q-Teleportation (Chapter 3) has tremendous relevance to the physics of p-Teleportation and psychotronics.

In the following I propose a parsimonious first-order hypothesis that can explain the gross features of both the Chinese p-Teleportation data and the other reported p-Teleportation phenomena. But I will refrain from including any role that might be played by quantum phenomena since the scientific community has not yet settled that particular issue. (However, it is apparent that quantum theory and quantum phenomena will likely play a key role in a formal physics theory of PK and psychotronics.)

First-Order Hypothesis:

- ☐ Fact 1: The mature discipline of mathematical geometry developed the properties of higher dimensional spaces (Reichenbach, 1957; Manning, 1977; Rucker, 1977). An example of one such property that is of relevance to the hypothesis: One can visualize a four-dimensional world by using color as the 4th dimension. We can think of a three-dimensional world, whereby objects pass through one another if their colors (i.e., four-dimensional locations) are different (Reichenbach, 1957). For example, color can be used as a 4th dimension to see how a knot in three-dimensions can be untied in a 4th spatial dimension without moving the ends of the cord. That is because a cord cannot stay knotted in four-dimensional space, because the extra degree of freedom will cause any knot to slip through itself. Two other interesting and relevant examples are that the links of a chain may be separated unbroken in the 4th dimension, and a flexible sphere may be turned inside out without tearing in the 4th dimension (Manning, 1977; Rucker, 1977).

- ☐ Proposition 1 and Fact 2: It has been proposed that our space actually possesses a slight four-dimensional hyperthickness, so that the ultimate components of our nervous system are actually higher dimensional, thus enabling the human mind/brain to imagine four-dimensional space (Hinton, 1888, 1904; Rucker, 1977). If this is the case, then the three-dimensional nets of neurons that code thoughts in our brain may form four-dimensional patterns to achieve four-dimensional thought. The "bulk" space in 3-brane theory (see Section 4.1), and experimental data from the Remote Viewing program (see Section 5.1), provide support for this concept. Can we see into the 4th dimension and have four-dimensional thoughts? Yes, we can. Proof (see, Rucker, 1977, 1984): If you look at a Necker cube for a while, it spontaneously turns into its mirror image and back again. If you watch it do this often enough, the twinkling sort of motion from one state to the other begins to seem like a continuous motion. But this motion can only be continuous if it is a rotation in four-dimensional space. The mathematician August F. Möbius discovered in 1827 that it is in fact possible to turn a three-dimensional solid object into its mirror image by an appropriate rotation through four-dimensional space (a.k.a. hyperspace rotation). Thus, it is actually possible for our minds to perform such a rotation. Therefore, we can actually produce four-dimensional phenomenon in our minds, so our consciousness is four-dimensional. Rucker (1984) shows another dramatic example of being able to see into the 4th dimension via a "Neck-A-Cube."

- ☐ Fact 3: Another property of higher dimensional geometry (Reichenbach, 1957; Rucker, 1977, 1984) is that one can move through solid three-dimensional obstacles without penetrating them by passing in the direction of the 4th (spatial) dimension. The 4th dimension is perpendicular to all of our normal three-dimensional space directions, and so our three-dimensional enclosures have no walls against this direction.

- ☐ Conclusion and Hypothesis: Therefore, the results of the Chinese p-Teleportation experiments can simply be explained as a human consciousness phenomenon that somehow acts to move or rotate test specimens through a 4th spatial dimension, so that the specimens are able to penetrate the solid walls/barriers of their containers without physically breaching them. No real dematerialization/rematerialization of the specimens takes place. The intensity fluctuations of the radio micro-transmitter specimen's electromagnetic signal, and the apparent blending of the other specimens with the walls of their containers, represent the passage of the specimens through a 4th spatial dimension. During teleportation the radio signals emitted by the micro-transmitter became weak/non-existent and fluctuated, because they were spreading out into the 4th dimension and became undetectable in our three-dimensional space. The weak signals that were ("barely") detected represent the leakage of a portion of the radio signal back into our three-dimensional space from the 4th dimension during teleportation. The observed blending of the other specimens with the walls of their containers is how the movement/rotation of the specimens through the 4th dimension was visually interpreted by the mind (along the lines of the Necker cube or Neck-A-Cube examples).

5.2 Conclusion and Recommendations

We will need a physics theory of consciousness and psychotronics, along with more experimental data, in order to test the hypothesis in Section 5.1.1 and discover the physical mechanisms that lay behind the psychotronic manipulation of matter. P-Teleportation, if verified, would represent a phenomenon that could offer potential high-payoff military, intelligence and commercial applications. This phenomenon could generate a dramatic revolution in technology, which would result from a dramatic paradigm shift in science. Anomalies are the key to all paradigm shifts!

•**Recommendations**:

- There are numerous supporters within the U.S. military establishment who comprehend the significance of remote viewing and PK phenomenon, and believe that they could have strategic implications. Bremseth (2001), a U.S. Navy SEAL, attended the Marine War College and studied the Remote Viewing program, and interviewed many of the former program participants. Bremseth then wrote his thesis on the topic, and concluded that the evidence supported continued research and applications of remote viewing. A research program improving on and expanding, or implementing novel variations of, the Chinese and Uri Geller-type experiments should be conducted in order to generate p-Teleportation phenomenon in the lab. The performances and characteristics of p-Teleportation need to be delineated in order to develop a refined hypothesis. Such a program should be designed so that an operational model for p-Teleportation can be developed and implemented as a prototype. An experimental program similar in fashion to the Remote Viewing program should be funded at $900,000 – 1,000,000 per year in parallel with a theoretical program funded at $500,000 per year for an initial five-year duration. The role of quantum physics theory and related quantum phenomena (i.e., entanglement and teleportation) in p-Teleportation and psychotronics should be explored in this program (see for example, the Biological Quantum Teleportation recommendation in Section 3.3). An experiment definition study should be conducted first to identify and propose the best experiments for this program, which should be funded at $80,000 for one year.

6.0 REFERENCES

1. Aczel, A. D. (2002), Entanglement: The Greatest Mystery in Physics, Four Walls Eight Windows Press, New York
2. Aharonov, Y., Reznik, B. and Stern, A. (1998), "Quantum limitations of superluminal propagation," Phys. Rev. Lett., 81, 2190-2193
3. Aharonov, Y. and Albert, D. (1981), "Can we make sense of the measurement process in relativistic quantum mechanics?," Phys. Rev. D, 24, 359-370
4. Alexander, Col. J. B. (2003), Winning The War: Advanced Weapons, Strategies, And Concepts For The Post-9/11 World, St. Martin's Press, New York, pp. 238 – 244
5. Alexander, J. B. (1996), "Uri's Impact on the U.S. Army," posted on http://www.urigeller.com
6. Alexander, Lt. Col. J. B. (1980), "The New Mental Battlefield: 'Beam Me Up, Spock,'" Military Review, vol. XL, no. 12
7. Alexander, Col. J. B., Groller, Maj. R. and Morris, J. (1990), The Warrior's Edge, W. Morrow Co., New York
8. Ambjϕrn, J. and Wolfram, S. (1983), "Properties of the Vacuum. I. Mechanical and Thermodynamic," Annals Phys., 147, 1-32
9. Antoniadis, I., Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (1998), "New dimensions at a millimeter to a fermi and superstrings at a TeV," Phys. Lett. B, 436, 257-263
10. Appelquist, T., Chodos, A. and Freund, P. G. O. eds. (1987), Modern Kaluza-Klein Theories, Addison-Wesley, Menlo Park
11. Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (2002), "Large Extra Dimensions: A New Arena for Particle Physics," Physics Today, 55, 35-40
12. Arkani-Hamed, N., Dimopoulos, S. and Dvali, G. (1998), "The hierarchy problem and new dimensions at a millimeter," Phys. Lett. B, 429, 263-272
13. Arkani-Hamed, N., Dimopoulos, S., Kaloper, N. and Dvali, G. (2000), "Manyfold universe," Journal of High Energy Physics (JHEP online physics papers), http://jhep.sissa.it/archive/papers/jhep122000010/jhep122000010.pdf0012
14. Aspect, A. (1983), Trois tests expérimentaux des inégalités de Bell par mesure de corrélation de polarisation de photons, Ph.D. thesis No. 2674, Université de Paris-Sud, Centre D'Orsay
15. Aspect, A., Dalibard, J. and Roger, G. (1982a), "Experimental Tests of Bell's Inequalities Using Time-Varying Analyzers," Phys. Rev. Letters, 49, 1804-1807
16. Aspect, A., Grangier, P. and Roger, G. (1982b), "Experimental Realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A New Violation of Bell's Inequalities," Phys. Rev. Letters, 49, 91
17. Aspect, A. and Grangier, P. (1985), Lettere al Nuovo Cimento, 43, 345
18. Aspelmeyer, M., et al. (2003), "Long-Distance Free-Space Distribution of Quantum Entanglement," Science, 301, 621 - 623
19. Banghui, W. (1990), "Evidence of the Existence of Abnormal States of Matter," Chinese J. Somatic Sci., First Issue, 36 [translated into English by the Defense Intelligence Agency]
20. Barcelo, C. and Visser, M. (2002), "Twilight for the energy conditions?," Int. J. Mod. Phys. D, 11, 1553
21. Barnum, H., Caves, C., Fuchs, C., Jozsa, R. and Schumacher, B. (1996), "Noncommuting Mixed States Cannot Be Broadcast," Phys. Rev. Lett., 76, 2818-2821
22. Bell, J. S. (1964), "On the Einstein Podolsky Rosen Paradox," Physics, 1, 195
23. Bennett, C. H., et al. (1993), "Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels," Phys. Rev. Lett., 70, 1895-1899
24. Bennett, C. H. and Wiesner, S. J. (1992), "Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states," Phys. Rev. Lett., 69, 2881-2884

25. Bennett, G. L., Forward, R. L. and Frisbee, R. H. (1995), "Report on the NASA/JPL Workshop on Advanced Quantum/Relativity Theory Propulsion," AIAA-95-2599, 31st AIAA/ASME/ASE/ASEE Joint Propulsion Conference and Exhibit, San Diego, CA
26. Birrell, N. D. and Davies, P. C. W. (1982), Quantum fields in curved space, Cambridge University Press, Cambridge
27. Blaauboer, M., et al. (1998), "Superluminal pulse transmission through a phase conjugating mirror," Optics Communications, 148, 295-299
28. Boschi, D., et al. (1998), "Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels," Phys. Rev. Lett., 80, 1121-1125
29. Bose, S. and Home, D. (2002), "Generic Entanglement Generation, Quantum Statistics, and Complementarity," Phys. Rev. Lett., 88, 050401
30. Bose, S., Knight, P. L., Plenio, M. B. and Vedral, V. (1999), "Proposal for Teleportation of an Atomic State via Cavity Decay," Phys. Rev. Lett., 83, 5158-5161
31. Bouwmeester, D., et al. (1997), "Experimental quantum teleportation," Nature, 390, 575-579
32. Bowen, W. P., et al. (2003), "Experimental investigation of continuous variable quantum teleportation," Phys. Rev. A, 67, 032302
33. Bowen, W. P., Treps, N., Schnabel, R. and Lam, P. K. (2002), "Experimental demonstration of continuous variable polarization entanglement," Phys. Rev. Lett., 89, 253601
34. Brassard, G., Braunstein, S. and Cleve, R. (1998), "Teleportation as a quantum computation," Physica D, 120, 43-47
35. Braunstein, S. (1996), "Quantum teleportation without irreversible detection," Proc. Royal Acad., 53, 1900-1903
36. Braunstein, S., Fuchs, C., Kimble, H. and van Loock, P. (2001), "Quantum versus classical domains for teleportation with continuous variables," Phys. Rev. A, 64, 022321
37. Braunstein, S. and Kimble, J. (1998), "Teleportation of continuous quantum variables," Phys. Rev. Lett., 80, 869-872
38. Bremseth, Cmdr. L. R. (2001), Unconventional Human Intelligence Support: Transcendent and Asymmetric Warfare Implications of Remote Viewing, Thesis, Marine War College
39. Buttler, W., et al. (1998), "Practical free-space quantum key distribution over 1 km," Phys. Rev. Lett., 81, 3283-3286
40. Center for Quantum Computation Web Site: http://www.qubit.org
41. Chan, H. B., et al. (2001), "Quantum Mechanical Actuation of Microelectromechanical Systems by the Casimir Force," Science, 291, 1941-1944
42. Chiao, R. and Steinberg, A. (1998), "Quantum optical studies of tunneling and other superluminal phenomena," Physica Scripta, T76, 61-66
43. Chown, M. (1990), "Can photons travel 'faster than light'?," New Scientist, 126, 32
44. CIA In-Q-Tel Web Site information: http://www.cia.gov/cia/publications/inqtel/
45. Cole, D. C. and Puthoff, H. E. (1993), "Extracting Energy and Heat from the Vacuum," Phys. Rev. E, 48, 1562
46. Davies, P. C. W. (1980), Other Worlds, Dent, London
47. Davis, E. W. (1999a), Research Summary Report #1 to Dr. Hal Puthoff, IASA: Brief Summary of "Lorentzian Wormholes From The Gravitationally Squeezed Vacuum", NASA Research Center Online Library – Interstellar Studies (available from the author)
48. Davis, E. W. (1999b), Research Summary Report #2 to Dr. Hal Puthoff, IASA: Brief Summary of "Gravitational Vacuum Polarization. Parts I – IV", NASA Research Center Online Library – Interstellar Studies (available from the author)
49. de Felice, F. (1971), "On the gravitational field acting as an optical medium," Gen. Rel. Grav., 2, 347-357
50. de Oliveira, E. C. and Rodriguez, W. (1998), "Superluminal electromagnetic waves in free space," Annalen der Physik, 7(7-8), 654-659

51. de Sabbata, V. and Schmutzer, E. eds. (1983), Unified Field Theories of More Than 4 Dimensions, Proc. Int'l School of Cosmology and Gravitation (Erice), World Scientific, Singapore
52. Deutsch, D. (1998), The Fabric of Reality, Penguin Books
53. DeWitt, B. S. and Graham, N. eds. (1973), The Many Worlds Interpretation of Quantum Mechanics, Princeton University Press, Princeton
54. DeWitt, B. S. (1970), Physics Today, 23, 30
55. DIA Report (1978), Paraphysics R&D – Warsaw Pact, Defense Intelligence Agency, Report No. DST-1810S-202-78, DIA Task No. PT-1810-18-76, Washington DC (authors' names redacted)
56. Dicke, R. H. (1961), "Mach's principle and equivalence," in Proc. of the Int'l School of Physics "Enrico Fermi" Course XX, Evidence for Gravitational Theories, ed. C. Møller, Academic Press, New York, pp. 1-49
57. Dicke, R. H. (1957), "Gravitation without a principle of equivalence," Rev. Mod. Phys., 29, 363-376
58. Ding, Y. J. and Kaplan, A. E. (1992), "Nonlinear Magneto-Optical Effect in Vacuum: Inhomogeneity-Originated Second-Harmonic Generation in DC Magnetic Field," J. Nonl. Opt. Phys., 1, 51-72
59. Ding, Y. J. and Kaplan, A. E. (1989), "Nonlinear Magneto-Optics of Vacuum: Second-Harmonic Generation," Phys. Rev. Lett., 63, 2725-2728
60. Drummond, I. J. and Hathrell, S. J. (1980), "QED vacuum polarization in a background gravitational field and its effect on the velocity of photons," Phys. Rev. D, 22, 343-355
61. Dür, W. and Briegel, H.-J. (2003), "Entanglement purification for Quantum Computation," Phys. Rev. Lett., 90, 067901
62. Einstein, A., Podolsky, B. and Rosen, N. (1935), "Can quantum mechanical description of physical reality be considered complete?," Phys. Rev., 47, 777-780
63. Evans, J., Nandi, K. and Islam, A. (1996a), "The Optical-Mechanical Analogy in General Relativity: New Methods for the Paths of Light and of the Planets," Am. J. Phys., 64, 1401-1415
64. Evans, J., Nandi, K. and Islam, A. (1996b), "The Optical-Mechanical Analogy in General Relativity: Exact Newtonian Forms for the Equations of Motion of Particles and Photons," Gen. Rel. Grav., 28, 413-439
65. Everett, H. (1957), Rev. Mod. Phys., 29, 454
66. Forward, R. L. (2001), Personal Communication, Salt Lake City, UT
67. Forward, R. L. (1999), Personal Communication, Los Angeles, CA
68. Forward, R. L. (1998), "Apparent Method for Extraction of Propulsion Energy from the Vacuum," AIAA-98-3140, 34th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, Cleveland, OH
69. Forward, R. L. (1996), Mass Modification Experiment Definition Study, PL-TR-96-3004, Phillips Laboratory-Propulsion Directorate, Air Force Materiel Command, Edwards AFB, CA
70. Forward, R. L. (1984), "Extracting electrical energy from the vacuum by cohesion of charged foliated conductors," Phys. Rev. B, 30, 1770-1773
71. Freedman, S. J. and Clauser, J. F. (1972), "Experimental Test of Local Hidden-Variable Theories," Phys. Rev. Lett., 28, 938-941
72. Friedman, J. et al. (1990), "Cauchy problem in spacetimes with closed timelike curves," Phys. Rev. D, 42, 1915-1930
73. Furusawa, A., et al. (1998), "Unconditional quantum teleportation," Science, 282, 706-710
74. Furuya, K., et al. (1999), "Failure of a proposed superluminal scheme," Phys. Lett. A, 251, 294-296
75. Geller, U. (1975), Uri Geller: My Story, Praeger Publ., New York
76. Giroldini, W. (1991), "Eccles's Model of Mind-Brain Interaction and Psychokinesis: A Preliminary Study," J. Sci. Explor., 5, no. 2

77. Gisin, N. (1990), “Weinberg’s non-linear quantum mechanics and superluminal communications,” Phys. Lett. A, 143, 1-2
78. Gisin, N., Scarani, V., Tittel, W. and Zbinden, H. (2000), “Optical tests of quantum non-locality: from EPR-Bell tests towards experiments with moving observers,” Ann. Phys. (Leipzig), 9, 831-841
79. Gissurarson, L. R. (1992), “The Psychokinesis Effect: Geomagnetic Influence, Age and Sex Difference,” J. Sci. Explor., 6, no. 2
80. Green, M. B. (1985), “Unification of forces and particles in superstring theories,” Nature, 314, 409
81. Greenberger, D. (1998), “If one could build a macroscopical Schrodinger cat state, one could communicate superluminally,” Physica Scripta, T76, 57-60
82. Groller, Capt. R. (1987), “Soviet Psychotronics – a Closer Look,” Military Intelligence, PB 34-87-1 (Test), pp. 43-44
83. Groller, Capt. R. (1986), “Soviet Psychotronics – a State of Mind,” Military Intelligence, 12, no. 4, pp. 18-21, 58
84. Hagley, E., et al. (1997), “Generation of Einstein-Podolsky-Rosen Pairs of Atoms,” Phys. Rev. Lett., 79, 1-5
85. Hald, J., Sørensen, J. L., Schori, C. and Polzik, E. S. (1999), “Spin Squeezed Atoms: A Macroscopic Entangled Ensemble Created by Light,” Phys. Rev. Lett., 83, 1319-1322
86. Hartle, J. B. and Hawking, S. W. (1983), “Wave function of the Universe,” Phys. Rev. D, 28, 2960-2975
87. Hasted, J. B. (1979), “Paranormal Metal Bending,” in The Iceland Papers: Selected Papers on Experimental and Theoretical Research on the Physics of Consciousness, Puharich, A. ed., Research Associates Publ., Amherst, WI
88. Hasted, J. B., Bohm, D., Bastin, E. W. and O’Reagan, B. (1975), “Scientists confronting the paranormal,” Nature, 254, 470-472
89. Haugan, M. P. and Will, C. M. (1977), “Principles of equivalence, Eötvös experiments, and gravitational red-shift experiments: The free fall of electromagnetic systems to post—post-Coulombian order,” Phys. Rev. D, 15, 2711-2720
90. Hawking, S. W. and Ellis, G. F. R. (1973), The Large-Scale Structure of Space-Time, Cambridge Univ. Press, Cambridge, pp. 88-91 and 95-96
91. Hegerfeldt, G. (1998), “Instantaneous spreading and Einstein causality in quantum theory,” Annalen der Physik, 7(7-8), 716-725
92. Heitler, W. (1954), The Quantum Theory of Radiation (3rd ed.), Oxford University Press, London, p. 113.
93. Herrmann, F. (1989), “Energy Density and Stress: A New Approach to Teaching Electromagnetism”, Am. J. Phys., 57, 707-714
94. Hinton, C. H. (1904), The Fourth Dimension, Sonnenschein, London
95. Hinton, C. H. (1888), A New Era of Thought, Sonnenschein, London
96. Hochberg, D. and Kephart, T. W. (1991), “Lorentzian wormholes from the gravitationally squeezed vacuum,” Phys. Lett. B, 268, 377-383
97. Hong, C. K. and Mandel, L. (1985), “Theory of parametric frequency down conversion of light,” Phys. Rev. A, 31, 2409-2418
98. Houck, J. (1984a), “Surface Change During Warm-Forming,” Archaeus, 2, no. 1
99. Houck, J. (1984b), “PK Party History,” Psi Research, 3, no. 1
100. Houck, J. (1982), “PK Party Format,” unpublished paper
101. IBM Press Release (2001), “IBM’s Test-Tube Quantum Computer Makes History,” http://www.research.ibm.com/resources/news/20011219_quantum.shtml
102. Jahn, R. G. and Dunne, B. J. (1987), Margins of Reality: The Role of Consciousness in the Physical World, Harcourt Brace Jovanovich, New York

103. Jammer, M. (1974), The Philosophy of Quantum Mechanics, Wiley-Interscience, New York, pp. 507-521
104. Jennewein, T., Simon, C., Weihs, G., Weinfurter, H. and Zeilinger, A. (2000), "Quantum Cryptography with Entangled Photons," Phys. Rev. Lett., 84, 4729-4732
105. Jinggen, H., Xinghai, Y. and Laijing, S. (1990), "Investigation into the 'Force' in Parapsychological Writing," Chinese J. Somatic Sci., First Issue, 32 [translated into English by the Defense Intelligence Agency]
106. Julsgaard, B., Kozhekin, A. and Polzik, E. S. (2001), "Experimental long-lived entanglement of two macroscopic objects," Nature, 413, 400-403
107. Kaku, M. (1994), Hyperspace: A Scientific Odyssey Through Parallel Universes, Time Warps, and the 10th Dimension, Anchor Books-Doubleday, New York
108. Kaku, M. (1993), Quantum Field Theory, Oxford University Press, New York
109. Kaku, M. (1988), Introduction to Superstrings, Springer-Verlag, New York
110. Kaluza, T. (1921), "Unitätsproblem der Physik," Sitz. Preuss. Akad. Wiss. Phys. Math., K1, 966
111. Kaplan, A. E. and Ding, Y. J. (2000), "Field-gradient-induced second-harmonic generation in magnetized vacuum," Phys. Rev. A, 62, 043805-(1-9)
112. Kim, Y.-H., Kulik, S. P. and Shih, Y. (2001), "Quantum Teleportation of a Polarization State with a Complete Bell State Measurement," Phys. Rev. Lett., 86, 1370-1373
113. Klein, O. (1926), "Quantentheorie und fünfdimensionale Relativitätstheorie," Zeits. Phys., 37, 895
114. Kongzhi, S., Xianggao, L. and Liangzhong, Z. (1990), "Research into Paranormal Ability to Break Through Spatial Barriers," Chinese J. Somatic Sci., First Issue, 22 [translated into English by the Defense Intelligence Agency]
115. Kuhn, T. S. (1970), The Structure of Scientific Revolutions, 2nd ed., Univ. of Chicago Press, Chicago
116. Kwiat, P. G., et al. (1995), "New high-intensity source of polarization-entangled photon pairs," Phys. Rev. Lett., 75, 4337-4341
117. Kwiat, P. G., et al. (1999), "Ultrabright source of polarization-entangled photons," Phys. Rev. A., 60, R773-R776
118. Lamas-Linares, A., Howell, J. C. and Bouwmeester, D. (2001), "Stimulated emission of polarization-entangled photons," Nature, 412, 887-890
119. Lamoreaux, S. K. (1997), "Measurement of the Casimir Force Between Conducting Plates," Phys. Rev. Letters, 78, 5-8
120. LaMothe, Capt. J. D. (1972), Controlled Offensive Behavior – USSR, Defense Intelligence Agency, Report No. ST-CS-01-169-72, DIA Task No. T72-01-14, Washington DC
121. Latorre, J. I., Pascual, P. and Tarrach, R. (1995), "Speed of light in non-trivial vacua," Nucl. Phys. B, 437, 60-82
122. Lee, H. C. ed. (1984), An Introduction to Kaluza-Klein Theories, Proc. Chalk River Workshop on Kaluza-Klein Theories, World Scientific, Singapore
123. Lee, T. D. (1988), Particle Physics and Introduction to Field Theory, Harwood Academic Press, London
124. Leggett, A. J. (1999), "Quantum Theory: Weird and Wonderful," Physics World, 12, 73-77
125. Leshan, C. (2002), "Proposal for Teleportation by Help of Vacuum Holes," in Gravitation and Cosmology: From the Hubble Radius to the Planck Scale, Proc. of a Symposium in Honour of the 80th Birthday of Jean-Pierre Vigier, Amoroso, R. L., Hunter, G., Kafatos, M. and Vigier, J.-P. eds., Kluwer Academic Publ., Boston, pp. 515-516
126. Leshan, C. (1999), "Thought Experiment to the Border of Universe," J. Theoretics, 1, no.4
127. Li, L.-X. and Gott, J. R. (1998), "Self-Consistent Vacuum for Misner Space and the Chronology Protection Conjecture," Phys. Rev. Lett., 80, 2980-2983
128. Lightman, A. P. and Gingerich, O. (1991), "When Do Anomalies Begin?," Science, 255, 690-695

129. Lightman, A. P. and Lee, D. L. (1973), "Restricted proof that the weak equivalence principle implies the Einstein equivalence principle," Phys. Rev. D, 8, 364
130. Maierle, C., Lidar, D. and Harris, R. (1998), "How to teleport superpositions of chiral amplitudes," Phys. Rev. Lett., 81, 869-872
131. Maire, L. F. and LaMothe, Capt. J. D. (1975), Soviet and Czechoslovakian Parapsychology Research, Defense Intelligence Agency, Report No. DST-1810S-387-75, DIA Task No. PT-1810-12-75, Washington DC
132. Mandel, L. and Wolf, E. (1995), Optical Coherence and Quantum Optics, Cambridge University Press
133. Manning, H. P. (1977), The Fourth Dimension Simply Explained, Peter Smith Publ., Gloucester, MA
134. Mavromatos, N. E., Mershin, A. and Nanopoulos, D. V. (2002), "QED-Cavity model of microtubules implies dissipationless energy transfer and biological quantum teleportation," http://arxiv.org/abs/quant-ph/0204021
135. McConnell, A. J. (1957), Applications of Tensor Analysis, Dover Publ., New York, pp. 163-217
136. Mead, F. B. and Nachamkin, J. (1996), "System for Converting Electromagnetic Radiation Energy to Electrical Energy," United States Patent No. 5,590,031
137. Milonni, P. W. (1994), The Quantum Vacuum: An Introduction to Quantum Electronics, Academic Press, NY
138. Mitchell, E. D. (1999), "Nature's Mind: the Quantum Hologram," National Institute for Discovery Science, Las Vegas, NV, http://www.nidsci.org/articles/naturesmind-qh.html
139. Mitchell, E. D. (1974a), "Appendix: Experiments with Uri Geller," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 683-686
140. Mitchell, E. D. (1974b), Psychic Exploration: A Challenge for Science, White, J. ed., G. P. Putnam's Sons, New York
141. Mittelstaedt, P. (2000), "What if there are superluminal signals?," Eur. Phys. J. B**,** 13, 353-355
142. Mittelstaedt, P. (1998), "Can EPR-correlations be used for the transmission of superluminal signals?," Annalen der Physik, 7(7-8), 710-715
143. Mittelstaedt, P. and Nimtz, G. eds. (1998), "Workshop on Superluminal Velocities," Annalen der Physik, 7(7-8), 591-592
144. Morris, M. S. and Thorne, K. S. (1988), "Wormholes in spacetime and their use for interstellar travel: A tool for teaching general relativity", Am. J. Phys., 56, 395-412
145. Mourou, G. A., Barty, C. P. J. and Perry, M. D. (1998), "Ultrahigh-Intensity Lasers: Physics Of The Extreme On A Tabletop," Physics Today, 51, 22-28
146. Naik, D. S., Peterson, C. G., White, A. G., Berglund, A. J. and Kwiat, P. G. (2000), "Entangled state quantum cryptography: Eavesdropping on the Ekert protocol," Phys. Rev. Lett., 84, 4733
147. Nash, C. B. (1978), Science of PSI: ESP and PK, C. C. Thomas Publ., Springfield, Ill.
148. Nielsen, M. A. (2003), "Simple Rules for a Complex Quantum World," Sci. Am., 13, 25-33
149. Nielsen, M. A. and Chuang, I. L. (2000), Quantum Computation and Quantum Information, Cambridge University Press
150. Nielsen, M., Knill, E. and Laflamme, R. (1998), "Complete quantum teleportation using nuclear magnetic resonance," Nature, 396, 52-55
151. Nimtz, G. (1998), "Superluminal signal velocities," Annalen der Physik, 7(7-8), 618-624
152. Opatrný, T., Clausen, J., Welsch, D.-G. and Kurizki, G. (2000), "Squeezed-Vacuum Assisted Quantum Teleportation," Paper No. 7thCEWQO/015, in Proc. 7th Central-European Workshop on Quantum Optics, Hungary
153. Opatrný, T. and Kurizki, G. (2001), "Matter-Wave Entanglement and Teleportation by Molecular Dissociation and Collisions," Phys. Rev. Lett., 86, 3180-3183
154. Overduin, J. M. and Wesson, P. S. (1998), "Kaluza-Klein Gravity," http://arxiv.org/abs/gr-qc/9805018

155. Pan, J.-W., et al. (1998), "Experimental entanglement swapping," Phys. Rev. Lett., 80, 3891-3894
156. Pease, R. (2001), "Brane new world," Nature, 411, 986-988
157. Peres, A. (2000), "Classical intervention in quantum systems. II. Relativistic invariance," Phys. Rev. A, 61, 022117(8)
158. Perry, M. D. (2000), "The Amazing Power of the Petawatt," Science & Technology Rev. (LLNL-DoE publication), March issue, 4-12
159. Perry, M. D. (1996), "Crossing the Petawatt Threshold," Science & Technology Rev. (LLNL-DoE publication), December issue, 4-11
160. Polchinski, J. (1995), "Dirichlet Branes and Ramond-Ramond Charges," Phys. Rev. Lett., 75, 4724-4727
161. Pratt, J. G. (1986), "Soviet Research in Parapsychology," in Handbook of Parapsychology, Wolman, B. B., Dale, L. A., Schmeidler, G. R. and Ullman, M. eds., McFarland and Co. Publ., Jefferson, NC, pp. 883-903
162. Preskill, J., Lecture Notes: http://www.theory.caltech.edu/people/preskill/ph229/
163. Puthoff, H. E. (2003), Personal Communication, Institute for Advanced Studies at Austin, Austin, TX
164. Puthoff, H. E. (2002a), "Polarizable-Vacuum (PV) Approach to General Relativity", Found. Phys., 32, 927-943
165. Puthoff, H. E. (2002b), "Polarizable-Vacuum Approach to General Relativity", in Gravitation and Cosmology: From the Hubble Radius to the Planck Scale, eds. R. L. Amoroso, G. Hunter, M. Kafatos, and J.-P. Vigier, Kluwer Academic Publ., Dordrecht, the Netherlands, pp. 431-446
166. Puthoff, H. E. (1999a), "Polarizable-vacuum (PV) representation of general relativity," http://arxiv.org/abs/gr-qc/9909037
167. Puthoff, H. E. (1999b), Personal Communication, Institute for Advanced Studies at Austin, Austin, TX
168. Puthoff, H. E. (1996), "CIA-Initiated Remote Viewing Program at Stanford Research Institute," J. Sci. Explor., 10, 63-76
169. Puthoff, H. E. (1993), "On the Feasibility of Converting Vacuum Electromagnetic Energy to Useful Form," Int'l Workshop on the Zeropoint Electromagnetic Field, Cuernavaca, Mexico
170. Puthoff, H. E. (1990), "The Energetic Vacuum: Implications for Energy Research," Spec. in Sci. & Technology, 13, 247
171. Puthoff, H. E., Little, S. R. and Ibison, M. (2002), "Engineering the Zero-Point Field and Polarizable Vacuum for Interstellar Flight," J. British Interplanetary Soc., 55, 137-144
172. Puthoff, H. E. and Targ, R. (1975), "Physics, Entropy and Psychokinesis," in Proc. Conf. Quantum Physics and Parapsychology (Geneva, Switz.), Parapsychology Foundation Publ., New York
173. Puthoff, H. E. and Targ, R. (1974), "PK experiments with Uri Geller and Ingo Swann," in Research in Parapsychology 1973, Roll, W. G., Morris, R. L. and Morris, J. D. eds., Scarecrow Press, Metuchen, New Jersey, pp. 125-128
174. Quantum Information: Special Issue (1998), Physics World, 11, no. 3
175. Radin, D. (1997), The Conscious Universe: The Scientific Truth of Psychic Phenomena, HarperEdge-HarperCollins Publ., New York
176. Raimond, J. M., Brune, M. and Haroche, S. (2001), "Colloquium: Manipulating quantum entanglement with atoms and photons in a cavity," Rev. Mod. Phys., 73, 565-582
177. Randall, L. and Sundrum, R. (1999a), "Large Mass Hierarchy from a Small Extra Dimension," Phys. Rev. Lett., 83, 3370-3373
178. Randall, L. and Sundrum, R. (1999b), "An Alternative to Compactification," Phys. Rev. Lett., 83, 4690-4693
179. Rarity, J. G. (2003), "Getting Entangled in Free Space," Science, 301, 604 - 605
180. Reichenbach, H. (1957), The Philosophy of Space and Time, Dover Publ., New York
181. Rhine, L. E. (1970), Mind over Matter: Psychokinesis, Macmillan, New York

182. Rubakov, V. A. and Shaposhnikov, M. E. (1983a), "Do we live inside a domain wall?," Phys. Lett. B, 125, 136-138
183. Rubakov, V. A. and Shaposhnikov, M. E. (1983b), "Extra space-time dimensions: Towards a solution to the cosmological constant problem," Phys. Lett. B, 125, 139-143
184. Rucker, R. (1984), The Fourth Dimension: A Guided Tour of the Higher Universes, Houghton Mifflin Co., Boston, pp. 45-49
185. Rucker, R. (1977), Geometry, Relativity and the Fourth Dimension, Dover Publ., New York
186. Sackett, C. A. (2001), Quant. Inf. Comput., 1, 57
187. Sackett, C. A., et al. (2000), "Experimental entanglement of four particles," Nature, 404, 256
188. Scarani, V., Tittel, W., Zbinden, H. and Gisin, N. (2000), "The speed of quantum information and the preferred frame: analysis of experimental data," Phys. Lett. A, 276, 1-7
189. Scharnhorst, K. (1990), "On Propagation of Light in the Vacuum Between Plates," Phys. Lett. B, 236, 354-359
190. Schein, F. and Aichelburg, P. C. (1996), "Traversable Wormholes in Geometries of Charged Shells," Phys. Rev. Letters, 77, 4130-4133
191. Schmidt, H. (1987), "The Strange Properties of Psychokinesis," J. Sci. Explor., 1, no. 2
192. Schmidt, H. (1974), "Psychokinesis," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 179-193
193. Schnabel, J. (1997), Remote Viewers: The Secret History of America's Psychic Spies, Dell Publ., New York
194. Schrödinger, E. (1980), Proc. Am. Philos. Soc., 124, 323
195. Schrödinger, E. (1935a), Die Naturwissenschaften, 48, 807
196. Schrödinger, E. (1935b), Die Naturwissenschaften, 49, 823
197. Schrödinger, E. (1935c), Die Naturwissenschaften, 49, 844
198. Seife, C. (2000), "'Spooky Action' Passes a Relativistic Test," Science, 287, 1909-1910
199. Shan, G. (2003), "A Primary Quantum Model of Telepathy," http://cogprints.ecs.soton.ac.uk/archive/00003065/
200. Shigemi, S., Yasuo, O. and Akihira, T. (1978), "Some Observations with Scanning Electron Microscope (SEM) of the Fracture Surface of Metals Fractured by Psychokinesis," Japan PS Soc. J., 2, no. 2
201. Shor, P. W. (1997), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," SIAM J. Sci. Statist. Comput., 26, 1484
202. Shor, P. W. (1994), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer", in Proc. 35th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, p. 124
203. Shoup, R. (2002), "Anomalies and Constraints: Can Clairvoyance, Precognition, and Psychokinesis Be Accommodated within Known Physics?," J. Sci. Explor., 16, no. 1
204. Shuhuang, L., et al. (1981), "Some Experiments on the Transfer of Objects Performed by Unusual Abilities of the Human Body," Nature Journal (Peoples Republic of China), 4, no. 9, 652 [Defense Intelligence Agency Requirements and Validation Branch, DIA Translation LN731-83, Intelligence Information Report No. 6010511683 (1983)]
205. Siegfried, T. (2000), The Bit and the Pendulum, John Wiley & Sons
206. Sørensen, J. L. (1998), Nonclassical light for atomic physics and quantum teleportation, Ph.D. thesis, Univ. of Aarhus
207. Srikanth, R. (July 1999), "Noncausal superluminal nonlocal signaling," http://arxiv.org/abs/quant-ph/9904075
208. Stanford, R. (1974), "Interview," Psychic, 7
209. Stenholm, S. and Bardroff, P. (1998), "Teleportation of N-dimensional states," Phys. Rev. A, 58, 4373-4376

210. Swann, I. (1974), "Scientological Techniques: A Modern Paradigm for the Exploration of Consciousness and Psychic Integration," in Proc. First Int'l Conf. on Psychotronic Research, United States Joint Publications Research Service, Document No. JPRS L/5022-1, Virginia
211. Targ, R. (1996), "Remote Viewing at Stanford Research Institute in the 1970s: A Memoir," J. Sci. Explor., 10, 77-88
212. Targ, R. and Puthoff, H. E. (1977), Mind-Reach: Scientists Look at Psychic Ability, Jonathan Cape Ltd.-Anchor Press, London
213. Tart, C. T., Puthoff, H. E. and Targ, R. eds. (2002), Mind at Large: Institute of Electrical and Electronics Engineers Symposia on the Nature of Extrasensory Perception, Hampton Roads Publ. Co., Charlottesville, VA
214. Terhal, B. M., Wolf, M. M. and Doherty, A. C. (2003), "Quantum Entanglement: A Modern Perspective," Physics Today, 56, 46-52
215. Thorne, K. S. (1993), "Closed Timelike Curves," GRP-340, CalTech, Pasadena, CA
216. Tittel, W. and Weihs, G. (2001), Quantum Inf. Comput., 1, 3
217. Tittel, W., Brendel, J., Zbinden, H. and Gisin, N. (2000), "Quantum Cryptography Using Entangled Photons in Energy-Time Bell States," Phys. Rev. Lett., 84, 4737-4740
218. Tittel, W., Brendel, J., Gisin, B., Herzog, T., Zbinden, H. and Gisin, N. (1998a), "Experimental demonstration of quantum correlations over more than 10 km," Phys. Rev. A, 57, 3229-3232
219. Tittel, W., Brendel, J., Zbinden, H. and Gisin, N. (1998b), "Violation of Bell Inequalities by Photons More Than 10 km Apart," Phys. Rev. Lett., 81, 3563-3566
220. Vaidman, L. (1994), "Teleportation of quantum states," Phys. Rev. A, 49, 1473-1476
221. Vaidman, L. and Yoran, N. (1999), "Methods for reliable teleportation," Phys. Rev. A, 59, 116-125
222. Vallee, J. (1997), Personal Communication, Science Advisory Board of the National Institute for Discovery Science, Las Vegas, NV
223. Vallee, J. (1990), Confrontations: A Scientist's Search for Alien Contact, Ballantine Books, New York
224. Vallee, J. (1988), Dimensions: A Casebook of Alien Contact, Ballantine Books, New York
225. van Enk, S. (March 1998), "No-cloning and superluminal signaling," http://arxiv.org/abs/quant-ph/9803030
226. Visser, M., Kar, S. and Dadhich, N. (2003), "Traversable Wormholes with Arbitrarily Small Energy Condition Violations," Phys. Rev. Lett., 90, 201102
227. Visser, M. (1997), Personal Communication, Washington University, St. Louis, MO
228. Visser, M. (1995), Lorentzian Wormholes: From Einstein to Hawking, AIP Press, New York
229. Visser, M. (1990), "Wormholes, baby universes, and causality", Phys. Rev. D, 41, 1116-1124
230. Visser, M. (1989), "Traversable wormholes: Some simple examples", Phys. Rev. D, 39, 3182-3184
231. Volkov, A. M., Izmest'ev, A. A. and Skrotskii, G. V. (1971), "The propagation of electromagnetic waves in a Riemannian space," Sov. Phys. JETP, 32, 686-689
232. Walker, E. H. (1974), "Consciousness and Quantum Theory," in Psychic Exploration: A Challenge for Science, Mitchell, E. D., White, J. ed., G. P. Putnam's Sons, New York, pp. 544-568
233. Weihs, G., Jennewein, T., Simon, C., Weinfurter, H. and Zeilinger, A. (1998), "Violation of Bell's Inequality under Strict Einstein Locality Conditions," Phys. Rev. Lett., 81, 5039-5043
234. Weinberg, S. (1992), Dreams of a Final Theory, Vintage Books, pp. 88-89
235. Weinberg, S. (1989), "Testing Quantum Mechanics," Ann. Phys., 194, 336-386
236. Weiss, P. (2000), "Hunting for Higher Dimensions: Experimenters scurry to test new theories suggesting that extra dimensions are detectable," Science News, 157, 122-124
237. Westmoreland, M. and Schumacher, B. (March 1998), "Quantum entanglement and the non existence of superluminal signals," http://arxiv.org/abs/quant-ph/9801014
238. Wheeler, J. A. (1962), Monist, 47, 40

239. Wheeler, J. A. (1957), Rev. Mod. Phys., 29, 463
240. Wheeler, J. A. and Zurek, W. H. eds. (1983), Quantum Theory and Measurement, Princeton University Press
241. Will, C. M. (1993), Theory and Experiment in Gravitational Physics (rev. ed.), Cambridge University Press, Cambridge, Section 2.6
242. Will, C. M. (1989), "The confrontation between gravitation theory and experiment," in General Relativity: An Einstein Centenary Survey, eds. S. W. Hawking and W. Israel, Cambridge University Press, Cambridge, Chapter 2
243. Will, C. M. (1974), "Gravitational red-shift measurements as tests of nonmetric theories of gravity," Phys. Rev. D, 10, 2330-2337
244. Wilson, H. A. (1921), "An electromagnetic theory of gravitation," Phys. Rev., 17, 54-59
245. Wineland, D. J., et al. (2002), "Quantum information processing with trapped ions," http://arxiv.org/abs/quant-ph/0212079
246. Wolf, F. A. (1988), Parallel Universes: The Search for Other Worlds, Simon and Schuster, New York
247. Wolman, B. B., Dale, L. A., Schmeidler, G. R. and Ullman, M. eds. (1986), Handbook of Parapsychology, McFarland and Co. Publ., Jefferson, NC
248. Wootters, W. K. and Zurek, W. H. (1982), "A single quantum cannot be cloned," Nature, 299, 802-803
249. Zbinden, H., Brendel, J., Gisin, N. and Tittel, W. (2000a), "Experimental Test of Non-Local Quantum Correlation in Relativistic Configurations," http://arxiv.org/abs/quant-ph/0007009
250. Zbinden, H., Brendel, J., Tittel, W. and Gisin, N. (2000b), "Experimental Test of Relativistic Quantum State Collapse with Moving Reference Frames," http://arxiv.org/abs/quant-ph/0002031
251. Zeilinger, A. (2003), "Quantum Teleportation," Sci. Am, 13, 34-43
252. Zhang, T. C., et al. (2002), "Quantum teleportation of light beams," http://arxiv.org/abs/quant-ph/0207076
253. Zubairy, S. (1998), "Quantum teleportation of a field state," Phys. Rev. A, 58, 4368-4372

APPENDIX A – A Few Words About Negative Energy

A.1 A General Relativistic Definition of Negative or Exotic Energy

We saw in equations (2.10a-c) that the surface energy and stress-tension densities of the material required to create and thread a traversable wormhole must be "negative." For surface stress-energy, and volume stress-energy in general, this is "negative" in the sense that the material we must deploy to generate and thread the traversable wormhole must have an energy density (ρc^2, ρ = mass density) that is less than the stress-energy density (τ), or we can write this condition as: *mass-energy* $\rho c^2 \leq$ *stress-energy* τ. On the basis of this condition, we call this material property "exotic." Therefore, the term "negative" is just a misnomer in this context. The condition for ordinary, non-exotic forms of matter that we are all familiar with is *mass-energy* $\rho c^2 >$ *stress-energy* τ. This condition represents one version of what is variously called the weak (WEC), null (NEC), average (AEC), dominant (DEC), strong (SEC) or "standard" energy conditions (that are mere hypotheses!), which allegedly forbid negative mass-energy density and gravitational repulsion (antigravity) between material objects to occur in nature. Hawking and Ellis (1973) formulated these energy conditions in order to establish a series of mathematical proofs in their study of the application of general relativity theory to cosmology and black hole physics.

However, there are general theorems of differential geometry that guarantee that there must be NEC violations (meaning exotic matter-energy is present) at a wormhole throat (Visser, 1997). In view of this, it is known that static radial electric or magnetic fields are borderline exotic when threading a wormhole, if their tension were infinitesimally larger, for a given energy density (Herrmann, 1989; Hawking and Ellis, 1973). Other exotic (energy condition violating) matter-energy fields are known to be squeezed quantum states of the electromagnetic field and other squeezed quantum fields (see Section A.2 for the discussion on squeezed quantum states), gravitationally squeezed vacuum electromagnetic zero-point energy (see Section 2.3 for the discussion on Gravitationally Squeezed Vacuum Energy), Casimir (electromagnetic zero-point) energy and other quantum fields/states/effects. These examples represent forms of matter-energy that possess negative energy density. Since the vacuum is defined to have vanishing energy density, anything possessing less energy density than the vacuum must have a negative energy density. With respect to creating wormholes, these have the unfortunate reputation of alarming physicists. This is unfounded since all the energy condition hypotheses have been experimentally tested in the laboratory and experimentally shown to be false - 25 years before their formulation (Visser, 1990 and references cited therein). Further investigation into this technical issue showed that violations of the energy conditions are widespread for all forms of both classical and quantum matter-energy such as planets, stars, black holes, neutron stars, people, space dust clouds, etc. (Davis, 1999b; Barcelo and Visser, 2002). In addition, Visser (1995) showed that all (generic) spacetime geometries violate all the energy conditions. Violating the energy conditions commits no offense against nature.

A.2 Squeezed Quantum States and Negative Energy

In quantum mechanics the energy (E) and frequency (ν) of a quantum oscillator system, such as electromagnetic radiation (or light), are interchangeable via the Planck relation $E = h\nu$ ($h = 2\pi\hbar$). And from the Heisenberg quantum uncertainty principle, we know that the conjugate variable to the frequency is the oscillator phase (φ), such that $\Delta\nu\Delta\varphi \geq \hbar$ is obeyed. Phase is difficult to measure and is ignored in characterizing complex quantum systems.

Recent theoretical and experimental work has shown that in many quantum systems the limits to measurement precision imposed by the quantum vacuum zero-point fluctuations (ZPF) can be breached by decreasing the frequency noise at the expense of increasing the phase noise (thus maintaining $\Delta\nu\Delta\varphi \geq \hbar$), while at the same time the variations in frequency, and therefore the energy, are reduced below the ZPF such that the energy becomes "negative." "Squeezing" is thus the control of quantum fluctuations and corresponding uncertainties, whereby one can squeeze the variance of one (physically important) observable quantity provided the variance in the (physically unimportant) conjugate variable is stretched/increased. The squeezed quantity possesses an unusually low variance, meaning less variance than would be expected on the basis of the equipartition theorem. We can exploit quantum squeezing to extract energy from one place in the ordinary vacuum at the expense of piling up excess energy elsewhere (Morris and Thorne, 1988).

Appendix B – *TH*εμ Methodology

In the formalism of the *TH*εμ methodology, the functions T and H are introduced by requiring that the Lagrangian for the motion of particles (with charge e_a and mass m_{0a} for the a^{th} particle), under the joint action of gravity and the electromagnetic field A_α ($\alpha \equiv$ spacetime vector components), be expressed in the canonical form

$$L = \sum_a \int \left(-m_{0a}\sqrt{T - H{v_a}^2} + e_a A_\alpha {v_a}^\alpha \right) dt + \left(8\pi\right)^{-1} \int \left(\varepsilon \mathbf{E}^2 + \mu^{-1}\mathbf{B}^2 \right) d^3x\, dt \qquad \text{(B.1);}$$

where the arbitrary functions T, H, ε, and μ are functions of the metric (a.k.a. gravitation field), v_a^α is the a^{th} particle four-vector velocity, and A_α is the electromagnetic field four-vector potential, ***E*** and ***B*** are the electric and magnetic field strengths, and (B.1) is in geometrodynamic natural units ($\hbar = c_0 = G = \varepsilon_0 = \mu_0 = 1$). The Lagrangian characterizes the motion of charged particles in an external gravitational field by the two functions T and H, and characterizes the response of the electromagnetic fields to the external gravitational field by the two functions ε and μ. For all standard (metric) theories of gravity, the four functions are related by

$$\varepsilon = \mu = \sqrt{\frac{H}{T}} \qquad \text{(B.2);}$$

and ***every*** metric theory of gravity satisfies this relation, such that the Einstein Equivalence Principle is satisfied.

AFRL-PR-ED-TR-2003-0034
Primary Distribution of this Report:

AFRL/PRSP (15 CD)
Dr. Frank Mead
10 E. Saturn Blvd
Edwards AFB CA 93524-7680

AFRL/PRSA (1 CD)
Dr. Jean-Luc Cambier
10 E. Saturn Blvd.
Edwards AFB CA 93524-7680

AFRL/PR (1 CD)
Dr. Alan Garscadden
1950 Fifth Street
Building 18
Wright-Patterson AFB, OH 45433-7251

AFRL/PR Technical Library (2 CD + 1 HC)
6 Draco Drive
Edwards AFB CA 93524-7130

Chemical Propulsion Information Agency (1 CD)
Attn: Tech Lib (Dottie Becker)
10630 Little Patuxent Parkway, Suite 202
Columbia MD 21044-3200

Defense Technical Information Center
(1 Electronic Submission via STINT)
Attn: DTIC-ACQS (Pat Mawby)
8725 John J. Kingman Road, Suite 94
Ft. Belvoir VA 22060-6218

Dr. Eric Davis (10 CD)
4849 San Rafael Ave.
Las Vegas, NV 89120

Dr. Dana Andrews (1 CD)
P.O. Box 36
Leavenworth, WA 98826

Dr. Jim Degnan (1 CD)
AFRL/DEHP
Kirtland AFB, NM 87117

Dr. Greg Benford (1 CD)
Physics Department
University of California
Irvine, CA 92717

Dr. Jim Benford (1 CD)
Microwave Sciences, Inc.
1041 Los Arabis Ln.
Lafayette, CA 94549

Dr. Gary L. Bennett (1 CD)
7517 West Devonwood Dr.
Boise, ID 83703

Dr. Mitat Birkan (1 CD)
AFOSR/NA
801 N. Randolf St.
Arlington, VA 22203

Dr. Jon Campbell (1 CD)
P.O. Box 295
Harvest, AL 35749

Mr. Michael Libeau
NSWCDD
17320 Dahlgren Rd.
Attn: Code G23 Libeau
Dalgren VA 22448

Dr. Phil Carpenter (1 CD)
US Dept. of Energy
Oak Ridge National Laboratory
P.O. Box 2008, MS: 6269
Oak Ridge, TN 37831

Dr. Brice N. Cassenti (1 CD)
Pratt & Whitney Aircraft
400 Main Street – MS: 163-07
East Hartford, CT 06108

Dr. Chan K. Choi (1 CD)
Purdue University
School of Nuclear Engineering
West Lafayette, IN 47907

Dr. Terry Kammash (1 CD)
University of Michigan
Nuclear Engineering Dept.
Ann Arbor, MI 48109

Dr. Ingrid Wysong (1 CD)
EOARD
223-321-Old Merylebone Rd.
London NW1 5th
United Kingdom

Dr. Robert Frisbee (1 CD)
JPL, MS 125-109
4800 Oak Grove Dr.
Pasadena, CA 91109

Dave Froning (1 CD)
Flight Unlimited
5450 Country Club
Flagstaff, AZ 86004

Geroge D. Hathaway (1 CD)
Hathaway Consulting Services
39 Kendal Ave.
Toronto, Canada, Ontario
Canada M5R 1L5

Clark W. Hawk, Director (1 CD)
Propulsion Research Center
University of Alabama in Huntsville
5000 Technology Drive, TH S-266
Huntsville, AL 35899

Alan C. Holt (1 CD)
NASA/Johnson Space Center
Code OD
Houston, TX 77058

Dr. Steven Howe (1 CD)
19 Karen Lane
Los Alamos, NM 87544

Mike Kaiserman (1 CD)
Raytheon Missile Systems Company
Bldg 805, M/S C3
Tucson, AZ 85734

Dr. George Miley (1 CD)
University of Illinois, Dept. of Nuclear Engr.
214 Nuclear Engineering Laboratory
103 South Goodwin Ave.
Urbana, IL 61801

Marc Millis (1 CD)
NASA Glenn Research Center
M.S. SPTD-2
21000 Brookpark Road, MS: 86-2
Cleveland, OH 44135

Dr. Jordon Kare (1 CD)
222 Canyon Lakes Pl.
San Ramon, CA 94583

Ron J. Kita (1 CD)
87 Shady Springs DR.
Doyelstown, PA 18901

Dr. Gerald L. Kulcinski (1 CD)
Nuclear Engineering Dept.
University of Wisconsin
1500 Johnson Dr.
Madison, WI 53706

Dr. Geoffrey A. Landis (1 CD)
Sverdrup Technology
21000 Brookpark Rd., MS 302-1
Cleveland, OH 44135

Dr. Michael LaPointe (1 CD)
NASA Lewis Research Center, MS: SPTD-1
21000 Brookpark Rd.
Cleveland, OH 44135

Dr. Sheldon Meth (1 CD)
DARPA
Tactical Technology Office
3701 N. Fairfax Dr.
Arlington, VA 22203

Dr. Michael M. Micci (1 CD)
Prof. of Aerospace Engineering
233 E. Hammond Bldg.
University Park, PA 16802

Dr. Hal Puthoff (1 CD)
Institute for Advanced Studies
4030 Braker Lane, West
Suite 300
Austin, TX 78759

Dr. Eric E. Rice (1 CD)
Orbital Technologies Corp.
402 Gammon Place, Suite 10
Madison, WI 53719

Dr. Aurthur Morrish (1 CD)
DARPA/ATO
3701 N. Fairfax Dr.
Arlington, VA 22203

Dr. Paul Murad (1 CD)
Sr. Analyst, Director for Intel Production
Missile & Space Intel Center
Defense Intelligence Agency
Washington, DC 20340-6054

Dr. Brian Palaszewski (1 CD)
NASA Glenn Research Center
21000 Brookpartk Road, MS: 5-10
Cleveland, OH 44135

Dr. Alan Pike (1 CD)
DSAS
1988 Crescent Park Drive
Reston, VA 20190

Dr. Dennis Pelaccio (1 CD)
SAIC
8100 Shaffer Parkway, Suite 100
Littleton, CO 80127

Ben Plenge (1 CD)
101 W. Eglin Blvd
Suite 342
Eglin AFB, FL 32542-6810

Dr. James Powell (1 CD)
Plus Ultra Technologies, Inc.
25 East Loop Rd.
Stony Brook, NY 11970-3350

Mr. Charles A. Yost (1 CD)
Electric Spacecraft Journal
73 Sunlight Drive
Leicester, NC 28748

Dr. George Schmidt (1 CD)
NASA HQ
300 E. Street SW
Washington, DC 20546

Steve Squires (1 CD)
Directorate of Applied Technology
Test and Simulation
STEWS-DATTS-OO
WSMR, NM 88002

Robert Talley (1 CD)
Topaz 2000, Inc
3380 Sheridan Dr.
Suite 172
Amherst, NY 14226

Dr. Kenneth D. Ware (1 CD)
Defense Nuclear Agency
Simulation Technology
6801 Telegraph Road
Alexandria, VA 22310

Dr. Feiedwardt Winterberg (1 CD)
University of Nevada
Desert Research Institute
Reno, NV 89507

Dr. Young Bae (1 CD)
1101 Bryan Ave.
Suite C
Tustin, CA 92780

Dr. Thomas M York (1 CD)
1215 Inverary Pl.
State College, PA 16801

Dr. Robert J. Barker (1CD)
AFOSR/NE
801 N. Randolf St.
Arlington, VA 22203

Time Travel: Time Dilation

Daniel Buzzo
University of the West of England
Department of Computer Science and Creative Technology,
Coldharbour Lane,
Bristol, UK
daniel.buzzo@uwe.ac.uk

In this paper I discuss the theory and actuality of the effects of Time Dilation, as predicted by Einstein's Special Relativity. Alongside this the research investigates the visible, experiential effects of most personal form of Time Travel from a personal, phenomenological perspective, illustrated through digital media.

The body of work, this paper and an accompanying book of modified digital images, recalls the Hafele-Keating experiment of October 1971 and parallels the Quantified Self experiments of diary photography and personal analytics in a holistic, interconnected way.

The work tracks 12 months of air travel and illustrates with practical data the physics and mathematics behind the theory and actuality of Time Dilation. The research combines the interdisciplinary worlds of Computer Arts with Data and Scientific Visualisation to create a tangible collision between the visualisation of the personal and illustration of theory

The observations are documented in ethnographic, photo essay illustrations as an investigation into the mathematics and physics of personal and interpersonal time. Using situated imaging presented as a practice-based body of work to communicate the abstract world of very small time periods and the direct relevance of Einstein's work on the personal and perceptual world we inhabit. Contrasting Newtonian mechanics and Einsteinian space-time the work seeks to illustrate the personal nature of how the reality of time travel influences every aspect of the interpersonal

Time dilation. Time travel. General relativity. Special relativity. Visual thinking. Photography. Visualisation. Practice-based research. Art.

1. INTRODUCTION

In 1971 two American researchers, J.C. Hafele and R. E. Keating, conducted an experiment.

They placed two caesium atomic clocks on passenger airplanes and flew them in opposite directions around the globe to test a hypothesis first posited in Einstein's theory of special relativity.

Their aim was to measure the effects of relativity on these highly accurate devices - the predictions indicated that the eastward flying clock would lose 40ns and the westward gain 275ns. A nanosecond is equal to one billionth of a second (10^{-9} or $^{1}/_{1,000,000,000}$ s). To compare, one nanosecond is to one second as one second is to 31.7 years.

> During October, 1971, four cesium atomic beam clocks were flown on regularly scheduled commercial jet flights around the world twice, once eastward and once westward, to test Einstein's theory of relativity with macroscopic clocks. From the actual flight paths of each trip, the theory predicted that the flying clocks, compared with reference clocks at the U.S. Naval Observatory, should have lost 40+/-23 nanoseconds during the eastward trip and should have gained 275+/-21 nanoseconds during the westward trip ... Relative to the atomic time scale of the U.S. Naval Observatory, the flying clocks lost 59+/-10 nanoseconds during the eastward trip and gained 273+/-7 nanosecond during the westward trip, where the errors are the corresponding standard deviations. These results provide an unambiguous empirical resolution of the famous clock 'paradox' with macroscopic clocks. (Hafele & Keating 1972)

The results showed that compared to stationary clocks in the laboratory the eastward clock lost 59ns and the westward gained 273ns. Whilst there has been discussion of the accuracy and therefore reliability of the measurements the experiment is commonly held up as one of the first to give practical evidence of the physical effects of time dilation.

2. SPECIAL RELATIVITY AND TIME DILATION

In special relativity Einstein States that the duration of time will pass at different rates when comparing an object at rest versus an object in motion. The greater the velocity of the object in motion, the closer to the speed of light, c, the greater the effect of what is known as 'time dilation'

The effect of this phenomenon is significant at speeds close to the constant of light speed but becomes exponentially less as the object in question approaches rest.

$$t' = \frac{t}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Figure 1: *The mathematical relationship of time, velocity and the speed of light describing Time Dilation*

Unless our velocity is a substantial fraction of the constant speed of light the dilation effect (γ) is approximately 1. This, combined with our extremely limited lifespan is why we don't immediately notice time dilation at ordinary speeds. The fastest speed most humans will achieve, an airplane travelling near the speed of sound, gives a time dilation where $\gamma = 1.0000000000005$.

In other words, 1 second for a passenger on the plane would pass as 1.0000000000005 seconds for a stationary observer.

Time effectively passes more slowly for the traveller relative to the observer.

3. THE REALITY OF THE EFFECT

Whilst there is discussion of the validity of the results of the 1971 Hafele & Keating experiment, particularly around the 'corrections' made to their data at publication, their initial observations are clearly aligned with the predicted results to be expected from the mathematics for time dilation.

As Reinhardt *et al* (2007) note:

> Time dilation is one of the most fascinating aspects of special relativity as it abolishes the notion of absolute time

In their 2007 paper they investigate the effects of time dilation with even more accurate atomic clocks, adding more experimental evidence consistent with special relativity. There are also numerous other experiments, such as those into muon decay, that provide additional practical evidence that the effect described is real and measurable.

Figure 2: *Over Dorset: Duration 55 minutes*

4. IMPLICATIONS FOR EVERYDAY LIFE

The practical implications for everyday life at first appear intangible when one considers the seemingly miniscule time periods being discussed. Even taking some of the most extreme examples one can imagine being feasible, such as that of a cosmonaut living in a fast low earth orbit, circumnavigating the earth every 90 minutes gives γ of approximately 1.0000000003.

If one considers our astronaut in orbit for two years, the Time Dilation due to special relativity would give an increase in lifespan of approximately 20 milliseconds. A millisecond is a thousandth (10^{-3} or $^{1}/_{1,000}$) of a second.

This gap of 20 milliseconds, or two hundredths of a second, are suddenly well within human scale perception and understanding. This is a common measured gap in motor racing and fast sports like downhill skiing. Even the record for the 100m sprint is now is 9.58 seconds, run by Usain Bolt at the World Athletics championships in Berlin in 2009.

Scales in milliseconds are spaces and durations that even humans with our modest lifespans can perceive and understand.

As Ed Lu, Science Officer on the International Space Station writes, from orbit, in his blog post on experiments in 'Relativity':

> After our 6 months in space, we will have actually aged slightly less than everyone else on the ground because of an effect called time dilation. It isn't by much (about 0.007 seconds), but it is one side benefit of flying in space!

Though the effects of time dilation at human scale speeds are often imperceptibly small they are nonetheless real. Seeing the implications of this reflected in everyday life reveals the truly personal nature of the time travel that we undertake in everything we do.

The realisation that time dilation is a tangible and observable phenomena that affects ourselves and everything around us was given heightened poignancy by spending large amounts of time flying at altitude and high speed. Observing the curvature of our planet and seeing the dark sky at the edge of space overhead, watching the mountains and deserts as they slid effortlessly under the wings of successive aeroplanes. Being able to observe the size of the individual against the backdrop of a macro-cosmic scale all the while reflecting on the fact that to a subatomic particle travelling near the speed of light, the universe is about four weeks old and it can be crossed in a matter of months.

> Nature can produce even larger particle energies. Some particles striking the Earth's upper atmosphere have energies that exceed 2*1020 eV. If such particles are protons (with mass of about 1 GeV), their speeds would be 0.999 999 999 999 999 999 999 995 c. For them, γ is 1011. Now the age of the universe is about 13 billion years for us, but for such particles, the age of the universe would be about (13 billion years/1011), ie about a month. Such a particle could cross the visible universe in a matter of months (their time).

This twist of perspective, the deliberate 'making strange' that Shlovsky (1917) describes occurs when translating the abstract of the macrocosmic to the intimacy of the personal. The act of comparing and contrasting macro-scale physics and mathematics with recordings of human scale observations evidence reveals useful and potentially significant insights into the nature of the world we actually inhabit.

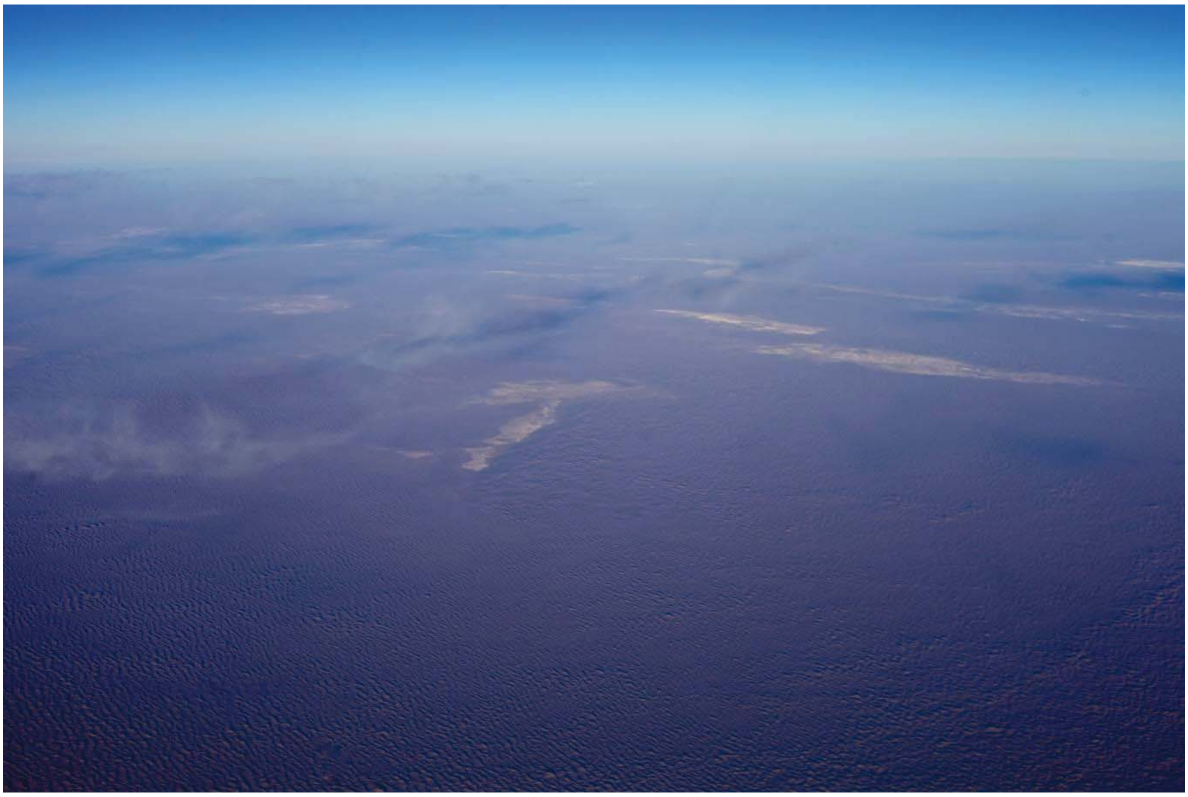

Figure 3: *Sunset over Northern China. Duration: 10 hours 50 minutes*

5. THE PICTORIAL AS PRACTICE-BASED RESEARCH

As Blevis (2011) so eloquently puts it, as an introduction to the extensive list of uses of digital imagery in design thinking.

> **Visual thinking** is the use of imagery and other visual forms to make sense of the world and to create meaningful content. Digital imagery is a special form of visual thinking, one that is particularly salient for HCI and interaction design

Alongside the descriptions of digital imagery as a *'form of information'* and a '*shared and externalised memory and cognition'* (Blevis 2012) I would argue that practice-based research, when used as an investigative tool, can assist in revealing and documenting subtle but important non-verbal evidence and also provide compelling and persuasive arguments within presentation of research findings and conclusions. The research approach taken in this area is explicitly practice-based, rooted within contemporary lens based digital media and video art. The investigative material presented as the counterpart to this research paper was collated from hundreds of minutes of digital video and several hundreds of digital still images taken while airborne on commercial passenger aircraft during January - December 2013. Taken over the course of 44 flights, totalling over 80 hours airborne the visual material is published as a 42 page visual essay, also containing text elements and descriptions taken from this academic paper.

***Figure 4**: Sunset over London. Duration: 55 minutes*

6. REALISATION OF EFFECT ON THE PERSONAL

Having recorded the number of flights, durations and distances the calculations based upon the mathematics of special relativity, as discussed earlier show that.

Given 44 flights totalling 85 hours and 20 minutes or approximately 5120 minutes flying time. gives 307,200 seconds of travel per year.

If γ = 1.0000000000005 when travelling at or near the speed of sound, gives a dilation of 0.0000001536 seconds.

To slip one second from base reference, such as the clock sat on my kitchen table I need to travel at my present annual rate for 6,510,416 Years.

Or to put it a different way, travel for a little over 260,000 years to slip one frame out of sync on a standard rate video camera.

7. CONSIDERING PHOTOGRAPHY IN EVIDENCE

When collating the increasing amounts of video and photographic material recorded whilst airborne, travelling at or near the speed of sound, the overwhelming sensation of velocity, of speed, of travel is ever-present. Viewed at a macro scale the visual material illustrates this distancing, sliding nature of objects moving in a larger framework. It supports the realisation that the terra firma so often taken as a baseline for observation is in itself only a small part of a larger perspective. The perspective that the proprioceptive nature of our observations of time, informed by the physicality of the form we have evolved and the life span perspective we consider as a norm and useful context are challenged by the linking of these measurable effects. The effects of the micro-scale variations in time seen in contrast to the macro-scale movement around the curve of the earth.

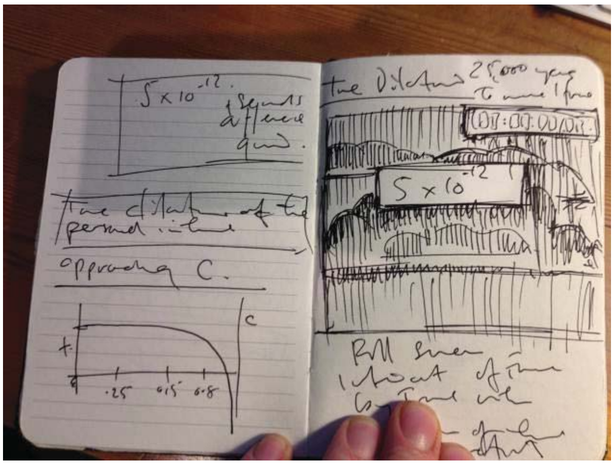
5 x 10^-12
seconds difference
time dilation of the period in time approaching C.
t
c
.25
0.5
0.8
5 x 10^-12

***Figure 5:** An airplane travelling near the speed of sound gives a time dilation where γ = 1.0000000000005*

The visual material, when presented as an adjunct to the evidence of the universal but intimately personal nature of the effects of time dilation help communicate the simple fact that we are all time travelling. Pursuing our own clearly delineated paths and durations relative to each other. Rather than there being one universal time that exists everywhere, for everyone, all at the same moment the personal implication that there is real evidence to prove that this is an illusion is both moving and compelling.

Sitting in an airline seat riding at the edge of space, looking down on the oceans, seas, cities, forests and deserts, seeing terra firma for the illusion that it is. Not a constant underlying grid of space and time - the universal yard stick and the universal clock, such as Newtonian physics promised as it replaced the Aristotelian events-make-time view of the world that had reigned before in popular consciousness.

Realising that the H&K caesium clock from 1971 could be sat beside me, ticking away in its own little timetrack, ticking to a completely separate rhythm than its counterpart clock sat on my kitchen table at home. Realising that the slipping and sliding of our lives against each other is our own intimate version of time travel, we all move forward as future becomes present becomes past but the durations we experience, and that actually occur, and when ***now*** happens is subtly different for everyone. The time we each experience actually does expand and contract as we move in space.

Figure 6: *Frost over Mongolia. Duration: 10 hours 55 minutes*

The illusion of the Newtonian ideal of a universal timescale, the idea of a 'universal now' is disappearing and being replaced by the ideas of general relativity. Of a universe full of individual 'nows' all unique and all moving independently. This change is akin to seeing the universe of unique, individual 'nows' at a human, personal scale and seeing a life travelling, sliding, back and forth in time against the 'nows' of those around one.

The important element this body of work seeks to describe is not physics or even a discussion of physics but an explanation of physics and the revelatory moment of seeing it in 'real' experienced life. This work seeks to relate that revelatory moment and articulate it with the artefacts, the book and the paper, telling the story of seeing how each and every one of us is travelling in time, independently, uniquely, every day.

> To see a World in a Grain of Sand
> And a Heaven in a Wild Flower,
> Hold Infinity in the palm of your hand
> And Eternity in an hour. (William Blake)

8. REFERENCES

Bergson, H. (1923) *The Creative Mind: An Introduction to Metaphysics*, Citadel Press [1992]

Blevis, E. (2011) Digital imagery as meaning and form in HCI and design: an introduction to the Visual Thinking Backpage Gallery. Interactions 18, 5 (September 2011), 60-65.

Blevis, E, Churchill. E, Odom. W, Pierce.J, Roedl.D,, and Wakkary.R. (2012) Visual thinking & digital imagery. In *Proc. CHI EA '12*. ACM, New York, NY, USA, 2715-2718.

Callender, C., and Edney, R. (2001). *Introducing Time.* Icon Books. ISBN 1-84046-592-1.

Hafele J.C., and Keating. R. E. (1972) Science 177, 166.

Lu. E, 'Expedition 7, Relativity' Blog Post. http://spaceflight.nasa.gov/station/crew/exp7/luletters/lu_letter13.html (retrieved 20 March 2014).

Muon Experiment in Relativity. http://hyperphysics.phy-astr.gsu.edu/hbase/relativ/muon.html (retrieved 5 March 2014).

Reinhardt, S. *et al.* (2007). Test of relativistic time dilation with fast optical atomic clocks at different velocities. Nature Physics 3 (12): 861–864.

Shklovsky, V. (1917) *Art as Technique*. In: Russian Formalist Criticism: Four Essays. Trans. Lee T. Lemon and Marion J. Reis. University of Nebraska Press, Lincoln, 1965.

The Quantified Self. http://quantifiedself.com/ (retrieved 4 December 2013).

Toothman, J. How Do Humans age in space? HowStuffWorks. http://science.howstuffworks.com/humans-age-in-space.htm (retrieved 24 April 2012).

US 20030230675A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2003/0230675 A1**

St. Clair (43) **Pub. Date: Dec. 18, 2003**

(54) ROTOR INDUCTANCE PROPULSION SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/170,847**

(22) Filed: **Jun. 12, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64G 1/40**; B64G 1/42

(52) **U.S. Cl.** **244/172**

(57) ABSTRACT

The invention is a spacecraft having a circular, domed hull around which dual electrically-charged rotors one above the other are counter-rotating on the edge of the hull. Embedded in the hull are three solenoids which create a positive vector potential at the rotors. The surface charge density times the radius times the vector potential times the area of the rotors creates an angular momentum in the vertical direction. This angular momentum produces a positive spacetime curvature over the dome of the hull and a negative spiking spacetime compression over the rotors. By machining circumferential grooves of decreasing height along the radius of the rotor, a negative surface inductance is generated. This negative inductance gradient times the negative spacetime compression time the rotor current density squared times the rotor area squared is a positive lift force on the spacecraft.

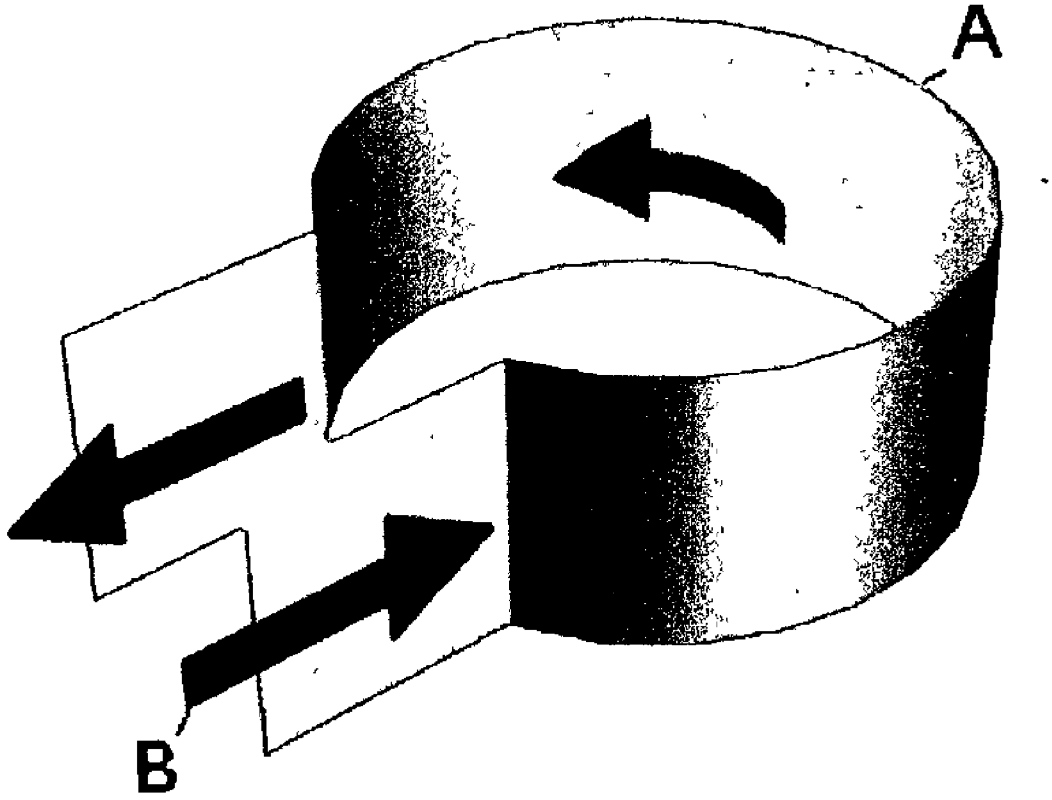

Figure 1

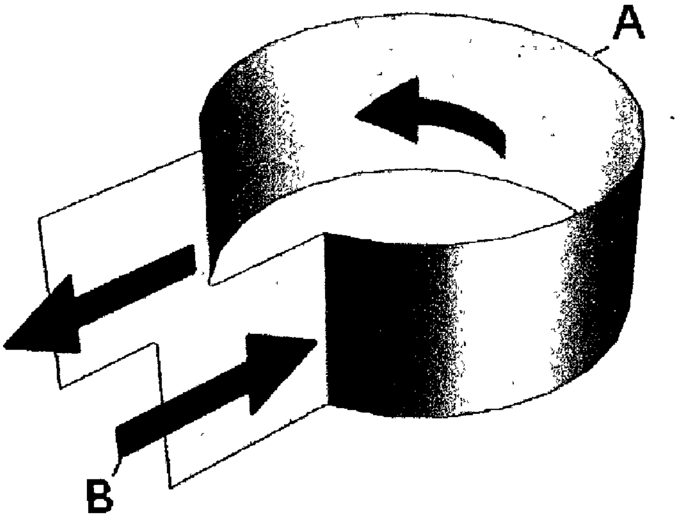

Figure 2

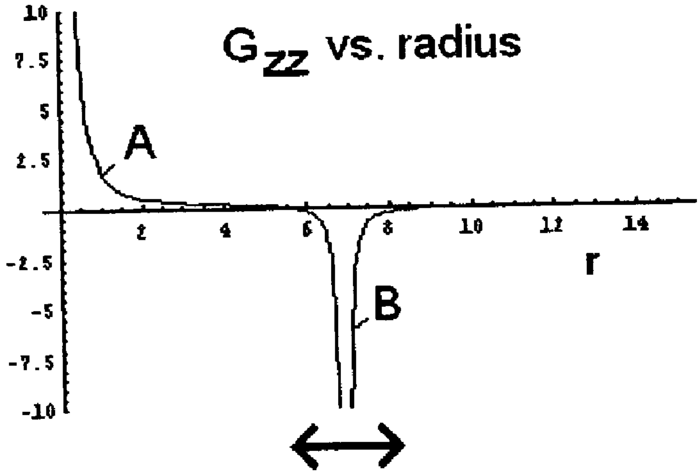

Figure 3

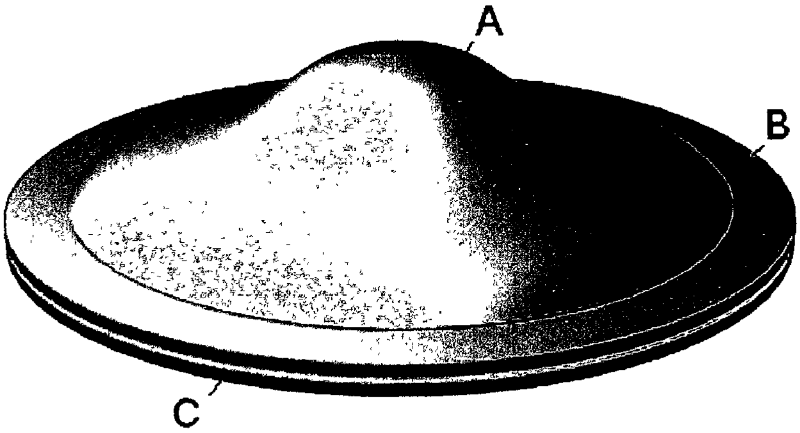

Figure 4

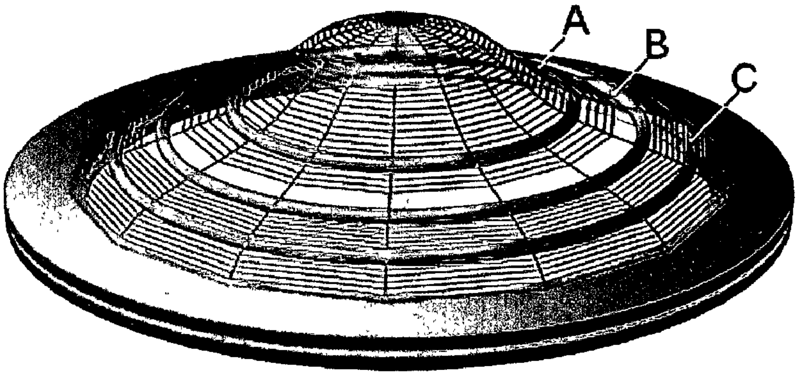

Figure 5

$$A = \frac{n I a^2}{2\varepsilon_0 c^2} \frac{1}{r'}$$

Figure 6

$$A \equiv \frac{1}{m} \frac{\text{coul}}{\text{sec}} m^2 \frac{m^2 \text{kg}\, m}{\text{coul}^2 \text{sec}^2} \frac{\text{sec}^2}{m^2} \frac{1}{m} = \frac{\text{kg}\, m}{\text{sec}\, \text{coul}}$$

Figure 7

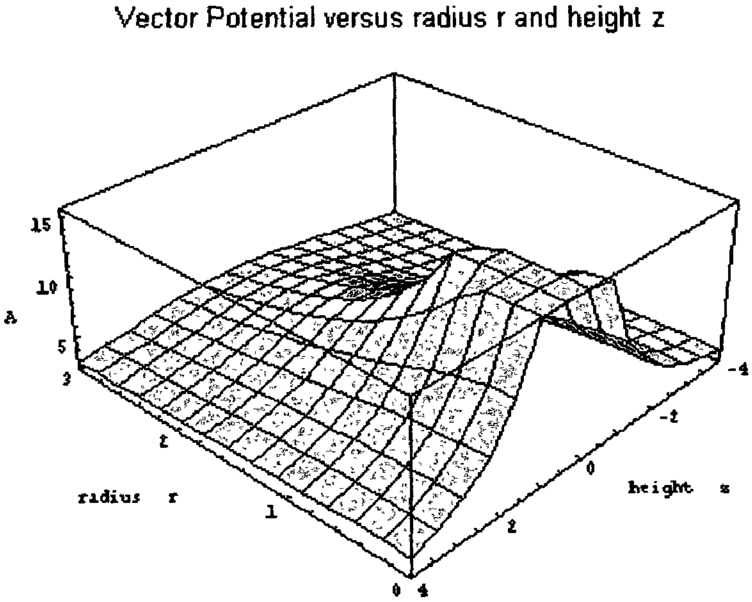

Figure 8

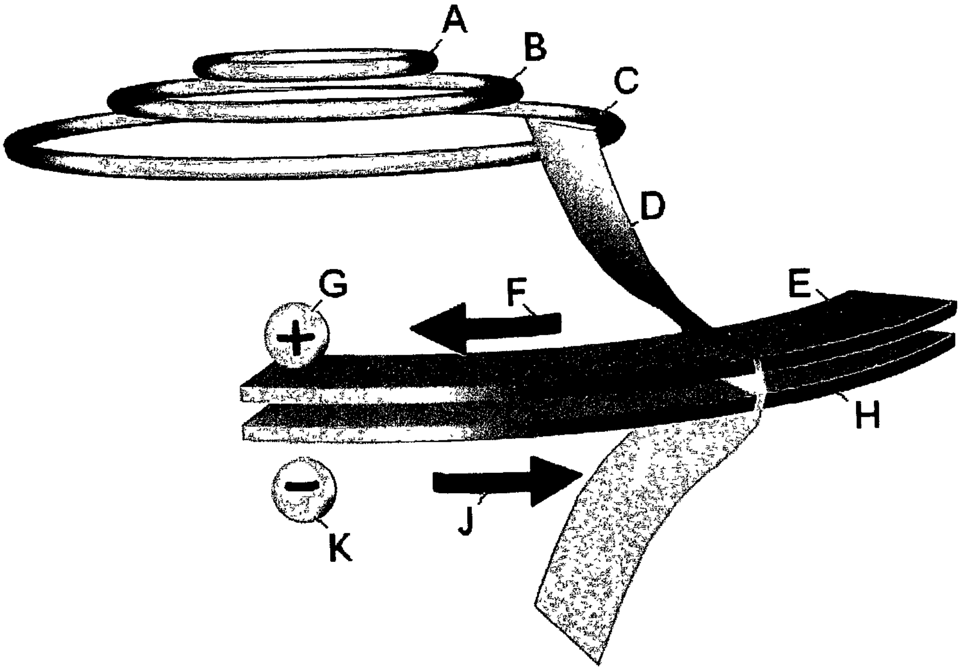
A
B
C
D
E
F
G
H
J
K
+
−

Figure 9

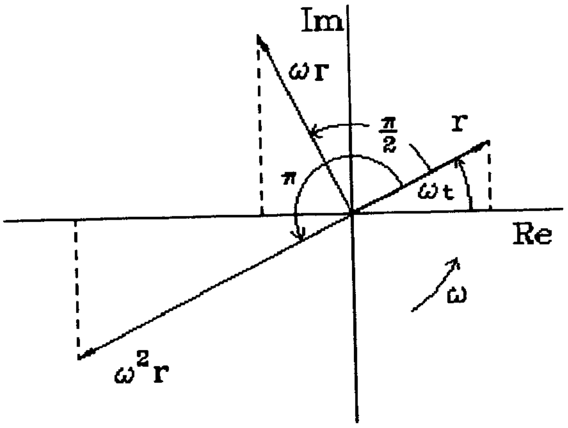
Im
ωr
$\frac{\pi}{2}$
r
π
ωt
Re
ω
ω^2r

Figure 10

$$S_k = \left[-\sigma r e^{i\omega t} + \sigma r e^{-i\omega t}\right] \frac{n I a^2}{2\varepsilon_0 c^2} \frac{n_z \text{area}}{r'}$$

$$= -\left[e^{i\omega t} - e^{-i\omega t}\right]\left[\sigma r \frac{n I a^2}{2\varepsilon_0 c^2} \frac{n_z \text{area}}{r'}\right]$$

$$= -\left[e^{i\omega t} - e^{-i\omega t}\right] K$$

Figure 11

$$S \equiv \frac{\text{coul}}{\text{m}^2}\,\text{m}\,\frac{1}{\text{m}}\frac{\text{coul}}{\text{sec}}\,\text{m}^2\,\frac{\text{m}^2}{\text{coul}^2}\frac{\text{kg}\,\text{m}}{\text{sec}^2}\frac{\text{sec}^2}{\text{m}^2}\frac{\text{m}^2}{\text{m}} = \frac{\text{kg}\,\text{m}^2}{\text{sec}}$$

Figure 12

$$e^{i\omega t} - e^{-i\omega t} = 2 i \,\text{Sin}[\omega t]$$

Figure 13

$$S_k = -2 i K \,\text{Sin}[\omega t]$$

Figure 14

$$ds^2 = \ldots - 4\frac{S}{r}dtd\theta = \ldots 8\frac{iK\,\mathrm{Sin}[\omega t]}{r}dtd\theta$$

Figure 15

$$g_{\mu\nu} = \begin{array}{c|cccc|} & t & r & \theta & z \\ \hline t & -1 & 0 & 4\frac{iK\,\mathrm{Sin}[\omega t]}{r} & 0 \\ r & 0 & 1 & 0 & 0 \\ \theta & 4\frac{iK\,\mathrm{Sin}[\omega t]}{r} & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 16

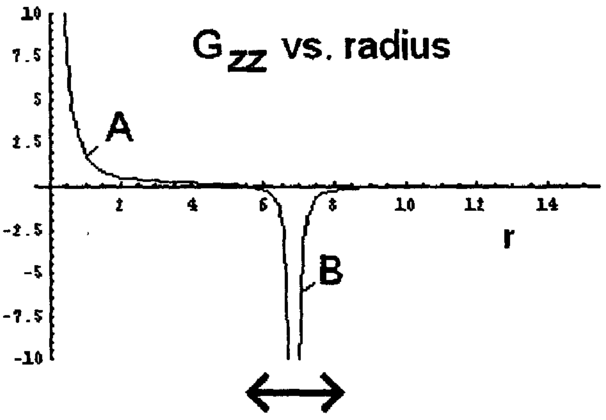

Figure 17

$$\text{Magnetic Energy} = -L\,K\,J^2\,A^2$$

Figure 18

$$\text{Magnetic Energy} \equiv -\frac{\text{kg}\,\text{m}^2}{\text{coul}^2}\frac{1}{\text{m}^2}\frac{\text{coul}^2}{\text{sec}^2\,\text{m}^2}\text{m}^4 = -\frac{\text{kg}\,\text{m}^2}{\text{sec}^2} = -\text{joule}$$

Figure 19

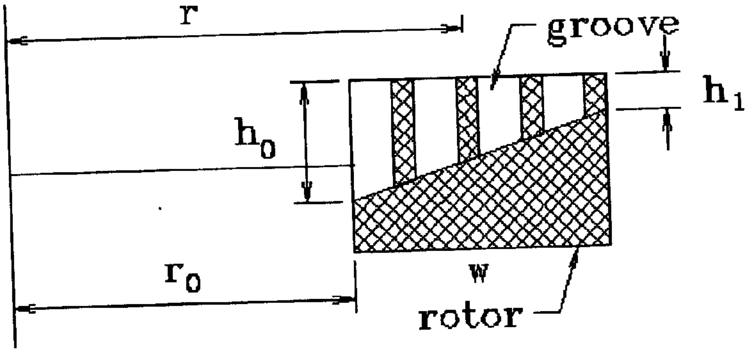

Figure 20

$$h = h_0 - (r - r_0)\left(\frac{h_0 - h_1}{W}\right)$$

Figure 21

$$\frac{dL}{dr} = \mu n^2 2\pi w \frac{\partial}{\partial r}\left(r\left(h_0 - (r - r_0)\left(\frac{h_0 - h_1}{w}\right)\right)\right)$$

$$\frac{dL}{dr} = \mu n^2 2\pi w\left(h_0 - \frac{(h_0 - h_1)(2r - r_0)}{w}\right)$$

Figure 22

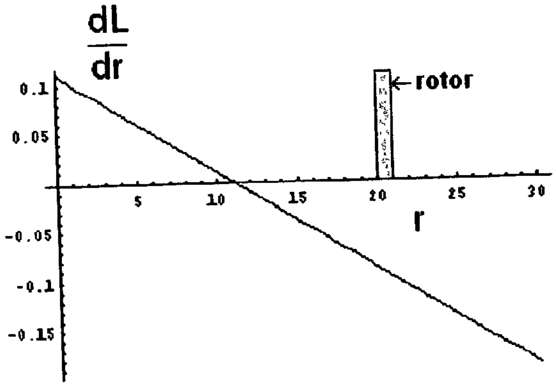

Figure 23

$$F = \frac{dL}{dr}K_c J^2 A^2 = \frac{kg\,m}{coul^2}\frac{1}{m^2}\frac{coul^2}{sec^2\,m^2}m^4 = \frac{kg\,m}{sec^2} = \text{newton}$$

ROTOR INDUCTANCE PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention, which is the object of my present application, is a spacecraft with a circular, domed hull around which are located dual electrically-charged counter-rotating rotors. The top surface of the upper rotor is etched with circular metallic grooves which give the rotor a surface inductance. The groove height decreases from the inside radius to the outside radius of the rotor giving it a radial inductance gradient. The surface charge density times the angular velocity produces a current density. The counter-rotating rotors produce a negative spacetime curvature over the rotors. The negative surface inductance gradient times the negative spiking spacetime curvature times the current density squared times the area squared is the positive lift force on the rotor.

REFERENCE WORKS

[0002] *Gravitation*, Wheeler

[0003] *Traveling-Wave Tubes*, The Bell System Technical Journal, Vol. XXIX, January 1950, No. 1, J. R. Pierce.

[0004] *The Magnetron as a Generator of Centimeter Waves*, The Bell System Technical Journal, Vol. XXV, April 1946, No. 2, J. B. Fisk, H. D. Hagstrum, P. L. Hartman.

[0005] *Lectures on Physics*, Richard Feynman

[0006] *Geometry of Electromagnetism*, Paul Hammond

BACKGROUND OF THE INVENTION

[0007] I was reading several articles about the development of the magnetron during World War II in the Bell System Technical Journal. I was trying to understand why the device resonates because it must contain a spring constant which would arise from an inductance and capacitance due to the geometry of the cavity. As given by Feynman, inductance of a solenoid is the permeability of space times the turns per length squared times the volume of the solenoid. Referring to **FIG. 1**, the magnetron cavity (A) has a circular region connected to a planar region. The electrical current flows on the sides of the cavity shown by the arrows (B). In this case, there is only one turn per height of the cavity times the volume of the cavity times the permeability which produces the inductance. The two planar regions produce a capacitance across the ends which creates the resonant frequency.

[0008] From my previous patent application Dual Rotor Propulsion System I know that the two rotors produce a current density in the angular direction along the rotor. If I spread out the magnetron cavity into a circular groove around a rotor, then the current would flow on the side walls enclosing the groove volume. The rotors also produce a spacetime curvature profile as shown in **FIG. 2**. Curve (A) is a positive spacetime curvature tension over the dome. Curve (B) is a negative spiking spacetime curvature over the rotors. The curvature is measured in inverse meter squared. So the surface inductance times the negative curvature times the current density squared times the rotor area squared is magnetic energy. The differential of the magnetic energy would be a force. So there has to be a gradient of the surface inductance. The volume element of a groove is equal to the circumferential length times the height times the width of the groove. The easiest is to vary the height with radius. Since the curvature is negative, the gradient has to be negative as well in order to get a positive lift force. Thus the height goes from large to small from the inside to the outside radius.

SUMMARY OF THE INVENTION

[0009] The invention relates to a spacecraft with a domed, circular hull of elliptical cross-section having dual electrically-charged counter-rotating rotors located one above the other on the edge of the hull. The upper rotor is positively charged and rotates clockwise with a negative angular velocity per the right-hand rule. The lower rotor is negatively charged and rotates with a positive angular velocity. The current density is the surface charge density times the velocity of the rotor. This particular combination of velocity and charge produces an angular momentum which creates a negative spiking spacetime curvature over the rotors.

[0010] The top surface of the rotor is etched or machined with circular grooves around the rotor. This creates a surface inductance which is equal to the permeability of space times the turns per length squared times the volume of the groove. In this case, there is only one turn per height of the groove. If the height of the groove decreases from one groove to the next, then there is a negative surface inductance gradient in the radial direction. So the lift force on the rotors would be the negative surface inductance gradient times the negative spacetime curvature times the current density squared times the rotor area squared.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0011] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0012] **FIG. 1**. Perspective of magnetron cavity.

[0013] **FIG. 2**. Spacetime curvature G_{zz} over hull and rotor.

[0014] **FIG. 3**. Perspective view of spacecraft with dual rotors.

[0015] **FIG. 4**. Wire frame view of three solenoids.

[0016] **FIG. 5**. Vector potential equation for solenoid.

[0017] **FIG. 6**. Units of vector potential.

[0018] **FIG. 7**. 3D graph of vector potential using three solenoids.

[0019] **FIG. 8**. Perspective of vector potential along rotors.

[0020] **FIG. 9**. Rotor mechanics diagram using exponential representation.

[0021] **FIG. 10**. Angular momentum equation due to vector potential and rotating charged rotors.

[0022] **FIG. 11**. Units of angular momentum.

[0023] **FIG. 12**. Exponential equation for twice imaginary sine of the angle.

[0024] **FIG. 13**. Angular momentum equation for g metric tensor.

[0025] **FIG. 14**. Angular momentum term for elemental line length ds^2.

[0026] **FIG. 15**. The g metric tensor containing the angular momentum.

[0027] **FIG. 16**. The spacetime curvature tension G_{zz} versus radius due to angular momentum.

[0028] **FIG. 17**. Magnetic energy equation.

[0029] **FIG. 18**. Units of magnetic energy.

[0030] **FIG. 19**. Cross-section of rotor showing groove height gradient.

[0031] **FIG. 20**. Equation for the height h of the rotor groove as a function of radius.

[0032] **FIG. 21**. Equation for the groove inductance gradient.

[0033] **FIG. 22**. Plot of the groove inductance gradient as a function of radius showing that it is negative at the rotor.

[0034] **FIG. 23**. Lift force on rotors due to inductance gradient, compression curvature, rotor current density and area.

DETAILED DESCRIPTION OF THE INVENTION

[0035] 1. Referring to **FIG. 3**, the spacecraft has a circular, domed hull of elliptical cross-section with dual electrically-charged counter-rotating rotors one above the other on the edge of the hull.

[0036] 2. Referring to **FIG. 4**, embedded within the hull are three solenoids or current loops carrying a constant electrical current in the positive sense per the right-hand rule.

[0037] 3. Referring to **FIG. 5**, Feynman has shown that the vector potential {A} of a solenoid is equal to the number of turns per length {n} times the current {I} times the radius of the solenoid {a} squared divided by half the permittivity of space {ϵ_0} times the speed of light {c} squared times the radius {r'} prime from the center of the coil to some location in space such as the rotors. It has been found by physicists Bohm and Aharanov that the vector potential field is not confined to an infinitely long solenoid as is the magnetic B field. The vector potential has units of kilogram-meter per second-coulomb as seen in **FIG. 6**.

[0038] 4. Referring to **FIG. 7**, the three solenoids of varying radius and area produce a positive vector potential at the centerline of the rotors as seen in the graph.

[0039] 5. This graph is then rotated ninety degrees so that it can be located in relation to the rotors as seen in **FIG. 8**. The vector potential (D), which is created by solenoids (A,B,C), passes through rotors (E) and (H). Rotor (E), which has a positive charge (G), is rotating clockwise (F), and rotor (H), which has a negative charge (K), is rotating counter-clockwise (J).

[0040] 6. Referring to **FIG. 9**, rotor mechanics uses the exponential function for the harmonic motion of the rotor. The radius is {$r\ e^{i\omega t}$} which when differentiated with respect to time becomes a velocity {$i\ \omega\ r\ e^{i\omega t}$} where the imaginary {i} is a 90° phase lead which makes the velocity tangential to the rotor.

[0041] 7. The rotor surface charge {σ sigma} is rotating around at some radius {r}. For the upper rotor the surface charge density is positive (+σ} but the rotor has a negative angular velocity {−ω}. For the lower rotor, the surface charge density is negative {−σ} but the rotor has a positive angular velocity {+ω}. So the combined surface charge rotation is {$-\sigma\ r\ e^{i\omega t}+\sigma\ r\ e^{-i\omega t}$}. This charge rotation times the positive vector potential due to the solenoids times the rotor surface area is equal to angular momentum S as shown in **FIG. 10**. The units are given in **FIG. 11**. The angular momentum is equal to the negative of the difference of the exponentials times a constant. I then recalled that this difference is equal to twice imaginary sine of the angle as shown in **FIG. 12**.

[0042] 8. Referring to **FIG. 13**, the angular momentum is equal to minus two times the imaginary number times a constant times the sine of the rotational angle. In most of my spacecraft designs, the electromagnetic fields determine the flow rate of angular momentum. When the flow rate is integrated with respect to time, the angular momentum becomes imaginary. In this invention, the angular momentum is imaginary due directly to the rotors. The importance of being imaginary is that the radius resonates with the angular momentum. In some spacetime curvature equations, the denominator has a term equal to the radius to the fourth power plus twice the square of the angular momentum. In spacetime units, angular momentum is viewed as meter squared. So the square of imaginary angular momentum is negative angular momentum equal to negative meters to the fourth power. So at some radius, these two terms are equal, the denominator goes to zero, and the spacetime curvature becomes infinitely-large, creating a huge spike.

[0043] 9. This next section calculates the spacetime curvature from the equation for the angular momentum.

[0044] 10. Referring to **FIG. 14**, the elemental length ds^2 is curved by the presence of the angular momentum which is one of the energies, such as mass, charge, and electromagnetic fields, that Einstein showed can curve spacetime. Because the angular momentum is pointing in the vertical z-direction, due to the direction of the unit normal vector to the surface of the rotor, it rotates around in the angular direction {dtdθ} as found in cylindrical coordinates {t,r,θ,z}.

[0045] 11. In gravitational physics there is a g metric tensor which is a measure of length in spacetime coordinates. It is a 4 by 4 matrix with rows and columns equal to the cylindrical coordinates. Referring to **FIG. 15**, the diagonal of the matrix is {$-1,1,r^2,1$} where the minus one corresponds to time which is Einstein's convention. Half the angular momentum goes in the {tθ} slot of the g metric tensor, and the other half goes in the {θt} slot.

[0046] 12. From this g metric tensor, Einstein's G curvature tensor can be calculated in the various directions. In Einstein's General Theory of Relativity, his equation is G=8πT where G is the spacetime curvature measured in inverse meter squared, and the T tensor is the stress-energy-momentum matrix containing all the electromagnetic pressures, mass and momentum components that curve spacetime. The spacetime curvature tension G_{zz} in the vertical direction, as a function of radius, is shown in **FIG. 16**. A positive curvature indicates that there is a spacetime tension over the hull which produces lift. The curvature has a large positive spike over the dome of the hull (A) which means that there is a large lift force over the center of the hull. The curvature then falls off and spikes with a negative spacetime curvature compression over the rotors (B). This curvature also oscillates back and forth a short distance due to the sinusoidal term. The problem is to convert this compression into a lift force which this invention solves.

[0047] 13. This next section shows how the spacetime compression over the rotors generates lift.

[0048] 14. As I mentioned, I have been reading some of the World War II magnetron scientific papers of the Bell System Technical Journal. It turns out that inductance of a solenoid is equal to the permeability of space times the number of wire turns per length squared times the volume of the solenoid. Imagine having a copper strip in the shape of the magnetron cavity in **FIG. 1**. The electrical current (B) flows around the height of the strip (A), so there is only one turn per the height. Then that is multiplied by the volume of the cavity and permeability to get the inductance. The inductor is storing magnetic energy equal to half the inductance times the current squared. In this invention, current density J in amp per meter is used rather than current. So taking a hint from this information, the magnetic energy would have to be related to the inductance times the square of the current density times the area squared of the rotor and times the negative spacetime curvature. The curvature is measured in inverse meter squared. That product produces a negative magnetic energy when in fact I want a positive force. This means that there has to be an inductance gradient in the equation instead of pure inductance and, furthermore, the gradient has to be negative in order to cancel out the negative sign of the compression curvature.

[0049] 15. The equation for the magnetic energy in terms of the inductance $\{\mathcal{L}\}$, curvature {K}, current density {J} and area {A} is shown in **FIG. 17** with the units in **FIG. 18**. The inductance in the equation is proportional to the volume of the circular groove in the top surface of the rotor. That volume is equal to {2π r h w} where {h] is the height of groove (depth) and {w} is the width located at some radius {r}. In terms of machining, it would be more difficult to machine a variable width groove rather than a deeper groove of a constant width. So I am going to say the gradient is in the radial direction with the height of the groove decreasing going from the inside radius to the outside radius as depicted in **FIG. 19**.

[0050] 16. **FIG. 20** gives the equation for the height of the groove along the radius of the rotor.

[0051] 17. **FIG. 21** gives the inductance gradient by differentiating the inductance with respect to the radius. The initial groove height ho is not very large, and the inductance gradient becomes negative at the rotor as seen in the graph of **FIG. 22**.

[0052] 18. Referring to **FIG. 23**, the lift force on the rotor is now positive due to the combined negative inductance gradient times the negative spacetime compression. The square of the negative current density is positive also. This means that that the top surface of the lower rotor can also have a surface inductance gradient which would double the lift force.

What I claim as my invention is:

1. A spacecraft having a circular, domed hull with dual electrically-charged counter-rotating rotors one above the other located on the edge of hull.

2. Said hull having embedded within it three or more solenoids which generate a positive vector potential at the rotors.

3. Said upper rotor having a positive surface charge density, and rotating clockwise in the negative direction per the right-hand rule.

4. Said lower rotor having a negative surface charge density, and rotating counterclockwise in the positive direction per the right-hand rule.

5. Said rotor surface charge density and velocity creating a negative current density on both rotors.

6. Said vector potential and rotating surface charge density on rotors generating an angular momentum in the vertical direction.

7. Said angular momentum, generating a spacetime curvature tension over the dome of the hull, and a negative oscillating spacetime curvature compression over the rotors.

8. Said rotors having circumferential grooves of decreasing height machined into the top surface of the rotors in order to create a negative surface inductance gradient.

9. Said negative surface inductance gradient times the negative spacetime curvature compression times the rotor current density squared times the rotor area squared generating a positive lift force on the spacecraft.

* * * * *

US 20060014125A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0014125 A1**

St. Clair (43) **Pub. Date: Jan. 19, 2006**

(54) WALKING THROUGH WALLS TRAINING SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
4A
52 KINGS COURT
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/890,635**

(22) Filed: **Jul. 14, 2004**

Publication Classification

(51) **Int. Cl.**
G09B 19/00 (2006.01)

(52) **U.S. Cl.** **434/258**

(57) ABSTRACT

This invention is a training system which enables a human being to acquire sufficient hyperspace energy in order to pull the body out of dimension so that the person can walk through solid objects such as wooden doors.

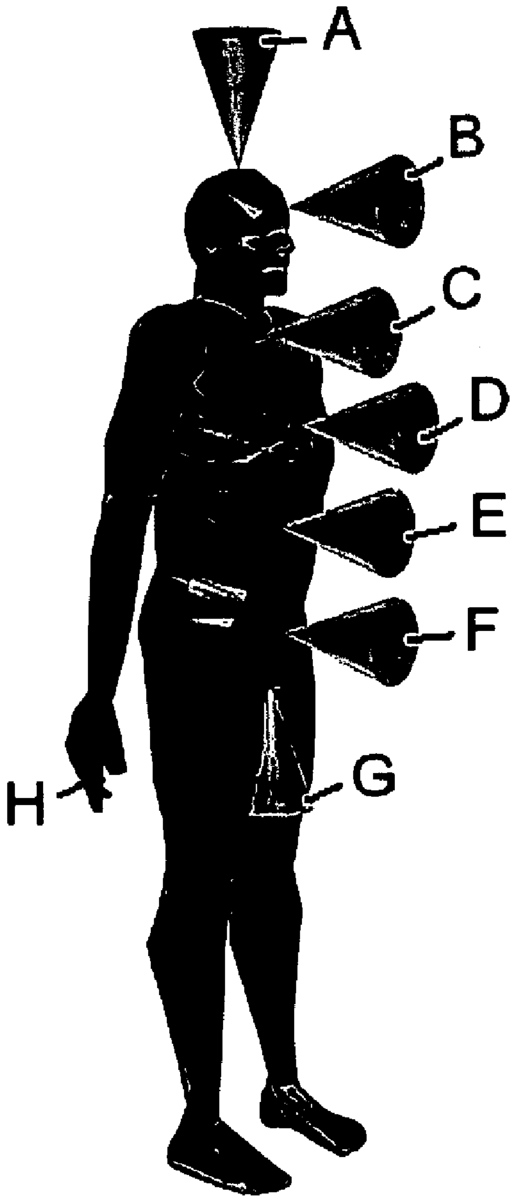

Figure 1

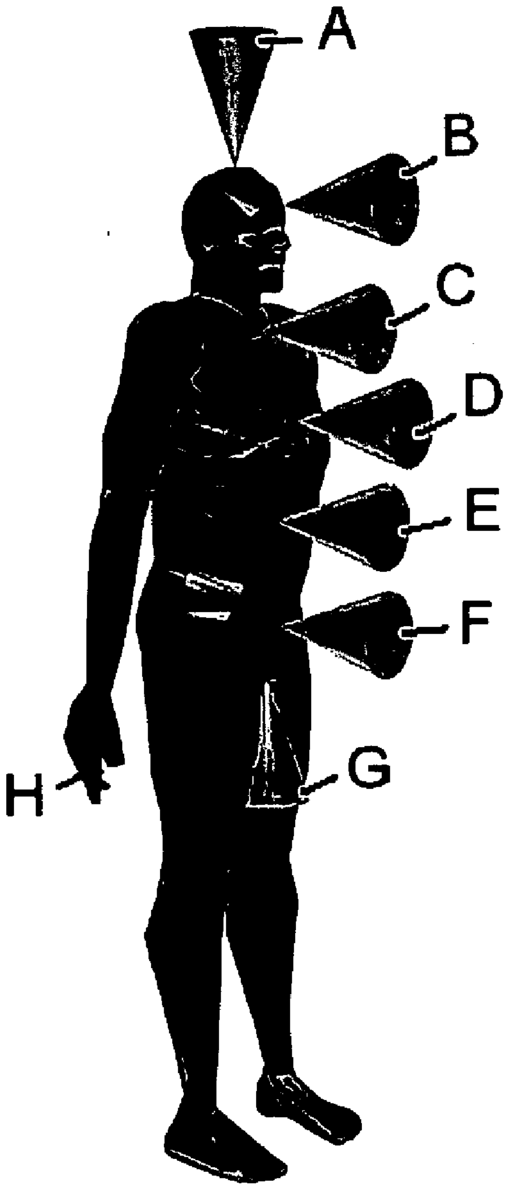

Figure 2

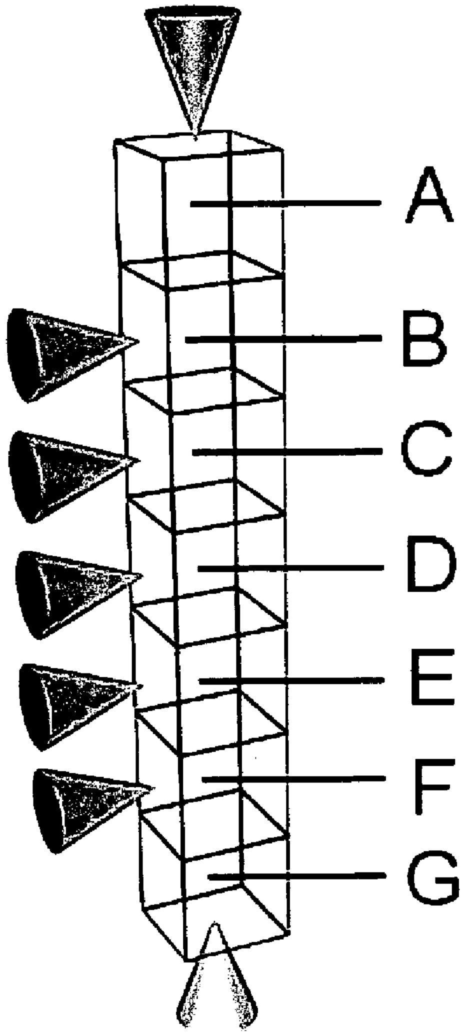

Figure 3

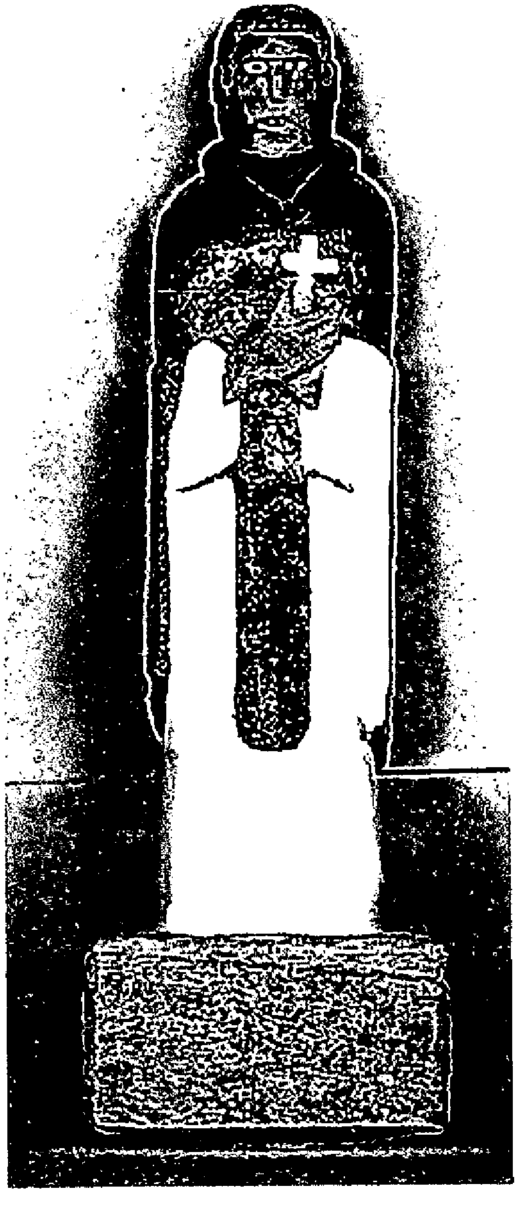

Figure 4

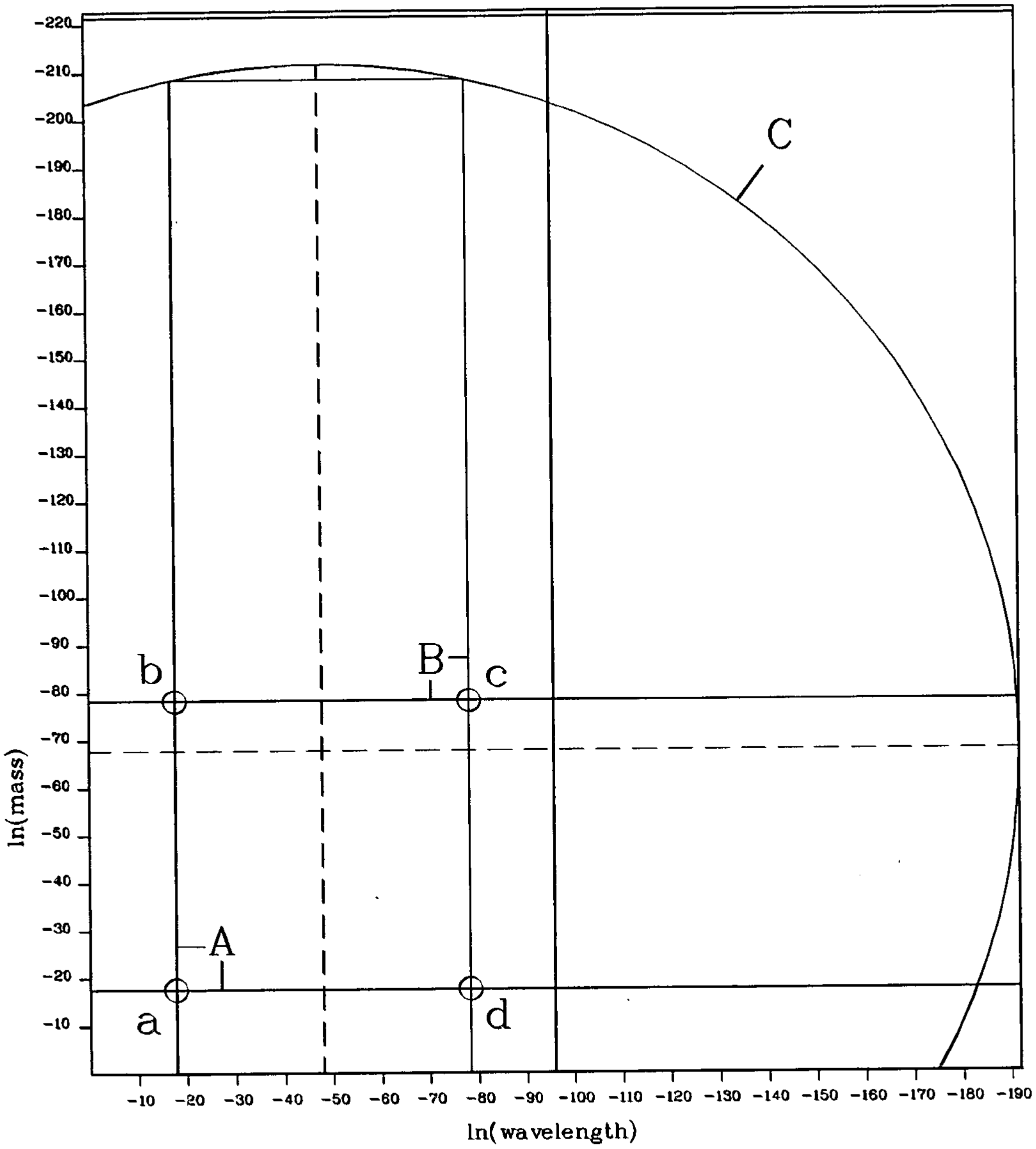
C
b
B
c
A
a
d
ln(mass)
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(wavelength)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190

Figure 5

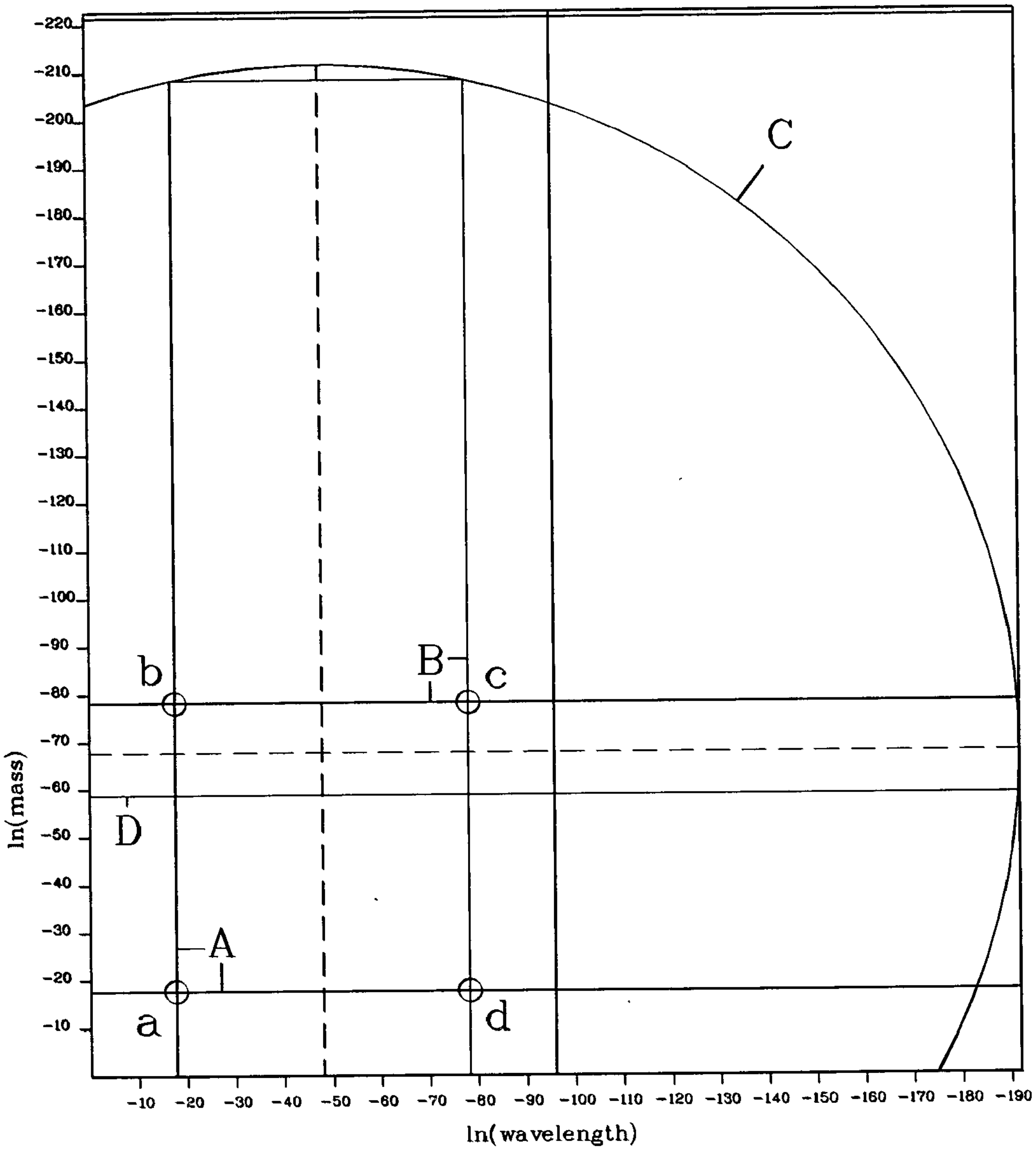
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
C
b
B
c
D
A
a
d
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 6

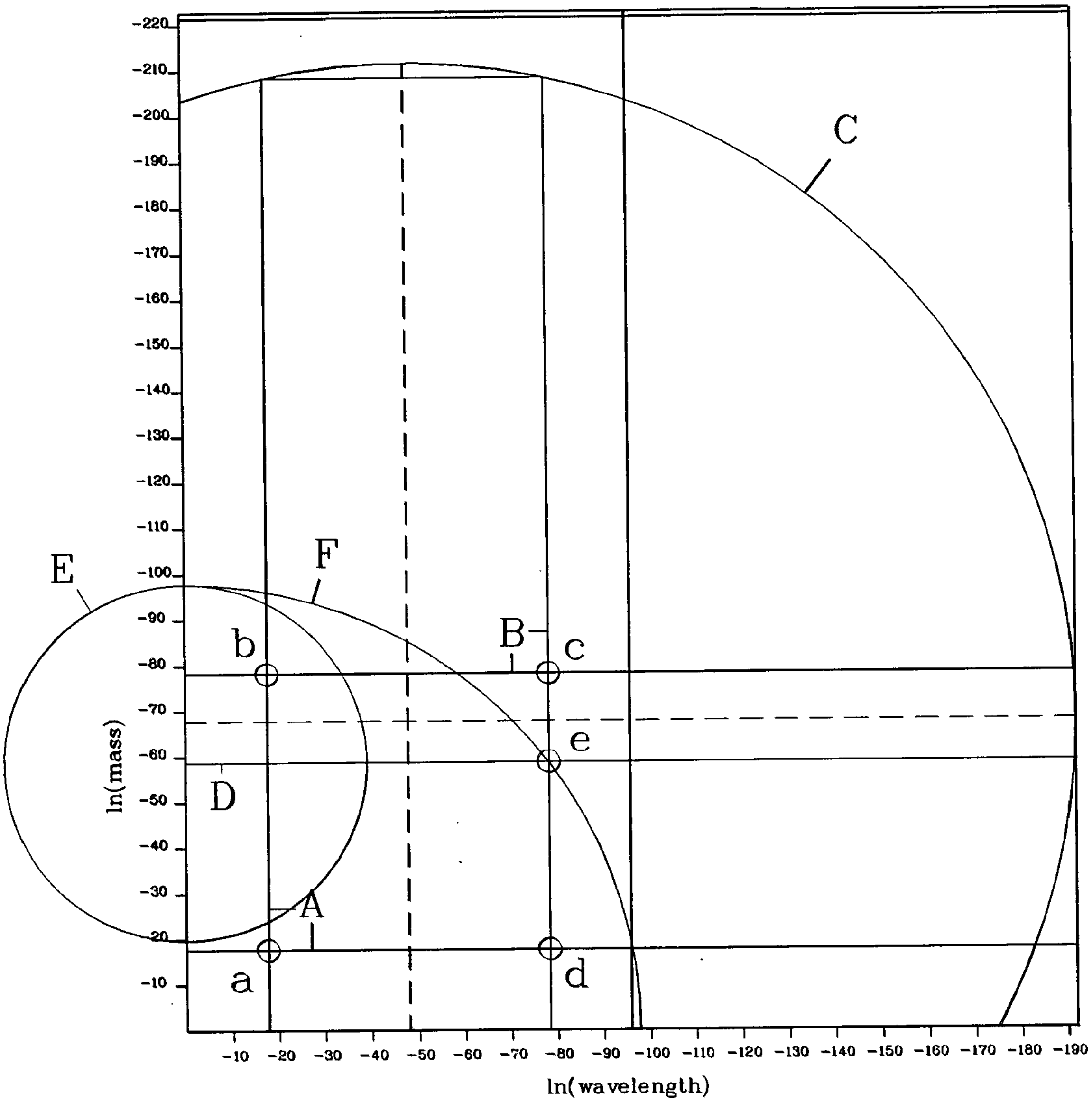

-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
C
E
F
B
b
c
e
D
A
a
d

Figure 7

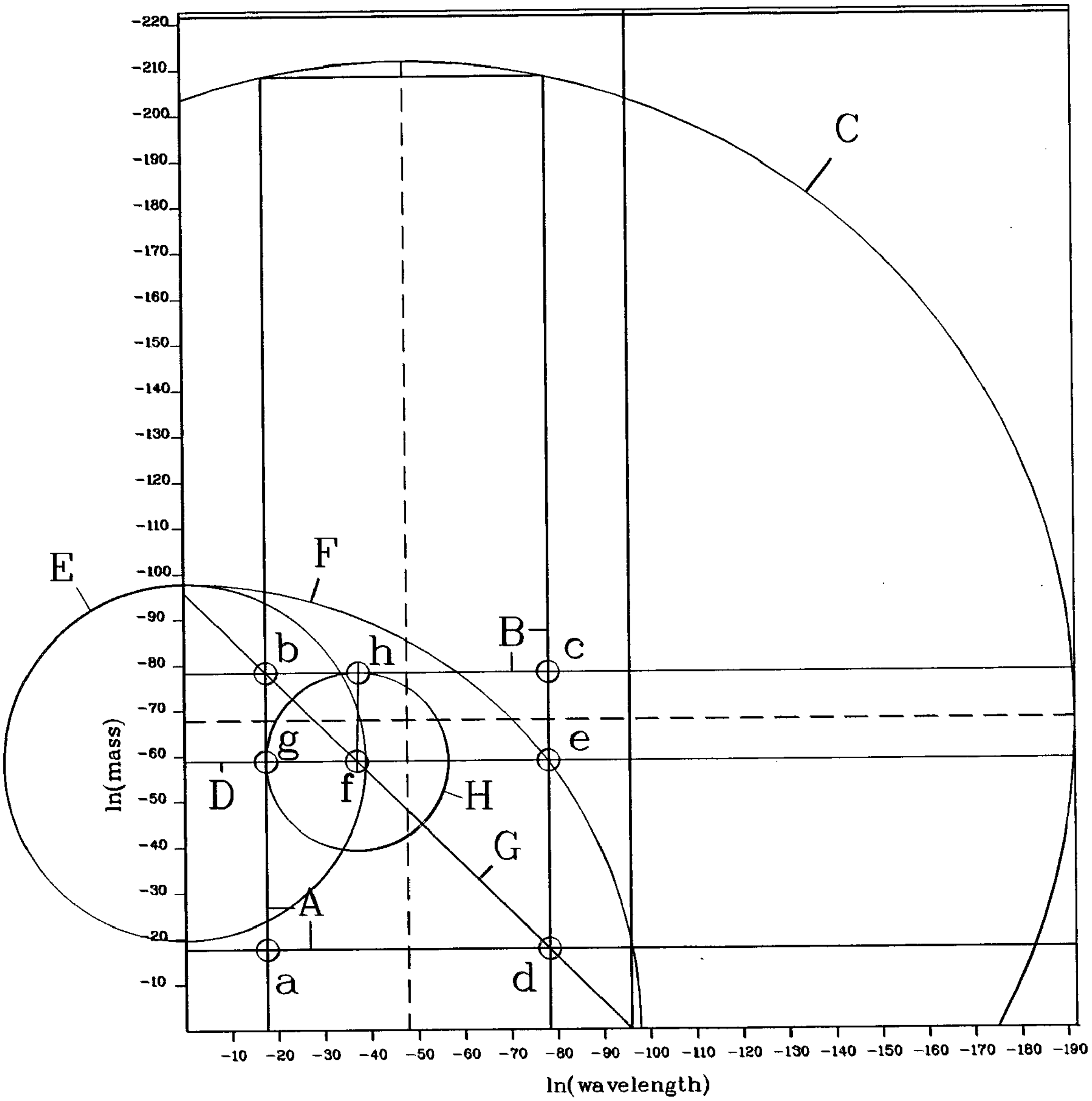
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)
C
E
F
B
b
h
c
g
e
D
f
H
G
A
a
d

Figure 8

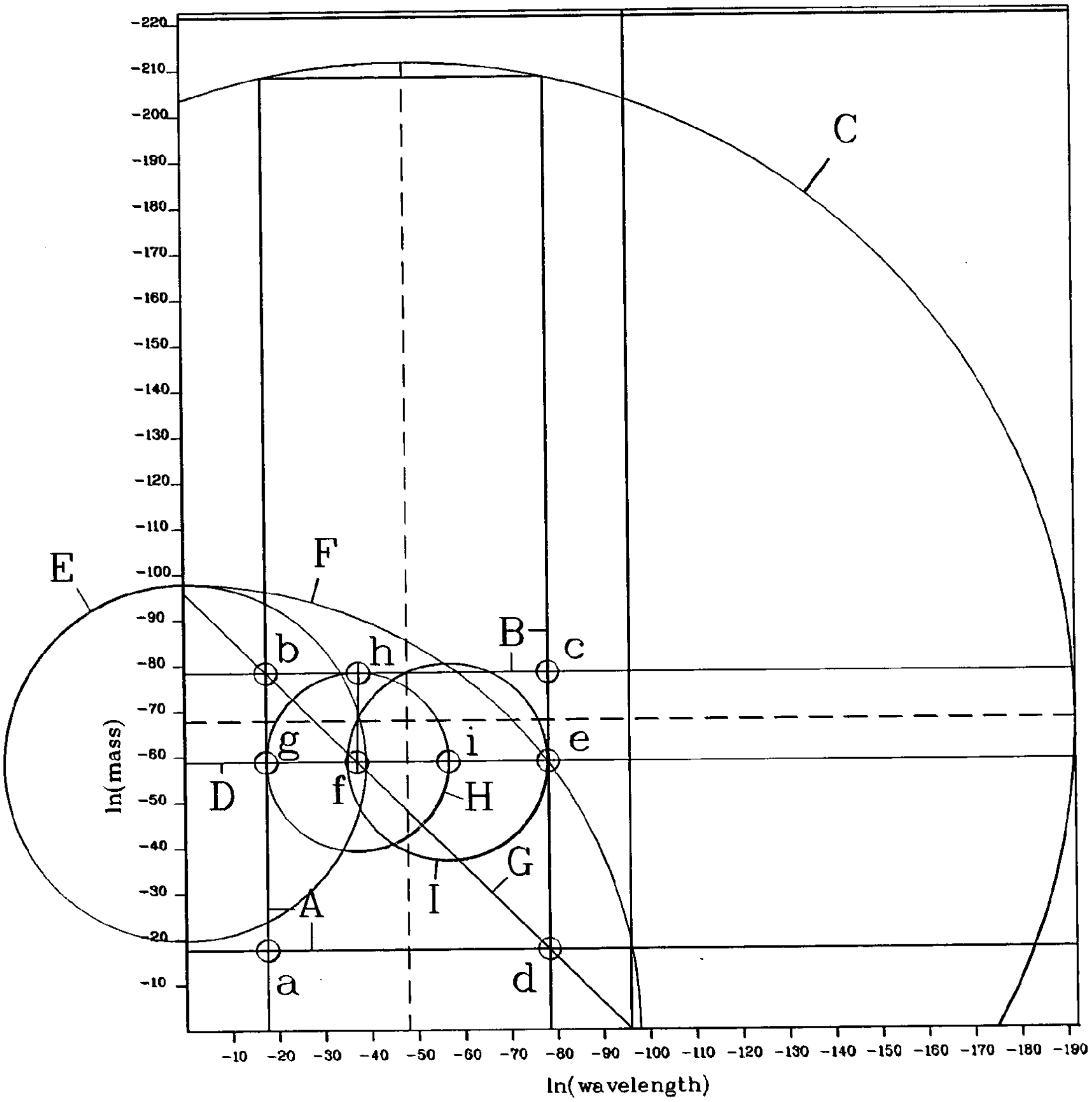
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
C
E
F
B
b
h
c
g
i
e
D
f
H
G
I
A
a
d

Figure 9

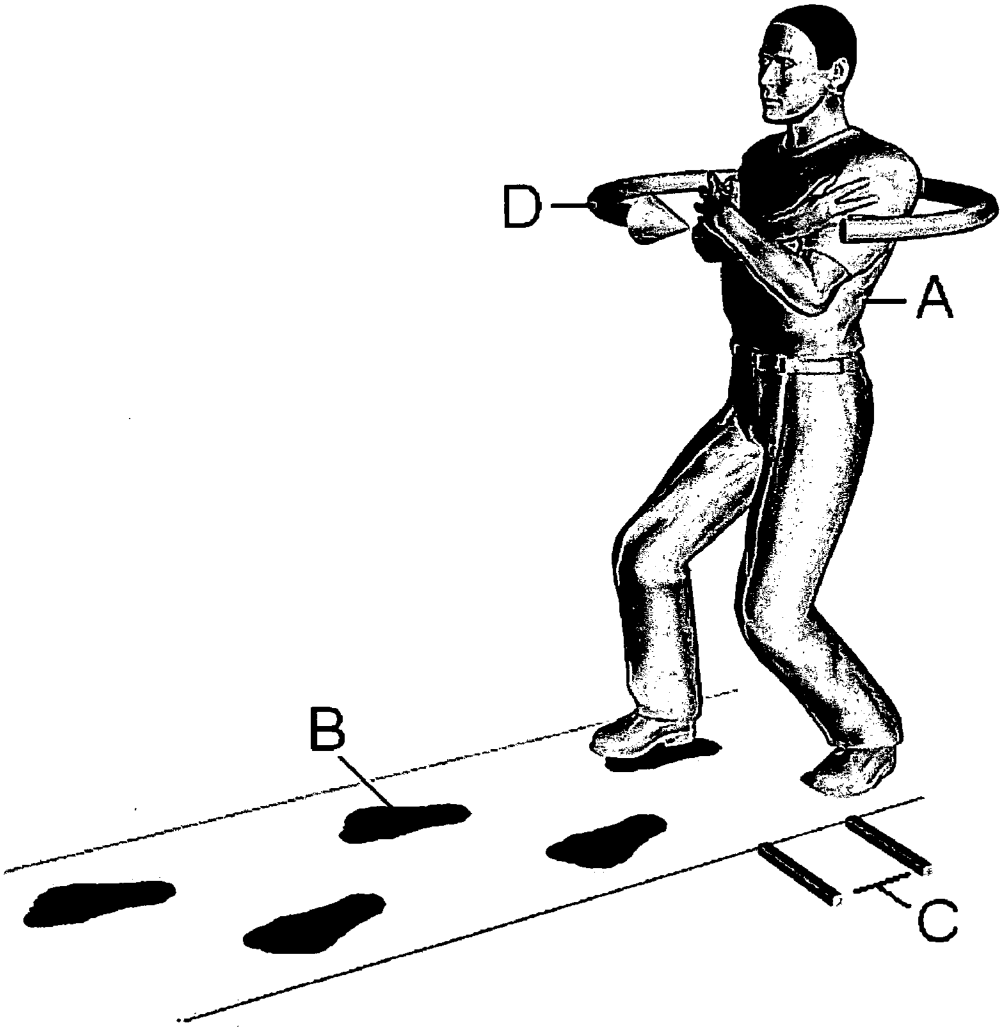

Figure 10

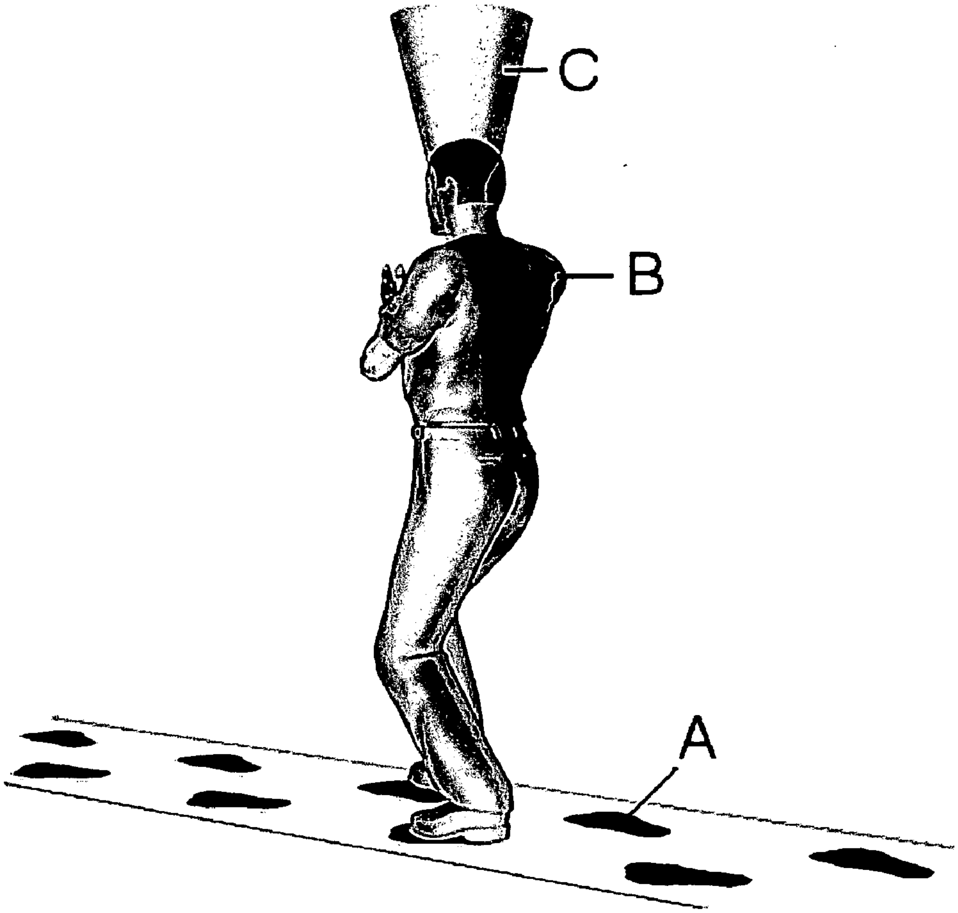

Figure 11

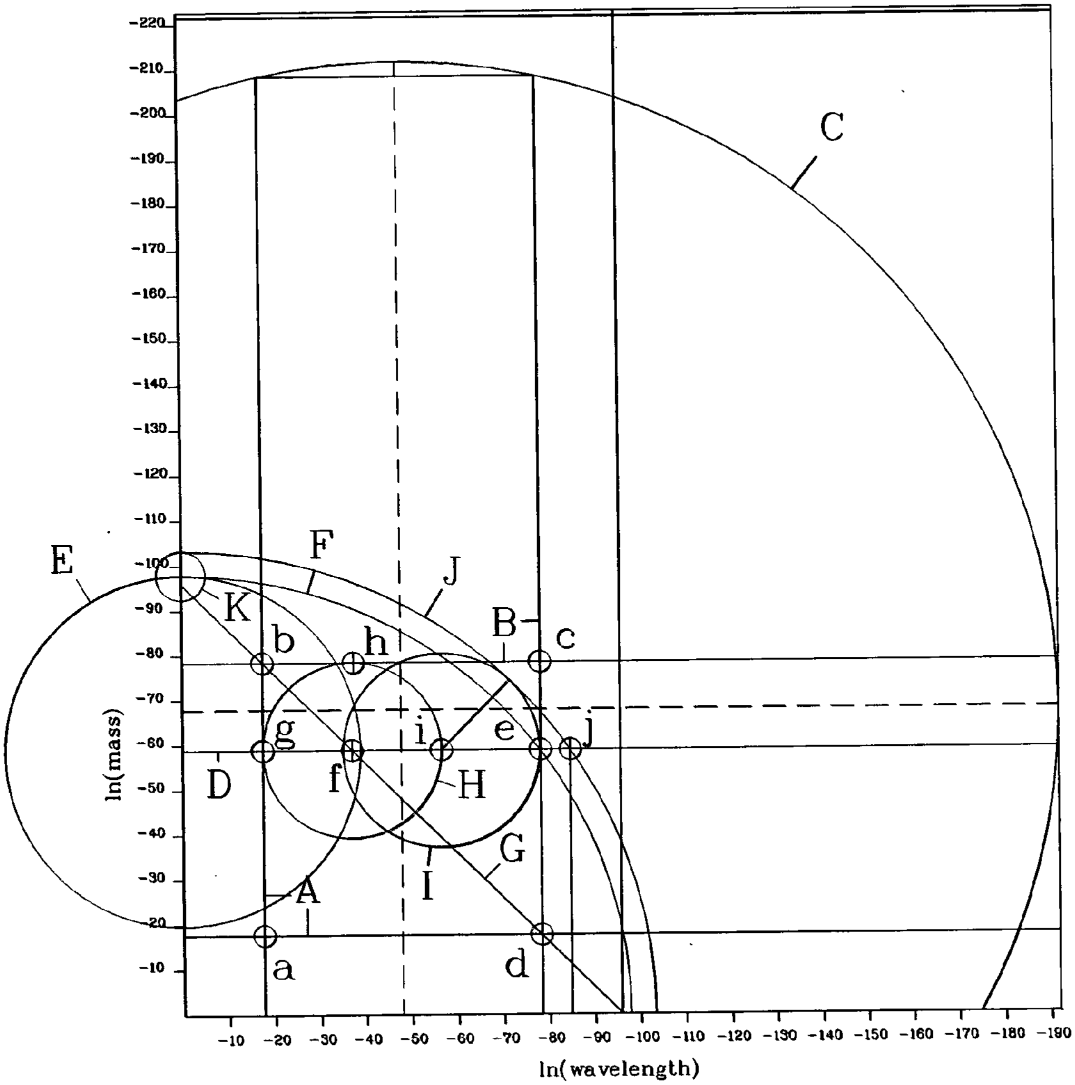
C
E
F
J
K
B
b
h
c
g
i
e
j
D
f
H
G
I
A
a
d
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)

Figure 12

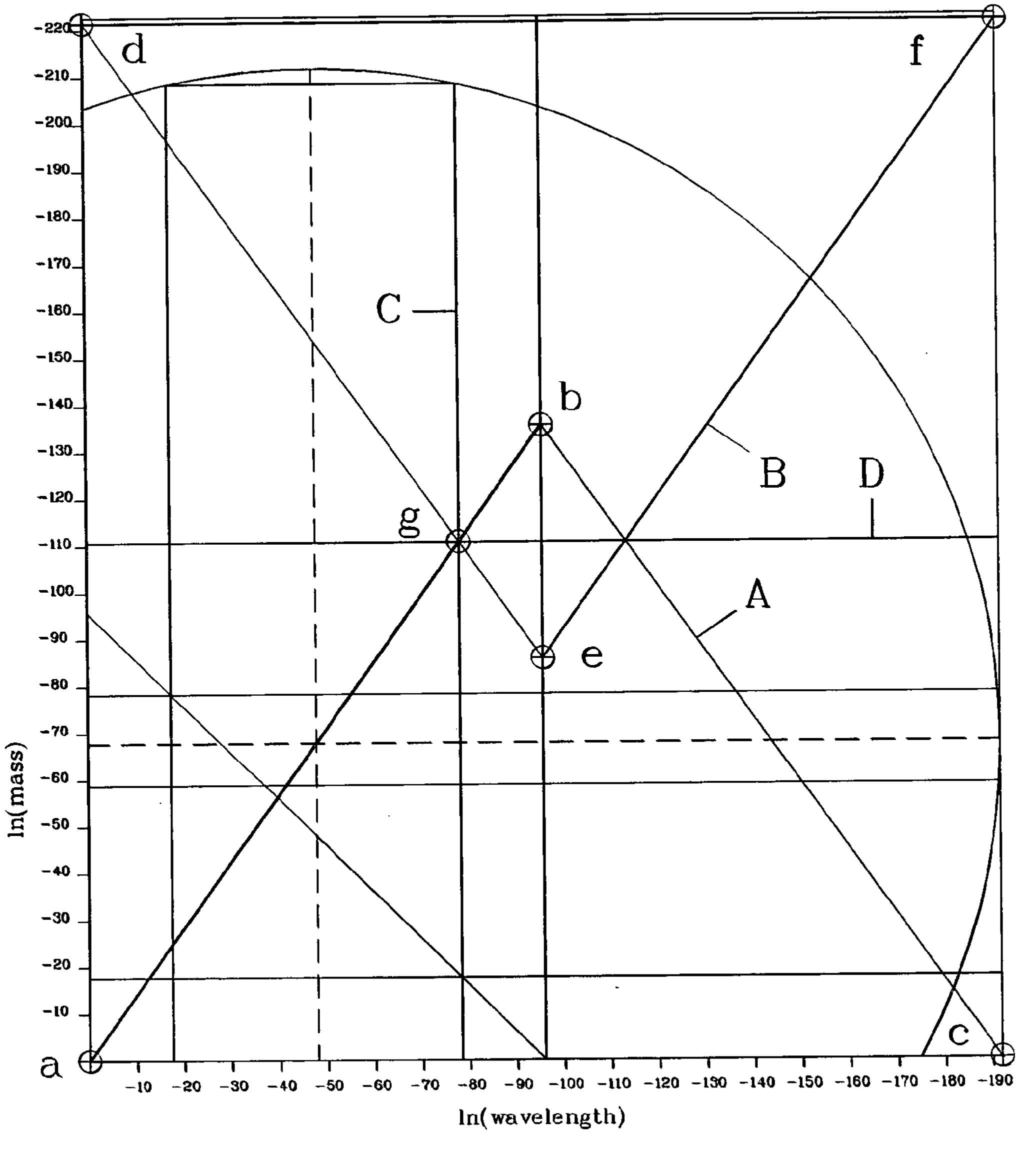
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
a
b
c
d
e
f
g
A
B
C
D
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 13

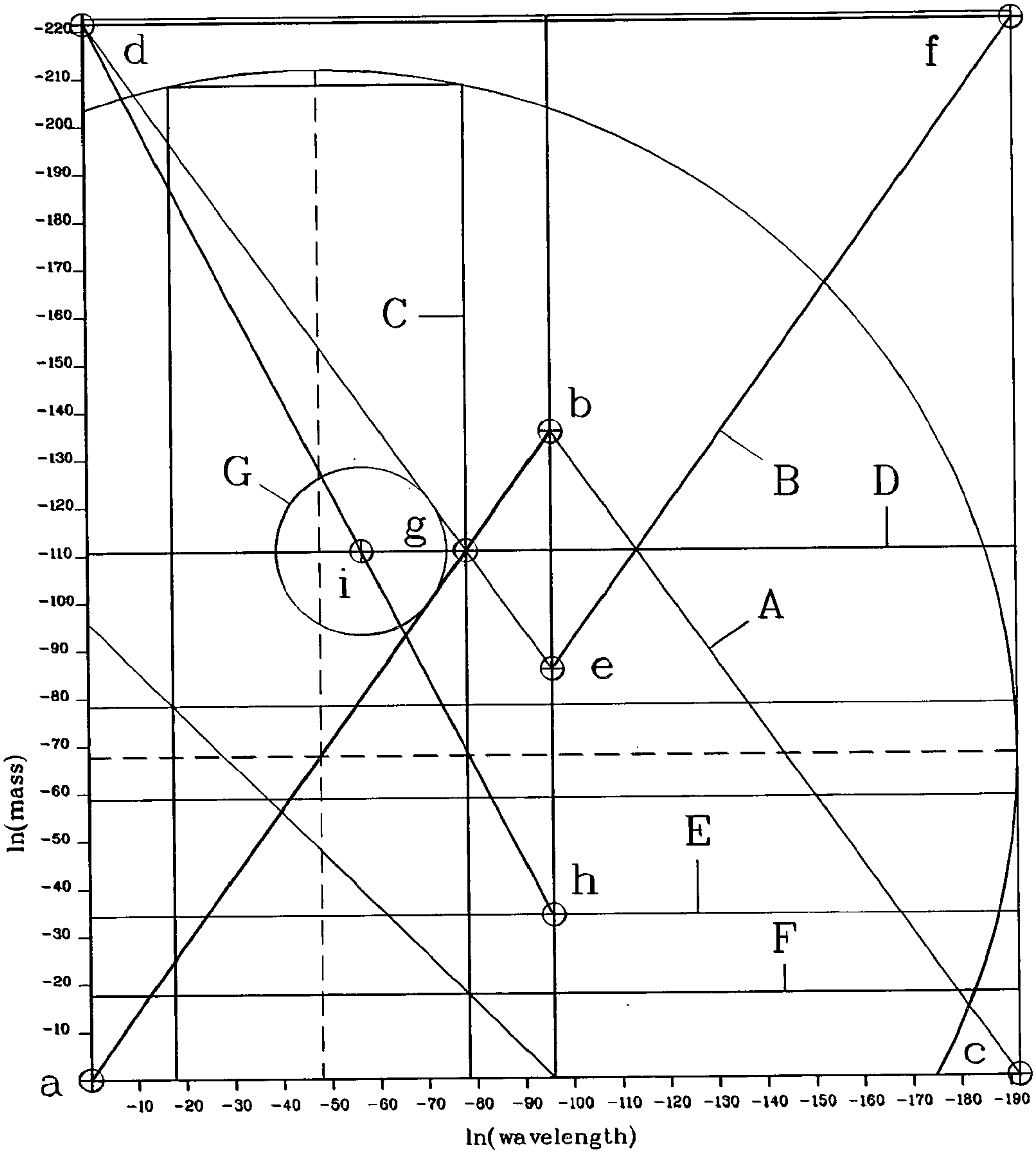
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
a
b
c
d
e
f
g
h
i
A
B
C
D
E
F
G

Figure 14

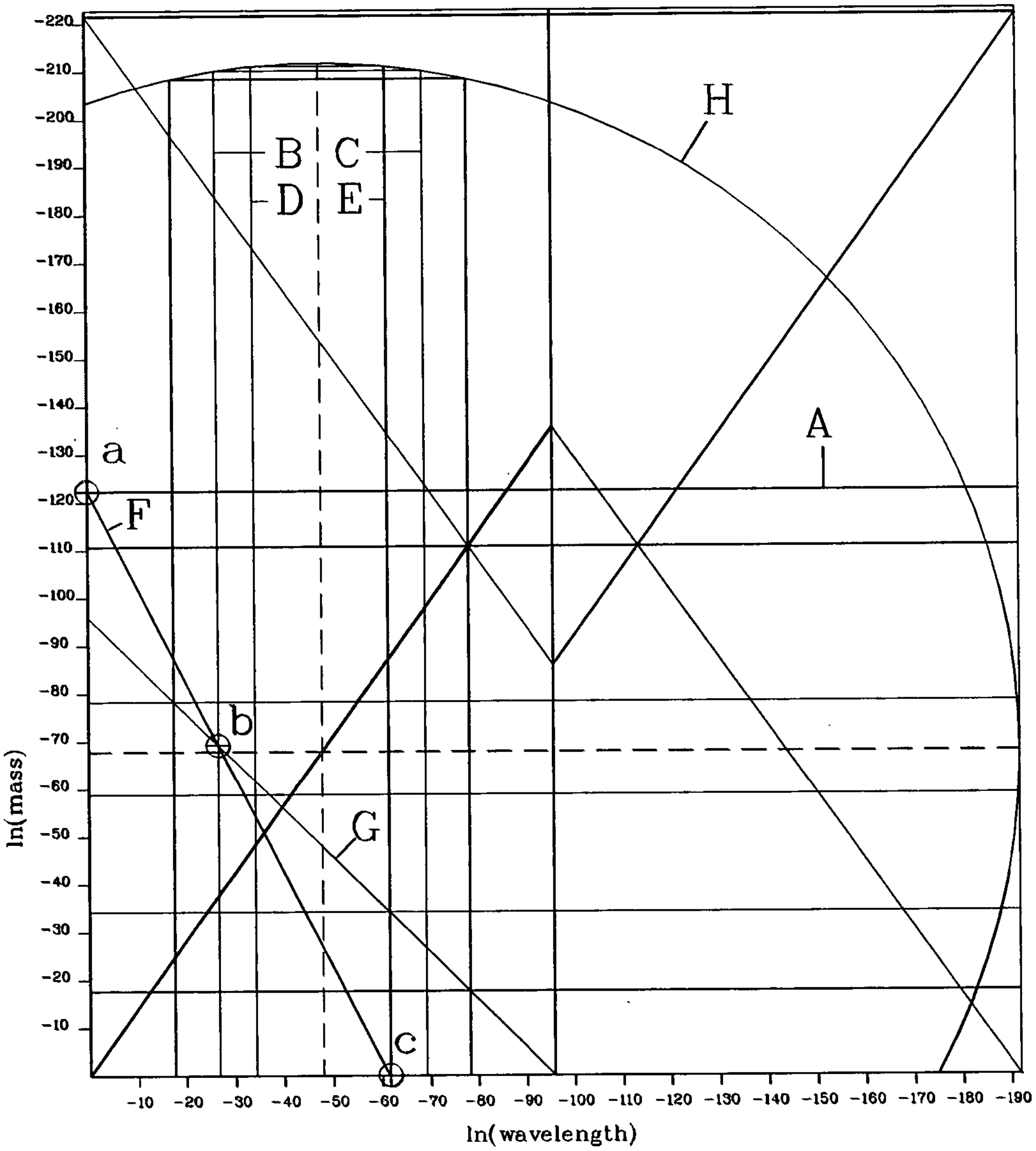
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)
H
B
C
D
E
A
a
F
b
G
c

Figure 15

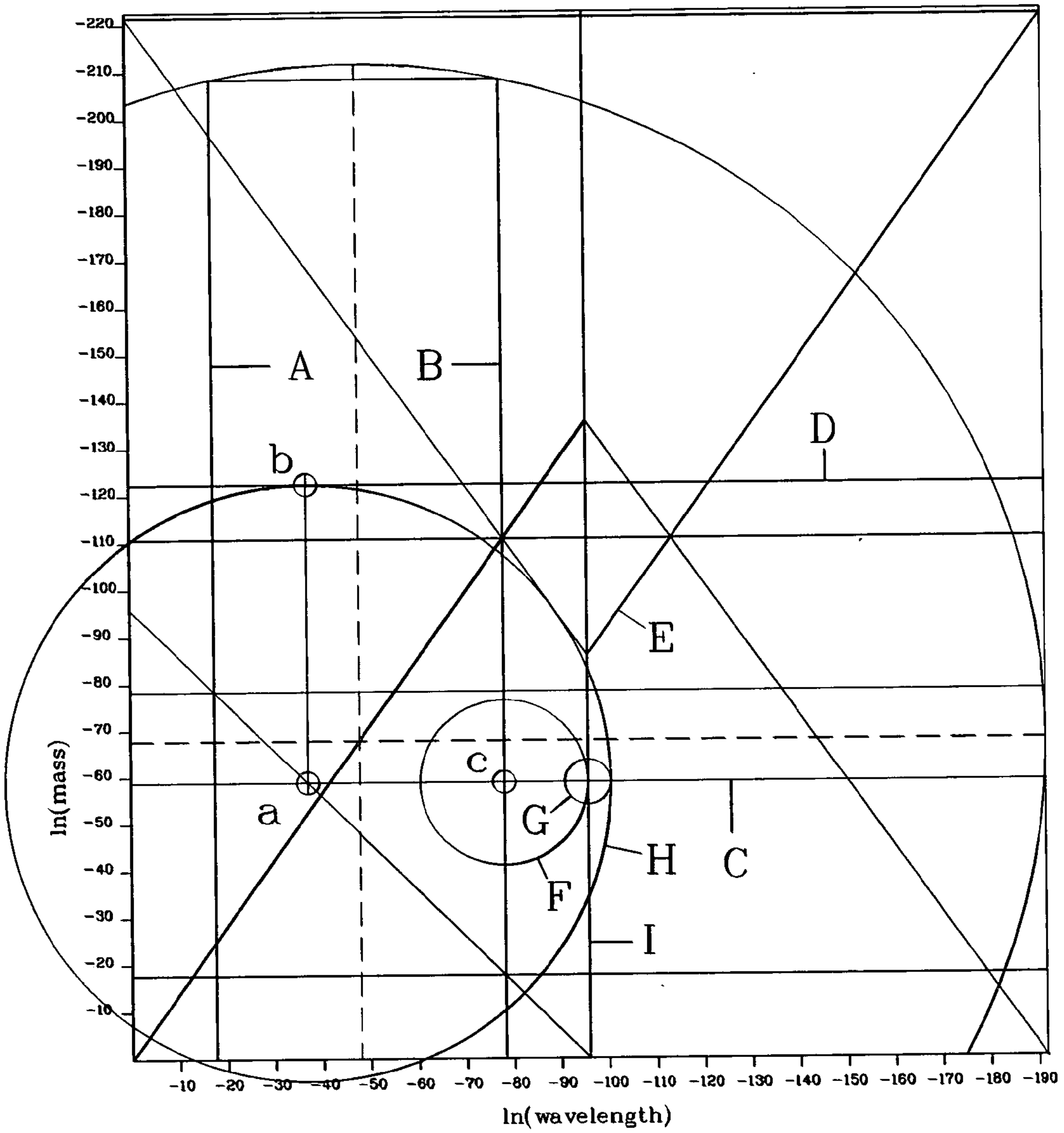
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
A
B
D
b
E
c
a
G
H
C
F
I

Figure 16

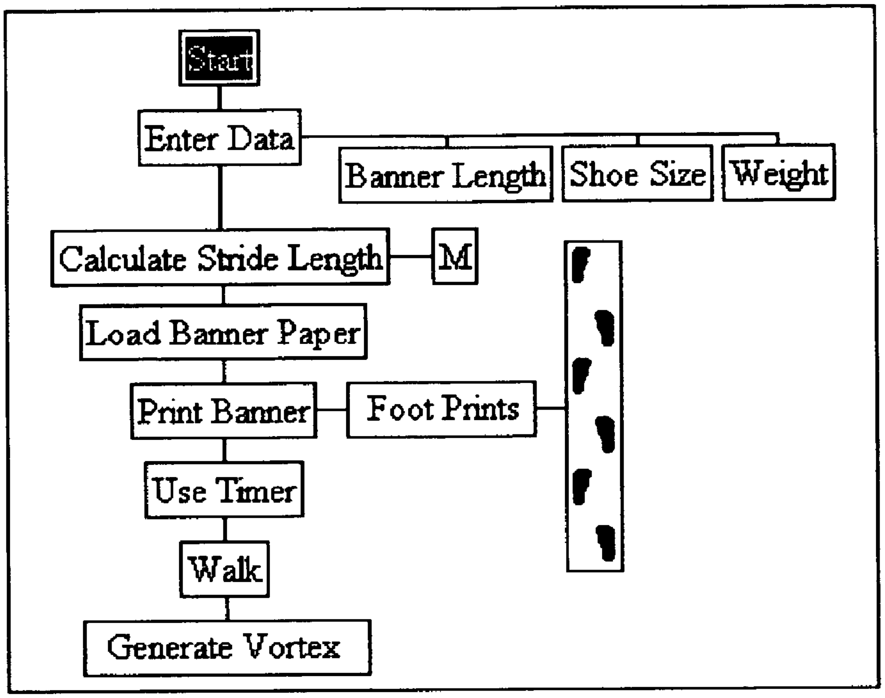
Start
Enter Data
Banner Length
Shoe Size
Weight
Calculate Stride Length
M
Load Banner Paper
Print Banner
Foot Prints
Use Timer
Walk
Generate Vortex

Figure 17

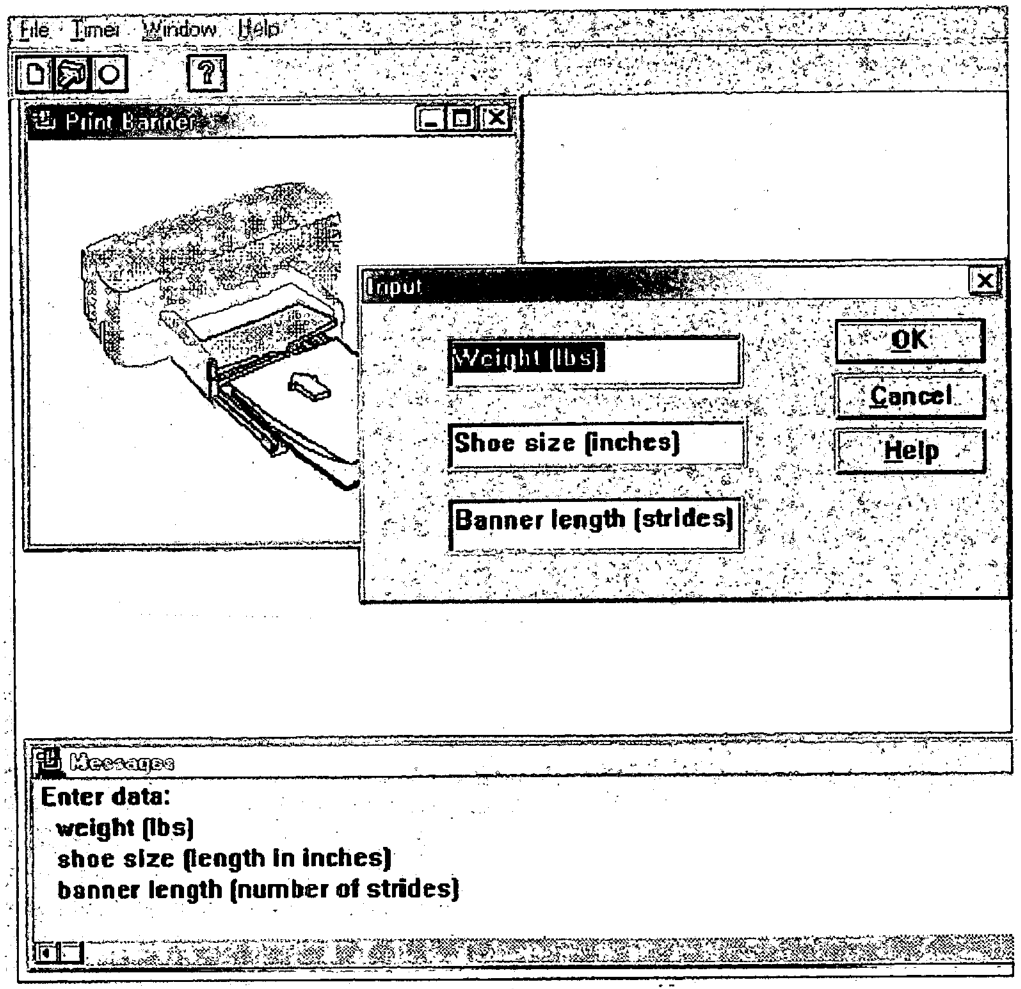
File Timer Window Help
Print Banner
Input
OK
Cancel
Help
Messages

Weight (lbs)

Shoe size (inches)

Banner length (strides)

Enter data:
weight (lbs)
shoe size (length in inches)
banner length (number of strides)

Figure 18

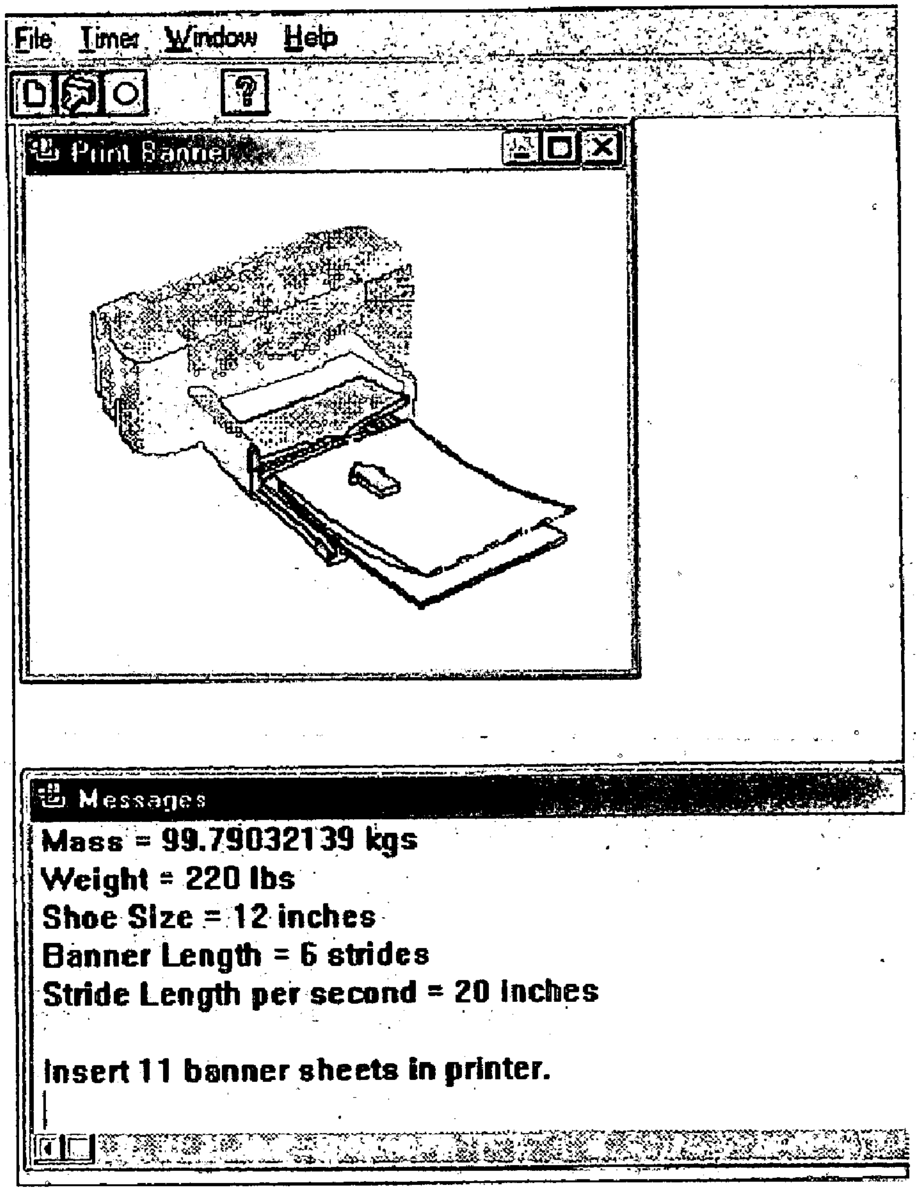
File
Timer
Window
Help
Print Banner
Messages

```
Mass = 99.79032139 kgs
Weight = 220 lbs
Shoe Size = 12 inches
Banner Length = 6 strides
Stride Length per second = 20 inches

Insert 11 banner sheets in printer.
```

Figure 19

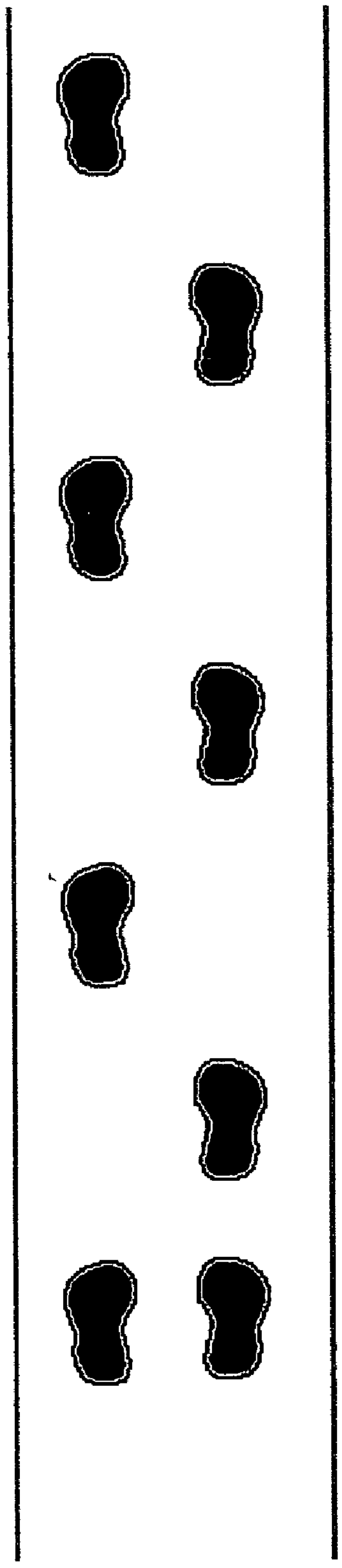

Figure 20

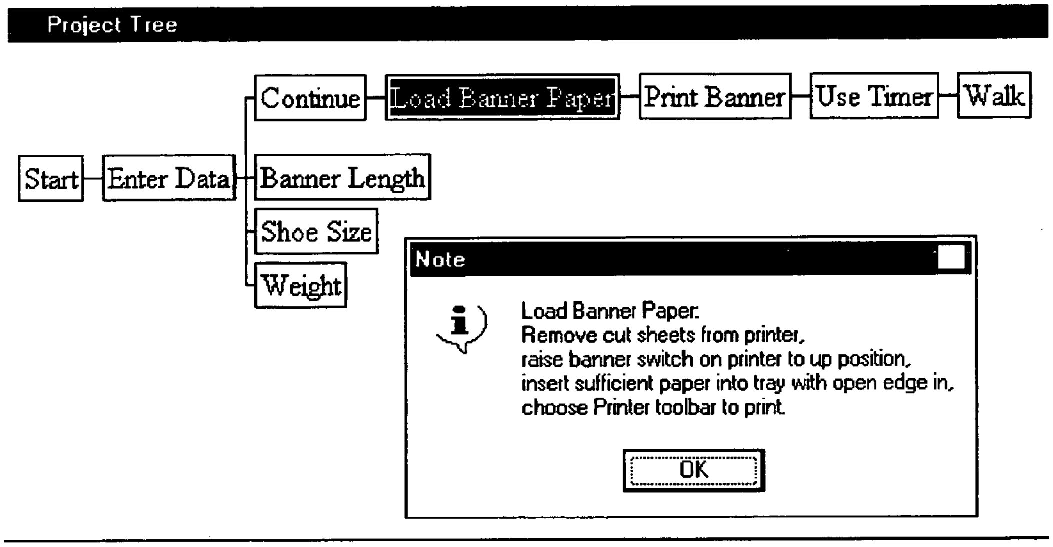
Project Tree
Start
Enter Data
Continue
Load Banner Paper
Print Banner
Use Timer
Walk
Banner Length
Shoe Size
Weight
Note
OK

Load Banner Paper:
Remove cut sheets from printer,
raise banner switch on printer to up position,
insert sufficient paper into tray with open edge in,
choose Printer toolbar to print.

WALKING THROUGH WALLS TRAINING SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a training system that enables a human being to acquire sufficient hyperspace energy in order to pull the body out of dimension so that the person can walk through solid objects such as wooden doors.

BACKGROUND OF THE INVENTION

[0002] A human being is a hyperspace energy being living in a physical container or body that is comprised of 67% water. This high percentage of water makes this invention possible. Referring to **FIG. 1**, the hyperspace energy being receives energy from our dimension through seven vortices that run the length of the body. Each vortex connects to a separate hyperspace dimension having its own particular frequency. This arrangement allows for the development of seven modular energy components corresponding to the mind, spiritual eye, voice, body, abdomen, plasma energy ball (battery), and ground connection.

[0003] Vortex (A), known as the top vortex, supplies energy to the mind and provides a channel of communication to other entities in the universe. This channel has been tested up to 100,000 light years which is the diameter of the galaxy.

[0004] Vortex (B), known as the spiritual eye, has a hollow cone-like appearance surrounded by white, misty, low-density hyperspace energy. Because the eye is modular, the mind can project it to vast distances, a process known as remote viewing.

[0005] Vortex (C), known as the voice and hearing module, can also be projected to enormous distances in order to communicate verbally with other entities. Using the proper remote viewing instruments, it is possible to project both the eye and verbal modules to see and talk at the same time.

[0006] Vortex (D), known as the heart vortex, provides protective energy to the upper internal organs, arms and hands.

[0007] Vortex (E) provides protective energy for the lower organs such as the intestines, liver and kidneys.

[0008] Vortex (F), known as the battery of the body, consists of an orange-colored plasma energy ball about one to two inches in diameter. The body becomes paralyzed if this ball is removed from the body. When the hyperspace energy being leaves the body, the vortices close and the battery takes over in order to keep the body functioning. This vortex also plays a role in the creation of the astral energy baby that attaches to the fertilized egg.

[0009] Vortex (G) is the ground vortex which rotates in the counter-clockwise direction in order to provide an energy ground for the electrical circuit. All the other vortices rotate in the clockwise direction as seen from the front such that energy flows into the vortex according to the right-hand rule of physics.

[0010] Vortex (H) is the hand vortex which rotates counterclockwise on the right hand and clockwise on the left hand as seen looking at the palm. Thus there is a rotating flow of hyperspace energy between the two hands when the palms are facing each other.

[0011] Referring to **FIG. 2**, each vortex feeds energy into its own hyperspace module shown by the lettered box. In terms of quantum mechanics, each box is actually a potential energy well in which each module develops separately. At the time of death of the body, these modules are joined together as a single energy being. The process is powered by the plasma energy ball battery which also contains the logical instructions for assembly.

[0012] Different people, as tested by the pendulum on the hand vortex, have different vortex sizes. Three people were tested. The first person had almost no rotational movement showing very little energy. The second person had a vortex radius of one inch. Another had a vortex radius of four inches which covered his entire hand. The latter also has the ability to lift another human being, lying prone on a table, off the table by flowing low-density hyperspace energy into the person through the hand vortices. He has also experienced walking through a solid wooden door with a dog at his side. What this means is that there is a way, as described in this invention, of creating large energy vortices which will enable a person to acquire sufficient energy to walk out of dimension through solid objects such as wooden doors.

[0013] Researching the historical records, referring to the statue in **FIG. 3**, a humble black Catholic Dominican friar of the Santo Rosario Convent, by the name of San Martin de Porres, living in Peru in the 1500's, also developed this ability to walk through doors. For his beatification, many witnesses came forward to recount his extraordinary abilities. For example, a witness, who worked in the Convent, went to the cell of San Martin to ask for something to eat. When he reached the cell, he saw San Martin leaving with some medications apparently to heal someone sick. The witness waited by the open door of the cell for his return. After having waited awhile without being distracted by anything else, he saw said venerable brother fray Martin come out from his cell from the inside and call him by name. The witness was terrified, not understanding how this was possible.

[0014] The woman who gave me the statue told me that he would walk miles and miles each day to visit the poor. The task of walking means that there is a velocity involved. Because the body has mass, then there is a mass times a velocity, or momentum, involved in this ability. Notice also that the statue shows him walking with his right arm crossed over his left arm in the form of an X.

[0015] What this walking momentum means in terms of physics is given in the following analysis. The human body consists of 67% water. A water molecule consists of two hydrogen atoms and one oxygen atom having the atomic formula H_2O. The atomic weight of one atom of hydrogen is 1.008 awu. The atomic weight of one atom of oxygen is 16.000 awu. The molecular weight of one atom of water is therefore:

Weight of two atoms of hydrogen 2 × 1.008 awu =	2.016 awu
Weight of one atom of oxygen 1 × 16.000 awu =	16.000 awu
	18.016 awu

The formula weight is just the atomic weight expressed in grams. Thus the formula weight of water would be 18.016

grams or 0.018016 kilograms. According to Avogadro's law, the formula weight contains N=6.02×10^{23} molecules. Thus the mass of one water molecule is the formula weight divided by the number of molecules:

$$\text{mass} = \frac{.018016\ \text{kg}}{N} = 2.99269103 \cdot 10^{-26}\ \text{kg}$$

$$\ln\left(\frac{\text{mass}}{1\ \text{kg}}\right) = -58.77103943$$

[0016] According to Einstein's Special Theory of Relativity, energy is equal to the mass times the speed of light squared. The energy of a photon is equal to Planck's constant h times the frequency f of the photon. Equating these two energies shows that

$$E = \text{mc}^2 = hf = h\frac{c}{\lambda} = \frac{h}{c}\frac{c^2}{\lambda}$$

$$\Rightarrow m\lambda = \frac{h}{c}$$

$$\Rightarrow \ln(m) + \ln(\lambda) = \ln\left(\frac{h}{c}\right) = \text{base} = -95.91546344$$

which says that the natural logarithm of the mass plus the natural logarithm of the wavelength is equal to the natural logarithm of Planck's constant divided by the speed of light c, known as the base constant in the tetrahedron diagram. This diagram plots the mass versus wavelength in natural logarithms. Notice that the left hand side of the equation is the sum of mass and wavelength, so the right hand side must also be the same.

$$\ln\left(\frac{h}{c}\right) = \ln\left(\frac{\Omega\Lambda\ 2\pi\Lambda\ c}{c}\right) = \ln(\Omega\Lambda) + \ln(2\pi\Lambda)$$

where ΩΛ, known as the Planck mass, is the linear mass Ω of the universe times the bottom dimensional limit of the universe Λ, and 2πΛ is the bottom dimensional wavelength, known as the Planck wavelength. That is, our dimension is bounded by the Planck box having sides Planck mass and Planck wavelength. These boundaries have values:

$\ln(\Omega\Lambda) = -17.64290101$

$\ln(2\pi\Lambda) = -78.27256243$

[0017] Referring to **FIG. 4**, these two lines are plotted on the tetrahedron diagram. The Planck mass line (A) reflects off the sphere (C) and returns as the Planck wavelength (B) which shows the dual nature of quantum physics. This creates the Planck box (a,b,c,d) which is the boundary of our dimension.

[0018] Referring to **FIG. 5**, the mass of the water molecule is plotted as horizontal line (D) on the diagram. The energy of the water molecule is the mass times the speed of light squared.

$E = mc^2$

$\ln(c) = 19.51860099$

$\ln(mc^2) = \ln(m) + 2\ln(c)$

[0019] Referring to **FIG. 6**, a circle (E), having a radius equal to the speed of light squared, centered on the mass of the water molecule at the vertical axis, generates a circle (F), centered at the origin, that intersects (e) the mass of the water molecule at the Planck wavelength. That is, this intersection point sits right on the Planck box boundary between space and hyperspace. The radius of circle (F) is actually the mass of the water molecule divided by the speed of light squared.

$$E = \frac{m}{c^2} = m$$

$$\Rightarrow \frac{1}{c^2} = 1$$

$$\Rightarrow c = 1\frac{\text{meter}}{\text{sec}}$$

$$\ln(c) = 0$$

Taking the positive square root, the speed of light is one meter per second at the Planck boundary. The experiments with brain hemisphere resonance show that the resonant frequency of the human energy field is between 1 Hz to 5 Hz which is well below the hearing threshold of 20 Hz. Because the traveling wave has a wavelength of 0.3048 m and the speed of light is unity at the boundary, the frequency should be

$$f = \frac{c}{\lambda} = \frac{1\frac{m}{\text{sec}}}{.3048\ m} = 3.28\ \text{Hz}$$

which is within the middle of the experimental male range. At this resonant frequency, the human energy being pops out of the body. This represents only a first stage in the development of the energy being. But what is really wanted is to have both the physical body and the energy field move out of dimension together as San Martin did.

[0020] Referring to **FIG. 7**, the 45° base line (G) is added to the diagram. Notice that the Planck mass intersects (b,d) the Planck wavelength on this line because they sum to the base constant. The mass of the water molecule crosses this line at point (f). A circle, centered (f), with a radius equal to the speed of light, is tangent to the Planck wavelength (h) and the Planck mass (g). Since mass times velocity is momentum, the diagram says that the momentum of the water molecule is tangent to the boundaries of the Planck box which separates space from hyperspace. In order to get to point (e) from the momentum of the water molecule, a second circle is added to the momentum.

[0021] Referring to **FIG. 8**, a circle (I), centered on the water molecule mass at the speed of light circle (i), is made tangent to the Planck wavelength at point (e). The momentum M of space is equal to the Planck mass times the Planck scale times the speed of light. At point (e), the speed of light is unity, so that the momentum is just the Planck mass in momentum units:

$$M = \Omega\Lambda c = 2.176634194 \cdot 10^{-8}\ \mathrm{kg}\frac{m}{s}$$

Circle (I) has a radius equal to

$$\text{ratio} = e^{21.572952} = 2338912700$$

Therefore the walking momentum in order to get to point (e) is the momentum of space M times this ratio

$$M_w = M \cdot \text{ratio} = 50.909573606\ \mathrm{kg}\frac{m}{\sec}$$

The stride length L per second that a person of mass W has to walk is the walking momentum divided by the mass W times a period T of one second

$$L = \frac{M_w}{W}T$$

For a person with a mass W of 99.79 kg (220 pounds), then the stride length L is 20.08 inches or one foot and eight inches. The person has to walk this length in one second on each foot.

[**0022**] Looking at the statue of San Martin, his arms are crossed over each other. The vortex of the right hand points backward, and the left hand vortex points forward due to the reversed rotation. Referring to **FIG. 9**, this creates a rotational energy channel (D) around his body (A). The stride length (C) is calculated according to the body mass, and then a banner printout is made showing where the footprints (B) are to be placed each second. The question is: "What happens when one walks the walk?".

[**0023**] On the very first experiment, referring to **FIG. 10**, what happens is that, after taking only six strides on the banner printout (A), a huge spinning vortex (C) develops over the top of the head and the vertex locks onto the heart vortex in the center of the chest (B). In everyday life, this vortex is not created because normal walking is much faster and the hands are held at the side of the body. The energy rush through the pineal gland is so intense that one feels immediately sleepy and starts yawning excessively due to the increased flow of melatonin.

[**0024**] After practicing with the banner printout, long walks were made through the park. In this case, a vertical white line rotated around a vertical axis located about six feet perpendicular to the path on the right side of the body. When the walking speed was correct, this white line would lock onto the centerline of the body. Speeding up or down caused the white line to lose synchronization and rotate away. This white line is related to the ability to levitate the body. San Martin had so much energy that, according to witness testimony, he could float horizontally in the air with his head resting against the bowed head of Christ on a carved wooden cross. Thus San Martin's energy sources were channeling energy from Christ, collective broom energy as described in a separate patent application, and the walking momentum vortex energy.

[**0025**] During the early part of the 20th century, a man's parents were lying in bed dying of tuberculosis. With their permission, he placed a weighing scale under each of their beds. When they passed away, he found that each scale registered a loss in weight of 2.5 ounces. This is equivalent to 0.071 kg, which is the mass of the human energy being.

[**0026**] After conducting a number of experiments with water vortices draining from a cylindrical tank, it can be stated from Bernoulli's theorem that the potential energy plus the kinetic energy is a constant

$$gz + \frac{1}{2}mv^2 = const$$

The shape of the inner surface of the water circulation has a velocity proportional to the inverse of the radius, so the shape of the surface is

$$(z - z_0) = \frac{k}{r^2}$$

which says that the height of the vortex is proportional to the inverse of the square of the radius.

[**0027**] The hand vortex area ratio between the second test subject and the third test subject is equal to the square of their radii:

$$\text{ratio} = \frac{(1\ \text{in})^2}{(4\ \text{in})^2} = \frac{1}{16}$$

Because the speed of light at the boundary was determined to be one meter per second, the energy of the third test subject is

$$E = \left(\frac{.071\ \mathrm{kg}}{16}\right)\left(\frac{1\ \mathrm{m}}{\sec}\right)^2 = 4.4375 \cdot 10^{-3}\ \text{joule}$$

$$\ln(E) = -5.417664124$$

[**0028**] Referring to **FIG. 11**, a circle (K), having this radius, is added to the energy of the water molecule (E), to produce augmented energy circle (J). This circle (J) intersects the mass of the water molecule outside the Planck box at point (j). This means that the increased hyperspace energy moves the water molecule, and hence the body, out of dimension. Furthermore, circle (J) is tangent to the walking momentum ratio circle (I) which keeps the geometry locked together.

SUMMARY OF THE INVENTION

[**0029**] It is the object of this invention to create a training system that allows a person to develop the ability to walk around out of dimension, passing through solid objects. This invention is based on one of the most remarkable relationships between the water molecule and the boundary between space and hyperspace. The mass of the water molecule is equal to the energy of the water molecule at this boundary. Because the body is composed of 67% water, the body sits on the boundary such that any additional increase in energy would move the body out of dimension into hyperspace. Because human beings are actually hyperspace energy beings living in physical bodies, the additional energy required to move the body out of dimension comes from increasing the energy of the hyperspace being. One source of this energy comes from walking cross-handed at the proper velocity in order to generate a large hyperspace energy vortex that flows energy into the potential wells of the hyperspace being. This increased hyperspace energy will then allow the person to walk around out of dimension through solid wooden doors. Because the door and the person are in two slightly different dimensions at the same moment, it appears that the person is walking through the door. After passing through the closed door, the person then returns to our dimension and emerges in the interior of the closed-door room.

[0030] This technique can be used in reverse to heal an infected hand instantaneously. A salve made from the St.Mary's herb is applied to the skin of the infected hand. The hyperspace energy then flows through the right-hand vortex such that the infected hand and the salve are taken slightly out of dimension. What happens is similar to when a short piece of straw is embedded in a hard wooden telephone pole as a tornado passes over the pole. The straw and pole are taken out of dimension such that they briefly merge together. As the tornado moves on, both objects come back to dimension merged together. Thus the salve (straw) is merged with the bacteria (pole) in hyperspace such that the bacteria is killed instantly. Removing the hand vortex brings the infected hand back into dimension cured.

[0031] Based on this information and the results of many experiments, this invention creates a large vortex by walking at a certain velocity with the hands crossed over the chest. The proper walking momentum is created by a computer program that inputs the person's weight, shoe length, and the number of strides to be taken. The program then prints out a banner showing the footprints where the person has to step each second. When a person obtains sufficient energy from these methods, the person is then tuned to the subspace geometry of the universe as will be shown using the tetrahedron physics diagram.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0032] **FIG. 1**. Energy vortices of the human body.

[0033] **FIG. 2**. Seven potential wells fed by the energy vortices of the body.

[0034] **FIG. 3**. Carved wooden statue of San Martin de Porres who could walk through solid wooden doors.

[0035] **FIG. 4**. Tetrahedron diagram showing boundaries of the Planck box of dimension.

[0036] **FIG. 5**. Tetrahedron diagram showing mass of water molecule.

[0037] **FIG. 6**. Tetrahedron diagram showing water molecule energy and mass are equal at the Planck box boundary.

[0038] **FIG. 7**. Tetrahedron diagram showing that water molecule momentum is tangent to the boundaries of the Planck box.

[0039] **FIG. 8**. Tetrahedron diagram showing the momentum ratio required to reach the Planck wavelength boundary from the water momentum.

[0040] **FIG. 9**. Perspective view of crossed-hand momentum walking using banner printout.

[0041] **FIG. 10**. Perspective view of vortex generated by momentum walking.

[0042] **FIG. 11**. Tetrahedron diagram showing how additional hyperspace energy supplied to the potential wells of the hyperspace energy being enables the human body to be pulled out of dimension.

[0043] **FIG. 12**. Tetrahedron diagram showing the inverted tetrahedrons whose crossing represents the merging of two worlds between space and hyperspace.

[0044] **FIG. 13**. Tetrahedron diagram showing that the proton wavelength is determined by the Planck mass tangent to the inverted tetrahedrons.

[0045] **FIG. 14**. Tetrahedron diagram showing that the mass of the universe determines the electron and proton elementary particles.

[0046] **FIG. 15**. Tetrahedron diagram showing that momentum walking together with the increased energy of the hyperspace energy being is tangent to the mass of the universe.

[0047] **FIG. 16**. Computer program block diagram for printing banner footprints.

[0048] **FIG. 17**. Computer program input dialog window.

[0049] **FIG. 18**. Computer program calculation of stride length per second.

[0050] **FIG. 19**. Six-stride screen banner printout for 220 lb. person.

[0051] **FIG. 20**. Project tree showing help information by double clicking on node.

DETAILED DESCRIPTION OF THE INVENTION

[0052] 1. Referring to **FIG. 12**, a tetrahedron (A, path abc) is added to the diagram. The tip of the tetrahedron (e) falls on the base constant which is equal to Planck's constant divided by the speed of light. A second tetrahedron (B, path def) is inverted around horizontal line (D) which has a geometrical relationship to the base constant. The line is located at

$$\text{centerline} = \frac{2}{\sqrt{3}}\ln\left(\frac{h}{c}\right) = -110.7536373$$

Notice that the intersection of the two inverted tetrahedrons (g) occurs at the Planck wavelength which is the boundary between space and hyperspace. Line (D) is referred to as the merging of two worlds or the connecting of two worlds, a phrase obtained by means of remote viewing. That is, it is the dividing line between space and hyperspace. This is the reason that the two boundaries intersect at this point.

[0053] Referring to **FIG. 13**, the proton wavelength (E) is added to the diagram. The proton wavelength has a value of the electron wavelength divided by 1836.1527

$$\ln(\lambda_p) = \ln\left(\frac{\lambda_e}{1836.1527}\right) = -34.26005901$$

A line (hd), from the base constant at the proton wavelength (h), to the corner of the inverted tetrahedron (d), intersects the merging of two worlds line at point (i). A circle, with a radius equal to the Planck mass (G), centered (i), is tangent to the inverted tetrahedrons. Thus the proton is defined by the base constant and the geometry of subspace. The reason that the proton is tangent to both tetrahedrons is because the electron and proton follow one single path between space and hyperspace. Thus there is only one single particle in nature. Because the particle enters our space at two different locations, we see the one particle as two distinct elementary particles. This relationship can be seen in Library of Congress tetrahedron diagram tet0565.

[0054] Referring to **FIG. 14**, the mass of the universe MU is equal to the linear mass Q of the universe times the radius R of the universe which is 10^{26} meters

$$\ln(MU)=\ln(\Omega R)=122.3347509$$

as shown on the diagram as line (A). The electron wavelength (B) reflects off the circumscribing sphere (H) and returns as the electron mass (C). The distance between reflection points is the hyperspace charge which is equal to the charge of space less the electron charge. So the electron goes from wavelength to electric charge to mass. The proton wavelength (D) reflects off the sphere and returns as the proton mass (E) which intersects the horizontal axis at point (c). A line (abc) from the mass of the universe at the vertical axis (a) to the proton at the horizontal axis (c) intersects the electron (b) which determines the electron's mass and wavelength since this point is on the 45° base line. What this means is that cosmology determines the values of the elementary particles.

[0055] Referring to **FIG. 15**, the mass of the water molecule (C) intersects the 45° base line at point (a). A circle (F), with a radius equal to the Planck mass is centered on point (c) at the Planck wavelength boundary (B) where the mass of the water molecule numerically equals the energy of the water molecule and where the speed of light is unity. The Planck mass is tangent to the base constant which is the vertical centerline (I) of the diagram. The large circle (H), centered (a) on the water molecule, is tangent to the mass of the universe (b) and tangent to the inverted tetrahedron (E). The difference between this circle (H) and the base constant (I) is the energy the hyperspace being has to acquire in order to be tuned to the mass and geometry of subspace. This difference, shown as circle (G), has a mass

$$m=e^{-4.792671}\ \mathrm{kg}=8.29\cdot10^{-3}\ \mathrm{kg}$$

Thus the hand vortex radius ratio has to be the square root of the mass of the hyperspace energy being divided by this tangent mass, or:

$$r = \sqrt{\frac{.071\ \mathrm{kg}\cdot 1\ \mathrm{in}^2}{e^{-4.792671}\ \mathrm{kg}}} \approx 3\ \mathrm{inches}$$

which is a vortex radius that is three times larger than that of a normal person, but one inch smaller than the third test subject who had a hand vortex radius of 4 inches. This is the reason that the third test subject was able to walk through walls and teleport to other locations because his energy was sufficiently large enough to cross over the inverted tetrahedron into a co-dimension of hyperspace. Notice also that dotted circle (J) with a radius equal to the mass of the water molecule, centered (c) on the boundary, is tangent to the tetrahedron (K). This makes the combined geometry tangent to the inverted tetrahedrons and the mass of the universe.

[0056] 2. Referring to **FIG. 16**, a computer program generates a banner printout with footprints spaced for walking according to the weight, shoe size and length of banner desired. Some banners could fit in a small room, or be placed on the floor of a long corridor. As shown in the block diagram the program inputs these three variables with error checking. Then the program calculates the stride length L per second from the equation

$$L = \frac{M}{W}T$$

[0057] 3. In the above equation, the value of the momentum M, as determined by the tetrahedron diagram, is preferably 50.9095736 kg m/s. The weight of the person is converted to mass W in kilograms. The stride period is preferably 1 second. Referring to **FIG. 17**, the data is entered in the dialog input window.

[0058] 4. Once the stride length has been calculated, the program displays the stride length and the required number of banner sheets in the message window, as seen in **FIG. 18**.

[0059] 5. The print banner menu is selected and the foot prints are printed on continuous banner paper. A six-stride scaled screen version of the banner is shown in **FIG. 19**. The banner paper is then placed on the floor, and a one-second beeping timer is activated from the toolbar or menu. The person then walks beside the printed foot prints, taking one stride per beep, which produces the correct walking momentum to generate the hyperspace vortex. The vortex, which forms in only six strides, brings additional hyperspace energy into the quantum potential wells of the hyperspace energy being.

[0060] 6. A help system consists of a project tree which explains the various steps in using the program. Double mouse clicking on a project tree node displays the help instructions in a dialog window as shown in **FIG. 20**.

[0061] 7. In summary, the purpose of the training system is to substantially increase the energy of a human being who will then have the capability of walking through walls, body levitation, instantaneous healing of infections, full-body teleportation to another location, remote viewing at vast distances in terms of light-years, and looking into hyperspace co-dimensions. The third test subject and I have been able to experience all the above phenomena. He did it through augmenting his energy, and I have done it through the application of electromagnetic fields, by spinning on my vortex accelerator machine and using this invention.

I claim:

1. A training method comprising the steps of:

generating a banner having a plurality of footprints spaced at regular intervals wherein the banner is placed on the ground;

generating a periodic audible signal, whereby the audible signal repeats at a regular interval of time equal to the period; and

walking on the banner by tracing the footprints spaced at regular intervals, wherein one step is made with each audible signal.

2. The method of claim 1, wherein the step of generating the banner further comprises:

providing a person's actual mass in kilograms; and

determining the stride length based upon the following equation:

$$L=(M/W)*T$$

where L is the stride length in meters, M is a constant of approximately 50.91, W is the mass of the person in kilograms, and T is the period of the audible signal in seconds, and

wherein the footprints on the banner are spaced at the stride length.

3. The method of claim 2, wherein the audible signal is a beep and the period of the audible signal is one second.

4. A training method for a person comprising the steps of:

generating an audible signal having a fixed period;

generating a banner having regularly spaced indicia for identifying preferred step locations, wherein the distance between adjacent indicia is determined by the following formula:

$$L=(M/W)*T$$

where the distance between adjacent indicia in meters is L, a constant of 50.9095736 is equal to M, the person's mass in kilograms is W, and the fixed period in seconds is T; and

walking on the marked path by stepping upon each of the regularly spaced indicia wherein one step is made with each period of the audible signal.

5. The training method of claim 4 wherein the fixed period is one second.

6. The training method of claim 5, wherein the regularly spaced indicia are footprints.

* * * * *

US 20030209636A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2003/0209636 A1**

St.Clair (43) **Pub. Date: Nov. 13, 2003**

(54) **BOBBIN ELECTROMAGNETIC FIELD PROPULSION VEHICLE**

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/142,583**

(22) Filed: **May 9, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64G 1/32**

(52) **U.S. Cl.** **244/166**

(57) **ABSTRACT**

This invention relates to a spacecraft which generates its own magnetic moment and magnetic field gradient in order to produce lift on the hull. The magnetic moment is generated by a large area solenoid located in the hull. A toroidal core wrapped with electrical bobbins at intervals along said core produces a traveling magnetic wave along its surface. This magnetic wave creates a spacetime curvature, similar to a tilted plate, which causes the formation of a magnetic field gradient. Power is not critical because the system uses a magnetic vortex wormhole generator to lower the speed of light in order to efficiently create highly relativistic fields due to Lorentz transformation.

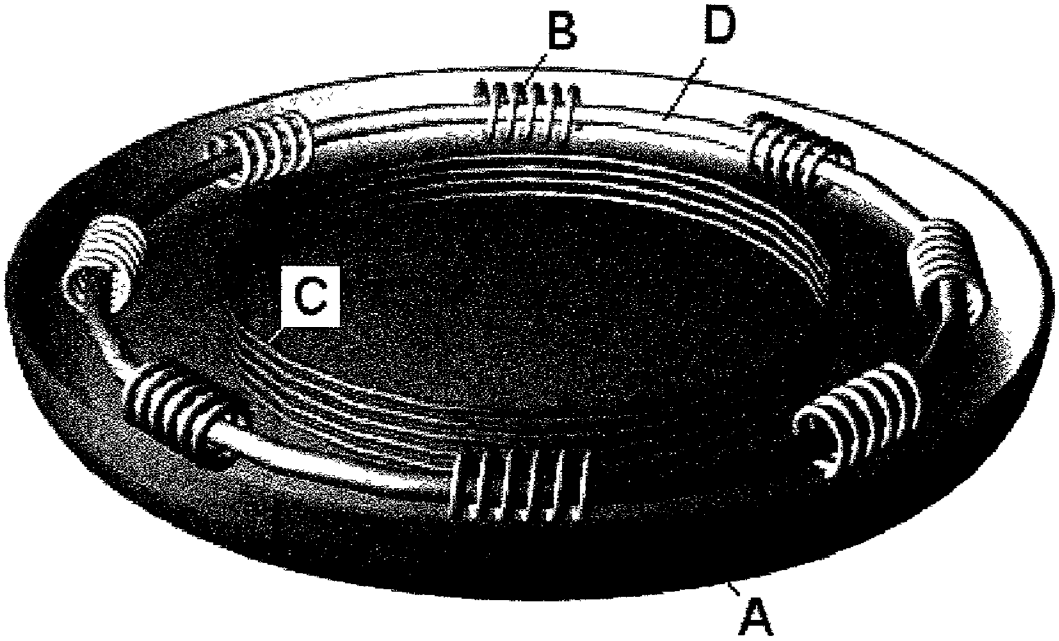

Figure 1

$$\text{Force} = \mu \cdot \frac{\partial B_z}{\partial z} = \text{amp} \cdot \text{m}^2 \frac{\text{kg}}{\text{sec} \cdot \text{coul} \cdot \text{m}} = \frac{\text{coul}}{\text{sec}} \cdot \text{m}^2 \cdot \frac{\text{kg}}{\text{sec} \cdot \text{coul} \cdot \text{m}} = \frac{\text{kg} \cdot \text{m}}{\text{sec}^2} = \text{newton}$$

Figure 2

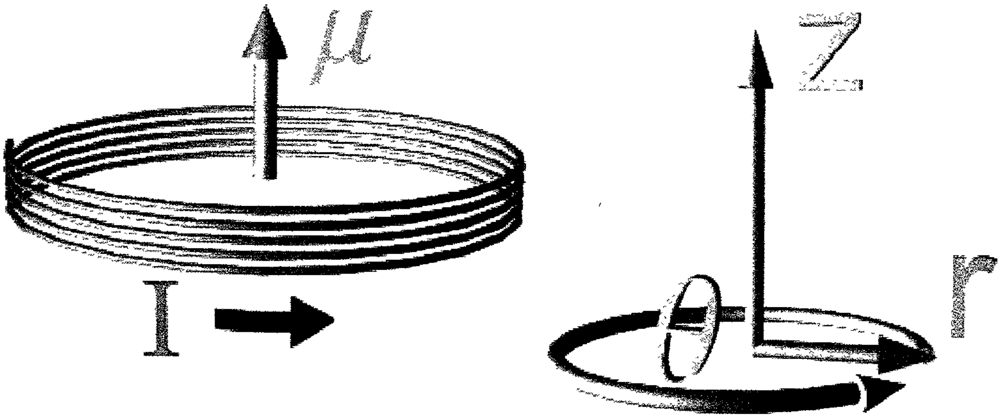

Figure 3

$$B_Z = \mu n I$$

Figure 4

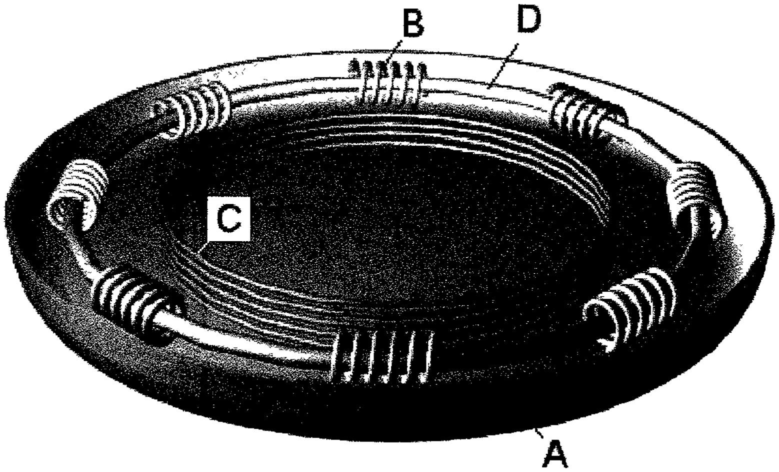

Figure 5

$$v = \sqrt{\frac{\omega}{\sigma\mu}}$$

Figure 6

$$g_{\alpha\phi} = \begin{array}{c|cccc|} & t & r & \theta & z \\ t & -1 & 0 & \mathrm{Sin}[\theta - t] & 0 \\ r & 00 & 1 & 0 & 0 \\ \theta & \mathrm{Sin}[\theta - t] & 0 & r^2 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

Figure 7

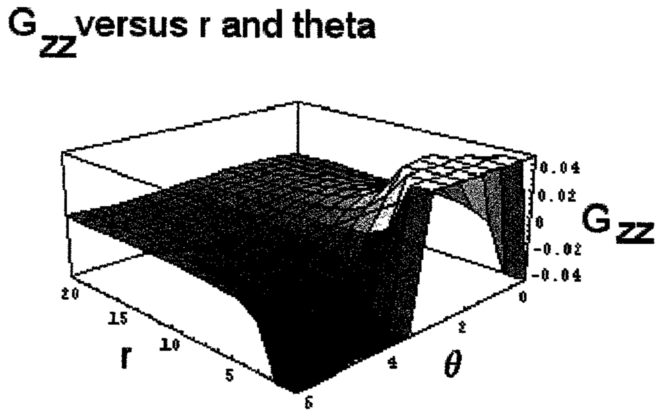

Figure 8

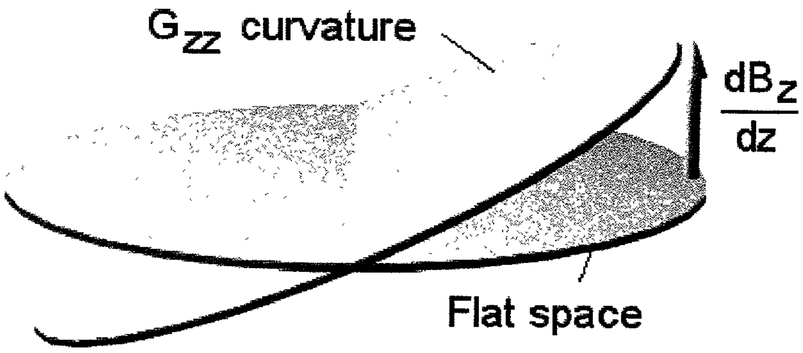
G_{zz} curvature
$\frac{dB_z}{dz}$
Flat space

BOBBIN ELECTROMAGNETIC FIELD PROPULSION VEHICLE

BRIEF SUMMARY OF THE INVENTION

[0001] This invention, which is the subject of my present application, is comprised of a toroidal core around which are wound a plurality of electromagnetic bobbins. The bobbins are pulsed electrically to create an amplified magnetic wave that travels around the core. At the same time, a circular electrical conductor carrying direct current creates a magnetic field around its area which results in the formation of a magnetic moment. This magnetic moment, together with the spacetime curvature distortion created by the traveling magnetic wave, produces a lift force on the vehicle.

REFERENCE PAPERS

[0002] Levitron, Hones, U.S. Pat. No. 5,404,062.

[0003] *Geometry of Electromagnetic Systems,* Paul Hammond, page 179.

BACKGROUND OF THE INVENTION

[0004] The idea for this invention comes from experiments I have done using thin transformer laminations wound in intervals with bobbins of wire connected to a frequency generator. Pulsing the bobbins electrically creates a slow traveling magnetic wave along the surface of the lamination. The velocity of the wave, as shown by Hammond in the reference paper, is the square root of the frequency divided by the conductivity and permeability of the material. From Einstein's General Theory of Relativity, this type of wave around the circumference creates a spacetime curvature distortion in the vertical direction that looks like a tilted plate. The magnetic field traveling around the circumference has to follow this curvature. This creates a magnetic gradient in the z-direction which together with the magnetic moment developed by a solenoid produces lift on the vehicle.

SUMMARY OF THE INVENTION

[0005] It is the object of this invention to create a magnetic field gradient and magnetic moment in order to produce a lift force on the hull of a vehicle. The magnetic field gradient is produced by a traveling magnetic wave which produces a tilting-plate spacetime curvature around the hull. The magnetic moment is created by a simple circular wire carrying direct current around its area.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0006] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0007] **FIG. 1**. The lift force on the vehicle is equal to a magnetic moment times the magnetic field gradient.

[0008] **FIG. 2**. Magnetic moment μ created by direct current I flowing counterclockwise around the wire coil. Cylindrical coordinates are shown to the right.

[0009] **FIG. 3**. The wire coil solenoid creates a magnetic field in the z-direction equal to the permeability of space times the number of turns per length of the solenoid times the current in the windings.

[0010] **FIG. 4**. Perspective view of bobbin spacecraft.

[0011] **FIG. 5**. Magnetic wave velocity along core.

[0012] **FIG. 6**. The g metric tensor in cylindrical coordinates with the sinusoidal wave function in the $\{t,\theta\}$ slots.

[0013] **FIG. 7**. The spacetime curvature G_{zz} in the vertical direction created by the magnetic wave traveling around the core.

[0014] **FIG. 8**. Tilted plate spacetime curvature showing magnetic gradient.

DETAILED DESCRIPTION OF THE INVENTION

[0015] 1. Referring to the equation in **FIG. 1**, the lift force on the bobbin spacecraft is equal to its magnetic moment times its magnetic field gradient. The magnetic moment has units of electrical current, measured in amps, times the area enclosed by the current. Thus the units of the magnetic moment are amp$-$m^2. The magnetic moment can be created by a large, circular coil of wire carrying direct current I as shown in **FIG. 2**. The spacetime cylindrical coordinates $\{t, r, \theta, z\}$ are shown to the right of the drawing where t is time, r the radius, the horizontal angle θ, and the vertical height z.

[0016] 2. The wire coil solenoid also produces a magnetic field in the vertical z-direction equal to the permeability of space μ times the number of turns per unit length of the solenoid n, times the current I in the winding. The equation is shown in **FIG. 3** which can be developed from Ampere's law that the magnetic field around a loop is equal to the current passing through the loop.

[0017] 3. Referring to **FIG. 4**, the bobbin spacecraft consists of a large, horizontal, highpermeability toroidal core (D) wrapped at intervals with electrical bobbins (B) which can be pulsed electrically to create a traveling magnetic wave around the core. Interior to this core is a large direct current solenoid (C) which produces said magnetic field in the z-direction. These devices are enclosed in a circular hull (A) containing the coils on the outer rim and a cabin area in the center.

[0018] 4. Referring to **FIG. 5**, the velocity of the magnetic wave on the surface of the core is equal to the square root of the wave frequency ω divided by the conductivity a times the permeability μ of the core material. When the first bobbin is pulsed, a wave starts to propagate along the surface of the core. As the wave passes the second bobbin in sequence, another electrical pulse is generated to amplify the wave. After many cycles, the wave traveling around the core becomes larger and larger in amplitude.

[0019] 5. A traveling wave has a wave function equal to a sinusoidal function with an argument of the angle θ around the periphery less the time t, or $\mathrm{Sin}[\theta-\omega t]$.

[0020] 6. In gravitational physics, there is a g metric tensor which is a measure of length in spacetime coordinates. When mass or electromagnetic fields are involved in a certain region of space, a curvature of space is created. The curvature of space can then be calculated directly from this metric tensor. The result is Einstein's G curvature tensor which shows the spacetime distortion. The g metric tensor is a 4 by 4 matrix having rows and columns equal to time t, radius r, angle theta, and vertical height z. The diagonal from top left to bottom right has a signature equal to $\{-1, 1, r^2, 1\}$.

[0021] 7. Because the magnetic wave traveling around the core is varying with time t in the theta direction, the wave function has to go into the $\{t,\theta\}$ and $\{\theta,t\}$ slots of the metric tensor, as shown in **FIG. 6**.

[0022] 8. From this metric tensor, Einstein's G curvature tensor is calculated using a general relativity software program. The spacetime curvature in the vertical z-direction is contained in the G_{zz} component. A 3D plot of this curvature with respect to radius and theta is shown in **FIG. 7**. The axis on the right side is the angle around the periphery, and the axis on the left is the radius which goes from 0 to 20 meters. The vertical axis of the graph is the spacetime curvature in the vertical z-direction. The center of the vertical axis is zero. At a small radius, there is a sinusoidal curvature which is positive from 0 to π and which is negative from π to 2π. What this looks like is a tilted plate as depicted in **FIG. 8**.

[0023] 9. In flat spacetime with no electromagnetic fields or mass, the curvature would be the horizontal plate as seen **FIG. 8**. Due to the traveling magnetic wave, the spacetime curvature looks like the tilted plate. The flat-space magnetic field was pointing up in the vertical direction, having no divergence and no gradient. In curved spacetime, however, the magnetic field becomes tilted just like the plate and a magnetic gradient dB_z/dz is created. This spacetime curvature gradient of the magnetic field times the magnetic moment of the second coil produces a lift force on the hull which is anchored to these coils.

What I claim as my invention is:

1. A spacecraft that generates a magnetic moment and a magnetic field gradient in the vertical direction in order to create a lift force on the hull.

2. A toroidal core wrapped with electrical bobbins at intervals around the core whose purpose is to create and amplify a magnetic wave that travels along the surface of the core.

3. A circular, direct-current carrying solenoid, located radially inside the toroidal core, to create the magnetic moment.

4. Said magnetic surface wave creates a spacetime curvature, similar to a tilted plate, which produces a magnetic field gradient in the vertical direction.

5. A circular hull, with interior cabin, electrical power supply to drive the coils, and a computer to calculate and sequence the activation of the electrical bobbins.

6. Magnetic sensors, in the form of small coils, located on the toroidal core which can detect the velocity and position of the traveling magnetic wave as feedback to the computer control system.

7. A magnetic vortex generator, either rotating magnet or dual coil, to produce a wormhole through which low linear mass and low speed of light hyperspace energy can enter the hull in order to create highly relativistic electromagnetic fields.

8. A variable current generator to modulate the direct current in the coil so as to produce a variable magnetic moment that can control lift, hover and descent.

9. An electrical power supply using mechanical flywheel stored energy together with a large area capacitor, resupplied with energy by solar cells located on the outer surface of the hull.

* * * * *

US 20040164824A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2004/0164824 A1**

St. Clair (43) **Pub. Date: Aug. 26, 2004**

(54) **HYPERSPACE ENERGY GENERATOR**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/372,556**

(22) Filed: **Feb. 21, 2003**

Publication Classification

(51) **Int. Cl.^{7}** **H01P 3/06**
(52) **U.S. Cl.** **333/243**

(57) **ABSTRACT**

This invention is a braided gold wire coaxial cable of micron size which generates hyperspace energy by coupling to the tetrahedral geometry of subspace, dimension and the Planck mass.

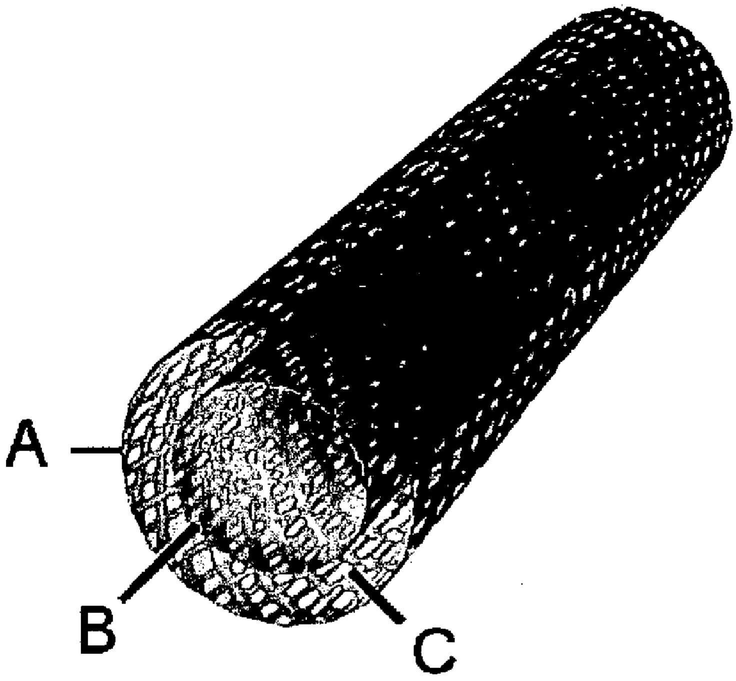

Figure 1

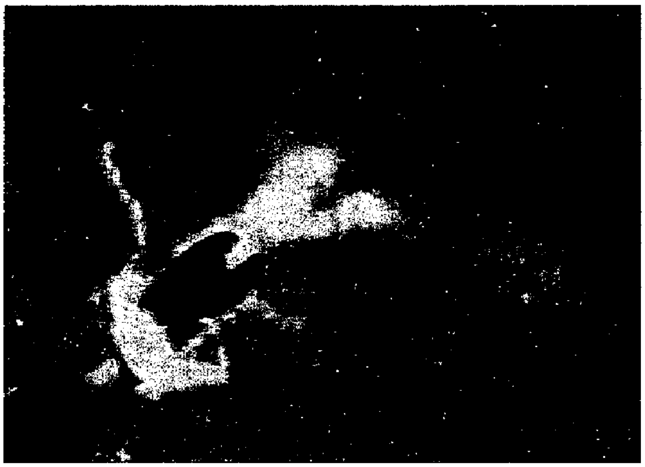

Figure 2

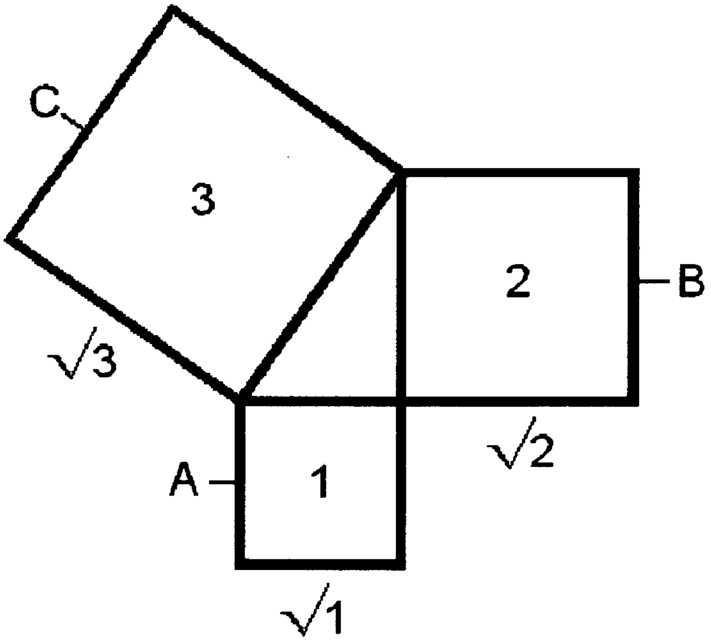

Figure 3

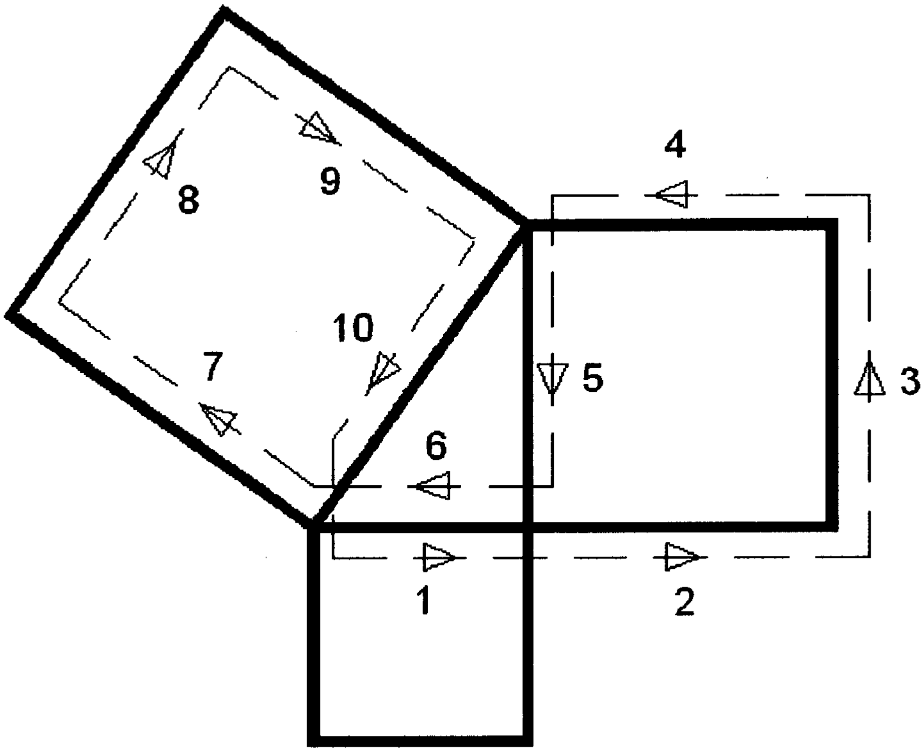
4
8
9
10
7
5
3
6
1
2

Figure 4

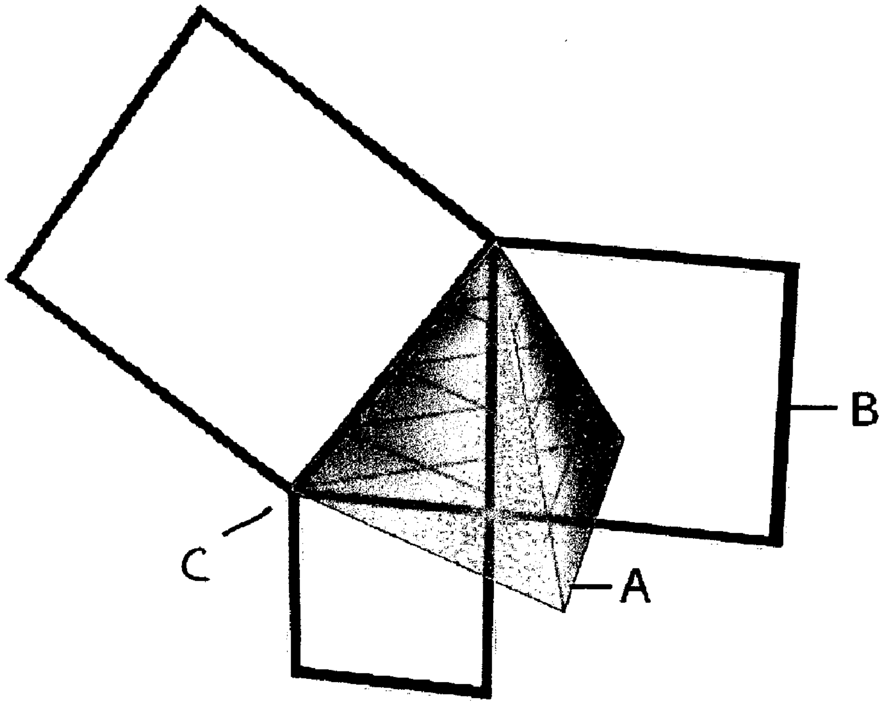

Figure 5

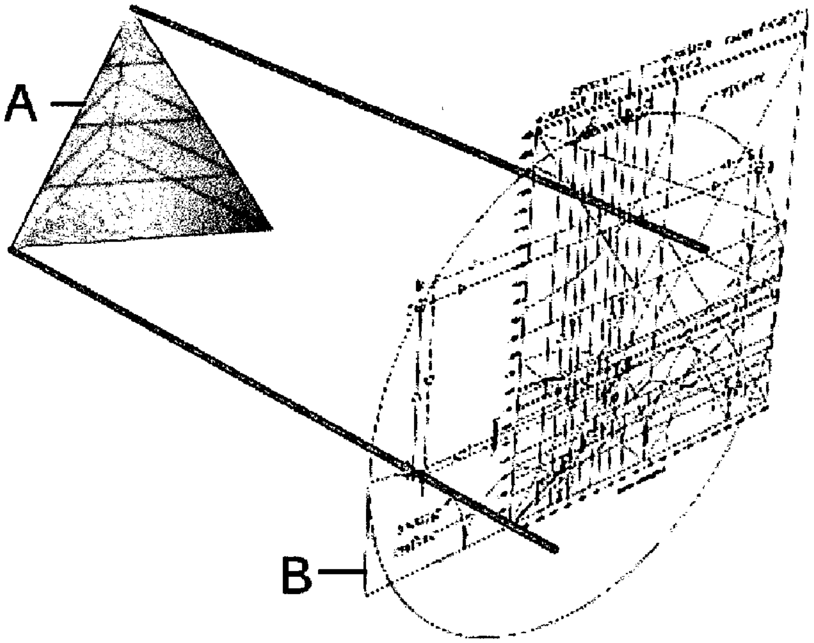

Figure 6

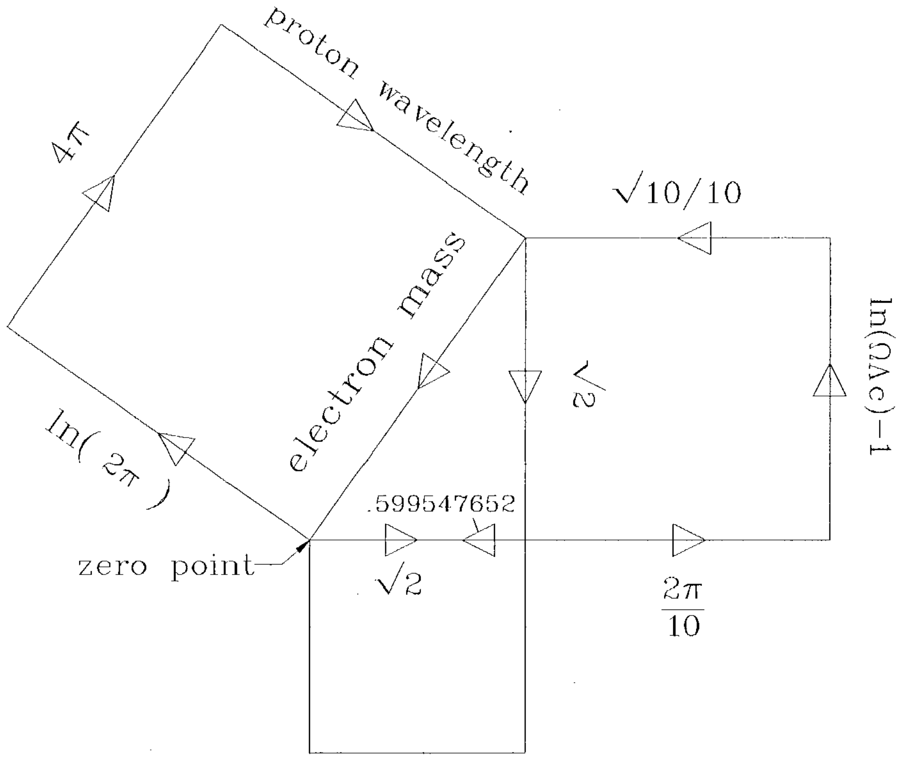
proton wavelength
4π
electron mass
ln(2π)
√10/10
√2
ln(ΩΛc)−1
.599547652
zero point
√2
2π/10

Figure 7

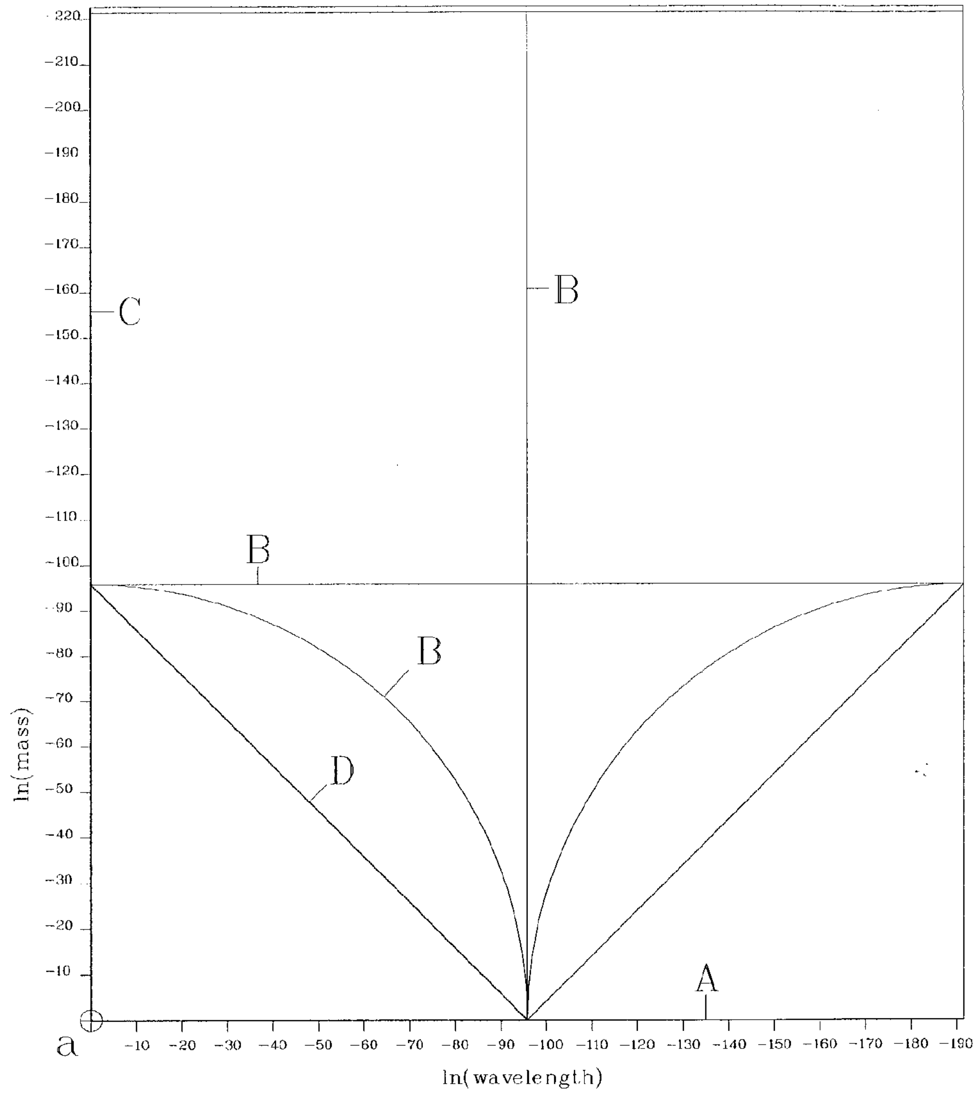
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
C
B
B
B
D
A
a
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)

Figure 8

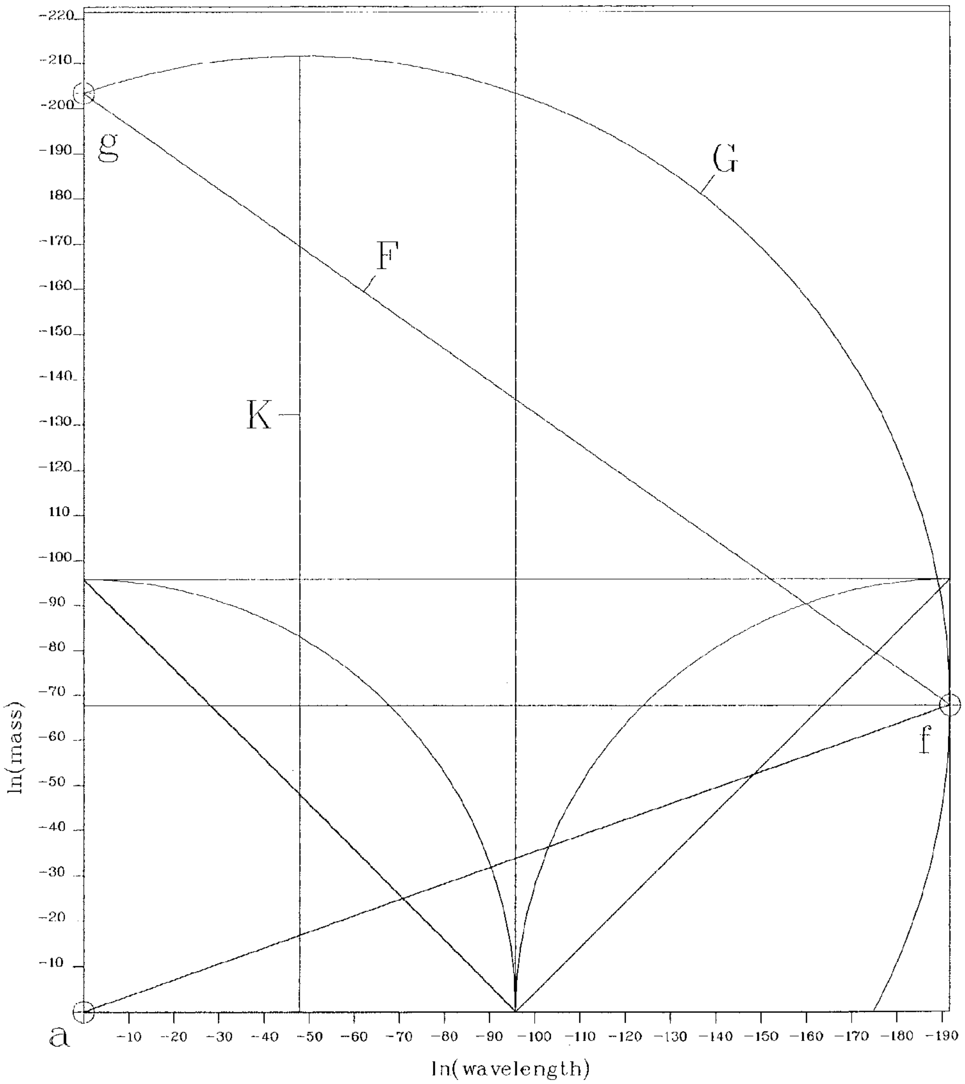
-220
-210
-200
-190
180
-170
-160
-150
-140
-130
-120
110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
g
G
F
K
f
a

Figure 9

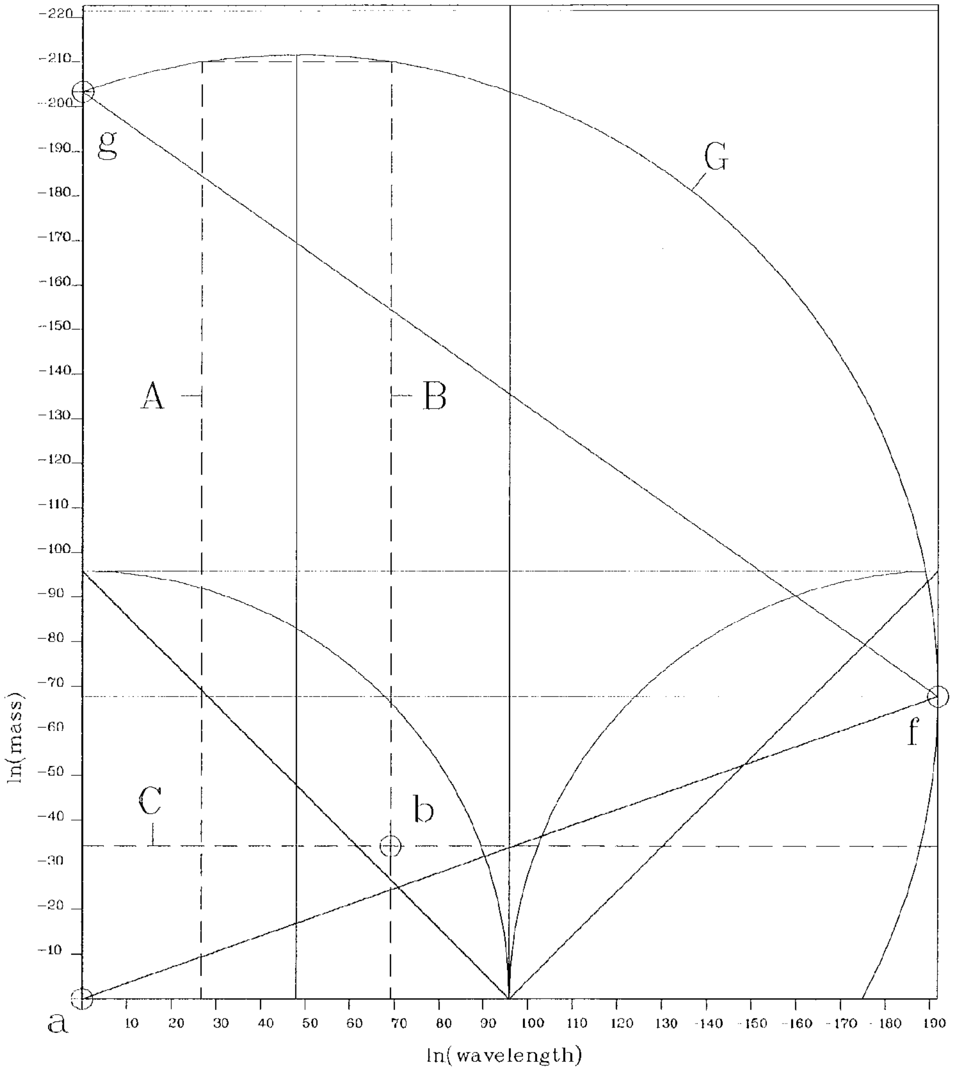
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
g
G
A
B
C
b
f
a
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
ln(wavelength)

Figure 10

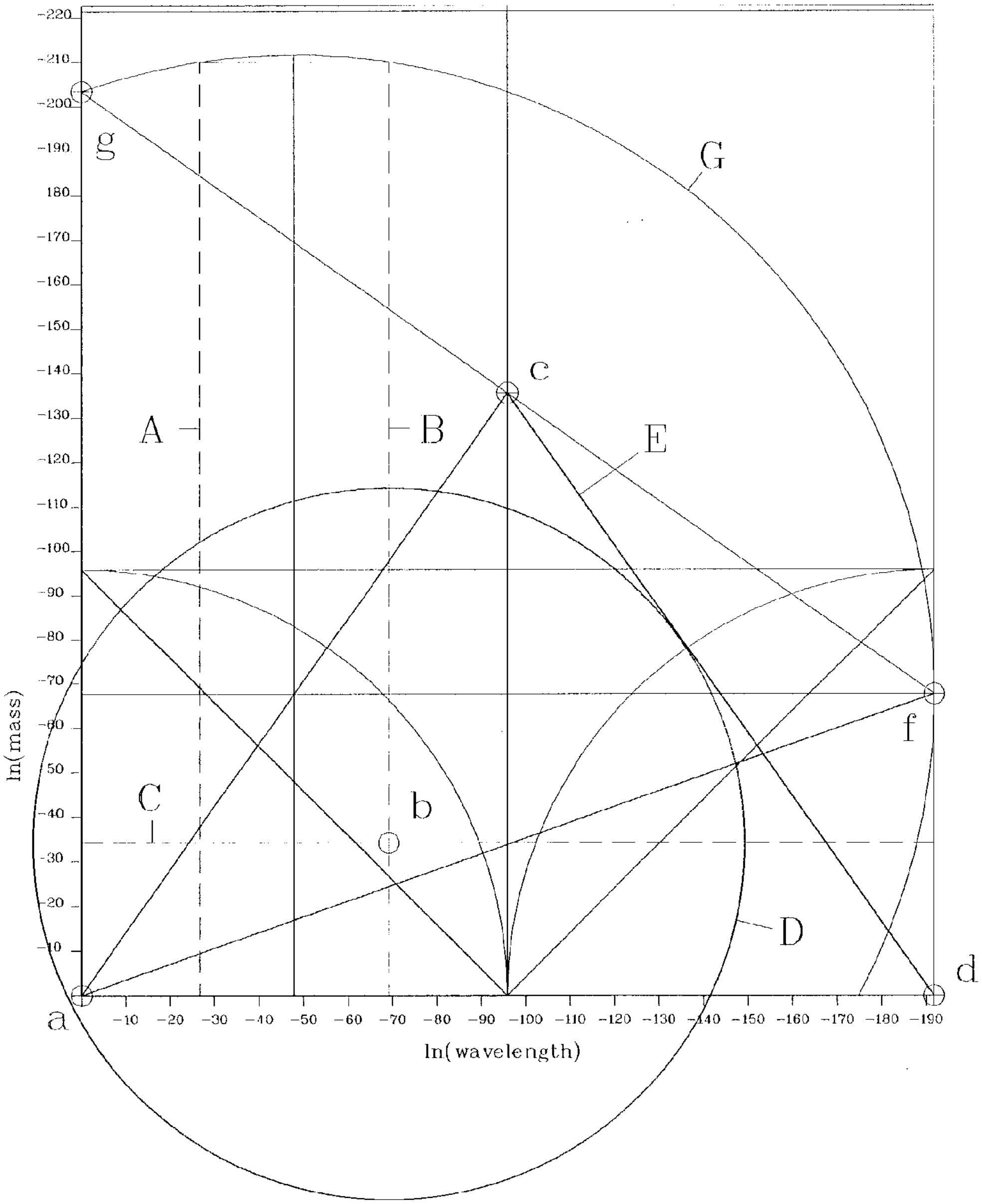
-220
-210
-200
-190
180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
g
G
c
A
B
E
f
C
b
D
a
d

Figure 11

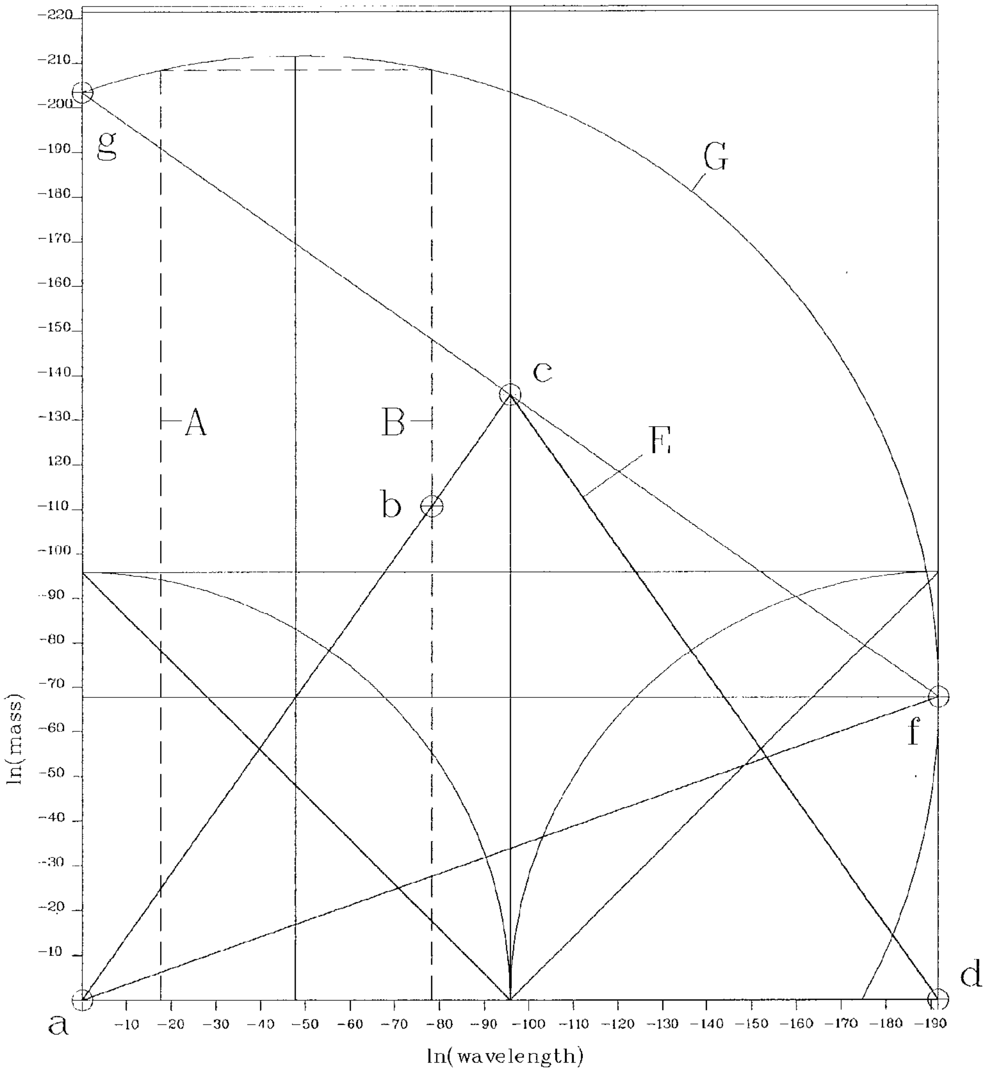
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
g
G
c
A
B
F
b
f
d
a
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 12

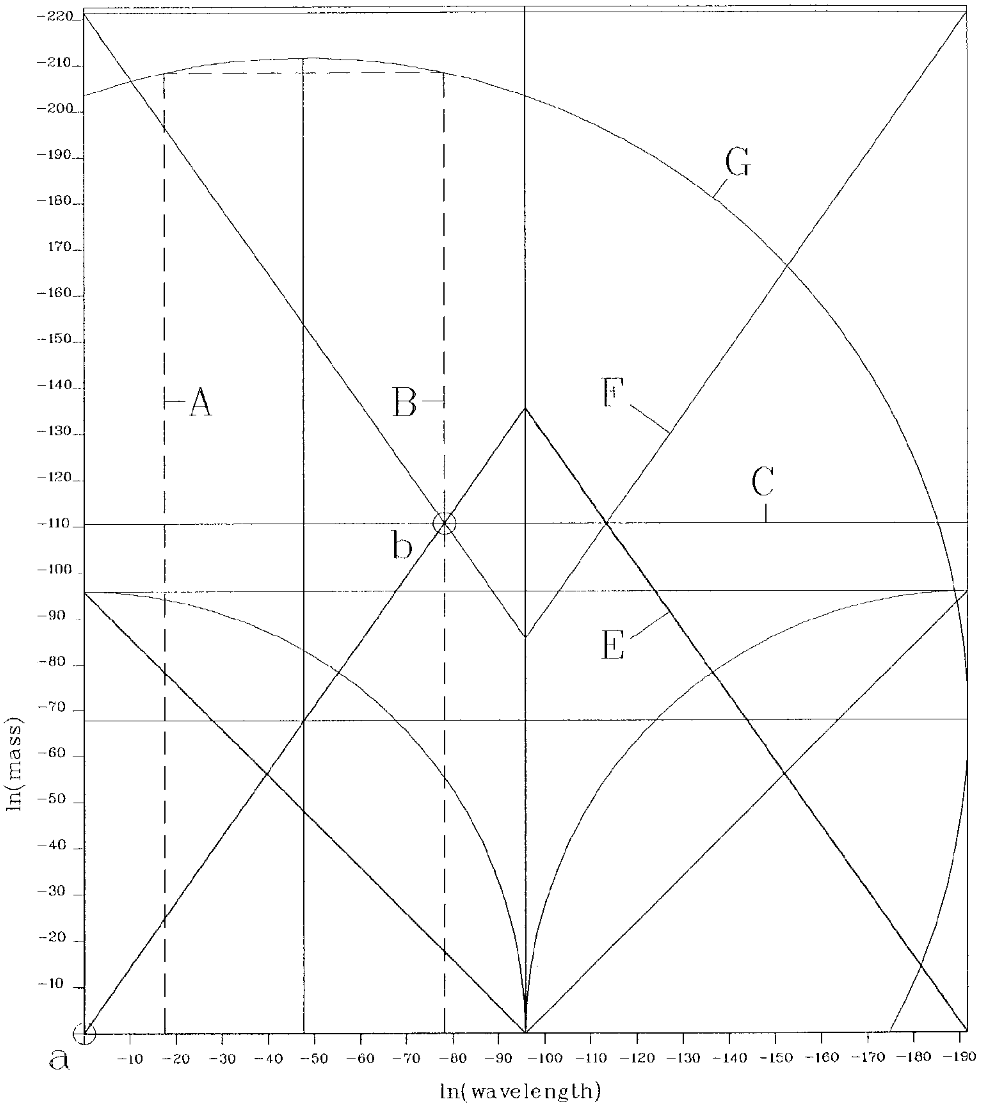
-220
-210
-200
-190
-180
170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
a
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
A
B
b
C
E
F
G

Figure 13

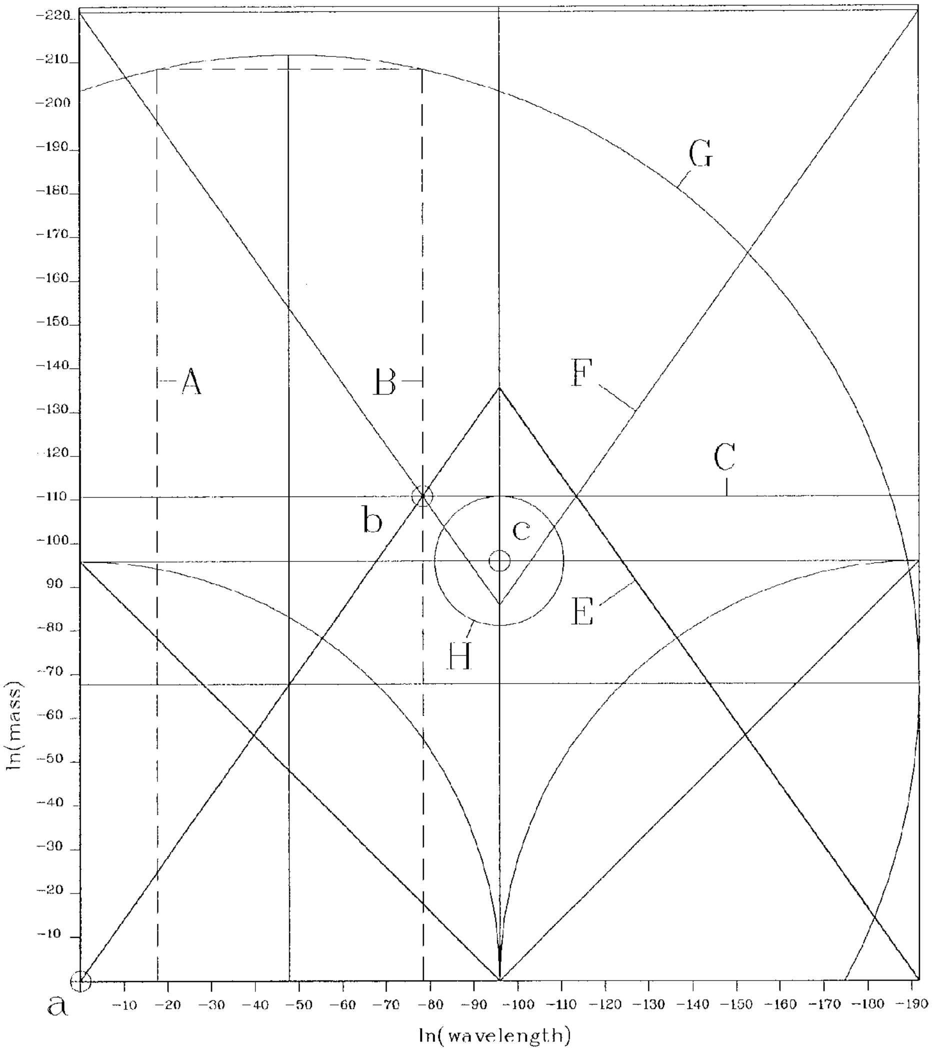
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
120
-110
-100
90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
A
B
C
E
F
G
H
a
b
c
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 14

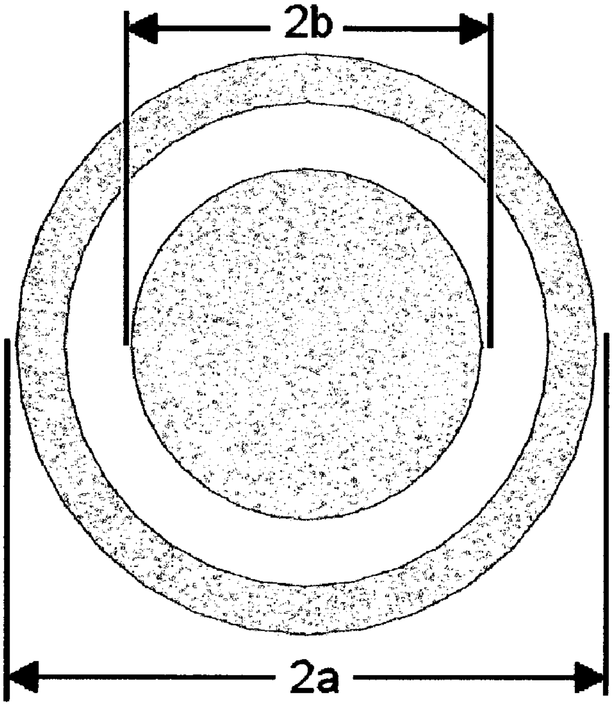

Figure 15

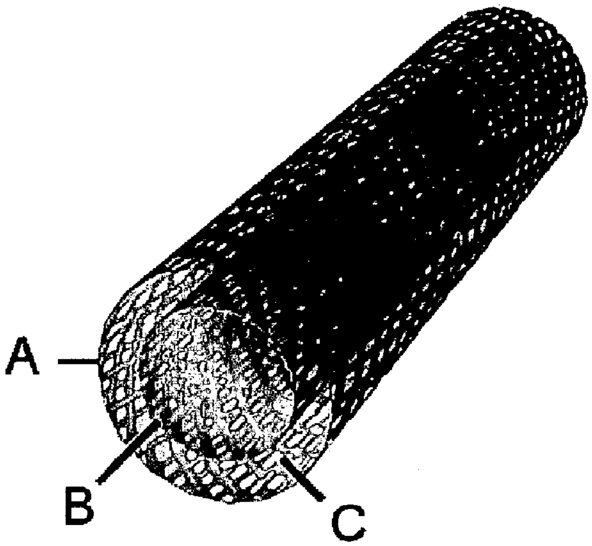

HYPERSPACE ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a braided gold wire coaxial cable of micron size which generates hyperspace energy.

BACKGROUND OF THE INVENTION

[0002] Electrical experiments with micron-sized braided gold wire coaxial cable show that it is capable of generating substantial amounts of hyperspace energy. Referring to the electron microscope photograph shown in **FIG. 1**, the white mist emanating from the cable is low-density hyperspace energy that is flowing in from a co-dimension of our universe. The dimensions of the cable are of such a particular size as to couple the cable to the tetrahedral geometry of subspace, the dimension of space, the Planck mass and the linear inductance of the universe.

[0003] According to physicist Dr. Edward Witten of Princeton University, space has twenty-four dimensions, of which ten dimensions are non-redundant. Imagine taking a path around the Pythagorean triangle, as known as the planar tetrahedron, with sides equal to $\{\sqrt{1}, \sqrt{2}, \sqrt{3}\}$ as shown in **FIG. 2**. There are three squares denoted the one-square (A), the two-square (B) and the three-square (C). Each square has four sides. The edge of each square can be traversed in two directions. Thus the total number of dimensions is

$$dim_{\text{total}}=2\cdot(4+4+4)=24$$

[0004] Referring to **FIG. 3**, there is a path starting at the corner of the triangle, along the one-square (**1**), around the two-square (**2,3,4,5**), back along the one-square (**6**), around the three-square (**7,8,9,10**) and back to the corner of the triangle. The numbering of the edges shows that there are ten edges. Because the path is traversed in only one direction, the number of reduced dimensions is

$$dim_{\text{reduced}}=1\cdot(10)=10$$

[0005] Referring to **FIG. 4**, the planar tetrahedron (B) forms one edge of the three-dimensional tetrahedron (A). Rotating the planar tetrahedron ±120° produces the other two edges. The tetrahedron has four faces which are equilateral triangles. The ten dimensional path starts and ends at (C), the corner of the tetrahedron known as the zero point.

[0006] Referring to **FIG. 5**, the projection of the 3D tetrahedron (A) onto a plane is called the tetrahedron diagram (B) which is the main diagram of the new geometrical physics known as Δphysics. All the constants of physics can be derived geometrically from the tetrahedron diagram and its associated planar tetrahedron. An example of this is shown in **FIG. 6** where the edges are given specific constants related to tetrahedral geometry, dimension, curvature, and the mass and wavelength of the elementary particles such as the electron and proton. The ten dimensional path includes the following constants

[0007] a. electron wavelength

[0008] b. proton wavelength

[0009] c. 47π The solid angle of the sphere. The tetrahedron is circumscribed by a sphere.

[0010] d. ln(2π) The natural log of the curvature. The subspace geometry is a logarithmic manifold. The tetrahedron diagram plots the logarithm of mass versus the logarithm of wavelength.

[0011] e. 0.599547652 A constant related to fractal dimension and the speed of light factor 2.99792458.

[0012] f. $\sqrt{2}$ The edge of the two-square.

[0013] g. $\sqrt{10}/10$ The square root of ten dimensions per 10 dimensions.

[0014] h. ln(ΩΛc)−1 The natural log of the momentum of space less one.

[0015] i.

$$\frac{2\pi}{10}$$

[0016] The curvature of space per 10 dimensions.

[0017] j. $\sqrt{2}$ The edge of the two-square.

[0018] The length of each edge is multiplied by the constant assigned to that edge. The ten edges have the order of $\{\sqrt{1}, \sqrt{2}, \sqrt{2}, \sqrt{2}, \sqrt{1}, \sqrt{3}, \sqrt{3}, \sqrt{3}, \sqrt{3}\}$. What subspace geometry does is to multiply the edge length, such as $\sqrt{1}$, times the square root of two $\sqrt{2}$. Then it takes the square root of that number and multiplies it by the next edge, which is $\sqrt{2}$, times the curvature per 10 dimensions

$$\frac{2\pi}{10}.$$

[0019] It then takes the square root of that number and so on. In equation form, this looks like the following calculation

$$\sqrt{3a\sqrt{3b\sqrt{3c\sqrt{3d\sqrt{1e\sqrt{2f\sqrt{2g\sqrt{2h\sqrt{2i\sqrt{1j}}}}}}}}}} = 80.1104395$$

[0020] where the letters correspond to those in the list of constants. The square root sum total is equal to the Planck scale Λwhich is the bottom dimensional limit of the universe. The sum of the ten constants per a speed of light circumference is equal to unity

$$\frac{\sum\limits_{n=10} s_n}{2\pi\ln(c)} = 1.000000000$$

[0021] where the log of the speed of light is

$$\ln(c)=\ln(299792458)=19.51860099$$

[0022] and multiplying by 2π is the circumference of a circle with a radius equal to the speed of light.

[0023] Referring to **FIG. 7**, the tetrahedron diagram plots the natural logarithm of mass on the vertical axis (C) versus the natural logarithm of wavelength on the horizontal axis (A). The reason for this is that the mass of the electron times its wavelength is equal to the mass of the proton times its wavelength which in turn is equal to Planck's constant h divided by the speed of light, known as the base constant (B). If two numbers multiply, they sum in logarithms. In subspace geometry, the sum of the logarithm of the mass of the electron plus the logarithm of the wavelength is equal to the logarithm of the base constant which has a value of −95.91546344.

$$\ln(m_{electron}) + \ln(\lambda_{electron}) = \ln\left(\frac{h}{c}\right) = -95.91546344$$

[0024] What this means is that the mass and wavelength slide on a 45° base line (D) which has end points on the vertical and horizontal axes equal to the base constant.

[0025] Referring to **FIG. 8, a** line (af) drawn from the origin at the tetrahedral angle of 19.47122063, equal to the asin(⅓), creates a tetrahedron (F) along path (afg). This tetrahedron is circumscribed by a sphere (G) with sphere diameter (K).

[0026] The Planck scale path calculation showed that the electron mass and the proton wavelength were the last two edges. The electron mass has a value of

$$\ln(m_e)=-69.17083217$$

[0027] and the electron wavelength has a value of

$$\ln(\lambda_E)=-26.74463127$$

[0028] Referring to **FIG. 9**, the electron wavelength (A) is plotted as a vertical line on the tetrahedron diagram. The wavelength reflects off the circumscribing sphere (G), and returns as the electron mass (B). So the diagram incorporates the concepts of both classical physics (point mass particles) and quantum mechanics (wave particles).

[0029] The proton wavelength has a value of

$$\ln(\lambda_p)=-34.26005901$$

[0030] The proton wavelength (C) is plotted as a horizontal line in order to get the intersection (b) with the electron mass.

[0031] Referring to **FIG. 10, a** circle (D) with a radius equal to the Planck scale is drawn centered (b) on the intersection of the electron mass with the proton wavelength, which are the last two edges of the Planck scale calculation. A line (ac) from the origin to the intersection of the base constant with the rotated tetrahedron creates the vertical tetrahedron (acd). As can be seen, the Planck scale is tangent to the tetrahedron on side (cd). This tetrahedron is the projection of the 3D tetrahedron shown before in **FIG. 5**. What this means is that the tetrahedral geometry of subspace determines the bottom limit of our universe. And this bottom limit, called the Planck scale, contains within itself the mass and wavelength of the elementary particles, curvature, dimension and planar tetrahedral geometry. Tetrahedron diagram tet0565 shows that the electron and proton are one and the same particle because the electron path rotates counterclockwise around the curvature and then returns clockwise as the proton path. This path occurs moving through space and hyperspace. Because the single particle enters our universe from hyperspace at two different positions, we see it as two distinct particles. Thus the tetrahedron diagram shows that hyperspace exists.

[0032] The speed of light is equal to the inverse of the square root of the permeability μ of space times the permittivity ϵ of space

$$c = \sqrt{\frac{1}{\mu\varepsilon}}$$

[0033] The permeability is linear inductance or inductance per length which you would find in a solenoid for example. The permittivity is linear capacitance or capacitance per length which you find in a capacitor. In an electrical circuit, the inductance and capacitance form a resonant circuit. The resonance frequency can be changed by changing the inductor or capacitor. In a similar manner, the speed of light is not constant, but can be lowered by increasing the permittivity. Hyperspace energy has a high permittivity and therefore a low speed of light. This low speed of light gives hyperspace energy a luminescent quality which is seen as a white mist (**FIG. 1**).

[0034] From Einstein's General Theory of Relativity, the stress pressure T on spacetime is proportion to the square of the ratio of the electric field E to the speed of light c.

$$T = \left(\frac{E}{c}\right)^2$$

[0035] Thus substantially lowering the speed of light creates an enormous spacetime pressure which can be used to generate the lift force on electromagnetic field propulsion vehicles. Furthermore, the electric field is subject to the Lorentz transformation

$$E = \frac{E_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

[0036] The electric field E_0 moving in a frame velocity of v, can quickly attain relativistic proportions because the speed of light could be 1 meter per second, rather than the enormous value in our universe of 299792458 meters per second. Thus one would like to permeate the hull of the electromagnetic field propulsion vehicle with this hyperspace energy in order to increase the electric field and hence the spacetime curvature around the hull which produces the enormous lift force on the vehicle. The method of bringing in this hyperspace energy is to use braided gold wire coaxial cable which is coupled to the geometry of subspace. The subspace geometry is contained in the Δphysics tetrahedron diagram.

[0037] Just as space has a linear inductance and linear capacitance, it also has a linear mass Ω or mass per meter.

Physicist Dr. John A. Wheeler of Princeton likes to invert this and call it "mom" for meter of mass. The Planck mass is equal to the Planck scale Λtimes the linear mass Ω

$$\ln(m_{\mathrm{Plank}})=\ln(\Omega\Lambda)=-17.64290101$$

[0038] Planck's constant h is equal to 2π times the Planck scale squared times the linear mass Ω times the speed of light c.

$$h=2\pi\Lambda^2\,\Omega c=2\pi\Lambda(\Omega\Lambda)c$$

[0039] which shows that Planck's constant is actually the circumference of a circle of radius Planck scale times the Planck mass times the speed of light. The base constant is therefore

$$\mathrm{base}=\frac{h}{c}=\frac{2\pi\Lambda(\Omega\Lambda)c}{c}=(2\pi\Lambda)(\Omega\Lambda)=-95.91546344$$

[0040] which is an area, known as the Planck box, bounded by the Planck wavelength (2πΛ) and the Planck mass. Everything outside the Planck box is hyperspace. Everything inside the Planck box is our universe. Thus the boundary between space and hyperspace is the Planck wavelength and the Planck mass. In logarithms, notice that the Planck mass and Planck wavelength, just like the electron and proton, sum to the base constant.

[0041] Referring to **FIG. 11**, the Planck mass (A) and the Planck wavelength (B) are plotted on the diagram and reflected off the sphere. The Planck wavelength intersects the tetrahedron at (b) which is the boundary between space and hyperspace known as the centerline of the diagram. The centerline has a value equal to the base times the square root of 4/3.

$$centerline=\sqrt{\frac{4}{3}}\,base=-110.7536373$$

[0042] Referring to **FIG. 12**, the centerline (C) is drawn on the diagram and the tetrahedron (E) is mirrored (F) across the centerline to indicate the co-dimensions of hyperspace.

[0043] Referring to **FIG. 13, a** circle (H), centered at the base at the base (c), tangent to the centerline (C), has a radius equal to base times the square root of 4/3 less one

$$R=\left(\sqrt{\frac{4}{3}}-1\right)base=14.83817383$$

[0044] This is the length that has to traversed in order to cross over the centerline from the base constant of our universe to the co-dimension of hyperspace. Furthermore, the Planck mass, which is the other boundary, has to be crossed in order to get to either axis as seen by the length between the vertical axis and line (A). It can be looked at also as the length needed to go from the Planck wavelength (B) to the tangent point of circle (H) in order to reach the center of the mirror tetrahedrons.

SUMMARY OF THE INVENTION

[0045] This invention is a braided gold wire coaxial cable of micron size that is coupled to the subspace geometry of the universe for the purpose of bringing in low-density hyperspace energy into our universe from the co-dimensions of hyperspace. The dimensions of the coaxial cable are of such particular size as to enable it to couple to the ten dimensions of space, the 3:1 geometrical ratio of the tetrahedron, the coaxial wave function based on the logarithm of the ratio of the outer radius to the inner radius of the cable, the length between the base constant of our universe and the centerline between mirror tetrahedrons, and the geometrical relationship between the Planck mass and the linear inductance of the cable.

[0046] Referring to **FIG. 14**, the coaxial guide has an outer braided conductor (A) and an inner braided conductor (B) in which the outer and inner conductors are at radii a and b respectively. The linear inductance L of the cable is equal to the natural logarithm of the ratio of the radii times the permeability μ of space divided by the curvature 2π

$$L=\ln\left(\frac{a}{b}\right)\frac{\mu}{2\pi}$$

[0047] It can be shown that the ratio of the area-to-volume ratio of the tetrahedron to the area-to-volume ratio of its circumscribing sphere is 3:1. It is also the ratio of the area of the three-square to the area of the one-square on the planar tetrahedron. It is also the tetrahedral angle asin(⅓) equal to 19.47122063°. It is also a maximum work condition between the velocity ratio of a fluid stream and a moving vane such as in turbomachinery. So the first constraint on the radii is

$$\exp\left(\ln\left(\frac{a}{b}\right)+1\right)\equiv 3$$

[0048] The second constraint is that the radius c of the circle, equal to the difference between the base constant and the centerline, is related to ten dimensions. The value of the radius projected into our universe is

$$c=\left(\sqrt{\frac{4}{3}}-1\right)\mathrm{base}=.3596358547\cdot\mathrm{microns}$$

[0049] and the dimensional constraint is

$$\left[\ln(c)+\ln\left(\frac{1}{a}\right)\right]^2=10$$

[0050] The coaxial wave function constraint is

$$\frac{\ln(a)^2}{2\pi\ln\left(\frac{a}{b}\right)} \equiv \ln(c)^2$$

[0051] The linear inductance of the coaxial cable has to be such that it gets geometrically across the Planck mass which is the second boundary of our Planck box

$$\exp\left(\ln(\Omega\Lambda) - \ln\left(\frac{\ln\left(\frac{a}{b}\right)}{2\pi}\mu\right) + 1\right) \equiv 3$$

[0052] With these four subspace constraints, the outer radius a and inner radius b of the cable are

[0053] a=8.342461828 microns

[0054] b=7.559058141 microns

[0055] so the cable has an outside diameter of roughly 16 to 17 microns.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0056] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0057] **FIG. 1**. Electron microscope photograph of the coaxial cable and hyperspace mist.

[0058] **FIG. 2**. The planar tetrahedron.

[0059] **FIG. 3**. The ten dimensional path around the planar tetrahedron.

[0060] **FIG. 4**. The relationship between the planar tetrahedron and the 3D tetrahedron.

[0061] **FIG. 5**. Projection of the 3D tetrahedron onto a plane to create the tetrahedron diagram.

[0062] **FIG. 6**. Dimension, geometry, curvature and elementary particles that make up the ten dimensional Planck scale path.

[0063] **FIG. 7**. Tetrahedron diagram showing base constant.

[0064] **FIG. 8**. Tetrahedron diagram showing rotated tetrahedron and circumscribing sphere with electron mass and wavelength reflecting off sphere.

[0065] **FIG. 9**. Tetrahedron diagram showing intersection of electron mass with proton wavelength which are components of the Planck scale path.

[0066] **FIG. 10**. Tetrahedron diagram showing Planck scale tangent to tetrahedron.

[0067] **FIG. 11**. Tetrahedron diagram showing Planck mass and Planck wavelength intersecting tetrahedron.

[0068] **FIG. 12**. Tetrahedron diagram showing mirror tetrahedrons across centerline, the boundary between space and hyperspace.

[0069] **FIG. 13**. Tetrahedron diagram showing distance between the base constant and the centerline used to calculate the dimensions of the coaxial cable.

[0070] **FIG. 14**. End view of coaxial cable showing radii a and b used in the subspace geometry constraints in order to couple to hyperspace.

[0071] **FIG. 15**. Perspective view of braided gold wire coaxial cable.

DETAILED DESCRIPTION OF THE INVENTION

[0072] 1. Referring to **FIG. 15**, the coaxial cable has a braided gold wire outer conductor (A) and a braided gold wire inner conductor (B) separated by a dielectric (C). The open braiding promotes the conduction of the electromagnetic wave while allowing the hyperspace mist to seep out of the braid and permeate the surrounding material in which it is embedded.

[0073] 2. The radius of the outer conductor a and the radius of the inner conductor b have the following values in order to couple the cable to the tetrahedral geometry of subspace.

[0074] a=8.342461828·microns

[0075] b=7.559058141·microns

[0076] 3. Even though the wire size is very small, the cable can be made in limited lengths using the new nanotechnology and silicon micromotors.

I claim:

1. A coaxial cable which has:

a) a braided gold wire outer conductor of radius 8.34 microns;

b) a braided gold wire inner conductor of radius 7.56 microns;

c) a thin dielectric separator between the two conductors; and

d) an open weave to allow the hyperspace mist to seep out of the cable and permeate the surrounding material in which the cable is embedded;

2. A specific relationship between the physical dimensions of the coaxial cable, given in items (**1***a*) and (**1***b*), to the following tetrahedral subspace couplings:

a) the ratio of the area-to-volume ratio of the tetrahedron to the area-to-volume ratio of its circumscribing sphere, equal to 3:1, with a coupling to the natural logarithm of the ratio of the radii of the conductors;

b) the ratio of the area of the three-square of the planar tetrahedron to the area of the one-square, equal to 3:1, with a coupling to the Planck mass and the linear inductance of the cable;

b) the distance between the base constant and $\sqrt{4/3}$ times the base constant, equal to −14.83817383 in natural logarithms, with a coupling to the outer radius of the conductor, the ten dimensions of space, and the co-dimensions of hyperspace; and

c) the wave function of the coaxial guide given in terms of item (**2***c*), the natural logarithm of the ratio of the radius of the outer conductor to the radius of the inner conductor, and the curvature 2π.

* * * * *

US 20060070371A1

(19) **United States**
(12) **Patent Application Publication**
St. Clair
(10) **Pub. No.: US 2006/0070371 A1**
(43) **Pub. Date: Apr. 6, 2006**

(54) ELECTRIC DIPOLE MOMENT PROPULSION SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/958,436**

(22) Filed: **Oct. 5, 2004**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)
(52) **U.S. Cl.** **60/203.1**

(57) ABSTRACT

A spacecraft propulsion system utilizing a rotating octagon of trapezoidal electrically charged flat panels to create an electric dipole moment that generates lift on the hull.

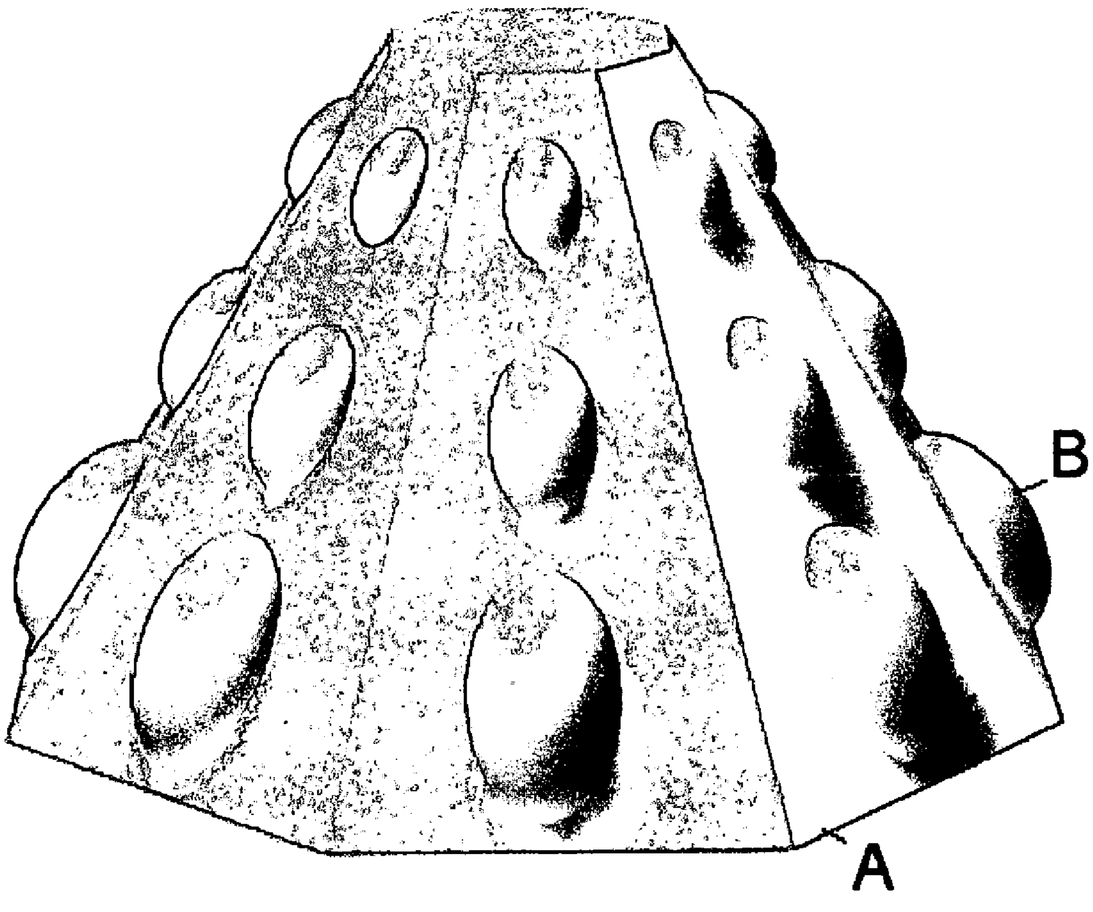

Figure 1

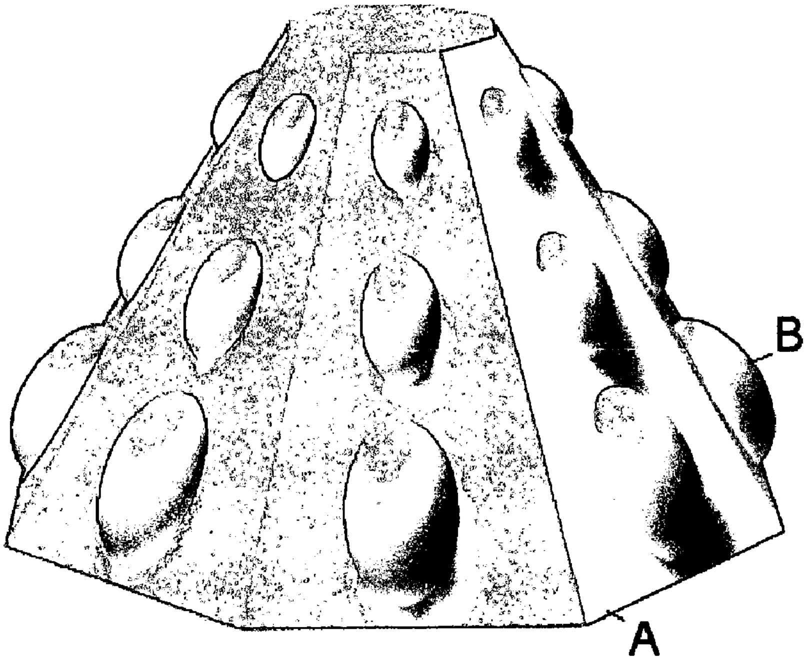

Figure 2

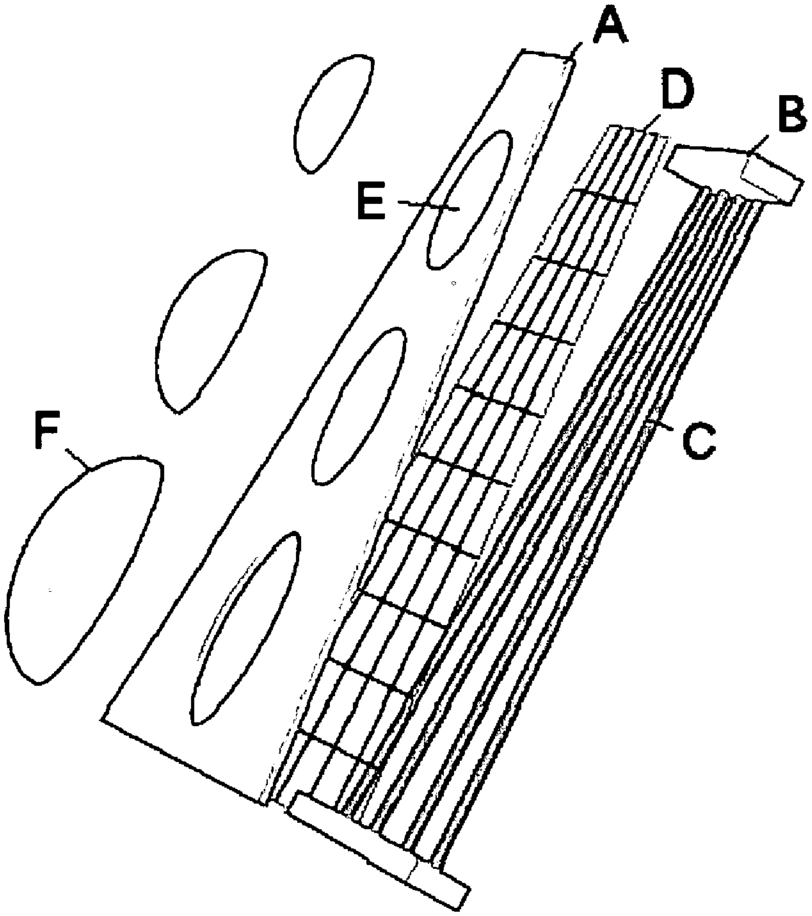

Figure 3

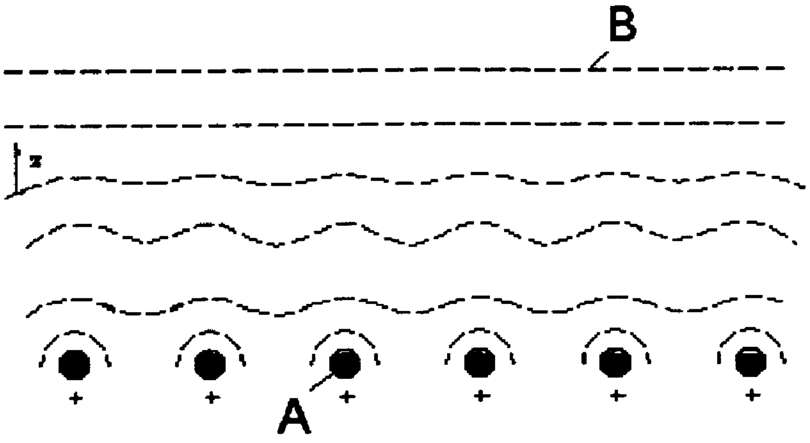

Figure 4

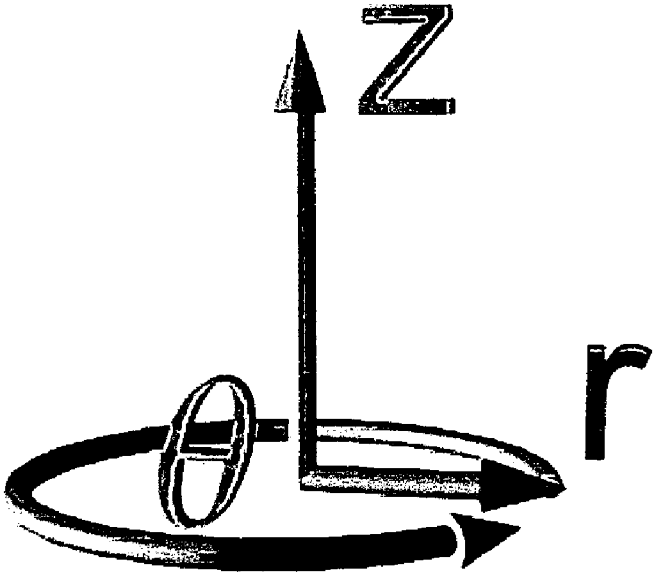

Figure 5

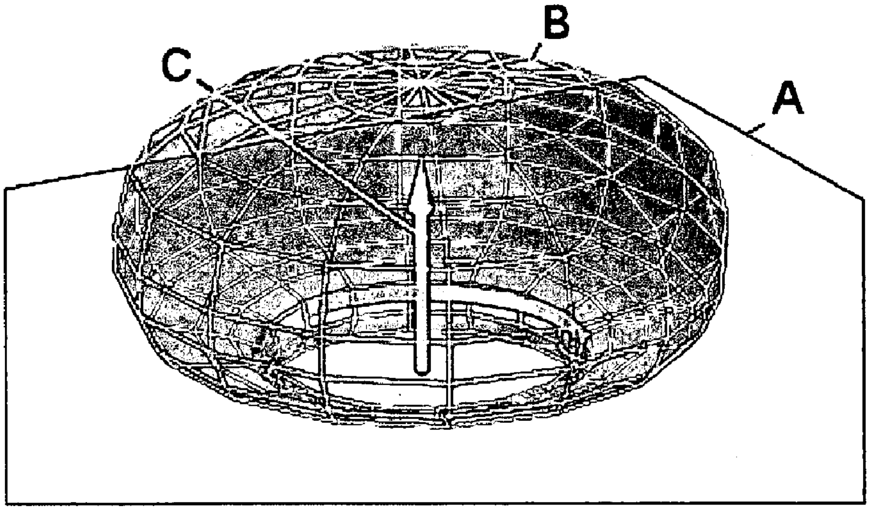

Figure 6

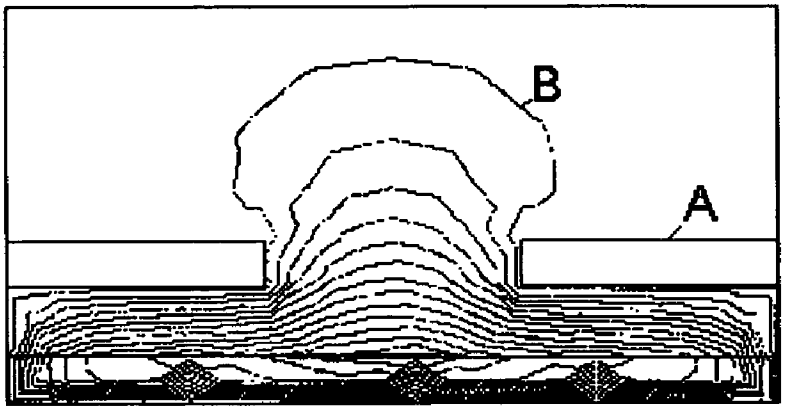

Figure 7

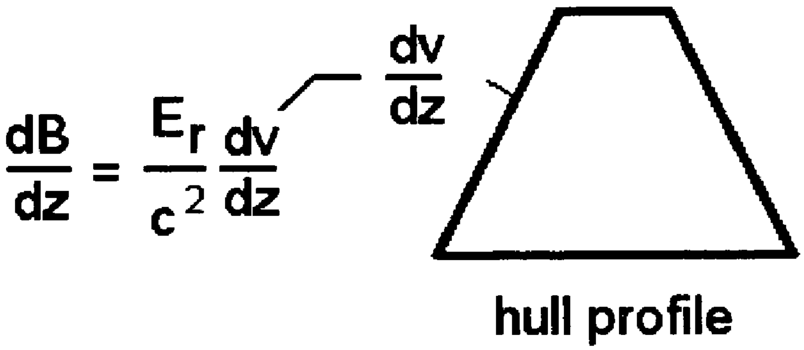
$\frac{dB}{dz} = \frac{E_r}{c^2}\frac{dv}{dz}$
$\frac{dv}{dz}$
hull profile

Figure 8

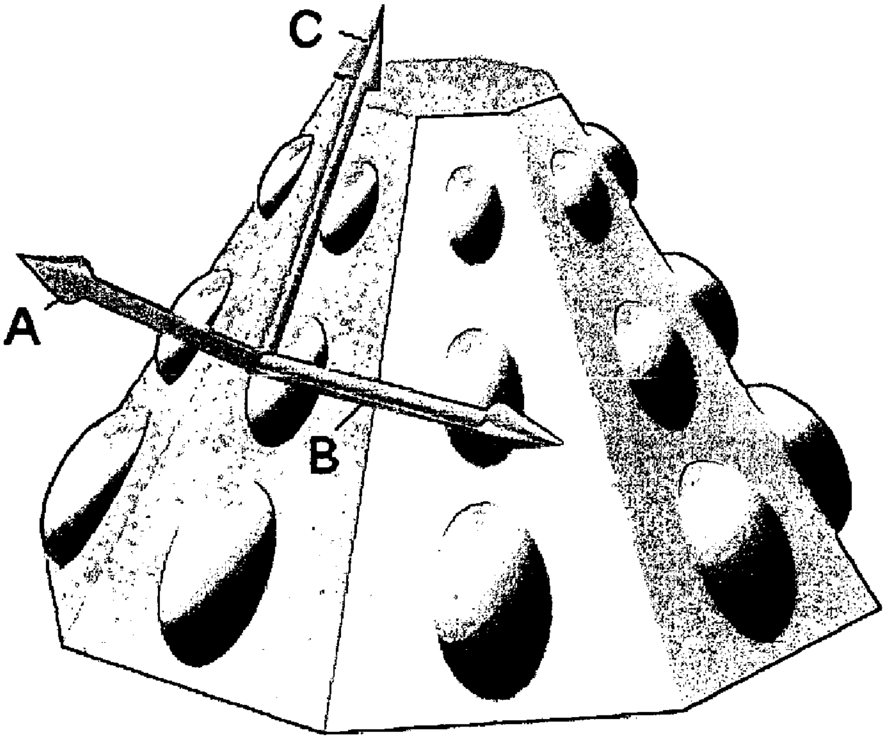

Figure 9

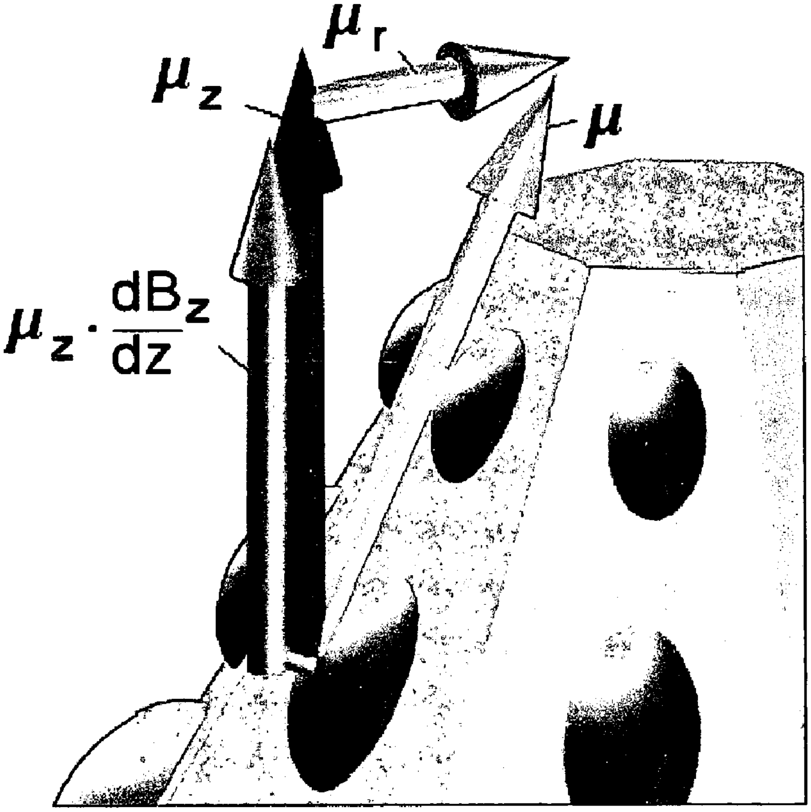
μ_r
μ_z
μ
$\mu_z \cdot \frac{dB_z}{dz}$

Figure 10

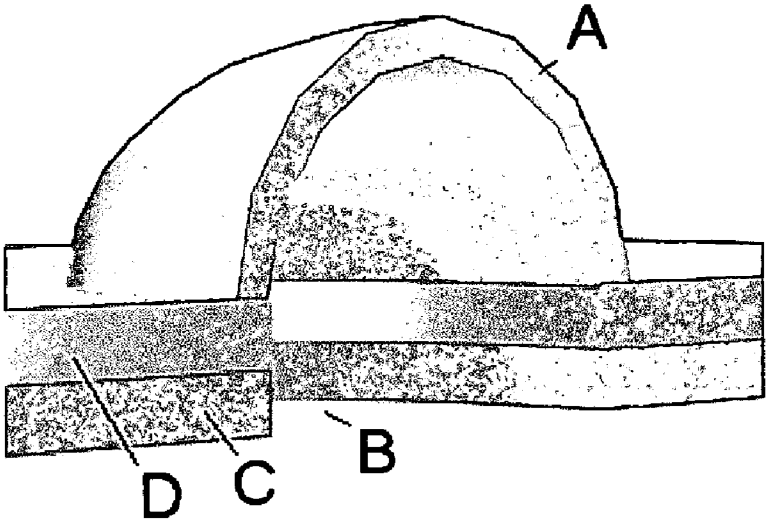

Figure 11

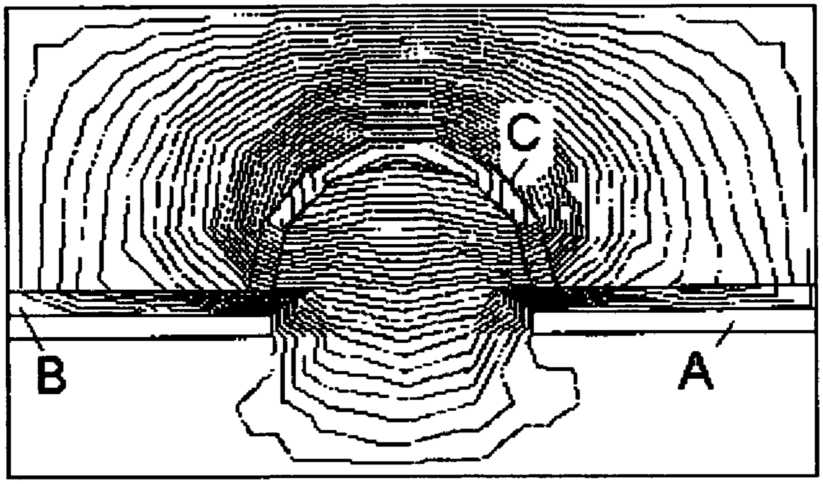

ELECTRIC DIPOLE MOMENT PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention is a spacecraft utilizing trapezoidal electrostatically charged flat plate panels which form a pyramidal hull. A panel contains three holes each of which produces a potential energy ellipsoidal bubble that creates an electric dipole moment. The rotation of the hull generates a magnetic moment and a magnetic field gradient in the vertical direction that produces a lift force on the spacecraft.

BACKGROUND OF THE INVENTION

[0002] It is known from electrodynamics that a hole in a conducting plane forms a potential energy bubble. This bubble creates an electric dipole moment from which it is possible to develop a magnetic moment. A rotating tilted hull produces a velocity gradient that generates a magnetic field gradient in the vertical direction. This combination produces a lift force on the spacecraft. A very large potential energy bubble is produced provided that the hole protrudes out of the plate in an ellipsoidal shape. Furthermore, a double cladding, in which each layer around the hole has a different permittivity, confines the field to the outside of the hull for even better results.

[0003] The planar potential energy is created by a grid of electrically charged wires or rods running the length of each panel. A circular potential energy from each rod very quickly sums to form a flat sheet of energy which emerges from the hole to form the potential energy bubble.

SUMMARY OF THE INVENTION

[0004] The invention relates to a spacecraft utilizing a rotating octagon of trapezoidal electrically charged flat plate panels to form a hull in the shape of a pyramid. Each panel has three protruding ellipsoidal bubbles that produce an electric dipole moment from a planar potential energy field created by a group of charged rods parallel to the panel. Because the panels are tilted and the hull is rotating, there is a tangential velocity gradient in the vertical direction. This creates the magnetic moment. Because the hull rotates, the radial electric field produces a magnetic field gradient in the vertical direction. This combination of magnetic moment and magnetic field gradient produces a lift force on the hull of the spacecraft.

[0005] On the underside of each panel is a group of high voltage electrically charged rods which run parallel to the panel. These wires or rods produce a planar electrical potential field underneath the holes in the panel. This potential energy field then bubbles out of the holes in the panel to create a large ellipsoidal potential energy field above the hull. The potential energy bubble carries an electric dipole moment which when rotated with the hull generates a magnetic moment in the vertical direction.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0006] **FIG. 1**. Perspective view of electric dipole moment spacecraft.

[0007] **FIG. 2**. Perspective exploded view of one panel with the ellipsoidal domes, flat hull panel with three holes, the charged rod grid and the planar potential energy field.

[0008] **FIG. 3**. Planar view of flat potential energy field produced by electrically charge wire rods.

[0009] **FIG. 4**. Perspective view of cylindrical coordinates $\{r, \theta, z\}$.

[0010] **FIG. 5**. Perspective view of ellipsoidal potential energy field emerging from hole in plate which produces an electric dipole moment.

[0011] **FIG. 6**. Planar view of field lines of potential energy bubble emerging from plate hole.

[0012] **FIG. 7**. Planar view of sloping hull profile needed to get a velocity gradient.

[0013] **FIG. 8**. Perspective view of hull showing electric dipole moment, the tangential velocity of the hull, and the magnetic moment.

[0014] **FIG. 9**. Perspective view of the magnetic moment components in the radial and vertical direction whereby the lift force is generated by the dot product of the vertical magnetic moment with the magnetic field gradient.

[0015] **FIG. 10**. Perspective view of cross section of dome showing two layer cladding with different permittivities to enhance potential energy field.

[0016] **FIG. 11**. Planar view of enhanced potential energy field with two layer cladding.

DETAILED DESCRIPTION OF THE INVENTION

[0017] 1. Referring to **FIG. 1**, the spacecraft is a rotating octagon of trapezoidal electrostatically charged flat panels which form a closed hull (A). Each panel has three ellipsoidal domes (B) of varying size centrally located along the major length of the panel. The purpose of the dome is to create a large ellipsoidal potential energy bubble over the hull which develops an electrical dipole moment. Because the hull is rotating, a magnetic moment is created in the vertical direction. A magnetic field gradient created by the rotating electric field on the hull in combination with the magnetic moment produces a lift force on the hull.

[0018] 2. Referring to **FIG. 2**, the trapezoidal hull panel (A) contains three ellipsoidal holes (E). A group of wires or rods (C) running parallel to and just underneath the panel are electrically charged to a high voltage at the end terminals (B). The rods produce a planar potential energy field (D) just under the holes in the panel. The field emerges from the holes in the shape of an ellipsoidal bubble and is amplified by an ellipsoidal dome (F) on the outside of the hull.

[0019] 3. Referring to **FIG. 3**, the group of parallel rods (A) are given a linear charge λ in units of charge per meter. The electric field E developed by the rod is the linear charge divided by the circumference of a circle of radius r around the wire times the permittivity ϵ of space. The analysis of this arrangement shows that within a few grid width spacings, the potential energy field ϕ due to the electric field E_0 has become planar (B) in the z-direction given by the equation

$$\phi = -E_0 z$$

[0020] 4. Referring to **FIG. 4**, the following analysis is done in cylindrical coordinates {r,θ,z}.

[0021] 5. Referring to **FIG. 5**, the ellipsoidal potential energy (B) emerges through the hole in the panel plate (A). In doing so it creates an electrical dipole moment (C) shown by the arrow normal to the hole area.

[0022] 6. Referring to **FIG. 6**, the bubble (B) emerges through plate (A).

[0023] 7. Because the bubble has the shape of an ellipse, the centroid y of the bubble would be four thirds the radius a divided by π as given by

$$\bar{y}_{ellipse} = \frac{4}{3}\frac{a}{\pi}$$

[0024] The electric dipole moment is then given as the charge q times the centroid y. The charge of the hole is equal to the permittivity E times the electric field E emerging from the hole times the area of the hole of radius a

$$q = \varepsilon_0 E\pi a^2 \frac{\text{coul}^2}{\text{m}^2\ \text{newton}}\frac{\text{newton}}{\text{coul}}\ \text{m}^2 = \text{coul} = \text{charge}$$

[0025] 8. The electric dipole moment p is the centroid y times the charge q

$$p = qy = \varepsilon_0 E\pi a^2 \frac{4}{3}\frac{a}{\pi} = \frac{4}{3}\varepsilon_0 E a^2\ \text{coul}\cdot\text{meter}$$

[0026] The electric dipole moment p times the hull velocity v is equal to a magnetic moment μ which is what creates the lift force on the hull

$$\mu = pv\ \text{amp}\cdot m^2$$

[0027] 9. The rotating hull creates the electric dipole moment velocity so that the entire hull develops a magnetic moment. In tensor notation, the magnetic moment μ is in the vertical z-direction because there is a radial component of the electric dipole moment times the velocity. The velocity is the radius r in the radial direction times the angular velocity ω in the z-direction

$$\mu^z = p_r x^r \omega^z$$

[0028] 10. The force F on the hull is the gradient of the dot product of the magnetic moment μ with the magnetic B field

$$F = \nabla(\mu \cdot B)$$

[0029] 11. By electrically charging the hull of the vehicle, a radial electric field is produced. By rotating the hull, the radial electric field changes with time. Thus Maxwell's equations will involve the curl of the magnetic field in the radial direction because the radial electric field is varying with time

$$(\nabla \times B)_r = \frac{1}{c^2}\frac{\partial E_r}{\partial t}$$

[0030] 12. The cross product involves the magnetic field in the theta direction which is zero

$$\frac{1}{r}\frac{\partial B_z}{\partial \theta} - \frac{\partial B_\theta}{\partial z} = \frac{1}{c^2}\frac{\partial E_r}{\partial t} = \frac{1}{r}\frac{\partial B_z}{\partial \theta}$$

[0031] Substituting the derivative of the electric field E

$$E_r = E_0 e^{i\omega t}$$

[0032] and integrating with respect to angle theta gives the vertical magnetic field B as the tangential velocity v times the radial electric field E divided by the speed of light c squared

$$B_z = \frac{v}{c^2} E_r$$

[0033] 13. The force on the hull is the gradient of the magnetic moment μ times the magnetic field B. In the equation for the magnetic field, the only available variable to work with in order to get a gradient of the magnetic field comes from the velocity.

$$\frac{dB_z}{dz} = \frac{E_r}{c^2}\frac{dv}{dz}$$

[0034] 14. Referring to **FIG. 7**, because the hull is in the shape of a pyramid, the velocity is a function of the height z of the hull. Using eight flat sides keeps the radial electric field pointing in the same direction in each panel. Each panel has three domes to produce the magnetic moment for a total of 24 magnetic moment generators.

[0035] 15. Referring to **FIG. 8**, the electric dipole moment (A) points in the radial direction, the rotating hull produces a tangential velocity (B), and the result is a magnetic moment (C) along the panel.

[0036] 16. Referring to **FIG. 9**, because the magnetic moment is parallel to the panel, there are vertical and radial components of the magnetic moment. The vertical magnetic moment creates the dot product with the magnetic field gradient, which is equal to the lift force.

[0037] 17. **FIG. 10** shows a cross-section of the dome (A) and the plate hole (B) with double cladding to enhance the field. The upper cladding (D) has a low relative permittivity in the range of 2 to 40, and the lower layer has a high relative permittivity in the range of 1200 to 4000.

[0038] 18. Comparing **FIG. 11** to **FIG. 7**, this dome and cladding configuration creates a much larger electric dipole moment compared to a hole in the plate. The wavy lines are the equi-potential energy lines from the dome (C) and the upper layer (B) and the lower level (A).

I claim:

1. A spacecraft propulsion system comprising:

a rotating octagon of trapezoidal electrostatically charged flat panels which form a closed sloping hull in the shape of a pyramid;

panels each having three holes covered by three ellipsoidal domes of varying size centrally located along the major length and axis of each panel; and

a grid of high voltage electrostatically charged rods located on the interior side of each panel such that a planar potential field is produced parallel to and under each panel hole.

2. The domes, holes and rotating charged hull of method 1 producing:

an ellipsoidal potential energy field emerging from the holes and generating an electric dipole moment on the outside of the hull;

a magnetic moment in the vertical direction due to the rotating electric dipole moment; and

a rotating electric field in the radial direction which generates a corresponding magnetic field gradient in the vertical direction proportional to the velocity gradient of the sloping panels of the hull.

3. A lift force on the spacecraft hull generated by:

the magnetic moment times the gradient of the magnetic field in the vertical direction; and

a dual surface layer hull cladding having different permittivities which enhance the electric dipole moment whereby the upper cladding has a low relative permittivity in the range of 2 to 40, and the lower layer has a high relative permittivity in the range of 1200 to 4000.

* * * * *

US 20060038081A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0038081 A1**

St. Clair (43) **Pub. Date: Feb. 23, 2006**

(54) **ELECTRIC DIPOLE SPACECRAFT**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/912,621**

(22) Filed: **Aug. 4, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) **ABSTRACT**

This invention is a rotating spacecraft that produces an electric dipole on four rotating spherical conducting domes perturbing a uniform spherical electric field to create a magnetic moment interacting with the gradient of a magnetic field that generates a lift force on the hull.

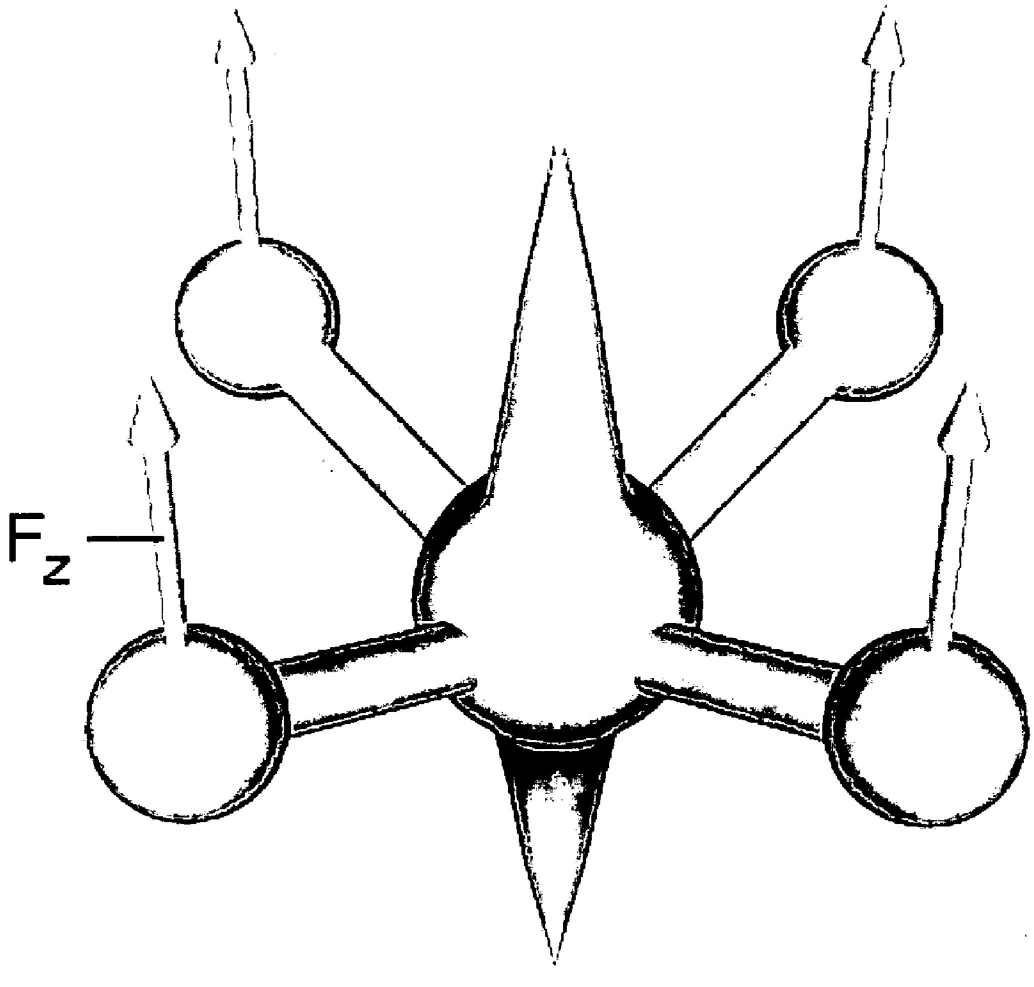

Figure 1

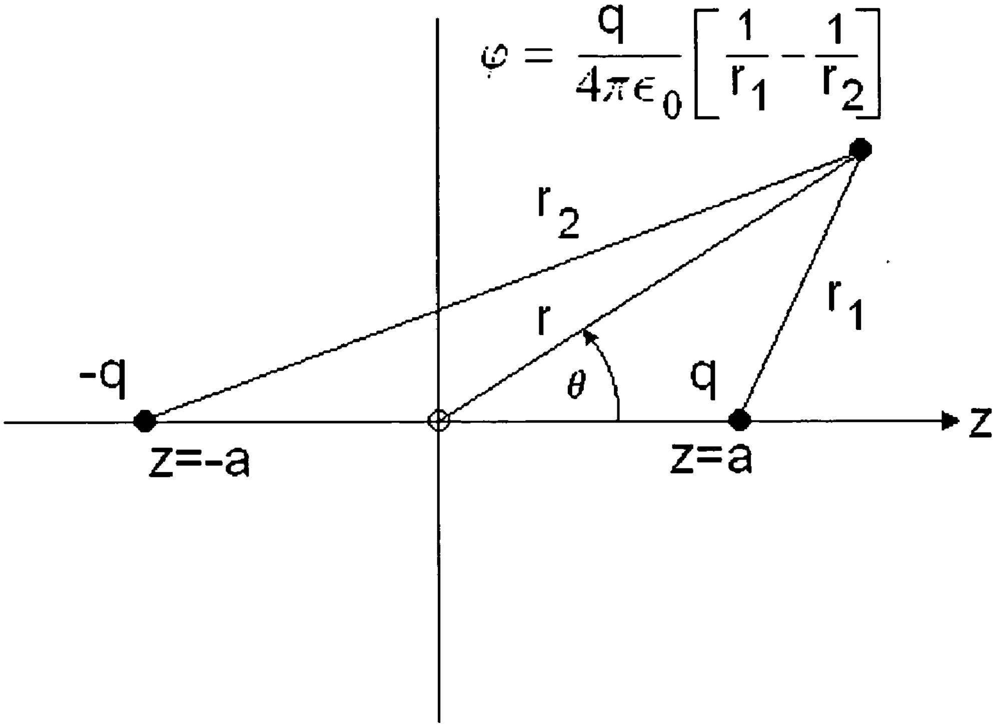
$\varphi = \frac{q}{4\pi\epsilon_0}\left[\frac{1}{r_1} - \frac{1}{r_2}\right]$
r_2
r
r_1
θ
-q
q
z=-a
z=a
z

Figure 2

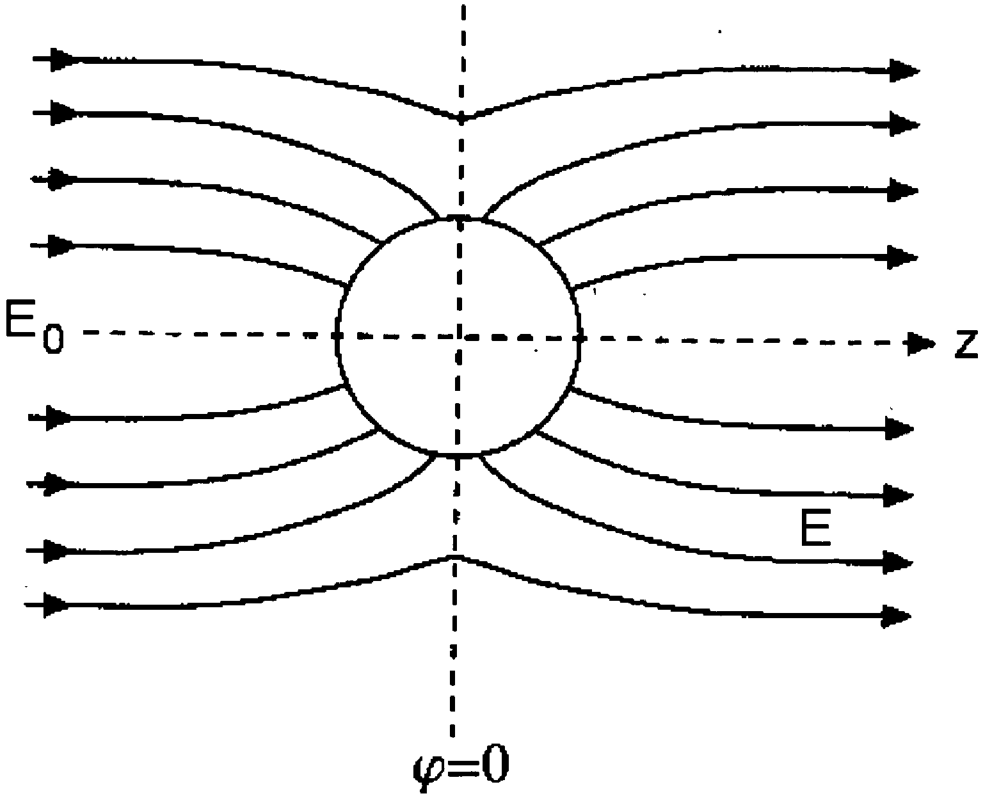
E$_0$
z
E
φ=0

Figure 3

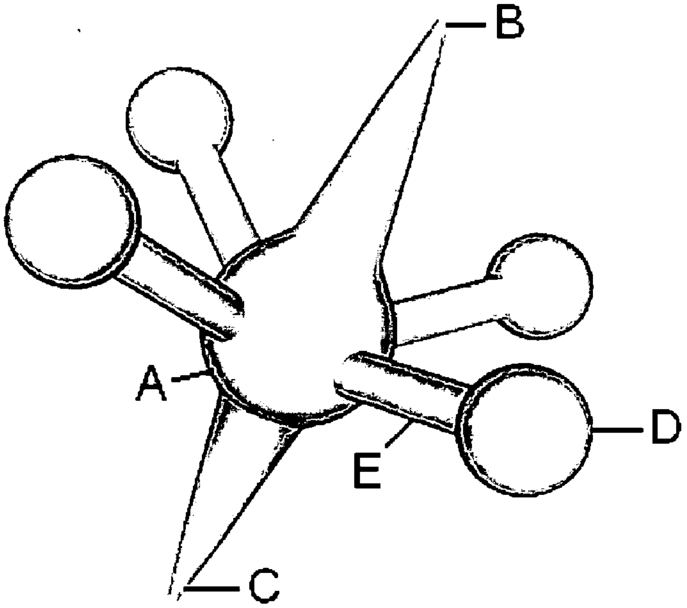

Figure 4

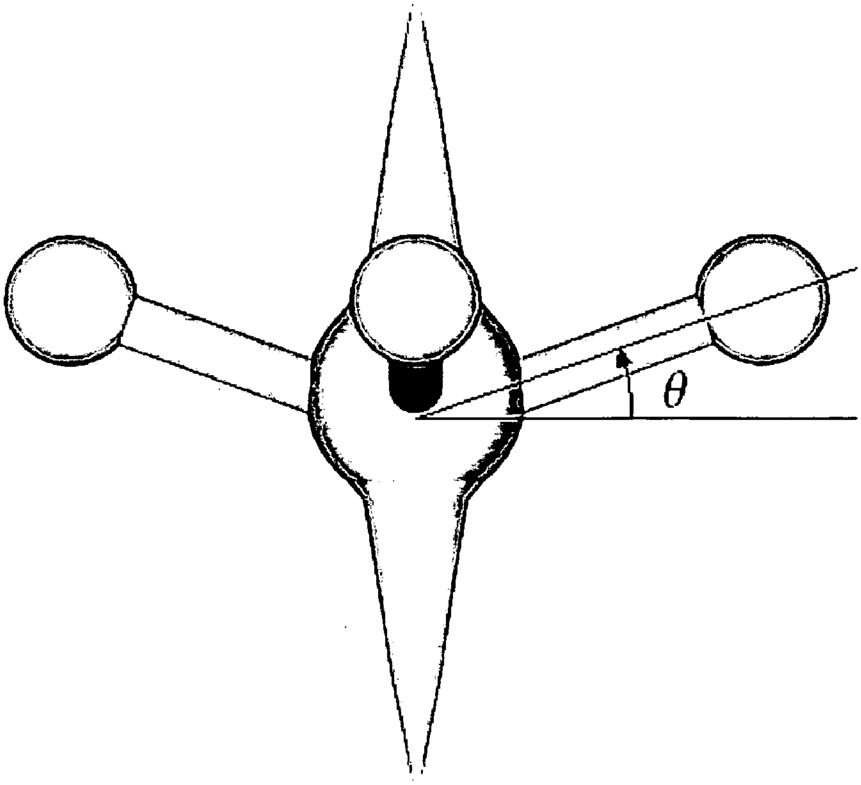

Figure 5

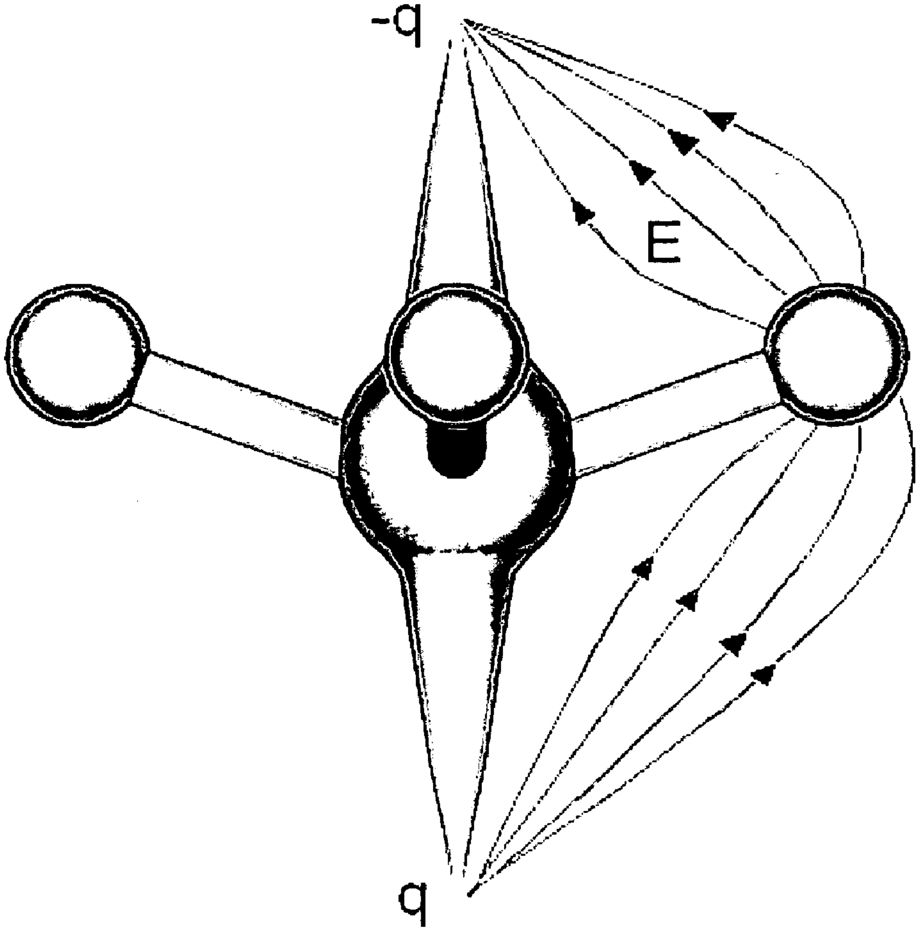

Figure 6

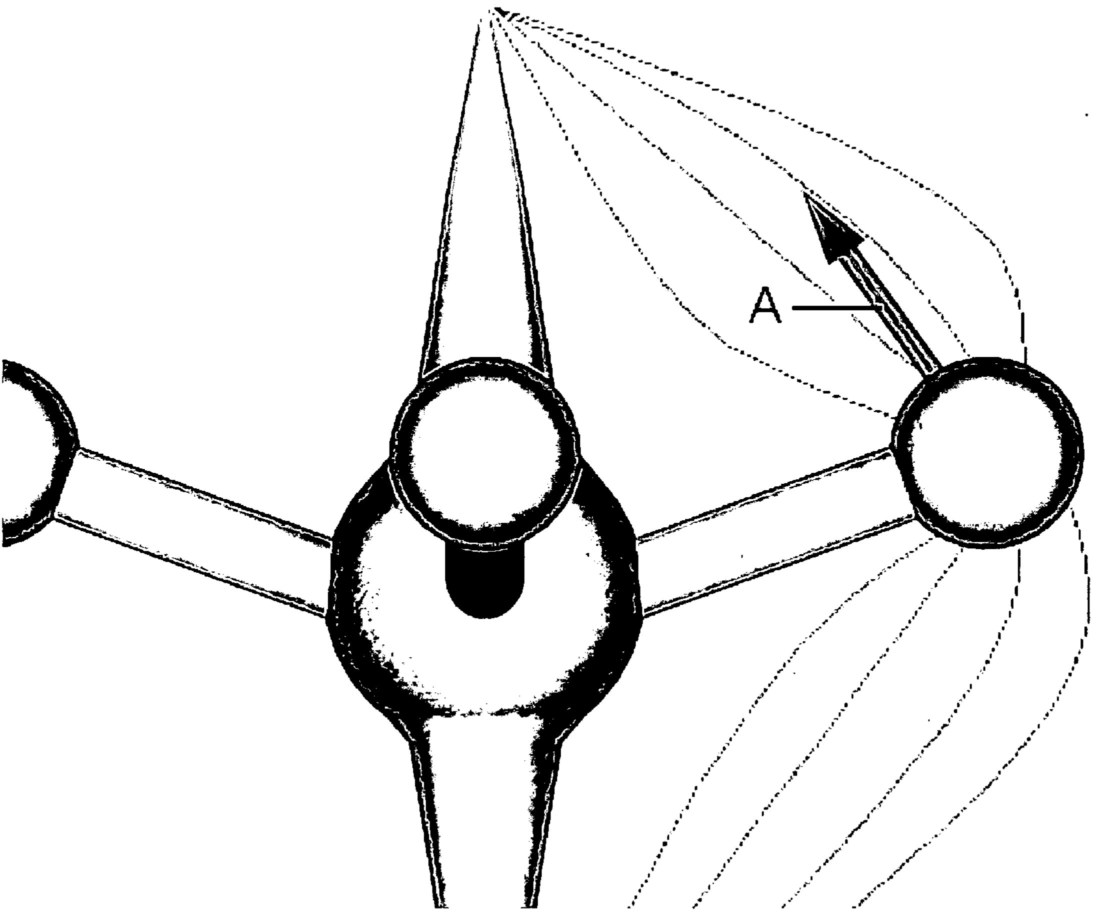

Figure 7

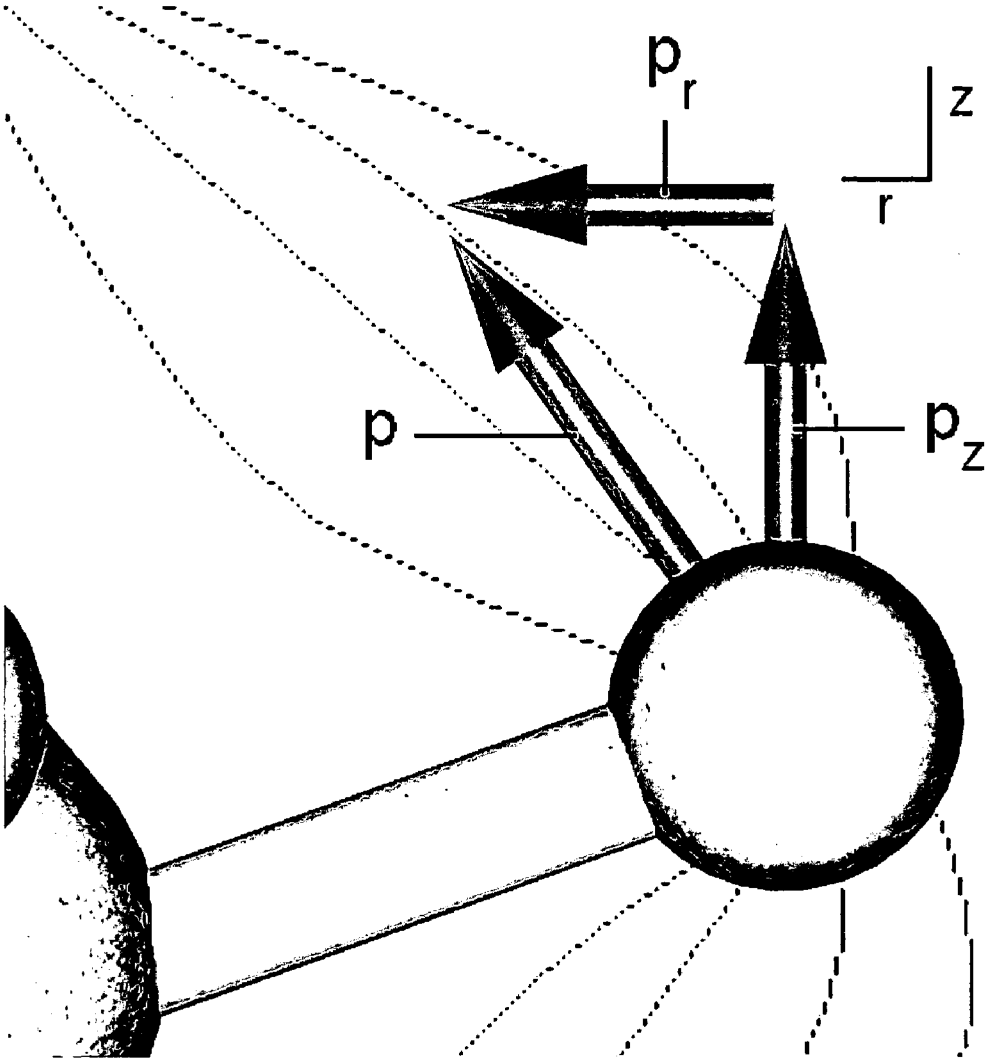

Figure 8

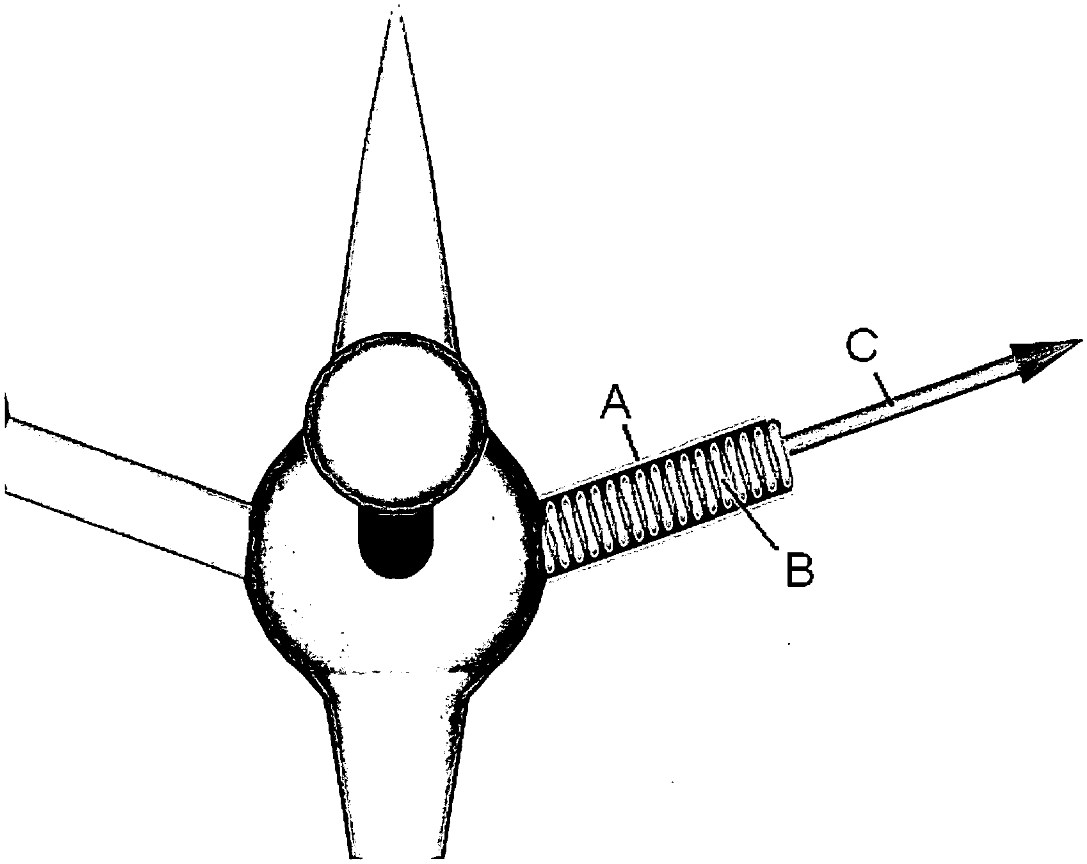

Figure 9

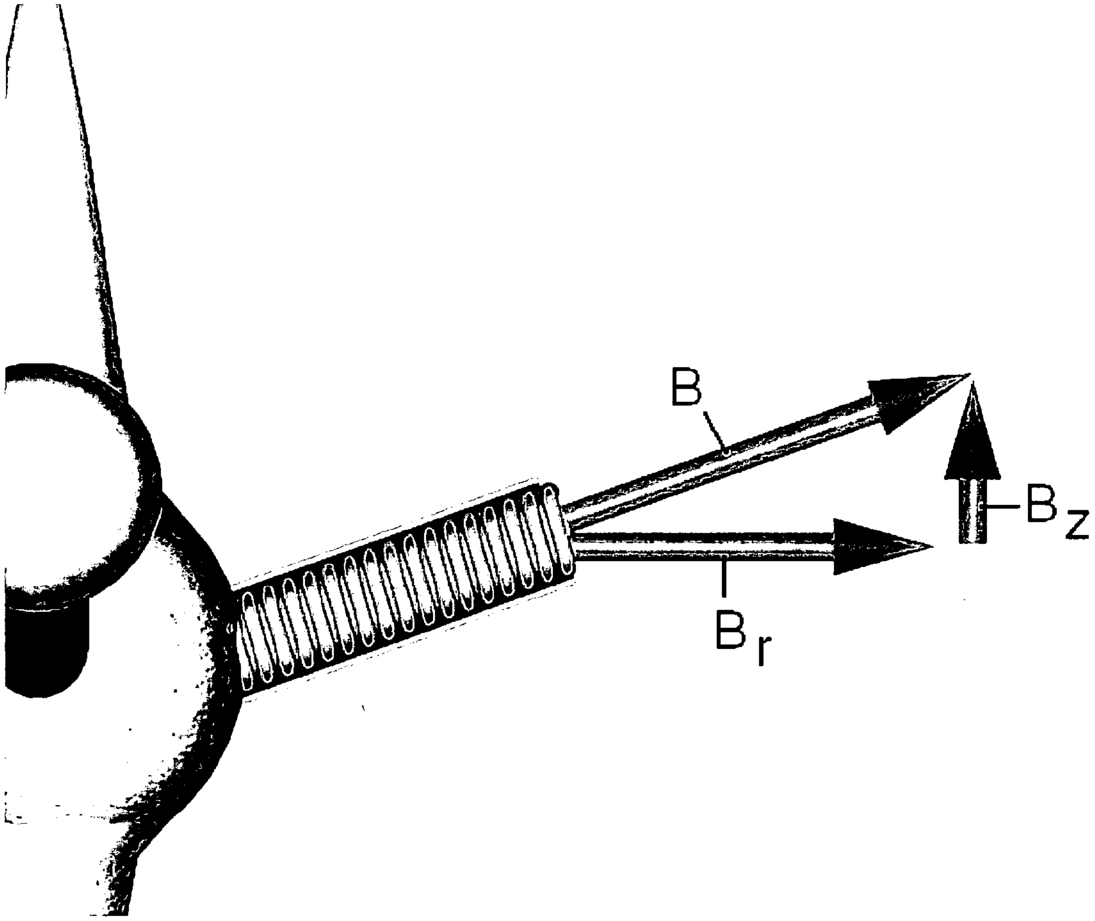

Figure 10

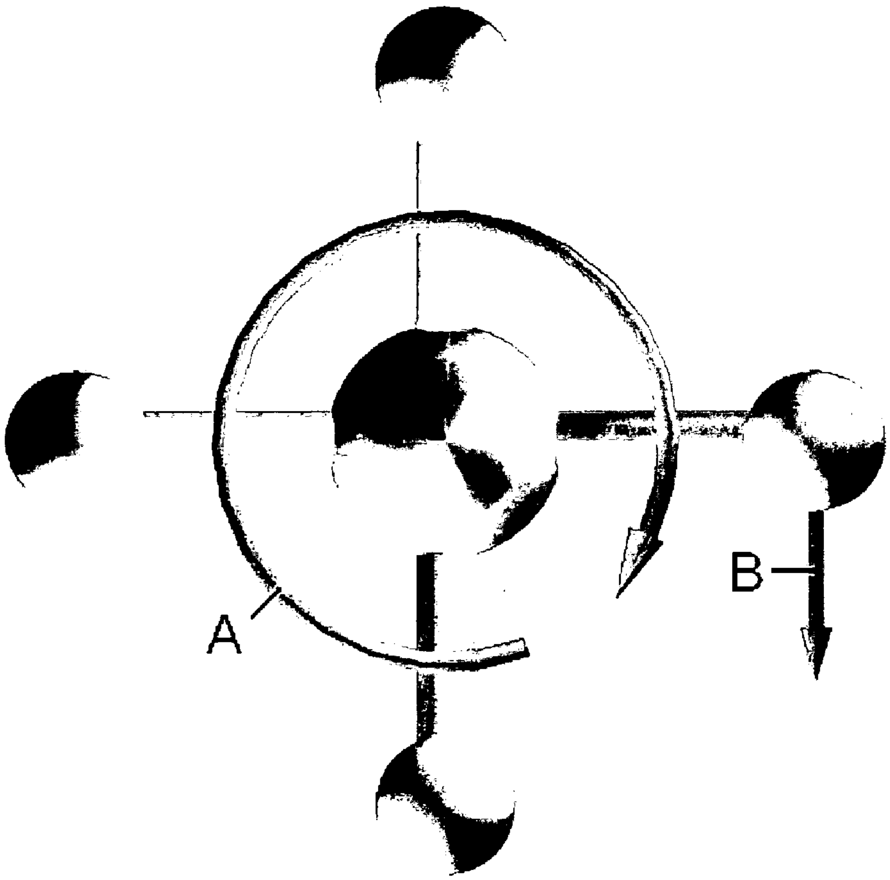

Figure 11

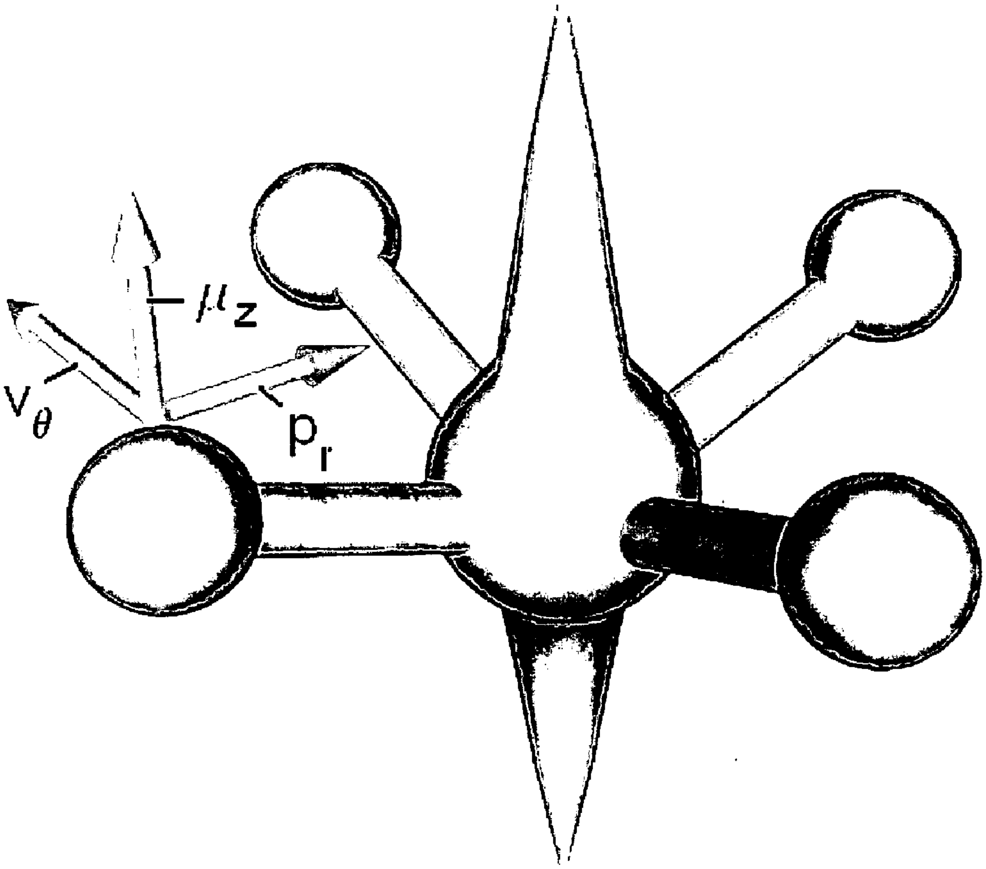
μ_z
v_θ
p_r

Figure 12

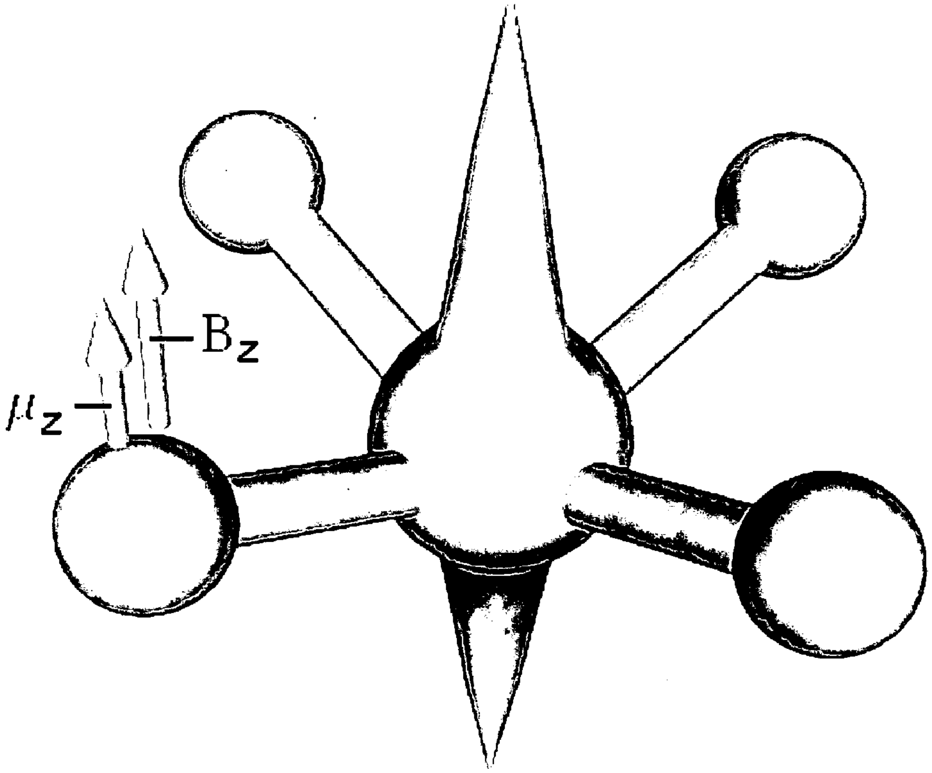

Figure 13

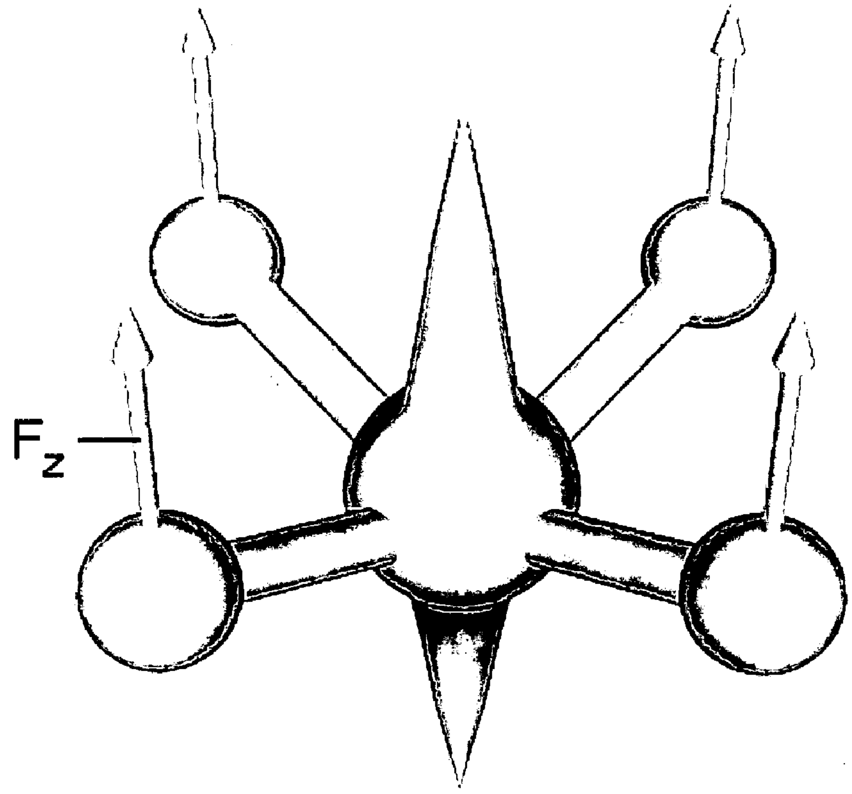

Figure 14

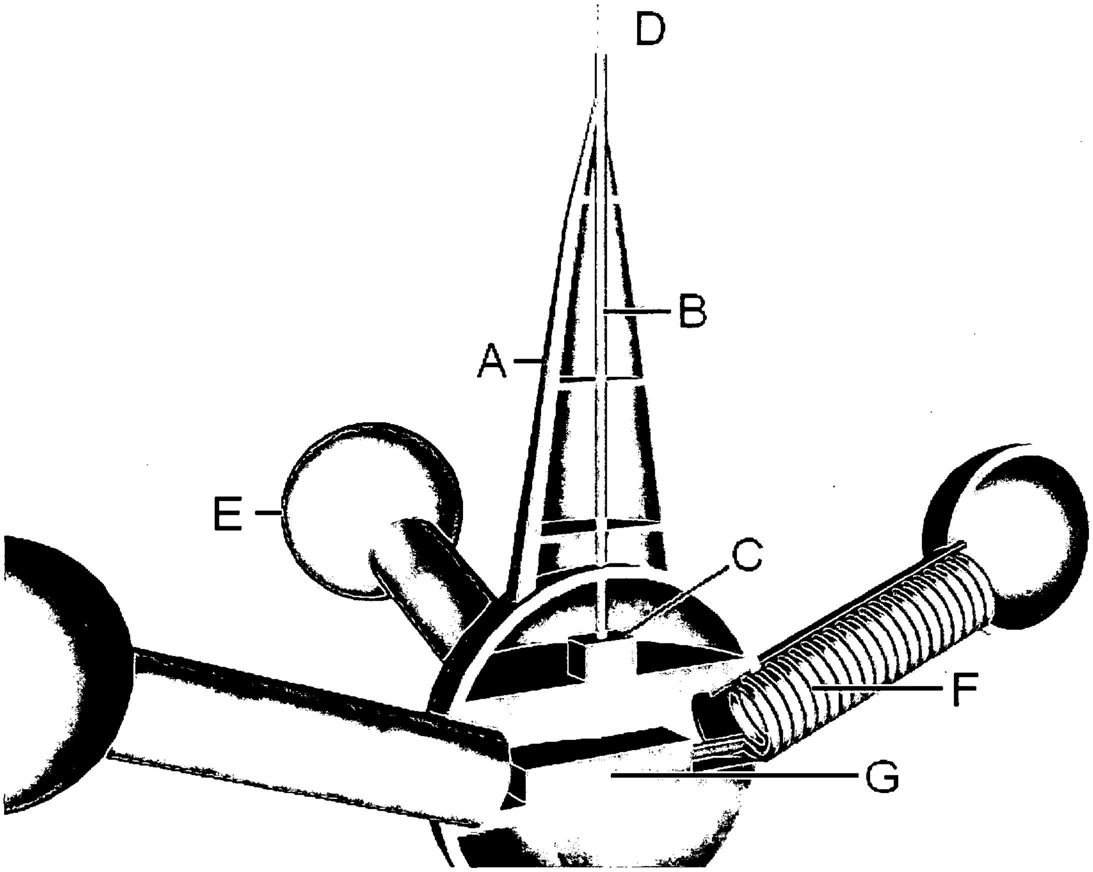

ELECTRIC DIPOLE SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a rotating spacecraft that utilizes four spherical conducting domes perturbing a uniform electric field in order to create a lift force by means of a magnetic moment times the gradient of a magnetic field.

BACKGROUND OF THE INVENTION

[0002] An electric dipole p is two electrical charges of opposite sign {q, −q} separated by a distance a.

$$p=qa=\text{coulomb}\cdot\text{meter}$$

[0003] If this dipole is moving with a velocity v, it produces a magnetic moment μ.

$$\mu = pv = qav = \text{coulomb meter}\frac{\text{meter}}{\text{sec}} = \frac{\text{coulomb}}{\text{sec}}\text{meter}^2 = IArea$$

[0004] The magnetic moment is equal to a current I circulating around an area. The magnetic field B has units of kilogram per second per charge coulomb.

$$B = \frac{\text{kg}}{\text{sec } coul}$$

[0005] The gradient of the magnetic field in the vertical direction z has units of

$$\frac{\text{dB}}{dz} = \frac{\text{kg}}{\text{sec } coul \text{ meter}}$$

[0006] This gradient interacting with a magnetic moment creates a force F measured in newtons.

$$F_z = \mu\frac{\text{dB}}{dz} = \frac{coul\ \text{m}^2}{\text{sec}}\frac{\text{kg}}{\text{sec } coul\ m} = \frac{\text{kg } m}{\text{sec}^2} = \text{newton}$$

[0007] In terms of vectors, the force is equal to

$$F=\nabla(\mu\cdot B)$$

which is the gradient ∇ of the dot product (·) of the magnetic moment with the magnetic field. This means that the magnetic moment has to be aligned with the field. The lift force on the spacecraft would then be the magnetic moment in the vertical z-direction μ_z times the magnetic field in the z-direction B_z. For constant magnetic moment, the gradient affects the magnetic field only, resulting in the same force equation

$$F_z=\nabla(\mu_z\cdot B_z)=\mu_z\nabla B_z$$

[0008] Referring to **FIG. 1**, the electric dipole has a positive charge q located on the z-axis at a distance a from the origin of the graph. A second negative charge −q is located at a distance −a from the origin. The positive charge produces an electrostatic potential ϕ_1 at a radius r_1 equal to the charge q divided by 4π times the permittivity of space ϵ_0

$$\varphi_1 = \frac{q}{4\pi\varepsilon_0}\frac{1}{r_1}$$

where the permittivity is linear capacitance, measured in farads per meter. The electrostatic potential has units of volts

$$\varphi = \frac{coul}{\left(\frac{\text{farad}}{\text{meter}}\right)}\frac{1}{\text{meter}} = \frac{coul}{\text{farad}} = \text{volt}$$

because the charge in coulombs held by a capacitor is equal to the capacitance, measured in farads, times the capacitor voltage. Because the second charge has the opposite sign, the potential ϕ_2 at a radius r_2 to the same point in space is

$$\varphi_2 = \frac{-q}{4\pi\varepsilon_0}\frac{1}{r_2}$$

[0009] The total potential ϕ at some point in space is equal to the sum of the two potentials, or

$$\varphi = \varphi_1 + \varphi_2 = \frac{q}{4\pi\varepsilon_0 r_1} - \frac{q}{4\pi\varepsilon_0 r_2} = \frac{q}{4\pi\varepsilon_0}\left[\frac{1}{r_1} - \frac{1}{r_2}\right]$$

[0010] As seen in the diagram, the point of space is a distance r from the origin. Using the law of cosines, radius r_1 can be written as

$$r_1 = (r^2 + a^2 - 2ar\cos(\theta))^{\frac{1}{2}} = r\left(1 + \left(\frac{a}{r}\right)^2 - 2\left(\frac{a}{r}\right)\cos(\theta)\right)^{\frac{1}{2}} = r(1 - 2xt + t^2)^{\frac{1}{2}}$$

where t is the ratio of the charge location over the radius, and x is cos(θ). The potential for positive charge q_1 can be written

$$\varphi_1 = \frac{q}{4\pi\varepsilon_0}\frac{1}{r_1} = \frac{q}{4\pi\varepsilon_0}\frac{1}{r}(1 - 2xt + t^2)^{-1/2}$$

[0011] Dropping the factor $q/4\pi\epsilon_0 r$, the square root can be expressed in terms of the Legendre polynomial P_n cos(θ) of the nth power

$$g(t, x) = (1 - 2xt + t^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x)t^n$$

where the absolute value of t is less than one. The polynomial coefficients of t^n can be obtained by using the binomial theorem to expand the generating function g(t,x) as

$$(1-2xt+t^2)^{-1/2}=\sum_{n=0}^{\infty}\frac{(2n)!}{2^{2n}(n!)^2}(2xt-t^2)^n$$

which evaluates to

$$\frac{0!}{2^0(0!)^2}(2xt-t^2)^0+\frac{2!}{2^2(1!)^2}(2xt-t^2)^1+\frac{4!}{2^4(2!)^2}(2xt-t^2)^2$$

$$1t^0+xt^1+\left(\frac{3}{2}x^2-\frac{1}{2}\right)t^2+\text{order}(t^3)$$

[0012] The first three Legendre polynomials are therefore

$$P_0=1$$

$$P_1=x$$

$$P_2=\frac{1}{2}(3x^2-1)$$

[0013] The electrostatic potential for both charges of the electric dipole is

$$\varphi=\frac{q}{4\pi\varepsilon_0}\frac{1}{r}\left\{\left(1-2\left(\frac{a}{r}\right)\cos(\theta)+\left(\frac{a}{r}\right)^2\right)^{-1/2}-\left(1+2\left(\frac{a}{r}\right)\cos(\theta)+\left(\frac{a}{r}\right)^2\right)^{-1/2}\right\}$$

[0014] The potential can be evaluated in terms of the Legendre polynomials as

$$\begin{aligned}\varphi&=\frac{q}{4\pi\varepsilon_0 r}\left[\sum_{n=0}^{\infty}P_n(\cos(\theta))\left(\frac{a}{r}\right)^n-\sum_{n=0}^{\infty}P_n(\cos(\theta))(-1)^n\left(\frac{a}{r}\right)^n\right]\\&=\frac{2q}{4\pi\varepsilon_0 r}\left[P_1(\cos(\theta))\left(\frac{a}{r}\right)+P_3(\cos(\theta))\left(\frac{a}{r}\right)^3+\dots\right]\end{aligned}$$

[0015] The first and most dominant term when the radius is much greater than location a is equal to

$$\varphi=\frac{2aq}{4\pi\varepsilon_0}\frac{P_1(\cos(\theta)0}{r^2}$$

which is the electric dipole potential and 2aq is the dipole moment

$p=2aq$

[0016] Now imagine a constant electric field E_0 which is perturbed by a conducting sphere of radius a. The unperturbed electrostatic potential outside the sphere would the negative of the electric field times the radius times the Legendre polynomial, or

$\Phi_1=-E_0 r P_1$

[0017] The electrostatic potential perturbed by the charges is the voltage E_0a times the radius a times the a of the dipole moment times the Legendre polynomial divided by the radius squared

$$\varphi_2=E_0 aaa\frac{P_1}{r^2}=E_0 a^3\frac{P_1}{r^2}$$

[0018] The total potential outside the sphere is the sum of the two potentials equal to

$$\varphi=-E_0 rP_1+E_0 a^3\frac{P_1}{r^2}=-E_0 P_1\left(r-\frac{a^3}{r^2}\right)=-E_0 rP_1\left(1-\left(\frac{a}{r}\right)^3\right)$$

[0019] Referring to **FIG. 2**, the previously uniform electric field is shown perturbed by the neutral conducting sphere. The center of the sphere is taken as the origin and the z-axis is oriented parallel to the original uniform field.

[0020] The electric field induces a surface charge density σ on the sphere equal to the negative of the permittivity of space times the gradient of the electrostatic potential

$$\sigma=-\varepsilon_0\frac{\partial\varphi}{\partial r}|_{r=a}=3\varepsilon_0 E_0\ \cos(\theta)$$

[0021] The electric field also induces an electric dipole moment on the sphere equal to the

$$p=\frac{qa}{r}\cdot\frac{2a^2}{r}=4\pi\varepsilon_0 E_0 a^3$$

with units of coulomb-meter. If this sphere is rotating around a central axis at a velocity v, it will create a magnetic moment μ equal to the dipole moment times the velocity.

$\mu=pv$

with units of ampere-meter^2.

SUMMARY OF THE INVENTION

[0022] As shown in the preceding background section, a neutral conducting sphere placed in a uniform electric field will generate a magnetic moment when rotated around a central axis. The electric field can be created by two points charges of opposite sign separated by a distance between them.

[0023] Referring to **FIG. 3**, the spacecraft has a spherical cabin (A) to which are attached cone-shaped electrostatic towers (B,C) above and below the cabin along the direction of travel in the z-direction. Because the electric field goes from the positive charge to the negative charge, the tip of the lower tower has a positively charged electrode, and the upper tower has a negatively charged electrode. Four equally-spaced neutral conducting spheres (D) are connected to the cabin by non-conducting tubes (E). The tubes make an angle with the cabin such that the distance (CD) is greater than distance (DB). The angle θ of the tube with respect to the cabin can be seen in side view **FIG. 4**.

[**0024**] Referring to **FIG. 5**, the charges create a uniform spherical field between the towers. The conducting spheres perturb this field such that the electric field (E) points toward the upper tower in a manner similar to that shown previously in **FIG. 2**.

[**0025**] Referring to **FIG. 6**, the conducting sphere produces an electric dipole moment (A) pointing at an angle toward the upper tower.

[**0026**] Referring to **FIG. 7**, by the law of addition of vectors, the electric dipole $\bar{p}$ can be represented by two orthogonal vectors pointing in the vertical z-direction p_z and in the inward radial direction p_r.

[**0027**] Referring to **FIG. 8**, the hollow tube (A) connecting the cabin with the conducting sphere contains a spiral-wound electrical solenoid (B) which produces a magnetic field (C). This magnetic field B can be decomposed into two orthogonal vectors pointing in the vertical z-direction B_z and in the outward radial direction B_r as shown in **FIG. 9**.

[**0028**] Referring to top-view **FIG. 10**, the spacecraft has a clockwise angular velocity ω (A) which gives the conducting sphere a velocity v as shown by the vector (B). By the right-hand rule of physics, the angular velocity vector points in the negative z-direction. The angular velocity in the z-direction crossed with the radius r in the radial direction produces a velocity v in the clockwise θ-direction using cylindrical coordinates {r,θ,z}.

$$v_\theta=w_z \times r_r=-\omega r$$

[**0029**] Referring to **FIG. 11**, the negative radial dipole moment p_r crossed with the negative velocity v_θ of the sphere produces a positive magnetic moment μ_z in the z-direction.

$$\mu_z=p_r \times v_\theta=(-p_r)(-v_\theta)=pv$$

[**0030**] Referring to **FIG. 12**, the magnetic field B_z in the vertical z-direction is dotted with the magnetic moment μ_z in the z-direction to produce a force F_z in the vertical z-direction on each conducting sphere (**FIG. 13**).

$$F_z=\nabla(\mu_z \cdot B_z)=\mu_z \nabla B_z$$

[**0031**] The magnetic field that is produced by the solenoid actually curves away and around. Thus there is a gradient of the field in the z-direction.

[**0032**] The force can also be expressed in tensor notation. The magnetic B field in the vertical direction is part of an electromagnetic 4×4 matrix Faraday tensor F

$$F^\alpha_\beta = \begin{array}{c|cccc|} & \beta=t & r & \theta & z \\ \alpha=t & 0 & 0 & 0 & 0 \\ \alpha=r & 0 & 0 & B_z & 0 \\ \alpha=\theta & 0 & -B_z & 0 & 0 \\ \alpha=z & 0 & 0 & 0 & 0 \end{array}$$

which shows that the magnetic field is located in slot $F^r{}_\theta$ of the Faraday tensor. In tensor notation the subscripts and superscripts have to match up on both sides of the equation. Matching subscripts and superscripts on the same side of the equation cancel. In this case, the electric dipole moment is in the radial direction p_r. The velocity can be represented as a time derivative of the θ-coordinate x^θ or

$$v^\theta = \frac{\partial x\theta}{\partial t}$$

[**0033**] Thus the force component in the z-direction becomes

$$F_z = p_r v^\theta F^r_{\theta,z} = (-p)(-v)\frac{\partial B_z}{\partial z} = pv\frac{\partial B_z}{\partial z}$$

where the angular and radial tensor components cancel and comma-z (, z) represents differentiation of the magnetic field in the z-direction.

[**0034**] The spacecraft design also has an inherent motion control system for moving in various directions. If the magnetic field of one solenoid arm is increased or decreased, the force on that sphere will be increased or decreased. Thus the spacecraft can turn in a particular direction.

A BRIEF DESCRIPTION OF THE DRAWINGS

[**0035**] **FIG. 1**. Electric dipole.

[**0036**] **FIG. 2**. Uniform electric field perturbed by electric dipole.

[**0037**] **FIG. 3**. Perspective view of spacecraft.

[**0038**] **FIG. 4**. Angle of solenoid tube.

[**0039**] **FIG. 5**. Electric field perturbed by conducting sphere.

[**0040**] **FIG. 6**. Electric dipole generated by conducting sphere.

[**0041**] **FIG. 7**. Orthogonal vector components of electric dipole.

[**0042**] **FIG. 8**. Magnetic field produced by tube arm solenoid.

[**0043**] **FIG. 9**. Orthogonal vector components of magnetic field.

[**0044**] **FIG. 10**. Angular velocity of hull.

[**0045**] **FIG. 11**. Magnetic moment produced by radial electric dipole and sphere velocity.

[**0046**] **FIG. 12**. Dot product of the magnetic moment with the magnetic field.

[**0047**] **FIG. 13**. Vertical lift force on all four conducting spheres.

[**0048**] **FIG. 14**. Perpective view of spacecraft interior.

DETAILED DESCRIPTION OF THE INVENTION

[**0049**] 1. Referring to the cut-away view **FIG. 14**, the construction of the spacecraft is a thin-wall insulating thermoplastic having a dielectric constant in the range of 20 kilovolts per millimeter (A). An insulated electrode (B) runs from the cabin power supply and high-voltage transformer

(C) to the tip of each tower (D). The four spheres (E) are silver plated to make them conducting. The tube solenoids (F) are driven by a direct current power supply (G).

[0050] 2. The present model uses 3D computer design software and stereolithography fabrication techniques to create the thin-wall, low-weight, hollow structure of the hull. The computer model is sliced into many thin horizontal slices. A laser, mounted on an x-y table, draws out the slice on a table immersed in a bath of liquid polymer. Due to its sensitivity to the light, the liquid polymerizes. The table is then lowered a few thousandths of an inch more and the process is repeated. Thus making hollow spherical and conical shapes is extremely easy to do. Parts can be designed and stored in *.STL stereolithography files for transmission by Internet e-mail to the service bureau machine shop which sends the finished parts back the next day by express mail.

I claim:

1. A spacecraft comprising:

a. a spherical cabin;

b. an electrostatic conical tower mounted on top of item (**1***a*), supporting a vertically-mounted negatively-charged insulated electrode at the tip of the tower;

c. an electrostatic conical tower mounted on the bottom of item (**1***a*), supporting a vertically-mounted positively-charged insulated electrode at the tip of the tower;

d. a vertical electric dipole created by items (**1***b*) and (**1***c*);

e. a high-voltage transformer to drive item (**1***d*), mounted in item (**1***a*);

f. four tubular arms, mounted at 90° around and extending at an angle from item (**1***a*);

g. four solenoids, each of which is mounted axially inside item (**1***f*);

h. a direct current power supply to drive item (**1***g*);

i. four silver-plated conducting spheres, each of which is mounted on the end of item (**1***f*);

2. an electrostatic lift system that:

a. produces a uniform spherical electric field by means of item (**1***d*) which envelopes item (**1***i*);

b. produces a perturbed electric field due to the presence of item (**1***i*);

c. produces an electric dipole moment in the direction of item (**1***b*) due to items (**2***a*) and (**2***b*);

d. produces a vertical magnetic moment due to the clockwise angular velocity of item (**1***a*) combined with item (**2***c*);

e. produces a vertical lift force on item (**1***i*) due to item (**2***d*) combined with the magnetic field gradient in the vertical direction produced by item (**1***g*); and

f. creates a motion control system by varying the current to item (**1***g*) in order to increase or decrease the effect of item (**2***e*) on a particular item (**1***i*).

* * * * *

US 20060144035A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0144035 A1**
St. Clair (43) **Pub. Date: Jul. 6, 2006**

(54) PHOTON SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/027,969**

(22) Filed: **Jan. 3, 2005**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)
(52) **U.S. Cl.** **60/203.1**; 60/204

(57) ABSTRACT

A spacecraft propulsion system utilizing photon particles to create negative energy over the hull in order to generate a lift force on the hull.

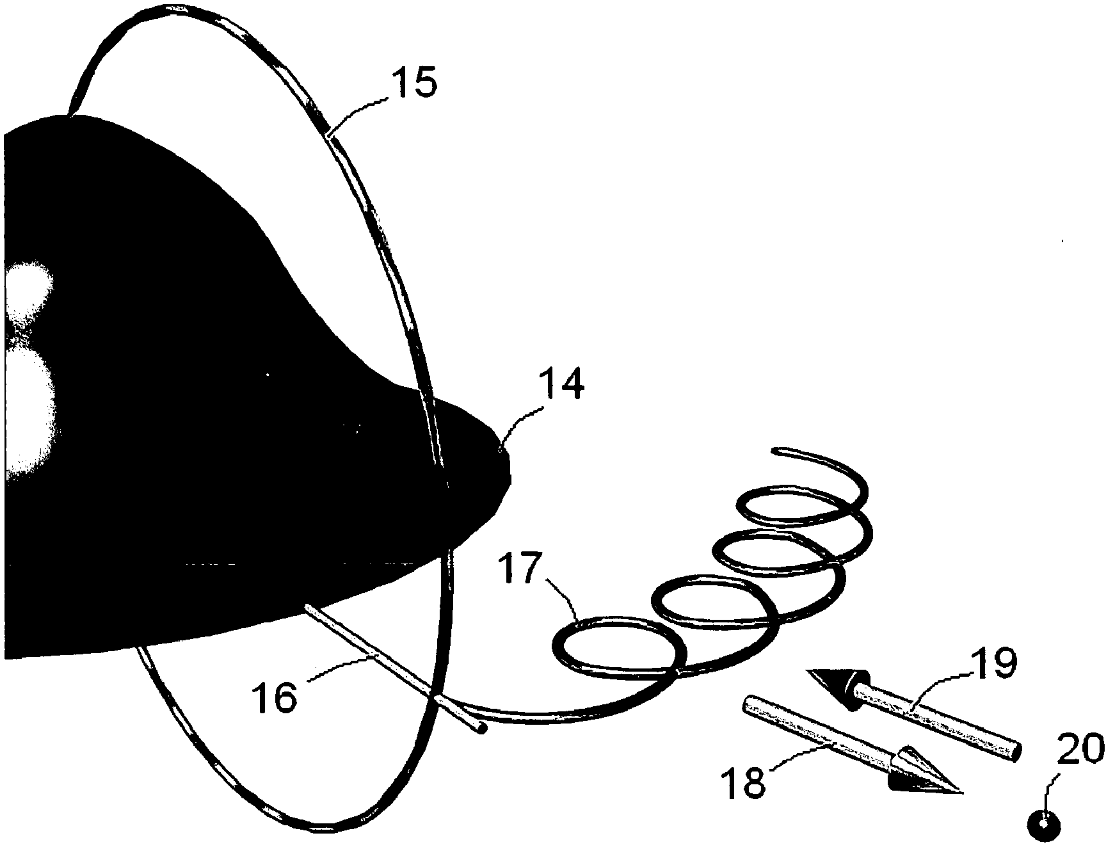

Figure 1

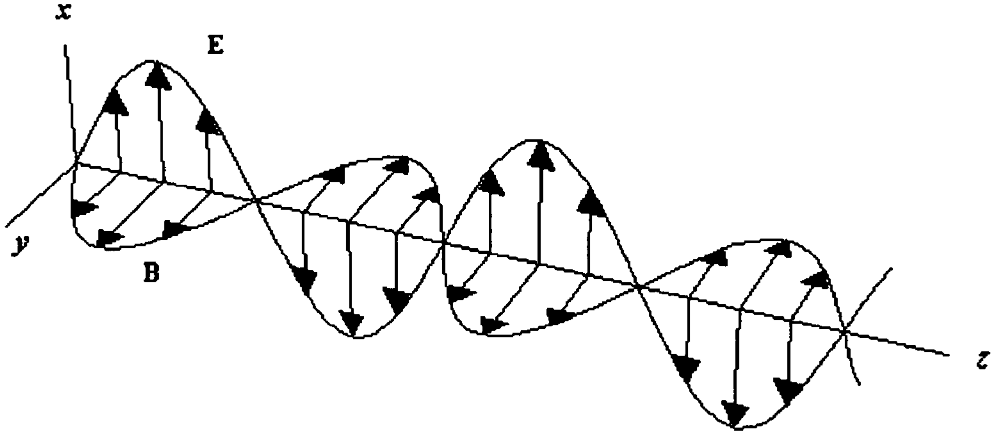

Figure 2

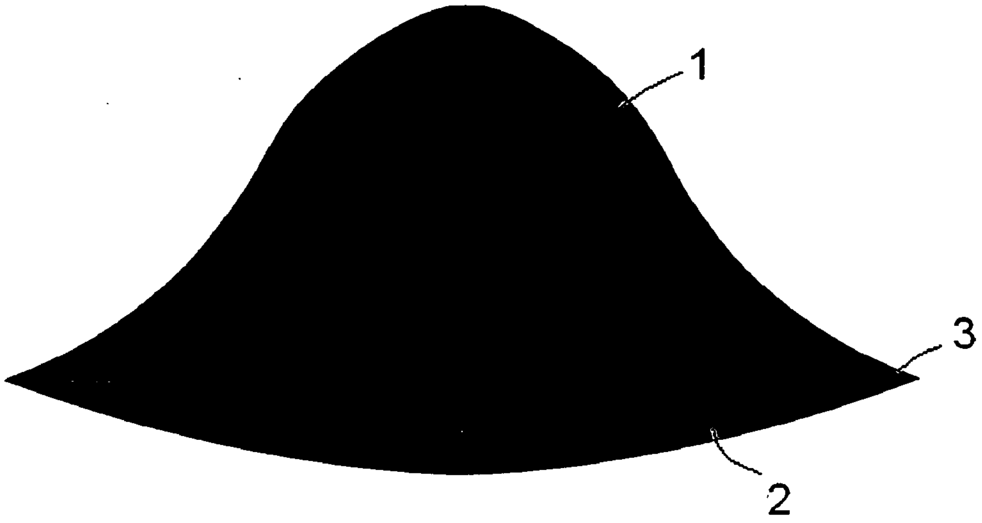

Figure 3

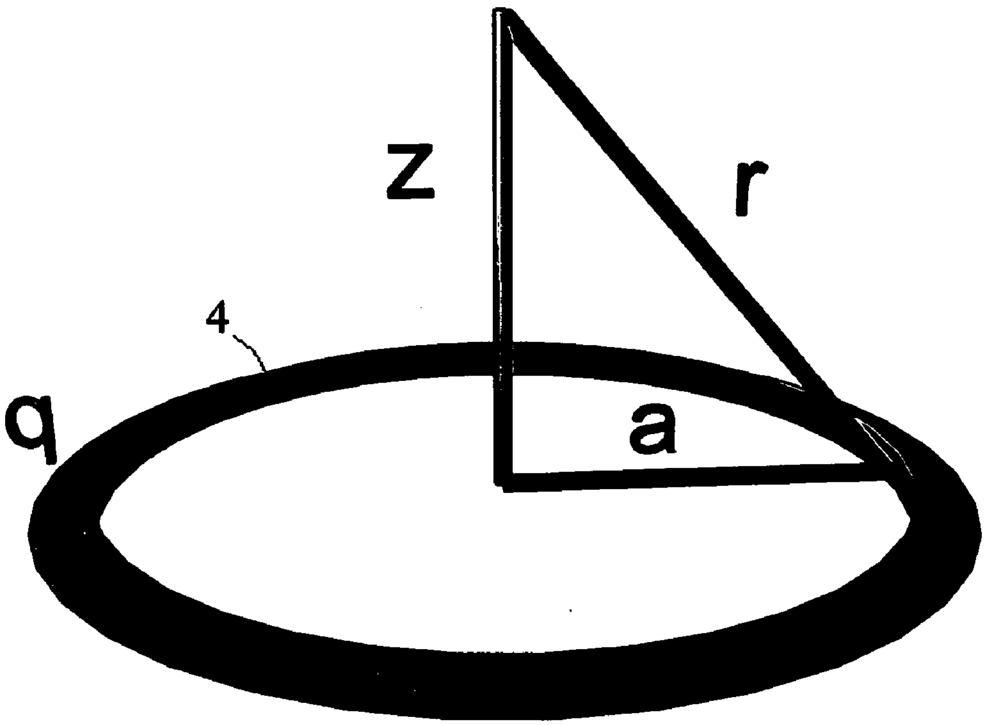

Figure 4

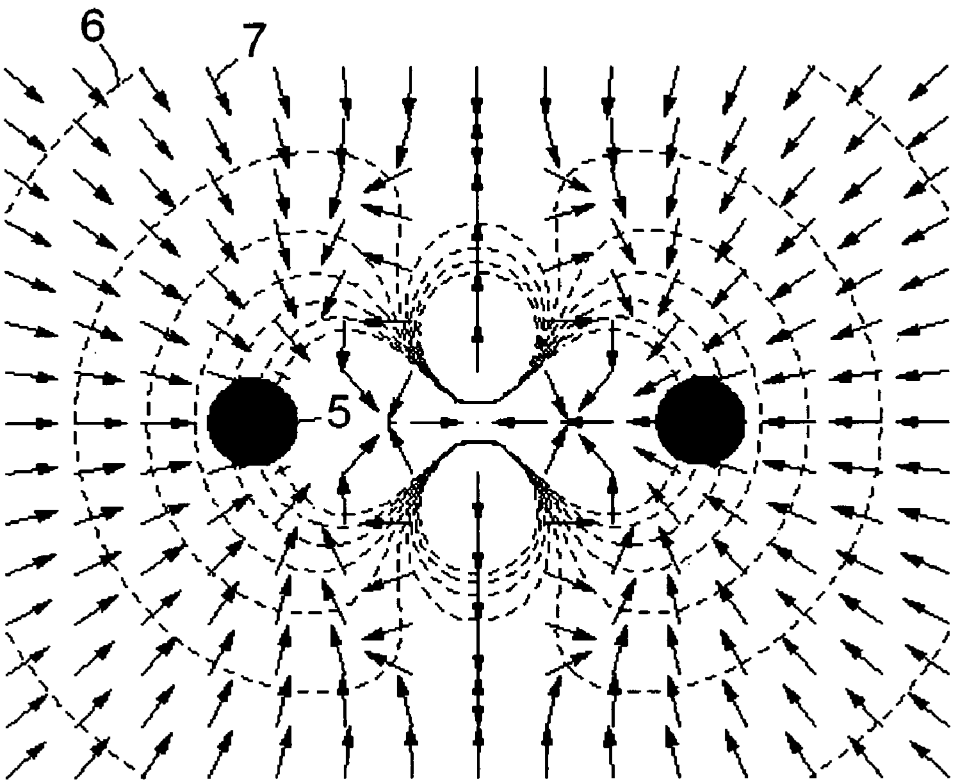

Figure 5

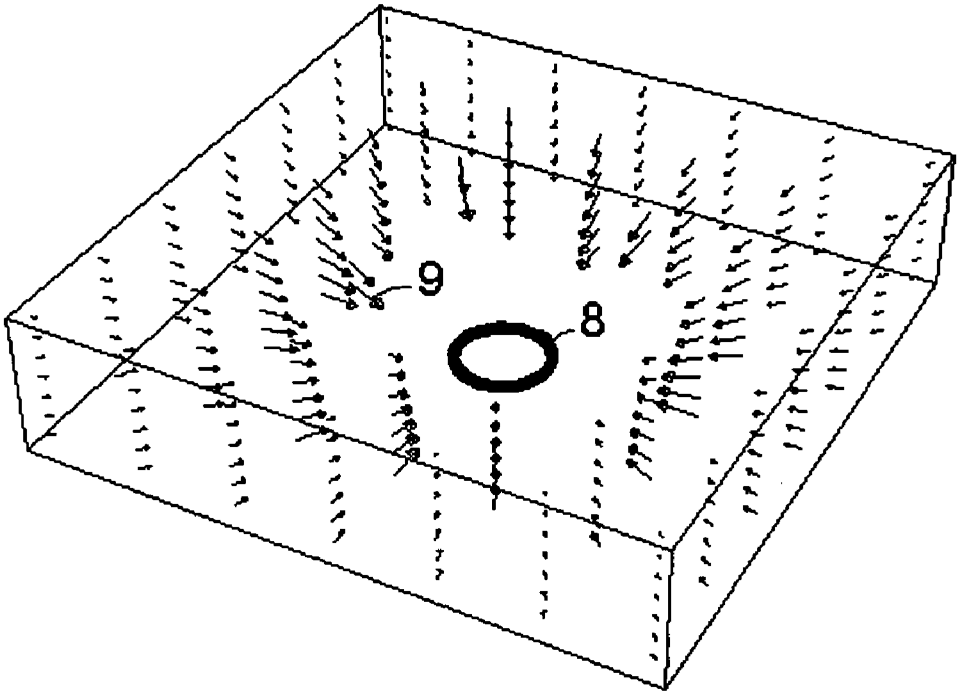

Figure 6

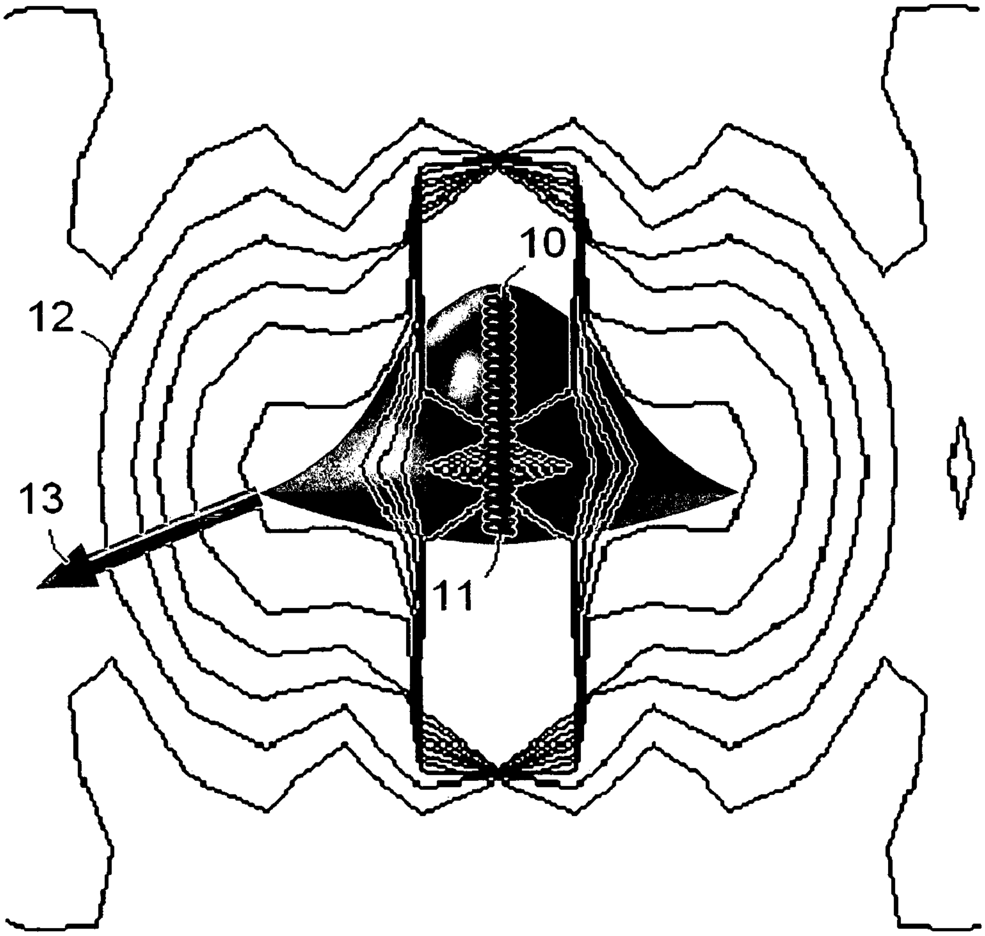

Figure 7

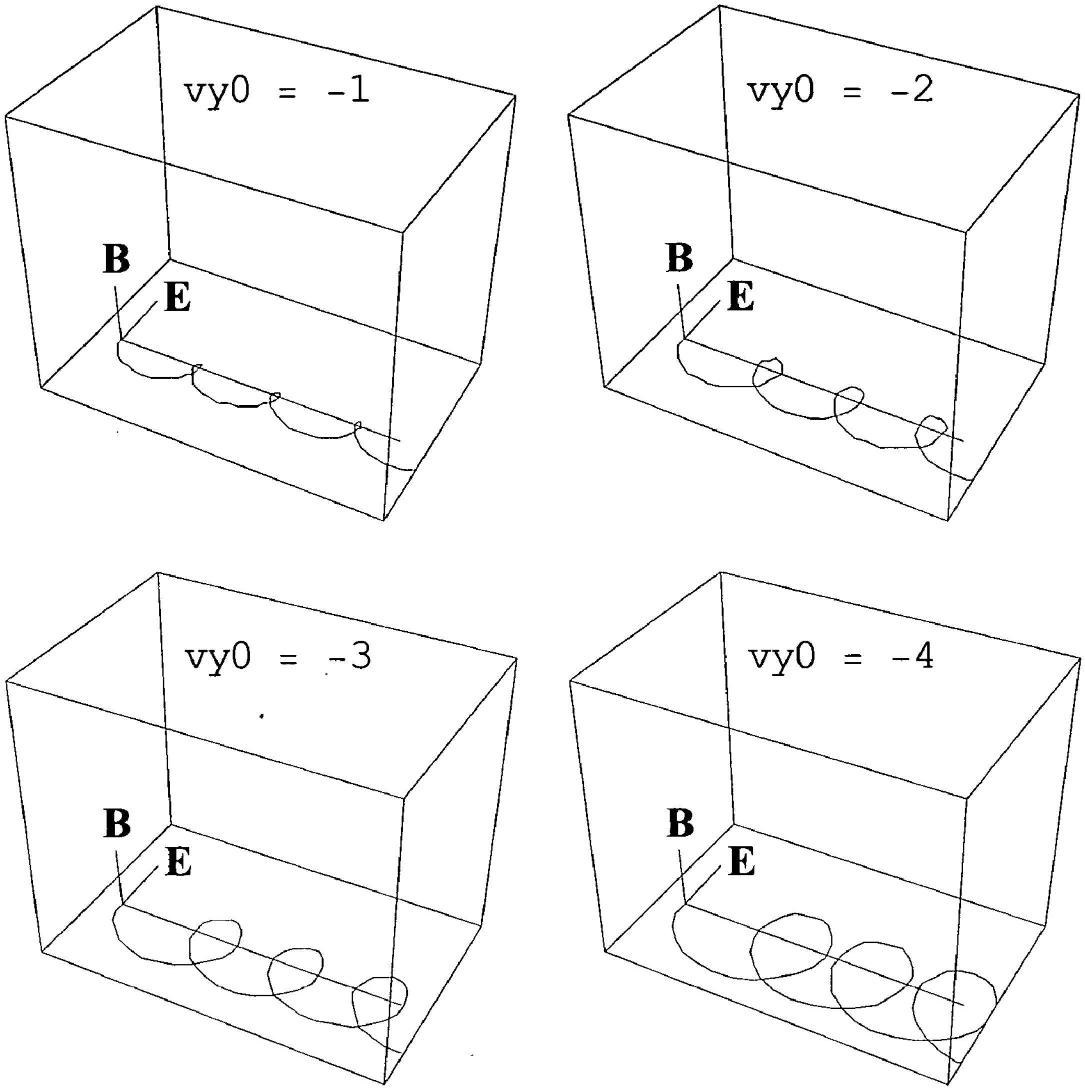
vy0 = -1
B
E
vy0 = -2
B
E
vy0 = -3
B
E
vy0 = -4
B
E

Figure 8

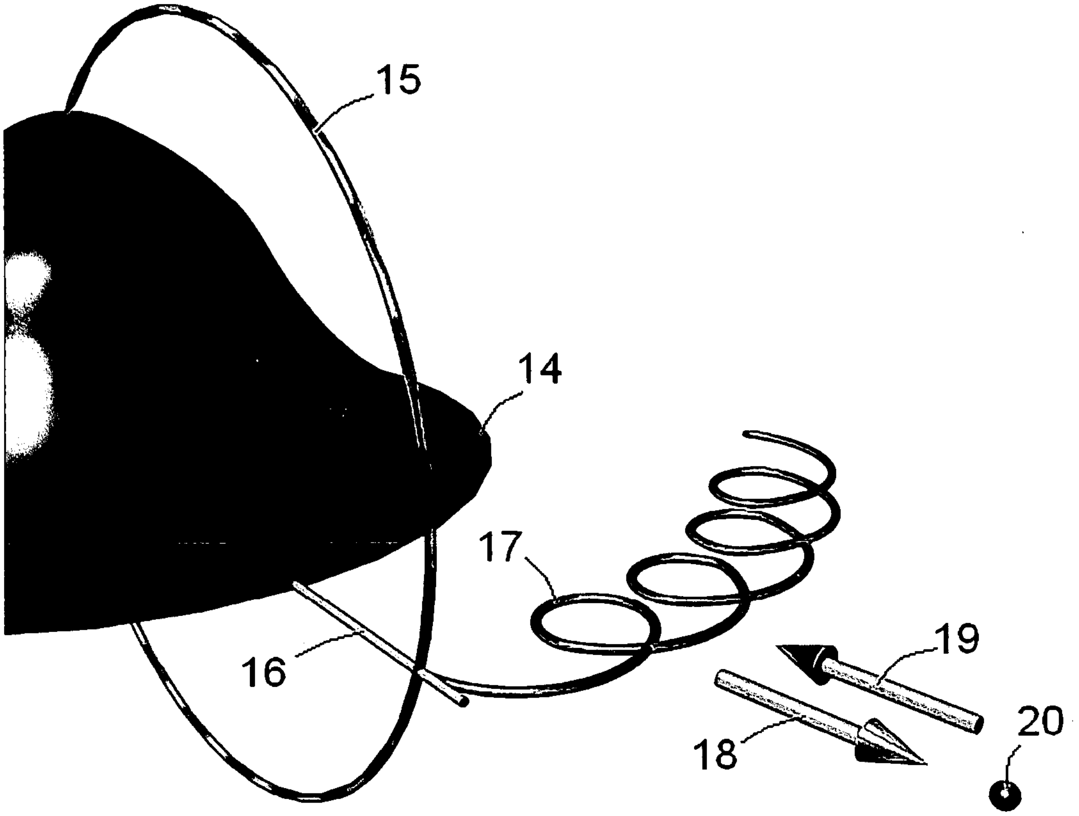
15
14
17
19
16
20
18

Figure 9

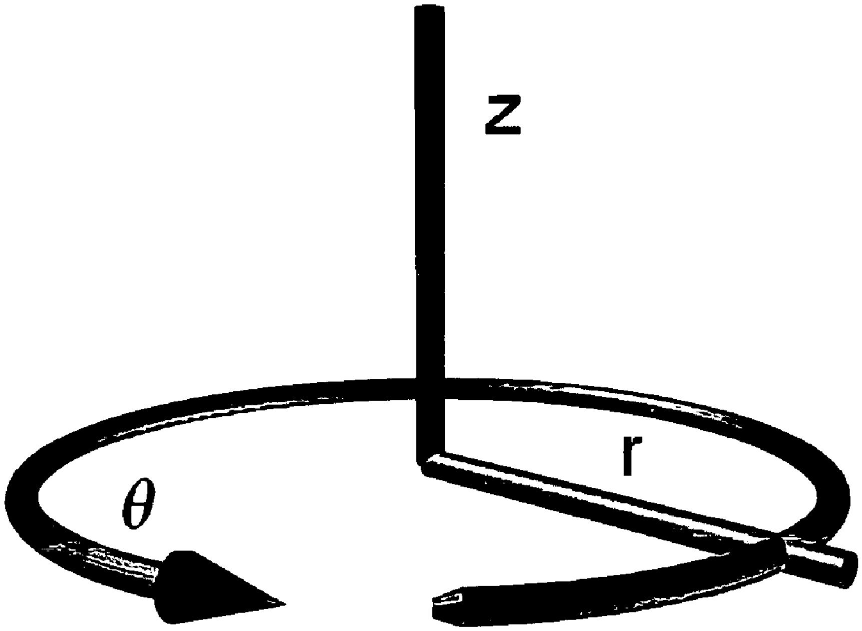

PHOTON SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft propulsion system that employs photon particles to generate a field of negative energy in order to produce lift on the hull.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, an electromagnetic wave traveling in the z-direction consists of an electric E field vibrating in the x-direction and a magnetic flux density B field vibrating at right angles in the horizontal y-direction. The energy-stress-momentum of this photon can be analyzed using Einstein's General Theory of Relativity and the Faraday F tensor. The Faraday tensor is a 4×4 matrix containing the electromagnetic wave components as shown here in general where c is the speed of light

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & \frac{E_x}{c} & \frac{E_y}{c} & \frac{E_z}{c} \\ x & \frac{E_x}{c} & 0 & B_z & -B_y \\ y & \frac{E_y}{c} & -B_z & 0 & B_x \\ z & \frac{E_z}{c} & B_y & -B_x & 0 \end{array}$$

For this particular photon, this tensor is

$$F^{\alpha}_{\beta}=\begin{array}{c|cccc|} t & 0 & \frac{E_x}{c} & 0 & 0 \\ x & \frac{E_x}{c} & 0 & 0 & -B_y \\ y & 0 & 0 & 0 & 0 \\ z & 0 & B_y & 0 & 0 \end{array}$$

[0003] The elemental spacetime length ds squared is equal to sum of the squares of the Cartesian elemental lengths

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

The coefficients of this equation, {−1,1,1,1} are the diagonal components of the g metric tensor

$$g_{\alpha\beta}=\begin{array}{c|cccc|} t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

The stress-energy-momentum tensor T can then be calculated for the photon using the Faraday tensor and the g metric tensor in the following equation from gravitation physics

$$4\pi T^{\mu\nu}=F^{\mu\alpha}F^{\nu}_{\alpha}-\frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

The stress-energy-momentum tensor indicates the curvature of space due to the application of electromagnetic fields, mass, angular momentum and charge. The mass of the Earth, for example, generates a negative curvature of spacetime such that objects fall toward the mass. The T tensor, which is also a 4×4 matrix, contains the momentum or flux terms in the first row and first column. The normal pressure stress is located along the diagonal. The shearing stresses are located off the diagonal. The energy term is in the upper left corner as depicted here,

$$T^{\mu\nu}=\begin{array}{c|cccc|} t & \text{energy} & flux_x & flux_y & flux_z \\ x & -flux_x & pressure_x & shear_{xy} & shear_{xz} \\ y & -flux_y & shear_{yx} & pressure_y & shear_{yz} \\ z & -flux_z & shear_{zx} & shear_{zy} & pressure_z \end{array}$$

[0004] Since $B^2=E^2/c^2$, the stress-energy-momentum tensor for the photon is therefore

$$T^{\mu}_{\nu}=\begin{array}{c|cccc|} t & \frac{-E^2}{c^2} & 0 & 0 & \frac{+E^2}{c^2} \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & 0 \\ z & \frac{-E^2}{c^2} & 0 & 0 & \frac{+E^2}{c^2} \end{array}$$

This remarkable result shows that the photon is actually a negative energy particle (top left corner) which is pushed along by a positive pressure wave (lower right corner). The particle has a positive flux (upper right corner) in the z-direction, as well as a balancing negative flux in the lower left corner so that the overall momentum of the universe remains the same. All four components cancel and we see the photon as a massless particle moving at the speed of light.

[0005] Thus the key idea behind this invention is that it is possible to cancel out the pressure term and leave a stationary vibrating electromagnetic field of negative energy over the hull of the spacecraft. The importance of negative energy is that it is a prerequisite to generating wormholes between space and hyperspace.

[0006] Hyperspace consists of the those co-dimensions which have different physics constants such as a low speed of light. The existence of hyperspace, which has a white misty look, is not a well-known scientific concept. Experiments with our magnetic vortex wormhole generators, hyperspace torque generator, full body levitation using Chi Kung breathing, arm levitation by spinning the co-gravitational K field, full body teleportation through hyperspace a distance of 100 meters using a pulsed gravitational wave, jumping into hyperspace, having a plate of toast enfold off the breakfast table and disappear into thin air, walking through walls and doors out-of-dimension, looking into other dimensions, remote viewing through subspace to distances of 100,000 light years, and other electromagnetic experiments carried out by co-researchers, have shown us the reality and existence of hyperspace.

[0007] Referring to **FIG. 2**, the spacecraft consists of an upper (**1**) and lower (**2**) hull attached by ceramic insulators to a circular ring (**3**). The ring provides support and is attached to an outer sharp-edged rim which is electrostatically charged to a potential −V. The purpose of the charged rim is to generate a radial electric E field around the vehicle.

[0008] Referring to **FIG. 3**, the radius of the ring (**4**) is equal to a. The distance from a point on the ring to the z-axis is r. The potential on the z-axis is therefore the charge divided by the distance,

$$potZ = \frac{q}{\sqrt{a^2 + r^2}}$$

This potential is expanded as a series in terms of inverse radius r

$$potZout = \frac{35qa^8}{128r^9} - \frac{5qa^6}{16r^7} + \frac{2qa^4}{8r^5} - \frac{qa^2}{2r^3} + \frac{q}{r}$$

The potential outside the ring can be written in terms of the Legendre polynomials P

$$Vout = \sum_{n=0}^{s} \left(\frac{a}{r}\right)^{n+1} A[n] LegendreP[n, \mathrm{Cos}(\theta)]$$

where s is the number of terms in the expansion. By equating the known particular solution potZout on the z-axis with the general Vout solution, the coefficients A[n] are found to be

$$A(0) = \frac{q}{a}$$

$$A(1) = 0$$

$$A(2) = \frac{-q}{2a}$$

$$A(3) = 0$$

which are substituted back into the Vout equation to get the potential outside the ring.

[0009] Referring to **FIG. 4**, the potential (dotted lines **6**) looking at a slice through the ring (**5**) is shown together with the electric E field. The negative gradient of the potential is the electric field (**7**) shown by the direction of the arrows. The importance of this diagram is that the electric field points in the radial direction toward the negatively charged ring. The force on an electron is the electron charge times the electric field

$$F=q_eE_r=-/q_e|(-|E_r|)=+F$$

Because the electron charge is negative and the radial field points in the negative direction toward the ring, the force on the electron is positive. Thus the electron moves away from the ring in the positive radial direction. A 3-dimensional plot of the ring (**8**) and the electric field (**9**) is shown in **FIG. 5**.

[0010] The stress-energy-momentum generated by a radial electric field is calculated using the Faraday F tensor

$$F_\alpha^\beta = \begin{vmatrix} 0 & E_r & 0 & 0 \\ E_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

The g metric tensor has to be given in spherical coordinates {r,θ,ϕ}

$$g_{\alpha\beta} = \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & r^2\sin(\theta)^2 \end{vmatrix}$$

where θ is the angle from the vertical to the radius r. The stress tensor T^{rr} along the radial direction is

$$T^{rr} = \frac{E_r^2}{8\pi c^2}$$

which shows that the pressure is negative along the radial line equal to the square of the radial electric field divided by the square of the speed of light. Because the field is squared, it doesn't matter that the electric field points in the negative direction. The square makes it positive, but the overall curvature pressure is negative. Thus this negative pressure cancels out the positive pressure propelling the photon along. The second key idea of the invention is how to generate this photon moving in the radial direction.

[0011] It has been known for a long time in physics that an electron moving in a circular path will emit photons in a process known by the German word Bremsstrahlung which is translated as "breaking radiation." There are several types of radiation such as classical Bremsstrahlung involving a charged particle making a collision with another charged or uncharged particle in which photons are emitted. The quantum mechanical Bremsstrahlung involves the sudden appearance or disappearance of a charged particle which also emits radiation. In space, having a field of wormholes in which the electrons are spiraling down into hyperspace would result in the emission of photons by the quantum mechanical method. Also, in the atmosphere, having collisions with air molecules results in emission of photons in the classical way.

[0012] In order to get the electrons to spiral around and emit photons, a crossed electromagnetic field is used as shown by the following equation

$$F=q(E_r+v_r \times B_\theta)$$

where the velocity v is in the positive radial direction due to the force of the electric field. The velocity crossed with a magnetic flux density B field in the θ-direction makes the electron move sideways back and forth in a wiggling motion.

[0013] Referring to **FIG. 6**, a direct current solenoid (**10**), represented by multiple current loops, running vertically through the center of the hull, generates a magnetic field that curves around the outside of the hull, as shown by contour lines (**12**). The north pole (**11**) is at the bottom of the hull. A radial arrow (**13**) from the electrostatically-charged rim is perpendicular to the magnetic field lines. The cross product in the force equation becomes the electron radial velocity times the magnetic field $v_r\ B_\theta$.

[0014] Referring to **FIG. 7**, the electric field is in the y-direction and the magnetic field is in the z-direction. The flat looping path in the x-direction is the motion of the electron. The electron, which has a negative charge, starts to move in the direction opposite to that of the electric field. In this particular diagram, the electron acquires a velocity in the negative y-direction. Then a sideways force in the x-direction is produced due to the cross product of the velocity with the magnetic field times the negative charge

$$-q(-v_y \times B_z)=+F_x$$

Depending on the magnitude of the velocity, various size loops can be produced.

[0015] In terms of the hull coordinates, because the flat loop is in the plane of the electric field which points in the radial direction, the electron emits light in the radial direction. This condition means that the negative radial pressure created by the electric field cancels the radial pressure of the photon. Thus the photon becomes a stationary vibrating quantum of negative energy. This has the appearance of a luminescent light source. The stress tensor for this condition is therefore

$$T_\mu^\nu = \begin{vmatrix} -\frac{E^2}{c^2} & 0 & 0 & \frac{E^2}{c^2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{E^2}{c^2} & 0 & 0 & 0 \end{vmatrix} = -\frac{E^2}{c^2}\text{residual negative energy}$$

[0016] residual negative energy

which leaves a residual negative energy per photon.

[0017] Referring to **FIG. 8**, the negatively charged rim (**14**) produces a radial electric field (**16**) that crosses the magnetic B field (**15**) of the solenoid. Electrons emitted by the charged rim then encounter this crossed field which makes them spiral (**17**) around the hull. Because of the tight loop, the electron emits Bremsstahlung radiation in the radial direction (**18**). The positive pressure field of the photon, which is directed in the radial direction, is canceled by the negative pressure field (**19**) created by the electric field. Because the photon energy is negative, a stationary vibrating electromagnetic quantum of negative energy (**20**) surrounds the hull.

[0018] This negative energy and the pressure stress created by the electromagnetic fields open up wormholes between space and hyperspace. The potential head is positive from hyperspace into space because the energy of hyperspace is more positive than the negative energy field. The low-density hyperspace energy fills the hull and its surrounding space with a white misty hyperspace energy which makes the spacecraft lighter in mass, and therefore lighter in weight within a gravitational field. The actual physics is more complicated still because the electrons find that the resistance of hyperspace is lower than the resistance of space. Thus they spiral down the wormholes which results in a sudden disappearance of charge. The quantum mechanical effect of this is to radiate even more photons which in turn produce even more negative energy.

[0019] The lift on the hull is generated by the radial electric field. In cylindrical coordinates, the g metric tensor is

$$g_{\alpha\beta} = \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Using this metric tensor, the pressure stress in the vertical direction T^{zz} is

$$T^{zz} = \frac{E^2}{8\pi c^2}$$

which is a positive curvature over the hull. The mass of Earth produces a negative curvature in which objects fall toward the mass. By counteracting this negative curvature with a more than positive curvature, lift is developed on the spacecraft. Because the negative energy lowers the effective mass of the vehicle, the acceleration is large with a modest electric field. Moreover, in our dimension, the speed of light is 299792458 meters per second. Hyperspace energy has a speed of light equal to one meter per second. Thus the stress is amplified by a factor of

$$A = \left(\frac{299792458\,m/s}{1m/s}\right)^2 \approx 9 \cdot 10^{16}$$

Because electromagnetic fields are relativistic, motion in a low-velocity-of-light energy field amplifies their strength.

SUMMARY OF THE INVENTION

[0020] It is the object of this invention to create a spacecraft propulsion system that produces wormholes between space and hyperspace using negative energy in order to generate lift on the hull. It was discovered in the Riemannian curvature calculations of gravitation physics that negative energy is required to keep open the throat of the wormhole. From experiments with the magnetic vortex wormhole generator, it is known that the proper combination of electromagnetic fields, together with this negative energy, can create a wormhole through which smoke can be blown into hyperspace.

[0021] Referring to **FIG. 9**, the directions of force, velocity, and electromagnetic fields are referred to in the cylindrical coordinate system $\{r,\theta,z\}$. An electrostatically charged sharp-edged ring in the θ-direction around the hull of the spacecraft produces a radial electric field. A vertical solenoid in the z-direction through the center of the hull produces a magnetic field which is perpendicular at the rim to the electric field. With the current in the solenoid flowing in the clockwise (−θ) direction, using the right-hand rule, the magnetic field points in the upward z-direction outside the rim. Because the rim is charged to a negative voltage, the electric field points toward the hull in the negative radial (−r) direction. Electrons emitted by the rim travel outward (+v) because the charge on the electron is negative which, together with the negative electric field, produces a positive radial force. The radial force on the electron causes it to acquire a velocity which interacts with the magnetic field. The cross product of the velocity (+v) with the positive (+B) magnetic field produces a sideways force on the electron in the negative θ-direction. However, because the charge on the electron is negative, the force is

$$F=-q\{v_r,0,0\}\times\{0,0,B_z\}=\{0,qB_zv_r,0\}$$

which is positive in the θ-direction. It is this sideways force that produces a flat spiraling or looping motion whereby the electron emits photons, known in German as Bremsstahlung radiation, in the radial direction. The photon, which is actually a quantum of negative energy, has a positive radial pressure which propels it along. Because the radial electric field produces a negative pressure in the radial direction, the two opposite fields cancel in the radial direction to form a residual stationary vibrating negative energy. Thus the hull becomes surrounded by negative energy which, together with the pressure stresses created by the electric field, generates wormholes between space and hyperspace.

[0022] The gravitational potential between hyperspace and space is positive because the hyperspace energy is more positive than the negative energy around the hull. Thus the low-density, low-speed-of-light hyperspace energy flows through the wormhole and fills the hull. This has the effect of reducing the effective mass of the hull. Because the electric field generates a positive pressure over the hull in the vertical z-direction, there is an upward force on the vehicle due to the pressure times the hull area. Since the vehicle has a low mass, there is a modest upward acceleration on the spacecraft equal to the force divided by mass.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0023] **FIG. 1**. Perspective view of an electromagnetic wave.

[0024] **FIG. 2**. Perspective view of spacecraft.

[0025] **FIG. 3**. Perspective view of charged ring.

[0026] **FIG. 4**. Planar plot of the radial electric field produced by charged ring.

[0027] **FIG. 5**. Perspective view of radial electric field around ring.

[0028] **FIG. 6**. Planar view of magnetic flux density field contour lines.

[0029] **FIG. 7**. Perspective view of electron motion in crossed electric and magnetic fields.

[0030] **FIG. 8**. Perspective view of production of negative energy around hull.

[0031] **FIG. 9**. Perspective view of cylindrical coordinate system {r,θ,z}.

DETAILED DESCRIPTION OF THE INVENTION

[0032] 1. The hull is made from a single sheet of aluminum which has been stretched to its yield point by hydraulic cylinders. An upper and lower die is CNC machined to the profile of the hull. The soft sheet is then clamped in the die where it takes on the smooth shape of the hull without any wrinkles. The hull is extremely rigid after forming and does not require any structural reinforcements.

[0033] 2. A section of the aluminum ring is made in a 3D computer graphics program. The model is stored as a stereolithography file (*.stl). The computer model is then sent via Internet e-mail to the stl server who prints the part in an ultraviolet light-cured polymer. The part is returned the next day by Express Mail. Using a rubber blanket mold to create several ring sections, the entire ring is assembled together in another wooden mold box having thin circular laminate-coated particulate wall boards on either side of the ring. Then a liquid rubber mold is poured on top of the ring and allowed to harden overnight at room temperature. Since the rubber mold is flexible, the ring can be extracted fairly easily. This ring model is then sent to the foundry where it is cast in aluminum using the lost wax process in which a wax mold evaporates out of the sand casting. We are also experimenting with non-magnetic copper casting metals containing beryllium having good conductivity.

[0034] 3. A 11.5 cm plastic pipe is mounted on a rotating fixture driven slowly by a microcontoller, stepper motor, and power electronics board. Using a large diameter insulated wire, such as a 17 AWG with a wire diameter of 0.127 cm, the wire is wound slowly on the pipe and expoxied so that the windings don't come loose. The solenoid is then mounted vertically in the hull supported by the support ring and driven by a current generator located nearby on the test rig.

[0035] 4. The ring is driven by a high voltage electrostatic generator similar to the night vision scope high voltage power supplies. The ring charge is isolated from the hull by ceramic insulators.

I claim:

1. A spacecraft propulsion system comprising the components:

an aluminum horizontal circular structural support ring;

an aluminum hull in the shape of a high dome on top and shallow dome on the bottom attached to the circular support ring using ceramic insulators;

an electrostatically negatively-charged sharp-edged circular ring, preferably of non-magnetic aluminum or copper, attached with ceramic insulators to the outside of the support ring;

a solenoid mounted through the center of the hull in the vertical direction and attached to the center of the support ring;

an electrostatic high-voltage generator to drive the outer electrostatic ring; and

a direct high-current generator to drive the solenoid.

2. The method of claim 1, wherein a negative radial electric field is generated around the hull by placing a negative potential on the sharp-edged electrostatic ring using the electrostatic generator.

3. The method of claim 1, wherein the current-driven solenoid generates a vertical magnetic field around the hull with the north pole of the solenoid facing down through the bottom of the hull which causes the magnetic flux density field to point up outside the rim.

4. The method of claim 1, wherein electrons are emitted radially by the sharp edge of the charged ring.

5. The methods of claims **2**, **3** and **4**, wherein the crossed electromagnetic fields cause the electrons to spiral around in flat loops during which photons are emitted in the radial direction.

6. The methods of claims **2** and **5**, wherein the negative radial pressure created by the electric field cancels the positive radial pressure of the photon to leave a residual quantum of negative energy per photon around the hull.

7. The methods of claims **2** and **6**, wherein the pressure stress created by the electric field, and the negative energy combine to form wormholes between space and hyperspace.

8. The method of claim 7, wherein low-density hyperspace energy of a higher gravitational potential flows through the wormholes to fill the hull and surrounding space around the hull with the effect of reducing the effective mass of the spacecraft.

9. The method of claim 2, wherein the electric field generates a positive pressure in the vertical direction over the hull which together with the hull surface area, generates an upward lift force on the hull.

10. The method of claim 6, wherein the negative energy, having a low light speed, amplifies the strength of the electromagnetic fields and pressure stress fields.

11. The methods of claims **4** and **7**, wherein the electrons spiral down the low resistance wormholes into hyperspace such as to create a sudden disappearance of electrical charge which quantum mechanically causes a large emission of additional photons.

* * * * *

US 20040200925A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2004/0200925 A1**

St.Clair (43) **Pub. Date: Oct. 14, 2004**

(54) **CAVITATING OIL HYPERSPACE ENERGY GENERATOR**

(76) Inventor: **John Quincy St.Clair**, San Juan, PR (US)

Correspondence Address:
John St.Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/372,557**

(22) Filed: **Feb. 21, 2003**

Publication Classification

(51) **Int. Cl.^{7}** **B64C 27/22**
(52) **U.S. Cl.** **244/10**

(57) **ABSTRACT**

A hyperspace energy generator that uses cavitating oil bubbles within a magnetic field in order to create wormholes between space and hyperspace for the purpose of permeating the hull of a spacecraft with low-density hyperspace energy.

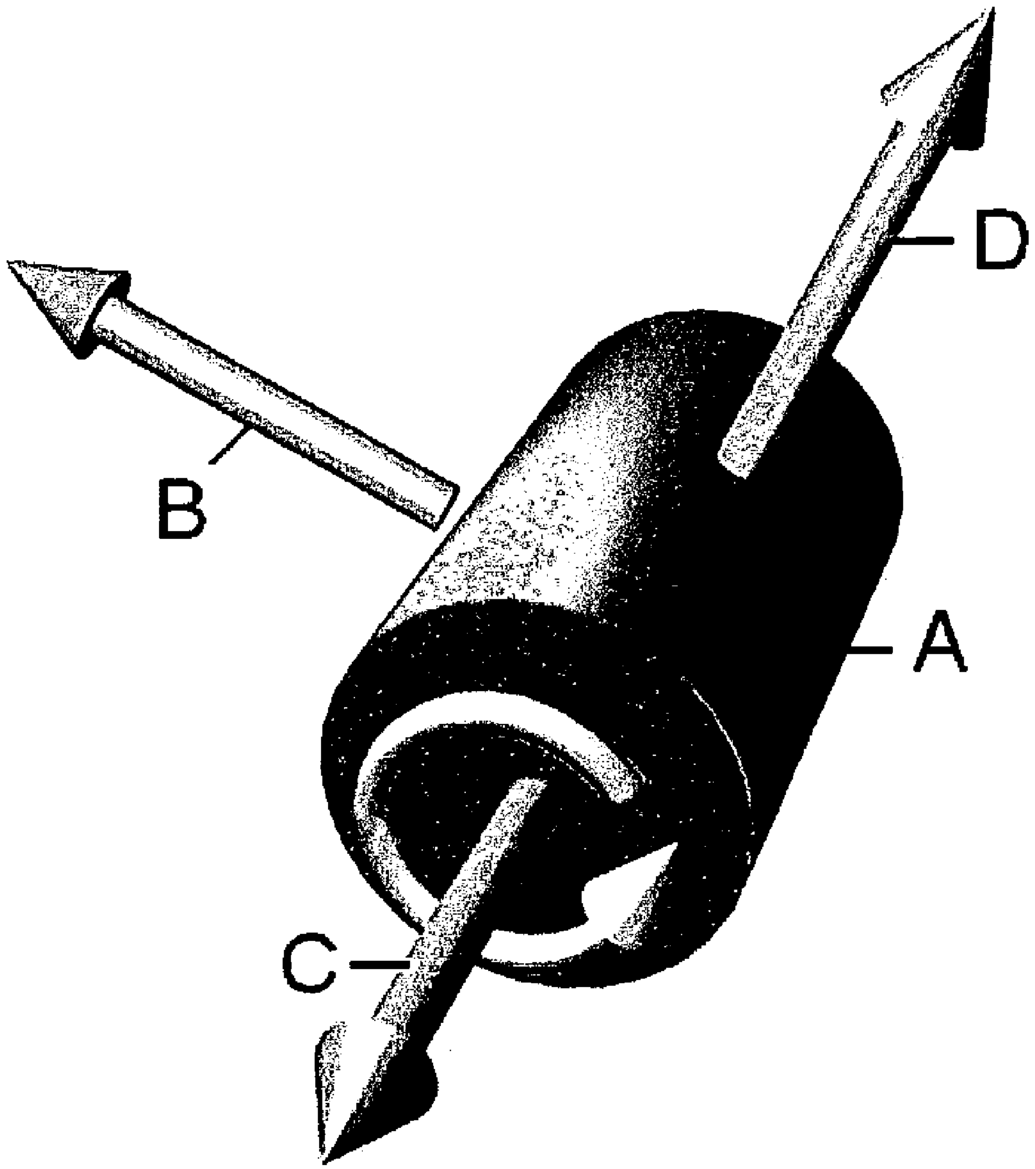

Figure 1

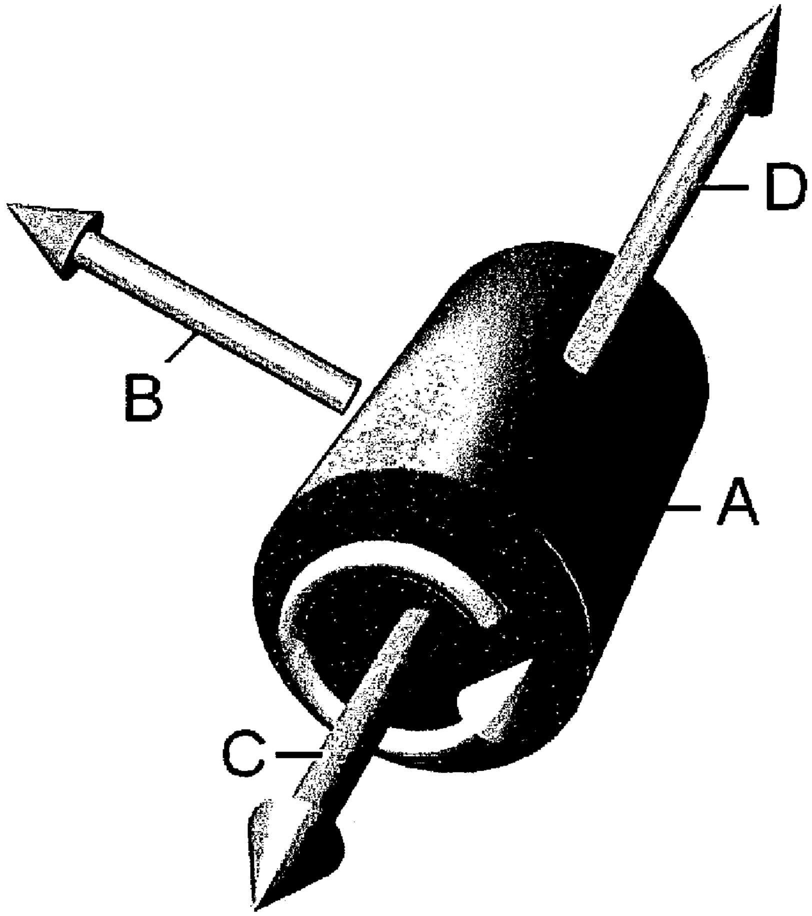

Figure 2

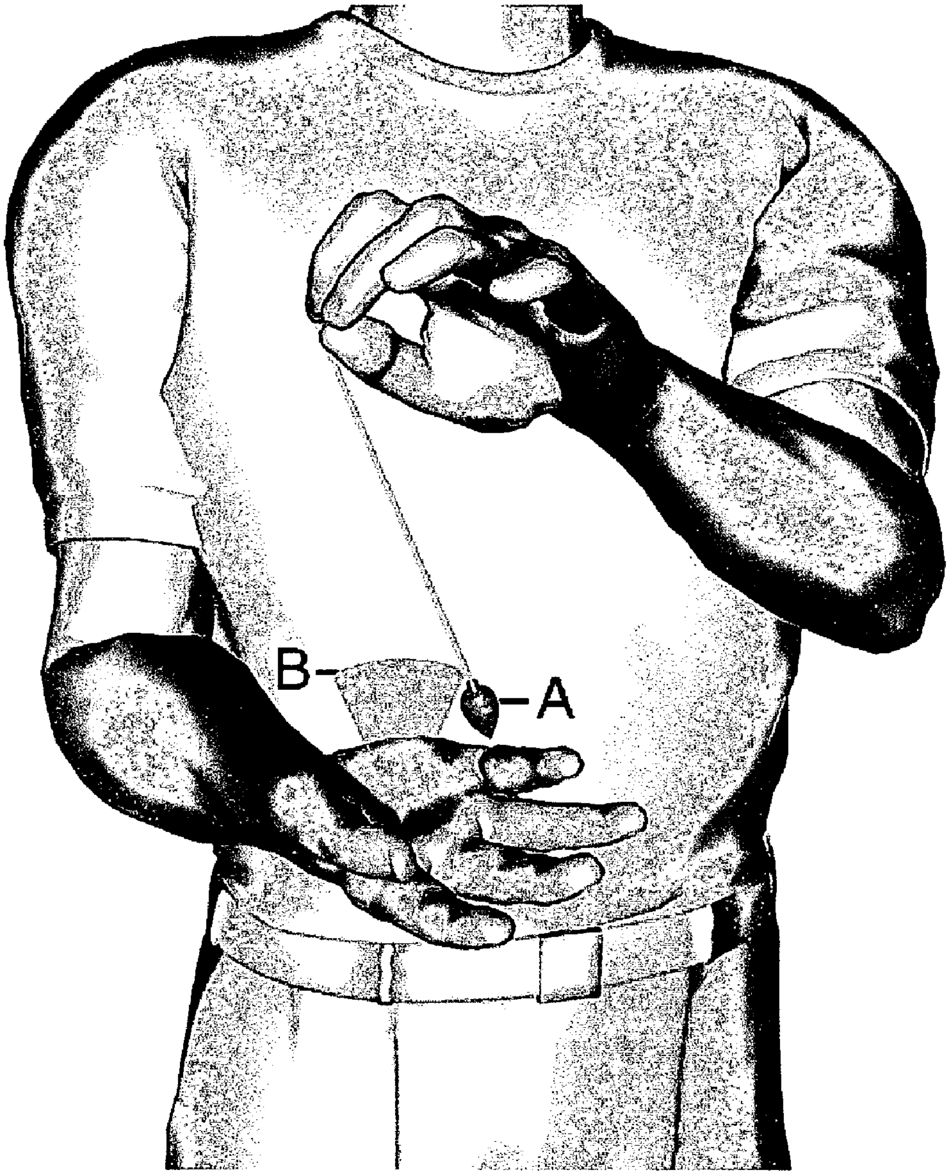

Figure 3

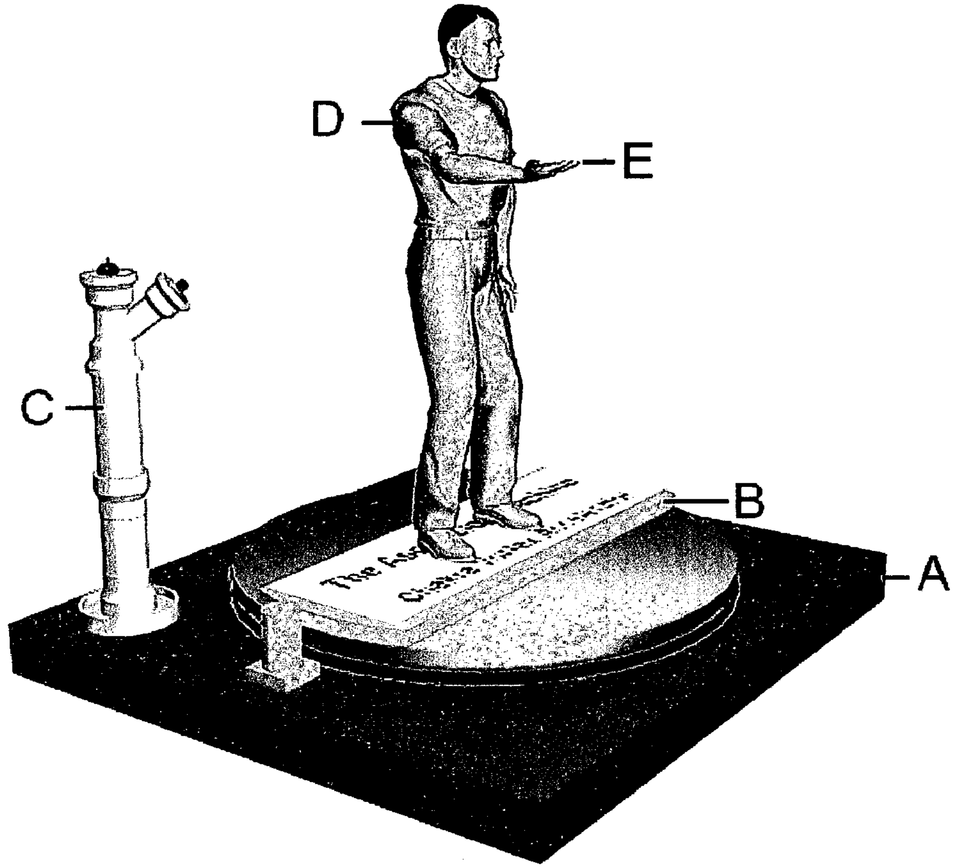

Figure 4

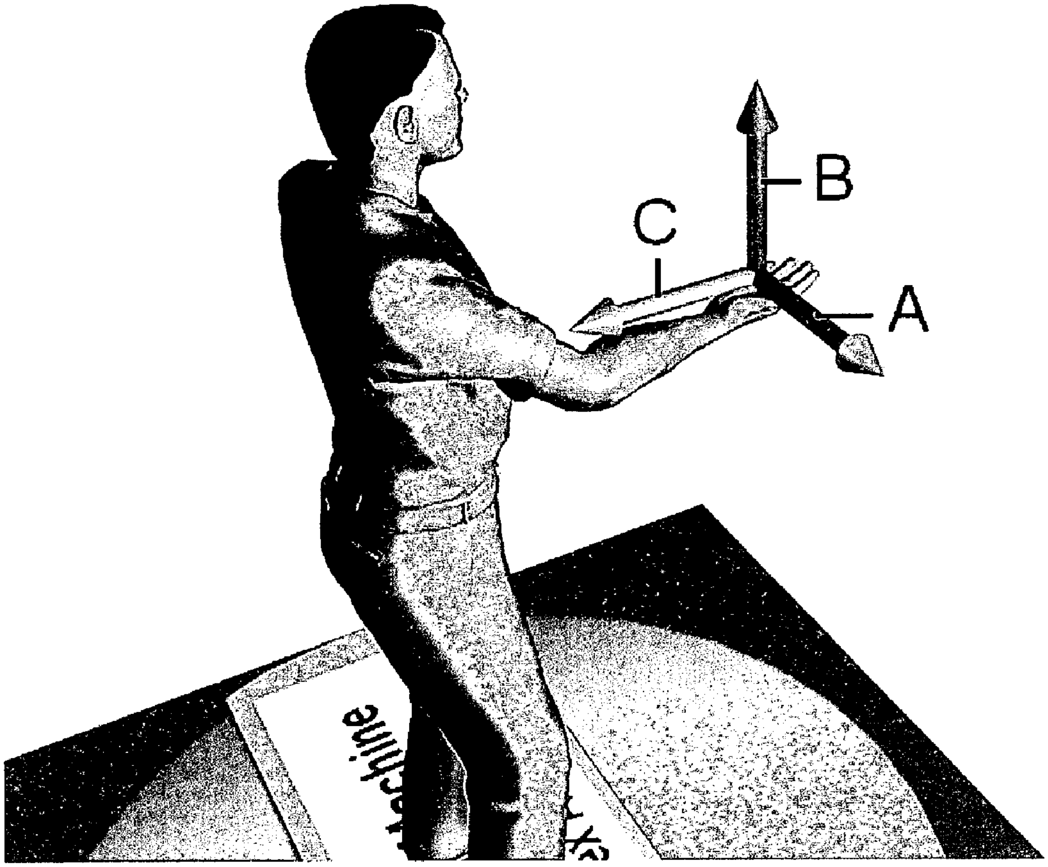
B
C
A
Machine

Figure 5

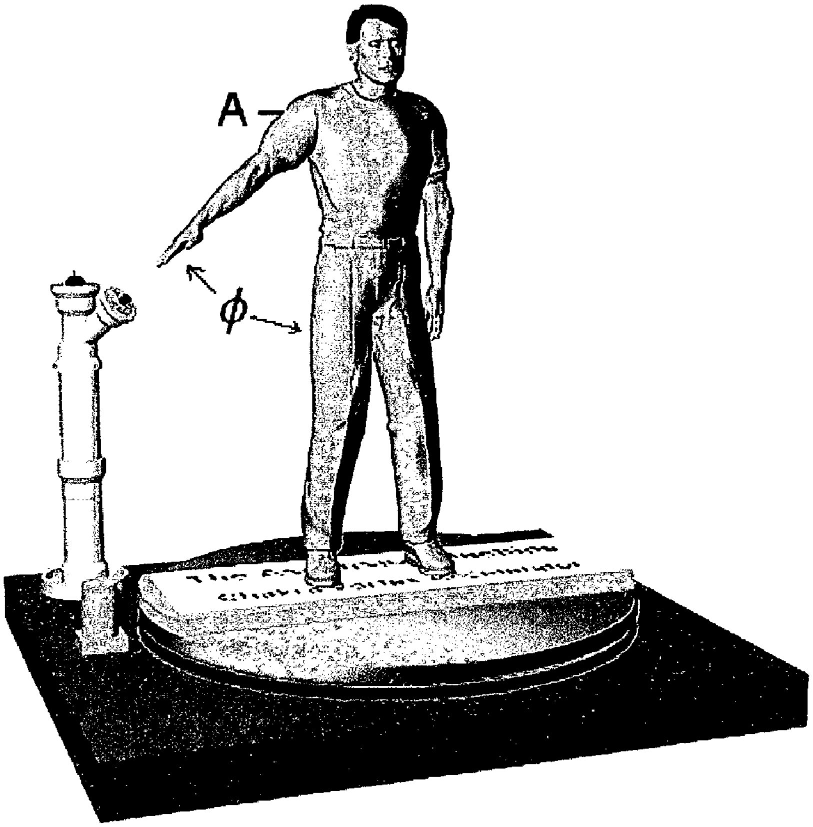

Figure 6

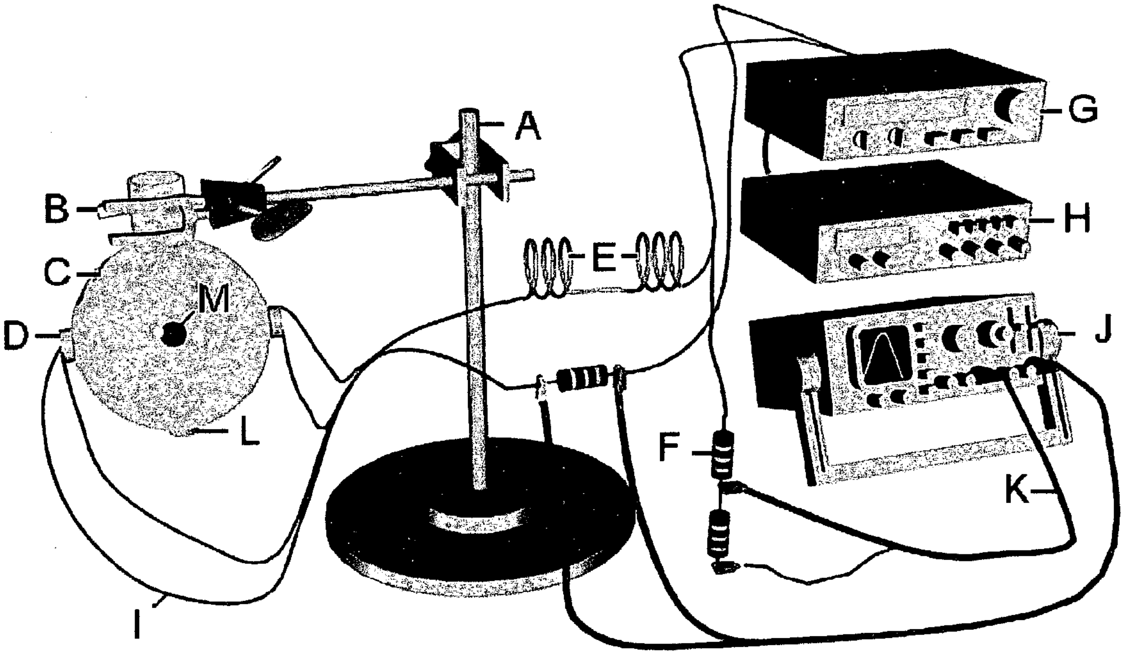
A
B
G
H
E
C
M
D
J
L
F
K
I

Figure 7

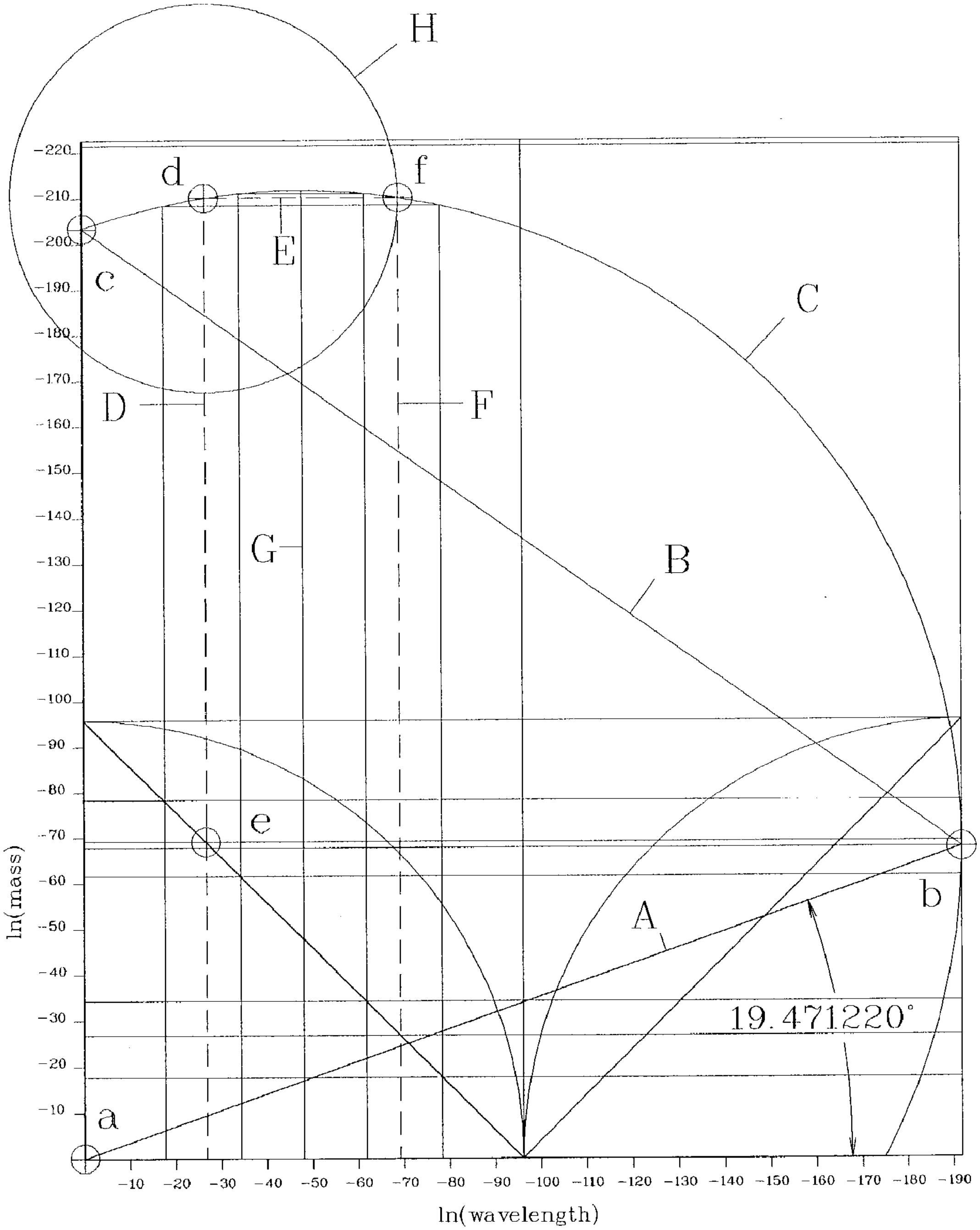
H
d
f
E
c
C
D
F
G
B
e
b
A
19.471220°
a
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)

Figure 8

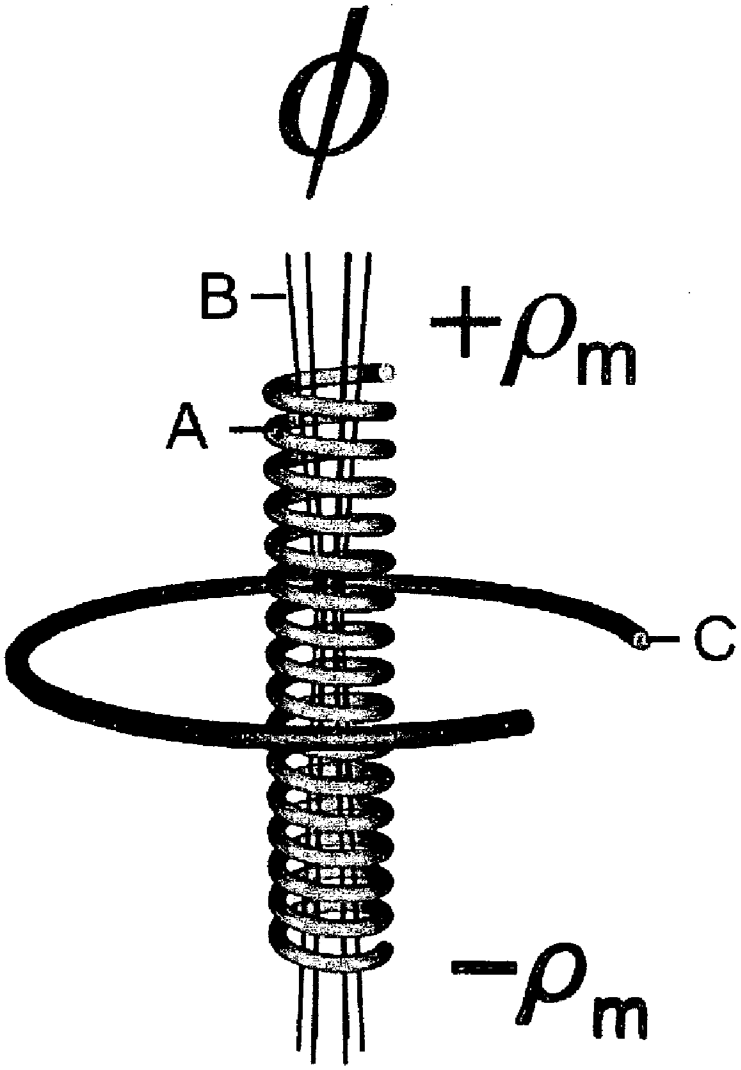
ϕ
B
$+\rho_m$
A
C
$-\rho_m$

Figure 9

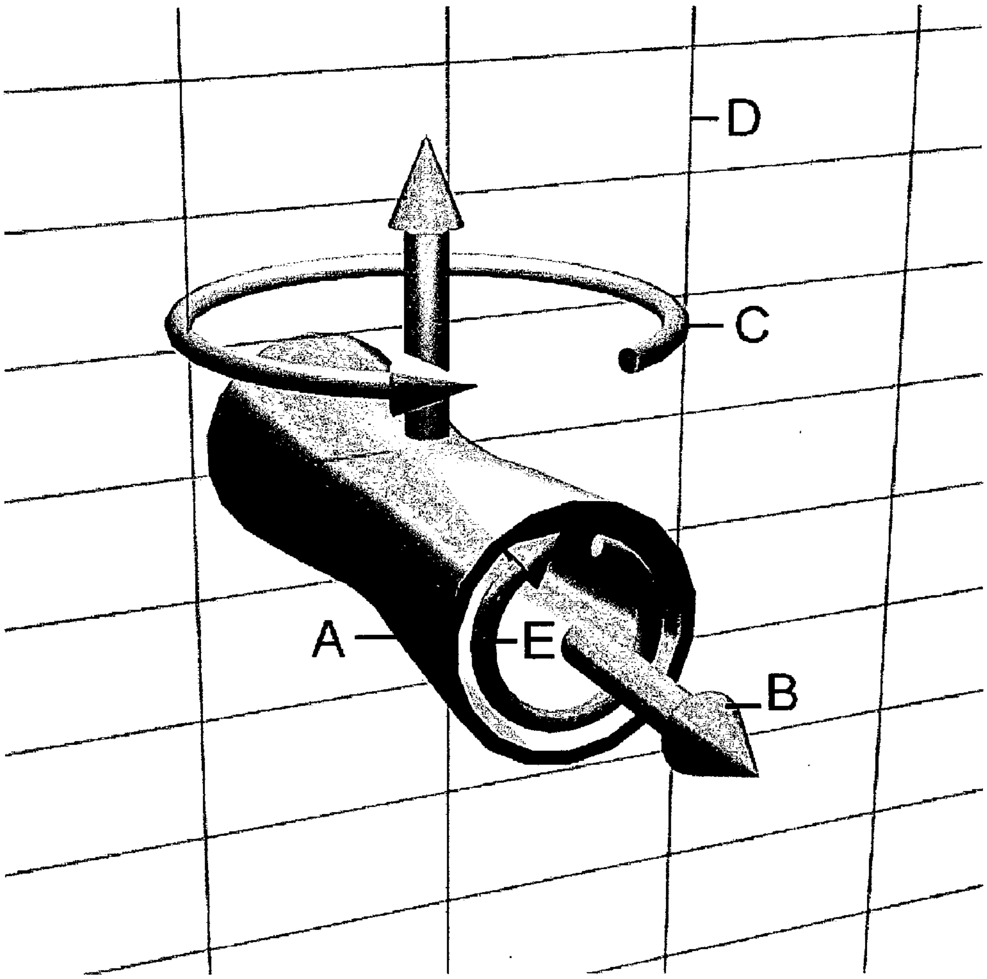

Figure 10

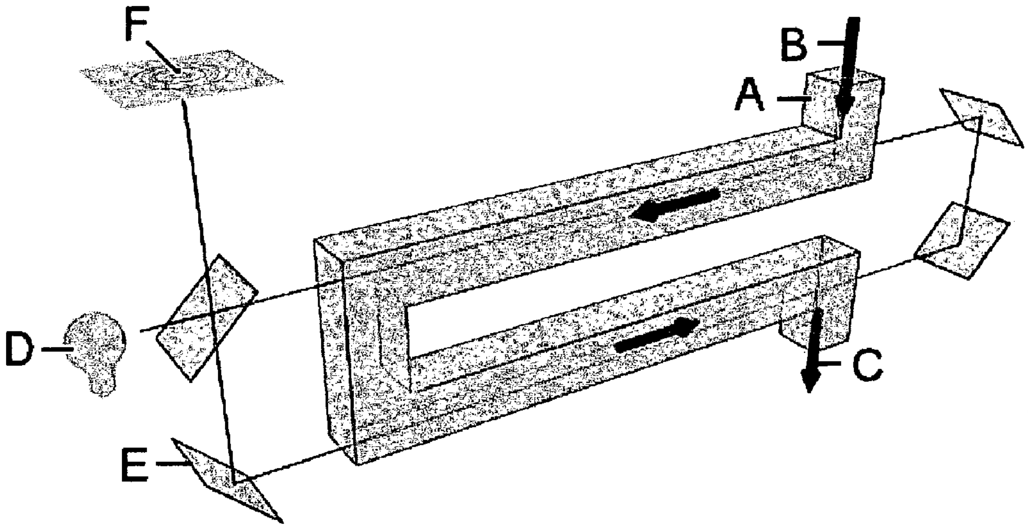

Figure 11

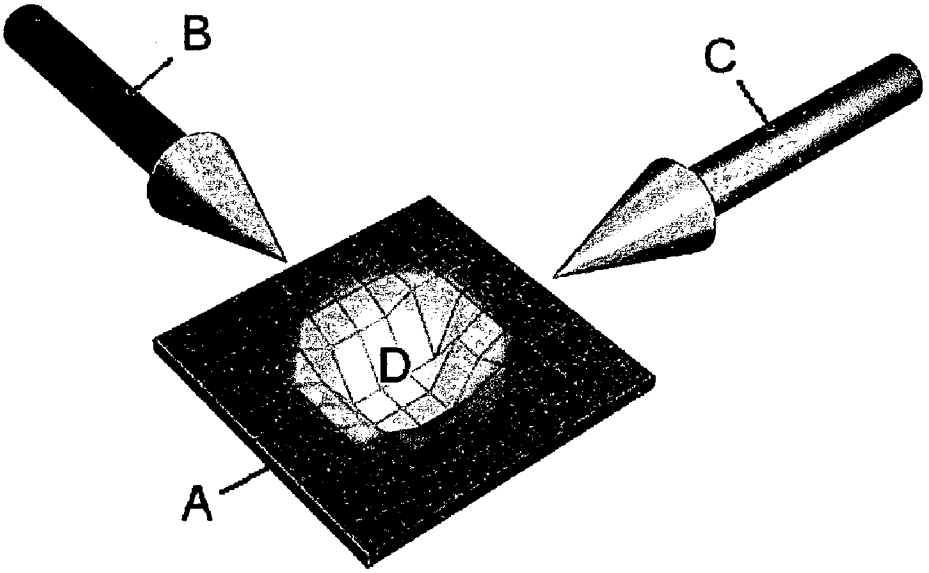

Figure 12

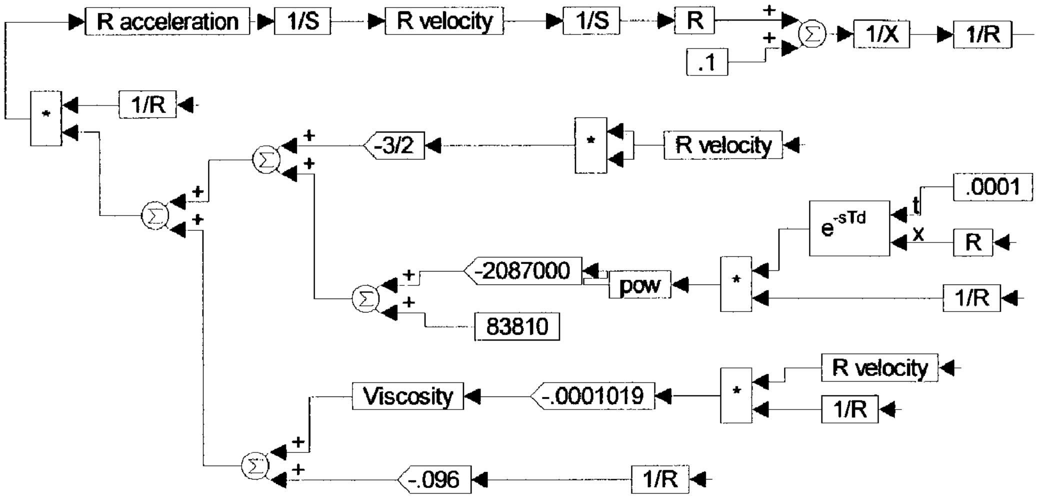
R acceleration
1/S
R velocity
1/S
R
+
+
.1
Σ
1/X
1/R
*
1/R
Σ
+
+
-3/2
*
R velocity
Σ
+
+
.0001
e^{-sTd}
t
x
R
-2087000
pow
*
1/R
Σ
+
+
83810
R velocity
Viscosity
-.0001019
*
1/R
Σ
+
+
-.096
1/R

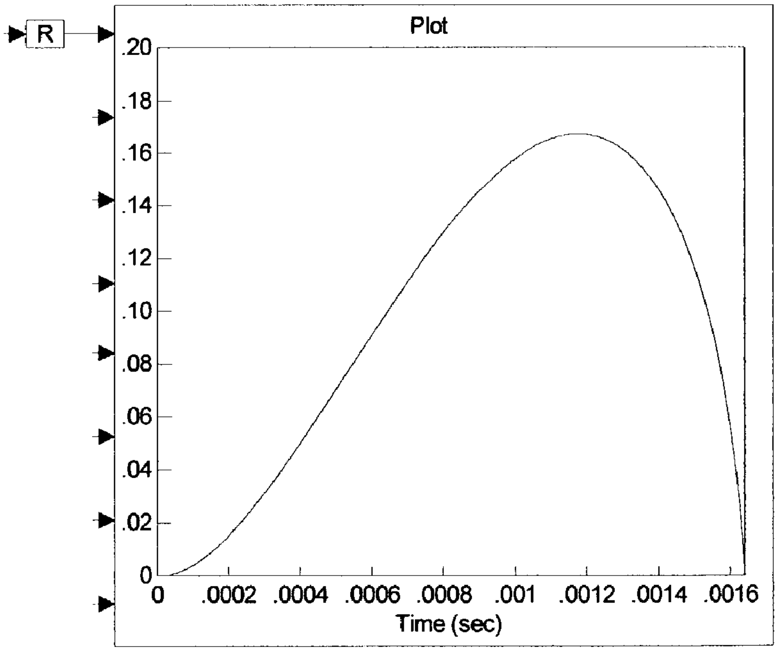
R
Plot
.20
.18
.16
.14
.12
.10
.08
.06
.04
.02
0
0
.0002
.0004
.0006
.0008
.001
.0012
.0014
.0016
Time (sec)

Figure 13

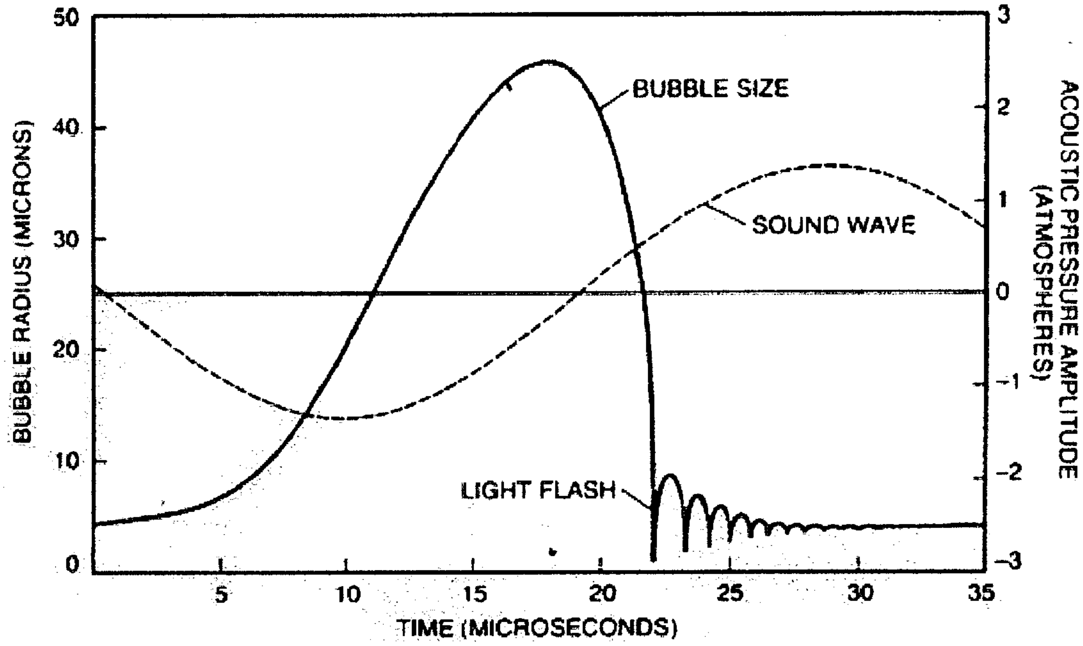
BUBBLE SIZE
SOUND WAVE
LIGHT FLASH
BUBBLE RADIUS (MICRONS)
50
40
30
20
10
0
ACOUSTIC PRESSURE AMPLITUDE (ATMOSPHERES)
3
2
1
0
−1
−2
−3
5
10
15
20
25
30
35
TIME (MICROSECONDS)

Figure 14

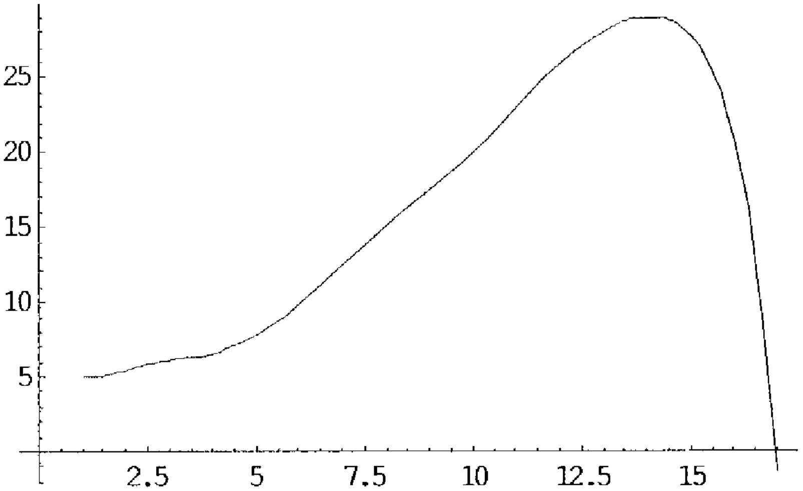

Figure 15

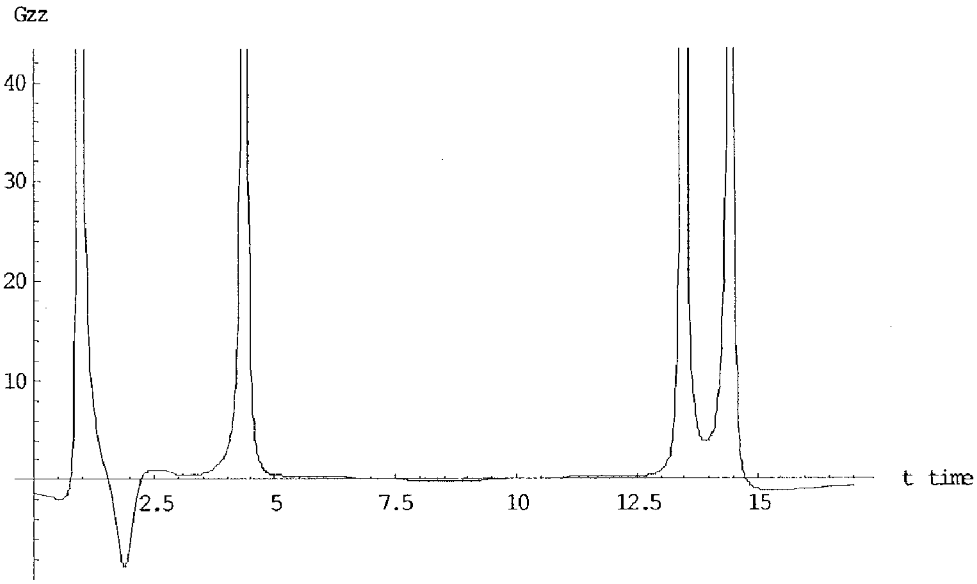

Figure 16

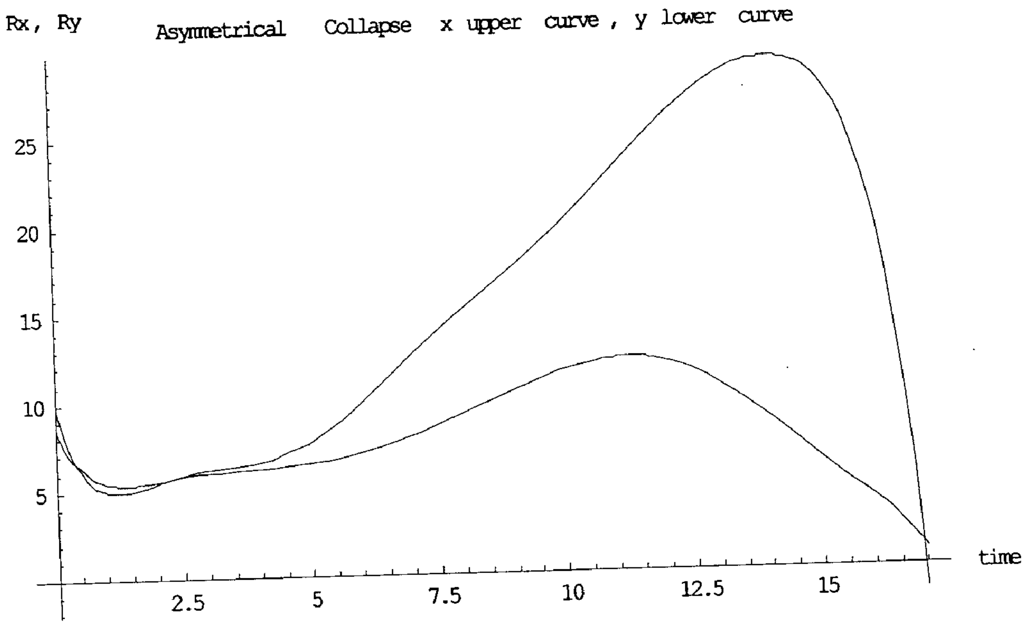
Rx, Ry
Asymmetrical Collapse x upper curve , y lower curve
25
20
15
10
5
2.5
5
7.5
10
12.5
15
time

Figure 17

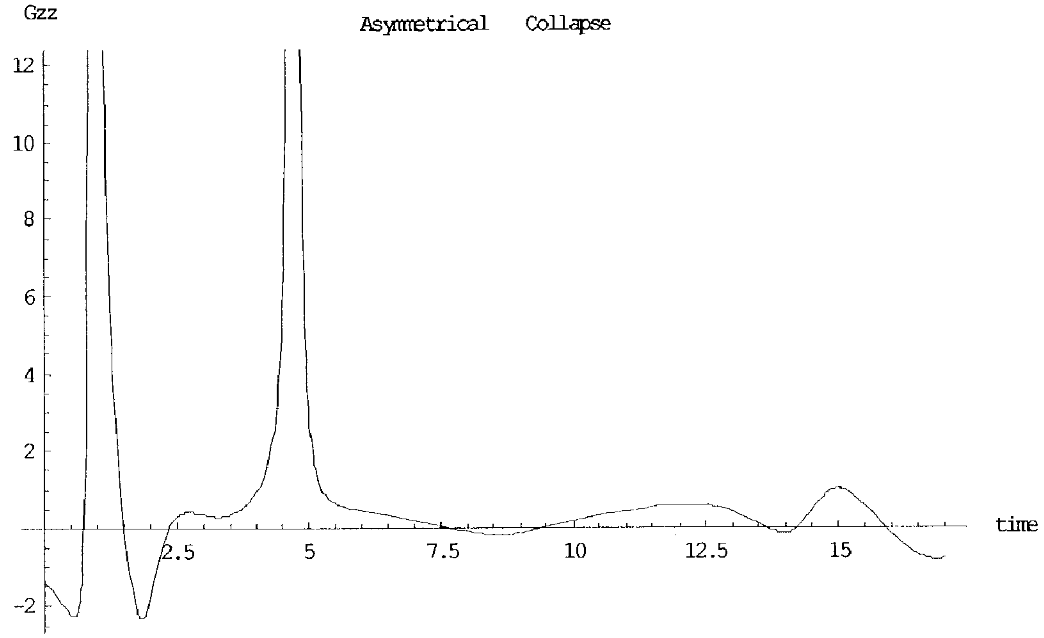

Figure 18

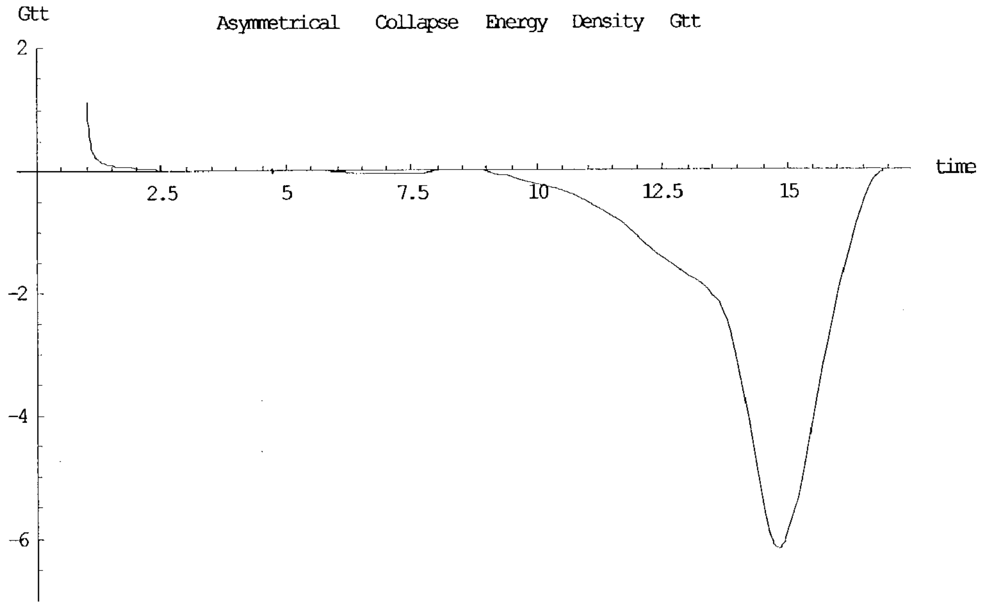
Gtt
Asymmetrical Collapse Energy Density Gtt
2
-2
-4
-6
2.5
5
7.5
10
12.5
15
time

Figure 19

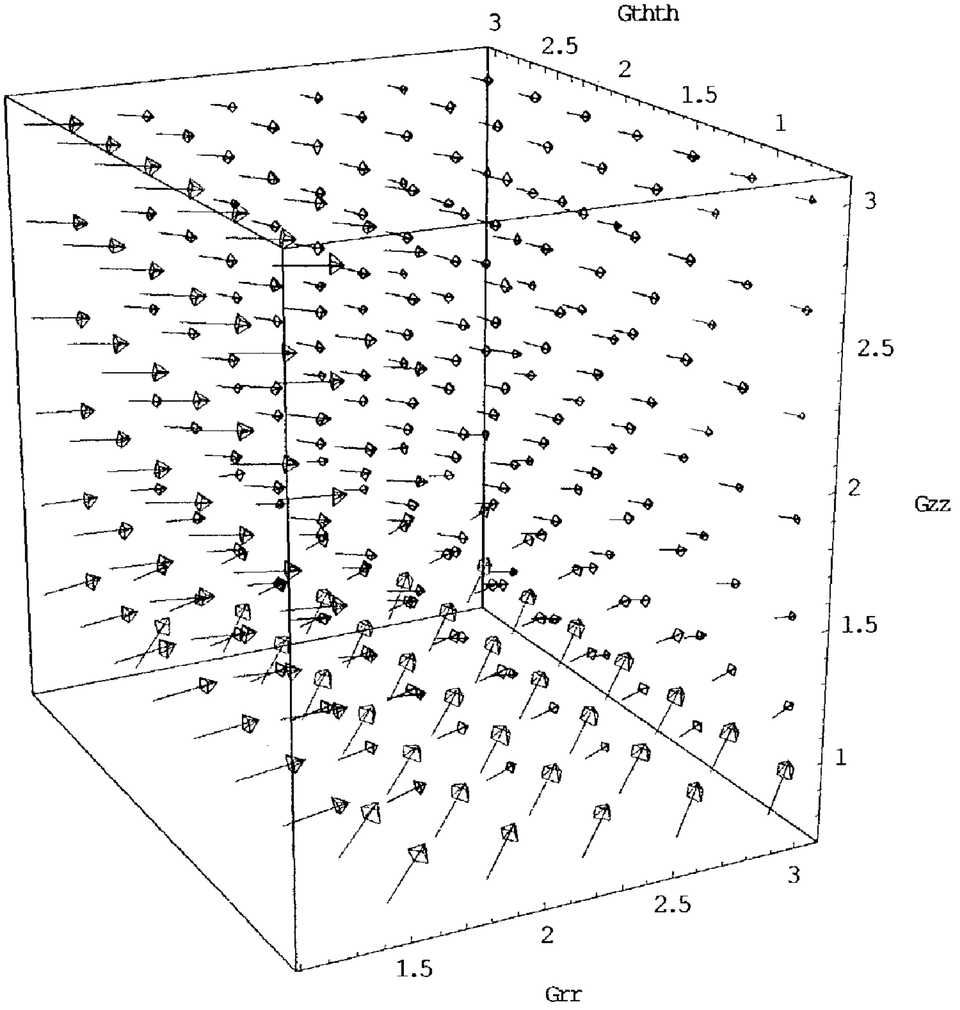

Figure 20

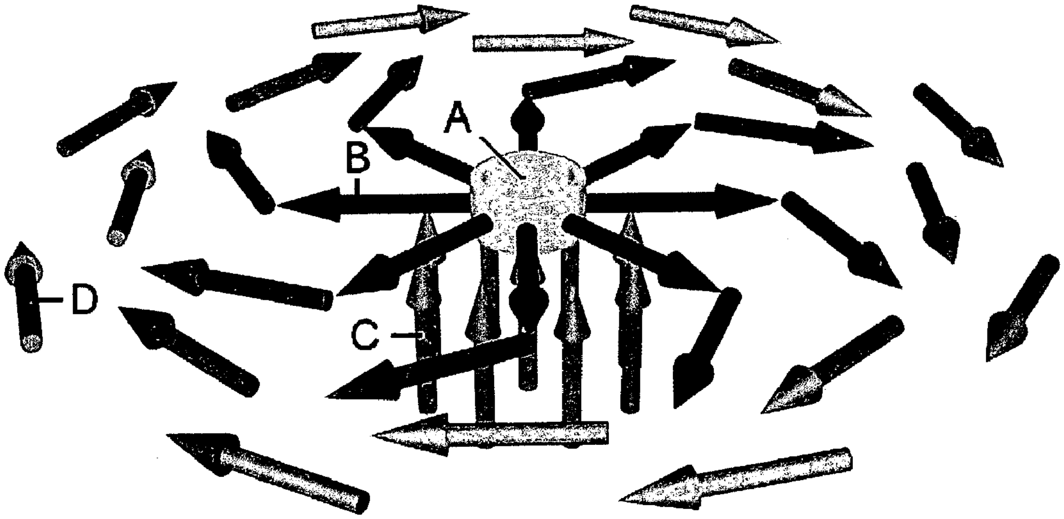

Figure 21

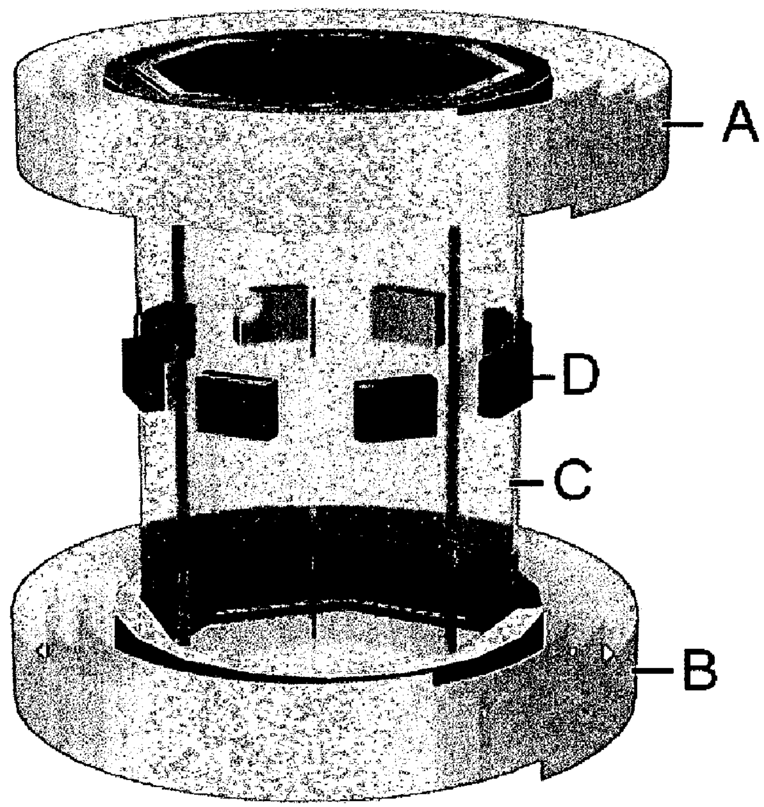

Figure 22

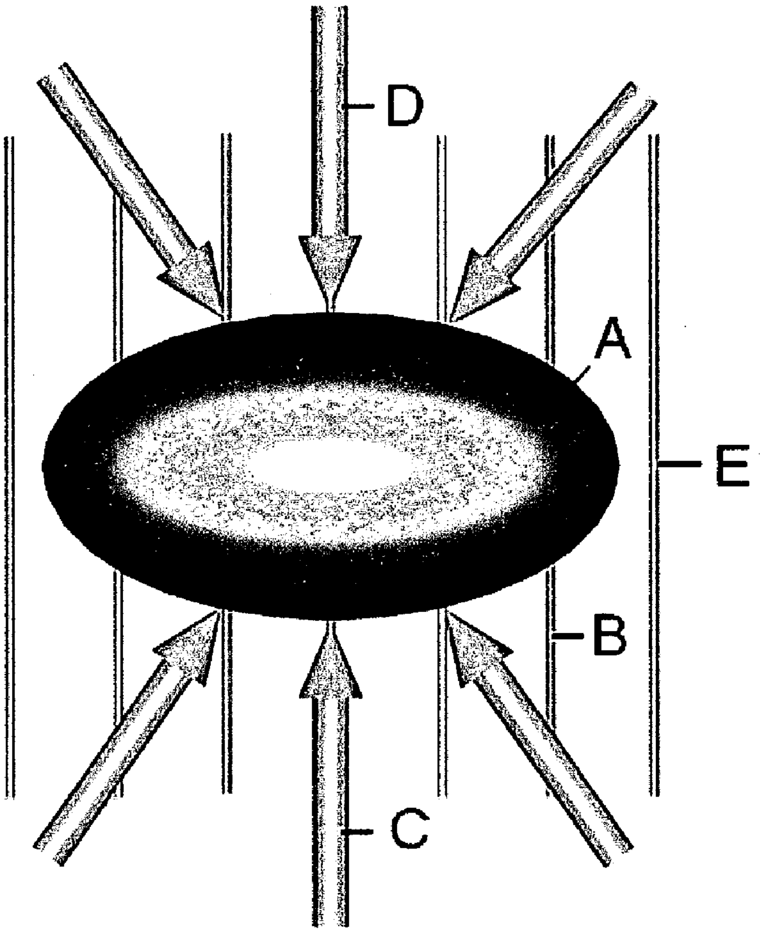

Figure 23

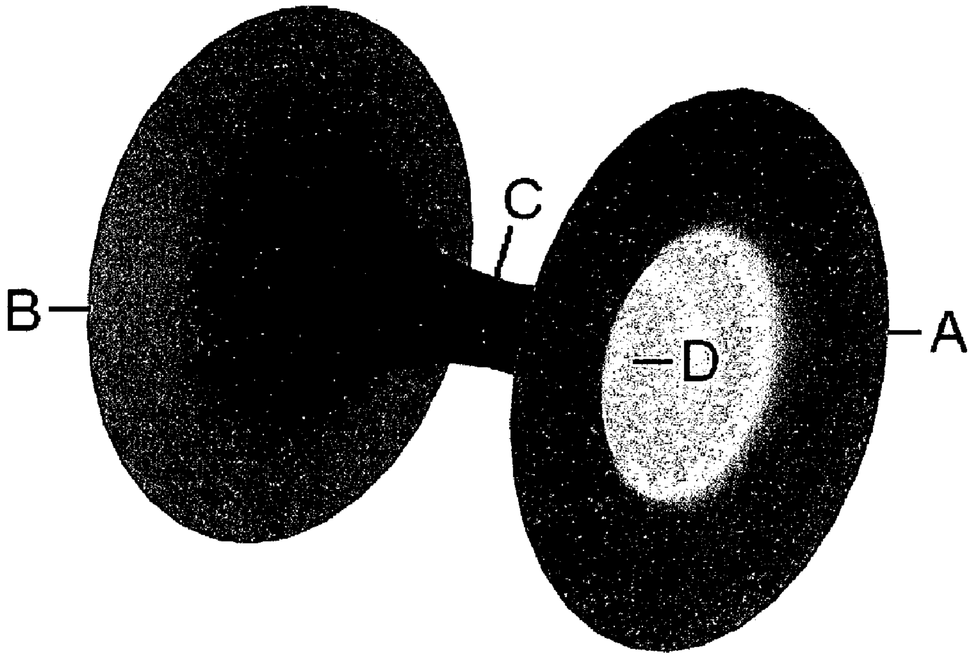

Figure 24

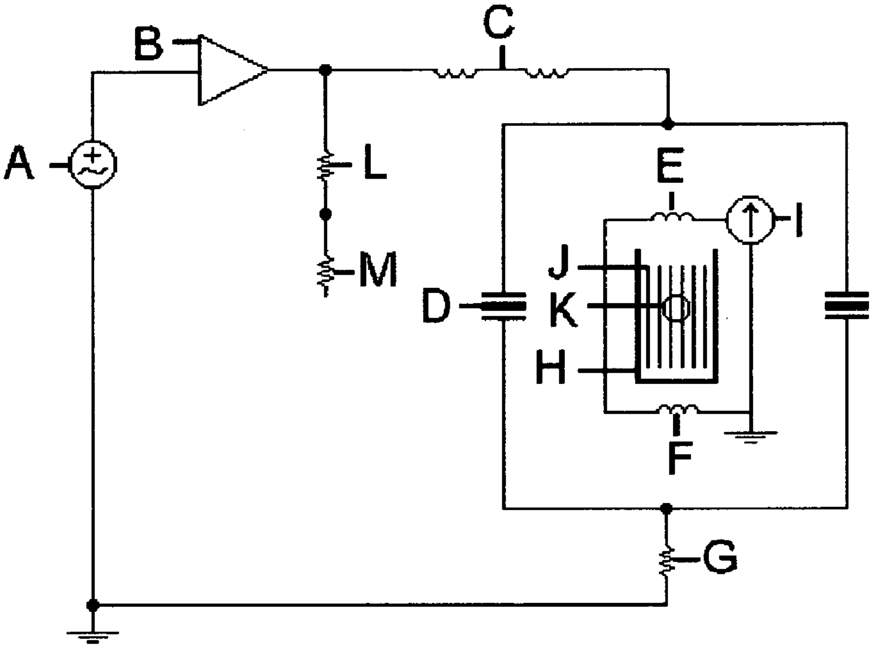
A
B
C
L
M
E
I
J
K
H
D
F
G

CAVITATING OIL HYPERSPACE ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a hyperspace energy generator that uses cavitating oil bubbles within a magnetic field in order to create wormholes between space and hyperspace for the purpose of permeating the hull of a spacecraft with low-density, low-speed-of-light hyperspace energy.

BACKGROUND OF THE INVENTION

[0002] As shown by physicist Dr. John Archibald Wheeler of Princeton University, there is a second term in Newton's gravitational equation which includes an additional force depending on whether or not the body is spinning.

$$F=mg+m(v\times\Omega)$$

[0003] where F is the force, m the mass, g the acceleration of gravity, v the forward velocity of the body, and Ω the angular velocity of the body. Referring to **FIG. 1**, a rotating cylinder (A) is moving with a velocity (B) while spinning counterclockwise as shown by the angular velocity vector (C). The velocity crossed with the angular velocity vector produces a force in the upward direction (D).

[0004] Referring to **FIG. 2**, a pendulum (A) held in the left hand over the right hand will spin in the counterclockwise direction due to the hyperspace energy vortex (B) emanating from the palm of the right hand. This vortex arises because human beings are hyperspace energy beings that live in physical containers located in this universe. The vortex might be considered a wormhole between space and hyperspace. Notice that the angular momentum vector of the vortex points up out of the palm of the hand. If the hand were spinning from left to right, then there would be a force exerted, as shown by the above equation, that would bring hyperspace energy into the right arm.

[0005] Referring to **FIG. 3**, in order to test this hypothesis, a Chakra Vortex Accelerator was built for the purpose of spinning this hand vortex. The motor-driven machine has a base (A) on which is mounted a warehouse palletizing ring and platform (B). A control column (C) has a speed control and on/off switch. A person (D) stands on the platform with his right arm outstretched and palm facing upward (E). The platform spins from left to right.

[0006] Referring to **FIG. 4**, the velocity vector (A) points to the right. The angular momentum vector of the rotating chakra energy vortex points up. The cross product A×B=C produces a force that brings in high permeability hyperspace energy into the arm.

[0007] Referring to **FIG. 5**, after spinning 99 times per day for a few months, it was noticed that the right arm (A) would remain out at an angle ϕ from the body without making any exertion to do so. The left arm, which was not receiving the energy, remained at the side of the body. That is, the low-density hyperspace energy in the right arm was making the arm lighter and producing a subsequent anti-gravitational effect. The conclusion was that it might be possible to invent some device which would produce large amounts of hyperspace energy. This would make possible relativistic electromagnetic fields which can be used to produce the lift force for electromagnetic field propulsion vehicles.

[0008] Then in the February 1995 issue of *Scientific American*, an article appeared entitled *Producing Light from a Bubble of Air* which gave the details for producing sonoluminescence from a cavitating bubble using a kit for sale containing the piezoelectric transducers. Referring to **FIG. 6**, the experimental setup includes a ring stand (A) with a three-finger clamp (B) holding a 100 milliliter flask containing degassed water produced by a Walsh vacuum/pressure pump. Piezoelectric drive transducers (D) are mounted on the sides of the flask in order to create sound pressure waves which create the cavitation bubble (M) in the center of the flask. The transducers are connected by insulated wire (I) to a pair of inductors (E) having a mutual inductance between them. These inductors are driven by a frequency generator (H) connected to an audio amplifier (G). The oscillation signals are picked up by the oscilloscope (J) which is connected to the resistors (F) using the probes (K). The piezoelectric drive transducers act electrically as capacitors. The inductors are wired in series with the capacitance such that the combination resonates at the frequency at which acoustic resonance occurs. Bubbles introduced into the flask tend to coalesce in the center of the flask Above a certain sound intensity, the bubble will collapse and emit a dim light visible to the unaided eye in a darkened room, a process called sonoluminescence.

[0009] Further research showed that there were some earlier doctoral theses available. One of them by Dr. Steven DeWitt Horsburgh entitled *Radial Instabilities of a pulsating air bubble in water* mentioned a most interesting observed phenomenon of the surface oscillations called a "beaconing effect." When a bubble exhibited a stable or long-lived shape oscillation lasting more than two minutes, the bubble had a tendency to slowly precess. The precession rate was on the order of a few hertz which can be taken to mean around 3 Hz. Visually, the beacon looked like a light house beacon. It was a very obvious effect and when the beacon crossed the photodiode, the output voltage increased by at least an order of magnitude.

[0010] Upon reading this, it became apparent that sonoluminescence is involved with gravitational hyperspace physics. Einstein's theory says that inertia is a manifestation of the geometry of spacetime. It also says that geometry is affected by the presence of matter to an extent proportional to the factor G/c^2 where G is the gravitational constant and c is the speed of light. The inverse of this factor is the linear mass Ω of the universe. A spherical shell of mass m with a radius R will create a drag angular velocity on a pendulum equal to

$$\omega_{drag}=\frac{4}{3}\frac{G}{c^2}\frac{m_{shell}}{R_{shell}}\omega_{shell}=\frac{4}{3}\frac{\frac{m_{shell}}{\Omega}}{R_{shell}}\omega_{shell}$$

[0011] The mass of the shell divided by the linear mass (mass per length) converts the mass into a length which is then divided by the radius of the shell which is also a length. The numerical factor 4/3 is found only by a detailed calculation. The calculation, starting from a flat background spacetime manifold, showed the effect of the moving current of mass on the metric. Expressed in polar coordinates, the metric acquires a non-zero coefficient $g_{\phi t}$. Inserted into the equation for geodesic motion, this off-diagonal metric coefficient gives rise to a precession according to the above equation. The numerical factor corresponds to a pendulum located anywhere inside the rotating shell of mass.

[0012] In other doctoral theses, books and physics articles on cavitation, light is emitted after the bubble starts to collapse. A very sharp peak of light is produced which lasts in terms of picoseconds. Thus this type of sonoluminescence is very different than the beaconing effect where the bubble remains stable for long periods of time. The beaconing effect is actually a rotating wormhole between space and hyperspace.

[0013] Briefly what happens is that there is ionized gas in the bubble which releases electrons. As the electrons spiral down through the wormhole from space into hyperspace, they give off light. This situation puts the electrons in contact with the charge of hyperspace. Space has a charge which is equal to the sum of the electron charge q_e and the hyperspace charge q_h. The charge of space q is equal to the Planck mass times a conversion factor between mass and charge. The Planck mass is equal to the linear mass Ω of the universe times the bottom dimensional limit of the universe known as the Planck scale Λ. The conversion factor is the square root of the gravitational constant G times the permittivity of space ϵ. The permittivity is linear capacitance or capacitance per unit length. Thus the charge of space is

$$q=\Lambda\Omega\sqrt{G\epsilon}=q_e+q_h$$

[0014] Because the electron comes in contact with the hyperspace charge, the drag angular velocity in the above equation has to be modified. Notice that if the angular velocity of the shell is modified by the electron fine structure constant α, and the mass involved is the mass of the electron, moving through a radius into hyperspace of the Planck scale Λ, then the drag angular velocity would be

$$\omega_{drag}=\frac{4}{3}\frac{\frac{m_{electron}}{\Omega}}{\Lambda}\omega_{shell}\frac{1}{\alpha}$$

[0015] The angular frequency of the shell is 2π times the electron frequency f_e.

$$\omega_{shell}=\omega_{electron}=2\pi f_e$$

[0016] The inverse of the electron fine structure constant α is equal to Planck's reduced constant ℏ times the speed of light c times the factor 4π times the permittivity of space ϵ divided by the square of the charge of the electron q_e.

$$\frac{1}{\alpha}=\frac{\hbar\, c\, 4\pi\varepsilon}{q_e^2}n=137.037916527$$

[0017] Planck's constant is equal to the Planck scale Λ squared times the linear mass Ω times the speed of light.

$$\hbar=\Lambda^2\Omega c$$

[0018] Notice that the fine structure constant has a square in it and Planck's constant also has a square. If the electron charge were traveling through the Planck scale into hyperspace, then there would be a linear charge equal to the electron charge divided by the Planck scale

$$\frac{1}{\alpha}=\frac{\hbar\, c\, 4\pi\varepsilon}{q_e^2}n=\frac{\Lambda^2\Omega\, c\, c\, 4\pi\varepsilon}{q_e^2}n=\frac{4\pi\Omega\,\varepsilon\, c^2}{\left(\frac{q_e}{\Lambda}\right)^2}n$$

[0019] The constant equal to the linear mass Ω times the permittivity times the square of the speed of light is actually the square of the linear charge χ of space. So the fine structure constant can be written

$$\frac{1}{\alpha}=4\pi n\frac{(\chi)^2}{\left(\frac{q_e}{\Lambda}\right)^2}=4\pi n\frac{\left(\frac{q}{\Lambda}\right)^2}{\left(\frac{q_e}{\Lambda}\right)^2}=4\pi n\left(\frac{q}{q_e}\right)^2=4\pi n\left(\frac{q_h+q_e}{q_e}\right)^2=$$
$$4\pi n\left(\frac{q_h}{q_e}+1\right)$$

[0020] So the electron fine structure constant is related to the ratio of the charge of hyperspace to the charge of the electron, which is the reason that the equation had to be modified. The electron moving across the Planck scale wormhole is modified by the hyperspace charge it is seeing in a hyperspace co-dimension. This confirms that the beaconing effect is a rotating wormhole into hyperspace.

[0021] The electron fine structure constant is actually the fastest velocity in the Bohr atom. The fine structure constant is the ratio of the electron orbital velocity to the speed of light.

$$\alpha=\frac{v_n}{c}$$

[0022] where the permitted quantized n orbital speeds are v_n/n or v/1, v/2, v/3 and so on. Using a value of v/3, the fine structure constant would be

$$\frac{1}{\alpha}=\frac{3c}{v}=4\pi n\left(\frac{q_h}{q_e}+1\right)=12\pi\left(\frac{q_h}{q_e}+1\right)$$

[0023] which when substituted into the drag angular frequency equation

$$\omega_{drag}=\frac{4}{3}\frac{\frac{m_e}{\Omega}}{\Lambda}2\pi f_e\,12\pi\left(\frac{q_h}{q_e}+1\right)=17.809\ \frac{\text{radians}}{\text{sec}}$$

[0024] The frequency of precession is

$$f_{drag}=\frac{\omega_{drag}}{2\pi}=2.83\text{ Hz}$$

[0025] which is close to the 3 Hz observed in the experiment. The ionized electrons are moving at a slower velocity in the bubble. The 1/3 ratio could also come from the subspace tetrahedron geometry where the tetrahedral angle is determined by this ratio

$$\theta=\operatorname{asin}(1/3)=19.47122063^\circ$$

[0026] This is an alternative possibility because this angle determines the charge of hyperspace according to the tetrahedron diagram of Δphysics. In geometrical physics, all the constants of physics are determined geometrically by the tetrahedron. The tetrahedron diagram plots the natural logarithm of mass versus the natural logarithm of wavelength. The tetrahedron is circumscribed by a sphere which reflects the electron wavelength into the electron mass. Thus the diagram combines the dual reality of classical physics, involving point mass particles, and quantum physics, involving wave-like particles. The scientific discovery that the electron and the proton are one and the same particle was made using this diagram. It also shows that hyperspace exists.

[0027] Referring to the tetrahedron diagram in **FIG. 7**, a line drawn at the tetrahedral angle produces the tetrahedral lines (A,B) along path (abc). This tetrahedron is circumscribed by sphere (C) with diameter (G). The electron wavelength (vertical dotted line D) reflects off the sphere (d,f) and returns as the electron mass (F). The distance (df) between reflection points is equal to the natural logarithm of the hyperspace charge. It can be shown that the electron charge is related to the geometry around the top of the sphere. Part of the geometry is outside the sphere which makes the charge immune to relativistic velocities.

[0028] Referring to **FIG. 8**, a current-carrying solenoid (A) produces a magnetic flux (B) through the center of the coil. The end of the solenoid behaves like a magnetic monopole. One end has a positive magnetic charge density $+\rho_m$ and the other has a negative magnetic charge density $-\rho_m$. One of Maxwell's electromagnetic equations states that the divergence of the magnetic field is equal to the magnetic charge density.

$$\nabla\cdot B=\rho_m=+\rho_m-\rho_m=0$$

[0029] Because the magnetic field is solenoidal, the sum of the charges in our spacetime is equal to zero, which is the usual Maxwell equation. In the case of the wormhole, one pole of the magnetic flux is in our spacetime and the other pole is in a co-dimension of hyperspace. In our universe, there is a net magnetic charge density, and Maxwell's equation becomes

$$\nabla\cdot B=\rho_m$$

[0030] When working with bar magnets, bringing two north poles together shows that there is a spring constant involved. The electrons on the surface of the pole spin in the same direction which creates an electrical current around the surface. The magnetic pole strength g is therefore a spring constant K divided by the current per area J.

$$g=\frac{K}{J}\equiv\frac{kg\,m}{s^2}\frac{1}{m}\frac{s\,m^2}{q}=\frac{kg\,m^2}{s\,q}$$

[0031] The magnetic charge density is then the pole strength per volume V or

$$\rho_m=\frac{g}{V}=\frac{kg\,m^2}{s\,q}\frac{1}{m^3}=\frac{kg}{m\,s\,q}$$

[0032] The divergence of the magnetic field has units of

$$\nabla\cdot B=\frac{1}{m}\frac{kg}{s\,q}$$

[0033] which are the units of the magnetic charge density. Notice that the units can also be formatted as a pressure per electrical current in the wire.

$$\rho_m=\frac{kg}{s\,q\,m}=\frac{kg\,m}{s^2\frac{q}{s}m^2}=\frac{\text{force}}{\text{area}}\frac{1}{I}=\frac{\text{pressure}}{\text{current}}$$

[0034] The current flowing through the coils of the solenoid, or around the pole face of the bar magnetic creates a pressure on spacetime. And it is this pressure that can punch through spacetime into hyperspace. Referring again to **FIG. 8**, an electrical charge winding n times around the end of the spacetime solenoid, as seen by path (C), is equal to the charge divided by Planck's constant times the pole flux

$$2\pi n=\frac{q}{\hbar}\int_{Surface}B=\frac{q}{\hbar}\Phi$$

[0035] which shows that the flux Φ is quantized by the winding number n

$$\Phi=n\frac{2\pi\hbar}{q}=n\frac{h}{q}$$

[0036] where the flux is equal to the winding number n times Planck's constant h divided by the charge. Since the flux is the pole strength in webers, the pole strength g has a value of

$$g=n\frac{h}{q}$$

[0037] The magnetic charge density is the pole strength divided by the Planck volume or

$$\rho_m=\frac{g}{\Lambda^3}=n\frac{h}{q}\frac{1}{\Lambda^3}=n\frac{2\pi\Lambda^2\Omega c}{q\Lambda^3}=\frac{2\pi n}{\Lambda}\frac{\Omega c}{q}=2\pi n\frac{B}{\Lambda}$$

[0038] The magnetic B field of space divided by the Planck length of the wormhole, acting as a linear magnetic field, times the winding number is equal to the magnetic charge density.

[0039] Referring to **FIG. 9**, the wormhole (A), connecting space and hyperspace (D), is precessing (C) in the counterclockwise direction with the angular velocity vector in the vertical z-direction. The magnetic flux (B) through the wormhole is in the radial direction. The mouth of the wormhole is an area with a normal vector in the radial direction also. This arrangement creates an electric field which circulates around the interior periphery of the wormhole. Maxwell's equation says that the curl of the electric field E is equal to the negative time rate of change of the magnetic B field.

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

[0040] This equation is difficult to manage in differential form. As both Maxwell and Feynman have said, it is better to use the integral form involving Stokes' theorem where the area da is converted into a line integral ds which goes well with Maxwell's idea of flux linkage.

$$\int_{surface} (\nabla \times E) \cdot n\, da = \oint E \cdot ds = -\frac{\partial}{\partial t} \int_{surface} B \cdot n\, da = -\frac{\partial}{\partial t}\Phi = -\Phi\, \omega_{drag}$$

[0041] The flux Φ, which is precessing around at the drag angular velocity calculated previously, is equal to the negative of the electric field around the interior perimeter (E) of the wormhole. This is the reason that the beaconing effect lasts for such a long length of time. The force on the electron is equal to the charge of the electron times the electric field. Because the charge is negative, the electron rotates counterclockwise in the opposite direction to that of the electric field. The electrons continuously circulate around the inside periphery of the wormhole. Quantum theory predicts that any radiated electromagnetic energy from an accelerated electron will give off one or more discrete quanta, or photons. The electron is accelerating because it is moving in a circular path. Because the water vapor is ionized in the cavitating bubble due to shock waves, surface oscillations and high temperatures, there are huge numbers of electrons available for producing an intense light which, as registered by the photodiode, increased by an order of magnitude when the beacon rotated by the photodiode.

[0042] From gravitational physics it is known that negative energy has to be produced in order to create a wormhole. Most experimenters are using collapsing spherical bubbles that do not produce negative energy and therefore do not produce the beaconing effect. This next background section shows that if the bubble collapses asymmetrically, then negative energy is produced.

[0043] Referring to an experiment carried out by Fizeau, **FIG. 10**, water flows (B,C) in opposite directions in two tubes of water (A). A light beam from a lamp (D) is projected through both tubes using half-silvered mirrors (E). Fizeau measured an interference (F) between light propagating with the flow and light swimming against the current, indicating that the flow of the medium does affect light propagation. A moving medium turns out to drag light to an extent quantified by a dragging coefficient

$$\alpha = 1 - \frac{1}{n^2}$$

[0044] where n is the index of refraction of the medium. Einstein's addition theorem of velocities states

$$v = \frac{v' + u}{1 + \frac{v'u}{c^2}} = \frac{\frac{c}{n} + u}{1 + \frac{cu}{nc^2}} = \frac{\frac{c}{n} + u}{1 + \frac{u}{nc}}$$

[0045] In the limit of slow flows compared with the speed of light in a vacuum, the effective velocity of light v in the medium flowing at a velocity u is

$$v \approx \left(\frac{c}{n} + u\right)\left(1 - \frac{u}{nc}\right) \approx \frac{c}{n} + \alpha u$$

[0046] The relationship between cavitating bubbles and light dragging is that when the bubble collapses, the surface has a tremendous acceleration and velocity which creates internal shockwaves that bounce back and forth within the medium. At the same time, the water vapor and air are ionized which creates short bursts of light. Thus there is light that is being dragged along by the collapsing surface of the bubble. This moving medium appears to light as a change in the metric of spacetime resembling, it turns out, a gravitational field. The velocity of the medium affects the elemental length ds^2 of spacetime as given by the metric for light dragging

$$ds^2 = c^2dt^2 - dx^2 + k(c^2dt^2 - u \cdot dx)^2$$

[0047] where k is equal to

$$k = \frac{\alpha}{c^2 - u^2}$$

[0048] and dx is the elemental length in the {dx,dy,dz} directions. The velocity u in the dot product is equal to the velocities {ux,uy,uz} in the x, y and z-directions. Expanding the equation produces the following metric

$$ds^2 = (1+k)dt^2 + (ku_x^2 - 1)dx^2 + (ku_y^2 - 1)dy^2 - dz^2 - 2ku_x dxdt - 2ku_y dydt + 2ku_x u_y dxdy$$

[0049] where velocity uz is zero and the speed of light is equal to one.

[0050] The dt^2 energy density term is modified by one plus the k factor. Thus the index of refraction of the medium, the speed of light and the overall light dragging velocity u affects the energy density. It can be shown that the speed of light in hyperspace is much lower. The difference of the squares of the velocities in the denominator creates the possibility for amplification of this term.

[0051] The factors of the elemental lengths are then inserted into the g metric tensor which is then used to calculate the spacetime distortion generated by the collapsing bubble. The g metric tensor is a 4×4 matrix having rows and columns that correspond to the elemental lengths {t, x, y, z}.

$$g_{\alpha\beta} = \begin{array}{c|cccc} & \beta = t & \beta = x & \beta = y & \beta = z \\ \alpha = t & 1+k & -ku_x & -ku_y & 0 \\ \alpha = x & -ku_x & ku_x^2 - 1 & ku_x u_y & 0 \\ \alpha = y & -ku_y & ku_x u_y & ku_z^2 - 1 & 0 \\ \alpha = z & 0 & 0 & 0 & -1 \end{array}$$

[0052] In Einstein's General Theory of Relativity, the G curvature tensor, which is a 4×4 matrix having units of inverse meter squared, is equal to a stress-energy-momentum T tensor that is calculated from the electromagnetic fields. The stress tensor T combines energy density, electromagnetic flux, and pressure terms in one matrix.

$$G=8\pi T$$

[0053] Most of the physics constants are linear constants such as linear mass which is mass per unit length. The permittivity of space ϵ is linear capacitance or capacitance per meter. The permeability of space μ is linear inductance or inductance per meter. The speed of light converts time into meters. What this means is that energy, pressure and momentum can all be converted into meters using these linear constants. The inverse squared of the length in meters is the curvature which is measured in units of m^{-2}.

[0054] Referring to **FIG. 11**, a square bubble (A) in the xy plane collapses with a velocity in the x-direction (C), and a velocity in the y-direction (B). This produces a curvature in the vertical z-direction (D). The curvature in the z-direction is the curvature tensor component G_{zz}. which can be computed directly from the g light dragging metric tensor. The result of the calculation is an equation in terms of the velocity, acceleration and the time rate of change of the acceleration, known as jerk.

[0055] Because the profile of the bubble radius expanding and collapsing with time has been published by those researching cavitation, it is possible to differentiate the profile to get the velocity, acceleration and jerk. These equations can then be inserted into the equation for the G curvature tensor.

[0056] The radius R of the bubble depends on the air density inside the bubble ρ, the viscosity of the water μ, the pressure of the P of the environment, the surface tension σ of the water, the ratio of specific heats γ, and the velocity $\dot{R}$ and acceleration $\ddot{R}$ of the bubble radius.

$$R\ddot{R} + \frac{3}{2}\dot{R} + \frac{4\mu}{\rho}\frac{\dot{R}}{R} + \frac{2\sigma}{\rho R} + \frac{P_0}{\rho} - \left(P_0 + \frac{2\sigma}{R_0}\right)\left(\frac{R_0}{R}\right)^{3\gamma} = 0$$

[0057] Referring to **FIG. 12**, the above equation for the bubble radius is shown in the feedback simulation program. The diagram starts by solving the equation for the acceleration which is then integrated twice (1/S) to get the velocity and position. These variables are then fed back to the input through the various constants involving the viscosity, surface tension, pressure and air density. The graph at the bottom of the simulation shows the bubble radius with time.

[0058] Referring to **FIG. 13**, the bubble profile with time is reproduced from the *Scientific American* article mentioned previously. This profile was then fitted with a polynomial equation which is plotted in **FIG. 14**. By differentiating the equation, the velocity, acceleration and jerk are obtained which can then be inserted into Einstein's G curvature tensor.

[0059] Referring to **FIG. 15**, the curvature component in the vertical direction G_{zz} is plotted with respect to time. In the *Scientific American* plot, the light flash occurs at the end of the collapse marked "LIGHT FLASH" around 22 microseconds (**FIG. 13**). This corresponds to the two intense positive curvature G_{zz} spikes at the end of the collapse. This represents a symmetrical spherical bubble collapse where the velocities are the same in the x and y directions.

[0060] The energy density component G_{tt} for this symmetrical collapse is given by

$$G_{tt} = \frac{-2u_y^2 a_x^2 + 4u_x u_y a_x a_y - 2u_x^2 a_y^2}{4(u_x^2 + u_y^2 - 2)}$$

[0061] Looking at this carefully, notice that if the velocity in the x-direction u_x is equal to the velocity in the y-direction u_y, and the acceleration in the x-direction a_x is equal to the acceleration in the y-direction, then the equation reduces to zero.

$$G_{tt} = \frac{-2u^2 a^2 + 4u^2 a^2 - 2u^2 a^2}{8(u^2 - 1)} = 0$$

[0062] which means that the symmetrical collapse with equal velocities does not produce any negative energy with which to create the wormhole. Only sharp pulses of light are emitted at the end of the bubble collapse where the intense positive curvature spikes are encountered. In order to get the rotating wormhole beaconing effect, the velocities have to be different.

[0063] Referring to **FIG. 16**, the bubble is given an asymmetrical profile as shown by the two curves. The upper curve is in the x-direction, and the lower curve is in the y-direction. The collapse in the y-direction is not as severe.

[0064] Referring to **FIG. 17**, the asymmetrical collapse has the curvature spikes at the beginning similar to the symmetrical case of **FIG. 15**. Notice that there are no spikes at the end of the collapse. However, the energy density is much different.

[0065] Referring to **FIG. 18**, the energy density G_{tt} is plotted as a function of time. Rather than being zero as in the symmetrical case, there is a large region of negative energy density as the bubble collapses. This negative energy is what creates the wormhole. Because there are no large curvature spikes at the end of the collapse, the wormhole bubble is stable and precesses for a few minutes.

[0066] Experiments with the pendulum show that some people have much larger auras than others. The pendulum swings over the entire hand. One particular person who has enormous vortices was working in a television repair shop. When a television set is brought in to the shop, the large capacitor has to be discharged first. The repairman with the large chakra vortices was standing behind a second repairman who was approaching the television set. As the second repairman got closer to the set, an enormous streaming blue spark sailed over his head and zapped the first technician. Hyperspace has a much higher permittivity than our spacetime. The resistance of space R is equal to the square root of the permeability μ over the permittivity ϵ.

$$R = \sqrt{\frac{\mu}{\varepsilon}}$$

[0067] Since the technician with the large chakra vortices was producing large quantities of hyperspace energy, his resistance was very much lower due to his larger permittivity. The spark grounded on his body. The free electrons, which are created by ionization of the hydrogen atoms of the water vapor, see a very low path of resistance and flow toward the wormhole. There they are trapped in accelerated motion due to the presence of the circular electric field and give off substantial numbers of photons which produces the beaconing lighthouse effect.

[0068] The elemental length ds^2 in cylindrical coordinates $\{t, r, \theta, z\}$ for a mass M is given by

$$ds^2 = -\frac{dt^2}{1-\frac{M}{r}} + \left(1-\frac{M}{r}\right)dr^2 + r^2 d\theta^2 + dz^2$$

[0069] The mass M distorts the radius r of the elemental length which is what gives the mass a curvature or gravitational field. Negative energy, which is produced at the end of the bubble collapse, is equal to a negative mass times the speed of light squared. Thus mass M in the equation goes to negative mass, or M→-M. Furthermore, negative energy inverts the mass factor so that the metric becomes

$$ds^2 = -\left(1+\frac{M}{r}\right)dt^2 + \frac{dr^2}{1+\frac{M}{r}} + r^2 d\theta^2 + dz^2$$

[0070] The g metric tensor contains the factors of the elemental lengths along the matrix diagonal

$$g_{\alpha\beta} = \left\| \begin{matrix} -\left(1+\frac{M}{r}\right) & 0 & 0 & 0 \\ 0 & \frac{1}{1+\frac{M}{r}} & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{matrix} \right\|$$

[0071] As before, Einstein's G curvature tensor in the radial G_{rr}, angular $G_{\theta\theta}$, and vertical z-direction G_{zz} is calculated from this negative energy g metric tensor.

[0072] Referring to the 3D G curvature plot in **FIG. 19**, the strength of the curvature is given by the length of the arrow, and the direction of the curvature is given by the arrow. The curvature in the radial direction G_{rr} is along the bottom axis from the box corner origin to the right. The curvature in the angular direction $G_{\theta\theta}$ is the bottom axis from the origin to the left. The height of the box is the curvature in the vertical direction G_{zz}. Notice in general that the curvature is perpendicular to the left face (θz-plane) and then starts to curve to the right and begins to point toward the right face (rz-plane). On the bottom face (rθ-plane) there is an upward vertical curvature.

[0073] In cylindrical coordinates, the θz-plane is actually a vertical tube of radius r. So at small r, the curvature is perpendicular to the tube. At larger radius, the curvature starts to point in the θ-direction, wrapping around the tube. At small radius, there is also a vertical curvature parallel to the sides of the tube.

[0074] Referring to **FIG. 20**, the curvature in cylindrical coordinates is seen to be the structure of the wormhole. The wormhole (A) is formed due to the radial pressure (B) which pulls open the wormhole. One end of the tube connects to hyperspace, and the other end connects to our spacetime. There is also a shearing pressure (D) along the sides of the tube which slices spacetime open and helps shape the tube. Moving outward along the radius, the curvature turns sideways and produces a shearing pressure (D) in the θ-direction which gives the wormhole a vortex appearance. These wormholes can also be very large, enabling one to peer into a co-dimension as attested to by the inventor.

SUMMARY OF THE INVENTION

[0075] As described in the previous section, the cavitating bubble has to collapse asymmetrically in order to generate the negative energy which creates the wormhole. In order to collapse asymmetrically, the forces on the bubble surface have to be unequal. The force on the bubble is given in terms of the a dragging coefficient, the bubble radius R, the surface charge density D_r, the magnetic B field in spherical coordinates $\{B_r, B_\theta, B_\Phi\}$. The fields are integrated over the entire solid angle Ω of the bubble to determine the force F.

$$F_r = -\left(1-\frac{1}{n^2}\right)\frac{R^2}{2}\oint d\Omega\left[\left(1+\frac{1}{n^2}\right)(D_r^2 + B_r^2 - B_\theta^2 - B_\phi^2)\right]$$

[0076] The integration of the tangential components of the magnetic field such as $\{B_\theta, B_\phi\}$ over the solid angle of the bubble will sum to zero. The surface charge density D_r and the radial magnetic field B_r are the only fields which can be used to create an unbalanced force. If there is no surface charge density, then the force equation reduces to

$$F_r = -\alpha\frac{R^2}{2}\oint d\Omega\left[(2-\alpha)\left(\frac{\chi^2}{\Omega_m}B_r^2\right)\right]$$

[0077] where the conversion constant is the linear charge squared divided by the linear mass of the universe χ^2/Ω_m.

[0078] In this invention, referring to **FIG. 21**, the magnetic field is created by two cylindrical solenoids (A,B) which are wrapped around the top and bottom of the glass container (C). The container has flat sides in the shape of an octagon. Each solenoid consists of a continuous roll of very thin sheet copper which are separated by beaded mylar line. The mylar beads separate the sheets electrically and provide air cooling to the coil which can get rather hot. This makes it easy to assemble the coil because the copper roll and mylar line are wound together as the solenoid is formed. The lines of magnetic flux flow vertically through the container yet do not interfere with the transducers (D) which are mounted on the flat sides of the container.

[0079] Referring to **FIG. 22**, the cavitating bubble which forms in the liquid in the container experiences a vertical magnetic field (B) due to the two solenoids. These flux lines create a radial force on the top and bottom of the bubble (D,C). The flux lines on the outside (E) do not create any radial force because they are tangential to the surface. The spherical bubble flattens into an asymmetrical ellipsoidal shape (A) which enables the creation of the wormhole and lighthouse beaconing effect.

[0080] Referring to **FIG. 23**, the spacetime curvature of the wormhole (D) creates a tunnel (C), known as the throat of the wormhole, between flat space in our universe (A) and flat hyperspace (B) which exists in a co-dimension of our universe. In general, the co-dimensions have different physics constants. Going back to the chakra vortices of the human energy field, the vortices remove energy from space and deliver it to several hyperspace co-dimensions. This energy builds up human aura modules at different frequencies in astral space. Astral space has a very low speed of light and small linear mass. This is the reason that human beings can leave their bodies and pass through solid concrete walls, or project their spiritual eye to remote locations for viewing. On the other hand, the linear mass and speed of light of space is very much greater than hyperspace. Therefore, the pressure of the space is much greater than the pressure in hyperspace. In terms of the wormhole with a throat area A, the pressure would be the linear mass times the speed of light squared over the area

$$P_{space} = \frac{\Omega_s c_s^2}{A}$$

[0081] The pressure of hyperspace on the other side of the wormhole is

$$P_{hyperspace} = \frac{\Omega_h c_h^2}{A}$$

[0082] Because the linear mass and the speed of light of hyperspace are lower than that of space, there is a positive net pressure going into hyperspace.

$$P_{\text{net}} = P_{\text{space}} - P_{\text{hyperspace}}$$

[0083] This is the reason that the human vortices can obtain energy from this universe. Because the cavitating bubble produces negative energy, the linear mass is negative on the side of space. The pressure equation is then given by

$$P_{net} = -\frac{P_s}{A} - \frac{P_h}{A} = -\frac{P_s + P_h}{A}$$

[0084] which shows that there is a reverse pressure going from hyperspace into our space. The asymmetrical cavitating bubble generates hyperspace energy. The hyperspace energy flowing into our universe also gives the light generated by the wormhole a soft white misty look.

[0085] There are many reasons for generating this hyperspace energy. Electromagnetic fields are subject to the Lorentz transformation which involves comparing the frame velocity with that of the velocity of light.

$$E' = \frac{E}{\sqrt{1 - \frac{v^2}{c^2}}}$$

[0086] If the ratio of the velocity v to the velocity of light c is close to unity, then huge relativistic fields can be produced. These fields can curve spacetime to such an extent that a large lift force can be produced on the hull of an electromagnetic field propulsion vehicle.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0087] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0088] **FIG. 1**. Rotating cylinder demonstrating additional lift force.

[0089] **FIG. 2**. Use of the pendulum.

[0090] **FIG. 3**. The Ascension Machine, Chakra Vortex Accelerator.

[0091] **FIG. 4**. Hyperspace energy brought into body by rotation of hand chakra.

[0092] **FIG. 5**. Antigravitational effect using low density hyperspace energy.

[0093] **FIG. 6**. Sonoluminescence experiment.

[0094] **FIG. 7**. Tetrahedron diagram showing hyperspace charge.

[0095] **FIG. 8**. Magnetic monopole.

[0096] **FIG. 9**. Precessing wormhole.

[0097] **FIG. 10**. Fizeau light dragging experiment.

[0098] **FIG. 11**. Asymmetrical velocity collapse of bubble radius generates wormhole.

[0099] **FIG. 12**. VisSim feedback simulation of the collapsing bubble.

[0100] **FIG. 13**. Cavitating bubble radius as a function of time.

[0101] **FIG. 14**. Digitized bubble radius of **FIG. 13**.

[0102] **FIG. 15**. Spiking spacetime curvature of collapsing bubble.

[0103] **FIG. 16**. Dual velocity profile required to produce negative energy.

[0104] **FIG. 17**. Spacetime curvature for dual velocity collapse.

[0105] **FIG. 18**. Negative energy generated from asymmetrical collapse of bubble.

[0106] **FIG. 19**. 3D plot of spacetime curvature using negative energy.

[0107] **FIG. 20**. Wormhole vortex.

[0108] **FIG. 21**. Perspective of cavitation tank equipped with magnetic coils.

[0109] **FIG. 22**. Flattening of bubble by magnetic field.

[0110] **FIG. 23**. Hyperspace energy generated by reverse pressure gradient of wormhole.

[0111] **FIG. 24**. Solenoid and transducer driving circuits.

DETAILED DESCRIPTION OF THE INVENTION

[0112] 1. Referring to **FIG. 24**, the transducer driving circuit uses a frequency generator (A) driving a linear amplifier (B) into a pair of variable inductors (C) which resonate with the capacitance of the transducers (D) attached to the sides of the octagonally-shaped glass container (H). Resistors (L,M,G) provides voltage measuring points for the oscilloscope.

[0113] A DC current source (I) drives the two solenoids (E,F) which produce a vertical magnetic field (J) through the container. The magnetic field produces an asynmetrical force on the bubble surface (K) such that the collapsing bubble produces negative energy. The spacetime curvature produced by the light dragging of the collapsing bubble surface, together with the negative energy, create a wormhole between space and hyperspace. The negative spacetime energy produces a negative pressure which brings low-density hyperspace energy into our dimension.

I claim:

1. A hyperspace energy generator having:

a) a glass container in the shape of an octagon acting as the resonating chamber;

b) two piezoelectric transducers, mounted and operating acoustically as a pair on opposite sides of item (1a);

c) an opaque oil-like liquid with a high index of refraction filling item (1a) acting as the resonating liquid in which the cavitating bubbles are formed;

d) two thin sheet copper solenoids, located on top and bottom of item (1a), that produce a vertical magnetic field through items (1a) and (1c);

e) two variable inductance coils for creating a resonant frequency with item (1b);

f) a linear amplifier driving items (1b) and (1e);

g) a variable frequency generator driving item (1e);

h) a direct current electrical generator driving item (1d);

2. a cavitation system which:

a) produces one or more cavitating bubbles that collapse asymmetrically due to the differential force exerted on the bubble surface by the presence of a vertical magnetic field generated by item (1d) referred to above;

b) generates negative energy due to light dragging of the collapsing bubble surface;

c) creates wormholes between space and hyperspace due to the presence of item (1b);

d) generates a negative pressure between space and hyperspace that forces low-density hyperspace energy into our dimension; and

3. optional dual electrically-charged plates that:

a) are mounted on the sides of item (1a);

b) create a differential electric charge density across the bubble surface for the purpose of creating an additional method of asymmetrical bubble surface collapse.

* * * * *

US 20030209637A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2003/0209637 A1**

(43) **Pub. Date: Nov. 13, 2003**

(54) ROTATING ELECTROSTATIC PROPULSION SYSTEM

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
John St. Clair
Hyperspace Research Institute
52 Kings Court, 4A
San Juan, PR 00911 (US)

(21) Appl. No.: **10/142,582**

(22) Filed: **May 9, 2002**

Publication Classification

(51) **Int. Cl.^{7}** **B64G 1/40**; B64G 1/42

(52) **U.S. Cl.** **244/172**

(57) ABSTRACT

This invention relates to a spacecraft propulsion system utilizing thrusters comprised of a motor-driven electrostatically charged cylinder rotating within an electrostatically charged annular ring for the purpose of creating a spacetime curvature stress-energy tension in the horizontal direction. The thrusters are augmented by magnetic vortex generators, either embedded in the cylinders or located above each thruster, for the purpose of increasing the permittivity of space by permeating each thruster with low density hyperspace energy generated by a wormhole created between our space and hyperspace. A combination of three thrusters mounted on the underside of the hull of the spacecraft provide thrust and yaw motion control.

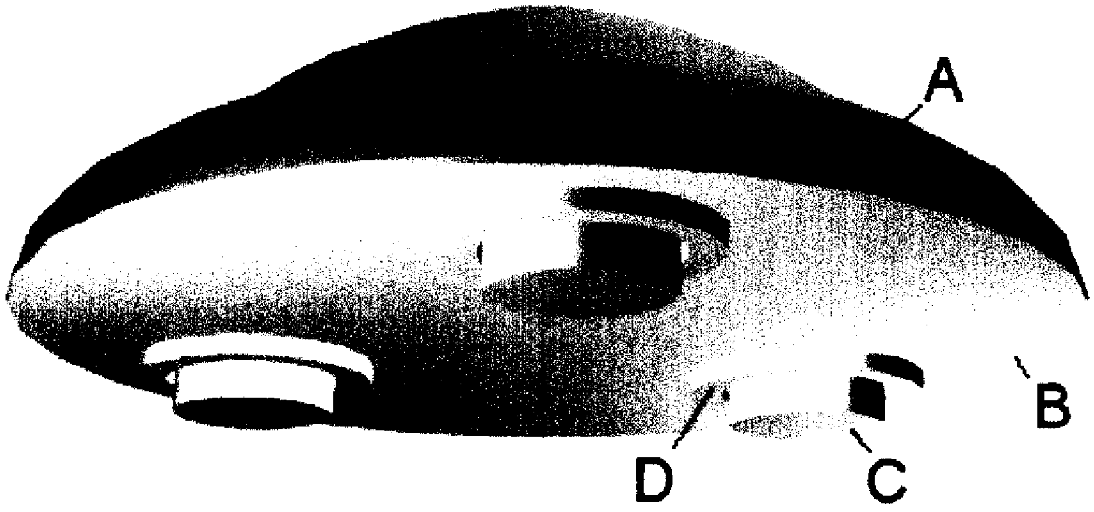

Figure 1

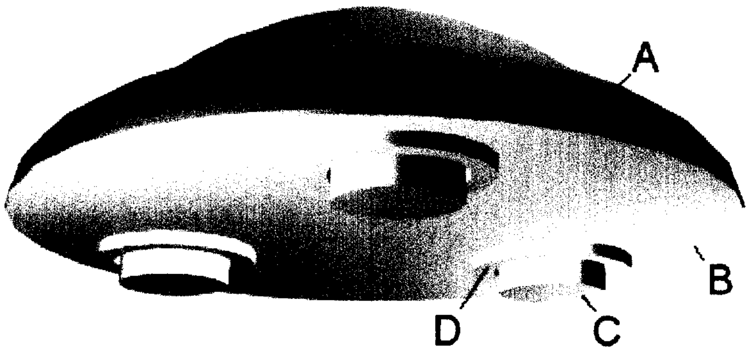

Figure 2

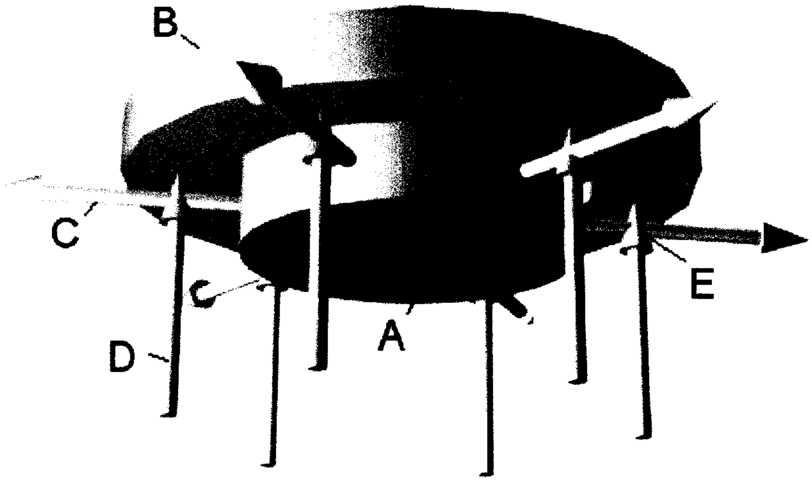

Figure 3

$$T^{zr} = -\frac{\varepsilon_0}{\Omega c^2}\frac{E_z E_r}{4\pi} = \frac{\mathrm{coul}^2}{\mathrm{m}^2\mathrm{n}}\frac{\mathrm{n}}{\mathrm{coul}}\frac{\mathrm{n}}{\mathrm{coul}}\frac{\mathrm{m}}{\mathrm{kg}}\frac{\mathrm{s}^2}{\mathrm{m}^2} = -\frac{\mathrm{kg\,m}}{\mathrm{s}^2}\frac{\mathrm{s}^2}{\mathrm{kg\,m}^3} = \frac{-1}{\mathrm{m}^2}$$

Figure 4

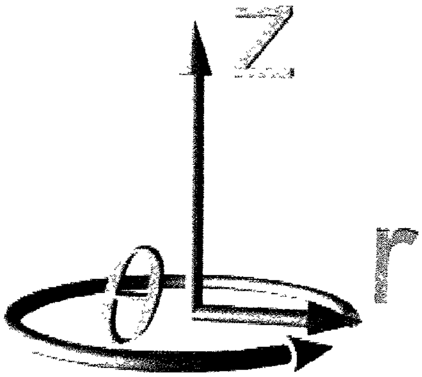

Figure 5

$$F^{\alpha}{}_{\beta} = \begin{array}{c|cccc|} t & 0 & E_r & 0 & E_z \\ r & E_r & 0 & 0 & 0 \\ \theta & 0 & 0 & 0 & 0 \\ z & E_z & 0 & 0 & 0 \end{array}$$

Figure 6

$$T^{zr} = -\frac{E_r E_z}{4\pi}$$

Figure 7

$$\frac{d}{dt} S_r = \varepsilon_0\, \varepsilon_{r\,t\,z} x^t T^{zr} n_r \text{Area} = \varepsilon_0\, t \frac{E_r E_z}{4\pi} n_r \text{Area}$$

Figure 8

$$\frac{\text{coul}^2}{\text{m}^2\,\text{n}} \sec \frac{\text{n}}{\text{coul}} \frac{\text{n}}{\text{coul}} \text{m}^2 = \text{kg} \frac{\text{m}}{\sec^2} \sec = \text{kg} \frac{\text{m}}{\sec}$$

Figure 9

$$\frac{d}{dt} S_{r,t} = \frac{\varepsilon_0}{4\pi} e^{i\omega t} E_r E_z \,\text{area} + \frac{i\varepsilon_0}{4\pi} e^{i\omega t} E_r E_z t\,\omega\, \text{area}$$

Figure 10

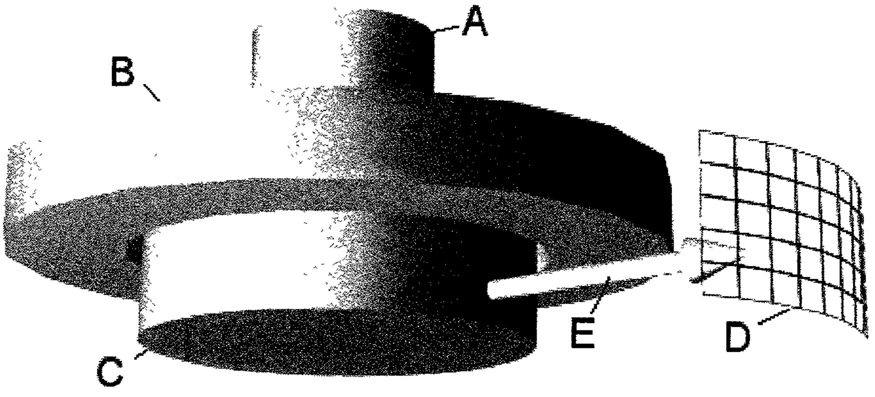

Figure 11

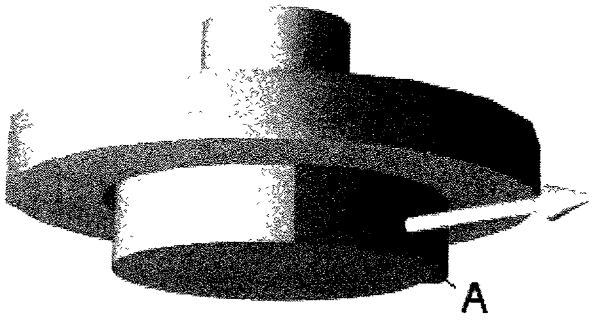

Figure 12

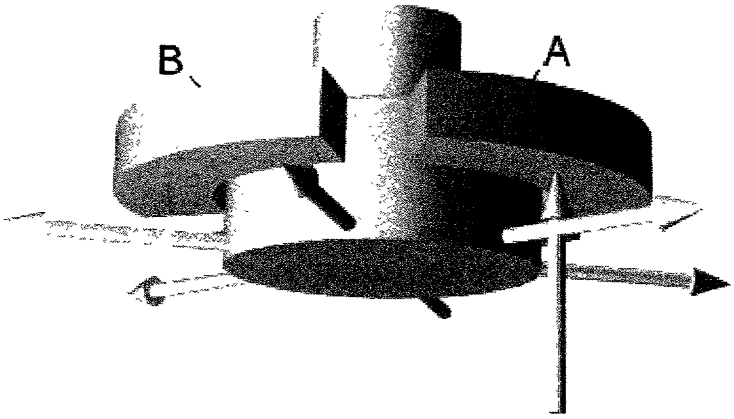

ROTATING ELECTROSTATIC PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] The invention, which is the object of my present application, is a spacecraft propulsion system which develops a spacetime curvature tension utilizing a combination of a rotating radial electrostatic field and a fixed vertical electrostatic field. The two fields create a stress-energy T^{zr} gradient in the radially direction which is equal to force. The radial field is created on the side of a charged rotating cylinder on the underside of the hull. The vertical field is created by an annular charged ring concentric with the cylinder. Three rotating cylinders are located in a triangle on the bottom of the hull in order to produce a force in any direction in the horizontal plane.

REFERENCE PAPERS

[0002] *Gravitation*, Wheeler, page 80.

BACKGROUND OF THE INVENTION

[0003] When working with Maxwell's equations in tensor notation, it became apparent that a tensor can change identity depending on what permutation of variables is involved. For example, one single equation can involve both charge density and current density. And all of Maxwell's equations can be reduced to just two equations.

[0004] In the tensor equation for momentum, if the lever arm is length then the equation is equal to the flow rate of angular momentum. If the lever arm is time, then you get linear momentum. And if the field rotates with time, then the time rate of change of linear momentum is a force which is the basis for this invention.

[0005] Einstein said that mass curves space and space tells mass how to move. In this sense, generalized mass can be mass, electromagnetic fields, charge or angular momentum which create a spacetime curvature that produces a force on the spacecraft.

SUMMARY OF THE INVENTION

[0006] The invention relates to a spacecraft utilizing a rotating electrostatically charged cylinder and a concentric annular charged ring to create a stress-energy spacetime curvature in the horizontal plane on the spacecraft's underside hull. A motor drives the rotating cylinder which extends below the hull. A charged surface produces an electric field in the direction normal to the surface. The vertical and rotating electric fields combine to create a rate of change of linear momentum which creates a horizontal propulsive force on the hull.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0007] Not Applicable.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0008] **FIG. 1**. Perspective view of spacecraft hull (A) with the three electrostatically charged rotating cylinders (C) surrounded by their annular charge rings (D) extending down on the underside of the hull (B).

[0009] **FIG. 2**. Perspective view of individual thruster showing rotating cylinder and ring.

[0010] **FIG. 3**. Stress-energy tension created by two electric fields and having units of curvature.

[0011] **FIG. 4**. Cylindrical spacetime coordinates {t, r, θ, Z}.

[0012] **FIG. 5**. Faraday electromagnetic F tensor containing electric fields in the radial and vertical directions.

[0013] **FIG. 6**. Stress-energy tensor T showing it is equal to the product of the two electric fields divided by 4π.

[0014] **FIG. 7**. The tensor equation for flow rate of angular momentum S.

[0015] **FIG. 8**. The units are linear momentum due to the time lever arm.

[0016] **FIG. 9**. The rate of change of linear momentum is the horizontal force produced by the two electric fields.

[0017] **FIG. 10**. The angular momentum flows through an area whose normal vector is in the radial direction.

[0018] **FIG. 11**. Cylinder with only one electrostatically charged segment.

[0019] **FIG. 12**. Annular ring with three individual segments which can be charged separately to create a force in a particular direction.

DETAILED DESCRIPTION OF THE INVENTION

[0020] 1. Referring to **FIG. 1**, the spacecraft comprises an upper hull (A) with three rotating electrostatically charged cylinders (C) with their concentric annular electrostatically charged rings (D) located on the spacecraft's bottom hull (B).

[0021] 2. In a closer view of one of the cylinders seen in **FIG. 2**, motor-driven rotating cylinder (A) has an electrostatically charged surface which produces an electric field (C) normal to said surface. The fixed, electrostatically charged annular ring (B), which is concentric with the cylinder, produces a vertical electric field (D) normal to its surface. This crossed field (E) creates a negative spacetime curvature tension which is the product of the two fields divided by 4π as seen in the equation, **FIG. 3**.

[0022] 3. Notice that the equation involves the permittivity of space ϵ_0 divided by the linear mass of the universe Ω and the speed of light. This produces units of inverse meter squared which is the spacetime curvature. In Einstein's General Theory of Relativity, the spacetime curvature tensor is equal to the stress-energy tensor or $G=8\pi T$ where G is the curvature, and T is the product of the electromagnetic fields. The problem with this equation, which has been resolved with this invention, is that the linear mass Ω times the speed of light c is an enormous number. Even with the square of enormous electric fields, the curvature would be too small even to notice, and little force would be generated.

[0023] 4. In another patent application of mine entitled Magnetic Vortex Generator, it was shown that a rotating cylinder containing embedded and stacked bar magnets produces a negative mass and negative spacetime spring constant. It can be shown that this combination produces a small wormhole or interdimensional connection between our space and hyperspace along the centerline of the rotating cylinder. Co-dimensions of hyperspace have different physics constants. A low pressure region of hyperspace has a very low mass density and a very low speed of light. The wormhole allows this low density hyperspace energy to enter into our space and permeate the cylinder and annular ring. The permittivity is proportional to the inverse of the speed of light squared. The hyperspace speed of light, obtained from my tetrahedron physics diagram, is 8971 meters per second. The speed of light in our dimension is 299792458 meters per second. As shown by the enclosed reference calculation, the hyperspace permittivity is about a trillion times larger. Because the force is equal to this new permittivity times the electric fields squared times the area around cylinder, the force is greatly amplified by this increase in the permittivity of space.

[0024] 5. It is pertinent to this invention how the stress-energy is created due to the two electric fields in the vertical and radial direction. In gravitational physics, there is a Faraday F tensor which contains all the components of the electromagnetic fields. It is a 4 by 4 matrix whose rows and columns correspond to the coordinates of spacetime which in cylindrical coordinates are $\{t, r, \theta, z\}$ where t is time, r the radius, theta the horizontal angle and z the vertical height. These coordinates are shown in **FIG. 4**.

[0025] 6. The radial electric E_r field and the vertical electric E_z field can be inserted into the Faraday tensor seen in **FIG. 5**. The sign of the vertical field is positive because it points in the positive z-direction due to the fact that the annular ring has a negative charge. The cylinder has a positive charge. This produces the negative stress-energy tensor T^{zr} as drawn in **FIG. 6**.

[0026] 7. In order to calculate the force on the cylinder, it is necessary to calculate the flow rate of angular momentum. Momentum is mass times velocity or mass meter per second. If the mass is moving in a circle, then there is a lever arm times the momentum which makes it mass meter squared per second. If this is differentiated with respect to time, then a flow rate of angular momentum is produced with units of mass meter squared per second squared. As mentioned previously, tensors have this dual nature where depending on the permutation of the variables, it means one thing or another. In this case, the lever arm will be time, rather than length which converts the flow rate of angular momentum into just linear momentum. If you change linear momentum with respect to time, then you get a force.

[0027] 8. The flow rate of angular momentum S is shown in **FIG. 7**. The force has to be against the area whose normal vector is in the radial direction which is also the direction of the momentum. So S has an r subscript indicating that it flows in the radial direction. The permutation tensor ϵ has three subscripts which keeps track of the tensor notation. The first subscript is the same as the momentum subscript. Permutations of the coordinate variables which are in order have a plus one sign. Permutations which are in reverse order have a minus one sign. Permutations in which the variable are repeated are zero. For example, $\epsilon_{tr\theta z}=-\epsilon_{rt\theta z}$ because the r and t are in reverse order in cylindrical coordinates. Because the permutation tensor starts with r, then we can have permutations such as $\{r,t,z\}$ which is a reverse order negative permutation. This negative sign cancels the negative sign of the stress tensor. The reason this permutation is chosen is because the first subscript on the stress-energy tensor is now z. Because the normal to the area is in the radial direction, then the stress-tensor matches the electric fields that we have available, which are also in the zr-direction.

[0028] 9. In this case, the second subscript of the permutation tensor is time, rather than length. So the units become, as shown in **FIG. 8**, those of linear momentum. To me, this was very surprising. Then I realized that the radial electric field rotates with time which means the differential of the linear momentum produces a force on the hull. In exponential notation, the radial electric field rotates with Exp[iωt]. This is multiplied by the time lever arm, so the term that has to be differentiated is $t\ e^{i\omega t}$. The time lever arm saves the differentiation by making one term real so that the force is real. This is shown in **FIG. 9**. The units of the first term are real newtons. In the second term, the time t multiplied by the frequency cancels out in terms of units, so the units are still force, but imaginary

[0029] 10. There is an area involved in the force equation which is depicted in **FIG. 10**. The charged rotating cylinder (C) located inside the charged annular ring (B) is driven by motor (A). The radial electric field (E) is normal to the area (D) whose normal vector is also in the radial direction. The angular momentum flows through this area which surrounds the cylinder. In doing so, it curves spacetime which produces the force.

[0030] 11. If the radial electric field is continuous around the cylinder, then the net force is zero. Referring to **FIG. 11**, one section (A) of the rotating cylinder is charged, which means that there is a force once per cycle in a selected direction depending on when the annular ring is charged. The other option, referring to **FIG. 12**, is that the annular ring is divided up into two or three sections with the feature that the charge can be turned on (A) or off (B) on a particular segment. As the rotating electric field goes around, one of the annular ring sections will have an electric field in order create a force on that side.

[0031] 12. Since there are three force cylinders, this allows for yaw motion control so that the hull of the spacecraft can pivot to change direction. After the directional change, the two back cylinders can be synchronized to produce thrust in the forward direction.

[0032] 13. There could also be a polarity change for the vertical electric field such that a positive stress-energy is produced which would reverse the direction of the thrust.

What I claim as my invention is:

1. A spacecraft propulsion system utilizing three electrostatically charged motor-driven cylinders each one of which rotates within a charged annular ring to produce a spacetime stress curvature tension in the horizontal plane on the underside of the hull.

2. Said rotating cylinder comprised of bar magnets embedded in the cylinder and stacked in groups at intervals around the periphery of the cylinder with the purpose of increasing the permittivity of space by permeating the cylinder and ring with low linear mass, low speed of light hyperspace energy by means of a wormhole between our space and hyperspace. The larger permittivity is to increase the force. Said technique is contained in my patent applications Magnetic Vortex Wormhole Generator and Magnetic Vortex Generator.

3. Said rotating cylinder and annular ring having a suitable metal surface for forming and maintaining the electrostatic charge.

4. Said rotating cylinder having one or more segments which can be electrostatically charged individually or together to produce a force in a particular direction.

5. Said annular ring having one or more segments which can be electrostatically charged individually or together to produce a force in a particular direction.

6. A combination of three such thrusters providing a force in any particular direction or for yaw motion control.

7. Electric polarity switching of the fields in order to reverse the spacetime curvature and therefore change the direction of thrust.

8. The use of a magnetic vortex generator located above each thruster in order to permeate the cylinders and rings with low density hyperspace energy which would substitute for the embedded magnets in each cylinder.

* * * * *

US 20060072226A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0072226 A1**

St. Clair (43) **Pub. Date: Apr. 6, 2006**

(54) REMOTE VIEWING AMPLIFIER

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/957,391**

(22) Filed: **Oct. 2, 2004**

Publication Classification

(51) **Int. Cl.**
G02B 1/00 (2006.01)

(52) **U.S. Cl.** **359/896**; 359/894

(57) ABSTRACT

An apparatus which enhances the ability of a person to perform remote viewing by connecting the human spiritual eye to the tetrahedral geometry of subspace.

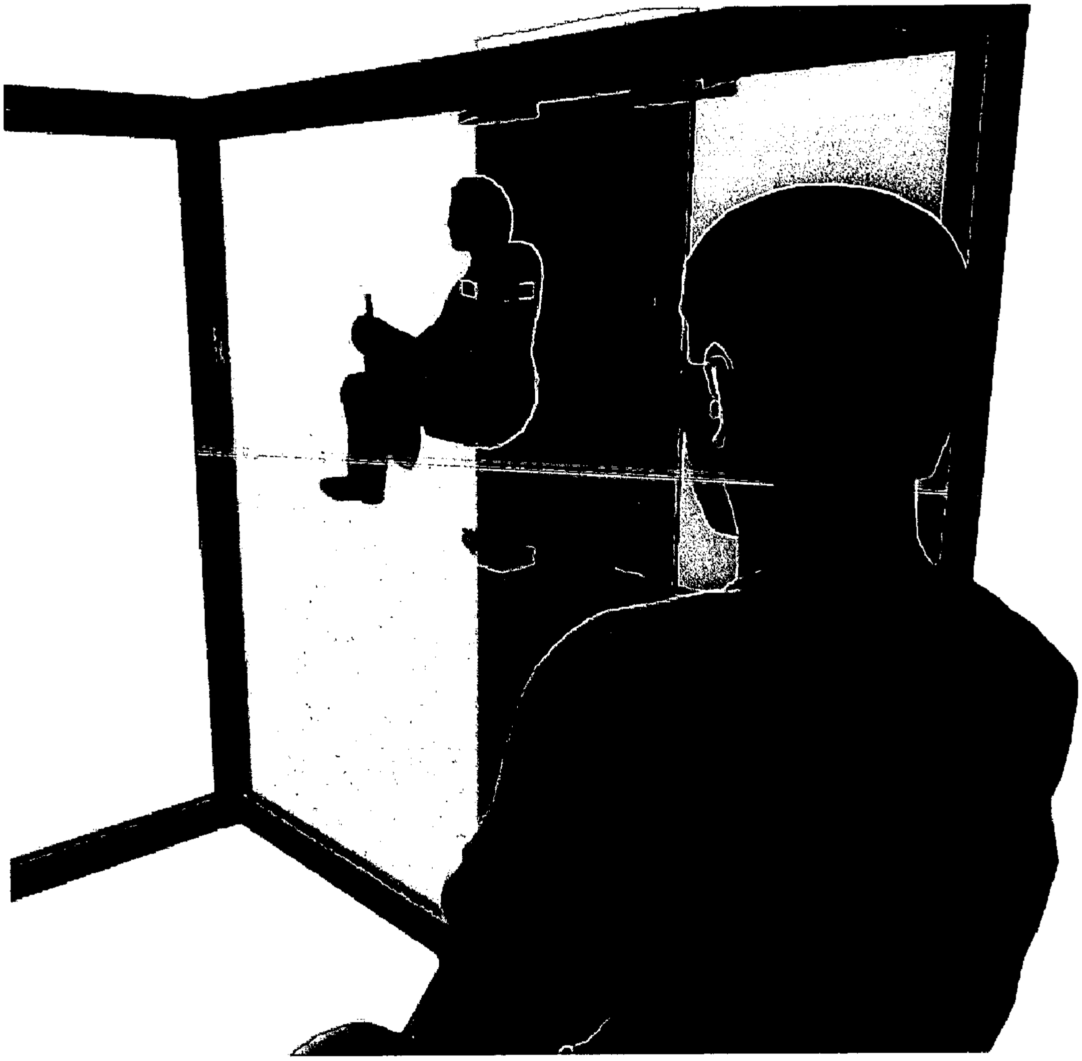

Figure 1

Figure 2

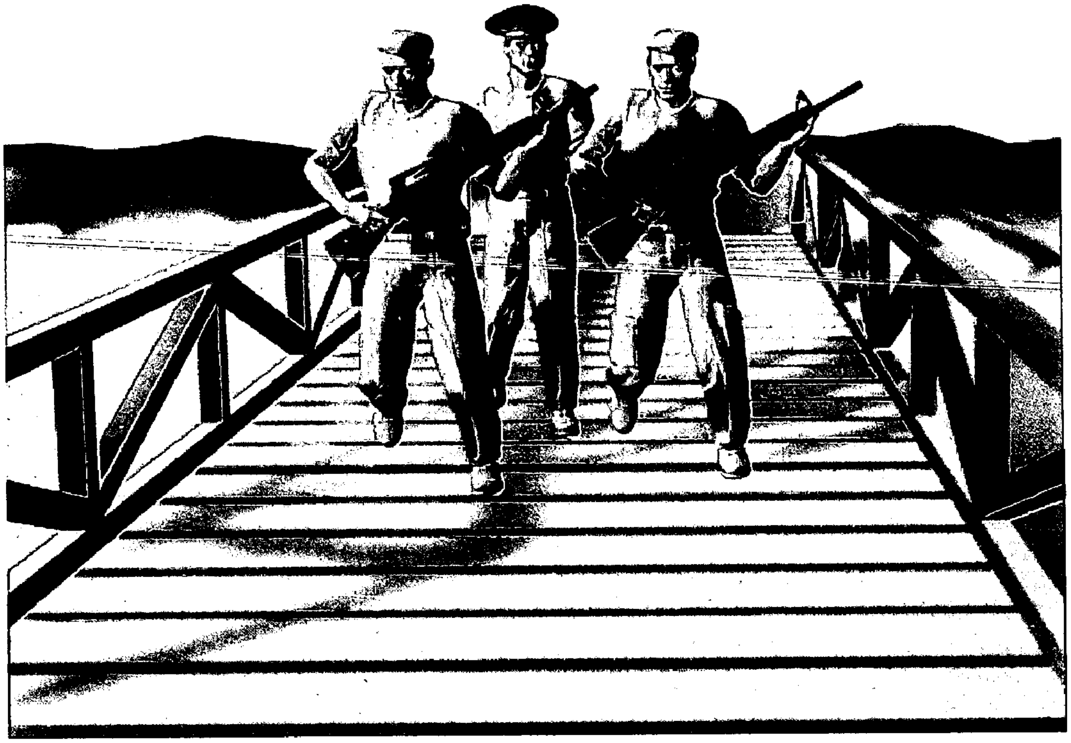

Figure 3

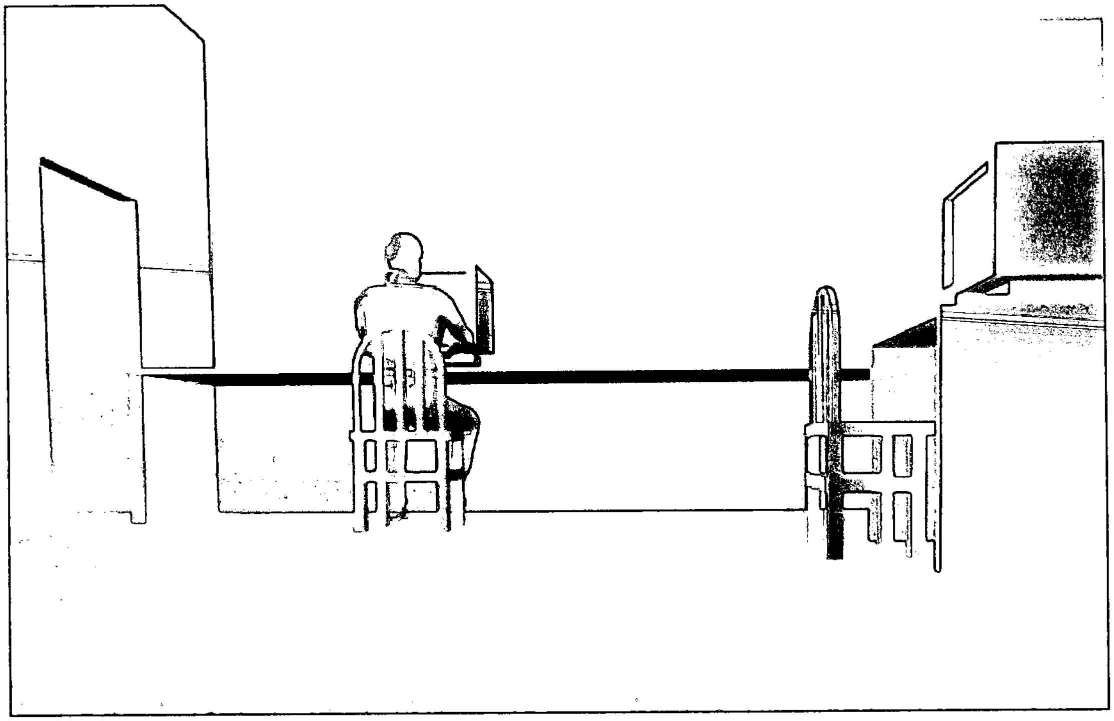

Figure 4

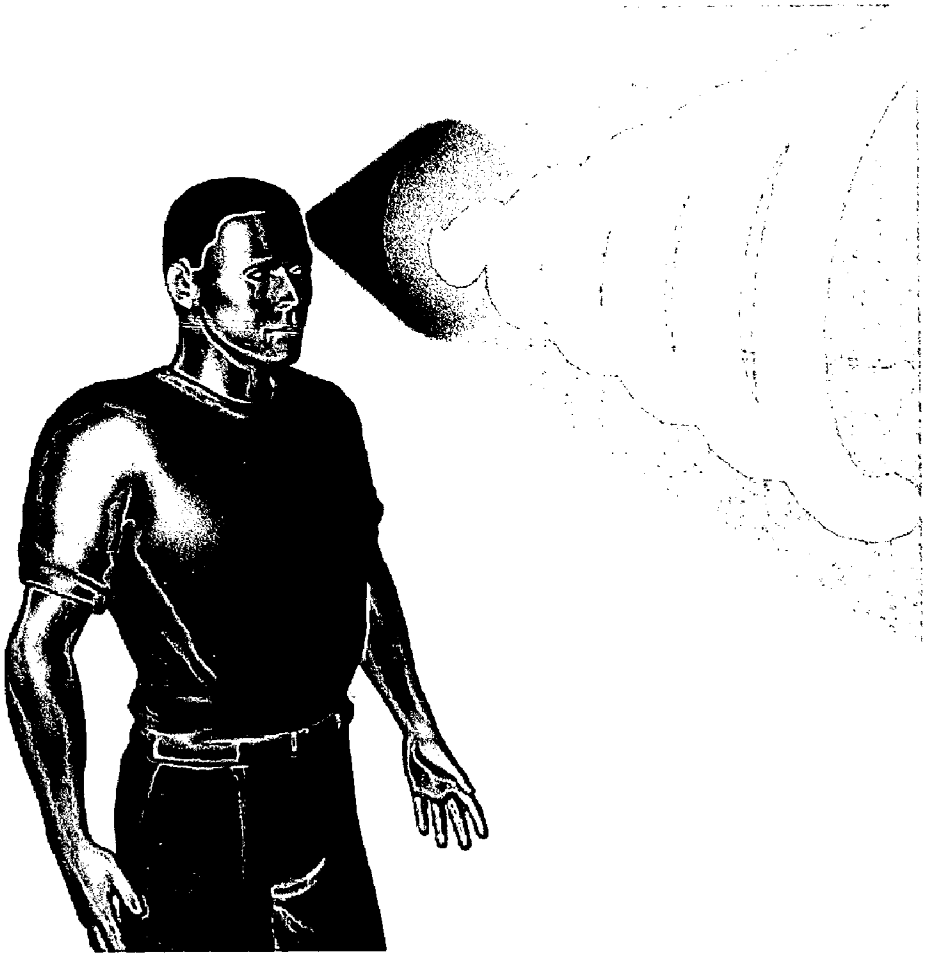

Figure 5

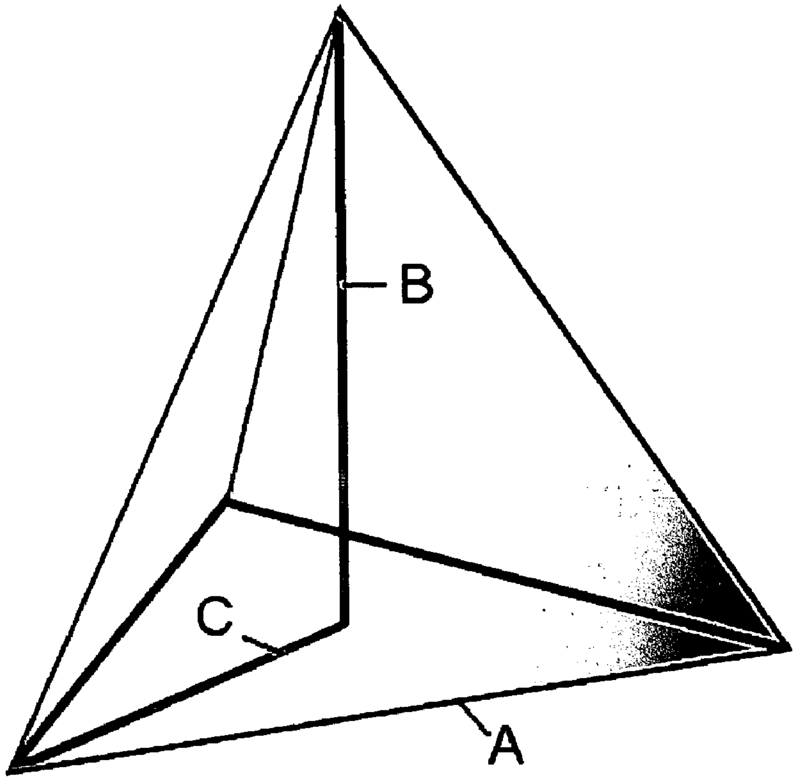

Figure 6

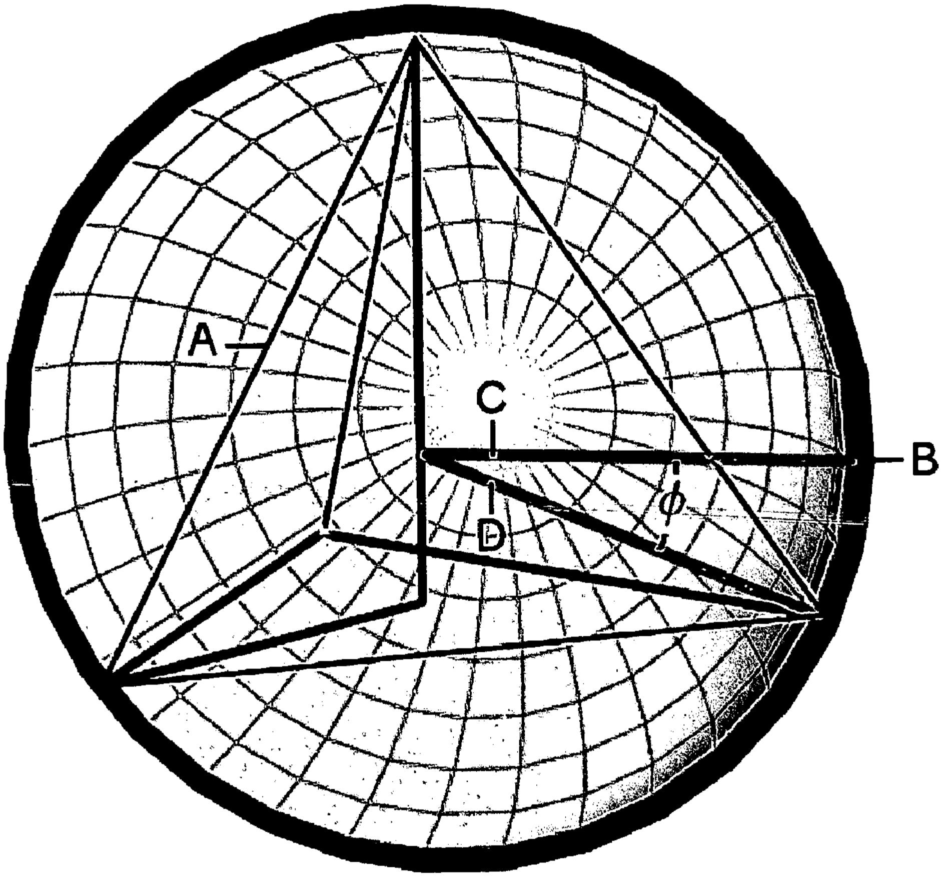

Figure 7

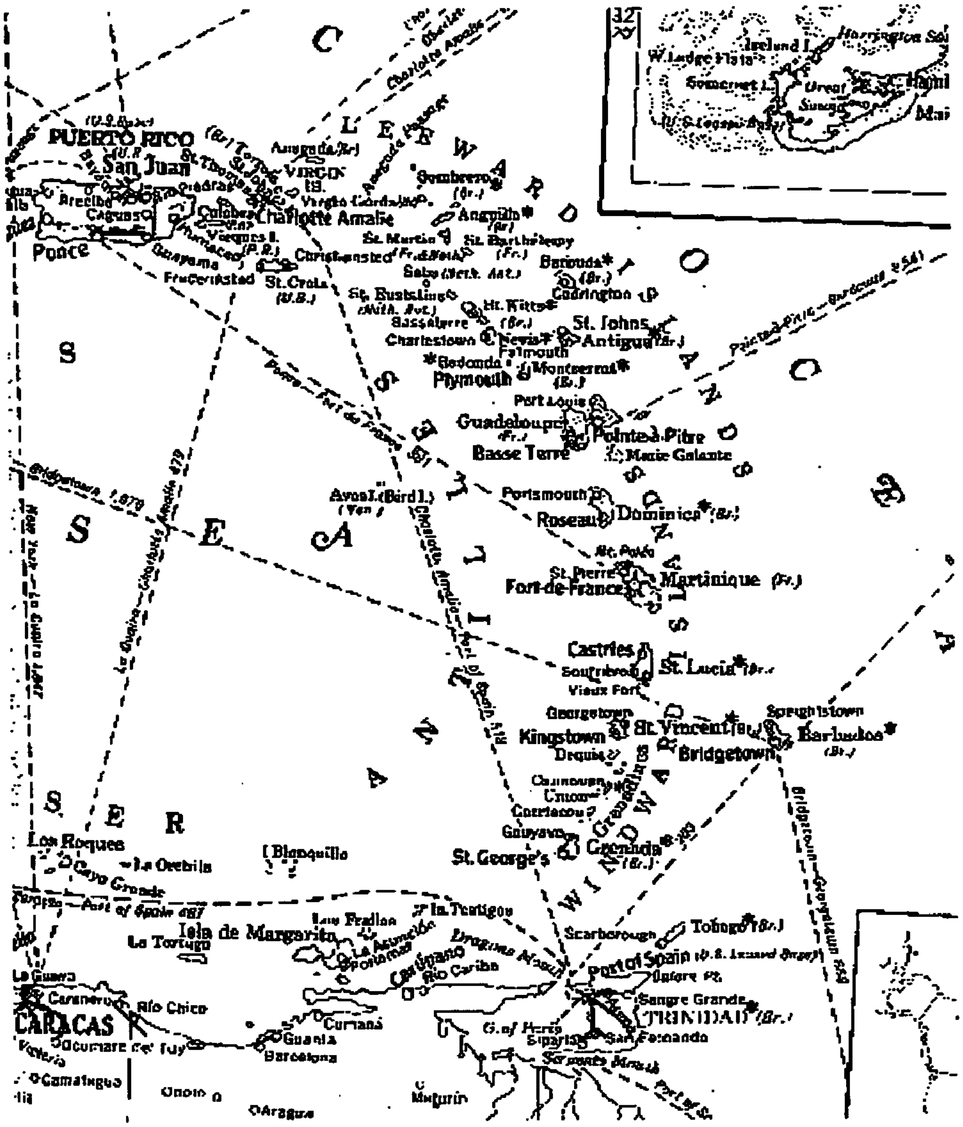
C
L E E W A R D
PUERTO RICO
San Juan
Arecibo
Caguas
Ponce
Culebra
Vieques I.
Charlotte Amalie
Christiansted
Frederiksted
St. Croix
Anguilla
Sombrero
St. Martin
St. Barthélemy
Saba
St. Eustatius
Barbuda
Codrington
St. Kitts
Basseterre
Charlestown
Nevis
St. Johns
Antigua
Falmouth
Redonda
Montserrat
Plymouth
Port Louis
Guadeloupe
Pointe-à-Pitre
Basse Terre
Marie Galante
Aves I. (Bird I.)
Portsmouth
Roseau
Dominica
St. Pierre
Fort-de-France
Martinique
Castries
Soufrière
St. Lucia
Vieux Fort
Georgetown
Kingstown
St. Vincent
Speightstown
Barbados
Bridgetown
Bequia
Carriacou
Gouyave
St. George's
Grenada
S
S E A
Los Roques
La Orchila
Blanquilla
Los Frailes
La Tortuga
Isla de Margarita
Los Testigos
Scarborough
Tobago
Port of Spain
Sangre Grande
TRINIDAD
San Fernando
Siparia
G. of Paria
La Guaira
CARACAS
Río Chico
Cumaná
Guanta
Barcelona
Maturín
Ocumare del Tuy
Camatagua
Aragua
Ireland I.
Somerset I.
Hamilton
32

Figure 8

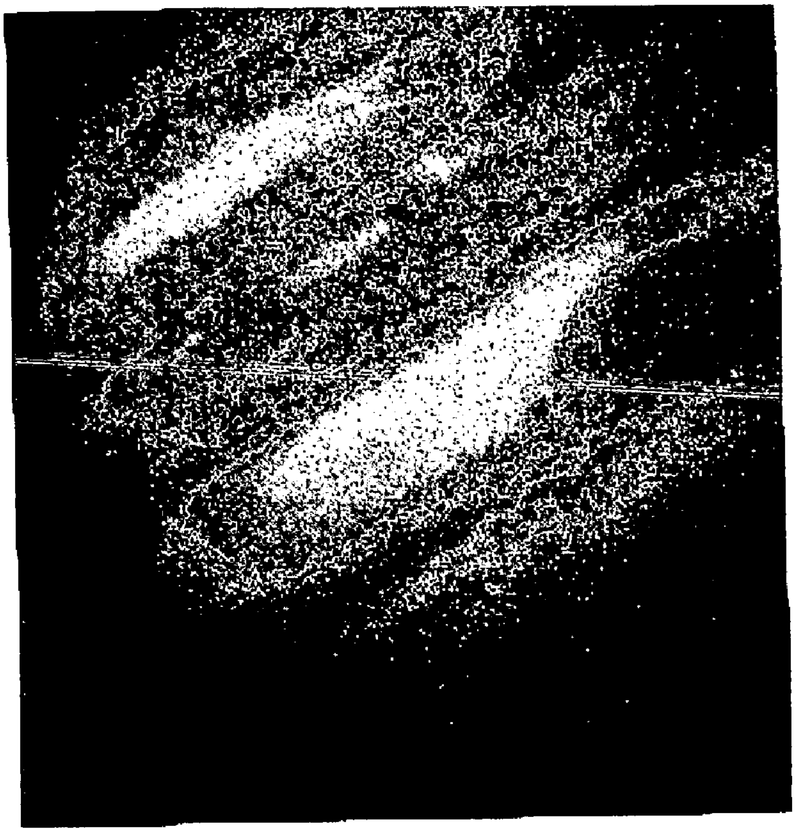

Figure 9

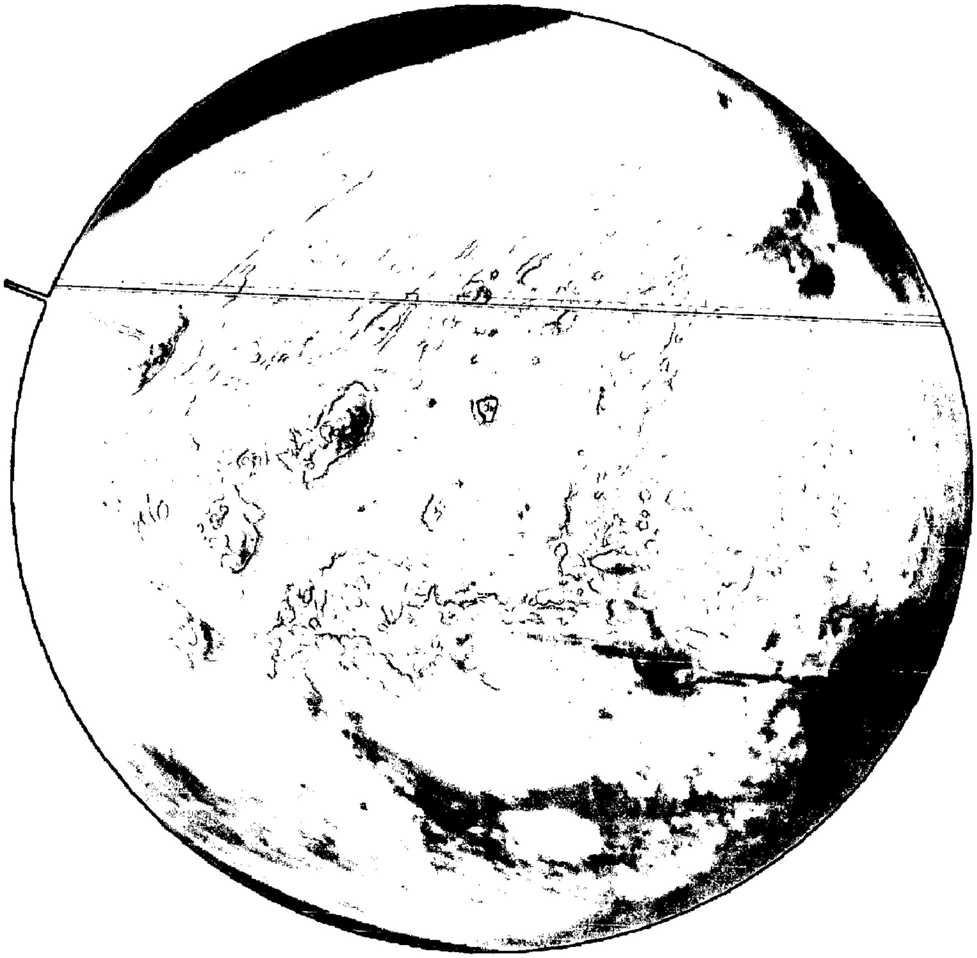

Figure 10

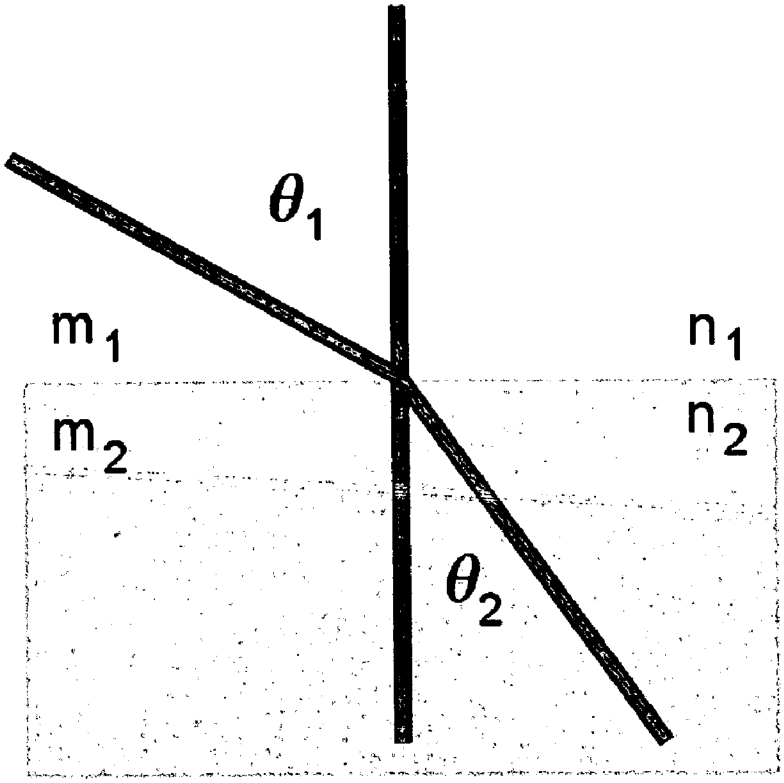
θ_1
m_1
n_1
m_2
n_2
θ_2

Figure 11

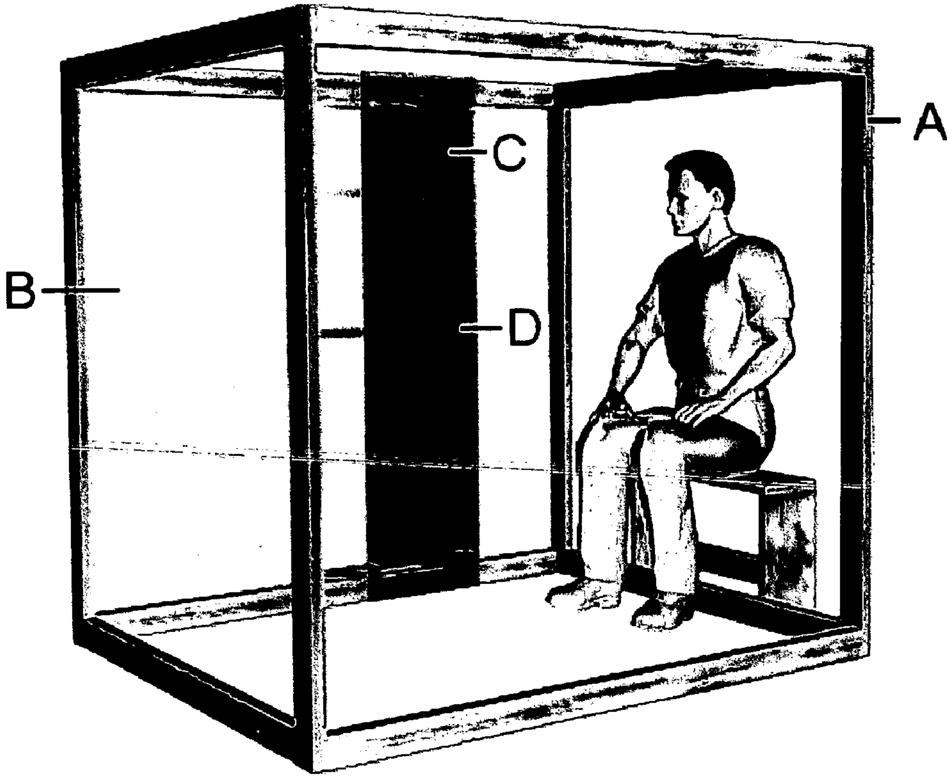

Figure 12

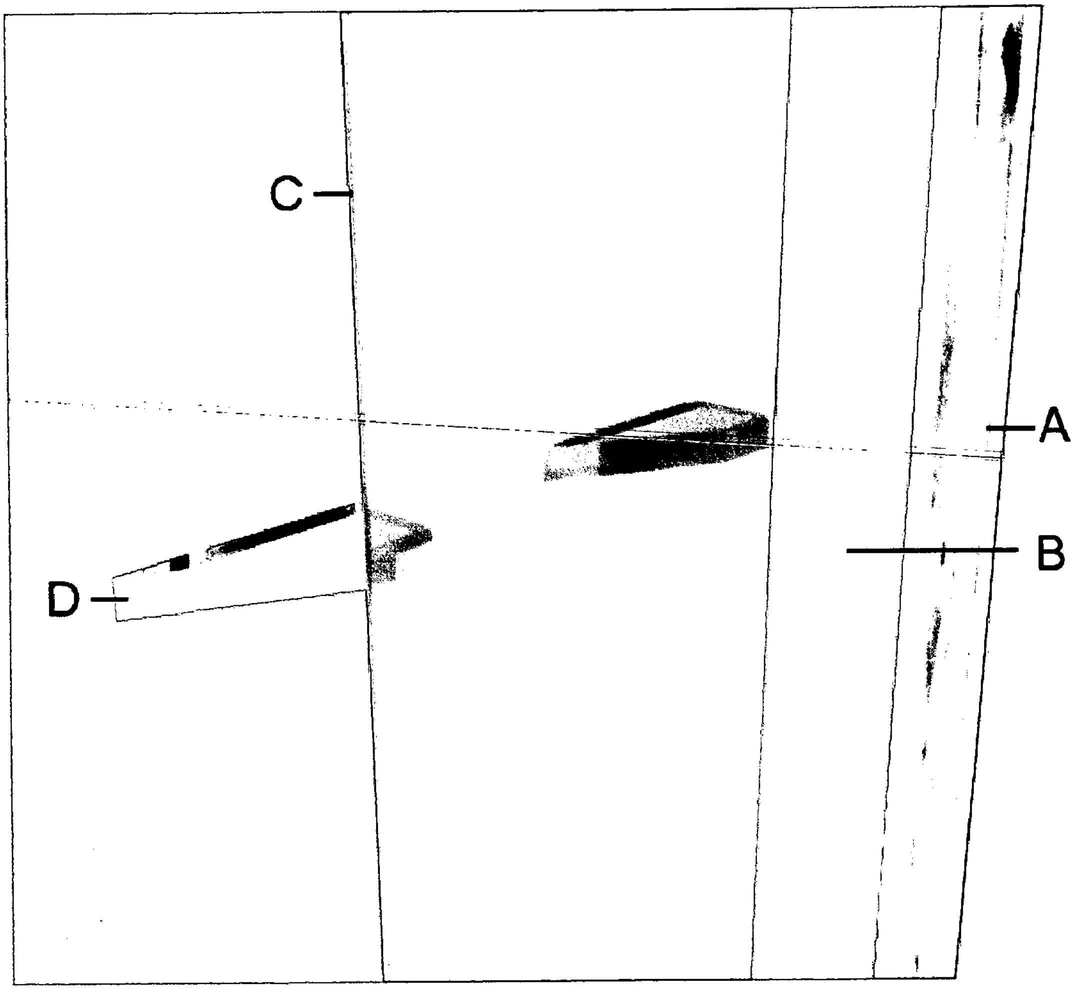

Figure 13

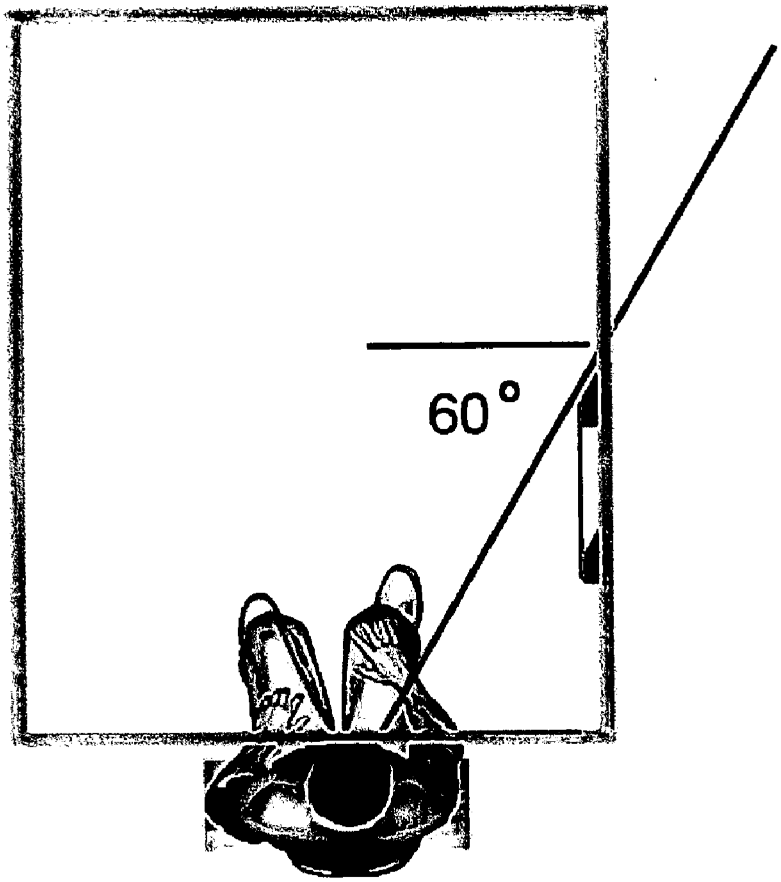

Figure 14

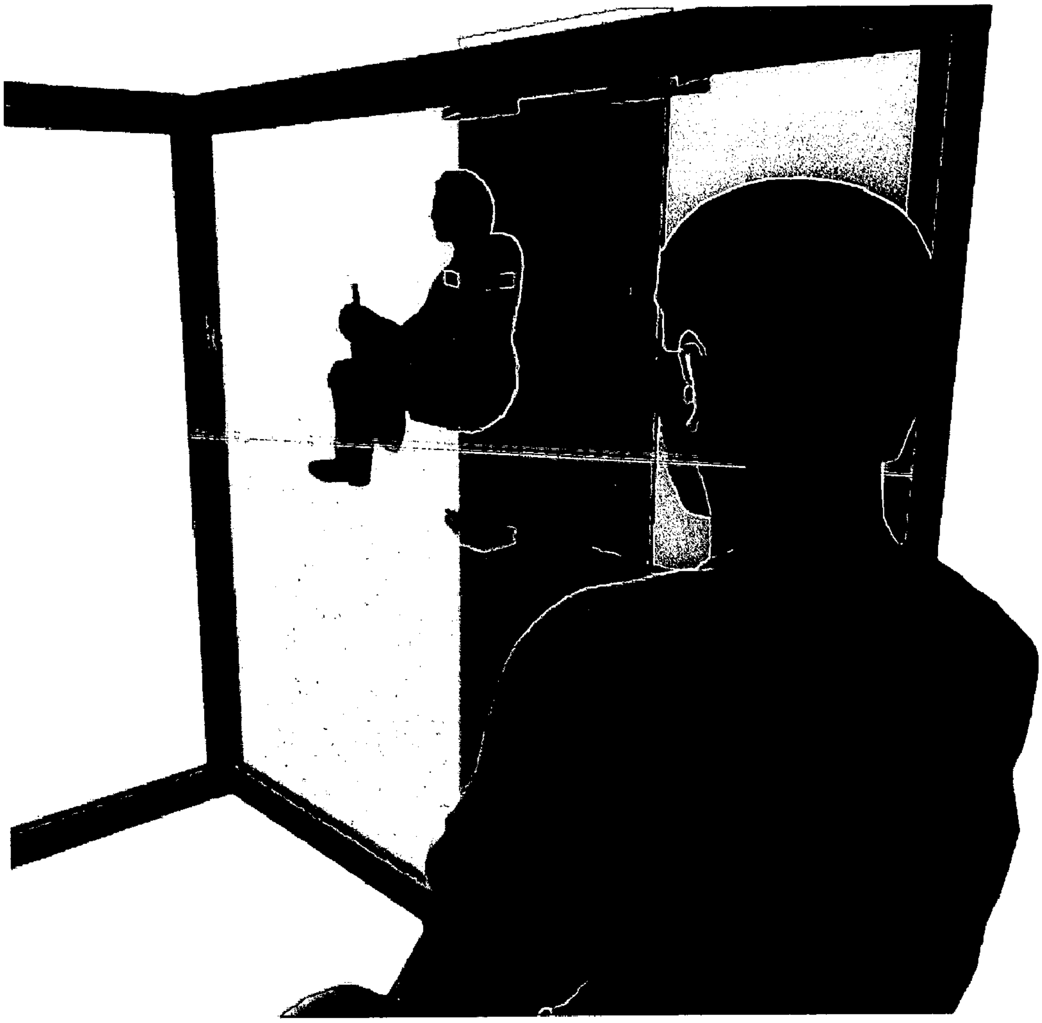

REMOTE VIEWING AMPLIFIER

BRIEF SUMMARY OF THE INVENTION

[0001] This invention enhances the ability of a person to perform remote viewing by connecting the spiritual eye to the tetrahedral geometry of subspace.

BACKGROUND OF THE INVENTION

[0002] Remote viewing is the projection of spiritual modules of the human energy field to distant locations in order to see, communicate and interact with other entities who live in subspace, space and hyperspace co-dimensions of the universe.

[0003] One of my first remote viewings was made at night to a distance of 10,000 miles on the sunlit side of the earth. My spiritual eye and body projected together while my mental facilities remained in my physical body. I found myself looking down on a palm tree from a height of about one hundred feet. The palm tree had several coconuts in it as seen in **FIG. 1**. I then gave the command to lower myself to the ground. At that moment I went sailing down past the coconuts, barely missing the tree! Finding myself on a pathway through the tropical forest, I then came to an extremely long wooden bridge which crossed over a river gorge. On the other side of the bridge I could see three soldiers running toward me as shown in **FIG. 2**. The two soldiers in front were carrying rifles and wearing light blue berets. The man running behind them was wearing an officer's cap with a red band. My first reaction was that I was going to be shot. I edged over on the right side of the wooden railing. They ran right past without seeing me. I then asked to see the building that these soldiers were guarding. Everything went dark, and then I found my spiritual eye peeking out of the floor of a computer room as seen in **FIG. 3**. There was one man using a computer on the opposite side of the room near an open door. He got up from his chair and came over to sit in front of a second computer located a few feet from where I was located. From the glare of the computer monitor, I could clearly see his face. Everything went dark as my spiritual eye and body projected back to my physical body.

[0004] Another time my spiritual eye, spiritual body and mind were standing outside the closed front door of my condominium. Upon patting my legs with my hands, I couldn't find the keys in my trousers. When I looked down, I realized I wasn't in my physical body. I then shot through two solid walls of concrete and returned to my awakening body.

[0005] What these two examples show is that the human spiritual energy system is modular. The reason it is modular is because there are seven hyperspace co-dimensions, each vibrating at a slightly different frequency, which receive energy from space through seven vortices located along the physical body. Over a lifetime, these vortices build up the human aura. Because all energy systems have to be grounded, the remaining six modules are the legs, body, arms/hands, voice, eye and mind. When the entire group is out-of-body as a single entity, then the soul energy powers the body in a manner similar to a battery. The soul looks like a two-inch diameter orange ball of plasma. If the soul is removed from the body, then the body becomes paralyzed except for a small movement of the eyelids. Upon death, all these separate modules are assembled into a single energy being. A light cord from the soul provides the battery power and information required to join these modules together. A second light cord from the pituitary gland transfers the modules into the energy being for assembly. This energy information transfer is the reason that people in a near-death situation say they saw their entire life flash before them.

[0006] Moving to a larger picture of things that are happening in the galaxy, I was able to make contact with the Pleiadian Federation which is located about 400 light years from earth. The Federation is a group of over one hundred intelligent beings that were brought to the Pleiades from around the galaxy. One member of the Federation calls itself the Intelligent Insect Beings. They are the ones who fly the black triangles over Belgium and France for the purpose of evacuating human beings back to the Pleiades for relocation on a planet called Earth II. The reason for this evacuation was that it was not known if it would be possible to win the battle of Revelations, which would take place about two years later here on earth.

[0007] As could be expected, the humans were angry and hysterical at being abducted. So the Intelligent Insect Beings asked me if I could calm them down. I was in telepathic communication with them, and they were in telepathic communication with their computer system, which meant that I could have my thoughts displayed to the humans on the computer monitor. It turned out that one woman was from Central America and only spoke Spanish. The Intelligent Insect Beings didn't speak Spanish so they thought there would be no way I could communicate with her. So I told them that I would spell the Spanish words letter-by-letter and she could then read my message. Since most Spanish people are Catholics, I thought a religious message would be of importance to her. I spelled out the phrase, "Que Dios te bendiga." which means, "May God bless you." I also asked the two beings to put their hands together in a form of prayer, and the woman followed suit. At that moment the Intelligent Insect Beings were reading her mind to understand her emotional state. They said, "She is crying tears . . . " After a slight but tense pause, they continued, "of joy!" They said the woman had a big smile on her face and was successfully transitioned into her new life.

[0008] One year later, the Blond Aliens of the hundred member council of the Pleiadian Federation remote viewed me so that I could celebrate with them the success of their mission to earth. The Blond Aliens, which is not their real name, fly the Beamship spacecraft. This hyperspace vehicle can teleport itself to any location in the galaxy. It dematerializes into the black void for an instance and then rematerialize anywhere in the galaxy. This ship can be seen in Billy Meier's video from Switzerland along with a picture of their envoy, Semjase.

[0009] Just recently I had a demonstration of the Beamship's ability. I heard an emergency distress call by a commander of a spacecraft who said that they had a fire onboard their spacecraft. I immediately remote viewed a Federation maintenance mothership, they got the frequency and location of the ship, and dispatched the Beamship in time to rescue the commander. The burned-out cables, which had caused the fire, were repaired by the Federation and the commander arrived home safely on his planet in his own spacecraft.

[0010] The Federation then received a message from the commander saying that he wanted his planet to join the Federation now that there was this type of communication available. So the Federation visited his planet for the signing ceremony, and I was invited to attend the proceedings by remote view. Word got around about this, and within three weeks another 20 planets joined the Federation for the same reason.

[0011] At one ceremony, which was attended by Admiral Third Class of the Pleiadian Defense Department, His Highness of the planet was signing the document of incorporation. At that moment, I could see him signing, so I exclaimed, "He is signing with his left hand." The Admiral was almost apoplectic at hearing this. After giving her the signed document, His Highness held up his hand and asked me how many fingers he was holding up. He had a fist so I said none. Then he put his index finger out and I said one. He then made a fist again and I said none. Then he held out all fingers and I said five. The Admiral said that he had a big smile on his face as he went to announce the agreement because he knew, even though his planet was located **90,000** light years away on the other side of the galaxy, he could instantly communicate any problems to the Federation. So this is the importance of developing inventions that can enhance our remote viewing ability because one day it will mean that we can become a vital part of the Pleiadian Federation.

SUMMARY OF THE INVENTION

[0012] Referring to **FIG. 4**, the spiritual eye of the human energy system is located at the pituitary gland in the forehead. It has the shape of a hollow cone which is composed of the misty white energy of hyperspace. Light coming into this vortex is then transferred by a light cord to a visual energy module which is located in a co-dimension of hyperspace. Because these modules are interconnected by light cords, the mind module is able to interpret the visual pattern the eye is seeing. More importantly, the mind can give logical instructions to this spiritual eye module for it to rotate around or move in a particular direction.

[0013] The reason that hyperspace has a white misty look to it is that the speed of light is very much less than the speed of light in our spacetime. The Lorentz transformation says that the distance L is shortened relativistically to a distance L' in a way related to the ratio of the velocity v of the object to the velocity of light c.

$$L' = L\sqrt{1 - \frac{v^2}{c^2}}$$

If the velocity of light is very low, then a small velocity creates an enormous contraction in length. By moving through hyperspace, therefore, enormous distances can be traversed. And this is the reason that it is possible to project the spiritual eye, voice and hearing to remote locations in the galaxy.

[0014] The universe is composed of subspace, space and hyperspace which are co-dimensions of each other. Subspace is defined by the geometry of the tetrahedron which is a four-sided solid whose faces are equilateral triangles having three 60° angles. Referring to **FIG. 5**, a tetrahedron whose sides are the square root of three (A), has a height of the square root of two (B), and base length equal to the square root of one (C). This forms the basic number set $\{\sqrt{1}, \sqrt{2}, \sqrt{3}\}$.

[0015] Referring to **FIG. 6**, the tetrahedron (A) is circumscribed by the sphere (B). Rod (C) is the sphere radius. A second rod (D), of equal length to rod (C), from the center of the sphere to the corner of the tetrahedron makes an angle ϕ of

$$\phi = \text{ArcSin}\left(\frac{1}{3}\right) = 19.47122063^\circ$$

So the four corners of the tetrahedron touch the sphere.

[0016] This tetrahedral geometry can be seen throughout the planets of the solar system Referring to **FIG. 7**, the islands of the Caribbean curve down from Puerto Rico to Venezuela forming an island vortex. The low density hyperspace energy releasing from the corner of the tetrahedron softens the rock mantle. The hot magma then rises through the rock with the least resistance. This creates a circular arc of volcanic islands along the edge of the vortex.

[0017] Referring to **FIG. 8**, the Giant Red Spot of Jupiter is located at a southern latitude of 19.5°. This vortex is so large that the entire earth can fit in it.

[0018] Referring to **FIG. 9**, the Olympus Mons volcano is located at a northern latitude of 19.5° as shown by the marker. This volcano is the size of France. Notice the fallen plume of volcanic debris toward the north east.

[0019] The double harmonic of the tetrahedral angle is twice 19.5° or 39° which is the location of the Silver Bridge in Point Pleasant, West Va. A large wormhole opened up around the bridge during Christmas rush hour when the bridge was full of cars. Due to the low density hyperspace energy, the rivets holding the cables down popped loose and all the cars were dumped into the river. A computer simulation using Schrodinger's quantum mechanics equation for a particle in a potential well shows that as the energy becomes less dense, the particle is no longer contained in the potential well. The electron jumps out. Thus the atomic bonds are broken which softens the rivets. This is the first time that there has been an understanding of the failure mechanism of this bridge.

[0020] After downloading from the Internet several pages of the index of refraction of a wide range of materials, I noticed that the index of refraction for Plexiglas was 1.50. Another source said it was 1.51. One of the Internet sites had a movable flashlight which showed the incident ray and the refracted ray. For Plexiglas, surprisingly enough, the incident ray was coming in at an angle of 60° to the normal, and the light was refracted at 35.26°, both of which are tetrahedral angles. The angle of the equilateral face of the tetrahedron is of course 60°. The angle at the top of tetrahedron is the arc-cosine of the ratio of the height over the edge length.

$$\theta = \mathrm{ArcCos}\left(\frac{\sqrt{2}}{\sqrt{3}}\right) = 35.26^\circ$$

According to Snell's law, the index of refraction n, times the sine of the angle $\sin(\theta_1)$ of the ray leaving material m_1, is equal to the index of refraction n_2 times the sine of the angle of refraction $\sin(\theta_2)$ of the ray entering material m_2. Referring to **FIG. 10**, the equation is

$$n_1 \sin(\theta_1)=n_2 \sin(\theta_2)$$

The index of air n_1 is equal to one. The index n_2 of Plexiglas is 1.50. If the incident ray is at $\theta_1=60^\circ$ to the normal, then the output angle is

$$\theta_2 = \mathrm{ArcSin}\left[\frac{n_1}{n_2}\mathrm{Sin}(\theta_1)\right] = \mathrm{ArcSin}\left[\frac{2}{3}\frac{\sqrt{3}}{2}\right] = 35.26^\circ$$

which is equal to the angle of the tetrahedron.

[0021] Then I recalled several months earlier that I had gone to the Subway restaurant to get a sandwich. I was sitting by the Plexiglas window communicating with the Admiral whose mothership was in earth orbit. She mentioned that they were bringing two people aboard. At that moment I looked out through the window and I could see both of them clearly and easily through my spiritual eye. To my amazement, I saw both of the captives start to pull out guns from behind their backs. I then projected by spiritual hands which resulted in preventing the attack on the security guards. To say the least, the Admiral was rather surprised at these events. She then asked me to look at the design of the window because she thought it had something to do with my enhanced remote viewing capabilities. Looking at the Plexiglas, I noticed that on the edge of the large window pane there was a shorter piece of Plexiglas which was mounted parallel to the window pane. This smaller panel acted to protect the yellow neon fluorescent tube. I took the measurements of the design using a piece of paper that I found near the table.

[0022] I then went home and designed up a mounting bracket with my 3D computer software. I had already installed the stereolithography software that converts the design to the *.STL file format. How stereolithography works is that it slices the design into many thin horizontal sections. The machine has a platform which is mounted in a bath of liquid polymer. An ultraviolet laser, mounted on an xy-table, then traces out the slice. Because the liquid polymer is light sensitive, it polymerizes immediately into solid plastic. Then the platform is lowered a few thousandths of an inch and the second slice is added. This process eventually builds up the complete 3D part. Using the Internet, the *.STL file is sent by e-mail to the stereolithography service provider who returns the part overnight. So the next day I had the piece from which I made a plastic mold and several additional pieces for mounting the Plexiglas bracket on a full sheet of Plexiglas that I ordered locally. It never occurred to me to measure the angles, so after I got the index of refraction for Plexiglas, I measured the incident angle and it turned out to be 61°. So then I realized that the spiritual eye was being diffracted across these two Plexiglas plates which connected it to the tetrahedral geometry of subspace. Because subspace is the foundation of space, this created a much more efficient route for remote viewing. The result is that this invention has allowed me to make contact with 430 alien civilizations. Since then I have been awarded the Aphysics prize for my work in the invention and elaboration of the tetrahedron diagram of which there are now over 4000 graphs. The scientific discoveries contained in the diagram are (1) the electron and proton are one and same particle, (2) the existence of hyperspace, (3) how mass can be taken out of dimension, (4) cosmology determines the elementary particles, (4) all the physics constants and the tetrahedral geometry are contained in the two 360° circles of the infinity symbol and (5) all the physics constants are determined geometrically and projected from a subspace manifold into our dimension. For my work in Revelations, I was awarded four beautiful galloping riderless white horses of the Apocalypse.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0023] **FIG. 1**. Remote viewing the top of a palm tree containing several coconuts.

[0024] **FIG. 2**. Remote viewing three soldiers running across a wooden bridge.

[0025] **FIG. 3**. Remote viewing computer building that soldiers were guarding.

[0026] **FIG. 4**. Spiritual eye of human aura.

[0027] **FIG. 5**. Tetrahedron.

[0028] **FIG. 6**. Tetrahedron circumscribed by sphere.

[0029] **FIG. 7**. Caribbean volcanic island vortex.

[0030] **FIG. 8**. Giant Red Spot on Jupiter.

[0031] **FIG. 9**. Olympus Mons volcano on Mars.

[0032] **FIG. 10**. Snell's Law of Refraction.

[0033] **FIG. 11**. Perspective view of remote viewing station.

[0034] **FIG. 12**. Wedge-shaped spacers for diffraction panel.

[0035] **FIG. 13**. Remote viewing angle.

[0036] **FIG. 14**. A remote viewing session taking place on a planet located 90,000 light years from earth on the other side of the galaxy showing His Highness signing with his left hand the document that allows his planet to become a member of the Pleiadian Federation.

DETAILED DESCRIPTION OF THE INVENTION

[0037] 1. Referring to **FIG. 11**, the remote viewing station is a rectangular box wooden frame (A) on which is mounted on one side a large sheet of Plexiglas (B). The diffraction panel (C), made of a shorter length of Plexiglas, is mounted with acrylic glue on the Plexiglas sheet using clear polyoptic molded plastic spacers (D).

[0038] Referring to **FIG. 12**, the wedge-shaped spacers (D) hold the diffraction panel (C) to the sheet of Plexiglas (B). The angle of the wedge is 30° which makes the incident angle 60° to the normal.

[**0039**] Referring to the top view perspective **FIG. 13**, sitting on the wide bench, the remote viewer can adjust his sight along the wedge for proper alignment at an angle of 60°. Due to the 1.50 index of refraction of Plexiglas, the spiritual eye is diffracted across the edge of the first panel and then refracted across the second panel at the tetrahedral angle of 35.26°.

[**0040**] Referring to **FIG. 14**, the remote viewing image is seen superimposed on the large sheet of Plexiglas which acts as the viewing screen.

I claim:

1. A remote viewing station comprising:

(a) a rectangular box frame made of wood having a length of six feet, a width of four feet and a height of six feet;

(b) a large sheet of quarter inch Plexiglas, having an index of refraction of 1.50, mounted on the right side of item (1a);

(c) several wedge-shaped clear plastic spacers, three inches in width and eight and a half inches in length, having a wedge angle of 30° that are mounted on the interior right side of item (1b);

(d) a shorter sheet of quarter inch Plexiglas, having a length of one and a half feet, mounted on item (1c) parallel to item (1b);

(e) a wide bench on which the remote viewer sits so that the remote viewer can align his sight along the wedge angle of item (1c);

2. A remote viewing amplifier that:

(a) diffracts the spiritual eye of the remote viewer across the edge of item (1d) at an incident angle of 60° to the normal;

(b) refracts the spiritual eye at 35.26° to the normal due to the refractive index of item (1a) according to Snell's Law of Refraction; and

(c) aligns the spiritual eye with the tetrahedral geometry of subspace due to the diffraction/refraction combination of item (2a) and (2b).

* * * * *

US 20060145019A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0145019 A1**

(43) **Pub. Date: Jul. 6, 2006**

(54) **TRIANGULAR SPACECRAFT**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/017,093**

(22) Filed: **Dec. 20, 2004**

Publication Classification

(51) **Int. Cl.**
B64G 1/40 (2006.01)

(52) **U.S. Cl.** **244/171.5**

(57) **ABSTRACT**

A spacecraft having a triangular hull with vertical electrostatic line charges on each corner that produce a horizontal electric field parallel to the sides of the hull. This field, interacting with a plane wave emitted by antennas on the side of the hull, generates a force per volume combining both lift and propulsion.

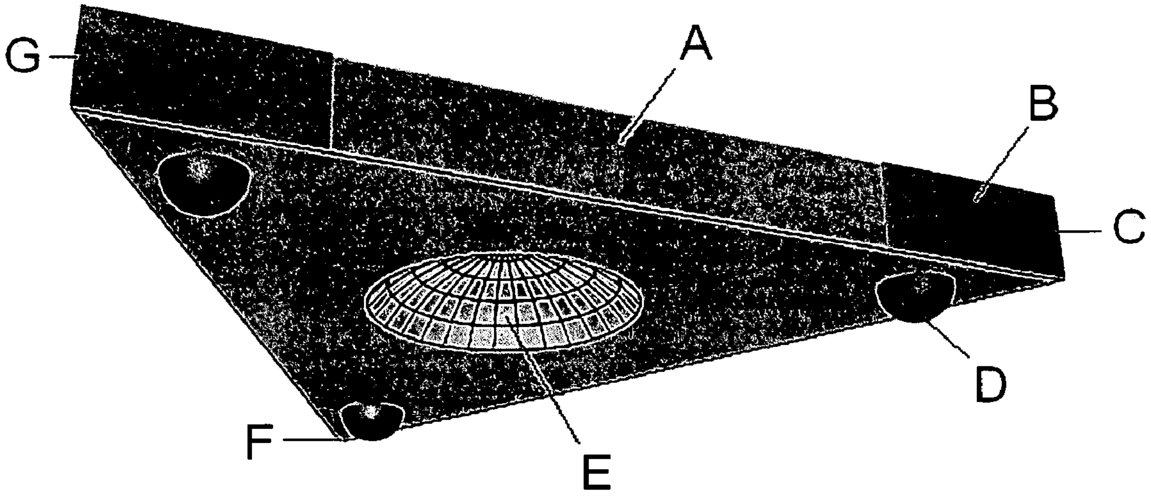

Figure 1

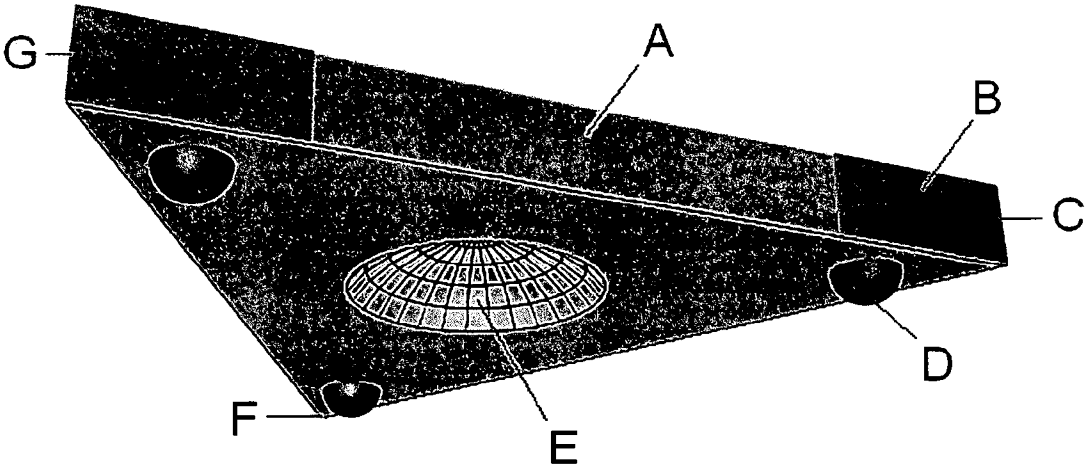

Figure 2

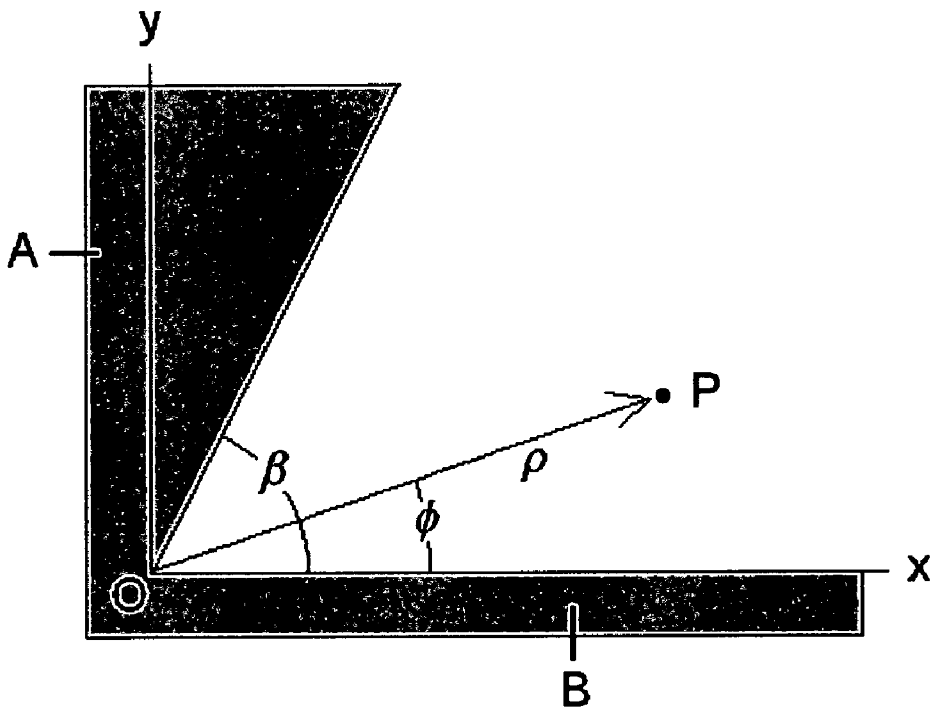

Figure 3

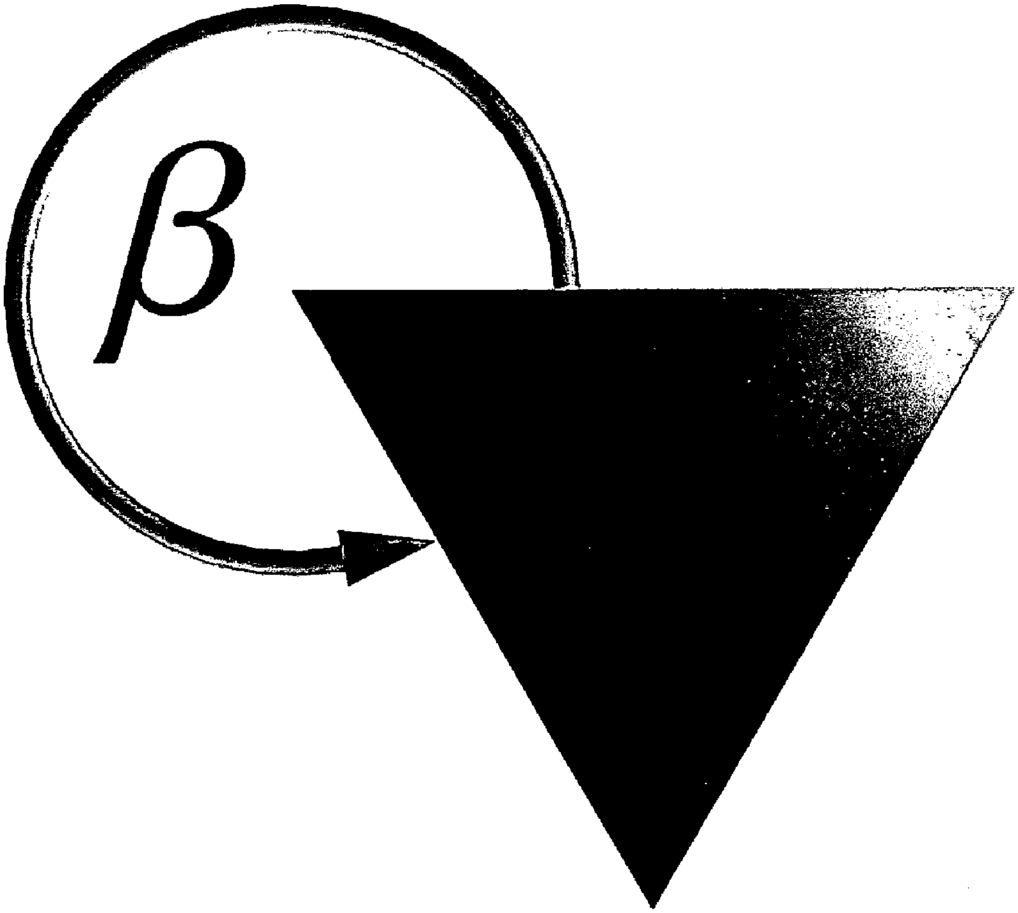

Figure 4

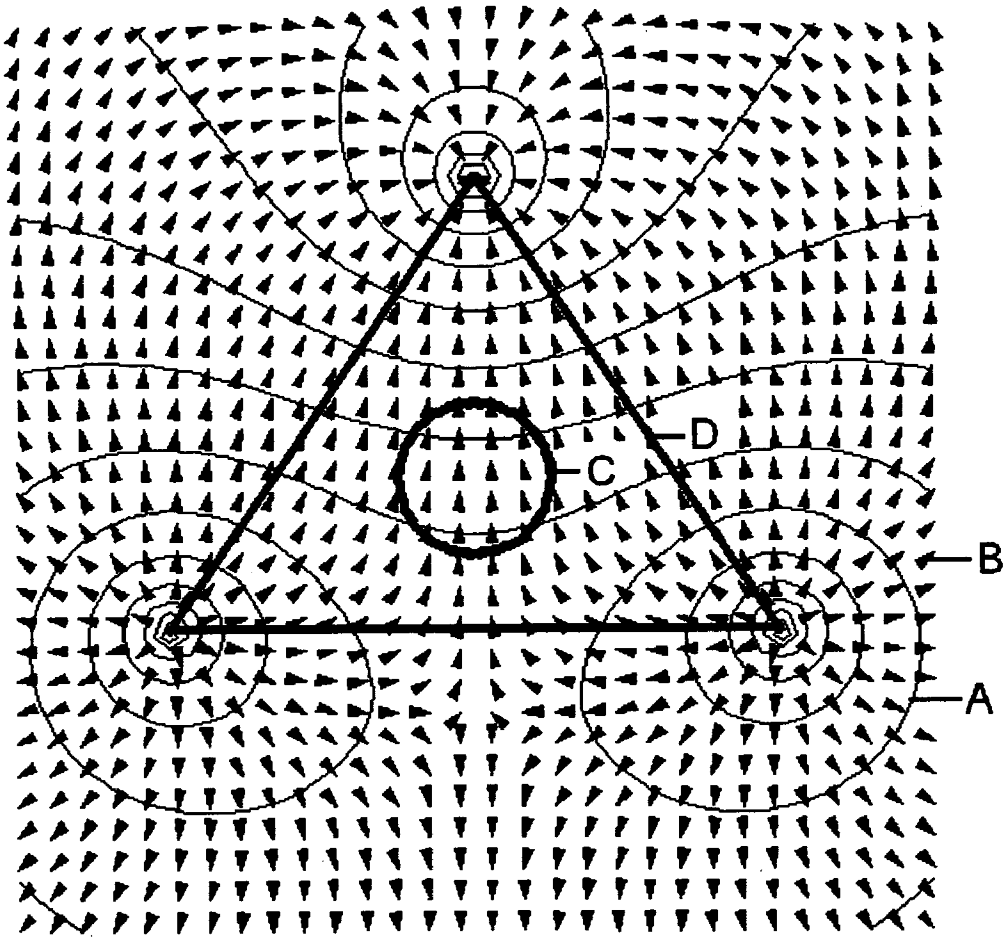

Figure 5

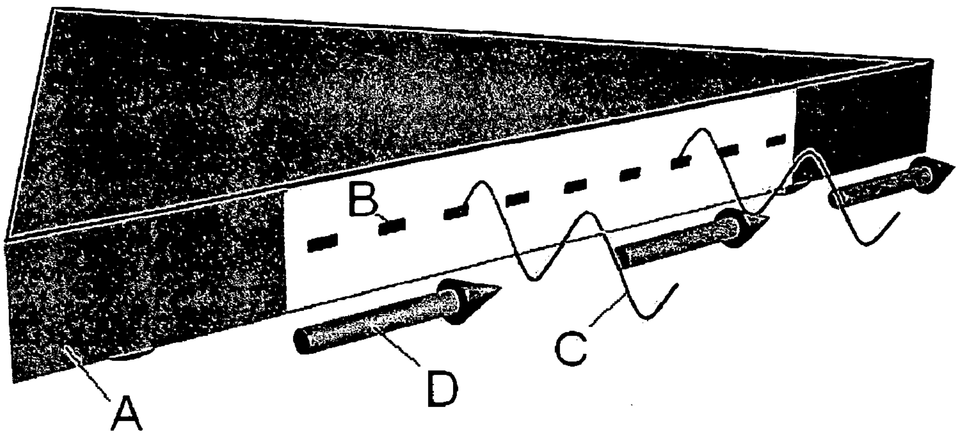

Figure 6

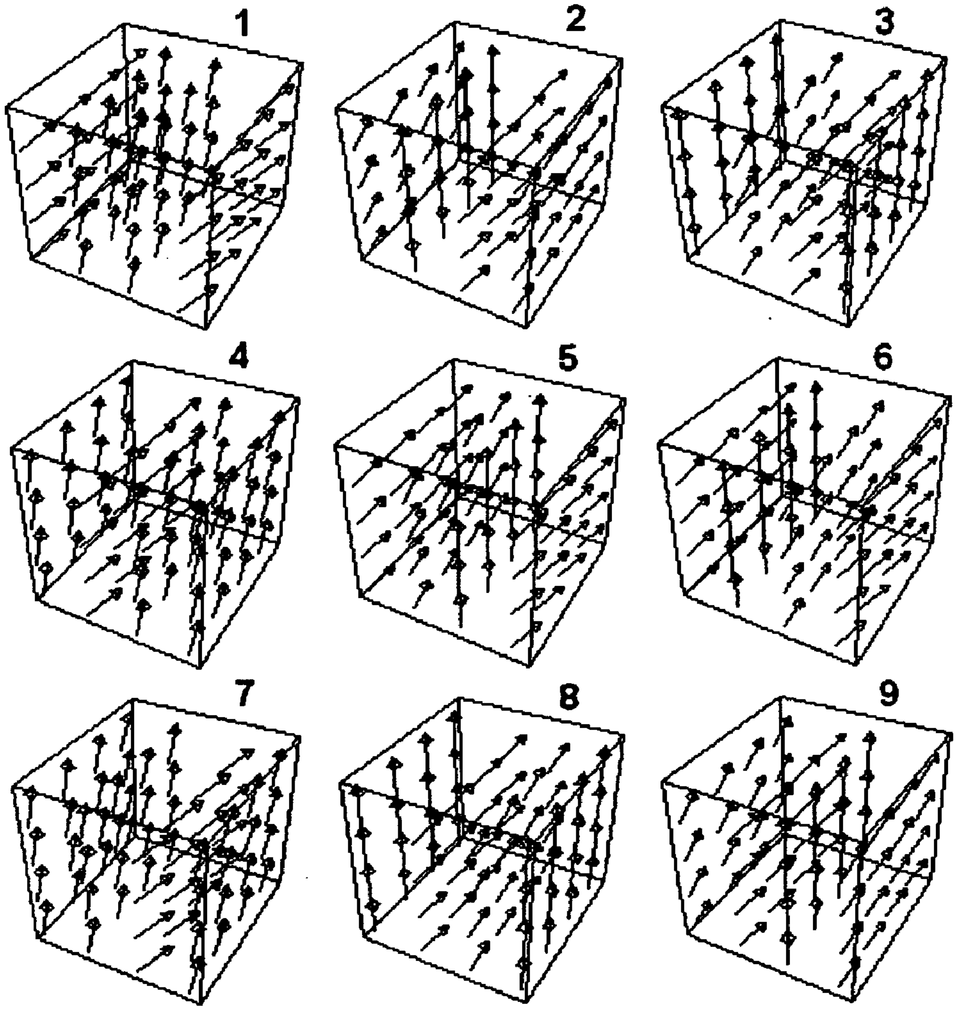

Figure 7

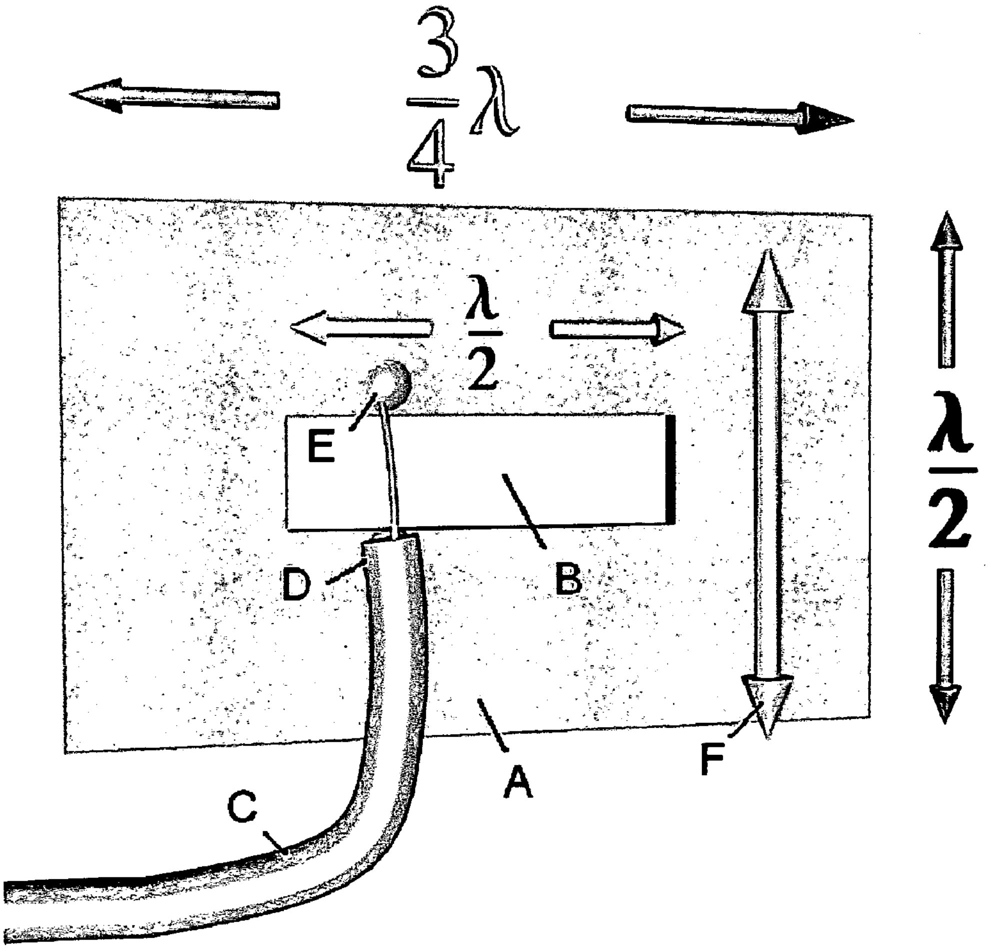
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
$\frac{\lambda}{2}$
E
B
D
C
A
F

TRIANGULAR SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft having a triangular hull with vertical electrostatic line charges on each corner. The line charges create a horizontal electric field that, together with a plane wave emitted by antennas on the side of the hull, generates a force per volume providing a unique combination of both lift and propulsion.

BACKGROUND OF THE INVENTION

[0002] Referring to **FIG. 1**, the spacecraft has a hull in the shape of an equilateral triangle. A parabolic antenna (E) is centrally located in the bottom of the hull. An array of horizontal slot antennas is located along the side of the hull (A). Each back corner (F,G) has a corner conducting plate which is charged to a positive voltage +V. The forward corner (C) has a conducting plate charged to a negative voltage −V. A motion control hemisphere (D) is located on the bottom surface in each of the three corners.

[0003] Referring to **FIG. 2**, two planes (A,B) intersect at the origin O at an opening angle β. Each plane (x,y) is charged to a voltage V. The potential at point P is determined in polar coordinates {ρϕ}. The Laplace equation for the potential Φ in polar coordinates is given by:

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\Phi}{\partial\phi^2}=0$$

Using a separation of variables solution, the potential is given as the product of two functions:

$$\Phi(\rho,\phi)=R(\rho)\Psi(\phi)$$

which when substituted into the Laplace equation becomes:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)+\frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=0$$

Since the two terns are separately functions of ρ and ϕ respectively, each one has to be constant with the sum of the constants equal to zero:

$$\frac{\rho}{R}\frac{d}{d\rho}\left(\rho\frac{dR}{d\rho}\right)=v^2 \quad \frac{1}{\Psi}\frac{d^2\Psi}{d\phi^2}=-v^2$$

These two equations have solutions:

$$R(\rho)=a\rho^{v+b\rho-v}$$

$$\psi(\phi)=A\cos(v\phi)+B\sin(v\phi)$$

The azimuthal angle ϕ is restricted to a value in the range $0\leqq\phi\leqq\beta$. The boundary condition is that the potential Φ is equal to V for any radius ρ when ϕ=0 and ϕ=β. This means that v has to be an integer value of π so that the sine function is zero:

$$\sin(v\beta)=\sin\left(\frac{m\pi}{\beta}\beta\right)=\sin(m\pi)=0 \quad m=1,2\ldots$$

which in turn means that the coefficient A of the cosine term has to be zero in the solution above. Choosing b=0 makes the general solution for the potential equal to:

$$\Phi(\rho,\phi)=V+\sum_{m=1}^{\infty}a_m\rho^{m\pi/\beta}\sin(m\pi\phi/\beta)$$

which shows that when the angle is zero, the sine is zero and the potential is V. If the angle is β, then there is a multiple of π such that the sine is zero again.

[0004] Because the series involves positive powers of the radius, for small enough ρ, only the first term *m*=1 in the series is important. Thus around ρ=0, the potential is approximately

$$\phi(\rho,\phi)\approx V+a,\rho^{\pi/\beta}\sin(\pi\phi/\beta)$$

[0005] The electric field component is the negative gradient of the potential:

$$E_\phi(\rho,\phi)=-\frac{1}{\rho}\frac{\partial\Phi}{\partial\phi}=-\frac{\pi a_1}{\beta}\rho^{(\pi/\beta)-1}\cos(\pi\phi/\beta)$$

The surface charge distribution σ at ϕ=0 and ϕ=β is equal to the electric field perpendicular to the surface times the permittivity of space ϵ_0:

$$\sigma(\rho)=\varepsilon_0E_\phi(\rho,0)=-\frac{\varepsilon_0\pi a_1}{\beta}\rho^{\frac{\pi}{\beta}-1}$$

Notice that if angle of intersection β is less than π, then the equation says that there is a very small radius to a positive power which means little charge density accumulation.

[0006] Referring to **FIG. 3**, the value of β, in the case of the triangular hull, is equal to 360° less 60° for a total of 300° or:

$$\beta=\frac{300}{180}\pi=\frac{5}{3}\pi$$

$$\rho^{\frac{\pi}{\frac{5}{3}\pi}-1}=\frac{1}{\rho^{\frac{2}{5}}}$$

which says that there is a charge density singularity to the two fifths power for small radius. Thus, the corner plates on the hull create a huge line charge density along the sharp vertical corner edge. The equation for the potential of a line charge density is given as:

$$\Phi(x,y)=-\frac{\lambda}{2\pi\varepsilon_0}\mathrm{Ln}((x-x_0)^2+(y-y_0)^2)$$

where λ is the charge per unit length in the vertical z-direction, and x_0 and y_0 are the location of the line charge in the xy-plane.

[0007] Referring to **FIG. 4**, the triangular hull (D) is plotted together with the potential contours (A) and the electric field arrows (B) created by the three corner line charges. The line charges are perpendicular to the paper. Notice that the electric field arrows are parallel crossing the center parabolic antenna (C). The electric field is also parallel to the sides (D) of the triangle.

[0008] Referring to **FIG. 5**, along the side of the triangle (A), an array (B) of horizontal slot antennas emit electromagnetic waves that have a vertically polarized electric E field (C). These traveling waves interact with the electric field (D) produced by the line charges on the corners of the triangle.

[0009] Using differential forms mathematics, this combination of fields is represented by the Hodge star of the differential of the wedge product of the two fields. The antenna electromagnetic field is a combination of a traveling magnetic field B_w, and electric field E_w. The stationary field E created by the line charges is perpendicular to the traveling wave.

$$*\,d(E \wedge (B_w + E_w \wedge dt))\frac{\varepsilon}{c} = \frac{\text{force}}{\text{volume}}$$

where ϵ is the linear capacitance of space and c is the speed of light. Thus there is a force per volume around the hull.

[0010] This combination of fields produces a spacetime curvature as determined by Einstein's General Theory of Relativity. The traveling electric field has an amplitude in the vertical z-direction and travels in the x-direction

$$E_{\mathrm{w=Ez}}\cos(x-t)$$

The Faraday electromagnetic tensor contains all the electric and magnetic fields in all the {x,y,z} directions. The first row and first column contain the two electric fields

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & E_x & 0 & E_z\cos(x-t) \\ x & E_x & 0 & 0 & 0 \\ y & 0 & 0 & 0 & 0 \\ z & E_z\cos(x-t) & 0 & 0 & 0 \end{array}$$

The stress exerted on spacetime occurs in the xx, yy and zz-direction as calculated from the stress-energy tensor T of gravitational physics

$$4\pi T^{\mu\nu} = F^{\mu\alpha}F^{\mu}_{\alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor for Cartesian space

$$g_{\alpha\beta} = \begin{array}{c|cccc|} t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

where the diagonal components are the coefficients of the elementary spacetime length ds squared

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

The calculation produces three stresses T^{xx},T^{yy} and T^{zz} in their respective {x,y,z} directions.

[0011] Referring to **FIG. 6**, these three stresses are plotted together as a **3**D vector field animated over time in nine frames. The graphs show that there is a lift force as depicted by the vertical arrows as well as a force of propulsion as shown by the interspersed horizontal arrows. With the passage of time, these vectors exchange places with each other so that the lift becomes the propulsion and vice versa, creating a wavy stress-energy field around the hull.

SUMMARY OF THE INVENTION

[0012] This invention is a spacecraft with a triangular hull having charged flat plates on the vertical corners of the three sides. The two rear corners are charged to a potential V. The forward corner is charged to a potential −V. The 60° angle on the corner creates a line charge density singularity that produces a huge horizontal electric field pointing from the back to the front of the craft which is also parallel to the sides of the triangle. An array of horizontal slot antennas located on the sides of the triangular hull produce an electromagnetic wave with the electric field polarized in the vertical direction. This combination of fields produces a spacetime force in both the vertical and horizontal directions such that the spacecraft receives a lift force and a force of propulsion.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0013] **FIG. 1**. Perspective view of triangular spacecraft.

[0014] **FIG. 2**. Drawing of the intersection of two charged plates in order to calculate the charge density in the corner.

[0015] **FIG. 3**. Perspective view of the corner angle β for the equilateral triangle.

[0016] **FIG. 4**. Planar **2**D graph showing the electric field produced by three line charges on the corners of the triangular hull.

[0017] **FIG. 5**. Perspective view of electric field produced by the linear charge interacting with the traveling electromagnetic wave produced by the slot antenna.

[0018] **FIG. 6**. **3**D vector animation of the lift and thrust force generated by the fields.

[0019] **FIG. 7**. Perspective view of slot antenna.

DETAILED DESCRIPTION OF THE INVENTION

[0020] Referring to **FIG. 7**, the antenna (A) is made out of sheet copper in which a rectangular horizontal slot (B) has been notched out using a die press and sheet metal fixture. A coaxial cable from the amplifier and frequency generator is attached across the slot by soldering the outer cable (D) to one side of the slot and the inner cable (E) to the other side of the slot. This creates the positive and negative charges across the gap which forms the vertical electric field (F) which radiates out perpendicularly to the copper sheet.

[0021] Although the invention has been described with reference to specific embodiments, such as a particular antenna system, those skilled in the art will appreciate that many modifications and variations are possible without departing from the teachings of the invention. All such modifications and variations are intended to be encompassed within the scope of the following claims.

1. A spacecraft comprised of the following components:

(a) a triangular hull in the form of an equilateral triangle;

(b) two copper plates attached on opposite vertical sides at each of the three corners of the hull (**1***a*) such that a sharp vertical edge is formed where they come together;

(c) an electrostatic generator used to charge the back two copper-cladded corners (**1***b*) to a high positive voltage, and the third forward copper-cladded corner to a high negative voltage;

(d) a horizontal slot antenna array mounted-on the sides of the hull; and

(e) a frequency generator, antenna and coaxial cables to drive the antenna array (**1***d*).

2. To create, by claims (**1***a*, **1***b*, **1***c*), an intense vertical line charge at the corners (**1***b*) and a horizontal electric field that that is parallel to the sides of the hull (**1***a*);

3. To create, by claims (**1***d*,**1***e*), an electromagnetic wave with a vertically polarized electric field traveling outward from the side of the hull (**1***a*); and

4. To create, by claims (**2**,**3**), an interaction of the electrostatic field (**2**) with the electromagnetic wave (**3**) such that a combined spacetime curvature pressure is generated on the hull in the upward and forward direction to produce lift and propulsion respectively.

* * * * *

US 20060112848A1

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0112848 A1**

(43) **Pub. Date: Jun. 1, 2006**

(54) **PERMANENT MAGNET PROPULSION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/001,217**

(22) Filed: **Dec. 1, 2004**

Publication Classification

(51) **Int. Cl.**
B61C 3/00 (2006.01)

(52) **U.S. Cl.** **105/49**

(57) **ABSTRACT**

This invention is a propulsion system for a train that uses permanent magnets mounted on a rotating iron cylindrical plate carrying a radial current in order to create a spacetime curvature distortion which pulls the locomotive along the track.

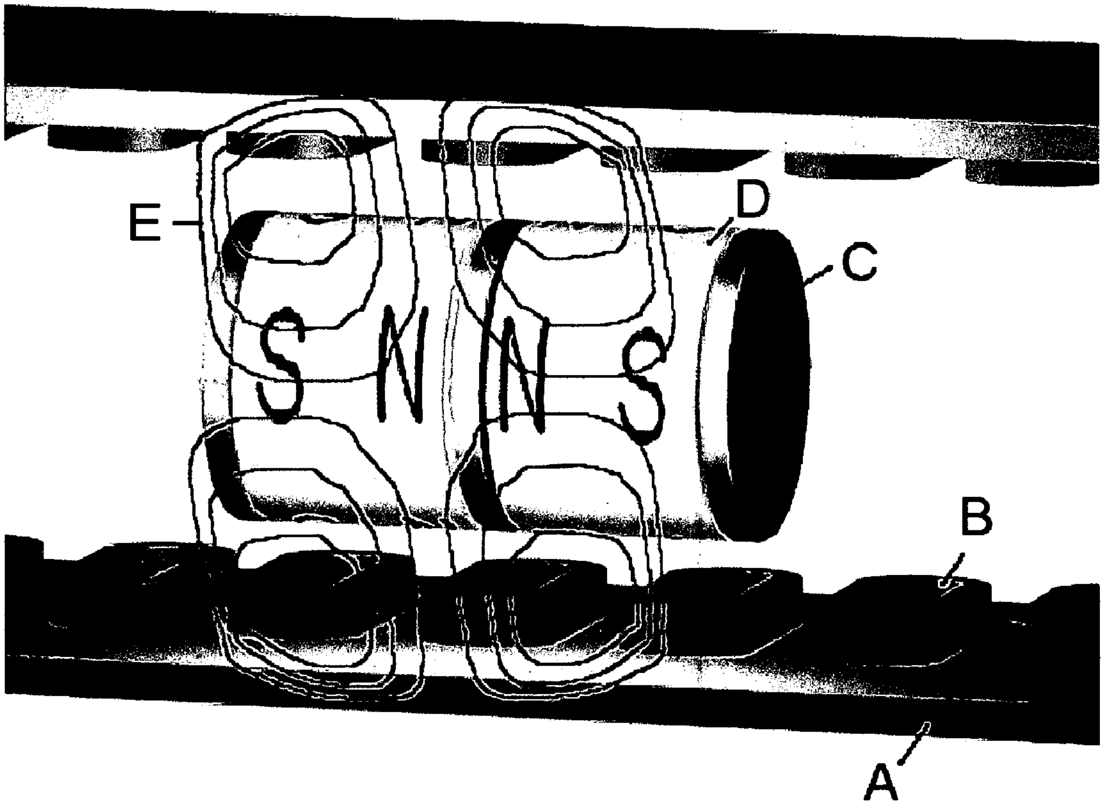

Figure 1

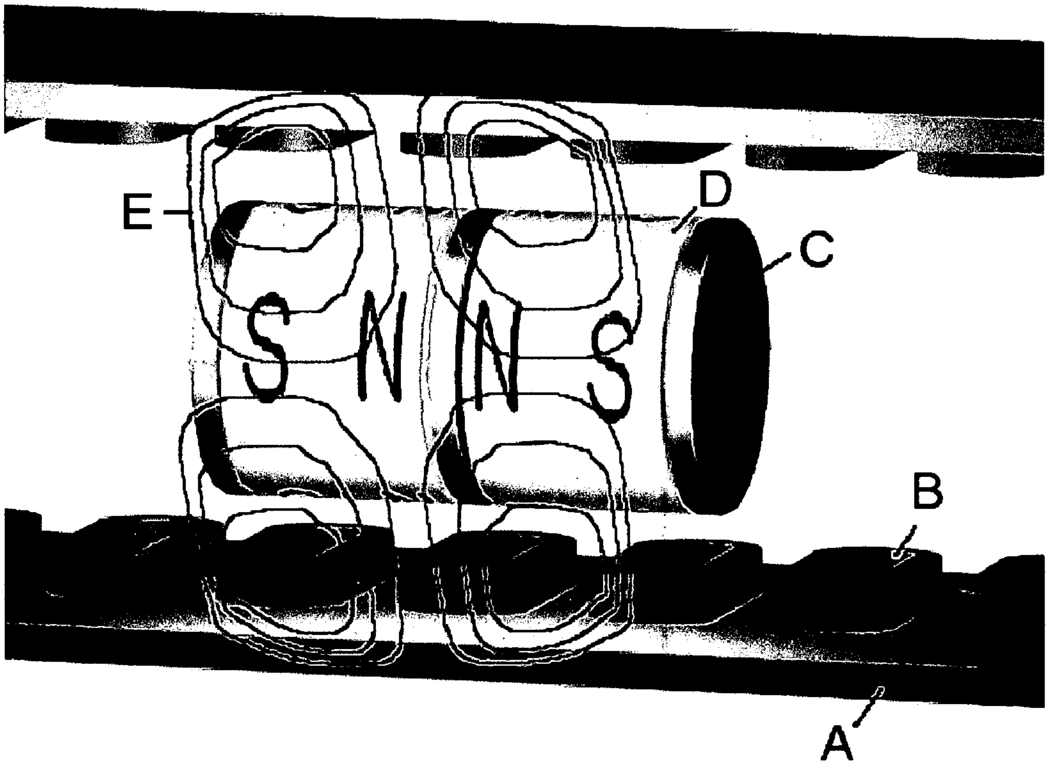

Figure 2

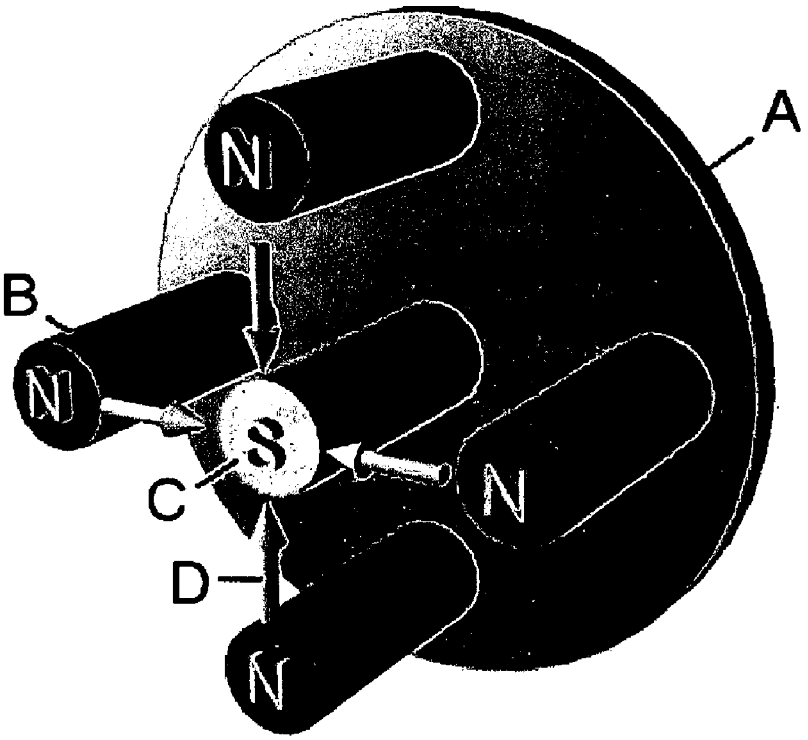

Figure 3

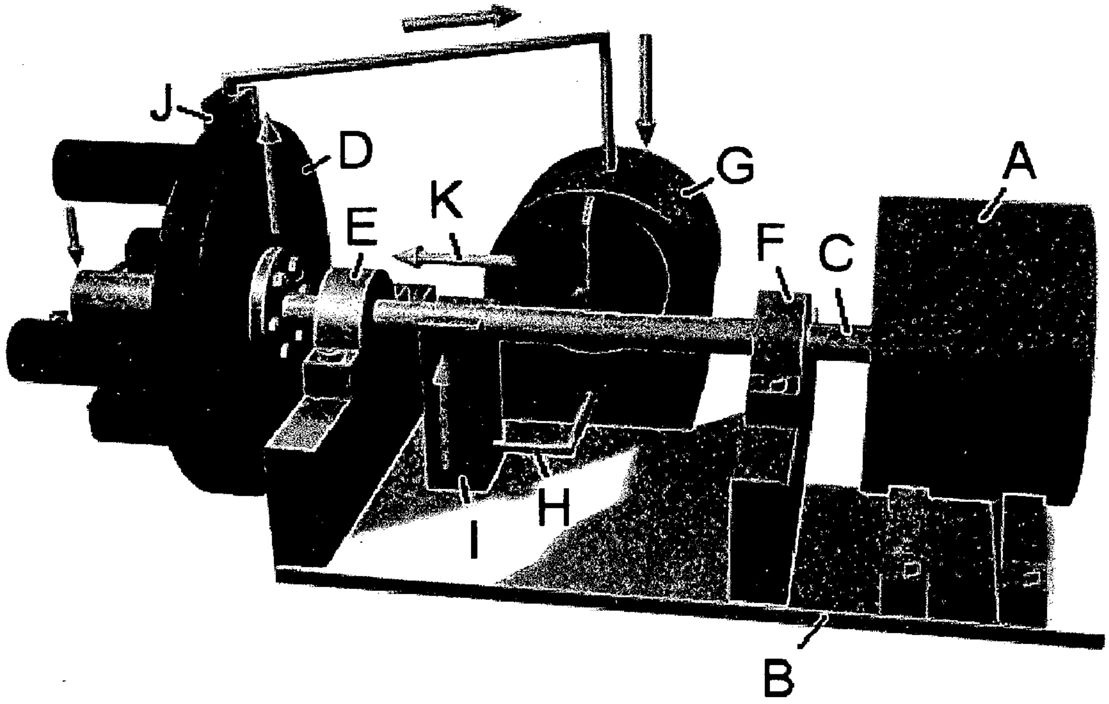
J
D
K
E
G
A
F
C
I
H
B

Figure 4

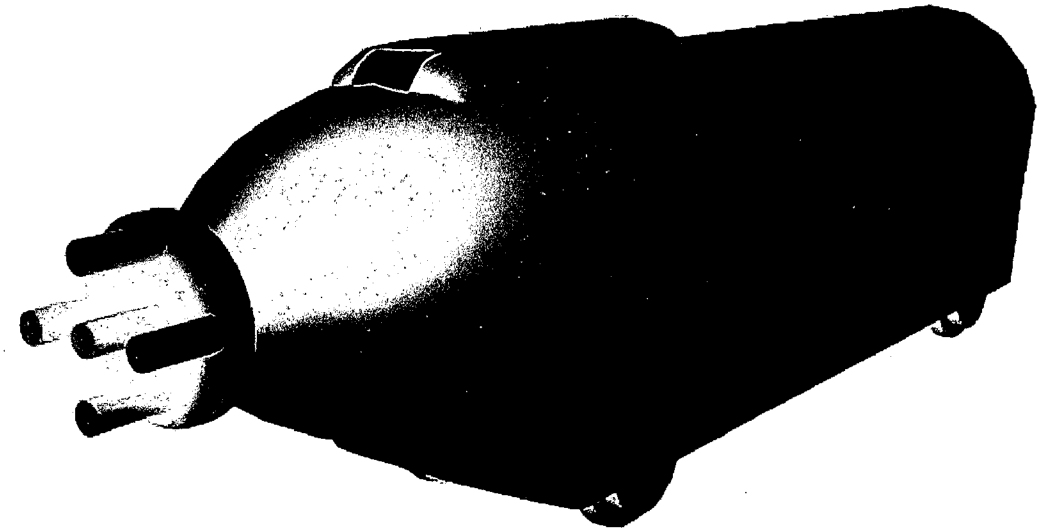

PERMANENT MAGNET PROPULSION SYSTEM

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a propulsion system for a train that utilizes spinning cylindrical magnets in order to create a spacetime pressure distortion ahead of the vehicle that pulls the locomotive along the track.

BACKGROUND OF THE INVENTION

[0002] At the present time, referring to **FIG. 1**, proposed permanent magnet propulsion systems use a dual railway track (A) supporting a series of coil windings (B) located along the track. The vehicle is attached to two permanent magnets (D) between steel pole pieces (C). The north pole of each magnet faces the interior pole piece such that the magnetic flux path (E) follows the center pole piece up through the railway bed and then back to the south pole of the magnet. As the magnets move along the track, the coil windings are activated at the correct time by Hall sensors. With the coil energized as a north pole, the permanent magnet north pole is repelled which drives the vehicle along the track The problem with this design, and other similar designs, is that it is not practical to wind huge numbers of sensor-activated electrical coils along a steel track.

[0003] From Einstein's General Theory of Relativity, it is known that a spacetime curvature pressure develops perpendicular to direction of vibration of the electric and magnetic field. As an example, the photon has an electric field vibrating in the vertical y-direction and a magnetic field vibrating in the horizontal x-direction. The spacetime curvature pressure is therefore along the z-axis of radiation which pushes the negative mass of the photon along. Thus in order to create a spacetime curvature pressure in the z-direction along the track which would pull the train forward, a magnetic flux density field is required in the radial direction.

[0004] Referring to **FIG. 2**, four equally-spaced north permanent magnets (B) surrounding a centrally-located south permanent magnet (C) are mounted on an iron cylinder which acts as the radial flux return path. The magnetic flux density field (D) is in the radial direction from the north pole to the south pole. In order to provide strength, the magnets are molded onto a steel shaft and coated with epoxy so that they don't rust. During the molding process, a capacitor-discharge magnetizer is used to create the magnetic field of the magnet.

[0005] In Cartesian coordinates {−ct,x,y,z}, the elemental spacetime length ds squared is the sum of the squares of the incremental lengths {cdt,dx,dy,dz}

$$(ds)^2=-(dt)^2+(dx)^2+(dy)^2+(dz)^2$$

where the speed of light c is unity. The coefficients (−1,1,1,1) of this equation make up the g metric 4×4 tensor

$$g_{\alpha\beta} = \begin{array}{c|cccc|} & [t & x & y & z] \\ t & -1 & 0 & 0 & 0 \\ x & 0 & 1 & 0 & 0 \\ y & 0 & 0 & 1 & 0 \\ z & 0 & 0 & 0 & 1 \end{array}$$

[0006] The Faraday electromagnetic tensor contains the magnetic fields which determine how the spacetime length ds is curved. For a magnetic flux density field in the x-direction, Bx, and a magnetic flux density field in the y-direction, By, the Faraday tensor is

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} & t & x & y & z \\ t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & -By \\ y & 0 & 0 & 0 & Bx \\ z & 0 & By & -Bx & 0 \end{array}$$

The stress-energy-momentum tensor T, which determines how space is curved, is calculated from the following equation

$$4\pi T^{\mu\nu} = F^{\mu\alpha}F^{\nu}_{\alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

The stress-energy in the z-direction ahead of the locomotive is

$$T^{zz} = \frac{B_x^2 + B_y^2}{8\pi} = \frac{B_r^2}{8\pi}$$

where the sum of the squares of the fields in the x and y directions is the radial B field. In Einstein's General Relativity Theory, the curvature G tensor is equal to the stress-energy tensor divided by 8π. The G tensor is the curvature of space having units of inverse radius squared.

$$G = \frac{T}{8\pi}$$

Therefore the curvature G_{zz} generated along the z-direction ahead of the train is proportional to the square of the magnetic flux density field

$$G_{zz} = \frac{1}{r^2} = \frac{G\varepsilon}{c^2}\frac{B_r^2}{8\pi} = \frac{1}{meter^2}$$

where G is Newton's gravitational constant (not to be confused with the curvature tensor), ε is the linear capacitance of space, and c is the speed of light. The linear mass of space Ω is the speed of light c squared divided by the gravitational constant G, so that the equation can be written as

$$\frac{G\varepsilon}{c^2}\frac{B_r^2}{8\pi} = \frac{\varepsilon}{\Omega}\frac{B_r^2}{8\pi} = \frac{1}{\frac{\Omega}{\varepsilon}}\frac{B_r^2}{8\pi}$$

where the conversion factor is the square of the magnetic vector potential A

$$\sqrt{\frac{\Omega}{\varepsilon}} = \frac{\text{kgm}}{seccoul} = A$$

which is actually the momentum per charge. Therefore the curvature equation can be written as

$$\frac{1}{r^2} = \frac{1}{8\pi}\left(\frac{B_r}{A}\right)^2$$

This equation shows that it is necessary to create a magnetic vector potential together with the radial magnetic flux density field in order to create a curvature of space. Looking at the units of A shows that it is a mass momentum per charge

$$A = \frac{\text{kg}}{sec}\frac{m}{coul} = \frac{m\omega^2 r}{I}$$

or a mass m rotating with angular velocity c) per current along the radius. In terms of the invention, what this means is that the mass of the iron cylinder has to be rotating and there has to be a radial electrical current I in order to produce the linear charge along the radius. The differential mass dm depends on the circumference times the differential radius dr, the mass density p, and the length L of the cylinder

$$dm = \rho 2\pi rLdr$$

so that the magnetic vector potential becomes

$$A = \int_0^R \frac{\rho 2\pi rL\omega^2 r}{I}\,dr = \frac{2}{3}R^3\rho\pi L\frac{\omega^2}{I}$$

The value of A for the iron cylinder is

$$L = .2m$$

$$\rho = 7866\frac{kg}{m^3}$$

$$R = 1m$$

$$\omega = 2\pi f = 6.28\ \text{sec}^{-1}$$

$$I = 3000000\ \text{amp}$$

$$A = .04335\frac{kgm}{\text{sec}coul}$$

$$Br = 1.2tesla$$

$$\frac{1}{8\pi}\left(\frac{Br}{A}\right)^2 = 30.47m^2$$

$$r_{curvature} = \sqrt{8\pi}\left(\frac{A}{Br}\right) = .181m$$

What makes this possible is that the new N-machines can easily generate a minimum of 6 million amps which is twice the value of the electrical current above.

[0007] Referring to **FIG. 3**, the assembly consists of a large induction motor (A) mounted on the train's base plate (B) driving a motor shaft (C) attached to the iron cylinder (D). The shaft is held in place by two thrust bearings mounted in two pillow blocks (E,F). The current-generating N-machine (G) is electrically connected by a copper bus (H) to a copper-beryllium brush (I) on the motor shaft with a similar return brush (J) on the edge of the iron cylinder. The current (K) flows through the motor shaft to the center of the rotating cylinder and then radially outward to the edge. The magnetic flux density flows from the north poles of the outer permanent magnets to the central south pole, along the central magnet to the center of the rotating cylinder and then radially outward to the south poles of the outer magnets.

[0008] The thrust F developed is the radius of curvature of spacetime r_c calculated above times the magnet flux density field times the current I

$$F = \frac{r_c B_r I}{\sqrt{8\pi}} \approx 30000lbf$$

Using conservation of tensor coordinates, the radius of curvature is in the z-direction, the magnetic flux density field is in the radial direction and the current is in the radial direction

$$F^z = x^z B_r I^r$$

where the radial indices cancel, leaving the z-index as the direction of the force.

SUMMARY OF THE INVENTION

[0009] It is the object of this invention to create a spacetime curvature in front of a train locomotive in order to pull the vehicle along the track It is known from gravitational physics that a spacetime curvature is generated perpendicular to the direction of vibration of the electric and magnetic field. A radial magnetic field, which can be produced by permanent magnets attached to the flat faces near the rim of a iron cylinder rotating about the z-axis, will create a curvature in the z-direction. Four cylindrical north-pole-oriented magnets produce a radial magnetic flux density with is channeled into a central cylindrical south-pole-oriented magnet. The flux lines then flow radially outward through the steel rotating cylinder and reconnect with the south poles of the four outer magnets. The rotating iron cylinder generates the equivalent of a magnetic vector potential when an electrical current flows from the center of the cylinder to the edge. This current is generated by an N-machine current generator. The square of the magnetic flux density divided by the magnetic vector potential is equal to the spacetime curvature. The square root of the inverse of the spacetime curvature is the radius of curvature. The thrust developed is this radius of curvature times the magnetic flux density field times the current.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0010] **FIG. 1**. Perspective view of proposed permanent magnetic propulsion system using coil windings on the steel track.

[0011] **FIG. 2**. Perspective view of permanent magnet rotor assembly.

[0012] **FIG. 3**. Perspective view of system showing motor drive, N-machine and permanent magnet rotor.

[0013] **FIG. 4**. Perspective view of locomotive and rotor/magnet assembly.

DETAILED DESCRIPTION OF THE INVENTION

[0014] 1. The permanent magnets are made of neodymium-iron-boron material which is heated to its melt temperature and injection molded around a steel shaft threaded at one end while at the same time a pulsed magnetic field is applied to the material using a charge-discharge magnetizer. Because of the iron in the material, a coat of epoxy is applied to the magnet in order to protect it from the environment. Holes are drilled into the iron plate 90° apart near the rim, threaded, and then the steel shaft with the magnet is then inserted. Another hole is drilled and tapped in the center of the circular plate for attaching the south pole magnet which is used as the return path for the magnetic flux.

[0015] 2. Another easier way to make the magnets is to purchase short lengths of tubular NdFeB magnets and then stack them on the steel shaft with a cylindrical iron pole piece on the end of the shaft. The pole piece then holds the magnets down in place when the shaft is threaded into the plate.

[0016] 3. Referring to **FIG. 4**, the propulsion system is mounted inside the train cabin such that the rotor/magnet assembly extends out in front of the locomotive where the spacetime curvature is generated.

I claim:

1. A train propulsion system consisting of the following components:

a. a rotating iron cylindrical plate rotor of high relative permeability driven by an induction motor and horizontal steel motor shaft mounted in pillow block thrust bearings;

b. four cylindrical magnets, each molded to a steel support shaft threaded into the iron plate at 90° intervals around the rim of the plate with their north poles facing away from the plate;

c. a fifth cylindrical magnet molded to a steel support shaft which is threaded into the center of the iron plate with the south pole facing away from the plate;

d. an N-machine current generator supplying a radial electrical current from the center of the rotating plate by means of a copper-beryllium brush on the motor shaft (**1***a*) and another similar brush on the outside edge of the rotor.

e. a locomotive train on which the components are mounted such that the rotor/magnet assembly extends out in front of the locomotive with the rotor's angular velocity vector pointing along the track.

2. a closed magnetic flux path along a radial path in air from the north poles of the four outer magnets (**1***b*) to the south pole of the central magnet (**1***c*), through the center magnet and then radially outward through the rotor (**1***a*), returning back through the four outer magnets, such that the flux and electrical current (**1***d*) flow in the same outward radial direction through the rotor.

3. the creation of a spacetime curvature due to claims (**1***a* through **2**) that produces a large force on the locomotive equal to the radius of the spacetime curvature times the flux times the current.

* * * * *

US 20060180473A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0180473 A1**

St. Clair (43) **Pub. Date: Aug. 17, 2006**

(54) **WATER ENERGY GENERATOR**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/060,037**

(22) Filed: **Feb. 17, 2005**

Publication Classification

(51) **Int. Cl.**
C25D 5/34 (2006.01)

(52) **U.S. Cl.** **205/210**

(57) **ABSTRACT**

A water energy generator that generates electricity using a magnetic vortex wormhole generator and a water injector/vacuum chamber to produce low density hyperspace energy that causes the hydrogen atoms of water molecules to decay into electron pairs.

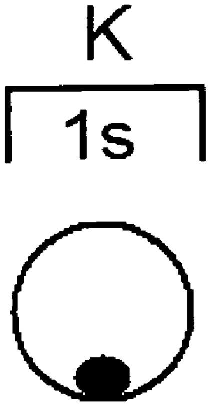

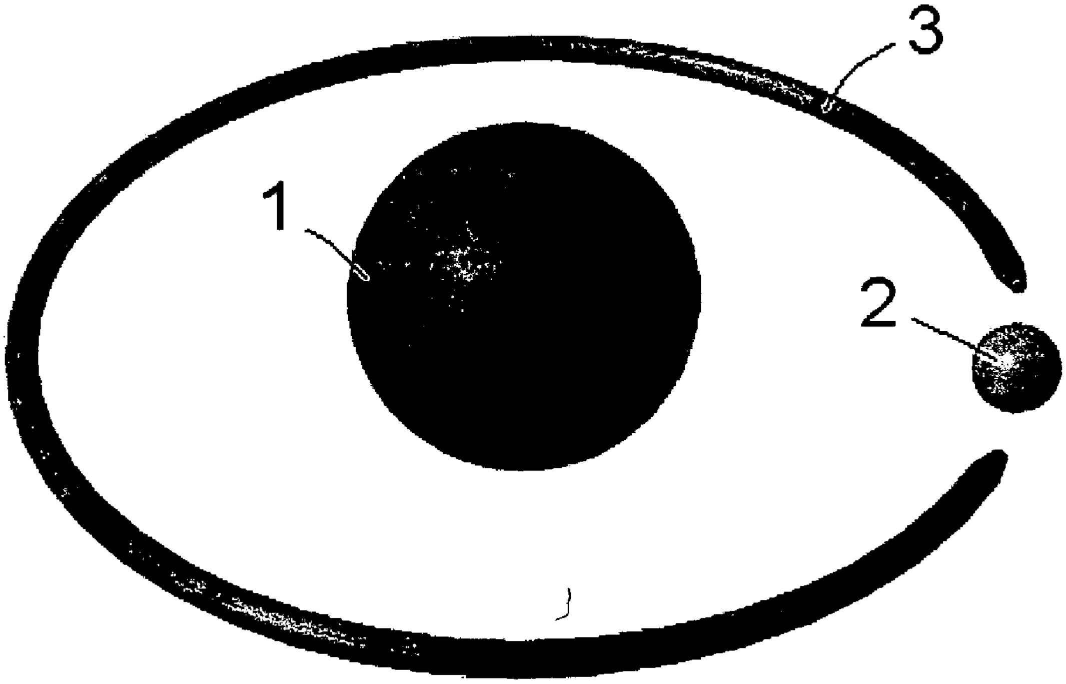

Figure 1

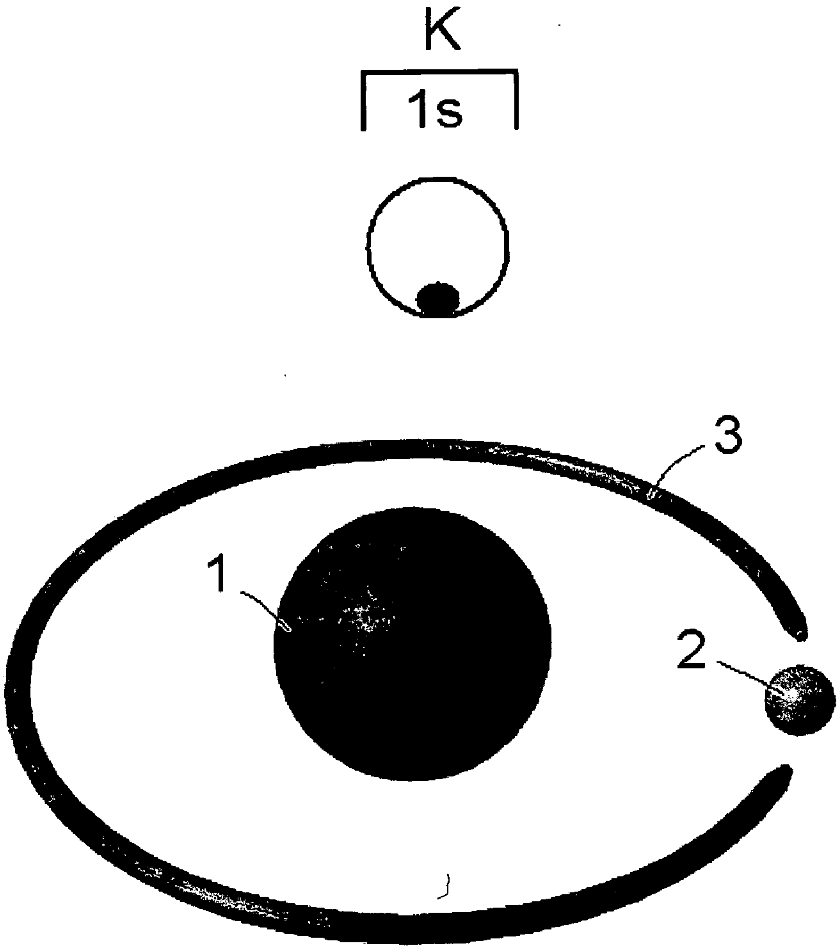

Figure 2

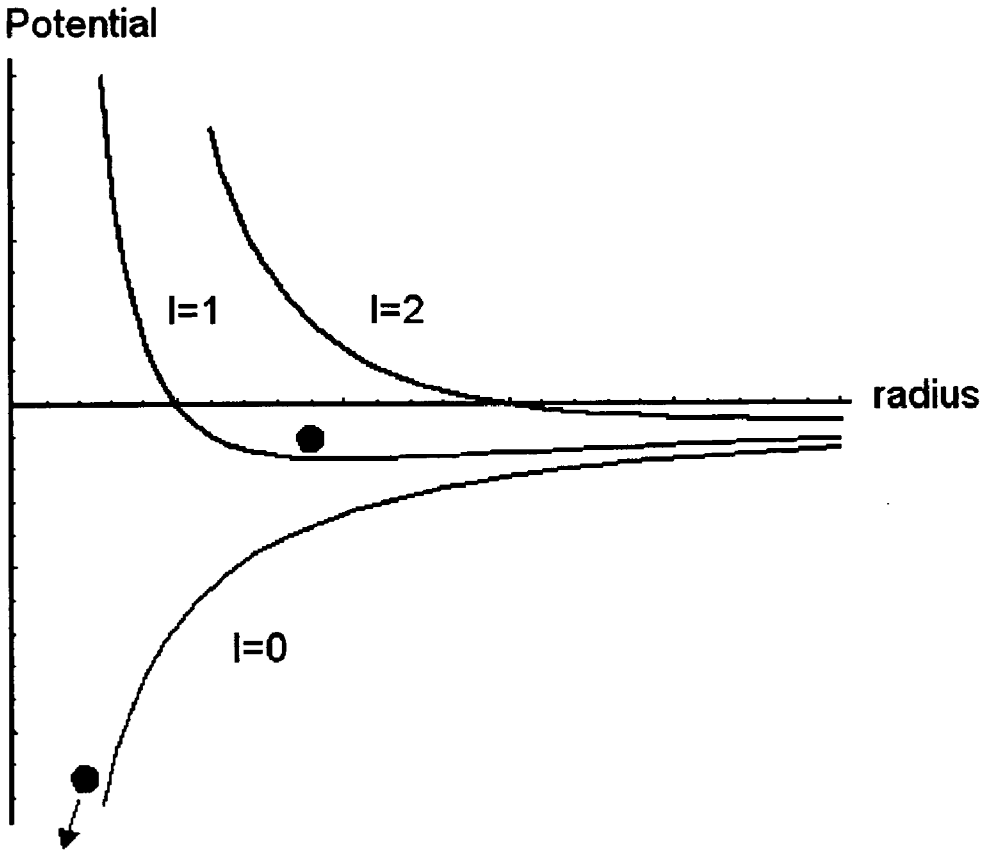

Figure 3

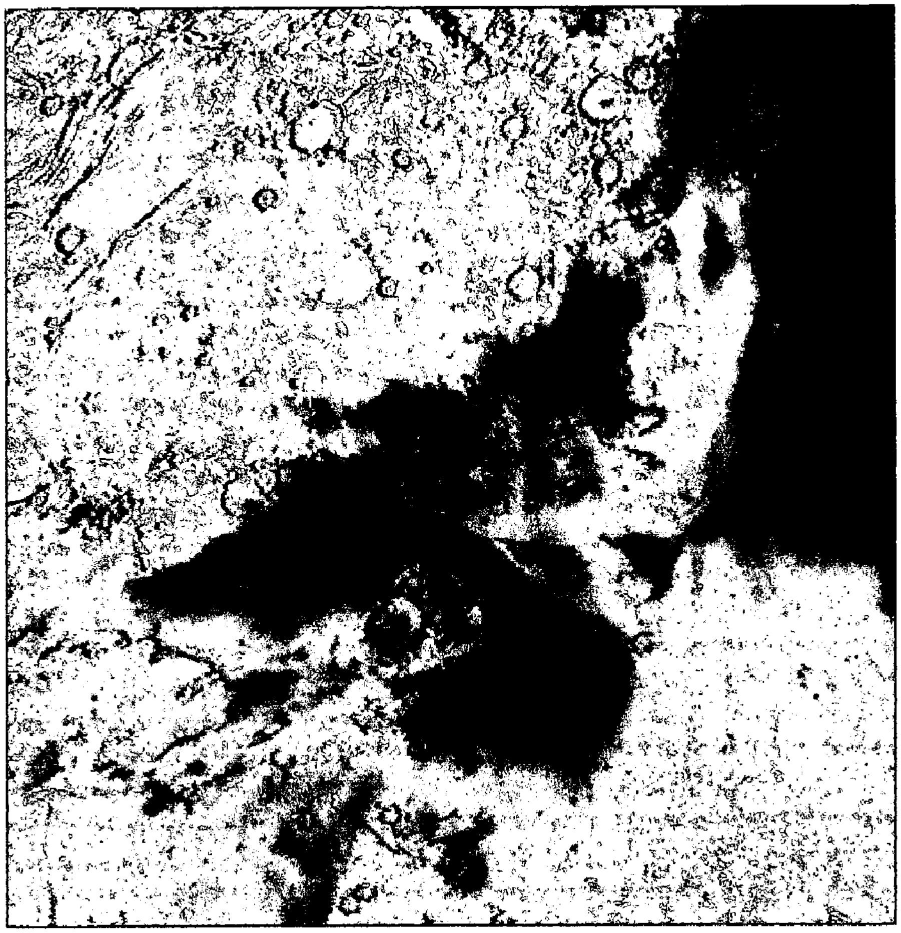

Figure 4

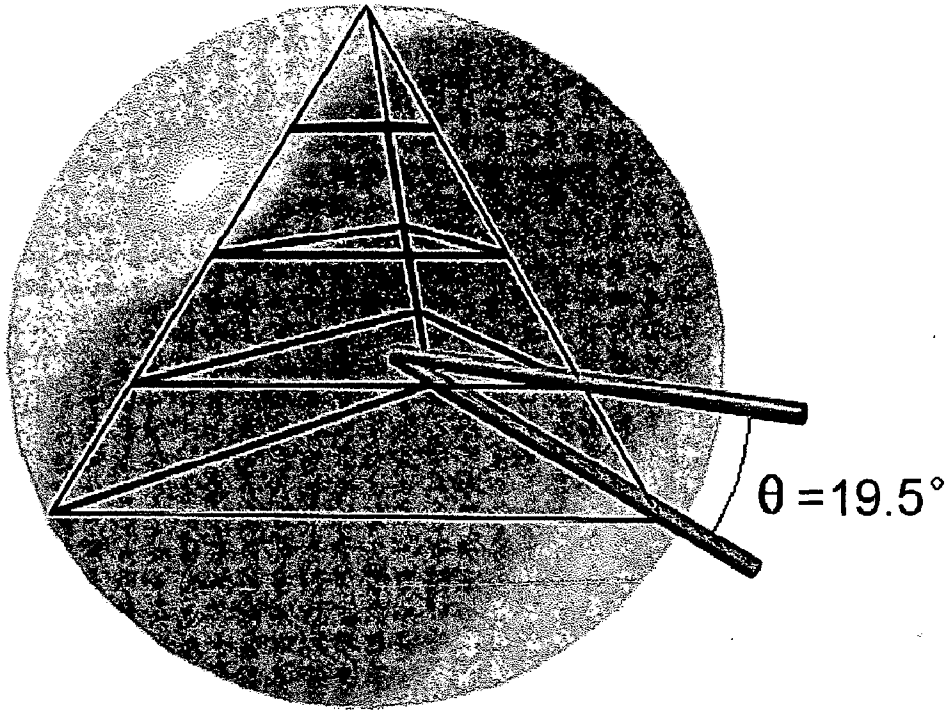

Figure 5

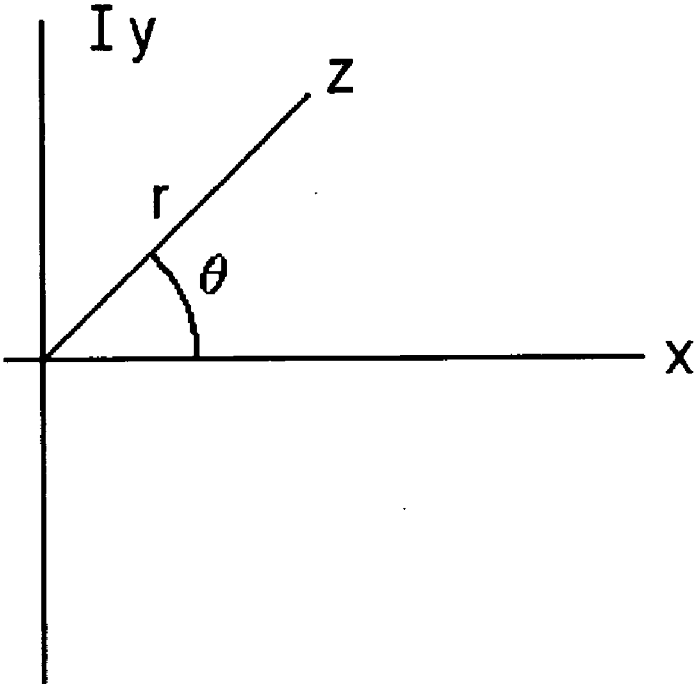
Iy
z
r
θ
x

Figure 6

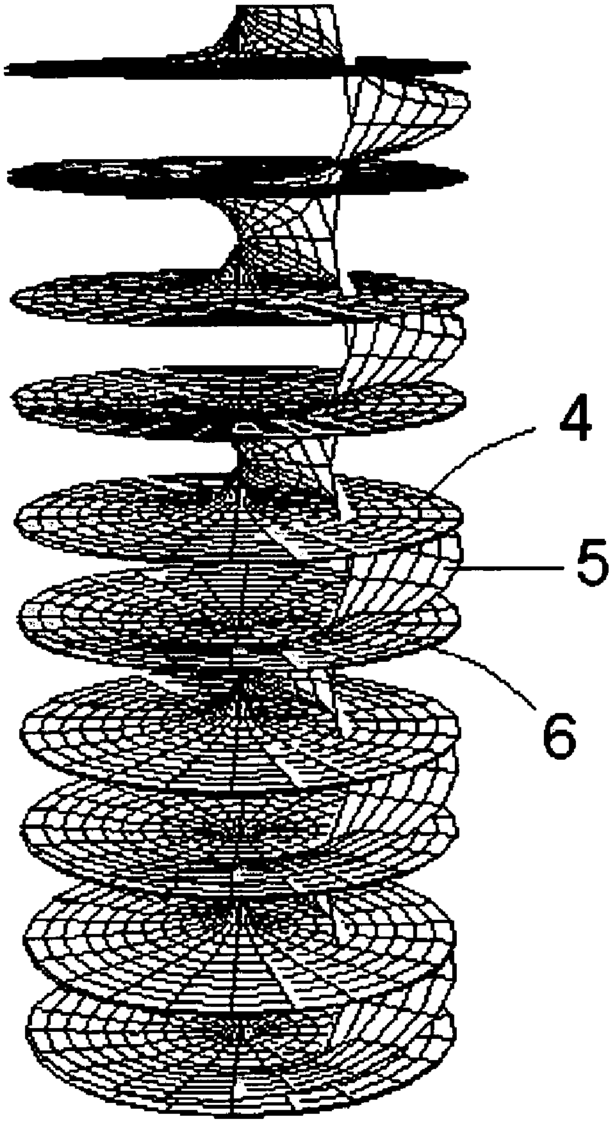

Figure 7

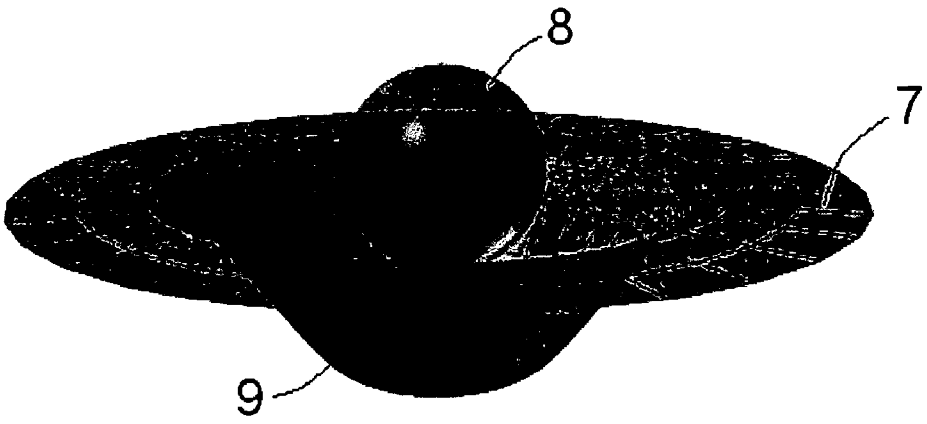

Figure 8

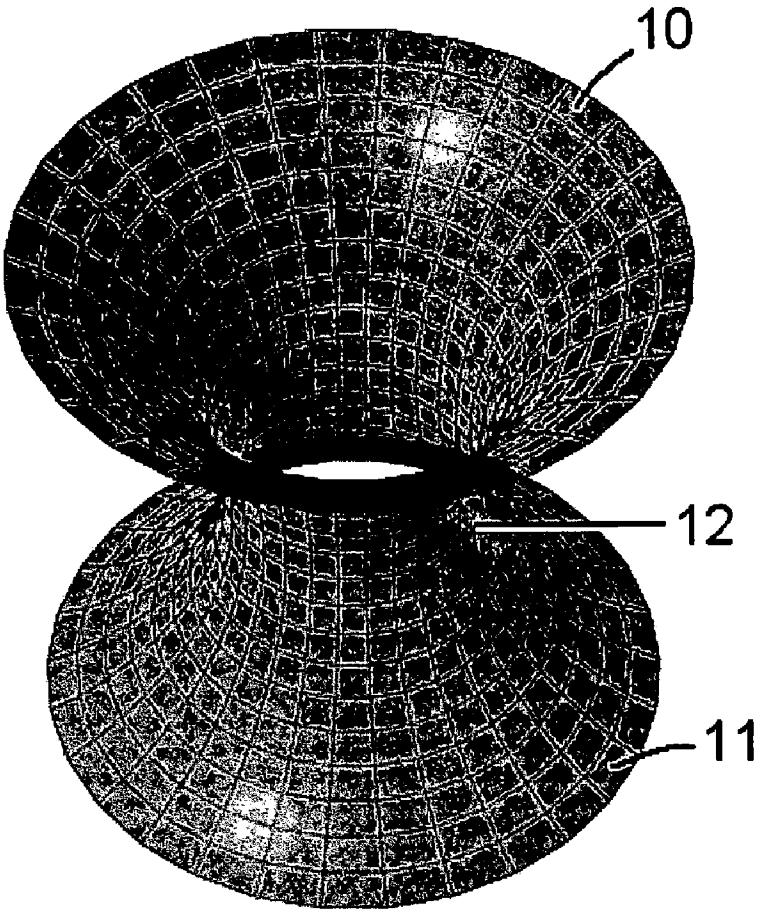

Figure 9

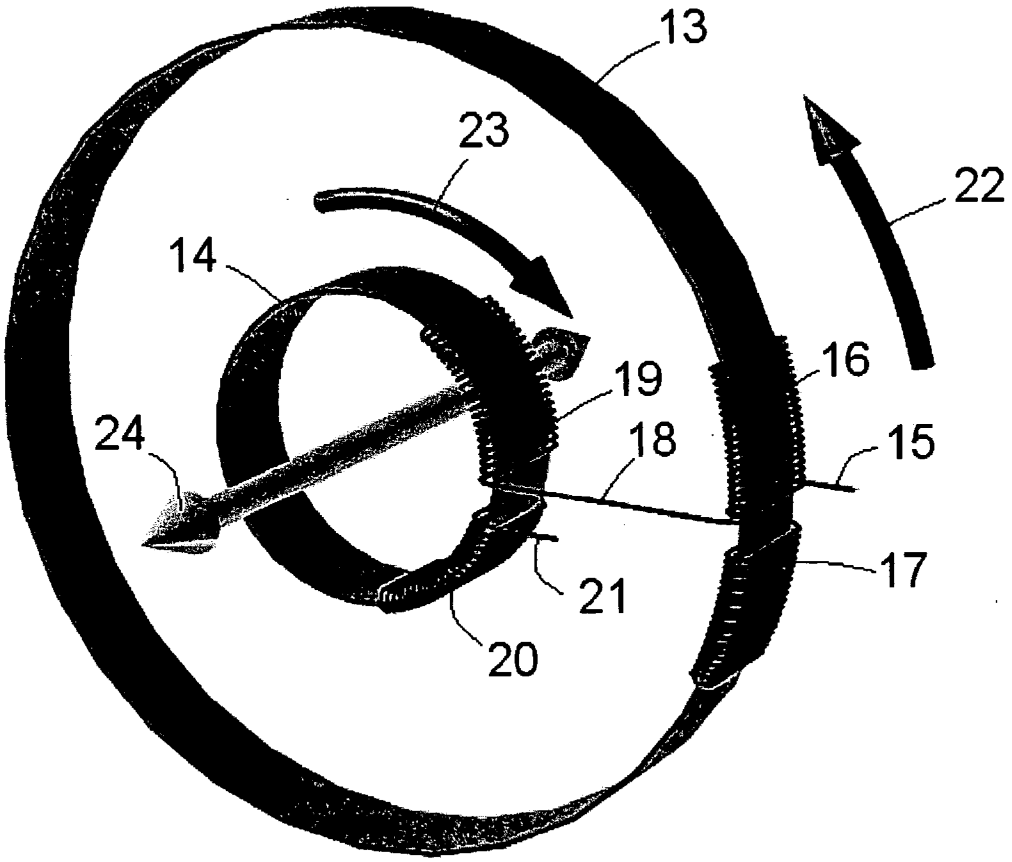
13
23
22
14
16
19
24
18
15
17
21
20

Figure 10

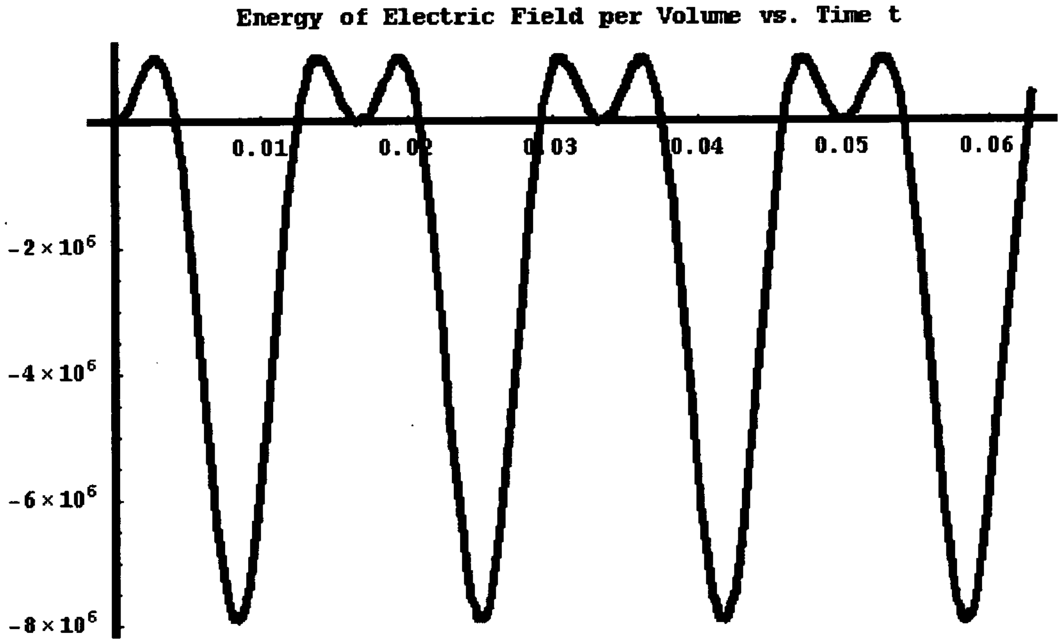
Energy of Electric Field per Volume vs. Time t
0.01
0.02
0.03
0.04
0.05
0.06
-2 × 10⁶
-4 × 10⁶
-6 × 10⁶
-8 × 10⁶

Figure 11

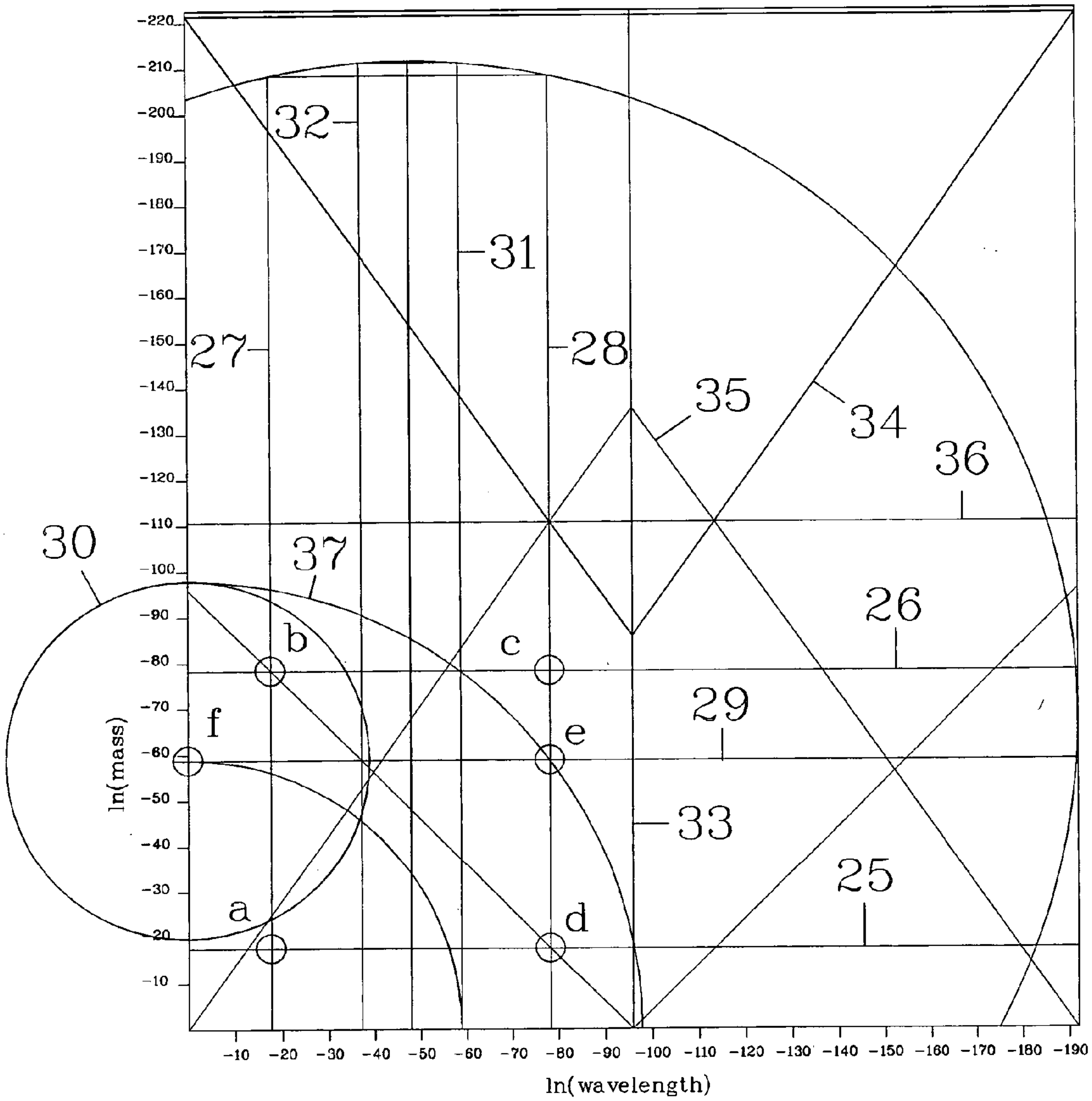
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
-150
-160
-170
-180
-190
ln(wavelength)
32
31
27
28
35
34
36
30
37
26
29
33
25
a
b
c
d
e
f

Figure 12

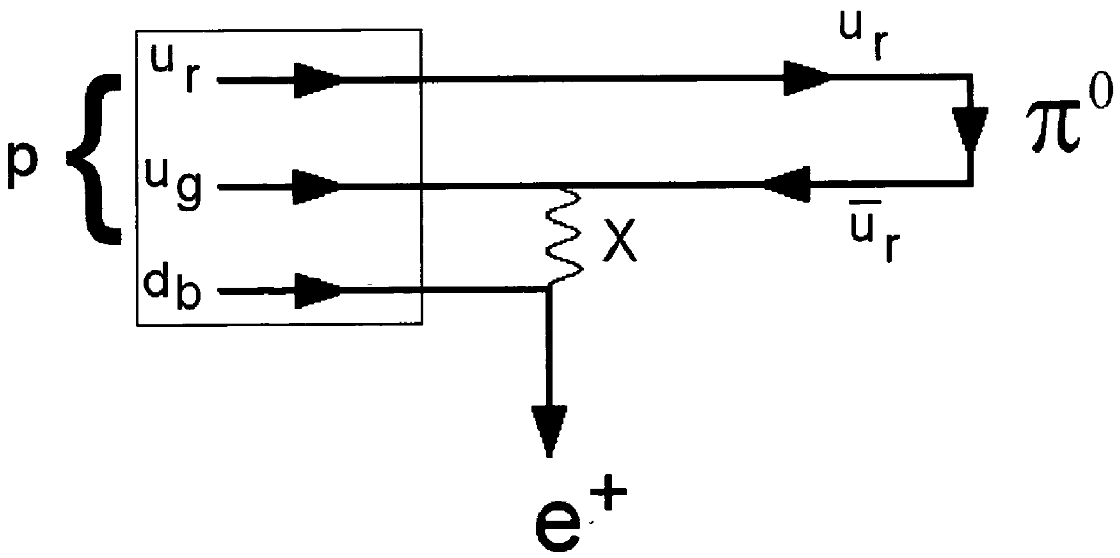
p
u_r
u_g
d_b
u_r
π^0
$\bar{u}_r$
X
e^+

Figure 13

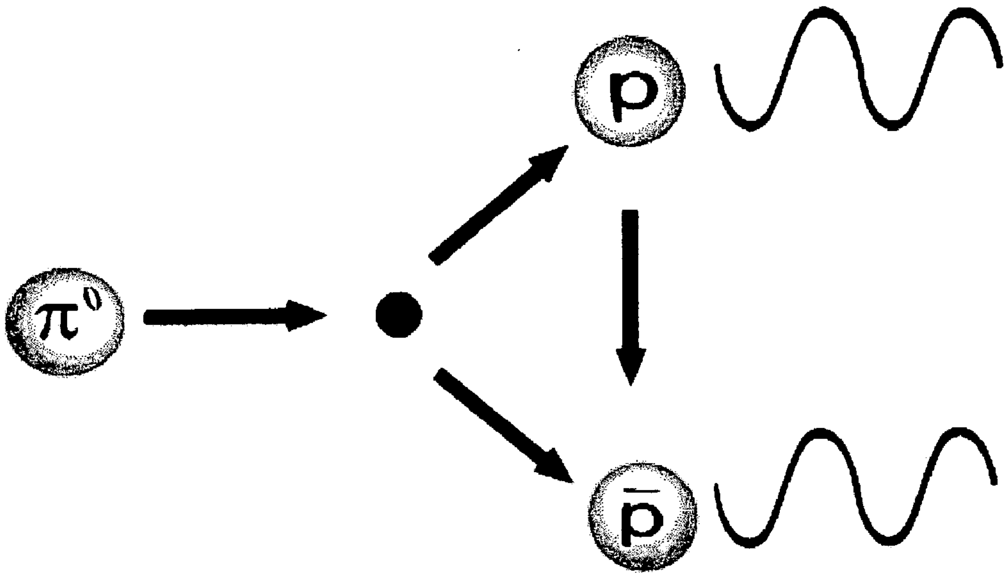

Figure 14

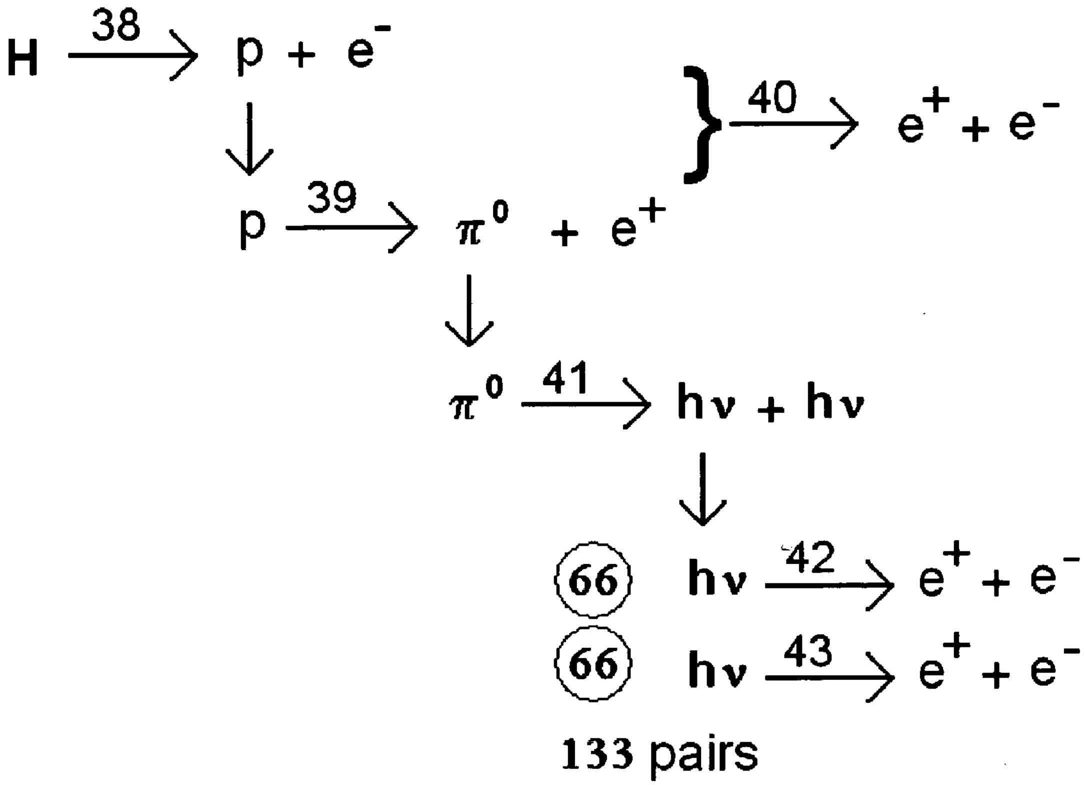
H
38
p + e⁻
40
e⁺ + e⁻
p
39
π⁰ + e⁺
π⁰
41
hν + hν
66
hν
42
e⁺ + e⁻
66
hν
43
e⁺ + e⁻
133 pairs

Figure 15

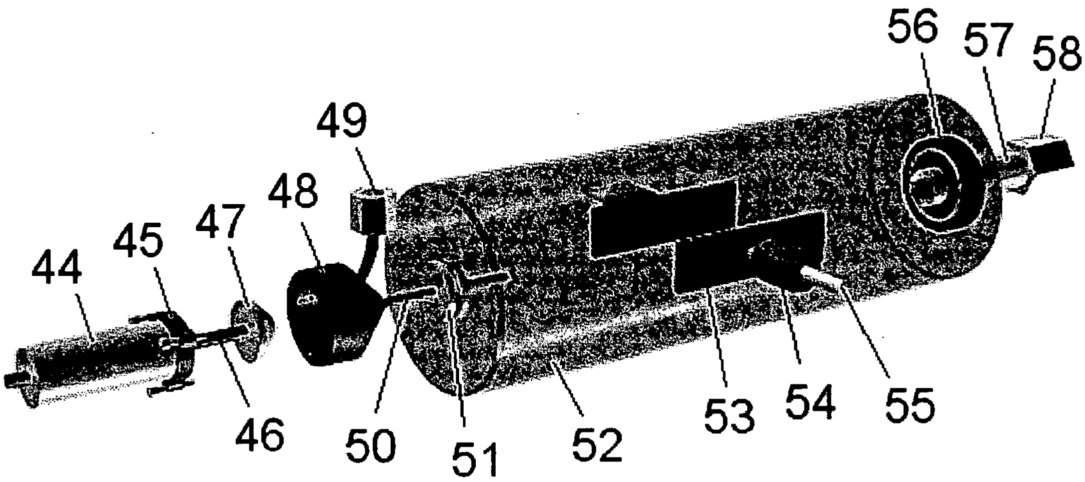
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58

Figure 16

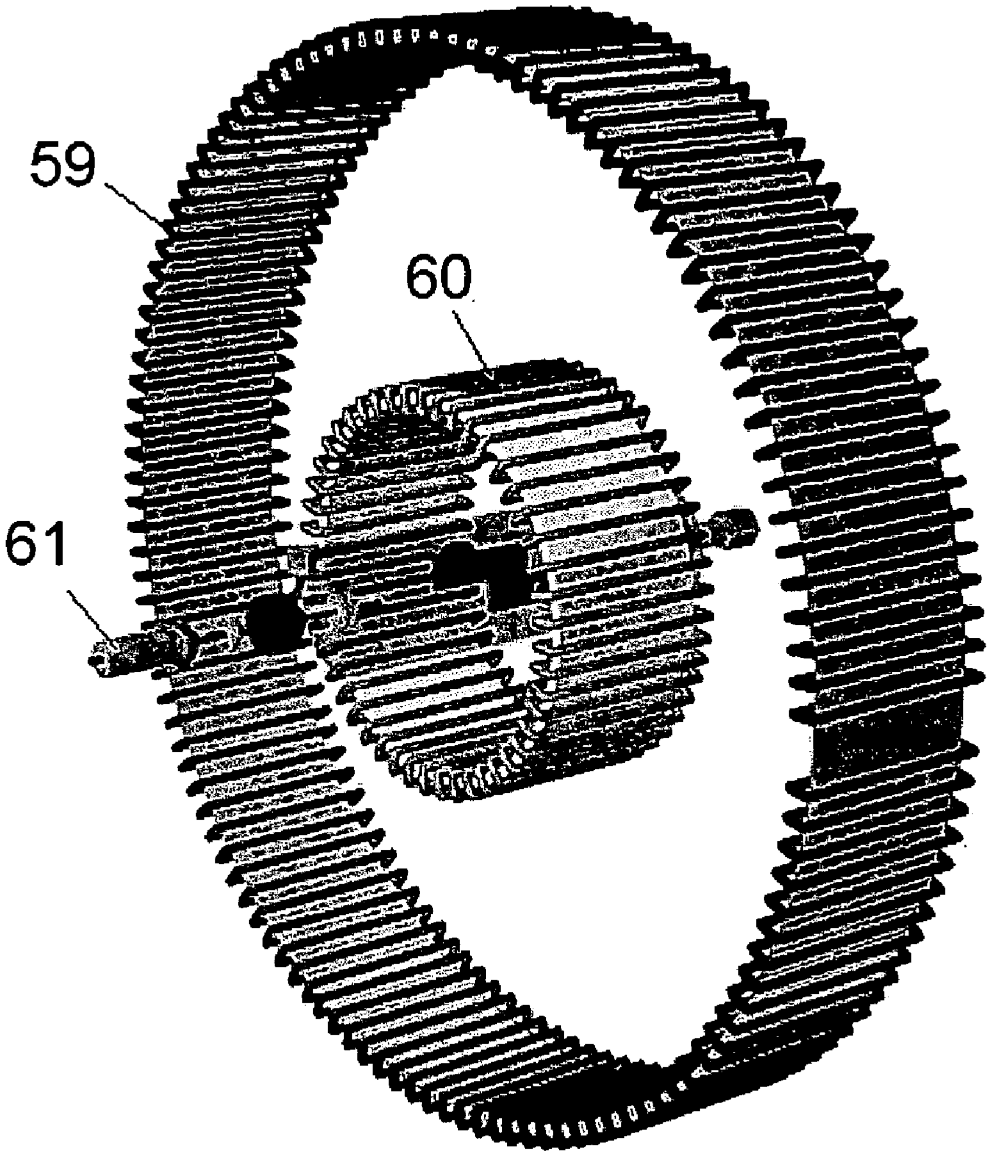

Figure 17

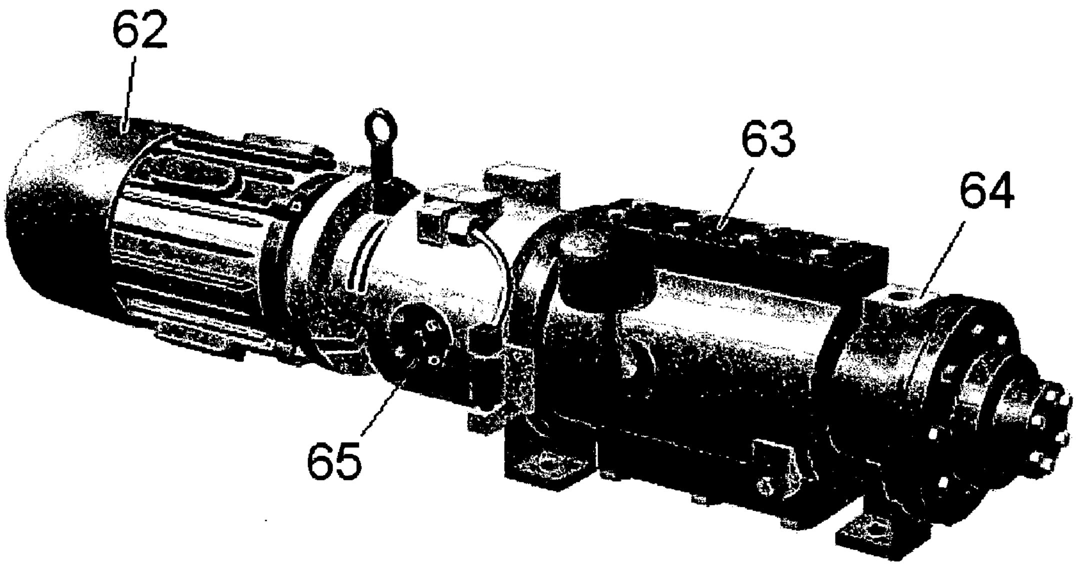

WATER ENERGY GENERATOR

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is an energy generator that uses the transition through wormholes of the hydrogen atoms of water molecules to break the bonds of the atoms and convert the protons into photons and electrons which can be collected for energy.

BACKGROUND OF THE INVENTION

[0002] A molecule of water consists of two hydrogen H atoms and one atom of oxygen O which has the chemical formula

H_2O

The hydrogen atoms can be separated from the oxygen atom by electrolysis. In this process, a direct current of electricity, such as from a battery, is passed through water decomposing it into hydrogen and oxygen. Pure water, however, is a poor conductor of electricity. It is therefore necessary to add some substance to form a solution that will conduct an electric current. Such a solution that will conduct electric current is called an electrolytic solution. A small amount of sulfuric acid or sodium hydroxide is added to the water to form an electrolytic solution. Water electrolyzed yields hydrogen plus oxygen

Water−>hydrogen+oxygen

$2H_2O \rightarrow 2H_2+O_2$

Because this invention is to be used on spacecraft, the oxygen can be used for breathing and the hydrogen can be used to produce energy that will be used to create the electromagnetic fields which provide lift and propulsion.

[0003] Referring to **FIG. 1**, a single hydrogen atom consists of one proton (1) in the nucleus and one electron (2) moving in an orbital (3) around the nucleus. In quantum physics notation, there is only one main K shell and one orbital containing a single electron in the 1 s subshell. An orbital is an energy level containing one or two electrons in a subshell of an atom. Only a total of two electrons may be placed in one orbital with the added constraint that the electrons spin in opposite directions. Looking at the 1 s orbital, it can be seen that only half of it is occupied. What this means is that it is possible to add another hydrogen atom in which its electron occupies the other position in the orbital to from the hydrogen molecule H_2.

[0004] In 1925, a physicist by the name of Edwin Schrodinger developed a wave equation, which bears his name, that models the hydrogen atom. Even though the classical picture of **FIG. 1** showing a distinct electron orbiting the proton is easy to visualize, in reality the electron is a wavefunction Ψ whose square indicates the probability of finding the electron at a particular point. This then was the start of quantum physics where it was found that the electron energy can only take on certain discrete values.

[0005] A traveling wave moving in the positive x-direction can be represented by the function $\Psi_1=\Psi_1(x,t)$ of the form

$$\Psi_1=A_1\cos(2\pi(x/\lambda-v\,t))$$

where x is the distance along the x-axis, A the wave amplitude, λ the monochromatic wavelength, v the velocity of the wave and t time.

[0006] A similar monochromatic wave moving to the left can be represented by

$$\Psi_2=A_2\cos(2\pi(x/\lambda+v\,t))$$

where the sign of the velocity is reversed. The superposition of these traveling waves results in a standing wave, or stationary waves, of the form

$$\Psi=\Psi_1+\Psi_2=A\cos(2\pi x/\lambda)\cos(2\pi v\,t)$$

which is now a product of a spatial-dependent term A cos(2πx/λ), and a time-dependent term cos(2πv t). Taking the partial derivative of Ψ twice with respect to x

$$\frac{\partial^2\Psi}{\partial x^2}=-\left(\frac{2\pi}{\lambda}\right)^2\Psi$$

The momentum p of a particle is equal to Planck's constant h divided by the mass m of the particle

$$p=h/\lambda=\frac{2\pi\hbar}{\lambda}$$

where h bar is the reduced Planck constant. Thus Schrodinger's equation can be written as

$$\frac{\partial^2\Psi}{\partial x^2}=-\frac{p^2}{\hbar^2}\Psi$$

For a particle whose potential energy is V(x), the total energy is the kinetic energy K plus the potential energy

$$E=K+V=(p^2/2m)+V$$

Therefore,

$$p^2=2\,m\,(E-V)$$

and Schrodinger's equation becomes

$$\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2}+(E-V)\Psi=0$$

The potential V is just the Coulomb potential of the product of two charges e divided by the radius r between them

$$V=\frac{e^2}{4\pi\varepsilon r}$$

where ε is the linear capacitance of space. A general wavefunction can be separated into a radial R part and a spherical harmonics part Y

$$\Psi=R_{n,1}Y_{1,m}(\theta,\phi)$$

where the first term is called a radial wavefunction which describes the "in-out" motion of the electron. When Schrodinger's equation is separated, it is found that the radial part of the wavefunction, R, must be a solution of the quantized differential equation

$$\frac{\hbar^2}{2mr}\frac{d^2(rR)}{dr^2}+\left(E-\frac{l(l+1)\hbar^2}{2mr^2}+\frac{e^2}{4\pi\varepsilon r}\right)R=0$$

This can be expressed in a simpler form by defining the function

$$f=rR$$

which is then found to satisfy the equation

$$\frac{\hbar^2}{2m}\frac{d^2 f}{dr^2}+(E-V_{eff})f=0$$

where the effective potential is given by

$$V_{eff}=\frac{l(l+1)\hbar^2}{2mr^2}-\frac{e^2}{4\pi\varepsilon r}$$

where the 1's are the quantum orbital numbers. For s orbitals equal to 1=0, the effective potential is just the electrostatic potential of the nucleus. For 1 greater than zero, the first term is equal to the kinetic energy owing to the angular motion of the electron at a distance r and with angular momentum √1(1+1)h.

[0007] Referring to **FIG. 2**, the effective potential is graphed as a function of radius r and the orbital numbers. For orbital number l=0, which is the lower of the three curves, the potential does not provide a stable position for the electron (black disk) and the negatively-charged electron just crashes into the positively-charged proton nucleus as shown by the arrow. For orbital l=1, the first term is called the centrifugal repulsion which together with the electrostatic potential provides for a stable position for the electron as seen in the middle curve. The potential energy is negative which creates a slight valley in which the electron obtains a stable orbit. Higher orbital numbers l=2 produce a similar valley potential further out on the radius. The l=1 orbital does not allow the electron to come near the proton which provides for a stable hydrogen atom. So the key to this invention is how to destabilize this hydrogen atom energy system and produce usable energy which can propel the electromagnetic spacecraft and run other hyperspace inventions. Refer to my patent applications such as Dipole Moment Spacecraft, Dual Potential Hull Spacecraft, Photon Spacecraft, Electromagnetic Field Propulsion System, Full Body Teleportation, Magnetic Vortex Wormhole Generator, Electric Vortex Wormhole Generator, Sulfur S8 Wormhole Generator, Cavitating Oil Hyperspace Energy Generator, Rotor Inductance Propulsion System and Triangular Spacecraft.

[0008] Many of these patent applications involve wormholes and hyperspace which are not well-known concepts in the scientific community. Hyperspace consists of those dimensions which are co-dimensional with our spacetime. The reason I know about hyperspace is because (1) I have been in hyperspace on a number of occasions and have experienced Einstein's time dilation according to his General Theory of Relativity, (2) I have experienced more than one full-body hyperspace teleportation over a distance of 100 meters, (3) we have been able to create a wormhole between space and hyperspace with the magnetic vortex wormhole generator in which smoke was blown through one side of the coil into hyperspace, a first contact verified by the Grey Aliens, (4) I have seen the green mist associated with moving out of dimension and crossing over into hyperspace, (5) I have looked into another dimension and have seen another building, a car and a man who waving at me in the presence of an artificially-intelligent Cyborg with the "high-tech look" from the Pleiadian Defense Department, (6) I can remote view through hyperspace subgeometry to distances of 100,000 light years to the edge of the galaxy and have made first contact with around 500 extraterrestrial civilizations involving the use of patent applications such as Remote Viewing Amplifier, Quantum Dot Energy Cylinder and Walking Through Walls Training System, (7) I am the only person on Earth who has communicated with the designers of the crop circles found in England and explained their design to them in terms of subspace geometry, (8) I can walk through walls as a hyperspace energy being, and (9) I have discovered how anti-gravity is possible using low density hyperspace energy, not to mention all the other research work on these electromagnetic field propulsion vehicles. For my work in developing the geometry of the subspace manifold known as the tetrahedron diagram, I was given the Aphysics award by the Admiral, who is third in line to the Admiralty of the Pleiadian Defense Department. For my work in Revelations, she awarded me four beautiful galloping white riderless horses of the Apocalypse. The reason that the Pleiadian Defense Department was involved was that the Admiral had the task of creating the energy being that would protect the subspace manifold during the battle of Revelations which took place in the year 2001. So these are some of my personal experiences in the field of hyperspace physics.

[0009] As to the scientific aspect of hyperspace physics, observations of the solar system have noted that large vortices occur on the planets at a latitude of ±19.5°. On the planet Jupiter, for example, the Giant Red Spot vortex, which is the size of two planets like Earth, is located at −19.5° latitude.

[0010] Referring to **FIG. 3**, the Olympus Mons volcano, which is the size of France, occurs in the northern hemisphere at 19.5°. The plume of volcanic ash can be seen being blown to the upper right.

[0011] Here on Earth, just north of me at +19.5° in the Caribbean, there is a slow moving rock mantle vortex that curves the islands down toward Venezuela. Since the tetrahedron has three corners, another corner occurs at 120° to the west where the Marshall Islands are located. The reason that the volcanic islands form is that low density hyperspace energy softens the atoms and molecules to such an extent that the atomic bonds are broken. The hot magma from the core has an easier path to the surface through the soft rock compared to the hard rock of surrounding areas. As the vortex rotates, the rising hot magma creates volcanoes which form the chain of islands.

[0012] As another example of this, the Silver bridge, which is located at a double harmonic of 39° between Ohio and West Virginia, collapsed because a large wormhole opened up and flooded the bridge with low density hyperspace energy. The metal rivets softened and pulled out of the plates holding down the suspension cables. Thus the roadway tilted to one side and all the cars fell into the river.

[0013] If you look at the quarried granite polygonal megalithic stone blocks that were used to build Machu Picchu, the blocks are put together as though they were soft putty. They fit together perfectly. Machu Picchu, or Great Picchu, is the Quichua name for a sharp peak which rises ten thousand feet above the sea. How this was achieved is described in my patent application *Rotating Magnetic Vortex Generator* which shows that rotating permanent magnets can create a wormhole between space and hyperspace. The low density hyperspace energy then floods the block as to make it weightless. The block is then floated up from the quarry which is 2500 feet below Machu Picchu. The huge 1000-ton blocks found in the platform at Baalbek in the Cedar Mountains of Lebanon were also lifted into position in this manner.

[0014] Using a technique called Chi Kung breathing, not necessarily known to the Chinese, one of our researchers has been able to levitate himself to a height of six feet. He might have gone higher but he lost his balance and his arm touched a nearby metal pipe. He then floated back to the ground. Contracting the abdominal muscles causes hot air to rise in the lungs while at the same time cold air is inhaled through the nose. This combination of hot and cold air mixing together creates environmental oscillators having a temperature fluctuation. This changes Planck's constant such that the body goes out of dimension and low density hyperspace energy fills the body.

[0015] The final example comes from Biblical times. Jesus and his father Joseph were both carpenters. A chair that Joseph had made for a client was found to be too small. So Jesus placed his hands on either side of the chair and stretched the chair to fit. The phrase in Spanish is "mas silla" which means that Jesus made more chair. In the dictionary, the Spanish word for putty is "masilla." The reason that Jesus could do this is that he passed hyperspace energy through his hand vortices which softened the wood. See my patent application called *Hyperspace Torque Generator* which shows this in more detail. The hand vortex is actually a planar co-gravitational K field that can rotate a pendulum around in circles. This solenoidal field corresponds to the magnetic H field. The electric field corresponds to the linear gravitational g field. The equations are identical for both sets of fields. I learned that Jesus actually reincarnated here on Earth in order to identify the planet on which the battle of Revelations would take place.

[0016] In summary, these are some examples as to how low density hyperspace energy can soften and break the bonds of atoms and molecules.

[0017] As mentioned previously, the astronomical observations suggest that the structure of the universe is related to the tetrahedron. Referring to **FIG. 4**, the corners of a tetrahedron circumscribed by a sphere, touch the sphere at an angle of 19.47122063°. The ratio of the area-to-volume ratio of the sphere to the area-to-volume ratio of the tetrahedron is ⅓ which is the magic ratio in physics. The arcsin(⅓) is 19.47°.

[0018] In order for all the physics constants to be projected into our universe and the co-dimensional hyperspace dimensions, there has to be a tetrahedral subspace manifold. The only mathematical function which allows multiple rotations of the complex plane is the logarithmic function Log[z].

[0019] Referring to **FIG. 5**, the complex number z in the complex plane is equal to x+I y where x and y are real constants and $I^2=-1$. The complex number z can also be written in terms of a radius and an angle

$$z=r\ \mathrm{Exp}[I\ \theta]$$

Because the angle can rotate more than once in 2π m multiples, the complex number z is actually

$$z=r\ \mathrm{Exp}[I(\theta+2\pi\ m\ )]$$

Taking the logarithm of z

$$\mathrm{Log}[z]=\mathrm{Log}[r]+I(\theta+2\pi m)$$

[0020] Referring to **FIG. 6**, it can be seen that it is not possible to go around multiple times on the same surface due to a branch cut along the origin where Log[0] is undefined. What happens is that as you rotate around on surface (4) and get to the origin, you go down along the branch (5) to the next hyperspace plane (6). Thus the universe is composed of many hyperspace co-dimensions. From personal experience, I estimate that the distance separating the two dimensions is about 3 meters when I was looking into another dimension at the man waving at me. The physics is even more complicated because there is a Lorentz dimensional transformation in which higher dimensions appear smaller and lower dimensions appear larger. In one case I was looking at a huge mothership at a much higher dimension and it looked like a tiny toy model spacecraft. They fired a laser cannon at me, and I then curved space which made the beam change course. You can see why I was awarded the four horses.

[0021] The antilog of Log[z] is

$$e^{\mathrm{Log}[z]}=z=re^{I\ \omega t}$$

where the angle is the angular frequency ω times the time t. The subspace geometry remains stationary or fixed, but the projection of the antilog into our dimension generates a system which is frequency dependent. This is why we get oscillating fields.

[0022] From Einstein's General Theory of Relativity, it is known that various kinds of energy can curve spacetime such as mass, electromagnetic fields, angular momentum and electrical charge. The elemental spacetime length ds in cylindrical coordinates {t,r,θ,z}, known as the Schwarzschild metric, shows that spacetime can be curved using mass M and charge Q

$$(ds)^2=-(dt)^2\left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right)+\frac{dr^2}{\left(1-\frac{2M}{r}+\frac{Q^2}{r^2}\right)}+(rd\theta)^2+(dz)^2$$

where you can see that the elemental time dt is dilated by the mass M and the elementary radius dr is reduced. I have actually experienced this time dilation in which, as I was jumping into hyperspace, I was hit by a car which broke my shoulder. When I came back into dimension which appeared to be a few seconds later, I found that I had an 8-inch blood ring down the right side of my chest. My shoulder blade, which sticks up about two inches, is still broken to this day.

[0023] Referring to **FIG. 7**, flat space (**7**) can be curved (**9**) by a massive body (**8**) such as a planet. For example, the mass of the Earth gives space a negative curvature such that objects tend to fall toward the center of the mass (bowl). On the other hand, using electromagnetic fields, it is possible to produce a positive curvature such that a spacecraft rises by falling upward.

[0024] Referring to **FIG. 8**, if the fields are intense enough, then a wormhole (**12**) forms between space (**10**) and hyperspace (**11**). This depiction is called an embedding diagram because there is no open space going through the wormhole. An object traversing the wormhole moves along the surface from one dimension to another. In order to open the throat (**12**) of the wormhole, negative energy is required. Referring to my patent application Dual Potential Hull Spacecraft, negative energy is produced by the interaction of a microwave beam with an oscillating magnetic H field. In terms of differential forms mathematics, this is given as the Hodge Star * of the differential d of the wedge ^ product of the magnetic H field with the electromagnetic wave {B,E}

$$*d(H\hat{}(B+E\hat{}dt)) = \frac{\partial(-\rho)}{\partial t}$$

which says that there is an increasing rate of change of negative energy (−ρ). Due the presence of negative energy together with the spacetime curvature pressure produced by the electromagnetic fields, wormholes open up between space and hyperspace. There is a positive gravitational potential between hyperspace and space because the low density hyperspace energy is more dense than the negative energy in this dimension. Thus the hyperspace energy flows into our dimension which reduces the mass of the spacecraft. The upward spacetime pressure stress over the hull due to the electromagnetic fields creates a lift force on the nearly massless vehicle. Because the hyperspace energy has a speed of light equal to one meter per second, the electromagnetic fields become relativistically strong since they obey the Lorentz transformation. Due to the low mass, high spacetime lift pressure and strong EM fields, the vehicle can attain very high rates of acceleration.

[0025] Another method to produce a wormhole is to use bucking magnetic fields which is described in my patent application *Magnetic Vortex Wormhole Generator*. In gravitation physics, the Faraday F tensor, which is a 4×4 spacetime metric {t,x,y,z}, contains all the components of the electromagnetic fields in the various spatial directions {x,y,z}

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & E_x & E_y & E_z \\ x & E_x & 0 & B_z & -B_y \\ y & E_y & -B_z & 0 & B_x \\ z & E_z & B_y & -B_x & 0 \end{array}$$

where contravariant index a refers to the rows, and the covariant index β refers to the columns. For example, the component $F^t_x=E_x$ is the electric field in the x-direction. If there were two magnetic bucking fields in the x-direction the Faraday tensor would be

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & B_x - B_x \\ z & 0 & 0 & -(B_x - B_x) & 0 \end{array} = \begin{vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{vmatrix}$$

which says that the Faraday tensor is zero. Thus no spacetime curvature is generated from two bucking magnetic fields at the same position. On the other hand, if the two bucking magnetic fields are concentric at different radii, then the Faraday tensor becomes

$$F^{\alpha}_{\beta} = \begin{array}{c|cccc|} t & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ y & 0 & 0 & 0 & B_x\delta(x_1) - B_x\delta(x_2) \\ z & 0 & 0 & -(B_x\delta(x_1) - B_x\delta(x_2)) & 0 \end{array}$$

which is not zero due to the presence of the Kronecker δ delta function which locates the fields at different positions. The spacetime stress-energy-momentum tensor T is then computed from the following equation

$$4\pi T^{\mu\nu} = F^{\mu\alpha}F^{\nu}_{\alpha} - \frac{1}{4}g^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}$$

where g is the metric tensor containing the coefficients of the elemental spacetime length ds. With the mass M and charge Q term equal to zero, there being just electromagnetic fields involved, the g metric tensor in cylindrical coordinates becomes

$$g_{\alpha\beta} = \begin{array}{c|cccc|} t & -\left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right) & 0 & 0 & 0 \\ \begin{matrix} r \\ \theta \end{matrix} & 0 & \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)^{-1} & 0 & 0 \\ z & 0 & 0 & r^2 & 0 \\ & 0 & 0 & 0 & 1 \end{array}$$

$$= \begin{vmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & r^2 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

[0026] Referring to **FIG. 9**, the magnetic vortex generator has two concentric, thin, flat cylindrical silicon-iron cores (**13**,**14**), each consisting of a stack of three 0.020 inch thick transformer laminations wrapped with insulating tape. Using insulated magnetic wire, a flat helical coil (**16**) is wrapped counter-clockwise around the outer laminations. The coil receives power through connection (**15**). The winding then continues around the core until coil (**17**) where the wire (**18**) is extended to the inner core and the wrapping of coil (**19**) proceeds in the clockwise direction. The inner core wrapping terminates on coil (**20**) with the extension of the second power connection (**21**). Because the coils are wound in opposite directions, the generator produces two bucking magnetic fields at different radii (**22**,**23**).

[0027] According to Maxwell's electromagnetic equations, the curl of the magnetic flux density B field times the square of the speed of light squared is equal to the rate of change of the electric E field

$$c^2\nabla xB = \frac{\partial E}{\partial t}$$

Multiplying both sides by the elemental area of the core and integrating over the area

$$c^2\int_0^{r_0}\nabla xBdA = \frac{\partial}{\partial t}\int_0^{r_0}E\cdot ndA = \frac{\partial}{\partial t}\int_0^{r_0}E\cdot n2\pi rdr$$

The curl of the field can be converted into a line integral around the core contour using Stokes' Theorem

$$c^2\int_0^{r_0}\nabla xBdA = c^2\oint_{r_0}B\cdot ds = c^2\int_0^{2\pi}Br_0d\theta$$

[0028] The magnetic flux density B field oscillates with angular frequency ω

$B=B_0e^{I\omega t}$

Substituting this into the equation and integrating over time for the inner core field

$$-c^2\int_0^t\int_0^{2\pi}B_0e^{I\omega t}r_0d\theta = \int_0^{r_0}E2\pi rdr = E\pi r_0^2$$

This can be solved for the electric E field that is produced by the oscillating magnetic flux density B field

$$E_0 = \frac{2IB_0c^2(-1+e^{I\omega t})}{r_0\omega}$$

The electric field for the outer core is the same equation except that the opposite sign of the magnetic flux density B field and the radius **r1**

$$E_1 = \frac{-2IB_0c^2(-1+e^{I\omega t})}{r_1\omega}$$

Adding these two fields together is the total electric field E

$$E = E_0 + E_1 = \frac{-2B_0c^2(-1+\mathrm{Exp}(I\omega t))(r_0-r_1)}{r_0r_1\omega}$$

The electrostatic energy of the field is equal to half the linear capacitance of space times the summation over the volume of the dot product of the electric field with itself

$$U = \frac{\varepsilon}{2}\int E\cdot EdV$$

Because the electric field points in the z-direction out of the coil, the dot product is actually the square of the electric field.

[0029] Referring to **FIG. 10**, the energy U per volume is plotted as a function of time with a radius ratio r_1/r_0 of 3/1. As the graph shows, the magnetic vortex wormhole generator produces mostly negative energy which is required in order to create the wormholes. Because the stress-energy-momentum T tensor is also the square of the electric field, this graph gives the spacetime curvature pressure. Thus the electric field produces both the pressure and negative energy required to open up wormholes between space and hyperspace.

[0030] Going back to the effective potential equation,

$$V_{eff} = \frac{l(l+1)\hbar^2}{2mr^2} - \frac{e^2}{4\pi\varepsilon r}$$

it can be seen that the first term is divided by the mass of the particle. In current gravitation physics, the mass of the particle is invariant with velocity. It does not obey the Lorentz transformation. The mass is related to the energy E of the particle and its momentum p by

$m^2=E^2-p^2$

In different inertial frames moving with a relative velocity v, the energy and the momentum obey the Lorentz transformation, but no matter what the relative motion, the mass of the particle is constant. In the first term of the effective potential, the mass is constant which leaves just Planck's constant.

[0031] Having worked for over ten years on the subspace manifold, known as the tetrahedron diagram, I found a most incredible intersection on the diagram when working with the water molecule. The water molecule has two hydrogen atoms and one oxygen atom as mentioned previously. The atomic weight of one atom of hydrogen is 1.008 atomic weight units (awu). The atomic weight of oxygen is 16.000 awu. Therefore the molecular weight of water is

$$\text{Weight of two atoms of hydrogen } 2\times1.008\ awu = 2.016\ awu$$

$$\text{Weight of one atom of oxygen } 1\times16.000\ awu = \frac{16.000\ awu}{18.016\ awu}$$

The gram molecular weight is the atomic weight expressed in grams, so there are 18.016 grams in Avogadro's number of molecules. So the mass per molecule in logs is

$$\text{Log}\left[\frac{18.016\ \text{gram}-\text{mol}\Big/\left(1000\ \frac{\text{gm}}{\text{kg}}\right)}{6.02\times10^{23}\ \text{mol}}\right]=-58.77103943$$

[0032] Referring to **FIG. 11**, the tetrahedron diagram plots the natural logarithm of mass versus the natural logarithm of wavelength. The reason for this is that mass times wavelength is equal to Planck's constant divided by the speed of light c, known on the diagram as the base constant. At the present time there are over 4000 diagrams which are copyrighted in the Library of Congress. In logarithms, the product of two numbers is the sum of the two numbers. This means that the sum of the mass and the wavelength are equal to the base constant which has a value of

$$\ln[m]+\ln[\lambda]=\ln[h/c]=-95.91546344=\text{base constant}$$

Our dimension has a lower limit on mass and length known respectively as the Planck mass and the Planck wavelength. The Planck mass is the linear mass Ω of the universe times the Planck scale Λ. The Planck wavelength is circumference of a circle of radius Planck scale. In terms of logs, the Planck mass and Planck wavelength are

$$\text{Planck mass}=\ln(\Omega\Lambda)=-17.64290101$$

$$\text{Planck wavelength}=\ln(2\pi\Lambda)=-78.27256243$$

[0033] When these values are plotted on the tetrahedron diagram shown in **FIG. 11**, the Planck box (abcd) is formed which are the boundaries of our dimension in subspace. The line numbering is as follows

Planck mass	25, 27
Planck wavelength	26, 28
mass of water molecule	29, 31
speed of light squared circle	30
base constant	33
inverted tetrahedrons	34, 35
centerline	36

The energy of the water molecule, circle (**37**), is equal to the sum of the water molecule mass (**29**) plus the speed of light squared circle (**30**). The energy circle (**37**) intersects the mass of the water molecule (**29**) at the Planck wavelength (**28**, point e), which is the boundary between space and hyperspace. What this means is that the mass is equal to the energy at the Planck box boundary. The only way that this is possible is if the speed of light c is equal to one meter per second

$$E=m\ c^2=m\ c=1\ \text{meter/second}$$

A water molecule traversing a wormhole into hyperspace undergoes a change in the speed of light from 299792458 m/s to 1 m/s.

[0034] Planck's constant $\hbar$ is equal to the Planck mass ΩΛ times the Planck scale Λ times the speed of light c.

$$\hbar=\Omega\Lambda\Lambda c$$

By having the speed of light go to 1 m/s, the orbital term in the effective potential V_{eff} is reduced by a factor of the speed of light squared equal to 9×10^{16}. This unbalances the equation to such an extent that only the Coulomb potential term remains.

$$\frac{l(l+1)\hbar^2}{2mr^2}\ll\text{coulomb term}$$

The electron is attracted to the proton nucleus because the centrifugal term no longer provides a stable orbit for the electron. Thus the atomic binding is destroyed and the water molecule becomes soft as putty.

[0035] Referring to **FIG. 12**, the collision of the electron with the proton together with the enormous change in the proton's energy causes the proton p to become unstable and decay. According to the Standard Model of particle physics, the elementary particles are composed of smaller particles known as quarks. The six quarks have been named up u, down d, strange s, charm c, top t, and bottom b. The subscript on the quark indicates one of three colors {red r, blue b, green g). As shown in the diagram, the proton is composed of three quarks $\{u_r, u_g, d_b\}$, two of which are up quarks of which one is red and the other green, and a third blue down quark. The proton p decays into a positron e^+ which is an electron with a positive charge, and a neutrally-charged pion π^0 particle through the exchange of an X boson particle. The pion has a mass between the electron and the proton.

$$p\rightarrow\pi^0+e^+$$

[0036] Referring to **FIG. 13**, the pion π^0 then decays into a proton p and antiproton $\bar{p}$ which annihilate each other to produce two photons shown on the right by the traveling waves. So the overall energy exchange is

$$p\rightarrow\frac{2hv}{c^2}+e^+$$

where hv is the energy per gamma photon with frequency v. The electron of the hydrogen atom would then annihilate the positron for additional photon energy.

[0037] Referring to **FIG. 14**, the hydrogen H atom is composed (**38**) of the proton and electron as seen in the upper left corner. The proton decays (**39**) into the neutral pion and a positron. The electron from the hydrogen atom and this positron form one electron pair (**40**). The pion then decays (**41**) into two gamma photons which produce an electron pair production energy cascade into 132 pairs (**42**,**43**) for a total of 133 electron pairs. These electrons can then be captured electrostatically and used for the production of electricity.

[0038] Referring to **FIG. 15**, the electrons are captured with the water droplet injector . The plunger (**46**) of a spring-loaded cylindrical solenoid (**44**) is attached to a tapered piston (**47**). By means of ring collar and bolts (**45**), the solenoid is bolted to the injector (**48**). A supply of purified water is attached to the water inlet connection (**49**). When the solenoid is activated, it pulls back slightly so that water can enter the valve. When the solenoid is deactivated, the piston forces the water droplet out through the nozzle (**50**) into a cylindrical glass vacuum chamber (**52**). Two cylindrical glass disks (**51**) hold the nozzle in place. On the other end of the vacuum chamber is the sealed-tube connection (**56**,**57**,**58**) to the vacuum pump. In the middle of the vacuum chamber, two metal plates (**53**) are attached through sealed glass collars (**54**) to electrical pins (**55**). The plates are electrostatically-charged with opposite charges so as to form a capacitor. This creates an electrical field between the plates which attracts the electrons to the positively-charged plate.

[0039] Referring to **FIG. 16**, the vacuum tube and water injector (**61**) are mounted along the centerline of the inner (**60**) and outer (**59**) magnetic vortex wormhole generator coils. The low density hyperspace energy traversing the wormhole along the centerline of the coils causes the injected water molecules to soften and decay into a cascade of electrons. The oscillating electric field along the centerline causes the electrons to vibrate back and forth. The crossed electric field between the charged capacitor plates causes the electrons to flow toward the positively-charge plate in order to produce electricity.

[0040] Referring to **FIG. 17**, the vacuum tube is connected to the vacuum pump through a hose connection to the pump air inlet (**64**). A 5 Hp electric motor (**62**) drives dual rotating flights of screws which trap the air and move it toward the exhaust outlet (**65**) shown with no muffler. On a spacecraft operating in the vacuum of outer space, this component would not be needed.

SUMMARY OF THE INVENTION

[0041] It is the object of this invention to generate electricity by using low density hyperspace energy to soften water molecules such that the atomic binding is broken which causes the hydrogen nucleus to decay into a cascade of electron pairs. These electrons are then collected on a positively-charged plate in order to produce electricity. The water molecules are softened by flooding them with low density hyperspace energy that is produced by a magnetic vortex wormhole generator. The generator creates negative energy and a spacetime curvature along the centerline of two concentric coils. This combination opens up wormholes along the centerline. Because the gravitational potential of low density hyperspace energy is greater than the negative energy, the hyperspace energy flows through the wormhole from hyperspace into our dimension. The hyperspace energy has a speed of light equal to one meter/second. This causes a change in Planck's constant h such that the proton orbitals of the hydrogen atom are unable to produce a centrifugal repulsion which keeps the electron in orbit. The Coulomb potential term dominates and the electron is attracted to the proton. Due to the vast change in the speed of light, and the collision of the electron with the proton, the proton becomes unstable and decays into a neutral pion and a positron. The pion then decays into two gamma photons which produce a large cascade of electron pairs.

[0042] A water injector, consisting of a solenoid-activated valve and nozzle, injects water droplets into a vacuum chamber which is positioned along the centerline of the two concentric coils where the wormholes form. Due to the low density hyperspace energy passing through the wormholes into our dimension, the water molecules soften and decay into electrons which are collected on an electrostatically-charged capacitor plate having a positive charge located in the glass vacuum chamber.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0043] **FIG. 1**. Perspective view of hydrogen atom K shell.

[0044] **FIG. 2**. Graph showing potential binding energy of hydrogen atom.

[0045] **FIG. 3**. Perceptive view of Olympus Mons volcano at +19.5° Mars latitude.

[0046] **FIG. 4**. Perspective view of tetrahedron inscribed in sphere.

[0047] **FIG. 5**. Graph showing complex plane.

[0048] **FIG. 6**. Perspective view of hyperspace co-dimensions of logarithmic manifold.

[0049] **FIG. 7**. Perspective view of embedding diagram showing curvature of space caused by a mass.

[0050] **FIG. 8**. Perspective view of wormhole embedding diagram.

[0051] **FIG. 9**. Perspective view of coils of magnetic vortex wormhole generator.

[0052] **FIG. 10**. Graph showing that generator produces negative energy.

[0053] **FIG. 11**. Tetrahedron diagram showing that the speed of light at the Planck box boundary at the water molecule is one meter/second.

[0054] **FIG. 12**. Perspective view of proton decay into neutral pion and positron.

[0055] **FIG. 13**. Perspective view of pion decaying into two gamma photons.

[0056] **FIG. 14**. Diagram showing decay of the hydrogen atom into electrons.

[0057] **FIG. 15**. Perspective view of water injector and vacuum chamber.

[0058] **FIG. 16**. Perspective view of vacuum chamber mounted along centerline of magnetic vortex wormhole generator.

[0059] **FIG. 17**. Perspective view of vacuum pump used to evacuate vacuum chamber.

DETAILED DESCRIPTION OF THE INVENTION

[0060] 1. The coils of the magnetic vortex wormhole generator are made of three stacks of 0.020 inch silicon-iron transformer laminates. These are washed to remove the oil, and then wrapped with insulating tape in order to keep the laminations together. Using a very long bench made of wooden planks and 2×4 sawhorses, the outer coil is wrapped counter-clockwise right to left using a large spool of 14 AWG magnetic wire. A thin spacer is used between windings in order to reduce the winding capacitance. Once the outer coil is wound, the wire is continued to the second inner coil which is wrapped clockwise, leaving enough wire between coils such that when the coils are mounted in the wooden frame, the coil is one continuous winding having an input and output connection. Using an inductance meter, the inductance of the coil is measured. Using a standard frequency of 60 Hz, the capacitance of a sheet metal capacitor is calculated such that the generator is electromagnetically resonant at this frequency. The generator is connected to the line voltage by a 1:1 isolation transformer which is connected to a small primary coil wrapped on a toroidal core whose similar secondary coil is connected to the sheet metal capacitor and inductance coil. Resonance is achieved by adjusting the spacing and overlap of the sheet metal.

[0061] 2. The vacuum chamber is made of a glass tube with sufficient wall thickness to withstand the vacuum pressure. A number of glass blowing techniques are used to make the glass-electrode connection for the capacitor plates. Then circular pieces of glass plate are cut out and ground to the inside diameter of the tube, fitted with the nozzle and vacuum connection, and then heat sealed to the chamber. The chamber and water injector are then attached to a wooden bracket mounting which is doweled and glued to the wooden frame of the generator.

I claim:

1. A water energy generator system comprising the components:

a magnetic vortex wormhole generator and driving resonant electrical circuit;

a water droplet injector;

a vacuum chamber and vacuum pump; and

an electrostatic electron capture system.

2. By means of claim (**1**), a magnetic vortex wormhole generator comprising two concentric cylindrical coils of different radii wound in opposite directions, made of thin sheet silicon-iron transformer laminations wound with one continuous length of magnetic wire providing a single input and single output connection to the driving electrical circuit.

3. By means of claim (**2**), a coil winding method and oscillating driving circuit producing bucking electric fields along the centerline of the generator normal to the coils which create a spacetime curvature pressure and negative energy.

4. By means of claim (**3**), the generation of wormholes between space and hyperspace along the centerline of the generator such that low density, low speed of light hyperspace energy flows through a positive gravitational gradient from hyperspace to space.

5. By means of claim (**1**), a water injector comprising a solenoid-activated water valve, water supply connection, seal and nozzle for injecting water droplets into the vacuum chamber.

6. By means of claims (**5**) and (**4**), the softening and particle decay of the water molecules by the hyperspace energy into neutral pions, positrons, gamma photons and finally a cascade of electron pairs.

7. By means of claims (**6**) and (**1**), the capture of the electrons on electrostatically charged capacitor plates located in the vacuum chamber for the purpose of producing electrical energy.

8. By means of claims (**1**) and (**2**), a resonant electrical driving circuit comprising a line isolation transformer connected to a primary coil wound on a toroidal coil whose secondary output coil is connected to a capacitor and the inductance coil of the generator such that the capacitance of the capacitor and the inductance of the coil form a highly resonant electrical circuit.

* * * * *

US 20060168937A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0168937 A1**
St. Clair (43) **Pub. Date: Aug. 3, 2006**

(54) MAGNETIC MONOPOLE SPACECRAFT

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **11/048,026**

(22) Filed: **Jan. 31, 2005**

Publication Classification

(51) **Int. Cl.**
F03H 1/00 (2006.01)
(52) **U.S. Cl.** **60/203.1**

(57) ABSTRACT

A spacecraft propulsion system that utilizes a dual method of providing lift on the hull by means of magnetic monopoles and electromagnetic spacetime curvature pressure.

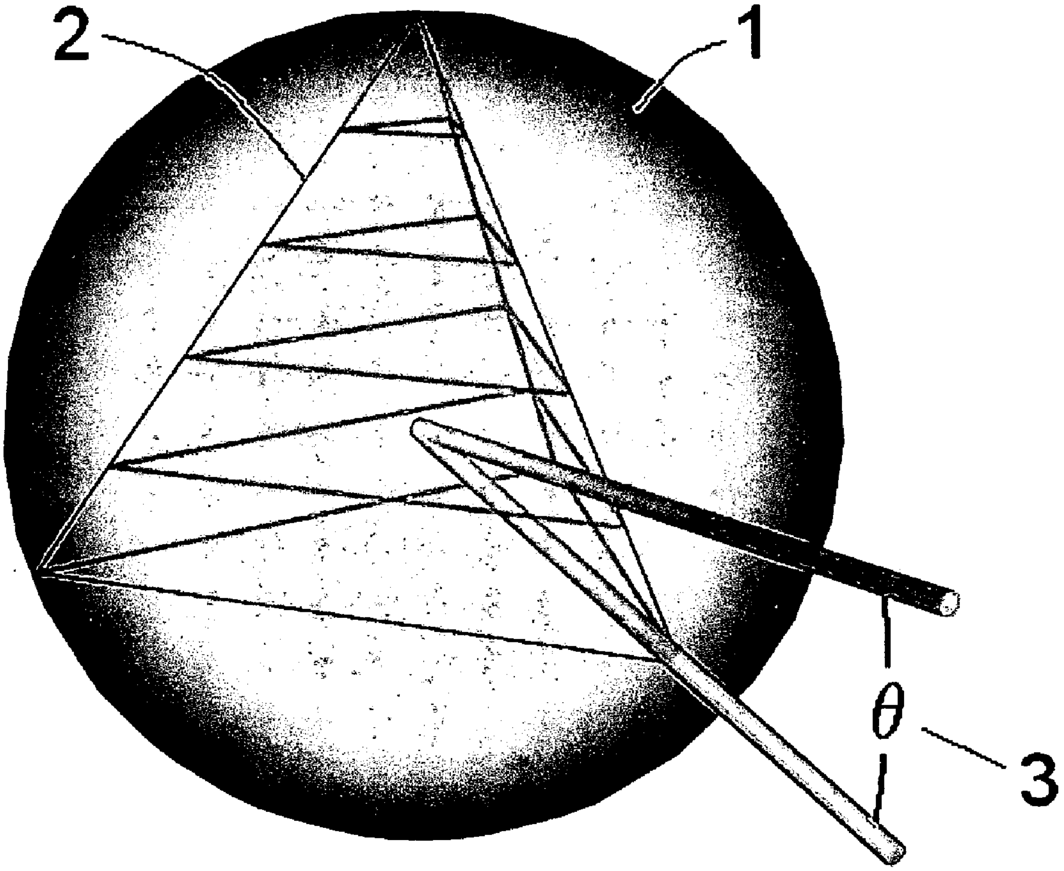

Figure 1

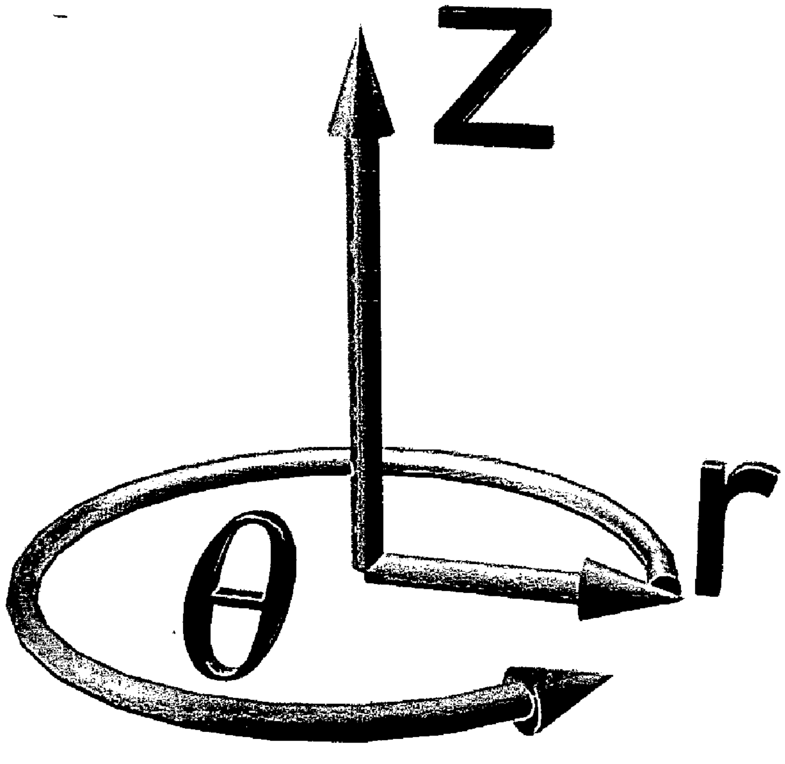

Figure 2

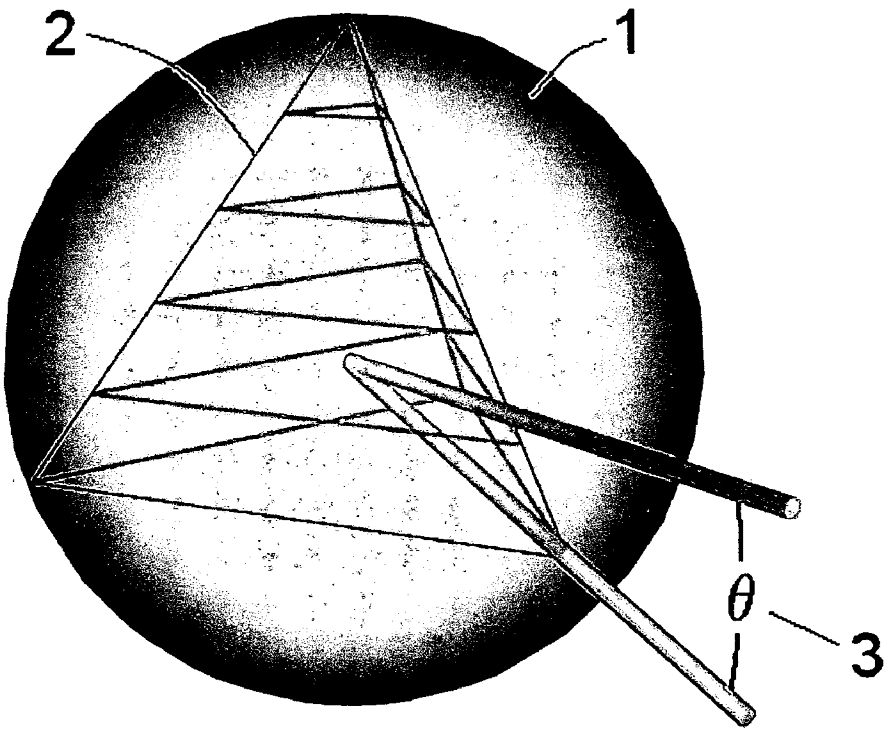

Figure 3

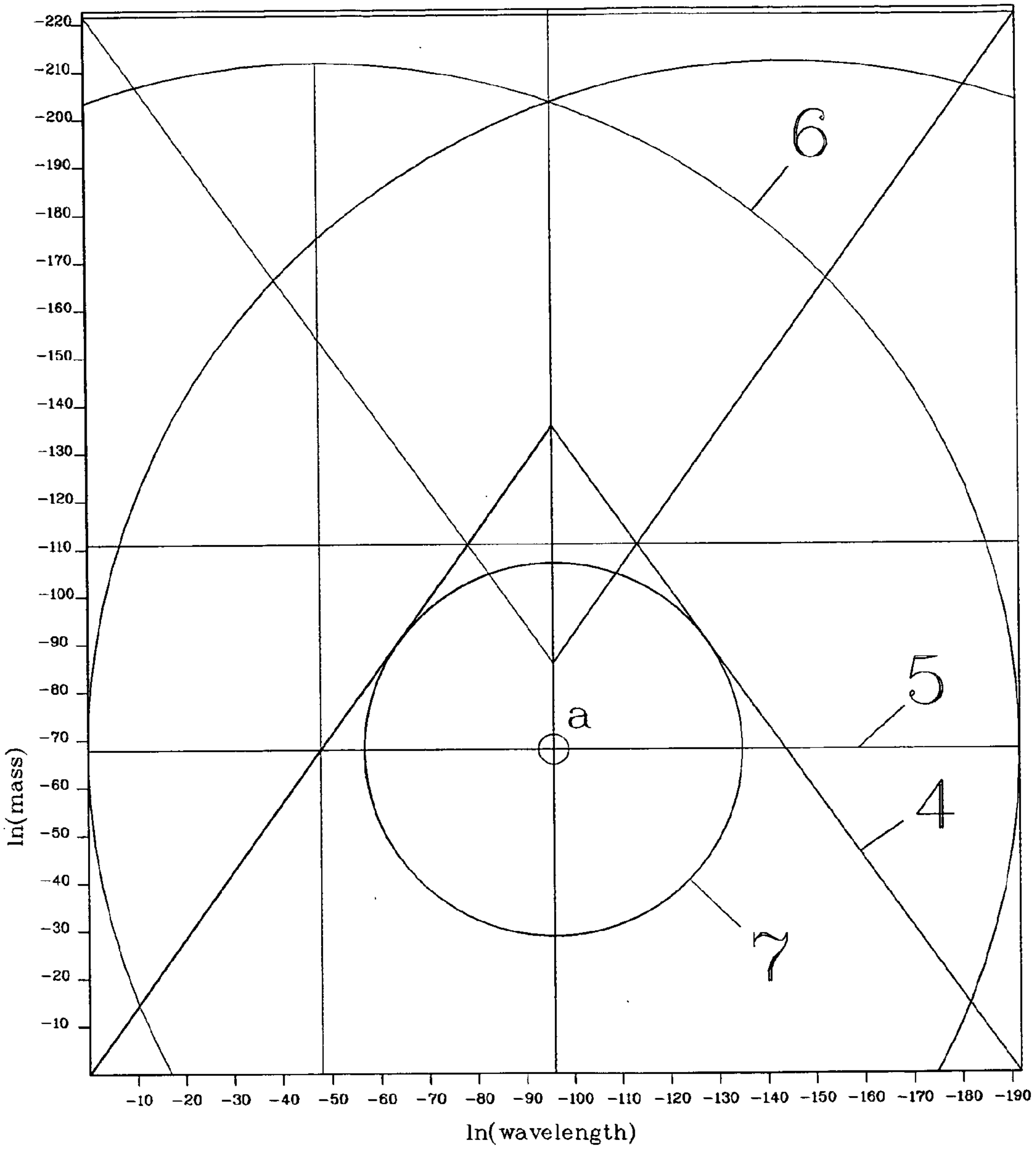
-220
-210
-200
-190
-180
-170
-160
-150
-140
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
ln(mass)
-10 -20 -30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130 -140 -150 -160 -170 -180 -190
ln(wavelength)
6
5
4
7
a

Figure 4

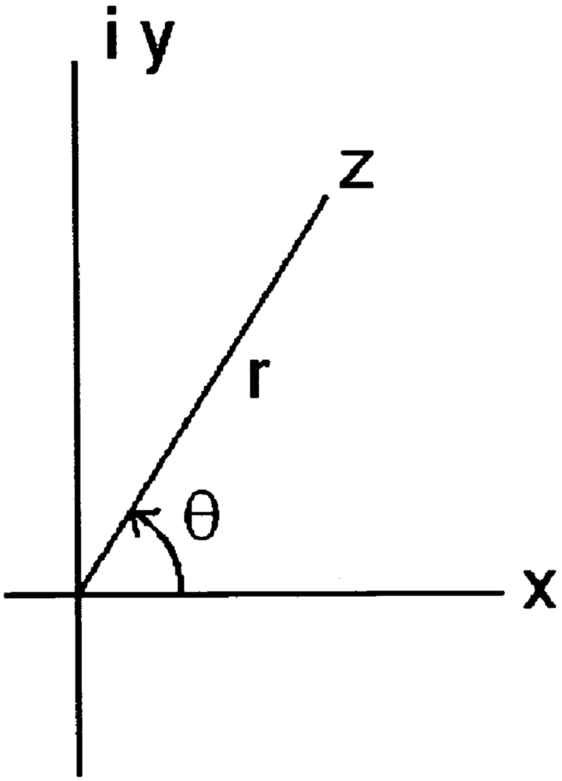

Figure 5

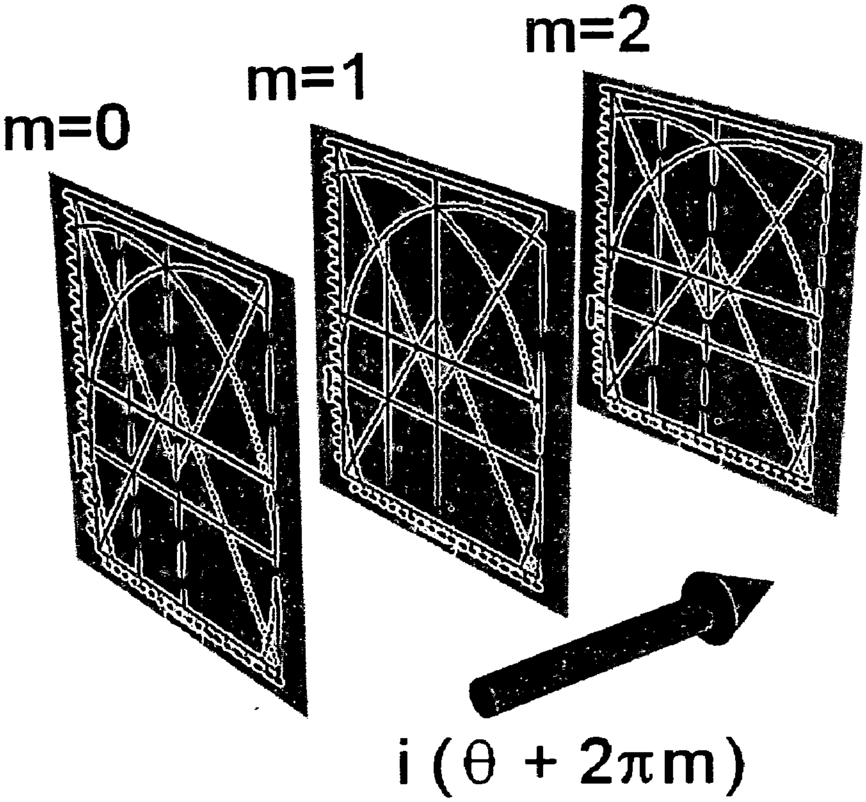
m=0
m=1
m=2
i (θ + 2πm)

Figure 6

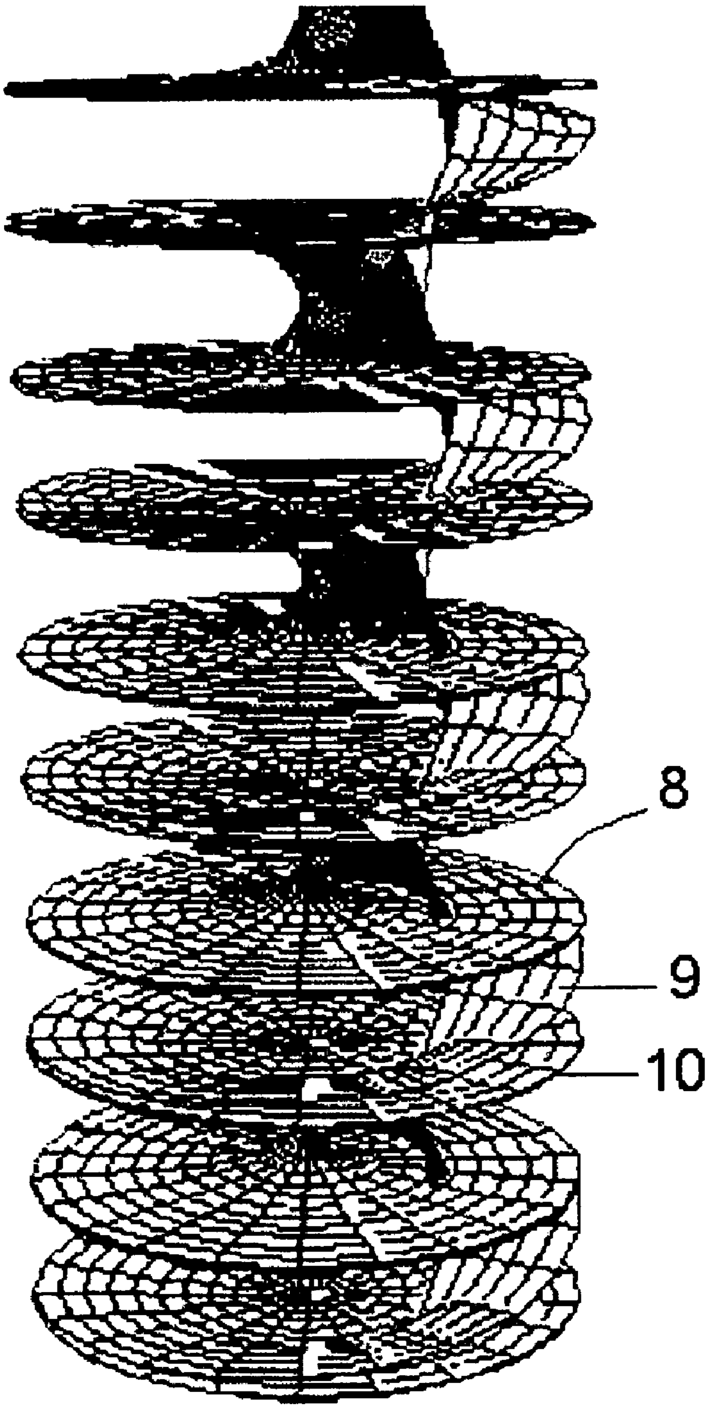

Figure 7

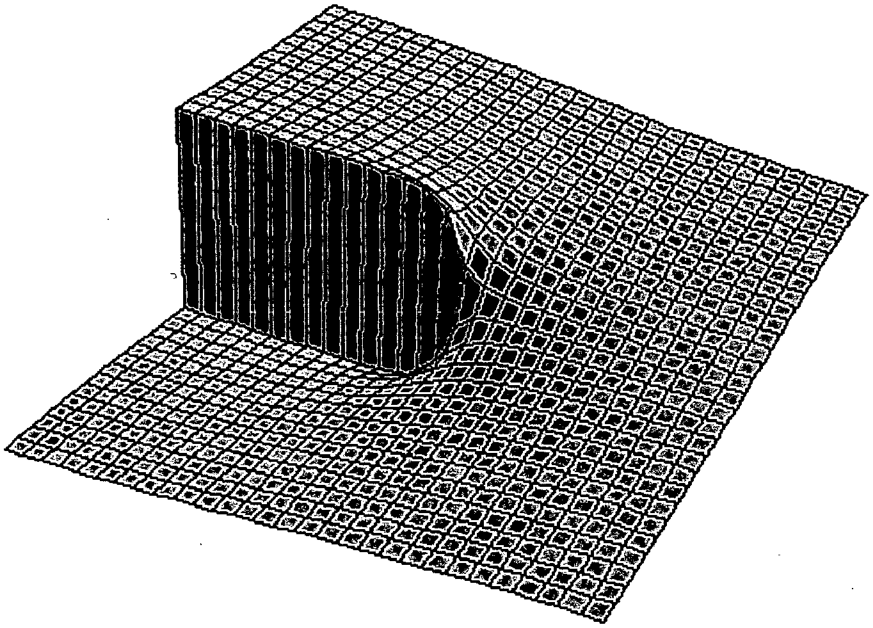

Figure 8

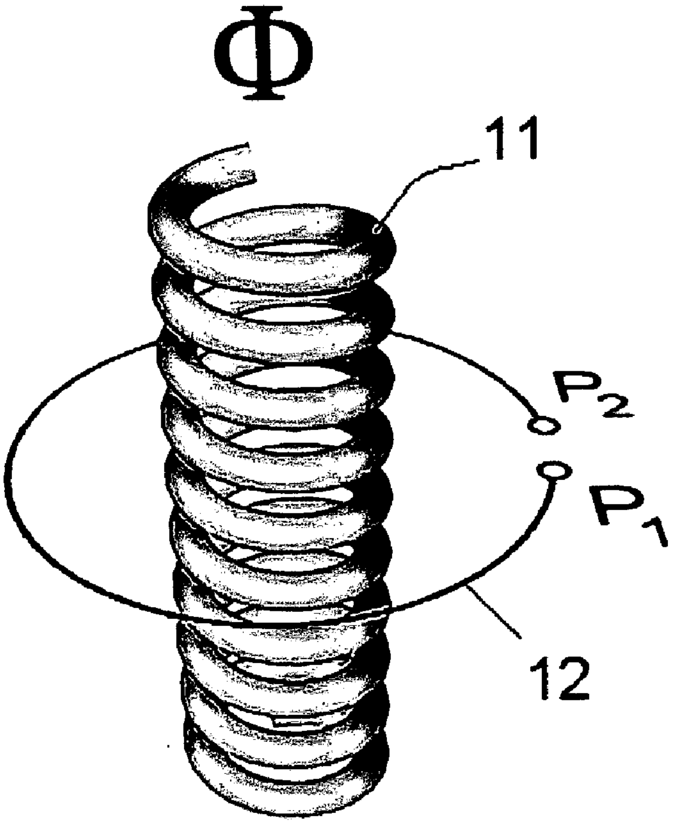

Figure 9

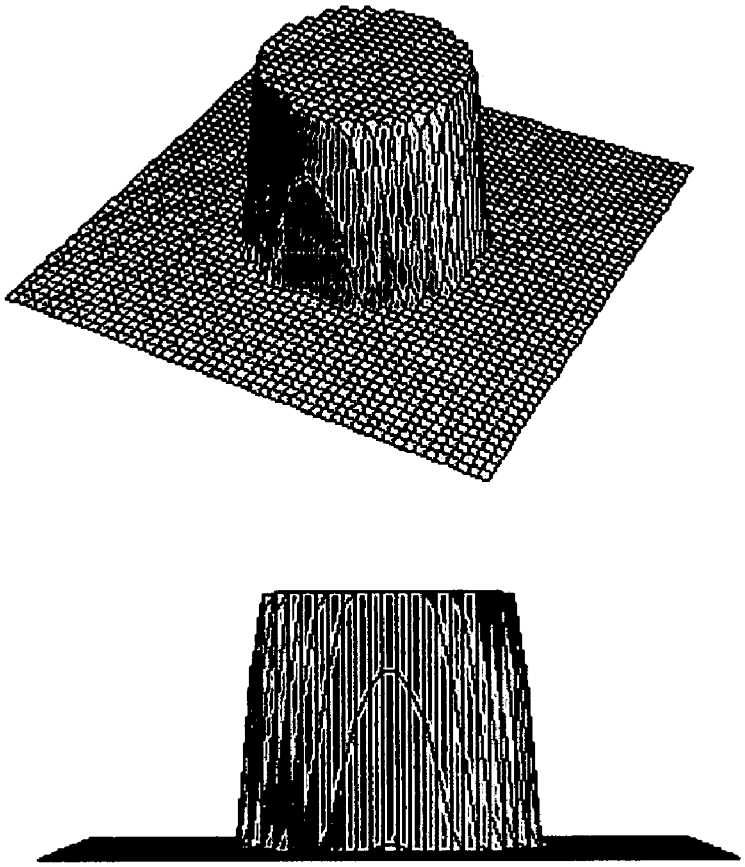

Figure 10

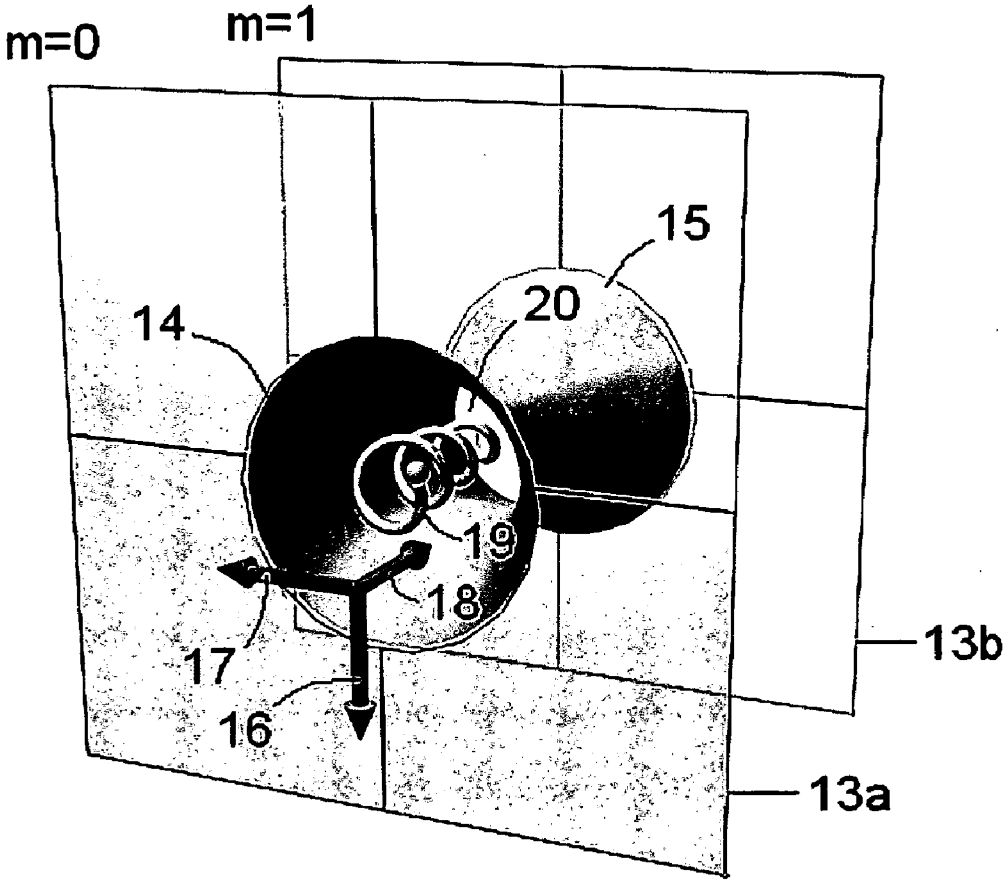
m=0
m=1
15
20
14
19
18
17
16
13b
13a

Figure 11

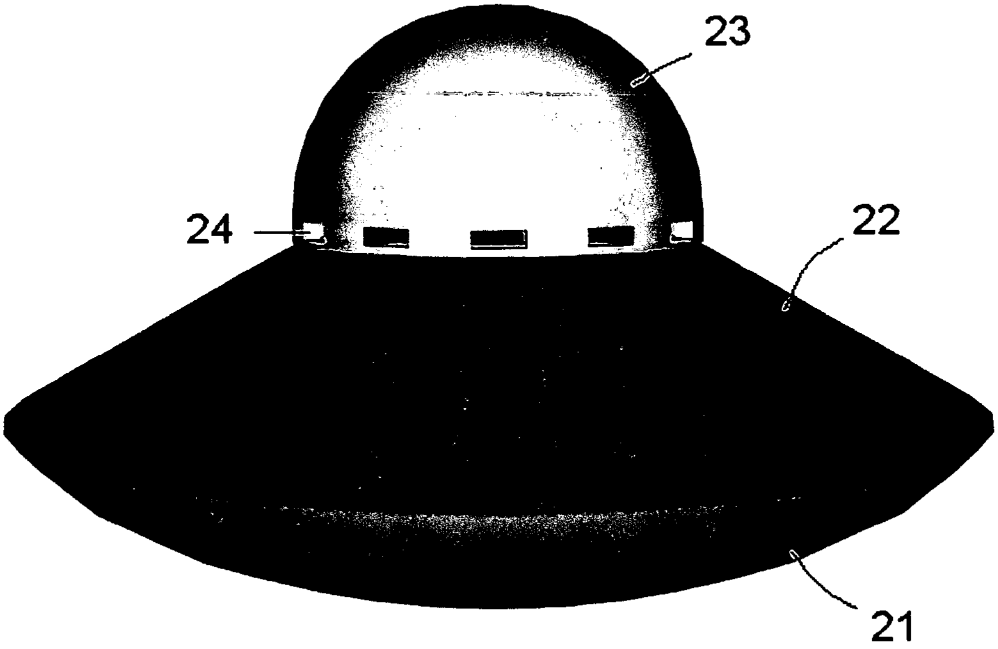

Figure 12

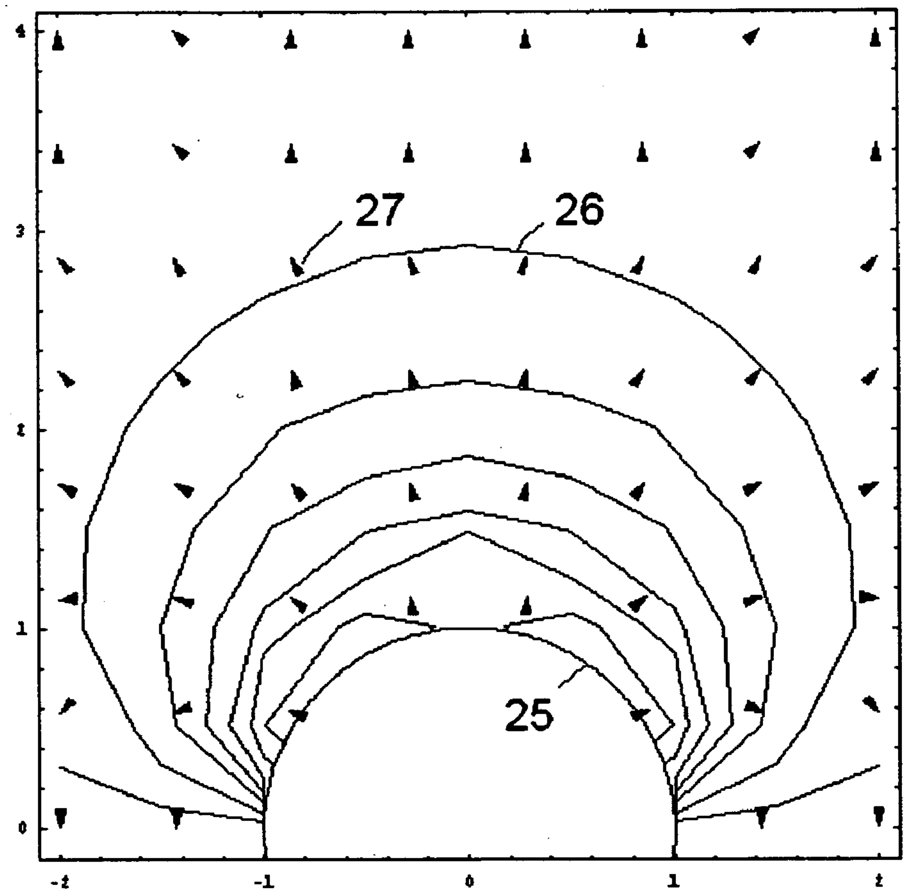

Figure 13

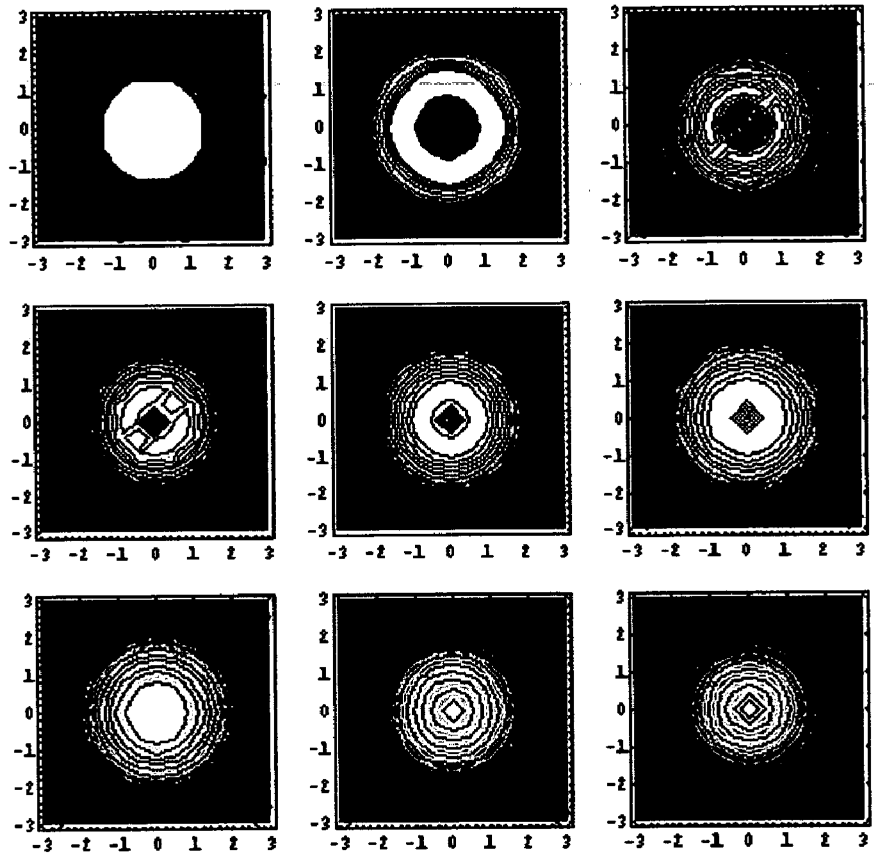

Figure 14

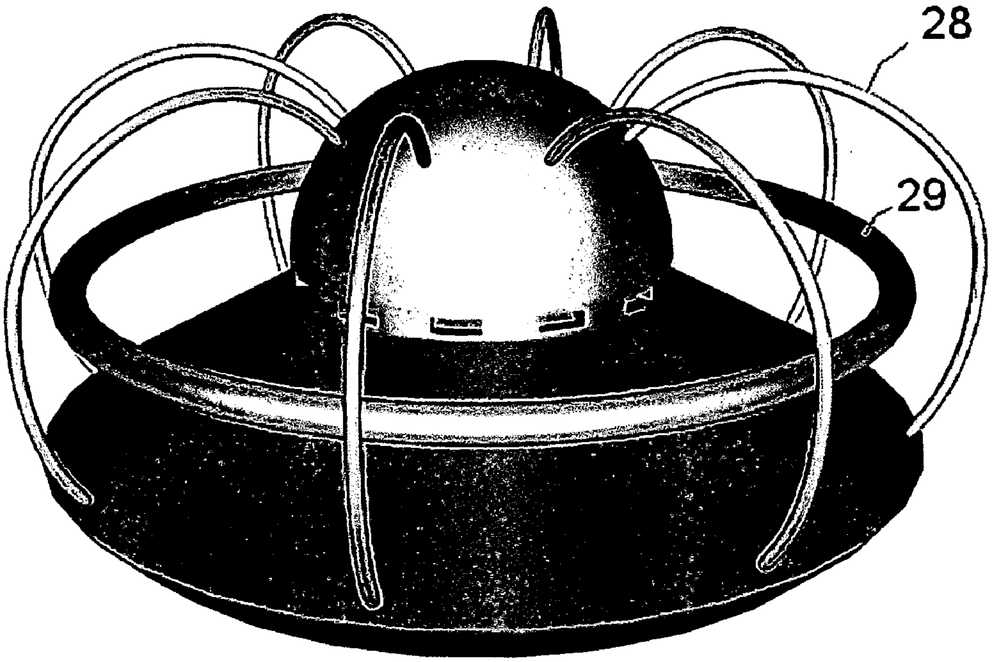

Figure 15

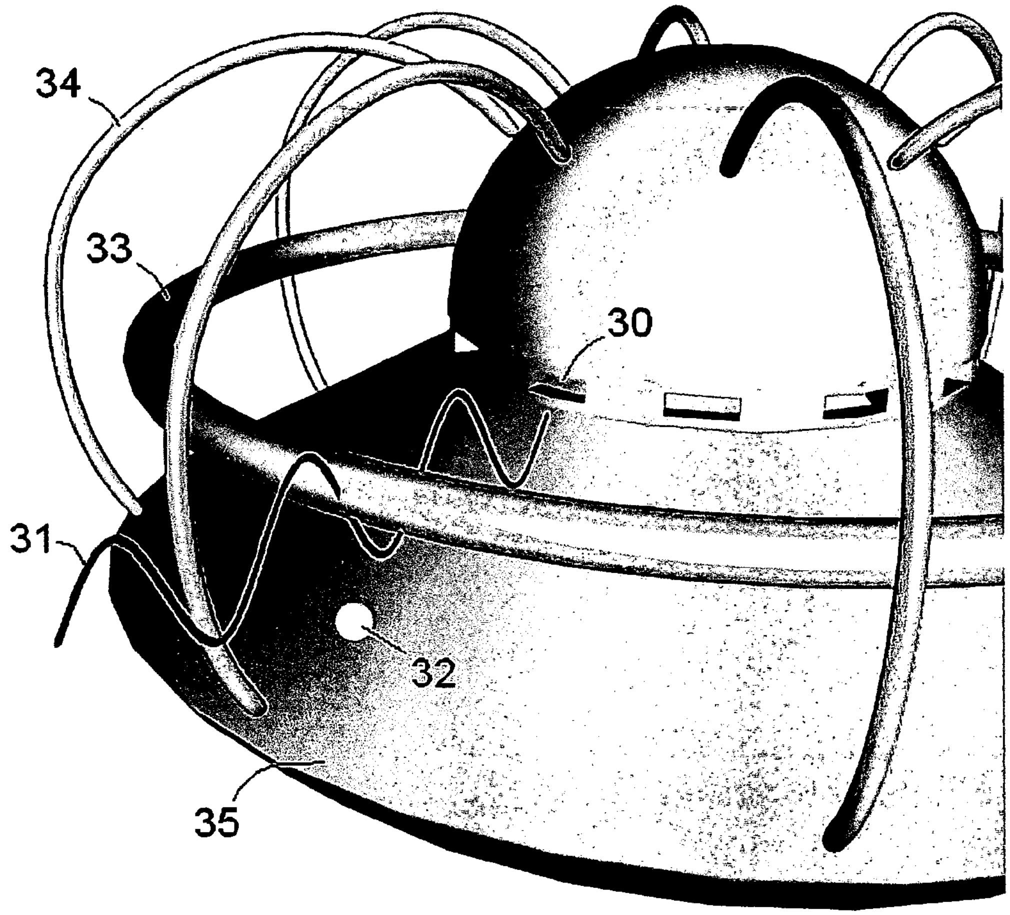
34
33
30
31
32
35

Figure 16

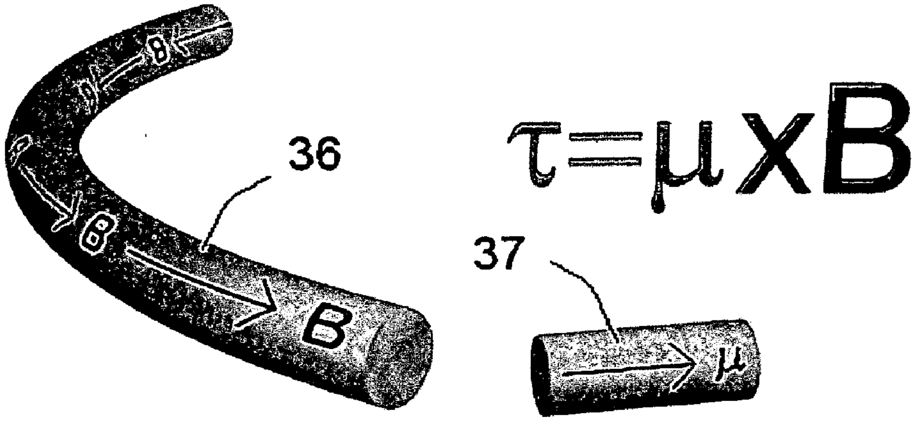
B
B
B
B
36
τ=μxB
37
μ

Figure 17

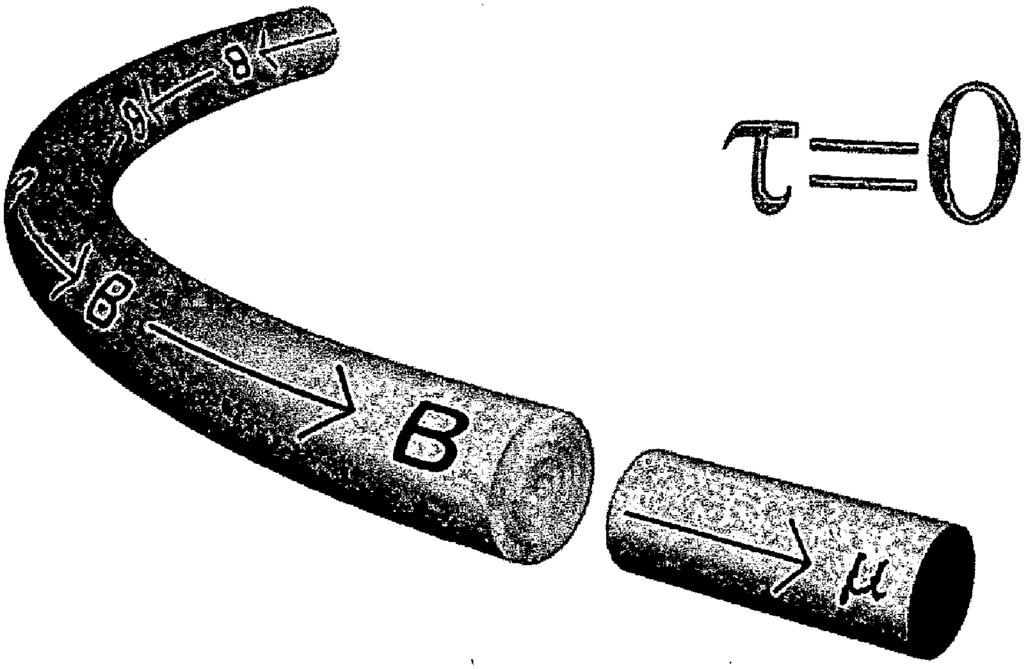

Figure 18

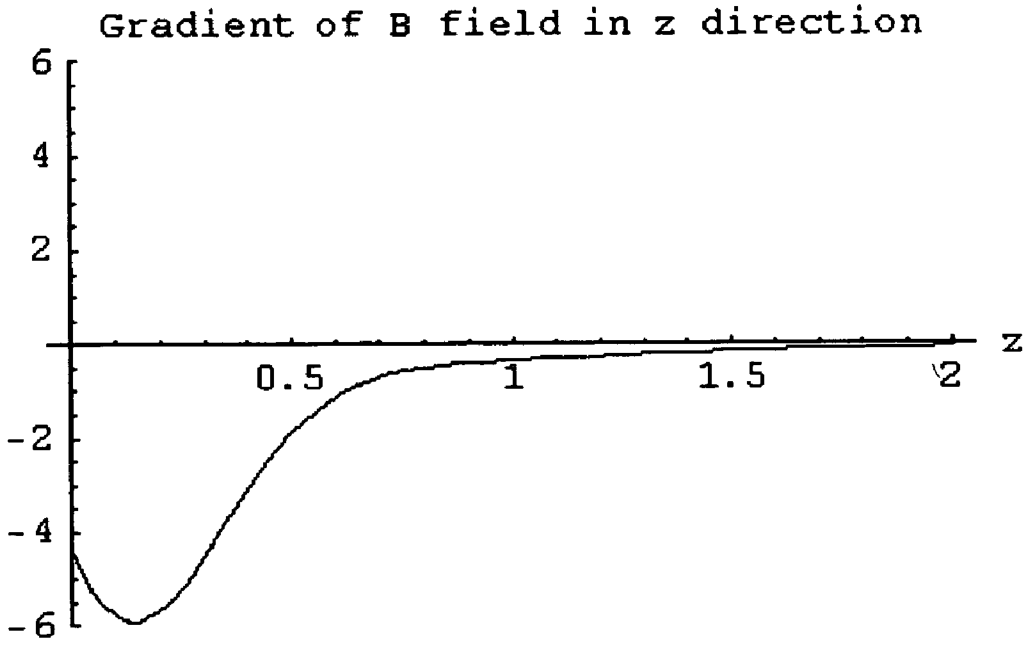

MAGNETIC MONOPOLE SPACECRAFT

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is a spacecraft propulsion system that generates a field of wormholes which are threaded with a magnetic field. Acting as two attracting magnets, the spacecraft's north magnetic field is attracted to the constantly regenerating south magnetic monopoles of the wormholes which provides lift on the hull.

BACKGROUND OF THE INVENTION

[0002] According to one of Maxwell's electromagnetic equations, the curl of the electric E field is equal to the negative time rate of change of the magnetic flux density B field.

$$\overline{V} \times E = -\frac{\partial B}{\partial t}$$

The curl can be thought of as a circulation around a closed loop specified by the right-hand rule where the fingers curl in the direction of the electric field and the thumb represents the changing magnetic flux density field through the area of the loop. At no time is the electric field diverging around the loop. That is, the divergence of the curl is zero which is a well-known vector operation

$$div\ \mathrm{curl}E = \frac{\partial}{\partial t}(divB) = 0$$

The partial derivatives of divB are zero at all points in space. Performing the integration, therefore, the divergence of B is equal to a constant

div B=constant

[0003] Referring to **FIG. 1**, the following discussion is made in cylindrical coordinates {r,θ,z}. In cylindrical coordinates, the divergence of the radial B field, div B, is equal to a constant C

$$Br'[r] + \frac{Br[r]}{r} = C$$

where the prime (') represents differentiation with respect to the radius r. The solution to this equation resides in determining the constant C. In the vacuum of space without any wormholes, the constant C is zero. Because the spacecraft is surrounded by a field of wormholes, there is a magnetic flux density field threading each one. That is, each wormhole is actually a magnetic monopole, and therefore the entire field of monopoles constitutes a large magnet with one pole in this dimension and the other pole in the hyperspace dimension.

[0004] The concept of the wormhole involves a new type of scientific thinking involving the creation of a gateway between our spacetime and that of a hyperspace co-dimension. The gateway is created electromagnetically as shown by my patent applications Rotating Magnetic Vortex Generator, Magnetic Vortex Wormhole Generator, and Sulfur 8 Wormhole Generator. The gateway can also be created ultrasonically through bubble cavitation as shown in my patent application Cavitating Oil Hyperspace Energy Generator. In one experiment, smoke was blown through one side of the coil of the magnetic vortex wormhole generator and no smoke came out the other side. The smoke was blown through the wormhole into another dimension.

[0005] The existence of hyperspace is not generally known in the scientific community. The reason it exists can be thought of in the following manner. Referring to **FIG. 2**, the corners of a tetrahedron (**2**) circumscribed by a sphere (**1**) touch the sphere at an angle (**3**) of −19.47°. Looking at the planets of the solar system, the Giant Red Spot vortex of Jupiter, which can hold two planets the size of Earth, is located at this angle. On Mars, the Olympic Mons volcano, which is the size of France, is located at north 19.5° Here in the Caribbean there is a slow moving rock mantle vortex at north 19.5° that curves the islands down toward Venezuela. So the geometry of space is related to the tetrahedron. What this suggests is that there is a subspace manifold whose tetrahedral geometry projects all the constants of physics into our dimension.

[0006] Referring to **FIG. 3**, this subspace geometry is shown in the tetrahedron diagram which plots the natural logarithm of mass on the vertical axis versus the natural logarithm of wavelength on the horizontal axis. With this diagram, it is possible to visualize the entire universe on a single sheet of paper. Triangle (**4**) is the tetrahedron whose vertical centerline shown by the small π circle (a) is equal to Planck's constant h divided by the speed of light c, known as the base constant. The base constant in logs is equal to −95.91546344. Line (**5**) is the diameter of the sphere (**6**). A circle (**7**), centered on (a), with a radius equal to the speed of light squared, is tangent to the tetrahedron. It can be shown that all the constants of physics, such as the speed of light squared and the Planck mass and the Planck wavelength, are determined geometrically by this logarithmic subspace tetrahedral geometry.

[0007] Referring to **FIG. 4**, the complex plane consists of a real horizontal axis, and a complex vertical axis where the value of the complex number z is given by a radius r and an angle θ

$$z=re^{i(\theta+2\pi m)}$$

The logarithm of z is

$$\mathrm{Log}[z]=\mathrm{Log}[r]+i(\theta+2\pi m)$$

where m is an integer m=0,1,2 . . . corresponding to multiple rotations of 2π. What this means in terms of the tetrahedron diagram, referring to **FIG. 5**, is that there are multiple diagrams separated by 2π rotations. Each multiple is another hyperspace dimension. Only the log manifold has this characteristic. Referring to **FIG. 6**, the hyperspace dimension, shown as 2π circular surface (**8**), has a cut (**9**) on the undefined Log[0] origin line such that another hyperspace dimension is created below it (**10**).

[0008] Referring to **FIG. 7**, this branch cut does not bring one back to the original surface nearing the origin. It takes one down to another level of the universe into another hyperspace dimension. I can attest personally that I have been able to look into another hyperspace co-dimension as well as jump into another dimension.

[0009] Furthermore, Dr. Stephen Hawking of Cambridge University has shown that our dimension is connected to a wormhole through complex time. That is, the hyperspace dimension is rotated forward by 90° which makes it orthogonal to us. While this is the mathematical explanation as to why there are hyperspace co-dimensions, I can attest personally to the fact, as described in my patent application Full Body Teleportation, that I was teleported through hyperspace and returned to our dimension over a distance of 100 meters. Because hyperspace exists, it is then possible, using electromagnetic fields, to open wormholes between our dimension and other hyperspace dimensions.

[0010] Referring to **FIG. 8**, the constant C in the magnetic flux density B field equation is determined as follows. The end of a solenoid (**11**) is equivalent to a local magnetic source of flux Φ which represents the wormhole. A charge (**12**) encircling the solenoid has the same value at P_1 and P_2 but there is a phase difference of 2 π n where n is an integer equal to the number of times the charge encircles the solenoid. The change in phase is equal to the charge q divided by Planck's reduced constant times the flux for a solenoid of radius r=R

$$\Delta\theta = 2\pi n = \frac{q}{\hbar}\Phi = \frac{q}{\hbar}B\pi R^2$$

Solving for the magnetic flux density field threading the wormhole

$$B = \frac{2\pi nh}{q\pi R^2}$$

The differential equation becomes

$$Br'[r] + \frac{Br[r]}{r} = C[2]$$

Solving the equation for the radial field Br

$$Br[r] = \frac{C[2]r}{2} + \frac{C[1]}{r}$$

Equating this Br field with the B field and solving for the C[2] constant

$$C[2] = \frac{-2[-2\hbar n + qrC[1]]}{qr^2}$$

Substituting for C[2], the C[1] constant drops out and therefore the wormhole magnetic flux density field becomes a constant inside the throat radius R

$$Br = \frac{2hn}{qR^2}$$

but the divergence moving from outside to inside the radius is a constant due to the discontinuity.

[0011] Referring to **FIG. 9**, the wormhole has a constant cylindrically-shaped magnetic field of radius R with a negative south pole due to the negative charge on the electron, Using a wormhole radius equal to one hundred times the electron radius with n equal to 10 turns, the magnetic flux density B field has a magnitude of 1.4×10^6 tesla.

[0012] Referring to **FIG. 10**, the wormhole surface (**14**) is a connection through a throat area (**20**) between space and hyperspace (**13**) to another surface in hyperspace (**15**). The electron path (**19**) is actually spiraling down the surface of the wormhole. The magnetic flux density field (**18**) points into the wormhole such that the cross product of the electron velocity (**16**) with the field (v×B) points toward the outside rim of the wormhole. Because the electron has a negative charge, the force (**17**) on the electron (−q v×B) is inward toward the centerline of the wormhole. Since one pole of the field is in space, and the other pole is in hyperspace, the wormhole appears to us as a magnetic monopole. The key to this invention is how to generate this magnetic wormhole field.

[0013] Referring to **FIG. 11**, the aluminum spacecraft is constructed of a flattened, shallow spherical lower hull (**21**), a circular upper hull with a flat sloping surface (**22**), a spherical dome cupola (**23**), and a cylindrical section (**24**) housing a circular array of radial microwave waveguides. Surfaces (**22**) and (**23**) are electrically charged, using high-voltage transformers, to an alternating electrostatic potential such that the potential on the dome is +V when the sloping hull has a −V potential and vice versa. This creates an electric field from the positively charged surface to the negatively charged surface.

[0014] Referring to **FIG. 12**, the dome (**25**) is charged to a positive potential. The spherical potential lines (**26**) emanate from the dome and curve around toward the sloping hull. The negative gradient of this potential field is the electric E field (**27**) which is perpendicular to the potential lines. The electric field lines from the dome then terminate on the sloping hull which can be seen in the lower right hand corner of the graph.

[0015] Looking down from above at the top of the dome, referring to **FIG. 13**, the oscillating electric field generates a circular oscillating magnetic field around the hull at various elevations. The last two graphs in the right-hand corner are near the top of the dome as seen by the smaller radius contour lines.

[0016] Referring to **FIG. 14**, the oscillating electric field lines (**28**) are between the dome and the sloping hull. The oscillating magnetic field (**29**) encircles the hull at the level of the microwave waveguides.

[0017] From gravitation physics, it is known that negative energy is required to open up the throat of the wormhole. In terms of differential forms mathematics, the negative energy is created by wedging the magnetic field with the radial microwave beams of the waveguides. This generates an increasing time rate of change of negative energy ρ as shown by

$$*d(B\wedge(B_{wave}+E_{wave}dt))=\frac{\partial(-\rho)}{\partial t}$$

where (*) is the Hodge Star operator, (d) the differential operator and (^) the wedge operator which joins the circular magnetic flux density B field with the electromagnetic wave (B_{wave}, E_{wave}).

[0018] Referring to **FIG. 15**, the interaction of the circular magnetic field (**33**), generated by the oscillating electric field (**34**), interacts with the radial electromagnetic field of the microwave beams (**31**) of the waveguides (**30**) to generate negative energy (**32**) over the sloping hull (**35**).

[0019] This combination of fields also creates the wormhole field over the hull. The spacetime curvature pressure T in the vertical z-direction is equal to the square of the circular magnetic flux density field

$$T^{zz}=\frac{B^2}{8\pi}$$

This stress-energy-momentum tensor can be thought of as a spacetime curvature proportional to the inverse of the radius squared, or as a pressure term which acts on the surface area of the hull. Thus there is the combination of a pressure stress and negative energy which creates the wormhole field over the sloping hull. Even though the magnetic flux density B field is oscillating, it is the square of the field which creates the stress. Thus the tension is still in the positive vertical z-direction.

[0020] Due to the low speed of light of hyperspace, the hyperspace energy is low density. Thus there is a positive gravitational potential between hyperspace and our dimension such that the hyperspace energy flows through the wormholes and onto the hull of the spacecraft. When this happens, a white mist forms over the hull. The effect of the hyperspace energy is to lessen the mass of the spacecraft, and relativistically increase the strength of the electromagnetic fields due to the smaller speed of light.

[0021] Referring to **FIG. 16**, the circular magnetic flux density field (**36**) is interacting with the magnetic monopole (**37**). The monopole acts as one pole of a magnet and therefore has a magnetic moment equal to the area times the electrical current circulating through it as the electrons move from our dimension into hyperspace. The great physicist Maxwell (**1870**) pointed out that the energy associated with charges and poles is potential energy and that therefore these objects tend to move in a direction that will decrease the potential energy, similar to a brick sliding down an inclined plane. Now to reduce the potential energy is the same as to reduce the field which gives a measure of the potential energy. If two like charges are brought together they strengthen one another's field, while opposite charges reduce one another's field. Thus like charges repel and unlike ones attract.

[0022] However, the energy associated with electric currents is not potential but kinetic in origin, because it is associated with moving charges. Now in mechanics it is well known that bodies move in a direction to increase their kinetic energy, if there is a source of external energy. The magnetic field of currents is a measure of kinetic energy and currents will try to move in a direction that will increase the field. Thus similarly-directed currents attract and unlike currents repel. In **FIG. 16**, the monopole will act so as to increase the field of the circular magnetic B field generated by the spacecraft. The monopole will then rotate into alignment with the flux tube. This increases the field strength of the flux tube. There is a torque τ on the monopole equal to the cross product of the magnetic moment with the B field

$$\tau=\mu\times B$$

[0023] Referring to **FIG. 17**, when the monopole is aligned with the flux tube, the cross product is zero and there is zero torque on the monopole. Notice that even if the oscillating B field points in the opposite direction, there is still zero torque since the torque depends on the sine of the angle between them. The sin(0°) or sin(180°) is the same zero value. With the monopole aligned with the flux tube, the kinetic energy is now maximized. That is, both the flux tube and the monopole point in the θ-direction.

[0024] The magnetic B field can be represented as a vector having three components. The only component is in the angular direction

$$B=\{B_r,B_\theta,B_z\}=\{0,\ \mathrm{B}_\theta(z),\ 0\}$$

where the field varies, as was seen in **FIG. 13**, in the vertical z-direction. The magnetic monopole field also points in the θ-direction

$$\mu=\{0,\mu_\theta(z),0\}$$

where there is a negative gradient of the monopole field in the z-direction due to the location of the microwave waveguides near the sloping hull.

[0025] The force F on the monopole is the gradient of the monopole's magnetic moment μ with the magnetic flux density B field

$$F=\nabla(\mu\cdot B)=\{0,\ 0,\ \mu_\theta\, B_\theta'(z)+B_\theta(z)\mu_\theta'(z)\}$$

which says that there is a force on the monopole in the z-direction equal to the magnetic moment times the gradient of the magnetic field in the z-direction plus the magnetic field times the gradient of the magnetic moment in the z-direction.

[0026] Referring to **FIG. 18**. differentiating the magnetic field in the z-direction shows that the gradient is negative outside the hull. This can also be seen visually in **FIG. 13** where the graphs decrease in intensity.

[0027] A negative gradient for both the magnetic moment and the field means that the force on the monopole is negative. The force on the tube connected to the hull is therefore the negative of a negative, yielding a positive lift force.

$$F_{monopole}=-\{0,\ 0,|\mu_\theta B_\theta'(z)+B_\theta(z)\mu_\theta'(z)|\}=-F_{tube}$$

$$F_{tube}=+\{0,\ 0,|\mu_\theta B_\theta'(z)+B_\theta(z)\mu_\theta'(z)|\}$$

which says there is an upward lift force on the hull due to the combination magnetic monopole and flux tube. This lift force is in addition to the vertical lift force generated by the spacetime curvature created by the electromagnetic fields themselves.

SUMMARY OF THE INVENTION

[0028] This invention is a spacecraft propulsion system that utilizes electromagnetic fields and microwaves to generate negative energy and a spacetime curvature over the hull. The hull consists of a hemispherical dome, a circular array of radial microwave waveguides, a sloping flat hull, and a shallow spherical hull on the bottom. Alternating current high voltage transformers connected to the dome and the sloping hull generate a curving oscillating electric field between the dome and the sloping hull. Due to this oscillation, a horizontal circular oscillating magnetic flux density B field is created around the dome.

[0029] By firing the microwaves at right angles to the B field, negative energy is created over the hull. The negative energy and spacetime curvature pressure generate wormholes between space and hyperspace. Because hyperspace has a low speed of light and positive gravitational potential, low density hyperspace energy flows through the wormholes and onto the hull. The effect of the hyperspace energy is to lessen the mass of the vehicle and to increase the strength of the electromagnetic fields. Because the resistance of hyperspace is less than the resistance of space, electrons spiral down the wormholes into hyperspace. This creates a magnetic field through the wormhole with one pole in our dimension and the other pole in hyperspace. Thus a field of magnetic monopoles is created over the hull.

[0030] The magnetic monopoles, which represent kinetic energy, align themselves with the magnetic flux tubes in order to maximize the total magnetic field. Because there is a gradient of the monopoles and field in the vertical direction, a negative force develops on the monopoles equal to the gradient of the dot product of the magnetic moment of the monopole with the B field. Thus the opposite reaction is a positive force on the flux tubes attached to the hull which is equivalent to bringing the north pole of a magnet together with the south pole of a second magnet. Because the hull constantly regenerates the wormhole field, the hull experiences a constant upward lift force. This is in addition to the lift generated by the spacetime curvature pressure which is proportional to the square of the magnetic flux density B field.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0031] **FIG. 1**. Perspective view of cylindrical coordinate system $\{r, \theta, z\}$.

[0032] **FIG. 2**. Perspective view of tetrahedron circumscribed by sphere.

[0033] **FIG. 3**. Tetrahedron diagram showing speed of light squared is determined by the tetrahedron.

[0034] **FIG. 4**. Complex number z representation in the complex plane.

[0035] **FIG. 5**. Perspective view showing multiple log manifold hyperspace dimensions.

[0036] **FIG. 6**. Perspective view of orthogonal hyperspace dimensions.

[0037] **FIG. 7**. Perspective view of Log[z] showing cut along origin.

[0038] **FIG. 8**. Perspective view of charge phase shift around a magnetic flux.

[0039] **FIG. 9**. Perspective view of wormhole magnetic flux density B field.

[0040] **FIG. 10**. Perspective view of magnetic monopole wormhole.

[0041] **FIG. 11**. Perspective view of spacecraft.

[0042] **FIG. 12**. Graph showing electrostatic potential and electric field over dome.

[0043] **FIG. 13**. Animation showing circular magnetic field around dome at increasing elevation.

[0044] **FIG. 14**. Perspective view of electric and magnetic fields around hull.

[0045] **FIG. 15**. Perspective view of generation of negative energy.

[0046] **FIG. 16**. Perspective view of monopole misaligned with flux tube with torque.

[0047] **FIG. 17**. Perspective view of monopole aligned with flux tube at zero torque.

[0048] **FIG. 18**. Graph showing negative gradient of flux tube in the z-direction.

DETAILED DESCRIPTION OF THE INVENTION

[0049] 1. The aluminum hull is made by a technique called stretch forming which uses hydraulic cylinders to stretch a large sheet of aluminum to its yield point. This makes the aluminum sheet soft and pliable. Using a die which has been CNC machined to the desired hull profile, the sheet is then die pressed into a very rigid, smooth and lightweight structure requiring no other support. There are actually three dies consisting of a spherical dome, sloping hull, and shallow spherical dome.

[0050] 2. The rest of the hull consists of a cylindrical hull with a radius equal to the upper dome. A segment of this hull is designed on a 3D computer graphics program and stored as a stereolithography *.stl file. The file is then transmitted over the Internet to a server who prints up the part on an xy-plotter with an ultraviolet laser and ultraviolet light sensitive polymer bath. The computer model is sliced by a special program into many thousands of slices which are printed one over the other until the part is completed. The server returns the part next day by Express Mail. Several parts are then molded using liquid plastic such as to form the complete ring. A sand mold is then constructed from all the molds to form a plastic cylindrical hull having the waveguide slots molded into it. The waveguide silver-coated aluminum boxes are then installed in the slots and connected to the frequency generators and amplifiers. The purpose of the plastic waveguide cylinder is to separate the electrostatic charges on the dome and the sloping hull In this particular case we used a dome from another spacecraft design which saved on the cost of the die.

I claim:

1. A spacecraft propulsion system consisting of the following components:

a lower hull made of aluminum sheet having a shallow spherical profile;

a circular flat sloping hull made of aluminum sheet attached to the top of the lower hull on the periphery;

an electrically-insulated plastic-molded tubular cylindrical hull containing slots for mounting an array of radial microwave waveguides, attached to top of the flat sloping hull;

a hemispherical cupola in the shape of a dome made of aluminum sheet mounted on top of the insulated cylindrical hull;

an array of rectangular microwave waveguides mounted in the waveguide slots of the cylindrical hull;

a high-voltage alternating current transformer with one side electrically attached to the upper dome and the other side electrically attached to the flat sloping hull; and

a frequency generator and amplifier to drive the microwave waveguides.

2. By means of claim (**1**), an oscillating electric field is created between the upper dome and the sloping hull using the high-voltage alternating current transformer.

3. By means of claim (**2**), an oscillating circular magnetic flux density field is generated around the sloping hull and upper dome.

4. By means of claims (**1**) and (**3**), negative energy is generated by the radial microwave beams of the waveguide array impinging on the circular magnetic flux density field around the hull.

5. By means of claim (**3**), a positive spacetime curvature pressure constituting a lift force is developed over the hull in the vertical direction proportional to the square of the field.

6. By means of claims (**4**) and (**5**), a field of wormholes between space and hyperspace are generated over the hull.

7. By means of claim (**6**), due to the positive gravitational potential between hyperspace and space, low-density hyperspace energy flows through the wormholes onto the hull to reduce the mass of the spacecraft and strengthen relativistically the electromagnetic fields.

8. By means of claim (**6**), electrons, emitted by the charged hulls, spiraling down the wormholes, generate a field of magnetic monopoles with one pole in space and the other in hyperspace.

9. By means of claims (**6**) and (**3**), the magnetic monopoles maximize their kinetic field energy by aligning with the magnetic flux tube.

10. By means of claim (**9**), the gradient in the vertical direction of the dot product of the magnetic moments of the monopoles with the magnetic flux density field is a negative force on the monopoles and an equal but opposite positive lift force on the magnetic flux tubes attached to hull.

11. By means of claims (**10**) and (**5**), a dual method of providing apositive lift force on the spacecraft is constituted.

* * * * *

US 20060073976A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0073976 A1**

Pohlman (43) **Pub. Date: Apr. 6, 2006**

(54) METHOD OF GRAVITY DISTORTION AND TIME DISPLACEMENT

(76) Inventor: **Marlin B. Pohlman**, Tulsa, OK (US)

Correspondence Address:
STITES & HARBISON PLLC
1199 NORTH FAIRFAX STREET
SUITE 900
ALEXANDRIA, VA 22314 (US)

(21) Appl. No.: **10/954,767**

(22) Filed: **Oct. 1, 2004**

Publication Classification

(51) **Int. Cl.**

F16C	***39/06***	(2006.01)
G01V	***7/00***	(2006.01)
G09B	***23/06***	(2006.01)
F41B	***6/00***	(2006.01)
H01S	***3/09***	(2006.01)

(52) **U.S. Cl.** .... **505/166**; 505/180; 434/300; 73/382 G; 505/164

(57) ABSTRACT

A method for employing sinusoidal oscillations of electrical bombardment on the surface of one Kerr type singularity in close proximity to a second Kerr type singularity in such a method to take advantage of the Lense-Thirring effect, to simulate the effect of two point masses on nearly radial orbits in a 2+1 dimensional anti-de Sitter space resulting in creation of circular timelike geodesics conforming to the van Stockum under the Van Den Broeck modification of the Alcubierre geometry (Van Den Broeck 1999) permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) which results in a method for the formation of Godel-type geodesically complete spacetime envelopes complete with closed timelike curves.

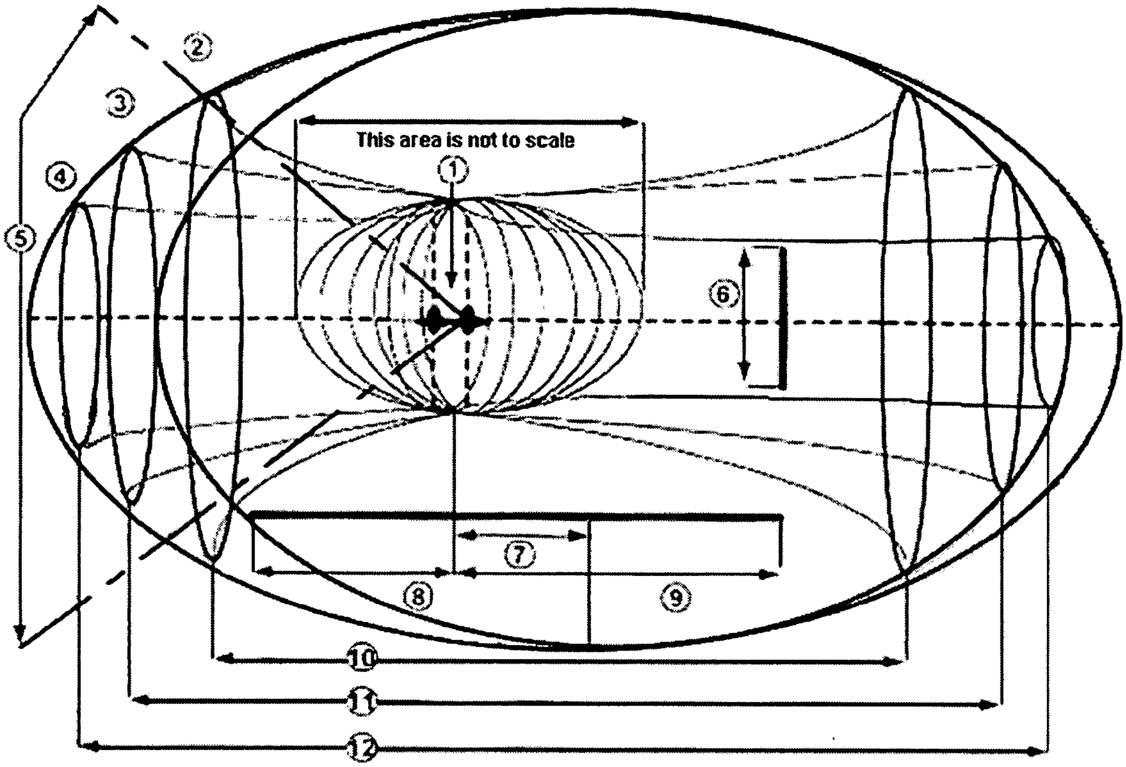

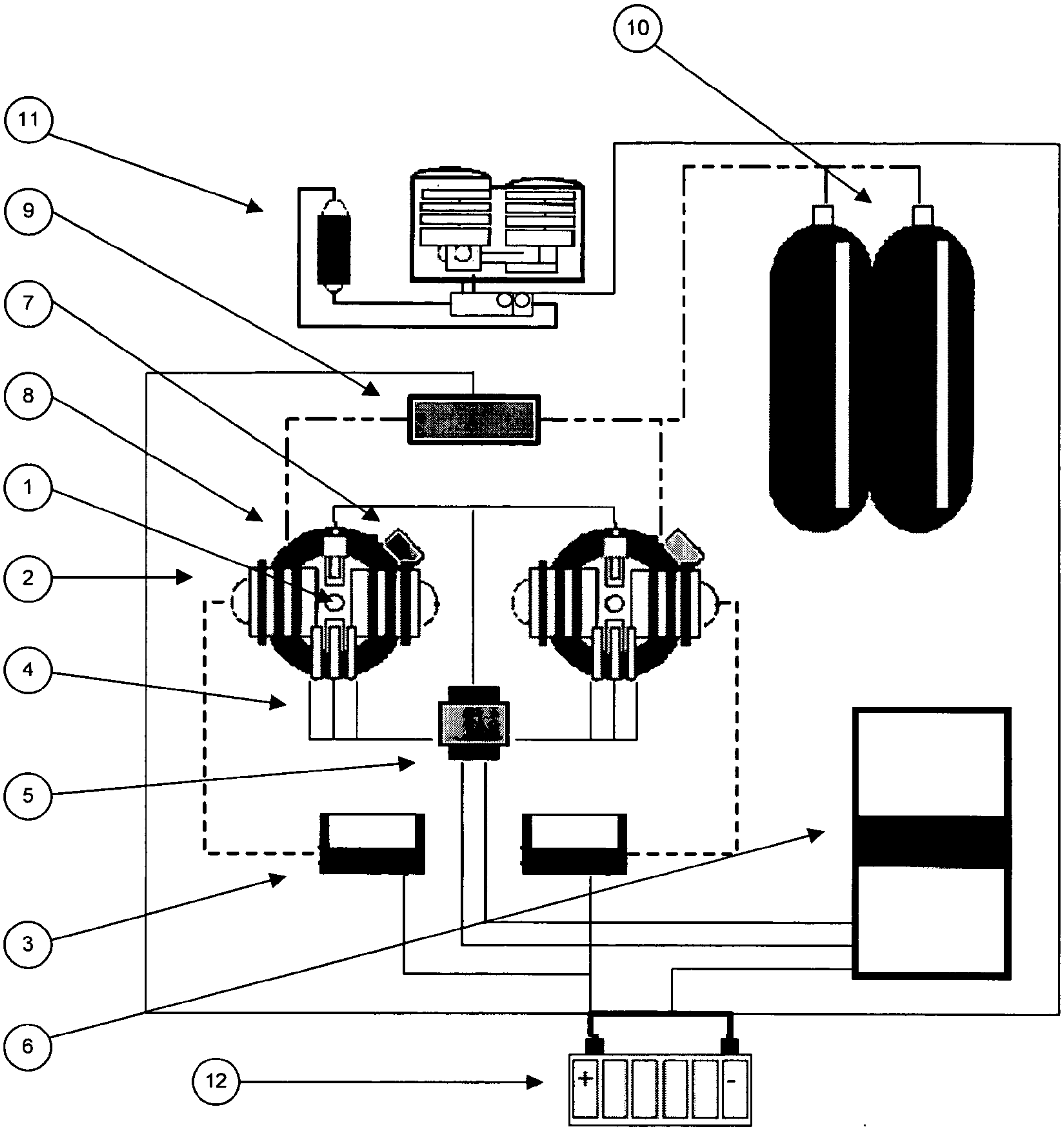
10
11
9
7
8
1
2
4
5
3
6
12
+
-

Figure 1

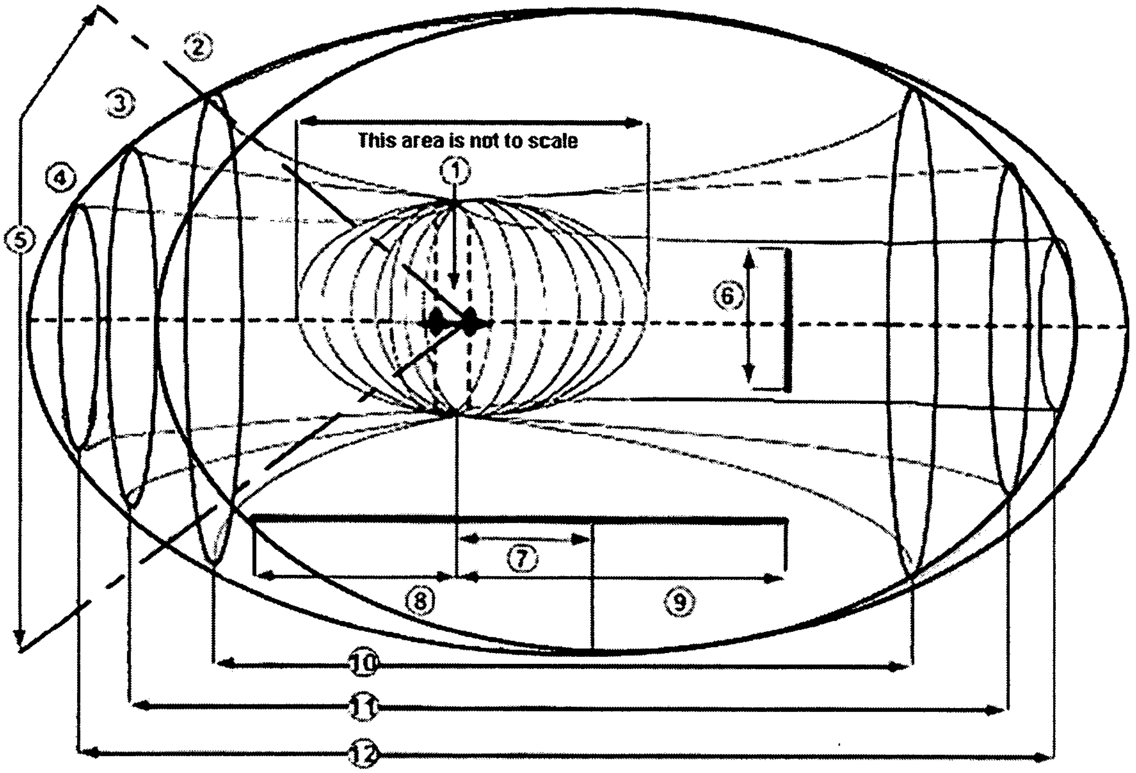
This area is not to scale
①
②
③
④
⑤
⑥
⑦
⑧
⑨
⑩
⑪
⑫

Figure 2

METHOD OF GRAVITY DISTORTION AND TIME DISPLACEMENT

FIELD OF THE INVENTION

[0001] The present invention relates to the use of technical time displacement devices, which operate by the modification of gravitational fields. These drive systems do not depend on the emission of matter to create thrust to take advantage of time dilation, but rather create a change in the curvature of space-time, in accordance with general relativity. This allows travel across topologies by warping space-time, to produce a topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967)

THEORETICAL BACKGROUND OF THE INVENTION

[0002] The concept of gravity should be examined in the light of quantum gravity and in turn as a component of quantum physics itself. The fundamental minimal quantum of energy in quantum physics is Planck's constant; h. Thus in accordance with the energy equivalence formula $E=mc^2$, the fundamental minimum quantity of mass (m_q) can therefore be derived, from known constants by; $m_q=h/c^2$ (1). Taking this minimal mass, it is possible to show that the formation of all matter, the forces of nature and indeed space-time itself derive from this single quintessential quantity.

[0003] Thus if the number of quintessences in a system is; $n_q=m/m_q$: then the total Energy of the system is more logically given by, the energy of a single quintessence (h); directly multiplied by the number of quintessences (n_q) in that system, thus

$$E=hn_q=mc^2 \qquad (1\ a).$$

[0004] Furthermore, this minimal mass, termed quintessence, can form the basis of the existence of a quantum gravitational field in the form of a space-time lattice, from which quantum gravity may be derived from first principles. Furthermore, the conglomeration of these quintessences also accounts for the formation of the elementary particles and the forces acting between them, as in superstring theory. This concept explains the formation of matter and the forces of nature on a quantum mechanical basis and directly explains the existence of wave particle duality. Thus as $n_q=m/m_q$; the frequency of light and matter (f) is determined, directly, from the number of constituent quintessences. This leads automatically to the fundamental equation, derived from (1), $f=n_q=E/h$, where n_q is the number of quintessences, which leads directly to the frequency of both light and matter. This in turn leads directly to a Universal wave equation for matter and light $\lambda=c/\beta n_q=hc/\beta E$ (2), where β is the relative directional velocity, v/c. As the momentum, $p=\beta\times E/c$, then this equation also gives the standard de-Broglie wave equation, $\lambda=h/p$ in agreement with current theory and experiments[1].

[0005] Using the Universal wave equation, the standard equation for special relativity, $m'=m_o/(1-\beta^2)^{1/2}$, derives from first principles. Also from these observations, a modified Dirac wave equation may be derived, $E\psi=(-j\beta\cdot\nabla+\beta m)\psi$ (2a), the results of which have been recently verified by a paper in which the orbitals of electrons were experimentally directly visualised[2]. Moreover, a fundamental equation for general relativity can be formulated, where G is the gravitational constant and r_q is the given radius of quintessence; $G=9(r_q)^2c^4/\lambda\beta E$ (3), such that the Universal wave equation is in direct agreement with general relativity[3]. Thus special and general relativity and quantum mechanics can be unified.

[0006] From here it is possible to proceed in a number of ways; the geometric structure of the electron and the forces of Nature may be derived from first principles and in turn the structure of the quarks, including the top and bottom, otherwise known as truth and beauty can be seen. Moreover, the presence of a space-time lattice results in an understanding of quantum EPR effects. By allowing a theoretical flow of energy through the space-time lattice it can be shown that:

[0007] Energy is not bound by space-time

[0008] Thus logically accounting for phenomena such as entanglement and quantum tunnelling. Quintessence can also be used to explain, logically, the inner physics of a black hole, the missing mass of the Galaxy, the continuing expansion of the Universe, Guth's inflationary theory and the Big Bang. Hence, it is now possible to understand the Universe, including space-time, matter and the forces of nature from the radius, mass and vibration of a single quantity, quintessence.

[0009] With this understanding of space-time, matter and the forces of Nature, and in particular gravity, it is possible to demonstrate that the modification of gravitational fields, and in turn the warping of space-time, can be technically readily achieved.

[0010] Using standard equations for special relativity, $m'=m_0/(1-\beta^2)^{1/2}$, it can be demonstrated that by differentially increasing the velocity of electrons, by applying a differential current, their mass can be increased in a specific way. In turn by increasing the mass of electrons, by general relativity, the number of gravitons emitted from these electrons can be modulated. By multiplying this effect using an ultracentrifugational device the differential graviton emission can be manifestly amplified. This in turn, in accordance with general relativity, will cause a change in the curvature of space-time.

[0011] This effective warping of space-time does not, of necessity, imply superluminal velocities, but does allow the creation of warp drive systems, which do not depend on the creation of thrust by the ejection of material as used in current space technologies.

Part 1—Fundamental Laws of Physics

[0012] Quintessential Mass

[0013] The quantum physical, minimum component of energy is Planck's constant; h. To define the minimal component of mass, using the standard energy equivalence formula; $E=mc^2$, such a minimal mass (m_q) would be required to have the value equivalent to; $m_q=h/c^2$ (1). The total mass of a system (m) would then be; $m=n_qm_q$, where (n_q) is the number of these minimal units. Thence, the total energy of a system can be derived from the minimal energy; h, multiplied by the number of these energy units (n_q). Thus as, $E=mc^2$, then also $E=m_qn_qc^2$ and substituting $m_q=h/c^2$, the energy equivalence formula has the more logical formulation; $E=hn_q$(1a). Thus the energy of a system is equivalent to the minimal energy unit; h, multiplied by the number of those minimal energy units (n_q).

[0014] Taking this minimal mass/energy, it is possible to show that all matter, the forces of nature and space time can be constructed from this single quintessential quantity. Moreover, using this quantity the laws of physics can be derived from first principles. Thus, a priori, all components of the physical universe, including space-time, can be constructed from this minimal mass component, termed quintessence.

[0015] Wave Particle Duality

[0016] If the presence of quintessence accounts for the structure of matter and if matter itself forms from the number of quintessences, then the frequency of matter and thus wave particle duality directly arises from first principles. Specifically the wavelength of matter derives from the vibration of quintessence from which it is constituted. Thus the frequency (f) and in turn the wavelength of light and matter is directly equivalent to the number of quintessences contained within it. We find that the actual frequency of light can be directly derived from first principles from the effective mass of the photon (m,) and thus by the number of quintessences (n_q) it contains.

[0017] Thus for light conventionally:

$$f=E/h$$

[0018] and if $E=mc^2$, and $h=m_q c^2$, then

$$f=m_\gamma c^2/m_q c^2$$

and

$$f=m_\gamma/m_q=n_q$$

Thus

$$f=n_q \tag{4}$$

[0019] Thus the formula for the frequency of light E=hf is now readily explained by the observation that the frequency is determined quite directly from the number of quintessences n_q within the photon.

[0020] The wavelength is thus also given by:

$$\lambda=c/f=m_q c/m_\gamma=h/p$$

[0021] We can now show that the frequency of matter also has the same derivation from quintessence, as has the frequency of light. The frequency of matter is again equivalent to the number of quintessences it contains. Thus the wave particle duality of matter itself can be explained by its composition from quintessence. The amount of quintessences contained within a electron sphere will depend on the number of quintessences constituting the electron and those passing through it as a result of its relative velocity β^2 (where $\beta=v/c$); effectively its relativistic momentum (p). The frequency will then be related to the total number of quintessences. Thus for matter,

$$f=\beta^2 n_q \tag{4a}$$

[0022] Thus it is possible to derive the conventional de Broglie wave equation for matter from first principles. Thus, as $\lambda=v/f$, we have:

$$\lambda=v/\beta^2 n_q \tag{5}$$

[0023] thus as $n_q=E/h$

$$\lambda=hc/\beta E \tag{2}$$

[0024] and as conventionally $\beta E/c=p$, then for matter:

$$\lambda=h/p$$

[0025] Provided that in the de Broglie equation, the momentum of the object is calculated using the relativistic mass, thus accounting for the total number of quintessences n_q in an object, this gives an accurate value for the wavelength of matter[1].

[0026] Thus the wavelength of matter follows directly from its constituents, quintessence. As matter is made of quintessence, similarly to light, its frequency depends on the number of quintessences n_q within it, traveling relative to the speed of light. Moreover, $\lambda=hc/\beta E$, underpins a fundamental relationship between wavelength and energy. Furthermore, this is mathematically the same as the term $\lambda=hv/\beta^2 E$, giving a relativistic expression for the wavelength of matter, from which the relativistic equations may be directly derived

[0027] Wave Equations

[0028] The derivation of wave particle duality from first principles also now allows the derivation of a modified wave equation for matter.

[0029] To derive his wave equation Shrodinger commenced with the de Broglie equation using momentum (p). For lower energies the momentum of an electron is conventionally derived from the kinetic energy of the electron and the mass of the electron m_0. Thus conventionally:

$$E_k=\frac{1}{2}mv^{1/2} \quad \text{and} \quad p=m_0 v$$

Thus

$$E_k=p^2/2m_0$$

then

$$p=\sqrt{(E_k\cdot 2m_0)}$$

[0030] and conventionally, the de Broglie equation can also be written as:

$$\lambda=h/p=h/\sqrt{(E_k\cdot 2m_0)}$$

[0031] In turn the Shrodinger wave equation directly derives from the square of the above classical non relativistic term for kinetic energy:

$$\lambda^2=h^2/E_k\cdot 2m_0$$

$$\text{thus } E_k=\frac{h^2}{2m}\cdot\frac{1}{\lambda^2}$$

$$\text{As } E=E_k+V$$

$$\text{then } E\psi=-\frac{h^2}{2m}\cdot\frac{d^2\psi}{dx^2}+V\psi=jh\cdot\frac{d\psi}{dt}$$

[0032] However, the Shrodinger equation, may be refined by taking into account relativity. Thus the true values for the energy are given by the relativistic momentum (p).

[0033] A fundamental relativistic wave equation for ψ, and its logical derivation may now be developed through the concept of quintessence as a fundamental constituent of matter.

[0034] The amount of quintessences in the electron is determined by the number of quintessences forming the electron at rest, plus the amount of quintessences passing through it due to its relativistic velocity, which will determine the relativistic momentum (p) of a particle.

[0035] The frequency of matter can now be readily calculated from first principles to give a more accurate result. Thus as matter is made of quintessence, similarly to light, its frequency is equal to the number of quintessences n_q within it. The wavelength will depend on its velocity travelling relative to the speed of light and thus multiplied by the relative velocity compared to c(β=v/c);

[0036] Hence for matter as previously shown:

$$\beta = v/\beta^2 n_q = hc/\beta E \tag{2}$$

[0037] And conventionally

$$E = (p^2c^2 + m_0^2c^4)^{1/2}$$

[0038] Using these equations, we can now, also, reformulate the Shrodinger wave equation, which has the advantage that relativity can be treated in a quantum mechanical way. Thus if the wave energy of matter is defined as:

$$E_\lambda = \frac{\beta\sqrt{p^2c^2 + m_0^2c^4}}{c^2}$$

thus

$$E_\lambda = \sqrt{(\beta^2p^2/c^2) + \beta^2m_0^2}$$

[0039] which in complex space generalises to

$$E_\psi = (-j\beta\cdot\nabla + \beta m)_\psi \tag{2a}$$

[0040] As the term

$$\alpha = \frac{e^2}{hc}4\pi\varepsilon_0;$$

also represents the ground state ratio of the velocity of the electron to c. Thus α=β=v/c=1/137.

[0041] Thus, also

$$E_\psi = (-j\alpha\cdot\nabla + \beta m)_\psi$$

[0042] This is thus the standard relativistic equation that Dirac was able to construct from the Shrodinger wave equation. This relativistic equation can be derived from the modified wave equation. This takes into account the relative mass energy which the quintessential wave equation contains.

[0043] Where importantly the term β·m is the mass m, multiplied by the ratio of the relative velocity to light β=v/c, and the term a is also essentially the relative velocity of the electron.

[0044] The Dirac equation was an empirical formula which worked mathematically, nevertheless even Dirac admitted it was not logically understood. The importance of these equations is that they show that the existence of quintessence allows the waveparticle duality of matter to be explained and mathematically derived from first principles, Thus the frequency of matter or even light is simply determined by the number of quintessences it contains.

[0045] Indeed, a recent publication in Nature has suggested that the direct visualisation of the orbitals of electrons shows that these are in very close agreement with theory. However, there is a significant departure from theory, in the interstitial molecular regions, suggesting that the higher velocities of the electrons obey the modified Dirac equation. Thus these orbitals were in keeping with the modified Dirac equation, which itself may be derived from the wave equation above, λ=hc/βE

[0046] The Shrodinger wave equation will approximate to the correct values until v approaches c. Indeed the Shrodinger equation will give similar answers as that derived from equation, under most experimental conditions.

[0047] However, equation 2 and its derivative may have advantages over standard Shrodinger theory with relativistic speeds. Furthermore, equation 2, conceptually shows that the wave particle duality of matter derives from the principle that the frequency of matter is directly equal to the number of quintessences it contains. Importantly it also mathematically allows relativity and quantum mechanics to be united.

[0048] With v≅c, the modified Dirac equation will yield more accurate results, particularly compared with the Schrodinger equation. We also find that the equation λ=hc/βE is equivalent to the de Broglie wave equation, λ=h/p, provided we use the relativistic mass in the de Broglie equation. Given this, these equations yield accurate experimental results.

[0049] Thus we find that the modified formulation of de Broglie wave equation λ=hc/βE leads directly to a modified Dirac relativistic wave equation and is supported by recent experiments which measure the wavelength of matter and demonstrate the electron orbitals experimentally from these wave equations for matter.

[0050] Wave Particle Duality and Relativity

[0051] From here it is possible to proceed in several ways using the relativistic wave equation. It is apparent that the reintroduction of the term for relative velocity into the wave equations will enable the reintroduction of special relativity into quantum mechanics. In particular we should now be able to derive the term

$$\sqrt{\left(1 - \frac{v^2}{c^2}\right)}$$

as a special case of quantum mechanics.

[0052] Thus if:

$$\lambda = hc/\beta E$$

As $E = \sqrt{p^2c^2 + m_0^2c^4}$, squaring

$$\lambda^2 = \frac{h^2c^2}{\beta^2\cdot(p^2c^2 + m_0^2c^4)}$$

-continued

Conventionally $p^2c^2 = \frac{E^2v^2}{c^2}$

then

$$\lambda^2 = \frac{h^2c^2}{\beta^2 \cdot ((E^2v^2/c^2) + m_0^2c^4)}$$

Thus as $\beta^2 = \frac{v^2}{c^2}$ and $m_0^2c^4 = E_0^2$, then:

$$\beta^4E^2 + \beta^2E_0^2 = \frac{h^2c^2}{\lambda^2}$$

hence $\beta^4E^2 = h^2c^2 \cdot \frac{1}{\lambda^2} - \beta^2m_0^2c^4$ thus

$$\beta^2 = \frac{h^2c^2}{\beta^2E^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $E^2 = m^2c^4$

$$\beta^2 = \frac{h^2c^2}{\beta^2m^2c^4} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

Substituting $h = m_qc^2$

$$\beta^2 = \frac{m_q^2c^6}{\beta^2m^2c^4} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $m_q/'m = 1/n_q$ (eq. 2)

$$\beta^2 = \frac{c^2}{\beta^2n_q^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4}{\beta^2E^4}$$

Thus if $f = \beta^2n_q$; (eq. 7a)

$$\beta^2 = \frac{v^2}{f^2} \cdot \frac{1}{\lambda^2} - \frac{\beta^2m_0^2c^4f^2}{\beta^2E^2}$$

As $1/\lambda^2 = f^2/v^2$

$$\beta^2 = \frac{v^2}{f^2} \cdot \frac{f^2}{v^2} - \frac{\beta^2m_0^2c^4f^2}{\beta^2E^2}$$

Thus

$$\beta^2 = 1 - \frac{\beta^2m_0^2c^4}{\beta^2E^2}$$

As $E^2 = m^2c^4$

$$\beta^2 = 1 - \frac{m_0^2}{'m^2}$$

Hence

$$m_0/'m = (1-\beta^2)^{1/2}$$

Thus

$$'m = m_0 \Big/ \left[1 - \frac{v^2}{c^2}\right]^{1/2}$$

[0053] Thus this derivation now allows relativity as a universal case of the quintessential wave nature of matter.

[0054] The original premises on which special relativity was based were: that the speed of light is a constant and that all observers are equal. As the speed of light has dimensions of length and time but not apparently of mass, the relativistic change in mass is not accounted for. Using quintessence logically and directly accounts for the relativistic mass changes.

[0055] Moreover, relativity can be derived from the de Broglie equation, and visa versa, directly, thus linking relativity and quantum mechanics by taking into account the existence of quintessence mass.

[0056] Hence, it is now possible to derive the relativistic equations for mass and in turn for space and time from the quintessential wave equation, thus deriving special relativity as a universal case of quantum mechanics and thus uniting special relativity and quantum mechanics. This now allows a further understanding of the nature of space-time.

[0057] The Space-Time Lattice

[0058] The understanding of the true nature of space-time and how it is formulated in three dimensions of real space is crucial. To simply assume that space-time exists, and thence not to question the nature of that existence, denies a deeper understanding of the universe.

[0059] In order to understand the nature of space-time itself, at the quantum level a further look at the nature light and the photon is necessary. Since Einstein's description of light as a particle (the photon) and the description of the photoelectric effect, the standard picture of light as simply a wave can, no longer be applied. If light was to exist as a photon, it could not exist in one dimension, as ordinary waves do, it would need to be three dimensional, with the addition of time. Let us suppose, in this case, that a photon is a three dimensional helical ringlet of light, travelling in the x vector, and spinning around the x-axis. Conventionally this ringlet has a radius; $r=\lambda/2\pi$. The ringlet itself would be vibrating in the y and z vectors. The vectors x, y and z would represent the photon, the substance of which, would be travelling in the x direction and oscillating in the y and z vectors, which would represent oscillatory energy. This in turn would allow it to act as a wave, and create oscillatory electromagnetic fields.

[0060] It is important to re-examine space-time itself in this light, this would have one directional vector with two vector dimensions of energy, one of capacitance and one of electrical permeability, thus accounting for the well known constants of free space; the permittivity of free space (ϵ_0) and the permeability of free space (μ_0) respectively. The vector dimension of direction x, would be the direction of travel and those "quintessences" travelling in an outwardly direction would account for none other than the expansion of the universe. Three of these quintessences would naturally constitute three dimensional visible space-time. These constituents of space-time would interact with the generations of the other vector dimensions reciprocally. Thus one quintessence would sweep out one vector of permeability and one vector of permittivity, through which the other two quintessences could travel, and vica versa, creating a three dimensional space-time lattice.

[0061] The permittivity of free space, (ϵ_0) which is equivalent to capacitance, would as with capacitance plates, be determined by the effective separation between quintessences. The permeability of free space (μ_0) is in fact a force, measured as $4\pi\times10^{-7}$ N/A^2, would result from the force produced by the vibration of quintessence and would be dependent on the density of quintessence. Hence these two parameters would be reciprocal and thus the product of these two would therefore be a constant, which is recognised as none other than the speed of light.

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

[0062] This space time lattice would in effect be created by quintessences travelling in all directions with a speed of c within the lattice. The quintessences of the space time lattice would in effect produce a non-static ether. A non-static ether is fully compatible with special and general relativity. Indeed such an ether explains how space time can be curved as in general relativity. Furthermore, the existence of a non-static ether, was espoused by Einstein in his University of Leyden lecture on general relativity of May 5, 1920. In Einstein's own words;

[0063] "According to the general theory of relativity space without ether is unthinkable."

[0064] Recent evidence from a number of sources now strongly support the presence of this non-static ether, in the form of quintessence. An editorial from a major journal states "combined with other observations such as those of distant Supernova, the QMAP results corroborate the prevailing theory of inflation with the twist that the Universe is only one third matter (both ordinary and dark) and two thirds quintessence, a form of energy possibly inherent in empty space".

[0065] If we take into account the existence of quintessence and as such a three dimensional space-time lattice, matter which is intrinsically made of constituents of charge would interact with this lattice to produce the effects of mass. Mass would be perceived as a result of matter (whose constituent particles appear to contain charge) interacting with this lattice directly due to the inhibition of motion by the lattice's electrical permeability and permittivity vectors, which would form the existence of complex space. These quintessences would in the direction in the y and z vectors produce small vibrations of the order of the Planck length (10^{-35} m), whilst passing through the vectors of permeability and permittivity, thus producing the effects of mass.

[0066] The vibration would endow quintessence itself a (non rest) mass m_q equivalent, to the minimal mass of:

$$m_q = h/c^2 = 7 \times 373 \times 10^{-51}\ \text{kg}\cdot\text{sec} \tag{1}$$

[0067] The presence and magnitude of Planck's constant (h) and especially the speed of light (c) is thus explained. Indeed, the speed of light

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$$

is not in itself a fundamental quantity.

[0068] As the energy equivalence formula is $E=mc^2$, the minimal mass of a single quintessence, would thus be the minimal mass, h/c^2, hence again:

$$m_q = h(\mu_0 \epsilon_0) = h/c^2 = 7.373 \times 10^{-51}\ \text{kg}\cdot\text{sec} \tag{1b}$$

or

$$m_q c^2 = h \tag{1c}$$

[0069] It is postulated by general relativity that the shape of space time itself can be altered, indeed the presence of the space time lattice now allows this to be altered by altering the density of quintessence. It is further clear that if quintessences underly the structure of the space-time lattice, they may also underly the structure of matter itself.

[0070] With regards a single quintessence, this passing through an energy vector of the space-time lattice would appear as a vibrating string. In a similar way to string theory, the conglomeration of these quintessences would produce the constituents of ordinary matter. Thus the general equation for the number of quintessences (n_q) in an object of mass (m) would be

$$m/m_q = n_q$$

[0071] The mass of the electron (m_e) for example, would be directly determined by the number of quintessences in the electron, multiplied by the mass of quintessence.

[0072] Quintessence and Complex Space

[0073] Quintessence is postulated to constitute the fundamental nature of space-time. Three quintessences each travelling in their respective x vectors at 90° to each other would create three dimensional real space-time. These quintessences would in the direction in their respective y and z vectors produce small vibrations of the order of the Planck length (10^{-35} m), this would create the vector dimensions of permeability and permittivity. The result would give space-time 9 dimensions of space as in superstring theory. However, unlike superstring theory the six hidden dimensions would not be "curled up so as to be so small as to be invisible" these six dimensions would be present in complex space. Thus, only three of these dimensions would represent ordinary three dimensional particulate space time i.e. three dimensional objects. The other six dimensions produced by the vibrations of quintessence would form complex space.

[0074] The mathematics of complex space, using imaginary $\sqrt{-1}$ or (j) numbers, is assumed in the standard formulation of the Shrodinger wave equation. Thus the presence of complex space is an integral part of quantum mechanics.

$$-\frac{h^2}{2m} \cdot \frac{d^2\psi}{dx^2} + V\psi = jh \cdot \frac{d\psi}{dt}$$

[0075] The mathematics of complex space is also an essential and integral part of the principles and application of modern electronic and control engineering. Indeed it has been well recognised for some time that each direction vector in electronic engineering can, be associated with complex vectors.

[0076] As this complex space consists of the vectors of permittivity and permeability it would only be "felt" by charged particles as in the electron. Nevertheless, as all particles are fundamentally composed of charged particles the effects of complex space would be felt by endowing these particles with mass and in turn kinetic energy.

[0077] In conventional complex space, a 2 dimensional Cartesian Argand diagram is mathematically used. However, in order to formulate the equations for particles a three dimensional Argand diagram is essential. This will have three dimensional vectors, one real vector and two imaginary vectors. Three of these diagrams will be required to fully describe the nature of particles, each with a real vector in the x, y and z vectors, respectively. Nevertheless, in the instance below the real vector is the x vector and the two imaginary vectors are given by $({}_{j}y,{}_{j}z)$

[0078] The Three Dimensional Argand Diagram

[0079] The beauty of a three dimensional Argand diagram is that the complex conjugate (i.e. the mirror image which confers mathematical reality on the coordinates) is formed by the value of the minus coordinate in the other complex vector dimension. Thus the complex conjugate of $(C_{x}^{1/2}+{}_{j}c_{y}^{1/2}+{}_{-j}c_{z}^{1/2})$ is $(c_{x}^{1/2}+{}_{j}c_{y}^{1/2}+{}_{j}c_{z}^{1/2})$. These two sums when multiplied thus give a real number solution.

[0080] Furthermore it is clear that nine dimensions of space time are necessary in the general relativistic equations. By including complex space we thereby create the nine dimensional spacial metrictensor and the metric energy tensor of matter necessary for computations for general relativity From here we can begin to understand the true structure of matter.

[0081] Energy and the Space-Time Lattice

[0082] The presence of numerous experimental data for quantum tunnelling. and indeed the recent observations by Nicholas Gisin, on the entanglement of distant photons now returns us to EPR experiments.

[0083] Using the quintessential modification of the de Broglie wave equation, gives us an insight into these teleportation and EPR effects.

[0084] As

$$\lambda = hc/\beta E \qquad (2)$$

and

$$E = hn_{q} \qquad (1a)$$

then

$$\lambda = c/\beta n_{q} \qquad (2b)$$

[0085] Importantly, as indicated by equation (2b), energy having no quintessence; would have a wavelength of infinity. Specifically pure energy containing no quintessences, would have a lambda of infinity. According to quantum mechanics an infinite wavelength would result in the probability of that energy being anywhere. As energy itself has no electrical charge it would not be impeded by the permittivity and permeability of the three dimensional space-time lattice. Moreover, energy would not be detectable in three dimensional space-time, unless it interacted with matter, as in the EPR experiments. Indeed, energy is not observed when not bound to any form of mass or particle. Thus equation 9d, takes us to our original assertion

[0086] Energy is Not Bound by the Space-Time Lattice

[0087] Thus, as the EPR experiments suggest the existence of energy separate from matter and thus separate from the three dimensional space-time lattice, it is interesting to find that experiment suggests the existence of free energy in a continuum separate from space time to produce the effects of quantum teleportation

[0088] This is not, however, teleportation across an additional dimension, this is a term to describe in partially familiar terms the dissociation of energy from the three dimensional space-time lattice. As time is inextricably linked to each dimension of space, the effects of energy would be inextricably linked to the events, such as the creation of virtual particles, we see interacting within space-time.

[0089] It is unlikely that observers have any direct day to day experience to explain quantum events. Nevertheless, quantum physics may have given us a window into the hitherto hidden workings of the Universe. Thereby, the mystery of the uniformity of the Universe, across distances which the speed of light could not apparently traverse, is readily explained by the fact that the free energy contained in the Universe is not bound by the space-time lattice.

[0090] In the case of light, due to the exceedingly small masses involved, there would be relatively easy exchange of matter with free energy within a photon. This would make the photon the ideal experimental tool to look for energy which is not bound by matter and in turn energy which is not bound in space-time. Indeed, very recently Furusawa et at. have reported to have observed the transference of energy as photons from A to B, without those photons traversing space-time. This finding which has been supported using other experimental techniques, is very important as it suggests the existence of such a quantum continuum.

[0091] We have already seen strong experimental data using photons, atomic spins and other data for quantum teleportation which have recently been published which support these findings. According to the above equations the teleportation would vary in a predictable fashion, as with photons, in line with the wavelength of the light used, relative to the size of vibration of quintessence. As regards matter, the results do confirm that the effect of quantum tunnelling is indeed dependant on the wavelength of matter and the size of that matter.

Part II—Particle Physics

[0092] Electron Structure

[0093] Understanding the electron is fundamental to the understanding of the elementary particles. The hidden nature of the electron may recently have been revealed through observations by Horst Stormer, Daniel Tsui and Robert Laughlin for which a Nobel prize has recently been awarded. They describe a quasi electron particle of charge ⅓e. This has been described on a quantum basis as a vortex of energy, bound as a quasi particle in one dimension x, but not bound in the other two dimensions y and z, allowing dispersion in space-time as a vortex. What is more intriguing are the experimental conditions in which this occurs. First of all a two dimensional electron gas is created and held between two capacitance plates. A magnetic force is then applied in the remaining dimension, virtually creating a one dimensional passage through which only a quasi electron appears to be able to pass.

[0094] Given the presence of charge of ⅓e, then three of these quasi electrons could form an entire electron in three dimensional visible space time. Nevertheless, each would have energy and hence a wave function which would be present in the other vectors. This electron could thus follow the probability functions as described by the Shrodinger wave equation for ψ (otherwise termed as "essence" by Shrodinger)

[0095] If the mass of the electron (m_e) is constituted from quintessence, using the formula:

$$m_e/m_q=n_q$$

[0096] Then an electron would be constituted from:

$$\frac{9.11\times10^{-3}\ \text{kg}}{7.373\times10^{-51}\ \text{kg}\cdot\text{sec}}=1.235\times10^{20}\ \text{quintessences/sec.}$$

[0097] Thus taking into account the mass-energy content of quintessence (m_q) it is independently possible to derive the magnitude of the charge of an electron (e) using the following equation.

$$e=\sqrt{\frac{m_q\varepsilon_0}{\left(\frac{4}{3}\right)\pi hc}}=1.61\times10^{-19}C$$

[0098] This is in close agreement with the experimentally observed charge on the electron of 1.602×10^{-19} C.

[0099] Interestingly substituting $m_q=h/3c^2$ in the above equation we have:

$$e=\sqrt{\frac{\varepsilon_0}{3\left(\frac{4}{3}\pi c^3\right)}} \tag{6}$$

[0100] This can also be written as

$$e=\sqrt{\frac{\varepsilon_0}{3\left(\frac{4}{3}\right)\pi c^3}} \tag{6a}$$

[0101] Equation (6) has a number of very special implications, if re-examined, firstly three of these quasi electron spheres appear to be required to constitute the charge of the electron. More intriguingly, it indicates that the charge is related to the volume of a sphere with an apparent radius of c. Thirdly it indicates that the square of the charge of an electron (e) is proportional to the permittivity of free space (so). The charge given from equation (3) is in close agreement with the measured charge of the electron. Furthermore a more exact value for the charge of the electron (to seven decimal places) can be deduced by taking into account the gravitational field of the Earth (see Gravity and the Charge of the Electron). Furthermore the charge of the electron (e) can now be derived from first principles. Thus, equation (3) corroborates the evidence that the electron is indeed composed of three quasi electrons in keeping with recent experimental findings.

[0102] The significance of the electron, composed of three spheres each with a radius of c, is not immediately clear, but can be understood if the frequency of rotation of the electron is taken into account. Thus if the diameter of the electron was approximately 10^{-19} m, then its spin would need to be

$$\frac{1}{c}\times10^{-15}$$

m approx. eq. 10^6 cycles/sec. Thus given a very high rotation rate an electron could have an effective radius of 1/c and still occupy subatomic sizes. Indeed these observations might be used to estimate the rate of rotation of the quasi electron and its size (see Appendix 1).

[0103] With regards a single quintessence, this passing through an energy vector of the space-time lattice would appear as a vibrating string. In a similar way to string theory, the conglomeration of these quintessences would produce the constituents of ordinary matter. The electron, for example, would be constituted from approximately 1.235×10^{20} quintessences.

[0104] The dimensions of the equation for the electron can be readily resolved by considering each of the three vector dimensions. The exact dimensions of the equation need to be considered in the light of the nature of space-time itself. These dimensional equations help explain the nature of matter. Indeed the equation for the electron may be necessary for the full understanding of gravity

[0105] Complex Space and Electron Structure

[0106] The presence of complex space also now further explains the conformation of the electron, and its formulation at the quantum level, and the presence of particles, anti-particles and their spin up and spin down characteristics.

[0107] Indeed the short form equation for the charge of the electron (−e) can now be rewritten as a metric tensor with three dimensions in real space and six in complex space.

[0108] Thus if three of the x, y and z vectors are in real space and six vectors in complex space, where c is the speed of light in the real space vector, $_{j}c$ is the speed of light in the complex vector and $_{-j}c$ is the complex conjugate of $_{j}c$, thus the electron can be mathematically represented by the equation:

$$-e=\varepsilon_{qe}\Big/\left(\frac{4}{3}\pi\right)^{1/2}\cdot\begin{array}{ccc}(c_x)^{1/2}\cdot & (jc_y)^{1/2}\cdot & (-jc_z)^{1/2}\\ + & + & +\\ (-jc_x)^{1/2}\cdot & (c_y)^{1/2}\cdot & (jc_z)^{1/2}\\ + & + & +\\ (jc_x)^{1/2}\cdot & (jc_y)^{1/2}\cdot & (c_z)^{1/2}\end{array}$$

[0109] Which now elegantly gives the real number solution

$$e = \frac{\varepsilon_{qe}}{3\sqrt{\left(\frac{4}{3}\right)\pi c^3}}$$

[0110] Where e=ϵ_{qe} is given as the permittivity of free space for a single quasi electron Equation 4 represents a "complex" tensor

[0111] Whilst the two dimensional Argand diagram has four quadrants, the three dimensional Argand diagram has eight cubic sectors. Two of these cubic sectors are diametric opposites and can represent "real" particulate objects. These have the primary coordinates x, y, −z; as in the electron described above, and the −x, −y, z, with the real vector x now having a minus sign. These two "real" cubic sectors, therefore, mathematically represent particles and their anti-particles.

[0112] The mathematical presence of the two primary diagonal mirror images (x, y, −z and −x, −y, z) now allow the introduction of the concept of antiparticles. This extension of the maths into a three dimensional Argand diagram thus results in the automatic formulation of the maths of antiparticles. Thus the charge of the positron ($^+$e) is formulated by the shortened form equation, where the real vectors now each have the minus sign, and therefore exist in the −x, −y, z sector of the three dimensional Argand diagram.

$$e = \varepsilon_{qe} \Big/ \left(\frac{4}{3}\pi\right)^{1/2} \cdot \begin{matrix} -c_x^{1/2} \cdot_{-j} c_y^{1/2} \cdot_j c_z^{1/2} \\ {}_j c_x^{1/2} \cdot -c^{1/2y} \cdot_{-j} c_z^{1/2} \\ {}_{-j} c_x^{1/2} \cdot jc_y^{1/2} \cdot -c_z^{1/2} \end{matrix}$$

[0113] The three dimensional Argand diagram also accounts for chirality and indeed the up and down spin of the electron. There are two other "real" primary coordinates in the Argand diagram, these are themselves the partial mirror images of the above coordinates (i.e. x, −y, z and −x, y, −z). In particular the y axis is of the opposite sign, thus in particles the y axis is in the downward direction, to form down spin particles and in anti-particles in the up direction, to form the antiparticle The up spin electron is given by eq. 8 and hence the down spin electron (−e|↓≦) is given by the equation

$$-el\downarrow\rangle = \varepsilon_{qe} \Big/ \left(\frac{4}{3}\pi\right)^{1/2} \cdot \begin{matrix} c_x^{1/2} \cdot_{-j} c_y^{1/2} \cdot_j c_z^{1/2} \\ {}_j c_x^{1/2} \cdot c^{1/2y} \cdot_{-j} c_z^{1/2} \\ {}_{-j} c_x^{1/2} \cdot jc_y^{1/2} \cdot -c_z^{1/2} \end{matrix} \tag{9}$$

[0114] Thus the three dimensional Argand diagram accounts directly for the presence of antiparticles and the spin up and spin down particles seen in nature. It also accounts for the necessity of the electron to form a square root spherical object, as complex space depends on √−1, otherwise known as j.

[0115] Electron Pairing and Superconductivity

[0116] As the quintessences making up the electron are in a square root conformation, each of these quasi electrons would have a tendancy to pair to form an entire sphere.

[0117] The square root sphere structure of electrons with up and down spins can now superimpose to produce a complete sphere of varying extents. This produces electron pairing as seen at the atomic and molecular levels. It also accounts for the Pauli exclusion principle. This pairing thus accounts for the reactivity of the valence electrons and the electron probability densities, which in turn accounts for the existence of chemistry.

[0118] Furthermore, it is possible to account directly for superconductivity from first principles. For if both the complex and real vectors of the electron combine completely, the product of an up and down spin electron form a perfect superimposed sphere with radius c, with a charge of 2.59×10^{-38} C, denoted by the formula:

$$e^2 = \frac{\varepsilon o}{3\left(\frac{4}{3}\pi c^3\right)} = 2.59\times10^{-38}C \tag{6b}$$

[0119] As with standard superconducting theory, superconductivity can be explained by the formation of "Cooper" electron pairs, where the electrons are forced to pair by the presence of positive crystal charge in particular formation, at supercooled temperatures. In addition the electron pair now forms a stable entity whose angular momentum cancels.

[0120] It additionally becomes clear that the charge of two separate electrons (2e) is 3.2×10^{-19} C, but the charge of the combined electrons (e^2) is 2.59×10^{-38} C. This electron pair thus appears to have 19 orders of magnitude less charge than the electron and in turn 19 orders of magnitude less resistance. It is this effective reduction in charge and in turn resistance, which may account for superconductivity. When observed directly any electrical interaction with the Cooper electron pair will, however, result in the release of the full charge of both electrons, so that the full electrical charge put in will be equal to that coming out of the apparatus.

[0121] The Fine Structure Constant

[0122] Intruiginty from our knowledge of the electron we can further define the term .alpha., the fine structure constant; from the structure of the electron. Thus as the standard term

$$\alpha = \frac{e^2}{hc\times 47\pi\varepsilon_0};$$

substituting the term

$$e^2 = \frac{\varepsilon_0}{3(4/3)\pi c^3} \tag{eq. 6}$$

and

$$h = m_q c^2 \tag{eq. 1}$$

we find:

$$\frac{2\pi}{\alpha} = m_q[3(4/3\pi c^3)]^2$$

*or

$$\frac{2\pi}{\alpha} = \frac{m_q e^4}{\varepsilon_0^2}$$

[0123] For brevity we may represent the quasi electron structure as (4/3)$\pi c^3 = \theta$; to signify its threefold symmetry, thus

$$\frac{2\pi}{\alpha} = m_q(3\theta)^2 \tag{10}$$

[0124] Indicating that the fine structure constant of the electron (a) is indeed related to its dimensional structure. Again taking into account the effects of gravity the fine structure constant can be derived from first principles to nine decimal places (see Gravity and the Charge of the Electron).

[0125] Fundamental Forces and Particle Structure

[0126] In order to understand the fundamental forces and the nature of fundamental particles, an overview is required. Thus, there are three major forces; strong, electro-weak and gravity, each mediated by three force particles the gluon, photon and graviton respectively. These in turn, influence three types of particle, the quark, lepton, and by general relativity space-time itself. Each of these are composed of particles with multiples of charge of ⅓, which are themselves in three generations, and are present in three dimensions of real space. It is important that a comprehensive view of nature explains this threefold symmetry.

[0127] Using the Standard Model of particles, it is well accepted there exist quark particle charges of −⅓, −⅔ and +⅓ and +⅔ in quarks and anti-quarks. Given that each particle is made up of three quarks the presence of these fractional charges support the association of the fractional charges in this way to form three dimensional charged particles. In stable particles each of the three quarks would have a vector in one dimension, giving the three quarks together an existence in three dimensional visible space time. The particles that bind the quarks (gluons) are themselves required, in stable particles, to have three different color charges, one color in each dimension, for the particle to exist in three dimensional space-time. Furthermore, there are three generations of quarks (and indeed leptons).

[0128] The Standard Model (or a modification of this) and in particular the observation of quarks and indeed quasi electrons with fractional charge of ⅓ and ⅔ in both cases, indicates that particles are constituted from the equivalent of three of these quasi particles to form an electron and quarks to form baryons. In the normal three dimensions the energy would be carried by the particle, However, because each particle is constituted of three quasi particles and in each quasi particle or quark one visible dimension would be the direction vector, in the other two hidden dimensions of each vector the waves would carry energy. Thus each particle would be associated with vibration, which would account for wave particle duality and Heisenberg's uncertainty principle in three dimensional visible space-time.

[0129] These observations lead us directly to the previous postulate that the structure of the electron is composed of none other than three (root) spheres, and that this equation for the electron allows the determination of the charge of the electron from first principles, thus:

$$e = \frac{\varepsilon 0^2}{[3(4/3\pi c^3)]} \tag{6}$$

[0130] In addition the mass of the proton (m_p) can be directly calculated from the ratio of the mass (m_e) of the electron, given by the equation:

$$\frac{m_e}{m_p} = 5.45 \times 10^{-4} = 3\frac{\pi}{\sqrt{c}} \tag{11}$$

[0131] Strictly we should write,

$$\frac{m_e}{m_p + m_e} = 3\pi\sqrt{c};$$

which is much more elegant.

[0132] Which now gives

$$\frac{m_e}{m_p} = \frac{1}{\sqrt{c}/3\pi - 1} = 5.4462 \times 10^{-3}$$

[0133] This is in very close agreement with the experimentally derived ratio of the proton to electron masses which is also; 5.4462×10^{-3}

[0134] Thus the correlation factor between theory and experiment has a maximum error <0.00001.

[0135] If we combine equation 3:

$$e = \sqrt{\frac{\varepsilon_0}{3(4/3)\pi c^3}}$$

and equation 13:

$$\frac{m_e}{m_p} = 3\frac{\pi}{\sqrt{c}}$$

the positive charge of the proton (e_p) is given by:

$$e_p = \sqrt{\frac{\varepsilon_0}{3(4/3)\pi c^3}} \times m_e \times 3\frac{\pi/\sqrt{c}}{m_p} = e \tag{12}$$

[0136] The stable nuclear proton conformation can thus be represented by the short form equation:

$$p=3^{+}\sqrt{3(4/3)\pi c^3}\times 3(\pi/\sqrt{c}) \tag{13}$$

[0137] This forms a stable 3×3 conformation as with the stable electron structure.

[0138] Importantly the term ($\pi/\sqrt{c}$) is the 90° solution to the Shrodinger wave equation for an electron confined in a space with radius c!.

[0139] Thus the standard equation for an electron confined in a one dimensional box is given by:

$$E\psi(x) = -\frac{h^2}{2m}\frac{d^2\psi(x)}{dx^2}$$

[0140] If the one dimensional box has a length 2L the quantum amplitude (A) can only be non zero between x=0 and X=2L and the standard solution for the amplitude is none other than:

$$A=(1/L^{1/2})$$

[0141] Thus in one dimension the standard solution to the Shrodinger wave equation is:

$$\psi(x)=(\pi/L^{1/2})\text{Sin } x/L$$

[0142] Thus not only is the electron charge derived from the equation for three spheres each with a radius of c (eq. 3); but the proton mass and charge can also be derived from the standard solution to the Shrodinger wave equation for a an electron confined in a space of radius c!.

[0143] The term ($\pi/c^{1/2}$) itself would thus most logically represent the gluon which is present in the proton. These gluons would bind the quasi electrons together to form the fundamental particles

[0144] The masses of all the known particles, including the up and down quarks, the W boson, the muon, charm, strange, the tauon, truth and beauty can thus also be derived from first principles in this fashion, and have the quasi electron as their basic constituent particle (see Appendix 1).

[0145] Thus the structure of the muon (μ) can also be derived from the ratio of the mass of the electron (m_e) and the mass of the muon (m_μ):

$$m_e/m_\mu=4.7\times 10^{-3}=c^{1/3}$$

Thus

$$\mu=\epsilon_0{}^{1/2}\times m_e/m_\mu\times 3^{-}(4/3=\pi c^3)^{1/2}\times(\pi/c^{1/3})$$

[0146] Where the charge of the muon is in this equation equivalent to that of the electron e. In this case ($\pi/c^{1/3}$) can be considered to represent a specific high energy photon. Thus the structure of the muon, written in short form is:

$$\mu=3^{-}(4/3\pi c^3)^{1/2}\times(\pi/c^{1/3}). \tag{14}$$

[0147] Moreover the structure of the tauon can be calculated from the ratio of the mass of the electron and that of the Tauon (1.79 Mev);

[0148] Thus

$$0.511\text{ Mev}/1.79\text{ Gev}=2.85\times 10^{-4}$$

$$m_e/m_r=(\pi/c)^{1/3}\times(\pi/c)^{1/9}=2.85\times 10^{-4}$$

[0149] As the charge of the tauon is equivalent to the charge of the electron, hence the structure of the Tauon is given by the above equation

$$e_\tau=\epsilon_{qe}\times(m_e/m_\tau)\times 3^{-}(4/3\pi c^3)^{1/2}\times(\pi/c)^{1/3}\times(\pi/c)^{1/9}=e$$

[0150] This equation accurately predicts the charge −1; and mass of the Tauon (.~1.78 Gev). Thus the structure of the Tauon can in short form be given by the equation

$$\tau=-3(4/3\pi c^3)^{1/2}\times(\pi/c)^{1/3}\times(\pi/c)^{1/9} \tag{15}$$

[0151] Furthermore a more exact value for the mass of the muon and tauon can be deduced by taking into account the gravitational field of the Earth in a similar way to identifying the exact charge of the electron. In addition it may be necessary to take into account a possible mass value of the neutrino to arrive at a precisely accurate mass value of the muon and tauon. Nevertheless, the mathematical proof of these short form equations lies in the fact that they can very closely identify the charge and the masses of these particles, from first principles, as in equations.

[0152] Overall the mathematical geometrical structure of all the particles can be derived from the quasi electron, which is in turn derived from quintessence. Thus, the short form particle structures can now be derived from first principles. This includes the quasi electron (qe) and electron (e), from which the quarks (u,d) and in turn the stable proton (p) and stable neutron (n) and alpha particle (a) respectively are derived. The general structure of the force carrying bosons the photon. (g) and the gluon (γ) and the intermediate vector boson (W) can be given. It will also intriguingly be possible to derive, according to their generation, the structure of the strange (s) charm (c), beauty (b, or bottom) and truth (t or top) quarks directly from the structure of the muon (μ) and Tauon (τ) respectively.

[0153] Using the term $\Theta=(4/3\pi c^3)$, where, −/+ represents the charge of the quasi electron, we find:

[0154] 1st Generation:

$$(q_e)={}^{-}\Theta^{1/2} \tag{6c}$$

$$e=3^{-1/2} \tag{6}$$

$$d={}^{-}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{16}$$

$$u=2^{+}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{17}$$

$$s={}^{-}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{18}$$

[0155] 2nd Generation

$$\mu=3^{-}\Theta^{1/2}\cdot(\pi/c^{1/3}) \tag{14}$$

$$c=2^{+}\Theta^{1/2}\cdot(\pi/c^{1/3})\cdot(\pi/c^{1/4}) \tag{19}$$

$$b={}^{-}\Theta^{1/2}\cdot(\pi/c^{1/9})\cdot(\pi/c)^{1/4} \tag{20}$$

[0156] 3rd Generation

$$\tau=3^{-}\Theta^{1/2}\cdot(\pi/c)^{1/3}\cdot(\pi/c)^{1/9} \tag{15}$$

$$t=2^{+}\Theta^{1/2}\cdot(\pi/c)^{1/3}\cdot(\pi/c)^{1/9}\cdot(\pi/c)^{1/4} \tag{21}$$

[0157] Particle Gluons (g):

$$g_1=(\pi/c^{1/2}) \tag{22}$$

$$g_2=(\pi/c)^{1/2} \tag{22a}$$

$$g_3=(\pi/c^{1/4}) \tag{22b}$$

$$g_4=(\pi/c)^{1/4} \tag{22c}$$

[0158] Particle Photons (.gamma.):

$$\gamma_1=(\pi/c^{1/3}) \tag{23}$$

$$\gamma_2=(\pi/c)^{1/3} \tag{23a}$$

$$\gamma_3=(\pi/c^{1/9}) \tag{23b}$$

$$\gamma_4=(\pi/c)^{1/9} \tag{23c}$$

[0159] Intermediate Vector Boson ($W^{+/-}$):

$$W^{+}=3^{+}\Theta^{1/2}\cdot 2(\pi/c^{1/6})^6 \tag{24}$$

$$W^{-}=3^{-}\Theta^{1/2}\cdot 2(\pi/c^{1/6})^6 \tag{25}$$

[0160] Stable* Proton:

$$p=3^{+}\Theta^{1/2}\cdot 3(\pi/c^{1/2}) \tag{13}$$

[0161] Stable* Neutron:

Stable * Neutron:

$$n = \begin{array}{l} {}^{+}\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot^{-}\Theta\uparrow^{1/2} \\ {}^{+}\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot^{-}\Theta\uparrow^{1/2} \\ {}^{+}\Theta\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot^{-}\Theta\uparrow^{1/2} \end{array}$$

* Stable nucleonic neutron and proton conformations differ slightly from the Standard Model, this is due to the sharing of quasi electron and quasi positron particles within the nucleus, which allows stabalisation of these particles by the formation of stable 3×3 structures. The Standard conformations which describe non-nucleonic neutrons and protons are additionally given in Appendix 1.

[0162] Alpha particle (α):

$$\text{Alpha particle }(\alpha):=\alpha=\begin{array}{l} [3^{+}\Theta]\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^{-}\Theta]\uparrow^{1/2} \\ [3^{+}\Theta]\uparrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^{-}\Theta]\downarrow^{1/2} \\ [3^{+}\Theta]\downarrow^{1/2}\cdot 2(3\pi/c^{1/2})\cdot[3^{-}\Theta]\uparrow^{1/2} \end{array} \tag{27}$$

[0163] The mathematical proof for these structures and their decay mechanisms is lengthy and is thus fully contained in Appendix 1. All the particle structures are accurately mathematically defined by the masses of these particles.

[0164] The structure of these particles all contain the quasi electron and thus the metric tensor structure necessary in the formulation of the gravitational equations is sustained. The respective forces created by the gluon and the photon are important as they tell us the behaviour of matter and also lead to the likely structure of the graviton

[0165] Particle Spin and Size

[0166] The significance of the electron, composed of three spheres each with a radius of 1/c, is not immediately clear, but can be understood if the frequency of rotation of the electron is also taken into account. Knowing the structure of the electron has led us to deduce its charge and thus may lead us estimate its size and spin. Thus these observations might be used to calculate the radius and rate of rotation of the electron.

[0167] Let us suppose, that nature is truly beautiful, and that the radius of the fundamental quasi electron is indeed 1/c, and in turn the radius was balanced by the velocity of rotation 2π/c. This can be directly confirmed mathematically by taking into account the known spin of the electron, h/4π. Thus the actual spin of the electron may be calculated form the known energy of the spin.

[0168] The radius of the electron is not up till now known, but the radius of a quark has been estimated, and this is the radius derived from deep inelastic collisions of the proton. These estimates reveal a radius of approx. $r_p=1.18\times10^{-15}$ m This value may be used to assist in confirming the spin of the proton in revolutions per sec. (revs) and in turn the spin and size of the electron. Firstly we may proceed to estimate the spin of the proton. Thus as h=E·t (Joules×sec) and h=E·t=F·d·t (Joules×sec), then the spin;

$$h/4\pi=F\cdot d\cdot t \tag{28}$$

[0169] As F=ma, where $\alpha=(\text{revs}\cdot 2\pi)^2 r_p$ and m=the mass of the proton, then

$$h/4\pi=m(revs\cdot 2\pi)^2 r_p\cdot d\cdot t$$

[0170] The actual distance (d) traveled in a circle of half integer spin in 1 second is: $\text{revs}\cdot\pi r_p$ thus:

$$h/4=m(revs\cdot 2\pi)^3 r_p^{\ 2}/2$$

[0171] Hence:

$$revs=[h/m(2\pi)^4 r_p^{\ 2}]^{1/3}$$

[0172] Taking the effective mass the proton as 1.6726×10^{-27} kg, then the rate of spin of the proton in revolutions/sec is:

$$revs=5.65\times10^{6}\text{ cycles/sec}$$

[0173] From the frequency of the specific rotation of the proton, given the half integer spin associated with the proton, we can thus mathematically confirm the relationship between the radius of a particle and its spin:

$$r_p\times revs/2=1/c \tag{29}$$

[0174] Furthermore, the fundamental radius of 1/c seen in geometric structure the quasi electron, is also reflected in the rotation rate and radius for the proton, thus as above 1/c÷½ revs=1.85×10^{-15} m. Moreover, this means the actual half integer velocity of rotation is none other than 2π/c in metres/sec. So that the particle is in harmonic balance.

[0175] Using the fundamental formula h/4π=F·d·t, it is possible to obtain accurate estimates of the radius and spin rates of the electron, or indeed any particle, using the same principle of harmonic balance. Using the formula:

$$revs=[h/m(2\pi)^4 r_e^{\ 2}]^{1/3}$$

[0176] It appears there are two unknowns, the radius if the electron and its revolution rate, however, in accordance with the equation, r_p=2/c.revs, which gives the revolution rate of the proton, the same principle may also be used for the electron, by substituting r_e=2/c.revs, such that:

$$revs=hc^2/4m_e(2\pi)^4 \tag{30}$$

[0177] Taking the mass of the electron 9.109382×10^{-31} kg, the rate of revolution of the electron is:

$$revs=1.048\times10^{10}\text{ cycles/sec}$$

[0178] Which gives a predicted radius of the electron as

$r_e=6.336\times10^{-19}$ cycles/sec

[0179] So the half integer rotation velocity (revs.πr) is 2π/c !, for the electron in keeping with the harmonic balance of the electron.

[0180] The same principle may be used to obtain an accurate estimate of the spin and radius of the muon, or any other particle. Using the above formula

$$revs=hc^2/4m_e(2\pi)^4 \quad (31)$$

[0181] Then as the mass of the muon is 1.8823×10^{-28} kg then the revs of the muon $f_\mu=5.070\times10^{-7}$ cycles/sec and the radius r_μ is thus 1.316×10^{-16} m.

[0182] It is now possible to begin to explain how the muon and the other subatomic particles are formed. If a quasi electron is complexed with another structure the total geometric structure needs to maintain harmonic balance. So the frequency of rotation would need to match geometric structure with which the quasi electron was complexed

[0183] Intriguingly we find asymptotic convergence for the formulas for frequency and mass occurs, when the geometric structure complexed with the quasi electron has the structure represented by ($\pi/c^{1/3}$) [giving the frequency divided by two, because the single integer spin of the force carrying particles compares to a half integer spin for the muon]. So that

$$(f_r 3\pi/2)^{1/3}=f_\mu$$

[0184] When the ratio of the masses of the electron (m_e) and muon (m_μ) are related, such that:

$$m_e(3\pi/c^{1/3})=m_\mu$$

[0185] Indeed we find that (allowing for the neutrino) this ratio is very close to the actual ratio of the mass of the electron to the mass of the muon, determined experimentally.

[0186] Furthermore, we have seen that these geometric structures, representing harmonics of the speed of light, which either match the frequency or the amplitude of vibration of the quasi electron, mathematically define the masses of the particles and the fundamental forces of Nature.

Part III—Quantum Gravity

[0187] Quantum General Relativity

[0188] Given the overall energy "complex" energy tensor structure of the electron and the metric tensor, assumed in general relativity, the quantum nature of gravity itself can now be explored. The spherical complex tensor for the electron and the positron give the mathematical quantum structure and energy tensor for all the other particles. Together with the time dimension these nine space dimensions account for the 10 parameters present in the metric tensor necessary to formulate the equations for gravity using Riemann geometry and thus forms the basis of quantum gravity. Intriguingly the metric tensor at each point in space time is required to consist of a collection of ten numbers, Consequently, ten dimensional space-time hypotheses, such as this or superstring theory, do automatically yield general relativity.

[0189] Furthermore, the mathematical representation of the graviton and the gravitational constant may be directly estimated from the knowledge of the mass and radius of quintessence. Thence, the force of the vibrations of quintessence lead directly to quantum gravity.

[0190] The radius of quintessence should be approximately in keeping with the Planck length estimate (r), which is conventionally derived from the standard dimensional equation:

$$r_q^2\cong Gh/c^3 \quad (32)$$

[0191] Given the nine spacial parameters present in the metric tensor, used in general relativity we find that the actual formula for r_q^2 is mathematically in agreement with theory when:

$$9r_q^2=Gh/c^3 \quad (33)$$

[0192] This again supports the 9 dimensional view of space and the size of the vibrations of quintessence can thus be estimated.

$$r_q=1.35\times10^{-35}\text{ m} \quad (33a)$$

[0193] This value is in agreement with the Planck length. Indeed if the above equation is correct then we find that we can derive the standard equation for the general relativistic increase in radius, r', (eq. 34) directly from first principles and arrive at a more fundamental equation for quantum gravity. As

$$r'=G\cdot M/3c^3 \quad (34)$$

[0194] By substituting eq. 33) into equation 34, a fundamental relationship between r' and M is obtained.

$$r'/3r_q^2=GMc^3/Ghc^3=Mc/h$$

[0195] And substituting the quintessential equation, $h=m^qc^2$ (eq. 1) then:

$$r'/3r_q^2=M/m_qc=n_q/c \quad (35)$$

[0196] Hence the ratio of the change in radius to that of the radius of quintessence squared, is proportional, by a factor of c, to the ratio of the mass M of an object to that of the mass of quintessence, effectively the number of quintessences. Thus the change in radius, r' due to gravitation, is related to none other than the ratio of the mass and radius of an object to the mass and the square of the radius of quintessence. Thus again the gravitational change in radius is directly related to the number of quintessences.

[0197] Naturally, this would be exactly what would be logically expected if quintessence, like the equation for the charge of the electron (eq. 6) forms from a root sphere. Thus the change in spacial radius of a normal sphere is dependant on the square of the quintessential radius.

[0198] This increase in apparent radius represents none other than the (gravitational) binding energy for quintessence.

[0199] The meaning of the above dimensional equation (33) might itself be further understood by substituting the mass of quintessence (where $m_q=h/c^2$) into the equation. Thus in nine dimensions the gravitational constant (G) may be more logically given as,

$$9(\pi r_q^2/m_q)=G\pi/c \quad (36)$$

[0200] Where πr_q^2 is the cross sectional area of quintessence and m_q is the effective mass of quintessence, and thus

($\pi r_q^2/m_q$) represents the effective mass per unit area which quintessence exerts. This equation reduces to:

$$9r_q^2/m_q=G/c \tag{37}$$

[0201] From this we may derive the standard general relativistic relationship for the apparent change in radius (r') around a mass (M), from an understanding of the mass m_q and number (n_q) of quintessences. As $m_q=M/n_q$, then:

$$3r_q^2=G\cdot M/3c\cdot n_q \tag{38}$$

[0202] Then if

$$n_q=r'c/3r_q^2 \tag{39}$$

[0203] thus directly substituting for n_q in eq. 38:

$$r'=G\cdot M/3c^2 \tag{34}$$

[0204] The importance of this is that the gravitational change in radius now logically derives from equation 36, which describes the gravitational force as resulting directly from the mass of quintessence exerted/per unit area of quintessence.

$$9(\pi r_q^2/m_q)=G\cdot\pi/c \tag{36}$$

[0205] Thus equation 34 is the conventional equation for the general relativistic increase in radius (r') in a gravitational field, which is here derived from the underlying nature of quintessence. Thus the gravitational constant is derived from the mass and radius of vibration squared of quintessence from first principles.

[0206] Indeed it is apparent that a more fundamental equation for gravitation now exists, for equation (39) is mathematically accurate and numerically agrees with eq. 34:

$$r'/3r_q^2=n_q/c \tag{39}$$

[0207] These equations may be readily mathematically verified. If in accordance with standard general relativity, the apparent increase in radius r' is:

$$r'=GM/3c^2 \tag{34}$$

[0208] Then given that the mass of the Earth is 5.9745×10^{24} kg;

$$r'=1.478\times10^{-3}\text{ m}$$

[0209] Accordingly if $r'=3r_q^2\cdot n_q/c$; (eq. 39). Given the number of quintessences n_q constituting the Earth is M_E/m_q, then

$$n_q=5.9745\times10^{24}/7.3725\times10^{-5}=8.104\times10^{74}$$

[0210] As $r_q^2=1.823\times10^{-70}$ (eq. 33a) then:

$$r'=1.478\times10^{-3}\text{ m}$$

[0211] Thus equation 39 gives the same answer as the standard equation and may be understood on a logical basis. Indeed the meaning of c in the equation may be understood as it has been previously shown as being the basis for the radius of matter (eq. 6). Hence the general relativistic change in radius, r', is none other than the effective binding energy for quintessence.

[0212] Quantum Gravity and Wave Particle Duality

[0213] Quantum gravity can now be readily linked with quantum mechanics, indeed any observations which are self consistent must be able to do so easily.

[0214] The frequency of light has been previously derived

$$f=E/h=n_q$$

[0215] Thus the formula for the frequency of light (E=hf) has previously been explained theoretically by the simple observation that the frequency is determined quite directly from the number of quintessences (n_q) within the photon. The same principle has also been shown to apply to matter.

[0216] Let us now follow these equations for matter by calculating the wavelength of a photon from the Gravitational constant as an example; and also as a test of these observations and to demonstrate that the gravitational equations can also apply to the quantum world.

[0217] If

$$n_q=r'c/3r_q^2 \tag{39a}$$

[0218] where r' is the general relativistic increase in radius, and r_q is the radius of quintessence (eq. 33). Where $f=E/h=n_q$, substituting for n_q, then the frequency of the photon f_γ(where $\beta=1$) is given by:

$$f_\gamma=r'c/3r_q^2$$

[0219] Using the standard equation, $r'=GM/3c^2$ (eq. 34); we may substitute for r', thus we have:

$$f_\gamma = GM/9r_q^2c\ \text{15 Thus}$$

$$f_\gamma = \frac{G}{9r_q^2c^3}\cdot m_\gamma c^2 \text{ and as } E = m_\gamma c^2;$$

$$f_\gamma = \frac{GE}{9r_q^2c^3} \tag{40}$$

Indeed as $9r_q^2 = Gh/c^3$, then $f_\gamma = E/h = n_q$

[0220] It is possible to also demonstrate that the same relationship holds for the wave equation for matter. If we take the relativistic wave energy of matter, which has been previously derived,

$$f=\beta^2 n_q$$

[0221] This includes the term for the number of quintessences flowing through the electron, in the complex vectors of space-time, to give the relativistic electron momentum (p) and a term for the rest mass, thus substituting into (40)

$$f_\gamma = \frac{GE}{9r_q^2c^3} \tag{40}$$

[0222] As $f=\beta^2 n_q$ for matter then the equation expands to:

$$f_m = \frac{G}{9r_q^2c^3}\cdot\beta^2E \text{ As } \lambda = v/f, \text{ then}$$

$$\lambda = \frac{9r_q^2c^3v}{G\beta^2E} \tag{41}$$

[0223] Then the equation again reduces to:

$$G = \frac{9r_q^2c^4}{\lambda\beta E} \tag{3}$$

[0224] Equations 3, 40 and 41 are important as they show that the quantum wavelength of any particle of rest mass m can be derived from the gravitational constant G. Thus linking quantum mechanics to quantum gravity.

[0225] It is therefore important to confirm the numerical accuracy of the above equation (40). We can do this by comparing the result to the standard computation of the de Broglie equation, in a range where de Broglie itself is likely to be most accurate; which according to these observations is in the low energy range (see section on Wave Particle Duality).

[0226] If we take an electron with an energy of 0.1 KeV the wavelength is conventionally given (where the kinetic energy of the electron E_k is given by the product of the charge of the electron (C) and the potential applied eV=0.1 Kev), by the standard equation:

$$\lambda = h/p = h/(E_k \cdot 2m_0)^{1/2} \text{ thus}$$

$$\lambda = 6.63\times10^{-34}/[1.602\times10^{-19}\times1\times10^2\times18.22\times10^{-31}]^{1/2}$$

$$\text{hence}$$

$$\lambda \cong 1.23\times10^{-10}m \text{ Using}$$

$$\lambda = \frac{9r_q^2c^4}{G\beta E} \text{ Where } E = \gamma m_0c^2 \tag{3a}$$

[0227] At 0.1 Kev, electron velocity is 6×10^6 r/sec, thus $\beta=2\times10^{-2}$ and $\gamma=1/(1-v^2/c^2)^{1/2}=1.0002$. Thus:

$$\gamma = \frac{9\times1.82\times10^{-70}\times80.78\times10^{32}}{6.76\times10^{-11}\times\beta\times1.0002\times9.11\times10^{-31}\times8.998\times10^{16}}$$

$$\lambda = 1.21\times10^{-10}m$$

[0228] Divergence between the de Broglie equation and the above equation (2) occurs at intermediate and high energies where it is generally accepted that the standard de Broglie equation may be less accurate. The values for eq. 2 and de Broglie are compared to recent experiments, which demonstrate a relativistic curvilinear plot for wavelengths of matter in keeping with eq. 40.

[0229] The de Broglie equation in the non-relativistic format yields a simple log/linear scale, which is not in keeping with relativity; whereas eq. 3 is dependent on relativity and mathematically accounts for both relativity in calculating the wavelength. Indeed recent experiment on quantum tunnelling through a wire mesh strongly suggests that the relationship between energy and wavelength is relativistically curvilinear(ref 1). Furthermore equation 3a suggests a fundamental relationship between energy (E), relative velocity (v/c=.β), gravity (G) and the quantum wavelength (λ)

$$\lambda = \frac{9r_q^2c^4}{G\beta E} \text{ Indeed as } {}^9r_q^2 = Gh/c^3\text{, then} \tag{3a}$$

$$\lambda = hc/\beta E \tag{2}$$

[0230] Equation 2 is the very same as the Universal wave equation derived form first principles for the wavelength of tight and matter, which allowed a relativistic solution to the equations for wave particle duality (see Wave Particle Duality). This now indicates that these quintessential equations are compatible with relativity, quantum mechanics and quantum gravity.

[0231] Graviton Structure

[0232] From these observations, if the value for the gravitational constant is substituted into the equation (35) we may now estimate the probable geometric structure of the gravitation, which is the force particle mediating gravity by acting on quintessence. Thus the Gravitational constant has been previously derived from the vibration of quintessence by the equation:

$$G\cdot(\pi/c)=9(\pi r_q{}^2/m_q) \tag{36}$$

[0233] This is in accurate agreement with the value for G(6.67×10^{-11} N m^2 kg^{-2}). This suggests that the most probable mathematical representation of the graviton (ϕ), the third force carrying particle is

$$\phi=(\pi/c) \tag{42}$$

[0234] Thus the gravitational constant (G) can be given by the mass and radius of quintessence and the structure of the graviton

$$G=9\pi r_q{}^2/\phi m_q \tag{43}$$

[0235] This shows the gravitational force to be related to the fundamental radius of quintessence space time, and the graviton.

[0236] Quantised General Relativity

[0237] The classical general relativistic formula, as given by Einstein is:

$$R_{\mu\nu}-1/2g_{\mu\nu}R=-\kappa T_{\mu\nu}$$

[0238] Where R is effectively the curvature of space-time, $R_{\mu\nu}$ denotes the contracted Riemann tensor of curvature and $T_{\mu\nu}$ is the “energy tensor” of matter.)

[0239] If we substitute the energy tensor matrix of the electron (eq. 9)×time, for the energy tensor of matter $T_{\mu\nu}$; and the metric tensor of the space-time lattice×time for the contracted Riemann tensor we can arrive at the same solutions for general relativity.

[0240] Furthermore, in his published paper on General Relativity, Einstein. defined the constant κ as:

$$\kappa=8\pi G/c^2$$

[0241] Therefore Einstein's equation should be written as

$$R_{\mu\nu} - 1/2g_{\mu\nu}R = \frac{8\pi G}{c^2}\cdot T_{\mu\nu} \tag{43}$$

[0242] Einstein himself was apparently not happy about the right hand component of the equation. However, we find that this part of the equation can now be explained and quantised by substituting the gravitational constant, G=9πr$_q$^{2}/ϕm$_q$, (eq. 42a),

[0243] Giving:

$$R_{\mu\nu} - 1/2g_{\mu\nu}R = -\frac{8\pi^2 9r_q^2}{\varphi m_q c^2} \cdot T_{\mu\nu} \tag{44}$$

[0244] By substituting m$_q$·c^2=h, and further substituting h=h/2π, we arrive at a quantised solution to Einstein's equations. Where A$_q$ is the surface area of quintessence (A$_q$=4πr$_q$^{2}); ϕ is the graviton [ϕ=(π/c)] and h is Plancks constant. thus:

$$R_{\mu\nu} - 1/2g_{\mu\nu}R = -\frac{9A_q}{\varphi\hbar} \cdot T_{\mu\nu} \tag{45}$$

[0245] The gravitational equation can now be further understood on a logical basis. The term A$_q$=(4πr$_q$^{2}), where represents standard term for the surface area of a sphere of quintessence for the 9 space dimensions of the space time lattice, h is the energy content of quintessence×time and ϕ is the graviton, thus the right hand term now represents a true "metric energy tensor" of matter.

[0246] This leads directly to the standard solution to the field equations, for the general relativistic increase in radius r' of an object, where A is the surface area of a sphere of a given mass M, such that

$$r = \sqrt{(A/4\pi)} - r = GM/3c^2 \tag{34}$$

[0247] Furthermore, although equation 45, gives the same solutions as Einstein's equation, which is essentially correct, the difference is that the equation is now dependant upon Planck's constant (h), and moreover the radius of quintessence, which now defines a quantised solution to the equations.

[0248] Graviton Force Characteristics

[0249] Similar to the photon, the previously derived equation (42) for the graviton [ϕ=(π/c)] appears to also mathematically represent a helical ringlet of quintessence, but with a spin of 2. For the photon, taking the direction of motion as the x vector and its axis of spin also as the x vector, would account for the electromagnetic force and its attraction and repulsion characteristics. In the case of the gluon component (π/c$^{1/2}$), if the direction vector is x, then the axis of spin would be in the y vector, the same as quasi electrons, accounting for the particle binding characteristics of the gluon force. In the case of the graviton, if the direction of motion was in the x vector, the graviton spin axis would be in the z vector thus, as will be demonstrated, accounting for the gravitational force.

[0250] The spin axis of the graviton can also be derived using the known characteristics of the electron. If an electron is travelling in the x direction, then its spin axis is determined by the by the sign of the $_j$y vector (up or down). This view is in agreement with conventional theory, which indicates that the electron spin is similar to a rotating planet orbiting the sun, (the electron even appears to have orbital precession). As the electron passes through the space-time lattice, this spin would generate the formation of gravitons. This would occur as a result of the ejection of the excess quintessence passing through the electron. As the electron spins, the ejection of these gravitons would occur at a tangent to the electron's direction of motion. The ejection of the gravitons would occur, similarly to the ejection of energy of a pulsar or quasar, through the equivalent of the north and south poles of the electron. Thus, propelling the graviton in the direction of the electrons y vector. The ejection of the graviton would re-orientate and impart a specific angular momentum to the gravitons which would thus end up spinning on its own z axis. If for instance the graviton is released from an up spin electron the graviton will be rotating clockwise and its leading edge will displace quintessence downwards. In turn this will provide an upwards force.

[0251] This picture accounts for Fleming's left hand rule, is logical and provides an explanation for the magnetic force around a wire. According, to the left hand rule if the direction of the current is in the x vector, the magnetic field is in the z vector, and the force is upwards, in the y vector, in accordance with the above model. Therefore, this particular spin axis and the structure of the graviton results in its force characteristics. As the graviton is very small compared to the electron and both have different rather rapid spin axis it is difficult for these to bind and interact. Nevertheless, because the graviton has a spin of 2, and as it spin axis is perpendicular to its direction of motion, in the z vector, it readily displaces space-time quintessence to produce gravity. Thus because the graviton is able to displace space-time, it is capable of escaping a black hole. How else could the effects of gravity be felt beyond a black hole?

[0252] Quantum Gravity and Electromagnetism

[0253] With the above electron model of graviton production the nature of magnetism can be understood from first principles, Furthermore, the presence of a space-time lattice links relativity, and the forces of gravity with the electromagnetic and other forces of Nature. Indeed, evidence for these links may first date back to the 1820's, when Andre Ampere first defined the Amp. The force of attraction between two parallel wires 1 metre apart each carrying 1 Amp in a vacuum was defined as none other than the permeability of free space (2×10^{-7} N per metre of conductor). Thus conventionally the magnetic field strength around a long straight wire is given as:

$$B = \mu_0 I/2\pi r$$

[0254] Where I is the current and μ_0 is the permeability of free space (4π×10^{-7} N A^{-2})

[0255] The attraction between two wires both carrying negative charge is, however, counterintuitive as negative charges should repel. A conventional explanation overcomes this by invoking the presence of a magnetic field which is created by the current by the production of virtual photons. Thus we appear to have an explanation for the effects of magnetism which involves virtual photons, however, these photons are not observed. More accurately, according to conventional special relativity the magnetic field is none other than the electric field viewed relativistically.

[0256] A more satisfactory explanation, therefore, lies in the interaction between the electrons and the space time lattice. The moving electrons in the two wires interact with the lattice to produce gravitons; which are in phase when both streams of electrons are traveling in the same direction. The gravitonic waves interact constructively to disperse the space time lattice between the wires and induce an attractive force between the two wires, which produces in effect the permeability of free space. Thus this force results from the vibration of quintessence itself.

[0257] Conversely in two wires with current going in opposite directions the graviton waves are in anti-phase and would interact destructively between the wires. The gravitonic waves traveling radially outward from the wires would, however, disperse the lattice outside the two wires and produce apparent repulsion between the wires, which is exactly what is observed. These effects of electricity suggest that gravitons act as waves and that phase is important.

[0258] This effect is also seen with the north and south poles of ferromagnets. Nevertheless, with matter other than iron, cobalt or nickel, the graviton emission cannot be phased as the atoms are unable to align and magnets do not appear to exist with other materials.

[0259] In ordinary magnetic system the release of gravitons from the north pole would be exactly balanced by those released from the south pole of the magnet and hence there would be no net force on the magnet until an external magnet or electrical current were applied.

[0260] Overall the magnitude of the forces in electrical systems where electrical conduction occurs are well defined by the permeability and permittivity of free space μ_0, and ϵ_0. Where v is the constant velocity of the charge and ϵ is the electric field produced by the charge.

$$B=[\mu_0\epsilon_0]v\cdot\epsilon$$

[0261] These observations suggest that the forces of electricity which produce magnetism are indeed related to the permittivity and permeability of free space and that these quantities are exerted by an apparent vacuum. Thus the effects of magnetism could be explained by none other than the phased effects of gravitational waves on the space time lattice.

[0262] Electromagnetism is of further interest to quantum gravity, particularly if we combine the standard equations, $B=\mu_0 I/2\pi r$ and $B=[\mu_0\epsilon_0]v\cdot\epsilon$, substituting for B we have:

$$2\pi r=I/\epsilon_0 v\epsilon \tag{46}$$

[0263] Thus $2\pi r$ is proportional to the inverse of ϵ_0. Thus as space time is dispersed by gravitons the permittivity field will increase in the same way capacitance increases with separation of plates. Because of the inverse relationship between ϵ_0 and $2\pi r$, as ϵ_0 increases the circumference of a circle and the apparent ratio of π is to r, will appear to diminish in accordance with general relativity. This not an actual diminution in the circumference of a circle but the effective reduction of the resistance to motion in a circular path in this field.

[0264] Incidentally, the above observations, also lead us directly to Schrödinger's formula for the average equilibrium distance (r) between an electron with charge (e) in orbit around a proton, which is conventionally given by:

$$r=n^2 4\pi h^2\epsilon_0/me^2$$

[0265] Where ϵ_0 is again the permittivity of free space, m is the mass of the electron and n is an orbital integer, h is Planck's constant and e is the charge of the electron. Furthermore if $e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}$ (eq. 3); then the equation at n=1, for the electron orbital radius elegantly simplifies to:

$$r=4h^2c^3/m$$

[0266] Hence the orbital radius of the electron is related to spin of the electron (h) and its mass (m).

[0267] Quantum Gravity and the Charge of the Electron

[0268] The equation for the charge of the electron (eq. 1) contains the term so (permittivity of free space) which according to these observations should vary in a gravitational field.

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2} \tag{6a}$$

[0269] If we combine the standard equations, $B=\mu_0 I/2\pi r$ and $B=[\mu_0\epsilon_0]v\cdot\epsilon$, substituting for B we have:

$$2\pi r=I/\epsilon_0 v\epsilon \tag{46}$$

[0270] Thus $2\pi r$ is proportional to the inverse of ϵ_0. Thus as space time is dispersed by gravitons the permittivity field will increase in the same way capacitance increases with separation of plates. Because of the inverse relationship between ϵ_0 and $2\pi r$, as ϵ_0 increases the circumference of a circle and the apparent ratio of π to r, will appear to diminish in accordance with general relativity.

[0271] Thus ϵ_0 rises when space-time is dispersed by the gravitons that produce the gravitational field, This occurs in a similar way to the process by which capacitance increases with separation of plates in a capacitor.

[0272] Nevertheless, as c is a constant and as $c=[\mu_0\epsilon_0]^{-1/2}$, then if ϵ_0 rises then μ_0 falls. This is entirely consistent as μ_0, which represents the force that quintessence exerts, would be reduced if the quintessence space time lattice is dispersed.

[0273] Furthermore, as $\mu_0=4\pi\times10^{-7}$ N A^{-2}; then as μ_0 falls, then the apparent ration π to r, also falls in a gravitational field. This is largely the same as stating, as does general relativity, that the apparent radius r', rises in a gravitational field. So this view is consistent with general relativity.

[0274] Nevertheless, to derive an exact value for the charge of the electron we must account for gravity in the above equation. We will take the specific example of the Earth's gravitational field in order to obtain the exact value for the electron. If in accordance with standard general relativity, the apparent increase in radius r' is:

$$r'=GM/3c^2 \tag{34}$$

[0275] Then given that the mass of the Earth is 5.9745×10^{24} kg; then

$$r'=1.47864\times10^{-3}\text{ m}$$

thus

$$2\pi r'=9.29057\times10^{-3}$$

[0276] Which is the incremental factor by which ϵ_0 must increase in Earth's gravitational field. So to correct ϵ_0 to account for gravity, ϵ_0 must be divided by the incremental factor, $2\pi r'$. Similarly as effectively π decreases in a gravitational field, to correct π to account for gravity it must be multiplied by this incremental factor. So the equation for an electron in a zero gravitational field is:

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}\div(1+2\pi r')=1.6022\times10^{-19}\text{ C} \qquad (6b)$$

[0277] This now gives the charge of the electron as measured in a zero gravitational field as 1.6022×10^{-19} C, which is the same as that measured on Earth. Notably these observations appear to suggest that the charge of the electron is the same irrespective of the gravitational field.

[0278] Virtually unlimited degrees of accuracy for the charge of the electron and for the fine structure constant (α), may be achieved by taking into account 2nd and nth order gravitometric effects. Thus if we take into account the effect of gravity upon the radius of the Earth it is also important to take into account an effect upon the instruments with which we measure quantities, this would be a second order gravitometric effect. Thus taking into account 2nd order effects (r"), we have a very small, but nevertheless relevant change, such that: r"=r'(1+2r'). Thus $2r''=9.3180486\times10^{-3}$, and thus:

$$e=[\epsilon_0/3(4/3\pi c^3)]^{1/2}\div(1+2\pi r'')=1.6021765\times10^{-19}\text{ C} \qquad (6c)$$

[0279] This agrees exactly to the nearest 7 decimal places with the maximum accuracy of the experimental value for the charge of the electron. Furthermore by taking into account the nth order gravitometric effect, it is theoretically possible to predict accuracy for the charge of the electron to 3n decimal places. This mathematically accuracy confirms the structure of the electron from first principles and indeed the theoretical effects of gravity on the permittivity of free space (ϵ_0).

[0280] This returns us directly to the fine structure constant for the electron which is conventionally given by: $\alpha=e^2/hc\cdot4\pi\epsilon_0$. If $\alpha=e^2=\epsilon_0/3(4/3\pi c^3)$, accordingly the quintessential equation for α is structurally given by: $2\pi/\alpha=m^q[3\Theta]^2$ (where $\Theta=4/3\ \pi c^3$; see The Structure of the Electron and Matter), we must now take into account the effects of gravity, as above, thus:

$$2\pi\alpha/m_q[3\Theta]^2\div(1\_2\pi r'')^2=0.007297353$$

[0281] Where the gravitational term for the increase in radius r" allows the mathematical derivation of $\alpha=0.007297353$, and the above equation is in agreement with the conventional experimental value for $\alpha=0.007297353$ to the nearest 9 decimal places.

[0282] Hence the term $(1\_2\pi r'')^2$ is in accordance with these observations for the effect of gravity on electromagnetic forces. To a maximum accuracy governed by current knowledge of the mass of the Earth and the Gravitational constant and thus the term for the gravitational increase in radius r'. These observations can also be used to accurately predict the magnetic moment of the electron

[0283] Thus the presence of the fine structure constant can now be further understood, by deriving the constant from first principles; specifically from the actual dimensional conformation for the charge of the electron: $e=[\epsilon^0/3(4/3\pi c^3)]^{1/2}$ (eq. 6).

[0284] Overall the fine structure constant α (allowing for the term r' which is the general relativistic increase in the radius of the Earth due to gravitation) is given by none other than the formula for the mass of quintessence and from the structure of the electron, which can now be derived from first principles to seven decimal places or more.

[0285] Quantum Gravity and the Electron Magnetic Moment

[0286] The theoretical origin and nature of magnetism remains obscure in current electromagnetic theory. An explanation suggests these magnetic effects are produced by photons, although no photons have ever been observed. To get round this difficulty it is postulated by physics that magnetism results from "virtual" photons. However, Maxwell's equation for electromagnetism states that the photon has no net magnetic effect.

$$\delta B_x/\delta B_y/\delta y+\delta B_z/\delta z=0$$

[0287] Thus magnetism could not, by the above standard equation, be derived from a photon real or virtual.

[0288] In addition observational data suggests that black holes have powerful magnetic fields and as in theory photons are unable to escape from black holes (except for small quantities in the form of Hawking radiation), it would be difficult to explain these magnetic fields on the basis of photon emission.

[0289] Einstein postulated that magnetism was merely due to special relativity .sup.(ref 17). The postulate for the nature of magnetism in these current observations, states that the magnetic force results from relativity due to none other than the phased emission of gravitons (why postulate two invisible forces, magnetism and gravity, when one, the graviton, will do). This view as previously discussed (Quantum Gravity and Electromagnetism) is entirely compatible with standard relativity$^{(\text{ref }19)}$. Thus with the graviton origin of magnetism, the equation for the magnetic moment of the electron should have an expression in terms of quintessence and in turn the gravitational force and in particular the graviton.

[0290] The standard term for the magnetic moment of the Bohr Magneton (SIB) is:

$$\mu B=eh/4\pi m_e$$

[0291] In standard quantum mechanics the Bohr Magneton, μB, however, needs to be corrected to agree with experiment. The "correction factor" is termed "ϵ"; where $\epsilon=(\alpha/2\pi)-0.328\alpha^2/\pi^2=0.001159641$. Thus theory reveals μ_e, the magnetic moment of the electron where:

$$\mu_e=(eh/4\pi m_e)[1+(\alpha/2\pi)-0.328\alpha^2/\pi^2]$$

[0292] The conventional derivation of the term ϵ above, is given from the fine structure constant, $(\alpha/2\pi)$ which is theoretically consistent. However, a rather arbitrary mathematical correction term; $0.328\alpha^2/\pi^2$ needs to be used in this standard equation. This appears ad hoc and needless to say, more accurate measurements show, the electron magnetic moment to the Bohr magneton ratio, $1+\epsilon=1.001159652$, which suggests the correction factor is indeed incorrect. Nevertheless, this correction factor is essential for "renormalisation" and thus for quantum mechanics to work.

[0293] Quantum gravity readily explains the discrepancy between the theoretical Bohr Magneton (μB) and the actual measured magnetic moment of the electron (μ_e). In accordance with the above chapter (Quantum Gravity and the Charge of the Electron)

[0294] Thus the significant mathematical discrepancies can be removed by accounting for the effects of quantum gravity.

[0295] Thus taking the charge of the electron (e), using the equation for the Bohr magneton and the effects of quantum gravity such that gravitational change in radius is r". The magnetic moment of the electron is given by:

$$\mu_e=(eh/4\pi m_e)(1+[\alpha/2\pi\div(1+r'')])$$

[0296] This gives an electron magnetic moment to Bohr magneton ratio of 1.00115968. Thus the mathematical term for the magnetic moment of the electron is given, avoiding the arbitrary and dubious term $0.328\alpha^2/\pi^2$ used in the standard equation, simply by accounting for quantum gravity.

[0297] It is now possible to unite the equations for gravity and magnetism by substituting the fundamental key equations of quantum gravity. Thus if: $h=3m_qc^2$ (eq. 1b) and $m_e=m_qn_q$ (eq. 2). Then we can express the magnetic moment of any particle with the charge of the electron, including the proton, in terms of the number of quintessences (n_q) in that particle.

$$\mu B=ec^2/(4/3\pi n_q); \tag{47}$$

[0298] Given that the postulated structure of the graviton is: ($\phi=\pi/c$) (eq. 42), then substituting we have

$$\mu B=3ec/4\phi n_q \tag{48}$$

[0299] Showing that the equations for the magnetic moment are compatible with the gravitational equations given earlier. Principally, the quintessential equations now allow the determination of the magnetic moment of any charged object from the equation for the graviton and directly from the number of quintessences it contains. In conventional physics the magnetic moment of the electron requires a correction factor, $(1+(\alpha/2\pi c)-0.328\alpha^2/\pi^2)$, to derive the correct experimental value. These observations herein, indicate that the correction factor is more logically (1+r"), where r" is the general relativistic increase in radius around a gravitational body. This suggests that magnetism is not only affected by gravity, but can, as shown as above, be derived using the quantum gravitational equations.

[0300] Quantum Gravity and Special Relativity

[0301] Ordinary matter passing through the lattice would produce gravitons which would interact with space-time as described by general relativity. The quantity of gravitons would be determined by the apparent mass and in turn these would apparently curve space time. The geometry of this "curvature" is elegantly described by general relativity using Riemann geometry, specifically using metric tensors. Intriguingly the metric tensor is not a single number, but at each point in space time it is required to consist of a collection of ten numbers, Consequently, ten dimensional space-time hypotheses, such as this or superstring theory, may automatically yield general relativity

[0302] General relativity is indeed very elegant, nevertheless there was a logical step yet to answer. That is, how do gravitons shape space time? This can now be readily answered by considering the interaction of a three dimensional space time lattice with gravitons themselves to produce the effects of gravity. The effects of gravity are as such to compel a body in motion towards the gravitational object and to a much smaller extent visa versa. This effect can only be produced If gravitons repel quintessence (the constituents of the 3D lattice). Indeed, it has been stated that in order to explain cosmic inflation and the "flatness" of the Universe that quintessence must shun (or be shunned by) matter.

[0303] In descriptive terms a body close to a large mass will have a tendency to move toward it because the three dimensional lattice would be less dense as it approached the surface of the large mass. Overall there would be less resistance to motion in the direction of the large mass, and the motion in this direction would be facilitated by the vibration of quintessence.

[0304] In general relativity the principle governing motion is the geodesic of least distance, this can be re-expressed using similar equations using least action. Furthermore, the concept of motion due to the vibrations of quintessence is more logically and experimentally compelling.

[0305] These observations can now be used to link general and special relativity. Thus as we approach the speed of light, the mass of an object travelling through the space-time lattice would approach infinity, directly because the number of quintessences passing through a body would increase with increasing velocity, hence the equation:

$$m'=m_0/(1-v^2/c^2)^{1/2} \tag{50}$$

or

$$m'=m_0/(1-v^{2\ [\mu}{}_0\epsilon_0])^{1/2} \tag{50a}$$

[0306] In turn this would generate increasing gravitons and accordingly this would explain the observed effects of special relativity. Time itself is due to passage through the space-time lattice, and where the space-time lattice is dispersed by gravitons, time and length are reduced with increasing velocity and hence increasing space-time lattice dispersion, similar to the way in which gravity alters space-time

[0307] As a result:

$$t'=t(1-v^2/c^2)^{1/2},\ l'=l(1-v^2/c^2)^{1/2}$$

[0308] Thus resulting in the effects of special relativity.

[0309] Quintessence and Black Holes

[0310] To address the relationship of the space-time lattice to gravity directly, it is important to discuss the concept of quintessence with regard to general relativistic equations. The standard general relativistic equation for the apparent increase in radius (r) due to the curvature of space time around a gravitational object, which has also been previously derived from first principles (eq. 36), is:

$$r'=GM/3c^2 \tag{34}$$

[0311] This can also thus be written as:

$$3r'=GM[\mu_0\epsilon_0] \tag{51}$$

[0312] This standard equation, is in keeping with the above observations. Specifically, as the mass increases, ϵ_0 increases, in turn the radius will appear to increase (relative to π).

[0313] The above observations now allow us to examine the effects with regard to the interior of black holes themselves. The event horizon would represent a critical density for quintessence, in which light could not escape. The Schwarzschild radius would now be given by:

$$R_s=2GM[\mu_0\epsilon_0]$$

[0314] The event horizon will occur at the point at which there is less resistance to circular motion than motion in a straight or partially curved line. Given that π is proportional to $1/\epsilon_0$. The event horizon should occur when the permittivity has increased by a factor of π.

[0315] Effectively because the permittivity of free space rises, π decreases. This is entirely in keeping with general relativity which predicts the effective change in the ratio of the radius to the circumference as given by the conventional equation, where r', is again the apparent change in radius.

$$r=GM/3c^2$$

[0316] Hence π will effectively decrease as we approach the event horizon of a black hole, When π decreases to 1, the circular circumference is equal to the diameter and moreover, inside this limit it is shorter for light to travel in a circle. Thus light cannot escape the event horizon.

[0317] This can give us great insights into the workings of space-time, for flat Euclidean space the standard equation is:

$$e^{i\pi}=-1$$

[0318] In accordance with general relativity, the ratio of the radius to the circumference changes in a gravitational field, and effectively π=1, at the event horizon, thus the boundary condition for the shape of space-time at the event horizon now has the direct equation:

$$e^{f}$$

[0319] Within a black hole as the permittivity of space increases by a factor of 2π an object within it will complete two rotations rather than travel in a straight line. In effect exceeding the speed of light by 2π. Hence, the condition for space-time is represented by the equation:

$$e^{i/2}$$

[0320] Thus an increase in the permittivity of free space by a minimum factor of π, to produce a black hole is estimated to result from an increase in mass by a factor of approx. 10^6 (the ratio of the mass of the earth and that of a putative black hole).

[0321] Continuing with the subject of a black hole, according to the model inside the black hole, the gravitons produced by the matter present would be in equilibrium with the density of the space-time lattice. Increasing the rate of rotation of the matter in the black hole for instance would thus increase the production of gravitons and its effective mass and increase the radius of the event horizon. A density gradient of the space-time lattice would continue to exist within the black hole. Progressively closer to the center of a black hole matter itself would be increasingly compressed and the spherical structure of the quasi electron would be predicted to collapse. This collapse would result in the formation of an exotic form of matter in the form of pure quintessence in a black hole.

[0322] This pure quintessence would produce the singularity at the centre of the black hole. The larger the black hole in terms of mass the more pure quintessence would exist at its core.

[0323] Quintessence and the Big Bang

[0324] Quintessence theory not only predicts the occurrence of the Big Bang, but allows a prediction for the value of the entire mass of the Universe, from first principles.

[0325] In accordance with quintessence theory the big bang resulted from the explosion of an immense black hole singularity, which was constituted from pure quintessence.

[0326] On the basis of quintessence, there will be a critical mass for Big Bang event; thus if entire space-time between quintessence is compacted so that no further quintessence can be accommodated, the addition of further quintessence would destabilize the immense black hole, resulting in the Big Bang.

[0327] It is possible to predict this critical mass, using the radius of quintessence as a benchmark. Given the nine spatial parameters present in the metric tensor, used in general relativity we find that the actual formula for the radius of quintessence; r_q^2, is mathematically in agreement with general relativistic theory when:

$$9r_q^2=Gh/C^3 \tag{33}$$

[0328] This again supports the 9 dimensional view of space (so crucial in superstring theory). Moreover, the size of the vibrations of quintessence can thus be calculated as:

$$r_q=1.35\times10^{-35}\ \text{m} \tag{33a}$$

[0329] The volume of each quintessence is thus:

$$4/3\pi r_q^3=1.0306\times10^{-104}\ \text{m}^3 \tag{33b}$$

[0330] So to be accommodated within unit volume of space time, with no intervening apparent space time, (given that each of 9 overlapping quintessences are required) would require approx.

$$9\times10^{104}\ \text{quintessences}$$

[0331] As the mass of quintessence is $m^q=h/c^2=7.373\times10^{-51}$ kg sec (eq. 1). Then the mass of the Universe, to two decimal places, is:

$$1.18\times10^{53}\ \text{kg} \tag{33c}$$

[0332] This is in close agreement with a recent estimate of the mass of the Universe from COBE and other satellite data, which estimates the mass to be 100 trillion trillion trillion trillion tonnes (10^{53} kg)

[0333] Moreover, the early formation of the galaxies can be readily explained, it is likely that in such a big bang some very small black holes might have prevailed and that these formed the seeds of the galaxies we see today.

[0334] The event horizon, calculated from the Schwartzschild radius, of such an immense black hole is about 10^{26} m, which would have allowed Guth's inflationary component to the early expansion of the Universe.

[0335] In addition, inflation may result directly from the observation that once electrons have formed from the primordial soup of quintessence, they emit gravitons which in turn repel space time, which might also result in another cosmic inflationary cycle.

[0336] Most importantly quintessence theory explains the Big Bang from first principles and is capable of accurately predicting the mass of the Universe.

[0337] The Nature of Energy

[0338] These observations allow a fundamental understanding of energy. The quantum physical, minimum component of energy is Planck's constant; h. To define the minimal component of mass, using the standard energy equivalence formula; $E=mc^2$, such a minimal mass (m_q)

would be required to have the value equivalent to; $m_q=h/c^2$ (1). The total mass of a system (m) would then be; $m=m_q n_q$, where (n_q) is the number of these minimal units. Thence, the total energy of a system can be derived from the minimal energy; h, multiplied by the number of these energy units (n_q). Thus as, $E=mc^2$, then also $E=m_q n_q c^2$ and substituting $m_q=h/c^2$, the energy equivalence formula has the more logical formulation;

$$E=hn_q \tag{1a}$$

[**0339**] Thus the energy of a system is equivalent to the minimal energy unit; h, multiplied by the number of those minimal energy units (n_q)

[**0340**] This leads directly to a deeper understanding of wave particle duality and the wave nature of matter.

[**0341**] This is encapsulated by the quintessential energy formulae

[**0342**] As conventionally $\beta \cdot E/c=p$, then

$$\lambda=h/p=hc/\beta E \tag{2}$$

and

$$E=hn_q \tag{1a}$$

then

$$\lambda=c/\beta n_q \tag{2b}$$

[**0343**] Importantly, as indicated by equation (2b), energy having no quintessence; would have a wavelength of infinity. Specifically pure energy containing no quintessences, would have a lambda of infinity. According to quantum mechanics an infinite wavelength would result in the probability of that energy being anywhere. As energy itself has no electrical charge it would not be impeded by the permittivity and permeability of the three dimensional space-time lattice. Moreover, energy would not be detectable in three dimensional space-time, unless it interacted with matter, as in the EPR experiments. Indeed, energy is not observed when not bound to any form of mass or particle.

[**0344**] Thus equation 2b, takes us to our original assertion regarding the existence of pure energy.

[**0345**] Energy is not Bound by the Space-Time Lattice

[**0346**] Thus, as the EPR experiments suggest the existence of energy separate from matter and thus separate from the three dimensional space-time lattice, it is interesting to find that experiment suggests the existence of free energy in a continuum separate from space time and matter to produce the effects of quantum teleportation.

[**0347**] This is not, however, teleportation across an additional dimension, this is a term to describe in partially familiar terms the dissociation of energy from the three dimensional space-time lattice. As time is inextricably linked to each dimension of space, the effects of energy would be inextricably linked to the events, such as the creation of virtual particles, we see interacting within space-time. It is unlikely that observers have any direct day to day experience to explain quantum events. Nevertheless, quintessence theory may have given us a window into the hitherto hidden workings of the Universe. Thereby, the mystery of the uniformity of the Universe, across distances which the speed of light could not apparently traverse, is readily explained by the fact that the free energy contained in the Universe is not bound by the space-time lattice.

[**0348**] In the case of light, due to the exceedingly small masses involved, there would be relatively easy exchange of matter with free energy within a photon. This would make the photon the ideal experimental tool to look for energy which is not bound by matter and in turn energy which is not bound in space-time. Indeed, very recently Furusawa et al. have reported to have observed the transference of energy as photons from A to B, without those photons traversing space-time. This finding which has been supported using other experimental techniques, is very important as it suggests the existence of such free energy.

[**0349**] Overall, quintessence theory gives an a priori explanation for the concept of mass, the elementary particles, the forces of nature and quantum effects. It can equally be used logically to explain the inner physics of a black hole, the missing mass in the Galaxy, the expansion of the Universe, Guth's inflationary theory and predicts the Big Bang, from first principles.

[**0350**] Part IV: Applied Theory with the Intent to Create Closed Timelike Curves

[**0351**] Electron Bombardment of the Photosphere to Induce Gravitational Shift

[**0352**] Leveraging the above relationship between mass and quintessence we derive that if an electromagnetic radiation with velocity v strikes the event horizon singularity of rest inertial mass mi, and U is the electromagnetic energy absorbed by the singularity, then, according to Maxwell's prediction, a momentum q=U/v is transferred to it. Mass shift $d \cdot m_b$, dependent on the external electromagnetic energy, equals the inertial mass shift dependent on the increment of energy in the particle. Since in this case the inertial mass shift does not depend on velocity V, i.e., it is related only to the momentum q absorbed, it can be obtained by making p=0 in variation $\Delta H=H'-H=c[q^2+(m_i c)^2]^{1/2}-(m_i c)^2$ from the particles inertial Hamiltonian. Consequently, the expression of d_{m_p}, is written as:

$$d_{m_p} = \Delta H / c^2 = m_i \sqrt{1 + U / m_i c^2 \left[\varepsilon_r u_{r/2} \left\{ \sqrt{1 + d / w\varepsilon^2 + 1} \right\}^2 = 1 \right.}$$

[**0353**] Comparing now the expression of m_i and m_g we have $m_g=m_i-2dm_b$. By replacing m_b in this equation, given by equation above, we obtain the expression of the correlation between gravitational mass and inertial mass, i.e.,

$$m_g = m_t - 2\sqrt{1 + U / m_t c^2 \left[\sqrt{\varepsilon_r u_{r/2} \{[d / we]^2 + 1\}^2 - 1} \right.} \cdot m_t$$

[**0354**] We see that only in the absence of electromagnetic radiation on the event horizon (U=0) is the gravitational mass equivalent to the inertial mass. Note that the electromagnetic characteristics, e, m and s do not refer to the singularity itself, but to the outside medium around the singularity (photosphere) in which the incident radiation is propagating.

[0355] Stable CTC Solution From Modified M-Theory

[0356] Our innovation is a method of creating an event where a dual membrane or dual boundary condition exists. We do this using generalization from a Misner space which has been modified if one analytically continues the maximal extended Misner Metric so that $ds^2=-du^2+dw^2+(dx^2)^2+(dx^3)^2$ to the Euclidean section so that u=iota zeta we obtain a Misner instanton on the section where w and zeta are both real. The Euclidean time, t, and the closed spacelike coordinate are both periodic, the later having a period of $2\Pi t^2$. Going back to the Lorentzian sector we find that the period of the closed coordinate becomes linearly dependent on the physical time. Using automorphic fields in the Hadamard function one can obtain a quantized condition for time. This gives us a figure on the order of the Plank time. This confines such a stable wormhole condition to the general area of the Plank scale in its modified form which is that area defined by the Membrane itself. As such, these Plank scale wormholes are the true source of the true virtual aspects of the vacuum and quintessence. The effects of the dual singularity system can be viewed as an overlap zone of two distinct space-times which have boundary conditions on both sides. As an object accelerates towards C it is this same boundary or horizon that object encounters when the Time defined horizon solution for the universe is imposed. At this point drawing upon Van Den Broeck's alterations for this space-time geometry to create a single closed Friedman-Robertson-Walker spacetime the space-time geometry can be represented by this equation $ds^2=dct^2-B^2[(dx-\beta fdct)^2]+dy^2+dz^2$ B can be any function that is large near the displacement device. We then consider this transformation as extended to four dimensional space-time with arbitrarily time dependent acceleration. We also present the device frame energy density T00 from a four dimensional calculation and note that the 4d classical calculation is everywhere finite.

[0357] Consider an Alcubierre interval given according to a remote frame's cylindrical coordinates by:

$$ds^2=(1-\beta^2f^2)dct^2+2bfdctdz-dz^2-dr^2-r^2df^2$$

where f is a function that is 1 at the location of the device and zero far from it.

[0358] Starting out with the first transformation z'=z−òct beta dct. Where b is first expressed here as a function of time ct. With some algebra for simplification this results in

$$ds^2=[1-\beta^2(1-f)^2]dct^2-2b(1-f)dctdz'-dz^2-dr^2-r^2df^2$$

Let g=1−f and this becomes

$$ds^2=[1-\beta^2g^2]dct^2-2bgdctdz'-dz'^2-dr^2-r^2df^2$$

[0359] Notice that this returned the original intervals form with a reversal on the sign of b and a reversal of the boundary conditions for g. Now we notice that at r=0, this interval becomes the interval for special relativity transformed to cylindrical coordinates. Thus, we have found a transformation to a frame based local to the device. One can also verify that in these coordinates the relevant affine connections vanish at r=0. a further proposed modification to this field we will reintroduce a time dilation term into the devices frame's interval. Only we will use different boundary conditions for it. We will keep A=1 both at the location of the device, and far from it, but allow it to become large in the warped region. This is achieved by the simplest means possible in the proposed field generation method. Since rotating kerr singularity under bombardment would produce an equal negative energy region. With the effect focused outward away from the device slightly by the relative alignment of the kerr singularity the actual inward going portion of such fields would overlap and cancel their effect out in the region of the device. This yields a space normal time region, which is again restored at the fringes of the outward going field. From the standpoint of a geometric picture of space-time around the device and extending outwards one has actually created a dual event horizon situation as far as time goes. The inner one is the shell of the canceled out field inside of which normal time flow is restored. The outer one is formed at the boundary where normal time resumes and within which we have a negative energy field.

SUMMARY OF THE INVENTION

[0360] The present invention is A method for the generation of a pseudo 2+1 dimensional anti-de Sitter space (DeDeo & Gott 2002) using two Kerr type positively charged rotating dilation singularities where one singularity is maintained as a axis of rotation or "reference" singularity, and the other "target" singularity is subjected to a differential electron flow so as to simultaneously pass above the photosphere of said singularity in its direction of rotation—prograde orbit—and contrary to its direction of rotation—retrograde orbit—to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the "target" singularity around the axis of rotation provided by the "reference" singularity, resulting in the creation of timelike curves in a compact time-oriented manifold permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) which results in a method for the formation of G odel-type geodesically complete spacetime envelopes complete with closed timelike curves.

BRIEF DESCRIPTION OF THE DRAWINGS

[0361] **FIG. 1** is a schematic representation of the mechanism employed to house the components necessary to generate a 2+1 dimensional anti-de Sitter space, resulting in the creation of timelike curves in a compact time-oriented manifold

[0362] **FIG. 2** is a schematic representation of the G odel-type geodesically complete spacetime envelope created by the mechanism complete with closed timelike curves

DETAILED DESCRIPTION OF THE INVENTION

[0363] Principles of Gravity Distortion Time Displacement Systems

[0364] The theoretical understanding of quantum gravity allows the design of time displacement systems from first principles. It is unlikely that gravitons or Kerr singularities can be controlled in a precise way using current technology. Nevertheless, an understanding of three dimensional space-time and matter, does allow the design of elementary displacement systems. That is, systems whose displacement rely on direct warping space-time as opposed to the ejection of material to provide thrust resulting in time dilatational effects.

[0365] The background for these systems are already partially understood and quintessence theory allows their further development. For this invention, the formation of black holes in the laboratory represents a crucial step in understanding the mechanisms that underlay gravitational physics and in turn the warping of space-time. The existence of black holes permits the localized application of the Axial torsion Spin-Rotation Coupling Effect (Zhang & Beesham 2002) resulting in a Rotating Frame with Relativistic Factor (Zhang 2003) which can be used in the creation of a Alcubierre space time bubble under the Van Den Broeck modification of the Alcubierre geometry (Van Den Broeck 1999) resulting in a method for the generation of a pseudo 2+1 dimensional anti-de Sitter space (DeDeo & Gott 2002) using two Kerr type positively charged rotating dilation singularities where one singularity is maintained as a axis of rotation or "reference" singularity, and the other "target" singularity is subjected to a differential electron flow so as to simultaneously pass above the photosphere of said singularity in its direction of rotation and contrary to its direction of rotation to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the "target" singularity around the axis of rotation provided by the "reference" singularity. In this instance the space-time lattice would be repelled by gravitons in such a way as to disperse space-time quintessence in a circular fashion around each of the singularities, producing multiple event horizons around the simulated axis of rotation provided by the reference singularity.

[0366] The release of gravitons from the target singularity is controlled by differentially governing the electron flow across the photosphere with the use of powerful electric currents. In turn the differential direction of flow across the photosphere of gravitons would determine the direction of motion through the space time lattice and the resulting time displacement from one spacelike boundary to the other in accordance with Geroch's theorem (Geroch 1967) when implementing this approach it is important to remember that the black hole is not excited by the smashing of the clump as it "hits" the horizon. The hole is rather excited when the metric perturbation associated with the clump is "felt" by the background metric. The excitation event therefore constitutes a smooth process whereby in-fall of a clump from .rmb and through r+ serves as a source in the Teukolsky (1973) equation for small perturbations to the Kerr geometry (with appropriate boundary conditions at r+ and r∞). This is an important distinction with a great deal of relevance to the practical engineering of the displacement unit since we need to gauge the "driving" of QNR modes in terms of an e_ective coupling from clump in-fall.

[0367] What results is a method for resonant driving of the quasi-normal ringing (QNR) wave modes of the Kerr geometry of the target singularity. The micro black hole hyper-accreting at rates ÿ M. 1 M□sec. 1 from a neutrino cooled disk is pushed through to oscillate near resonance of its (l,m=2, 2) quadrupole QNR frequency due to the in-fall of compact mass over-densities from the cusp in e_ective potential on a dynamical time scale. This mode is induced via induced magneto-rotationally induced fluid dynamics in the ultra-relativistic region of the flow bounded from below by the marginally bound orbit radius: rmb If the QNR modes are fed resonantly for a few seconds of hyper-accretion, the enhanced amplitude of the oscillations yields a very high rate of energy deposition into gravitational waves. Indeed, the integrated energy deposition is large enough to "evaporate" the equivalent of a factor of a few times the total rest mass-energy of a single clump into gravitational waves, which in turn interact with the reference singularity

[0368] Application of the method described in the previous section results in translation outside of the cosmological horizon, where F(R)<0. The topology of the resulting geodesic, for large constant R, is mathematically equivalent to a Euclidean cylinder of the condition R×Sn where T is the coordinate along the cylinder. I± are located outside the future/past cosmological horizons, where R is timelike and T is spacelike. In the case of a rotating Kerr black hole, there are two circular photon orbits that can exist in the equatorial plane and be exploited to create a tipler sinusoid. One is a prograde orbit moving in the same direction as the black hole's rotation, while the other is a retrograde orbit moving against the black hole's rotation. Their radii are respectively given by

$$r1=2M(1+\mathrm{Cos}(2/3\ \arccos(-|\alpha|/M)))$$

$$r2=2M(1+\mathrm{Cos}(2/3\ \arccos(|\alpha|/M)))$$

where a is the angular momentum per unit mass of the black hole. The orbits fall in the range M greater or equal to r1 greater or equal to 3M greater or equal to r2 greater or equal to 4M The fact that a prograde photon or in or case cooper pair with Bose characteristics orbits the black hole at a smaller radius than a retrograde one can be attributed to the well-known Lense-Thirring effect, i.e., the dragging of inertial frames due to the black hole's rotation which we control and influence via electromagnetic induced load coupling with the open magnetic field lines threading the BH horizon. This dragging would cause charged cooper pairs to revolve around the black hole relative to a static observer at infinity. Thus, to such an observer, a prograde cooper pair would have to orbit at a smaller radius to compensate for the 'extra' angular momentum acquired, while a retrograde one would have to orbit at a larger radius to compensate for the 'lost' angular momentum. Indeed, in the limit of zero rotation, these two orbits coincide at r D 3M, giving the single circular orbit of the Schwarzschild black hole. Now, recall that orbits around the Schwarzschild black hole are necessarily confined to a plane passing through its center, because of the spherical symmetry of the space-time. However, the Kerr black hole space-time has only an axial symmetry (in addition to being stationary), and this raises the possibility of non-planar orbits. One could, for example, contemplate the existence of spherical Boseon orbits—orbits with constant coordinate radii that are not necessarily confined to the equatorial plane—around the Kerr black hole. Such orbits would be a nontrivial generalization of the two circular photon orbits that lie in the equatorial plane. At first it may seem a little surprising that such spherical orbits could even exist, but there is an interesting reason as to why they are possible. Note that an object in a spherical orbit would, in addition to moving around the black hole in the azimuthal direction, be undergoing some periodic motion in the latitudinal direction. This is only possible if there is a conserved quantity associated with motion in this direction, just as angular momentum is necessarily conserved by its rotational motion in the azimuthal direction. (This result can be seen, for example, using action-angle variable) Now, because the Kerr space-time has only axial symmetry, geodesics in it should have only two constants of motion, namely energy and angular momentum. However, Carter discovered the remarkable fact that geodesics in the Kerr space-time possess a third constant of motion. It turns out that Carter's new constant governs the motion of geodesics in the latitudinal direction, although it is not related to any obvious space-time symmetry. Thus, spherical timelike orbits, which assume eigenlike properties around the Kerr black hole, are possible and when coupled with charged Bosons may be exploited to manipulate the relative event horizons of the "target" singularity. The location and the temperature of the modified event horizon depend on the time, charge and angle of incidence of the cooper pairs. The Fermionic spectrum of Dirac particles displays a spin-rotation coupling effect due to the interaction between the particles with spin-1=2 and the black holes with rotation. The effects arise from the interaction between the spin of Dirac particles and the rotation of the evaporating black holes. The feature of this spin-rotation coupling effect is its dependence on different helicity states of coupled particles with spin-½ and its irrelevance to the mass of particles.

[0369] In order to design a mechanism for time displacement exploiting the Carter asymettry cited above we utilize two positively charged top spin rotating kerr type black holes aligned on demand to create a simulated rotational effect around a central axis provided by the "reference" singularity. In the case of the Kerr blackhole (singularity), this is accomplished by an inverse Blandford-Znajek (BZ) process utilizing the magnetic flux of open field lines connecting the horizon and an induced remote load. (Ding-Xiong Wang, Kan Xiao & Wei-Hua Lei, 2001) permitting the microscopic blackhole (singularity) to be rotated along its horizontal axis in at relativistic centrifugational speeds. A differential current is then applied in such a way as to pass through the entire photosphere in the desired direction. As a result the gravity field can be manipulated by three factors that affect it in distinct ways. Adding electric charge to the singularities increases the diameter of the inner event horizons. Adding mass to the singularities increases the area of gravitational influence around the singularities. Rotating and positioning the polar axis of the singularities affects and alters the resulting gravity sinusoid

[0370] In order to create a sinusoid capable of inducing a topology change from one spacelike boundary to the other in accordance with Geroch's theorem The electric charge in the upper half of the photosphere would be maximised. The electrons will have a vector in the left to right direction as the singularity spins clockwise. If a maximised current is applied to the singularity in the same direction this will result in a increase in the velocity of the electrons relative to the centre of gravity of the singularity, due to the flow of current. In turn, according special relativity and to the space-time lattice model, this wilt result in an increase in the relativistic mass if the electrons and in turn by general relativity an increase in the release of gravitons.

[0371] Conversely in the lower half of the singularity the electrons will have a vector of motion in the right to left direction due to the spin of the singularity. This will be relativistically slowed by the differential current applied in the same direction as the current above, and hence in the opposite direction to the direction of rotation. The charge can be separately applied and adjusted to ensure that the electrons are relativistically stationary relative to the centre of gravity. In turn this will minimize the relativistic mass and result in a decrease in the release of gravitons for the lower half of the singularity.

[0372] The overall result will be a greater release of gravitons in one direction and a lesser release of gravitons in the converse direction. The effect will be enhanced by the use of a multi-phasic current simultaneously applied. This will result in the release of multi-phasic gravitons which will disperse space time in the singularity with increase in density in the converse this effect can also be produced and supplemented with the use radio frequency pulses, The radio frequency pulses must be designed to produce a change in the spin of the particle to enhance the release of gravitons in the desired direction

[0373] With the use of large currents the drift velocity of the electrons across the photosphere could be greatly increased. Within this region the electrical resistance is virtually eliminated. Thereby allowing large currents to be induced with minimum total power output.

[0374] The result is the production of cooper paired electrons of high speed and hence high relativistic mass in the desired half of the singularity, whilst producing low speed and thus low mass paired electrons at the converse of the singularity, in accordance with special relativity. The imbalance in the rotating singularity will be continuously present creating a dynamic warping of space-time. In effect, the differential current flow, will produce differential graviton production and in turn, by general relativity, the warping of space-time

[0375] As the cooper pairs cross the photosphere, accretion of the Bose particles results. Near-hole accretion across the target singularity is then motivated by magneto-rotationally induced, ultra-relativistic disk dynamics in the region of the flow bounded from below by the marginally bound geodesic radius rmb. As the particles impelled have high spin values, a largely coherent magnetic field in this region has the dynamical implication of compact mass segregation at the displacement nodes of the non-axisymmetric, MRI modes. This results in prolific gravitational wave emission coincident with the gamma-ray stage. The gravitational wave emissions are then manipulated to influence the reference singularity to produce a variable gravitational sinusoid which is then used to mathematically approximate the gravametric distortion, inducing a topology change from one spacelike boundary to the other in accordance with Geroch's theorem.

BIBLIOGRAPHY

[0376] Mesoscopic behaviour of the neutral Fermi gas ^{3}He confined in quantum wires. S. L. Phillipson, A. M. Genault, S. N Fisher, G. R. Pickett, P. J. Y Thibault. Nature, 395: 578-580.

[0377] Direct Observation of d-orbital Holes and Cu—Cu Bonding in Cu_2. J. M. Zuo et al. Nature 401: 49-52.

[0378] The Foundation of the General Theory of Relativity, A Einstein. Translated from "Die Grundlage der Allgemeinen Relativittstheorie". Annalen der Physik, 49, 1916. Published in, Einstein, The Principle of Relativity, Dover publications, 1952.

[0379] Sidelights on Relativity. Ether and the Theory of Relativity Albert Einstein. Dover Publications 1983.

[0380] Inflation is dead long live inflation. George Musser. Scientific American, July 1998, p 9.

[0381] The flip side of the Universe. George Musser. Scientific American, September 1998, p 13.

[0382] The Fifth Element, Marcus Chown. New Scientist, Apr. 3, 1999, pp 29-32

[0383] The Quantum Universe. Tony Hey and Patrick Walters. Cambridge University Press 1987. (Solving the Shrodinger equation pp. 165-167).

[0384] Complex Numbers, Mathematics for Engineers. W. Bolton; Longman Scientific & Technical; Chap. Introducing Complex Numbers, p 1-11.

[0385] Unconditional Quantum Teleportation. A. Furusawa, I. Sorensen, S. L. Braunstein, C. A. Fuchs, H. J Kimble, E. S. Polzik. Science, Oct. 23, 1998. Vol. 282: 706-709.

[0386] Complete quantum teleportation using nuclear magnetic resonance. M. A. Neilsen, E. Knill, R. Laflamme. Nature, Nov. 5, 1998. Vol. 396: 52-55.

[0387] Experimental quantum teleportation. D. Bouwmeester et al. Nature, 390: 575-579 (1997).

[0388] Experimental realization of teleporting an unknown pure quantum state via dual classical and Einstein-Podolski-Rosen channels. D. Boschi, S. Branca, F. De Martini, L. Hardy, S. Popescu. Phys. Rev. Lett. 80: 1121-1125 (1998).

[0389] W. Tittel, J. Brendel, H. Zhuniden, N. Gisin, Phys. Rev. Lett. 81: 3563-3566. (1998) into the Vortex. Charles Seife. New Scientist, Oct. 24, 1998. p 7.

[0390] Direct observation of a fractional charge. R de Picciotto, M Reznikov, M. Heibium, V Umansky, G. Bunin, D. Mahalu. Nature, 389: 162-164.

[0391] Tables of Physical and Chemical Constants. Kay & Laby. Longman Scientific & Technical.

[0392] Physics in Focus. Michael Brimicombe. Section, Wave Mechanics; p 521. Nelson Special Relativity. A. P. French. Nelson. Chapter 8, Special Relativity and Electricity, pp 256.

[0393] Universe In Balance. Jeff Peterson. New Scientist, 168: p 27

[0394] The Accelerating Universe. Mario Livio. John Wiley and Sons, Inc. (2000)

[0395] Escape from Earth, Charles Seife, New Scientist, Feb. 6, 1999. p 6.

[0396] The Discovery of the Top Quark. Tony Liss and Paul Tipton. Scientific American; September 1997, pp. 26-41.

[0397] The Mystery of the Nucleon Spin. Klaus Rith and Andreas Schaifer. Scientific American: July 1999, pp 42-47.

[0398] And you're glue. Frank Wilczek. Nature; 400, pp 21-23.

[0399] The warp drive: hyper-fast travel within general relativity Miguel Alcubierre —Classical and Quantum Gravity, vol. 11, L73-L77, (1994)

[0400] Photon propagation in a stationary warp drive space-time: Claes R. Cramer (1995)

[0401] Some thoughts on the Implications of Faster-Than-Light Interstellar Space Travel: I. A. Crawford, Quarterly Journal of the Royal Astronomical Society, vol. 36, 205-218, (1995)

[0402] Physical and Cosmological Implications of a Possible Class of Particles Able to Travel Faster than Light: Luis Gonzalez-Mestres: Contribution to the 28th International Conference on High Energy Physics, Warsaw (Poland), (1996)

[0403] Warp drive and causality: Allen E. Everett, Physical Review D, vol. 53, 7365-7368, (1996)

[0404] A Superluminal Subway: The Krasnikov Tube, Allen E. Everett & Thomas A. Roman, Physical Review D, vol. 56, 2100-2108, (1997)

[0405] Quantum effects in the Alcubierre warp drive spacetime, William A. Hiscock, Classical and Quantum Gravity, vol. 14, L183-L188 (1997)

[0406] The unphysical nature of "Warp Drive", Michael J. Pfenning & L. H. Ford; Classical and Quantum Gravity, vol. 14, 1743-1751, (1997)

[0407] On the Possibility of a Propulsion Drive Creation Through a Local Manipulation of Spacetime Geometry; Vesselin Petkov; Presented at the 34th

[0408] AIAA/ASME/SAE/ASEE Joint Propulsion Conference, (1998)

[0409] 'Operational' energy conditions, Adam D. Helfer; Classical and Quantum Gravity, vol. 15, 1169-1183, (1998)

[0410] No warp drive; D. H. Coule; Classical and Quantum Gravity, vol. 15, 2523-2527, (1998)

[0411] Quantum Inequality Restrictions on Negative Energy Densities in Curved Spacetimes, Michael John Pfenning; Doctoral Dissertation, (1998)

[0412] Hyperfast Interstellar Travel in General Relativity; S. V. Krasnikov; Physical Review D, vol. 57, 4760, (1998)

[0413] Superluminal travel requires negative energies; Ken D. Olum; Physical Review Letters, vol. 81, 3567-3570, (1998)

[0414] Hyper-fast travel without negative energy, Eric Baird; (1999)

[0415] Warp drives, wavefronts and superluminality, Eric Baird; (1999)

[0416] A traversable wormhole, S. Krasnikov; (1999)

[0417] Speed Limits in General Relativity; Robert J. Low; Classical and Quantum Gravity, vol. 16, 543-549, (1999)

[0418] Null geodesics in the Alcubierre warp drive spacetime: the view from the bridge; Chad Clark, William A. Hiscock & Shane L. Larson; Classical and Quantum Gravity, vol. 16, 3965-3972, (1999)

[0419] A 'warp drive' with more reasonable total energy; Chris Van Den Broeck; Classical and Quantum Gravity, vol. 16, 3973-3979, (1999)

[0420] On the warp drive space-time, Pedro F. Gonzalez-Diaz; Physical Review D, vol. 62, 44005-44012, (2000)

[0421] On the (im)possibility of warp bubbles; Chris Van Den Broeck; Summary of talk delivered at STAIF-2000, (2000)

[0422] Reduced Total Energy Requirements for a Modified Alcubierre Warp Drive Spacetime; F. Loup, D. Waite & E. Halerewicz Jr. (2001)

[0423] Warp Drive With Zero Expansion, Jose Natario; Classical and Quantum Gravity, vol. 19, 1157-1166, (2002)

[0424] A Causally Connected Superluminal Warp Drive Spacetime; F. Loup, R. Held, D. Waite, E. Halerewicz, Jr., M. Stabno, M. Kuntzman & R. Sims, (2002)

[0425] Weak Energy Condition Violation and Superluminal Travel; Francisco Lobo & Paulo Crawford (2002)

[0426] On the Problems of Hazardous Matter and Radiation at Faster than Light Speeds in the Warp Drive Space-Time; C. B. Hart, R. Held, P. K. Hoiland, S. Jenks, F. Loup, D. Martins, J. Nyman, J. P. Pertierra, P. A. Santos, M. A. Shore, R. Sims, M. Stabno & T. O. M. Teage, (2002)

[0427] 2+1 gravity, chaos and time machines; Bengtsson I and Brannlund J, J. Math. Phys. 42 (2001) 3565

[0428] Time Travel in Einstein's Universe: the Physical Possibilities of Travel through Time, Gott J R; Houghton Mifflin Company, Boston, 2001

[0429] Gravitational memory of natural wormholes, Anchordoqui L A Trobo M L Vucetich H Zyserman F Mod Phys Lett A 15(6): 429-438, 2000

[0430] A traversable singularity-free wormhole. The gravitational memory phenomenon Gott time machines, BTZ black hole formation, and Choptuik scaling, Birmingham D, Sen S; Phys Rev Left 84 (2000) 1074-1077

[It is shown that the Gott time machine provides a mechanism for black hole formation]

[0431] Negative energy, wormholes and warp drive, Ford L H, Roman, T A; Sci Am: 46-53, 2000 Jan (reply: 2000 May p. 12)

[0432] Warp drive space-time; Gonzalez-Diaz P F; Phys Rev D62: 044005 (2000)

[0433] Misner-brane cosmology, Gonzalez-Diaz P F; Phys Lett B486: 158-164 (2000)

[0434] Kinks, energy conditions and closed timelike curves; Gonzalez-Diaz P F; Int J Mod Phys D 9(5): 531-541, 2000. (gr-qc/0007065)

[0435] Will we travel back (or forward) in time? Gott J R; Time, Apr. 10, 2000 pp 68-70

[0436] Charged universes of the Godel type with closed timelike curves; Klepac P and Horsky J, Class Quant Grav 17(13): 2547-2555, 2000

[0437] The Ori-Soen time machine, Olum K D; Phys Rev D61 (2000) 124022

[0438] How much energy do closed timelike curves in 2+1 spacetimes need?, Tiglio M H, Phys Rev D61 (2000) 08 1503

[0439] Energy conditions and their cosmological consequences, Visser M and Barcelo C gr-qc/0001099

[0440] Cylindrically symmetric spinning Brans-Dicke spacetimes with closed timelike curves Anchordoqui L A, Perez Bergliaffa S E, Trobo M L, Birman G S; Mod Phys Lett A14 (1999): 1105-1112

[0441] From spinning to non-spinning cosmic string spacetime; De Lorenci V A De Paola R DM Svaiter N F, Class Quant Grav 16(10): 3047-3055, 1999

[0442] Bounds on negative energy densities in static space-times, Fewster C J, Teo E; Phys Rev D 5910(10): 4016, 1999

[0443] Decohering a charged scalar field in a time-machine wormhole background, Flagga M SN; Int J Theor Phys 38(8): 2121-2161, 1999

[0444] Rotating black holes in unified theories; Gibbons G W. Progress of Theoretical Physics Supplement. (136): 18-28, 1999

[0445] Supersymmetric rotating black holes and causality violation; Gibbons G W and Herdeiro C AR, Class Quant Grav 16 (11) 3619-3652, 1999

[0446] Perdurance of multiply connected de Sitter space; Gonzalez-Diaz P F, Phys Rev D 59, 1999, 123513

[0447] Nonorientable spacetime tunneling; Gonzalez-Diaz P F, Garay L J, Phys Rev D 59 (1999) 064026

[0448] Can the universe create itself?; Gott J R, and Li X L, Phys Rev D 58, 1998, 023501. Also: astro-ph/9712344

[0449] Supersymmetric rotating black holes and causality violation; Holst S, Matschull H-J Class Quant Grav 16, 1999, 3095

[0450] Godel-type universes in string-inspired charged gravity; Kanti P, Vayonakis C E, Phys Rev D 60, 1999, 103519

[0451] Quantum field theory and time machines; Krasnikov S., Phys Rev D 59: 4010, 1999 Jan 14 Dec. 1998

[0452] Time machines constructed from anti-de Sitter space; Li L X, Phys Rev D 59 (1999) 084016

[0453] Wormholes in string theory, Vollick D N., Class Quant Grav 16(5):1599-1604, 1999

[0454] Time machines and the breakdown of unitarity; Antonsen F and Bormann K, Int J Theor Phys 37(9): 2383-2393, 1998

[0455] An exterior for the Godel spacetime; Bonnor W B, Santos N O, MacCallum M A H, Class. Quant. Grav. 15 (1998) 357-366

[0456] All the worlds a time machine, Chown M, New Scientist 157(2124): 38-41, 1998 Mar. 7.

[short exposition of M. Hadleys idea that an elementary particle is a local ST warp that it permits closed timelike curves. Hence general relativity could reproduce (explain) the effects of QM]

[0457] Quantum time machine; Gonzalez-Diaz P F Phys Rev D 58 (12): 4011, 1998

[0458] A design for a quantum time machine, Gray A quant-ph/9804008

{misleading title. On QM measurement problem; quantum "seeing into future."}

[0459] Time machines with non-compactly generated cauchy horizons and "handy singularities"; Krasnikov S V, gr-qc/9711040 (To appear in the proceedings of the 8th Marcel Grossman Conference)

[0460] A singularity-free WEC-respecting time machine, Krasnikov S V, Class Quant Grav 15 (1998) 997-1003

[0461] Self-consistent vacuum for Misner space and the chronology protection conjecture, Li L X and Gott J R, Phys Rev Lett 80, 2980, 1998

{Misner space stable to quantum flutuations; against CPC}

[0462] Decoherence caused by topology in a time-machine spacetime; Mensky M B and Novikov I D, Int. J. Mod. Phys. D5 (1996) 1-27

[0463] Godel metric as a squashed anti-de Sitter geometry, Rooman M, Spindel Ph, Class Quant Grav 15 (1998) 3241-3249

[0464] Testing causality violation on spacetimes with closed timelike curves; Rosenberg S, Phys Rev D 57(6):3365-3377, 1998

{politzer TM used to investigate the scattering experiment; interesting results on agreement with the no CTC case if there is enough TM's}

[0465] A new non-Hausdorff spacetime model for resolution of the time travel paradoxes; Sharlow M F Annals of Physics 263(2): 179-197, 1998

[0466] Particle creation near the chronology horizon; Sushkov S V, Phys Rev D 58, 1998, 044006

[0467] Divergences in the effective action of acausal spacetimes, Cassidy M J, Class Quant Grav 14 p 523, 1997 {shows that a self-consistent quantum vacuum for Misner space exist, allowing time travel}

[0468] The 1-loop instability of Misner space; Cassidy M J, Class Quant Grav 15 (8),

[0469] Superluminal subway—the Krasnikov tube; Everett A E and Roman T A, Phys Rev D 56 (4): 2100-2108, 1997

[0470] Time machines and the principle of self-consistency as a consequence of the principle of stationary action (ii)—the Cauchy problem for a self-interacting relativistic particle, Carlini A and Novikov I D, Int J Mod Phys D 5(5): 445-479, 1996 Oct.

[0471] Warp drive and causality, Everett A E, Phys Rev D 53 (12) pp 7365-7368, 1996 Ringholes and closed timelike curves, Gonzalez-Diaz P F Phys Rev D 54(10): 6122-6131, 1996

[0472] Time machine as four-dimensional wormhole; Guts A K gr-qc/9612064

[0473] Gott time machines in the Anti-de Sitter space; Holst S, Gen Rel Grav 28 (1996) 387-403

[0474] The quantum stability of the time machine, Krasnikov S V, Phys Rev D54 (1996) 7322-7327

[0475] Must time machine be unstable against vacuum fluctuations?, Li L X, Class Quant Grav 13 (1996) 2563-2568

[0476] Three-dimensional billiards with time machine, Mensky M B and Novikov I D, Int J Mod Phys D 5(2): 179-192, 1996

[0477] Decoherence caused by topology in a time-machine spacetime, Mensky M B and Novikov I D, Int J Mod Phys D 5(1): 1-27, 1996.

[0478] Improved time-machine model; Soen Y and Ori A, Phys Rev D 54(8): 4858-4861,

[0479] Unitarity restoration in the presence of closed timelike curves; Anderson A, Phys Rev D51 (1995) 5707-5715

[0480] Problems in time-machine construction due to wormhole evolution; Antonsen F and Bormann K, Int J Theor Phys 34 1995, p 2061

[0481] Time machines: the principle of self-consistency as a consequence of the principle of minimal action; Carlini A, Frolov V P, Mensky M B, Novikov I D, and Soleng H H, Int J Mod Phys D4 (1995) 557-580; Erratum-ibid. D5 (1996) 99

[0482] Nonlinearity in quantum theory and closed timelike curves; Cassidy M J, Phys Rev D52 (1995) 5676-5680

[0483] Unitarity of Quantum Theory and Closed Time-Like Curves; Fewster C J, Wells C G, Phys Rev D52 (1995) 5773-5782

[0484] Time machines without closed causal geodesics; Low R J, Class Quant Grav 12(5): L 37-L 38, 1995 May.

[0485] A remark on kinks and time machines; Yurtsever U, Gen Rel Grav 27(6): 691-693, 1995

[0486] The warp drive: hyper-fast travel within general relativity; Alcubierre M, Class Quantum Grav 11, L73 (1994)

[0487] Energy-momentum restrictions on the creation of Gott time machines; Carroll S M, Farhi E, Guth A H and Olum K D, Phys Rev D 50 pp 6190-6206, 1994

[0488] Kinks and time machines; Chamblin A, Gibbons G W and Steif A R, Phys Rev D 50(4) pp R2353-R2355, 1994

[0489] The breakdown of quantum mechanics in the presence of time machine, Goldwirth D S, Perry M J and Piran T, Gen Rel Grav 25 pp 7-13 1994

[0490] Quantum propagator for a nonrelativistic particle in the vicinity of a time machine; Goldwirth D S, Perry M J, Piran T and Thorne K S, Phys Rev D 49(8): 3951-3957, 1994

[0491] Time machine, self-consistency and the foundations of quantum mechanics; Gonella F, Found Phys Lett 7(2): 161-166, 1994

[0492] Time walk through the quantum cosmic string; Gonzalez-Diaz P F, gr-qc/9412034

[0493] Unitarity and causality in generalized quantum mechanics for nonchronal spacetimes; Hartle J B, Phys Rev D 49 (12) pp 6543-6555, 1994

[0494] (2+1)-dimensional spacetimes containing closed timelike curves; Headrick M P, Gott J R; Phys Rev D 50, 1994, 7244-7259

[0495] Time travel, Jackiw R.; Ukrainskii Fizicheskii Zhurnal. 39(1): 6-14, 1994.

[0496] On the classical stability of a time machine; Krasnikov S V, Class Quant Grav 11 (11): 2755-2759, 1994 Nov.

[0497] Isometries between Gott's two-string spacetime and Grant's generalization of Misner space, Laurence D, Phys Rev D 50 pp 4957-4965

[0498] Negative Energy Densities in Extended Sources Generating Closed Timelike Curves in General Relativity with and without Torsion; Soleng H H, Phys Rev D49 (1994) 1124-1125

[0499] Multiparticle solutions in 2+1 gravity and time machines; Steif A R, Int J Mod Phys D 3(1): 277-280, 1994

[0500] VanVleck determinants—traversable wormhole spacetimes; Visser M

[0501] Algebraic approach to quantum field theory on non-globally-hyperbolic spacetimes; Yurtsever U, Class Quant Grav 11 (1994) 999

[0502] Cosmic strings and chronology protection; Grant J DE, Phys Rev D 47, 2388 (1993)

[0503] Complex geometry, quantum tunneling, and time machines; Li L X, Xu J M and Liu L, Phys Rev D 48(10): 4735-4737, 1993

[0504] Quantum field theory in spaces with closed timelike curves; Boulware D G, Phys Rev D46 (1992) 4421-4441

[0505] Global structure of Gott's two-string spacetime; Cutler C, Phys Rev D 45 p. 487-494, 1992 {Cauchy horizons for Gott's machine (two-string space-time)}

[0506] "Cosmic flashing" in four dimensions; Ford L H and Roman T A, Phys Rev D 46, 1328-1339 (1992)

[0507] Unitarity of interacting fields in curved spacetime; Friedman J L Papastamatiou, Phys Rev D 46 4442, 1992

[0508] Feilure of unitarity for interacting fields on spacetimes with closed timelike curves; Friedman J L Papastamatiou, and Simon J Z, Phys Rev D 46 4456-4469, 1992

[0509] Particle creation for time travel through a wormhole; Kim S W, Phys Rev D 46 2428-2434, 1992

[0510] Time machine and self-consistent evolution in problems with self-interaction; Novikov I D, Phys Rev D 45 p. 1989, 1992

[0511] Structure of the singularity inside a realistic rotating black hole, Phys Rev Lett} 68, 2117-2120; Ori A (1992)

[0512] Simple quantum systems in spacetimes with closed timelike curves; Politzer H D, Phys Rev D 46 4470-4476, 1992

[0513] Quantum mechanics near closed timelike lines; Deutsch D, Phys Rev D 44 (10) p 3197-3217 1991

[0514] Billiard balls in wormhole spacetime with closed timelike curves: Classical theory; Echeverria F, Klinkhammer G and Thorne K S, Phys Rev D 44 (4) p. 1077-1099, 1991

[0515] Closed timelike curves produced by pairs of moving cosmic strings: Exact solutions; Gott J R, Phys Rev Left 66, 1126, (1991)

[0516] Inner structure of a charged black hole: An exact mass-inflation solution; Ori A, Rev Left 67: 789, 1991

[0517] On the outcome of Kerr-like collapse; Charlton N and Clarke C J S, Classical and Quantum Gravity 7 No. 5 (May 1990)

[0518] Cauchy problem in spacetimes with closed timelike curves; Friedman J, Morris M S, Novikov I D, Echeverria F, Klinkhammer G, Thorne K S and Yurtsever U, Phys Rev D 42 (6) 1915-1930, 1990 {QM fixes the TM classical billiard paradox (Polchinski's prdx. "CTC's may be not as nasty as people have assumed". The principle of self-consistency in time travel reviewed.}

[0519] Traversable wormholes from surgically modified Schwarzschild spacetimes, Visser M, Nucl Phys B 328 pp 203-212

[0520] Traversable wormholes: some simple examples; Visser M; Phys Rev D 39: 3182-3184 1989

[0521] Minimal acceleration requirements for time travel in Godel spacetime; Malament D, Journal of Mathematical Physics 26, 1985, p 774-777

[0522] Timelike nongeodesic trajectories which violate causality: A rigorous derivation; de Felice F, Nuovo Cimento 65 B pp 224-232, 1981

[0523] Rotating cylinders and the possibility of global causality violation; Tipler F J, Phys Rev D 9 pp 2203-2206, 1974

[0524] Global structure of the Kerr family of gravitational fields, Carter B, Phys Rev 174, p 1559, 1968

[0525] Complete analytic extension of the symmetry axis of Kerr's solution of Einstein's equations, Carter B, Phys Rev 141, p 1242, 1966

[0526] Acta Physica Polonica, Staruszkiewicz A 24 1963, p 734

[0527] An example of a new type of cosmological solutions of Einstein's field equations of gravitation, Goedel K, Rev Mod Phys 21 p. 447, 1949

[0528] Gravitational field of a distribution of particles rotating about an axis of symmetry; Van Stockum W J, Proc R Soc Edin 57 pp 135-154, 1937

[0529] Appendix A: The Synthetic creation of micro singularities. Throughout this patent application the creation of micro singularities has been maintained. While at first glance this may seem to be a deficiency in the patent application it is in fact an engineering problem that is in the process of being solved. In the International Journal of Theoretical Physics 41 (11): 2073-2090, November 2002 L. J. Garay published "Black Holes in Bose-Einstein Condensates". In this article Garay showed that there exist both dynamically stable and unstable dilute-gas Bose-Einstein condensates that, in the hydrodynamic limit, exhibit a behavior completely analogous to that of gravitational black holes. This discovery coupled with creation of color glass condensate (CGC) an extreme form of nuclear matter in which a nucleus travels at near-light (relativistic) speed which flattens like a pancake in its direction of motion and spawns a large number of gluons hold the key to high energy singularity formation.

I claim:

1. A method for the generation of a pseudo 2+1 dimensional anti-de Sitter space comprising the steps of:

creating two Kerr type positively charged rotating dilation singularities, including the steps of

maintaining one of the singularities as a axis of rotation reference singularity,

maintaining the other of the singularities as a target singularity, and

subjecting the target singularity to a differential electron flow so as to simultaneously pass the differential electron flow above a photosphere of said target singularity in a direction of rotation thereof and contrary to the direction of rotation thereof, in order to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the target singularity around the axis of rotation provided by the reference singularity.

2. A method of generating a force around a body, comprising the steps of:

employing sinusoidal oscillations of electrical bombardment on the surface of one Kerr type reference singularity in close proximity to a second Kerr type target singularity to take advantage of the Lense-Thirring effect, wherein the electrical currents employed in the bombardment are passed simultaneously across the photosphere of said reference singularity in its direction of rotation and contrary to its direction of rotation to release a directed flow of gravitons in a sinusoidal oscillation simulating a rotational effect of the target singularity around the axis of rotation provided by the reference singularity;

creating timelike curves in a compact time-oriented manifold of Godel-type geodesically complete spacetime envelope under the Van Den Broeck modification of the Alcubierre geometry, resulting in the creation of timelike curves in a compact time-oriented manifold permitting topology change from one spacelike boundary to the other in accordance with Geroch's theorem.

* * * * *

US 20070285325A1

(19) **United States**
(12) **Patent Application Publication**
St. Clair

(10) **Pub. No.: US 2007/0285325 A1**
(43) **Pub. Date: Dec. 13, 2007**

(54) **CHI ENERGY AMPLIFIER**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
52 KINGS COURT, 4A
SAN JUAN, PR 00911

(21) Appl. No.: **11/448,456**

(22) Filed: **Jun. 7, 2006**

Publication Classification

(51) **Int. Cl.**
H01Q 1/00 (2006.01)
H01Q 13/10 (2006.01)

(52) **U.S. Cl.** **343/720**; 343/767

(57) **ABSTRACT**

A Chi energy amplifier that utilizes a co-gravitational K field to generate a flow of hyperspace energy into or out of a ceramic dome containing a slot antenna.

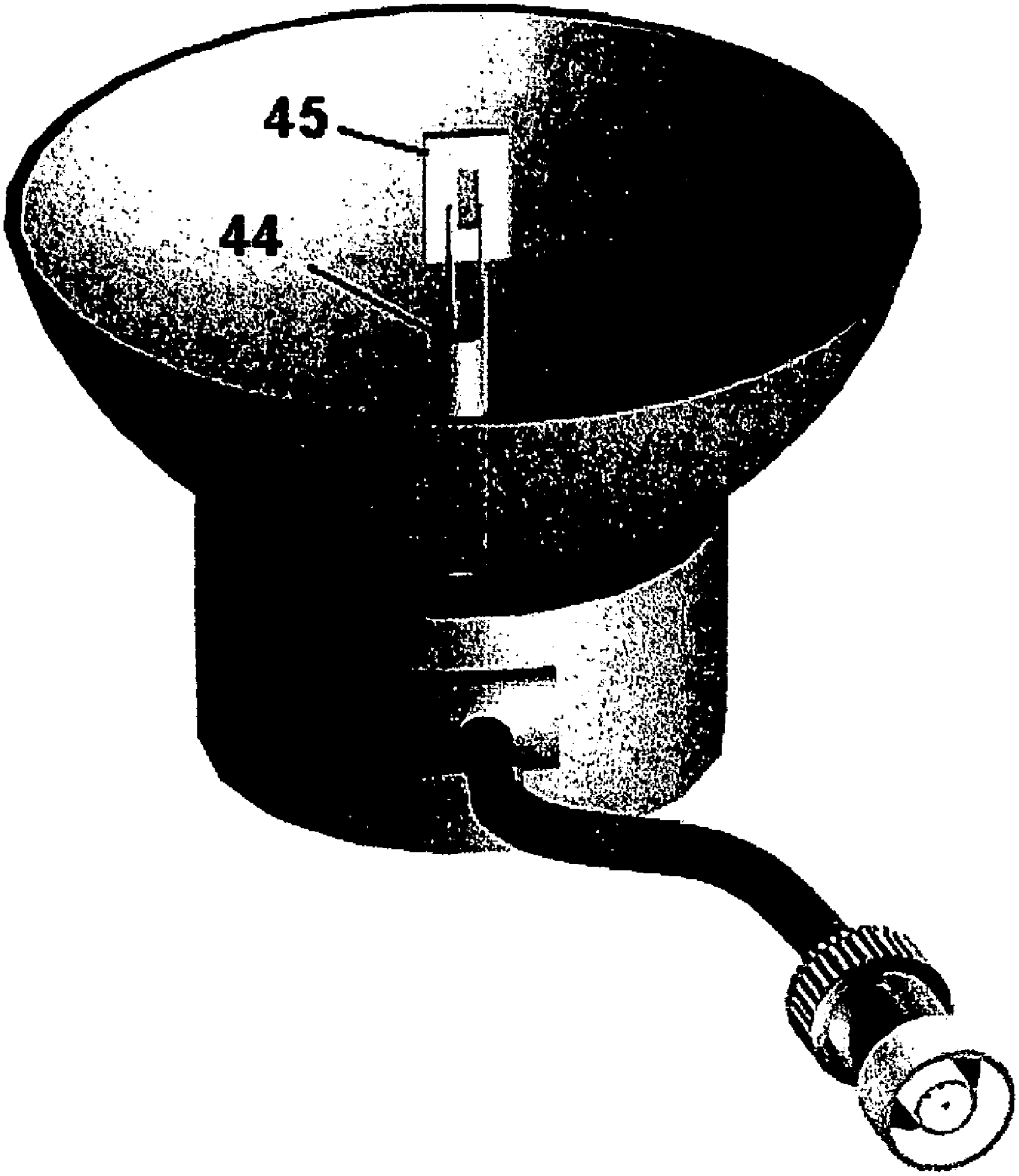

Figure 1

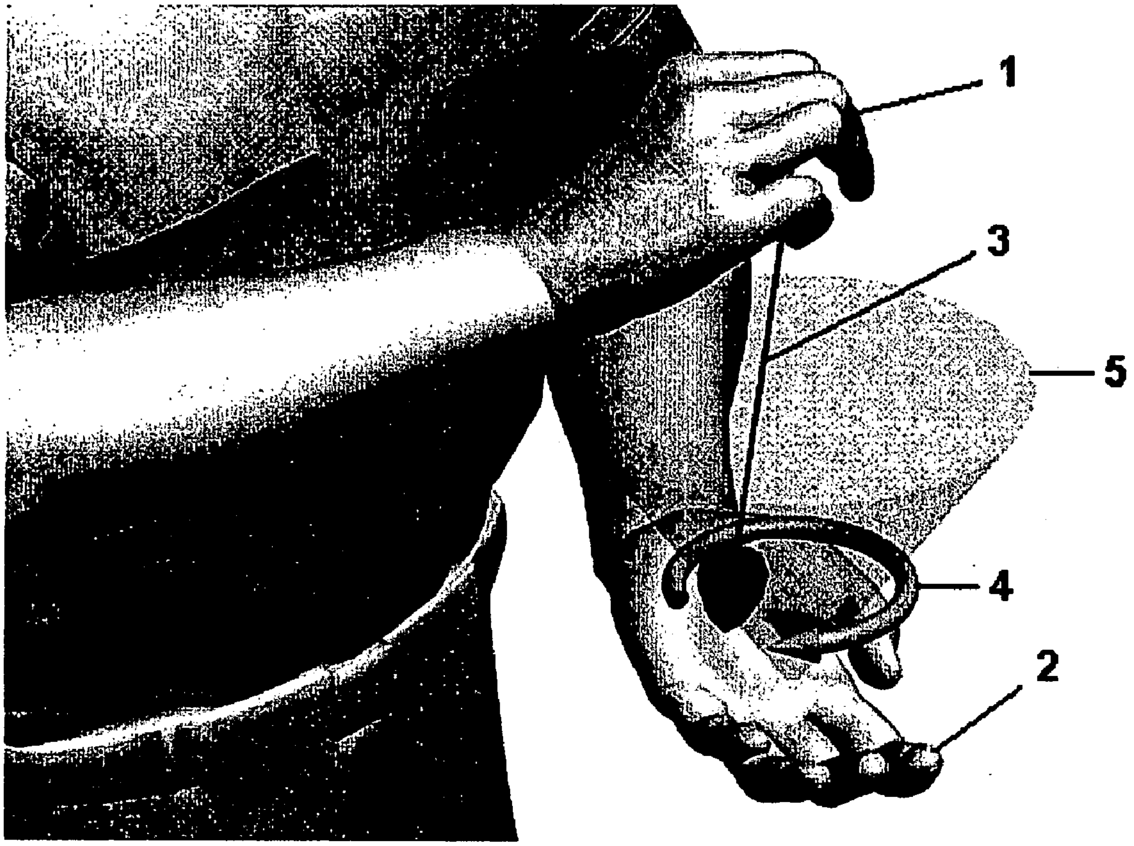

Figure 2

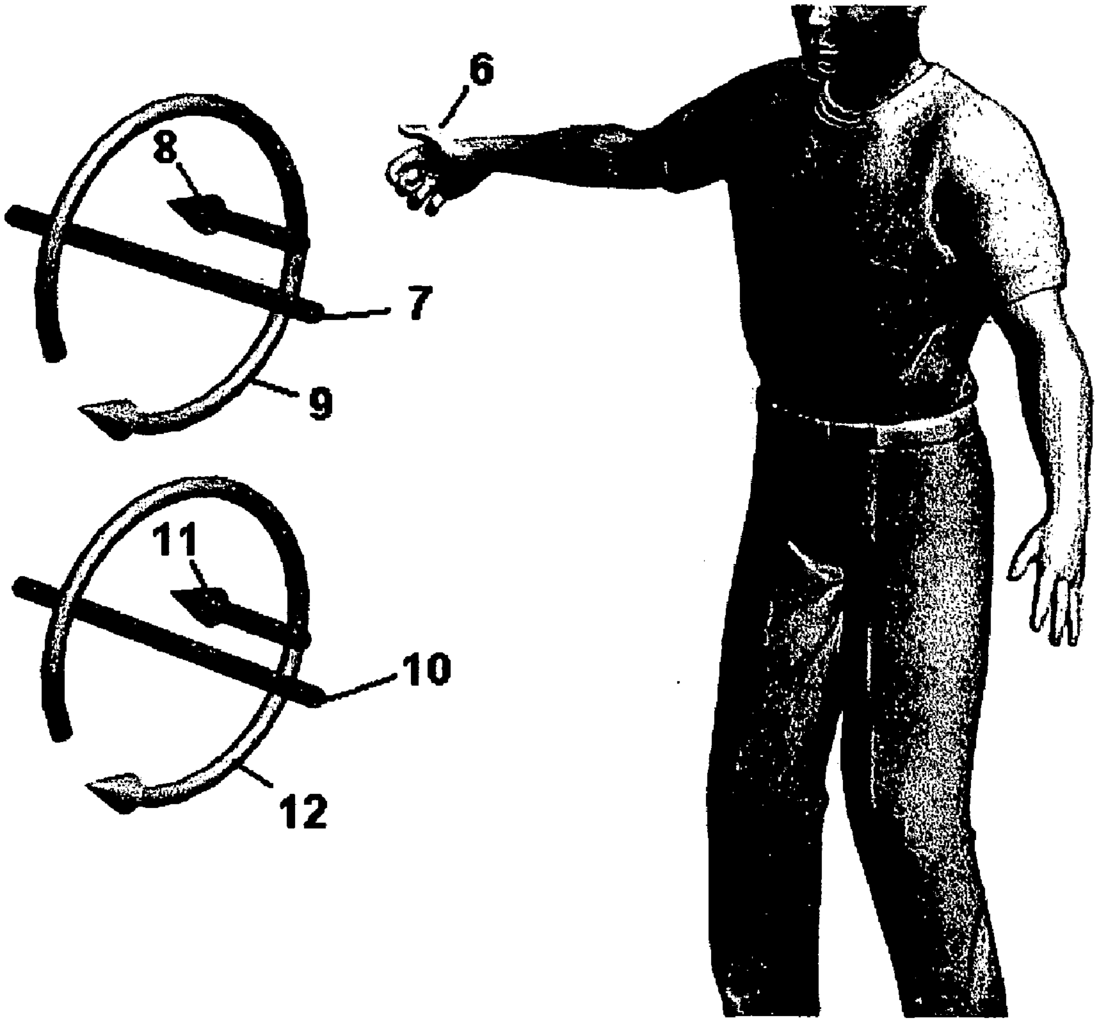
6
8
7
9
11
10
12

Figure 3

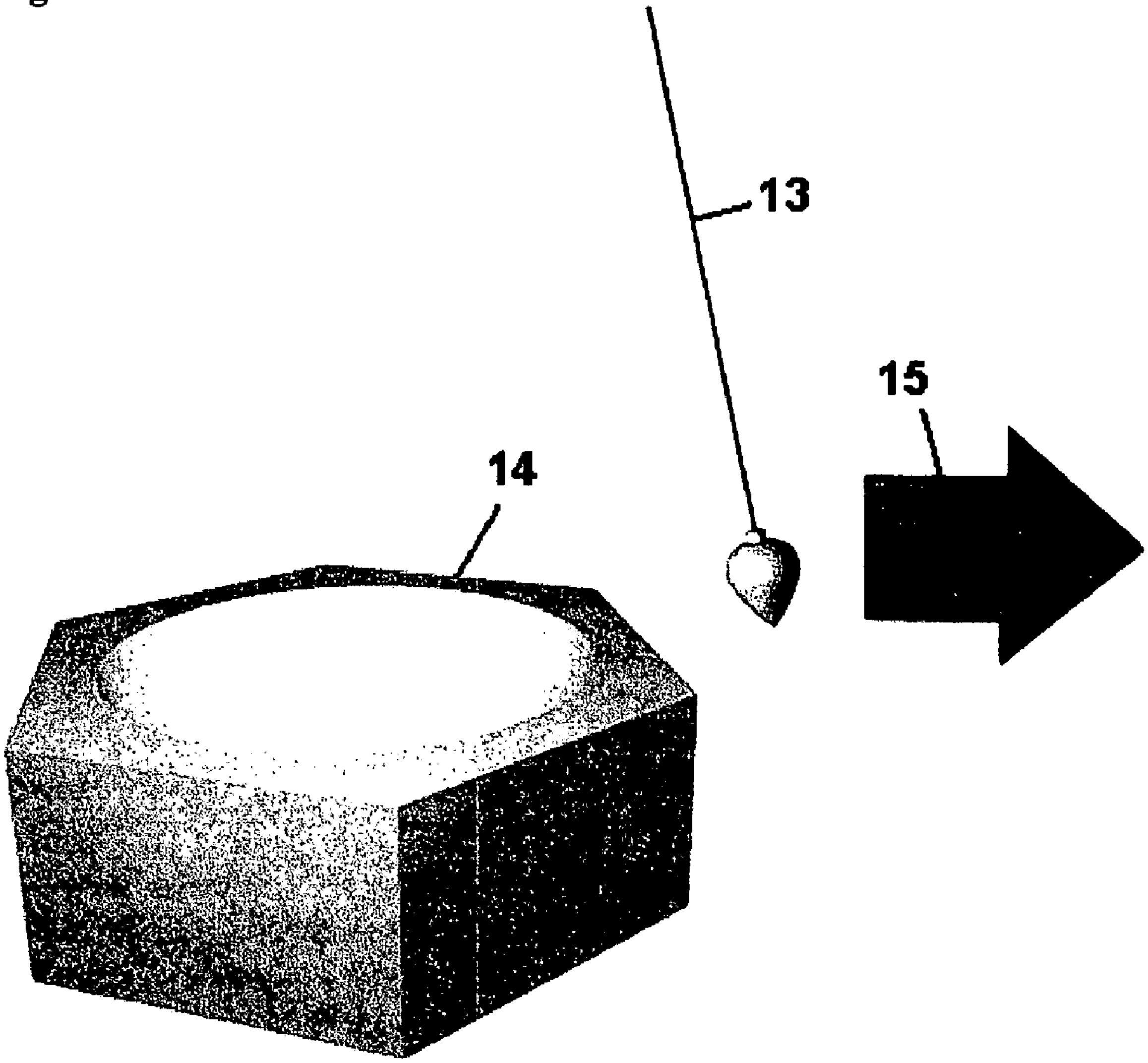

Figure 4

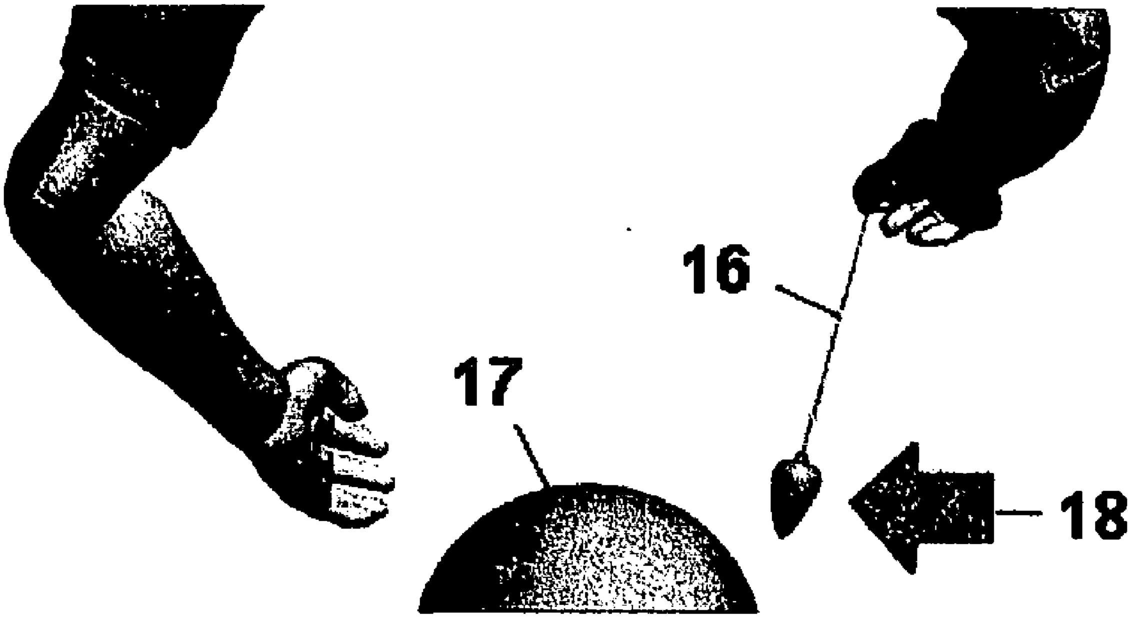

Figure 5

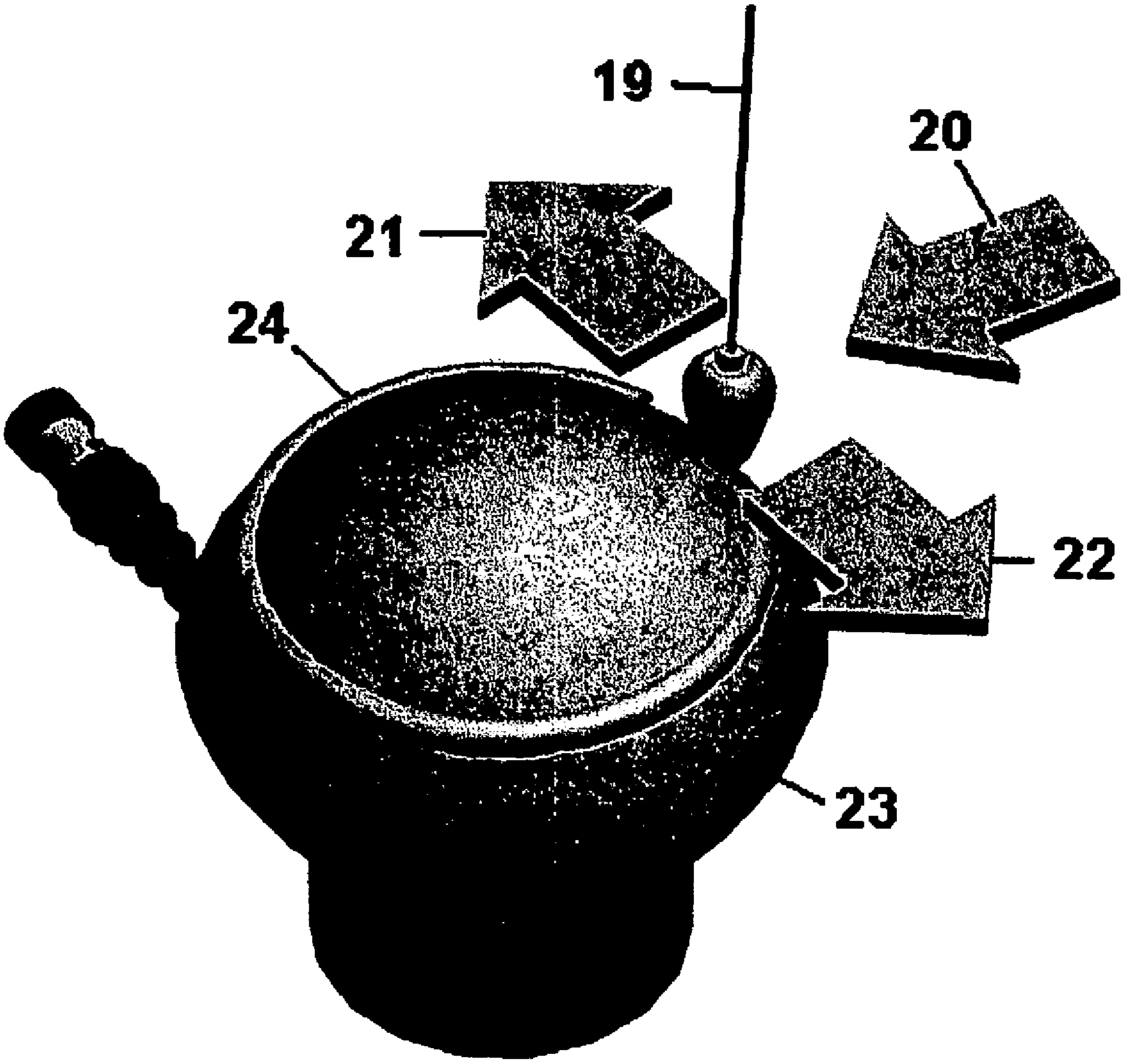
19
20
21
24
22
23

Figure 6

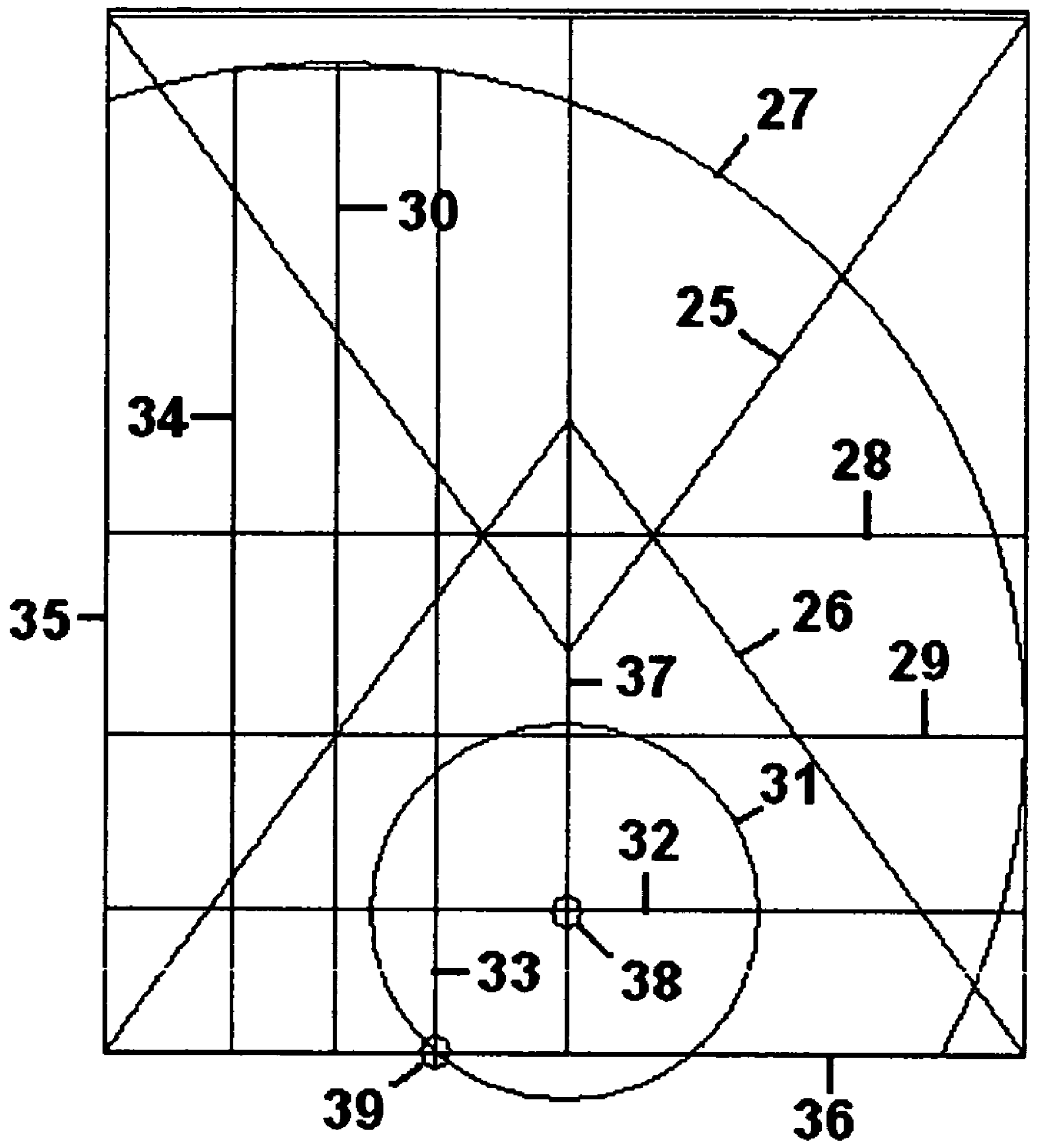
27
30
25
34
28
35
26
37
29
31
32
33
38
39
36

Figure 7

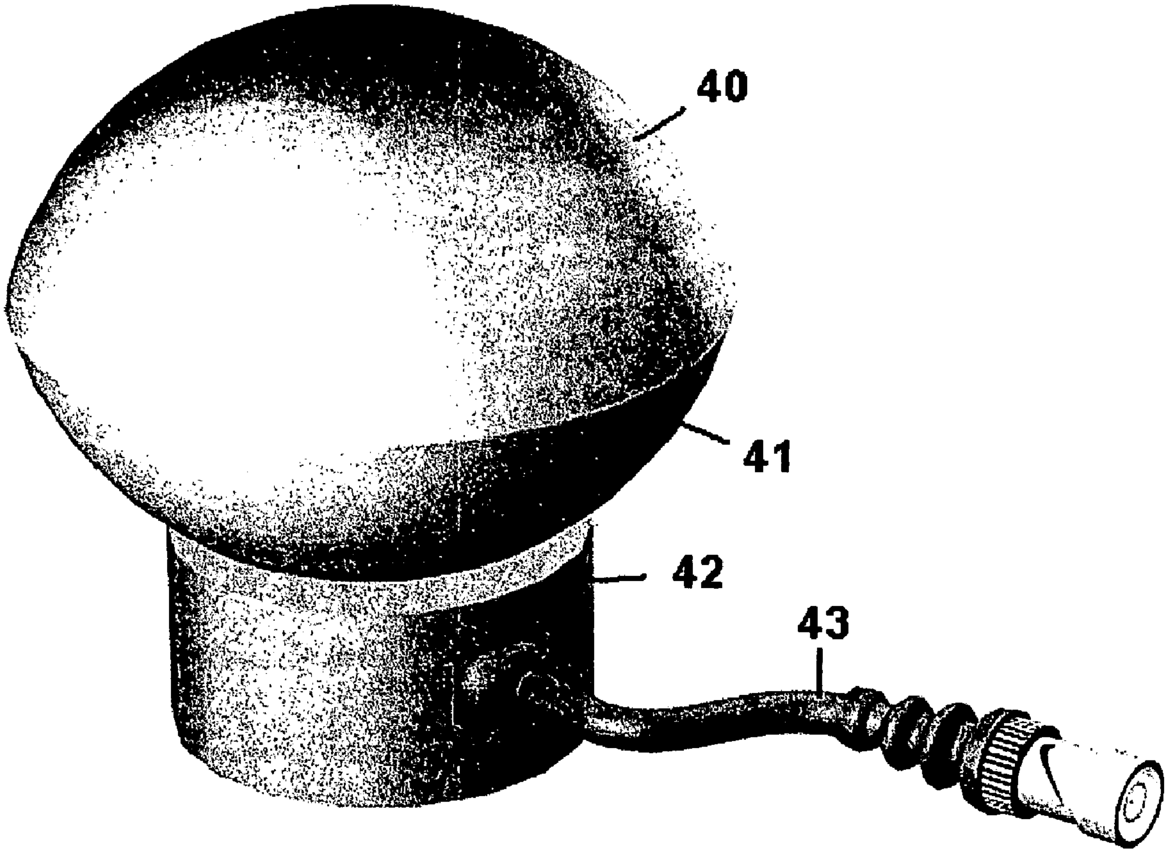

Figure 8

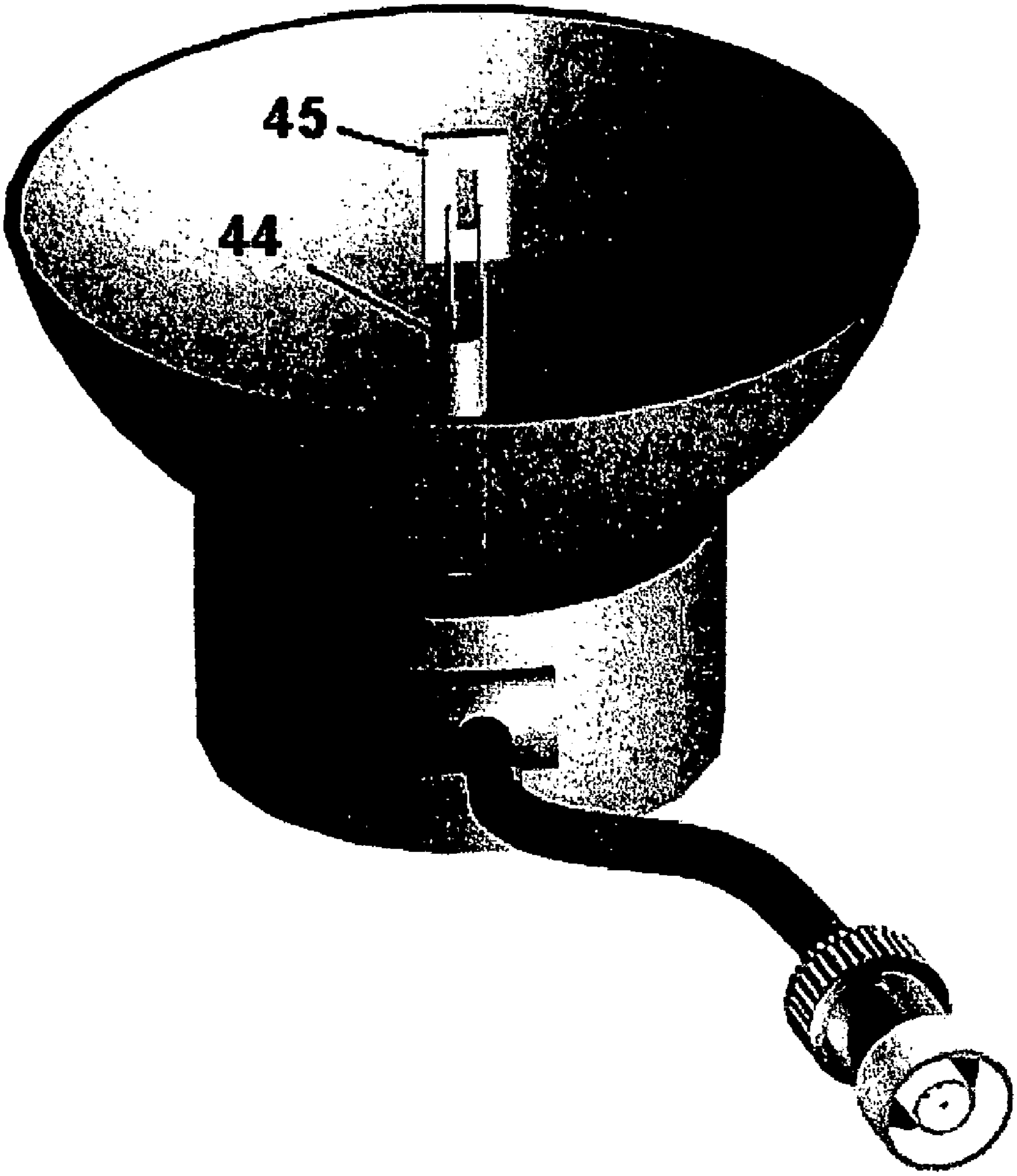

Figure 9

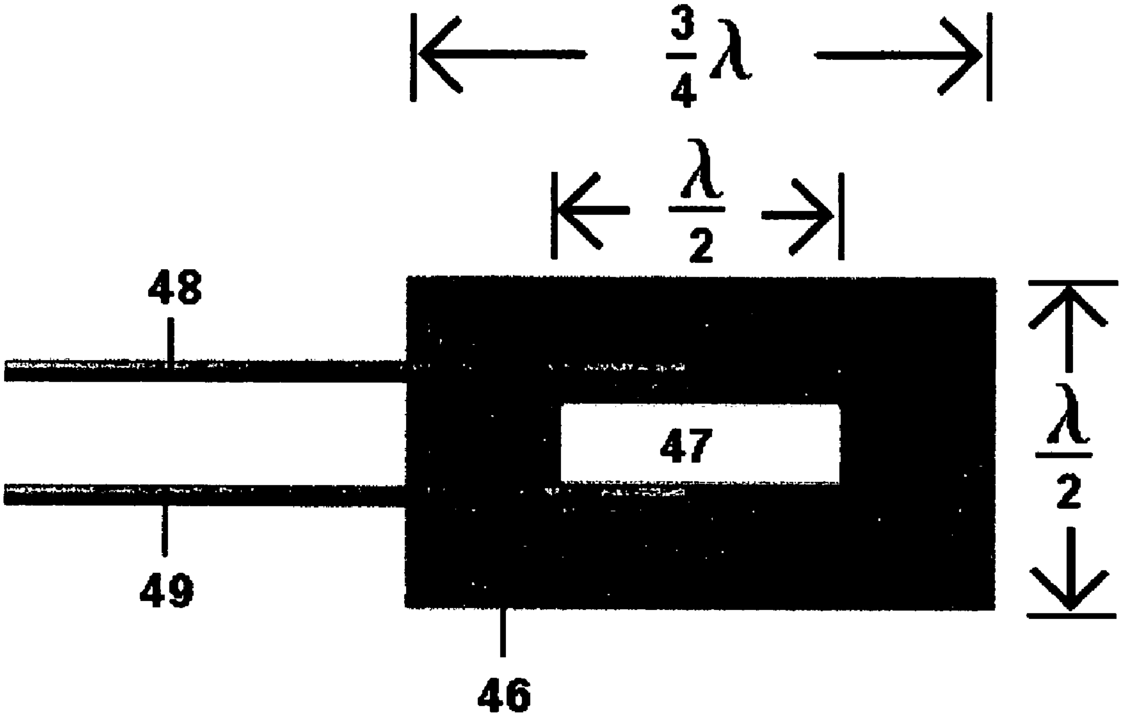
$\frac{3}{4}\lambda$
$\frac{\lambda}{2}$
48
47
$\frac{\lambda}{2}$
49
46

Figure 10

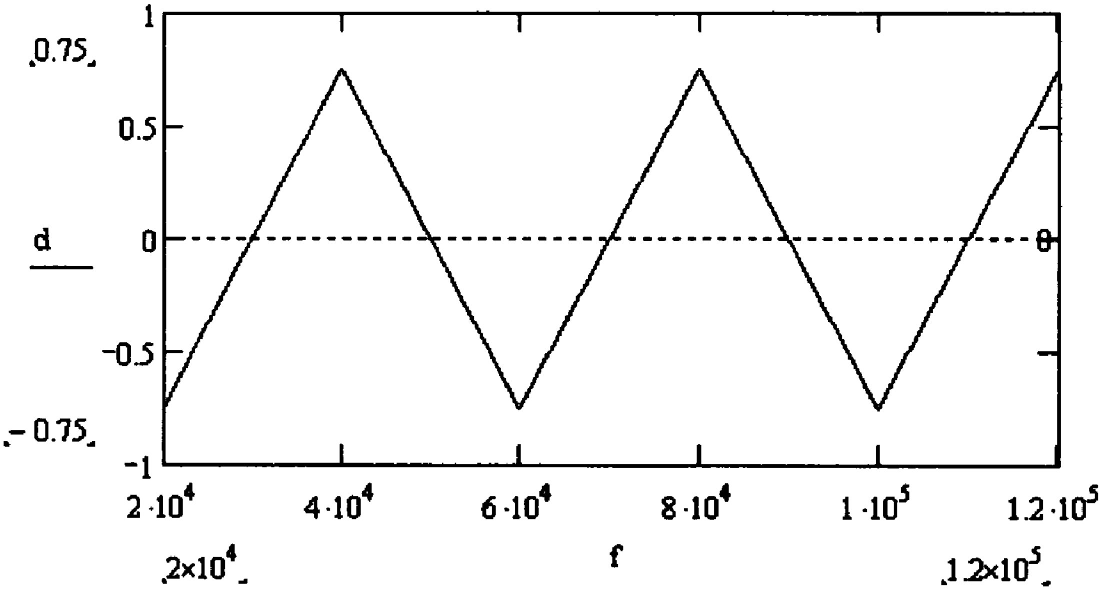
1
0.75
0.5
d
0
-0.5
- 0.75
-1
2·10⁴
4·10⁴
6·10⁴
8·10⁴
1·10⁵
1.2·10⁵
2×10⁴
f
1.2×10⁵
0

Figure 11

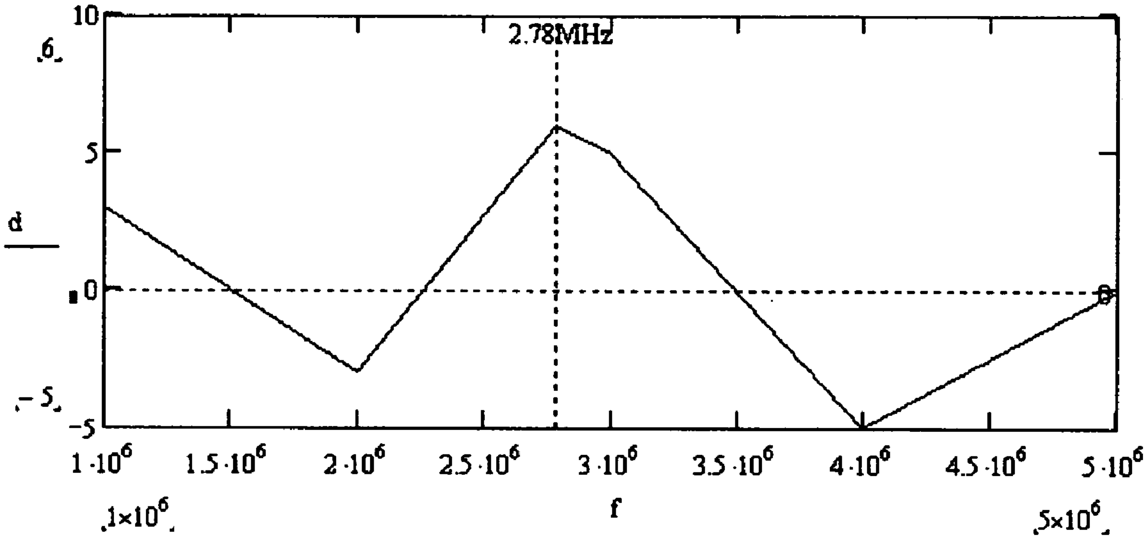
10
.6.
5
d
—
0
.−5.
−5
2.78MHz
1·10^6
1.5·10^6
2·10^6
2.5·10^6
3·10^6
3.5·10^6
4·10^6
4.5·10^6
5·10^6
.1×10^6.
f
.5×10^6.

Figure 12

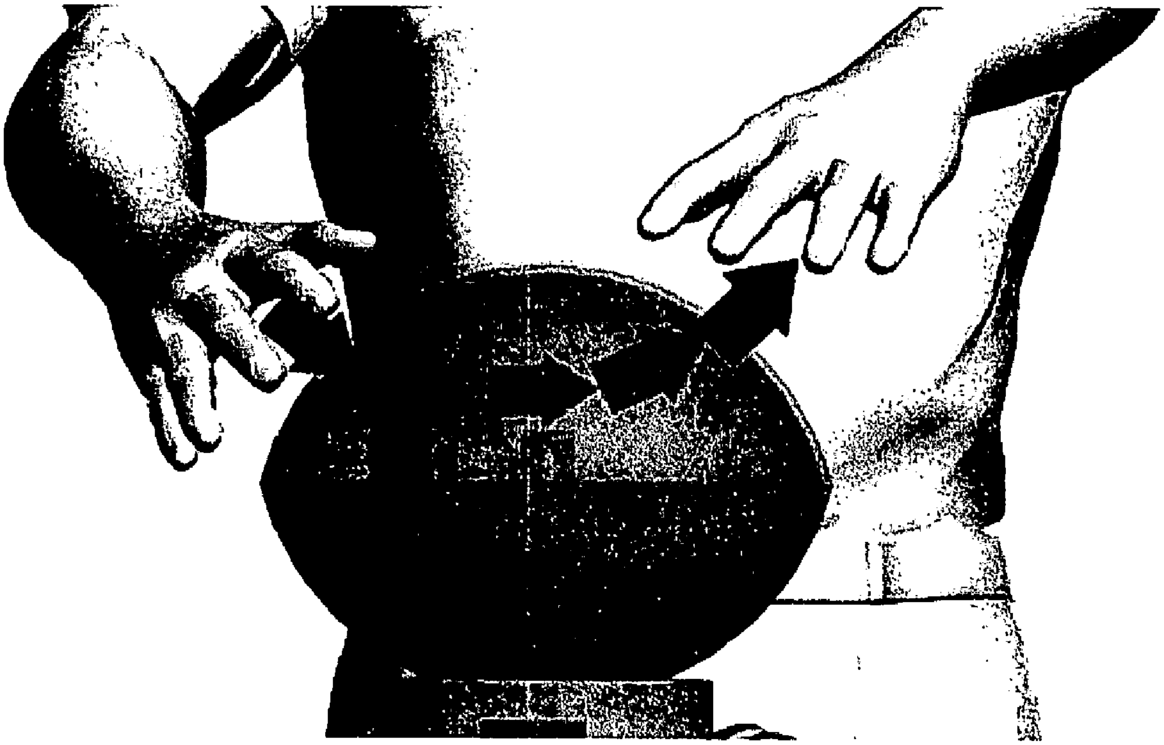

CHI ENERGY AMPLIFIER

BRIEF SUMMARY OF THE INVENTION

[0001] This invention is an energy amplifier that controls the direction of the second co-gravitational K field by means of a slot antenna located in a resonating sphere. The hyperspace mass flow rate entering into this dimension is determined by the frequency of the antenna. This energy flow is used to increase the energy of the human energy field known as Chi.

BACKGROUND OF THE INVENTION

[0002] Referring to FIG. **1**, a pendulum suspended by a string (**3**) is held by the string with the right hand (**1**) over the upright palm of the left hand (**2**). The pendulum swings in circles shown by the clockwise direction of the arrow (**4**). The pendulum swings at a constant frequency at about 1 to 2 Hz. What this means is that there is a second gravitational K field (**5**) that can cause a mass to rotate in circles.

[0003] Everyone is familiar with the linear gravitational g field which is defined as Newton's gravitational constant G times the mass of the earth divided by the square of the radius of the earth. This is the field that causes objects to accelerate radially toward the earth's surface. Newton's gravitational constant G is equal to the speed of light c squared divided by the linear mass Ω of the universe. The speed of light is

$$c = 299792458\ \frac{\text{meters}}{\text{second}}$$

$$\Omega = 1.346812891 \cdot 10^{27}\ \frac{\text{kilograms}}{\text{meter}}$$

$$G = \frac{c^2}{\Omega} = 6.673200002 \cdot 10^{-11}\ \frac{meter^3}{second^2\ \text{kilogram}}$$

The mass and radius of the earth are

[0004]

$$ME = 5.977 \cdot 10^{24}\ \text{kilograms}$$

$$RE = 6371.03\ 10^3\ \text{meters}$$

$$g = G\frac{ME}{RE^2} = 9.82\ \frac{\text{meters}}{second^2}$$

[0005] From electromagnetism, the electric E field is linear and the B magnetic field is circular or forms closed loops. The electric field starts and ends on electric charges. Since there are two electromagnetic fields, it makes sense that there are also two gravitational fields. A flow of electric current through a straight wire causes a circular magnetic B field to form around the wire. In a similar manner, a flow of mass through a channel causes a circular gravitational K field to form around the channel.

[0006] Referring to FIG. **2**, the right hand rule (**6**) shows that if the flow is along the direction of the thumb of the right hand, then the field curls around in the direction of the fingers. If an electric current is flowing through the wire (**7**) in the direction of arrow (**8**), then a counter-clockwise magnetic field circles the wire shown by arrow (**9**). In the bottom view, a mass flowing through a channel (**10**) in direction (**11**) generates a counter-clockwise gravitational K field shown by arrow (**12**). This field is the reason that the pendulum swings in circles. The gravitational K field only exists if there is a mass flow. In the wire there is also an electric field driving the electrons, so it must be that there is a g gravitational field driving the mass along the channel.

[0007] The correspondence between electromagnetic and gravitational constants must be the following:

Electric	Gravitational
Charge q	Mass m
Electric Field E	Gravitational Field g
Magnetic Field B	Co-gravitational Field K
Linear Electric Charge λ	Linear Mass Ω
Permittivity of Space ϵ_0	$-1/4\pi G$
Permeability of Space μ_0	$-4\pi G/c^2$
Convection Current Density J	Mass Current Density J

[0008] Maxwell's equation for the curl or circulation of B involves the current density J and an electric field E changing with time t

$$\nabla \times B = \mu_0 J + \frac{1}{c^2}\frac{\partial E}{\partial t}$$

The corresponding gravitational equation is therefore

[0009]

$$\nabla \times K = -\frac{4\pi G}{c^2}J + \frac{1}{c^2}\frac{\partial g}{\partial t}$$

[0010] If the g field is constant with time, then the circulation of the K field only depends on the mass density flow through the channel. Notice the minus sign in front of the first term which says that if energy is entering our dimension through a channel, then the K field is in the counterclockwise direction as indicated by the pendulum. As seen in FIG. **1**, the energy is leaving through the left hand because the rotation is clockwise. On the right hand, the pendulum rotates in the counterclockwise direction indicating that energy is coming from the hand. Thus there is a flow of energy between the hands known as the Chi energy. The purpose of this invention is to amplify this energy.

[0011] The g gravitational field is due to a negative space-time curvature created by the mass of the earth. In many physics books this is depicted as a bowl. So the idea was to acquire a ceramic kiln which could produce ceramic bowls using earthenware clay and a bowl mold. The curvature created by the bowl, if any, could be detected by the pendulum.

[0012] Referring to FIG. **3**, a pendulum (**13**) held in the right hand with the left hand near the side of the concave mold (**14**) starts to swing toward the right, away from the plaster mold as shown by arrow (**15**).

[0013] Referring to FIG. **4**, a pendulum (**16**) held in the right hand over a positive curvature dome (**17**) swings toward the top of the dome as shown by arrow (**18**).

Supposedly the K gravitational field rotates in circles, but these experiments showed that the pendulum movement was radial.

[0014] Referring to FIG. **5**, the pendulum (**19**) swings radially (**20**) toward the top of the dome (**23**) because the forces (**24**) around the dome counteract each other as shown by arrows (**21**, **22**). Thus there is no sideways pressure on the pendulum.

[0015] The next question was what is the value of the co-gravitational K constant? In electromagnetism the electric field is the speed of light times the magnetic field. So the g gravitational field should be the speed of light times the co-gravitational K field.

$$\begin{aligned} K &= \frac{g}{c} \\ &= \frac{1}{c} G \frac{MU}{RU^2} \\ &= \frac{1}{c} \frac{c^2}{\Omega} \frac{MU}{\frac{MU}{\Omega} RU} \\ &= \frac{c}{RU} \end{aligned}$$

$$\ln(K) = -40.34861143$$

[0016] where the radius RU of the universe is 10^{26} meters and c is the speed of light. The speed of light divided by the radius is angular frequency measured in Hz. Recall that the pendulum over the hand has a frequency f of 1 Hz which makes the angular frequency o equal to 2π radians per second since $\omega = 2\pi f$ So the question is whether or not this value of K means anything in terms of hyperspace physics?

[0017] The universe has a geometrical sub-manifold based on the tetrahedron. All the physics constants of the universe are determined by this tetrahedral geometry. As shown in diagram tet0565 stored in the Library of Congress, the electron and proton are one and the same particle. There is a continuous clockwise path from the electron wavelength, through electric charge to electron mass and returning out-of-dimension back along a counterclockwise path as the proton. Because the paths are in opposite directions, the proton has a positive charge and the electron has the opposite negative charge. Because this path crosses into our dimension from hyperspace, we see two different particles. Thus Nature only has one particle. The tetrahedron diagram also shows that our dimension is offset from the origin of the diagram by the Cabibbo angle which is found in particle physics. Our dimension is determined by the Planck mass and the Planck wavelength which are the bottom limits of our dimension, known as the Planck box. If the proton mass is centered on the Planck mass, a tangent line to the circle can only be made by using the down quark and up quark which comprise the proton (duu). And the tangent line is drawn at the Cabibbo angle. So the diagram verifies some important experimental physics data.

[0018] Referring to FIG. **6**, the tetrahedron diagram is drawn with a vertical axis (**35**) corresponding to the natural logarithm of mass. The horizontal axis (**36**) is the natural logarithm of wavelength. The inverted tetrahedrons (**25**, **26**) cross at the centerline (**28**) which is known through remote viewing as "the merging of two worlds." The centerline is the separation point between space and hyperspace. The circumscribing sphere (**27**) has a horizontal diameter (**29**) and a vertical diameter (**30**). The electron wavelength (**34**) reflects off the sphere and returns as the electron mass (**33**).

[0019] Because the K co-gravitational field is related to circular or spinning motion, as well as vorticity per area, it might be imagined that it has something to do with the spinning electron. The energy E of the electron is equal to the mass m_e of the electron times the speed of light c squared.

$$\begin{aligned} m_e &= 9.1093897 \cdot 10^{-31} \text{ kilogram} \\ c &= 299792458 \cdot \frac{\text{meter}}{\text{second}} \\ E &= m_e c^2 \\ \ln(E) &= -30.13363019 \end{aligned}$$

[0020] The tips of the inverted tetrahedrons pass through the base constant, vertical line (**37**), which is equal to Planck's constant h divided by the speed of light.

$$h = 2.210260685 \cdot 10^{-42} \text{ kilogram meter}$$

$$\text{base} = \ln\left(\frac{h}{c}\right) = -95.91546344$$

[0021] On the tetrahedron diagram, the co-gravitational circle K (**31**) is centered on the base (**37**) at the electron energy (**32**) shown by the small π circle at the intersection (**38**). As can be seen, the K circle determines the electron mass (**33**) at the horizontal axis (**36**) as shown by intersection (**39**). Because the electron is the proton, it also sets the constants of the proton such as wavelength, charge and mass. Thus the K co-gravitational field determines our elementary particles. Because these particle paths go in and out of dimension, it means that hyperspace exists. And because the K field is circular, depending on the direction of the energy flow, another purpose of this invention is to control the direction of the field.

SUMMARY OF THE INVENTION

[0022] Referring to FIG. **7**, the Chi energy amplifier consists of two ceramic domes (**40**, **41**) resting one on the other such as to form a hollow internal clam-like structure. The domes are mounted on a cylindrical ceramic base (**42**) containing a reinforced passageway for the coaxial cable and BNC connector (**43**). The BNC connector plugs into the frequency generator (not shown) which has a frequency range of 0.4 Hz to 5 MHz.

[0023] Referring to FIG. **8** with the upper dome removed, the coaxial cable (**44**) extends vertically through the base into the lower dome. The cable is soldered to a slot antenna (**45**). The center conductor of the cable is soldered to the left side of the slot and the ground shielding is soldered to the right side of the slot. Thus there is a voltage difference on the two sides which produces an oscillating electric field across the slot which radiates electromagnetic energy into the hollow dome.

[0024] A close-up of the slot antenna is shown in FIG. **9**. The antenna is a 0.050" thick copper sheet (**46**) with a width of ¾ wavelength by a height of ½ wavelength. A slot (**47**) of ½ wavelength is electric-discharge machined into the copper sheet (**47**). The center conductor (**48**) from the coaxial cable is soldered to the top of the slot. The ground wire (**49**) is soldered to the bottom of the slot. This configuration produces a voltage difference between the inner top and bottom surfaces of the slot. The frequency generator produces an oscillating voltage across the slot which generates an electric field that radiates from the antenna into the ceramic cavity.

[0025] The wavelength of the antenna and the dimensions of the ceramic cavity have to be tuned to the geometry of our dimension. Newton's gravitational constant G is equal to

$$\ln(G) = -23.4303342$$

$$f = e^{23.4303432} * \text{Hz} = 15\ Ghz$$

$$\lambda = \frac{c}{f} = .020\ \text{meter} = .787\ \text{inch}$$

$$\text{slotwidth} = \frac{\lambda}{2} = .393\ \text{inch}$$

$$\text{antennawidth} = \frac{3}{4}\lambda = .590\ \text{inch}$$

The frequency of the antenna is the inverse of the gravitational constant. The reason for this is that when a circle of radius natural logarithm G is centered on the proton on the tetrahedron diagram, the circle intersects the corner of the Planck box which bounds our dimension between space and hyperspace. It is the low-density hyperspace energy that the amplifier will bring into this dimension.

[0026] The size of the ceramic cavity based on the angular frequency ω=2πf is

$$\frac{\frac{\omega^2}{c^2} + \frac{1}{ab}}{\frac{\omega^2}{c^2} + ab\left(\frac{\omega^2}{c^2} - \frac{1}{a^2}\right)\left(\frac{\omega^2}{c^2} - \frac{1}{b^2}\right)} = \frac{\tan\left(\frac{\omega}{c}(b-a)\right)}{\omega} \frac{c}{b-a}$$

where b is inside radius of the cavity, a is the radius of an object at the center of the cavity and c is the speed of light. Using the object radius as a=0.020 meter to account for the antenna, the inside radius of the cavity is b=4.25 inches which is the size of the 8.5 inch diameter ceramic dome using ¼ inch thick clay. The frequency is then equal to

$$f = \ln\left(\frac{\omega}{2\pi}\right) = 23.45$$

which is close to the inverse of the gravitational constant G of −23.43. Thus the amplifier is tuned to the gravitational constant of the universe both in size of the cavity and the frequency of the slot antenna. Using a SMD surface mount capacitor and inductor in the picofarad and nanohenry range, it is possible to get up to this high frequency, but it was found that it was not necessary because the amplifier works at lower frequencies that are pulsed.

[0027] Referring to FIG. **10**, the swinging movement of the pendulum measured in inches at the side of the dome is plotted against a range of frequencies from 20 kHz to 120 kHz. At 20 kHz, the pendulum swings ¾ inch away from the dome. Then at 40 kHz, the pendulum swings ¾ inch toward the dome. This pattern is repeated until 120 kHz. What this means is that the amplifier can control the direction of the co-gravitational field using frequency. The tetrahedron diagram shows that this phenomenon is related to the two axes of the circumscribing sphere of the tetrahedron.

[0028] Referring to FIG. **11**, the swinging of the pendulum is plotted for a range of frequencies from 1 MHz to 5 MHz. At a frequency of 2.78 MHz the pendulum was swinging strongly from the vertical to the horizontal, a distance of six inches. On the tetrahedron diagram, the distance from our base constant to the centerline between inverted tetrahedrons is equal to 14.838168. This corresponds to a frequency of

$$f = e^{14.838168} * \text{Hz} = 2.780574169\ \text{MHz}$$

which is close to the experimental value.

[0029] Referring to FIG. **12**, the Chi energy flowing between the hands, shown by the arrows, is amplified by placing the hands across the dome. This energy from the right hand mixes with the hyperspace energy entering the dome from hyperspace. The combined energy is then absorbed in the left hand vortex. The effects of this amplified energy are simply amazing and have to be experienced to appreciate what it means.

A BRIEF DESCRIPTION OF THE DRAWINGS

[0030] FIG. **1**. Perspective view of pendulum movement over left hand vortex.

[0031] FIG. **2**. Perspective view of right hand rule of physics showing similarity between electromagnetism and gravity.

[0032] FIG. **3**. Perspective view of plaster mold cavity showing negative curvature.

[0033] FIG. **4**. Perspective view of dome showing positive curvature.

[0034] FIG. **5**. Perspective view of dome showing radial pendulum movement.

[0035] FIG. **6**. Graph of co-gravitational K field on tetrahedron diagram.

[0036] FIG. **7**. Perspective view of Chi Energy Amplifier with BNC coaxial cable.

[0037] FIG. **8**. Perspective view of slot antenna in dome.

[0038] FIG. **9**. Perspective view of slot antenna showing dimensions.

[0039] FIG. **10**. Graph of pendulum movement versus frequency of antenna, 20 KHz.

[0040] FIG. **11**. Graph of pendulum movement versus frequency from 1 MHz to 5 MHz.

[0041] FIG. **12**. Perspective view of amplifying Chi energy.

DETAILED DESCRIPTION OF THE INVENTION

[0042] The dome of the amplifier is made of red earthenware clay from Minnesota having a cone 06 kiln firing temperature of 1828° F. over a period of 7.5 hours. The clay is placed between two hardwood ¼ inch thick slats. It is then rolled flat with a rolling pin. The sheet of clay is cut in half and one half is placed in a 9-inch diameter bowl mold using a soft sponge to push it into position. The second half is then added to the first with the seam between worked flat with a metal kidney and elephant ear sponge. The top of the mold is scrapped flat to create the bowl shape. The moist clay is left to dry for one day after which it falls out of the mold due to shrinkage. After completely drying, the dome is then placed in a kiln which runs the computer-controlled cone 06 temperature firing profile. A preheat at 180° F. for one hour is required to make sure the dome is completely bone dry. It takes 7 to 8 hours to fire the dome with another 12 hours to cool down naturally.

[0043] The base is also made of ¼ inch clay. A circular shape is cut out and fitted with a ribbon of extruded clay from a clay gun to make the cylindrical wall. Using a hole cutter made of hollow tubing, a hole is made in the side of the base for the coaxial cable. The second lower dome has a hole cut in the bottom when the clay is leather hard.

[0044] A 0.050" thick copper sheet is EDM machined to cut a half wavelength slot in the copper. The inner conductor of the coaxial cable is soldered using silver epoxy to the one side of the slot. The ground shielding of the cable is silver epoxied to the other side of the slot. The other end of the coaxial cable has a BNC connector which plugs into the frequency generator.

[0045] The pendulum is made out of soft brass rod and machined on a lathe into a plumb bob shape. A hole is drilled in the stem to hold the 6" long string.

I claim:

1. A Chi energy amplifier comprising:

a. a hollow ceramic dome in the shape of a clam shell with a hole in the lower half that provides access to a coaxial cable;

b. a cylindrical ceramic base on which the dome is mounted with holes that provide access for a coaxial cable into item (a);

c. a slot antenna soldered to the coaxial cable, located midway inside item (a), such that the inner conductor of the cable is soldered to one side of the slot and the ground shielding is soldered to the other side of the slot;

d. a frequency generator connected to the other end of the coaxial cable by means of a BNC connector for the purpose of providing an oscillating voltage to item (c);

e. an electromagnetic wave generated by items (c, d) that resonates inside item (a);

f. a co-gravitational K field generated by items (a, b, c, d, e) whose direction can be controlled by the frequency of item (d); and

g. a flow of hyperspace energy into or out of the dome due to item (f).

* * * * *

455-619 AU 233 EX
FIP8106 OR 685,958

55

No. 685,958. Patented Nov. 5, 1901.

N. TESLA.

METHOD OF UTILIZING RADIANT ENERGY.

(Application filed Mar. 21, 1901.)

(No Model.)

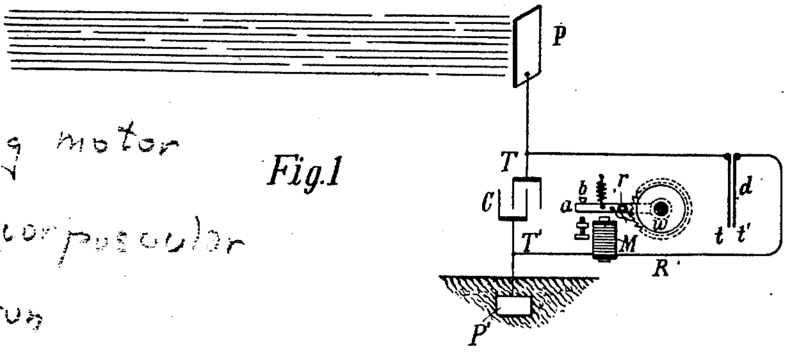

Fig.1

Electric steping motor
energized by corpuscular
energy from sun

No refs.

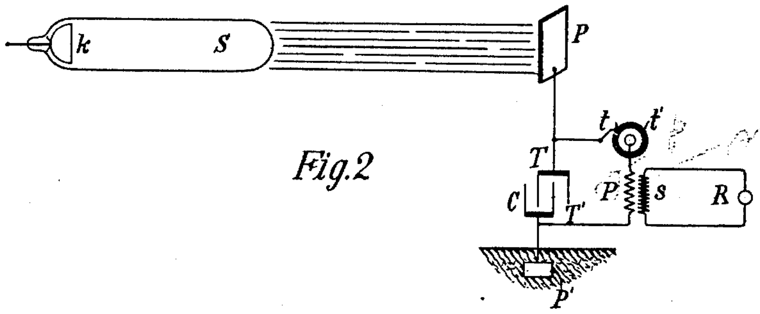

Fig.2

Witnesses:
Raphaël Netter
M. Lawson Dyer

Nikola Tesla, Inventor
by Kerr, Page & Cooper Attys

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

UNITED STATES PATENT OFFICE.

NIKOLA TESLA, OF NEW YORK, N. Y.

METHOD OF UTILIZING RADIANT ENERGY.

SPECIFICATION forming part of Letters Patent No. 685,958, dated November 5, 1901.

Application filed March 21, 1901. Serial No. 52,154. (No model.)

To all whom it may concern:

Be it known that I, NIKOLA TESLA, a citizen of the United States, residing at the borough of Manhattan, in the city, county, and State of New York, have invented certain new and useful Improvements in Methods of Utilizing Radiant Energy, of which the following is a specification, reference being had to the drawings accompanying and forming a part of the same.

It is well known that certain radiations—such as those of ultra-violet light, cathodic, Roentgen rays, or the like—possess the property of charging and discharging conductors of electricity, the discharge being particularly noticeable when the conductor upon which the rays impinge is negatively electrified. These radiations are generally considered to be ether vibrations of extremely small wave lengths, and in explanation of the phenomena noted it has been assumed by some authorities that they ionize or render conducting the atmosphere through which they are propagated. My own experiments and observations, however, lead me to conclusions more in accord with the theory heretofore advanced by me that sources of such radiant energy throw off with great velocity minute particles of matter which are strongly electrified, and therefore capable of charging an electrical conductor, or even if not so may at any rate discharge an electrified conductor either by carrying off bodily its charge or otherwise.

My present application is based upon a discovery which I have made that when rays or radiations of the above kind are permitted to fall upon an insulated conducting body connected to one of the terminals of a condenser, while the other terminal of the same is made by independent means to receive or to carry away electricity, a current flows into the condenser so long as the insulated body is exposed to the rays, and under the conditions hereinafter specified an indefinite accumulation of electrical energy in the condenser takes place. This energy after a suitable time interval, during which the rays are allowed to act, may manifest itself in a powerful discharge, which may be utilized for the operation or control of mechanical or electrical devices or rendered useful in many other ways.

In applying my discovery I provide a condenser, preferably of considerable electrostatic capacity, and connect one of its terminals to an insulated metal plate or other conducting body exposed to the rays or streams of radiant matter. It is very important, particularly in view of the fact that electrical energy is generally supplied at a very slow rate to the condenser, to construct the same with the greatest care. I use by preference the best quality of mica as dielectric, taking every possible precaution in insulating the armatures, so that the instrument may withstand great electrical pressures without leaking and may leave no perceptible electrification when discharging instantaneously. In practice I have found that the best results are obtained with condensers treated in the manner described in a patent granted to me February 23, 1897, No. 577,671. Obviously the above precautions should be the more rigorously observed the slower the rate of charging and the smaller the time interval during which the energy is allowed to accumulate in the condenser. The insulated plate or conducting body should present as large a surface as practicable to the rays or streams of matter, I having ascertained that the amount of energy conveyed to it per unit of time is under otherwise identical conditions proportionate to the area exposed, or nearly so. Furthermore, the surface should be clean and preferably highly polished or amalgamated. The second terminal or armature of the condenser may be connected to one of the poles of a battery or other source of electricity or to any conducting body or object whatever of such properties or so conditioned that by its means electricity of the required sign will be supplied to the terminal. A simple way of supplying positive or negative electricity to the terminal is to connect the same either to an insulated conductor, supported at some height in the atmosphere, or to a grounded conductor, the former, as is well known, furnishing positive and the latter negative electricity. As the rays or supposed streams of matter generally convey a positive charge to the first condenser-terminal, which is connect-

RADIANT ENERGY.

EXAMINER

ed to the plate or conductor above mentioned, I usually connect the second terminal of the condenser to the ground, this being the most convenient way of obtaining negative electricity, dispensing with the necessity of providing an artificial source. In order to utilize for any useful purpose the energy accumulated in the condenser, I furthermore connect to the terminals of the same a circuit including an instrument or apparatus which it is desired to operate and another instrument or device for alternately closing and opening the circuit. This latter may be any form of circuit-controller, with fixed or movable parts or electrodes, which may be actuated either by the stored energy or by independent means.

The rays or radiations which are to be utilized for the operation of the apparatus above described in general terms may be derived from a natural source, as the sun, or may be artificially produced by such means, for example, as an arc-lamp, a Roentgen tube, and the like, and they may be employed for a great variety of useful purposes.

My discovery will be more fully understood from the following detailed description and annexed drawings, to which reference is now made, and in which—

Figure 1 is a diagram showing typical forms of the devices or elements as arranged and connected in applying the method for the operation of a mechanical contrivance or instrument solely by the energy stored; and Fig. 2 is a diagrammatical representation of a modified arrangement suitable for special purposes, with a circuit-controller actuated by independent means.

Referring to Fig. 1, C is the condenser, P the insulated plate or conducting body, which is exposed to the rays, and P′ another plate or conductor, all being joined in series, as shown. The terminals T T′ of the condenser are also connected to a circuit including a receiver R, which is to be operated, and a circuit-controlling device *d*, which in this case is composed of two very thin conducting-plates *t t′*, placed in close proximity and very mobile, either by reason of extreme flexibility or owing to the charater of their support. To improve their action, they should be inclosed in a receptacle from which the air may be exhausted. The receiver R is shown as consisting of an electromagnet M, a movable armature *a*, a retractile spring *b*, and a ratchet-wheel *w*, provided with a spring-pawl *r*, which is pivoted to armature *a*, as illustrated. The apparatus being arranged as shown, it will be found that when the radiations of the sun or of any other source capable of producing the effects before described fall upon the plate P an accumulation of electrical energy in the condenser C will result. This phenomenon, I believe, is best explained as follows: The sun as well as other sources of radiant energy throw off minute particles of matter positively electrified, which, impinging upon the plate P, communicate an electrical charge to the same. The opposite terminal of the condenser being connected to the ground, which may be considered as a vast reservoir of negative electricity, a feeble current flows continuously into the condenser, and inasmuch as these supposed particles are of an inconceivably small radius or curvature, and consequently charged to a relatively very high potential, this charging of the condenser may continue, as I have found in practice, almost indefinitely, even to the point of rupturing the dielectric. Obviously whatever circuit-controller be employed it should operate to close the circuit in which it is included when the potential in the condenser has reached the desired magnitude. Thus in Fig. 2 when the electrical pressure at the terminals T T′ rises to a certain predetermined value the plates *t t′*, attracting each other, close the circuit connected to the terminals. This permits a flow of current which energizes the magnet M, causing it to draw down the armature *a* and impart a partial rotation to the ratchet-wheel *w*. As the current ceases the armature is retracted by the spring *b* without, however, moving the wheel *w*. With the stoppage of the current the plates *t t′* cease to be attracted and separate, thus restoring the circuit to its original condition.

Many useful applications of this method of utilizing the radiations emanating from the sun or other source and many ways of carrying out the same will at once suggest themselves from the above description. By way of illustration a modified arrangement is shown in Fig. 2, in which the source S of radiant energy is a special form of Roentgen tube devised by me having but one terminal *k*, generally of aluminium, in the form of half a sphere with a plain polished surface on the front side, from which the streams are thrown off. It may be excited by attaching it to one of the terminals of any generator of sufficiently-high electromotive force; but whatever apparatus be used it is important that the tube be exhausted to a high degree, as otherwise it might prove entirely ineffective. The working or discharge circuit connected to the terminals T T′ of the condenser includes in this case the primary *p* of a transformer and a circuit-controller comprising a fixed terminal or brush *t* and a movable terminal *t′* in the shape of a wheel with conducting and insulating segments which may be rotated at an arbitrary speed by any suitable means. In inductive relation to the primary wire or coil *p* is a secondary *s*, usually of a much greater number of turns, to the ends of which is connected a receiver R. The terminals of the condenser being connected as indicated, one to an insulated plate P and the other to a grounded plate P′, when the tube S is excited rays or streams of matter are emitted from the same, which convey a positive charge to the plate P and condenser-terminal T, while terminal T′ is continuously receiving negative electricity from the plate

P'. This, as before explained, results in an accumulation of electrical energy in the condenser, which goes on as long as the circuit including the primary p is interrupted. Whenever the circuit is closed, owing to the rotation of the terminal t', the stored energy is discharged through the primary p, this giving rise in the secondary s to induced currents which operate the receiver R.

It is clear from what has been stated above that if the terminal T' is connected to a plate supplying positive instead of negative electricity the rays should convey negative electricity to plate P. The source S may be any form of Roentgen or Lenard tube; but it is obvious from the theory of action that in order to be very effective the electrical impulses exciting it should be wholly or at least preponderatingly of one sign. If ordinary symmetrical alternating currents are employed, provision should be made for allowing the rays to fall upon the plate P only during those periods when they are productive of the desired result. Evidently if the radiations of the source be stopped or intercepted or their intensity varied in any manner, as by periodically interrupting or rythmically varying the current exciting the source, there will be corresponding changes in the action upon the receiver R, and thus signals may be transmitted and many other useful effects produced. Furthermore, it will be understood that any form of circuit-closer which will respond to or be set in operation when a predetermined amount of energy is stored in the condenser may be used in lieu of the device specifically described with reference to Fig. 1, and also that the special details of construction and arrangement of the several parts of the apparatus may be very greatly varied without departure from the invention.

Having described my invention, what I claim is—

1. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other armature by independent means, and discharging the condenser through a suitable receiver, as set forth.

2. The method of utilizing radiant energy, which consists in simultaneously charging a condenser by means of rays or radiations and an independent source of electrical energy, and discharging the condenser through a suitable receiver, as set forth.

3. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other by independent means, controlling the action or effect of said rays or radiations and discharging the condenser through a suitable receiver, as set forth.

4. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations and the other by independent means, varying the intensity of the said rays or radiations and periodically discharging the condenser through a suitable receiver, as set forth.

5. The method of utilizing radiant energy, which consists in directing upon an elevated conductor, connected to one of the armatures of a condenser, rays or radiations capable of positively electrifying the same, carrying off electricity from the other armature by connecting the same with the ground, and discharging the accumulated energy through a suitable receiver, as set forth.

6. The method of utilizing radiant energy, which consists in charging one of the armatures of a condenser by rays or radiations, and the other by independent means, and effecting by the automatic discharge of the accumulated energy the operation or control of a suitable receiver, as set forth.

NIKOLA TESLA.

Witnesses:
M. LAWSON DYER,
RICHARD DONOVAN.

New Energy Technologies

Published by Faraday Laboratories Ltd. Issue #3 (6) 2002

Research on advanced propulsion systems and new energy sources

Time Control experiments

Also in this issue:

* New experiments: magnet monopole beam cures cancer
* Electrogravity experiments
* High efficiency ceramic heater
* New aether experiments
* Instantaneous telecommunication

Editorial

Dr. Chernobrov and NET editor Olga Leontyeva.

The head of the Moscow research group "Kosmopoisk" Dr. Vadim A. Chernobrov (our readers certainly remember his articles on experiments with time) has recently visited the editorial office of New Energy Technologies in Saint-Petersburg, Russia. His visit was caused by discussion on the Time Machine project (investigation of active properties of time), started under agreement with Faraday Labs Ltd. Our magazine will inform readers on the topic in later issues.

NEW ENERGY TECHNOLOGIES
New Book

Dear Reader!
All articles of the previous 6 NET issues will soon be published as a new book in English (400 pages, more than 100 photos, figures and drawings, format A4).
You can order by e-mail: book2001@faraday.ru
or fax 7-812-3806564
Price with airmail delivery is 83 USD.

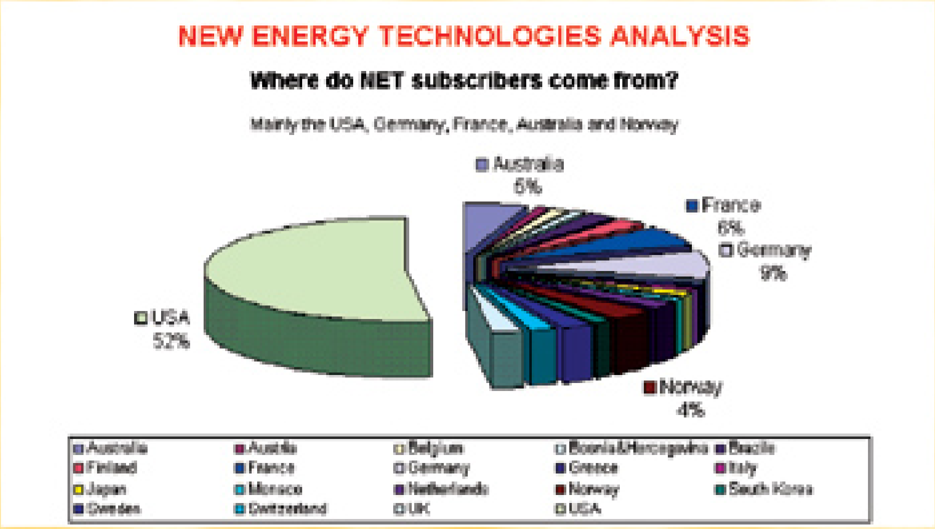

Publisher Faraday Lab Ltd

Editor-in-Chief Alexander V. Frolov
Editor Olga O. Leontyeva
Designer Ekaterina F. Yakovleva
Translator Antonina V. Nikolaeva

Scientific Advisory Board
Kyrill P. Butusov
Ivan M. Shakhparonov

New Energy Technologies Magazine
is published by Faraday Lab Ltd
six times a year

Correspondence Address:
POBox 37, St.Petersburg,
193024, Russia
Phone/fax: 7-812-380-6564
net@faraday.ru
www.faraday.ru

Subscription: $90 any foreign
Air mail delivery included

PRINTED IN RUSSIA
Copyright © 2002 by Faraday Lab Ltd.
All rights reserved.

LUTEC 1000

News from Australia

Magnet motor-generator

www.lutec.com.au

John Christie (left) is the co-founder, CEO and chairman of Lutec (Aust.) Pty Ltd.

Ludwig (Lou) Brits (right, above) is the co-founder, managing director and head of development of Lutec (Aust.) Pty Ltd. He is Chairman and director of TTI Pty Ltd. TTI Pty Ltd is a state government recognised R.A.P. (remote area power) system provider.

"There are a couple of other major factors that we won't go into here, suffice it to say that our current prototype demonstrates 1500% more " out" than "in"... "

"The Lutec 1000 will be sold throughout Australia within three years. The majority of buyers are expected to be the home owners, although there is strong support for our product throughout the entire strata of society and business. Market leadership should be maintained for the next five years. The product is sold through commission only sales agents and our staff at the plant, mainly to anyone who needs or uses electricity."

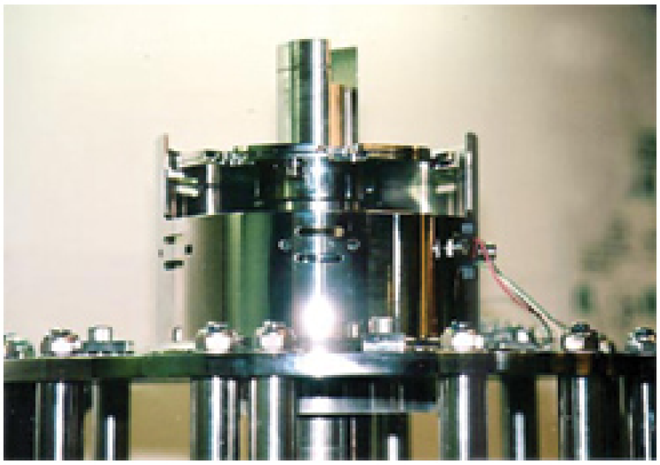

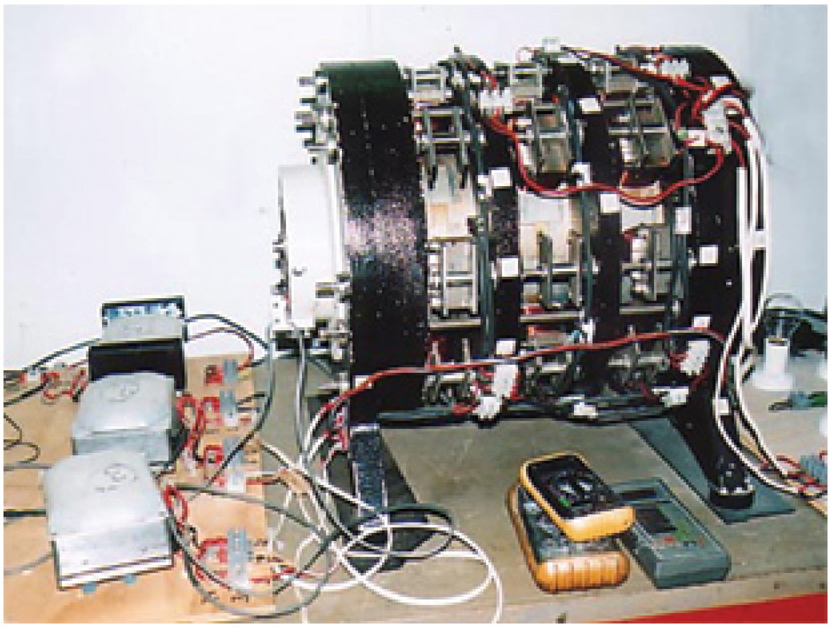

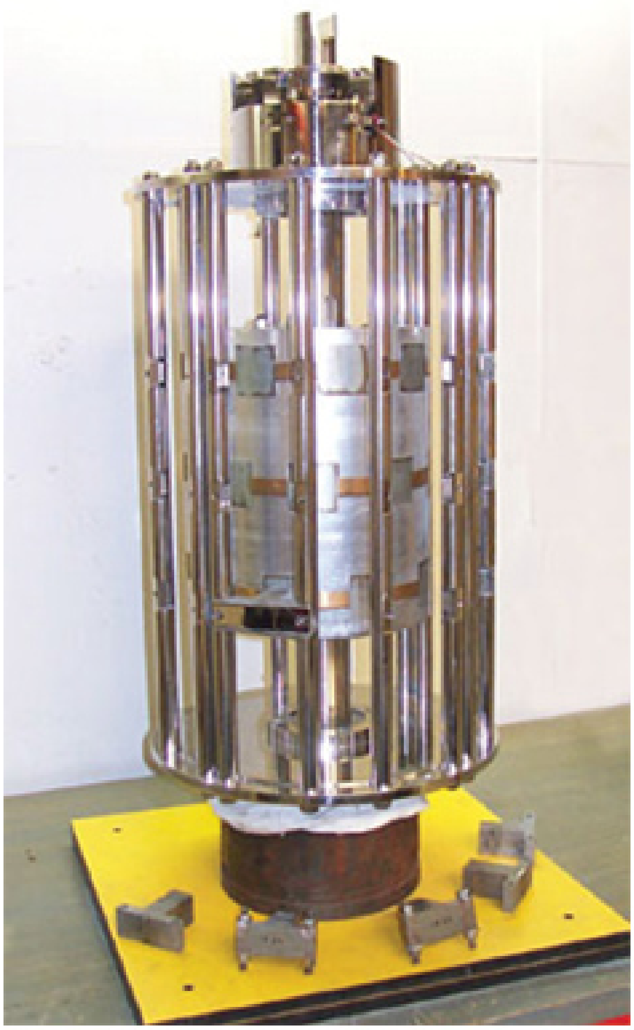

Electrogravity

by Tim Ventura
tventura@attbi.com

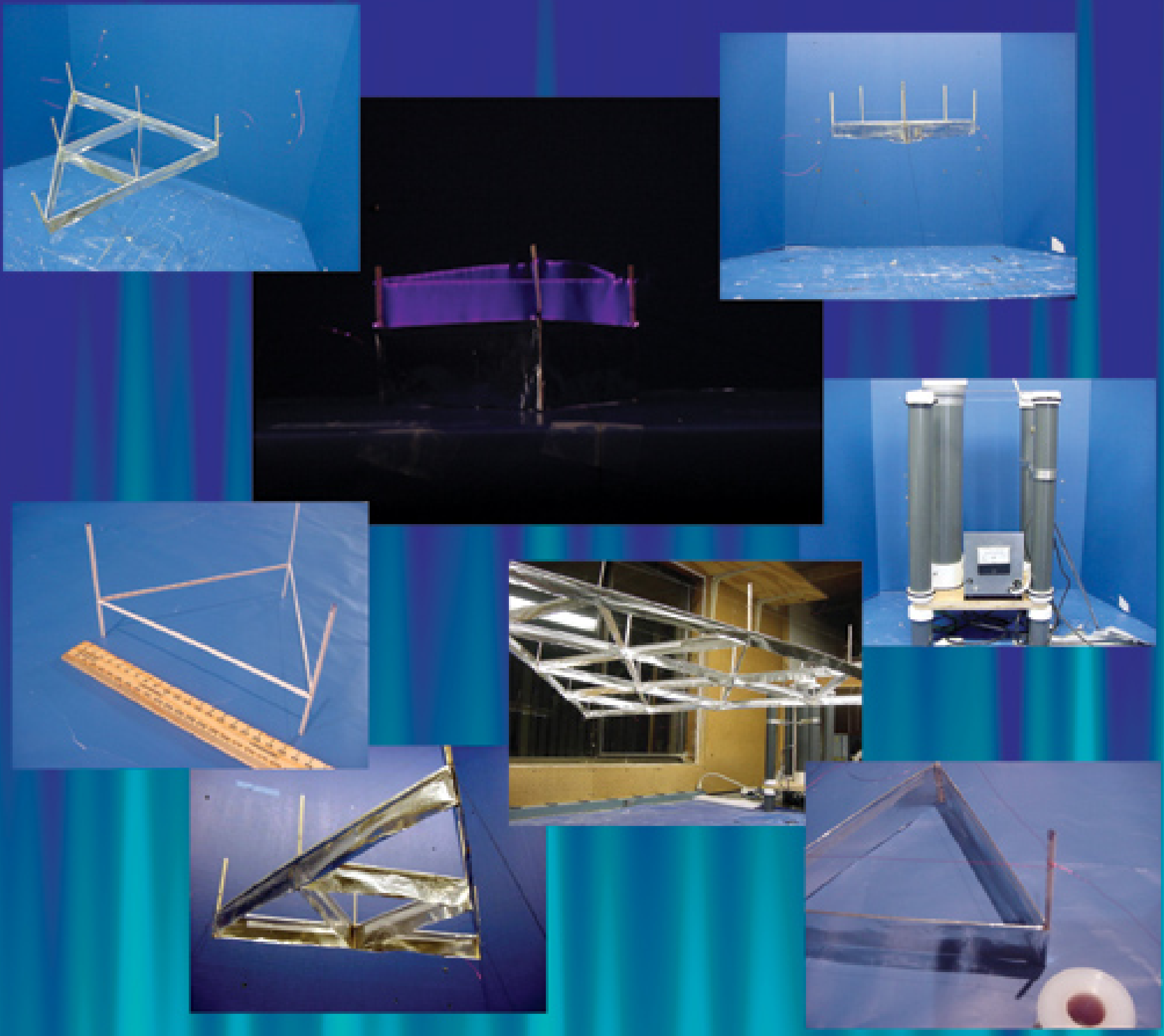

In this issue a report by Tim Ventura, USA on experimenting with asymmetrical capacitors. www.americanantigravity.com

NEW ENERGY TECHNOLOGIES #6

1. Large-Scale Shakharov condition, **David Noever and Christopher Bremner**
2. Matter as a resonance longitudinal wave process, **Alexander V. Frolov**
3. Physical Principles of The Time Machine, **Alexander V. Frolov**
4. Time Machine Project by **Alexander V. Frolov**
5. Kozyrev-Dirak radiation, **Ivan M. Shakhparonov**
6. The Electrical Vortex Non-Solenoidal Fields**, S. Alemanov**
7. Physical Mechanism of Nuclear Reactions at Low Energies, **V.Oleinik, Yu. Arepjev**
8. The Evolution of Lifter Technology, **T. Ventura**
9. Reality and consciousness in education and activity, **A.Smirnov**
10. Old new energy, **Y. Andreev, A. Smirnov**
11. On the influence of time on matter, **A. Belyaeva**
12. Life without diseases and ageing-preventive electrical bio-heater features, **A. Belyaeva**
13. Technical report, on Belyaeva's high efficient ceramic heater, **Sh. Mavlyandekov**
14. Fundamental properties of aether, **A. Mishin**
15. Effect of Magnetic Blow Wave Field on Wine Systems, **I. Shakhparanov and others**
16. Nikola Tesla and Instantaneous Electric Communication, **V. Korobeynikov**
17. The Unitied Gravitation theory, **I. Kuldoshin**
18. New Sources of Energy from the Point of View of Unitary Quantum Theory, **L.G. Sapogin, Yu.A. Ryabov, V.V. Graboshnikov**
19. Antigravitation Force and antigravitation of matter. Methods of its creation, **A. K. Gaponov**
20. The capacitor, which has energy of atomic bomb (Review of A. Gaponov's research)

consequence of many commonly accepted concepts and dogmas of the modern "scientific perspective of natural phenomena". This crisis situation in modern physics is a direct consequence of many conservative scientific viewpoints, unfortunately supported and protected by modern official academic science. The evolution of our consciousness has been influenced by many undoubtedly well known experts and has been evolving for a long time in the environment of specific scientific vacuum and requires immediate revival. Even methods used for dissemination of new knowledge should be improved, if one actually wishes to accelerate the progress of Humankind.

The perspective for practical applications of new previously unknown scientific phenomena and effects looks very attractive, and they may be achieved by cooperative efforts of the human intellect. New breakthrough technologies of the 21st Century will require serious changes of many commonly accepted concepts and dogmas in fundamental physics. This process of progressive development cannot be stopped.

REFERENCES

1. Richard P. Feynman, Robert B. Leighton, Matthew Sands. The Feynman Lectures on Physics, Addison-Wesley, 1964, Vol. 2, Ch. 1. Paragraph 6 "Electromagnetism in Science and Technology" (the very end of paragraph)
2. J. Maxwell, Selected Works on the Electromagnetic Field Theory, Gostekhizdat, Moscow (1954).
3. G. V. Nikolaev, Non-contradictory Electrodynamics. Theories, Experiments, and Paradoxes, Publishing House of the Tomsk State University, Tomsk (1997).
4. A. S. Kompaniets, in: Theoretical Physics, State Technical and Theoretical Press, Moscow (1957), pp. 126-128.
5. R. T. Sigalov, T. I. Shapovalova, Kh. Kh. Karimov, and N. I. Samsonov, New Research of Forces of the Magnetic Field, FAN Press of the Uzbekskaia SSR, Tashkent (1975).
6. Ya. I. Frenkel, Electrodynamics. Vol. 1, United Scientific and Technical Presses, Leningrad/Moscow (1934).
7. G. V. Nikolaev and B. V. Okulov, Inertial Properties of Electrons, deposited at VINITI, No. 4399-77, Moscow (1978).
8. Observations of the Aharanov-Bohm Effect, Nature, No. 7, 106 (1983).
9. G. V. Nikolaev, Scientific Vacuum. Crisis in Basic Physics. Is There Any Way Out?! Publishing House Kursiv, Tomsk (1999).

Large-Scale Sakharov Condition

David Noever and Christopher Bremner

NASA Marshall Space Flight Center,
Space Sciences Laboratory
Mail Code: ES76, Huntsville AL 35812

Editor's note: This article was presented by the autors for publication in New Energy Technologies. For the first time it was published in 1999 by the American Institute of Aeronautics and Astronautics, Inc. All copyrights belong to the authors.

Abstract

Recent far reaching theoretical results have used the quantum vacuum noise as a fundamental electromagnetic radiation field to derive a frequency (ω) dependent version of Newton's gravitational coupling term, $G(\omega)$. This paper reconciles the cut-off frequency with the observed cosmological constant, and then briefly puts forward a realizable laboratory test case in the 10 - 100 MHz frequency range. One analogy is drawn between the classical vacuum energy experiments with attraction between two closely spaced plates (Casimir cavity) and the arbitrarily dense material boundaries possible in Bose condensates, such as irradiation at MHz frequencies of superfluid helium or superconductors.

Theoretical Background

Zel'dovich [1] first suggested that gravitational interactions could lead to a small disturbance in the (non zero) quantum fluctuations of the vacuum and thus give rise to a finite value of Einstein's cosmological constant in agreement with astrophysical data. Using dimensional analysis and the suggestion by Zel'dovich, Sakharov [2] derived a value for Newton's gravitational constant, G, in only one free parameter, frequency, ω:

$$G \sim c^5 \Big/ h\int \omega d\omega \sim 1 \Big/ \int \omega d\omega$$

where c is the speed of light and h is the Planck constant. The free parameter in frequency when integrated over all values from zero to high frequencies must contain the usual integration cutoff value (Planck frequency on observable electromagnetic phenomenon).

Puthoff [3] and others [4 5] have extended Sakharov's condition in a relativistically consistent model to determine constants of proportionality. His model derives an acceleration term in first order expansion (in flat space time), then equates inertial and gravitational mass (by the equivalence principle) to make contact with the gravitational constant, G, directly as:

$$G = \left(\pi c^5 / h\omega_c^2\right) \sim 1/\int \omega d\omega$$

which is the Sakharov condition [2,3]. This paper revisits the meaning of the cutoff frequency, ω_c, for radiation interactions, of which the quantum vacuum [6-10] and

Planck frequency are only the leading terms, and for which linear combinations of forces can introduce other plausible frequencies. One purpose of this reexamination is whether the resulting gravitational coupling constant, G, can be reconciled with the anticipated energy density of the universe [11] without resorting to extreme space time curvature and thus yield enough critical density to contain the expansion of the universe. Finally we particularize the case to the high-density fluctuations possible in Bose condensates [12], a potential experimental test case for how the effects of vacuum noise might manifest observably.

One far-reaching consequence of the vacuum energy model is the attractive force of gravity becomes reducible to the radiative interaction between oscillating charges, e.g. the zero point field (ZPF) applied to subatomic charges. Mass and inertia arise from the fundamentally electromagnetic ZPF oscillations.

This random background gives the usual quantum mechanical energy spectrum from particle field effects:

$$\rho(\omega)d\omega \sim \omega^3 d\omega$$

a very important dimensional relationship, since the third power in frequency avoids anomalous Doppler shifts from velocity boosts, or stated alternatively is the correct spectra for a Lorentzian (non accelerated) invariant radiation field [13].
More specifically, the energy spectrum [3] can be written as:

$$\rho(\omega)d\omega = [\omega^2 / \pi^2 c^3][h\omega / 2]d\omega =$$
$$= h\omega^3 / 2\pi^2 c^3 d\omega \sim \omega^3 d\omega$$

which is an expression in the first parenthesis of the density of the normal modes and in the second parenthesis of the average energy per mode. When this energy density is integrated over all frequencies, the ω^3 divergence produces well known infinities in the integration limit of high frequencies, thus an assumed cutoff frequency (appropriate to experimental observation limits at the Planck frequency), is usually introduced:

$$\omega_\rho = \left(c^5 / hG\right)^{1/2}$$

For mass, m, moving in an accelerated reference frame $g = -a = Gm/r^2$, the resulting energy spectrum includes a gravitational spectral shift [3],

$$\Delta\rho'(\omega)d\omega = h\omega / 2\pi^2 c^5 [Gm / r^2]^2 d\omega \sim 1 / r^4 d\omega$$

a kind of short range ($1/r^4$) gravitational energy shift, but electromagnetic in origin when zero point fluctuations are included. (N.B. To account for equal gravitational mass effects in neutrons and protons, the ZPF oscillations must involve subatomic charges, or 'parton' effects. The assumption derives from high frequency interactions of ZPF wherein these subatomic particles are asymptotically free to oscillate as independent or free particles as quantum noise).

A further far reaching consequence [3] is mass itself becomes interpretable as a dependent quantity derived from a damped (with decay constant Γ) oscillation driven by random ZPF:

$$m = \Gamma c^3 / G = 2h\Gamma / \pi^2 c^3 \int \omega d\omega$$

with the only two free parameters, the damping factor Γ, and again the frequency, ω. The internal kinetic energy of the system contributes to the effective mass.

This leads to an overall average spectral density, written in terms of mass as:

$$\Delta\rho'(\omega) = m^2 c^5 \omega / 2h\omega_c^4 r^4$$

for the electromagnetic field distribution near ($1/r^4$) to the mass, m, which in detail is half electric and half magnetic.

One additionally attractive feature is the correspondence between this derivation and the view of gravity as a dynamical scaleinvariance breaking model (e.g. symmetry breaking near the Planck mass energy [14]). A final result includes the force calculation between two ZPF radiation oscillators of the correct form yielding Newton's average force law

$$< F > = -Gm^2 / r^2$$

Thus, for a Newtonian force to first order in a flat space time, Sakharov [2] could be credited for proposing gravity as not a fundamentally separate force and Puthoff [3] and co workers [4-5] applied the vacuum electromagnetic field to equate gravity to a long-range radiation force (e.g. van der Waals like force). Higher order oscillatory gravity modes vary as $\left(\sin[\omega_o / \omega_c]\right)^2$.

To first order, a weak G coupling constant, $G = \left(\pi c^5 / h\omega_c^2\right)$, appears for high frequency cutoff at the Planck scale. A corollary in analogy to electromagnetic shielding by ordinary matter can be rationalized as the problem of frequency mismatch at high Planck frequencies, e.g. ZPF cannot be fundamentally shielded. In other words, frequency mismatch precludes gravity shielding by matter.

The purpose here is to revisit the only free parameter, the frequency cutoff, more in the spirit of a mass resonant frequency. The motivation for this approach can be summarized as: 1) the generality of other complementary radiation effects without relying on ZPF alone (e.g. other isotropic, homogeneous radiation sources); 2) the weak coupling constant, G , yields a vastly smaller than observed size of the universe (e.g. too small cosmological constant) when the Planck frequency is used as a cutoff value; and (3) the particle mass, $m = \Gamma c^3/G$, can be viewed as a renormalized or 'dressed' mass with a resonant interaction potential that is frequency dependent in its coupling constant, G , and with 'bare' mass that is large, $m_o \sim \left(m_\rho^2 / m\right)$, where the experimentally unobservable, $m_\rho = \left(hc / G\right)^{1/2}$ is the Planck mass.

In particular, why this large 'bare' mass does not generate a large gravitational field is not a unique anomaly in the Sakharov derivation, since similarly large vacuum point energies are common to field theories. The important point is that the derivation $G(\omega)$ is general however to any isotropic radiation field with the Lorentz invariant energy spectra $\left[\rho(\omega) \sim \omega^3\right]$, thus the candidates for the cutoff frequency of the particular radiation source can be interpreted as a Planck scale only if the rest mass, m_o , is not composed of many terms, rather than just the ZPF leading term. Since the ZPF is akin to a van der Waals force [3 5], polarizability (in charge and mass) must be considered, but without also excluding any number of linear combinations that might have alternative cutoff frequencies, ω_c , or damping terms, Γ, 'ala particle physics interpretations for resonant masses during renormalization. In other words, once a gravitational energy spectrum, $\rho(\omega)$ is postulated that is Lorentzian invariant, many fundamental sizes (or corresponding frequency values) are smeared (or dressed) by any number of characteristic frequencies between zero and the high frequency electromagnetic (Planck) cutoff ω_ρ . Quite simply, is the expression, $\omega_c = \omega_\rho$, a requirement for all radiation sources?

Many types of particle oscillations may satisfy the general requirements of a Sakharov condition, each having a characteristic mass (and energy) as in calculating the mass of any fundamental particle at its resonant frequency (including underlying partial charges or dense bosons). This brings the calculation to a consideration of the high density fluctuations characteristic of a Bose condensate [15 19]. While the high density variation may intrinsically be of interest, the exploration has more to do with reconciling the ZPF interpretation of the Sakharov condition with the observed cosmological constant [14].

A "top down" view of calculating the cutoff frequency imposes the self consistency test for the cosmological constant, Λ, from the outset. To calculate, the total frequency integrated energy density of the universe must be included:

$$\rho(E) = \int \rho(E) dE = h\omega_c^4 / 8\pi^2 c^3$$

which must have a mass equivalent, contribute to the universe's curvature, and thus have a fundamental relation to the critical density to contain the expansion of the universe [14 15]. The mass - equivalent ZPF to reach the universe's critical density [15],

$\rho \sim 10^{-29}$ g cm^{-3} would necessarily limit the cutoff frequency for gravity to the value, $\omega_c < 7 \cdot 10^7 s^{-1}$, or between 10 -100 MHz.

A higher frequency greatly overshoots the cosmological constant, Λ , and induces extreme curvature in the universe. This problem has been cited frequently and stated most bluntly, as either ZPF or the cosmological constant requires revision. The relevance here arises from similarly large positive coupling terms in quantum gravity [15], which also generate a local gravitational Instability for typical upper limits on the cosmological constant, $\Lambda/8\pi G < 10^{12}$ cm^{-4}.

Rather than to dwell on the inconsistencies that plague attempts to reconcile quantum gravity, we particularize the problem to a case where the restriction to Planck scale becomes less clew, namely the high density fluctuations and universal scaling introduced in a Bose condensate. A Bose condensate, such as superfluid helium or superconductors [15 19], becomes of potential interest, mainly because of its arbitrarily dense boundaries and the classic Casimir experiment [20 22] which allows such dense material boundaries (two closely spaced conducting plates), if available, to modulate the background quantum fluctuation of ZPF. In other words, the matter-ZPF interaction becomes measurable by the observed attraction between two material boundaries. What dense boundaries might generate in Bose condensates remains a subject of great interest.

The significant case to investigate is whether Casimir-like interactions [20 22] will not only couple to ZPF radiation at a scale comparable to the quantum noise (or other radiation field), but also alter the value imposed by the Sakharov condition for G. It remains an open question whether this potential coupling interaction shares, as in ordinary critical phenomenon, the density correlation function, Φ, that is both independent of the coupling strength (or universal in renormalization) and consistent with the observed average energy density of the visible universe.

Thus the purpose here has bow to restate the Sakharov condition in the gravitational coupling constant, G, based on its only free parameter, a frequency cutoff, ω_c . Any potential relevance arises from similarly large values for the positive coupling term in quantum gravity, which generate conditions for a local gravitational instability for typical upper limits on the constant, $\Lambda/8\pi G < 10^{12}$ cm^{-4}.

To restate the Sakharov condition, matter in the vacuum provides boundaries for reduced 'Casimir like' modes available for otherwise isotropic radiation from quantum fluctuations (broad spectral noise). That this view reproduces Einstein gravity has been examined, including the full relativistic derivation [4-5]. The details of the appropriate mass, however, remain buried in the kinetic energy of general internal particle ('parton') motion [3]. Any appeal to a specific parton representation is limited only by essentially free particles with high frequency interactions, including underlying partial charges or dense bosons. The basis of considering arbitrarily high-density fluctuations in Bose condensate in analogy to the ZPF-Casimir experiment remains both an empirical and theoretical case to examine. There exist laboratory scale cases [15-19] where resonant radiation in the required 10-100 MHz range appear to produce anomalous effect for such Bose condensates as superconductors, but further work to confirm these results would be needed. In other contexts, these effects have been discussed as the Schiff-Barnhill effect for superconductors interacting with a gravitational field [23], but for the static rest moss rather than an effective mass in a conduction band.

Experimental Propositions

J. Weber [24,25] proposed the use of a superconducting Bose condensate for gravity wave detection, principally because of its potentially higher signal to noise ratio in carrying electrical signals upon length dilations in a relativistic framework for gravity waves travelling near the speed of light. W. Weber and Hickman [26] derived an experimentally testable relation based on torquing of a charged capacitor parallel to a gravity field, with

$$\tau = 2E_g / \pi[\alpha / (1-\alpha)^{1/2}]$$

where the capacitor will rotate relative to the gravity vector, for $\alpha = 2GM / rc^2$, is Schwarzschild radial coordinate $[dR = dr(1\text{-}\alpha)^{1/2}]$, E_g is dependent on the capacitor charge and geometry of the plates, $E_g = [Q^2 d/2\varepsilon WL(1\text{-}\alpha)^{1/2}]$, for a plate separation, and radial dimensions,W and L, charge Q, and ε the permittivity of free space. For plate separations of 2 mm on Earth, the maximum torque is approximately $\tau = 10^{-12}$ Nm, when charged to 2/3 dielectric breakdown. While not entirely promising for detection of such low torques, the large separation (2 mm) distance between capacitative plates naturally prompts generalization to the classic Casimir force [21] experiments only recently confirmed experimentally [20]. In particular, we rewrite the torque values to include the frequency terms derived with the Sakharov condition

$$[G = (\pi c^5 / h\omega_c^2)]:$$

$$\alpha = 2M\pi c^3 / h\omega_c^2 r$$

The appeal of this formulation is that a frequency dependent torque is derived, which further makes contact with proposals to modulate the Casimir capacitative plates for continuous extraction of energy [27]. This result requires further investigation experimentally, particularly to compare with previous reports for anomalies in AC- tuned electrical capacitors [28].

References

1. Ze1'dovich Ya. B. JETP Letters, 6, 345, 1967.
2. Sakharov A. Vacuum quantum fluctuations in Curved Space and the Theory of Gravitation, Sov. Phys. Reports, 12,1968,1040 1041.
3. Puthoff H. E. (1989) Gravity as a zero-point fluctuation force, Physical Review A, 39(5): 2333 2342, March 1, 1989.
4. Haisch B., Rueda A., Puthoff H.E., (1994) Inertia as a Zero Point Field Lorentz Force, Physical Review A, 49:678 694.
5. Haisch B., Rueda A., Puthoff H.E.,"Inertia as a Zero Point Field Force" Physical Review A 49, N 2, 678 (1994).
6. Ambjorn J. and Wolfram S. (1983) Properties of the Vacuum, 1. Mechanical and Thermodynamic, and Properties of the Vacuum, 2. Electrodynamics, Annals of Physics
7. Ambjorn J. and Wolfram S. (1983) Properties of the Vacuum. 1. Mechanical and Thermodynamic, Annals of Physics, 147:1 32.
8. Fulcher et al., "The Decay of the Vacuum," Sci. Am., vol. 241, p. 150, Dec. 1979
9. Puthoff, H.E. "Source of Vacuum Electromagnetic Zero Point Energy" Physical Review A 4 0, 4857 Nov 1 (1989); Errata and Comments, Physical Review A 4 1, March 1(1990); Physical Review A4 4, 3382, 3385 (1991)
10. Senitzky I.P., "Radiation Reaction and Vacuum Field Effects in Heisenberg Picture Quantum Electrodynamics", Phys. Rev. Lett. 31(15), 955 (1973). As pointed out by Puthoff [3] the relativistic results for the Sakharov condition have so far been encouraging, while the consequences for nuclear interactions in all coordinate frames have not been fully explored.
11. da Costa L. N., Freudling W., Wegner G., Giovanelli R., Haynes M. P., and Salzer J. J. (1996) The Mass Distribution in the Nearby Universe, Astrophysical Journal Letters, 468: LS L8 and Plate L 1
12. Modanese G. (1996) Theoretical analysis of a reported weak gravitational shielding effect, Europhy. Lett., 35(6):413 418.
13. Shupe M.A. The Lorentz invariant vacuum media, Am. J. Phys. 53, 122 (1985). A cautionary note is that lower frequency cutoffs can violate Lorentzian invariance, thus allowing a moving detector to reveal absolute motion by recording Doppler sifted frequencies. Standard methods might treat such effects like the cancellation of terms that remove anomalous ZPF infinities from field theories, but these topics remain to be explored.
14. Zee, A. Phys. Rev, Lett, 42,417 (1979); Phys. Rev. D. 23, 858, (1981).
15. Torr D. G. and Li. N. (1993) Gravitoelctric Electric Coupling vVia Superconductivity, Foundations of Physics Letters, 6(4): 371 383.
16. Unnikrishan C. S. (1996) Does a superconductor shield gravity? Physics C, 266:133 137.
17. Podkletnov E. and Nieminen R (1992) A Possibility of Gravitational Force Shielding by Bulk $YBa_2 Cu_3 O_{7-x}$ Superconductor, Physics pp.203: 441 444.
18. Li N, and Torr D. G. (1992) Gravitational effects on the magnetic attenuation of superconductors, Physical Review B, 46(9): 5489 5494. A simple consequence of the Sakharov condition $G = (\pi c^5 / h\omega_c^2) \sim 1/\int \omega d\omega$, can be written for the gravitomagnefic permeability as: $\mu_g = 4\pi G / c^2 = 4\pi^2 c^3 / \left(h \int_\circ^{\omega c} \omega d\omega\right) \sim 1/\int \omega d\omega$ which suggest that the same frequency resonance implied by the ZPF derivation will share similar consequences for vector gravity

effects. See also, DeWitt, B. S. Superconductors and Gravitational Drag, Phys. Rev. Lett. 16, 102(1966).

19. Li N., Noever D., Robertson T., Koczor R., and Brantley, W. (1997) Static Test for a Gravitational Force Coupled to Type II YBCO Superconductors, Physics p., 55, 287.
20. Lamoreaux S. K. (1997) Demonstration of the Casimir Force in the 0.6 to 6 mm Range, Phys. Rev. Letters, 78:5 8.
21. Milonni P.W. et al., "Radiation pressure from the vacuum: Physical interpretation of the Casimir force", Phys. Rev. A, Vol. 38, No. 3, 1621 August 1988.
22. Milonni P. W. (1994) The Quantum Vacuum, Academic Press, San Diego, CA.
23. Schiff L.I. and Barnhill M.V. Bull. Am. Phys. Soc. 11, 96, (1966) and refn. 18.
24. Weber J. (1960), Detection and Generation of Gravitational Waves, Physical Review, 117(1)306 313.
25. Weber J. (1966) Gravitational Shielding and Absorption, The Physical Review (The American Physical Society), 146(4): 935 937.
26. Weber W. and Hickman H. (1997) A possible interaction between gravity and the electric field, Spec. Science Tech. 20, 133 136
27. Forward R.L. "Extracting electrical energy from the vacuum by cohesion of charged foliated conductors" Phys. Rev. B, Vol. 30, No. 4,1700 August 1994
28. Woodward, J. F. (1992) A Stationary Apparent Weight Shift From a Transient Machian Mass Fluctuation, Foundations of Physics Letters, 5:425 442.

The Problem of Electron and Physical Properties of Time:

To the Electron Technologies of the 21st Century

V.P. Oleinik

[*]Department of General and Theoretical Physics,
National Technical University of Ukraine
"Kiev Polytechnic Institute",
Prospect Pobedy 37, Kiev, 03056, Ukraine [**]Institute of
Semiconductor Physics, National Academy of Sciences,
Prospect Nauky 45, Kiev, 03028, Ukraine;
e-mail: yuri@arepjev.relc.com

> ***"...it is necessary to periodically subject to the deepest revision the principles, which were recognized as final and were no longer discussed".***
> ***Louis de Broglie***

Abstract

The results of an approach based on the synthesis of standard quantum electrodynamics and of the ideas of self-organization in physical systems are briefly outlined. The quantum model of electron as an open self-organizing system is constructed, with the physical mechanism of self-organization consisting in the back influence of the own field created by electron on the same electron. The own field is considered as a physical property of electron, intrinsically inherent in electrically charged matter, which is included in the definition of the particle from the very beginning. The own field of electron endows the particle with wave properties and represents a bearer of superluminal signals, which can be used for the creation of qualitatively new communication systems. Because of the inseparable link between space and time, the force in relativistic mechanics is the cause of change not only of the velocity of particle, but also of the course of time along the particle's trajectory. For this reason the flow of time in some area of space depends on the character of physical processes, occurring in it, and, therefore, time can be controlled by slowing down or accelerating its course with the help of material processes. The conclusions of the paper are not in conflict with the special theory of relativity (STR); they are a direct consequence of relativistic equations of motion and represent an essential development of the generally accepted notions about space and time. At present all the necessary prerequisites are available, both theoretical and technical, for the practical mastering of the own fields of particles and of the physical properties of time.

1. Introduction. The Problem of Electron and Future Outlook

Electrodynamics, what is this? What is its value for man? Electrodynamics is the theory of electromagnetic interaction, one of four interactions existing in nature. Its role in the life of society is seen from the fact that the most part of natural phenomena, which we encounter at every step, is of electromagnetic origin: it is due to the interaction of electromagnetic field with electrically charged particles entering into atoms and molecules. It is fair to say that electromagnetism plays a crucial role in the life of mankind as it determines the ways of technical advance of society [1].

The key problem of quantum electrodynamics is the problem of electron, which can be formulated as follows: to construct from the first principles a non-contradictory model of electron, which takes into account experimental facts, i.e. to find the dynamical equation capable of describing the unique physical properties of electron, its internal structure, its behaviour when it interacts with electromagnetic field.

Electron was discovered a little more than 100 years ago, in 1897. With discovering the electron the revolution in physics began, which has resulted in unprecedented technical advance of society. The summit of development was reached in the middle of the 1950s and then the long period of evolutional development followed, when new physical principles were used to describe various physical processes and phenomena. The violent development of physics became slower in the 1970s and was replaced by stagnation in the subsequent years. The stagnation in electrodynamics continuing already over a period of several decades is gradually giving place now to a new ascent. The new scientific revolution is starting, which is associated with electron again, much as it happened hundred years ago. The reason is that electron is the most unique particle storing in itself the deepest mysteries of nature and the degree, to which they are disclosed, determines the wants to stay anonymous, until his patent application is done and university verification tests will be done). The claims are: 1200 Watts coil out with about 1076.4 Watts in into the driving motor at 3450 RPM. 8 amps 117volts at no load 9.2 amps 117 volts at full load. The output of about 1200 Watts is already a total overunity operation! As they just increase the input power by about 140 Watts only between idle and load state and they get 1200-Watts output it seems indeed a case, where Lenz law is violated! This generator also has NO motor effect! If you supply current to the coil, the permanent magnet in the center will not rotate; cause the flux just stays inside the toroid core! There you can see, that the back drag does not influence the mechanical rotation of the magnet!" Stefan used very good criterion to prove high efficiency of the design: **There is no back-torque effect! It is most important aspect of Gramm's generator.** You can contact directly Stefan Hartmann: Keplerstr. 11 B, 10589 Berlin, Germany. Tel: +49 30 345 00 497, FAX: +49 30 345 00 498 email: harti@harti.com info@ccard.net (Please, note: Dr. Harman referred to my old web site www.time-machine.spb.ru which is closed now).

So, basic principles of MEG and Φ-machines are the same. It was patented more than 100 years ago. Primary magnetic flux is topologically separated in two (or more) fluxes, which are mutually compensated in the ring core. Advantages of MEG are absence of moving parts since special input coils produce changes of primary flux. Also level of saturation in ferromagnetic material obviously should be corresponding to intensity of primary magnetic field, which is created by the permanent magnet, Fig.4.1.

Besides MEG the same principle can be (and already was!) realized in many other systems. So, there is no any news in the USA patent #6,362,718 granted for "The Motionless Magnetic Generator". What did they claim? You can find it in the patent: "The first input coil and the first output coil **extend around portions of the first magnetic path**, while the second input coil and the second output coil **extend around portions of the second magnetic path."** Yes, it is the same bi-directional principle we discussed above: two parts of the magnetic flux and each coil produce effect to reduce flux due to this superposition.

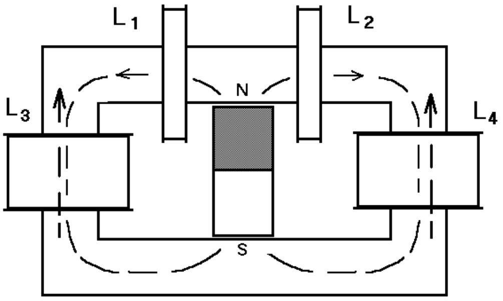

Fig.4.1

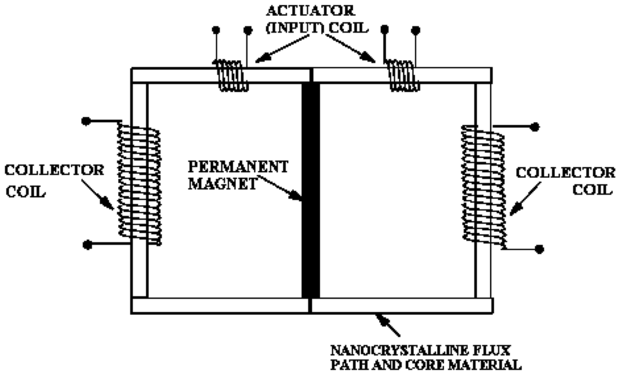

Fig.4.2
Diagram of prototype by Bearden.

In conclusion I'd like to confirm our sincere interest to develop joint work with all new energy research teams if they are not trying to obscure the issue of the technology by means of complex theoretical constructions and common words about zero point energy.

Matter as a Resonance Longitudinal Wave Process

Alexander V. Frolov

Abstracts

There is experimental data on gravitation anomalies for cases of resonance irradiation of the Bose condensates (superfluid helium or superconductor) at 10-100 MHz frequencies. It is developed by the author in frames of his aether theory that can be used for practical applications in aerospace and new energetics.

ZPF or aether fluctuations

The fundamental electromagnetic radiation field (Zero Point Field) ZPF or the quantum vacuum noise is a recently accepted physical notion. In the article [1] David Noever and Christopher Bremner used it to derive a frequency – dependent version of Newton's gravitational coupling term G. On the other hand we can consider the quantum vacuum noise as aether fluctuations. Dr. Alexander Mishin [2] described experiments on registration of these processes by means of special equipment. Both approaches (ZPF and aether fluctuation) allow to conclude that mass and inertia arise from these oscillations. However if we are considering the oscillation as some aether process then we can assume and describe some physical mechanism of this process.

One of consequence of the vacuum energy model, which is described in [1] is that "the attractive force of gravity becomes reducible to the radiative interaction between oscillating charges..." Let's clarify which kind of radiation can be created by oscillating electric charges.

There are many different sources to find the answer on this question and one of them is the article by Prof. Kirill P.Butusov [3] on symmetrization of Maxwell's equations and practical methods of generation of longitudinal waves in vacuum. So, ZPF model has a direct relation with the aether model since indirectly it leads to the question of longitudinal waves in vacuum. Physically they are waves of density of energy and in the aether model the waves are areas of more dense and more rarefied aether. Let's note that there are standing waves besides moving waves.

To consider the interaction of some mass particles and the fundamental field the notion of subatomic charges "partons" was introduced [1]. So, the mass itself "becomes interpretable as a dependent quantity derived from a damped oscillation driven by random ZPF" [1]. The authors wrote about "internal kinetic energy" of the mass particle and it can be considered as a function of ZPF oscillation frequency. In the aether theory of mass there is a similar notion of "aether vortex", which represents some cyclical process of some frequency and it is possible to calculate its kinetic energy. This aether vortex model of matter elements allows to assume real methods to change parameters of vortex and to get changes in parameters of existence of the matter. On the other hand we can discuss the possibility to change some physical parameters of aether in areas of the vortex to get the same result. This possibility follows from the well-known N.Kozyrev's experiments, which were named "investigation of active properties of time". N.Kozyrev used chronal (temporal) approach in his theory. We have to change his notion "the density of time" to "the density of aether" to get a direct link between his experiments and the aether theory of mass.

N.Kozyrev and others have [4,5,6] experimentally demonstrated that irreversible processes in matter produced changes of aether density in the area of the experiment. Detectors of different type can register this change. It is obviously that any matter element (i.e. the aether vortex) in this area of changed aether density should get more inner (kinetic) energy or slow the inner motion. From the chronal point of view these are changes of inner time of this matter element.

Gravity shield

One more interesting point that is discussed in the article by Noever and Bremner [1] is a problem of gravity shield. The authors show that resonance interaction with ZPF produces "the particle mass" and it can be viewed as "a renormalized or "dressed" mass with a resonant interaction potential. Similar resonance approach is used in the conception of de Broglie's matter waves. Also the authors [1] mentioned the existence of an experimentally unobservable mass. In this case ZPF cannot be fundamentally shielded by matter since "frequency mismatch precludes gravity shielding by matter" [1]. The only way to get screening of ZPF fluctuations seems to be very complex: it is necessary to provide frequency matching for whole wavelength band of the oscillations.

I think some specialization is necessary here to explain experimental gravity anomalies with Bose condensates experiments (superfluid helium or superconductors): **special process in matter can be used as the gravity screen and this approach does not involve the frequency-matching problem.**

We have concluded above that any matter element is a resonance process and its energy is derived from ZPF. It is useful to note that these are longitudinal wave oscillations of energy density in aether. In this case, the gravity shield problem can be solved in frames of the aether vortex conception of matter.

The longitudinal wave is a moving (or standing) areas of rarified and thickened aether. Let's consider the moving wave, which is responsible for gravitation attraction effect. How can we stop, re-direct or reflect longitudinal wave in aether by means of aether vortexes (matter elements)? **We can produce interaction with this wave only by means of other longitudinal waves.**

In macro-level this idea can be realized as longitudinal wave generator. Electromagnetic processes, which can be used as sources of directed longitudinal waves, are known and some of them are described in [3]. In other way the gravity shield can be produced as longitudinal waves generated by natural aether vortexes (i.e. by matter elements) if the matter exist in a special exited state, for example for cases of resonance irradiation of superfluid helium or superconductor at 10-100 MHz frequencies.

Matter element as resonance process

In [1] the authors wrote that it is possible to calculate "the mass of any fundamental particle at its resonant frequency." There is the question: what is the general basis of whole spectrum of stable elements masses?

In 1996 the author published the article "The concept of mass process" [7]. At first in this work physical sense and notion of 3-dimensional curvature was introduced. By analogy with known mathematical notion of linear curvature $\rho_1 = \frac{1}{r}$ (where r is radius) and uniform surface curvature $\rho_2 = \frac{2}{r}$ it was proposed to calculate curvature of a 3-dimensional space as

$$\rho_3 = \frac{3}{r} \qquad (1)$$

The radius r in this case means that in a 3-space there is some periodical process. In other words, 3-dimensional matter is a resonance process.

Further, de Broglie used formulations E=hf and E=pc (where p is momentum, h is Planck constant, f is frequency and c is velocity of light) to derive the following:

$$hf=pc \qquad (2)$$

that allows us to get the well-known formulation

$$\lambda = \frac{h}{p} \qquad (3)$$

There is another logical branch of this idea that leads to the understanding of the mass properties of matter as a resonance process. Instead of E=pc in [7] it was proposed to use $E=mc^2$. In strength of the wave-particle duality we can write the equation

$$mc^2 = hf \qquad (4)$$

and from this equation the mass can be presented as resonance electromagnetic oscillations

$$m = \frac{h}{c^2} f \qquad (5)$$

Let's note that f=1/T, where T is some period of oscillation. So, we can write the following

$$m = \frac{h}{c^2}\frac{1}{T} \qquad (6)$$

where $\frac{h}{c^2}$ is new constant between mass and period of time.

There is an important conclusion: any mass is a process and there is some period of time, which corresponds to this mass. In other words, there is no physical sense of time separate from some process of existence of mass. Product mass and period is a constant value, which was named as a chronal constant

$$mT = \frac{h}{c^2} = const \qquad (7)$$

The chronal constant is a parameter of some real space and it is equal to $0.73725 \cdot 10^{-50} [Js^2 / m^2]$

Also in this work [7] there was a demonstration of several examples of newly discovered physical law: spatial curvature of some natural objects (proton, planet, DNA molecule) is a whole number. There is some analogy with the nuclear physics notion of wave number. From this fact we can assume that main natural matter elements exist in main resonance states. For example, if Bohr radius is 0.52917 Angstrom, then we can find the wave-length $l = \pi d$ and the linear curvature is $\rho = 1/l = 3.0075 \cdot 10^9$ (m) and 3-dimensional curvature of this object is $\rho = 3/l = 1.0025 \cdot 10^9$ (m) that is unit of mater, corresponding to simplest atom, i.e. unit matter engine. Let's note that it is near the unit and some distortion of 0.0025 means non-ideal resonance state of the system.

Calculations for planet Earth in [7] were based on the known period of orbital rotation T=31557600 sec that corresponds to frequency of electromagnetic oscillations

$$f = 1/T = 3.168861 \cdot 10^{-8} (1/s) \qquad (8)$$

and wave-length

$$\lambda = c/f = 9.46... \cdot 10^{16} (m) \qquad (9)$$

The curvature (if this wave-length is considered as radius of the resonator) is following wave number

$$\rho = 1057.00 \cdot 10^{-20} (1/m) \qquad (10)$$

Also we can use other known data about the planet. Daily rotation period of our planet is known T=86400 sec and we can calculate its wavelength $\lambda = 3469{,}82(m)$ and corresponding curvature (wave-number). Sure, it is also a whole number with a good accuracy:

$$\rho = 2882 \cdot 10^{-7} (1/m) \qquad (11)$$

The laws of physics in macro cosmos and micro cosmos are similar. From these calculations it was assumed that whole formation of mass spectrum of stable chemical elements of matter is determined by similar physical mechanism.

Creation of mass

In short we can summarize that technology of longitudinal waves in aether is a real basis for creation of matter with mass and inertia properties. N. Tesla used this method to produce different objects: from ball lightning up to electrons. Velimir Abramovic says in his article [8]: "The principle of resonance and harmonic oscillation of aether seems to be so clear that all problems of modern physics, especially a problem of energy conversion, will be solved with its development. By means of his vacuum tube Tesla got protons, electrons and neutrons directly from aether and reproduced them at any distance. Instead of giving a possibility to the bundle of protons to move through space to some place, he created conditions for momentary appearance of arbitrary quantity of particles in the given place."

Any objects can be classified as aether vortex and parameters of this vortex determine its mass, electric charge and other properties of matter.

The "parton" as element of matter in [1] is a useful tool for description of physical properties of aether.

Longitudinal waves in Woodward's experiment

In [1] the authors state that resonant radiation in the required 10-100 MHz range appears to produce anomalous effects for such Bose condensates as superconductors. In my opinion it is a particular case of discussed above technology of longitudinal waves in aether due to possibility of transformation of transverse electromagnetic waves in longitudinal waves in the superconductors. This transformation in plasma is a well known physical mechanism.

More facts to prove this idea: by Woodward [9] there is a special requirement, i.e. the frequency of mechanical vibrations should be twice the frequency of electrical oscillations in the capacitor, which demonstrates the weight anomalies. But from the other hand it is a common rule for creation of longitudinal weaves in plasma! Also it is a necessary condition for generation of parametrical oscillations! So, we can assume that basis of the effects in [1] and [9] is a generation of longitudinal wave in aether.

Conclusion

Any element of matter can be considered as resonance process of aether oscillations, which are longitudinal waves. There is an analogy with description of these longitudinal waves and well-known matter waves by de Broighl. Experimenting on the longitudinal waves generation and especially experiments on standing waves to get gradient of aether pressure allows to develop gravity control technology.

References

1. D. Noever and C. Bremner, NASA Marshall Space Flight Center, Large Scale Sakharov Condition, 35th AIAA/ASME/SAE/ASEE joint Propulsion Conference and Exhibit, 1999, Los Angeles, Ca, USA.

2. Alexander M. Mishin, The physical system of artificial biofield (experimental research of ether), New Energy Technologies, #1, 2001, p.45.

3. Kirill P.Butusov, Symmetrization of Maxwell-Lorenz equations, New energy Technologies, #2(5), 2002, p.14.

4. Kozyrev N.A. Selected Works, 1991, Leningrad, publ. by Leningrad State University.

5. L. Shikhobalov, N. Kozyrev's ideas today, New Energy Technologies #2 (5), 2002, p. 20.

6. Alexander V. Frolov, Kozyrev on possibility of decrease of mass and weight of the body, New Energy Technologies #2 (5), 2002, p. 35.

7. The Concept of Mass Process, Frolov A.V., Proceedings of the congress "New Ideas in Natural Sciences", 1996, St.Petersburg, published by PiK Co., p.123-134.

8. Tesla, Velimir Abramovic, New Energy technologies, #1(4) 2002, p.17.

9. Woodward, J.F. A Stationary Apparent Weight Shift From a Transient Machian Mass Fluctuation, Foundations of Physics Letters, 5, p.425-442, 1992.

Gerlovin's Theory of Activation

Alexander V. Frolov

It is a review of the famous book by Ilia L. Gerlovin "Basis of unified theory of all interactions in matter"published in 1990, St.Petersburg, Russia. We hope this article let you discover some new aspects of physical vacuum structure to develop more new experimental methods. Comments made by Alexander V. Frolov, Editor.

In [1] the author wrote about different methods to activate water solutions: mechanical, thermal, acoustic, magnetic and electrical. One of the known methods is an activation by means of electrohydraulic method. There is also some information about activation of other mediums, mainly liquids, but also some gases and solid bodies.

There are no theoretical explanations of these facts to explain all aspects of these phenomena. Furthermore, complexity of interpretation of these phenomena in frames of common physical notions induced some scientists to announce these phenomena as non-existing and "illegal".

Ilia L. Gerlovin formulated the physical principles of theory of activation of mediums on the basis of new physical theory, the Theory of Fundamental Field (TFF). These principles are based on his two important conclusions from the TFF:

a) "Space around us is not empty, physical vacuum consists of material physical objects, i.e. elementary particles of vacuum (EPV). These particles are responsible for main activation processes;
b) Force interactions between atoms in molecule, between molecules in crystals has not spherical symmetry in the crystals of solid bodies, but an axial symmetry and the interactions are changing in time with very high frequency of about 10^{18} Hz. This feature of force interactions also makes its own contribution to the activation of mediums." [1, p. 314]

So, it was assumed that the phenomenon of activation of mediums can be defined as anisotropy of force interactions, which leads to **"meta-stable state, which can be called *structurally activated state of the given structure*".**

Here is some difference in principle between chemical term "activation", which characterizes a transformation of molecule or atom in some active state with an increased energy, which is sufficient to provide a chemical reaction. It is ***energy activation*.** **Gerlovin described new notion, a *structural activation:* "This phenomenon can be classified as some change of structure of activation object. With this, energy of molecule can have no changes, and *active properties***

Coupled with aetherodynamics time conception, which was suggested by Alexander V. Frolov, the works on the control of space-time parameters gain the possibility for development and commercial application. As a theoretical basis there are those N. Kozyrev works where his conception of "time density" are replaced by that of "aether density" according to Frolov.

In September 2002, Faraday Labs Ltd Company plans to complete testing of the first experimental system, and to start the patenting and research of applied aspects, first of all in medicine.

Physical Principles of the Time Machine

Alexander V. Frolov

Experimental success of research team headed by Dr. Vadim A. Chernobrov, Moscow was reported in [1]. The time course can be controlled as rate of any process in local space-time (inner space of the Time Machine). It can be decelerated or accelerated by means of special "converging electromagnetic waves". Ordinary waves move from the source whereas special "converging waves" move to some central point, i.e. into the focus of the system. In Chernobrov's design of the Time Machine this process is organized by means of several spherical envelops, which consist of several electromagnets. Electronic control unit controls the processes in this design. Dr. Chernobrov reported about 3% change of the time course in 4th version of the system, which was tested with a human inside. The goal of Dr. Chernobrov's work is to research the medical aspects and experimental investigation of the principles. Several important conclusions were obtained from the project: the time course can be controlled and character of the changes is different for acceleration and deceleration mode.

Other known publication and research projects on the same topic seem to be very far from any commercial and practically useful application. Obviously the topic is very new and fantastic for most of scientific community and at first we have to clarify the physical principles of the time control project, which is started by Faraday Labs Ltd.

In this project we believe that notion of time is one of possible description of real physical properties of our Universe. So, it is not mathematical abstraction but some aspect of physical reality and we can discover some physical properties of time. Russian astrophysicist N. A. Kozyrev [2] developed a theory of active properties of time and according to his point of view there are two properties: time course and time density. Prof. Kozyrev demonstrated experimentally that time density in area of some process (changes of matter) is dependent on entropy parameters of the processes. In [3] it was demonstrated that Kozyrev's experiments could be interpreted in aether theory and it has led to simple physical conclusions and clear experimental perspectives: **time course and its density can be explained and controlled as parameters of aether.** Directions of aether flow and density of aether are subjects of experimenting with non-reversible changes in matter, for example, in crystallization or melting processes. Also it is possible to use special electromagnetic processes, for example, Chernobrov's "converging waves" or other longitudinal waves as methods of aether compression or rarefaction. If we assume that process of existence of elements of matter physically can be explained as aether vortex processes then its rate is a parameter of **aether income/outcome balance** (aether inflow in element of matter and aether outflow from the element of matter). It was also described in Time Rate Control (TRC) theory [3]. To control this balance it is necessary to develop technology of longitudinal waves generation, its focusing and resonance effects. The previous research and experimenting on the topic has been made by N. Tesla.

Let's assume that we have some technology to change parameters of time course. How should we organize this local space-time (what is spatial topology of the design)? There is a very interesting experiment to get the answer: rotation of a heavy cone (for example, lead cone) entrains surrounding aether, so a vortex appears, which is a toroidal formation of aether (rings). The rings can exist in space for a long period. The further question is: Why does the beam of light (laser beam) directed to the cone by tangent create a luminous ring? We can assume that due to natural properties of photons (light propagates along the geodesic line in space) some autonomous closed toroidal space should be created in such experiment. The next thought is: since space and any matter exist in time then we can speak about some autonomous time. The general conclusion is to be the following: autonomous 4-dimensional space-time can be created as toroidal aether vortex.

Here is point to note some aspects of research project by Prof. Robert Mallett, Connecticut University, USA. In fact, sometime next year, he hopes to produce the first piece of technology that eventually will allow him to build a time machine. By Mallett it will be a device that employs lasers "to twist space". Why is he going to close the beam of light? His theoretical background is knowledge about black holes, i.e. understanding of the connection between gravity and curvature of space-time. In Einstein's theory both matter and energy can bend space and time. So Prof. Mallett assumes that curvature of space-time can be changed not only by mass (like a black hole) but it can be affected by energy of photons. This has led Prof. Mallett to consider the possibility of using a circulating beam of light to twist space and to create closed loops in time. It is predicted that a spinning neutral particle, when placed in the ring, is dragged around by the resulting gravitational field [4]. From the first view it is the same approach we have considered above (experiment with aether toroidal rings). But proposals by Prof. Mallett differ in principle from the aether conception.

The main aspect of this technology is a creation of autonomous (self-closed) toroidal space-time. Autonomous geodesic world line of this space-time is self-closed. Any photon should be circulating in this system due to its properties: photon is always moving along the straight line of the space.

More deep understanding of this technology follows from the explanation of photon as oscillation of aether. Any photon can be considered as result of relative motion of the matter (observer) in absolute space (immovable aether). Usually a photon is considered as moving object in space. But we can assume that observer is in the motion and the photon is oscillations of the absolute space (immovable aether). Which approach is more real one? Sure, it is more easy to consider a photon as moving object but let's remember fact of our real motion in the Universe and fact of the Universe expansion.

So, ideas by Prof. Mallett are very far from the **aether nature of the time phenomenon.** He follows the black holes theory and general understanding of space-time distortion due to mass or energy presence. Also he knows that a light beam should be closed in a ring. However Prof. Mallett is very far from physical basis of the effects. **The key of time rate control is technology of artificial aether flow, creation of aether vortex systems (AVS), management on density and direction of aether flow.** There are several technical methods to produce it. Any light beam should be curved in self-closed "light ring" if it is placed in a toroidal aether vortex and we can say that this system has own space-time.

What does "some changes of time course" mean? We can measure it as some changes of standard rate of oscillation process, for example, some stable wavelength of laser beam or quartz oscillations. There is a well-known experiment with two atomic clocks (one of the clocks is placed on the roof of some building and another one is placed on surface of planet). Due to vertical component of gravity the time course should be different and it can be measured. How can we organize difference in these measurements if both atomic clocks are placed in the same altitude?

It is necessary to consider gravity nature in frames of the aether conception. Two atomic clocks demonstrate difference in measurements due to difference in aether flow density. Hence, by means of aetherodynamics methods it is possible to control the rate of oscillation processes in the atomic clocks and in any matter (i.e. time course itself).

The aetherodynamics methods have a clear analogy with electrodynamics: **motion of charge produce field and there is the induction law**. Really, classical electrodynamics can be considered as particular case of the aetherodynamics. So, physical sense of any field is stress or deformation (it is some static field) or oscillations of aether.

Let's introduce the notion of chronal (temporal) charge to consider some technical aspects. In electrodynamics we assume an electric charge as element of matter with positive or negative electric properties and we have to compare it with some reference (zero charge or test charge). Let's note that in any case we have to consider "charge of some particle" but not an "abstract charge". So, we can postulate that any element of matter has **zero chronal charge** if it is moving from Past in Future with standard (usual for measurements of surface of our planet) time course. If the time course (i.e. existence of some element of matter) is decelerated then it can be measured as decrease of standard oscillation frequency of the matter. Time course acceleration means some increase of standard oscillation frequency of the matter. Let's determine that in the first case it is negative chronal charge and in the second case it is some positive chronal charge. Atomic clock is one of possible methods to measure zero chronal charge or to find some relative positive or negative difference.

It is predicted here that motion of chronal charge should produce a chronal field. Some provisional data was received by Frolov from simple experiments on the rotation of a heat source. Accelerated motion of chronal charge (changes of density of chronal current) should produce aethero-induction effect that is an analogy (or more general case) of Faraday's induction effect. This effect can be detected as secondary (induced) deceleration of time course in nearest area of accelerated time matter. Another case is a secondary (induced) acceleration of time course in the nearest area of decelerated time matter.

Technical realization of aethero-induction method seems to be very close to idea, which is described in classical epic "Back to the Future". At first, it is necessary to create or to collect some chronal charge in a "flux condenser" and then to accelerate it in space up to some velocity. According to the aether conception, this creation of the chronal charge is a real technical process.

It is assumed that estimated chronal effects are demonstrated as some threshold field, i.e. space-time has some stable discrete energy levels and changes of its curvature should have discrete threshold mode. All new aspects disclosed in this paper are the subject of a patent process. Faraday Labs Ltd organizes experimental program on the topic. Practical application of this technology is new energy systems and propulsion methods.

References

1. Chernobrov V.A. Experiments with a man in the Time Machine, New Energy Technologies, #3 November-December 2001, p.6.
2. Kozyrev N.A. Selected works, Publ. by Leningrad State University, 1991.
3. Frolov A.V. Practical application of the Time Rate Control theory, New Energy Technologies, #3 November-December 2001, p. 15.
4. Mallett, R.L., Weak gravitational field of the electromagnetic radiation in a ring laser, Phys. Lett. A, 2000.

investigated completely yet. It was found also that harmful effect on biological systems is not related to the process of movement in Time itself but is a result of the difference of the Time rate value in various parts of a body (a biological system).

Inside of the laboratory setup it was also discovered that Time could be changed with some inertia. Areas of space having different Time rates have vague borders. With sufficient difference in Time rate the human can see an area with a different Time rate as some white mist. Higher the difference – the mist is denser, that can be used as an alarm signal for biological systems. It is possible to consider Time-travel as possible and (after experiments with mice) there are reasons to suppose it will be safe for travelers if they follow certain rules. It is especially necessary to emphasize: the trips through Time (due to new discovered properties of Time) can't affect the Past and they can't change our past history. All the so-called paradoxes for the traveler in Time (for example when "he meets himself in the Past" or "he kills his grandfather in his childhood" have clear solutions in 3-dimensional Time.

It is possible to consider as a proven fact that Time has more than one dimension, i.e. O. Bartini's theoretical calculations are confirmed by these experiments: Time has 3 dimensions. Hence our Earth world can be considered as a 6-dimensional object: length, width, height, age or date of Time, variant of a History or erosion of Time, density or rate of Time. The concept of "the Arrow of Time" as fourth dimension (moment of Time) is a particular case of the concept of sixth dimension (rate of Time) that leads to the physical concepts of gravitation and energy and they are simultaneously connected. Concepts of the " Einstein-Rosen bridges" known since 1916 or "worm-holes" introduced into science by John Willer in the 50's, are travels in 5th and 6th dimensions, i.e. the "classical" Time travels, which were described by H. Wells.

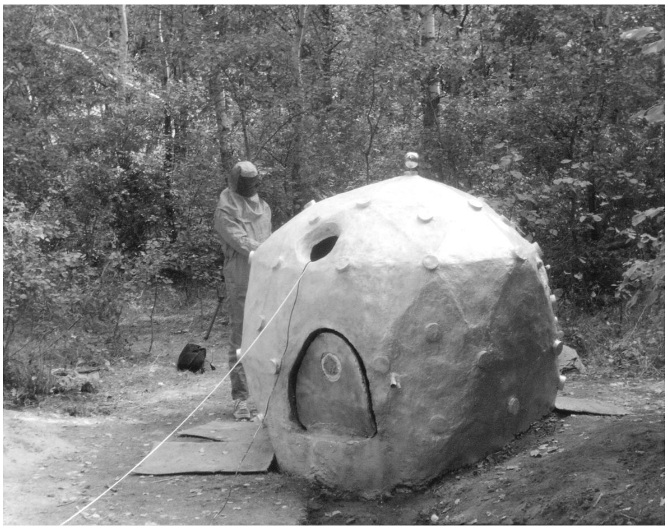

Editor's: As the reader could note, the author does not disclosure the secrets of the TM design. From the photo you can see the electromagnets, which form the regular stereometrical construction as well as the cables from the TM to the control unit. Dr. Chernobrov mentioned the converging electromagnetic waves only. So, to understand how it works, it is necessary to get a clear notion of the converging electromagnetic waves. Let's imagine the ripple effect created by a stone in the water. The waves move from a central point to periphery. The converging waves are just an opposite process: the waves move from periphery to the central point. Is it possible in Nature? Yes, sure. Dr. Chernobrov wrote: "Let's throw a hoop on the water and inside of the hoop we'll see converging waves." The Time Machine technology by Dr. Chernobrov is based on the similar principle.

Time Machine Project

Alexander V. Frolov

Scientific Expert of the Russian Physical Society,
General Director, Faraday Lab Ltd
Tel/fax: 7-812-380-6564 Tel: 7-921-993-2501
Email: director@faraday.ru alex@frolov.spb.ru

May 29, 2002

Faraday Labs Ltd and Dr. Vadim Chernobrov have signed the agreement on scientific-research work on investigation of active properties of time.

In the course of the previous experimental works, carried out by Dr. Chernobrov's research team during the period from 1984-2002, four versions of Time Machine had been made and tested. At these devices (the biggest system is about 1 meter in diameter) the effects of deceleration and acceleration of time course were created and measured. The principles of control of time course velocity were based on the interconnection of electromagnetic processes and physical properties of space-time. Special electromagnets, operating in pulse mode, are placed at the spherical frame. They create the so-called "converging wave", which by Alexander Frolov is a longitudinal wave in nature.

Alexander V. Frolov, General Director Faraday Labs Ltd and Ph. Dr. Vadim A. Chernobrov have just signed the Contract

Coupled with aetherodynamics time conception, which was suggested by Alexander V. Frolov, the works on the control of space-time parameters gain the possibility for development and commercial application. As a theoretical basis there are those N. Kozyrev works where his conception of "time density" are replaced by that of "aether density" according to Frolov.

In September 2002, Faraday Labs Ltd Company plans to complete testing of the first experimental system, and to start the patenting and research of applied aspects, first of all in medicine.

Physical Principles of the Time Machine

Alexander V. Frolov

Experimental success of research team headed by Dr. Vadim A. Chernobrov, Moscow was reported in [1]. The time course can be controlled as rate of any process in local space-time (inner space of the Time Machine). It can be decelerated or accelerated by means of special "converging electromagnetic waves". Ordinary waves move from the source whereas special "converging waves" move to some central point, i.e. into the focus of the system. In Chernobrov's design of the Time Machine this process is organized by means of several spherical envelops, which consist of several electromagnets. Electronic control unit controls the processes in this design. Dr. Chernobrov reported about 3% change of the time course in 4th version of the system, which was tested with a human inside. The goal of Dr. Chernobrov's work is to research the medical aspects and experimental investigation of the principles. Several important conclusions were obtained from the project: the time course can be controlled and character of the changes is different for acceleration and deceleration mode.

Other known publication and research projects on the same topic seem to be very far from any commercial and practically useful application. Obviously the topic is very new and fantastic for most of scientific community and at first we have to clarify the physical principles of the time control project, which is started by Faraday Labs Ltd.

In this project we believe that notion of time is one of possible description of real physical properties of our Universe. So, it is not mathematical abstraction but some aspect of physical reality and we can discover some physical properties of time. Russian astrophysicist N. A. Kozyrev [2] developed a theory of active properties of time and according to his point of view there are two properties: time course and time density. Prof. Kozyrev demonstrated experimentally that time density in area of some process (changes of matter) is dependent on entropy parameters of the processes. In [3] it was demonstrated that Kozyrev's experiments could be interpreted in aether theory and it has led to simple physical conclusions and clear experimental perspectives: **time course and its density can be explained and controlled as parameters of aether.** Directions of aether flow and density of aether are subjects of experimenting with non-reversible changes in matter, for example, in crystallization or melting processes. Also it is possible to use special electromagnetic processes, for example, Chernobrov's "converging waves" or other longitudinal waves as methods of aether compression or rarefaction. If we assume that process of existence of elements of matter physically can be explained as aether vortex processes then its rate is a parameter of **aether income/outcome balance** (aether inflow in element of matter and aether outflow from the element of matter). It was also described in Time Rate Control (TRC) theory [3]. To control this balance it is necessary to develop technology of longitudinal waves generation, its focusing and resonance effects. The previous research and experimenting on the topic has been made by N. Tesla.

Let's assume that we have some technology to change parameters of time course. How should we organize this local space-time (what is spatial topology of the design)? There is a very interesting experiment to get the answer: rotation of a heavy cone (for example, lead cone) entrains surrounding aether, so a vortex appears, which is a toroidal formation of aether (rings). The rings can exist in space for a long period. The further question is: Why does the beam of light (laser beam) directed to the cone by tangent create a luminous ring? We can assume that due to natural properties of photons (light propagates along the geodesic line in space) some autonomous closed toroidal space should be created in such experiment. The next thought is: since space and any matter exist in time then we can speak about some autonomous time. The general conclusion is to be the following: autonomous 4-dimensional space-time can be created as toroidal aether vortex.

Here is point to note some aspects of research project by Prof. Robert Mallett, Connecticut University, USA. In fact, sometime next year, he hopes to produce the first piece of technology that eventually will allow him to build a time machine. By Mallett it will be a device that employs lasers "to twist space". Why is he going to close the beam of light? His theoretical background is knowledge about black holes, i.e. understanding of the connection between gravity and curvature of space-time. In Einstein's theory both matter and energy can bend space and time. So Prof. Mallett assumes that curvature of space-time can be changed not only by mass (like a black hole) but it can be affected by energy of photons. This has led Prof. Mallett to consider the possibility of using a circulating beam of light to twist space and to create closed loops in time. It is predicted that a spinning neutral particle, when placed in the ring, is dragged around by the resulting gravitational field [4]. From the

Kozyrev-Dirak Radiation

Its influence on animals

Dr. Ivan M. Shakhparonov

International Academy of Energy-Informational Sciences

In the experiment with animals, that were made in the Center of Oncology Researches (COR) at the Russian Academy of Medical Sciences (RAMS), on the applying of Kozyrev-Dirak's Focused Radiation (KDFR), it has been found that KDFR decreases the quantity of glucose in the blood, reduces its tenacity, promotes the strengthening of immunity and the rise of the quantity of marrow cells.

Introduction

This time researchers in Russia and abroad experiment on ball lightnings by means of nonoriented circuits, which are similar to the electric analogues of Mobius band, also by means of Klein bottle and their combinations. Non-oriented fields are investigated very intensively now. Accordingly, organisms of the researchers, who observe the interactions of such fields with a matter, are also changed, thus they should take it into account on making such experiments. The aim of the article is to show in which way the fields of nonoriented circuits influence on animal and human organism. Besides the article has for its object the prevention of negative consequences, which can appear for experimenters through the research process.

Experiments with animals that were carried out in 1992-1993 in Russian Academy of Medical Sciences (RAMS) had not been published in proper time because there were no quantitative methods of radiation detection. Later, in 1996 they were developed [1] and KDFR parameters were measured in that geometry, which were applied in RAMS. In 1998 powerful and super-powerful KDRF sources were obtained. These sources were applied (and are applied now) in the researches at the controlled radioactive decay [2]. Kozyrev and Nasonov [3] and later also Lavrentyev with the collaborators [4,5] have proved experimentally that the Sun and some stars generate the radiation, which has early unknown properties. We suggest that the radiation, discovered by Kozyrev [3], and the radiation, which is researched by us and by other experimenters with nonoriented circuits, are of the same phenomenon. At first, it should be noted that on interaction between Kozyrev-Dirak radiation (KDR) and a matter made it colder. As it was demonstrated above [1], cooling effect can be explained by matter re-magnetization under the influence of KDFR beam (adiabatic demagnetization). According to the still unpublished data, KDFR bunch destroys matter lattice by the way of it's moving. However, after a couple of week matter reconstructs it to the almost tabulated points, without defects, blockness and other damages, which are peculiar to other natural crystal structures. **In the definite sense nature demonstrates the way to rejuvenate compound structures**. As it is well known, vital functions of biological systems on the Earth depend on the structure and composition of water. Therefore, we have a right to expect considerable changes in the vital functions of biological organisms under the influence of KDFR.

Experimental Devices

In experiments with animals there were applied the devices, which concentrated KDR (KDCR) and had 50 Wtt aggregate electrical power. The description is presented in [2].

KDFR indication was obtained by calorimetric method [1], along the way of movement of the main bunch (with 10 cm across diameter) and at angle of 45° from the geometrical axis of a device.

Researches of Bleeding Duration

Let us consider KDFR influence on the blood composition of animals. At the experiment 24-28 gram weighting, pelletized fed male mice were used. In the process of the experiment it was discovered that 3 and 4 hour processing of mice with KDFR at the distance 2.5 m and at the presence of animals in the sphere of maximum radiating power, caused some changes of fibrillation system. The bleeding duration was determined according to Duke method. Two groups of animals were used at the experiments: a group with 4-hour duration of KDRF processing and a group with 6-hour duration. Time of bleeding was considered in dynamics at 1, 2, 7, 14, 21, 28 and 35 day (Fig. 1). The bleeding duration of the intact animals was determined by the value 128±11 sec. After the applying of KDFR there was noticed some increase of bleeding duration to 261±15 sec and 223±21 sec on the first day after the stopping of the influence. In the subsequent periods bleeding duration gradually decreases up to the level of physiological norm. The whole normalization of the index is observed at the animals, which were processed by KDFR during 4 and 6 hours, on 28-35 day up to 115±12 and 133±18 sec correspondingly. In the process of observations at the animals, the correlation between time of fibrillation and periods of KDRF processing of the animals has not been revealed (Fig. 1).

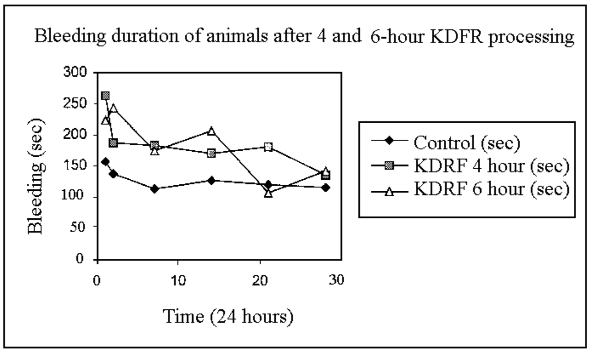

Fig.1

In the course of the experiment the strongly marked chronometric hypocoagulation was discovered due to the extension of the parameter "K" or, probably, because of the change of aggregation properties of platelets (Table 1).

For the determination of biochemical indexes serum was obtained from 5-8 ml of rats' venous blood. Beforehand (24 hours before the slaughter), forage was taken away from the animals. The determination of biochemical indexes was provided by means of biochemical analyzer HITACHI. As a result of the experiment it was determined that at the first day after influence of KDRF there was a tendency of decrease of the glucose content (Table 2). Other indexes varied in the limit of physiological norm.

Table 1 Parameters of thromboflexogramm after KDFR, 4 hours

Animal #	Parameters of thromboflexogramm			Fibrio gene	Fibrinal activity
	R (sec)	K (sec)	Ma (mm)		
1	72	∞	10	-	-
2	102	∞	18	-	-
1	90	∞	10	275	75
2	180	150	52	315	90
1	180	∞	5	-	100

Table 2 KDFR influence on the glucose content in blood of the rats

Time (days) after the experiment	KDFR 4 hours (mmole/l)	Test (mmole/l)	KDFR 6 hours (mmole/l)	Test (mmole/l)
1	3.14	6.12	6.39	7.27
10	7.59	9.35	8.90	6.69
30	6.05	6.69		

Research of haemopoiesis system

Several criteria were considered: the dependence of biological effect on the distance, on the power flux density, on the duration of processing. Besides, KDRF influence on mice survival was considered.

At the experiment 24-28 gram weighting, pelletized fed male mice were used. The marrow was examined in the dynamic at 1, 3 and 7 day after KDFR influence. Six animals were taken on each point. After the decapitation of the mouse their thighbones were taken out and after that the absolute number of myelokaryocytes was calculated by the standard method in Goryaev chamber.

Dependence of the biological effect on the distance

In all experiments the maximum flux density along the geometrical axis of KDCR device was a constant. There were used four temporal modes of the influence (1, 2, 3, 4 hours) and three points of long distance between KDSR and the biological object (0.5; 1.5; and 2.5 m). At 0.5 m distance there were no differences in the number of marrow cells in comparison with the control cells. With the increase of distance between KDCR and the object from 0.5 to 1.5 m some tendency to the increase of the number of marrow cells up to the 7th day was observed. Four-hour KDRF processing caused the increase of the number of karyocytes up to $29.99 \pm 1.25 \times 10^6$ ($P<0,001$). Thus, any dependence of the biological effect on the exposition has not been revealed. For instance, at the 7th day after one hour of the exposition the number of karyocytes was equal to $28.45 \pm 1.87 \times 10^6$ at the same time after 3-hour processing it came to $27.65 \pm 0.74 \times 10^6$. Alongside with the change of the distance to the biological object from 1.5 to 2.5 m the tendency towards the increase of the number of marrow cells has kept within the same limits $28.27 \pm 1.32 \times 10^6$ and $29.57 \pm 0.88 \times 10^6$.

Dependence of the biological effect on the power flux density

The comparative investigation of KDFR influence on the biological object in the coverage of KDCR (along its geometric axis) and outside the coverage has demonstrated that alongside with the increase of the radiation intensity there was a tendency towards the decrease of stimulative influence of KDFR on haemopoiesis.

Dependence of the biological effect on the duration of processing

On processing the animals at distances up to 2.5 m from KDCR and on increase of exposition to 3-4 hours it is possible to obtain reliably significant difference in the number of marrow cells from the physiological norm to the 7th day.

KDRF influence on the survival of mice

The experiments, determining the survival reaction of animals, were made by means of gamma radiation. 30-day survival is the criterion of determination. Conditions for the experiments are the following: in the coverage of KDCR and aside the coverage, (the distance between the KDCR and the object is 2.5 m in the coverage of KDCR and 0.5m outside the coverage). Time of influence is 4 hours. Animals of both sex were used. Two groups of animals were used. The test group was put to the gamma radiation in the diapason of doses, which caused marrowy syndrome, i.e. from 7.5 to 8.5 Gr. The second group of animals after the irradiation in the same diapason of doses was repeatedly processed with KDFR. Time of the influence is 4 hours at 7.5 Gr gamma radiation on 15 mice in one bath and 8.5 Gr on 15 mice in another bath. Total gamma radiation of the animals was made by means of the source ^{137}Cs with the dosage rate of 5.2 Gr/min. Gamma radiation in this dose diapason causes death of the animals during the development of the marrowy syndrome, i.e. from the 6th to the 20th day along with the aplasia of haematogenic tissue. Combination of gamma radiation and KDFR sometimes leads to the slight increase of the number of survived animals. If the animals are irradiated by gamma rays at first and then by KDFR, the death control at 7.5 Gr radiation is equal to 5.5% from the total number of the animals and at the following KDFR processing 16% of the irradiated animals die. However, 67% of the animals in the tested group have died after KDFR processing and after the coming next gamma radiation with 8.5 Gr total dose. And in the group, which was processed with KDFR, only 46% of the animals died.

Immunity strengthening

For the investigation of KDFR influence the following tests were chosen: activity of natural killers and T-killers, which had been obtained by the immunization in vitro in the unidirectional mixed culture of lymphocytes and also in the reaction of blast transformation on the specific mitogen [6, 7]. All tests were made on the 7th day after a single KDFR influence. Unfortunately, data have been obtained with the applying of radioisotope preparations. Though the experiments of this kind were successful and though they have demonstrated the increase of some immune reactions' level, there is a certain doubt in the relevancy of radioisotopes application [2]. Thus the series of experiments was made. These experiments were aimed at the investigation of KDFR influence on the development of the swelling process. The aim of the experiment is the investigation of KDFR influence on the development of Ehrlich cancer and sarcoma-37, which were repeatedly inoculated to mice. At the first stage of the experiment there was a single KDFR influence on the mice repeatedly inoculated with sarcoma-37, on the 2nd day after the repeated inoculation of the swelling cells to the animals. The repeated inoculation was made intramuscularly in a right thigh, in a dose 10^6 of cells per a mouse. Time of KDFR influence for the first group of animals was equal to 1 hour; for the second one it came to 2 hours; and for the third one it came to 4 hours. Each group consisted of 6 animals. The test group consisted of six mice with repeatedly inoculated sarcoma-37 and which had not been processed with KDFR. As a result, the average lifetime of tested animals was equal to 9 days. The average lifetime of animals of KDFR groups came to 48 days (for 1-hour KDFR influenced group); to 12 days (for 2-hour group); and to 31 days (for 4-hour group). Thus, the average lifetime of the experimental group came to 29 days. Besides, in the group, which has been processing with KDFR during 1 hour, **the half of mice had survived (three of six mice).**

At the second stage of the experiment the repeated (5 times during 2 hours) KDFR influence on the mice was applied. These mice have been inoculated with sarcoma-37 at seven days before the beginning of the influence. As a result, the average lifetime of the animals was equal to 27 days, and for the mice, which were processed with fivefold KDFR influence, the average lifetime was equal to 76 days. The obtained results are the evidence of inhibition of swelling development for the animals, which were processed with KDFR influence. This leads to the increase of lifetime of such animals in comparison with the test. **Thereby, at a great extent the results of the previous experiments on the strengthening of immune status after KDFR influence were confirmed.**

Results and discussion

Let us make a conclusion. At the KDFR influence on animals' organism the following effects were observed: decrease of blood viscosity; strongly pronounced hypocoagulation; decrease of contents of glucose. Increase in the number of karyocytes and the extended lifetime of the animals, infected with Ehrlich cancer and sarcoma-37 were also observed.

As for human being, the researches in this area have not been carried out yet and they are still confined to the single observations. It is possible to give an example from the author's practice. In 1975 nonoriented circuit of 3kWtt power was examined. Field strength was measured. The author of the article had been working in the field for about 8 hours. And after five hours after the experiment I had felt bad. That time it was nothing known about the influence of the new radiation on human organism. The arrived ambulance has quickly diagnosed that I was close to hypoglycemic coma. On several hours after the intravenous glucose injection, my state has become normal. **Now we know that before the experiments with powerful KDFR bunches it is necessary to eat sugar.** Thus we believe that the data, which were obtained after the experiments with animals, can be applied to a human being. We can suggest that the manifestation of the symptoms of the KDFR influence on human organism depends on the power of the applied source, on the total mass of the organism and on the time of it in the coverage of the irradiation. From aforesaid it is clear that the experiment with powerful KDFR sources is far from being harmless and it is better to make it distantly after exclusion of man presence near experimental stands and devices. At the same time it is quite obvious that on applying of small capacity and fixed time of irradiation it is possible to develop methods for curing of human diseases, which are considered now as incurable (for instance of diabetes, some diseases of haematogenic system, of cancer and possibly of AIDS.

References

1. Proceeding of the International Scientific Conference "New Ideas in Natural Sciences" Problems of Modern Physics, p. 176-187
2. Journal of new energy, Vol. 3, #4,1999, I.M. Shakhparonov "Interaction between Kozyrev – Dirak radiation and radionuclides", p. 85-89
3. Kozyrev N.A. Selected works, Leningrad State University, 1991, part 3
4. Lavrentyev M.M. Eganova I.A. Luzet M.L., Fominykh S.F./ / Reports by AS USSR. – 1990, Vol. 314, #2, p. 352-355
5. Lavrentyev M.M., Gusev V.A., Eganova I.A., Luzet M.L., Fominykh S.F.// Reports by AS USSR. – 1990, Vol. 315, #2, p. 368-370
6. Methodological materials on experimental pharmacological and clinical trials of immune modulating effect of pharmacological remedies. - Ministry of Health USSR, M., 1984
7. Talmadge J.E. and Chiragos M.A. Comparison of immunomodulatory and immunotherapeutic properties of biologic response modifiers. Springer Seminar Immunopathol., 1985, 8, 429-443

Effect of Magnetic Blow Wave Field on Wine Systems

I.M.Shakhparonov (Corresponding author), S.A.Grin,
S.R.Tsimbalaev, L.N.Kreindel, V.N.Kocheshkova,
A.I.Podlesny, S.Yu.Gelfand

AGD Firm, Peschanyi pereulok, House No.20P korpus No.1, Lfl. 33
125252, Moscow, A- 252, Russian Federation

Russian Institute of Canning Industry, Shkolnaya Street. 78.
142703 Vidnoe 3. Moscow Region, Russian Federation

Introduction

Authors communicate the data on influence of Magnetic Blow Wave (MBW) field on several wineproducts. It was found, that MBW did not lead to significant changes in the major components of the wineproduct (sugar, organic acids, minerals). At the same time the taste and aroma of treated wine become more pleasant; content of heavy alcohols and wine stone in the treated samples was less than in non treated ones. A mechanism of transformations was also discussed.

Keywords: *Magnetic Blow Wave (MBW), Wineproduct, GLC of aroma compounds and ethanol, HPLC of sugars, Atomic Absorption Spectrometry (AAS) of minerals, Heavy alcohols and aldehydes, Wine stone, Turbidity tendency, Organoleptic evaluation*

Magnetic Blow Wave (MBW) was obtained for the first time during the investigations on ball lighting generation under the laboratory conditions (Shakhparonov 1994). MBW as a physical object is interesting because of some facts, which suggest that MBW is a magnetic monopole. The MBW can also interact with the matter and transforms it in a definite way. Typical example is an elementary carbon in the form of graphite, which is transformed by such magnetic treatment into ferromagnetic substance (ibid). The graphite, which is initially diamagnetic, transforms to paramagnetic one with general radiation doze of about $7 \cdot 10^{19}$ neutrons/cm^2. Other types of radiations could not affect this way (Svoistva 1975). So one unit of MBW can be considered as $1 \cdot 10^5$ of neutron masses. This fact may be regarded as an indirect evidence for assuming that MBW and magnetic monopole are the same things. In the absence of excited radioactivity a slow MBW [$v/c < 1 \cdot 10^4$] occurs, which does not ionize atoms (Devons, 1963). Therefore, their interaction with the matter can be observed only indirectly. No data exist on the interaction of MBW with organic substances. The experiments and results reported in the present communication may be a starting point for development of technology and to formulate the methods for vintage wine and best quality spirit production.

Materials and Methods

Assuming that MBW and magnetic monopole are the same things, a number of conditions were selected for all experiments. The MBW source and the samples were placed in the same axis and the axis was oriented according to magnetic meridian direction. Such magnetic orientation is appropriate, as the energy of magnetic monopole theoretically increases in a magnetic field (Devons, 1963). All of samples were placed at 250 cm distance from MBW source, in hermetically closed glasses. It should be noticed that MBW could penetrate through many other barriers, for example into cast iron reservoir with wall thickness of 5 cm (Amaldi, 1970).

The quality investigations were made by using of standard equipment. HPLC, equipped with refractometric detector was used for sugars estimation. Separation of organic acids in forms of their ethyl esters and acid esters was carried out chromatographically using a column packed with polyethylenglycol succinate and the following temperature option: initial temperature is 120°C, final temperature is 220°C, temperature growth rate: 8°/min. GLC was also employed for determination of ethanol. Minerals content conclusion: in order to obtain complete information about any system, it should be destroyed. However, destruction of tissues of the man in order to get information about their state is a too high price to pay for the information about his health.

However, the above Van Hoven's criterion can be satisfied with the minimum influence, when the cells are not destroyed and the atoms of these cells, being primary sources of torsion spectrums to be registered, are bring into the non-equilibrium state by means of outer disturbing influence.

In order to choose the frequency of the disturbing torsion influence correctly, it is necessary to take into account the role of water in physical and biochemical organization of tissues of the human organism.

At the same time, it is necessary to take into account the resonance torsion frequencies of various human organs. Finally, it turns out that the signal of torsion disturbance should be rather sophisticated considering both these factors. The TORDI system is a ready-to-use production device. Nevertheless, it is important to understand that the model is not the limit of scientific and technical potential incorporated in it and that enhanced variants of the system will appear with the course of time.

Summing up, I would like to draw your attention once more to the fact that work on torsion technologies is not limited by the directions that were discussed here. Actually, as it was pointed out in the beginning, ongoing development includes all branches of economy, industry, agriculture and medicine, as well as all problems of everyday life. Technologies that we mentioned are the forerunner of the fact that the mankind is on the threshold of the age of torsion technologies, which, we believe, will change our life in the 21st century more than all the scientific and technical revolutions of the 20th century.

Reference

1. Cartan E. Compates Rendus. Akad.Sci., Paris, 1922, v.174.
2. V.Melnikov, P.Pronin. Problem of gravitation constant stability and additional interactions. Itogi Nauki I Tekniki, ser. Astronomy, v.41, Gravitation and Astronomy, Moscow, VINTI, 1991.
3. G.I.Shipov Theory of the Physical Vacuum. Nauka, Moscow 1997.
4. I.Ternov, V.Bordovitsin. On the modern interpretation of classical spin theory of Ya.Frenkel. UFN, 1980, v.132, No.2.
5. V.Bagrov, V.Bordovitsin. Classical spin theory. Izvestiya VUZ, Phys.Series, 1980, No.2.
6. F.I.Belinfante. On the spin augular momentum of mesons. Physica VI, 1939, v-6, No.9.
7. M.Markov. The very early universe. Proc of the Nuttfield Workshop, Eds. G.V.Gibbson, S.W.Hawking, S.T.Siklov, Cambridge, 1988.
8. Hideo Uchida.A method apparatus for detecting a fluid. Patent England, No 511662, May 24, 1978.
9. Anatoly Akimov. Heuristic discussion of the problem of finding long-range interactions. EGS-concepts. CISE VENT, preprint N7A, Moscow, 1992.
10. IITAP RANS, TORTECH USA, Horizonts of the Science and Technology XXI age Proc, Editor A.E.Akimov, Folium, Moscow, 2000, vol.1 (in Russian).
11. N.Kozyrev, Astronomical observations by means of the physical properties of time. In "Flarestarse" International Symposium in Byurakan, 1976,Armenian Academy of Sciences Publ., Yerevan, 1977 (in Russian)
12. Bouwmecster D. et.al. Nature, v.390, 11 dec, 1997.
13. G.I.Shipov. Theoretical and Experimental Research of Inertial Mass of Four-Dimensional Gyroscope. ITTAP RANS, preprint N10, Moscow, 2001, (in Russia).
14. The way of correction of metal alloy microstructure. Patent Russian, RU 2107105, 1998.
15. A.Dolgov, Yu.Zel'dovich, M.Sazshin. Cosmology of the Early Universe. MGU Publ., Moscow, 1988.
16. I.A.Wheeler. Einstein's vision. Springer Verlag, 1968.
17. Convegno Internazionale:Quale Fisica per 2000, Proc. Bologna, 1991. see: The Manual of Energy Devices and Systems.
Complied D.A.Kelly, D.A.K. WLFUB, Burband, California, 1986, Publ. N1269/F-289.
18. Daytlov V.L. Polarization Model of the Heterogeneous Physic Vacuum. Institute Mathematical, Sibirians Academic Science, 1998 (in Russian).

The Electrical Vortex Non-Solenoidal Fields

Sergey B. Alemanov
E-mail: alemanov@inbox.ru Phone 7 (095) 323-6848

A mistake was found in the electrodynamics: it is detected that all electrodynamics' postulates corresponds to the experimental facts, but vortex electric fields has unclosed inductive lines.

When the magnet is moving, then the current of magnetic induction is moving together with it. From known velocity of motion **v** and the value of magnetic induction **B**, it is possible to calculate the intensity **E** of appearing vortex field according to electrodynamics formula of transformation of fields **E=vB**.

If to change the E=vB on induction $D=\varepsilon_0 E$ in formula of fields' transformation, that will get $D=\varepsilon_0 Bv$, where D is electric induction, B is magnetic induction, v is velocity of motion, ε_0 is electric constant.

Herewith the appearing electric induction is always transverse to the direction of motion. It is possible to formulate the rule of origin for electric induction under the condition of rectilinear motion: if to dispose the right hand palm so four fingers shows the motion direction of the magnetic flow (the field), connected with moving magnet, and the vector B fells into palm, then the moved aside big finger will indicate the direction of vector D. The given rule is like the rule for Lorenz' force, but on the contrary (the difference is in frame). In the first case the charge moves, but the magnet rests. Here the magnet moves, but the charge, which points the direction for lines of force of electric induction, is immovable. So, there it is the rule for left hand, but here, on the contrary, it is the rule for right hand. Thereby, if the charge moves, but the magnet is immovable, then the rule of left hand uses for determination of the force. But if the magnet moves, but the charge rests, then the rule of right hand uses for determination of the force.

The origin of electric force is connected with that, the vortex electric field $D=\varepsilon_0 Bv$ appears around moving magnet (the magnetic field does not act on immovable charges).

In common literature on electrodynamics there is no any difference between electric vortex field and solenoidal field, but these are different notions. The sign of solenoidal field is the closed lines of electric induction (the flow of vector D through the closed surface is a zero), but for the vortex field the sign is following: the work of forces can be different from zero under the condition of motion along a closed line. That is to say, the vortex fields can agitate the rotational currents.

From the electrodynamics textbook: "The work of forces of vortex electric field can be different from zero, when the electric charge is moving along a closed line."

For instance, when the magnet moves, the vortex electric field appears and this field can be solenoidal or not, depending on magnet's orientation. Let's take such example: the magnet moves evenly, rectilinearly, and it's poles are oriented transversely to direction of motion. According to the rule of origin for electric induction ($D=\varepsilon_0 Bv$ that is the rule of right hand), the appearing vortex electric flow is not a solenoidal, since the lines of electric induction are not closed. Its begins in one conditional area of disturbance (+), accompanies the moving magnet, and it finish in another area of disturbance (-). For presentation it is enough to consider only two areas (+) and (-), represented on Fig.1. These dissimilar areas of disturbance appears because that flow of magnetic induction inside the magnet has the inverse direction, that outside the magnet.

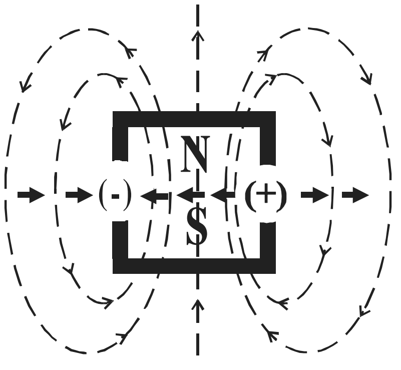

Fig.1

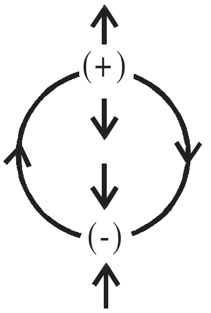

Fig.2

That moving disturbance of electric and magnetic fields presents itself as transverse electromagnetic disturbance. Also, it is necessary to notice, that under such magnet's motion, the appearing vortex electric field is not closed, but the current of electric displacement, connected with it, is closed (a currents are always closed). In given example, for clarity, it is possible to present a intensity of electric field through the Lorenz' force, if to take the frame, in which the magnet rests, and the test charge moves.

On the Fig.1 the moving magnet is conditionally represented (motion is toward to the text, magnet is moving away). N and S are poles of magnet. The direction of lines of electric induction, appearing when the magnet is moving, specified by arrows → and ←. Part of the lines begins in positive area (+) and finishes in negative area (-), the areas are placed on the ends of magnet. **The flow of electric induction through closed surface is not a zero; that is to say, these areas of disturbance are moving electric charges.**

From the electrodynamics textbook again: "The flow of vector D through any closed surface is equal to algebraic amount of external charges, covered by this surface. In the electrodynamics these postulates has the same role, as Newton' laws in classical mechanics."

Thereby, according to postulate, it is necessary to consider the appearing dissimilar areas of disturbance (+) and (-) to electric charges, or it is necessary to change the postulate.

It is interesting, that a part of lines of electric induction, which placed frontal and behind magnet, starts and finish at infinity, since the distribution of magnetic induction around magnet has not determined borders.

For clarity, it is possible to make following calculation. For instance, the coil (loop or turn) with current, as a magnet, moves evenly and rectilinearly, but its magnetic poles are oriented transversely on motion direction. Under such motion the lines of electric induction are not closed, and the dissimilar areas of electric field's disturbance appears in space on the edges of this coil.

On Fig.2 the moving coil with current is conditionally represented. It moves from left to right side of the page. The arrows on the coil indicate the direction of current. The appearing dissimilar areas of disturbance of electric field are marked by signs (+) and (-). Knowing, that in medium of the coil $B=\mu_0 I/2r$ and according to $D=\varepsilon_0 Bv$, it is possible to find the electric induction, appearing in the center, between two dissimilar areas $D=\varepsilon_0\mu_0 Iv/2r$, where I is current in the coil, r is radius of the coil, v is velocity of motion, ε_0 is electric constant, μ_0 is magnetic constant. The electromagnetic disturbances in transverse electromagnetic waves has the similar field construction, there also dissimilar areas of disturbance of electric field exists, that is to say the lines of electric inductance are not closed. Only the currents of electric displacement and magnetic induction are closed.

Let's consider another example: magnet moves rectilinearly, but its poles are oriented longitudinally to direction of motion. According to the rule for origin of electric induction ($D=\varepsilon_0 Bv$ is the rule of right hand), the appearing rotational electric flow is solenoidal, since in this case the inductive lines become closed lines. **Usually in books on the electrodynamics such moving magnet is considered, and the wrong conclusion is thereof done, that vortex electric field is always**

solenoidal, herewith it is forgotten, that poles of the magnet can be oriented not only along the direction of motion, but across also.

From the electrodynamics textbook: “The vortex electric field differs from electrostatic field that it is not related with any electric charges and its lines of intensity are closed lines.”

From theory and from experiments it follows, that under transverse motion of magnet the lines of disturbance of vortex electric field can be unclosed and, accordingly, the flow of induction through the closed surface is not a zero. Then there is a direct discrepancy to facts in modern electrodynamics. It is strange, but for the whole history of researches in magnetism the transverse magnet’s motion was not is considered. It leads to revising of electrodynamics’ postulates, which plays such role in electrodynamics, as the Newton’s laws plays in classical mechanics. The postulates, giving invalid belief about field processes, accordingly, do not allow to make some correct calculations. Fallaciousness of these postulates was one of the reasons, on which the electrodynamics could not to consider and to calculate the discrete electromagnetic waves (photons), where the magnetic field also is the transverse field (the field construction and calculation of photons are represented on the page http://www.comail.ru/~alemanov). **That is to say, not only particles has the charges, but areas of disturbance of field (without particles) are the charges also,** where the flow of electric induction through the closed surface is not a zero. Thereby, the vortex electric fields can be not only as closed flows of induction, but as well as inducted electric charges, accordingly, the laws for electric charges are valid for induced electric charges also. For instance, in the law of conservation of charge: if somewhere the area of disturbance with positive sign appears, that negative area appears also.

From the electrodynamics textbook: “The vortex electric field is generated by the variable magnetic field. Its force lines are always closed, like force lines of magnetic field.”

But before this fundamental postulate, confirming, that force lines of vortex electric field are always closed, it was necessary to consider all variants of change for the magnetic field, including the variant of the transverse motion of the magnet. That is to say, the consideration of physical processes could not be unilateral. Faraday considered the longitudal motion of magnet and discovered the electromagnetic induction, but the transverse motion of magnet that have the principle importance for understanding of field processes in electrodynamics was not considered. Thereby, the longitudal motion of magnet brings to arising a vortex electric field with closed force lines, but transverse motion of magnet brings to arising a vortex electric field, where the lines of forces are not closed. In this case it lead to induced electric charges. It is necessary to notice, that this is first mistake, detected in electrodynamics postulates for all time of existence of electrodynamics.

From the electrodynamics textbooks: “...Gauss’ theorem is valid not only for electrostatics, but also for electrodynamics, which using a variable in time electromagnetic fields. We are not sure if this hypothesis is valid or it is not valid... Only the experiment can give the answer on this question. The whole collection of experimental facts speaks in favor of this hypothesis.” But, unfortunately, the experiment with transverse motion of magnet was not considered seriously in this textbook.

(Editor’s note: Well-known Searl’s experiments and Godin & Roshchin’s experiments are based on such transverse motion of magnets (rollers). In Alemanov’s article it was demonstrated that in this case the experiment should lead to induced electric charges. Really it was detected in experiments. Hence this missed aspect of electrodynamics is very important for development of the new energy technologies.)

Gravito-Inert Mass

J.A. Asanbaeva

720000, Kyrgyzstan, Bishkek, Kadyrov’s Scientific Center
+996 (312) 47-25-40, +996 (312) 65-02-83
bondarenko@nazaraliev-centre.com

Nature of mass is one of the important problems of modern physics. It is accepted to consider that the mass of elementary particle is determined by fields, which are connected with it (electromagnetic, nuclear and others). However, we didn’t create any quantitative theory of mass. There is no theory to explain why masses of elementary particles form a discrete spectrum of values and to allow determining this spectrum.

Mass (m) is a physical value, one of characteristics of matter, which defines its inert and gravitational properties. Accordingly, we distinguish inert mass (m_i) and gravitational mass (m_g).

Inert mass (m_i) characterizes dynamical properties of a body, its property to accelerate under the action of the force $\left(\vec{F}_i\right)$ and according to the second Newton’s law is considered to be constant coefficient of proportionality for the given body between $\vec{F}$ and acceleration $\vec{a}$.

$$\vec{F}_i = m_i\vec{a} \qquad (1)$$

Gravitational mass (m_g) is a source of gravity field. Every body creates its gravity field, which is

18. Kozyrev N.A. *Selected Transactions* (Leningrad University Press, Leningrad, 1991) (in Russ.).

19. Lavrent'ev M.M., Eganova I.A., Medvedev V.G., Olejnik V.K., and Fominykh S.F. *On Scanning of Celestial Sphere by Kozyrev's Sensor.* Doklady AN SSSR, 323 (4), p.649-652 (1992) (in Russ.).

20. Akimov A.E., Kovaltchuk G.U., Medvedev B.P., Oleinik V.K., Pugatch A.F. *Preliminary Results of Astronomic Observations of the Sky by the Kozyrev Procedure.* Academy of Sciences of Ukraine, Central Astronomic Observatory. Preprint CAO-92-5P, 1992, p.16 (in Russ.).

21. *Oleinik V.P. Superluminal Transfer of Information in Electrodynamics.* SPIE Material Science and Material Properties for Infrared Optoelectronics. 3890 p. 321-328 (1998) (http://www.spie.org /).

22. Oleinik V.P. *Faster-than-Light Transfer of a Signal in Electrodynamics.* Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999), p.261-281.

23. Eganova I.A. *The World of Events Reality: Instantaneous Action as a Connection of Events through Time.* Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999).

24. Lavrent'ev M.M. and Eganova I.A. *Physical Phenomena Predicted and Revealed by N.A.Kozyrev, in the Light of Adequacy of Space-Time to Physical Reality.* Philosophy of Science, 1(3), p.34-43 (1997) (in Russ.).

25. Logunov A.A. *Lectures in the Theory of Relativity and Gravity.* A Present-Day Analysis of the Problem. (Nauka, Moscow, 1987) (in Russ.).

26. Oleinik V.P. *The Newest Development of Quantum Electrodynamics: Self-Organizing Electron, Superluminal Signals, Dynamical Inhomogeneity of Time.* Physical Vacuum and Nature. 4 p.3-17 (2000) (in Russ.).

27. Jefimenko O.D. *Electromagnetic Retardation and Theory of Relativity* (Electret Scientific Company, Star City, 1997).

28. Oleinik V.P., Borimsky Ju.C., Arepjev Ju.D. *Time, What is this? Dynamical Properties of Time.* Physical Vacuum and Nature. 5 p.65-82 (2001); *New Ideas in Electrodynamics: Physical Properties of Time.* Semiconductor Physics, Quantum Electronics and Optoelectronics. 3 N4 p.558-565 (2000). E-print: quant-ph/0010027.

29. Borimsky Ju.C., Oleinik V.P. *The Course of Time in Classical and Quantum Systems and Dynamical Principle.* Physical Vacuum and Nature. 6 (2001) (in print).

The Principle of Self-Organization, which can be formulated as follows: any material object represents an open self-organizing system whose internal structures are formed with the participation of the whole universe. Apparently, the Principle of Self-Organization, incorporated in nature as one of the integral properties of matter, is nothing more nor less than a spirit (or absolute idea, or creator) which operates the world and creates all its variety.

Physical Mechanism of Nuclear Reactions at Low Energies

V.P. Oleinik* and Yu.D. Arepjev

> *Tell me what the electron is,*
> *and I shall explain to you everything else.*
> *W. Thomson*

Abstract

The physical mechanism of nuclear reactions at low energies caused by spatial extension of electron is considered. Nuclear reactions of this type represent intra-electronic processes, more precisely, the processes occurring inside the area of basic localization of electron. Distinctive characteristics of these processes are defined by interaction of the own field produced by electrically charged matter of electron with free nuclei. Heavy nucleus, appearing inside the area of basic localization of electron, is inevitably deformed because of interaction of protons with the adjoining layers of electronic cloud, which may cause nuclear fission. If two or more light nuclei occur "inside" electron, an attractive force will appear between the nuclei that may result in the fusion of nuclei. The intra-electronic mechanism of nuclear reactions is of a universal character. For its realization it is necessary to have merely a sufficiently intensive stream of free electrons, i.e. heavy electric current, and as long as sufficiently a great number of free nuclei. This mechanism may operate only at small energies of translational motion of the centers of mass of nuclei and electron. Because of the existence of simple mechanism of nuclear reactions at low energies, nuclear reactor turns out to be an atomic delayed-action bomb, which may blow up by virtue of casual reasons, as it has taken place, apparently, in Chernobyl. **The use of cold nuclear reactions for production of energy will provide mankind with cheap, practically inexhaustible, and non-polluting energy sources.**

Introduction

Nuclear reactions at low energies, occurring in physical and biological systems, and, in particular, the cold fusion (CF) of nuclei, attract ever increasing attention (see review articles [1,2]). This is explained by the fact that research on CF (in what follows, by cold fusion we shall understand any nuclear reactions at low energies) opens up the way to the solution of the problem which was set more than 50 years ago in the field of controlled thermonuclear reactions (CTR) and which has not been solved that is the problem to provide mankind with cheap fuel. An important point is that CF allows to create not only cheap, but also non-polluting energy sources, as nuclear reactions at low energies are not accompanied by radiations dangerous to health (γ-radiations, streams of fast neutrons and other particles). Note that the energetic problem facing mankind is presently of special interest in connection with the fact that, according to expert evaluations, the oil-and-gas resources in the world will suffice only for some decades. For this reason the study of CF is among the most important problems of physics.

It is necessary to note that, relying on the standard theory of nuclear reactions describing nuclear processes in vacuum, experts in the field of nuclear physics, engaged in CTR, reject the very possibility of existence of nuclear fusion at low energies. Two basic objections are raised against CF:

1. at low energies the penetrability of Coulomb barrier around nuclei is so small that the probability of nuclear fusion is practically equal to zero;
2. distinction between the atomic and nuclear energy scales is so great that the energy, which might be evolved as a result of nuclear fusion, could not be transferred directly to atomic lattice; therefore the energy above should be emitted in the form of streams of γ-quanta, fast neutrons and other particles. However, such streams of sufficient intensity have not been registered.

The answer to the first objection against existence of CF is that at the heart of CF are nuclear processes occurring in environment, and the basic role is played here, apparently, by collective effects caused by interaction of nuclei with particles of environment in which the nuclear reaction takes place. The laws governing the behavior of interacting nuclei in vacuum are inapplicable to the description of CF of nuclei [3]. Nuclear reactions occurring at low energies submit to completely different laws, which can be established only provided that collective effects mentioned above are taken into account. For this reason the standard theory of nuclear reactions in vacuum can by no means refute the existence of CF.

As to the impossibility of transferring the energy between levels of various scales, we can give an example of the phenomenon of sonoluminescence (luminescence of a liquid when a sound wave causing cavitation passes through it) [4], in which the energy transfer from an acoustic wave to electromagnetic field occurs with appreciable probability in spite of the fact that the distinction between energies of acoustic phonons and quanta of light reaches 11 orders.

As early as 10 years ago J. Schwinger, the Nobel winner and the known expert in the field of the theory of elementary particles and quantum electrodynamics, asserted that it is impossible to deny the reality of CF phenomenon [3,4]. Since then the CF phenomenon for nuclei was repeated hundreds times in laboratories all over the world, tens of patents on the ways of energy generation on the basis of CF were registered and enormous number of experimental works were published, which not only confirmed the existence of effect, but also contained its detailed analysis.

The most convincing evidence for the existence of nuclear reactions at low energies seems to give the mass-spectrometric research of reaction products [5] as well as research on biological systems [6]. Detailed study of electric explosion of foil made of especially pure materials in water, described in [5], suggests that at electric discharges transformation of chemical elements occurs. Study of optical spectrum of plasma arising at discharge and of the mass-spectrometric analysis of sediments, which remained after the discharge, shows that in plasma there is an appearance of a significant number of chemical elements which were not presented in the initial material of explosive foil and electrodes and also that the isotope structure of the foil material changes appreciably. The change of experimental conditions, for example, of energy contribution in foil, its mass and dimensions results only in redistribution of intensity of plasma spectral lines, i.e. in the change of statistical weight of chemical elements in plasma, but the composition of chemical elements remains unchanged and it essentially depends on the material of foil. As it is seen from the received results, nuclear reactions, which take place at electric discharge, are not accompanied by the occurrence of a neutrons stream and γ-radiation and proceed at low energies of atomic nuclei.

The research mentioned above as well as many others, carried out by different researchers in different laboratories, **allow to draw a conclusion that existence of nuclear reactions at low energies is reliably established.**

The development of research on CF is hampered by the absence of theory of the phenomenon. As noted by Schwinger [3,4], the situation in CF is closely parallel to that one in high-temperature superconductivity: reality of the last, as a result of careful experimental research, is completely established, though theory of the phenomenon is absent till now.

In [5], to account for the transformation of chemical elements, the hypothesis is put forward that at the electric explosion of foil in the plasma channel magnetic monopoles are formed which may overcome the Coulomb barrier even at insignificant kinetic energy due to the great magnitude of their magnetic charge. The monopole, appearing not far from a nucleus, causes its polarization: those nucleons of the nucleus, which are situated more close to the monopole, experience stronger influence of the last, than the nucleons situated on the opposite side of the nucleus. As a result, a deformation of the nucleus arises (the nucleus is lengthened), which may result in nuclear fission.

Obvious drawback of this mechanism of nuclear reactions is that magnetic monopoles have yet to be found out in nature.

Numerous attempts to construct a consistent theory of CF (see reviews [1,2]) have not been crowned with success. As it was noted above, for the CF to be described, the account of the collective effects may be important caused by interaction of nuclei with environment, in which nuclear reaction takes place. But does it suffice to take into account these effects in order that the theory of the phenomenon is constructed? The analysis of the experiments on transformation of chemical elements at low energies and on the CF of nuclei suggests that the discussed phenomenon does not fall within the domains of exotic ones: it seems to occur in nature constantly, at every step, in both physical and biological systems. Therefore, it is natural to expect that nuclear reactions at low energies should have a simple physical explanation.

However such explanation, which is not beyond the scope of existing representations, is yet to be found. Does not it mean that we are facing here the situation similar to that which has arisen in physics at the end of the 19th century and which has been figuratively described in the words: on the light sky of physics there are only two small dark clouds – the radiation of absolutely black body and the Michelson experiments? Let us remind that in order for these clouds to be removed, it has taken the revision of physical notions about electromagnetic field as well as about space and time.

As is noted in [8], there is a simple physical mechanism of nuclear transformations at low energies which existence follows from the quantum theory of electron as an open self-organizing system [9]. If two or the greater number of light nuclei appear inside free electron, more precisely, inside the area of basic localization of the particle, because of interaction of nuclei with electrically charged matter of electronic cloud, a force of attraction appears between the nuclei which may result in fusion of nucleus. This means that **cold nuclear reaction represents an intra-electronic process** which character is defined by physical properties of the own field produced by electrically charged matter of electron. The purpose of this paper is more detailed consideration of the mechanism above stemming from the spatial extension of electron.

In section 2 physical ideas are formulated and basic results are schematically presented of quantum theory of electron as an open self-organizing system. The theory outlined is necessary to elucidate the origin of the mechanism resulting in the occurrence of nuclear reactions of fusion and fission at low energies. The essence of the developed approach consists in that **the own field created by electron is treated as a congenital, integral physical property of electron, intrinsically inherent in the particle by the very nature of things and for this reason the own field and self-action are included in the definition of the particle** at the initial stage of formulating the theory. As is seen from the received results, electron represents a quantum (elementary excitation) of the field of electrically charged matter. It is a solition, which physical and geometrical properties are described by the non-linear and non-local dynamical equation similar to the known Dirac equation.

In section 3 the application of quantum model of self-organizing electron to nuclear reactions at low energies is considered. It is noted that because of the presence of simple physical mechanism of nuclear reactions at low energies, which is of a universal character, nuclear reactors represent, in effect, nuclear delayed-action bombs, which from time to time may blow up by virtue of the casual reasons. Hence, though nuclear stations may provide mankind with energy, however atomic engineering is a very dangerous way of energy production. The only acceptable way of solving the energetic problem consists in the use of nuclear reactions at low energies.

Quantum model of electron as an open self-organizing system

The basis for the standard formulation of quantum electrodynamics (QED) is the hypothesis that electron is a structureless point particle which does not experience self-action. This assumption results in serious difficulties – the divergences of mass and charge of electron and the impossibility to explain stability of the particle (see, for example, [10-12]).

The difficulties mentioned above are very serious. According to Dirac, **the difficulties of QED "in view of their fundamental character can be eliminated only by radical change of the foundations of the theory, probably, radical to the same extent as transition from the Bohr orbits theory to modern quantum mechanics"** ([13], p. 403). "Correct conclusion", Dirac emphasizes, "is that the basic equations are incorrect. They should be changed in such a way that divergences do not appear at all".

The main reason of occurrence of difficulties is the assumption that electron is a point-like particle. Therefore, abandonment of this hypothesis is inevitable. As an analysis of the problem shows, the key to constructing a consistent quantum theory of electromagnetism lies in taking account of the Coulomb self-action of electron, i.e. the back action of the own field created by charged particle in environmental space upon the same particle. In the special case that the particle is at rest in an inertial reference frame, own field of the particle turns into static Coulomb field.

E.Schrödinger who suggested the historically first physical interpretation of quantum mechanics put one of the boldest ideas concerning the problem of electron forward. According to Schrödinger's hypothesis, the quantity $e|\Psi(\boldsymbol{r})|^2$ (e and $\Psi(\boldsymbol{r})$ are charge and wave function of electron, respectively) is the density of spatial distribution of electron's charge and, consequently, **the linear sizes of electron are the same as those of atom** [14,15]. However, they did not succeed in substantiating the interpretation and, for this reason, it was rejected by the majority of physicists [16].

An important step to the correct understanding of the physical nature of electron was made by A. Barut and by his collaborators [16-18] who formulated and developed quantum theory of electromagnetic processes on the basis of self-energy picture (the Self-Field QED). Using expression for the total own energy of electron, they managed to calculate the Lamb shift and other radiative corrections and to show that radiative phenomena may be described in terms of the action function, without using the second quantization method. As is pointed out by Barut [17], "the correct quantum equation of motion for radiating electron is not the Dirac or the Schrödinger equation for bare electron, but an equation containing an additional non-linear self-energy term".

New lines of approach to the problem of electron are offered in [9, 19-24]. The formulation of electrodynamics is considered which represents a synthesis of standard quantum electrodynamics and ideas of the theory of self-organization [25]. The physical mechanism of self-organization of electron consists in self-action. Taking into account the self-action means that electron is treated as a feedback system.

Let us outline schematically the results of the formulation of quantum electrodynamics in which electron is an open self-organizing system.

Editor's note: The authors develop mathematics by using Lagrangian functions, 7 equations. You can contact the authors for more information about.

Thus, the negative result is received: we have tried to take into account self-action of electron in a natural way by supplementing the Lagrangian function with the self-energy term, but we came to an equation that has no reasonable physical solutions at all. This result seems to mean that the standard theoretical scheme reaches here the limits of its applicability and so, remaining in its framework, it is impossible to solve the problem of electron and elucidate the physical nature of electromagnetic interaction.

Essentially new point, which is introduced in [9] into quantum mechanics consists **in the replacement of the model of isolated system described by harmonic oscillator with the model of open system**. Let us advance the arguments indicating the inevitability of using the model of open system as a basis of the description of interaction between microparticles [26].

Note, first of all, that quantum particle theory based on the use of the models of isolated system is, strictly speaking, physically meaningless. Really, any observation conducted on a system represents a process of interaction of the system with the means of observation. But in case of microparticles (quantum particles) this interaction is not weak and consequently it is inadmissible to neglect it, i.e. microparticles should be necessarily considered as essentially non-isolated systems.

A starting point of the standard formulation of quantum mechanics is the physical idea that interaction between physical fields can be reduced to collision of the particles corresponding to these fields, the particles before and after collision being considered as free ones. According to these representations, quantum mechanics is based on the notions of "bare", non-interacting particles, with the interaction between them being considered as an additional factor which can only insignificantly alter the physical properties of non-interacting particles. However, such an approach to interaction between physical fields is obviously of an idealized character because particles constantly interact "with vacuum as with some kind of physical medium in which the particles move" [27]. Interaction of particles with vacuum fluctuations is not small and it cannot be removed.

It is well also to bear in mind that the necessary intermediary at studying micro-objects are the means of observations (the devices) with the classical field corresponding to them which should be taken into account in consistent quantum theory [28]. Inclusion in theoretical scheme of arbitrarily weak classical external field results in occurrence of non-zero width Γ of energy levels of "dressed" particles. The basic impossibility to isolate a real particle from vacuum fluctuations of the field and from the classical sources connected to the means of observation is indicative, thus, of necessity to take into account the non-zero width of energy levels of real particles [26].

The use of the harmonic oscillator model, when describing the interaction of electromagnetic radiation with substance, seems to be the main source of serious difficulties of the standard formulation of quantum theory, as such an approach means apparent neglect of those physical processes which, proceeding constantly, are responsible for inseparable coupling of real physical system to surrounding medium. Introducing artificial notion about switching on and switching out of interaction of oscillator with radiation field, we are able to calculate within the framework of existing theory the width of energy levels of oscillator, but we cannot assert with certainty that such an approach results in correct description of interaction.

From the reasoning given above it is seen that they are the models with energy levels of non-zero width that should form the basis for the description of interaction of radiation with substance. It is necessary to formulate such a quantum theory, which would take into account the energy levels of non-zero width Γ. The case in point is that one should introduce an infinitesimal damping Γ into the initial set of equations describing interaction of charged particles with electromagnetic field. Such an approach means the violation in infinitesimal of homogeneity of physical system relative to translations in time. Necessity of violating the homogeneity of time follows from that fact that in the usual approach (with $\Gamma = 0$) the states of the system of interacting fields have degeneracy of infinitely large multiplicity in relation to time translations. According to the fundamental Bogoliubov's concept of quasi-averages [29], when describing the behavior of degenerate systems, one should include into Hamiltonian an infinitesimal term removing degeneracy. In the theory presented here degeneracy of states of quantized fields relative translations in time is removed by introducing the infinitesimal damping Γ into Lagrangian. Thereby the degeneracy under study is removed already in the initial, zero-order approximation, which is of fundamental importance for the approach based on perturbation theory.

Formulation of the physical idea that quantum friction arises at the very elementary level - at the level of one particle is given in monograph [26]. Impossibility to isolate real particle from the surrounding world is that property which should be taken into account already in the one-particle theory (for each kind of particles), even before switching on the interaction with other particles. Model of the particle as an open system ($\Gamma \neq 0$) is attractive owing to the fact that from the very beginning the degeneracy of states relative to time translations is absent in it, the degeneracy, which is removed in standard approach by taking into account the interaction of particle with vacuum field fluctuations and classical fields. The basis for the developed formulation is the fundamental concept of quasi-averages supplemented with the requirement that the equations of motion of the particle with $\Gamma \neq 0$ follow from the action principle. It should be emphasized that the non-zero damping Γ is introduced into electrodynamics with the aim to establish the structure of the Lagrangian function, which takes into account the property of openness of physical system. After establishing the structure, the limiting transition $\Gamma \to 0$ is fulfilled.

In our opinion, the development of quantum theory will be inevitably connected with the use of models of open system; as such models reflect more completely the physical essence of interrelations in the real world. It is necessary, thus, to define more exactly the concept of openness of physical system, which, on the one hand, would describe real system accurately enough and, on the other, would be simple enough to describe the particular physical processes.

As open system has the richer physical contents in comparison with isolated system, some essentially new mathematical ideas are needed for its description. First of all, it is necessary to increase the number of independent dynamical variables describing the particle as open system. In papers [9,19-24], as a basis for the description of self-acting electron, the simplest model of open system is used which can be described by the Morse-Feshbach-Bateman Lagrangian function [30,31] and which was successfully used for the description of dispersive medium (the review of articles, in which applications of the model of open system to electrodynamics of dispersive medium are considered, is given in monograph [26]). In this model the number of dynamical variables is doubled as compared with the isolated system, namely, to each dynamical variable of "bare" particle, Ψ, there correspond two dynamical variables, which are denoted by ψ and $\tilde{\psi}$. These quantities are considered as components of the wave function describing the quantum state of self-acting particle. One of them, say, ψ, corresponds in a sense to the particle alone (to the "bare" particle) and the other, $\tilde{\psi}$, to the surrounding medium, in which the particle moves.

Editor's note: You can contact the authors directly for more information (8-16 equations).

Equation (16) coincides in its appearance with the usual Dirac equation for charged particle in an external field described by 4-potential . However, in reality, it differs essentially from Dirac's equation. The distinction consists in that **equation** (16) **is non-linear and non-local**, with the non-locality being of both spatial and time character. Potential ($A_{||}$) and vortex ($A_{\perp}$) components of the 4-potential, entering equation (16), differ from each other by their physical nature: the former describes the Coulomb field and is expressed quadratically in terms of the wave function components of electron, and the latter describes transverse electromagnetic waves and is expressed in terms of vortex electromagnetic field. As a detailed analysis shows, solutions to the basic dynamical equation describe the clots of self-acting electrically charged matter, localized in space, i.e. the particle is a soliton.

The internal energy spectrum of electron is discrete with an indefinitely large number of levels, and to each value of internal energy E_k (k is the set of quantum numbers) there correspond certain linear dimensions and geometrical form of the region of localization of electron's charge. Dimensions and the number of extreme of wave function increase with increasing the value of energy E_k. The distribution of electric charge of atomic electron in the ground state consists of the range of basic localization with the linear dimensions of the order of Bohr radius a_0 ($a_0 \sim 10^{-10} m$) and of the tail stretching up to infinity. It is essential that because of non-linearity of the dynamical equation of electron, wave function does not obey the superposition principle. By virtue of this, electron acquires the properties of absolutely rigid body: the perturbation acting on electron at an instant of time in the range of basic localization becomes known at the next instant $t + 0$ at any distance from the particle.

In Fig. 1 the results of calculation are represented schematically, carried out on the basis of equation (13), of the distribution of electric charge in atomic and free electrons in the ground (a) and first excited (b) states.

According to [9,19], the atom represents a system of nuclear and electronic solitons interacting with each other, the internal energy spectrum of the hydrogen atom, due to electromagnetic interaction, being of a zoned character. The occurrence of zoned structure of energy spectrum of hydrogen atom is explained as follows. Free nucleus, because of existence of Coulomb self-action, has a discrete internal energy spectrum. As the interaction of nucleus with electron is small in comparison with the energy of Coulomb self-action of the nucleus, it can be taken into account by perturbation theory. From here it follows at once that each energy

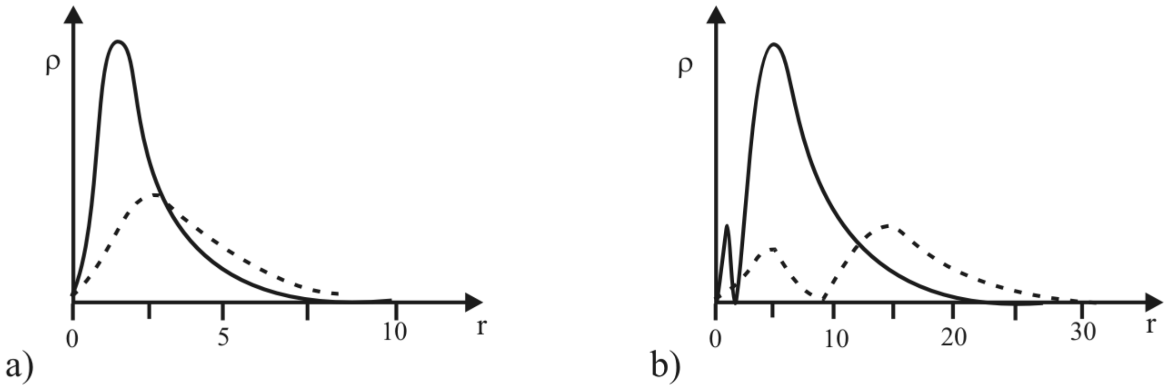

Fig. 1. Density of electric charge (ρ) of electron in the ground state (a) and in the first excited state (b): the continuous lines correspond to electron in the hydrogen atom, and the dotted ones to free electron, r is the distance from the center of mass of electron measured in Bohr radii.

level of free nucleus is split in a zone. There are indefinitely many zones (Balmer's replicas) and in each of them there are indefinitely many energy levels. The lowest zone coincides with the usual Balmer spectrum.

Physical mechanism of nuclear reactions at low energies

The quantum theory presented above schematically of electron as an open self-organizing system is indicative of the existence of the following mechanism of nuclear reactions at low energies [8].

If there occur in the region of basic localization of free electron, which linear sizes in the ground state of the particle are several times as large as those for hydrogen atom (see Fig. 1), two or the greater number of nuclei, each of them attracts on itself the adjoining areas of electronic cloud, resulting in compression of the electronic cloud as a whole. As a result, there appears automatically an attraction of the nuclei, which proved to be "inside" electron, on each other (see Fig. 2).

Calculation shows that the Coulomb barrier around nuclei is deformed, its height decreases and the probability of penetration through the barrier accordingly increases due to tunnel transition. Under certain conditions this process may result in fusion of nuclei. Obviously, the process in question can occur only at small energies of translational motion of the centers of mass of electron and nuclei: nuclei should be "inside" electron long enough for them to have time to come nearer to each other as a result of electron-nuclear interaction. This mechanism of nuclear fusion is of a universal character. In order for it to be realized, it is necessary to have only a stream of free electrons intensive enough, i.e. heavy electric current, and as long as sufficiently great number of free nuclei.

If heavy nuclei appear "inside" free electron, owing to their interaction with the electronic cloud there occurs polarization of nuclei. Because the own field of electron interacts with protons more strongly than with neutrons, nuclei are deformed (become extended), and this process may result in the decomposition of nuclei to fragments (in nuclear fission).

As is noted in [7], the official version of the reasons for Chernobyl accident contains serious contradictions, a number of facts concerning the accident has no convincing explanations, and this circumstance forces to search for the true reasons for the happening, since

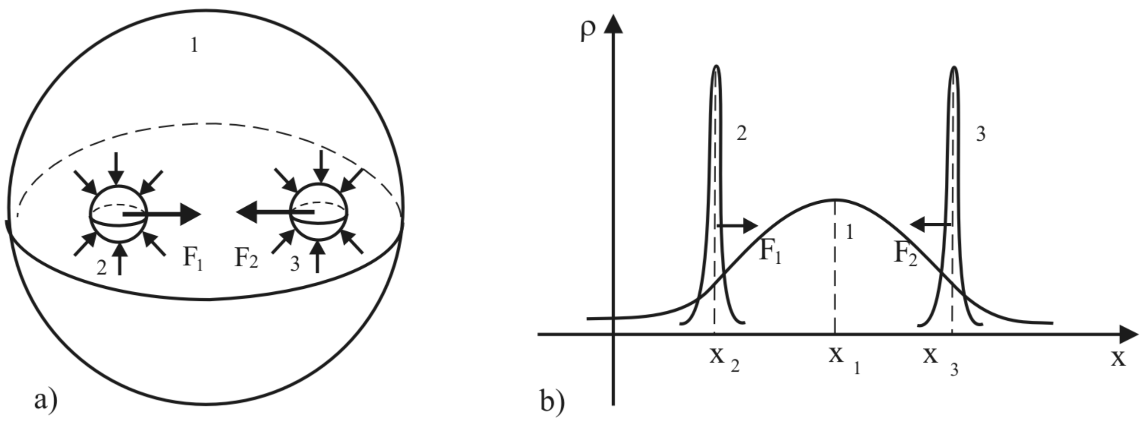

Fig. 2.
The schematic image of interaction of nuclei with electronic cloud: (a) 1 is the region of basic localization of electron, 2 and 3 are nuclei, F_1 and F_2 are the attractive forces between nuclei, which appear at the expense of electronic cloud compression induced by Coulomb forces; (b) ρ is the charge density, 1 is electronic soliton, 2 and 3 are nuclear solitons, X_n (n=1, 2,3) are coordinates of the centers of mass of particles.

"not having understood the mechanism of the one tragedy, we sooner or later shall become witnesses of the other". The authors hypothesize that the reason of the accident was penetration into the nuclear reactor of magnetic monopoles, which have caused the decay of nuclei ^{238}U, and this has resulted in production of delayed neutrons, growth of power output of the reactor and explosion. As an argument in favor of the assumption, the fact is presented that nucleus ^{238}U are disintegrated under the action of "strange" radiation appearing at explosion of foil.

In the opinion of the authors of [5,7], "strange" radiation is created by those magnetic monopoles, which form bound states with nuclei of atoms. These compound particles give the abnormally wide tracks similar to those of a creeping caterpillar, and also the tracks of complicated shape reminiscent of spirals and gratings. Character of tracks changes when imposing magnetic field, which, as the authors believe, is an argument in favor of the assumption above. There are also some special tracks very similar to scratches and ink spots. "Strange" radiation is of spherical form, it resembles a ball lightning, and its duration is more than ten times as great as that of the current pulse arising at electric discharge. With the course of time the luminous sphere (the ball-like plasma formation) is dividing into many small "balls".

It is our opinion that "strange" radiation is caused by free electrons in excited state arising in the area of electric discharge. According to [9, 19], linear sizes of the region of basic localization of such electrons can make many tens of sizes of atom. The heavy nucleus, for example, the nucleus ^{238}U, appearing inside the electronic cloud, is inevitably deformed because of interaction of protons with adjoining layers in the distribution of electric charge of electron, and this deformation can cause nuclear fission. If two or the greater number of light nuclei appears "inside" electron, then attractive forces arise between nuclei, which may result in fusion reaction. When electric discharge is strong enough, the areas of basic localization of some electrons can overlap, and if a nucleus lands in the area of overlap, because of Coulomb attraction of nucleus on the adjoining layers of electronic clouds, a bound state may be formed, of two electrons and the nucleus, characterized by the relative stability and significant spatial extension.

Obviously, if the concentration of free electrons is great enough, there may be formed some relatively stable bunch of plasma consisting of great number of free electrons and nuclei, which in virtue of chaotic movement of nuclei and because of the absence of preferred directions should have approximately spherical form. Let us note that atomic electrons, belonging to additional energy zones of atom (Balmer's replicas associated with nuclear self-action, see Section 2) can contribute to "strange" radiation.

As is seen from above, to account for the reasons for Chernobyl accident, there is no need to involve magnetic monopoles. The scenario of development of events during the accident, described in [7], seems to be quite plausible if only to understand by initiators of nuclear fission not hypothetical monopoles but free electrons, which powerful pulse might arise as a result of electric discharge in the region of turbo-generators.

The existence of simple physical mechanism of nuclear reactions at low energies, indicated in this paper, implies that nuclear reactors are, in effect, nuclear delayed-action bombs, which will blow up from time to time. Explosion of nuclear reactor may take place because of casual short circuit at an electric subcircuit, owing to which there appears an intensive stream of free electrons. This stream, having got for any reasons in nuclear reactor, may initiate explosion of the reactor. It follows from here that though nuclear stations may provide mankind with cheep energy, atomic energetics represents a very dangerous way of producing energy (as well as the energetics using controlled thermonuclear fusion). The only acceptable way of resolving the energetic problem consists in the use of nuclear reactions at low energies.

According to the results obtained, nuclear reactions at low temperatures occur "inside" electron under the action of own field of particle. Hence, to elucidate physical mechanism of CF, it is necessary to study in detail intra-electronic processes and physical properties of own fields of particles. Note that the own field, by its physical properties, essentially differs from the field of electromagnetic waves: this is the field of standing waves of matter, it is of purely classical character and may not be reduced to the set of photons. The own field of charged particle plays in nature a special role, consisting in that it transforms environmental space into the physical environment (physical vacuum) with the properties of absolutely rigid body [32].

As it was repeatedly noted in the literature [1,2], experiments on CF are badly reproduced, and this fact gives rise to doubt the very existence of the phenomenon. Bad reproducibility of results seems to be explained by the fact that CF depends upon great number of parameters: upon electric current density, concentration of free nucleus, concentration of impurities and dislocations in samples, sizes of samples etc. In order to obtain reproducibility of results, it is necessary that all these parameters, describing the environment in which nuclear reactions occur, be the same in various experiments, but to achieve this as a difficult task.

In conclusion we shall dwell upon the problem of linear dimensions of electron, which is of special interest in connection with the mechanism of nuclear reactions indicated here. The inference that the dimensions of electron in the ground state of atom are of the order of Bohr radius, i.e. of the order of atomic dimensions, following from dimension considerations [9,19] and confirmed by quantum model of electron, seems completely unexpected. At first sight, it is in conflict with both the theory of quarks and experimental data on scattering of electrons. According to quark models, the radius of electron corresponding to its quark structure makes up the quantity of the order of 10^{-22} m [33]. It is necessary to emphasize, however, that the above-mentioned magnitude of linear dimensions of electron refers to the internal structure induced by Coulomb field. The last is long-distance and consequently the linear dimensions of internal structures produced by it (i.e. spatial inhomogeneities in the distribution of electric charge in various quantum states) should considerably exceed the dimensions of quark structures connected with electron. There seems to exist a hierarchy of internal structures of particle produced by Coulomb forces, nuclear forces, inter-quark interactions etc. characterized by the smaller and smaller linear sizes.

As to the experiments on scattering of high energy electrons, according to which the internal structure of electron is not manifested up to distances of the order of $10^{-16} \div 10^{-17}$ m, two arguments, at least, can be adduced in favor of that there is no contradiction here with the experiment. Firstly, in experiments on scattering, investigators were trying to register the details of internal structure of electron within intervals much smaller than Bohr radius, which is why it is not surprising that results of experiments proved to be negative: at high energies electrons behave like point particles, their internal structure has no time to be manifested. Secondly, the results of experiments were analyzed from the point of view of standard representations about electron, which refer to a point particle, but are obviously inapplicable to real, self-acting electron. According to the predictions of quantum theory of electron as an open self-organizing system, real electron is a special object - soliton, i.e. such a cloud of electrically charged substance which, when interacting with other particles, tends to keep its sizes and geometrical form.

At present there is as yet no scattering theory of this kind of particles and for this reason it is impossible to predict with certainty how can the internal structure of electron be manifested in experiments on scattering.

References

1. Storms E. A *Critical Review of the "Cold Fusion" Effect.* J. Sci. Explor., **10**, #2, p.185 (1996). See also: http://www.jse.com/storms/1.html
2. Storms E. *Cold Fusion Revisited.* Infinite Energy, 4, #21, p.16(1998).
3. Schwinger J. *Cold Fusion: A Hypothesis.* Z. Nat. Forsch. A 45, 756 (1990); *Cold Fusion: A Brief History of Mine.* Infinite Energy, 1, #1, p.10 (1995).
4. Schwinger J. *Nuclear Energy in an Atomic Lattice I.* Z. Phys. D15, 221 (1990); Prog. Theor. Phys. 85, 711 (1991); *Energy Transfer in Cold Fusion and Sonoluminescence.* http://www.geocities.com/CapeCanaveral/Lab/4520/theory.htm
5. Urutskoev L. I., Liksonov V. I., Tsinoev V. G. *Experimental Detection of "Strange" Radiation and Transformation of Chemical Elements.* Applied Physics, M., 2000, p.83-100. Urutskoev L.I., Liksonov V.I., Tsinoev V.G. *Observation of Transformation of Chemical Elements During Electric Discharge* http://arxiv.org/ftp/physics/papers/0101/0101089.pdf
6. Vysotskii, V. I., Kornilova A. A., and Samoylenko I. I. *Experimental Discovery of Phenomenon of Low-Energy Nuclear Transformation of Isotopes (Mn55=Fe57) in Growing Biological Cultures.* The Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy (Ed. M. Okamoto) Oct. 13-18, 1996, Hokkaido, Japan, Vol. 2, p. 687; Infinite Energy, 2, #10, p.63 (1996).
7. Urutskoev L.I., Geras'ko V.N. *On the Possible Mechanism of Chernobyl Accident*. http://www.recom.hotmail.ru/CHER_INC_1.doc
8. Oleinik V.P. *To Electronic Technologies of the 21st Century: on the Threshold of Revolution in Communication Systems. Collection of Reports*, Millenium 2002, International Conference "To Innovations in the 21st Century", Odessa, April 13, 2002, p.268-273.
9. Oleinik V.P. *The Problem of Electron and Superluminal Signals*. (Contemporary Fundamental Physics) (Nova Science Publishers, Inc., Huntington, New York, 2001).
10. Berestetsky V.B., Lifshits E.M., Pitaevsky L.P. *Relativistic Quantum Theory*, part 1. (Moscow, Nauka, 1968).
11. Medvedev B.V. *Foundations of Theoretical Physics*. (Moscow, Nauka, 1977).
12. Dirac P.A.M. *Relativistic Wave Equation of Electron*. Progress in Phys. Sciences, 129, #4, p.681–691 (1979).
13. Dirac P.A.M. *The Principles of Quantum Mechanics.* (Moscow, Nauka, 1979).
14. Schrödinger E. *Quantisierung als Eigenwertproblem*. Vierte Mitteilung. Ann. der Physik, Bd. 81, S.109–139 (1926).
15. Schrodinger E. *Selected Works on Quantum Mechanics*, Edited by Polak L.S. (Moscow, Nauka, 1976).
16. Barut A.O. *Schrodinger's Interpretation of as a Continuous Charge Distribution*. Ann. der Physik, Bd. 45, S.31-36 (1988).
17. Barut A.O., van Huele J.F. *Quantum Electrodynamics Based on Self-Energy: Lamb Shift and Spontaneous Emission without Field Quantization*. Phys.Rev., A32, #6, p.3187–3195 (1985).
18. Barut A.O., Dowling J.P. *Quantum Electrodynamics Based on Self-Energy*: Spontaneous Emission in Cavities. Phys.Rev., A36, #2, p.649–654 (1987).
19. Arepjev Yu.D., Buts A.Yu., Oleinik V.P. *To the Problem of Internal Structure of Electrically Charged Particles. Spectra of Internal Energy and Charge Distribution for the Free Electron and Hydrogen Atom.* Preprint of the Inst. of Semiconductors of Ukraine, N8-91 (Kiev, 1991) 36 p.(in Russ.).
20. Oleinik V.P. *Quantum Electrodynamics Describing the Internal Structure of Electron*. Quantum Electronics. #44, p.51-59 (1993) (in Russ.).
21. Oleinik V.P. To *the Theory of the Internal Structure of Electron. Second Quantization and Energy Relations.* Quantum Electronics. #45, p.57-79 (1993) (in Russ.).
22. Oleinik V.P. *Quantum Theory of Self-Organizing Electrically Charged Particles. Soliton Model of Electron*. Proceedings of the NATO-ASI "Electron Theory and Quantum Electrodynamics. 100 Years Later." (Plenum Press, N.-Y., London, Washington, D.C., Boston, 1997), p.261-278.
23. Oleinik V.P. Nonlinear *Quantum Dynamical Equation for the Self-Acting Electron*. J. Nonlinear Math. Phys. 4, #1-2, p.180-189 (1997).
24. Oleinik V.P. *Quantum Equation for the Self-Organizing Electron.* Photon and Poincare Group (Nova Science Publishers, New York, Inc., 1999), p.188-200.
25. Nicolis G., Prigogine I. *Self-Organization in Non-Equilibrium Systems* (Wiley-Interscience, 1977).
26.Oleinik V.P., Belousov I.V. *The Problems of Quantum Electrodynamics of Vacuum, Dispersive Media and Intense Fields.* (Kishinev, Shtiintsa, 1983).
27. Bogoliubov N.N., Shirkov D.V. *Introduction to the Theory of Quantized Fields.* (Moscow, Nauka, 1976).
28. Bohr N. *Selected Scientific Works*, V.2. (Moscow, Nauka, 1971).
29. Bogoliubov N.N. *Quasi-Averages in the Problems of Statistical Mechanics.* In the Book: Statistical Physics and Quantum Field Theory. (Moscow, Nauka, 1973).
30. Morse P.M., Feshbach H. *Methods of Theoretical Physics*, V.1. (Moscow, Foreign Literature, 1958).
31. Dakker H. *Classical and Quantum Mechanics of the Damped Harmonic Oscillator.* Phys. Reports, 80, # 1, p.1-112 (1981).
32. Oleinik V.P. *Superluminal Signals, the Physical Properties of Time, and the Principle of Self-Organization.* Physics of Consciences and Life, Cosmology and Astrophysics, #1, p.68-76 (2001).
33. Dehmelt H. *Experiments with an Isolated Subatomic Particle at Rest.* Progress in Phys. Sciences, 160, #12, p.129–139 (1990).

The Evolution of Lifter Technology

Tim Ventura

tventura6@attbi.com

INTRODUCTION

Readers of the electric-spacecraft journal might know a little about the Lifter technology popularized recently be Jean-Louis Naudin, but they probably don't know the whole story. In the short amount of time that has transpired since the publication of that article, this technology has both literally and figuratively taken off – going from a "proof-of-concept" prototype by Naudin to an international group of researchers investigating how to give the lifter higher-performance and greater efficiency. With the first commercial products now on the horizon, if you haven't taken the time to read up on lifter technology, this is the perfect time to do so. . .
To give you a complete up-to-date overview of where this technology is, where it is going, and what I think it is capable of, let me start with the basics – an overview of how I became involved with Electrogravity research and what eventually led me to become involved with lifter technology.

MY BACKGROUND

I started college at 16 years old, back in 1992 – at the same time, I purchased a kit containing "hoverboard plans" from Hovertech, Inc. The moment that I received that $20 white-manilla envelope in October 1992 was the moment that I became involved with what has now been nearly 10 years of electrogravity research.

I worked with Bill Butler – the president and chief-scientist of Hovertech – on a variety of different antigravity, Electrogravity, and levitation ideas from approximately 1992 through 1996. While putting in my college time, I was also taking distinct advantage of the enormous college library at Western Washington University to read up on everything that might possibly relate to Electrogravity. I read books on standard electronics and physics theory alongside with books by the masters of this science, such as TT Brown and Nikola Tesla.

Bill and I played with several different ideas – many of them only peripherally related to Electrogravity. For instance, I published a manuscript initially in 1996 describing Tesla's theory on how to reliably produce Ball-Lightning using a standard Tesla coil – the information courtesy of WWU's excellent library. Bill also assisted me with obtaining video footage of a Searl-effect conference that he attended in Denver in the early 90's – this footage was an excellent overview of Searl's design and construction concepts for what he believes is the next major technological step in aviation and space travel.

Bill and I eventually found different paths, and in some ways drifted apart. Bill moved into Geomagnetic levitation research and started intense investigation on the patents of How Wachspress and the magnetic dipole levitator. I went to more traditional technologies – eventually becoming a UNIX system administrator for AT&T Wireless.

I hadn't heard from Bill Butler in about 6 months when he sent me a short email containing the words "hey, check this out" – and a link to Jean-Louis Naudin's "Lifter Experiments" home-page. I visited the site, watched all of the video clips, and then watched them again. This was the technology that I had been waiting for!

LIFTER TECHNOLOGY

I can say without a doubt that the lifter technology is completely revolutionary, but you might not realize how profoundly revolutionary it is until you've stopped to think about it for a bit. What is it about the lifter that makes it so unique, especially when so many inventions claim to produce more and better electromagnetic thrust? The answer is simple – the lifter works repeatedly.

Jean-Louis Naudin started a figurative bonfire when he decided to replicate a "proof-of-concept" experiment by a small Huntsville, AL aerospace contracting firm. The lifter initially came into being in the mind of Jeff Cameron – the chief scientist of Transdimensional Technologies – in the 1970's from experiments conducted with high-power military and research-grade lasers. A device in the lasers called a "pre-ionizer" was used to apply a high-voltage to the lasing-medium to facilitate better performance. Repeated operation of the pre-ionizer had a common side-effect of horribly twisting the wire and foil combination out of shape, which required a decent amount of work to repair.

Jeff Cameron realized that the torsional effect on the pre-ionizer was a side-effect of some unknown force acting on the pre-ionizer apparatus, and he began a long-term investigation into what was causing the apparatus to deform. His eventual results indicated that a force in the foil collector in the pre-ionizer was causing a net-thrust in the entire pre-ionizer apparatus that was making it twist and move on its mounts within the laser – the lifter came to him later as a three-dimensional device to demonstrate this force.

Naudin's genius became readily apparent not through a giant breakthrough in technology, but rather in a more subtle fashion – he replicated the lifter experiments of Transdimensional Technologies and published videos, articles, and complete construction plans on his website to allow others to do the same. In a manner similar to the open-source software movement, Naudin had taken an incredible scientific find that might have otherwise been overlooked and done and incredibly charitable and intelligent thing – he gave it away for others to play with. By following Naudin's instructions, inventors all over the globe began to slowly replicate the

Transdimensional Technologies experiments and thereby validate the proof of concept that Jeff Cameron had created to show that his "mystery force" was real after all. Naudin of course took advantage of these replications of the experiment by showcasing them on his own website – which in turn lends additional credibility to his research.

As far as technology goes, the lifter demonstrates that science and engineering have more than their share of humorous irony. For the years that I researched Electrogravity and antigravity claims, all of the devices that I had seen required something "magic" to make them work. For instance, Bob Lazar's UFO-claims could have been reverse-engineered except that they require 'element 115' to make them work – an element chemically related to Bismuth that is theorized to potentially have electrogravitic properties. I will come back to the possible electro-gravitational properties of Bismuth in a bit, as it turns out that this element may in fact provide some use for future lifter technology.

The Searl-effect disc is an even better example of the "magic" usually involved with building a working Electrogravity device. Searl's ideas seem valid enough, but although he supposedly demonstrated several working prototypes in the 1950's, he is currently pursuing millions of dollars in research funding in to replicate those experiments in a modern-day setting.

The irony involving lifter technology is that while inventors all over the world have been searching for the perfect electro-gravitational device for decades, the possible working proof of concept for many of these theories has been sitting in front of us the whole time – the lifter costs less than $10 in parts to build, and none of them are magic – in fact, for my experiments, all of them were at stores within 2 blocks of my house — balsa wood from the craft store, aluminum foil from the supermarket, 30-gauge magnet wire from the local Radio Shack, and an old computer monitor for the high-voltage power-supply.

LIFTER PHYSICS

Whether or not Jeff Cameron knew it at the time he constructed his lifter prototype, what he was actually building was a 3 dimensional representation of a drawing on a patent application by TT Brown in the 1950's. In the patent application, the drawing shows a positively charged wire suspended over a grounded foil body which was meant to demonstrate the most basic Biefeld-Brown effect generator. While Brown's drawing is a little different than Jeff's design, the resemblance is uncanny enough to indicate that both of these men had the same basic force in mind.

TT Brown's patent indicates that this Biefeld-Brown effect generator works due to a gradient electrostatic-field between the wire and the foil – in essence, these two elements compose a low-efficiency, high-voltage air-gap capacitor in which the difference in geometries between the two capacitive elements generates a net-directional force from the larger element towards the smaller element. Jeff Cameron seems to have a practical axiom that goes along with this scientific philosophy, which is that there must be both a leakage current and a capacitance between the wire and the foil in order for the lifter to function.

Conventional physics says that two capacitor elements of different sizes will not generate a net-directional force, so what gives? This is actually the thinking that convinced me to abandon my research into Biefeld-Brown effect technology in 1996 – physics says it doesn't work. What the books say will happen is that since the wire can only maintain a lower-capacitance than the foil, the overall capacitance between the two elements will be reduced to be equivalent to that on the smallest element (or plate) in the capacitor. This, of course, assumes a 2-element series-wired capacitor, such as the lifter.

I can give you the conventional physics answer to this small riddle by simply saying that the lifter uses a manifestation of ion-wind. This would state that the electrons crossing the air-gap cause a breeze that causes thrust – since the breeze would be traveling down from the wire to the foil, the thrust would be up, as demonstrated in testing. In the ion-wind explanation, the electrons are emitted from small-diameter of the positively charged wire in such great abundance that they move a significant airflow down to the foil where they are absorbed and transported electrically back to the HV power-supply's electrical ground.

Conventional physics would seem to have the theoretical answer to why the lifter causes lift, but in the experimental setting, which is what we now have an abundance of thanks to Jean-Louis Naudin, the conventional physics explanation doesn't suffice. Experimentally, there are several deviations from the ion-wind explanation that seem to invalidate it. For instance, if you completely contain the lifter in a plastic-enclosure, it will still generate lift – this would not be the case if a breeze was responsible for lifting the device. How could it be, if the breeze is limited to the inside of an enclosure which itself is levitating?

A more compelling proof that Biefeld-Brown is something other than ion-wind comes from Purdue University, where the lifter experiment was replicated inside a vacuum-enclosure with positive results. While ion-propulsion can work in space, it usually assumes that there is argon, krypton, or other noble gas to be used as the propellant – the vacuum enclosure showed that with no gas available for transport the lifter showed a moderate improvement in performance.

The vacuum enclosure tests are definitely compelling evidence that something else is going on other than ion-wind – at least compelling enough for NASA to file patent number 6,317,310 – "Apparatus and Method for Generating Thrust using a Two Dimensional, Asymmetrical Capacitor Module". The NASA patent description – which can be accessed from Naudin's lifter website – is as vague is it is compelling in that NASA is basically requesting a patent on any technology that generates force using two geometrically dissimilar capacitive plates. Disregarding the fact that this patent was issued nearly 50 years after TT Brown's patent using nearly identical descriptions and pictures, and also disregarding the fact that NASA also doesn't understand why the lifter generates thrust, it seems apparent the this phenomena is gaining credibility in engineering circles while physicists seemingly continue to deny that anything is going on.

THE EVOLUTION OF LIFTER TECHNOLOGY

Every good movie always has a sequel, and in technology, if at first a major government agency 'liberates' your idea, it may seem that a sequel is in order. In the case of the lifter, it would appear that the NASA patent would cover this technology to at least some degree – at least until someone overturns this patent under the prior-art rule – which means that the next generation has to be considerably more advanced to escape having the research and development be forfeit to the government.

The pursuit of more advanced versions of the lifter technology is currently underway by several independent inventors, as well as Transdimensional Technologies themselves. Most of the private research by inventors has delved into improving the current lifter design to produce a greater force output and utilize less power to do so. Because the lifter is so simplistic in design, many of these enhancements have been of a very basic nature.

Jean-Louis Naudin was the first independent inventor to do serious work with improving the technology behind the lifter – and even so, the majority of his work has utilized similar materials in more complex arrangements. Naudin has demonstrated dramatically increased lifting forces by building a "lifter inside a lifter" for demonstration purposes. Naudin has also done a great deal of work in taking breaking up the concept of the single triangular lifter into a parallel series of lifting cells – which means that these cells, working in parallel, can contributed to greater stability and higher force output than any single lifting element.

Saviour – an independent inventor working with Jean-Louis Naudin – has done some of the most interesting improvements on lifter design since those by Naudin himself. Saviour's concerns have not focused around the "bigger is better" philosophy that many inventors have stuck by – he has done several experiments to determine the radiation output, remote-controlled applications development, and materials analysis and improvement on the lifter that others have not had the time or expertise to conduct.

A recent experiment by Saviour demonstrates just how this gentleman's foresight is helping other experimenters – Saviour substituted nichrome heating wire for the common lightweight wire used for the emitter, and demonstrated that the lifting force greatly increased when a higher potential 12-volt charge was used to heat the emitter wire in conjunction with the standard high-voltage charge coming off it.

Transdimensional Technologies – the developers of the initial lifter design – are taking the approach to optimizing lifter performance to another level. They are currently not-so-secretly working on a 2nd generation lifter, which will consist of a 1-piece layered material to replace the current wire and foil design.

The layered material approach to the lifter is an idea that Jeff Cameron may or may not have had after some lengthy discussions with Travis Taylor – the man responsible for testing some anomalous materials known as "Art's Parts".

Art's Parts were some pieces of material sent by an unknown person to the Art Bell radio talk-show with a note stating that the they were pieces of UFO wreckage taken from the often-cited "Roswell crash" in 1947. Whether or not the pieces of material actually came from that crash is unknown, but Art Bell did the honorable thing by sending them to an acquaintance in US Army research named Travis Taylor for a professional scientific investigation.

Taylor, who apparently tested the materials after-hours in a world-class research lab to avoid potential classification by his superiors, used an electron-microscope to determine that the layered materials were actually pieces of metal – containing several hundred microscopically thin layers of magnesium and bismuth. Taylor also tested the layered-metal with a high-voltage apparatus, which seemed to indicate that when a voltage was applied to the material, the layered metal would move – and in some cases levitate.

Taylor reported his findings to Art Bell and sent video clips of his high-voltage experiments, which eventually made it back to a permanent home on the Art Bell radio show website. In addition, Taylor conveyed his belief that the only manner in which the pieces of metal could properly be produced was through an advanced form of electron-deposition technology, due (apparently) to an absence of oxygen-molecules between the different layers of metals. Additionally, the layers of metal were too thin to have been mechanically produced.

Jeff Cameron indicated that Transdimensional Technologies maintained some contact at one point in time with Travis Taylor, apparently as professional colleagues in the defense community in Huntsville, AL. I am not an expert on this relationship, other than to say that to the best of my knowledge these two individuals knew and contacted each other, and that this is how Jeff Cameron might have come up with the 2nd generation lifter idea.

ADVANCED LIFTER TECHNOLOGY

As an inventor, I couldn't care less whether or not the idea for the technology came from a crashed UFO. To be perfectly honest, I'm not what you would call a "believer" anyways, although I have often wondered about it. My point is not to attempt to lend any credibility to "Art's Parts", but rather to tie in the properties of the anomalous material's high-voltage movement with the underlying theory of lifter operation.

Even mentioning a UFO in a respected publication or article is the kiss of death in today's world – and I wouldn't do it if it wasn't an intricate part of the story. The other interesting thought is that the layered material is once again partially composed of Bismuth – which is thought to possibly have some of the same electro-gravitational properties as Bob Lazar's Area 51 "element 115". Is there a similarity, or merely a coincidence between a claim that hasn't gained credibility and a technology currently under development?

The lifter in its own right is essentially a layered material. One of those layers is the emitter wire, which is highly charged with about 30kV worth of electrons, another layer is the air-gap, which is approximately 3 cm in height, and the final layer is an electrically-grounded "skirt" of aluminum foil that surrounds the lifter. It is also reasonable to expect that there are only two possible forces at work in the lifter – one of which being a possible ion-wind effect moving down from the emitter to the foil, and the other being a possible Biefeld-Brown effect, moving up through the foil to the emitter.

There are a few shortcomings in the lifter as a design that might be overcome if we could transition the layered material from one containing an air-gap to one that does not. For instance, the lifter is currently a rather delicate object, in that having a wire under tension as the emitter makes construction difficult for future automated assembly. Additionally, because the air-gap requires struts to support the emitter wire, a trade off involving the weight versus the strength of the struts is additionally involved in any current implementation of lifter technology.

Some of the other changes that would be helpful to implement when transitioning lifter technology from one type of air-gap to another are changes in the materials used to increase the dielectric capacity. High-K dielectric materials may be used to increase the displacement of electrons in the material to enhance charge transport. And since increasing the dielectric potential of the layered materials also increases the breakdown resistance, it means that thinner materials can be used.

Designing a lifter without an air gap would accommodate lower voltage requirements between the foil and the emitter. The voltage would not have to create the large e-field gradient to create a leakage current across such a large void. Therefore the overall voltage across the device could be greatly reduced, without much cost in thrust. A lower operating voltage in turn means that a lower-output power-supply can be used for a given amount of current, which increases the overall efficiency.

Transdimensional Technologies recent research is utilizing the layered materials approach to eliminate the air-gap and substitute for it high-k dielectric materials that may allow higher overall performance. Although they have not yet released details about the exact composition or thickness of the materials that they are working with, they claim to currently have a 10% reduction in weight using a low-voltage current across the thickness of their newest device.

FUTURE LIFTER TECHNOLOGY

Thanks to the tremendous amount of research being done on lifter technology by Transdimensional Technologies and a loosely affiliated group of inventors around the world, the future of lifter technology seems very bright at this point.

Transdimensional hopes to release some breakthrough research to allow replication of their newest 2nd generation experiments in the very near future, and along with that stands the massive body of research and advancements being done by inventors and researchers such as Jean-Louis Naudin, Saviour, the Lifters-group, and myself.

My personal goals are to attempt to assist Transdimensional Technologies in popularizing this technology to increase awareness of it and help "spread the word" about what it is and how it can potentially help the world.

Imagine if instead of getting in your car and driving through the usual maze of thoroughfares and side streets you were able to simply type in your destination and have a flying vehicle take you there automatically. The lifter technology offers to potential to transform the current transportation market by offering point-to-point aerial transport without the need for roads or freeways.

Additionally, unlike the magnetic-levitation ("Maglev") technologies that are currently being promoted as the future of transportation, the lifter does not require a specially constructed and exorbitantly expensive track to operate – the greatly reduces the per-unit cost on the technology and opens the door for wider adoption by the general public for transportation solutions.

Other individuals are currently working to see if lifter technology may offer cost-effective methods of transport into space, which would reduce the cost greatly and allow a one-piece, reusable method of moving things into orbit.

LIFTER RESOURCES

All of the research involved with the lifter technology is available to the public on the internet. The list of resources below are some of the better and more common resources to obtain detailed lifter information.

American Antigravity
→ http://tventura.hypermart.net

→ The author's website that includes video clips, complete instructions, and other related lifter information.

Jean-Louis Naudin's "Lifter Experiments Website"
→http://jnaudin.free.fr
→ A very in-depth website containing video clips, complete instructions,

World-Wide Lifter Replications
→ http://jnaudin.free.fr/html/lftwrld.htm
→ An overview with photos and video from many of the independent inventors who have replicated the lifter experiments.

Transdimensional Technologies, Inc
→http://www.tdimension.com
→ The home page for Transdimensional Technologies, the developers of the lifter design.

Blaze Labs (Saviour's Research Website)
→http://bel.150m.com
→An excellent site on research into lifter enhancements, radiation testing, sealed devices, power supplies, and other topics relating to lifter technology.

Lifter Builders Group
→ http://groups.yahoo.com/group/Lifters
→ An email group for the exchange of research findings for those interested in building lifters or staying current on the state of the technology.

NASA Patent #6,317,310
→ The NASA patent regarding obtaining thrust from an asymmetrical two-dimensional capacitor, grant Nov 13, 2001.

Research on the Capacitance Converter of Environmental Heat to Electric Power

N.E. Zaev

143970, Moscow region, village Saltykovka,
Granitchnaya Str., 8 529-9664

Nickolay E. Zaev works on creation of the prototypes of converter energy, which do not require any fuel. The direct conversion of environmental heat to electric power is possible in the processes of "charge-discharge" in non-linear condensers or by means of "magnetization-demagnetization" of ferrites. Such converters of energy create cold and electric power without any fuel.

Theory of the converter, results of early experiments on the generation of microwatt power, methods and features of research are given in this article. The methods of generation of a few watts power are described in details. The possibilities and difficulties of creation of powerful capacitance converters are discussed in this article.

I. Grounds of research.

1.1. From positions of orthodox physics there is no subject of research. It is evident that the energy of charging (C) A_c condenser C_x is always equal or more than the energy of discharging (D) A_d, i. e. always $A_c \geq A_d$. Only the advanced analysis shows that it is not always true. Exactly, in C_x, where $\frac{\partial C}{\partial V} < 0$ an inequality $A_d > A_c$ is possible, and in C_x, where $\frac{\partial C}{\partial V} < 1$, then the work $A_c > A_d$. Therefore we should discuss the nonlinear capacitors (NC). In the end of 1969 I noticed a systematic inequality $A_d > A_c$ during the measurement of A_c and A_d of many capacitors with different dielectrics. Theoretical grounds and results of measurements of this phenomenon are given in the publications in 1984 [1], [2, page 73]. On the industrial standards NC (varicond), ceramic condensers VK2-ZSH, $4 \cdot 6{,}8 \cdot 10^{-9}$μF with an optimal voltage about 95 V it was stated that $\frac{A_d}{A_c} \sim 1{,}21$ with the power to about $98 \cdot 10^{-6}$ Wt and "generated" extra power is equal to $21 \cdot 10^{-6}$ Wt.

1.2. In [1] and [2] the strict theoretical proofs of realization of $A_d > A_c$ (there are four of them) are given. On 1m^3 of dielectric $|A_d| - |A_c| = -\frac{1}{2} a \cdot \varepsilon_0 \cdot E_c^{\ 3}$ (E_c is an intensity of the field, V/m; ε_0 is a dielectric constant of vacuum, a is a coefficient of nonlinearity of the capacitor). Below we state one more proof more connected with the parameters of circuit.

It is well known that with the charge of a linear capacity from the source of constant voltage V_0=const through the resistor R=const it gets an energy $A_c = \frac{C \cdot V_0^{\ 2}}{2}$ exactly equal to the output energy in the time of charging t_c. The output energy irradiated from the load R is a Joule heat $\Theta = R \cdot \int_0^{t_c} i^2 \cdot dt$ [3, page 546]. If NC (nonlinear condenser) is charged, then there are no proofs of such equation. The NC are the variconds or other capacitors, which have $\frac{\partial C}{\partial V} > 0$ in the interval $V = 0 \div V_k$. For the variconds V_k is some voltage, which corresponds to the maximum C_{v}. If $V > V_k$, then $\frac{\partial C}{\partial V} < 0$. For some other capacitors V_k is a voltage breakdown.

For further consideration let's believe that in the operating area of the given sample of varicond a function

Reality and Consciousness in Education and Activity

A.P. Smirnov

Vice president of International Club of Scientists
190031, Saint Petersburg, Kazanskaya str., 36
Tel: +7 (812) 312-0508
E-mail: science@shaping.org

Relation of thought to existence is the main question of philosophy as science on general laws of Nature was formulated but it still did not interpreted and solved in the frames of generally accepted logic standards. The ways to solve it lead to futile discussions of materialists and idealists, to senseless disputes of determinists with eclectics and apologists of the "chance". This discussion lost its sense without a determination of terms under discussion and condemned debaters to have subjective "gustatory" senses, which were changing while aging and depended on the extent of received and conceived knowledge. Such is the situation in this link of World studying, which does not allow creating a logic chain of reasoning in the understanding of cognizable things.

A paradoxicality of all things that happen is connected with incorrect translations and interpretation of wisdom of ancient philosophers and scornful attitude both to the knowledge of distant past and classical heritage, which highlighted the elements of natural-science approach to Weltanschauung.

According to Plato, **an ideal thing** is a visual thing, which can be felt by our organs of sense. Therefore, the understanding of objective reality is mediated by the crowd of our feelings in such a way that perception of reality by means of these feelings gives us a notion of the World. Hence, **our notions about reality are the subject of research in science, but not the World itself, i.e. the World outside of our consciousness.** So, what should be studied in our notions about the World? Let us refer to the wisdom of ancient scientists again: " The World is given in motion and its laws are the laws of motion". Then, we should speak about laws, order, i.e. about relation and interrelation in the phenomena of motion. This is the distinctness in notions and actions (determinism) to predetermine further development of reality cognition logic, i.e. what has an influence on us and determines specific character of our perception. Further we can speak about formation of ideas about reality, which require some premises, principles to organize these ideas. These principles are given in classical heritage, in "Dialogues" by G. Galilee [1] and "Mathematical principles of natural philosophy" by I. Newton [2]. A notion of force as a measure for momentum was introduced, which manifests in action and disappears from the body after the action is over, and the body keeps its new state due to the inborn "inertia force". **But the force itself cannot do anything without its application with a certain speed.** Then we introduce a notion of action as a product of acting force F_A and the speed of action V_A.

We offer a law of interaction, which determines the interaction between action of the cause and the effect appeared during this action as a reaction, i.e. the product of the force of reaction F_R and the speed of reaction V_R. Thus, this interaction between the cause and the effect is determined by the transfer of action from one object to another in equal quantity, but with appearance of new quality, which is determined by specificity of interacting objects according to fundamental law of interaction:

$$F_A V_A = -F_R V_{R.}$$

Unfortunately, an incorrect interpretation of interaction manifestation as an opposite counteraction became strong in our mind. This manifestation is perceived as a compensation of cause by action of the effect. Moreover, the incorrect way of writing of the mathematical form of Newton's third law manifestation established in textbooks and scientific literature due to the incorrect translation as $F_A = -F_C$. This very tragic situation for the science suppressed the development of logic in description of processes. Chance and statistic approach to the description of phenomena has taken place in our perception. This approach is based on the model of non-interacting elements, in which there is no order stipulated by the interrelation of elements. The science has developed this model and its properties, and this fact predetermined the evolution of notions about real World.

This ideology penetrated in mathematics, which for sake of physics began to study properties of objects, but not operations with them. Moreover, a possibility to reflect specific character of real physical processes in the interconnection of cause-effect relations by mathematical operations is not realized. It is essential, that fundamental law of interaction establishes manifestation and description of elementary act of cause and effect interrelation, **the law of manifestation of a Fact.** It means that order in the World is conceived through manifestation of concrete facts. The action of law of interaction lies in the basis of these facts.

So, there is a conclusion: the World is perceived through the discrete manifestation of motion forms evolution. Hence, the discrete mathematics of finite discrete aggregate can be applied to describe the World, but not the continual mathematics, which lies in the basis of traditional orthodox physics. All these circumstances lead to numerous problems and difficulties in description of our notions of reality, to the plenty of used principles, which are in contradiction to each other, as R. Feinmann noticed once [3].

And what we can get from determinism, which is based on fundamental law of interaction, law of cause and effect interrelation? **The change of force value in a reaction takes place, i.e. the change of value of the potential gradient, i.e. the change of energy**

concentration. This circumstance is visually demonstrated by the operation of Archimedean lever as well as in all phenomena of the real World. This is Archimedean lever, where the loss of speed takes place, but there is a gain in force. And the load raised on a lower height than the way, which was made by the applied force, will give a huge power during its free fall. This power is higher in so many times, in how many times the time of the load fall is less than the time of action spent on its raising! And this is the fact, which determines specific character of **creation in the real World**. We should attribute both quantitative and qualitative characteristics to energy. This is the side of energy manifestation, which is reflected in Plank's formula: energy is proportional to frequency.

Manifestation of fundamental law of interaction also lies in the basis of general universal regularity of evolution of real many-particle systems with the change in external conditions. This process develops in multistage way, and on the each stage the logarithm of the ratio between the event happened and the event to happen always is equal to the work of external forces. In other words, the relation of the event happened to the resource is in exponential dependence on the initial conditions and extent of external influence. Exponential character of development of processes is the evidence that Nature develops according to the law, which conserves itself during evolution. This regularity, which manifests everywhere, can be naturally called **the Principle of Order.**

Fundamental law of interactions and Principle of Order appeared to be enough to describe and understand phenomena in the observed World. And it is natural to expect that this principle of Nature manifest in finer World also. This World includes lower and higher frequencies, which are not available for us yet to watch this wide-range frequency-wave emanating Universe.

From all aforesaid we should make a conclusion that the logic, which exists in the traditional physical tool, appeals to the model and principles of the World of non-interacting elements using the range of regularities, which also reflect some features of the real World, but they do not include fundamental law of interaction and Principle of Order, which are necessary and sufficient to describe reality. Descriptions existing in traditional physics are phenomenological ones and concern only those aspects of the phenomena under investigation, which do not include possible qualitative changes during development of processes, because the main property of real processes of interactions (creation of new energy property) was excluded.

The current situation in physics had a strong influence on formation and development of other sciences, other fields of knowledge, since the logic of reflection of cause-effect links was initially excluded. These are the links to determine existence, i.e. existence of constant creation of the World. All these circumstances give grounds to fundamentally revise educational programs, first of all, in physics, philosophy, mathematics, chemistry and biology. A change to the offered logic of cognition, which is based on the Principle of Order and fundamental law of interaction, will fundamentally change our notions about the World as well as will open big opportunities for new technique and technology. A Man has got huge opportunities in cognition and existence, but due to his immorality and features of incorrect aims in the logic of cognition he cannot use these gifts of Nature. We present wider and deeper view on the World and a Man in it, which allow analyzing, watching and operating with those fields of reality, which manifest in finer World, World of higher-frequency energies and other structures of fields. Logic of cognition had not touched these structures yet.

References

1. Galilee Galileo. Selected works in 2 volumes. M.: Nauka, 1964.
2. Newton Isaac. Mathematical principles of natural philosophy. News of Nickolaevskaya Marine Academy, issue IV, V, Petrograd, 1915-1916. Volumes I, II and III, 620 p.
3. Feinmann Richard. Character of physical laws. M.: Nauka, 1967, 160 p.

Old New Energy

Y. I. Andreev, A.P. Smirnov

St.Petersburg, Russia
E-mail: science@shaping.org
Internet: http://www.physical-congress.spb.ru

Two kinds of energy, accumulated energy [1] and free energy [2], are considered as an inexhaustible source of natural energy created by Nature itself. It is ecologically clean and possible to be renewed in natural conditions.

The energy accumulated in substance is released as a result of partial decay of substance in elementary particles. At that, the acquired defect of mass is so small that it does not change chemical properties of substance and is compensated in natural conditions. Physical mechanism of energy-release lies in the fact that an electron in plasma layerwise takes sufficiently smaller elementary particles (electrino) from positively charged atoms or fragments of substance (ions). Electrino give their kinetic energy to plasma, heat it up and move beyond the bounds of reaction zone in the form of thermal and optical radiation. There is no substance, which could not take part in such process of energy-release, i.e. phase transfer of higher form (PTHF). The most appropriate, available and low-cost substances are air and water, which play the role of nuclear fuel in PTHF. It is turned out that usual combustion is also a process of PTHF, in which oxygen is a nuclear fuel and organic fuel is a donor of electrons. In the process of combustion oxygen atoms get the defect of mass equal

concentration. This circumstance is visually demonstrated by the operation of Archimedean lever as well as in all phenomena of the real World. This is Archimedean lever, where the loss of speed takes place, but there is a gain in force. And the load raised on a lower height than the way, which was made by the applied force, will give a huge power during its free fall. This power is higher in so many times, in how many times the time of the load fall is less than the time of action spent on its raising! And this is the fact, which determines specific character of **creation in the real World**. We should attribute both quantitative and qualitative characteristics to energy. This is the side of energy manifestation, which is reflected in Plank's formula: energy is proportional to frequency.

Manifestation of fundamental law of interaction also lies in the basis of general universal regularity of evolution of real many-particle systems with the change in external conditions. This process develops in multistage way, and on the each stage the logarithm of the ratio between the event happened and the event to happen always is equal to the work of external forces. In other words, the relation of the event happened to the resource is in exponential dependence on the initial conditions and extent of external influence. Exponential character of development of processes is the evidence that Nature develops according to the law, which conserves itself during evolution. This regularity, which manifests everywhere, can be naturally called **the Principle of Order.**

Fundamental law of interactions and Principle of Order appeared to be enough to describe and understand phenomena in the observed World. And it is natural to expect that this principle of Nature manifest in finer World also. This World includes lower and higher frequencies, which are not available for us yet to watch this wide-range frequency-wave emanating Universe.

From all aforesaid we should make a conclusion that the logic, which exists in the traditional physical tool, appeals to the model and principles of the World of non-interacting elements using the range of regularities, which also reflect some features of the real World, but they do not include fundamental law of interaction and Principle of Order, which are necessary and sufficient to describe reality. Descriptions existing in traditional physics are phenomenological ones and concern only those aspects of the phenomena under investigation, which do not include possible qualitative changes during development of processes, because the main property of real processes of interactions (creation of new energy property) was excluded.

The current situation in physics had a strong influence on formation and development of other sciences, other fields of knowledge, since the logic of reflection of cause-effect links was initially excluded. These are the links to determine existence, i.e. existence of constant creation of the World. All these circumstances give grounds to fundamentally revise educational programs, first of all, in physics, philosophy, mathematics, chemistry and biology. A change to the offered logic of cognition, which is based on the Principle of Order and fundamental law of interaction, will fundamentally change our notions about the World as well as will open big opportunities for new technique and technology. A Man has got huge opportunities in cognition and existence, but due to his immorality and features of incorrect aims in the logic of cognition he cannot use these gifts of Nature. We present wider and deeper view on the World and a Man in it, which allow analyzing, watching and operating with those fields of reality, which manifest in finer World, World of higher-frequency energies and other structures of fields. Logic of cognition had not touched these structures yet.

References

1. Galilee Galileo. Selected works in 2 volumes. M.: Nauka, 1964.
2. Newton Isaac. Mathematical principles of natural philosophy. News of Nickolaevskaya Marine Academy, issue IV, V, Petrograd, 1915-1916. Volumes I, II and III, 620 p.
3. Feinmann Richard. Character of physical laws. M.: Nauka, 1967, 160 p.

Old New Energy

Y. I. Andreev, A.P. Smirnov

St.Petersburg, Russia
E-mail: science@shaping.org
Internet: http://www.physical-congress.spb.ru

Two kinds of energy, accumulated energy [1] and free energy [2], are considered as an inexhaustible source of natural energy created by Nature itself. It is ecologically clean and possible to be renewed in natural conditions.

The energy accumulated in substance is released as a result of partial decay of substance in elementary particles. At that, the acquired defect of mass is so small that it does not change chemical properties of substance and is compensated in natural conditions. Physical mechanism of energy-release lies in the fact that an electron in plasma layerwise takes sufficiently smaller elementary particles (electrino) from positively charged atoms or fragments of substance (ions). Electrino give their kinetic energy to plasma, heat it up and move beyond the bounds of reaction zone in the form of thermal and optical radiation. There is no substance, which could not take part in such process of energy-release, i.e. phase transfer of higher form (PTHF). The most appropriate, available and low-cost substances are air and water, which play the role of nuclear fuel in PTHF. It is turned out that usual combustion is also a process of PTHF, in which oxygen is a nuclear fuel and organic fuel is a donor of electrons. In the process of combustion oxygen atoms get the defect of mass equal to 10^{-6} %, which constitutes the so small value that it cannot change chemical properties of oxygen and does not call killing radioactive emanation.

There is a possibility to use energy properties both of oxygen and nitrogen of free air in the process of PTHF. To do this it is necessary to destroy nitrogen molecule at least in atoms or smaller fragments by some initiating influence. It is achieved by electrical discharge, magnetic flow, explosion and other means. These means consume much less energy than produced in PTHF. In particular, such processes were achieved in combustion engines. Such nitrogen mode of operation and combustion is accompanied by oxidation to H_2O, but not to CO_2, which is more effective in energy and ecological aspects. Accordingly, the power of engine increases and organic fuel is saved. Exhausts from this process mainly contain water vapor [3].

PTHF processes with excessive power release (more than consumed power) were also obtained in heat-generators operating with water.

Free energy diffused in the surrounding space could be transformed into mechanical, electrical or another kind of energy by means of vibration-resonance, electromagnetic and other energy systems. Classification of these systems as well as physical mechanism of energy transformation is given in [2]. The known Searl's engines can serve as an example of energy systems working with free energy.

The developed physical mechanisms of energy-release processes will allow to create industrial, stably operating, ecologically clean energy systems, which do not consume organic and nuclear kinds of fuel, harmful for humankind.

References

1. Andreev Ye.I., Smirnov A.P., Davydenko R.A., Klucherev O.A. Natural enegetics. – SPB, Nestor, 2000, 126 p.
2. Andreev Ye.I., Andreev S.Ye., Glazyrin Ye.S. Natural energetics –2. – SPb, Nevskaya Zhemchuzhina, 2002, 104 p.
3. Patent 2179649, Russia, 2000 / Andreev Ye.I., Smirnov A.P., Davydenko R.A.

On General Nature of Forces

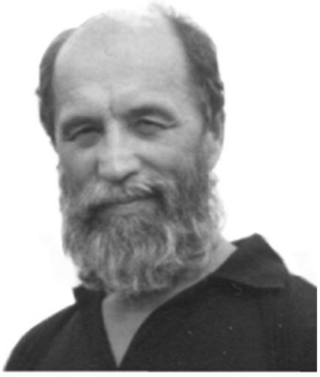

Dr. Evgueni D. Sorokodoum

Entrepreneur and General Director, Vortex Oscillation
Technology Ltd, Volochaevskaya Street, 40-B, Flat 38,
109033, Moscow, Russia
Telephone: 7-095-362-8084
e40dum@dol.ru http://vortex.viptop.ru

Introduction

In techniques and in our life we got used to certain physical notions concerning force. We usually use these notions in creation of automobiles, airplanes, rockets and other techniques, but we don't think about **the origin of forces** in general. Usually appearance of force in continuum is connected with presence of momentum gradient.

A number of works, which describe various versions about origin of a force appeared [1, 2, 5, 8, 17, 21, 22, 23, 25, 30, 35, 36, 38, 39]. Different mechanisms of appearance of force are considered in these articles. Usually they consider origin of a force in one of the fields, which are: electromagnetic, gravitational and others. **There was a theoretical attempt to connect the force initiation with energy gradient [33]. Experimental proof of force initiation due to energy gradient was obtained in the works [7, 38].**

Below we made an attempt to show the general regularity of force initiation, which is connected with non-uniform distribution of energy in space. With this process, physical nature of any kind of energy and specific mechanism of force initiation does not play any role. These are only particular cases of general nature of force initiation.

General nature of forces

We are surrounded by space, which is full of energy. Here we mean the energy of any nature: mechanical, thermal, electromagnetic and others. Energy is related with material world and its value is connected with the volume.

Any particle (volume) of continuum has energy:

$$A = A(x, y, z, t) \qquad (1)$$

where *x, y, z* are Eighler's coordinates of the center of particle, *t* is time.

Transmission of energy from one point of space to another one can take place by various methods, both in connection with energy transmission by material particle itself (which is a "carrier" of energy in this case) and without such transmission (for example, with wave motion). For the volume degenerated in ideal point the energy will be zero. That's why it is more comfortable to operate with the energy density concluded in the increasing constantly. We are the first who analytically got the law of gravity of the masses from the known equation of thermal conductivity. Appeared that on the relatively small distances (in the bounds of the Sun System) the law of gravity by Newton remains valid, but on the larger distances the sudden decrease goes on (Gauss integral), which naturally solves the famous Zelinger's paradox of gravity.

As a conclusion we should note that in the bounds of a stable galaxy of a spiral kind there is the circulation of ether. Ether moves from the periphery of the galaxy to its center (nucleus) by two spiral branches. This becomes apparent as a weak magnetic field (8-10 micro Gauss). In the nucleus of the galaxy there is the impact of two strings as well as there is formation of the spiral toroidal circles (protons). Then the protons form the adjoined vortexes around themselves (electron shells) and from the proton- hydrogen gas the stars are forming, which are moving to the periphery by the same branches. There they dissolve in ether at the periphery since the protons will loose their energy and stability due to the viscosity. Ether which have got the freedom will return to the nucleus of the galaxy and this process is going on in our galaxy for hundreds milliard years and it will keep going until the new center of vortex formation will begin to concentrate ether. Then the new galaxy will appear and our galaxy will disappear. But it will not happen soon and we have enough time to understand that we should return to the concept of ether in modern science.

Experimental Demonstration of Cosmic Influence on the Earth Life in N.A. Kozyrev's Researches

("On the Influence of Time on Matter")

Alexandra L. Belyaeva

Bishkek, 720075, Russia 8th Location, 46, apt.80
Tel.: 7-996-31-41-25-79
E-mail: artem_belyaev@mail.ru
Internet: www.newphysics.h1.ru

Editor's note: this article represents a part of the big scientific conception "World models in the new scientific progress". On applying of this conception a great number of practical technical devices have been created (as an example of such device we offer the description of universal electrical bio-heater, which was created by the group of researchers from Bishkek, Kyrgyz Science Technical Center "Energy" during the *work on ceramic electroconvector).*

We have to note that the position of our editorial board concerning "time" and Kozyrev's work is not in a good correlation with the authors' one.

Nicolay Alexandrovich Kozyrev scientifically and experimentally discovered the action of relations' interconnection, which was falsely named as time. Time cannot cause action because it is absolute and does not have any physical sense (Samat Kadyrov. Monograph "Theory of unified field").

Author's note: relations' interconnection is an interaction of structurally similar objects. It is a nuclear resonant gain-frequency process: in a stationary electric field, which is modeled by systematic organization, there is a development of similar to structural one, in-focus rays of powerful regular coherent radiations. These coherent radiations are determined by properties of chemical components of interrelated substances.

According to N.A. Kozyrev, it is ought to expect not identical density of relations' interconnection in space. Some processes decrease density; others on the contrary increase density of relations' interconnection. Action of the increased density is weakened according to the law of reversed squared distances; it is shielded by a solid matter, at thickness about 5cm, and is reflected by a mirror, according to the familiar optics law. The action of the decreased density on a detector is shielded, but does not reflected by a mirror. Properties of a matter can be changed under the influence of relations' interconnection. In this sense there is a big advantage in changes of electric current conductivity of resistor, which is brought into Witson bridge and is located near some process. For instance, in order to increase density it is useful to realize the process of evaporation of a volatile liquid; and for density decrease the process of cooling of a warmed-up agent can be realized. Due to these processes, change of conductor resistance is actually realized with opposite signs. Increase of density of the conductor with positive temperature coefficient leads to decrease of its resistance. At negative temperature coefficient there is an effect of the opposite sign, in the direction of changes, caused by temperature changes. Such correspondence to fall in temperature should be observed at changes of other properties of a matter, **because disorder in a matter structure is reduced along with fall in temperature**. The researches have shown the following results at the resistor, which was situated near processes of acetone evaporation on cotton wool and of solution of sugar in water. The relative resistance change of resistor was observed at the 6th or 5th digit after comma (or even at the 4th digit if resistors had especially high temperature coefficient).

There is now a possibility to study the Universal World not only by means of the investigated spectrum of electromagnetic oscillations, but also through physical properties of relations' interconnection.

At many researches the influence of relations' interconnection on resistor electroconductivity was investigated. Acetone evaporation (at 10-15 cm distance from the resistor) was applied there as the process, which controls sensitivity of a system. However, the process of evaporation can influence on the resistor not only with density increase, but also due to temperature increase that occurs at evaporation. In order to take into consideration this cooling effect, (in the area of evaporating acetone) temperature was measured by Beckman mercurial thermometer with 0.01°C multiplying factor. The first experiments (without thermal protection) have shown the fall in temperature by several hundredth of degree. This fall was enough to cause the changes of resistor electroconductivity. However, the thermometer had been keeping on the demonstration of practically the same fall in temperature at thermal insulation of the resistor. The thermometer reacted on the radiation of relations' interconnection at acetone evaporation.

The part of the thermometer with a placed in a pasteboard tube mercury tank was laid round with cotton wool and put into a glass retort. The experimental process was fulfilled near the retort, and the reading of mercury altitude in capillary was determined by the scale of the thermometer through the closed window in the next room. The mercury altitude was decreased at dissolution of sugar in water (with steady temperature) and it was increased at the release of the squeezed spring, which was placed near the thermometer.

The radiation of the relations' interconnection was observed from many stars. It is caused by the inner processes, which take place on these heavenly bodies. The Sun (with its turbulent processes) radiates the relations' interconnection besides the searched electromagnetic radiation. Actually, if sunlight is recovered with a thin screen, the significant influence on the resistor will be discovered. The influences of the Sun to the Earth through the relations' interconnection become doubtless. These influences of the Sun should have a particular significance in vital functions of organisms, because it brings the beginning for life support. The totality of the researches demonstrates the dependence of matter state from the changes of the general background of the relations' interconnection. The drift of the devices (that show daily changes) usually stops about at midnight and then changes its direction. As for the seasonal course, there is a density decrease of the relations' interconnection in spring and summer; and there is an increase of it in autumn and winter. It is connected with the absorption of the relations' interconnection by the vital functions of plants and with the return of it at their fading. There are indications at the seasonal changes of chemical processes. For instance, reaction of polymerization has more difficulties in its realization in springtime. V. Zhvirilis observations of minimum and maximum light admission by means of the crossed Nickolya prisms can be explained by the crystalline reconstruction of these prisms.

By Kozyrev, as being invisible, vital source is disseminated everywhere in Nature, thus possibility of its accumulation is the only necessary thing. Such a possibility is realized in vital organisms because all vital functions counteract to the usual course of systems' destruction. The ability of organisms to keep and accumulate this counteraction is the reason, which determines the great role of biosphere for the Earth life. But even if we assume, that spreading of life in Space is one of its peculiar properties, biosphere will not have a decisive significance.

Cosmic bodies (and first of all stars) can serve as the reservoir, which gathers vital source. Enormous stocks of energy flow out of stars in a very weak degree through the radiation of comparatively cold external layers. Inner stars energy is preserved so well, that even at the lack of supplement, matter of the Sun would become cold only at one third degree per year. For the Universe the creative source carries the relations' interconnection. Thus cosmic bodies are necessary for support of life.

Author's note: We apprehend relation's interconnection as natural radioactive background. In fact, it is a nuclear resonance gain-frequency interaction of inertial masses that depends on living systems, especially on its rituals and that regulates its survival. Cosmic bodies regulate this process. Humanity is able to control nature only obeying to natural laws. In-focus beams of powerful laser streams are formed in the electric field of living system organisms. The creation of proton-antiproton pair in the living cells, alongside with the process of the absolute release of energy serves as a creative vital force. The process of radiation, support, absorption of energy by the organization (assembly of particles) is realized through the relation's interconnection and regulates its total mass.

Humanity is able to control nature only obeying to natural laws.

Life without Diseases and Old Aging Preventive Electrical Heater with Programmed Features

Alexandra L. Belyaeva

Bishkek, 720075, Russia 8th Location, 46, apt.80
Tel.: 7-996-31-41-25-79
E-mail: artem_belyaev@mail.ru
Internet: www.newphysics.h1.ru

Universal electrical bio-heater is intended for heating of rooms and preventive clearing of an air atmosphere from disease-producing organisms at continuous exposition (continuous work). The principle of its work fundamentally differs from those of the existing analogues. Carbon crystals are in the basis of bio-heater, which makes it environmentally appropriate.

Bio-heater represents a range of ceramic cylinders, jointed with metal plates on top and underneath. These plates play the role of load-carrying structure. It is used in production areas and living rooms for heating alongside with destruction of pathogen microorganisms. One bio-heater with 0,2 kWtt power is oriented for heating of the area with volume 35-45m^3 (in the future production of modernized models powered from solar cells is planned).

As distinct from the usual oil heater, preventive electrical bio-heater destructs agents of infectious diseases, whereas, according to the researches, oil heater stimulates their reproduction.

Absolute ecological cleanness is obtained by release of the quarters from the effect of increased atmospheric dampness with the temperature, appropriate to sanitary code. Any type of mold or fungus disappears in the quarter and in the future these forms do not renew their existence (even after removal of bio-heater).

The absence of injurious radiations is attained by the following: features of raw material, which is used during the process of electrical bio-heater production; radiation is normal during bio-heater working. Pollution-free temperature influence is attained by favorable infrared radiation.

Among the other properties of electrical bio-heater there are following: fire-safety; explosion proof; chemical inertness; enormous effectiveness from the point of view of electric energy demand. Structural simplicity facilitates its durability; there is nothing in bio-heater to be broken.

Technical aspects (applied Know How): In the process of technologic production of ceramic cylinders, from which bio-heater is consisted, **diamond-like cellular ceramic structure** with superimposed combination of atoms of lattice elements is created. Rhythmic work of cells, which form ceramic mixture, leads to resonance and creates a kind of blow wave (at micro level). This blow wave physically destroys microorganisms that have no calcium framework. It is related only to those microorganisms that are agents of infectious diseases, such as: staphylococcus, enterococcus, enterobacterium, etc. Thus parameters of the evoked blow wave coincide with vibration frequency of the definite types of bacterium and elementals. These blow waves cause the similar effect in room near of bio-heater, e.g. colonies of microorganisms are noticeably decreased there (even at the absence of bio-heater in the nearest room).

Due to its self-organization, bio-heater works in the range of living systems, it is approached to them. There is a realization of active connection with living coaly forms of biological systems. Actually the work of bio-heater is adjusted to them. Bio-heater properties can be programmed at the process of its production.

Bio-heater is a patented product. Patent KR #464 MKI C 04 V 33/24 “Ceramic mixture, possessing heat-radiating properties”. Application #20010075.1 at Patent KR #464 MKI C 04 V 33/24 “The way of creation of energy, renewable, programmed hard-phase ceramic-carbon mass structure”. Application at Patent KR #464 MKI C 04 V 33/24 “Technology of producing of electrical heaters with anti-resonant air prophylactic effect”.

Finale product (FP) purchase is not more expensive than those of existent models of electrical heaters. Cost value is noticeably brought down on organization of the scaled production. It is ought to take into consideration that from all existent types of heating, from the customer’s point of view, this one is the most energy-efficient. Manufacturing of such bio-heaters can be organized on the base of acting industrial production of ceramic fabrics. It will require some expenses. Moreover production service is rather cheap because there is no need in maintenance staff.

Electrical bio-heater can be applied everywhere, where there is a need in: a) economical heating; b) decrease of air moisture; c) disinfections of rooms.

As for the life cycle of bio-heater it does not become obsolete morally and technically. It is produced from the materials, which are not liable to wear.

The invention has a certificate of KR Gosstandart. From the end of 1998 the first unimproved modification of bio-heater (with power 0,6 KWtt) were put into serial production in Bishkek (with small test production runs). This time bio-heaters are readily used as medical equipment in hospitals and maternity hospitals in Bishkek. **Inventor: Alexandra L. Belyaeva.**

Technical Report

The comparison of quantity of heat energy, required for heat of rooms, and of heat quantity, which is produced by Belyaeva's electroconvector.

Mavlyanbekov Sh.Yu.

Deputy Director KSTC "Energy"
Internet: www.newphysics.h1.ru

Editor's note: this calculatious demonstrates the advantages of the device, which at 340 Wtt energy consumption produces about 700 Wtt of heat power.

The calculation of heat output, coming from the ceramic electroconvector to a room, was based on the basis of research statement of EVNA-0.2/220 electroconvector's influence on air micro flora of industrial rooms at 23.10.01. - 06.11.01. period.

The researchers were carried out in the arbitrary room in a four-storied large-panel building. This room was on the 3rd floor, with facing east windows. The room was of 52.5 m^3 air-space, 3.5 m height and 15 m^2 area. The calculation of heat, was made on the basis of "Methods for calculation of the requirement in heat and electric energy of buildings". These methods were registered by Department of Justice of Kyrgyz Republic on 08.09.2000, #154. According to the normative data, temperature of inner air (t_{in}) in the room must be equal to +20° C. In Bishkek planned specified temperature of external air (t_{ex}) for heating is minus 23°C. The average temperature of heating period is t_{av}= -0.9°C, specific heat characteristic of the building is:
q=0.4 Kcal/m^3 h °C.

Medium quantity of heat energy, which is required for heating, is determined by the formula:

$$Q_{av}heating = q \cdot V \cdot (t_{in} - t_{ex}) \cdot 1.12 \cdot$$
$$\cdot [(t_{in} - t_{av}) \div (t_{in} - t_{av})] \text{Kcal/h}$$

$$Q_{av}heating = 0.4 \cdot 52.5 \cdot (20+23) \cdot 1.12 \cdot$$
$$\cdot [(20+0.9) \div (20+23)] = 492 \text{Kcal/h}$$

Thus at the average annual temperature of the heating period, which is: t_{av}= - 0.9°C, the quantity of heat energy required for this room, comes to 492 Kcal/h.

According to the research statement, the trials of the electroconvector with 200Wtt power were carried out at the following external air temperature: +10.2°C; +8.5°C; +10°C; +6.6°C. The calculation data and results of its examination are brought together in a table. The parameters of electroconvector with 340Wtt power are demonstrated in the same table.

Table 1
Table of determination of heat entry and heat consumption's correspondence in the experimental room

#	External air temperature, °C	Inner air temperature °C	Normative heat consumption KWtt (Gcal/h)	Experimentalheat consumption KWtt (Gcal/h)	Percentage depending on normative heat consumption %	Economy of heat energy kWtt (Gcal/h) comparing with power consumption of the device 0.2 kWtt (0.000172 Gcal/h)	Economy of heat energy kWtt (Gcal/h) comparing with power consumption of the device 0.34 kWtt (0.000292 Gcal/h)
1	- 0.9	+20	0.572 (0.000492)		100	-	
2	+10.2	+20	0.267 (0.00023)		100	-	
	+10.2	+16		0.158 (0.000136)	59	0.2-0.158=0.042	0.34-0.158= =0.182
3	+8.5	+20	0.314 (0.00027)		100		
	+8.5	+17		0.233 (0.000200)	74	0.2-0.233=-0.033	0.34-0.233= =0.107

4	+10	+20	0.273 (0.000235)		100		
	+10	+23		0.355 (0.000305)	130	0.2-0.355=-0.155	0.34-0.355= =-0.015
5	+6.6	+20	0.366 (0.000315)		100		
	+6.6	+19		0.338 (0.000291)	92	0.2-0.338=-0.138	0.34-0.338= =0.002

Calculation data demonstrate a considerable economy of heat energy at daily unevenness of external air temperature.

Heat productivity of the new structure of electric convector with 340Wtt power was calculated on the assumption on the suggestion that heating of the room is carried out by the irradiation at the process of heat exchange.

$$E = \varepsilon \cdot C_o \cdot T^4 \cdot 10^{-8} \quad \text{Wtt/m}^2$$

where: C_0 =5.67 Wtt/m^2 K^4 is a radiant emittance of blackbody, ε=0.93 is an emissitivity factor of the surface of earthenware duct tube; T=70°C=343 K is the temperature of the surface of earthenware duct tube.

On substitution of the known values into the formula we get:

$$E = 0.93 \cdot 5.67 \cdot 343^4 \cdot 10^{-8} = 727 \quad \text{Wtt/m}^2$$

As the area of irradiation surface is equal to S=0.96 m^2, then quantity of heat, which is evolved by the convector, comes to:

$$E_k = S \cdot E = 0.96 \cdot 727 = 698 \quad \text{Wtt (or 600 Kcal/h)}$$

The quantity of heat, which is required for the heating of the room, is 492 Kcal/h (at the external air temperature equal to minus 0.90 and temperature in the room equal to plus 200).

Thus, electric convector with 340 Wtt power is able to heat totally the room with 60m3 area.

Editors note: 340 input and 700 output!!!

Longitudinal Waves in Vacuum: Creation and Research

Ph. Dr. Kirill P. Butusov

190121, Saint Petersburg, Angliysky prospect, 5-18
Tel: (812) 113-8511

The author presents a new elegant system, which is the symmetrized Maxwell's equations. In practice it gives a possibility to create the longitudinal waves in vacuum. This system is of great importance in telecommunications and aerospace technigue.

There is a stable paradigm in electrodynamics that the existence of the longitudinal waves in vacuum is impossible. This paradigm played its negative role preventing scientific minds from solving this problem. However, Maxwell was not as categorical in his opinion on this question as his following were.

Particularly he wrote: "Science of electromagnetism as well as optics is not able to confirm or deny the existence of longitudinal oscillations."

Maxwell's dynamic equations are usually considered as partial derivatives in time. However, the total derivative in time includes the so called substantial derivative, which was shown in the equations for the moving coordinate system. In particular, one of these equations was written by Maxwell himself to explain the phenomenon of electromagnetic induction discovered by Faraday. This induction takes place in the conductor moving across the field lines of electromagnetic field:

$$\vec{E} = \vec{V} \times \vec{B}\,; \qquad \text{(I)}$$

Other equations were obtained later by other scientists. In the table I below Maxwell's equations are given in a split form. Their static and dynamic parts are given separately as well as the equations for moving and fixed coordinate systems. Such matrix concept of Maxwell's equations allowed finding their incompleteness. Really, the analysis of the matrix shows its high symmetry. However, full symmetry of the system of equations is broken by the absence of the equation (X). It seems to be strange and calls a desire to remove this defect in such an elegant system of equations.

A new equation is introduced in the Table 1 for the full symmetry of the matrix:

$$\nabla \cdot \rho = -\frac{1}{c^2} \cdot \frac{\partial \vec{j}}{\partial t}\,; \qquad \text{(X)}$$

Fundamental Properties of Aether

Alexander M. Mishin

Author's note: In the article the principles determining major properties of aether are formulated on the basis of an empirical material.

Real aether [1-6], the primary and superfine essence of which is still a secret, has turned out to be absolutely non-standard superfluid three-dimensional material medium, which simultaneously is at solid, liquid and gas phases. The first master phase of aether is a specifically solid absolute space or an energetical "bottom" of the Universe ("celestial stronghold"). At that the solid phase is considered as mesomorphic vortical-wave structure, which has particular holographic properties. Classical matter represents to be one of the stable and energetic space-time levels of the Universe. Aether vortexes exceed all conceivable space scales, have quasi-material properties and create a great number of stereo-dynamic subspaces (parallel worlds).

The first basic principle, to which aether entirely follows, is the principle of the least disturbance (the least action). Many well-known and unknown physics laws are the subsequent of this principle. In particular, any motion in macroscopic aether happens in such a way to minimize the interaction with the matter of our world, with zero moment of the disturbance momentum. In the classical physics this principle has been reflected as Le Shatelye principle, as variation principle, laws of thermodynamics etc.

The second principle is the principle of fractality, which confirms the similarity of forms and properties of quantum aether vortex structures regardless of their space scale. This principle also determines the Universe as stereodynamically multivariate system in the form of hierarchy of vortical-wave structures of the unified aether (fractal matreshka). On the researching of the macroscopic objects of the Universe it is possible to make a conclusion about microcosm structure if taking into account the changes of frequencies and velocities of action transmission.

In the third place there is a principle of physical autonomy, which confirms that any solitary mass (for example a planet) creates aether system. The particular principle of relativity, which reflects one of the fractal properties of the Universe, can be applied to this system. Such autonomous mass becomes similar to the miniuniverse with its aether subspaces, which repeat the basic phases of the Universe spectrum in more narrow (which depends on the size of mass) frequency band of space-time frequencies. Thus, in the local system of the Earth solid aether reproduces the structure of gravitational field with energy "bottom" in the mass center. As the result such spherical body occurs to be an energy drain and warms up from within.

The fourth is the principle of interaction between matter and vortex-wave forms, which do not depend to the spectral part of the Universe, that is quasimatter. This is the principle of new interaction in nature. The value of energy interaction in each experiment diminishes in time according to exponential law that is explained by the forming of energy informational or adaptation barrier, which separates parallel worlds and reflects the properties of vortex tenacity of aether as superfluid medium. At that, time of interaction is proportional to the size of quasimatter and the barrier for the earthly conditions is lowered at the indefinite period, on the assumption of only thrice-repeated observation of forces (triad law).

According to this principle, aether dynamic experiments in the earth laboratory do not have classical repeatability that, from the one hand, gives occasion to doubts in the objectivity and scientific character of the non-traditional experiments and from another hand it is the most reliable test feature of macroscopic aether motions. Biosystems have special relations with this principle.

The fifth is the principle of many-dimensional autobalance of forces. All vortex and linear motions of macroscopic aether organize themselves in the way that in the band of space-time spectrum of the local system (usually with the aid of fluid and gas aether) occurs to be self-balanced, that is they have zero resulting impulse and the moment of impulse due to the existence of the proportionate antivortexes and antistreams of another spectral structure at the same space volume. The self-balanced vortex structures and streams are practically closed for the outer watch from the direction of our material world, at least with respect to the methods of classical physics. The principle of autobalance of forces reflects aether properties as unified synergetic system and has a significant applied meaning.

Let call the principle of viability of aether dynamic systems as the sixth principle. Only a stereodynamic multivariate system is a viable one, that is a system, which during a definite period of time has the opportunity, called as life cycle, to realize interconcerted self-oscillating processes of vortex-wave character simultaneously at different phase states (subspaces, layers) of aether. The most important features of such a system are its space-time quasimaterial (vortex-wave) broadbandness and finite time of existence, which is determined by the conditions of creation of the energy-informational barrier. Self-oscillation regime demands the presence of an energy source, oscillatory circuit (a pendulum) of any character, intensive process (of negative tenacity) and a channel of positive feedback (negative entropy).

In the sense, referred above, any material system is viable and occurs to be a big system in the form of coordinated community of multivariate subsystems. In its turn each big system as a part of the hierarchy is a constituent of bigger system, until everything is embraced by the Biggest System, that is the Universe.

The seventh principle of the universal energy interchange is the physical realization of the law of unity and struggle of oppositions. This principle determines spontaneous creation of thermodynamic and antigravitation potentials. Any local matter mass (a body), situated in the open space, creates an exchange process with the surrounding aether volume in the way that more fine-structure fluid aether is absorbed by the body, and the less power-consuming gas aether is radiated. As the result the body as a heat engine gets energy due to the cooling of aether exteriors. At that, antigravitation forces acts between bodies and aether exteriors, which have different temperature.

This principle, which establishes the existence of antipodes of the second law of thermodynamics and Newtonian attraction, is realized mainly in cosmic scales and explains in which way the energy is created in the bowels of planets and stars and why the Universe is stable as regards to gravitation. Obviously, the most unexpected for the modern Physics is the discovery of non-traditional nuclear processes where conditional reactions of decay and fusion occur at the usage of quasimatter.

More deep research of new experimental results and of the stated above scientific principles lets to determine the priority-driven strategic tendencies in Physics, to open more entirely the laws of mechanics and thermodynamics of many-dimensional aether, including the theory of non-traditional waves and new types of electromagnetism. At that, the supreme aim is the research of differences in aetherodynamics laws on the Earth (in a laboratory) and in outer space, the unknowing of these differences has caused logical insularity, false all-sufficiency of classical physics, which had refused as "not wanted" the aether conception and fundamental Universal laws.

References

1. Mishin A.M. On the new properties of physical vacuum, gravitational field and mass. DD USSR, 1988, p. 44
2. Mishin A.M. Experimental results on the registration of aether wind // New ideas in Natural Sciences. Series: Problems of research of the Universe, issue 18. – SPb: RAS, 1995, p. 24-33
3. Mishin A.M. The Aether Model as Result of New Empirical Conception. New Ideas in Natural Sciences. (On materials of the International Conference). Part I – SPb: RAS, 1996, p.95-104
4. Mishin A.M. The physical system of artificial biofield // "New Energy Technologies" – SPb: Faraday Labs Ltd, 2001, issue #1, p. 45-50
5. Mishin A.M. Antigravitation and new energy processes // "New Energy Technologies" – SPb: Faraday Labs Ltd, 2001, issue #2, p. 37-41
6. Longitudinal thermomagnetic effect // New Energy Technologies – SPb: Faraday Labs Ltd, 2002, issue #2(5), p. 38-41

Irving Langmuir and Atomic Hydrogen

Nicholas Moller

PO Box 201
34008 Eretria
Greece

Editorial

In this paper Dr. Nicholas Moller describes the history of development of Atomic Hydrogen technologies in details. It is remarkable that this technology can be applied not only for welding processes but also as a clean free energy source. It is important to note that in this case the hydrogen process does not involve a consumption of hydrogen, which is not combusted in the process. Atomic hydrogen is not really a fuel but rather a medium, gateway or a super-conductor of ZPE form the vacuum of space, converting ZPE radiation and ultra-high frequency electrical energy into infrared (heat) radiation.

Irving Langmuir.

This is the story of Irving Langmuir who was the first to develop a theory on Atomic Hydrogen on the basis of empirical research and experimentation. His work in this field lasted from 1909 to 1927. During this period he was employed by the Research Laboratory of General Electric Company. Patents and discoveries developed by Langmuir during his time with General Electric were to a considerable extent instrumental in laying the foundations for what is today one of the largest corporations in the world.

The question that gave birth to this article, is why his work and discoveries on Atomic Hydrogen were the only work that received hardly any attention at all and why his revolutionary breakthrough was deprived of world attention for almost 100 years? This question becomes even more relevant when taking into consideration the high standing he enjoyed with his contemporaries (including being awarded the Nobel Prize in Chemistry) and it is better to make it distantly after exclusion of man presence near experimental stands and devices. At the same time it is quite obvious that on applying of small capacity and fixed time of irradiation it is possible to develop methods for curing of human diseases, which are considered now as incurable (for instance of diabetes, some diseases of haematogenic system, of cancer and possibly of AIDS.

References

1. Proceeding of the International Scientific Conference "New Ideas in Natural Sciences" Problems of Modern Physics, p. 176-187
2. Journal of new energy, Vol. 3, #4,1999, I.M. Shakhparonov "Interaction between Kozyrev – Dirak radiation and radionuclides", p. 85-89
3. Kozyrev N.A. Selected works, Leningrad State University, 1991, part 3
4. Lavrentyev M.M. Eganova I.A. Luzet M.L., Fominykh S.F./ / Reports by AS USSR. – 1990, Vol. 314, #2, p. 352-355
5. Lavrentyev M.M., Gusev V.A., Eganova I.A., Luzet M.L., Fominykh S.F.// Reports by AS USSR. – 1990, Vol. 315, #2, p. 368-370
6. Methodological materials on experimental pharmacological and clinical trials of immune modulating effect of pharmacological remedies. - Ministry of Health USSR, M., 1984
7. Talmadge J.E. and Chiragos M.A. Comparison of immunomodulatory and immunotherapeutic properties of biologic response modifiers. Springer Seminar Immunopathol., 1985, 8, 429-443

Effect of Magnetic Blow Wave Field on Wine Systems

I.M.Shakhparonov (Corresponding author), S.A.Grin,
S.R.Tsimbalaev, L.N.Kreindel, V.N.Kocheshkova,
A.I.Podlesny, S.Yu.Gelfand

AGD Firm, Peschanyi pereulok, House No.20P korpus No.1, Lfl. 33
125252, Moscow, A- 252, Russian Federation

Russian Institute of Canning Industry, Shkolnaya Street. 78.
142703 Vidnoe 3. Moscow Region, Russian Federation

Introduction

Authors communicate the data on influence of Magnetic Blow Wave (MBW) field on several wineproducts. It was found, that MBW did not lead to significant changes in the major components of the wineproduct (sugar, organic acids, minerals). At the same time the taste and aroma of treated wine become more pleasant; content of heavy alcohols and wine stone in the treated samples was less than in non treated ones. A mechanism of transformations was also discussed.

Keywords: *Magnetic Blow Wave (MBW), Wineproduct, GLC of aroma compounds and ethanol, HPLC of sugars, Atomic Absorption Spectrometry (AAS) of minerals, Heavy alcohols and aldehydes, Wine stone, Turbidity tendency, Organoleptic evaluation*

Magnetic Blow Wave (MBW) was obtained for the first time during the investigations on ball lighting generation under the laboratory conditions (Shakhparonov 1994). MBW as a physical object is interesting because of some facts, which suggest that MBW is a magnetic monopole. The MBW can also interact with the matter and transforms it in a definite way. Typical example is an elementary carbon in the form of graphite, which is transformed by such magnetic treatment into ferromagnetic substance (ibid). The graphite, which is initially diamagnetic, transforms to paramagnetic one with general radiation doze of about $7{\cdot}10^{19}$ neutrons/cm^2. Other types of radiations could not affect this way (Svoistva 1975). So one unit of MBW can be considered as $1{\cdot}10^5$ of neutron masses. This fact may be regarded as an indirect evidence for assuming that MBW and magnetic monopole are the same things. In the absence of excited radioactivity a slow MBW [$v/c < 1{\cdot}10^4$] occurs, which does not ionize atoms (Devons, 1963). Therefore, their interaction with the matter can be observed only indirectly. No data exist on the interaction of MBW with organic substances. The experiments and results reported in the present communication may be a starting point for development of technology and to formulate the methods for vintage wine and best quality spirit production.

Materials and Methods

Assuming that MBW and magnetic monopole are the same things, a number of conditions were selected for all experiments. The MBW source and the samples were placed in the same axis and the axis was oriented according to magnetic meridian direction. Such magnetic orientation is appropriate, as the energy of magnetic monopole theoretically increases in a magnetic field (Devons, 1963). All of samples were placed at 250 cm distance from MBW source, in hermetically closed glasses. It should be noticed that MBW could penetrate through many other barriers, for example into cast iron reservoir with wall thickness of 5 cm (Amaldi, 1970).

The quality investigations were made by using of standard equipment. HPLC, equipped with refractometric detector was used for sugars estimation. Separation of organic acids in forms of their ethyl esters and acid esters was carried out chromatographically using a column packed with polyethylenglycol succinate and the following temperature option: initial temperature is 120°C, final temperature is 220°C, temperature growth rate: 8°/min. GLC was also employed for determination of ethanol. Minerals content was examined with Atomic Absorption Spectrometry (AAS). Electronic spectra of samples were obtained with double beams UV Vis spectrophotometer equipped with permanent wavelength scanning. Redox potential was measured with EV-74 potentiometer.

The aroma alterations in the wine samples were investigated by GLC method after preliminary concentration of aromas by solid phase adsorption. The concentration was carried out by barbotation of inert gas (nitrogen) through liquid and consecutive catching of volatiles with tube trap, filled by Polysorb 1 sorbent (Lur'e 1972). The well-known analogue of Polysorb 1 is Porapak Q. The tube may be regarded as a short chromatographic column, and volatiles go through it according to their retention times. The choice of sorbent was motivated by the fact, that retention times of water and ethanol was rather small (ibid). Thus, a concentration process can be ended at the moment, when water and ethanol have passed through the column, as the other volatiles remained bonded. The aroma desorption was made with ethyl ester. The analysis of the concentrates obtained was carried out with gas chromatograph equipped with flame ionisation detector (FID), column 3 m x 3 mm, packed by Carbovax M on the Supelcoport. Temperature for the analysis was programmed from 100 to 190°C with increase of 1°/min. Isothermal conditions in the borders had durations of 2 and 40 min respectively. The "mild" conditions of separation were also employed (initial oven temperature was 80°C with isothermal condition duration 5 min, temperature growth rate 1°/min, final temperature 150°C and isothermal condition duration 40 min).

Optical activity was tested with Spectropol at D line of Na (580 nm). The samples were evaluated organoleptically by a group (12 persons) of workers from Russian Institute of Canning Industry. Turbidity tests were made under the methods of Valuiko et al (1987). In some cases, qualitative tests were completed by MPL turbiditymetric measurements. Before testing samples were filtered. Determinations of heavy alcohols and aldehydes contents were carried out in accordance to National Standard (GOST, 5363-67) as follows below. Determination of the constituents of "heavy spirits"(i pentanol, i butanol) was based on reaction of the sample with salicylic aldehyde in a presence of H_2S0_4. Rose colour develops if sample contains the heavy alcohols. The density was measured with Vis-photometer and the quantative determination was carried out using standard graph made with mixture solution of i pentanol and i butanol. A method for determination of aldehydes content is based on a reaction of fuchsine sulphite. The developed colour was measured with Vis-photometer. Calibrating plot constructed basing on typed solutions was used for quantification.

Results and Discussion

Investigations of wine quality changes after MBW treatment were performed using two samples of portwine ("Zemfira") type wine. Sample 1 was a reference (non treated) and sample 2 was treated with MBW.

In both samples, the fructose and glucose levels were practically the same and amounted to 43.8 ± 3,32.22.5 g/*l* respectively. Sucrose and maltose were absent. Total sugar content was 76.0 g/1 though the level marked on the label was 80 g/*l*. It is thus apparent that the treatment of wine with MBW does not lead to noticeable changes of sugars content. Results of organic acids determinations are given in the Table 1.

Table 1
Main organic acids content, g/1

Acid	Treated wine	Initial wine
Lactic	0.0265	0.00187
Oxalic	0.010	0.0088
Succinic	0.209	0.18
Malic	4.56	4.22
Tartric	0.0805	0.0895
Citric	0.401	0.483

Standard deviation for the determination method was estimated as 7 %. This fact shows that differences in organic acids content are not significant. It should be noted that a tendency of slight increase in light acids (up to malic) in the treated wine was observed in contrast to noticeable change in more heavy acids. The ethanol content of both of samples was 181 and 184 g/*l* for non treated and treated samples, respectively, though the label on the bottle indicated 190 g/*l* concentration. Standard deviation was 5 %. Thus, MBW treatment does not lead to significant changes in alcohol content.

Atomic Absorption Spectrometry (AAS) data indicated that the samples were practically identical in terms of K, Na, Ca, Mg, Fe, Cu and Zn contents (data are not shown).

Similarly, spectra of treated and non treated wines, diluted 150 times before photometring, were practically identical, thereby pointed out that polyphenols are unchanged.

When wine is industrially treated with IR or microwave heating, ultrasonic, ultraviolet and g radiation, different reactions occur and there include redox reaction, esterification, condensation, hydrolysis, Maillard reactions, etc (Kishkovsky 1988). Most of reactions are accompanied by redox potential changing. Increase in Redox potential points out the increase in concentration of oxidants, i.e. oxygen, peroxides, and other compounds, which are electron acceptors. Redox decrease is a result of oxidation processes (ibid). Redox potential was practically constant (ΔE= 145 mV and 150 mV in samples # 1 and 2 respectively). Evidently, oxidation processes, like they occurred during heat treatment, were absent during the MBW treatment.

One of the important reactions to be considered is the Maillard reaction. Essentially it appears in form of browning, decrease in reducing sugars and amino acids, and new aromas formation. While our result evidences on absence of irrelevant aromas, alterations of wine colour, and sugar content, thereby indicating insignificant contribution of Maillard reaction on wine quality changes due to MBW treatment. Technological treatment often leads to esters accumulation that improves wine aroma. It's well known that the most important in this context are the esters of C_6-C_{14} fatty acids (Kishkovsky 1988). During heat treatment, storage and other physical influences, different kinds of acid esters accumulate. These have weaker aroma than esters of fatty acids. But their appearance proves the existence of esterification processes. A comparison of aromas chromatograms of the samples 1 and 2 proves occurrence of changes in concentrations of the individual substances (increasing of peaks length with retention times of 13.10, 100.9; decreasing of peaks length of 54.85 min). An order of peaks exit of different volatiles is given in the Table 2.

Table 2

Exit order of different volatiles

Exit order, from published data for Carbovax 20 M	Retention time in our experiments, min	Exit order, from published data for Carbovax 20 M	Retention time in our experiments, min
Acet aldehyde	3.8	i-Pentanol	23.56
Ethyl acetate	4.71	i-Amyl butyrate	
Diacetyl	4.82	Acetone	
Methanol	4.91	n-Pentanol	27.92
Ethanol	6.01	i-Amyl valerate	
n-Propanol		Ethyl lactate	
i-Butanol	13.58	Ethyl caprilate	52.6
Butyl acetate	13.59	Acetic acid	
i-Butyl acetate		Diethyl succinate	
Ethyl valerate		Ethyl laurate	
n-Butanol	17.73	Phenyl ethanol	
Amyl acetate	20.70	Diethyl malate	

A comparison of retention times of components with peaks of standard substances of wine aroma indicates that butyl acetate and i-butanol are very close to peak 2. Data on chromatographic separation with mild condition showed that i butanol and butyl acetate peaks exited simultaneously. Organoleptic evaluation recorded a nice smell in the treated wine, thereby due to the formation of butyl acetate. Data indicated the presence of ethyl malate, ethyl tartrate and ethyl citrate in the samples, in addition to two peaks corresponding to ethyllactate and ethyl oxalate. The large experiment error does not allow any inference on changes of their height after magnetic treatment. Identification of peaks with retention time factor especially in such complex system is not unquestionable. However, the best way is to use the chromato mass spectrometer, which allows inference according to their individual mass-spectrum.

Organoleptic evaluation can depend on aliphatic alcohols content. Determination with GLC shows (Table 3), that their quantity in the both samples is rather small with respect to average values taken from literature for this type of wines. Thus such changes can not be recognized with such evaluation. For both of samples pH was equal to 4.0.

Table 3
Aliphatic alcohols content, mg/1

Alcohol	Sample #1	Sample # 2	Average literature values (Kishkovsky 1988)
Methanol			80-350
i-Propanol			0,3-3
n-Propanol	less than 20		5-50
i-butanol	less than 20	less than 20	20-100
n-Butanol	less than 10	less than 10	2-10
i-Pentanol	less than 20	less than 20	100-250

The results of optical activity measurements indicated, that both samples are not optical active. Filtration, clarification and dilution could not change the optical activity. Perhaps, there is a compensation of different forms of D- and L- compounds in the samples, thus total activity was very close to zero, and magnetic influences could not change equilibrium between the forms.

Organoleptic evaluation of more delicate taste and aroma of the treated sample with respect to non treated one, MBW treated sample as more complete, harmonic, noble, and natural in contrast untreated sample was recorded as excessively bitter and sour in spite of practically the same pH of samples.

It is interesting to test the tendency of wine to make a different kind of turbidity after the MBW treatment. Data showed that both samples were not positive for protein turbidity. In term of reversible colloid turbidity formation, after storage at 7.5°C for 1 day, the MBW treated sample was homogeneous, in contrast to the formation of different phases with different refractometric numbers in untreated sample. Both the phases in untreated sample were liquid, with a density very close to each other, but the borders of phases were like broken lines when crystallization begins in crystallization process. This alteration in untreated sample may be due to micelle state changes or of structurization of product.

The tendency test for polysaccharide turbidity based on the reaction with phenol in presence of H_2S0_4 and determination of the derivative formed by photometrically, indicated, that difference in concentrations of polysaccharide in the both samples are very small, the levels being 119 and 106 mg/*l* for untreated and treated samples respectively. These values are close to range of polysaccharide stability (150-200 mg/*l*), and thus do not allow any conclusion on changes of relative stability of the samples. A tendency for polyphenols turbidity, due to polyphenols associates precipitation upon addition of salt did not show differences. Turbidity, as determined in MPL apparatus, was 15 FEM as against value of 0.2 FEM before testing in untreated sample. These numbers were respectively 14 and 0.3 FEM for treated sample. Thus it indicates that both the samples are very stable with respect to polyphenols turbidity and that the magnetic treatment does not lead to alteration in the polyphenols stability.

The data on the colloid stability indicate, that both the samples showed rather high resistance against protein, polysaccharide and polyphenols turbidities. Besides, treated sample showed higher stability with respect to reversible colloid turbidities.

It is interesting to investigate as to how heavy alcohols and aldehydes, which are often produced, when low-grade technology is used, are affected by magnetic treatment. For these studies, a system of simple mixture, consisting only of spirit and water, was used. Commercial vodka bottled in standard 0.5 l bottles and artificial solutions, containing 40 % of food derived rectified spirit were used. Data showed, that MBW treatment significantly influenced the heavy alcohols content, as the reduction in heavy alcohol was more than two times. In addition, it reduced aldehydes by more than 3 times in vodka, and more than 30 % in rectified spirit. Data indicate that efficiency of aldehydes removal is higher when the sample contained higher level of aldehydes. Thus, the MBW treated vodka and rectified spirit will be better than untreated one. It is however stressed that untreated samples were also recorded as good by sensory panel. So, limits for aldehydes are usually present in high quality vodka established by National Standard (GOST 5363- 67) are 6 - 15 mg/*l*. Thus the organoleptic evaluation of samples does not allow to find difference in aldehyde levels in these samples.

Table 4

The main results of heavy alcohols and aldehydes determination (mg/1) in vodka and solution, contained 40 % of rectified spirit

Substance	Non-treated vodka	Treated vodka	Non-treated spirit	Treated spirit
Alcohols	8.7	2.55	3.38	1.5
Aldehydes	1.5	0.4	0.6	0.4

Sediments formation and its character were also evaluated. The sediment in treated grape juice was dense and more dark, the formless, non crystalline sort, and gel like form. The volume of the sediment occupied up to 30% of total volume. The sediment did not sink or float, nor it stick to the walls of glass. It was found that 100 ml of juice gave about 155 mg of dry sediment. Microscopic investigations showed an absence of any kind of bacteria or fungi in the sediments.

The effects of high energy of magnetic influences on sediment were also investigated. The experiments were carried out with “Portwine Erevanski, vol. 0.5 1, white, spirit content 19 vol %, sugar 10 %, prepared according to GOST (National Standard) 7208-84”. Crystalline sediment appeared on the walls and especially on the bottom of the bottle after the BMW treatment. An amorphous precipitate was also presented, and it can be separated by decantation. Crystalline sediment, after washing with ethanol and drying to constant weight, weighted 69.2 mg, and was of bright brown colour. A tartrate content as a tartrate acid, of the sediment was 59% mass. If it is considered as a tartar (a wine stone) of potassium sodium tartrate, then tartar content in sediment works out to be 86%. If it is considered as a tartar of dipotassium tartrate, tartar content in sediment will be 93%.

Generalization of data shows the positive effect of magnetic treatment on the wine samples, leading to harmonic taste of treated wine and absence of non pleasant tastes. Most of the changes were found to bein the flavour and taste components, which were minor substances in the product. For example esters concentration changes during the treatment. At the same time the content of major components, such as sugars, organic acids, particularly, heavy organic acids, and especially ethanol remain constant. It seems logical from kinetic point of view, when simple processes, like esterification, are preferable with respect to many stages reactions, and reactions with high activation energies, which can go at hard conditions. Also, it seems logical that magnetic treatment may influence on electrical state of colloid species. Thus magnetic treatment can be considered as mild, selective in the comparison with many other physical methods. Nevertheless, the changes lead to acceptable energetical and nutritious value of the product. Data show that difficult problems, such tartar removal, can be solved by MBW treatment.

References

1. Amaldi E. et al in: Preprint CERN, Report 63 13. Search of Dirak Magnetic Pole, 1970.
2. Devons S. Search for Magnetic Monopole, Sci., Progr. (No.204), 601 (1963).
3. GOST (National Standard) 5363 67. Vodka. Metody ispytaniy (Testing Methods).
4. Kishkovsky Z.N., Skurikhin I.M., Khimiyavina (Wine Chemistry), Moscow: VO "Agropromizdat", 1988, 253 p. (in Russian)
5. Lurle A.A. Sorbenty i khromatograficheskie nositely (Handbook on Sorbents and Chromatographic Supporters), Moscow, Khimiya, 1972, 320 p.(in Russian)
6. Shakhparonov I.M. in: Sharovaya molniya v laboratorii (Ball Lighting in the Laboratory) / Collection of Articles. Moscow: Khimiya, 1994, 400 p.
7. Svoistva konstrukzionnykh materialov na osnove uglegoda (The Properties of Constructional Materials Based on Carbon) / Handbook/ Nagorny V.G., Kotosonov A.S., Ostrovsky V.S., Dymov B.K., Lutkov A.I., Anufriev Yu.P., Barabanov V.N., Belgorodsky V.D., Kuteinikov A.F., Virgelev Yu.S., Sokker G.A. Moscow, Metallurgy, 1975, P. 73 77 (in Russian)
8. Valuiko G.G., Zinchenko V.I., Mekhuzla N.A. Stabilizatsiyavina (Wine Stabilization), Moscow: "Agropromizdat", 1987, 160 p. (in Russian).

The Fundamentals of the New Principle of Motion

By The Group Studying Inertialess Natural Processes (GSINP)
123430, Moscow, Mitinskaya Str., 40-1-244 Email: gibip@mail.ru

P. Sherbak

The concepts of active and passive interaction between the moving object and the space form the basis of the new principle of motion.

So as to be more understandable, let's consider what is the old principle of motion. For this we will use the concept of a moving object and the space in which the object is moving. Naturally, material objects and the space can't interact between each other directly, because the space is the philosophical category. In this case we can understand physical essence of natural phenomena easily. In our view, the material objects interact with some fundamental energy of space (FAM), which fills all space with a different density. Thus the energy (FAM) is inalienably connected with the space. One of the first names of this energy is "ether" in the early scientific works. So, for the simplicity we will accept that the object and the space interact between each other.

Thus, all existing methods of motion which have been invented by mankind till the present time are based on activity of the material object that means the one expends some energy to produce the motion, and at the same time space is passive, it means that space does not need to spent any energy to move the object. And so in common case space tries to keep the object in the former state interfering with accelerated movements of the object (in accordance with **the 1st , the 2nd and the 3rd Newton's laws**).

It should be noticed that such method of motion (for the speed, which is much less than the speed of light) takes place both in animate and inanimate natures. In this case the level of energy of motion and reaction of space (or an environment) are not very high. Incidentally, the energy of object can be of different types: electrical, chemical, biochemical, mechanical etc. The common consequence of this type of motion is existence of the ***inertia***. The classical physics can't answer the question: "what is inertia?» The same situation is applied to the concept of ***mass,*** which is closely connected with inertia. The classical physics says that ***the mass is a measure of inertia.***

There is the new principle of motion of material object: the object is passive and space is active. In this case it's more favorably for space in the energy aspect to move the passive object and to spent some power then to keep the object in the present place in the former state of immobility (in accordance with **the 1st, the 2nd and the 3rd Newton's laws**). And so we should introduce **the 4th law of** Newton's **mechanics**. It says that ***there are the systems of coordinates in which the body is moving not rectilinearly with acceleration when this body is in the state of immobility***.

The basic and the main differences of the offered principle of motion from the existing methods at the end of the XX century are the following:

1) The absence of inertia of motion;

2) There are no limits for the speed of motion;

3) The absence of "fuel reserves" "on board" of the moving material object.

Nikola Tesla and Instantaneous Electric Communication

Vladimir I. Korobeynikov

Russia
E-mail: elen@mail.infos.ru

Nikola Tesla (1856-1943), an outstanding inventor, was and still remains one of the most mysterious persons in the history of electrophysics. Whereas the most scientists were moving together in direction of microparticles investigations, as the basis of matter structure and of nature itself, he was going in opposite direction. He had a keen interest in the investigation of electric charge of the Earth as a whole. He was looking for the ways to influence on it, to control its state and methods of its regulation.

Therefore, exactly, the most of his searches, experiments, the purpose of constructions and buildings, created according to his conceptions, cause perplexity and misunderstanding of scientists even in nowadays.

The most mysterious of his main experiments were made in USA after 1904. After Nikola Tesla death in 1943, all his diaries and records over a period from 1904 year had mysteriously disappeared. Probably they were stolen (it was known, what to take). Lost records could "cast light" on one of the most "strange" of his buildings in the form of the enough tall tower, on the top of which a specially created toroidal transformer was placed. This transformer could create there a huge electric potential up to the billion volts.

Nikola Tesla switched on this tower-device, what caused the fright and even panic in mind of people from nearby settlements. Of course! Because of very high electric potential there began air ionization, which spread very high to the atmosphere accompanying by the effect of color play. Such luminous, color-playing sky caused even a horror of people, who knew nothing about the experiment made and its goals. They did not guess that Tesla by means of the electric charge, created of the tower, was influencing on the electric charge of the Earth as a whole (about 600000 Coulomb). There was a global scale in Nikola Tesla's investigations.

There is no point in detailed analysis of the fact that the potential of the tower top influenced on the Earth charge. Interaction of charges-balls with the distortion of field lines, distortion-distribution of charge on their surfaces, induced charge, is beautifully described even in school physics textbooks. In Nikola Tesla investigations the Earth had the role of one of the charged balls. It was possible by changing of charge on the tower to deform electric charge distribution on the whole Earth surface at once. This deformation (electric currents) could be fixed at once in every point of the Earth surface. It is alluring to use this effect for data transfer telecommunication, both on the Earth, and in space.

After such introduction the question "How does the system of instantaneous electric communication for any distance look like and work?" is still opened. First of all, the readers need to know, that such instantaneous communication is possible in principle. The proving it theoretical calculations, are rather difficult for popular interpretation. Some part of readers can take it on trust, and those who are most interested in can apply to works of Oleinik V.P. (quantum physics) the professor from Kiev Polytechnic University. At the minimum there are two necessary works: Oleinik V.P. "Faster-than-light transfer of a signal in electrodynamics. Instantaneous action-at-a-distance in modern physics" (Nova Science Publishers. Inc. New York. 1999) and Oleinik V.P. "Latest development of quantum electrodynamics: self-organizing electron, faster-than-light signals, dynamical heterogeneity of time." (Physical vacuum and nature. 4. 3-17. 2000).

"PC" magazine has devoted a rather significant article entitled "Computers and teleportation" to V.P. Oleinik works, concerning instantaneous electric communication ("PC" #6, 2000). Note, that the author of the given article has also found the possibility of instantaneous electric communication, but by means of materialistic methods, absolutely different from Oleinik's ones, what is most important – two different solutions point to the possibility of this communication. "PC" #6, 2000 in the article "Circles on fields" cited mathematical formulae of the structure of electron electro-magnetic field as an illustration (it refers to the Earth too) that the author of this article has got.

The most attentive readers of that article could notice, that one vector Hz absolutely "ignores" Special Theory of Relativity, since its mathematical expression does not include the velocity of light, whereas it presents in other vectors as a product of electric and magnetic conductivity. Magnetic line of this Hz vector goes to infinity and returns back from infinity. It surrounds the whole Universe. It is alluring to use exactly this (Hz) line for the **instantaneous** communication for any distance.

It is not so difficult to do it. In the Fig. 1 the easiest and most available for understanding line of the **instantaneous** electric communication is shown. A rotating charged dielectric ball (an "electron", isn't it?) is used as transmitter. The ball can be electrically charged up to the limit of charge flow-out into the ambient space. Around the charged rotating ball there appears electro-magnetic field, entirely analogous to the electro-magnetic field of the Earth (and of the electron too). The central magnetic line Hz goes to the infinity and returns

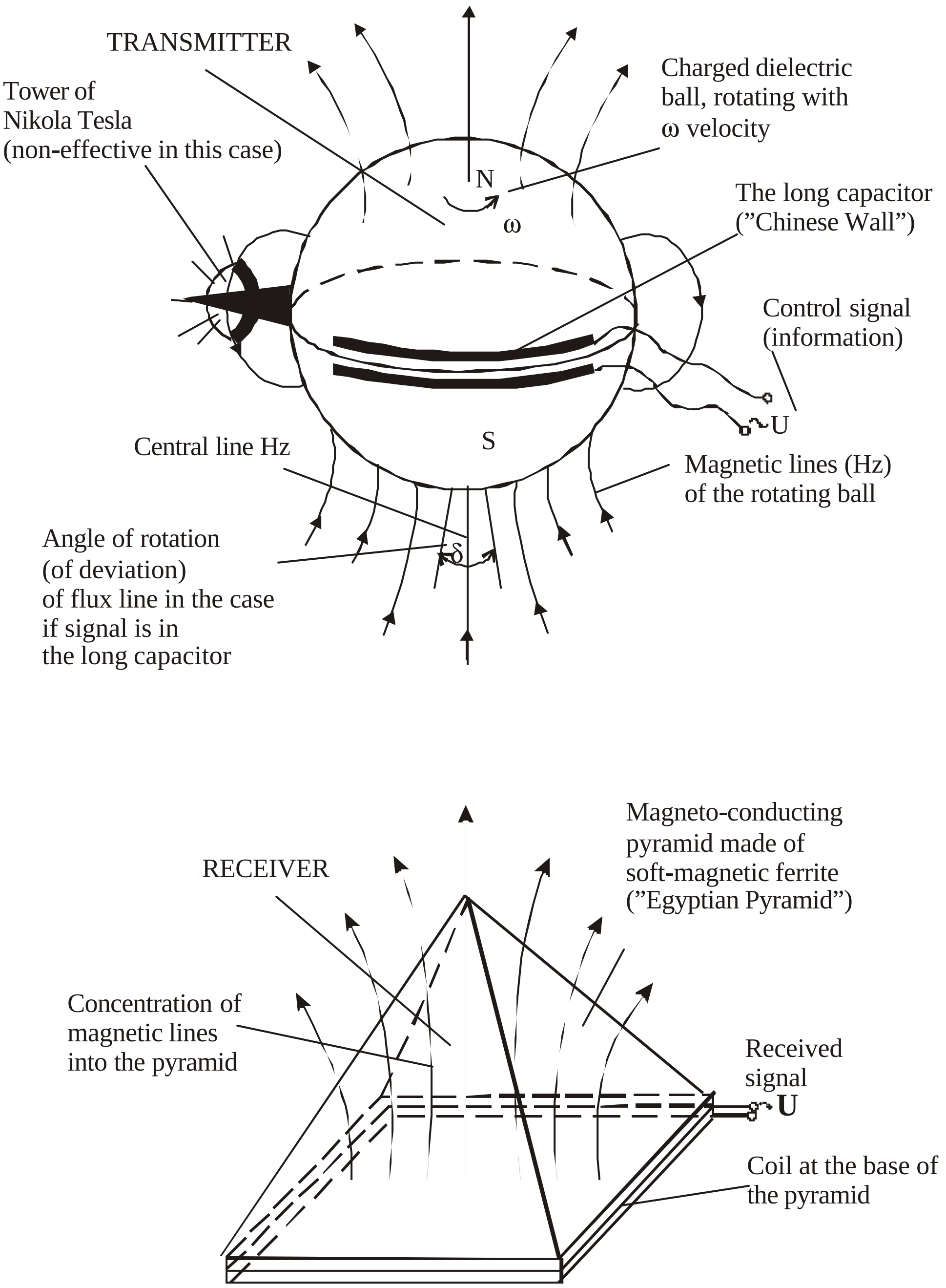
TRANSMITTER
Tower of
Nikola Tesla
(non-effective in this case)
Charged dielectric
ball, rotating with
ω velocity
N
ω
The long capacitor
(”Chinese Wall”)
Control signal
(information)
U
S
Central line Hz
Magnetic lines (Hz)
of the rotating ball
Angle of rotation
(of deviation)
of flux line in the case
if signal is in
the long capacitor
δ
RECEIVER
Magneto-conducting
pyramid made of
soft-magnetic ferrite
(”Egyptian Pyramid”)
Concentration of
magnetic lines
into the pyramid
Received
signal
U
Coil at the base of
the pyramid

Fig.1

Line of the instantaneous electric communication on the basis of rotating charged ball and pyramid

back from it to the opposite side of the ball. In the same way the central magnetic line of the Earth (Hz) goes from one pole to the infinity through the whole Universe and returns from it to the center of the opposite pole.

If by the information to force the rotating ball (electron) to “wag by tail” (by Hz vector) which stretches through the whole Universe, then this “wagging” can be controlled instantly in every point of the Universe. While the rotating ball has a steady distribution of the surface charge, the line Hz does not change its dynamic position in the Universe.

If such distribution of the surface charge is broken, space position of the line Hz also will change. On mounting the Tower of Nikola Tesla on the surface of rotating ball and measuring the potential on this tower in time with an information it is possible to change the charge distribution on the ball, and, respectively, the space position of central magnetic line (Hz) in the whole Universe at once. Big disadvantage of the Tesla tower is that maximum influence on charge is executed in the point under the tower, and farther it began decreasing roughly (exponentially), according to physics laws.

Hence it is advisable to influence on the whole surface of charged ball, but not on some point of it. It is possible to influence at once on the very big part of surface by the long capacitor, placed on the perimeter (equator) of the charged rotating ball. Because of optimality reasons, this capacitor length should not exceed a quarter of the ball perimeter (equator) length. Charging and discharging this long capacitor on the ball equator by the data signal, only the position (angle 5) will be changed, not a value of the infinitely long magnetic line (Hz) in the Universe. It is a data transfer.

The natural question appears: "How to make on the Earth the most powerful transmitter for the **instantaneous** electric communication?" The answer suggests itself: "It is necessary to use the Earth itself as a rotating charged ball." It is not effective to use Nikola Tesla's Tower to deform the Earth electric charge. To place on the Earth very long (about thousands km) capacitor is quite easier. However, it must be placed not on the equator exactly, but moved a little bit because of the initial heterogeneity of the Earth surface charge distribution, caused by the presence of continents and oceans. It will be necessary to find the line of electric equator, where the amount of charge north and south of it is similar. This line will not be ideally straight and will be situated near the 30th parallel.

As a matter of fact, this grand capacitor is already built, but is half-broken. This capacitor is very well known – it is a Great Chinese Wall. The ancient, powerful Chinese Tzcin' Shi Huandi empire adapted and used it (capacitor) for the protection from nomads incursions. How unexpectedly and originally it is! In this case the electric iron would be the best tool for spiking. It is clear enough that the charged ball (as well as the Earth) will "wag by tail", which stretches through the whole Universe and does not change its energy, but only changes its position in space in time with information. Now we can go on to the question, concerning the way to control the Earth "wagging by tail" in the Universe, and thus to read information **instantaneously** in any point of the Universe.

In the Fig. 1 it is shown the input device of the electric communication receiver, made of the magnet sensitive material (it can be soft-magnetic ferrite) in the form of pyramid, with the proportions of well-known Egypt pyramids. Magnetic field lines of the far space pass through the pyramid from the top to the base and are concentrated by pyramid. If there is no signal (the "wagging by tail" of the far planet-transmitter is absent), then the magnetic flow, coming through the pyramid, does not change, and induced voltage in the coil, placed in the base of the pyramid, is absent (no information). If "wagging by tail" begins, then the magnetic flow, coming through the pyramid, will change, and it will cause the appearance of voltage on the coil in the base of the pyramid in time with the information.

Thus, the signal is received instantly. Here it is necessary to remind once again the difference between the instantaneous and usual radio transmission. The usual radio transmitter for the transmission of the information uses the **energy** distortion of space by the information. This energy change in space happens with the velocity of light and hence there is the loss of time for information passing. In the considered case there is no **energy** change in space, there is only a change of magnetic lines position (Hz).

This is exactly the vivid and fundamental difference between the usual electric communication and the instantaneous one. **In other words, in usual transmitter during the fixed time interval there is the change of signal energy (instantaneous value), whereas in instantaneous transmitter there is no this change (only information). This is exactly the fundamental difference.**

Evidently, to receive instantly the signal from the opposite part of our Galaxy, we need rather big pyramid, in order to concentrate a big amount of field lines into the oscillatory circuit under the pyramid. The question can appear: why the pyramid, why not a cone? The point is that lines of the Earth magnetic field (the very lines that compass needle reacts on) in the any place of the pyramid horizontal section have the same density of distribution and are directed strictly parallel to the pyramid base. The cone in its horizontal section cannot provide such uniformity of distribution that is why it is not advisable to use it. From the space magnetic field lines pass through and concentrate in the pyramid strictly at right angle to the pyramid base.

This is the riddle of pyramids wonderwork. Any person coming into a pyramid, at the same moment feels the change of mental and physical condition of organism; whish is very different from that it was before the entering into a pyramid. Of course! Visitors come inside, into concentrated magnetic field lines of the powerful and functioning magnetic core of the receiving electric circuit, what is absent outside the pyramid.

It is strange, but most of tourists are afraid of the ill effect, which can be produced by electrical systems on their health, but there they stand in a queue to feel this effect in pyramids. Concentration and division of magnetic field lines are the easy and effective way to reject a noise, created by the Earth magnetic field.

It is clear, pyramids should be oriented very thoroughly, so that lines of the Earth magnetic field would be strictly parallel to the base and to the opposite (East-West) sides of pyramid. To get such exactness of orientation in modern conditions is very problematically.

The most convenient place to build a pyramid (pyramids) is on the electric equator, in the place of its intersection with the electric meridian. Such place is located in Egypt, near its capital Cairo. And again we meet a paradox: such pyramids are already built on the Earth, but they are half-broken. And Egypt was not less powerful than the ancient Chinese empire.

The Egyptian dynasty of Pharaohs has “completed” and adapted pyramids to burial-vaults, where mummies of dead Pharaohs were buried. Perhaps, it is even more incredible than in China. The impression is given that ancient powerful civilizations on the Earth had a competition between themselves, who will use radio-engineering constructions for instantaneous galactic communication in the most incredible way. Let’s give to a reader an opportunity to select a “winner”.

It must be noted that «PC» already published information that the Chinese Wall and Egyptian Pyramids are radio engineering constructions, intended for the instantaneous galactic communication (PC #114, 1997, etc).

There appears an interest in the possibility to produce very simple and manufacturable systems of instantaneous electric communication right now. Radio-electronic industry can produce them, but still does not guess about it.

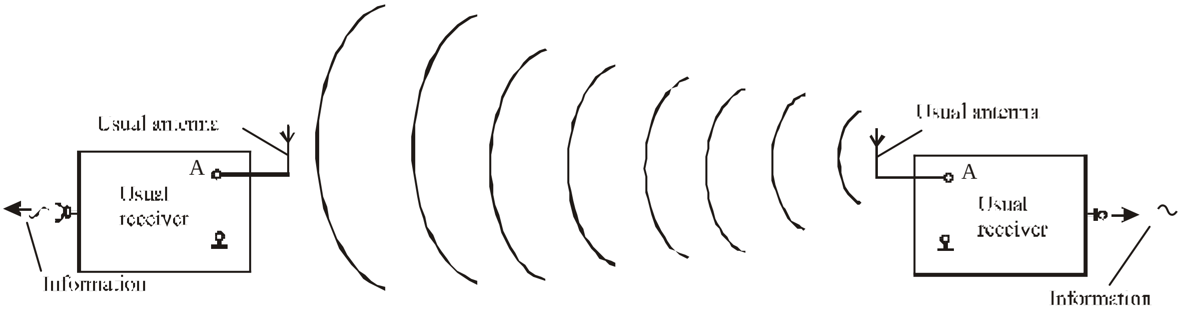
Usual antenna
A
Usual receiver
Information
Usual antenna
A
Usual receiver
Information

Line of the usual (Einsteinian) radio communication

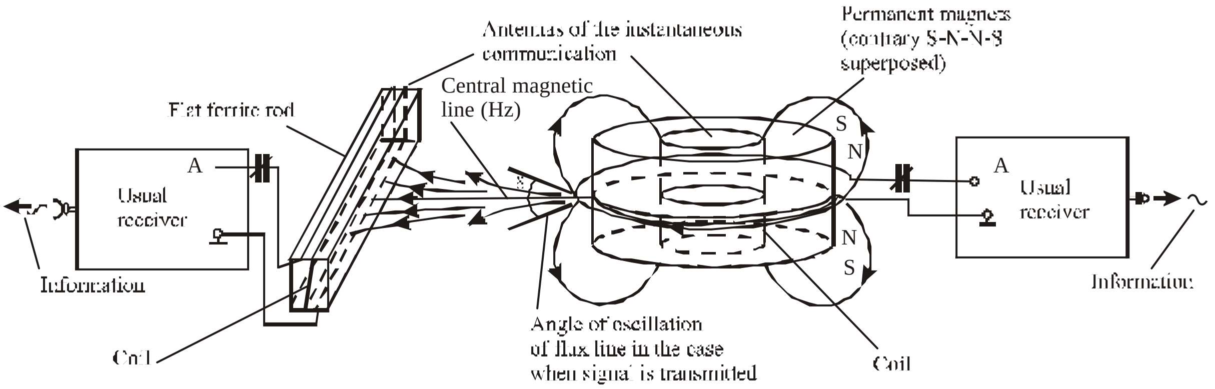
Antennas of the instantaneous communication
Permanent magnets (contrary S-N-N-S superposed)
Central magnetic line (Hz)
Flat ferrite rod
A
Usual receiver
Information
S
N
N
S
A
Usual receiver
Information
Coil
Angle of oscillation of flux line in the case when signal is transmitted
Coil

Line of the instantaneous (Maxwellian) radio communication

Fig. 2

Transformation of the usual line of communication of usual receiver-transmitter into the instantaneous one

Furthermore, such systems of instantaneous electric communication can be created at home, and even senior pupils are capable to use them. In the Fig. 2 there is shown the construction of instantaneous electric communication line in comparison with the usual one. It can be produced even at home conditions. Two permanent magnets, connected between each other by analogous poles, are used as transmitting circuits.

Permanent magnets can be replaced by electromagnets. In the magnet connection point there is a coil, which while the signal passing through it will change its position (angle 5) in the space of the central magnetic line (Hz), coming out from the place of two magnets connection. Receiving circuit is available to be made of the flat ferrite, but coil must be winded along, through butt-ends of core, so that the whole internal part of the coil would be maximal (in area extent) filled by ferrite.

The obtained coil can be completely “winded” (screened) by flat ferrite of big size. For more clearness of the experiment the central magnet line of transmitting part must be directed strictly along the axis of the receiving coil.

Now, if we give the alternating voltage (information) from transmitter to the transmitting coil, fixed on the permanent magnet, then the receiver, connected to the circuit of the transmitting coil, placed on the flat ferrite, will detect an alternating voltage (information). Maximum effect is achieved at the resonance (coincidence of transmitter and receiver frequencies).

It is checked. It works. The dullest experts in radio-electronic (after the reading above) can rejoin, without making an experiment, that it is an absolute nonsense that any communication is out of the question. Coils with absolutely perpendicular axes, besides one of them is screened, do not interact with each other.

And here the most interesting thing starts. In the Fig.2 as it was mentioned above, the usual communication line and the instantaneous one were compared. The usual transmitter cannot generate the vector Hz that is why systems of usual and instantaneous communication cannot see each other in strict sense. What does it mean? It means, that in the same city it is possible to transmit on the one-carrier frequency ("what a nightmare!") two absolutely different television channels without any noises of one to another.

Usually by frequency match of a transmitter to the working frequency of another one, the radio communication is broken, but here it does not happen. Here some additional explanation should be given. As the vector Hz, which "ignores" the theory of Einstein is received from Maxwell equations, it follows that the usual (Einsteinian) system and the instantaneous (Maxwellian) one work on mutually perpendicular electromagnetic field lines (vectors).

In the Fig. 2 such difference is shown clearly. These are just "jokes" of complex numbers, when one value is absolutely perpendicular to another and nevertheless together they form a single whole. In other words it means, that two greatest persons in science Einstein and Maxwell as a matter of fact are something like "Siamese twins", completely grown together at the angle of 90 degrees, even by heads. On the one hand every one is on his own, but nevertheless they are the common (complex) organisms.

Hence there are a lot of misunderstandings on happened phenomena. How many scientists tried to find some mistakes of Einstein? They produced very convincing proofs concerning instantaneous interactions in nature. These scientists did not suspect that time and still do not guess now that they already for a long time are "walking" in the complex physics, which still does not exist. Einstein and Maxwell ("Siamese twins"), each occupies his own part of the complex number (complex physics) and they cannot be already taken off from there.

The only third, free "vacancy" is left to throw on the both of them at once the common "collar" and "reins", i.e. to fasten them ("twins") together by module and argument as any complex number. In this case no matter how the one part of complex number "ignore" the other one, only its argument will change, and module always will be equal to the conditional unit. Only this single fact in principle changes the conception about "appearance" and "disappearance" of elementary particles! Even considered instantaneous and usual communications on the same receivers and transmitters in a complex conception (complex physics) eliminate appeared "misunderstandings" of all kinds.

Advantages of the instantaneous (Maxwellian) electric communication are especially evident during the connection with long-distance spacecraft. At present in the interval between sending of control signal to the station in region of Solar System peripheral planets and getting the reply it is possible to have a small break for dinner (it is very convenient).

In the case of the instantaneous electric communication use, duty operators will have "no dinner". Moreover, the system of instantaneous electric communication can realize two-way communication underwater and from underwater to overland. It is clear that input and output circuits of such system must be covered by slushing composite for the protection from aggressive effect of the salt sea-water. Such systems of instantaneous communication are very required to submarines.

Now, when readers know and understand the principle of operation of instantaneous (Maxwellian) electric communication systems and their advantages over usual ones (Einsteinian), we can only wait, when radio-electronic industry will start to produce these very required systems.

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors, fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908
fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com
website: http://www.ipsgeneva.com

The Unified Gravitation Theory

(The unified super-principle, which controls the Universe)

I. P. Kuldoshin

Orenburg, Neftyanikov str., h. 2, apt. 9,
460019, Russia

(Editor's comments by Alexander V. Frolov)

A forum of the leading USA physicians took place in the White House in March 1998 in presence of President Clinton. There was only one question: "When will the nature of Gravitation be opened?"

The well-known USA physician-astrologer S. Hoking declared that it possibly would occur in twenty years and it would be **the Unified Theory of All.** So, the scientific world by default called it the greatest discovery of the future.

Some time later a new hypothesis pretending to this discovery has got its birth in Orenburg. Despite this fact this hypothesis would gain recognition and status of the **Greatest Discovery of Mankind** only by 2018 that was predicted by S. Hoking.

To present day there have been written a lot of hypotheses on this problem but they haven't been recognized. Many scientists consider our Universe as living and functioning according to the unified and rigid laws in Macro and Microworld, which provides automatic regulations of all its processes due to **circulation of radiant energy of the Universe life in cosmic space**. This energy is inexhaustible and environmentally clean, and Mankind may learn using it in the nearest time for the welfare and for prevention of **contradirectional irreversible ecological** catastrophe.

There is no alternative for humankind to escape and it will not appear in the future. Only cosmic energy will save us. From the book "Secret Doctrine" by E.P. Blavatskaya we can get complete information about the fact that a highly developed civilization of Atlases on the Earth had a "**General Theory of All**" yet 10-12 thousand years before our civilization. They had no automobiles, but instead they had flying objects (aircrafts) "Vimana" of various types as well as ships and submarines, on which they also used **Cosmic energy.**

While reading an abstract in General Soviet Encyclopedia, I got acquainted with the theory of "Aether wind", which was abolished in the beginning of the XX century, and then I understood that this theory contains a deposit to discover the nature of gravitation.

The nature of gravitation is the only one and there are no alternatives in theoretical as well as in physical sense. When scientific world of entire planet abolished the theory of "Aether wind", it lost the possibility to discover the nature of gravitation. The XX century was marked by a revolutionary development of scientific and technical progress, but there was an almost 100-year stagnation in cognition of the Universe elements.

The theory of "Aether wind" supposed that all Cosmos is filled with aether particles flying with the speed of light (these particles are "neutrino" according to modern understanding). **The role of gravitation, carrier of light and retarding medium in Cosmos** was attributed to this motion of particles.

But this theory allowed chaotic motion of particles, which is impossible in mechanism of the Universe, which is adjusted up to automatic mode. Besides, motion of these particles is not possible without an absolute buffer unit, which prevents their head-on collision at the speed of 600000 km/sec (it is thermonuclear explosion and death of matter, i.e. the Universe). E.P. Blavatskaya wrote that Cosmos is filled with radiant energy of the Universe life, luminophore, electromagnetic aether. Thereby she predicted a ready solution to make correction in the uncompleted theory of "aether wind". On the basis of above stated and due to the **un-assumed dawning up,** the theory of "aether wind" was completed. It was the ground to develop a hypothesis of radiant "aether wind". Particles of this wind (neutrino) are electromagnetic particles and move with the speed of light in all directions as **contradirectional paired single-stream flows** (like electrical current in twin-wire cable). Due to this, an absolutely stable concentration of these beams in cosmic space is provided according to the principle "what has come in, the same has gone out".

The hypothesis formulates new views on the problem of structure of elements of the Universe material world.

Some separate conclusions do not match the views of modern scientific thought on the problems of physical principles of material world structure and functioning of the Solar system.

List of topics of the hypothesis

1. The hypothesis disclosures the operating environment of a super-mechanism, which controls the Universe (it is a radiant "aether wind").

2. It disclosures the nature of retarding mechanism of flying objects in Cosmos (its name is Lorenz-Fitzgerald compression).

3. It proves the absence of Universal gravity and beams of light as we usually conceive it. (The beams of aether wind collide and compress matter. An alternative to the notion about beams of light is a temperature wave impulse on the beam of aether wind. It explains why the speed of light doesn't depend on the speed of the source of light. Light is a "passenger" on the beam of "aether wind").

4. This hypothesis disclosures the mechanism of stablization of rotary and orbital movement of the Universe matter in macro- and micro world due to retarding medium in Cosmos.

5. It disclosures the mechanism of reverse rotation of Venus due to the forces of autorotation.

6. It disclosures the mechanism of reverse orbital movement of planets and satellites of planets. (Such a planet had not been opened yet, but there are 6 satellites in the Solar system, which move counter to the others, and it is not an occasion, but **a particular case of the effect of aether wind beams**).

7. It disclosures a real nature of Tungusska catastrophe. (There were about 100 hypotheses, but **neither of them** was recognized to be true).

8. It disclosures **the nature of gravitation** and gives an explanation that gravitation can be:
   - usual (vertical);
   - horizontal;
   - circular

It is important to note: not the entire matter takes part in gravitation, but 1/3, i.e. 33,3% of matter.

9. It disclosures the **nature of Levitation** and proves that 1 liter of water on the surface of the Earth can have the weight from 0 up to 3 kg.

10. It disclosures the role of gyroscope effect in life support of the Universe. The gyroscope effect allows transformation of translation energy of radiant "aether wind" to the rotational energy for practical needs of humankind.

11. This hypothesis gives scientific and technical recommendation for creation of cosmic energy converters.

12. It gives scientific and technical recommendation for producing of levitation effects for any technical systems.

13. It disclosures the possibility of cosmic flights with the super-light speed.
14. It explains experiments on metering of horizontal gravitation (The first experiment was made on February 27, 1999).

15. It disclosures the particle (neutrino) of original matter of the Universe and gives its characteristic. (Ancient thinkers called modern "neutrino" as "Aether", and it was not occasionally, because its diameter is in 10^{25} times smaller than atom's diameter.

All matter of the Universe consists of the same indivisible particles "neutrino" presented by three groups:
- "energy" group, which is in the beams of "aether wind";
- building group, which forms the part of any micro particle;
- free group (neutral-reserve) as a building material for new matter and operating environment of all electromagnetic processes.

All neutrino of three groups rotate with the speed of 3×10^{43} rps (equatorial speed of neutrino is equal to the speed of light).

Fields are formed in every particle as a result of rotation:
- strong field of a small volume doesn't allow particles to close up;
- weak field of a big volume is a general mechanism of gravitation.

As scientists write at the present time, the World is subdivided on a dense world (which we can see) and fine world (invisible). At that the density of such world is in 10^{15} times less than density of water.

It is known in science that all matter of the Universe both great and small rotates and is a gyroscope. Particles of matter get rotation with their birth, thus the fields are born in them simultaneously. **Matter cannot exist without rotation, which generates fields.**

All mechanism of interaction between three groups of particles is based on the mutual repulsion. This is the only mechanism, which always and automatically is able to create the necessary stable interval between the particles and only this mechanism provides the function of gravitation.

Many scientists of the late XX came close to the discovery of the nature of gravitation, but they didn't accept a thought to conceive **the motion of aether particles as a pair-counter flow**. And there are three necessary conditions to realize gravitation:

1. The particles should have the fields of repulsion.
2. The contradirectional flows should envelop the particle of matter from two sides.
3. While one beam is passing a matter mass then the force of fields should decrease and gravitation effect should appear.

Mechanism of gravitation

Gravitation appears due to the intersection *(Editor's note: interference)* of fields, produced by beams particles and fields of the visual matter. As it was mentioned above, the beams are paired and contradirectional. Usually the beams in cosmos are mutually balanced and they do not call gravitation effects.

But on the surface of the Earth the contradirectional beams are not similar in their power. The powerful beams come from above, i.e. they only penetrate the atmosphere, and the weakened beams come from below, i.e. they penetrated all the Earth. Thus, gravitation appears.

Gravitation is a unique property of "aether wind" beams to loose part of their power during penetrating of matter mass. **Gravitation** is the difference of forces of contradirectional beams. *(Editor's note: Really other authors reported this idea also. I cannot find who was the first in discussion about gradient of aether as nature of gravitation.)*

Horizontal gravitation

As a particular case, there is **horizontal** gravitation on the surface of the Earth. It appears on the boundary between lowland (of the sea) and plateau. In this case one beam goes above the surface of the Earth (water), and the counter beam penetrates mountain range and weakens. The first measurements of horizontal gravitation effect were made on February 27, 1999 on the route Orenburg – Samara at 49 km before Syrtinskiy slope.

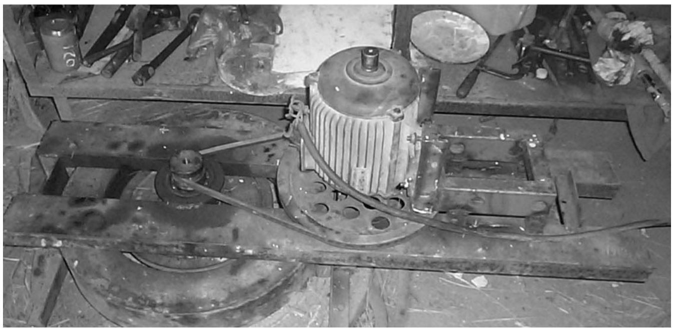

A leaden load (0,5 kg) on the float (a piece of foam plastic) moved on the water surface (not in the sea but in basin) towards the mountain.

Horizontal gravitation is much more weaker than usual gravitation, but it can reach the value that makes water to flow at some angle upwards.

Circular gravitation

Only fast-rotating bodies can create circular gravitation.

All bodies rotate by their orbits around the Sun in the open space of Solar system **due to circular gravitation guided by rotating the Sun.** Furthermore; circular gravitation always is direct (co-directional to the Sun rotation) and reversed gravitation on the periphery of Solar system. A planet with reverse orbital movement had not been discovered until now, but 6 satellites of planets in the Solar system have reversed orbital movement.

Here is the proof of the fact that circular gravitation appears only around the fast-rotating bodies and **slow-rotating bodies, for example, the planets Venus and Mercury cannot form circular gravitation**, that's why they have no satellites.

Our Sun is a prototype of mechanism to transform translation energy of aether wind beams into rotary energy.

(Editor's note: According to Kozyrev, any star is a transformer of time (chronal type of energy) into heat energy. Really, the aether wind can be considered as the chronal type of energy in our understanding and for our usual three-dimensional measurement equipment. To my mind it is a clear link to notion of 4-dimensinal objects, i.e. the time. ***Time can be described by parameters of the aether wind, i.e. its velocity, direction and density. So, we can say that quantitatively time can be described by formulations for kinetic energy of the aether movement. From the other hand it is equivalent of heat energy, which can be measured by usual methods after transformation of the longitudinal waves of the aether in transverse electromagnetic waves).***

Therefore, any mechanical disk rotating very fast will create a circular gravitational field, which is able to rotate all bodies in the direction of the disk (for example, a rim mounted on its bearing co-axially with rotating gyroscope). I designed and tested a similar device in January 2000. A gyroscope (of 200 mm diameter and 3 mm thick) was over-speeded up to 18 thousand rpm. Rotation of gyroscope called slow (but with a good momentum) rotation of the rim of 15 kg weight.

The gyroscopes with the mass of 0,5 kg, 15 kg and 90 kg were tested during summer of 2001. All them called rotation of the rims.

(Editor's note: There are other experimental facts. Fast rotation of mass should produce rotation of some part of nearby aether. Self-closed aether forms vortex and if photon is trapped by this vortex, then

experimenters can see "ring of light" near rotating mass. The rings or self-closed photos can exist in the same place after the mass was stopped or removed away.)

Nowadays gyroscopes in military devices are over-speeded up to hundreds of thousands of rpm. The more rates the gyroscope has, the more energy the rim will produce if it is connected to some generator. But these research works led to single-valued conclusions that gyroscopes themselves cannot produce big quantity of additional cosmic energy not jointly with permanent electromagnets. The Sun as well as planets has natural electromagnetism and their circular gravitation increases in many times due to the presence of electromagnetic fields.

(Editor's note: I think it is obviously that a rotating magnet can involve into the rotation much more quantity of aether than any simple rotating mass. In some theories any magnetic field is considered as circulation of aether particles.)

Electromagnetic fields are the unique boosters of circular gravitation. So, the gyroscopes themselves cannot produce necessary quantity of cosmic energy per unit mass of gyroscope without using of electromagnetism.

In October 2001, I got a copy of 24 patents description. There were patents on "perpetual motion machines". But since such "perpetual motion machines" cannot exist in reality, then we can explain them as gyroscopical transformers of cosmic energy. Efficiency of these transformers varies from 150% up to 10^6 % and practically all of them work using gyroscope. But nowadays only the transformer (Bauman's machine) works in Switzerland, in Maethernitha theological community, Linden city. Some systems have been working from 1980 and producing total power of 750 kWtt, the gyroscopes of 2m diameters are provided with constant magnets.

Besides, there are ready transformers of cosmic energy in Russia. The Professor of Moscow State University, Academician of Russian Academy of Natural Science Leonid Leskov spoke about them in the first half of 2001. He actually said that Mr. Chubais does not allow innovation of energy transformers, which are ready for commercialization (see newspaper "Raduga", Samara, July 2001).

I assume that any kinds of such transformers work on the energy produced by "Aether wind" beams. Perhaps our earth ancestry (Atlases) used this energy to fly as well as extraterrestrials. I remember information about flying platforms, which were designed in Germany in 1943-1945. Nowadays there are publications that there are not less than 10 captured extraterrestrial's spacecrafts on the Earth, and some samples were tested in Russia and the USA.

On the basis of all above-mentioned it becomes extremely clear that the main secret of Nature was discovered, and let's representatives of conservative science don't pull the wool over people's eyes to prove that "it is impossible". It is possible! Physics is an experimental science in its main part, and there is no completed theory until now.

As a result, I'd like to make some conclusions: The secret of Gravitation nature was discovered not in connection with new scientific investigation, but due to dawning up and understanding of the fact that gravitation since earliest times was produced by "Aether winds", which fill all cosmic space. Instead of improvement of "Aether wind" theory, academician science abolished it and forgot it such as some scientists of nowadays don't have an idea of it. While abolishing of "Aether wind" theory, scientific world spent 100 years in vain to find an alternative to it. A real Cosmic scientific and technical progress was slowed down during this term. Without this progress all humankind will kill environment of the Earth in 30-40 years!

Rush hours for humankind to turn to cosmic energy came, we have not even an hour to wait, and otherwise we will loose a chance to survive. Today the scientific and technical level is such that taking into consideration the buildup made by inventors – enthusiasts, who created more than 50 types of Cosmic energy transformers, it is possible to begin repetition work in one year. Now there is the only barrier to do it, i.e. market relations in energetics developed during last 100 years.

Let's look into near Future. The process of energy resources (coal, oil, gas) formation in bowels of the Earth took hundreds million years. There was period of clean ecology in the World Ocean, on land and in the atmosphere. **And all it catastrophically had been diminishing during 2-nd half of the XX century**. **There are about 40 years for our civilization to reach the boundary of having no chance to support normal life on the Earth.** An irreversible process of struggle for survival using underground environment and protection from mortal ecology will begin. Our close posterity will not forgive us this betrayal.

Is there any solution? Yes, there is.

It is necessary to publish the descriptions of all "perpetual motion machines" models as well as unprofitable publication of short technical documentation in the Internet and magazines, which will give a chance to many companies, research groups and individuals to re-produce them. But at first we should choose the models, which are the most reasonable in technology and prime cost. Such a way of replication of the models will give people confidence, interest and reliable information on existence of inexhaustible salutary cosmic energy. And the victory will be the reward for courageous, enterprising and advanced people.

which generates electrical voltage $E_{\updownarrow} = \ell \frac{dB_{rot}}{dt}$ in the structure of vacuum. This voltage generates Gravitational Impulse itself $G = 4\pi E_\sigma S \cdot (\Delta r_g)^2$, where $\Delta r_g = e_0 \frac{E_{\updownarrow}}{b}$.

The supposed "Gravitational Impulse" in the experiment by Podkletnov is modeled by a quarter of cosine curve.

Duration of this curve is determined by the decrease of magnetic field "trapped" into the superconductor due to the partial heating of semi-conductor emitter after plasma passed the discharge of 2MV with the current strength of 10000 A. The formula of the model is the following:

$$X'' = Ae^{-2\pi f_0 D_0 t} \cos(2\pi f_0 \sqrt{1 - D_0^2}\, t) \qquad (1)$$

the calculation is made for the frequencies of 30, 3, 0,3 and 0,03 Hz and acceleration of 12 m/sec^2, which appears for the mass of the pendulum 30 g with the force horizontally to gravitation 0,03·12=0,36 N.

It can be supposed that it is necessary to make more careful solution of the problem to find the effect on the pendulum by its reaction, which is known from experiment. We should apply the more correct use of spectral method of solution of differential equation for the pendulum with setting of impulse effect. Further, having the recording of temporal function of magnetic field by Hall-effect devices and using Maxwell formulas, we should find electrical field acting in physical vacuum. This field will give us the force of gravitational impulse.

The experiment by V. Roshchin and S. Godin is simpler for physical modeling (Editor's note: the author assumes it is simpler than Podkletnov's effect). All input and output parameters are known to the authors, i.e. force of the magnets, frequency of variable magnetic field in the local place of space vacuum, change of gravity. Furthermore, there are known cylindrical formations of magnetic "loops" around the device and their approximate arrangement with the intervals divisible by the half of rotor radius. Effects of temperature decrease at 8° C in cylindrical atmospheric formations can be simply explained by adiabatic decrease of air pressure due to the decrease of gravitation between molecules of air. Formulas for estimation of decrease of gravitational and inertial forces are the same that for Podkletnov's experiment:

$$E_{\updownarrow} = \ell \frac{dB_{rot}}{dt}, \qquad (2)$$

$$G = 4\pi E_\sigma S \cdot (\Delta r_g)^2, \text{ where } \Delta r_g = e_0 \frac{E_{\updownarrow}}{b}. \qquad (3)$$

References

1. Podkletnov E., Giovanni M. Impulse Gravity Generator Based on Charged Superconductor with Composite Crystal Structure.
2. Roshchin V., Godin S. Experimental research on physical effects in dynamic magnetic system // Letters to the Journal of Theoretical Physics, 2000, vol. 26, issue 24.
3. Rykov A.V. Principles of full-scale physics. // Institute of Earth Physics RAS, M.: 2001, 58 p.

New Sources of Energy from the Point of View of Unitary Quantum Theory

L.G. Sapogin,

Department of Physics, Technical University (MADI)
Leningradsky pr. 64, A-319, 125829, Moscow, Russia
E-mail:sapogin@cnf.madi.ru, sapogin@cnf.madi.msk.su

Yu.A. Ryabov,

Department of Mathematics, Technical University (MADI)
Leningradsky pr. 64, A-319, 125829, Moscow, Russia
E-mail:ryabov@vmat.madi.ru

V.V. Graboshnikov

Representative in Moscow of Sceptre Electronics Ltd.
Millennium House Business Center
12, Troubnaya Street, Moscow 103045, Russian Federation.
E-mail: sceptre@orbita.ru

Abstract

The Unitary Quantum Theory (UQT) is a new version of the field quantum theory, which has been developed by the principal author (Prof. L.Sapogin) of this paper for over 25 years. The theory is directly related to the problem of new energy sources, and this paper can be of interest for Journal of New Energy for it is the UQT (and not the classical Newton mechanics or the modern standard quantum mechanics) that provides a theoretical basis for the development of new sources of energy and for the explanation of the operation principles of the existing and functioning over unity devices.

The fundamental provisions of the UQT and a number of results received on the basis of it were published in many scientific journals and reported at international conferences (see [1-6], etc.). Generally, the UQT as expressed by the language of formulae and equations represents a new mathematical model of interaction and movement of elementary particles in the form of a complicated system of non-linear integral-differential equations, an important property of this model principally defines the trajectories and velocities of the particle movement in space (unlike the standard quantum theory, which directly defines only the probabilities of the presence of the particles at a certain point in space). Another, and the most essential (for the problem of new energy sources) property of the UQT is the absence of the energy conservation laws and the impulse for single particles in it. That is why the UQT makes theoretically possible processes of energy generation as if from nothing, if they are regarded from the classical mechanics point of view or the standard quantum theory (while the UQT is able to explain the phenomenon), as well as creation of a device with efficiency above 1. In other words, the UQT provides for a theoretic possibility of making a perpetual mobile!

In the 1970's, when the UQT started to be developed, there was nearly no data of the observed phenomena, or any experimental results confirming this unusual theory. Today, such data are abundant. For example, such processes can be named as generation of excessive heat energy during cavitation of very small water bubbles; generation of excessive electric energy in an anomalous gas discharge; excess generation of electric energy when electric current passes through proton-conducting ceramics, etc. Besides, and still more important, operating devices that have been created much more energy than it was necessary for these devices functioning: electric current generators "Testatica" (Switzerland); thermal cell CETI (J.Patterson, USA); heat generators (Yu.S. Potapov, Moldavia, J.Griggs, USA); electric current generators (P.Correa, A.Correa, Canada); electric engines on magnetic ceramics (Japan), and others. The said phenomena and operation principles of the above-mentioned devices can be explained with the help of the UQT.

In this paper we will also touch upon such an important problem as cold nuclear fusion. The feasibility of this nuclear process, which is categorically denied by the standard quantum theory and nuclear physics specialists, was predicted by the author of the UQT as far back as in 1983. This phenomenon was discovered in 1989 (electrochemical experiments, M.Fleischmann, S.Pons). Many subsequently received experimental data confirmed the existence of nuclear reactions under very small energies, of nuclear transmutations in plants and biological objects, very slightly connected with generation of energy [7-8]. From the point of view of the UQT, which provides an explanation of the cold nuclear fusion mechanism, this process can be applied in practice (after the relevant devices are designed) for generation of energy, for production of isotopes, and for nuclear waste liquidation.

Introduction

Inventors, as well as swindlers of all kinds, had long ago been trying to construct, or at least design, a *perpetual mobile*, i.e. an imaginary machine that produced work without any outside energy. Peter the Great even founded the Imperial Russian Academy of Sciences for such research, but the modern Russian Academy of Sciences does not like to recollect this circumstance. On the other hand, the French Immortals in 1755 decided not to consider any *perpetual mobile* projects at all, and, as we would see, were quite right as regards the Newton mechanics. The brilliant success of the classic mechanics has strengthened still more the sacred belief of the mankind in the Divine Infallibility of the Conservation Laws, and today it is nearly indecent to express any doubts about these laws.

Let us first of all find out the origin of the conservation laws in ordinary mechanics. Practically any textbook will tell you that the Energy Conservation Law (ECL) follows from the homogeneity of time, the Impulse Conservation Law from the homogeneity of space, and the Angular Momentum Conservation Law from the isotropy of space. That is why many people have an impression that the conservation laws themselves follow only from the quality of time and space, which is today an undoubtedly relativistic notion. But, for example, the angular momentum is not a relativistic notion. So, such a narrow approach is not altogether correct, and it is necessary to turn to the second Newton law, or the equation of relativistic dynamics and the system insularity. However, *the qualities of the time and space ensue exactly from the analysis of the Newton mechanics*, though they are often construed incorrectly. Let us remind you the correct interpretation.

Homogeneity of time suggests that *if at any two moments of time two similar experiments are made in similar closed loop systems, the results thereof will not differ.*

Homogeneity and isotropy of space mean that *if a closed loop system is moved from one part of space to another, or is oriented differently, nothing will change.*

The making of the fundamental energy and impulse conservation laws from the Newton equation is very simple. Let us put down the main equation of dynamics as

$$\mathbf{F} = \frac{d\mathbf{P}}{dt}$$

For closed loop system $\mathbf{F}=0$ (no external forces operating) and the equation integral will be $\mathbf{P} = Const$ the impulse conservation law.
Now let us take the main equation of dynamics as:

$$\mathbf{F} = m\mathbf{a} = m\frac{d\mathbf{v}}{dt}$$

and multiply it scalarwise by $\mathbf{v}$

$$\mathbf{F}\cdot\mathbf{v} = m\frac{d\mathbf{v}}{dt}\mathbf{v} = \sum_{i=1}^{3} m\frac{dv_i}{dt}v_i = \sum_{i=1}^{3} m\frac{d}{dt}\left(\frac{v_i}{2}\right) = \frac{d}{dt}\left(\frac{mv^2}{2}\right)$$

where v is the module of the velocity vector $\mathbf{v}$. For closed loop system $\mathbf{F}=0$, and the equation integral will be

$$\frac{mv^2}{2} = Const$$

one of the forms of the energy conservation law. From the definition of the angular momentum for a particle,

$$\mathbf{L} = [\mathbf{r} \times \mathbf{P}]$$

Differentiating both parts by *t*, we get

$$\frac{d\mathbf{L}}{dt} = \left[\frac{d\mathbf{r}}{dt} \times \mathbf{P}\right] + \left[\mathbf{r} \times \frac{d\mathbf{P}}{dt}\right]$$

Since the impulse vector is parallel to the velocity vector, the first bracket will equal to zero. On the basis of the resultant equation and the definition of central force as not creating any momentum, we get

$$[\mathbf{r} \times \mathbf{F}] = 0 \quad \text{or} \quad \mathbf{L} = Const.$$

In case of the central force in an unclosed system, the angular momentum is preserved by value and direction. The angular momentum conservation law for a closed loop system results in the same way as the impulse conservation law from the equation of the rotary motion dynamics:

$$\mathbf{M} = \frac{d\mathbf{L}}{dt}$$

For a closed loop system, the momentum of external forces $\mathbf{M}=0$ and the integral of the equation will be the angular momentum conservation law

$$\mathbf{L} = Const$$

In relativistic dynamics the emergence of the energy and impulse conservation laws separately can be easily received from the relativistic ratio for energy and impulse

$$E^2 = P^2c^2 + m^2c^4$$

Term m^2c^4 is an invariant, i.e. the same in all reference systems. In other words, it is a certain constant. This equation can be represented in a slightly different form

$$E^2 - P^2c^2 = Const$$

For the equation to be valid, it is required that

$$E = Const \text{ and } P = Const$$

and this none the other but conservation laws for energy and impulse.

Strictly, relativistic mechanics has a conservation law for 4-impulse vector p^{μ}, but we will not dwell on these details, because small energies are what we are interested in.

In the classical theory, the energy conservation law states that the energy of a closed loop system remains unchanged, so, if the energy of such a system is designated at a moment *t=0* as E_0, and at the moment *t* as E_t, $E_0 = E_t$.

Conservation laws in ordinary quantum mechanics

The standard quantum theory formulates the energy conservation law in the same way. In quantum mechanics we have the same movement integrals as in classic mechanics. A certain value *L* will be a movement integral, if

$$\frac{d\hat{L}}{dt} = \frac{\partial \hat{L}}{\partial t} + \left[\hat{H}, \hat{L}\right] = 0 \qquad (1)$$

Since $\left[\hat{H}, \hat{L}\right]$ is defined by commutator of operator $\hat{L}$ and of Hamilton's operator, any value *L*, *not depending explicitly on time*, will be the movement integral, if its operator commutes with the Hamilton's operator. When value *L* does not explicitly depend on time, the first item in (1) turns to zero. There remains

$$\frac{d\hat{L}}{dt} = \left[\hat{H}, \hat{L}\right] = 0 \qquad (2)$$

and for the movement integrals not explicitly depending on time the Poisson quantum bracket equals zero. From (1) and (2) it follows that the average value of the movement integrals does not depend on time:

$$\frac{d}{dt}(L) = 0$$

All good papers on the quantum theory prove that probability $w(L_n, t)$ to find at any moment *t* any value of the movement integral, i.e. L_n, does not depend on time. Further, is constructed as the movement integrals not explicitly dependent on time. Since operators $\hat{\mathrm{L}}$ and $\hat{\mathrm{H}}$ commute, they have common proper functions, which are functions of stationary states. Let us note that the latter follows from solutions of the equation without time, which was received from the full equation with imposition of requirement

$$\Psi(\mathbf{r}, t) = \Psi_0(\mathbf{r}) \exp\left(i\frac{E}{t}\right)$$

equivalent to search of only periodic solutions. Further, quite naturally, *there appeared an equation without time with actually imposed conservation laws, because now nothing depends on time*. Expansion by such proper functions looks as follows:

$$\hat{L}\Psi_n = L_n\Psi_n \qquad\qquad \hat{H}\Psi_n = E_n\Psi_n$$

where

$$\Psi(x,t) = \sum_n c_n \Psi_n(x) \exp\left(-i\frac{E_n}{\hbar}t\right) = \sum_n c_n(t)\Psi_n(x) \qquad (3)$$

$$c_n(t) = c_n \exp\left(-i\frac{E_n}{\hbar}t\right) = c_n(0)\exp\left(-i\frac{E_n}{\hbar}t\right)$$

Since (3) is expansion into proper functions of operator $L_{n'}$ probability does not depend on time:

$$w(L_n, t) = |c_n(t)|^2 = |c_n(0)|^2 = Const$$

Since energy is a movement integral and probability *w (E,t)* to find at a moment *t* an energy value equaling E, does not depend on time, then:

$$\frac{dw(E,t)}{dt} = 0$$

Let us note once again that it is *the probability to find a certain value that does not depend on time, but not the value itself, which for any separate event is accidental and can assume a wide range of values.*

The quantum energy conservation law in the above form suggests a possibility of defining energy at a given moment without subjecting it to uncontrolled change, which raised no doubts in classic mechanics. But in the quantum theory the energy, without changing its value, can only be measured to

$$\Delta E \geq \frac{\hbar}{\tau}$$

where r - is measurement duration. Formally, it does not present any difficulties for the energy conservation law, since energy is a movement integral, and we have much time to make long measurement. For example, let us make measurements during time r, then leave the system to itself for time *T*, and then define energy again. The classic quantum energy conservation law states that the result of the second measurement will coincide with the result of the first measurement to $\Delta E \approx \frac{\hbar}{\tau}$.

But even in the ordinary quantum theory all this is not consistent enough. For the real vacuum fluctuations can interfere, that always influences the results of a single process, but their influence disappears after the passage to an ensemble of events. Here we have a violation of the conservation law due to vacuum fluctuations, though existence of movement integrals, unlike in the Unitary Quantum Theory (UQT).

The generally accepted quantum theory carefully avoids the question of conservation laws for individual events in the case of small energies. This question is either not discussed at all, or it is said that *the quantum theory does not describe individual events. Yes, it does describe individual events, but it can only predict a probability of this or that result.* It is clear that in this case there are no conservation laws for individual events (it is wrong to speak about it in case of an accidental result of an individual event), and they appear only after the averaging by large ensembles of events. Essentially, it can easily be proved that classic mechanics follows from quantum mechanics after summing up by a large number of particles, because for a sufficiently big mass, the length of the de Broglie wave becomes much less than the body dimensions, and no quantum-wave qualities can be talked about.

Conservation laws in Unitary Quantum Theory

In the UQT [1-14] any quantum particle is not a point, but a source of field like in the ordinary quantum mechanics, but it represents a bunched field (wave packet) of a certain unified field. The dispersion equation of such a nonlinear field turned out to be such that the wave packet (particle) during its movement periodically appears and disappears, and the envelope of this process coincides with the de Broglie wave. Numerous particles during their periodic disappearance (spreading in the Universe) are repeated appearance from vacuum fluctuations. A theory of quantum measurements has been built, and the probability interpretation follows from the mathematical formalism of the quantum theory [10,11], and it is not postulated as in conventional quantum mechanics. Unfortunately, the main UQT equation turned out very complicated, for it is a system of 32 nonlinear integral-differential equations, which could require for their solution some new mathematical methods. But from this the relativistically invariant Hamilton-Jacoby equation, and the Dirac equation system strictly follow.

Papers [13,14] give a solution of the simplified scalar integral-differential UQT equation, which gave a localized solution for the form of a wave packet representing a particle. It turned out that the integral from a bilinear combination of such a solution for the whole volume gives with the precision of 0.3% the value of a non-dimensional elementary electric charge [13,14], which was essentially its first theoretical calculation. Then, this solution in the form of a periodically appearing and disappearing wave packet (which square describes the density of a spatial charge) can be replaced by an oscillating charged particle [15-18], the movement whereof will be described by the conventional Newton equations:

$$m\frac{d^2\mathbf{r}}{dt^2} = -2Q\mathbf{GRAD}U(\mathbf{r})\cos^2\left(\frac{mt}{2\hbar}\left(\frac{d\mathbf{r}}{dt}\right)^2 - \frac{m\mathbf{r}}{\hbar}\frac{d\mathbf{r}}{dt} + \varphi_0\right) \quad (4)$$

$$m\frac{d^2\mathbf{r}}{dt^2} = -2Q\mathbf{GRAD}U(\mathbf{r})\cos^2\left(-\frac{m\mathbf{r}}{\hbar}\frac{d\mathbf{r}}{dt} + \varphi_0\right) \quad (5)$$

where *m, Q,* $\mathbf{r}$ - mass, charge, and radius-vector of the particle, *U*(**r**) – external potential, φ_0 - initial phase.

Since $\mathrm{E} = -\mathrm{GRAD}U$, and a magnetic field also exists, the Lorenz force should also be calculated for $\mathbf{F} = \frac{Q}{c}[\mathbf{v}\times\mathbf{H}]$, but in the electromagnetic wave **E** and **H** are equal, and for small energies value $\frac{v}{c} \to 0$, and force **F** can be ignored. Both these equations produce qualitatively similar results for different problems, but the first non-autonomous equation evidently does not have any movement integrals at all, and any hope for analytical solutions is very unreal. But for the second autonomous equation such hope still exists. Let us note that these equations describe more accurately the experimental results of scattering on the coulomb potential than the classic Rutherford formula! Application of these equations for the tunnel effect and scattering on short potential also produces correct results, but in this case passage through a high barrier (tunnel effect) will be defined by the initial phase. Of greatest interest, however, is the harmonic oscillator problem.

It is possible that a change in the properties of a material point in the process of its movement is just another step in the material point movement theory. In conventional mechanics this idea is not altogether new. There are Meshchersky's equations for bodies with a changing mass, and Tsiolkovsky's equation for a rocket. But so far, in the conventional quantum theory, the particle has a permanent and stable in space and time set of properties, and in the UQT all the parameters of the particle are changed and oscillate during movement.

It should be noted that *Newton did not introduce the notion of a material point at all,* and it would be ridiculous to think that he was not able to have this natural and rather trivial idea. Most probably, and it is not by chance, for today many troubles of the field quantum theory are rooted in the approach to the particle, as to the point, the most vivid example being a large bouquet of divergences. Nevertheless, this approach is very convenient and should only be used correctly. Let us also remember, that in accordance with the Newton corpuscular theory, beams of light were to be regarded as a flow of certain particles. They are emitted by a shining body in all directions and move in an empty space or a homogeneous medium evenly and straight, i.e. in the same way as the ordinary material particles do in the absence of any external or interaction forces. Newton explained reflection and refraction of light beams on the surface of border between two homogeneous mediums by the effect of certain forces on this border, in the direction perpendicular to the surface. These forces changed the normal velocity component, but did not touch upon the tangential one, which allowed to derive the reflection and refraction laws. However, the inability of such a theory to account for the light partial reflection and passage phenomena, as well as the Newton rings (which he himself discovered), led him to bouts (or fits) theory, which is quite modern, although nearly forgotten. Newton believed that for full explanation of all the processes it was necessary to suggest that some light particles could experience reflection bouts, and others – passage bouts. Let us imagine light falling to a flat surface, which is partially, reflects and partially passes. With quantum description of this phenomenon, a particle connected to the falling wave at the time of hitting the surface has a certain probability of passing or being reflected, and Newton simply used the word "bouts" in place of the word "probability".

It is absolutely clear that all descriptions of processes by the equation with an oscillating charge will be an approximation, because it is evident that no movement equations for a material point can describe even the simplest interference processes on a semi-transparent mirror, during which a material particle should be divided in two parts which will later eliminate each other by destructive addition. It is surprising, but the numerical solution of the problem of scattering on a short potential (the Ramsauer effect) for equations (4) and (5) gives the correct diffraction picture! But if we want to describe an individual particle correctly in the conventional quantum mechanics, the picture becomes inexact and purely probabilistic. At every given moment of time a particle can exist in only one of the mutually incoherent states, because one particle cannot move in different directions simultaneously (it cannot have many impulses at the same time). Nevertheless, there seems to exist a whole class of processes, where description with the help of equations (4) and (5) have certain sense. It is well known that in all experiments the local energy and impulse conservation law in individual quantum processes are true only under high-energy values. But under small energy values it is not so, at least because of the ratio of uncertainties and the probabilistic character of all the quantum theory predictions, and the idea of a global, not local ECL, is invisibly present in the quantum mechanics, and is certainly far from new.

In the strict UQT and the quantum measurement theory, a great role belongs to unavoidable vacuum fluctuations. It is clear that these fluctuations are totally unpredictable and non-invariant in relation to space and time translations. The same can be said otherwise: there are no habitual properties of time and space in this theory. Space-time is now not homogeneous and not isotropic. For example, if the system is transferred to a new point in space, or a certain experiment is repeated at another time, at the point where particle parameters are studied, and it interacts with the macro-device, a new value of vacuum fluctuations (different from the previous one) can appear and produce a different result. Of course, all this is only true for small energies and individual events.

Still more destructive is the UQT for the notion of a closed system. For individual events under small energy values this notion is simply unacceptable for the following reason: vacuum fluctuation at the location of the particle (e.g. in a potential pit) can be sharply changed at any moment. It can be caused by different factors – the nature of vacuum fluctuations itself, or the tunnel effect of another random particle.

Sometimes it is stated that conservation laws follow from the Nether theorem, though these results are present in the works by D. Gilbert and F. Klein. For any physical system, the movement equations from which can be received from the variation principle, each one-parametric continuous transformation that leaves the variation functional invariant, corresponds to one differential conservation law, and there exists a clearly conserved value. It is easy to see, however, that vacuum fluctuations imposed on the varied function (integral of Lagrangian) do not in sum remain unchanged during parametric transformations (at least today it seems so), and this consideration does not work without preliminary of ensemble.

And now we are in for a little philosophy. The local Energy Conservation Law (ECL) in individual processes follows from the Newton equations for closed systems. It would be naive to think that its local formulation will be preserved forever, and would be a bad mistake to transfer the ECL from the Newton mechanics to the quantum processes without any changes, because the latter are more fundamental.

References to the first principle of thermodynamics are, strictly speaking, groundless, because this principle is a postulate. For example, well-known Russian mathematician N. Luzin, in a letter to an inventor wrote that the first principle of thermodynamics is the result of unsuccessful attempts of the mankind at building a *perpetual mobile*, and strictly follows from nothing. *Today it may be said with a great degree of certainty that no sophisticated machine in the framework of the Newton mechanics can be a perpetual mobile,* and the decree of the French Academy of 1755 not to consider any *perpetual mobile* projects is still valid. We will only add that now it is valid only for those projects that are based exclusively on the Newton mechanics.

There is the tendency in modern physics to reduce ECL, especially in theory, to the rank of a secondary derivation from the movement equations (movement integrals). Some physicists restrict the ECL to the framework of the first principle of thermodynamics, others, like D. Blokhintsev [37], think it quite probable that with the development of a new theory the form of the ECL will undergo certain changes. F. Engels wrote in his "Dialectics of Nature": "…none of the physicists actually regard the ECL as an eternal and absolute law of nature, a law of spontaneous transformation of the movement forms of the matter and the quantitative constancy of this movement in all its transformations". But many people do not share this opinion. M. Bronstein in his book "Structure of Matter" wrote: "The ECL is one of the main laws of the Newton mechanics. Nevertheless, Newton did not ascribe to this law the general character that this law actually possesses. The reason for this *erroneous* (italicized by authors) opinion of Newton of the ECL is very interesting…". It is now clear that in view of the above, such an opinion was not at all erroneous. Let us remind you that Newton predicted many things, even the UQT, in his "bouts theory".

On the other hand, the authors of quantum mechanics realized that there was no conservation law for single quantum processes under small energies at all. The idea that the construction of ECL, together with the second law of thermodynamics, was a statistical law, true only on the average and inapplicable to individual processes with small energy, first occurred to Schroedinger, and later to Bohr, Kramers, Slater, and Gamov. In 1923 Bohr, Kramers and Slater made a desperate attempt to develop the theory, where the energy and impulse conservation laws in case of scattering would be true only statistically, on an average for long periods of time, but would be inapplicable to elementary events. Lev Landau even called it "Bohr's wonderful idea".

Later, however, the authors gave up this approach and, besides, this idea at that time did not follow from the quantum theory equations, and the authors, *to come out of the predicament, simply declared that quantum mechanics did not describe individual events at all*. Thus, the most vivid paradox of the quantum science was removed by a simple ban on thinking about it! But the ingenious idea that conservation laws do not apply to individual quantum processes and emerge only after the averaging by the ensemble of particles remains alive. This idea might have been a little premature, and, possibly, should be a little different.

The Unitary Quantum Theory (UQT), on the contrary, individual particles, and the difference in their behavior is accounted for by the initial phase of the wave function. In this case, local conservation laws do not exist for a single particle, and measuring the initial phase or some other parameters for an individual particle is quite a different matter. It is not true that the UQT has given up probabilistic description. Probabilistic interpretation remains, but the probability now is strongly dependent on the initial phase. Although the equations with an oscillating charge can determinately predict a particle's behavior, the measurements can be made only with the help of a macro device, which will give only a probabilistic result. Impossibility to determinate measurements does not change anything, for the UQT provides for a possibility of influencing the probability value, which was earlier unavailable. The existing Von Neumann theorem about hidden parameters does not effect our result, but the relevant discussion is too cumbersome, and we will leave it out.

*In other words, **all the requirements**, wherefrom the classical conservation laws follow, are now absent. We can hardly expect the conservation laws for individual particles to be preserved under small energies in such a situation.* Today we are convinced that the classical energy, impulse and angular momentum conservation laws for individual quantum objects are not valid under small energy values because of periodic appearances and disappearances of the particle. All direct experimental tests of the conservation laws were made for large energy values, and for small energies of an individual particle only probabilistic results can be received, and, in this case, it would be indecent even to recall the conservation law.

Energy generation and perpetual mobile

Let us make the following imaginary experiment. For simplification purposes we will use in our reasoning a certain quantum ball-particle. When a classical ball approaches a wall (perpendicularly for simplification), the speed of the reflected ball is always equal to the initial speed (we ignore friction and regard the ball and the walls as absolutely elastic). In the case of a quantum ball, the speed of the reflected ball will acquire in different experiments with absolutely equal initial conditions a whole range of values: some balls will be reflected at a speed greater than the initial speed, others – at a speed equal or lower than the initial speed, and all this is described by quantum mechanics.

Let us ask the following question: what if a second wall is found, parallel to the first one, in order for the ball to increase its speed after each reflection from the wall? Then we will have increased ball energy without any special efforts on our part. Such phenomena appear in the problem of particle oscillations in a potential pit (not necessarily parabolic) on the basis of equations (8) and (9), when four types of solutions are possible, three of which are most important for us: stationary, "maternity home", and "crematorium". In the two latter solutions traditional conservation laws do not work. These solutions are presented in Fig. 1. Such oscillator behavior explains many experimental facts. From the physical point of view, it means that in stationary solutions with fixed discrete energies (conventional quantum mechanics) the speed of the particle reflected from the wall will be equal to the speed of the falling particle. If the speed of the particle is decreased after each reflection, it will mean the "crematorium" solution, and if it increases, the "maternity home". Scenarios for situations will depend on the initial phase of the wave function and the particle energy. In ordinary situations the "crematorium" and "maternity home" solutions always compensate each other, and we find conservation laws.

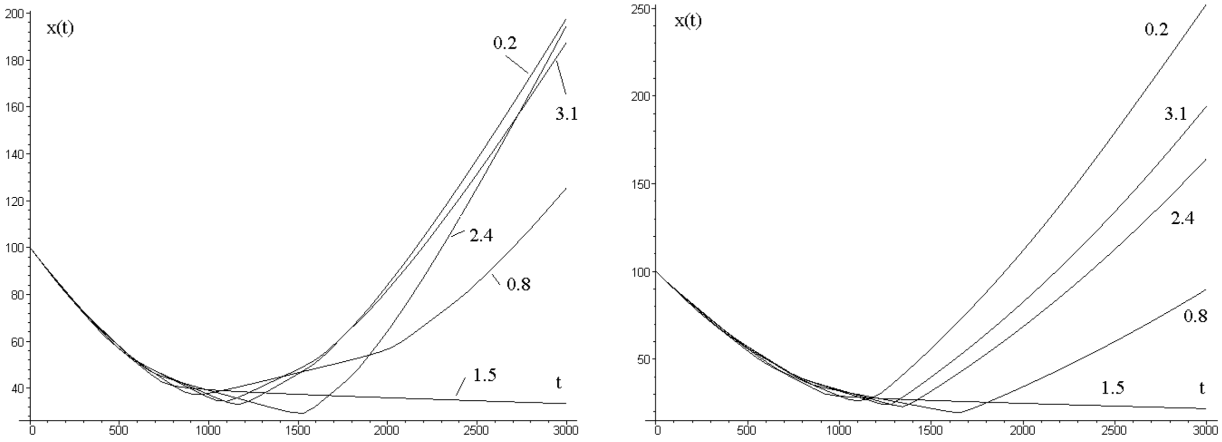

Fig. 1. Dependence of the distance between the moving charge and the nucleus on time for autonomous and non-autonomous equations.

The task of the future developers of new energy systems of the 21st century will consist in creating such initial conditions for a great number of particles making up a body that only the "maternity home" solution would be realized, and the "crematorium" solution would, if possible, be suppressed.

It follows from the above that if the unitary quantum theory ideas are applied correctly, there is no fundamental taboo for a perpetual mobile. Such a taboo, as it was shown, does not formally exist even in conventional quantum mechanics (no conservation laws for individual processes with small energies), and, in order to generate energy, they should be somehow accumulated (all random processes with excess energy should be grouped together). But conventional quantum mechanics refuses to describe individual events and is unable to offer any ways for such grouping. The unitary quantum theory seems to offer such an opportunity.

However, the great idea of free energy generation was distorted by effort of some research associations interested to keep their stability degree, that everyone who started speaking about it was considered to be a crazy man.

Modern experimental physics has verified the correctness of conservation laws either for very large energies in individual quantum events, or for big macroobjects, when automatic averaging by ensemble is made, but the area of very small energies for individual events today is a *terra incognita*.

In order to see how the conservation laws for reflection (repulsion) of an individual particle from the Coulomb heavy nucleus with different values of the initial phase are violated, we have solved numerically one-dimensional equations (8) and (9) under the different initial conditions:

$$\hbar = 1, m = 1, \ 2Zze^2 = 1, \ x_0 = 100, \ V_{x0} = -0.1$$

In the Fig. 2 the distances between the moving particle and repulsive nucleus are shown as a time function, for different initial phases in cases of non-autonomous and autonomous equations.

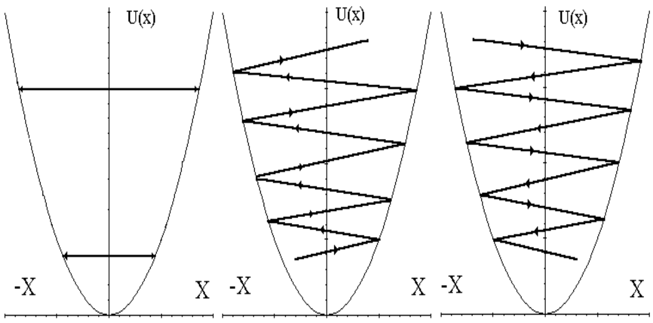

Conventional solution Mathernity home solutions Crematorium solutions

Fig.2 Three types of solutions for oscillator

It is evident from the calculations that the speed of a reflected particle can be equal, lower or higher than the speed of a falling particle. This situation seems to be true for all potentials.

Calculations were also made for other potentials: harmonic oscillator, Yukawa, Gauss, dipole, hyperbolic secant, and Wood-Saxon, and the quality results were nearly the same. If we sum up the impulse of all the particles falling with different phases and compare it with the summarized impulse of all the reflected particles, the summarized reflected impulse, for example, for the Coulomb potential, will be several percent higher than the summarized impulse of the falling particles. For other potentials such a small deviation can even be in the opposite direction. On the whole, this problem is very complicated and requires additional research, because all this is also dependent in quite a complex way on the initial conditions (initial speed, phase and distance).

Philosophically, any categorical taboos, like the impossibility of creating a *perpetual mobile*, are absolutely unacceptable. If everyone is convinced of it forever, the conservation laws and *perpetual mobile* taboos will remain unshakable as long as the human civilization exists. Of course, the funeral of the Conservation Laws can be very prolonged. Anyway, we are not going to do it, and our article might be just a cleanup for the future tomb, and the splendid funeral with all the necessary honors will be organized by future generations. *On the other hand, these laws will never die out completely and will surely be applied, but such spheres of science and technology will appear, though small at first, where these laws are not valid.*

The truth should be accepted irrespective of where it comes from. Words of F. Engels from the “Dialectics of Nature” will be quite appropriate here: “When the solar system ends its life circle and shares the fate of all the finite things, when it falls victim to death, what will happen next? Thus, we come to a conclusion that the heat emitted into the universe should have an opportunity, in a way yet to be established by the natural sciences, to turn into another form of movement, where it can be accumulated again and start functioning. And in this casethe main problem that prevented transformation of dead suns , back to red-hot nebulas will disappear”.

The question of whether the conservation law exists in global form (we have already proved its not being local) remains open, because nothing leads to it except the inertia of the human mind. This inertia was based on the Newton laws, which were replaced by quantum laws. This mental inertia leads to a situation, when in case excessive energy is generated during solution of movement equations, a question arises how it can happen, and where it comes from. Of course, if a particle (e.g. a photon) falls on a semitransparent mirror, the packet is divided into two halves, which, due to imposition of vacuum fluctuations, will be recorded by photomultipliers as full-fledged photons [1-5]. The result is that energy is taken as if from vacuum: two photons appear in place of one. Another photon can be divided on the mirror into two halves, but they will not be recorded by the meters, and the energy will allegedly pass into vacuum. So, at one time we borrowed energy from vacuum, and then gave the same amount of energy back to vacuum at another place. You can think like that, and this process might take place. But if we consider the equation with an oscillating charge, the energy and impulse conservation laws are not valid there for solution of the movement problem, and vacuum fluctuations have nothing to do with it at all. As for the question of where the energy comes from, it is the result of our mental inertia, and is, essentially, an atavism imposed by the Newton mechanics. But the latter appears as a result of an extreme passage from quantum mechanics, which is more fundamental.

It is interesting to note that there is a bomb in the logical definition itself of the energy conservation law. If energy is something that cannot appear or disappear and always simply passes from one form to another, the only value satisfying this condition is zero. We are far from assuming that energy does not exist. But the problem of existence is solved differently in different philosophical systems, and the mathematical approach seems to be the most correct one: *an object exists if it is free from contradictions*. Energy has bad luck in this case, for under such an approach it should be zero.

Some cosmologists (for example, British prof. Fred Hoyle) are very willing to have a process, in accordance with which the Universe has certain places where energy appears from other certain places, in which it is eliminated. Besides, any philosopher at least a little bit familiar with astronomy, looking at the bright night sky, will see the birth of matter and its expansion into a still greater space. But for this purpose the Global Energy Conservation Law is superfluous and only denies what is observed. The head reels…

Cold Nuclear Fusion and Nuclear Transmutation.

Let us approach the epoch-making experiments made by Fleischmann and Pons in March 1989 [30] from the positions of the equation with an oscillating charge. One of the authors predicted in 1983 [9] the possibility of such nuclear reactions under very small energies. Without going into well-known details, we will sum it up very briefly: cold nuclear fusion exists, and there are no people or theories capable of giving a clear explanation. The chain of various mechanisms meant to explain this intriguing phenomenon is growing, but few really believe in them. The reason is as follows.

When a charged particle interacts with the nucleus, the potential energy is like in Fig. 3, where the right top part of the curve is conditioned by mutual Coulomb repulsion between the nucleus and the charged particle.

The repulsion potential will be

$$U(r) = \frac{Zze^2}{r}$$

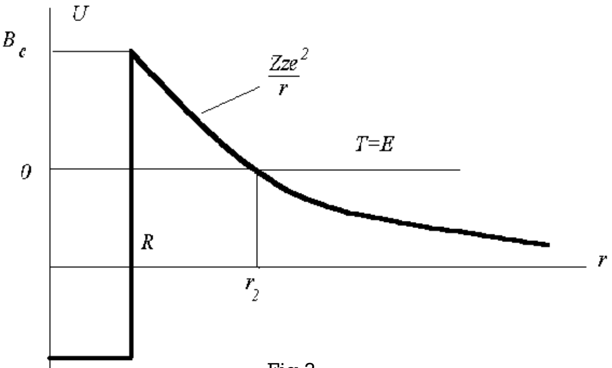

Fig.3

where Z – charge of the nucleus, and z – charge of the approaching particle, e – charge of the electron, r – distance between the particle and nucleus. When $r=R$ (critical distance), then the potential energy curve goes sharply down, which is due to the emergence of intense nuclear gravitation, the potential whereof today appears more complex than could be imagined mathematically. If the charged particle overcomes the Coulomb barrier with a height of

$$B_c = \frac{Zze^2}{R} \approx \frac{Zz}{\sqrt[3]{A}} MeV$$

it will further get into the nuclear gravitation area and a nuclear reaction will take place.

Let us look at the nuclear interaction of a charged particle with kinetic energy $T < B_c$. From the point of view of classical mechanics, there will be no nuclear reaction in this case, because the particle will approach the nucleus and at a certain distance $r < R$ from the top of the Coulomb barrier will turn back and be reflected from it. However, from the point of view of quantum mechanics, there exists a tunnel effect, and the probability of such a tunnel passage, or transparency of potential barrier D is described by a well-known formula:

$$D \approx \exp\left(-\frac{2}{\hbar} \int_{r_1}^{r_2} \sqrt{2\mu(U-T)}dr \right) \qquad (6)$$

where $\mu = \frac{Mm}{M+m}$ is reduced mass. The bottom integration limit r_1 coincides with the nucleus radius R, and the top limit r_2 can be found from the condition $T = \frac{Zze^2}{r_2}$. After integration we will get

$$D = \exp(-2g\gamma)$$

where $g = \frac{R}{\lambda_{B_c}}$; $\gamma = \sqrt{\frac{B_c}{T}} \arccos\left(\sqrt{\frac{T}{B_c}}\right) - \sqrt{1-\frac{T}{B_c}}$,

and value $\lambda_{B_c} = \frac{h}{\sqrt{2mB_c}}$, the de Broglie wavelength corresponding to the kinetic energy of the particle equal to the barrier height $T = B_c$. If $T << B_c$, expression (6) is easily transformed to look as

$$D = \exp\left(-\frac{2\pi RB_c}{\hbar v} \right) = \exp\left(-\frac{2\pi Zze^2}{\hbar v} \right) \qquad (7)$$

where v is velocity.

Let us now see what the shocking cold nuclear fusion will look like on the basis of the above considerations. The deuteron energy in an ordinary electrolytic Fleischmann-Pons cell will be about 0.025 *eV*, and the height of the Coulomb barrier for this case is $B_c = \frac{Zze^2}{\sqrt[3]{A}} = 0.8 MeV$. In classical mechanics it would be just naive to talk about overcoming such a barrier with a height dozens times greater than the kinetic energy. Let us now see how the tunnel effect will improve the situation. Let us assess the value of g and γ for the case of collision between two deuterons with such energy:

$$g = \frac{R\sqrt{2mB_c}}{\hbar} = 1.9;$$

$\gamma = \sqrt{\frac{B_c}{T}} \arccos\left(\sqrt{\frac{T}{B_c}}\right) - \sqrt{1-\frac{T}{B_c}} \approx 8883$ and the probability of such a process will be $\exp(-2 \cdot 1.9 \cdot 8883) \approx 10^{-7328}$, i.e. practically pure zero. The fusion cross-section will be defined by the product of nuclear cross-section and the tunneling probability:

$$\sigma = \sigma_{nucl} D$$

and, in the case under review, is also a very small value. If the clash parameter of deuterons is not zero, the emergence of centrifugal potential

$$U = \frac{\hbar^2 l(l+1)}{2mr^2}$$

will still further lower the probability of such interaction.

It is these very circumstances that make the nuclear physics scientist think that there is no cold nuclear fusion as such. For example, such a serious and responsible edition as Encyclopedia Britannica 2001 found no place for the notion of cold nuclear fusion at all. Such an official position can be understandable only from the point of view that quantum mechanics is absolutely true and unshakable. Despite this, for the 12 years since the Fleischmann-Pons experimental discovery, nearly 30 international conferences have been devoted to this subject, there are lots of books and magazines on this subject, and the number of articles on the problem is nearing ten thousand. Today the situation is gradually developing in the positive direction, and the research in the field of hot nuclear fusion, which has already wasted over $90 billion for 45 years, is slowly coming to naught.

But today there exist well known experimental data on cold nuclear fusion. They are numerous and various. We will dwell only upon the most important and sufficiently reliable results. Thus, the classical view of electrolysis of a palladium cathode saturated with heavy hydrogen in heavy water identifies an anomalous quantity of heat energy up to 3 kWt/cm^3, or up to 200 Mj per small sample. Products of nuclear reaction have also been found: tritium (10^7 - 10^9t/s), neutrons with energy of 2.5 MeV (10-100n/s), and helium. Absence of He^3 among the reaction products shows that heat is not generated by reaction $d+p$. Besides, emission of charged particles ***(p, d, t, γ)*** is observed. Similar processes are observed in case of a gas discharge on a palladium cathode, of phase passage in different crystals saturated with heavy hydrogen, irradiation of deuterium mixture with a powerful sound or ultrasound flow, in cavitating microbubbles in heavy water, in a tube with palladium powder saturated with heavy hydrogen under a pressure of 10-15 atm., etc. In certain reactions (e.g. $d + t \rightarrow \alpha + p$) neutrons of 14 MeV are absent, and such a strange situation occurs in other cases too. Activity of Li^6 , Li^7 in reactions with heavy hydrogen and protons failed to be discovered, whereas reaction

$$K^{39} + p \rightarrow Ca^{40}$$

was well recorded even in biological objects. But the most intriguing fact of all these processes is the shortage of nuclear reaction products for explanation of the emerging heat effects. Thus, in certain cases the number of nuclear reaction products (tritium, helium, neutrons, quanta) should be millions of times greater in order to account in some way for the quantity of the generated heat. Generation of such a big amount of energy cannot be accounted for by either chemical or nuclear reactions, or by phase passages. The well-known interaction $d+d$ goes along three channels:

D + D —> T (1.01) + p (3.03) (Channel 1)
D + D —> He (0.82)+ n(2.45) (Channel 2)
D + D —> He + γ(5.5). (Channel 3)

All these reactions are exothermal. The third channel has a very small probability. It was experimentally discovered that they could occur under very small energies. In a molecule D_2 the equilibrium position between atoms is 0.74A and in accordance with the conventional quantum theory, these two deuterons could accidentally enter a nuclear fusion reaction. But the interaction value is very small $\lambda_{D_2} = 10^{-64} c^{-1}$. There is a known estimate that in the water of all seas and oceans there are 10^{43} deuterons, and in 10^{14} years there will be only one fusion.

It follows from the aforesaid that the main problem impeding the occurrence of the $d+d$ reaction lies in the existence of a very high Coulomb barrier. Our approach allows for this problem to be solved, and there is such an opportunity in the UQT. The UQT equation solutions show that the distance to which deuterons can approach each other is strongly dependent on the phase of the wave function (by the way, it is absolutely clear intuitively).

Let us consider the one-dimensional problem [15-18,31]. There is a stationary nucleus with charge *Ze* at the point of origin, and another nucleus is approaching it along axis x (charge *ze*, mass *m*) at a certain initial velocity. The non-autonomous and autonomous equations of such a problem will look as follows:

$$m\frac{d^2x}{dt^2} = -\frac{2Zze^2}{x^2}\cos^2\left(\frac{m}{2\hbar}\left(\frac{dx}{dt}\right)^2 t - \frac{m}{\hbar}\frac{dx}{dt}x + \varphi_0\right) \quad (8)$$

$$m\frac{d^2x}{dt^2} = -\frac{2Zze^2}{x^2}\cos^2\left(-\frac{m}{\hbar}\frac{dx}{dt}x + \varphi_0\right) \quad (9)$$

Since an analytical solution was not found for all the areas of initial phases , numerical methods were applied with the following initial values: $Z=z=1$, $e=1$, $m=1$, $x_0=-10$, $\hbar=1$ for different initial velocities and initial phase values. As had been expected, braking or acceleration of the particle happens only when the charge is large. But at the last stage, under certain initial phases close to $\frac{\pi}{2}$, a wonderful process occurs: velocity, charge and repulsing force are very small. Due to phase ratios, the small charge is not changed for a long time, which means that the particle (or rather what is left of it) is not influenced by any forces, and it is crawling at a permanent small speed for a very long time ("the snail effect") inside the field of another particle, and can come very close to the center. Such movement with a very small charge and a small speed can last for several hours, and disconnection of the external field will not effect this movement. This process reminds of quiet and invisible scout penetration into the enemy territory. This phenomenon occurs only in certain phase areas, and can be conveniently called a phase hole, which is illustrated in Fig. 4 resulting from the solution of equation (8).

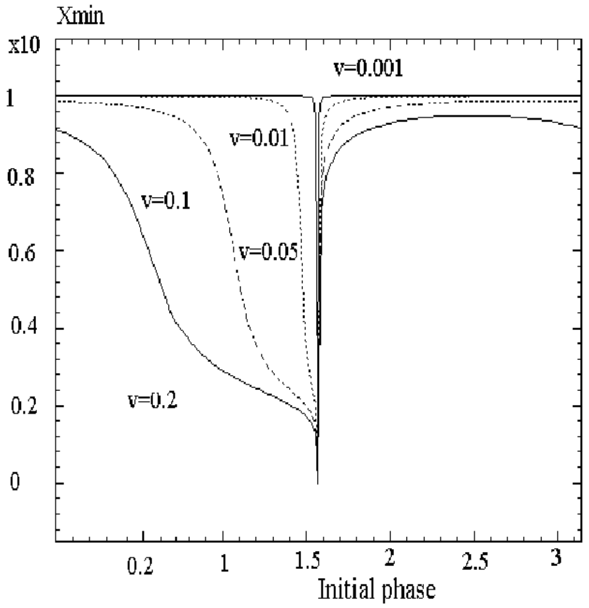

Fig.4. Distance to the turning point of the moving charge depending on the initial phase value for different initial velocities.

Let us note in passing that now we can account for one of the nuclear physics anomalies, which has a tendency to be totally ignored. Under a nucleon energy of 1 MeV, its velocity is 10^9 cm/s, nuclear radius is 10^{-12} cm, and the passage time of the nucleus is 10^{-21}s, but the time period in which the nucleon passes is usually anomalously long - 10^{-14} and even more, and it is absolutely unclear what the nucleon is doing in the nucleus so long. In our model it is easily explained by the "snail effect".

For the same equation, the minimum distance was calculated between the charges dependent on velocity (Fig. 5) for different initial phase values. For comparison, Fig. 5 also shows the result of the classical calculation based on the Coulomb law. It is obvious from Fig. 4 and 5 that the minimum distance to which charges can approach each other is nearly independent of kinetic energy, but with reduction of speed the initial phase area width is reduced as well. In other words, reduced energy brings also reduced probability of a nuclear reaction.

The same results are true for the autonomous equation (9). Under the conventional quantum theory, the ratio of the reaction speeds in the tritium and neutron channels should be close to unity: $\frac{t}{n}=1$. But in many experiments on cold nuclear fusion this value is very different from unity and equals $\frac{t}{n}=10^9$. In different experimental situations it is reproduced by different experimental groups with a very high accuracy. This very intriguing problem has so far received no simple explanation. Let us dwell on a possible cause for such aphenomenon. With a small velocity in the phase hole, neutrons are affected by nuclear gravitation forces, and protons are affected by electrostatic repulsion forces. Under the effect of this momentum, the deuteron will have enough time to turn in such a way that the neutron parts of the deuteron would faceeach other. After the neutron gravitation the nuclear forces will be saturated, which will weaken the proton connection, and one of the protons will leave the system. This reaction can be conditionally presented in the following form:

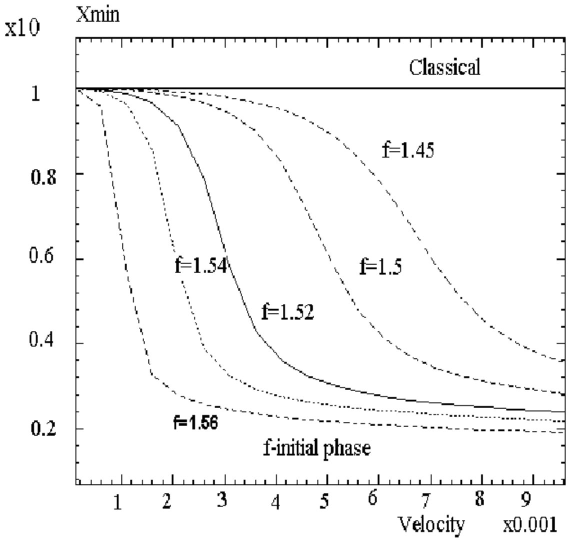

Fig.5 Minimum distance between charges depending on the initial velocity for different initial phase values.

$$d+d \to p+(n+d) \to p+t$$

This reminds of the Oppenheimer-Phillips effect.

It is well known, however, that under big energies, the probabilities of the first and second reaction channels are the same, and this phenomenon should somehow be accounted for. Increased probability of the neutron channel with growing energy can be connected with the appearance of secondary neutrons in the reaction *T + D = He + n (14.1 MeV).* In a deuterium-rich environment, a big part of the resultant tritons will pass to neutrons in the process of this reaction, which has a cross-section of 5 barns under an energy value of 70 KeV. According to assessments in [32], the number of such secondary neutrons per one triton is $7.9\cdot10^{-12}, 1.7\cdot10^{-9}, 2.7\cdot10^{-6}$ for energy tritons 10, 20 and 100 KeV respectively. Thus, the prevalence of $\frac{t}{n}>10^6$ can be expected only in those reactions, where tritium is born with energies over 40 KeV.

It should not be assumed, however, that the phase hole phenomenon in its whole area leads to a nuclear reaction. It can be assumed that reduction of the Coulomb repulsion is followed by reduction of strong interaction. But how? Today nobody knows the exact equation of the strong interaction potential. Besides, the particle approaches the turning point $X_{\min}$ is rather "thin". Will it be able to take part in a full-fledged nuclear reaction, or will it fly through, like it happens with the electron in the s – states of the atom? There are very narrow phase areas, when soon after the particle stops the charge grows quickly and is sharply accelerated. The charge can even be maximum in the nuclear force effect area. May be, it is this narrow phase area that is responsible for cold nuclear fusion, and in case of strong interactions the phase hole mechanism must be operating as well.

It was discovered long ago that nuclear transmutations have a mass character (especially in plants and biological objects), but they have little to do with energy generation. Examples of such reactions:

$$Mn^{55} + p \rightarrow Fe^{56} \ ; \ Al^{27} + p \rightarrow Si^{28} ;$$
$$P^{31} + p \rightarrow S^{32} ; K^{39} + p \rightarrow Ca^{40}$$

In reactions of this type, a very slow proton (with practically zero kinetic energy) penetrates the nucleus in the above-mentioned way and remains there. *No intranuclear energy is generated, because both before and after the reaction the nucleus remains a stable object.* In classical nuclear physics, the nucleus usually became unstable after it was penetrated by a charged nucleon with a large kinetic energy and always broke into parts, and the nuclear debris had an even greater kinetic energy. Reactions of the above type were considered impossible under small energies and for this reasons were not studied by classical nuclear physics. It seems to be *a completely new type of nuclear transmutations, not recognized by modern nuclear science, but experimentally discovered rather long ago.* Today there is a great deal of experimental material confirming mass nuclear transmutation phenomena. Moreover, there are many projects of neutralizing nuclear excess with the help of this technology. Journals *Infinite Energy, New Energy, Cold Fusion, Fusion Facts*, etc. and Internet are full of such projects.

Of course, a change in the nuclear charge will result in restructuring of electronic atom shells, but the energy related to this process will be about several electron-volts and is nothing in comparison with the energies of nuclear reactions from several to hundreds million electron-volts. By the way, nuclear engineers are accustomed to such energy ranges in nuclear reactions. It was this circumstance that made them deny *a priori* all nuclear processes in biology, because under such energy values of the debris dozens and hundreds thousand of complex biological molecules will be destroyed.

Quite a long time ago, Lois C. Kervran [33] wrote a book about nuclear transmutations in biology, and now, nearly 20 years later, its second edition was published! It gives, evidently for the first time, numerous *experimental* data of the above-mentioned phenomena. The reaction of the official science is very interesting. For example, well-known physicist Karl Sagan, after reading the book about such *experimental* data, advised Kervran to read elementary textbooks in nuclear physics!

Some time later a research was made by Panos T. Pappas [58], who studied one of the well-observed nuclear reactions in biological cells:

$$Na_{23}^{11} + O_{16}^{8} = K_{39}^{19}$$

Classical biology has long known about the existence of equilibrium, when the ratio between the number of *K* and *Na* ions is maintained with greatest accuracy despite the shortage or even absence of K ions in food. Later, in work [59] this nuclear reaction was even called the life equation, and the existence of such nuclear reactions in biological objects was proved by M.Sue Benford with the help of direct physical methods.

All thermonuclear fusion programs are based on blunt heating and compression of the reacting material. Despite the progress achieved, the head of the works in England, Dr. Alan Gibson [34], established several years ago that the model reactor design would be created not earlier than in 50 years. Today, this point of view is generally accepted. Even if the reactor is once made (although the authors have grave doubts about it), it will be very complicated, expensive, and harmful for the environment.

Classical approaches have so far not yielded any positive results, despite multi-billion investments and a great number of physicists, engineers, service personnel, and managers involved. It is only natural that this army of researchers is a potential impediment for all alternative projects of new power engineering. It has been noted that *"viability" of any idea is proportionate to the number of people involved and investments made.* For these reasons, the Fleischmann-Pons works were given a hostile reception in the USA and other countries.

All the controlled thermonuclear fusion programs are accompanied with the adjective "controlled", although there is no control whatever. It is simply that the initial quantity of the reacting substance is *prudently* made very small. For example, a ball of lithium deuteride during laser reduction has a diameter of several mm. So far no one has been seriously considering the question of utilizing the energy of an explosion of such a ball, which is approximately equal to the energy of an explosion of a box of antitank grenades.

The straightforward approach to fusion used by the modern science is very natural, because quantum mechanics has no methods of influencing this process. The future of really controlled nuclear fusion systems may be not on the way of primitive and blunt method of heating and compression of the material, but following UQT on the way of using collisions of nuclei having small energy and corresponding fine adjustment of the wave function phase.

It is essentially possible in case of imposition of the external controlling electromagnetic field on the reacting system, which contains quasi-fixed ordered deuterium atoms and free deuterons. The same properties can be demonstrated by special atomic grid geometry. Diffraction scattering of a deuteron flow on such grids will lead to automatic deuteron selection by energies and phases.

It seems that in the Fleischmann-Pons electrochemical experiments such an ordered system existed in the *Pd-D* grid, and some phasing occurred, which accounts for the results of these experiments [30].

Today it appear to us that the cold nuclear fusion processes will be effectively used for nuclear waste liquidation and production of isotopes.

Many researchers [35,36] discovered that the quantity of heat generated in the process of electrolysis of ordinary water on nickel electrodes (there is no hope for nuclear reactions in such systems) is the same as in the electrolytic cell with heavy water. It confirms other measurements, which showed that the quantity of nuclear reaction products is millions times less than is required for such an amount of generated heat, and its origin remains a mystery.

Further we will give certain concrete data demonstrating the phase values of a deuteron with an oscillating charge, under which the deuteron can approach the nucleus to a critical distance of 10^{-12} cm or less, i.e. giving the data to estimate the value of the above-mentioned phase "hole" in the interval $(0,\pi)$ of the phase change.

Assume that the stationary nucleus with the charge q is placed to the coordinate origin x=0 and the deuteron with the same charge q is placed at the initial moment t=0 to the point $x_0 < 0$on the x-axis, and the deuteron velocity equals $\dot{x}_0 = v_0 > 0$. The units of mass, length and time are chosen in such a way that $m = 1, \hbar = 1, c = 1$ (*m* - deuteron mass, *c* - light velocity). Charge *q* equals 0.085137266. Our units are connected (to 4 significant figures) with the system (kg, m, s) as follows:

1 mass unit = 3.345×10^{-27}kg,
1 length unit = $1.049\cdot 10^{-16}$ m,
1 time unit = $3.502\cdot 10^{-25}$ s.

The electron velocity corresponding to its energy of 1eV equals $5.931\cdot 10^{7}$ cm/s. The deuteron velocity corresponding to such energy will be assumed to be 3680 times less, and in our units it will be $5.372\cdot 10^{-7}$ (if $c = = 3\cdot 10^{10}$ cm/s). Then the deuteron movement towards the nucleus is described by the equation

$$\ddot{x} = -\frac{2q^2}{x^2}\cos^2\left(\frac{1}{2}(t+t_*)\dot{x}^2 + x\dot{x} + \varphi_0\right), \qquad (10)$$

where the parameter t_* is defined under the condition that the argument of cosine equals φ_0 for $t = 0, x = x_0, \dot{x} = \dot{x}_0$ (thus $t_* = -(2x_0)/\dot{x}_0$), and this parameter may be considered as the initial moment of so called local time.

In the interest for us are namely solutions of eq. (10) under very small deviation ε from phase φ_0 and so we put $\varphi_0 = \frac{\pi}{2} + \varepsilon$ and rewrite eq.(10) in the following form:

$$\ddot{x} = -\frac{a}{x^2}\sin^2\left(\frac{1}{2}(t+t_*)\dot{x}^2 + x\dot{x} + \varepsilon\right), \qquad (11)$$

where *a* = 0.0144967 Let the initial x_0 to be equal - 500000 of our length units (i.e. approximately $5\bullet 10^{-9}$ cm) and the initial deuteron velocity v_0 to be equal to the velocity v_{00} corresponding to the deuteron energy of 1 eV or less. But it turned out that the precision of numerical integration of this equation under such initial conditions and under values $|\varepsilon| = 10^{-15}$and less is small and besides the interval of the integration must be very large. That is why this equation also had to be transformed by passing to "slow" time $\tau = |\varepsilon| t$ to the equation relative to the variable $w = \left(\frac{dx}{d\tau}\right)^2$ as a function of *x:*

$$\frac{dw}{dx} = -\frac{2a}{x^2}\left\{\frac{1}{\varepsilon^2}\sin^2\left[|\varepsilon|\left(\frac{1}{2}(\tau+\tau_*)w + x\sqrt{w} \pm 1\right)\right]\right\}, \qquad (12)$$

where $\tau_* = -(2x_0)/\sqrt{w(x_0)}$ and +1 if ε >0, and -1 if ε < 0. It must be added also the equation for τ as a function of *x* :

$$\frac{d\tau}{dx} = \frac{1}{\sqrt{w}}. \qquad (13)$$

The system of equations (12, 13) is, so to say, a "model" system describing fairly accurately the deuteron movement under all values of |ε| from 10^{-24} to 10^{-6}.
Numerical integration of this system was fulfilled under different values of e and under following initial conditions:

$$w(x_0) = 2.103, \tau(x_0) = 0, x_0 = -500000, \tau_\bullet = 689573.18 \qquad (14)$$

It may be noted that the initial deuteron velocity v_0 equals 1.450172 (following the relation $\dot{x}_0 = |\varepsilon|\sqrt{w(x_0)}$) for given initial $w(x_0)$ and for |ε| = 10^{-7} , i.e. such velocity is approximately 3.7 times less than velocity v_{00} corresponding the deuteron energy of 1eV. If $|\varepsilon|=10^{-6}$ then the velocity v_0 is approximately 2.7 times greater than velocity v_{00} .

It turned out that the numerical tables for values of w, τ obtained under different values of $\varepsilon < 0$ in the interval (-10^{-24}, -10^{-6}) don't differ essentially from each other. The following table is true to three-four significant figures for τ and $\dot{x}/|\varepsilon|=\sqrt{w}$:

x	τ	$\dot{x}/\lvert\varepsilon\rvert$
-500 000	0	1.450
-50 000	$1.426\cdot10^{6}$	0.0493
-500	$1.002\cdot10^{7}$	0.000489
-200	$1.067\cdot10^{7}$	0.000440
-100	$1.090\cdot10^{7}$	0.000425
-80	$1.100\cdot10^{7}$	0.000423

If reducing the table values of x to centimeters, we obtain the following corresponding approximate values:

$$5\cdot10^{-9}, 5\cdot10^{-10}, 5\cdot10^{-12}, 10^{-12}, 0.8\cdot10^{-12}$$

The time interval ΔT , in which the deuteron reaches the critical distance 10^{-12} *cm* from the center, is $67350/|\varepsilon|$ of our time units or $(1.090\cdot10^{7}/|\varepsilon|)\cdot3.502\cdot10^{-25}$ seconds. If nuclear forces are not taking into account then the deuteron may approach the distance less 10^{-12} *cm* .

We present here the table, where are given the initial deuteron velocities v_0 in velocities shares v_{00} and the corresponding time intervals ΔT (in seconds) for different values of ε.

ε	$\dfrac{v_0}{v_{00}}$	ΔT (s)
-10^{-6}	2.7	$3.82\cdot 10^{-12}$
-10^{-7}	0.27	$3.82\cdot 10^{-11}$
-10^{-22}	$0.27\cdot 10^{-15}$	$3.82\cdot 10^{4}$ (≈ 10.6 hours)
-10^{-23}	$0.27\cdot 10^{-16}$	$3.82\cdot 10^{5}$ (≈ 10.6 hours)

Let us note that the given data change essentially under positive values of ε (10^{-6}, 10^{-7}, etc.). There is some asymmetry of solutions behavior under negative and positive values of ε. The calculations show the minimum distance $|x|_{\min}$ more than 500 of our lengths units even for relative big initial $w(x_0)$ = 10000. Thus, if we limit ourselves to the condition that the deuteron energy is not over $(0.27)^2$ eV at a distance of $5\cdot10^{-9}$ cm from the central nucleus, and the whole process of deuteron movement towards the nucleus does not exceeds approximately 10.5 hours, then the interval $(\frac{\pi}{2}-10^{-7}, \frac{\pi}{2}-10^{-23})$ is approximately the sought phase "hole" in the whole interval ($(0,\pi)$) of phase change φ_0 in eq. (10).

If many deuterons with the energy not more than $(0.27)^2$eV at the distance $5\cdot10^{-9}$ *cm* from the nucleus are equally distributed along their phases φ_0 , the ratio of the length of this "hole" to π, equaling approximately $0.3\cdot10^{-7}$, is equal to a share (or a relevant percentage of $0.3\cdot10^{-5}$) of deuterons overcoming the Coulomb barrier.

The above figures express at least the order of probability of the cold nuclear fusion occurrence, and this order is absolutely incompatible with the figures in the classical quantum mechanics mentioned above. Let us note once again that a one-dimensional problem was solved, and in case of an accurate analysis (not zero sighting distance will be taking into account) this probability will be lower. Let us also pay attention to the large time intervals ΔT calculated if $|\varepsilon|$ is very small. It explains well the effect (observed by many researchers) of continuation of cold nuclear fusion reactions during even many hours after the disconnection of the voltage in the electrolytic cell. This effect was named even "life after death".

As for the analysis of the deuteron movement with the help of the autonomous equation, the calculations lead to initial velocities v_0 , exceeding the above mentioned numbers, although the general motion picture is the same. But the autonomous equation is interesting, because in the area of those values $x, \dot{x}$, under which the product $x\dot{x}$ is modulo small, it is possible to replace $\sin(x\dot{x})$ with $x\dot{x}$, and the eq.(11) under ε=0 to replace with simplified equation (describing the deuteron motion from initial point $x_0 > 0$ to center)

$$\ddot{x}=a\frac{(x\dot{x})^2}{x^2}=a\dot{x}^2$$

This equation has a very simple analytical solution. Without giving very simple calculations, we will present the final formulas.

Let us take the following initial conditions:

$$x(0)=x_0>0, \quad \dot{x}(0)=-v_0<0$$

Then

$$\dot{x}(t)=-\frac{v_0}{1+av_0t}, x(t)=x_0-\frac{1}{a}\ln(1+av_0t)$$

It follows from these formulas that the velocity of a particle moving in accordance with the initial equation never turns to zero, and under

$$t=t_*=\frac{\exp(ax_0)-1}{av_0}$$

$x(t_*) = 0$, i.e. the particle reaches the center of the nucleus, its velocity at this moment being

$$\dot{x}(t_*) = \frac{-v_0}{1 + av_0 t_*} = -v_0 \exp(-ax_0) \text{ ,}$$

so that it *passes through the nucleus and moves further.*.

For example, let $a=0.0144967$, $x_0 = 1000$ ($\approx 10^{-11}$ cm), $\dot{x}(0) = 5.37 \cdot 10^{-10}$ (≈ 16 cm/s).

Under such initial data, the product $x\dot{x} = -0.0000537$, so it is quite possible to replace $\sin(x\dot{x})$ with $x\dot{x}$.
In this case,

$$t_* \approx 2.3 \cdot 10^7 \qquad (\approx 8 \cdot 10^{-18} \text{ s}),$$

$$\dot{x}(t_*) \approx -29.9 \cdot 10^{-17} \qquad (\approx 9 \cdot 10^{-6} \text{ cm/s})$$

These figures fit well into the reasonable framework, so the autonomous model can also be of use for the movement analysis in the problem under review. The phenomenon of particle passage through the Coulomb potential accounts very well for the existence of pendulum orbits in the Bohr-Sommerfeld model, when in states *1s,2s,3s* etc. the electron passes through the nucleus. Such states in the strict theory and experiment have no impulse, so in the Bohr-Sommerfeld model they were discarded as absurd. Now they have a right to existence. Further, the experimental data for angular distribution of non-elastic scattering by nuclear reactions (including reactions with heavy ions) reveal the big amplitude of the scattering forward. It is impossible to explain such effect by the formation of intermediate nuclei but it is may be explained from the viewpoint of our UQT.

General Principles of Creating New Energy Sources

In the ancient classical *perpetual mobile* idea it is supposed that energy is just created and not taken from outside (impossibility of *a perpetual mobile* is the first law of thermodynamics). There have appeared lately many articles and even books dwelling on the idea of energy generation from vacuum. We are not in complete agreement with many of these works, and we will dwell only on some of them, which, in our view, can be of interest. One of the main ideologists of this completely new sphere in science are Daniel C. Cole and Harold E. Puthoff, and their first serious work entitled «Extracting energy and heat from the vacuum» was published in Physical Review E, vol. 48, #2, (1993). In this work authors use the Casimir forces [60] making them produce useful work. The appearance of such forces in vacuum is understandable intuitively: if in a stormy sea we put vertically into the water two big parallel plates, on the outside part of these plates the waves will hit them at random, and between the plates there will be no waves. Then, the hitting of the waves outside the plates will produce a gravitation force between them (the Casimir force discovered experimentally long ago), which the authors of this interesting work are going to exploit. It is easily seen that in this idea energy is generated from vacuum fluctuations.

Our approach is altogether different. When the equation with an oscillating charge was solved for the quantum oscillator, 4 types of solutions were discovered. For us only two of them matter – "crematorium" and "maternity home". In one solution ("crematorium") the particle slowly falls to the bottom of the pit and finally turns into a "specter" (under the strict unitary quantum theory it disappears, spreads about the Universe and contributes to vacuum fluctuations everywhere). In the other solution ("maternity home") the particle can even be born of a very small fluctuation, or accumulate a sufficiently big energy. Let us underline once again that both these processes are not at all logically connected. In other words, there are such systems where energy will disappear completely (electrolytic baths), or increase unlimitedly (it might be our Universe).

It is the energy conservation law that presents the strongest impediment in all cosmological approaches. However, universes with birth of matter have long existed in scientific cosmology independently of us. There is known the theory of British astronomer Fred Hoyle based on the idea of continued creation of matter from nothing. The question of whether such an approach is realized in nature and whether the energy emitted by quasars is the result of work produced by a certain gigantic pit, is the most intriguing question of the future.

It is yet unclear whether the values of appearing and disappearing energy in these solutions are equal. But neither in the strict UQT nor in the equation with an oscillating charge vacuum (as a big set of random oscillations) is needed for energy generation. Of course, UQT admits of such an energy exchange with vacuum. For example, during split of a photon on a semitransparent mirror, at one time both halves of the photon will not be registered and will give their energy to the vacuum and disappear for the observers for good, at another time there will appear two photons out of one, and the lacking energy will be taken from vacuum. But *the movement equations (4) and (5) themselves know nothing about vacuum and can generate energy due to their nature* (they are noninvariant relative to the coordinate translations) and the conservation laws we are so accustomed to do not exist for them.

Let us remind you once again that the latter follow from the Newton equations, and the Newton equations result from averaging by a big number of events, while for individual events of small energies no conservation laws in quantum physics exist.

In other words, it can be said philosophically that a motion of a small wave packet, once started, will give birth to other movements (energy) and, consequently, to matter. Since most various and breath-taking speculations are possible, up to the creation of a universe, we will stop here.

Thus, the generated or disappearing energy in our approach can be manifest not only in the changes of the particle velocity during movement in a certain potential, but in the appearance or disappearance of the particles themselves as well. A change of the particle velocity in movement is most easily discoverable, and it is the velocity increase that can be used for generation of heat or electric current. There can be energy systems, which exploit the fact itself of charge oscillation and the consequence of it. It is very probable that these phenomena, contradicting the most fundamental laws of modern science, have been long discovered and even applied. But these are the very phenomena that are the easiest to be exploited at the first stage of development of such new energy technologies.

When an energy generation mechanism is used, crematorium-type solutions should be suppressed. But all the quantum processes are built on the basis of elementary acts, and each of them is impossible to be controlled separately. But if the probabilities of such processes are controlled, they, being multiplied by the great number of participants in the process, automatically become macroscopic variables of quantum kinetics, and the process itself becomes possible. It can easily be achieved, if process participants with correlated initial phases are selected.

Let us remind you that the Newton and relativistic classical mechanics follow from the strict UQT, while the Newton movement equations with the resultant energy and impulse conservation laws follow from the oscillating charge equation with averaging by the particle ensemble composing a classical body (material point). But these conservation laws are nonexistent for individual microparticles in our theory, and they appear only in case of averaging by the ensemble of particles. Thus, if the energy-generating processes are accumulated, and the processes where energy disappears are suppressed, a classical *perpetual mobile* can be created.

But the UQT and the oscillating charge equation have other differences not only from the equations of classical mechanics, but also from some equations of electrostatics and electrodynamics.

There is a fundamental theorem of circulation for the electric field. Let us dwell on it in more detail. Let us have a vector field **E,** which can be an electrostatic or a gravitation field.

$$\mathbf{E} = \mathrm{P}(x,y,z)\mathbf{i} + \mathrm{Q}(x,y,z)\mathbf{j} + \mathrm{R}(x,y,z)\mathbf{k}$$

Line integral

$$\Gamma = \oint_l (\mathrm{P}\,dx + \mathrm{Q}\,dy + \mathrm{R}\,dz) = \oint_l \mathbf{Edl} \qquad (15)$$

is called circulation of vector field **E** by contour *l*. Of course, circulation depends not only on **E**, but also on the passage direction accepted in contour *l*; by changing the passage direction we will change the circulation sign. Form (15) is convenient for mathematicians, but for our purposes we will present equation (15) in a different way. If both parts of equation (15) are multiplied by electric charge *q,* on the right we will get the integral of force $q\mathbf{E}$ by way **dl,** i.e. work for moving the charge along a closed loop contour. It is well known that this value is zero.

$$\int_l q\mathbf{Edl} = \mathbf{0} \qquad (16)$$

If this value were not zero, an energy source could be created. For this purpose a charge should be moved in electric field **E** from point *a*, located in the high voltage area of the field, to point *b,* located in low voltage area of the field, and then back, but along another route. The values of work from $a \to b$ and from $b \to a$ would be different, and we could extract work from the field without making any changes in the system. When the charge is constant, it is certainly true, so for a macroscopic constant charge this theorem is an analog of the energy conservation law. The authors have not come across such an interpretation of the energy conservation law in other works. If the charge is microscopic, then in the UQT it changes, depends on time, coordinate and velocity, so work from $a \to b$ and from $b \to a$ will be different, in this case work can in principle be extracted from the field without any changes made in the system.

Discussion of Experimental Results

Let us now get down to explaining some very unusual experimental results, which the authors have nothing to do with, and which they sometimes regard rather skeptically. The point is that the sphere of new energy sources is the headache of all the human civilization, and in this sphere, like nowhere else, the dividends can be exorbitantly high, and for this reason there are in this sphere a lot of swindlers (even among the theoreticians) and simply erring people. The official science of the world does not so far believe in such research, but the most suspicious fact is the great multiplicity of such works. The authors are not inclined to regard all these people as swindlers or erring, because the UQT can offer a beautiful and simple interpretation of certain phenomena.

There are strange plants with the efficiency over 100%. They are even manufactured in small quantities and are rated among energy-saving devices already termed over unities. Japanese researchers take these problems very seriously, and the leading role in studying this problem belongs not to the USA, but to Japan, which even finances many US institutes in this framework. The total Japanese expenditures for this research exceed $200.000.000 a year. It can be forecast that with the Japanese mentality and the state policy of exporting not natural resources, but superhigh technologies and intellect, Japan will find itself among the leading countries early in the 21st century. We think that our readers will not be surprised to hear that Russia has not allocated a single cent for this program, and all research was made on pure enthusiasm.

In the USA such works do not get official governmental support either (like, for example, the dying out hot nuclear fusion problem), but a great number of private firms and individual businessmen are conducting large-scale research. The following US journals are devoted to the subject: *Journal of New Energy, Infinite Energy, Cold Fusion, New Energy News, Fusion Facts*, and *NET-Journal* (Switzerland).

Switzerland, Italy, Germany, and France are also among the countries where the new energy problems are seriously researched from the cold nuclear fusion point of view.

A very young sphere of power engineering has emerged and is quickly developing, which researches many new energy sources. In future those new energy-saving sources will first be used, which will considerably differ from the existing ordinary energy transformers in that they will generate additional energy that can be used in the interests of the mankind. The development of civilization will then be limited not by long-expected reduction of natural fuel resources, but by heat pollution of the environment.

Let us enumerate just a few of the new energy directions:
1. The Patterson fuel cell (CETI).
2. Supermagnet-superengines of Takahashi, Aspden and Adams
3. Swiss plant Testatika.
4. Engines operating on water.
5. Hypersound Griggs pump, the Potapov and Schaffer heat generators.
6. Schoulder and Fox cluster systems.
7. N – machines of Farade, Bruce de Palma, Newman, Searl, Tewari, etc.
8. PAGD reactor of Canadian researchers P. Correa and A. Correa.

This list can be complemented with the surprising experimental results received by physicists A. Samgin and A. Baraboshkin (Russia, Institute of High-Temperature Electrochemistry under the Russian Academy of Sciences, Ekaterinburg) [24,25] and T.Mizuno [26] (Japan). They appear to have used, totally independently of each other, special proton-conducting ceramics, which, when electric current runs through them, generate a thousand times more heat energy than the electric energy consumed. In some experiments by T.Mizuno this value even exceeded 70000(!). T. Mizuno in a personal talk with one of the authors of this report said that he feared very much the radiation sickness. But no α, β, γ radiation or nuclear debris was found, and the nuclear processes are not responsible for such energy generation. Such proton-conducting (or, to be more exact, deuteron-conducting) ceramics was made using the power metallurgy methods by agglomeration under high temperatures. In other words, all the chemical processes in it had long been over. The origin of such an amount of excessive energy is absolutely incomprehensible in the framework of conventional science, for they cannot be accounted for either by nuclear or chemical reactions, or by phase passages. At first the authors of this experiment supposed nuclear fusion reactions of the D+D type. At our request, A. Samgin replaced heavy hydrogen (deuterium) during ceramics production with ordinary hydrogen. If the effect of such huge energy generation was connected with the nuclear D-D reactions, all the anomalous heat effects would have disappeared, but they persisted. After such a large quantity of energy was generated, the tablet disintegrated into powder.

These effects can easily be explained by UQT from the harmonic oscillator theory point of view. When the tablet is agglomerated, there remain in it some caverns of a size of hundreds Angstrom units. When direct or alternating current flows through it, the protons and deuterons in their movement (there are few electrons in such ceramics) get into these caverns, and a process can start which is described by the “maternity home” solution. A particle accumulating energy, oscillates in such a pit, and finally the energy will be sufficient both for heating and for destruction of the pit walls (tablet turning into powder). The same processes seem to be taking place in a palladium electrolytic cell with heavy water, and in a nickel electrolytic cell with ordinary water, which accounts for anomalously large heat generation, not related to nuclear processes.

It would be good to verify experimentally the dependence of the tunnel effect on the initial phase. But it seems us that it is more important for our opponents, since both cold nuclear fusion (CNF) and discovery of nuclear transmutations *(which, from the point of view of modern science, are even more absurd than the existence of CNF*) evidently cannot be accounted for in any other way. Besides, such a direct experiment is of a fundamental value. There are today a lot of people and groups in the world, who pin great hope on exploiting the nuclear transmutation phenomenon for the purposes of processing and recycling of nuclear wastes, and the question of industrial generation of tritium for military purposes using CNF methods was under consideration in Los-Alamos. Internet magazines are full of such information. We are not giving Internet addresses here, because everything is constantly changing in this live system.

Let us analyze some of the above-mentioned devices. The first, the oldest and the most mysterious information was information about internal combustion engines operating on water.

Let us give just one example. When we were students, one of our teachers, the late Professor G.V. Dudko (1959) told us that in 1951 he had participated in the testing of an internal combustion engine [39,55-57]. The device represented a hybrid of a diesel and an ordinary carburetor engine, where a gas of petrol was needed to start it and then ignition was switched off, and an ordinary fuel pump sprayed into the cylinder warmed up and strongly compressed water with special additives (which the inventor himself put into the tank in small quantities, and which, as we now understand, represented the principal secret). The engine was installed on a boat. The researchers were riding for two days in the Azov Sea, and only water vapor was the engine exhaust. Professor Dudko himself drew the water fuel overboard and poured it into the tank. They needed much water, several buckets a day, but there was no shortage of it… The question of why, if everything was so great, these engines are still not in use, can occur only to a person who has never lived in Russia.

From the point of view of the solutions of the harmonic oscillator problem, the following theoretical possibility exists [40,44,47,55-57]: if water with the necessary additives (which, evidently, represent the secret of many invented engines operating on water) is compressed and sprayed into the cylinder, each drop of water, when it gets into the cylinder after being compressed, will start dilating and will pass by inertia the equilibrium position. As a result, caverns (empty volumes) can be formed in it, with a size of several dozen of Angstrom units. If a free proton (or some other microparticle) gets into such a cavern in the required phase (it is supposed that the task of the additive is exactly this), the "maternity home" solution will be realized and some of the drops will explode… Later we heard and read many times about various Russian inventors, who had successfully created and tested engines operating on ordinary water with some mysterious additives.

Of course, the possibility of catalytic water decomposition with small energy consumption before spraying into the cylinder is not at all excluded. There are films and information in Internet about testing of cars operating on water, which is catalytically (with small energy consumption) decomposed into oxygen and hydrogen. Such power engineering would be ecologically absolutely clean, and the only restriction would lie in heat pollution of the environment.

An ideal solution for the motor transport could also lie in use of some new types of electric energy generators. The UQT even admits of the possibility, which was long observed in the experiments of Nicolas Tesla and in those made by Canadian physicists the Correas, who even received a patent for a system generating energy from vacuum fluctuations (as they believe) [45]. The readers could have got acquainted with our detailed theory of these processes in [46]. But the ideal system for the automobile would certainly be Testatika.

Any imagination will be amazed at the thermal cell CETI created by James Patterson, USA [27], in which takes place the electrolysis of specially made nickel balls in ordinary water. The US paper «Fortean Times» [1] 85, 1995, wrote about it: "December 4, 1995 will go down into history. On this day a group of independent experts from 5 US universities was testing a new source of energy with a stable output heat power of 1.3 kWt. The consumed electrical energy was 960 times smaller". All experts note that the generated heat is of mysterious origin and cannot be explained by chemical or nuclear reactions, as well as by phase passages. The US ABC TV showed on February 7 and 8 1996 in the «Nightline» and «Good Morning America» cycles of programs about the development by Patterson of a new energy source generating hundreds of times more energy than it consumes. The mysterious nature of the generated heat was again underlined. It is interesting [34] that Motorola tried to buy the CETI patent from its authors for $20.000.000, but met with a refusal. We are sure that Motorola had invested a certain amount of money into the study of this problem before making such a serious offer. All that happens within the Patterson element has nothing to do with nuclear reactions (although Patterson told one of the authors that he was of a different view), and, in our opinion, can be accounted for by exactly the same processes as were described above for proton-conducting ceramics.

The sonoluminescence phenomenon, when certain liquids start shining if weak ultrasound is run through them, also looks very mysterious. No satisfactory explanation has so far been found for this experimentally proved phenomenon, discovered by Moscow University Professor S.N. Rzhevkin in 1933. As Nobel Prize winner Professor Yulian Schvinger said, "it has no right to exist, but it does exist" [38]. This phenomenon can also be explained from the above-mentioned positions.

There are also heat generators (Yu. Potapov [21-23], Moldavia, James J.Griggs [28], and Huffman [29], Schaffer - USA). In them many cavitating bubbles are formed during circulation of ordinary water, in which excessive energy is generated, with the output to input energy ratio approaching 1.7. In these experiments and plants no chemical or nuclear reactions can take place, and thousands of Potapov's heat generators have been manufactured for heating homes. In such devices (they are very different in appearance) a great number of cavitating bubbles are created in a flow of water. This is achieved either with the help of interrupting the water flow with a special rotor (J.Griggs, Huffman, Schaffer), or the water flow is twirled by a special helix and then enters the zone of sharp dilation, where cavitating bubbles are formed (Yu. Potapov). In general, it should be said that cavitation remains a great puzzle for theoretical hydrodynamics and science. For example, forged multi-ton screw propellers of big nuclear submarines under certain operation modes and geometry of the surrounding forms can be destroyed by cavitation within only a few hours. It happens because of huge energy generated in cavitating bubbles.

Under certain values of phase and energy, a particle in the pit, each time reflecting from the walls, will have a greater velocity than that of a falling particle (this is within the uncertainty relation), and after many reflections will accumulate a fairly big energy which will be generated in the form of heat or bremsstrahlung when the pit is destroyed, and, finally, the energy of the oscillations of such a particle accumulated in the pit will always be transformed into heat in an ordinary solid body or a liquid. This physical idea immediately accounts for both sonoluminescence (although for sonoluminescence in general this mechanism is less primitive), and energy generation in proton-conducting ceramics, nickel during electrolysis in ordinary water (CETI element), and water bubbles of commercial heat generators. The theory predicts that the samples should be fissure due to increased pressure on the walls of the potential pit with the growth of energy, which fact also takes place, since both ceramic samples and nickel balls in the CETI element finally disintegrate. It is evidently for these reasons that any metal containing much hydrogen in its grid becomes fragile and is quickly destroyed, which fact is well known to engineers.

The small number of experiments does not so far allow for making concrete conclusions as to what particles generate energy in pits (microbubbles). Besides, for at least an electron to disappear a pit of about 0.5 MeV is required, while in a solid body the pits are about several eV deep, and what seems to happen is only loss of kinetic energy, and not disappearance of particles. The fact that this process requires very deep potential pits, which do not exist in a solid body, does not change the essence of the matter.

Of course, under ordinary conditions, both competing solutions usually take place at once: “maternity home” and “crematorium”, which compensate for each other and the energy is preserved. For energy generation, the “maternity home” solution should prevail. Both these processes take place simultaneously and compete with each other, but, formally, they are not connected in space and time. The complexity of the energy generation problem lies in suppressing the “crematorium” solution by a careful selection of different parameters and promoting the “maternity home” solutions. So far we cannot say for sure what the optimum dimensions of such cavitating bubbles are, or which object oscillates in them, because for this purpose special experiments are needed, which so far have not been staged.

Of course, the inexorable Robber in the form of the Carnot principle stands in the way of transformation of the heat generated in a heat generator or ceramics into electrical or mechanic energy. In accordance with this principle, all mechanic or electrical energy can be transformed into heat, but the reverse process is always connected with big losses.

If there are experiments and plants in which energy generation contradicting the conventional conservation laws is discovered, there should also exist opposite ones, where energy disappears completely, i.e. the “crematorium” solution prevails. It proved to be true. There are such modes during electrolysis in electrolytic baths, under which the temperature of the solution in the bath is strongly reduced for unaccountable reasons, and this fact has no explanation at all. This phenomenon long ago was noted by attentive industrial engineers, and it is called the “bath-freezing” mode [49,50]. Chinese physicist Swe-Kai Chen from Taiwan in his experiments [48] stably observed the same phenomena. It is quite easily explained: a particle with a velocity exceeding the most probable velocity in this distribution gets into caverns on electrodes and after some oscillations reduces its velocity, which becomes smaller than the most probable one, and then the particle leaves the cavern at a small speed, and the same process can happen to another energetic particle. This leads to the cooling of the cell in the case of such mass processes.

The problem of ferromagnet magnetization (the Easing model) can also be reduced to the orientation of a magnetic doublet by the external magnetic field, and then it is essentially the harmonic oscillator equation with a slightly different return force ($F = \frac{\mu}{r^3}$) and all the conclusions made earlier remain in effect. That is why magnetization should also produce energy generation effects. This proved to be true. For the general public everything began on May 17, 1996, when Frode Olsen from the research group “Free Energy” showed on the Norwegian TV (TV2) a surprising film about a “dynamic sculpture” made by artist and sculptor Reidar Finsrud from Skaarer, Norway. The author of this “dynamic sculpture” *had no idea about physics* and had been making it for 12 years. Einstein’s idea of how discoveries are made conveniently comes to mind at this point: everyone knows that a certain thing cannot be done, but there is a man who does not know it, and it is he who makes the discovery.

This “dynamic sculpture” accompanied by an “explaining” poster «*perpetual mobile*» represents an iron well-polished ball with a diameter of 2.7 inches weighing about 2 pounds. The ball is rolling along a circle on close guides resembling two parallel skids with a diameter of 25 inches past the poles of three permanent magnets, where it is magnetized. In the area of three permanent magnets three more mobile magnets are installed on special mobile 5-inch long levers, and these magnets, when the ball passes them, are slightly inclined (due to the ball gravitation) and, after the ball passes them, are raised by the holding springs (sway like yokes). The ball makes a complete turn in 3 seconds. All this magic (***they say the ball had been rolling along the close contour for more than a year***) does not have any sources of energy and is installed for everyone to see in a Norwegian picture gallery on a special stand covered with a glass cover. The authors only saw a good TV film about this installation and were mostly surprised at the fact that the ball had not stopped during uninterrupted shooting (about 20 minutes).

We are well acquainted with circus tricks, but it is absolutely incomprehensible how such a trick could be staged using some secret methods. It is clearly seen that the ball in its movement always partially transfers its energy to the three long swaying pendulums, but there is no way to use them for pushing the ball and making up for friction, this being the only trick that could, in our view, be applied here. All the rest is clearly visible and contains nothing suspicious.

Let us estimate the generated energy. At an initial speed of about 1m/s the ball stops after 30 seconds, if all the magnets are removed. It means that the energy consumed in 30 seconds is about 0.5 joules, or 1/60 Watt. The total energy generated in a month is 43,200 joules, and this is huge energy, much greater than that of a good shell!

It is clear (if the word is relevant here at all) that when the ball is approaching the permanent magnet and the process of magnetization is going on, it is accelerated, but when it mechanically gets past the equilibrium position and, moving away from the magnets, becomes demagnetized, the gravitation (which now starts slowing the ball down) will be slightly less than it was at the moment of the ball's acceleration. This small difference in forces provides for small positive work to overcome friction. Energy generation and similar things during magnetization had been *predicted* by one of the authors in magazines *Infinite Energy* vol.1, No.2, p.38, (1995); *Proceedings of the ICCF*5, p.361, April 9-13, (1995), Monte-Carlo; *Cold Fusion*, No 11, p.10, (1995); *Chinese Journal of Nuclear Physics* (vol. 19, ^{1}2, 1997). The quantum-mechanic processes are very complicated, but some of them can be understood.

All keen physicists were quick to understand it, and J. Naudin in France made a similar, but much simpler experiment. A ball of a soft magnetic material is swaying along parallel U-formed skids in a system of four magnets. Near the bottom of the U-form there is a small smooth step. It may have been made to make the magnetization and demagnetization processes different in time, which is very important. If there are no magnets, nothing interesting happens and oscillations are quickly (in a few seconds) damped. If the magnets are present, oscillations go on up to 3 hours 27 minutes. It appears that in this case the author failed to find good material and parameters of the plant, so friction was not compensated completely. In all these experiments demagnetization of permanent magnets does not happen, because the experiment is repeated many times with the same results.

And now a few vague words about demagnetization processes. During magnetization of the ball, the magnetic moments of its atoms are oriented (like the hands of a compass) along the field lines. When the ball leaves the magnetic field area, the atom magnetic moments are disoriented under the influence of the heat motion, and it becomes demagnetized. In the unitary quantum theory the share of the oriented magnetic moments in the external field can be bigger than in the conventional quantum mechanics (the "maternity home" solution), and the ball gravitation can be stronger due to it. Disorientation of these moments happens similarly in both theories. It seems to be for this reason and due to the difference in magnetization and demagnetization time that a difference in magnetic forces occurs when the ball approaches the magnet or moves off from it.

The scientists of the older generation will remember that a similar toy was shown in the 30's to David Gilbert, who said it was the most interesting thing he had ever seen. A question arises as to why it has not yet been realized. We do not know a physical-mathematical answer to this question, and it is not our task to analyze the social reasons of this phenomenon. Japan has a different mentality, and there is a governmental program for generating energy from permanent magnets. Takahashi [51] even seems to have made an electric engine with an efficiency of up to 318%!

Still more mysterious is the long-known problem of energy shortage in many biochemical reactions with ferments (enzymes). For example, in the well-studied reaction of disintegration of polysaccharides in the presence of lysozyme the following happens: a polysaccharide molecule gets into a special cavern in a big lysozyme molecule, and some time later its debris are thrown out of it (Fig. 6). The broken binding energy of the polysaccharide is about 3 eV, while the energy of the heat movement is only 0.024 eV. From the standard science point of view, it is absolutely unclear where lysozyme takes the energy to break the polysaccharide. No satisfactory mechanism for explanation of such reactions (and they are very numerous) was found, and all this was "swept under the carpet", as physicists say. The UQT provides for a completely new look at the catalytic processes, which has an incomprehensible source of energy reducing the molecule activation energy. From our point of view, this process is a variant of the "maternity home" solution for oscillator.

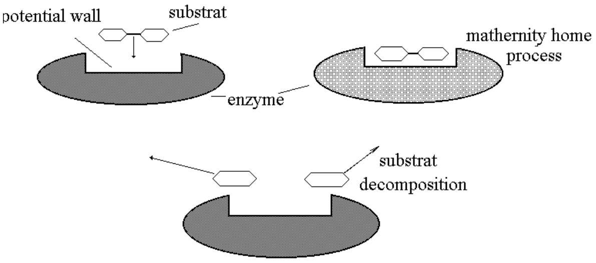

Fig.6. Break of polysaccharide molecule by lysozyme.

The most surprising thing is that in all the cases generation of excessive energy cannot be accounted for by chemical reactions or phase passages. If nuclear reactions do sometimes happen (which should not be according to modern science), they can account for only a hundredth or a thousandth share of the generated heat energy. There is no doubt that all these are effects of new physics, for in the framework of the old physics all this is simply unexplainable.

But the existence of a plant that produces *out of nothing* about 10 kilowatt of direct current electric energy with a voltage of 300 V seems nearly impossible. The story was described by one of the authors in three different magazines, and we will just give a brief resume [52-54].

In summer 1999, at the invitation of Swiss physicists (Director of the Institute of New Energy Sources in Egerkinhen Adolph Schneider), one of the authors visited several research organizations. It is interesting that there is such an institute in small Switzerland, and there is none in big Russia. The purpose of the invitation was very simple: to explain the operation of a plant generating energy out of nothing, i.e. a *perpetual mobile*. In Switzerland such plants are called Testatik Machine M/L Converter from religious group «Methernitha» (Address: Methernitha, CH-3517 Linden, Switzerland, phone: ++41 31 97 11 24).

Such machines exist today in the Swiss town of Linden near Bern. Part of the town belongs to the Religious Christian Community, which is fenced and heavily guarded. There are about 250 members of the Community, many of them are physicists, graduates of the universities of Geneva, Lozanne, Bern. It is not only a research laboratory, they have their own TV center, a film studio, a small furniture plant, shops, garages, residential blocks, and support services. You will probably have guessed that this community does not consume any energy, and this is the most accurate fact in the whole story, for the inquisitive journalists have found out that no money from them comes to the accounts of the local power station, which provides power for all the town. In a cellar of one of the houses they have a power station that produces energy… out of nothing. The author of this inexhaustible source of free direct current energy is Swiss physicist Paul Baumann. Let us briefly describe these fantastic plants: they are of four types (sizes) with capacities of 0.1, 0.3, 3 and 10 kWt. Externally, the plant resembles very much the standard electrostatic machine with Leyden jars often used in physical demonstrations. There are two acryl disks with 36 pasted narrow sectors of thin aluminum, which rotate in different directions. In the first samples ordinary gramophone records were used for disks. The machine is started by pushing the disks in different directions by fingers. The rotation speed is 50-70 turns per minute. After the start disks rotate independently and can be easily stopped by hand, the direct current voltage is about 300-350 V, and the current is up to 30A. The mechanical energy used for rotation (only 100 mWt, according to measurements made by Austrian Professor S. Marinov) is hundreds of thousand times smaller than the generated electrical energy. The biggest plant for 10 kWt has plastic disks with a diameter of over 2 m, the smallest one – 20 cm, the weight of the plants is small enough, the 3-kWt machine weighing about 20 kg.

The charge separation process (which consumes energy!) practically does not slow down the disks. Connection of a load in the form of a 200-Wt bulb does not change the rotation speed either. No cooling or heating of the air or machine parts during long operation takes place, only a slight smell of ozone is felt. The system is noiseless, compact, environment-friendly, and can be installed anywhere.

The Community management thinks, and quite rightly, that wide spread of such systems in the world will lead to a heat explosion, because all the energy generated by the mankind finally finds itself in an energy dump (is transformed into heat), and all this can finally lead to overheating of the environment. They absolutely do not believe (and not without grounds) in the capability of the mankind as a whole to negotiate reasonable use of this invention, and they think that the harm caused by it will be greater than from nuclear, bacteriological, or conventional weapons. Their main idea for the mankind is to live in balance with the environment and to make full use of the energy of the wind, the sun, the water, etc. For this reason the Community is heavily guarded, and they are not going to donate their main discovery to the mankind.

Professor Stephan Marinov visited the Community twice (in July 1988 and in February-March 1989). He was even given such a plant with a capacity of 100 Wt (300 V, 0.3A), which he studied in his laboratory. As far as we now know, even the inventor of this machine does not fully understand its operation principle, so he contacted Marinov out of sheer curiosity of a scientist.

In 1989 Professor Marinov published a book “Thorny Path to Truth – Documents of Violation of Conservation Laws” in International Publishers East-West. The book contains a lot of photos, a measurement report, and a description of the plant. He also organized a research group called “Free Energy” within the Community (Methernitha Group Stephan Marinov Free Energy).

There are very interesting words in this book: “I can state without any doubts that this machine is a classical *perpetual mobile* in its pure form. After the initial push, it goes on rotatingby itself for an indefinitely long time, constantly producing electrical energy in the amount of 100 Watt… It is still unclear, however, how it all can happen…”. As far as we know, nobody has managed to build a similar plant elsewhere.

We have an approximate idea of how the plant operates. The idea is as simple and ingenious as that of the wheel, which is absent from the surrounding nature, so the inventor could not borrow the idea. We will just show that the existence of such a plant is in full conformity with the UQT. It is natural that the plant operates on the basis of the charge separation principle. Let us have two metal spherical surfaces with a hole, isolated from the earth and from each other. If, with the help of an insulated stick, we transfer the first electron from Ball A to the internal surface of Ball B through the hole, a difference of potentials will occur, and if we transfer the second and the subsequent electrons, Ball A will attract the transferred charge, while Ball B will repulse it, and energy will have to be spent during the transfer of charges (Fig. 7).

Let us remind you that under the existing circulation theorem (16), the charge transfer work will consume the same amount of energy as will later be generated during the passage of electric current resulting from charge separation. But in the UQT the circulation theorem (16) for *an individual elementary charge* is not valid. Thus, we can select the time and route, along which the charge will be transferred in such a way, that

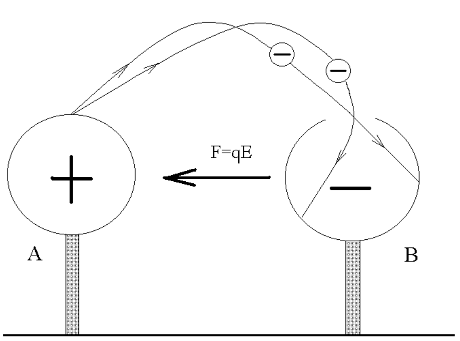

Fig.7. Work for moving the charge depends on method of movement and route.

the charge value during the transfer will be close to zero, and, consequently, the electrostatic force and the charge transfer and separation work will be close to zero too. For example, instead of selecting the route you can wait for the charge to be reduced to zero and then transfer it quickly, and when the charge increases, immediately stop the transfer and fix the charge. Or you can duly select the route and velocity. There are many options. This was evidently realized by Paul Baumann, who is *so far* practically unknown to the official science, and who can find consolation in the idea that the inventor of the wheel will never be known at all. The problem of simple arrangement of all this is just a matter of technique.

You cannot help, thinking that all these might be just tricks. The history of *perpetual mobile* abounds in evidence of downright swindling and frauds, and not a single positive result before, and who can guarantee that the information given above will not prove to be another swindle?

First of all, if all the people always piously believe in the unquestionable stability of the energy conservation law, there will never be any progress in this sphere, and it is then unexplainable how man got down from the palm at all. Secondly, to justify the proposed rebellious position, the following idea comes in mind: if 30 years ago somebody had told the authors (who were then already professors) that at the beginning of the next millennium they would deal in such research, it would have seemed not only a silly joke, but an absolutely impossible thing as well. But, as Voltaire said, “He is silly who does not change”.

In conclusion we wish to express with certainty that the time of theoretical recognition and of practical universal using of overunity devices will come soon and become the epoch of new energetics. The people of our planet will regret that so much oil, coal and gas was burned causing terrible ecological losses.

The authors thank astronaut V.A. Dzhanibekov and Professor A.P. Buslaev.

References

1. Sapogin, L.G. Unitary Field and Quantum Mechanics. In: Investigation of Systems. Academy of Sciences of the USSR, Vladivostok, No 2, p. 54-84, 1973 (in Russian).
2. Sapogin L.G. On Unitary Quantum Mechanics. Nuovo Cimento, vol.53A, No 2, p.251, 1979.
3. Sapogin, L.G. A Unitary Quantum Field Theory. Annales de la Fondation Louis de Broglie, vol.5, No 4, p.285-300, 1980
4. Sapogin, L.G. A Statistical Theory of Measurements in Unitary Quantum Mechanics. Nuovo Cimento, vol.70B, No 1, p.80, 1982.
5. Sapogin, L.G. Statistical Theory of Detector in Unitary Quantum Mechanics. Nuovo Cimento, vol.71B, No 3, p.246, 1982.
6. Boichenko, V.A. and L.G. Sapogin On Equation of Unitary Quantum Theory. Annales de la Fondation Louis de Broglie, vol.9, No3, p.221, 1984.
7. Sapogin, L.G. and V.A. Boichenko On Solution of One Non-linear Equation. Nuovo Cimento, vol.102B, No 4, p.433, 1988.
8. Sapogin, L.G. and V.A. Boichenko On Charge and Mass of Particles in Unitary Quantum Theory. Nuovo Cimento, vol.104A, No 10, p.1483.
9. Sapogin, L.G. Clearcut Picture of Microworlds. “Journal Technique for the Young” (Tekhnika Molodezhi). Moscow, No 1, p.41, 1983 (in Russian).
10. Sapogin L.G., “Statistical Theory of Measurements in Unitary Quantum Mechanics,” Nuovo Cimento, vol. 70B, No.1, p.80, 1982.
11. Sapogin L.G., “Statistical Theory of Detector in Unitary Quantum Mechanics,” Nuovo Cimento, vol. 71B, No. 3, p. 24M, 1982.
12. Boichenko V.A., Sapogin L.G., “On Equation of Unitary Quantum Theory,” Annales de la Fondation Louis de Broglie, vol. 9, No. 3, p.221, 1984.
13. Sapogin L.G., Boichenko V.A., “On Solution of One Non-linear Equation”, Nuovo Cimento, vol. 102B, No.4, p.433, 1988.
14. Sapogin L.G., Boichenko V.A., “On Charge and Mass of Particles in Unitary Quantum Theory,” Nuovo Cimento, vol. 104A, No.10, p.1483, 1991.
15. Sapogin L.G. “Deuteron Interaction in Unitary Quantum Theory”, and “On Mechanisms of Cold Nuclear Fusion”. In: Proceedings of the Forth International Conference on Cold Fusion, vol.4, Theory and Special Topics Papers TR-104188-V4, July 1994, p.171-178, Hawaii. 1994.
16. Sapogin L.G. “Deuterium Interaction in Unitary Quantum Theory”, and “On Mechanisms of Cold Nuclear Fusion”. In: Fusion Source Book. International Symposium on Cold Nuclear Fusion and Advanced Energy Sources, Belarussian State University, Minsk, May 24-26, p.91-98. 1994.
17.Sapogin, L.G. “Cold Nuclear Fusion and Energy Generation Processes in Terms of theEquation”. Chinese Journal of Nuclear Physics vol.19,#2, p.115-120, 1996 .
18. Sapogin, L.G. «Cold Nuclear Fusion and Energy Generation Processes in Terms of the Schrödinger Equation». Infinite Energy [E.Mallove, editor], vol.1, No 5,6, p.75-76, 1996.
19. Sapogin, L.G. «Energy Generation Processes and Cold Nuclear Fusion in Terms of the Schrödinger Equation». In: Proceedings of the Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18, 1996, Japan, vol.2, p.595-600.
20. Sapogin, L.G. “Energy Generation Processes in Terms of the Schrödinger Equation.”Proccedings îf the 2nd Russian Conference CNFNT (in Russian) p.18-24, Sochi, September 19-23,1994.
21. Potapov J.S. Patent of the Russian Federation No 2045715 Heat Generator and Device for Heating of Liquids. Registered on the 10-th of October 1995; priority from the 26-th of April 1993 (in Russian).
22. Potapov J.S. Water as a Source of Life and Energy. Enerjia-takarekossadi, Revu. p.25- 29, September 1998, Budapest.
23. Potapov J.S., MD, Patent No 649 Instalatie pentru obtiinerea enerjiei electrice si-temice. Buletin Oficial de Proprietate Industriala, No12, p.18-19, Chisinau (in Moldavian).
24. Samgin A., Baraboshkin A. et al. Influence of Conductivity on Neutron Generation Process in Proton-Conducting Solid Electrolytes. In: Proceedings of the 4th International Conference on Cold Fusion. Palo Alto, USA, v.3, p.51-57, 1994.
25. Samgin A. Cold Fusion and Anomalous Effects in Deuteron Conductors during Stationary High-Temperature Electrolysis. In: Proceedings of the 5th International Conference on Cold Fusion. April 9-13,1995, Monte-Carlo, p.201.
26. Mizuno T., Enio M., Akimoto T. and K. Azumi Anomalous Heat Evolution from SrCeO$_3$-type proton conductors during absorption/ desorbtion of deuterium in alternate Electric Field. Proceedings of the 4th International Conference on Cold Fusion, December 6-9,1993, Hawaii, vol.2, p.14., EPRI, Palo Alto, USA, 1994.
27. Patterson J.A System for Electrolysis, U.S. patent No 5,494,559,27 Feb.1996; Miley G.H. and J.A. Patterson in: Proceedings of the 6th International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18,1996, Japan, vol.2, p.629-644.

28. Griggs J. Calorimetric Study of Excess Heat Production within Hydrosonic Pump System Using Light Water. Fusion Source Book. International Symposium on Cold Fusion and Advanced Energy Sources, Belarussian State University, Minsk, Belarus, May 24-26, p.248-253, 1994.
29. Huffman M.T. From a Sea of Water to a Sea of Energy, Infinite Energy, vol.1, No 1, p. 38-45, 1995.
30. Fleischmann M., Pons S. Electroanal. Chem., v.261, p.301, 1989.
31. Sapogin, L.G. and I.V.Kulikov "Cold Nuclear Fusion in Unitary Quantum Theory". Chinese Journal of Nuclear Physics, vol.17, No 4, p.360-370, 1995.
32. Cryz W.: Rivista Nuovo Cimento, 1, Special No, 42, 1969.
33.Kervran Lois C. Biological Transmutations. Swan House Pub. Co, NY, 11223, 1972.
34.Private Communication.
35. Notoya R., Noya Y., Ohnisi T. Fusion Technology. vol. 26, p. 179-183, 1993.
36.Swartz M. Journal of New Energy vol.1, #3, 1996.
37. Blokhintsev D.I. On Energy Conservation Law. In: Works on Methodological Problems of Physics, p.51, 1993, Print of Moscow State University. (In Russian).
38. Schwinger J. Casimir "Energy for Dielectric". In: Proceedings of the National Academy of Sciences, vol.87, p.8370-8372, 1990, "Cold Fusion: Does it Have a Future?" Journal "Cold Fusion", vol.1, #1, page 14-17,1994.
39. Sapogin L.G. «What Can Our Power Engineering Be Like in the Next Millennium», Journal Business-Match, 14, 1998 (In Russian).
40. Sapogin, L.G. On One of Energy Generation Mechanisms in Unitary Quantum Theory. Infinite Energy [E.Mallove, editor], vol.1, No 2, p.38-39, 1995.
41. Sapogin, L.G. On One of the Energy Generation Mechanisms in Unitary Quantum Theory. Proceedings of the ICCF5, p.361, April 9-13,1995, Monte Carlo.
42. Sapogin, L.G. Energy Generation Processes and Cold Nuclear Fusion in Terms of the Schrödinger Equation. In: Proceedings of the Sixth International Conference on Cold Fusion, Progress in New Hydrogen Energy, October 13-18, 1996, Japan, vol.2, p.595-600.
43. L.G. Sapogin, "On One of Energy Generation Mechanisms in Unitary Quantum Theory". Proceedings of the 2nd Russian Conference CNFNT (in Russian) p.18-24, Sochi, September 19-23, 1994; Cold Fusion, No 11, p.10, 1995.
44. L.G. Sapogin, "On One of Energy Generation Mechanisms in Unitary Quantum Theory". Cold Fusion, No 11, p.10, 1995.
45. Correa Paulo and Correa Alexandra XS NRG in Technology, Infinite Energy, vol.2, #7 p.18-38, Nr 8 p.10-15, #9 p.33-37, 1996. US Patents, numbers: 5.416.391, 5.502.354, 5.449.989.
46. Sapogin L.G. Theory of Excess Energy in PAGD Reactor (Correa reactor). In: Proceedings of ICCF-7, Vancouver, April 1998; Infinite Energy, No 20, 1998, p.49.
47. Sapogin L.G. "New Source of Energy?" Journal "Acknowledgement and Physical Reality", Moscow, vol. 2, #1, page 34-40,1997, (in Russian).
48. Swe-Kai Chen, Chu-Yung Liang «Observation of Cell Temperature Drops». In: Proceedings of ICCF-7, Vancouver, April 1998, p.68-72.
49. Jakimenko L.M. Electrolysis of Water. Chimia Press, Moscow, p.p. 33, 86, 90-114, (1970) (in Russian).
50. Pfleiderer N. Electrolysis of Water. pp. 12, 17-18, 1935, Leningrad (in Russian).
51. Rothwell J. Yasunori Takahashi´s Supermagnets, Infinite Energy, vol.1, No 5,6, p.33, 1996.
52. Sapogin L.G. "Is This Really True?", Infinity Energy, N 28, 2000.
53. Sapogin L.G. «Perpetual Mobiles Operating in Switzerland»,magazine «Chudesa I Prikluchenia» # 2, 2000.
54. Sapogin L.G. «They Say There are No Perpetual Mobiles. Then what is it?", magazine "Samolet", 4, 2000.
55. Sapogin L.G. XXI Century - New Sources of Energy? In: Chudesa i Prikliuchenia, Moscow, No 11, p.32-35 (In Russian) 1996, and No 3, 1998.
56. Sapogin L.G.,Kulikov I.V. "Neue Quantenfeldtheorie und prozesse zur electromagnetischer und thermisher energie mit overunity effekt", DVR-Mitglieder-Journal 2/2000.
57. Sapogin L.G. "The 21st Century: Will it Bring a New Quantum Picture of the Universe and New Energy Sources?", Journal of New Energy,vol.2, #3/4,1999.
58. Panos T. "Electrically Induced Nuclear Fusion in the Living Cell ", Journal of New Energy vol.3, #1, 1998.
59. M.Sue Benford, R.N. M.A. "Biological Nuclear Reactions: Empirical Data Describe Unexplained SHC Phenomenon" Journal of New Energy vol.3, #4, 1999.
60. Schwinger J. "Casimir Energy for Dielectric". In: Proceedings of the National Academy of Sciences, vol.87, p.8370-8372, 1990, "Cold Fusion: Does it Have a Future?" Journal "Cold Fusion", vol.1, #1, page 14-17,1994.

KOZYREV-DIRAC EMANATION.

INTERACTION WITH MATTER and METHODS OF DETECTING

Dr. Ivan M. Shakhparonov

125252, Russia, Moscow,Pestchanny Pas. 20-1-33
phone/fax 8·095-198-2012

In this paper the authors show the possibility of creation of a new kind of emanation. The magnetic monopole beam can be made in space as a result of focusing of some natural substance. Special devices based on the Moebius band elements make the given focusing. This emanation is able to magnetize graphite and organics, decrease the radioactivity, and influence the oncology diseases .The time reverse technology is realized in such devices.

Experimental data, which allow making a conclusion about existence of previously unknown emanation, are presented in this report. Here are descriptions of experiments and methods of measurement. The effects of interaction between new type of emanation and matter have been obtained.

Till the present moment theoretical physics didn't pay attention to the nonoriented configurations and spaces. The reason of this situation is the fact, that from the philosophic point of view it is not possible to determine and locate the area of the nonoriented topological structures in our world. We (eight scientific teams) joined our forces and we needed more than 30 years to solve this problem by an experimental approach.

The fundamental tenet of the casual mechanics developed by Kozyrev can be formulated as follows. There are two types of energy in the Universe. The positive or «right» energy acts as a factor of the entropy increase. The negative, or «left» energy tends to decrease the entropy, i.e. it acts as a factor, which regulates the entropy increase. The «right» energy is transformed to the «left» one and this fact may be interpreted as a course of time from the past to the future. When the energy is transformed from the «left» to the «right» form, time is reversed. Kozyrev supposed [1] that through revolving of a body together with a

6. Activation by nuclear magnetic resonance;
7. Activation by electronic paramagnetic resonance;
8. Activation by electrochemical force.

All these methods can be used as possible way to high efficient energy systems. Gerlovin wrote: "Usually 1, 6 and 7-th methods of structural activation are realized in catalysis simultaneously. Besides, catalysis differs from macroscopic methods because it has the most minimal distances from the sources of activator fields to the activated molecules. And finally, an active participation of force fields created by nuclei of atoms and significantly more active participation of disturbed EPV is possible in catalysis. That's why catalysis is the most effective method of structural activation. The detailed account of this method exceeds the limits of this article and we can only annotate it." [1, p.333]

Information stated above is only a small part of the questions appeared under consideration in Gerlovin's theory of fundamental field (TFF). Other important questions should be considered with a new experimental data.

References

1. Ilia L. Gerlovin, "Foundations of unified theory of all interactions in matter", published in 1990, St.Petersburg, Russia.
2. To the question of multipolarity, Alexander V. Frolov, New Energy Technologies, #1 (4) 2002, St.Petersburg, Russia.
3. N. A. Kozyrev, Selected works, St.Petersburg, LGU, 1991.
4. Practical application of time rate control (TRC) theory, Frolov A.V., New Energy Technologies, #3, 2001, p.15, St.Petersburg, Russia.

Antigravitation Force and Antigravitation of Matter. Methods of its Creation

Anatoly K. Gaponov

Sadovaya Str. 195, Novosibirsk, 630009, Russia

Part I

For a long time there is an opinion in physics about antimatter as a possible source of antigravitation, but the researches on this subject came into a dead end. The existent presentations and formulas forbade the conclusion about antigravitation, but our conducted investigations brought us to the possibility to get antigravitation of substance and to the paradoxical conclusions concerning the next:

1. Two types of space exist:
    a) The Absolute space
    b) The Relative space

2.a The Gravitation Field is the relative space, which has accelerated motion, directed to the center of a planet.

2.b The Antigravitation Field is the relative space, which has accelerated motion, directed from the center of a planet.

3. Gravity force does not depend on mass of a body! The mass can be presented in three versions:

a) m_k –mass as amount of atoms.

b) $W_{m_A^e}$ - electronic-atomic energy in mass.

c) $W_{m_G^m}$ – mechano-gravitational energy in mass.

On the basis of the stated notions we offer to revise the essence of force not only in Coulomb's formula, but in Newton's formula too.

$$F = K\frac{q_1 \cdot q_2}{R^2} \qquad F = P\frac{m_1 \cdot m_2}{R^2}$$

It is well known, that mechanical energy can be bringing in electrostatic charge, where mechanical energy runs (turns) into energy of electric field, where

(mechanical energy) $F \cdot R \rightarrow \frac{E_2 \cdot V_{m^3}}{2}$ (electrical energy)

Similarly it is also possible to insert mechanical energy into mass of a body. As the result, the mechanical energy will turn into energy of gravitational field, where

(mechanical energy) $V \cdot F \cdot t \rightarrow \frac{g^2 \cdot V_{m^3}}{2}$ (gravitational energy)

Since the volume of the Earth is constant, the acceleration of gravitational field will be increased.

It should be logical to expect, that when removing the mechanical energy from mass the inverse process will occur, that is to say acceleration reduction of gravitation field will occur.

In his works I. Newton affirmed about existence of two spaces:

The Absolute space - is an immovable non-rotatable space, which represents a limited cube, with our planet in the center.

The Relative space – is a movable space. It can move with acceleration in the absolute space.

Editor's note: In aether conception this means two parts of aether: some part is involved into the motion with the mass, but another part of aether is immovable.

The main mistake in search of aether consisted in the following: Maikelson's experiments were aimed on search of relative velocity between bodies and space. However, it was the **relative acceleration** between bodies and space that was necessary to search for.

To quote the conclusions of I. Newton: “Body can keep the quiescent mode or mode of rectilinear uniform motion ...” By this, he postulates, that the relative linear velocity between solids and space does not exist. But we know that for rotation it exists (the famous experiments with revolving pail of water).

The gravitational field is the accelerated “falling” relative space, which represents a spherical form. If relative space moves, thus the question appears: where does it move? There is only answer: it moves in the absolute cubic space.

In Einstein’s theory there is notion of unified and curved space in gravitational field, but the contradictions appear here, and on concerning that N. Tesla writes: “Only by presence of force field it is possible to explain the observed motion of celestial bodies, but thus the hypothesis of curvature of space is not necessary. The whole scientific literature on this subject is futile and doomed on oblivion”. [1]

The fact that gravitation is the accelerated moving relative space can be proved by observation of accelerated moving rocket, where the acceleration in rocket is equivalent to the acceleration in gravitational field. Accelerated movement of rocket is relatively, that allows speaking about either acceleration of rocket motion in immovable space, or accelerated motion of space in immovable rocket!

The anti-gravitational field is the relative space, which has accelerated motion from the center of a body (for example: rotating cylinder, Earth satellite and etc.) But it is possible to create the model of anti-gravitation without rotations. On the basis of analogy between mechanical and electric energy comes to conclusion that gravity between bodies does not depend on mass of the body, but on mechanical-gravitation energy, contained in this mass, which is possible to contribute or to extract from. **Therefore, this is the internal gravitation energy.**

Part II

The “Mass” can be considered as a measure of three different conditions of matter:

m_k - as a measure of amount of atoms, representing a “framework” or “container”, in which two types of independent energies are concentrated.

$W_{m_A^e}$ - as a measure of electric energy, which can be either accumulated or extracted, and it have a “compressed” form.

An example of accumulation of electric energy in mass is a big cylinder, rotating with linear velocity, close to velocity of light, in this cylinder the mass of electric and magnetic fields of atoms increases. There are another possible ways to contribute and to extract the said energy from mass.

And finally $W_m M_G$ is the mass, which can be a measure of mechanical energy, or it can be either inserted or extracted from the matter (it can be identified as the gravitational mass). This gravitational mass is what we put our attention on, because it affects upon gravitation and it is able to create antigravitation.

“In his time N. Tesla worked on more general problem, which is the problem of matter and energy. And he has found, as he believed, the new physical principle, on the ground of which he brought forth his gravitational theory that was named dynamic gravitation. But he did not tell about it until almost the end of his life”. [2] **Really, dynamic gravitation is the energy of motion.**

Let’s take the following indications:

V – mechanical velocity
F – force
t – time.
In this case the product $W = V \cdot F \cdot t$ has the dimension of energy. Hereinafter, let’s take

I – strength of electric current
U – difference of potentials
t – time.

Then $W' = I \cdot U \cdot T$ **has the dimensional of energy.** Thereby, $W \sim W'$ that is to say, the following products are accepted as equivalent:

1. $V \cdot F \cdot t \sim I \cdot U \cdot T$

2. In previous materials it was reported about untraditional way for accumulation of energy, under the condition, in which at constant current I the product $q = U \cdot t$ will depend on amount of inserted energy in unchangeable circuit L = const, in which the energy can be accumulated by untraditional way not only in electric capacity, but also in inductance.

Similarly the energy can be accumulated by untraditional way in a moving body, under the condition V=const and m_k=const (the product $qr = F \cdot t$ will depend on inserted energy and have unlimited value). **Exactly this charge will create the powerful gravitational fields.**

3. Let’s take: F is mechanical force, R is distance. Then the product $F \cdot R$ has the dimensionality of energy. For the uniform electric field the product $E_a \cdot \frac{E^2 \cdot V}{2}$ also has the dimensionality of energy. In this case E_a is constant, E is intensity of electric field, V is volume. Thereby, $F \cdot R \sim E^2 \cdot V$. Similarly, $V \cdot R \cdot t \sim g^2 \cdot V$

We have received the correlations of resemblance for heterogeneous physical values, on the ground of which the following physical experiments can be offered:

On the grounds of the above-mentioned analogies it can be assumed that the accumulation of "compressed" energy is possible in mechanics, as well as in electricity. Since the velocity is relative, that the mass can have zero velocity relatively a observer, who moves with this mass, but the force field will remain unchangeable, since it depends on already invected mechanical energy.

Let's note that:
When the "compressed" electric energy is accumulated, the power field does not change.

When the "compressed" mechanic energy is accumulated, the power field increases.

Now we have come to the amazing conclusion that the gravitational force does not depend on mass of matter, but it depends on mechanic energy, which is included in this mass. This energy is unstable and at contact with land it is disappearing, and at zero gravity it can be saved for a long time.

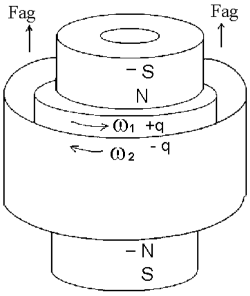

Fig. 1 The first way to obtain the antigravitational force.

1. The magnets are not revolved.
2. The cylinders of charged capacitor are revolving in different directions.

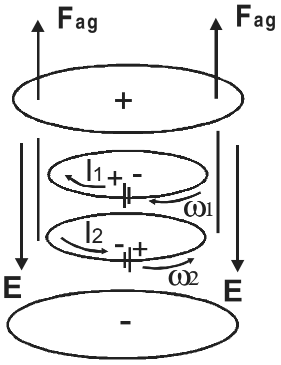

Fig. 2 The second way to obtain the antigravitational force.

1. The capacitor plates are charged and not revolving.
2. The current circuits are revolving in different directions.

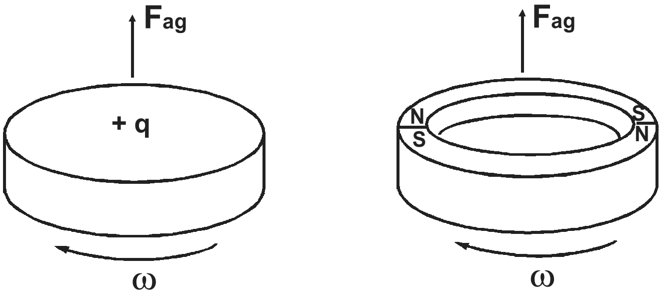

Fig. 3. The third way to obtain the antigravitational force

a) electric
b) magnetic

1. The disk and the ring are made from electrical current conductive material.
2. When these disks rotate, the currents, which emit the mechanic-antigravitational energy in the manner of heat, are formed there.

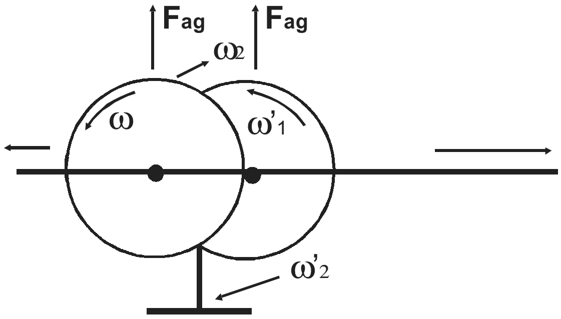

Fig. 4. The fourth way to obtain the antigravitational force.
Mechanical method.

1. This is an extraction of energy from matter. It was reported in details on the 10 international symposium in Volgo-Donsk, Russia.
2. The difference with electric circuits is that it is possible not only to extract the mechanical energy, but also to insert additional energy in the system.

References

1. Magazine Inventor and Rationalizator, Russia, #9, 1979.
2. Magazine Inventor and Rationalizator, Russia, #9, 1979, p.28

The Capacitor, which has the Energy of an Atomic Bomb

(Review of Anatoly K. Gaponov's research by Eugenie and Marina Golomolzins)

Is it possible to place a pail of water into a one-liter jar? At the first look the answer is obvious: certainly not! However, the inventor from Novosibirsk, Anatoly Gaponov thinks differently. He does not "press" water, but electrical energy, placing an energy equivalent of atomic bomb into ordinary electrical capacitor.

Everybody using ordinary batteries knows its defect: they need frequent recharging. Gaponov's capacitor is slightly smaller than a matchbox. Just come home by electrical automobile, take out the capacitor from engine, and then put it into the pocket. For home needs you can just insert the capacitor into plughole to power the light, boiler, and TV system. In general, each electronic device can have its own capacitor, then an electrical wiring is not necessary. After one

On the grounds of the above-mentioned analogies it can be assumed that the accumulation of "compressed" energy is possible in mechanics, as well as in electricity. Since the velocity is relative, that the mass can have zero velocity relatively a observer, who moves with this mass, but the force field will remain unchangeable, since it depends on already invected mechanical energy.

Let's note that:
When the "compressed" electric energy is accumulated, the power field does not change.

When the "compressed" mechanic energy is accumulated, the power field increases.

Now we have come to the amazing conclusion that the gravitational force does not depend on mass of matter, but it depends on mechanic energy, which is included in this mass. This energy is unstable and at contact with land it is disappearing, and at zero gravity it can be saved for a long time.

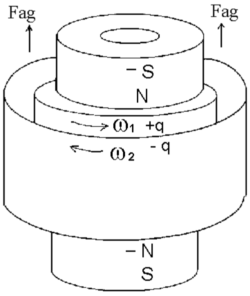

Fig. 1 The first way to obtain the antigravitational force.

1. The magnets are not revolved.
2. The cylinders of charged capacitor are revolving in different directions.

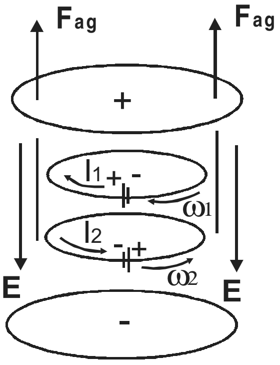

Fig. 2 The second way to obtain the antigravitational force.

1. The capacitor plates are charged and not revolving.
2. The current circuits are revolving in different directions.

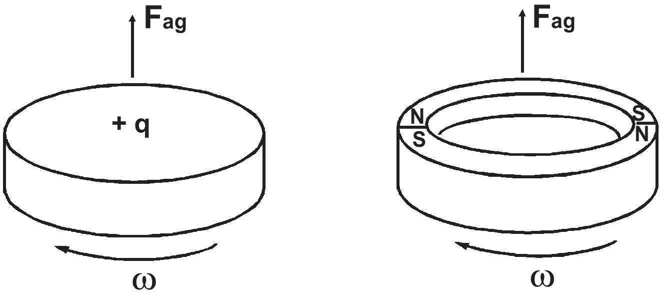

Fig. 3. The third way to obtain the antigravitational force

a) electric
b) magnetic

1. The disk and the ring are made from electrical current conductive material.
2. When these disks rotate, the currents, which emit the mechanic-antigravitational energy in the manner of heat, are formed there.

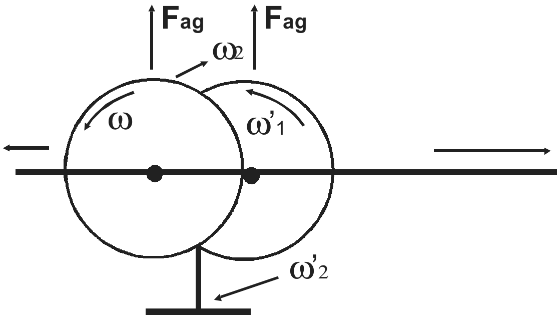

Fig. 4. The fourth way to obtain the antigravitational force.
Mechanical method.

1. This is an extraction of energy from matter. It was reported in details on the 10 international symposium in Volgo-Donsk, Russia.
2. The difference with electric circuits is that it is possible not only to extract the mechanical energy, but also to insert additional energy in the system.

References

1. Magazine Inventor and Rationalizator, Russia, #9, 1979.
2. Magazine Inventor and Rationalizator, Russia, #9, 1979, p.28

The Capacitor, which has the Energy of an Atomic Bomb

(Review of Anatoly K. Gaponov's research by Eugenie and Marina Golomolzins)

Is it possible to place a pail of water into a one-liter jar? At the first look the answer is obvious: certainly not! However, the inventor from Novosibirsk, Anatoly Gaponov thinks differently. He does not "press" water, but electrical energy, placing an energy equivalent of atomic bomb into ordinary electrical capacitor.

Everybody using ordinary batteries knows its defect: they need frequent recharging. Gaponov's capacitor is slightly smaller than a matchbox. Just come home by electrical automobile, take out the capacitor from engine, and then put it into the pocket. For home needs you can just insert the capacitor into plughole to power the light, boiler, and TV system. In general, each electronic device can have its own capacitor, then an electrical wiring is not necessary. After one

or two years you will just have to come into electric service station and charge your magic capacitor like a gas balloon. Meantime, this research work began from hypnosis.

Anatoly Konstantinovich Gaponov (by birth from Kaluzhskaya region) is Ziolkovsky' countryman. In his youth Gaponov was brought by fate into Sakhalin, where he showed hypnotic abilities. As an inquisitive person Gaponov had organized a research group, started experiments and soon he understood that human brain had incredible possibilities.

A mental prick was made distantly to the hypnotized man, and he uttered a cry of pain. The ability to see people through, to define and to avoid organism's faults, was revealed in a hypnotic trance. It was possible to inspire pleasant emotions, to force "watching" a film on the given subject, as if on the screen. An uneducated person became an erudite, as if being connected to a certain global information database. Thus an idea to make an amazing experiment was appeared.

In one of the experiments Gaponov hypnotized the person with four classes education, and asked, if it was possible to transmit the electric current without wires? The hypnotized person gave the answer, that it was possible. For that it is required to convert the electric energy in x-ray radiation. And what afterwards? Afterwards it is required to focus that rays. By what? By the lens made from quartz glass, gold coated. It was a miracle! The person told about things that in usual condition he had no idea of! The information was received from somewhere outside.

Further quite an amazing thing has occurred. Gaponov asked the hypnotized person, if it was possible to intensify the abilities of hypnotist's brain?

He answered, that he could. "He turned me round and stared at the back of my head, - recalls Anatoly. - And suddenly the smile began to tear my mouth. I could not do anything with myself. When my mouth was sprawled literally from ear to ear, the hypnotized person in some inhuman voice declared that experience could not be continue since the cerebral hemorrhage would occur. I was hardly able to give the order to stop the experiment".

Thereby, the experiments with hypnosis gave the beginning to the thirty-years period of inventions in the field of accumulation and transmission of energy. After the return to native Kaluzhskaya region, Gaponov was occupied with physics, development of logical thinking and became the town champion in chess.

The necessary books fell into his hands by themselves: some time a certain acquaintance gave it to read; another time he found the last copy in a bookstore. As a majority of self-taught inventors, Anatoly preferred practical experimentation. In quest of laboratory for realization of his own ideas, he moved to Novosibirsk. **As a result, in 1980 Gaponov has made experimental system for compression of energy.**

From the school Physics we have known the notion of "electric arc" – it is a small blue lightning between two electrodes. Gaponov has tamed this lightning in such a way, that having drawn apart two wires, which executed the role of electrodes, by hands and got the arc by length up to half meter. Anatoly confirms that in principle, it is possible to create an arc of any desired length under any amperage.

One of the experiments found out one more enigmatic characteristic of electric discharge. During electric photography of arc a person happened to be between the camera and the system. On typing pictures, the researchers have found with surprise, that the electric arc was perfectly seen through the person. That is to say, it created the invisible field, for which material object was not an screening obstacle, and which was fixed on the film.

The further experiments with electric arc have allowed to get a new source of energy, as well as to open the possibility of setting light and sound on fire! Just imagine, you ring up a bell, it's sound waves spread at once in all directions, and then flash up with bright blaze.

(Editor's note: this experimental facts are rare modern evidences of possibility to create longitudinal electric waves. It is clear analogy here with sound waves in air since they are longitudinal waves also. Alexander V. Frolov)

When the problem of energy source was solved, Gaponov turned to the problem of energy accumulation. According to Gaponov, he has provedexperimentally the possibility of charging of an ordinary capacitor with any amount of energy. This statement sounds paradoxically: how it is possible to place the unlimited amount of contents in limited volume? However, this is not a simple way.

Gaponov believes that energy "placing" occurs not in space, but in time by means of his system! In what way? Imagine, that you fill one-liter jar with water. But already after an instant the water-filled jar is in past, and that present one is once again ready to be filled. And so ad infinitum. Water as if it fills a certain "time reservoir", and a jar is just a neck of this "time reservoir".

(Editor's note: This method is described in other articles also but usually it is pure mathematical discussion about Minkovsky space-time and theoretical proposals. Gaponov's experiments are realization of fantastical idea to take power from the time flow, i.e. from Past or from Future to get over-unity in Present space. Alexander V. Frolov)

"It is possible to demonstrate one more example, - Anatoly Gaponov adds. – Let's charge the capacitor with the expectation, that it will supply the light bulb

for one second. Thereby, on the Earth this light bulb will be on only for an instant. But if the same capacitor with light bulb is placed in rocket and dispersed around the Earth at the velocity, closed to velocity of light, time on the board of rocket will be so slowed that the light bulb on rocket will be glowing infinitely long for an observer from the Earth. It means, that in any case it is the same energy quantity, but in one case it's action is sprawling for a second, and in another one it is sprawling for eternity! It is possible to say, that in my system I have created the condition corresponding to this hypothetic rocket".

The system for accumulation of electric energy could be charged by ordinary wall plug 220 VAC. Time period of charging is different and depends on the certain scheme of the system. By the way, sea electric slopes are the certain natural analogues of such capacitor. Some elements of internal device of these sea creations reminds the "pump" elements for placing of electric energy into "temporal jar".

Finally, the third Gaponov's invention is the system for transmitting of energy without wires. As well as in two previous cases, there is an experimental device. Anatoly Gaponov speaks that he has succeeded in getting the essence of experiments for transmitting of energy, which were conducted by Tesla.

It is clear, that the main advantage of this method is an absence of wires and losses of electric energy. The electricity could be transmitted directly into any point, where receiving equipment placed, let say from Kaluga to Sahara. However, this is not so interesting for anybody, since for the present day Anatoly Gaponov's inventions don't have demand.

" The first system was created twenty years ago", - says Mr. Gaponov. – "Now I am fifty five, but things have not budged an inch". He adds dreamily: "Eh, if only I had a laboratory and some money...".

Gritskevitch's Hydro-Magnetic Dynamo

Oleg V. Gritskevitch

RUSSIA, 690002, VLADIVOSTOK,
Okeansky prospect, 99 - ap.112
phone/fax: (7-4232) 424-674 Email: ogri@mail.primorye.ru
Russian Academy of Energy and Information,
Russian Academy of Natural Sciences

Editorial: The article presents construction and operation of Oleg V. Gritskevitch's hydro-magnetic dynamo, which is an example of very powerful new energy system. The prototype in Armenia has been produced over 1500 KWtts power during several years.

The author was born on 14 August 1936 and grew up in Vladivostok, Russia. He is married and has a son Boris. Gritskevitch is a physicist by education. He worked in the Far - East branch of the USSR Academy of Sciences. Since 1985 he has been working independently as an inventor. He has more than 70 patents on inventions ranging from household engineering up to high technologies, which he has been trying to apply in our country and met big difficulties. After numerous attempts to receive the patents the author was convinced that outflow of the information occurred. Therefore he has received the state certificates as on know-how (on a French way of patenting), for all his inventions.

Introduction

During the Institute for New Energy 1999 Symposium, I lectured on my hydro-magnetic dynamo. This paper is my attempt to explain the construction and operation of my dynamo.

To fool investigators of my secrets, I have an occasion provided misleading information. For example, the drawing accompanying the Russian patent referenced below shows a cylinder across the toroid to fool readers. The real dynamo only has the toroid without the cylinder. Even its name "hydro-magnetic dynamo" is somewhat deliberately misleading.

I have some familiarity with the new energy field. Nearly all purported new energy devices are fairly small electrical generators. The dynamo may be the only new electrical generator which most nearly meets all the requirements of an ideal large-scaled electrical generator. My dynamo really is the single most valuable invention the world has ever known.

Alexander V. Frolov of St. Petersburg recommended me to contact with Dr. Patrick Bailey, Institute for New Energy since Pat has lots of contacts who could possibly help me with patenting my invention of a new source of energy in USA.

I conducted the work on the theory and creation of the electrostatic generator-converter «Hydro-magnetic dynamo» about 20 years. (See dynamo history below.) The first primitive equipment was created when I worked in Academy of Sciences. During that time various changes were introduced in the generator and in the theory of its work. It is now possible to manufacture, install, and apply it in industry.

For the first time I made the public report on this work in 1991 on a symposium in Volgodonsk city. The report received the positive replies and reviews of the experts of a nuclear industry in USSR. The same year I was accepted in International Nuclear Society. In these years I offered development of this technology to different state bodies and private enterprises. But there was the only answer: "It is very interesting and perspective project, but there is no money for it".

NEW ENERGY TECHNOLOGIES #7

1. Gravitonics is Electronics of the XXI Century, Spartak M. Poliakov, Oleg S. Poliakov
2. Experimental Research on Gravitational Propulsion System. Review
3. Beamship Technology: A Re-working of Early 20th century Discoveries, Russell Anderson
4. Commercial Antigravity, Tim Ventura
5. Action without Reaction. New Gravidynamic Paradox, Yuri N. Ivanov
6. New Possibilities of Vortex Electric Power Devices, Stanislav A. Lisnyak
7. Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels, D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko
8. Microcosm – Universe - Life, Alexey G. Shlienov, Ernest L. Petrov
9. Hypothesis of a Theory of Everything, Jack P. Gibson
10. The Charge and Mass of a Photon, Dzabrail Kh. Baziev
11. The Homopolar Motor: A True Relativistic Engine, Jorge Guala-Valverde, Pedro Mazzoni, Ricardo Achilles
12. Weight Reductions Generated by Bucking-Field Permanent Magnets, William C. Simpson
13. Advanced Nuclear Waste Decontamination Technologies, Mark Porringa
14. Experimental Investigations of the Radioactive Isotope Half-Value Period Changing in the Local Volume ofCause-Effect Relations, Igor A. Melnik
15. Tilley Electric Vehicle
16. The Problem of Time: Force as the Cause of Change in the Course of Time, Valentin P. Oleinik
17. Time and its Physical Relationships, Andrew Michrowski
18. Time as Phenomenon of the Expanding Universe, Michael H. Shulman
19. The Experiments on Thermo-Gravitation. Review
20. Letters to the Editor from Hector D. Perez Torrez

New Energy Technologies

Published by Faraday Lab **Issue #4 (7) 2002**

Research on advanced space propulsion methods and new energy systems

History of invention: Poliakov's vortex drive

Patented by Research Institute of Space Systems named by Krunichev, Russia

Also in this issue:

- * New gravidynamic paradox
- * Experiments on weight reduction
- * New possibilities of vortex energetics
- * Experiments on homopolar motor
- * Physical relationships of time

Gravitonics is Electronics of the XXI Century

Hypothesizes, Conclusions, Speculations

Spartak M. Poliakov,
Oleg S. Poliakov

60-let SSSR str., 1-167, Friazino,
Moscow area, 141120, Russia
Tel: (096) 564-65-67

(This article was published in "ELECTRONICS: Science, Technology, Business" magazine 5/2000 p. 8-13)

Today we can easily insist that ways to solve main gravitonics problems are already defined, at that the practical realization of any of them will mean the break through in engineering. So, what successes has the Russian science already achieved in the area of gravitonics, and what priorities can we lose here in the nearest future?

Introduction to Gravitonics

The electronics of the "past century" uses electron as a ball, which has weight, radius, electric charge and magnetic moment. These very parameters define electron behavior in the electrostatic, magnetostatic and electromagnetic fields. But electron abilities are not limited by it; electron spin and internal microstructure features remain unaccounted and unclaimed. Evenly speaking, General Theory of Relativity (GTR) of Einstein was a power impulse for mathematical physics development and gave birth to many productive ideas. But the main problem, that is the secret of gravitation, remains undisclosed… In works of K.P. Stanjukovich [1] and A.Z. Petrov [2], who carefully followed GTR, it was shown that this theory described neither energy, nor impulse of gravitational radiation, i.e. it can not explain gravitation. About 20 years ago V.B. Braginsky, today's RAS Corresponding Member, came up with an idea: "if the propagation speed of gravitational signal is higher than the velocity of light, there will be already another theory, not GTR!" Maybe, the reason is the postulation of equality of gravitational and electromagnetic radiation velocities?…

Today the approximate theory of gravitational radiation sources can be built on the basis of the following simple considerations: if during annihilation of "electron-positron" pair there creates the pair of gamma-quantums with energy about 0.511 MeV, then the pair of back gamma-quantums with the energy about 0.511 MeV, could create "electron-positron" pair. Is it possible to assume, that electron, positron and gamma-quantum with the energy about 0.511 MeV are just three stages of one and the same object?! If it is possible, then for the rational description of the given object we will have to suppose the existence of subparticles, named by us uniquantums [3], or named by other authors microleptons [4].

On the basis of Heisenberg uncertainty relation conformably to the energy and duration of quantum of electromagnetic radiation, measured by laboratory means, it is possible to calculate the minimal "electrical length" of photon (i.e. quantum geometrical extension in free space in wave-length units), which is equal to 137λ, and in the uniquantum theory it is equal to 137 uniquantum-antiuniquantum pairs. On the basis of these conceptions it is possible to construct the spatial microstructure model of electron. So, what kind is it?

We think, that electron can be represented as thin-walled spheroid, walls of which are two light (C) barriers, separating the "internal" part of electron from the "external" one. From the traditional physics point of view "over-barrier" space is an "imaginary" one. This very space can contain the gravitational mass of electron. The radius of the gravitational spheroid is equal to the half of the classic electron radius, and its imaginary weight is 137 times more than the rest mass of electron. Being "cut" off by the double light C-barrier, uniquantums of the spheroid internal part are as if non-existent for the outer world, and the rest mass of electron is formed by magnetic energy of three uniquantums on the external orbit with the classic electron radius. This very spheroid, rotating with the tangential velocity C, let us get the precise value of the electron spin.

The study of presented model shows, that:

- The "electromagnetic" rest mass of electron is "magnetostatic";
- The gravitational mass of electron is an imaginary value and it is 137 times more than the rest mass of electron;
- The gravitational radius of electron is two times less than the "classical" one;
- The "internal" gravitational radius of electron is 45.7 times more than the external one, i.e. the internal space is compressed per 45.7 times (!);
- The spin is equal to the classical one, but this value is imaginary one (!);
- The value of the "effective" electron charge is three times more than the classical tabulated value;
- The native magnetic field of electron is equal to $8{,}9 \cdot 10^{13}$ Oersted;
- The gravitational constant is equal to 10^{33} $cm^3/g.s^2$, i.e. it is about 10^{40} more than the "world" gravitational constant of the Earth;
- The gravitational energy of electron is equal to $137 \cdot 0.511$ MeV, i.e. 137 times more than the equivalent energy of the rest mass of electron.

The model is paradoxical. But it can be tested experimentally! Comparing "electromagnetic" rest mass of electron with the relation of electromagnetic energy to gravitational one, it is possible to determine the connection between magnetostatic and gravitational energy of electron, and, therefore, with energy of the magnetized ferromagnetic.

Gravitational constants of the Earth and of the electron differ in about 10^{40} and can be described by the same simple equation:

$$\gamma_{loc} = 2k\gamma_0 \cdot \omega^{3/2},$$

where k - is a parameter of the gyroscope shape, γ_0 - is absolute universal constant, equal to 1/137 and ω - is native rotation frequency of the gyroscope.

Let us assume, that gravitational constants of all objects should be described by this equation. By substitution of the new gravitational constant into the known equation of the gravitational energy $W = \gamma m^2 / r$ we will get the equation of the gravitational energy of rotating gyroscope with any size (from electron up to the Galaxy!). Thus, the main point of the "non-Einsteinian" theory of gravitational energy sources comes to the thing that any rotating object and any magnetized ferromagnetic have their own gravitational energy, and the sources of gravitational radiation can be only nonlinearly moved objects, or objects which are in the state of change of phase (for example, permanent magnet during its demagnetization). It is the gravitational theory and explanation of "strong" and "weak" interactions!

Laboratory test of the equations

Magnetostriction

J.P. Joule found the effect of change of ferromagnetic linear sizes and volume during magnetization as early as 1842. Magnetostriction is widely used in modern technique, but in the physical encyclopedia of 1963 there is the following honest acknowledgement: "For the most ferrites both longitudinal and transverse magnetostriction is negative; the reason of it is still unclear."

In the scientific literature magnetostriction is usually defined as $\lambda=\Delta L/L$. However, during the change of external field to some arbitrary and enough small value ΔH, it is advisable to define magnetostriction as $\lambda=1/L\cdot\Delta L/\Delta H$, since in magnetostriction experiments the value $\Delta L/\Delta H$ (or $\partial L / \partial H$) is changed. By means of the suggested equation $W=137(BHV)=BHV/\alpha$, which connects magnetic energy with the gravitational one, it is possible to get enough simple equation for the magnetostriction:

$$\lambda = 1/L \cdot \partial L/\partial H = \alpha \cdot k/(B \cdot H)_s \cdot H^2 \cdot \partial\mu/\partial H$$

where $(B\cdot H)_s/\alpha$ - is the density of gravitational energy in the point of magnetic saturation, k – is the parameter of share of gravitational field in the magnetostriction effect, H - is magnetic bias, $\partial\mu/\partial H$ - is differential magnetic conductivity.

The new equation qualitatively corresponds to four known features of magnetostriction [5], namely:

- The magnetostriction sign is defined by the sign of $\partial\mu/\partial H$, i.e. by the course of the magnetization curve, measured in the direction of calculated component of the linear magnetostriction;
- Graphical sum of three linear components of magnetostriction, calculated by three main axes of the anisotropy form of the model, is always negative and numerically close to the value of the volume magnetostriction;
- Magnetostriction is an even effect, since the equation includes squared value of the external magnetic field;
- Dependence of magnetic conductivity μ from the filed H and hence dependence $\partial\mu/\partial H$ has a hysteresis nature. Therefore, the magnetostriction is a hysteresis phenomenon too.

So we have the right to "close" the question of physical encyclopedia on the cause of magnetostriction. **Magnetostriction is the secondary gravitational effect of ferromagnetic "self-constriction" in its own gravitational field.**

Gravitational-optic effects of GTR

Distortion of the light beam, passing near the Sun and the photon frequency bias in the field of terrestrial gravity (the Nobel experiment of Paunda and Rebki) are the main arguments in favor of GTR canonization. It is very attractive to repeat these experiments in laboratory conditions, basing on our conception of the origin of gravitational field.

The acceleration of gravity, used in experiments with ferromagnetic, reached the value $4.72\cdot10^{15}$cm/s^2, i.e. about $4.8\cdot10^{12}$g. At such values of acceleration there is no necessity to introduce a definition "space masses". In these experiments there was used the optically transparent ferromagnetic, which was the saturated solution of manganese chloride in water at room temperature. The experiment on the beam distortion was made in 1975 [7]. It was shown, that this effect is the result of two simultaneous processes. The first is an intense drift of magnetic ions, which forms the gradient of index coefficient that causes the light beam distortion. Another process is a relatively weak gravitational beam distortion, for which, nevertheless, the relation of deviation angle to the track length (the length of the dish is about 100 mm) is turned out to be about 10^{10} more than in "Einsteinian" gravitational-optic experiments.

The experiment on bias of the optic radiation frequency [8] was made in 1978-1980 and was repeated in 1983. With use of heterodyne and interferometrical methods of measurement we were succeeded to observe effects of "red" and "blue" frequency biases in the non-uniformly magnetized ferromagnetic by means of simple displacement of the working dish (with the length about 40 mm) from one side of the magnet gap to another. The maximum displacement is about 10^{-5}, that is about 10^{10} more than in the experiment of Paunda and Rebki.

Problem of the propagation speed of gravitational radiation

There are still only few publications about such fundamental parameter as the propagation speed of gravitational radiation; it is able to speak only about pages, or even lines! Let us refer to major sources:

I. Newton: *"The propagation speed of gravitational interaction is equal to infinity."* It is an argument, because otherwise we would have to bring the "delay" parameter Δt into the Law of Gravity, what is not noticed in real conditions of star observations [9].

P.S. Laplace in 1787, taking into account observation errors of that time, showed, that gravitational interaction speed was about $50 \cdot 10^6$ times more than the light propagation speed, i.e. it was about $1{,}5 \cdot 10^{18}$ cm/s [10].

A. Einstein: *"The propagation speed (of gravitational interaction) is equal to the light velocity".* This statement is postulated.

Even during the change of propagation speed of gravitational radiation between the Earth and the Moon it is impossible to define the signal delay about 10^{-11} s, i.e. we cannot measure directly the propagation speed of gravitational radiation (supposing that we have both generators of the gravitational radiation and receivers of it). But this speed can be estimated by the reflection impulse, what exactly was made in 1987 [3]. And its value is about $9 \cdot 10^{20}$ cm/s!

On the basis of conservation law of impulse of unidirectional radiator with arbitrary energy type we can get a simple equation:

$$F/(dW/dt) \cong 10^4 \cdot V/C^2 \qquad \text{[g/Wt]}$$

where V – is the speed of radiation propagation, F – is tractive force in grams, dW/dt – is power of radiation in Watts and C – is velocity of light.

For making the experiment there were constructed, produced and adjusted: sensible scales with one degree of freedom (sensitivity of balance is about 1g at the oscillator mass together with the moving element of scales which is about 50kg); the indication system of small mass changes (phase-meter receiver); gyroscopic system, changing the mass in the dynamic mode (there are 16 possible operating modes – from the rotation with steady and variable angular speed up to the forced precession with the variable angle of precession, with the "right" and "left" rotation of all load-bearings elements at option); power sources and commutation automated system. The period from idea up to its realization took about two years (1985-1987) [3]. Taking into account the real parameters of the system, the program of calculation was drawn and propulsive burns were calculated. The results of machine computation can be compared with real impulses, demonstrated on the screen of the oscilloscope.

If strange speed value 177 C^2 is discarded, then the middle speed value is close to C^2, i.e. to $9 \cdot 10^{20}$ cm/s! Of course, we would like to think that this is the second fundamental matter speed of our world, which we has approached experimentally …

Gravitational receiver

During the creation of gravitational antennas and receivers there appear almost insuperable difficulties from the modern fundamental science point of view. That is why it is advisable to look at this problem from another side. At first, it is necessary to consider gravitational radiation interaction not with the mass, which it goes through without losses, but with the gravitational field of independently gravitating mass, when the interaction must be the most effective because of the principle of physical processes reversibility. At second, it is necessary to choose some critical parameter of auto-gravitating receiver as a value, which is directly measurable by gravitational detector. For example, angular velocity of free rotation of thin disk with big diameter, the frequency of magnetization precession during NMR (nuclear-magnetic resonance) or NFMR (non-linear ferromagnetic resonance) etc. can be chosen as such a value.

In 1987 there was the first successful attempt to receive the gravitational impulse. The source of external signal was gyroscopic precessing system with the variable angle of precession (the propagation speed of gravitational radiation was measured by it). Double gyroscope, setting in motion by one electric motor, but with the opposite directions of rotation, was used as a detector. Between disks there was placed the source of light, impulses of which, passing through disks openings, were registered by photodiodes. Their signal came into differential circuit of data processing. The memory oscilloscope reproduced impulses of gravitational radiation. At that radiating system and memory oscilloscope was started up simultaneously. During the work process there appeared a problem of exciting of slow auto-oscillations of gyroscope-detector. This problem together with the low frequency of auto-oscillations of mechanical system led to a conclusion that this research direction is not very promising. However, the fact of detection was proved!

Gravitational engine of continuous action

Only about nine years passed since the appearance of the idea about engine up to its realization! In 1997 the engine was produced and tested. The engine with weight about 28 kg was made "weightless" on the magnetic hanger, and longitudinal draft, appearing in accordance with the impulse conservation law, was measured by micrometer detector of longitudinal shifts (sensitivity is about 50g/point). Such engine could be built still in the beginning of the last century… However, it has a secret **that is a gyroscope with the variable radius, working in the continuous mode.**

The engine power is defined by the formula

$$\frac{dW}{dt} = 5k\gamma_0\omega^{3/2}\left(\frac{m^2}{r^2}\right)\cdot\frac{dr}{dt}$$

In June of 2000 there were made experiments with the model of gravitational engine, which represents a gyroscope with the variable radius (see photo on the 1st cover page). The mercury was used as rotating fluid. Tests were made in the Research Institute of Space Systems named by Krunichev. In three experiments, at a certain speed of rotating fluid there was fixed a decreasing of the engine weight (38,5 kg) up to 1.0-1.5 kg (2-3%). The specific impulse of the engine was equal to 2.5-3.0 kg per kilowatt of electric power. Analysis shows, that the increasing of propulsion force is possible at optimization of design and operating modes.

Some preliminary resume

In the magazine "Foreign Literature" #1, 1967 the article "For hundred years forward…" by Jack Marabini was published. There were made some conclusions about prognostic work of firm Rand Corp., including the area of gravitational technique. Namely:

- Development of communication facilities on gravitational waves in 2000;
- Creation of spaceships with antigravity engines in 2050;
- Transformation of gravitational energy into electric one in 2100.

In the article it was noted, that the most "fantastic" predictions of this firm, as a rule, come true passing ahead.

According to our crude estimations, the propagation speed of gravitational radiation is "C" times as much than the velocity of light, but we know neither laws of attenuation and propagation of gravitational waves, nor laws of their reflection and refraction, nor laws of their interaction with the substance… The large routine work is expected: making of measurements and investigations, tabulating of obtained data, publishing and society familiarization of the results, their "popularization". It is necessary to learn to use gravitational radiation and to protect oneself from its accidental influences, to design standards and dosimeters, etc., i.e. to repeat the way of radio engineering and nuclear physics comprehension.

For that we need generators and receivers of gravitational radiation. It means that the financial support is necessary. And engineers are sure to be ready to pay the highest price for the chance to give to the Mankind spaceships, systems of instantaneous communication with them and real perspectives for the very long history.

Conclusion

We have already passed the long way, if not in space, then in time. We have made:

- Gravitational engines of continuous action with the specific impulse about 2.5 kg/kWt [11];
- Transformers of gravitational energy into thermal and electrical ones [12];
- Communication system based on gravitational waves [13];
- Receivers of gravitational (microlepton) radiation of biological and mineral objects [5];
- Devices for control of "laboratory time" flow (time machine) [14].

The main goals of the authors were to attract readers' attention to the problems, which demand an urgent solution. Some questions were decided, and even seemed to be clear. It also seems to be clear what to do further. And what do you think about it?

References

1. Stanjukovich K.P. Dr. Phys.-Math. Sc. CTR and gravitation // Seminar of the Moscow nature investigators society. // M. – 1966.
2. Petrov A.Z. Some peculiarities of CTR // Seminar of gravitational chair of Kazansky SU. (Preprint / Ukraine Academy of Science) // Kiev – 1971.
3. Poliakov S.M., Poliakov O.S. Introduction into experimental gravitonics. // M.: Prometey – 1991.
4. Patent #2113000 RF. Method to search minerals by their own radiation, device for its realization and microlepton detector / / Ohatrin A.F., Ohatrin A.A., Sizov V.S. Priority from 21.07.1997.
5. Kirensky L.V. Magnetism. // M. Nauka – 1967, p.141.
6. Vonsovsky S.V. Magnetism // M. Nauka – 1971, p.404.
7. Poliakov S.M., Martynov V.F. Method of deflection and focusing of optic radiation. A.n. #2187534, 10.11.1975.
8. Poliakov S. M., Olihov I. M. Dorofeev V. A. and others. Displacement of optic radiation frequency in non-uniformly magnetized ferromagnetic // Proceedings of the Moldova Academy of Science. Series of Physicotechnical and mathematical sciences. // Kishinev – 1983, #2, p. 57-59.
9. Isaac Newton. Mathematical beginning of Natural Philosophy // Transl. Moscow-Leningrad – 1936.
10. Laplace P. S. Statement of the world system. // Leningrad, Nauka, vol. 4, p. 197-198.
11. Poliakov S.M., Poliakov O.S. The beginnings of experimental gravitonics // In: Proc. of Int. Conference "New Ideas in Natural Science" // S.P. – 1996, p. 529-536.
12. Potapov Ju.S., Fominsky L.P. Eddy energy // Cherkassy, Kishinev OKO-Plus – 2000, p. 387.
13. Akimov A.E, Shipov G.I. Torsion fields and their experimental manifestations // Proc. of Int. Conference "New Ideas in Natural Science" // S.P. – 1996, p.22-225.
14. Chernobrov V.A. Secrets of time // M., AST – 1999.

● ● ● ● ● ●

About the authors:

Spartak M. Poliakov. Graduated from Kishinev State University. Profession is physicist-experimenter. He works in electronic industry for about 47 years. S. M. Poliakov is the author of more than 50 scientific works. One of his latest books is "Introduction into experimental gravitonics". Interests: microwave engineering, gravitational electronics, faster-than-light communication transformation of gravitational energy into electric one.

Oleg S. Poliakov. Graduated from Moscow Institute of Electronics. Profession is "Semi-conductor electronics". Interests: computer engineering, industrial gravitonics. He is a co-author of "Introduction into experimental gravitonics" and "Self-tutorial of computer work".

BEAMSHIP TECHNOLOGY

by Russell Anderson
russanderson3@hotmail.com

Read in this issue a report by Russell Anderson, USA, on beamship Model Flying Craft

www.americanantigravity.com

Publisher Faraday Lab Ltd
Editor-in-Chief Alexander V. Frolov, Editor Olga O. Leontyeva,
Designer Ekaterina F. Yakovleva
Scientific Advisory Board Kyrill P. Butusov

New Energy Technologies Magazine, is published by Faraday Lab Ltd, six times a year. Correspondence Address:
POBox 37, St.Petersburg, 193024, Russia
Phone: 7-812-380-6564, net@faraday.ru, www.faraday.ru

Subscription: $90 any foreign. Air mail delivery included
PRINTED IN RUSSIA **Copyright © 2002 by Faraday Lab Ltd.**
All rights reserved.

Experimental Research on Gravitational Propulsion System

*Editor's: It is a review of the article by **V. A. Menchikov**, the Director of Research Institute of Space Systems, named by Krunichev, Russia. The article was published in"Polyet" magazine #10,2001, p.38-39, Russia. It scrutinizes the matters on development of propulsion systems based on the unconventional approach to the problem of gravity, i.e. gravitational engines. It also cites the results of the gravitational engine model research made by means of the experimental facility, created in the Khrunichev Research Institute of Space Systems.*

The device, transforming rotary motion into unidirectional motion, looks like S.M. Poliakov's one. It also operates with rotation of liquid, which causes the propulsive force. Truly speaking, Poliakov had an agreement with Research Institute of Space Systems named by Khrunichev in 2001. Some funds were assigned to develop the device, however the project, into which Poliakov had put a lot of work, still remains unrealized. Besides, the scientist's name is not even mentioned in the patent.

***

Scientific and technological advance opens to mankind more and more wide abilities to use space for the solution of global problems. In many respects the complete realization of these abilities will be defined by the developement of means used for delivery of payloads into the space. In the XXI century the dominating use of reactive chemical and electrical propulsion systems in rocket-space technique as well as low application level of engines with other physical principles can be the factor of an "inhibitory" influence for the development of such techniques. It is caused by the fact that created rockets practically do not correspond to noticeably increased standards of safety, operating costs, costs for transport operations execution and ecological influence on the environment.

Thus, there becomes to be urgent the problem of development of alternative approaches towards the creation of propulsion systems, made for the rocket-space technique on the base of unconventional ideas and engineering solutions. A rather old-established idea of creation of gravitational engine should be concerned as one of such ideas. It is based on the unconventional approach to the problem of gravity. Nowadays many countries take part in solution of the gravitational problem, namely Russia, USA, Japan, etc., and if till recently only some scientists and inventors showed the interest to this problem, then now it arouses interest of research-and-production majors. Unfortunately, now it is not possible to speak about sufficient theoretical or practical development of this idea. However, the interest is so considerable, that practically separate experiments on this subject were made earlier and they are still made nowadays. After all, stakes are very high and are defined by applied nature of the problem (the ability to create qualitatively new engines for the rocket-space technique), as well as by its scientific significance.

One of the directions to solve the problem of the creation of gravitational propulsion systems is the realization of associated theoretical and experimental methods of the search of physical processes, leading to the antigravity effects appearance, which cannot be adequately described by existed theoretical conceptions. V. Shauberger's patent, based on the postulate of gravitational energy radiation by "disturbed rotating mass" can be considered as an example of such practical realization. Taking into account a number of known experimental results, a model of gravitational engine and experimental system for estimation of this model parameters were made in Research Institute of Space Systems named by Khrunichev to provide the practical realization of Shauberger idea (Fig.1). It is the metal construction, which provides the model displacement in upward direction with the ability of its rotation around vertical axis.

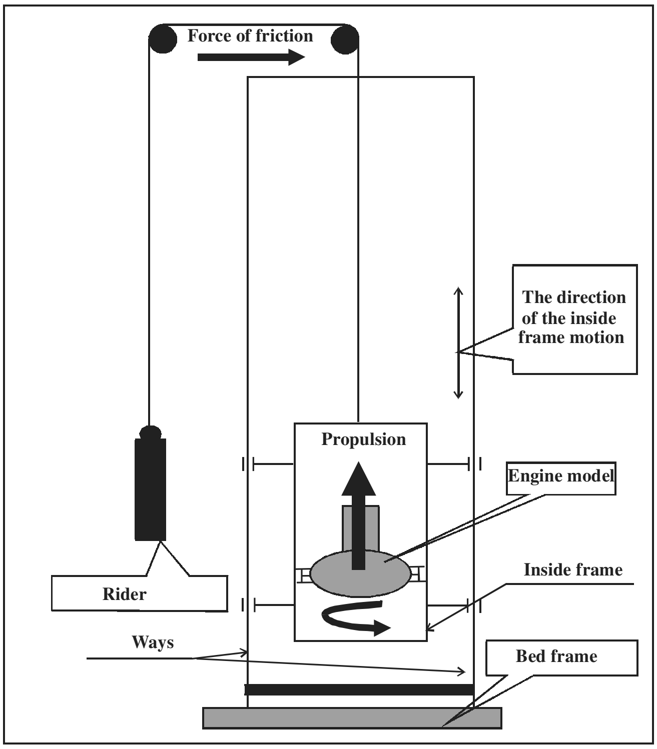

Fig.1
Schematic circuit of the experimental system

Later, to increase the system sensitivity it was improved and the block system of suspension was replaced by the lever frame.

Turning angle of the stand frame, where the model of gravitational engine is suspended, depends on the following: weight and geometry characteristics of the frame; weight characteristics of the engine (of counterweight); engine propulsion and frictional forces in bearings. Laser indicator of frame turning angle and vertical ruler let increase the gauge of lifting height of the gravitational engine model proportionally to the arm of light beam.

On using this system there was made a wide cycle of tests, which qualitatively confirmed the presence of propulsion force. Tests were recorded by video. The analysis of experimental results shows, that acting time of propulsion was about 12 s at each switching. During repeated switching of the gravitational engine model in different conditions it is able to create the propulsion, the value of which at 40…50 s of operation can be about 3 standard units of propulsion force (1 standard unit of thrust is about 10gs), and while using the powerful electromotor it can be about 80 standard units of propulsion at the intervals up to 4 s.

Research Institute of Space Systems named by Khrunichev, works on automation of experimental researches and on development of laboratory resources for factor analysis of appearance of the propulsion vector with the usage of the described model of gravitational engine.

Patent

The patent was published in the Bulletin of Patent Information in 2001.
(11) 20946
(51) 7 F 16 H 27/00
(21) 2001121237/20
(24) 01.08.2001

Editorial: V.A. Menchikov together with A.F. Akimov, A.A. Kachegan and V.A. Svetlichnyi have got this patent. Dr. Spartak M. Poliakov, being the author of the principle, is not mentioned at all in the patent.

NEWS REVIEW

Boeing Tries to Defy Gravity

According to Jane's Defence Weekly (UK), http://www.janes.com, Boeing, the world's largest aircraft manufacturer, has admitted it is working on experimental antigravity projects. These projects are able to overturn a century of conventional aerospace propulsion technology and alter the entire aerospace business. Boeing uses researchers by Yevgeny Podkletnov, who claims to have developed a device, which can shield objects from the Earth's gravity. Many conventional scientists, who have not been able to reproduce Dr Podkletnov's results, view his project, named «GRASP» (Gravity Research for Advanced Space Propulsion) with suspicion.

Dr Podkletnov claims to have countered the effects of gravity in an experiment at the Tampere University of Technology in Finland in 1992. The scientist says he found that objects above a superconducting ceramic disc rotating over powerful electromagnets lost weight. The researches have shown that the reduction in gravity was small, about 2%, but the implications - for example, in terms of cutting the energy needed for a plane to fly - were immense.

His devise, named "impulse gravity generator" is capable to produce a beam of "gravity-like" energy that can exert an instantaneous force of 1,000g on any object – enough, in principle, to vaporize it, especially if the object is moving at high speed. Laboratory installation has already demonstrated the 4in (10cm) wide beam's ability to repel objects a kilometer away and that it exhibits negligible power loss at distances of up to 200km.

Applications of the device can include space launch systems, artificial gravity on spacecraft, aircraft propulsion and "fuel-less" electricity generation ("free energy"). However, observers say that Podkletnov's device could be engineered into a radical new weapon, for example, adapted for use as an anti-satellite weapon or a ballistic missile shield.

Documents, obtained by reliable sources, show that Boeing is taking Dr Podkletnov's research seriously. It is also possible, Boeing admits, that "classified activities in gravity modification may exist". The paper points out that Podkletnov is strongly antimilitary and will only provide assistance if the research is carried out in the "white world" of open development.

Boeing is the latest in a series of high-profile institutions trying to replicate Dr Podkletnov's experiment. The military wing of the UK hi-tech group BAE Systems is working on an anti-gravity programme, dubbed Project Greenglow. The US space agency, Nasa, is also attempting to reproduce Dr Podkletnov's findings, but a preliminary report indicates the effect does not exist.

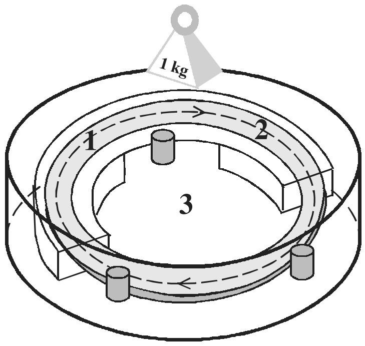

1. Solenoids create magnetic field
2. Spinning, super-conducting ceramic ring
3. Liquid Nitrogen acts as coolant

BEAMSHIP TECHNOLOGY

by Russell Anderson
russanderson3@hotmail.com

Read in this issue a report by Russell Anderson, USA, on beamship Model Flying Craft

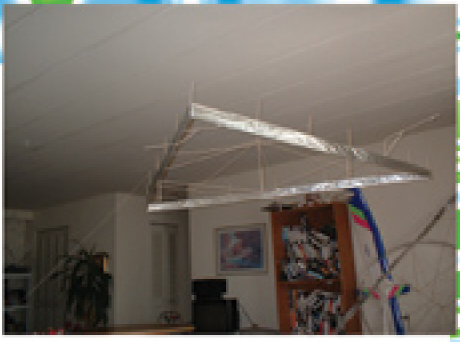

www.americanantigravity.com

Publisher Faraday Lab Ltd
Editor-in-Chief Alexander V. Frolov, Editor Olga O. Leontyeva,
Designer Ekaterina F. Yakovleva
Scientific Advisory Board Kyrill P. Butusov

New Energy Technologies Magazine, is published by Faraday Lab Ltd, six times a year. Correspondence Address:
POBox 37, St.Petersburg, 193024, Russia
Phone: 7-812-380-6564, net@faraday.ru, www.faraday.ru

Subscription: $90 any foreign. Air mail delivery included
PRINTED IN RUSSIA **Copyright © 2002 by Faraday Lab Ltd. All rights reserved.**

Beamship Technology: a Re-working of Early 20th century Discoveries

Russell Anderson

Applied Electrogravitics 377 Avon Rd., Ste. D-119
Devon, PA 19333 russanderson3@hotmail.com
484-255-1072

Some Basic Background

The concept of an electric aero-spacecraft with no moving parts was initiated by the Yugoslavian electrical wizard **Nikola Tesla**, who lit the entire world 100 years ago, at the turn of another century, with his revolutionary AC electric current. In 1916-17, Dr. Francis Niepher performed meticulous mass-deflection experiments under rigorous scientific conditions with lead spheres suspended by wires with shielded and unshielded containers. An accounting of this important series of experiments is in TRANSACTIONS OF THE ACADEMY OF SCIENCE OF ST. LOUIS VOL.23, 1916 and 1917. Related article is in THE ELECTRICAL EXPERIMENTER, March 1918.

Before 1905, **George S. Piggot** was routinely suspending small silver balls to water globules, corks, wood, using the electrostatic field from a specially designed Wimshurst machine in a glass container under several atmospheres of pressure to raise the current level. Output voltage was typically 500KV. The field was propagated by a charged sphere. A small curved conducting plate on the floor acted as a ground. He observed unusual patterns of blue dots with filaments over the suspended objects, sometimes with an anomalous 1/2 cm "dark band" on the suspended objects. Piggot states, "It is my firm conviction that that somewhere on the outer confines of our planet there exists a similar contracting belt thru which naught but the gravitational vibrations of the sun penetrate, and these vibrations absolutely annihilate or absorb all other less powerful ones". If the force was Coulombic in nature, objects would be first attracted, and then strongly repelled by the charged metal sphere. After the objects were suspended, Piggot found he could remove the conducting ground plate, and the objects still floated, suspended. The phenomenon of levitation was accompanied by "luminous halos".

In 1925-27, **Albert Einstein** released his scientific "gem", his "zur Einheitlichten Feldtherie", or the Unified Field Theory for Gravitation and Electricity, to the press and the scientific community. It combines electricity, magnetism, and gravitation into a single mathematical expression, showing how High-Voltage/Low Current electricity (Electrogravity) -and conversely Low-Voltage/High-current (magnetogravity) "acceleration-fields" (G-field) could be produced using then-available relatively LOW-technology. Indeed, a very simple technology. The unifying field is the electrical field (because it can produce gravitation and repulsion fields, as well as magnetism). His Crowning work was released with much press write-ups and fanfare, then it was quickly forgotten as if the scientific community and the world had suffered some kind of collective amnesia!

All of the readers of this magazine need no introduction to the pioneering work of American Scientist **Thomas Townsend Brown**, who was playing around with an X-ray tube around the same year as Einstein's Unified Field Theory was released. He filed his first patent for this newly and accidentally discovered "electrogravitational-effect" which causes motion in a high-voltage condensor or capacitor configuration. He was only 17 at that time. The discovery that high-voltage/low amperage electrostatic potentials applied to an object causes motion in the direction of the positive pole, and electrical charges naturally move to the OUTER surface of an enclosed charge-conductor, held strong prospects for what Brown would later name the "space-car", and wrote an article "HOW I CONTROL GRAVITATION". His pioneering work, and demonstration of devices in Hawaii during World War II, drew attention from the department of Naval Intelligence. He was invited to work on "Project-Rainbow" (the Philadelphia Experiment for Electromagnetic Stealth) because of his pioneering work on what was starting to be understood very covertly as a true WARP DRIVE. Experiments with certain new and classified arc-welding apparatus at the Philadelphia Navy Yards to weld armor-plate for battleships was (by use of banks of primitive but powerful avalanche-discharge capacitors) producing anomalous and unexplained effects, such as disappearing tools and other apparatus in the heavily shielded welding chamber. These strange effects were accompanied by a strange "blackout -zone" which, like Piggot's early work, was not optical in nature. TT Brown's devices in his AH Bahnson Labs home movies lift more than their own weight and move inside vacuum chambers in these films. TT Brown later founded NICAP in 1956, which became the most respected UFO data gathering and hard scientific organization in the world, besides the US department of Naval Intelligence itself, and the Foreign Technology Division at Wright-Patterson Air Force base in Ohio.

My background and work

I primarily have a background and degree in computer programming, electronics, most fields of science, Flying Saucer Technology research (almost 30 years worth), Radio/Control fixed and rotary-wing aircraft since 1972. I have been experimenting and working with high-energy and electrogravitic devices and systems since 1987. I built my first small High-Voltage generators starting around this time. I built kits from Information unlimited and elsewhere.

In late January 1990 I built my first working 2-foot flying discs, which were a direct replication of Thomas Townsend Brown's most important representation of his electrogravity-propelled scale-model vehicular concept, from US Patent #2,949,550. In January 1992, I built a 120KV high-voltage/low-current electrostatic generator from an Information Unlimited kit, primarily for force-field propulsion research.

In June 1999, I built a tower and rotor apparatus to complete the experiment, and I powered it with the output from a 100KV generator I built from an Information Unlimited Kit. The results were spectacular, and taught me a great deal about what was involved in producing and maximizing the Biefeld-Brown Electro-gravitational effect. **I suspected from my research, and my experiments, that the basic effect was not due to current-flow and resulting ion-wind. When there was current-flow, the effect is attenuated, power consumption goes up, and thrust goes DOWN.** In late June of 2000, I presented this working TT Brown Electrokinetic Apparatus with larger 1-meter discs at the 2nd Antigravity Conference in Reno, NV, hosted by Jim Cox. A VHS videotape of this working and spectacular presentation at the first part of the conference is available from www.soundphotosynthesis.com

Now that I had mastered producing horizontal thrust, vertical thrust, or antigravity, was the next goal. About this same time, there was buzz all over the Internet about claims of two or three individuals who sounded credible at the time who had successfully replicated the many multi-layered "gravity-warp capacitor" or "electric rocket". Oddly, these claims could not be verified, and the individuals making the claims disappeared back into the woodwork. Such actions are bizarre and hinder the progress of true science, which is undergoing a shift in paradigms right now, if not a change in dogma.

I spent months die-cutting hundreds to thousands of tinfoil and aluminum-foil circular-notched conductor plates and wax paper and mylar rings. I did some initial testing with a Tin and wax-paper 400-layer gravity warp-capacitor heap, according to plans I had acquired from H & A Industries in 1992, and what was on Bill Beatty's amateur science site. No one else has come forward with positive results on this tedious and time-consuming device. So much work for so little effect! And, if you short out the stack with too much power, you must tediously and laboriously search thru hundreds of layers to find the dielectric layers with the telltale carbonized holes. The Electric rocket has been recently successfully replicated and tested in hard vacuum and patented recently by Hector Serrano. The Serrano effect is identical to the so-called Biefeld-Brown Electrogravitational effect. They are one in the same thing. I may dust off my completed 400-layer grav-cap, but I hardly find it worth the time and effort, because of my recent work starting in early October 2001.

The Lifter and The Evolution to Beamship Model Flying Craft

Although I had attempted a few small "Hagen" patent-type antigravity (VTOL) models in the early 90s, I found their performance poor at best and their power consumption high. In late summer, 2001, someone, I forget who, on the JLN's lab list of researchers and anomalous science-experiment and technology enthusiasts ran across a website owned by Transdimensional Technologies, of Huntsville, Alabama (famous for NASA research facilities, the late Dr. Rolf Schaffranke, author of the important ETHER TECHNOLOGY, under the pseudonym "Rho Sigma", and Dr. Tom Bearden) had produced a hovering device. From my previous work, I recognized it immediately as TT Brown's Electrokinetic Apparatus that I had successfully replicated and demonstrated before a live audience years earlier. I noticed the capacitors were made from Aluminum FOIL, not the thin-but-heavy Aluminum sheet stock from Home Depot that I had been using for years, (I had assumed that to make my 3-foot discs hover and ascend vertically, I would have to use voltages in the hundreds of kilovolt range, and generate high x-ray, UV, and possibly gamma-ray emissions as a by-product, in other words, a typical flying saucer with all the associated radiological effects that have been documented for over half a century) so they could lift their own weight. The result matches almost exactly the simplest graphical representations of TT Brown's patent from 1960, and De Seversky's Ionocraft patent from 1964, which was a thin foil cathode plate with a thin anode wire separated from the cathode by stand-off insulator posts. I was eager to reproduce these devices (I don't know how I overlooked this simple solution, it was all sitting in those old 1960s Brown and De Seversky patents I have studied for 15 years previously) and many people around the world, especially the webmaster of the JLN Lab's site French researcher Jean-Louis Naudin, who began replicating many different types of larger and more sophisticated devices, some of which resembled model spacecraft, and began amassing tables of very useful data, that researchers could use as basic guidelines to follow. I replicated the first hovering device, the "Lifter" (so-called by Transdimensional Technologies) as a 1-foot triangle, with 2-inch foil cathode and #42 enameled copper magnet wire. To energize it I used a commercial power supply from Gamma High-Voltage Research that I had acquired from Ebay some years ago. It was perfect for antigravity research, having full metering, and variable voltage from 0 to 40KV, and current limiting from 0 to 1.5 milli Amperes of current. The heavy 1-meter discs of the Electrokinetic Apparatus were too heavy and the rotor-friction to great for this low-powered device (60 Watts, maximum), but for the lifter, it proved ideal.

My first "lifter" antigravity device worked, but its performance was less than ideal. It had to be stripped of its lower balsa-wood frame and some of its foil before it would degravitate (counterbary), and it "maxed-out" the current-limited power supply at 33KV I 1.5mA, for a stable hover (actually this is an upward flight configuration, because the device is tethered to the lab table with 3 sewing threads). That is 49 Watts. The concept of a hovering TT Brown Electrokinetic Apparatus had been proven to my satisfaction, however, and I initiated more research into past works and patents to raise efficiency to workable levels. The performance was slightly better than my early 1990s "wire-grid" type devices. I found this slightly encouraging.

After a couple months reading and research (why reinvent the wheel, its all been done before), I started to replicate larger models in February and March 2002, but kept coming up against a size-barrier with the Multicellular (grid) approach that many researchers had assumed would raise thrust, and efficiency. This approach obviously did neither, as no one seemed to be able to produce hovering devices above a certain size, the current consumed (adding to total wattage consumed) was prohibitive with the low-powered (still high-voltage, low-current) devices that most of the mostly amateur researchers were using. Researchers around the world started to replicate different versions of the basic lifter 1 (an 6 to 12-inch equilateral triangle). The lifters are always tethered to the testing surface with 3 strings to keep them from going dangerously unstable and possibly short-circuit when they reach the limit of the umbilical supplying power to the device.

From my previous Biefeld-Brown effect replications years earlier, and from carefully reading Brown's EK Apparatus patent, I knew that increasing the diameter of the wire would reduce leakage current created by coronal discharge, mostly coming from the forward electrode, which in the 2 and 3-foot saucers consisted of an arc of copper tubing in the front quadrant of the saucer, or disc. Corona robs power (amperage) from the disc that otherwise would be used to "propel" the disc. Increasing the diameter of the copper tubing, as per Brown's patent if the effect was due primarily to ion-wind, more current and current flow between the electrodes would be desired to effect more air movement. But this is not what I saw in the saucers. **There was apparently another, far more powerful but subtle force effecting silent propulsion of the saucers that had nothing to do with charge-transfer and ion-momentum.**

In February of this year, I undertook an effort to replicate and improve performance and reduce power consumption of the lifter device, based on data from my electrogravitic work of years past. I started by using thicker diameter enameled copper magnet wire, #35 to #30 diameters. I first built a 1-foot equilateral triangular basic "Lifter-1", weighing only 3.5 grams. On March 16th, I built a lifter with the thicker #35 enameled copper wire.

I made the three sides 1-foot long and exactly 2-inches high. After experimentation, I found the optimum spark gap for my High-Voltage power supply (Gamma High Voltage Research 40KV with current limiting to 1.5mA). The small silver-colored device leapt off the test table and pulled violently against its anchor strings to a distance of about a foot. This seemed like a great deal of force for such low power. The large discs of my TT Brown EK apparatus required a good deal higher voltage to initiate motion in the direction of the anode. The device consumed 26KV I 0.56mA DC, which calculates out to 14.56 Watts. I was getting more excited, because this was the best efficiency seen of any result yet posted.

On March 19th, I tested 2 lifters glued together in a "diamond" shaped configuration.

This 2-foot device weighed 6.0 grams, with the same #35 wire and a 2 and 5/8" air gap. It took 25KV to nullify the weight of the device, and it achieved a stable hover at 35KV I 0.8mA. That is 28 Watts. This is about what I had initially expected, double the power for double the Watts. Still, this was far less overall power going into the device to achieve a stable hover than my first primitive and radically shorn and trimmed device. After lift off to the extent of the anchors, I found I could reduce power slightly and maintain a stable hover. On march 22nd at 3:49 pm I got the diamond lifter to achieve a stable hover with a 2 and 5/8" air gap at 29.5 KV I 0.32mA. This was only 9 Watts! This was unheard-of efficiency. I was further encouraged to build and test larger hovering devices to see how large I could get them with my low-powered commercial power supply.

I then built a "lifter-2", which consists of three 1-foot triangular capacitor cells taped together. It weighs 11.4 grams. March 30th at 3:22 pm, the device achieved a stable hover at 38KV I 0.57mA for 19.76 Watts total power. The larger device was more energy efficient than a device 1/3 the size. I wanted to see how far this could go, so I added three more lifter cells to make a 6-cell device, 3 feet on each of its three sides. I was eager to check the performance of this fairly large device. This was the diameter of my horizontally propelled TT Brown Discs.

This device weighed 21.6 grams. I kept the spark gap the same distance on this device. However this device failed to achieve counterbary (lift). It just sat on the test table, filling the air with the smell of ozone and making a sizzling sound (corona noise). I noticed that the current maxed-out on the power supply at a fairly low voltage and would not go any higher.

I concluded in my disappointment that all that wire from all the inter-connections to the cells was causing corona leakage and robbing current, which otherwise would be used by the device for propulsion.

The idea then hit me that perhaps I could make a device with the same outer diameter as the 3-foot device, but have greater efficiency because of a much shorter length of wire. I built basically a 3-foot (1-meter) version of the first 1-foot device. This device weighed 16 grams. It lifted off the table with amazing force and hovered stably with I 52mA 30KV which is 15.6 Watts. Not only had I achieved a larger-size device, but far better power efficiency for a much larger and heavier device. I was overjoyed! I saw that I had a great deal of lifting force to spare. I had not even come near the limit of my power supply. I added extra bracing at the corners and extra balsa and a triangular paper "payload -tray" in the center of the device, supported by three 1/16"x1/16-inch balsa stock. The extra bracing and payload area added approximately 2 more grams. With a 5 gram payload, the device consumed 39.9 KV I 0.99mA, for total power consumption of 39.5 Watts. I was really encouraged at that point, because I knew that these results were unheard-of, in terms of energy efficiency. I had solved the problem of decreasing efficiency by dispensing with a "grid-based" device. Increasing the area of the capacitor plate was one of the factors that increased performance and efficiency, lessening the input power requirements with increasing size. Now the Biefeld-Brown effect could be properly studied, now that most of the ion-flux had been eliminated, resulting conservation of energy by the device, and resulting in greatly increased propulsive force.

Since I now knew the limit of payload for the device at the power level I was using, I added a balsa framework that approximated a central cabin area, and three small styro-foam spheres on the center of the straight sides on short lengths of balsa. The device no longer looked like a test device, but now looked like a scale model spacecraft. I remembered the Edouard "Billy" Meier UFO contact case, and knew that all his original photographs and movie footage of extraterrestrial spaceships the extraterrestrials themselves called "beamships" (there are several styles and variations, all with different specific functions and capabilities, some manned, some remote-controlled "telemeter discs" that had a tri-hemispherical undercarriage that I knew from past research were propulsion condensors) and that the original un-tampered photos all passed rigorous analysis using the latest and most sophisticated computer and other equipment, case detractors not withstanding. Also the spiritual messages of these genetic brothers of Man and their accounting of humankind's history and origins from far across space rang true and struck a chord with me.

I decided to name this new 1-meter model spacecraft Beamship Variation I. The three sides of the device performed the same function on this device that the three spherical or hemispherical capacitors often seen on the underside of full-size "beamships" (Daylight-disc-type UFOs), which illustrates a similar if not identical propulsion methodology to full-size 3 to 7 meter and larger "off-Earth-built" aero-spacecraft. Clearly the propulsion methodologies were exactly the same in the model as in the full-size flight device.

I immediately built a 4-foot diameter model with a full cabin framework and internal payload area and achieved even greater performance and efficiency. I was ecstatic. This 4-foot device I dubbed Beamship Variation II. I received a suggestion from Mr. Tim Ventura of American Antigravity that performance could be increased by using small diameter stainless-steel wire. It seemed unlikely to me that smaller diameter wire would increase performance, it contradicted Brown's patent, and my own past research with large electrogravitic discs. But Stainless steel has a high number of free electrons in the outer valence atomic shells (electron orbits).

So I obtained some #40 stainless locally and the results confirmed Mr. Ventura's suggestion. Corona noise was heard at a much higher power level, and was greatly attenuated in volume. Leakage current was less, and the two Beamships now had more thrust with less power input. They even carried more payload at less power input. Variation II weighs 21 grams and loft a payload of 6 grams at 40KV I 1mA for 40 Watts total power. Again, this was unheard-of efficiency. The anode wires sang a strange harmony as the Beamships floated in the air, stably at any altitude, from floor to ceiling, without any fuel or visible means of support. This was very Beamship-like.

I thought that now since corona discharge on the anode wire was less, I could decrease the spark-gap distance without creating a spark (which kills lift). Thrust seemed initially to increase, but efficiency went down because there was current-flow now, and current consumption went way up. The supply would now max-out at 37KV I 1.5mA and would not increase because of the current limiting. The Beamships now were noisy, as the foils chattered loudly because of all the ion-wind that was now rushing downward along and past the foils. I used a concert fog machine to observe the ion-flux vector, and filmed it digitally with my Logitech web-cam, and with VHS analog video.

Analysis of the fog -tests showed a circular vortex of air surrounding the anode wire that flared out into a downwash of air below the Beamship. I was disappointed, because I thought then that the thrust action of the device was due to simple ion-transfer. A useful-enough effect, but of questionable use in the vacuum of space without an ionizing medium.

Beamship Variation III

I reasoned I had just about enough power in the supply to build and fly a 6-foot (2-meters) Beamship. Since the balsa came in 3-foot lengths, this was simple. As with all the lifter devices and the more evolved and efficient Beamship-series model aircraft, construction techniques are extremely simple and require little skill to assemble. Weight of the Beamship Variation III is 42 grams, with 6-feet of length on its 3 sides. Height of the foils was still 2". Full frame and cabin, with Searl "IGV"-type landing legs, to support the weight of this heavy and very large device. I set the spark gap at 2 and 1/2 inches. At 12:15pm EDT, May 12th, 2002 the Beamship was weightless at 32KV with current maxed-out at 1.5mA.

The device barely lifted off, and "hopped", across the floor once or twice at full power. It had the same loud rattling of the foils due to the terrific downwash of electrified air. I needed to raise the power level. I increased the distance of the spark gap to 2 and 3/4". Now the Beamship took off straight up with power to spare, as if it was one of the smaller craft. Beamship Variation III is weightless at 30KV I 0.85mA (25.5 Watts), and airborne into a stable hover at 35KV I 1.35mA. That is only 47.25 Watts. It can carry a payload of 5 grams, or 5 grams worth of additional framework and structure, to the limit of the power supply, which is 60 Watts (40KV I 1.5mA).

The 42-gram, six-foot model aero-spacecraft only consumes 47.25 Watts at hover, but my first small and trimmed device ate up 49 Watts! Clearly, using the single-cell Beamship methodology had a huge advantage over the "multi-cellular" design that other researchers had built and tested, seemingly reaching an impasse in terms of size and efficiency, which my large single-cell Beamship technique had seemingly solved. In early April my 1-meter Beamship, weighing 22 grams including 5-gram payload, consumed 39.6 Watts. So the 42-gram, 6-foot Beamship used only 7.65 Watts more total input power at stable hover that the 1-meter Beamship. Power-to-weight ratio for the 2-meter Beamship V. III works out to 1.125 Watts to lift 1 gram stably of scale model electric spacecraft. With little to no ion wind, the Beamship had plenty of upward force and achieved stable counterbary at greater efficiency than I had ever heard or read about. Also, I was not aware of any devices in scientific history that has achieved this type of counterbary for this little input power and this colossal size. I performed additional fog-tests with two red semiconductor-emitted laser beams in the plane of one side of the device, one above the wire, one below the foil. The Beamship without all the ion-wind was nearly silent again, except for "singing" and softly "thrumming" anode wires. These laser-beam tests further confirmed the marked absence of ion-wind with a larger spark gap.

At 1.125 Watts-per gram at 57KV I 1.4mA (78.8 Watts) would lift 89.775 grams worth of electric spacecraft. So not only is the larger size in a single cell far more efficient that the "grid" design, in thrust and power consumption, but the reduction in current from increasing the spark gap raised power level to the device, while dropping power consumption of the device. My initial suspicion of ion-wind producing most of the thrust in the Biefeld-Brown effect had been disproven also because of the great weight of the device. I never would have discovered this important fact if I had stayed with smaller devices, trying to raise their efficiency. And I never would have discovered the efficacy of the Biefeld-Brown effect if I had stayed with the multi-cellular "lifter" methodology. One of the factors that raises the level of propulsive force (if "propulsion" is the right word) is increasing the area of the plate, according to TT Brown's patents. So the larger size single-cell capacitor's ability to reduce power consumption and effect greater propulsive force and upward acceleration, was easily explained by the Biefeld-Brown Effect. Brown had been vindicated. My gut feeling had seemingly been confirmed: this was our first warp-drive (reaction-less drive).

My experiments had yielded greater efficiency, and greater size and weight of VTOL hovering models than any that I had ever heard or read about. I still am having difficulty taking in these facts; and it is very awe-inspiring to see such a large device de-gravitate and hover stably at any altitude, from floor to ceiling.

Next for me is a higher-powered supply (60KV), moving up to a three-meter Beamship Variation IV, improvements to the cathode such as a thin, symmetrical airfoil shape, as Brown suggests in his patents, tungsten wire, and then carbon-wire for the anode, and full heat-shrink coverings on the frames, running lights, onboard lasers, onboard digital/proportional Radio/Control, and now that we know the power requirements, and have a good handle on efficiency, onboard power generation. I have already begun designing with my associates our own custom-made outboard and inboard battery-powered power supplies, and finally will cut the power umbilical to the model electric spacecraft permanently, and Beamship technology advances further. If the on/off duty cycle is pulsed at a low frequency, power input can be reduced by two-thirds, at least. Experiments conducted by Jean-Louis Naudin last fall (2001) confirm this phenomenon, suggested by Brown and De Seversky in their patents. Clearly, onboard power can easily be effected, using modern miniaturization and circuitry. Pitch and roll vector control can be achieved by electrically isolating the anode wires on the three sides of the ship, and varying pulses to these three wires. Yaw control can be achieved by simply installing a horizontal-double sided capacitor inside the ship near one corner. Simple full 4-channel flight control is thus achieved.

The Beamship series aircraft are fascinating research and entertainment devices (see cover page), and are the vanguard of a whole new generation of radio/controlled antigravity model aircraft with no moving parts and dead silent propulsion. But they are more than that. The Beamships, if allowed, could probably rise up at any speed thru the atmosphere, right up to near-Earth-orbit, and probably keep on going out into limitless space. No need to achieve ballistic escape velocities of miles-per-second. This is non-ballistic flight. They even have a certain amount of wind resistance outdoors and indoors because the electrical field causes air to flow AROUND the model flying craft, not into it. This is such a safer, environmentally cleaner, vibration and nearly silent and more pleasant method of aero- space travel than carrying tons of explosive reaction mass, which can and does explode. No more use of heat energy to effect transportation.

The 21st century has begun in earnest!

Beamship series aircraft are available for sale for research and hobby/entertainment use right now through the American Antigravity website: www.americanantigravity.com. Look for the Applied Electrogravitics antigravity technology website late spring, 2002. You can contact me, Russell Anderson for details on pricing, and new and improved variations, and power supplies for outboard and onboard drive, which are currently in design stages.

Editor's note: More ideas on development of T.T. Brown's patents are on our web site: http://www.faraday.ru. Read about T-capacitor!

Data Table 1

Antigravity device	Weight of device	Wire type	Voltage/current	Total Watts	Payload	Payload/power
1-foot "Flyer-1"	3.2 grams	#42 enameled copper	33KV I 1.5mA	49	None	————
1-foot "lifter-1"	3.5 grams	#35 enameled copper	26KV I 0.56mA	14.56	None	————
2-foot "diamond-lifter"	6.0 grams	#35 enameled copper	29.5KV I 0.32mA	9	None	————
2-foot, 3-cell "lifter-2"	11.4 grams	#35 enameled copper	38KV I 0.57mA	19.76	None	————
3-foot, single-cell "Beamship Variation I"	16 grams	#35 enameled copper	30KV I 0.52mA	15.6	None	————
Beamship Variation I-fully-rigged	18.5 grams	#35 enameled copper	————	39.5	5 grams	39.9KV I 0.99mA
4-foot "Beamship Variation II"	21 grams	#40 stainless-steel	————	40	6 grams	40KV I 1mA
6-foot "Beamship Variation III"	42 grams	#40 stainless-steel	35KV I 1.35mA	47.25	None	————

References

1. Ether Technology, by Rho Sigma,
2. Lost Science, by Gerry Vassilatos,
3. The Philadelphia Experiment, by William Moore,
4. High Energy Electrostatic Research Vol. 1, Antigravity and UFOs, by Raymond A. Nelli, Homemade Lightning, by R.A. Ford,
5. Antigravity – the Dream Made Reality, by John Searl, The Law of the Squares Book 4, by John Searl,
6. The Antigravity Handbook, compiled by D.H. Childress,
7. The Coming Energy Revolution, by Jeanne Manning,
8. Agnew H. Bahnson Labs home movies (1957-60), JLN labs website, Bill Beatty's "weird-science" website, American Antigravity website (www.amercanantigravity.com).

Patent references

1. A.H. Bahnson, Jr. Electrical Thrust Producing Device, USP # 2,958.790, filed May 12,1958, granted Nov. 1, 1960.
2. Marcel Jean-Joseph Pages (France), # 1,253,902 Petit-Viton, filed Jan 5, 1960, granted January 9th, 1961
3. G.E. Hagen, Flying Apparatus, #3,120,363, filed Sept. 11th, 1958, granted Feb. 4th, 1964
4. T.T. Brown, Electrokinetic Apparatus, #2,949,550, filed July 3rd, 1957, granted Aug. 16th, 1960
5. T.T. Brown, Electrokinetic Generator, #3,022,430, filed July 3rd, 1957, granted Feb. 20th, 1962
6. A.P. De Seversky, Ionocraft, # 3,130,945, filed Aug. 31st, 1959, granted April 28th, 1964
7. V. Gradecek, Electric Aerospace Propulsion System, # 3,177,654 filed Sept. 26th 1961, granted April 13th, 1965

SmartPAK Technology™

invention by William Alek

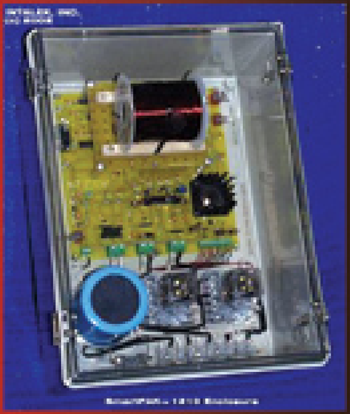

the world's first commercial all solid-state system that provides a "standard" platform for experimenters, researchers, and developers to do energy-related practical applications, experiments, and perform exploration of highly efficient alternate energy system

The theory of operation is based on the amount of energy that is required to magnetize and de-magnetize ferromagnetic materials utilizing a core/coil/magnet assembly. It has been discovered that it takes MORE energy to magnetize a suitable core assembly than to de-magnetize it. The SmartPAK system is designed to measure and collect the difference, and store the excess energy for later use.

The SmartPAK device is controlled by a 68HC908GP32 micrcontroller programmed to measure input/output voltages and currents, calculate COP, and contains software algorithms for a complete "turn-key" power management system. The device features a "standard" user interface, which allows the user to design their own custom core/coil/magnet "head assemblies", and immediately test and display in real-time its' performance.

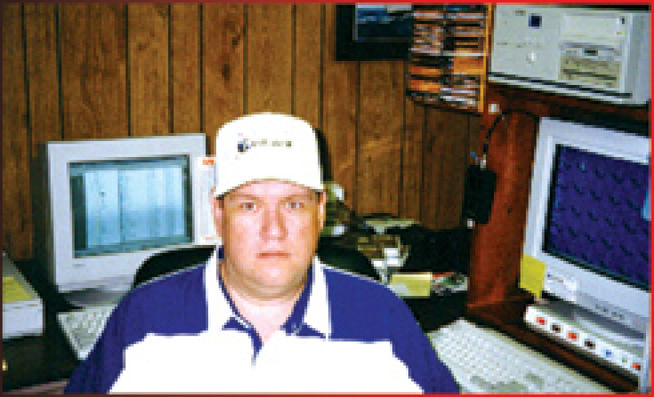

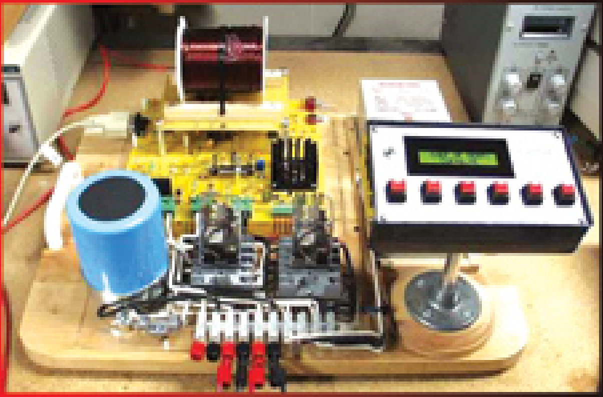

William Alek
President and CEO of INTALEK, INC

3506-43rd. Place
Highland, Indiana 46322-3129
e-mail: wsalek@intalek.com

Correspondence Address: POBox 37,
St.Petersburg, 193024, Russia
Phone: 7-812-380-6564,
net@faraday.ru, www.faraday.ru

Commercial Antigravity

Tim Ventura

tventura6@attbi.com

Introduction

Let me begin by posing a simple question that I would like the reader to keep in mind throughout this article: How far away is commercial antigravity? I'm not talking about a laboratory experiment where a giant magnet is used to levitate a frog, or secret UFO experiments that the government isn't sharing with the business world, but a real, viable antigravity solution to what I consider to be the most pressing issue facing the world today – transportation.

Who can answer a question like that? How far away is commercial antigravity? The author has read numerous scientific texts on the subject, and is familiar enough with contemporary theories of gravity, antigravity, and electromagnetism to suggest that most scientists believe that commercial antigravity is at least 100 years away from existence. But the author believes that most scientists are wrong.

Defining terms

Any article about antigravity would not be complete without properly defining the terms to be used. In this article, antigravity is not used in the strict sense of the word. The author's intent is to discuss a method of propulsion, which for all intents and purposes can be considered antigravity, and may include antigravity – but also may include several other forms of similar propulsion. The reasoning behind this is that experience has shown that the majority of people in the world don't care how something works – they care what it can do for them. This article is about the effect of Antigravity – not the cause.

Real Antigravity would consist of an apparatus used to either reduce the apparent mass of an object or reduce the effects of gravitational attraction between the Earth and an object. An example of an apparatus that may in fact do this is the Podkletnov superconductor apparatus currently being tested by NASA.

This Podkletnov device essentially consists of a spinning superconductor that self-levitates above a pool of liquid nitrogen and supposedly creates a "beam" or "shaft" of antigravity (or reduced gravity) directly above it as it operates. The levitation of the superconductor itself is not antigravity – it is a well-known side effect of ceramic-superconductors called the "Meissner Effect". The Meissner effect is simply a side effect of the superconductor's interaction with the Earth's magnetic field, and is easily explained by physics.

Podkletnov claims that when he worked with a team of researchers investigating superconductors in Russia around 1991, the smoke from the tobacco pipe of a fellow researcher began to climb steeply in a column directly above the superconductor. The researchers began to think that they were on to something, and Podkletnov subsequently performed several follow-up experiments that led him to the conclusion that the levitating superconductor produced a shielding effect between the Earth and anything positioned directly above the superconductor. He reasoned that gravitational shielding would provide a "column" of reduced gravity above the superconductor that should extend up and away from the Earth indefinitely. Podkletnov calculated that with a rapidly spinning levitating superconductor he had achieved a 2% loss in weight for anything directly over the superconductor.

Podkletnov's research is interesting and compelling, and it would fall into the category of "real" antigravity – but I am writing about Podklentov's type of research as well as enormous amount of research and theory available on electromagnetic propulsion systems. These can be considered "effective" antigravity.

The Harrier jet fighter can swivel its engine exhaust-nossels to create vertical lift, which resembles antigravity in that it is VTOL takeoff. However, the Harrier does not use effective antigravity because it has all of the functionality and side effects of an aircraft. A helium blimp would be a closer example to "effective" antigravity, but it too is not – because it works on basic aerodynamic principles.

Aerodynamics is not effectively antigravity – aerodynamics is instead expensive, difficult to manufacture, prone to explosive failure, and highly unreliable. This is not to suggest that a jet aircraft is unreliable, because it has a variety of backup systems, but that the technology itself is unreliable in that a jet is adversely effected by the medium that it uses to propel itself. Air pressure, humidity, temperature, and strong winds all cause a degree of unreliability. In addition, turbine engines stretch the limit of what mechanical engineering can achieve – which is why they are prone to break if even something as small as a bird gets sucked in during flight.

Antigravity is not about moving the air around — it is about a medium-agnostic means of air transportation that produces vertical and/or directional lift without relying on air-pressure like a wing or blimp. Antigravity is an electromagnetic or electrogravitic system for reducing the weight of an apparatus to allow it to lift more easily. Antigravity is pushing a button and having your vehicle take off without runways, noisy engines, minimum flight-speeds, propellers, or any of the other drawbacks that limit conventional aircraft from achieving popularity similar to what an automobile might have.

The author's definition of Antigravity for the purposes of this article is confined to electromagnetic or electrogravitic devices that reduce the weight of an object to enable it to take off without conventional thrust-producing apparatus. This definition might also fit many of the classical characteristics known at one time as "the electric spacecraft".

1. *Business Analysis versus Scientific Analysis*

The author disagrees with the majority of scientists as to when commercial antigravity will become possible for some very basic and obvious reasons. To begin with, the majority of credentialed physics-related scientists come from a theoretical school of thought, which tends to limit their world-view to only contain those things that are currently or potentially explained by theory. The author, however, comes from an experimental school of thought that seeks to capitalize on existing observational data without the rigorous need to explain every last detail of its functionality. The author is an engineer, not a scientist – and engineers don't need to totally understand how something works in order to make it better.

This difference between the engineering point of view and the physicists is also different in the manner in which they seek out observational data. A physicist looks towards naturally observable data, and in the event that none exists they look towards current theories to explain potential future observations. The engineer is more open to ideas that are less rigorously tested from the perspective of scientific method, but are currently observed as potential solutions to real-world problems.

2. *Potential Technologies Overivew*

Phycists currently tend to dismiss the entire concept of Electrogravity, and the reasoning behind their logic is very sound. To begin with, Electrogravity is not observable in nature. In addition, many of the claims by those persons who submit Electrogravity and antigravity devices for public review are faked, exaggerated, or just plain wrong.

Physicists are responsible for maintaining a working body of theoretical knowledge, and if they were to admit results such as Schnurer's without skeptical scrutiny it would undermine the very fabric of technology itself. If the Podekletnov results were to be accepted as fact at face value without rigorous prrof, imagine the amount of money that would be wasted in attempts to build enormous Antigravity vehicles based on this theory.

In the middle of the spectrum lies the concept of Maglev, which is mentioned here only for the purpose of specifying that Maglev is not commercial antigravity. It has been mistakenly thought of as antigravity by many because it utilizes a magnetically-levitated train to improve the velocity of the train and reduce transit time between stops. In reality, Maglev is not really a vehicle at all.

The definition of a vehicle would be a device that transports itself as well as its passengers and cargo between two points. This is why an automobile is considered a vehicle but an escalator or elevator is not – the automobile transports its entire propulsive apparatus to another location, but an escalator or elevator does not move – it merely repositions its cargo between points. The Maglev train is not really a vehicle at all – it is actually a very long electromagnetic armature that transports people and cargo between its ends at high speed. While it may serve a commercial need, it is not to be confused with Antigravity.

On the opposite end of the spectrum is the author of this article – who has built and successfully tested over 30 electromagnetic "Lifters" at the time of this writing. The Lifter is a device based on research by Transdimensional Technologies and related to research by Thomas Townsend Brown that demonstrates an antigravity effect when a High-voltage DC current is applied to it.

Currently, the exact method of propulsion for the Lifter is being debated. It is thought to be one of two things – either an effective form of "ion-wind" propulsion, or else a form of field-effect propulsion based on an as-yet unknown force. While the debate about the exact nature of this propulsion is important with regard to future research, in reality it does not change the observational data that demonstrate that this technology works perfectly, consistently, and reliably.

The Lifter design was demonstrated by the author in a continuous mode of operation for over 7 hours straight on Sunday, April 21st, 2002, at the Seattle Center "EarthDay and Renewable Energy Exhibition". During this seven hour period of time, the author's Lifter hovered at a tethered height of 12 inches from the surface of the table, powered by a 30 watt load from a simple computer monitor.

This article is not meant to get into the details of methods of antigravity, only to suggest that it already exists in the form of electromagnetic propulsion systems if nothing else. The author is confident that in time physicists will find a theoretical reason for why the Lifter operates as it does, but for the time being the fact of its operation overshadows the method of its operation.

3. *Market Needs*

Commercial Antigravity doesn't require a 2% loss in weight to operate – it will require something akin to a 200% loss in weight. A commercial antigravity device will have to demonstrate exceptional performance to gain market acceptance, but not for the reasons that might immediately come to mind.

One might believe that skepticism from the scientific community would prevent antigravity technology from gaining the scientific acceptance needed to become a commercially accepted engineering discipline. The long term view, however, shows that this is not the case – engineering and market forces drive innovation, and formal science plays a supporting role in explaining and quantifying the innovations that engineers have already commercially qualified as valid.

The real roadblock to success for commercial antigravity is market acceptance. The author's demonstration of the Lifter technology at the Renewable Energy Exhibition helped him to realize that the vast majority of consumers have no idea what antigravity technology could be used for, much less what they themselves could use it for. The same thinking was apparent at the dawn of the age of personal computing, when the idea of having a computer in the home was a completely foreign concept.

So in brief, a market does not exist for antigravity technology, which is why inventors working with this technology have been unable to find appreciable support for their work. Many inventors look at this technology and ask, "how could the public not understand how valuable a technology like this is?" – but that isn't the problem. The problem is that most innovators with an interest in antigravity are so closely tied to the science behind the technology that they fail to review and address the business needs that drive the market acceptance of a new technology. In other words, people don't buy antigravity – they buy solutions. People don't buy cars to simply have a car – they buy cars because people need transportation needs that they have to fulfill. People don't buy computers because they want to have a computer – they buy computers because they want to share and process information and communications.

Marketing Requirements

How will antigravity technology gain the market acceptance to become a commercially viable technology? There are a variety of ways in that antigravity technologies will become commercially viable, but only after antigravity is no longer sold as antigravity – it needs to be sold as a personal or business solution.

The solutions that antigravity technologies are best prepared to provide at the moment are in the realm of transportation technology. This includes moving people and cargo to destinations in a similar manner to conventional transportation technologies such as aircraft or automobiles.

With regard to providing transportation solutions, antigravity has the ability to incorporate the best features of both contemporary automotive and aerospace technologies into a single technology that will serve point-to-point transportation needs better than either of the two aforementioned technologies could by itself.

For a moment, assume that a person wants to travel from Los Angeles to New York in a short period of time. Currently, the most convenient method of transportation to accomplish this would require the person to take an automobile to the airport, and from there take an aircraft from the Los Angeles airport to the New York airport. After departing at the New York airport, the passenger must then take another vehicle to their intended destination.

Commercial antigravity technology could serve a dual-purpose short and long-range transportation role, taking on the aspects of both ground transport as well as air transport.

Product Delivery Requirements

In order to deliver commercial antigravity as a viable solution to business needs, a variety of work will need to be completed on the various component systems of this technology to turn it from what is currently a "proof of concept" into a commercial reality.

Let us assume for a moment that we have developed a working device based on Antigravity or some method of Field-Effect Propulsion. While this is the critical stepping stone to success, this is by no means the end al be all of the development cycle.

To begin with, the technology must be perfected to the point of being both economical and reliable. As it stands now, the market already has technologies in place that fulfill some or all of the requirements for the technology that Antigravity is being developed to replace. In order to serve as an effective replacement for these technologies, antigravity technology must then demonstrate that it both costs less in terms of operation and manufacture, as well as being more reliable than conventional air-transportation solutions.

I mention reliability in light of the recent negative media attention surrounding several recent commercial airline crashes. From a marketing perspective, air-travel disasters provide a great deal of negative publicity for the airline industry. Since the airline industry has a mostly successful track record of delivering passengers and cargo, people are for the most part willing to forgive the occasional air-disaster. However, with a new technology such as antigravity-based air-transport, there is not a long enough track record to permit public acceptance of air-disasters. One substantial disaster in the early days of antigravity could serve to forever damage the credibility of this new technology.

With regard to being economical, any type of antigravity system that intends to surpass existing methods of air-transport must be able to do so at a less-expensive rate to own and operate, and must have a vehicular lifespan at the very least similar to conventional air-transport devices. This would allow the total cost of ownership (TCO) to be less for an antigravity vehicle than it would otherwise be for a conventional craft.

There is one caveat to acceptance of antigravity technology as compared to conventional aircraft, which is simply that if antigravity vehicles are able to operate in an environment or manner that precludes conventional aircraft, then they should be able to gain a market niche without immediately having to surpass conventional aircraft in the area of TCO.

Assuming that we can develop a propulsion system that is both more reliable and less expensive to operate for the transportation of passengers and cargo, we then have to build up the skeleton of a vehicle compatible with this form of propulsion around the actual propulsion system.

For instance, a conventional aircraft has pitch, yaw, and several other flight controls, but for an antigravity vehicle there is a high likelihood that some or all of these controls will not be required, thereby changing the dynamic of flight associated with the craft. This will require new methods of pilot certification and flight-qualification, as well as requiring a control-philosophy to be created surrounding how the craft will operate.

I use the phrase "control-philosophy" instead of simply "control layout" because one of my assumptions about antigravity propulsion systems is that they will allow more flexibility in the design process for engineers to determine how the craft "should" fly, as opposed to an aircraft or helicopter, in which the components determining speed and handling are based primarily on an interaction between the design of the craft and the atmosphere.

In brief, an antigravity cargo-transport may have very different needs for flight than perhaps a lightweight passenger vehicle would, although there would also need to be a consistency between the control-systems of these devices to reduce the need for extra pilot training and competency testing.

Therefore, it should be apparent from the last few paragraphs that not only are there several propulsion-system related challenges involved with developing a commercial antigravity device, but there are also several challenges in the design, training, support aspects of this technology that also factor into the requirements to be complete before a complete product can be delivered.

Conclusion

At the beginning of this treatise, I posed the simple question of **"how far away is commercial antigravity"**. The reader, I expect, probably interpreted that question in terms of time, which is the usual measurement of questioning when new technologies will become part of our lives.

However, as I have attempted to demonstrate throughout this article, **the time component is much less important to the development process than is the distance component – that is, how far away from commercial antigravity we are.** When I use the word distance, I mean specifically what tasks must be completed in order for antigravity to go from being a proof-of-concept approach to a new form of propulsion-system to being a completed vehicle ready for manufacture.

I have attempted not to address the legal implications of antigravity technology with regard to certification for general or specific use – my thought on this is that the discussion of legal ramifications of antigravity is best left for another time. This is due primarily to the size and scope of that discussion, which is beyond what I am attempting to analyze in this article.

So, in finale, how far away from commercial antigravity we are depends not so much on time as on the rate at which we can perform the work required to provide the underpinnings on top of which the technology can be built. This seems important to me, as it underscores how close we appear to be to a working method of antigravity propulsion, and how we might consider focusing resources and goals to achieve the realization of this common dream that we share.

EDITORIAL: PERPETUAL MOBILE OF 1902

There is one more interesting example of perpetual mobile, which was described in the collected articles [1]. The motor shown in Fig.1 was invented in 1902. The vessels *b, c, d* and *e* are mounted on a shaft *a*, and have one side *f* tangential to the shaft, and the other side radial. Compressed air is forced into each vessel through the valves *p*. It is stated that under "the action of the internal pressure of the vessels, and after a slight impulse has been given to same, in the direction of the arrow, the whole apparatus will begin to move and continue to do so without ever stopping, the velocity corresponding to the pressure established within the vessels".

Really simple... Let's try to examine it.

Reference

R.A. Ford. The Perpetual Motion Mystery. A Continuing Quest, 1987 Lindsay Publications, Inc., Bradley, IL 60915

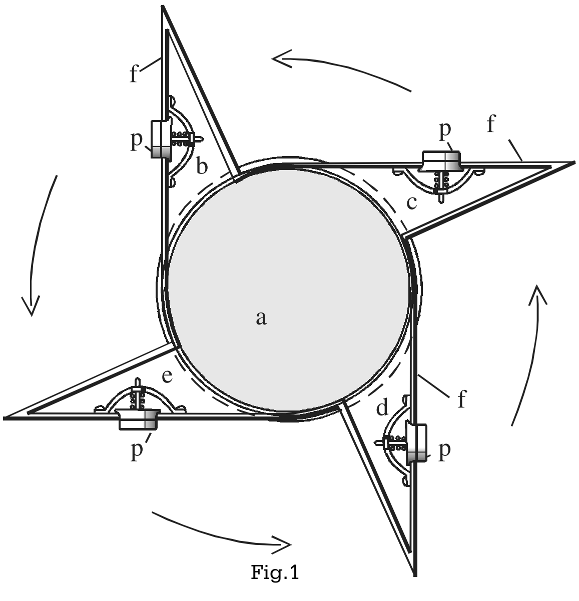

Fig.1

Action without Reaction

New Gravidynamic Paradox

Yu. N. Ivanov
Academician of Russian Academy
of Natural Sciences

Saikina, 11/2, app.4, Moscow, Russia
Tel: (095) 277-33-76/276-50-39
mirit@narod.ru
http://www.mirit.org

Editorial: Yu.N. Ivanov discovered a new unusual physical phenomenon that is a gravidynamic paradox. The sense of the paradox is the existence of the situation, which is not forbidden by physical laws and which allows any man (even a child) easily hold suspended or carry things with 100kg or more weight by means of some simple device. This phenomenon is just a part of the significant applied topic and a demonstration of the possibility to cancel weight characteristics of any material object. Per se it is a work on the obtaining of the methods to eliminate aircrafts weight.

We can use only resistant things as a support!

Galilee-Newton's Laws are in the basis of classical mechanics. Thought concerned fundamental, these laws do not explain the main thing, i.e. the internal processual essence of the phenomena, described by them. There is another scientific line – the quantum mechanics, which tries to discover the essence of deep processes. There is no connection between these lines, therefore they as if exist independently. There are also long-time conversations about necessity to combine the quantum and the classical approaches. We suppose, that practically the combining has already happened, since thanks to rhythmodynamics, there appeared phase, frequency and velocity of light in formulae of classical mechanics (without these attributes the quantum mechanics is impossible). It was revealed, that phase-frequency method of late mechanics formulae presentation gives a real physical sense both to formulae themselves, and to concepts which seemed vague before *(namely: cause of motion, force, velocity, inertiality, gravitation)*. In this sense, the upcoming rhythmodynamics fills up the gap between the main physical lines and draws them together. But let's change the vector, since **the subject of the given article is to concern cases in which the third Newton law is not directly valid.**

Newton formulated the third law in the following way:

"Action always has equal and opposite interaction, in other words, actions of two bodies on each other are equal and directed in opposite sides".

This law represents the fact that one-way action of one body to another cannot exist in nature, but there is only an interaction between them, i.e. there is no action without reaction [1].

We must note, that the third Newton law is valid only for systems with 100% feedback, appeared in the interaction period. For example, during the magnet influence on the iron object it is revealed, that this object also starts to influence upon the magnet, i.e. to attract it with the same force. It happens because in the period of action of magnetic field the iron object itself becomes the source of magnetic field (Fig.1). Here the magnetic field, as an especial environmental condition, acts as a mediator, i.e. exists by itself in the interval between object and magnet .

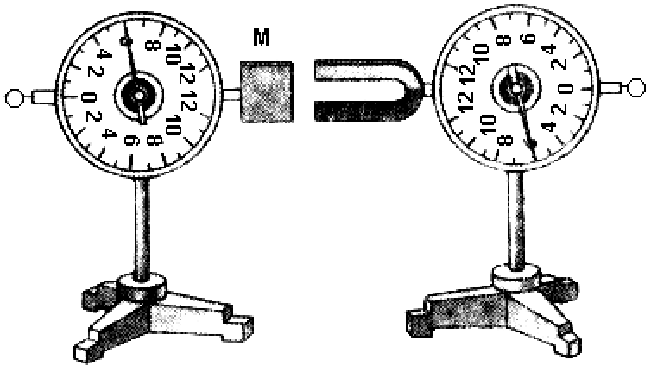

Fig. 1

Interaction of magnet and metal body M. Here the action is equal to reaction, i.e. the third Newton law is valid.

It is considered to be that magnetic field, created by electric current, spreads with the velocity of light. If current impulse is short, then the magnetic field spreads independently from the conductor and no matter if there is current in the conductor at present, or not. If the distance up to the object is big, then the magnetic field remains between source and object for some time, and has no influence upon the object. In this sense the portion of magnetic field, placed on the path between source and body, can be considered as moving independent "entity", i.e. the space by means of itself carries its changed state (magnetic properties) from one place to another. In this period magnetic properties in the form of quantum can influence neither on the source, from which the magnetic quantum is already detached, nor on the object, which the magnetic quantum still does not reach.

But having reached the object, magnetic field changes the state of this object. If as the result of influence the object becomes a source of magnetic field, then the part of the field is reradiated in the direction of source, i.e. the object itself becomes the source for some time, and by this it is able to influence on the first primary source. In this situation the principle of action and reaction works, because the feedback takes place.

If the feedback does not appear during the period of influence of one body to another, then the action and reaction law is not valid in the system. Let's demonstrate it by the example of the mechanical experiments, in which there is a feedback between objects.

Let there is a device (Fig. 2), which throws off two water (air) streams in the opposite directions in such a way, that reactive forces completely compensate each other. At that the thrown down stream compensates the gravity too. In this case the device will fly without falling, i.e. it will have zero weight.

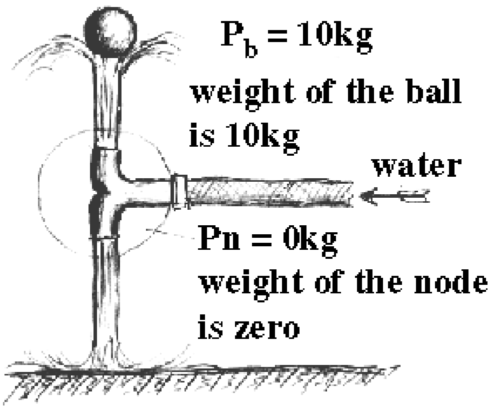

Fig. 2

The ball, suspended in the water stream, does not put pressure on the source. It happens because there is no feedback between the ball and the source. If you have such a device, then, lifting the ball by means of it, you will not feel the weight of the ball. The reason: the ball hangs due to the kinetic energy, which the stream passes to it and and this ball can not influence on the source through the water stream.

The appearance of the ball with the weight P_b does not influence on the weight P_n of the device. And even in the case if we press the ball down, i.e. essentially displace its location in the stream, the device will feel nothing.

The absence of device weight is evident for us; therefore we can easily move the device if its original (vertical) orientation remains unchanged.

Let's change the situation and place the ball (body with the shape, which is steady for hanging) in the upper stream. This ball with the weight about 10 kg is placed in a way to be kept by this stream at some distance, for example at 0.5 m. Will the device react on the changed situation, i.e. on the ball appearance? Will the weight of the ball, hanging in the stream, be added to zero weight of the device (0 + 10 kg)?

The calculation shows that it will not: $\Sigma P = 0 + 10 = 0$ [kg]. There is only an illusion that the ball is supported by the stream and that it is a part of the system. As a matter of fact, the ball is detached from the system and hangs due to the water kinetic energy. It is easy to check it by making an experiment in a bathroom: "Remove the douche sprayer, then direct the stream up and holding it by the one hand with the aim of weighting, try to influence on the source by the other hand through the stream. You will be surprised to feel that the hand, which is supported by the stream or any object, which is hanged in the stream, does not influence on the source".

At first sight the situation seems to be a paradox, however it very illustrates the possibility of the force action without reaction. So, the action can be one-way, i.e. calling no reply reaction in the form of pressure on the action source.

In such an unusual way it is possible to keep the ball of bigger weight (100 kg and more), at that to move it easily by means of the device and at the same time not to feel the presence of additional weight in the stream. **It is possible only in the case when there is no feedback between body and source,** i.e. the body, hanged in the stream, has no ability to act on the force source. Thus, we can both hold and move the heavy ball without additional efforts, and also lift it up to any height (for example, 100 m). As a matter of fact, we deal with the **new gravidynamic paradox**.

It is interesting then, how to solve the task of the following type: "Let the body with 100 kg weight is hanged in the water (air) stream in such a way, that it does not change the velocity and direction of liquid outflow from the source nozzle. How much energy the operator must spend to lift this body to the height about 10 m?" (Here it means that the operator must take the device, which supports the ball, and, moving upstairs, lift the body, which hangs in the stream, by means of this device to the height about 10 m). If to solve this problem correctly (the condition is that the operator lifts himself together with the device and the ball), we will find out, that operator's energy is spent only for lifting of the device, which creates water streams. The operator, lifting together with the device, will not even notice that in the stream there is a body with about 100 kg weight **(this is the sense of the paradox).**

The situation only seems to be absurd, and even paradoxical for theorists, but it is not a hopeless one.

We can also observe the effect of action without reaction in ultrasonic field of the source. If the source is fixed on scales (Fig.3) and body is hanged, as it was made in the stream, the scales will show only the precise weight of the source and will not react on the weight of the hanged body, no matter how heavy it is.

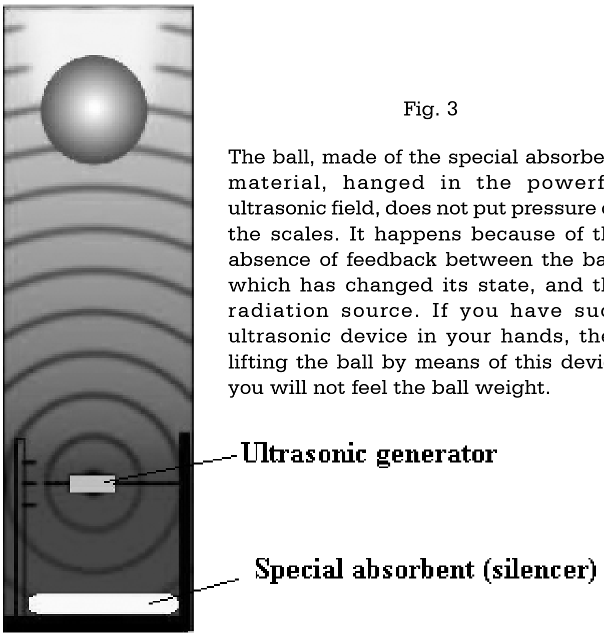

Fig. 3

The ball, made of the special absorbent material, hanged in the powerful ultrasonic field, does not put pressure on the scales. It happens because of the absence of feedback between the ball, which has changed its state, and the radiation source. If you have such ultrasonic device in your hands, then lifting the ball by means of this device you will not feel the ball weight.

The experiment of such type was made in one of the secret institutes. Water was used as an acoustic environment. The heavy ball easily floated up from the vessel bottom i.e. it behaved as if it had no weight.

The similar phenomena but such of the other organizational level, sometimes become apparent in nature. The phenomenon has its very name because the event happens unexpectedly and is not analyzable through laboratory research [3].

"A spherical ball lightning about the size of a football ball flew over a village Galtsovka of Altaisky Krai at a height of twenty to thirty meters. The first shed in its way, with ferroconcrete poles, was crushed and collapsed. Flying over a slated house, the ball lightning tore away the roofing slate together with nails, raised it into air and pulled it along, scattering its parts all over the village. Flying over a tractor station, the ball lightning crushed a frame, welded of metal angles and covered with tarpaulin. On the approaching of the ball lightning to another frame, it was at first dragged along the ground, and after the ball lightning had passed it, the frame was lifted and carried at a distance of 300 meters. The weight of the frame was no less than 100 kg".

Further the author analyzes the situation from point of view of the known physical laws: "The frame of hundreds kilograms weight was carried by fields of the flying ball lightning. However, for some reason the ball lightning kept on flying straight and did not note that some metal frame was caught to it. If the ball lightning, as it is usually considered, has a density of air and it is a weightless formation, then why the frame with the weight not less than 100 kg could not change its flight trajectory even in some extent?"

It draws attention that in some cases the ball lightning field pushed objects away, and in other cases it attracted them. It can be explained only by fact that in each case the ball lightning field specifically influenced on internal properties of objects, and then objects themselves somehow reacted on their new state (they changed their motion) . The feedback absence (in other way it happens between magnet and iron object) allows the ball lightning not react upon the things, which take place in its field. If these objects themselves became sources of similar field, i.e. reradiated it, they would influence on the ball lightning trajectory. **Most likely, in the given case, there was realized the situation when action caused no reaction.**

Conclusions

In the context of known physical laws the particular problem of direct action without feedback was formulated and solved. The solution of this problem provides deep understanding of how to control weight characteristics of material objects in open systems. As it was shown by the example with water (which is only an illustration of more fundamental processes) we can "deceive" the nature, but only through the deep understanding of the processual character of the concerned events.

By the example of ultrasound we show, that in principle there can exist such field flows, which are able to influence on bodies without the feedback effect. There were defined some conditions and criteria of bodies and methods of influence, at which the third Newton law is not directly valid (it is not published in this article).

The described type of influence can be called as pressure. But during consideration of processes at atomic or deeper level, i.e. from the position of rhythmodynamics, we find out something of another kind, that is phase-frequency one. It prompts to us, what kind of technologies will exist and how our aircrafts will look in nearest future. But not everyone is able to understand it at once.

Moscow 15.05.2002

For sponsors: mirit@yandex.ru
Our website: http://www.mirit.org

References

1. Physics textbook for the 9th class of secondary school // M., Prosveschenie, 1982
2. Ju.N. Ivanov, Rhythmodynamics // M., Noviy Tcentr, 1997
3. G.V. Nikolaev, Scientific vacuum // Tomsk, Kursiv, 1999

Comment

In the gravitation field we have a balanced open system, in which appearance of the new body at first sight must cause this system reaction such as weight increase. However, this is not the case.

The given problem is solved in two independent steps:

1) examination of interaction between water source and water (reactive effect, which is compensated by counter flow);

2) examination of interaction between water, detached from the source, and the body (unit time impulse per unit area).

It is impossible to examine the interaction between the source and the body because water has no rigidity. There are no means to influence on the source through the intermediate stream, therefore the principle of action and reaction is valid for each step individually, but not for both! But in this case the third Newton law must be developed as following: "If there is a 100% feedback between two bodies, their interactions are equal and inversely directed. And if there is no feedback, the action of one body to another causes no reaction, i.e. the action is not equal to reaction". And this is already another law!

New Possibilities of Vortex Electric Power Devices

Stanislav A. Lisnyak

Prospect Stoletiya Vladivostoka, 112-37
690069, Vladivostok, Russia
Tel/Fax: (7-4232)310-554
Email: Stasvladcomru@mail.ru

The earlier published article [1] was devoted to the explanation of the excess energy output at the rotation of flow medium in vortex tube (VT). The researches on VT demonstrated by the real examples that the excess energy output is a real fact, which is determined by natural physical processes. These processes take place in the fluid at its vortex (rotational) motion in the closed volume by means of phase transformations (skips) of the fluid state. However, the practice shows that for obtaining of patents and benefits for the real vortex devices, which are made for heat generation, it is necessary to write a little about their real technical characteristics, though technical and constructive possibilities of vortex devices can be much better. Temperature of water of these devices could not be higher than 120°C, and value of fluid pressure is confined within 5…6 kg/cm^2. Circulation pumps of pressure over 6 kg/cm^2 are used to increase thermal power output of vortex heat generators (VHG). It causes the discredit of VHG since at such a method their generative possibilities reduces to 100% value and less! We set a problem to discover the ways and engineering solutions for the essential increase of the coefficient of VHG energy conversion to not less than 200% value.

Our experience in development and exploitation of deep-sea physical devices has given some engineering solutions in this way. Actually, we have the opportunity to use the generally applicable circulation pumps, which have pressure not more than 4.0 kg/sm^2. At the same time we can raise the temperature of the heat carrier to 300°C and higher! and the pressure can be increased to 1000kg/sm^2! Such fluid parameters allow to use this fluid as a heat carrier in powerful hot-water generators of steam engines, turbines etc. By the way, this method allows to make the value of temperature of water enough to supply its PYROLESIS! The main point of the method is the following: all the closed system of water-filled VHG works at pressure with the value under 1000kg/sm^2. It is provided by means of the special device, the so-called automatic pneumohydraulic block (APHB). This lets to raise the temperature of working fluid to 1000°C with no change in its aggregative state. The practical functional scheme of such a device is shown in Fig.1.

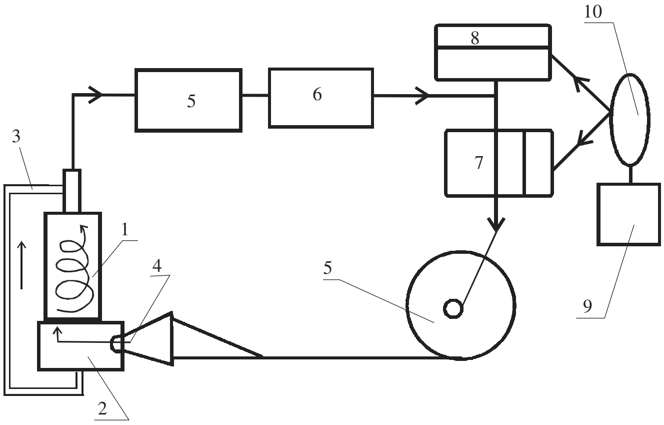

Fig. 1

1 – vortex tube; 2 – entry vortex chamber; 3 – by-pass; 4 – upstream end; 5 – circulation pump; 6 – heat-exchangers; 7 – APHB; 8 – volume of fluid leakage compensation; 9 - compressed-air flask; 10 - gas pressure regulator

The device works in the following way. The whole device is filled with working fluid without entrapped gas. At the operation of the circulation pump (5), fluid through the upstream end (4), placing in the entry vortex chamber (2), swirls, accelerates and gets into the vortex tube (1), where its "energy saturation" is realized. Then heated fluid gets into the heat-exchangers (6) for heating or for other purposes. After passing through the heat-exchangers, fluid gets into the pneumohydraulic device (7), where there is kept up the proper level of fluid pressure in the whole system. To prevent the system breakdown at the accidental fluid leakage, there is a device (8), which compensates such leakages. Compressed-air flask (9), with the volume under 1000kg/cm^2, and gas pressure regulator (10) keep up the selected level of the pressure in the system. The use of blocks (7,8,9,10) excludes evaporation in the system and prevents the breakdown of the circulation pump (5).

This VHG scheme allows to double the effectiveness of YUSMAR devices [2] only at the expense of rise in working fluid temperature in 2…3 times.

Observation of any fluid swirls and film documents of windspouts arrive to conclusion that all vortex structures are rotation bodies, created by the lines of the second order: $Y=aX^2$. In other words, as a result of the rotation in the swirl, air or fluid mass gets the acceleration of the second order. Taking into account the aforesaid, it is evident that in order to form the classical swirl in the VT, the very VT should be a tubular body of rotation. This body is created by the curve (see below Fig.2).

In YUSMAR and similar devices the vortex fluid motion take place in straight cylinder and 1/3 of this area is used for fluid deceleration that causes vortex flew disruption. This results in the impossibility to increase additional heat in the straight tube in more than 1.54 times. It is caused by the fact that the main vortex formation takes place only in the vortex chamber and the flows separation does in the tube itself. Then, this vortex formation is right away disrupted by different plate brakes! It suggests itself that output part of VT should be made in the spiral form, expanding at the flow passage.

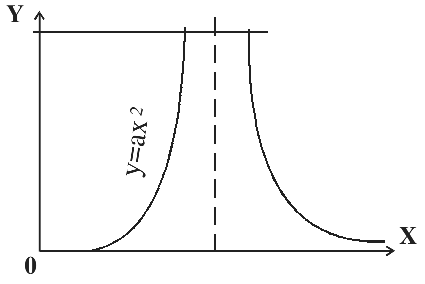

Fig. 2
Generatrix of the wall of vortex tube

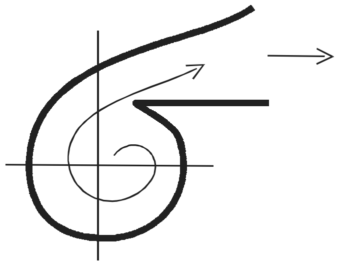

Fig. 3
Upper part of the tube

Such section, made on the outlet of the vortex tube, helps to slow down the flow avoiding its disruption, since axial-cold flow, placed along the tube axis, should stay intact. Two of such engineering solutions allow raising the level of the additional VHG energy up to 180…200%.

YSMAR devices are designed for one tube with the assigned heat output. However, if to take into consideration and to use the aforesaid conclusions, then we can raise energy conversion coefficient (ECC) above 200% at the series connection of two and more VT. The scheme of such a device is shown in Fig. 4.

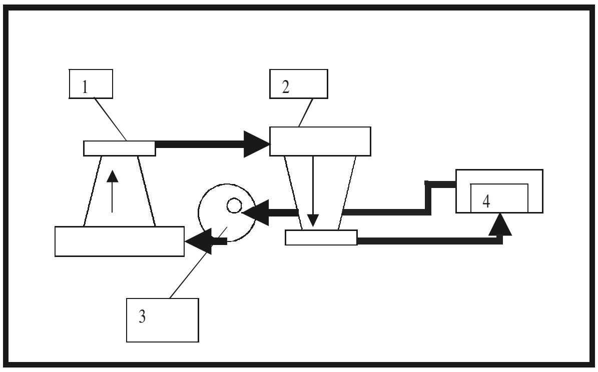

Fig. 4

Where: 1 – the first VT; 2 – the second VT; 3 – circulation pump; 4 – heat exchanger

According to Fig.4, the total energy output of the presented device is:

$$Q=K^2$$

Where: Q is the total output of the device; K is energy conversion coefficient of one VT; 2 – quantity of serially connected VT; P – the assigned power of the circulation pump.

In our earlier article [1], concerning VT application, we supposed that there is a straight transformation of vortex motion of ionized fluid into electric current. Under studying of many articles, devoted to the methods of water ionization, we discovered a significant one. It demonstrates that at the determined temperature, pure water increases its ionization capacity up to 3 orders without changing in the aggregative state [3]. The diagram in Fig.5 shows such dependence.

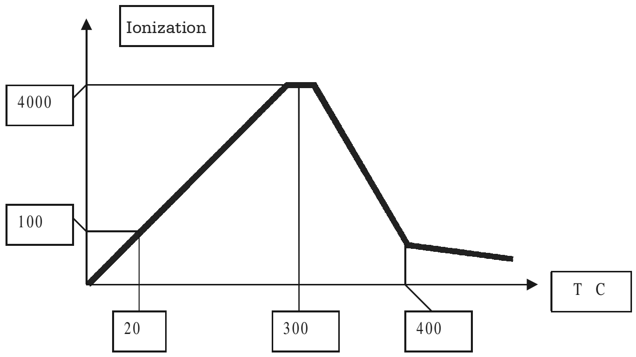

Fig.5
Diagram of the dependence of pure water ionization from temperature

The diagram demonstrates that water at 300°C temperature increase the quantity of ions in 4000 times as compared to 0°C and in 40 times as compared to 20°C. If to take into consideration that first experiments on VHG modernization help us to find the ways of fluid heating up to the practically any temperature without changing of its phase state, then on using the dependences (see Fig.5) it seems possible to create vortex fluid electric generators of the forward transformation. The aforesaid pneumohydraulic block, which is used for VHG functioning, allows to keep up water temperature at 300°C and pressure at about 90kg/cm^2 without threat of water evaporation! Presence of any rotation of fluid medium always causes appearance of two vortex flows. **These flows always move and rotate in opposite directions and if we do not put obstacles for their motion then they transform into each other and can exist without energy supply for an indefinite period of time.**

The presented VT form (see Fig.6) provides the producing of two fluid flows in the tube. These flows do not influence to each other and there is only their reciprocal overflow without disruption of the flows in the central part of VT. The tube presents a tubular body of rotation, which is created by hyperbola. There is positive angular acceleration of fluid in the lower part of the tube; maximum speed of rotation of the axial and peripheral flows is in the middle part and negative angular acceleration, i.e. deceleration and transfer of kinetic energy into heat energy, is in the upper part of VT.

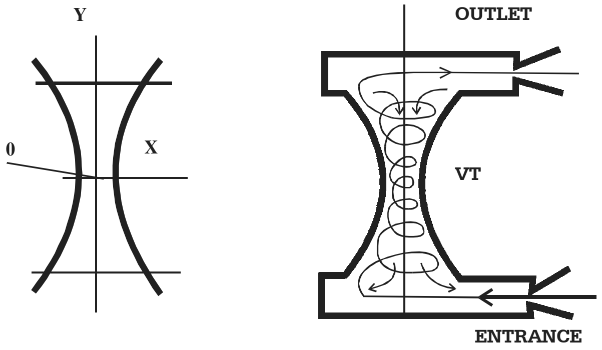

Fig.6

The maximum heating of fluid and its polarization is in the point Y=0. The polarization potential in the point can reach 10.000 V at 5m/sec flow speed for 10 cm diameter in the point. The cold flow, which comes along VT axis, is opposite to the peripheral hot flow. These flows close into one system in lower and upper chambers of VT. Thus, the presented VT classically modeled "rolling" and "unrolling" of water area.

There are no single vortexes in nature. Two vortexes, rotating in opposite directions, always appear while fluid mediums rotation! Science is not still able to describe energy overflows from one vortex center to the center of another one by means of mathematics. However we believe that this moment is close. There are maximum energy transformations of rotating water in the point Y=0.

The potential of the inner flow is equal to the outer flow by its value but is opposite by its sign. Potential difference is maximum in the point Y=0, in which removal of electric charges is the most effective.

It is appropriate mention here the name of Romanian engineer and researcher Henry Koanda, who in 20th of the last century discovered the so-called "Koanda effect" (attachment of fluid jet to the surface of usual kettle at pouring of it out the cups). Basing on this effect he suggested to make new type of aircrafts, which could have advanced bearing capacity and maneuverability. Unfortunately, money and conservatism of aircraft-industrialists did not allow realization of these engineering solutions.

At the same time, Koanda charged his disciple Patrick Flanagan with the job to research all **water properties**. The result was amazing! It was turned out that water actually has infinitely many phase states and when it is moving, then it can trap energy from the environment by some way. The publications on the point appeared in our press in the early eighties, late nineties of the last century. Most likely, they had become the basis of the invention, made by Potapov, i.e. vortex tubes for water!

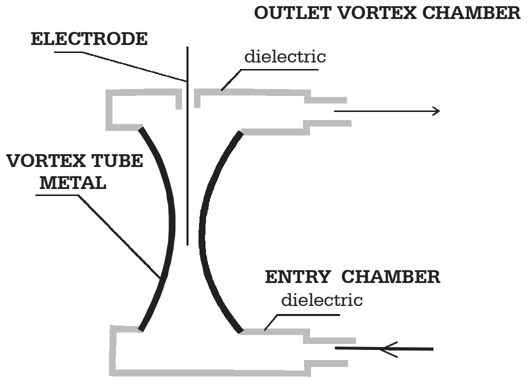

Fig.7
Scheme of vortex electric generator

In the presented scheme of electric generator the entry and outlet vortex chambers are made of dielectric material in the spiral form. This material certainly should withstand not less then 300°C of fluid temperature and about 100kg/sm^2 of pressure. Fluoroplastic or ceramic can be used as such a material.

From aforesaid it seems to be possible to submit the new scheme of electric generator for readers' consideration (see Fig.8).

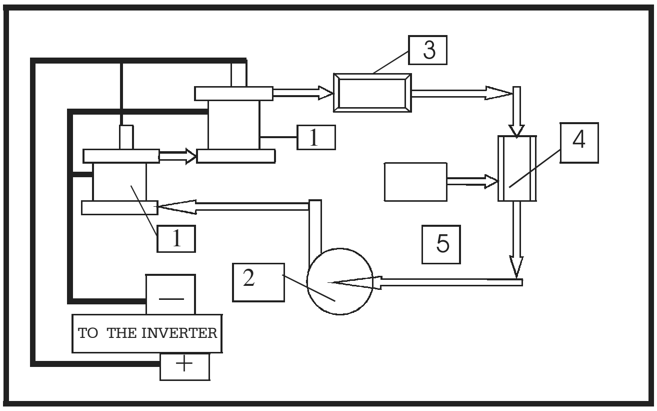

Fig.8
Generalized scheme of vortex electric generator
1 – vortex tubes; 2 – circulation pump; 3 – heat exchanger; 4 – automatic pneumohydraulic block; 5 - compressed-air flask

Actually we could finish the article if it were no disputes about excess energy, which is released at vortex motion of fluid in tubes. Having used a simple experiment, which any inquisitive person can repeat, we found the positive solution (see the scheme of the experiment in Fig. 9).

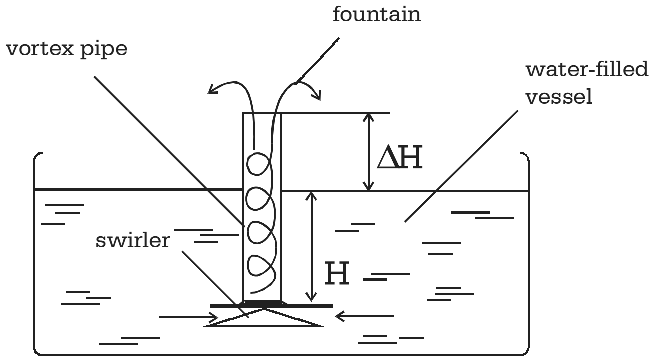

Fig.9
Demonstration of energy increase of fluid, rotating in vortex tube

A pipe with not less than 60 cm length and about 1-inch diameter is placed in a reservoir with not less than 50 cm depth. The lower part of the pipe is formed as an arm circular generator with about 15...17cm diameter. The gap width is about 2 cm. If we place the pipe with closed top end, which then will be open, into the reservoir, then water is splashed out above 20 cm of the reservoir level. Thus at the least 20% energy increase of energy mgH is observed.

By the way, we suggest orthodoxies of physics and other who have doubts to make one more simple experiment. Try to boil water in a can. When water starts to boil, swirl it with a spoon to the right or to the left. Then you will suddenly discover that all water surface is calm and only in the center of the can there is a water-steam mixture, which is about 20% higher than the level of water surface. You will also see that water is absolutely calm at walls of the can. This experiment can be reproduced anytime and anywhere, even in space, since it is a demonstration of **vortex** fluid motion and its actual influence on everything, which is around!

All presented engineering solutions are practically reproducible in any laboratory. The authors of the article would welcome the opportunity to co-operate with science and industry representatives for realization of the ideas in real serial production devices.

References

1. O.V. Gritskevitch, S.A. Lisnyak, On possibilities of vortex energetics //New Energy Technologies. 2002, #1(4), p.2-3
2. Patent RF #2045715, 1993
3. Chemical and Engineering News. 2000 #1, p.26

Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels

D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko

All-Russian scientific Research Institute of Electrification
of Agricultural Economy
2 Vishnyakovsky passage, 1 Moscow, 109456, Russia,
energy@viesh.msk.su

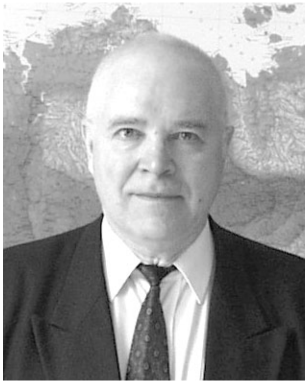

D.S. Strebkov

(Editor's notes by Alexander V. Frolov)

It has been found that at low frequency (1-25 kHz and higher) electric power can be transmitted with low losses from generator to receiver along single channel made of non-metallic conductive media such as water in plastic tube carbon thread, layer of damp soil, ITO films on glass substrate, laser and electronic beams. Transmitted power, as well as for traditional three phase lines, is limited by natural power of the transmission line and line capacity may reach at high voltage and pulse and operation modes the value 10^9 Wtt.

Introduction

The well-known methods of electric power transmission are based on transmission of active energy by means of conductivity currents in closed circuit. Electromagnetic energy spreads along power transmission lines (PTL) as progressing waves of electromagnetic field or field of charge [1]. Line wires made of aluminum or copper are conductive (guide) channels. Electromagnetic energy stream moves along these channels from generator to energy receiver and backwards to the generator. Maximum transmission possibility of 3-phase PTLs is limited by losses on the line resistance, by peak voltage (which is determined by electric strength of the insulation) and by electromagnetic stability of the line.

The modern approach to provide the electromagnetic stability consists in rigid regulation of line parameters by means of high-speed shunt reactor and consequence capacitive compensation for the purpose to except changes of electromagnetic power flows and to suppress resonant properties of a line [2].

In Tesla works [3] and in the researches of Russian scientists [4] a method of active power transfer was offered. This method supposes to transfer active power by means of electromagnetic capacitive current assisting with resonant properties of a single-wire line (SWL), made of a metal conductor. The purpose of the present work is a research of an opportunity to use non-metal conducting mediums for transmission of electric energy.

A pipe with not less than 60 cm length and about 1-inch diameter is placed in a reservoir with not less than 50 cm depth. The lower part of the pipe is formed as an arm circular generator with about 15...17cm diameter. The gap width is about 2 cm. If we place the pipe with closed top end, which then will be open, into the reservoir, then water is splashed out above 20 cm of the reservoir level. Thus at the least 20% energy increase of energy mgH is observed.

By the way, we suggest orthodoxies of physics and other who have doubts to make one more simple experiment. Try to boil water in a can. When water starts to boil, swirl it with a spoon to the right or to the left. Then you will suddenly discover that all water surface is calm and only in the center of the can there is a water-steam mixture, which is about 20% higher than the level of water surface. You will also see that water is absolutely calm at walls of the can. This experiment can be reproduced anytime and anywhere, even in space, since it is a demonstration of **vortex** fluid motion and its actual influence on everything, which is around!

All presented engineering solutions are practically reproducible in any laboratory. The authors of the article would welcome the opportunity to co-operate with science and industry representatives for realization of the ideas in real serial production devices.

References

1. O.V. Gritskevitch, S.A. Lisnyak, On possibilities of vortex energetics //New Energy Technologies. 2002, #1(4), p.2-3
2. Patent RF #2045715, 1993
3. Chemical and Engineering News. 2000 #1, p.26

Investigation of Electric Energy Transmission Processes in non-Metallic Conducting Channels

D.S. Strebkov, A.I. Nekrasov, S.V. Avraamenko

All-Russian scientific Research Institute of Electrification
of Agricultural Economy
2 Vishnyakovsky passage, 1 Moscow, 109456, Russia,
energy@viesh.msk.su

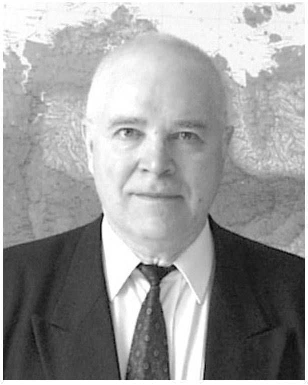

D.S. Strebkov

(Editor's notes by Alexander V. Frolov)

It has been found that at low frequency (1-25 kHz and higher) electric power can be transmitted with low losses from generator to receiver along single channel made of non-metallic conductive media such as water in plastic tube carbon thread, layer of damp soil, ITO films on glass substrate, laser and electronic beams. Transmitted power, as well as for traditional three phase lines, is limited by natural power of the transmission line and line capacity may reach at high voltage and pulse and operation modes the value 10^9 Wtt.

Introduction

The well-known methods of electric power transmission are based on transmission of active energy by means of conductivity currents in closed circuit. Electromagnetic energy spreads along power transmission lines (PTL) as progressing waves of electromagnetic field or field of charge [1]. Line wires made of aluminum or copper are conductive (guide) channels. Electromagnetic energy stream moves along these channels from generator to energy receiver and backwards to the generator. Maximum transmission possibility of 3-phase PTLs is limited by losses on the line resistance, by peak voltage (which is determined by electric strength of the insulation) and by electromagnetic stability of the line.

The modern approach to provide the electromagnetic stability consists in rigid regulation of line parameters by means of high-speed shunt reactor and consequence capacitive compensation for the purpose to except changes of electromagnetic power flows and to suppress resonant properties of a line [2].

In Tesla works [3] and in the researches of Russian scientists [4] a method of active power transfer was offered. This method supposes to transfer active power by means of electromagnetic capacitive current assisting with resonant properties of a single-wire line (SWL), made of a metal conductor. The purpose of the present work is a research of an opportunity to use non-metal conducting mediums for transmission of electric energy.

Methods

For realization of the experiments a single-wire energy system (SWES) was used. Its electric circuit is shown in Fig. 1 a,b. SWES consists of the following parts: high-frequency generator (1) of 28V voltage AC and 1 kWtt power; transmitting (2) and receiving (3) Tesla transformer with conductive channel (4) between them; rectifier (5) and electric load (7) as an incandescent lamp or electric motor (220V, 1 kWtt power). High-voltage winding of Tesla transformer is made in the form of cylindrical winding on the ferrite core with 50-100mm in diameter, 4000-6000 turns. The internal end of a high-voltage winding is connected to the conducting channel, and the external end is not connected at all (free end). A low-voltage winding that consists of 40-60 turns provides power supply of Tesla transformer. At the submission of electric power from the high-frequency generator to the winding of power supply zero potential appears on the free end of the high-voltage winding, and voltage with corresponding to the generator 1-25 kHz frequency is formed on the conducting channel. Besides, Tesla transformer as the spiral antenna generates electromagnetic waves of 5-10 cm length , which corresponds to the turn length of a high-voltage winding. Current resonance mode is formed in a supply circuit of Tesla transformer. At the same time, voltage resonance mode (of frequency equal to the generator frequency) is formed both in the reception and transmitting circuit of high-voltage windings and in the conducting channel (1).

The following materials were used as the conducting channels: filled with water or sea water polythene tubes with 10 mm diameter and 1,5 m long; plastic tray with soil layer of 150×10×400mm; film of tin dioxide – indium oxide (ITO) on 0,3 micrometer thick and 300 Ohm resistant glass substrate; graphite thread of 0,1 mm diameter, 500 mm long and with 100 Ohm resistance. For comparison steel and copper wire of 0,1 mm diameter and 5 m long was used also.

The voltage on the conducting channel changed in limits 1-10 kV, generator frequency changed from 1 up to 25 kHz. Voltage, current and capacity were measured on SWES output and on the load by standard electric measurement equipment.

Results and discussion

The replacement of the metal conductive channel to the channel made of non-metallic conductive materials does not effect to any reduction of possibilities of transmitting SWES or heating of a material of conducting channels. The consecutive connection of the channels did not resulted in reduction of transmitted power. Circuits break in the conducting water channel by the creation of an air gap led to the occurrence of arc discharge of reactive capacitor current. However this discharge did not cause increase of water temperature at transmitted power 300 Wtt and voltage 4,5 kV within 1 hour, that confirms the absence of energy losses in the conducting channel. The increase of water temperature did not result in decrease of transmitted power. Water PH reduction from neutral value up to 4 was detected.

The increase of sea salt concentration in water up to the level of 5-7 g/l did not increase transmitted power in comparison with tap water. However the replacement of tap water to dionized water with 16 MOhm resistances resulted in 100 % decrease of transmitted power.

Thus, it is experimentally shown, that conducting channels, produced from non-metal materials, have quasisuper-conductive properties in SWES at resonant mode. Possible explanation of this effect is the absence of active conduction current in the channel and the dominant role in the process of energy transfer belongs to displacement current, for which Joule Lentz law is not valid [11]. In the supply circuit of Tesla transformer current is practically reactive, and in resonance conditions active values of inductive and capacitive currents are equal. Vectors of these currents are opposite in their phases. Current of the high-frequency converter is spent for losses (component less than 2 %) in supply circuit wires and in the core of Tesla transformer and also for creation of reactive current in the conducting channel. In the mode of voltage resonance we have measured voltage active values of high-voltage inductance windings and conducting channel, interturn capacitance of windings and the capacities of the conducting channel. **We have discovered that they were equal and their phases were opposite to each other.** Losses from transmission of capacitive current through active resistance of the conducting channel are insignificantly small. Corona discharged losses and leakage current could be decreased by isolation of conducting channel. In this case active current and magnetic field of the line are equal to zero. Electric field of the line has maximum value. As well as in usual PTLs, maximum transferred power is limited by charge power of the line. Angle between vectors of voltage is equal to zero in the beginning and at the end of a line. Quality factor of SWES at frequency 5 kHz is in 100 times above than usual PTLs at frequency 50 kHz. In the conditions of resonance it leads to the significant increase of voltage along the conducting channel and it also leads to transmission power.

In usual PTLs voltage changes along the line are insignificant. The angle between vectors of voltage in the beginning and in the end of PTL constitutes the value, which is proportional to the wavelength of line.

On the basis of the researches the methods and devices are offered for transmission of electrical energy through plastic water guide, electro-insulated from ground, through irrigational channels and through isolated pipelines, which are used for gas, oil, hot and cold water transportation. Also the energy can be transferred through fiber-optic cable with conducting film on the surface, through all-carbon composite cable and through electro-insulated part of a ground and water surface, including highway parts. There are also methods and devices, designed for the following transmissions. These methods can be applied for transfer of power to stationary and mobile units. There are also generated requirements to electrical safety and to restriction of use of drinking and hot water from pipelines, which are under electrical voltage. These requirements and restrictions are generally the grounding of pipeline parts, which are located on certain distance from the generator. This distance is equal to the whole number of half-waves and for it SWES voltage is equal to zero. In the case of a side pipeline it is necessary to ground the parts of the pipeline that are placed on the distance of odd number of quarter-wavelength from the main pipeline. For 5kHz frequency the quarter-wavelength is equal to 15000 m.

N. Tesla grounded one end of high-voltage windings of his transformer on the receiving and on the transmitting end of SWES. Tesla considered this condition as the necessary one for the transfer of power along the Earth. The results of our researchers demonstrate that it is not necessary to use metal self-closed conductor (and current lines in the Earth) for transmission of electric energy on low frequency (1-25 kHz).

For this frequency energy could be transferred from the generator to the receiver if we have single-wire guide system created as non-metal conducting channel. By the similar way electromagnetic energy is transferred by laser beam or microwave-beam. But in our case we can obtain high degree of efficiency that is caused by slight losses on energy absorption and energy emanation. Thus one of the ends of high-voltage winding at the energy generator will have zero potential and remain free. The symmetric end of a high-voltage winding on the reception end should be connected to some natural capacity 6 (Fig. a), which can represent the case of a balloon or frame of a tractor. In our experiments we used metallic safe-box as such natural capacity.

Editor's note: In 1887, October 11, the famous Russian scientist Pavel Yablotchkov got the France Patent #120684, which described the method to increase efficiency of electrical circuit by means of "atmosphere electricity". It was confirmed in many experiments that output power can be twice more than power provided for the circuit from primary electric generator. For that it is necessary to use a single-plate capacitor. The special feature by Yablotchkov, that provides maximum efficiency, is the high degree of air ionization. For this Yablotchkov proposed to use the special capacitor, which consists of a big number of metallic needles. By its view this construction reminds of a hedgehog. Thus we can assumed; that above described natural capacity (balloon or frame of a tractor) serves as a collector of free electrons. By Yablotchkov the efficiency of such systems can be increased by means of maximization of ionization process. It will not lead to the increase of losses if the second end of the high-voltage winding is not grounded.

In the other method of energy transfer, a condenser-diode block 8 was connected to the conducting channel on the receiver end. This block is one of the known circuit of voltage doubling, Fig.1 (b). On the condenser 8 electric energy is transferred through electronic switch 9 to load 7. In this case the entire length of conducting channel 4 and Tesla transformer winding 2 at generator must be equal to odd number of quarter wave-lengths.

Non-metal conducting channel (for instance, fiber-optic or coal-plastic cable) can be used for transfer of electric energy not only along but also as perpendicular to the Earth (for example, to relay aerostat or sounding balloon).

SWES conducting channel can be also created by ionization of air ions with laser beam [13]. Neodymium laser with double frequency and with energy 1 Joule in impulse is able to create 10^{15} cm^{-3} ion concentration in air. This concentration is sufficient for streamer initiation and for transfer of electric energy through the conducting channel. Ionization potential, time of ion existence and of excited molecule state, coefficient of multiquantum absorption, all these determine the limiting length of conducting channel in atmosphere that is equal to 300 km and its wave resistance at 200-400 Ohm. Voltage which is necessary for SWES comes to the quantity 0,5 MV – 15 MV, that depends on the length of a channel.

We suggest to use relativistic electron bunches of high energy as the conducting channel out of the atmosphere. As distinct from laser bunches they do not have divergency. In this connection the Moon or artificial conducting body, where the energy receiver is placed, can be used as natural capacity 6. Whereas energy generator can be installed on the Earth or on its satellite. Transmission range of electric energy is determined by the length of the generated conducting channel. The entire length of the conducting channel in the beginning and in the end must be equal to the whole number of half-waves. Here the length of high-voltage windings of two Tesla transformers must be taken in considerations. Electric energy, transmitting through the conducting channel, can exceed the energy of electron and laser beams generators in 10-100 thousand times. These generators play the role of a directing system (of usual SWEG wires), along which the transfer of electric energy proceeds.

It was offered to use colliding and crossing electron and laser beams with conducting transitional bodies as conducting channels for transfer of energy from the Space to the Earth and back. On the heights up to 30 km compositional coaly and fiber-optic cables can be used. To create the global energy system of the Earth it was also suggested to apply single-wire energy system and conducting layers of the Earth ionosphere as the conducting spherical channel [14].

Thus for electric energy transfer at the frequency 1-25 kHz and higher in the resonance mode a single-wire channel from the following non-metal conducting mediums can be used: water, damp soil, coal-plastic, oxide film, ionized air channels, that are created by laser beams in the atmosphere, ionosphere conducting layers, and also beams of relativistic electrons out of the atmosphere. These non-metal conducting channels in the resonance mode have negligibly small resistance losses especially if to be compared with metal conductors, which are used in the known non-resonance methods of energy transfer by means of active conduction currents in the closed circuit. Electric energy in the resonance mode can be transferred with small losses from the generator to the receiver along the single-wire channel, made of non-metal conducting materials. The transfer can be realized at the frequency 1-25 kHz and higher, to any distance and to any direction relatively to the Earth. The transmission capacity is limited by charge power of a line as well as in the usual PTL. At high voltage the transmission capacity can reach the quantity from 10 Wtt to 10^9 Wtt in the pulse and streaming modes.

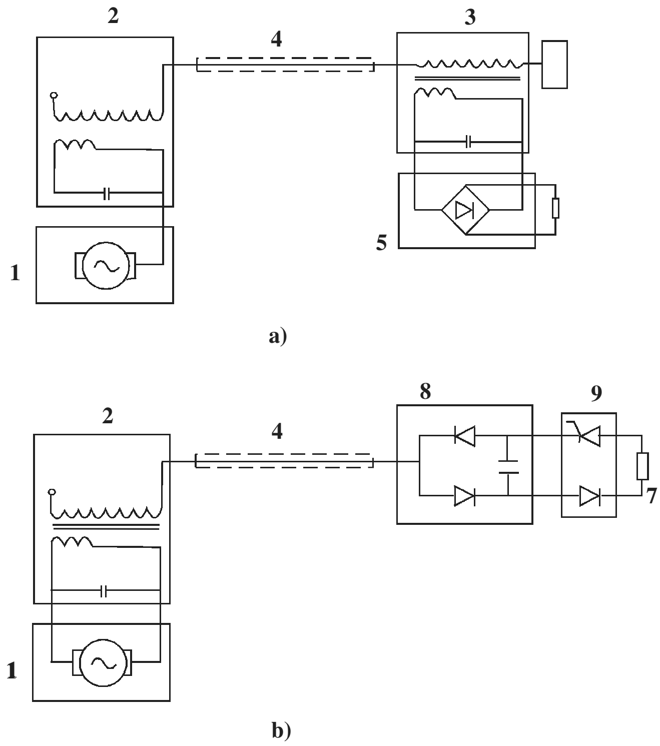

Fig. 1
Electric circuit of single-wire energetic system with non-metal conducting channel
a) SWES with symmetric array of Tesla generators

b) SWES with diode-capacitor block in the end of conducting channel

References

1. A.M.Kalashnikov, Ya.V Stepuk. Bases of radio engineering and radiolocation. Oscillating systems. M: Department of USSR Defense 1965, - p.127-158.
2. G.N.Alexandrov, S.V. Smolovic, Flexible lines for electric energy transmission over long distances. V Symposium "Electrical Engineering-2010", October 12-22, 1999, Moscow Region P., 35-42.
3. N.Tesla Apparatus for transmission of electrical energy. US Pat #349621, 15.05.1900.
4. S.V. Avraamenko Mode of electric devices supply and device for its realization. Pat. RF #210649 11.04.1995. Publ. 10.04.1998. Bulletin #10.
5. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Device for power supply of mobile electrical aggregation. Pat. RF #2158206, 15.04.1999. Publ. 27.10.2000. Bulletin #30.
6. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Mode of supply of electric transport devices and device for its realization. Pat. RF #2136515, 26.08.1998. Publ. 10.09.1999. Bulletin #25.
7. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov One-trolley system of power supply of mobile electrical aggregations. Materials of the international scientific and technical conference on automation of agricultural manufacture. Minsk 6-8 July 2000, - p. 65-66.
8. D.S.Strebkov, S.V.Avraamenko, A.I. Nekrasov Research on single-wire system of electric energy transmission. International conference Ecology and agriculture. Technique. – Saint-Petersburg, 2000, - p.50-55.
9. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Single-wire electric system for renewable-based electric grid. Third European Solar Congress "Eurosun 2000", June 19-22, 2000, Copenhagen, Denmark p., 10.
10. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. SWEPS – a novel electric power transmission technology. Planetary Association for Clean Energy Newsletter, 2001. – vol. 11, - 3, - p. 7-11.
11. I.E. Tamm. Bases of electricity theory. – M.Science, 1976. – p. 133, 397-400.
12. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2172546, 24.01.2000. Publ. 20.08.2001. Bulletin #23.
13. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2143775, 25.03.1999. Publ. 27.12.1999. Bulletin #36.
14. D.S. Strebkov, S.V. Avraamenko, A.I. Nekrasov. Mode and device for electric power transmission. Pat. RF #2161850, 14.07.1999. Publ. 10.01.2001. Bulletin #1.

Microcosm, Universe, Life

Alexey G. Shlienov
Budapeshtskaya, 66-1-77,Saint Petersburg, 192286 Russia

Ernest L. Petrov
Marshala Zaharova, 50-809, Saint Petersburg, 198330 Russia,
Fax.: (812) 587-9349

The authors of this article have prepared the 3rd edition of the book "Microcosm, Universe, Life" and now they are looking for a publisher to edit Russian and English version. In particular, the mechanism of energy transformation in the Universe (Fig.1) is described in the book. Matter radiates energy in the form of photons (as well as neutrinos). In general this energy is observed in the form of Cosmic Microwave Background Radiation (CMBR). Photons and neutrinos transfer the most of energy to de Broglie longitudinal photons. Matter absorbs energy mainly in the form of longitudinal photons. Thus, a cycling process of energy transformation (as well as matter transformation) occurs in the Stable Universe.

The 3rd edition is devoted to the description of 6 World Systems (Table 1), at that the 6th System is based on Unified Field Theory, which is developed by the authors, and proceeds from the following:

(Editor's: The understanding of the energy transformation mechanism allows to develop new fuel-less energy sources based of mutual transformations: longitudinal photons⟷transverse photons).

Table 1

System	Principal Ideas
The 1st – Geocentric (Ptolemaic system)	Geocentrism and Anthropocentrism
The 2nd – Classic (Copernicus – Newton's system)	Description of the Universe on basis of Newton's Law of Gravity
The 3rd – Compromise (Tycho Brahe's system)	The compromises between two first systems
The 4th – Fractal (Charlie's system)	Non-heterogeneity of Large-Scale distribution of matter in the Universe
The 5th – Relativistic (Einsteinian system)	Description of the Universe on basis of gravitational field equations. Anthrop principle
The 6th – based on Unified Field Theory	Electromagnetic nature of all physical interactions. Irreversibility of all elementary micro-processes in combination with circular character of transformations in Large Scale of space ($c/H=R$ order) and time ($t_H=I/H$ order).

1. For theoretical results, which agree with the facts, it is necessary to use assigned inertial system (as Lorentz did), which, as we have known, is connected with CMBR. Obtained by this way decisions can be applied to the other systems, in particular, with use of Lorentz transformation for mass, energy, momentum, time and length.

2. Theory must agree with the whole known collection of facts, starting with such established empirical generalizations as basic principles of Natural science (Giordano Bruno, Lyell) and conservation laws (Mayer, Joule, Helmholtz, Faraday, Newton, Huygens, Vernadsky).

3. All physical objects could be presented as the systems, consisting of quanta of positive electric charge (protons), quanta of negative electric charge (electrons in fermions or boson states, which are part of neutrons

and atomic nuclei) and quanta of energy, impulse, moment of momentum (apeirons).

4. The leading role in the Universe energetic belongs to de Broglie longitudinal photons (of the first class, i.e. with the spin I=0), energy of which is twice as much as apeiron energy and is equal to *h H*, where *h* – is Plank's constant, *H* – is Hubble's constant.

5. Longitudinal photons, polarizable and depolarizable by matter particles, are the carriers of gravitational, magnetostatic and strong interaction. Electromagnetic interaction is carried by photons with spin I=±1 and weak interaction is carried by neutrinos and apeirons with spin I=±0.5.

6. All elementary microprocesses are irreversible, that correlates with circular character of transformations in Large Scale of space and time.

7. In particular, there is an irreversibility of the process of photons and neutrinos motion. At each segment, equal to the length of de Broglie wave, photons and neutrinos lose energy *h H*, which is equal to longitudinal photon energy. In such a way, Hubble's law can be presented as the following:

$$\lambda = \lambda_1 \exp\left(\frac{r}{R}\right) = \lambda_1 \exp(Ht),$$

$$Z = \frac{\lambda}{\lambda_1} - 1 = \exp\left(\frac{r}{R}\right) - 1 = \exp(Ht) - 1,$$

where λ, λ_1 – are observable and laboratory wave lengths, *r* – is distance, *R* – is radius of gravitation interaction, which is equal to the radius of Metagalaxy, *t* – is time, *Z* – is cosmological red shift.

8. As resulted upon this process, the excess of longitudinal photons is absorbed by matter. At that, mass is considered as measure of capacity, which is absorbed by matter in form of longitudinal photons. Thus, energy of the matter should be increased according to the law:

$$E(t) = m_\circ c^2 \exp\left(\frac{\varepsilon'}{c^2} t\right) \approx m_\circ c^2 + m_\circ \varepsilon' t,$$

at that,

$$\frac{\varepsilon'}{c^2} = \frac{H}{137^2},$$

where ε'- is the capacity, which is absorbed by unit mass in form of longitudinal photons, *c* – is velocity of light. In particular, electron of the 1st Bohr orbit in hydrogen atom (i.e. in this case at de Broglie wave length) absorbs energy *h H* at 1 period.

9. Being in ionization state, intergalactic matter is the general portion of average density of matter in the Universe ρ_{av}, besides, the average value of absolute velocity of baryon component is close to c/137. Let's consider this correlation to be exactly executed.

CMBR is the most powerful cosmic radiation. Its specter is close to that one of black body at temperature T_F=2.726K, i.e. its spectral density is maximal at frequency ν=160GHz. Cosmological red shift causes the increase of spectral density in radio-region (ν<160GHz) and the decrease of it in microwave region (ν>160GHz). Each of these processes is compensated by the inverse Compton effect, i.e. by the dispersion of radio-photons at matter corpuscle, first of all at protons. Calculations demonstrate, that average energy of absorbent radio-photons is equal to $0.45 \cdot 10^{-15}$erg (ν=68GHz) and average energy of radiated microwave photons is equal to $2.17 \cdot 10^{-15}$erg (ν=330GHz). There is one re-radiated microwave photon per one absorbed radio-photon. At that, the concentration of photons and CMBR spectrum remain unchangeable. Syunyaev and Zeldovich concerned the close inverse Compton-effect of CMBR dispersion on electrons in clusters of galaxies. Actually such an effect was discovered in 2 clusters of galaxies.

Thus, on transferring of energy to photons, matter corpuscles must fill the deficiency of energy by receiving it from longitudinal photons. In fact, there are observed demonstrations of longitudinal photons, they are "static fields". In this case it is a cosmic magnetic field, which accelerates charged particles of matter (Alfven). More detailed consideration let us to find a virial correlation between 4 main components of energy density:

$$\frac{\rho_{av} c^2}{137^2} = E_F = 2E_M = 2E_{kin},$$

where $r_{av}c^2$- is an energy equivalent of mass density of matter; E_F- is energy density of CMBR; E_M - is average energy density of magnetic field; E_{kin} - is average density of kinetic energy. Thus, some kinds of energy circularly transfer to another, which are interrelated (Fig. 1).

$$\frac{\varepsilon_{av}\rho_{av}V}{1.26} = E_F HV = 2E_M HV = 2E_{kin} HV = $$
$$= \frac{(\varepsilon_{av} - \varepsilon')\rho_{av}V}{0.26} = \varepsilon'\rho_{av}V,$$

where ε_{av} - is average capacity, radiated by unit mass in form of photons; *V* – is volume of 10^{81} cm^3 order, according to which the averaging is made.

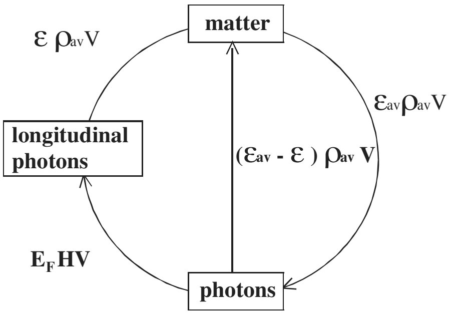

Fig.1
Energy transformation in Large Scale ε_{av}=0.0942 erg gr^{-1}sec^{-1} – is average energy, which matter unit radiates in unit time; ε'=0.07476 erg gr^{-1}sec^{-1} – is energy, which matter unit absorbs in form of de Broglie longitudinal photons in unit time; $\rho_{av}=0.8730 \cdot 10^{-29}$ gr cm^{-3} – is average matter density in the Universe; $E_F = 4.18 \cdot 10^{-13}$ erg cm^{-3} – is energy density of cosmic microwave background radiation (CMBR); $V \approx 10^{81}$cm^{-3} –is volume at which the averaging is made; $H = 1.562 \cdot 10^{-18}$ Hertz.

10. These correlations include pressures, densities and temperature T_F, i.e. we have the equation of the state of intergalactic matter, and thus of Metagalaxy and the Universe. Let us present this equation in simple form:

$$E_F = \frac{\varepsilon' \rho_{av} R}{c} \approx \frac{\varepsilon_{av} \rho_{av} R}{c}$$

and let us compare it with correlation for density of radiant energy on surface of a star * or of star system

$$E_* = \frac{\varepsilon_* \rho_* R_*}{3c} \approx \frac{\varepsilon_* \rho_* R_*}{c}.$$

It is an especially amazing analogy between Metagalaxy (and the Universe) and large-scale cosmic system *g*, for which

$$E_g \approx \frac{\varepsilon_g \rho_g R_g}{c},$$

at that $E_g \approx E_F,\ \varepsilon_g \approx \varepsilon_{av} \approx \varepsilon'.$

This equation agrees with the whole observed data of globular clusters, galaxies, groups and clusters of galaxies, in particular, with empirical correlations, which are magnitude - angular diameter.

11. The stated approach lets to determine the values of many fundamental constants by different ways. As a case in point, the results of definite values ε_{av} and ε', erg g^{-1} c^{-1} are presented in the Table 2.

Table 2

#	Data	ε_{av}	ε'
1.	The boarder between two parts of the Main Sequence of stars	of 0.1 order	of 0.1 order
2.	The results of recalculation of observed star characteristics as respects to their centers	of 0.1 order	of 0.1 order
3.	The local minimum of star luminosity function near the Sun (according to G.A. Starikova's data)	of 0.1 order	of 0.1 order
4.	The correlation: mass - luminosity of white dwarf stars	of 0.1 order	of 0.1 order
5.	The correlation: mass - luminosity of neutron stars	of 0.1 order	of 0.1 order
6.	The correlation: mass - luminosity of globular clusters	less than 0.5	less than 0.5
7.	The correlation: mass - lumonosity of elliptic galaxies	more than 0.06	more than 0.06
8.	The correlation: mass - luminosity of spiral galaxies and irregular galaxies *Ir I*	less than 0.5	less than 0.5
9.	The correlation: mass - luminosity of galaxies as a whole	of 0.1 order	of 0.1 order
10.	The correlation: mass – luminosity of clusters of galaxies	of 0.1 order	of 0.1 order
11.	Empiric values of H, ρ_{av}, E_F constants	of 0.1 order	of 0.1 order
12.	The most exact values of h, e, G, E_F constants	0.0942	0.07476

12. Uncontradictory description of Microcosm and the Universe promotes the better understanding of Life, which is inseparably unified with them. The statistical data manipulation of more than 100 catalogues of microphysics and cosmic objects let us to get more than 1000 empirical correlations and diagrams, and to determine, that they agree with theoretical correlations, which were received according to the ideas on circular character of energy transformations (Fig. 1) and of matter in the Universe.

The Russian edition of the book is mailed out: the 1st edition (1995) at the cost of $3; the 2nd edition (1998) at the cost of $6.

The authors are thankful for valuable discussions to Yaroslav G. Klyushin and to Alexander V. Frolov.

HYPOTHESIS OF A THEORY OF EVERYTHING

Jack P. Gibson

www.crikcritter@infoave.net

This work grew out of an interest in curious occurrences, physical, mathematical and metaphysical. Many things just do not fit normal knowledge or are of unknown origin. Items such as gravity, time, mathematical equations of higher order, ghosts, e.s.p., etc. After a number of years, the problem began to appear to be not of this physical dimension but of a grander scale. A multi-dimensional space-time format seems to offer solutions to understanding most of the problems. In the following hypothesis all known physics and mathematics are considered to be valid. Metaphysical aspects, while not accounted for by physics and mathematics, are considered to be present in nature, either real or as a defect in physical processes, simply because they are experienced by so many people. The hypothesis will be kept simple because the details require the work of many specialists beyond my abilities and there is an endless list of speculative problems to which the hypothesis could be applied. To save time and distractions to the reader the following is stated as fact, knowing full well it may not be so, in order to present the most concise statement of the hypothesis.

The Universe, from the smallest to largest unit, is a quantum probability based fractal illusion. All matter exists in three-dimensional space, both internal and external. Any grouping can be considered a fractal space. It is both real and imaginary at the same time acting according to Eulers equation:

$$(e\,\hat{}\,(i*Pi))+1=0$$

more useful in this instance is the general form:

$$e\,\hat{}\,ix = -1=\cos x + i \sin x = -\tan 45 \text{ deg}$$

which allows use of a variable where Pi , a constant, is used in the original. The key to understanding of multi-dimensional space time is the realization that Pi is a variable! As we know Pi is the lock on three-dimensional space. Because of its mathematical nature it can neither grow nor diminish, thereby locking in the three dimensional form of the fractal unit under consideration. Such a fractal unit is considered to be at rest. But Einstein's equation $E=mc^2$ includes speed or movement. When objects are moving relative to a fixed point, the moving object is foreshortened in the direction of movement. A circle becomes an ellipse and Pi becomes indeterminate letting the circumference fluctuate, according to its relative speed, between two and Pi in a circle of unit diameter. If one goes beyond the value Pi the Universe grows out of bounds without limit becoming unstable. If the value goes below two then the Universe blinks out. I.e., one has exceeded the value of the constant c. Since all matter moves relative to a fixed fractal point and there are infinitely many points in the Universe of constantly changing velocities, the Universe is constantly adjusting its reality to the relative speeds of its components. Thus you get at times some weird things. Since not only time changes but the structure and observable mass also, one finds such things a matter appearing and disappearing as the small units of chaos move through the Universe. Changes in the local probability structure cause matter or reality to change in the image structure of the Universe.

To delve further into the aspects of what is happening, let us consider the following. Einstein's familiar formula gives the relationships for our three-dimensional fractal space. The inverse of the formula gives the speed with which any fractal space can change, including the Universe. There is both a positive and negative component to the change. If we consider our fractal space to be positive then we may consider any other point to be negative. There are an infinite number of such points and while the effect on our space may be small from any one point; the effect adds up. Any moving or changing of mass must effect the entire Universe. Mass and inertia are tied together in such a system. Measured mass is the attempt of the mass to return to zero relative speed in its associated fractal space according to the laws of physics, i.e. return to its lowest energy state. Inertia is the same effect of the changing state of the mass from one-dimensional state to another. Both are the result of efforts to change the dimensional state of mass. The positive aspect of the square root of the ratio of energy/mass is the view we behold from our position in space while the negative is the view from the opposite position.

In addition to the view of the Universe as being made up as fractal space, all moving at different relative speeds to any other fractal space, one must consider the make up of space having at the same time dimensional space, according to Euler's formula given above. There are at any moment an infinite number of spatial dimensions, defined by the relative speed of the fractal units involved, extending throughout the Universe. Each fractal unit is constantly changing its position in dimensional space. There are thus constantly changing energy states throughout the Universe both in relative position in real space and in dimensional space. While on one hand the distances across the Universe are immense in relative space, many points across the Universe are present in the same dimension at the same moment, possibly allowing for instance travel across space without the time penalty. This results in what might be considered a tuned circuit for the Universe or its components.

If we consider the structure of the relative Universe to be stored energy analogous to magnetic fields and the dimensional fields to be capacitive, then we can plot the results on a conventional two-dimensional chart where x is the horizontal axis and the vertical is the imaginary axis. A well-known construction in electrical physics and which is known to very often follow relativity closely. That throws the negative side of i into the second quadrant and the positive into the fourth quadrant, if we consider the Universe to be reflective and x to be negative when i is positive. Euler's equation does not reflect a totally positive or negative result. One wonders at this point as to other equations where unity is set to one or higher dimensional equations where the three dimensional components are factored out leaving another component. The fourth is considered to be time but how do the remaining fit into a Universe as this hypothesis describes? The line of thought can be followed further into electrical analogy, which I leave to the reader. The important result is that there should be a resultant Q, figure of merit, of the Universe resulting in nodes, or peaks, of probability. The resultant tuned frequencies of the Universe can be calculated from known factors of the Universe. Such an effect should show nodes of reality where matter occurs.

Properly applied the above hypothesis gives speculative answers to most of the problems facing science today. Questions such as:

Is the Universe open or closed?

Both. At each fractal component of the Universe there is a separate universe each reaching limits beyond which it cannot exist. At the same time there is somewhere a top Universe that must ever remain open.

What is the structure of time?

Time results from the ever changing fractal Universe and goes forward because all justification of the Universe causes a corresponding change in reality. Time moves slower at higher relative speeds and as all time is perceived to be slower in other fractal units, changes in our time are perceived to always go forward.

What are such phenomena as ghosts, spirits, apparitions, etc?

Events such as these are fragments of probability left behind due to sudden changes in the probability of the illusion we perceive to be reality. Having little energy they are perceived but are not able to interact with the more forceful real reality. They float free in probability space interacting at random intervals with the more forceful realities of the real world.

Where is the missing matter of the Universe?

There may be clouds of matter floating about in deep space, however, most of the approximately 90% of matter missing will be found in the layers of dimensions making up space and which, though interacting, do not appear in our reality. Calculations show only about ten dimensions are active in any one reality. The rest curl up in a ball or knot. While an essential part of space or reality they are not perceivable from the fractal Universe of which they are a part.

Parapsychology?

Events of this type can be attributed to interconnections on the dimensional, or imaginary, level where there can be interactions between fractal units though not directly associated with each other in relative space. The manifestation varies or is temporary due to the constantly changing and adjusting due to interactions of the probabilities of the Universe. Due to the innate probabilities of each individual, one may be more susceptible or sensitive than other individuals.

Bible, prophesies, angels, demons, aliens, etc.

If one takes the Bible at face value, with some leeway for it's age and many translations, it pretty well describes what it purports to describe. Historic predictions and conditions of today. If we assume this hypothesis has any value then the passage in which God says "Let us make man in our image." takes on new meaning. If we assume the probability basis of the Universe and the constant changing due to justification of reality and time then the miracles listed do not seem so in violation of the physical world. The miracles only require some ability to control reality. Certainly well within the providence of a creator. It speaks of heavenly beings capable of traveling through the Universe in real time and some who interfered with mankind on the Earth and that they are still doing so today. If this hypothesis is correct then it is readily seen how such events could occur. Creation events closely compare with the way todays computers are made and organized. If man can do it, why could not the original creator? If one ignores a creator, then one is hard pressed for an answer as to how the Universe was created.

UFOs

If one accepts any part of the above then the functioning of UFOs begins to be understandable. They move through space by making spatial jumps through fractal nodes. Their appearing around magnetic and electrical sources maybe due to some, on our part unrealized, easy entry and exit to other space at those points. They are able to make right angle turns at high speeds because the speed is only from our viewpoint. From the viewpoint of the craft it is simply changing locations in space. The turning on edge may be only a different orientation of space at that point.

Particles "Out of the Vacuum"

As we shoot atomic matter at higher and higher speeds into nuclei we are able to fracture the construction of the building blocks of matter. The high speed fractal construction of probabilities of the unit used smashes into the target with enough speed to mix the probabilities resulting in new nodes from the total probabilities present. Various probabilities are ejected resulting in short lived particles in unstable nodes. At times fragments of probabilities appear out of nowhere as they form up to combine into larger nodes. Travel distance and speed observed may give a clue to the basic frequency of the Universe.

Black Holes

The midpoint of the tuned circuit of the relative structure and dimensional structure of the Universe. As matter spirals in, from the relative structure, to be torn apart into its basic probabilities it passes through to the dimensional structure of space which holds 90% of the matter. There it is available for recollection and use by the relative space.

Probability. Basic structure of the energy of space.

Everything operates at random but within preferred patterns set by the frequencies involved. If you look long enough anything can be found. As a result we find odd bones of prehistoric creatures that may have never existed. Weird things can happen. Odd pieces of structure can occur such as metaphysical events. The Q of space is sometimes sharp and sometimes rounded, spreading out to include things not really meant to be in the overall scheme of the Universe. Christ said "If you had the faith of a mustard seed, you could say to the mountain move and the mountain would move". (For those not of the Christian faith, this may be considered not as any proof but as a representative statement of the meaning of this paragraph.) That is in keeping with the quantum statements that nothing exists until it is seen and that we have some control over the probabilities. If the structure of space is based on probability then the next important question is 'What is a probability'. We look to probability as a simple mathematical work. But, what makes probability work?

• •

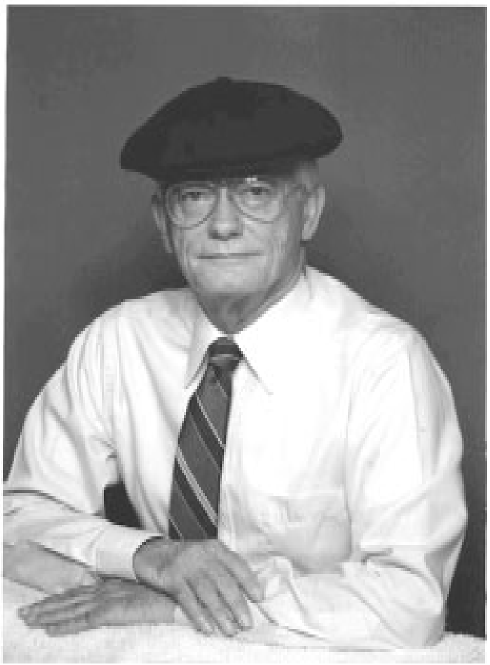

JACK P. GIBSON

About the Author:

EDUCATION: University of North Carolina at Raleigh (Nuclear & Electrical Engineering), Air University, Gunter Air Force Base (Radio Fundamentals, Motorized Equipment) Western Electric Co Engineering Center

The entire Universe is based on mathematics but we have no idea what is underneath the mathematics. Are dimensional planes perhaps better represented by other number bases? Are some of the unsolvable mathematical problems solvable in other number bases?

Gravity

All forces in the Universe are the same. Gravity happens to be the one that works on our fractal space. Others, molecular, atomic, and nuclear are the same but work in different fractal space. All are the result of matter trying to reach its lowest potential.

The above are only random questions chosen without any order. Most other problems can in some way be answered in line with the hypothesis, if not in detail. Too many questions can be fitted to the hypothesis not to take a serious look at it even though it may seem a little far out. The number of answers from one simple statement of the structure of the Universe defends the idea better than details.

One possible proof, and a relatively simple one, occurs to mind. Einstein's theory of relativity was proved when NASA flew a clock in space and then determined that time did slow down. To test the above hypothesis I would like to see four clocks used in the following manner.

Four atomic clocks in sets of two each. Two to be left on the Earth and two to be placed in space for an appropriate time. At the end of that time one of the clocks on the Earth to be taken into space and compared to the two already in space. Bring one of the clocks in space down to the Earth and compare to the clock left on the Earth.

The clock taken to space should read slower compared to the clocks in space because the ones in space were at rest compared to the one on the Earth.

The clock brought back from space should read slower, the same as the first one flown by NASA, because the clock on the Earth was at rest compared with the one in space.

The two clocks left in space should be returned after an appropriate time and compared to the ones on the Earth. The original in space should now be slower than the original on the Earth and its mate from space. The fourth clock is unpredictable.

The author apologizes for any errors in theory, mathematics, etc. There was no one to consult on the hypothesis which is an original work of my own and drawn on many references of others from the past If it has any value I am indebted to the work of all the others whose work I relied on. Any errors are entirely my own and not attributable to others. This work is freely published in the public domain to be used by any and all who wish to do so. It is not to be copyrighted or patented in any manner so as to restrict others rights to the hypothesis or it's use.

The Charge and Mass of a Photon

Dzabrail Kh. Baziev

Sirenevy blv., 65 building 4, apt.55
105484, Moscow, Russia
(095) 461-9398, (866-22)-5-10-11

Abstract

The discovery of true elementary particle named **electrino**, which has a constant mass and constant positive charge and which is a material carrier of magnetic field, electric current and all kinds of emanation, was firstly practically applied at the modernization of wireless telephone. In the review V. Anpilogov writes: "The question on the influence of low intensity microwave radiation on human health still remains open for discussion for more than 50 years" [3]. Discussion on this question has been already inappropriate still it was definitely proved in the patent application "Device, which is made to put away the charged particles flux from the head of the user of mobile wireless phone". During the talk on mobile wirelessphone $2{,}6\cdot 10^{22}$ particle/sec pass through user's brain, whereas their energy is $7{,}65\cdot 10^{-19}$ Joule, and the energy of binding of molecules in protein polymers ranges at $4{,}3\text{-}7{,}6\cdot 10^{-19}$ Joule according to L. Poling. From that we can conclude that microwave radiation of the radiotelephone with top antenna leads to the destruction of user's brain tissue. The author of the patent application D. Baziev (#2001105456, 28.02.01. in Russia and #PCT/RU02/00054 of the international application) offers to fix the antenna on the low part of the radiotelephone and to produce radiotelephones with minimal length within 10 cm. In this case the diagram of the directional radiation pattern is on the level of the user's chin and the brain is out of the direct exposure area. This harmless of the microwave radiation could not be proved before the discovery of "electrino".

Introduction

A systems analysis of all experimental and observation material gathered in physics, astronomy, and astrophysics from the times of Galileo let us reveal the following:

1. Experimental materials do not agree with the existing physical theory.
2. There is a certain fundamental disadvantage of the experimental material, which prevents to build a consistent theory.
3. This fundamental disadvantage consists in the absence of a charge antipode of electron, which is in the form of a true elementary particle with a positive charge and finite mass.
4. The proton and positron are not true elementary particles and neither of them can be a charge antipode of the electron because they are subjected to splitting.
5. Discovery of the second true elementary particle with a positive charge could restore the charge symmetry in physics, thus leading to a radical revision of the existing theoretical physics and resolving its current crisis state.

Searching for this particle required to ascertain physical nature of Planck's constant. This became possible only after the structure of a light beam had been understood. Namely, it was the photon sector velocity, known as Millikan constant μ, rather than the speed of light *c*, that proved to be a constant, viz.:

$$\mu = \lambda_i^2 \nu_i = 119.916984 \text{ m}^2/\text{s} = \text{constant}, \qquad (1)$$

where λ_i and ν_i are the wavelength and frequency of the *ion* monochromatic beam in the light beam.

This new quantity elucidated the physical nature of Planck's constant:

$$h = m_\varepsilon \mu \frac{\sqrt[3]{4\pi/3}}{2} = \\ = 6.6262681\cdot 10^{-34} \text{ kg m}^2/\text{s} = \text{constant}, \qquad (2)$$

where m_ε is the mass of the second (after electron) true elementary particle to be called "electrino". From this expression we have

$$m_\varepsilon = \frac{2h}{\mu\sqrt[3]{4\pi/3}} = \\ = 6.85575729963\cdot 10^{-36} \text{ kg} = \text{constant}. \qquad (3)$$

The electrino has a positive charge ε determined by

$$\varepsilon = \frac{m_\varepsilon n_\varepsilon e}{m_u - n_e m_e} = \frac{-3.229526609098\cdot 10^{-54}}{1.6578584539\cdot 10^{-27}} = \\ = 1.98764431671\cdot 10^{-27} \text{ C}, \qquad (4)$$

where $m_u = 1.66057 \times 10^{-27}$ kg is the mass of an elementary atom accepted as a mass equivalent of one atomic unit; $n_e = 3$ is the number of electrons in one elementary atom; $e = 1.6021892 \times 10^{-19}$ C is the charge of an electron; $m_e = 9.038487 \times 10^{-31}$ kg is an improved value of electron mass; $n_\varepsilon = 2.418198867 \times 10^8$ is the number of electrinos in an elementary atom.

Thus, it is obvious that Planck's constant is the angular momentum of the electrino. Moreover, it was Planck's constant that concealed the second true elementary particle, which is the charge antipode of the electron discovered by J.J. Thompson as far back as in 1897.

The solution of Planck's constant has become a basis for the synthesis of the new theory of physics [1]. This theory in particular shows that the electrino is the carrier of the magnetic field and electrical current. It is a photon of radiation of all ranges, and serves as a universal carrier of energy and information. The electrino plays the role of a neutrino in moving along the first order trajectories.

The first experiment

An extraordinary importance and novelty of the new theory required **an experimental proof of the electrino**. For that several experiments were made in the Institute of General and Inorganic Chemistry, Moscow. The experiment was based on the following effects predicted by the theory.

1. If assume that electrino exists and that light beam is a flux of particles having positive charge and finite mass [1] we can conclude the following. At the discharge of the dc source through an incandescent lamp in which the current is converted to light and irreversibly emitted, the source weight in charged state must differ from its weight in discharged state. If we prove this difference experimentally we may say that light does consist of material particles of finite mass and a dc charge carried away by light is positive because an incandescent lamp (W=15 Wtt) does not emit electrons, which are the carriers of negative charge.
2. The second effect to prove was that the weight of a discharging dc source is *increasing* whereas its weight when charged is *decreasing*.

To prove the validity of these predictions, several sealed containers with different dc sources inside were fabricated. The electrodes were brought out through glass insulators. The batteries were discharged through an electric lamp radiating in the visual and infrared ranges. The weight of containers was measured before and after discharge process with accuracy $\Delta W=\pm 0{,}02$ mg; balance error was equal to $\Delta = \pm\ 0{,}05$ mg the standard deviation of the measurements was within $\sigma = \pm 0{,}03$ mg; the buoyancy was calculated for each measurement of weight. In this paper, we present test results of only one container with four generally marketed GP rechargeable cells connected in series. The total battery voltage reached 5400 mV at 6000 mA/h charge capacity. The discharge was interrupted when the voltage dropped to 4000 mV, the duration of the discharge was measured accurate to one second. Two series of experiments were run: one in air, the other, under argon. Each series had ten charge-discharge cycles (Table 1 and Fig. 1). The total amount of the experiments and detailed discussion of results have been summarized in a recently published brochure [2].

The results of the above tests allow us to make the following conclusions:

1. Both galvanic and rechargeable cells during a discharge through an electric lamp show sufficient changes in their weight and charge thus proving that photons have a finite mass and a positive electric charge.
2. A new elementary particle, named electrino, derived from Planck's constant in August 1982, and published in May 1994, thus gets a complete and absolute experimental confirmation.

The second experiment

One of the concepts of the new theory is that the speed of light in vacuum is a function of photon frequency along the beam axis, according to the proportions:

$$c_i = \sqrt{\mu \nu_i} \quad [\text{m/s}], \qquad (5)$$

$$\nu_i = \mu / \lambda_i^2 \quad [\text{s}^{-1}]. \qquad (6)$$

According to the new theory, for the velocity of monochromatic light (solar light or mercury-discharge lamp, but not a laser) with a wavelength of $\lambda_r=6.8\times10^{-7}$ m (mid-point of the red spectral line), we have

$$\nu_r = \mu / \lambda_r^2 = 2.59336038 \cdot 10^{14}\ \text{s}^{-1}, \qquad (7)$$

$$c_r = \sqrt{\mu \nu_r} = 1.76348505882 \cdot 10^{8}\ \text{m/s}, \qquad (8)$$

which is 58.823% of the speed c_v = 2.9979246×10^8 m/s of a violet beam with a wavelength of 4 × 10^{-7} m.

We have to account that, according to this theory, the laser beam is not a true light beam though it is created of electrinos. The speed of laser beam is equal to the speed of beam plus the speed of current in the conductor, viz.,

$$v_0 = 2.8992629 \cdot 10^{8}\ \text{m/s} = \text{const} \qquad (9)$$

If we select a monochromatic beam of ultra-violet light with a wavelength of λ_1 =4 · 10 $^{-8}$ m then its velocity will be c_1=10 C:

$$c_1 = \mu / \lambda_1 = \frac{119{,}916984 m^2/\sec}{4 \cdot 10^{-8} m} = \\ = 2{,}9979246 \cdot 10^{9} m/\sec$$

$$\nu_1 = \mu / \lambda_1^2 = \mu / 16 \cdot 10^{-16} m^2 = 7{,}4948115 \cdot 10^{16} \sec^{-1}$$

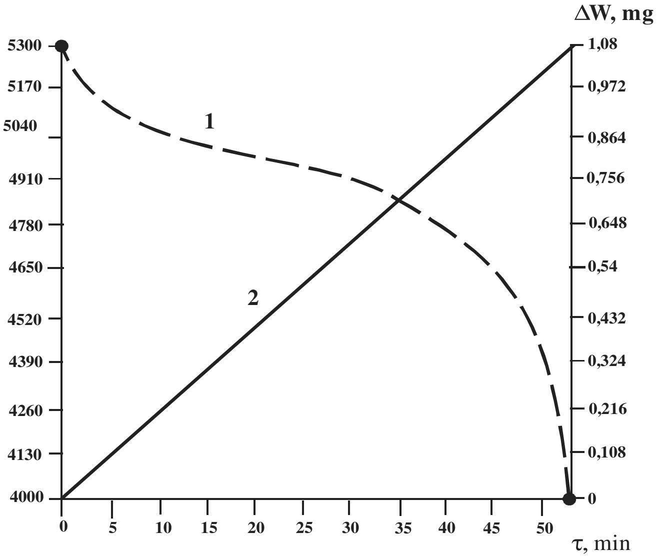

Fig. 1
Voltage drop [mV] of a battery and (2) weight increment [mg] of container #6 in an argon atmosphere during the second discharge cycle [minutes]. X-direction – is spark duration in minutes; Y-direction – is voltage of a battery; auxiliary Y-direction – is weight incensement (mg).

$$c_1 = \sqrt{\mu \cdot v_1} = \sqrt{8{,}987551907 \cdot 10^{18} m^2 / c^2} = 2{,}9979246 \cdot 10^9 m/\sec$$

This experiment has been run yet, and it is offered for experimental verification with further publication of experimental results.

Table 1

Weight of container #6 (under argon) in charge-discharge experiments

		Charged battery		Discharged battery			Charge
Run	Measured value $W \pm \sigma$, mg	Buoyancy Gmg	Real weight $W_0=W+G$	Measured value $W \pm \sigma$, mg	Buoyancy Gmg	Real weight $W_1=W+G$	weight $\Delta W_z = W_1 - W_0$, mg
1	126825.13±0,02	85.031	126910.166 ±.0.02	126825.901 ±0.01	85.002	12691.903 ±0.01	0.737
2	126825.107±0.02	86.572	126911.679 ±0.02	126826.221 ±0.01	86.538	126912.759 ±0.01	1.080
3	126825.21±0.01	86.782	126911.992 ±0.01	126826.279 ±0.01	86.560	126912.839 ±0.01	0.847
4	126825.187±0.01	86.563	126911.749 ±0.01	126826.493 ±0.02	86.385	126912.878 ±0.02	1.128
5	126825. 65±0,04	86.290	126911.941 ±0.04	126826.65 ±0.01	85.836	126912.941 ±0.04	0.770
6	126827.28±0.00	85.187	126912.467 ±0.00	126827.990 ±0.01	85.204	126913.194 ±0.01	0.727
7	---	---	---	---	---	---	---
8	126826.98±0.00	86.182	126913.162 ±0.00	12682 7.897 ±0.02	86.308	126914.205 ±0.02	1.042
9	126826.95±0.00	86.307	126913.257 ±0.00	126827.757 ±0.01	86.402	126914.159 ±0.01	0.902
10	126827.25±0.00	86.294	126913.544 ±0.00	126828.35 ±0.00	85.729	126914.079 ±0.00	0.535

References

1. D.K. Basiev, Osnovy obyedinyonnoy teorii fiziki (Foundations of Unified Theory of Physics), Pedagogika, Moscow, 1994
2. D.K. Basiev, Zaryad i massa fotona (The charge and mass of a photon), Pedagogika, Moscow, 2001
3. V.P. Anpilogov, "The century of quality", #3, 2001, p.60

● ● ● ● ● ● ● ● ● ●

About the Author

When a child Dzabrail Kh. Baziev became a political exile along with all Balkarian people; in 1956 he was rehabilitated and in 1957 graduated from the secondary school with steady purpose to become an outstanding chemist. His first higher education is biological and chemical (two-profile department of Kabardino-Balkarian State University). In 1965 Baziev also completed his post-graduate study in Biological Department of Moscow State University.

At 1979 Dz.Kh. Baziev had published 25 scientific articles on ornithology and ecology, however, this period became a sudden turn in his scientific work. In order to elaborate the basis of theoretical biology, he came to the conclusion that biological process could not be described without understanding of its physical essence. On the other hand, it had become evident that there was no any solution for physical phenomenon, which is the basis of biological process.

In 1983 the scientist gave descriptions of his first fundamental discoveries. He found the solution of the physical essence of Plank's constant and pioneered the use of it in gases analysis. It was a revolutionary step in science since the synthesis of science was realized at the fundamental level for the first time.

As the result of this considerable work there was a new interdisciplinary theory of physics. The author created new thermodynamics of real gases and new electrodynamics, which differs from maxwellian one and it considers *electrino* as the carrier of magnetic field and electric current.

Dzabrail Kh. Baziev has succeeded in the systematic analysis of the vast experimental material and discovered that ***electrino,*** **as the true elementary particle with positive charge, is able to provide the connection of physics with biology, chemistry and other parts of the fundamental science. Thus it leads us to the radical reconsideration of all conceptions, existing in natural science.**

The Homopolar Motor: A True Relativistic Engine

Jorge Guala-Valverde

Norpatagonica-R&D Dept. S.Fe 449 Neuquen, Argentina
Q8300BG1, Confluencia Tech University Neuquen, Argentina
gual@ieee.org

Pedro Mazzoni

Fundación Julio Palacios Neuquen, Argentina
fundacionjuliopalacios@usa.net

Ricardo Achilles

Confluencia Tech University Neuquen,Argentina,
RA Biosystems Neuquen, Argentina
achilles@ieee.org

This article discusses experiments, which enable the identification of the seat of mechanical forces in homopolar-machines. Authors provide a suitable variation on a recent work "The Unipolar Dynamotor: A Genuine Relational Engine" [3], where "relational" implies "absolutely relativistic". The authors' view agrees with both Weber's recognition in the 19th century of the importance of relative motion in electromagnetic phenomena [4] and Einstein's 1905 statement concerning electromagnetism [5].

The Faraday disk: a reversible engine

The essential components of the homopolar machine, first conceived by Faraday in 1832, are shown in Figure 1. A conducting disk, free to rotate in the neighborhood of a permanent magnet, is attached to the end of a shaft. A closing wire provides a conducting path between two arbitrary points of the disk. Such a device exhibits *reversible* behavior.

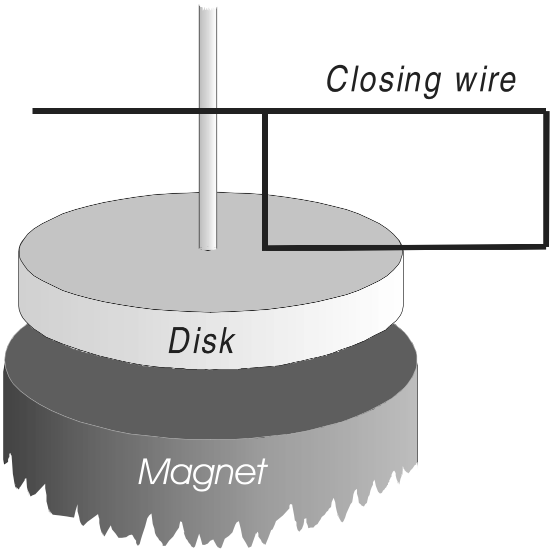

Fig.1
Faraday's setup magnet, disk and closing wire

A radial current path of length L takes place in a region of the disk when direct current (*dc*) from an external source is injected into the closing wire. The interaction of the current with the magnetic field produces a Laplace force [6]

$\boldsymbol{F} = \int_{a}^{a+L} I(d\boldsymbol{r}x\boldsymbol{B})$ causing the rotation of the disk. This set-up is the *motor configuration.*

When the disk is spun by an external source of mechanical energy, an *emf* appears in it. The displacement of free charges is produced in this case by the Lorentz force $\boldsymbol{f} = q(\boldsymbol{v}x\boldsymbol{B})$, converting the conducting disk into an *emf* source able to drive *dc* through the whole disk plus closing-wire circuit. This set-up is the *generator configuration*.

A seemingly curious fact occurs in the motor configuration, when *dc* is injected into the circuit with the disk attached to the magnet. Both disk and magnet turn together.

Two rival theories, a relativistic and an absolutistic one, have been applied to understand the observed facts:

In relativistic view, generator configuration makes sense only when there is relative motion of the magnet with respect to either the disk or the closing wire. Also, a motor configuration will only take place if the *possibility of relative motion* between magnet and either disk or closing wire is enabled.

Thus, in the relativistic framework, with the magnet attached to the disk, the closing wire becomes the "active" part for the production of mechanical forces or *emf*. In this case the disk itself behaves as a "passive" element providing a closing-circuit current path.

Conversely, in the eyes of an absolutist, a generator configuration is enabled only because of the disk or closing-wire *absolute* motion. Here, absolute means "relative to a frame where the preponderance of the mass of the universe is at rest" [7,8]. In our case, the lab frame acts as an acceptable absolute-motion reference. Thus, from an absolutistic view, the magnet's rotation with ∂ B/t = 0 in each point of the surrounding space is unable to produce an *emf* on nearby conductors. When in a motor configuration, *dc* is injected in the circuit, and the absolutist assigns the observed rotation to the magnet "dragging" by the conductor. Here, the closing wire acts as a "passive" circuit element.

New experimental work, complementary to that currently known on the subject, introduces arguments in favour of the relativistic viewpoint. The related experiments, whose underlying physics rests upon a modified version of the original Faraday setup, are described in the following sections.

The asymmetrical rotor

Figure 2 shows the disk-shaped ceramic permanent magnet creating the axial magnetic field ***B***. The removal of a 12° sector introduces a field-reversion region. Outgoing and ingoing ***B*** field lines are represented by the and symbols, respectively.

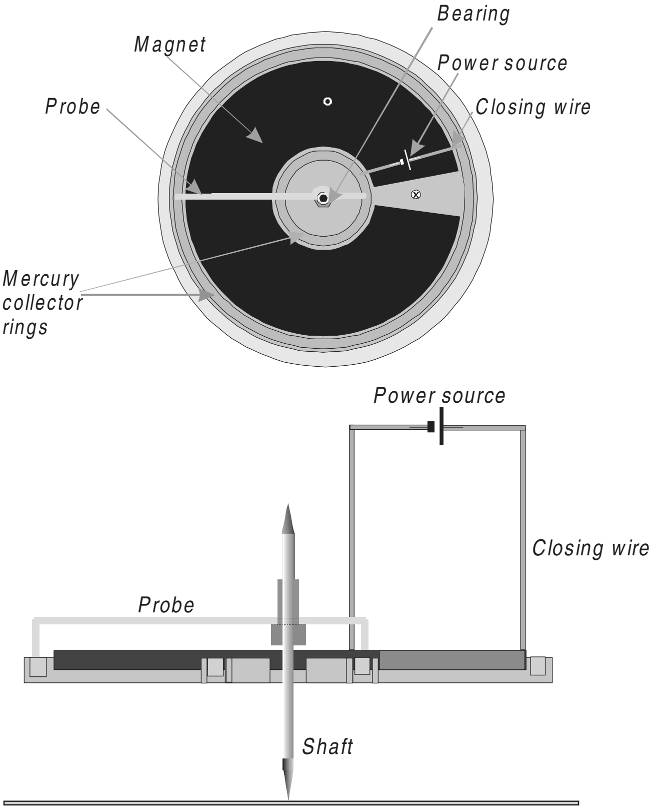

Fig.2
Layout of the Asymmetrical Rotor applied to the experiments

Two mercury collector rings are embedded in a wood cylinder. One is located close to the hollow-disk magnet inner rim and the other in the proximity of the outer rim. The magnet's inner and outer radii are 25 and 75 mm, respectively, and its height 25 mm. Its average flux density 2 mm above the magnet has been estimated to be 0.05 T based on a generator experiment with a rotating copper disk. The magnet-and-wood-cylinder body (the asymmetrical rotor from here on) is firmly anchored to a vertical shaft terminated in sharp points at both ends. While the lower one lays on a hard-polished surface, a conical bearing, enabling its almost frictionless rotation, centers the upper one.

Unlike the series-connected conductors diametrically anchored to the shaft in the Guala-Valverde case [3], only one radial conductor wire, a probe located 2 mm above the magnet's face, was considered. By mounting it on a bearing, its free rotation is permitted with its ends remaining in contact with both collector rings. A 12V lead-acid battery applied to the closing wire feeds the probe through the collector rings. In the first four experimental cases presented the closing wire remains firmly anchored to the lab. In two complementary experiments, rotation of the closing wire mounted on two shaft-centered bearings is allowed. Its behavior as a probe occurs by the injection of *dc* from an additional closing-circuit wire anchored to the lab.

Experimental

Six experiments performed are described below:

1. Rotor anchored to the lab, probe free to rotate above the magnet's upward magnetic-field region: A radially-ingoing injected *dc* in the 0.2 A range was enough to overcome conductor-bearing and mercury-wire contact friction. A net *counterclockwise rotation* of the probe took place.

2. Probe anchored to the rotor above the magnet's upward magnetic-field region, with both free to rotate: A radially-ingoing injected *dc* in the 5 A range was enough to overcome conductor-plus-rotor inertia and friction. A net *counterclockwise rotation* of the probe took place.

3. Rotor anchored to the lab, probe free to rotate above the magnet's downward magnetic-field region: A radially-ingoing injected *dc* in the 0.2 A range was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the probe took place.

4. Probe anchored to the rotor above the magnet's downward magnetic-field region, both free to rotate: A radially-ingoing injected *dc* in the 5 A range was enough to overcome conductor-plus-rotor inertia and friction. A net *counterclockwise rotation* of the probe took place.

5. Rotor anchored to the lab, closing wire free to rotate above the magnet's upward magnetic-field region: A 0.4 A *dc* injected in the inner collector ring was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the closing-wire took place.

6. Rotor anchored to the lab, closing wire free to rotate above the magnet's downward magnetic-field region: A 0.4 A *dc* injected in the inner collector ring was enough to overcome conductor-bearing and mercury-wire contacts friction. A net *clockwise rotation* of the closing-wire took place.

Discussion of results

Experiments (1) and (3) can be explained using either absolutist or relativistic viewpoints because of the coincidence of the probe motion relative to the lab with the probe motion relative to the magnet.

Experiment (2) can be explained by a trivial absolutist argument founded on a hypothetic probe "dragging effect" on the magnet. A relativistic viewpoint recognizes the "active" rotational torque on the closing wire rather than on the probe where, hinging on Newton's third law, the whole action may be split in two:

Magnet-probe. The magnet produces a counterclockwise torque on the probe, and the probe exerts an equal but opposite torque on the magnet.

Magnet-closing wire. The magnet exerts a clockwise torque on the closing wire, and the wire an equal but opposite torque on the magnet.

With the probe attached to the magnet, there is no chance for relative motion between them. Consequently, due to the action-reaction cancellation, rotation is forbidden. Conversely, with the closing wire mechanically decoupled from the magnet, relative motion of the latter is permitted. The torque exerted by the closing wire on the magnet is responsible for the observed rotation.

Experiment (4): Due to its similarity with (2) a trivial relativistic explanation is applicable to the counterclockwise torque exerted by the closing wire on the magnet. There is no known plausible absolutistic explanation for it. As quoted above, the hypothetical dragging effect would produce a clockwise rotation in this case. The consideration of the experiments (2) and (4) suffices to reject the dragging hypothesis.

Complementary experiments (5) and (6) confirm the short-range extension of the field-reversion region founded on the closing-wire *clockwise* rotation (6). Briefly speaking, the closing wire is not sensitive to the field reversion and the magnet's *counterclockwise* reaction explains at once the outcome of (4). Clearly, experiments (5) and (6) show that the torque on the closing wire is independent of its location on the magnet.

Figure 3 depicts the two rotational torques involved in (2) and (4).

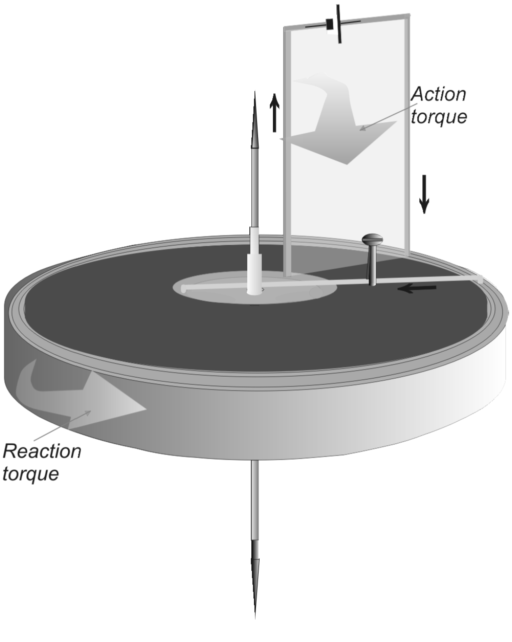

Fig.3
Rotational torques acting on the magnet and on the closing wire

Topological and miscellaneous considerations

One of the keys to the success of the above described experiments lies in the dynamotor's magnet design (see Fig.4). The short-range field reversion region allows the inversion of the Laplace force on the probe, making the force on the closing wire insensitive to that ***B**-field* reversion.

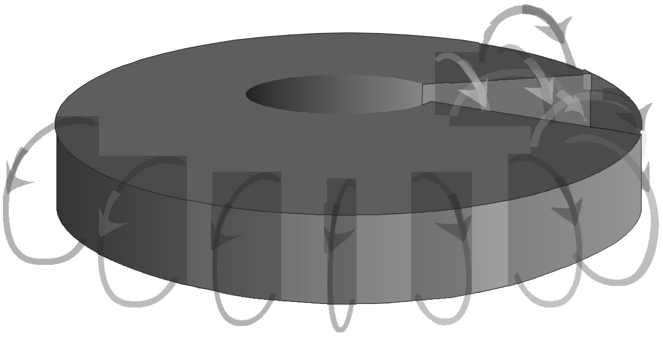

Fig. 4
The magnet's field-revision region

In all the above cases the electromagnetic forces between probe and closing wire were neglected because of its small magnitude compared to the predominant magnet-wire interaction forces.

The observed torques became, in all the experiments, independent of the location of the contact points between closing wire and collector rings. Also, the closing wire shape exhibited no noticeable influence on torques. These observations can be easily explained using the $div\mathbf{B} = 0$ fundamental law, Laplace force, and some elementary topological considerations.

Kennard [1], Bartlett [1]· Panosky [7,8], Muller [9], Wesley [10] and some of this article's authors took absolutistic viewpoints when dealing with homopolar phenomena [11,12]. On the contrary, Weber [4], Assis [13], and Kelly [14] adopted a relativistic framework on the issue from the beginning.

By attaching the magnet to the disk in the original Faraday setup, the relative rotation between disk and closing wire remains unchanged. Therefore, in a generator configuration, the disk plus magnet rotation at with the closing wire at rest in the lab is entirely equivalent to the closing-wire rotation at – with the disk plus magnet at rest. This fact introduced a correct but physically "colorless" weak relativism to the homopolar generator description: the "unipolar generator really has three components, the magnet, the cylinder and the meter (including the contacts). A relative motion of the last two, not the first two, is required" [1].

A growing interest in basic electromagnetism [15,27] can not be ignored, and from time to time some authors, attempting to catch "free energy" from the space, have claimed the design of homopolar engines with efficiency greater than unity, as can be checked by searching for *homopolar motor* on the Internet. The strict application of Newton's third law precludes the above non-physical possibility.

It is worthwhile to stress that the homopolar machine is a famous example where Faraday's flux rule fails. This fact worried Faraday himself and is clearly discussed by Feynman [28] who emphasized that the correct physics is always given by the Lorentz force law and the Maxwell fundamental equation *curl E = -B/t.* Homopolar induction is fully understood using only the Lorentz force. Our experiments enhance the relativistic structure of the Lorentz force because the only relevant velocity is the velocity of the conductor relative to the magnet.

Acknowledgments: To Profs. C.N. Gagliardo and A. Ipohorski-Lenkiewicz for the conceptual comments on this development.

References

1. J. Guala-Valverde and P. Mazzoni, "The principle of relativity as applied to motional electromagnetic induction". Am. J. Phys 63, 228-229 (1995).
2. J. Guala - Valverde, P. Mazzoni and Ro Blas, The quasilinear generator and = 0". Am. J. Phys. 65, 147-148 (1997).
3. J. Guala-Valverde and P. Mazzoni, "The unipolar dynamotor: a genuine relational engine". Apeiron 8, 41-52 (2001).
4. A.K.T. Assis, Weber Electrodynamics (Kluwer, Dordrecht, 1994).
5. A. Einstein, "Zur elektrodynamik bewegter körper". Annalen der Physik 17, 891-921 (1905). See also The Principle of Relativity. A.Einstein, H.A. Lorentz, H. Weil and H. Minkowsky, pp. 35-65 (Dover, New York, 1952).
6. G. Cavalleri et al. "Experimental proof of standard electrodynamics by measuring the self-force on a part of a current loop". Phys. Rev. E 58, 2505-2517 (1998).
7. W. K. H. Panofsky. Private communication (letter to J.Guala-Valverde, dated Jan 30, 1995).
8. W. K. H. Panofsky and M. Phillips, Classical Electricity and Magnetism. (Addisson Wesley, NY, 1955).
9. F. Müller, "Unipolar induction experiments and relativistic electrodynamics". Galilean Electrodynamics 1, N°3, 27-31 (1990).
10. J.P. Wesley. Private communication (letter to J.Guala Valverde dated Nov 01, 2001). See also "Progress in Space-Time Physics", Benjamin Wesley, Blumberg (1987).
11. J.Guala-Valverde and P. Mazzoni, "The motional electromagnetic induction revisited". Russian Academy of Sciences, 3rd Conference, pp.186-188. St. Petersburg (1995).
12. J. Guala-Valverde and P. Mazzoni, "The unipolar Faraday generator revisited". Galilean Electrodynamics, 4, 107-108 (1993).
13. A.K.T. Assis and D.S. Thober, "Unipolar induction and Weber's electrodynamics". Frontiers of Fundamental Physics, Plenum Press, NY, pp. 409-414 (1994).
14. A.G. Kelly, "Experiments on unipolar induction". Physics Essays 12, 372-382 (1999).
15. A.E. Robson and J.D. Sethian, "Railgun recoil, Ampere tension, and the laws of electrodynamics". Am. J. Phys. 60, 1111-1117 (1992).
16. A.K.T. Assis, Relational Mechanics (Apeiron, Montreal, 1999).
17. P. Graneau, "Electromagnetic jet-propulsion in the direction of current flow". Nature 295, 311-312 (1992).
18. T. E. Phipps and T.E. Phipps Jr., "Observation of Ampere forces in mercury". Phys. Lett. A 146, 6-14 (1990).
19. P.Graneau and N. Graneau, Newtonian Electrodynamics (World Scientific, Singapore, 1996).
20. V.F. Mikhailov, "The action of an electrostatic potential on the electron mass". Annales de la Fondation Louis de Broglie 24, 161-169 (1999); 26, 33-38 (2001).
21. A.K.T. Assis, "Comment on experimental proof of standard electrodynamics by measuring the self-force on a part of a current loop". Phys. Rev. E 62, 7544 (2000).
22. P. Graneau and N. Graneau, "Electrodynamic force law controversy". Phys. Rev.E 63, 058601-2 (2001).
23. T.E.Phipps and J. Guala-Valverde, "Simulation of amperian current elements by magnetic toroids". 21st Century Science & Technology 11, 55-65 (1998).
24. M. Bueno and A.K.T. Assis, Inductance and Force Calculations in Electrical Circuits (Nova Science Publishers, Huntington, N.Y. 2001).
25. H. Aspden, "Anomalous electrodynamic explosions in liquids". IEEE Trans. on Plasma. Sc. 14, 282-285 (1986).
26. G.M. Graham & D.G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo". Nature 285, 154-155.
27. V. Dvoeglazov, "Essay on the nonmaxwellian theories of electromagne-Tism". Hadronic Journal Suppl. 12, 241-245 (1997).
28. R. Feynman, The Feynman Lectures on Physics, II. (Addisson Wesley,NY, 1964).

Weight Reductions Generated by Bucking-Field Permanent Magnets

LAB REPORT ON SmCo RING MAGNET EXPERIMENTS

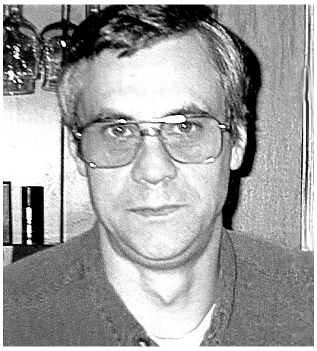

Experiments conducted by:
William C. Simpson

New Horizons Research 600 Meridian Street Extension,
#302 Groton, CT 06340
Tel. (860) 405-1157

The following experiments were conducted at the *Coastal Environmental Laboratory* (CEL) at Avery Point, Groton, Connecticut. (41° 19‘ 0.17" N. latitude x 72° 3‘ 50.27" W. longitude x 35 feet elevation above mean sea level) I wish to thank the personnelat the CEL for their generosity for providing the use of their *Mettler Toledo*® Model AG104 electronic scale for the measurements taken in the proceeding experiments. The AG104 electronic scale is an enclosed pan unit with a maximum mass range of 101-grams with 0.0001-gram readability.

The purpose of these experiments was to see if there is any detectable weight change when permanent magnets are forced together with their like-poles facing each other. The magnets were weighed individually, in both directions, with their field poles oriented vertically. The sums of the two individual magnet weights (magnet #1 and magnet #2) in each vertical orientation were compared to the weight measurements taken when they were assembled using the nylon bolt and wing nut depicted in **DIAGRAM 1**. The specifications for the two Samarium Cobalt magnets used in the following experiments are shown in **DIAGRAM 1**.

The first set of experiments with the SmCo Ring magnets were conducted January 14, 2002. An inverted paper cup was used to raise the test sample magnets 2.75" above the AG104 electronic scale pan in order to minimize possible magnetic interaction with the scale-sensing element, as depicted in **DIAGRAM 2**. The tare adjustment was used to set the scale readout to 0.0000-gram with the cup in place. The magnets were weighed individually. Magnet #1 weighed 9.9450-gram with the N pole facing up and 9.9397-gram with the S pole facing up. Magnet #2 weighed 9.9520-gram with the N pole facing up and 9.9443-gram with the S pole facing up.

The second set of experiments with the SmCo Ring magnets were conducted February 4, 2002. These experiments were shielded with Mu 80 magnetic

DIAGRAM 1

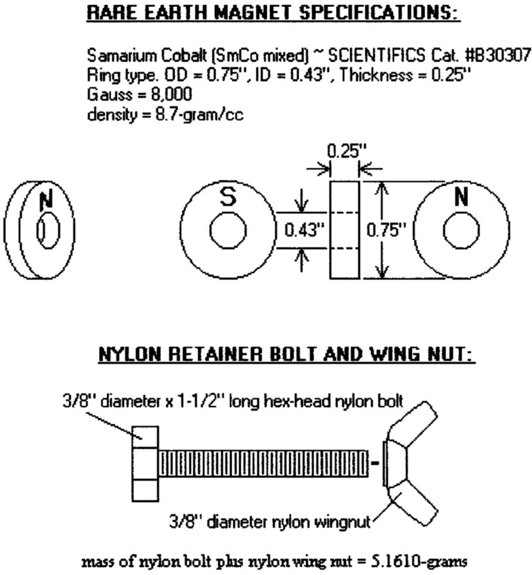
RARE EARTH MAGNET SPECIFICATIONS:
0.25"
N
S
0.43"
0.75"
N
NYLON RETAINER BOLT AND WING NUT:
3/8" diameter x 1-1/2" long hex-head nylon bolt
3/8" diameter nylon wingnut

Samarium Cobalt (SmCo mixed) ~ SCIENTIFICS Cat. #B30307-30
Ring type. OD = 0.75", ID = 0.43", Thickness = 0.25"
Gauss = 8,000
density = 8.7-gram/cc

mass of nylon bolt plus nylon wing nut = 5.1610-grams

Editorial: This article is one more link between magnetizm and theory of aether, to my mind. It is possible to assume that in his experiments the author creates small but detectable changes in density of aether, that demonstrates itself as the weight changes.
Alexander V. Frolov.

shielding material as depicted in **DIAGRAM 3**. The scale was tare adjusted to 0.0000-gram with the entire set of Mu 80 shield pieces in place. Then the magnets were weighed individually. Magnet #1 weighed 9.9483-gram with the N pole facing up and 9.9486-gram with the S pole facing up. Magnet #2 weighed 9.9527-gram with the N pole facing up and 9.9542-gram with the S pole facing up. The Nylon bolt and wing nut were placed in the Mu 80 shield can (without the magnets) and the scale was tare adjusted to 0.0000-gram. Therefore, the readouts would only be reading the weight of the bucking magnets.

The first column in **TABLE 1**, the vertical measurements, is the distance of separation d, or air gap, of the magnets. The second column shows the January 14, 2002 weight measurements of the two magnets, as shown in **DIAGRAM A**. The third column shows the January 14, 2002 weight measurements of the two magnets, as shown in **DIAGRAM B**. The fourth column shows the February 4, 2002 weight measurements of the two magnets, as shown in **DIAGRAM C**. The fifth column shows the February 4, 2002 weight measurements of the two magnets, as shown in **DIAGRAM D**.

The horizontal measurement, as depicted in **TABLE 2**, **TABLE 3**, **TABLE 4** and **TABLE 5**, are through the four basic magnetic compass headings; North, East, South, and West respectively. They are referenced to **DIAGRAM E** and **DIAGRAM F** for the January 14, 2002 experiments and **DIAGRAM G** and **DIAGRAM H** for the February 4, 2002 experiments. The corresponding graphs of the force change plots, **GRAPH 1**, **GRAPH 2**, **GRAPH 3**, **GRAPH 4**, and **GRAPH 5** accompany each table. The forces were converted from the mass readings, which are a scalar measurement, to dynes. The convention used for the force vector was chosen as plus (+) for up, or a weight reduction, and minus (-) for a weight increase.

DIAGRAM 2

01/14/2002 EXPERIMENNTS

CUP HEIGHT: 2.75" above scale pan. 1.75" upper diameter, 2.5" lower diameter

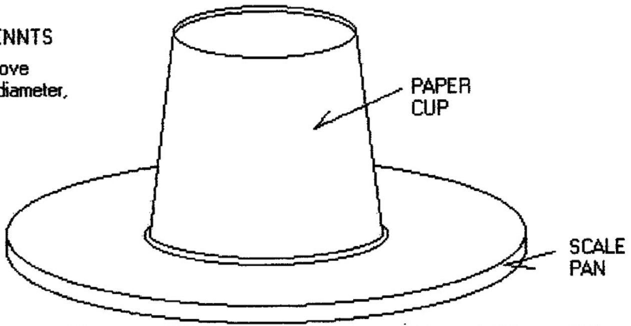
PAPER CUP
SCALE PAN

DIAGRAM 3

02/04/2002 EXPERIMENTS

MU METAL SPECIFICATIONS:

Source: NATIONAL ELECTRICAL ALLOYS
Oakland, NJ
SPECIFICATION: MIL N 14411C COMP 1
GRADE: HY MU 80 SHIELDING ALLOY
COLD ROLLED BRIGHT
DESCRIPTION: COIL
HEAT # 982011207
CHEMICAL ANALYSIS:
Ni: 80.16%
Mo: 4.6%
S: 0.0007%
C: 0.010%
P: 0.004%
Mn: 0.486%
Si: < 0.010%
Fe: BALANCE
MECHANICAL PROPERTIES:
HARDNESS: HV1 = 159
GRAIN SIZE: 8.5
COERCIVE STRENGTH:
HC = 0.0080 Oe
PERMEABILITY: MUMAX = 355000
SATURATION: 7900 Gauss
THICKNESS: 0.010"
DENSITY: 8.747-gram/cc

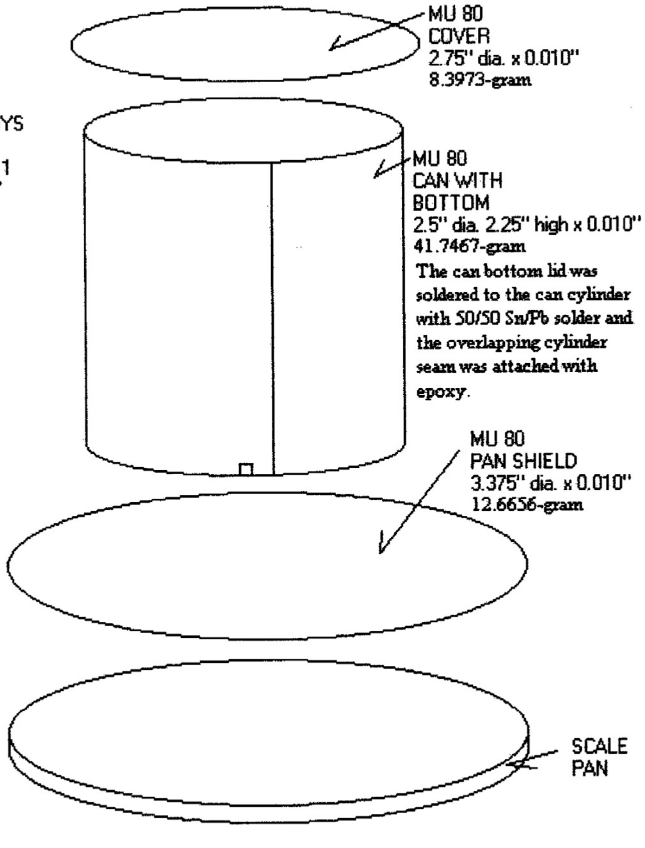
MU 80
COVER
2.75" dia. x 0.010"
8.3973-gram
MU 80
CAN WITH
BOTTOM
2.5" dia. 2.25" high x 0.010"
41.7467-gram
The can bottom lid was soldered to the can cylinder with 50/50 Sn/Pb solder and the overlapping cylinder seam was attached with epoxy.
MU 80
PAN SHIELD
3.375" dia. x 0.010"
12.6656-gram
SCALE
PAN

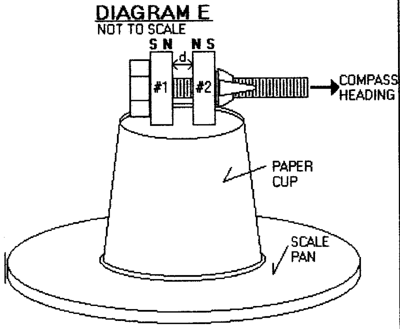
DIAGRAM E
NOT TO SCALE
S N
N S
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

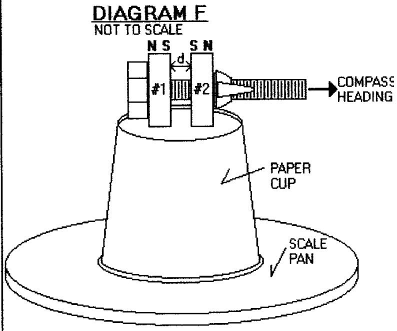
DIAGRAM F
NOT TO SCALE
N S
S N
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

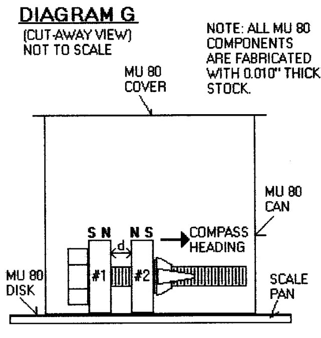
DIAGRAM G
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.
MU 80 COVER
MU 80 CAN
S N
N S
d
COMPASS HEADING
#1
#2
MU 80 DISK
SCALE PAN

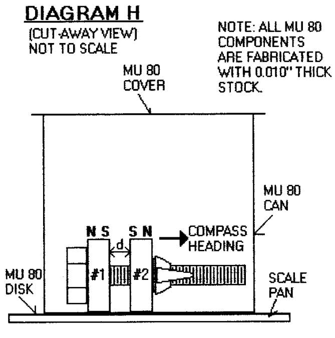
DIAGRAM H
(CUT-AWAY VIEW)
NOT TO SCALE
NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.
MU 80 COVER
MU 80 CAN
N S
S N
d
COMPASS HEADING
#1
#2
MU 80 DISK
SCALE PAN

ELEVATION VIEW OF THE VERTICAL EXPERIMENTS

NOTE:

January 14, 2002 Experiments: Start time: 13:27 hours EST, End time: 15:40 hours EST.

February 4, 2002 Experiments: Start time: 13:18 hours EST, End time: 14:50 hours EST.

TABLE 1:

	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap d	**Diagram A** N to N vertical 19.8893-gram (Sum of 1 & 2)*	**Diagram B** S to S vertical 19.8917-gram (Sum of 1 & 2)*	**Diagram C** N to N vertical 19.9025-gram (Sum of 1 & 2)*	**Diagram D** S to S vertical 19.9013-gram (Sum of 1 & 2)*
0.0 Inch	19.8759-gram 0.0134-gram weight reduction (0.06737%)	19.8757-gram 0.0160-gram weight reduction(0.08045%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8758-gram 0.0255-gram weight reduction(0.12821%)
1/8 Inch	19.8761-gram 0.0132-gram weight reduction(0.06637%)	19.8751-gram 0.0166-gram weight reduction(0.08346%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8763-gram 0.0250-gram weight reduction(0.12570%)
1/4 Inch	19.8763-gram 0.0130-gram weight reduction(0.06536%)	19.8748-gram 0.0169-gram weight reduction(0.08497%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8768-gram 0.0245-gram weight reduction(0.12318%)
3/8 Inch	19.8765-gram 0.0128-gram weight reduction(0.06436%)	19.8753-gram 0.0164-gram weight reduction(0.08246%)	19.8760-gram 0.0265-gram weight reduction(0.13324%)	19.8777-gram 0.0236-gram weight reduction(0.11866%)
1/2 Inch	19.8774-gram 0.0119-gram weight reduction(0.05983%)	19.8753-gram 0.0164-gram weight reduction(0.08246%)	19.8757-gram 0.0268-gram weight reduction(0.13475%)	19.8809-gram 0.0204-gram weight reduction(0.10257%)
5/8 Inch	19.8776-gram 0.0117-gram weight reduction(0.05883%)	19.8754-gram 0.0163-gram weight reduction(0.08195%)	19.8779-gram 0.0246-gram weight reduction(0.12368%)	19.8806-gram 0.0207-gram weight reduction(0.10408%)
11/16 Inch	19.8777-gram 0.0116-gram weight reduction(0.05832%)	19.8748-gram 0.0169-gram weight reduction(0.08497%)	19.8773-gram 0.0252-gram weight reduction(0.12670%)	19.8832-gram 0.0181-gram weight reduction(0.09100%)

* Magnets #1 and #2 were individually weighed in the orientation used in each experiment and their separate weights were added together.

The following equation was used to calculate the weight changes, in dynes, in the proceeding graphs.

$$\overrightarrow{\Delta F} = k \cdot \Delta mass \cdot g$$

where

$$k = 980.665 \cdot dyne \cdot gm^{-1}$$

and

$$g = 9.80665\ m \cdot sec^{-2}$$

which is the local rate of gravitational acceleration.

The product of the measured changes in mass, Δmass, and g is denoted as follows on the graphs:

$$NNopen_{n,1} = \Delta mass \cdot g$$

for N-to-N pole facings in the open (or unshielded).

$$SSopen_{n,1} = \Delta mass \cdot g$$

for S-to-S pole facings in the open (or unshielded).

$$NNshield_{p,1} = \Delta mass \cdot g$$

for N-to-N pole facings shielded with Mu 80 shielding.

$$SSshield_{p,1} = \Delta mass \cdot g$$

for S-to-S pole facings shielded with Mu 80 shielding. In the subscripts, $_n$ and $_p$ refers to the respective number of data points per plot. The subscript $_1$ refers to the vertical change in force (weight change) axis and $_0$ refers to the horizontal distance d axis. In **GRAPH 1**, $_n = _p$.

GRAPH 1:

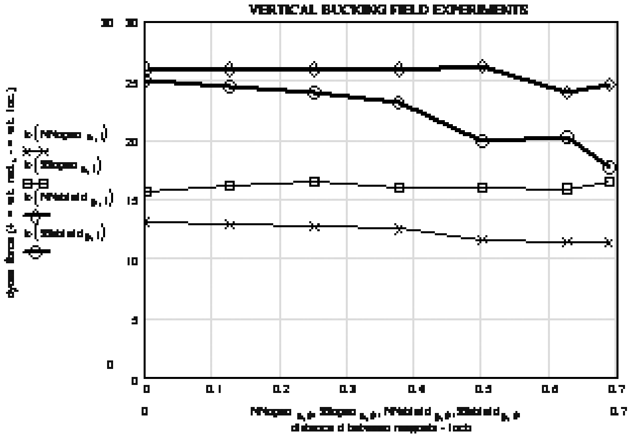

ISOMETRIC AND ELEVATION VIEW OF THE HORIZONTAL EXPERIMENTS

DIAGRAM E
NOT TO SCALE

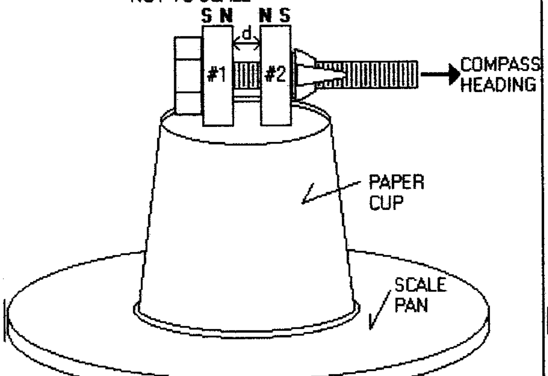
S N
N S
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

DIAGRAM F
NOT TO SCALE

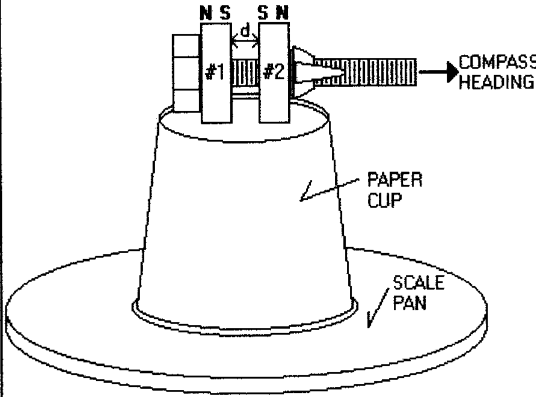
N S
S N
d
#1
#2
COMPASS HEADING
PAPER CUP
SCALE PAN

DIAGRAM G
(CUT-AWAY VIEW)
NOT TO SCALE

NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.

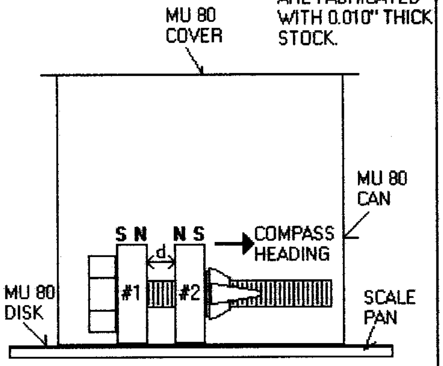
MU 80 COVER
MU 80 CAN
S N
N S
d
#1
#2
COMPASS HEADING
MU 80 DISK
SCALE PAN

DIAGRAM H
(CUT-AWAY VIEW)
NOT TO SCALE

NOTE: ALL MU 80 COMPONENTS ARE FABRICATED WITH 0.010" THICK STOCK.

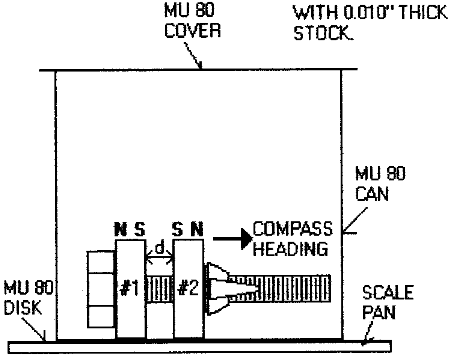
MU 80 COVER
MU 80 CAN
N S
S N
d
#1
#2
COMPASS HEADING
MU 80 DISK
SCALE PAN

MAGNETIC DECLINATION FROM THE LOCAL TOPOLOGICAL MAP:

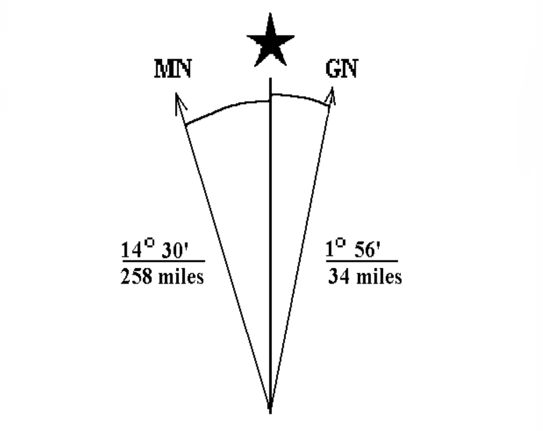

1984 Magnetic Declination~ U. S. Geological Survey
New London, Connecticut Quadrangle Topological Map

TABLE 2

North # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap d	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8735-gram 0.0170-gram weight reduction (0.08547%)	19.8778-gram 0.0127-gram weight reduction(0.06385%)	19.8890-gram 0.0129-gram weight reduction(0.06479%)	19.8810-gram 0.0209-gram weight reduction(0.10497%)
1/8 Inch	19.8728-gram 0.0177-gram weight reduction(0.08899%)	19.8790-gram 0.0115-gram weight reduction(0.05782%)	N/A **	N/A **
1/4 Inch	19.8721-gram 0.0184-gram weight reduction(0.09251%)	19.8798-gram 0.0107-gram weight reduction(0.05379%)	N/A **	N/A **
3/8 Inch	19.8705-gram 0.0200-gram weight reduction(0.10055%)	19.8800-gram 0.0105-gram weight reduction(0.05279%)	N/A **	N/A **
1/2 Inch	19.8698-gram 0.0207-gram weight reduction(0.10407%)	19.8811-gram 0.0094-gram weight reduction(0.04726%)	N/A **	N/A **
5/8 Inch	19.8693-gram 0.0212-gram weight reduction(0.10658%)	19.8818-gram 0.0087-gram weigh t reduction(0.04374%)	N/A **	N/A **
11/16 Inch	19.8696-gram 0.0209-gram weight reduction(0.10508%)	19.8827-gram 0.0078-gram weight reduction(0.03921%)	19.9190-gram 0.0171-gram weight increase(0.08588%)	19.9029-gram 0.0010-gram weight increase(0.00502%)

Compass Heading is approximate

* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.

** Due to time constraints, these measurements were not taken.

GRAPH 2:

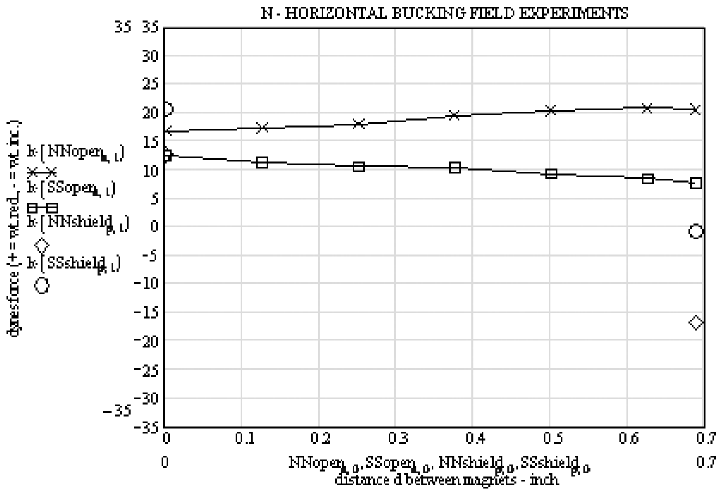

NOTE:
The calculated mass of each magnet was determined by the following formula based upon the manufacturer's dimensions and density value.

$$\text{volume} := \pi \cdot \left[\left(\frac{0.75 \cdot \text{in}}{2} \right)^2 - \left(\frac{0.43 \cdot \text{in}}{2} \right)^2 \right] \cdot \left(\frac{1}{4} \right) \cdot \text{in}$$

$$\text{volume} = 1.214963\ \text{cm}^3$$

$$\text{density} = 8.7 \cdot \text{gm} \cdot \text{cm}^{-3}$$

$$\text{mass} = \text{density} \cdot \text{volume}$$

$$\text{mass} = 10.570177\ \text{gm}.$$

This is higher than the magnetized mass of each magnet.

However, some tables give a lower density for the SmCo magnet, 0.300lb/in^3, which equals:

$$\text{density} = 8.303971 \cdot \text{gm} \cdot \text{cm}^{-3}$$

$$\text{mass} = 10.089017\ \text{gm}..$$

This is still higher than the magnetized mass of each magnet. Does the SmCo material become slightly lighter in weight when it is magnetized?

TABLE 3

East # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8763-gram 0.0142-gram weight reduction (0.07139%)	19.8755-gram 0.0150-gram weight reduction(0.07541%)	19.8928-gram 0.0091-gram weight reduction(0.04570%)	19.8731-gram 0.0288-gram weight reduction(0.14464%)
1/8 Inch	19.8777-gram 0.0128-gram weight reduction(0.06435%)	19.8750-gram 0.0155-gram weight reduction(0.07793%)	N/A **	N/A **
1/4 Inch	19.8782-gram 0.0123-gram weight reduction(0.06184%)	19.8743-gram 0.0162-gram weight reduction(0.08145%)	N/A **	N/A **
3/8 Inch	19.8779-gram 0.0126-gram weight reduction(0.06335%)	19.8722-gram 0.0183-gram weight reduction(0.09200%)	N/A **	N/A **
1/2 Inch	19.8792-gram 0.0113-gram weight reduction(0.05681%)	19.8715-gram 0.0190-gram weight reduction(0.09552%)	N/A **	N/A **
5/8 Inch	19.8814-gram 0.0091-gram weight reduction(0.04575%)	19.8712-gram 0.0193-gram weight reduction(0.09703%)	N/A **	N/A **
11/16 Inch	19.8815-gram 0.0090-gram weight reduction(0.04525%)	19.8720-gram 0.0185-gram weight reduction(0.09301%)	19.9338-gram 0.0319-gram weight increase(0.16021%)	19.9030-gram 0.0011-gram weight increase(0.00552%)

Compass Heading is approximate
* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.
** Due to time constraints, these measurements were not taken.

GRAPH 3:

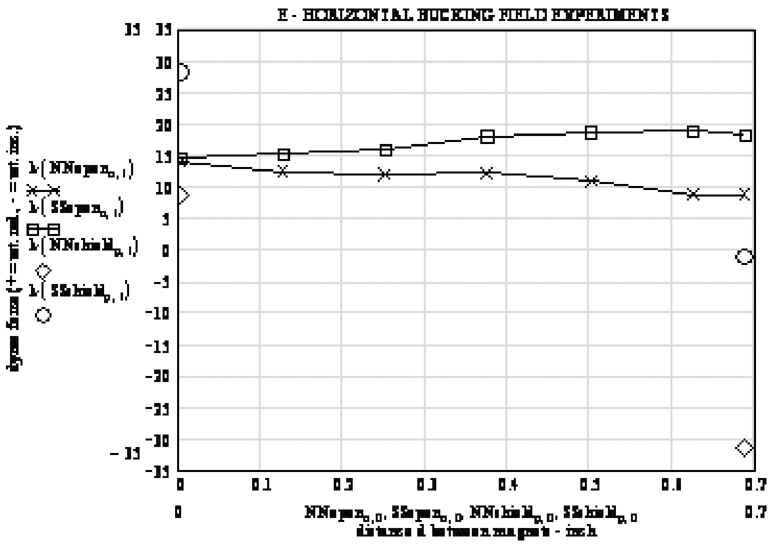

TABLE 4

East # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8746-gram 0.0159-gram weight reduction (0.07994%)	19.8769-gram 0.0136-gram weight reduction(0.06837%)	19.8896-gram 0.0123-gram weight reduction(0.06178%)	19.8660-gram 0.0359-gram weight reduction(0.18030%)
1/8 Inch	19.8720-gram 0.0185-gram weight reduction(0.09301%)	19.8788-gram 0.0117-gram weight reduction(0.05882%)	N/A **	N/A **
1/4 Inch	19.8709-gram 0.0196-gram weight reduction(0.09854%)	19.8797-gram 0.0108-gram weight reduction(0.05430%)	N/A **	N/A **
3/8 Inch	19.8704-gram 0.0201-gram weight reduction(0.10105%)	19.8803-gram 0.0102-gram weight reduction(0.05128%)	N/A **	N/A **
1/2 Inch	19.8711-gram 0.0194-gram weight reduction(0.09753%)	19.8814-gram 0.0091-gram weight reduction(0.04575%)	N/A **	N/A **
5/8 Inch	19.8698-gram 0.0207-gram weight reduction(0.10407%)	19.8825-gram 0.0080-gram weight reduction(0.04022%)	N/A **	N/A **
11/16 Inch	19.8703-gram 0.0202-gram weight reduction(0.10156%)	19.8817-gram 0.0088-gram weight reduction(0.04424%)	19.9423-gram 0.0404-gram weight increase(0.20290%)	19.8810-gram 0.0209-gram weight reduction(0.10497%)

Compass Heading is approximate
* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.
** Due to time constraints, these measurements were not taken.

GRAPH 4:

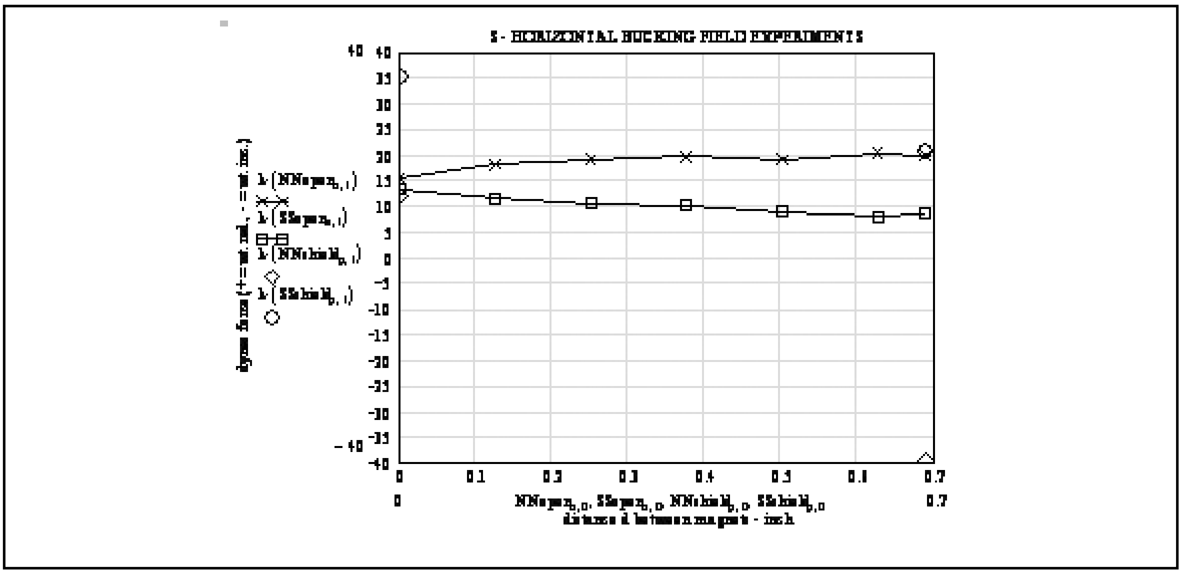

TABLE 5

West # **Heading**	01/14/2002 EXPERIMENTS OPEN, WITH PAPER CUP		02/04/2002 EXPERIMENTS WITH MU 80 SHIELDING	
Air gap D	**Diagram E** N to N horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram F** S to S horizontal 19.8905-gram (Sum of 1 & 2)*	**Diagram G** N to N horizontal 19.9019-gram (Sum of 1 & 2)*	**Diagram H** S to S horizontal 19.9019-gram (Sum of 1 & 2)*
0.0 Inch	19.8777-gram 0.0128-gram weight reduction (0.06787%)	19.8748-gram 0.0157-gram weight reduction(0.07893%)	19.8933-gram 0.0086-gram weight reduction(0.04319%)	19.8724-gram 0.0295-gram weight reduction(0.14816%)
1/8 Inch	19.8767-gram 0.0138-gram weight reduction(0.06938%)	19.8737-gram 0.0168-gram weight reduction(0.08446%)	N/A **	N/A **
1/4 Inch	19.8777-gram 0.0128-gram weight reduction(0.06435%)	19.8733-gram 0.0172-gram weight reduction(0.08647%)	N/A **	N/A **
3/8 Inch	19.8783-gram 0.0122-gram weight reduction(0.06134%)	19.8727-gram 0.0178-gram weight reduction(0.08949%)	N/A **	N/A **
1/2 Inch	19.8806-gram 0.0099-gram weight reduction(0.04977%)	19.8722-gram 0.0183-gram weight reduction(0.09200%)	N/A **	N/A **
5/8 Inch	19.8811-gram 0.0094-gram weight reduction(0.04726%)	19.8715-gram 0.0190-gram weight reduction(0.09552%)	N/A **	N/A **
11/16 Inch	19.8803-gram 0.0102-gram weight reduction(0.05128%)	19.8723-gram 0.0182-gram weight reduction(0.09150%)	19.9427-gram 0.0408-gram weight increase(0.20491%)	19.8976-gram 0.0043-gram weight reduction(0.02160%)

Compass Heading is approximate
* Magnets #1 and #2 were individually weighed with pole faces oriented vertically, with N up then with S up, and the results were averaged and added.
** Due to time constraints, these measurements were not taken.

GRAPH 5:

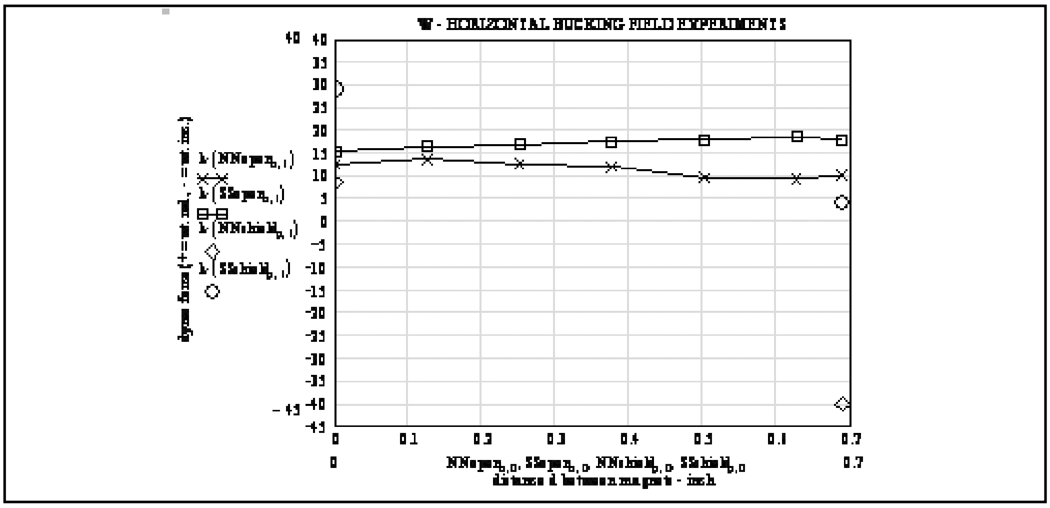

RING MAGNET #1, N UP: 9.9483-grams
RING MAGNET #2, N UP: +9.9527-grams
TOTAL WEIGHT: 19.9010-grams
WEIGHT INCREASE WITH OPPOSITE POLES ATTRACTING, AS DEPICTED IN **DIAGRAM I**, IN MU 80 SHIELD where 19.9861-grams was the measured value:
19.9861-grams – 19.9010-grams = 0.0851-gram weight increase.

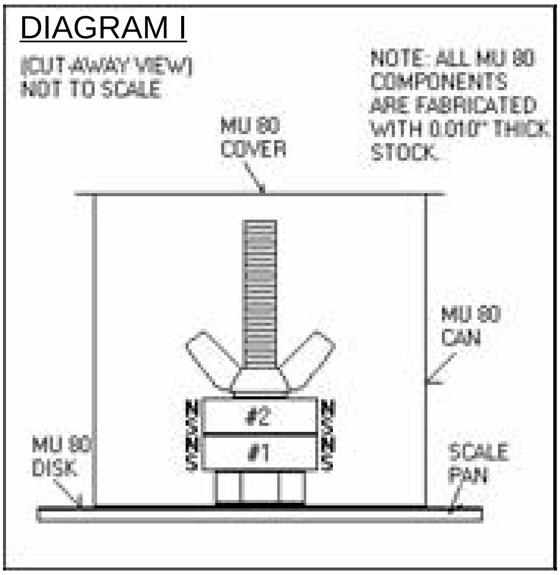

THREE RIVERS COMMUNITY COLLEGE, THAMES VALLEY CAMPUS LAB REPORT ON SmCo RING MAGNET EXPERIMENTS

The following experiments were conducted at the *Thames Valley Campus* (TVC) of the *Three Rivers Community College*, Room #207 Chemistry Laboratory, in Norwich, Connecticut on March 8, 2002. (41° 30' 34.62" N. latitude x 72° 6' 13.63" W. longitude x 115 feet elevation above mean sea level) I wish to thank the instructors at *Three Rivers* for their generosity for providing the use of their *Sartorius*® Model # 2442 analytical balance for the measurements taken in the proceeding experiments. The *Sartorius* Model # 2442 analytical balance is an enclosed pan unit with a maximum mass range of 200-grams with 0.0001-gram micrometer readability and a precision of 0.05-mg standard deviation.

The purpose of these experiments was to compare the Avery Point vertical measurements conducted on the *Mettler Toledo*® Model AG104 electronic scale, in **TABLE 1**, with the *Sartorius* Model # 2442 analytical balance measurements recorded in **TABLE 6**. The same two Samarium Cobalt magnets (magnet #1 and magnet #2) weighed individually in each vertical orientation were compared to the weight measurements taken when they were assembled using the nylon bolt and wing nut depicted in **DIAGRAM 1**. The specifications for the two Samarium Cobalt magnets used in the following experiments are shown in **DIAGRAM 1**.

The first column in **TABLE 6**, the vertical measurements, is the distance of separation d, or air gap, of the magnets. The second column shows the weight measurements of the two magnets, as shown in **DIAGRAM A**. The third column shows the weight measurements of the two magnets, as shown in **DIAGRAM B**. An inverted paper cup was used to raise the test sample magnets 2.75" above the *Sartorius* Model # 2442 balance scale pan in order to minimize possible magnetic interaction with the balance, as depicted in **DIAGRAM 2**. The fourth column shows the weight measurements of the two magnets, as shown in **DIAGRAM C**. The fifth column shows the weight measurements of the two magnets, as shown in **DIAGRAM D**. These experiments were shielded with

Mu 80 magnetic shielding material as depicted in **DIAGRAM 3**. The resultant data of **TABLE 6** is plotted on **GRAPH 6**. The results of the previous experiments at *Avery Point*, from **GRAPH 1**, and the recent *Thames Valley* experiments, from **GRAPH 6**, are plotted on **GRAPH 7** for comparison. The Mu 80 magnetically shielded experiments on the *Sartorius* Model # 2442 analytical balance at *Thames Valley* are in close agreement with the data collected with the AG104 electronic scale at Avery Point. However, the *Thames Valley* data collected for the unshielded experiments is somewhat smaller in weight reduction. I attribute this to external interference. The *Thames Valley* setup included a nonferrous tabletop, as did the Avery Point setup. However, the *Thames Valley* balance table consisted of a steel frame and legs, which may have altered the readings. The Mu 80 shielding provided a more intrinsic method for accurate data collection.

The horizontal measurements were not taken in this set of experiments due to time constraints. A final experiment was attempted to replicate the relative weight increase with the opposite poles of the ring magnets "stuck" together, as depicted in **DIAGRAM I**; however, the magnets shattered during assembly!

TABLE 6

	03/08/2002 TRCC EXPERIMENTS, @ TVC: OPEN, WITH PAPER CUP		03/08/2002 TRCC EXPERIMENTS, @ TVC: WITH MU 80 SHIELDING	
Air gap d	**Diagram A** N to N vertical 19.8734-gram (Sum of 1 & 2)*	**Diagram B** S to S vertical 19.8726-gram (Sum of 1 & 2)*	**Diagram C** N to N vertical 19.8912-gram (Sum of 1 & 2)*	**Diagram D** S to S vertical 19.8929-gram (Sum of 1 & 2)*
0.0 Inch	19.8699-gram 0.0035-gram weight reduction (0.01761%)	19.8696-gram 0.0030-gram weight reduction(0.01510%)	19.8691-gram 0.0221-gram weight reduction(0.11110%)	19.8706-gram 0.0223-gram weight reduction(0.11210%)
1/8 Inch	19.8699-gram 0.0035-gram weight reduction(0.01761%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8688-gram 0.0224-gram weight reduction(0.11261%)	19.8707-gram 0.0222-gram weight reduction(0.11160%)
1/4 Inch	19.8695-gram 0.0039-gram weight reduction(0.01962%)	19.8695-gram 0.0031-gram weight reduction(0.01560%)	19.8687-gram 0.0225-gram weight reduction(0.11312%)	19.8712-gram 0.0217-gram weight reduction(0.10908%)
3/8 Inch	19.8695-gram 0.0039-gram weight reduction(0.01962%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8692-gram 0.0220-gram weight reduction(0.11060%)	19.8730-gram 0.0199-gram weight reduction(0.10004%)
1/2 Inch	19.8689-gram 0.0045-gram weight reduction(0.02264%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8695-gram 0.0217-gram weight reduction(0.10909%)	19.8738-gram 0.0191-gram weight reduction(0.09601%)
5/8 Inch	19.8686-gram 0.0048-gram weight reduction(0.02415%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8702-gram 0.0210-gram weight reduction(0.10557%)	19.8746-gram 0.0183-gram weight reduction(0.09199%)
11/16 Inch	19.8680-gram 0.0054-gram weight reduction(0.02717%)	19.8699-gram 0.0027-gram weight reduction(0.01359%)	19.8707-gram 0.0205-gram weight reduction(0.10306%)	19.8756-gram 0.0173-gram weight reduction(0.08697%)

* Magnets #1 and #2 were individually weighed in the orientation used in each experiment and their separate weights were added together.

NOTE:
March 8, 2002 Experiments: Start time: 12:12 hours EST, End time: 13:47 hours EST.

GRAPH 6:

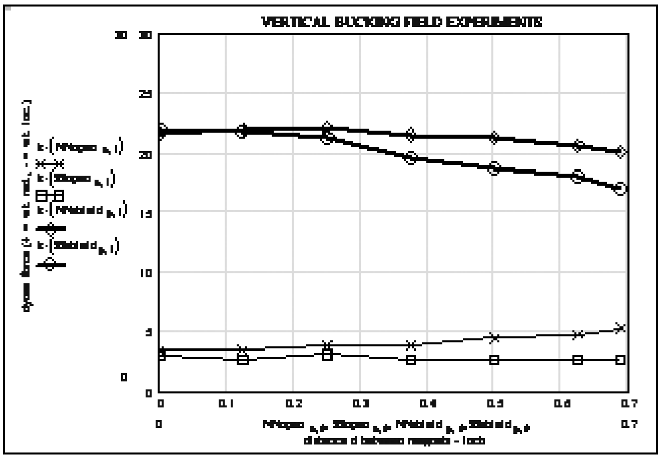

MAGNETIC DECLINATION FROM THE LOCAL TOPOLOGICAL MAP

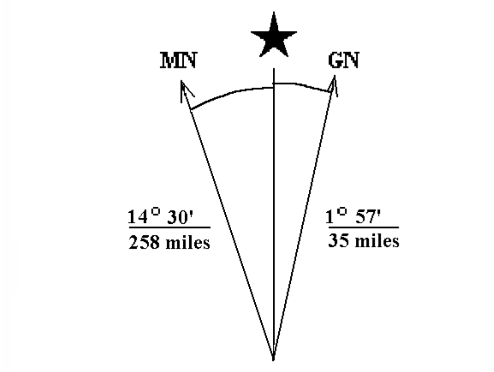

1983 Magnetic Declination~ U. S. Geological Survey
Norwich, Connecticut Quadrangle Topological Map

GRAPH 7:

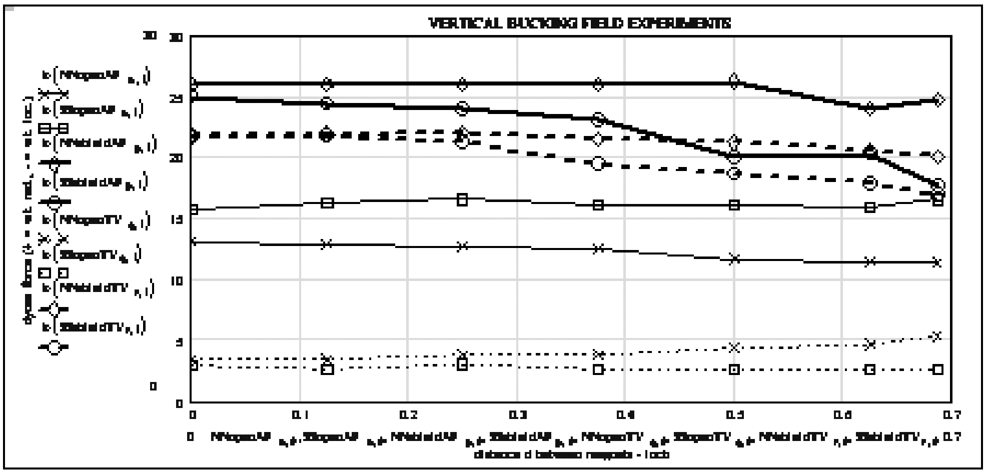

KEY TO **GRAPH 7**:

For N-to-N pole facings in the open (or unshielded) at Avery Point (AP):
NNopenAP
For S-to-S pole facings in the open (or unshielded) at Avery Point (AP):
SSopenAP
For N-to-N pole facings shielded with Mu 80 shielding at Avery Point (AP):
NNshieldAP
For S-to-S pole facings shielded with Mu 80 shielding at Avery Point (AP):
SSshieldAP
For N-to-N pole facings in the open (or unshielded) at Thames Valley (TV):
NNopenTV
For S-to-S pole facings in the open (or unshielded) at Thames Valley (TV):
SSopenTV
For N-to-N pole facings shielded with Mu 80 shielding at Thames Valley (TV):
NNshieldTV
For S-to-S pole facings shielded with Mu 80 shielding at Thames Valley (TV):
SSshieldTV

The subscripts for the data points $_{n} = _{p} = _{q} = _{r} = 7$ are all the same value in **GRAPH 7**.

Advanced Nuclear Waste Decontamination Technologies

Mark Porringa

Zeropoint Techtonix Inc, 430 Bass Lake Road, R R # 1,
Deep River, Ontario K0J 1P0, Canada
(613) 584-2960 fax: (613) 584-4616 porringam@aecl.ca

It is a review of nine alternative, peer-reviewed techniques as candidates for the global clean-up of nuclear waste.

The following is a reasonably comprehensive list of potentially effective nuclear waste treatment methods that might be employed to treat the entire range of radioactive wastes that have proven to be such a daunting and horrendously expensive problem for the nuclear industry (in all its forms) with major, long term implications for the environment.

A wide variety of methods will probably be required to accommodate the many different radioactive waste sources including high and low level, solids, liquids and gases. Process names used here are in some cases just convenient labels used to categorize and set them apart from each other.

Theories on several of these processes are still quite speculative and solid evidence that would pass conventional peer review is still lacking. This is after all a very new field of science.

Some of these technologies are already well protected by international or national patents, with additional US and international patents pending, and further patents may be obtained on new developments as they are made.

The Brown's Gas-Metal Matrix Process:

The BG-MMX process utilizes a patented electrolysis cell of the Australian Prof. **Yull Brown**'s design that is said to produce a stoichiometric mixture of monatomic hydrogen and oxygen or possibly a quasi-stable water molecule raised to a high-energy state. This gas has some very peculiar properties including the ability to sublimate tungsten (6000°C) with an implosive flame that burns cool in air with a temperature of only 130°C.

The gas is used to heat a proprietary mixture of metals and/or metal oxides including the radwaste to be neutralized. A highly exothermic radiant reaction appears to result in the immediate reduction of radioactivity approaching 95% of the original levels judging from preliminary tests, within seconds of treatment. The process is conjectured to be effective with high level solid wastes and possibly gasses, but probably not liquids. The high temperatures involved may also preclude the processing of more volatile wastes.

Since 1991, this technology has been successfully demonstrated, on a small scale, at least 50 times to US, Chinese, Japanese and United Kingdom officials on a variety of nuclear waste products including Americium, Cobalt, Uranium, and Plutonium. The technique can be applied for the immediate decontamination of stockpiles of nuclear waste materials being held near nuclear power plants. The process is very simple, safe, and inexpensive to develop further into robotics application for on-site treatment with no foreseen environmental effects.

Photoremediation

The *Photoremediation* process of the American Dr. **Paul Brown** is essentially conventional physics, *albeit* applied in a new and novel way. The process involves the use of a high-energy electron beam impinged on a target, which in turn produces a monochromatic gamma radiation that is tuned to induce *Photofission* and *Photoneutron* reactions in the target material causing rapid neutralization of radioactive isotopes. The efficiency claimed exceeds 500% due to the high cross-section reactions in the Giant Dipole Resonance region. The 10 MeV electron beam produces typical fission reactions in the 200MeV range effectively turning high level solid wastes such as spent fuel into an energy source. The process is apparently intended for on-site treatment with some waste-partitioning required, an aspect which may not be desirable in certain countries.

While this idea is similar in topology to a system being developed by **Los Alamos National Labs**, Dr. Paul Brown's approach offers several advantages: no need for extensive chemical pre-processing and the energy required to effect transmutation is greatly reduced. No new technology needs to be developed, yet the engineering of such a photon reactor must be completed and it could itself become a practical method for generating power.

ZIPP Fusion

The ZIPP fusion process, identified by **Mark Porringa**, induces a wide variety of fusion reactions, resulting from the radial compression of individual diatomic and other simple molecules dissolved or suspended in a light water, carbon arc electrolysis cell. A variety of other cell configurations are envisioned.

The process appears to produce only stable isotopes, which should therefore make it capable of stabilizing a wide variety of radioactive waste materials. The theory on the process draws from Condensed Charge phenomena, Brown's Gas implosion, cavitation bubble collapse and sonoluminesence - all variations of the Casimir effect - which is believed to cohere the Zero-point energy of Quantum Vacuum Fluctuations. Transmutations using variations of this basic process may be applicable to a wide variety of nuclear wastes and appears capable of operating with an efficiency exceeding 100%.

A major implication of this process is that the Strong force of the nucleus is understood as an ultra close range Casimir effect. **Oakridge** Nuclear Laboratories in the US in conjunction with several international collaborators have just (this month, in fact) announced a deuterium cold fusion process based on the essential elements of the ZIPP Fusion process first reported in 1998. The process is very simple and inexpensive to develop.

RIPPLE Fission

The RIPPLE Fission process is an adaptation of existing potential technology utilizing a supersonic ionized gas to aerosol a counter flow heat exchanger that envelopes the radioactive waste aerosol in a vacuum induced plasma vortex which appears to disrupt the matter stabilizing influence of the Quantum Vacuum fluctuations resulting in "gentle" low recoil fission reactions which produce only stable fission products, with excess neutrons being prompt converted to protons via quenched Beta emissions. The process is apparently proven with conventional non-radioactive wastes and is believed applicable to the entire spectrum of radwaste without the need for waste partitioning. This process is also conjectured to operate with over-unity efficiency.

The LENTEC Processes

The Low Energy Nuclear Transmutation Electrolytic Cells of the Cincinnati group produce a variety of transmutation reactions using a variety of exotic electrolysis cell designs that generally produce condensed charge clusters composed primarily of up to 10^{11} electrons each. These electron charge clusters produced with the use of special electrodes can penetrate the nuclei of larger atoms in solution and transmute these atoms into stable elements.

The range of design and operating protocols and potential applications are essentially limitless provided for the waste that is dispersed in the electrolyte. The reported transmutation of thorium to stable titanium and copper by the Cincinnati Group and by the Salt Lake City group is one of the most dramatic examples of this type of treatment process. Application to other high-level liquid transuranic fissionable wastes such as surplus Plutonium seems likely. The glaring absence of normal fission yield energies is perplexing but probably explicable as another form of low recoil fission reaction, similar to RIPPLE fission.

The Plasma Induced/Injected Transmutation - PIT Processes (also known as HDCC)

Plasma Induced/Injected Transmutation processes run include a gamut from recent achievements dating back to the **Oshawa-Kushi** cold plasma transmutations reported in 1964. The patented high-density charge cluster process (HDCC) was first discovered by **Kenneth Shoulders** and added on to by **Harold E. Puthoff**. Later, the late **Stan Gleeson** discovered HDCC in properly processed solutions. Still later, **Alexander Ilyanok** of Belarus discovered HDCC, followed by **Vasiliy Baraboskin** in Russia.

The production of Condensed Charge Clusters and various plasma glow discharge phenomena in a variety of gaseous atmospheres is again implicated as the underlying cause with what should be by now an obvious connection with the coherence of Zero-point energy from the Quantum or Stochastic vacuum.

Desk-top high energy particle accelerators have also been envisioned, based on the "piggy back" principle, in which the clusters permit acceleration of "piggy-backed" heaver +ions to extremely high energies capable of causing fusion and transmutations in target materials including those in solution and the materials of which the electrodes are composed. Brown's Gas implosion and cavitation bubble collapse reactions are also believed to be prevalent in these types of cells due to the prevalence of electrolysis.

A high-density charge cluster technology was discovered and used by Stan Gleeson to stabilize radioactive liquid wastes and has been developed further in the last 4 years by a group led by **S-X Jin** and **Hal Fox**. Best results for radioactive liquids have been demonstrated in the processing of thorium for a 30-minute period and achieving a reduction of radioactivity of about 90% from a liquid sample.

Kervran Reactions

The very compelling evidence compiled by French Nobel Candidate Dr. **Louis Kervran** has identified a wide range of nuclear transmutations in biological systems that have not been adequately explained. Coherence of Zero-point energy via Casimir effects within the *Somatid* particles identified by the Canadian **Gaston Naessens** is implicated as a possible cause. A wide variety of *in vitro* and *in vivo* reactions are believed to be possible as proven in nature and numerous experiments typically involving a reaction medium composed of a dielectric fluid such as water. Highly radiation resistant microorganisms have been found thriving in the core of nuclear reactors indicating the possibility of microorganisms being capable of transmuting some bioactive nuclear wastes in the course of the normal metabolism of such organisms.

The Monti Process

The Italian **Roberto A. Monti's** process involves confined explosions involving proprietary mixtures of materials that include radioactive waste. Ignition of such mixtures causes nuclear transmutations resulting in reduced radioactivity (to near-background levels) following combustion, gradually over 1 to 4 days.

This technique has been confirmed by the Italian **ENEA** and is supported by the French CEA scientists as a serious candidate for treatment of waste stockpiles. The system, as currently designed, required waste to be inserted into a chamber.

Higher group symmetry electrodynamics

Extremely weak, non-classical, higher group symmetry electromagnetic fields were found during a 1991 experiment made by **Glen Rein** to alter significantly the level of radioactivity in materials, even those in the environment. The experiments suggest that higher group symmetry electrodynamics modulate the quantitative and /or qualitative properties of radioactive species. If the non-classical fields directly affect the radioactive species, it is likely that the appropriate field parameters will be discovered to neutralize radioactive emissions. In 1999, a theoretical basis for the phenomenon was developed by the Welsh physicist, **M. W. Evans**, with the participation of Lt. Col. (retired) **Thomas E. Bearden**.

The technology is extremely simple and could be applied with minimum logistics for treating massive structures, *in-to* outdoors, such as the Chernobyl disaster site.

Psitronics Group Systems, International

Robert "Paul" LeBreton, 2901 Hwy. 6, HC 77 Box 42, Laguna, NM 87026 USA

E-mail: wizzard9@earthlink.net Phone:1-(505)-836-7534

Psi/Group's Magnetic Motor - Funding for New Prototype Sought:

As the engineer who designed the self-sustaining magnetic motor being advertised by Psitronics Group Systems, International. It perhaps falls upon me to explain (as well as I am able) the methods used by me in this motor's design:

The actual "picture" description is totally Intellectual Property; Protected as a Trade Secret.

However: This magnetic motor is a "perpetually imbalanced configuration" of permanent magnets; an unequal number of magnets on the rotor & stator; that revolves in a "self-sustaining" manner as the magnets seek balance... There are no electrical components in the design unless one wished to insert alternator windings in the stator to provide an electrical output...

We have discovered: through a prior unsuccessful prototype that using a para-magnetic material (we used aluminum) for the rotor & stator was self-defeating; as with the powerful rare earth magnets used in the design aluminum destroyed the magnetic fields... On further analysis of the failed prototype it was deduced that a "latch up" condition would occur between rotor & stator magnets; unless the rotor magnets were canted or skewed at an angle from the stator magnets... The plastic prototype contemplated corrects both of these problems....

We are seeking an investor willing to put up $4000 (DUS) for the prototype and who can offer $40,000 return on investment "if it proves successful" ...We have a Global Distribution 'ready to go" as well as a pledge of $ Millions towards Manufacturing & Marketing from EarthTech, International on submission of a running motor...

Very Respectfully Yours,
Paul LeBreton

http://home.earthlink.net/~wizzard9
(Psi/Group Website)

Inquiries should be emailed to:
wizzard9@earthlink.net

Experimental Investigations of the Radioactive Isotope Half-Value Period Changing in the Local Volume of Cause-Effect Relations

Igor A. Melnik

Pr.Frunze 232, Tomsk, 634021, Russia, Ph. (3822) 244555
breg@mail.ru

Nuclear-design methods of the elemental analysis are based on the property of the radioactive isotope decay rate constant. Half period as a constant quantity is defined by the time feature (the time flow is uniform in the every space point). The given postulate is confirmed by the long-term results and raised no doubts. But in last years some researches, investigating enough fine effects by change of the registered radiation intensivity, came to the conclusion about influence of space cycles on the time flow [1]. In his turn, N. A. Kozyrev during the investigation of active physical properties of time came to the conclusion about violation of its uniform flow in the local volume of the cause-effect relations, created by the cyclic motion (rotation, oscillation) of bodies [2,3,4].

Thus, the author get an idea to use a radioactive isotope cezium-137 as a "sensor", measuring the changes of time flow uniformity in the certain local volume of cause-effect relations. In basis there are following arguments: owing to a conception of time uniformity and considerable cezium-137 half-value period, the source activity must be permanent during the experiment time. By the time flow (period) change, i.e. changes of uniformity in the local volume of cause-effect relations, the half-value period – T is changing in the direct proportion. Source activity, respectively, is changing in the inverse proportion according to the law exp (1/T) in relation to the external space volume. Hence, registered gamma-quantum intension (amplitude impulse distribution) in the absence of the cause-effect relations is proportional to the function $N_0 \sim \exp(\ln 2/T_0)$. Then the proportion $N_0/\Delta N$, where $\Delta N = N_0 - N$ is made, and by means of it a half-value period difference is defined by the formula

$$1/\Delta T = \left| 1/T_0 + \ln(\Delta N/N_0)/\ln 2 \right| \qquad (1)$$

N_0 – selective average amplitudes of impulses at the case of static liquid (in the absence of the cause-effect relations);
N - selective average amplitudes of impulses at the case of rotation of liquid.

In this case, there were investigations of the determination of the gamma-quantum intension change (i.e. change of a half-value period) dependence on the angular velocity of the activator rotation, and also on the coordinates location and amount of cause-effect relations (Fig. 1).

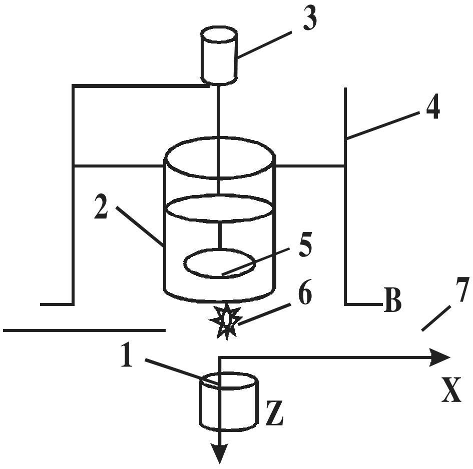

Fig.1

The following devices were used as measurement equipment: semi-conducting detector (SCD-63V) (1), preamplifier (PAG-2K), amplifier (BAI-3K) and analyzer (AMA-02F1). Energy gamma-line of 662 KeV was measured with the resolution 4 KeV. The gamma-quantum source was glued to the detector housing at the distance of 100 mm from its surface; so, any spatial change through coordinate axes was executed together with the detector to avoid even a tiny change of the source-detector geometry.

The vessel with a liquid (2) was placed above the source (6), vortex fluid motion was created by an activator (5), placed on the rod of electric motor (3). The glass with a liquid, connected with the motor and frame (4), was displaced regarding to the center of the revolved activator, what was the reason of the glass vibration in horizontal line, and, hence, one of frame sides vibration in vertical line (v). In its base, the frame was a square with the side of 200 mm. To assign the direction of the space cause-effect relations, one of frame sides was vibrating, and the opposite one (on the X axis) was fixed on the table (7). An experiment was made in such a way, that vibrating parts did not adjoin with the detector. The radius of glass is 50 mm, the distance from central axis up to the vibrating frame part is 100 mm. N-selection of the every value on the diagram (Fig. 2) corresponded to twenty measurements, roots from the average variance of numbers distribution $D \approx 160$ impulses.

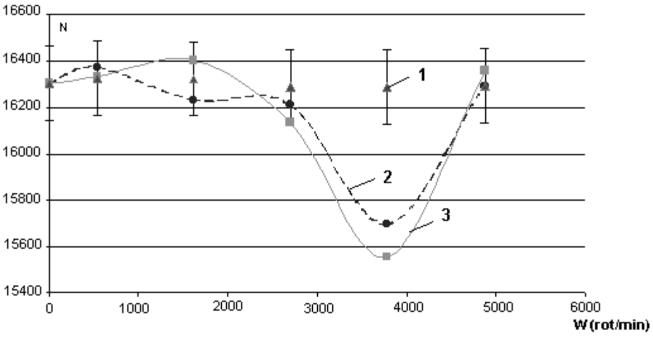

Fig. 2
Dependence of photopeak-N area from the angular velocity of rotation – W, where 1 – is photopeak area at static liquid, 2 – is photopeak area at clockwise rotation, 3 – is photopeak area at anticlockwise rotation

During two months eighteen experiments were made (at the one geometry) of the given effect fluctuation study, and there were no considerable deviation. An effect of the time deceleration always appeared at the angular speed of the activator rotation W=3780 r/min (during the vibration of the whole perimeter of the frame base). If there was a vibration on the only one side of the frame base (on the X axis), then this effect appeared at the higher speed of rotation. Unfortunately, the maximum motor speed of rotation W_{max}=4880 r/min did not let to define the precise position for the next rotation point. During the experimental results extrapolation W_2≈5930 r/min ($W_2 \approx \pi W/2$) was obtained.

In the table there are selective middle amplitudes of impulses -$N=\sum N_n/n$ roots from the numbers distribution average variance D, dispersion of the simple average G at the different directions of the liquid rotation (for the rotation point W=3780 r/min). The half period is estimated by the equation (1) and formula $T=T_0+\Delta T$ where T_0=30.2 year (the cezium-137 half-value period).

Liquid motion	N, imp.	D	G	T, year
Static liquid	16304.0	160.7	35.9	30.200
Clockwise	15696.0	162.0	36.2	30.412
Anticlockwise	15553.0	165.0	36.9	30.423

During the investigation of the activator angular speed of rotation dependence on the registered gamma-quantum intension, **there was discovered the time deceleration effect in the fixed point of ambient space of the glass with the liquid.** The effect was versatile, during the further study of the reasons of the effect disappearance, the following regularities were found:

1. Intension change appeared only at the case of asymmetrically fixed glass, when there appeared horizontal oscillations, transmitted to the vertical oscillations of the frame;
2. An effect disappeared, if the gamma-quantum source was placed in the glass center (~ 2 ÷ 3 mm down from the glass bottom);
3. At the same speed of rotation but without liquid, with an eccentric activator, the time deceleration effect disappeared.

Will consider the system with liquid, its internal chain consisted of three cause-effect relations:

1. Activator-liquid;
2. Liquid-glass (frame);
3. Frame-table.

Glass oscillations were transmitted to the frame through the hard cohesion. There was only one cause-effect relation – an activator (a frame) – a table. Thus, amplitude of the time flow changing was, most likely, influenced on by the amount of cause-effect relation, and also, there was observed a quantum effect, dependent on the activator and liquid rotation frequency.

The time deceleration maximum amplitude appeared during the source displacement from the central axis of glass along the X coordinate to the distance ~ 5 mm. There appeared the necessity to check the given effect on space points along the all coordinate axes in the radius R ≈ 100 mm from the central point of the glass bottom (an extreme vibration point along the X axis was placed on this distance).

On this purpose there were made some experiments, results of which are shown on the (Fig. 3), (Fig. 4).

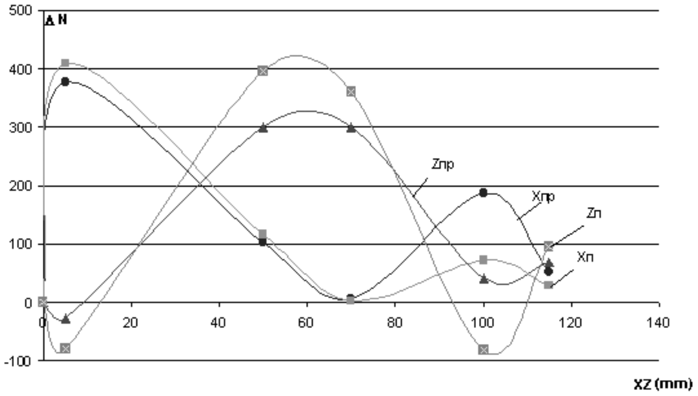

Fig. 3
Dependence of the increase of photopeak -ΔN area from the spatial interval in Cause-Effect Relations by X and Z coordinates, where X_cZ_c – is clockwise rotation, $X_{ac}Z_{ac}$ – is anticlockwise rotation

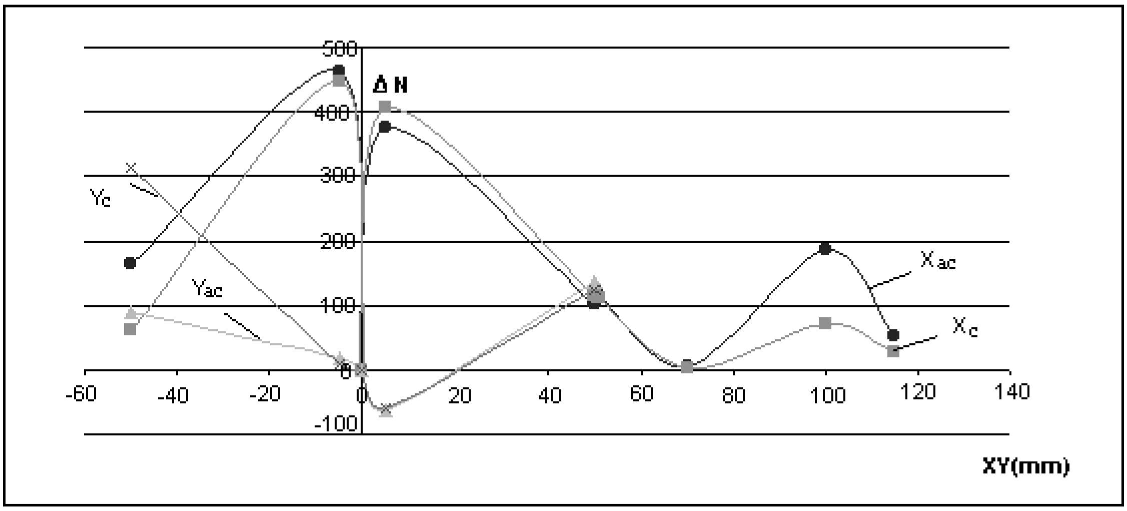

Fig. 4
Dependence of the increase of photopeak -ΔN area from the spatial interval in Cause-Effect Relations by X and Y coordinates, where X_cY_c – is clockwise rotation, $X_{ac}Y_{ac}$ – is anticlockwise rotation

Let us suppose that positive axis Z is directed from the glass bottom center to the ground and axis X – from the central axis of glass to the vibrated part of the frame (North-West), in this case Y is directed to the North-East. Along the all coordinates, in every adjusted space points on the distance R_i from the glass bottom center (R_i = 5; 50; 70; 100; 115 mm), there were forty measurements. At that in every point after ten measurements connected with the rotation there were measurements at the static liquid. It was made to avoid any systematic errors. The amplitude of the sample average impulses at the static liquid - $N_0=10550$ imp. In this case the root from the dispersion $\Delta=120$, dispersion of the simple average G=19. Amplitude increment was calculated by the formula $\Delta N=|N_0-N|$. Following regularities were found:

1. If considering the difference of impulse amplitude (in the positive coordinate space) in the first cause-effect relation evidence (activator-liquid), i.e., to put the difference between the initial (5 mm) and final (50 mm) points of amplitude registration (their middle evidence between "clockwise" and "anticlockwise") $\Delta N_5^{50}=|N_5-N_{50}|$ into the formula (1), then relations $\Delta T_Z/\Delta T_X=\Delta T_X/\Delta T_Y\approx1.11$ ($\pi/2\sqrt{2}\approx1.11$)

2. Change of impulse amplitude difference along X is the antiphase of changes along Z and Y;

3. At the rotation clockwise and anticlockwise there takes place a periodical inversion of amplitudes difference relation;

4. At the liquid rotation clockwise and anticlockwise, there appears an obvious distinction in evidences of the increment ΔN in negative region of coordinate axes;

5. Along Y and Z-axes in points (5; 100 mm) there is observed insignificant time acceleration.

The whole cause-effect relation system was defined as an internal (activator-table), as an external one (ground - system center of gravity). **In the internal space volume time at the certain conditions breaks its own uniformity, at that the time period change is nonlinear and is defined by its quantum nature.** Therefore, standard clocks in causal relations must be considered regarding to the center of gravity and location of the measurement point in the internal system space.

References

1. Shnol S.E., Zenchenko T.A., Zenchenko K.I., PozharskyE.V., Kolombet V.A., Konradov A.A. Regular change of the statistical distributions thin structure as a result of cosmophysical causes. Kushnirenko E.A., Pogozhev I.B. Comment to the article of Shnol S.E. and Co. Progress of physical sciences. Vol. 170, #2, p.213-218, 2000.

2. Kozyrev N.A. Causal or asymmetric mechanics in the linear approximation. Pulkovo, 1958, p. 90.

3. Kozyrev N. A. Astronomic observations by means of time physical properties. Symposium transactions, Erevan, 1977, p. 209-227.

4. Kozyrev N.A. About possibility of bodies mass and weight under the influence of time active properties. Eganova I.A. Analytical survey of modern time-keeping ideas and experiments. Novosibirsk, 1984, p. 92-98.

5. Melnik I.A. Experimental researches of the time flow local deceleration. VINITI 5.02.92, N1032-mg92, #4, p.70.

TILLEY ELECTRIC VEHICLE

Tilley Foundation, Inc.
131 Hiwassee Road
Lebanon, TN 37087, USA
615-443-5315
E-mail: cktilley@bellsouth.net

Nowadays there stays urgent the development of technologies, which could possibly be combined into a reliable, commercially successful advanced self powered vehicle with additional desirable but currently commercially unavailable features.

The advantages of electric vehicles seem to be evident – they are clean, quiet, powerful, require much less maintenance than gasoline or diesel-fueled vehicles, and are inherently much simpler and easier to manufacture. Their drawbacks have been a short range, long battery recharging time, and a heavy, bulky battery pack.

After several years of personal accomplishments in the alternative energy industry, **Carl B. Tilley, the inventor from USA,** was convinced that it was possible to build an electric car that could be powered without the help of external power to keep the battery charged.

The concept to produce a useful electric performance car that would last more than a few hours and would be economical to run, safe to drive around town or across the United States and never use a drop of fuel challenges the future of transportation as we know it today.

With the establishment of the **Tilley Foundation, Inc.**, in the year 2001, Carl Tilley set out to prove it could be done. It was an ambitious project and it broke ground on the facility in Tennessee that would build the first self-generating electric car.

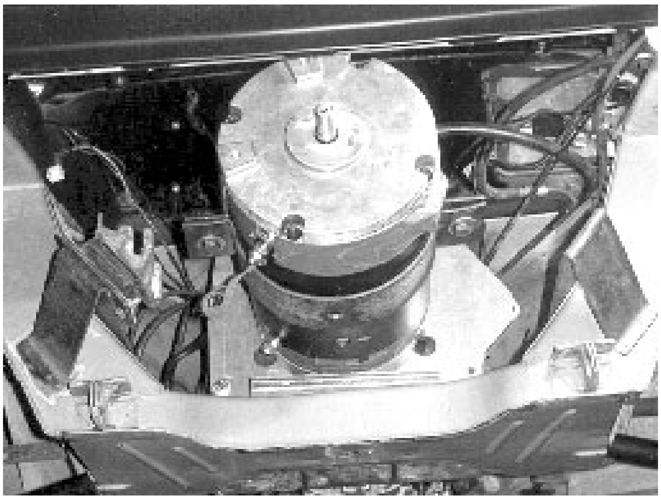

Construction of a 1,800 square foot building, that was powered with another recently developed electric device, began in the year 2002. Electricity for this car is provided from a different energy invention, which was void of any outside power supply. It is ironic that one alternative energy device actually built the invention to power and build the electric car.

The **Tilley Electric Vehicle** (**TEV**), converted from a 1981 DeLorean, energizes the imagination and defies what has been accepted as a standard in the area of transportation for years.

From the selection of the proper car to be converted, to the advanced technology, which is on board, the **TEV** performs comparably to gasoline-powered vehicles.The difference is you have no need for fuel and you do not have to stop the vehicle to charge it after driving. There is no pollution and you can cruise the highways at the same speed as any other vehicle.

There is a new car on the road today. A car built with technology that defies the concept of fossil fuel powered cars, and can run coast to coast without ever relying on the battery being charged from an outside source.

The long awaited transportation revolution and the end of our reliance on fossil fuel has now begun **….the Tilley Electric Vehicle.**

The demonstration of a DeLorean powered by an electric motor and 12-volt standard car batteries is supposed to be on September 7, 2002, at the Nashville Superspeedway, USA. The battery bank is kept in a charged condition by the "on board" charger which is the device invented by Carl B. Tilley. Racing legend Bobby Allison is one of the guest drivers for this demonstration.

Details and results of the demonstration read in our next issue

The Problem of Time: Force as the Cause of Change of the Course of Time

Valentin P. Oleinik

Department of General and Theoretical Physics,
National Technical University of Ukraine "Kiev Polytechnic
Institute", Prospect Pobedy 37, Kiev, 03056, Ukraine;
e-mail: yuri@arepjev.relc.com
Internet: http://temporology.bio.msu.ru/OLEINIK/oleinik.htm.

Abstract

Material processes occurring in a physical system under the action of a force field necessarily influence the course of time along the trajectory of motion of particle. A general relationship is obtained which relates the course of time on one path section of a particle when moving in a force field to that on the other path section in the same inertial reference frame. According to the results obtained, the force in relativistic mechanics is not only the cause of acceleration of particle relative to an inertial frame of reference, but also the cause of change in the course of time along the particle's trajectory. ***Therein lies the physical content of the dynamical principle underlying the special theory of relativity (relativistic mechanics).*** *The applications of the theory developed to homogeneous fields - to the field of gravity and electromagnetic field, and to the gravitational field produced by a point mass particle are considered. Physical properties of the state of imponderability of particle in an external force field are investigated. It is noted that the change in the course of time in a force field is in no way connected with the change in space-time metric and is a direct consequence of the causality principle of relativistic mechanics.* ***The existence of dependence of the course of time on the state of motion of particle in a force field points to the feasibility of controlling the course of time using force fields.***

Time is among the most common concepts, which are used constantly both in everyday life and in science. This is because all the events and material processes in the world happen in space and develop in time and, hence, the laws that govern space-time connections are the most general and hold for all the forms of matter. Nevertheless, time remains one of the most mysterious concepts of physics; its physical essence is not adequately revealed up till now [1-4]. The concept of time with difficulty yields to logical analysis.

From the point of view of common sense the essence of time is that time characterizes the duration of events and processes, indicates their natural sequence, at which the present, going away to the past, gives place to the future.

Isaac Newton gave a clear-cut characteristic of the concept of time, to which the majority of physicists adheres: "The absolute, true, and mathematical time in itself and by virtue of its nature flows uniformly and regardless to any other object". Though, according to Newton, time flows equally and uniformly and does not depend on the processes, occurring in the world, the daily experience speaks in favour of the fact that the course of time is not uniform. Depending on circumstances in our history, it seems to us that time either flies swiftly or hangs heavy on our hands; sometimes it even changes suddenly, by leaps. In connection with these speculations the question arises of whether the subjective sensations of non-uniformity in the course of time familiar to everyone have an objective basis.

In Newtonian mechanics time is of an absolute character, it does not change as one passes from one inertial reference frame to another and represents merely a parameter, whose change at the will of explorer results in the change of state of a mechanical system in accordance with the equation of motion.

In relativistic mechanics time remains a parameter describing the development of system. But now time and space are intimately linked with each other to form a single system, i.e. the 4-dimensional space-time. In going from one inertial frame of reference to another time gets entangled with spatial coordinates, so that time in one reference frame represents a "mixture" of time and coordinates in the other. Time ceases to be universal, the same in all inertial reference frames; it takes on a relative character.

The indissoluble association of time and space takes on special importance in the light of the concept of physical field, which was called by Einstein the most important discovery in physics after Newton. According to this concept, the occurrence in space of a force field means that space turns into a physical environment, which is capable to interact directly with other bodies and gains, thus, physical properties, becoming an active participant of physical processes. In view of the fact that **space and time are indissolubly related to each other, the presence of a force field in some area of space must necessarily result in the appearance of physical properties of time caused by the motion of body in this area.**

Thus, from the synthesis of the notion of space-time and of the idea of physical field it follows with necessity that **the course of time in a given region of space should depend on physical processes in this region, i.e. time, as well as space, should have physical properties [5-8].**

It should be emphasized that in special theory of relativity (STR) time and spatial coordinates are independent and formally equal in rights quantities, which determine the position of elementary events in space-time. On the other hand, time stands out in relation to spatial coordinates. The special role of time is due, from the viewpoint of geometry, to the pseudoeuclidity of geometry of the 4-dimensional space. From the physical point of view, it is associated with the dynamical principle (causality principle), according to which the state of motion of a physical system at an instant of time t uniquely defines its behaviour at the next instant of time $t+0$. The significance of dynamical principle lies in the fact that it relates the temporal evolution of system to the physical processes caused by force fields and in doing so it allows one to determine the course of time in the system, its possible dependence upon the character of physical processes, and not just the sequence of events and their duration.

The idea about the existence of the physical properties of time belongs to N. Kozyrev [9]. By introducing into mechanics an additional parameter taking into account the directivity of the course of time, Kozyrev has formulated causal (asymmetrical) mechanics from which it follows that time has physical properties. According to the results of theoretical and experimental investigations conducted by Kozyrev and his followers [9-13], events can proceed not only in time, but also with the help of time, **information being transmitted not through force fields, but via a temporal channel, and the transfer of information happens instantaneously. The appearance of additional forces, associated with the physical properties of time and capable to fulfill work, testifies that time can serve as a power source.**

In the papers by I. Eganova [12] and M. Lavrent'ev and I. Eganova [13] it is stated the problem of direct experimental research of the physical properties of time to ascertain the relations of a new type between phenomena and to discover new methods to change the state of substance. In [14] O. Jefimenko investigated the dynamical effect of the slowing-down of time.

According to [6-8], the conclusion that physical properties of time exist follows strictly from relativistic mechanics, without introducing any additional hypotheses. The physical properties of time are of purely dynamical nature: their existence results from dynamical principle. **The availability of physical properties of time is manifested in that time has a local inhomogeneity: its course along the trajectory of motion of a point particle in a force field is continuously changed, and this change in the course of time is a result of the action upon the particle of a force field in the inertial reference frame, in which the motion is considered.**

Editor's note: The author gives a detailed consideration of the physical content of the local dynamical inhomogeneity of time. Considering the motion of a classical point particle under the action of a force field in the inertial reference frames, that moves relatively to each other, Valentin P. Oleinik derives 73 equations, that help him to obtain the following conclusion (mathematical details and physical reasoning may be found in http://temporology.bio.msu.ru/OLEINIK/oleinik.htm.*)*

The elucidation of the physical nature of time is one of the most important problems of theoretical physics. The purpose of research on the problem of time is to study the physical properties of time, i.e. to ascertain the possible interrelation between time and material processes. In particular, it is of interest to find out

- whether the flow of time depends upon physical processes and whether the back influence exists (i.e influence of the change of the time course on physical processes);
- what mechanisms of the change of the course of time are available;
- what factors are capable to speed up or to slow down the flow of time.

In papers [5-8] on the basis of Lorentz transformations relating to coordinates of points, lying on the trajectory of motion of particle in a force field, the phenomenon of local dynamical inhomogeneity of time is predicted. The main result consists in the proof that material processes occurring in a physical system under the action of a force field necessarily influence the course of time along the trajectory of motion of particle. The case in point is the change of the course of time along particle's trajectory in one inertial reference frame as compared with that in the other.

In this paper the next step is made: the relationship is obtained which relates the course of time on one path section of a particle when moving in a force field to that on the other path section in the same inertial reference frame. The main idea underlying the approach developed results from the analysis of Lorentz transformations and consists in that **the course of time of a particle moving by inertia, i.e. of a particle being not exposed to force, should be uniform.**

As is well known [17,18], the existence of dependence of the course of time upon the gravitational field potential is predicted with the general theory of relativity (GTR). According to GTR ([17], p.303), time flows differently at the different points of space in one and the same reference frame. Since "gravitational field is nothing more nor less than a change of the space-time metric" ([17], p.313), one can assert, apparently, that the change in the course of time is due, in the view of GTR, to the change of the 4-space metric. It should be emphasized that in the present paper gravitational field is considered as an ordinary force field, and the particle motion is supposed to occur in pseudo-Euclidian space-time. The main formulas of the article, (22) and (25), describe the change in the course of time in an arbitrary force field at different spatial points in one and the same inertial reference frame. As is seen from the results received, the change in the course of time in a force field is in no way connected with the change of space-time metric. It is conditioned by the force field action on particle in inertial reference frame and is a direct consequence of the dynamical principle underlying relativistic mechanics.

It should be emphasized that the existence of dependence of the course of time on the state of motion of particle in a force field points to the feasibility of controlling the course of time using force fields.

Note an important peculiarity of the non-inertial reference frame, in which the imponderability state of a particle is attained: **there is such a space-time region in which the reference frame at hand can be approximately considered as inertial.** In connection with the fact that such reference frames (it is natural to call them **quasiinertial** in contradistinction to the **true inertial** reference frames) are, generally speaking, not equivalent to each other (see previous section), the derivation of a rigorous criterion for inertial reference frame acquires especial significance. The dynamical criteria for defining the inertial and non-inertial states are considered in the papers by B. I. Peschevitsky [19]. The heliocentric reference frame seems to be among the quasiinertial reference frames, being inertial with adequate accuracy only in a restricted region of space (for example, within the limits of our Galaxy) [16].

The author is grateful to Yu.D. Arepjev for his interest to the paper and stimulating discussions.

References

1. Blokhintsev D.I. Space and Time in the Microworld (Nauka, M., 1970) (in Russian).
2. Prigogine I. From Being to Becoming: Time and Complexity in the Physical Sciences (W. H. Freeman & Co., San Francisco, 1980).
3. Prigogine I. and Stengers I. Order Out of Chaos (Bantam Books, New York, 1983).
4. Logunov A.A. Lectures on the Theory of Relativity and Gravitation: Modern Analysis of the Problem (Nauka, M., 1987) (in Russian).
5. Oleinik V.P. The Newest Development of Quantum Electrodynamics: Self-Organizing Electron, Superluminal Signals, Dynamical Inhomogeneity of Time, Physical Vacuum and Nature, 4, 3-17 (2000).
6. Oleinik V.P., Borimsky Ju.C., Arepjev Ju.D. New Ideas in Electrodynamics: Physical Properties of Time. Semiconductor Physics, Quantum Electronics & Optoelectronics, 3, #4, 558-565 (2000). E-print: quant-ph/0010027.
7. Oleinik V.P. Superluminal Signals, Physical Properties of Time, and Principle of Self-Organization, Physics of Consciousness and Life, Cosmology and Astrophysics, 1, 68-76, (2001); Borimsky Yu.C., Oleinik V.P. The Course of Time in Classical and Quantum Systems and Dynamical Principle, Physical Vacuum and Nature, 6, (2001) (in print).
8. Oleinik V.P. The Problem of Electron and Superluminal Signals. (Contemporary Fundamental Physics) (Nova Science Publishers, Inc., Huntington, New York, 2001).
9. Kozyrev N.A. Selected Transactions (Leningrad University Press, Leningrad, 1991) (in Russian).
10. Lavrent'ev M.M., Eganova I.A., Medvedev V.G., Olejnik V.K., and Fominykh S.F. On Scanning of Celestial Sphere by Kozyrev's Sensor, Doklady AN SSSR, 323(4), 649-652 (1992) (in Russian).
11. Akimov A.E., Kovaltchuk G.U., Medvedev B.P., Oleinik V.K., Pugatch A.F. Preliminary Results of As-tronomic Observations of Sky by the Kozyrev Procedure. Academy of Sciences of Ukraine, Central Astronomic Observatory. Preprint CAO-92-5P, 1992, p.16 (in Russian.).
12. Eganova I.A. The World of Events Reality: Instantaneous Action as a Connection of Events through Time, Instantaneous Action-at-a-Distance in Modern Physics (Nova Science Publishers, Inc., New York, 1999).
13. Lavrent'ev M.M. and Eganova I.A. Physical Phenomena Predicted and Revealed by N.A.Kozyrev, in the Light of Adequacy of Space-Time to Physical Reality, Phylosophy of Science, 1(3), 34-43 (1997) (in Russian).
14. Jefimenko O.D. Electromagnetic Retardation and Theory of Relativity (Electret Scientific Company, Star City, 1997).
15. Khaikin S.E. Forces of Inertia and Imponderability (Nauka, M., 1967).
16. Sivukhin D.V. The General Course of Physics. Mechanics (Nauka, M., 1979).
17. Landau L.D. and Lifshits E.M. The Classical Theory of Field (Pergamon Press, Oxford. 1985).
18. Pauli W. Theory of Relativity (Nauka, M., 1983).
19. Peschevitsky B.I. Dynamical Criteria for Inertial and Noninertial States, p.1 and 2. Phylosophy of Science, 1(7), 79-84 (2000); 1(9), 99-109 (2001) (in Russian).

EAGLE-RESEARCH

Energy Solutions
since 1984

A Research Organization
that Develops & Distributes
Practical Energy-Saving
Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

Raum&Zeit

EhlersVerlag GmbH, Geltinger Str.
14e, 82515 Wolfratshausen
Telefon: 08171/41 84-76, 08171/41 84-66
Internet: http://www.raum-und-zeit.com

Institut für Raum-Energie-Forschung (IREF)
i.m. Leonard Euler

raum & zeit
Laser harmonisiert Blut und Wasser
El Niño

Global Scaling Theory
Gravitation theory
Mathematic Biology
Gravitation energetics
Hydrogen-Helium Energetic
Information technologies
Telecommunication technologies
Transport technologies

Time and its Physical Relationships

Andrew Michrowski

President, The Planetary Association for Clean Energy, Inc., 100
Bronson Avenue, # 1001, Ottawa, Ontario K1R 6G8 Canada.
(613) 236-6265; fax: (613) 235-5876
pacenet@canada.com

> *Time cannot be absolutely defined, and there is no inseparable relation between time and signal velocity.*
> *Albert Einstein [1]*

Time does not exist by itself

The phenomenon of time emerges in relationships – as an expression of properties of physical bodies and changes that occur to them.

Time is a factor of energy. Time has to do with the *increase* and *decrease* of energy. For example, as energy is brought down to a "zero level", time is "eliminated", so in a sense, time cannot be "compressed" - only eliminated. In the zero-energy level, electrons occupying this level in unlimited numbers are available through state transitions for the building of matter and the vacuum [2]. So it is the extent and the nature of energy flow that determines the characteristics of time.

How do we know all this?

In his seminal foundations of physics work of the early 1900s, **Sir Edmund T. Whittaker** demonstrated by canonical quantization that there exist physical, time-like and longitudinal photons *in vacuo* [3,4,5]. The 2 scalar potential functions, F and G which completely characterize an electrodynamic field due to electrons in the ether are:

$$F\,(x, y, z, t) = \sum \frac{e}{4\pi} \sinh^{-1} \frac{\bar{z}' - z}{\left\{(\bar{x}' - x)^2 + (\bar{y}' - y)^2\right\}^{1/2}}$$

$$G\,(x, y, z, t) = \sum \frac{e}{4\pi}\ \tan^{-1} \frac{\bar{y}' - y}{\bar{x}' - x}$$

In these equations, for the fundamental case in which a field is due to any number of electrons moving in any way, we observe that time emerges only through the displacement of energy. Another way of putting it would be that time emerges through change in energy.

These photons have an independent physical existence. Whittaker himself observed, after computation that the "total disturbance at any point, due to this system of waves, is independent of the time, and is everywhere proportional to the gravitational potential due to the particle at the point" [6]. **A. D. Sakarov** admitted that the gravitational field is a "conglomerate of loose things and not a fundamental field of nature at all [7].

Everything electromagnetic, and probably gravitational, starts from these potentials, not fields, and under certain circumstances, there may exist photons without fields being present at all. In the *vacuo*, the longitudinal light photon travels in the direction of the beam, like an energy capsule, as a scalar four-potential-function energy standing-wave field, with many different frequencies, with an internal symmetry based on circular polarization [8,9], an energy field or nexus that "has an end, but no beginning". The time-like and space-like parts of the four potential are photons with spin –1, 0 and +1 that are longitudinally directed, and which are observed in the ***Compton*** and the photoelectric effect [10].

Movement of light affects time

Philip S. Callahan designed an elegant experiment that shows how variations in the movement of light affect time [11]. Changes in exposure settings of photographs of same objects, including coherent light laser spots results in shift of position of images. The more coherent the light, the less apparent is a shift in time. He suggested that time is neither absolute nor independent of photon activity of space.

We can also state, as a corollary, that the *movement* of light generates time. After all, the electrodynamics is associated with photons. This is particularly significant in time engineering, as will be seen later.

Time and life

Now, it is well known that the ability of cells to sense the presence of light is a primary function of life itself. If a cell can sense light, it is alive; if it cannot, it is dead [12]. Callahan was able to observe that time increments were detected on light-detecting surfaces of living organisms such as the cuticles of leaves.

The Russian time researcher, **Nikolay A. Kozyrev** considered that living systems "consume" time for their life-energy [13]. **Velimir Abramovich** suggests that living organisms are naturally-driven "time machines". They each have an inbuilt time that serves as a "code" to structure their own physical totality and to regulate their own functioning. Thus their "local time" acts as a "time operator" ("time condition") frequency [14]. The nature and level of electromagnetic fields in living systems could therefore be considered as indicators of their "energy level" and how it affects and adjusts the inherent rate of time-flow.

Time and consciousness

We append to these statements the notion that, yes, time is "consumed" by living organisms – but only if they are conscious. Essentially, time can be perceived by measurement, which always requires a conscious observer.

When we measure, we observe the differentials of energy flow. It is the measurement of these differentials - as we note from the Whittaker equations of scalar fluxes - that allows us to the measure the equivalence of time. When we measure differentials, we are effectively creating our notion of time.

It is difficult for humans to conceive that time does not exist when humans think about it. **Saint Augustine** speaks for all humans with his observation:

> *While I do not think about time, I know that time exists. When I begin to think about it, I stop understanding what it is.*

With such conditioning, humans cannot come to grips with the notion that the Universe has no beginning and no end. It is like the paradigm of absolute nothingness – people cannot deal with the concept of no time because, effectively, consciousness, like life, consumes time. One could advance the notion that thought itself is time in motion. When we think, we are creating time. Life is movement – and the brain registers energy differentials and their associated fluxes of scalar potentials.

Time and causal mechanics

Since time is the result of movement of energy (and accordingly, light), then time is associated with the cause-effect processes linked to a first position of energy to the next. In other words, time is innate to causal mechanics.

Variations in energy flow, logically, lead to variations in the characteristics of time. They also permit variations in causality.

Conversely, changes in time lead to changes in physical structure [15]. This phenomenon was proven by Kozyrev and has been labelled as the ***Kozyrev effect*** [16]. He demonstrated that changes in the course of time affect the performance of electronic components. Figuratively speaking, a transistor is like a corral in which light behaves in a certain way. By changing the way that light behaves, a transistor becomes valueless.

To create an extra-temporal causality (with linkages to another level/dimension) one must change the light movement. Changing the light patterns means changing the behaviour of a wave, and its concordant photonic reality. By altering the photons, we make them go to another dimension.

Changes in time can effect change in living systems [17]. The engineering of "time-polarized" waves has the promise to target and amplify natural healing procession in living organisms. It is these natural activities that would restore diseased cells to their original and healthy condition [18]. Such time engineering may also include the reversal of the effects of AIDS, smallpox, anthrax, and most bio warfare agents, with treatment times of only a few minutes per week, with no more than 3 treatments required in all. Eventually it may also be possible to reverse the effects of genetic disorders, effect limb regeneration, and cure spinal cord injuries [19].

Engineering causality

Variation in the density of energy determines variation in the "the course of time". By the "speed of course of time" is meant the rate of causal transformation and the input of additional forces into systems (including mechanical tensions). According to Russian scientists [20, 21,22] there are interesting possibilities in deploying engineering for causal mechanics.

In preventive time engineering, one could delay the approach of a known cause and to artificially close the consequence loop, and thereby annulling it from ever achieving an effect. In other words, one can make the effect "happen" before its normal time, disrupting space structure with its related "speed of the course of time". One can also make it "happen" after its expected deadline. This technology could have interesting implication in strategic situations (preventing a extra-planetary body from attaining a collision course).

According to the Russian experience, when the spatial structure is disrupted by time-engineered causal mechanics, the affected region undergoes relative greater entropy (or, less order). The volume of space is forced out to somewhere else, generating torsion fields, much like a balloon will drift away from denser air. A similar phenomenon occurs when the velocity of a mass increases, the force emerges against it, called "inertia" [23]. Now, to increase velocity, the energy of the mass increases. In the case of time, as the movement of energy increases, interaction or reaction with another dimension ensues to compensate for that primary change [24].

The boundary layer between the two states of space can then act as a mirror and the approaching agent may be reflected back to its source. For example, a light beam may be reflected back to its emitter, in full, or in part, depending on the engineering.

With space-time engineering, we could develop teleportation systems [25]. The course of time goes from the past to the future, in the direction of greater disorder. Going back to the past represents deceleration. The etheric continuum is perceived by Kozyrev is containing variations of "density" of structural elements. The "denser" the etheric region, the slower is the course of time. A zone of accelerated time course would be forced out into "rarefied" ether. A zone of slowed time would be forced into denser ether, hence the basis for a teleportation technique.

In vivo experiments with a Time Machine

For over 15 years, a Russian association of scientists has conducted experiments with acceleration and deceleration of time with 4 prototypes of time machines

[26]. Light-heartedly, the units are called "muskrat traps" since the experiments conducted in remote forest were disguised as a high-tech electromagnetic technology for trapping muskrats. The time machine units are spheres ranging in diameters of 30cm, 1 meter and 2.1-meter. The shells are encased with coils designed to produce convergent waves.

Team leader **Vadim A. Chernobrov** describes converging electromagnetic waves as moving from a periphery to a central point. They are observed when a hoop is thrown into the water and inside the hoop the waves converge. If a potential is applied to do work and to initiate the energy differential process, the other reverse direction scalar (the reverse-time energy flow) must react. Thus, compensation of time - in the form of the deceleration or acceleration of the rate of time - can take place [27].

The first trials involved mice, in which most (25 out of 31) died. Eventually, there were successful 2-hours runs of time travel. An experiment with a dog that was clearly frightened also showed no ill effects. This led to experiment with humans, the first being **Ivan Konov** who, on August 26, 2001 decelerated into the past by 3% of planetary time during a half-hour trial. Dozens of others have experienced the phenomenon and report such sensations as: quicker pulse, giddiness, itching skin, body twisting, numbness at extremities and a case of an out-of-body experience. Harmful effects on living systems do not appear to be linked to the change of the rate of time, but rather to the variations of the time rate value among regions of a living organism.

Some individuals reported visual experiences such as "starry sky", "luminous vortices" and colour spots. Individuals observing outside of the time travel machine noted headaches. The most interesting phenomena occurred just before the start-up: significant presence of ozone for several hundreds of meters around the machine, which was located in a forest. Also noted were strange lighting effects in the sky above the apparatus, accompanied by sounds that inexplicably appeared to generated from inside.

Factors affecting the rate of time

The experiments yielded interesting observations: the phenomenon of the rate of change of time varies according to the hour of the day and according to the lunar phase. The rate can be influenced by a variety of external inputs, including mechanical vibration.

The transition into the future differs from than into the past. It is like movement from any point in a tree – where downwards represents past time. There are many paths possible towards the future – upwards, along the branches, but only one towards the past – downwards to the trunk. The return from the past time is possible only if the time traveler does not interfere with occurring events – or the possibility of returning to another branch of the tree. However, a return move from any variant of future time is possible regardless of the behaviour of the traveller.

The Russian time-travel experiments point to a relationship between inertia and time. In changes of rates of time, the region adjacent to the spheres develops a boundary layer effect, appearing as an aura of "white mist". The greater the time differential, the denser is the mist.

A similar phenomenon has been observed, and captured on film, with some experiments conducted by **John Hutchison** involving remote-controlled lifting and disruption of objects [28,29]. It may be that the ***Hutchison effect*** involves causal mechanics.

Time and frequency

Time may be viewed as a process – or "change-of-space" in any direction that does not exist in our dimension [30,31,32]. In physics, new properties are commonly acquired as the result of change in some property: charge, current, induced magnetic field, etc. Here, the new property becomes a new dimension. For the frequency of oscillations, the formula is $f = 3/\lambda$ [l/s], where λ is wavelength in [m]. Here, the velocity of light is equal to 3 (the 10^8 mathematical degree is omitted since it is a question of scale of measurement only).

Therefore, the analogy between our dimension and frequency gives us a new notion, which is curvature ρ: $\rho=3/r$ [l/m] and $f = 3/\lambda$ [l/s]. The 3-dimensional radius is represented by $R=\lambda/\rho= r/3$ [m] and time as a period of oscillation has the relationship: $T = \lambda/f$ [s]. **Time can be considered to be equivalent to the radius of curvature ρ if the linear radius r and the wavelength λ are the same. This is a condition for the spatial resonance effect.** Note that in this analysis, *"m"* and *"s"* are unlike when length is measured in meters in our dimension. However, for a new dimension it is possible to use equal units for (m) of 4th dimension and conventional "second". Furthermore, light (photons) is possible in our dimension only as a process in such a spatial resonator.

Experiencing time dilation

Based on the above discussion, it is proposed that we experience changes in rate of time in our daily life. During sleep, as our energy level decreases, so it can be argued that time decreases and we "go" into another dimension. Aspects of what is observed during the dream-state do not obey the rules of our familiarity of physical existence, and conventional causality - because the rate of time is different. As we start our dreaming, and as we emerge from dreaming, the recalled experiences resemble more our regular experiences. Before and after sleep, our brain frequency tends to resemble the daytime's. In effect, understanding our dream state's sense of time may be a reference key to comprehending the physics and causality mechanics of time engineering.

Time and planetary gravity

The Earth's mass appears to be in continual growth. 250 to 350 million years ago, our globe may have been half size - with all of the continents as one landmass [33]. In ancient sediments, the natural angles of slope in sand beaches greatly exceeded those of today, indicating that gravity has on our planetary surface has increased 8-fold, several times, during the last 1.5 billion years. Yet the planet's average density may not have changed – only the acceleration of free fall [34,35].

There could also be another explanation possible for the change of slope of beaches. The pull of the moon may have been different in the past.

Also, the planetary magnetic fields could have increased over the millions of years, through interactions with the solar flux of hydrogen atoms. The sun is in constant explosion – production of energy differential. Earth and other planets could be responding with harmonics to compensate for the solar activity, leading to an overall increase in magnetic fields. Such a phenomenon may give the illusion that the physical body is growing larger. In other words, with time, the force of gravity would alter planetary mass and energy.

Assuming that the Earth's density has not increased, it is possible account for, mathematically, the relative increase – growth processes - of the nuclei of terrestrial atoms, including the doubling of mass of nucleons and of electrons [36]. Such calculation also accounts for the emission of 2 different photons by hydrogen atoms (also known as the "red shift" described by **E. Hubble**).

The phenomenon can be explained with the time-like and longitudinal photons described by the bi-directional 2 scalar potential functions [37,38]. Could it be suggested that matter is gravity minus time? In other words, gravity is related to accumulation of energy differential (time) in mass.

The value of gravity varies throughout the planet, in part because of the poles and in part due to local density of matter. **Kirill P. Butusov** has noted a correlation between places of civilization and regions of greater gravity. A faster rate of evolution may be associated with such gravity zones. In these zones there would be a greater influx of scalar potentials. Time would be more "authenticated" by these energy flows into conscious beings. Butusov reminds us that time has a "positive energy" and flowing into nuclei of atoms or a "negative energy" flowing out of the nuclei of atoms [39].

The outflow would be representative of gravitational energy. Longitudinal waves are known to be able to enter and to leave nuclei. Such flows of time must come from other dimensions. After all, the surface of any elementary particle separates our dimension from another. This leads to an interesting possibility in which all time-associated processes between dimensions are synchronized.

Time reconsidered

Arguments have been laid that suggest that time is equal to energy differential (including movement of light). Living systems "consume" time as part of their consciousness and measuring processes.

It is possible to engineer causality. An understanding of higher-order electrodynamics is required. Techniques exist for the generation of scalar potentials. Causality technology has many applications. They include: therapeutics, energy generation, consciousness technology, inter-dimensionality, defence systems, teleportation, and of course, "time travel".

Appreciation is expressed for guidance in this
discussion by Thomas E. Bearden and
Bernard de Montréal.

References

1. Einstein, Albert. How I created the theory of relativity. [Translation by Yoshimara A. Ono.] /Physics Today. Volume 35 (8), August, 1982. p. 45-7. [Lecture given by Einstein in Japan, on learning of his Nobel Prize.]
2. Phipps, Jr. Thomas E. Heretical Verities: mathematical themes in physical description. /Classic Non-Fiction Library, Urbana, 1986. p. 397-405
3. Whittaker, Edmund T. On the partial differential equations of mathematical physics. /Mathematische Annalen. Volume 57, 1903. p. 333-355.
4. Whittaker, Edmund T. On an expression of the electromagnetic field due to electrons by means of two scalar potential functions./ Proceedings of the London Mathematical Society. Volume 1, 1904. p. 362-372.
5. Whittaker, Edmund T. A history of the theories of aether and electricity. Volume 1 and 2. London: Thomas Nelson & Sons, 1953.
6. Whittaker, Edmund T. On the partial differential equations of mathematical physics. /Mathematische Annalen. Volume 57, 1903. p. 354.
7. Sakarov, A. D., Vacuum fluctuations in curved space and the theory of gravitation. /Soviet Physics Doklady. Vol. 12, 1968. p. 1040-1. [English translation.] Original in: Doklady Akad. Nauk. USSR, Vol. 177, 1967. p. 70-1.
8. Evans, Myron W., Physica B. Volume 182, 1992. p. 227
9. Evans, Myron W., and S. Kielich (editors). Modern nonlinear optics. [Special topical issue of Ilya Prigogine and Stuart A. Rice: Advances in Chemical Physics. Wiley, New York 1992, 1993, 1997, and 2001] Volume 85 (2).
10. Institute for Advanced Study (Budapest). On Whittaker's F and G fluxes, Part III: the existence of physical longitudinal and time-like photons. // Higher symmetry electrodynamics: a collection of AIAS papers. [Special Issue. Journal of New Energy.] 2000. p. 7-1 – 7-5.
11. Callahan, Philip S. and Kevin. Sinewaves of spacetime. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 8-9.
12. Jibu, et al. Evanesecent (tunnelling) photon and cellular "vision". /Biosystems. Vol. 43, 1997. p. 65-73.
13. Kozyrev, Nikolay. A. Selected works. University of St. Petersburg. 1991. [In Russian]. Also: On the way to understanding the time phenomenon: Construction of time. A. P. Levich, editor. World Scientific Press. 1995.
14. Abramovich, Velimir. Experimenting with Time. /New Energy Technologies. Volume 3, November-December, 2001. p. 13-5
15. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001. p. 15-18.
16. Kozyrev, Nikolay A. The possibility of experimental study of the properties of time. /National Technical Service, US

Department of Commerce, Springfield, Virginia. (JPRS 45238). [Undated.]

17. Bearden, Thomas E. Renaissance healing, a concept whose time has come. /Proceedings of the United States Psychotronics Association. Volume 9, 2000. p. 1-2.
18. Bearden, Thomas E. Vacuum engines and Prioré's methodology: the true science of energy-medicine. I and II, /Explore! For the Professional. Volumes 6 (1), and (2) 1995.
19. Bearden, Thomas E. Private communication. February 2002.
20. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001 p 15-8.
21. Kozyrev, Nikolay. A. Selected works. University of St. Petersburg. 1991. [In Russian]. Also: On the way to understanding the time phenomenon: construction of time. A. P. Levich, editor. World Scientific Press. 1995.
22. Kozyrev, Nikolay A. and V. V. Nasonov. On some properties of time found by means of astronomical observations. / Collection: Problems of research of the universe. Volume IX, M-L, 1980.
23. Puthoff, Harold E. The energetic vacuum: implications for energy research. /Speculation in science and technology. Volume 13 (3). p. 247.
24. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
25. Frolov, Alexander V. Practical application of time rate control (TRC) theory. /New Energy Technologies. Volume 3, November-December, 2001 p 15-8.
26. Chernobrov, Vadim. Experiments with a man in the Time Machine. /New Energy Technologies. Volume 3, November-December, 2001 p. 6-8.
27. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
28. Calante, Pelayo and Andrew Michrowski (compilers). The Hutchison File. The Planetary Association for Clean Energy, 1996. p. 16-7.
29. Hathaway, George D. The Hutchison effect – a lift and disruption system. /New Energy Technology. The Planetary Association for Clean Energy, 1990. p. 77-103.
30. Frolov, Alexander V. Trilateral spacetime effect. / Newsletter of the Planetary Association for Clean Energy. Volume 9(2&3), April 1997. p. 10-13.
31. Ouspensky, Pavel D. A new model of the universe. New York, 1971.
32. Frolov, Alexander V. The application of potential energy for the creation of power./ New Energy News. Volume 2(1), May, 1994.
33. Kirilov, I. V. Increase in volume of rocks is one of the reasons of tectonic deformations. /Academy of Science of the USSR Geologist. Volume 1, 1963.
34. Smirnov, L. S. and Ju. N. Lubina. On the possibility of studying the change in gravity together with geological time. /Reports of the USSR Academy of Science. Volume 187, Number 4. 1964.
35. Maxlow, James. Earth expansion – the definitive proof. / Nexus. Volume 8 (3), May 2001. p. 47-51.
36. Butusov, Kirill P. Time is a physical substance. /New Energy Technologies. Volume 3, November-December, 2001. p. 8-13.
37. Bjerknes, K. A. Vorlesungen ueber hydrodinamische fernkraefte. Leipzig, 1900-2.
38. Yarkovsky, I. O. Universal gravitation as a result of formation of ponderable matter inside heavenly bodies. Moscow, 1889.
39. Butusov, Kirill P. Time is a physical substance. /New Energy Technologies. Volume 3, November-December, 2001. p. 8-13.

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad

B/W - $400

Color - $800

Half Page Ad

B/W - $200
Color - $400

Quarter Page Ad

B/W -$100
Color- $200

Ask about discounts for multiple ads

New Energy Technologies is distributed to scientific laboratories in 20 countries around the world.

To advertise contact us:
Faraday Labs Ltd, Phone/fax 7-812-3806564,
http://www.faraday.ru, office@faraday.ru

Time as Phenomenon of the Expanding Universe

Michael H. Shulman

7 (095) 357-80-33, 7 (095) 421-11-31
E-mail: shulman@dol.ru

Preface

In 1992 the author of this words, having been stimulated by Kozyrev's book [1], started to reflect on Time meaning and Time and Space asymmetric roles during the Universe Expanding. The development of these ideas has led to the work [2], it is available now (the Russian version only) on the website of the Institute of time nature explorations (grant #00-07-90211 of the Russian fund of basic research).

I would like to thank the Chairman of the Russian Interdisciplinary Temporology Seminar Dr. A. Levich from Moscow State University for his constant support and friendly interest. Also I would like to thank A. Moskowsky for the 20-year discussion of the physics history and philosophy.

The said work [2] pretends to revise radically a number of basic physical concepts of the Space - Time, Motion and Energy nature. It includes a detail analysis and mathematical calculations. Only a brief account of the main part of this work is presented below.

Introduction

The nature of time is not yet enough clear for natural science. In Newton mechanics time was presented as some universal formal parameter. Its value rises steadily at every point of the Universe by unknown for us reason. Each physical process occurs in space in correspondence with the time course.

In the Special Relativity (SR) time and space are integrated to the common 4D-continuum. However, in this theory the time component having imaginary factor seems also to be "exotic". In this concept the increase of time is also implied in each reference frame.

The General Relativity (GR) allowed linking the time properties with gravitational fields and the space geometry. The time currency started to be associated with a spatial expansion of the Universe.

The theoretical physics traditional approach to the process description is based upon the considering of time course as primary (original) one. There are also in the modern physics [3] several attempts to deduce the time concept as secondary one from different fundamental (microscopic) concepts. However, the third way (inverse to the first one) is possible and forms the basis of this paper. A starting point of this way is the following question: "Does any universal process exist which could *generate* physical time?"

(Editor's note: The same question was formulated and the answer was proposed by Alexander V. Frolov in 1996, report "Matter as process", Scientific congress "New Ideas in Natural Sciences". It was assumed that similar process can be produced by special technical methods also.)

Such fundamental cosmic process really exists and it is well known in the modern science. **It presents the Universe expansion** and was opened at the first third of the 20th Century by the American astrophysicist E. Hubble and others [4]. It means the general increase of distances between all 3D-bodies. The same scattering of two-dimensional-figures happens on the surface of some spherical balloon during air incoming. The centre of this sphere does not belong to the surface; all points of the sphere (the Universe) are equivalent.

Some time earlier the theoretical physics had come to the same results. As it is well known, the Einstein's GR was published in the 1916. After that Friedmann (1922) proposed the concept of the expanding Universe. For example, in the book [5] a description of the basic cosmic model is given. Hereinafter this model is called "Einstein-Friedman model", or "EF-model". In this model the Universe is presented as 3D- hyper surface of a 4D- sphere with increasing radius. Of course, the curvature of the 3D-hyper surface increases with time too.

Basic hypothesis relative to time nature

Some simple and pictorial views consist a basis of the new concept. Hereinafter it is called briefly as "The Spherical Expanding Universe Theory (SEUT)".

In the SEUT, as well as in the EF-model, in every time the Universe represents the 3D-hyper surface of a 4D-sphere. However, there is one very important difference. In the Einstein's theory the spatial components of the metric tensor are opposite in sign to the time's one. For example, we may consider time as imaginary quantity, then spatial coordinate as a real one. On the contrary, in the SEUT the 4D-continuum is considered as purely

Euclidean, all the four coordinates are real quantities. The usual spherical geometry can be used on a surface of the 4D-sphere.

As it is well known, in the GR the Age of the Universe can be *calculated* using the EF-model or a similar one. Usually, the radius-age dependence is not a direct proportionality in such GR models. However this result may be deduced if to neglect the global pressure of matter that fills the Universe. In [2] it is demonstrated, that an account of the static pressure of matter follows to the Universe radius linear dependence on its age (see below).

On the contrary, in the SEUT the time universal course is *manifested*. The Universe age is *identified* with a current Universe radius divided by the velocity of light. Numerous important consequences may be deduced from this statement. On the other hand, it allows avoiding many other arbitrary postulates in the model.

Mechanical motion and maximum velocity in the SEUT

The SEUT states, there is no unlimited set of independent mechanical motions. Only world lines of moving bodies exist. Each of them has some inclination relative to the time line, which presents a normal to the hypersurface of the 4D-sphere. It is *an inclination angle that defines the spatial motion velocity*. At the increase of sphere radius the intersection point of word line and current hypersurface "moves" with exact correspondance with a modern physics prediction.

In particular, immovable objects (stars) have zero inclination, their world line are normal to the hypersurface. Hence they "scatter" according to the Hubble law, their mutual velocity is proportional to mutual distance. If a body world line has some inclination relative to normal, the angle is more than zero. But it can't exceed 90°, therefore *maximum* mechanical motion *velocity* appears *naturally*, it is equal to the velocity of light.

Let us consider three variants of motion (see Fig.1).

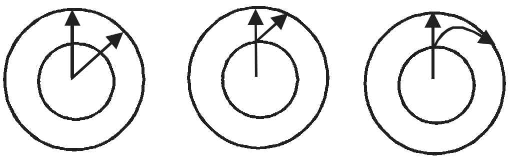

Fig. 1
Phenomenon of a "motion" of point on a sphere surface for immovable (at the left), uniformly moving (at the centre) and non-uniformly moving (at the right)

The left picture illustrates the Hubble effect. The right picture presents a general case of motion with acceleration. The central picture corresponds with an inertial motion; its world line is direct. In this case the moving body displacement increases proportionally to the Universe radius increment. So, *inertial* motion is not postulated in the SEUT, it appears as natural model consequence.

At a large 4-sphere radius values all the relationships of SR and usual mechanics laws are applicable approximately in the SEUT. A Special Relativity light cone transforms to all the hyper surface of the 4D-sphere. But the analogy is not complete, because an absolute remote SR area degenerates to this 3D- hyper surface in the SEUT (see Fig. 2).

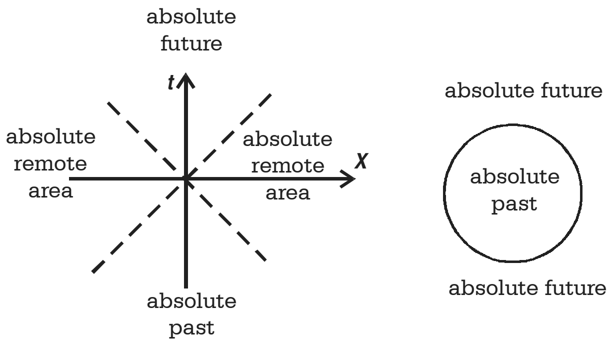

Fig. 2.
4D- continuum areas in the SR (at the left) and in the SEUT (at the right)

SEUT and Minkowski geometry

Let us consider small increments of time and space coordinates along body world lines during the Universe expanding. It is enough to consider a small area of the Universe, so we can neglect its curvature. Then it is acceptably to replace approximately concentric hyper surfaces ("isochrones") by parallel hyper planes. The space-hold corresponds with a representative point "drift" perpendicularly to isochrones, an inertial motion corresponds with displacement along inclined direct lines between isochrones.

At each time the Universe is represented by a certain isochrone that contains all the real spatial points. Let us accept that 4D-sphere radius increment divided by velocity of light presents invariant measure of (absolute) time increment. We will also state that this quantity has the same value in each *inertial* reference frame, i.e. at a motion along each *direct* world line.

Let the angles of world line inclinations from normal direction are enough *small*. Then metric relationships like Minkowski geometry ones appears in our purely Euclidean 4D-continuum. In particular, well known relationship

$$c^2 ds^2 = c^2 dt^2 - dr^2$$

can be deduced from the Pythagorean theorem. It connects a spatial component *dr* with a time component *dt* (at moving reference frame) through velocity of light *c*. Here *ds* is an absolute time interval (between two 4D-events at a immovable reference frame). **Hence, if velocities aren't very high, the Lorentz transform is correct in different inertial reference frames.**

On the Einstein's relativity principle

If the world lines inclinations from exact normal direction cannot be accepted as small, then Minkowski geometry relationships are correct approximately only. It means that Einstein's relativity principle is correct (in the SEUT) only for reference frames that move with velocities enough small relative to *selected* reference frame. Such reference frame is linked hardly with a body at (absolute) rest, i.e. drifting along radial world line.

The selected reference frame existing reminds of old ether theories that contradict to the Special Relativitiy views. It seems, these theories became a thing of the past irretrievably. In fact, the velocity of light in vacuum is constant everywhere and everywhen. However, the reference frame existing can be detected in principle as a light signal frequency bias, i.e. with the help of Doppler effect. Well, this phenomenon is really detected by the modern astrophysics!

The temperature diagram of the Cosmic Microwave Background Radiation (CMBR) coming to the Solar system from all the sides of the Universe is presented on Fig.3. The data was registrated during 4 years by the Cosmic Background Explorer (COBE) satellite (NASA Goddard Flight Centre, COBE Science Working Group).

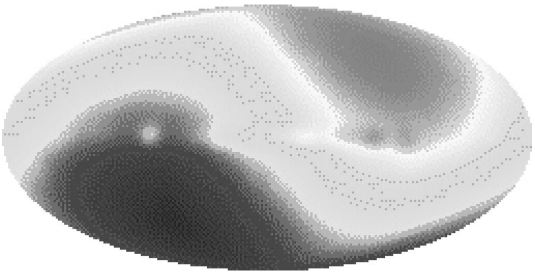

Fig. 3
The anisotropy of the Cosmic Microwave Background Radiation (CMBR)

The well-known Russian scientist J.B. Zeldovitch in the Editorial Addition to [6] in connection with earlier experiments notes that careful measurements **allowed to find out some anisotropy of CMBR.** An antenna oriented to the Lion constellation detects that the radiation temperature is 0.013% more, than mean temperature. The radiation temperature in the opposite direction is 0.013% less, than mean. **Generally, a temperature varies continuously between these two values.** The isotropy presents only for some imaginary observer. ***The Solar system, Earth move to the Lion constellation relative to this observer having velocity 390 ± 60 km/s.*** Hence, as a result of the Doppler effect, a incoming radiation seems to be more hot, and an overtaking radiation seems to be more cold. This example shows that for observer of any point of the Universe this CMBR is isotropic. We may consider this observer and the connected reference frame as selected one. The selected reference frame existence at the Universe every point looks like the physicists commonly held view preceding to Relativity. They thought that the light presented ether oscillations occupying whole the Universe. They thought also that a reference frame connected with ether was preferable, or selected. They tried to detect the Earth motion relative to ether. We know that these experiments gave the negative result: any ether doesn't exist. But the Universe evolution follows that when CMBR is observed (and only in this case!), the selected reference frame (called sometime "new ether") appears. ***The new ether or CMBR just realises the motion according to Hubble's law.***

The modern NASA's data allows to put the relation of the Solar system velocity to the velocity of light equal to 0,15%. It is enough small value justified Special Relativity and Minkowski geometry relationships application. But can we believe this phenomenon to be an exhaustive proof of the SEUT accuracy?

To test it we propose a not complicated observational experiment. If the CMBR anisotropy is due to the real selected frame existence, then it may be detected for any electromagnetic radiation. In particular, an anisotropy of solar radiation has to exist at the different year periods. It has to be detectable in August, when the both solar radiation and CMBR come to the Earth from the Lion costellation side (see Fig. 4). In February these sourses are opposite in disposition relative to the Earth, therefore the solar radiation anisotropy direction has to be opposite. The expected effect value (with account of the Lion constellation straight ascendancy and obliquity of the ecliptic) is approximately equal to 300 km/s, i.e. nearly 0,1% of the velocity of light. In November and May the anysotropy has to be practically absent.

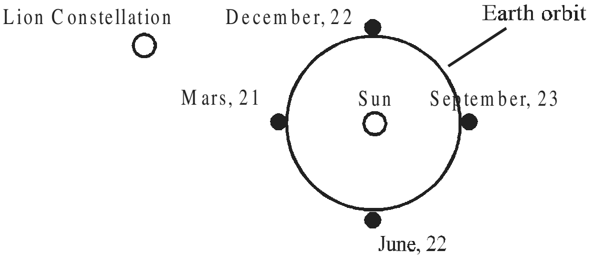

Fig. 4
The Sun and Lion constellation disposition relative to the Earth

Probably, an analogous SEUT test in a laboratry is realizable with help of artificial radiation sources.

Particle mass, energy and impulse

So, we consider the Universe as expanding 3D-hypersurface of a 4D-sphere. Mass localisation places in the Universe present the points of the hypersurface intersection by world lines. So, **these world lines have a real physical meaning, not abstract illustrative** this one. We may expect this physical meaning to be more essential than simple word expression.

Particularly, while the Universe global analysis is making, we may suppose that such fundamental particle feature as its mass at rest presents some *relative* value. Such relation (some kind of a quantum number) may include, for example, 4D-sphere (the Universe) diameter and some characteristic size like de Broglie wave period that is inversely proportional to the mass. This hypothesis may make clear *inertia* nature as two characteristic times *relation*. It may also explain the *rest-energy* notion.

Meanwhile, the Universe radius increases with time. Well, what happens to mass? If a de Broglie wave period increased proportionally to time, we couldn't generally detect the Universe expansion, including famous "red shift". But if particle wave periods are constant, then *matter mass has to rise proportionally to the Universe age and size*.

In the Relativity (like Minkowski geometry) we use vectors having imaginary projection to time axis and real projections to space ones. Particularly, it is true for velocity, acceleration, and energy-impulse 4-vectors. As against, vectors having all the real components are only used in the SEUT. At that, a 4-interval value (length in pseudo-Euclidean space) of some relativistic vector answers the absolute time axis projection of a corresponding SEUT-vector, and imaginary component of a relativistic vector (time of motion) answers the corresponding vector length in purely Euclidean continuum of the SEUT. For example, the energy-impulse vector module presents such quantity. Its projection to absolute time axis is energy at rest divided by velocity of light, and its projections to spatial axis are impulse components. This quantity is *constant* while the particle movement is *inertial* one.

A jump to *non-inertial* motion in the SEUT is connected with a corresponding state vector *changement* law. So, if particle motion velocity changes, its *energy at rest* doesn't change, therefore full acceleration at a time interval can be calculated using the difference between new and old impulse values. Thus, the non-uniform motion equation in the SEUT can be found like SR as time derivative of an impulse expression.

The force-acceleration relation depends on a mutual orientation of force and velocity vectors in the both SR and SEUT. But in Relativity a reference frame velocity can be choised arbitrary, for example it can be zero, then the relation will be equal to one.

On the contrary, in the SEUT an *absolute* velocity is presented, it is defined by the world line inclination relative to the normal. Let the Earth move with any velocity relative to the immovable (selected) reference frame. Then we will be able to detect the absolute velocity using two measurements, the first one along the world line, and the second one in a perpendicular direction.

If this absolute velocity is really defined by direction and value following from CMBR anisotropy effect, then we can expect a relative difference near 2,25 $\cdot$ 10^{-6} between longitudinal acceleration and transversal one.

Local gravitational fields of particles

What does the SEUT talk about body gravitational fields? Let us imagine all the bodies as immovable and drifting exactly along the *radial* world lines. If there is a mutual gravity attraction effect between two bodies in such Universe, an observer will detect some curvature of their world lines. They will seem to be bending one to another instead of a radial divergence. In essence, in this case we may replace a world line by a gravitational field line. Then the analogy allows us to identify an Universe isochronous intersection with an equal potential surface that these field lines have to be *normal* to this surface. So, we arrive to a presentation that a Universe isochronous intersection is not strictly concentric hypersurface. It is perturbed by some kind of craters (see Fig.5), that centres correspond with gravitating bodies.

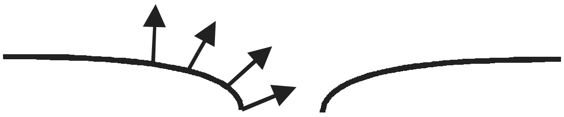

Fig. 5
Local body gravitation field

The inclination angle of a crater profile relative to non-perturbed sphere hypersurface is equal exactly to the inclination angle of a normal relative to strict radial direction. Hence, a local gravitation field intencity measure agrees practically in each point with body velocity measure that we used earlier. It authorizes energy concept using for both mechanical motion and gravitation phenomenon.

SEUT and General Relativity

Let us discuss some GR's aspects. Is it acceptable to neglect pressure of matter? When Einstein searched for his early cosmological static model solution, he had to introduce a cosmic constant in his equation. This constant answered a ***negative matter pressure***, that Einstein could not determine a meaning. In a non-stationary model a solution exists independently on cosmic constant presence, therefore it may be put often as zero. As rule, bodies' velocities may be put as zero too; therefore (*dynamic*) pressure is usually neglected.

However, we insist on necessity to account a *static* pressure of gravitating matter. Really, it can be ignored in the case when Einstein's relativity principle is applicable. Accordingly with it a gravitation field can be *always* replaced by reference accelerated frame. In this case a purely kinematical side is only accounted. However, *not every* field may be considered (even locally) as uniform one (see Fig. 6). Let the radius of a field source (or a *probe particle*) have the same order that the mutual distance. Then the Einstein's equation connecting space geometry with matter physical features seems to be incomplete. More precisely, it is incorrect to put exactly equal to zero a static pressure in the matter density tensor, it is necessary to introduce its (unknown, calculable) value accounting **the *continuum deformation energy*.**

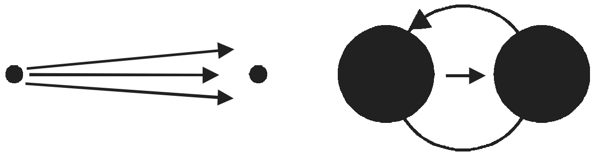

Fig. 6
At the left a locally uniform gravitation field is presented, the right field can't be considered as uniform one even locally

Are the proposed amendments important? Specialists know, that the Metagalaxy gravitation radius is *in fact* comparable with its real size. **In [2] we have showed that the Universe radius was *less* than its gravitation radius.**

It is shown also in the same work [2], that a resulting gravitational pressure in the Universe is *negative* and it is responsible for the Einstein's cosmic constant. This fact seems to be evident because a gravitation force aims to *gripe* any matter accumulation. **But it is possible to conclude it by a non-trivial way.** Let us consider a matter pressure as a reactive force operating within a uniform sphere of incompressible liquid. A pressure dependence on internal density is presented in [7]. We can see [2], **that if the sphere gravitation radius exceeds its geometric radius not more than approximately 1% (or it is still less), then a sudden negative change of the pressure appears at the internal abroad. This phenomenon may be explained as a volume "expansion" due to a metrics perturbation.**

The consideration of static pressure allows not only to find out a *new* (linear in time) cosmologic solution, but also to calculate a dependence of the Universe gravitational pressure on the radius of the Universe. This *negative* value has the representation like that one for a non-relativistic sphere (a star or a planet, for example). Also we would like to note, the formulation of density-Universe radius in the SEUT is exactly the same that the formulation of so-called *critical* density in the EF-model.

Two very important circumstances are clarified for all that. First, the **Universe mass was turned out as *linearly increasing* function of the 4D-sphere radius, and it is not a constant. A famous Einstein's programme is realized unexpectedly in the SEUT: a matter features (density) are reduced to a space features (curvature).** In other words, a necessity to introduce a mass distribution in the equations externally ("by hands") is eliminated in the SEUT. This operation is need in the GR to find out a spatial metrics changement law.

Secondly, the seeming paradoxicality of University mass (and energy!) non-conservation make us to reflect on conditions that the *accomplishment* allows to the energy conservation law correctness (see Fig.7).

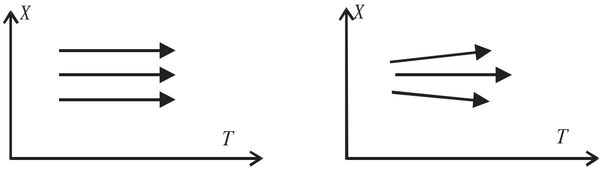

Fig. 7
If space features *do not depend* on time, then the closed system energy is *constant* (at the left). But if space features *evolve* in time, then the isolated system energy *have not to be constant*.

It is evident, we believe, that the energy can be exactly constant only in such physical system (or in whole the Universe) for which space features (in particular, a curvature) are strictly constant in time also. However, the both modern physics and SEUT issue from the opposite concept. Hence, this state can only be accomplished approximately, moderately of a bit of the modern rate of a relative space curvature evolution. This rate has order 10^{-10} per year for the modern Universe.

N.A.Kozyrev [1] basing on astrophysical observations stated the common star radiation origin that has to be due to the time-energy transformation. Accordinly with the SEUT, the relative increment of star mass (and its rest energy) is equal to the Universe age relative increment. It is interesting that the Sun mass relative lost due to the radiation consists in 10^{-15} per year, i.e. five orders less than mentioned above energy increment.

Universe origin and closed geometry

The cosmological EF-model could not say anything on the Universe origin. On the contrary, the work [2] approach allows an obvious way to study the problem.

As it is noted in [7], the metrics of any sphere area having a non-zero density is perturbed relative to Euclidean one, its geometry agrees with 4D-sphere hypersurface geometry. For non-collapsing sphere its gravitation potential relief is like a very small *"pit"* that gravitation radius is much less than its geometric size. However, when the density rises, the metrics perturbs more and more, and the pit transforms to some kind of *"crater"*. The crater is connected with the external surface by a narrow neck. Only this neck or its part is visible for an external observer, and the gravitation insurmountable barrier transmutes an object central area into "a lost world".

From the point of view of the external world, the central area presents a "black hole" absorbing irreversibly all the matter and radiation. On the other hand, for our Universe inhabitant the "navel-string" connecting with the external world has to seem a spherical "white hole", to which a matter and radiation are coming continuously (and, may be, carry out an information on the external world features). There is an old Russian fiction science book called "Sannikov Land", where an internal gigantic trench concerning central small star is described. Our model seems to be like this picture.

Is it possible that we live in such a black hole? The present hypothesis answers affirmatively this question. The negative sign of the matter pressure due to continuously increasing of our world size allows to such conclusion. And the University self-closing can be physically explained by the way.

On the other hand, as it is shown in [2], in spite of matter average density negligibly small, the Universe gravitation radius is more than its geometrical radius; hence, it presents a black hole. This condition accomplishment agrees the Universe self-closing, the boundary absence in spite of its finite volume. Also, this fact confirms our assumption that the Universe rest energy localisation area does not exceed the gravity operation area.

A specific model of star collapse is created in the modern General Relativity. In general, it can study in *three* different reference frames. As a rule, the "point mass" models are used. *The first* model is linked with an external observer; the *second* one accompanies a matter falling to the black hole. *The third* model presents an internal reference frame, i.e. an observer within collapsing object.

From an external observer's point of view the matter falling time to the black hole is infinitely large. However, in an accompanying reference frame it is finite. Since in the accompanying reference frame time and space coordinates are expressed through the both types of external reference frame coordinates. What is more, in the internal reference frame the time and space coordinates quite trade places, the metrics tensor components are depending on time. Further, any matter point history in this accompanying reference frame starts at the zero moment and finishes after same universal time period in a special (singular) point, after which nothing exists ("time barrier").

As we believe, another lacing between internal and external collapse pictures will be possible, if we consider a non-point collapsing object. Nobody wonders now at a situation, when a time period can be finite in one reference frame and infinite in another one. **Therefore, we can believe, unlimited black hole *collapse* in the external Super-Universe may seem to present unlimited *expansion* of our Universe observing inside.** This expansion seems to *start* from a singular point, and the same point is the history end of all the matter of the external Super-Universe that fall to the black hole. I would like to note especially, it does not mean that internal time pass in opposite to external one. **Rather, it is possible to state, time within a black hole passes ortogonally to external one.**

References

1. Kozyrev N.A. Selected works. - Leningrad University, 1991.
2. Shulman N.H. Site of the Institute of time nature exploration (Russian-language version): www.chronos.msu.ru , Project of Chair-laboratory of time as phenomenon of expansion, 2001.
3. Vladimirov Yu. S. Relational theory of space-time and interactions. Part 2. Theory of physical ineractions. Moscow university, 1998.
4. Sharov A.S., Novikov I.D. A man, who discovered the Univerce explosion. Edvin Hubbl's life and work. Moscow, Science, 1989.
5. Einstein A. The essence of Relativity. Moscow, IL, 1955.
6. Wineberg S. First three minutes. The modern approach to the Univerce origin. Moscow, Energoizdat, 1981.
7. Tholmen R. Relativity, thermodynamics and cosmology. Moscow, Science, 1974.

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

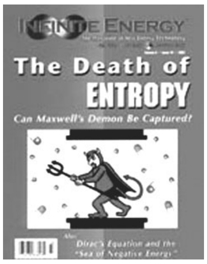

Infinite Energy Magazine
❊ Cold Fusion ❊ New Energy
❊ New Science
❊ New Technology

Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign

Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

ELECTRIFYING TIMES

an online and published magazine about Electric, Hybrid, Fuel Cell Vehicles, advanced batteries, ultra capacitors, fuel cells, microturbines, free energy systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd, Bend Oregon, 97701
541-388-1908, fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com

website: http://www.ipsgeneva.com

The Experiments on Thermo-Gravitation

This review of the article by Alexander P. Schegolev, the scientist from Saint Petersburg, was received by NET editors from our reader vladrim@mail.ru and presents his own view about Schegolev's experiment [1]. The opinion of our editorial board is not fully in agreement with the author's conclusion. Also we have no information who is vladrim@mail.ru.

It is known from the science history that the very experiment gives a push to revising of old knowledge; it also checks the new designs and conclusions. The knowledge of physical theories is based on the experiment; the experiment confirms hypothesizes or refuses them. Making experiments, we ask questions to the nature. And it always answers us on the clear pointed question. However…

Michael Faraday was trying in vain to find out the relationship between gravitation and electricity. These experiments were repeated on the modern level. Further, there was made an attempt to screen gravity force, the influence of environment on gravity was also trying to be researched. Many efforts were spent on finding out of the gravitational radiation falling to the Earth from depths of the Universe, as well as artificial generation of gravitational waves.

The experimental studies of weight change must be under attention. In particular, the attempt to find the differences in weight after sharp turning of body from the quiescent state into the state of rotation around its own axis was undertaken. The experiments of the influence of strong magnetic field and temperature on the weight of a body were made.

But alas! No changes of weight were discovered in previous experiment *(Editorial: Perhaps, the author has no information about sucessful results. We reported about such results in our magazine)*. Analyzing these experiments and theoretical premises, the author of the given article has made an unusual experiment imitating the heat motion in the Earth from core to peripheries. For that experiment there was used the steel ball with 100 mm diameter. The cone hole was made in the ball up to its center. The ball was installed by the hole upwards on laboratory scales with 50 mg scale factor and laser beam was directed inside (into the hole). The directed heat flow outgoing from the center of the ball was created by this way. While the surface temperature increasing, the arrow of scales became to be rejected aside weight decrease. The temperature was measured by the contact thermocouple. After an hour and half, approximately, when the temperature was 300 degrees Celsius the laser was switched off. The difference (the decrease) in weight against initial value (in the cold state) was four grams per 4200 grams of the test body.

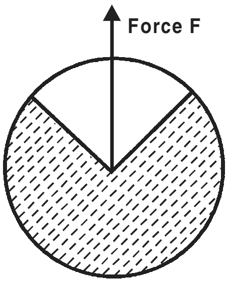

Fig. 1
The ball with cone hollow

Further, the ball was getting cold slowly on the scales and its arrow was creeping to the initial position. For acceleration of this process, the ball was periodically blown on by pressure air. It was the check, if the air (heated by the ball) influences on scales or not. The stability of its data independently on actions of pressure air stream excludes such suggestion. When the temperature of the ball became the same with room temperature, its weight return to the value nearly initial. However it is necessary to add, that this "nearly" is about 200 mg. This confirms that the change of the weight was not occurred because of the evaporations of metal from action of laser beam, and we were the witnesses of absolutely unusual phenomenon. *(Editorial: Perhaps, it is necessary to take into account the deformation).*

That is all concerning the question, which was given to nature by the author of the given article. However, one more doubt remains: did the heat radiation (coming from the ball) affect on mechanism of scales? To check this, the control experiment was made, under which the same ball was heated by usual way in the electrical furnace. In spite of its temperature was about 600 degrees Celsius and the ball was placed on scales for a long time (until full cooling down), the arrow did not move from the initial position. Thus, we have really turned out face to face with a phenomenon requiring an explanation.

If the weight change in this experiment has happened in consequence of gravitational interaction, then, therefore we must revise some fundamental concepts. Today it is difficult even to assume the consequences. The only thing, which is possible to say certainly, is **that divergent on radius uni-directional (single-vector) heat flow** is the object for observation, absolutely unknown to us or unnoticed before, which holds the ensemble of surprises. Up to now, we have dealt only with **chaotic** heat motion, which was researched by the thermodynamics and heat transmission.

For analogy, it is possible to give the example **of the electric current arising, which is possible to get only under directed electrons moving.** In general, more than ten experiments were made, and all of them have given the same result: the weight of the body was decreased. Who will solve this enigma?

Editor's note: The question is presented above. The answer is the aetherodynamics theory. Any directed (unidirectional) heat flow is also a flow of aether. The mass changes are the demonstration of natural mechanism of existence of this mass as aether vortex (Alexander V. Frolov).

References

1. Collected reports "Untraditional scientific ideas about nature and its phenomena", #1, club FENID, Gomel, 1991.

UP-TO-DATE INVENTION

The RotoVerter from arkresearch

Editor's: The description of this invention was received from Hector D. Perez Torrez, designer of the "RV" Rotoverter and TV transverter OU "transformation" devices.

Here is the first prototype, used to test a 3PH generator. Light bulb was a 100W 120V, AC generator required 380W for excitation to 120VAC plus the light bulb load for a total of 480W. Input to rotoverter was 240W for a total of 200% efficiency from the prime mover. Friction loss and prime mover loss was not calculated, PF unity, 120VAC unto a purely resistive load excitation pure DC120VDC3.1666667A (380W) + 100W = 480W. -240W=240 excess.

It is a common 3PH dual winding 230/460V motor wired to 480V run at 120v as a ROTOPHASE converter. That's the beauty of it, OFF the shelve components preferably totally enclosed motors (so fan can be removed) low friction bearings, from 3HP to 7.5HP. Re-rating hp can be from 1/7 to 1/3 the original HP (some motors can run up to 1/2 HP capacity but will require extreme care and cooling.... and exact tuning to load).

System requires starting and running capacitance; the perfect starting capacitance is the one that gives same plate current value as 480V but running at 120V. The best running capacitance is the one that gives the lowest Amperage reading at programmed load. (ALL capacitors are AC 370V OIL "no electrolytic").

Some motors are better than others; also for generator use you must choose the over-unity generator and motor combination to obtain OU. It looks easy to gain it, but requires quite a good choice of items and common sense. In this Research we have found even a few turns or a run weld in stator can change drastically the results. Up to date best performance has being obtained from US (motors 7.5 HP, BALDOR 3,5, HP, GE 3HP totally enclosed, dual winding 230/460V). Higher HP will require professional expertise in industrial electricity and power eng. (not recommended for amateur experimenter). Anyone engaged in this experiments must be familiar with safety procedures and basic electric and mechanic knowledge. Voltages and rotating machinery must be handled with great care. SEE basic diagram and connection.

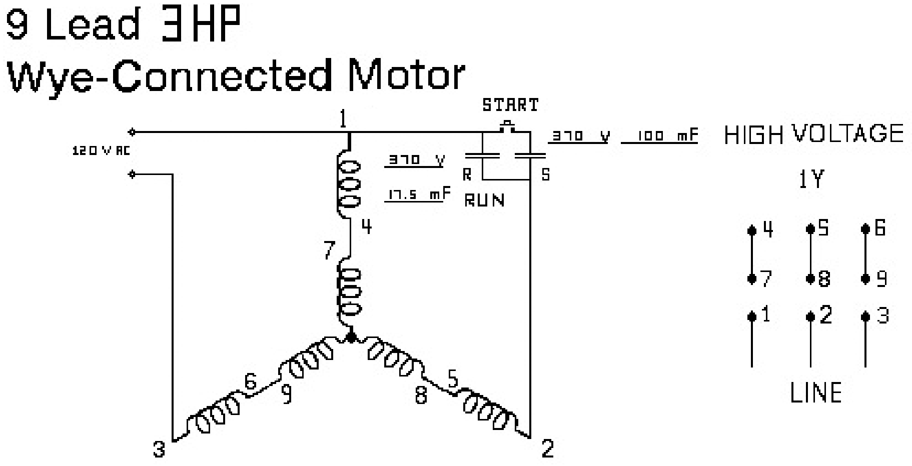

Fig. 1

Latest UPDATE BELOW 7-01-02

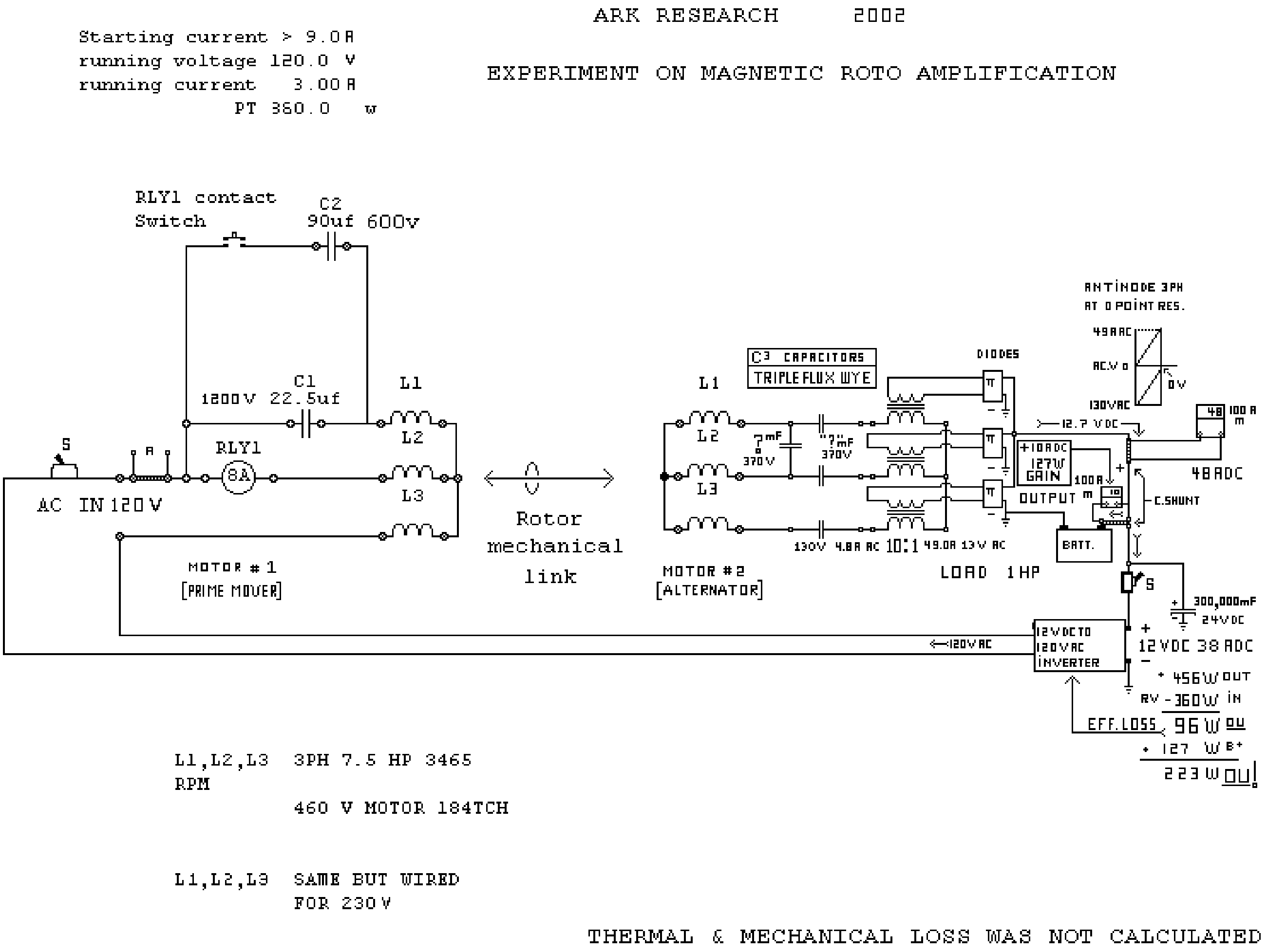
ARK RESEARCH 2002
EXPERIMENT ON MAGNETIC ROTO AMPLIFICATION
Starting current > 9.0 A
running voltage 120.0 V
running current 3.00 A
PT 360.0 w
RLY1 contact Switch
C2 90uf 600v
C1 1200V 22.5uf
L1
L2
L3
S
A
RLY1
8A
AC IN 120 V
MOTOR # 1
[PRIME MOVER]
Rotor mechanical link
L1
L2
L3
MOTOR # 2
[ALTERNATOR]
C3 CAPACITORS
TRIPLE FLUX WYE
?mF 370V
"?"mF 370V
130V 4.8A AC 10:1 49.0A 13V AC
DIODES
ANTINODE 3PH AT 0 POINT RES.
49 A AC
AC.V 0
130VAC
0V
12.7 VDC
48
100 A m
+10 ADC 127W GAIN
OUTPUT
100 A m
10
48 ADC
C.SHUNT
BATT.
LOAD 1HP
S
300,000mF 24VDC
12VDC TO 120VAC INVERTER
120VAC
12VDC 38 ADC
+ 456W OUT
RV - 360W IN
EFF.LOSS 96 W OU
+ 127 W B+
223 W OU!
L1,L2,L3 3PH 7.5 HP 3465 RPM
460 V MOTOR 184TCH
L1,L2,L3 SAME BUT WIRED FOR 230 V
THERMAL & MECHANICAL LOSS WAS NOT CALCULATED

Fig.2

Subject: RV Basic level 1 Full MODEL Disclosure (Public domain)

This is the RV (Level 1) plan of a full looped RV system:

Battery provides primary power for 12VDC to 120VAC. As to run prime mover (Roto-converter), second motor acts as ansquirrel cage self-exited generator, a triple flux-capacitor LC tank tuned to best Standing wave condition as to create standing wave current node internally in battery at 0 voltage to battery "negative resistor". At 0 volts "voltage" a negative current is created as to maintain a reverse flow (charge) to battery exceeding the forward drain of the inverter demand, detuning system with a forward charge at 10 amperes with a voltage rise of approximately 0.83333 V over the battery. Voltage charge produces OU transform from the 0 point standing wave component. System gains energy from stochastic resonance within the LC tank components draining energy from "thermal" signature of the Aether and K thermodynamic-thermoelectric ambient heat (Electron spin). This is a full disclosure of an operational and tested device. System is made of standard off shelve items. Tuning is made by changing capacitor values and the proper selection of standard items for its construction, 3PH motors, 10:1 12V or 5:1 24V transformers with the proper core and winding values (standard) off shelve, diode bridges capacitor (all standard).

Warning: System opens (NON standard) space-time anomaly as time is reversed due to aether energy transform....
Extreme warning: Do not exceed 10KW Nor use of trans-uranic elements near unit as they may reach critical mass by regressive "inversed" decay. PB 206, U235, U238 must be maintained as far as possible from unit.
Extreme Warning: Dangerous Toy, Eldridge, Event horizon and hell raiser effects may be created by protoplasmic fields at extreme potentials PK.
Use wisely. You Can create Paradise or open the gates to Armageddon.

The choice is yours now....

Hector D Perez Torrez

Lifter Project
by Jean-Louis Naudin
http://www.jlnlabs.org

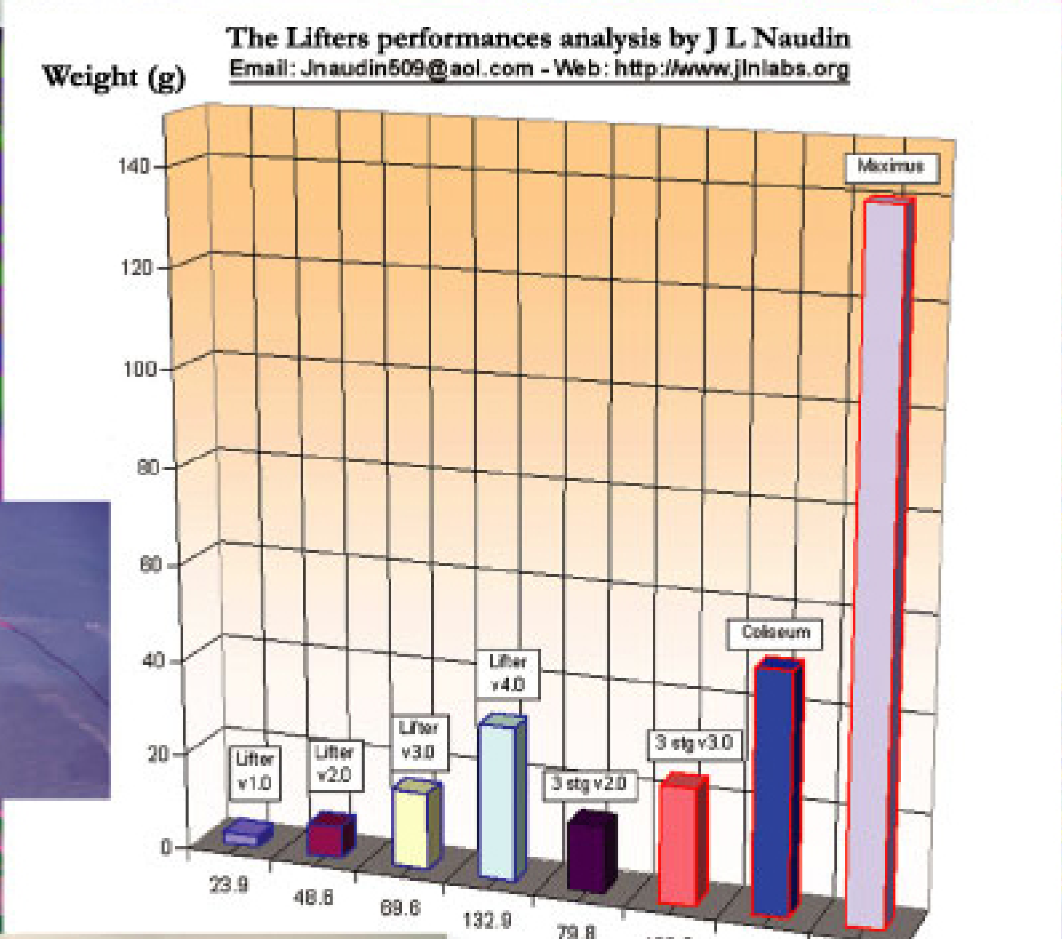

Mouse aboard!

The 1st Worldwide
Electronaut Lifter Project

ISSN 1684-7288

New Energy Technologies

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

Published by Faraday Laboratories Ltd. *Issue #1 (10), 2003*

Research on advanced aerospace systems and new energy sources

Single-Wire Electric Power Transmission
20 kWt!

News from France
Electrostatic propulsion: 60g payload

Also in this issue:

- Cold Fusion by Plasma Electrolysis of Water
- News of Hydrogen Energetics
- Unusual Permanent Magnet Motors
- New Energy and Propulsion Systems

New Energy Technologies magazine
Back issues 2001-2002

Julia V. Sukhova
Translator

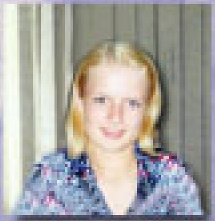
Kate M. Pomerantseva
Designer

Olga O. Leontyeva
Editor

New Energy Technologies
History of invention: Poliakov's vortex drive
July-August 2001
Antigravitation system by V.Roshchin
Report on Russian version of J.Searl's experiments: 7 KW output 35% weight loss
Time Machine
Time Control experiments
Also in this issue:
September-October 2001
Acad. Oleg Gritskevich's Hydro Magnetic Dynamo: 1500 KWt power without any fuel!
Also in this issue:
New energy source created in Germany
MEG 500%
Happy New Year!
2003
Ball lightning

Kirill P. Butusov
Scientific Advisor

Alexander V. Frolov
General Director
"Faraday Laboratory" Ltd

Searl Effect Technology

http://www.sisrc.com

Experimental craft construction (self-rotating permanent magnet system)

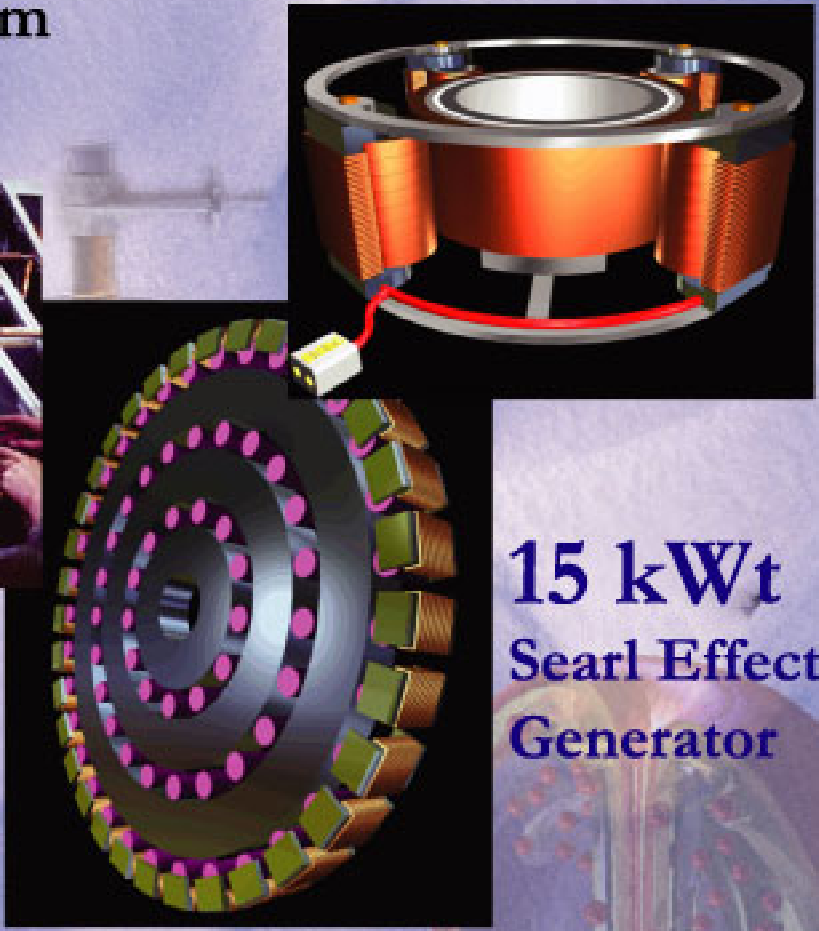

15 kWt

Searl Effect Generator

LENR Experiments

(Low Energy Nuclear Reactions) "Cold Fusion"

http://www.lenr-canr.org

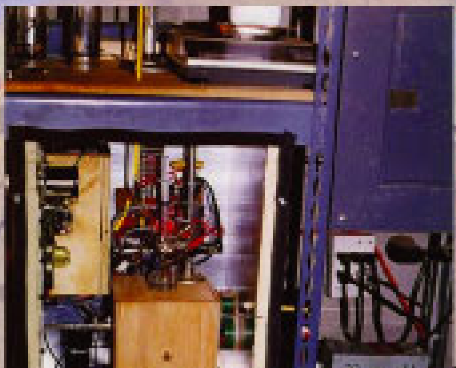

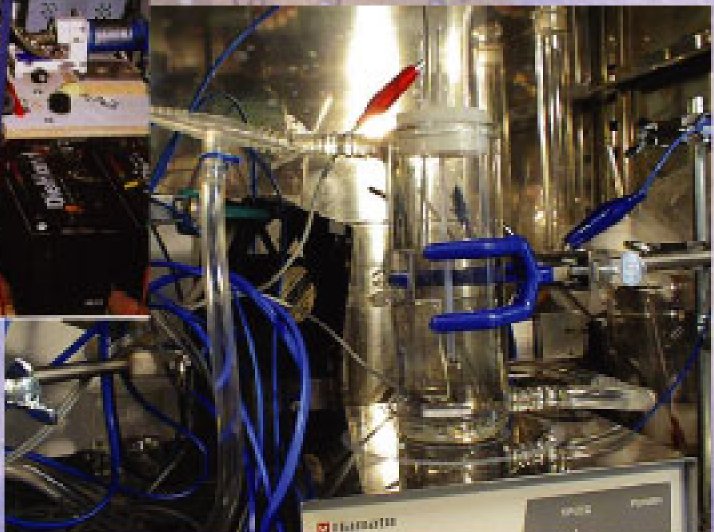

Read in this issue!

NEW ENERGY TECHNOLOGIES #10

1. Testatika. Review of well-known electrostatic system.
2. Cold Fusion by Plasma Electrolysis of Water. Ph.M. Kanarev, T. Mizuno
3. Cold Fusion: What is it and what does it mean to science and society? E. Storms
4. Technical Introduction to LENR-CANR (Low Energy Nuclear Reactions). E. Storms
5. It started in 1989… (History of Cold Fusion Development). P. Hagelstein
6. A Science Tutorial (Cold fusion). T. Chubb
7. LENR (Low Energy Nuclear Reactions) Experiments.
8. On Electrogravitation. A.V. Lemeshko
9. Avalanchedrive Propulsion Device. M.L. Filho
10. Hydrogen-Powered Vehicles. P. Behr, G. Schneider
11. Unusual Permanent Magnet Motors. T. Valone
12. Hydrogen Energetics. A. Pashova
13. On the Longitudinal Electromagnetic Waves. A.V. Frolov
14. Space and Terrestrial Transportation and Energy Technologies. T.C. Loder
15. Commercialising the "Searl Effect".
16. Interesting Publications of the Last Year.
17. "Lifter" Project
18. New Electric Fire Technology. V.D. Dudyshev
19. New Effect of "Cold" Evaporation. V.D. Dudyshev
20. Conceptual Hurdles to New Millennium Physics. D. Reed
21. Resonant Phenomena Occurring in Alternating Current Circuit. A.V. Frolov
22. Books Review

TESTATIKA

Review of well-known electrostatic system

Editor comments by Alexander V. Frolov
General Director, Faraday Lab Ltd

A running "free energy" machine coming from Switzerland, Europe. It was developed over a 20 years Research period by a religious group called Methernitha. This group lives in Linden, Switzerland. The head engineer of this superb machine, Mr. Paul Baumann discovered its principles while observing nature. He claims its running principle was found by studying the lightning effects from nature. The documents being offered here shed light on this energy marvel.

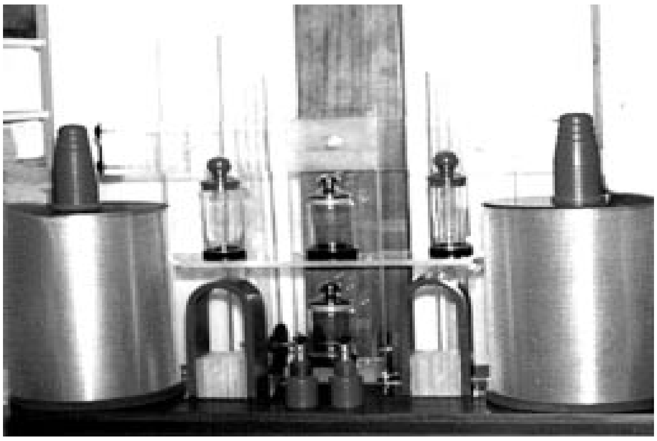

These documents are the result of work of people who have witnessed this wonder machine over the years... And here are the facts:

- The Testatika is a rotary-type machine that runs on it's own, once started by revolving it's two discs by hand!
- The machine not only runs on it's own energy, but produces also a huge amount of excess power, at least 3 KWtt of power! This is almost enough to supply a small house with one machine!
- The machine is about 70cm wide, by 40cm deep and is about 60cm in height!
- It delivers from 250 volts to about 320 volts direct current, depending on the humidity in the air! At this voltage it can supply at least 10 Amperes of pulsating direct current!
- Testatika is not a perpetual motion device, but an energy machine that collects it's huge amount of energy from the ions contained in air. However, there are some technological secrets implemented to overcome the normal drag-resistance of conventional generators. This is but one of it's secrets held by the Methernitha group.

There are many different sized Testatika machines, some of the smaller units deliver only about 200 to 300 watts, none of which are mass-produced. These devices are still laboratory prototype units.

Methernitha is a spiritual community (Christian Alliance) http://www.methernitha.com. They wrote: "We are fortunate in gaining the experience that paradoxically the most beautiful and useful results can be achieved by just using the most simple means. Never did we use any borrowed capital because we want to stay free Swiss citizens and do not want to be hindered or even bound in any way in the pursuance of our aims."

The two contrary-rotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid electrodes the charges are bound. After that they are collected by non-contacting so-called antenna keys and then sorted.

After being initially turned on by hand, the discs rotate by themselves according to the electrostatic laws about attraction and repulsion. A rectifying diode keeps the cycles in steady state. Otherwise the impulses of attraction and repulsion would accumulate and cause the discs to run faster and faster. The correct speed is of great importance and for optimal power generation the discs have to run quite steady and slow.

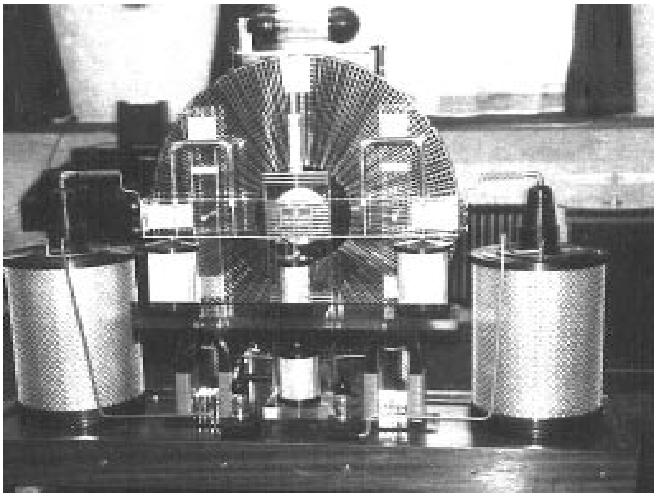

By means of grid condensers the energy is stored and then uniformly discharged, at the same time reducing the high voltage and building up power with additional devices. Finally the machine supplies a uniform direct current, which varies according to the size of the model. The machine furnishes about 3-4 kWt permanent output, depending on humidity, whereby the electric potential ranges from 270 to 320 Volt. High humidity of the atmosphere prevents the build-up of electric potential. The drier the air is, the better.

(Editor: It is necessary to note that the analogues method was used by Russian inventor Pavel N. Yablochkov who in 1877 take out a patent #120684 "The system of distribution and amplification of electrical currents by means of atmosphere electricity..." As it was shown by Yablochkov devices, power doubled in lamps, though current of consumption did not increse.)

No doubt, through the so far achieved results one main objective has been reached, namely to prove that it is possible to use Free Energy. Nevertheless the research work is not yet completed.

To the educated physicist many things of this machine may seem impossible, maybe even crazy. Maybe he is also offended by the conceptions used to explain the whole. Only partly we could use the concepts of conventional physical terminology to explain and define only approximately the functions and properties of the various parts of the machine.

After all it will be necessary to create some more new concepts like the one we have already used before, when we termed the non-contacting collectors of electric charges as antenna keys.

This machine puts experts, which are just trained in conventional physics to a very hard test, because its mode of action is not explainable with the state of the art of officially accepted physical knowledge, or at the most only partially explainable. However also a trained specialist should remain free and independent in his thinking, and should avoid to be limited by the temporal framework of publicly admitted knowledge in any science.

It has to be noted that the established science was already many times forced to change or give up some of its very fundamental concepts. Think about Galilee, to name only one example. Our human society almost condemned this man as a sorcerer and magician, just because he investigated and discovered a truth that seemed unacceptable by the established science of the days.

The book knowledge of any times is not wrong, but it is incomplete, and therefore allows to draw wrong conclusions.

There was used information from the official web site: http://www.methernitha.com

Official adress of Methernitha: Methernitha Genossenschaft Administration, Moosbuehlweg 2, 3517, Linden, SWITZERLAND

Official E-mail of Methernitha: info@methernitha.com

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Color - $800

Half Page Ad
B/W - $200
Color - $400

Quarter Page Ad
B/W -$100
Color- $200

Ask about discounts for multiple ads

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

COLD FUSION BY PLASMA ELECTROLYSIS OF WATER

Ph.M. Kanarev
The Kuban State Agrarian University,
Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Tadahiko Mizuno
Faculty of Engineering, Hokkaido University,
Kita-ku, North 13, West-8 Sapporo 060-8628, Japan

Abstract: It has been disclosed that transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material takes place during plasma electrolysis of water.

Key words: atom, nucleus, proton, neutron, electron, cathode, low-current.

INTRODUCTION

Cold nuclear fusion is the first hypothesis of a source of additional energy in heavy water electrolysis. Fleischmann and Pons, the American electrochemists, are the authors of this hypothesis [1]. They reported about it in 1989. Since that time a large number of experiments has been carried out in order to obtain additional energy from water [2], [3], [4], [5], [7], [8], [9], [10], [11], [12]. We continue to discuss this problem.

THE FIRST EXPERIMENTAL PART

In order to check this hypothesis, the following experiments were performed. Two cathodes were made of iron with mass of 18.10 g and 18.15 g. The first cathode operated during 10 hours in KOH solution; the second cathode operated during the same period in NaOH solution. Mass of the first cathode remained unchanged; mass of the second one was reduced by 0.02 g. The voltage by plasmaelectrolysis process was 220 V and the current (0.5-1.0) A (Fig.1). The indices of the consumption of the solution and the gases being generated were as follows (Table 1).

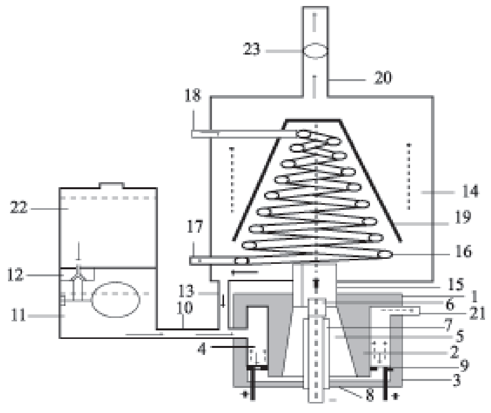

Fig. 1.
Diagram of gas generator. Patent # 2175027:
1 - lid of the reactor; 3 - body of the reactor; 6 - the cathode; 9 - the anode; 11 - solution dosing unit; 16 - cooler; 20 - pipe for gas release; 23 – anemometer

Table 1
Experimental results

Indices	Water consumption, kg	Volume of gases,m^3	Energy expenses, kWh/m^3
KOH	0.272	8.75	0.28
NaOH	0.445	12.66	0.21

In order to increase safety of experimental results, the volume of the gases introduced with the help of anemometer is reduced twofold.

It is known that from one litre of water it is possible to produce 1220 litres of hydrogen and 622 litres of oxygen. Quantity of the gases generated by the plasma electrolytic process is much greater than it is possible to get from consumed water (Table 1) [6]. It gives the reason to think that not only water molecules, but also the nuclei of alkaline metals and the atomic nuclei of the cathode material serve as a source of these gases. The analysing experiment has been performed in order to check this fact.

Tadahiko Mizuno, the famous Japanese scientist (the co-author of this article), who works at the Division of Quantum Energy Engineering Research group of Nuclear System Engineering, laboratory of Nuclear Material System, Faculty of Engineering, Hokkaido University, Japan, kindly agreed to perform chemical analysis of the cathode samples with the help of the nuclear spectroscopy method (EDX). Here are the results of his analysis. The content of chemical elements on the surface of non-operating cathode is as follows (Table 2).

Table 2
Chemical composition of the cathode surface prior its operation in the solution

Element	Fe
%	99.90

The new chemical elements have appeared on the working surface of the cathode, which works in KOH solution (Table 3).

Table 3
Chemical composition of the surface of the cathode, which operates in KOH solution

Element	Si	K	Cr	Fe	Cu
%	0.94	4.50	1.90	93.00	0.45

The chemical composition of the surface of the cathode, which operates in NaOH has proved to be different (Table 4).

Table 4
Chemical composition of the surface of the cathode, which operates in NaOH solution

Element	Al	Si	Cl	K	Ca	Cr	Fe	Cu
%	1.10	0.55	0.20	0.60	0.40	1.60	94.00	0.65

Thus, the hypothesis concerning the participation of the nuclei of alkaline metals and the atomic nuclei of the cathode material in the formation of gases during plasma electrolysis of water has experimental confirmation. Let us carry out the preliminary analysis of the data being obtained (Tables 2, 3, 4).

THE FIRST THEORETICAL PART

In any of these cases, the atoms and the molecules of hydrogen are formed. The part of it is burned and the other goes out with the steam. We have already shown that the processes of fusion of the atoms and the molecules of hydrogen and its isotopes result in occurrence of additional thermal energy [6]. Numerous experiments show that up to 50% of additional thermal energy are generated during the plasma electrolysis of water, it is less than the results of the calculations originating from the existing cold fusion theories [6]. That's why it is necessary to analyse energetics of the particle creation process during the atomic nucleus transmutation.

Having considered the model of the electron we have found out that it can exist in a free state only when it has a definite electromagnetic mass [6]. Being combined with the atomic nucleus it emits a part of energy in the form of the photons, and its electromagnetic mass is reduced. But stability of its condition does not become worse, because the energy carried away by the photons is compensated by binding energy of the electron in the atomic nucleus [6].

If the ambient temperature is increased, the electron begins to absorb the thermal photons and to pass to higher energy levels of the atom reducing binding with it. When the electron becomes free, it interacts with the atom only if the ambient temperature is reduced. As this temperature is reduced, it will emit the photons and sink to lower energy levels [6].

If the electron is in a free state due to an accidental external influence on the atom and the environment has no photons, which are necessary for it to restore its mass, **it begins to absorb the ether from the environment and to restore its constants in such a way**: mass, charge, magnetic moment, spin and radius of rotation. The electron acquires the stable free state only after it has restored its all constants [6].

Thus, if an interchange of the free state and binding state with the atom takes place due to the accidental influences on the atom, **the electron restores its electromagnetic mass every time due to absorbing the ether. It means that actually it plays the role of a converter of the ether energy into the thermal photon energy.**

The Japanese investigators Ohmori and Mizuno [4] registered neutron radiation during plasma electrolysis of water and reported that not only the nuclear process, but also the process of the electron capture by the free protons can be the source of this radiation.

As hydrogen plasma is generated during the plasma electrolytic process of water electrolysis, there exists a tendency of the capture of the free electrons by them.

It is known that rest mass of the electron is $m_e = 9.109534 \times 10^{-31}\ kg$, rest mass of the proton is $m_p = 1.6726485 \times 10^{-27}\ kg$, and rest mass of the neutron is $m_n = 1.6749543 \times 10^{-27}\ kg$. The difference between the mass of the neutron and the mass of the proton is equal to $\Delta m_{np} = 23.058 \times 10^{-31}\ kg$. It is $23.058 \times 10^{-31} / 9.109 \times 10^{-31} = =2.531$ of the mass of the electron. Thus, the proton should capture 2.531 electrons in order to become the neutron. The question arises at once: what will happen to the remained of electron mass $(3.0-2.531)m_e = 0.469m_e$? The disturbed balance of masses in this process is explained by modern physics in a simple way: a neutrino is created [6].

As the neutrino has no charge, it is very difficult to register it. If the neutrino takes the excess mass away or replenish the lacking one, can the elementary particles execute this process by themselves?

As the photons are emitted and absorbed only by the electrons, the proton, which absorbs the electrons, cannot convert the remainder of mass of the third electron into the photon. If the electron is absorbed by the third one and gives more than a half of its mass to the proton in order to convert it into the neutron, the remaining part of mass ($0.469m_e$) of the electron, which has no possibility to become the photon, is converted into a portion of the ether, which "is dissolved" and mixed with the ether in the space. The fact that plasma has no photons with the mass corresponding to the part of mass of the third electron, which has not been absorbed by the proton during its conversion into the neutron, can serve as a proof of such affirmation. Let us calculate energy of such photon [6].

The difference the mass of the neutron and the proton is equal to $\Delta m_{np} = 23.058 \times 10^{-31}\ kg$. If we subtract this value from the mass of three electrons, we'll get mass m_F, from which the photon should be formed [6]

$$m_F = 3m_e - \Delta m_{np} = 3 \times 9.109534 \times 10^{-31} - 23.05810^{-31} = 4.270602 \times 10^{-31}\ \text{kg} \qquad (1)$$

If the photon is formed from this remainder of mass m_F, its energy will be [6]:

$$E_{ph} = m_F \times C^2 = \frac{4.270602 \times 10^{-31} \times (2.997924 \times 10^8)^2}{1.602189 \times 10^{-19}} = 23.956126 \times 10^4\, eV \qquad (2)$$

This value of energy corresponds to roentgen spectrum, that's why the creation of each free neutron should be accompanied by the creation of one roentgen photon. If it does not take place, we have two opportunities: the first one – we should think that in the case when the neutron is created, the neutrino was formed from mass $m_F=4.270602 \times 10_{-31}$ *kg* and flew away in the unknown direction; the second one – there were no conditions for the formation of the photons in the process being considered, and mass, which failed to be formed as a particle, "was dissolved" in the ether. Which variant is closer to the truth [6]? There is no exact answer, but it is known that the Japanese scientists registered only neutron radiation with intensity of 50,000 neutrons per second, and they failed to register roentgen radiation [4].

If in this process the roentgen photons were created, they would not exceed heat efficacy of the plasma electrolytic process, because they would not be the thermal photons. The thermal photons are radiated and absorbed when the electrons make the energy transitions to the energy levels, which are the most remote from the atomic nuclei, where the infrared photons and neighbouring ones from the optical range of the spectrum with energies of ≈(0.001-3.3) eV are generated (Table 5) [6].

Table 5
Electromagnetic spectrum bands

Bands	Wave-length, m	Energy, eV
1. Low- frequency band	$\lambda \approx (10^{7}...10^{4})$	$E \approx 10^{-15}...10^{-11}$
2. Broadcast band	$\lambda \approx (10^{4}...10^{-1})$	$E \approx 10^{-11}...10^{-6}$
3. Microwave band	$\lambda \approx (10^{-1}...10^{-4})$	$E \approx 10^{-6}...10^{-3}$
4. Relic band (maximum)	$\lambda \approx 1 \times 10^{-3}$	$E \approx 1.2 \times 10^{-3}$
5. Infrared band	$\lambda \approx (10^{-4}...7.7 \times 10^{-7})$	$E \approx 10^{-3}...1.6 \times 10^{-2}$
6. Light band	$\lambda \approx (7.7 \times 10^{-7}...3.8 \times 10^{-7})$	$E \approx 1.6 \times 10^{-2}...3.27$
7. Ultraviolet band	$\lambda \approx (3.8 \times 10^{-7}...10^{-9})$	$E \approx 3.27...1 \times 10^{2}$
8. Roentgen band	$\lambda \approx (10^{-9}...10^{-12})$	$E \approx 10^{2}...10^{5}$
9. Gamma band	$\lambda \approx (10^{-12}...10^{-18})$	$E \approx 10^{5}...10^{9}$

Thus, the neutron fusion processes in plasma electrolysis of water will not generate additional thermal energy. But the appearance of the neutrons in plasma will promote the formation of the nuclei of deuterium and, possibly, of tritium. As the balance of masses remains almost unchanged, we have no reason to expect that **additional energy will take place** when deuterium and tritium are formed. But it is sure to appear during **fusion of the atoms of deuterium and tritium, i.e. the hydrogen atoms [6].**

In order to become a proton, the neutron should radiate something, which mass is $\Delta m_{np}=23.058 \times 10^{-31}$ *kg*. Let us convert this mass into energy [6].

$$E_{ph}= \Delta m_{np} \times C^2 = \frac{23.058 \times 10^{-31} \times (2.998 \times 10^{8})^2}{1.602 \times 10^{-19}} = 1.294 \times 10^{6}\, eV \qquad (3)$$

This energy corresponds to the gamma range photons, i.e. not to the thermal photons, and this process does not give additional energy. Thus, if the process of the formation of the helium atoms takes place during plasma electrolysis of water, it should be accompanied by gamma radiation. If there is no such radiation, but the helium atoms are formed, the neutrino takes away the above-mentioned portion of mass Δm_{np} or this mass, which has no opportunity to be formed as the photon, "is dissolved" in the environment, i.e. it is transferred into the state of the ether [6]. As the roentgen photons and the gamma photons are not the thermal ones, this process gives no excessive thermal energy [6].

Another variant is possible. When the atoms of alkali metal bombard the cathode atoms, they are destroyed completely and destroy the atoms of the cathode materials. Under the notion "completely" we'll understand such state when both the atom and the nucleus are destroyed. In this case, the protons of the destroyed nuclei begin to form the hydrogen atoms. The process of fusion of the atoms and the molecules of hydrogen generate additional thermal energy [6]. But one should bear in mind that if plasma disintegrates water molecule into hydrogen and oxygen and if these gases contact plasma, hydrogen is combined with oxygen, and water is formed. Noise generated by plasma is hydrogen microexplosions. Taking into consideration the above-mentioned fact the larger the volume of hydrogen burnt in plasma, the smaller its volume in the gas-vapour mixture. It means that such reactor operation modes are required when quantity of burnt hydrogen is minimal one. Our theory allows us to have such results.

As iron is the cathode material, the nuclei of its atoms are the targets of the atomic nuclei of potassium, alkaline metal. During the transmutation of the iron nuclei (Fig. 2 b), the atomic nuclei of chromium (Fig. 2 a) and the atomic nuclei of copper (Fig. 2 c) are formed [6].

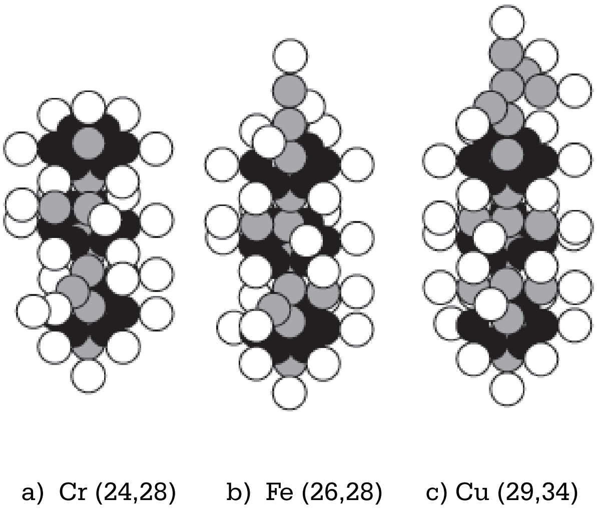

a) Cr (24,28) b) Fe (26,28) c) Cu (29,34)

Fig. 2.
Diagrams of the atomic nuclei of: a) chromium, b) iron, c) copper

When the atomic nucleus of iron (Fig. 2 b) pass into the atomic nucleus of chromium (Fig. 2 a), two protons and two neutrons are released; two atoms of deuterium or one atom of helium can be formed from them. If the neutrons pass into the protons, four atoms of hydrogen are formed.

It is easy to see (Fig. 2) that the atomic nucleus of iron (Fig. 2 b) should lose two upper protons and two neutrons in order to pass into the atomic nucleus of chromium (Fig. 2 a).

Three additional protons and six neutrons (total 9 nucleons) are required for the formation of the atomic nucleus of copper (Fig. 2 c) from the atomic nucleus of iron. As on the cathode surface (Table 3) the number of chromium atoms, which probably are formed from the atomic nuclei of iron, four times more than the number of atoms of copper, then the solution is sure to have superfluous protons and neutrons of the destroyed atomic nuclei of iron, and we can determined their approximate relative quantity.

Let us suppose that four nuclei of the iron atoms pass into the nuclei of the chromium atom. The total quantity of free protons and neutrons (nucleons) is equal to 16. As one atom of copper falls on each four atoms of chromium, 9 nucleons are spent for the formation of one nucleus of the copper atom, and 7 nucleons remain free.

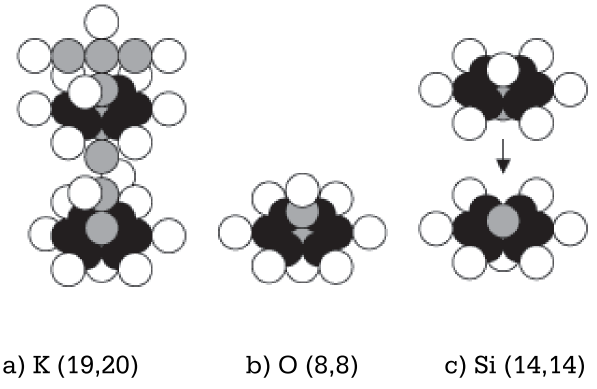

a) K (19,20) b) O (8,8) c) Si (14,14)

Fig. 3.
Diagrams of the atomic nuclei of:
a) potassium, b) oxygen, c) silicon

Let us see what is formed when the nucleus of the potassium atom is destroyed. Potassium is situated in the first group of the fourth period of the periodic law. Its nucleus contains 19 protons and 20 neutrons (Fig. 3 a) [6].

In Fig. 3 a, we can see a weak link of the nucleus of the potassium atom [6]. It is situated in the middle of its axis neutrons. When the transmutation of the nuclei of the potassium atoms takes place, the nuclei of the oxygen atoms can be formed (Fig. 3 b) as well as its isotopes and the nuclei of the silicon atoms (Fig. 3 c).

The analysis of the structure of the nuclei of the potassium atom (Fig. 3 a) shows that it is the most probable source of the nucleus of the silicon atom (Fig. 3 c), which atoms appear on the cathode (Table 3).

It is easy to count that during the destruction of one nucleus of the potassium atom and the creation of one nucleus of the silicon atom 5 free protons and 6 free neutrons, i.e. 11 nucleons, are formed.

Thus, the transmutation of the nuclei of the iron atoms and the potassium atoms results in the formation of free protons and neutrons. As the protons cannot exist in free state, the hydrogen atoms are created from them. If the protons are connected with the neutrons after the destruction of the nuclei of the iron atoms and the potassium atoms, the formation of deuterium, tritium and helium is possible.

Let us pay attention to the main fact – absence of the sodium atoms in the cathode material. It is natural that the potassium atoms have appeared on the cathode, which operated in KOH solution (Table 3). Why are no sodium atoms on the cathode, which operated in NaOH solution? The answer is as follows: the nuclei of the sodium (Fig. 4,a) atoms are completely destroyed during the plasma electrolytic process. The presence of potassium on the surface of the cathode, which operated in NaOH solution (Table 4), can be explained by insufficient ablution of the reactor after the operation with KOH solution.

As free protons and neutrons appear during the destruction of the nucleus of the sodium atom (Fig. 4,a), some nuclei of this element begin to form the atomic nuclei of aluminium (Fig. 4, b), chlorine (Fig. 4, c) and calcium (Fig. 5).

But not all free protons and neutrons are spent for the construction of the atomic nuclei of aluminium, chlorine and calcium. A part of them is spent for the hydrogen atom formation.

If we knew the total quantity of transmutating atomic nuclei of iron, potassium and sodium as well as the exact composition of the gases generated during the plasma electrolytic process, it would be possible to determine the atomic nuclei being formed from additional nucleons. Now we can only suppose that the majority of new nuclei are the protons, i.e. the nuclei of the hydrogen atoms. The increased volume of the gases generated during the plasma electrolytic process is explained by it [6].

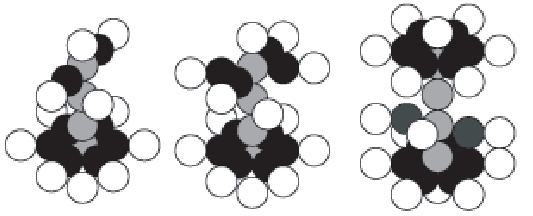

a) Na (11,12) b) Al (13,14) c) Cl (17,18)

Fig. 4.
Diagrams of the atomic nuclei of:
a) sodium, b) aluminium, c) chlorine

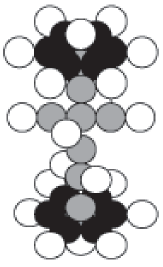

Ca (20,20)

Fig. 5.
Diagram of the nucleus of the calcium atom

The analysis of these Tables shows that transmutation of the nuclei of iron, of which the cathodes are made, results in the formation of chromium and copper in both cases. Apparently, aluminium, chlorine and calcium are formed from the destroyed sodium nuclei. In any case, free protons and neutrons are formed.

But not all free protons and neutrons are spent for the formation of the atomic nuclei of copper, aluminium, chlorine and calcium. A part of them is spent for the formation of the hydrogen atoms. In any case, the atoms and the molecules of hydrogen are formed. The analysis has shown that plasma electrolytic process extracts not more than 0.005 kg of alkaline metal from one litre of the solution. It appears from this that if all neutrons of the atomic nuclei of the molecules of water and alkali metals are transferred into the protons and the atoms and the molecules of hydrogen are formed, the formed volume of gas will be considerably less than the one registered during the experiment (Table 1). A question arises: where do additional gases come from? In order to get the answer on this question we made the next experiment.

THE SECOND EXPERIMENTAL PART

First of all we take into account, that high temperature of plasma forms the conditions when a set of various processes takes place at the cathode. First of all, water is boiled and evaporated. At the same time, one part of water molecules is disintegrated with a release of the atomic hydrogen; another part of the molecules forms the orthohydrogen molecules. A part of water molecules is disintegrated completely and is released at the cathode together with hydrogen and oxygen. A part of hydrogen is combined with oxygen again generating microexplosions (noise) and forming water.

During plasma electrolysis of water, water vapor, hydrogen and oxygen are released simultaneously. If vapor is condensed, gas mixture is released. In order to measure gas flow rate the electronic anemometer have been used. Diameter of the electronic anemometer was equal to internal diameter of the gas make tube (23, Fig. 1). Its readings were registered and processed by the computer. The experiment was performed dozen times, and each time its readings were reproduced with small deviations [11]. But we had no hydrogen analyzer, that's why the results being obtained cannot be considered as final ones. We admonished it in all editions of the book Water is a New Source of Energy with such a phrase: "We abstain from lending an official status to these results with the hope to get necessary financing and to repeat them with a complete set of the necessary devices" [12, page 176].

In the middle of the year of 2002 we received small financing, which allowed us to make a new reactor and to buy some measuring instruments, in particular the scales with the measurement limit up to 600 g and accuracy of 0.02 g. Careful preparation allowed us to increase duration of continuous operation of the reactor (to 10 and more hours) and to register solution consumption for gas production.

The main difficulty of operation with the hydrogen is in the fact that its mixture with air (4-74)% or oxygen (4-94)% is combustible, and the fact was emphasized more than once during the experiments that made the researches be very careful. The second difficulty during hydrogen quantity measurements generated by the plasma electrolytic reactor is in the fact that its molecule has the smallest dimensions, that's why it penetrates easily to the places where the molecules of other substances do not penetrate. Molecular hydrogen diffuses easily even into metals. For example, one volume of palladium absorbs up to 800 volumes of hydrogen.

Gas flow speed was measured with the help of various anemometers, its readings being registered with the help of the computer. Numerous measurements and numerous analysis of gas flow speed measurement accuracy with the help of the anemometers showed that error of a conventional anemometer can be 100%.

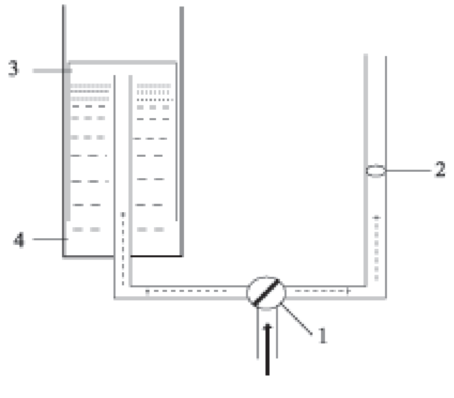

Fig. 6.
Diagram of measurement of flow rate of the gas and its volume:
1 - tap for gas flow movement direction switching,
2 – anemometer, 3 – graduated tank, 4 – water tank

It is known that it is possible to produce 1220 litres of hydrogen and 622 litres of oxygen from one litre of water. Quantity of the gases generated by the plasma electrolytic process is much greater than it is possible to get from consumed water (Table 1). It was a strong reason for a search of the measurement error. For this purpose, the diagram of measurement of flow rate of the gases and their quantity was used (Fig. 6).

The results of the measurements were as follows. The anemometer showed that 200 litres of gas mixture penetrated through it during 10 minutes. Nearly one litre of gases was in the graduated tank during this period. Thus, the measurement of gas flow with the help of the anemometers distorted the result 200 fold. It should be mentioned that the reactor operated in the production mode of hydrogen and oxygen in the cathode zone. As a result, their mixture burst. The pulses of these explosions increased the readings of the anemometer.

It has become necessary to return to the reactor operation modes when no oxygen is released in the cathode zone. Our theory allows us to do this easy.

PROTOCOL
of tests of the first model of low-current Electrolyzers

It is known that it is possible to produce 1.22 l of H_2+ 0.622 O_2= 1.843 (H_2+O_2) from 1 ml of H_2O

Table 6
Experimental results

Indices	1	2	3	Average
1-duration of experiment, hour	1	1	1	1
2-voltage, V	70	70	70	70
3-current, A	0.038	0.080	0.098	0.072
4 – power, W	2.7	5.60	6.44	4.91
4-volume of consumed solution, ml	1.67	3.98	4.32	3.32
5-density of the solution, kg/l	1.04	1.04	1.04	1.04
6-volume of consumed water, ml	1.60	3.83	4.15	3.19
7-volume of the gas mixture being produced, l	2.95	7.06	7.85	5.95
6-volume of hydrogen being produced, l	1.95	4.67	5.07	3.80
7-energy consumption per 1 l of hydrogen, W·h/l	1.38	1.20	1.27	1.28
8-energy consumption per 1m^3 of hydrogen, kWh/m^3	1.38	1.20	1.27	1.28
9-existing energy consumption for production of 1 m^3 of hydrogen from water, kWh/m^3	4.00	4.00	4.00	4.00

CONCLUSION

Transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material during plasma electrolysis of water existed. Plasma electrolytic process opens new prospects in study of matter on the nuclear, atomic and molecular levels. The low-current electrolysis allows us to get the inexpensive hydrogen from water.

REFERENCES

1. M. Fleischmann, S. Pons and M. Hawkins. Electrochemically Induced Nuclear Fusion of Deuterium. J. Electroanal. Chem. 261, 301 (1989),
2. ICCF-7 ACCEPTED ABSTRACTS. Infinite Energy. V. 4, Issue 20, p. 59…69.
3. Harold L. Fox. Cold nuclear fusion: essence, problems, influence on the world. View from USA. Production group SVITAX. M.: 1993, 180 pages.
4. T. Ohmori, T. Mizuno. Strong Excess Energy Evolution, New Element Production, and Electromagnetic Wave and/or Neutron Emission in Light Water Electrolysis with a Tungsten Cathode. Infinite Energy. Issue 20, 1998. Pages 14-17.
5. T. Mizuno. Nuclear Transmutation: The Reality of Cold Fusion. Infinite Energy Press. 1998. 151 pages.
6. Ph. Kanarev. Foundations of Physchemistry of Microworld. Krasnodar, 2002. 330 pages.
7. Kenneth R. Shoulders, “Method of and Apparatus for Production and Manipulations of High Density Charge”, U.S. Patent 5,054,046, issued Oct 1, 1991.
8. Ken Shoulders & Steve Shoulders, “Observations on the Role of Charge Clusters in Nuclear Cluster Reactions”, J.of New Energy, vol. 1, no 3, pp 111-121, Fall 1996, 7 refs, 22 figs.
9. Hal Fox, Robert W. Bass, & Shang-Xian Jin, “Plasma-Injected Transmutation”, J. of New Energy, vol. 1, no 3, Fall 1996, pp 222-230, 23 refs, 4 figs.
10. Shang-Xian Jin & Hal Fox, “High Density Charge Cluster Collective Ion Accelerator”, J. of New Energy, vol. 4, no 2, Fall 1999, pp 96-104, 47 refs, 4 figs., 3 tables.
11. Ph.M. Kanarev. Water is the Main Power Carrier of Future Power Engineering. Journal of New Energy. An International Journal of New Energy Systems. Vol. 6, No.2. Pag. 101-121.
12. Ph.M. Kanarev. Water is New Source of Energy. The Third Edition. Krasnodar 2002. 194p. (In English).

LENR
(Low Energy Nuclear Reactions)

http://www.lenr-canr.org

Editorial: There is a collection of some papers on LENR (Low Energy Nuclear Reactions), also known as Cold Fusion. CANR, Chemically Assisted Nuclear Reactions, is another term for this phenomenon. These original scientific papers are reprinted with permission from the authors and publishers.

Website http://www.lenr-canr.org features a growing library of scientific papers about LENR and an extensive bibliography of journal papers, news articles and books about LENR.

COLD FUSION: What is it and what does it mean to science and society?

Edmund Storms

Cold fusion is important because it promises to be a new source of pollution-free, inexhaustible energy. In addition, it is important because it reveals the existence of a new way nuclei can interact that conventional scientific theory predicts is impossible. What then is this phenomenon that suffers such promise and rejection?

Energy can be obtained from the nucleus in two different ways. On the one hand, a large nucleus can be broken into smaller pieces, such as is experienced by uranium in a conventional nuclear reactor and by the material in an atom bomb. This is called fission. On the other hand, two very small nuclei can be joined together, such as occurs during fusion of deuterium and tritium in a Hot Fusion reactor and in a hydrogen bomb. This process, called fusion, also takes place in stars to produce much of the light we see.

The fission reaction is caused to happen by adding neutrons to the nucleus of uranium or plutonium to make it unstable. The unstable nucleus splits into two nearly equal pieces, thereby releasing more neutrons, which continue the process. As every one now knows, this process produces considerable waste that is highly radioactive. The uranium used as fuel also occurs in limited amounts in the earth's crust. As a result, this source of energy is not ideal, although widely used at the present time.

The normal hot fusion reaction requires two deuterium or tritium nuclei to be smashed together with great energy. This is accomplished by raising their temperature. However, this temperature is so high that the reactants cannot be held in a solid container, but must be retained by a magnetic field. This process has proven to be very difficult to accomplish for a time sufficient to generate useable energy. In spite of this difficulty, attempts have been under way for the last 40 years and with the expenditure of many billions of dollars. Success continues to be elusive while the effort continues.

Cold fusion, on the other hand, attempts to cause the same process, but by using solid materials as the container held at normal temperatures. The container consists of various metals, including palladium, with which the deuterium is reacted to form a chemical compound. While in this environment, the barrier between the deuterium nuclei is reduced so that two nuclei can fuse without having to be forced together. Because the process causing this to happen is not well understood, the possibility is rejected by many conventional scientists. Difficulty in producing the process on command has intensified the rejection.While this difficulty is real, it has not, as many skeptics have claimed, prevented the process from being reproduced hundreds of times in laboratories all over the world for the past 13 years. As you will see by reading the reviews and papers in our Library (see http://www.lenr-canr.org/LibFrame1.html), the process continues to be reproduced with increasing ease using a variety of methods and materials.

What is the nature of this process and why has it been so hard to understand? To answer this question, a person needs to understand the nature of the barrier that exists between all nuclei. Because all nuclei have a positive charge in proportion to their atomic number, all nuclei repeal each other. It is only the surrounding electrons that hold normal matter together, with the nuclei being at considerable distance from each other, at least on the scale of an atom. When attempts are made to push the nuclei closer, the required energy increases as the nuclei approach one another. However, when deuterium dissolves in a metal, it experiences several unique conditions. The surrounding metal atoms produce a regular array that is able to support waves of various kinds. These waves can be based on vibration of the atoms (phonons), vibration of the electrons, standing waves of electromagnetic energy, or a wave resulting from conversion of the deuterium nuclei to a wave. In addition, the high density of electrons can neutralize some of the positive charge on the deuterium nuclei allowing a process called tunneling, i.e.. allowing passage through the barrier rather than over it. The mechanism of this neutralization process is proposed to involve a novel coherent wave structure that can occur between electrons under certain conditions. All of these wave processes have been observed in the past under various conventional conditions, but applying them to the cold fusion phenomenon has been a subject of debate and general rejection.

While the debate based on wave action has been underway, people have proposed other mechanisms. These include the presence of neutrons within the lattice. Normally, neutrons are unstable outside of the nucleus, decomposing into a proton, an electron, and a neutrino. Presumably, this reaction can be reversed so that neutrons might be created in a lattice containing many free electrons and protons. Having no charge, the neutron could then interact with various atoms in the lattice to produce energy. These neutrons might also be hidden in the lattice by being attached to other nuclei in a stabilized form, to be released when conditions were right. Several particles normally not detected in nature also have been proposed to trigger fusion and other nuclear reactions.

While search for a suitable mechanism has been underway, an understanding of the environment that triggers the mechanism has been sought, the so-called nuclear-active-environment. Initially, this environment was thought to exist in the bulk of the palladium cathode used in the Pons-Fleischmann method to produce cold fusion. It is now agreed that the nuclear reactions only occur in the surface region. Recent arguments suggest that this surface layer does not even require palladium for it to be nuclear-active. Nuclear reactions have now been produced in a variety of materials using many methods. The only common feature found in all of these methods is the presence of nano-sized particles of material on the active surface. If this observation is correct, four conditions seem required to produce the nuclear reactions. First, the particle must have a critical small size; second, it must contain a critical concentration of deuterium or hydrogen; third, it must be constructed of certain atoms; and fourth, it must be exposed to a source of energy. This energy can take the form of a sufficiently high temperature, a significant high flux of hydrogen through the particle, application of energetic electrons or charged particles, or application of laser light of the proper frequency. Until, the importance of these factors is understood, the effect will continue to be difficult to replicate.

Technical Introduction to LENR-CANR

Edmund Storms

At low energies, the Coulomb barrier prevents nuclei from coming together and fusing to form a single nucleus. To initiate a nuclear reaction, several methods are used. Nuclear reactions are normally initiated by pushing two atoms together with enough force to overcome the Coulomb barrier by brute force, or by using neutrons which penetrate the nuclei without seeing a barrier. (Neutrons have no electrical charge, so the Coulomb barrier does not stop them.) These forces are normally provided by high-temperature plasma or by accelerating ions to high energies. In contrast, LENR describes the mechanism and conditions that cause a variety of nuclear reactions to take place with relatively low activation energy. These unique conditions reduce the need for excessive energy. The normal method forces the nuclei together, while the new method encourages them to come together. The challenge has been to understand the unique characteristics of the necessary solid structure such that this structure could be generated at will.

Because the proposed method is unique, at odds with current nuclear theory, and is still difficult to reproduce, support for studies in many countries, but not all, has been very limited. Nevertheless, considerable information has accumulated over the last 13 years since Profs. Stanley Pons and Martin Fleischmann showed the world the possibilities inherent in this phenomenon. Much understanding is buried in conference proceedings and reports that are not available to a serious student. This information will, as time permits, be made available on this site. Students of the subject are also encouraged to use this site to interact with other people in the field and provide objective critiques of the work published here.

PHENOMENA DISCUSSED IN SOME OF THE PAPERS

At least 10 ways have been demonstrated to produce anomalous heat and/or anomalous elemental synthesis. A few of these methods will be described here. For course, not all of the claims are worthy of belief nor are they accepted by many people. Nevertheless, the claims will be described without qualifications in order to provide the reader with the latest understanding.

conditions. The surrounding metal atoms produce a regular array that is able to support waves of various kinds. These waves can be based on vibration of the atoms (phonons), vibration of the electrons, standing waves of electromagnetic energy, or a wave resulting from conversion of the deuterium nuclei to a wave. In addition, the high density of electrons can neutralize some of the positive charge on the deuterium nuclei allowing a process called tunneling, i.e.. allowing passage through the barrier rather than over it. The mechanism of this neutralization process is proposed to involve a novel coherent wave structure that can occur between electrons under certain conditions. All of these wave processes have been observed in the past under various conventional conditions, but applying them to the cold fusion phenomenon has been a subject of debate and general rejection.

While the debate based on wave action has been underway, people have proposed other mechanisms. These include the presence of neutrons within the lattice. Normally, neutrons are unstable outside of the nucleus, decomposing into a proton, an electron, and a neutrino. Presumably, this reaction can be reversed so that neutrons might be created in a lattice containing many free electrons and protons. Having no charge, the neutron could then interact with various atoms in the lattice to produce energy. These neutrons might also be hidden in the lattice by being attached to other nuclei in a stabilized form, to be released when conditions were right. Several particles normally not detected in nature also have been proposed to trigger fusion and other nuclear reactions.

While search for a suitable mechanism has been underway, an understanding of the environment that triggers the mechanism has been sought, the so-called nuclear-active-environment. Initially, this environment was thought to exist in the bulk of the palladium cathode used in the Pons-Fleischmann method to produce cold fusion. It is now agreed that the nuclear reactions only occur in the surface region. Recent arguments suggest that this surface layer does not even require palladium for it to be nuclear-active. Nuclear reactions have now been produced in a variety of materials using many methods. The only common feature found in all of these methods is the presence of nano-sized particles of material on the active surface. If this observation is correct, four conditions seem required to produce the nuclear reactions. First, the particle must have a critical small size; second, it must contain a critical concentration of deuterium or hydrogen; third, it must be constructed of certain atoms; and fourth, it must be exposed to a source of energy. This energy can take the form of a sufficiently high temperature, a significant high flux of hydrogen through the particle, application of energetic electrons or charged particles, or application of laser light of the proper frequency. Until, the importance of these factors is understood, the effect will continue to be difficult to replicate.

Technical Introduction to LENR-CANR

Edmund Storms

At low energies, the Coulomb barrier prevents nuclei from coming together and fusing to form a single nucleus. To initiate a nuclear reaction, several methods are used. Nuclear reactions are normally initiated by pushing two atoms together with enough force to overcome the Coulomb barrier by brute force, or by using neutrons which penetrate the nuclei without seeing a barrier. (Neutrons have no electrical charge, so the Coulomb barrier does not stop them.) These forces are normally provided by high-temperature plasma or by accelerating ions to high energies. In contrast, LENR describes the mechanism and conditions that cause a variety of nuclear reactions to take place with relatively low activation energy. These unique conditions reduce the need for excessive energy. The normal method forces the nuclei together, while the new method encourages them to come together. The challenge has been to understand the unique characteristics of the necessary solid structure such that this structure could be generated at will.

Because the proposed method is unique, at odds with current nuclear theory, and is still difficult to reproduce, support for studies in many countries, but not all, has been very limited. Nevertheless, considerable information has accumulated over the last 13 years since Profs. Stanley Pons and Martin Fleischmann showed the world the possibilities inherent in this phenomenon. Much understanding is buried in conference proceedings and reports that are not available to a serious student. This information will, as time permits, be made available on this site. Students of the subject are also encouraged to use this site to interact with other people in the field and provide objective critiques of the work published here.

PHENOMENA DISCUSSED IN SOME OF THE PAPERS

At least 10 ways have been demonstrated to produce anomalous heat and/or anomalous elemental synthesis. A few of these methods will be described here. For course, not all of the claims are worthy of belief nor are they accepted by many people. Nevertheless, the claims will be described without qualifications in order to provide the reader with the latest understanding.

The most studied method involves the use of an electrolytic cell containing a LiOD electrolyte and a palladium cathode. Current passing through such a cell generates D+ ions at the cathode, with a very high effective pressure. These ions enter the palladium and, if all conditions are correct, join in a fusion reaction that produces He-4. Initially palladium wire and plate were used, but these were found to form microcracks, which allowed the required high concentration of deuterium to escape. Later work shows that the actual nuclear reaction occurs on the surface within a very thin layer of deposited impurities. Therefore, control of this impurity layer is very important, but rather difficult. The use of palladium is also not important because gold and platinum appear to be better metals on which to deposit the impurity layer. This method is found, on rare occasions, to generate tritium within the electrolyte and transmutation products on the cathode surface. Different nuclear reactions are seen when light water (H_2O) is used instead of D_2O, although the amount of anomalous energy is less when H_2O is used. These observations have been duplicated hundreds of times in dozens of laboratories.

Application of deuterium gas to finely divided palladium, and perhaps other metals, has been found to generate anomalous energy along with helium-4. Both palladium-black as well as palladium deposited as nanocrystals on carbon have shown similar anomalous behavior. In both cases the material must be suitably purified. Palladium deposited on carbon can and must be heated to above 200/260°C for the effect to be seen. When deuterium is caused to diffuse through a palladium membrane on which is deposited a thin layer of various compounds, isotopes that were not previously present are generated with isotopic ratios unlike those occurring naturally.

A plasma discharge under H_2O or D_2O between various materials generates many elements that were not previously present. When the electrodes are carbon and the plasma is formed in H_2O, the main anomalous element is iron. This experiment is relatively easy to duplicate.

Several complex oxides, including several superconductors, can dissolve D_2 when heated. When a potential is applied across a sheet of such material, the D+ ions are caused to move and anomalous heat is generated.

If deuterium ions, having a modest energy, are caused to bombard various metals, tritium as well as other elements not previously present are generated. These ions can be generated in a pulsed plasma or as a beam.

When water, either light or heavy, is subjected to intense acoustic waves, collapse of the generated bubbles on the surrounding solid walls can generate nuclear reactions. This process is different from the fusion reaction claimed to occur within a bubble just before it disappears within the liquid because neutrons are not produced in the former case, but are produced in the latter case. This method has been applied to various metals in heavy water using an acoustic transducer and in light water using a rotating vane which generates similar acoustic waves...

HOW TO EXPLAIN THE CLAIMS

A major problem in deciding which model might be correct is the absence of any direct information about the nature of the nuclear-active-environment. At this time, two important features seem to be important, the size of the nanodomain in which the reactions occur and the presence of a deuterium flux through this domain. The domain can apparently be made of any material in which hydrogen or deuterium can dissolve. Until the nature of the nuclear-active-state (NAS) is known, no theory will properly explain the effect and replication of the claims will remain difficult.

When fusion is initiated using conventional methods, significant tritium and neutrons are produced. In addition, when other elements are generated, they tend to be radioactive. This is in direct contrast to the experience using low energy methods. These products are almost completely absent and, instead, helium-4 is produced. When radiation is detected, it has a very low energy. This contrasting behavior, as well as the amount of anomalous energy, has made the claims hard to explain using conventional models. This difficulty has been amplified by a failure of many skeptics to recognize the contrasting effect of the environment, a plasma being used in the older studies and a solid lattice of periodic atoms being present as the new environment.

Over 500 models and their variations have been proposed, some of which are very novel and some are variations on conventional ideas. Most models attempt to explain the nuclear reaction once the required environment has been created, without addressing what that unique environment might be like. These models involve conversion of a proton (deuteron) to a neutron (dineutron), creation of an electron structure that is able to neutralize the barrier, conversion of deuterium to a wave which interacts without charge, and the presence of otherwise overlooked neutrons and/or novel particles. Many of the models will have to be abandoned or seriously modified once the nature of the nuclear active environment is understood.

It Started in 1989 . . .

Peter Hagelstein

Many of us recall the controversy surrounding the announcement of claims of observations of fusion reactions in a test tube that were made in 1989. At the time, these claims were greeted with considerable skepticism on the part of the physics community and the scientific community in general.

The principal claim of Pons and Fleischmann

The principal claim of Pons and Fleischmann in 1989 was that power was produced in palladium cathodes that were loaded electrochemically in a heavy water electrolyte. The evidence in support of this was a measured increase in the temperature in the electrochemical cell. There was no obvious evidence for nuclear reaction products commensurate with the claimed heat production. Fleischmann speculated that perhaps two deuterons were somehow fusing to He-4 through some kind of new mechanism.

Rejection by the physics community

This claim was not accepted by the physics community on theoretical grounds for several reasons:

First, there was no mechanism known by which two deuterons might approach one another close enough to fuse, since the Coulomb barrier prevents them from approaching at room temperature.

Second, if they did approach close enough to fuse, one would expect the conventional dd-fusion reaction products to be observed, since these happen very fast. Essentially, once two deuterons get close enough to touch, reactions occur with near unity probability, and the reaction products (p+t and n+He-3) leave immediately at high relative velocity consistent with the reaction energy released. To account for Fleischmann's claim, the proposed new reaction would seemingly somehow have to make He-4 quietly and cleanly, without any of the conventional reaction products showing up, and would somehow have to arrange for this to happen a billion times faster than the conventional reaction pathway. Most physicists bet against the existence of such a magical new effect.

Third, the normal pathway by which two deuterons fuse to make He-4 normally occurs with the emission of a gamma ray near 24 MeV. There was no evidence for the presence of any such high energy gamma emission from the sample, hence no reason to believe that any helium had been made.

Finally, if one rejects the possibility that any new mechanisms might be operative, then the claim that power was being produced by fusion must be supported by the detection of a commensurate amount of fusion reaction products. Pons and Fleischmann found no significant reaction products, which, given the rejection of new mechanisms, implied an absence of fusion reactions.

An alternate explanation is proposed

The physicists decided in 1989 that the most likely reason that Pons and Fleischmann observed a temperature increase was that they had made an error of some sort in their measurements. When many groups tried to observe the effect and failed, this led most of the physics community to conclude that there was nothing to it whatsoever other than some bad experiments.

The claim of Jones

A second very different claim was made at the same time in 1989 by Steve Jones. This work also involved electrochemistry in heavy water and the observation of reaction products corresponding to the conventional dd-fusion reactions. The initial publication showed a spectrum of neutron emission that Jones had detected from a titanium deuteride cathode loaded electrochemically. The response of the physics community was skeptical, as the signal to noise ratio was not particularly impressive. Given the polarization of the physics community in opposition to the claims of Pons and Fleischmann (which were announced essentially simultaneously), the physicists were not of a mood to accept much of any claims that fusion could happen in an electrochemical experiment at all. Jones went to great lengths to assure fellow scientists that his effect was completely unrelated to the claims of Pons and Fleischmann, and was much more reasonable.

Also rejected

Physicists had reason to be skeptical. Theoretical considerations indicated that the screening effects that Jones was relying on were not expected to be as strong as needed to account for the fusion rates claimed. As this experiment could not seem to be replicated by others at the time, it was easy for the physics community to reject this claim as well.

Cold fusion, weighed and rejected with prejudice

Cold fusion, as the two different claims were termed, was dismissed with prejudice in 1989. The initial claims were made near the end of March in Utah, and the public refutation of the claims was made at the beginning of May. It only took about 40 days for the physics community to consider the new claims, test them experimentally, and then announce loudly to the world that they had been carefully weighed and rejected.

Following this rejection, physicists have treated cold fusion rather badly. For example, Professor John Huizenga of Rochester University was selected to be co-chair of the DOE ERAB committee that met to review cold fusion and issue a report. Shortly afterward, he wrote a book entitled Cold Fusion, The Scientific Fiasco of the Century, in which he discusses the claims, the experiments, and the extreme skepticism with which the new claims were greeted. Robert Park discusses the subject in his book entitled Voodoo Science. You can find many places where physicists and other scientists happily place the cold fusion claims together with claims of psychic phenomena.

A Science Tutorial

Talbot Chubb

First it is important to recognize that there are four distinct types of energy production:
1) chemical energy, that powers our cars and most of our civilization;
2) nuclear fission energy, as used to generate about 15% or our electricity;
3) hot fusion nuclear energy, which powers the sun and most stars;
4) cold fusion nuclear energy, which appears as unexplained heat in a few experimenter's laboratory studies and which most scientists believe is impossible.

The three types of nuclear energy produce 10 million times as much heat per pound of fuel than occurs with chemical energy. How do these types of energy differ? To understand this question you need to know some chemistry and physics.

Lesson 1

Nature has provided us with two types of stable charged particles, the proton and the electron. The proton is heavy, normally tiny, and has a positive charge. The electron is light, normally large and fuzzy, and has a negative charge. The positive charge and the negative charge attract each other, just like the north pole of a magnet attracts the south pole of a magnet. When you bring two magnets together with the north pole of one facing the south pole of the other, they pull together, bang! When they bang into each other they release a little bit of energy in the form of heat, but it is too small an amount to easily measure. To pull the magnets apart you have to do work, which is another way of saying you have to use up energy. It's almost like pulling a rock back up a hill. Rolling the rock down a hill actually creates a little heat, and pulling the rock back up the hill takes energy. In the same way the positive charge of the proton pulls on the negative charge of the electron and they stick together releasing energy in the process. The result is a hydrogen atom, designated H. A hydrogen atom is nothing but a fuzzy electron hugging a compact proton. The proton is the nucleus of the hydrogen atom. If you knock the electron off the hydrogen atom you get a positive ion H+, which is nothing more than the original proton. An ion is the name applied to an atom or molecule that has lost or gained one or more electrons, hence is no longer electrically neutral.

Lesson 2

As you know, nature has provided us with more than one type of atom. We have oxygen atoms, nitrogen atoms, iron atoms, helium atoms, etc.. How do these atoms differ? The answer is that they all have different types of nuclei (plural of nucleus, from the Latin). And these different nuclei all have different numbers of protons inside them, which means they all have different plus charges. The nucleus of the helium atom has 2 protons inside it, hence has plus 2 charge, and requires 2 electrons to neutralize its charge. When 2 electrons stick to it, it becomes a helium atom. The oxygen nucleus has 8 protons and has charge 8. When 8 electrons stick to it, it becomes an oxygen atom. The nitrogen atom has 7 electrons, and the iron atoms something like 26. But all the atoms are built more or less the same way, with a compact positively charged nucleus embedded in a cloud of fuzzy electrons. The difference in size between the compact nucleus and the fuzzy electrons is enormous. The sun has a diameter only about 100 times that of the earth. The electron cloud on an atom has a diameter which is about 100,000 times that of the nucleus. Cube these numbers to get the difference in volumes.

Lesson 3

We now are in a position to understand what chemical energy is. The atoms, all electrically neutral, can actually join with each other and release more energy. This is another way of saying that they can join into more stable configurations. The electrons in an atom try to configure themselves so as to get as close as possible to their nucleus, but their fuzzy nature requires that they take up a certain volume of space. However, if they join together with the electrons of another atom they can usually find a tighter configuration that leaves them closer to their beloved beginning of May. It only took about 40 days for the physics community to consider the new claims, test them experimentally, and then announce loudly to the world that they had been carefully weighed and rejected.

Following this rejection, physicists have treated cold fusion rather badly. For example, Professor John Huizenga of Rochester University was selected to be co-chair of the DOE ERAB committee that met to review cold fusion and issue a report. Shortly afterward, he wrote a book entitled Cold Fusion, The Scientific Fiasco of the Century, in which he discusses the claims, the experiments, and the extreme skepticism with which the new claims were greeted. Robert Park discusses the subject in his book entitled Voodoo Science. You can find many places where physicists and other scientists happily place the cold fusion claims together with claims of psychic phenomena.

A Science Tutorial

Talbot Chubb

First it is important to recognize that there are four distinct types of energy production:
1) chemical energy, that powers our cars and most of our civilization;
2) nuclear fission energy, as used to generate about 15% or our electricity;
3) hot fusion nuclear energy, which powers the sun and most stars;
4) cold fusion nuclear energy, which appears as unexplained heat in a few experimenter's laboratory studies and which most scientists believe is impossible.

The three types of nuclear energy produce 10 million times as much heat per pound of fuel than occurs with chemical energy. How do these types of energy differ? To understand this question you need to know some chemistry and physics.

Lesson 1

Nature has provided us with two types of stable charged particles, the proton and the electron. The proton is heavy, normally tiny, and has a positive charge. The electron is light, normally large and fuzzy, and has a negative charge. The positive charge and the negative charge attract each other, just like the north pole of a magnet attracts the south pole of a magnet. When you bring two magnets together with the north pole of one facing the south pole of the other, they pull together, bang! When they bang into each other they release a little bit of energy in the form of heat, but it is too small an amount to easily measure. To pull the magnets apart you have to do work, which is another way of saying you have to use up energy. It's almost like pulling a rock back up a hill. Rolling the rock down a hill actually creates a little heat, and pulling the rock back up the hill takes energy. In the same way the positive charge of the proton pulls on the negative charge of the electron and they stick together releasing energy in the process. The result is a hydrogen atom, designated H. A hydrogen atom is nothing but a fuzzy electron hugging a compact proton. The proton is the nucleus of the hydrogen atom. If you knock the electron off the hydrogen atom you get a positive ion $H+$, which is nothing more than the original proton. An ion is the name applied to an atom or molecule that has lost or gained one or more electrons, hence is no longer electrically neutral.

Lesson 2

As you know, nature has provided us with more than one type of atom. We have oxygen atoms, nitrogen atoms, iron atoms, helium atoms, etc.. How do these atoms differ? The answer is that they all have different types of nuclei (plural of nucleus, from the Latin). And these different nuclei all have different numbers of protons inside them, which means they all have different plus charges. The nucleus of the helium atom has 2 protons inside it, hence has plus 2 charge, and requires 2 electrons to neutralize its charge. When 2 electrons stick to it, it becomes a helium atom. The oxygen nucleus has 8 protons and has charge 8. When 8 electrons stick to it, it becomes an oxygen atom. The nitrogen atom has 7 electrons, and the iron atoms something like 26. But all the atoms are built more or less the same way, with a compact positively charged nucleus embedded in a cloud of fuzzy electrons. The difference in size between the compact nucleus and the fuzzy electrons is enormous. The sun has a diameter only about 100 times that of the earth. The electron cloud on an atom has a diameter which is about 100,000 times that of the nucleus. Cube these numbers to get the difference in volumes.

Lesson 3

We now are in a position to understand what chemical energy is. The atoms, all electrically neutral, can actually join with each other and release more energy. This is another way of saying that they can join into more stable configurations. The electrons in an atom try to configure themselves so as to get as close as possible to their nucleus, but their fuzzy nature requires that they take up a certain volume of space. However, if they join together with the electrons of another atom they can usually find a tighter configuration that leaves them closer to their beloved nuclei. For example, 2 hydrogen atoms can join together into a more compact configuration if each hydrogen atom contributes its electron to a 2-electron cloud, which the separate protons share. In this manner they form a grouping of the 2 electrons in a single cloud, together with the 2 isolated protons spaced apart from each other but still within the electron cloud. The result is a heat-producing chemical reaction $H + H => H_2$. (The => means "goes to" or "becomes".) The H_2 configuration is the hydrogen molecule, and when you buy a tank of hydrogen gas, H_2 molecules is what you get. Furthermore, the 2 electrons of the H_2 molecule and the 8 electrons of the O atom can find a still more compact configuration by combining their electrons to create the water molecule H_2O, plus heat. The water molecule is really a single cloud of electrons in which are embedded the three point-like nuclei to form a minimum energy configuration. So when we burn oil or coal we reconfigure the electrons to produce more stable configurations of point-like nuclei embedded in electron clouds, liberating heat. So much for chemical energy.

Lesson 4

We have slid over one point in the above discussion. How does Nature make a nucleus containing two or more protons in the first place. After all, each of the protons has a positive charge, and the positive charges repel each other very strongly when they are separated by a tiny distance, equal to the distance across a nucleus. The repulsion of like charges is just like the repulsion between the north poles of two magnets when they are pushed together the wrong way. Something must overcome this repulsion, or else the only kind of atoms we would have would be those of hydrogen. Fortunately, this is not what we observe. The answer is that there is a second kind of force which acts on protons. This is the nuclear force. The nuclear force is very strong but requires particles to almost sit on each other to have any effect. Also, there is a second kind of heavy particle, which is just like a proton, except that it has no positive or negative charge. It is not pushed away by the proton's plus charge. This other kind of particle is called the neutron, since it is electrically neutral. A peculiar fact of life is that it exists in stable form only inside a nucleus. When not in the nucleus it changes into a proton, an electron and a very light anti-neutrino in about 10 minutes. But it lasts forever inside a nucleus. Anyway, the neutron and the proton very strongly attract each other once they get close enough together, and then they combine to form a highly stable pair called a deuteron, which we designate $D+$. The single deuteron, when it combines with a single electron, forms the heavy hydrogen atom called deuterium, designated D. A second nuclear reaction occurs when two deuterons make contact. When they can be forced together so as to make contact, the 2 deuterons fuse, making a doubly charged particle. The grouping of 2 protons and 2 neutrons is even tighter than the proton-neutron grouping in the deuteron. The new particle, when neutralized by 2 electrons, is the nucleus of the helium atom, designated He. Larger groupings of neutrons and protons exist in nature and serve as the nuclei of carbon, nitrogen, oxygen, and iron, etc. atoms. All of these groupings are made possible by the very strong nuclear force, which is felt between particles only when they are in contact or share the same nucleus-size volume of space.

Lesson 5

We can now understand normal nuclear energy, which is really nuclear fission energy. During the early history of the universe massive stars were formed. In the explosion of these massive stars, lots of different types of nuclei were formed and exploded back into space. Second and later generation stars and planets were formed from this mix, including the sun. In the explosion process probably every possible stable configuration of protons and neutrons was produced, plus some almost-stable groupings, such as the nucleus of the uranium atom. There are actually 3 different types of uranium atom nuclei, called uranium-234, uranium-235, and uranium-238. These "isotopes" differ in their number of neutrons, but they all have 92 protons. The nuclei of all uranium atoms can go to a lower energy configuration by ejecting a helium nucleus, but this process occurs so rarely that the Earth's uranium has already lasted over 4 billion years. But the uranium nuclei are unstable in another way. In general, groupings of protons and neutrons are happiest if they have about 60 protons-plus-neutrons. The uranium nuclei contain more than three times this number. So they would like to split in two, which would release a lot of heat. But nature doesn't provide a way for them to split apart. They have to first go to a higher energy configuration before splitting in two. However, one of the three forms of uranium nucleus found in nature called uranium-235 and designated ^{235}U, gains the needed energy if it captures a neutron. The energized nucleus that results from neutron capture then splits apart with the release of an enormous amount of energy, and incidentally with release of additional neutrons. The additional neutrons can then split more uranium-235 nuclei, keeping the reaction going. This is what happens in nuclear power plants, where the heat, which is the end product of the nuclear splitting process, is used to boil water, generate steam, and turn electrical generators. (One also gets lots of radioactive products, which are a nuisance to dispose of safely.)

Lesson 6

We are now also in a position to understand hot fusion nuclear energy. As mentioned in lesson 5, the groupings of protons plus neutrons is most stable when the numbers of neutrons and protons approximate those found in the nucleus of an iron atom. Just as uranium has too many neutrons plus protons to be comfortable, so the light elements like hydrogen, helium, carbon, nitrogen and oxygen have too few. If the nuclei can be made to make contact under proper conditions, they can combine to create more stable groupings, plus heat. This is the process of fusion. Nature has found a way of doing this in stars like the sun. All Nature has to do is heat compressed hydrogen hot enough and wait long enough and hot fusion will occur. If Nature were to start with deuterium, which already has a paired proton and neutron, the task would be relatively easy in a star. Temperature is a measure of how much speed an atom of a given type has as it bangs around inside a cloud of such atoms. The higher the temperature, the higher the speed and the closer the atoms get to each other momentarily during a collision. In a star the temperatures are high enough that all the electrons quickly get knocked off the atoms, so one is really dealing with a mixed cloud of electrons and nuclei. At very high temperature the nuclei occasionally get close enough during collisions for the pulling-together short range nuclear force to turn on. Then the nuclei can stick together and go to a lower energy grouping of protons plus neutrons, releasing heat. Hot fusion nuclear energy is an attempt to carry out this process in the lab, using deuterium and mass-3 hydrogen (whose nucleus is a compact grouping of 1 proton and 2 neutrons) as the gas. Hot fusion requires that the gas be contained at temperatures of hundreds of millions of degrees, which can be done with the help of magnetic fields, but only for 1 or 2 seconds. The hope is to contain the gas for longer times. During the period of high temperature containment nuclear reactions occur during collisions. The main form of energy release is ejection of high energy neutrons and protons. The proton energy quickly converts to heat. The neutron energy can also be converted to heat but makes the equipment highly radioactive. It then becomes difficult to repair the equipment, which could make hot fusion a poor candidate for commercial power production. In any case hot fusion power is a dream that is still probably at least 50 years away. But most scientists view hot fusion as the only way to achieve fusion power. Hot fusion produces less radioactivity than fission power, is environmentally benign, and has a virtually limitless fuel supply on earth (many millions of years at present energy usage rates).

Lesson 7

So now we come to cold fusion. Cold fusion may provide an easier and non-radioactive way of releasing nuclear fusion energy. Cold fusion relies on a different way of letting the protons and neutrons in one nucleus make contact with those in another nucleus, so that the nuclear force can bring them into a more stable configuration. The requirement for any nuclear reaction to occur is that the reacting nuclei occupy the same volume of space. This condition is called particle overlap. In hot fusion particle overlap is brought about briefly by banging the nuclei together so as to overcome momentarily the repulsion of the two positive charges which try to keep the particles apart. In cold fusion particle overlap conditions are achieved by making deuterium nuclei act as fuzzy objects like electrons in atoms, instead of like tiny points. When either light or heavy hydrogen is added to a heavy metal, each hydrogen “atom” occupies a position inside the metal where it is surrounded by heavy metal atoms. This form of hydrogen is called interstitial hydrogen. With interstitial hydrogen the electrons of the hydrogen atom become part of the pool of electrons of the metal. Each hydrogen nucleus oscillates back and forth through a negatively charged electron cloud provided by the electrons of the metal. They can be thought of as moving back and forth like the pendulum in a grandfather clock. This vibration exists even at very low temperature, due to a peculiarity of a branch of physics called quantum mechanics. The vibration is called zero point motion. The nucleus then becomes a fuzzy object, like the electrons in an atom. But this amount of fuzziness is not enough to permit a hydrogen nucleus to make contact with another hydrogen nucleus. To get two or more hydrogen nuclei to share the same volume one must go one step further. In a metal electrical current is carried by electrons that act more like vibrating matter waves than like point particles. If electrons did not become wave-like inside solids, there would be no transistors and no present day computers. This wave-like kind of electron is called a Bloch function electron. The secret of cold fusion is that one needs Bloch function deuterons. One needs wave-like deuterons inside or on the surface of a solid in order that two or more deuterons share the same volume of space. But once the Bloch function deuterons are created, the nuclear force comes into play and the protons and neutrons making up the deuterons can rearrange themselves into the more nuclearly stable Bloch function helium configuration, with release of heat. To study cold fusion the experimenter has to force deuterons to assume the wave-like form and keep them in the wave-like state. Cold fusion experiments demonstrating release of excess heat show that this can be done. But at present no one knows how to do it reliably. Since cold fusion promises millions of years of energy without the problems of global warming or radioactivity, a real effort should be made to learn how.

For more tutorials go to www.hometown.aol.com/cffuture1 and www.hometown.aol.com/cffuture2

LENR (Low Energy Nuclear Reactions) Experiments

Review from
http://www.lenr-canr.org

Have you ever wondered what a physics laboratory looks like? They are seldom spacious or organized the way they are shown in movies. Most LENR researchers work at universities or home laboratories, with tight budgets in a crowded space. They keep old, broken equipment on shelves to scavenge parts for new experiments. In this section we present some photographs of equipment provided by researchers, and close up pictures of equipment. The actual cells, cathodes and other equipment used in electrolysis experiments often have an ad-hoc, home-made appearance, because they are made by hand. They have to be; they are unique, one-of-a kind prototypes. Nothing quite like them has ever been made before.

A visitor seeing a LENR experiment the first time may feel disappointed. It looks like any other electrochemical experiment. The heat or neutron flux produced by the experiment are so small they can only be detected with sensitive instruments. A null cathode that produces no effect looks exactly like an active cathode. The difference between one cathode and another is in the microscopic structure, or the traces of elements mixed in with the palladium. Only one kind of cold fusion looks dramatic i. e. the glow discharge reaction.

Here are a few photographs of cold fusion cells and devices (also see the cover page).

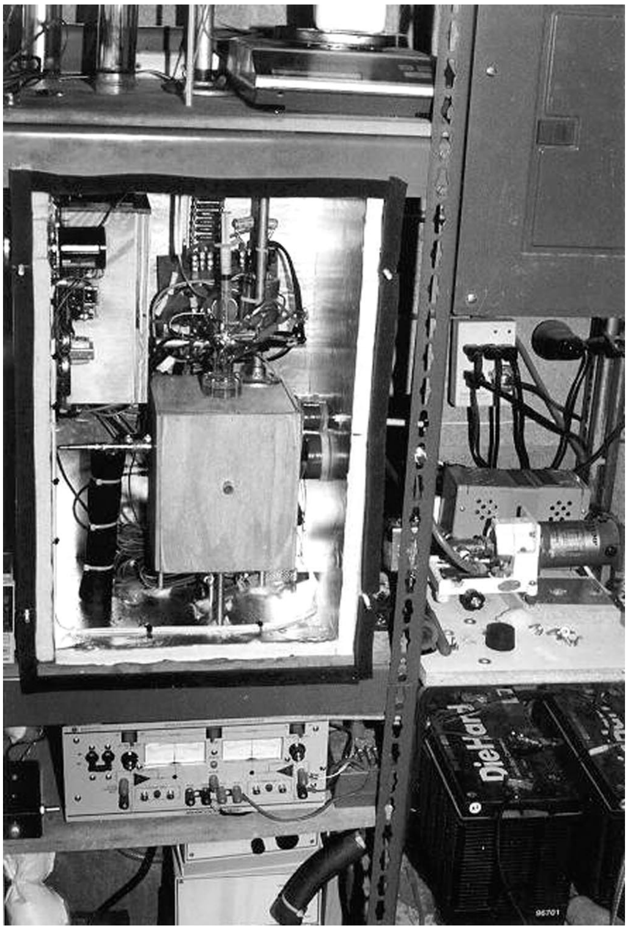

Fig.1
Box Calormeter

On Fig. 1 there is a calorimeter constructed by Edmund Storms. Note the DieHard® battery, lower right, that serves as an unninterruptable power supply. A power failure can ruin an experiment. Whenever possible, inexpensive, ordinary materials and instruments are used. However, experiments are never cheap, and they cannot be done on a shoestring. The equipment, arranged for another experiment (see Fig. 2), costs about $40,000.

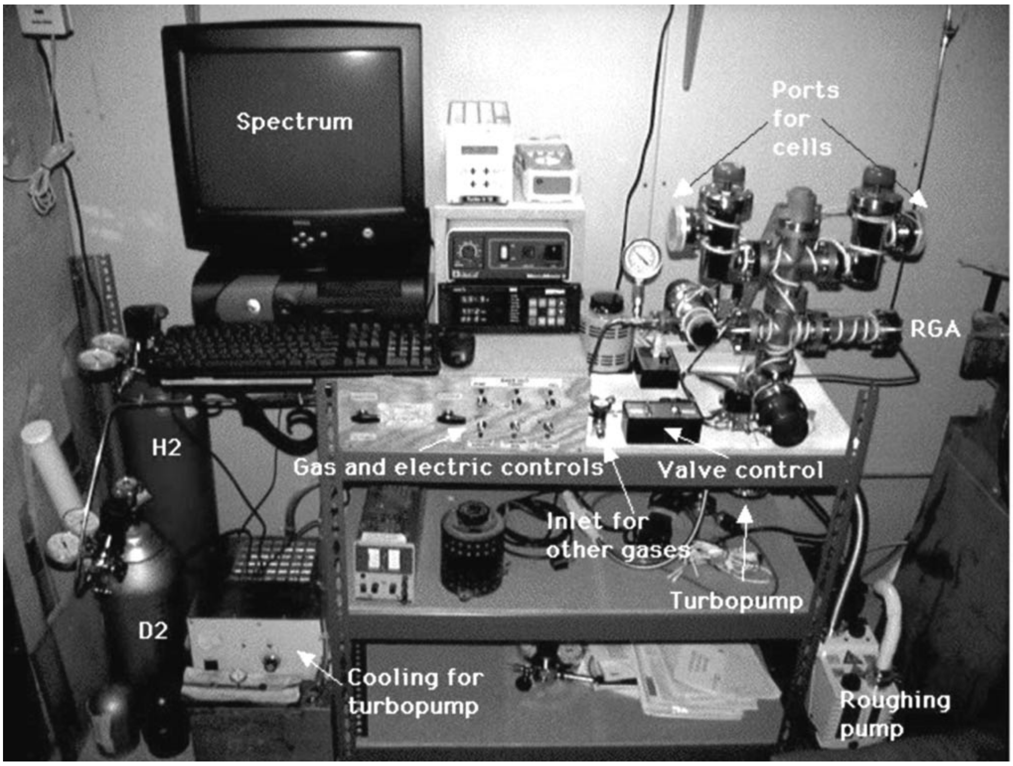

Fig.2 (On the left)
Vacuum system to prepare particles for gas loaded cold fusion cells, courtesy E. Storms.

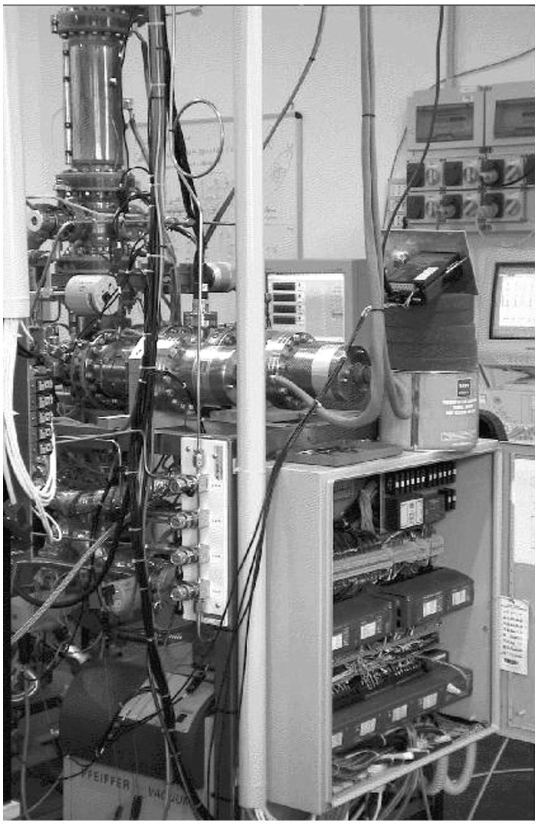

Fig.3

A high resolution mass spectrometer used for on-line helium detection during a cold fusion experiment at C. R. ENEA Frascati. (http://www.frascati.enea.it/nhe/)

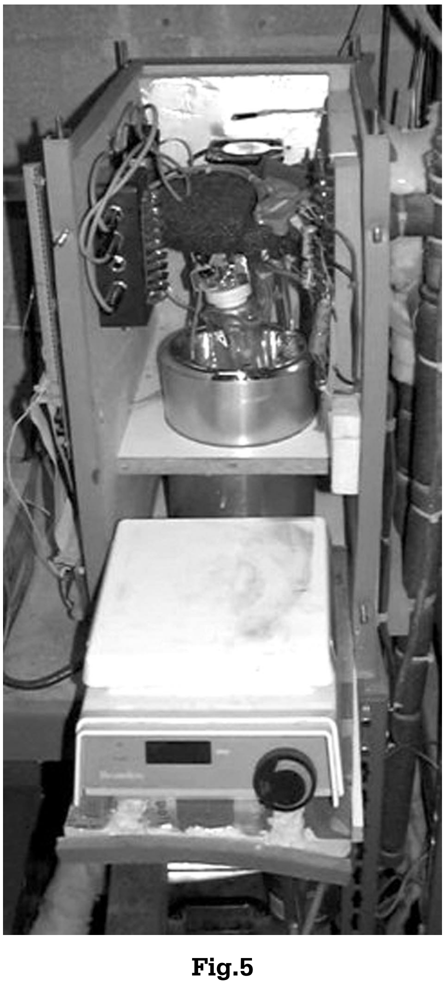

Fig.5

A flow-type cell, courtesy E. Storms

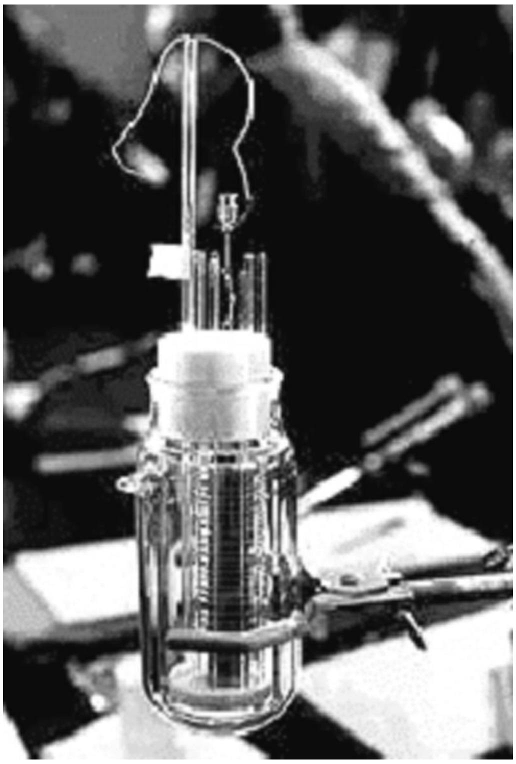

Fig.4

A cell at ENEA Frascati

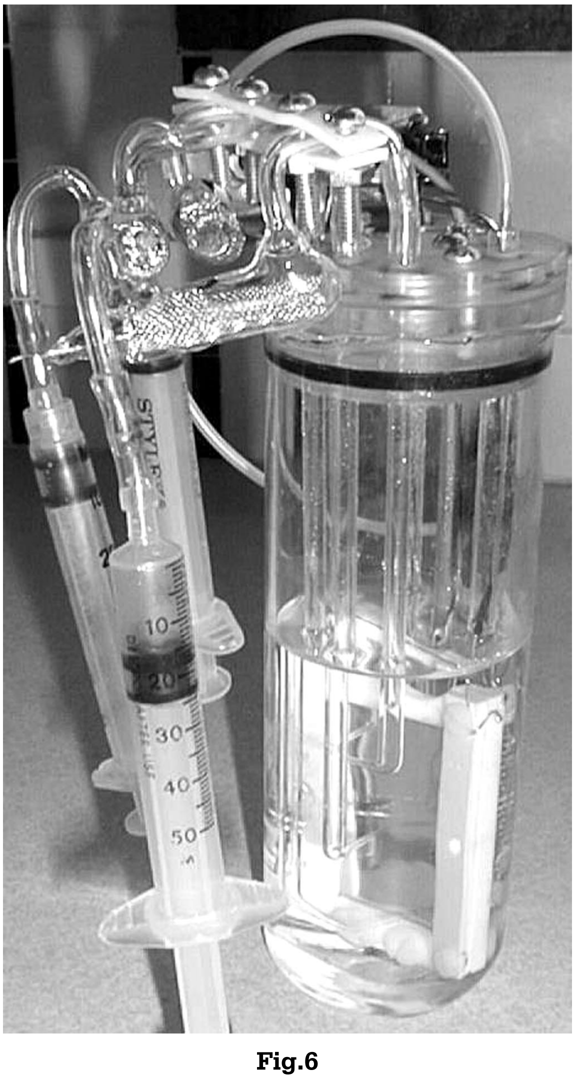

Fig.6

Close up of a Miley-style cell, courtesy E. Storms

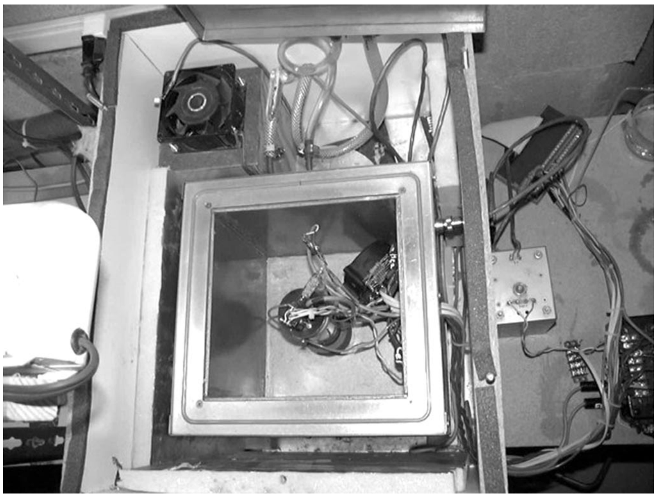

Fig.7

A cell installed inside a Thermonetics Seebeck calorimeter with the lid removed, courtesy E. Storms

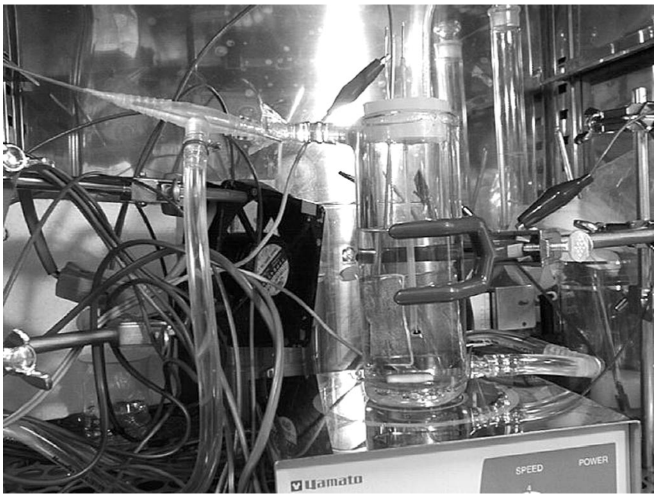
Yamato
SPEED
POWER

Fig.8

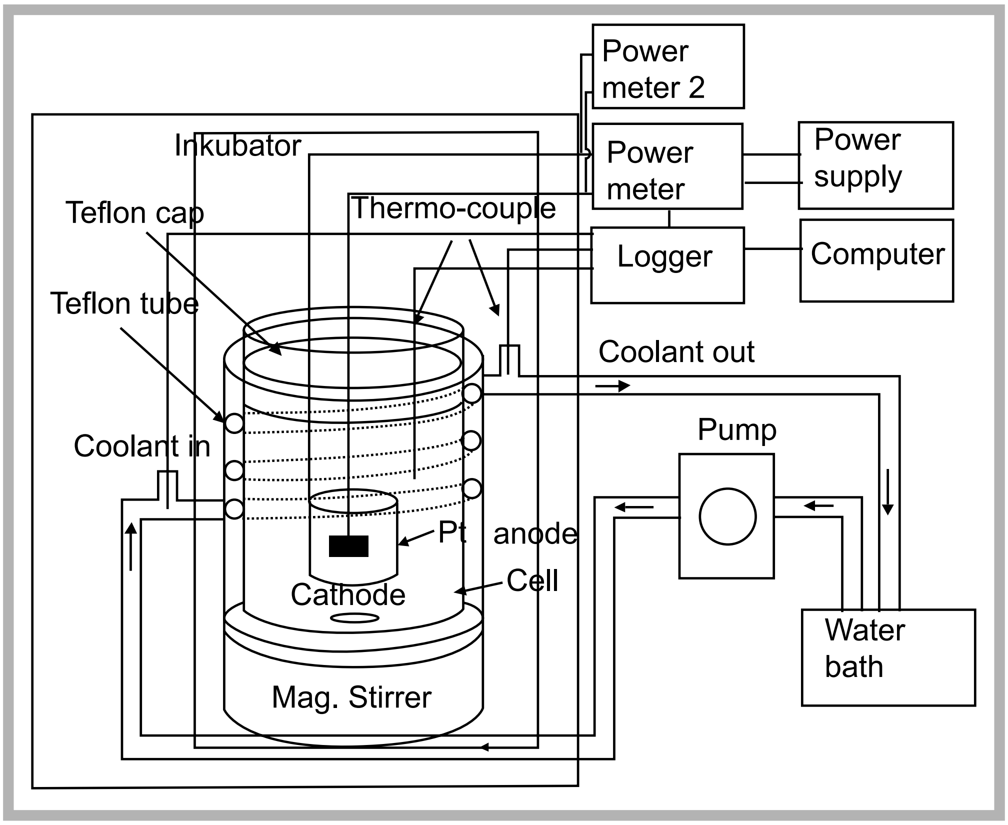
Power meter 2
Inkubator
Power meter
Power supply
Teflon cap
Thermo-couple
Logger
Computer
Teflon tube
Coolant out
Pump
Coolant in
Pt
anode
Cathode
Cell
Water bath
Mag. Stirrer

On the Fig. 8 you can see a glow discharge electrochemical cell at Hokkaido University, courtesy T. Mizuno. The cell is installed inside a crowded constant temperature air-cooled chamber. It placed on a magnetic mixer. Cooling water is pumped through the plastic tubes attached to the top and bottom. The muffin fan at the back circulates the air in the chamber

Fig.9

A schematic of the calorimeter shown above.

Hydrogen power engineering

Faraday Laboratories Ltd, Moscow, and Spectrum Investments Ltd, London, started joint R&D project on hydrogen power engineering. Photo: Alexander V. Frolov, Faraday Laboratories Ltd and Nicholas Moller, Spectrum Investments Ltd. The project includes designing and building of prototype to use hydrogen recombination process for heat generation.

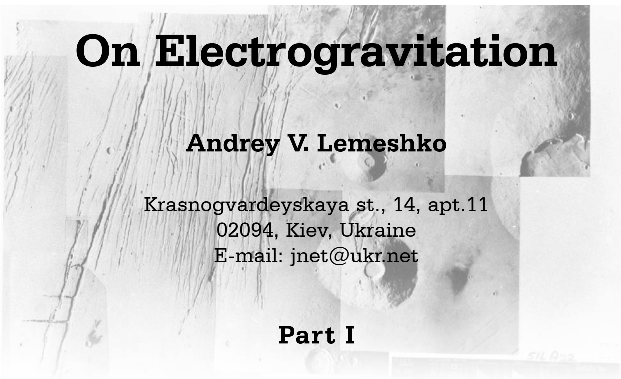

On Electrogravitation

Andrey V. Lemeshko

Krasnogvardeyskaya st., 14, apt.11
02094, Kiev, Ukraine
E-mail: jnet@ukr.net

Part I

"let's go to the Mars!"

Nowadays it becomes evident that exploration of deep space by means of rocket equipment is impossible. Aircrafts, which run on jet thrust, are too expensive and unreliable. Besides, they are too slow-speed if were considered on a scale of solar system. For example, the flight to the remote areas of our planetary system can drag on tens years. But the main disadvantage of rockets is their low load-carrying capacity. Thus, the most powerful transport system "Progress" can deliver less than 100 ton of load to the orbit of the Earth, and only 1 ton to the orbit of the Mars. Rather natural question arises as to whether any alternative ways of space exploration exist? Are there any transport systems, which can compete with rocket equipment on the interplanetary routs of future? Actually, at present there is only one technology, which is alternative to rocket ones – it is the so-called light sailers, using light pressure for acceleration. Laser and maser can be also used for their acceleration from satellites. However, there are also some disadvantages here.

There is also one more transport system, which is undeservedly consigned to oblivion. It uses electricity in acceleration aggregate. This transport system or the electric propulsion system is a result of researches and investigations of two little-known, but extremely talented American physicians and inventors, Thomas Townsend Brown and Paul Alfred Biefeld (the last one was also a professor of Astronomy).

Thomas T. Brown was the first who discovered electric principle of this propulsion system. Thus, in twenties of the XX century Thomas T. Brown, being the student of secondary school, became interested in X-rays. He wanted to reveal if the rays, outgoing the X-ray unit, could take positive effect. For his researches the inventor had got Coolidge Tube, the device, which radiates X-rays and consists of a glass tube. Deep vacuum is created in this glass tube. Anode and cathode are placed in the tube. Cathode radiates electrons and anode brakes them. X-rays are created at striking of electrons against anode. Brown used Coolidge Tube in the way nobody had done before. For ease of handling the researcher suspended the device on wires, which run to anode and cathode parallel to the Earth. During the operation of the device he noticed that every time when current fed to the tube, the tube was moved aside and tended to make slight progressive motion. Subsequent experiments demonstrated that the deviation value depended only on the value of voltage, which is between anode and cathode. The higher voltage is, the more deviation. X-rays were neither here not there. Much later at participation of Paul A. Biefeld, the same tendency to move was revealed at plane disc capacitors. The reason, which causes this motion, is the same, i.e. voltage, which exists between plates and not between anode and cathode as in Coolidge Tube.

Having run series of experiments, the researchers proved that electric energy could be directly transformed to mechanical. They also devised the physical principle, which were included in the scientific annals as Biefeld-Brown effect.

Editor's: Let us note that earlier Ampere and Faraday have been working on Electrokinetic effects which are easily detected at strong current. Modern researches on this point are known as "Sigalov's effect" (force appears in electric conductor which is bent at angle and at current impulse).

The essence of this effect lied in the fact that charged disc electric capacitors tend to move in the direction of their positive pole. At formulating of the essence of this effect the scientists consciously did not mention Coolidge Tube in order to demonstrate that according to their experimental observations, electron streams were not involved in generation of electromotive force. This effect of the direct transfer of electric energy to mechanic one (Biefeld-Brown effect) is observed, when current is fed to the Coolidge Tube or to the plane disc capacitor.

However, all this was realized much later, when in 1923 Brown became a student of Dennis University in Greenville, Ohio, where he met his teacher and later on a coauthor, Paul A. Biefeld. In school time the inventor came to the conclusion that the reopened by him phenomenon has nothing in common with X-rays and that high voltage, used for rays generation, is the basis of this phenomenon. Besides, Brown designed a device named "Gravitor". It looked like a usual bakelite box and was made as a simple construction, which consisted of several flat aluminum plates. The plates were placed as coins in a rouleau and separated with dielectric. It had only to place the box on the scale and connect it to some energy source with 100 kV voltage as the device either lost or put on about one percent of its weight that depended on polarity. An outside observer, who was not familiar with the essence of this effect, could think that mass was really changed. *(Editor's: Weight but not the mass is changed here. Weight is the force. There is a possibility of partial compensation of the weight by another force.)* But this was not the case. For better understanding of the processes, which take place at gravitor operating, let us make a mental experiment. Let us take a weight, put it on the scale and then raise it a little. The scale will certainly show that the weight has become lighter. After that let us press the weight then the scale will show that it has become heavier. However, in fact mass does not change. In the same way motive force, caused by high voltage, slightly raises the gravitor and at changing of polarity presses it. At that it seems that there is some change in weight. Thus, as if by chance, the scientist has discovered effect of direct transition of electric energy into mechanical one. *(Editor's: Let us specify the author's terminology. In T.T. Brown's Patent of 1927 it was mentioned that there* *was generated a force which partially compensated weight. The matter was always about the generation of x-force while gravitational field remained unchangeable).*

Having graduated from the University, from 1926 up to 1952 Brown already by himself was being occupied with development of his electrical propulsion system. At the same time whether in chase of fame or in hope to obtain money for his researches, T.T. Brown began asserting that he had discovered something new in physics, viz "electrogravitation", i.e. some connecting effect between gravitation and electricity. *(Editor's: In my opinion, T.T. Brown has never posed a problem in such a way. The matter always concerned the fact that X-force appears while gravitational field does not change.)* As for the development of the invention, it came to the decrease of weight of the device and increase of DC voltage, which fed the plates of the "gravitor". As a result the devices could fly and lift weight which greatly exceeded their own. Thus in 1953 in his laboratory Brown demonstrated the flight of aircraft which had a disk form. The device made 51 m/c (180 km/h), flying at circular course of 6 m diameter. This device was nothing but very lightweight capacitor in which by means of the special construction electromotive force was redirected in two planes upward and sideward. It made the device rise in air and fly around the mast. Current of 50kWtt voltage was supplied to the central mast by the wire. Then speed of "flying disks" became higher, the payload capacity was increased but that was again resulted by decrease of weight and increase of the supplied voltage. T.T. Brown so succeeded in decreasing of the thickness of capacitor plates and in making his devices superlight, that the witnesses of these demonstrations called these devices as "air films".

Later on all works in this area were placed on secret list and then closed because of lack of prospects. Thus, accounts and experiments demonstrated that it was impossible to overcome terrestrial attraction entirely and go to outer space by means of the gravitor because there are no such natural materials which could take huge electric and heat loads. Besides, sufficiently massive electric power station would be needed to produce required currents. *(Editor's: Thereupon the following question appears: why the author looks for "currents" while the essence of the technology is in applying of high voltage but not of the currents?!)* Moreover some alternative was made, i.e. jet engines. Gravitor could not compete with them by above-mentioned and quite objective reasons. In time, since the mechanism of transition of electric energy into mechanical one was not disclosed and Brown's populist idea has received some publicity, this effect became surrounded with idle talks and guess-works. However the scientists who researched this very effect, has developed only three theory wh ich could explain its existence.

The first theory was suggested by T.T. Brown, the discoverer of this effect. To his dying day Brown asserted that he had discovered the effect which could connect gravitation and electricity, i.e. "electrogravitation". But this theory can be easily refuted by practice. It is enough to put Brown's bakelite box to the scale in such a way as to place the plates of the "gravitor" at right angle to the surface of the scales. Then poles of the "gravitor" will be placed at the same level and parallel to ground and as a consequence electromotive force will influence in no way on the scale, since this force is directed sideward but not upward or downward. The scale will accurately show that there are no changes in weight and hence there is no **electrogravitation**. *(Editor's: The author will obtain propulsive force and it is directed not upward and downward but sideward. However the effect exists in any case. The changing of weight appears only when propulsive force generated by the device is directed against the weight force, i.e. upright.)*

The second theory was suggested by T. Brown's opponents. They asserted that electromotive force appears as a result of the so called "electron wind", i.e. electron streams which exist between the plates of the "gravitor". In their argumentations Brown's opponents rest upon the following facts which are well-known to every physicist. Fast electrons, as well as photons, have dual properties, i.e. interacting with matter and environment, they behave both as a wave and as a material particle. Since light or photons carry energy, then they heat bodies which are in their way and put pressure on the illuminated surface etc. Similarly electron streams carry heat or kinetic energy and hence they can put pressure on the anode of Coolidge Tube and get it moving forward. This theory seems to be interesting, however it explains appearance of electromotive force only in Coolidge Tube, but there is no explanation why discoid capacitors tend to move. Point is that there are no free electron streams in the capacitors. Usually between the plates of capacitors there is a dielectric which brings electron streams to nothing. If there is no electron wind, then what makes capacitors move?

Editor's: There is always a surface charge and reactive outflows of ions. But it is just a partial explanation of Biefeld-Brown's effect.

The third theory is based on the following fact. It is well known that electrons can have heat or Brownian chaotic motion. Free electrons, placed in a conductor, are in the similar chaotic heat motion before voltage is applied to this conductor. The higher temperature of the environment is the higher speed of this heat motion. According to school course of physics, heat energy or energy of chaotic motion of molecules (electrons) can be entirely transferred into mechanical one. If all molecules (electrons) start to move simultaneously and in the same direction, and if they push some piston toward this direction then heat energy of molecules (electrons) will be entirely transferred into mechanical work. In other ways, this piston will move until molecules (electrons) calm down. It is considered that this process is unrealizable in practice.

However, it seems that this is not quite the case, i.e. it is possible to regulate chaotic heat motion of electrons. Obviously this process takes place in the capacitor and in Coolidge Tube. It is no secret that when negative electrons try to move to the neighbor positive charged electrode, then dielectric layer which is placed between the plates of the capacitor prevents it. Electrons rest upon this layer and start to heat. Their internal energy is also increasing. The higher voltage is the higher heating of electrons, i.e. their capacity to have chaotic heat motion increases. However, chaotic motion of electrons is not observed on the capacitor plates, since the directed force (voltage) influences electrons. Electrons tend to move upward, downward, to the right and to the left, but voltage prevents it. Thus they just are forced to move in space in one direction and with the same speed. Capacitor or gravitor has to move alongside with electrons because plates of the capacitor (gravitor) become a piston which is pushed by heated electrons. This piston is in progressive motion. It starts to move under the influence of electrons, or rather together with electrons. The higher voltage is, the higher temperature of electrons and hence the higher their speed. Therefore speed of capacitor (gravitor) increases at voltage increase. There is a usual chain, i.e. electric energy transfers into heat energy and heat energy transfers into mechanical one. Approximately the similar process takes place in Coolidge Tube, but there the heat of electrons is caused by thermoelectric emission on the cathode. Electrons evaporate from the cathode surface. Voltage, existing between anode and cathode, makes electrons move in one and the same direction. They move towards anode, which serves as piston and takes heat push upon itself, i.e. it transfers heat energy of electrons into mechanical work. It is clear that the higher voltage is the more intensively electrons are heated. Therefore electrons more intensively influence anode and the suspended tube is deviated to a greater extent. (*Editor's: Nowadays this very explanation is the most admitted one.)*

Resting upon Biefeld-Brown's theory and experimental results and at sufficient financing, already within the next few years it would be possible to create a unique electric engine which can operate on hot or quick electrons. Besides there will appear a possibility to orbit the first spaceship equipped with a propulsion system similar to gravitor one's. In near future scheduled cargo-and-passenger service could be organized within planetary system. First of all such transportations can be organized between the Mars, Moon and Venus orbits that will let to start active colonization of these planets.

Thus, more than 200 flights to the Moon and back and about 70 flights to the Mars are possible by means of the gravitor which is made of modern fusions and dielectric materials and which is placed on the Earth orbit. At that there is a minimum risk for its breakdown during the flight, because gravitor has no movable parts and no explosives are used. Negative consequences for the crew in an improbable emergency are insignificant. Even if high voltage causes disruption of dielectric layer between the electrodes and voltaic arc appears then the spaceship, equipped with such a propulsion system, will mechanically arrive its destination. The source of energy will be its only trouble spot and it will happen only if nuclear reactor instead of solar batteries is used. There is also an idea to attach dwelling capsules and cargo bays, equipped with such electric engine, one after another, like carriages are attached to a locomotive. It will let to equip this interplanetary vehicle in several stages by compactly packed modules which are delivered from the Earth. Unfortunately, this project exists only on paper and support of government or financial institutions is not expected in the near future. Faint hope is rested upon enthusiasts but there are too few of such people, besides almost all of them are rather hardly related to space flights and science. Most of scientists even hear nothing about Biefeld-Brown Effect, because it has never become widely known, since all works of this field, except the very first experiments from the USA, were run and then closed in secrecy.

Part II

"...it's time to go to the stars"

Due to the theory of relativity there is an ingrained belief of several generations that it is impossible to move with more than 300 000 m/c speed, i.e. to move faster than light, since the postulates of this theory read that mass is directly depended on speed. Calculations demonstrate that while speed of some accelerated body verges towards light speed its mass will increase and in the point of equilibrium (V=c) it will unrestrictedly increase. Hence acceleration will tend to zero and speed will not virtually increase, no matter how long acceleration speed acts on the object. In other words, the accelerating mass dissipates speed. In principle it is correct but another conclusion can be also drawn from the dependence of mass and speed, for example, the following one: to accelerate an object to the speed which exceeds light speed, mass of the accelerated body should decrease and its speed should increase. Any physicist will say that this statement is correct. At first sight it is practically impossible to realize this idea therefore it was wrongly give to writers of science fiction. They added this idea to their armory and the hypothetical device, which realizes it, was named as gravitational engine. Let us realize this idea from the practical point of view and basing only on facts and logic.

Thus let us apply to the facts and give the description of one experiment. Professor of Pulkovo Observatory N.A. Kozyrev is the author of this experiment, he was also the first to discover the phenomenon of lunar volcanism. Kozyrev used a gyroscope, i.e. the device consisting of two rings of different diameter. These rings are placed athwart one in another and movably jointed. Top is attached to the inner ring through the cardan. The scientist put usual thermos filled with hot water near the beam scale. The top of the gyroscope was preliminarily spun up counterclockwise. At that balance indicator showed that gyroscope top at 90 gram weight became 4 milligram lighter in weight. Then Kozyrev started to pour water of room temperature inside the device through the vinyl chloride tube fixed into the hole, which was made in the plug of the thermos. Seemingly it is impossible to influence upon the gyroscope operating and its weight by a balloon filled with hot water, which is being cooled. Besides the thermos consists of a double-walled container which practically entirely excludes heat exchange with the environment. However balance indicator moved at one or two points, i.e. such connection existed.

This experiment can be divided into two steps. At the first stage the thermos filled with hot water is placed near the gyroscope and we can register the decrease of the top weight. At the second stage water of room temperature is poured into the thermos and the scale again shows that the weight has been changed. Somehow Kozyrev explained the second stage of this experiment. He assumed that when we pour cold water in the thermos, then the equilibrium of the system is broken because some irreversible processes take place there. Cold water can not cause temperature increase of water in the thermos. Until the system come to equilibrium at a new level, i.e. until the same temperature is set in the whole volume of the thermos, this process will cause the condensation of time which "additionally" influences the top. However the scientist missed that the weight of the top had decreased long before the moment when cold water was poured into the thermos, i.e. before irreversible processes appeared there. *(Editor's: Evidently, the author of the article is not well familiar with Kozyrev's researches and with works of other experimenters who investigated this problem.)* In other words, at the first stage of the experiment there is nothing to condense time and that insignificant weight condensation, which was equal to 4 milligram, absolutely does not fallen into his theory. Some other process is presented there.

It is quite possible that at the first stage of the experiment we deal with nullification of the top weight, i.e. mass of the top tends to zero. Some conclusion arises from this assumption, i.e. hot water can influence the mass of the top in some still incomprehensible way. *(Editor's: The point is not in water temperature but in the process of changing of the entropy.)* Thus there is a simplest and primitive model of the gravitational engine, viz there are hot water along with gyroscope and the required decrease of mass at the "outlet". Certainly, it is very bold assumption but it can be easily checked. Near the gyroscope there should be placed thermos filled with not boiling water (of 100°C temperature) but with water of 50°C temperature. If the weight of the top decreased by one half (for example, by 2 milligram instead of 4 milligram), then we can surely ascertain that the hotter water in the thermos is the more decrease of mass. Besides, each 100°C of water heating will decrease mass of the top by 4 milligram. It is not very difficult to calculate the temperature at which mass of the top will verge toward zero and the top will start accumulating negative mass.

Let us now assume that our experiment was successful then water as well as any other matter cannot be heated to such high temperature. However it is possible to heat electrons. They are very compact and have insignificant mass and can be heated to very high temperature. It is necessary to take several plates then to place a dielectric between them and to apply voltage. Electrons rest on the dielectric and start heating. The top can be rotated anti-clockwise by means of electric engines, i.e. capacitor and gyroscope should be combined in one system with common center of gravity. It is not very difficult, as the saying goes, it is just an engineering problem. The higher temperature of electrons is the more decrease of the top mass. Finally there is a moment when mass of the capacitor-gyroscope system will reach zero and this electric machine will rise to the surface of the gravitational field of the Earth or any other planet. It will be Biefeld-Brown's force that will make our apparatus taking off the gravitational field of the Earth. This very force will start to accelerate our apparatus in the space vacuum and top of the gyroscope will lose its mass which appears at acceleration. Thus it is quite possible to reach speed, which is close to the light one, or even to overcome the threshold of light speed.

Part III

"...what is new?"

What is implied by gravitational waves? This question has been exciting the scientists from the half of the XX century. But still nowadays, in spite of all attempts of the scientists to dissolve the problem, these waves were not discovered. There is a standard way to discover them, which is based on general relativity. Basing on this theory, scientists assume that gravitational waves should change in some way the weight of material objects. According to this assumption the scientists suspend very heavy balls to very precision scale and then try to trace changes of their weight. There were only negative results. It seems that mass changes so quickly and to such short period of time that scale fails to react upon these changes. However there is another way which seems to be very promising.

Dependence of time and gravitation is a well-known fact. Thus speed of time course depends on force of the gravitational field of an object. Stars and planets can be considered as an example. The more intense attraction near them the more slowly time passes there. In other words, the greater mass of the planet the more slowly time passes near it. It is possible that oscillations of masses, which generate gravitational waves, in some way change speed of the time course along the whole way of the propagation of gravitational waves. In other words, gravitational waves are always attended with slight time fluctuations. Time starts to pass a little faster or a little more slowly as compared with its usual course. These changes can be traced by means of usual quartz plates. Let me remind of the fact that quartz plates are used at some models of clocks to keep time. Thus, oscillations of gravitational masses produce in space both gravitational waves and time fluctuations (chronowaves), which can be easily detected. Possibly these chronowaves are the part of gravitational waves. This assumption is already a scientific fact which is unfortunately ignored by more than one generation of researchers.

Professor N.A. Kozyrev was the first who surmised about the existence of time ways (chronowaves). He called them as "time flows". The scientist worked out and run a simple experiment to confirm his hypothesis. Telescope and quartz plates were used. Quartz plates were placed in focus of the telescope, which was pointed to some bright star. Its objective lens was covered with black paper or tin plate in order to exclude the influence of light beams. At that quartz plates reacted upon the presence of chronowaves. Thus there was changed the frequency of oscillations of plates which were placed in focus of the telescope.

Besides the scientist discovered that in focus of the telescope there were changed electroconductivity and volume of some matters. It is explainable since some parameters of space as well as speed of passing of some physical processes and phenomena depend on the speed of time course. Outwardly it appears in change of electroconductivity and volume. In spite of the fact that during the experiments tin plate (thick metal cover) was used some skeptics insisting that the point is in infrared radiation which is produced by the heated cover. However they was silenced by Kozyrev's researches. Basing on the fact that we usually see stars not where they really are, but at the place they were at the moment of emanation of light signal, the scientist calculated the location of Procyon star, which is the nearest to the Earth. Kozyrev pointed the telescope to this "clear" part of the sky zone that allowed to avoid the influence of infrared radiation and to fix the location of the star. Thus it is nothing but chronowaves generated by the star that can change oscillation frequency of quartz plates.

After Kozyrev's death this field of research is not considered by official science and the very existence of chronowaves is denied. Only some theorists warily impute the above-listed effects to the action of some torsion fields and waves the existence of which is rather questionable. However chronowaves exist and they can be used in practice, for example, to get information from any point of our galaxy in several seconds. **It is quite possible since chronowaves have no mass and thus they are propagated more quickly than light waves.** However it is still a question what will transmit this information to us because the transmitter of these waves yet has not been invented. *(Editor's: It is interesting what mass of light waves the author means. It is a pity that the author of the article did not read 9 previous issues of NET magazine! There was a detailed demonstration of the connection of Kozyrev's theory and experiments with the theory of longitudinal waves in aether. If to replace Kozyrev's notion of "waves of time density" by the notion of "waves of aether density" then we will get a real experimental approach which can help to solve the problem of antigravitation and time control.)*

Of course theoretically chronowaves can be also generated without participation of great gravitational waves (it seems that heat processes can also generate them). However no particular success was achieved in this field. Thus by means of chronowave transmitter it is possible just to "listen" to remote stars the light of which will reach the Earth in many years. This receiver is very simple and it can be made by anyone who knows a little about radio engineering. *Under home conditions it is better to use a miniature incandescent bulb (lamp with filament tungsten) as a receiver and usual ohmmeter as measuring instrument.*

If in usual receivers radio waves are excited by weak electric currents existing in antenna then chronowave antenna is constantly under tension. Chronowaves do not excite current there but change some characteristics of it, influencing on electroconductivity of the matter which was used for the antenna (let us remember Kozyrev's experiments). Later on these insignificant changes are transferred in audio signal or fixed by any other means. That is the principle of operation of the device.

Editor's: These experiments on the registration of flicker effect in more detail are described in A.M. Mishin works which were published in previous issues of our magazine. As the reader could notice, the editors' opinion in many respects does not coincide with the author's point of view. (Alexander V. Frolov)

References

1. *G.Ya. Myakishev, B.B. Buhovtsev,* Physics, *M.: Prosveschenie, 1991*
2. *A.S. Kuzovkin, N.N. Nepomnyaschy,* What happened to torpedo-boat "Eldridge"? *M.: Znanie, 1991, Series "Znak voprosa" #3, p.37*
3. *S.N. Zigunenko,* How the time machine is made? *M.: Znanie, 1991, Series "Znak voprosa" #5, p.35*
4. A. Einstein and gravics, *Collection of articles, M.: Mir, 1979*
5. *A.D. Chernin* Physics of time, *M.: Nauka, 1978*
6. *A.I. Veinik* Thermodinamics of natural processes

Time Machine

Alexander V. Frolov gave this interview to the Spain magazine "Mas Alla de la Ciencia" ("Science"). The interview concerns some questions about scientific-research work on investigation of active properties of time. Let us remind that in 2002 Faraday Lab Ltd and Vadim A. Chernobrov (Moscow) started the join Time Machine Project.

Miguel Segui: How does it value the results obtained by the prototype of time machine?
Alexander V. Frolov: Usually it is about 3% changes, i.e. 3 sec per 100 sec of experiments.

M.S.: How other devices are preparing inside the project Time Machine?
A.F.: Now we have project with special electromagnets, which are emitters of the longitudinal waves. So, other devices are electronics parts to provide impulse current. But it is not the only one version of the system. If the emitters of the longitudinal waves are based on other principles so all device has other design.

M.S.: Do you believe that some day the time trips will be possible?
A.F.: Yes, sure. We have the clear theory for it and it is based on understanding of physical sense of the aether, its relation with notion of time and with concept about existence of elements of matter in space-time. According to aether-vortex theory any element of matter is created from aether and its temporal parameters depend on density of aether.

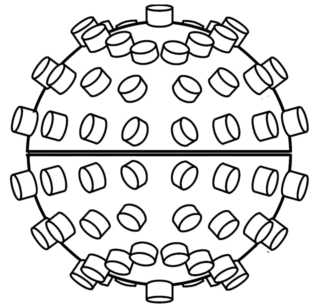

M.S.: What practical applications can have the prototypes that are developing?
A.F.: Any change in the time course is the method of influence onto biological systems and the medical aspect is the main application. Also any physicist know relation between G (the gravity constant) and time course (rate of time). For example, time is slower on the surface of Earth than in space. So, local time changes (which are changes in density of aether or direction of aether wind) can be used to get propulsion force without reactive mass

flow. Also we can hope that some new materials can be produced in special area of slowed or accelerated time course.

M.S.: Which are the main difficulties of developing of these devices at the moment?
A.F.: For us now it is a part of the work that is related with designing of the powerful current electronics circuits. Also there are no known analogies for patent work, but we hope to complete our patent claim in March of 2003.

M.S.: Is it foreseen that they carry out experiments with human beings and the machine of modification of the time, just as V. Chernobrov carried out?
A.F.: We, i.e. Faraday Laboratories Ltd, are developing small unit for testing of the principles. In future we'll work with other design and with participation of the pilot (human passengers). But it will be the system of other kind. Let me explain: now we are working to get small local change of the aether density, so called "chronal charge". It is analogy with electric charge, which produces electric field in space. If the electric charge is moving then it will produce magnetic field. According to our plans after confirmation of the principles we are going to create design with moving "chronal charge" to generate "chronal field". There are suppositions that this field can be used practically for antigravitation propulsion transport.

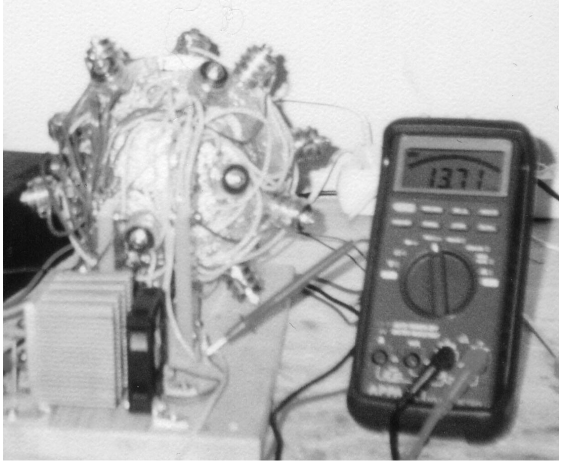

"FRCM" – Avalanchedrive

Fractioning and Replacement of a Constant Mass

Murilo Luciano Filho, Brazil

murilo@intrumenti.com.br, avalanchedrive@hotmail.com

For the while, this project has only these two names, as above. There is still a lot to do upon this development that involves free energy and / or "over-unity". It is designated to be of free use and application by everybody and everywhere, without any royalties just like it was a very old idea, which is now opened for entire community use. The inventor has enough strong reasons for this will.

This is a kind of donation, a personal donation, for the very best future of the Earth. Pioneers are needed (late pioneers?) Any person may be involved in this development to make money since there are no barriers for that. This means that this is an open idea and in this way it must stay. The principles are new and original and the energy involved is quite huge. In the inventor's mind there is realized the possibility to create an expert hands-on group to finish the project. In this article there is a rough draw which seems to be enigmatic but in fact it can clear the understanding of the principle and show the situation at which body may offer controlled spatial variation of length for the same weight.

The short text and a schematic draw (see Fig.1) are the way to show the principles and a few possible construction solutions. The photo shows a part of the device which can clear the so called weight localized rarefaction. Possibly the understanding of capacity for an observer will be a little forced. If necessary author can send the full patent requirement text, which contains 54 pages and 32 figures, not in English but in Portuguese.

FRCM – "AVALANCHEDRIVE"

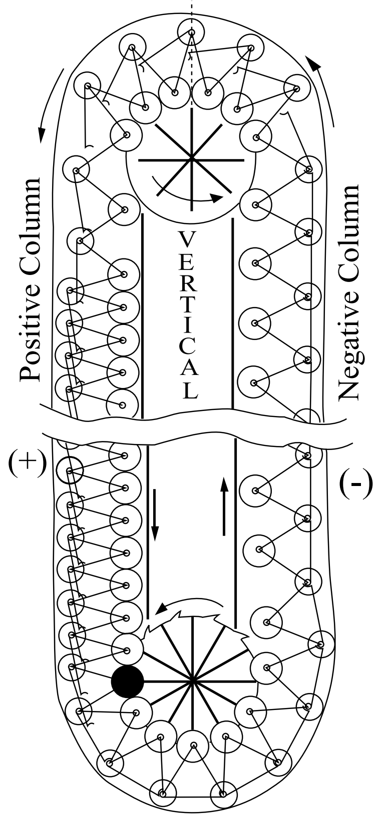

Fig. 1

Understanding of **FRCM / avalanchedrive** is a challenge in itself. When you start to learn about FRCM you will find complete original problems to deal. **Free your mind basing on some mechanical and physics knowledge and then avalanchedrive will come to make a part of your life. Feel free to work.**

PATENT REQUIREMENT RELEASE

This patent requirement is basically the application of a new physical principle which could be defined and found at least 400 years ago, so simple it is. Surely it is one of such things which was let behind by human mind and concern. FRCM is odd and amazing. The key for this project is a special body that can behave like a solid and like a fluid, at the same time, with combined characteristics just as designed and required.

The mass potential of a specially designed segmented body is hold and managed, and throughout proper means is sequentially fractured while suspension balance is kept. This causes, after start, a constant and cyclic movement, since arrangement goes, and the conducting and forcing form of the mass is not interrupted.

Also called "avalanchedrive", this principle work is based on just three main parts or components: 1st is the special segmented body – a chain; 2nd is the stationary and external assembling to hold and conduce the chain; 3rd is the straight vertical, in order to achieve the maximal weight storage position in the arrangement.

The chain is an endless repetitive zigzag construction with weights and defined angles that looks like an entire and elongated ellipse. Its design makes easy the change of its profile and form (mutant profile) that will be contracted or elongated, according to the way it is organized, supported and driven. When contracted or expanded, the chain assumes its maximal or minimal weight, in comparatively the same heights and level, but at different cross sections (tunneling?).

For distribution of the chain track the external assembling forms a kind of duct or rail, which is also elliptical. It allows to the chain to be contracted and expanded, to change direction, move, and pass away, as desired. The chain follows the way that the rail allows, in four different zones: mass-united zone, or positive column, or weight capacitor; the lower reversion input zone; the liberation zone, or negative column, with same height of the positive one; and the upper reversion zone, and its "zero point", or non-resistance top.

Both reversion zones are composed mainly of wheels. However, while the lower zone has to hang, open and bolt the chain, in order to change its profile and natural falling-down trajectory, the upper has a very passive function, that is just to let the chain cross over, and avoid any turn back possibility.

Fact is that in all circuit the chain follows with passivity the "voids" of trails, but assumes a very strong positive and pressure action in mass-united zone, especially in the button, where to the wheel there is applied all the active potential weight of the device, formed by the entire stack of the full contracted state, forcing movement.

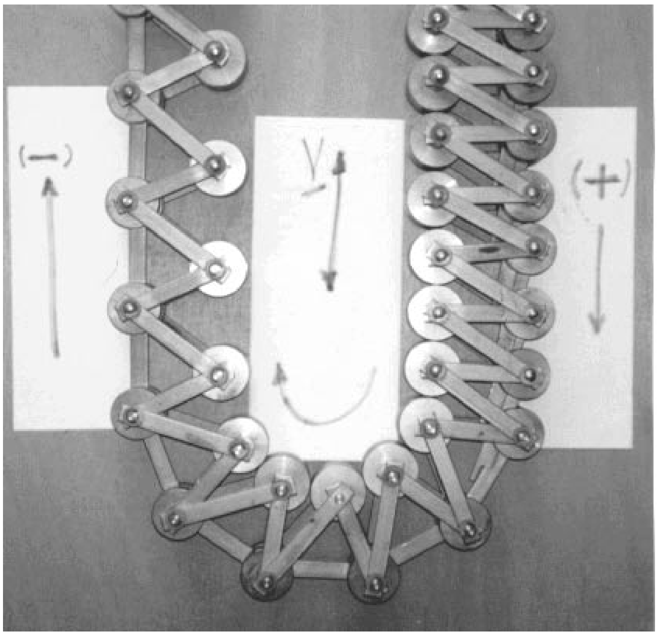

Photo of FRCM (Avalanchedrive)

It is easy to understand how and why FRCM works. Weight of the positive column is directly connected – as a compressed spring – against the relatively lighter negative column, through lower reversion wheel, where happens a kind of flux (the "solid flux") of the chain, that never can find a state of rest. Cycles of "breaking" of the chain can beconsidered as the instant of touch of each corner of the chain to the lower step wheel. To complete the circuitry, the chain excess, which is much lighter, passes over the upper wheel and turns back to the positive column top. It is impossible to turn the full subtracted potential to its source but mechanically the easier job to be done in FRCM.

FRCM manages and converts mass in energy...

The chain flows from the positive column, naturally finds changed profile and reverses to opposite way, and then it is pushed up to the negative column and to top. Finally it is collected, friendly, to restore the potential, at non-stop run! The only alternative to the chain is to move and move, and apply the positive surplus of weight to wrench under axle of the wheel.

This FRCM principle allows develop the equipment which is going to be able to take energy, work and force from gravity potential, so as it happens from wind and water at electrical generator machines, today. The potential is kept in such critical situation, that the only way for it is to escape and move. Besides in consequence of the suspending mass state the linear constant repetition of event will suffer. Water and other fluids do not offer "mutant profile" with fixed value of pressure, just as solid materials. There are many arguments for this view.

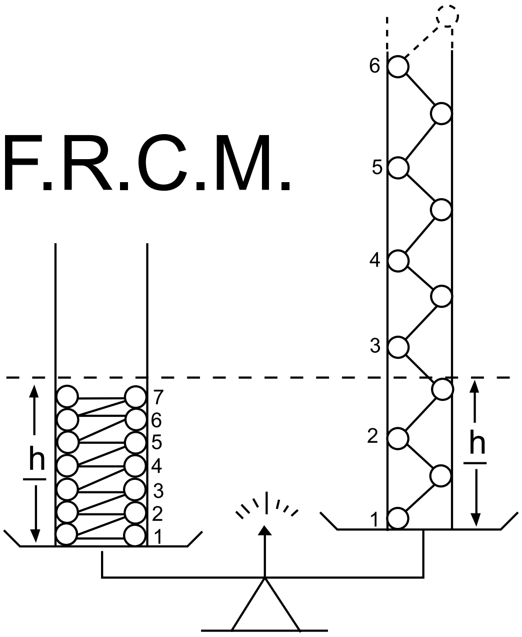

Fig. 2
Segment of a chain

Considering all mechanical losses, a part of the mass will be very active and applied in eccentric, or radial position to the wheel, even with the risk to cause too high speed. FRCM manages and converts mass in energy.

See a non-scale schematic draw (Fig. 2) and photo of a chain segment to compare densities between columns at right and at left.

A theoretical case can be calculated: at total chain weight - 100, general losses - 20, total negative resistance - 30, active or positive surplus is 50 – 30 = 20. This surplus that may be of 2kg, 200kg or 2,000kg, depending on scale, will force and free edge of the wheel, while the total weight keeps confined in the device. In the above case the difference of balance is 1.66. This means that the flowing velocity of the negative pile is 1.66 times higher than the falling of opposite positive pile. Just like a "solid venture", this change and relation is fixed, and in other cases may be varied or calculated to 3, 4 or even more! In all cases the speed must be hold that is not very easy job while one faces gravity acceleration.

The author of FRCM conceptions feels like these ideas have been around for many centuries, and that they are now discovered just like old draws. And this can be the bigger and the best moment to release this. Many thousands of different designs and solutions are coming in the future at looking for the better performance, costs and durability. Author thinks it to be strange that previous guesses could not reach this macro-physical concept. To build FRCM circuits, even re-applied or used old pieces, for example from ships and trains, will be useful.

This is only a start to a "new-old" development and clue to the sources of the modern mechanical engineering. Certainly some oppositions should be overcome. It is expected that other names are also going to be created and proposed to FRCM, but "avalanchedrive" seems to be a good one.

Author expects to get the invention and concept credits but not royalties and would also be glad to keep working on this concern. So, this message is a kind of invitation...

Editor: Unfortunately we have no photo or video documents of the operative device. We wait for readers' comments on this invention.

Complete text of the patent is avaible in Pourtuguese only.

ELECTRIFY
an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy systems,
EVents and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20,
Switzerland
Tel. 41-022-733.88.76, Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com
website: http://www.ipsgeneva.com

Hydrogen-Powered Vehicles at Least a Decade Away

Peter Behr
Greg Schneider

Washington Post
http://www.washingtonpost.com/wp-dyn/articles A635122003Jan29.html
Thursday, January 30, 2003; Page A09

This information is forwarded as a courtesy from:
http://www.integrityresearchinstitute.org

President Bush's vision of a hydrogen-powered, non-polluting "Freedom" car for the next generation of American motorists pulled out silently from a Newport Beach, Calif., garage with Gregg Kelly at the wheel, bound for his office 10 miles away.

Kelly, president of a California robotics company, happens to drive a Toyota prototype of a hydrogen-fueled car, one of a handful in the United States today.

It will take at least a decade before a host of technological, economic and political barriers are overcome, permitting a fleet of these revolutionary vehicles to appear on U.S. highways, industry experts said. But by including the goal of hydrogen fuels in his State of the Union message Tuesday, Bush has opened the door to a fledgling movement that has already attracted a surprising coalition of supporters.

Environmentalists, automakers, oil companies and coal producers, engineering labs and strategists have seized on hydrogen as an almost too-good-to-be-true power source. It is abundant in water and air, it burns cleanly and it could free the nation from its dependence on Middle Eastern oil.

"For President Bush to frame the goal as he did is significant," said Jeremy Rifkin, consultant and author of a book advocating a transition from oil and gas to hydrogen. "How much is made of this, time will tell."

The president said he hoped that Americans born today would learn to drive in hydrogen-powered cars, a schedule that auto experts said could technically be met. But whether these vehicles will be commercially available depends on a huge array of variables.

First, the technology is still incomplete and unaffordably expensive. The specially equipped Toyota Highlander that Kelly drives has no price tag. The Japanese manufacturer, after investing millions of dollars in research, lent the vehicle to a University of California research project that Kelly's company supports. "My checkbook isn't fat enough," Kelly said.

The car is powered by electricity generated in a fuel cell by chemically combining hydrogen and oxygen. The engine spits out water drops instead of the carbon dioxide and other pollutants generated by burning gasoline.

With foreign manufacturers committed to press ahead, Detroit's carmakers have had to accelerate their research programs. In June, Ford Motor Co. will unveil a prototype car that uses hydrogen to power an internal combustion engine – part of a "bridging strategy" to help ease hydrogen into the marketplace until fuel cells are fully developed.

General Motors Corp. has developed a fuel cell-powered, car-sized "skateboard" – four wheels attached to a platform less than a foot thick, to which any kind of car body could be buckled.

Rather than use fuel cells, BMW has refitted 10 of its $70,000 Model 745 sedans with hybrid engines that burn either gasoline or liquid hydrogen directly. It could be mass producing them by the end of the decade at a "reasonable" cost for its customers if there were enough hydrogen fueling stations to power them, said spokesman Gordon Keil. "We're trying to get [fuel suppliers] interested in hydrogen. We've not met with a lot of enthusiasm."

As daunting as the engineering challenge is the need for a national hydrogen fuel infrastructure – factories to produce the fuel, pipelines and trucks to distribute it and stations to store and sell it. Environmentalists dream of a totally "green" strategy in which solar or wind power is used to separate hydrogen from water – an approach whose costs now would be prohibitive. A nearer prospect is producing hydrogen from natural gas or coal, however in either case, the carbon dioxide byproduct would have to be injected underground to avoid a huge increase in greenhouse-gas emissions, experts say.

Rifkin argues that an eventual scarcity of oil and gas, decades ahead, will push prices of these fuels up to a point where hydrogen becomes cost-competitive. "It isn't a problem that will yield to technology alone," agreed David M. Nemtzow, president of the Alliance to Save Energy.

While all the major automakers are developing fuel-cell technology, most are cautious about hyping it. "We don't want to get too exuberant about it in that sense, overselling it," said Greg Dana, vice president for environmental affairs at the Alliance of Automobile Manufacturers.

Fresh in some minds is the experience of the Clinton administration, which launched a high-profile, $1.5 billion research venture with the

Detroit automakers a decade ago to produce an 80 miles-per-gallon family car. No cars emerged, and the Bush administration halted the venture in favor of its hydrogen strategy.

Some of the president's political opponents contend the hydrogen option is a way of deflecting criticism over administration policies favoring energy production over conservation. "The president seems content with the auto industry's approach: 'Don't make us do anything today'," said the Sierra Club's Daniel Becker. Others say it does not go nearly far enough. Sen. Byron L. Dorgan (D-N.D), chairman of the Democratic Policy Committee, said recently, "It's moving in the right direction. But his proposal is rather timid. I think we need a bolder plan."

Bush's spending plan for the hydrogen project, $1.5 billion over five years, represents a $500 million increase over his current budget. The administration proposes to earmark $273 million for the 2004 fiscal year, but did not offer many specifics yesterday. The funding would support research on fuel cells, vehicle technology and distribution issues.

The magnitude of the goal demands an effort on the scale of the Apollo Moon project, Dorgan said yesterday. "You have to set benchmarks for five, 10 years out."

UNUSUAL PERMANENT MAGNET MOTORS

Thomas Valone

Integrity Research Institute,
www.integrityresearchinstitute.org
1220 L St. NW, Suite 100-232, Washington, DC 20005
Email: iri@erols.com

Abstract

Permanent magnet motors that try to achieve unusual overunity efficiencies with changes in wiring geometry, electronic switching schemes and magnetic configurations often are not successful. There are some designs that should be regarded as conventional and others as promising. Hopefully this article will help the reader to tell the difference before investing or accepting investment. Note: patents can be viewed for free at www.uspto.gov and also http://gb.espacenet.com/espacenet.

Introduction

An article about permanent magnet (PM) motors would not be complete without first reviewing the basic configurations that are present on the market today. Commercial PM motors are necessarily DC motors since their magnets are permanently polarized before assembly. Many PM motors which use brushes are switching to brushless motors that promise less friction and wear. Brushless motors include electronic commutation or step motors. A step motor, often used in the automotive industry, offers more continuous duty torque per unit of volume than any other electric motor but it is often a lower speed motor. The electronic commutation design is applicable to the switched reluctance (SR) motor. The SR motor substitutes soft iron in the place of higher cost permanent magnets for the outer stator and instead has an inner PM rotor. Brushless motors in general produce torque from current in the armature by the application of Faraday's Law. The ideal PM motor has a linear torque vs. speed curve. There are both outer rotor and inner rotor designs that are standard in PM motors.

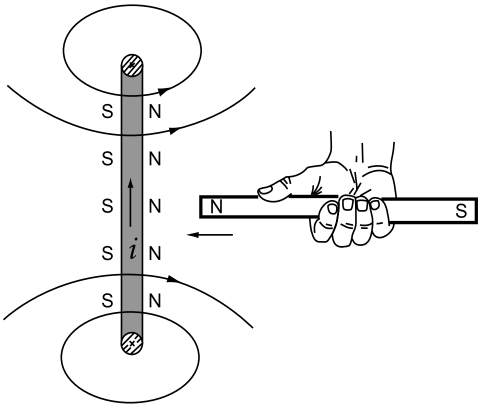

Fig.1

Lenz's Law
Induced B-field opposes motion.

To point out the focus of many of the problems with analyzing motors, the *Motion Control Handbook* (Designfax, May, 1989, p. 33) says that there is a "very important relationship between torque and back emf that is sometimes not understood." This relates to the electromotive force (emf) that is produced by the application of a changing magnetic field (dB/dt). In engineering terms, the "torque constant" (N-m/amp) equals the "back emf constant" (V/radian/sec). In physics, the motor terminal voltage

Detroit automakers a decade ago to produce an 80 miles-per-gallon family car. No cars emerged, and the Bush administration halted the venture in favor of its hydrogen strategy.

Some of the president's political opponents contend the hydrogen option is a way of deflecting criticism over administration policies favoring energy production over conservation. "The president seems content with the auto industry's approach: 'Don't make us do anything today'," said the Sierra Club's Daniel Becker. Others say it does not go nearly far enough. Sen. Byron L. Dorgan (D-N.D), chairman of the Democratic Policy Committee, said recently, "It's moving in the right direction. But his proposal is rather timid. I think we need a bolder plan."

Bush's spending plan for the hydrogen project, $1.5 billion over five years, represents a $500 million increase over his current budget. The administration proposes to earmark $273 million for the 2004 fiscal year, but did not offer many specifics yesterday. The funding would support research on fuel cells, vehicle technology and distribution issues.

The magnitude of the goal demands an effort on the scale of the Apollo Moon project, Dorgan said yesterday. "You have to set benchmarks for five, 10 years out."

UNUSUAL PERMANENT MAGNET MOTORS

Thomas Valone

Integrity Research Institute,
www.integrityresearchinstitute.org
1220 L St. NW, Suite 100-232, Washington, DC 20005
Email: iri@erols.com

Abstract

Permanent magnet motors that try to achieve unusual overunity efficiencies with changes in wiring geometry, electronic switching schemes and magnetic configurations often are not successful. There are some designs that should be regarded as conventional and others as promising. Hopefully this article will help the reader to tell the difference before investing or accepting investment. Note: patents can be viewed for free at www.uspto.gov and also http://gb.espacenet.com/espacenet.

Introduction

An article about permanent magnet (PM) motors would not be complete without first reviewing the basic configurations that are present on the market today. Commercial PM motors are necessarily DC motors since their magnets are permanently polarized before assembly. Many PM motors which use brushes are switching to brushless motors that promise less friction and wear. Brushless motors include electronic commutation or step motors. A step motor, often used in the automotive industry, offers more continuous duty torque per unit of volume than any other electric motor but it is often a lower speed motor. The electronic commutation design is applicable to the switched reluctance (SR) motor. The SR motor substitutes soft iron in the place of higher cost permanent magnets for the outer stator and instead has an inner PM rotor. Brushless motors in general produce torque from current in the armature by the application of Faraday's Law. The ideal PM motor has a linear torque vs. speed curve. There are both outer rotor and inner rotor designs that are standard in PM motors.

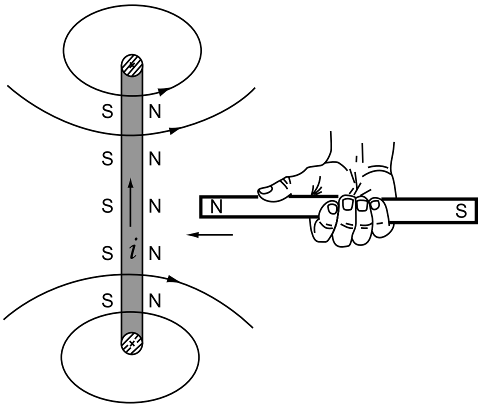

Fig.1

Lenz's Law
Induced B-field opposes motion.

To point out the focus of many of the problems with analyzing motors, the *Motion Control Handbook* (Designfax, May, 1989, p. 33) says that there is a "very important relationship between torque and back emf that is sometimes not understood." This relates to the electromotive force (emf) that is produced by the application of a changing magnetic field (dB/dt). In engineering terms, the "torque constant" (N-m/amp) equals the "back emf constant" (V/radian/sec). In physics, the motor terminal voltage is equal to the back emf minus the IR drop due to internal resistance. (Example: V = 8.3 v, back emf = 7.5 v, IR drop = 0.8 v.) This physics principle, also referred to as **Lenz's Law**, was discovered by Friedrich Lenz in 1834, three years after Faraday invented the homopolar generator. The oppositional nature of Lenz's Law, and its back emf, is built into a physical law called **Faraday's Law**, which is at the root of motor drive. The back emf is the reaction of the changing current in the coil. In other words, the changing magnetic field naturally creates a back emf because they are equivalent.

Therefore, it is recommended that Faraday's Law be carefully reviewed first before proceeding. An article such as "Faraday's Law—Quantitative Experiments" (*Amer. Jour. Phys.*, V. 54, N. 5, May, 1986, p.422) will help convince the valiant new energy experimenter that the change in flux which causes a back electromotive force (emf) is INHERENTLY equal to the back emf. It cannot be avoided or circumvented for excess energy benefit, unless the amount of magnetic flux change per time is also altered. They are two sides of the same coin. The energy into an inductive coil style of motor will naturally equal the energy out. Also referred to as "electrical induction," the changing flux "induces" a back emf.

Switched Reluctance & Field Switching Motors

To explore an alternative method for inducing motion, the "Permanent Magnet Motion Conversion Device" by Ecklin, patent #3,879,622, uses rotatable shutters for alternately shielding the poles of a horseshoe magnet. Repeated again in the Ecklin #4,567,407 "Biased Unitized Motor Alternator with Stationary Armature and Field," the idea of switching the magnetic field with a "flux switch" is common to these types of motors. To illustrate the underlying principle, Ecklin states, "The rotors of most of today's generators are repelled as they approach a stator and are attracted back by the stator as soon as the rotor passes the stator in accordance with Lenz's law. Thus, most rotors face constant nonconservative work forces and therefore, present generators require constant input torque." However, "the steel rotor of the unitized flux switch alternator actually aids the input torque for half of each rotation as the rotor is always attracted and never repelled. This construction makes it possible for some of the current or power fed to the motor windings to magnetically feed through a solid magnetic path to the AC output windings ..." Unfortunately, Ecklin still to this day has not achieved a self-running machine.

Also related is the Richardson patent #4,077,001 which discloses a low-reluctance keeper physically moving in and out of engagement with the ends of a magnet (p.8, line 35). Lastly, the Monroe patent #3,670,189 uses a related principle but accomplishes gating with the passing of rotor poles between permanent magnet stator poles. Monroe's claim 1, seems by its length and detail, to have almost guaranteed its patentability but of course its utility remains questionable.

It seems unlikely that as a closed system the Field Switching Motor can become self-running. In many examples, a small electromagnet will be necessary to help push the keeper into a synchronized rhythm. The Magnetic Wankel from Popular Science (June, 1979) can be compared in a basic manner to this type of invention. Also, the Jaffe patent #3,567,979 can also be compared (see abstract). The Minato patent #5,594,289 is also of a similar type as the Magnetic Wankel and quite intriguing to many people.

It has been found with inventions such as the Newman motor (U.S. Patent Application Serial No. 06/179,474), a nonlinear effect such as an impulse voltage is advantageous for overcoming the Lorentz force conservation effect of Lenz's Law. Also similar is the mechanical analog of the Thornson inertial propulsion device which uses nonlinear impact to transfer momentum along an axis perpendicular to the plane of rotation. A magnetic field contains angular momentum which only becomes apparent under certain circumstances such as Feynman's Disk Paradox, where it is still conserved. The impulse technique may possibly be used to advantage in this Field Switching Motor if the field switching can be done fast enough, with a rapid rise time, but more research is needed.

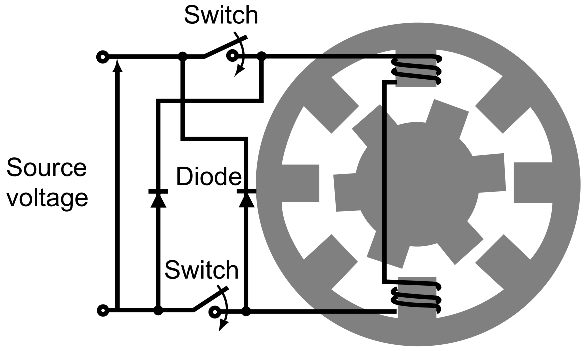

Fig. 2
Switched Reluctance Motor (*IEEE Spectrum* 1/97)

The best Switched Reluctance Motor that also has full accommutation is the Dr. Harold Aspden patent #4,975,608 which optimizes the performance of the coil input and operating above the knee of the B-H curve. Switched reluctance motors are also explained and praised in *IEEE Spectrum* (1/97).

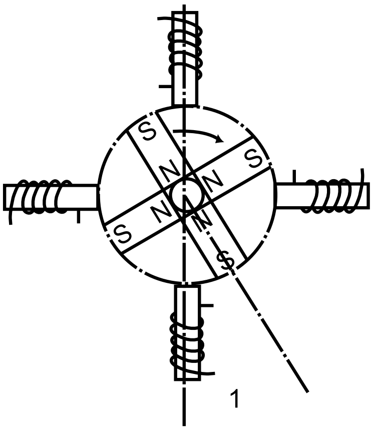

Fig. 3

Adams Motor

The Adams motor has attracted many followers including an endorsement from *Nexus* magazine as the best free energy motor they have seen. The performance of the machine, however, can be fully explained by Faraday's Law. The pulsing of adjacent coils which moves a magnetized rotor is actually following the same configuration as a standard switched reluctance motor. The delay that Adams speaks of in an Internet posting of his motor can be understood from the exponential voltage (L di/dt) of the back emf. The latest addition to this category, which gives credit to the Adams motor, comes from down under with PCT WO 00/28656 awarded to Brits and Christie in May, 2000. The simplicity of this motor is immediately obvious with the switchable coils and permanent magnet on the rotor. The patent also makes it clear that the "input DC current as supplied to the stator coil produces the magnetic repulsion force and is the only outside input to the overall system for total movement...." It is a well-known fact that all motors work on this principle. The key to their design is on p.21 of their patent where the inventors want to "maximize the influence of back EMF which tends to maintain rotation of the rotor/armature in a single direction." All of the motors in this field-switching category try to achieve this effect. Figure 4A of Brits and Christie disclose the voltage sources "VA, VB, and VC." Then, on page 10 it is stated, "At this time current is applied from the power source VA and continues to be applied until the brush 18 is no longer in contact with one of the contacts 14 to 17." There is nothing unusual about this design compared with the more sophisticated attempts listed previously in this section. All of these motors require an electrical power source and none of them are self-running.

When pulsing a coil with the passing of a permanent magnet, a suggestion that would help prove the claim for free energy is not to use battery power for the coil current. Instead, the amazing Weigand wires are recommended (*Pop. Sci.*, May, 1979) that exhibit a huge Barkhausen jump of magnetic domain alignment and a very well-defined pulse shape. Having a coil wrapped around a Weigand wire produces a substantial pulse of several volts with a changing external magnetic field passing a certain threshold. No electrical input power is required for this pulse generator.

Toroidal Motor

As compared to motors on the market today, the unusual design of the toroidal motor is similar to the Langley patent #4,547,713 with a two-pole armature in the center of the toroid. If a single-pole design is chosen, with for example North poles at each end of the armature, this would resemble the radial magnetic field for the armature which the VanGeel patent #5,600,189 uses. The Brown patent #4,438,362 assigned to the Rotron company, utilizes varying magnetization segments for a rotor in a toroidal air gap. The best example of a carousel toroidal motor is the Ewing patent #5,625,241, which also resembles the Langley patent mentioned above. Based upon magnetic repulsion, the Ewing invention uses a microprocessor-controlled carousel, basically to try and take advantage of Lenz's law and get a jump ahead of the back emf. The Ewing invention may be seen in operation, with co-inventor David Porter, in the commercial video, "Free Energy: The Race to Zero Point." Whether it may be more highly efficient than other motors on the market remains an open question. As the patent states, "it is also possible to operate the device as a motor using a pulsed direct-current power source" (col. 7, par. 30). It also contains a programmable logic controller and power control circuit which the inventors thought would send it over the top of 100% efficiency.

Unless a prototype proves to be successful in achieving a torque or force conversion linkage, the internally propelled magnet may be left without a practical application. Commercialization of these types of motors may not be favorable, since many competing designs are currently available on the market, with high flux linkage.

Linear Motors

The area of linear induction motors is well known in the literature. Schaum's Outline Series, *Theory and Problems of Electric Machines and Electromechanics* (McGraw Hill, 1981), explains that these are the same as cutting the rotor and stator of a standard induction motor and laying them out flat. The late Dr. Laithwaite, author of

Motion Without Wheels, was famous for monorail designs for trains in England based on linear induction motors.

The Hartman patent #4,215,330 is an example of one that achieves a linear motor transportation of a steel ball up a magnetized incline of approximately 10 degrees. Another invention in this category is the Johnson patent #5,402,021, which uses permanent arc magnets on a four-wheel cart, exposed to a parallel track of alternating permanent magnets which are in a fixed position. An even more amazing permanent magnet patent is the Johnson #4,877,983 which an eye witness has seen operating at the Johnson home in a closed loop for hours. It is reasonable to assume that a pickup coil could be positioned nearby so that each trip would result in a pulse of electricity to charge a battery. The Hartman patent could also be arranged in such a circular track so that perpetual motion of the first kind can finally be demonstrated.

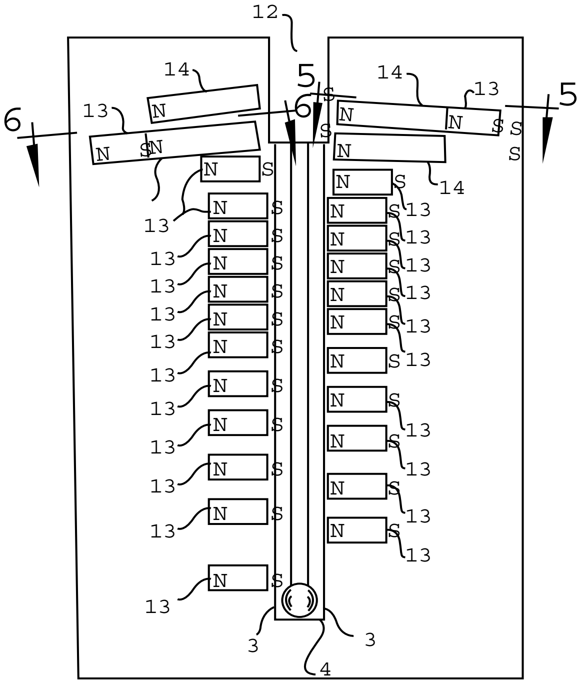

Fig. 4

Hartman patent #4,215,330

The Hartman patent is based upon the same principle as the famous electron spin experiment in physics called the Stern-Gerlach experiment. With an inhomogeneous magnetic field (one that is non-uniform) the force on an object with a magnetic moment is the gradient of the potential energy. Every physics textbook points out that this type of field, that is strong at one end and weak at the other end, will result in a unidirectional force on the magnetic object equal to dB/dx. That is exactly what the Hartman patent possesses (note spacing of magnets). Therefore, the resulting force propelling the ball up a ten degree incline, in the x direction, is in keeping with the laws of physics.

With state-of-the-art magnets, including ambient temperature superconducting magnets which are now finishing the development stage, a demonstration of impressive cargo weight will be shown to be transportable without maintenance electricity costs. Superconducting magnets have the unusual property of retaining the initial magnetized field for years, without the need for periodic energization to restore the initial field strength. Examples of the state of development of the superconducting magnet market can be found in the Ohnishi patent #5,350,958 (lack of cryogenics and lighting system output) as well as the reprinted article from *IEEE Spectrum*, July, 1997 on magnetic levitation.

Static Electromagnetic Angular Momentum

In a provocative experiment with a cylindrical capacitor, Graham and Lahoz (*Nature*, V.285, No.15, May, 1980) have expanded upon the proof published by Einstein and Laub in 1908 that the Lorentz force needs an additional term to preserve action and reaction. The article they cite has been translated and published in my book, *The Homopolar Handbook* (described below). Graham and Lahoz emphasize that there is a “real angular momentum density to r x (E x H)/c^2” and suggest how to see this energy effect in permanent magnets and electrets.

This is encouraging work, with an impressive source of Einstein and also Minkowski for its information. It is possible that it may have a direct application for the homopolar generator as well as the magnetic energy converter mentioned below since both have an axial magnetic field and a radial electric field like the cylindrical capacitor experiment of Graham and Lahoz.

Homopolar Motor

My book, *The Homopolar Handbook* (HH), covers experimental tests and history of the Faraday discovery, including Tesla’s contribution to it. Recently however, there have been new developments into a multi-rotor design of a homopolar generator, similar to the invention of John R. R. Searl.

Recurring interest in the Searl device, as pictured on the cover of *Antigravity*, the biography of Searl by John Thomas, should also center on the homopolar generator (HG). Preliminary analysis reveals that there are actually two separate HG phenomena occurring simultaneously, one which can be called the "revolution" effect (#1) and the second that could be called the "rolling" effect (#2). The first effect can be visualized as magnetized segments of an imaginary solid ring revolving around a common center. As suggested by drawings in HH, p.141-2, there are precedent designs that allow for segmenting an HG rotor.

With this model in mind, the #1 effect can be calculated, for 1 Tesla strength magnets, magnetized axially, adjacent to a single ring 1 meter in diameter, to produce more than 2 volts emf across each roller, (E-field directed radially from outer diameter of rollers to outer diameter of the adjacent ring) with say, 500 RPM. Note that this #1 effect is independent of any rolling of the magnet. The magnetic field in an HPG is tied to space and not to the magnet so rolling will not affect this large scale homopolar generator's Lorentz force effect (HH, p.10).

The #2 effect, located within each roller magnet, is the one noted in *Electric Spacecraft Journal*, Issue 12, 1994, (HH, p.160) where each roller, is a small homopolar generator. This effect is found to be somewhat weaker as it generates electricity from the center of each roller to its periphery. This design is like Tesla's HG (HH, p.81) where a rolling belt is contacting the outer edge of a circular magnet. With rollers in the vicinity of a tenth of a meter in diameter rolling, without slipping, around a 1 meter ring, approximately a half of a volt will be generated. The Searl design of ring magnetic material will normally strengthen the roller's B field.

It is important to realize at this point that the principle of superposition applies to these two effects. The #1 effect is a uniform E field across the diameter of the roller. The #2 effect is a radial effect as stated above (see HH, p.6-8). However, only the emf in the section of a roller between the two contacts, say at the center of the roller and its edge which contacts the ring, will actually cause current flow in any external circuit. This realization means that the effective voltage from the #1 effect will be half of the available emf, or a little more than 1 volt, which is still about double of the #2 effect. Upon applying superposition in the limited region indicated, we also find that the two effects oppose each other and the two emfs must be subtracted. The result of this analysis is that approximately one half of a volt of regulated emf will be present to generate electricity from a single set of rollers and one ring about 1 meter in diameter. As current is drawn, a Ball Bearing Motor effect will also take place (HH, p.54) that actually pushes the rollers along, assuming the roller magnets have a reasonable conductivity *(Thanks to Dr. Paul La Violette for this reminder)*.

In a related work, (*Tech. Phys. Lett.*, V. 26, #12, 2000, p.1105-07), Roshchin and Godin have published experimental results of their one-ring device, called a "Magnetic Energy Converter," with rolling magnets on bearings. It was designed as an improvement to the Searl device. Though my above analysis does not depend upon the ring being made of magnetic material, Roshchin and Godin did so. Their findings are encouraging and detailed enough for researchers to find renewed interest in this type of magnetic motor.

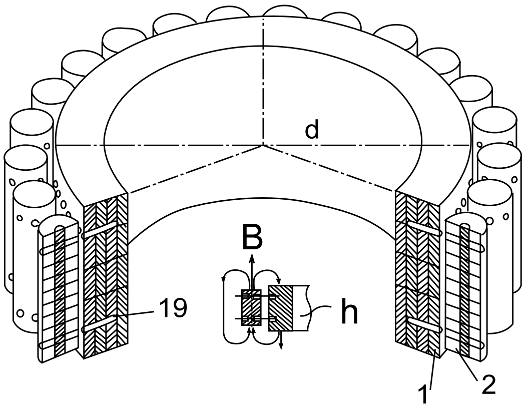

Fig.5

Magnetic Energy Converter in the experiment
by Roshchin and Godin (Russia)

Conclusion

So far, a couple of permanent magnet motors may have achieved perpetual motion, which exceeds 100% efficiency. Of course, conservation of energy concepts have to be considered and the source of the alleged extra energy examined. If permanent magnet field gradients do offer a unidirectional force, as the textbooks predict, then it is about time for its conversion toward useful work. The roller magnet geometry, now called a "magnetic energy converter" is also a unique style of magnetic motor. Exemplified in the Russian patent #2155435 by Roshchin and Godin, it is a magnetic motor-generator that shows potential for excess energy output. Since it relies upon the circulating cylindrical magnet rolling around a ring, the design is actually a generator rather than a motor. However, as they utilize the torque produced by the self-sustained motion of the magnets to run a separate electrical generator, it is working as a motor.

Reprint from Proceedings of Institute for New Energy Conference, 2001

Hydrogen Energetics

Review by Alla Pashova

A Russian writer Ludmila Ulitskaya, who is rather far from energy problems, says: "I'm quite sure that cheap electric energy sources have been already worked out and are hidden in oil king's safes. However such projects won't be taken out of the safes until the last oil drop is combusted. They don't want redistribution of money, world, power, and soon". Experts are sure to agree with the writer.

Now only ideas which can coexist together with the traditional oil power engineering and with a system of world oil distribution can escape the safes. Some expensive scientific toys promising future universal welfare are created to reassure ecologists, viz automobiles equipped with hydrogen engine. According to estimation of Japan representatives of motor car industry, installation of the hydrogen engine into such a car as a classical Sedan now comes to about $700 000. Approximate cost of a hydrogen bus named Kitaro achieves the astronomic amount of $1.25 million.

The hydrogen engine consists of several fuel elements, which are known as fuel cells. Hydrogen under pressure of 1.5-2.7 atmospheres comes to an anode. A cavernous catalyst splits the molecules into electrons and protons. The electrons set the electric engine in motion, and the protons come through a membrane acting to a cathode as an electrolyte. Here the catalyst combines them with the working electrons and with atmosphere oxygen into water molecules. It should be noted that the Europeans and the Americans are extremely charmed by this water streamlet which is exhausted instead of poisoning gases. Moreover, output of the fuel cell is 3-5 times more than output of the petrol engine.

The first hydrogen power station inaugurated in Las Vegas, Nevada, in 2002, (see Fig.1, 2) represents a system of the fuel cells with the proton penetrable membrane as an electrolyte (they are designed and produced by experts of Plug Power Inc.). The station produces hydrogen transforming and cleaning natural gas and generates electric power combining hydrogen stored in reservoirs with atmosphere oxygen. The USA Energy Department and Air Products Inc. together paid $ 10.8 million for this project.

It is necessary to note that the priority branch of the hydrogen engineering is providing new means of transport with fuel, and electric power production is just the second aim. The mentioned station produces electric power in addition to its main function. Nevertheless, if fuel market begins to decrease then the electric power payments will provide the "hydrogen magnates" with stable profit.

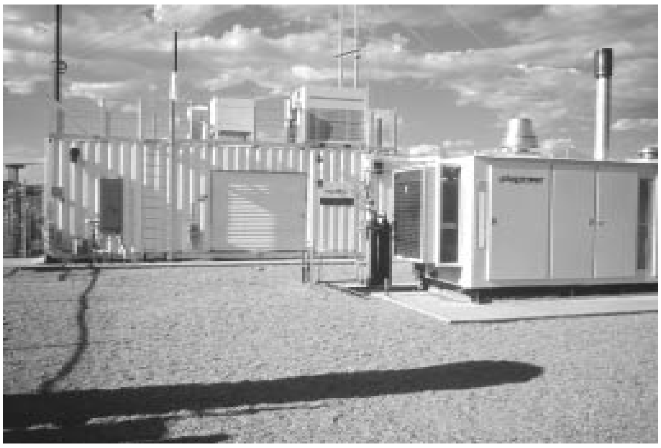

Fig. 1

Hydrogen power station

Nowadays the hydrogen automobiles allow oil kings keep the system of centralized fuel distribution and the infrastructure of service stations which will provide the means of transport with methanol, natural gas with hydrogen blend, and liquefied or gaseous hydrogen. Thus interested companies control the hydrogen engineering and scientific thought.

Moreover, oil can be applied as it is useful for board reforming (conversion), i.e. it can be used as a hydrogen source. The systems based on conversion have been known for a long time but there was need to work at them.

Russian experts improved the conversion system by replacement of the catalyst by plasma. Thus the conversion occurs in gas discharge of HVF appearing in the blend which needs to be converted. According to experts' opinion the Russian system is portable and has high output. In 2001 a group of experts of Russian Centre of Automobile Engine represented experimental prototype basing on Chevrolet Pickup S 10.

The ideal of modern hydrogen engineering is the following: "There should be minimal changes in the system of fuel distribution, non-polluted air, and thin streamlet of distil exhausted". However, evidently, the exhausted water is an ideal source for hydrogen fuel reproduction by means of high effective electrolyzing. If the closed system is worked out we would obtain an energy source powered by continuous chain of electrolyzing and recombination whose efficiency would be more than 100 %.

What are the known ways of hydrogen production? The greatest amount of the gas is produced on the base of catalytic conversion of hydrocarbon with water vapor. Temperature of the process depends on the catalyst composition. It is known that temperature of propane reaction can be decreased to 370° using bauxite as a catalyst. Approximately 95% of by-product carbon monoxide is further used in the reaction with water vapor.

A significant part of the total hydrogen production is given by a method of water gas. The essence of the method lies in reaction of water vapor with coke that is followed by the producing of carbon monoxide with hydrogen mix. The reaction is endothermal, it occurs at 1000°C. The heated coke is processed by the vapor; extracted purified mix of gases contains some hydrogen, some carbon dioxide, and great share of carbon monoxide. Further vapor processing of the carbon monoxide at 370°C increases the hydrogen extraction. The amount of carbon dioxide increases as well however it is easy to be removed by passing the gases mix through a scrubber sprayed by a water contraflow.

AIR PRODUCTS

Fig. 2

There is a well-known iron-vapor method according to which vapor is passed above iron at 500-1000°C. Hydrogen obtained by this method is usually applied to hydrogenise fats and oils. Composition of the iron oxide depends on the process temperature. Iron-3 oxide (Fe_3O_4) prevails at the temperature below 560° C. If the vapor is passed above the iron at the temperature above 560° C then iron-2 oxide (*FeO*) share increases. A slight admixture of carbon monoxide is removed when the heated mix passes above a catalyst. At this time the carbon monoxide is transformed into methane.

In the processes of oxidation and thermal cracking of hydrocarbons at soot production hydrogen is produced as a by-product.

The next effective method is methanol-vapor one. It is an endothermal reaction which is proceeds in common iron reactors at 260° C temperature and at 20 atmospheres pressure.

There should be also mentioned a catalytic reaction of ammonia decomposition at which nitrogen and hydrogen is produced. The reaction can be profitable if it is necessary to produce great amount of hydrogen. The method is useful if the obtained hydrogen is directly applied.

Electrolysis is the oldest method to produce hydrogen. Direct current comes to electrodes, the cathode collects hydrogen, and the anode accumulates oxygen. The technology makes hydrogen to be too expensive energy carrier. There is often used a technology of hot processing of water vapor at 700-900° C accompanied with light petrol and heavy liquid fuel which bleeds oxygen. This method is quite expensive as well.

As it is known, **wasteful power consumption of classical electrolysis** is caused by the fact that it is used to overcome forces of hydrate bound of ions with water molecules and to compensate endothermal effect of water decomposition. Therefore, ions reduction at corresponding electrodes requires more current intensity than in the case of absence of this physical phenomena. Hence production of 1 cubic meter of hydrogen requires 18-21.6 MJ, and total power consumption exceeds 50 MJ (taking into account the electric power production). Therefore hydrogen becomes too expensive.

In 1888 D.A. Lachinov patented a method of non-contact electrolysis. Improvement of this method promises huge amount of cheap hydrogen and appearance of high effective energy source operating as a closed cycle. The main problem of Lachinov's method is a gas layer covering the electrode in some time and impeding the reaction. It was proposed by I. A. Goryachev to prevent the gas layer using pulsating electric field. A.V. Frolov has proposed to increase electrolysis efficiency by means of rotation. Centrifugal force causes more swift detach of gas blebs from electrodes surfaces. Power consumption to overcome frictional force in the construction and to generate potential electric field is insignificant as compared with output. Hence efficiency of the system may exceed 100%.

There is also described the experimental device for production of hydrogen and oxygen from water where modulated voltage (not dc voltage) is applied to the electrodes. Modulation frequency is in connection with proper oscillation frequency of water molecules as well as with spatial structure. Possible efficiency of water electrolysis in spark discharge which removes oxygen admixture should be investigated . Water electrolysis at the radiation by electromagnetic waves of light range needs to be investigated as well.

Ikar Research Centre informs that plasma electrolysis can be the most probable source of cheap hydrogen. In 1987 a group of Russian scientists patented the first plasma-electrolytic reactor. Since the reactor had been worked out at a defense establishment then the patent was restricted and not covered by press. The patent contains a structural scheme of the reactor and some results of cleaning and disinfection of water by plasma. The patent does not represent any information of additional energy generated by plasma or hydrogen production.

In April of 1989 American scientists Pons and Fleishman published their experimental results on additional energy production at common water electrolysis. They claimed cold fusion to be source of this energy. However there is no reliable proof for this phenomenon.

In 1996 Yu. Beklyamishev, one of the co-authors of the first plasma-electrolytic reactor, published his experimental results. They demonstrated that there was additional energy in the plasma-electrolytic process. However he did not explain the energy source.

In 1998 there were published new experimental issues of appearance of additional energy in the plasma-electrolytic process. A group of Russian scientists tested one of the plasma-electrolytic reactors and officially fixed the additional energy. A report of the testing session was published in issue #22 of "Infinite Energy" magazine. In May of 1998 the third edition of a book named "Crisis of Theoretical Physics" was published by Prof. F. M. Kanarev. It contains data of additional energy production at plasma electrolysis of water defining the energy source. Soon Ohmori and Mizuno (Japan) published their results in works of Vancouver conference on cold fusion and in issue #20 of "Infinite Energy". Ohmori and Mizuno observed neutron radiation occurring at a plasma process as well as iron, chrome, nickel, and carbon appearance on a wolfram cathode. That seemed to be a strong evidence of cold nuclear fusion at plasma electrolysis of water. Ohmori and Mizuno have explained neutron radiation as a result of electrons captured by protons. Nevertheless their conclusions raise doubts as cold fusion would have produced much more additional energy than it was fixed.

In 1996 there appeared the first publications where hydrogen atoms fusion instead of nuclear fusion was supposed as the additional energy source at usual electrolysis as well as at plasma source. In Russia first experimental results demonstrating energy consumption decrease for hydrogen production at plasma electrolysis of water were published in 1999 by Prof. F.M. Kanarev. (See F.M. Kanarev's article in this issue of NET).

Fig. 3

Honda FCX-V3 uses hydrogen as a fuel. Hydrogen is fed under high pressure in the engine.

Nevertheless we have to note that high efficiency electrolysis is possible! O. Hvolson in his "Physics" (Berlin, 1923) explains a method of designing of a machine which can produce heat due to positive difference between energy of hydrogen combustion and energy consumed to hydrogen production by means of water electrolysis. In this case we do not consume energy to split the atoms. It occurs due to intermolecular forces at water dissociation by sulphuric acid ions. Energy is consumed only to neutralize charges of existent hydrogen ions and acid residue ions. Hydrogen combustion produces the energy which would have been consumed to split the atoms in the air. **Hence 67.54 Kcal of energy can be produced at 5 Kcal consumed. (F. Lepehin)**

High effective electrolysis application is clear future of energetics. Nowadays Stuart Energy Inc applies water electrolysis as a hydrogen source in a frame of hydrogen-fuel project. Experts of Stuart Energy Inc have designed a refueling unit which produces hydrogen and stores it in a 104-pound reservoir. Thus several auto are provided with fuel and it takes a couple of minutes to refuel.

It is quite simple to predict social and economical consequences of innovation of high effective energy source operating as a closed cycle energy system. Economic activity and private life of people will be independent of cities, of their industry and power engineering. People will move out of the cities and they will use compact and powerful energy sources there. Hence, decentralization of world economics will happen.

Regions of any country become more free from the central authority, the power of which nowadays is based on centralized fuel energetics.

On the Longitudinal Electromagnetic Waves

Alexander V. Frolov

General Director, Faraday Lab Ltd
Phone/fax: 7-812-380-3844
Email: office@faraday.ru

Concerning acoustic waves, we deal with longitudinal oscillations of the medium, since degree of air (medium) compression and rarefication changes along the direction of the wave spreading. Notion of the medium of electromagnetic waves propagation, i.e. notion of "aether", as well as the very possibility of the existence of longitudinal electromagnetic waves is disputable for some scientists. Let's clarify the situation and demonstrate that usual photon is longitudinal wave of energy density, which can be described as Umov-Poynting vector.

The transverse character of electromagnetic waves means that vector of electric field and vector of magnetic field are directed across the wave line. However, these vectors are just our subjective way to describe the process. Notions of electric and magnetic fields can be replaced by the notion of **unified helical field**, since to describe the motion of the point along the helical line there should be concerned the linear transfer (it is detected as electric field) and the process of rotation (it is described with the magnetic field). Thus notions of electrodynamics are just one of the ways to describe reality. Energy density of the given point of space is the real (objective) parameter of the wave. This energy density is described by the known Umov-Poynting vector $\overline{S} = \overline{E} \times \overline{H}$.

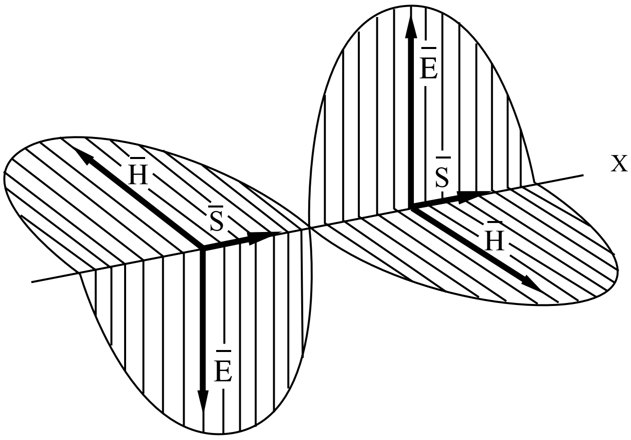

Fig.1

Let us consider electromagnetic wave in classical representation (Fig.1). Then we discover that direction of Umov- Poynting vector coincides with the wave line. The vector is unidirectional and "pulsating", because its quantity changes from zero to some maximal value and then it is reduced to zero, besides it takes place at half of the period of transverse wave. So, it is the explanation of the fact that frequency of longitudinal oscillations is twice as much than frequency of transverse oscillations (Fig. 2). **Previously it was known from the mechanism of energy interchange between longitudinal and transverse plasma oscillations and from the phenomenon of parametrical resonance. However the physical meaning of this phenomenon was not disclosed.**

From the presented here conception it follows that from the physical point of view, electromagnetic waves are **longitudinal oscillations of energy density**. In usual case these oscillations are unidirectional pulsations that determines the ability of the photon to be moving unidirectionally. Practically it is possible to create other types of photons, i.e. types of oscillations of energy density with qualitative new physical properties. It can be achieved by setting of definite functions and superposition of $\overline{E}$ and $\overline{H}$ vectors.

For example, in 1996 a conference "New Ideas in Natural Science" was organized by us in Saint-Petersburg, with the participation of 30 foreign guests and more than 100 Russian scientists. The report made by Academician Ignatyev, Krasnoyarsk, aroused great interest. During the experiments with rotating of crossed vectors $\overline{E}$ and $\overline{H}$ (see Fig. 3) there was created Poiting's vector, which corresponds to propulsion force of 60 N (about 6 kg).

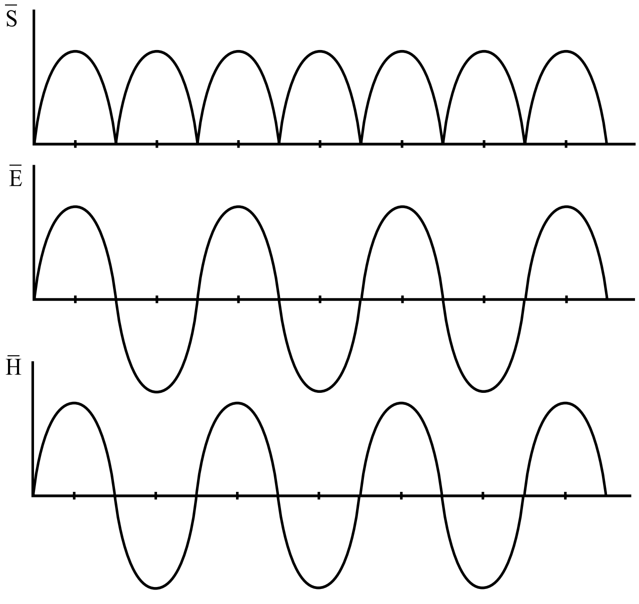

Fig. 2

The experiment is demonstrated on the photo (see Fig. 4), where the diameter of the device is equal to 4 meters and ends of the coils are to toroidal capacitors. This experiment was made in Krasnoyarsk.

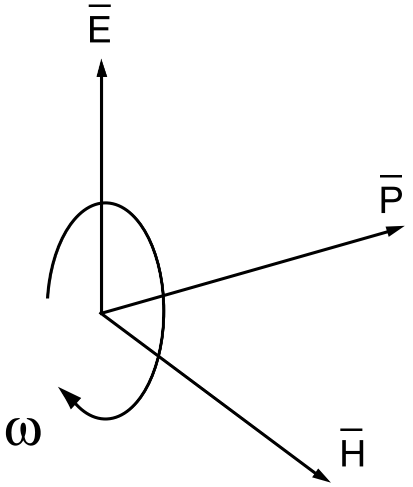

Fig. 3

Since it is impossible to consider vacuum, where some form of energy can exist, as "emptiness" then we can speak about vacuum as about some medium. Faraday and Maxwell, classics of electromagnetic theory, wrote exactly about deformations, stresses and expansion of aether. From this point of view electromagnetic waves are analogues of the waves of longitudinal deformation of elastic medium. More than 60 years ago Nicola Tesla wrote: "I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light." [1]

Nowadays, the development of engineering and the new views upon the phenomena of electromagnetism, impel us to consider physical **vacuum as material medium of the special type**, which has the well-known properties, in particular electrical and magnetic properties. Moreover, this medium has energy, and energy density can be changed in case of propagation of any photon. Thus, as Tesla stated in his article "The True Wireless": "The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history".

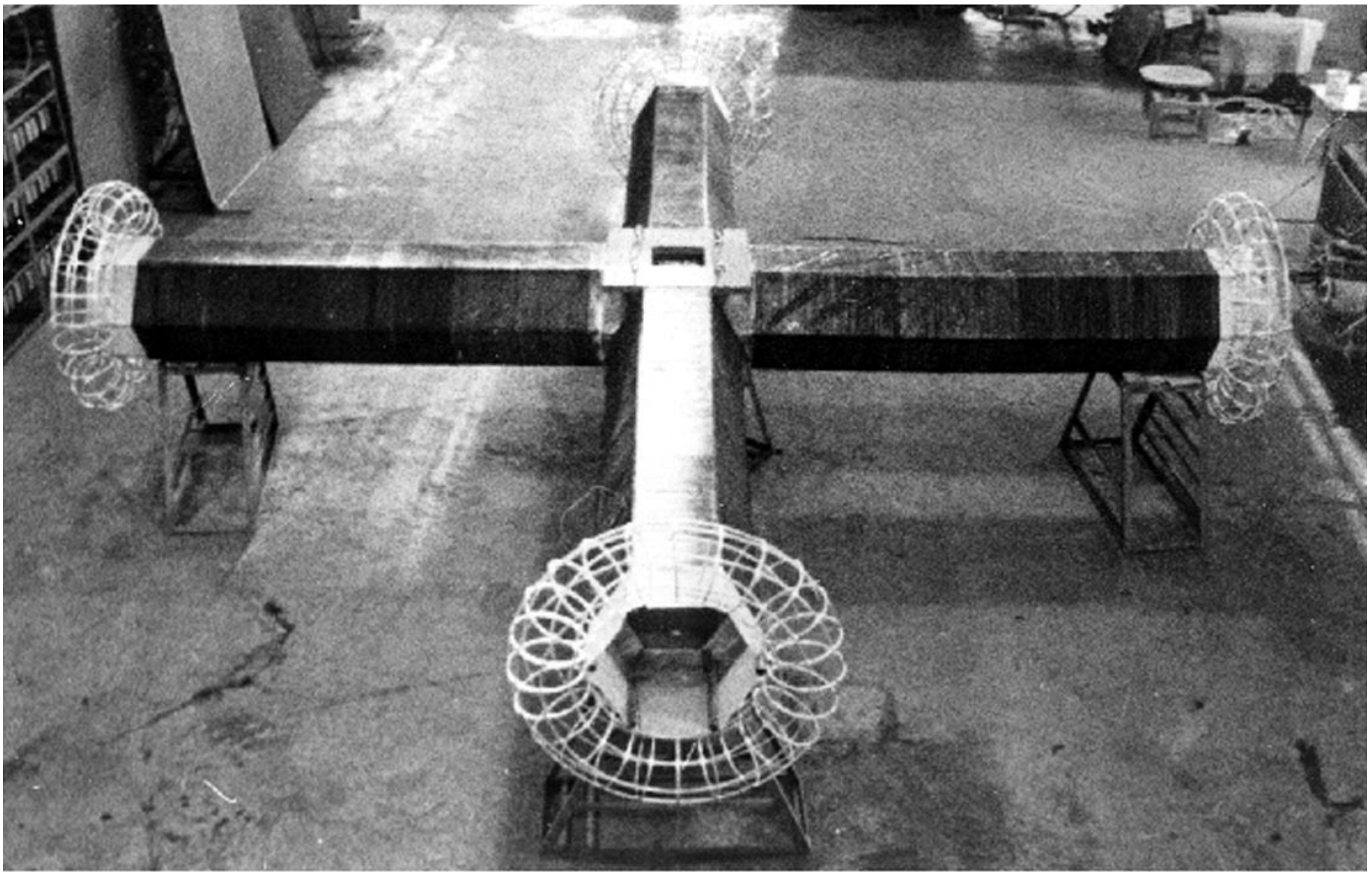

Fig. 4

References

1. "Pioneer Radio Engineer Gives Views on Power", New York Herald Tribune, Sept. 11, 1932.

“OUTSIDE THE BOX” SPACE AND TERRESTRIAL TRANSPORTATION AND ENERGY TECHNOLOGIES FOR THE 21ST CENTURY

Theodore C. Loder

tedloder@aol.com

Institute for the Study of Earth, Oceans and Space
University of New Hampshire
Durham, NH 03824

Abstract

This paper reviews the development of antigravity research in the US and notes how research activity seemed to disappear by the mid 1950s. It then addresses recently reported scientific findings and witness testimonies - that show us that this research and technology is alive and well and very advanced. The revelations of findings in this area will alter dramatically our 20th century view of physics and technology and must be considered in planning for both energy and transportation needs in the 21st century.

Historical Background

Townsend Brown’s Technology of Electrogravitics [1]

In the mid 1920’s Townsend Brown [2] discovered that electric charge and gravitational mass are coupled. He found that when a capacitor is charged to a high voltage, it has a tendency to move toward the positive pole. His findings, which became known as the Biefeld-Brown effect, were opposed by conventional minded physicists of his time.

The Pearl Harbor Demonstration. Around 1953, Brown conducted a demonstration for military top brass. He flew a pair of 3-foot diameter discs around a 50-foot course tethered to a central pole. Energized with 150,000 volts and emitting ions from their leading edge, they attained speeds of several hundred miles per hour. The subject was thereafter classified.

Project Winterhaven. Brown submitted a proposal to the Pentagon for the development of a Mach 3 disc shaped electrogravitic fighter craft. Drawings of its basic design are shown in one of his patents. They are essentially large-scale versions of his tethered test discs.

Review of Issues from the 1950s

In 1956, a British research company, Aviation Studies (International) Ltd. published a classified report on Electrogravitics Systems examining various aspects of gravity control. They summarized the pioneering work of Townsend Brown and then described the use of electrogravitic thrust as follows:

“The essence of electrogravitics thrust is the use of a very strong positive charge on one side of the vehicle and a negative on the other. The core of the motor is a condenser and the ability of the condenser to hold its charge (the K-number) is the yardstick of performance ” [3].

In one of their conclusions, based on Brown’s work, they suggested that: “Electrostatic energy sufficient to produce a Mach 3 fighter is possible with megavolt energies ” [4].

In spite of Brown’s solid research, they later stated that, “One of the difficulties in 1954 and 1955 was to get aviation to take electrogravitics seriously. The name alone was enough to put people off” [5]. It seems that is as true today as it was in the 1950s.

A report by another British company, Gravity Rand, Ltd. in 1956, agrees with this assessment and states: “To assert electrogravitics is nonsense is as unreal as to say it is practically extant. Management should be careful of men in their employ with a closed mind or even partially closed mind on the subject.” [6]

However, a trade press magazine, The Aviation Report, made numerous references to antigravity projects and listed many of the companies pursuing research in this area. Quotes from The Aviation Report listed in the Aviation Studies (International) Ltd. Report [7] are suggestive of what was going on behind the scenes.

In 1954 they predicted that: “... progress has been slow. But indications are now that the Pentagon is ready to sponsor a range of devices to help further knowledge.”... “Tentative targets now being set anticipate that the first disk should be complete before

1960 and it would take the whole of the 'sixties to develop it properly, even though some combat things might be available ten years from now." (Aviation Report, 12 October 1954) [8].

During this time period many of the major defense and technology companies were cited as either having research projects or activities in this new field. For example: "Companies studying the implications of gravitics are said, in a new statement, to include Glenn Martin, Convair, Sperry-Rand, and Sikorsky, Bell, Lear Inc. and Clark Electronics. Other companies who have previously evinced interest include Lockheed, Douglas and Hiller." (Aviation Report, 9 December 1955) [9].

Others of these reports mention: AT&T, General Electric, as well as Curtiss-Wright, Boeing and North American as having groups studying electrogravitics.

During the same time period, the Gravity Rand report notes that: "Already companies are specializing in evolution of particular components of an electogravitics disk." [10]

However, in the area of predictions, the Aviation Report stated the following based on an extrapolation of technology development: "Thus this century will be divided into two parts – almost to the day. The first half belonged to the Wright Brothers who foresaw nearly all the basic issues in which gravity was the bitter foe. In part of the second half, gravity will be the great provider. Electrical energy, rather irrelevant for propulsion in the first half becomes a kind of catalyst to motion in the second half of the century." (Aviation Report, 7 September 1954) [11].

Looking back it is easy to say that they missed the mark. Did they really miss it by a half a century? Reading through these reports it is quite obvious that there was much interest in antigravity among a number of very high profile companies, as well as in the Department of Defense. What happened to this interest and why was it all downplayed during the following four plus decades? After all, T. Brown had shown that there is a demonstrable connection between high voltage fields and gravity. Why has it taken until the 1990s for more than just a few scientists to look at these results and publish on them in the open literature? A review of recent statements by former military personnel and civilians connected to covert projects begins to shed light on research activity in these areas over the last half century. **And it appears that there had been significant breakthroughs during this time period, well shielded from both the scientific and public eye.**

Recent Scientific Developments

In this section we consider developments in the antigravity field since the late 1980s and why the confluence of scientific findings and the testimony of witnesses associated with the military and covert groups indicates that a gravity solution with technological implications has been found.

Although general relativity has not been able to explain Brown's electrogravitic observations, or any other antigravity phenomenon, the recent physics methodology of quantum electrodynamics (QED), appears to offer the theoretical framework to explain electrogravitic coupling. Recent papers by members of the Institute for Advanced Study Alpha Foundation are putting a solid theoretical foundation onto the antigravity effects within the theory of electrodynamics and include papers by Evans [12] and Anastasozki et al [13].

Earlier in a 1994 breakthrough paper, Alcubierre showed that superluminal space travel is, in principle, physically possible and will not violate the tenants of the theory of relativity [14]. Puthoff [15] later analyzed these findings in light of the present SETI (Search for Extraterrestrial Intelligence) paradigms that insist that we could not be visited by extraterrestrial civilizations because of the speed-of-light limitations dictated by the general relativity theory. He suggests that superluminal travel is indeed possible. This leads to reduced-time interstellar travel and the possibility of extraterrestrial visitation, which our limited understanding of physics and scientific arrogance has "forbidden" in some sectors for most of the 20th century.

The second aspect of these physics findings deals with the zero point or vacuum state energy shown by the Casimir effect [16], which predicts that two metal plates close together attract each other due to imbalance in the quantum fluctuations. The implications of this zero point or vacuum state energy are tremendous and are described in several papers by Puthoff [17] starting during the late 1980s. Bearden [18] and colleagues have also written extensively on the theoretical physics of zero point energy and additionally have described various technological means of extracting this energy (for example see the recent paper by Anastasozki et al [19].). A theoretical book on zero point energy (and antigravity) was published by Bearden in 2002 [20]. There is significant evidence that scientists since Tesla have known about this energy, but that its existence and potential use has been discouraged and indeed suppressed over the past half century or more [21].

The coupling of the electrogravitic phenomena observations and the zero point energy findings are leading to a new understanding of both the nature of matter and of gravity. This is just now being discussed in scientific journals (though some evidence suggests that it has been understood for decades within the black project covert community). The question that is being addressed is: what keeps the universe running? Or more specifically, where do electrons get their energy to keep spinning around atoms? As electrons change state they absorb or release energy, and where does it come from? The simplistic answer is that it is coming from the vacuum state. Puthoff [22] describes the process as follows: "I discovered that you can consider the electron as continually radiating away its energy as predicted by classical theory, but **simultaneously absorbing a compensating amount** of energy from the ever-present sea of zero-point energy in which the atom is immersed. An equilibrium between these two processes leads to the correct values for the parameters that define the lowest energy, or ground-state orbit (see "Why atoms don't collapse," NEW SCIENTIST, July 1987). Thus there is a DYNAMIC EQUILIBRIUM in which the zero-point energy stabilizes the electron in a set ground-state orbit. It seems that the very stability of matter itself appears to depend on an underlying sea of electromagnetic zero-point energy."

Furthermore, it appears that it is the spinning of electrons that provides inertia and mass to atoms. These theories, linking electron spin, zero point energy, mass, and inertia have been presented in a number of recent papers, such as those by Haisch [23] and colleagues and provide us with a possible explanation of the Biefield-Brown effect. It appears that an intense voltage field creates an electromagnetic barrier that blocks the atomic structure of an atom from interacting with the zero point field. This slows down the electrons, reducing their gyroscopic effect, and thus reducing atomic mass and inertia, making them easier to move around.

Evidence of Extensive Antigravity Technology

The B-2 Advanced Technology Bomber

In 1993, LaViolette wrote a paper [24] discussing the B-2 bomber and speculating on its probable antigravity propulsion system, based on a solid understanding of electrogravitics, [25] the aircraft's design and the materials used in its manufacture. It appears that the craft is using a sophisticated form of the antigravity principles first described by T. Brown. Support for this thesis came from the Aviation Week and Space Technology (March 9, 1992), which reported that the B-2 bomber electrostatically charges its leading edge and its exhaust stream. Their information had come from a small group of former black project research scientists and engineers suggesting the B-2 utilizes antigravity technology. This information was supported by Bob Oechsler, an ex-NASA mission specialist who had publicly made a similar claim in 1990. These findings support the contention that there have been major developments in the area of antigravity propulsion which are presently being applied in advanced aircraft.

LaViolette later states the obvious that "the commercial airline industry could dramatically benefit with this technology which would not only substantially increase the miles per gallon fuel efficiency of jet airliners, but would also permit high-speed flight that would dramatically cut flight time." [26]

The Hunt for Zero Point [27]

This recent book contains some of the strongest evidence yet for major efforts and success in the field of antigravity technology. The author, Nick Cook, who for the past 15 years has been the Aviation Editor and Aerospace Consultant for Jane's Defense Weekly, spent the last 10 years collecting information for the book. This included archival research on Nazi Germany's antigravity technology and interviews with top officials at NASA, the Pentagon and secret defense installations. He shows that America has cracked the gravity code and classified the information at the highest security levels. **Because antigravity and its allied zero point energy technologies potentially offer the world a future of unlimited, non-polluting energy it has been suppressed because of the "huge economic threat".** His findings support those reported by many of the Disclosure Project witnesses cited above.

Antigravity Technology Demonstrations

Although T. Brown reported many of his findings nearly a half century ago, other experimenters have just recently begun to reproduce his work and report on it in the open literature and on the WWWeb. For example, Davenport [28] published the results of his work in 1995 supporting the findings of T. Brown, while Bahder and Fazi [29] in 2002 described their assessment of the forces associated with an asymmetric capacitor. Transdimensional Technologies [30] in the USA and J. Naudin [31] labs in France have posted on the WWWeb: diagrams, web videos, and data on their versions of antigravity "Lifters" based on an extension of Brown's work. It is a sad commentary on this whole area of research to see that public science is requiring us to demonstrate principles that were demonstrated nearly fifty years ago.

There have also been a number of other demonstrations of "antigravity" phenomena by researchers throughout the world. This includes the work of Brazilian physics professor, Fran De Aquino, and such devices as: the Searl Electrogravity Disc, the Podkletnov Gravity Shield and Project Greenglow, the Zinsser Kineto-baric Field Propulsion and the Woodward Field Thrust Experiments on Piezoelectrics. All of these are described in more detail by Greer and Loder. [32]

Implications of This Research

Antigravity and zero point energy research and their applications are finally being addressed by some of the open scientific community. This means there will have to be a rewriting of textbooks in this area so our new generation of students can apply this "new knowledge." Its application will lead to major breakthroughs in transportation technologies both earthside and in outer space. The implications are that we have the potential for human exploration of our solar system and beyond, if we have the will, within our lifetimes. It also means that the majority of 20th century space technology will be obsolete and in fact may already be so.

The zero point or vacuum state energy source is seen as a totally non-polluting energy source, which has the potential to replace all the fossil fuels on this planet. It also will provide the energy needed for long range space flights. This means that fuel cells and solar cells in common use today for space flight energy applications will only be needed until we transition to these new energy technologies.

Based on an analysis of trends in antigravity research over the last half-century and the information provided by numerous witnesses, it appears that there is both good and bad news. The good news is that it appears that we (at least covert projects) have already developed the theories of antigravity, and additionally have developed working spacecraft based on these principles. The bad news is that these technologies have been developed for at least several decades, at the public's expense and that human kind has been deprived of these technologies, while continuing to waste energy using less efficient and pollution enhancing technologies.

Supporting this contention is the following quote from Ben Rich, former head of the Lockheed Skunkworks. Just prior to his death, he stated to a small group after a lecture [33] that: "We already have the means to travel among the stars, but these technologies are locked up in black projects and it would take an act of God to ever get them out to benefit humanity..." He further went on to say that, 'anything you can imagine we already know how to do.' Strong words from a knowledgeable deep insider and words that support what a number of the witnesses stated as well.

As the reality of this knowledge begins to be understood, there will be an outcry among space scientists not on the inside for release of these technologies to allow all of us to explore space. There will be major changes in the way that NASA does its business, though predicting these changes is difficult.

Not only has space exploration in the public sector suffered, but our planet's environment has suffered as well. Thus as this knowledge begins to sink in there will be an outcry among all concerned citizens on this planet for release of these technologies to allow all of us to reduce and ultimately eliminate global warming and environmental pollution that so threatens our way of life. These technologies will not only affect space travel technologies, but will also have a profound effect on transportation and energy production on the earth's surface.

In conclusion, we might consider the observation made by Halton Arp [34]: "We are certainly not at the end of science. Most probably we are just at the beginning!"

Acknowledgements

I thank the following people for many discussions while preparing and writing this paper including: S. Greer, A. Craddock, T. Bearden, P. LaViolette, M. McCandlish, D. Hamilton, T. Valone, E. Mallove, T. Loder, C. Loder, S. Patel and many of the courageous Disclosure Project witnesses.

References Cited

1. Quoted from: LaViolette, P. A. 2000. Moving Beyond the First Law and Advanced Field Propulsion Technologies. In T. Loder (ed.), "Outside-the-Box" Technologies, Their Critical Role Concerning Environ-mental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
2. Brown, T. T. 1929. How I control gravity. Science and Information Magazine, Aug. 1929. Reprinted in Psychic Observer 37 (1): 66-67.
3. Aviation Studies (International) Ltd. 1956. Electro-gravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 14. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
4. Ibid. p. 27.
5. Ibid. p. 19.
6. Gravity Rand Ltd. 1956. The Gravitics Situation. p. 54. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
7. Aviation Studies (International) Ltd. 1956. Electrogravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 11. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
8. Ibid. p. 34.
9. Ibid p. 41.
10. Gravity Rand Ltd. 1956. The Gravitics Situation. p. 47. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
11. Aviation Studies (International) Ltd. 1956. Electrogravitics Systems: An examination of electrostatic motion, dynamic counterbary and barycentric control. p. 32. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.
12. Evans, M. W. 2002. The link between the Sachs and O(3) theories of electrodynamics. In Evans, M.W. (ed.), Modern Nonlinear Physics, Pun 2. 2nd ed., Advances in Chemical

Physics 19:469-494.
13. Anastasovski, P.K., T.E. Bearden, C. Ciubotariu, W.T. Coffey, L.B. Crowell, G.J. Evans, M.W. Evans, R. Flower, A. Labounsky, B. Lehnert, M. Mészáros, P.R. Molnár, S. Roy, and J.-P. Vigier. 2001. Anti gravity effects in the Sachs theory of electrodynamics Foundations of Physics Letters. 14(6):601-605
14. Alcubierre, M. 1994. The Warp Drive: Hyper-fast travel within general relativity. Classical and Quantum Gravity, 11, L73.
15. Puthoff, H. E. 1996. SETI, The Velocity-of-Light Limitation, and the Alcubierre Warp Drive: An Integrating Overview, Physics Essays 9:156.
16. Lamoreaux, S.K. 1997. Demonstration of the Casmir force in the 0.6 to 6 μm range. Physics Review Letters. 78:5.
17. Puthoff, H. 1989. Gravity as a Zero-Point Fluctuation Force." Phys. Rev A., 39(5):2333-2342. Puthoff, H. 1989. Source of Electromagnetic Zero-Point Energy." Phys. Rev A, 40(9):4597-4862.
18. See the Tom Bearden web site for an extensive listing and copies of his papers at: www.cheniere.org.
19. Anastasovski, P.K., T.E. Bearden, C. Ciubotariu, W.T. Coffey, L.B. Crowell, G.J. Evans, M.W. Evans, R. Flower, A. Labounsky, B. Lehnert, M. Mészáros, P.R. Molnár, J.K. Moscicki, S. Roy, and J.P. Vigier. 2001. Explanation of the motionless electromagnetic generator with 0(3) Electrodynamics. Foundations of Physics Letters, 14(1):87-93.
20. Bearden, T. 2002. Energy from the Vacuum: Concepts and Principles. Cheniere Press, Santa Barbara, CA.
21. Valone, T. 2000. The Right Time to Develop Future Energy Technologies. in T. Loder (ed.). "Outside-the-Box" Technologies, Their Critical Role Concerning Environmental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
22. Puthoff, H. 1990. Everything for Nothing. New Scientist, 28 July 1990, pp. 52-55.
23. Haisch, B., Rueda, A. and Puthoff, H. 1994. Beyond E =mc2; A First Glimpse of a Postmodern Physics, in which Mass, Inertia and Gravity Arise from Underlying Electromagnetic Processes. The Sciences, 34:26. Haisch, B., Rueda, A., and Puthoff, H. 1997. Physics of the Zero-Point Field: Implications for Inertia, Gravitation and Mass. Speculations in Science and Technology, 20:9. Haisch, B. and Rueda, A. 1998. An Electromagnetic Basis for Inertia and Gravitation: What Are the Implications for 21st Century Physics and Technology? in El-Genk, M.S. (ed.), Space Technology and Applications International Forum-1998, DOE CNF-980103, CP420, p. 1443. Haisch, B. and Rueda, A. 1999. The Zero-Point Field and the NASA Challenge to Create the Space Drive. Proc. NASA Breakthrough Propulsion Physics Workshop, NASA/CP-1999-208694, p. 55.
24. LaViolette, P. 1993. The U.S. Antigravity Squadron. p. 82-101. In Valone, T. (ed.), 1994. Electrogravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005.p.82-101.
25. LaViolette, P. A. 1992. Electrogravitics: Back to the future. Electric Spacecraft, Issue 4, pp. 23-28. LaViolette, P. A. 1993. A theory of electrogravitics. Electric Spacecraft, Issue 8, pp. 33-36
26. LaViolette, P. A. 2000. Moving Beyond the First Law and Advanced Field Propulsion Technologies. in T. Loder (ed.). "Outside-the-Box" Technologies, Their Critical Role Concerning Environmental Trends, and the Unnecessary Energy Crisis. Report prepared for The U.S. Senate Environment and Public Works Comm. www.senate.gov/~epw/loder.htm.
27. Cook, N. 2001. The Hunt for Zero Point: Inside the World of Antigravity. Broadway Books, NY. 256pp.
28. Deavenport, L. 1995. "T.T. Brown Experiment replicated. Electric Spacecraft Journal. Issue 16. Oct. 1995. (Reprinted in: Valone, T. (ed.), 1994. Electro-gravitics Systems: Reports on a new propulsion methodology. Integrity Research Institute, Washington, DC 20005)
29. Bahder, T.B. and Fazi, C., "Force on an Asymmetric Capacitor," Army Research Laboratory, v51 ARL-TR.nb. Avail. at: http://xxx.lanl.gov/abs/physics/02110. 2002.
30. Transdimensional Technologies, 906-E Bob Wallace Ave., Huntsville, AL 35801.
31. http://jnaudin.free.fr
32. Greer, S.M. and T.C. Loder III. 2001. Disclosure Project Briefing Document, 492 pp. Available on CD from: The Disclosure Project, P.O. Box 2365, Charlottesville, VA 22902. Also available from: www.disclosureproject.org.
33. Lecture given at UCLA Engineering Department on March 23, 1993
34. Arp, H. 1998. Seeing Red: Redshifts, Cosmology and Academic Science. Montreal, Aperion. (p. 249).

Congress "The Time Machine"

Faraday Labs Ltd invites to participate in scientific congress devoted to experiments on changing of space-time properties. It is planned April 12, 2003, Moscow. Main topics of the congress are time and gravitation in the context of eatherodynamics, experiments and applied aspects of these technologies. The main report is "Method to Control Temporal Parameters of Physical Processes" by Alexander V. Frolov.

Organizing committee: Vadim A. Chernobrov (KOSMOPOISK research center) and Alexander V. Frolov (Faraday Labs Ltd). The registration fees are not required.

Please, contact us http://www.faraday.ru or **email** congress@faraday.ru **Phone/fax 7-812-380-3844**

Please send this pre-registration form by post: P.O. Box 37, St. Petersburg, Russia 193024 **or e-mail:** congress@faraday.ru

***PRE-REGISTRATION FORM* for participation in the congress "The Time Machine"**

Last name:
First name:
Title/ Function:
Company/ Organization:
Address:
Postcode:
City:
Country:
Tel.:
Fax:
e-mail:

☐ I would like to submit an oral presentation (report). (Please, enclose one-page abstract).

☐ I intend to participate as a guest

☐ I'm interested in further information.

Date Signature

COMMERCIALISING the "SEARL EFFECT" for ENERGY and PROPULSION SYSTEMS

Review from web site http://www.sisrc.com/
Email address: admin@sisrc.com

The "Searl Effect", discovered by J.J.R. Searl, is a method of extracting energy. Some current expressions for the source of this energy being The Space Fabric, Quantum Space Field, and Zero Point Energy.

SISRC Ltd has been formed to develop and license Searl Effect Technology (SET) on a global basis.

THE COMPANY

SISRC Ltd is the company formed to administer the implementation of the Searl Effect Technology, (SET) invented by J. R. R. Searl. SISRC Ltd designs, develops and implements Searl Effect Technology as it is initiated in various applications and territories. SISRC Ltd is the administrative hub of the group and will continue to be based in the UK. SISRC Ltd will grant production and marketing licenses for Searl Effect Technology applications to different companies in specific territories. Currently proposed related companies include:

* SISRC -Germany, SISRC -Iberia, SISRC - Sweden, SISRC - Australia, SISRC New Zealand.
* SISRC -AV (Audio Visual) provides computer graphics presentations of the SET.

BACKGROUND

The Searl Effect Generator (SEG) technology, as applied to the commercial market, had been previously developed to the point where a few prototype SEG generators were made, and used for electricity generation and motion. Commercial interest at that time focused on the SEG's transport potential and, under commercial pressure to deliver a fully functioning system, the original generators were used and lost in a series of vehicular propulsion experiments and demonstrations. Funding was insufficient to continue with the manufacture of the required larger-scale pressurized cabin machines - resulting in the termination of the project at that time.

Although all the operating principles, precise proportions and weights of the required materials are known for three of the four required operating materials, the precise data of the original magnetic layer is uncertain. The objective of the current R&D programme is to re-establish the original magnetic layer using modern and more efficient materials.

Originally, the layered materials were constructed and magnetized by the now-defunct Midlands Electricity Board under the direction of John Searl (see colored photo of the experimental craft construction on the cover page).

Modern magnetic materials have advanced considerably, and old ones discontinued, so a series of tests need to be conducted to establish the optimum materials and processes. These tests need to comply with the working criteria required and must lead to a cost-effective manufacturing process.

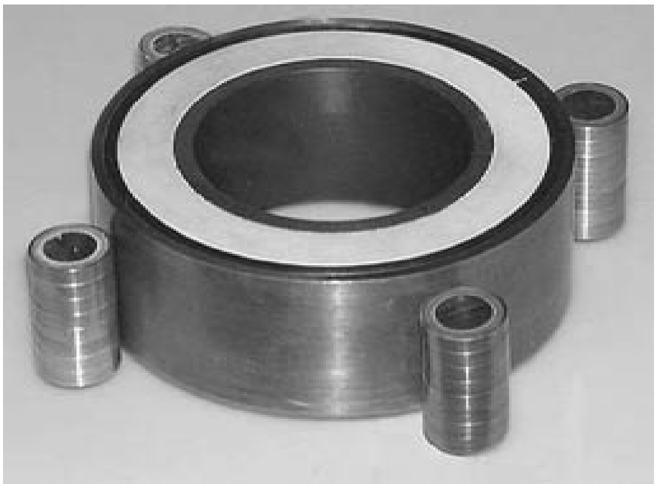

Fig.1
Experimental ring and rollers

In recent times SISRC has been re-establishing the original research. Due to the very limited funding that has been available, only a partially functioning demonstration prototype of the SEG principles has been possible. This prototype consists of the innermost of the three composite rings required and several rollers.

We would like to hear from anyone who worked on or was involved with Searl Technology prior to 1983. Also anyone who has any old technical data, photographs or films relating to the technology. Such information could greatly assist this technology to reach the marketplace for the benefit of all and would be dealt with in the strictest of confidence.

TECHNICAL DESCRIPTION

Physically the Searl Effect Generator (SEG) consists of three concentric rings each made of a composite of four different materials which are also concentrically attached to each other. The three rings are fixed to a base. Surrounding each of the rings, and free to rotate around them, are rollers - typically 10 on the first ring, 25 on the next and 35 on the outer ring. Surrounding the rollers on the outer ring are coils which are connected in various configurations to supply either AC or DC current at a variety of voltages.

Multiple magnetic poles are imprinted on the rings and rollers to form frictionless magnetic bearings. These also arrange the static charge into opposing charge clusters which cause the rollers to rotate around the circumference of the ring (as shown in Fig. 2).

Some expressions currently in use to describe the source of the energy for the SEG are The Space Fabric, Quantum Energy field of Space or Zero Point Energy. This is an unlimited and constant source of energy which can be made to flow when the correctly proportioned masses concerned are stimulated by the correct frequencies creating an 'open system'.

The idea of utilising this source of energy is currently the subject of various devices and experiments such as the 'Lamb shift', 'Casimir Effect' and the work of the Russian Nobel Prize winner Ilya Prigogine. However these devices and experiments tend to only prove the existence of the energy and not a method to create a coherent, ordered flow to produce useful power.

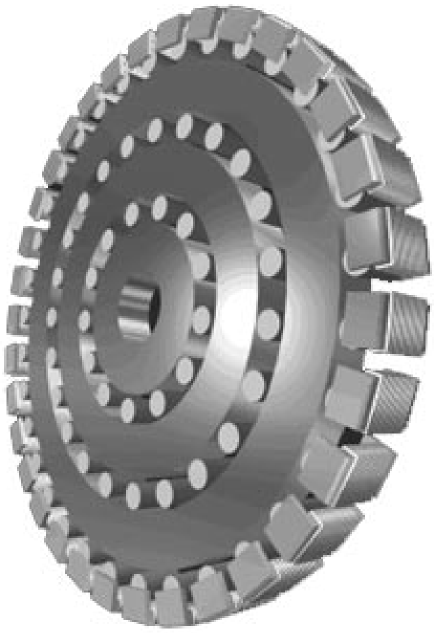

Fig. 2

15 KWt Searl Effect Generator (SEG)

In contrast, John Searl has discovered that, in order to create a steady and stable flow, all the masses of the device (and the stimulating frequencies) must conform to precise values determined mathematically by the 'Law of the Squares'. A machine constructed to these principles produces a stable and useful power output.

DEFINITION OF A SURFACE OPERATING S.E.G. (J. Searl) 15.09.00

A Linear motor operating on a magnetic bearing with the characteristics of an auto-transformer. The S.E.G. is defined as a device, which is constructed from 2,124 component parts, which make up 3 plates and 66 roller sets (see figures on the cover page). 12 of the 2,124 components create the 3 plates, which act as reaction components. 2,112 components are employed to create 66 roller sets, which act as active components. Each roller set consists of 8 segments constructed from 32 components. Each segment cannot be less than 34 grams in weight, as the law of the squares defines that value as the lowest value suitable for the reasonable generation of electricity. Therefore the smallest roller set must weigh 272 grams, making a grand total of 2,244 grams per roller set for the smallest surface bound SEG.

Tolerance should be about 0.05 gram per roller set. The larger the error the greater is the loss of power within the unit. Therefore the total tolerance for the smallest SEG roller set must not be greater than 3.30 grams - above this factor the SEG will not function. **For the best results, the tolerance across the total 66 roller sets should not exceed 0.05 grams!**

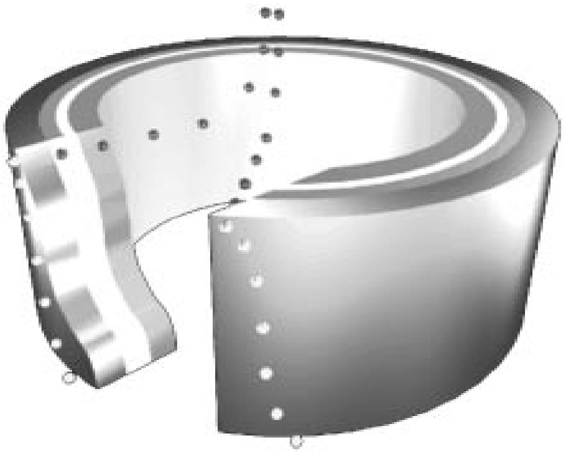

Fig.3

The S.E.G. is a step up rotary transformer. In fact it's a 'three rotary transformers' operational system, within a unit. The inner transformer output is fed in to the input of the second plate which increases the output of the second unit, which output feeds as an input to the final unit, thereby increasing its finaloutput to a very high voltage. Therefore the S.E.G. is just a prime mover for the production of clean electrical energy. The output windings must be designed to step down this output to 240 volts at 15KWt's.

During 1968 it being upon this issue, as development progress both here and in the USA, new findings are being created, and updates will be added to this page, as they are certified. When the SEG rollers are brought into close proximity to the SEG Ring, the Searl Effect resonant magnetic field causes negative ions and electrons to be drawn into and accelerated through the machine. This process is assisted by the highly electron-attracting rare earth metal Neodymium.

The unique mechanical and material arrangement of the SEG pulses the neodymium to continually release and replace the surplus electrons to provide electrical or mechanical power, or both.

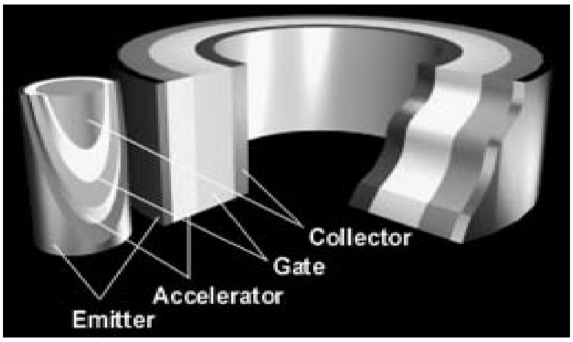

Fig.4

DEVELOPMENT

Prior to the production of complete 15KWt Searl Effect Generators (SEGs), a minimum of two demonstration Ring & Roller prototypes are to be produced.

The "Ring" is typical of the innermost of the three rings required to make a complete SEG. The 69-roller sets in a complete three ring SEG are identical. Six roller sets will be made for the prototype inner rings so that the Searl Effect can be demonstrated.

The production of a self-running Ring & Roller prototype will establish the precise sequence and parameters of electrical, magnetising and material processes required for the SEG and will demonstrate electrical power generation. An operating single inner ring with 1-8 rollers will also determine if a particular material will function correctly and at what level of power output.

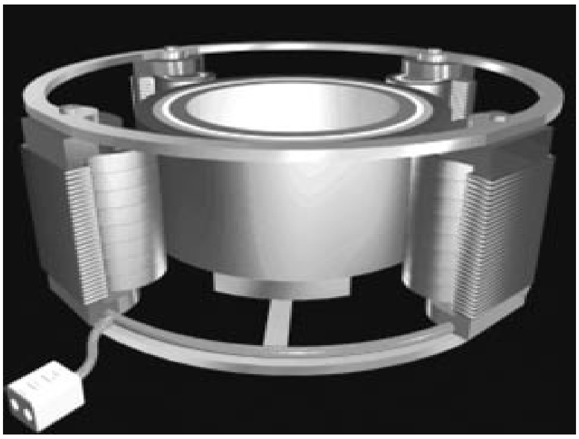

Fig. 5
Demonstration ring and rollers

The five phases of the development programme that commenced in October 1999 have been completed and the rollers successfully magnetised with 'Searl Effect' magnetic fields. A sixth phase is now required to apply the technology developed in Phase five to the "Ring". This sixth Phase resulting in a demonstration Ring & Roller prototype has duration of 6 months.

NEWS REVIEW

Scientific Breakthrough Liberates Energy Users from Fossil Fuel Dependence

Information from
http://www.genesisworldenergy.org/genesis_world_energy.htm

Story originally published by Diana Echeverria, USA

Director of Public Relations
Guy Rome & Associates, Inc.
208-345-4143 208-602-0325 (cell)

Technology breakthrough harnesses energy from the molecular structure of water

BOISE, ID - Dec. 5, 2002 - Genesis World Energy, a privately funded consortium created by a group of military and space program research and development specialists, today unveiled a scientific breakthrough that allows consumers to easily access the energy contained within the hydrogen and oxygen molecular structure of ordinary water. This scientific breakthrough provides a limitless, low cost and environmentally clean source of energy that can be implemented with minimal cost and effort. The viability of using water as an energy source, previously a theoretical concept, is now a reality. "Water has always been the source of life on this planet, now it will also transform the way we create energy" said Charles Shaw, corporate counsel and spokesperson for Genesis World Energy. "The implications for worldwide energy generation and consumption are nothing less than staggering."

The Edison Device

The first application of this technology is represented in the "Edison Device", a self-contained, self-sustaining energy generation unit. Roughly the size of an outdoor air conditioning system, the power source can be quickly and easily installed in any home or business to provide virtually unlimited energy from any available water source. The Edison Device utilizes the existing electrical wiring and natural gas plumbing in a home or business to replace the energy provided by utility companies. The home version of the Edison Device produces approximately 30 kWtt of combined gas and electrical energy per day. By comparison, the typical home uses between five to six kilowatts per day. The commercial model is capable of producing 100 kilowatts of energy per day. The energy generation portion of the devices has no moving parts. In fact, the only "mechanical" aspects of the equipment are small circulation pumps and micro-valves, making the Edison Device both silent and virtually maintenance-free. A minimum amount of water is used over an estimated 20+ years of service life.

Interesting Publications of the Last Year

Discovery by a Group of Scientists Headed by Valerian Sobolev

Some time ago mass media announced that a sensational discovery was made by a group of Russian scientists headed by Valerian Sobolev. This discovery aroused ambiguous attitude of physic community. We have written about this fundamental and, to a certain extent, revolution discovery. Let us revise the fact that Sobolev claimed 7 scientific discoveries which were made, i.e. the process of depletion (a special electrochemical process), magnetic discharge, a new energy source, a method of low temperature plasma generation and superconductor. A laboratory system of the experimentally disclosed process as well as its industrial prototype is easy to be realized in practice for creation of new energy sources and wide-ranging production of ultrastrong materials. In many cases these materials can replace existing constructional materials.

The materials produced as a result of the process are multielement chemical compounds which are new states of matter. This new state of matter has **a time-changing magnetic field that can be EMF source in a coil of a generator.** Ordered structures of matters which are in the new modified state are nothing but a magnet charge. Materials containing the magnet charge and representing a continuous matters are **new energy sources.** Due to the new state matters become able to produce electric power as well as to generate low temperature plasma. It is easy to be realized in the process of common technological procedures. That promises designing of propulsors for "unsupported" transportation systems in near future.

Basing on realization of the disclosed process and using new materials the group of Russian Scientists headed by Sobolev together with American businessmen has created superfine and flexible glass for packing. The glass was called as "strong glass". It seems to demonstrate higher pressure stability than steel. A method of production of these materials was patented in 1999. In the patent the scientists of Valerian Sobolev's group are represented as authors of this new method, and the owner of the patent is Dynelec Corp (Columbus, Ohio, USA). You can read about the patent at the Website of http://patft.uspto.gov/netahtml/srchnum.htm (patent #5,964,913, October 12, 1999).

Sobolev's group has appeal to the Russian government for sponsorship; however only foreign investors have provided funds for new energy sources developing. According to Russian Information Agency "News", Sobolev's group has signed a $168 million contract with one of Canadian companies for developing industrial production of energy sources. Manufacturing of such energy sources **can change energy supply system all over the world**. According to the contract the Canadian company is going to finance building of at least two plants. One of the plants is going to be built in Russia, and the other one is planning to be established in Canada. Each plant is supposed to produce 70 thousand of the self-running energy sources a year. Power output of the sources will come to 3-10 kWt. In other words, in the nearest future every person will be able to buy such an energy source and to apply it in household.

Single-Wire Electric Power System

(see photo on the cover page)

Experimental Results of Russian Scientists

Scientists of the All-Russia scientific research institute of electrification of agriculture (VIESH) academician D.S. Strebkov, engineer S.V. Avramenko, dr., A.I. Nekrasov, post-graduate student O.A. Roshchin developed a new method and the equipment for transmission of electric energy on a single-wire line using resonant idle operation mode and reactive capacitive currents for transmission of active electric power. Tesla transformers and frequency converter were used at the experiments.

First there was created the experimental sample of single-wire electric power system (SWEPS) with **10 kW** electric capacity and 3000 V voltage. As a result, obtained SWEPS electric parameters a hundred times exceeded parameters of a usual two-wire or three-wire alternative and direct current line.

In the end of 2002 there was made an attempt to increase transmitted capacity by using of more powerful condensers, which were installed in resonant circuit. There was developed the electric technique of **20kW** single-wire electric power system (SWEPS) with two Tesla transformers. As a result parameters of a low-voltage winding of the step-down transformer were changed. They **exceeded parameters of a usual two-phase or three-phase ac line or dc line in two hundred times.** At room temperature effective specific transmitted electric power was 4 MW/mm^2 and specific current density was 600 A/mm^2. These parameters could be achieved for existing methods of electric energy only at use of special materials in a mode of low temperature superconductivity. The measurements which were made at wire with diameter of 1mm, 100 microns and 80 microns, demonstrated that parameters did not depend on diameter of a wire. There was also experimentally confirmed the property of a single-wire line to transfer active power without essential losses on line resistance. Irradiation losses at low frequency 3,4 kHz were small. In single-wire power system the 80 microns wire plays a role of directing system for an electromagnetic energy flow, which runs from the generator to the load.

"Lifter" Project

We present update information on the experiments with Lifters or asymmetrical capacitors, which use High Voltage to produce a thrust. ***By Jean-Louis Naudin it was demonstrated with "Maximus" experiment, that a Lifter can be scaled up and also that such a device is able to lift up to 60 g of payload (total weight: 194 g). According to Naudin, it is now possible to build a craft which will use the Biefeld-Brown effect to fly silently and without moving parts only powered by electrical energy.*** *It was also declared that on January 8th, 2003, there was successfully done two historical flights with a mouse as a test pilot of the Lifter "Maximus"* ***(propellantless electrokinetic craft)****. Below there is an analysis of electrogravitation experiments made by Jean-Louis Naudin and Tim Ventura. (See color photos on the cover page).*

Review of Electrogravitation Experiments made by Jean-Louis Naudin and Tim Ventura

Tim Ventura

Tim Ventura Website: http://www.americanantigravity.com
Email: TVentura@seattle.telecomsys.com, tventura6@atbi.com

There is a spreadsheet containing a jpeg-snapshot of data that I have assembled based upon Jean-Louis Naudin's published results (JLN Labs Website: http://jnaudin.free.fr. and http://www.jlnlabs.org). In the graphs that I've created, it seems to indicate that current plays more of a role in propulsion than voltage does — i.e.: for a given amount of power in watts, raising the current and decreasing the voltage seems to create a higher level of thrust.

In light of this, I have been increasing the thrust of my own Lifters by creating an electrical-bypass of the "load-resistor" on my power-supply's high-voltage output. This has an interesting result:

I use a current-driven power-supply, which means that if no load is connected to it the voltage will build up on the high-voltage output until arcing occurs. In my case the voltage will build up to around 65kV, at which point the power-supply will automatically shut down. Conversely, if a very light load (such as a short circuit) occurs, the opposite effect happens and the power-supply will deliver much higher current at a much lower voltage until the supply is overdrawn and once again shuts down.

Normally the output load-resistor constrains the current, and in doing so it maintains the voltage on the high-voltage output at the nominal 50kV output level. However, bypassing the output load-resistor allows the power-supply to deliver voltage and current that are based almost entirely on the type of load connected to it.

When I connect a normal Lifter to the bypassed power-supply, the voltage will build across the air-gap until ions begin to flow in a conduction-current across the air-gap. Although this begins to happen at approximately 15kV, it seems to have some type of "peak efficiency' at approximately 22.5 kV. The amount of current climbs to approximately 11mA at this voltage.

This method of experimental setup seems to allow the Lifter to "find its own sweet-spot" for operation, instead of using the output-resistor to "force" a specific voltage on it for operation. This appears to maximize the thrust output during operation.

Lifter Efficiency Spreadsheet

Description	Lift capacity (grams)	Length (cm)	Efficiency (g/m)	Lifter Weight (grams)	Voltage (kilovolts)	Corona Air-Gap (centimeters)	Current (mA)	Power (watts)
Lifter 1	3.3	60	5.5	2.3	41.9	3	0.57	23.9
Lifter 2	9.6	180	5.3	6.6	43.35	3	1.12	48.5
Lifter 3	20	360	5.6	16	27.5	3	2.53	69.5
Lifter 4	36	720	5	32	44	3	2.01	132.9
3-Stage Lifter 3	54	1080	5	24	30	4.5	8.06	240
Coliseum Lifter	90	2160	4.16	50	31.769	4.5	8.06	254.15

New Electric Fire Technology

Valery D. Dudyshev

Doctor of Technical Science, Professor,
Corresponding Member of Russian Ecological Academy
ecolog@samaramail.ru
http://www.intrasoft.ru/nizot

Abstract

The article proves importance and prospects of development of combustion technologies. In particular there is considered application of a new electric fire technology for environmentally appropriate combustion of any matters and gases. The technology uses electric fields as combustion catalyst. There is observed application of technology in heat-and-power engineering, transport heat engines, disposal units etc. [1, 2]. It is demonstrated that the new technology allows intensify combustion processes and increase their controllability (i.e. the control of temperature, gases pressure, gradient of heat conductivity, and so on). There is made a conclusion about availability of the technology to solve critical energetic and ecological problems of the civilization.

We hope that this article will excite interest of a wide circle of readers, i.e. of professionals in the area of combustion and heat technologies, physicists, ecologists, and those people who are interested in new scientific ideas.

Global Ecological Problems are the Problems of Ineffective Combustion of Hydrocarbon Fuel

The ecological problems become more and more serious and threaten to grow into world ecological catastrophe. The main cause of the atmosphere pollution in megapolises is imperfection of combustion technologies (of heat-and-power engineering, heat machines, transport engines, waste combustion…). It has been proved that their share of pollution of the planet atmosphere comes to 70-80%. The combustion technologies mean any technologies of combustion of fuel, matters and gases. The combustion technologies are the most popular technologies in the world. The modern civilization without the combustion technologies is impossible. There are many industries, which apply these technologies, such as heat-and-power engineering, transport, metallurgy, food industry, oil-and-gas refining industry, chemical industry, waste combustion neutralization.

Therefore the global ecological problems will not be solved until the mankind develops the combustion technologies. The article describes and discloses the essence of the new technology of ecologically appropriate and effective combustion of fuel and wastes of any type.

Why is it difficult for ecologists to conserve nature?

Modern methods and technology used for solving of the ecological problems consist in **analysis** of extent and sources of the environment pollution (ecological monitoring), and, moreover, in **refinement of atmosphere, water and soil, which are used in different technologies or have been polluted due to the imperfection of the technologies** (it is refinement from such toxic components as drinking and run-off water, exhaust and waste gases etc.).

Unfortunately, methods of measuring of the huge spectrum of toxic matters, which are thrown into the atmosphere at matters combustion, as well as the methods of their utilization are impartially difficult, expensive and imperfect. Even using of an ideal technics for measuring of the atmosphere pollution is not effective since it is a struggle with **consequences** of the combustion and other technologies imperfection but not with **causes** of the atmosphere pollution.

Ecological and energetic effectiveness of the known combustion technologies

It is known that real effectiveness of transformation of chemical energy of fuel in the combustion process is low. For example, in heat engines it comes to 25%, effectiveness of transformation of heat energy into electric energy in thermoelectric power stations does not exceed 40%. If energy consumptions of mining, processing and delivery of the fuel to consumers are taken into account then the summary efficiency of the existent combustion technologies (of heat machines and devices) comes to no more than 10 – 15%! It means that more than the half of chemical energy of the fuel is transformed into heat and different toxic matters and waste gases, which pollute the planet atmosphere. They cause acid precipitation, a global “greenhouse” effect, which threaten to cause the climate warming, the world flood, and final poisoning of the living nature. Thus the share of imperfect power engineering in the nature pollution comes to 70 – 80%! How is it possible to cleanse the planet atmosphere, especially the environment of world megapolises, in the situation when the amount of toxic matters produced by transport and industries is comparable with the amount of rest pure air?

Until the mankind learns burn matters and gases by effective and ecologically appropriate way the planet atmosphere will remain polluted and in near future it can become uninhabitable. Thus we will not solve the ecological problems until we gain the understanding of the combustion processes and processes of combustion transformation of matter chemical energy into heat energy, and then into other useful types of energy (i.e. electric energy, mechanical energy, light energy).

Physical Essence and Problem of Classical Combustion of Matters

Combustion is one of the most difficult phenomena, which are known by the humankind. From the scientific point this phenomenon is a chain reaction of sequential fragmentation of fuel particles into smaller charged radicals; it is physical chemical processes of transformation of chemical energy of intermolecular connections as well as combustion also includes physical processes of transformation of energy into heat and light on molecular and atomic levels. Many other processes, which proceed simultaneously, are involved.

From school years we know that combustion is a process of interaction of fuel with an oxidant that is accompanied by heat and light energy generation. In higher school the words of “as well as by cryptic energy of chemical connection of waste gases” are added to the school definition. The combustion processes are studied and improved by scientists and experts of different areas (chemists, physicists, heat-and-power engineering specialists, thermalphysicists etc.). There are known fundamental investigations of combustion chain reactions made by such Russian scientists as N.N. Semenov, Ya.B. Zeldovich and their followers.

Until now intensiveness of fuel combustion is increased by air blowing into the combustion zone that increases the amount of waste toxic gases thrown into the atmosphere. **Let us arouse several questions, which seem naive at first sight. Why is an oxidant (air or oxygen) needed for matter combustion? Is it possible to do it without any oxidant?** How does the combustion process begin and proceed? There are a lot of vague questions in physics of combustion. For example, how can the temperature and the intensiveness of the combustion be regulated? Can the heat conductivity of the flame be controlled? How can heat motion of particles be regulated in the flame and in the waste gases, and what can it cause? There is another problem of combustion. It is the very hydrocarbon fuels, which are applied at modern heat processes. The great Russian scientist, D.I. Mendeleev stated that to use oil is the same thing as to stoke a stove with banknotes.

Since the hydrocarbon fuels are complicated chemical matters and the combustion processes are imperfect then in the process of their combustion a great amount of different by-product matters and toxic gases are produced. They waste unused self-energy of fuel into the atmosphere and pollute our planet.

Physical essence of the new electric combustion technology

How is it possible to burn the hydrocarbon fuel by environmentally appropriate way? How can this environmentally appropriate technology be realized in practice? A brief answer is following**: it is necessary to make such conditions of combustion of organic fuel, and to introduce such a combustion catalyst that the energy of intermolecular and intramolecular connections of the organic fuel becomes absolutely free. At that the chemical energy can be transformed into energy of electromagnetic radiation, for example, into light energy and heat energy without generation of by-product polluting matters and gases. Then a working body (for example, water) is effectively heated by this directed concentrated electromagnetic radiation. In this process effects of electromagnetic waves reflection and concentration are used, or the electromagnetic radiation is directly transformed into electric energy.** In this case there is a new opportunity to regulate flame temperature and to extremely decrease quantity of the oxidant, i.e. to create organics combustion with minimal quantity of waste gases.

Is it principally possible to burn organic matters and gases without the atmosphere pollution using electric field as a catalyst? It is possible if parameters of this field are correctly chosen. More precisely it is possible if huge Coulomb forces of the electric field are correctly regulated by interaction of electrically charged particles of the fuel and the oxidant with force lines of the electric field.

One variant of an experimental device is represented in Fig. 1.

Devices for approbation of the new electric combustion technology

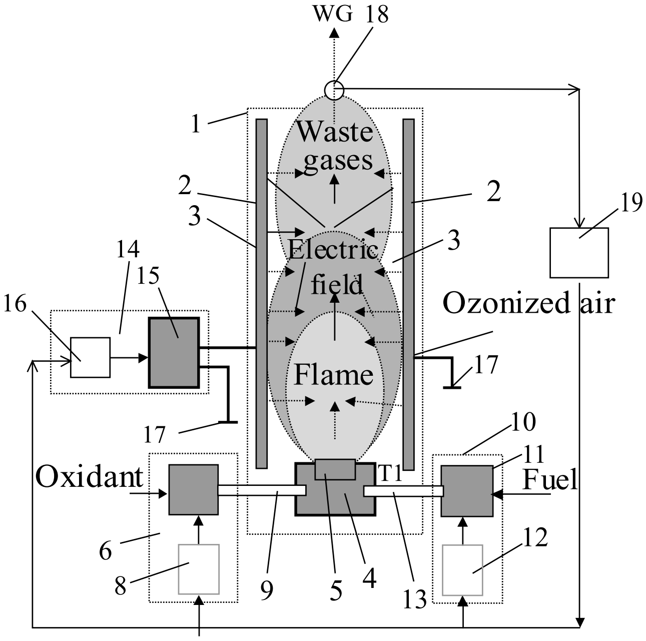

Fig.1 a

Plan of the experimental device

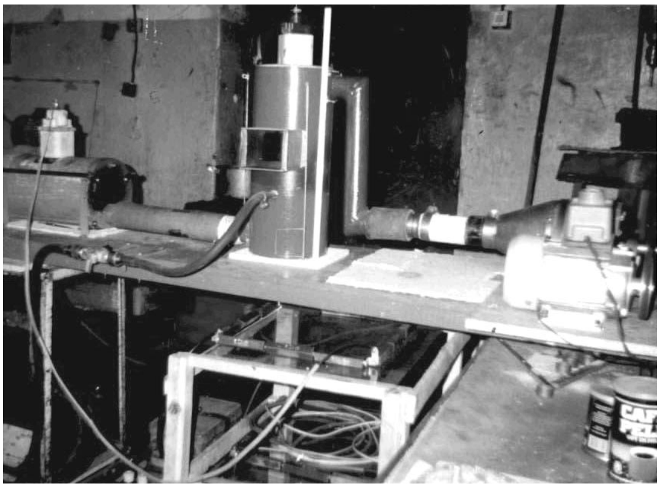

Fig. 1 b
Photo of the experimental system

A structure chart of the experimental device, which was designed to investigate the new electric combustion technology, is demonstrated in Fig.1a. The device contains a body (1) equipped with flat electrodes, which are insulated from the body (2), (they can be placed in either horizontal flat or vertical flat). The electrodes (2) are located on opposite inner walls of a combustion chamber (3). The device is equipped with a mixer (4) with a device for supplying of blended fuel to the zone of combustion. At the same time the device (5) is used for electrical combustion of the blend. The described device is equipped with an air track (6), which contains an oxidant activator (7) joined with a regulator (8) by a control circuit and with the mixer (4) by an air pipe (9). The regulator (8) serves for regulation of the oxidant activation extent. At the end of the air pipe it is possible to place a vortex device (it is not shown in the figure). Also the device is equipped with a fuel track (10) containing a fuel activator (11) and a regulator (12) of the activator and consumption of fuel. The block (11) is joined with the mixer (4) by a fuel pipe. The device is equipped with an electric combustion activator (14) containing a high-voltage transformer (15) of intensity and frequency joined to its regulator (16) by a control circuit. An electric outlet of the block (15) is joined to one of the electrodes (2), and the second outlet is safely electrically grounded by a grounding mat (17). The high-voltage wire of the outlet of the block (15) is connected with the electrode (2) through a bushing insulator (it is not demonstrated in Fig.1a). The device is equipped with a gas toxicity calculator (18) joined to the inlet of a mode optimizer (19), which is a control system for interacted regulating of all the parameters of combustion. For that the outlet of the mode optimizer (19) is joined to the inlets of control of the regulators (8), (12), (16).

The device operates in the following way. At first an activated oxidant of O1 is applied through the air track (6), activator (7) and air pipe (9) to the mixer (4). Then an activated fuel of F1 is applied through the fuel track (10) and the activator (11) to the same mixer (4). A blended fuel is prepared in the mixer (4) and then is atomized and combusted by a sparkle of the block (5). In the combustion chamber (3) flame and waste gases are processed by strong alternating electric field, which is generated by the block (14) in a gap between the electrodes (2). In the process of combustion waste gases toxicity is measured by the special toxicity sensor (18). Depending on current toxicity the parameters of flame combustion are regulated by the mode optimizer (19). More precisely the oxidant consumption and the extent of its activation are changed by the regulator (6), the fuel consumption and the extent of its activation are changed by the regulator (12), the intensity and frequency of the alternating electric field generated by the block (14) in the combustion chamber (3) is changed by the regulator (16). Let us note that the electric field in the combustion chamber (3), viz in the gap between the electrodes (2), catalytically influences on both the flame and the waste gases. The essence of the process lies in the fact that the processes of fragmentation and oxidation of fuel radicals and of toxic oxide molecules are accelerated. Intensiveness of combustion and of toxic gases refinement increases as well as the intensity of this field and its frequency do. As a result of the interconnected regulation of all the listed parameters fuel of any kind can be fully, intensively, “cleanly” combusted.

In Fig.1b there is a photo of the operating experimental device designed to investigate the processes of influence of electric field on the process of matters combustion and of cleaning of waste gases. The photo represents following things.

1. In the left side there is a furnace with a high-voltage electrode in a higher part of its body.
2. In the centre there is a vertical column of the electric combustion cleaning (after-burning) of toxic gases in the electric field; on the top of the column a high-voltage electrode can be seen. Gas pipes designed for toxic gases supply–withdrawal are connected with the column and located to the right and to the left from it.
3. A regulable blower is represented in the upper right side of the picture. It is designed for toxic gases withdrawal and connected with the gas pipe by the column.
4. At the foot of the picture there is a regulable high-voltage tension source (the electric field source), which is connected with the furnace and the column of toxic gases cleaning by high-voltage wires.

It is experimentally proved that combination of two stages of the combustion activation (i.e. in the furnace and the column) ensures ideal cleaning of toxic gases at combustion of any toxic matters.

Some Results of Experiments

Our experiments and investigations of many-sided influence of electric and high-voltage electromagnetic fields on the combustion process have proved that such practically ideal condition of fuel and matters combustion is possible to be realized in practice [1].

Low-powered static and alternating electric fields (i.e. the fields of constant sign and of variable sign) of more than 1kV/cm intensity were used as combustion catalysts as well as high-frequency electromagnetic fields of low power with some frequency of oscillations of molecules in the flame (flame power is 0.1 – 1% of heat power of the flame of the combusted organic fuel).

Peculiarities of Oxidants Application in the New Technology

It is known from the thermodynamics and combustion theory that optimal ratio of the oxidant mass, for example air, and the fuel is approximately 1:16 in an average combustion process. In our experiments with the electro-field catalyst the ecologically appropriate combustion of the hydrocarbon fuel (mazut, straw oil) was achieved at oxidant deficiency (for example, at the ratio of oxidant mass and the fuel of 1:1). For experts it means that there is a real possibility for 10 – 15 times decrease of amount of waste gases of any heat machines and to while their former power is the same. As the experiments demonstrate carbon and hydrocarbon are absolutely removed from the waste gases. In the experiments oxides of hydrogen and nitrogen were 4 – 8 times decreased and flame existence was 5 – 10 times increased.

The issues of the experiments prove the hypothesis about possibility of effective environmentally appropriate combustion as a process of direct transformation of chemical energy of organic fuel into electromagnetic radiation energy of the flame (including heat and optic diapasons). The transformation is accompanied by removal of polluting toxic components from the waste gases that occurs due to many times intensification of the combustion process under action of electromagnetic catalysts.

Operations of regulable activation of fuel oxidants (simultaneous or separate), which are introduced into the combustion technology, ensure additional improving of the combustion process. Especially it occurs at heavy oil fuel combustion and water-fuel emulsion combustion that was experimentally tested by us. The process becomes especially effective if the flame of the combusted activated blended fuel is additionally processed by alternating electric field. Due to introduction of the operation of all the combustion parameters regulation (of consumption of fuel, oxidant, of their activation extent, and of their combustion intensiveness extent), according to the information of the waste gases toxicity, it is possible to achieve effective combustion of fuel and waste products of practically all kinds.

Our experiments demonstrate that combination of the very flame procession with procession of the waste gases and air (i.e. an oxidant) by electric field is very effective for the waste gases cleaning. The essence of this additional cleaning of the waste gases lies in fragmentation of carbon particles and exhaust opacity by electric forces of an alternating field as well as in after-oxidation of some toxic oxides in the medium of an ozonized oxidant. Energy consumption for activation of the flame combustion by strong electric fields is small and does not exceed 1–3% of heat energy of the flame. Advantage of this invention is universality of the application for combustion of any inflammables. It is possible due to widening of the diapason of the electric field parameters regulation (of intensity and frequency), especially in the mode of their interconnected regulation.

The essence of catalytic action of the alternating electric field on the flame combustion process consists in effective breaking of dipole radicals of fuel by an activated (dipole) oxidant. Moreover, the essence lies in better mixing of layers of combusting flame with the oxidant that occurs due to removal of a doubled electric layer from the limit line of the flame. Thus this technical solution allows achieve new positive effects due to its significant peculiarities, i.e. extending of application area of the known electric combustion method over the combustion process of any inflammables as well as it allows significantly increase controllability of the flame combustion process.

Let us note that in the experiments on the combustion of organic fuel in strong electric fields the regulation of the flame temperature and of its existence was achieved at unchanged consumption of fuel and of oxidant. That was achieved by changing of the parameters of the combustion electric-field catalyst (intensity and frequency) of a longitudinal electric (electromagnetic) field. Rotation and stabilization of the flame was achieved by a rotating transversal electric field. Changing of the flame height was realized by the longitudinal electric field.

The investigation, which has been performed by us, proves that directly acting on the flame the very electric field which directly influences on the flame and emission of electrons flow (ideal type of oxidation) into the flame can most effectively intensify the combustion process and make it harmless for the humankind and for the environment!

It has already been experimentally proved that energy consumption required for generation and regulation of this electric field and of the electrons flow for intensification and environmental appropriateness of the combustion is quite small relatively the combustion energy and comes to fractions of a percent of the flame energy. Thus our experiments prove that the best **"oxidants" and combustion catalysts are** not superfluous air but **the electron and the electric field!**

Regulation of Heat Conductance of the Flame and the Heat Flow

As our experiments have demonstrated, the electric field can act as an effective combustion catalyst as well as a regulator of its intensiveness; moreover, it can control even a vector of heat conductance. It is proved by the experiments that the parameters of this field can be regulated as well as temperature of the flame and gradient of the flame heat conductance. Interesting experimental results were obtained measuring full heat of combustion of the same quantity of fuel at the usual method of fuel combustion and at the method, which uses combustion electric catalysts, even in the case of oxidant deficiency. In the last case the energy of fuel combustion almost 1.5 times increases that can be explained by fuller transformation of fuel chemical energy into electromagnetic radiation. At the usual combustion methods the chemical energy of the organic fuel was not completely used and remained as a cryptic summary chemical energy of intermolecular connections of many toxic waste gases, which were exhausted into the atmosphere by heat devices. Basing on the performed experiments it may be supposed that, evidently, specific heats of matters are 20-50% higher at this method than at their usual combustion method. The essence of the new electric combustion technologies consists in this new physics of combustion.

The author has already got patents for invention of Russian Federation [3 – 12] for the methods of regulation and intensification of matters combustion processes.

Some peculiarities of atomization, inflammation and combustion of organic fuel in electric fields

We have not discussed all the potential capabilities and advantages of the new electric combustion technology for different areas of technics. Let us demonstrate them in more details.

One of characters of the new electric combustion technology is **effects of oxidant ozonization, of electrostatic atomization and electrostatic injection of electrically charged particles of fuel and the oxidant that is followed by generation of the finest aerofuel opacity on a molecular level in a combustion chamber**. Naturally, such fine atomization of the fuel causes its easier evaporation, inflammation and combustion especially in the medium of the ozonized oxidant. **The mechanism of fuel electrostatic fragmentation is physically explained by Coulomb interaction of repulsion from each other of electrically likely charged fuel drops (particles) that is accompanied by their progressing fragmentation and corresponding decrease of their mass and electric charge**. As the experiments have shown the extent of the fuel drops (particles) fragmentation depends on initial electric potential of fuel charging and initial size of injected fuel drops (particles), which for their part depend on configuration and size of a fuel-injection nozzle as well as on pressure occurring in a fuel line of the heat machine.

Generated by dipolar high-voltage potentials catalyzing electric field can be introduced into the atomization zone or (and) the combustion zone of the combustion chamber of an engine. Its Coulomb forces accelerate motion of electrically charged particles of fuel and oxidant to an opposite electric potential, which is connected with, for example, a piston of an engine. Thus electric field acts as an electrostatic pump that allows decrease pressure in a fuel line as well as allows practically ideally atomize the fuel and mix it with the oxidant. Moreover, it lets simplify and improve the systems of injection and inflammation of the blended fuel in heat machines and devices, for example, in internal combustion engines or in boiler plants. Their application can cause additional improving of energetic and ecological characters of the combustion technologies. According to our information, such high-performance fuel-injection nozzles have yet been unknown in technics therefore they have not been applied in heat machines (injector internal combustion engine).

Is it possible to combust water? Water as a fuel!

More amazing experimental results were obtained at 40-80% dilution (emulsification) of liquid organic **fuel**, for example diesel oil, with **habitual water**. Summary energy as heat and light, which is generated in the process of such blend combustion, was not practically changed that can be explained by liberation of energy of chemical connections of water as well as of fuel. In the process of the experiments on combustion of the emulsion at first **it was transformed into the finest water vapor on the molecular level by means of capillary electroosmosis and of electrostatic atomization. Then dipolar water molecules, which had been electrified by field, were fragmented into hydrogen and oxygen, and then the hydrogen was effectively combusted in the medium of the ozonized oxygen.** Let us note that at this mechanism of water "evaporation" and of following splitting of water molecules into hydrogen and oxygen the electric field expends energy, and heat of organic fuel combustion just accelerates (catalyze) this process.

Editor: It is necessary to note that electric field can not expend energy. If there are no conduction currents then the field source does not decrease difference of potentials. Hence it can be concluded that effectiveness of such power systems is possible to be UNRESTRICTEDLY HIGH. The author writes about it below as about an "incomprehensible phenomenon".

The most amazing and incomprehensible fact is that the electric field transforms water into "vapor" and breaks the water molecule into hydrogen and oxygen practically **without any expenditure at minimum of energy consumption**. Electric power of the intensity high-voltage transformer, which is necessary for strong field generation, lies in the interval from several watts to tens of watts. As the experiments have demonstrated electric consumption of the high-voltage source practically is not changed at correct shunting of the flame by the electric field (i.e. through an air gap). Current consumption practically did not increase depending on mode of fuel atomization and its combustion (of flame height, of combustion intensiveness...).

There are only two conditions of effective work of static electric field as an electrostatic pump-atomizer of fuel and as a catalyst of blended fuel combustion. The first one is intensity of the field in the zone of fuel atomization and in the zone of flame combustion. The second one is enough emission of electrons into the blended fuel flow. In this mode Coulomb forces make work of electrostatic pressure generation and of electrostatic fragmentation of particles and molecules of fuel and water. The forces repulse likely charged drops of water. Heat motion of fuel radicals, which are combusted and fragmentized in the flame, impedes their chemical recompounding into water molecules in the zone of combustion. Moreover, it causes proceeding of a very physical chain reaction of hydrogen combustion in the flame.

Therefore additional light and heat energy, which is generated in the flame by combustion of hydrogen generated from water molecules that occurs in the ozonized oxygen, does not decrease summary energy of combustion of blend of fuel with water but increases it.

Thus using of electric and electromagnetic fields as the strongest catalysts of combustion of organic fuels and any matters clears great **prospects** of combustion technologies improving. Moreover it allows create ecologically appropriate heat machines and devices and 20-50% increase their effectiveness due to fuller transformation of chemical energy of inflammables into heat or light energy.

Ways of radical improvement of heat machines or about a new mechanism of transformation of heat energy into mechanical energy and into kinetic one

For clear realization of causes of extremely low output of modern heat machines that has generally lead the civilization to the ecological catastrophe it is necessary to understand their common principle of operation as well as a cause of heat energy non-effective using in them.

A mechanism of transformation of heat energy, which is generated by fuel combustion, is the same for all known heat machines. It consists in generation and transformation of superfluous pressure of a heated working body (i.e. gas, vapor etc.) into kinetic energy of a working element of the heat machine (i.e. a piston of an internal combustion engine, a turbine of an aviation engine, a reaction jet flowing out a rocket nozzle etc.) in special chambers of the machines.

It is also known that the higher the temperature of the heated working body, for example of a gas, the higher initial pressure in the working chamber of a heat machine. **Why output of heat machines is so low?** Any expert of thermal physics, thermodynamics, and heat machines can answer that efficiency (output) of heat energy transformation into mechanical one is defined by the second law of thermodynamics and is clearly illustrated by Carno heat cycle. According to these postulates, **output** of an ideal heat engine does not depend on a working matter and on the engine construction but is defined by temperatures of the working body in the starting point and in the final point of the cycle, i.e. by the temperatures of the heater and of the refrigerator of the **heat machine**. Real output of the heat machines is limited by heat stability of materials and by imperfection of engine constructions.

Nevertheless known thermodynamic processes and lows of heat motion of particles and molecules are significantly changed in strong electric fields and demonstrate new prospects for improving of heat machines.

A solution of the problem can be briefly formulated as: to increase output of heat machines it is necessary to use a new mechanism of redistribution of heat energy into energy of directed pressure of a compressed working body (i.e. gas, vapor, etc.) on the working element of the heat machine. At the minimal pressure it should be redistributed on sidewalls of the working chamber.

Can it be realized in practice? On the face of it the technical solution is unrealizable nevertheless it exists. It is possible to redistribute and to regulate heat motion (pressure) of heated gas inside a closed reservoir and, hence, its temperature in a certain direction. To achieve this aim **it is necessary to introduce static electric field into the combustion chamber of a heat machine and to orient heat motion of electrically charged and dipolar molecules of expanding working bodies (i.e. gas, vapor) along force lines of the electric field by electric Coulomb forces**. Something similar occurs, for example, in liquid crystal cells of electronic timer, in Kerr cells (electric optic effect) at periodical application of electric field to them that causes dipolar molecules turn along a field vector as well as change of light penetrability of these matters.

There is a difference between this known in electrooptics phenomenon and our case. The difference lies in the fact that polarized gas molecules remain movable along force lines of the electric field in contrast to liquid crystals, which do not. This fact causes redistribution of parameters of heat energy (i.e. pressure, temperature, and heat penetrability) of dipolar molecules of heated gas just along force lines of the field. The more the intensity of the field at the initial temperature of the working body is the more the difference of pressures on the sidewalls of the chamber and on the working element, for example, on an engine piston.

Heat motion of polarized particles of heated gas can be decelerated by combination of three electric fields of constant sign in accordance with all three coordinate axes; hence it is possible to significantly and quickly decrease temperature and pressure of the gas. In this case the heat energy of the heated gas makes jump transformation into electromagnetic radiation.

Editor: Historically this method, i.e. plasma retention by electric field, was proposed by Oleg Lavrentiev in 1948. He had 7-grade education and served as a sergeant in Armed Forces in Sakhalin. After he sent Stalin a letter containing a phrase, as "I know a secret of hydrogen bomb creation" he was invited to Moscow to set his ideas out to academicians. His ideas were not lost nevertheless the scheme of plasma electrostatic retention was not applied ("Expert" magazine #23, 18 of July of 2001). In 1950 Sakharov and Tamm proposed a scheme of a toroidal magnetic thermonuclear reactor which has been developed until now. Why is it so? It is difficult, expansive and practically unreal…

Thus introduction of strong electric fields into heat machines allows significantly increase effectiveness of transformation of heat energy of a working body into mechanical and kinetic energy of working elements of the machines by means of directed regulation of temperature and pressure of heated gas, for example, in the starting point and in the final point of a working stroke of a piston. In other words it allows increase output of the machines.

APPLICATION AREA OF NEW ELECTRIC COMBUSTION TECHNOLOGY

Practical application of the new combustion and thermal technologies

The new technology is universal and applicable in practically all spheres of technics. Therefore we believe that further development and application of this new technology of fuel and waste products combustion is very important for radical improving of all the heat machines and of all thermal technologies. As the civilization uses hydrocarbon fuel and matters in its life circle then just this new technology can solve critical ecological and energetic problems of the civilization.

It is impossible to view in details all proposed perspective technical solutions based on this technology [1-25] within one article. Therefore we give only one striking example.

Environmentally appropriate engine for motor transport

Since the transport, which uses thermal internal combustion engine (ICE), is most harmful for the environment then let us discuss ways of their ecological and energy improving. Several years ago the author patented "method of intensification of ICE operation" [7]. The essence of the invention consists in introduction of controllable electric field into combustion chambers during the whole operation period of an engine by special monoelectrode spark-plugs (Fig.2).

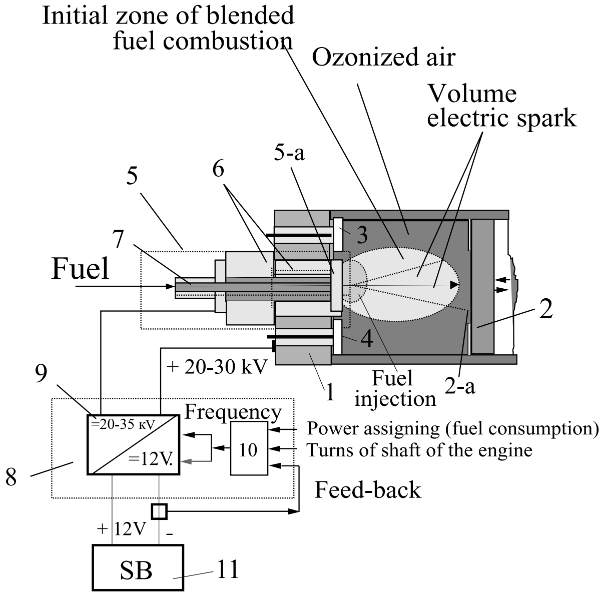

Fig.2a

Plan of an environmentally appropriate ICE
(for transport)

1. – combustion chamber;
2. – piston, 2a –heat-resistant surface of the piston (zone of electric-spark discharge)
3. – admission valve (its canal is not indicated)
4. – exhaust valve (its canal is not indicated)
5. – combined mono-electrode spark-plug; 5a – fuel nozzle with hollow central electrode and end disk electrode
6. – electric insulator of the spark-plug – fuel nozzle
7. – fuel pump (for example, an electrostatic one)
8. – high-voltage regulable transformer (electric field source)
9. – force block of tension high-voltage transformer (=12V/25kV)
10. – system of regulation of electric field source (9)
11. – board current network =12V (SB is a storage battery of an automobile).

This structure chart briefly demonstrates basic components of a new ICE, where there is directed pressure of gases on a piston. In particular, many inventions of the author are realized in practice here, i.e. semi-digital spark-plug, electrostatic fuel nozzle (5), electrostatic fuel pump (7). A new mode of volume inflammation of blended fuel has been realized in practice, i.e. "a spark coming from a disk electrode (5-a) to a piston (2-a)". Since it is possible to regulate the electric field intensity generated by a high-voltage source of tension then an advance angle of blended fuel inflammation can be regulated, gases pressure on the piston at explosion stroke of ICE operation can be regulated as well.

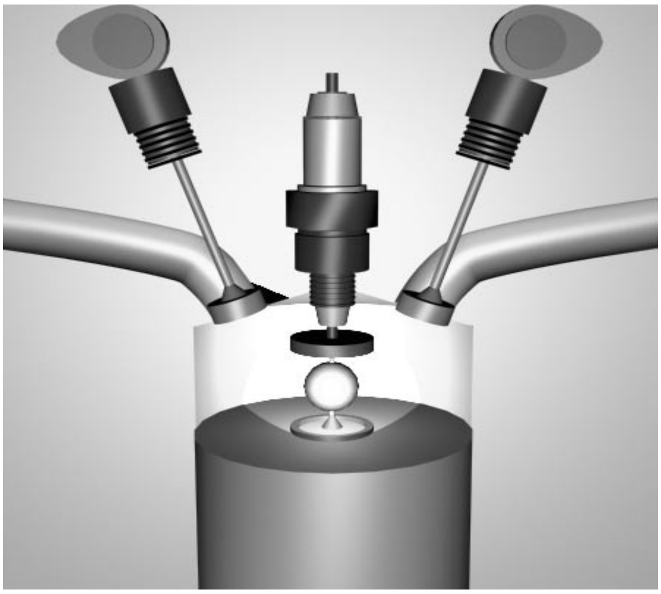

Fig.2b

Photo of a simplified construction of the new ICE

Fig.2b demonstrates simplified three-dimensional construction of the proposed energetically and ecologically perfect ICE. There are observable elements such as a combustion chamber, a piston, valves, a monoelectrode spark-plug, and an electrostatic fuel nozzle (in the centre on top), which is connected with the spark-plug. In the centre of the combustion chamber it can be seen volume inflammation of the blended fuel at spark appearance from end electrode to the piston.

As a result a summary positive effect of ecological, construction and energetic improving of ICE is achieved, i.e. effective cleaning of exhausted gases directly in the combustion chambers of the engine, significant simplification and perfection of a fuel injection system, improvement of system of distribution and electric inflammation of the blended fuel. In this engine a distributor (as well as its analogues) is removed at all since sparking and intensive inflammation of the blended fuel automatically appears between central electrode of the spark-plug and the piston which comes to the upper "dead" point. The central electrode of the spark-plug is constantly under high tension. The advance angle of ignition is regulated by change of the electric field intensity. Powerful multipoint ignition causes simultaneous intensive inflammation of the blended fuel that occurs throughout the whole chamber. The electric field as a powerful combustion catalyst intensifies the process of the blend combustion at the explosion stroke of the engine and at after-burning of waste gases directly in the combustion chambers at the following output stroke of the engine operation. Directed along the axis of the piston in the combustion chambers this electric field serves as a transformer of gases heat energy into mechanical energy of the thermal engine pistons. That is caused by the fact that the field orients heat motion of the expanding gases along the axis of the pistons in combustion chambers of the engine at explosion stroke of ICE just. It causes redistribution of heat energy and increase of the gases pressure on the very pistons that significantly improves effectiveness of transformation of heat energy of the fuel combustion into mechanical energy of the piston motion (theoretically the improvement is by two-three times), i.e. it two-three times increases output of a classical thermal engine making it come to 70-80%.

The electric field, which is introduced into the combustion chambers of an internal combustion engine, ensures significant economy of fuel (up to 30-40%) at saving of its working characteristics. It occurs due to fog electrostatic spray of fuel and to the fuel electrization as well as due to oxide ozonization. Moreover it occurs as a result of deep after-burning of hydrocarbon components of fuel, combusting blend and waste exhaust gases. This method allows effectively regulate temperature of the blend combustion in the chambers while required compression in cylinders is the same, for example, to minimize nitric oxide generation in the exhaust gases. As a result there is no necessity in external devices for cleaning of exhaust gases of internal combustion engines; ecological appropriateness improvement can be simultaneously achieved. Additional aerosol hydro-alkaline processing of these gases can allow totally clean **exhaust gases** of the transport, which operates on thermal engines.

Workability of such an environmentally appropriate engine in two-stroke and four-stroke variants is experimentally demonstrated. There are also experimentally shown decrease of all toxic components in exhaust gases, 20-30% increase of useful power of the engine, and simultaneous decrease of fuel consumption.

Thus practical application of the new electric combustion technology can exert revolutionary influence upon propulsion engineering development and upon increasing of ecological compatibility and effectiveness of the transport, which operates on thermal engines.

Diagram of operations of technology of effective using of refinery wastes in heat-and-power engineering

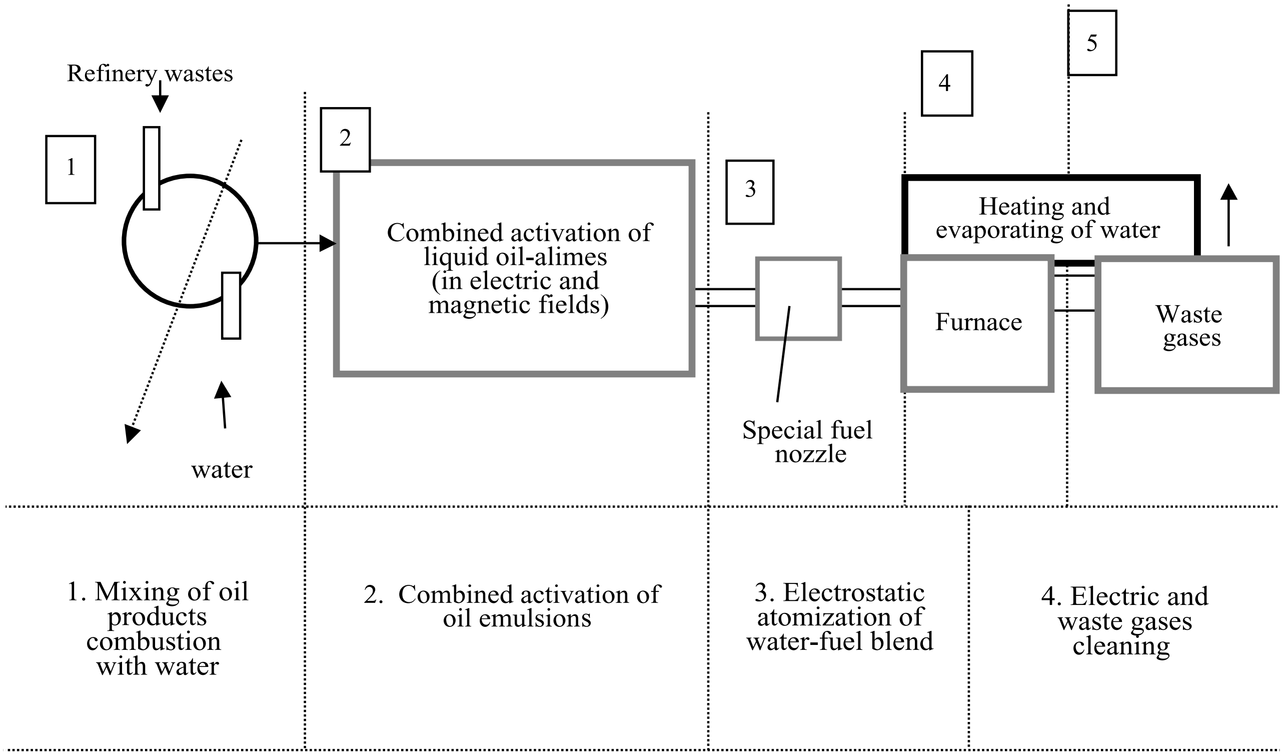
Refinery wastes
1
2
Combined activation of liquid oil-alimes (in electric and magnetic fields)
3
4
5
Heating and evaporating of water
Furnace
Waste gases
Special fuel nozzle
water
1. Mixing of oil products combustion with water
2. Combined activation of oil emulsions
3. Electrostatic atomization of water-fuel blend
4. Electric and waste gases cleaning

Fig.3

Technology of environmentally appropriate combustion of refinery wastes as fuel for boiler plants

The author has patented other technical solutions of improvement of different devices, which are based on these technologies. These solutions concerning boiler plants, gas turbine engines and jet engines and many other heat machines and devices allow increase their effectiveness and simultaneously significantly improve their ecological factors. As an example of application of this technology in heat-and-power engineering there is a diagram of operations of the technology of the environmentally appropriate combustion of refinery waste as a fuel for boiler plants (Fig.3). Unfortunately the limits of one article do not allow discuss these inventions in details nevertheless the author will realize it in practice in future works.

Conclusions:

1. The existent methods of solving of the ecological problems are ineffective as well as work of organizations, which deal with ecological and nature conservation activity. It is caused by the fact that they are oriented to search and remove consequences of ecological pollution of the nature instead of their causes.
2. The article demonstrates real causes of global ecological problems, which consist in imperfection of technologies of generation and transformation of energy.
3. Using of strong electric and electromagnetic fields of low power as combustion catalysts allow significantly improve ecological factors of combustion technologies and of devices, which realize the technologies in practice. In other words it allows realize in practice "clean combustion of fuels and wastes and possible existence of "clean" heat technics".
4. Application of electric and electromagnetic fields in the combustion technologies allows significantly increase effectiveness of transformation of chemical and heat energies of fuel into mechanical and electric energies.
5. The electric combustion technology allows realize in practice principally new ways to control combustion process as well as to control many thermal and kinetic processes (i.e. processes of pressure, heat conductance, temperature, etc.), i.e. allows increase output of heat machines.

6. The new electric combustion technology is a universal method of solving of ecological and energetic problems of the civilization and a universal way of radical improvement of technics (i.e. of the transport, boiler plants, oil-and-gas processing plants, combustion wastes recycling plants, etc.).

References

1. V.D. Dudyshev Electric combustion technology is an effective way to solve energetic and ecological problems //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #3 1997;
2. V.D. Dudyshev How is it possible to save the civilization and the nature from global ecology pollution? //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #11 1997;
3. V.D. Dudyshev Method of combustion control. Patent of Russian Federation #2071219 of 19.04.1994;
4. V.D. Dudyshev Method of intensification and regulation of combustion. Patent of Russian Federation #2125682 of 06.06.1995;
5. V.D. Dudyshev Method of electric combustion of refuse and wastes. Patent of Russian Federation #2117870 of 06.06.1995;
6. V.D. Dudyshev Method of electric combustion cleaning of waste gases. Patent of Russian Federation #2125168 of 15.11.1995;
7. V.D. Dudyshev Method of combustion intensification in a fire-chamber of a boiler plant. Patent of Russian Federation #2079766 of 14.06.1995;
8. V.D. Dudyshev Method of flame extinguishing, Copyright of USSR #1621234 of 12.03.1988;
9. V.D. Dudyshev Method of decrease of toxicity of exhaust gases of motor transport. Patent of Russian Federation #2117179 of 20.02.1996;
10. V.D. Dudyshev New method of environmentally appropriate matter combustion as a radical way of solving of global ecological problems. // Collected reports of II international congress of "Environment for us and future generations: ecology, business and human rights in new conditions", Russia, Samara, 1997;
11. V.D. Dudyshev New method of refuse electric combustion // Collected reports of III international congress of "Environment for us and future generations: ecology, business and human rights in new conditions", Russia, Samara, 1998;
12. V.D. Dudyshev Problems of ecological education and upbringing of the youth // Collected reports of II international practical conference of "Pedagogical activity as a cultural activity", Russia, Samara, 1999;
13. V.D. Dudyshev Environmental safety of motor transport //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #5 1997;
14. V.D. Dudyshev Perspective technical works out and inventions of ecological improvement of motor transport. //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #12 1998;
15. V.D. Dudyshev Method of ICE exhaust gases cleaning, i.e. electric combustion cleaning and electric filter at ICE output. Patent of Russian Federation #1404664;
16. V.D. Dudyshev Problems and ways of ecological improving of native motor transport. //"Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine, #11 1998;
17. V.D. Dudyshev Method of internal cleaning of exhaust gases of ICE. Patent of Russian Federation #2165031;
18. V.D. Dudyshev Method of intensification of ICE operation. Patent of Russian Federation #2135814;
19. V.D. Dudyshev Method of decreasing of toxicity of ICE exhaust gases and a device of its application. Patent of Russian Federation #2135814;
20. V.D. Dudyshev Method of intensification and regulation of combustion. Patent of Russian Federation #2125682;
21. V.D. Dudyshev Method of heat machine control. Patent of Russian Federation #2134354;
22. V.D. Dudyshev Environmentally appropriate ICE. // "New technologies" ("Novye tekhnologii") magazine #2 2001 Samara;
23. V.D. Dudyshev New electric combustion technologies. // "New technologies" ("Novye tekhnologii") magazine #3 2001 Samara;
24. V.D. Dudyshev New electric combustion technologies. // "New technologies" ("Novye tekhnologii") magazine #4 2001 Samara;
25. V.D. Dudyshev New electric technology of fire non-contact extinguishing and ignition prevention. // "New technologies" ("Novye tekhnologii") magazine #9 2002 Samara;

Infinite Energy Magazine

* Cold Fusion * New Energy
* New Science * New Technology

Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

EAGLE-RESEARCH
Energy Solutions
since 1984

A Research Organization that
Develops & Distributes
Practical Energy-Saving
Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

New Effect of “Cold” Evaporation and Dissociation of Liquids Based on Capillar Electroosmosis

Valery D. Dudyshev
Doctor of Technical Science, Professor,
Corresponding Member of Russian Ecological Academy

ecolog@samaramail.ru
http://www.intrasoft.ru/nizot

There is discovered new electric physical effect of intensive “cold” evaporation and dissociation of liquids and aqueous solutions into fuel gases **without any energy consumption** due to high-voltage capillar electroosmosis [1].

Prospects and problems of hydrogen engineering

Effective obtaining of hydrogen from water is a long-standing tempting dream of the civilization. That is connected with the fact that there is a huge amount of water on the planet, and hydrogen engineering promises the possibility to obtain unlimited quantity of “free” energy obtained from water. All the more, the very process of hydrogen combustion in oxygen medium, which is obtained from water as well, ensures ideal high-calorie and environmentally appropriate combustion.

Ecological and energetic problems are very actual nevertheless they have not been effectively solved. All known methods and devices of production of hydrogen and other fuel gases are ineffective since there is no a real high-performance technology of evaporation and splitting of liquid molecules. The main cause of ineffectiveness of the analogous consists in their difficulty and in energy consumption for breaking of intermolecular connections at dissociation of water liquid fractions.

Physical-chemical structure of even habitual tap water is quite complicated since there are numerous intermolecular connections, chains and other molecular structures in water. In particular, in habitual tap water there are different chains of oriented water molecules, which are peculiarly connected with admixture ions (cluster formations), its various colloidal compounds and isotopes, mineral matters as well as various dissolved gases and admixtures.

It is a paradox but in the living nature there is a long-standing effective way of electric capillar delivery and “cold” evaporation of liquid, which allows transform it into gaseous state without heat energy and electric energy supply line. This natural effect is realized in practice by plants, which deliver aqueous solution and make its “cold” evaporation by capillar electroosmosis. It is quite comprehensible that this natural energetically perfect technology is applicable in methods of liquids transformation into fuel gases. The author of this article has designed such experimental devices of cold electric capillar evaporation of liquids according to electric pumps of trees (Fig.1-3).

The simplest operating device, which experimentally realizes in practice the effect of high-voltage capillar electroosmosis of the “cold” evaporation and dissociation of water molecules, is demonstrated in Fig.1.

New Electroosmotic Capillar Method of Obtaining of Fuel Gas from Water

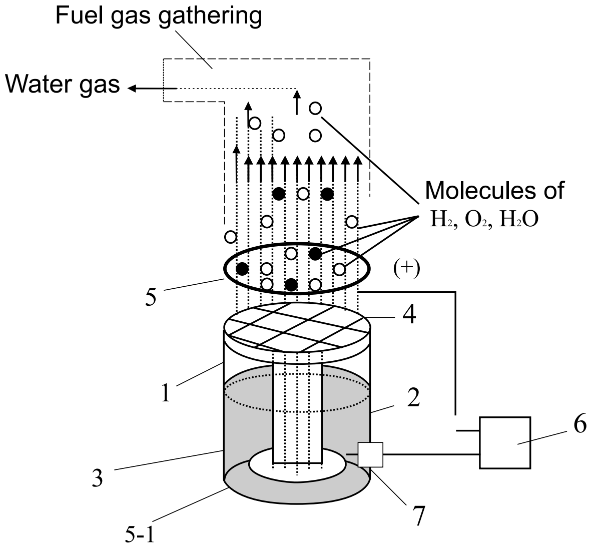

Fig.1
Simplest device of capillar electroosmosis of liquids

First experiments on the electric capillar dissociation of liquids were made with using of habitual water as well as its solutions and water-fuel emulsions of various concentrations as liquids. In all these cases fuel gases were successfully obtained in spite of the fact that these gases greatly varied due to their composition and heat capacity. The experiments on the electroosmotic evaporation and dissociation of liquids are realized in practice by the following way. First a wick (3) and a porous evaporator (4) are moisten with a water-fuel blend (emulsion) (2) then the blend (2) is poured into a reservoir (1). Then a high-voltage source of tension (6) is switched on and high-voltage difference of potentials (about 20 kV) is supplied to the liquid at some distance from the capillaries (i.e. from the wick (3) and the evaporator (4)). The source of electric field is joined to the device by electrodes (5-1) and (5). A plate perforated electrode (5) is placed above the evaporator (4) surface at a distance, which is enough to prevent an electric breakdown between the electrodes (5) and (5-1). Electrostatic forces of longitudinal electric field acts on the liquid. As a result dipolar polarized molecules of the liquid move along capillaries of the wick (3) and evaporator (4) from the reservoir to an opposite electric potential of the electrode (5) (i.e. electroosmosis occurred). Liquid molecules are detached from the evaporator (4) surface by these forces and transformed into a visual fog, i.e. the liquid is transformed into another aggregative state at minimal energy consumption of the electric field source (6). After that they provide electroosmotic elevation of this liquid. In the process of detaching and collision of evaporated liquid molecules with molecules of air and ozone as well as with electrons there is occurring partial dissociation between the evaporator (4) and the upper electrode (5) in an ionization zone. At the process a fuel gas is produced, which can come thorough a gas collector (7), for example, into combustion chambers of motor transport engine.

It has been experimentally shown that change of intensiveness of process of evaporation and dissociation of vapor molecules depends on change of distance from the electrode (5) to the evaporator (4). Moreover, this dependence is conditioned by the following factors, viz by changing of the evaporator area, kind of the liquid, quality of capillar material of the wick (3) and the evaporator (4), parameters of the electric field as well as the source of intensity (of power) (6).

The author's first experiments on this simplest device, which were organized in 1986, demonstrated that "cold" water fog (i.e. gas) appears in capillaries from liquid (i.e. water) at high-voltage electroosmosis without any observable energy consumption but just using potential energy of the electric field. This conclusion is evident since in the process of experimenting consumed electric current of the field source was the same and equal to the consumed current of the source idling. This current was not changed depending on the appearance of the liquid evaporation.

The experiments demonstrated that capillar electroosmosis evaporated quite significant quantity of water (1 liter) without any energy consumption for 10 minutes at a 10-centimeter-diameter capillar cylinder. That is to say that the consumed electric power (10 Watts) of the electric current source, i.e. of the tension high-voltage transformer (20 kV), was unchangeable and did not depend on mode of its operation. It has been experimentally stated that the whole consumed energy supplied by the current network is insignificantly small comparably with the energy of liquid evaporation. It can be explained by the fact that power was consumed only to generate the electric field and did not increase at liquid capillar evaporation that occurred due to work of an ionic pump and of a polarized pump. Hence the effect of cold electric-capillar evaporation of liquid is very economical in respect to the process energy consumption.

In spite of the fact that energetic essence of this process has not been disclosed it is evident that both "cold evaporation" and water dissociation are realized in practice by potential energy of the electric field. More precisely, the visual process of evaporation and water splitting into H_2 and O_2 at the capillar electroosmosis is realized by powerful Coulomb forces of this strong electric field.

Editor: It is not the most surprising fact. The most astonishment is aroused by school-day stereotypes that work of a field at a closed cycle is equal to zero. These stereotypes have been kept in mind of people for a long period of time. Everybody understands that a field can do work but if a body falling from some height is accelerated in a potential field and its kinetic energy increases then it requires energy consumption to relevate the body up to this height. Nevertheless, the analogy of a gravitation field with an electric one is not one-valued since the electric field may be generated only at a part of trajectory of the accelerated body motion. An electric field can be pulsating, it can be screened or it is possible to change its direction at the reverse part of the trajectory in such a way that the field constantly accelerates the body. Hence a principal conclusion can be made: summary work of a potential field may not be equal to zero. This conclusion has earlier been proposed by A.V. Frolov in his article published in the USA (Newsletter of the Institute for New Energy. May 1994. p. 1-4).

In principle this uncommon electroosmotic pump-evaporator-splitter is an example of the perpetual motion machine of the second type. Thus the high-voltage capillar electroosmosis of aqueous liquid provides really intensive and energetically free evaporation and splitting of water molecules into the fuel gas (H_2, O_2, H_2O) by means of using of potential energy of the electric field.

To produce more complete dissociation of water molecules into the fuel gas it is necessary to make the whole water molecules collide each other and be split into molecules of H_2 and O_2 in an additional transversal alternating field (Fig.2).

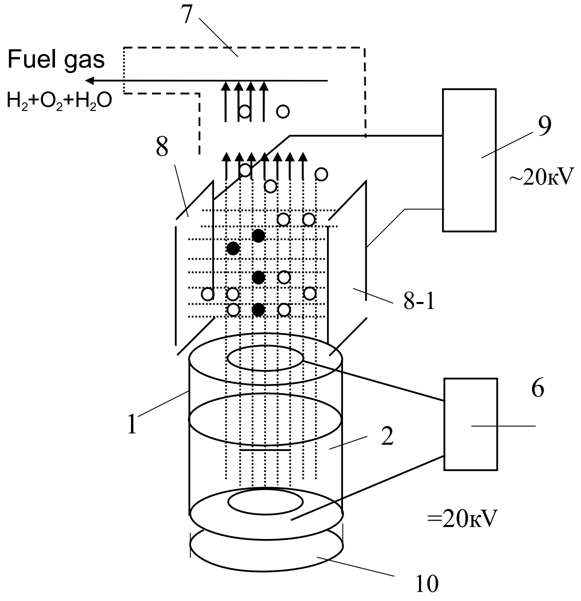

Fig.2

Device for production of fuel gas by electroosmosis. The device is equipped with an additional high - voltage splitting center of liquid molecules by electric alternating field.

At the second stage of water dissociation the energy of the second electric field is used, more precisely, powerful electrostatic forces are used to intensify oscillation resonant process of "collision-repulsion" of electrified water molecules represented as water gas. The result of this process is complete breaking of liquid molecules and generation of fuel gas molecules.

Conditions of optimal dissociation vary due to a kind of the liquid, to capillaries properties, and to the field parameters. These conditions are caused by required productivity of the process of dissociation of concrete liquid. Fig.2 demonstrates in details functional structure and composition of the device equipped with two sources of the electric field.

In the case of preliminary division of initially chemically neutral water into chemically active fractions (i.e. acid fraction and alkaline fraction) realization of the technology of production of fuel gas from water becomes possible at temperature below zero (up to –30°C). In winter it is quite important and useful for motor transport. This "fractional" electrically activated water does not freeze at degree of frost; hence the device designed for hydrogen production from such activated water can operate at environment temperature below zero and at degree of frost.

This principle of additional chemical activation of water (or liquid) is realized in practice in the device

(Fig.3). In contrast to the devices mentioned above (see Fig.1, 2) this device is supplied with an electrochemical activator of liquid (3) with evaporators (4).

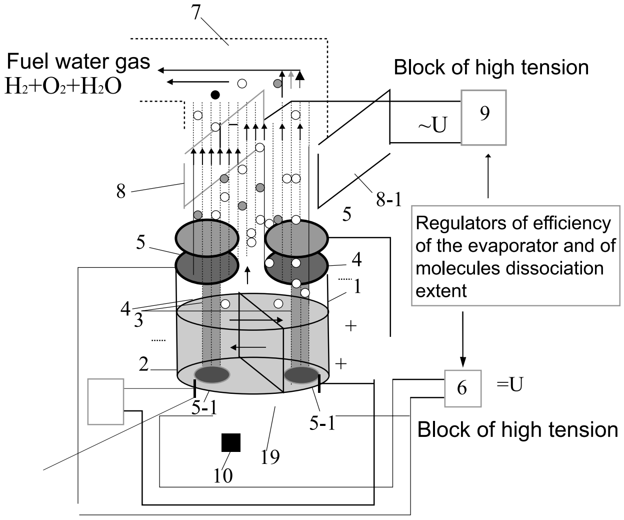

Fig.3

Hybrid device equipped with two sources of electric field and liquid electric activator

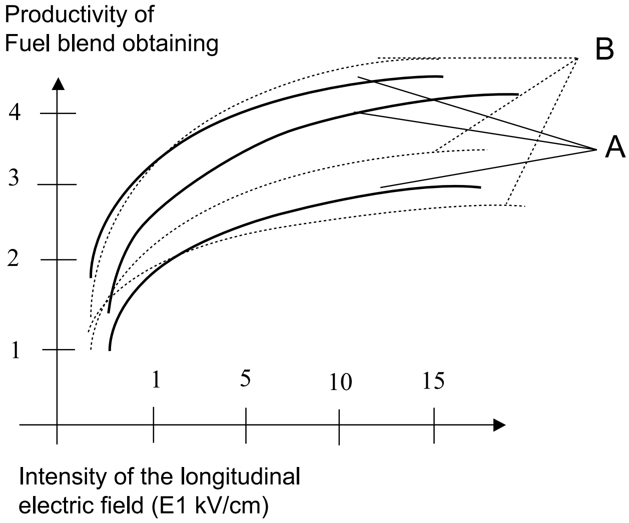

Fig.4

Productivity of the method increases as well as oscillation frequency of the second transversal electric field (A) and evaporating area (B) increase.

There is a known and low-expenditure progressive electric technology by Stanly Mayer, which has been worked out to obtain fuel gas from water (Mayer cells). Our technology is more progressive and ensures higher productivity than Stanly Mayer's technology does [3]. It can be explained by the fact that this electroosmotic effect of evaporation and liquid dissociation combined with a mechanism of the electrostatic pump and the ionic pump ensures intensive evaporation and dissociation of liquid as well as effective detachment of gas molecules from the dissociation zone. The process of gas molecules detachment is accompanied by acceleration from the upper end of the capillaries. Therefore in our case the effect of screening of the working zone of molecules electric dissociation is not produced. All these processes proceed at minimal energy consumption, which is similar to analogous energy consumption. Moreover, the process of fuel gas production does not slow down as the process in Mayer's technology does therefore gas productivity of our method is significantly higher than that of this progressive analogue at similar minimal energy consumption.

Some technical and economical aspects of the new technology realization

In the near future production run of these high-effective electroosmotic generators of fuel gas from practically any liquids (including tap water), which are based on the proposed new technology, can be established. At the first familiarization level it is especially easy and economically appropriate to realize in practice a variant of the device of transformation of water-fuel emulsion into fuel gas. A prime cost of the production-run device for generation of fuel gas from water of 1000 m^3/hour productivity comes to approximately 1 thousand of US dollars. Consumed power of such an electric generator should come to no more than 50-100 Watts. Therefore such compact and effective fuel electrolyzers can operate practically in any motor car. As a result heat engines can work on any hydrocarbon liquid or even on habitual water. Mass application of these devices for the motor transport can cause immediate energetic and ecological perfection of the motor transport as well as designing of an environmentally appropriate and economical heat engine. Approximate financial expenditures for working out and designing of the device for fuel gas obtaining from water, and for bringing the investigation of the first testing device of 100 m^3/sec productivity to an experimental-industrial model come to about 450-500 thousands of US dollars. That contains the expenditures for projecting and investigation, for designing of the very experimental device and of a testing bed, which is necessary for approbation and engineering development of the device. The author is interested in business and creative cooperation with those companies, which can provide this project with investments to bring the device to the experimental-industrial model and introduce the perspective technology into practice.

Conclusion

Electroosmotic "cold" evaporation and dissociation of water and aqueous solutions through capillaries is a perspective way of highly productive fuel gas production at minimum of energy consumption.

References

1. V.D. Dudyshev New electric combustion technology is an effective way of solving of energetic and ecological problems // "Ecology and Industry of Russia" ("Ekologia I Promyshlennost Rossii") magazine #3 1997;
2. "New Energy News" magazine, May 1994. P. 1 – 5;
3. Stanly Mayer Method of production of fuel gas. Patent of the USA #4.936,961;

Conceptual Hurdles to New Millennium Physics

Donald Reed

E-mail: torsionpower@yahoo.com

Abstract

Time, space, energy and mass form the four-fold conceptual basis for gauging physical reality. The following paper, as a follow-up to a previous exposition [1] which focused on the necessity to change the current paradigm for time, continues in this same vein, but will also consider in greater depth the interrelations of this phenomenon with the other three yardsticks cited above. Again, information garnered from a wide variety of sources will be considered. We hope to provide, through just such a unique eclectic format, the beginnings of a possible fresh understanding of the workings of nature and perhaps ultimately furnish a conceptual basis for extending the structure of current physical theory to compatibly encompass the elements of a unified framework of physics and metaphysics.

Introduction

In the former article cited above, by investigating the following varied sources and research [2-17], we came to the inescapable conclusion that the phenomena of time and space are considerably more intimately related than is currently suspected by modern science, and will require a drastic re-working to fit the conclusions of the various evidence cited. Towards this end, the adoption of a more expansive paradigm for these yardsticks was suggested, which incorporates a fluid-field nature for time and space where both are derivatives of the fundamental ground-form of energy in flux or oscillation. These new approaches involve the inclusion of a Kozyrev-type of "substantial"(active) time-flow as opposed to the conventional "relational"(passive) concept of time, where it is used as a static parameter signifying duration. The substantial aspect of time presupposes that it is an essence which can and does affect physical processes, and that those same physical systems can cause a reverse action on time [8]. Such active fluctuating (deformable) types of time or space, implies the establishment of a revolutionary notion: positing completely *non-scalable* metrics for both time and space. This is in direct contrast to all contemporary orthodox models for physical reality, either in Einsteinean relativity (Special or General theory), quantum field theory, or even superstring/supersymmetric theory, all of which continue to consider both of these yardsticks from their relational (length or durational) standpoints. Consequently, all these formalisms require scalable metrics of some sort for their proper description (for refreshing viewpoints on this matter, see [7]). For instance, in the standard equations of quantum theory, time is regarded merely as an unchangeable static parameter. When we later consider applying these novel substantial aspects of time to the edifice of quantum theory, we will see that time must then become a "hidden variable". It will be seen that by considering time in this manner, a clearer and more tractable explanation of the inevitable probabilistic aspects of quantum theory, evidenced in the signature Uncertainty Principle, complementarity and non-locality will arise, without having to invoke the counter-productive and ill-conceived Copenhagen Interpretation.

Oscillatory Models for Time and Space

Upon accepting non-scalable metrics for time and space, we observe that a similar field nature for **energy** and **mass** must in turn also be postulated. Moreover, from the Killick description of how (sub-atomic) tachion-pairs operate [6], shuttling their energy back-and-forth in a free-wheeling but purposeful manner and creating oscillatory features of time and space by their action, we should consider the possibility of another unprecedented concept: a "value-motivated" energy might be at the foundation of the structural integrity of physical matter. This is conceivable when considering Killick's description of the tachion-pair dynamic/evolutionary cycle as necessarily including the 3-step process of observation – reflection – action (**trinitivity** of motion). Along the same lines, from the Smith book **The New Science** [5], we learn that a so-called "tempic field" energy exists, which could be described as the parent structure out of which our standard conception of "clock-time" (entropy changes) results. The tempic field is essentially a scalar (but not static) function which has vectorial nature only in terms of its distribution or *gradient* in space. Because of the nature of the tempic field, its derivatives – the electric and magnetic fields can operate on each other in a specific geometric/topological manner, to produce a local change in the *time-frame* of matter, to use a term coined by the entity Kryon [4]. This theoretical description of the alteration of inertial/gravitational mass and time-frame has possibly seen recent actual demonstration in the dramatic experiments of John Hutchison [10] and Rudolf Zinsser's "kinetobaric" effect [11], as well as unpublished research of both Wilbert Smith and Ken Killick with electric caduceus-wound coils [5,6].

New Models for Relativistic-Fluidic Vacuum-Structure and Possibility of its Manipulation

All of this evidence allows us to conclude that measurable changes in relativistic parameters of time, space, mass and energy might not only be a feature accompanying rapid uniform movement of physical objects (Special Relativity), or representative of large gravitating astrophysical objects (General Relativity). Indeed, by relying on current limited paradigms, contemporary physics may have missed ascertaining the possibility of being able to alter these same relativistic parameters by the artificial technological manipulation of the tempic (vacuum) fields of sub-atomic particles in stationary matter, by use of specific electromagnetic fields. Tom Bearden, for one, has articulated on these various "futuristic"-type technologies in his many writings over the years [15]. We have recently seen that some of the novel theoretical conceptions of the more visionary physicists such as David Hestenes on the zitterbewegung (vacuum "jitter") phenomenon exhibited by the electron [12-14], has finally caught up with some of these advanced ideas by presenting new mathematical demonstrations (using geometric Clifford-algebraic) manipulations of the Dirac equation. In summary, the Hestenes' study concludes that the Dirac wave function and its properties, including the Dirac equation and relations to physical observables such as energy-momentum, spin and position probability current, all possess heretofore obscured important geometric relations. These results imply that probabilistic features of the quantum theory of the electron/positron arise principally from the electromagnetic interaction of the accompanying zitterbewegung-spin field of these particles with the ambient dynamic vacuum. This is, of course, in direct contrast to the conventional received view that ascribes wave-particle duality as a property of matter that is completely independent of the nature of its interactions. This revolutionary geometric interpretation of electron dynamics incorporates in its model an electron spin which arises from a helical or spiral world-line in space-time. The essential unprecedented feature of the Hestenes' zitterbewegung idea is the association of the spin with the local circulatory-helical motion characterized by the *phase* of the electron wave function. Thus, we reach the conclusion that the complex phase factor of the electron wave function can be directly associated with an objective helical motion of the electron which is, in turn, a derivative of the zitterbewegung. One intriguing feature of this structure is a frequency of oscillation that is inversely proportional to the scalar radius of curvature of the particles' helical world-line. Moreover, the Clifford-algebraic analysis reveals that this oscillation frequency is identical to electron/positron *mass*, revealing a possible key variable particle mass-energy (*frequency measure*), which is in inverse relationship to particle size. It is clear that this new model has important classical implications not yet considered by established physics which nevertheless directly correspond to some of the key features of sub-atomic behavior enumerated above and previously. Here, we refer to the tachion-pair dynamics [1,6], and also to the microscopic dynamics of elementary particles revealed by the selected esoteric sources previously considered [1-5].

Another researcher who echoes many of these conclusions is C. Sano [21]. Sano built on the work of A.P. Smirnov [22] who postulated that Newton's Third Law of action/reaction actually modeled "screw" structures, implying that elementary particles (specifically electrons/positrons) possess chiral-spiral field configurations. Accordingly, Sano posited that all actions/reactions are transmitted between actors and reactors by parallel or perpendicular clutching of the rotating chains of electromagnetic spirals of the hidden electrons and positron-pairs of the vacuum. Also, similar to the Hestenes' development cited above, Sano postulated that the radius of the outer electromagnetic spiral surrounding the electron/positron pair can change, oscillating in size around either particle, and thus producing the particles' electric or magnetic character. Supporting his theory, Sano cites the key research of I.M. Shakhporanov [23] whose experiments claimed that magnetic monopoles were generated, by employing an electrical circuit based upon Moebius band topology. Some of the unusual phenomena demonstrated by this apparatus, tending to support the magnetic monopole hypothesis, was transformation of diamagnetic substances into paramagnetic, the ferromagnetization of normally non-magnetizable materials (graphite, etc.), acceleration of chemical reactions, acceleration or deceleration of the decay of radioactive materials, etc. Sano also claimed that extraction of energy from the vacuum was possible using rotational action/reaction using magnetic monopoles.

Referencing this research to the current exposition, we recall that the key element of Killick's hypothetical tachion-pair operation was the similar non-linear/non-orientable Moebius-type dynamic cycle which results in the corresponding deformation (compression or rarefaction) of time/space/energy parameters [1,6]. In this regard, the remarkable similarity of this model to the research of Sano and Shakhporanov may be much more than mere coincidence. In fact, later we shall examine new research outlining an apparatus (yet untested), which claims to overcome the Coulomb barrier in low energy nuclear reactions (LENR), and achieves nuclear fusion by producing local time-dilation of soft photons in proximity of the deuteron reactants.

Additional Key Russian Research

To motivate the reader's sensibilities towards consideration of some of these unorthodox ideas, we defer to the previous article [1] in which many of these notions and mentioned research are explored in greater depth, as well as consider the following corroborative research.

One intriguing example of such experiments is the work of V. Chernobrov, from which he has claimed to demonstrate both acceleration and deceleration of local time-rate, within a small spherical enclosed volume conducted since 1988 [18]. The experimental system to produce these effects, was a set of electromagnets, connected in series and parallel and installed inside the globe-shaped surface in several layers. In various arrangements including up to 3 to 5 of such layers, these electromagnetic working surfaces (EWS) of various diameters were installed inside each other (similar to the Russian toy doll "matrioshka") with the maximum EWS diameter was about 1 meter, and the minimum (internal) diameter of 115 mm. With this configuration, Chernobrov claimed to measure small but detectable deceleration (-30 sec/hr.) and acceleration (+30 sec/hr.) of time within the sphere to time monitored outside its field of influence. One interesting difference was noted between the slowing down vs. the speeding up of time. The deceleration occurred considerably more smoothly and steadily whereas with acceleration, sharp discontinuous jumps were observed. These instabilities accompanying time-acceleration was observed in connection with cycles of the moon, diurnal fluctuations, and also operator presence. Chernobrov also noticed another phenomenon that also was reported in connection with the alleged legendary Philadelphia experiment, where matter in different time-frames apparently exhibited strange relative optical phenomena [see 1,4]. Specifically, the human eye in the time-frame exterior to the experiment perceives matter within the time-frame created by the apparatus as either transparent or surrounded by a vague white mist. We note with interest that transparent "shimmering" effects of substances in the target area were also occasionally a feature of Hutchison effect [10].

Academician A. Chernetsky produced what he termed a self-generating discharge (SGD) in a plasma that exhibited longitudinal energy density waves from a structured vacuum. Some unusual effects noted were over-unity energy generation (C.O.P. > 1), as well as change of the electrical conductivity of matter (reduction of the resistance of resistor) placed between the capacitor plates of the SGD. A local structured vacuum was definitely produced since the resistance value remained unchanged even when the generator was turned off. Here reference must be made to the Kozyrev experiments [1,8], in which a type of similar *memory* process was activated ostensibly within the vacuum. Specifically, it turned out that in experiments with a vibrating torsion balance (or pendulum), at points of support the emerging additional forces did not disappear when the irreversible process (vibration) was stopped, but remained in the system for an appreciable time. The SGD plasma device may have also produced a detectable change in local time-flow rate as evidenced by decrease in frequency of a quartz oscillator placed in the discharge of the generator. The Chernetsky generator thus possibly caused anomalies in time-rate at a local space position. There were also basically anecdotal phenomena where psychic ability of personnel in proximity to the SGD was apparently enhanced [19]. The latter effects are not inconceivable once the intimate connection of consciousness to space and time is ascertained, from consideration of not only hypothetical "value-motivated" tachion-pair dynamics, but from the more prosaic studies as we shall see next.

Pathological Studies in Time-Perception

We can motivate further understanding for the claimed intimate connection of time to space by considering the results of a recent study of patients possessing a disjointed sense of time by Metod Saniga of the Slovak Academy of Sciences [20]. Saniga discovered the brain is hard-wired to perceive space and time as interconnected. Specifically, time pathology is apparently always accompanied by space pathology, in a sense that space either loses dimensions or acquires other dimensions. To quote Saniga "When time seems to stop, people often feel as if space becomes 2-dimensional. On the other hand, when the subject feels they perceive the past, present and future all at once, they simultaneously have the impression that space has infinite dimensions". This phenomenon is apparently not culturally endemic since Saniga illustrates that both pathologies cross cultural lines, evidencing similar studies cited from Italian, German and English psychological journals. In his report, Saniga combined mathematical models ("pencil-conics") and pathology reports of schizophrenic, drug-induced and other abnormal perceptions of time. His current work also encompasses studies of near-death experiences. He found that most of those who claim to have allegedly crossed over to the "other side" and back, tell similar tales. For that brief moment of near-death, the universal subjective experience of all individuals in this state of consciousness is that time loses its meaning. Although such evidence can at best be considered as anecdotal, since subjective conscious experience transcends the possibility of scientific proof, nevertheless perhaps these results from pathological and near-death studies also give us a hint towards new conceptions of time and space that necessitate both phenomena to be inextricably linked.

Investigation of the Possible Internal Properties of Time

One inescapable conclusion that results from all the above varied studies, either scientifically or psychologically based, and which cries out for future experimental verification, is the possible existence of an internal structure for time. Such a structure also implies the necessity for constructing a new edifice for physical reality that incorporates non-scalable metrics for the four fundamental yardsticks. We shall first investigate supporting evidence for internal time structure and subsequently the various implications of space, time, energy or mass that is non-scalable.

First, accepting an internal time structure, we are led to conclude that in quantum theory, time must then attain "hidden variable" status. One recent study that brilliantly articulates this notion, is the thought-provoking paper by X. Chen [24]. By positing 3 time variables as quantum hidden variables, Chen derives the Dirac equation classically. Moreover, he demonstrates that the non-intuitive property of "spin" of an electron or positron , arises naturally as a topological property of 3-dimensional time + 3-dimensional space. By extrapolating from this model, Chen then concludes that the inherent probabilistic aspects of quantum theory itself, as evidenced in wave-particle duality, Uncertainty Principle, quantum non-locality ("entangled" particles), etc., can be interpreted as the behavior of a single particle in 3 + 3 time-space. Chen postulates that the three dimensions of time geometrically form a "time sphere", with a generic point on the sphere possessing the following 3 coordinates: time radius, and two "time angles". The mathematical formalism arising from the process of stereographic projection from the north pole of this sphere, having radius of one-half, to any point on the spherical surface, incorporates both the positive-energy (north hemisphere) and negative-energy (south hemisphere) solutions to the Dirac equation (see Fig. 1 for clarification; note Z is a four-component spinor wave function). Also, by considering the evolution of a single particle, from the standpoint of this model of internal time-space, we can derive a picture of different paths on the time sphere of different weights, and on each path causality is satisfied. Through this process, Chen derives a purely classical explanation of the originally quantum-mechanically canonized Feynman Path Integral concept. This result emerges since each path from the time-sphere center to the surface corresponds to each Feynman path, and the surface of the sphere corresponds to the "surface" of the wave function. Chen also classically explains the processes inherent to Bose-Einstein-condensation (BEC) and superconductivity with this model. In such phenomena, two particles with the same spatial coordinates, cannot possess the same time angle, and thus will not have any interaction with each other and occupy the same quantum state as evidenced in BEC and superconductivity. Finally, Chen's solutions of the Dirac equation for a free particle correspond to Hopf bundles in monopole theory, and each Hopf fiber corresponds to each plane wave with different momentum states. This picture corresponds to wave packet diffusion in quantum theory. When a particle is in a fixed momentum state, each space point can contain only one Hopf bundle (one time angle), and various different time angles will be distributed in the whole space but with the same Hopf bundle. Consequently, that particle can be found everywhere. This picture appears to describe the Uncertainty Principle classically using 3 + 3 dimensional time-space. Along similar lines R. Kiehn [25] has also recently underscored the important but overlooked classical connection between spinors, minimal surfaces and the Hopf map.

Recently, Chen's work has been further developed by H. Kitada [26] who, by introducing both 3-dimensional time and energy operators, sheds new classical light on the uncertainty relation that holds between these two parameters as well. Moreover, Chen is not the only reseracher to derive the Dirac equation classically. R.A. Close has published a paper which presents a classical mathematical description of circularly-polarized waves in 3-dimensions [27]. It turns out that these chiral space waves are satisfied by a four-component wave function which satisfies a Dirac equation. Furthermore, much like the Hestenes' zitterbewegung interpretation of quantum theory, the term normally associated with electron mass in the Close equation, introduces a rotation or oscillation frequency of the propagation direction.

Classical models for nuclear processes have also recently been proposed which incidentally also imply internal time structure. Notable among these, C. Cagle claims to have developed a fusion energy device employing LENR based upon such a classical theory [28]. Normally, Deuterium nuclei are diffuse in momentum space (having high relative velocities). The process Cagle outlines produces a compactification of their momentum space (low relative velocities), so that the deuterons develop a common de Broglie wavelength that is greater than or equal to the inter-particle distance. This is claimed to be accomplished by passing soft x-ray photons near the region of two deuterons in a lattice structure of Lithium Deuteride. **The soft x-ray, when absorbed , produces a region of total *time-dilation*, causing two effects: first, ionization of atoms by strongly repelling any associated electrons in the area, and secondly and most importantly, the physical extent of the time-dilation causes two adjacent normally repelling Deuterium nuclei to overlap in a common momentum space; that is, their relative velocities achieve very low values.** Consequently, due to the time dilation, for a very short time their common de Broglie wavelength exceeds their inter-particle distance. Thus, they become strongly attractive and undergo nuclear fusion.

Cagle also claims to clarify misconceptions about the fusion process in a thermonuclear bomb. He maintains that it is not due to extreme kinetic energies (high temperatures) of particles as supposed in current paradigms, but due to nuclear processes caused by absorption of a time-dilated soft photon flux, produced by Compton scattering of hard x-rays from a fission igniter, passing through a foil of depleted Uranium. Again, just like the controlled nuclear fusion process, this causes overlapping of deuterons in momentum space (low kinetic energies) for a few pico seconds, and subsequent fusion with normal exothermal processes ensuing.

Furthermore, as a fundamental component of his LENR fusion device, which ostensibly produces time dilation and controls energy production from the fusion process, Cagle incorporates a topological field structure termed an electromagnetotoroid. The toroidal field current oscillates between toroidal and poloidal modes in a dynamic cycle which is amazingly virtually identical to Killick's description of the oscillation dynamics of a single unit toroidal tachion [1,6](also see Cagle's website for animated simulation).

Regarding dynamics of astrophysical structures, Cagle further claims that the stellar jet core star of HH30 is also a gigantic electromagnetotoroid. During the poloidal current mode, vast quantities of matter are produced and ejected along the poloidal axis when the mode changes from poloidal to toroidal. This leads one to consider the possibility that the source of superluminal gamma ray bursters, which have heretofore mystified astrophysicists and defied explanation in terms of current paradigms, might be the result of large- scale abrupt changes in the time-frame of celestial vacuum regions, subsequently causing a boost in light velocity.

Implications of Non-scalable Metrics

When postulating fluid-field natures for the four yardsticks of physical reality: time, space, energy and mass, we must assume the existence of a *non-scalable* metric for our vacuum structure. Such a non-scalable vacuum, absent of a defined reference frame, must of necessity possess a *dual* nature. That is, the primary feature of a dual vacuum structure implies that infinitely small quantities must be treated on an equal footing with those that are infinitely large. In other words, information encoded into infinitesimal elements, is also instantly present in the unbounded infinite extents as well. This is tantamount to the existence of a holographic information encoding and transmitting vacuum field, which would structure events in space and time as a non-Markovian chain. In a non-Markovian chain of events, the prediction about the next link in the chain, requires a knowledge of **all** links, not just the one preceding it.

Metod Saniga's research into mystical perceptions of superconscious reality appears to support such a primordial non-Markovian holographic mapping of time and space. To quote from a transcript of such a vision: "I wake up in a whole different world…a different space…This space was distinct from the one we all know. It had different dimensions, everything contained everything else. One was situated in a state of being in which the 'will be' (future) and the 'vanishing' (past) were already included, and this being was my consciousness. It contained it all. The 'being contained' was present very vividly in a geometric way in the form of circles of different sizes which again were all part of a unity since all of the circles formed exactly one circle. The biggest circle was a part of the smallest and vice versa."[20].

Such a structure of time and space can best be topologically encoded in a non-orientable structure such as the Moebius band or Klein bottle. Similar to the data presented in the above mystical revelations and in the previously treated Killick tachion-pair dynamics [1,5], in such configurations "inside" and "outside" lose their meanings and meld into one another; ordinary dichotomic relations distinguishing thesis from antithesis are sublated and supplanted by a higher unity defining the evolutionary process of the system. In the mathematical edifice abstracted from such a vacuum structure, duality rather than exclusivity of field structures holds sway. This in turn, presupposes a static-dynamic triality of fundamental field structures, characterized by the three magnitudes: zero, infinity and the mediating factor of unity, the latter being an indication that each of the other two field extremes are in perfect balance. Accordingly, we will assume that the stability (coherency) of non-scalable fields results whenever more than half their reality is in common.

Further development pertaining to a specific algebraic-geometric structure that would inherently incorporate this field model, will be carried out in a future paper. Nevertheless, the interested reader can skirmish on this frontier by consulting the relevant recent references to a Clifford-algebraic structure which encodes the projective duality of "space" and "counterspace" into a model for mechanics on the quantum level [29]. This promising model accounts classically for the phenomenon of quantum non-locality. It also heralds a new version of quantum field theory which, by treating electron-positron pairs as topologically non-orientable, is free of the detrimental divergences in self energy and charge, thus obviating the necessity for the ad hoc prescription of renormalization so endemic to current theory [30].

However, due to the limited scope of the current treatment, here we shall only make general comments on the significance of the above-mentioned concept of field triality. Although this model may appear foreign to current physical theory, we underscore that exactly such a system was delineated by W. Smith in the book, The New Science [5]. Observing, as we have [1], that this book in some parts of its exposition can cause confusion, we carefully choose the following selected passages that are most instructive in clarifying the notion of non-scalable metrics and in pointing the way to new paradigms (my comments in parentheses). General comment on non-scalable metrics: "Unity is half way between zero and infinity and always remains the fulcrum about which all other values regardless of scale pivot". On the electric field: "Between the two limits of zero and infinity and through unity there is divergence which we recognize as the electric field, but with the exception that there is no point charge at zero. Our awareness establishes the charge at radius unity with exactly half of it being 'inside' and the other half being 'outside'". On tempic field dynamics: "The tempic field being purely scalar in nature merely contributes to the manner in which changes can occur in the system. If, however, through some characteristic of configuration, the tempic field within a composite particle displays some irregularity, this will manifest as a 'vibration', and if the magnitude of the vibration is sufficient as to cause the interpenetrating fields to exceed the half-and-half point , the whole system will come apart: radioactivity. When a system does come apart, those field which were coherent and interpenetrating in the correct proportion will remain so and take off down the tempic field gradient and become radiated energy". On velocity of light/Planck's constant: "Within the universe we perceive, we are reasonably satisfied that the maximum value of all the coherent fields involved in our particles and radiated energy have the same value. Planck's constant is the numerical expression of this value, and the velocity of light is the expression of the numerical value of the (local) tempic field intensity". On nuclear fusion: "If a region is selected in which there are two fields of the same kind, same magnitude, same direction, such that very nearly half the reality of each is within the region, then the two fields are just on the verge of becoming coherent. If the fields are not coherent, the total energy in the region is the sum of the energies of the two fields; i.e., twice the square of the field intensity of each integrated over the region, or twice the energy of the two fields incoherent. This represents the 'packing energy' of bits and pieces of atomic nuclei, and also points the way to the precipitation of energy out of the cosmic background (vacuum engineering)". Notice how this description, written 45 years ago, of how a tempic field gradient (time-stress) operates between atomic nuclei to produce coherency (half-in half-out condition), perfectly parallels the recent Cagle view, which claims the fusion process takes place only as a function of time-dilation [28]. Finally, on gravitation: "With the exception of skew electric fields (electric field possessing tempic field gradient) all the other fields of an aggregate mass may be considered as static because they have no component of the tempic field at right angles to them. The skew fields (spiral-helical?) on the other hand, are dynamic because they do have a quadrature tempic field component. Furthermore, skew fields are largely incoherent (gravitation cannot normally be shielded) simply because the usual almost random orientation precludes their meeting the half-in requirement of form to become coherent. Since the gravitational field is due to the skew electric field, or stated differently, to the induction from the 'motional magnetic field' (longitudinally moved magnetic field), this is the logical region to explore for the mechanism by which fields may be produced to combine with the gravitational field (of Earth) to produce a resultant more to our liking (alteration of local gravitational potential)".

This last quote describes the gravitational field of a mass as a residual field phenomenon, similarly to the Sakharov and Puthoff conclusions which attribute gravitational and inertial mass to the interaction of accelerated or gravitating matter with the zero-point-vacuum-fluctuations (ZPF) causing a vacuum reaction force [31]. Moreover, similar to the Hestenes' claims, in their recent papers [32], Haisch/Rueda/Dobyns propose that, via this new interpretation of inertial mass as an acceleration-dependent electromagnetic (Lorentz) force, that a former postulate of quantum mechanics appears to be derivable classically via the interpretation of rest mass as the energy of the ZPF driven zitterbewegung; that is, the de Broglie wavelength of a moving particle, may be derived from Doppler shifts of the Compton frequency oscillations associated with zitterbewegung that occurs when a particle is placed in motion.

For further studies of physical theories with non-scalable metrics, the reader is directed to the excellent work of A.A. Nassikas [16]. Like few theories hitherto postulated, Nassikas posits a fundamental probability density function for vacuum energy, out of which oscillatory-deformable sub-atomic level physical aspects of time and space then emerge. This is the reverse to most contemporary paradigms, which view energy as a derivative of matter, time and space, even at the quantum level. Finally, A. Frolov has used Nassikas' theory to explain over-unity energy generation that has been demonstrated in some LENR electrolytic cells of the Pons-Fleischmann variety [17].

Editor: Actually Prof. Nassikas wrote in this patent that the idea which is described in the patent was proposed by Alexander Frolov. In 1996 these questions were discussed by us in Saint-Petersburg. Specifically, a Palladium cathode over-saturated with protons produces an imbalance of the vacuum engine in this area, with an accompanying change in the local time-rate and subsequent breakdown in the Coulomb barrier of the nuclei as per schemes articulated by C. Cagle [28] and T. Bearden [15].

Conclusions

In this journey through the latest research and speculations involving new interpretations of the four physical yardsticks: time, space, energy and mass, we hope the reader and technical specialists/theorists have been inspired to continue research along the lines suggested in the above dissertation. In this regard, certain open-ended questions present themselves that focus squarely on the viability of technological future developments in the new energy field. First, could the missing element in our eventual understanding of all devices featuring documented over-unity energy generation, change in mass (gravitational potential), LENR-based transmutation of elements and the amelioration of radioactivity, etc., be due the alteration of local time-flow in sub-atomic vacuum fields within the apparatus? Secondly, could the frustrating feature of the capriciousness of over-unity power and the like to yield to testable replication in many cases, be due to currently unperceived uncontrollable aspects of time and the vacuum which must be addressed in order to solve these problems? Thirdly, could incorporation of non-orientable topological structures such as the Moebius band, Klein bottle, etc. in new models of electrodynamic field structure, be instrumental in finding the answers to both of the previous questions? With sufficient probing for possible answers to such questions, the present author has the firm belief that new paradigms are certain to soon result that will crystallize these thoughts into viable scientific hypotheses subject to experimental testable verification. As an offshoot and definite bonus of this process, possibly a new more expansive understanding of the role of the vacuum in both nature and consciousness will be in the offing.

References

1. D. Reed, "A New Paradigm for Time: Evidence from Empirical and Esoteric Sources", Explore!-for the Professional, vol. 4/5 July/Sept., 2002.

2. J. Roberts, unpublished remarks as transcribed by W.H. Kautz.

3. J. Roberts, The Seth Material, Prentice-Hall, Inc., 1970, pp. 292-298.

J. Roberts, The "Unknown" Reality, Prentice-Hall, Inc., 1986.

4. L. Carroll, Alchemy of the Human Spirit (Kryon Book III), The Kryon Writings, 1995.

L. Carroll, Letters From Home (Kryon Book VI), The Kryon Writings, 1998.

L. Carroll, Passing the Marker (Kryon Book VIII), The Kryon Writings, 2000.

5. W. Smith, The New Science, Fern-Graphic Pub., 1964.

6. D. Reed & K. Killick, "Tachion Energy Theory", Energy Unlimited, 1978-1983.

7. A. Correa & P. Correa, "Consequences of the Null Result of the Michelson-Morley Experiment: The Demise of the Stationary Aether, the Rise of Special Relativity and the Heuristic Concept of the Photon", Infinite Energy, no. 38, July-Aug. 2001, pp. 47-64. A. Correa & P. Correa, "The Sagnac and Michelson-Gale Experiments", Infinite Energy, no. 39, Sept.-Oct. 2001, pp. 32-49.

8. N. Kozyrev, "On the Possibility for the Experimental Investigation of the Properties of Time", Time in Science and Philosophy, Prague, 1971, pp. 111-132.

N. Kozyrev, "An Unexplored World", Soviet Life, 1965.

A.P. Levich, On the Way to Understanding the Time Phenomenon (Part II: The Active Properties of Time According to N. A. Kozyrev), World Scientific, Singapore, 1996.

9. D. Reed, "Torsion Field Research and Implications for New Physics and Energy Technologies", Jour. New Energy, vol. 4, no. 2, Fall, 1999, pp. 151-164.

10. J. Hutchison, "The Hutchison Effect Apparatus", Proc. of the 1st Symp. On New Energy, Denver, May, 1994, p. 199. J. Hutchison, "Inside the Hutchison Effect", Extraordinary Science, vol. III, issue 4, Oct-Dec. 1991, pp. 23-26.

11. T. Valone (ed), Mechanical Energy from Gravitational Anisotropy, Integrity Research Institute, Washington, DC, 1996.

W. Peschka, "Kinetobaric Effect as Basis For a New Propulsion Principle", Raumfahrt-Forschung, Feb, 1974 (in German); trans. By D. Reed, Infinite Energy, vol. 4, issue no. 22, 1998, pp. 52-56.

D. Reed, "Translator's Analysis and Comments on the Zinsser Effect Device, Infinite Energy, vol. 4, issue no. 22, 1998, pp. 57-59.

12. D. Hestenes, "Quantum Mechanics from Self-Interaction", Found. Phys., vol. 15, no. 1, 1985, pp. 63-87.

13. D. Hestenes, "The Zitterbewegung Interpretation of Quantum Mechanics", Found. Phys., vol. 20, no. 10, 1990, pp. 1213-1232.

14. D. Hestenes, "Zitterbewegung Modeling", Found. Phys., vol. 23, no. 3, 1992, pp. 365-387.

15. T. Bearden, Extracting and Using Electromagnetic Energy from the Active Vacuum, Association of Distinguished American Scientists, Huntsville, Alabama, 2000.

T. Bearden, "Giant Negentropy in the Common Dipole", Proc IC-2000, St. Petersburg, Russia, 2000(in Press)

16. A.A. Nassikas, " The Hypothesis of the Unified Field and the Principle of Its Dual Interpretation", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.

17. A. Frolov, "The Work Created by Means of a Potential Field", Proc of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.

18. V. Chernobrov, "Experiments on the Change of the Direction and Rate of Time Motion", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996.

19. A.V. Chernetsky, "Processes in Plasma Systems with Electric Charge Division", Deutsche Vereinigung fur Schwerkraft-Feld Energie e. V. no. 25, Mar. 1991, no. 27, Feb. 1992.

20. M. Saniga,"Unveiling the Nature of Time-Altered States of Consciousness and Pencil-Generated Space-Times", Int. Jour. Of Transdisciplinary Studies, vol. 2, no. 2, pp. 8-17.

21. C. Sano, "Twisting and Untwisting of Spirals of Aether and Fractal Vortices Connecting Dynamic Aethers", Jour. New Energy, vol. 6, no. 2, Fall 2001, pp. 178-184.

22. A.P. Smirnov, "Hyper-Physics - Physics of Reality", Kneeling Before the Truth, pp. 62-100, Materials of the 2nd Int. Scientific Clarification Conf.-2000, St. Petersburg, Russia,(2000).

23. I.M. Shakhparonov, "Interaction of Kozyrev-Dirac Emanation Methods of Detection and Interaction with Matter", Proc. Of III Int. Conf. On Problems of Space Time and Gravitation, Russian Academy of Sciences, St. Petersburg, Russia, 1996, pp. 175-187.

24. X. Chen, A New Interpretation of Quantum Theory – Time as a Hidden Variable, Apr. 2001, xxx.lanl.gov/abs/quant-ph/9902037.

25. R.Kiehn, Spinors, Minimal Surfaces, Torsion, Helicity, Chirality, Spin, Twistors, Orientation, Continuity, fractals, Point Particles, Polarization, the Light Cone and the Hopf Map, www.uh.edu/~rkiehn.

26. H. Kitada, Three Dimensional Time and Energy Operators and an Uncertainty Relation, July 200, xxx.lanl.gov/abs/quant-ph/0007028.

27. R.A. Close, "Torsion Waves in Three Dimensions: Quantum Mechanics with a Twist", Found. Phys. Lett., vol. 15, no. 1, Feb. 2002, pp. 71-83.

28. C. Cagle, (2001), www.singtech.com/.

29. O. Conradt," Mechanics in Space and Counterspace", Jour. Math Phys., vol. 41, no. 10, 2000, pp. 6995-7028.

30. B. Fauser "Projective Relativity- Present Status and Outlook", Gen. Rel. Grav, vol. 33, 2001, pp. 875-887.

31. H.E. Puthoff, Phys. Rev. A, vol. 39,(1989), p. 2333. B. Haisch , A. Rueda & H.E. Puthoff, Phys. Rev. A, vol 49,(1994), p.678.

32. B. Haisch, A. Rueda & Y. Dobyns, "Inertial Mass and Quantum Vacuum Fields", Annalen der Physik, (Sept. 2000). B. Haisch & A. Rueda, "On the Relation Between a Zero-Point-Field-Induced Inertial Effect and the Einstein-de Broglie Formula", Phys. Lett. A, vol. 268, (Apr. 2000), pp. 224-227.

Resonant Phenomena Occurring in Alternating Current Circuit

Alexander V. Frolov

Tel/fax: 7-(812)-380-38-44
E-mail: office@faraday.ru

There is a classical (according to manuals) notion that resonant phenomena can not influence on efficiency of a transformer or a motor as in a capacitive circuit or in an inductive circuit power produced at minimal power at the input of an oscillation circuit is reactive (let us remind that active power is measured at active resistance).

At the beginning of the last century this aspect was viewed differently. Advantages of a resonant case in alternating current circuits were used in practice. Let us appeal to a rare book by J. Claude-V. Ostwald named "Electricity and its applications by popular language" (I.N. Kushnerev Press, Moscow, 1914, p. 463).

"The phenomenon proceeds in a corresponding electric circuit as well as it occurs in hydraulic model: if self-induction and capacity parallel connected with each other are under influence of an alternating electric propulsion force then the total current coming through the system is equal to the difference of currents which pass two certain paths.

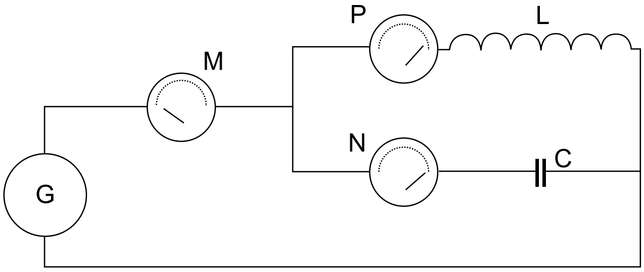

Fig.1

… join ammeters to the general circuit (M) and to either path (P and N). If P demonstrates 100 Amperes and N demonstrates 80 Amperes then the total current will be equal to 20 Amperes instead of 180 Amperes. Thus alternating current has its own "summary" so we should consider its condition. …capacity introduction compensates self-induction action in some way… let us begin to change the self-induction by moving a core into. What will happen if the current coming through the coil achieves 80 Amperes, i.e. is equal to the current observed in the path with a capacitor?

As you guess, since the total current is equal to difference between currents passing the paths then now it will be equal zero. It is an incredible case:the device produces zero current dividing into two paths and in each of these paths current is equal to 80 Amperes. It is a good example for the first acquaintance with alternating current, isn't it?"

Andrey A. Melnichenko (Moscow) is one of modern investigators who research this phenomenon. Any motor of alternating current can be considered as inductance. A circuit consists of motor coils and some capacitor connected in series with the motor winding. Thus if we take the circuit, consisting of motor coils and a capacitor which is connected in series to motor winding and adjust it in resonance, then mechanical power occurring on the motor shaft is produced at zero (minimal) power consumption of an alternating current source.

Melnichenko applied a simple method of voltage increase by meanse of resonance: he succeeded in obtaining of normal voltage for operation of standard motors of 50 Hz 220 V from a source of 50 Hz 110 V and 70 V. The circuit consumption for overcoming of its active resistance (of coils) can be considered as insignificant. Low frequency currents require a large capacitor. Nevertheless even at higher frequency, for example at 400 Hz, the system can be compact and effective. This method is worth to be applied in a scheme consisting of an alternating current motor in resonance mode and an electric generator which has a stable load. Change of load causes change of rotation speed therefore the system requires to be readjusted for resonance.

Experiments with powerful alternating current motors (about 100 Amperes as J.C. Ostwald wrote) working in resonance mode should demonstrate all advantages of the resonance mode application.

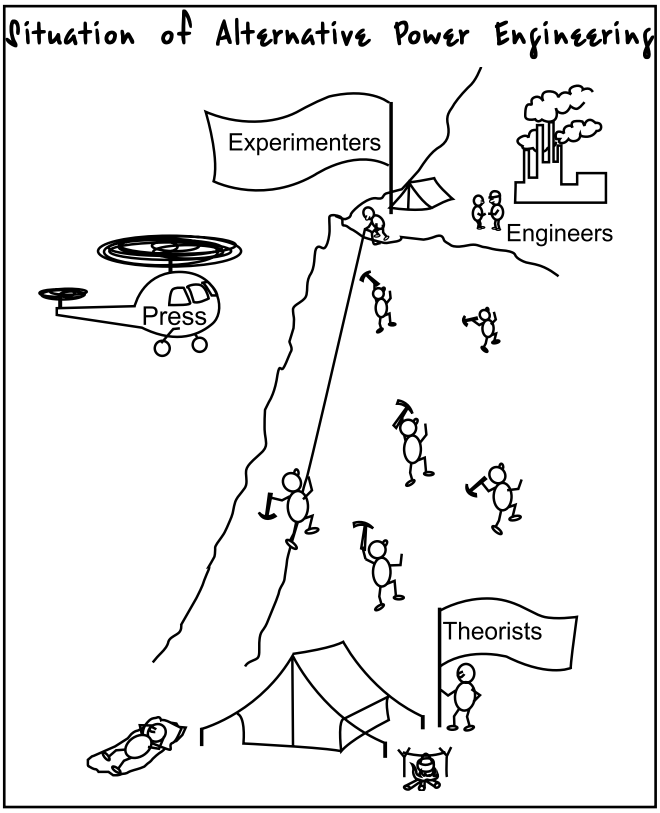
Situation of Alternative Power Engineering
Experimenters
Engineers
Press
Theorists

Space Power Generator and Paramahamsa Tewari Update

Submitted by Toby Grotz, USA

http://www.tewari.org

A new generation of Space Power Generators is being built by Paramahamsa Tewari. The new design is expected to achieve substantial gains over previous models which has been measured over unity by experts in the utility industry as well as by independent consultants. Details are presented on the web site http://www.tewari.org.

From the Electron to a Perpetual System of Motion

Paramahamsa Tewari, B.Sc. Engg

As is well known, an electron, despite high-speed interactions with electric and magnetic fields and other particles of matter, remains unaffected structurally—maintaining its mass, charge, inertia, and locality after the interactions. This single fact itself is sufficient to postulate on the existence of some unique universal entity constituting the electron, such that the structure explains the known properties and behavior of the electron – as to how it sustains the collisions with the other particles and, yet, remains permanently indestructible without any reduction in the quantities of its basic properties. A theory that explains this enigmatic fact is framed and described in detail elsewhere. It reveals that the electron is a perpetually rotating system, a space-vortex (vortex of absolute vacuum) that through the process of motion in electric current, and interaction with external magnetic field can also lead to the development of a system partially violating Lenz's law and, thereby, capable of achieving perpetual motion in a machine. Such a conclusion has a new theoretical basis, as well as experimental confirmation – briefly outlined in this paper.

The basic understanding of the interaction of electric current with magnetic fields requires deeper knowledge of the fundamental nature of the electric current as well the magnetic field; this, in turn, leads to an inquiry on the nature of "electric charge" and its distribution in the electron structure. The origin of electron mass and charge, electrostatic and electrodynamic forces are described below with the vortex structure of electron, which further facilitates explanation of atomic structure and, thereafter, an over-unity efficiency generator. Due to brevity, predominantly qualitative-descriptions of the physical processes involved have been presented.

The full theory and diagrams are presented in From the Electron to a Perpetual System of Motion at http://www.tewari.org.

Universal Principles of Space and Matter (A Call for Conceptual Reorientation)

A new book, *Universal Principles of Space and Matter* has just been published by Paramahamsa Tewari. Arangements are being made with the publisher to provide these books outside of India. The price is expected to be about $30 depending on shipping costs from India.

A new theory of matter, that confronts with the existing concepts of space in the contemporary physics.

The basic phenomena dealt with are:

* Structural interrelationship between space and matter
* Origin of mass, inertia, and electric charge
* Creation of the electron from space

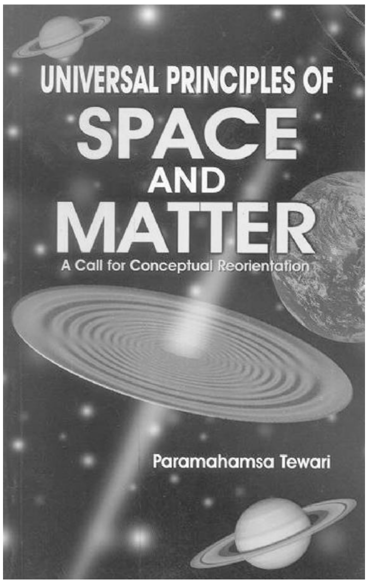

* Discovery of new fundamental equations on mass and charge
* Derivation of the presently known universal constants (gravitational, Coulomb's constant, dielectric, magnetic permeability, electron charge, electron mass, Planck's constant etc.) with the postulate of a single universal costant – the speed of light in the absolute vacuum
* Discovery of an inward force in nuclear structure against the Coulomb repulsive forces, hitherto unknown
* Relationship between light and gravity
* Theoretical derivation of the surface gravity of the Earth, Sun and the planets
* Discovery of electrical repulsive forces between the Sun and the planets, unknown in celestial mechanics
* Derivation of the planetary orbits
* Prediction of the continuous creation of universal matter at glactic centers and existence of electrical forces of interaction btween the stars and galaxies, so far least suspected
* Identifying the fundamental particle of matter
* Revelation of conceptual errors in the modern understanding of the basic nature of light
* Vindication of Einstein's conclusion of the speed of light as the limiting speed
* Vindication of Descartes' principle of property-less ether.

New Reviews of P. Tewari's work

New reviews of P. Tewari's work have been received from Dr. John A. Wheeler and Prof. Umberto Bartocci. A copy of Universal Principles of Space and Matter was forwarded this summer to Dr. John A. Wheeler at his summer residence in Maine.

Dr. Wheeler saw the development of modern physics while working with Neils Bhor at age 27 and was a colleague of Albert Einstein's at Princeton. Credited with being the "father of the Black Hole Theory", Dr. Wheeler is well schooled in quantum mechanics and relativity.

After receiving his copy of Universal Principle of Space and Matter Dr. Wheeler called the book "a unique gift". Further reviews of P. Tewari's work by Dr. Wheeler can be found at: http://www.tewari.org/Critical_Reviews/critical_reviews.html.

Prof. Umberto Bartocci

Universita' di Perugia
Dipartimento di Matematica e Informatica Via Vanvitelli
06100 PERUGIA - ITALY
Tel.: (39)-075-5002494
E-mail: bartocci@dipmat.unipg.it

This is a very singular book, in front of the current paradigm of contemporary Physics, a "cartesian-inspired" work which calls for reorientation in the foundations, by wisely warning that:

The conclusion of the modern physics that absolute space, time, simultaneity, and space filling media are discredited ideas is certainly premature (p. 178).

The books deals with arguments such as: Discovery of Charge and Mass Equations; Fundamental States of Cosmic Energy, Fields and Forces; Gravitation; Universal Constants; Motion of Electron; Atomic Structure; Light; Creation of Cosmic Matter (about Tewari's physical conceptions see also the second section of this Episteme's special issue http://www.dipmat.unipg.it/~bartocci).

JNET

New Energy Technologies collection of articles Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

raum & zeit
Laser harmonisiert Blut und Wasser

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory - Hydrogen-Helium Energetic
- Gravitation theory - Information technologies
- Mathematic Biology - Telecommunication technologies
- Gravitation energetics- Transport technologies

Rotating device powered by permanent magnets

Minato Wheel

http://fdp.hemsida.net

Experiments made in Sweden

by Eric Vogels

Read about launch vehicle "Lightcraft" in the review "Commercial exploration of space"

http://www.lightcrafttechnologies.com

"Lightcraft" is made
for space flights
by means of megawatt
laser beam

Lightcraft Technologies, Inc.

Ball Lightning control

Experiments by Russian scientists

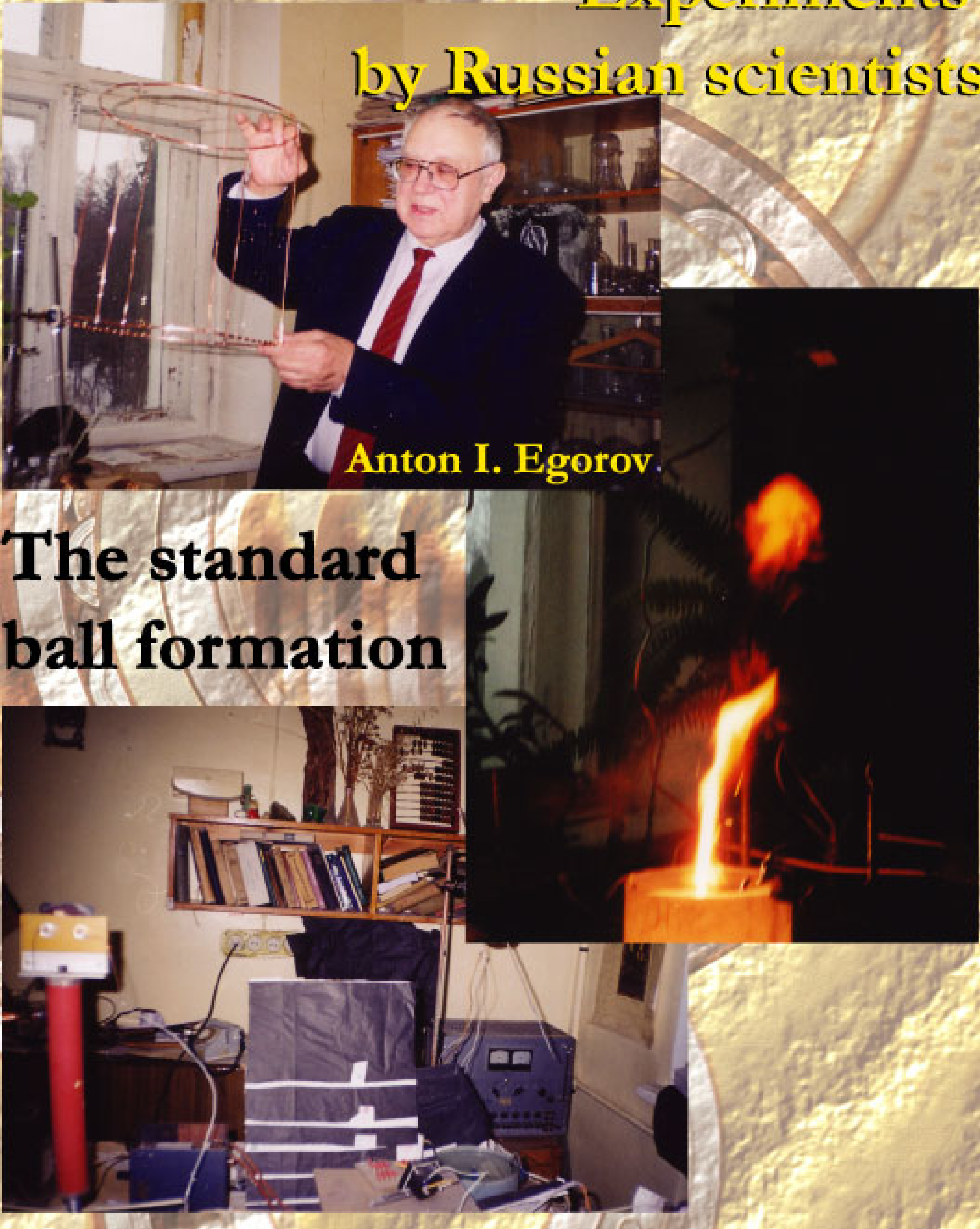

Anton I. Egorov

The standard ball formation

Read in this issue!

Tilley Foundation, Inc.

http://www.tilleyfoundation.com

K. Tilley
Electric Vehicle
(TEV) - automobile
of the future generation

ISSN 1684-7288

New Energy Technologies

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

Published by Faraday Laboratories Ltd. *Issue #2 (11), 2003*

Research on advanced aerospace systems and new energy sources

Commercial exploration of space

Also in this issue:

- New experiments on magnetic motors
- Ball lightning in laboratory environment
- Alternative aircrafts and space propulsion systems

NEW ENERGY TECHNOLOGIES #11

1. Minato Wheel. E. Vogels (Sweden)
2. Mechanism of Drive-Free motion. S.A. Gerasimov (Russia)
3. Production of Ball Lightning in Laboratory Environment. Report by A. Pashova (Russia)
4. Electric Discharge Produced into Air Half-Space. G.D. Shabanov, O.M. Zherebtsov (Russia)
5. Free Energy. A.V. Frolov (Russia)
6. Some General Properties of Matter and Energy Sources. E.D. Sorokodum (Russia)
7. Commercial Exploration of Space. Review by O. Leontyeva (Russia).
8. Prospects of Hydrogen Energetics. Ph.M. Kanarev (Russia)
9. Heater Producing Healthy Effect. A.L. Belyaeva and others (Kirghizia)
10. Inertial Propulsion Device. V.E. Senkevich (Russia)
11. Unlimited Accelerated Nonreactive Motion. Review.
12. Revolutionary Current. S.K. Baliyan (India)
13. Tilley Electric Vehicle. Review.
14. Trends in the Application of Motor-Generators. A. Akau (USA)
15. Electric Vehicle in Russia. Review by A. Pashova (Russia)
16. The Richard Clem Motor and the Conical Pump. R. Koontz (USA)
17. Spontaneous Flow Heat Pump. J.-L. Brochet (France)
18. Registration of Gravitational Waves. V.N. Petrov (Russia)
19. High-Energy Hyper-Low-Frequency Electric Field. Yu.V. Ivanko (Ukraine)
20. To the Question of Work Made by Electrostatic Field. A.V. Frolov (Russia)
21. Comments to Electro-Reactive LIFTER News.
22. Letters from Readers.
23. Advertisements (books)

Minato wheel. Experiments in Sweden

Eric Vogels, Sweden

E-mail: fdp@hemsida.net
http://fdp.hemsida.net

The self-running bicycle wheel was shown to the public at the First International New Science Symposium in 1997, held in Korea. The inventor of the wheel, Mr. Kohei Minato, was a speaker during the symposium. On www.keelynet.com you can find the quote of an eye-witness:

"The motor is actuated by moving the N pole of a large permanent magnet (the drive magnet) toward the wheel. As this magnet is moved toward the wheel, the wheel starts to spin. As the magnet is moved closer to the wheel it spins faster.

The acceleration of the wheel is rapid. So rapid in fact, as to be startling. To put it another way I was very impressed. The motor works. And it works very well. In the film clip (see http://fdp.hemsida.net) a slight pumping action of Minato's hand holding the magnet is apparent. When I braced my hand so that there was no pumping action, the motor still ran. In fact it seemed to run better.

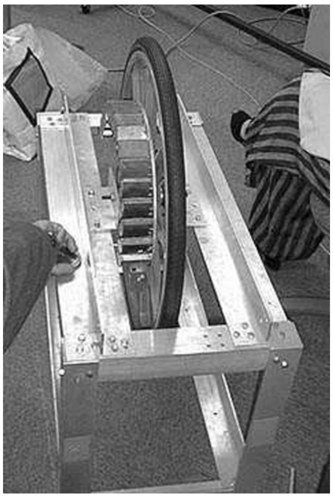

Fig. 1

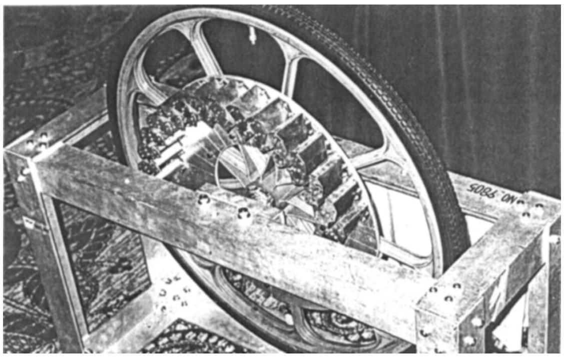

Fig. 2

Pumping action by the hand held magnet **is not** the power that drives the motor. When the drive magnet is moved away from the wheel it coasts rather quickly to a stop and comes to rest in a manner typical of any spinning bicycle wheel. Again when the wheel is at rest and a large magnet is moved up to the wheel it starts to spin. At no time it is necessary to touch the wheel to get it rotating."

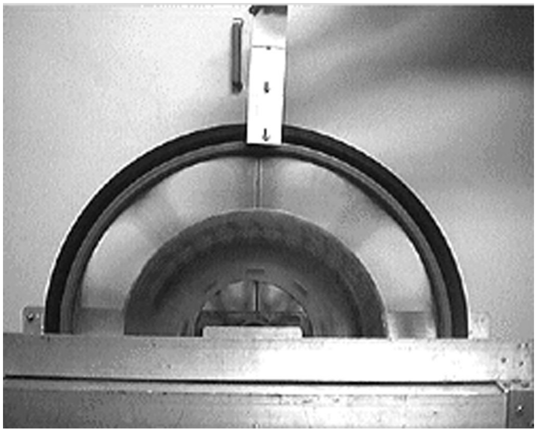

Fig. 3

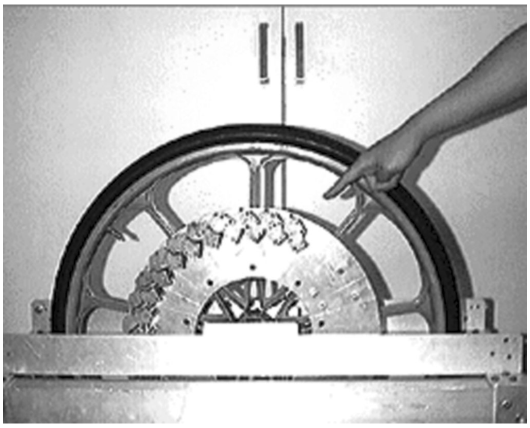

Fig. 4

Since the symposium in 1997 and the first publications in the Internet several people have tried to replicate the wheel, as it seemed to be a **perfect energy source for the future**. Some claims are made that replication was successful but no proof in the form of pictures or video clips are given.

Because I have a great interest in magnets and free energy and the fact that I did a lot of experiments to find a device that could be rotated only by means of permanent magnets, I decided to try to replicate the wheel by myself.

This journey started by searching on the Internet and collecting all the material that was available. A good way to sort all the found material was building my own website at http://fdp.hemsida.net where all my findings are presented.

One of the devices I build was made from a 'how to build a Minato wheel'-manual that I found on the Internet.

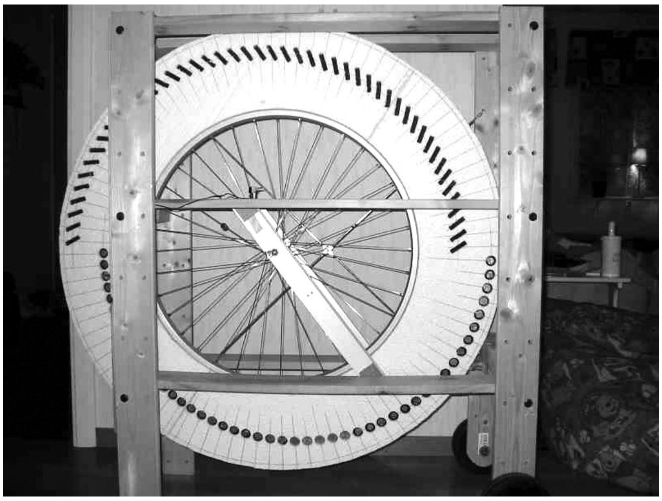

Fig. 5

Unfortunately this wheel did not show any movement at all. I mounted the magnets at several angles and distances from the rim without achieving any positive results. Since the size of the wheel (1 meter diameter) was not very practical to handle and since the results were negative I searched for an easier way for testing. I found that the use of Styrofoam, an old turntable and Neodymium disc magnets kept the whole project within the edges of my kitchen table. I developed a computer program to help me design wheels, since I want to be able to replicate my own wheel, in case I find a model that works. I do not believe in an engine that works after several weeks of tuning and that fails to function as soon as one magnet is replaced. Individual tuning is okay to optimise a working design but if a design does not show the potential to rotate by itself it's a flawed design.

After optimising the way to build the small wheels, I could design and make a wheel in 30 minutes. I made a test scenario and the results for every test were posted on the web site.

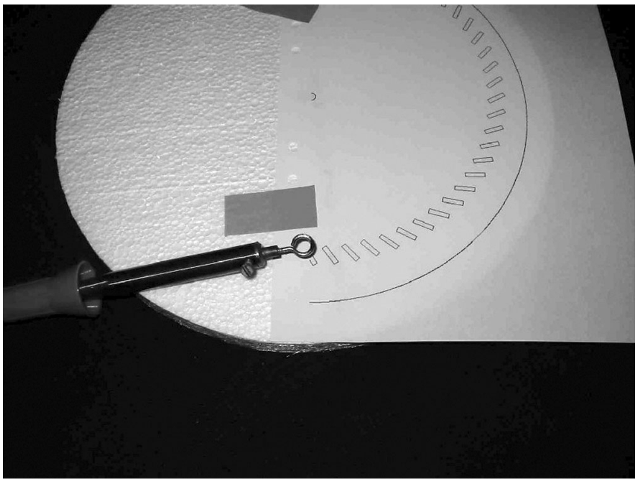

Fig. 6

I searched for the 'perfect' angle, number of magnets and position. The set-up with the strongest thrust was used as the base for the other experiment. By this time I realized that I was not really trying to rebuild the Minato wheel but that I wanted **a rotating device, powered by permanent magnets.**

When for the first time I decided to split a track of magnets that covered 180 degrees of the wheel in a number of smaller tracks, **the results became much better**.

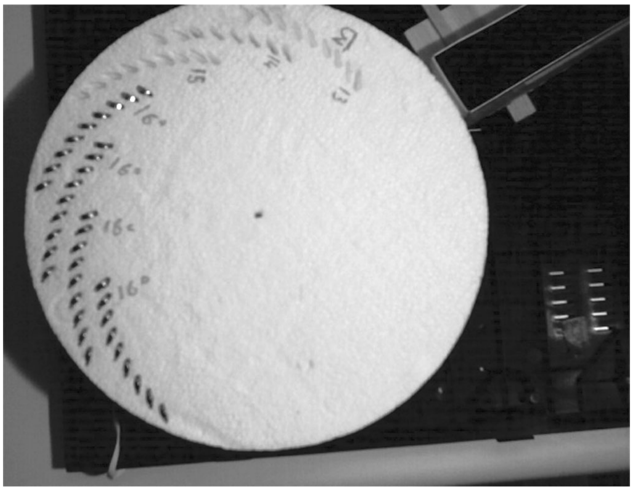

Fig. 7

By using this way of placing the magnets, every track differs a little bit from the next track, depending on its place on the wheel. The angle of the magnets influences on the thrust that is given by the stator magnet, depending on the speed that the track is entering the magnetic field of the stator. In other words: the track at the beginning of the wheel, with magnets covered side of the wheel, gives less thrust than the track at the end. This is important since a track with a lot of thrust has a big 'sticky spot' that should pass the stator. This 'sticky spot' stops the wheel during its rotation just before the first track enters the magnetic field of the stator.

During the last experiments the stator magnet changed place and pushed against the magnetic fields of the tracks from above. Simply because more thrust was developed by this way. The next wheel will have the tracks mounted on a different level. It means that the first track will have a bigger distance to the stator magnet than the last track. This should make the sticky spot weaker and the final 'kick off' stronger. The final wheel will be mounted vertically, like the Minato Wheel.

Thanks to the invention of the Neodymium magnet in 1983, magnets became much more powerful while the size and length decreased. Before 1983 a lot of the experiments could not be done since the size of the magnets made the positioning of several magnets in tracks like this was impossible. It is my personal believes that the strength of these magnets will lead to results that are impossible according to the physical law books, because these magnets did not exist when those books were written.

Until the moment I write this I have not found the perfect energy source yet but by exploitation all the material and findings on my web site I hope that more people get interested and will help searching for a great energy device that will not pollute the air which our children and grand children have to breath.

Mechanism of Drive-Free Motion

Sergey A. Gerasimov, Russia

Physics Department, Rostov State University, Rostov-on-Don, 344090
Email: GSIM1953@mail.ru

The article presents results of the experimental study of the motion created by vibrations of an internal mass of the system of bodies accompanied by impacts of the unbalanced load with the external body. This type of propulsion drive is known by a number of names. Among these are the ***vibrational propulsion device*** *and* ***reactionless machine****. Sometimes it is called the* ***inertioid*** *[1, 2]. The vibrational transposition is proved to go on by means of internal forces of a system of bodies [3-5]. The Tolchin's inertioid [6] is considered to be the first device that used the forces of inertia to create the reactionless infinite motion in space. We are not in a position to be a judge of this. There exist a number of projects of such machines but experimental results concerning such kind of motion are very limited. Below there are the experimental results on average velocities of such a motion and the description of a propulsion device a main particularity of which is absence of wheel-drive.*

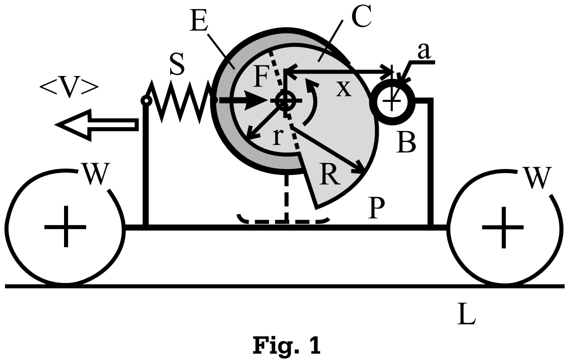

Fig. 1

Scematic representation of a vibratory-impact self-transposition

The unbalanced load in this device is an electric motor E of mass m that executes undumped vibrations relative to a platform P of mass M with four wheels W which can roll on a horizontal surface L with rolling friction the coefficient of which is k. The frictional force is determined to be a force which adjusts to keep the cart from motion across a surface. A disk cam C at one end of the axle of the motor ensures a prescribed character of the vibrations and elastic impacts between the load m and the follower B of radius a. In present experimental device, the cam C consists of two semicircles of different radii r and R as it is shown in Fig.1. In this experiment r=0.03m, R=0.05m and a=0.01m. The electric motor and the platform are coupled by a connecting spring S. The force F by means of which the electric motor is pressed to the follower varies linearly from F=4N at x=0.04m to F=4.8N at x=0.08m. The compressed spring S is necessary not only for creating close contact between the cam and the follower. The restoring force of the spring produces the transposition of the cart in a direction opposite to **F** when the follower moves without contact from x=2R-r+a up to an impact at $x=r+a$. The impact suppresses the transposition. Another mode of transposition is also possible when the impact of the cam on the follower is a reason of the transposition of the cart in the direction of the force **F**. In this case the frictional force extinguishes the transposition of the cart.

Friction between the load of mass m and the cart is negligible since the corresponding effective coefficient is less than 0.001. The same is for the friction between the cam and the follower. The mass of the spring is m_S=0.007 kg , and the mass of the cam is m_C=0.019 kg. The total mass of the wheels is m_W=0.082 kg. This value can be useful for a theoretical

During the last experiments the stator magnet changed place and pushed against the magnetic fields of the tracks from above. Simply because more thrust was developed by this way. The next wheel will have the tracks mounted on a different level. It means that the first track will have a bigger distance to the stator magnet than the last track. This should make the sticky spot weaker and the final ‘kick off’ stronger. The final wheel will be mounted vertically, like the Minato Wheel.

Thanks to the invention of the Neodymium magnet in 1983, magnets became much more powerful while the size and length decreased. Before 1983 a lot of the experiments could not be done since the size of the magnets made the positioning of several magnets in tracks like this was impossible. It is my personal believes that the strength of these magnets will lead to results that are impossible according to the physical law books, because these magnets did not exist when those books were written.

Until the moment I write this I have not found the perfect energy source yet but by exploitation all the material and findings on my web site I hope that more people get interested and will help searching for a great energy device that will not pollute the air which our children and grand children have to breath.

Mechanism of Drive-Free Motion

Sergey A. Gerasimov, Russia

Physics Department, Rostov State University, Rostov-on-Don, 344090
Email: GSIM1953@mail.ru

The article presents results of the experimental study of the motion created by vibrations of an internal mass of the system of bodies accompanied by impacts of the unbalanced load with the external body. This type of propulsion drive is known by a number of names. Among these are the ***vibrational propulsion device*** *and* ***reactionless machine****. Sometimes it is called the* ***inertioid*** *[1, 2]. The vibrational transposition is proved to go on by means of internal forces of a system of bodies [3-5]. The Tolchin's inertioid [6] is considered to be the first device that used the forces of inertia to create the reactionless infinite motion in space. We are not in a position to be a judge of this. There exist a number of projects of such machines but experimental results concerning such kind of motion are very limited. Below there are the experimental results on average velocities of such a motion and the description of a propulsion device a main particularity of which is absence of wheel-drive.*

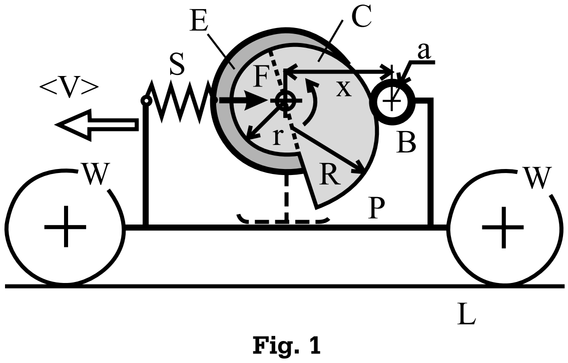

Fig. 1

Scematic representation of a vibratory-impact self-transposition

The unbalanced load in this device is an electric motor E of mass m that executes undumped vibrations relative to a platform P of mass M with four wheels W which can roll on a horizontal surface L with rolling friction the coefficient of which is k. The frictional force is determined to be a force which adjusts to keep the cart from motion across a surface. A disk cam C at one end of the axle of the motor ensures a prescribed character of the vibrations and elastic impacts between the load m and the follower B of radius a. In present experimental device, the cam C consists of two semicircles of different radii r and R as it is shown in Fig.1. In this experiment $r=0.03$m, $R=0.05$m and $a=0.01$m. The electric motor and the platform are coupled by a connecting spring S. The force F by means of which the electric motor is pressed to the follower varies linearly from $F=4$N at $x=0.04$m to $F=4.8$N at $x=0.08$m. The compressed spring S is necessary not only for creating close contact between the cam and the follower. The restoring force of the spring produces the transposition of the cart in a direction opposite to **F** when the follower moves without contact from $x=2R-r+a$ up to an impact at $x=r+a$. The impact suppresses the transposition. Another mode of transposition is also possible when the impact of the cam on the follower is a reason of the transposition of the cart in the direction of the force **F**. In this case the frictional force extinguishes the transposition of the cart.

Friction between the load of mass m and the cart is negligible since the corresponding effective coefficient is less than 0.001. The same is for the friction between the cam and the follower. The mass of the spring is $m_s=0.007$ kg , and the mass of the cam is $m_c=0.019$ kg. The total mass of the wheels is $m_w=0.082$ kg. This value can be useful for a theoretical analysis. Thus, the total mass of the unbalanced body is chosen to be much larger than the mass of the cam or the spring: $m = m_E + m_C + m_S \approx m_E$.

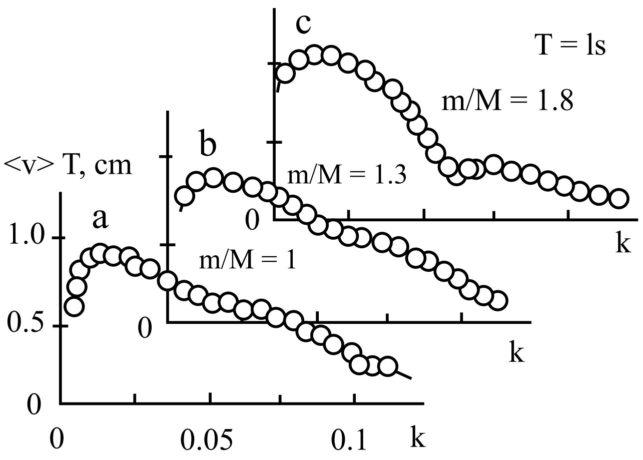

Fig. 2
Frictional coefficient dependence of average velocity of the platform at the period of vibrations T=1s for various mass ratios m/M:
(*a*) M=0.9 kg , (*b*) m=1.2 kg , (*c*) m=1.6 kg

At least in the case of irreversible mode, one could expect that the value of average velocity of transposition $<v>$ must be proportional to the number of impacts per unit of time. In the other words the path traveled by the cart per one impact must not depend on frequency of vibrations. This is a reason why the dependence of the product $<v>T$ versus the value of the friction coefficient k and ratio of masses d = m/M is investigated in this work. Measurements were carried out for two magnitudes of rotation period T= 1s and T=6s of the cam. The obtained results are presented in Fig. 2 and Fig. 3. First of all, the assumption mentioned above is not confirmed.

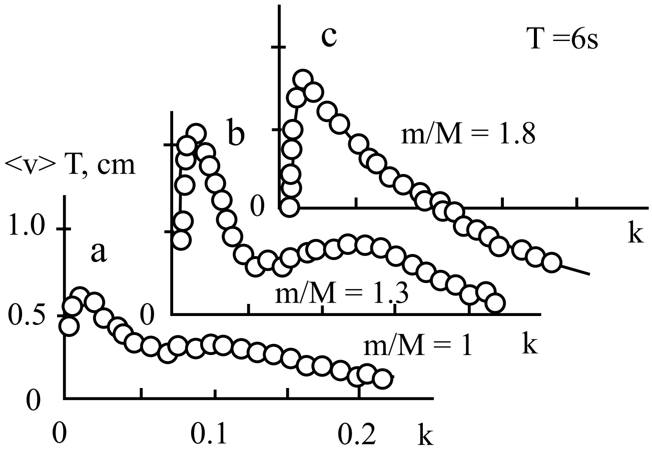

Fig. 3
Average velocity of the platform $<v>$ as a function of frictional coefficient k at T=6s for various mass ratios m/M: (*a*) M=0.9 kg , (*b*) m=1.2 kg , (*c*) m=1.6 kg

The value $<v>T$ for the period of the rotation T=1s sufficiently differs from that for T=6s measured at the same conditions. Besides, when the mass ratio is large the cart can change its direction of transposition. The reason of these results is not clear but this device provides a challenge to explain and investigate this kind of motion that enables us to discover the reactionless motion if it is probable. One should pay attention to a fact that the infinite transposition of the platform takes place even if the friction coefficient k is small.

References

1. Gulia N.V. *Inertia.* // Moscow: Nauka. 1982.
2. Aspden H. Anti-gravity Electronics. // *Electron and Wireless World.* 1989. V. 95. No 1635. P. 29-31.
3. Nappo F. Sulla Influenza Indiretta Delle Forze Inertne sul Moto del Baricentro. // *La Ricerca Scientifica (Rendiconti).* 1965. V. 8. No 1. P. 3-14.
4. Gerasimov S.A. Anisotropy of Frictional Force and Vibrational Transposition. // *Problems of Applied Physics.* 2001. V 7. P. 85-88.
5. Gerasimov S.A. Self-Similarity of Vibrational Motion in a Resistant Medium. // *Journal of Applied Mechanics and Technical Physics.* 2002. V. 43. No 1. P. 90-92.
6. Tolchin V.N. *Inertioid.* // Perm: Perm Book Publisher. 1977.

About the Author

Sergey A. Gerasimov graduated from Faculty of Experimental and Theoretical Physics of Moscow Physical Engineering Institute in 1976. Received Degree of PhD in Physics and Mathematics in 1987. Author of about 70 articles on Astrophysics, Atomic and Radiation Physics, Classical Electrodynamics and Mechanics. In present - Associate Professor of Department of General Physics of Rostov-on-Don State University. The fields of scientific interests are questionable problems in physics: self-interaction, self-transposition, unipolar induction.

ELECTRIFYING TIMES
an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy systems,
events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

Production of Ball Lightning in Laboratory Environment

Report by our correspondent Alla Pashova

Three years ago a working device producing a ball lightning in laboratory environment was designed in Saint-Petersburg Institute for Nuclear Physics (SPINP). The produced ball lightning is accessible for detail investigation and quite stable. The lifetime of the ball lightning comes to about one second that is rather significant for alike artificial formations. The experiments on the device made by scientists of SPINP, i.e. A. E. Egorov, G. D. Shabanov, S. Stepanov, are not supported or financed. Let us note that every scientist of the group searches proofs for his own hypothesis of nature and structure of the ball lightning at all.

A leading expert of SPINP, Anton I. Egorov, pays attention to dethronement of scientific myths:

— There is a myth of ball lightning that is created by mass media. The mythical ball lightning is a concentrate of mysterious energy which is extremely dangerous for a human. It destroys houses, kills animals, pursues people. After meeting it a human can lose his hair or teeth and different misfortunes begin to happen. Supposing a simple story of a farmer-eyewitness: "It thundered, and a fist-sized fire ball rolled down along a drainpipe. It fell into a barrel of water, the water gurgled. I came up to it and put my hand into the water. The water seemed to become warmer…". After republishing of the story by several newspapers a dramatic story about a ball lightning which has evaporated a barrel of water appears. No wonder that such familiar attitude to facts causes hundreds of hypotheses of ball lightning nature.

— **What is your hypothesis about ball lightning structure?**

— At the beginning of 90th I. D. Stakhanov, a member of Institute of Magnetism (IZMIRAN), developed a special method to interview eye-witnesses that resulted a right notion on ball lightning phenomenon. According to Stakhanov, ball lightning is a clot of hydrated plasma which is generated in wet air at electrical discharge.

Water as a chemical compound is remarkable for its anomalous properties: combining of two lightest elements does not generate gas but produces a high-boiling liquid. This is caused by extremely irregular distribution of electrons in a water molecule. Due to this property it acquires properties of an electric dipole. Water molecules interact with charged ions, aerosol particles, and with each other in a special way.

If a positive ion and a negative ion are introduced simultaneously into a puff of warm wet air then water dipoles immediately produce hydrate shells around the ions. When the hydrated positive ion approaches the hydrated negative ion additional water molecules are drown into gaps between them. As a result, there is a stable cluster in which the charged ions are conserved. The cluster consists of two ions of opposite charges and the hydrate shell. Water molecules prevent the ions' approaching and recombination, hence, lifetime of the ions in the cluster increases up to tens of minutes, i.e. in 12-13 orders. Interaction of clusters causes open-chain structures followed by space structures. That is to say, there is appeared a clot of cold hydrated plasma that accumulates great energy (up to 1 kylojoule per liter). The clot of plasma loses this energy at recombination of ions.

— **Could you, please, tell about the design of the device? What processes proceed when the device is operating?**

— Our task is to introduce an abundant population of ions into a puff of warm air saturated with water vapors. A base of the device for laboratory reproduction of ball lightning is a capacitor bank which is able to be charged up to 5.5 kV. The positive pole of the capacitor bank is connected to a ring electrode by means of a copper bar. The ring electrode is placed on a bottom of a polyethylene reservoir filled with water. The negative pole of the capacitor bank is connected to a carbonic electrode which is placed in the centre of the reservoir near water surface. A quartz pipe encloses the electrode in such a way that it is possible to drop water or to put some natural matter on it.

To generate a ball lightning 2-3 drops of water are put on the electrode. When impulse discharge occurs a bright plasma spout escapes the centre of the electrode that is accompanied by a quiet plop. A glowing plasmoid which is an artificial ball lightning parts from the plasma spout. It comes up slowly in the air and then disappears falling to pieces in 0.2-0.3 seconds.

We have made thousands of experiments for investigation of ball lightning properties, i.e. defining size, lifetime, colour, average temperature, excessive charge, content of a dust component.

It was ascertained that the artificial ball lightning is generated in a narrow interval of breakdown tensions. The average size of such a ball lightning is 12-20 cm, and its lifetime comes to 1 second. Temperature of the ball lightning is not very high, i.e. 50° C. This can be defined if to take into account speed of rising of the ball lightning. If the plasmoid is accepted as a puff of warm wet air of 14-centimeter diameter which in atmosphere comes up at 293 K at speed of 1-1.2 m/sec then, consequently, its temperature should not exceed 330 K.

Colour of the lightning varies greatly and depends on presence of aerosol of matter trapped in the moment of discharge. Usually the lilac central part of the plasmoid is surrounded by a diffusive yellowish layer. Some admixture of natrium salt and calcium makes the core of the plasmoid yellow or orange.

If the central carbonic electrode is replaced by an iron or copper or aluminum one then the fundamental character of the phenomenon does not change. However, colour of the plasmoid depends on a radiation spectrum of excited atoms of the electrode, i.e. iron plasmoids are whitish, copper plasmoids are greenish, aluminum ones are white with reddish shimmer.

— **A generated ball lightning exists for about 1 second. How can it be made more stable?**

— Lifetime of an artificial ball lightning depends on many conditions, i.e. size and geometrical form of the central electrode, voltage between the electrodes, value and duration of a current impulse, temperature and electroconductivity of water which is put on the central electrode. Besides, lifetime of the plasmoid can be changed by introducing an additional dispersive phase into it. We have tested tens of matters and begun to investigate suspensions of colloidal graphite and fine-dyspersated ferric oxide.

A suspension of 3 g of colloidal graphite, 8-10 ml of acetone (which played role of penetrating agent), and 90 ml of water is put on the central carbonic electrode. When an electrical discharge occurs a layer of the suspension forms a flying spherical plasmoid. It comes up slowly and disappears in 0.3-0.8 seconds. The core of the plasmoid has a colour of flame, i.e. colour of burning carbon.

To prolong the existence of the generated ball lightning without application of aerosols it will be possible to use the so-called "Faraday's cylinder" whose production has been already begun. For the same purpose G.D. Shabanov proposes to put a stopping potential on a probe of detention.

— **There is an opinion that physical nature of ball lightning is similar to the process of controlled thermonuclear fusion. In this case, if your work on generation of a stable ball lightning is successful then you will be a competitor of the expensive project of controlled thermonuclear fusion**.

— I think that it is totally incorrect. Hydrated plasma is the first enemy of the thermo-nuclear fusion as water molecules do not allow neutrons approach each other. Effective cold fusion should be realized in organic liquids, for example, in heavy acetone or in a water-free medium. Somehow, it should be an absolutely "dry" process. No experiments on real "dry" cold fusion have been realized in practice. It has not been also examined the surfaces on which the combination of heavy hydrogen atoms produces maximal heating.

Scientists should pay attention to two most effective cold fusion processes. The first one expects association of two atoms of deuterium occurring on a totally dry deuterated surface which consists of, for example, zirconium deuteride. In the moment of fusion of a deuterium molecule local heating appears, and neutrons depart. The other perspective method of realization of the cold fusion process requires an absolutely "dry" organic liquid, i.e. liquid acetone in which hydrogen atoms are replaced by deuterium atoms (C_3D_6O) or by atoms of cyclic compound of $C_6(D_2)_6$. A tellurium or zirconium tip of an ultra-sonic dispersant is placed into a reservoir of this liquid. Cavitation blebs are formed on the surface of the dispersant. Neutron output comes to 10^4 particles. Maximal neutron output, which was achieved by Lipson, an American experimenter, consists of 10^8 neutrons at a desired result of 10^{13}. Certain quantity of neutrons can be obtained during ultrasonic cavitation which is accompanied by a phenomenon of sonoluminescence. Due to sound resonance the only cavitation bleb is generated in acetone. When the bleb collapses weak glowing is observed. The cause of this phenomenon consists in gas heating occurring in the bleb that is the result of high pressure produced by its collapse. The burst can last from 1/20 up to 1/1000 sec. Light intensiveness depends on quantity of gas in the bleb. If gas is absent in the bleb then the glowing does not occur. Light emission of the bleb is very weak, it becomes visible if it is strengthen or in absolute darkness.

— **Is it rightful that cold fusion is the future of world power engineering?**

To my mind another direction seems to be more perspective, i.e. extraction of uranium from sea water and then its burning in heavy hydrogen reactors like one which exists in Canada. Photo voltaic accumulators can also become a successful approach of the alternative traditional fuel engineering. By the way, a working model of such a device to utilize free solar energy has recently been created in our Institute (official web site: http://www.pnpi.spb.ru).

Editor: Read the publications on this theme in following issues of our magazine. Below there is a description of other attempts to generate a ball lightning in laboratory environment or at home. Besides, we publish an article dedicated to the problem of laser control of ball lightning. In the article there are photos (also see the cover page) and a scheme of the working device designed by this scientific group.

Ball Lightning Experiments

Information from
http://www-personal.umich.edu/~reginald/ball_l.html

Ball Lightning in the Microwave

This is a great experiment to try at home. It requires a microwave oven, a candle, a toothpick and a lighter. First you put the candle (~1-2 inches in height) in the microwave (remove glass plate from bottom so candle sits on metal). Put the toothpick in the candle sticking straight up. Light the toothpick on fire so that flames are leaping off the tip. Shut the door quickly and turn on oven full blast. There will be loud popping noises and then balls of fire will leap from the toothpick and fly around inside the microwave while making a buzzing sound like a bee. If it doesn't work at first try to move the position of the candle in order to find the "hot spot" in the oven for it to work.

Real Ball Lightning Generated by Pulsed Power Inductor

This experiment is very DANGEROUS. The author did this one a few years ago. The Idea was got from an article about a guy generating ball lightning using a high current transformer (TBA). The author had an idea of using an inductor to store large amounts of electrical energy. The conductor on the end of the stick touches the metal ring. This completes the circuit, and start the inductor. Then there was used the air compressor nozzle to blow out the arc as the wire was pulled away. The copper wire explodes and creates rapidly rotating and burning molten balls of copper. These 'Balls of Fire' exhibit many of the properties of real ball lightning.

Electric Discharge Produced into Air Half-Space

Gennady D. Shabanov, Russia

Tel. +7(271)46477
E-mail: discharge@gtn.ru

Oleg M. Zherebtsov, Russia

Tel. +7(271)46272
E-mail: omzher@mail.ru

In the article a possibility of streak lightning control by means of low intensive laser emission is observed.

During investigation of an electric discharge into air half-space [1] it was noticed that this discharge can be controlled by a low intensive laser beam. Now there is a problem of laser control of lightning discharge [2]. However, numerous works on this approach demonstrate that "the hope to get a quick solution of the problem of lightning control by laser emission has not been confirmed" [3].

Careful observation of this problem in [2] has shown ways out the situation. The authors of [2] consider plasma channel produced (by means of laser) in free atmosphere at a possible greatest height to be of doubtless interest of the science of lightning. Finally, creation of the plasma channel should be of benefit for lightning protection. The authors of [2] give notice that "there are many difficulties of fundamental and practical importance on this path".

Weak theoretical understanding of lightning generation causes pessimistic estimations of solution of the problem. The article [2] notes particularly that "there are neither adequate theory, nor numerical calculations and qualitative understanding of the phenomena defining the speed of a leader… The situation of a theory of the leader channel is little better (from quantitative point of view)…".

Realistically the following statement can refer to the lightning discharge: "The electric discharge appeared to be very "unhandy" for theoretical description but the most interesting phenomenon in the experimental aspect" [4].

Experimental Part

In this work a capacitor bank with 0.6 mF capacity, which can be fed up to ~5 kV, was used to produce impulse discharge into air half-space. A scheme of the device is presented in Fig.1. At connection/disconnection of a discharger 5 a "spout" is let out of an electrode 3. The spout carries the potential of the cathode (virtual cathode) at a significant height into air half-space. A probe placed at the height of ~15 cm fixes a potential which is similar to the potential occurring at the cathode. The researches have demonstrated that the produced formation continues to glow for several hundreds of milliseconds (the glow is fixed from a zone located at 15-45 cm above the cathode). Typical time of the discharge comes to 100±20 msec and depends on the cathode material. Electric field generated in the spout comes to less than $8\ \mathrm{V\ cm^{-1}}$.

Ball Lightning Experiments

Information from
http://www-personal.umich.edu/~reginald/ball_l.html

Ball Lightning in the Microwave

This is a great experiment to try at home. It requires a microwave oven, a candle, a toothpick and a lighter. First you put the candle (~1-2 inches in height) in the microwave (remove glass plate from bottom so candle sits on metal). Put the toothpick in the candle sticking straight up. Light the toothpick on fire so that flames are leaping off the tip. Shut the door quickly and turn on oven full blast. There will be loud popping noises and then balls of fire will leap from the toothpick and fly around inside the microwave while making a buzzing sound like a bee. If it doesn't work at first try to move the position of the candle in order to find the "hot spot" in the oven for it to work.

Real Ball Lightning Generated by Pulsed Power Inductor

This experiment is very DANGEROUS. The author did this one a few years ago. The Idea was got from an article about a guy generating ball lightning using a high current transformer (TBA). The author had an idea of using an inductor to store large amounts of electrical energy. The conductor on the end of the stick touches the metal ring. This completes the circuit, and start the inductor. Then there was used the air compressor nozzle to blow out the arc as the wire was pulled away. The copper wire explodes and creates rapidly rotating and burning molten balls of copper. These 'Balls of Fire' exhibit many of the properties of real ball lightning.

Electric Discharge Produced into Air Half-Space

Gennady D. Shabanov, Russia

Tel. +7(271)46477
E-mail: discharge@gtn.ru

Oleg M. Zherebtsov, Russia

Tel. +7(271)46272
E-mail: omzher@mail.ru

In the article a possibility of streak lightning control by means of low intensive laser emission is observed.

During investigation of an electric discharge into air half-space [1] it was noticed that this discharge can be controlled by a low intensive laser beam. Now there is a problem of laser control of lightning discharge [2]. However, numerous works on this approach demonstrate that "the hope to get a quick solution of the problem of lightning control by laser emission has not been confirmed" [3].

Careful observation of this problem in [2] has shown ways out the situation. The authors of [2] consider plasma channel produced (by means of laser) in free atmosphere at a possible greatest height to be of doubtless interest of the science of lightning. Finally, creation of the plasma channel should be of benefit for lightning protection. The authors of [2] give notice that "there are many difficulties of fundamental and practical importance on this path".

Weak theoretical understanding of lightning generation causes pessimistic estimations of solution of the problem. The article [2] notes particularly that "there are neither adequate theory, nor numerical calculations and qualitative understanding of the phenomena defining the speed of a leader... The situation of a theory of the leader channel is little better (from quantitative point of view)...".

Realistically the following statement can refer to the lightning discharge: "The electric discharge appeared to be very "unhandy" for theoretical description but the most interesting phenomenon in the experimental aspect" [4].

Experimental Part

In this work a capacitor bank with 0.6 mF capacity, which can be fed up to ~5 kV, was used to produce impulse discharge into air half-space. A scheme of the device is presented in Fig.1. At connection/disconnection of a discharger 5 a "spout" is let out of an electrode 3. The spout carries the potential of the cathode (virtual cathode) at a significant height into air half-space. A probe placed at the height of ~15 cm fixes a potential which is similar to the potential occurring at the cathode. The researches have demonstrated that the produced formation continues to glow for several hundreds of milliseconds (the glow is fixed from a zone located at 15-45 cm above the cathode). Typical time of the discharge comes to 100±20 msec and depends on the cathode material. Electric field generated in the spout comes to less than 8 V cm^{-1}.

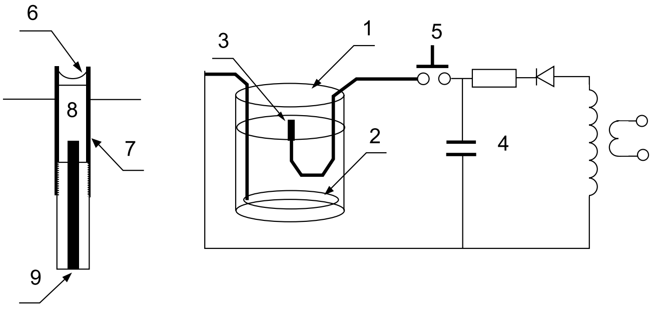

Fig 1
A scheme of a device made for producing of long-living plasmoids

1 – a polyethylene reservoir, 2 – a ring electrode, 3 – a central electrode, 4 – a capacitor bank of 0.6 mF capacity, 5 – a discharger, 6 –water or aqueous suspension drop, 7 – a quartz pipe, 8 – a carbonic or metal electrode, 9 – a copper bar.

The spout was acted by weak laser emission of less than 1 mW at 22-centimeter height. The laser beam was directed perpendicular to the discharge axis, the spout achieving the laser beam moved along the beam towards the laser.

In Fig.2 the spout has moved ~8.5 cm along the laser beam, and a usual sphere formation was generated. Due to the horizontal motion it was distorted if to be compared with usual sphere formation. The usual sphere formation is represented in Fig.3 (also see the cover).

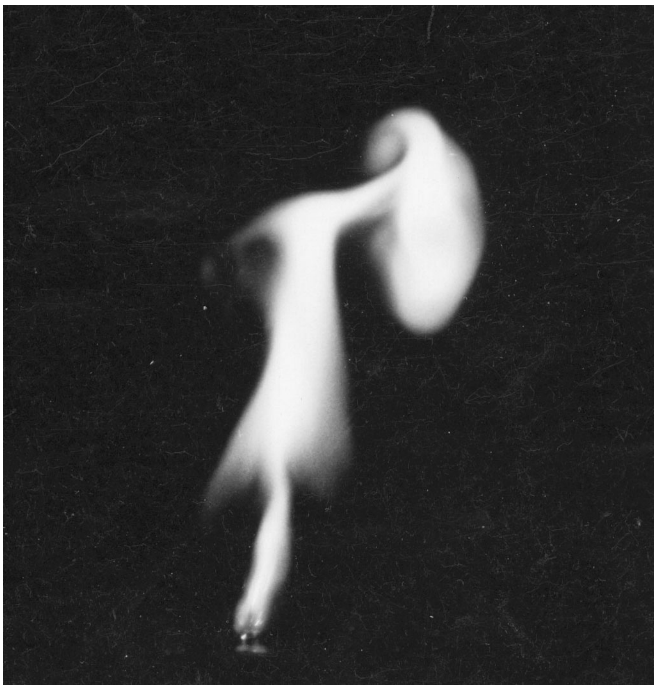

Fig. 2
Motion of the leader channel (spout) along the laser accompanied with the sphere formation

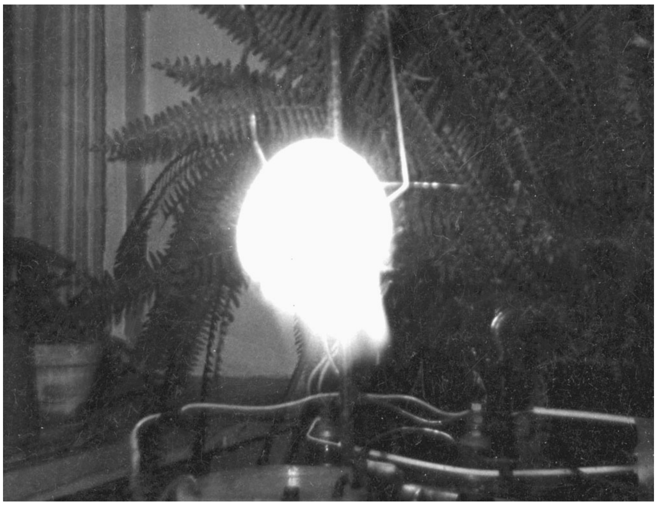

Fig.3
The standard ball formation

Discussion

As it is evenly mentioned in [5], results of model experiments are difficult to be applied directly to laser control of lightning because in short intervals characteristics of the discharge differ from characteristics of lightning discharges. Actually, at comparing considered current and the field existing in a lightning (i~100 A, ~3 V cm^{-1}) with these phenomena occurring in a laboratory spark (~1 A, ~300 V cm^{-1}) incorrectness of modeling of this process becomes obvious. From the other hand the authors of [2] note that "the leader channel is like a channel of electric arc… At current strength of ~100 A plasma of the channel of the arc is supported by fields the densities of which come to several volts by a centimeter. The lightning has such leader currents".

Even an air arc of atmosphere pressure has a field of about 100 V cm^{-1} at currents of about 1 A. The generated discharges (spout) accompanied by the field of less than 8 V cm^{-1} (maximal current appearing in the discharge gap is 50-60 A) are more appropriate for modeling the leader channel of the streak lightning than typical discharges used for investigation of these processes, for example, like [6].

At first approximation the leader channel is considered as an ideal conductor in the article [2]. We make research to define the field in the discharge (spout) more exactly. The article [2] makes the following consideration about a cause of occasional generation of new leader heads: "the surface of *equipotential* plasma conductor (channel) *has* a property of *instability*. There is an occasionally generated sharp jut. An intensified field appears at the jut along the edge. Under the influence of the field the jut becomes to grow in any direction including at the significant angle to the weak outer field". The aforesaid seems to explain why our leader channel changes its direction and runs at right angle to its initial motion. (Fig.2). Weak harmonic action of the laser beam to the leader channel causes generation of a new head which continues motion "on the significant angle".

The effect of plasma motion towards a light beam is a demonstration of a general tendency of propagation of the discharges to an incident electromagnetic field [7]. "Area occupied by plasma usually increases towards the laser emission" [8]. Leading character of motion of this formation is confirmed in [9] as well as the interaction with the laser beam is confirmed in [10]. It should be taken into account that in the works [9-10] experiments were performed at devices which generated such a discharge. However, that discharge was by 2 orders weaker than the discharge generated by the device represented in this work. Naturally, the results were less defined. Comparison of the discharges is presented in [11].

According to our data, the leader channel (spout) has a very abrupt bound (less than 1 mm). In this layer the field can come to ~30 kV cm^{-1} (at height of 15 cm).

Conclusions

Due to the assumption [2] about *instability* of the surface of the *equipotential* leader channel there has been successfully performed the "control" action on an electric discharge which models a streak lightning. It has been achieved by means of weak harmonic oscillations made by the laser. This mechanism is supposed to be applied for streak lightning control.

References

1. G. D. Shabanov. Optical properties of glowing formations // Letters in the Magazine of Technical Physics. 2002, V.28, I.4, P. 81-86;
2. E. M. Bazelyan, Yu. P. Rizer. Mechanism of attraction of lightning and problem of lightning laser control (Nowadays Physics) // the UPS. 2000. V. 170, #7, P.753-769;
3. A.D. Starikov, Yu. A. Rezunkov. Lightning discharge control by laser emission. // the Optic Magazine. 1999, V. 66, #3, P. 15-16;
4. B. E. Meyerovich on the path to realization of electromagnetic collapse // the UPS. 1986. V. 149, #2, P.221-257;
5. Miki Megumu, Uada Acushi, Shindo Takatoshi. Characteristics of laser controlled discharges occurring in long air gaps // the Optic Magazine, 1999, V. 66, #3, P. 25-29;
6. P. Rembo, J. Biggert, V. Kubechek, J. Schwartz, A. Berstein, J. Dilet, R. Bernstein, K. Shtalkopff. Laboratory investigations of lightning discharge initiated by laser // the Optic Magazine, 1999, V. 66, #3, P. 30-35;
7. Yu. P. Rizer Laser spark and distribution of discharges. Moscow: Science (Nauka), 1974. P.308
8. A. A Mak, N. A. Solovyev. Introduction into physics of high temperature laser plasma. Leningrad: Publishing House of Leningrad University. 1991. P. 152;
9. S. E. Yemelin, A. L. Priozersky, V.S. Semenov, G. E. Skvortsov Characteristics of dynamic state propagation in stream of capillary discharge. // Letters in the Magazine of Technical Physics. 1997, V.23, I.19, P. 54-59;
10. R. F. Avramenko, A. Yu. Gridin, A. I. Klimov, V. I. Nikolaeva. Experimental investigation of power-consuming compact plasma formations. // Collected articles Ball lightning in laboratory. Moscow: Chemistry. 1994;
11. G. D. Shabanov, O. M. Zherebtsov, Experimental modeling of ball lightning analogue. // Theses of 10th Russian conference on ball lightning. Moscow (Dagomys). 2002. P. 17-20.

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Color - $800

Half Page Ad
B/W - $200
Color - $400

Quarter Page Ad
B/W -$100
Color- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Free Energy

Alexander V. Frolov, Russia
Faraday Laboratories Ltd
http://www.faraday.ru, tel/fax 7-812-3803844

In general the conception was formulated in 1995 and today we can say that the concept of physical vacuum, which is a new source of energy, finds more and more supporters. The fundamental works about nature of "zero point energy" are published, for example, Andrew D. Sakharov [1], Hal T. Puthoff [2] and many other interesting works. It is clear now that space or "physical vacuum" has its internal structure, therefore it can be used as a new source of energy if we organize the process of changing of this structure. Moreover, according to some theories, the existence of reality (i.e. the World of some certain parameters of space and time) is defined by the function of probability of energy density. For example, Dr. Nassikas, Greece [3], proves that it is not possible to consider the space without its energy, and there is no space without energy. The internal structure of reality is coexistence of two forms: gravitational energy and electromagnetic energy. Any local increase of the first one should produce decrease of the second one, so the sum amount of the change is zero.

In some other articles about energy transformation processes (gravitational form of energy into electromagnetic heat radiation of mass, for example) we can find that both increase of entropy and the inverse processes (decrease of entropy) are possible and in this case the electromagnetic energy can be converted into the gravitational form, that is shown in details in the works of Nobel laureate I. Prigozhin, Order and Haos. Man's new dialog with Nature, London, 1984. Since **the direction of time** (the time course) **and direction of the entropy change** (increase or decrease) are related notions then free energy technologies are considered by Kozyrev as methods of practical application of natural time course [4] that is presented in aether-dynamics by Frolov as aether flow of some density and this aether density determines the time rate as hardness of cause-effect connections for any process, and also for the process of existence of matter in space-time of this aether density.

According to this theory, in any point of space it is possible to get power by means of energy transformations without any consumption of mass-fuel. We can say also that in this case some change of the energy density of space should be detected.

Let's clarify some determinations:

1. The Potential (lat. Potentia that means "force"). In physics this is scalar parameter, it's gradient expresses the intensity of field of a certain force. In common sense, the potential is possibilities, which exist for execution of some task, for completion of some work.
2. The Work is a quantitative parameter of energy transformations. The transformation means here change of form.
3. The Energy (Greek "Energie" means "action, activity") is a quantitative characteristic for different forms of motion.
4. The Power is amount of work per unit of time.

According to the given determinations by The Soviet Encyclopedic Dictionary, edition of 1988, Moscow, **the fact of presence of potential (scalar) field, for example, electrical or gravitational filed, is real possibility to produce some work if we can organize change of energy forms. Let's note that power source is not required to keep the potential field in force. It is free.**

...the potential (scalar) field can produce real work!

The example of this work, which is produced by the field: body falls in gravitational field and when it strikes on the ground then some part of its potential energy is transformed to heat, that is the work as transformation of energy forms. So, there is a conclusion: the potential (scalar) field can produce real work! But we have considered only a half of cycle and in classical case in the second half of the cycle it will be necessary to produce the same work against the field to raise the body to the initial point.

Let's formulate the task to produce the work periodically and to get the power in load from this process. Usual mistake is to accept the particular case (the same body returns back in the same field) as a single possible case. But in special case changes of the system are possible, for example, the field intensity is not a constant but some variable value (alternating or pulsing), or the body changes its own parameters. In this case in each of half-cycle of the process the field can produce real positive work to accelerate the body.

The main technological solutions are obvious: **it is necessary to create gradient of field in space** (full or partial screening of trajectory of the body, which is moving in the field) **or gradient of field in time** (pulsing mode of field). This is quite easy for electric and magnetic fields, but for system, which uses gravitational field to produce the work, we can assume changes of parameters of the body only.

It is possible to understand why professional physicists dislike the question about possibility to use scalar (potential) fields to produce useful work in a load since they think about The Law of Energy Conservation. To find mutual understanding it is necessary "to upgrade" this Law for 4-dimentsional case of real physical system. It is necessary to introduce **the notion of the structure of space-time** where the considered process is created. In other words, for real practical engineering tasks it is necessary to consider **space-time of the real process** but not an abstract space-time. Usually the energy density of technical processes is small in comparison with natural energy density of space (aether density), which is a result of real astrophysical processes, i.e. motion of planet, star, galaxy. If we discover structure of this real rhythm of the space-time of our planet then we'll be able to design it in our technical devices to use aether-dynamics as theoretical basis.

So, in orthodox physics there is the unchallengeable formulation: the work of potential field on closed trajectory of motion is equal to zero. Yes, it is right for only case: if one part of work is positive (the acceleration) and another part is negative (the deceleration). But one part of the trajectory or a part of periodic process with the negative work can be excluded by different methods: by means of spatial superposition, pulsing mode and change of interaction polarity or by the screening of electromagnetic interaction.

For example, minor changes in vacuum tube design (the grid is located under cathode) allows increasing the kinetic energy of electrons and to increase emission current by means of the potential on the grid only. Also vector potential of magnetic field or gravitational potential can be used by similar way.

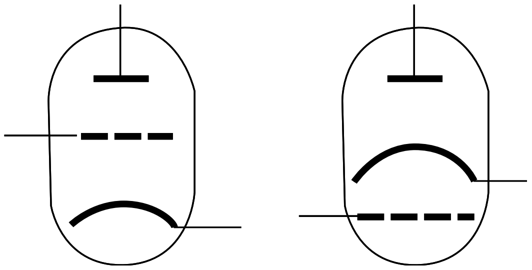

Fig. 1

Besides this way it is possible to use pulsed mode and switch-off the primary energy source before the emission electrons will reach the anode, Fig.2. In this case there is not the conductivity current between anode and cathode and the primary source is not discharging during its work.

Therefore, it is not a news that potential field can produce real work. In any textbook there are examples of positive and negative half-cycles that result to zero net work. But for the case of consequent execution of two processes the parameters of one of the processes can be changed, that **the work, which is produced on the closed trajectory can be formed as two or more parts of positive work.** Important aspect is following: it is necessary to determine the point (moment) of change the sign of the interaction and at that point (moment) to change the parameters of the process accordingly.

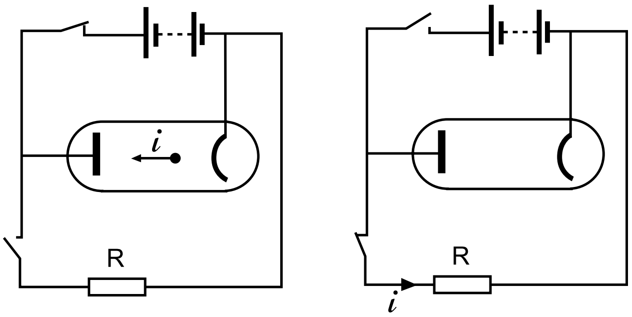

Fig. 2

The good sense requires the answer: what is the source of the power to produce this work if this proposed concept is right? Let's notice that before the considered examples, the existence **of power interactions of potential fields** must make the same question. For example, how the ordinary permanent magnet (i.e. its vector potential) holds a piece of metal making work against force of gravity without any fuel? How the potential field can move and accelerate ions, i.e. kinetic energy of ions can be increased without any fuel? What provides the forces of elasticity? Let's try to find answers from consideration of inner structure of electric and magnetic field, especially let's develop **our understanding of the notion of gradient, which describe properties of space-time design in area of this field.**

The concept of potential as bi-directional flow of photons and anti-photons was proposed by English mathematician E. Whittaker, and then it was developed by Dr. Thomas E. Bearden [5]. In this concept, the generalization of the Third Newton's law looks as the requirement of complementary pair to the process of the electromagnetic radiation. Since "process" means a change of information in time, this paired anti-process is reversed in time. Of course, it is development of process to its own future, but from our point of view, it goes from future to past. According to generalized Third Newton's law the radiation of photon is paired with anti-photon. In this case, the internal space-time structure of electric potential field is formed by two contrary flows of energy: photons spread from charge source and the anti-photons "inflow" into the charged mass. Let's notice that this concept defines the relationship of charge and mass. The charge without mass does not have any sense.

The flow of energy outgoing from charged mass is responsible for all phenomena of radiation. Incoming flow of energy is responsible for the gravitational interaction. Therefore, the notions "radiation" and "gravitation" can be considered as complementary pair; that is considered in details in concept of Josef Hasslberger [6].

...it is impossible to get "something" from "nothing"

There is a quite correct question: since "something" can not be created from "nothing" then what and where will be changing if we create the process of free power dissipations in local area of our space-time by means of asymmetrical potential interaction?

Let's try to present a "process" as some change of "information" in time and from this point we'll get the conclusion about results of this energy disbalance of the natural internal structure of potential field: in this case the entropy process in direct time (heat dissipation in load) should be balanced with equal anti-entropy process in reversed time. In other words, the disbalance is change of the space-time curvature and changes in the time rate.

So, it is impossible to get "something" from "nothing" but if we have understanding of the balance (process in time and anti-process in reversed time) then in the Universe the 4-dimentsional balance is not violated. It is assumed here that it is impossible to create one process but it is possible to create two opposite balanced processes, and each of them produce real work in the load. In astrophysical consideration it was stated in 1964 by Academician Gustav Naan, Estonia, Tartu. He wrote [7]:

"In our real world we could extract any amount of energy from vacuum if some technology provides at the same time the extracting of the same amount of energy for anti-world. Total sum value of energy is equal to zero."

Let's note also that similar conception of "dynamical zero" was applied in ancient India mathematics, i.e. "zero" is not "nothing" but it is hidden possibilities (balance of forces).

One more interesting question appears: what is about reality of the particle of matter after we have claimed that its mass-parameters are parameters of certain process? With this idea the material world as a whole is certain process and "stability" is not a property of some object but a parameter of its existence, which is a process in space-time of some certain energy density and some certain structure. By analogy the vortex in liquid is a process but it is not a liquid.

Thereafter, as electron was presented by Shredinger as wave packet and by Whittaker as function of two scalar potentials, the old idea of instability or **dynamic structure of matter** has the possibility of experimental testing and development as a teleportation technology. This idea is technology of space-time engineering, i.e. creation of the space-time with some certain parameters where curvature (or rate of time) determined rate of existence of the matter in space. It is related with some energy density, which can be increased or decreased. Since by this way it is possible to consider any material particle of substance as certain process, which is balanced by corresponding time-reversed process, so there is no theoretical difficulties to develop technologies of chemical transmutation, materialization, dematerialization, teleportation etc.

So, the using of potential energy to create a process of dissipation of power in load does not violates the Law of Conservation, and this Law can be generalized: the total energy of four dimensional system is amount of energy of processes in time and energy of processes in reversed time, it is constant and it is equal to zero.

Now let's try to review some free energy projects (the systems to produce work without consumption of fuel). Russian Peter the Great had intention to visit Germany in 1725 to test the Orferius' device. There is a very old description of "perpetually rotating wheel", which was made in India by inventor Bhaskar in 1150! From that times the mind of inventors is developed to other modern systems but due to the efforts of scientific groups, which are interested to keep in force the ideas of primitive materialism, the great idea of free energy transformation is distorted to such extent that any person, who began to speak about free energy, had a chance to get the name of "mad". Why? The reason is common understanding of the "power", which is "some work per unit of time" and generally it can be presented only as result of some transformation of matter structure, i.e. disintegration, chemical reaction, nuclear decay, nuclear synthesis or any change of structure of a matter. In any case, a material (the firewood, oil products or nuclear fuel) is considered as the fuel, regardless of its transformation method.

Some concepts did not consider the fields (electromagnetic, gravitational and others) as a kind of matter. So hypotheses, and even successful experiments on transformation of "non-material" type of energy in energy of material object (into the work) were not taken into consideration. The physics is a study about measurable and tangible quantities. New measurement methods let us work with a new physical phenomenon. So, we can see that **real situation in alternative energy is changing due to experimenters efforts** but not from the great theoretical team.

Some time ago the electric energy was not considered as a material object, but gradually people have been able to refuse the gas pipes, which were real material source of power for the gas light lamp, in favor of electric wires for electrical illumination. In a short time, I think, it will be possible to refuse the wires and we'll consider the aether (physical vacuum) as unlimited energy source if we organize **the process of transformation of space-time parameters.** It will be necessary to drop the old conception of "primary source of power", which should be connected with the consumer by some method and to develop conception of free energy source in any place "on-demand".

Let's consider, what the term "free energy" means today. The energy in general sense means the "ability of body to produce some work". The energy of closed system is constant. Certain device can look like "perpetual mobile", but nobody will be shocked because the way of energy "inflow" is known. For example, solar panel battery is obtaining its power from external source of light. But in general case, 3-dimensional observer can see nothing similar to the inflow of energy into the system if multi-dimensional energy balance is not analyzed. In other words, "perpetual mobile" is a right name since for its description it is necessary to use notions of "Time", "Eternity", "Causality" and other categories, which are more related with philosophy and religion but not with modern physics.

The "perpetual mobile idea" is really value to be the purpose of serious scientific work. In the book "Great experiments in physics" published by "World", 1973, Professor G. Lipson wrote: "Joule was the person of a very practical kind of mind and he was enthusiast of the perpetual mobile idea". I hope that it is not necessary to explain here who was Joule. However, "practicality" of free energy generators is obvious only for energy customers, but not for energy producers, who created the centralized system of energy distribution. This is the main reason of absence of alternative fuel-less energy systems in the modern market.

Let's consider the existing classification of "perpetual mobiles":

1. "Perpetual mobile" of the first kind is a design, which can "create energy". It is disputable aspect and all patent offices refuse to consider the patent claim of such type. They answer that "energy can not be created or destroyed" but energy can be transformed from one type into another type.

2. "Perpetual mobile" of the second kind is the collector of environmental heat. It is not a "thermo-pair" which uses temperature difference to produce electricity, but it is the "heat pump". This type of devices works with negative entropy, or more exactly, with sintropy. The terminology still is not defined but we can say that in entropy systems the produced work is equivalent to the dissipated heat and the work, which is produced by sintropy systems is equivalent to some absorbed environment heat.

3. The "perpetual mobile" of the third kind is a demonstration of perpetual motion without friction. The analogue is electric current in superconductors. Development of this idea is creation of systems with negative friction. In electromagnetic systems it corresponds to the known cases of the "negative conductivity", i.e. some electrical circuits in this mode can generate output power.

Additionally, let's show that "creation of energy" is possible in theory, for example: two processes of equal power compensate each other $A + B = 0$. Let's assume that some technology creates another (inverse) process as: $0 = A + B$ i.e. two processes of some power in sum create zero result. In general principle, also there is possibility to use many processes balanced situation as $A + B + C + \ldots = 0$ and this interesting conception is known as theory of multipolarity.

Let's consider some quite real (to my mind) examples from the history of development of free energy technologies. Nikola Tesla's investigations are not well known to modern scientists and engineers. In his works on development of wire-less telecommunication Tesla used flat spiral coils as a secondary winding of transformer. The magnetic field of such coil is radial and it is placed in the plane of the coil. In 1995 the author of this article experimented with similar flat spiral coils. I have to claim that when such coil is operating as secondary winding and the solenoid is the primary winding of the transformer then we can measure asymmetrical mutual induction, i.e. connection of active load (lamp) to output circuit of the transformer is free from input power in the primary winding. It is the simplest example of technical realization of the asymmetrical cause-effect connection, described by Kozyrev in his theory of active properties of time.

... "creation of energy" is possible in theory...

Another Tesla's invention is his resonance transformer. Modern electrical engineering describes transformer with forced electrical oscillations and radio engineering considers operation of resonance systems mainly. Tesla put the question on transformation of power in resonance transformer and by this way the efficiency can be more than 100%. With high frequency currents and high power level Tesla used single-wire terminals as loads, i.e. the lamps and other single-wire receivers of power (motors) were powered from high frequency changing electric field. Such single-wire terminal does not consume any power from primary source because it uses the change of potential in point of connection to conductor (let's note also that for maximum efficiency this point of connection should be one of the maximums of standing wave).

The notion about "free vibrations" belongs to Tesla and this term describes sinusoidal oscillations in electric circuit that is created after short non-sinusoidal impulse due to real vibrations of the free electrons. Resonance mode of the free vibrations can be the real way to excess power output.

Original Tesla’s approach to electrodynamics allowed him to build in 1934 fuel-less car with electromotor, which was powered from some 12 vacuum tubes generator of unknown design.

The standing waves of electric field that were observed by Tesla during thunderstorm brought him to conclusion about possibility of the system to power remote energy consumers from energy generator without any transmission lines and without radiation methods. He assumed that it is necessary to create special standing wave of electric potential (or variable in time electric potential field) around the generator, then the unlimited number of loads (lamps, motors) in area of this potential field can be powered if they are tuned in resonance with oscillations of the generator. Let’s notice that in each receiver the power can’t be more than power of this generator but it is possible to install many independent “receivers” without mutual interference.

The modern investigations on these problems sound as sensation, because Tesla’s works are unknown for the modern generation of scientists. Of course, modern electronics components and the tools facilities allow to create real “miracles” in comparison with the past age experiments. For example, engineer Avramenko described his work on single-wire power transmission in Journal of Russian Physical Ideas, 1991, No.2, and in journal “Inventor and rationalizator” 1992, No. 5, 6. The light bulb (or ventilator) was used as load of the single-wire power transmission line. This line can be made of high resistance material, for example, tungsten, but the power can be transmitted without heating of wire! It is possible to say that in this experiment a wire does not transmit power from the generator to the load, but the wire is the conductor of information signal, which is created by polarizational current opened and described by M. Faraday.

It is not difficult to repeat the experiments with single-wire line: it is necessary to place two diodes on the end of a line, which is connected to secondary windings of high voltage transformer (I used television set high voltage unit) by such a way that different poles of the diodes are connected to the line, Fig.3.

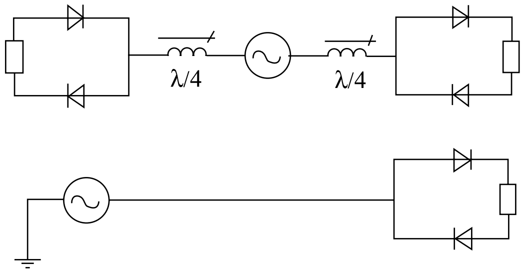

Fig.3

Two other poles of diodes create the source of potential difference (voltage), from what it is possible to charge the capacitor or to power the load (luminescent bulb). This diodes scheme is named as “Avramenko’s diodes plug”. In my experiments it was determined that high efficiency mode is resonance mode in real line and in this case the maximum change of potential is created in the point of connection of the “Avramenko’s diodes plug”. There is well known formulas to calculate this resonance as quarter-wave antenna vibrator system. Of course, high frequency and high potential allow creating high power in the load. In 2001-2003 New Energy Technologies magazine have published new experimental results on single-wire power transmission reported by research team from Moscow (Prof. Strebkov, Avramenko, Nekrasov and others). For example, they built and tested lines from 20 to 100 Kwtt power level!

... the electric potential field can create non-compensated force in the system and to produce a work without any power input.

One more well known researcher on the subject of free energy was Thomas Townsend Brown. He considered creation of reactionless propulsion force by means of electric forces only. Ionization is not considered here! According to his works, the electric potential field can create non-compensated force in the system and to produce a work without any power input. English patent by T. T. Brown #300,311 of August 15, 1927 describes the method to create propulsion force and power from electric energy source only. In this first patent it was claimed that in ordinary flat electric capacitor (two flat plates and dielectric between them), which is charged up to 50 kilovolts and more, Brown discovered propulsion force. This force moves the capacitor to positively charged plate direction. It is only one of his ideas and in his other patents of 1930 - 1965 Brown has described many new methods to create propulsion force and free power in load by means of electric field only (scalar potential field as a source!).

... the efficiency of “electric system can be million to one”

We have to exclude ideas on electrokinetic apparatus since it is just a reactive method and propulsion force is result of ionization flow. Especial case is Brown’s **idea to create the asymmetry of electrostatic forces** by means of some special form of surface, USA patent #3187206 of June 1, 1965, application of May 9, 1956 (Fig 4).

The schemes and descriptions by the Brown’s patents are undoubtedly the work of great practical value. As was mentioned by Brown, the efficiency of “electric system can be million to one” because the potential field can produce real work, for example, rotate some electro generator, but it does not change the primary source of field.

In 1927 T. Brown demonstrated the devices in Ohio, later he worked in France. His works in France were stopped and he returned in the USA. Detailed information on his works can be received from descriptions of patents http://www.soteria.com/brown and from the book “Electrograviticis Systems” by Thomas Valone, Integrity Research Institute, USA.

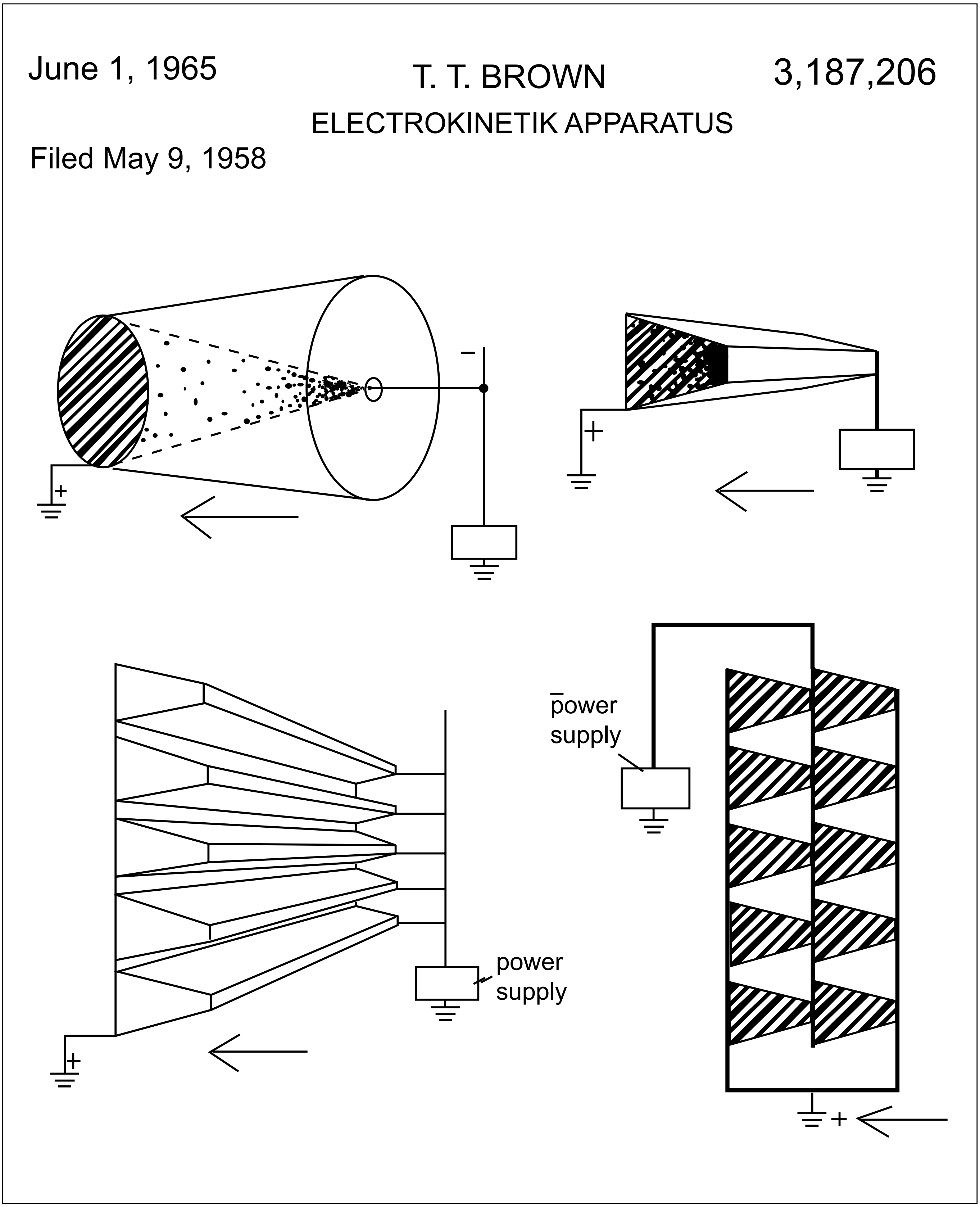
June 1, 1965
T. T. BROWN
3,187,206
ELECTROKINETIK APPARATUS
Filed May 9, 1958
−
+
+
−power supply
power supply
+
+

Fig. 4

Electrokinetic apparatus by T.T. Brown

Besides, there is the simplest example of creation of real work by means of potential field that is also resulting from the Brown's works. Usually two plates of electric capacitor are equal to each other. However, if one of the plates is small and another one is large then the electric field between them is not isotropy field and in this case there is some **gradient of intensity of the field**. In such field any dielectrical object, for example a ball, will be polarized unevenly and due to this fact some force should make it to be moving aside, where the field has the greater intensity. I have to note for mathematicians that since the intensity is "gradient of potential" then gradient of intensity is the **gradient of the gradient**, i.e. it is the second derivative of potential. This idea is an example of the well-known rule: the change gives the new quality.

The example of creation of the propulsion force by means of potential field is also connected with "perpetuum mobile" task, since conductivity currents in the capacitor are very small and it almost does not require power consumption (after being charged once time) and the system can produce mechanical work permanently, in this case it is the work against the gravity force. In general case, if the design allows asymmetrical energy transformations, then the surplus output power and reactionless propulsion force can be created in this system.

Most likely future aerospace systems, which are based on the electrogravity, are the most perspective direction of free energy technologies. Why not for the power engineering? It is clear that now there are some existing fuel heating systems, heat stations and power industry to provide by some traditional way all current needs of the society and due to this fact the innovation of any free energy technology is related with hard competition. However, for commercial programs for cosmos the reactive rockets principles are not acceptable more. Each satellite during its operation should produce profit to compensate large expenses for the rocket-carrier. Only fuel-less propulsion systems can allow developing space commercial programs that are new infinite market. Let's note that the gravitational (reactionless propulsion) technologies are not related with cosmodrome and their cost can be quite acceptable to realize the project with private funds.

... future aerospace systems, which are based on the electrogravity, are the most perspective direction of free energy technologies.

There are some known examples of free energy systems.

In 1921 The Seattle Times, as well as Denver Post of August 8, 1921 published the articles about inventions of Alfred Earl Hubbard. His device included the central core with coil and eight remote coils placed around the central core. After primary impulse, the impulses in all coils were powered and rotating magnetic field was created in central coil. The power produced in the central coil was quite sufficient for self-excitation of the system and for producing of useful work in the load (motor). The boat and the car with electromotor, which was powered from the Hubbard's generator, were demonstrated.

In 1928, Lester Hendershot invented the electric generator of 300-Watts power. This device was designed of details used in radio-receiver to get oscillator (500 kilohertz) and non-inductive coil. Later, in 1970, William Cooper experimented with non-inductive bifilar coils. He used induction phenomenon in the case of zero magnetic component (two-wire winding or flat spire coil).

The Cooper's USA patent 3610971 of 1971 describes **the principle and the device to create power in secondary circuit without reaction on primary circuit**, as well as method to get reactionless propulsion force for aerospace application. Cooper also has found that specially designed coils can produce the field, which can not be screened and this field has some common parameters with the gravitational field.

The gravitation is considered by Cooper as a polarization of atoms in gravity field of planet. So he declares in description of his patent: "the electronic generator... of super high frequency creating the pulsing electric field of single polarity... acting in opposite direction to the Earth gravitational field... that to depolarized the atoms and to release them from the gravitation".

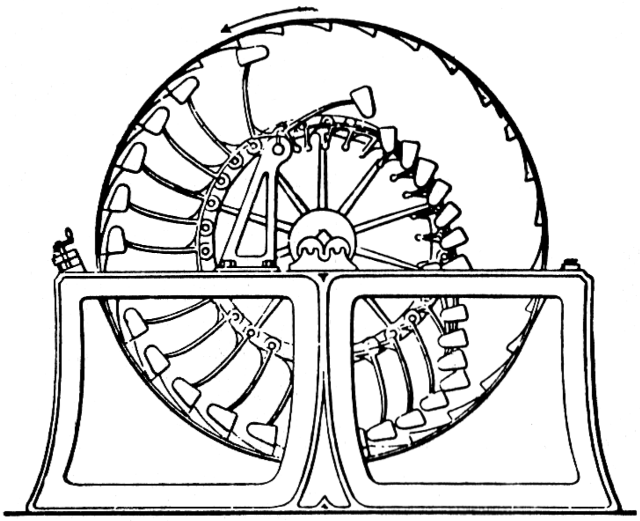

Fig.5

By the way, the gravitational field itself can be used to get a power. "Unbalanced wheel" is a well known design. The weights on the one side of the wheel, which is rotating in vertical plane, can be organized to be moving to axis, but on the another side the same weights can be organized to be shifted from the axis to the periphery of the wheel. By this way there is some constant shifted center of gravity in the system and it can be permanently rotating system. One of the inventors who built such wheel is Hugo R. Fraga, Havana, Cuba (The Perpetual Motion Mystery. R.A. Ford, Lindsay Publications Inc., Bradley, IL 60915, USA, 1987, see Fig. 5).

The earliest information about such systems is dated of 1150 and the name of inventor is Bhaskar, India. He used the tangential disposed pipes, which were half-filled with water. In France (1235-1240) William de Onnecure demonstrated self-rotation wheel with seven weights. In Italy (1438) Mariano de Jacopo has built the system made of eight bars disposed in the plane of rotation, and the bar can be fold up in the middle like elbow joint to provide rotation. One of the well-known and documented events of real demonstration of perpetual rotation wheel was mentioned in 1620. Edward Somerset (Second Marquis of Worcester), author of the book "Century of Inventions", 1963, has built and tested the wheel of about 4 meters in diameter, 14 weights of 25 kilograms each. The test of this machine was organized in London, in witness of King Karl, Grand Duke Hamilton and Grand Duke Richmond, and there are files in royal archives about this test. The descriptions of these and other systems are published in the book "Perpetual mobiles: past and present time" by Brodiansky, Moscow, Energoatomisdat, 1989.

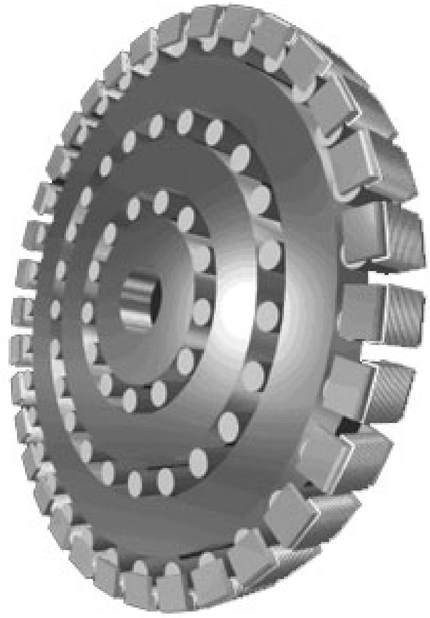

Fig.6

In different idea, which is known from Leonardo Da Vinchi drawings, lifting of water is produced by the helical "Archimedes' screw" of small diameter with large centrifugal acceleration, which reduces weight, but lowering of water was organized with the screw of other (large) diameter, so the force of weight is working to rotate this screw by the weight of the falling water. The paradox of these systems will be removed if to consider them as systems of variable topology, as it was done at the beginning of the present article. Practically, the **cycle is separated here into two processes** (the lifting and falling of mass) in the system with different parameters (the topology) for the first stage and the second stage of the cycle, but herewith **it is necessary to consider two different physical systems, not one and the same.** In that case, the classical theory can explain the work created by potential gravitational field of our planet as energy exchange between two different systems.

Other well known topic is research projects by John Searle, i.e. the "Searle's disks" (Fig. 6). It is necessary to note that the inventor mentioned in his articles anti-gravitational effect and free energy output also. The address is: John Searle, 13 Blackburn Lower Strand, Graham Park Estate, London MW9 5 NG, United Kingdom.

In several words we can say that rotor makes free electrons to be displaced to the peripheries of the system. With sufficient velocity it was mentioned that there is phosphorescence and ionization around the disk. It is possible to assume that main effect is based on well known Lorenz forces and understanding of the Poynting vector, which is circulating in this system. The ionization currents are closed through the space from periphery to the center, and self-rotation of the disk is provided by the classical Lorenz force, since the current interacts with magnetic field of the rollers.

In Russian experiments of 1992 Roshin and Godin built similar system of 7KWtt power output and they claimed that 100 kg axial force and areas of decreased temperature in environmental also were detected. Fig.7 demonstrates main parts of the system by Godin and Roshin.

It is necessary to note that similar anti-gravitational effects appearing for the case of over-unity operation were observed by different inventors independently. For example, in 1990 Floyd Sweet demonstrated his invention named as "vacuum triode amplifier" VTA. The barium magnets were pre-conditioned by special method to be used in special "trigger mode". This "bi-stable condition of magnet" provides possibility of transition from one direction of field to another direction due to the weak control signal, which was provided from external generator. It is known that if the material was pre-conditioned by the magnetic switching of 60 Hz frequency then its control signal must have the same 60 Hz frequency. A part of output power was closed to provide feedback and additionally some power can be used in output coil for the load. Tom Bearden studied the scheme of vacuum triode amplifier and confirmed that it demonstrates the work with negative energy. It means that the work in the load is connected with use of negative time. In this negative time, according to Bearden, the gravity is repulsing force. The experiments on VTA demonstrated that VTA decreased its weight according to level of the power, which is extracted from vacuum. Additionally we can

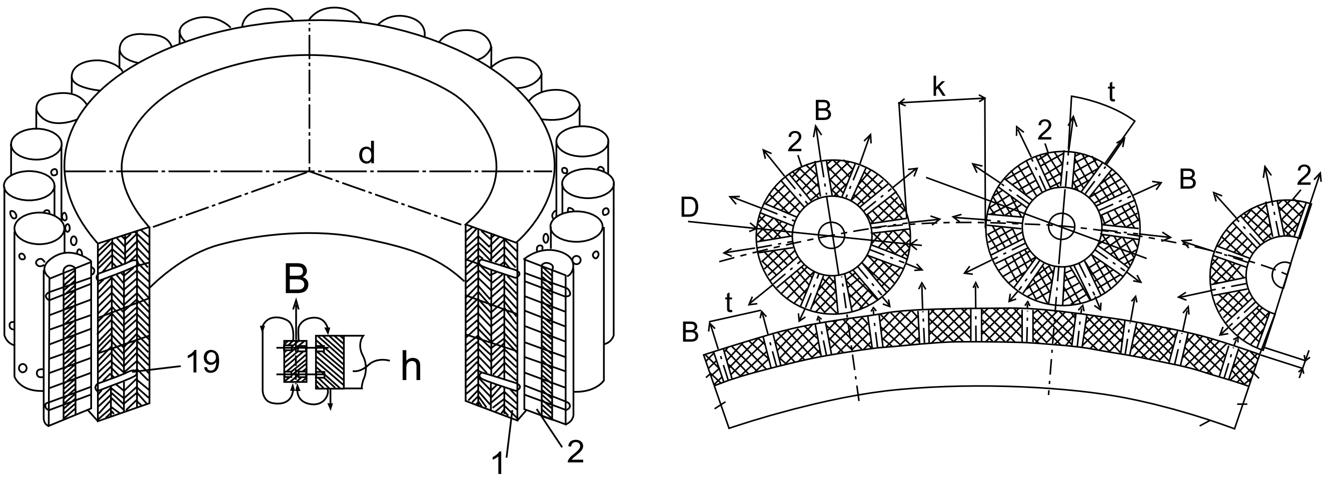

Fig.7

say that permanent magnets and coils of VTA were self-cooling during operation and the temperature difference was about 20 degrees in contrast with the environmental temperature.

One of VTA schemes includes two sets of magnets 4 x 6 x 1 inch, placed on two walls of frame. The attraction is created between them. Output and controlling coils are located between them. The axes of output coils are parallel to the magnetic field lines, but the axis of control coils are placed under 90 degrees angle. The secret of system is the “conditioning process”, which “brings the magnets to the special condition”. It is necessary to create great number of micro-cracks in the magnet due to re-orientations of magnetic domains. In such “half-magnet” the domains get the ability to be orientated in the same direction in a weak “control” magnetic field. In fact, they are not domains but large parts of magnet’s material, separated by micro-cracks, that is to say acoustic domains. Many researchers repeat Floyd’s works. Let’s note that arc discharge of alternating current through magnet barium ceramics directly can provide the best results of the “conditioning” process. The coil of “conditioning” is not necessary in this case. The frequency of alternating current must be corresponded to the frequency of the control signal. Thereby, Sweet was creating **the bi-stable solid-state condition of magnetic substance.** The acoustic resonance is the reason of oscillations with the frequency according to the control weak magnetic field. Floyd Sweet died on July 5, 1995 at age of 83 years old. It is known, that his widow sent the archives to some great Automobile Corporation.

... the extraction of space energy is result of transformation of vacuum energy but from the other hand the energy density in this case determines so called “time rate”

It is interesting to note again that the magnets of VTA demonstrated self-cooling during operation, up to 20 degrees difference in comparison with the environmental temperature. It is one more example of relation between notion “density of time” and “energy”. We can say that the extraction of space energy is result of transformation of vacuum energy but from the other hand the energy density in this case determines so called “time rate”.

Kozyrev’s articles [4] on causal mechanics theory and experiments describe possibility to use “time course for producing of useful work”. Kozyrev introduced notion “density of time” and he demonstrated experimentally several methods how to change the density of time, which depends on irreversible processes intensity. Powerful “generator” of such processes is biosphere of our plane and it creates **season and daily changes of the density of time**. It is known that VTA power output also was variable in different time of day and night. It can be explained by Kozyrev. Next step in logical development of this idea is **to change notions from the “density of time” to “density of aether”** [10].

Relation between magnet phenomena and aether circulations was known from the beginning of the electrodynamics and now we can assume that VTA was real example of asymmetrical cause-effect connection, where the hardness of this connection depends on the density of aether. In this case the season and daily variation of the aether density are reason of VTA output power variations. So, we can make a conclusion: conception of **asymmetrical cause-effect connections** should be used as theoretical basis of all over-unity systems.

Generation of extra power in nonlinear materials (ferrites and dielectrics) was considered by Nikolay E. Zayev, Journal of Russian Physical Ideas, #1, 1991. Discovery was claimed as “Cooling of some dielectrics by changing electric field with generation of energy”, Russia discovery #32-OT-10159, November 14, 1979; the inventions were also claimed as “Method of transformation of heat energy of dielectrics into electric energy", Russian patent claim #3601725/07 (084905), of June 4, 1983 and "Method of transformation of heat energy of ferrites into electric energy", Russian patent claim #3601726/25 (084904), of April 3, 1983. This theory is not about some transformation of space-time structure but in any case the practical application of conversion of environmental heat is very important because this work is real basis of free energy systems.

In one of Zaev's articles, which were published in Journal of Russian Physical Ideas, he quoted from K. Zialkovsky: "If the heat can be transmitted from cold body to hot body, then it must have enormous importance, and not only philosophical or scientific, but also practical importance. The Clausius postulate is not confirmed in this case. The gravitational force, as well as other reasons (number of the reasons is unknown) break this postulate... heat can be transmitted from cold body to hot body but only as the result of some exclusive conditions", published in Russian, "The Second beginning of thermodynamics", Kaluga, Russia, 1914. So, the inventor of free energy system should provide this "exclusive conditions" to organize collection and transformation of environmental heat in his free energy system.

Another well-known free energy system is Swiss electrostatic machine. In spiritual commune Methernitha, Linden in Switzerland, since 1980 several free energy devices have been generating total power of 750 kilowatts. From the technical point of view, these devices are modernized electrophore generator. Permanent magnets also included into design of these devices. The machine of 20 cm diameter produces about 200 watts, and a big machine has the disk of 3 meters and it produces about 30 kilowatts, Fig. 8.

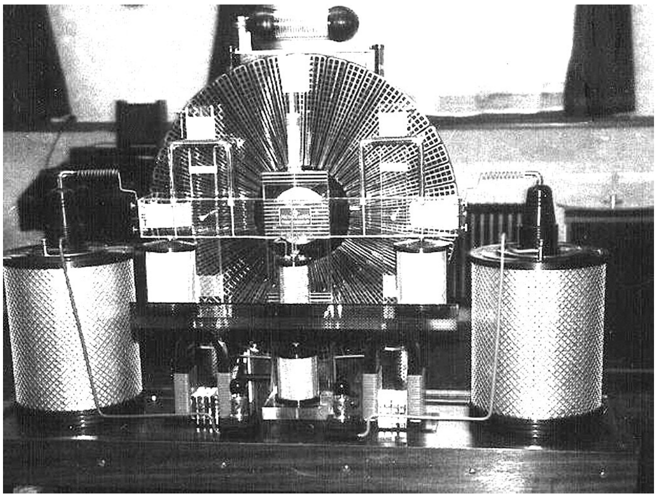

Fig.8

One of the modern technical decisions, which are very close to this Swiss machine is patent USA No. 4897592 by William Hide, January 30, 1990. This device is "system to generate power from electric field". It is one more example of work, which can be produced by potential field, in particularly by means of electric field that can be used as free source of power. The energy (the potential energy) is the possibility to produce work and the power is a work per unit of time, i.e. it is a process. The properly organized process, for example acceleration of rotor in Hide's machine, uses the potential field on the part of positive work (acceleration) and the author of the patent provides meta screen on the part of the trajectory, where the field decelerates the rotor.

... it is "perpetuum mobile" and it can not be patented even if it really works!

One more example is Reed's motor, which uses energy of permanent magnets. By the description of 1991 it is made of four disks (two immovable discs and two rotating discs), and eight magnets are placed on them. Howard Johnson used similar method, USA patent #4151431, Fig.9

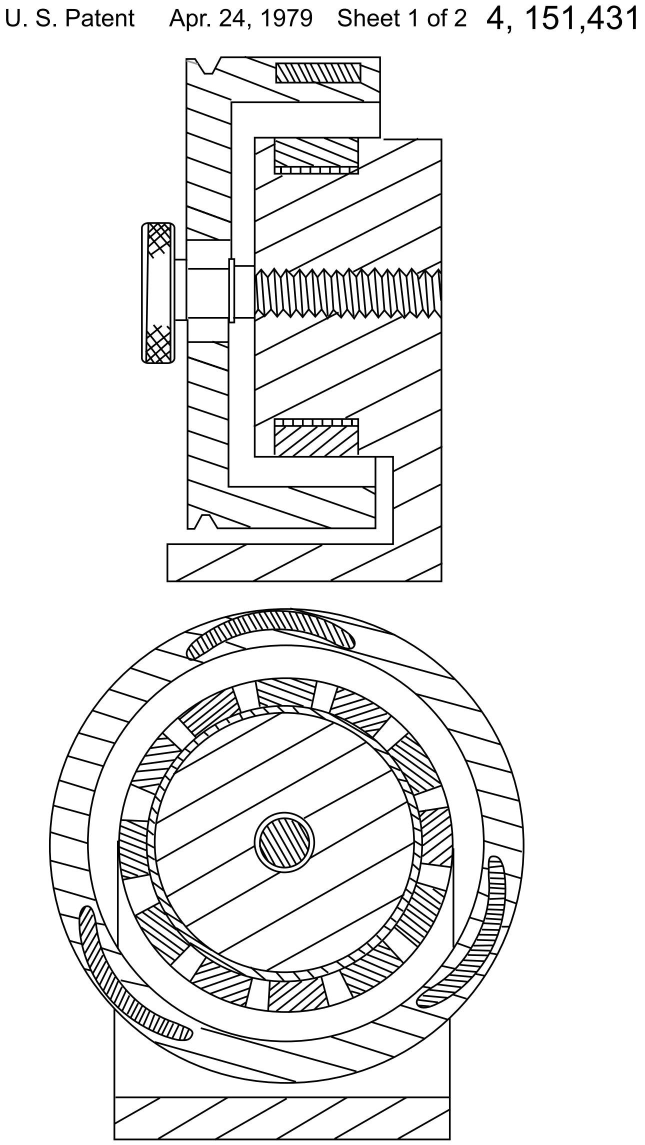

Fig.9

Journal "Science & Mechanics" of 1980 describes this invention. Hovard filled the application in 1973 but he has got the confirmation only in 1979! The reason of delay is very clear: it is "perpetuum mobile" and it can not be patented even if it really works! Johnson found clear and simple description of generation of power in his device and explained it as "extraction of power from inner energy of electron spin in ferromagnetic material".

From text of his description it follows that USA patent #4,151,431 of April 24, 1979, "Permanent magnet motor", author Howard R. Johnson, application #422,306 December 6, 1973 was developed from the prototype that is USA patent #4,074,153 of 1978, class of international classification H02K 41/00E:

"The invention is directed to the method of utilizing the unpaired electron spins in ferromagnetic and other materials as a source of magnetic fields for producing power without any electron flow as occurs in normal conductors, and to permanent magnet motors for utilizing this method to produce a power source. In the practice of the invention the unpaired electron spins occurring within permanent magnets are utilized to produce a motive power source solely through the superconducting characteristics of a permanent magnet and the magnetic flux created by the magnets are controlled and concentrated to orient the magnetic forces generated in such a manner to useful continuous work, such as the displacement of a rotor with respect to a stator. The timing and orientation of magnetic forces at the rotor and stator components produced by the permanent magnets to produce a motor is accomplished with the proper geometrical relationship of these components."

...conception of asymmetrical cause-effect connections should be used as theoretical basis of all over-unity systems.

It is reported that functioning Johnson's model produced about 5 kilowatts power free of any primary source of energy. Let's note that Johnson writes in his patent about permanent magnet as about system with "super conductive parameters". The currents of electrons in permanent magnet are manifestation of real superconductivity and for this case it is not necessary to provide the cooling for zero ohmic resistance. Moreover, the "resistance" must be negative since the magnet could save and renew its magnetized condition. Thereby, any permanent magnet is the example of perpetuum mobile of the third kind on the micro level. Also we can say the same about each atom.

Also, let's note that in general case motion (rotation) can be created due to the discussed above gradient of field, which in Johnson's device is result of asymmetry in "rotor-stator" system. By the similar way the gradient of velocity of airflow above and below wing creates the gradient of pressure and this fact produces great lifting power in airplanes.

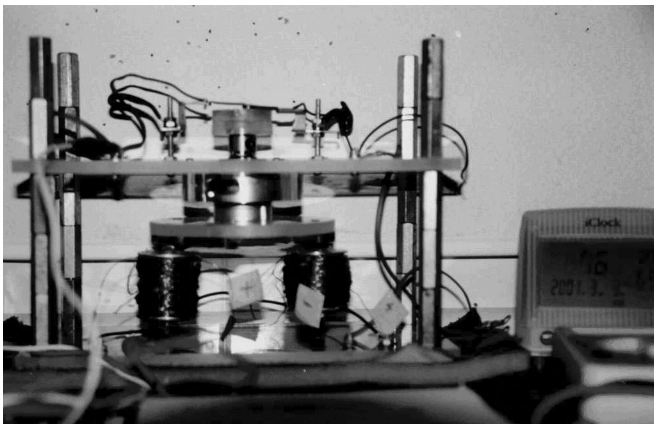

Fig.10

Well-known variant of permanent motor-generator is Adams motor, Fig.10. The rotor with radial orientated (the same pole outward) permanent magnets is rotating and creating inducted currents in stator coils, which are placed around rotor in the plane of rotation. From the point of traditional electrical engineering, any motor-generator without closed magnet flux is not high efficient device. However, exactly open magnetic flux of the Adams motor allows to take off power without deceleration of the rotor. We can assume that in this case phenomenon of electromagnetic induction is not important but in this design there is magnetic induction only, i.e. magnetization and demagnetization of cores in the field of the moving magnet. It is a perfect analogy with phenomena of electric induction that is "electrization by influence". Similar "magnetization by influence" differs from electromagnetic induction and secondary magnetic field in winding of generator is not related with deceleration of the rotor. Robert Adams works with Harold Aspden under patenting of their system. Adams is more than 70 years old but from our correspondence with him we can say that he is going to build demonstration version of 10 Kwtt generator.

...any permanent magnet is the example of perpetuum mobile of the third kind on the micro level.

There is also special name "alternators" for this class of devices, which use interruption of magnetic flux, for example it is the device by USA patent of John Echlin #4567407.

The experiments to investigate the alternator principles were organized also by the author of this article and it was demonstrated that ferrite core of the generator coil is self-cooling. The simplest experiment is based on electro motor, which rotates iron plate and it periodically appears in the gap between magnet and coil. But it is necessary to note that change of the flux in the coil area should be organized by such a way to decrease the flux of the field in the rapprochement half-cycle and to increase the flux for the moving off half-cycle. In this case the rotor is accelerated by the secondary field (back-EMF). From 1994 to 2003 several experiments were produced and main principles were claimed in the patent description, Fig.11.

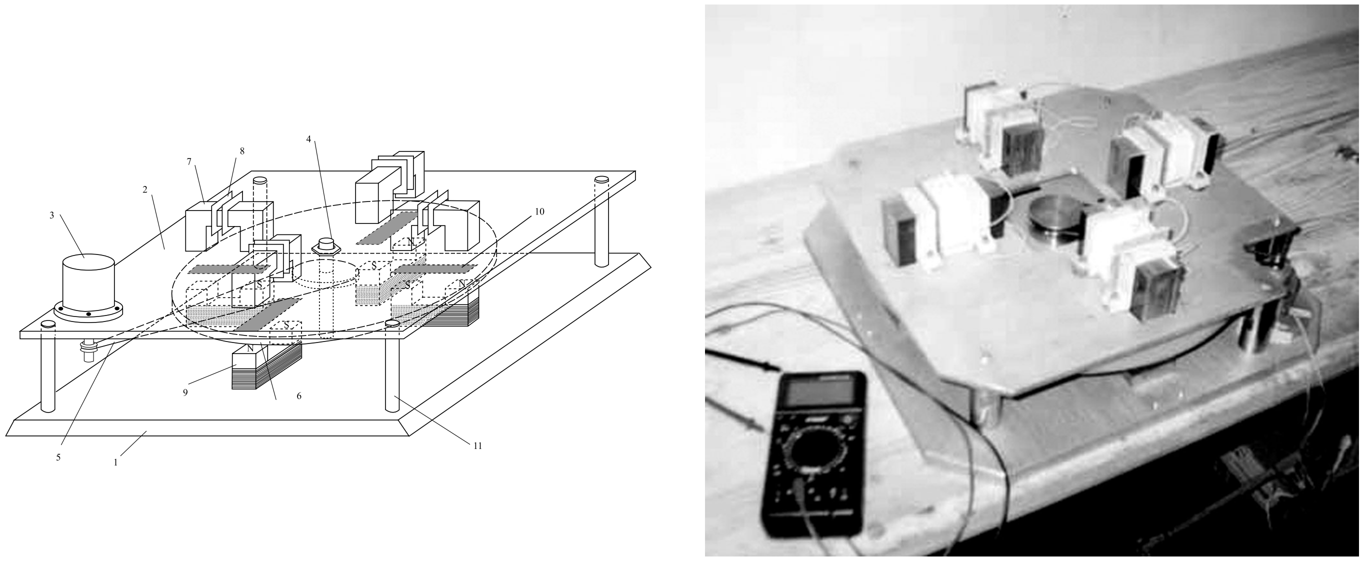

Fig.11

One more topic is "extraction of power from air" and Josef Swenson has conducted the series of simple experiments to develop it. The frequency of natural pulsations of electric field of planet is about 7.5 Hz and it is well known from Tesla's age. Swenson works with frequency 375 kilohertz and antenna of 10 meters. Please, contact for more details: Josef Swenson 423 North 15th Street, Moorhead, Minnesota 56560, USA. However, everybody remembers from the school story about simplest electrical experiments by Lomonosov and Rihman, who investigated arc discharge in gap between iron wire from a roof ("antenna") and ground wire. Let's include the resonance circuit "inductance-capacity" and diode rectifier to get some useful work in the load "from atmospheric electricity", Fig.12.

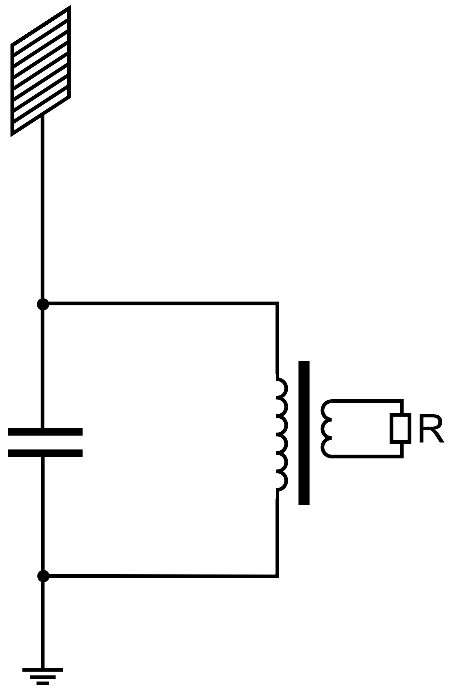

Fig.12

In 1900-1930 a lot of articles were published in technical press about Henry Moray. His demonstration systems produced more than 50 kilowatts free power output. It is known that Mr. Yakovlev (from USSR Foreign Department headed by Mr. Molotov) visited Morey in November of 1929 in New York to test his devices. The devices consist of capacitors, coils and special electronic-vacuum lamps.

In 1990 journal "Magnets", 2 (3) published article, which describe analogy between Moray's devices and Hubbard's coils, which can extract power by means of inner energy of nucleuses of materials if special resonance is created. In 1978 Cospray Research Institute has published the well-known book "The Sea of Energy" by T. N. Moray, in which theory of Moray is presented most completely.

RQM Raum-Quanten-Motoren Corporation (Schmiedgasse 48, CH-8640 Rapperswil, Switzerland, fax 41-55-2125209) offers for free energy devices of different power level: RQM 25 kilowatt and RQM 200 kilowatt. The principle of work is based on an invention by Oliver Crane and his theory. Web site http://www.rqm.ch. But I have to note that in present time they are developing capitalization of the company mainly (actives are above 650 million dollars) than innovation of the new technology.

One more historical example: in 1925-1945, Hans Koler demonstrated his free energy devices. The system was built in Germany and it produced 60 kilowatts of free power. The description includes six permanent magnets and coils, which were superposed in plane of hexagon form. Each magnet is core of the coil.

Another interesting topic is unipolar induction effect, which is well-known from Faraday. This effect creates electro motive force (EMF) in rotating disk if axial magnetic field is applied to the disk. One of the well-known practical developments in this area is device by Bruce de Palma. In 1991 he has published the results of his tests, from which it follows that deceleration of the rotor due to back EMF in the case of unipolar induction is less than in traditional electro generators. So output power of system can exceed the power, which is necessary to rotate the rotor.

Really, motion of electrons in magnetic field, which is perpendicular to plane of rotation, is the reason of the Lorenz force and this force acts in radial direction that can not be the reason of deceleration. Indian research on this topic is developing by P. Tewari. In New Zealand there is research group headed by Dr. Ashley Gray. In 1994 the leader of Japanese market MITI published report about progress in 40 KWtt unipolar generator, which uses superconductors for its electromagnets. The interest of Japan to alternative energy projects can be explained by the position of Japan on fuel market.

There is a well known rule: Demand is related with Proposals. It is easy to imagine the prospects of local introducing of free energy systems in one or several countries, if some producers of product will be able to exclude expenses on electricity and fuel from cost sales. Other countries of their own rich natural resources (for example, oil) will be in problematic position on the new international market, mainly due to the fact that their industry and transport are oriented to conversion and consumption of oil fuel that increase the cost of all products.

... increase of the spark gap produces more surplus power in the load of the circuit.

One more modern free device was invented by Wingate Lambertson, USA. In his device free electrons get the additional energy passing through the number of thin metal-ceramic composed layers. The units were designed by the authors and each unit can generate 1600 watts, and it is possible to connect them in parallel. The address of author: Dr. Wingate Lambertson, 216 83rd Street, Holmes Beach, Florida 34217, USA.

Especially the researches on free energy with plasma processes should be noted here. In 1980-1990 Alexander Chernetsky, Yury Galkin and others have published the results of experiments on creation of "self generated discharge" SGD. The electric arc was placed consecutively into secondary circuit of electromagnetic transformer and it produces real increase of power in load and reduction of consumption power in primary circuit of transformer. The author of the present article produced simplest experiments to investigate the arc (electric discharge) in electric circuits and possibility to create the mode of "negative resistance" in this circuit was confirmed. One of the effects was demonstrated in 1996 during the conferences "New Ideas in Natural Science", St.-Petersburg.

Adjusting parameters of arc (distance between two electrodes) it is possible to see that consumption current is decreasing until zero and then it can change its direction i.e. this system begins to generate the power. During similar experiment of 1971 by Dr. Chernetsky substation transformer in Moscow Aviation Institute was destroyed in result of strong "reversed current" impulse, which exceeded consumed power in 10 times more. According to Chernetsky's concept, the reason of this mode is well known phenomenon of plasma instability and **pinch-effect for great currents.** However, the author of this article tested device, which demonstrated similar effect (switching on the load in secondary circuit of transformer and in the presence of arc in this circuit, consumption power does not increase, but reduces) for small currents about 300 mA.

Since for pinch-effect it is necessary hundreds Amperes then it was offered another explanation: the surplus power in this circuit appears due to the acceleration of electrons in the gap between electrodes, i.e. particles of plasma are **accelerated by means of electric potential field between two electrodes**. It is noted during the experiments that increase of the spark gap produces more **surplus power in the load of the circuit.** To avoid mistakes the measurements of the consumed power were organized in DC (direct current) battery circuit and therefore there is no any reason to speak of phase shifts mistakes to try to explain skeptically this effect.

... waves of density of time are used by organisms for their vital activity.

Today theory and experiments on self-generating discharge are quite well developed to build free energy systems of any power scale. The reason of delay in its practical development is a complex problem: this work leaves the frames of classical physics. In his book "About physical nature of bio-energy and its simulation", Moscow, Publ. VZPI, 1989, Dr. Chernetsky considered the structure of biological fields and bio-energy processes in living organisms from the point of **longitudinal waves conception.** Self-generating discharge in the mode of negative resistance produces such longitudinal waves and they are self-sustaining (self-powered energetically) and it is considered as field of living object.

Really experimentators of the Chernetsky's group, who worked with SGD device detected influence of biologically active radiations and fields, which can be not screened by usual methods. It was reported by Chernetsky that parameters of this radiation can be selected to accelerate the development of plants and biomass or to suppress it. So, we should speak about artificial living biological system or biological form of energy for this class of free energy devices. Perhaps by the same plasma oscillation way all living organisms provide its vital activity, since long time ago it was known that metabolism and food can not provide enough energy for vital processes.

Nikolay A. Kozyrev also wrote about "reason of life" and he confirmed that waves of density of time are used by organisms for their vital activity. Between "waves of density of time" and "waves with longitudal component" there is a clear analogy. Kozyrev and Chernetsky experimentally demonstrated methods of creation of such waves.

...any free energy system should change the causality in surrounding space-time.

New Energy News magazine, March 1996, wrote about attempts to innovate for American aerospace stations special power source, which uses similar plasma technology: anomalous electric discharge. The USA patents #5416391 and #5449989 belong to Dr. Paulo Correa and Dr. Alexandra Correa, Canada. In their report on free energy Denver conference it was claimed that efficiency is about 483% .

Let's note one more free energy topic: high efficient electrolysis. Classical electrolysis as decomposition of electrolyte in electric field is a wonderful example of work, which is produced by field and it can be organized by such a way to be free from the primary energy source (battery). The traditional scheme uses closed circuit of current through electrolyte and the battery, so the battery is discharged during the operation with electrolytic cell. However any physics textbook confirms that **ions in electrolyte are moving due to electric field only, i.e. work to organize the displacement of ions and heat power, which is connected with this work, are produced by the potential field and expenses of the primary power are not required.**

The current through the battery, which is created in usual closed electric circuit destroys the primary difference of potentials in the battery but it is not some necessary condition. For correct organization of the experiment the products of electrolysis (gases) can be created almost free and their utilization (burning) can provide more heat power than input electric power.

Prof. Latchinov, the real member of Russian Physic-Chemical Society, who patented his method of electrolysis in 1888, mentioned that in some cases the electrolytic cell is freezing during its operation. This is an effect of the law of Conservation, which forces to return the surplus power by means of heat environmental energy. In the experiment by Latchinov the electrolytic cell can create gases of high pressure and consumed power is the same as for the case of low pressure gases. However, it is clear that high pressure gases can produce more work than low pressure gases do. This question was sensational problem in scientific societies of 1888 and it is not solved yet completely.

One of other variants of high efficient electrolysis was investigated by Igor Goriatchev, Russia. Instead of usual 3 Volts level of electrolysis, Goriatchev uses 0.2 Volt pulsing mode. He claimed ratio output/input as 1500% efficiency and he hopes to increase this value twice in more perfect design.

Professor Kanarev from Krasnodar wrote about his experiments on plasma electrolysis and has proved that in electrolysis of water output power can be more than input power.

Real example of "perpetuum mobile of the second kind" is invention by Russian engineer Albert Serogodsky (Moscow) and German engineer Bernard Sheffer (Berlin). They have patented new system for direct transformation of environmental heat into electricity, Germany patent #4244016. Retro-condensation of mixture of benzine and water is organized in closed system under temperature of 154 ^{0}C degrees. You can try to get more info from: Werkstatt fur Dezentrale Energleforschung, Pasewaldtstrasse 7, 14169 Berlin, Germany.

The fundamental theoretical researches on direct transformation of environmental heat to useful work were provided by Real Member of The Russian Physical Society Mr. Gennady N. Buynov, St.-Petersburg. The description of his project on "The Mono-thermal device" was published in journal "Russian Ideas" #2, 1992.

In 1995 the scientific journal of Russian Physical Society #1-6 published article "Perpetuum mobile of the second kind (paired gas-chemical cycle)". The author Gennady N. Buynov proposed to consider that **entropy function can be abortive**, i.e. it can be undetermined in some place if reversible chemical reactions are presented in the system.

Herewith, the circular integral of entropy is not a zero and in this case the heat function but not entropy function (according to the Gess law) becomes the function of condition.

Buynov offers to use four–oxide of nitrogen as working matter for this cycle. His works are excellent example of scientific enthusiasm, which (in combination with financial interest of the customers) could produce real results for Russia many years ago. Let's note that priority of this discovery belongs to Russia in spite of attempts of other scientists to claim and develop this idea now as their own work.

We can make a very interesting conclusion if we analyze history of so called "cold fusion" discovery. According to de-classified materials of 1960, priorities of Russia in this topic are obvious. In 1989 Pons and Fleshman have reported about results of their experiment. In 1995 Russian Journal "Inventor and Rationalizator", #1 has published article about invention made by Ivan S. Filimonenko that was named in 1957 as "warm nuclear syntheses". In 1957 he detected **extra power output in process of heavy water electrolysis** and he mentioned several applied aspects, for example, propulsion force and possibility to reduce radioactivity by means of this process. In 1960 Kurchatov, Korolev and Zhukov have supported the ideas of the author. Russian Government adopted secret resolution on this work:

1. Investigation of energy generation.
2. Development of propulsion without reactive mass flow.
3. Research on protection from radioactivity.

But in next several years this work was suppressed by people of nuclear power plant team. The only system of such type named as TOPAZ was innovated for Russian aerospace systems. World-wide innovation of this technology is a real way to introduce "warm reactors of syntheses" and it is not necessary to wait for results of high-cost "Tokomak" project and other thermonuclear researches. Let's note that secondary effects (propulsion force and influence on radioactivity) are possible due to using of "free energy" aspect if output power is result of change of space-time parameters.

In 1994 Journal "Russian Ideas", #1-6, it was published an interesting document "Conclusion of The Moscow City Council Commission on the question about Ivan Filimonenko's discovery and its development". In this document it was recognized that it is vitally necessary to renew this works. But in 2003 we still have nothing new on this topic. Why? It can be assumed that the Problem of innovation for this technology is possible military application of the methods since influence on radioactivity (for example, remote reduction of radioactivity of some object) is area of interests of the Defense Department. The fact that energy generators by Filimonenko can be used for quick restoration of the present ecological balance is not so important in this case. The same conclusion is about propulsion method, which was proposed by Filimonenko. Mr. Korolev knew about this method; however present space programs are still based on rockets and reactive principles, and anti-gravity flying machines we can see only in fantastic movies.

At the same time, development of commercial cold fusion projects was started in some countries, for example: Patterson Power Cell is introduced in Texas, USA (Clean Energy Technologies Inc., Dallas, Texas, fax 214-458-7690). More than thirty patents were owned by ENECO Corporation, which is collecting the main technological solutions in this area. The production of electrolytic thermal cells was started by Nova Resources Group., Inc., Colorado.

In August of 1995 Atomic Energy of Canada, Ltd. Company, which is member of The Planetary Association for Clean Energy, has published the review on modern methods of conversion of nucleus wastes and deactivation. Two new technologies were offered for introduction: contact processing by "Brown's gas" and remote processing by scalar (torsion) fields. Let's note that the technology proposed by Canadians and Filimonenko's technology demonstrate the effect of influence on the rates of radioactive decay.

These examples are only small part of real situation. Main references on publications are foreign and it can lead to wrong conclusion that Russia is delayed in this direction of new technologies development. In fact, Russia has more talented inventors and researchers than any other country. However, condition for work, patenting and publications of ideas are not the same, and usually Russian technologies cannot reach the level of international market. This problem depends only on real and official state policy with respect to inventors and scientists. In real life financial support of scientific institutes is mainly subsidy for management of the institutes, but not for science. The inventions and discoveries always were made by a certain real person, but not by Institute or some scientific team.

In Russia of last age and in the other world on the whole the institutes and laboratories have been creating for a new scientific problems, discoveries or new directions in science. It was necessary to claim about discovery and provide priority for own country to get official support. Main schools of thought appear by this natural way. By the same natural way the necessary in existing of some scientific institute can be removed when the idea grows up to serial production stage. If there are no any fresh ideas in this school then the institute should be transformed in design office of the production plant on this topic. It is almost impossible to create a new research institute in modern Russia, so really new ideas (if they are not ranged in frames of some existing scientific directions) can not be developed and they can not create a new school of thought in Russia.

... it is possible to make a conclusion about grandiose misinformation of society...

People (carriers of the new ideas) have to leave Russia to realize their sensational ideas on free energy and antigravitation topics. Why they do not work with Russian Academy of Sciences? It is the rhetorical question. Usually after many years of correspondence with patent office or with official science bodies the authors can find that their ideas are published in “serious scientific magazines” but without any respect to their names...

Attachment 1 presents some information about patented technologies. When we study old and modern patent documents, it is possible to make a conclusion about **grandiose misinformation of society**, which leads to appearance of two different Worlds, separated by the level of knowledge: evident and hidden knowledge. The achievements of the second hidden world could change our planet, give a chance to remove all energy and ecological problems. Besides, we have discussed that some free energy systems (for example, self generated plasma discharge) have also the medical-biological aspects. This “influence” from free energy system is related with negative entropy processes in area of the operation of the system so some components of biological systems can be changed to less entropy. The design of free energy system defines the type of this influence. It was noted before that operation of any free energy system should be considered in high topology space-time as multidimensional system to see what the reason of the effect is.

So, since the rate of time course is determined by Nikolay A. Kozyrev as rate of cause-effect transformations, then any free energy system should change the causality in surrounding space-time. In this case we can assume that it creates quantum physics effects on the macro level. It is interesting to assume Heisenberg’s uncertainty, tunnel effect and manifestation of wave characteristics (up to diffraction of objects) for macro level.

Some experimental data by Kozyrev is related with quantum effects, which were detected in his experiments with rotating and vibrating gyroscopes. This technology is the basis to design real teleportation systems, which will change parameters of space-time to allow changes of position of some material object from one point of space to another point of space (without transference on the distance between these points) by means of combining these points in the same place of space for some time.

The process of study of new technologies in alternative power industry and gravitation is developing actively in the world. Besides secret programs and institutes, it is possible to make a conclusion that activity of this work in countries of limited fuel resources is more efficient. Now Russia has rich **natural materials and oil resources but in the short time this advantage will not play any role in development of economy.**

Industrial and defense power of any country will depend on free energy technologies, knowledge on the biologically active energy and reactionless propulsion methods.

The new technology creates not only a new technical systems but also **new economical systems**. It is not physics but policy... Development of great Russian territories with industry of free energy technologies for home and commercial application is the way to change world economical balance in favor of Russia. It is difficult to say now what financial-economical group of modern Russia is most interested in development of this direction of science. The purpose of any financial-economical group is power on the market of energy resources, but when the free energy technologies are developing, then people and industrial or agricultural producers will be more independent from centralized system of energy and oil distribution, therefore, they will be more independent from influences of central authorities.

From the point of view of serious business, there is only one real argument in favor of developments of any really new technology: it should lead to increase of profit and expansion of the market. In this sense, it is possible to compare the new energy technologies with beginning of the steam machine age or with appearance of electric machines and illumination. This means the super profits and serious fight with competitors. To develop this way any efforts of scientists are not sufficient work. It is necessary to join them with the efforts of large business structures, which are interested to create new market of energy and power engineering in parallel to present monopoly fuel-energy market or to develop commercial exploration of space by means of new propulsion principles. In particular, telecommunication space satellites and projects on colonization of space could be the nearest commercial directions of the new market.

References

1. Sakharov A.D. “Quantum Fluctuations of Vacuum in Curved Space and Theory of Gravitation”, Reports of Academy of Sciences USSR, vol. 12, 1968, page 1040.
2. Puthoff H.T., Ph. Review E, vol.48, num.2, p.1562-1565, “Extracting energy and heart from the vacuum”; Physical Review A, vol. 39, num. 5, “Gravity as a zero-point-fluctuation force”.
3. Nassikas A.A., The Hypothesis of the Unified Field and the Principle of its Dual Interpretation, III International Conference “Problems of Space, Time, Gravitation”, St.-Petersburg, 1994.
4. N.A. Kozyrev, Selected works, published by Leningrad State University, 1991, p.202, in Russian.
5. Gravitobiology, by T.E. Bearden, 1991, Tesla Book Company, P.O. Box 121873 Chula Vista, CA 91912, USA. Library of Congress Catalog Card Number 86-50553.
6. Hasselberg Josef, New Beginning of Thermodynamics, http://hasslberger.com, Hasslberger, Josef “The inventor and society” in Raum&Zeit #4, October 1989, Hasslberger, Josef “Vortex, the natural movement” in EXPLORE, No.5, Vol.3, 1992.

7. Gustav Naan, Symmetrcial Universe, Tarty Astronomical Observatory, Vol. XXXIV, 1964.
8. New Results of Development and Testing of Single-Wire Electric Power System, Prof. Dmitry S. Strebkov, Eng. Stanislav V. Avramenko, Dr. Aleksey I. Nekrasov, Eng. Oleg A. Roschin, New Energy Technologies, #5(8), 2002.
9. An Experimental investigation of the Physical Effects In a Dynamic Magnetic System. V.V. Roschin and S.M. Godin, New Energy Technologies, #1, 2001.
10. Practical application of time rate control theory. Alexander V. Frolov, New Energy Technologies, #3, 2001.

Attachment 1

Methods and devices to create power without external sources and reactionless motion systems

3913004 October 14, 1975, Method and equipment to increase electric power, Robert Aleksander.
4975608 December 4, 1990, Motor with switchable magnetic resistance, Harold Aspden.
5288336 Converters of heat into electricity, Harold Aspden, see also patents 5,065,085 and 5,101,632
4622510 November 11, 1986, Parametric electric machine, Ferdinand Kap.
2912244 1959, Gravitational System, Ottis Karr.
4006401 February 1, 1977, Electromagnetic generator, V. Rivas
3811058, 3879622 Motors with permanent magnets.
1835721 December 8, 1931, Motor with permanent magnets, A. Povel.
1963213 June 19, 1934, Magnetic motor, G. Pose.
1859643 May 24, 1932, Motor with permanent magnets, G.L. Worsington.
1859764 May 24, 1932, Magnetic device, G. Baugon.
2982261 Mac Klintok's Air motor.
4595843 June 17, 1986, Transformer of rotation magnetic flux, Robert Del Vechio.
4567407 January 28, 1986, Motor - alternator, John Eklin.
3368141 January 6, 1968, Transformer in combination with permanent magnets, K. Garon.
3890548 June 17, 1975, Motor with pulsing capacitor discharge, Edwin Gray.
4595852 June 17, 1986, Electrostatic generator, Robert Gandlach.
4831299 May 16, 1989, Homopolar generator of alternating current, Enakishy Khasaka.
4249096 February 3, 1981, Electric dynamo, Barbara Nikoks.
3610971 October 5, 1971, Electromotive generator of electric field, Williams Couper.
4897592 January 30, 1990, System producing power from energy of electrostatic field, Williams Hyde.
4151431 April 24, 1979, Motor with permanent magnets, Hovard Johnson.
4806834 February 21, 1989, Electric circuit of inductive conductors, transformers and motors, Erl Kening.
3374376 March 19, 1968, Electric generator, Raymond Kromry.
3977191 August 31, 1976, Power source, Robert Brett.
3670494, Method of converting of atomic energy in kinetic energy.
4709323 November 24, 1987, Converter of parallel resonance, Charles Lien.
5146395 September 8, 1992, Power source using two accumulating circuits, Richard Mac Kee.
4210859 June 1, 1980, Inductive device with two orthogonal windings, Paul Meretsky.
4500827 February 19, 1985, Linear electric generator, Thomas Merit.
4904926 February 27, 1990, Electric generator of magnetic motion, Mario Patsishinsky.
4945273 July 1990, High effective electric machine, Josef Pinkertone.
4883977 November 28, 1989, Converter of magnetic power, Dennis Regan.
4077001 Electromagnetic converter, Frank Richardson.
5018180 May 21, 1991, Conversion of energy, Kennet Shoulders.
4652771 March 24, 1987, Transformer, Theodore Speach.
477 2816 September 20, 1988, Conversion of energy, Jefry Spens.
4748311 May 31, 1988, Inverter, Fridrikh-Verner Thomas.
International patent H02K 31/00, 39/00 dated June 24 1982, Closed part of unipolar machine, Adam Trombly.
4687947 August 18, 1987, Electric circuit for conserving of power, Melvin Kobb.
4772775 September 20, 1988, Generation of plasma flux in electric arc, Sam Lich.
The USA patents 5416391 and 5449989, Paulo Correa.
4432098 and 4429280, Transmission of information by means of magnetic vector potential, Raynolds Gelinas.
Great Britain, #547668, January 30 (September 7) 1942, Motor with permanent magnets, Stenly Hichkok.
Great Britain, application # 2282708A, Motor with permanent magnets, Robert Adams, Harold Aspden.
4394230 USA patent, July 19, 1983, Henry K. Pukharich.
2251775 Great Britain patent, April 20, 1994, Thermoelectric conversion, Harold Aspden.
5288336 USA patent, Thermoelectric conversion, Harold Aspden.

The USA patents on electrogravitation

1363037 Goddard, December 21, 1920;
2004352 Simon, June 11, 1935;

2210918 Karlovitz, August 13, 1940;
2588427 Stringfield, March 11, 1952;
2231877 Bennet, 18 February, 1941;
2279586 Bennet, 14 April, 1942;
2305500 Slayter, December 15, 1942.
English patent 300,311 August 15, 1927, Townsend Braun.
French patent 1003484 November 1951, Electrogravitation;
3187206 June 1, 1965, Electrokinetic equipment, Townsend Brown.
3022430 February 20, 1962, Electrokinetic generator, Townsend Brown.
3018394 January 23, 1962, Electrokinetic converter, Townsend Brown.
2949550 August 16, 1960, Electrokinetic equipment, Townsend Brown.
1974483 September 25, 1934, Electrostatic motor, Townsend Brown.

Attachment 2

Organizations

- Faraday Lab Ltd, Lev Tolstoy Str. 7, St. Petersburg, 197376, Russia.
- The Russian Physics Society, 141002, Moscow reg., Mytishy, B. Sharapovskaya str. 3. Fax 095-2926511
- Academy for Future Sciences, P.O. Box FE, Los Gatos, CA 95031, USA.
- AERI, Advanced Energy Research Institute, 14 Devonshire Mews West, London W1N 1Fp, Great Britain.
- ADAS, Association of Distinguished American Scientists, P.O. Box 1472, Huntsville, AL 35807, USA.
- Borderland Sciences Research Foundation, P.O. Box 429, Garberville, CA 95440-0429, USA.
- Center for Action, P.O. Box 472, HCR 31, Sandy Valley, NT 89019, USA.
- Electrodynamics Gravity, Inc., 35 W. Tallmadge Ave., Akron, Ohio 44310, USA.
- Fusion Information Center, P.O. Box 58639, Salt Lake City, Utah 84158-0369, USA
- Gravity Power Research Association, 36 Mountain Road, Burlington, MA 01803, USA.
- GRI, Group Research Institute, P.O. Box 438, Nelson, New Zealand. Dr. Ashley Gray.
- High Energy Enterprises, P.O. Box 5636, Security, CO 80931, USA. Fax 719-4750582.
- Institute for Advanced Studies at Austin, 4030 Braker Lane W., Suite 300, Austin, TX 78759, USA.
- INE, Institute for New Energy, 1304 South College Avenue, Fort Collins, CO 80524, USA.
- Integrity Institute, 1377 K Street, NW, Suite 16, Washington DC, USA. Fax 202-543-3069.
- Orgone Biophysical Research Laboratory, Inc., P.O. Box 1395, E1 Cerrito, CA 94530, USA.
- Quantum Biology Research Laboratory, P.O. Box 60653, Palo Alto, CA 94306, USA.
- SEA, Space Energy Association, P.O. Box 11422, Clearwater, FL 34616, USA.
- Tesla Book Company, P.O. Box 121873, Chula Vista, CA 91912, USA.
- Tesla Incorporated, 760 Prairie Av., Craig, CO 81625, USA. Fax 303-824-7864.
- Journal Explore, P.O. Box 1508, Mount Vernon, Washington 98273, USA.
- Electric Spacecraft Journal, P.O. Box 18387, Asheville, NC 28814, USA. Fax 704-683-3511.
- Nexus New Times Magazine, P.O. Box 30, Maplepton Old, 4560, Australia. Fax 074-429381.
- Journal Cold Fusion Times, P.O. Box 81135, Wellesley Hills MA 02181, USA.
- Journal Infinite Energy, P.O. Box 2816, Concord, NH 03302-2816, USA.
- Journal 21st Century Science & Technology, P.O. Box 16285, Washington, DC, 20041, USA.
- Journal Cold Fusion, 70b Route 202N, Petersborough, NH 03458, USA.
- Brown's Gas International, 5063 Densmore Av., ENCINO, California 91436 USA.
- ENECO Inc., 391-B Chipeta Way, Salt Lake City, Utah 84108, USA. Fax 801-5836245.
- "Robert Adams and Company" 46 Landing Road, Whakatane, Bay of Plenty, New Zealand. Robert Adams,
- Methernitha, 3517 Linden, Switzeland. Manager Francis Bosshard.
- Swiss Association for Free Energy, P.O. Box 10, 5704, Egliswilli, Switzeland.
- Space Research Institute, Box 33, Uwajima, Ehime 79, Japan. Dr. Shinichi Seike. Fax 895-24-7325.
- Nuclear Power Corporation, 581 400 Karnataka, India. Project Director Dr. Paramahamsa Tewari.
- Cosmic Energy Foundation, Neptunuslaan 11, 3318 E1 Dordrecht, Netherlands. Dr. Martin Holwerda
- Sabberton Research, P.O. Box 35, Southampton S097BU, England, Dr. Harold Aspden.
- Labofex, Experimental and Applied Plasma Physics, Ontario, Canada, Dr. Paulo N. Correa.

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory
- Gravitation theory
- Mathematic Biology
- Gravitation energetics
- Hydrogen-Helium Energetic
- Information technologies
- Telecommunication technologies
- Transport technologies

Some General Properties of Matter and Energy Sources

Devoted to memory of Academician A.A. Harkevich

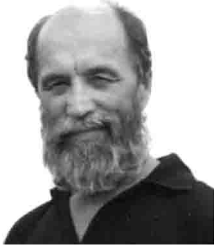

Dr. Eugene D. Sorokodum

Volochaevskaya Street, 40-b, Flat 38,
111033, Moscow, Russia
Telephone: (7) - (095) -362 80 84
Email: e40dum@dol.ru http://vortex.viptop.ru

Introduction

Nowadays there are appear more and more attempts to disclose new methods of obtaining energy from the ambient space, i.e. from heat, electromagnetic, gravitational fields, physical vacuum, aether, etc. (See, for example, publications in New Energy Technologies [1-3]). The publications demonstrate that researchers meet various incompletely investigated physical processes. These scientific areas are difficult to be investigated. All this makes obstacles for development of works on creation of new energy sources. **In this work there is made an attempt to demonstrate general properties of energy sources of any physical nature. This attempt is based on a general analysis of properties of matter.** A theory of transformers is used to investigate general properties of matter.

Action of all physical, technical and biological systems is expressed in energy transformation. Numerous theories based on particular (specific for some system) methods are applied for description of these systems. However, if the number of degrees of freedom and of elements inside a system increases then many theories are not able to describe operation of the systems. First these problems appeared in electrical engineering, radio engineering, automation and acoustics. The theory of transformers is applied to these fields of science. The theory represents a complicated system as a "black box" having several inputs and outputs. Operation of the numerous elements occurring inside the box is represented as some equivalent functions reduced to the inputs and outputs.

In the last decades methods of solution of mechanical dynamic tasks by the method of complex resistances are developed as well as representation of elements as linear transformers [11] and finite elements in liquid [14]. This tendency can be applied to the mechanics of liquids and gases. Now there are successfully developed those concepts which assume observation of models having very few degrees of freedom to be enough for analyzing processes in hydrodynamic systems [10]. However, a mathematical apparatus for description of transformers operation which is well-developed in these areas of science and engineering has a special view and is applicable only for these scientific areas.

Academician A.A. Harkevich developed a theory of a linear transformer up to the level of the general theory of transformers which is applicable for transformation of any types of energy [24]. The general theory of transformers proposed by A. A. Harkevich is applicable for description of various energy sources, flying and swimming objects, functioning of different animals' organs, and technological processes. In this work some general properties of matter and energy sources are investigated, according to the general theory of transformers.

General theory of energy transformer

The whole ambient space, from the microworld to the macroworld, is filled with energy. According to different theories, space is represented as a compact medium (i.e. having distributed parameters) or a medium consisting of a limited number of discrete elements (i.e. having concentrated parameters). On the analogy of hydrodynamics [10, 14] the compact medium can be represented as an equivalent system with a limited number of degrees of freedom. Hence, the whole space can be represented as some system consisting of elements and communications between the elements with a limited number of degrees of freedom. Energy exchange occurs due to the degrees of freedom. Energy transmission occurs if energy gradient is presented in the ambient space. Due to the energy gradient a force tends to realize transmission in the space [4]. The elements can be systems as well.

Hence, the more we observe dividing the elements on the systems and the systems on the elements the more we will penetrate into the microworld (i.e. atoms, elementary particles, physical vacuum, aether, etc.). The more we combine the elements in the systems, and the systems in the new larger systems the more we observe the macroworld (the Solar System, galaxies etc.). All the systems and the elements are interconnected. The systems and their elements are transformers of energy. Energy motion occurs in the smallest part of space. Consequently, the whole space can be represented as a system consisting of the energy transformers. In the general case due to every type of energy limits of a transformer consist of outer limits (i.e. communications with the macroworld) and inner limits (i.e. communications with the microworld). Dividing into the macroworld and the microworld occurs in relation to the size of a transformer for every type of energy. Types of energy coming through communications of a transformer and inside it may differ, i.e. mechanical, heating, electromagnetic, chemical and other known and unknown types of energy.

The processes in which a great number of interacting elements and different types of energy participate are very complicated and do not allow describe the processes accurately by modern mathematical methods. Hence, there is a problem to find such methods of solution of the tasks which without disclosing all the communications inside the element can give the understanding of the way an element moves in the system. The following premises can be made for the method of solution of the task of a system and its elements' motion:

1. Motion of all the elements in the space is characterized by energy exchange occurring among them;
2. We are interested in a certain limited area for every individual type of energy. The chosen limited area will be called as **a transformer**;
3. The transformer has degrees of freedom both on its limits (sides) and inside it (inner degrees of freedom);
4. Further this limited area (the transformer) which has or is able to have a limited number of degrees of freedom (sides) at its limit will be observed;
5. Interaction between this transformer and the ambient space occurs only through these degrees of freedom (sides) by means of energy exchange;
6. Motion of energy between the elements of the transformer occurs according to its degrees of freedom inside the transformer. There can be a limited or unlimited number of degrees of freedom.
7. All the space is full of the transformers. All the transformers adjoin each other without gaps. Energy exchange between the transformers occurs through their sides which do not have a size but reflect general kinematical and dynamic characteristics of the energy transferred through these sides.

Editor: The full variant of the article includes mathematical description of operation of energy transformers which is followed by these conclusions:

- There can be any number of different types of elements having a corresponding number of degrees of freedom and any type of energy inside the transformer. However, on every side of the transformer a generalized force has the same value and depends only on change of energy according to this degree of freedom. Therefore, two transformers are considered to be equal if the generalized forces (energy changes) on all the sides are equal. In this case it is not necessary for the inner structure of the transformer, number of inner degrees of freedom, and energy types of these two transformers to be equal. This is **a principle of equivalency of transformers at the equivalency of the forces acting at the sides of transformers**.

- In the general case the generalized force at any side depends not only on energy exchange occurring at this side but on energy exchange occurring among the sides or at other sides of the transformer. The generalized forces are produced by energy distribution in the space.
- A transformer is characterized by the fact that there are different types of energy at its different input sides (or the energy can be of the same type but having other characteristics of motion).

Resistances of the interaction provide information about physical properties of the transformer, in particular, about physical interaction between the sides. At that a number of inner degrees of freedom in this transformer as well as the reactions at the other sides do not play any role. Proper resistances of the sides and resistances of the interaction provide information about inner physical characteristics of the transformer reduced to equivalent values at the sides. Hence, it is possible to have two transformers having equal equivalent resistances of interaction for all the sides. However, according to their geometrical and constructional characteristics and types of energy, the transformers will be different. The resistances may be a function of kinematical characteristics **(a nonlinear transformer)** or of time **(a parametrical transformer)**. Moreover, it can have constant values **(a linear transformer)**.

It is appropriate to consider physical properties of matter included in the volume of a transformer only due to resistances at the sides of the transformer.

Structure of the transformer

Every inner degree of freedom can be represented as a series circuit. By analogy with electric circuits [6, 7, 11, 16-18] the transformer can consist of various circuits. An unlimited number of degrees of freedom allows represent a transformer consisting of an unlimited number of chains. The circuits have series, parallel, or mixed junctions. A part of a circuit whose elements have the same generalized displacement is called as a branch. The branch can consist of one or several elements. A place where three or more branches are joined is called as a multiple junction. A circuit is considered to be a closed path including several branches and multiple junctions. All the elements of a chain connected in series have equal generalized displacement (as well as equal speed and acceleration). The generalized force acting on the whole circuit is equal to the sum of forces acting on the elements of the series circuit.

The elements connected in series can be replaced by one element in such a way that the generalized displacement and the summary generalized force would not change. In the case of a parallel connection all the branches of the circuit are joined to the same pare of multiple junctions and are under influence of the same generalized force. Parallel circuits can be replaced by series equivalent circuit and vice versa [6, 7, 11, 17, 18]. Hence, a number of inner degrees of freedom and a quantity of chains may be decreased as well as increased. In this case resistances occurring on all the sides will be constant.

Structure of the linear transformer

The simplest transformer is a one-side transformer having one inner degree of freedom…

If to the input of the transformer we deliver generalized displacement (or force) which depends on time as a pure sinusoidal signal then the linear transformer will have sinusoidal reaction of the same frequency. For the sinusoidal signal with fixed frequency the linear transformer (including a transformer having an unlimited number of inner degrees of freedom) can be represented as an equivalent transformer reduced to this input as one chain. An equivalent transformer placing at this input at other fixed frequency of sinusoidal disturbance will be represented as a chain having one degree of freedom but different values of chain elements.

If the signal at the input of the transformer is of complicated time-periodical form then it can be decomposed to Fourier series. Every harmonics of the Fourier series will produce its own chain. The general complicated signal will produce the complex chain consisting of compound sum of elementary series chains. This complicated chain consists of elementary chains interconnected in series or in parallel and combined in multiple junctions and circuits. The construction principles of this complicated chain can be based on methods of identification and synthesis which are well developed in electrical engineering, automation for linear and non-linear transformers [9, 12, 13]. These principles are formally useful for transformers which transform energy of any physical nature. Identification produces so much elementary chains as the number of inner degrees of freedom. Every elementary chain reflects motion by one inner degree of freedom. Physical analogy occurring among motions of different nature is actual here [11, 16, 18]. Values of the generalized masses, elasticities, dissipative and active elements can depend on kinematical variables (non-linear chains), on time (parametrical chains), or they can be constant (linear chains).

If a series circuit includes mass and elasticity then it will have resonant frequency. At this frequency reaction of mass and elasticity will be absent in the summary reaction. It is equal to the fact that we will know nothing about the value of mass (and elasticity) and, moreover, about its presence at all. The mass can have huge value, but it will be absent for us. Hence, if we assume a transformer as an atom then the atom may include elementary chains having huge masses (much more than the mass of the Earth), but we will not know about it until we apply such frequency to the input which explicates this mass.

Any energy transformer including a transformer with distributed parameters can be represented as a system of elementary series and parallel chains connected in a certain way. A concrete transformer can have various types of these connections. All of them can be included into another combination of connections that is accompanied by change of values of masses, elasticities, dissipative and active elements. However, in this case equivalency of reaction should be fulfilled at a certain diapason of frequencies on all the sides of the transformer.

There is a special spectrum of resonant frequencies for every combination of connections of the transformer. There will be an unlimited number of degrees of freedom and, hence, an unlimited number of resonances for the transformer with distributed parameters (for example, an elastic nail, a compressible fluid having a limited size). Our influence on the transformer usually has a certain diapason of frequencies. As the result of a limited diapason of excitation frequency the transformer will represent a system having a limited number of inner freedom, even if the transformer has distributed parameters.

Values of equivalent elements depend on oscillation frequency. Moreover, according to known laws, several parallel circuits can be transformed into a series one and vice versa. Taking it into account the following conclusions can be made:

1. The generalized mass, elasticity and dissipative elements do not have constant values in nature. Their value depends on frequency of the process, i.e. they depend on the character of time-change of the disturbance acting on the transformer.
2. The generalized mass, elasticity and dissipative elements are products of time-space change of energy.

Types of energy transformers

Transformers can be nonlinear, parametrical and linear. The transformer can be **active** (having an internal source) and **passive** (having an external source). The passive transformer can never be an energy source (by definition). A transformer can accept or transform one type of energy (monoenergetic transformers) or several types of energy (polyenergetic transformers). The polyenergetic transformer includes a mechanism and corresponding elements and chains which allow transform one type of energy into another one. Fundamentally, all the transformers are polyenergetic transformers of energy, however many of them can be represented as monoenergetic ones since other types of energy participate weakly in them. Designing a transformer it is more convenient to put the energy source out the transformer. Let us further classify the passive transformers.

The passive transformers can belong to the following types:

1. Simple transformers

In this case energy is produced at the output due to transformation of energy coming to the inputs into energy of another type or having other characteristics. The following transformation can serve as the example.

An electric transformer: resistance of one value comes to the input, and the resistance of another value comes from the output (a type of energy is the same).

A furnace for combustion of fuel: energy of chemical connections comes to the input and heating energy comes from the output (energy of radiation may be neglected).

A Wind Generator, Hydroelectric Power Station: kinetic energy of moving air or water comes to the input, and electric energy comes from the output.

2. Energy intensifiers

In this case a transformer intensifies energy coming to one of the inputs due to energy coming to the other inputs (or to the other input). Below there is an example for a *monoenergetic transformer*. Energy (an electric signal having certain parameters) comes to the input of a transformer, and the output energy has the same characteristics but it is more intensive. It can occur due to electric energy coming to the other input of the transformer.

A heat pump represents an example for a *polyenergetic transformer*. Low potential heat energy of the ambient space (air, water, ground) comes to the input of the heat pump, and electric energy from a power network comes to the other input. The output heat energy has value which is equal to the sum of the coming heat energy of the ambient space and electric energy coming from the power network through the electric engine of the compressor. The output heat energy excesses energy coming from the power network by several times. The output energy is always lower than the sum of the low potential heat energy and the electric energy coming to the input. Hence, the output of a heat pump is always less than unit.

Efficiency of the heat pump consists in the fact that a consumer pays for electric power produced by power network only (use of the low potential heat of the ambient space is free). However, the output heat energy is several times more than the energy which can be produced by the electric power coming to the input.

3. Active oscillator

An active oscillator is a device of a certain type. Energy of time-constant characteristics comes to its input, and at output there is energy of time-periodical characteristics. Energy with time-periodical characteristics can come to the input as well since these characteristics are not connected with the time-periodical characteristics of the output energy. The transformer can operate in a mode of an active oscillator if it has the following features:

1. Indirect connection as a series circuit: one of the outputs is connected to an additional transformer whose output is connected to one of the inputs of the transformer. The additional transformer provides certain changes of the input signal (time delay).
2. Energy comes to the input (it can be time-constant or time-periodical).
3. The transformer must have a certain type of non-linearity.

Examples of the active oscillator

Internal combustion engine: the indirect connection is the system of combustion, the non-linear element is the system of admission and exhaust valves, and the energy source is energy of chemical connections of fuel.

Various electric generators: all of them have the indirect connection, the non-linear element, and the energy source.

The active oscillators differ from the electrical intensifiers qualitatively. The difference consists in the fact that the active oscillator can operate and receive energy from the ambient space without additional energy coming to it (from a power network, for example).

Theorists and experimenters try to create a self-supporting self-organizing system in their works on creation of alternative sources. This fundamentally correct tendency requires using theory of active oscillators. There is the developed theory in electrical engineering, automation, and other scientific fields [25]. To simplify application of these theories it is necessary to use analogy between equal processes which take place in different physical fields.

4. Rectifiers

Time-periodical energy comes to the input of the transformer. The output energy has time-constant or practically time-constant characteristics. If the transformer has a non-linear element or elements of a certain type (a diode or a one-sided valve) then it can operate in the mode of a rectifier. The examples are electrical rectifiers and pumps of a valve type.

Properties of energy transformers

General properties:

1. A transformer can transform energy of one type as well as transform one type of energy into another type.
2. The generalized kinetic, potential and dissipative energies are specific peculiarities of energy change in the space and time.
3. The generalized masses, elasticities and dissipative elements are products of energy change in space and time.
4. Values of the generalized masses, elasticities and dissipative elements (existing inside the transformer as well as reduced to the equivalent values on the sides of it) depend on a temporal character of changes of kinematical or dynamic disturbances. Values of the generalized masses and elasticities at the sides of the transformer are lower than the corresponding values inside the transformer.
5. Energy interchange between macrostructures and microstructures can occur through a transformer.
6. One area of space can "know" about another area of space through energy exchange only, i.e. through generalized dynamic and kinematical interactions. In the light of the fact that chains of different transformers can have similar physical properties (for example, an equal resonant structure) interference of these transformers seems to be possible including interference occurring at a great distance.

Properties of the passive monoenergetic transformer:

1. Transformation of characteristics of the same energy type.
2. The output energy can be both equal or lower than the sum of energies coming to all the inputs. The sum of the output energies decreases due to dissipative losses occurring inside the transformer.

Properties of the passive polyenergetic transformer:

1. One energy type coming to an input can be transformed into another energy type coming from the output.
2. The sum of energies of all the outputs can be equal to the sum of energies of all the inputs or less than it. The sum of the output energies decreases due to dissipative losses occurring inside the transformer. The output will always be lower than unit. If the energy coming from certain of the outputs (within the limits of one output) is considered as a positive effect then the output will be less than the mentioned value.
3. Energy of one and the same type comes to the input and from the output while other energy types can come from other outputs. This property can allow increase one type of energy coming from an output at the expense of the other types of energy coming to the inputs.

Analysis

This article is aimed to provide general properties of the energy transformers and to analyze some problems of creation of renewed energy sources, according to these properties. There are made various attempts to design a renewed energy source which would produce energy without fuel consumption. In the light of the above mentioned results, the researches on creation of the renewed energy source may be represented by the following way. Energy can be produced from the macrostructures and the microstructures.

In the case of the macrostructure, energy can be produced by a simple transformer, an electric intensifier or an active oscillator. Methods of production of energy from the ambient space by means of simple transformers (thermoelectric and hydroelectric power stations, wind generators, etc.) are developed sufficiently by the humankind. Works on producing energy from the ambient space by means of energy intensifiers are on its initial stage. Heat pumps, which utilize low potential heat energy by means of Freon-compressor heat pumps, are successfully introduced. One of the perspective approaches is replacing such a heat pump by a vortex oscillating heat pump which does not contain Freon and a compressor [3, 5, 19, 22]. In future it will be possible to replace the energetic pump by a new energy transformer of the type of an active oscillator. In this case the energy source will always obtain energy from the ambient space without using additional energy from a power network. Energy sources of the type of the active oscillator are internal and external combustion engines, stream engines, etc.

In the general case a transformer can contain different energy types and has communications with **outer microstructures**. It should be noted that the outer microstructures can be included in the geometrical size of the transformer. Energy of this microstructure can be huge; hence, this energy can be obtained from them for a very long period of time. Atomic reactors can serve as an example for this fact. The types of energy in these microstructures can be represented by known types (electromagnetic energy, nuclear energy) and unknown types (whose numerous quantity can be considered). These energy types exist in the ambient space. However, almost all the transformers of energy produced by a human (except the nuclear reactors) do not interact with the energies, and we do not obtain energy from these microstructures.

The task of creation of a new energy source consists in designing such an energy transformer which could obtain energy from the microstructures and transform it into the type of energy which is needed to a consumer. In this case all the designed transformers will not have over unit output. Efficiency of these transformers will be defined by the cost of the energy produced by them. This energy depends on a ratio of energy coming to one of the inputs (electric power) (for example, in heat pumps) to the output energy. If the output energy is much more than the artificial energy which comes to one of the inputs then this transformer can be very useful for a consumer.

New high effective energy sources are often claimed to be created. According to the observed facts, existence of such sources is practically possible. Nevertheless, realization of this source will depend on our knowledge of physics of the microstructures and methods of transformation of energy of the microstructures into such type of energy which can be used by a consumer. Physics of microstructures is investigated badly, and many types of energy are totally unknown. Therefore, modern scientists try to create the new energy sources at random. In this case the following results can be achieved:

1. The high effectiveness of an energy source (over unit output) does not correspond to reality.
2. The output energy of the energy source is actually higher than the input energy (the source of the energy intensifier type).
3. The energy source actually produces energy without artificial input energy (the source of the active oscillator type). It can occur if an inventor failed to recognize the source of energy coming from the microstructures. It can be connected with particular or total lack of exploration of physics of the microstructures. Moreover, the fact is caused by transcendent difficulty of disclosing energy sources from the macrostructures.

Conclusions

It is principally possible to create an energy source whose output energy would be higher than its input energy (an energy intensifier) or an energy source which would produce energy without additional input energy (an active oscillator).

1. Various energy sources or transformers of energy of the macrostructures and microstructures can be designed.
2. Practical value of the energy source is defined according to cost of its output energy instead of the output value.
3. Designing new energy sources it is useful to consider them from the point of view of transformers. It is worth to define all the communications of the transformer and to state experimentally energy at all the accessible communications of the transformer.
4. Developing theories of sources it is expedient to use physical analogy and to take into account ready results of works on theories of transformers, chains, etc. of electrical engineering, radio engineering, automation and other scientific fields.
5. In the light of the fact that it is planned to obtain energy from the microstructures it is necessary to pay attention to new probable types of radiation and to their influence on a human.

References

1. Frolov A.V. The work Advanced Energy and Propulsion Systems based on Chronal Reaction Method // New Energy Technologies. - 2001, 2. p.51-55.
2. Frolov A.V. Reactionless Propulsion and Active Force // New Energy Technologies. - 2001, 2. p.56-58.
3. Sorokodum E. Recycling of energy of the atmosphere by means of vortex // Russian technologies for industry “Renewable Energy” IWRFRI’2001,may 28-30, 2001, Saint-Petersburg. Abstracts. - 2001, p. 92-93.
4. Sorokodum E. On general nature of forces. New Energy Technologies. 1(4), 2002, p.30-36.
5. Sorokodum E. D. Vortex heat-generators. New Energy Technologies Issue. Vol.2(5), 2002, p.17-76.
6. Andreev V.S. Theory of nonlinear electric circuits. M.: Radio and Communication. 1982, p.280
7. Atabekov G.I. Theory of linear electric circuits. M.: Soviet Radio. 1960, p. 712
8. Berezkin E.N. Lectures on theoretic mechanics. Part 11. Dynamics of the system, dynamics of solid body, analytical dynamics. M.: MSU - 1968, p.314
9. Blakyer O. Analysis of nonlinear systems. Transl. from English. Edited by Hohlov R.V. M.: Mir. – 1969, p. 400
10. Gledzer E.B., Dolzansky F.V., Obuhov A.M. Systems of hydrodynamic type and their application. M.: Nauka – 1981, p. 366
11. Druzinsky I.A. Mechanical circuits //L.: Mashinostroenie. – 1977, p. 240
12. Kononenko V.O., Plahtienko N.P. Methods of identification of nonlinear mechanical oscillating systems
13. Zakrzhevsky M.V. Oscillations of essentially-nonlinear mechanical systems. Riga, “Zinatne”. – 1980, p.190
14. Ladikov Yu.P. Stabilization of the processes in continuums. M.: Nauka. – 1978, p. 432
15. Merkin D.R. To the question on forces’ structure // Applied mathematics and mechanics. - 1975, 39, 35. p. 929-932
16. Oltson G. Dynamic analogies. M.: State Edition. Foreign Literature. - 1947, p.223
17. Popov E.P. Theory of linear systems of automatic regulation and control. M.: Nauka. – 1989, p.302
18. Puhov G.E., Vasilyev V.V., Stepanov A.E., Tokareva O.N. Electrical modeling of the tasks of structural mechanics. Edition of Academy of Science of Ukrainian SSR. - 1963
19. Rodionov B.N., Sorokodum E.D. Vortex energetic. Constructional materials, equipment, technologies of the XX century. 3 (26). 2001, p. 28, 29
20. Sorokodum E.D. Theory of transformers for nonlinear biomechanical oscillating systems. Taganrog institute of radio engineering. – Dep. in VINITI, 7.05.1980, #1948-80 Dep. – 1980, p. 38
21. Sorokodum E.D. On the application of theory of nonlinear transformers for the calculation of complex mechanical oscillating systems. Reports at the scientific conference “Problems of nonlinear oscillations of mechanical systems”. Kiev. - 1978
22. Sorokodum E.D. Vortex source of energy. International scientific and social conference. Prospects of preservation and developing of the united civilization of the planet. Culture. Ecology. Space. 26-31 of May, 2002
23. Strelkov S.P. Introduction to the theory of oscillations. M.: Nauka. – 1964, p. 440
24. Harkevich A.A. Theory of transformers. State Energetic edition, M.-L.: 1948, p. 191
25. Harkevich A.A. Auto-oscillations. / Harkevich A.A. Selected works in 3 volumes, V. 2, M.: Nauka – 1948, pp. 253-379.

Commercial Exploration of Space

Review
by Olga Leontyeva, Editor
http://www.faraday.ru

Of late years the mankind has greatly advanced in space exploration. New spaceships are created, automatic apparatus are launched to the planets of solar system, space stations are orbited. A man has passed to outer space and begun exploring the planets of solar system. More and more specialists and people of different professions are involved in the development of space industry. However to present day only a few of them has succeeded flying into space.

At the present state of affairs space flight is very expensive and a serious reason should exist to forward a man into space if he or she does not relate to the aims of professional cosmonautics. However in last years there has appeared a certain breakthrough in the solving of the problem. Nowadays space technologies gradually turn from the sphere of experimental and scientific researches to the area of practical application. The time has come for a man to realize the real space flight without being the professional cosmonaut.

What are the ways to solve the problem? No doubt that special attention should be paid to the development of new space technologies, search of new types of fuel and attraction of investments to the sphere of space tourism.

April 28, 2001 can be considered as official date of birth of space tourism, when there was launched the space ship “Soyuz TM-32” with the first space tourist aboard. Almost in a year, on April 25, 2002 the space tourist #2 South African Republic person Mark Shattlword started on his space journey.

Today many companies give the opportunity to make a real space flight for all comers who have enough money and health. It is promised the providing of the most modern space technologies and the most perfect space equipment. It can be created a furor by the “Minimum program” of the Russian company “Atlas Aerospace” which is made by the members of Yu.A. Gagarin Center of training of cosmonauts.

In recent years at the international market there appear more and more companies which deal with search and encouragement of inventors who work on creation of alternative propulsion systems. Activity of “X Prize Foundation” company can be considered as an example.

The X PRIZE Foundation

722-A Spirit of St. Louis Blvd
St. Louis, Mo. 63005
Tel: 636-519-9449, Fax: 314-533-6502
http://www.xprize.org
E-mail: press@xprize.org

The X PRIZE is a $10,000,000 prize to jumpstart the space tourism industry through competition between the most talented entrepreneurs and rocket experts in the world. The $10 Million cash prize will be awarded to the first team that:

- Privately finances, builds & launches a spaceship, able to carry three people to 100 kilometers (62.5 miles)
- Returns safely to Earth
- Repeats the launch with the same ship within 2 weeks

The X PRIZE competition follows in the footsteps of more than 100 aviation incentive prizes offered between 1905 and 1935 which created today’s multibillion dollar air transport industry.

For more than 30 years, the general public has waited for an opportunity to enjoy the space frontier on a first-hand basis. The X PRIZE Foundation is working to make space travel possible for all. The spaceships that compete for the X PRIZE are designed to carry passengers.

Since its inception in May 1996, the X PRIZE Foundation has registered more than 20 teams from seven countries to compete for the prize. The X PRIZE is fully funded through January 1, 2005, through private donations and backed by an insurance policy to guarantee that the $10 million is in place on the day that the prize is won. Additional funds are still being raised by the X PRIZE Foundation to implement the competition (judging, media, event management, etc.) and continue the Foundation’s education mission.

The X PRIZE was inspired by the early aviation prizes of the 20th Century, primarily the spectacular trans-Atlantic flight of Charles Lindbergh in The Spirit of St. Louis which captured the US $25,000 (US$) Orteig prize in 1927. Through a smaller, faster, better approach to aviation, Lindbergh and his financial supporters, The Spirit of St. Louis Organization, demonstrated that a small professional team could outperform a large, government-style effort.

The Societal Benefits of the X PRIZE include:

* Creation of a new generation of heroes
* Inspiring and educating students
* Focusing public attention and investment capital on this new business frontier
* Challenging explorers and rocket scientists around the world; and,
* Vehicles built for the X PRIZE will eventually serve four different industries:
- Space Tourism
- Low-cost satellite launching
- Same-day package delivery
- Rapid point-to-point passenger travel.

In Fig. 1 it is demonstrated the typical X Prize trajectory.

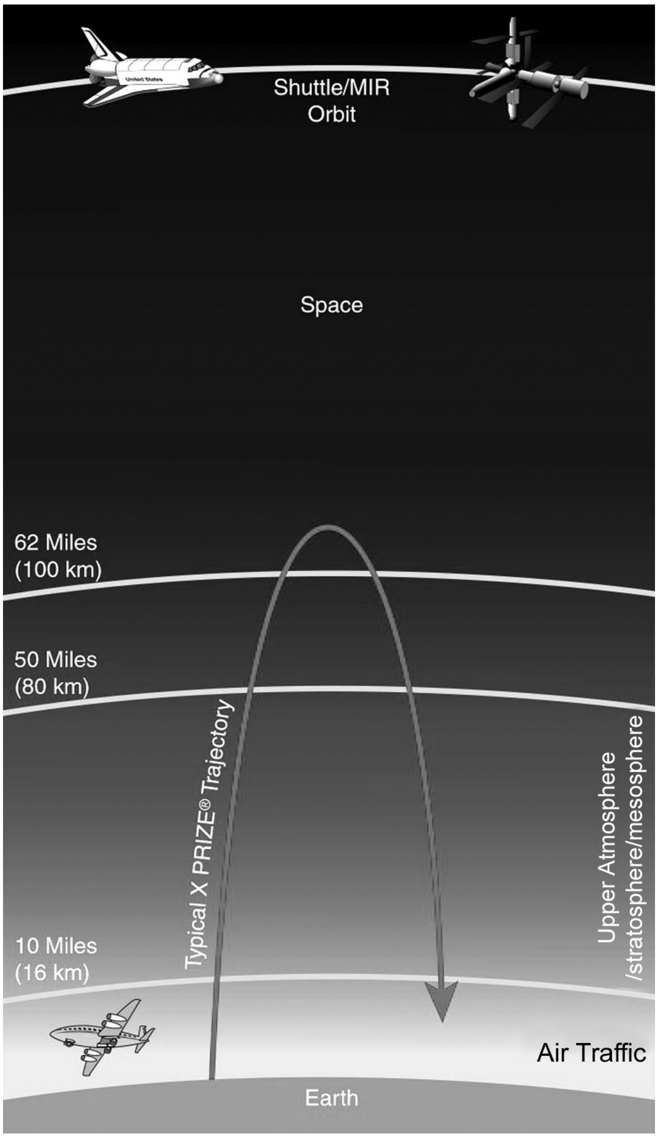

Fig. 1

Editor: Below we publish photos and brief comments about some official X PRIZE registrant teams. More detailed information you can find at http://www.xprize.org/imagefacts/photo1.html.

The Da Vinci Progect

http://www.davinciproject.com

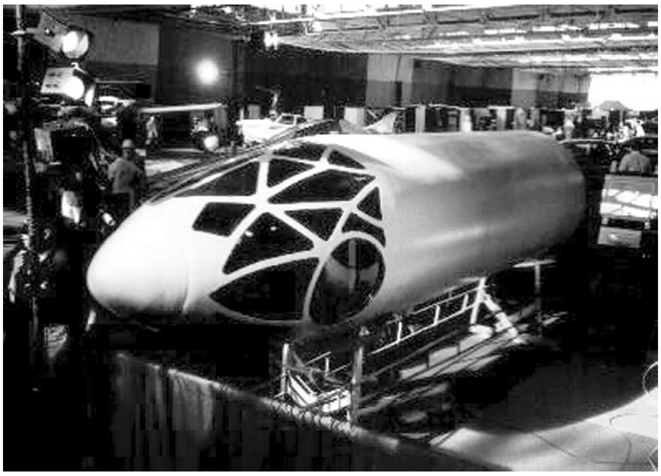

Fig. 2
"Wild fire"

The **da Vinci Project** will launch its spacecraft ("Wild Fire") from the world's largest helium balloon. The 3,270 kg (7,200 pound) rocket will be tethered 720 meters (2,400 feet) below the balloon and lifted over the course of an hour to an altitude of 80,000 feet. The 10,000 pound thrust, liquid oxygen, kerosene engines will fire the first stage and the rocket will fly an initial angular trajectory to clear the balloon. The spacecraft then will transition to vertical flight to its apogee of 120 km in space. The rocket will reach a maximum speed on both its ascent and re-entry of Mach 4, or 4,250 kph (2,650 mph).

An innovative ballute will protect and stabilize the rocket on re-entry. A flyable parachute will be deployed at 25,000 feet and the rocket will descend under control, guided by GPS, to a predetermined landing zone. The da Vinci Project has already successfully conducted full-scale rocket motor test and has built a full-scale mockup of their vehicle.

"Discraft" corporation

http://www.xprize.org/teams/teams.html

John Bloomer, the team leader, is an aerospace engineer and he has worked on many aerospace progects, including Apollo and holds more than 60 patents on a disc plattorm aircraft. Bloomer's ship utilizes "Blastwave" Pulsejets.

Flight Sequence

Fixed, 7850-ft^2-area, laminar-flow wing take-off at about 60 mph within about 150 ft, featuring climb with gradual air-breathing acceleration (according to a fixed program) at fixed angles, to exit the atmosphere at Mach 10 on an unpowered ballistic arc to reach 75-mi. altitude: return on down-leg of same unpowered arc to gradual power-on flare-out re-entry of the atmosphere in simple reverse sequence of the take-off velocity profile. Range above 100,000 ft is about 480 mi. which is covered in about 5 minutes.

"Kelly Space & Technology"

http://www.kellyspace.com

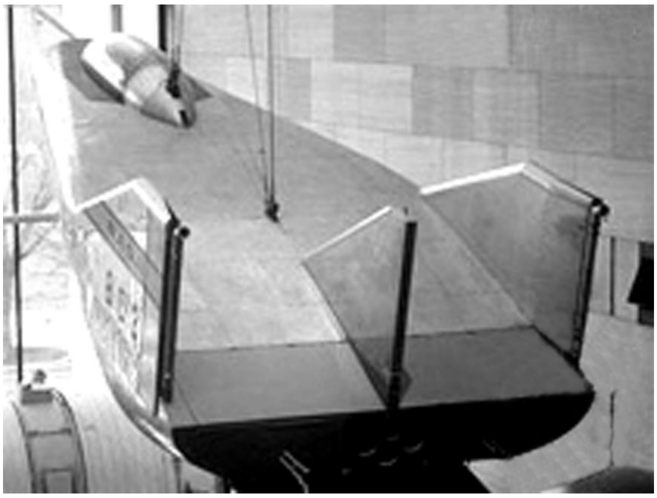

Fig. 3
"LB-X"

The vehicle is a rocket-powered delta wing glider with a liquid oxygen and kerosene liquid rocket engine. The spacecraft is prepared and fueled at the takeoff airport. The spacecraft is towed to release altitude behind a conventional jet powered aircraft such as Boeing 747 aircraft. Upon release from the tow aircraft, the main engine is throttled up for boost phase. The vehicle nominally coasts to an apogee altitude of 100 kilometers. The vehicle then glides to a landing at the takeoff airport for checkout and refueling prior to the next flight.

Flight Sequence

The lifting body will be towed to launch altitude behind another aircraft, and the rockets will be ignited. The craft will return to the landing site and make an unpowered, horizontal landing.

"Lone Star Space Access"

http://www.dynamicar.com

The Cosmos Mariner employs air-breathing jet propulsion for take-off and landing from conventional airports and rocket propulsion for ascent from cruise flight in the stratosphere to 30 or 40 nautical miles altitude. From there, the vehicle coasts to a target altitude of around 65 nautical miles. The air frame is designed to interface with two jet engines (turbofan or turbojet) each with 20,000-lbs static thrust or less. For rocket propulsion, the Cosmos Mariner will use the Aerojet AJ26-NK31A, a staged-combustion kerosene engine. The vehicle is designed to take off and land from conventional runways.

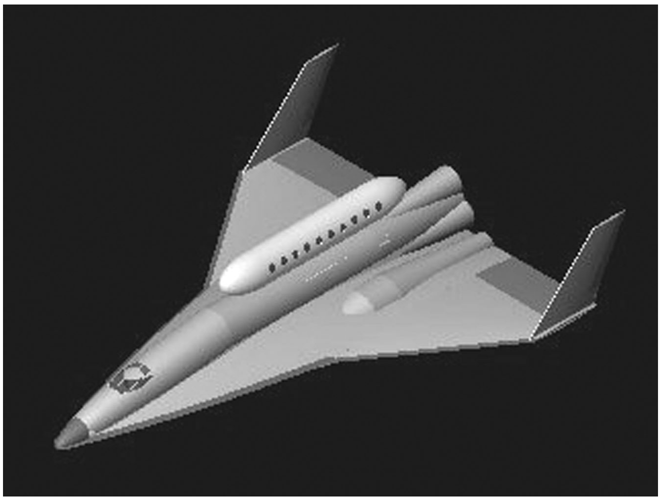

Fig. 4
"Cosmos Mariner"

Along with X Prize Foundation it should be mentioned ***ALLTRA*** *Company (Germany). ALLTRA consists of a small group of space experts. The main objective of ALLTRA is to 'sell' the space idea to a broad public and to identify future commercial opportunities in the space sector.*

On the official ALLTRA website http://www.alltra.de you can find collection of artist's views of the projects which are aimed at decrease in space ships value by means of using of new types of engines. Besides the Company deals with the development of space hotel projects.

JNET

New Energy Technologies
collection of articles
Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

Alternative Aircrafts and Space Propulsion Systems

Editor: Nowadays all over the world there are made numerous attempts to create new types of aircrafts and alternative space propulsion systems. Below we publish the review of some interesting devices which are presented in the modern market.

"Arbortech Pty.Ltd"

http://www.airboard.com.au

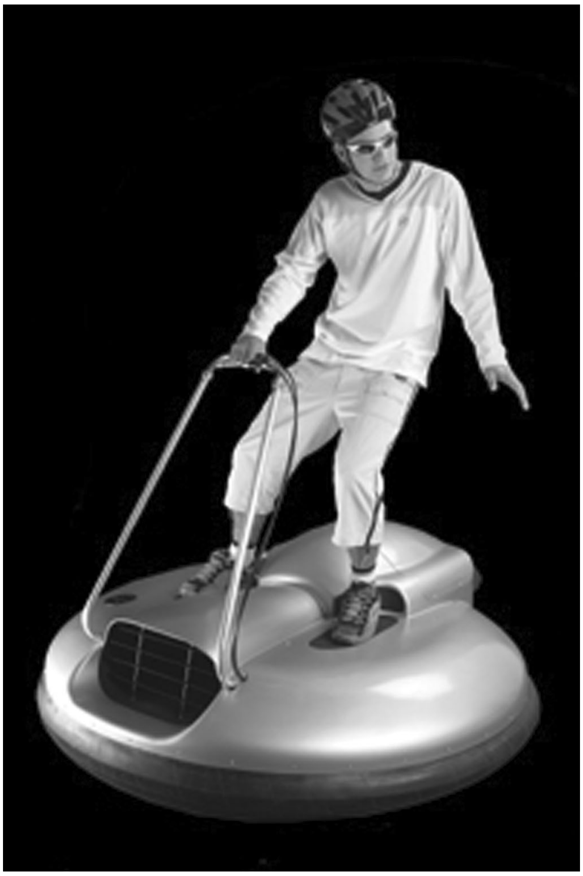

Fig. 1
Airboard 2000

General Technical Specifications

Total payload, including rider — 100 kg (220 lb)
Operating time — 1 hour on full tank of fuel
Construction — Fiber glass/High-impact plastic shell, Aluminum frame, Rubber skirt
Starting — Electric key-start, Battery included
Engine — Briggs & Stratton 4-stroke
Fuel tank capacity — 5 litre (1.3 US gal)
Fuel type — 85 Octane unleaded

How does it work?

Airboard uses Hovercraft air cushion principles to glide just above the ground. The air cushion is generated from a purpose designed engine and fan which are suspended below the Airboard shell. In addition to providing an air cushion for the Airboard to ride on the fan also provides a fast stream of air exiting from the rear of the craft to provide forward thrust – just like a hovercraft. To help provide better acceleration the Airboard also includes a unique friction drive wheel at the rear. When the rider wishes to accelerate forward, weight transfer is used to bring the friction drive into contact with the ground. By engaging the friction drive clutch the Airboard can be accelerated forward without losing the feel of hovering above the ground. The control of the Airboard is provided by weight transfer of the rider, similar to surfboards, skate boards and snow boards. In operation the Airboard can be started, stopped and steered in a controlled manner and this means that riders can perform stunts, trick maneuvers and race around tight tracks against each other. In this unique way the Airboard is the first ever vehicle to deliver the advantages of hovercraft vehicle without the disadvantages of proof acceleration and handling.

"Trek Aerospace, Inc."

http://www.solotrek.com

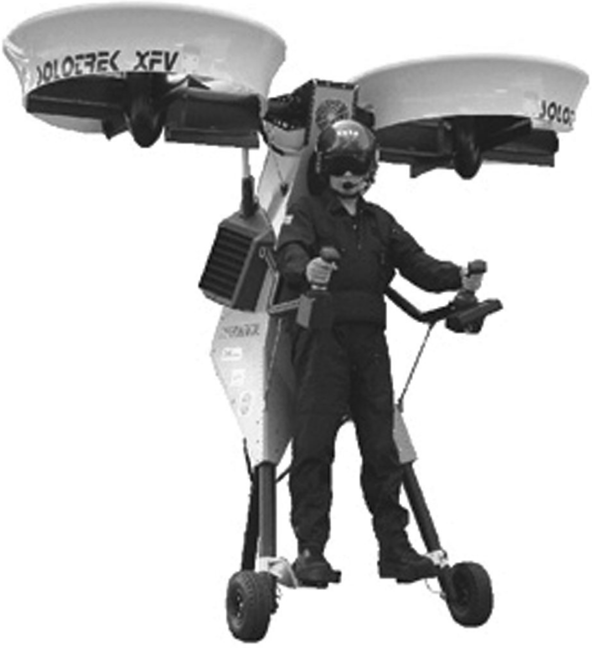

Fig. 2

Trek Aerospace

Preliminary Specifications and Predicted Performance

Normal Gross Take Off Weight	800 Lbs.
Fuel (15 U.S. Gallons)	98 Lbs.
Mission Payload, net of fuel	352 Lbs.
Takeoff/Landing Distance	0 (VTOL)
Maximum Speed	70 Mph
Range	120+ Miles
Hover/Loiter Endurance	2+ Hours
Engine Type	Advanced Int. Combustion
Fuel Requirements	Heavy-Fuel or Gasoline

Special Features: 100% fly-by-wire control system with electronic stability augmentation; Intuitive, easy to fly safely; Minimal field service requirements; Rugged, efficient power train.

DM AeroSafe

http://dmaerosafe.freeservers.com

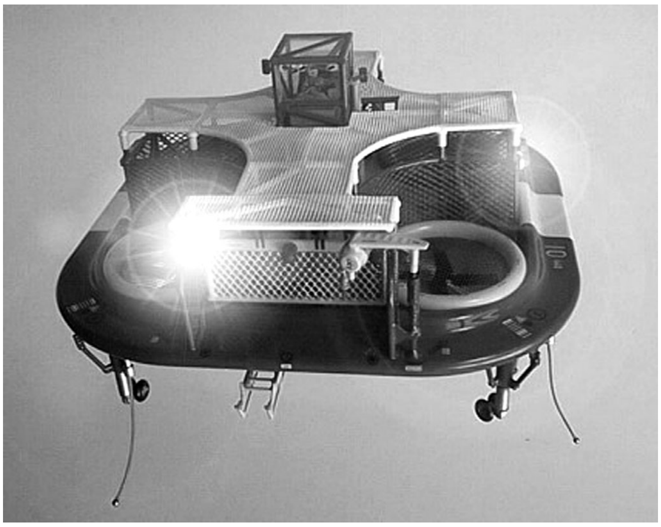

Fig. 3

EAGLE vertical take-off and landing
aerial rescue platform

DM AeroSafe is a small research and development team, which has developed a totally new high-rise rescue technology to retrieve trapped people from areas which cannot be reached by conventional aerial ladder, conventional helicopter or a helicopter equipped with a Heli-Basket.

This technology could give rise to a new class of air transportation means, used for safe close-in maneuvering around tall structures, even inside the highly populated areas.

Eagle Aerial Rescue Platform's Performance Summary

Dimensions:

Length - 42 feet (12.8 m)
Width - 42 feet (12.8 m)
Height - 16 feet (4.8 m)

Weights:

Max. Gross weight - 5000 pounds (2268 kg)
Payload (Crew of two plus 10 rescued people) - about 2000 pounds (about 900 kg).

Propulsion System:

Four (4) variable collective pitch Ducted Propellers
Propeller Diameter - 7.8 feet (2.4 m)
Engines - four (4) four-cylinder radial piston aero engines with forced air cooling system 250 hp each (1000 hp total)

Performance:

Max. Duration - about 5 hours without refueling
Max. Airspeed - 40 knots (75 km/h)
Max. operating altitude - about 7000 feet (about 2200 m).

"LTAS/CAMBOT, Inc."

http://www.lvcm.com

Passenger Craft "Ltas 30-Xb"

The ships have rigid monocoque hulls, hybrid powered vectored thrust control and full active buoyancy control. (LTAS US Patent Pending).

This small 2-3 person craft at 70 to 80 feet in diameter will demonstrate ALL production systems and is designed for the LTAS FAA Type Certification program.

NASA's Langley Research Center

http://science.nasa.gov

NASA researchers are studying insects and birds, and using "smart" materials with uncanny properties to develop new and mind-boggling aircraft designs.

The "personal aircraft" that replaces the beloved automobile in people's garages may still lie in the realm of science fiction or Saturday-morning cartoons, but researchers at NASA's Langley Research Center (LaRC) are developing exotic technologies that could bring a personal "air-car" closer to reality.

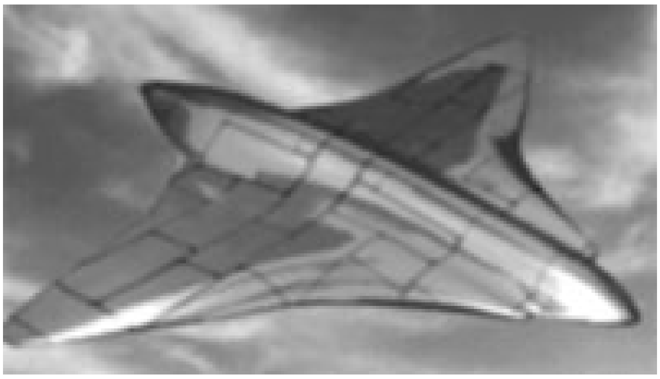

Fig. 4

And air-cars are just the beginning! Self-healing wings that flex and react like living organisms, versatile bombers that double as agile jet fighters, and swarms of tiny unmanned aircraft are just a few of the science-fiction-like possibilities that these next-generation technologies could make feasible in the decades ahead.

Laser Beam Flight

Lightcraft Technologies, Inc. (LTI)

http://www.lightcrafttechnologies.com

Lightcraft Technologies, Inc. is a new company committed to providing low-cost access to space through the use of beamed energy propulsion. (See color photos on the cover page).

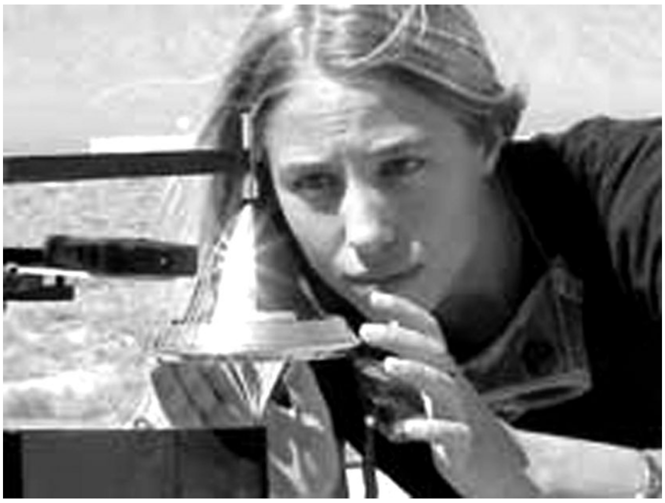

Fig. 1

In 2000 at the High Energy Laser Systems Test Facility (HELSTF), Lightcraft Technologies, Inc. (LTI) set a new world's altitude record of 233 feet (71 meters) for its 4.8 inch (12.2 cm) diameter laser boosted rocket - in a flight lasting 12.7 seconds.

Although much of the flight was spent hovering at 230+ feet, the Lightcraft sustained no damage and will fly again. Besides setting the new altitude record, the craft demonstrated the longest ever laser-powered free flight and the greatest "air time" (i.e., launch-to-landing/recovery). LTI launched a total of seven vertical flights between 8:30 am and 11:30 am with three Lightcraft weighing less than 1.8 ounces (51 grams). Two of the flights by Lightcraft #3 reached 159 and 184 feet with the same propellant load!

...set a new world's altitude record of 233 feet!

The record flights were powered by the 10 kW pulsed carbon dioxide laser named "PLVTS" by the organization that owns it: the Directorate for Applied Technology, Test and Simulation (DATTS). Even though PLVTS was suffering from an arcing or grounding problem that caused it to run erratically, the laser power was still adequate to propel the craft to record altitudes.

What is a Lightcraft?

A Lightcraft is a 1kg launch vehicle, made from high temperature ceramic materials, that flies into space on a megawatt laser beam.

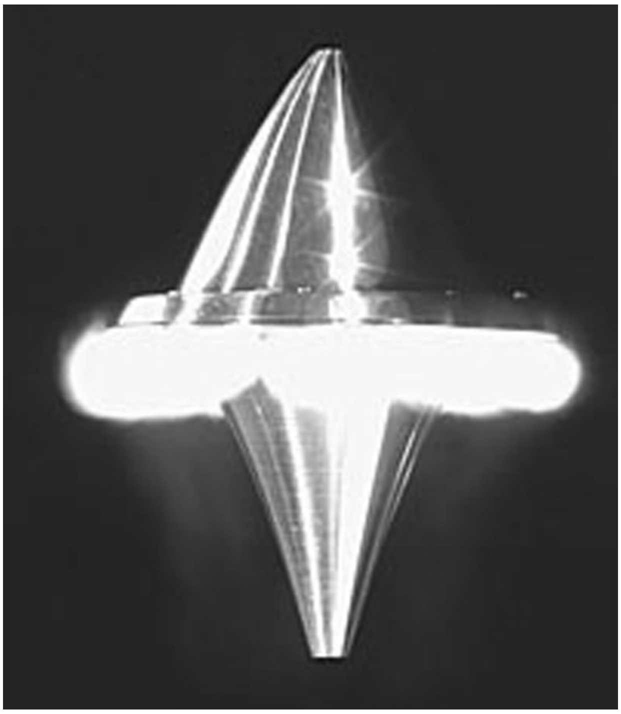

Fig. 2

The Lightcraft is both a single-stage-to-orbit launch vehicle and a satellite. If you have any further questions or comments, write or call LTI headquarters in Bennington,VT. The company representatives would be delighted to send you more information, or answer any inquiries over the phone.

How does it work?

A ground based laser is the power source that propels the Lightcraft into orbit. Lightcraft can deliver payloads into space for a fraction of the cost of traditional rockets because most of the engine stays on the ground, thereby unburdening the craft from having to lift the energy source for its propulsion system.

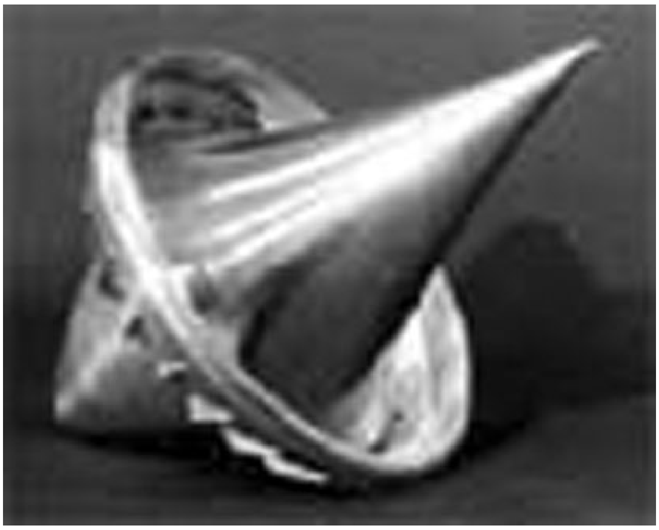

Fig. 3

The back side of the craft is a large, highly polished parabolic mirror that is designed to capture the laser beam projected at it from the ground. The mirror focuses the beam, rapidly heating the air to 5 TIMES the temperature of the sun, creating a blast wave out the back that pushes the vehicle upward. As the beam is rapidly pulsed, the vehicle is continuously propelled forward, on its way to orbit.

History of Lightcraft

1987 — Prof. Leik Myrabo invents Lightcraft for SDIO.

1997 — First successful wire-guided tests at WSMR. Solved flight stability difficulties, much like the Wright brothers did with the airplane.

1997 — Lightcraft broke Goddard's 41 ft., 1926 first successful rocket flight - but this time with no on-board fuel.

1998 — Record flight of 99 ft. with air breathing Lightcraft engine.

1999 — Record flight of 128 ft. with first rocket Lightcraft engine.

2000 — LTI sets new world record for highest flight (233 ft), longest flight time, and heaviest vehicle.

How can LTI reduce launch costs?

Conventional Launch — $175,000,000
The Lightcraft — $46,000

Chemical Rockets:

- carry massive propulsion source on board
- are expendable
- extremely costly
- prone to explosion due to fuel on board

Laser Propulsion:

- propulsion energy source remains on the ground!
- Lightcraft are inexpensive to manufacture and extremely light weight
- highly reusable power source is never subjected to the risks of flight

Lightcraft Technologies, Inc.

1914 Walloomsac Rd.
Bennington, VT 05201
Office: 802-447-6275
FAX: 802-447-8216
E-mail: tmyrabo@lightcrafttechnologies.com

Alternative space propulsion systems Star Drive

Mark R. Tomion, USA
http://www.stardrivedevice.com

Re: U.S. Patent **6,404,089** for the **Electrodynamic Field Generator**, **(EDF),**
issued June 11, 2002 to Mark R. Tomion.

The 'official' name of the "StarDrive device", per the U.S. Patent and international PCT Applications, is **Electrodynamic Field Generator.** The EDF Generator uses banks of permanent magnets and rotating Field Coils to produce a very-high DC rotor voltage, and plane-parallel ring electrode arrays to electrostatically expand and control that voltage *as applied to the hull*, so that huge quantities of external Field electrons may be accelerated to energy levels that are usually reached only with a particle accelerator! It's somewhat like a glorified arc welder whose output is deliberately shorted to its own housing, and the DC voltage and current across the emitter and collector housing sections can be thermionically increased to values that are generally observed only in lightning: but the Field's *current density* is limited to a value which falls short of damaging the hull!

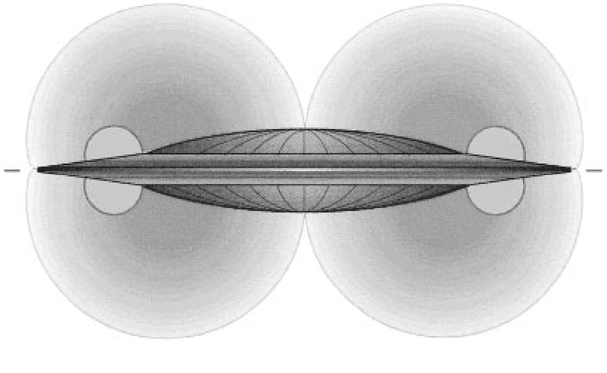

Fig. 1
StarDrive device hull & electrodynamic
Field configuration

As depicted above (Fig. 1), electrons circulating in the external Field envelope can achieve an impact velocity at the central collector sections which is very nearly that of light, and applied magnetic fields allow broad modulation of the Drive Field current's properties. The *Primary Arrays* shown in the generalized schematic diagram below (Fig. 2) have control grids which allow an arc resistance imbalance to be imparted to the otherwise symmetrical Field current, so that they *render the two relativistic current impulses variably non-isometric:* thereby yielding thrust that is essentially **reactionless**! And the simple DC *Primary Power System*, like the early Faraday disk dynamo, *is wholly rotor-based* **. . .**

Note: It can be seen that the propulsive thrust developed by a StarDrive vessel is essentially brute-force in nature – it's produced simply by means of a controlled variable imbalance in the continuous *physical impact* of the two external hemitoroidal electron current streams with the collectors! If these two Field currents were of equal magnitude, no net force would be developed. However, if the "lower" current stream is stronger than the "upper", the vessel will be propelled away from the stronger current – in the "upward" direction. Since there's no 'backward' exhaust produced in the process, this type of thrust is truly and demonstrably reactionless in nature.

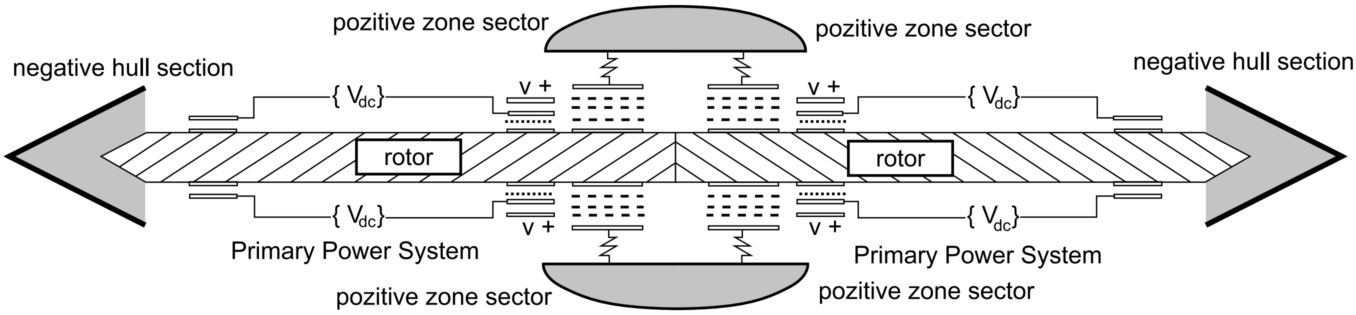

Fig. 2
Generalized schematic of rotor & dual induction ring assemblies

Needless to say, tremendous quantities of heat are produced in the StarDrive device's electron "targets" or collector housing sections, and liquid sodium must be pumped through each Primary Array's ceramic resistor network as a coolant. However, this excess heat in ground-based units may be used in the *commercial generation of electric power and desalinization of seawater*.

In fact, because an intense arc discharge field has the unique capacity to absorb vast quantities of quantum background energy, the EDF Generator is so efficient that the latter task may become truly cost-effective for the first time! And not only will large ***over-unity*** StarDrive Dynamo units be able to produce electric power at 60 to 720 MW output levels, they'll be able to do so for many years before the permanent magnet banks must be remagnetized!! *The only truly external input energy required* in the interim *is that necessary to initially bring the rotor up to speed . . .*

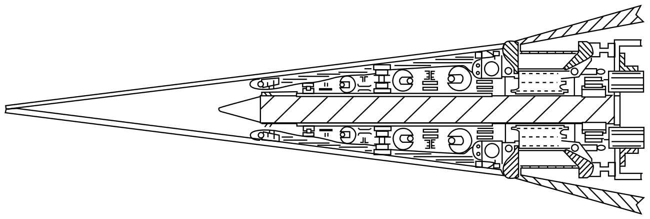

Fig. 3
A denumbered version of Fig. 1
from the EDF Generator Patent

Those of you who find this advanced technological prospect for the 21st century as exciting and fascinating as we do should consider making the inventor's book *StarDrive Engineering* a valuable addition to your personal or reference library.

A Layman's Description of the StarDrive Device

Nearly anyone who endeavors to gain an initial understanding of the *Electrodynamic Field Generator*, even a scientist or engineer, is likely to ask for a concise layman's explanation of "how it works". This isn't easily accomplished, for this device represents an unorthodox and multidisciplinary technology. However, the overview provided below will discuss the operative characteristics of the machine's simplified *electric power generation variant* in what is hoped to be the most direct manner possible. In this case, when people ask the question above, what they generally mean is: "By what means is *over-unity* operation achieved in a 'StarDrive' Dynamo?". It is assumed here that the reader has not only already become acquainted with certain basic aspects of the design, but also understands that over-unity operation is an absolute prerequisite for any viable system of light-speed interstellar propulsion.

By way of further background, the formative mechanical design basis of the EDF Generator is of course the original Faraday disk dynamo. For whatever reason, no one seems to have bothered trying to develop this simple machine into a more sophisticated *and patentable* form before now. The principal limitation of Faraday's disk dynamo in its original form was that, when one or more permanent magnets were used to pass flux directly through the plane of a solid conductive rotor, a rather high-loss heavy current at very low voltage was produced. The first design improvement was therefore to use rotor-mounted toroid field coils in order to generate a much higher-voltage primary current, and to segment the rotor to reduce "eddy current" losses. Also, the toroid coil configuration absolutely minimizes Lenz losses, or the magnetic "drag" that is experienced by any conventional generator's rotor.

To eliminate the use of brushes, it was necessary to apply traditional vacuum tube design and operating principles. Fortunately, most of the original patent work in this field has passed into the public domain, so incorporating certain aspects of that work into the EDF Generator presented no impediment to its patentability. By using plane-parallel electrode arrays instead of brushes to charge the rotor, it is possible to limit that portion of the induced rotor current which passes through the field coils to a very low level – effectively isolating them from the actual output circuit. More importantly, however, it then becomes possible to thermoelectrically charge the Generator's housing itself in such a way that *it* carries the device's full output circuit current instead!

The reason for this unorthodox design parameter is that it was desired to actually incorporate a standing electric arc field into the output circuit, to take advantage of the electron's inherent ability to absorb quantum background energy [including *zero point energy*, if and as necessary]. As a result of a thorough study of lightning, the inventor of the StarDrive device reasoned that the electrons comprising any naturally-occurring bolt of lightning had to *recover* (or absorb) an amount of ambient photonic energy equal to that which they expended in transit – in order to satisfy the conservation of energy principle. In such a case, it is not necessary that "we" do the *work* of moving charge *against* a potential gradient; the work may be done by the charge itself in being attracted *along* the potential gradient (or voltage level).

And since it is known from the field of welding that it takes less energy to sustain an arc than it does to initiate it, it therefore becomes possible to create an electrical circuit that outputs more energy than it requires as input. This is exactly what the EDF Generator does – by incorporating a standing arc field in its output circuit. Using the StarDrive device as our mechanism, "we" only provide the work-energy required to establish and maintain the external field's potential gradient, by initiating rotor rotation and bringing the thermoelectric elements up to temperature. The electrons in the electrodynamic field do all the rest . . .

Further Notes from the Inventor

In conjunction with one of our *StarDrive Generator* prototype project funding proposals, we are presently working on an interim proof-of-concept experiment for our proposed 24 kW air-cooled EDF Generator prototype. This full-scale mock-up of the 30"-dia. Generator's rotor and dual induction ring assemblies is intended to demonstrate the fundamental design principal discussed at the close of the Technical Overview (linked to our website's Method of Operation Summary page), whereby the voltage electrostatically induced on the rotor anode rings which power each Primary Array should be roughly one-third (1/3) of the Field Coil voltage (because of the capacitive dual induction ring geometry of the Primary Power System). For safety reasons, the Field voltage in all *air-cooled* StarDrive *Generators* will be limited by design to 850 VDC, and to 1,400 VDC in the larger *liquid-cooled* StarDrive *Dynamos*.

The experiment will also assist greatly in the derivation of remaining production model specifications, and will in fact incorporate production-quality rotor segments and electrode rings. Should this proof-of-concept experiment be successful, not only will the ability of our over-unity 24 kW Generator prototype to deliver large-scale DC output that's compatible with standard AC inverters (for utility grid distribution or off-grid conventional use) be virtually assured, but a major milestone incentive in our existing funding proposals will have been fulfilled as well. Further updates like those below will be forthcoming on the *News* page of http://www.stardrivedevice.com.

01/28/03 - International Patent Filings Secured!: We are very pleased to report that we were successful in our efforts to secure numerous international (PCT) Patent Application filings for the *Electrodynamic Field Generator* by the final deadline of January 21, 2003! These important filings were effected in Australia, Canada, the People's Republic of China, the European Union (including France, Germany, Italy, Spain, and the United Kingdom), India, Japan, Mexico, the Russian Federation, and South Africa.

The securing of these key Patent Application filings adds immeasurable value to our StarDrive Engineering Project overall, as it will greatly protect and enhance our investors' upside global market potential (assuming, of course, that our *24 kW StarDrive Generator prototype* is successful)! Be sure to check back from time to time for further updates as we proceed into the development phase of the Project.

11/07/02 - Academic Reference: Those of you kind visitors to our website who would like to have the benefit of an informed and unbiased academic opinion regarding the Electrodynamic Field Generator are welcome to contact John J. Tulip, Ph.D., Exec. Vice-Pres. of *American International University*. This consideration also applies of course to those parties who may be interested in securing a direct participation in our forthcoming EDF Generator Prototype Project. Dr. Tulip has not only expressed much-appreciated support of our efforts to introduce this important new technology, but has also had the opportunity to review our technical manual *StarDrive Engineering*. You may contact Dr. Tulip via e-mail at TulipJJ@aiuniversity.edu although we ask that you expect him to field serious and respectful inquiries only. Should you desire to speak with him by telephone, please be assured that he will endeavor to return your call whenever circumstances permit if you provide him with the proper phone number.

10/07/02 - Joint Venture Agreement Announced!: Mark Tomion, founder and president of Archer Enterprises and inventor of the recently-patented *Electrodynamic Field Generator*, is very pleased to announce that he has signed a Joint Venture Agreement with Affirm Technology Partners of Carlsbad, California to build a working prototype of his over-unity 'StarDrive' device's electric power output variant. The co-developers are planning to commence construction of a small air-cooled **StarDrive Generator** unit with a projected output rating of 24 kW and a housing diameter of only 30 inches, at a total weight of under 50 lbs., before the end of November 2002.

Should this exciting project be successful, it would represent an historic milestone in the development of over-unity electric power generation technology. Not only is this small prototype EDF Generator expected to demonstrate *a minimum Coefficient of Performance in excess of 20:1*, but the design employed is completely linearly-scalable in a very broad range of sizes that would include liquid-cooled ***StarDrive Dynamo*** units with output ratings of up to 1 gigawatt! The most remarkable feature of these large Dynamo units is that an amount of recoverable thermal energy comparable to their respective electrical outputs will be made available for desalinating seawater, or for use in centralized municipal and industrial hydronic heating systems. And this capability would make the large-scale desalinization or distillation of water truly cost-effective for perhaps the first time ever.

Interested parties are welcome to contact Mr. Tomion at office@stardrivedevice.com (585-526-6817) for further information.

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

INFINITE ENERGY
The "Lifter" Phenomenon

Infinite Energy Magazine

* Cold Fusion * New Energy
* New Science
* New Technology
Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

Institute for Planetary Synthesis

P.O. Box 128, CH-1211 Geneva 20, Switzerland

Tel. 41-022-733.88.76,
Fax 41-022-733.66.49
E-mail: ipsbox@ipsgeneva.com

website: http://www.ipsgeneva.com

Prospects of Hydrogen Energetics

Ph.M. Kanarev, Russia

E-mail: kanphil@mail.kuban.ru

Hydrogen is the only pollution free and inexhaustible energy carrier. But the implementation of such attractive properties of hydrogen is restrained by large expenses of energy for its production from water. Modern level of knowledge gives the opportunity to reduce these expenses [1], [2], [3].

It is known that a gram-atom is equal numerically to atomic mass of the substance, and a gram-molecule is equal numerically to molecular mass of the substance. For example, the hydrogen gram-molecule in the water molecule is equal to 2 grams, and the oxygen gram-atom is equal to 16 grams. The gram-molecule of water is equal to 18 grams. As the mass of hydrogen in the water molecule is 2x100/18=11.11% and the mass of oxygen atom is 16x100/18=88.89%, the ratio between quantity of hydrogen and oxygen is preserved in one litre of water as well. It means that 1000 grams of one litre of water contain 111.11 grams of hydrogen and 888.89 grams of oxygen.

One litre of hydrogen has mass of 0.09 grams, one litre of molecular oxygen has mass of 1.47 grams. It means that from one litre of water it possible to produce 111.11/0.09=1234.44 litres of hydrogen and 888.89/1.47=604.69 litres of oxygen. Thus, one gram of water contains 1.23 litres of hydrogen [1].

Now energy consumption for production of 1000 litres of hydrogen from water is 4 kWh and of one litre – 4 Wh. As it is possible to produce 1.234 litres of hydrogen, then 1.234x4=4.94 Wh are spent for production of one gram of water now.

Instruments and Equipment Used for the Experiment

A special experimental low current electrolyzer, a voltmeter of the highest accuracy (accuracy class 0.2, GOST 8711-78), an ammeter of the highest class of accuracy (accuracy class 0.2, GOST 871160), a balance with value of a division of 0.10 grams and 0.010 grams, and a stopwatch with value of a division of 0.1s.

Experimental Results

Indices	Sum
1 – duration of electrolyzer operation connected to the supply line, in 6 cycles t, min	6x5=30.0
2 – readings of voltmeter V, volts	13.6
3 – ammeter readings I, amperes	0.02
4 – power consumption (P=VxIxτ/60), Wh	0.136
5 – duration of electrolyzer operation disconnected from the supply line, in 6 cycles t, min	6x55=330.0
6 – solution mass change m, grams	0.44
7 – mass of evaporated water m', grams	0.02x6=0.12
8 – mass of water converted into gases m''=m-m', gram	0.320
9 – power consumption per gram of water converted into gases P'=P/m'', Wh/grams of water	0.425
10 – existing power consumption per gram of water converted into gases P'', Wh/grams of water	4.94
11 – reduction of power consumption for hydrogen production from water, K=P''/P', times	11.62
12– quantity of released hydrogen, ΔM =0.320x1.23x0.09=0.035, grams	0.035
13 – power content of hydrogen being produced (E=0.035x142/3,6)=1.397, Wh	1.397
14 – energy efficacy of water electrolysis process (Ex100/P), %	**1027**

Note: Gas output is clearly observed during many hours after the electrolyzer is disconnected from the electricity supply.

Conclusion

Low ampere water electrolysis is a way for production of inexpensive hydrogen from water and hydrogen energetics.

References

1. Ph.M. Kanarev. The Foundation of Physchemistry of the Micro World. Krasnodar, 2002. 320 pages
2. http://book.Kanarev.innoplaza.net
3. hppt://www.n-t.org/tp/ns/if.htm.

Heater Producing Healthy Effect

The First Device of Healthy Life Support in Fuel-Energy Industry

A. Belyaeva, L. Savelyeva, O. Bondarenko, Kirghizia

http://www.leteco.h1.ru
Email: leteco@mail.ru

Nowadays problems of rational use of resources including energy resources are very important. Lack of energy-efficient technologies causes consumption of great amount of electric power, coal, and mineral oil. From another hand, in the XXI century one of the main approaches of human society development is turned to healthy life-style and development of ecologically appropriate technologies. Hence, basic direction of heating systems development should correspond to, at the least, two requirements, i.e. effectiveness (energy-saving) and ecological appropriation (ecological functioning). Moreover, it is necessary to take into account additional requirements for modern works:

- Relatively low prime cost of the systems;
- Minimal operating costs;
- Availability of a system of temperature mode control;
- Use of domestically produced materials for the systems.

The aim of the work is creation of a universal electric device of natural ecologically appropriate materials. The device should maintain relative air humidity of natural atmosphere in a quarter, make its ecology healthier, meet all the listed requirements, and represent new generation energy-saving system designed for heating of industrial and domestic areas according to features of technical solution.

A.L. Belyaeva is the author of this work. The invention of this heating device was acknowledged as the best invention of Kirghiz Republic of the last two years. A.L. Belyaeva was called as a laureate of a competition of "The Best Inventing in Kirghiz Republic in 2001-2002".

Solving this problem the author based on work experience existing in the area of semi-conductors production. Actually, the model was worked out in the common area of electric engineering and industrial crystals growing. Using of knowledge and skills of the both areas made creation of this device possible.

It should be noted that initially the invention of the heating device, which is discussed here, was connected with necessity to improve microclimate and air atmosphere in a semi-conductor shop since the industrial process required optimal conditions. Hence, the device was observed as an element of the industrial process. However, soon other positive characteristics of the device were disclosed, i.e. when it was installed in a quarter all visible mould disappeared. Laboratory investigations stated that the operating device annihilated all pathogenic microflora. This rediscovered feature allowed using the device both in domestic and industrial area. After series of following laboratory investigations and bench tests construction of the device was changed and improved that allowed prepare it for line production.

Ceramic electroconvector: general characteristics

The working name of the final version of the model is *Ceramic Electroconvector TY 2971-006-22997241-2002.* Ceramic Electroconvector is an industrial and domestic electric heater of direct stationary action. It has high effective heat emission and satisfies ecological, sanitary, medical, and fire-prevention requirements.

At the same time it should be mentioned that by several characteristics this electroconvector differs from other known models of electroconvectors as well as from other existent heaters. It may be observed as a representative of an independent, specially created group of heating devices.

One of the main characteristics of the electroconvector is presence of constructive heating carbonic elements made of ecologically appropriate natural *non-metal* materials. Essentially, the electroconvector construction does not contain metal (the only metal part of the device is its supporting construction).

Production of heat of physiologically comfort zone requires 0.3 kW/h energy consumption which is 3-10 times less than energy consumption of known models of heating engineering. The Ceramic Electroconvector influences positively on ecology of a room. Room heating occurs better and more softly as conducting carbonic elements can be heated maximally up to 100° C. As a result oxygen is not burnt, and air is not overdried in a room. Achieved minimization of metal content in the constructive elements of the device increases the level of ecological compatibility both of electroconvector constructions and operation.

The device accumulates no static electricity, neutralizes harmful magnetic field generated by alternating current in the conducting element (it is typical for all the other electrical household appliances). Therefore, **the additional positive effect** is produced and higher ecological characteristics of functioning of this electroconvector are confirmed.

Insulating strength of ceramics prevents electrical shock accidents. Ecological compatibility of the electroconvector is provided by materials of the construction. The base of the device is a studied natural silicate fiber which has quantitative and qualitative content of useful chemical elements which are the closest ones to the group of medical adsorbents listed in a medical encyclopedia. The ceramics is adjusted to emit electromagnetic waves only in infra-red spectrum.

The electroconvector produces heat waves in the average IR spectrum (8.4-8.6 mkm) which is maximally approximated to the diapason of heat waves generated by a human (9.37 mkm). It annihilates humidity of buildings independently of outer space humidity. At the same time it neither burns oxygen nor overdries air.

Healthy effect is produced by all the constructive elements made of ecologically appropriate natural materials. Electric power is transformed into heat emission by conducting elements. This process causes a mode of generation of a continuous heat spectrum of radiation. The heat radiation is similar to heat spectrum of radiation generated by a human. At the same time, this feature together with resonant oscillations of the crystal lattice of the ceramic cylinders produces a destroying effect on pathogenic and conditionally pathogenic microorganisms.

From the point of view of room ecology the proved healthy effect of the operating electroconvector becomes very significant. The effect considers continuous presence of a human in the room, i.e. risk of pathogen infection through respiration objectively decreases.

Application of the ceramic electroconvector is especially actual in patient care institutions and children's institutions, in special precision industries and space technologies. It can also be used in saunas with dry vapor.

The ceramic electroconvector is designed for unsupervised continuous work.

Structure of the electroconvector producing the healthy effect

Shortcomings of known electric heaters are: great electric power consumption (0.75-3.0 kW/h), big number of metal details, complex technical performance, and use of a necessary additional blower as a ventilator. Big number of metal details decreases ecological compatibility and productivity of the heaters. The listed devices have to use high temperatures on heat-release surfaces for warming up rooms up to the level of physiological comfort. This causes increase of energy consumption. Moreover, using of metal heating elements influences on air and relative humidity in the room. According to available data, no existent electroconvectors have healthy influence on ecology of a room.

The represented ceramic electroconvector contains a carrying frame with horizontal boards which have convective gaps. The carrying frame has heat-release monolithic hollow ceramic cylinders whose walls contain longitudinal through holes. Heating carbonic conducting elements and de-energized carbonic rods are built in the through holes. The heating conducting elements are connected in parallel-series circuit at the output of the cylinders with it's ends placed into insulating supports of the carrying frame.

Constructive heating elements are the main differences of the device. All the heat generating constructive elements, i.e. conducting and de-energized heating elements, as well as heat-release surfaces of ceramic cylinders have contiguous spectra of infra-red radiation.

Energy-efficient effect

The electric scheme of connection of the conducting elements provides different modes of work of the device. According to these modes, energy consumption is in the interval of 0.05-0.3 kW/h. The mode of 0.05 kW/h is calculated on a supporting level of warming-up of a room. Maximal energy consumption (0.3 kW/h) corresponds to the superior limit of the temperature mode of heating of working heat-release surfaces of the ceramic cylinders. In this case temperature of the heat generating elements, i.e. de-energized carbonic rods and conducting carbonic elements) lies in the limit of max 100°C. This produces a significant potential resource of electric strength and of durability of the used elements. The de-energized carbonic rods function as heat accumulators at switching on and switching off the device. Heating the conductors the de-energized rods accumulate heat through the ceramic walls of the cylinder till their temperature becomes equal to the temperature of the conducting elements. Appearing electromagnetic resonance between the conductors and the de-energized carbonic rods intensifies infra-red radiation of the ceramic wall. Accumulative heat potential of the de-energized carbonic rods allows maintain uniform radial heating of the heat-release ceramic surface of the cylinder without decreasing energy consumption. In this case additional conducting elements of infra-red radiation become unnecessary.

The heating efficiency of the ceramic electroconvector was estimated by independent experts. Surface density of the radiation flow was calculated by formula of Stefan-Boizmann distribution law. Taking into account heating of the cylinders' surfaces up to 70° C this value came to 727 W/m^2. **Total heat generated by the electroconvector per hour comes to 600 kcal or 698 W** at the total area of the radiation surfaces of 0.96 m^2 (on the basis of 1 kcal=1.163 W/h according to [1]). The electroconvector consumes 300 W/h and produces 698 W/h. That is to say that it effectively transforms electric energy into heat energy. Operational modes of the electroconvector are based on analysis of the heating effect produced by different devices, i.e. a tube metal heater (TMH) having a conducting metal element of Nichrom, a ceramic cylinder having a Nichrom conducting element, and a ceramic cylinder having a conducting element of carbonic ribbon. The carbonic ribbon produces high heating efficiency. Temperature on the surface of the ceramic frame is up to 80° C and temperature on the conducting element is 100° C. Hence, the device equipped with the conducting element of the carbonic ribbon does not produce excessive heat radiation.

The electroconvector consumes 300 W/h and produces 698 W/h.

After one-year operation of the preproduction models in a private school it was noticed that number of respiratory illnesses and influenza among the pupils slumped, condition of skin became better, attacks of bronchial asthma among the teachers stopped, and allergic itch left the patients who suffered from allergy. A side effect was that flowers began to grow better, and those which had not blossomed began to blossom at least. Obviously, a combination takes place there: air cleaning of microorganisms, humidity normalization, warming comfort, and influence of pyramidal ceramic structures on water vapors which are transported by warm blasts.

Applied Know How

Ambient air in rooms is a complex substance including various chemical compounds, ions, dust parts, water vapors, infectious and potentially infectious microflora, etc. The electroconvector is represented by a ceramic hardphase crystal structure. Cold air blast moves close to the ceramic surface. It meets combined oscillation of crystal microlevel structures of ceramics. *Filtering electromagnetic screen* appears that breaks shell of pathogens and potential pathogens. It is a performance of **sanitation properties**. Atmospheric water contacts with structures of ceramics, then it is cleaned and structured. After that water cleans air which is breathed in by a human. Hence, the organism becomes healthier. It is a performance of the **healthy effect**. A clean room warmed uniformly improves human's health.

Main Performance Attributes of the Ceramic Electroconvector Producing Healthy Effect

Nominal required power	0.05-0.3 kWt/h
Voltage of the feeding network	220 ± 22 V
Frequency of alternating current	50 Hz
Temperature of the conducting element in operation mode	50…110° C
Temperature of the heat-release surface of the cylinder	39…90° C
Surface density of radiation flow	727 Wt/m^2
Quantity of radiated heat	600 kcal/h
Class of protection	1
General area of heat-release surfaces of the cylinders	9600 cm^2
Sizes	410 x 400 x 75 mm
Weight	18-21 kg
Durability	30 years

The ceramic electroconvector is designed as a floor construction for repetition work.

References

1. Heat Engineering Reference-book, Moscow, Energy (Energia), 1975, V. 1, p. 12.

About the author

After graduating from Polytechnic University of Frunze city Alexandra L. Belyaeva (1953) participated in launching of semiconductor production. She was given a rank of Perfect Inventor and Rationalyzer for invention and rationalization action. In 1987 her invention named Method of Production of Monosilicon Seed Crystals was introduced into manufacture, according to close copyright of USSR. Since that year this introduction has allowed grow monosilicon having better semiconductor properties. She was given a rank of Inventor of USSR for the invention.

Alexandra L. Belyaeva has acknowledged and given a rank of Inventor of Kirghiz Republic. Since 1994 in collaboration with Russian specialists of ceramics she has been dealing with organization of ceramic productions in Kirghizstan. Here are the inventions by Belyaeva: Method of Production of Seed Crystals and Its Variants (Patent #42 KG), Ceramic Mass for Insulators' Production (Patent #43 KG), Ceramic Mass Having Heat-Radiating Properties (Patent #464 KG).

In 1992 Belyaeva was rewarded with a Silver medal of VDNH (Exhibition of Achievements of National Economy, Moscow) for development of technologies for production of constructive nitride-cadmium items of semiconductor wastes.

Her Method of Production of Nonexpendable Seed Crystals is at introduction stage. The seed monocrystals produced by the new method will allow grow ribbon monocrystals having certain properties. It will be possible to produce items made of the monocrystals which will meet requirements of energy-saving and ecological appropriation. These devices represent non-aging powerful energy sources, various semiconductive devices with no inner microdeffects and with unlimited durability, structural water filters.

Since 2002 she has been dealing with organization of production of ecologically appropriate industrial-domestic appliances causing healthy effect. The invention of Ceramic Electroconvector Producing Healthy Effect is now introduced. Energy-saving of the new ecologically appropriate electroconvector exceeds energy-saving of all the existent heating systems.

Belyaeva is a laureate of competition of The Best Inventing in Kirghiz Republic in 2001-2002.

Inertial Propulsion Device

Vitaly E. Senkevich, Russia

E-mail: hhhhha@mail.ru

A body is at rest or moves linearly and uniformly until it is not influenced by an external force.
(School course of physics)

Forget everything that you were taught at school.
(Arkady Raykin)

This propulsion device consists of an engine and a body. The engine (see Fig.1) is remarkable for its movable stator (S) which can make free rotation as well as a rotor (R). When the engine starts the stator and the rotor begin to rotate in opposite directions. Thus the engine has two ends and one of them is connected to a flywheel (F). This flywheel begins accelerated rotation.

A cross-beam (CB) is connected to the second end of the engine where a rotating momentum appears. Under the influence of this momentum the cross-beam presses down one of the supports placed on the body (for example, Sup.1). As a result a force which is compensated by acceleration of the flywheel (F) appears on the support. When the flywheel is accelerated up to certain speed a control system (CS) switches the windings of the engine to change the direction into the opposite one (reverse). At that the cross-beam also tends to turn to the other side and presses down the second support (Sup.2). Thus the cycle repeats. It should be noted that forces acting at the supports are **codirected** and they move the whole device.

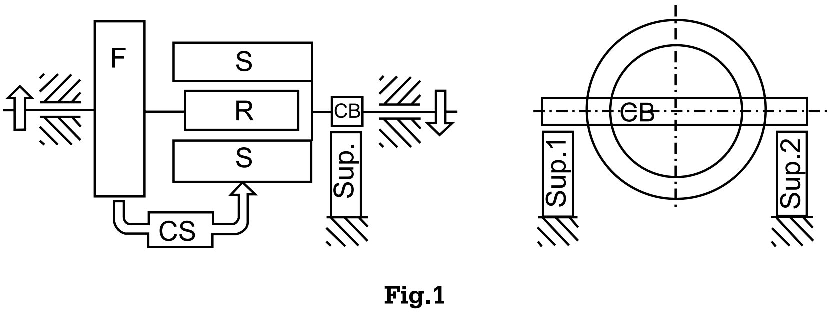

Fig.1

About the author

After graduating from Polytechnic University of Frunze city Alexandra L. Belyaeva (1953) participated in launching of semiconductor production. She was given a rank of Perfect Inventor and Rationalyzer for invention and rationalization action. In 1987 her invention named Method of Production of Monosilicon Seed Crystals was introduced into manufacture, according to close copyright of USSR. Since that year this introduction has allowed grow monosilicon having better semiconductor properties. She was given a rank of Inventor of USSR for the invention.

Alexandra L. Belyaeva has acknowledged and given a rank of Inventor of Kirghiz Republic. Since 1994 in collaboration with Russian specialists of ceramics she has been dealing with organization of ceramic productions in Kirghizstan. Here are the inventions by Belyaeva: Method of Production of Seed Crystals and Its Variants (Patent #42 KG), Ceramic Mass for Insulators' Production (Patent #43 KG), Ceramic Mass Having Heat-Radiating Properties (Patent #464 KG).

In 1992 Belyaeva was rewarded with a Silver medal of VDNH (Exhibition of Achievements of National Economy, Moscow) for development of technologies for production of constructive nitride-cadmium items of semiconductor wastes.

Her Method of Production of Nonexpendable Seed Crystals is at introduction stage. The seed monocrystals produced by the new method will allow grow ribbon monocrystals having certain properties. It will be possible to produce items made of the monocrystals which will meet requirements of energy-saving and ecological appropriation. These devices represent non-aging powerful energy sources, various semiconductive devices with no inner microdeffects and with unlimited durability, structural water filters.

Since 2002 she has been dealing with organization of production of ecologically appropriate industrial-domestic appliances causing healthy effect. The invention of Ceramic Electroconvector Producing Healthy Effect is now introduced. Energy-saving of the new ecologically appropriate electroconvector exceeds energy-saving of all the existent heating systems.

Belyaeva is a laureate of competition of The Best Inventing in Kirghiz Republic in 2001-2002.

Inertial Propulsion Device

Vitaly E. Senkevich, Russia

E-mail: hhhhha@mail.ru

A body is at rest or moves linearly and uniformly
until it is not influenced by an external force.
(School course of physics)

Forget everything that you were taught at school.
(Arkady Raykin)

This propulsion device consists of an engine and a body. The engine (see Fig.1) is remarkable for its movable stator (S) which can make free rotation as well as a rotor (R). When the engine starts the stator and the rotor begin to rotate in opposite directions. Thus the engine has two ends and one of them is connected to a flywheel (F). This flywheel begins accelerated rotation.

A cross-beam (CB) is connected to the second end of the engine where a rotating momentum appears. Under the influence of this momentum the cross-beam presses down one of the supports placed on the body (for example, Sup.1). As a result a force which is compensated by acceleration of the flywheel (F) appears on the support. When the flywheel is accelerated up to certain speed a control system (CS) switches the windings of the engine to change the direction into the opposite one (reverse). At that the cross-beam also tends to turn to the other side and presses down the second support (Sup.2). Thus the cycle repeats. It should be noted that forces acting at the supports are **codirected** and they move the whole device.

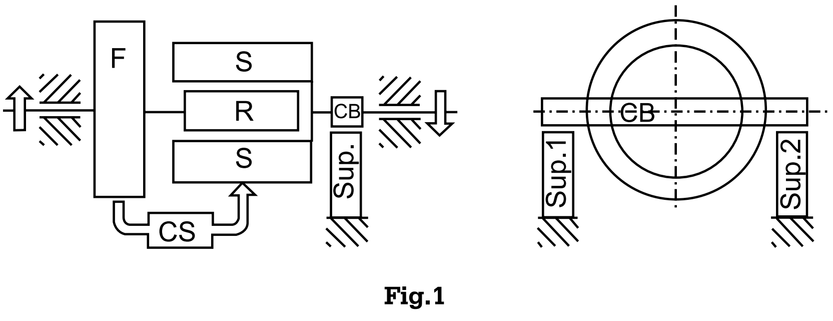

Fig.1

Unlimited Accelerated Nonreactive Motion

by G.P. Ivanov

Information from http://tts.lt/~nara/ruspopul.htm

Bias of centre of inertia of the closed system is a phenomenon exceeding the bounds of modern scientific conceptions. At the same time it is a way to the new marvelous world of earlier unknown laws and phenomena of nature. Researches made by G.P. Ivanov, Russia, have allowed him come to the following conclusion: it is a quite realizable task to create technical devices which can move under the action of nonreactive forces. However a purposeful scientific-research spade-work is required for the reliable registration of them by modern experimental facilities. According to the author during the whole XX century the known idea of latent impulse were misleading the scientists from the serious research of impulse-energy processes existing in systems at the presence of quasi-stationary electric and magnetic fields. According to G.P. Ivanov, it is related with the fact that the notion of "latent impulse" has nothing in common with the real momentum, since the very existence of "latent impulse" and "latent energy" which attends it, would make it impossible, for example, to adjust radio equipment since all effective capacitances change their value a hundred and thousand times as much (it depends on their orientation with respect to the magnetic field of the Earth).

Fig. 1, 2 demonstrate the patented by G.P. Ivanov method of realization of nonreactive motion (G.P. Ivanov, Yu.G. Ivanov. Method for production of propulsion. Patent #2172865, M., 2001). Fig. 1 demonstrates a device which consists of magnetized core with the attached metal electrodes.

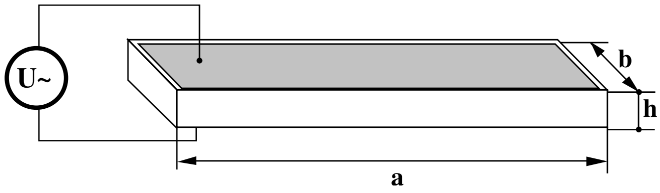

Fig. 1
The simplest "nonreactive" element

When alternating voltage is applied the device together with the center of inertia of the whole system (including power source and lead) will oscillate under the action of nonreactive force. It will move along the direction which is perpendicular to the vectors of electric and magnetic fields inside the core.

On Fig. 2 there is an analogous device supplied with a cylinder core. Magnetization of the core is defined by current of the coil which is wound around it (see Fig.2).

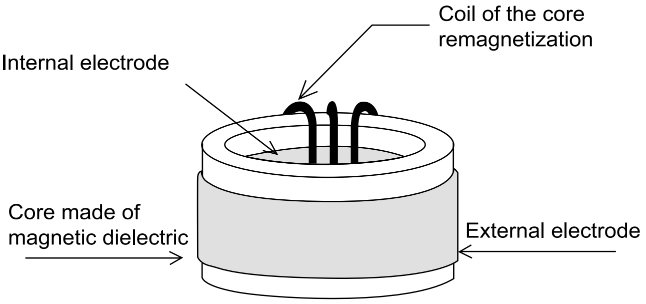
Coil of the core remagnetization
Internal electrode
Core made of magnetic dielectric
External electrode

Fig. 2
Diagram of the device which can be in unlimited accelerated nonreactive motion

Let us supply the electrodes of the device with such voltage which is enough to change field density in the core (for example, according to the law $E = E_0 \sin\omega t$), and the coil with voltage enough to change the magnetization (according to the law $M = M_0 \cos\omega t$). Then nonreactive force which is constant in its direction will act on the device along the axis of the cylinder. The average of the force is $F_{cp} = \pi f E_0 M_0 V/c^2$, where E_0, M_0 are the amplitudes of density values of the electric field and amplitudes of the core magnetization, **f** is frequency, **V** is core capacity. As a result the device can either move with acceleration or act against external forces. The researches allow the author to prove the validity of the following theorem: **At motion of the open-loop system (device) the work made by nonreactive force could not be realized by means of decrease of energy of the proper (appurtenant to the system) power source.** Where this energy comes from if there are no artificial power sources outside the device? However everywhere there is such form of matter as electrovacuum. It allows us come to a conclusion that nonreactive forces make work by means of decrease of electrovacuum energy.

Existence of electrovacuum does not contradict to the modern physical picture of the world. On the contrary providing the realization of laws of momentum and energy conservation, this idea originates from and organically supplements it. The theory of electrovacuum opens quite realizable prospects in different fields of human activity which seem to be fantastic and impossible. It also proposes unusual ways to solve many problems which are considered as insoluble. For example, it becomes possible to create nonreactive cosmonautics and private aircrafts. Energetics will be changed beyond recognition since power sources which require no material fuel will occupy the place of big and small modern electric power stations. The analogous list could be continued. There are presented quite realizable prospects related to inexhaustible reserves which are hidden in the ambient space.

Revolutionary Current

Suresh Kumar Baliyan, India

E-mail: suresh_baliyan@rediffmail.com
suresh_baliyan@yahoo.com

*Electricity is one of the most wonderful inventions of the 18th century. It can be produced by various methods in which different types of energy are used such as mechanical or chemical energy. In this article we discuss a new type of methods in which **electrostatic energy** is used.*

In this method we use **electrets** as an energy source. Electrets are permanent polarized dielectric material which is made by cooling dielectric material in a high intensity electric field. When we place an electron in the electric field of a point charge, it is influenced by a force. If the direction of the force is such that the electron moves in a closed path then it represents a current in the opposite direction along the closed path. Here closed path means a metal wire loop in which the electron movement is responsible for the current.

The charge of the electrets remains constant for many years (100) and there is no loss of energy because we use only the property of electrets that they apply force when a charge is placed in its electric field. Since here we create energy in the form of electric current then it violates the Law of conservation of energy.

Let us discuss the method in three steps:

1) Fundamental
2) Equivalent circuit
3) How we can use it as a potential source of electricity.

1. When we put a metal rod in the electric field of a point negative charge then the electron is influenced by repulsion force which moves in arrow direction and the electric field is cut by putting an earthed metal plate from any direction. When we place four charges on the corner of the metal frame and shield the electric field in particular direction as shown in Fig.1 then the force on the electron which is placed in metal frame will be in the direction of arrow and this makes a loop. The force influencing on the electron will be continuous, so the current which moves in direction opposite to the electron movement will induced in the metal frame.

2. The equivalent circuit of Fig.1 is shown in Fig2. It demonstrates that a 'V' volt battery and a resistance 'R' can be replaced by an arm of the metal frame. The polarity of the battery is shown in Fig. 1. A current of magnitude (V/R) will be induced in the circuit by this method.

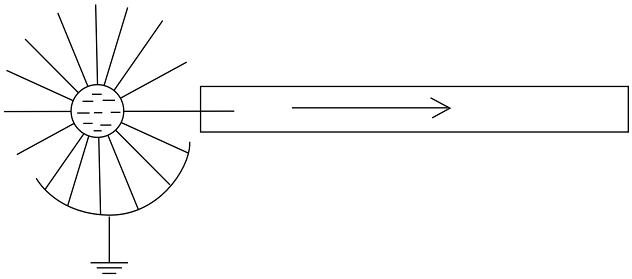

Fig. 1a

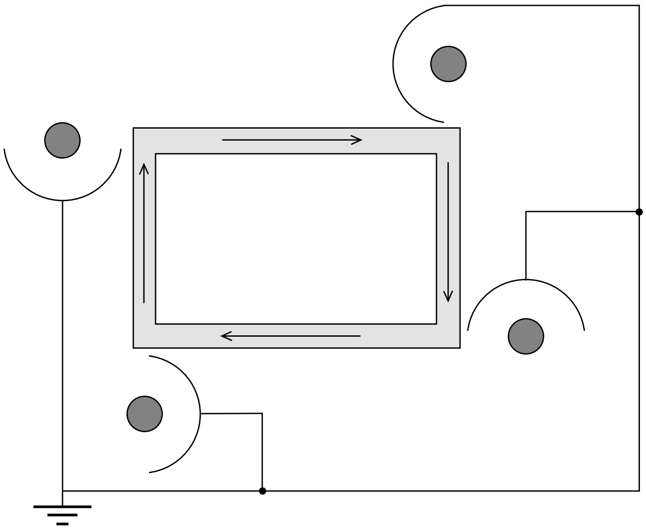

Fig. 1b

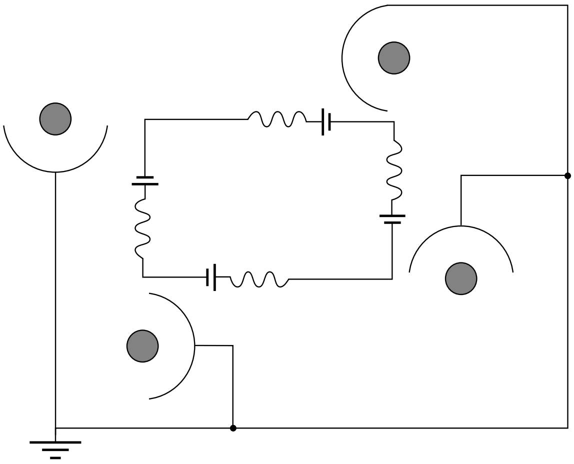

Fig. 2

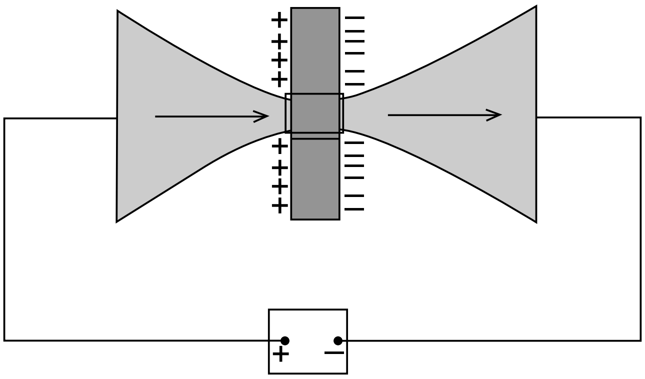

Fig. 3

3. We can use another device as a potential source (see Fig3.). Let us take an electret and make a hole at the centre. A hollow thin walled metal cylinder is fitted in the hole and a metal frame is taken as shown in the Fig. 3. We can use this metal cylinder as an electric field shield which may affect the direction of the force influencing on the electron. We may use this device as a potential source of electricity which will give electricity until the charges on the electrets vanish. The charge on the electrets remains constant for many years. Thus we can create energy and the problem of energy shortage can be solved by this method forever. We are trying to make a prototype of this but due to lack of equipment facility we are not able to demonstrate a working model. However we are trying to modify it.

NEWS REVIEW

Russian scientists are preparing new application of energy of geothermal waters

http://www.sciteclibrary.ru

A way for economical use and utilization of energy of geothermal waters was found in the Institute of Geothermal Problems (Russian Academy of Sciences). The new technology can be applied for heating in different purposes, and can compete with the power engineering, which uses the conventional energy carriers.

Geothermal springs are ideal for generating electricity by means of using of their permanent heat. This is especially important for Kamchatka, which is rich in these springs; however they are almost never used in mass power engineering, but only serve as an attraction for tourists, who admire beauty of Kamchatka geysers. At the same time the region itself freezes every year because of shortage of fuel for electric power stations.

What prevents us from using the geothermal springs?

The problem is that the storages of most of geothermal water fields have low and medium temperatures. This does not allow providing their competitiveness as regards conventional energy carriers. During tens of years this fact was an obstacle to the development of geothermal energy.

A solution was found in the Institute of Geothermal Problems. It was brought out that at many exploited geothermal fields wellhead overpressure exceeds 5-10 MPa and more. These waters contain fair quantity of dissolved organic gases. The scientists have found out that methane content of these waters exceeds 90 %. And until now at exploitation of the fields these types of energy have not been properly utilized.

The scientists of the Institute of Geothermal Problems have developed a technology of the optimum utilization of energy of geothermal waters and have increased the thermodynamic effect of the process.

This aim was fulfilled by means of transmission of thermal energy of geothermal water through the intermediate heat exchangers to the secondary heat carrier. Chemical energy of dissolved gases was used as an additional source of energy. The transmission was held by the use of primary and secondary separators. The distinctive feature of their technology is that associated potential energy of geothermal water is used as an additional source of energy. Expander and compressor placed on the same shaft are used as potential energy converters. Gas-holder and gas control point are used for utilization of energy of dissolved gases.

Thus on the bases of this technology geothermal energy can be effectively developed that will be competitive with fuel hydrocarbon energy.

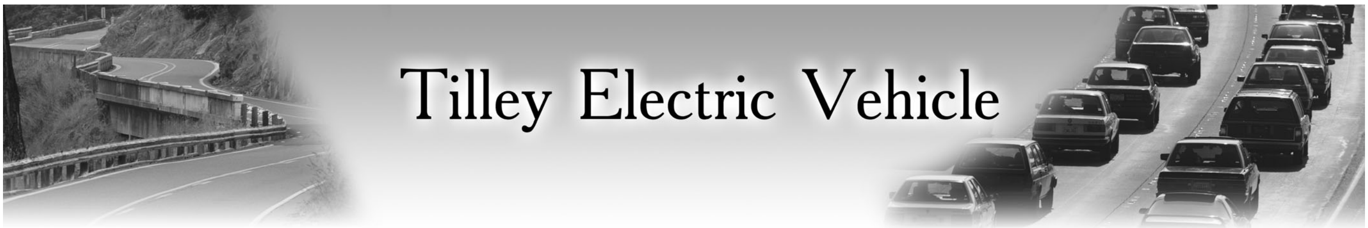

Tilley Electric Vehicle

Tilley Foundation, Inc.

131 Hiwassee Road
Lebanon, TN 37087
http://www.tilleyfoundation.com
E-mail: cktilley@bellsouth.net

Editor: The advantages of electric vehicles were proved more than once. No fuel is required to set them in motion. Along with evident advantages of electric vehicles, such as their powerful characteristics and harmlessness for environment there is a side benefit, i.e. much less maintenance as compared with gasoline or diesel-fueled vehicles is required. Moreover it is easier to manufacture such vehicles.

In the previous issues of New Energy Technologies magazine we have already acquainted our readers with Tilley Electric Vehicle designed by Carl B. Tilley (USA). Below there is some up-to-date information and photos from the inventor. Besides we also publish an article devoted to the similar types of fuel-less motors.

After several years of personal accomplishments in the alternative energy industry, Carl B. Tilley was convinced that it was possible to build an electric car that could be powered without the help of external power to keep the battery charged.

The concept to produce a useful electric performance car that would last more than a few hours and would be economical to run, safe to drive around town or across the United States and never use a drop of fuel challenges the future of transportation as we know it today.

With the establishment of the Tilley Foundation, Inc., in the year 2001, Carl Tilley set out to prove it could be done. It was an ambitious project and it broke ground on the facility in Tennessee that would build the first self generating electric car.

...you have no need for fuel and you do not have to stop the vehicle to charge it after driving.

Construction of a 1,800 square foot building, that was powered with another recently developed electric device, began in the year 2002. Electricity from the building built the car from a different energy invention, that was void of any outside power supply. It is ironic that one alternative energy device actually built the invention to power and build the electric car.

From the selection of the proper car to be converted, to the advanced technology which is on board , the **Tilley Electric Vehicle TEV** performs comparably to gasoline powered vehicles. The difference is you have no need for fuel and you do not have to stop the vehicle to charge it after driving. There is no pollution and you can cruise the highways at the same speed as any other vehicle.

The rear mounted electric motor provides over 130 SAE net horsepower at 5,500 rpm. It offers a 3 speed automatic transmission that is smooth shifting and totally silent running. All this is combined with rack and pinion steering and a 35/65 rear weight bias that enables fast, sensitive handling and needs no power assistance.

...battery system will be fully charged at all times while in use.

It has 4 wheel disc braking for fast progressive, fade-free stopping. Counterbalanced gull-wing doors need only 14 inches of clearance. The rear sporty louvers are aerodynamic designed so that it almost eliminates any drag effect. All this with a stainless steel body makes for a great car.

Control center for the battery bank only allows what is needed to keep the batteries charged while in operation no matter what the speed or discharge from the battery bank. Your battery system will be fully charged at all times while in use. Simply get in, start the car and drive like any other vehicle.

A 1981 DeLorean was converted as the Tilley Electric Vehicle.Conversion of the car began in late June of 2002. State of the art metal fabrication to construct support for the electric motor, battery bank, control center and the TEV device was completed in July of 2002.

Several tests were made to validate the TEV technology. One of the last tests was made on September 7th, 2002. It has been demonstrated that after 17.3 miles driven on the Superspeedway at speeds ranging from 80MPH to 96MPH independent engineer certified batteries were full.

Assembly Pictures (see also the cover page)

Photos by Robert Gaither

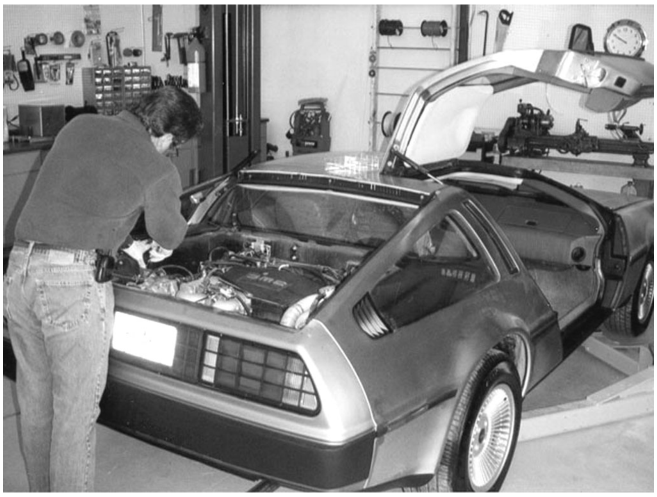

Fig. 1
Removing Gas Motor

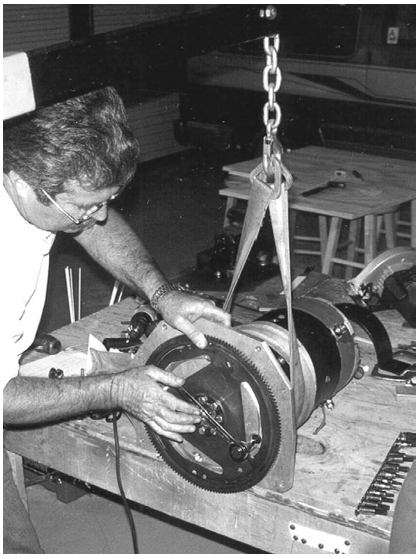

Fig. 4
Fitting Parts

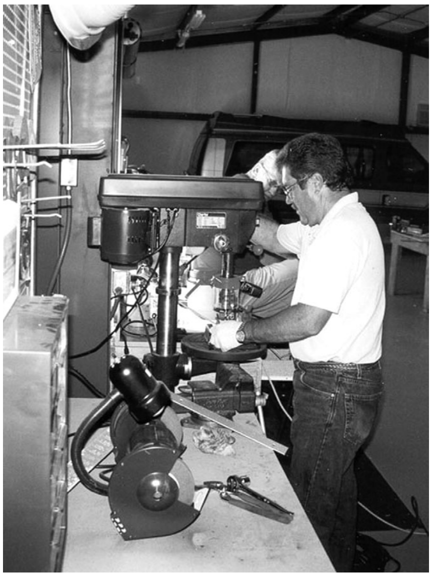

Fig. 2
Fabrication

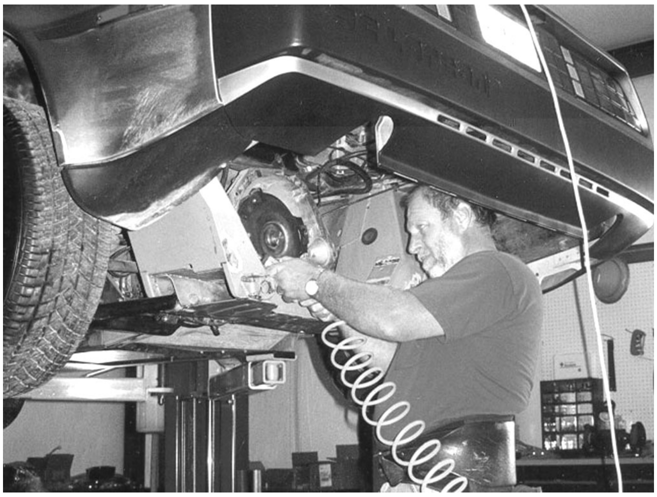

Fig. 5
Transmission Work

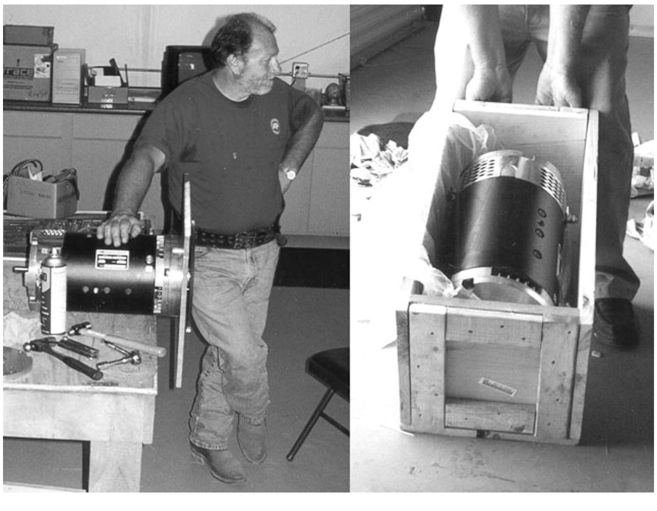

Fig. 3
Drive Motor

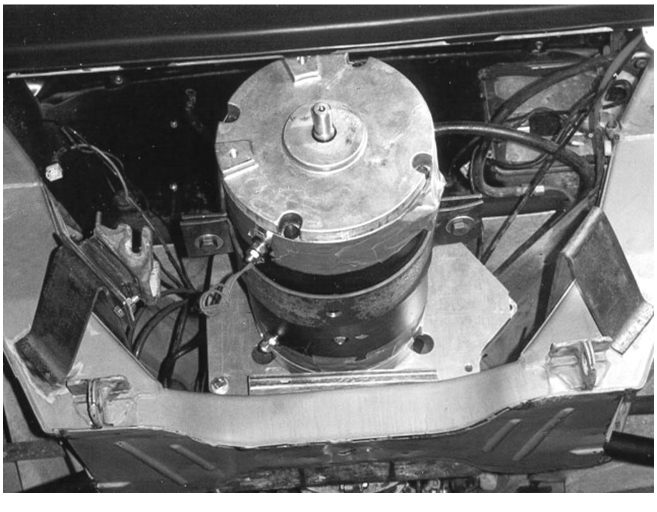

Fig. 6
Motor Mounted

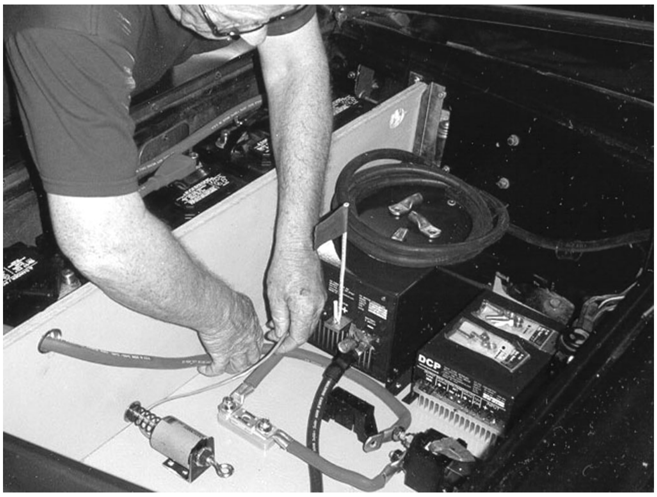

Fig. 7
Making Connections

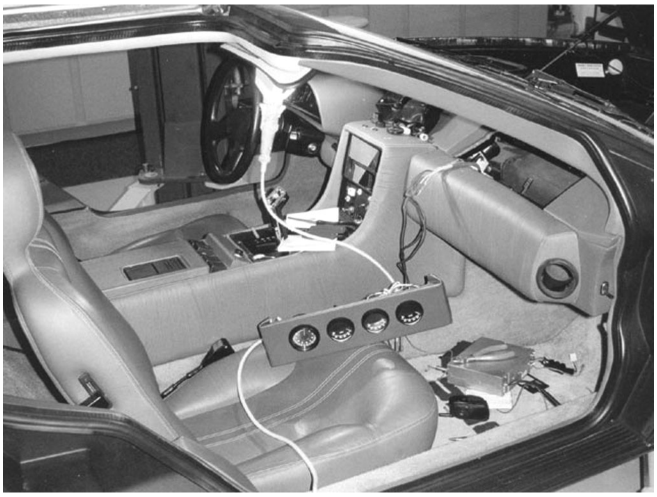

Fig. 9
Special Wiring &Guages

Fig. 8
About Ready

Fig. 10
Car Unveiling

Trends in the Application of Motor-Generators

Adrian Akau, USA

E-mail: adrianakau@aol.com

If the Tilley generator-motor should enter into the world market, some curious consequences may result, the most important being the transition of the Honda and Toyota hybrid cars into full fledged electrics.

First of all, a comparison should be drawn between the Tilley and the two hybrid cars presently being sold, the Honda Civic and the Toyota Pirus should be made.

The Honda Civic and the Toyota Pirus both use charging systems with a gasoline motor for better milage. The Tilley motor-generator is a stand-alone electric vehicle. The Honda motor-generator is just 60mm thick and provides (10kW or 13hp). The Toyota Pirus has a 44hp unit (American Version) which feeds power into electric motors at the wheels. Both Honda and Toyota gasoline motors are off at 0 mph. Only the electric system is used until the power demand reaches 10 kW; then the gasoline motor automatically kicks in. The Pirus is able to get higher milage in city driving than in country driving from the fact that the ratio of the power from the electric part to power to the gasoline part of the motor is greater at lower than at higher speeds, that is, less power is needed to run the car in city driving than in country driving because of the lower speeds and air resistance.

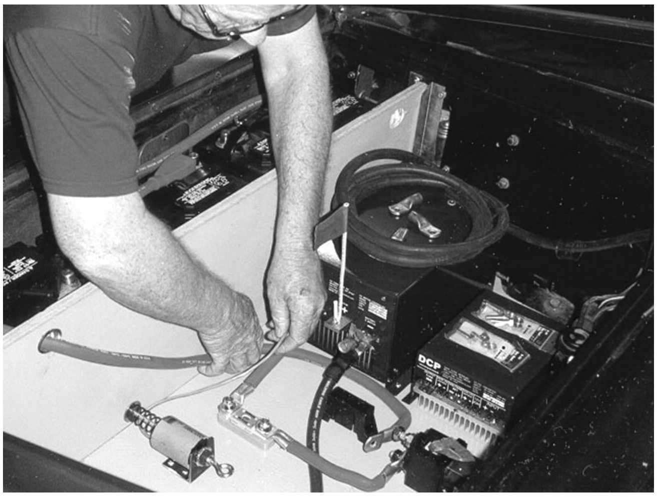

Fig. 7
Making Connections

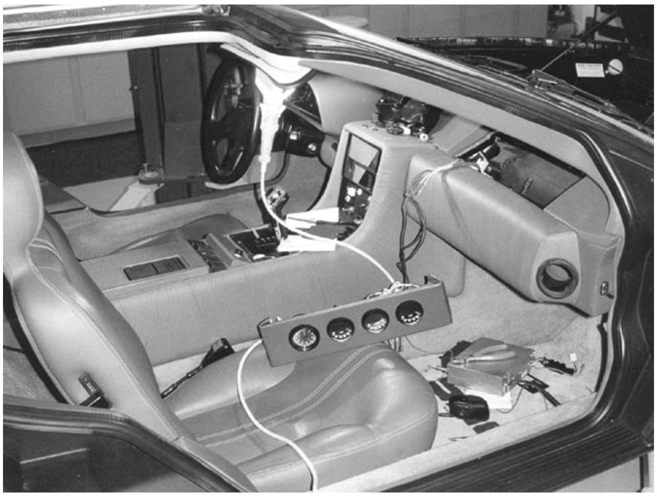

Fig. 9
Special Wiring &Guages

Fig. 8
About Ready

Fig. 10
Car Unveiling

Trends in the Application of Motor-Generators

Adrian Akau, USA

E-mail: adrianakau@aol.com

If the Tilley generator-motor should enter into the world market, some curious consequences may result, the most important being the transition of the Honda and Toyota hybrid cars into full fledged electrics.

First of all, a comparison should be drawn between the Tilley and the two hybrid cars presently being sold, the Honda Civic and the Toyota Pirus should be made.

The Honda Civic and the Toyota Pirus both use charging systems with a gasoline motor for better milage. The Tilley motor-generator is a stand-alone electric vehicle. The Honda motor-generator is just 60mm thick and provides (10kW or 13hp). The Toyota Pirus has a 44hp unit (American Version) which feeds power into electric motors at the wheels. Both Honda and Toyota gasoline motors are off at 0 mph. Only the electric system is used until the power demand reaches 10 kW; then the gasoline motor automatically kicks in. The Pirus is able to get higher milage in city driving than in country driving from the fact that the ratio of the power from the electric part to power to the gasoline part of the motor is greater at lower than at higher speeds, that is, less power is needed to run the car in city driving than in country driving because of the lower speeds and air resistance.

American car companies are beginning to develop their own starter-generators to "save an extra 10% on the gas milage". Are they doing this because they don't want to be left behind in the milage competition or because they really understand the nature of the starter-generator motors used by the companies producing the hybrids?

With the advent of the Tilley motor-generator, a new factor enters into the situation. If the system proves valid, does this mean that Honda and Toyota are out of the running? Hardly not! It is likely that these companies have allowed or even planned for this contingency with larger capacity motor generators which would do a similar job. With their fine cars and superb engineering, the transition from a hybrid to a pure electric might be costly but not too difficult to engineer. Each of the two hybrid companies may already have been built, tested vehicles of this nature. However, since the primary purpose of the car companies is to make money it is doubtful if this full-fledged fossil fuel-less type will be placed on the market unless competition so requires. A car using the Tilley motor-generator might certainly push them in this direction.

Should the Tilley remains undeveloped, one might expect the second level hybrid car to arrive in a few years from Honda and Toyota. These cars should have greater electric generator-motor capacities and smaller gasoline engines. These systems could give small cars much higher gas milage and could also be placed in larger, more elegant vehicles such as the Honda Accord and the Toyota Camrey.

The transition to fuel-less motors should hopefully be slow and painless for both the pocket book as well as the human psych. If scientists cannot be convinced by studying conventional scientific laws on how these motors work, perhaps they will better see the need for the promulgation of new laws in this area.

We know that at the present time, both Honda and Toyota motor generators (M.G.) have a powers assist function while the Tilley stands alone. Let us look at the present situation (we will use the Toyota U.S. M.G. because it is the more powerful version):

Honda and Toyota classify their M.G.'s as permanent magnet types; Carl Tilley says his is and advanced D.C. motor.

Power output:

Honda is given at 10 kW at 3,000 rpm or 13 hp
Toyota is 33 kW at 1,040-5,600 rpm or 44 hp
Tilley is up to 135 hp depending on rpm's

Voltage/Battery (Ni-MH=Nickle Metal Hydride, AH=amp hours)

Honda: 144V, 120 Ni-MH cells@1.2V ea., 6.5 AH
Toyota: 274V, 228 Ni-MH cells@1.2V ea.
Tilley: 144V,12-12V lead acid (Wal-mart), 1200AH

Physical Characteristics of each M.G.

Honda: 60 mm wide
Toyota: unknown
Tilley: 9" dia., 146 pounds

The main effort should be to implement the transition away from the use of fossil fuels. We know this can be done but it takes proper attitude. Proper attitude takes knowledge and knowledge takes willingness to learn. Willingness to learn takes desire to learn and desire to learn takes humility. Unfortunately, humility is a characteristic that many people, especially learned ones lack. *Editor's: Carl Tilley's comments are also given below.*

Very Good…you seem to know what we already were aware of…before we could mass produce our Tilley car I am sure some auto company would come out with basically the same thing…I do believe they already have it and they are just waiting for the right time. The good news is that at least the Tilley car lighted their fire to get them started.

Carl Tilley

Read in the next issue!

In April 12 of 2003 scientific conference "The Time Machine" was organized by Faraday Laboratories Ltd in Moscow. There were presented the reports by V.A. Chernobrov, A.V. Frolov, A.V. Rykov, E.D. Sorokodum, V.J. Kosyev, A.N. Solonyi, V.A. Atsukovsky. There was discussed design of the device made by Faraday Laboratories Ltd. as well as main principles of control on temporal characteristics of physical processes by changing of density of space energy, i.e. aether density according to patent claim.

Details are in the next issue!

Electric Vehicle in Russia

Review by Correspondent Alla Pashova, Russia

The director of *EV* (*Electric Vehicles*) company located in Tolyatti (Russia), Alexander Mukhanov states: "If during 2 centuries intellectual efforts of a scientific thought were concentrated on an electric engine instead of the internal combustion engine then now we would have driven electric vehicles, and the greenhouse effect would not have exist".

The electric vehicle has driving-wheels which are set in motion by an electric engine, and it is powered by an electric battery. The electric vehicle appeared in Great Britain and France in 1880 and it is significantly older than the automobile equipped with the internal combustion engine.

At the beginning the speed and the fuel distance of the electric and gasoline vehicles were approximately the same. The main disadvantage of electric vehicles was a complicated system of recharge since there were no usual transformers of alternating current into direct one. Hence, the electric vehicle was recharged in a quite difficult way. An electric engine operated on alternating current was used to recharge it. The electric motor rotated a shaft of the generator which was connected to batteries of the electric vehicle. However in 1906 a rectifier which was simple to operate was invented. Despite its principle of operation was quite complicated (mercury vapor was used for transformation of alternating current into direct one) it has given a great stimulus for development of automobile industry.

There were no essential proves of advantage of gasoline vehicles as compared with electric vehicles. Nevertheless, even at that time there was an opinion that electric vehicles run more slowly and to shorter distances. However "gasoline" records were also insignificant and comparable with results produced by the electric vehicles. At the same time producers of electric vehicles did their best to demonstrate the advantages of these vehicles. If you remember, Walter Baker, one of the most famous American constructors and producer of the electric vehicles, achieved the speed of 130 km/h driving his car. An electric vehicle produced by Borland Electric company run from Chicago to Milwaukee (the distance of 167 km/h along non-asphalted road) without recharge. Next day after recharging the electric vehicle returned to Chicago by its own power without meeting any incidents. During the journey its speed came to 55 km/h.

In 1899-1900 a hereditary Russian nobleman Ippolit Romanov made an experiment on designing of the first Russian electric vehicle. Since these vehicles had been designed to run in Saint Petersburg then the city council required to produce them in the same place. In 1899 Romanov's first electric vehicle was made. Its general design was borrowed from English cabs in which a cabman sat on a high dickey placed behind passengers. Romanov's electric cab had two passenger seats situated in front of the dickey in a half-cabin having side and back windows. The cabin was equipped with a hood. The seat of a driver was placed behind and above the cabin; a box containing an accumulator was located under the driver's seat. The vehicle was four-wheeled. Diameter of the front wheels was smaller than those of the rear wheels. The front wheels were fixed by elliptical springs and were driving ones. They were connected to two independent electric motors by a roller chain drive. The regulated rear wheels had a smaller diameter and were fixed by spiral springs.

This first electric vehicle was equipped with lead accumulator which had 36 banks. It required recharging every 60 versts (1 verst – 3500 feet). The total power of both the electric motors was equal to 4 hp. Design of the vehicular part of Romanov's electric vehicle was borrowed from the models produced by an American company named *Morris-Salom*. This company had produced electric vehicles since 1898; however those models had wheels of greater diameter as they were equipped with pneumatic tires, while Romanov's vehicle had wheels with light rubber rims. Both the electric vehicles had two electric motors, but the distinction of Romanov's electric vehicle consisted in the presence of 6-row ball bearings in the wheels.

Romanov's second cab was built in 1900. This model had entirely closed and glazed cabin for passengers. The chain drive was replaced by a gear; however the basic sizes remained the same. This model was equipped with an accumulator designed by Romanov. The general weight of the electric vehicle was 45 poods (750 kg), and the weight of the accumulator came to 22 poods (362 kg). It should be noted that weight of American and French electric vehicles exceeded these figures. The speed of both Romanov's models of the electric vehicles came to 15 km/h.

Romanov's first electric bus was built in the same 1900. The electric bus weighted 100 poods (1600 kg) could make speed up to 10 km/h. For reasons beyond Romanov's control, traffic of the electric vehicles in Saint Petersburg was not organized.

A vehicular company named *"Frezer and Co."* participated directly in production of Romanov's electric vehicles, i.e. a running gear of these vehicles was created by this company. In January 1900 *The Cyclist* magazine reported that this company had built an electric vehicle which had already been tested. A picture of this first electric vehicle by *"Frezer and Co."* can be found in advertising publications of those years. Various firms proposed the innovation for sale and described its technical characteristics.

Frezer electric vehicle was four-seated, had two engines whose total power came to 7 hp. The advertisement booklets reported as following: "The accepted system of accumulator is remarkable for its solidity and life-time. Capacity of the elements is about 15 Amperes/ hour per a kilogram of electrodes. The size of the battery is defined according to calculation of 120 W/h per one ton and per one-kilometer of the distance; its weight is equal to 30-40% of the total weight of the vehicle including a payload". Weight of Frezer electric vehicle came to 70 poods (1120 kg). It made the speed of 15-18 km/h and required recharging every 35-50 versts of the path. The second variant of Frezer electric vehicle differed from the first one for less weight of the accumulator which was located above the rear axle.

In addition to the four-seated vehicles, *Frezer* Company produced two-seated electric vehicles equipped with engines of 3.5-hp power. The two-seated electric vehicle was remarkable for its more perfect steering tube. There is known one specimen of this model purchased by somebody from Riga (Latvia). Moreover, on the territory of *"Frezer and Co."* there was a station for recharging accumulators of the electric vehicles.

In the period of the Soviet Union pilot batches of electric vehicles began to be produced only after the II World War. In general there were small lorries applied for transportation inside great buildings, for example, in a main post office or an airport.

Now there is a great probability that Moscow will be the first Russian city where the electric vehicles can find mass application. Russian capital suffers from gas pollution which is especially heavy in the center of the city. At the same time the budget of the city allows acquire own ecologically appropriate transport. Now financing of production of the electric vehicles is realized through the Fund of Ecologization of Moscow Transport. The means are assigned due to income taxes coming to the budget of the city.

To tell the truth, citizens of big cities know electric vehicles very well, i.e. usual trolleybuses and trams are electric vehicles despite they are not autonomous. An autonomous public electric transport has one advantage which is especially valuable in the conditions of the cities overcrowded by cars. This advantage is their mobility. Trolleybuses are inseparable with their wires that causes decrease of their maneuverability. For example, they cannot pass several cars which have been parked wrongly. In this case a maneuverable and autonomous electric vehicle will run to the center of the road and pass the obstacle with ease.

A Moscow company named *Eltran* which deals with exploitation of electric transport has developed a model of EV which runs the distance of 60 kilometers without recharging. Total recharging of this electric vehicle takes about 5 hours, however a half of required energy they can obtain for 1 hour. Hence, if the accumulators are recharged before their full discharge then this vehicle can operate constantly. However, a developed infrastructure of charging stations is necessary for this aim.

From the economical point of view, EV does not yield to trolleybuses and trams which house more passengers but run twice more slowly than the autonomous and compact EV. Therefore EV does not experience difficulties in jams in the conditions of the traffic of the city. It maneuvers easily in thick traffic current.

Unfortunately, now EV cannot replace automobiles equipped with internal combustion engines. Imperfection of its accumulator batteries does not allow it to become a public vehicle. The battery imperfection causes a small run on one recharging, long cycle of recharging, and high price of the electric vehicle.

To tell the truth, EV has important advantages. Maintenance charges of EV are lower than the charges of the standard automobile requiring expenditures for support of cooling systems, powering systems and exhausting systems. The lifetime of the electric engine comes to 10 thousands hours. Quantity of operations for maintenance of the engine is minimal. For example, it is necessary to change brushes in the dc motor from time to time. As for a modern three-phase electric motor and synchronous ac motor they are practically maintenance-free.

EV is easy to drive. To start the car you should only insert the key, turn it and press the accelerator pedal. You do not need to make any manipulations with adherence or gearshift.

Use of EV is justified in so called zones of higher environmental requirements that are cities, parks, vacation resorts. Electric buses, developed by *Eltran* run in the public green space of All-Russia Exhibition Center. There are about ten compact commercial EV operating in Moscow.

Taking care of environmental sanitation the management of AZLK (big automobile plant) has developed EV on the basis of two production automobile models "Moskvich-2141" and "Moskvich-pikup-2335". "Moskvich-Elektro" (Fig. 1-2) at fully loaded mass of 2060 kg accelerates up to 60 km/h at 15 seconds. Maximum speed of the load-free vehicle is 110 km/hour. Cruising range at one battery charging is 100 km, and it comes to 80 km with 400 kg of load. At present Electrical Vehicle "Moskvich" is filled with imported component parts, which are cheaper but functional. In time it will be supplied with Russian equipment that will make it cheaper.

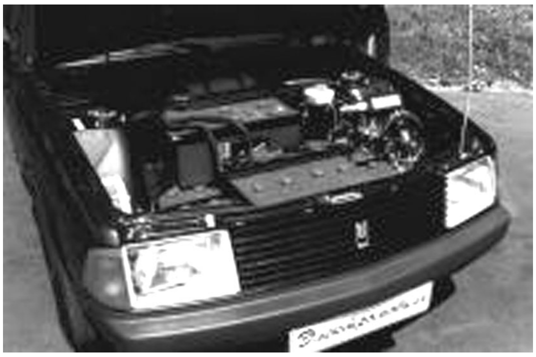

Fig. 1

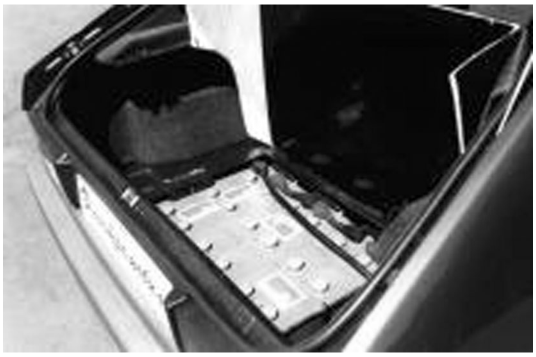

Fig. 2

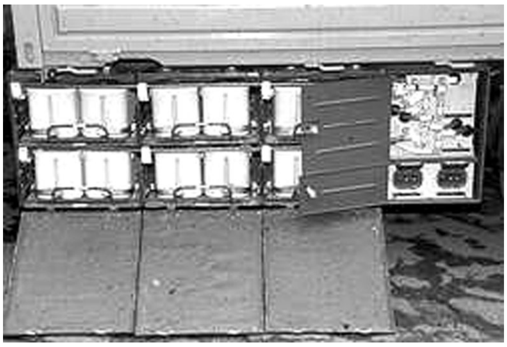

Fig. 3

The model of truck equipped with electric motor "ZIL-Electro", the authors of which being called "AVEKS" and "Optimum-electro", also enjoys the support of the Fund of Ecologization of Moscow Transport. Various privileges and free passage to any location in public green zones are promised to the future owners of "ZIL-Electro".

Forty-eight lead-acid traction batteries of roll type (Optima battery) are hanged in pairs along the ramps of the truck. They are united into 4 sections (Fig.3). Capacity of these batteries is enough for 70 km running. Probable charging station for these machines is the loading/unloading point. Charging time varies from 30 to 60 minutes. Substantial disadvantage of the model is that batteries take away a great percentage of carrying capacity. Among the advantages of the model there are small internal resistance, fast recovery (to 400 A), discharging to zero level without serious consequences.

A traction induction unit "ATAD-Optimum 50/120" serves as the engine for "ZIL-Electro". Its weight is only 100 kg and it doesn't require any diesel generator or transmission that compensates a little the loss of carrying capacity.

The engine is very simple. It is a non-contact rotor of a "squirrel cage" type rotating on the bearings. No brushes are used. Thus expensive maintenance is postponed for uncertain date. On the other hand there is asynchronism. It means that some equipment is needed for converting into three-phase current and further frequency and amplitude adjustment. However it is expensive and complicated.

In spite of all existing disadvantages, Electric Truck is considered to have a bright future ahead of it. There several reasons for that:

1 The vehicle is manufactured from cheap production component parts.
2 The system of optimum vector control of the drive is introduced. (The torque and shaft speed of rotation is regulated precisely, energy of the batteries is used rather efficiently.)
3 Microprocessor system with feedback coupling and great number of sensors (current sensor, heat sensor, velocity sensor, voltage sensor) minimizes losses, prevents the motor and batteries from overheating, protects mains from short circuit in case of an accident.
4 The vehicle is reasonably priced (for such an exotic object) – about $26 000 USD subject to serial production.

Volzhsky automobile plant has already been engaged in research of EV for a quarter of a century. Over 10 original models have been designed and produced for these years. They have been appreciated abroad as well as in Russia. These are electric vehicles "Poni", "Oka", "Elf", "Gnom", "Niva", "Rapan" "Lada-golf", etc.

The models manufactured on the basis of VAZ-1111 "Oka" became the most popular among them. These EV received many different rewards, and became world-renowned. For example, VAZ-111E "Oka", manufactured in 1993, kept all merits of a production midget car. Use of electric motor makes it a non-toxic and noiseless vehicle. Efficient fast-acting short circuit protection eliminates current rush, providing absolute car operating safety.

In EV of VAZ manufacture two direct-current motors are commonly used as a power-generating set. The first is a 25kW capacity motor with torque of 110 Nm.

Another one has 40 kW capacity and torque of 190 Nm. Motors of the first type are usually mounted on light vehicles, such as "Golf", "Oka-Electro", "Elf", while high-power motors are usually used for VAZ-2108, VAZ-2109, "Niva".

The first models of electric vehicles used to be equipped with nickel-zinc batteries. But as life time of these batteries is not long, it was decided to start using nickel-cadmium batteries, manufactured at Saint-Petersburg plants "Rigel" and "Istochnik". Nickel-cadmium batteries are power-consuming, they withstand temperature of –40°C. The only drawback of these batteries is their cost, and as a consequence a high cost of the vehicle itself. Because of that the final choice has been made in favor of lead-acid storage batteries, which were mounted at "Gamma Golf" in particular, thus reducing its price by several times. The batteries are located under the driver and passenger seats. They allow discharge of 80-85% and provide cruising range without recharge of up to 100 km within city area and several times more on a highway. Power current is transmitted from the electric motor to the front wheels through single-reduction gear unit, which substitutes the transmission.

For the present time electric vehicles are not developed to take the place of automobiles with internal-combustion engine in all industries and spheres of application, but only as a specific vehicle for highly specialized utilization. They are utilized when use of internal-combustion engine is objectionable or impossible.

It is considered that so-called hybrid models of EV would be able to replace automobiles with internal-combustion engines. These are vehicles with two motors, an electric motor and an internal-combustion engine usually running on diesel oil. In this case the combustion engine works constantly, but under the operation modes of little air pollution. Besides, fuel consumption is also much reduced.

In this direction real commercial success is possible. For example, "Toyota-Pirus" has already been chosen as a personal vehicle by tens of thousands of customers, this being a good result for such an original construction.

Nowadays three circuits of hybrid propulsion system are generally known. The simplest one is a series circuit, in which combustion engine operates together with a generator, and battery or generator powers electric motor providing propulsion. Actually this is the same circuit of electric transmission that is used at quarry trucks. Its main disadvantage is great losses of energy at transmitting the torque to the wheels. In parallel circuit the output shafts of the motor-generator and combustion engine are rigidly bound, and it doesn't allow working in steady-state mode (i.e. with permanent rotations and load).

In the third circuit, which is also the most commonly used there are an electric motor, combustion engine and a generator. The output shafts of all these units are bound by planetary differential. This allows almost lossless distribution of power between the units, and provides sustained performance of the combustion engine. The disadvantage of this circuit is the extreme complexity of coordination of units operating.

A hybrid EV was produced in Russia by automobile plant "Izhmash" though it has not been put in mass production yet. At the developing of the propulsion system instead of unique components there were used units of Russian mass automobile production that is one of the fundamental design solutions. "Orbit" Izh-21261 was used as a base for the device. It was equipped with gasoline engine (of 30 hp power and 650 cm^3 capacity), drive dc motor ПТ-125-12 (of 120V voltage and 49 H/m torque) and eight accumulators 6CE-55. Twin-engined propulsion system works according to the parallel circuit of engine joint. Electromotor serves as starter and at the acceleration it operates simultaneously with internal-combustion engine. On reaching some definite speed the electromotor changes its operation mode to the mode of generator and charges the accumulators. When the loads at gasoline engine increase then the electromotor begins to operate together with it. Operation of the hybrid propulsion system is controlled by starting regulation electronic equipment. Control unit, speed transducer of crankshaft of the internal combustion engine, detector of the accelerator pedal etc. are responsible for coordinated work of the engines.

During the tests fuel consumption of the hybrid vehicle "Orbit" was less by 20% than that of the usual model. For Russian prototype it is a good qualitative characteristic, however foreign designers reach better results, for example, "Crysler ESX3", which was presented in 1998 at Geneva auto show, consumes 3.3 liter of diesel oil per 100 km of path.

Representatives of specialized Moscow companies producing EV assert unanimously that these are hybrid electric vehicles which should change public transport with internal combustion engine. To encourage the automakers it is useful to create the proper laws drawing on foreign experience. For example, in California, USA, automakers whishing to present their production in the state market should put out 2% of automobiles with null blast. Legal system also supports an automobile user who is provided for free parking and free or cheap recharge. The user is granted with maximum of discounts to encourage him buying EV which is still twice as expensive. However it is said that in one of the villages situated near Moscow where the so called new Russians live, everybody drive electric vehicles leaving their cars with gasoline engine outside the gates. It would be quite good if EV turns to the real vehicle from the extravagancy of nouveau riches and beloved creation of ecologists.

The Richard Clem Motor and the Conical Pump

An Investigation of the Clem Motor

Robert Koontz, USA
rdkoontz@attbi.com

In December of 1992 Jerry Decker posted an article on the KeelyNet BBS, about a self-running motor that developed excess useable power. The information, gathered from newspaper and individual sources, gave an anecdotal account of the motor invented in 1972 by Richard Clem of Flower Mound, Texas. New information has since been added and can be found on http://www.keelynet.com/energy/clem1.htm.

Richard Clem worked with heavy machinery for the city of Dallas. He used asphalt-spraying equipment, which pumped liquid asphalt. He noticed the asphalt pump would continue to run for up to 30 minutes after the power was turned off. It was this discovery that led to the development of the motor. Modifications he made eventually resulted in a substantial 350 horsepower output from a 200-pound motor. Clem is said to have often driven a car, powered by this motor, up and down Central Expressway in Dallas. He claimed it didn't use any fuel, and only needed a change of oil every 150,000 miles.

The motor had only one moving part, a cone shaped rotor mounted vertically on a hollow shaft. Spiral channels cut into the cone wound around its length and feed into peripheral nozzles at its large end. When fluid flowed through the spiral channels it was ejected out the nozzles and caused the cone to spin. At a certain velocity, the rotating cone became independent of the starter pump and began to operate by itself. At an operating speed of 1800 to 2300 RPM the fluid heated up to 300 F, requiring a heat exchanger. Vegetable oil was used because at 300 F water boils and conventional engine oil breaks down. A 12-volt battery was the only other power source. Clem never applied for a patent because his motor design was derived from the asphalt pump that was already patented. Fifteen companies turned him down before a large coal company offered to back him and signed contracts to sell the motor. Soon after the deal was signed, Richard Clem died of a heart attack

The above account contains only what I considered to be relevant for analysis of the Clem motor. Visit http://www.keelynet.com/ for the original material.

The gear pumps, typically used for asphalt spraying, do not match the description of the pump used by the city of Dallas back in 1972. There should be public records showing what equipment manufacture the asphalt sprayer was purchased from. Since the asphalt pump was patented, I searched for a pump patent that met the following criteria:

1) Patent issued on or before 1972.
2) Delivered pressure equivalent to a positive displacement gear pump.
3) Cone shaped rotor with spiral channels.
4) Self-propelling action.
5) Capable of pumping a viscous fluid like asphalt.
6) Large heat transfer to pumped fluids.

The following illustration (Fig.1) is from US Patent 3,697,190 (Truncated Conical Drag Pump). The patent was issued October 10, 1972 (criteria 1) and appears to match the description of the asphalt pump that Clem converted into his motor.

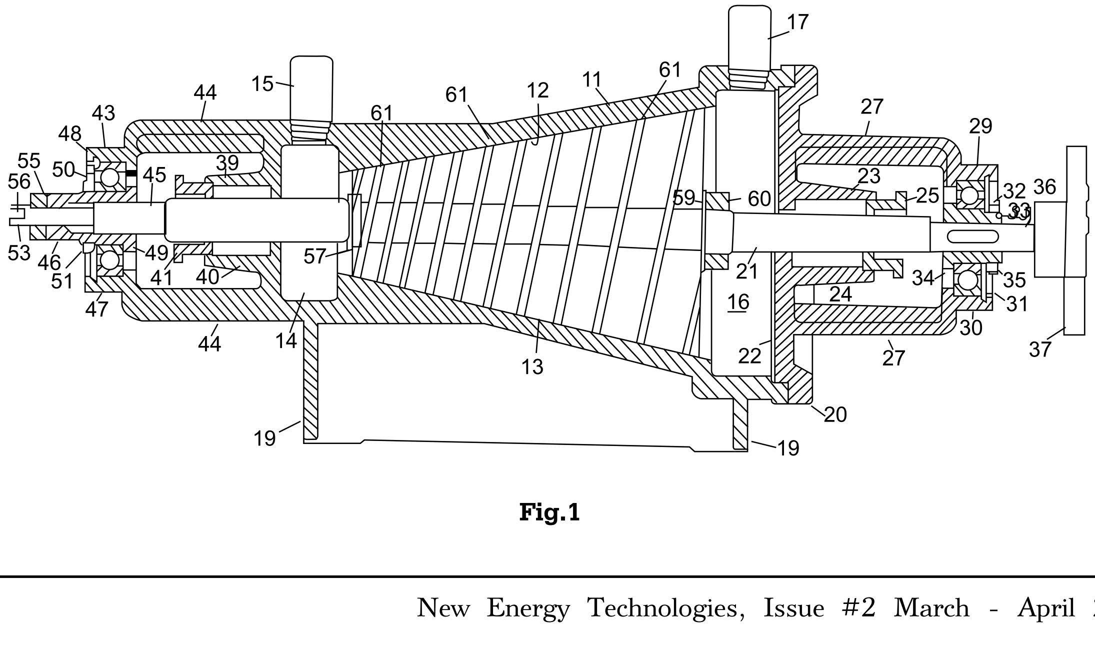

Fig.1

Housing 11, Conical interior wall 12, Conical rotor 13, Inlet chamber 14, Inlet pipe 15, Outlet chamber 16, Outlet pipe 17, Support feet 19, Detachable end cap 20, Rotor shaft 21, End cap wall 22, Boss 23, Packing 24, Adjustable gland nut 25, Bracket arms 27, Bearing boss 29, Bearing 30, Snap ring 31, Inner race 32, Sleeve 33, Shoulder 34, Retainer nut 35, Reduced diameter outer end 36, Coupling 37, Packing 39, Retainer 40, Gland nut 41, Bearing boss 43, Integrally formed bracket 44, Shaft reduced diameter 45, Bearing sleeve 46, Bearing 47, Snap ring 48, Inner flanged 49, Inner race 50, Nut 51,Shaft reduced diameter 53, Lock nut 55, Flat faces 56, Snap ring 57, Washer 59, Nut 60, Helical channel 61, Channel base 63, Channel sidewalls 64.

This is a high-pressure, low volume drag pump that can be used in place of conventional positive displacement pumps (criteria 2). It has a conical rotor that has a close fit clearance with the stationary housing wall. Delivered pressure is limited by back flow across the radial clearance and is inversely proportional to the square of the clearance. As a result, even a small increase in radial clearance would rapidly reduce pressure. The rotor is cone shaped so that the clearance can be controlled by axial adjustment of the rotor relative to the housing wall.

The conical rotor has two helical channels (criteria 3), in the form of square threads, spaced 180 apart for balance. The channel depth decreases as the rotor diameter increases. Fluid enters the channels at the small end of the rotor. The fluid is induced to rotate with the channel by boundary layer drag. The boundary layer is the thin layer of fluid adhering to the channel surface. Molecular cohesion tends to drag the adjacent fluid with the boundary layer. The fluid is also in contact with the housing wall. The boundary layer drag against this stationary wall slows the rotation of the fluid in the channels. Because the fluid rotates slower than the rotor, it is forced through the channels towards the large end of the rotor. In addition the fluid is forced towards the large end by centrifugal force.

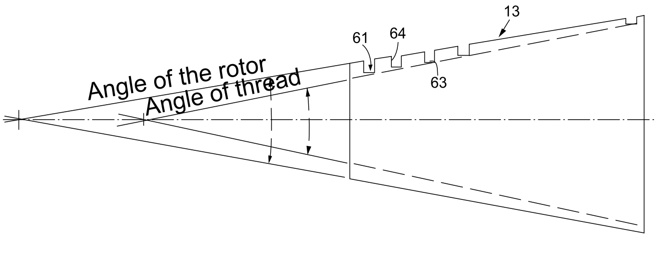

Fig.2

The above drawing illustrates the proportional decrease in channel depth as the rotor diameter increases. Why was this done? Note that as the diameter doubles so does the circumference. This means the fluid has to travel twice as far in the same time to maintain a constant slip velocity. By reducing the channel depth in half (cross-section area = depth x width) the fluid velocity is doubled thereby keeping the slip constant.

The spiral channels could be thought of as very long convergent nozzles. The increase in fluid velocity is in the opposite direction of the rotor spin. We should expect a reaction force from the acceleration of the fluid. This thrust would be directed tangent to the circumference and would increase the spin torque on the rotor. Even without the peripheral nozzles, that Clem later added, the pump rotor experiences a thrust force in a direction that would self-propel it (criteria 4).

Because fluid drag is the primary pumping force, it is well suited for viscous fluids like asphalt (criteria 5). The long channels also represent a large sliding surface area with frictional losses that would transfer heat to the pumped fluid (criteria 6).

All six of the patent search criteria have now been met. Of course this doesn't prove that it is the asphalt pump Richard Clem worked with.

A peculiar condition indicated by the patent is that as the velocity increases in the channels the pressure also increases. Bernoulli's Law requires the pressure to drop proportionally as the velocity increases. Assuming an ideal fluid without losses, when the channel depth is reduced in half, the cross section area is also half and this doubles the fluid velocity and the fluid pressure should drop in half. So what is going on here? There is a centrifugal component that would add to the fluid pressure.

My guess is it's too small to overcome the predicted pressure drop. Here is what I think may be going on. As the diameter and velocity increases the drag force propelling the fluid through the channel is proportionally greater. Energy is being added all along the length of the channel. Whatever the reason, if this high-velocity, high-pressure fluid is feed into tangent peripheral nozzles at the rotor large end, the energy will be converted to shaft horsepower.

The Clem motor is producing 350 shaft-horsepower and a large heat energy component. Where is this huge amount of energy coming from? Recent quantum mechanics zero-point field (ZPF) theories may point to the answer. From an article available at "BEYOND $E=mc^2$" Bernhard Haisch, Alfonso Rueda & H.E. Puthoff [1]:

"Our work suggests inertia is a property arising out of the vast, all-pervasive electromagnetic field we mentioned earlier, which is called the zero-point field (ZPF). The name comes from the fact that the field is held to exist in a vacuum-what is commonly thought of as "empty" space-even at the temperature of absolute zero, at which all thermal radiation is absent."

ZPF researchers theorize that mass, inertia and gravity are not intrinsic properties of matter but the interaction of matter with the zero-point field. By all pervasive it is meant that the ZPF exists not only in empty space but it is passing through your body right now and everywhere else. When you throw a stone you are interacting with this field since the ZPF resists change in motion. In essence the ZPF is the modern day aether.

The amount of energy making up the ZPF is thought to be enormous. Is the fluid acceleration in the Clem motor interacting with the ZPF in such a way as to rectify it and draw energy from it? Is it a hydraulic aether-diode? The fluid, in the Conical Drag Pump, flows through long convergent channels. Disregarding the boundary layer, is this accelerated flow laminar? Would such a long orderly flow entrain the aether energy?

From the perspective of the rotating channels the fluid appears as the discharge from a long nozzle. To exaggerate, if the fluid was held fast to the housing wall, the rotating channel would travel through the stationary fluid. This would be equivalent to achieving 100% efficiency. In reality the fluid is slipping against the stationary housing wall so that the rotating channel (nozzle) is moving faster than the fluid discharge velocity. Assuming the reaction thrust as the only propelling force, this would give efficiency greater than 100%. So, as the slip increases the reaction thrust decreases, but the efficiency increases.

Assuming the Conical Drag Pump is the pump Clem used, can it answer the following?

1) Why was a hollow shaft used?
2) Why was the cone mounted vertically?
3) Why was a starter pump needed?
4) How were the peripheral nozzles added?
5) How was the motor RPM regulated?
6) How did a large coal company get involved?
7) Was this kind of pump ever used in asphalt sprayers?

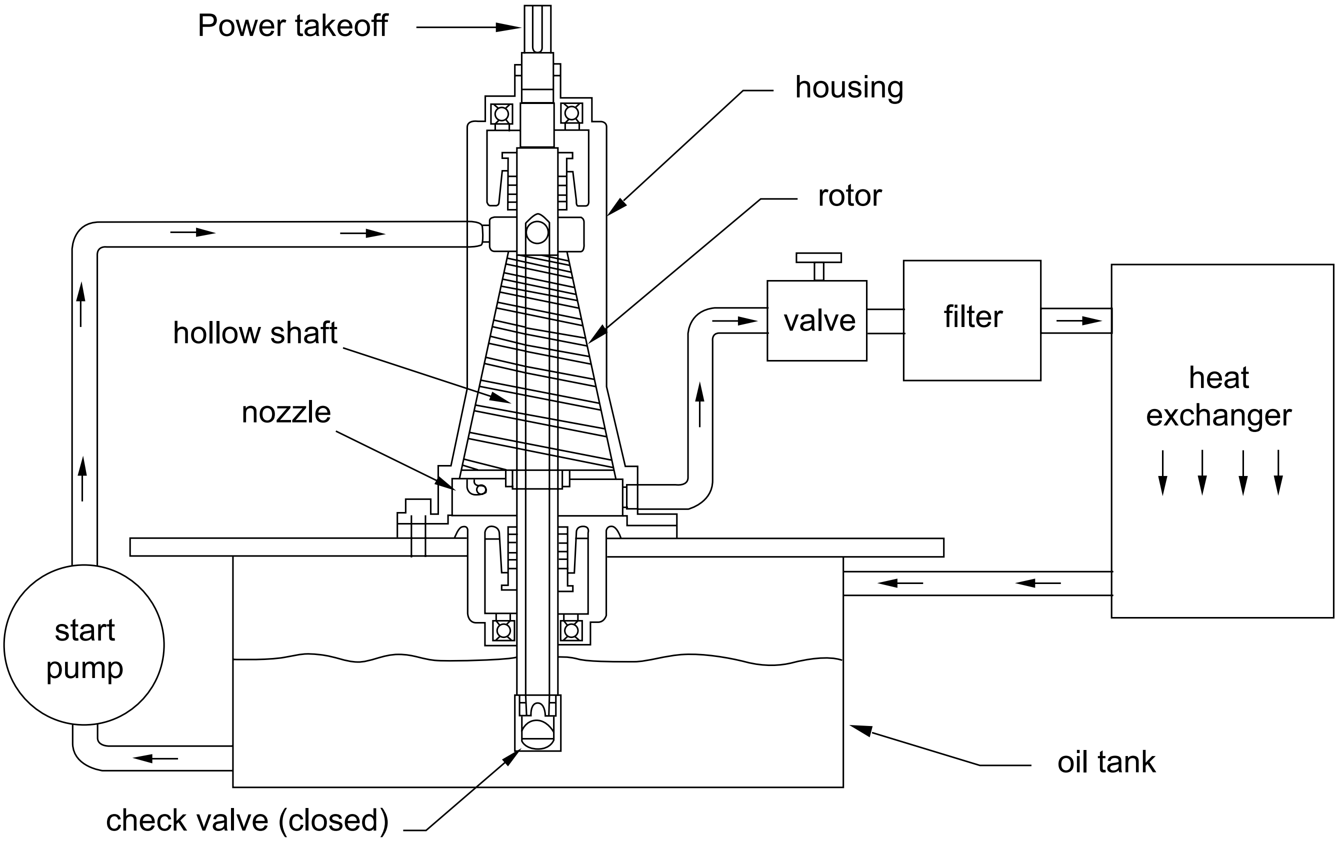

Fig.3

Clem motor shown in start mode
The arrows show oil flow

The above drawing shows a hypothetical Clem motor based on the Conical Drag Pump. The motor is mounted vertically so that the check valve on the hollow shaft is submerged down in the oil tank. The hollow shaft extends from the oil tank through the rotor into the inlet chamber. The start pump draws oil from the tank and forces it up the external feed line connected to the inlet chamber at the small end of the rotor. This fills the hollow shaft and forces the check valve closed. The oil flows into the spiral channels and out the peripheral nozzles. The reaction thrust of the nozzles spins the rotor. The oil flows through the return line, through the valve, filter, and heat exchanger and back into the tank. The start pump is most likely a standard hydraulic gear pump. It continues to pump until the rotor spins up to its operating speed. The combinations of a start pump and check valve would be a simple way to both prime the motor and spin up the rotor.

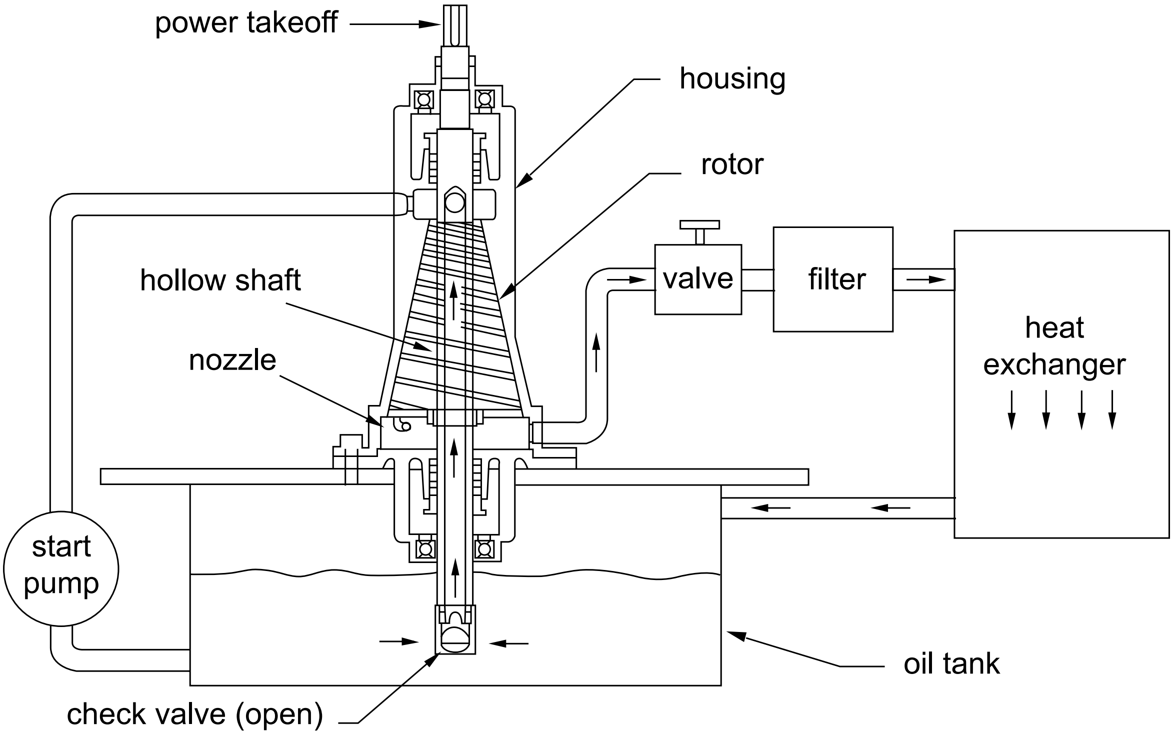

Fig.4

Clem motor shown in operation mode
The arrows show oil flow

Once the start pump is shut off the check valve is free to open. Oil is drawn up into the hollow shaft to the inlet chamber at the small end of the rotor. The spiral channels pump the oil down towards the large end of the rotor. A plate is attached to the large end of the rotor and fits with a close clearance with the housing wall. Nozzles attached to the outer edge receive high-pressure oil from the spiral channels. The jet reaction thrust from the nozzles delivers shaft horsepower to the power takeoff at the shaft top. Adjusting the valve to create hydraulic backpressure regulates the motor RPM. Closing the valve stops the motor.

When I first read about the Clem Motor I found it odd that a deal had been made with a coal company. Was there a connection with the pump? After finding the Conical Drag Pump patent, I wanted to contact the inventor Walter D. Haentjens of Barrett, Haentjens & Co., Hazleton, Pennsylvania. Otto Haentjens founded Barrett Haentjens & Co., in 1916. The business began in the coalmines of Pennsylvania with Otto Haentjens original patent on the balanced opposed impeller multi-stage volute pump. The company still supplies pumps to the coal industry. They have expanded to other markets and their pumps are installed in many industries worldwide. It is now known as Hazleton Pumps Inc., after its acquisition by The Weir Group.

I contacted Peter Haentjens, the VP/General Manager of Hazleton Pumps, by e-mail to find out if this pump had ever been put into production. He replied that they had not done anything with the patent.

An unusual pump design would have a tough time competing in the market with an industry standard like gear pumps. The Dallas asphalt sprayer may have been a one of a kind field test of the pump design. Or the pump manufacturer offered it for testing to an asphalt equipment company in the hopes of generating interest in it.

References

Bernhard Haisch, Alfonso Rueda & H.E. Puthoff, "BEYOND $E=mc^2$" //The Sciences, Vol. 34, No. 6, November / December 1994, pp. 26-31 copyright 1994, New York Academy of Sciences.

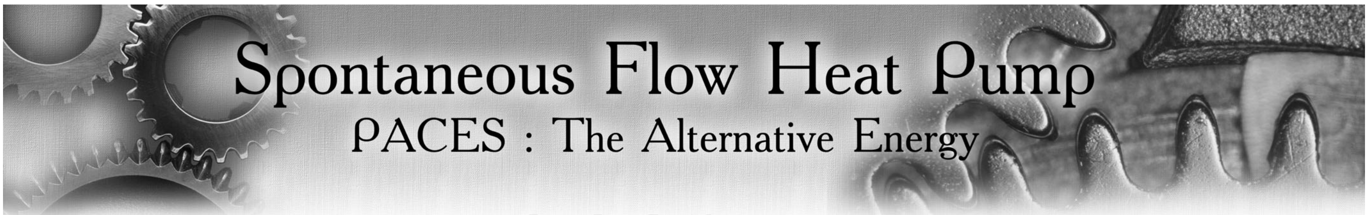

Spontaneous Flow Heat Pump

PACES : The Alternative Energy

Jean-Luc Brochet, France

http://www.new-energy-paces.com
Email: brochet.jean-luc@wanadoo.fr

"Paces" is a new high performance electronic process. It is not a traditional heat pump fitted with some electronic gadgets to improve its output, but a truly revolutionary process totally unused up to now. There are no moving parts, not even a moving fluid in it. "Paces" is an electronic process because for its development the pump's fundamental element uses micro-electronic techniques.

...an average of 9000MW could be produced, that is to say the equivalent of several nuclear power stations.

The principle consists in attracting the gas neutral molecules onto a so-called "electrostatic" plate, to accelerate them and heat the plate. An intense electric field is applied in a capacitor with micro-pores covering the plate. The neutral molecules are polarized and accelerated when they penetrate these micro-pores where the intense electric field exists. Each molecule accelerates and absorbs electrical energy when it enters into a pore. But in turn it decelerates and gives off an equal amount of energy upon its exit from the pore.

The molecules, by going back and forth between the plates, continually give off and take energy, thus resulting in a total consumption equal to zero.

The gas fluid does not undergo any change of thermodynamic state, contrary to all other heat pump processes. A thermodynamic cycle remains at the level of each gas molecule, but not at the level of the entire set of molecules.

There is more to this new process, it is not just heat pump process added to an already long list: **these new characteristics also enable some new applications.** Besides the traditional heat pump applications, due to its excellent coefficient of performance this process makes it possible **to produce electricity by cooling the environment.**

For instance, if the Rhine was equipped with heat exchangers into which diverted water could flow, thus cooling its temperature by just one degree, an average of 9000MW could be produced, that is to say the equivalent of several nuclear power stations. Please note that this process would also enable **both electricity and fresh water to be produced from seawater,** an interesting application for areas of the world where water is a scarce resource.

Thus we have at our disposal an ecological process for producing electricity, it would become conceivable to produce hydrogen in great quantities; vehicles would be hydrogen powered. Ozone, CO2 or any other type of pollution would become a thing of the past.

It will most probably be the energy source of future generations, ecological and excellent, inexhaustible, powerful yet completely harmless, stable and steady (contrary to the sun and wind); an invention capable of re-launching the world economy, in particular giving impetus to the currently depressed new technology sector.

However, you have probably noticed that this invention requires advanced technology, and the few French laboratories capable to develop this process have proven inaccessibility to do it for independent researchers such as the author of the article. Besides, the invention's claimed qualities, the revolutionary aspect was very often interpreted by examiners as 'utopian', even before they read the document! Having tired of battling against concept fixed for the past two centuries, the author of the article **has now decided to place this project within the hands of the international scientific and technological community,** in order to give the best chance of success to the invention.

It will most probably be the energy source of future generations!

Initial development requires the construction of a prototype; three possibilities are described on the website (http://www.new-energy-paces.com). As there is no longer legal protection abroad, anyone is free to experiment and to market, as he or she wishes, in all other countries outside of France.

Please help me to spread the word about this project, for example through scientific forums on the Internet. This will allow you to receive feedback and gather other opinions. Your participation, even the most modest, will be of use. To those of you who help this project to see the light of day, all of humanity will be indebted.

Registration of Gravitational Waves

Vladimir N. Petrov, Russia

173001, apt. 76, 10 Bolshaya Saint-Peterburgskaya st., Veliky Novgorod
E-mail: petvn@km.ru

Registration of gravitational waves is one of fundamental problems of physics. At the beginning of the last century A. Einstein's works predicted existence of these waves. Nevertheless, despite numerous attempts are made in different countries of the world these waves have not been fixed up today. Josef Veber, an American physicist working in 60th of the XX century, is sure to be called as the first experimenter who disclosed gravitational waves. Nowadays lots of many-million projects on registration of gravitation waves are developed and realized (LIGO (USA), GEO-600 (Germany-England), VIRGO (Italia), and TAMA (Japan)). However no positive results were achieved.

Failures in registration of gravitational waves can be explained by the fact that researchers have erroneous notion of the essence of these waves. This erroneous notion is laid in an experimental project that causes negative results.

Many researchers believe that influence of gravitational waves on the bodies should be resulted in deformation of these bodies (i.e. in change of mutual place of particles of a material body). All the attempts to disclose the gravitational waves are based on this notion. Now two types of ground gravitational antennas are supposed as priority.

The first type supposes the registration of mechanical oscillations of a massive test body that is initiated by influence of gravitational waves on it.

The second can register changes of a distance between space-apart free masses when gravitational waves act upon them. Great attention is paid to the second type of antennas in the developed projects. Failure attempts of registration of gravitational waves can be explained by their low intensiveness and insufficient accuracy of existing measuring equipments. Therefore, now there are made various attempts to register gravitational waves produced by such great cosmic phenomena as confluence of black holes. More accurate instruments and measuring equipments, for example, a laser interferometer, are developed. Nevertheless, positive results are hardly to be obtained.

Here is one of the fallacious explanations of action of gravitational waves on a body: *"Passing of gravitational waves changes an interval between objects transferring them one from another like two boats sometimes are approached sometimes moved from each other by sea waves".*

Following this example to explain influence of the gravitational waves on bodies placing on the Earth, it can be said that all bodies are in one boat and raising or sinking of the boat on sea waves does not cause change of the distance between the bodies. However, this example seems to be very simplified for explanation of influence of the gravitational waves on bodies. In fact, the case is much more difficult.

According to A. Einstein's theory, bodies moving with variable acceleration will radiate gravitational waves. **Gravitational waves represent the essence of changes of gravitational field occurring at light speed.** Since the gravitational field is a deformed space-time then the gravitational waves cause change of deformation of the space-time. Change of deformation of the space-time is accompanied with change of deformation of geodesic lines (i.e. lines by which all the bodies move).

A body moving by a geodesic line whose deformation is changed changes its acceleration. Change of acceleration of the body will change gravity applied to the body in this area of space. Hence, influence of gravitational waves on the body occurs through gravity which influences on it.

It is known that a fundamental difference between gravity and other forces appearing at a direct contact with a body consists in the fact that gravity provides all the elements of the body (all its points) with equal acceleration (deformation can not occur), and other forces influence on certain parts of the surface of a body and so cause its deformation.

Hence it is senseless to try to measure deformation of a body or change of a distance between test bodies occurring under influence of gravitational waves. It is necessary to note that, in particular, when a body is placed on an unmovable support relatively to the Earth, value and the direction of gravity coincides with weight of the body. Therefore, in this case gravitational waves can be registered at measuring of changes of the body weight.

An experiment on registration of the gravitational field of the Earth by measuring change of weight of a test body was made. Before the beginning of the experiment there was set a problem to register gravitational waves produced by the Mercury and the Venus at approaching to the Earth. The gravitational field in which the Earth moves changes due to superposition of gravitational waves of other planets on it. These changes were necessary to be registered.

The experiment was as follows: a test body (a 100 grams weight) was weighted on an electronic balance in equal time intervals of 14 days. The experiment was lasting for 12 months. In the experiment it appeared that change of a distance between the Earth and the mentioned planets causes change of weight of the test body. The change of weight of the body occurred in proportion to the distance between the planets. The test body was weighted in the same conditions and in the same place. The performance of the experiment is so easy that every person can made it.

It is known that weight of a body changes (decreases) at moving along the Earth surface from its pole to the equator. Change of weight of a body locating in the same place has not been fixed until now.

The fixed change of weight of the test body proves change of gravity applied to it and, consequently, change of the gravitational field in which the test body moves as well as the Earth does. **These changes of the gravitational field were caused by motion (superposition) of the gravitational waves produced by the mentioned planets at their approaching towards the Earth at the distance of the existent gravitational field.**

Hence, the Earth and all the bodies locating on it move in the constantly changing gravitational field, therefore, constantly change their weight. Nevertheless, as it is known, change of the gravitational field is the influence of gravitational waves on it. Hence, measuring changing of weight of bodies we can speak about registration of gravitational waves.

In search of the 4th dimension

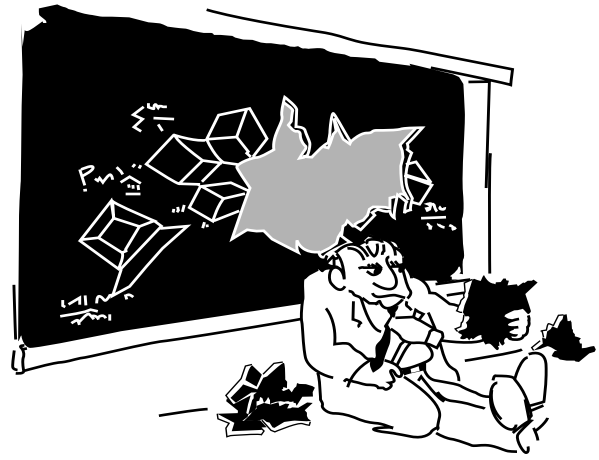

From http://home.ural.ru/~meop/index.html

High-Energy Hyper-Low-Frequency Electric Field

Yury V. Ivanko, Ukraine
E-mail: ux01w@hotmail.com

The article represents a practical and theoretical research on possibility of existence of a high-energy hyper-low-frequency electric field which is interpreted by modern measuring instruments and subjective perception as a static field. Besides there are observed a problem of the field registration, a hypothesis of relic origin of the registered field and a hypothesis of wave nature of the Universe.

A passenger on board a ship sailing in the ocean will never notice action of waves of high and low tide. At the same time, weathering waves of several numbers produce a dangerous storm. However, an ocean liner is designed for such a storm therefore the passenger would just experience some discomfort. It is another case if tide wave hides reefs… I wish our ship, i.e. the Earth, the Solar System, the Galaxy, and the Universe, not to be lead to such a place by any ignorant "navigator"!

One of the most mysterious persons in the history of electricity physics was an outstanding inventor Nicola Tesla (1856-1943). When the majority of scientists developed researches of microworld particles he followed the opposite scientific path. He was interested in the electrical potential of the whole Earth. He researched ways to influence on it, control its state and methods of its regulation. Therefore many of his patents, experiments as well as a purpose of constructions and devices built according to his ideas arouse perplexity and misunderstanding of modern scientists.

a test body was made. Before the beginning of the experiment there was set a problem to register gravitational waves produced by the Mercury and the Venus at approaching to the Earth. The gravitational field in which the Earth moves changes due to superposition of gravitational waves of other planets on it. These changes were necessary to be registered.

The experiment was as follows: a test body (a 100 grams weight) was weighted on an electronic balance in equal time intervals of 14 days. The experiment was lasting for 12 months. In the experiment it appeared that change of a distance between the Earth and the mentioned planets causes change of weight of the test body. The change of weight of the body occurred in proportion to the distance between the planets. The test body was weighted in the same conditions and in the same place. The performance of the experiment is so easy that every person can made it.

It is known that weight of a body changes (decreases) at moving along the Earth surface from its pole to the equator. Change of weight of a body locating in the same place has not been fixed until now.

The fixed change of weight of the test body proves change of gravity applied to it and, consequently, change of the gravitational field in which the test body moves as well as the Earth does. **These changes of the gravitational field were caused by motion (superposition) of the gravitational waves produced by the mentioned planets at their approaching towards the Earth at the distance of the existent gravitational field.**

Hence, the Earth and all the bodies locating on it move in the constantly changing gravitational field, therefore, constantly change their weight. Nevertheless, as it is known, change of the gravitational field is the influence of gravitational waves on it. Hence, measuring changing of weight of bodies we can speak about registration of gravitational waves.

In search of the 4th dimension

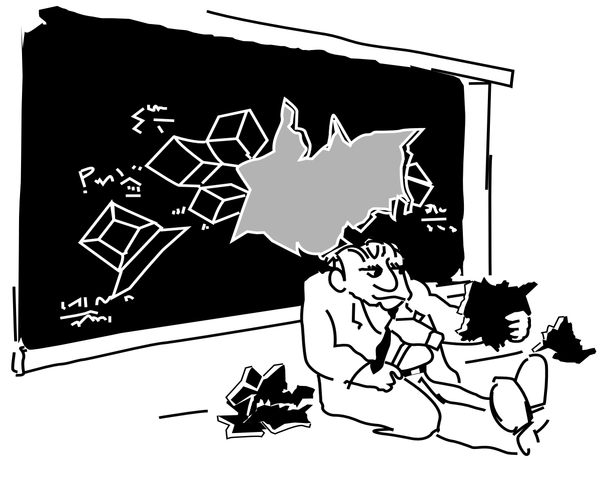

From http://home.ural.ru/~meop/index.html

High-Energy Hyper-Low-Frequency Electric Field

Yury V. Ivanko, Ukraine
E-mail: ux01w@hotmail.com

The article represents a practical and theoretical research on possibility of existence of a high-energy hyper-low-frequency electric field which is interpreted by modern measuring instruments and subjective perception as a static field. Besides there are observed a problem of the field registration, a hypothesis of relic origin of the registered field and a hypothesis of wave nature of the Universe.

A passenger on board a ship sailing in the ocean will never notice action of waves of high and low tide. At the same time, weathering waves of several numbers produce a dangerous storm. However, an ocean liner is designed for such a storm therefore the passenger would just experience some discomfort. It is another case if tide wave hides reefs… I wish our ship, i.e. the Earth, the Solar System, the Galaxy, and the Universe, not to be lead to such a place by any ignorant "navigator"!

One of the most mysterious persons in the history of electricity physics was an outstanding inventor Nicola Tesla (1856-1943). When the majority of scientists developed researches of microworld particles he followed the opposite scientific path. He was interested in the electrical potential of the whole Earth. He researched ways to influence on it, control its state and methods of its regulation. Therefore many of his patents, experiments as well as a purpose of constructions and devices built according to his ideas arouse perplexity and misunderstanding of modern scientists.

Having formulated my own hypotheses I often think why such a Great Genius as Nicola Tesla took his invention away with him.

Danger of this invention lies in simplicity of production of devices for generating (detecting) energy.

Practical Prerequisites

Prerequisites of this and other hypotheses were some undeclared effects occurring in practical wireless communication as well as investigation of devices for distant transmission of electric power.

Distant energy transmission was demonstrated by Nicola Tesla in 1892 in London and in 1893 in Philadelphia. However, even now the way he did it remains unknown.

In 1977-1985 I had to work a lot with radio transmitting equipment of different power of middle-wave (MW) and short-wave (SW) diapasons. According to fundamentals of physics, intensive electromagnetic field is produced around antenna curtain connected to a radio-frequency transmitter. There are brightly glowing gas-discharge electrovacuum devices at the distance up to several meters in this field. Attention was paid to the fact that more than ten daylight lamps of 80 W were successfully "glowed" at 40 W power of a transmitter. In this case a distant correspondent did not note significant decrease of a level of a received signal at the indicator of the receiver. It had to return to this fact in many years. The data of this experiment are represented in a particular document.

There is one more interesting effect which is disclosed when transmitting equipment (TX) and antennas (ANT) are adjusted. An incandescent lamp, which is connected to a certain place of a break of the antenna curtain or of a power feeder (L), glows brightly. The antenna circuit is not closed as galvanic element. (See Fig. 1).

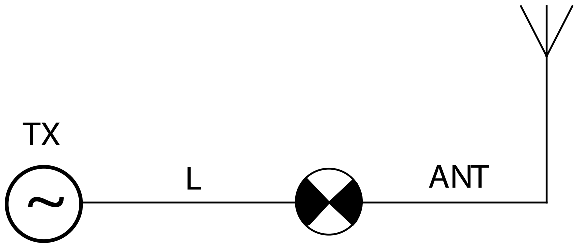

Fig. 1

Hence, the power line of the incandescent lamp is as if it is single-wire. It also had to return to investigation of this effect in many years. The data of this experiment are represented in a particular document. As in the case of the incandescent lamps, the correspondent did not note significant decrease of the level of the received signal at the indicator of the distant receiver.

S. V. Avramenko designed a plug of semiconductor diodes. Connecting this plug to a generator of alternating voltage of 10-10000 V by means of a single-wire line in the load of the plug he discovered current, which was continuous by the direction but pulsating by the value. Electric current measured in the single-wire line is very little, hence, the impression of superconductivity is produced, i.e. it is possible almost to avoid energy losses in the wires.

My experiments also represent a visual demonstration of the effect. For example, several links of light-emitting diodes connected inverse-parallel glow in the single-wire line. Garlands of light-emitting diodes and lamps having a single-wire power line can be produced. It is interesting that in this case power consumed by the generator do not increase practically (See Fig. 2, 3, 4). It was noted that if a standing wave is created around the generator "then an unlimited number of consumers would be able to apply a change of a value of this field to produce power in a load in their place" *(Editor: This phrase is in inverted commas since it is a part of the well-known article "Free Energy" by Alexander V. Frolov (Read in this issue, p. 15)).*

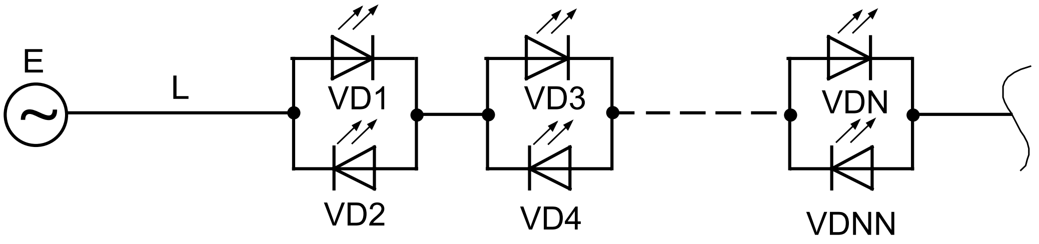

Fig. 2

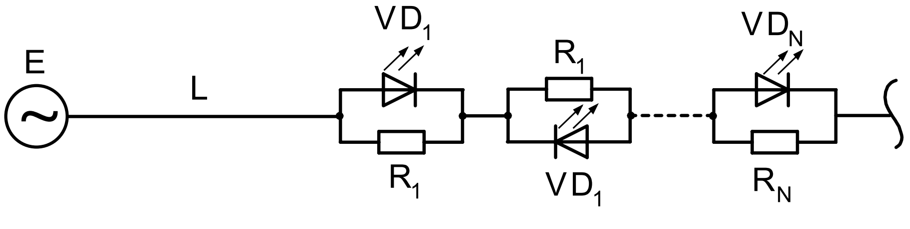

Fig. 3

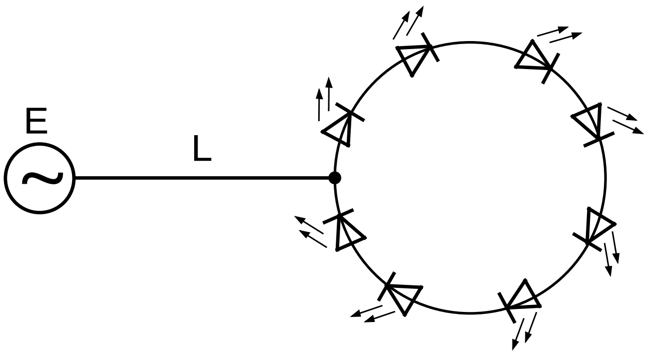

Fig. 4

Russian scientists investigated this phenomenon and proposed an explanation of producing active power in the load by means of reactive capacitive current. This was made applying resonant properties of the single-wire line made of a metallic conductor. Further, possibilities of application of nonmetallic conducting mediums for transmitting electric power were investigated. However, the researchers pay attention for the fact that this explanation is insufficient.

A task group named "Analysis" from Voronezh city (Russia) made a qualitatively new explanation of results of S. V. Avramenko's experiment. The works of this group from Voronezh are devoted to the comparison of properties and characteristics of so-called inertial and inertialess currents occurring in circuits at low frequencies. Their general and particular features have been disclosed. In the works there are demonstrated conditions at which conductivity caused by inertialess discharges predominates in conductors and conditions at which currents caused by conduction electrons predominate. There are demonstrated a description of S. V. Avramenko's experiments and the conclusions made by the researchers. The authors cite data of their test experiments which prove a part of conclusions made by S. V. Avramenko and his colleagues.

However this explanation does not have a right to exist as it does not have variants for practical application.

Results of my theoretical and experimental researches allow provide a **fundamentally new (!)** explanation of energy "transmission" having occurred in the experiments performed by S. V. Avramenko, investigations by D. S. Strebkov, V.A. Kuligin and other investigators. *(Editor: See publications of these works in previous issues of our magazine*). Moreover, they allow explain origin of many physical phenomena and processes including the origin of non-inertial charges and currents and their visualization.

First Hypothesis

Recently a number of scientists all over the world have declared a thought that surplus power is produced in the systems having overunit output due to inability to identify correctly all the existent energy sources. And a so-called phenomenology is observed due to the inability to deal with different types of energy (or fields).

In physics a negation of an overunit device follows a well known state which is considered to be true in other fields of knowledge, i.e. energy conservation law. It should be noted that it is formulated for isolated (closed) systems. I have devoted some time to the investigation of this state and to practical experiments and come to the following conclusion: there are **no isolated (closed)** systems in the nearest cosmos which can be observed by us!

Particular physical, chemical and other processes may actually be observed in closed circuits for making calculations with sufficient approximation.

Show me an isolated system, and I will prove that it is open.

Hence, according to fundamental physical laws, energy does not appear from nothing and not disappear to nowhere. However, the source of oscillations, i.e. our generator, does not provide a load with energy; nevertheless, power is produced in the load. Energy conservation law must be fulfilled! The classical formulation of the energy conservation law is as following: total energy of an isolated system is not changed in the course of time. According to this, the only fact remains valid, i.e. the single-wire system is not isolated from outer influence, hence, energy is consumed from some other source which is outer in relation to the system?! In this case our generator is one of sources of information about an amplitude, frequency, and phase.

Calculations and experiments have allowed made a following supposition. Avramenko's diode plug is a particular case of a device known in radio engineering for a long time, i.e. AM detector (a peak-detector, a limiter, a frequency mixer). I have examined the other known circuits of AM detectors and frequency mixers. The investigations are continued and by this time the supposition has been confirmed.

Let us accept an artificial generator (Tesla's transformer or any other one) as a main generator (or a main oscillator) and a hyper-low-frequency (HLF) oscillation of a huge amplitude as an envelope curve. Let us name it as RW ("relic wave") without placing emphasis on the sense of the name.

Let us suppose that the field can be of:
- local origin (artificial radio transmitters);
- atmospheric origin;
- geomagnetic Earth origin;
- the Moon – the Earth;
- the Sun – the Earth (the Sun as a nuclear reactor producing electromagnetic radiation, the Solar System);
- galactic origin;
- Universal origin – the Relic Wave; in contrast to the relic radiation occurring in the diapason of SHF (i.e. in

contrast to the convenient notion) it is the diapason of HLF.
- Etc.

The researches have demonstrated that the field of local and atmospheric origin may be excluded. The other variants require to be developed. Nevertheless, the variant of the Universal scale seemed to be the most probable. There is some reason for this supposition: the modern notion of the birth and the state of the Universe, i.e. the Big Bang and the expanding Universe. The known variants of scenarios of the Big Bang are based on a nuclear explosion of enormous power. The nuclear explosion is accompanied by an electromagnetic impulse (disturbance). It can be supposed that the Big Bang caused forming of a field having enormous amplitude values and long (great) period of oscillations. Probably, we deal with this field.

The second hypothesis about inconstancy, i.e. about wave nature of the Universe

A Russian proverb says: "Everything passes, everything changes". Constant electric voltage does not exist (!). That, what is considered as the constant voltage, really is a constant component (occurring at a certain time interval). This component is formed by combining of variable sinusoidal oscillations and/or is a quite low frequency fundamental. Hence, it is a non-sinusoidal "changing" having amplitude linear part at the time interval of the observer.

For example, in reality a usual storage battery of "direct current" can not support invariable voltage, i.e. when a discharge appears the voltage decreases (changes in time, oscillates with a non-sinusoidal oscillation and a quite low frequency fundamental), when a charge appears the voltage increases.

It is necessary to observe the constant voltage occurring after a rectifier and even a stabilizer as a constant component existing in the time interval from switching on to switching off. This will correspond to that is proposed by Fourier!

Show me a device producing constant voltage, and I will prove that the voltage is not constant.

Conclusions

If the supposition is totally confirmed then the statement that an electric field plays a fundamental role in the Universe will be valid.

We live in an electric field of a huge potential but we do not notice it. Since **our reference point is the amplitude value of RW field in our point of space** then the potential is equal to zero for our perception and for measuring instruments. Besides, **RW has an amplitude linear part at the time interval of the observer**. Hence, we consider RW field existing as unrevealed or as weak interactions in natural processes. Nevertheless, this influences directly on the whole surrounding material world.

Artificial production of heterogeneities in the RW field can cause its significant visualization.

According to the concept of the physical field, which was called by Einstein as the most important discovery in physics since Newton's time, every body generates a force field in the ambient space (for example, the electric field occurring if the body has an electric charge). This force field becomes apparent when a test body located in a certain point is influenced by a force.

Discovery of the fields of this scale allows define this concept more accurately. I would like to restate the information mentioned above in the following way: every non-homogeneity of physical vacuum (aether), which is registered by our perception or by measuring instruments as a physical object or a physical field, interferes with RW (according to other interpretations, it absorbs or rereflects RW, or becomes excited) producing new wave fields. This perceived physical field has frequency which is higher than RW frequencies. It is perceived as static in the time interval of observation. Hence, the field generates a charge and not the reverse. Any linear or non-linear transmission (motion) of a material or field object changes non-homogeneity which is generated by this object.

A question about measuring of parameters of the Universe remains open. Can an oscillation of the period of T=14 billions years and of the amplitude of A~T be seen, perceived, registered or measured?

The used literature and references are not cited in the article due to its great volume. The reader can familiarize with them at the web-site of http://www.efir.com.ua. The complete text of the article and other materials can be received from the author.

EAGLE-RESEARCH

Energy Solutions
since 1984

A Research Organization that Develops & Distributes Practical Energy-Saving Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

To the Question of Work Made by Electrostatic Field

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./Fax.: 7-812-380-3844
Email: office@faraday.ru

We know some simple examples of work making due to the forces, which appear in electrostatic field. Since the field itself does not require energy consumption from the source (not taking into account the charge leakage), then free energy can be received with these forces. The first phenomenon that we are going to consider is a phenomenon of **electrostriction**. Mechanical forces appearing in dielectric under the influence of electrical field tend to deform it. Sometimes these forces are small, but they can be huge in special dielectrics. The character of deformation also depends on non-linearity of electrical field. Suffice it to remember that heating of a body takes place during deformation, and we can make a conclusion on possible using of the given effect in practice. Periodical deformation can be created due to the rotor or any other methods to change the field.

The second effect is not so popular. It was found by Hertz in 1881 that there is a rotation of dielectrical cylinder (or ball) in permanent electric field if it is created in some liquid or in gas (see Fig.1).

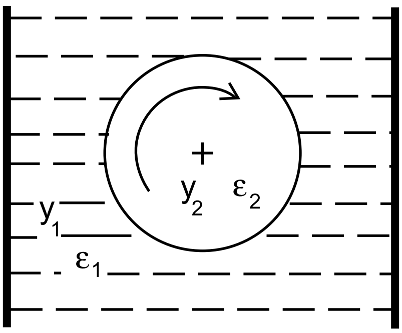

Fig.1

Here $\varepsilon1$ is a dielectrical permittivity of the liquid, $\varepsilon2$ is a dielectrical permittivity of the rotor, $\gamma1$ is a conductivity of the liquid and $\gamma2$ is a conductivity of the rotor. This effect was discovered by Hertz, then it was described by G. Quincke, Germany. Also Japan scientist I. Sumoto studied the effect in 1955. Modern research work was made by K.M. Polivanov, Moscow. There is the so called equation by Polivanov describing special conditions, which are necessary to begin rotation ($\varepsilon1/\ \varepsilon2=\gamma2/\gamma1$ is the Polivanov's condition). Forces of electrostatic field produce this rotation and it is a real free energy system that can trap energy from inner structure of potential field. Let's try to describe the mechanism of this effect and then we'll be able to increase the power of the system up to the level of industrial application. So, why does it work? To my mind, there are some differences in conditions for polarization of the rotor and molecules of liquid or gas those are surrounding the rotor since there is a difference in permittivity and conductivity.

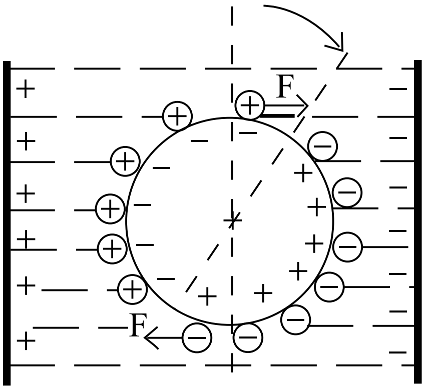

Fig.2

Due to this reason the molecules are polarized both by the field of electrodes and by the field of rotor. The rotor is polarized in the electrical field of electrodes, and molecules of liquid on its surface are polarized by the electric charge of the rotor, but not according to direction of the field. So, the rotor is surrounded by "screen" of molecules on its surface. The electric charge of this screen partially compensates the field of electrodes.

When some angle is created by the first initial turn, there is some part of the rotor surface, where molecules of liquid on the surface of the rotor are attracted to electrodes and it is the reason for future rotation. After some angle of the turn the polarization in this point of the rotor surface is changed but new molecules are incoming in the so called "sector of attraction" and the rotation is always accelerated. This well-known effect of 1881 is a very good example of possibility to produce useful work in load by means of electrostatic field only.

One more interesting effect is known as Faraday effect. In 1836 Faraday noted that raising of liquid upwards takes place in condenser with liquid dielectric (plates are installed vertically). During this process electrostatic field makes the work against gravity. A. Gyemant in 1926 established that value of this force is proportional to the square of voltage on condenser. Later in 1955 this phenomenon was studied in details by I. Sumoto, Japan. That's why in physical encyclopedias this effect is called "Faraday-Sumoto effect". In Tareev's book [1] it is pointed: "Under high voltage this phenomenon leads to flowing and turbulent boiling of liquid". I should remember for those, who see nothing unusual here, that it does not necessary to consume the power from initial source. Creating an initial field in electric capacitor, then we can use the received mechanical work. It is evident that the gradient of electrical field creates the conditions for movement of liquid dielectric. A task to create conditions for liquid circulation is more complex, but it has solution, if we take into account the presence of one more static field, i.e. gravitational field. Joint action of electrostatic and gravitational fields stipulates the circulation of liquid dielectric even in the simplest constructions. Besides, electrostatic filed can be partially screened.

The analyzed examples should draw attention of experimenters to these obvious ways of how to create useful work by means of electrostatic field.

References

1. Tareev B.M. Physics of dielectrical materials, Moscow, Energoisdat, 1982, p.199 – 200.

Editor: Below we publish the comments of our reader.

Letter On
Hertz-Quincke-Sumoto Effect

from Doug Marett
Email: doug.marett@sympatico.ca

…As far as I can tell, the original reference by Hertz is "On the Distribution of Electricity over the Surface of Moving Conductors", Wiedemann's Annalen, 13, pp.266-275,1881. Hertz's discussion of this phenomenon is cursory at best, having very little experimental work included. The real phenomenon appears to have been discovered by W. Weiler in 1893 (Zeitschrift fur den Physikalischen und Chemischen Unterricht, Vol.6, pp194-195). Weiler observed that a glass cylinder placed in a poorly conducting liquid between two spherical electrodes began to rotate when the electrodes were connected to an electrostatic generator. In 1896, George Quincke reported the same phenomenon and published a comprehensive report on it. This was in Annalen der Physik, Ser.3, Vol. 59, pp.417-486. Subsequent investigators have tended to attribute the discovery to Quincke, when in fact Weiler was the first to pioneer the work.

I do know that P.E. Secker et al. performed work on this field, references of which are available in the English language. These are:

P.E. Secker, et al., (1968) Journal of Applied Physics, Vol. 39, pp.2957-2961, and
P.E. Secker, et al., (1970) Journal of Physics D: Applied Physics Vol.3, pp 216-220.

I hope that this information might be of use to you.

BOOKS REVIEW

WWW.FARADAY.R
WWW.FARADAY.RU

The Foundations of Physchemistry of Micro World

Ph. M. Kanarev

The Second Edition

The book was published in Russian and English; you can also read it at http://book.Kanarev.innoplaza.net

The new axiomatic of natural sciences is given in the book; on its basis, quantum physics and quantum chemistry have been returned to the classical way of development. The first steps are made on this way, which have led to discovery of the structure of the photon, the electron, and the principle of the formation of the atomic nuclei, the atoms and the molecules.

The Planck's Law of radiation of perfect blackbody is given on the basis of classical concepts, and the connection of quantum phenomena with the laws of classical physics is proved.

The application of the new theoretical results to the solution of practical energy tasks on the basis of plasma electrolysis of water is shown. Due to this electrolysis, additional heat energy and the energy containing gases (hydrogen and oxygen) are generated. Cold Nuclear Transmutation of the atomic nuclei of alkaline metals and the atomic nuclei of the cathode material takes place during plasma electrolysis of water.

The book is intended for physicists, chemists and other specialists who are seeking the new directions for understanding the foundations of the micro world and the new energy sources.

Comments to Electro-Reactive LIFTER News

Alexander V. Frolov, Russia

General Director, Faraday Laboratories Ltd
E-mail: office@faraday.ru
http://www.faraday.ru

March 21, 2003 we got news from France about electro-reactive propulsion system.

Blaze Labs reached 100g PAYLOAD level.

You will find all the details of the Saviour's experiment at: http://blazelabs.com/exp14.htm.

This electro-reactive ionization-flying model requires 46 kVolt, 4 mA, i.e. about 200 watts to fly with 100 g useful load.

It is interesting to note that Jean-Louis Naudin's previous opinion of 1997 – 1999 about priority of the Frolov's asymmetrical capacitor now is changed. Sometimes people wish to re-write the history. It is possible and it is easy. Naudin's web page about http://jnaudin.free.fr/lifters/story.htm (see Fig. 1) is not started from the real beginning of the story. He forgot our discussion of 1997 and also photos and MPEG video I have send to him when he has started his way in electrograviticis.

If you remember the name of Naudin's first asymmetrical capacitor "Frolov's Hat" (see Fig.2) then you know why there is this name of this technology. But really it is not important to try for some serious reply from Jean since he has no personal post address, phone, or his photo on the web site…

Also it can be useful to visit this page http://www.faraday.ru/t-cap.html to know about something **more important than electric ionization flying models (Lifters),** since it is just a reactive ionization way, which is similar to rockets.

Ideas now developing by us are not aimed to increase the power level (from 1 gram to 100 gram propulsive force). It is **a qualitatively new idea**; it is really electrograviticis instead of "electro-reactive Lifter".

So**, what is about real "history of the question" instead of the French version?** The question here is not about a priority, it is nonsense after T.T. Brown's patents. As an example the NASA patents on asymmetrical capacitors can be mentioned since the ideas were opened before the NASA patent and there is nothing new in it. The problem is that people who develop reactive ionization Lifters develop only primitive reactive technology. **It is not essence of T.T. Brown research but it is the distortion of his ideas.**

So, the team in France is working to develop and support ideas of secondary and wrong method. Why? Perhaps they either do not see the real way or have payment for the wrong way **to mask a real work.**

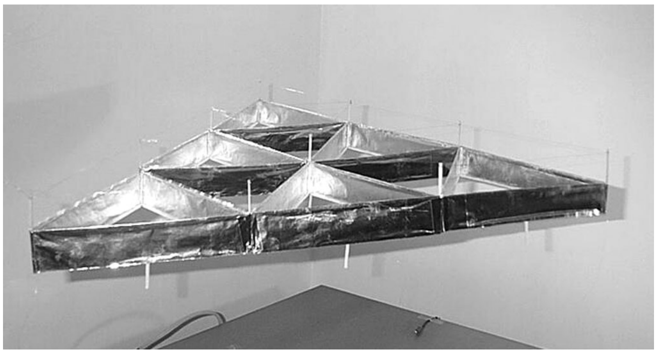

Fig. 1

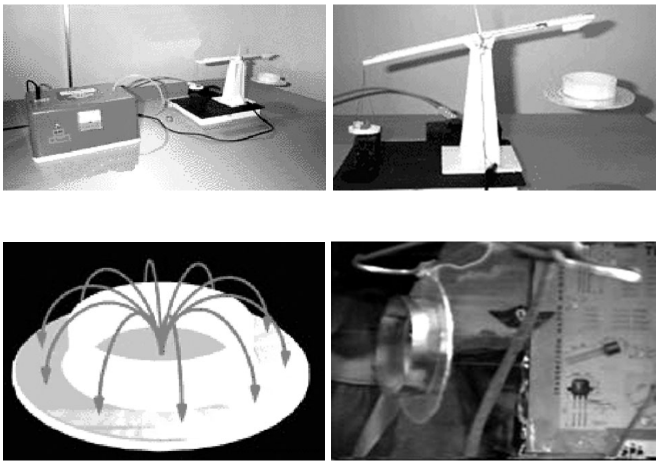

Fig. 2

Questions & Answeres

Josh Werrmann (Email: jsh111@yahoo.com): I wanted to ask you a few questions about an article titled "Highly Efficiency Electrolysis" by A.V. Frolov. In this article it is stated that Dmitry A. Latchinov patented a method for electrolysis, where one of the electrodes is insulated completely from the water. It is also stated that another scientist, Igor Goryachev, used a method of pulsed electric fields. I don't know if you know, but inventor Stanley Meyer has a patent that covers both of these processes together, in a way. His patent is U.S. patent number: 4936961. Take a look at this. In the patent it states EXPLICITLY that there is a pulsed electric field, where a condition of NO CURRENT is preferred. This is very similar to the above to Russian scientists, except that the patent states a condition of resonance is necessary. I am wondering do the two scientists that are described require resonance in their devices? Or just plain pulsed electric fields, where one electrode is insulated from water.

Another patent I would like to refer you to is Archie Blue's electrolyser, U.S. patent number: 4124463. This device is similar to A.V. Frolov's idea, to get the oxygen and hydrogen bubbles off the electrodes, except he uses a blower to blow air in the cell, instead of rotating the cell like A.V. Frolov does. His invention also uses the conductivity current to work.

Alexander V. Frolov: As far as I know resonance is not created in this case. The main idea is the removal of gaseous film which appears on the surface of electrodes by rotation or some other methods.

NEWS REVIEW

Giant magnets lead the world to the new source of energy

http:// www.sciteclibrary.ru

The system of 150-tons magnets which have been developed by scientists of the USA, Japan and Russia opens a new era of safe and unlimited energy. In this system nuclear fusions serve as a source of energy. They accumulate heavy chemical elements not by means of nuclear reactions but by ultrahigh plasma pressure.

After supplementary research is held which aims to reduce manufacturing costs of the giant magnet weighing 925 tons, this magnet will be produced and demonstrated at the opening of the International Thermonuclear Experimental Reactor (ITER). This magnet in its turn will become a part of a bigger system, that is combined into a magnet weighing approximately 10 000 tons. The task of ITER is the demonstration of the process of nuclear fusion, as an energy source. During the process of nuclear fusion the light elements are combined by immense pressure, thus producing heavier elements. During this production process a lot of energy is emitted. The task of giant high-power magnets is to create magnetic fields, which must hold and control plasma, or to charge electrically the gas, in which the fusion is being held.

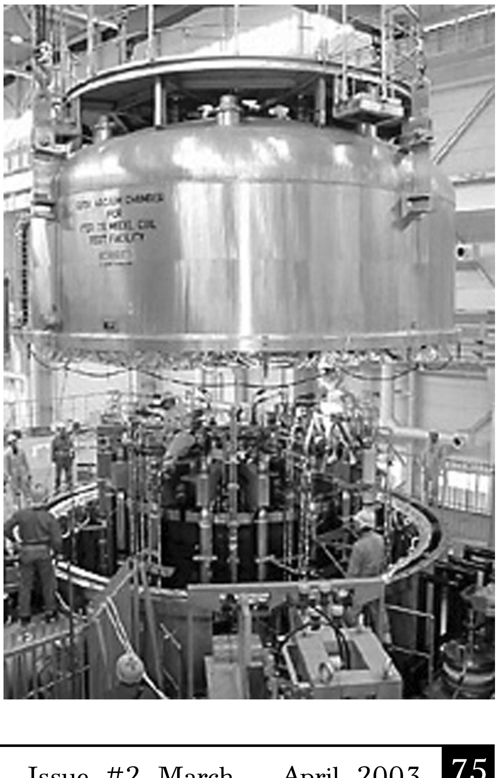

In Japan the system of 150-tons magnets serves as adjustment system of 925-tons magnets, which will be finally put into operation and heat up the ITER plasma. Two additional giant magnet systems will confine the plasma and will control its form. The model for one of them is presently being examined in Germany; the model of the other is in project.

Cylindrical 150-tons magnets have three main parts: external module, built by the Japanese team, internal module, built by the American team and a thin rod in the core which is equipped with process control equipment. Three different rods were separately checked, two of them were built in Japan, another in Russia.

BOOKS REVIEW

Harnessing the Wheelwork of Nature

(340 pgs. Adventures Unlimited Publishers, 2002)

Thomas Valone
Email: iri@erols.com

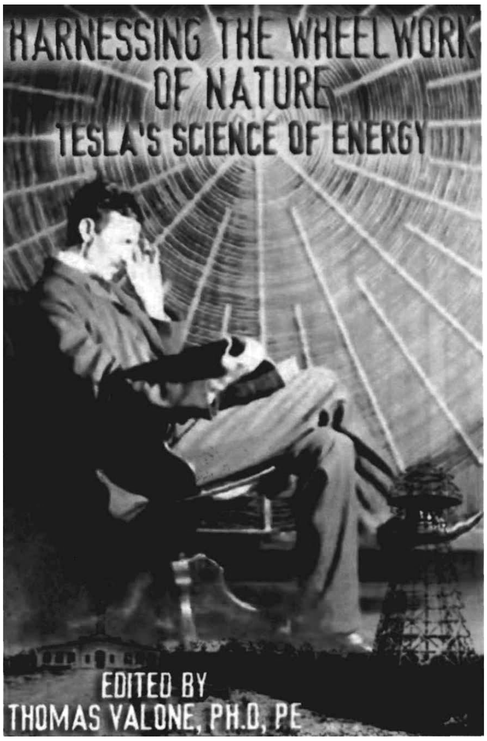

HARNESSING THE WHEELWORK OF NATURE
TESLA'S SCIENCE OF ENERGY
EDITED BY
THOMAS VALONE, PH.D, PE

It is a new book by **Thomas Valone**, who edited this anthology in time for the **Wardenclyffe Tower Centennial (1903-2003)**. This book presents for the first time, the feasibility argument for Tesla's most ambitious dream, the wireless transmission of power. Pictured on the book's cover near his feet, the 187-foot Wardenclyffe Tower was Tesla's means to deliver natural 8 Hz electricity anywhere in the world, by longitudinal waves.

Unknown to most electrical engineers, Nikola Tesla's dream answers the energy crisis worldwide, saves electrical conversion losses, and provides a real alternative to transmission lines. In Dr. Corum's contributed papers, he explains Tesla's magnifying transmitter, which Tesla compared to a telescope. Corum points out that "the tuned circuit of his magnifying transmitter was the whole earth-ionosphere cavity resonator." This fact helps explain why Tesla stated, "When there is no receiver there is no energy consumption anywhere. When the receiver is put on, it draws power. That is the exact opposite of the Hertz-wave system...radiating all the time whether the energy is received or not." Thus, with Tesla's futuristic transmission of power, source dissipation will only be experienced when a load is engaged in a tuned receiver somewhere on the earth. This fact alone represents **a major leap forward** in electrical transmission efficiency, even one hundred years later.

Why wasn't the prototype of Wardenclyffe finished in 1903? Tesla offered this visionary conclusion: "The world was not prepared for it. It was too far ahead of time. But the same laws will prevail in the end and make it a triumphal success... Let the future tell the truth and evaluate each one according to their work and accomplishments. The present is theirs; the future, for which I really worked, is mine."

Up until now, there has been a general malaise regarding the lack of scientific comprehension of Tesla's greatest dream. For example, the Serb National Federation notes, "With the exception of the first biography of Tesla by John J. O'Neill, science editor of the *New York Herald Tribune*, and published in 1944, unfortunately no biographer since has had the necessary scientific/engineering academic credentials to discuss Tesla's work in the various fields." Contributors to ***Harnessing the Wheelwork of Nature*** are primarily physicists and engineers who are experts in Tesla technology. Their wealth of knowledge demonstrates their mastery of this extraordinarily progressive and technical subject. Finally, the best academic credentials have been brought to bear on the world's greatest electrical futurist.

This is a very readable and profusely illustrated reference volume on wireless transmission of power, besides being an excellent biographical gold mine of Tesla history. Nick Cook, editor of *Jane's Defence Weekly* and author of *The Hunt for Zero Point* says, "Tesla is one of the great overlooked geniuses of science and electricity. His full story deserves to be told. Tom Valone sheds important new light on his life and work."

Integrity Research Institute will hold a Tesla Energy Conference & Expo on November 8-9, 2003 to celebrate the Wardenclyffe Tower Centennial (1903-2003) in conjunction with the Tesla Wardenclyffe Project, Inc. (www.teslascience.org).

See www.IntegrityResearchInstitute.org for more details.

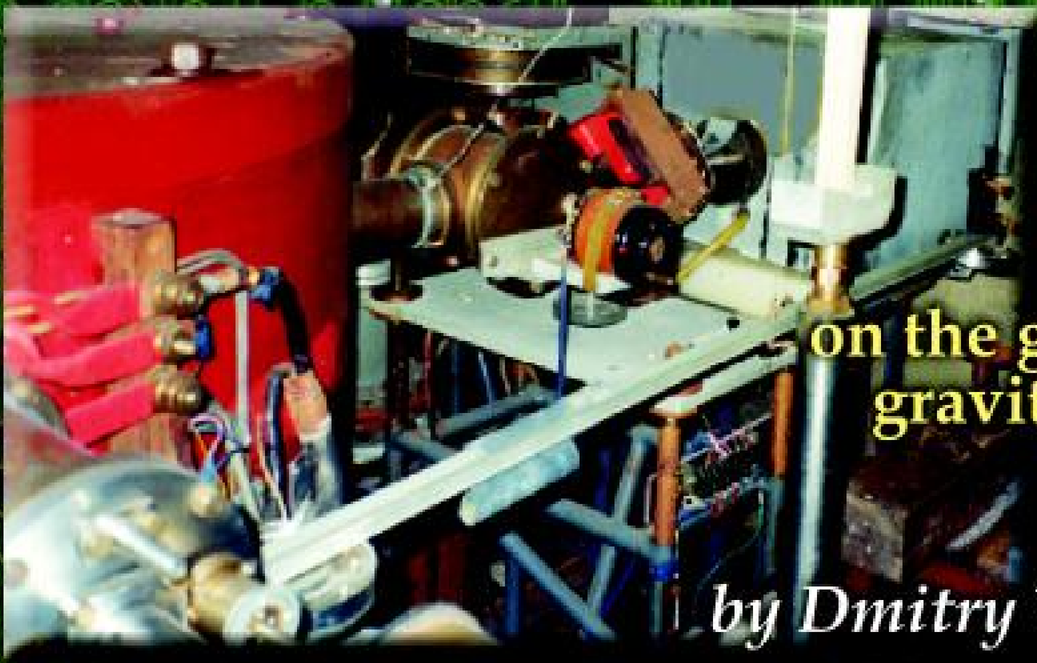

Experiment on the generation of gravitational field

by Dmitry Yu. Tsipenuk

Sergey S. Abramov

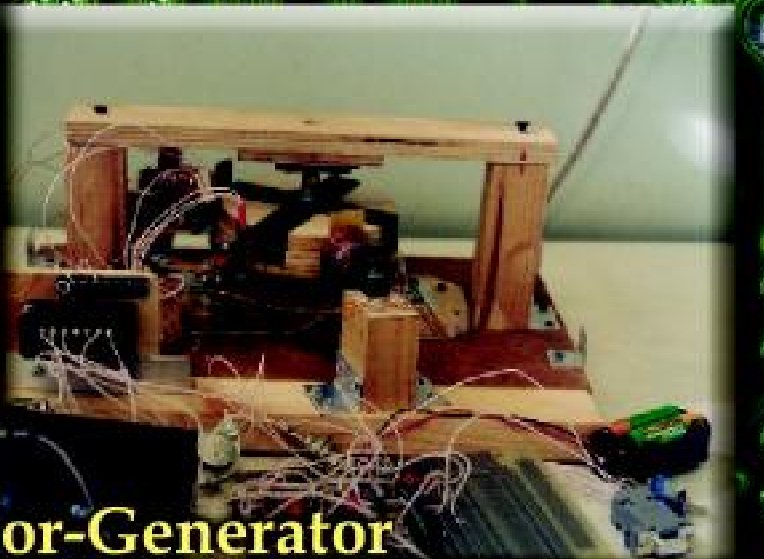

Model of Adams Motor-Generator

Electrolytic photoaccumulator for the XXI century energy systems

Новая Энергетика

Read the new article by Anton I. Egorov

Flynn Generator

Original version of photomaterials is presented at http://www.flynnresearch.net

Joe Flynn Laboratory, USA

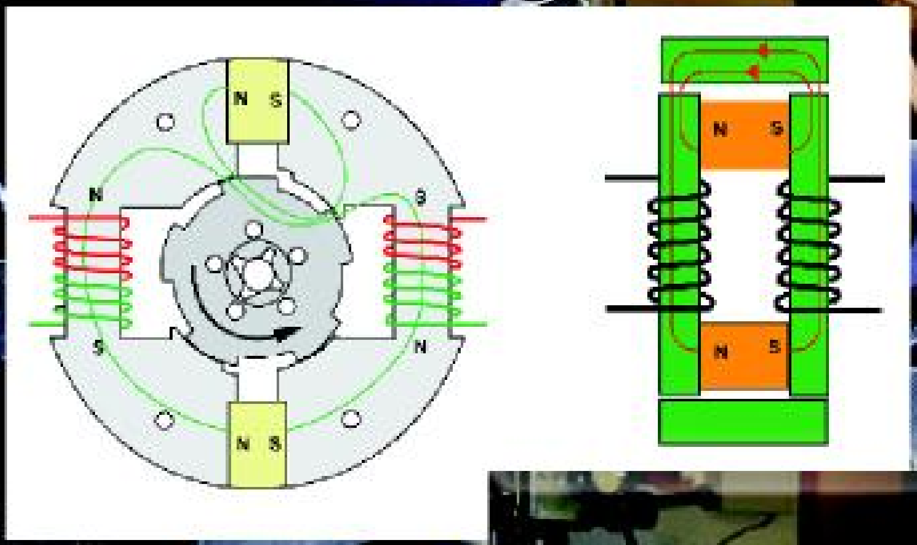

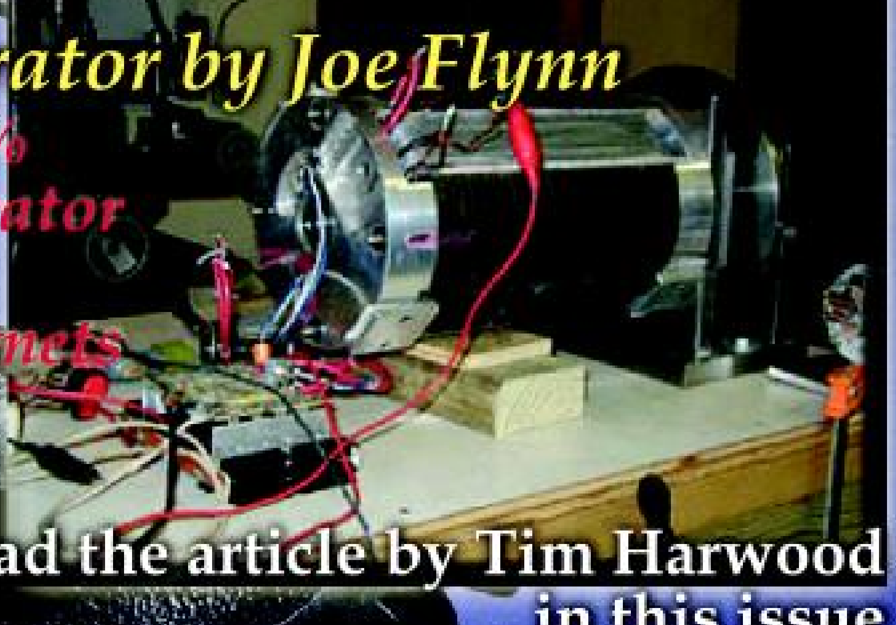

Electric generator by Joe Flynn

More than 350% efficiency generator operating on permanent magnets

ISSN 1684-7288
9 771684 728801

Read the article by Tim Harwood in this issue

New Energy Technologies

Printed version ISSN 1684-7288
Online version ISSN 1684-9698

Published by Faraday Laboratories Ltd. Issue #3 (12), 2003

Research on advanced aerospace systems and new energy sources

Time Machine

Conference in Moscow: discussion of the patent application on the method and device to control temporal parameters of matter

Time Machine

Experiments by research group "KOSMOPOISK"
Summer, 2001

Device designed
by Faraday Lab Ltd, 2002

Alexander V. Frolov
Vadim A. Chernobrov
April 12, 2003, Conference hall
Hotel "Cosmos", Moscow

NEW ENERGY TECHNOLOGIES #12

1. Works on the Designing of Time Machines. Vadim A. Chernobrov (Russia)
2. Control of Temporal Parameters of Physical Processes. Alexander V. Frolov (Russia)
3. Etherodynamics as a New Field of Physics. Vladimir A. Atsukovsky (Russia)
4. Electromagnetic Gravitational Interaction Vadim Ya. Kosyev (Russia)
5. Medium For Existing of Matter in Nature. Anatoly V. Rykov (Russia)
6. Adams Motor. Sergey S. Abramov (Russia)
7. Field Transformation in the Model of Extended Space. Dmitry Yu. Tsipenyuk (Russia)
8. Joe Flynn's Parallel Path Magnetic Technology. Tim Harwood (USA)
9. Flynn Laboratory Photos
10. Flux-machine and Its Analogies. Review
11. Russian Patents on Alternative Energetics.
12. Gravitational Spaceships. Gregory R. Uspensky (Russia)
13. On Velocity Of Drive-Free Motion. S.A. Gerasimov , V.V. Stashenko (Russia)
14. Electromagnetic Self-Action. S.A. Gerasimov, A.V. Volos (Russia)
15. Uranium Photoaccumulator. Anton I. Yegorov (Russia)
16. Global Energy Prize (reportage)
17. Teleportation. Alla V. Pashova (Russia)
18. "Mass Defect" in Home Conditions. Prokofy V. Sherbak (Russia)
19. Aether as Unified Field. Alexander M. Mishin (Russia)
20. Global Energy. Philip M. Kanarev (Russia)
21. Fusion Processes Of Molecules Of Oxygen, Hydrogen And Water. Philip M. Kanarev (Russia)
22. Plasma Energy Power Generation. Bruce A. Perreault (USA)
23. News
24. The Marcus Device Controversy. Tim Ventura (USA)
25. Systems of conversion of thermal energy to mechanical one. Review
26. I. Prigozin
27. Letters

CONFERENCE "THE TIME MACHINE"

Editorial: On April 12 of 2003 a scientific conference "The Time Machine" was organized by Faraday Laboratories Ltd in Moscow, Russia. It was devoted to the experiments on control of space-time physical properties. At the conference there were discussed problems of time and gravitation in the context of etherodynamics, experiments and applied aspects of these technologies. Below we publish a review of the main reports presented at the conference.

The First Results of the Works on the Designing of Devices to Control Parameters of Physical Processes (of Time)

Further Plans on Designing of the Time Machine

Vadim A. Chernobrov, Russia

"KOSMOPOISK", Nagatinskaya 19A, 111533,
Moscow, Russia
E-mail: chernobrov@kosmopoisk.org

To conduct the experiments on the influence upon physical Time (density of space energy) some special devices were used as a general method of such influence. These devices can create converging waves which can cause the appearance of quasi-monopole in the confined space. *Quasi-monopole* is a part of space which has some parameters of hypothetical unitary monopole or bunch of such particles (in particular, it allows registering one magnetic pole by means of measuring equipment from the outside at some distance from this pole).

The pilot experiments have shown that it is very difficult (if not impossible) to create long-living quasi-monopole by means of permanent magnets or electromagnets operating on direct currents (in this case quasi-monopole represents a space with one outer and one inner magnetic pole). It can be explained by the fact that lines of force of the "inner pole" invariably find a weak spot in the heterogeneous surface of magnets and break out. As a result, at the device along with one "outer" magnetic pole there is a local output of magnetic lines of the "inner" pole.

During the designing of new devices there was a task to create a quasi-monopole situation in the confined space. This situation should be created not uniformly but transiently by pulsation method. Frequency of work of electromagnetic oscillators first of all was selected depending on linear dimensions of the devices. Selection of the frequency was made in such a way that one period of pulsation does not exceed the period of time which is necessary for electromagnetic waves to reach the center and opposite waves of the device.

At the designing of laboratory devices, which generate converging longitudinal waves, several different principles to obtain required parameters were considered in the multi-layer quasi-monopole, which works in high-frequency mode. Several types of the devices were realized at different degrees of development and with different success. It was shown that the simplest devices were that ones which used electromagnetic (solenoid) oscillators connected in series and in parallel. In different experiments between 3 and 5 such surfaces were used. These surfaces were called electromagnetic work surfaces (EWS). All layers of EWS of different diameters were mounted in series in each other (like matreshka). The outer layer was either mounted at force shell or simultaneously represented such a shell by itself.

The size of the maximal EWS was about 0.9 m, the diameter of the minimal (inner) EWS was equal to 115 mm that was enough to place laboratory animals inside the control detection devices. Laboratory animals were used to determine the consequences of the influence of converging spherical electromagnetic waves. Payload, (i.e. that which was directly used for transfer in Time-Space), included the afore-mentioned detection devices and (occasionaly) laboratory animals. The term "payload" was coined by analogy with the term which is used in cosmonautics. Volume of payload section was placed in the center of symmetry of the Time Machine (TM). In all the earliest Machines (except the 7th model) this volume still has not exceeded the volume of a football. The device with an outer diameter of 2.1 m and inner payload section of 1 m has the maximum size. It allows making human-aided experiments.

At different stages of the experiments the calculations were made by all available known modern methods to fix time. All types of electronic, quartz, mechanical and several specially made doubled quartz generators were used (there were compared frequency readings of measuring and etalon heat-insulated generators which were placed at a distance). Certain experiments used lightguiding diodes and some other methods. Before and after the experiment (more rarely during the experiment) readings of measurement clocks were periodically compared with those of an etalon clock and with signals of exact time which were transmitting by radio. Other physical factors causes side effects upon some types of measuring devices, for example, upon quartz-crystal clocks. However, doubling of measurement methods allowed essentially decreasing inaccuracy of measurements.

At some operating modes (which were not always predicted) the change of Time speed was attained (Prof. Nikolay A. Kozyrev called it density of Time t/t_0). This change came to about a fraction of a second per an etalon hour. Let us take usual "etalon" earth time as $t_0=+1$ then it becomes clear that it is speed range of $+0.99<t/t_0<+1.01$ which is researched in the experiments. Thus being placed inside the device, the inanimate objects and animals were transferred in the Future (with "above-zero" speed) more slowly or faster than those around them. It occurred at all operational modes of the device (deceleration or acceleration).

Difference of Time (gradient of Time speed or curvature of Space-Time field) was observed not only inside the TM, though there is no doubt that a maximal value of changed Time was registered inside the smallest "matreshka". As was expected, during the experiments Time change was also registered outside the device. However such change which had an opposite sign was smaller than the change inside the device approximately by an order of magnitude. It entirely corresponds to geometric laws, i.e. in proportion to the cube of the distance).

In other words, TM influences not only its inner part and pay load but also the environment. It bears a strong resemblance to jet propulsion but in Time and not in Space. It is a flight which is realized by rejection of Time instead of mass.

Thus it was determined that the processes of deceleration and acceleration of Time distinctly differ in their nature and consequences. In such a way the deceleration was considerably smoother and more stable. During acceleration there were observed sudden changes in readings and this operational mode was characterized by general instability and dependence on any (or many) external factors. Particularly, acceleration instability also consisted in the following: at fixed power the value speed of Time depended on the time of day and situation of the Moon. Possibly it also depended on some other factors including the presence of an operator or other people near the device. Even insignificant external influence, for example, mechanical shaking, caused the change of speed value which sometimes was rather significant.

Inside the laboratory devices there was also registered that **Time can change with some sluggishness**. After the changed Time speed influenced some physical object (for example, soil) then residual effects were observed at it for some period of time. These effects can be eliminated only by the influence of another speed of Time.

Preliminary general conclusions

The Present is the transfer or transformation of the polyvariant easily changeable Future into the univariant unchangeable Past. Thus flights in the Past (at "negative" density-speed t/t_0) and in the Future will happen in different ways. They can be compared with the motion of an ant along the tree: from any point of the tree (i.e. from the Present) there is only one way downwards open to the ant (i.e. in the Past) and many different ways upwards (i.e. in the Future). However among all the ways to the Future undoubtedly there are most probable, low-probable and almost improbable variants for development of the Future. The less probable this variant of the Future the more unstable and energy-intensive the motion to the Future will be. According to the "law of the crown of tree", return to the Present is possible only if when being in the Past the traveler does not interfere in the course of events and does not change the course of the past History. Otherwise the chronotraveler will return to the parallel Present from the Past by another branch of History. Penetration to the Future from the Present is hampered by the choice of the branch for transference. However return from any variant of the Future into the Present is possible at any scenario if there are no fusions of different variants of History…

In other words, we have a circumstantial evidence of the assumption that Time has more than one dimension. Thus there is an affirmation of the theoretical inferences of R. Bartini who believed that Time has 3 dimensions. Hence we can consider our terrestrial globe as 6-dimensional where the dimensions are: length; width; height; age or date of Time; variant of History or blur of Time; density or speed of Time. Thus the notion of "Arrows of Time" is completely absent in the fourth dimension (date of Time) but it is a special case of the notion of the sixth dimension i.e. speed of Time. At the same time the notion of speed of Time also relates to the physical notions of gravitation and energy. Thus the notion of "Einstein-Rosen bridge", which was introduced in 1916, or notion of "worm course", which was introduced by John Willer in the end of the fifties, are connected with transference in the 5th and 6th dimensions.

Method and Device to Control Temporal Parameters of Physical Processes by Means of Changing of Energy Density of Space

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./fax: 7-812-380-3844
Email: office@faraday.ru

This invention belongs to the methods and devices to provide control on rate of physical processes (that includes the process of the existence of matter in space-time) by means of increasing or decreasing of energy density of space (i.e. energy density of physical vacuum or density of the aether).

Let us consider the history of the invention:

Earlier there were proposed some methods and devices to influence the rate of physical and chemical reactions, biological processes or period of oscillation of the system. In the works by N.A. Kozyrev [1] there is a description of the experiments on the influence of some process (for example, process of evaporation or crystallization of matter) upon the period of another process, which serves as a detector and can be compared with reference oscillation process. In one case, the rate of oscillations of the detector decreases in the surrounding area near the process of matter evaporation. In another case, the rate of detector oscillations increases in the surrounding area near the process of matter crystallization. If we use a term of "entropy" then it is possible to say that the processes which are accompanied with entropy increasing (for example, conversion of matter from solid state into liquid one) influence on the matter (surrounding processes) in such a way that entropy of systems decreases. In another case, for example, near the process of crystallization, entropy of systems increases in the surrounding area near this process. Kozyrev used the term "wave of density of time" and he made a conclusion that in addition to "directivity" of time (time course) there are active properties of time, for example, "density of time".

To develop this approach for applied purposes it is necessary to use in-depth analysis of the physical sense the "law of the crown of tree", return to the Present is possible only if when being in the Past the traveler does not interfere in the course of events and does not change the course of the past History. Otherwise the chronotraveler will return to the parallel Present from the Past by another branch of History. Penetration to the Future from the Present is hampered by the choice of the branch for transference. However return from any variant of the Future into the Present is possible at any scenario if there are no fusions of different variants of History…

In other words, we have a circumstantial evidence of the assumption that Time has more than one dimension. Thus there is an affirmation of the theoretical inferences of R. Bartini who believed that Time has 3 dimensions. Hence we can consider our terrestrial globe as 6-dimensional where the dimensions are: length; width; height; age or date of Time; variant of History or blur of Time; density or speed of Time. Thus the notion of "Arrows of Time" is completely absent in the fourth dimension (date of Time) but it is a special case of the notion of the sixth dimension i.e. speed of Time. At the same time the notion of speed of Time also relates to the physical notions of gravitation and energy. Thus the notion of "Einstein-Rosen bridge", which was introduced in 1916, or notion of "worm course", which was introduced by John Willer in the end of the fifties, are connected with transference in the 5th and 6th dimensions.

◆

Method and Device to Control Temporal Parameters of Physical Processes by Means of Changing of Energy Density of Space

Alexander V. Frolov, Russia

General Director, Faraday Lab Ltd
Tel./fax: 7-812-380-3844
Email: office@faraday.ru

This invention belongs to the methods and devices to provide control on rate of physical processes (that includes the process of the existence of matter in space-time) by means of increasing or decreasing of energy density of space (i.e. energy density of physical vacuum or density of the aether).

Let us consider the history of the invention:

Earlier there were proposed some methods and devices to influence the rate of physical and chemical reactions, biological processes or period of oscillation of the system. In the works by N.A. Kozyrev [1] there is a description of the experiments on the influence of some process (for example, process of evaporation or crystallization of matter) upon the period of another process, which serves as a detector and can be compared with reference oscillation process. In one case, the rate of oscillations of the detector decreases in the surrounding area near the process of matter evaporation. In another case, the rate of detector oscillations increases in the surrounding area near the process of matter crystallization. If we use a term of "entropy" then it is possible to say that the processes which are accompanied with entropy increasing (for example, conversion of matter from solid state into liquid one) influence on the matter (surrounding processes) in such a way that entropy of systems decreases. In another case, for example, near the process of crystallization, entropy of systems increases in the surrounding area near this process. Kozyrev used the term "wave of density of time" and he made a conclusion that in addition to "directivity" of time (time course) there are active properties of time, for example, "density of time".

To develop this approach for applied purposes it is necessary to use in-depth analysis of the physical sense of the "time density" notion. Connection of notions about "time directivity" and "entropy of the system" was demonstrated in the work "Introduction into thermodynamic of irreversible processes", 1964 by Ilya Prigozhin [2]. In the work "Quantum fluctuations of vacuum in curved space and theory of gravitation" by A.D. Sakharov [3] the conception of vacuum structure was suggested. In the USA the work "Can the vacuum be engineered for space flight applications?" by H.E. Puthoff [4] is well-known. The author considered the applied aspects for study of vacuum structure and described the method and device to obtain propulsive force by means of changes in vacuum properties.

A substantive conception of time and methods for creation of waves of energy density were also considered by Prof. K.P. Butusov in the work "Time is a physical substance", 1991 [5]. In the book "What is The Time?" by Yu. G. Belostotsky [6] the connection between the notions of time and aether was demonstrated. This connection was considered from the point of view of astrophysics there.

We can also say that the modern conception of aether is successfully developed by V.A. Atsukovsky in his works [7].

In my articles, for example, "Physical principles of the Time Machine" [8], it was demonstrated that to develop experimental works on the topic it is useful to clarify the terminology and to consider **"waves of time density" as longitudinal waves of energy density in space.** In this case the notion of "time density" has a physical sense of energy density (aether density).

This approach can be realized practically by means of classical electro-technical and radio engineering methods and it is a development of aether-dynamical conception on the nature of electricity and magnetism by M. Faraday, "Experimental researches on electricity", volume 3, [9].

Let us consider a usual bipolar magnet from the point of view of aether-dynamics. Then it is possible to say that it is inflow and outflow of aether, i.e. it is a balanced energy system which does not change energy density in space. In this case it is evident that creation of magnetic monopole or modeling of quazi-monopole by means of electro-dynamical methods is a technical basis to create some local change of energy density in space.

Electric processes could also be used alongside with the magnetic phenomena. For example, in another book "Symmetrization of Maxwell-Lorenz equation" by Prof. Butusov [10] the creation of longitudinal wave was also considered. It was demonstrated that an electrically charged sphere can radiate longitudinal wave when the sphere radius is changing, i.e. when its surface is changing while the value of electric charge is constant.

One more well-known method is described in the book "Experimental gravitonics" by Polyakov [11]. There is a consideration of the generation of gravitational waves at the high-frequency magnetization and demagnetization of ferromagnetic material, i.e. at the powerful volume magnetostriction. Since at this phenomenon there are changes of matter density (i.e. changes of energy density in space, which is occupied by matter), then volume magnetostriction is a special case of changes of energy density.

Earlier Vadim A. Chernobrov had described a method and device to control temporal characteristics of physical and chemical processes by means of creation of the magnetic monopole (quasi-monopole). In this magnetic mono-pole there is a convergent wave, which is created by several sources situated in the spherical frame. According to this method in the multilayer spherical structure where every layer (the so called "electromagnetic work surface") is an assembly of electromagnets, by means of series connection of the layers the wave is created, which converges to the center of the device. The device has the same outside magnetic poles of the electromagnets (and the same inner poles) and thus a model of macroscopic magnetic monopole is created.

We assume that at in-phase operation of all sources of waves, interference of longitudinal waves provides some change of value of energy density of space in the focus of the system.

Experimental facts prove that detectors installed in the center of the device (for example, mechanical or electromagnetic oscillators) show change of period in their own oscillations. We have protected them by shielding them from heat radiation as well as from other kinds of electromagnetic influence. Thus it is possible to assert that the detectors decelerate or accelerate their oscillation period depending on the energy density, which is generated in the center of the device.

However for the experiments made by means of such a device, accurate adjustments of all wave sources are required to provide their in-phase operation. At the same time, operational stability of the system depends on the operational stability of each of the wave sources. Increase of the impulse frequency causes increase of the effect; however, it is limited by the parameters of electromagnets and generator of impulses. Besides, to increase the effect it is necessary to use more powerful energy sources since the current in the windings of electromagnets determines the value of the magnetic field of the created quasi-monopole.

Since the efficiency of such systems directly depends on the frequency and the value of changes of energy density in space, then in the next version of realization of this technology we suggest using of plasma shells instead of electromagnetic working surfaces. This will allow significant improvment in the specific parameters of the device.

Thus let us consider general engineering principles of operation and outline the ways to develop this method. Fig. 1 represents a three-layered electromagnetic emitter. This electromagnetic emitter is designed according to the invention in which the directed radiation of wave of energy density is created along the axis of the device.

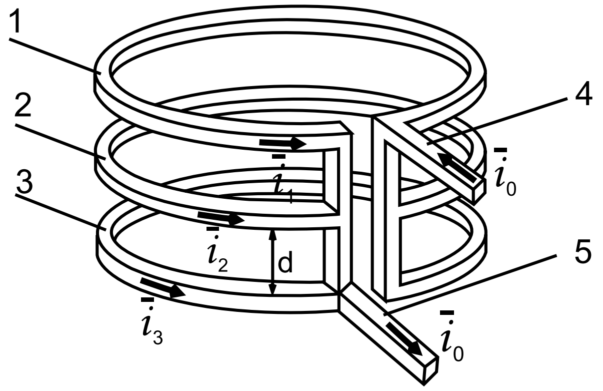

Fig. 1
Three-layered electromagnet

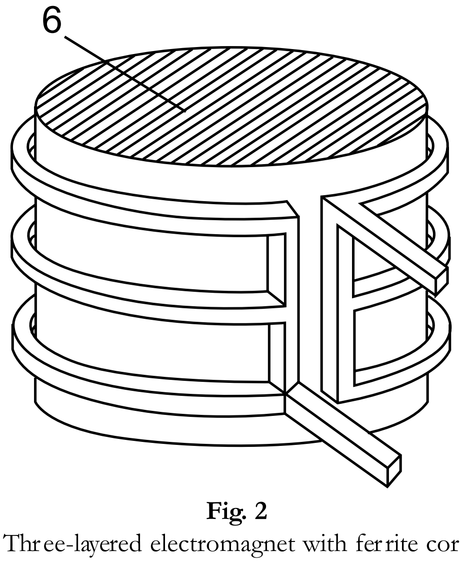

Fig. 2
Three-layered electromagnet with ferrite core

The device is designed according to the idea by Vadim A. Chernobrov for creation of the directed wave of energy density by means of phase shift in propagation of impulse front in three current branches, namely i_1, i_2, i_3. These branches are displaced along the electromagnet axis at some distance d.

The device works in the following way. When the pulsed generator is activated, front of current pulse i_0 appears at the output 4. Impulse front at branch 1 advances impulse front at branch 2 that is caused by spatial shift of current branches 1, 2, 3 relatively to each other along the electromagnet axis at the distance d. Impulse front at branch 2 in its turn advances impulse front at branch 3 for a certain time T. The second output of the electromagnet 5 is placed in such a way that impulse front at branch 1 will phase lag behind the impulse front at branch 2 (which in its turn will phase lag behind the impulse front at branch 3) for the same period of time T. Therefore at branch 5 the united impulse front is generated again.

Time T can be calculated in the following way:

$$T=d/c \text{ (seconds)} \qquad (1)$$

where c is a constant of propagation of impulse front. This constant is known as velocity of light.

At each impulse the T (i.e. the value of relative lag of impulse front) is a constant value. Thus high-frequency consequent excitation of layers of the electromagnet appears at each impulse. The frequency of the excitation is calculated in the following way:

$$f=1/T \qquad (2)$$

where T is relative lag of impulse front in seconds.

There is an example of frequency calculation: for the shift distance d=7 mm we can calculate a lag $T = (7/2.997924) \times 10^{-11} = 2.335 \times 10^{-11}$ (seconds) and frequency $f=1/T$ approximately comes to 4.28×10^{10} (Hertz).

Thus this design of three-layered electromagnetic emitter allows creating the waves of super-high-frequency band (for example of millimeter range) without the use of a semiconductor or other radio components.

It is very expedient to use the electromagnets designed with magnetostrictive material cores, that will significantly increase energy density of the longitudinal wave, which is generated by the multilayer electromagnet. In Fig. 2 there is an emitter with the core. In the case of high-frequency ferromagnetic magnetostrictive materials 6 the efficiency of emitter operation significantly increases.

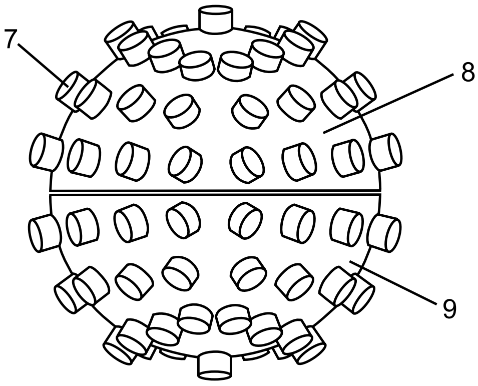

Fig. 3
Spherical design

Fig. 3 represents the spherical distribution of emitters 7 at the upper 8 and lower 9 hemisphere of the frame which could be opened in order to place detectors and different objects inside. It can also allow to establish in what way the changes of density of space energy influence the properties of different materials, velocity of physical and biological processes as well as chemical reactions. The installation of detectors 10 inside the device is shown on Fig. 4.

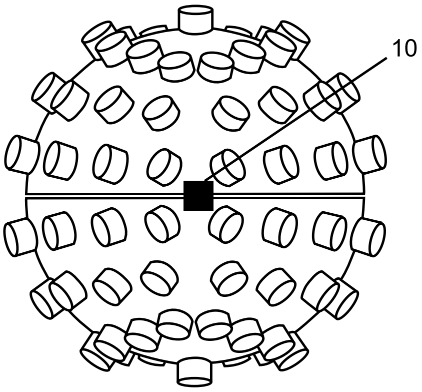

Fig. 4
Detectors inside the system

Another version of design is represented in Fig. 5, where the suggested method is realized by means of a spherical electric capacitor with three coats 11, 12, 13. Each capacitor coat is connected to the outlet of three-phase pulsed generator 14.

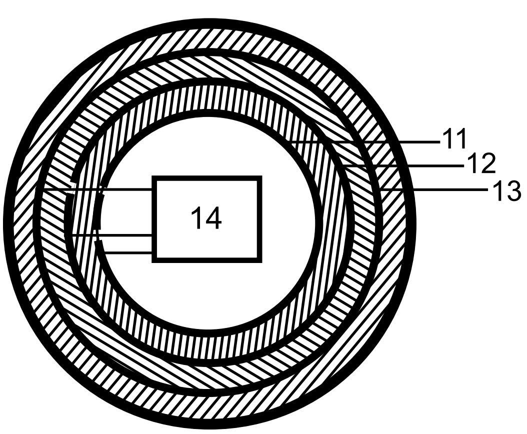

Fig. 5
Three-layered spherical capacitor

In this case the wave of energy density is created without electromagnetic emitters and this principle is not related to the modelling of magnetic quasi-monopole. The device operates due to the control unit, which provides the high-frequency changes of electric potential at each coat in such a way that the mode of high-frequency converging or diverging wave of energy density is created. In fact it is a standard three-phase generator but it produces not the rotation of the rotor of some electromotor but **"compression" or "decompression" of aether.** Aether is "pumped" in the center of the device or "pumped out" of the center.

In this case there is no need to tune separate sources of waves to make the device work in in-phased mode. It ensures reliability of the device operation if to be compared to the quasi-monopole. Besides, much less energy is necessary for the processes of charging and discharging of multi-layer spherical electric capacitor than for creation of magnetic field by means of conductivity currents.

Since the efficiency of such systems directly depends on the frequency and value of changes of energy density in space, we suggest the use of plasma shells instead of electromagnetic work surfaces for the next version of the device. It will allow significantly improvment in the specific parameters of the device. For that it is enough to place the electrodes of the multi-layers capacitor in a low-pressure gas area and these electrodes should be made as gauze electrodes. At that the wave is created in plasma, which is excited layerwise by several gauze electrodes placed in a spherical space between the inner and outer spherical bodies of the device. Therefore this version of design of the device can be considered as **the manipulation of the plasma method**.

Fig. 6 is a plan of one more design version. It is made as an inner frame 15 and outer frame 16, the space between them is filled with some gas 17. Three electrodes 18, 19 and 20 are connected to the three-phase pulsed generator 21. **Consequent excitation of plasma layers by electrodes 18, 19 and 20 creates the wave of energy density.** Propagation of this wave can be directed both to center of the device and from the center of the device.

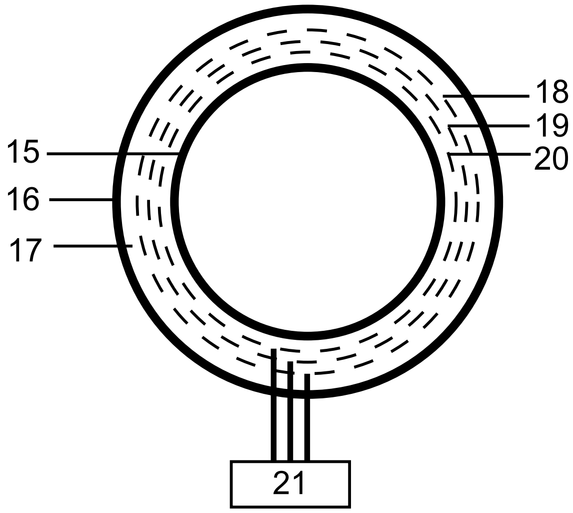

Fig. 6
Wave in plasma

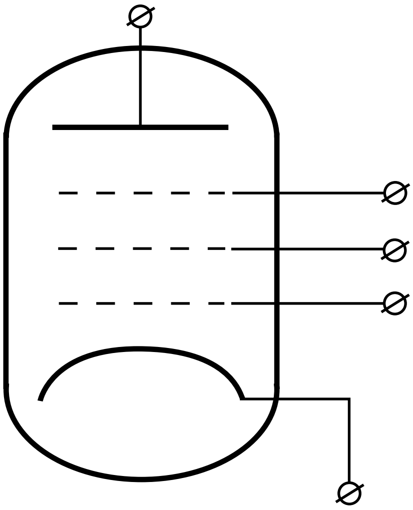

Fig. 7
Schematic electric circuit

In Fig. 7 there is a schematic electric diagram of the design.

To conclude. The work on time control has just started. We have clearly determined the main physical principles of operation of such devices, which can change the aether density in some given volume of space and thus influence temporal parameters of any physical process. The small experimental results today allow us to make real positive conclusion on availability of this method and on the possibility of its practical application in applied aspects. The first aspect is antigravitation propulsion technology and we are developing the methods to detect mass (weight) changes in the time control experiments to prove this applied possibility. Another area is medical applications of the changes in the aether density.

Russian Federation patent claim #2003110067 was filled April 9, 2003. At present time we are interested in marketing for this technology as well as in search of additional investment and partners.

References

1. N.A. Kozyrev, Selected works, Leningrad State University, 1991
2. I. Prigozhin, Introduction in thermodynamics of irreversible processes. M., 1964
3. A.D. Sakharov, Quantum fluctuations of vacuum in curved space and theory of gravitation, Reports of USSR Academy of Sciences, vol. 12, 1968, p. 1040
4. “Can the vacuum be engineered for space flight applications & Overview of theory and experiments”, Dr. H.E. Puthoff, Infinity Energy, July-November, 1997
5. K.P. Butusov, Time as physical substance. Collection of articles “Problems of space and time in modern natural science”, Issue 14, L. 1990, pp. 301-311
6. Yu.G. Belostotsky “What is The Time?” Saint-Petersburg, 1991
7. V.A. Atsukovsky “General etherodynamics”, M., 1990
8. A.V. Frolov, Physical Principles of The Time Machine, New Energy Technologies #3(6), pp. 8-10, Saint-Petersburg, 2002
9. M. Faraday, Experimental researches on electricity, vol. 3, M. 1959
10. K.P. Butusov, Symmetrization of Maxwell-Lorenz equation. Problems of Space and Time in modern natural sciences, vol. 15, L. 1991
11. S.M. Polyakov, O.M. Polyakov “Introduction in experimental gravitonics”, pp. 58-59, M. 1991
12. V.A. Chernobrov “Mysteries of Time”, M. 1999

Aetherodynamics as a New Field of Physics

Theory and Experiments

Vladimir A. Atsukovsky, Russia
Email: atsuk@dart.ru

A crisis exists these days both in Natural science on the whole, and, in its basis, i.e. physics. This situation is caused by fundamental errors in the methodology of its theory. The postulate method and predominance of mathematics above physical sense can be also considered as the reasons for this. Crisis situation has also taken place in the past and the way out was always to use a deeper level of matter organization than the accepted one.

A large amount of material on elementary particles of matter is collected in physics. It was discovered that all these particles can be transformed in each other. Besides that, there was discovered the ability of vacuum to create such particles at strong electromagnetic fields. In this connection it becomes clear that all these particles and vacuum have a common building material, i.e. a physical medium which fills all the world space. This medium was renamed **aether** and the element of the medium was renamed **amer**.

The worked out methodology of aetherodynamics allowed determining that aether is a usual viscous coercible gas for which all laws of usual dynamics of gases are valid. It has allowed to determine its characteristics for near-Earth space and then to describe the structures of primary stable elementary particles (i.e. proton, neutron, electron, photon, atomic nuclei, atoms and some molecules) as well as physical essence of general fundamental interactions (i.e. strong and weak nuclear interactions, electromagnetic and gravitational interactions) and some physical phenomena.

Also developed was the model of aether circulation in the Universe within the limits of perpetually existent matter, usual Euclidean space, and evenly flowing time. Formation and decay of matter, formation and decay of celestial bodies and galaxies as well as functional classification of galaxies are also included there.

To confirm some theses the author and his research team conducted several experiments mainly on electromagnetism that gave positive results. The research was conducted on testing of the presence of aether wind. These research works have confirmed the results obtained by D. Miller in 1925.

Report by Vladimir A. Atsukovsky

Electromagnetic Gravitational Interaction

Vadim Y. Kosyev, Russia

http://uft.h1.ru; http://uic.nnov.ru/~kovy2
Email: uft@inbox.ru

In the article some mechanisms of electromagnetic gravitational interaction will be considered in the view of the Unified Theory of Field, Space and Time. See the complete description of The Unified Theory of Field, Space and Time at http://uic.nnov.ru/~kovy2; http://uft.h1.ru or in the book [1] which has the same title.

First of all, let us consider the structure of Space-Time where the radiation propagates. In our World gravitation is the unique substance which exists everywhere and gets over any obstacles. There are no known methods of insulation of gravitational field in modern science. It is impossible to imagine space and gravitation separately. Gravitation exists everywhere where there is some space. The gravitational field created by all masses of our metagalaxy is the aether in which cosmic objects moves and electromagnetic oscillations are propagated. The space surrounds us since the whole matter carries gravitational charge of only one sign. From astronomical research of cosmic space it follows that radius of metagalaxy is equal to the critical (gravitational radius) and hence gravitational potential is equal to $c^2 \sim 10^{17}$ $[m^2/c^2]$ in each point of space.

In our space the sum potential of the electric field is equal to zero. All bodies and space-time as a whole are electrically neutral. This assertion follows from two unique properties of gravitational space-time:

1. Values of electric charges of different signs of elementary particles are exactly equal.
2. There is an equal number of elementary particles which carry electric charges of different signs.

If electric potential is equal to zero in the given point of space then energy of alternating magnetic field passes entirely to the electric field and radiation gains the parameters of electromagnetic oscillations. However electric potential of positive and negative signs can prove its properties at the simultaneous charging of different local spatial regions. In the electric field the character of radiation differs from that one which is observed in ideal gravitational space-time (without electric potential).

In the presence of electric potential some part of magnetic energy is spent to create gravitational alternating field. Radiant energy is absorbed. Electrically charged spatial regions are perceived by us as filled with a strong absorbing substance. At the same time if the potential of the electric field can be compared to the potential of the gravitational field then spreading of radiation becomes impossible.

Change of both electric and gravitational field results in the creation of a magnetic field in the region of space-time which has a dual electrogravitational nature. Change in the magnetic field results in the creation of both electric and gravitational fields. The amplitude of electromagnetic and magnetogravitational constituents of the unified electromagnetic gravitational oscillations depends on field potential of opposite nature. The electromagnetic constituent is determined by gravitational potential and the magnetogravitational one is determined by electric potential. Transference of gravitational masses of matter in electrogravitational field-aether causes the creation of the proper magnetic field. Coming from the direction of magnetic field some force influences upon the moving electrically neutral masses. This force is similar to Laurence force.

Electromagnetic gravitational converter designed according to the Searl Effect (Fig. 1) visually demonstrates the mechanisms of electromagnetic gravitational interaction. A detailed description of the experiment is presented in the article [2] by Vladimir Roshschin and Sergey Godin (see http://www.n-t.org/tp/ts/dms.htm).

The device consists of a cylindrical stator of about 1 meter in diameter which is surrounded by 24 cylindrical rollers. Stator and rollers are made of magnetic material and they are magnetically linked (there is no contact between them). Vector of the magnetic field of the stator and rollers is vertically oriented along the axis of the cylinder but it has opposite direction. Rollers are mounted on the movable separator which circles round the stator. Each roller rotates on its axis in the same direction. The weight of the device is 350 kg. During rest, the device is electrically neutral and has proper gravitational charge, inertial mass, magnetostatic field of the stator and rollers.

On rotating of the magnetic system the following effects were observed:

- Depending on the frequency (up to 35%) the weight of the device was decreased.
- At a frequency of 550 rpm there is a spontaneous acceleration of the system and turns of the rotor abruptly increase (in quadratic dependence). To stabilize the mode it had to take off the excess energy to active load (up to 7 kWtt) by means of electromagnetic transducers.
- Magnetic field was distributed in a room as concentric surfaces of 5-8 cm thickness and 50-60 cm period. Depending on speed of rotation of the magnetic system there was a fixed temperature decrease from 0 to 8° within the regions of space which were curved by magnetic field.
- The glowing (corona discharge) appeared around the working converter.

Appearance of all mentioned effects is caused by the simultaneous presence of electric, magnetic and gravitational fields in the area of the converter.

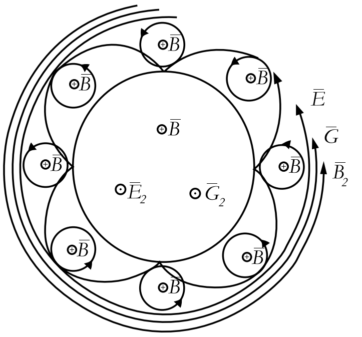

Fig. 1
Electromagnetic gravitational converter

Because of the Searl Effect a very complicated configuration of electric, magnetic and gravitational fields appears in the device. Rollers having their own field B rotate around the stator which has its field B. Cycloidal motion of numerous magnets which are built in the rollers causes change in the magnetic flux. EMF E appears in the area of rotation of the rollers. EMF is directed along the perimeter around the device. The electric field creates stationary waves which are registered around the device as concentric electromagnetic gravitational walls. Magnetic rollers should be made non-conducting in such a way as to allow electric field penetrating inside. In the presence of the electric potential change of the magnetic flux causes appearance of gravitational field G. This field is directed along the perimeter around the device. In the presence of electric potential the Lorenz force influences the gravitational masses of the rollers. These gravitational masses move in the external magnetic field of the stator. The Lorenz force is directed transversely to the velocity along the cycloid and in the direction of the field B. The value of the Lorenz force depends on the electrical potential, magnetic intensity, mass of the rollers and their traverse speed. The electric potential in its turn depends on speed of rotation of rotor of the magnetic system. As a result the Lorenz force depends on speed according to square law. Rotating gravitational masses of the rollers generate magnetic field B_2. Field B_2 generates derived fields E_2 and G_2 etc.

Spontaneous acceleration of the system does not depend on the direction of rotation. Directions of fields and forces change at change of the direction of rotation that corresponds to acceleration in the opposite direction. When the direction of rotation is changed the weight of the device changes. When rotation is in one direction the field G_2 is directed upwards (decrease of weight), when rotation is in another direction, it is directed downwards (increase of weight). In the presence of electrical potential many other interesting but still unstudied phenomena are observed (for example, curve of space, change of speed of time-flow, changes of ambient temperature…). Magnetogravitational converter transforms gravitational energy of space-time into energy of electric field which in its turn is transformed into mechanical energy of the rotating magnetic system and electric energy of the load. Energy resources of the future are not coal and gas reserves or nuclear power. They are in the inexhaustible energy of gravitational aether of our space-time.

Reference

1. V.Ya. Kosyev. Unified Theory of Field, Space and Time. Nizniy Novgorod, "Arabesk", 2000.
2. New Energy Technologies #1, 2003, p.35.

Medium For Existing of Matter in Nature

Anatoly V. Rykov, Russia
Email: rykov@uipe-ras.scgis.ru

By medium for existing of matter in Nature we understand physical vacuum which is a category determining all spheres of matter "life" (from particles of microworld to gravitational interactions in the Universe). Matter can not exist without this medium. Thus a question emerges, viz: how physical science can be developed without taking this fundamental aspect into consideration?

According to the Einstein postulate, in a void light always propagates with limited speed, which does not depend on the motion mode of a radiating body. It is a statement of deep physical sense which lies in the fact that only some physical medium can possess such a property. Indeed, if velocity of light, which was radiated by a body, further does not depend on speed of this body then it is possible only in a certain medium. For example, in air sound propagates with a certain speed which does not depend on the speed of the source and is determined only by density and volume elasticity of air. Thus, according to the Einstein postulate, there is a physical medium instead of void and velocity of light depends on the parameters of this medium (as is well known, velocity of light is equal to the square root from the product of inverse values of electrical and magnetic conductivity of vacuum).

The author of this article made an attempt to research this physical medium [1]. The author used a well-known experimental fact that at interaction of a photon with a charged real particle this photon transforms to electron-positron pair. Besides, it was taken into consideration that photon is an electromagnetic phenomenon. The research allows the following conclusion: the medium named in physics as vacuum (before the XX century it was called aether) is created by electric dipoles from (+) and (-) elementary charges. These dipoles are surrounded by magnetic (mass) continuum. There were determined dipole intervals, elasticity of decompression and ultimate strength of this interval. What does this model of medium mean?

1. This model is a physical validating for the "birth" of electron-positron pair at energy which is necessary to break dipole and create the "clot" of mass of these two particles. A certain model for the "birth" of mass must exist. It is not clear as yet but can be represented as a vortex of magnetic continuum inside the charge shell of electron and positron.

2. Let us suggest that values of (+) and (-) charges differ at 7.848981x10^{41} Coulomb, then this difference is sufficient for the medium to be a source of gravitation and inertia. According to Newton's law of gravity, a weak electric charge of medium (all material bodies exist in this medium) causes all bodies to attract one another (under Coulomb's law). On the other hand, weak medium charge of a like sign creates repulsive forces which manifest themselves in the form of the expanding Universe. Thus it becomes clear the amazing uniformity of gravity and forces of negative pressure for the Universe. The mentioned difference of values of (+) and (-) charges of medium dipoles is not used as arbitrary parameter but it logically follows from its electrical structure [2].

3. Hypothesis of nature of gravitation is confirmed by calculations of the deflection angle of electric waves by the Sun. Calculated angle differs from the experimental value only in the fifth sign, which depends on the accuracy of several physical values used in calculation formula. There is practically an absolute coincidence with Einstein's theory. The difference is that it is a concept of space and time (i.e. geometry) which prevails in the general theory of relativity while in the nature of gravitation physical basing is used. Velocity of light is unstable and determined by medium state which depends on electric, magnetic and gravitational potentials.

4. The existence of "black holes" relates to medium structure and nature of gravitation. At the edge of "black holes" the ultimate acceleration from gravity is realized. It causes breakdown of connections in electric dipoles of medium, creation of matter and antimatter (the so called "evaporation" of black holes which was theoretically predicted by E. Hoking, England). However at the border of a black hole the velocity of light is equal to zero since its propagation

medium disappears. According to Einstein's theory it is such a deceleration of time that electromagnetic oscillations gain zero frequency.

5. Naturally the medium is a carrier of all types of electromagnetic radiation, beginning at stationary electric voltage and ending with super-high-frequency "photons", which can be determined not as electromagnetic but as magnetoelectric phenomenon. In the latter case magnetic continuum of medium has a crucial importance since it determines very small degrees of magnetoelectric disturbance. These degrees are thousands of times less than those of a hydrogen atom. It is reasonable that such small degrees creates illusion that a photon possesses properties of particles.

6. Medium structure directly leads to the notions of Quantum Mechanics, beginning at quantification of electronic "orbits" in atoms. This medium determines the "allowed" spots for electrons to be placed around a nucleus. Thus medium is a necessary place for the existence of all matter or matter of the Universe. The so called Compton length of electron wave is one of the validations of this statement. It is directly calculated to a high accuracy according to electric structure of vacuum.

7. Thus Plank's constant is not a mysterious "quantum of action", on the contrary it is entirely determined by medium parameters. Thereby the crisis (which conventional physics is accused of) can be logically overcame by the introduction of medium which is capable to take radiation in electromagnetic region only by quanta. Plank's constant is always presented at all quantum approaches. It is additional evidence in favor of the necessity to take into account the medium as natural place for existence of all the matter in Nature.

8. It is still a question what processes take place in the centers of galaxies. The observations show that the centers of galaxies create star matter. They often flow out the centers and settle themselves at approximately the same plane. It is an evidence of the fact that centers of galaxies quickly rotate and the favorable conditions for stars to reject matter are created. Similarly planetary systems are created around the rotating stars. It is thought that centers of galaxies are gigantic black holes. In the context of the concept of medium for matter existence it can be set up a hypothesis that being at some special state the medium creates stellar systems, i.e. galaxies.

9. Astrophysicists more and more trend to accept the existence of unknown "dark" matter, which occupies approximately 70% of all matter. It is said that due to antigravitation property this dark matter is responsible for the expansion of the Universe. Hypothesis on the nature of gravitation gives the affirmative replies to such suggestions as: expansion of the Universe is caused by weak electric charge of the medium and distributed mass of magnetic continuum is the very dark matter.

10. All elementary particles (electrons, positrons, mesons, protons, neutrons etc.) are in the medium and interact with it. In the case of electrons and positrons this interaction causes the great extension of dipoles of medium which directly adjoins to borders of particles. This extension helps photons to break the particles. In case of protons, there is such a breakdown of the medium at their borders that protons turn to be dressed in "coats" of virtual electrons and positrons. And medium is shown to be in entire state only when reaching the first Bohr orbit. This interaction of the medium and particles leads to the factors which determine the life time of some of them. Thus life time of a neutron is up to 30 minutes depending on the state which it has while leaving the nucleus. For mesons "the tearing" Coulomb forces, which exist between the medium and particles, are very powerful that makes the life time of mesons very short. However if the particles moves with a high speed relatively to the medium, then these forces noticeably decrease and if the speeds of motion is close to velocity of light then these forces become very small. The life time of such particles noticeably increases. In Einstein's theory it is said about "deceleration" of time. Finally the life time of particles is determined by the inner steadiness and outer influence of the medium. Protons have fantastic binding energy and they cause such great external destructive effect that their life time becomes immense.

11. When particles move in medium with high speeds, there is a concentration of magnetic continuum which is recognized as increase of mass of particles at increase of their speed. Evidently, we can continue the list of physical phenomena which relate to the presence of medium for existing of matter in Nature. However it seems to be enough to admit that medium (physical vacuum, aether) is of a great importance for the very existence of the Universe and for the processes which take place there. "Mechanism" of gravitation and inertia has a special importance since

it is the only thing which can make clear the real mechanism of Nature. It is possible that the methodology, which is accepted in physics, is necessary for more accurate description of phenomena but it is not enough for understanding of the nature of "space" and matter. Research of the medium for existing of matter will be able to meet the conditions sufficiency.

The understanding of vacuum structure, which is created by electric lattice from charge dipoles, surrounded with magnetic continuum, gives an opportunity to control the forces of gravitation and inertia. The medium of vacuum can be influenced by:

1. The radiation which has frequencies coming to the frequency from point #11 of the list of parameters.
2. Electric voltages which exist in vacuum (it is not very promising because of real matter breakdown).
3. Magnetic intensities (magnetic flux density). It is the most promising influence method (1-10 Tesla is enough to compensate terrestrial gravity).
4. Transformations of low accelerations to high accelerations of impact type.

Potential electric and magnetic energy is immense in vacuum medium. Reasoning from energy of one dipole, which is equal to 1.6x10 (-13) Joule, we will derive the value of electric energy equal to about 10 (+31) Joule per one cubic meter of the medium that is equivalent to mass annihilation of 10 (+15) kg!

References

1. Rykov A.V. Foundations of aether theory// RAS, M., 2000, p. 54
2. Rykov A.V. Hypothesis on the nature of gravitation// MSU, M., 2001, pp. 59-63

Spartak M. Polyakov

(09 January, 1931 – 04 June, 2003)

Spartak M. Polyakov, a legendary physicist-experimentalist, known for his research in the field of gravitation, has died. Polyakov was the author of more than 50 research works, among them "Introduction into experimental gravitonics». Polyakov was the creator of the new model of photon. He also conducted numerous research in the field of electrodynamics, gravitational phenomena and superlight connection.

Spartak M. Polyakov graduated from Kishinev State University with degree in "General Physics". For many years he worked at the top secret military engineering department "Istok" (Russia).

Plyakov's alternative physical theory allows explaining such phenomena as magnetostriction and optical magnetization. The basic postulate of this theory states that the speed of propagation of gravitational waves is dozens of times faster than velocity of light regarded by the traditional science as the limit. In his experiments Polyakov demonstrated such methods of generation of gravitational waves as precession of gyroscope and remagnetization of ferromagnetic.

While working for "Istok", Polyakov designed on his own a device which is possibly the first gravitational motor in history. Total specific thrust produced by rotation of gravitating mass comes to 2.5 kg/kWt. This index is close to that produced by the engines of modern helicopters, which is equal to 8 kg/kWt. Practical application and further development of Polyakov's ideas can provide humanity with new kinds of communication, gravitational engines and free energy devices.

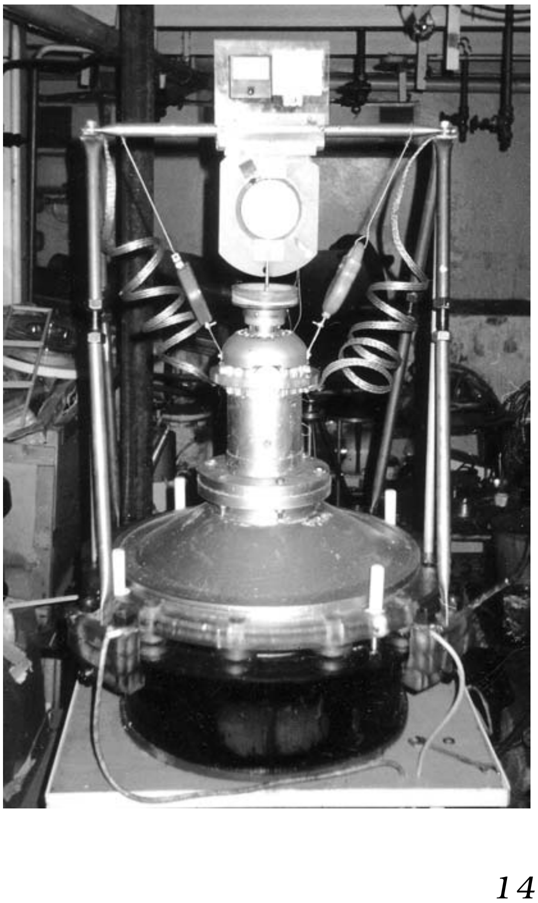

Adams Motor

Principle of Operation and Experimental Data

Sergey S. Abramov, Russia

Email: a_serge@cards.lanck.net

Adams' motor-generator represents the type of devices which use, as their creators claim, so-called "free energy". The term "Zero point energy" is also used in some sources [1, 2]. Due to quanta-mechanical fluctuations, this energy exists even at zero temperature. Adams' motor-generator belongs to the group of Switch reluctance motors [3]. Robert Adams, **former Chairman of the Institute of Electrical Electronics Engineers, USA (New Zealand section)**, designed this machine in the late sixties of 20th century. Similarly to all the devices utilizing free energy (referred to in come sources as "overunity devices"), Adams' device remains practically unknown to the general public. **The device is rather simple to assembled even at home, which I experimentally proved. However, it is necessary to be very careful while choosing the model parameters. As to the latter I have managed to collect quite a big number of instructions from existing sources; the summary of these instructions is presented below in this article. Based on these instructions, a low-power model can be assembled even without a mathematical analysis and modeling of electromagnetic field. Such model would certainly facilitate optimization of the device.**

Adams' motor is most frequently a DC machine; however, it can also use an AC source through a rectifier. In the latter case only the adjustment of the device and its control system can be provided.

Editorial: We disagree with some assertions of the author; however this article is of great interest.

My experiments on my own model do not yet allow me to make an unambiguously positive conclusion concerning the possibility to generate excessive energy. Experiments with my new control system designed on the base of AVR controller AT90s2313-10PI (it is produced by Atmel company (http://www.atmel.com) will allow to be more specific. Below there is a general analysis of the motor principle of operation and a number of recommendations concerning the construction and technology. I do not propose to take this as compulsory rules to follow; other technical solutions are possible.

Basing on the principle described in this article, R. Adams (the link to his articles is available at: http://www.aethmogen.com/wri/intro.shtml) created a few DC motor-generators which operate on permanent magnets. Some of them, according to information found on the Internet, have manifested 690% electrical efficiency and 620% mechanical efficiency. These devices operate at room temperature without overheating. My device has shown between 1 and 3 degrees overheating after an hour of functioning. However, it is easy to prove that such overheating is predictable for an average current of 0.15 A in coils of 35 mm long and 25 mm in cross sectional diameter. I have not been able to prove the data published on the Internet concerning the Adams motor capability to operate when the stator temperature is a few degrees lower than that of the environment. The temperature of the coil and of the power transistor is a good indicator of correctness of the circuit set-up and of functioning of the control circuit. There were cases when transistor and coil were noticeably heated after adjustment. Usually this was explained by a bad choice of points of transistor switching or by too extensive current impulses in the stator (which must amount to approximately 25% of period length). After the required adjustment the motor continued to operate almost without overheating.

Adams' motor was first mentioned in Australian Nexus Magazine in 1992. Later, Harold Aspden (Britain) proposed a slightly improved version of the motor and received Great Britain patent No. 282708 [4], which strongly reminds of the original version published by the above magazine. **Adams' device** represents an electrical motor and/or generator consisting of a rotor with radially directed permanent magnets and of a stator also constructed with a few radially directed and periodically magnetized cores with winding. (Fig. 1). In some models axial orientation of magnets and coils are also used.

Permanently magnetized poles of rotor can include any number of poles, even an odd number. Analogous poles of the magnets (all N poles or all S poles) are directed outside. A version with alternating poles is also possible; such model allows the torque to increase. In this case, after passing a rotor pole the stator is demagnetized by the current impulse and begins to be attracted by the magnet of different polarity. This circuit requires a more complicated control; on the other hand, it manifested rather good results in certain models.

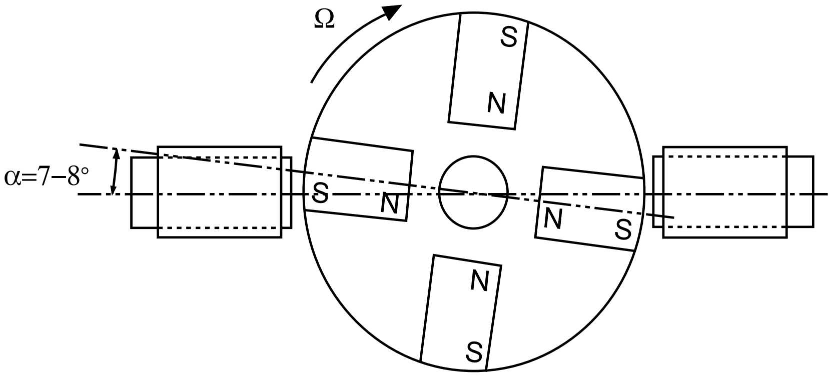

Fig. 1

Poles with winding placed on the rotor are radially oriented in order to obtain a supply of energy emerging as a result of the influence of counter emf from the rotor poles. Steel or iron cores are used for the poles of the stator with winding. It is also possible to use other materials, at that the core must have high magnetic inductivity and low level of magnetization reversal losses. The stator winding consists of a few hundred turns. The current inducted by the magnet in this winding will have the polarity which will cause repulsion of the magnet. Since the electromagnetic state of stator changes quite significantly and quite fast, then the stator core can be considered to be the most crucial element of the device. Ignoring this fact was one of the most frequent mistakes made by those who tried to reproduce Adams motor.

The current inducted in the stator is the function of:

- field size,
- number of winding turns,
- speed of flux changes.

Resultant parameters of this device cause each pole to be attracted or repulsed by stator poles when the rotor is in certain position in case when the rotor is unbalanced. To achieve this effect it is required to switch the input current in control coils after the signal from sensor of rotor position. R. Adams used a mechanical switch as a sensor. My device and a number of other devices use the signal from two Hall sensors. However, according to experimenters' information, better results are achieved if a position optical sensor is used.

Time of switching of impulses is determined by the size of the motor itself, i.e. the speed of motor rotation, location of rotor magnets towards the stator windings and the distance the rotor magnets pass while moving by the poles with stator winding.

It is necessary to take into account that any part of this motor can be modeled based on the existing electromagnetic theory and no part of the motor is in conflict with any laws of electromagnetism. There are so many ways to construct Adams' motor that any version may be considered to be correct.

One can say that the frequently pulsating electromagnetic process in the stator core is what allows Adams' device to function as a kind of diode which borrows energy from the field of permanent magnet but then does not return that energy in full.

5 stages can be defined in the periodical process which takes place in stator:

1. The magnet is attracted to the stator core. The permanent magnet is attracted to the iron core of the stator with winding. While doing so no consumption of electrical current takes place. It is as if kinetic energy is borrowed from an internal ferrite magnetic source and is supposed to be returned into the stator.
2. Stator core is magnetized. During the period when the magnet is positioned in front of stator core they both comprise a single magnetic conductor with an air gap and the stator core becomes an extension of the magnet side it faces. It is usually supposed that the energy "borrowed" on the first stage is getting back now.
3. Stator core is demagnetized. When the stator core becomes rotor magnet extension, the circuit closes and current impulse gets to stator windings. I have seen the instructions saying that the angle between

stator axles and rotor magnet must amount to 7-8 degrees as shown in Fig. 1. However, my model made it clear that at increase of speed of rotation it is necessary to start the coil a little earlier, when the magnet axle has not reached the stator axle. Probably this applies to the circuit with a Hall sensor only and if an optical switching is used the making angle will be different.

Magnetic field of this current acts to compensate magnetization of the stator, which is caused by the field of rotor magnet. Consequently, the summed current significantly compensates attractive force between rotor and stator and the rotor can freely rotate by using the inertia obtained at stage 1. This process is characterized by the fact that this current impulse is amplified by the current inducted in the stator winding by rotor magnet which, in accordance with Lentz law (1834), counteracts the power which induced it. Consequently, kinetic energy obtained as a result of attraction of the rotor to the stator at stage 1 is transformed into electrical demagnetizing impulse in stator winding during the period when the rotor and stator directions are congruent. This is the unique overunity characteristic of this model. However it is obvious that instead of returning this energy the motor transforms it into electromagnetic demagnetizing field.

4. Restoration: when the rotor is removed from the stator attraction zone the latter looses energy and returns to its initial demagnetized state. Decreasing electromagnetic field creates a current wave of reverse polarity which can be stored in the capacitor.
5. Reiteration of the process: This periodical process is renewed as described in stage 1 during the next magnetization of the stator, excluding the fact that emf preliminarily stored in the capacitor, on term of presence of suitable electrical circuit, can be used for facilitating stator demagnetization or even used to supply the load.

It may be briefly summarized that the frequently pulsating electromagnetic process in the stator core is what allows Adams device to function as a kind of diode which borrows energy from the field of the permanent magnet but does not return that energy in full. The important characteristic of such motors is that the stator windings are used for demagnetizing and not for magnetizing as it could seem from the first sight.

It is noteworthy that there is a small pause between attraction to the stator and repulsion from the stator. The effect of attraction to the core takes place a split second before the repulsion effect manifests clearly. This pause being the reason of electromagnetic asymmetry creates conditions necessary for achieving overunity effect. If the attraction to the core and repulsion by means of Lentz currents were taking place simultaneously and with the same power there would not be any overunity characteristics. That is why the rotor must be as lightweight as possible. From this point of view, T. Harwood's model is the most lightweight of the known models. In Harwood's device the magnets are mounted between two CD disks fixed on the shaft by means of plastic washers and glue. My model is heavier, which can be considered as one of its disadvantages.

The principle of operation of Adams' motor is based on the balance, which creates the electromagnetic asymmetry. To get the motor to operate the magnet must be attracted to the stator core which must have a smaller cross-section area in order to create attraction without any significant repulsion effect from the stator windings mentioned above. When stator and rotor axles are congruent the Lentz induced current must be sufficient for compensating the natural attraction of the magnet to the stator core. Consequently, stator windings must have enough turns for demagnetizing effect, but not to the extent that this effect fully manifests before the rotor reaches the stator axle when Lentz current has its maximal value.

During my experiments at 12 V voltage and on using two independently controlled stator coils the speed of rotation reached 3400 rpm. Please note that while reproducing such device it is necessary to take certain measures in order to ensure safety in case of possible breakdown. **The magnet disconnected from the rotor may be dangerous!**

Technological recommendations are as follows:

1. The device must be low power. It is better not to try to begin with a motor functioning in kilowatt range. This is feasible only on condition of having all necessary technological documentation which is not available at the moment.
2. The preferable voltage for the first model is 12 V. If the voltage is less the speed of rotation is too slow for indicating the expected characteristics of the device.
3. The best magnets are ferrite ones with dimensions 4x4x5 (where 5 is the magnet length). The practice has shown that at 12 V voltage neodymium-iron-boric (NdFeB) magnets cause a jerky rotation of the rotor.
4. The side of the stator core facing the rotor must

Control circuit with hall sensors

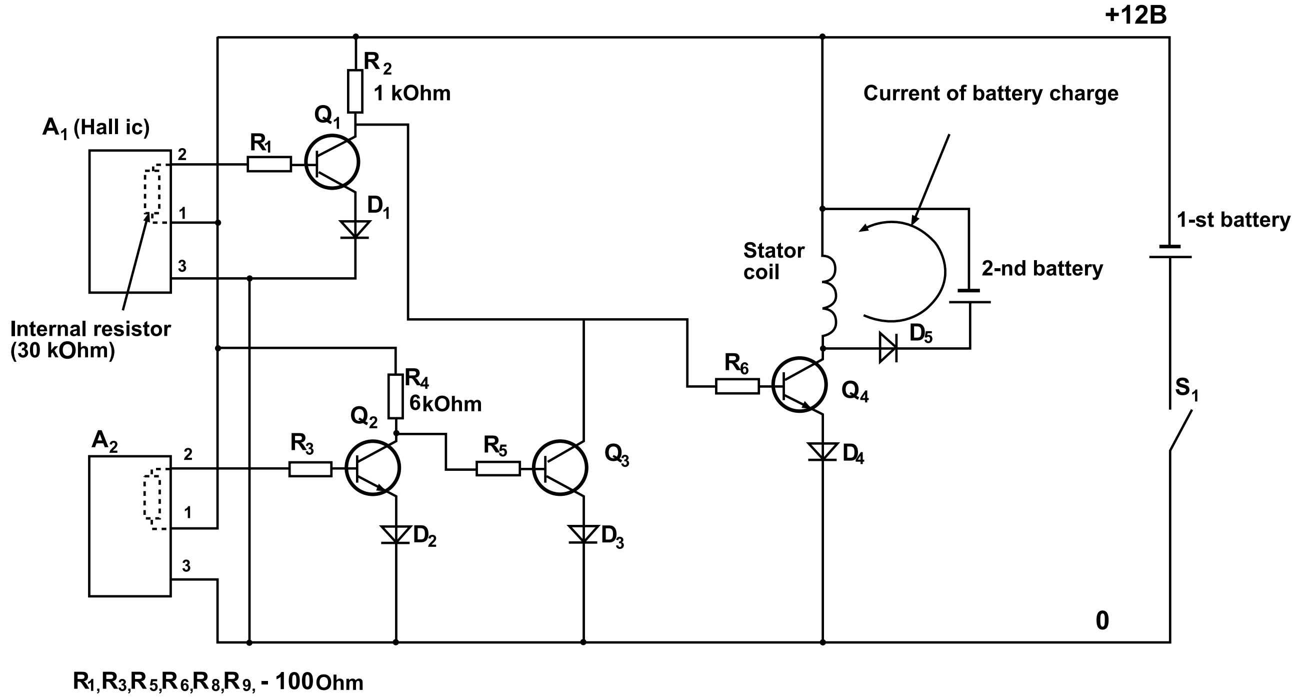

Fig. 2

have the size which is 4 times smaller than the corresponding side of the magnet. If cross-section area of the stator is larger, a bigger part of magnet field comes to the stator core when their axles coincide and thus there is nothing to induce Lentz' currents in the stator windings. It is a common point of view that it is necessary to try to minimize the volume of a device and to obtain the maximum efficiency at minimal material expense. In practice, it has always been the goal while designing electrical machines to decrease losses in stator I^2R. As for the Adams motor, it requires not only the use of disproportionate magnets but also stator windings with disproportionate number of turns specially designed to obtain maximum Lentz' currents that is achieved by hundred of turns of winding.

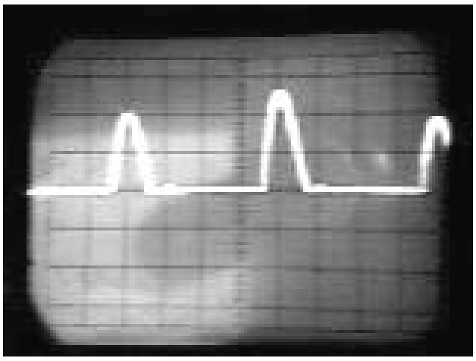

5. It is necessary to define approximately the effective zone of the magnet field. If practically applicable magnet field is equal to, for example, 8 cm, and the stator has a 10 cm long winding, then more than 20% of turns will not be efficiently crossed by magnetic force lines and will only create an excessive mass of the device. To define the degree of effective action of the magnet we can put a paperclip on the table and move it gradually in direction of the magnet until the paperclip is attracted to it. Actually if we take into account friction losses then the magnet influence zone will be a little larger. That is why the stator winding in the axle direction can be 10% larger than that in this experiment. The description of this test has been found on Tim Harwood's website (http://www.geocities.com/theadamsmotor/cdmotor.html).

6. The air gap clearance between the stator and the rotor must not exceed 1 1.5 mm.

7. Use as little metal in the device as possible. It is preferable that the metal is used in the core and stator windings only.

In order to increase the efficiency of this motor, it is necessary to build it in a manner, which allows removing the counter emf from the stator windings. To do that, this emf may be taken off and stored in the capacitor. The article published by Nexus magazine and Great Britain patent No. 282708 consider special generator windings; however, no sufficiently detailed data on operating rules are provided. Michael Smith (Australia, http://www.Fortunecity.com/greenfield/bp/16/content1.htm) has unequivocally informed me in one of his letters that

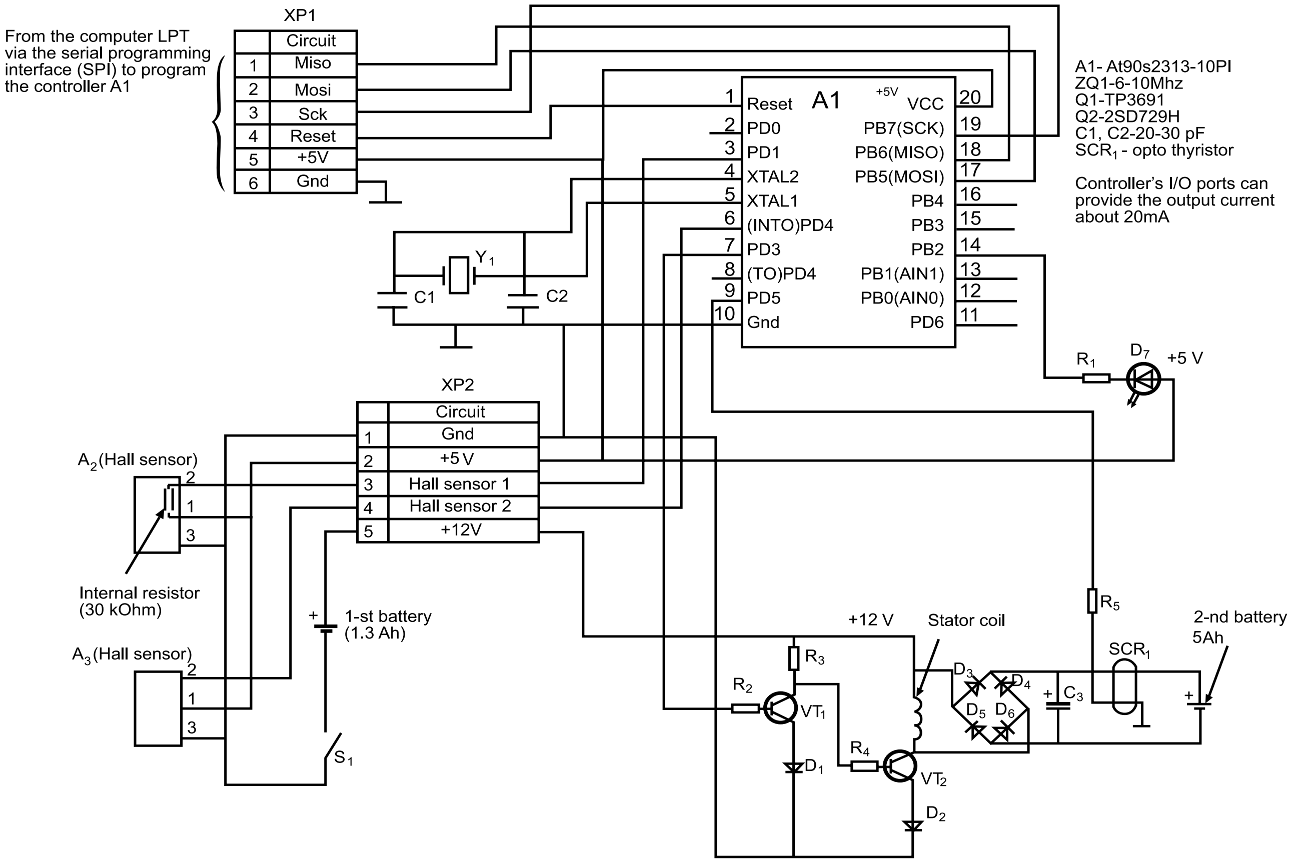
From the computer LPT via the serial programming interface (SPI) to program the controller A1
XP1
Circuit
1 Miso
2 Mosi
3 Sck
4 Reset
5 +5V
6 Gnd
A1
+5V
1 Reset
2 PD0
3 PD1
4 XTAL2
5 XTAL1
6 (INTO)PD4
7 PD3
8 (TO)PD4
9 PD5
10 Gnd
20 VCC
19 PB7(SCK)
18 PB6(MISO)
17 PB5(MOSI)
16 PB4
15 PB3
14 PB2
13 PB1(AIN1)
12 PB0(AIN0)
11 PD6
A1- At90s2313-10PI
ZQ1-6-10Mhz
Q1-TP3691
Q2-2SD729H
C1, C2-20-30 pF
SCR1 - opto thyristor
Controller's I/O ports can provide the output current about 20mA
Y1
C1
C2
R1
D7
+5 V
XP2
Circuit
1 Gnd
2 +5 V
3 Hall sensor 1
4 Hall sensor 2
5 +12V
A2(Hall sensor)
Internal resistor (30 kOhm)
A3(Hall sensor)
1-st battery (1.3 Ah)
S1
+12 V
Stator coil
R5
2-nd battery 5Ah
SCR1
R3
R2
VT1
D1
R4
VT2
D2
D3
D4
D5
D6
C3

Fig. 3

he has not achieved generation of excessive energy in such a system. During his experiments a two-battery circuit proved to be more effective. In this circuit, during a part of periodical process the energy is stored in the capacitor and then a command is transferred to a thyristor, which discharges it into the second battery. At that capacity, the second battery must be no less than 4 times more than the capacity of the first one. Otherwise the excessive energy does not have enough time to be stored during the discharge. I have also managed to provide the charge of the second battery, but the control circuit has not been assembled fully yet, as it is shown below in Fig. 4. Better results are expected when its final adjustments together with AT90s2313 controller program have been made.

In order to increase the efficiency of this motor, it is necessary to build it in a manner, which allows removing the counter emf from the stator windings.

I managed to start my first model of the motor in April, 2002. Afterwards, I spent approximately six months increasing its rotational speed from 750 to 3200-3400 rpm, decreasing vibrations and improving control circuit. In order to achieve a better rotor balance I had to re-construct it twice. It is very important to align the axles of the bearings with maximum precision; otherwise a considerable decelerating torque will manifest. The shaft penetrates the lower moving base. It is possible to rotate the base at a small angle for the precise alignment of axles and then to fix it with screws and nuts.

In order to decrease aerodynamic losses, two veneer parts are mounted between the magnets. Thus I managed to increase the speed, although the rotor weight has also been increased.

Stator cores are made of plates taken from a disassembled radio transformer. Tim Harwood used nails with winding, however, my own results with such core proved to be poor. Dimensions of the core are 10x11x50 mm.

The type of power source is also important. First, I connected a 9 Volt accumulator of "Krona" type in series with three metal-hydride 1.2 V accumulators. The speed of rotation did not exceed 1500 rpm. But when I used a lead/acid accumulator with 1.3 Ah capacity, the speed increased to 2600 rpm if there was one coil on the stator.

Four magnets mounted on the rotor have the dimensions of 20x20x35 mm and are fixed on a 105 mm glass fiber laminate disk. The current impulses proved to be too wide, by up to 40%. In order to make them shorter (down to 25-30%), I had to use the control circuit (Fig. 2) with two Hall sensors. At the signal from the first sensor the stator current is started, the second sensor switches it off. Many experimenters used a timer to control the impulse length while working with Adams' motor that is more practical since impulses are supposed to be wider during the start. I took this factor into account while designing the controller circuit. Current impulses are shown in Fig. 3. Their fronts are supposed to be shorter; it is probable that the coil has more inductivity than necessary. The impulses amplitudes are slightly different which is explained both by the difference in volume of induction of the magnets and by difficulty in achieving similar air gaps while working at home.

I have provided the charge mode of the 2-nd battery at my two-battery design. After 75 minutes of operation of the device the source lost 0.17 Volt whereas the second battery was charged at 0.36 Volt. The capacity of both batteries in this experiment was equal. Besides after such charging the second battery started to discharge quickly. The circuit where stator current charges the battery directly has to be considered ineffectual (see Fig. 1). To evaluate the charge, which is gained in non-hermetic accumulators, the density of electrolyte can be measured.

To obtain a more uniform torque I added the second stator which is controlled independently. This required installing two more Hall sensors and an additional power transistor. The angle between axles of the coils amounts to 135 (180-90.2=135) degrees. When the current is present in one part of the stator it is absent in another and vice versa. The speed has increased up to 3200-3400 rpm, and I deemed the further increasing of this characteristic unnecessary.

Increasing the number of circuit elements does not seem a good idea. Futhermore the adjustment process becomes more complicated. In order to improve the circuit of battery charge adding a timer circuit is required. Thus, I decided to use a controller circuit. A simple Basic program has been developed for AVR controllers. This program operates in mode similar to transistor circuit, but its capability can be significantly increased due to the built-in processor timers. Today program improvement is the most efficient way to solve the task of generation of excess energy in this circuit.

This article is meant to elucidate the principles of operation of one of the simplest devices which pretend to become an "overunity device", "free energy machine", "perpetual motion machine", whatever you call it. Probably someone will try to create such motor-generators on their own. I hope that my article will serve as a guide and will help to avoid the mistakes made by many experimenters (including myself) before they managed to build their own model.

References

1. Puthoff H.E. Source of vacuum electromagnetic zero-point energy. Phys. Rev. A 40.4857-4862; also 44. 3385-3386
2. Bernard Haich, Alfonso Rueda, Puthoff H.E. Physics of the zero-point field: implications for inertia, gravitation and mass. Speculation in Science and Technology, Vol.20, p.p. 99-114, 1997 (preprint version)
3. Miller, Switched Reluctancy Motors & Their Controls, ISBN: 10881855-02-03
4. Patent No. GB 2,282,708 (from NEN, Vol. 4, No. 8, December 1996, pp. 1-7).

INFINITE ENERGY
Cold Fusion and New Energy Technology
Including Conventional and Renewable Energy

Infinite Energy Magazine

Infinite Energy
The "Lifter" Phenomenon

* Cold Fusion * New Energy
* New Science
* New Technology
Subscriptions, 6 Issues Per Year
$29.95 North America
$49.95 Foreign
Single Sample Copy
$5.95 North America, $10.00 Foreign

Infinite Energy Magazine
P.O. Box 2816-FV
Concord, NH 03302-2816
Phone: 603-228-4516 Fax: 603-224-5975
http://www.infinite-energy.com

Prediction and Experiment on Field Transformation in the Model of Extended Space

Dmitry Yu. Tsipenyuk, Russia

119991, Vavilov str., 38, Moscow
E-mail: tsip@kapella.gpi.ru

A series of preliminary experiments on checking the possibility of generation of gravitational field at deceleration of charged massive particles in matter was carried out.

Introduction

In previous works [1-3] generalization of the special theory of relativity (STR) for the five-dimensional extended space with metric (+;-,-,-,-) was offered.

The model of extended space (ESM), combining electromagnetic and gravitational interactions, was made. For this, there was made (1+3)-dimensional space extension $M(T;\vec{X})$ of Minkovsky to (1+4)-dimensional space $G(T;\vec{X},S)$. Let us call it extended space. As the 5-th additional coordinate the value already existing in Minkovsky space, i.e, S interval

$$s^2 = (ct)^2 - x^2 - y^2 - z^2 \qquad (1),$$

is used.

Let us note that attempts to combine gravitation and electromagnetism have a substantial background.

Modern approaches to this problem trace back to the work of F. Klein [10] in which he proved that classical Hamiltonian mechanics can be represented as optics in the space of a great number of dimensions.

Then T. Kaluza tried to generalize Einstein's theory of gravitation to include electromagnetism in this theory as well [11]. He proposed to consider (1+4)-dimensional space with metric depending on potentials of the electromagnetic field. Kaluza's idea was evolved by O. Klein [12], G. Mandel [13] and V. Fock, and the model they had created got the name of the Kaluza-Klein theory. They proved that the trajectory of a charged particle has the form of a geodesic line with zero-length in 5-dimensional space.

In his works on 5-optics Y. Rummer [15] proposed to assign action dimensionality to the new dimension and to consider it periodical with the period equal to Plank's constant. Note that rest mass of particles in all these constructions unlike the model of extended space evolved in the works [1-7] was considered a fixed value. Subsequent development of multidimensional theories is given in the monograph [16].

A separate approach is represented by multidimensional constructions in the theory of strings and superstrings [17].

Approach to construction of (1+4) dimensional space evolved in [18] is close to the proposed model of extended space. Here it is proposed to use mass (matter) as the 5-th coordinate. However, in this model, as its originators admit, it is impossible, for example, to create the energy-pulse tensor. There is no such disadvantage in the extended space model [8].

Mechanics of a material point [1, 2, 7] and electrodynamics [1, 8] were made in the introduced extended space. Besides Lienar-Vihert potentials [6, 19] corresponding to such a model were also considered and properties of solutions of Maxwell augmented system of equations which are in conformity with these potentials were analyzed there.

Gravitational effects in extended space, such as the second orbital velocity, red shift and light deflection [4, 20, 21] were considered. It is proved that the formulas received in the general theory of relativity for calculation of values of these effects can be received by an absolutely different method within the framework of the extended space model as well.

It was proved [6, 19] that the fields in the model of extended space can change their signs. Such change of field-intensity sign and, consequently, change of sign of Lorenz force can be associated with radiation reaction of these fields which occurs when charged particles move with acceleration.

Thus, on the one hand, it was proved in the model of extended space that it is possible to get certain formulas describing gravitational effects of the general theory of relativity [4,20] using the technique of turns in extended space. On the other hand, it was proved that the electromagnetic field can be a source of gravitational field [6, 19].

Besides, a moving massively charged particle under deceleration can create a variable gravitational field around itself [9, 6, 19]. The following experiment was offered for experimental check of the latter assumption. In this experiment probable occurrence of gravitational field at deceleration of relativistic electrons was determined by change of oscillations of a massive torsion pendulum.

Experimental device

A narrow bunch of relativistic electrons from a microtron 1 (average power of the bunch is 450 Wt, electrons energy is 30 MeV) was directed to a brake target (position 2 or 3) made of wolfram exactly where deceleration of accelerated electrons took place.

A special torsion pendulum suspended on a vertical suspender 5 made of a springy metallic string with 1,8 mm diameter was placed near the brake target to register gravitational field which could probably appear at electrons deceleration. The length of the suspender made is 85 cm. The pendulum could rotate freely on the suspender only in horizontal plane.

The pendulum consisted of a light aluminum rod 4 (with a length of 120 cm) on the ends of which massive loads 6 and 7 made of non-magnetic material were fixed. The weight of each load was equal to 4kg. In the center a pendulum was fastened to a vertical suspender 5 by a special mounting preventing slippage during turns. To reduce the influence of magnetic inducings the pendulum was grounded and additionally screened by metallic grid from all sides. The period of free oscillations of the pendulum made were about 40 s.

Rigidity of the pendulum vertical suspender could be changed by means of limiting the length of effectively operating part of the suspender. As a result, the period of oscillations could be continuously changed within the limits of 40 to 27 s.

To reduce the influence of mechanical noise and to introduce additional attenuation in pendulum oscillations two liquid dampers 10 and 11 located near the pendulum massive loads were used.

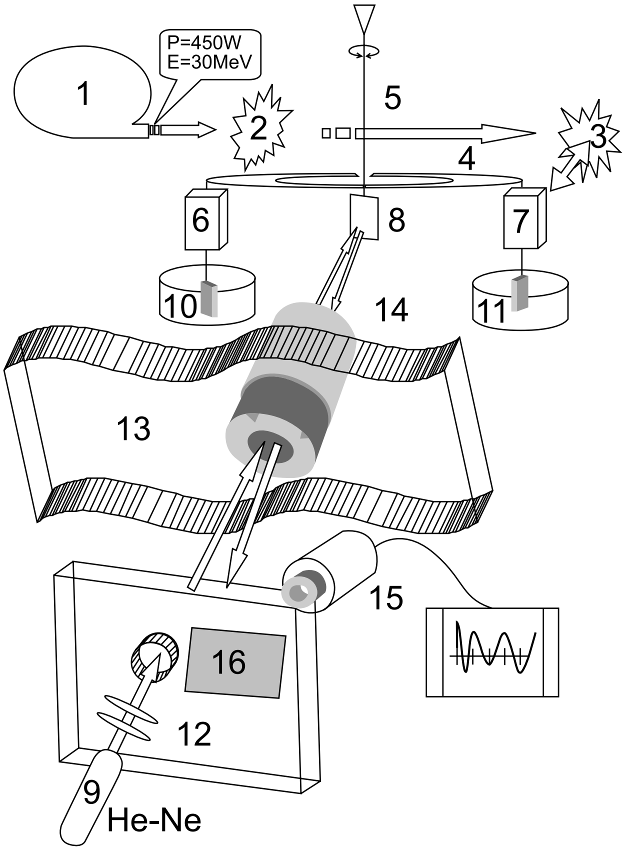

Fig. 1

Experimental device

1 – microtron, brake target made of wolfram – positions 2 or 3, 4 – rod, 5 – vertical suspender, 6 and 7 – massive loads, 8 – mirror, 9 – He-Ne laser, 10 and 11 – liquid dampers, 12 – optical system, 13 – concrete protection, 14 – observation channel, 15 – video system, 16 – screen.

Pendulum deflections were observed on a graduated screen by deflection of a laser beam reflected from a flat mirror 8. For this, the beam from a continuous He-Ne laser 9 through the optical system 12, which constricts the divergence angle of the laser beam, was directed to the mirror through a special narrow channel 14, located in concrete protection 13 around the microtron. By means of a video system 15 the beam reflected by the mirror was registered on the screen 16 located at a distance of 500 cm from the mirror. The video system allowed remote checking of vibrations of the laser spot and additionally enlarged the visual angle up to 12 times. The diameter of a focused laser beam on the screen was made 0.15 mm. The maximum turn angle of pendulum for the reflected beam to remain within the receiving channel was approximately 2 degrees. The accuracy of turn angle registration of the whole system was 5x10^{-4} degrees.

The pendulum was placed in such a way that one of the massive loads were close to the brake target at a distance of about 20 cm. There was also an opportunity to move the brake target from one end of the pendulum (position 2) to the other (position 3). This allowed changing the place of deceleration of electron bunch at constant parameters of all unaccounted mechanical noise and magnetic inducing. Thus, direction of pendulum torsion was changed under probable exposure of occurring gravitational radiation.

Below there is a photo of the experimental plant (see also the colored photo on the cover page).

Calibration Measurements

To make measurements it was necessary to select optimal parameters of the pendulum (masses of loads, suspender rigidity and the value of oscillations attenuation). On the one hand, while carrying out the measurement it is desirable that the amplitude of pendulum oscillations should be as maximal as possible. On the other hand, the beam reflected from the mirror should not go outside the limits of observation, restricted by the diameter of the narrow observation channel in radiation protection around the accelerator. Besides, the typical operating time of the loaded accelerator usually is 10-15 minutes. Necessity to accumulate the minimum of sufficient statistical data within this time limited oscillations period and the time of setting of the pendulum in a new equilibrium position at outside influence. All these requirements were as far as possible taken into account during selection of the final setting parameters.

An example of free oscillations of the pendulum in the presence of minor mechanical vibrations caused by the operating of vacuum pumps is given in Fig. 2, series I (the experiment took place on 31 May, 2001). The diagram shows the amplitude of laser beam oscillations on the screen 16 (upper and lower rows of values) depending on the number of oscillation. The laser beam is reflected from the mirror 8 which was fixed to the pendulum. The diagram also shows the current central equilibrium position (the central row) calculated by these amplitudes. Series I represents oscillations at influence of the background mechanical noise. Series II represents the response of the pendulum to minor permanent outside force. Accuracy in determination of position of the center of a light spot was 0.1 mm.

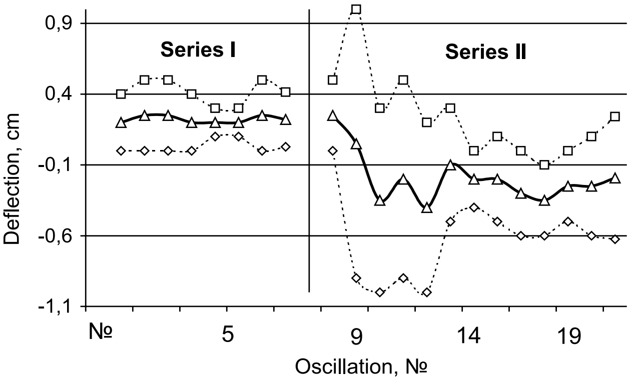

Fig. 2

Calibration Measurement 31 May, 2001

Measurement was carried out with one liquid damper in service and with increased rigidity of suspender (it was made by means of limiting the effectively operating suspender length). The period of free oscillations in these series was 29 sec. In this case there were set continuous oscillations of the pendulum around the average value of balance of 2.2 mm with the average amplitude of values fluctuation of about 0.2 mm.

Air cooling of one of the massive loads by a very light continuous air flow was carried out to study response of the pendulum to a minor constant external force. In this case (Fig. 2 Series II) noticeable change of pendulum oscillations already took place after 3-4 periods. Absolute setting of a new balance took place after 7-8 oscillations.

In the case of another series of calibration measurement (held on 7 June, 2001) both liquid dampers were used and rigidity of suspender was decreased. The period of free oscillations of pendulum was about 40 sec.

Addition of the second damper and decrease of suspender rigidity caused, on the one hand, increase amplitude of pendulum oscillations at influence of the external force and, on the other hand, in that case noticeable change of the equilibrium position of the pendulum took place after 1-2 oscillations.

Periodical checking of the invariability of initial central position of pendulum balance in time was also carried out. Thus, for example, in the series of measurement of 7 June 2001 measurement of equilibrium position was checked not only before the start of the main series of measurement, but also 2 hours after completion of the main works.

Experimental Results and Their Analysis

Within the period from 17 May, 2001 to 7 June, 2001 7 measurement series were carried out at various operation modes of the accelerator and at various pendulum parameters.

A record of check of equilibrium position of the pendulum was carried out before and after switching off the electron bunch (as well as during calibration measurement). At the same time, all electric inducing and mechanical noise remained stable within the whole measurement period. This was achieved by means of additional switching on all the devices which were used during measurement (water and vacuum pumps, magnetron, deflecting magnets, etc.) and their switching off only on completion of the measurements.

In Fig. 3 there are results on measuring the central position of the pendulum when brake target is in position 3 (see Fig. 1). Series I and III on the diagram correspond to check measurement directly before switching on and several minutes after switching off the electron bunch. Series II-A and II-B totally reflect pendulum oscillations at the time when the accelerator is operating (for about 10 minutes) and for some time after the bunch is off. Additionally, a trend line is drawn (averaging by 3 points).

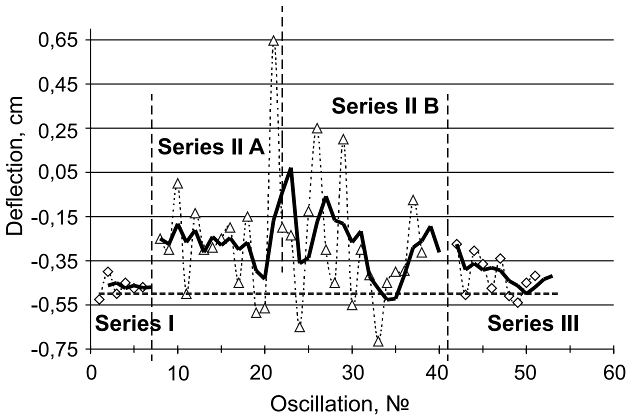

Fig. 3

Measurement of central position of pendulum equilibrium when brake target is in position 3.

Results of a similar experiment are given in Fig.4. The only difference in this experiment is that the brake target is in position 2. Also a trend line is added (averaging by 3 points). Series I and III are check measurements made directly before switching on and after switching off the electron bunch. Series II are pendulum oscillations at the time when accelerator is operating.

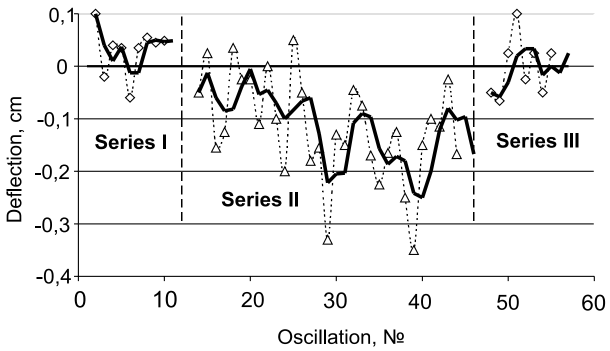

Fig. 4

Brake target in position 2

From qualitative comparison of trend lines (Fig. 3 and Fig. 4) it can be derived **that there is a correlation between switching on the electron bunch and the average deflection of the pendulum from equilibrium position if compared to checked series before and after switching on. At that, direction of deflection changes depending on what pendulum load the brake target is close to.**

Unfortunately, due to circumstances beyond the control of the author it seems to be difficult to improve experiments accuracy or to accumulate much of statistical data by now. Estimate of the value of the force which may cause such a shift of position of pendulum equilibrium was carried out. In the experiments this deflection did not exceed 1-2 mm (in the units of registering scale). Calibration of a rigid suspender (if it applies to a massive load at a pendulum end) gives the upper boundary of this force of not more than 10^{-6} N.

Conclusion

A series of experiments were made on checking the possibility to generate a field at deceleration of charged massive particles in matter.

Electrons accelerator was used as a source of charged particles. A narrow bunch of relativistic electrons (average beam power is 450Wt, electrons energy is about 30 MeV) was directed at a brake target made of wolfram where deceleration of accelerated electrons took place.

Measurement proved appearance of statistically reliable deflection of a torsion pendulum, one of massive loads of which was located close to a brake target by the time of deceleration of relativistic electrons.

Change of direction of pendulum torsion at shift of a brake target from one end of the pendulum to the other was also registered. The value of the force which causes pendulum deflection has the upper boundary of N.

Of course, these first experimental results on checking the predictions made on the basis of development of the model of extended space are of preliminary nature and need more thorough checking. That will be the basis of future experiments.

References

1. Tsipenyuk D. Yu. Andreev V. A, “Extended space and model of combined interaction”, Short reports on physics FIAN, No.6, pp. 23-24, (2000); gr-qc/0106093
2. Tsipenyuk D. Yu., Andreyev V. A., “Explored in Russia” E-magazine,60, (1999); http://zhurnal.ape.relarn.ru/articles/1999/060.pdf
3. Tsipenyuk D. Yu., Andreyev V. A., “Structure of extended space”, Preprint IOFAN, 5, 25p., Moscow, (1999)
4. Tsipenyuk D. Yu., “Field transformation in the model of extended space: prediction and experimental test”, Gravitation&Cosmology, Vol.7, No.4(28), pp336-338, (2001); physics/0203017
5. Tsipenyuk D. Yu., “Field transformation in the extended space model: prediction and experimental test”, Short reports on physics FIAN, No, 7 p.39-49, (2001); physics/0107007
6. Tsipenyuk D. Yu., Andreyev V. A., “Lienar-Vihert potentials for the extended space model”, Short reports on physics FIAN, No 6 p.3-15, Moscow,(2002); physics/0302006
7. Tsipenyuk D. Yu., “Explored in Russia” e-magazine, 81, p.907-916, (2001); http://zhurnal.ape.relarn.ru/articles/2001/081.pdf
8. Tsipenyuk D. Yu., Andreyev V. A., “Electrodynamics in extended space”, Preprint IOFAN, 9, 26p., Moscow, (1999)
9. Tsipenyuk D. Yu., Andreyev V. A., “Interaction in extended space”, Preprint IOFAN, 2, 25p., Moscow,(2000)
10. Klein F. Zeits.f.Math.a.Phys., 375, (1901) (see translation in collection “Variation principles of mechanics”, M.: Physmatgis, 1960)
11. Kaluza Th. Sitz.d.Preuss.Akad., 966, (1921) (see translation in collection “Albert Einstein and gravics”, M.: Mir, 1979)
12. Klein O. Zeits.f.Phys., 37,895, (1926)
13. Mandel H. Zeits.f.Phys., 39,136, (1926)
14. Fock V. Zeits.f.Phys., 39,226,1926
15. Rumer Yu. B. Investigations on 5-optics, M., Gostehizdat, 1956
16. Vladimirov Yu. S. Dimensionality of physical space-time and integration of interactions. M., MGU, 1987.
17. Green M., Schwartz G., Vitten E. Theory of superstrings, v.1, 2, M.: Mir, 1990
18. Wesson P.S. Space, time, matter: Modern Kaluza-Klein theory. Singapoure, World Scientific publishing Co.Pte.Ltd, 1999
19. Tsipenyuk D. Yu., Andreev V.A., “Lienar-Vihert potentials and Lorenz force in extended space”, Preprint IOFAN,1,20p., Moscow,(2001)
20. Tsipenyuk D. Yu., Andreev V. A., “Gravitational effects in extended space”, Preprint IOFAN, 4, 15p., Moscow,(2001)
21. Tsipenyuk D. Yu., “Field transformation in the model of extended space: prediction and experimental test”, Preprint IOFAN, 5, 24p., Moscow, (2001)
22. “Tsipenyuk D. Yu., report “Electromagnetic and gravitational interaction in extended space” at the Conference of Nuclear Physics Department of the Russian Academy of Sciences (RAN) “Physics of Fundamental Interactions”, 27 November-1 December 2000, ITF, Moscow, http://www.itep.ru/
23. Tsipenyuk D.Yu. and Andreev V.A., “Interval as the fifth coordinate” proc. of 5-th Int. Conf. on Gravitation and Astrophysics of Asian-Pacific Countries, PFUR, Moscow, 2001, pp.29-30. http://rgs.da.ru/
24. Tsipenyuk D.Yu., “Direct transformation of the electromagnetic field to gravitational field in the model of extended space: predictions and possible way of its experimental test”, proc. of 5-th Int. Conf. on Gravitation and Astrophysics of Asian-Pacific Countries, PFUR, Moscow, 2001, p.28. http://rgs.da.ru/
25. Tsipenyuk D.Yu., “Model of extended space: Prediction and test experiments on possible transformation of electromagnetic field into gravitational field”, proc. of XVIth Workshop on High Energy Physics and Quantum Field Theory (QFTEP), Moscow, Russia, 2001, pp.398-405. http://theory.sinp.msu.ru/~qfthep/2001/Proceed2001.html
26. Tsipenyuk D.Yu., “Model of extended 5-d space and possible checking experiments”, presented at 5-th Moscow international ITEP school of physics (30-th ITEP winter school of physics), 20-28 Febr., Moscow, 2002, http://www.itep.ru/
27. Tsipenyuk D.Yu., report “Second experiment on checking prediction on field transformation in extended space” at the scientific conference of nuclear physics section OFN RAN “Physics of Fundamental Interactions”, ITF, 2-6 December 2002, Moscow, http://www.itep.ru/
28. Urbah V. Yu., Mathematical statistics, M., USSR Academy of Sciences Publishers, 1963.

Joe Flynn's Parallel Path Magnetic Technology

Tim Harwood, USA

Email: timharwood@usa.net

There is a widespread opinion common to the mainstream academic community and also to various alternative scientific forums, that some kind of exotic new physics will be required to design and implement over-unity technologies. That is to say electrical motors, electrical generators, or other apparatus, which produce an excess of magnetic force or energy above the value actually inputted.

However, this has recently been experimentally demonstrated not to be the case, as I have validated myself in simple $20 experiments undertaken at home with parts bought from the local hardware store. It is the purpose of this article to educate readers that with care, thought, and a little work, it can be demonstrated that existing textbook physical law, freely allows for the extraction of excess electrical energy from magnetic systems.

The Flynn Research Project

Joe Flynn has been engaged in magnetic flux research for over 25 years now. His work is long standing, comprehensive, and in later years, well funded. It is reported $7m has been spent to date, with over $1m alone developing a revolutionary high performance magnetic motor. His equipment is validated, and apparently already in mass production for select customers. Yet few researchers have heard of Joe Flynn. This surprising situation will hopefully shortly be changed, and Joe Flynn awarded the scientific accolades he deserves, for being the genius master mind behind one of scientific history's most outstanding research projects. Since many lines of research have been formulated and explored by Joe Flynn, the following article presents only a brief summary of some of his best art apparatus, but is nonetheless sufficient to convey the basic ideas.

The first illustration (Fig. 1) is taken from Joe Flynn's US patent 6,246,561, and explains a simple magnetic force multiplication experiment, which forms the basis for the Flynn magnetic art. If the windings on either side of the central magnet, which are normally connected in series, are properly pulsed, the field of the permanent magnet in the center will be diverted to the opposite side of the core flux path provided. Or in alternative language, the side of the core that is pulsed is demagnetized, relative to the field of the permanent magnet used in the apparatus. This is elementary textbook physics anyone can understand.

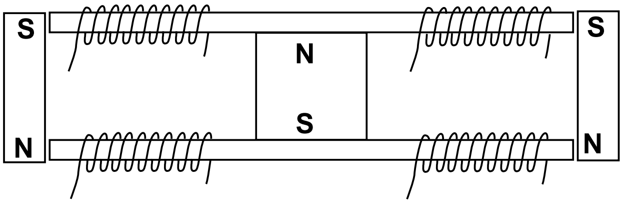

Fig. 1

1.75 times more force is delivered to the legs of the core than is provided by the electrical input to the control coils

So what is surprising about this apparently simple apparatus is that the armature on the side of the flux core will contain 1.75 times more units of magnetic force, that could be manifested by the electrical input to the apparatus alone. Since the ability to arbitrarily move the force from one point to another is the basis for motion or work, however simplistic, we therefore have a basis for a system that can be developed for practical technological purposes. Expressed in alternative language, we also have the capability to engineer a time varying magnetic field, without the need for moving parts, which will allow development of systems that output electrical energy. Both capabilities are highly desirable, and offer substantial opportunity for technical development.

Following on from this basic experiment, there is a second simple and logical improvement in layout illustrated in Fig. 2, which should be obvious, but has been shown not to be the case. In this instance, the pulse is centrally located, and a dual flux field layout employed, which both demagnetizes the core relative to one magnet, and magnetizes it relative to the other. Since the two actions are complementary, the input required to manifest the flux switching effect is halved, therefore doubling 'efficiency.'

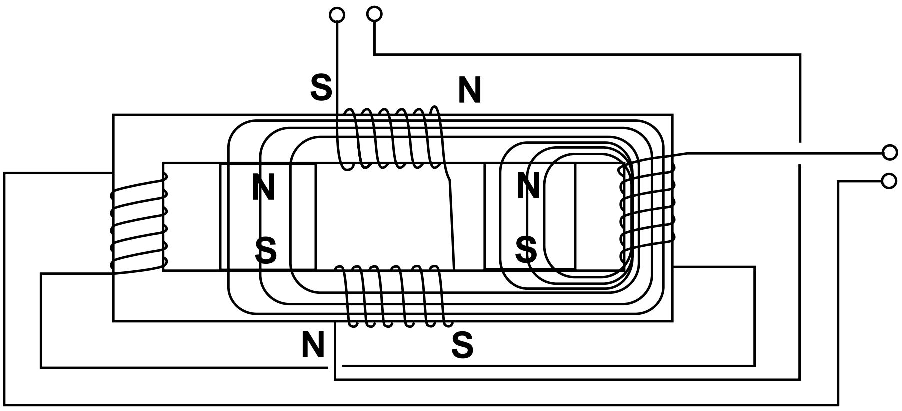

Fig. 2

3.47 times more force is delivered to the legs of the core than is provided by the electrical input to the control coils

It should be noted that while the efficiency is doubled, the absolute output may not be significantly improved. This is because the major weakness of this effect and technology is flux saturation of the core, with values depending upon the specific properties of the B-H curve of the core material employed, limiting the absolute output of both layouts the same.

The previous statements are not required to be taken on trust, and simple experiments have been proposed by Joe Flynn, such that anyone can validate this effect for themselves. Figure 3 is a simple experiment taken from the Flynn website (http://www.flynnresearch.net), that can be used to validate the principals put forth in this article.

Simple Magnetic Force Multiplication Experiments

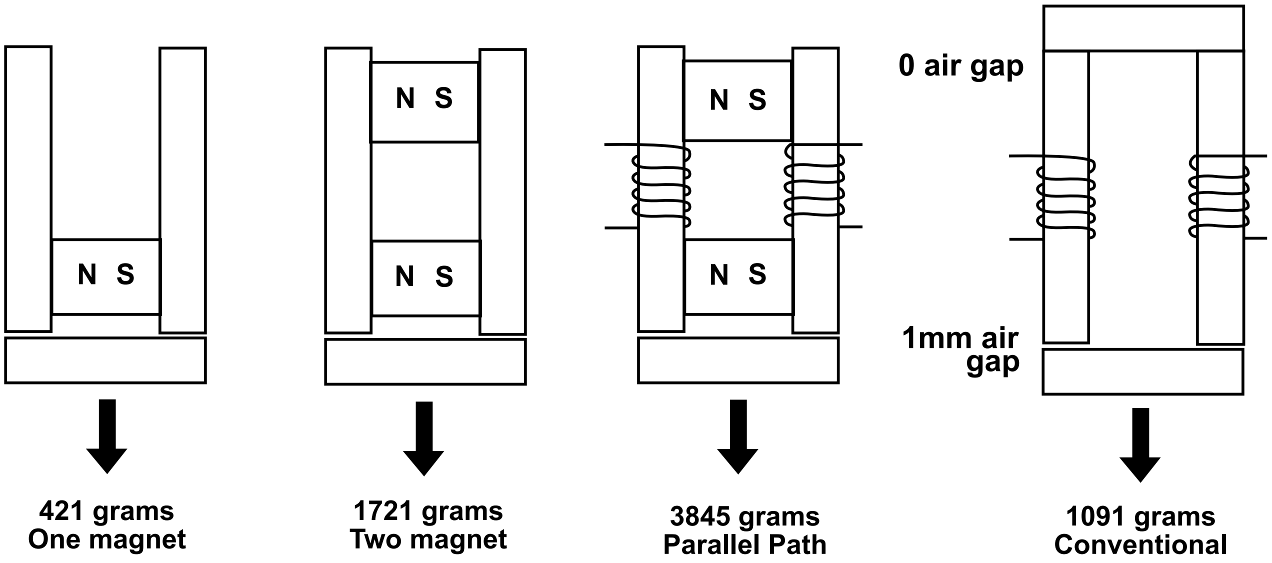

Fig. 3

Parallel Path can deliver 3.47 times more force to the legs of cores than any competing conventional technology

An even simpler non electrical flux experiment was proposed by GM in the Parallel Path Egroup. My apparatus is illustrated below in Figure 4. It is no more than magnets and steel staple strips, bought from a local hardware store for a total of under $10. The Parallel Path effect can be replicated with identical apparatus, at only a slight increase in cost and complexity, with the addition of a simple 12V polarity reversible power supply, such as those commonly sold to power computer speakers, among other applications.

Simple Magnetic Force Experiments

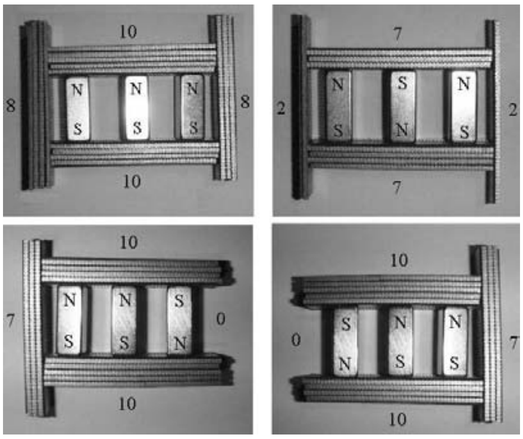

Fig. 4

Relatively small changes in layout produce large changes in force

Conservation of Energy / Field Potential

One of the aspects of the Flynn technology people find most difficult to understand, is how you can have a device that delivers 3.47 times more magnetic force than is electrically inputted, yet not violate accepted principals of text book physics, as stated in the introduction. The key point here is conservation of energy. Magnetic fields do not gain energy – they are conservative. You can only ever obtain less energy from a magnetic field minus losses, than is in fact present. I feel this apparent puzzle can not be better explained, than by reference to Joe Flynn's own words:

"Since the Parallel Path System produced 3.47 times more force than the conventional system, with the same electrical input, it appears to violate conservation, this is only true when observed from a traditional view point. The system contains three flux producing sources (2 magnets and an electromagnet) which together are capable of producing a far greater force than is actually produced. All of the flux sources together can produce a force of 13.11 units, therefore in the physical sense a loss of 1 - (9.01 / 13.11) = 31% is realized."

So the system is 350 % efficient, in terms of delivered magnetic force compared to net electrical input, yet still conforms to the accepted physical principals of energy conversation, by being only 69 % efficient, in terms of the fields present in the system. However surprising this result may appear, the analysis presented is in outline correct, with the difference between fields present in the system, and net electrical input, being the important concept presented.

Losses in the System and Optimisation

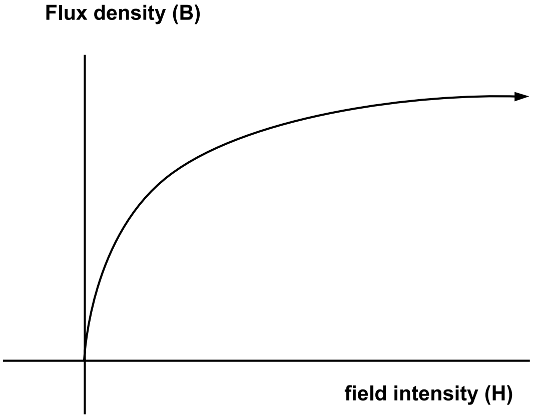

Fig. 5

In order to properly optimise flux cores, an appreciation of the physics that underlies the transfer of flux within a core is required. The normal magnetization curve, or B-H curve, is a mathematical relationship between applied field intensity H, and resultant flux density manifested in the core B. It varies according to core material, and the curve will shift, if there is a starting magnetism within the core, such as that provided by the field of a permanent magnet. If the start magnetism is excessive, the core is saturated, and will not properly respond to the applied force H. A simple B-H curves is illustrated in Fig. 5.

Hysteresis is a delay between applied magnetic force H, and resultant flux density B, that again varies according to material type. It also manifests as a delay between the termination of force H, and the manifestation of flux density B. So, the system will not turn on instantly, and will not turn off instantly, in simple terms. This is because the magnetic memory of the core, means a flux vector remains within it, even when the application of magnetic force H has been terminated. If we apply a reversed force H to the core, the basic B-H curve is now expanded as in Fig. 6, with the memory effect also illustrated.

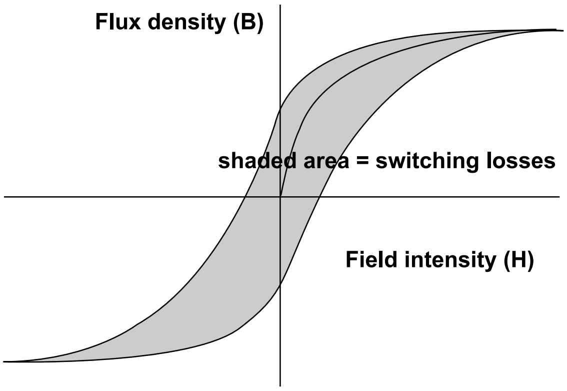

Fig. 6

Thus as can be seen, to return to the initial switched state, the remnance magnetism must now be overcome, hence input once in operation, will be greater than that required for the very first pulse. The area within the hysteresis curve gives a rough estimate for the amount of wasted energy, and along with other conventional sources of losses resultant in flux transfer within a core, is what reduces the efficiency of flux cores from maximum values of 2, or 4, down to values such as 1.75 or 3.47, typically.

Motor Apparatus

Simple Parallel Path Technology Demonstrator Motor

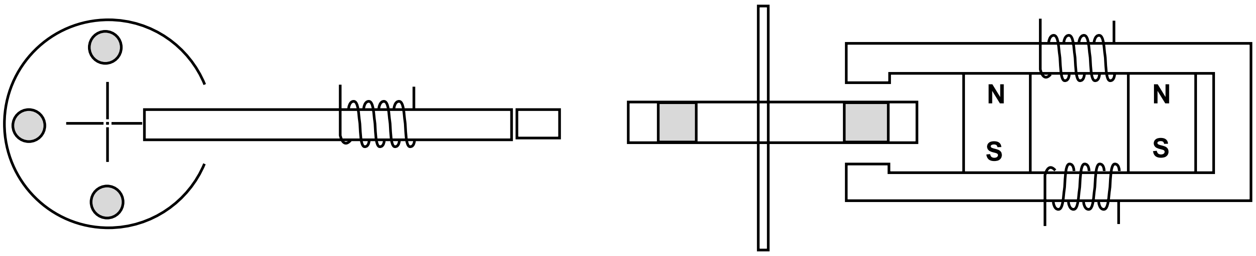

Fig. 7

Designed to demonstrate core principals, not provide over-unity

Although numerous practical applications abound for this effect, electric motor design remains the most outstanding opportunity. To this extent, again a few simple images, should be sufficient to explain how the basic flux switching apparatus, can be turned into a highly efficient electrical motor.

The first motor shown in Fig. 7 is one I have proposed to validate the flux switching effect at a most basic level. It illustrates the point made in the Flynn patent, that the armature of the core can be removed, and replaced with a motor flux path. This first motor is not claimed to be highly efficient, but it helps one to understand how the transition from simple flux core to motor takes place.

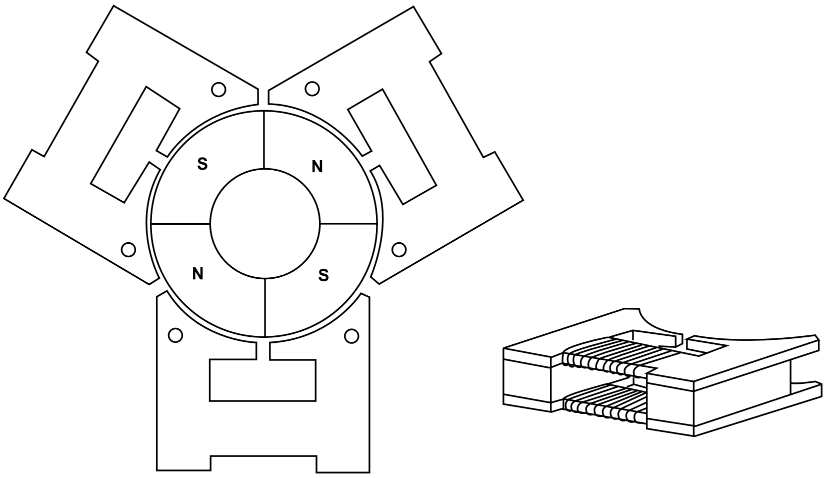

Fig. 8

Magnetic force is alternately switched from one leg to the other, imparting motion to the rotor sections

The next motor shown in Fig. 8 is again taken directly from the Flynn patent, and illustrates the next intermediate step to motor design. The fields of the permanent magnets are alternatively switched from one side of the surrounding flux cores to the other, alternately interacting with N and S poles on the rotor, imparting motion to the central rotor shaft.

With proper financial support, and the facilities to have metglas cores custom moulded, Joe Flynn was able to develop his final best art, shown in Fig. 9. No detailed performance numbers have been released for this motor, whose precise performance characteristics remain proprietary to Joe Flynn at this time. But the optimisation is so expert, it is stated to posses certain exotic properties, such as cool ambient operation, even during prolonged periods of continuous load. This 'cold running' is said to be of great interest to the American military, as it offers excellent stealth performance characteristics.

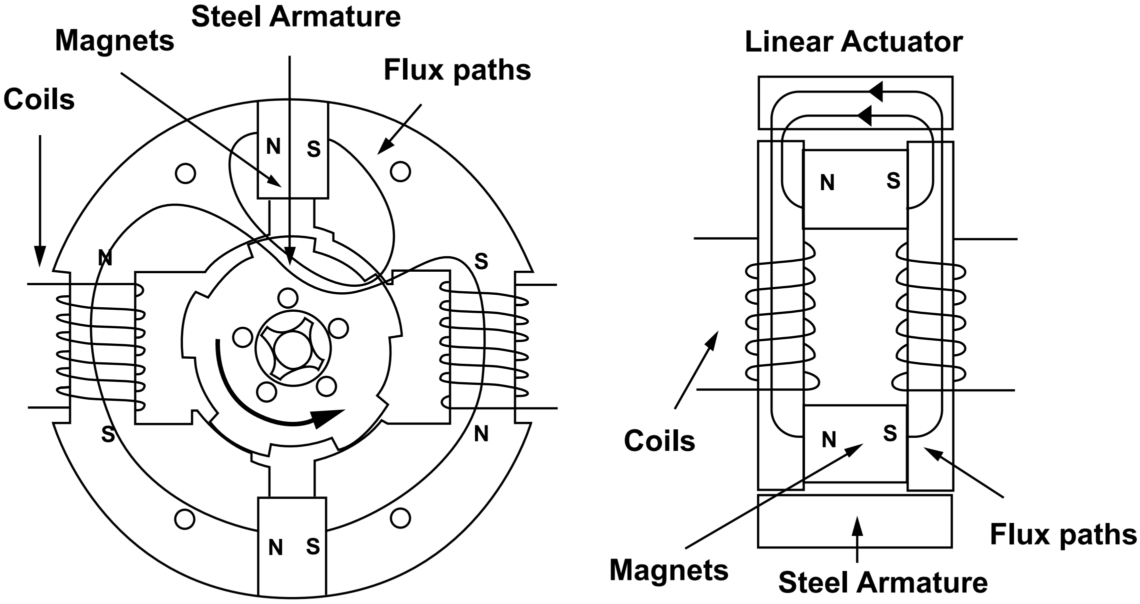

Fig. 9
Optimal Flux Core Motor Construction

Electrical Apparatus

Many readers will no doubt have noticed the similarity of the first illustration presented in this document, to the so called 'Tom Bearden MEG.' This is fair comment, and Joe Flynn has always highlighted this issue. However, it has been commonly stated Joe Flynn has simply developed mechanical apparatus, and the MEG with its electrical functionality, is distinct art, more advanced than the mechanical Flynn apparatus. However, this is shown not to be the case by a careful examination of the Flynn patent, in which the following is stated in the 'Power Conversion' section:

'The construction (shown in Fig.10 A) utilizes four control coils and a single permanent magnet and the construction (shown in 10 B) uses two control coils and two permanent magnets. The flux that would normally be supplied by a primary winding is supplied by the static flux of the permanent magnet or magnets and the control coils convert this static flux into a time varying flux in a novel way. Both arrangements use two secondary coils, the secondary coils are placed in the region of the continuous flux path that would be occupied by an armature or rotor in the linear or rotary arrangements. The regions of the flux paths that perform work are the same in all cases.'

'By alternating the polarity of the control coils during one cycle, one working region experiences an increasing flux and the opposite region experiences a decreasing flux and during the next cycle the opposite occurs. This results in the induction of a voltage in the secondary coils that is decided by the magnitude of the change in flux in the working region and the time in which this change occurs. The novelty of this discovery is that the primary flux inducing the voltage in the secondary coils is supplied by the permanent magnet or magnets and is far greater than the flux supplied by the control coils.'

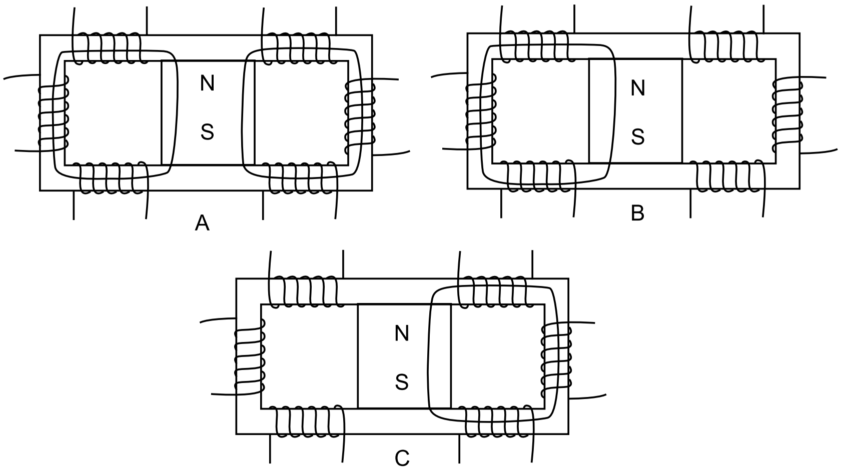

Fig. 10

Fig.10 taken directly from the Flynn patent makes the point even clearer. As can be seen, the device illustrated is in all functional respects absolutely identical to the so called 'Tom Bearden MEG.' This identical prior art, therefore calls into question the intellectual property rights of Magnetic Energy LTD, as the same invention can not be patented more than once. Joe Flynn has also stated that his intellectual property rights will be robustly defended, by legal action if necessary, and he regards himself and his company as being in possession of exclusive rights to the so called 'MEG' unit.

the input and output circuits must be closed in series. The disclosure of this technique amounts to putting the basic MEG methodology fully into the public domain.

As regards replication of electrical output orientated flux core devices, certain important details need to be stated. For example grade 8 ceramic magnets should be used, so as to avoid flux saturation of the core. A basic error, many early experimenters wasted time on. The requirement for strong magnets to obtain over-unity results, is as much of a myth, as the idea new physics is required.

But perhaps the greatest trade secret of the electrical devices, one which several lengthy non disclosure agreements are required to be signed before it can be disclosed, is that the input and output circuits must be closed in series. The disclosure of this technique amounts to putting the basic MEG methodology fully into the public domain.

The reason for this circuitry requirement is obvious enough, with only a little analysis. If the output circuit is closed when the input circuit is activated, then the input energy simply leaks into the output circuit, as in an ordinary transformer. So no flux switching effect is manifested, and the field of the permanent magnet is static in time. Thus you have an ordinary transformer, with reduced efficiency, because of the core flux saturation effect provided by the permanent magnet.

This is one of the most important point to make about the Flynn apparatus. If you approach it as if it is a normal piece of scientific equipment, then proper optimisation is not greatly problematic. For example more turns on the output coils, simply means more voltage and less current, exactly as standard textbook equations predict. Generally, problems only occur, if you imagine the effect is based upon exotic scalar type or vacuum energy physics, when in fact it is ordinary flux manipulation within a core.

Present Status of the Flynn Project

Initially Joe Flynn was remarkably open about his work and research. However, since performing a working demonstration of various advanced hardware samples for the American Department of Defence, little has been heard.

I want to clearly emphasize I do not speak for Flynn research, nor am I in any way connected with Flynn research, and by consequence, have no inside information whatsoever as to the present status of the project. But we all sincerely hope, that the project has not been swallowed whole by the American deep black military industrial research complex.

However, even if this is the case, it does not mean the technology is lost. Extensive and generous details have been provided by Joe Flynn of his research, both in his patent deposition, website, and other comments, such as to enable persons of scientific training and skill, to replicate the effects stated.

While replication of the electrical effect remains extremely demanding, the mechanical apparatus is very easy both to understand and replicate. There is no reason why scientists and home tinkers together, can not build Flynn type flux core motors, and explore over-unity flux manipulation for themselves. The future has arrived, and it is simpler and cheaper than anyone imagined to be possible.

Raum&Zeit

EhlersVerlag GmbH,
Geltinger Str.14e,
82515 Wolfratshausen
Telephon: 08171/41 84-76,
Telefax: 08171/41 84-66
Internet:
http://www.raum-und-zeit.com

Institut fur Raum-Energie-Forschung (IREF) i.m. Leonard Euler

- Global Scaling Theory
- Gravitation theory
- Mathematic Biology
- Gravitation energetics
- Hydrogen-Helium Energetic
- Information technologies
- Telecommunication technologies
- Transport technologies

Flynn Research Company

http://www.flynnresearch.net

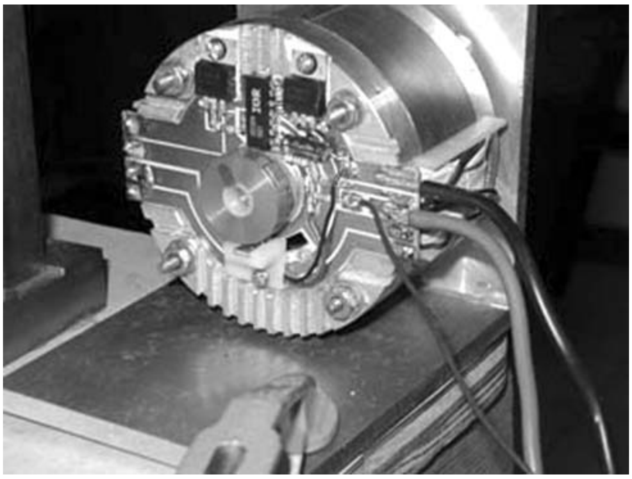

Photo 1

3.3" 1.2" stack height 1/2 hp weight 59 oz

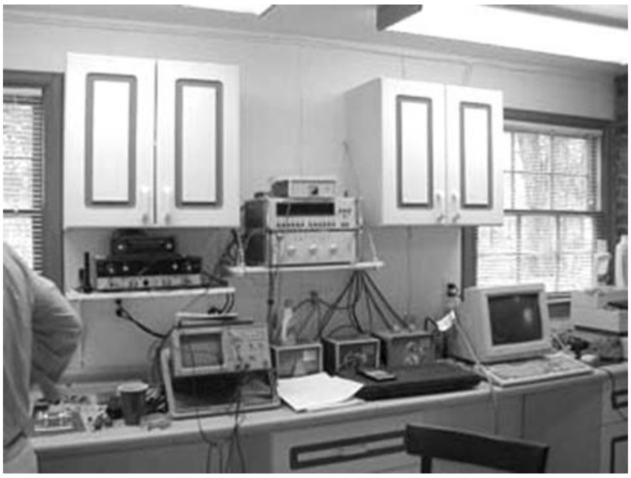

Photo 2

Dynamometers, Power Analyzer etc.

Editor: Below we publish information and photos, represented on Flynn Research Company website (http://www.flynnresearch.net), and a description of the patent on methods for controlling the path of magnetic flux.

Flynn Research is dedicated to research focused on the novel application of permanent magnet and electromagnetic technology. Magnetism is a field that has potential for tremendous growth from both a science and application perspective. The researchers of Flynn Research Company believe that magnetism is one of the least understood and most misinterpreted of the natural forces, with the exception of gravity.

Everything known about magnetism is based on "effect and relationships" without a clear understanding or provable definition of what is a line of magnetic flux.

Even though much work has occurred in the advancement of magnetic materials, little advancement has occurred in the basic understanding of magnetism. The field of magnetism is based on both theory and 'myth' and is generally described as a 'phenomenon'.

We still do not know what a line of flux is and how it travels through space. We know flux "appears" to originate within a permanent magnetic material, forms a loop from one end to the opposite end of the magnetic material, appears to occur without a time constant, produces a force (attractive and repulsive) between other permanent magnets and "magnetically permeable materials and has a relationship to current flow. Everything known about magnetism is based on "effect and relationships" without a clear understanding or provable definition of what is a line of magnetic flux.

The research is based on the "magnetic force of attraction effect" that is related to magnetic flux by the number of lines of flux occupying a given area or flux density squared. The simple fact that if the number of lines of flux occupying a given area "doubles" the force of attraction becomes four

Photo 3

8" dia x 8" height motor presentation

Photo 5

Technology Discussion

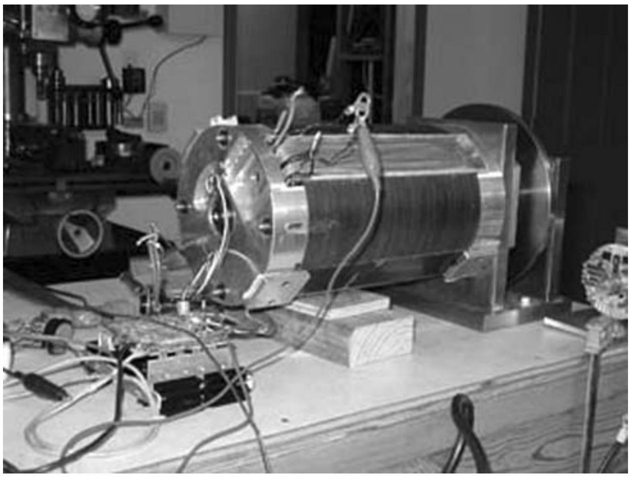

Photo 4

8" X 8" motor testing

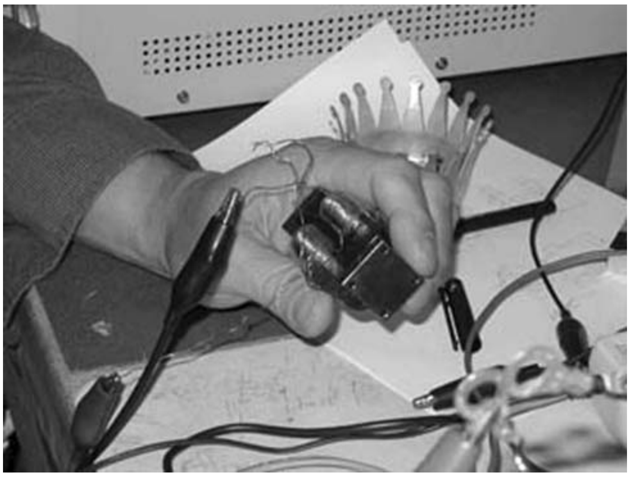

Photo 6

Parallel Path Test Apparatus

times greater. The basic model shows that with specific structures electromagnetism acting with permanent magnet materials can divide and multiply force in a manner that cannot be explained with conventional physics.

Rotary devices, linear devices, reciprocating devices and power conversion devices are now a patented product of this research. The goal of the Company is to commercialize their devices for use in an energy dependent world. Flynn Research is a research & development company and in cooperation with Magnetic Revolutions LLC licenses their technologies for use in the electric motor and power conversion industries.

The researchers do not construct their own test equipment nor write capture and data evaluation programs but use industry / scientific accepted products. Close to $1.5 million has been expended developing, protecting and evaluating the technology. Many fine minds have been involved in the development and evaluation of the technology. Many presentations have been given both to motor companies and academic groups. The prototypes have all been professionally built, which include a 1/2 hp 3.3" dia 1.2" stack height, the assembled motor weighs 59 oz and a motor 8" dia with a stack height of 8" currently undergoing testing.

United States Patent No. 6,246,561
June 12, 2001

Methods for controlling the path of magnetic flux from a permanent magnet and devices incorporating the same

Inventors: **Flynn; Charles J.** (Greenwood, MO)
Assignee: **Magnetic Revolutions Limited, L.L.C** (St. Louis, MO)
Appl. No.: **127056**
Filed: **July 31, 1998**

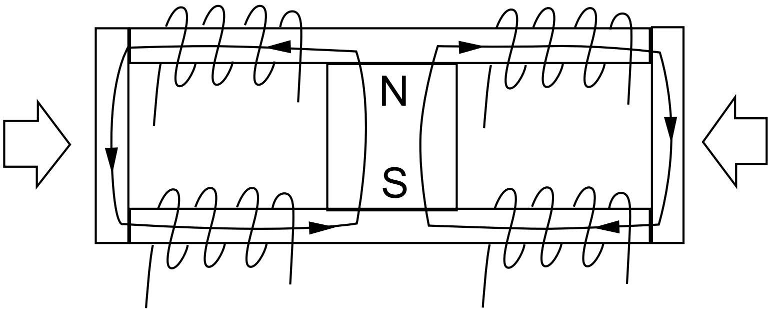

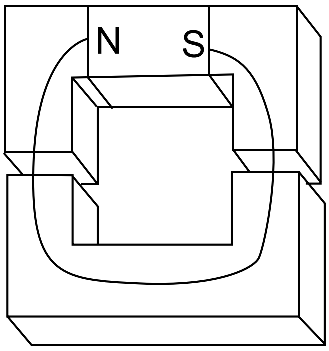

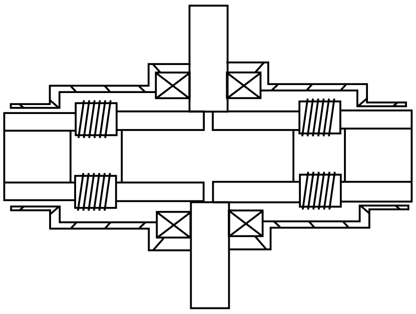

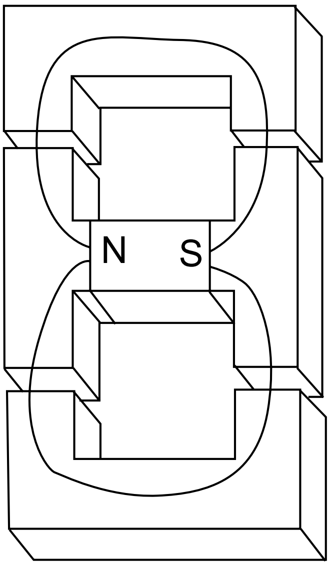

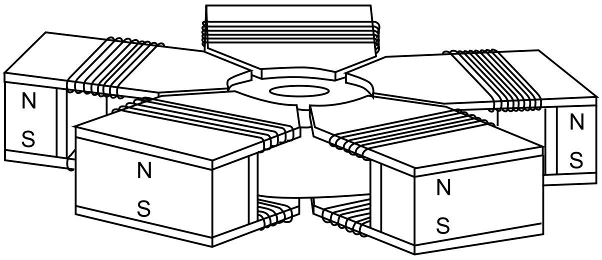

A permanent magnet device includes a permanent magnet having north and south pole faces with a first pole piece positioned adjacent one pole face thereof and a second pole piece positioned adjacent the other pole face thereof so as to create at least two potential magnetic flux paths. A first control coil is positioned along one flux path and a second control coil is positioned along the other flux path, each coil being connected to a control circuit for controlling the energization thereof. The control coils may be energized in a variety of ways to achieved desirable motive and static devices, including linear reciprocating devices, linear motion devices, rotary motion devices and power conversion.

Editorial: The principle of designing of the device (two counter parts of the flux are used) resembles Φ-machine or Gramm's generator. Read in this issue more details on these devices.

Flux-machine And Its Analogues

Review prepared by correspondent Alla Pashova, Russia

Flux-machine (or Φ-machine, since configuration of its field resembles the Russian letter "Φ") was designed by Alexander V. Frolov during the first half of 1990s. Φ-machine description was published more than once on the pages of scientific periodicals (1994, Institute of New Energy, Newsletter, June 1994, p.9.; 2002, New Energy Technologies, 2002, p.3), and it is well known to many researchers concerning themselves with such kinds of engineering.

Let us consider the arrangement of Φ-machine; see Fig. 1 (first published in 1994). A report was presented at the conference «New ideas in natural science», 1996 in Saint Petersburg (see photo in Fig. 1). The primary coil is mounted in the center of generator; two secondary ones are diametrically wound on a ring core. An air gap between ring and central magnetic circuits are of particular importance. **Two magnetic fluxes from two coils are balanced, and, thereby, there is no reaction in the primary circuit.**

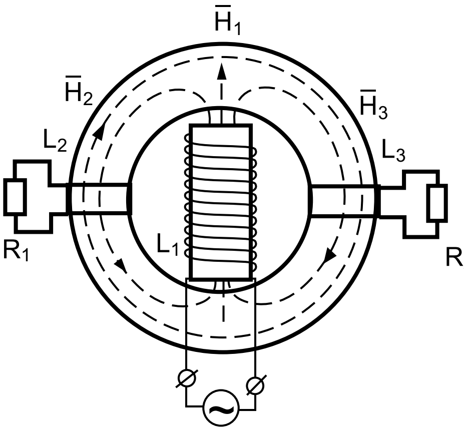

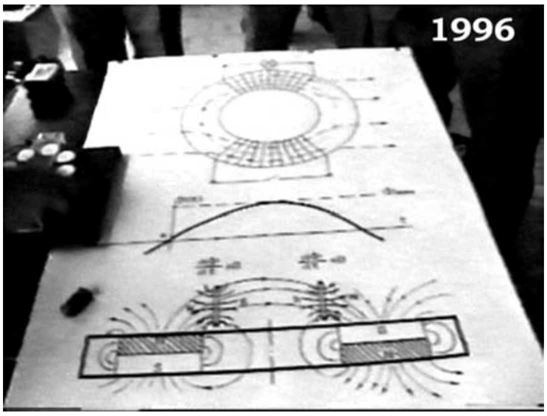

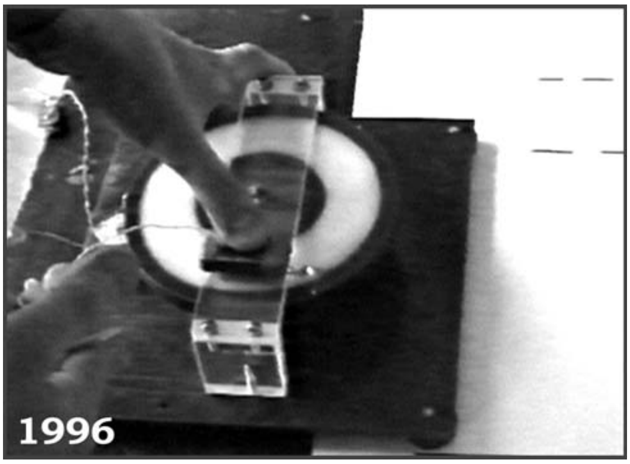

Fig.1
See also Fig. 7, p.29 – the analogy is evident

The device bears some similarity to the Gramm's generator (Zenob Theophyl Gramm, 1826-1901, Belgium-France, took out a patent for an electric ring-rotor generator in 1869). In Gramm's generator (Fig. 2) the ring rotor with a toroidal winding rotates. The toroidal winding touches two diametrically located conducting brushes. It turns out that in the ring of the rotor all winds of one rotor half create a field which is directed towards a field created by the other rotor half.

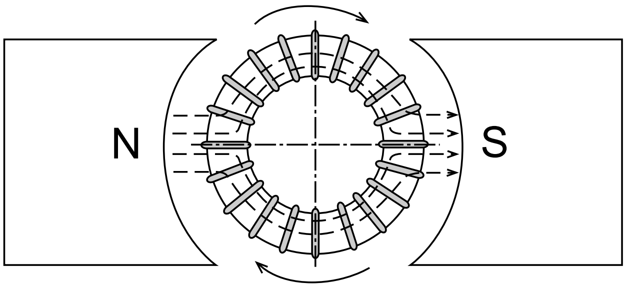

Fig. 2

An example of analogous device is represented in Fig. 3 (information is from the website: http://www.skif.biz/energy/arhiv1-6.shtml), and also in Fig. 4 (we have written about this invention in New Energy Technologies, Issue #5 (8), 2002, article by V.I. Boryak, Email: spin@i.com.ua).

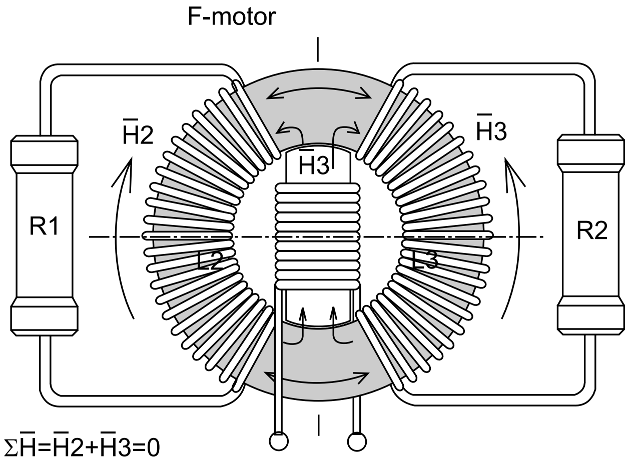
F-motor
$\overline{H}2$
$\overline{H}3$
$\overline{H}3$
R1
R2
L2
L3
$\Sigma\overline{H}=\overline{H}2+\overline{H}3=0$

Fig. 3

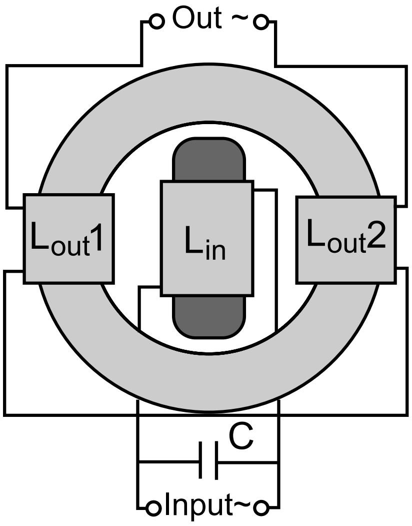
Out ~
$L_{out}1$
L_{in}
$L_{out}2$
C
Input~

Fig. 4

In August 1999, according to Stephen Hartman (Hartman Multimedia Service, Email: hart@harti.com, info@ccard.net), a German research team designed a toroidal generator. As stated by the researchers, it obtained more than 1.200 W by load while efficiency exceeded 100%!

This subject has become very popular among different researchers. Fig. 5 illustrates a design with a magnet rotating in the center (designer - Olaf Berens, Email: olaf.berens@prognost.com).

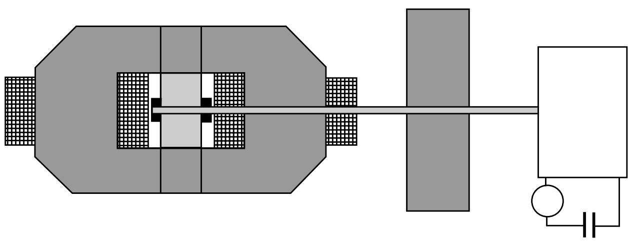

Fig. 5

One more example from the USA: in February, 2003 Donald Hofmann (USA) filed an application for patenting “Generators and transformers with toroidally wound stator winding». Below we publish the description of the patent.

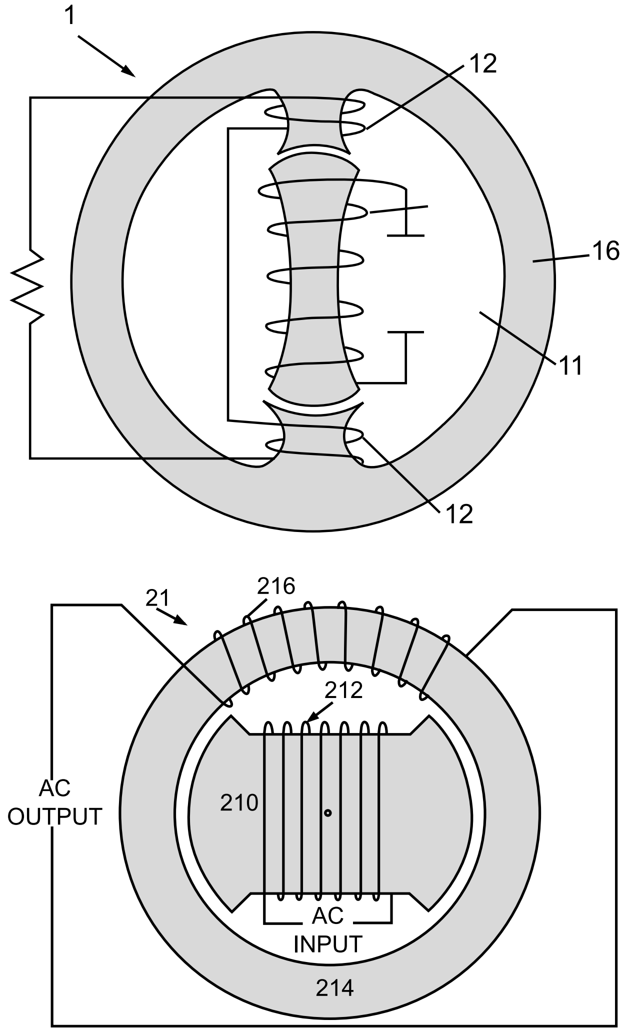
1
12
16
11
12
21
216
212
AC
OUTPUT
210
AC
INPUT
214

Fig. 6

Inventors: **Steven L. Sullivan, USA; David L. Goulet, USA; Donald Hofmann ,USA.**

Electrical generators are provided with toroidally wound stator windings electrically connected in series; a high permeability stator core, preferably an amorphous magnetic alloy material, glassy metal or HYPERCO™ 50 laminations; and a rotor. The toroidally wound stator coils and the stator core trap essentially all of the flux fields generated by the stator coils within the stator core. Since there is essentially no magnetic field leaving the stator, there is essentially no flux field interaction with the field generated by the rotor. The reduction of flux field interaction also reduces counter torque.

According to another embodiment of the invention, open positions of a rotor may be filled with magnetic material, preferably iron. Filling open portions of the rotor may assist in reducing undesirable flux leakage from the rotor. By trapping substantially all of the flux within the rotor, interaction between rotor and stator flux is reduced. In this manner, counter torque is reduced thereby increasing the overall efficiency of the device.

A prototype of this embodiment was constructed using a standard 10 kWtt gas-powered generator. The stator was removed from the generator and its housing. The stator windings were removed from the stator and rewound by hand using a No. 10 size copper wire with 180 winds wrapped in a toroidal fashion around approximately 180 degrees of the stator. The stator was then reinstalled in the modified housing and re-attached to the gas-powered generator. A standard 12 volt car battery was attached to the input of the rotor. A 12.92 volt input with 0.8 Amps was measured as the input to the rotor. The output was measured as 6.5 volts with a 40 Amp output. A series of input and output measurements were taken for various inputs.

Editor: Thus input power is 10 Wt, and output power is 260 Wt.

So, the old well-known ideas have found new ways of embodiment, though their essence remains unchanged. Let us note that Faraday Labs Ltd has been working on creating a prototype of ▯-machine with a toroid of 200 mm in diameter.

Please, read more about this in the next issue!

Editorial: We publish the specification of several interesting patents, which were taken out by Russian inventors for recent years.

Russian patents on alternative energetics

- №93001754, 1996.06.27, Converter of gravitational forces into energy, V.V. Mironov and others
- №92008720, 1995.02.10, Slavic method to produce energy by conversion of gravitational forces, V.V. Mironov
- №93006696, 1995.04.30, Supporting converter of gravitational forces into energy, V.V. Mironov
- №2001121071, 2003.04.10, Method to create energy in liquid, heat-generator and heat-and-power device, A.Yu. Baurov and others
- №2132109, 1999.06.20, Generator of mechanical energy, A.Yu. Baurov and others
- №2001125794, 2003.05.10, Generator of static electricity, S.I. Danilov
- №2001120796, 2003.04.10, Electric power device, D.M. Beliy
- №98116077, 2000.05.10, Heat engine (monotherm, "perpetual mobile of second-type"), V.I. Likhachov and others
- №93048971, 1996.04.20, Float engine, I.P. Tchinarev
- №94023174, 1996.03.27, Gravitational-hydrostatic engine, I.P. Tchinarev
- №98112650, 2000.03.20, Water engine, V.A. Gylchuk
- №2001115953, 2003.03.20, Potential engine by Prokopenko, V.F. Prokopenko
- №2001115954, 2003.03.20, Potential amplifier by Prokopenko, V.F. Prokopenko
- №2001124364, 2003.04.20, Drum of gravitation (gravity), I.A. Strelnikov
- №95107181, 1997.02.20, Method to convert energy of physical vacuum into energy of physical space, O.K. Polev and others
- №92008695, 1995.04.20, Method to get overunity efficiency, Yu.V. Karasev and others
- №93006851, 1995.04.30, Principle, method and device for circulation of mass by force of gravitation and inertia, G.S. Kirichenko
- №93039885, 1996.05.20, Hydro-gravitational electric power station, L.S. Konkov
- №93036371, 1995.11.27, Method to convert energy created in result of interaction of earth's gravity force and air bearing capacity into electric energy, V.I. Kovalenko
- №93025690, 1996.07.27, Method and device to convert heat energy into mechanical one, A.I. Lendyaev
- №2001122885, 2003.04.20, Rotary method to convert magnetic energy into mechanical one, L.N. Mezentsev
- №2001106783, 2003.02.10, Gravitational source of energy, A.D. Yurik and others
- №2001115968, 2003.03.20, Method to convert energy, V.S. Gorelyuh

According to another embodiment of the invention, open positions of a rotor may be filled with magnetic material, preferably iron. Filling open portions of the rotor may assist in reducing undesirable flux leakage from the rotor. By trapping substantially all of the flux within the rotor, interaction between rotor and stator flux is reduced. In this manner, counter torque is reduced thereby increasing the overall efficiency of the device.

A prototype of this embodiment was constructed using a standard 10 kWtt gas-powered generator. The stator was removed from the generator and its housing. The stator windings were removed from the stator and rewound by hand using a No. 10 size copper wire with 180 winds wrapped in a toroidal fashion around approximately 180 degrees of the stator. The stator was then reinstalled in the modified housing and re-attached to the gas-powered generator. A standard 12 volt car battery was attached to the input of the rotor. A 12.92 volt input with 0.8 Amps was measured as the input to the rotor. The output was measured as 6.5 volts with a 40 Amp output. A series of input and output measurements were taken for various inputs.

Editor: Thus input power is 10 Wt, and output power is 260 Wt.

So, the old well-known ideas have found new ways of embodiment, though their essence remains unchanged. Let us note that Faraday Labs Ltd has been working on creating a prototype of Φ-machine with a toroid of 200 mm in diameter.

Please, read more about this in the next issue!

◆

Editorial: We publish the specification of several interesting patents, which were taken out by Russian inventors for recent years.

Russian patents on alternative energetics

- №93001754, 1996.06.27, Converter of gravitational forces into energy, V.V. Mironov and others
- №92008720, 1995.02.10, Slavic method to produce energy by conversion of gravitational forces, V.V. Mironov
- №93006696, 1995.04.30, Supporting converter of gravitational forces into energy, V.V. Mironov
- №2001121071, 2003.04.10, Method to create energy in liquid, heat-generator and heat-and-power device, A.Yu. Baurov and others
- №2132109, 1999.06.20, Generator of mechanical energy, A.Yu. Baurov and others
- №2001125794, 2003.05.10, Generator of static electricity, S.I. Danilov
- №2001120796, 2003.04.10, Electric power device, D.M. Beliy
- №98116077, 2000.05.10, Heat engine (monotherm, "perpetual mobile of second-type"), V.I. Likhachov and others
- №93048971, 1996.04.20, Float engine, I.P. Tchinarev
- №94023174, 1996.03.27, Gravitational-hydrostatic engine, I.P. Tchinarev
- №98112650, 2000.03.20, Water engine, V.A. Gylchuk
- №2001115953, 2003.03.20, Potential engine by Prokopenko, V.F. Prokopenko
- №2001115954, 2003.03.20, Potential amplifier by Prokopenko, V.F. Prokopenko
- №2001124364, 2003.04.20, Drum of gravitation (gravity), I.A. Strelnikov
- №95107181, 1997.02.20, Method to convert energy of physical vacuum into energy of physical space, O.K. Polev and others
- №92008695, 1995.04.20, Method to get overunity efficiency, Yu.V. Karasev and others
- №93006851, 1995.04.30, Principle, method and device for circulation of mass by force of gravitation and inertia, G.S. Kirichenko
- №93039885, 1996.05.20, Hydro-gravitational electric power station, L.S. Konkov
- №93036371, 1995.11.27, Method to convert energy created in result of interaction of earth's gravity force and air bearing capacity into electric energy, V.I. Kovalenko
- №93025690, 1996.07.27, Method and device to convert heat energy into mechanical one, A.I. Lendyaev
- №2001122885, 2003.04.20, Rotary method to convert magnetic energy into mechanical one, L.N. Mezentsev
- №2001106783, 2003.02.10, Gravitational source of energy, A.D. Yurik and others
- №2001115968, 2003.03.20, Method to convert energy, V.S. Gorelyuh

- №2002131190, 2003.04.20, Method to convert heat energy into useful work, G.R. Umarov and others
- №2001122676, 2003.05.27, Device to get energy by conversion of energy from one state to another, T.A. Trushina and others
- №2131636, 1999.06.10, Fuel-less engine, V.E. Alekseenko
- №5037775, 1996.10.10, Magnetic engine, V.E. Alekseenko
- №94003087, 1995.12.27, Gravitational engine, A.A. Voznyukov
- №93058095, 1996.06.27, Perpetual mobile, Yu.S. Parshutin
- №93056706, 1996.07.27, Perpetual mobile, A.V. Remezovich
- №93039002, 1996.01.20, Gas-dynamic perpetual mobile, A.I. Lendyaev
- №94011000, 1995.11.27, Perpetual mobile by Makagyan, V.S. Makagyan
- №97114829, 1999.06.27, Perpetual mobile by Makagyan, V.S. Makagyan
- №95114965, 1997.08.10, Electric perpetual mobile by Abrakitov, V.E. Abrakitov
- №9512209, 1998.02.20, Perpetual mobile, V.V. Sharov
- №99111572, 1999.09.27, Perpetual mobile, N.M. Rotar
- №98116343, 2000.05.10, Engine-electric saw (perpetual mobile), V.V. Smirnov
- №2000101256, 2001.11.20, Perpetual mobile, V.D. Rykov
- №2000120592, 2002.07.20, Perpetual mobile, K.I. Harazov
- №2001117783, 2003.03.20, Magnetic perpetual mobile, S.N. Verevkin
- №2001123502, 2003.04.20, Magnetic engine, A.E. Ryumin
- №5032711, 1995.02.20, Ferromagnetic engine, V.P. Levkin
- №94019782, 1996.01.10, Device for demonstration of conversion of permanent magnet energy, M.F. Ostrikov and others
- №95103846, 1996.11.27, Magnetic generator, M.F. Ostrikov and others
- №95112010, 1997.06.27, Method to convert permanent magnet energy into rotary motion, A.V. Starostin and others
- №2143170, 1999.12.20, Method to convert energy of mechanical loading into energy of repeated transference, V.I. Rahovsky and others
- №2000119415, 2002.06.20, Method to convert energy of fields of permanent magnets into mechanical work, Yu.S. Pilipkov
- №2001109811, 2003.02.27, Method to capture and convert energy of physical vacuum into electric energy, S.N. Kovalev
- №99121374, 2001.11.10, Perpetual mobile, E.M. Volodin
- №94026259, 1996.05.20, Device to convert magnetic field energy into kinetic energy, P. Imrish

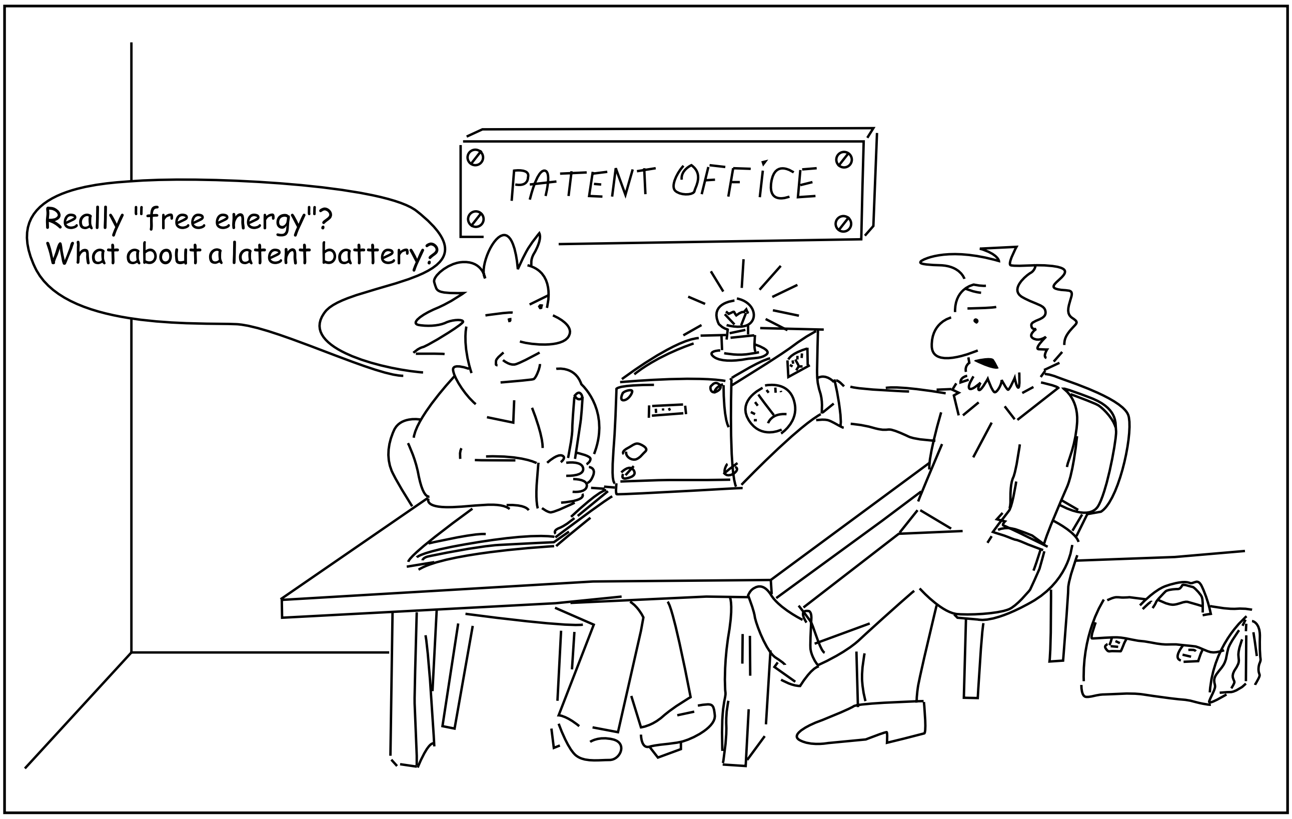
PATENT OFFICE
Really "free energy"?
What about a latent battery?

Device for conversion of magnetic field energy into kinetic energy

No. 94026259, 1996.05.20

Author: Pavel Imrish

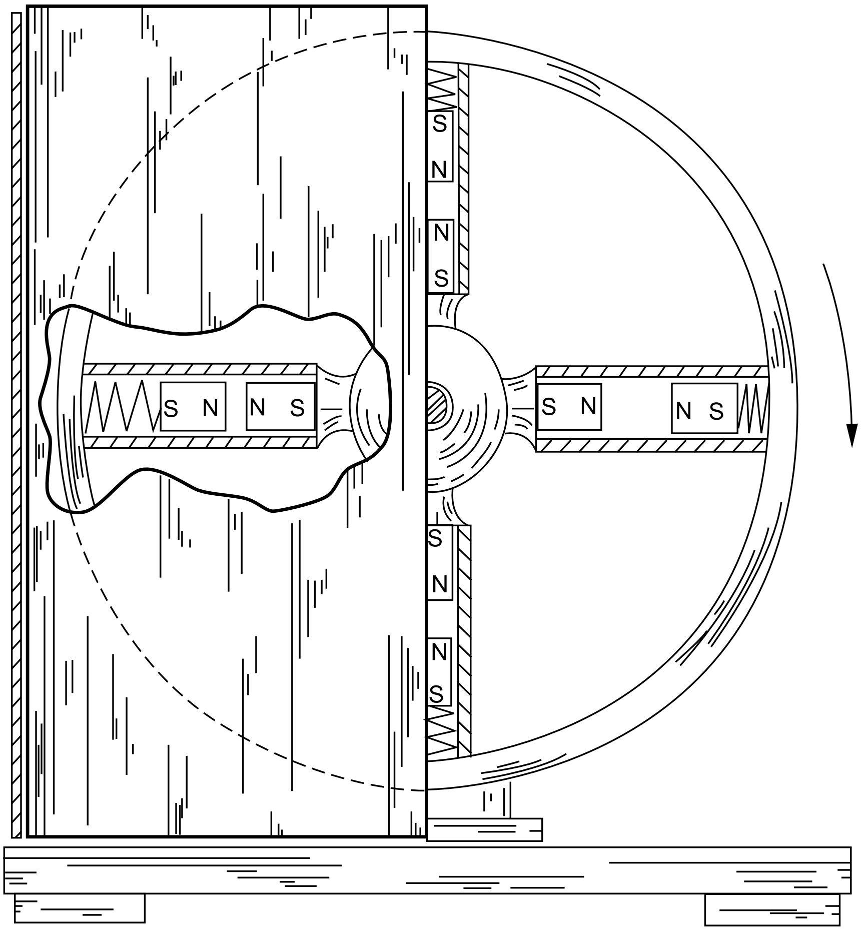

Fig. 1

The invention refers to a device for conversion of magnetic field energy into kinetic energy. The device has a rotating wheel with at least two radially opposing non-magnet holders used as guide ways for at least two permanent magnets, respectively. The permanent magnets are mounted in the holders so that there is a possibility to remove them. From the both sides of the wheel plates made of magnetic material are mounted axially; the plates surround the wheel partially. Depending on direction of the both poles, namely on location of the similar and opposite poles towards each other, compression element is situated on the holders, or between the wheel and the outer end of the permanent magnet located at the furthest radial distance from the axle of the wheel, or between two permanent magnets. (The device operation principle is obvious from Fig. 1).

Editor: We have found an example of a similar device whose description is presented in the book by R. Ford The Perpetual Motion mystery [1].

In this case, there is an analogous process, during which half of the cycle is screened and the rotor is supposed to rotate permanently (see Fig. 2).

1. RA. Ford, The Perpetual Motion mystery. Lost Technology Series, p.23.

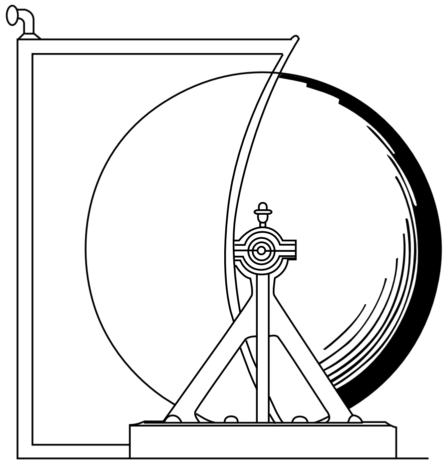

Fig. 2

Gravity spaceships

Review on Grigory R. Uspenskiy's works, Russia

http://space21.boom.ru/gravity.htm

The level of development of modern astronautics is primarily determined by the energy capabilities of space carriers. So, nowadays, the single-mission carrier rockets equipped with chemical engines allow space exploration within the solar system. These carriers have delimited the near-earth space in the form of a geostationary orbit, where the practical-purpose space complexes operate.

Newer and higher levels of quality will be reached by cosmonautics upon its mastering gravity power engineering. Flights towards the nearest stars will become possible and by the end of the next century this will be followed by travels throughout Galaxy.

Gravity engines are structurally simple (**two bodies of different density are rigidly connected**). To form some practicably significant value of thrust it is required to use absolutely new technology regarding creation and retaining of matter of high density which can be compared to that of atomic nuclei. For instance, at an engine mass of about 2 t, it is possible to attain 10 N thrust using bunches of aluminum and lead nuclei. The nuclei are approached to each other at a distance comparable with sizes of these bunches.

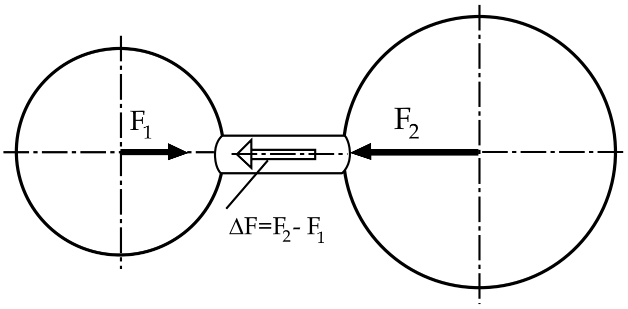

Fig. 1

Use of matter which is denser than nuclei bunch, increase of mass of the gravitating bodies, and reduction of the distance between the bodies causes increase of thrust. So, for a level of density of gravity holes (i.e. for extremely dense state of matter in which it is not attracted by other bodies) limit value of acceleration of the gravity engine reaches the order of 10^{100} ms^{-2}.

In engineering, we usually use units of distance as meters, centimeters, millimeters, and microns. Let us consider one of them – millimeters. With such a size of the gravitating bodies and distance between them, it is possible to obtain substantial thrust forces of the gravity engine and, accordingly, accelerations of its motion.

So, using an engine of 20 ton mass we will get the thrust of 2x10^4 N, with mass of 200 ton we will get the thrust of 2x10^6 N, and so on. Thus, by increasing the engine mass by one order of magnitude, we obtain increase of thrust by two orders, while sizes of the bodies remain unchanged. But acceleration, in this occasion, remains constant and equal to approximately 1 m/s^2.

Decrease in size of gravitating bodies of the engine leads to the corresponding increase of thrust by an order. Decrease of size is realized by increase of density of these bodies by 3 times up to 0.3 mm at the same masses. The acceleration they produce is increased by an order as well, i.e. up to 10 m/s^2. Decrease of sizes of the gravitating bodies and the distance between them down to 0.1 mm results in further increase of engine thrust by one more order and also increase of acceleration up to 100 m/s^2. If size of the bodies is 1 micron then acceleration increases up to 10^6 m/s^2.

It is apparent that even with up-to-date technological possibilities of miniaturization the gravity engine with masses of dozens and hundreds of tons is capable to create great thrust forces and form accelerations considerable enough to fly across the Galaxy. Therefore, the problem of creating a gravity engine is, mainly, the problem of producing and retaining the high-density matter. It will be possible after studying the mechanism of interaction between matter & gravitational substance followed by developing techniques for synthesizing of high-density, large-mass, and small-size substance from this matter.

It is possible to control the value of thrust of the gravity engine changing the distance between gravitating bodies. Thrust direction can be changed by turning the rigid connection of the gravitating bodies. The engine can be activated by bringing the gravitating masses together, and switching-off by moving them apart.

In a 24-hour period, such a ship can reach a speed of 10^6 m/s and cover a distance of about 10^{10} m. In a month, with a top speed of $3x\ 10^7$ m/s the ship will cover the distance of 10^{13} m; in a year, with a speed of $4x\ 10^8$ m/s the ship will cover a distance of 10^{15} m; in 10 years - $4x\ 10^9$ m/s and 10^{17} m, respectively. Apparently, the ship with sufficient over-load can be used for flights within the solar system and for manned flights towards the nearest stars.

When transporting unbreakable and non-urgent cargoes, the acceleration can be increased up to a few dozens of unities. The flight duration might be also increased by several tens of years. This extends the area of utilizing the gravity ships with over-loads of several unities. Besides these ships can be used for transporting operations within near vicinities of the Galaxy.

The high-speed flight will demand protection against approach flow of matter. So, per second (with a ship speed of 10^{10} m/s) 10^{-12}kg of matter will approach the ship. At that density of interstellar material is 10^{-24} kg/m^3 and midship area is 10^2 m^2.

The great speed of approaching matter will create a substantial resisting force. At a speed of 10^{10} m/s, this force will come to the order of 10^{-2} N, whereas with a speed of 10^{18} m/s it will come to 10^{14} N. This is a great value, but as compared to thrust of such a ship which is to equal to 10^{17} N, the former value is a small one. That is why it is possible to overcome such resisting force.

Intercommunication with these ships might be, most likely, realizable by means of distortion of gravitational field. It is possible that the gravitational field distortion will be formed, on the contrary, by generating the matter from the gravitational field, and, most probably, by a method as yet unknown.

New books and DVD

Monograph **"General Etherodynamics. Modeling of matter structures and fields on the basis of conception of gas-like aether" by Vladimir A. Atsukovsky.** 2nd edition, M., Energoatomizdat, 2003; Brochure **"12 experiments on etherodynamics"** Zhukovsky, Publisher "Petit", 2003.

On questions about purchasing, please, refer to Email: atsuk@dart.ru

◆

Floyd Sweet Bench Test DVD

In order to demonstrate the reality of tapping free energy from the vacuum, we have now made available a very reasonably priced DVD of the bench tests of the late Floyd Sweet's Vacuum Triode Amplifier, with commentary by Tom Bearden.

During these two tests, the unit, which weighed about 6 lbs., can be seen to be putting out well over a million times more power than was put into it.

This DVD makes an excellent educational or instructional tool, and is priced at USD 9 including US domestic shipment. Running time is about 20 minutes.

Note: the full length Sweet DVD, which includes Sweet's proprietary conditioning process for the magnets, is also still available.

Order from http://www.cheniere.org/sales/order_by_credit_card.htm

On Velocity of Drive-Free Motion

S.A. Gerasimov , V.V. Stashenko, Russia

Physics Department, Rostov-on-Don State University,
Zorge St. 5, 344090, Rostov-on-Don, Russia

***Editorial:** We publish a summary of the article. The original text you can receive from the authors.*

A drive-free system made on the basis of a planar mechanism is described. The experimental results on average velocities of the drive-free propulsive system are presented in a scaled view.

During usual motion, a system is repealed from the surface in the case of its motion on a plane or from the medium when it moves in a resistance medium. This kind of motion is achieved by using a drive mechanism. In contrast to the usual motion, the drive-free propulsive system moves due to interaction of a body of a system with another body of the same system. Sometimes such machines are called inertioids [1] or vibration propulsive devices [2]. To produce such kind of motion it is sufficient to provide anisotropy of the resistance force [3-5] or asymmetry of the internal force [6]. One of the simplest ways to do it is to use a simple two-link planar mechanism. A device for such a motion is shown in Fig. 1.

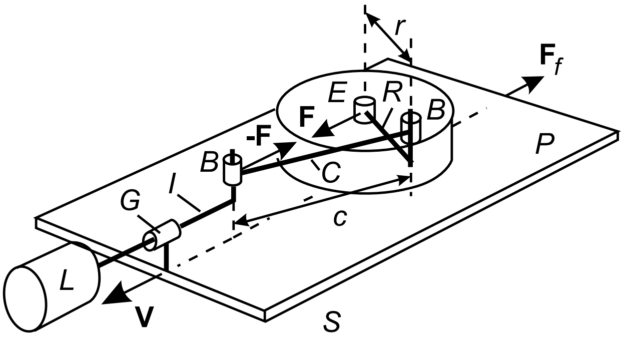

Fig. 1

Experimental device

This device with a total mass of M+m consists of a platform P on which an electric motor E rotating a drive rod R of length r is mounted. This is only one part of the planar mechanism. Another link is the connecting rod C of length c which connects the unbalanced load L and the drive rod R by means of two bell bearings B. The unbalanced load L of mass m on the rod I slides in guide G. The basic distinction of this mechanism is in that this variant of drive-free machine allows us to analyze the experimental data on average velocities of motion. Actually the only forces, which act on the platform and are collinear to the moving direction, are the force **F** caused by vibrations of the load L and the frictional force $\mathbf{F}_f$. The platform starts to move when the force acting on the platform becomes greater than the frictional force.

Let us discuss possibilities of the real propulsive system shown in Fig. 1. Measurements were made at r=0.01 m and at three mass ratios m/M. The values of frictional coefficients were measured for each parameter of the system and vary from k=0.25 to k=0.4. The system starts moving when theoretically the drift does not take place. When motion becomes reversible, the velocity of drive-free drift does not decrease.

Thus drive-free motion is possible even if the frictional coefficients are very small. When increasing the frequency of vibrations velocity does not decrease. We can not discuss the existence of some third force acting on the platform (we can suggest the existence of frictional force $\mathbf{F}_f$ and the internal force **F** caused by vibrations). Let somebody else take the liberty of making such a conclusion!

References

1. Tolchin V.N. Inertioid. // Perm: Perm Book Publisher. 1977. Blekhman I.I. Vibrational Motion. // Moscow: Nauka. 2000. Nappo F. Sulla Influenza Indiretta Delle Forze Inertne sul Moto del Baricentro. // La Ricerca Scientifica (Rendiconti). 1965. V. 8. No 1. P. 3-14.
2. Gerasimov S.A. Anisotropy of Frictional Force and Vibrational Motion. // Problems of Applied Physics. 2001. V 7. P. 85-88.
3. Gerasimov S.A. Self-Similarity of Vibrational Motion in a Resistant Medium. // Journal of Applied Mechanics and Technical Physics. 2002. V. 43. No 1. P. 90-92.
4. Gerasimov S.A. A Real Model of Vibratory-Impact Propulsive Device. // Problems of Applied Physics. 2000. V 6. P. 117-118.
5. Langhaar H. L. Dimensional Analysis and Theory of Models. // New York: Willey. 1951.

ELECTROMAGNETIC SELF-ACTION

S.A. Gerasimov, A.V. Volos, Russia

Physics Department, Rostov-on-Don State University,
Zorge St. 5, 344090, Rostov-on-Don, Russia

The torque of a self-action exerting on the mobile part of the reactionless engine is measured. It is shown that there exists a value of the height of the mobile electrode at which the torque of self-action is maximal.

There is an opinion that a body can not act on itself [1]. The existence of the so-called force of self-action is believed to contradict to the law of conservation of momentum. This is nothing but prejudice. Though the experimental confirmations of existence of self-action [2-4], conventional science is ready to refuse modern electrodynamics and make up a new theory of electromagnetism [1, 5] in order to save Newton's third law in use. It is comprehensible. First of all, these are the theoretical [5] and experimental [3] errors. The equivalence [5] of the Biot-Savart force is mathematically coarse since in this case the highest terms of expansion of force of self-action are cast out. This equivalence violates law of action and reaction and Ampere force law for which the principle of equality and collinearity of action and reaction forces is valid. Non-linearity of dependence of the torque of self-action on the current intensity [3] makes us doubt in the validity of these results. Moreover, typical values of force [2, 4] and torque [3, 6] are too small even if direct current in the circuit is significant. It is required to pass through the direct current of hundred amperes to produce the considerable displacement or turn of a body. As a result, small values of the effect caused ambiguity in the explanation of the phenomenon [1].

In fact, searching the reactionless propulsive devices we usually forget about the magnetic self-action which arises when a body consisting of a magnet and incomplete electric circuit can move violating law of action and reaction [7]. This violation is proved theoretically [8] but peculiarities of this motion are not well discussed. The principle of such a motion is shown in Fig. 1. This is the magnetic interaction between two current elements one of which is a part of a closed circular loop *L* and another is perpendicular to the first one. Since the force $d\mathbf{F}_{jm}$ acting on any current element of the loop *L* is perpendicular to the density of current $\mathbf{j}_m$ then the **Z**-component of the torque $d\mathbf{N}_m=[\mathbf{r}_m \mathrm{x}\, d\mathbf{F}_{jm}]$ is equal to zero while torque which exerted on the second current element by the magnetic field of the loop, does not.

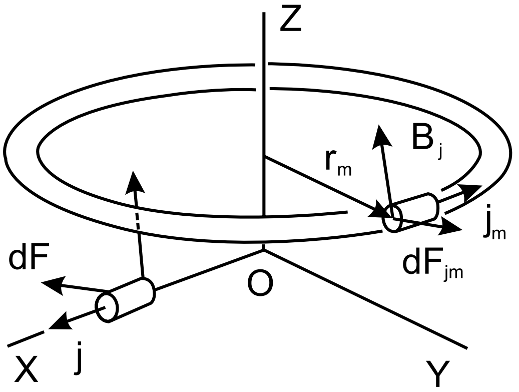

Fig. 1

Violation of law of action and reaction.
The force $d\mathbf{F}_{jm}$ with which the magnetic field $\mathbf{B}_j$ of another current element acts on the current of density $\mathbf{j}_m$, is not equal nor opposite to the force $d\mathbf{F}$ with which the magnetic field $\mathbf{B}$ of the first current element acts on the current of density $\mathbf{j}$.

An experimental device is shown in Fig. 2. A commercially available ring-shaped carbon-steel magnet *M* (inner radius – 20 mm; outer radius – 55 mm, height – 25 mm) of $2.2\mathrm{x}10^5$ A/m magnetization is located on the cover of a cylindrical electrode *E* suspended by thread *T*. The electrode *E* is furnished with a central electrode *C* of the same height *h* and 5 mm diameter. The radius of the electrode *E* is 130 mm. The open end of electrode *E* and the central electrode *C* are submersed in a conducting liquid L_+ and L_- placed in two volumes of a vessel *V* so that the depth of submergence of these parts is about 5 mm. In this experiment the conducting fluid is 10% solution of copper sulfate ($CuSO_4 5H_2O$). The vessel *V* consists of two thin coaxial plastic cylinders of 2 mm thickness. The radii of the plastic cylinders having the common bottom are equal to 10.5 cm and 18 cm. The height of the vessel is 12.5 cm. A disk and a ring-shaped electrodes S_+ and S_- of 1.5 mm thickness are placed on the bottom of the vessel to supply the direct electric current of intensity *I*. All conducting parts of the device are made of copper. The thickness of the electrode *E* is 1.5 mm.

When the direct current *I* of density **j** flows through the electric circuit, the magnet and the electrode *E* rotate in the direction of the magnetic force *d***F** which is proportional to the vector product [**j**x**B**] as shown in Fig. 2. This is a force of self-action by means of which the mobile part of this device consisting of the magnet *M* and the electrode *E* acts on itself. However there is no force of reaction which could cause such a rotation. This is really true since a ring-shaped magnet is equivalent to two cylindrical surfaces with the surface current of density $\mathbf{j}_m$. The force of self-action *d***F** acting on the cover of the electrode is compensated by the force of self-action $d\mathbf{F}_*$ acting on the cylindrical part of the electrode *E*. Magnetic induction field $\mathbf{B}_*$ "does its part" as shown in Fig. 2. Therefore, it would be appropriate to find out how the height of the electrode *E* influences the value of the torque **N**. Such an influence is demonstrated in Fig. 3.

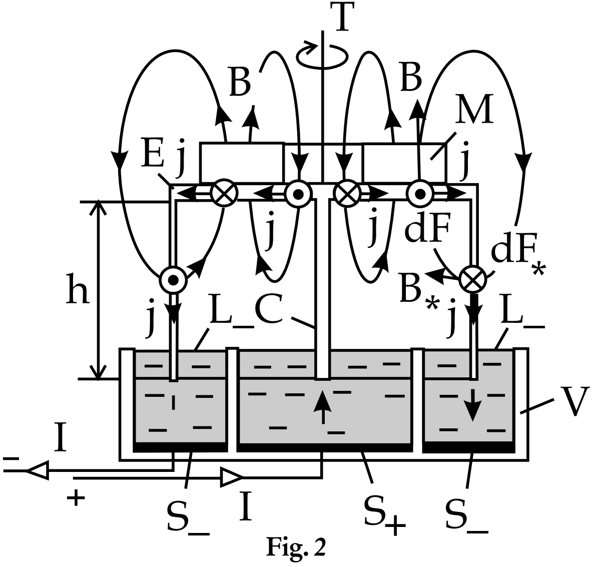

Fig. 2
Experimental device
(•) and (x) – directions of magnetic forces acting on various parts of the mobile electrode *E*

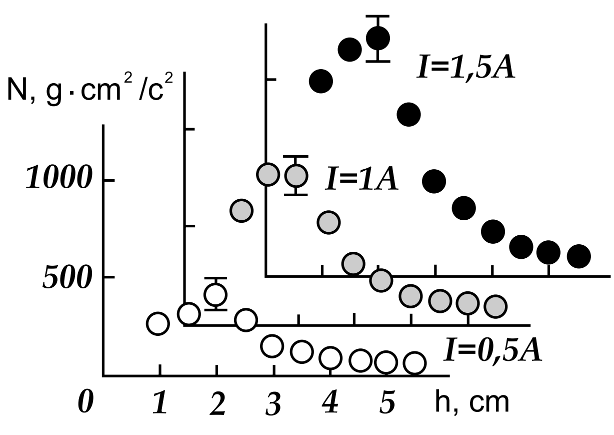

Fig. 3
Dependence of the torque *N* on the height of the electrode *h* at various values of the direct current *I* in the circuit

Thus there exists the value of the height *h* at which torque of self-action is maximal. This is the first result of the present work. At *I*= 1 A and *h*= 2 cm the electrode *E* turns at angle of 2.2 radian (126 °) with respect to the equilibrium position. Such a turn corresponds to 3.4x10^2 gxcm^2/s^{2}rad, i.e to the constant of torsion of the thread. The second result is that the maximum value of the torque of self-action *N* at the current *I*= 1 A is even larger than the value of the torque produced by the traditional unipolar device [3] at the current 50 A.

In fact, this work represents an attempt to draw attention to a more effective and simple practical application of the self-action. It remains only to transfer electric current in an unclosed electrical conductor and such a possibility really exists.

References

1. Graneau N. The Finite Size of the Metallic Current Element. // Physics Letters A. 1990. V. 147. N. 2-3. P. 92-95.
2. Sigalov R.G., Shapovalova T.I., Karimov H.H., Samsonov N.I. Magnetic Fields and Their New Applications. // Moscow: Nauka. 1976.
3. Das Gupta A.K. Unipolar Machines. Association of the Magnetic Field with the Field-Producing Magnet. // American Journal of Physics. 1963. V. 31. N 6. P. 428-430.
4. Cavalleri G., Bettoni G., Tonni E., Spavieri G. Experimental Proof of Standard Electrodynamics by Measuring the Self-Force on a Part of a Current Loop. // Physical Review E. 1998. V. 58. N 2. P. 2505-2517.
5. Christodoulides C. Equivalence of the Ampere and Biot-Savart Force Law in Magnetostatics. // Journal of Physics A. 1987. V. 20. N 8. P. 2037-2042.
6. Serra-Valls A., Gago-Bousquet G. Conducting Spiral as an Acyclic or Unipolar Machine. // American Journal of Physics. 1970. V. 38. N 11. P. 1273-1276.
7. Gerasimov S.A., Volos A.V. On Motion of Magnet in Conducting Fluid. // Problems of Applied Physics. 2001. V. 7. P. 26-27.
8. Gerasimov S.A. Self-Interaction and Vector Potential in Magnetostatics. // Physica Scripta. 1997. V. 56. N 3-4. P. 462-464.

JNET

New Energy Technologies
collection of articles
Japanese Version

Collection of articles from New Energy Technologies magazine, 2001-2002 is published in Japan.

On purchasing of the book, please, contact

Eiichi Yamamoto,
President Yama Trans Co. Ltd.

Email: admin@yamatrans.co.jp or
eyama@yamatrans.co.jp

Uranium Photoaccumulator

Anton N. Yegorov, Russia

St.Petersburg Institute of Nuclear Physics, Russian Academy of Sciences, Gatchina, 188350, Leningrad region, Russia
http://www.pnpi.spb.ru, Email: bti@isppd.pnpi.nw.ru

Photoaccumulators are galvanic elements consisting of two communicating vessels, one of which is exposed to light and another is kept in darkness. Ionic composition of electrolyte, which fills these vessels, depends on intensity of illumination. If an inexpensive electrode is placed in each vessel, then a potential difference will appear. If the electrodes are closed at external working resistance the electron stream will begin to equalize ionic composition and the difference in illumination intensity will tend to disbalance it. As a result, a certain working mode will be established which will allow to utilize the photoaccumulator for production of industrial electrical energy. As distinct from the solar battery, which works while exposed to the light, the photoaccumulator stores some part of solar energy in its electrolyte which is supplied with electrical energy after sunset too.

The idea of photoaccumulator is not a new one. It was established in the end of 19th century that it is feasible to create a photoaccumulator containing electrolyte made of the following mixture: $Fe^{2+} + Hg^{2+} = Fe^{3+} + Hg^{+}$.

This photoaccumulator was not adopted in practice because of high toxicity of the electrolyte and very low EMF (approximately 0.018). Development of technology is accompanied by a widening range of materials and matters which can be used in large scale production and power industry. Among such materials are, in particular, titanium, depleted uranium and rare-earth elements. Nuclear power industry uses U235 isotope only. Naturally occurring compound contains only 0.72% of U235. Hundreds of thousands of U238 from which U235 has been extracted are kept as useless stock.

In the end of sixties in one of radiochemical laboratories of LINP (Leningrad Institute of Nuclear Physics) P.N. Moskalyov repeated the same experiment over a long period of time. In the morning he placed a tightly closed retort filled with uranyl-chloride soluted in compound of water, spirit and hydrochloric acid on the window-sill. The solution in the retort changed its color from yellow to emerald-green. Before leaving the laboratory Moskalyov removed the retort with green solution from the windowsill to a closed chest. In the morning Moskalyov retrieved the retort with the solution (which was already yellow) from the chest and placed it on the sill. The solution became green again and the whole process repeated daily.

Chemical processes in uranium electrolyte are rather complicated: first, under the light influence the uranyl ion oxidizes ethyl alcohol, which is then transformed to aldehyde: $UO_2^{+2} + C_2H_5OH^+ = U^{iv}(OH)_2 + C_2H_4O$.

In the dark ions of urany-4 interact with aldehyde and are transformed to UO_2 uranyl ion and spirit again. Thus, **this cyclic process can take place over a significant period of time being supported by the energy provided by light quanta of visible light which are transformed into the heat.**

The idea of utilizing this process in order to obtain electrical energy emerged in the late nineties. The mechanism of uranium photoaccumulator is very simple (Fig. 1): two glass vessels, one made of transparent glass and another of non-transparent glass, are connected with a black acid-proof rubber tube. The communicating vessels are filled with aqueous electrolyte containing 40% of spirit, 80-100 g/l UO_2Cl_2 and 0.2 m HCL. The electrolyte is previously activated by multiple cycles of exposing it to light and then placing it in darkness in order to obtain a significant volume of aldehydic fraction. The major difficulty of designing uranium photoaccumulator is connected with the choice of the material for electrodes. In galvanic elements carbonized platinum electrodes are usually used. For practical use platinized titanium gives the best fit. Titanium foil covered with superfine layer of platinum is currently widely used in electrochemical industry and is procurable.

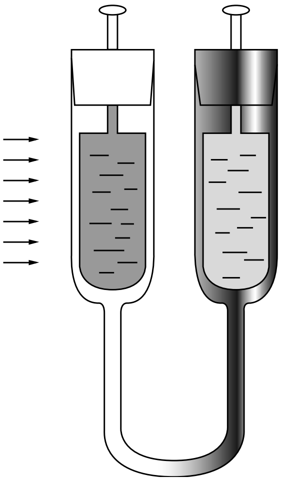

In the transparent vessel the titanium petal-shaped electrodes are parallel the light flow, so the light quanta move freely through the solution.

In working condition each photoaccumulator produces 10 mA of current at 0.3 volt voltage, so they are to be connected in large batteries, both in series and in parallel.

It is also necessary to consider the aspect of safety. Until now, the depleted uranium has been used for one purpose only, namely, for manufacturing of cores for armor-piercing shells. When used in this manner, the uranium is pulverized, which leads to environmental damage. The battery of photoaccumulators is placed in hermetic metal chest; the front panel of the latter must be made of solid safety glass. Such a system will protect the photoaccumulator battery from malicious intent and the security staff from weak uranium radioactivity.

International Prize on Power Engineering "GLOBAL ENERGY"

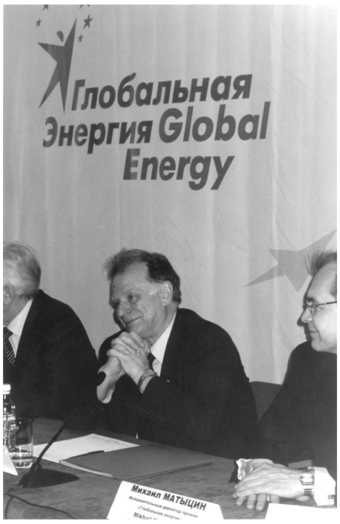

Zhores I. Alferov

International Prize "Global Energy" was first presented in St. Petersburg on June 15, 2003. The bonus fund of $900,000 was shared between three scientists "whose work has been and is still important both to the last century and future research", said Zhores Alferov, Chairman of the Global Energy Prize Award International Committee.

Nick Holonyak, Professor at the University of Illinois (USA), was awarded for "fundamental contributions to the development of power silicon electronics and invention of the first semi-conducting light-emitting diodes in a visible part of the spectrum". Thyristor invented by Holonyak allows the transformation of direct current into alternating and vice versa, to direct this process, to store and accumulate energy. This device is used for every electric locomotive, for frequency transformation. Such devices help to save significant amounts of energy, thus, about 30 percent of all energy produced in the world is processed through thyristors.

Nick Holonyak's other invention, i.e. semi-conducting light-emitting diodes in a visible part of the spectrum, has led to development of a new field of the modern power engineering. Supposedly, efficient and safe light-emitting diodes will replace daylight lamps in the next ten years.

"Although Nick Holonyak's invention dates back to the end of the 50ies of the last century, it remains current" said the Chairman of Expert Commission, Academician Vladimir Fortov. "This is one of the key inventions leading to the search of new ways of sustainable energy saving".

Gennady Mesyats, Academician of Russian Academy of Sciences, and **Ian Douglas Smith**, Senior Scientist at The Titan Pulse Science Division, were awarded for giving birth to a new direction in the field of power engineering, i.e. power pulse energy. Due to their research, commutation of high-level currents at megavatt levels of voltage has become possible; a number of powerful current choppers have been created, futhermore, there have been developed transformer constructions which allow avoiding energy loss in power transmission lines. These research and innovations resulted in the creation of a number of unique pulse power machines and devices widely used in Russia and abroad ("Aurora", "Helia" (USA), "Sinus", "Gamma", "Maus", "Pik" (Russia) and others).

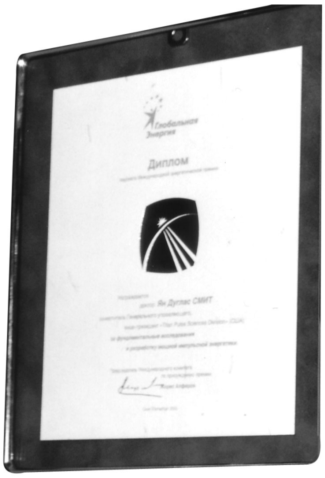

Diploma of Global Energy Prize laureate

Глобальная
Энергия Global
Energy
Геннадий МЕСЯЦ
Gennady MESYATS
Жорес АЛФЕРОВ
Zhores ALFEROV
Михаил МАТЫЦИН
Mikhail MATYTSIN

Teleportation

Review prepared by correspondent Alla Pashova, Russia

For conventional science, the term "teleportation" is not worthy of serious consideration unless ascertained by compulsory "quantum" teleportation. Thus, teleportation is unconditionally referred to microworld phenomena and, in fact, comes to distant information transfer. Spatial transference of a macroobject of definite mass has been excluded.

Teleportation of states

What does **quantum teleportation** mean? During active development of quantum theory, in 1935, the so-called EPR-paradox (Einstein-Podolskiy-Rozen paradox) was formulated in the well-known work "Can quantum-mechanical description of reality be full?" written by Albert Einstein, Boris Podolskiy, and Natan Rozen.

The gist of the paradox is as follows. There are two particles interacting for some time, thereby forming a common system. From the position of quantum mechanics, this coupled system can be described with a certain wave function. When interaction is over and particles scatter within arbitrarily large distances, they still will be described by the same function as before. At that, state of each separate particle cannot be known in principle that is apparent from uncertainty relation. Only when one of the particles enters a receiver, which records its parameters, the relevant characteristics of the other one emerge (exactly emerge, but not become known!). Thus instant unlimitedly distant "transmission" of quantum state of the particle is possible. Therewith, teleportation of the particle itself and transference of mass do not take place.

Einstein and his colleagues believed that existence of such particles predicted by quantum mechanics prove theory incompleteness. Thereof, the scientists inferred the necessity of other parameters (besides wave function) to describe quantum states. Otherwise, from the local viewpoint, correlations between elements of such a system could not be understood. It was far much later, when Bell showed that some of measurements could define these correlations and exclude any local hidden parameters. It was not until early 1980-s that famous experiments were performed finally eliminating a possibility of local hidden parameters.

In 1980, Alan Aspect experimentally proved that EPR-paradox in quantum world takes place indeed. Special measurements of state of EPR-particles indicated that EPR-pair not only has a common origin, but in addition, that one of the photons somehow "get to know" the way the second one was changed. In further experiments, existence of EPR-paradox was affirmed, even if particles of EPR-pair were removed from each other over a distance of 10 kilometers or so.

In 1993, Charles H. Bennett and his colleagues worked out a method to transfer the quantum state of some object of the microworld to another quantum object by means of EPR-pair and called this method "quantum teleportation". In 1997 a group of experimentalists under the direction of Anton Zeilinger for the first time implemented quantum teleportation of the photon state in the University of Innsbruck.

In such a way, researchers keep on improving the process of quantum teleportation. In 2001 Danish scientists managed to link gas particles spaced at a substantial interval from each other, by transmitting information about quantum state from one particle to another by means of laser. The quantum teleportation between two gaseous clouds were attained by Eugene Polzik and his colleagues in Orkhus University. They succeeded in coupling about million of cesium atoms, whereas the previous record was only four atoms.

Scientists of Australian National University destroyed a laser beam and nearly instantly recreated it in another point in space; in other words, they teleported photons of the laser beam. In contrast to previous similar experiments, the physicists managed to obtain the required result in 100 percents of cases. Ping Koy Lam, the head of the task group of Australian University, claimed that the first atom of solid substance was likely to be teleported within near three to five years. However, as most scientists admit, a task to teleport a human remains almost impracticable. Even teleportation of atoms, as compared to that of photons, is much more complicated process. It is even harder when dealing with molecules. It is basically possible (though practically very difficult) first to transfer a molecule to a minimum-energy state (ground state) causing it to radiate a certain sequence of photons. These photons will find themselves in a certain superposition containing all the "quantum" information, which was available in the molecule. Thereafter, it is possible to teleport photon states by means of EPR-pairs. Moreover, it is also required that the classically measured information about a molecule. If molecule of minimum-energy state is present in the receiver, then this molecule, by interacting with teleported photons in a required order, will transform to the quantum state identical to that of the initial one. Consequently, the quantum state of molecule of a certain material will be transmitted, actually with velocity of light. When that happens, the quantum state at the transmitting side will be destroyed.

The human organism comprises about 10^{27} atoms. To save and transmit information on properties of that number of particles seems to be practically unachievable. "Theoretically, nothing prevents us from doing that, but complexity of the problem is such that now no one seriously thinks about the solution" – states Ping Koy Lam.

Being a method of information transmission, quantum teleportation has found its application in quantum computers, whose information is stored in the form of a set of quantum states. Impossibility to wiretap and copy transmitted information is considered to be an advantage of such computers. Those researchers, who, nevertheless, wish to answer the question of "How to teleport matter, but not its state?" have to seek for more perspective theories and techniques.

Teleportation of material objects

Those people, who really want to realize instant spatial transportation of objects, i.e. teleportation, should refer to studying properties of Space and Time. Quantum teleportation has a certain finite velocity that cannot exceed that of light. The genuine teleportation assumes that an object should set off from a starting point to a finishing point (these points differ by a certain distance X), at that the transference time comes to zero. The object to be teleported is not changed or taken to atoms to be gathered later at a distant point of space according to information transferred to this point. (Fig.1).

The object disappears from one place and simultaneously appears in another place. How is that possible? A body will disappear from point A and appear in point B if to bend space in such a way as to let point A and point B coincide. Then the object will instantly appear in point B since there is no interval between points A and B. Teleportation could be realized by a device, which would make it possible to superpose points A and B.

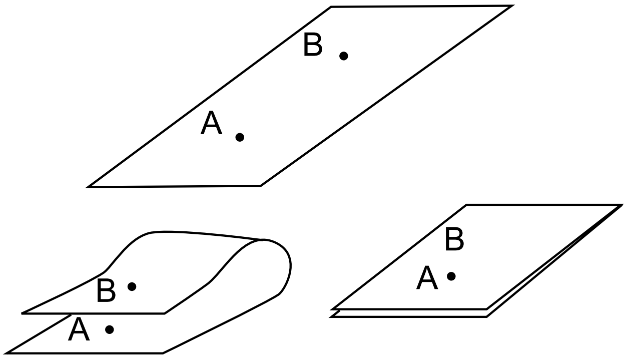

Fig. 1

Space is supposed to be unbendable without enormous energy consumption. However, Alexander V. Frolov points out that this issue is not so unambiguous: "Normally we consider space in connection with "natural" course of time existing in it. That is some degree of space curvature. Let us assume that it can be changed. The space curvature can be considered as acceleration or deceleration of time. Acceleration always demands, for instance in mechanics, some energy consumption. But if we "decelerate" time, energy is liberated and it can be accumulated for subsequent use".

K.Z. Leshan suggests surrounding a transferred object with a closed surface consisting of vacuum holes. Inside such a "hole sphere" there is geometry similar to that of a black hole. This place is absolutely isolated from the external Universe. No radiation is able to penetrate through the hole in space and time. For an observer, who is inside the hole sphere, the distance between sphere center and its border is infinitely large, since spatial metrics is ever-varying from the center to the border. Distances between the points is continuously shortened so that the distance between any couple of points comes to zero at the very hole surface. A matter transmitter can have inner or outer hole surface. In the first case a teleportation station has a spherical chamber to place the object. Equipment to produce holes should be located upon the outer side of the sphere. Such a station is capable to provide hundreds of launches a day by instantly expulsing spaceships to deep space over distances of millions of light years.

From microcosm towards macrocosm

Simeon Bocharov (member of Chemistry and Biochemistry Department, University of Delaver, Newark, USA) considers teleportation phenomenon using microobjects as an example and applying an interesting conception of protomatter.

In S. Bocharov's opinion, many paradoxes of the present-day science could be solved under condition of considering the whole existing matter to be a single continuum, i.e. protomatter. At this approach, the microcosm objects are not independent ones, being represented in the form of distortions of the very continuum and its manifestations here and now. Protomatter, whose distortion degree is beyond modern possibilities of detection, corresponds to vacuum. In the present-day paradigm redistribution of protomatter distortion corresponds to motion of particles. The important peculiarity is a refusal of such concepts as structure, dimensions, mass and other macroscopic characteristics with respect to microcosm objects, since here they are not considered as separate entities.

For experimental justification of his theory, S. Bocharov suggests to consider teleportation of microobjects incorporated in fullerens under low temperatures and pressure.

As a basis for reasoning he takes the fact that identical or similar conditions, whereon distortions are located, bring them to the state of identical or similar by lability/diffuseness. And vice versa, distortions, which are similar by lability/diffuseness, influence their near surroundings, creating, as a result, similar macroscopic states. In such a case protomatter distortion (in some spatial domains and under similar conditions) will cause appearance of similar distortions in another domain, whose degree of removal depends on presence of other distortions and their characteristics. In other words matter will be teleported from one spatial domain to another.

Possibility to realize teleportation of microobject causes the well-known phenomenon, which is nowadays described as embedding of particles into inner hollows of spheroid fulleren molecules without reacting with them. As object of teleportation there can be used distortions classified as elementary ones, such as hydrogen, helium or stable matters (noble gases). In both cases the influence of objects upon surroundings is minimized.

The researcher believes that in case of successful experiment, new pattern of matter structure will be confirmed, and valuable teleportation of microobjects (with prospects of such possibility for macrocosm objects) will be demonstrated.

One more approach to teleportation of physical macroobjects was considered in the work "Practical application of time rate control" (New Energy Technologies No. 3, 2001) by Alexander V. Frolov. It is assumed that density of space (aether) energy determines rate of passing of any processes including the very process of matter existence. Changes of aether density (increase or decrease) must result in the emergence of a force analogous to that of buoyancy, though acting towards the fourth dimension. This "chrono-motive force" (CMF) is also an analogue of electromotive force (EMF) and can be generated by analogy with electrodynamics. According to Alexander V. Frolov, laws of quantum mechanics as to discretization of levels of energy of material system, which exist in the domain of increased or decreased aether density, are valid in the macrocosm as well. Teleportation (as a transition from one state of the system to another one) can be studied with electron transference from one orbit to another as an example with the only difference that for the teleported object not only its location, but also the very spatial properties are changed. In aether of different density (after discrete transition) the same object will have different space around it, wherein time is decelerated or accelerated. Experiments of this field have already been in development stage.

Therefore, modern theoretical physics has handed the problem of teleportation to researchers-experimentalists, who possess sufficient breadth of mind. Perhaps, using teleportation technologies, they will manage to fundamentally alter the process of space exploration and raise our civilization up to radically new development level.

ELECTRIFYING TIMES

an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra
capacitors,
fuel cells, microturbines, free energy systems,
events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

"Mass defect" in Home Conditions

Prokofy V. Sherbak, Russia

Mitinskaya Str., 40-1-244, Moscow, 123430, Russia
Tel.: (095)950-4135
Email: Prokofy@mail.ru

The idea to conduct experiments on "mass defect" demonstration appeared after a cycle of seminars, which took place in1993-1996 at the Moscow Aviation Institute under the direction of V. I. Patrushev, Doctor of Technical Sciences, Professor of "Designing of aerohydrodynamic systems" Department (the author of the article is also a graduate of the Department). At the seminars, an enterprising group studied and developed an applied theory of "Displacing fields" by Alexis V. Murlikin and an associate "Theory of energy exchange processes". Later the group was called **"Group studying non-inertial transference"** (non-inertial natural processes).

The experiment was aimed to test some theoretical suppositions, namely, that it is possible to influence on existent energy flows which input and output to the matter. At that the simplest way to influence indirectly is to change the medium the flows come through. In general, optic experiments on luminous flux refraction show the same. But light refraction and change of speed of light propagation at passing through different mediums is one thing; and change of mass of a material body is something new that has not ever been dealt with. Moreover you can consider light (i.e. electromagnetic radiation) to be the secondary manifestation of output energy flow. Besides mass defect is very important for nuclear physics because it "helps" to develop this branch and stimulates drawing "energy dividends" in splitting reactions, decay reactions and nuclear fusion reactions. But unfortunately we can not touch element nucleus and even million nuclei while we can easily touch metallic balls of 1 gram weight.

Let us note that a concept of mass is one of the most uncertain in physics. We can not measure mass directly. It is possible to do it only indirectly, by gravity, using scales (it does not matter what kind of scales are used), or by kinetic momentum (in this case mass is a measure of inertia) by means of dynamometers.

Thus it is evident that change of WEIGHT shown by scales is not MASS change as measure of matter quantity. Since one of the most important points of the theory of "Displacing fields" is creation of a vehicle of new generation, then one of the major tasks is to overcome gravity (to decrease weight). And this "home experiment" shows the way to solve it. However we can call it "home" only in part, because for weighing there were used electronic scales (mass-comparator with measurement accuracy of 0.1 microgram. (See Fig. 1).

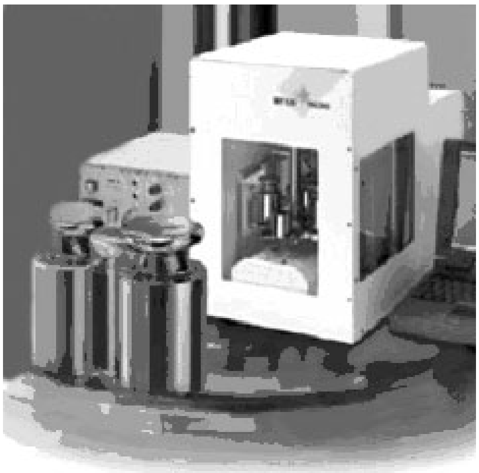

Fig. 1

Precision scales METTLER TOLEDO allows making:
High-accuracy weighing in the range
from 0.1 microgram to several tones
Measurement with resolution up to milliard points

Two sets were designed for weighing. The first one had an aluminum (Al) ball-kernel which weighed 1 gram and a lead (Pb) casing-shell consisted of two parts (half-spheres) which also has weight of 1 gram. The second set had on the contrary a lead kernel and an aluminum casing. The sets were disassembled and then hermetically assembled to make a single device with no gaps (See Fig. 2).

Since it had to be done with great accuracy and it was a piece article, not a Moscow product company undertook to produce it. And only one of the students, Nicholas Sorokin, a young boy with skilful fingers, had helped. He invented a technology which allowed making balls of the necessary size at home conditions. He was literally pickling every micron of metal with acid and finally got the necessary size and weight.

So, the weight of a non-assembled set was: 1 gram (weight of kernel) and 1 gramm (weight of shell). Weight of an assembled set, according to the "Theory of energy exchange processes" by Murlikin, is not equal to the simple sum of the non-assembled components, and mass defect would be different for various sets right up to the change of sign.

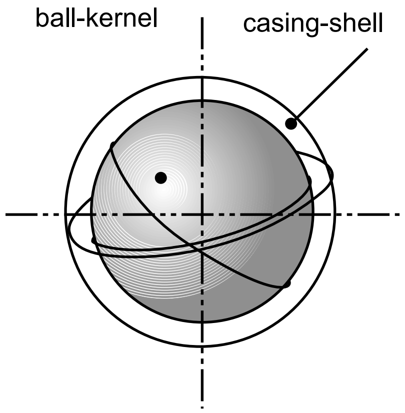

Fig. 2

Set of experimental balls

More than 50 experiments was carried out on weighing for every set. Assembled sets as well as non-assembled sets were weighed. **It was confirmed that there was an interaction between input and output flows for MATTER MAINTENANCE. It became apparent in the form of mass defect. Defect appeared in the sixth sign, i.e. it comes to the range from 0.0002% to 0.0007% of "kernel" mass (or "shell" mass). At that the lesser mass defect corresponds to aluminum shell, while the greater one corresponds to lead shell. Accuracy of weight measuring by the mass comparator exceeded the acquired result by 10 times. So we could not tell about inaccuracy of measurements.** So we had to apply to the "Russian center of testing and certification" which was very famous all over the world (Rostest-Moscow).

In general that is all, we have got an answer for ourselves, and the period of testing is behind. We have to advance; very much is still to be done. In addition let us mention one more surprise we got during the experiments. The matter concerns some time delay in receiving final results of every weighing. A qualitative picture of it is represented in Fig. 3. We did not expect that this effect would appear at such a rough level of measurement of energy exchange processes.

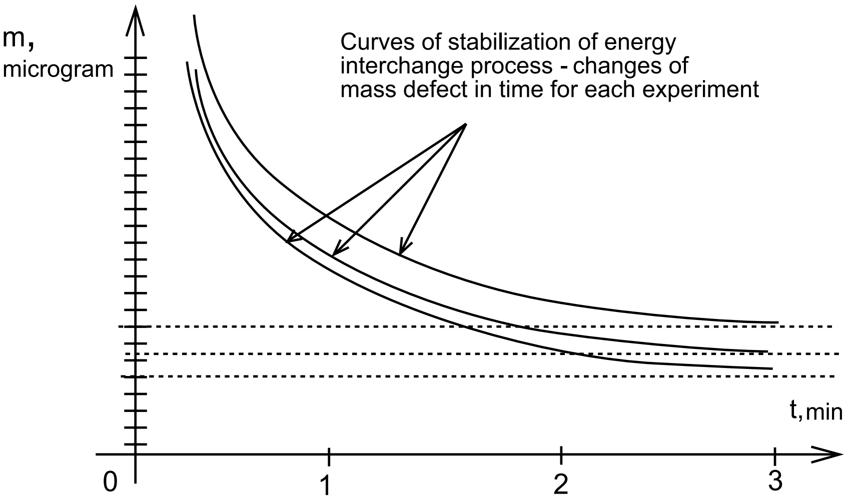

Fig. 3

Time stabilization of the process for receiving final weight

The subsequent report and detailed discussion of the experimental results by specialists (including physicists of various specializations) resulted in an unexpected non-recognition of facts and distrust. The main argument was a demand to increase the number of experiments up to 5-10 hundreds. And only then scientists would be ready to accept facts. It was quite a predictable reaction for the orthodox science in the middle of 90s. It remains to hope that if the discussion took place today it would have a paved way. However it is not a fact.

In conclusion let us note that this experiment on revealing mass defect ranks with famous experiments, in which rotating gyroscopes also demonstrate mass defect (decrease of weight) and even one of higher order than in our experiment. But in return we have no rotation, no chemical and nuclear reactions, while mass defect is presented.

Let us add that it is possible to complicate the task: to make a double casing and to choose other materials for "kernel" and "shell". (See Fig. 2). In our case low prices and availability made us use Al and Pb. The results will be more interesting, for example, with Li and Os (and with other supertransuranic elements).

Aether as Unified Field

Alexander M. Mishin, Russia

Planernaya Str., 79-208, St.Petersburg, 197373 Russia
E-mail: samish@mail.ru

Based on the original observations and experiments, the article gives more precise definitions of aether model, than which was previously elaborated. General laws of the Universe aetherodynamic are considered. There is described a new physical reality in the form of quasi-real aetherodynamic structures ("topological harmonics") generated by dynamics of the substantial world.

At the present-day stage of physics development words of Maugham, an English philosopher, take on particular topicality: "The great verities are too important to be new". Indeed, we are increasingly making certain that our remote ancestors were familiar with fundamental laws of the Universe; furthermore, they knew these better than we do. But which methodology of cognition of Nature and what kind of information sources did the ancient scientists have? Today, we can openly say: they had not ordinary experience only, but also esoteric, pretersensual one, without feeling any necessity to be divided into materialists and idealists. For more details on evolution of physical ideas, refer to work [1].

In discovering all over again, I also had to elaborate a "cellular taction" strategy, which enables the most important characteristics of aetherodynamic processes to be sensually perceived and evaluated. The main thing is that all my sensory observations are definitely reproduced at laboratory experiments while the simplest physical equipment is used.

We may get reliable knowledge about Nature from Nature itself only. Thus foundation of classical physics describing usual substantial world is also based on empirical facts studied by Galilei, Newton, Faraday and many other physicists-natural philosophers. I also would like to defend great Newton against attacks on his adherence to inductive methods which really brought imperishable values into science.

A new convolution of cognition is possible on condition of introduction into science of new concepts based on empirical study of physical entities previously unknown. Thereupon, I share the thought of Elena Ventzel, professional mathematician: "Mathematical methods are not useful, but harmful until phenomenon is not studied at a pre-mathematical, humanitarian level". It is ideological and logical language that may be used today to discuss the aether problem, whose solution calls for comprehension of natural-science facts, which are beyond traditional notions settled over centuries.

The long-term investigations have confirmed that exhaustive definition of a fundamental natural entity was given by Rene Descartes: "Universe does not contain anything but aether and its vortexes". This is not merely a postulate; this is an objective summary of empirical observations of Nature phenomena. In any academical model of microcosm there is no answer to the question: what kind of substance are electrons, quarks, strings made of? That is the very substance our forefathers called aether. Aether is unified field indeed; and it appears that for nearly his whole life Einstein had been searching for something that was under his foot.

It must be recognized that there is no undistorted, quiescent aether in Nature; as well as no zero-size vortexes exist. Therefore, in a customary sense, the aether concept is pushed behind transcendental horizon and turned into an inconceivable entity whose mystique is tantamount to religious ideas. However, all quantum vortex-wave distortions of the superfluid three-dimensional aether, which have an appearance of tori, threads, solitons and occupy spatio-temporal spectrum from zero to inter-universe scale, are material and contain energy under scientific interpretation. Any spaces such as turbulent fields, physical vacuum, gravitational fields and, of course, matter are material. All of them are etherodynamic structures of the real world. Spatial hierarchy of aether vortices is referred to as quantum stairs, fractal structure or "matreshka" [2-5].

To understand basic laws of aetherodynamics it is necessary to give a general definition for classical matter and other aether states. As a first approximation, let us give the name of matter to aether microscopic **left-screw** self-oscillating vortex-wave systems (elementary particles), whose existence and parameters are conditioned by higher hierarchical systems. This is our displayed world, with its electromagnetic fields and fundamental constants. Let us refer to vortex-wave structures of any vortex size, which in the considered space domain are **balanced** by a number of left- and right-screw vortices corresponding to matter and antimatter with prefix "quasi" to a free aether. Stable vortex lattices with alternating direction of helicity can be also considered in this case. Under natural conditions everything but classical matter or electromagnetic and gravitational fields can be related to the free aether. In the first place, these are the so-called physical vacuum and vortex fields, which attend gravitation. They represent anisotropic nonlinear subspaces and occupy definite places on spatio-temporal spectrum of the world aether. Particular free subspace creates an aggregate of topological harmonics [5] attending all the physical processes. Satellite of the gravitational field is distinguished because it belongs to specific class of stable vortex lattices.

Apparently, in general case, energy of free aether is virtual for our world due to its helical symmetry and quasi-matter, from the view of difference in sizes of typical vortices which correspond to protomatter. But this is not the entire point. There is also superfluidity of aether, whose vortex viscosity has marvellous properties. As a result, there are created phenomena of inertia and energy-information barrier separating material world from constant influence of quasi-real structures or turbulated space.

Classical inertia, as vortex viscosity, arises in the microscopic and more fine-structure free aether (in Newtonian space) and ever regularly reveals itself during acceleration of material bodies. However everything is different at stage of macrovortex aether, where inertia has properties of adaptation. Here it is affirmed that inertia phenomenon is "spread" throughout the spatial-temporal spectrum of vortex distortions of aether field. At that nature of inertia is changed. Within topoharmonic subspace in laboratory scales, the inertia law "diminishes" in time according to exponential dependence, whose time constant is determined by specific conditions and varied over wide limits (from seconds to many days). On the expiry of this time, inertial intercoupling transforms to the law of flicker effect. It may be said that flicker effect specify penetrability of energy and information barrier after expiry of time of paraphysical regular phenomenon [6]. Consequently, most of time the interaction with free macroaether is of random character and should be evaluated by mean observation.

To extend conceptual scope of our investigation it may be added that in classical hydrodynamics a liquid, along with ordinary inertia, has determined vortex macroinertia, which is created by vortices of the liquid itself. Therefore, it is necessary to keep on comparing processes at all levels of the unified field of aether. For instance, electric inductance is a kind of specific inertia, and topological harmonics of electromagnetic and other phenomena are subjected to the energy-information barrier (under laboratory conditions).

In the above arguments, the principle of relativity is latently and unconsciously presented. However, this complicated question needs to be separately considered. Physical nature of the energy-information barrier and reason for dependence of the inertia properties upon the size of vortices of free aether remain unclear, too. However, it can be supposed that classical elementary particles, as distinct from protomatter, have self-oscillating properties and are commensurable with vortices of the surrounding free world; and that the above-mentioned energy-information barrier does not exist for them on microscopic section of the spectrum. The gist of the barrier itself comes to "memorizing" topoharmonic processes by Earth's gravitational fields and laboratory equipment. Turbulent fields (physical vacuum) do not have such a "memory".

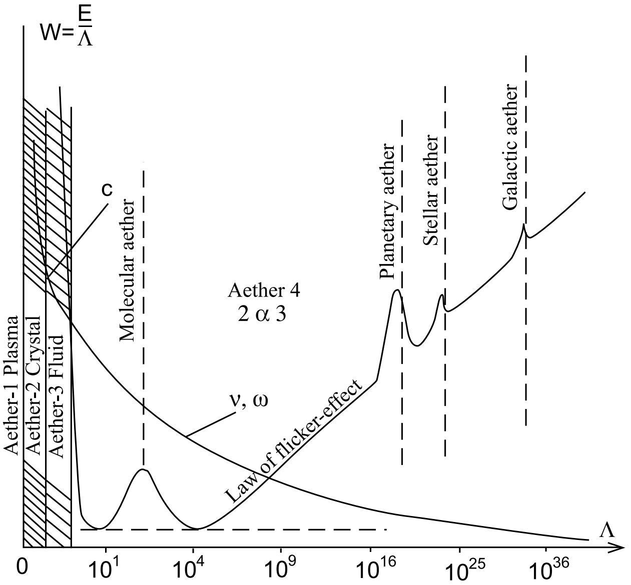

Fig.1

Fig.1 depicts diagrams of spectral density of vortices' energy; diagrams of rate of interaction transfer v and circular frequency ω depending on sizes of vortices of the world aether Λ (in picometers). As we can see, velocity of distortions propagation v is equal to velocity of light C in the physical vacuum, i.e. in electron sizes, exceeds C in nanocosm and tends to zero toward megacosms. Circular frequency of vortices rotation and course of physical time have the analogous qualitative dependence. However, vortices energy W has a feature and grows toward both ways from the lab scales; the aether numbers have been assigned conditionally, but all the specified phase states reflect some physical reality.

One of the most complicated questions is a definition of vortex matter in each particular case. The diagram demonstrates energy of free aether in the form of protomatter (simplest elementary particles – vortexes), and energy of classical elementary particles as helically definite self-oscillating emissions of aether field is localized in the most nonlinear subspace (physical or substantial vacuum). Unification of the elementary particles into atoms and macrobody does not denote direct transition to parallel world with different fundamental constants, whereas change of vortex size in the diagram exactly implies such a transition, though at a level of protomatter. This is a very intricate and crucial point. For example, a star is, unconditionally, an elementary particle of the relevant level of the Universe, even though comprising the ordinary matter.

Left wing of energy spectrum reflects property of the turbulent field of aether to gravitate, to sink down to the "bottom" transferring energy to the vortices of lesser scales. This phenomenon of distributed "gravitation" is available in every space point and explains a least-action principle. An illustrative analogue of this regularity is represented by the Earth's gravitational field and structure. Dense plasma assumed as iron core is situated in the center, and higher - there are magma, solid crust, water and air. These phase states exist in free aether within universe scales as well. Consequently, in femtocosm, and maybe at an even deeper energetic "bottom", there must be a plasmic subspace; evidently, it is assignable that aether translated from Greek means "fire".

Increase of energy to the right is connected with extending a dynamic band of the vortices spectrum, as the scale increases. In our world, this increase follows the flicker effect law, but this is only what we perceive across the energy-information barrier. Actually, at this section the vortex energy must increase well more strongly, because **our observation capabilities are limited** by accessible matter processes. Turning back to stars, one may affirm that modern science is unable to evaluate their real energy (evaluations are considerably underrated). The Galaxy vortex entrains such deep levels of multidimensional space-aether to its rotation that this causes violation of the Newton's law of inertia. Notion of "hidden mass" is unnecessary for explaining dynamics of galaxies. It is quite enough to take into account that differential rotation entrains the fine-structure aether (inertial subspace), and this process results in decreasing the inertia and external centrifugal forces. Unfortunately, Fig. 1 reflects only a characteristic size of the galactic vortex and does not visually demonstrate its active band of the spectrum. The finer is the space structure, the more inertial and closer to the absolute reference frame is the space. Obviously, the non-homogeneous, stereodynamically multidimensional and multiphase aether requires a perfectly new theory of random fields, which would also make allowance for the phenomenon of the material world.

Energy motion processes or spectral energy-cascades along abscissa axis are of decisive importance for existence of our world and the Universe. These processes have unusual properties. The "violet" cascade is attended with reduction of vortices size, absorption of matter energy and tends to increase energy concentration at the "bottom" of the world aether (near ordinate axis on Fig.1). This regularity experimentally confirmed in both spatial and temporal domains is a direct consequence of the above-mentioned phenomenon of distributed gravitation and clarifies an existence of the left wing of the spectrum. In classical physics "the violet" cascade effect is reflected in the least-action principle as well as in the thermodynamics laws.

The "red" energy cascade (right wing of the spectrum) corresponds to increase of vortices size and is accompanied with energy liberation. To start it up under earthy conditions special technologies are called for [7]. This has been confirmed with laboratory experiments and underlies the "perpetual motion" theory, where energy is pumped from nanocosm to the classical material world. Conception of artificial biofield [6], I have developed, considers methods of creating the aether vortex-wave macroscopic-scale distortions in physical systems. As a matter of fact, we deal with **ways of extending of natural spatio-temporal spectrum of aether (which settled into itself) by creation of a subspace of the topological subharmonics of the classical elementary particles, atoms and molecules**

(physical bodies), within the context of the etherodynamics fundamental law [5].

Even within action of fluctuating energy-information barrier, the spectrum right wing exhibits existence of natural processes in Nature that, supposedly, violate the known laws of thermodynamics [8]. But, as far as Nature cannot conflict with itself, we have to recognize that there is another physics, which "works" in geographic and greater scales, where role of the energy-information barrier demands to be reappraised. To all appearance, the universe vortex partialyl entrains plasmic subspace of aether into its rotation, and conditions for realizing "red" energy-cascade and generating matter are created on the sections of the maximum density gradient ("spectral" non-linearity). But the same way as a fish in ocean does not know reasons for storms and currents, we will never learn which force untwists the Universe handwheel, whose energy feeds our substantial world.

There are two components of free aether field that are the most essential for the present-day physics. The first one (left wing) is positioned on the spectrum beside the "bottom" in the form of dense liquid-crystalline "plasmatic" medium transforming to the Newtonian inertial subspace (physical vacuum). Since inertia is as regular as electric inductance is, the first subspace proves to be especially material and not such free, because it has properties which were imposed by the Universe rotation and distributed "gravitation". Let us conventionally limit the spatial spectrum of the first component with a nucleon size. The second component of aether space (right wing of the spectrum) stretches towards larger scales.

Except for known cosmic objects, this is an unknown ocean of the macrovortex and really free aether, or topoharmonics, which are born by dynamics of the substantial world. It is an experimentally proven fact that topoharmonics are generated by photons, electrons, and all the moving bodies [5, 9].

The second component has an ability to conversely affect the matter it was generated by as well as other material objects via "flickering" forces [10, 11]. This special subspace turns out to be doubly virtual, a genuine phantom. That is why, it has been winning the status of a real physical entity with such a difficulty, while Chinese philosophers were familiar with the same thousands years ago, separating entities "yang" and "yin".

The laboratory experiments indicate that topoharmonics have not only properties described in [5], but also an ability to interact with the space, creating a static domain of the excited aether in the course of propagation of the vortices-waves. At that physical bodies placing within this space volume are "charged" [2, 9]. The mentioned active domain (horizontal route) is "separated" by the Earth's gravitational field as follows: quasi-substantial constituent rapidly settles down creating vertical flow, but some phase structures are disintegrated very slowly (days, months, years).

The surprising thing is that biosystems equally belong to both aether components, not experiencing any constraint from the side of the energy-information barrier. The so-called human biofield is a topoharmonic part of physical body. It occupies its place on the spatio-temporal spectrum (Fig. 1). Quasi-substantial constituent (yang) and antiquasi-substantial one (yin) are strictly balanced, and acupuncture points and chakras are an analogue of Hartmann's geophysical grid, which is well-known to paraphysicists.

Conceptual generalization of physical ideas in the context of unified field of aether gives a constructive, cognitive impulse in solving of different problems, such as Nature's choice of exclusive laws for existence of "left-screw" material world as well as the problem of flicker-effect. At that, researching into the physical processes responsible for proton positive charge is of fundamental importance.

In conclusion, let us take note of conventional character of dividing the aether field into two components by the form of energetic spectrum. As a matter of fact, topoharmonics are generated by matter towards both larger spatial frequencies (subharmonics) and smaller ones (harmonics) and fill all levels of the Universe. The aforesaid is aimed at making perception of the most studied aspect of the **new physical entity** (vague due to diversity of hypothetical descriptions which exist in publications on this subject) more accessible.

Reference

1. A.P. Smirnov, I.V. Prokhortsev. Order principle. – Saint-Petersburg: ZAO «P and K», 2002. - 296 p.
2. A.M. Mishin. The aether model as a result of the new empirical conception. New ideas in natural sciences. (On materials of International Conference). Part 1 "Physical". - St.-Petersburg: RAS, 1996, p. 95-104.
3. A.M. Mishin. Fundamental properties of aether. Fundamental Problems of Natural Science. Series «Problems of the Universe research», fascicle 21 – Saint-Petersburg: RAS, 1999, p. 176-176.
4. A.M. Mishin. Physics of parallel worlds. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23 - Saint-Petersburg: Publishing

House of the Saint-Petersburg State University, 2001, p. 258-269.
5. A.M. Mishin. Etherodynamics basic law. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 206-218.
6. A.M. Mishin. Physical system with artificial biofield. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23, - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2001, p. 258-269.
7. V.F. Zolotarev, V.V. Roschin, S.M. Godin. About space-time structure and some interactions. - Moscow: Publishers «PREST», 2000. - 309 p.
8. V. Starr. Physics of phenomena with negative viscosity. - Moscow: MIR, 1971.
9. V.V. Kvartalnov, N.T. Perevozchikov. Discovery of a "non-physical" component of OKG radiation. Journal «Parapsychology and Psychophysics». - Moscow: 1999, No. 2 (28), p. 64-67.
10. A.M. Mishin. Longitudinal thermomagnetic effect. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 219-228.
11. A.A. Efimov, A.D. Shpitalnaya, Y.A. Zakoldaev. Eruptive prominences and earthquakes in the view of global space anisotropy. / «New Ideas in Natural Science». Series «Problems of the Universe research», fascicle 19, part II. – Saint-Petersburg: RAS, 1996, p. 403-413.

GLOBAL ENERGY

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Lately, the notion "Global energy" has become widely spread as a prize corresponding to this notion has been instituted in Russia. Since the Euclidean times, an unwritten law exists in science: to give definition to the notions, which are involved in scientific analysis. It is done for the purpose that everybody who uses this notion can understand the sense implied in it.

The Global energy problems have been discussed for a long time; they are well known. Depletion of natural energy carriers (oil, gas and coal) is the first problem. Environmental safety of energy carriers is the second problem. It is generally recognized that the solution of the issues connected with the aggregate of these two problems is of a global character. The results of scientific research, which should deal with global energy, originate from it. First and foremost, they are such scientific investigations, which result in the possibility to use an inexhaustible and environmental friendly energy carrier. It has been known for a long time. It is hydrogen produced from water.

Why hydrogen and why from water? At combustion of hydrogen, water is regenerated; thus, the energy carrier is inexhaustible. As far as inexhaustibility and environmental safety are concerned, hydrogen has no competitors. However implementation of these qualities is restricted by large energy expenses for hydrogen production from water. For hydrogen production, modern electrolyzers use from 10 to 20% more energy than it is produced at combustion of hydrogen.

It is easy to imagine what financial and intellectual resources of the world are included in the search of the ways to reduce energy expenses for hydrogen production from water. In Russia many scientific institutions of applied research and educational institutions deal with this problem. There is a Research Hydrogen Institute.

The associations of scientists on hydrogen energetics have been established in USA and Europe. This year, they carry out their regular scientific conferences, to which the author of this article has been invited.

Editorial: Let us note that it is not necessary to coin new scientific term "global energy", since quite for a long time the notion of "free energy" is used. This term represents the idea of the possibility to freely use energy by a consumer.

Energy problem is global not due to depletion of oil and gas, but due to their environmental unsafeness. However the confirmations exist that the world owners of energy resources do not bother about the problem of environmental unsafeness of modern energy carriers.

House of the Saint-Petersburg State University, 2001, p. 258-269.
5. A.M. Mishin. Etherodynamics basic law. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 206-218.
6. A.M. Mishin. Physical system with artificial biofield. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 23, - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2001, p. 258-269.
7. V.F. Zolotarev, V.V. Roschin, S.M. Godin. About space-time structure and some interactions. - Moscow: Publishers «PREST», 2000. - 309 p.
8. V. Starr. Physics of phenomena with negative viscosity. - Moscow: MIR, 1971.
9. V.V. Kvartalnov, N.T. Perevozchikov. Discovery of a "non-physical" component of OKG radiation. Journal «Parapsychology and Psychophysics». - Moscow: 1999, No. 2 (28), p. 64-67.
10. A.M. Mishin. Longitudinal thermomagnetic effect. / Fundamental Problems of Natural Science and Technics. Series «Problems of the Universe research», fascicle 26, part III. - Saint-Petersburg: Publishing House of the Saint-Petersburg State University, 2003, p. 219-228.
11. A.A. Efimov, A.D. Shpitalnaya, Y.A. Zakoldaev. Eruptive prominences and earthquakes in the view of global space anisotropy. / «New Ideas in Natural Science». Series «Problems of the Universe research», fascicle 19, part II. – Saint-Petersburg: RAS, 1996, p. 403-413.

GLOBAL ENERGY

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

Lately, the notion "Global energy" has become widely spread as a prize corresponding to this notion has been instituted in Russia. Since the Euclidean times, an unwritten law exists in science: to give definition to the notions, which are involved in scientific analysis. It is done for the purpose that everybody who uses this notion can understand the sense implied in it.

The Global energy problems have been discussed for a long time; they are well known. Depletion of natural energy carriers (oil, gas and coal) is the first problem. Environmental safety of energy carriers is the second problem. It is generally recognized that the solution of the issues connected with the aggregate of these two problems is of a global character. The results of scientific research, which should deal with global energy, originate from it. First and foremost, they are such scientific investigations, which result in the possibility to use an inexhaustible and environmental friendly energy carrier. It has been known for a long time. It is hydrogen produced from water.

Why hydrogen and why from water? At combustion of hydrogen, water is regenerated; thus, the energy carrier is inexhaustible. As far as inexhaustibility and environmental safety are concerned, hydrogen has no competitors. However implementation of these qualities is restricted by large energy expenses for hydrogen production from water. For hydrogen production, modern electrolyzers use from 10 to 20% more energy than it is produced at combustion of hydrogen.

It is easy to imagine what financial and intellectual resources of the world are included in the search of the ways to reduce energy expenses for hydrogen production from water. In Russia many scientific institutions of applied research and educational institutions deal with this problem. There is a Research Hydrogen Institute.

The associations of scientists on hydrogen energetics have been established in USA and Europe. This year, they carry out their regular scientific conferences, to which the author of this article has been invited.

Editorial: Let us note that it is not necessary to coin new scientific term "global energy", since quite for a long time the notion of "free energy" is used. This term represents the idea of the possibility to freely use energy by a consumer.

Energy problem is global not due to depletion of oil and gas, but due to their environmental unsafeness. However the confirmations exist that the world owners of energy resources do not bother about the problem of environmental unsafeness of modern energy carriers.

In history of science, the facts of annihilation of the scientists who have won success in reduction of energy expenses for production of hydrogen from water have already been registered. The requesters of these actions think that they will lose their profits with the coming of hydrogen energetics. They do not understand that this process cannot be abrupt. It is impossible to replace the infrastructure of the existing energy carriers by infrastructure of hydrogen energetics within one or even ten years. Besides, infrastructure of hydrogen energetics will not be created on a blank place. It will be integrated into the existing infrastructure of energetics gradually; and its owners will automatically become the owners of hydrogen energetics. The first step has already been made in this direction. In the USA, a decision has been taken to equip all filling stations with pumps to fill the cars with hydrogen.

It is known that a priority of results of theoretical investigations is their publication in press. Usually, such priority is a personal one. Generally a patent is a priority of the results of experimental investigations. As a rule, this patent belongs to a group of authors. A published patent is a genie released from a bottle. No finesse of the authors to hamper a reproduction of experimental data given in a patent without the participation of the authors can stop the process of their implementation. Thus, the authors or a group of the authors who have filed an application for a patent are deprived automatically of the opportunity to influence the process of practical realization of their ideas.

In Russia, the state is a main patent holder of the global patents. Due to the well-known reasons, it has proved to be unable to control the course of scientific investigations and to forecast significance of their results.

It is known that if it becomes possible to reduce energy expenses for hydrogen production of water fivefold, it will be the cheapest energy carrier. Russia has already got technology, which reduces these expenses tenfold and more. However another way of looking is more perspective. Is there any use to decompose water into hydrogen and oxygen and to use hydrogen as fuel for heating, for example, of water in heating systems? Is it possible to make water generate heat? It turned out that, it is possible.

In Russia, three firms ("Yusmar", "Termovikhr" and "Noteka") sell cavitation heating equipment with energy performance index up to 150%. Official science looks awry at this activity, since such results conflict with one of the main laws of physics: law of conservation of energy. But market profit is stronger than this law.

Meanwhile, engineering practice has already proved that additional energy in the form of heat is generated in the ventilation systems and in the systems of water cavitation. Thorough scientific analysis of this problem shows that physical vacuum is the most probable source of additional energy in the systems of ventilation and water cavitation. Valence electrons of destroyed molecules of water take energy from physical vacuum and release it during repeated fusion of these molecules.

Why is additional energy generated in the air systems of ventilation and in systems of water cavitation? Because they are mechanical systems and mechanical destruction of chemical links requires half energy as compared with thermal destruction of these links. This is the main reason why one fails to increase energy performance index of cavitation processes over 200%.

Certainly, an increase of efficiency of any process by 30 or 50% is a good result; if it is obtained, it is possible to get even better one. But what if a water molecule is destroyed not mechanically, but electrodynamically? In this case, it becomes possible to find resonance frequencies of influence on the molecules and therefore to reduce considerably expenses of electrical energy for their destruction. Subsequent fusion of destroyed molecules will release determined quantity of energy unavoidably. It is a simple idea, and it has already been implemented. At electrodynamic influence on water molecules, electrical energy is converted into thermal energy with tenfold energy performance index. It means that if we spend 1 kWh of electric energy, we will get 10 kWh of thermal energy. To obtain such results, it is necessary to have knowledge of physchemistry of the microworld, which corresponds to knowledge of the 21st century, and it has already been published. Every month more than 1000 foreign scientists become familiarized with this knowledge at
http://Kanarev.innoplaza.net and
http://book.physchemistry.innoplaza.net.
The Russian speaking readers get this information from:
http://www.n-t.org/tp/ns/if.htm,
http://www.ikar.udm.ru/sb28-2.htm and
http://www.n-t.org/tp/ts/eb.htm.

I hope that now the readers will know the essence of the notion "global energy" and will understand those results of scientific researches which can really solve the related problems.

Energy Balance of Fusion Processes of Oxygen, Hydrogen and Water Molecules

Philip M. Kanarev, Russia

The Kuban State Agrarian University, Department of Theoretical Mechanics
13, Kalinin Street, 350044 Krasnodar, Russia
E-mail: kanphil@mail.kuban.ru

There is revealed the cause of appearance of additional energy during formation of covalent bonds in the fusion processes of oxygen, hydrogen and water molecules, besides the source of this energy is described.

Introduction

Engineering practice connected with servicing of ventilation systems allows revealing appearance of excessive thermal energy in circulated air. Similar phenomenon has been registered in systems of water circulation with the devices for its active cavitation. The results of our investigations explain not only a cause of these phenomena, but they give an opportunity to perform quantitative calculations for energy processes, which generate additional thermal energy [1], [2], [3], [4], [5].

Theoretical part

An oxygen atom is the eighth element of the periodic table. It is situated in the sixth group. The structure of its nucleus is given in Fig. 1 [1], [2], [3].

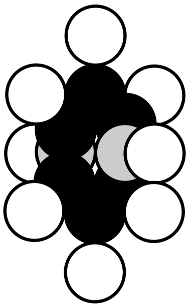

Fig. 1
Diagram of nucleus of oxygen atom: light – the protons, dark and grey – the neutrons

In Fig. 2 a diagram of the oxygen atom originating from the structure of its nucleus is given (Fig. 1). It has eight electrons. The electrons situated on the axis of symmetry are the most active ones (1, 2). Other six electrons situated in the plane, which is perpendicular to the axis line (a line of symmetry), by means of their total electric field remove electrons 1 and 2 from the nucleus at a large distance at that forming conditions for their large activity during the interaction with the electrons of the neighbouring atoms [1], [2], [3].

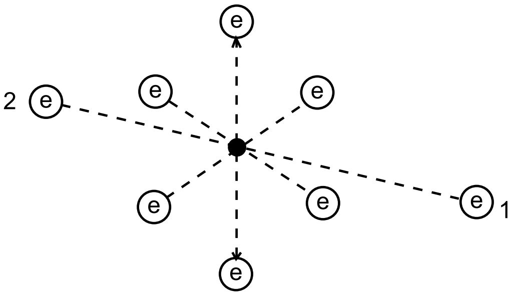

Fig. 2
Diagram of the oxygen atom

The least ionization energy of the electron of oxygen atom is equal to E_i =13.618 eV. Binding energy of this electron with the atomic nucleus corresponding to the first energy level is equal to E_1 =13.752 eV. Let us call this electron the first one. The calculation of energy indices of this electron, including its binding energies E_b with the atomic nucleus, according to the formulas (1) and (2), gives the following results (Table 1) [1], [2], [3].

$$E_{ph} = E_i - \frac{E_i}{n^2} = E_i - \frac{E_1}{n^2} \tag{1}$$

$$E_b = \frac{E_1}{n^2} \tag{2}$$

Table 1
Spectrum of the first electron of the oxygen atom

Values	n	2	3	4	5	6
E_{ph} (exp.)	eV	10.18	12.09	12.76	13.07	13.24
E_{ph} (theor.)	eV	10.16	12.09	12.76	13.07	13.24
E_b (theor.)	eV	3.44	1.53	0.86	0.55	0.38

The oxygen molecule structure is given in Fig. 3, a. It is formed by means of a connection of unlike magnetic poles of axis electrons of two oxygen atoms [1], [2], [3]. It is known that the fusion process of the oxygen molecules is accompanied with a release of 495 kJ/mole of energy, or in calculation for one molecule

$$E_b = \frac{495 \cdot 1000}{6{,}02 \cdot 10^{23} \cdot 1.602 \cdot 10^{-19}} = 5.13 eV. \quad (3)$$

What principle does the Nature follow by distributing energy of 5.13 eV between the electrons of oxygen molecule (Fig. 3, a)? Energy of 5.13 eV is a thermal binding energy between the electrons 1 and 2' of two oxygen atoms (Fig. 3, a). When the oxygen molecule is formed, it is emitted in the form of the photons by the electrons, which enter into the bond. Hence it is equal to an amount of energies of two photons emitted by these electrons. Consequently, each contacting electron emits a photon with energies of 5.13/2=2.565 eV=E_b (Fig. 3). According to Table 1, in this case the valence electrons are situated between the second energy level and the third one [1].

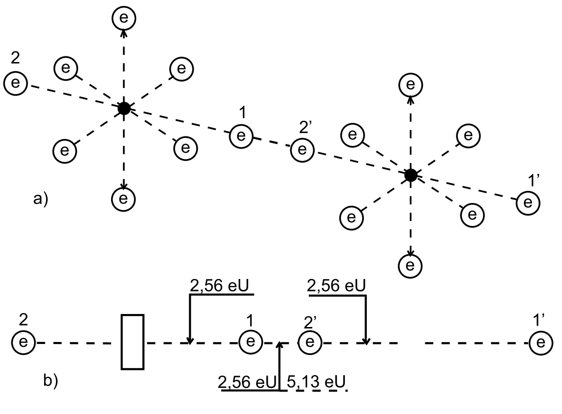

Fig. 3
Diagram of binding energy distribution between the electrons in the oxygen molecule

Two oxygen atoms are combined into a molecule in an excitation state. The excitation state is the state of an atom when its valence electrons are situated at such distances from the nuclei when the binding energy E_b between them is reduced to the thousandth of fractions of an electron-volt. In such state the atom can loose an electron and become an ion. Otherwise, without loosing electrons it is combined with an electron of the neighbouring atom by the valence electron, and a process of formation of oxygen molecule begins. It is an exothermic process when the axis valence electrons 1 and 2' emit photons, descend on lower energy levels and release 2.565x2=5.13 eV.

Let us pay attention to the fact that energy 5.13 eV is released by two electrons, which form a bond with energy of E_b=2.56 eV. In modern chemistry this bond is called a covalent bond. In order to break this bond it is necessary to use 2.56 eV of mechanical energy. For thermal cleavage of this bond, double quantity of energy is required, i.e. 5.13 eV. It is explained by the fact that the photon energy of 5.13 eV is absorbed by two electrons simultaneously. Only in this case, both electrons will be transferred to the highest energy levels with minimal binding energy E_b when they are disconnected, and each oxygen atom becomes a free one.

Thus, energy expenses for destruction of oxygen molecule depend on the method of influence upon the bond. During thermal action upon the bond it is destroyed when energy is 5.13 eV. During mechanical effect upon the bond, it is necessary to spend 2.56 eV of energy in order to destroy this bond. Therefore energetic of fusion process of the oxygen molecule depends on method of its destruction.

After thermal destruction of the oxygen molecule process of its formation begins from emission of the photons with energies of 2.56 eV by both valence electrons, and the previous electrodynamics binding energy (E_b=2.56 eV) is restored between the electrons of both atoms.

Thus during thermal destruction of the oxygen molecule the same amount of thermal energy is spent than that which is released during its further formation. No additional energy appears during thermal dissociation of oxygen molecule and at its further fusion.

If oxygen molecule is destroyed by a mechanical method, then it is necessary to spend 2.56 eV of mechanical energy for this purpose. Valence electrons of oxygen atoms are in a free state at lack of energy, which corresponds to such state, as there is no process of absorption of 2.56 eV of energy by each of them. The electrons cannot remain in such state; they should replenish immediately the energy, which they have failed to receive during a mechanical break of the bond between them. Where should they take it from? There is only one source: the environment, i.e. the physical vacuum filled with aether. They convert aether into energy of 2.56 eV immediately. The next stage is a connection of two oxygen atoms, whose valence electrons have replenished the reserves of their energy by means of aether. This process is accompanied by emission of the photons with energies of 2.56 eV by two electrons. Thus energy of absorbed aether is converted into thermal energy of the photons. If we spend 2.56 eV of mechanical energy for destruction of oxygen molecule, we will get double quantity of energy (2.56x2=5.13) eV during further fusion of this molecule. Additional energy is equal to 2.56 eV.

Much experimental data show that in ventilation systems thermal energy of circulated air exceeds electric energy spent for a fan drive. Now we know that this energy is generated at mechanical destruction of covalent bonds in the molecules of gases, which the air consists of.

Using the above-mentioned method, let us analyse energetic of water molecule, which sometimes generates additional thermal energy. A water molecule consists of one oxygen atom and two hydrogen atoms. Binding energies E_b of the hydrogen atoms with its nucleus are given in Table 2 [1], [2], [3].

Table 2

Spectrum of hydrogen atom

Values	n	2	3	4	5	6
E_{ph} (exp.)	eV	10.20	12.09	12.75	13.05	13.22
E_{ph} (theor)	eV	10.198	12.087	12.748	13.054	13.220
E_b (theor.)	eV	3.40	1.51	0.85	0.54	0.38

It is known that combination of hydrogen and oxygen is accompanied by an explosion, but its cause remains unknown. Let us try to find it.

Energy of fusion of hydrogen molecule is equal to 436kJ/mole, or 4.53 eV per a molecule. As the molecule consists of two atoms, then the above-mentioned energy is distributed between them. Thus energy of one bond E_b between the hydrogen atoms is equal to 2.26 eV (Fig. 4). At mechanical destruction of this bond 2.26 eV is enough. At thermal destruction of this bond double quantity is required (2.26x2=4.53 eV) [1].

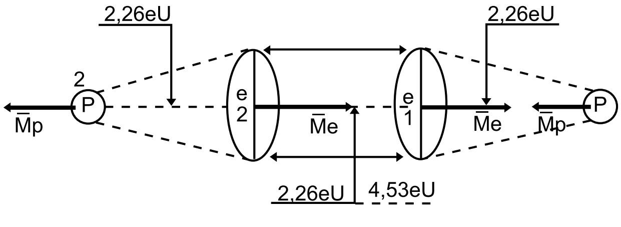

Fig. 4

Hydrogen molecule

In order to form two water molecules, it is necessary to break two hydrogen molecules and one oxygen molecule into atoms. At mechanical destruction of covalent bonds 2.26x2=4.53 eV is required to break two hydrogen molecules and 2.56 eV to break an oxygen molecule. Sum of these energies is equal to 7.13 eV. If the destruction processes of the above-mentioned molecules are carried out with a thermal method, then 4.53+4.53=9.06 eV is required for the destruction of two hydrogen molecules, and 5.13 eV is required for the destruction of one oxygen molecule. In total, it is equal to 14.19 eV. The difference between the energy spent for mechanical and thermal destruction of covalent bond of hydrogen and oxygen molecules is almost double.

It is known that during fusion of one mole of water 285.8 kJ or 285.8x1000/6.02x10^{23}x1.6x10^{-19}=2.96 eV per a molecule are released. As a water molecule consists of one oxygen atom and two hydrogen atoms, 2.96/2=1.48 eV falls per the bond (Fig. 5). Hence the electrons of hydrogen and oxygen atoms in water molecule are between the forth energy level and the fifth one at the usual temperature (1.48/2=0.74 eV = E_b), Table 1,2 [1].

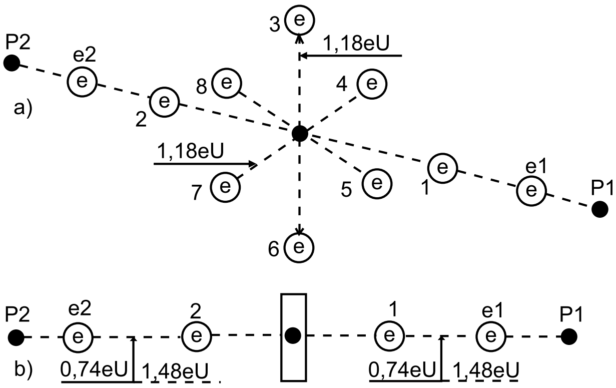

Fig. 5

Diagram of water molecule:
1, 2, 3, 4, 5, 6, 7, 8 are the numbers of the electrons of oxygen atom; P_1, P_2 are the nuclei of the hydrogen atoms (the protons); e_1 and e_2 are the numbers of the electrons of hydrogen atoms

Thus when two hydrogen molecules $2H_2$ and one oxygen molecule O_2 are destroyed by the thermal method, 14.19 eV are spent. As a result of fusion of two water molecules ($2H_2O$), 2.96x2=5.98 eV is released. There is some disbalance here since fusion process of water molecule is an exothermic one and 2.96 eV is released by one molecule. The given calculation shows that (14.19-5.98)/2=4.10 eV is absorbed during fusion of one water molecule. What is the cause of this contradiction?

The oxygen atom in the water molecule should reduce its volume when the transition from gaseous state into liquid state takes place. It will happen when the ring electrons of oxygen atom descend on lower energy levels (nearer to the nucleus). They will emit the photons and their total energy will be equal to energy spent to destruction of two hydrogen molecules and one oxygen molecule, i.e. 14.19 eV. Since two water molecules have 12 ring electrons, each of them will emit 14.19/12=1.18 eV=E_b (Fig. 5). It is more than energy (E_b=0.74 eV) of binding of axis electron with the nucleus, and it shows that the ring electrons are situated nearer to the nucleus than the axis ones.

In this case quantity of energy produced due to fusion of two water molecules (14.19+5.98) eV exceeds energy, which was spent for the destruction of two hydrogen molecules (9.06 eV) and one oxygen molecule (5.13 eV). Energy difference of 5.98 eV is divided between two water molecules. It means that 5.98/2=2.99 eV or 285.8 kJ/mole fall per a molecule. It corresponds to the existing experimental data completely [1].

The above-mentioned facts clarify a cause of the explosion, which takes place when hydrogen is combined with oxygen. Simultaneous transition of six ring electrons of each oxygen atom in the nascent water molecules to lower energy levels is accompanied by simultaneous emission of the photons, which generate explosion phenomenon.

Let us pay attention to the fact that two binding energies E_b between valence electrons e2 and 2 and between 1 and e1 are shown in Fig. 5, b. Energy of one electrodynamics bond is equal to E_b= 0.74 eV. If this bond is destroyed by the thermal method, 0.74x2=1.48 eV is required. This energy will be released during further fusion of the water molecule from hydrogen atom H and hydroxyl ion OH^-. In this case, no additional energy is generated.

Therefore the given bond is destroyed by the mechanical method spending 0.74 eV per a bond, each electron will have energy deficit equal to 0.74 eV after bond destruction. This energy will be immediately absorbed from the environment and will be emitted during the repeated fusion of the water molecule from the hydrogen atom H and the hydroxyl ion OH^-. At mechanical destruction of one bond of water molecule, the covalent chemical bond forms E_b= 0.74 eV of additional thermal energy, which is registered in systems of water cavitation constantly (as we have already noted) [1], [2], [3].

It is known that water molecules combine and form clusters. If the bonds between the molecules in the clusters are covalent ones, mechanical destruction of these bonds should be accompanied by a release of additional thermal energy as well [1], [2], [3].

Experimental Part

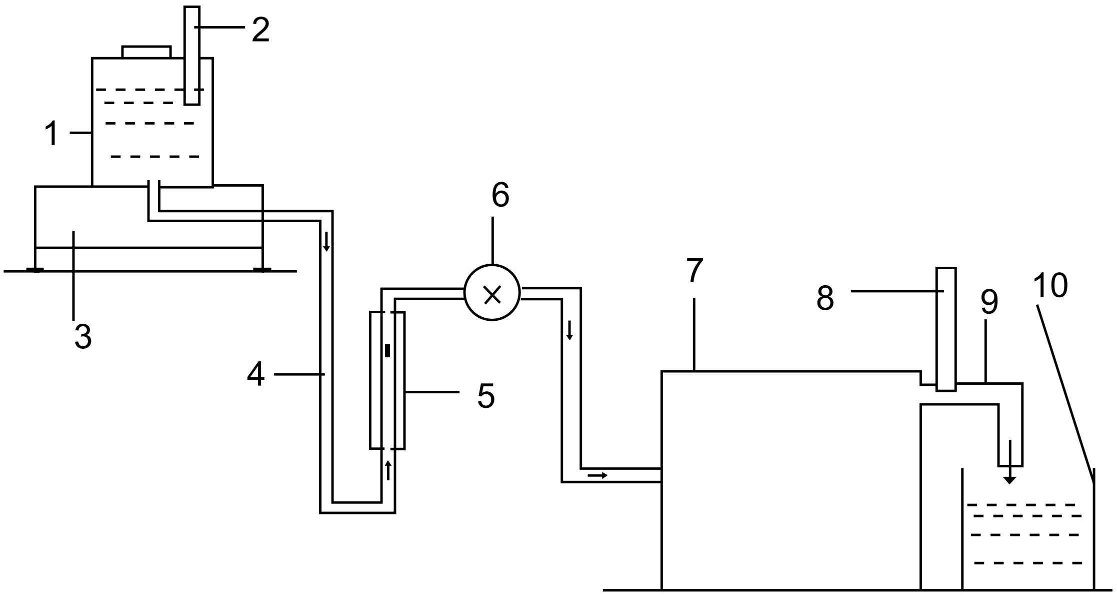

Fig. 6

Diagram of the experimental device: 1 - reservoir for solution; 2 - thermometer; 3 - electronic scales; 4 - solution supply duct; 5 - rotameter; 6 - feed solution regulator; 7 - a special thin plasma reactor is in the process of patenting; 8 - thermometer; 9 - discharge of heated solution; 10 - inlet reservoir

Thus, chemical bonds between the atoms in the molecules and the molecules in the clusters can be destroyed mechanically, by electrodynamic and thermal influence. We have already shown that the mechanical way of destruction of such bonds requires half energy as compared with thermal energy. It appears from this that energy expenses for electrodynamic destruction of these bonds should be less than thermal expenses as well. Electrodynamic impact on the bond gives the opportunity to form the resonance modes where energy expense for the destruction of these bonds is reduced to greater degree. In order to check this hypothesis a special experiment was carried out. It was connected with electrodynamic destruction of chemical bonds of water molecules with a changing frequency of impact. To test this hypothesis the check experiment was prepared and carried out by (besides the author of this article) A.I. Tlishev, G.P. Perekotiy, D.A. Bebko, D.V. Korneev. A diagram of the experimental device is given in Fig. 6. The results of this experiment are given in Table 3.

Table 3

Protocol of control test

Indices	1	2	3	Mean
1 – mass of the solution, which has passed through the reactor m, kg.	2.112	2.153	2.118	2.128
2 – temperature of solution at the input of the reactor t_1, degrees	24	24	24	24
3 – temperature of the solution at the output of the reactor t_2, degrees	33.5	33.5	33.5	33.5
4 – differential temperature of the solution $\Delta t = t_2 - t_1$, degrees	9.5	9.5	9.5	9.5
5 – durability of the experiment $\Delta\tau$, s	300	300	300	300
6 – reading of voltmeter V, B	25.0	25.0	25.0	25.0
7 – reading of ammeter I, A	1.40	1.40	1.40	1.40
8 – electric power consumption according to indices of voltmeter and ammeters, $E_2 = IxVx\Delta\tau$, kJ	10.50	10.50	10.50	10.50
9 – power spent for heating of the solution, $E_3 = 4.19xmx\Delta t$, kJ	84.10	85.70	84.31	84.70
10 – reactor efficiency index according to the reading of voltmeter and ammeter $K = E_3/E_2$	**8.00**	**8.16**	**8.03**	**8.06**

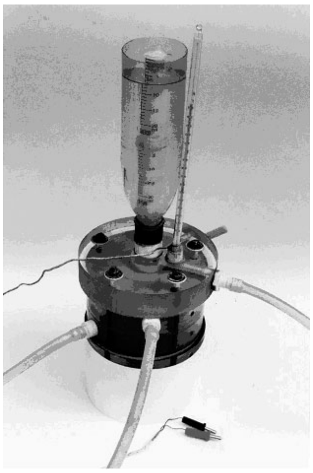

Fig. 7
Photo of heat reactor

In the Russian market three firms (Yusmar, Termovikhr and Noteka) sell cavitation water heating equipment with energy efficiency index of 150%. Soon, an air heating devices with the same efficiency will be produced. The processes of mechanical destruction of covalent bonds of the air gas molecules, molecules and clusters of water and their further fusion serve as a source of additional energy generated by these devices [1], [2], [3].

Conclusion

Analysis of energy balance of the molecules with covalent bonds shows the possibility of additional thermal energy formation during mechanical and electrodynamics destruction of these bonds.

References

1. Ph.M. Kanarev. The Foundation of Physchemistry of Micro World. Krasnodar. 2002. 320 pages. (In Russian).
2. Kanarev Ph.M. The Foundation of Physchemistry of Micro World. The second edition. (In Russian). http://www.ikar.udm.ru/sb28-2.htm
3. Kanarev Ph.M. The Foundation of Physchemistry of Micro World. The second edition. (In English). http://book.physchemistry.innoplaza.net
4. Kanarev Ph. M. Energy Balance of Fusion Processes of Molecules of Oxygen, Hydrogen and Water. (In English) http://Kanarev.innoplaza.net
5. Kanarev Ph. M. Energy Balance of Fusion Processes of Molecules of Oxygen, Hydrogen and Water. (In Russian) http://www.n-t.org/tp/ts/eb.htm.

The Fourth State Of Matter (Plasma Energy) Power Generation

Bruce A. Perreault, USA

http:// www.nuenergy.org
Email: nuenergy@cyberportal.net

Under the proper configurations and in the correct pre glow discharge environment more electrons can be released than what is required to trigger the release. This release of energy attains a high frequency oscillation that is indicative of the metal or metals involved in its release.

The release of high frequency energy from metals and semi metals (semi conductors) through the mechanism in the pre-glow discharge leads to the generation of radiant energy. This correlation has been identified through my findings. The release of high frequency energy from the atoms of electrical conductors is credited to my identification of the mechanism in the pre-glow discharge, or fourth state of matter. This has lead to extremely efficient alternative energy circuits.

The release of high frequency energy from the atoms of electrical conductors is credited to my identification of the mechanism in the pre-glow discharge, or fourth state of matter. This has lead to extremely efficient alternative energy circuits.

The pre-glow discharge condition occurs when a sufficient amount of charge difference is applied across the gap of two electrical conductors. It makes a hissing sound if the electrodes are surrounded by air. The electrical properties between this gap change: electrons become "stripped" from their respective atoms and are liberated. In this state, air is **ionized** and is transformed into a **plasma** and is no longer a gas. It is now in the fourth state of matter, the other three being solid, liquid, and vapor (gas). The fourth state of matter is a good conductor of electricity, its average resistance being much lower than that of the same substance in its gaseous state.

A good readily available spark gap combination is lead and iron. This combination generates plenty of radiant energy and causes noise to be heard on a radio in its vicinity. As a general rule, the better the thermoelectric difference between the dissimilar couple the more radiant energy that will be generated when it is exposed to a plasma field.

Early in the summer of 1981, the pre glow had first caught my attention. This is where I had observed it between two plastic insulated hook-up wires from radio shack. The plasma between the two wires was called into existence because of a high voltage power supply that I had obtained for my electrogravitic experiments. The two wires were twisted together and given a charge of around ninety kilovolts.

It was not until 1989 when I observed a high frequency signal on my oscilloscope coming from the pre-glow discharge phenomenon. These signals were coming from across the room and originated from a crude component that I had built. This crude component consisted of a guitar wire that was centered through a section of PVC tubing. Around this tube was wrapped a coil of insulated transformer wire. This was my first ion-valve. It was excited with the same power supply where I had observed the plasma effect several years previous.

I knew that my ion valve was generating high frequency oscillations using air ions but did not find any practical applications for it until six years later. It was in this year (1995) when I realized the importance of the ion valve and its associated high frequency ion field, known as radiant energy to Moray, it being the essential principle behind his energy receiver.

In one experimental set-up, it was confirmed that when a charged capacitor is discharged through a spark gap that the stored energy is transferred by way of high frequency electrical oscillations. It was found that these oscillations occur directly before the discharge of the capacitor, thus the term, pre-glow discharge. It was found that during the pre-glow discharge plateau, energy is transferred from a primary to a secondary electrical coil. Directly after pre-glow discharge, a surge of current is measured. At the point of current surge is where most of the stored energy in the capacitor is released as wasted heat. It was apparent that by maintaining the pre-glow discharge plateau a very efficient mode of energy transfer and conversion can be obtained. Logically what is required is a control mechanism. This is where my ion valve finds its niche. Upon further investigation, it is found that it keeps pre-glow discharge oscillations at their plateau, and suppresses current surge in a spark gap. It is a self-regulating quenching device. When it is placed in series with an air-core transformer and arc gap, its plasma hisses smoothly, the arc in the spark gap is almost silent and is very small. Continuous oscillations can be seen on an oscilloscope. A light bulb connected to the secondary will glow brightly. This result is not possible without the ion valve included in the circuit. The ion valve has the added feature of preventing back electromotive force, generated from inductive loads, from being discharged back through the circuit. Without the ion valve, energy is normally lost when it is discharged and is transformed into heat. Using the ion valve therefore has the effect of conserving energy.

Fundamental Radiant Energy Device

"Standard" scientific principles are not being violated. There is no fringe science involved with radiant energy power generation. This is a very old source of energy being extracted and harnessed in a unique way through my discoveries. What I have discovered is practically a limitless source of energy. Radiant energy has existed since the beginning of time. What I hold claim to is an improved method to generate and to convert radiant energy, kinetically active ions, cosmic energy, call it what you will, into useful electrical power. Think of the radiant energy power generator as a type of energy detonator that liberates great quantities of energy with only a small exciting spark. No laws of physics are being violated. No new laws are being implied. They are being expanded. The concept is not that much different from how a lighted matchstick is able to start a bonfire.

The power generated from my circuits comes from the transformation of matter into radiant electricity. It does not come from the splitting of atoms. To obtain high wattage from a spontaneously radioactive substance would require unsafe amounts of radioactive material.

Under the right conditions, ordinary matter can be made to generate intense surges of radiant energy that can be heard on a radio receiver as static noise. Build a device that can efficiently capture this energy and convert it into useful electrical currents and you will have yourself a powerful source of electrical power. This device will be powered by artificially disintegrating matter as described by Gustave Le Bon in his book "The Evolution of Matter" and in his book "The Evolution of Forces".

Energy and matter are two distinct entities of the same manifestation. Matter represents a stable condensed form of energy. Heat, light, electricity, etc., are uncondensed vibrations of matter oscillating at differing rates. "Cosmic Energy" is the term that Le Bon used to define matter and energy as being one and the same manifestation. He theorized that when stable matter is disintegrated it is transformed into energy that we recognize as heat, light, electricity, radioactivity, etc.

Semiconductors will generate intense surges of electron oscillations that become powerful sources of radiant energy. A simple **ion valve**, or call it what you will, can be used to generate and convert this form of radiant energy into useful electromagnetic oscillations. A tuned transformer can be used to directly convert these oscillations into to a practical voltage and amperage. This unique valve is shown in Fig. 1. Before it can function it must be connected to a suitable voltage source. The circuit in Fig. 2 demonstrates the full working concept. Other embodiments are also possible. The proof of concept circuit reveals how radiant energy can be generated and converted into useful electrical currents. Many additional stages can be added for more power. Many other circuit and component configurations can also be used but the fundamental conversion principle remains the same. The circuit and components shown should give you a very clear idea of how the technology works. The actual mechanics and electronics of building and validating a radiant energy prototype are relatively simple. Contrary to what some people might think, dangerous levels of radioactivity are never used in my devices.

Ion-valve Converter Technology Explained

The ion-valve converter (ion-valve) shown in Fig.1 has an axial negatively charged tungsten cathode wire that extends the length its cylinder and is capable of emitting secondary electrons. The anode cylinder is positively charged and is made from a semiconductive material that will readily capture electrons.

Within a few milliseconds the accumulated negative ions are attracted to the positively charged onrushing atomic ions. When the negative and positive charges collide they neutralize each other generating high frequency electrical oscillations.

There appears to be a common thread shared between several alternative energy devices. It is the pre-glow discharge. The report on the **Hans Coler** device released by the British Government indicates that there is excess energy released when electrical contacts are opened and closed. The **Lester Hendershot** device utilized a buzzer circuit that opened and closed its electrical contacts. In the **Alfred Hubbard** coil pre-glow discharge flowed through electrical contacts, a distributor cap and radium soaked spark plug. The **Joseph Newman** motor used a sparking commutator. **Thomas Moray** invented a glowing, cold cathode discharge tube that was the heart of his radiant energy device. **Hermann Plauson** was granted U.S. Patent No. 1,540,998 that used spark gaps to convert atmospheric energy. **Frank Wyatt Prentice** was granted Canadian Patent No. 253,765 that detailed his invention, which lighted 50 sixty-watt carbon lamps with an input of only 500 watts. His invention utilized a spark gap driven high frequency tuned resonant system. **Chancy Britten** used ion-valves constructed with a central wire that was surrounded by a coil of wire which is described in his US Patent No. 1,826,727. Britten's valve was said to have lit up his home in the 1930's according to a local newspaper article of that time period. **Alexander Chernetski** experimented with what appears to have been a type of ion-valve that was filled with hydrogen gas. It is said that he got up to five times more energy out of his device than what he put into it. **Edwin Gray** was granted U.S. Patent No. 3,890,548 for his efficient spark gap driven capacitive-discharge motor. He improved on this patent by replacing the spark gaps with a pre-glow discharge switching tube. His U.S. Patents No. 4,595,975 and No. 4,661,747 describes this tube in detail. Gray's patents claim to conserve battery power by sending unused energy back to the supply batteries. On close examination, we also find that his switching tube is actually a quenching device.

Radiant energy was generated during the pre-glow discharge cycle that also contributed to recharging the batteries. **Paulo N. Correa** and **Alexander N. Correa** obtained patents to a pulsed pre-glow discharge system that recovers energy and recharges a battery.

I have found that radiant energy is generated when a plasma field is in contact with the atoms of an electrical conductor. Most importantly, the quantity of radiant energy is greatly increased when the plasma field occurs between the inter-electrode couple of two differing electrical conductors. That, the wattage produced will depend largely on the type of electrode materials that are used. I do not suspect that the above referenced inventors were aware of this amplifying mechanism.

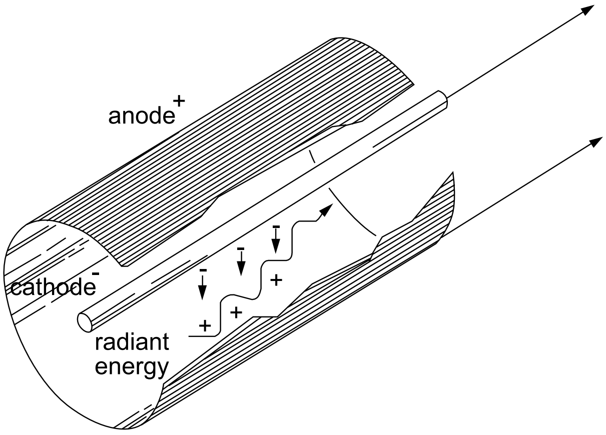

Fig. 1

Ion Valve

In the embodiment in Fig. 1 the negative charge on the wire in the ion-valve negatively ionizes any gas that contacts it. These ions rush towards the positively charged cylinder. When a metal ion that carries an excess electron hits head-on with an elementary differing positively charged onrushing metal ion a violent union of the two metals occurs. This causes the newly formed bimetallic alloy to violently oscillate, break up (dissociate) and to liberate a burst of electrons. The **"sea electron model"** helps to explain this effect. According to the sea electron model metals are bonded to each other through electron sharing. The model suggests that metal atoms be bathed in a sea of valence electrons. If this model is taken one step further it can be seen that when metallic atoms are dissociated from each other excess electrons are released in the form of electrical oscillations of high frequency (radiant energy). This occurs because the electrons no longer take part in the inter-atomic binding force that existed before the dissociation took place. It becomes clear that the freed electrons will add amperage to the output circuit to which it is connected. Henceforth, the equation **I x E = P** holds true in this system. Where, "I" represents the electrons (amperage), "E" electromotive force (ionic voltage), and "P" the power generated.

Obtaining Electrical Energy from the Transformation of Cosmic Energy

Matter is cosmic energy in a condensed state according to Le Bon and Moray. What this means is that matter can be excited its plasma state and caused to rapidly disintegrate transforming itself into electricity.

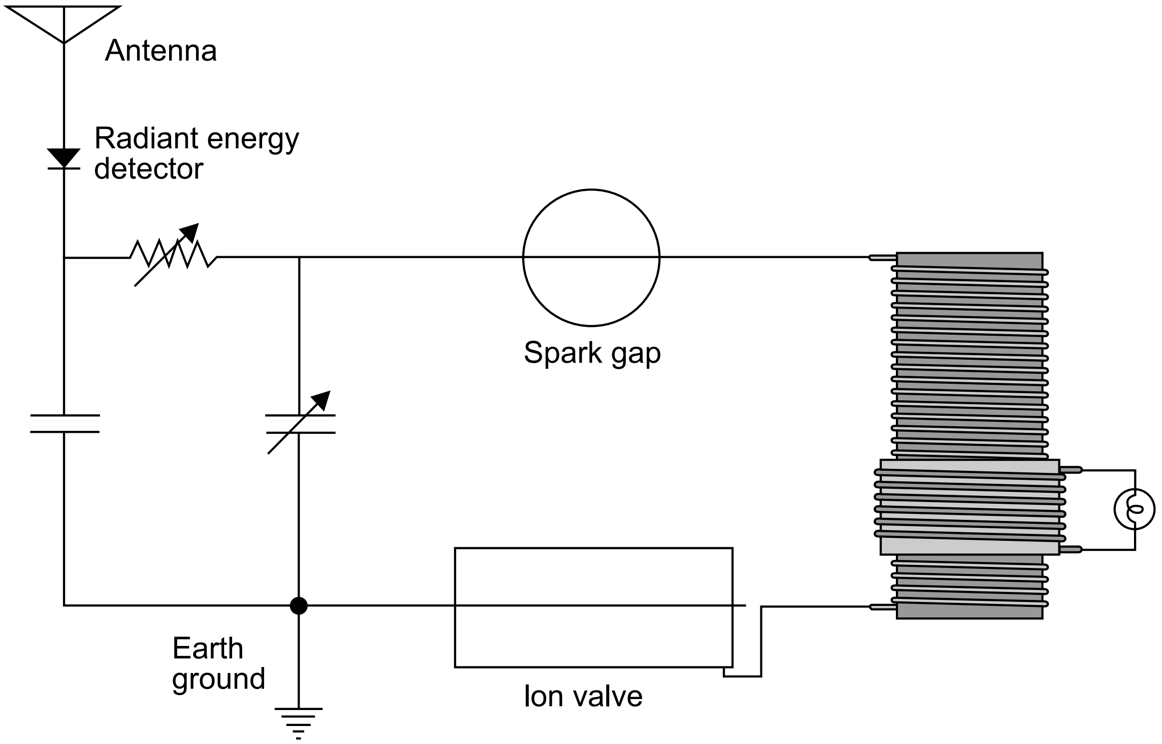

Fig. 2

Proof of Concept

The proof of concept shown will obtain energy from the dissociation of matter bringing the circuit to life. It does this by creating a plasma field between differing electrical conductors. Electrons are unleashed and feedback to the circuit where they are converted into useful power.

Plasma oscillating on the correct wavelength will act like a spark on a mass of explosive material but releasing electrical particles instead of heat. This is to say that the excited fourth state of matter (plasma) causes condensed energy (solid matter) to become uncondensed (radiant energy). The reaction will be far superior to the force that invoked it in the first place because the energy stored in matter is released. The energy that is condensed in the elements of matter is immense. The result is that an enormous amount of energy is released with only a slight loss of matter. Gustave Le Bon proved this when he demonstrated that the action of solar light and from electric arcs on bodies produced electric particles similar to those of uranium. He showed that it caused all bodies to disintegrate to different degrees.

the excited fourth state of matter (plasma) causes condensed energy (solid matter) to become uncondensed (radiant energy). The reaction will be far superior to the force that invoked it in the first place because the energy stored in matter is released

According to the law of conservation, when we give to a material body a determined quantity of energy, this energy might be transformed, but the body will never give back a quantity in excess of what it received. This principle is considered too self-evident to be disputed. It makes sense that matter can only give up energy that is given to it and is unable to create excess energy. Without violating this law, matter can be excited into giving up its stored interatomic energy. Condensed matter can become uncondensed if its oscillations become violent enough and henceforth transforming itself into radiant energy. No laws have been violated. They have been expanded. The first law of thermodynamics describes the principle of the conservation of energy. It states that "energy is not created or destroyed; it merely changes form". The fact is that the creation or destruction of energy is a result of matter being broken down or built up. They both go hand in hand.

Naturally occurring radioisotopes were created through the billions of years of **cosmic ray** bombardment of ordinary matter. Their matrixes became unbalanced. The correct trigger will cause these isotopes to seek their original balanced state. Enormous amounts of electrical energy can be obtained from the transformation of stored cosmic energy. These isotopes contain a very old energy reserve that can be released in a unique way with my discoveries.

Have you ever wondered about spontaneous radioactive decay? Is it really a super charged state of matter? If an atom can become ionized by either gaining or by losing electrons then why can there not be a nuclear ionization too? I hypothesize that atoms do become ionized on the nuclear level by gaining or losing neutrons. This may appear incidental but could be at the very heart of an ultra-chemistry. It is highly probable that this hypothesized phenomenon is behind natural spontaneous radioactive decay and explains many unanswered questions about nuclear science.

Liberated Energy

With a slight excitement from a pre-glow discharge in a spark gap, or even with no excitement at all, as we observe in spontaneously disintegrating radioactive bodies, such as U235, we can obtain large quantities of energy. Clearly, we did not create this liberated energy, since it already exists in matter, but we release it under the right conditions. This is being done without violating the law of energy conservation. The idea that matter could be transformed into energy was absurd before the acknowledgement of nuclear transformations.

A Nu Science is on the horizon. It involves the means of transforming matter into energy without splitting atoms. This science recognizes several isotopes of matter that spontaneously liberate energy as observed in naturally occurring radioisotopes. My research indicates that it is also possible to artificially speed up the natural decay process of condensed energy (matter) using a minute plasma field, as from a pre-glow discharge, etc… With a very small quantity of energy, we will be able to produce a very large quantity of energy without splitting the atoms.

Capture Capacitor

Nature offers us cosmic energy that manifests itself in many different forms. Electricity is only one of its manifestations. From this knowledge, we can obtain electrical power with no moving parts. The natural world contains many storehouses of this cosmic energy. Energy is all around us just waiting to be transformed. The capture capacitor is one such example. It is called a capture capacitor because it captures and transforms electrical charge into electrical current. It does this with the aid of naturally occurring radioactive matter (N.O.R.M.). Put a pair of dissimilar metal electrodes together with a porous ceramic dielectric material sandwiched between them, along with a weak electrolyte and you have built a capture capacitor. In a low power capture component there exists a minute impurity of N.O.R.M. It is always present in clays and is usually present in ceramic materials.

If you want a capture capacitor to generate more power all you have to do is to add additional amounts of radioactive material to the dielectric. Lead–210 is the optimum choice because it has a half-life of around 22.3 years and is a pure beta (electron) emitter. Its half-life is almost twice as long as tritium. This means that it is feasible to build a component that will put out power for several years with little maintenance. Lead-210 is a decay product of radon gas. The parent source of radon is uranium. Therefore, uranium ore can be powdered and mixed into the ceramic material.

The atomic ions emitted from naturally occurring or artificially induced radioactive transformations can be directly converted to electrical power. The circuit shown can be used to convert radiant energy into useful electrical currents. Additional stages can be added for more power. The circuit shown gives a general idea of how such an energy conversion device functions. By no means is this technology limited to this one circuit, configuration, or source of radiant energy.

EAGLE-RESEARCH
Energy Solutions
since 1984

A Research Organization that
Develops & Distributes
Practical Energy-Saving
Methods & Devices

4 Energy Way, P.O. Box 118
Porthill, Idaho 83853
FAX: 250/492-7480

Technical questions to:
wiseman@eagle-research.com

NEWS REVIEW

◆ A new type of high field superconducting magnet from Oxford Instruments has enabled physicists at the University of Nottingham to balance the force of gravity and levitate heavy and dense materials such as lead, gold and diamonds. The patented Minimum Condensed Volume™ (MCV™) Cryofree™ superconducting magnet, developed and manufactured by Oxford Instruments in collaboration with the University of Nottingham, is the first of its kind in the world. Unlike conventional superconducting magnets, the MCV magnet never needs to be refilled with liquid helium and yet can generate a field over 300,000 times stronger than the Earth's magnetic field. MCV has a very small footprint and, operating at 4.2 K with a room temperature bore, can generate highly stable magnetic fields up to 16 Tesla.

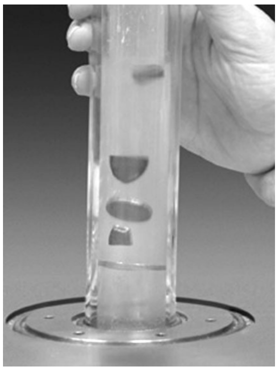

http://www.oxford-instruments.co.uk

◆ Polish inventor Zygmunt Orlowski called attention of the scientific society to his perpetual motion machine **METOZ**, which according to the author after further improvement will be able to extract energy from the earth's gravitation.

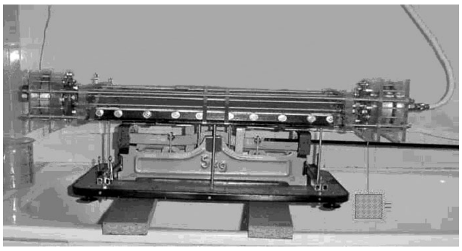

http://www.nets.pl/~metozor/perpetual_motion_machine.htm

◆ It is necessary to change behavior of people in order to decrease environmental damage. This was the message of **"Green Week 2003"** which took place since June 2 through June 5 in Brussels.

The conference's key issues were three aspects of United Europe environmental policy: sustainable production and consumption, renewable energy and water supply.

Green Week is the first environmental forum of such a scale. Three thousand participants from all over the world represented governments and funds, industry and mass media. The exhibition organized in the framework of Green Week included 62 displays from 14 European countries and were dedicated to the newest environmental technologies and projects. During the Green Week 26 conferences and seminars took place, as well as a great number of press conferences and actions for young people.

◆ **British researchers (**Gardner Watts company) **believe that they have made a groundbreaking scientific discovery after apparently managing to "create" energy from hydrogen atoms.**

The device represents "thermal energy cell" which appears to produce hundreds of times more energy than that put into it. If the findings are correct and can be reproduced on a commercial scale, the thermal energy cell could become a feature of every home, heating water for a fraction of the cost and cutting fuel bills by at least 90 per cent.

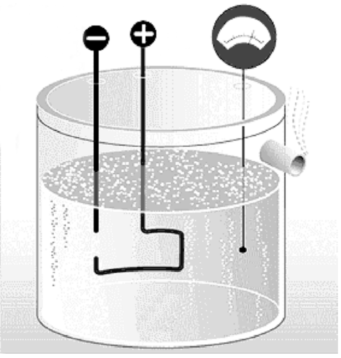

http://www.telegraph.co.uk

◆ **Gravity Conversion Rotary Device**

The below figure is from http://www.theverylastpageoftheinternet.com/newclaims/GravityMotor/gravity_motor.htm

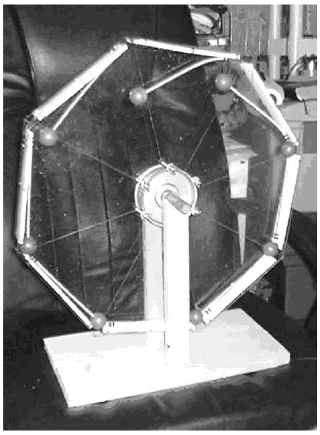

Does it really work?

THE MARCUS DEVICE CONTROVERSY

Tim Ventura, USA

Email: tventura6@attbi.com
http://www.americanantigravity.com
http://www.americanantigravity.com/marcus.html

Editorial: This overview is devoted to the fantastical device created by inventor Marcus Hollingshead that encountered more than its share of publicity and controversy between November 2002 and March 2003. You can participate in discussion on aggroup@yahoogroups.com.

The problem with Marcus has been that he was pressured by too many people to produce too fast results. I started talking to him in November, and he is always sounded like an honest, rational man with a good heart. The problem is that once people started to hear his private claims that he could lift something like 2 tons of weight, things kind of went south.

While people in general have been very well-meaning, they are desperate to solve many of our current energy-problems, and when Marcus appeared with a potential solution it got out of control. Pretty soon film crews were showing up at his house unannounced, and he was getting unwanted phone calls from people who had looked up his number online.

In my opinion, it is really a problem of a "social dynamic" at work here – too many people, with too much access, all at once. This also correlates to the type of person that different people are. For instance, I am a very open person, and I really do not value my privacy a heck of a lot — therefore, when TV crews drop by with 5 minutes notice (and they have), I tend to just let them in for a demo. However, Marcus, who is much more of a private person, seems to have become a bit overwhelmed at this proposition.

Also, keep in mind that I had initiated working with the media because I felt that it was time for the excellent work in the Lifters groups to see the light of day. There was some debate about this, but the Lifters group as a whole has greatly benefited from being about to tell people about this new technology.

Marcus had not actually "gone public" about his work. He had leaked information about his research to a few people, and from there it was leaked to several others, who in turn told their friends about it. Unlike the Lifter technology, which has hundreds of people that can talk about how it works and where it came from, Marcus technology was represented solely by himself. That means that everything surrounding his device came back ultimately to him alone. Without any real group ownership of the expertise behind the device

As I said before, Marcus was new to the publicity, and did not have time to prepare the resources that the rest of us have had to answer questions, provide demo materials, etc. Naudin and Saviour have a masterful manner with regard to teaching people about Lifter technology, but they have had years to prepare this information and perfect a method of presenting it to people.

Most of us in the AG groups tend to proselytize for the technologies that we believe in. I know that I do, and if you haunt the newsgroups much you will notice that everybody there believes in something and can go on for months non-stop telling the world about it (again, I do this). Marcus, however, does not really have this personality — he is more of a "take it or leave it" type of person.

For me, this was another indicator that he was the "real deal". You have to understand that I have never seen working prototype pictures of the Marcus device, and all that I had to go on was the feeling that I got from interacting with him on the phone and online. The fact that he did not try to "sell me" on the idea was a big plus. He told me about the device, and his story was always consistent. The other people that he talked to also had a consistent story – I have not talked to anybody and heard a different version than I heard from Marcus himself. That is a big plus from the credibility perspective.

Marcus initial goal (from when I started talking to him in November) was to gradually ease into working with the public to tell them about the technologies he was working on. Unfortunately, I probably aggravated the issue a great deal, because his early descriptions of the technology might have gone unnoticed if it had not been for some of the things he had mentioned about his research.

When I read Marcus' firsts posts on the device, my very first thought about his claims was that they were bogus. Judging from the reaction of the Antigravity newsgroup, I gathered that a few other people also had the same initial reaction. The reason might be a little selfish – after you hear claims but do not see results for a long enough period of time, you tend to protect yourself emotionally by becoming skeptical to the idea that somebody can be lifting hundreds of pounds of weight using a completely new and revolutionary technology.

Nonetheless, something that Marcus had said resonated with me. I remember that he was talking about his results being based on a study of the Earth's fields, which is something that I had read about myself during the college years. Telsa had done a lot of work involving the Earth's electrical and magnetic-fields, and I had always felt that this research might come in handy later as a potential basis for Antigravity research. However, I did not have any specific ideas, and began asking Marcus questions offline to find out more about his research.

In reality, I am nobody special, so I think that perhaps this would have happened in any event no matter what my reaction would have been. Certainly I am not the only person that took an early interest in his work, and from what I learned later he had already been a little bit too honest in the BBC online newsgroup before he ever became involved with the Antigravity newsgroups.

From January to mid-February Marcus became like somewhat of an addiction in the newsgroups. This was a strange phenomenon to watch, but if he did not answer questions in the newsgroups for even a day people started asking whether he had been kidnapped or the government had assassinated him.

You have to understand that most inventors tend to segregate their time into "public" and "private". I do, Naudin does, and everybody does this. When you talk to the public about what you have built, you really get into it — but when you are actually working on it, you tend to hole up in the garage and put the entire world on "ignore" until you have finished with whatever it is that you are doing.

Marcus really started to get publicity in late November 2002, and it built up over a few week period of time. By mid-December, he had a healthy following in the Antigravity newsgroup and people in the 5 or 10 other AG newsgroups were developing more than just a passing interest. Great stuff for Marcus – he had thought that he was the only person doing this research, and it turned out that lots of people shared this common interest with him. We had lots of really valuable and productive exchanges.

Marcus bided his time working on the projects – he had initially stated that he could lift around 200 pounds of weight with his November version (prototype #161, if I remember correctly). He did not have a digital camera, and he did not want to publish photos until he had a better prototype anyways. That was OK with me – I am used to work with a variety of inventors and came to realize a long time ago that everybody has their own style. I was content to wait until Marcus had additional data, and the only thing that I insisted upon was that sooner or later he ante-up and provide some photographic evidence.

Despite lacking photos or video evidence, Marcus did have documentation. In fact, he wrote more documentation about his device than I had ever hoped to receive. He cranked out about the equivalent of 30-pages of "real" documentation in a very detailed form within 1-month, and had the email equivalent of many more pages in postings about how the device worked online. Marcus had already been publishing details online for about 2.5 months.

Despite the incredible amount of published data that Marcus was able to complete, in some ways it made things worse for him than not having anything at all. The reason was the nature of his research. My work, currently focused on Lifters, was in a technology that is easy to explain – the device is built like this, it works like that, and it can be described in a manner similar to some other thing ....basic, easy to explain, and easy to compare.

After all, Lifters have been accused by science of being ion-wind devices. If nothing else, rebutting this claim gives us someplace to start – something to compare against to provide people with an idea about how the device works. Additionally, we had tons of video and photographic evidence to back up our claims. In Marcus case, the documentation that he provided made things worse because his technology is so utterly different. It is not ion-wind, nor the Biefeld-Brown effect, nor apparently the Searle effect. In fact, the more than you tried to pin down exactly what it was, the harder it became to describe it.

Also, its not like he could simply fall back onto describing the construction of the device **– reason being that it contained a series of interrelated coils with a unique configuration that were intrinsically tied to the operation of the device.** Every time he wrote a detailed summary of the construction of the device, we had come away with more questions than we had started out with. After a time I gave up trying to understand how it was constructed, because I only had bits and pieces.

Marcus was using an arrangement of 6-coils, and what are described as "bifilar windings". Essentially, this means that you wind two wires together around a coil and then tie one of the ends to another, making the two wires essentially one long wire wound "against itself" down the length of the armature its on. There are 6 of these armatures on the device, mounted around a central RP (or reference point).

Marcus created all of the terminology for these parts – this lends additional credibility to the idea that he had had 7 years of experience with it. These are ideas that you just do not "make up" for publicity. For instance, the bifilar windings on the armature-coils are self-cancelling fields, and even a basic-electronics education will tell you that. However, the interesting thing is that once you really start to get into the advanced "post-scientific" research on magnetic fields you start to see self-cancelling magnetic coils all over the place. The reason is apparently that while the "B-field" *(Editor: i.e. magnetic field)* of the coils cancels, the "A-field" *(Editor: i.e. vector potential)* that Tom Bearden keeps talking about does not, which means that you can isolate the really interesting effects from the coil without having to deal with the high field-intensity effects from normal electromagnetism.

Marcus had described these six armatures as being activated in a series, or order, to obtain results. Each of the coils was spun by an electric motor mounted on the armature, and when the coil was spun up to a certain speed and "stubs" mounted a periodic intervals on the side of the armature were fired in sequence the antigravity effect was created.

Marcus Antigravity effect was a directional force that was applied in a unique manner depending on which coils were activated. This means that by activating (for instance) coils 1 and 3, he could create an antigravity effect and move the device left – or something to that effect. Turning on and off each set of coils gave him movement on one axis of thrust, which corresponded to X, Y, or Z in a standard 3-D geometric graph. Since you can move in either direction down any given axis of thrust, this means that you have 6 axis of movement height (up and down), width (left or right), and depth (forward or backward). His device provided movement on any axis by either one coil or an arrangement of coils.

For a while Marcus did not want to talk about side effects, but I pressured him into it. The reason was that I have seen side-effects that were pretty darned strange in a number of different experiments, and I was willing to bet that if I pressured him a bit he would talk about what he had seen but did not want to reveal (for fear that people would not take him seriously). It turns out that I was right.....

Marcus had talked about the RP, or "Reference Point" upon several occassions. This is a multi-layered device that acts somewhat like a capactive element. It sits in the direct center of the Marcus device, and it is the reference-point for the entire device. The RP is manufactured from cast-iron, and Marcus says that he has a local company build them for a few bucks each by pouring iron into a mold based on one of his designs.

You can active the RP at the same time that you activate the coils. You have coils rotating around the RP, with "stubs" on the coils firing periodically at points corresponding to spots on the surface of the RP. Meanwhile, you also have an electrical charge on the RP itself, which means that there are a lot of elements interacting at once to create a very complex dynamic.

One of the interesting side effects was a darkening and "blurring" effect of the RP, as if light was being reflected off it. Another was a Star-Trek like "force shield" around the RP. These only occurred when all of the coils were activated at once. Also note that since the coils are opposing in nature (one for each direction on each axis), that when you are creating the force field effect the device cannot be levitating. However, the device is operational, and you have a force-field that you can apparently bounce a heavy hammer off without being able to penetrate it (Marcus description).

Another interesting note is that the force-field effect can be modified to create a vacuum. I am not sure how

Marcus noticed this, but he had said that it appeared to create a vacuum inside of the field's boundaries during operation. Interestingly, it also ran very cold – apparently down to a hundred degrees below zero, but only within the localized boundaries of the field.

I am not a professional physicist, but this is something that has been reported with the Searl effect and several other experiments, and the reason that I believe it occurs is because you are taking the kinetic energy of the device as well as some of the kinetic energy of its molecules (i.e.: heat energy), and injecting it into another dimension. After all, anytime we directly modify a gravitational field we are creating a dimensional effect, but in normal life we do not notice it because the boundaries between gravitational fields are very gradual.

Well, in the Marcus device the boundary between gravitational fields is not a several thousand mile-long gradient like the Earth's field is – instead of you have a gradient perhaps 10-times more intense focused within a 3 to 4 inch area. That means that you are creating essentially a rift in time and space (self-healing) that is kept open by the energy that you are pumping into the device.

the energy input is used only in putting the device into another frame of reference, but that any interaction between us and our "normal" frame would have to be modified for any interactions with the "modified" frame.

To the best of my knowledge, this is the most accurate description that I have seen yet for why his device does this. Marcus himself could not explain why these effects occurred, and to be honest I scared him more than a little when I told him my theory.

In the hypothesis that I just put forth on the origin and functionality of the Marcus device effect, the interesting this is that there really is not a direct correlation between input energy and the effect itself. That is because you are not using the input energy to actually cause an effect – you are essentially using the input energy to translate the device into another dimension (although it is still partially within ours).

Einstein called these "frames of reference". The idea is that the energy input is used only in putting the device into another frame of reference, but that any interaction between us and our "normal" frame would have to be modified for any interactions with the "modified" frame.

Again, this takes place in normal physics – but usually it involves motion, and normally it takes place over a vastly larger gradient of time/space/whatever. The difference between two frames is the difference between two locomotives moving at different speeds – the energy input goes into accelerating each locomotive up to that speed, but the actual work is performed if those two trains interact is solely a product of the frame of inertial reference (mass & velocity = energy).

Another excellent example is a helium-balloon. You put the energy into condensing the helium and pumping it into the balloon, but it is not the helium that creates lift – it is the surrounding atmosphere. In a very real sense a helium balloon gets its energy from the surrounding environment. The energy that you put into "maintaining flight" in a helium balloon is only the strain on the fabric used in holding the helium into the confines of the bag that encloses it. With the Marcus device, I would bet money that it is not the input energy creating these effects – it is instead the difference in frames of dimensional/inertial reference between "our" environment and the "device's" environment, whatever that difference may be.

You see, even though this sounds like a load of BS, magnetism is not real. This is not an idea that I am making up – in fact, it is an entire chapter in the physics 101 textbook kicking around downstairs in my house. Take two magnetic fields, and project them in the same direction at the same speed. Field A can not interact with Field B, because neither of the fields really exists – they are merely ripples on a pond, and that pond is the background of the time-space continuum. People that believe in relativistic gravitational effects take for granted that things like this happen – after all, gravity is not a force per se in relativity – only a modification of the dimensions of time-space. Well, most people do not realize that Einstein based his theory of gravitational force on his study of magnetism. Magnetism is the same thing – it is a modification of the fabric of time-space.

Take two magnetic fields that did not interact from a moment ago, and now change the direction that one of them is traveling in. Suddenly, the two fields that could not even see each other a moment ago now create an incredible "torque" – this is the same thing that we see in bar-magnets and electric motors. This is the "torsion field" research.

Bar magnets are different than pure fields because they are composed of the discrete fields of thousands of tiny magnetic "domains". That is why the fields from bar-magnets are never "invisible" to each other – it is because that the fields in bar magnets are too jumbled to ever really "line up" to the point of being invisible. Pure fields can do this, however, reinforcing the supremacy of Einstein's relativistic effects with relation to electromagnetism.

Anyhow, that is about it for now – I will try to come up with a more detailed picture for you on this if you like, but I expect that you will probably have more questions about Marcus theory and device in near future. It is a very complex idea underlying his research, but the basis of it is the idea of "rotating magnetic fields".

If you want to do additional research on your own into related fields of study, I would definitely look up "relativity and magnetism" – it provides an excellent example of how to visualize relativity theory in the study of magnetic field interactions. This is something that electronics engineers are not taught, which is why you have never heard of this in an engineering sense before. However, in physics is a well-known concept, although most physicists prefer to work with quantum mechanics rather than relativity these days.

In terms of rotating magnetic fields, you can look up "torsion fields" – again, the Russians have the best understanding of advanced torsion field physics, although their best material probably has not been translated yet. Additionally, rotating magnetic fields are the basis for the Searle Effect and the Hamel device.

Marcus was intrigued by Searle's claims, because they appear to have matched many of his own results. Interestingly, Marcus had developed his work within an "intellectual vacuum" and had never heard of Searle before I told him about the similarities. Nevertheless, force-fields, ice-cold operating temperatures, and antigravity seem to be a pervasive similarity between all manners of experiments that have been reported using rotating magnetic fields. I had assumed that Searle's claims were bunk, but after hearing about similar effects from Marcus research it led to reconsider some of the criticism surrounding Searle lately.

One a final note – one group that may lend additional credibility to this research is Godin and Roshchin in Russia. They obviously benefit from the years of knowledge that the Russians accrued in magnetic field systems theory, and they attempted a replication of the Searle effect device some years back (and claimed to get some results). I hear that they have been working on a revised version of their experimental setup recently and hope to have even better results in the very near future.

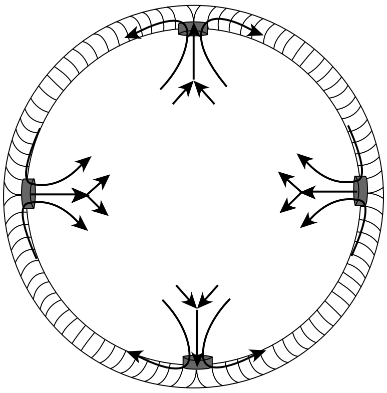

Fig. 1

Magnetic flow for a bifilar-wound toroid that is split at four points on the toroid. This configuration occurs when the coils in the nubs are charged to match the charge on the toroidal coil. The configuration is a quadra pole with a minimum magnetic field at the center

Fig. 2

Marcus recommends the Kikusui 6000L and 18000L multipurpose AC power-supplies

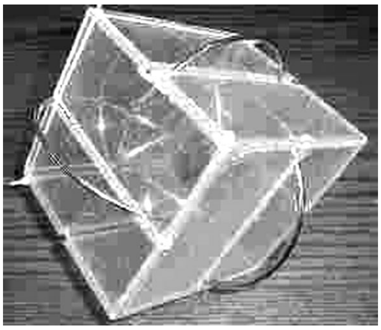

Fig.3

A scale-model mockup of the ring-arrangement within the Marcus device, showing the interlinking coils

History pages

Systems of conversion of thermal energy to mechanical one

Review prepared by correspondent Alla Pashova, **Russia**

There is a wide class of closed-cycle heat engines using differential temperature of water and ambient air. This kind of systems enables thermal energy to be directly converted to mechanical one so as to perform a useful work. We publish patent specifications of two analogous devices of this kind.

PATENT FOR INVENTION

Specification

of the engine, which uses differential temperature of water and ambient air.

To the patent by **P.I. Dobrokhotov** filled on October 22, 1928 (#34248).

Patent grant is published on March 31, 1930.
Patent is valid during 15 years from March 31, 1930.

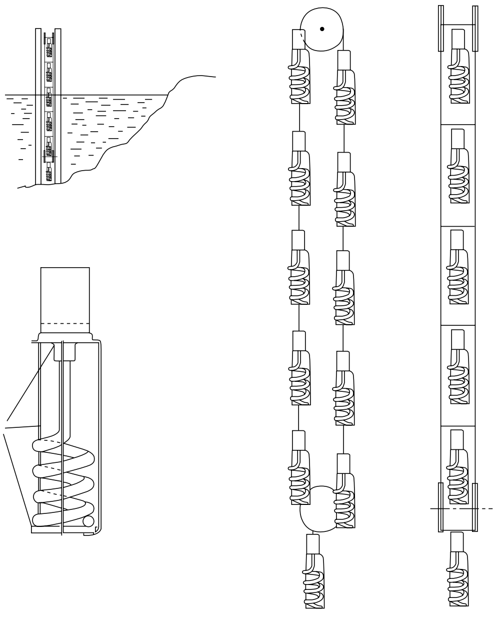

In 1928 **P.I. Dobrokhotov** proposed a design of such an engine. It represents an endless chain thrown over a pair of blocks and partly immersed into water. The chain consists of coiled pipes filled with a volatile liquid (such as sulfurous anhydride or ammonia). One end of each coil is fastened to a frame, while another end is fastened to a cylinder piston mounted on the same frame. When overwater elements are heated by means of ambient air, the liquid inside the coil evaporate. Vapor put pressure upon the pistons which being in motion compress air in the cylinder, and thereby reduce volume of each element. When submerging, the element gradually cools down, thereupon causing a decrease in pressure of vapor of the liquid, which is enclosed within coils. The pistons move in backward direction, and the volume of air available inside the cylinder increases. Increase of volume results in increase of water buoyancy force that sets the device in motion.

In 1980, **Y.S. Grigoriev** developed his forerunner's idea by improving the system described above. The inventor aspired to enhance the effectiveness of this method of conversion of heat energy to mechanical one. He suggested fixing the volume of chambers filled with a heat-sensitive actuating medium at its minimal magnitude before immersing of those chambers into warm water. In water the actuating medium heats and partially evaporates. At a maximum depth of submersion, the chambers are released from fixation to be expanded during emersion. Thus one may avoid expansion of the chambers during their submersion. Futhermore, there is a possibility to activate the device and get the useful work without preliminary spinup from an external drive.

INVENTION SU1404681 A1

23.06.88
Yu.S. Grigoryev
Prototype: USSR patent #13534
F 03 G 7/06, 1928

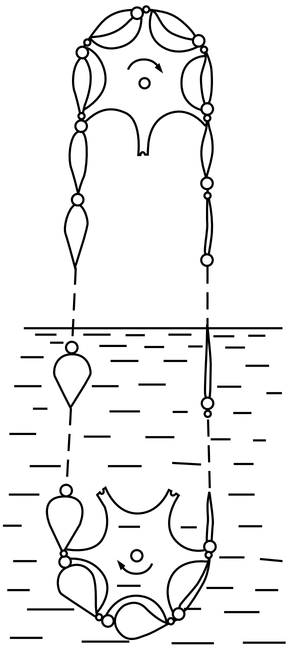

Iliya R. Prigozhin

(25 January, 1917 – 28 May, 2003)

Iliya R. Prigozhin, aged 86, Nobel Prize Winner, Belgian scientist of the Russian origin, died on 28 May in Brussels. Prigozhin made a number of prominent discoveries in the field of thermodynamics and statistical mechanics of nonequilibrium processes. In particular, he developed a concept of irreversibility as applied to thermodynamics. Prigozhin was studying time as a physical phenomenon. In 1977, he was awarded the Nobel Prize for chemistry "for works on thermodynamics of nonequilibrium processes".

The scientist formulated one of the major theorems of the theory of nonequilibrium processes. This theory was later named after him. According to Prigozhin's theorem, stationary state of the system corresponds to minimal entropy generation. The outstanding physicist was also an initiator of applying methods of theory of nonequilibrium processes in biology.

Thermodynamics principles were enunciated in the middle of the XIX century after the invention of a steam-engine, when interaction of heat, electrical, and mechanical work aroused a higher interest. In accordance with the first law of thermodynamics representing an energy conservation principle, energy neither disappears nor appears in any closed system, but transforms from one form to another.

The Prigozhin's theorem sounds as follows: stationary state of the system (under conditions impeding attainment of equilibrium state) corresponds to the minimal generation of entropies. If there are no such impediments, then production of entropy reaches its bare minimum, i.e. zero. The theorem was proved by Prigozhin in 1947.
In thermodynamics the nonequilibrium specifically open systems were of most interest for Prigozhin. In such systems either or both (matter and energy) interchange in reactions with an environment. At that, quantity of matter and/or energy increases or decreases in the course of time.

It is obvious that human society as well as the biological environment is an example of dissipative and nondissipative structures. In the 60s and 70s, Prigozhin developed his theory of dissipative structures and described the formation and development of embryos. Critical points of bifurcation in his mathematical model are correlated with a point, where a biological system becomes consecutive and stabilized.

For his scientific career time, the physicist was awarded over 40 scientific rewards and admitted as an honorary member into national academies of many countries all over the world.

New Energy Technologies

PRINT EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Color - $800

Half Page Ad
B/W - $200
Color - $400

Quarter Page Ad
B/W -$100
Color- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Faraday Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Russian Academy of Sciences
Russian Physical Society
Nuclear Society of Russia
Mendeleev Chemical Society of Russia
Moscow Lomonosov State University
Russian People's Friendship State University

The 11th Russian Conference on Cold Nuclear Transmutation of Chemical Elements and Ball-Lightning (RCCNT&BL-11) is to be held during September 28- October 5, 2003. The place of the Conference is in Dagomys near the city of Sochi.

The program of the Conference includes the following subjects:
- Experimental research in cold nuclear transmutation (fusion) and ball lightning;
- Theoretical models with respect to cold nuclear transmutation (fusion) and ball-lightning effects;
- Applied to these problems technologies and devices.

The Organizing Committee of the Conference is pleased to invite you to attend the Conference. The terms of your participation are as follows:

The full cast is $900, which will include the registration fee, hotel reservation and living, three daily buffet meals, conference proceedings, transportation from the Sochi airport and back, social dinner and special excursion or entertainment.

The languages of the Conference are Russian and English.

The total cost can be reduced to $800 if transferred before August 10 to the account of the Organizing Committee, which is to be announced on having received from you a confirmation about your decision to attend the Conference.

If you make a decision to take part in the Conference, please, let us know before July 1 about the title and abstract of your report.

Contact telephone: (7) (095) 196-9476 (ask Mr. Igor Goryachev), Fax: (7) (095) 196-6108
Email: gnedenko@kiae.ru; Bazhutov@erzion.madi.ru; bazhutov@izmiran.rssi.ru; bychvl@orc.ru
Address: I. Goryachev, Russia, 123182 Moscow, 1 Kurchatov Sq.,
Research Institute of High Technologies at RC "Kurchatov Institute"

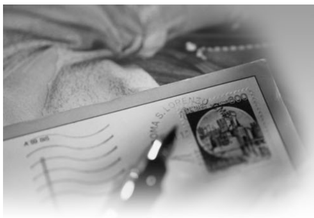

Letters from our readers

Thank you for publication of my article. Your magazine becomes more and more popular.

My special thanks to Editor-in-Chief Alexander V. Frolov. Whatever they say or write in the press, he has done a great work and it is not only my own opinion. He has managed to do that everybody had been waiting for a long time. My colleagues (not only from Rostov-on-Don State University but also from other Rostov Institutes of higher education) ask me personally send the best regards to the editorial staff of your magazine.

Sergey A. Gerasimov

Free energy?

The scientific community says that it cannot exist. Where did we hear that before? Academician Valery Sobolev from the Academy of Sciences, the head of the research group working on the experiment says: "This is not some new physics, but just experiments that have been performed very seriously. Make surge of electrons from the environment, remove them, deplete the fused electrolyte with basic typical metals; what you derive from it is what cannot be produced by chemical methods. This is a new state of substance." As the researchers say, first they discovered a particular electrochemical process where high-temperature materials in a new state are the product. This in its turn allowed to discover a new state of substance, new class of materials, a new source of energy, a new method of cold plasma generation, a new superconductor. See whole article at: http://english.pravda.ru/science/19/94/379/9675_energy.html

Josef Hasslberger

Personal home page on physics, energy technology, social and economic issues: http://www.hasslberger.com

Interference disc generator with permanent magnets

inventor Alan Francoeur,
Canada

Carl Reich

Bud Johnson

Vortex heat pump

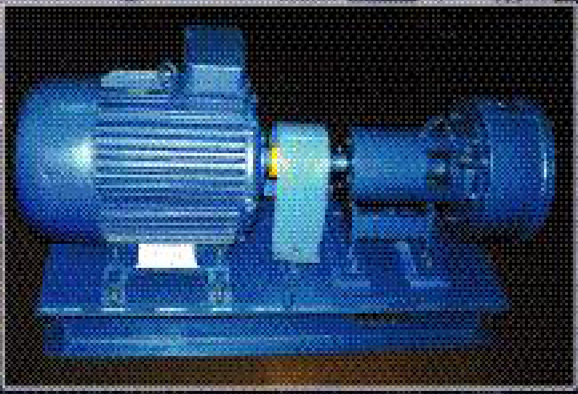

Efficiency
200% and more!

Inventor: Vladimir Susorov
(812) 342-2652
8-901-976-5151
susorow@rol.ru

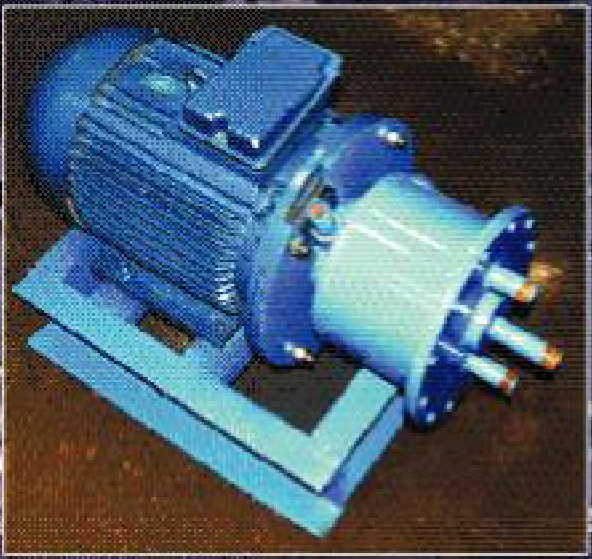

ISSN 1684-7288
9 771684 728801

Gravity-heat power system

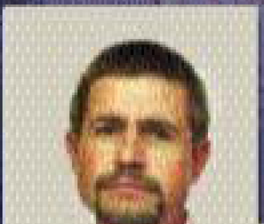

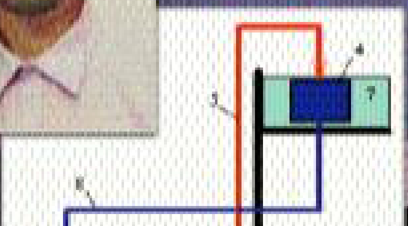

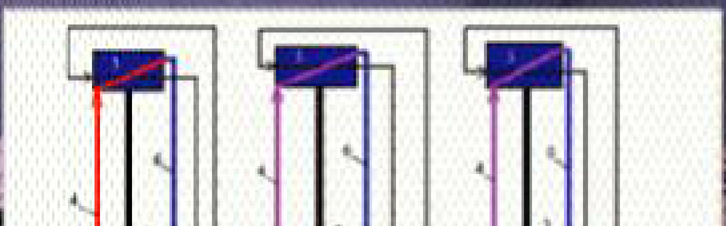

Author: Vladimir Sukhanov Wladimir.Sukhanov@soros.ksu.ru

New Energy Technologies

ISSN 1684 - 7288

Published by Faraday Lab. *Special Christmas Issue #5-6, 2003*
Research on advanced aerospace systems and new energy sources

In this issue:

- *atomic hydrogen in power engineering*
- *magnetic generators*
- *capillary perpetuum mobile*
- *Hutchison effect*

*Photos from Faraday Laboratory:
test bench for measuring the effectiveness of
hydrogen dissociation and recombination processes*

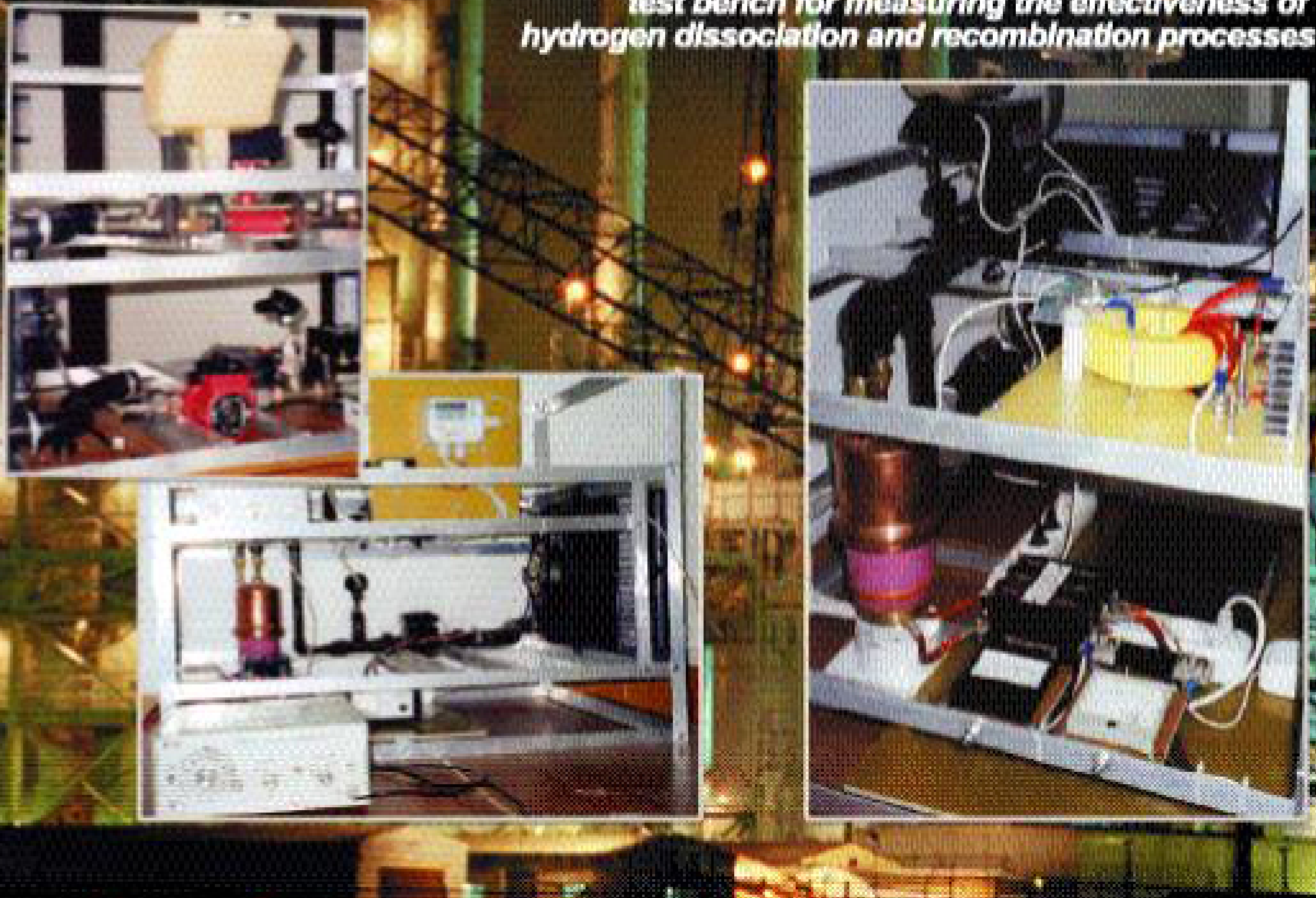

Bogomolov's Invention

Over-unity from elastic forces of spring, compressed by rotation.

Read in this issue!

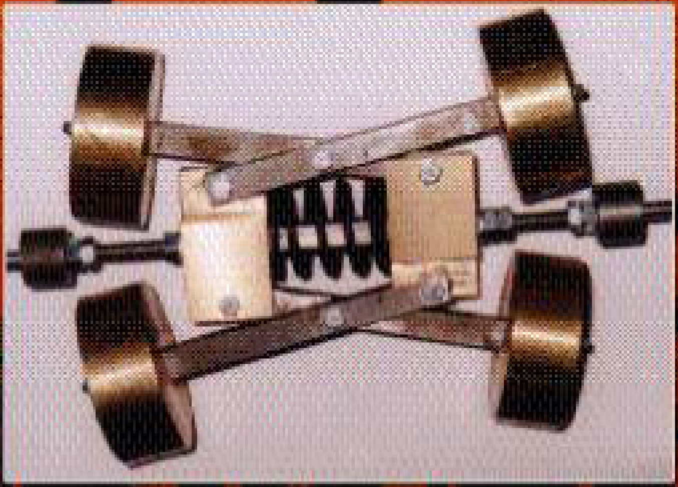

Spin Wave Laser of Ines Espinoza, USA

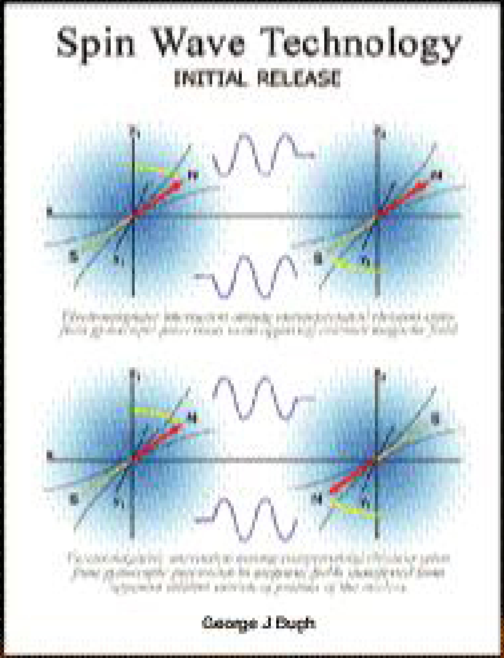

Spin Wave Technology

INITIAL RELEASE

George J Bugh

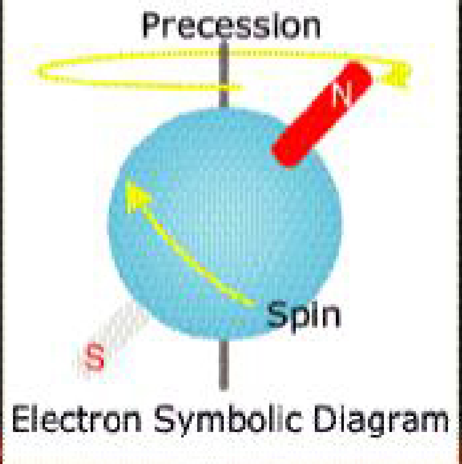
Precession
N
Spin
S
Electron Symbolic Diagram

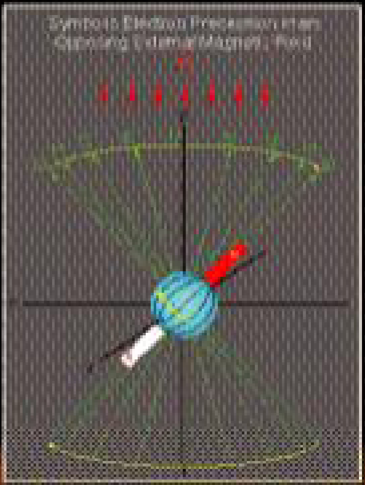

NEW ENERGY TECHNOLOGIES #14-15

1. News from our company
2. Novozhilov's motor, Y. Novozhilov, Russia
3. Self-action in the case of magnetic rotation, S. Gerasimov and others, Russia
4. Capillary motor, I. Elshansky, Russia
5. Gravity-heat energy system, V. Sukhanov, Russia
6. Modern perpetual mobile, Review
7. Free energy and antigravitation, (spin wave lasers), Ines Espinosa, USA
8. Mechanical analogies, Y. Ivanko, Ukraina
9. Cell of fine-plasma electrolyses, Ph. Kanarev, Russia
10. Smith Generator, Review.
11. Strategy of Over-unity Devices Marketing, R. Wood, USA
12. Hutchison effect, Mark Solis, USA
13. Heat energy conversion by means of non-linear dielectric, A Ries and others, Brasil
14. Gates motor, A. Akau, USA
15. "Cold electricity", A.Akau, USA
16. Gravity conversion rotary device
17. Torsion technologies, China
18. News
19. Antigravitation, B.Smith, USA
20. Electric levitation, MacLeod, Canada
21. Perpetual mobile, V. Likhatchev, Russia
22. Interference Generator, A.Francouer, Canada
23. New Paradigma of Time, D. Reed, USA
24. On the rotating permanent magnet system, A.Frolov, Russia
25. Inventor's week, USA
26. Patents on inertial drive topic
27. Bogomolov's experiment with spring drive, Russia
28. History of perpetual mobile
29. Tesla Energy Science Conference, USA
30. Perendev company, Review
31. Internet pages about free energy
32. Matveev's electrical generator, Russia
33. Perpetuum motor with magnets, France
34. The most interesting of 2003
35. Hydrogen power engineering, Review
36. News on hydrogen fuel cells
37. Electric discharge as part of the over-unity device, Review
38. Subscription page

Faraday Lab News

One of the most interesting research projects of our company in the year 2003 has been the research on hydrogen dissociation and recombination processes. It has taken us about half a year to build a special vacuum tube and in October, 2003 we obtained some experimental results on the heat output and efficiency of the system. The system we designed in our laboratory is shown in Fig.1 and also on the cover. The water-cooling system and the heat measurement system make it possible to measure the heat output within the range of 100 W to 5 kW with pinpoint accuracy. In the year 2004 we are planning to go on with our research work.

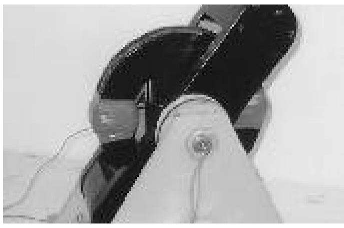

Fig.2

We have completed the second stage of designing and developing a high-efficiency magnetic generator with the neodymium permanent magnets of 1.25 Tesla. We confirmed the possibility of double output increase from the usage of two magnetic induction counter fluxes produced by the two coils wound on the generator core (Fig.2).

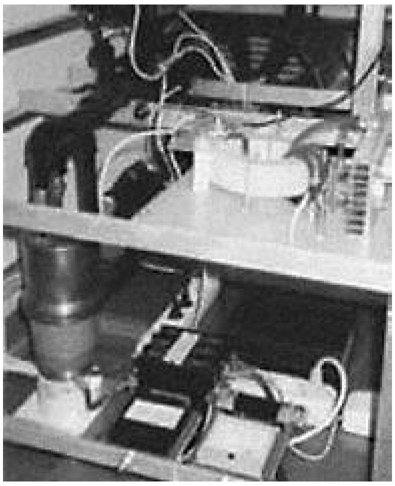

Fig.1

Our laboratory continues designing and perfecting a powerful “alternator”-type generator with permanent magnets and a drum-type rotor. We will give our readers more detailed information after we are aware of the patent examining result.

Our research on “The Control of Temporal Characteristics of Physical Processes by means of Ether Energy Density Change” is also awaiting the results of the patent examining group. This research on time and the construction of the “time machine” is not a pure theory but it has quite real prospects of being applied to the aerospace industry as a propulsion method.

It is worthy to note that now the alternative energy prospects in Russia do not entirely depend on financial factors. The cooperation is important between researchers, investors and organizations interested in the innovative development of Russian industry rather than in the development of Russia as the world primary natural resource is also important. Progressive political and public organizations should develop and introduce alternative energy programmes as well as render support to local researchers. The successful innovation of new technologies is impossible without this support and lobbying government. Let us hope that the year 2004 will bring along new opportunities and promising contacts.

Alexander V. Frolov
General Director Faraday Laboratory Ltd
http://www.faraday.ru

Novozhilov's Motor

Yury N. Novozhilov, Russia
Bratislavskaya Str., 25-6, 390035, Ryazan,
Tel.: (80912) 72-12-72

An interesting but a very simple motor has been designed. It operates on the basis of heating differences of its components due to temperature variations of contacting environments or due to variable radiation intensity.

Motor structure

The motor structure (see Fig.1) includes a metal or plastic wheel (1). In the center of the wheel there is an axis (2), on which spoke ends are fixed (3). Other spoke ends come through a hole in the wheel rim. On the spoke end there are weights, for example, in the form of small balls (4).

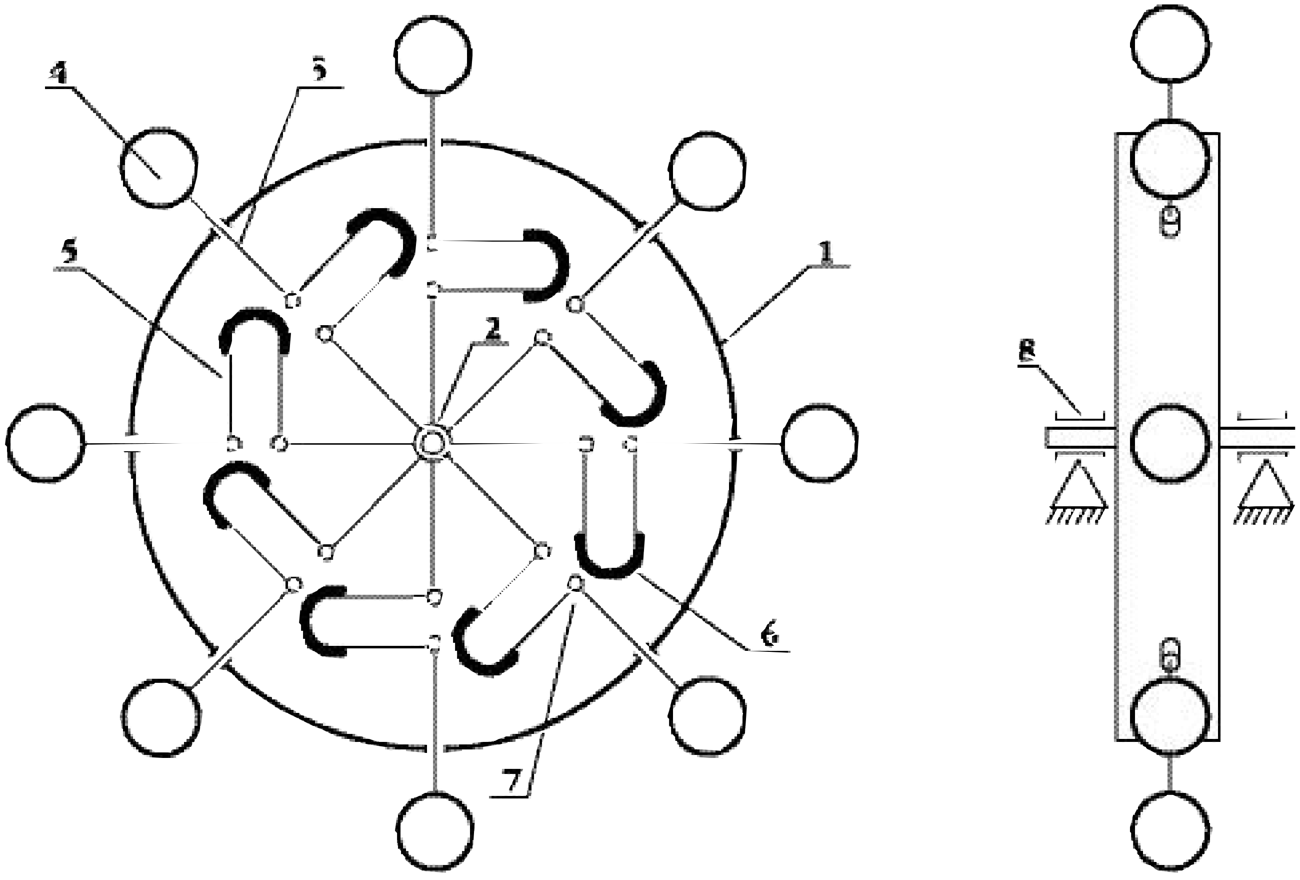

Fig.1
Motor

The structure peculiarity is preconditioned by a loop-shaped part placed on each spoke (5), which has a form of a metal strip or core. An arched bimetal component is built in the center of each loop-shaped part (6). Joints can be fixed at the joining point of loop-shaped parts and spokes to simplify the motor movement (7).

In the motor structure described arched bimetal components become straight when heated. The motor axis ends are placed in bearings (8).

To initiate the motor movement the loop-shaped parts need to be in mediums (zones) with different temperature. For instance, the lower part of the wheel may be placed into hot water (9) and the upper part will stay in the air (10).

Motor operation

When the motor is in a medium with identical temperature, for example, in the air, the temperature of all bimetal components will also be the same; therefore, they have identical form. In this case all spokes will similarly project over the wheel rim and all weights have similar distance from the wheel axis.

This state makes the whole system balanced and immovable. But if the motor wheel is partly placed into water with the temperature exceeding that of the ambient air **the bimetal components in water will become straight and shift the axis weights from the wheel axis.** This phenomenon initiates motor operation.

Let us consider this process in more detail.

On touching water, the bimetal component in the loop-shaped part becomes straight and shifts a weight from the wheel axis. The wheel becomes unbalanced, obtains torque and turns clockwise. This rotation makes the following bimetal component on the next loop-shaped part of the following spoke touch water. This bimetal component also becomes straight and shifts a weight from the wheel axis, generating an increased torque that turns the wheel clockwise.

At the same time a bimetal component comes from hot water on the left side of the wheel in zone “A”. The component cools and restores its initial form, i.e. it curves more and pulls the weight on the spoke to the rotation axis. This leads to a decrease of its counterclockwise torque that affects the wheel.

The loop-shaped parts on the left side (A) and the right side (B) of the motor wheel will cause asymmetry in the number of weights projected by the spokes. On the motor left side (A) the weights will be projected from the wheel axis when the spokes are located in corner “C”. On the right side (B) the weights on the spoke axes will be shifted from the axis when the spokes are located in corner “D”. Fig.2 shows that corner “D” is far larger than corner “C”. This means that more weights will be shifted on the right side of the wheel in zone “B” than on the left side in zone “A”.

Fig.2 shows that the wheel weights generating a clockwise torque are projected on the right side. On the left side in zone “A” only one weight is projected and it generates a counterclockwise torque. More weights shifted from the wheel axis on the right side will generate a greater clockwise torque than the weights shifted from the wheel axis on the left side. This is the cause of the wheel clockwise rotation.

This interesting effect is achieved due to the loop-shaped parts on the wheel spokes with bimetal components placed on them. When the next wheel spoke reaches corner “D”, the bimetal component connected with it touches hot water, becomes straight and shifts the corresponding weight further from the wheel rotation axis. The motor wheel continues rotation, making the next bimetal component touch water. It is also heated, becomes straight and shifts the corresponding weight from the wheel rotation axis. The wheel continues rotation. Bimetal components are touching water in succession. On the left side, in zone “A”, the bimetal components successively come out of hot water, cool in comparatively colder air, return to the initial sharply curved form and pull the weights on spokes to the wheel rotation axis. This is the reason for the wheel torque to decrease.

Mass movement from the axis in a rotating object commonly causes reduction of rotation speed. And, on the contrary, the rotation speed increases when the mass comes close to the axis. In motor zone “A” the weights come close to the rotation axis after the corresponding bimetal components come out of hot water, and the same number of weights move from the center in zone “B”, i.e. these effects on the rotation wheel compensate each other.

In essence this is a heat machine. The heat transfer from a warmer medium – water – to the cold medium – air – is achieved owing to the heat capacity of the band-form bimetal component. This is the reason for fast heating and cooling processes. Due to the fact that bimetal components are not large, not so much heat is needed for their heating.

The motor will operate when the air temperature exceeds that of water but in this case it rotates counterclockwise.

There is another model of this motor when nitanol components are used instead of bimetal ones. Nitanol is a memory metal alloy that changes shape when reaching certain temperature. In this instance, the nitanol component is normally curved but becomes straight when heated (for example, up to 50^0 C), and the weight on the spoke end moves from the wheel rotation axis initiating rotation of the wheel. That means similar operation of a nitanol-based and bimetal-based motor.

The motor structure is very simple and its operation does not require power or fuel supply: it is enough to place its wheel into the water the temperature of which is different from that of the air.

The motor structure may be further developed: for example, a focused sun ray can be used for the heating purposes; the bimetal component can be heated with a common electric lamp on one side of the wheel, the required heat may come from a heating radiator, stove or burning gas-jet. Such motor may be heated when installed on a mantelpiece, etc.

This article describes the motor model that can be used as a toy. Being supplied with looking-glass units it may be applied to get various luminous effects or may be installed in the shop showcases, or else at exhibitions. This motor is also able to operate in a gravity force environment, even on the Moon. There direct sun rays may heat bimetal components to 200^0 C, and the temperature of the components located in the shadow almost reaches absolute zero, i.e. the temperature difference in this instance is over 400^0 C, which enables constant efficient motor operation.

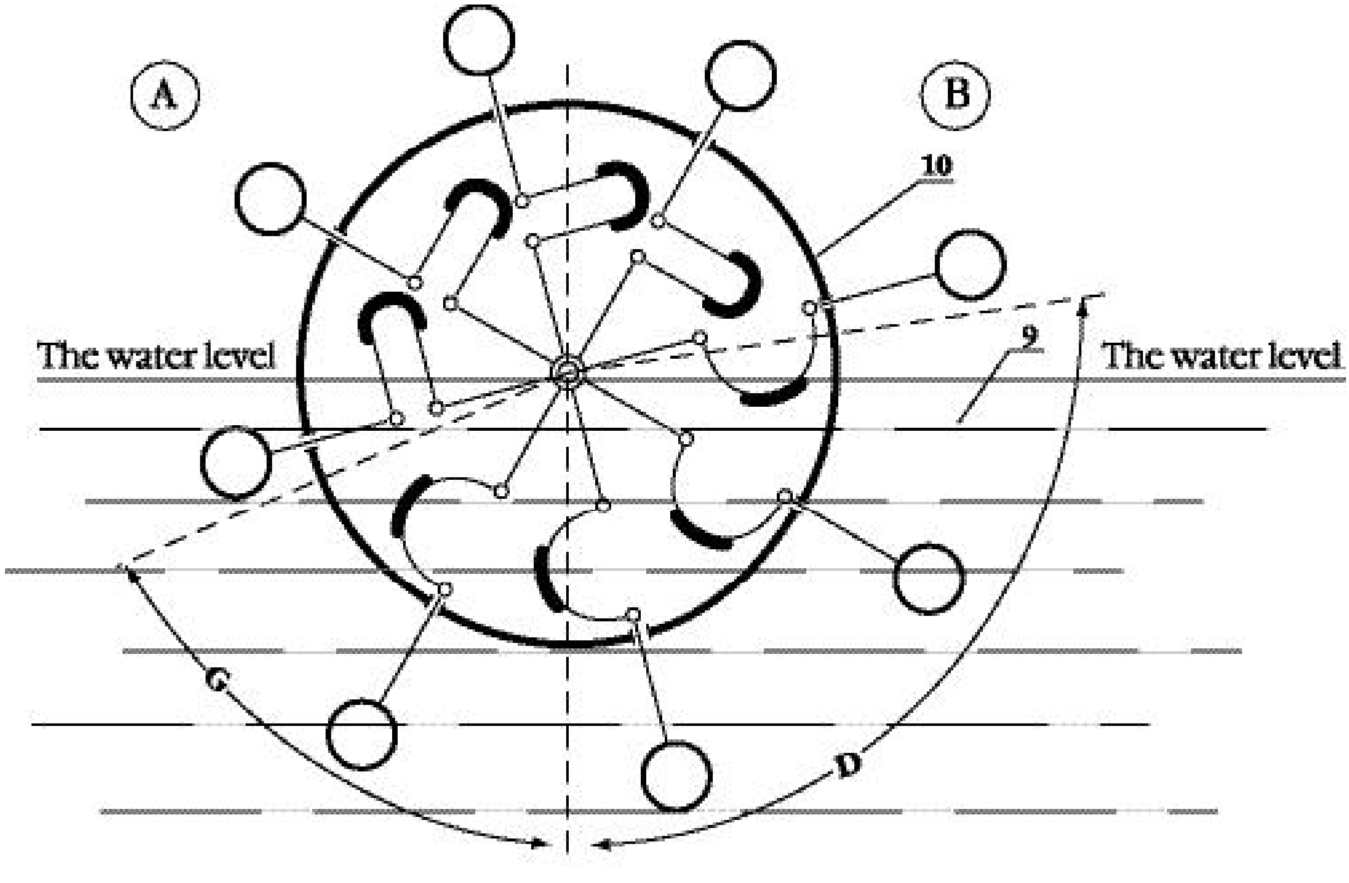

Fig.2
The motor in operation

Self-Interaction in Electromagnetic Rotation

S.A. Gerasimov , V.V. Stashenko, Russia
Physics Department, Rostov State University, Rostov-on-Don, 344090
Email:GSIM1953@mail.ru

Introduction

The electromagnetic rotation is the motion of a magnetized body ina direction perpendicular to both the vector of magnetization J_m and the direction of the electric current j in a liquid inside which the magnet is located (Fig. 1).

Editor: Readers can compare this vector composition and the device scheme of Godin and Roschin with the Searl experiments. The similarity is evident. *A. V. Frolov*

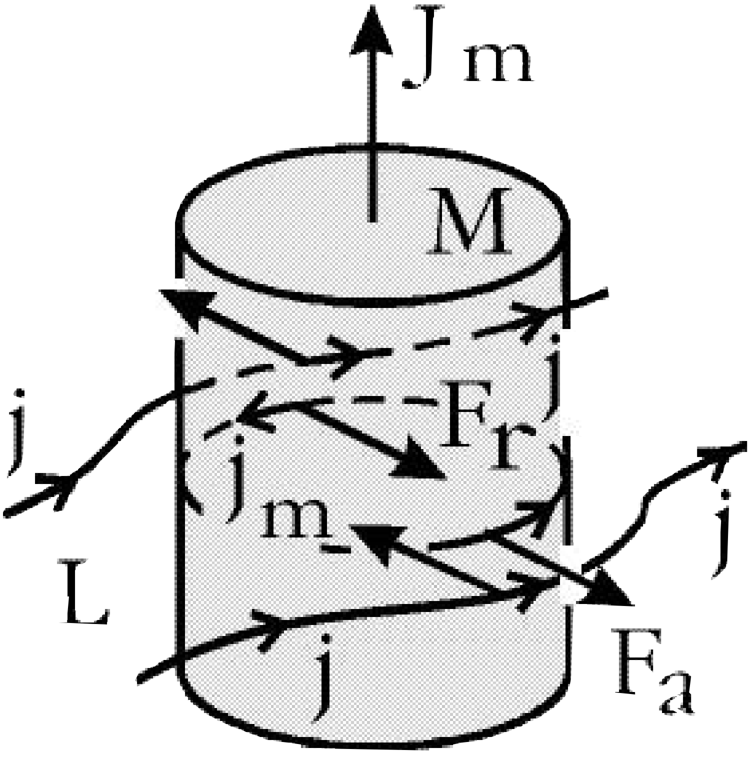

Fig.1
Electromagnetic rotation. A magnet M of magnetization J_m that is partly or entirely submerged in a conductive liquid L with a direct current of density j is moving in the direction of the force F_a+F_r

The force producing such a motion is sufficiently weak and, normally, the phenomenon is observed at rotation. Although, this is quite a well-explored phenomenon [1-6], nobody knows what makes the magnet rotate [4]. There is an opinion [2,3] that the motion of the magnet in the conductive liquid is caused by the forces of attraction $\mathbf{F}_a$ and repulsion $\mathbf{F}_r$ between currents of density $\mathbf{j}$ flowing in the liquid and surface magnetization currents $\mathbf{j}_m$ of the magnet, often called the Ampere currents as shown in Fig. 1.

The external magnetic field of a long cylindrical magnet originating far from either pole is zero. Therefore, no force is exerted on charges moving in the liquid far from the ends of the magnet. Nevertheless, the magnet still moves in the liquid with the direct current even if it is only a thin magnetized needle [5]. Thus, the deeper the magnet is immersed in the conductive liquid, the more liquid gathers around it. The deeper the immersion is, the higher the total force must be. That should be so.

In reality, it is quite different. During intermediate immersions when the magnet is partly submerged in the liquid, the torque N does not depend on the depth of immersion, in the other words, currents flowing round the cylindrical magnet do not influence the exerting force acting [6]. Moreover, during shallow immersions, the magnet moves together with the liquid. **This is none other than *the self-interaction* [7].** The conductive liquid under the magnet experiences the magnetic action. The direction of this force coincides with the direction of the magnet motion. **Thus, the part of the conductive liquid moving in this direction under the action of the magnetic field B, affected by the frictional forces in the liquid causes the motion of the magnet in the same direction.** The only way to find out the role of the self-interaction in this phenomenon is experimental investigation. It would be appropriate to exclude the currents flowing under the magnet. In this case the magnet must rotate in the opposite direction.

Experimental system

A commercially available cylindrical magnet M of magnetization J_m=1,95•10^5 A/m and a balanced load *P* connected by a rocker R are suspended by a thread *T* as shown in Fig. 2.

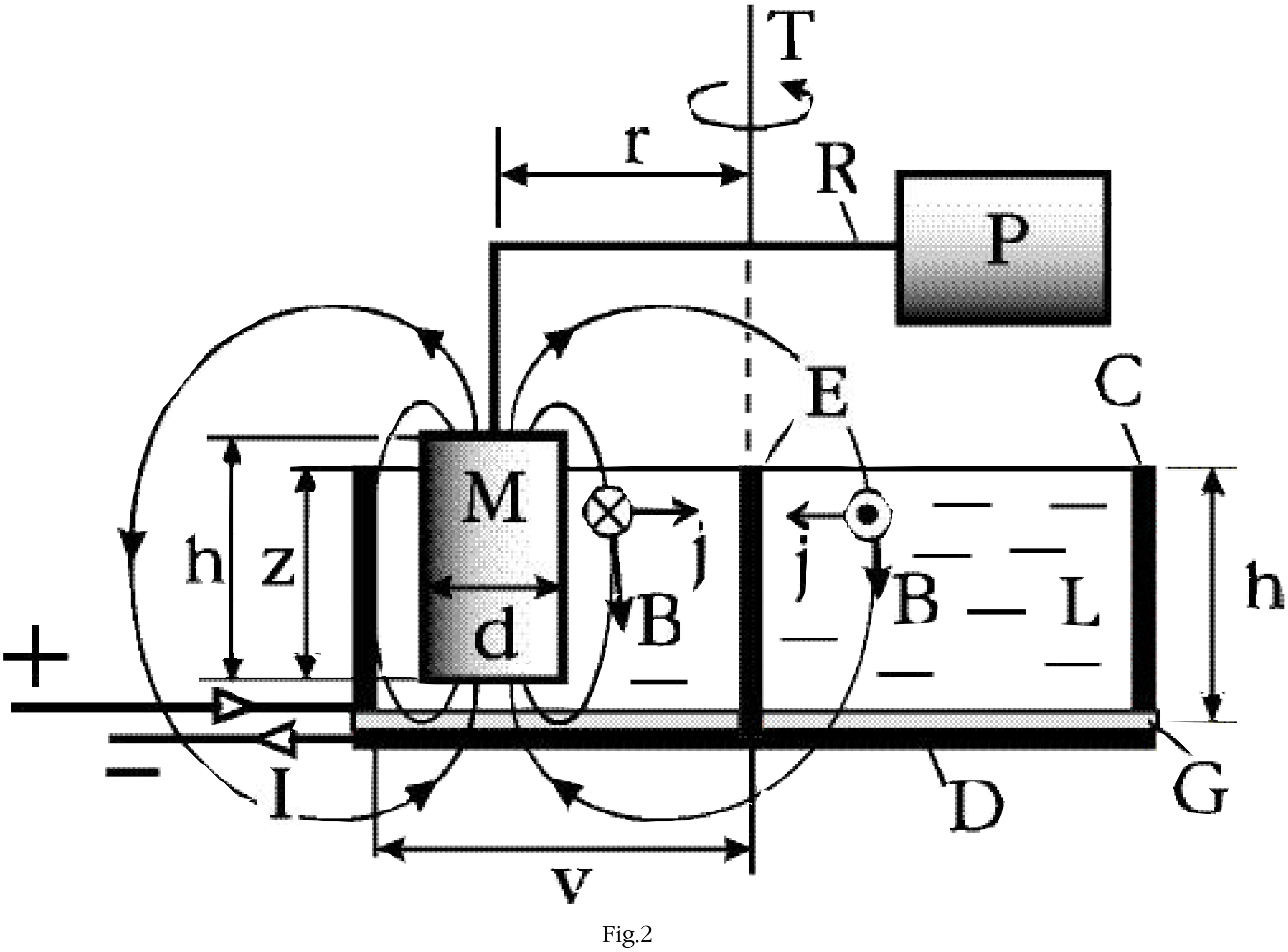

Fig.2
The experimental set and its parameters: h=50 mm, r=35 mm, d=25 mm, v=70 mm, the diameter of the central electrode E is 5 mm, the thickness of the bottom G is 2 mm. (×) and (·) are directions of the force acting on current elements of current density j in the magnetic field of inductance B

The magnet is immersed in the 5% copper sulfate solution ($CuSO_4$) so that the depth *z* of immersion can vary. A vessel containing the conductive liquid *L* is large enough for the magnet to be almost entirely immersed. The cylindrical surface *C* of the vessel and the central electrode *E* of the system are made of non-magnetic materials. The bottom of the vessel *G* is, of course, insulator. The magnet is coated with an insulating moisture-proof varnish. The length of the electrode *E* equals the height h of the vessel. A direct current of strength *I* passing through the cylindrical electrode *C*, the conductive liquid *L*, the central electrode *E* and the disk electrode *D* generate a torque of electromagnetic origin which results to the rotation of the magnet. The disk electrode *D* is intended to provide the system with the symmetrical supply of the current.

Experimental results

The size of this experimental device differs from that described previously [6]. **As a result we have obtained the torque N, which is about ten times more intensive than the former one.** But the main result is that the torque does not change its sign when the magnet is being immersed in the conductive liquid (Fig. 3).

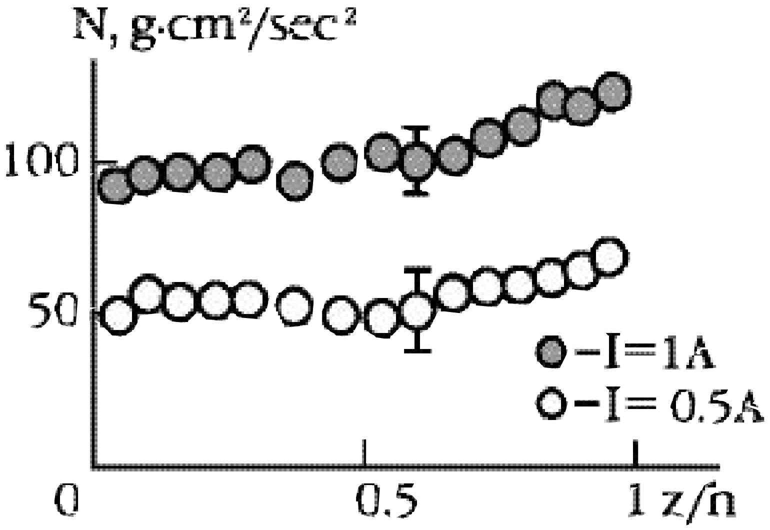

Fig.3
A typical experimental dependencies of the torque *N*
acting on the magnet depending on the depth of immersion *z*

When the magnet is almost entirely submerged in the liquid the depth of which equals the height of the magnet, no currents are flowing under and above it. In this case the rotation of the magnet is expected to be the result of drag action of the rotating liquid located in the intermediate field relative to the magnet. This part of the conductive liquid is rotating in the direction opposite to that of the motion of the magnetized body. It seems to be the right way. The magnet pushes the liquid away and, therefore, is moving in the opposite direction. No, this is wrong. If it was correct, this effect would also occur during the shallow immersion. At small *z* the magnet and the liquid rotate together. So far we still do not know what makes the magnet rotate.

S.A. Gerasimov

References

1. Faraday M. Experimental Researches on Electricity. // London: Ed. by R. and J.E. Taylor. 1839.
2. Ampere A.M. Theorie Mathematique des Phenomenes Electrodynamiques. // Paris: Blanchard. 1958.
3. Sigalov R.G., Shapovalova T.I., Karimov H.H., Samsonov N.I. // Magnetic Fields and Their New Applications. // Moscow: Nauka. 1976.
4. Gerasimov S.A., Ershov A.V. What Makes a Magnet Rotate? // Physics Education (Moscow). 2002. V. 22. P. 70-75.
5. Grabovsky M.A., Mlodseevsky A.B., Telesnin R.V., Shaskolskaya M.P., Yakovlev I.A. Lecture Demonstrations on Physics. // Moscow: Nauka. 1965.
6. Gerasimov S.A., Ershov A.V. Mechanism of Electromagnetic Rotation. // Problems of Applied Physics. 2001. V. 7. P. 34-35.
7. Gerasimov S.A. Self-Interaction and Vector Potential in Magnetostatics. // Physica Scripta. 1997. V. 56. N 3-4. P. 462-464.

Capillary Motor

Iosiph I. Elshansky, Russia
Koshtoyantsa St., 1a - 75, 119454 Moscow
Email: ewrik@bisinter.ru

For several decades the author has been organizing a contest for young inventors on the radio, television, in mass media and the Internet. There have been accumulated thousands of engineering solutions that belong to schoolchildren and youth. There are many original ones, including perpetual mobile. Though unfeasible in principle, they have interesting structure and competent engineering solutions.

For example, a child's dream to create an oblique roof over a city or field and direct the rain flow to a turbine to obtain mechanical work is of little use. But it's quite another matter if this child takes a pencil, a sheet of paper and tries to make a sketch or a drawing of such construction. Even if the construction is primitive and unfeasible, the child will get his first engineering experience and self-reliance.

I can still remember one peculiar case. Once the editors were instructed (it was in the times of the USSR) to assist an honored inventor of new machines. It was a Stalin Prize laureate, creator of a special type of lime that was widely applied in construction (I will not disclose his name). The inventor was given (that was not common practice) a large room in a shared apartment situated in a prestigious block of apartments in Moscow. It turned out that all room space, except for narrow passages at the walls, was occupied by a huge table. On the table up to the ceiling there were some pipelines, glass and metal vessels, retorts and other laboratory ware. The most impressive thing was retorts heated by gas coming through a hose from the shared kitchen. All that staff was bubbling, hissing and steaming. The inventor was creating a perpetual mobile!

I will give one more example of a situation common for the editors of the "Inventor and Rationalizer". Once we had a respectable-looking visitor who made an odd request to get the President of the French Academy of Sciences to respond to his letters about a perpetual mobile plan. The essence of Alexander Rodionov's (Maloyaroslavets, Russia) invention was that "in accordance with Newton's and Guerin's laws fluid flows up through capillaries and turns a wheel when flowing down (see Fig.1).

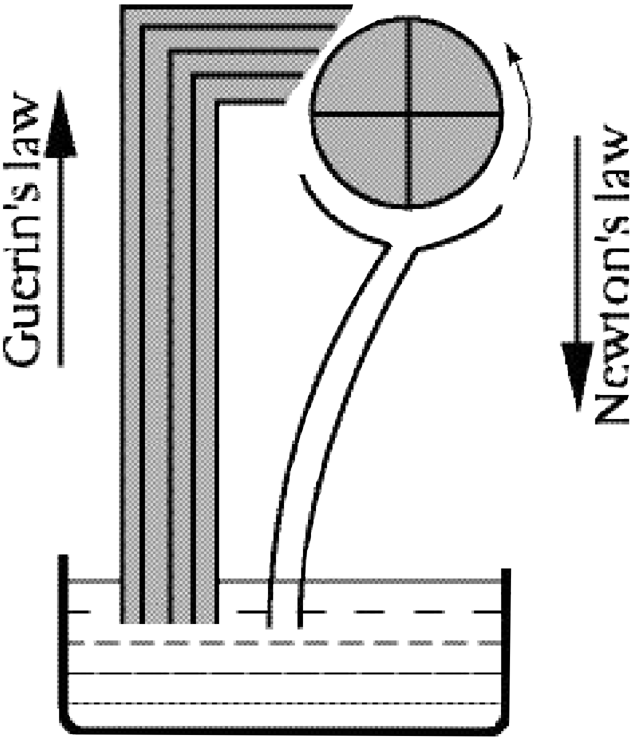

Fig. 1
Perpetuum-mobile

There are many similar projects in the history of engineering. Hence, this case may be considered as typical. We all know that the molecular adhesion forces (moistening) that push the fluid upwards will stick to it and the capillary motor will never work. But is the idea totally unfeasible? Since false modesty is inappropriate when it comes to perpetual mobile, I will present my ideas on this not so honorable subject for the reader's judgment.

It seems that all perpetual mobile inventors start not with a pencil and a sheet of paper but with experimenting on magnets, oblique plane, wicks, wheels, springs and other materials at hand. My first heat (perpetual) motor was also the result of an experiment.

The "Capillary" idea of Alexander Rodionov haunted me. I recalled that the great Lomonosov devoted lots of time to studying molecular adhesion and capillarity. They are present in everyday life, let alone the nature, especially in plants that would not be able to exist without them. How else would moisture flow up the trunks and stems? But, on the other hand, according to Mikhail V. Lomonosov's data, water flows up only for the maximum of several dozens of millimeters even through the thinnest capillary. But trees may have the height of several dozens of meters! If, as it is customarily considered, moisture flows from one capillary to another of its own accord, why not suppose that a capillary perpetual mobile is feasible? Explanations that moisture in plants flows up due to root pressure do not sound plausible. So where is the truth?

Years have passed since that visit of A. Rodionov. Now, just like I used to sympathize with, as it then seemed, a rather odd inventor, now my relatives sympathize with me. One can not blame them: for weeks and months there have been flowerpots all around the apartment - on the tables, window-sills and bookstands. With bottles, tubes and self-made water manometers to follow.

Once, when I was assembling a device, I could not find two similar glass tubes and had to put in the tube of thin polyethylene. But no matter how hard I worked the water in the communicating vessels was at different levels. In the glass tube the water level was always higher. It could not be otherwise, if you come to think of it, but isn't it the right time the words "made of equally wettable material" were inserted into the communicating vessels law?

Here is a primitive experiment: make a hole in the bottom of a plastic flowerpot containing soil and a plant, and insert there one end of an elastic tube. The other end is inserted into the bottom of a plastic bottle containing water (see Fig. 3). In accordance with the communicating vessels law, the water will flow from the bottle into the flowerpot. When we see that the soil has been made wet, the experiment shall get more complicated: elevate the flowerpot so that its bottom is higher than the liquid level in the bottle. In some days we shall notice that the soil has not dried up and the water level in the bottle has lowered. One should not be a Solomon to guess that the soil or the roots of the plant transfer the water from the lower to the higher level. Here is a typical natural perpetual mobile!

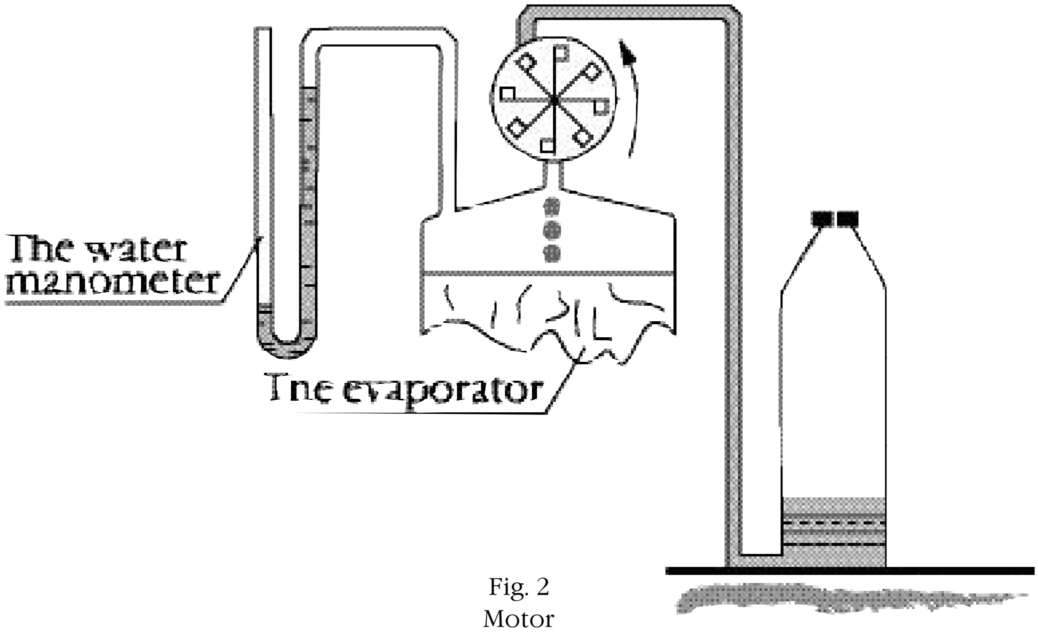

Fig. 2
Motor

It is no use referring to the molecular adhesion forces since they can not be applied at such height differences. What is it then? A molecular collision as in a boiling kettle? But a leaf has neither high temperature nor much space. Nevertheless, water evaporates both from the soil and leaf surfaces. At the same time there is rarefaction under the evaporating surface. Can it be the rarefaction that pulls up new portions of water?

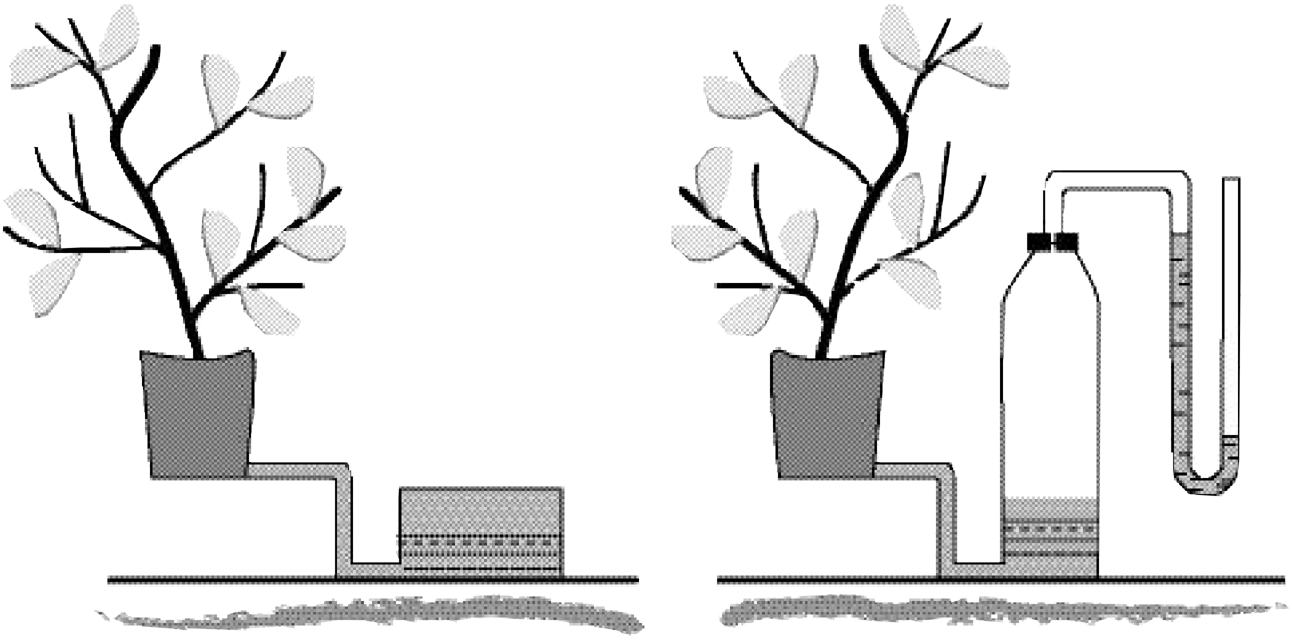

Fig. 3
Rarefaction

Let us make the experiment still more complicated. Put the leg pipe of a liquid manometer into the bottle cork. (Fig. 3) Almost immediately the water level in the closest leg to the bottle will start rising. It means that there is rarefaction in the bottle. Can it be the rarefaction that "rules" the plants?

Most likely, the essence of the mistake made by Rodionov and other authors of capillary motors consisted in attempts to get the water outflow from the capillary. **If it is not flowing out but evaporating, as it is common for the soil or plants, then, probably, vacuum will manifest itself and the capillary perpetual mobile will start working.**

I made my first capillary perpetual mobile many years ago (see Fig. 4). And it worked: though slowly, the wheel was steadily turning. No doubt, it was due to the heat inflow from the air. Therefore, such motor will work if there is a temperature difference and relative humidity of surrounding air is below 100%.

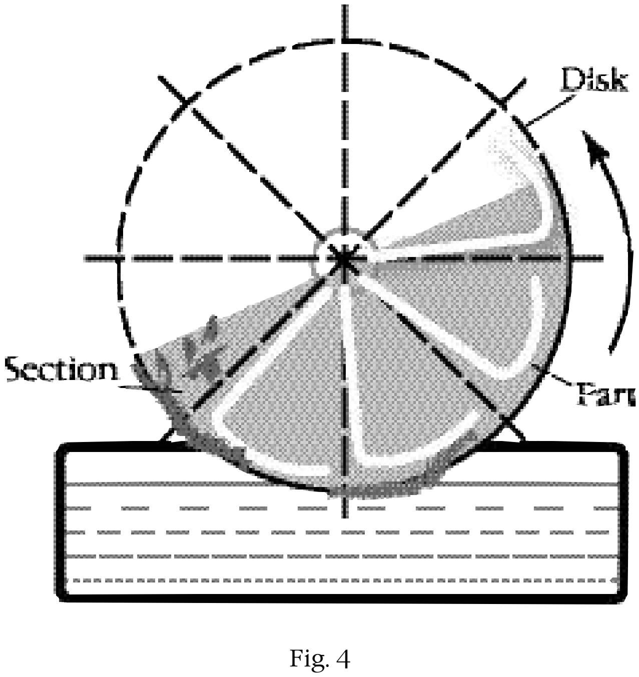

Fig. 4
Capillary Motor

The heat motor operating principle is clear from this figure. The disk (many disks on the common axis) is divided into isolated sections. Their surface is covered with a moisture-proof piece of material (filtered paper, cloth, etc.). The part of each section somewhat gets under the base of the next one. Hence, a small part of the base of one section and a part of the part following the section can get into water simultaneously. Due to the capillary moistening the section and the part will start to absorb water. Since the section is shifted about the center of gravity, the turning moment is bound to appear. The section that is out of water starts to dry up, the moment will increase and the wheel will start turning.

Having determined that, though slowly, the motor is turning, the author applied to a patent authority for a patent on the invention of the "Perpetual mobile", meaning it as a joke. I hoped that the experts (I knew some of them personally) will understand this joke. But, unexpectedly, the device was considered an invention, though under the name of the "Heat motor" (certificate of authorship USSR # 1455040).

Basically, the motor is similar to the project of A. Rodionov. I would not be surprised if he protested. The same capillaries and the spontaneous rise of liquid on the moisture-proof material. The only difference is that the liquid does not flow out but evaporates from the capillaries.

Almost without any reason - one can not predict his eccentricities - I have recently made one more capillary motor (see Fig. 3). Perhaps, it resulted from my experiments with plants. This motor directly and spontaneously lifts water to a higher level, which is more convincing. Moreover, water flows in the form of drops (or a stream) and turns the propeller! The motor is quite operable and has even been shown on TV. But do not try to verify it on your own - such work is useless without certain know-how. The author would not like to disclose the "secret" yet. I could still make a patent application.

NEWS

The Continuous Rotation Device

The essence of this invention is that in a closed circle the rotation of the object occurs without any outer influence or any power sources (electric energy, oil, diesel oil, etc.), which results in the rotation of the rotating part of the device. The torque can be easily enhanced by means of transmitting it to the reduction system. At that rate, the necessary rotation speed and rotation power can be acquired.

The device is small, easy to transport, mobile (both, in operation and switch-off), simple and inexpensive to make and assemble. It is economical, ecological and weather independent.

The device can be used in any industrial or household appliance field where the continuous rotation is required. It is especially important to note that this device can be used to generate energy on a large industrial scale as well as in small laboratories, in mountains and remote districts, in tunnels, at sea, etc. for it does not require any power lines.

Patent Claim Priority:
#20030059, April 19th 2003

For more information
contact
the author of the invention:
Eldar Sariyev,
Azerbaydzhan, Baku,
Tbilisskiy Prospekt, 75-26;
Tel: + (994 12) 92-47-73, 98-95-02;
Email: eldar_sariyev@yahoo.com

Gravity-heat Power System

Vladimir N. Sukhanov, Russia
Kosmonavtov Str., 35-39, 420061 Kazan
Email: Wladimir.Sukhanov@soros.ksu.ru

Natural hydrologic cycle is a widely recognized phenomenon. Water evaporates from the ground and water reservoir surfaces, generally, due to the Sun heat, then moves to the superstructure atmospheric layers where it is condensed, and precipitates onto the ground surface, but on higher levels compared with evaporation surfaces. In this process water is also purified and desalted. From the higher levels water flows to its main evaporation points, forming streams like brooks and rivers where water power plants could be established. Flowing to the lower surface layers with lower potential energy, water in the Earth's gravitational field performs work that can be utilized.

Natural hydrologic cycle principle is used in the gravity-heat power system (USA patent No. 3953971, International class. F03G7/04 of May 4, 1976).

In this system the temperature difference between ground surface layers (the mountain foot and its top) is used to produce electric power and obtain fresh water. The temperature difference does not change much with change of height, and the system effectiveness is quite moderate, which can be named as its drawback.

The system utilizes free environmental energy. Its efficiency (converted to the process maintaining energy) tends to infinity.

The author proposes a new gravity-heat power system that can substitute modern heat power plants. It contains all installation components as per USA patent No. 3953971 except for the heat exchanger on the ground surface whose functions are performed by a conventional steam boiler.

Gravity energy component **G** in the proposed system:

$$G = X_m \cdot T\, g\, H\,,$$

where

- X_m – fluid flow,
- **T** – time of operation,
- **g** – acceleration of gravity,
- **H** – tower height between the steam boiler and the freezer (condenser).

This formula does not allow for vapor density of operating fluid.

The vapor density is insignificant as compared with the operating fluid density.

Energy **P**, required to maintain operation:

$$P = X_m \cdot T\, q\,,$$

where **q** – specific fluid vaporization energy

Efficiency **K** – of the system should not be lower than 80%

To obtain efficiency that equals 120% the following is required:

$$H = 2{,}2q/g$$

Or **H** = 25,4 kilometers.

Construction of such towers is too complicated for modern technology. Therefore, a stratospheric balloon that is linked to the ground surface by a flexible twin-core hose rope may be used. One core is designed to send down fluid and the other to send up vapor. Flowing down to the ground surface through the hose, the fluid will generate fluid column pressure. The fluid will cross water-turbine generator cascades, after each of which the fluid pressure will decrease and the fluid will continue its way down.

There is an alternative. To make the tower several times lower the evaporation (at the tower foot) and condensation (at the tower/mountain top) should be phased (in the form of cascade) at various pressure values and with utilization of the same heat energy. For this purpose a heat carrier is used that will transfer heat from the tower (mountain) top to its foot. This operation will enable the tower height segmentation into **H/n** sections where **n** is the number of grades in a cascade.

Cascade structure is shown in Fig. 1

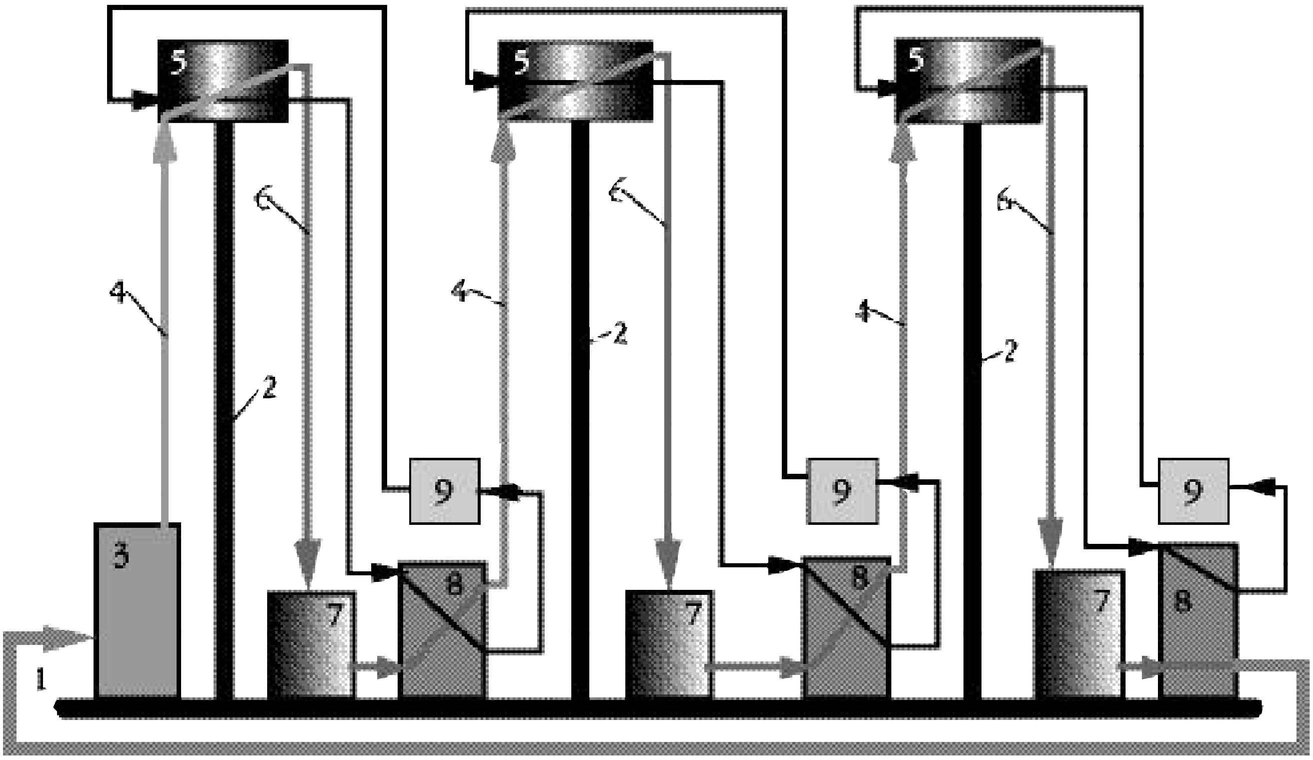

Fig. 1
Cascade

where

1 – ground surface,
2 – tower component or the smoke-stack of the steam boiler,
3 – steam boiler,
4 – steam pipe,
5 – heat exchanger and condenser,
6 – pressure pipeline,
7 – water-turbine generator,
8 – heat exchanger and evaporator,
9 – circulation pump,
10 – pipeline of the heat carrier circulation.

Cascade (operating fluid, water) with 500 meters of height can contain several dozens of grades (50 and more). **In this case the cascade efficiency may approach 120%.**

Decreasing the tower height to 500 meters (for water) will allow for utilization of conventional water-turbine generators and standard equipment.

In the proposed cascade the heat energy is required to maintain the cascade working capacity when Earth's gravitational field energy is utilized. In the author's opinion there are still other ways to use gravitation.

The system scale is its main drawback. Therefore, partial utilization of the proposed principle may seem more attractive. The proposed engineering solution was simple.

In a thermal power station the condenser is installed on the top of the tower (mountain). After the condenser the water lift in the water tower finished at the steam boiler pressure pipe (omitting the pump that was no longer necessary). Energy required for pressure pipe operation was saved. In other words, the water tower was utilizing the Earth's gravitation at the conventional juncture of vapor condensation in the steam plant.

Such proposal was made on June 17, 1982 in the patent application USSR N3453603/06, (101161) expert code 060701KH. However, the utilization of free and environmentally friendly gravity energy did not arouse interest even on the part of patent experts.

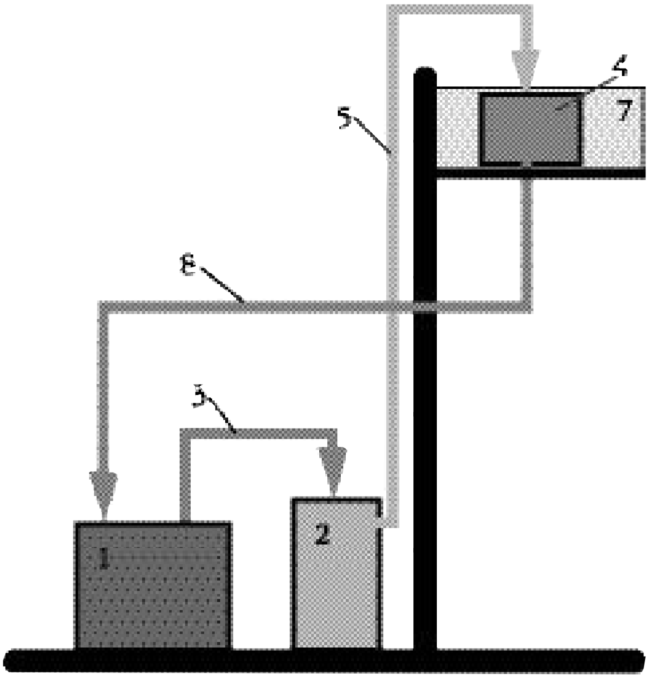

Fig. 2
System

The author is looking for serious partners.

The Alternative Energy Potential in Russia

According to the International Energy Agency (IEA) calculations, total investments into the energy industry constitute approximately 330 billion dollars annually. Almost half of this sum (150 billion annually) is allotted to the development of the electric power industry. This information only partially reflects the potential of the alternative energy development and does not include the market growth factor upon the introduction of autonomous power system technologies. Vast territories of the planet have not been sufficiently developed yet due to the absence of local energy resources and power lines. From this point of view, Russia is the country with an enormous consumer demand potential provided that the mass production of autonomous fuel-less energy systems guarantees low production prices.

Alexander V. Frolov
Editor-in-Chief

Faraday
Lab Ltd

Fuel-less energy R&D
R&D in alternative astronautics

Read our
New Energy Technonlogies
magazine

Lev Tolstoy Str. 7-601
St.Petersburg
197376, Russia
Tel: 7-812-3803844
www.faraday.ru

Modern Perpetual Mobile

Review prepared by correspondent Alla Pashova, Russia.
Based on the Internet pages and other information sources

It is common knowledge that any experimental information, which is against a conventional scientific paradigm, is considered to be false, extracted from the society and finally is turned into something **classified and elite.**

Politics-oriented science as a whole and natural science in particular often result in misrepresentation and falsehood of **popular knowledge** that turns it into the fighting weapon of socio-political organizations, states and their groups.

World-outlook and political contradictions of any society often determine science development strategies. In 1885 Nicola Tesla presented his transformer in action by means of which he lit carbon lamps in a 25 miles circle without any cords, conductors or switches using the 5000-HP turbine of the Niagara hydroelectric power station. Afterward, he obtained financial support for one of his energy projects.

On a special testing ground Nicola Tesla built vacuum-energy systems. However, in 1885 the systems, as well as the testing ground, were destroyed as it became evident that if they went on the mankind would never need organic fuel again. But the destruction of the systems and testing ground does not at all mean the destruction of papers and documents...Ever since that moment people have been trying to find "free energy".

The experiment of distant lighting of carbon lamps without using any cords was repeated only once by the physicist Filippov. He managed to light the lamps in Tsarskoe Selo employing the electric generating systems in St. Petersburg. Filippov died in 1914 under mysterious circumstances.

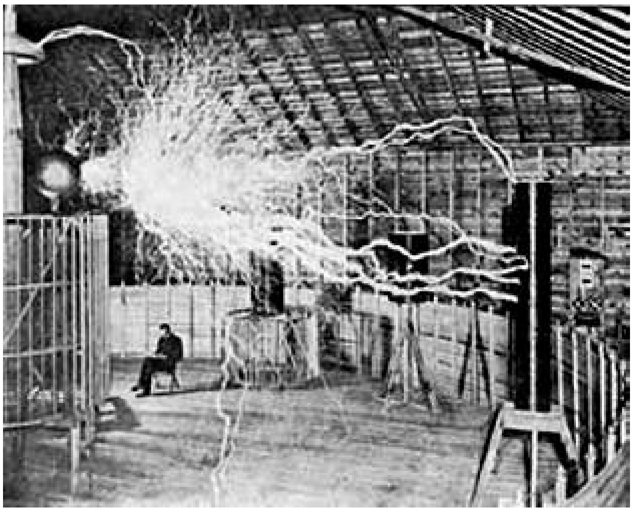

In 1917 the Portuguese American Andres invented non-convetional fuel for an internal combustion engine. The essence of his invention was in some inexpensive chemicals, which he dissolved in water (some drops of chemicals per pail of water). A special state committee tested the new fuel during the New York - Washington - New York automobile races. After the races one of the largest oil monopolies in the U.S. offered Andres 2 million dollars in cash for the documents and all rights for his invention and later locked all the information in their safes. Two days later after getting the money Andres disappeared. (The information was disclosed by V. Vasilevsky, the former chief of the scientific and technical intelligence department of the KGB of the USSR, who had been in charge of the department since 1930).

The above-mentioned examples show that there were a lot of attempts to develop revolutionary energy technologies. But what then are the grounds for the 'energy crisis' complaints? The answer is simple: the desire of a group of financial and energy multinational corporations for an economic, political and, finally, for a total control of the world. That is why there is no room for "free-energy" hunters in our science, not even as laboratory assistants, not to mention some higher positions.

Practice often negates scientific dogmas, especially when it comes to the efficiency of different devices. In general, the traditional understanding of efficiency is quite limited and should be forgotten. Instead of the traditional notion "efficiency" we should consider the notion "energy-conversion efficiency" (ECE), which has no limits. In 1974 in the USA a six-stroke internal combustion engine was designed which had a double ECE. The fifth stroke of the engine operation included water injection; the sixth was characterized by the water vapor introduction. Firstly, the engine had a higher ECE than that of the Carnot cycle. Secondly, taking the 55 % ECE of the contemporary internal combustion engine, the ECE of a six-stroke engine exceeds 1.

Practice often negates scientific dogmas, especially when it comes to the efficiency of different devices.In general, the traditional understanding of efficiency is quite limited and should be forgotten. Instead of the traditional notion"efficiency" we should consider the notion "energy-conversion efficiency"...

Before the war there appeared "studebekkers" with the fuel consumption of 5.5 l per 100 km. But the fuel-economy record was set by the Japanese. In 1986 they built up an automobile, which used only 0.055 l of fuel per 100 km (about 44 g). Unfortunately, now we do not have plants, which could manufacture automobiles of the kind.

In 1832 Tomson came up with the idea of **a heat-pump unit**. Stating the unity and interconvertibility of substance movements, he proved that the mechanical energy consumption can restore stray heat. At that time cross-Atlantic ships had low-temperature steam engines, which employed exhaust steam and ether vapors. The efficiency of those engines was already higher than the Carnot cycle efficiency.

Now with the help of the facts acquired during the last century we will prove the scientific inconsistency of the orthodox model of the second stage of thermodynamics. In 1941 the English physicists O. Hougen and K. Watson published their work featuring the experimental fact that the ammonia-saturated vapor pressure in liquids was higher than the general pressure. For example, with the general pressure of 50 atm and the temperature of 0^0 C, such pressure was 17 % higher. This means that having placed a semitransparent partition and a turbine between two containers it could have been possible to build up a natural-heat vapor-converter. This method was used by **von Platen, the Nobel prizewinner from Sweden, who constructed a self-rotating centrifugal machine, which provided a 1000 atm pressure when ammonia interacted with water. The heat emitted was enough to compensate friction losses**. It is also known that the Carnot principle cannot be applied to closed cycles. W. Vielstich, 'Combustion Cells', 1968: **"If an entropy change of a reaction is negative, it can exceed 1**. It means that a certain amount of energy can be obtained from the natural heat. Such an effect is possible in electrochemical generators, which employ a direct oxidation process before oxidation and dioxidation".

A demonstrative example of scientific experiments withdrawn from science is discovery # 13 "The knock power-transfer regularity" made on December, 18th, 1962, which makes it possible to create a mechanical "perpetual mobile". The experiment shows that the conventional "knock theory" does not work in practice: **the bouncing energy of a body after it has been knocked can be higher than its energy before it has been knocked**.

Trying to get some recognition, **E. Aleksandrov, Doctor of Engineering Science, made his demonstrative experiment in front of many different commissions: a chilled-steel ball when falling from a 10 m height onto a firmly-fixed chilled-steel plate made a 14-15 m-high bounce**. That was the notorious 'perpetual mobile' as it is. It went on like this until somebody decided to explain the result of the experiment by a metal lattice internal energy. At once, everything was clear. Since that moment the above-mentioned principle has been successfully employed in percussion-action machines. This principle can also be used in simple electric-power systems.

Now we can see that electro-technicians had already constructed demonstrative 'perpetual mobile' machines 10 years before politicians and economists started to speak about the 'energy crisis'.

In 1921 the mass media wrote about A. Hubbard who invented an electromagnetic generator without any external energy supply. The Hubbard generator was used as a boat motor.

> *...now we can see that electrotechnicians had already constructed demonstrative 'perpetual mobile' machines 10 years before politicians and economists started to speak about the 'energy crisis'...*

In 1928 L. Nidershot invented a 300 W electric generator, which did not require any external energy supply. The device consisted of a radio-technical oscillation generator (500 kHz) and a coil. After 68 years there appeared publications about A. Melnichenko who repeated the experiment.

In 1927 T. Brown (England) obtained a patent on the ways of originating a moving power and energy using an electric field. Later, in 1955 while working in France, he presented a system with the speed-capability of several miles per hour using a 2000 eV field. After that he had to stop his experiments and was sent to the USA.

In 1943 N. Tesla presented an electric-motor automobile. The energy was generated by a previously-unknown generator.

In 1960 Stovbunenko (whose research work results were subject to a special verdict of the Military-Industrial Complex) presented his motor, which made it possible to move in his old automobile 'Moskvich' around the city for the whole day utilizing the power of a standard automobile battery.

In 1980 there was a qualitative break through in electro-technical "perpetual mobiles". Baumann's electrostatic machines of the total capacity of 750 kW started functioning in the religious community of Liden (Switzerland), serving all daily needs of the village. (See the photos below)

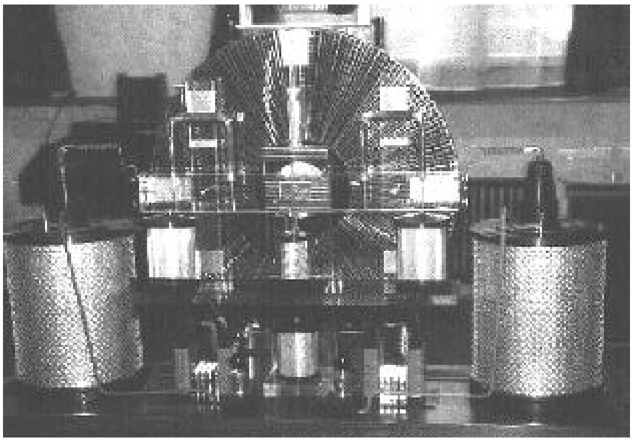

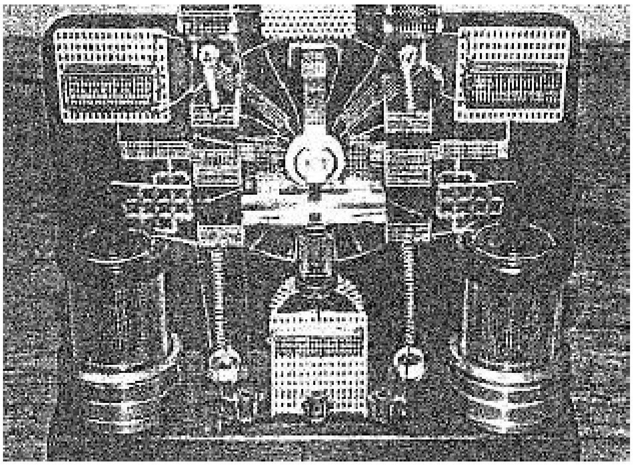

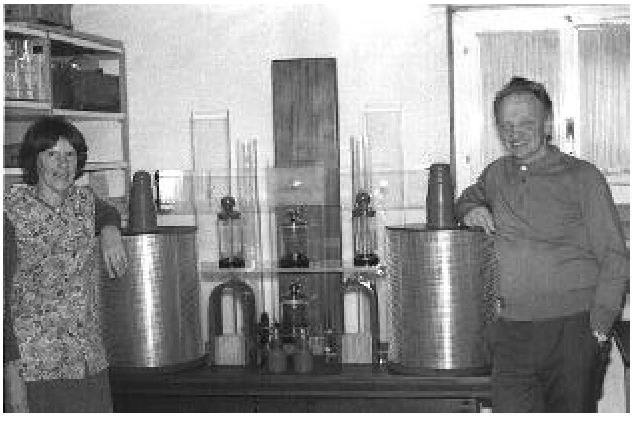

The photos are taken from www.free-energy.cc

Thus, in 1980 there appeared a community with no energy problems, no need for organic fuel and no fear for the 'energy crisis'.

In 1881 N. Sluginov discovered an energy imbalance that occurred in the water electrolysis process. In his experiments he got that the output energy was 30 % higher than the input energy. In 1980 American scientists recreated this energy imbalance and proved that while employing the steam turbine rejected heat, the water electrolysis efficiency reaches 120%.

This information introduces a story about an energy system invented in 1957 under the guidance of I. Filimonenko. The system did not simply generate energy in the form of a high-pressure vapor and output hydrogen and oxygen but also absorbed radiation. In 1960 the USSR Central Committee and Cabinet Council gave a special secret ruling, known as “The Three 'K's'” (Keldysh, Kurchatov, Korolyov) to go on with the research in this system. However, after Kurchatov's death the project was "oppressed" and later on after the death of Korolyov the project was completely cancelled. The special committee of the USSR Academy of Sciences stated that the system was functioning against "the Law of Nature". I. Filimonenko was dismissed and expelled from the party. Later, in 1980-1991 the experiments were partially renewed. Several testing systems were constructed in the Chelyabinsk region but the construction was never finished and the use of a mobile system for eliminating the damage caused by the Chernobyl accident was refused. This story is a demonstrative example of the traditional rejection of promising projects by science, which could be very beneficial for Russia.

Theoretical Background of 'Perpetual Mobile'

Originally, the term 'entropy' was used only in connection with the reasoning about limited or eternal nature of the Universe. Later it came to be used to describe the functioning principles of thermal machines. Now there exist more than 600 contradictory definitions for 'entropy', which are actually the results of a long and unfinished discussion between the supporters of Dekart and Leibnitz, who defined 'motion' in different ways. It is these permanent contradictions in the theoretical fundamentals of physics that make it impossible to explain the existence of a 'perpetual mobile'. However, standard physics has never denied the possibility of creating energy technologies allowing the ECE higher than 1.

We should remember that the basis of the so-called variational calculation is the mathematical apparatus technique of the conventional mechanics. It seems that if system behavior equations directly depend on time, this system cannot fall under any conservation laws. **This is the proof of a complete failure of the conventional energy conservation law!**

> *It is these permanent contradictions in the theoretical fundamentals of physics that make it impossible to explain the existence of a 'perpetual mobile'*

In time-independent thermodynamics the existence of 'perpetual mobile' can be explained by the system internal energy usage. There exist two kinds of thermodynamic systems. Some of them when being heated or affected in some other way start increasing their internal energy while other systems show the internal energy decrease. The latter are called negative absorption systems. The standard version of the first law of thermodynamics does not cover all the processes occurring in such systems. The output of a negative absorption system can be higher than its energy. The additional output is acquired by the gradual internal energy loss.

To cap it all, there are more than a hundred experimental facts that show the limited nature of the Maxwell theory. They were obtained by scientists from different research institutes and construction bureaus, were tested and registered. Thus, for example, in 1973 in the USSR the acoustic magnetoelectricity effect was discovered. The author of the discovery proved the interaction of electrons with ultrasonic waves with a 1000-times energy increase. This is against the Maxwell theory, which absolutely denies effects of the kind. But the limited nature of the theory was never accepted.

However, the examples of the machines in action prove that since 1834 when the first refrigerator was invented, when the cross-Atlantic ships were equipped with low-temperature ether-vapor steam-engines, the contemporary science had no right to introduce 'the second holy principle', especially in the orthodox form of 'the Carnot cycle efficiency'. The modern theoretical physics cannot provide grounds for the negation of creating over-unity devices. That is why we should rely on the physicists-experimentalists who develop the science in a practical rather than theoretical way.

Alekseenko's Fuel-less Motor

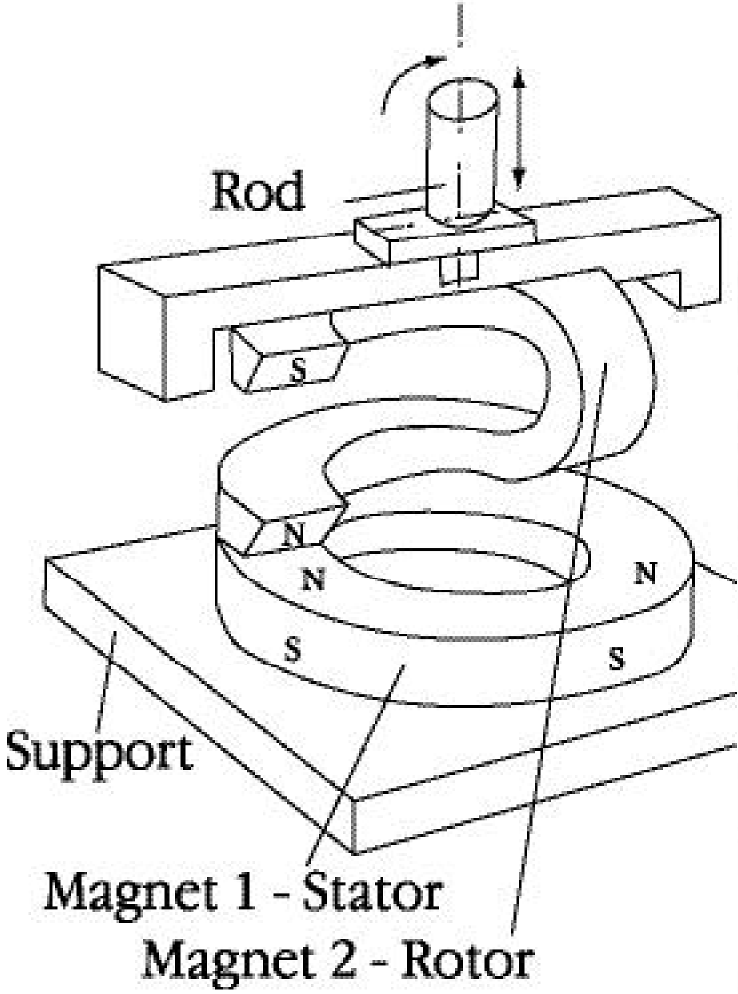

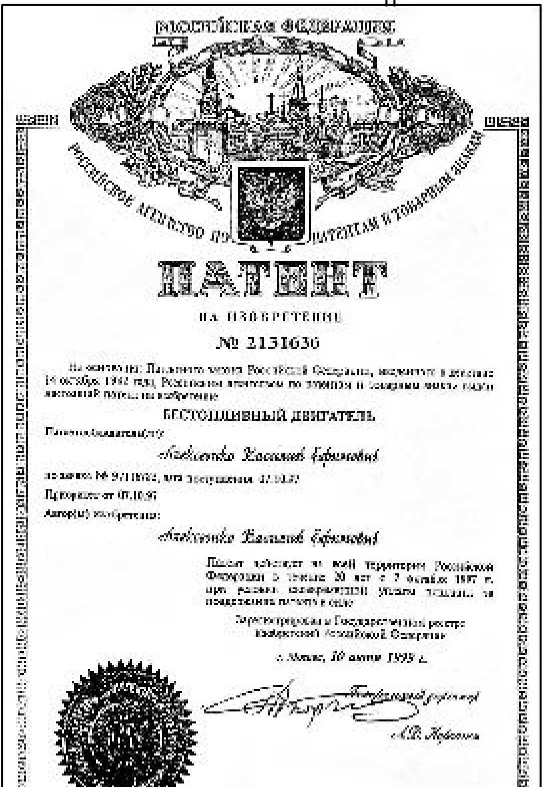
РОССИЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ИЗОБРЕТЕНИЕ

№ 2131636

БЕСТОПЛИВНЫЙ ДВИГАТЕЛЬ

Patent # 2131636
Vasiliy E. Alekseenko

Aerospace Defense Research Finds Free Energy and Antigravity Possible

Dr. Ines Espinoza, USA
email: dr.ines@vasantcorporation.com

Editor: it is worth mentioning that the first gravitational-wave physical mechanism and the mechanism of spin-effect "grazer" (gravitational lasers) designing were described by S. M. Polyakov and O.S. Polyakov in the book "An Introduction to the Experimental Gravitational Studies", Moscow, 1991.

After a 7-year research study, a senior staff aerospace defense engineer, George J. Bugh, concludes economically free energy and even antigravity are possible.

A 7-year study of "free energy" devices, sometimes called "over unity" devices, was done to determine if any of these devices generated power by transferring energy from unknown sources and if so to determine where the energy was coming from. Included in the study is research of related devices with claimed antigravity effects. The study attempted to determine validity of claims, commonalities of device characteristics and to determine how these devices could work. **The results conclude that some devices can generate economically free energy. This study also concludes there is a possible link between gravity and electromagnetism that can be exploited to generate antigravity or electrogravity effects.**

The majority of the study was to come up with a theory to explain how the devices could work. In his research, Mr. Bugh used mostly classical electrodynamics rather than quantum electrodynamics. In Quantum theory, the wave-like characteristics of matter are described using abstract probability waves. However, Bugh proposes that the wave characteristics of matter may also be described as coming from a very real sea of unseen electromagnetic standing waves among all matter. There is a slide show presentation at the website that explains the differences between the Quantum and Classical way of explaining particle interactions.

According to classical electrodynamics, all electrically charged particles, like quarks of protons and neutrons as well as orbital electrons for example, **should radiate away energy from precessional and precessional plus orbital motion**.

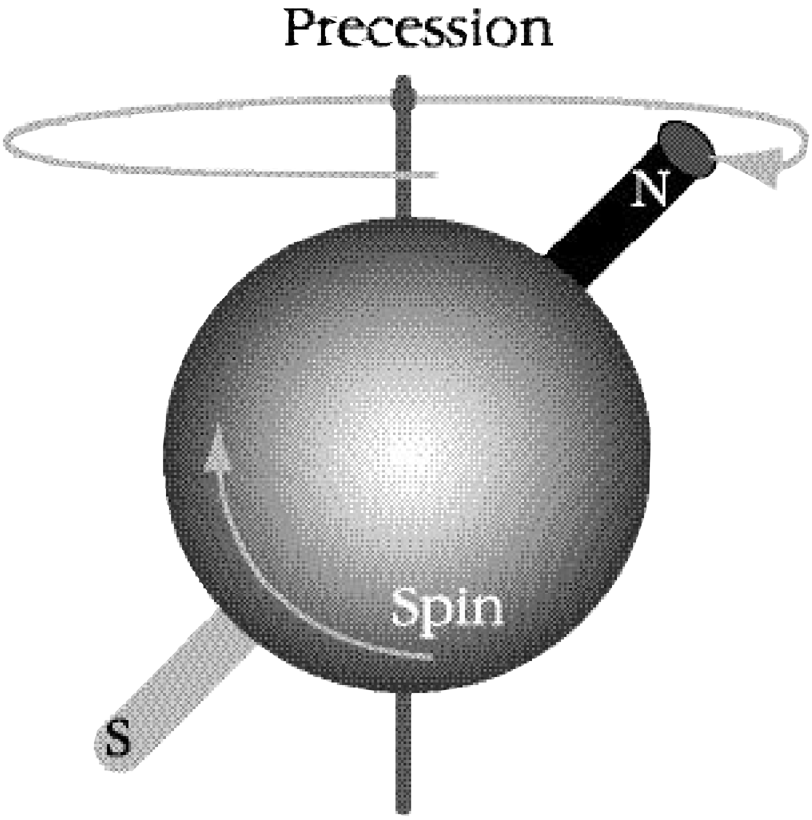

Fig.1
Electron motion

If in fact this really happens then all electrically charged particles can be radiating away energy all the time. However, all particles can also absorb just as much energy from all other radiating particles. The absorbed energy applies electromagnetic forces that naturally move all similar type particles into harmonious precessional motions with all other particles. This results in a vast sea of electromagnetic standing waves among all matter. Even free particles would move into precessional motions that are in sync with the established sea of standing waves.

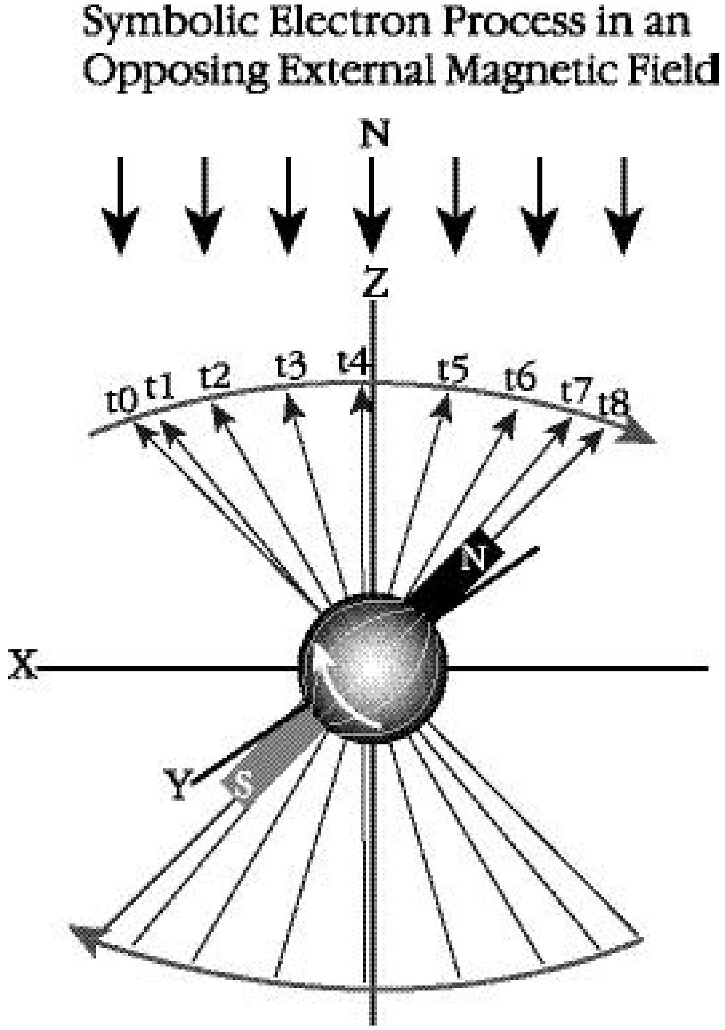
Symbolic Electron Process in an Opposing External Magnetic Field
N
Z
t0
t1
t2
t3
t4
t5
t6
t7
t8
N
X
Y
S

Fig. 2
Electron precession

There can be a hidden yet strong tendency towards harmony among all matter in the universe due to these unseen standing waves and spin interactions among all matter. This tendency can overcome to a great extent the tendency towards chaos and heat death of the universe. This tendency can also be exploited to perform work.

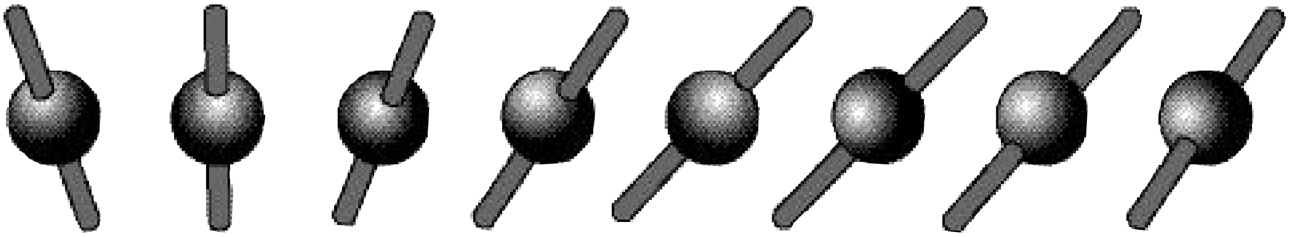

Fig. 3
Spin wave

This is an interaction among all matter that Ernst Mach alluded to as necessary to cause matter's characteristic of inertia. Einstein later called this Mach's Principle. Einstein studied Mach's ideas while developing his theory of General Relativity.

Bugh describes inertial resistance to acceleration as caused by electromagnetic forces. Changes in position of a mass will cause phase differences to develop between the precessional motions of the particles of that mass relative to the sea of standing waves. This in turn causes electromagnetic force that resists a mass from changing its position.

The research papers are published in a book and a CD titled "Spin Wave Technology Initial Release".

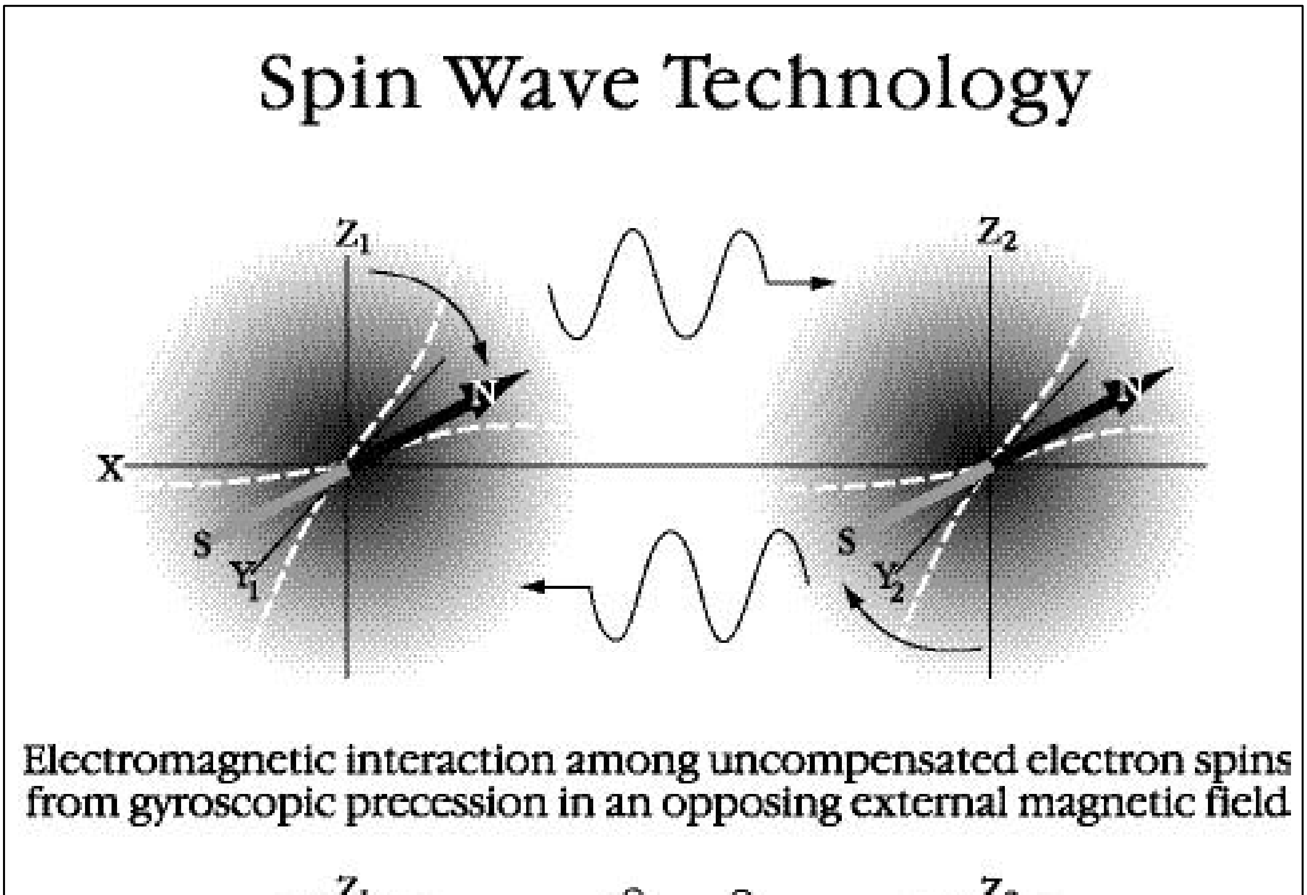
Spin Wave Technology

Electromagnetic interaction among uncompensated electron spins from gyroscopic precession in an opposing external magnetic field

Electromagnetic interaction among compensating electron spins from gyroscopic precession in magnetic fields manifested from apparent relative motion of protons of the nucleus

Fig. 4.
Spin wave technology

At the end of the book it suggests that it should be possible to create computer simulations of particle interactions using classical electrodynamics to demonstrate both inertial resistance and gravitational attraction and it should be possible to simulate devices that manipulate these particle interactions to demonstrate antigravity effects.

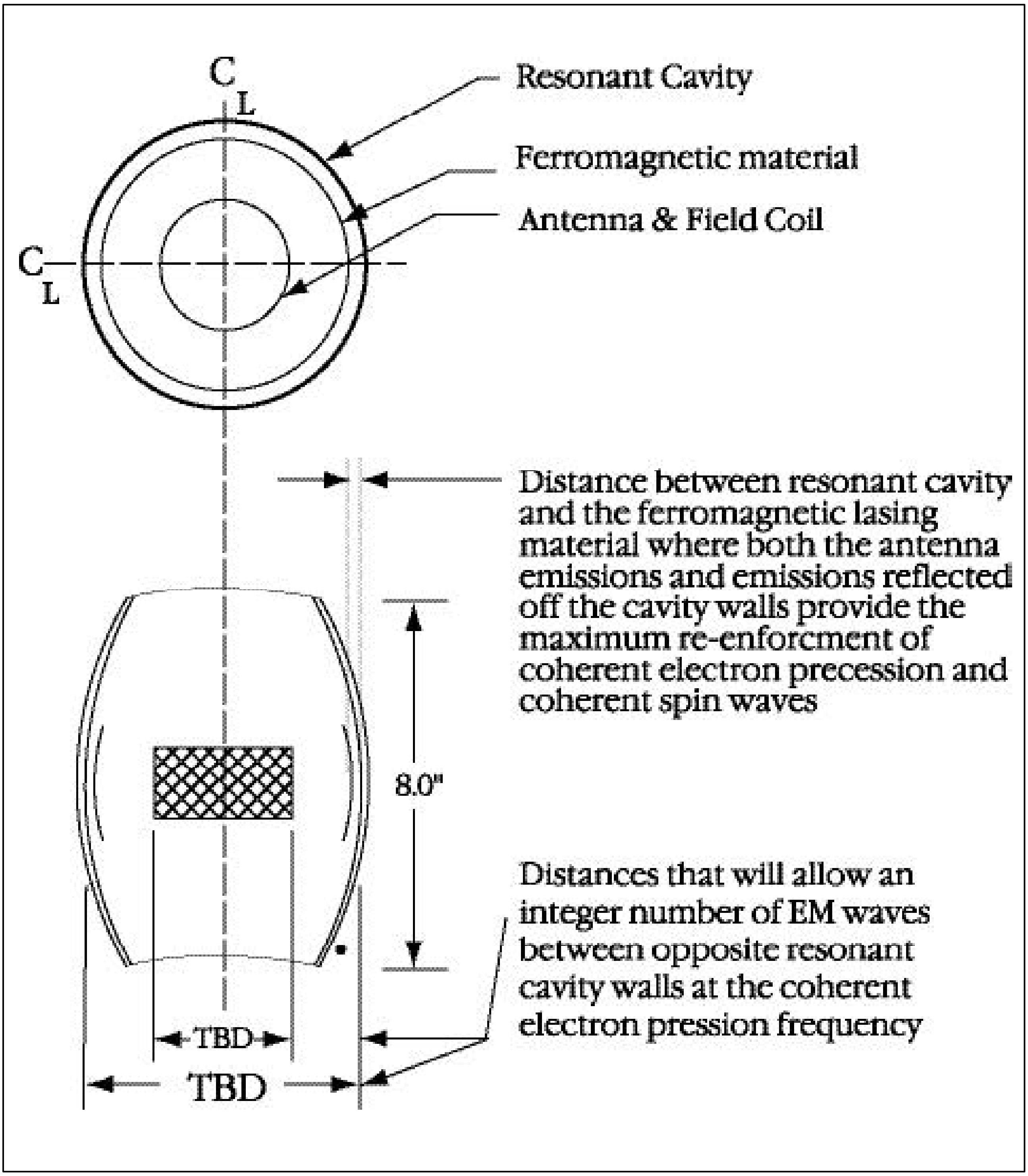
C L
Resonant Cavity
Ferromagnetic material
Antenna & Field Coil
C L
Distance between resonant cavity and the ferromagnetic lasing material where both the antenna emissions and emissions reflected off the cavity walls provide the maximum re-enforcment of coherent electron precession and coherent spin waves
8.0"
Distances that will allow an integer number of EM waves between opposite resonant cavity walls at the coherent electron pression frequency
TBD
TBD

Fig. 5
Spin wave laser

More information is available about the research results at the website: www.vasantcorporation.com.

On Physics of Electromagnetism

Mechanical Analogue or Pure Mechanics?

Yury V. Ivanko, Ukraine
E-mail: ux0lw@hotmail.com

After publication of the article "Study of possibility of a high-energy hyper-low-frequency electric field" in the "New Energy Technologies" magazine, Issue #2, 2003 I am often asked why I have mentioned the Big Bang in my hypothesis.

I would like to reassure atheists since I do not back the Big Bang hypothesis. But I had to base myself on something. At present the origin of the Universe as a result of the Big Bang has priority and is recognized by the academic science.

Both theorists and experts have long noticed a deep analogue between vortex motion of air currents and electromagnetism. My vision of the physics of electricity and electromagnetism is fully based on existence of a vortex capable physical environment. What is considered to be and is measured **as electric voltage, electric current, magnetic or gravitational interactions is nothing but similar-originating phenomena - the result of vortex flow environment affecting our detectors - dissimilar formations for these vortex flows**.

For instance, let us consider three vectors for the flow along the skin-layer of the line conductor - aerial emitter, together with alternating voltage attached to it (Fig.1):

1) longitudinal - showing as electric voltage, the gradient of potentials between the emitter feeding point and the emitter cold end, U;

2) tangential - showing as amperage - electric current, I;

3) radial - showing as the interacting force of two flows - magnetic field (attraction/repulsion), H (not to be mixed with the commonly used term of magnetic intensity).

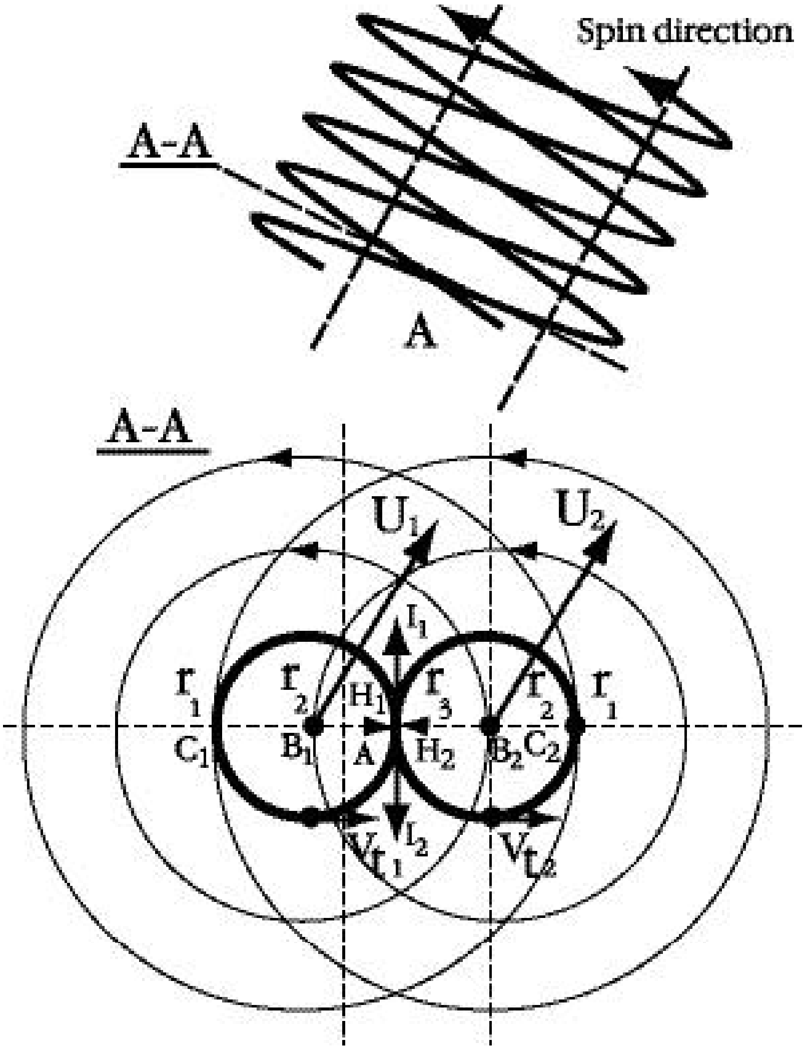

Fig.1
Interaction of two left-handed homogeneous parallel flows.

Let us consider it step by step:

Consider line conductor AB with length L in free space (Fig. 2).

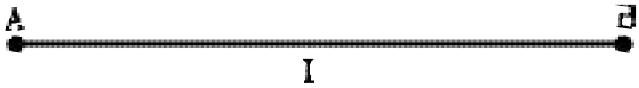

Fig.2
Line conductor in free space.

Let us assume that he is not affected by and fields. Then the gradient of potentials between A and B equals zero. Let us apply sinusoidal voltage from the generator output to point A (Fig. 3).

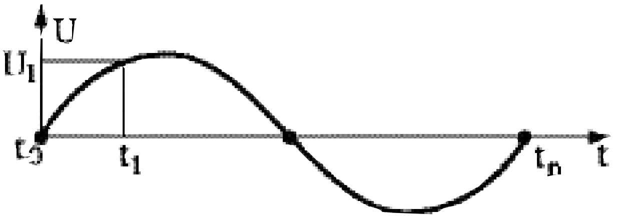

Fig.3
Sinusoidal voltage at the generator output.

Let us consider the initial time point t_0. The generator output voltage is U=0. At time point t_1 the generator output voltage has changed and took on value U_1.

It is necessary to note that (See Fig. 2) the propagation speed is finite: c=300000 km/sec. Point B will "learn" about the voltage alteration only in time $t_B=t_1+L/c$. Therefore, at time point t_1 between points A and B appears gradient of potentials $\Delta\varphi = U_1$. So what is the direction of vector U_1? Right, along conductor AB.

What practice says

One should not be a physicist or radio mechanic to notice small electro-shock of ungrounded metal objects close to aerials of powerful transmitters. Specialists term them as "electromagnetic inducers". Has an expert ever measured the actual characteristic of the field intensity distribution along a line conductor? Why should he? Everyone knows from school lessons how alternating voltage is distributed along the conductor, the length of which is comparable with the wave length (Fig. 4).

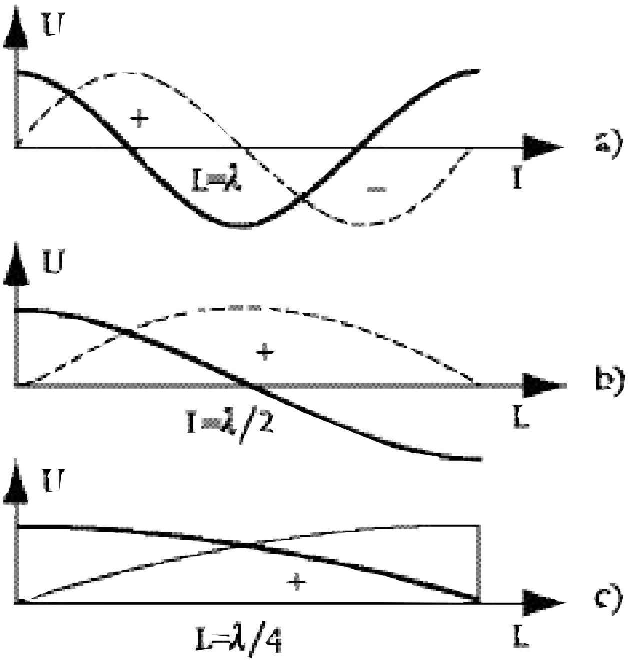

Fig.4
Classical representation of voltage distribution along a linear vibrator a) for λ, b) for $\lambda/2$, c) for $\lambda/4$.

Now let us consider what experimental measurements suggest.

An experiment of measuring the field intensity distribution along the line conductor is shown below. A classic field indicator (FI), assembled as per scheme in Fig.5, was used as an intensity measuring instrument.

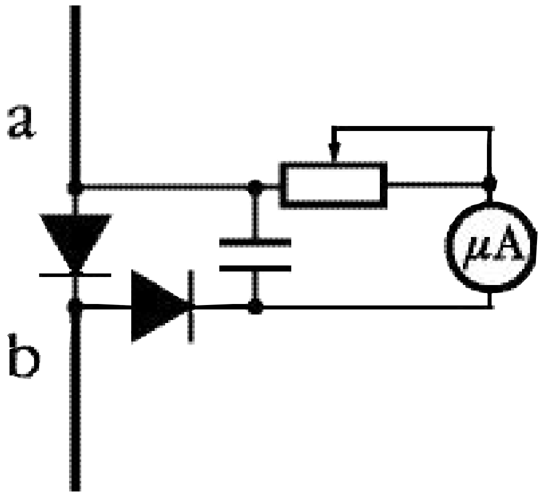

Fig.5
Field indicator principle diagram.

FI was located in immediate proximity to the vibrator but had no galvanic contact with it. Experimental frequencies were selected in accordance with the standing wave coefficient minimum (SWC) for λ, $\lambda/2$, $\lambda/4$. Sinusoidal signal generator was connected to one end of the vibrator (Fig. 6a).

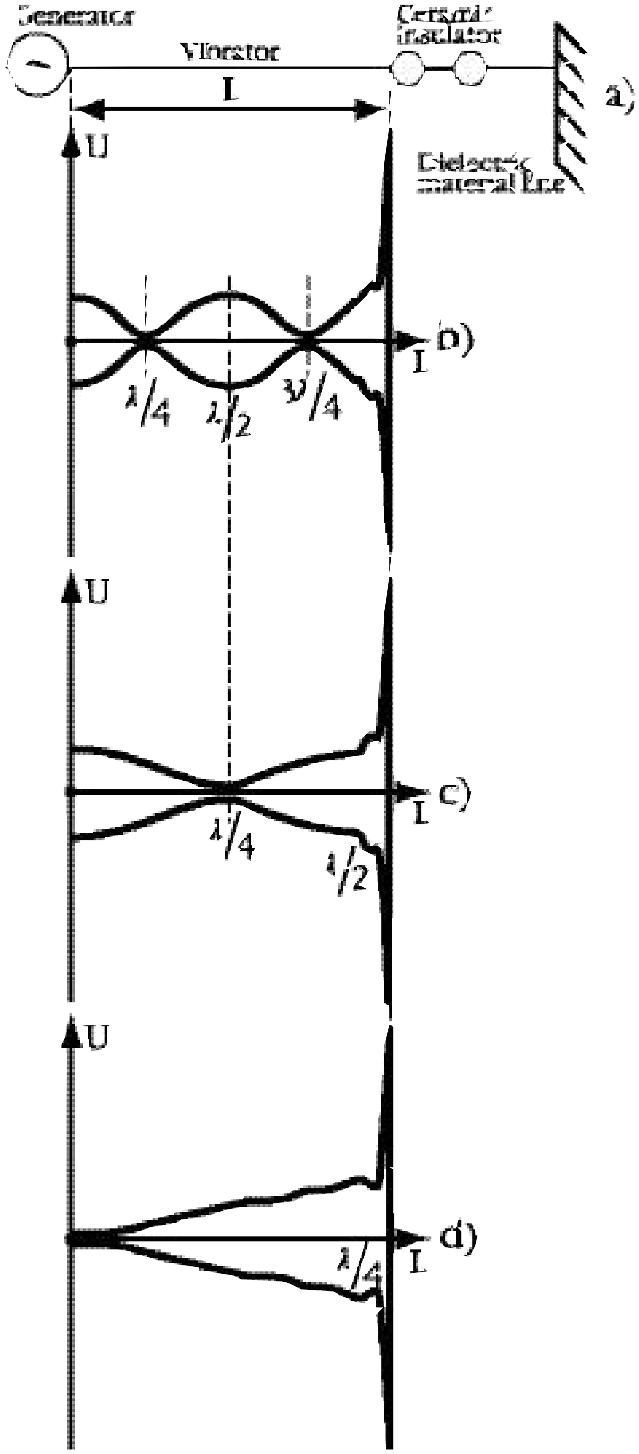

Fig.6
Measured field intensity
along a linear vibrator.

To better understand diagrams in Fig. 6 curves are drawn at the upper and lower ends of each axis L that corresponds to the vibrator location. The curves were run with FI horizontally located on both sides of the vibrator.

From Fig. 6d it is clear that the diagram is axisymmetric and looks like a conic funnel.

Meanwhile, the vibrator directional diagram measured by the monitor receiver far more distant than λ, shows the characteristic that coincides with the simulated one through the Matlab-6 system (See Fig.7). As you can see from this figure, the spatial characteristic of the directional diagram is in the form of a toroid.

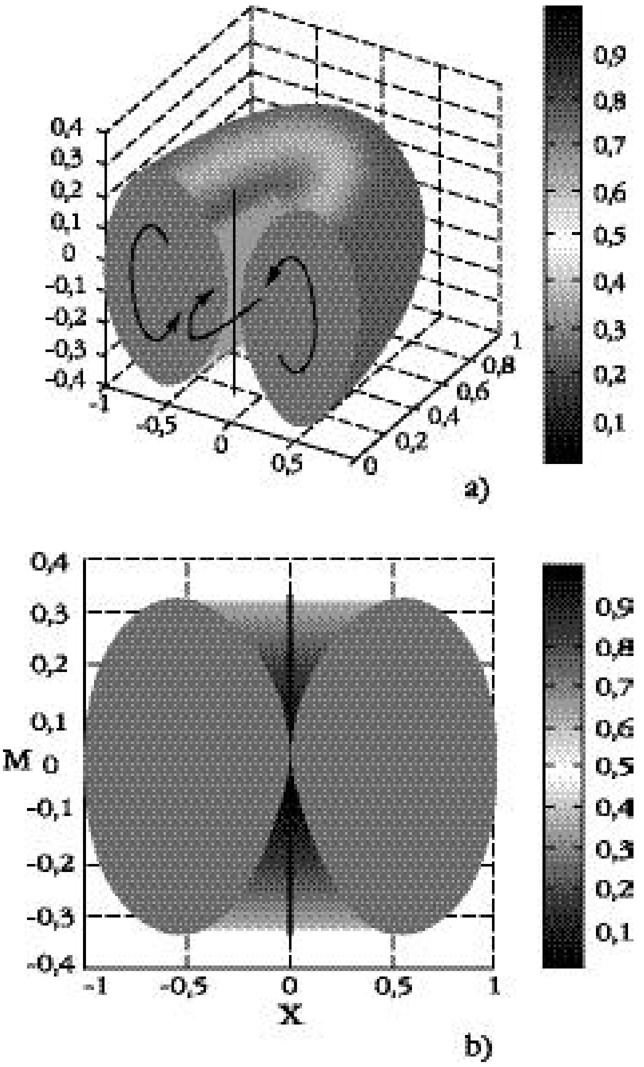

Fig.7
Matlab-6-simulated directional diagram of the semi-wave vibrator

A series of alterations was made to explore vortex ether dynamic processes with variable wave length and frequency in line conductors, solenoidal and flat coils, etc.

For certain purposes the solenoidal coil may be considered as a shortcut linear vibrator. On respective frequencies the field intensity distribution along the solenoid corresponds to the linear vibrator. This data may prove useful to many modern engineers of the Tesla transformers and generators. To increase voltage on the cold end a toroid capacitor is set in the Tesla transformers. Nuclear physicists use the hemisphere capacity for the particle accelerators. Thus, a voltage of millions of volts may be achieved.

Flat coils are best explored in terms of field intensity distribution (See Fig. 8)

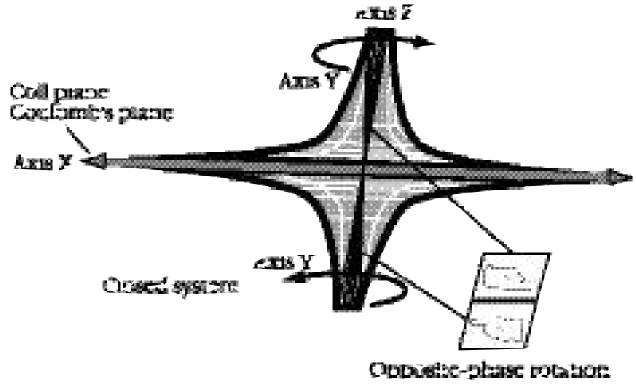

Fig.8
Field intensity distribution for a flat coil in a closed resonance circuit.

Field intensity distribution on a resonance frequency is correct.

The practical experiment for the case in fig. 6d has shown that the neon lamp that is connected with the vibrator end will be off. And this is with a 100 W generator! It is a paradox, one would say. The measurement diagram shows a "splash" of the field intensity at the vibrator end but we feel nothing. Indeed, when the FI is located as shown in Fig.9 for $\lambda/4$, the "electrodynamic vacuum" will generate. The device showed "0" intensity, which conforms neither with the diagram in Fig.6 nor with the diagram in Fig.7.

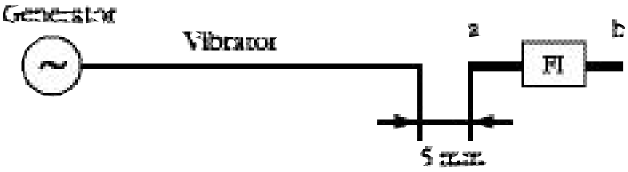

Fig.9
Field indicator shows "0" - "electrodynamic vacuum".

The same "vacuum" in the center of tornado was observed by people in the epicenter.

Remember what is measured with the field indicator (an ampere meter) along the linear vibrator. The Ampere force!

So, both people and their environment are in rotation process. The Earth rotates around its axis and around the Sun. The Sun rotates around the Galaxy center. The Galaxy rotates around the center of the Universe. The Universe rotates...

Every space point has an inceptive rotation impulse (torque). Let us remember the analogy. Water flowing out of the tub rotates counterclockwise in the northern hemisphere and clockwise in the southern hemisphere. If you set a right-handed torque in the northern hemisphere, the water will continue to flow counterclockwise.

The electromagnetism is the same. It has a natural left-handed rotation. Nevertheless, practical radiotechnology is familiar with right-handed fields artificially generated by transmitters.

The commonly used term of magnetic interactions is easier to understand if we consider the electromagnetic field as a vortex flow. Two left-handed homogeneous parallel flows are shown in Fig.1. At the point of interaction A of vector projection, the speeds of flow propagation on section plane **A-A** have opposite direction and compensate each other. The actual flow density is $\rho_1 < \rho_2 < \rho_3$. Reduced pressure occurs at point A. The flows are attracted. The resulting attraction vector for the whole flow is radially directed. The counter parallel flows will have composition, i.e. overpressure, and will repel.

One should note that graphic presentation in Fig.1 for vectors H and I for the etherdynamic systems is possible only in special cases. In dynamic vortex flows the current intensity vector is actually directed towards the flow movement and coincides with the vector direction of its propagation speed (V). The resulting vector H remains a radius-vector only for the considered point of interaction of two flows. The flow energy parameters are defined by its propagation speed V (Fig.1).

I am convinced of existence of ether vortex flows in the Universe. Modern astronomic and cosmological surveys confirm my hypotheses of aether Hyper-fields existence on a Universal scale.

The commonly used term of magnetic interactions is easier to understand if we consider the electromagnetic field as a vortex flow.

Even an amateur in aerodynamics and dynamic of gas vortex may make cosmological conclusions:

- vortex initiation implies the initial gradient flow of at least two counter flows;
- energy parameters of the emerging vortex flow can not exceed those of the parent flows
- there exist ether flows with energy exceeding that of our Universe
- our Universe has not originated from a point and is not finite. It is the spiral (vortex) structure of our Universe that testifies against the Big Bang hypothesis.

Conclusions:

Practical radiotechnology is familiar with the key difference between the field intensity distribution diagrams of aerial emitters in near and far zones. After combining diagrams 6d and 7b (Fig.10) it is possible to draw certain conclusions.

When transverse waves (radio waves) are formed around the aerial vibrator, the double transformation of the initial flow geometry starts:

- **the first flow is a homogeneous spiral flow on the conductor skin-layer;**
- **the second flow is a non-homogeneous vortex flow of the near zone;**
- **the third flow is a toroidal flow of the far zone.**

The toroid volume alteration in the far zone (as per the transmitter vibration frequency) forms volumetric planar transverse vibrations (spherical waves) that are called electromagnetic or radio waves. The gradient of potentials alterations (as per the transmitter vibration frequency) provoke changes of the Coulomb plane, which form longitudinal vibrations with a narrow directional diagram.

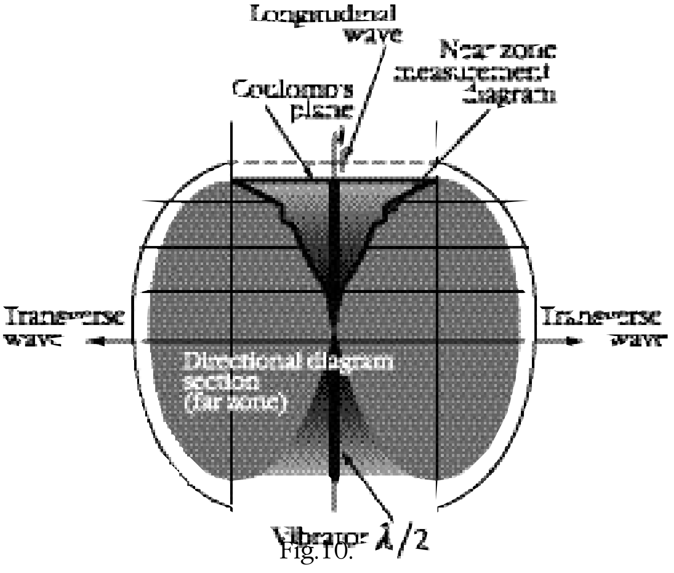

Fig.10.

The longitudinal flow is assumed to be formed as a chain of toroids (like smoke rings)

Ether-dynamic approach provides grounds for more substantial conclusions and assumptions. A new approach to energy and communication systems is made real. Further results of the experiments with transmitters and aerials, assembled on ether-dynamic principles, enabled to draw the following conclusions:

- all reasons, preventing transformation of the transmitter power into the power of transverse radio waves, are evident;
- all reasons for the transverse wave phase lagging with the increase of distance from the emitter, are also evident;
- besides the toroid that forms the transverse wave, the second flow generates a crucially new radiation that has a very narrow directional diagram along its longitudinal axis;
- the aerial ignores wave geometry for the second flow and its longitudinal derivative;
- the radiation along the second flow axis has a very high penetrating power;
- the radiation along the second flow axis has no signs of polarization.

The list of references is not shown in this article due to its size. It can be looked up at **www.efir.com.ua**.

NEWS

In October 2003 the second stage of the research on the "time control" was completed. In the photo below you can see Chernobrov V. A. and Frolov A. V. in the laboratory, October 31st 2003.

The second stage conclusions are quite interesting, however the usage of timers (chronometers) as detectors is proven to be inappropriate for they are subject to magnetization and their indication variations cannot be reliable in this case.

The next stage will feature the following method of detection of the time rate changes (the rate of the matter existence) that is the radiation wave-length measurement. If it is possible to achieve significant results then we will be able to detect the change of the laser ray colour in the area of the effect and its linear path deviations. The experiments will be held to register the weight changes of the detector at the expected influence of the produced effect on the ether density.

Our company looks for cooperation with corporations, which are interested in the application aspects of these technologies.

Frolov A. V.
General Director, Faraday Lab Ltd
7-812-3803844
http://www.faraday.ru

CELL OF THE THIN PLASMAELECTRIC HEAT GENERATOR

Ph. M. Kanarev, A.I. Tlishev, Russia
E-mail: kanphil@mail.kuban.ru

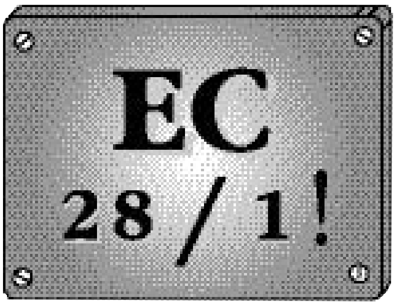

The main task of the experiment was to check the hypothesis: "Electrodynamic influence on the water molecules gives the possibility to significantly reduce energy expenses on destruction of their chemical bonds; further fu-sion of these molecules considerably increases the output of additional energy in the form of heat".

In order to solve this task, special experiments were carried out as regards the electrodynamic destruction of chemical bonds of water molecules by electric pulses of various frequencies.

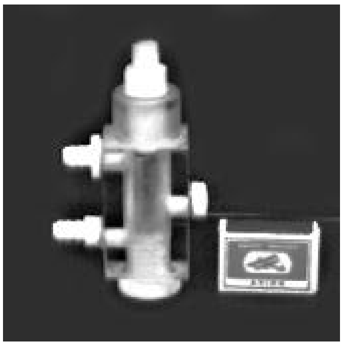

Photo of the Thin Plasmaelectric Heat Generator

CHECK TEST RECORD FOR PLASMAELECTRIC HEAT GENERATOR

Table 1

	Indices	1	2	3	Avg.
1	mass of the solution, which has passed through the reactor m, kg.	0.470	0.432	0.448	0.450
2	temperature of solution at the input of the reactor t_1, degrees	22	22	22	22
3	temperature of the solution at the output of the reactor t_2, degrees	66	66	65	65.67
4	temperature difference of the solution $\Delta t = t_2 - t_1$, degrees 44	44	43	43.67	
5	durability of the experiment $\Delta\tau$, s	300	300	300	300
6	reading of voltmeter V, V	4.50	4.50	4.50	4.50
6’	**Reading of oscillograph , U', V**	**4.47**	**4.47**	**4.47**	**4.47**
7	reading of ammeter I, A	2.1	2.1	2.1	2.1
7’	**Reading of oscillograph , I'**, A	**2.2**	**2.2**	**2.2**	**2.2**
8	electric power consumption according to indices of voltmeter and ammeters, $E_2 = I \cdot V \cdot \Delta\tau$, kJ	2.84	2.84	2.84	2.84
9	power the heated solution, $E_3 = 4.19 \cdot m \cdot \Delta\tau$, kJ	79.64	80.01	80.72	80.46
10	**reactor efficiency index K = E_3 / E_2**	**28.04**	**28.17**	**28.42**	**28.21**

Supply voltage and current were measured with the help of a voltmeter, an ammeter and an oscillograph (Fig. 1-4)

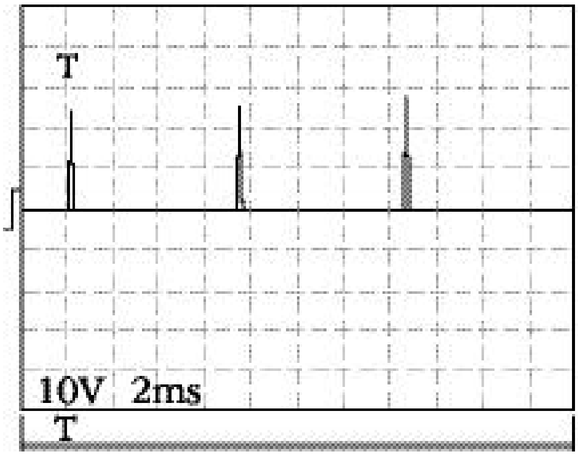

Fig. 1
Voltage

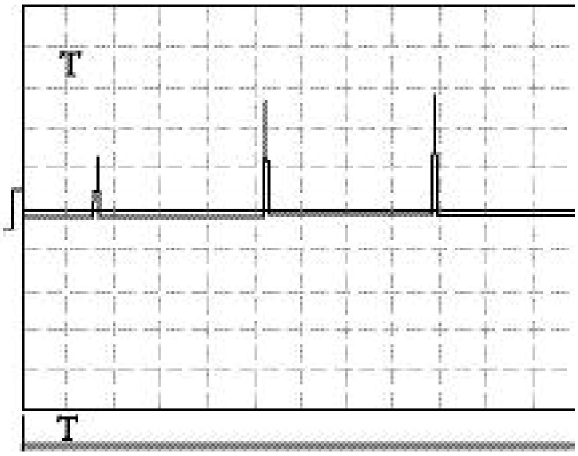

Fig. 2
Voltage

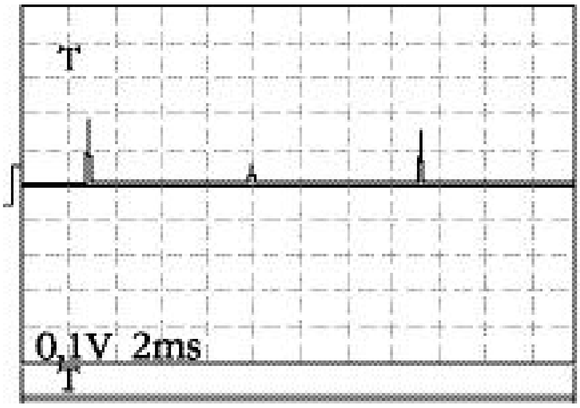

Fig. 3
Current

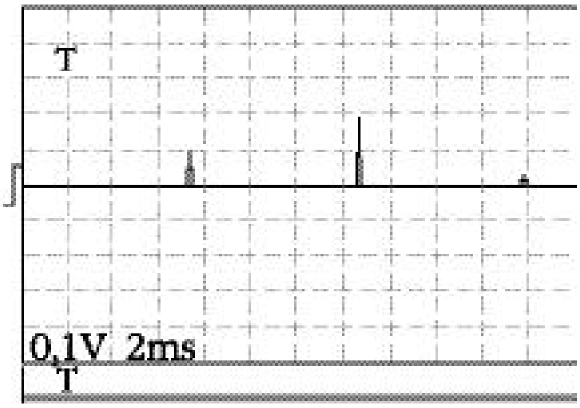

Fig. 4
Current

Process parameter calculation according to the oscillograms (Fig. 1-4) to the check test record (Table 1) gave the following results:

- Pulse scale 10.

- Average voltage amplitude according to Fig.1 and Fig. 2: Ua_{av} = (23+25+28+10+26+29) x 10/6 = 235 V.

- Average current amplitude according to Fig. 3 and Fig. 4: Ia_{av} = (20+6+17+7+10+19+3) x 10/7 = 117 A.

- Pulse repetition period T = 7.4 ms. Pulse duration t_p = 0.28 ms.

- Pulse frequency f = 1000/7.4 = 135.1 Hz. Relative pulse duration S = 7.4/0.28 = 26.32.

- Space factor Z = 0.5/26.32 = 0.019.

- Average pulse voltage U_{av} = 0.019 x 235 = 4.47 V.

- Average current in pulses I_{av} = 0.019 x 117 = 2.22 A.

Thus, it is possible to consider that the experimental test of energy efficiency of the water electric heat generator with the help of two methods gives practically identical results and confirms the above-mentioned hypothesis concerning the possibility to generate additional energy in the processes under consideration. It should be noted that as during measurements the pointer instruments of high class of accuracy of 0.2 have been used (relative conventional gauging error does not exceed 0.2%), and oscillographic measurement accuracy is much lower (usually, about 5%), the readings of the voltmeter and the ammeter should be considered more accurate.

Commercial efficiency of the water electric heat generator will depend on pulse generator economy. Since effi-ciency of powerful pulse generators can be near one unit, energy efficiency in industrial plants that use the considered heat generators should not differ greatly from the data obtained during laboratory research.

Simplicity and one hundred per cent reproducibility of the described experiments open a prospect for quick commercialization of the water electric heat generator.

REFERENCES

1. Kanarev Ph.M. The Foundation of Physchemistry of Microworld (the second edition). (In Russian) http://www.ikar.udm.ru/sb28-2.htm
2. Kanarev Ph.M. The Foundation of Physchemistry of Microworld. The second edition. (In English). http://book.physchemistry.innoplaza.net
3. Kanarev Ph.M. Water Electric Generator of Heat. http://Kanarev.innoplaza.net

Professor Kanarev looks for co-projects with prospective investors.

• Fantastic Projects •

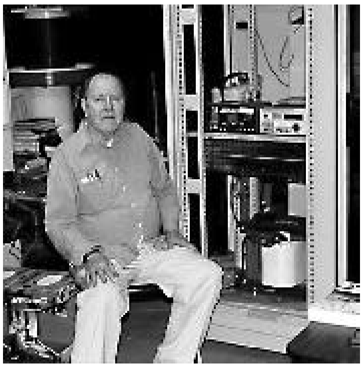

Smith's Generator

Review

TransWorld Energy,
8110 Bent Oak Lane Spring, Texas 77379, USA
Email: donsm1@earthlink.net
http://altenergy-pro.com

Don Smith is a University degreed professional in science & engineering. Served in three wars; WWII, Korea and Vietnam. He has enjoyed a long and successful profession as an executive in the petroleum industry. His life long outside interests include electronics and engineering. A popular walk on substitute for University Professors in physics, chemistry, biology and computer assisted drafting. Teaching style similar to Richard Feynman. He enjoys celebrity status in Japan and Saudi Arabia.

About 20 years ago, the book "Inventions, Researches and Writings of Nicola Tesla", resulted in his reconstructing, as a way of understanding the many theories and devices shown in the book. From this encounter, a very strong bonding resulted. The object then became putting his thinking inside that of Tesla, such that expanding upon and extrapolating areas not yet reached by Tesla. Technological advances in materials provide sourcing for devices not possible at earlier periods. For example, magnetic permeability which is the counter part of negative resistance is an open field for experimentation today.

Don does not sell his inventions, but trades licenses for shares in companies which then incorporate the new technology into their marketing systems. As such, he is on the Board of directors of several well established ventures located in Japan, Brazil, Mexico and Saudi Arabia.

Successful reproductions of Don's technology are present in Finland, Saint Petersburg - Russia, Yugoslavia, Romania, Japan, Hong Kong, and numerous other Countries. His book "Resonance Energy Methods" has a circulation of 40,000 copies in worldwide in many languages.

generators can be near one unit, energy efficiency in industrial plants that use the considered heat generators should not differ greatly from the data obtained during laboratory research.

Simplicity and one hundred per cent reproducibility of the described experiments open a prospect for quick commercialization of the water electric heat generator.

REFERENCES

1. Kanarev Ph.M. The Foundation of Physchemistry of Microworld (the second edition). (In Russian) http://www.ikar.udm.ru/sb28-2.htm
2. Kanarev Ph.M. The Foundation of Physchemistry of Microworld. The second edition. (In English). http://book.physchemistry.innoplaza.net
3. Kanarev Ph.M. Water Electric Generator of Heat. http://Kanarev.innoplaza.net

Professor Kanarev looks for co-projects with prospective investors.

• Fantastic Projects •

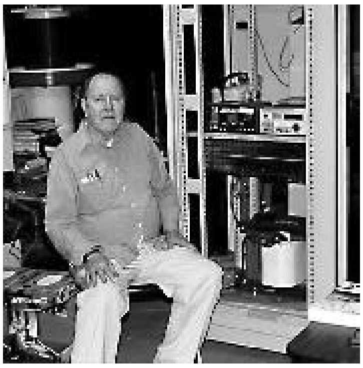

Smith's Generator

Review

TransWorld Energy,
8110 Bent Oak Lane Spring, Texas 77379, USA
Email: donsm1@earthlink.net
http://altenergy-pro.com

Don Smith is a University degreed professional in science & engineering. Served in three wars; WWII, Korea and Vietnam. He has enjoyed a long and successful profession as an executive in the petroleum industry. His life long outside interests include electronics and engineering. A popular walk on substitute for University Professors in physics, chemistry, biology and computer assisted drafting. Teaching style similar to Richard Feynman. He enjoys celebrity status in Japan and Saudi Arabia.

About 20 years ago, the book "Inventions, Researches and Writings of Nicola Tesla", resulted in his reconstructing, as a way of understanding the many theories and devices shown in the book. From this encounter, a very strong bonding resulted. The object then became putting his thinking inside that of Tesla, such that expanding upon and extrapolating areas not yet reached by Tesla. Technological advances in materials provide sourcing for devices not possible at earlier periods. For example, magnetic permeability which is the counter part of negative resistance is an open field for experimentation today.

Don does not sell his inventions, but trades licenses for shares in companies which then incorporate the new technology into their marketing systems. As such, he is on the Board of directors of several well established ventures located in Japan, Brazil, Mexico and Saudi Arabia.

Successful reproductions of Don's technology are present in Finland, Saint Petersburg - Russia, Yugoslavia, Romania, Japan, Hong Kong, and numerous other Countries. His book "Resonance Energy Methods" has a circulation of 40,000 copies in worldwide in many languages.

Abstract

Mr. Don Smith has discovered that Ohm's Law of Resistance does not apply to Magnetic Resonance that travels unrestricted for great distances. Therefore, multitudes of electrons are disturbed. The Magnetic moments are translated into usable electric energy. The Magnetic Dipole provides an unlimited source of electrical current. Mr. Smith's "Dipole Transformer" has been pending since last September.

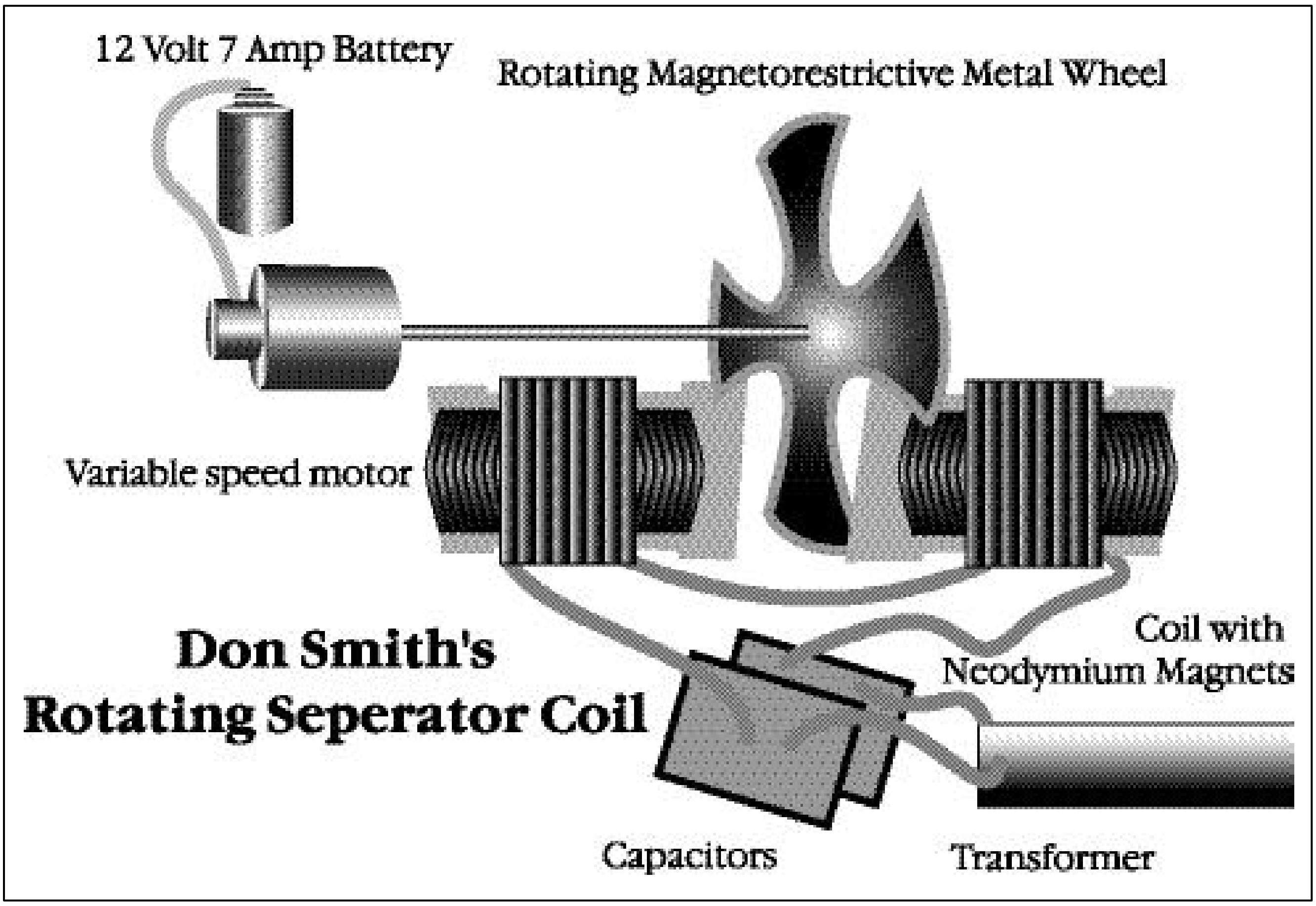

Fig.1
Smith's Energy Receiver

Useful energy occurs as the result of imbalances of ambient background energy and it is a transient phenomenon. In conventional circuits, the electrical field is in a closed system that is damped out with heat loss, which severely limits its utility. **The flip side of the electron generates magnetic waves, which is an open system, not subjected to heat loss. These being unrestricted are the universal source of energy. Universally present electrons are flipped to their electrical position, resulting in useful energy.** When properly constructed an electrical circuit can become self sustainable once started. An obvious corollary would be that the number of radio or television receivers in operation does not deplete the magnetic wave source providing the service. The key source of unlimited energy is Magnetic Resonance. To understand this requires putting a stake through outdated physics. An obvious example is the piano where the key impacts the one note giving one sound, which resonates with its two side keys providing a much higher level. Magnetic Resonant Energy clearly amplifies demonstrating more energy out than in. **Excess energy consequent to a reaction is called Free Energy and can be very large as is in the case of Magnetic Resonance.** The intentionally ignorant physicist makes a habit of ignoring this fact.

Useful electricity is obtained from disturbed electrons, which radiate magnetic fields and waves. The generator disturbs and collects this

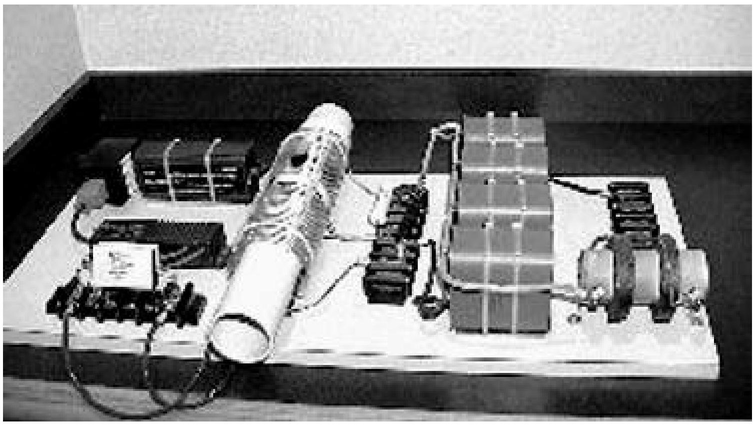

Fig.2
D. Smith's Energy Receiver

emitted electron energy. These electrons remain fully in tact and undiminished until the end of time. This Energy is free and the cost of accessing it depends on the ignorance of the collector. With an appropriate collector system an unending source of environmentally benign, inexpensive, available energy everywhere, is provided.

Editor: in September 2003 Donald Smith took part in the annual "Inventor's Week-end" Conference where he presented the report on his invention. In this issue you can find more details on the conference.

It is worth of a note that we have received ***a number of negative comments on this inventor's work****. However, we thought his articles to be quite interesting and decided to publish them. We would be glad to get readers' comments, who are familiar with the work and inventions of Donald Smith.*

NEWS

News from Farady Lab

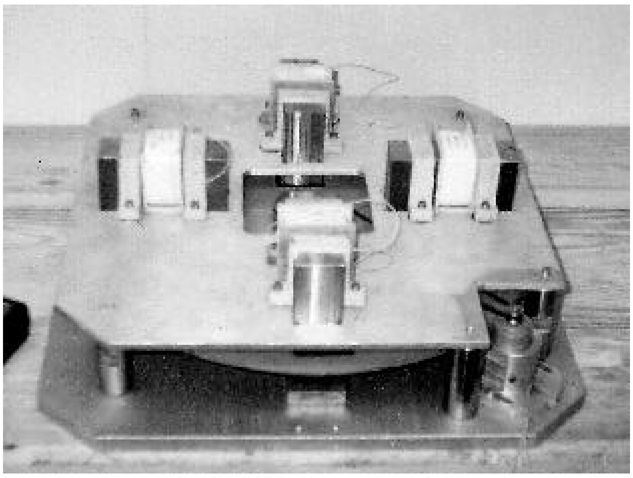

Disk permanent-magnet generator
Faraday Lab Ltd

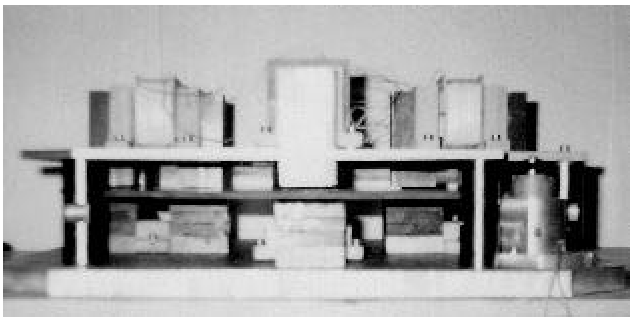

In our lab we are still working at the optimization of the alternator design (electromagnetic generator with the permanent magnets). We have introduced the drum rotor instead of the disk rotor into the new design of the alternator. The operation mechanism is the same, namely, the alternation of the magnetic flux near the generator coils occurs as a result of the rotating elements operation of the rotor, which "shunt" the magnetic flux.

Frolov A.V.

Strategies for Launching a New Over-Unity Device

Ryan S. Wood, USA
14004 Quali Ridge Drive, Broomfield, CO 80020

Problem description: With any singly successful over-unity energy device comes a host of business, technical and financial problems and opportunities. A public announcement permanently will lead to a "Pons & Fleishman" effect that at best delays public success and at worst derails the entire technological effort. Careful attention needs to be applied to the marketing launch, replication, patents, legal defense and finances.

Potential Solution: With a few working devices that generate100-500 horsepower it is relatively easy to hook these units to the electrical gird and generate monthly checks from the utilities. The law states that major utilities must purchase electricity from any supplier at "cost avoidance" rate now typically, between 6 and 10 cents per kilowatt-hour. This law, California Public Utilities Commission - Rule 21, for example, is designed to help small hydro-electric producers and solar or wind farms. There is a safety requirement of grid transfer switches, but they are not terribly complicated or expensive. The entire equipment set-up must be approved by the utility, however, if you start with a diesel electric generator or small hydro plant it can be approved, operating and stable before the clandestine switch or hybrid operation with an over-unity machine. The law is strongly supportive of the energy producer and disconnection for non-safety or maintenance reasons is exceeding rare.

A key benefit of this approach is that you have established repeatability by having multiple working sites, you have established that it can generate money and the technology does real work and you have created several independent locations where press and scientists can visit to validate the over-unity effectiveness in-situ.

Typical revenues would be as follows:

Generator Horse Power	100	200	300	400	500
Annual $, 90% uptime @ 6 c/kWh	$ 35,100	$ 70,199	$ 105,299	$ 140, 398	$ 175, 498
Monthly payment @ 6 c/kWh	$ 2,92	$ 5,850	$ 8,775	$ 11,700	$ 14, 625
Annual $, 90% uptime @ 10 c/kWh	$ 58,499	$ 116, 999	$ 175, 498	$ 233,997	$ 292, 496
Monthly payment @ 10 c/kWh	$ 4,875	$ 9,750	$14, 625	$ 19,500	$ 24, 375

It will take money without the strings of investors to champion and defend a new breakthrough technology into the energy marketplace and this strategy offers a quiet way to prepare for the ultimate launch and resulting scientific and media frenzy that will ensue. It will take years for the business to form, products to be built and marketing and acceptance and industry confidence to be established. All during those years this strategy can generate cash to support engineering and operations.

Please call if there are questions at 303-941-9663 (cell),
Home: 720-887-8071
Ryan S. Wood

The Hutchison Effect – An Explanation

Mark A. Solis, USA
http://www.geocities.com/ResearchTriangle/
Email: your_neighbor_2000@yahoo.com

People often ask, "What exactly is the Hutchison Effect?" This brief essay is an attempt to answer that question to the satisfaction of the majority. First of all, the Hutchison Effect is a collection of phenomena which were discovered accidentally by John Hutchison during attempts to study the longitudinal waves of Tesla back in 1979. In other words, the Hutchison Effect is not simply a singular effect. It is many.

The Hutchison Effect occurs as the result of radio wave interferences in a zone of spatial volume encompassed by high voltage sources, usually a Van de Graff generator, and two or more Tesla coils.

Editor: We can call this effect "The H. Wells Effect" because he was first to describe it in his book "The Invisible man". The main character of the book used two sources of electro-magnetic vibrations and Wells said that they were not Hertz waves but something else. "But the essential phase was to place the transparent object whose refractive index was to be lowered between two radiating centers of a sort of ethereal vibration..."("The Invisible Man", chapter 20).

A. V. Frolov

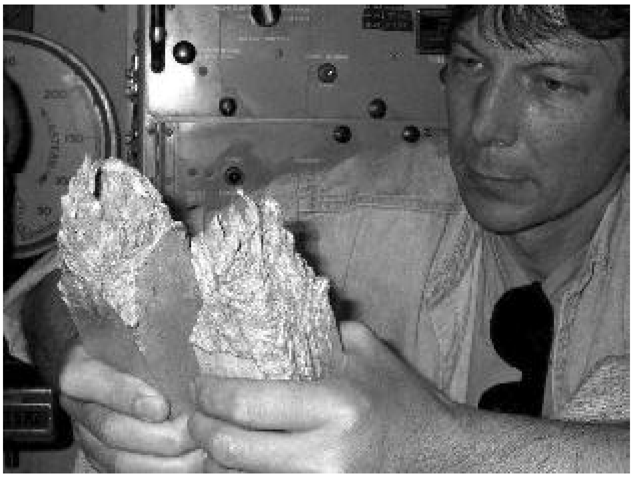

The effects produced include levitation of heavy objects, fusion of dissimilar materials such as metal and wood, the anomalous heating of metals without burning adjacent material, spontaneous fracturing of metals (which separate by sliding in a sideways fashion), and both temporary and permanent changes in the crystalline structure and physical properties of metals.

Fig.1 represents the proof of Hutchison Effect based on metal blocks. Deformation occurred at room temperature as a result of complex interaction of electromagnetic fields.

The levitation of heavy objects by the Hutchison Effect is not - repeat not - the result of simple electrostatic or electromagnetic levitation. Claims that these forces alone can explain the phenomenon are patently ridiculous, and easily disproved by merely trying to use such methods to duplicate what

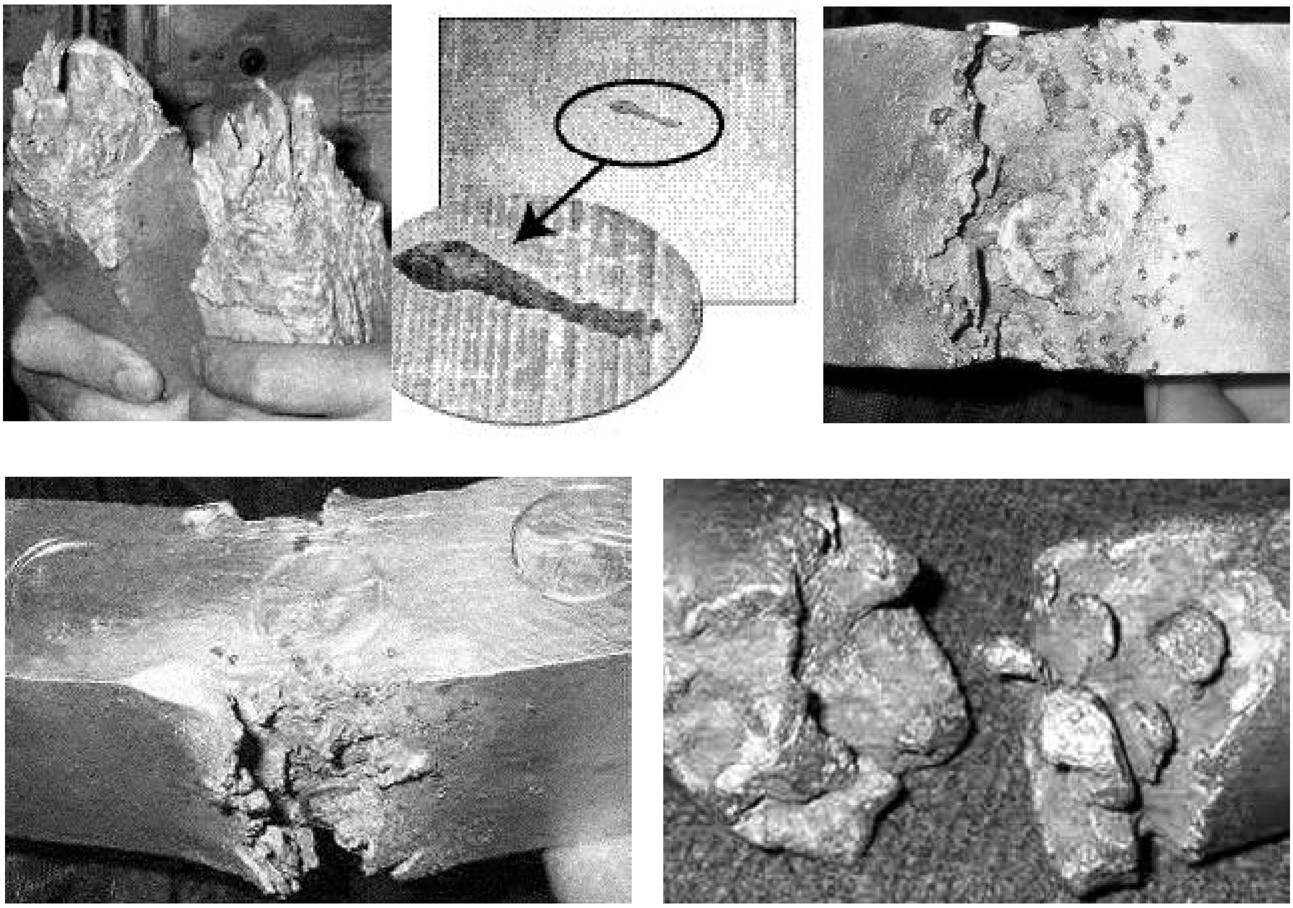

Fig.1

Cold melting of metals

The upper left: steel, The lower left: aluminium with coin prints and with a coin inserted in a partially open split; The lower right: fully torn aluminium bar; The middle upper and right: a section part of the aluminium block, into which a wood block is fused (brown matter)

http://www.rumormillnews.com/JOHN-HUTSHISON2.htm

the Hutchison Effect has achieved, which has been well documented both on film and videotape, and has been witnessed many times by numerous credentialed scientists and engineers. Challengers should note that their apparatus must be limited to the use of 75 Watts of power from a 120 Volt AC outlet, as that is all that is used by Hutchison's apparatus to levitate a 60-pound cannon ball.

The fusion of dissimilar materials, which is exceedingly remarkable, indicates clearly that the Hutchison Effect has a powerful influence on Van der Waals forces. In a striking and baffling contradiction, dissimilar substances can simply "come together," yet the individual substances do not dissociate. A block of wood can simply "sink into" a metal bar, yet neither the metal bar nor the block of wood come apart. Also, there is no evidence of displacement, such as would occur if, for example, one were to sink a stone into a bowl of water.

The anomalous heating of metal without any evidence of burning or scorching of the adjacent materials (usually wood) is a clear indication that possibly the nature of heat may not be completely understood. This has far-reaching implications for thermodynamics, which hinges entirely on the presumption of such knowledge. It should be noted that the entirety of thermodynamics is represented by the infrared portion of the electromagnetic spectrum, which is insignificant in a context of 0 Hz to infinite Hz. The anomalous heating exhibited by the Hutchison Effect shows plainly that we have much to learn, especially where thermodynamics and electromagnetism meet.

The spontaneous fracturing of metals, as occurs with the Hutchison Effect, is unique for two reasons: (1) there is no evidence of an "external force" causing the fracturing, and (2) the method by which the metal separates involves a sliding motion in a sideways direction, horizontally. The metal simply comes apart.

Doctor John Hutchison

Some temporary changes in the crystalline structure and physical properties of metals are somewhat reminiscent of the "spoon bending" of Uri Geller, except that there is no one near the metal samples when the changes take place. One video shows a spoon flapping up and down like a limp rag in a stiff breeze. In the case of permanent changes, a metal bar will be hard at one end, like steel, and soft at the other end, like powdered lead. Again, this is evidence of strong influence on Van der Waals forces.

The radio wave interferences involved in producing these effects are produced from as many as four and five different radio sources, all operating at low power. However, the zone in which the interferences take place is stressed by hundreds of kilovolts.

It is surmised by some researchers that what Hutchison has done is tap into the Zero Point Energy. This energy gets its name from the fact that it is evidenced by oscillations at zero degrees Kelvin, where supposedly all activity in an atom ceases. The energy is associated with the spontaneous emission and annihilation of electrons and positrons coming from what is called "the quantum vacuum." The density of the energy contained in the quantum vacuum is estimated by some at ten to the thirteenth Joules per cubic centimeter, which is reportedly sufficient to boil off the Earth's oceans in a matter of moments.

Given access to such energies, it is small wonder that the Hutchison Effect produces such bizarre phenomena. At the present time, the phenomena are difficult to reproduce with any regularity. The focus for the future, then, is first to increase the frequency of occurrence of the effects, then to achieve some degree of precision in their control.

> *The spontaneous fracturing of metals, as occurs with the Hutchison Effect, is unique for two reasons: there is no evidence of an "external force" causing the fracturing, and the method by which the metal separates involves a sliding motion in a sideways direction, horizontally. The metal simply comes apart.*

Editor: On November 17th 2003 we got to know that the inventor John Hutchison can eventually lose his home laboratory. The property can be confiscated upon court order due to neighbours' complaints. Apparently, it is dangerous to conduct experiments of the kind at home and even within a town. We wish him success!

A. V. Frolov

A SIMPLE DEVICE FOR HEAT TO ELECTRICAL ENERGY CONVERSION BY MEANS OF FERROELECTRICS

A. Ries, A. Z. Simoes, M. A. Zaghete, J. A. Varela,
Institute of Chemistry, Universidade Estadual Paulista, Rua Prof. Francisco Degni s/n,
14801-970 Araraquara-SP, Brazil
Email: ries@posgrad.iq.unesp.br

E. Longo,
Department of Chemistry, Universidade Federal de Sao Carlos, Rodovia Washington Luis - km 235, 13565-905
Sao Carlos-SP, Brazil

Introduction

A few years ago, Zaev [1, 2] demonstrated experimentally the possibility of heat to electrical energy conversion by means of ferroelectrics. A capacitor filled with a ferroelectric material was charged and discharged periodically at a temperature just a little lower than the Curie point. According to Zaev, the capacitor must have a nonlinear capacitance $dC/dV > 0$, which means the capacitance increases with rising voltage. Besides these experimental results, it was proven by a theoretical calculation that the electrical energy obtained during discharging can reach up to 1.35 of the energy introduced to charge the capacitor. The present paper presents an inexpensive and simple electric circuit which can be used to prove heat to electricity conversion experimentally. Moreover, the physical mechanism of heat to electricity conversion is explained in the form of a thermodynamic cycle.

Thermodynamic cycle for energy conversion

The following thermodynamic cycle is composed of four reversible steps. It works only at a temperature just a little lower than the Curie point, where the dielectric constant is highly temperature dependent. Fig. 1 shows the relative dielectric constant of a representative ferroelectric ceramic as a function of temperature.

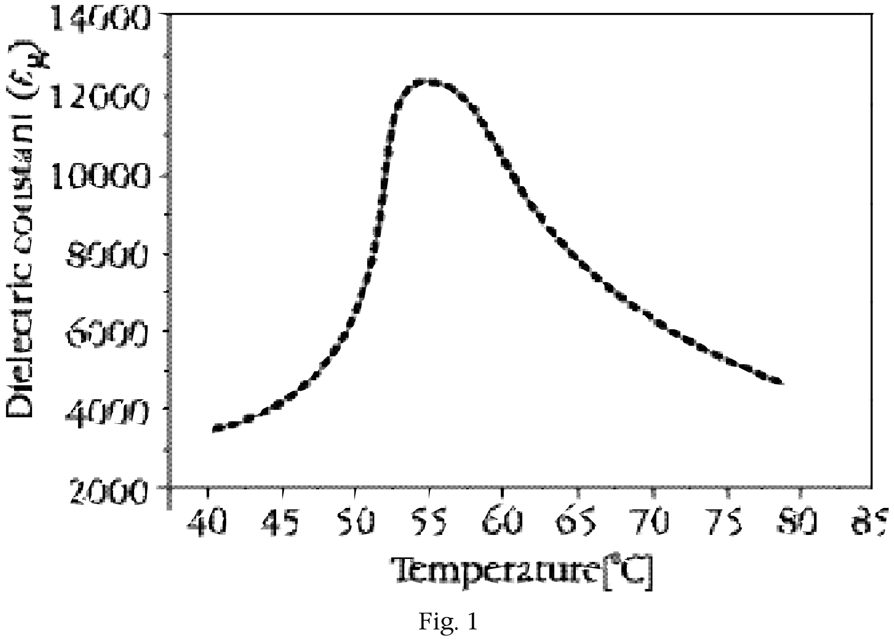

Fig. 1
Relative dielectric constant versus temperature for a barium strontium titanate ceramic (80% Ba, 20% Sr) prepared by the author. Curie point = 55° C.

As it can be seen, just below the Curie point in the range from 50 to 55^0C, small changes of the temperature cause high changes of the dielectric constant.

Furthermore, to understand the conversion mechanism, it is necessary to understand the electrocaloric effect. When an electric field is applied to a dielectric medium, the latter is polarized. For ferroelectric materials, the parallel domain alignment is the most important part of the total polarization. During the successive alignment of the domains, some material parameters in particular entropy, heat capacity and temperature change. Since the material possesses less degrees of freedom in the polarized state, its entropy and heat capacity are reduced. As a consequence, provided that no heat exchange with the environment is possible (adiabatic conditions), the temperature increases. It is very important to understand that the energy for warming of the dielectric is not taken from the polarizing electric field. The increase in temperature is a result of the decreased heat capacity only.

From this, it can be derived easily what means "nonlinear capacitance".

When a capacitor is charged adiabatically which means its voltage increases faster than any heat exchange with the environment is possible, due to the electrocaloric effect, the temperature of the dielectric must also increase. As it can be seen from Fig.1, below the Curie point, an increasing temperature leads to a higher dielectric constant. Assuming that both temperatures, the one before charging and the other after charging are still below the Curie point, one can say that the dielectric constant increases with an increase in voltage.

Since the capacitance is proportional to the dielectric constant, the capacitor has a nonlinear characteristic $dC/dV > 0$. In the same way it can be concluded that adiabatic charging just a little above the Curie point reveals a nonlinear capacitor characteristic with $dC/dV < 0$.

Quantitative experimental measurements on the electrocaloric effect in some representative ferroelectric and antiferroelectric materials were published by Thacher [3].

STEP 1: Adiabatic charging of the capacitor

We consider a capacitor filled with a ferroelectric medium at an environmental temperature a little lower than the Curie point.

An adiabatic charging of a capacitor (no heat exchange) effects an increase in temperature and capacitance due to the electrocaloric effect. We assume that this increase in temperature is so small that the temperature remains below the Curie point. Finally, the capacitor is completely charged while the dielectric medium has a higher dielectric constant and therefore a higher capacitance than it posseses at the environmental temperature.

To go further to step 2, the capacitor is hold under conditions where no discharge is possible, e.g. it is disconnected from the power supply.

STEP 2: Thermal equilibration

After a short time, the charged capacitor has cooled down to the environmental temperature. While cooling down, the dielectric constant and capacitance decrease. But the charges of the capacitor plates remain constant. As a consequence, the voltage and the electrical energy increase. This can be readily recognized from the general capacitor equation $q = C V$ (q = charge, C = capacitance, V = voltage):

$$q = \text{constant} = C_1 V_1 = C_2 V_2$$

STEP 3: Adiabatic discharging of the capacitor

While discharging the capacitor the temperature and capacitance decrease (again heat capacity changes) due to the electrocaloric effect, leading to a further increase in the available electrical energy. The ferroelectric medium posseses now a final temperature below the temperature of the environment.

STEP 4: Thermal equilibration

In order to proceed with step 1, it is necessary to reach the environmental temperature. An inflow of heat from the environment to the dielectric material is required.

Electric circuit to observe energy conversion

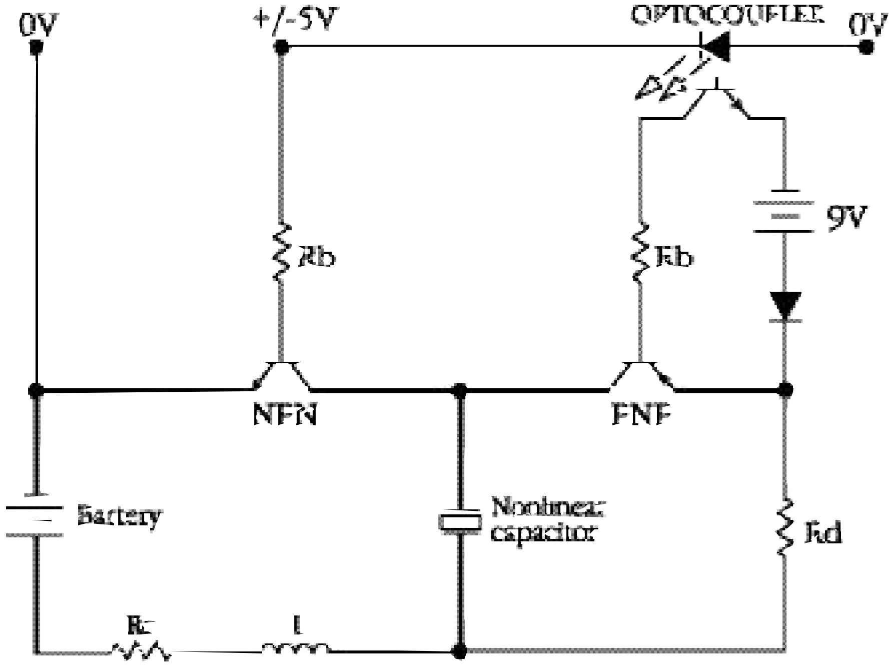

Fig. 2
Electric circuit for testing the energy conversion effect. The connections to the square wave generator are labeled with 0V and +/- 5V, the resistors Rb are to adjust the maximum base current of the transistors.

Assuming the experimenter has a nonlinear capacitor, the set-up shown schematically in Fig. 2 is proposed as a possible conversion device. The circuit contains two different types of transistors (NPN and PNP) which fulfill here the function of a switch.

The NPN junction transistor consists of two n-type semiconductors (called the emitter and collector) separated by a thin layer of p-type semiconductor (called the base). On the other hand the PNP junction transistor consists of a thin layer of n-type semiconductor lying between two p-type semiconductors. The base is the ON/OFF switch for the transistors. If a current flows to the base, there is a path from the collector to the emitter, where a current can flow (switch is ON). If there is no current flowing to the base, then no current can flow from the collector to the emitter (switch is OFF).

The charging-discharging process is controlled by a square wave generator which produces an output signal switching from +5V to -5V. The switching frequency is adjustable from 0.1 Hz up to at least 10 kHz. This output signal is applied to the bases of both transistors.

During the time the square wave generator output signal is +5V, a base current flows through the base-emitter junction of the NPN transistor which creates a low resistance path between the collector and the emitter. As a consequence, the capacitor is charged by the battery. The collector-emitter junction of the PNP transistor has a very high resistance, because here no base current flows. Therefore the capacitor is not discharged at the same time. When the output signal switches to -5V, a current flows to the base of the PNP transistor and not to the base of the

NPN transistor. Now the capacitor is discharged.

In order to understand the function of the inductance in Fig. 2, we analyze the thermodynamics of the charging process.

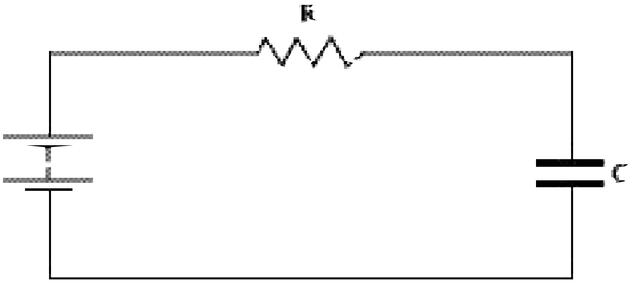

Fig. 3
Simple circuit for charging a capacitor

If a linear capacitor of capacitance C is charged to a voltage V through the load R according to Fig. 3, the energy $W = 0.5\,C V^2$ is stored in the capacitor. During the charging process, a time-dependent current I(t) flows:

$$I(t) = V/R \exp(-t/RC)$$

This current develops the heat Q in the load:

$$2Q = R \int I^2(t)dt = 0{,}5CV^2$$

Here the integration limits are from zero to infinity. This calculation shows, that the energy taken from the battery is the energy stored in the capacitor after charging plus the heat produced in the load, also 2•C V2. One can see that the heat Q is not dependent on the resistance value R.

The above used equations are only valid in case of a one-step charging process.

If the capacitor is charged stepwise to the final voltage V, e.g. using a ramp generator, and each voltage step effects a voltage increase , the total heat produced in the load is (N = number of voltage steps):

$$Q = N\,0{,}5\,C\,(V/N)^2 = 0{,}5\,C/N\,V^2$$

Now the energy taken from the source is:

$$W = 0{,}5\cdot(1 + 1/N)1/2\,CV^2$$

If the number of steps N tends to infinity, the heat Q tends to zero.

This energy loss in the load dramatically influences the efficiency of any conversion device and must be taken into account, otherwise all experimental attempts to observe energy conversion would fail.

A more detailed discussion on charging a capacitor and the unavoidable energy losses was given by Heinrich [4] and Gupta et al. [5].

The inductance in the circuit (Fig. 2) effects a slower rising of the current during the charging process (due to self induction) and can reduce the heat generated in the load dramatically.

Acknowledgments

Financial support of this work was provided by the Brazilian agency FAPESP (www.fapesp.br) and the German Academic Exchange Service (DAAD), (www.daad.de).

References

1. N.E. Zaev, Yu.S. Spiridonov, Capacitance as a converter of heat to electrical energy, Russian Electrical Engineering, Vol. 69, (2000) pp. 69-72
2. N.E. Zaev, Research on the Capacitance Converter of Environmental Heat to Electric Power, New Energy Technologies, issue 2, (2001)
3. P.D. Thacher, Electrocaloric effects in some ferroelectric and antiferroelectric Pb(Zr, Ti)O3 compounds, Journal of Applied Physics, Vol. 39, (1963), pp. 1996-2002
4. F. Heinrich, Entropy change when charging a capacitor: A demonstration experiment, American Journal of Physics, Vol. 54, (1986), pp. 742-744
5. V.K. Gupta, G. Shanker, N.K. Sharma, Reversibility and step processes: An experiment for the undergraduate laboratory, American Journal of Physics, Vol. 52, (1984), pp. 945-947

The Gates Motor

A. Akau, USA
Email: adrianakau@aol.com

The Gates motor company has produced a motor that does not run on fossil fuel or on electricity. It comes in a 28 HP and a 5,000 HP version. When linked to an electric generator, the larger motor will produce 2,200 kW.

My name is Adrian Akau. I am a recently retired school teacher, having taught science and math in the public school system, state of Hawaii for the past 36 years. One of the classes I taught was on energy.

Our world should be reaching the half-way mark in easily accessible oil by about the year 2016. Since the rate of oil usage is becoming higher each year, it is unlikely that the oil supply will last as long as the 140 years since the first well was drilled in Pennsylvania. We will become more dependent upon OPEC with all the political pressures and problems that accompany this type of dependency. As the easy oil is consumed, we must look for ways to extract less accessible oil such as from shale. The added expense will be passed on to the consumer.

The burning of coal is not the answer either because coal cannot supply convenient energy as does oil. An oil-based economy cannot transfer over readily to natural gas; there is not enough natural gas available and it is again, not a convenient and concentrated source as is oil. We also have serious pollution problems with the burning of fossil fuels (natural gas being the exception) but the burning of all fossil fuels have contributed to the serious problem of global warming.

It may take all of 50-80 years to transfer to a non-fossil fuel economy. The transfer must be made slowly so as not to disturb the infrastructure of our economy which is presently based upon fossil fuels. We must gradually create a new infrastructure that is non-fossil fuel dependent.

> *The motor incorporates a unique spring configuration to provide rotation*

The production model of the motor is 24 inches high by 48 inches in length. Connected to a generator it will provide sufficient power for about eight hundred US homes. It is designed to be tough and to last a long time. It uses a high grade airplane synthetic engine oil to prevent friction and the build up of heat. Only the oil filter needs to be changed. The motor has special needle bearings and portions of the motor are made of high quality steel.

The Gates motor is designed to run 24 hours/day, 365 days/year without incurring any fuel costs. The larger size motor would come in pairs; the spare motor being used as a back-up unit should the regular motor be turned off for maintenance such as to change the oil or in case there should be any problems with the first motor. The company is willing to enter into a contract for maintaining the motor after installation.

The Gates motor represents a major breakthrough in motor design technology and provides practical solutions to environmental and fuel conservation concerns. **Finally, a completely mechanical motor has been developed.** It can deliver the power needed in a wide variety of service applications without any waste products or emissions to adversely affect the environment. This revolutionary new motor utilizes spring power technology. Unlike conventional motors which must reach a maximum rpm level before the desired horsepower, the Gates motor provides maximum horsepower instantly by virtue of the torque stored in the springs.

The basic operating principle of the Gates motor involves a series of springs configured to provide the required motor rotation and power delivery. Consider, if you will, the operating principle of a grandfather clock. Once the clock is wound, it continues to operate until it runs down and stops. The springs must then be rewound so it will operate again. Suppose the clock is continually being rewound as it operates. Then the clock would run continuously. This condition is what forms the basis for the operating principle of the Gates motor. The motor incorporates a unique spring configuration to provide rotation. The springs inside the motor are wound and preset at the factory at the time the motor is assembled. The amount of torque wound into the springs at the factory determines the horsepower of the motor. Inside the motor are a multiple of springs positioned horizontally in a circular arrangement. As the motor operates, the springs are unwound a set number of degrees. A double ratchet system at one end of the motor rewinds the degrees used back into the springs. As the springs are unwinding, **the power generated from all the springs reset each spring (one at a time) during each revolution of the motor**. It is this reset action that produces and delivers power from the springs to the flywheel located at the opposite end of the motor. The flywheel serves to ensure the smooth operation of the motor and to convert the springs into useable horsepower. The smaller motor has 50 foot pounds of torque at a shaft speed of 3000 RPM and will produce 28 HP. The HP can be altered by factory adjustment. The speed of the engine is controlled by a hydraulic pump which provides pressure upon the power shaft.

Contacts: Gates Motor Corporation,
P.O. Box 715, Hauula, HI, USA 96717
Email:gagates@earthlink.net

Editor: We tried to get the video of this device operation from Gates Motor Co., but after a long consideration it was not provided since the authors are competitor-conscious.
Alexander V. Frolov

Cold Electricity

Adrian Akau, USA
Email: adrianakau@aol.com

Dr. Peter Lindemann in his book "The Free Energy Secrets of Cold Electricity" presents the story of this type of electricity which really is not electricity as electricity is normally meant to be understood. Cold electricity behaves differently than normal electricity. Regular electricity may be used in the process to generate this cold form but this transformation is not the type of change in voltage and current that occurs in a transformer. Rather, it is the extraction of a form of current from "normal" electricity by a process that uses high voltage.

In discussing the Edwin Gray motor, Dr. Lindemann relates a demonstration by Gray. Gray used a 6 volt car battery with lead wires running to a system he had devised using a voltage booster and a series of capacitors which raised the voltage to 3,000 volts. He then closed a switch that ran the high voltage into two electromagnets which caused a loud popping sound and cause the top electromagnet weighing a pound and a quarter to be propelled over two feet into the air. Gray claimed that only 1% of the system's energy was used with 99% going back to the battery. He said that he had 'split the positive' with his system.

Other evidence of using this fractionated electricity was Gray's use of a small motorcycle battery (15 amps which would normally produce W=V x A = 6 V x 15 A= 90 watts) to run concurrently six 15-watt electrical light bulbs, a portable 110 volt T.V. set and two radios. A glowing 40 watt light bulb running off the system was dropped into water without the glass breaking; the bulb gave off light but not the heat that accompanied a bulb run by regular electricity. This would mean that tungsten in the filament of the light bulb was not resisting the flow of this fractionated electricity and that this "cold electricity" was causing the luminous glow by some other means.

According to Gray's Patent #4,595,975, low voltage was chopped into pulsating DC by passing it through a multi-vibrator (buzzer like a door bell). Then the pulsed DC was sent through the low voltage winding (primary) of a transformer which changed it to pulsating high voltage DC current at the secondary side. The pulsating high voltage DC was rectified by a full wave bridge and changed into high voltage DC.

The high voltage DC was used to repeatedly charge a capacitor as the current was sent to briefly discharge across a spark gap (rated at 3000 volts). The discharge of the spark across the gap had to be in one direction only and its duration was controlled by the size of the capacitor and by the strength of a magnetic field encompassing the gap. This magnetic field had a quenching effect because it caused a back EMF each time the discharge occurred. The current from the discharge across the gap was then sent through a resistor and then to a vacuum tube (the conversion switching element tube).

Another of Edward Gray's patents "Efficient Electrical Conversion Switching Tube Suitable for Inductive Loads" (April 1987, patent 4,661,747) described the switching tube. It utilizes a low voltage anode (positive plate), a low voltage anode (positive plate) and one or more electro-static or charge receiving grids (located between the positive plates and the cathode or negative plate supplying the electrons). This tube's function was to 'split the positive'. A normal vacuum tube usually uses only a single anode at a particular voltage to control the voltage flow within the tube. The function of the charge-receiving grids was to collect the "cold electricity".

The most unusual part of this Power Supply circuit consisted of devices specifically designed to remove excess energy when the Switching Tube was functioning. A spark-gap protection device, served to "protect the inductive load and the rectifier elements from unduly large discharge currents". As stated, the rectifier elements are the two anodes and the grids in the Switching Tube. In a normal vacuum tube, no protection would be necessary because the maximum power of the tube would be based upon the electricity provided to heat the tube filament and to charge the anode (voltage x amps). Therefore, the excess energy comes from the process occurring as the spark gap affected electrical pulses pass through the Switching Tube. The process of extracting the "cold electricity" places danger of overload upon the system. (Please take note that the "electricity" from the Switching Tube is no longer the normal type of electricity we have from our batteries or from an electrical outlet but rather a completely new form with entirely different characteristics.)

The description continues, "Should the potentials (voltages) within the circuit exceed the predetermined values fixed by the mechanical size and spacing of the elements within the switching tube, the energy is dissipated (bypassed) by the protective device to the circuit common (electrical ground)" by two strategically placed diodes (devices that permit the passage of electricity in one direction only). The question now arises "What is the cause of this large amount of excess energy which must be dumped to ground through device acting like a lightning rod in order to prevent the circuit from burning out?"

To examine the cause of the excess energy source, we must go back over a hundred years to 1889. Heinrich Hertz had just announced in 1887 that he had discovered electromagnetic waves and Nicola Tesla was attempting to duplicate Hertz's experiments. Tesla used abrupt and powerful electric discharges produced with banks of capacitors charged to very high potentials and was able to explode thin wire (copper bus bars). He came to the conclusion that Hertz had mistakenly associated electrostatic inductions (electrified shockwaves in air) for true electromagnetic waves.

The explosion of the copper bars by means of the "disruptive discharges" from the capacitor bank produced sharp shockwaves which struck Tesla with great force across the entire front of his body. Tesla said they felt more like powerful gunshots rather than electrical sparks. They produced effects similar to lightning or to those produced by high voltage DC generators previously discussed; the simple closing of a high voltage DC generator caused a stinging shock. AC generators were not in use in this pre-AC era and it was shown, later on, that AC generators did not produce these effects.

This shock was first assumed to be the result of residual static charging. It stood straight out of highly electrified conductors, seeking ground paths which included workmen and switchboard operators. In long cables, Tesla estimated that this electrostatic concentration was several orders in magnitude greater than any voltage the DC generator could actually produce. It caused a hedge or crown of bluish needles or spicules to point at right angles to the cable or straight from the electrical cable line into the surrounding space. The bluish needles appeared the very instant the switch was closed and disappeared a few milliseconds later, after which the system functioned normally. However, anyone through whom the blue needles passed, especially in large regional power systems which used high exceptionally high voltages, usually did not survive. Generators rated at a few thousand volts produced hundreds of thousands or even millions of electrostatic volts during the start-up pulse. Highly insulated, heavily grounded relay switches had to be installed to protect workers from certain death.

Engineers at this time assumed **the effect to be caused by a "bunching" action which occurred when a powerful force was not able to move charge sufficiently quickly through a system.** (A similar "choking effect" also happened with large steam engines; if steam was introduced too rapidly, the steam engine could explode.) The metal of which the wire was composed somehow offered resistance to the charge carriers before they were able to move away from the generator terminals. It seemed that the wire acted as a blocking force to the electrons or to some unknown part of the electrons for the first few milliseconds, rather than as a conductor. The powerful, deadly bluish spikes sprang from the line until the current charges actually "caught up" with the applied electrical field. It was as if the spicules were a form of relieving pressure on the system by converting the voltage impulse into some other form.

Tesla began wondering why the electrostatic field moved faster than the actual charges. He believed that this effect could help him find electric waves better than his capacitors because the line resistance caused the electrostatic charges to bunch together into a density much greater that obtainable with his capacitors.

Tesla understood that ordinary capacitor discharges were oscillating currents or spark currents which "bounced" between each capacitor plate until their stored energy was dissipated. The high surge voltage of the DC generator exerted such a high one way pressure on the dense line charges, due to the millisecond resistance of the metal, that possible voltage "bounces" were also oscillations of current as was true with the capacitors. Tesla used every means possible to block the "back-rush" (bouncing) of the oscillating current in order to prevent the supercharge to prematurely decline; he wanted to maintain this powerful "bunching" effect as long as possible in order to study it and to make use of it.

His face and hands experienced a penetrating shockwave; a sharp pressure and an electrical irritation were brought on at the sudden closing of the switch. His face and hands were especially sensitive to these shockwaves which caused a "stinging" effect at close range. Tesla believed that he was being struck by material particles approaching the vapor state as they were thrust from bluish crowned wire.

Later on, he was to discover they were not gas particles. **He placed himself behind a glass shield but to his amazement, he still felt the shockwaves and stinging effects; the glass shield did not protect him.** These stinging rays could be felt at great distances from their sources and, to his amazement, would penetrate shields of both glass and copper. Tesla knew that normal electrostatic charges spread over the surface of a metal (copper, for example) shield so that the stinging effect could not have an electrostatic origin and therefore was not electrostatic in nature but was from some other yet unknown source.

Joseph Henry in 1842 had noticed the magnetization of steel needles by the discharge of a Leyden Jar (type of primitive condenser made with a glass bottle). The Leyden Jar was at the top floor of a building while the needles were in the basement. Dr. Henry, noting the passage of these magnetizing rays through brick walls, oak doors, stone and iron flooring and tin ceilings believed that the spark given off by the discharge released "light-like rays" that passed through the material obstacles with ease before magnetizing the needles.

Elihu Thomson, a physics instructor in 1872 had been attempting to make sparks from a Ruhmkorrf Spark Coil more clearly visible for his physics students. He attached one pole of the coil to a cold water pipe and noted that the previously blue spark changed to white. He then attached the other pole to a large metal table top and produced a silver-white spark that would be clearly visible to all attending his lecture. He went to the door of the room to notify his colleague but received a strong shock from the door knob. Turning off the Ruhmkorrf coil prevented the brass knob on the oak door to stop shocking. He returned with his friend, again turned on the coil and discovered that all metal objects in the entire building, no matter how distant from the coil or insulated from the floor, produced long and continuous white sparks upon the touch of a penknife or screwdriver.

The device Tesla perfected was far superior to the Ruhmkorff Spark Coil. The effect of his disrupter device was so strong that a single wire placed in an oil bath produced what he had originally but mistakenly thought to be gaseous streams so powerful that they visibly depressed the oil into a depth of about two inches. Tesla concluded that "besides the air, another medium is present".

> *...he placed himself behind a glass shield but to his amazement, he still felt the shockwaves and stinging effects; the glass shield did not protect him. These stinging rays could be felt at great distances from their sources and, to his amazement, would penetrate shields of both glass and copper...*

Tesla had been able to perfect his device by placing a capacitor between the switch and the D.C. generator thus increasing the power and protecting the generator windings in much the same manner as the diodes Edwin Gray's Switching Tube protected Gray's circuit. Tesla also raised the voltage and quickened the "make-break" rate of the switch to increase the power level by placing a powerful permanent magnet crosswise to the discharge path of the high voltage output of the DC generator wires. The magnetic field caused the discharge arc to automatically "blow out"; the charge passing through the wires in the presence of the magnetic field built up reverse emf (electromotive force or voltage).

Tesla believed he had discovered a new type of electricity with special characteristics. This form of electricity did not consist of alternating waves. **They were longitudinal waves** composed of successive shocking waves with effects that could be seen and felt at a distance. **Vector components of these shock waves were unidirectional. They were able to force charges in the direction of their propagation.** In his patent #787,412, "Art Of Transmitting Electrical Energy Through The Natural Mediums" (April 18, 1905), Tesla calculated the mean velocity of the waves propagated by his device to be 471,240 Km/sec. Knowing that the velocity of light is 300,000 Km/sec would indicate that the type of transmission to which Tesla is referring is different than standard electromagnetic radiation; these special longitudinal "Radiant Energy" waves moved faster than the speed of light.

After conducting hundreds of experiments, Tesla found that "Radiant Energy" longitudinal waves **could penetrate all materials and cause "electronic responses" in metals such as copper and silver.** Impulses exceeding 0.1 millisecond duration produced effects such as pain, mechanical pressures, explosion of thin wires and vibration of objects. Impulses of 1.0 microseconds caused the sensation of physiological heat; at still shorter lengths white light would fill the room. Impulses less than 100 microseconds were safe to work with and Tesla planned to use them for his power broadcast system since they could pass through all matter.

In 1890, Tesla discovered that placing a long single turn copper helix near his magnetic disrupter became covered with an envelope of white sparks. Effects were strongest when the helical coil was placed within the disrupter wire circle. In this "shockzone", the coil was surrounded by a blast of long, fluid like silvery streamers which clung to the surface of the coil, flowing over the coil at right angles to the windings. Tesla hypothesized the electrostatic-like effect was due to radiant transformation rules requiring measurements of discharge lengths and attributes of the helix used (number of turns, diameter, etc.). (Editor: resonance)

The new induction law he discovered showed that radiant shockwaves became much stronger when encountering segmented objects. The radiant shockwaves "flashed over" the outer skin of the helix from end to end without passing through the windings of the coil; shockwaves of 10,000 volts introduced to a 24 inch coil rise to 240,000 volts. The greater the resistance in the helix turns, the higher the maximum voltage. This was completely different from magneto-induction. Remember that he was not working with electricity but with radiant shockwaves.

> *Aether particles were incompressible and could easily pass through space and matter with a speed much greater than the speed of light. This was pure radiant matter but, at the same time, pure Radiant Energy. Cold electricity is one of the forms of this Radiant Energy*

This transformer he invented used radiant shockwaves to produce pure voltage without current. Each transformer had to be "tuned" by adjusting the disrupter to a specific impulse duration. At that point, voltage impulses could flow smoothly, flowing over the copper surface much like a stream of water in a pipe. No amperage could be detected but if the stream was aimed at a distant metal plates, "current" was produced which reached several hundred or even thousands of amperes. Tesla then began to wonder what made up this white, currentless stream.

Tesla determined that normal charge carriers (electrons) could not travel as quickly as the radiant pulse; no current moved the coil because the electrons were choked in the metal lattice of the coil. **The radiant pulse moving over the surface of the coil was not electronic in nature.** He placed the legs of a heavy U-shaped copper bus bar directly to the disrupter primary and then connected the short-circuited system to several incandescent lamps. These lamps glowed with a brilliant cold white light as with Gray's light bulb, proving that the power for the light was not electrical in nature. Tesla believed that the electrons were blocked from flowing through the wire while the radiant pulse was released over the coil surface as a "gaseous" pulse (splitting the positive, according to Gray).

Tesla came to believe that voltage could be viewed as streams of aether under various states of pressure and that his transformers affected the aether as to produce the luminous effects he observed. In his patents, he describes his "light-like rays" as tightly constricted aether streams propelled from his transformers along infinitesimal ray lines along which an incompressible movement occurs instantly through space on all points along its path. Aether streams were being drawn in through his transformer at higher natural pressure and then accelerated in the electrical discharge. The voltage in his transformer could control the brilliance of light in a room but this type of light was almost impossible to register on film. He could heat up a room or cause cool breezes by controlling the voltage in impulse duration in his transformer.

Tesla saw electrical current was really a complex combination of aether and electrons. Through the application of a disrupter, the electrons were removed from the gap by the magnetic field while the aether steams continued to flow through the circuit. He considered aether particles highly mobile with infinitesimally small mass and cross-section as compared to electrons. They were incompressible and could easily pass through space and matter with a speed much greater than the speed of light. This was pure radiant matter but, at the same time, pure Radiant Energy. Cold electricity is one of the forms of this Radiant Energy.

Edwin Gray, Dr. Nicola Tesla and Dr. Thomas Henry Moray all used radiant energy technologies. It is important to understand that the Laws of Thermodynamics and Maxwell's equations do not pertain to "Radiant Energy" technologies.

• Fantastic Projects •

Gravity Conversion Rotary Device

Review is prepared by editor Tatiana Ezhova

Here another fuelless perpetuum mobile is presented.

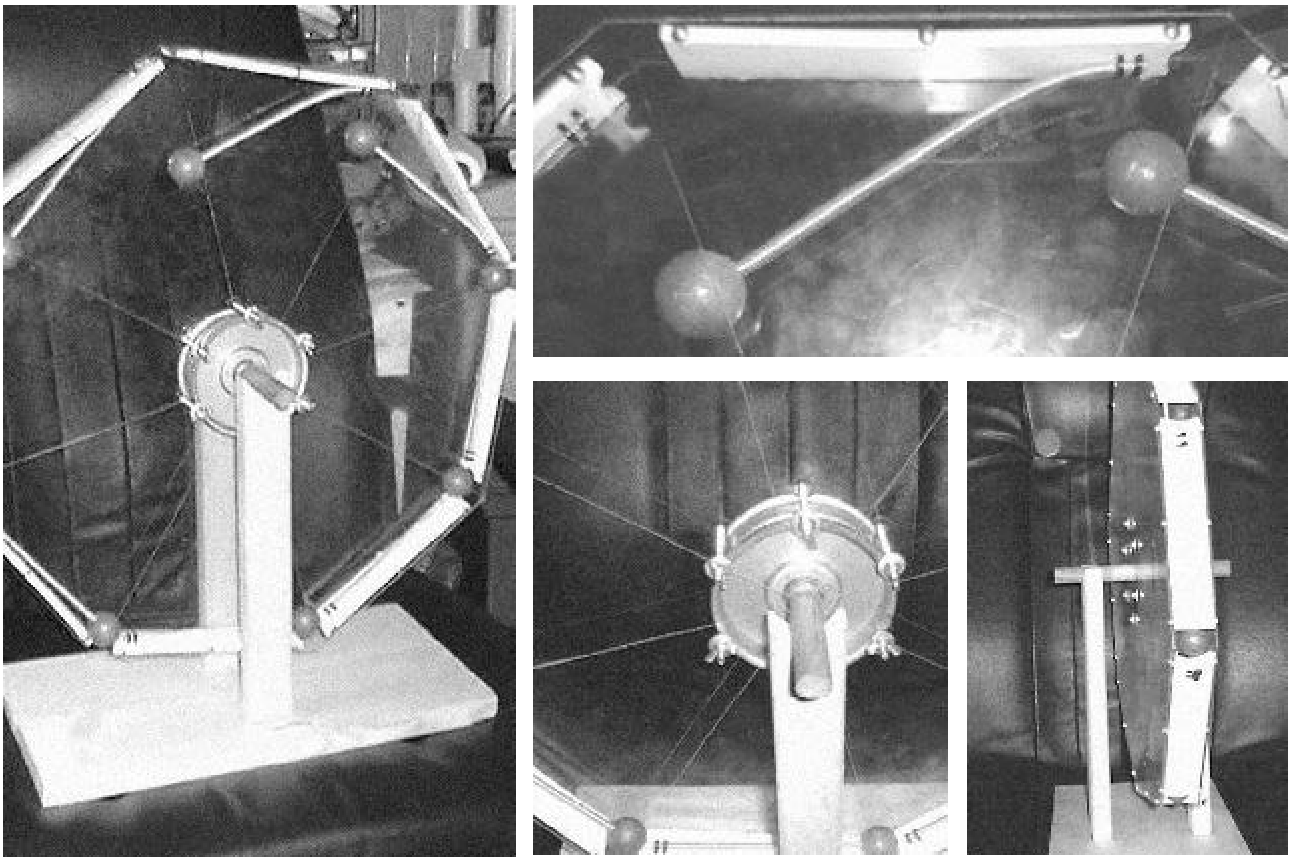

The system design

The design includes a wheel with spring spokes. Pegs may be fixed inside the wheel to balance movement of weights on springs. All parts should be homogenized as to dimension and weight. The author used 4 oz. lead balls and the 3/16" springs. The wheel measures 9" in radius as measured from hub center to center of each lead ball at outer periphery. This device is actuated and stopped by the hand. Its speed is self governing, it ramps up to speed quickly and simply stays there.

Note: you can find more detailed information about the device at http://www.greaterthings.com/News/FreeEnergy/Directory/GravityMotors/photos/. Also read in this issue article "Novozhilov's motor" that describes another wheel motor that does not require any fuel for its operation.

Torsion Technollogies

Xing-liu Jiang, Xiong-wei Wen, China
Science School,
Beijing University of Aeronautics and Astronautics, Beijing 100083
and
Li-jun Han, China
Department of Materials Science and Engineering,
Beijing University of Aeronautics and Astronautics, Beijing 100083

ABSTRACT

Highly localized nuclear activation in electrochemical systems and other electrical discharge processes have been observed by many laboratories in the world. This report is an attempt to explain such anomalous phenomena by using torsion field theory and axion model. Anisotopic behaviours of radiation products, burst character, "heat after death" of excess energy release in electrical discharge systems are considered to be interpreted by the torsion coherence of vortex dynamics with the zero-point energy induced by localized intense field emission of micro-protrusion of the cathode, and the dynamic Casimir effect of transient evolution of triple region of gas, liquid solution, and electrode protrusion. Axion model and Primakoff's effect are proposed for explanation of nuclear transmutation without noticeable gamma radiation.

Nuclear products with high concentration, unidentified tracks with highly collimated lines of low energy nuclear reactions in the electrochemical systems were recorded by CR-39 solid detectors and photo-films, and localized spots with chemical alterations were observed at our laboratory. It is suggested to carry out intensive study of vortex dynamic for explaining the anomalous phenomena in wide area of nature and laboratories. Analysis of vortex dynamics with wide range from pitting corrosion of electrochemical system, laboratory plasma, tornado, to quasar spiral model with extremely high energy cosmic rays in the center region, leads to a conclusion for that vortex dynamics creates torsion fields responding to the anomalous effects.

I. INTRODUCTION

Many laboratories in the world have observed nuclear reactions and **excess heat** in electrochemical systems. The mechanism of such anomalous phenomena is not being well understood according to normally accepted physics.

Nuclear products with high concentration and tracks with highly collimated lines of low energy nuclear reactions in the electrochemical systems were recorded by CR-39 solid detectors and films at our laboratory [1,2]. These facts suggest that quasar model with spiral structure and extremely high enrage cosmic rays in the center could be used to explain the mechanism. It is supposed to use the concept of torsion field to interpret the observed phenomena, typically, the properties of axion acceleration, memory effect, and the polarized nuclear reactions with torsion effect [3].

II. PHENOMENA IN NATURE AND AT LABORLATORIES

The vortex and spiral structures are the archetype that appears at all levels of nature and laboratories, for example, atom structure, vortex lattice in superconductors, dense plasma focus, lightning, quasar etc.

Recent report of dark matter annihilation at the galactic center describes that the cold dark matter near the galactic center is accreted by the central black hole into a dense spike [4]. Particle dark matter annihilation makes the spike to be a compact source of photons, electrons, positrons, protons, antiprotons and neutrinos. It reminds us for that there is a similarity among pitting corrosion with electrochemical noise, laboratory plasma pinching (dense plasma focus, for example), fast laser induced ion beams [5] and quasar spiral model with high-energy cosmic rays in spiral center in spite of large dimension difference. Comparing the experimental results of electrochemical cells with excess heat and nuclear transmutation to astrophysics phenomena, it is supposed that the investigation of vortex dynamics of torsion coherence with the zero-point energy is essential for tapping the zero-point energy.

As it is predicted by quantum mechanics that the vacuum is seething with active energy, even at temperature of zero point Kelvin. This zero-point energy (ZPE) can be thought as an infinite number of virtual photons that are popping out of the vacuum and going back in, but should be a measurable effect en masse. To exam the origin of ZPE background, the interaction of matter with the ZPE can be treated on the basis of charged point particles interaction with a background of electromagnetic zero-point radiation with spectral-energy density.

Based on observation in the experiments, the following features are considered in order to understand the experimental results with electrical discharge systems.

A. ELECTROCHEMICAL DOUBLE LAYER

In an electrolytic cell, the electrolysis with high conductivity and the electrochemical double layer with large layer-capacitance lead to a typical structure of the cathode potential distribution similar to the cathode drop of glow discharge in low gas pressure. For a compact layer the thickness of the double layer is equal to one ionic layer, across which there is a linear fall of potential. Thus, high electric field exists in some regions on the surface of the cathode. The local enhancement of the electric field on the cathode surface with the double layer is related to the protrusions and cracks similar to the tip discharge in air or in a vacuum. The current distribution depends strongly on the surface roughness and the work function of the electron emission. A high transient current density ($> 10^8$A/cm^2) could be expected due to enhanced field.

B. ENERGY CONCENTRATION

On the cathode surface, the high persistent electrical fields ($>10^7$V/cm) and large equivalent capacitance ($>25\mu$F/cm^2) lead to a high energy concentration in the double layer [6]. The concentrated field on the tips of the protrusions or cracks after a long-loading period with deuterium on the palladium cathode surface creates a high transient electron flux because of the large distributed capacitance and the negligible inductance in a localized discharge mini-network. The experimental data show that the reactions take place only in some restricted areas that have specific properties. The idea of micro fusion due to the results of energy concentration and the high deuteron flux could be used to explain the nuclear transmutation.

C.TORSION FIELD AND THEIR EXPERIMENTAL MANIFESTATIONS

Elementary particles have the moment of quantity of motion, i.e. spin. If in any substance the spins of particle have a preferable direction, then it is interpreted as spin polarization of the substance. Every substance creates a torsion-field (or called spin-field or axial field) in the space surrounding it when polarized by spins [7]. The superposition of torsion field, generated by the atomic and nuclear spins of each molecule, determines the intensity of torsion field in the space surrounding each molecule. Torsion field has strong penetration ability and does not interact with the crystal lattice of substances. The torsion field created by rotation of some sort of matter is concentrated in two opposite beams propagating along the rotation axis. The intensity of torsion-field with some lower constant value can be retained for several weeks after the rotation is stopped. Such property of vortex mater has been recently observed in type II superconductors with magnetic flux line lattice [8]. The studies of vortex matter of type II superconductors have shown a number of puzzling phenomena associated with vortex motion, including: low-frequency noise and slow voltage oscillations; a history-dependent dynamic response, and memory of the direction, amplitude duration and frequency of the previously applied current.

Some behaviors of the torsion-field effect have been observed in the experiments of electrolysis regarding the so-called cold fusion, such as when the gas bubble chains come out from the protrusions of the cathode surface a long time after switching off the electrolysis potential. The more surprising thing is the heat after death, which has been recorded by many labs. This phenomenon could be explained by the persistence effect of the torsion-field produced by vortex dynamics of tip effect.

D. EXPERIMENTAL RESULTS IN THE ELECTROCHEMICAL CELL

CR-39 plastic films possessing with a high degree of optical clarity and isotope in track response, and are sensitive to neutron, proton, tritium, alpha, and other charged particles, were used to detect the products of nuclear reactions. The films of CR-39 were immersed in the NaOH electrolyte of heavy water and placed adjacent to the tips of the cathode [9]. After 110 hours, electrolysis experiment with the applied voltage 1 V and current 2 ma, the solid detector was etched by 6.25 N NaOH solution in 70^0C for 11 hours. The photomicrograph shows the cluster of tracks with a cycle crater of 100 μm in diameter and 25 μm in depth. According to the etching condition, the energies of the most of the particles, P, T, α are estimated roughly to be in the range of 1-4 MeV. On the backside of same region of the CR-39 film detector, one can clearly see a few of tracks in the circle area or nearby. Such tracks are believed to be created by recoil-protons of forward incident neutrons. High concentration of the cluster of nuclear tracks could be explained by the quasar spiral model and by the crystal channeling effect [Fig.1]. The experimental results of the generation of tritium with single crystal metal of Pd show that none of the generation of tritium has been revealed by using non-single crystal electrodes [10]. The importance of the crystal channeling effect for nuclear reactions in the electrochemical systems can be inferred.

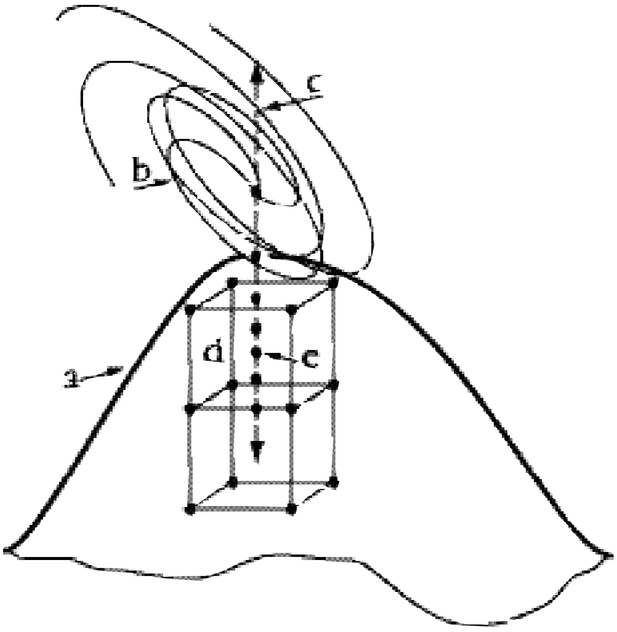

Fig.1
Schematic presentation of a micropinch spiral by tip effect in the electrolysis cell:
a) tip of the electrode; b) spiral structure of micropinch; c) electron beam; d) crystal channel; e) ion beam

To determine the spatial distribution of radiation active sites (RAS), Black-white 135 films of 27 DIN have been used to image the position of the RAS. After a 1,5 year of deposition in glass tube of finishing electrolysis experiments with light water electrolyte for more than 200 hours running, the patterns of RSA were clearly formed on the films after exposure of 100 hours [2]. The bright spots corresponding to the tips of palladium cathode edges can be seen due to the tip effect. The effect of magnetic field on the traces of the charged particles has been observed while the Pd samples exposing to the films, which were folding and wrapping up the sample, were inserted between couple magnets. The tracks stretching along the film surfaces confirm that the tracks were created by charged particles, electrons for example, with low energy about some keVs. Highly oriented tracks can be observed by autoradiography by using normal films locally[Fig.2].

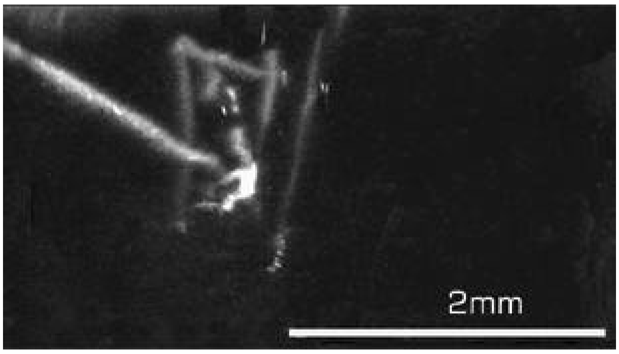

Fig.2
Autoradiography of charged particle tracks of beta delay isotopes on the surface palladium cathode. Some tracks of beta particles are paralleling the cathode surface.

E. SONOLUMINESCENCE AND "BUBBLE NUCLEAR FUSION"

Some scientists of Oak Ridge National Laboratory in America reported their articles of bubbles experiments in Science. Experimental results show that the radiation lights of sonoluminescence possess three characters: short duration with picoseconds; wide continual spectrum; highly oriented thin beams. **Based on those characters, vortex dynamics with axial acceleration of bubble collapse could be inferred.** Nuclear reaction with abnormal gamma radiation was observed [11]. Dr.Claudia Eberlein describes her conclusion on sonoluminescence that only the ZPE spectrum matches the light emission spectrum, which must be a ZPE phenomena [12]. The effect of torsion field on nuclear reactions along the axis of vortex should be taken into consideration for the low ratio of the nuclear products of n/T due to spin polarization of reaction particles.

Axion model and the Primakoff effect are proposed for explanation of nuclear transmutation without noticeable gamma radiation. Many laboratories in the world have been engaged in several experiments to search for axions, light neutral pseudoscalar particles yet to be discovered. The axion would be produced in the solar core through the Primakoff effect if its mass is a few electronvolts and could be detected in the laboratory.

III. CONCLUSION

Researchers of new energy study of "cold fusion" type should pay great attention to the general processes of electrolysis to find the key points, which could play a major role in the transit from the electrochemical processes to processes of torsion and the nuclear processes. From the authors' point of view, the evolution of double layers is of importance to understand the anomalous effect, typically on the protrusions of cathode. The change of space-time near the tips due to torsion field generated is expected to delivery the zero-point energy, and dynamic Casimir effect for the evolution of gas bubbles on the tips are expected to generate photons and excess heat [1]. The yields of transmutation products are related to the current distribution on the cathode surface. The cathodes of thin wire were benefit to the generation of nuclear reactions and excess heat for electrochemical systems. Careful examination of the evolution of electrochemical double layer will lead to a good understanding of pitting corrosion with electrochemical noise, and further to recognize the anomalous excess heat and nuclear reactions.

Heat after death was observed at many laboratories. It is believed that the persistence behaviors of torsion field can be used to explain such anomalous phenomena.

The contact between beads coated with a thin metallic layer or multilayered film and palladium black particles could be regarded as point-contact similar to the tip-effect [13].

We may say that it is expected to reveal the puzzles of the mechanisms of high-Tc superconductivity with pinning vortices by the torsion field theory [14,15].

REFERENCES

[1] Xing-liu Jiang, Jin-zhi Lei, Li-jun Han, Dynamic Casimir Effect in an Electrochemical System, J. New Energy, Vol.3, No. 4, 47(1999).
[2] Xing-liu Jiang, Chang-ye Chen, Li-jun Han, Tip Effect and Nuclear Active Sites, Proc. of the 7th Intern. Conf. on Cold Fusion, Vancouver, April, 1998, pp.175.
[3] Don Read, Excitation and Extraction of Vacuum Energy Via EM-Torsion Field Coupling Theoretical Model, J. New Energy, Vol. 3, No. 2/3, p.130(1998).
[4] P.Gondolo, J.Silk, Dark matter annihilation at galactic center, Phys. Rev. Lett. 83, 1719(1999).
[5] BPS, Laser light in , stream of protons out, Physics Today, Jan.2000,p.9.
[6] G.Korluem and J.OM. Bookris, Textbook of Electrochemistry Vol. II. Elsevier Publishing company Amstorsan , 1951. P.364
[7] A.E.Akimov, G.I.shipov, Torsion fields and their Experimental Manifestations, Journal of New Energy, 2(2), 67(1999).
[8] Y.Paltiel, E.Zeldov Y.N.Myasoedov el. al. Dynamic instabilities and memory effects in vortex matter, Nature 403, 398 (2000).
[9] X.L.Jiang, L.J.Han and W.Kang. Concentrated energy and Micro Nuclear Fusion, ICCF6, Oct 1996, Japan p.580.
[10] R.L.Matlock, F.E.Collins, G.R.Bancher, Anomalous tritium found in the recombined off gasses during electrolysis using crystal cathodes. Elem. Energy(cold fusion), 26, 28(1998).
[11] Taleyarkhan R P, West C D, Cho J S, Lahey Jr R T, Nigmatulin R I, Block R C. Evidence for Nuclear Emissions During Acoustic Cavitation. Science, 2002 295: 1850~1862.
[12] Claudia Eberlein. Theory of quantum radiation observed as onoluminescence, Phy.Rev.Lett.53,2772(1996)
[13] G.H. Miley, G.Narne, M.J.Williams, J.A.Patterson, J.Nix, D.Cravens, and H. Hora, Quantitative Observation of Transmutation Products Occurring in Thin-Film Coated Microspheres During Electrolysis, Proceedings of ICCF-6, OCT. 1996, Japan, p. 629.
[14] T. Matsuda, K. Harada, H. Kasai, O. Kamimura and A. Tomomura, Observation of Dynamic Interaction of Vortices with Pinning by Lorentz Microscopy, Science, 271, 1393(1996).
[15] G.W. Crabtree and D.R. Nelson, Vortex Physics in High Temperature Superconductors, Physics Today, April 1997, p.32.

New Energy Technologies
PRINTED EDITION ADVERTIZING RATES

Full Page Ad
B/W - $400
Colour - $800

Half Page Ad
B/W - $200
Colour - $400

Quarter Page Ad
B/W - $100
Colour- $200

New Energy Technologies is distributed in the leading Scientific Laboratories in Russia and has subscribers in more than 20 countries all over the world

To advertise contact us:
Farady Labs Ltd, Phone/fax: 7-812-380-3844
http://www.faraday.ru, office@faraday.ru

Fantastic Projects

The Invention of Roman Solomyanny

Our previous issue featured an article about Roman Solomyanny who came to our office and told us amazing things. He said that during his army years (according to his words, he was in the military intelligence) he made a radio technology invention and later on built a free energy generator. The inventor claimed that his device had been working for a certain period of time without any fuel and had been producing power enough for heating and lighting his countryside house. We could not ignore such an intriguing piece of information. Moreover, Roman promised to build and present his device in 2003. We paid him for this interesting story to be published in the issue and prepared for the continuation... Alas, to no avail... His mobile phone was silent and we started to worry about the inventor's health. Fortunately, we managed to contact his parents on his home telephone (Roman lives in St. Petersburg) and they told us that the chap was OK. It seems he's lost his interest in us. But we still hope that Roman did invent something really exciting and not only polished up his talent for writing science fiction articles. Those people, who want to know the details about this invention or discuss it with the inventor himself, can get in touch with us for his contact information. Anyway, we will be very pleased to see him again full of new exciting stories. We have already introduced the 'Humour' section into the magazine and are planning to start the 'It Just Can't Be So' section.

Editor

Physicists are Joking

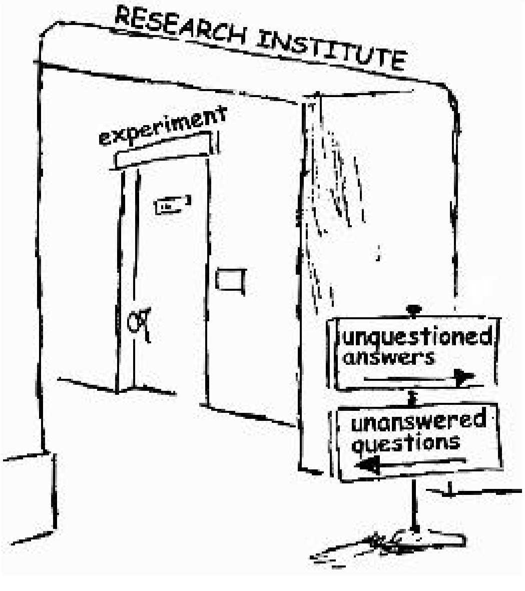

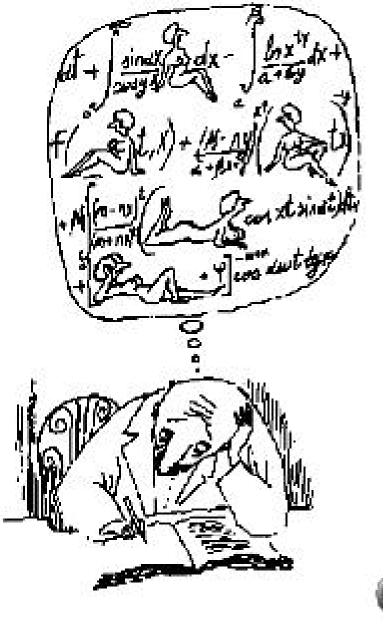

Anti-Gravity: The Holy Grail of the 21st Century

Bruce A. Smith, USA
PO Box 1676, Yelm, Washington 98597
Email: brucesmith@rainierconnect.com

A Primer on the Role of Electromagnetic, Electrostatic, and Torsion Fields in Anti-Gravity and Field-Effect Propulsion

Before visiting a University of Washington physics professor to talk about electrostatic propulsion, and hopefully anti-gravity, I realized: *Birds defy gravity. So do 747's for that matter. They apply the laws of physics and lift off the ground. That's anti-gravity, isn't it?* Yes, that's true, I suppose, in a metaphorical sense. But seagulls, jumbo jets and space crafts manifest anti-gravitation effects strictly speaking. What I want to address here is not the overcoming of gravity but the neutralizing of it.

Dr. Eugene Podkletnov, one of the foremost researchers in anti-gravity, and whose work is sought by NASA, Boeing and British Aerospace (now known as BAE Systems) describes the hunt for anti-gravity the greatest scientific quest of this century. He calls for an international effort, akin to the Manhattan Project that developed the atomic bomb, to conquer the secrets of anti-gravity, and usher in a new era of scientific understanding whose technological development will be at a scale so vast that the potential outcomes are merely hinted at by our previous achievements.

Just getting such a project off the ground will require unprecedented international cooperation, and public disclosure as well; the potentials are that vast, that scary, and that dangerous. Dr. Dan Marckus, noted British avionics expert, states in The Hunt for Zero Point, the seminal work-to-date on anti-gravity written by Jane's Defence Weekly aviation editor Nick Cook, that the secrets of anti-gravity in the wrong hands will make thermonuclear weapons look like firecrackers.

The secrecy surrounding anti-gravity research is phenomenal. Boeing refuses to publicly acknowledge any activity in anti-gravity development despite the fact that its competitor and sometime sub-contractor, British Aerospace (BAE Systems) is, the latter providing funds for four university research efforts as part of its Project Greenglow, one of which was a Podkletnov replication experiment headed by Dr. Clive Woods at the University of Sheffield,. Further, Nick Cook publicly, and privately to me in an email, states quite directly that George Muellner, former director of Boeing's ultra secret Phantom Works, claims Boeing sought the services of Dr. Podkletnov to unlock the secretes of his gravity-shielding research. Cook says that Muellner states Boeing was denied Podkletnov's services due to the objections of Russian officialdom, which the Russian-born Podkletnov must pay attention to, apparently, despite the fact that he works in Tempere, Finland. Dr. Podkletnov, wisely perhaps, chooses not to clarify these particulars despite our several emails.

Perhaps Boeing can deny any activity on anti-gravity because NASA is doing its own research, and as a prime contractor to NASA, such as by running the Space Shuttle Program, Boeing probably knows what NASA knows. NASA spent $600,000 recently in its Breakthrough Propulsion Physics program (BPP) to purchase Podkletnov replication equipment. Inexplicably, that equipment sits in boxes in NASA's Marshall Research Center in Huntsville, AL, awaiting more funding; according to an email I received from NASA propulsion researcher, Ron Koczor.

But enough with this business; what do we know about anti-gravity?

The search for that answer has taken me to some exciting and obscure places in this world, like the Aeronautics and Astrophysics lab at the Seattle campus of the University of Washington. I called those folks because Nick Cook in The Hunt for Zero Point, mentions that UW received a NASA contract to study theories of inertia as part of its BPP program. *That's a good place to start,* I thought, but it took backtracking to BPP Project Director Marc Millis at NASA's Glenn Research Center in Cleveland to find Dr. John Cramer at the UW Physics Department. His mission was to confirm with Dr. James Woodward the latter's 1996 preliminary research into the loss of gravitational mass in a targeted piece of metal from oscillating capacitors. Although Woodward's initial data appeared encouraging, NASA's Millis told me that their funding dried up before completing their research. Furthermore, the entire BPP became unfunded in 2002 and now in 2003 has become a hazy, privatized version of its former NASA subset self.

However, UW is continuing related research, such as magnetically confined fusion energy generators and that, for me, by using electromagnetism to contain an inner field makes it a close cousin of anti-gravity and field-effect propulsion. I spoke with Professor Uri Shumlak who told me that he and other UW staff from the Department of Aeronautics and Astronautics, along with a bevy of their grad students, are building a prototype of a fusion generator called HIT, which stands for Helicity Injected Torus.

This donut ring-shaped torus encloses a roundish chamber. Within that chamber a vacuum is first created, and then a volume of hydrogen gas is introduced and heated to a few million degrees Celsius, which separates the electrons and protons from their atoms turning the whole stew into a quasi-neutral foam of plasma. Then the torus envelopes the plasma with a magnetic field to keep it away from the sides of the chamber enabling the plasma mass to stay hot, and keep the rest of Seattle cool. (While I was standing next to his little eight foot long gizmo, Prof. Shumlak assured me there was no danger of a couple million degrees of heat escaping. The heat density of the plasma was "too low" for me to, well, break a sweat over. His quote was, "There's no more heat mass inside that chamber than what's contained in a cup of coffee." I sure hope you're right, Doc.)

Then, once the plasma field is contained, the magnetic field squeezes the plasma, fusing the nuclei of one hydrogen atom into another. As the hydrogen couples combine, a helium atom is created and a neutron is released, along with lots of energy in the form of heat. One day, such a generator will give us unlimited amounts of electricity, as the heat can produce electric voltage.

Lots of electrical power on the cheap the UW predicts; and the Department of Energy agrees, once the details of building reliable magnetic field generators are solved. What does magnetic fusion have to do with anti-gravity? Two things: first, magnets. Electromagnetism seems to be one of the major players in anti-gravity, particularly the use of electromagnetic fields to contain other fields, such as plasma fields in the HIT, or torsion fields, but more about that later. Secondly, the HIT works, or is about to. It's real and mainstream science embraces it, while anti-gravity is, well, a little more out there and reliable data is harder to obtain. So the technology of HIT lays a base that other research can build upon, such as not only containing other fields, but also building field effect propulsion systems, the most elementary of which is electrostatic propulsion, and aspects of that are already being applied by NASA.

Electrostatic propulsion uses electrical fields differently than electromagnetism does. In EM a current flows and creates a field, while in electrostatic systems the current is static and a charge builds up a field, such as in a capacitor.

These theories are utilized on NASA's Deep Space I, a probe bound for the outer reaches of our solar system. On Deep Space I, the propellant, a tankful of xenon gas, is excited electrostatically into positive ions. The engine has a negative charge at the exit end, so the charged xenon rushes out the tail pipe with a greater thrust than if it was just using conventional chemical propellants. In fact, the electrostatic propulsion system on Deep Space I allows it to fly at 60,000 mph, or 10,000 mph faster than it would with a conventional rocket. In addition, only 82 kg of xenon is needed for its entire mission, so with its smaller mass and weight Deep Space I will fly along side its intended target, a comet, and drag race on equal footing while filming and conducting studies. Again, not anti-gravity per se, but electrically charging Deep Space I's fuel-field sets the stage for a closer look at electrostatic propulsion.

Taking that closer look is Tim Ventura and his fellow researchers at American Antigravity, an organization based in Kirkland, WA. Ventura and his crew use electrostatic asymmetrical capacitors to create a field that levitates objects, such as their small, kite-like "lifters." These lifters are very light, weighing only a few ounces, and have balsa wood struts that support the capacitors. When two capacitors of different size receive their share of a 30,000 volt charge, the lifter lifts- no motors or wings.

How, no one really knows in my judgment; and the phenomenon is replete with controversy and mystery. But as one who has seen a lifter fly, let me tell you what one looks like and what I saw when Tim Ventura's took off.

Tim has been building lifters since he was a kid and has perfected a four-foot, by-four-foot, by four foot triangular lifter which has flown so many missions in his garage that the silver aluminum foil has turned white. The thin, chopstick-like balsa wood ribs that hold the aluminum foil in place are joined every few inches by a vertical strut (much like a telephone pole on a HO model railroad set) which sticks up and secures the copper or stainless steel wire of the upper capacitor. The ribs are intersected every ten inches or so by the strut of an interior triangle, since the whole lifter is composed of interconnected isosceles triangles which give the necessary strength to the balsa wood frame. All told there is about 30 linear feet of aluminum foil and a similar run of wire.

The lower and larger capacitor is a strip of aluminum foil stretched between the horizontal balsa wood struts. The second capacitor is a thin strip of 50 gauge wire mounted about one inch above the aluminum foil. As capacitors they store electrical charge but don't pass it on in a current.

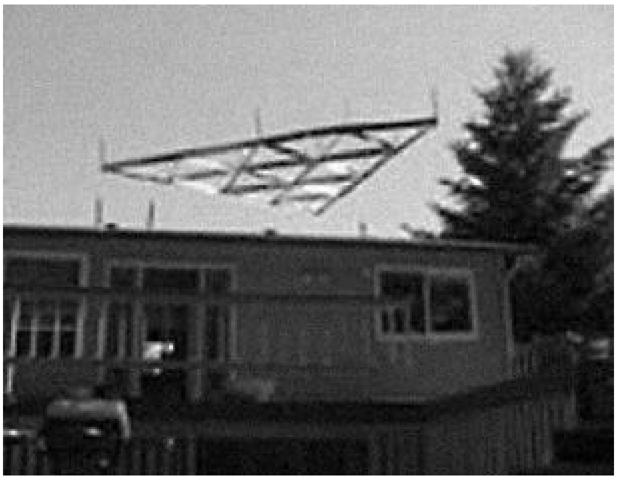

Fig.1
Lifter

The negative lead goes to the lower aluminum foil and the positive lead is attached to the upper wire. The three corners of the lifter are tethered to the work table so that the electrical leads from the power source are not broken off in flight.

The power source kicks out 15,000 volts at 250 watts. Tim uses a voltage generator made by Information Unlimited, Inc, but before the current reaches the capacitors, the voltage is stepped up to 30,000 volts by Tim's homemade voltage multiplier stack. At full throttle the lifter is straining at the tethers, bending the balsa wood frame near the point of fracture.

Throw the switch and at around 17,500 volts the lifter begins to quiver in take-off. At the full power of 30,000 volts the lifter is roaring and a noticeable downward breeze is observed. Many physicists call it "ion wind," and say that it is how the lifters fly. But what exactly is ion wind, and can it be the cause of flight?

"Ion wind is the movement of ionized air particles," according to Ventura, "which flow downward according to electrical charge." Here's his theory:

The positively charged wire on the top part of the lifter steals elections from the surrounding air, leaving the effected air molecules positively charged. These positively charged air molecules, or ions, then head downward toward the large source of negatively charged electrons generated by the aluminum foil. These air molecule ions are bigger and heavier than the electrons seeking them, so there is a net thrust downward pushing the whole lifter up. That's the theory, and frankly all I can do to verify the theory is to tell you what others tell me. Before I do that, though, let me tell you what I experienced standing next to a levitating lifter.

In flight the lifter emits a high whining, hissing buzz, and I could feel a good breeze coming up at me from the work table underneath the lifter. Also, standing next to the lifter but not touching it, the hair on the back of my head started to rise up in electrostatic-like fashion.

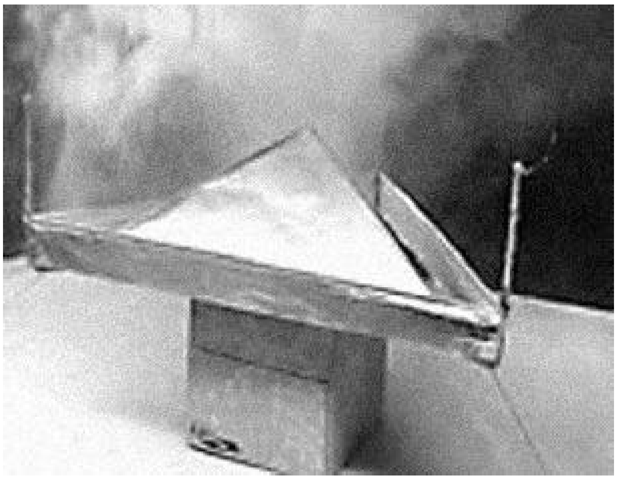

Fig.2
Lifter

To analyze the air currents Tim blew baby powder at the top of the lifter. The majority of the particulate cloud was drawn into the middle area of the lifter and then sucked downward. A kind of vortex was created at times, for intermittently I could see a cloud forming into an organized column beneath the lifter and then spreading out in 360 degrees once it hit the work table surface.

Electrostatic propulsion uses electrical fields differently than electromagnetism does. In EM a current flows and creates a field, while in electrostatic systems the current is static and a charge builds up a field, such as in a capacitor

Is that ion wind? Well, there certainly was a breeze, and it sure felt like air, but how would I know if it was ionized? Something definitely sucked the baby powder down, but was it more than just regular air blowing past me? Again, I don't know.

Is the movement of wind why lifters fly, regardless of whether it is ionic or not, or is the wind just a by-product and not the propulsion? Could the capacitors be creating a field that neutralizes gravity, allowing the craft to levitate? Or are they creating some kind of new field that is localized and the surrounding ambient field pushes this "field-bubble" up, much like a helium balloon is pushed up by the surrounding heavier air trying to fill the emptier "field" of the lighter helium?

Ventura thinks at least two phenomena are at work. Ion wind is definitely one he feels, for the breeze is self-evident. However, he thinks a second effect is at work, too, and many agree with him. Most speculation concerns what is called the Biefeld-Brown effect, the "Brown" being T. Townsend Brown, whose name is well known in early quantum research and whose work is prominently discussed in Nick Cook's The Hunt for Zero Point.

The Biefeld-Brown effect, according to Ventura, is the theory that high voltage, air-gapped capacitors with different or asymmetrical capacities generate a net directional force upwards from the larger element to the smaller element, which on the lifter is from the aluminum foil to the wire. This force then pushes against the ambient energy field of the surrounding area, perhaps pushing against a more rigid energy field of the zero point energy field.

Brown apparently made his case for these electromagnetic effects, receiving patents in the 1960's for his research. NASA's Dr. Jonathan Campbell at Marshall Research Center in Huntsville confirmed to me that he also, has received a patent recently for his research into the thrust effects of asymmetrical capacitors.

Editor: By the way, this patent is meaningless from the point of view of a prime claim. Before that there had been many publications on the subject.

A.V. Frolov.

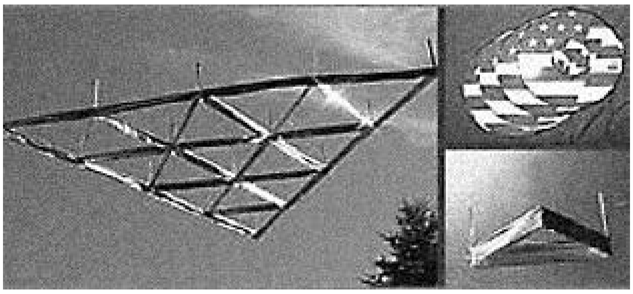

Fig.3
Lifter

However, prominent physicist Hal Puthoff, whose research is a broad swath across the fields of the 'new physics', featured in both The Field and The Hunt for Zero Point, and who was also the military's "Top Psychic" as the twelve-year director of the CIA's remote viewing squadron, has a different perspective: "I'm quite certain at this point that the so-called 'lifter' phenomena is just an electrostatic ion wind phenomena, not 'antigravity.'"

But Dr. John J. Rusek, Adjunct Professor of Aeronautics and Astronautics at both Purdue University and the United States Air Force Academy, says that "Initial findings of 'classroom' experiments with lifters show ionic wind to be way too small a factor, by three orders of magnitude." Dr. Rusek has formed a technological company, Swift Enterprises, to continue this research and bring it to the level that is "presentable to the mainstream physics community."

Along these lines, Jean-Louis Naudin shows on his extensive web site, not only how to build a lifter, but also several photos of research into the ion wind issue. Naudin's team has wrapped test lifters in plastic yet they still produce anti-gravitational effects.

Others may have a clue to the second or even a third force at work. Researcher Fran De Aquino, professor of physics at Maranhao State University in Sao Luis, Brazil, is described in the literature as showing that "bubbles of localized space-time" can exist in variance to the surrounding fields. Anecdotal experience suggests that the lifters may be undergoing such space-time anomalies.

Editor: At this point we need to interrupt the author as he has already completely confused the reader. I would recommend referring to Thomas T. Brown's patent: USA patent #3,187,206, 1965.

Alexander V. Frolov

The Hunt for Zero Point states that NASA sought the services of Dr. Eugene Podkletnov, and although their replication research languishes, The Hunt claims that researcher Ning Li, of Huntsville, AL is pursuing this line of research as a private contractor to NASA.

Another Huntsville operation, Transdimensional Technologies, is exploring these multi-faceted phenomena as well, and its extensive web site shows it to be a frequent contractor to NASA, including research into "asymmetrical capacitive propulsion," and capacitor-based devices to test "ion wind" forces. Jeff Cameron, of Transdimensional, is said by Ventura to be "the father of the lifter," having developed them while exploring anomalous torsional effects of high energy lasers. The lasers twisted and at the time it was considered a nuisance. But the unknown forces at work later led Cameron to found Transdimensional, develop lifter technology to a commercial level, and subsequently patent many pieces of related technology. Unfortunately, I have been unable to reach Jeff

Cameron or anyone at Transdimensional for any kind of confirmation.

Nevertheless, how would gravitons be blocked, or gravity shielded?

Dr. Hal Puthoff says there are two ways of looking at it. First, one can look at the issue from a quantum point of view, that there is a particle exchange between the gravitons and something else, and the net effect is anti-gravity. The how's and why's of that are speculative, so Puthoff turns to a classical approach for answers. He prefers looking at "engineering the vacuum." To do that one must first look at the vacuum.

As I understand it we are all in the vacuum, everything is. The "vacuum" is the matrix that contains all matter and all energy. It is the engineering perspective of the zero point energy field, or the *"Field"* as popularized by Lynn McTaggart in her masterpiece The Field. Puthoff shared with me statements from fellow researcher Dr. T. D. Lee that state: "The vacuum is the seat of energetic particle and field fluctuations, and ... is the seat of space-time structure ... that encodes the distribution of matter and energy.... The vacuum is energetic in its own right."

Thus energy can be drawn from the field; and spacecraft can have "vacuum propulsion systems, or propellant-less propulsion," in other words, field effect propulsion.

At any rate, more and more physicists are thinking that the vacuum can give them a whole lotta oomph, enough to propel spacecraft; and when they learn how to corral it, a whole bevy of new phenomena may be encountered, including anti-gravity. This new potpourri of research is being called by many the "new physics." And although his approach is classical, Dr. Hal Puthoff seems to be sensing what's out there waiting to be discovered.

Puthoff's current research as been to explore "the perturbation of atomic or molecular ground states, hypothesized to be equilibrium states involving dynamic radiation/ absorption exchange with the vacuum fluctuations. In this model atoms or molecules ... are expected to undergo energy shifts that would alter the spectroscopic signatures of excitations involving the ground state." Puthoff says he's had no success so far with this approach, but his words remind me of De Aquino's speculation that objects lose mass as they absorb energy. Pull energy from the field around you and you lose weight. Bingo, lift-off. **But how does one pull energy from the field?**

Editor: Some solve this problem by means of the ether density changes, which are performed with the help of vortex longitudinal-wave technologies.

A.V. Frolov

Torsion fields might play a role here according to many, and the literature on anti-gravity is filled with the phrase, "torsional effects." But what exactly is a torsion field? "It has something to do with spin," Nick Cook told me on the phone. "You have a torsion field when you spin something. Add a little electromagnetism and you might have anti-gravity." That's the short-hand version of it and here's a deeper look.

Mike Wright, resident physics expert at BeyondTheOrdinary.Net web stream radio, told me this: "When forces create curvature (such as rotation) in more than two planes, a torsion field results. Not only does the object go around, but it goes around and 'down' or 'up', and the up/down movement is an additional acceleration in that dimension. EM and gravitational fields differ by having a magnitude of force and only one direction of movement.

"A tornado is a structure of air in air. A whirlpool is a structure of water in water. So, because more than two planes are involved, objects can be created from 'nothing'; that is to say that objects can be created from the medium of the environment, such as a tornado from two air masses of differing temperature."

So spin plus movement is the key. Again, Tim Ventura is on the hunt. He demonstrated to me **that spinning magnets will cancel out their magnetic fields sufficiently so that two magnets facing each other with like poles will not push each other away if one of the magnets is rotating**

perpendicular to the force of opposition. It's not anti-gravity, but it gets us closer to the heart of the matter.

Further, Russian physicists have been researching the torsional effects of both subatomic particle spin, and the loss of gravitational mass in planets from the angular momentum of their orbits. Spinning makes something happen, but what? Tornadoes and Mother Nature might have a few clues.

Tornadoes spin, in a sense, although no one is Oklahoma who has spent a night in a storm shelter during an F5 event would describe the tornadoes in the night sky as *spinning*. Nevertheless, tornadoes have anomalous effects that are legendary: blades of grass stuck into mirrors, a piece of straw embedded flawlessly into a tree trunk. How? It seems as if the laws of mass, gravity and inertia are melted as winds swirl at speeds up to 300 mph in an organized vortex pattern. Is this a clue to melting the pull of gravity?

Getting information to this question has not been easy. Many scientists claim not to have even heard of torsion fields, including particle physicists at major US universities. So, again I turn to Nick Cook and The Hunt for Zero Point

Dr. Dan Marckus says that if "you generate a torsion field of sufficient magnitude the theory says you can bend the four dimensions of space around the generator. The more torsion you generate the more space you perturb. When you bend space you also bend time."

Marckus continues, "If you dipped ...one of these whirlpools ... into the zero point energy field, the seething mass of latent energy that existed on an almost undetectable level all around us [in the field would] ... react in an almost magical way by directing that energy."

The torsion field, in effect, is "a pump, a 'coupling' device that could dip into, and then direct, energy out of the zero point energy field." "But," Marckus continues, "the vortex wasn't a three-dimensional phenomena or even a four-dimensional one. It couldn't be. For a torsion field to be able to interact with gravity and electromagnetism it had to be endowed with attributes that went beyond the three dimensions of left, right, up and down, and the fourth-dimensional time field they inhabited; something that the theorists for convenience sake labeled a fifth dimension-hyperspace."

Cook concluded that the torsion fields "bind with gravity ... to produce a levitation effect -- an antigravity effect," but "it wasn't doing so in the four dimensions of this world, but somewhere else." That somewhere else is hyperspace. **So how do we activate torsion fields and enter hyperspace?**

Dr. Eugene Podkletnov may have a clue.

Podkletnov, the Russian researcher working in Finland, has studied the gravity shielding effects of superconductors. Again, Nick Cook in The Hunt, relays vital information. Cook says Podkletnov claims, "If the superconductors are rotated considerably faster than 5,000 rpm ... the disc experiences so much weight loss that it actually takes off." Thus, torsion field creates levitation. I emailed Dr. Podkletnov to find out more about this issue. He replied: "[A] fast rotating object can, under certain conditions, cause the polarization of the volume that it occupies in space and around it. This polarization causes the gravitational effect as it modifies [the] local gravity field. The vortex of the polarized particles will create a vertical thrust with a certain force and spatial momentum. Some scientists call these polarized particles gravitons. The term graviton is an artificial one and at present we are not sure if it is a wave or a particle and what type of particle. Maybe it is a usual tachyon or a superluminal neutrino (a faster-than-light-particle). Polarization of the media means that the spins of electrons, protons, neutrons and of small subatomic particles that constitute the fabric of space or vacuum would be parallel. Then a kind of gravity well is formed and the objects tend to fall into this well. We observe this picture as an object rising to the sky. Polarization of the media (of space) causes some glow around the object as it acquires additional energy and because of it, the glow around some objects is observed."

What I understand from Dr. Podkletnov is that gravity is the effect of spin -- the spin of all subjected particles, from the sub-atomic level and up being parallel; thus they are all aligned to fall into the gravity well of earth. And spinning objects, such as his superconducting discs, when influenced additionally by an electromagnetic field, will experience a shift in the spin of the sub-atomic and atomic elements. They will be turned and not be aligned in parallel. Thus, they are able to levitate.

But how to polarize the media and get things spinning? Enter Dr. Marcus Hollingshed, an enigmatic figure allegedly from Cambridge University. Dr. Hollingshed claims to have built a six-ringed toroidal coil antigravity device, which achieved great effect using rotating magnetic fields. In January 2003 he announced on the Internet that he has developed a 160kg vehicle able to lift in excess of 2000kg and that it has both horizontal and vertical drive features. His device cannot only go up and down and sideways, but it can push things away and pull objects to it.

In addition, the field that the device purportedly generates is capable of being broadened and weakened, or narrowed and amplified in a lensing effect, with the field producing an absolute vacuum of 2.2m spherical diameter. Best of all, when it's cranked up the core of it goes invisible, although the term Dr. Hollingshed uses is that there is a "loss of reflected light." There are no reports of independent confirmation, and Nick Cook says he hasn't been invited to see it, so, he's skeptical.

Where does this leave us? Perhaps Dr. Podkletnov's words sum up our current situation. "Modern theoretical physics cannot give you the direct answer to your questions (levitation, torsion fields, etc.,) and a scientist who would agree to give you the answer cannot be regarded seriously, softly speaking. If you had asked Dr. Einstein if he were an expert on gravity, the answer would be NO. I can repeat his words: No, I am not a magician, yet; I am still learning."

ELECTRICAL LEVITATION

Wayne Macleod, USA
Email: dynamars2100@yahoo.ca, cwleod@shaw.ca

Gravity is the incremental slowing of time as we move closer to its source. If we imagine panes of glass stacked on top each other, one pane slightly denser than the one immediately above, a ray of light coming from above will constantly refract in a curved beam downward as it enters the stack. The same is true of a ray of light entering a gravity field. The analogy is not too far fetched because refraction is caused by a slowing of light in the denser medium. Of course, gravity affects material objects as well as light, but this is explained by every object traveling on a 'world line' in spacetime. We can use light as in the glass example because light is its own world line.

Why the world lines of objects take the paths they do can best be understood by compressing our normal 3 dimensional space into 2 dimensions, length and height only. When we throw a stone into the air it rises and falls in a parabolic arc in these two space dimensions. That path in space is a complete mystery until we consider another dimension, time, a dimension we can imagine measured at a right angle to the plane of the 2 space

energy and because of it, the glow around some objects is observed."

What I understand from Dr. Podkletnov is that gravity is the effect of spin -- the spin of all subjected particles, from the sub-atomic level and up being parallel; thus they are all aligned to fall into the gravity well of earth. And spinning objects, such as his superconducting discs, when influenced additionally by an electromagnetic field, will experience a shift in the spin of the sub-atomic and atomic elements. They will be turned and not be aligned in parallel. Thus, they are able to levitate.

But how to polarize the media and get things spinning? Enter Dr. Marcus Hollingshed, an enigmatic figure allegedly from Cambridge University. Dr. Hollingshed claims to have built a six-ringed toroidal coil antigravity device, which achieved great effect using rotating magnetic fields. In January 2003 he announced on the Internet that he has developed a 160kg vehicle able to lift in excess of 2000kg and that it has both horizontal and vertical drive features. His device cannot only go up and down and sideways, but it can push things away and pull objects to it.

In addition, the field that the device purportedly generates is capable of being broadened and weakened, or narrowed and amplified in a lensing effect, with the field producing an absolute vacuum of 2.2m spherical diameter. Best of all, when it's cranked up the core of it goes invisible, although the term Dr. Hollingshed uses is that there is a "loss of reflected light." There are no reports of independent confirmation, and Nick Cook says he hasn't been invited to see it, so, he's skeptical.

Where does this leave us? Perhaps Dr. Podkletnov's words sum up our current situation."Modern theoretical physics cannot give you the direct answer to your questions (levitation, torsion fields, etc.,) and a scientist who would agree to give you the answer cannot be regarded seriously, softly speaking. If you had asked Dr. Einstein if he were an expert on gravity, the answer would be NO. I can repeat his words: No, I am not a magician, yet; I am still learning."

ELECTRICAL LEVITATION

Wayne Macleod, USA
Email: dynamars2100@yahoo.ca, cwleod@shaw.ca

Gravity is the incremental slowing of time as we move closer to its source. If we imagine panes of glass stacked on top each other, one pane slightly denser than the one immediately above, a ray of light coming from above will constantly refract in a curved beam downward as it enters the stack. The same is true of a ray of light entering a gravity field. The analogy is not too far fetched because refraction is caused by a slowing of light in the denser medium. Of course, gravity affects material objects as well as light, but this is explained by every object traveling on a 'world line' in spacetime. We can use light as in the glass example because light is its own world line.

Why the world lines of objects take the paths they do can best be understood by compressing our normal 3 dimensional space into 2 dimensions, length and height only. When we throw a stone into the air it rises and falls in a parabolic arc in these two space dimensions. That path in space is a complete mystery until we consider another dimension, time, a dimension we can imagine measured at a right angle to the plane of the 2 space dimensions. By multiplying the time of the object in flight by the speed of light, we have a three-dimensional coordinate system, not of space but of spacetime. We can then graph the world line of the object. So why does the object fall? It is because spacetime near massive bodies is curved. A remarkable fact is that the curved world lines of all objects in the same gravity field, whether of a thrown ball or fired bullet, have the same radius of curvature. Hence all objects fall with the same acceleration.

To neutralize gravity we must somehow 'straighten' the world lines of objects traveling in curved spacetime. It would be as if we had in our stack of glass panes a bubble, around which the density of glass became less on approach to the bubble. Then the ray of light, when close to the bubble, would refract opposite to the path it would normally take when traveling thought the rest of the glass. Similarly, if we could manufacture a 'time bubble' in a gravity field opposite to the time slowing known as gravity, we would have any object embedded in the time bubble isolated from that gravity field.

The Electrical Levitation theory can best be understood by first considering an analogy. Let us imagine a wheel spinning on an arm like a child's propeller toy. The arm also rotates, in a direction opposite to the spinning wheel. We consider the rate of spinning of the wheel from the point-of-view of two observers, one observer stationary on the ground, the other observer rotating with the arm. Obviously the two observers will not see the same rate of spinning on the wheel. Because the arm is rotating opposite the spin of the wheel, its rotation must be subtracted from the rate of wheel spin as seen by the stationary ground observer. This is not true of the observer rotating with the arm, who will see the spin of the wheel as if there were no arm rotation.

The concept is easier to envision with linear motion, such as of a man throwing a ball on a moving train. The velocity of the ball seen by the pitcher on the train will not be the same as its velocity seen by a stationary observer on the ground. As the stationary observer sees it, that velocity will have the train's velocity subtracted from it if thrown against the train's velocity. It is the same with rotary motion: the velocity of the wheel for the stationary observer is slower because the arm's rotation is subtracted from it.

But suppose the wheel rotation is *the same* for the two observers! Something must be different between them, and that is time. As explained by Special Relativity for linear motion, time is not a universal constant; it differs between observers depending on their relative motion. The property of nature that *is* constant, that produces this relative time difference, is the speed of light. So we must look for a similar constant in nature for a relative time difference between observers in our rotary example. We have one in electron angular momentum, known as electron "spin," $h/4\pi = 5.28 \times 10^{-35}$ kg-m^2/sec, where h is Planck's constant. Like the speed of light, this quantity is constant for all observers, whether the observer is on a rotating system or stationary on the ground. Here is the tool for producing our time 'bubble'.

Let us now imagine a series of concentric rings, all rotating in the same plane and in the same direction. Electric current is pushed through these rings in the same direction as the ring rotation. (Current here is considered the flow of electrons, not conventional positive current.) The rotating rings are sandwiched between two magnetic plates, the function of which is to maintain the angular momentum of the current electrons oriented properly with ring rotation.

Analogous to the above example, each electron takes the place of the spinning wheel and the rotating arm is replaced by the rings. Thus, because electron angular momentum is a universal constant, an observer of the electrons in this rotating system will not have the same time as a stationary observer outside it. If each current electron has its "spin" oriented opposite the rotation of the rings, time on the rotating system would run faster than for a stationary ground observer, the same as in empty space relative to the Earth. The rotating system would therefore have the world line of empty space, not that of the gravity field. Its spacetime would not be curved. Since gravity is a time phenomenon, such a system in a gravity field could not have the behavior of a normal object. This conclusion may seem erroneous considering that ring rotation could never reach relativistically significant velocities, but we have an analogy with magnetism that is caused by a Lorentz contraction of the distance between moving electrons. If we considered the diminutive drift velocity of electrons alone we would never conclude that magnetism was possible, but the effect is *accumulative* over trillions of electrons. The hypothesis here is that the same is possible with the trillions of free electrons oriented with a spinning ring, to produce an accumulative time effect.

We now come to the controversial part of this theory because the above treats electron angular momentum the same as angular momentum of a normal physical object, whereas the electron is a quantum particle and quantum particles have their own realities with bizarre consequences when analogized with our macro universe. For instance, the electron has quantum spin number 1/2, and such a particle would have 720 degrees in one rotation, not 360. Nevertheless, the electron *does possess* di-pole magnetism. It *does behave* as a spinning ball with negative charge. Electrical Levitation is therefore an empirical theory. The property of the electron that gives electron di-pole magnetism cannot be an actual physical rotation, but whatever that quantum property, if it produces di-pole magnetism there is reason to expect it to produce other macro physical phenomena as if it were.

The theoretical finding of General Relativity that time runs slower in a gravitational field was confirmed by the Pound-Rebka experiment in 1959. **It is not that gravity causes time slowing, gravity *is* time slowing.** Since time and energy are reciprocal, more time on our rotating ring system would mean less energy seen by an observer in that frame of reference than seen by a stationary observer on the ground, the opposite of the red shift of a gravity field. This energy difference must equal the energy of the mass in a gravity field that is to be levitated, its energy of weight, and lost. Experimenters should therefore be aware that their device might radiate. But this is not a free energy machine. The energy of levitation comes from its magnetic field, and the electron magnetic moment energy turns out to be $V_e = mc^2$ divided by the electrical current, m is the mass to be levitated and c is the speed of light. This is an enormous amount, but becomes practical if enough electrical current can be sent through the rings. That means the electrical resistance of the ring material must be very low. For copper it is not. No material currently exists with sufficiently low electrical resistance at room temperatures, but **superconducting materials** exist that at cryogenic temperatures experience a dramatic loss of electrical resistance.

An experimental device can therefore be envisioned using a superconducting disc substituting for the electric rings. The disc would serve as a conductor for an electric current and would therefore need to be sliced along one radius with an insulator placed in the notch. Both edges of the notch would be connected to a power source by brushes rotating with the disc. With a counterclockwise disc rotation as seen from the top, to have clockwise electron "spin" the magnetic field between the plates would have to be up. Only the moving free electrons of the current will be available for any time alteration effect. These will also **produce a magnetic field, which it is reasonable to assume would be of the same time alteration effect as the electrons, thus producing the required alternate time "bubble".**

The resources required for an experiment using superconducting material at cryogenic temperatures are beyond this writer's means and an experiment has not been attempted, but an experiment at the Tampere University, Finland, 1992, using a superconducting disc suggests that gravity shielding is possible. Owing to the immense advantages gravitational shielding would give to the present interest in space exploration, effort toward its development would be logical. This essay may offer clues on how to begin.

Editor: I find the article by Wane McLeod quite interesting as his approach is very similar to the idea I first presented in my report at the "Space, Time and Gravitation" Conference which took place in St. Petersburg in 1998. The report was included into The Proceedings of the Conference, Part I, 1999. Before that the brief version of the article was published in English in the 'ELECTRIC SPACECRAFT' magazine, Leicester, North Carolina 28748 USA, Issue 27, 1997 p.30-31.

I managed to demonstrate that the ideas of Thomas T. Brown, especially his USA patent # 3,187,206 of 1965, are something more than just the force asymmetry in the electric capacitor. According to the concept presented in my report, by creating a matter property gradient (in particular, the dielectric permittivity gradient) we actually change the curvature of the electric force line in space. Normally, the natural space curvature accounts for the electric field potential decrease with the distance increase from the surface of the charged object. By creating the dielectric permittivity gradient (described by T.T. Brown in his patent of 1965) we change the natural distance potential gradient law. We can both increase and decrease this change and even reverse it. With the dielectric permittivity change square function, the natural space curvature is completely compensated and with more extent it is reversed and can be turned to negative. At that rate, the potential is not decreased but increased with the increase of the distance from the surface of the charged object. This is the essence of my concept that has never been considered before by any author. By creating the gradient described by the quartic function we get the same distance potential gradient law as in the natural conditions only with a different sign!

T.T. Brown, who discovered a force in capacitors with a special dielectric, offered the practical application of this concept. However, he did not find the source of this force, which, according to my concept, is conditioned by two factors: the corresponding dielectric property gradient function and the elastic properties of the dielectric material, which is of no less significance. Normally, dielectric particles are attracted to the charged surface but when the reversion of the curvature of electric field force lines occurs, they are repulsed from it and we can use this propulsion force. Elastic deformations (including those occurring during the pulsating operation of the field) account for the reaction forces equivalent to the generated propulsion force. Thus, the momentum conservation law is in action, however, the device is still moving.

T.T. Brown also considered other ways of generating a force, such as the material magnetic property gradient. Logically, by analogy with the electric field we can consider the gravity field, which can be "designed" and "reversed" by means of creating a matter with the density gradient. At that rate, the gravity potential must change in accordance with a certain law when the distance from the gravity field source being increased.

We should also note that the local space volume levitation effect with the accelerated or decelerated time rate was first described in my articles "Physical Principles of the Time Machine", NET #3 (6), May-June 2002 and "Practical Application of the Time Rate Control (TRC) Theory", NET #3, November-December 2001.

Our company conducts experimental researches on creating new materials that possess the properties described. We are interested in serious business contacts with companies in the aerospace industry to further discuss eventual cooperation projects in this field, including *co-patenting*.

Alexander V. Frolov

Perpetuum Mobile Not in the Past but at the Present

Vladimir I. Likhachev, Russia
Krasnodarskaya str., 38/20-34, Moscow, 109386

In several issues of the "Nature and man. Light" magazine in 1995-97 my colleague E.Oparin and I wrote on the unscientific character of absolutizing the law of degradation of energy, on the important issue of creating the perpetuum mobile of the second kind utilizing infinitely available *environmental heat energy*. We even send a letter to Yury S. Osipov, the Russian Science academy President, but received no reply. Notwithstanding many official discussions, there are no well-reasoned objections to our arguments or proof on the part of "high" official science, because, in fact, we are right.

I have recently received one more confirmation: bright and feasible data on the negentropy cycle (the cycle with spontaneous reduction of entropy) with a chemically active working substance that will be described later.

After the publication of Victor M. Brodyansky's book "Perpetuum Mobile in the Past and at Present. From Utopia to Science and from Science to Utopia", Second Edition, one can not but return to these issues. More so because the foreword of V.A. Fabricant, member of the USSR Academy of Pedagogical Sciences, claims the law of degradation of energy to be the "law of nature" and supports shamefully the deformation of the monothermists, especially P.K.Oschepkov, by academicians P.Kapitsa, L.Artsimovich and I.Tamm ("Pravda" of November 22, 1959), followed by E. Velikhov, A.Prokhorov and V.Sagdeyev ("Pravda" of June 22, 1987).

After the publication of Brodyansky's book I met the author several times (notably at a special seminar in The Moscow Energy Institute) and expressed my opinion about his book: the book avoids serious discussion of the monothermists, stresses their failures which are quite normal in the initial development of perpetuum mobiles of the second kind.

For example, Brodyansky knew about Tsiolkovsky's discovery. Guay's book "On the Little-Known Hypothesis of Tsiolkovsky" is listed among the First Edition references. But Vladimir Brodyansky avoided analyzing this discovery both "in the past" and "at present". The Second Edition references even list the "Physical Thought in Russia" magazine, issue #1 of 1991, that contained for the fist and only time the modern edition of the "Law of Degrading of Energy" by Tsiolkovsky. "At present" V. Brodyansky knows my analysis of Boltsman's mistakes but he paid no scientific attention to my analysis either.

In Brodyansky's book the description level of chemical reaction cycles is even lower than that of Schpilrine. Schpilrine at least tried to present his arguments but here we see unsupported statements "justified" by the phrase "the detailed discussion would take too much space". Vladimir Brodyansky's analysis of anti-Stokes luminescence (pp.216-219) is also superficial. Dissipative and entropy processes prevail in it. But if there is evidence supporting higher frequency quanta than the frequency of radiating flux, **it is the evidence of negative entropy**. Vladimir Brodyansky not only is familiar with it and also shows it by the example of investigating the Sun and the Earth (p.247).

Vladimir Brodyansky's analysis of Rank's tube (pp.235-237) displays the same tendency. In the Rank's tube the dissipative processes also prevail. Its effectiveness as a refrigerator is lower than in the traditional schemes.

But one can not deny the existence of negentropy processes in it. It is these processes that "originate" the cold gas flow. And Brodyansky's "origination" on p.236 "The vortex tube in fact divides the coming gas into two fluxes: heated and refrigerated" - is totally false. That can be performed only by "Maxwell's demon" that does not exist in substance.

The refrigerated flux in the Rank's tube is formed by the "Tsiolkovsky's flux" by way of heat transfer "from cold to hot", from the tube center to its circumference by **centrifugal accelerations**. On its basis (to be more exact, on the basis of Finko's tube) the perpetuum mobile of the second kind can be created. And we have this engineering solution. E.Oparin and I made the corresponding patent application for an invention but the Federal Institute of Industrial Property again refused to consider such applications. Here one can see similarity with the causes of re-edition of Vladimir Brodyansky's book. And one need not be a big politician to admit the reality and understand the sources.

Environmental and, especially, anti-nuclear movement is invincibly expanding. Danger to human life and threat of ecological catastrophe are quite real. The Second Edition of Brodyansky's book protects behind scientific arguments those who use Russia as a source of easy money and then flee.

But we like our country - vast and austere. This austerity and space form REAL PEOPLE and support humanity and materialistic humanism of the energetic and talented, not just the rich.

In modern Russia the powerfully spread alcoholism, debauchery and drugs are the instruments of slavery. But the monothermy and development of alternative power engineering will become basis for Russia's revival and unprecedented prosperity.

Russians learn slowly, sometimes, very slowly, even disgracefully slowly. But we will learn. And then no one will stop us. The following tips are meant for those who are ready to risk their efforts and facilities in order to build alternative energy knowledge base: dissipative and negentropy processes are inseparable and very often their mechanism is the same. In our environment the dissipative processes are more effective and hide the negentropy ones. No dissipative process combination can generate the negentropy cycle. Look for examples in Brodyansky's book. To create the negentropy cycle and use the environmental energy at least one negentropy process is required that has effectiveness and negentropy exceeding entropy growth at all levels of the cycle.

News from "IntAlek" Company

William Alek [alekws@intalek.com] has updated the latest and greatest ZPOD system drawings:

hhttp://www.intalek.com/Index/Projects/SmartPAK/Projects/ZPOD/ZPOD_System05.pdf

This is release 5.0

Here is the latest electrical schematic: http://www.intalek.com/ZPOD/ZPOD_System05.gif

Based upon his "preliminary" tests thus far, he estimates the COP is around 2.0. This estimate is interpreted from the following scope traces:

http://www.intalek.com/ZPOD/in.GIF

Alan Francoeur's Generator
Canada

Alan L Francoeur, Canada
Email: al.f@shaw.ca

This is my writing and development about the Interference disc generator, and the permanent magnet dynamo machine, and other related inventions. I have always been intrigued with electricity and magnetism and aetheric energies starting at a young age. Back in 1980 I wanted to do something to help get my large vehicles better fuel economy. With this goal in mind, I experimented and designed my first heat exchanger vaporizer system to do my part to clean up our cars and trucks, fig. 1, 2, and 3 are two earlier vaporizer units under construction. For more information about this fuel system, please refer to my vaporizer fuel system report. The main body of this writing is referring the permanent magnet dynamo system that has been in the making since 1987, starting after the development of the interference disc generator.

Fig. 1

Fig. 2

Fig. 3

I invented the Interference Disc generator concept while I was employed at Giant Yellowknife Mines NWT from 1983 to early 1986. The idea of using metal blocking plates to shield magnetic fields stirred in my mind while working around heavy electrical mining equipment, and 1000hp and other electric motors with exposed coils that have heavy metal around the outer casing. I did a search at the time and found John Ecklin's work very interesting with his application of the shielding effect. During my off time in Yellowknife, I was experimenting with the interesting effect of producing induction with a stationary coil mounted beside a stationary magnet with a small air gap maintained between the magnet and the coil. During the construction of my first disc machine, I called it the Interference Shielding effect because of the interference the blocking disc does to cause induction, thus the Interference Disc Generator fits the description. The first transformer coils I used to test the interference shielding effect were used from old radios, and the magnets were alnico types. Back in 1986 in Calgary Alberta, a friend Bud Johnson and I constructed the first test disc generator model using alnico magnets, fig. 4, and later ceramic magnets were installed. Months later I designed and began construction of another larger Interference Disc machine with horseshoe shape neodymium 35 grade magnets mounted on the outside, and with the N and S magnetic poles facing toward the coils in the center portion of the machine. The coils and magnets in this larger machine are mounted in reverse of the first Interference Disc generator, fig. 5, 6.

Fig. 4

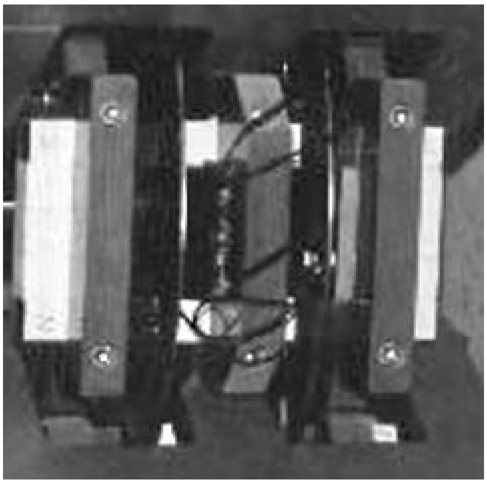

Fig. 5

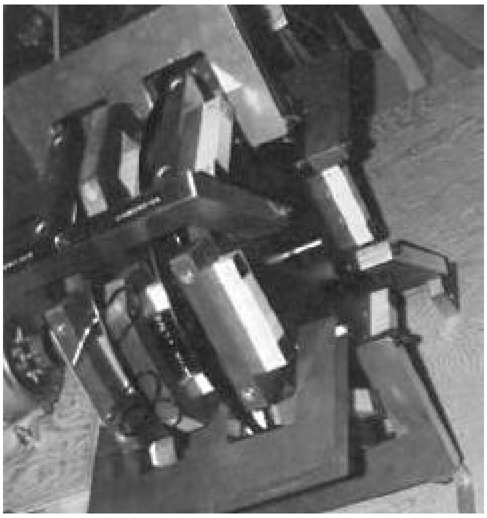

Fig. 6

General description of the interference disc generator

The Interference Disc electrical generator is made of a stator (114) consisting of an array of even number parallel mounted bar magnets (112) arranged and supported (114) in a circular fashion equidistant from each other, where the polarity of the magnets (112) are alternating when viewed at either end. Fig. 7

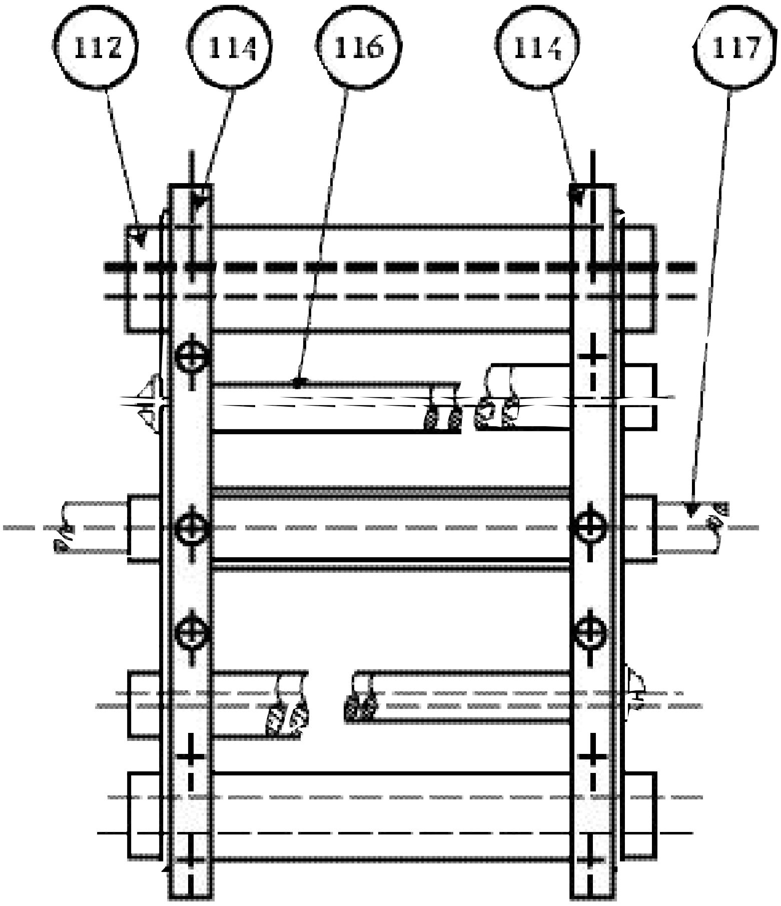

Fig. 7
Magnetic bars assembly (rotor)

Also in the stator fig. 8, (102) there are twice the number of coils (111) as bar magnets (112) wound in pairs on U-shaped cores (110). One half of the coil pairs mounted at one end of the bar magnets (112) in the same circular fashion, each corresponding to and aligned opposite to a pair of bar magnets (112) and separated from it by an air gap. On the other end of this pair of bar magnets (112) there are another set of coils (111) aligned likewise.

The rotor is made of two magnetically susceptible circular plates (109A or 109B) (such as iron or steel) Fig. 9, centrally mounted on an axle (107) fig. 7, which is in respect to the array of magnets and coils Fig. 9a.

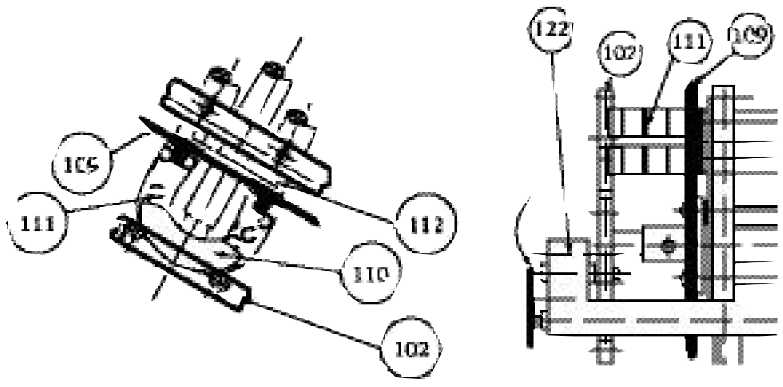

Fig. 8

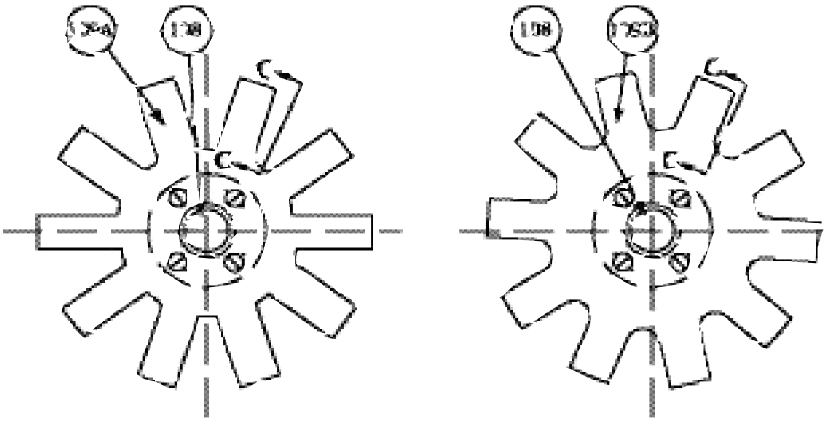

Fig. 9

The interference discs (109A or 109B) are separated from each other so that they can rotate in the air gaps between the magnets (112) and the coils (111). The interference discs (109A or 109B) have a number of **equidistant opening, either more or less in number than the number of the bar magnets (112)**.

The arrangement and size of the openings are such that when the discs (109A or 109B) are rotated by an outside motive force, they open the magnetic field between adjacent bar magnets (112), and the opposing coils (111) at both ends of the stator (114) simultaneously. Hence inducing an alternating electric current in the stationary coils (111) fig. 10.

By opening and closing the magnetic influx to the core (110) of the coils (111), an alternating current will be generated in the coils

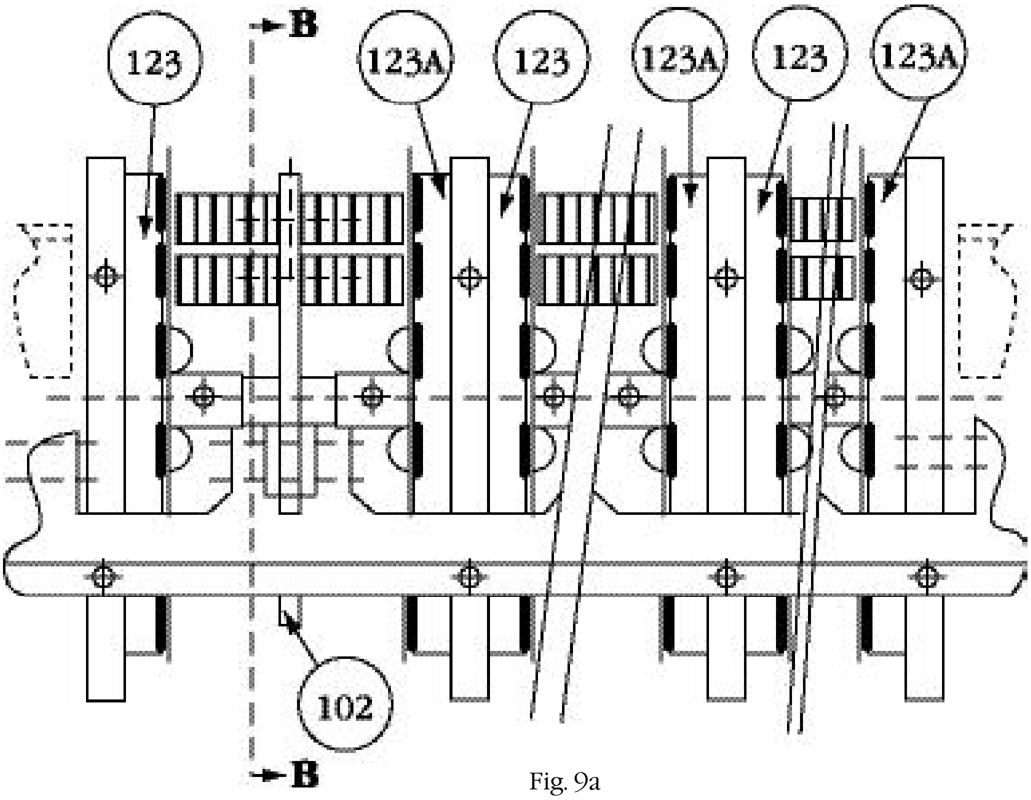

Fig. 9a
Magnetic bars assembly (stator)
Option II: multable magnetic coil-disk assembly

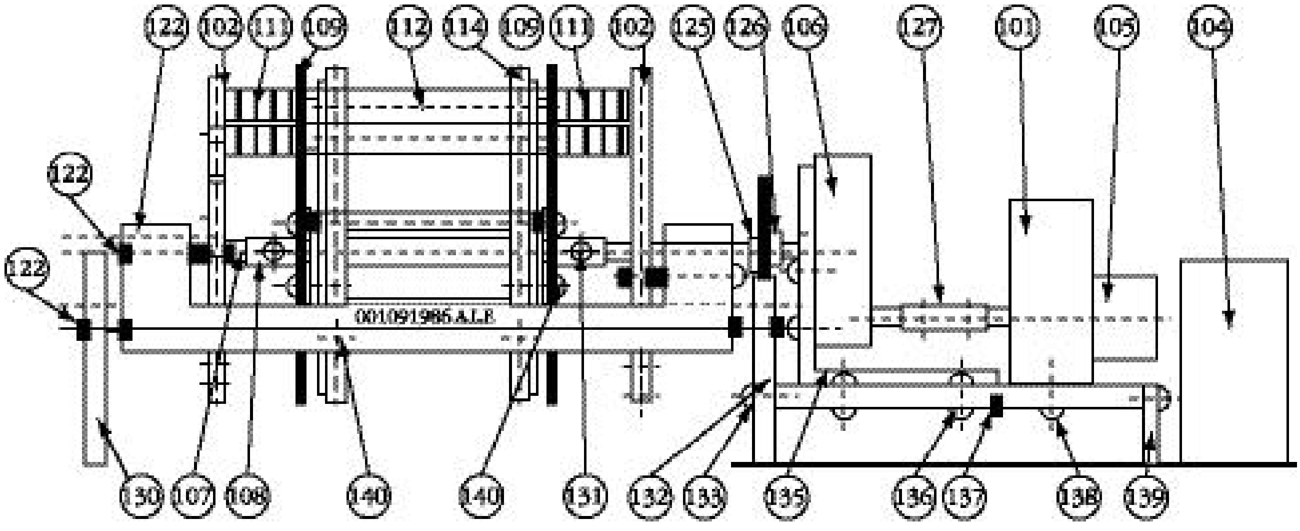

Fig. 10

(111) in a retro-order to the rotational direction of the interference discs (109A or 109B). Depending on the number of magnet-coil assemblies three phase AC current can be obtained. The regulation and attenuation of these currents can be achieved by known electric engineering methods (Fig. 10).

The efficiency of the unit would vary greatly depending on the speed of rotation, the width of the air gap, the strength of the magnets, and the materials used to construct the coils. I used nickel base amorphous metglass in the coils in both the smaller and larger machines, which as a result, shows greater efficiency as compared to conventional core laminates. Less energy is needed to spin the balanced disc rotor while causing a larger moving magnetic field to cause induction in the stationary coils. With this effect, it is easy to visualize the advantages this design has over conventional generators and motors. New testing of the coils will be conducted after the machine goes through all the upgrades to eliminate the problems with the warping interference discs.

The Interference Discs Generator described in this document **has the potential to produce more than unity effects** by easy turning of a magnetically balanced (non-warping) interference disc. Optimizing the device with closer tolerances and correctly selected materials and incorporating internal magnetic balancing, will add to the success and improved performance of this unique generator design.

Improvements can be made with this design while the operating principal remains the same. Looking at the photographs included you will see the 1st proof of concept generator frame is constructed mainly from aluminum, this metal was used only because it was inexpensive material and easy to work with fig. 11. Non magnetic and non-conductive materials should be used to construct the generator frame supports to eliminate the eddy current losses. All of the nuts and bolts used to hold the disc generators together worked well for making the test machines capable to have adjustable air gaps, however, they can all be eliminated with design changes that preset the gaps and tolerances to optimize the conditions of the interference disc effect.

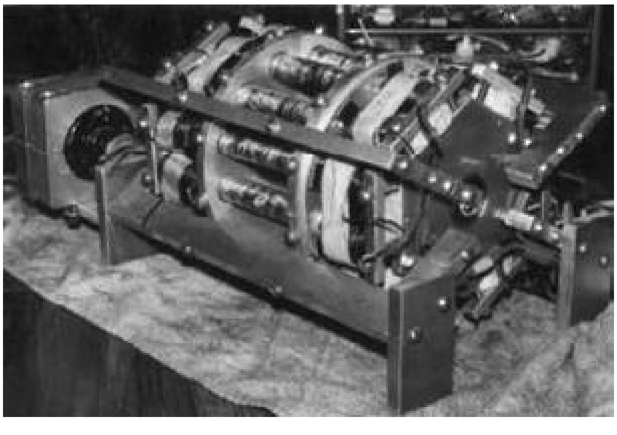

Fig. 11

Many improvements were adopted in the next larger #2 version of the Interference Disc generator fig. 12. Phanolic materials (non conductive and non magnetic) were used in the frame supporting the magnets and the coils, thus eliminating eddy current losses within the frame support. Large aluminum E bars were used to support the magnet frame support and the coil frame support with no noticeable eddy current losses. The aluminum used is positioned further away from the coils and magnets in the larger generator so as not to be affected by the magnets. Larger more powerful neodymium magnets and I shaped transformer coils were installed and tested in the larger machine which showed an increase in output power over the 1st machine fig. 13. Using more powerful neodymium magnets in this larger disc machine created some problems with the interference discs. Not only did they warp the discs they caused them to vibrate and chatter causing the fins to hit the magnets and the coils. I increased the air gap to over half an inch and moved the disc further away from the magnets to prevent damage. The effect was a reduction of magnetic flux reaching the coils that reduced the power output. The discs must be constructed much stronger and rigid if using neodymium magnet grade 35 or greater, then the air gap can be closed so more magnetic flux will reach the coils allowing for a greater electrical output.

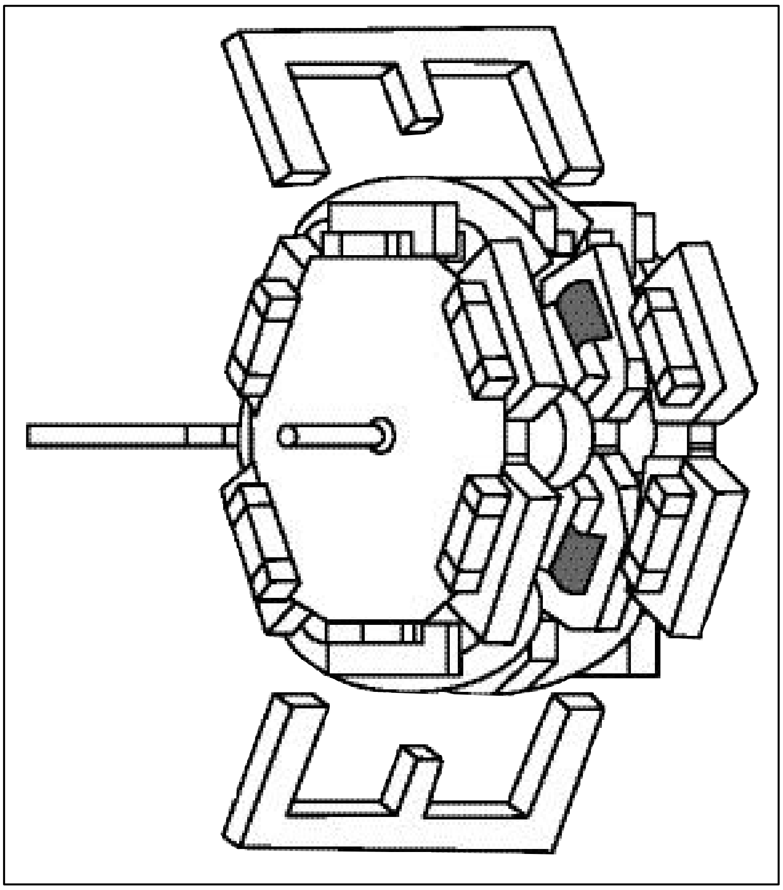

Fig. 12

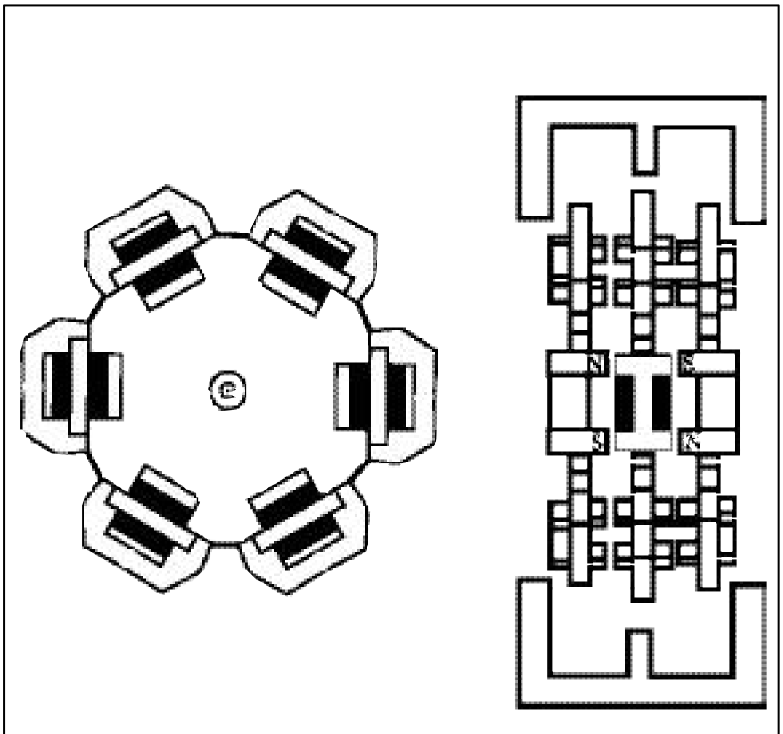

Fig. 13

The wiring of the coils circuit can be constructed by known engineering methods to achieve AC or DC power, and like any conventional transformer, wind the coils with heavy wire for more amps and more turns for more volts. The coils are stationary in this machine which makes it easy to harness AC power directly from the coils without brushes. All coils that are in phase can be wired in series or parallel, the number of the blocking plates on the interference disc determines the phasing of the coils.

The diagram in fig. 14, 14a shows the path of the magnetic flux movement in relation to the position of the blocking plates of the interference disc. The magnetic field in the core/coil is switching polarity as the interference disc fins alternately moves in and out of the air gaps on each side of the coil. As you can see, depending on the position of the disc, positive and negative induction takes place for AC output of the coils. John Ecklin has achieved this effect using rotating blocking plates. I have achieved the same effect using balanced rotating interference discs with multiple stacks of isolated coils and poles mounted in a circle equal distance from each other.

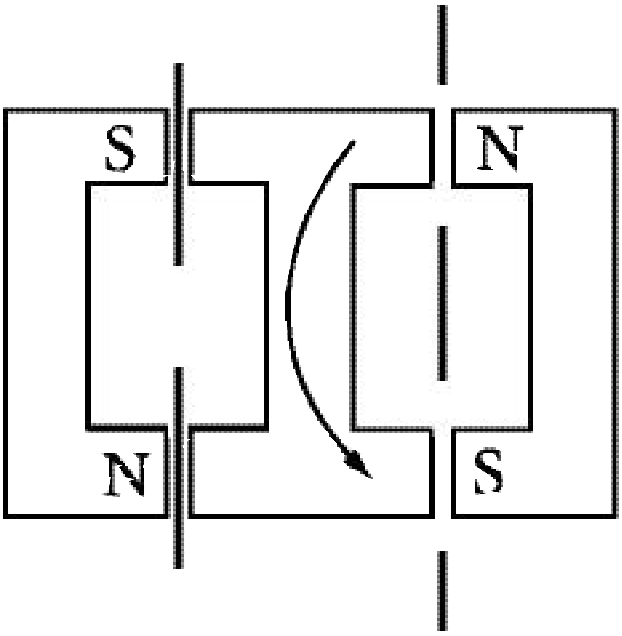

Fig. 14

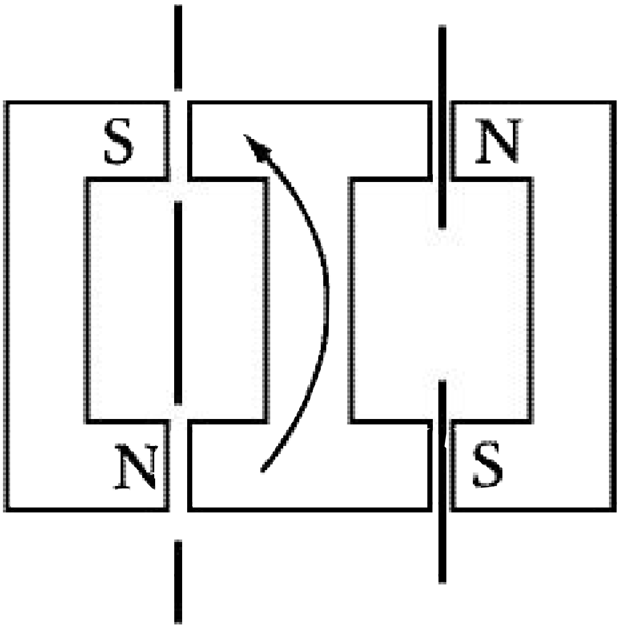

Fig.14a

The effect I reproduced with Gary Wesley's permanent magnet motor is related to the Interference Disc generator in the fact that blocking plates are used to shield the magnetic field, only we differ on methods to balance the blocking shields. Gary Wesley used springs to carefully balance the blocking plate, and I used a different number of interference fins working with the magnets to balance the blocking plates. The magnets in my Interference Disc machine become the springs.

More photos of both Interference Disc machines are shown in fig.15, 15a and 15b below.

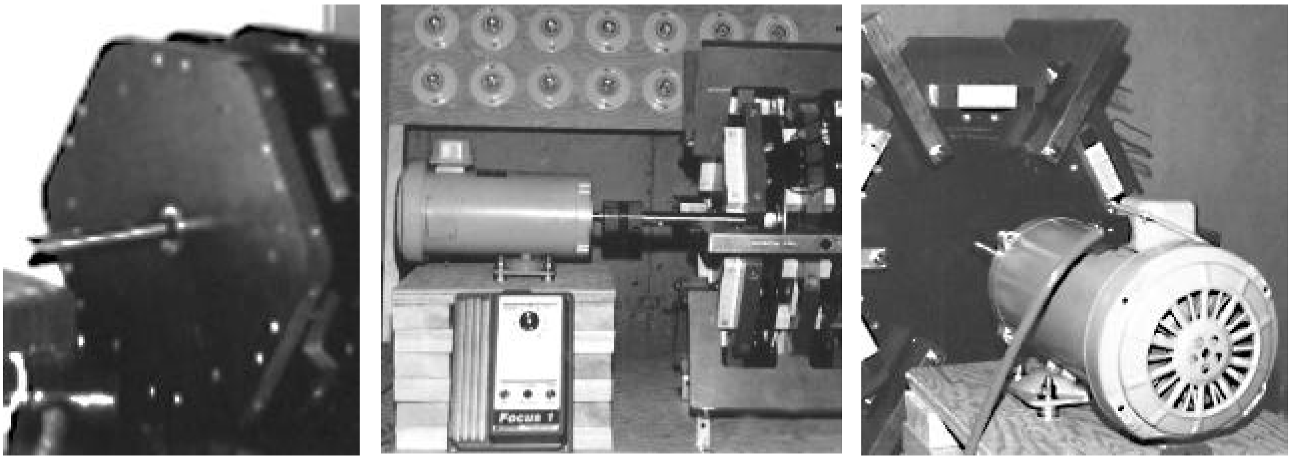

Fig. 15

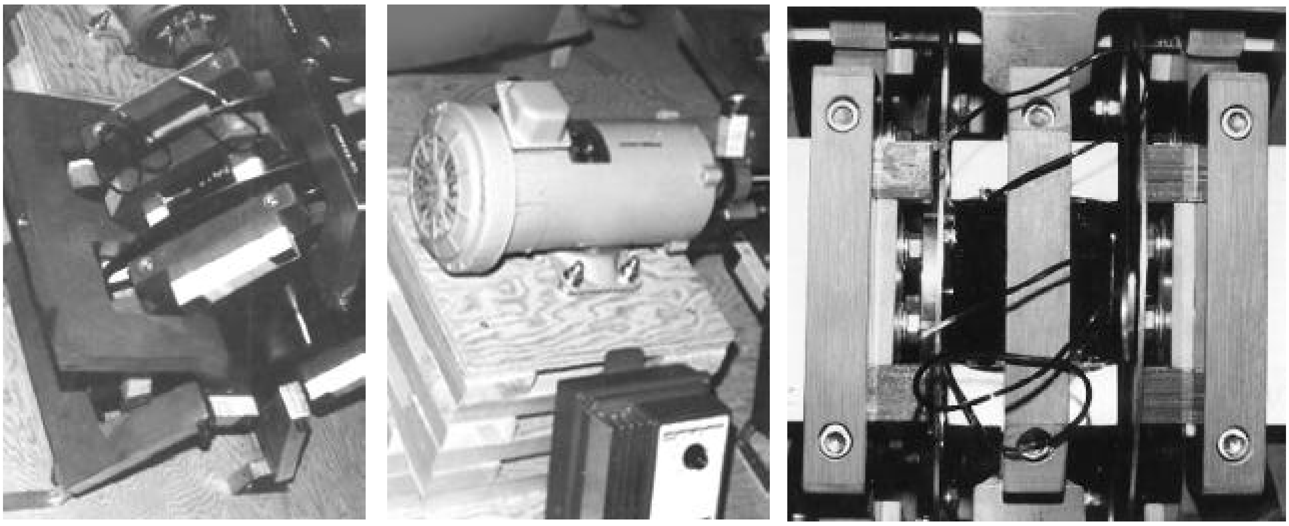

Fig. 15a

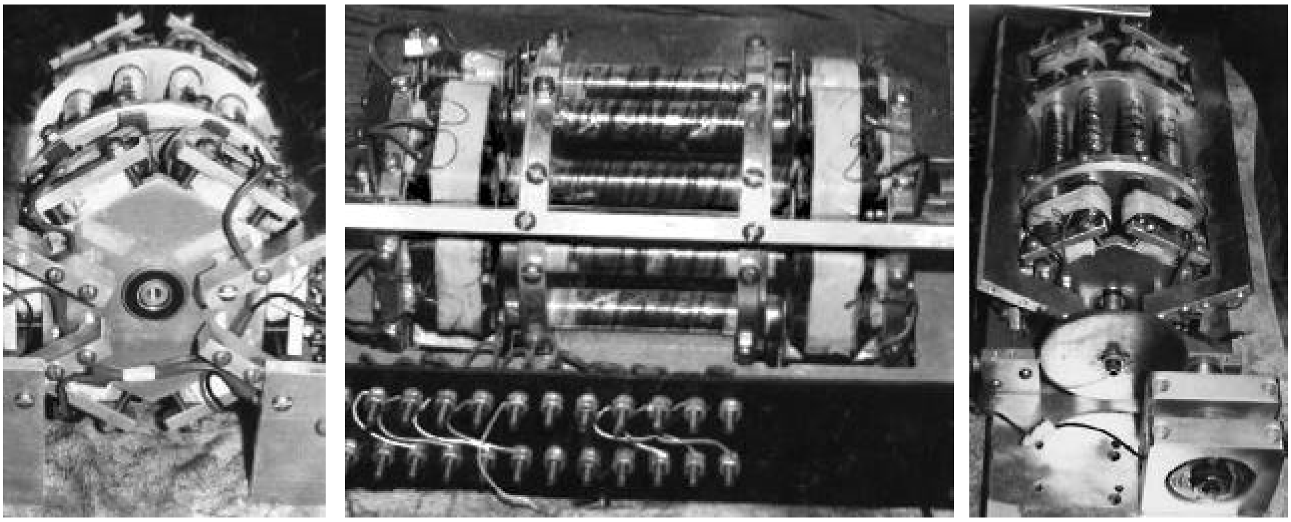

Fig. 15b

Closing comments on the disc generator

In closing this discussion of the Interference disc generator, I would like to go over and point out some of the concepts that I incorporated into the design of my permanent magnet dynamo.

As I made further progress and completed the later stages of the Interference disc generator, it became clear to me to use certain concepts of the machine and incorporate them into the design of my multiphase permanent magnet dynamo. Some of these concepts are explained here.

I have constructed several interference discs that have a different number of blocking shields, some with 10 blocking fins and some with 11 blocking fins and others with 12 blocking fins. They are all interchangeable on the machine to change the phase relations between all the isolated coils. The 12 finned disc will produce a single-phase dynamo with all the coils in phase, and the 10 finned disc on the machine will produce a three phase dynamo. One phase represents four groups of isolated coils, and the second phase represent the other four coils, and the four coils left are the third phase, for a total of twelve coils. All just by changing the number of fins on the interference disc and or adding more layers. Using 12 magnet poles and the 10 or 11 interference fins on the disc, the rotor discs becomes magnetically balanced and easy to rotate in spite of the load on the coils. The 12 finned disc has a larger holding force because of the same number of fins as magnetic poles. However, it can be reduced to almost nil by adding more layers of interference disc and balance them against each other. Many in phase coils can then be added in series or in parallel to build your power levels to any desired potential.

Both machines have stationary coils, and the disc generator has 12 magnetic poles (6N and 6S), and the permanent magnet dynamo also has 12 magnetic poles (6N and 6S). The disc generator uses 10 or more fins on the interference shielding plate per rotor, and the permanent magnet dynamo has 10 or more coils around each rotor. There is the same number of interference blocking plates as there are induction coils on each of the two difference machines and the number of blocking fins and or coils are interchangeable. The back emf effects are different on both machines, the Interference disc generator demonstrates easy turning of the discs while the stationary coils are loaded, and when producing alternating current the permanent magnet dynamo rotor slows down when a load is placed on the stationary coils. Both machines produce AC power out of the coils and the Interference Disc generator has a closed magnetic circuit through the coils, and the permanent magnet dynamo has open magnetic circuit through the coils.

The Permanent Magnet Dynamo

I would like to begin this topic at this point by describing my permanent magnet dynamo machine.

I designed this machine to be completely interchangeable into many different motor generator combinations. These combinations involve different number of magnetic poles and different number of interchangeable coil poles in the same machine, two versions are possible, one with isolated coils and another with the coils that share a common core fig. 16.

We can change the magnetically balancing effect (*holding force*) by changing the number of coil poles or (*interference disc fins*) in relation to the magnetic poles. This turns the machine into a multipurpose, multiphase, variable output, and variable frequency ac dc permanent magnet motor generator system. I first designed this unique multi purpose dynamo in 1987 after I built my third Interference Disc generator.

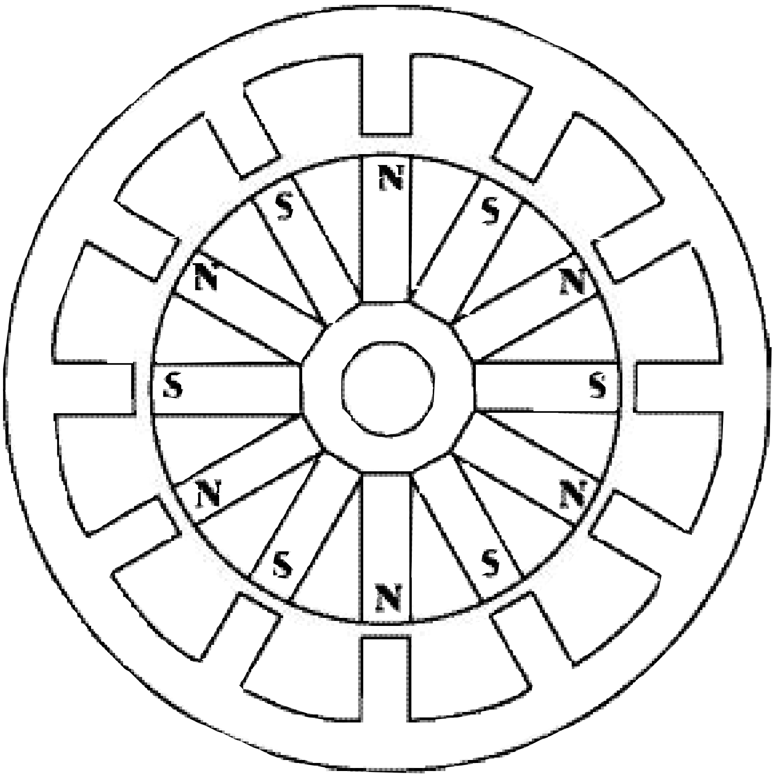

Fig. 16

Dynamo description:

Rotor:

It is well known that a bar magnet has its magnetic field concentrated at the N and S pole ends, and the half way line between the poles of a magnet is the neutral zone, this is the region where the magnetic field is canceled or neutralized. The *ferrous* keeper ring I designed for this rotor is mounted near the axle and it has machined flat decks so the magnets attach flush to the deck of the

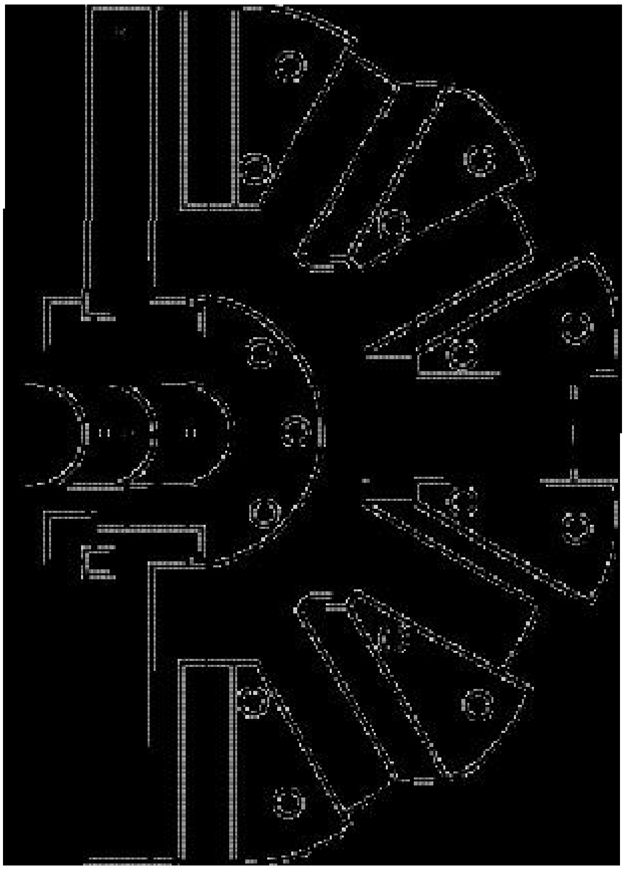

Fig. 17

keeper ring. Because of this feature, flux leakage is reduced at the keeper ring and the magnetic field concentrates at the rotor poles. With the poles of the magnet positioned near the axle and attached to the keeper ring in the rotor, the magnetic field that was there is now forced to add magnetic field density to the outer magnetic poles on the rotor rim. The density of the magnetic field at the poles is now greatly increased which makes a very powerful small size rotor. The *ferrous* inner keeper ring (*active iron*) that is installed inside this rotor design, is also the *neutral zone* of the magnetic field when viewing all the combined permanent magnets in the rotor. Shown is a cut away view diagram of the rotor, fig. 17.

I utilize rare earth rectangular shape neodymium grade 35 bar magnets with the dimensions of 3" long and 1"x1" on the end poles. There are 12 bar magnets mounted 30 degrees apart in NSNS configuration, and the bar magnets are attached to a precision machined inner keeper ring which makes the active iron the region of the middle of the magnetic field (the neutral zone). Another way to view this is to imagine two 3" long bar magnets attached together, you now have a 6" long magnet. Bend this magnet in half and install iron at the U and it now becomes a horseshoe magnet with two 3" long pole ends. It's a powerful rotor design in the fact that each of the 3" long bar magnetic poles has the field strength of a 6" long bar magnet. More poles on the rotor will allow for higher frequency ac generation at lower rpms, twelve magnetic poles on each rotor is shown in fig. 18, 19.

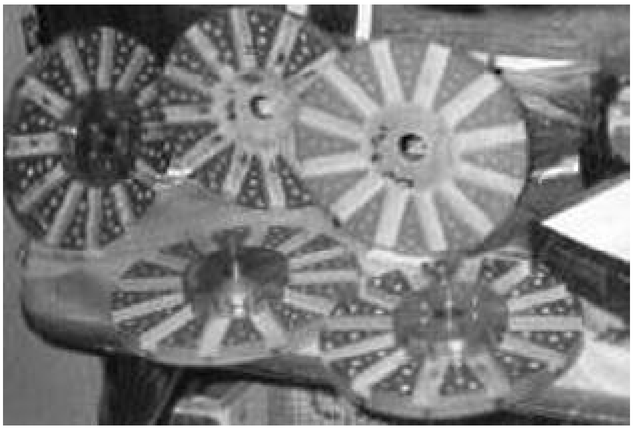

Fig. 18

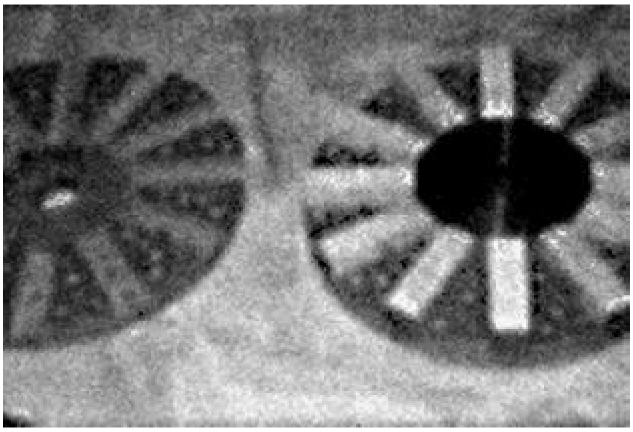

Fig. 19

The drawing in fig. 20 and picture in fig. 21, shows twelve magnet poles in each rotor in relation to the position of the coils. Ten *amorphous* cores are shown mounted around each of three rotors before they were wound with magnet wire fig. 21. The machine I have constructed houses a total of three rotors side by side that are mounted 10 degrees apart in relation to their *magnetic poles*, each rotor has twelve magnets installed for a total of 36 magnetic poles shown infig. 20. This dynamo concept can be constructed to any size with unlimited numbers of rotors and coil combinations to fit any power requirements.

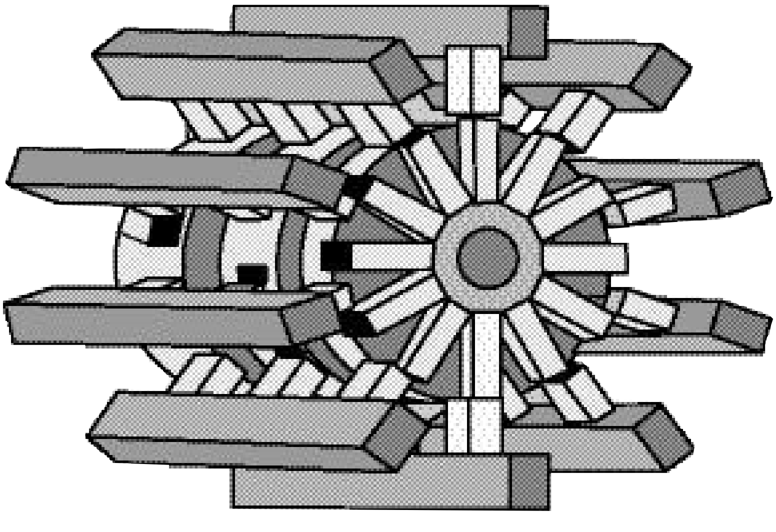

Fig. 20

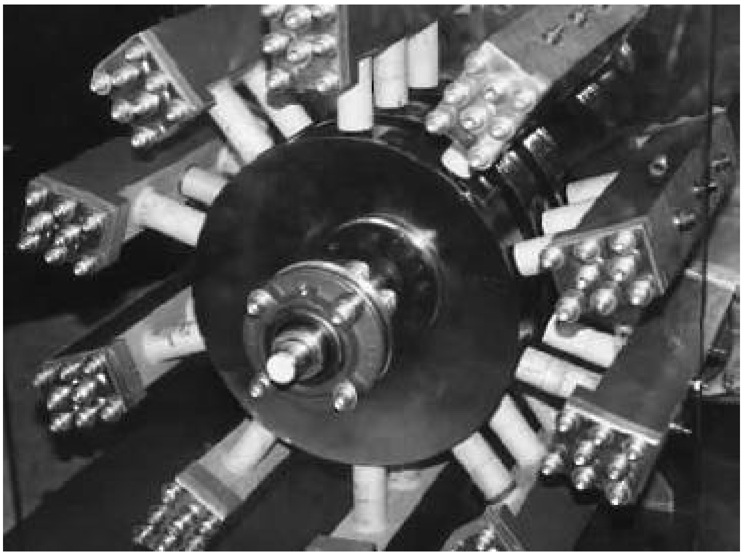

Fig. 21

It is well known that a magnet exerts a force of attraction to iron bringing them closer together, after they have come together, they are in their **static position**. A force must be exerted to either the magnet or the iron in order to pull the magnet and the iron apart, removing them from their static position. The term **static position** is used to describe the point when the magnet and the iron bar are at their nearest point.

As in the Interference Disc Generator, this Permanent Magnet Dynamo I am describing also uses an internal magnetic balancing effect which reduces startup torque on the rotor.

In this system, one set of magnets is in their static position in relation to the coils, such as the impulse coils. Another set of magnets on one side of the rotor is not in the static position, and there is another set of magnets that is not in its static position on the opposite side. One side is half in to the coils, and the other side is half out the coils, allowing the rotor to have minimal start up torque in turning the rotor. The magnets function in two jobs, the first with its velocity is to cause induction into the transformer coils producing ac power.

The second function, is the opposite pair magnetic fields on the rotor balancing each magnet out of their static position in relation to the coil positions at any given time, they cancel their holding force which puts the rotor at equilibrium. This reduces the amount of start up energy required to rotate the rotor resulting in higher efficiency. It allows the rotor and or multiple rotors to turn easily and efficiently despite the inherent powerful holding force of the rare earth neodymium magnets. With the three rotor machine configuration, I have connected 6 impulse coils in three phase to function as the motor impulse circuits with a hall effect brushless motor controller fig. 22.

Star wound three phase motor impulse coils.

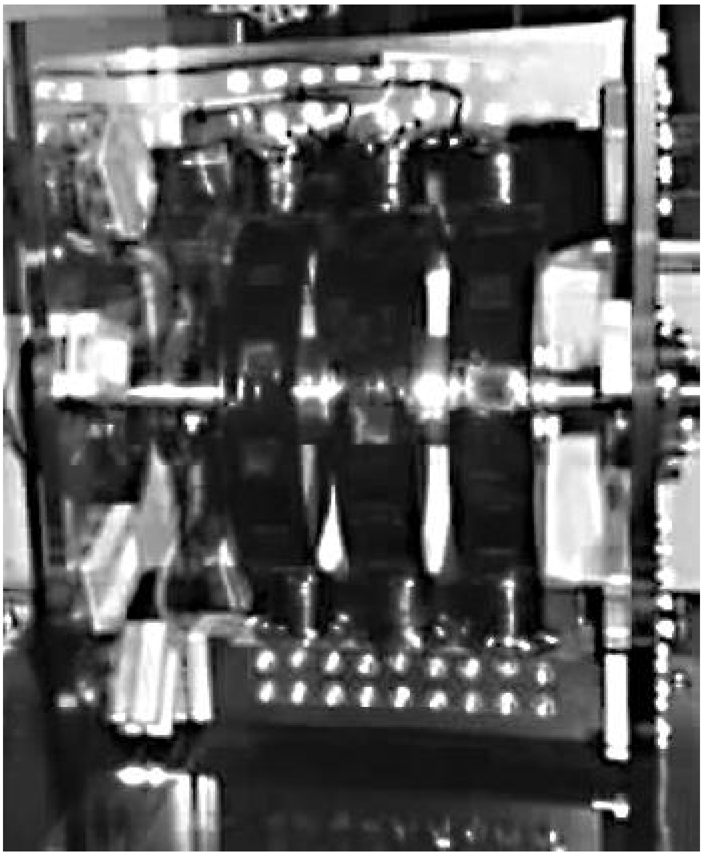

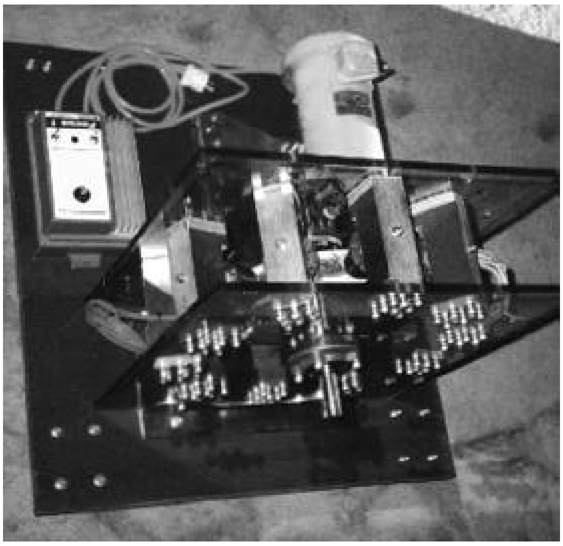

Fig. 23

Fig. 22
Star wound
three phase motor impulse coils

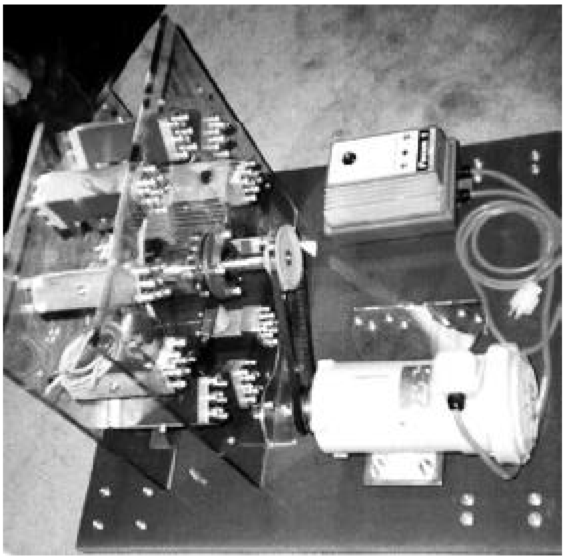

Fig. 24

Two more pictures of the one rotor dynamo option are shown with a 1hp dc motor turning the rotor with an 800 watt load on the coils, six 100 watt AC light bulbs, and two 100 watt 12 volt DC light bulbs, Fig. 25, 25a.

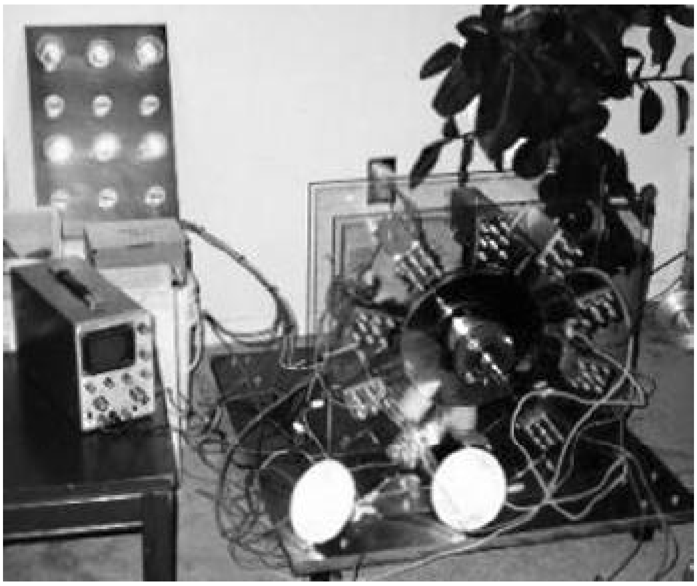

Fig. 25

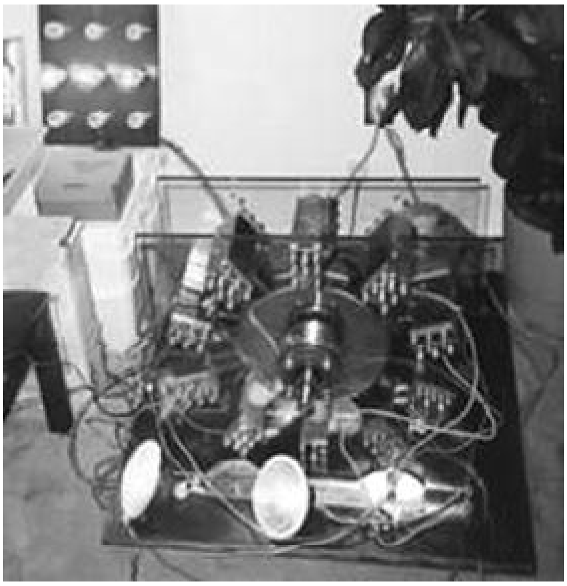

Fig. 25a

The efficiencies of this dynamo are so far inconclusive with regards to producing an endless self running machine. Our research is continuing to improve the efficiency by eliminating all conventional type ac and dc motors from this dynamo, and incorporating high voltage dc impulses of very short time duration's as Ed Gray did with his technology years ago. However, a description of my early self running attempts is included near the end of this dynamo report. These tests have manifested an abundance of heat in the 24 volt dc drive motor.

Another picture of the three-rotor dynamo configuration is shown with 12 regenerating ac to dc coils, and with the motor impulse coils removed fig. 26.

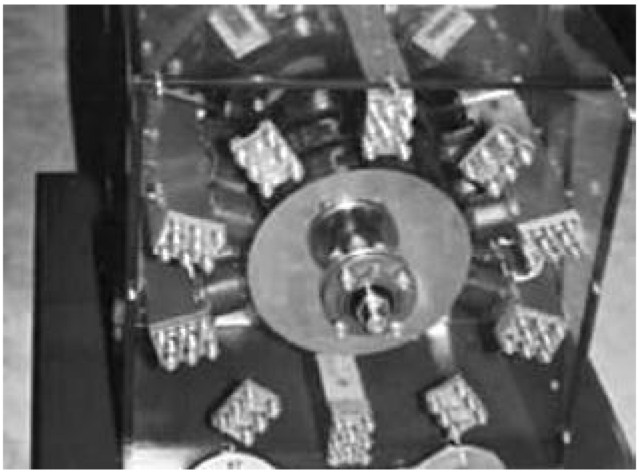

Fig. 26

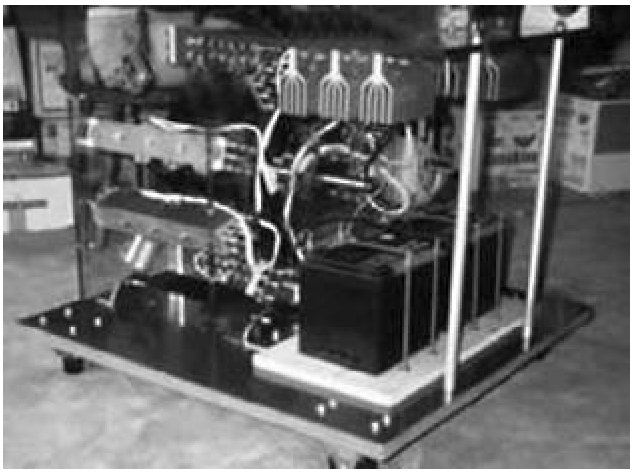

Fig. 26a

Another close up of the regenerating coils and three phase motor impulse coils is shown on the dynamo fig. 27.

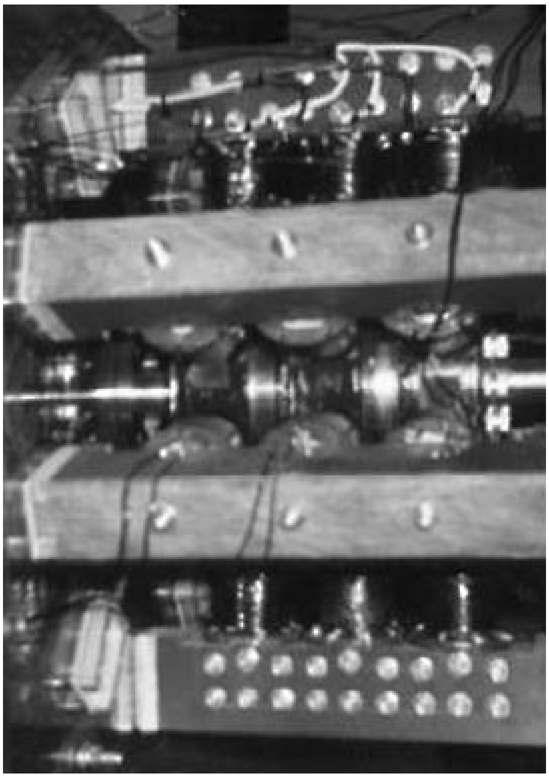

Fig. 27

Tests to produce an abundance of heat

I have testing the machine so as the output coils are connected to regenerate a 12 volt battery as the motor is turning the rotors. I used a bran new conventional permanent magnet 1/4 hp 24 volt dc motor that is running on 12 volts to turn the 110 pound 3 balanced rotors, and or the 35 pound 1 balanced rotor version.

The regenerating output coils produce AC to DC power with no noticeable ripple on the scope. The generated direct current from the coils is connected to go back into the battery, in which the motor uses to produce the torque needed to rotate the rotor. As this happens, the regenerated dc power from the coils is again going back into the battery first, and then to the motor. The result is that the battery supplies less power to the motor when the regenerating DC current is going into the battery. The motor takes on a load from two sources, one is from free wheeling the rotor with out loading or drawing current from the coils, and the other is the load induced on the motor when the coils are loaded to produce current.

The dc permanent magnet motor has shown that during one of many switch mode tests, it consumes 15 amps at 12 volts from the battery while a chosen number of regenerating coils supply ? the amount of current going back to the battery. The battery supplies 50% of the amps and the regenerating coils supply the other 50% amps. This effect can be changed depending on how many open isolated regenerating banks of coils are activated and loaded.

This test shows a 15 amp draw the motor while it is turning the rotors under load while regenerating current from the coils. It may be they balance each other and both battery and coils supply 7.5 amps each that make the total 15 amps draw on the motor. This accounts for the 15 amps total load that is shown on the dc drive motor. It produces an excessively large amount of heat in the dc motor that is useable output energy. On a larger scale the heat generated in a given motor can be enormous. More accurate measurements are needed to determine the true energy in verses energy out.

However, this electric heat pump effect can be altered to produce more or less heat depending on the size and type of the drive motor, and the amount of regenerating coils that are placed under load. Cooling tubes wound around the motor with water flowing through them is a way to pull the heat out for useful applications. Hot water from the drive motor and electric power from the coils at the same time.

After running the machine, the amorphous regenerating coils remain cool to the touch where as the conventional dc motor gets very hot with higher amps going through it. The machine also has a longer running time when the coils are regenerating the battery, as compared to running an outside equal load with the same amp draw, in that case the battery drains much faster.

This sheds light with the observation of high heat build up in the motor side (less efficient side) of the circuit while being cooler on the generator side (amorphous side) of the circuit. After hours of run time the battery begins to drain, when this happens, the battery begins to show tangible temperature changes from one side of the battery to the other. The whole negative side of the battery remains cool to the touch (temperature drop), while the whole positive side of the battery gets warmer to the touch (temperature rise). The effect is repeatable and only happens when the regenerating coils are connected to charge the battery as it is running. It may be approximately 20 degrees difference in temperature, depending on how many regenerating coils are activated sending its current to the battery. There must be exothermic reactions and endothermic reactions going on inside the battery motor generator combination that is causing this temperature difference. The load and heat generated can be changed the on the motor from the flick of a switch that activates and loads the output coils, which in turn dumps it's energy back into the battery pack or outside load. The efficiency ratio of the machine changes depending on how many and which banks of regenerating coils are activated, which again also effects the temperature difference manifesting on each side of the battery, motor, and amorphous coils.

Other tests

An Automotion dc brushless 3 phase motor controller with hall sensors was used to test the machine in fig. 26a. Tests have shown the 1hp dc motor and the three phase brushless dc motor controller will not be the answer to closing the loop on this motor generator machine. The brushless controller has a 50% duty cycle which will not manifest the electro radiant effect, and it also has a constant draw on the battery as a regular dc motor does. It is much less efficient than using very fast and short unidirectional dc impulses in the motor impulse coils to kick the rotor around. This brings the subject of Nikola Tesla's magnifying transmitter and Ed Gray's radiant energy impulse motors to light, and a possibly answer to closing the loop for a real auto-rotating machine.

Before I go further with this, I want to point out the many phase possibilities this dynamo has. If you look at the coil arrangements in the 10 coil machine, you can clearly see the coils that are in phase with each other. When a second and third rotor is installed with all the coils attached, one pair of coils on the first rotor is in phase with another pair of coils on the second rotor, and another pair of coils is in phase on the third rotor. With the ten coil to twelve magnet pole rotor assembly, the machine has five phases of isolated banks of coils. If twelve coil mounting bars are attached around the three rotors, and the three rotors are spiraled off set by 10 degrees, you will have a powerful three phase dynamo. If the rotors are not off set by 10 degrees and are lined up in phase with each other, all the coils on the machine will be in phase for a powerful single phase dynamo. The volts and or amperage can be tailored for an application by the winding of the transformer coils, and the machine produces 60hz at 600 rpm. This dynamo machine concept is also interchangeable into many other motor generator combinations, and sizes, and any volts or amps requirements, it all depends on the needs of the end user.

Cooling tubes wound around the motor with water flowing through them is a way to pull the heat out for useful applications. Hot water from the drive motor and electric power from the coils at the same time.

It is possible to have super efficiencies with this system using unidirectional hv dc impulses of very short duration's at the make brake contact at the closure of the switch.

I have acquired two of Ed Gray's radiant energy impulse motors #4 and #5 to revive an almost lost technology. I am proposing to combine a special permanent magnet dynamo with less back EMF that is tailored to produce 5kv dc under 2000rpm. This dynamo will energize a 5kv dc capacitor bank and replace Gray's 12 volt to 5kv dc converter. The energized capacitor bank will discharge through the conversion tubes in the very same manner as in Gray's original circuit, only the capacitors will get energized from the high voltage dc dynamo instead of from a battery through an inverter. This version of a high voltage dc dynamo will be connected to and turned by Ed Gray's EMA motor, a starter motor will disengage after it begins running.

There are three banks of high voltage coils that are connected 120 degrees out of phase from each other, each bank represents 12 - 100 watt coils for a total of 1200 watts per phase. All isolated coils in each phase are wired in series to produce the high voltage dc potential to energize the capacitor bank. The isolated capacitor bank will discharge their potential through its conversion tubes in 120 degree increments each time the impulse coils are at their closes point toward each other. One isolated capacitor bank will discharge through the conversion tube with it disconnected from the dc dynamo circuit. The second isolated capacitor is at this point beginning to receive its charge from the dc dynamo, while the third capacitor phase is fully energized just before the switching circuit disconnects it from the dc dynamo for it discharge phase. This fully aligns all the isolated 3 coil phases in this permanent magnet dynamo system to the 120 degree three phase radiant energy impulse system of Ed Gray's EMA motor #5.

The plan is to eliminate all the batteries and converter and use a rotating high voltage dc dynamo in its place. Nikola Tesla used high voltage dc dynamos with capacitors as part of his magnifying transmitter to produce radiant electricity. So it seems logical to do the same thing, merging Ed Gray's radiant energy motor and our version of a high voltage dc permanent magnet dynamo. Could it regenerate a car, or power a home, and or a factory?

Shown below are pictures of Ed Gray's motors in fig 28, 29.

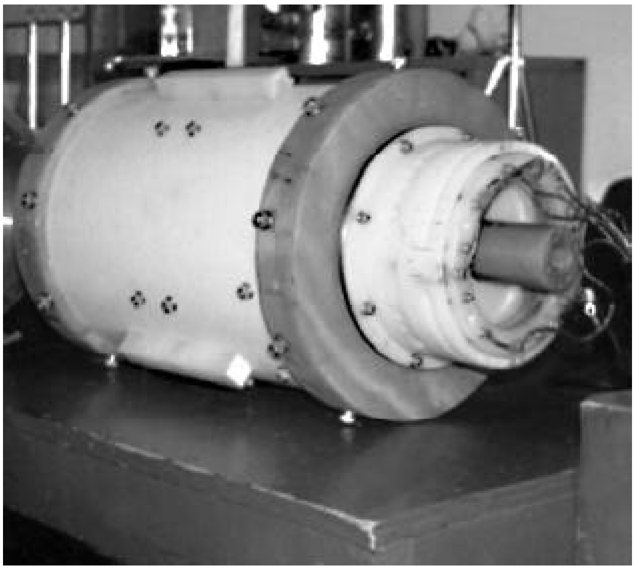

Fig. 28

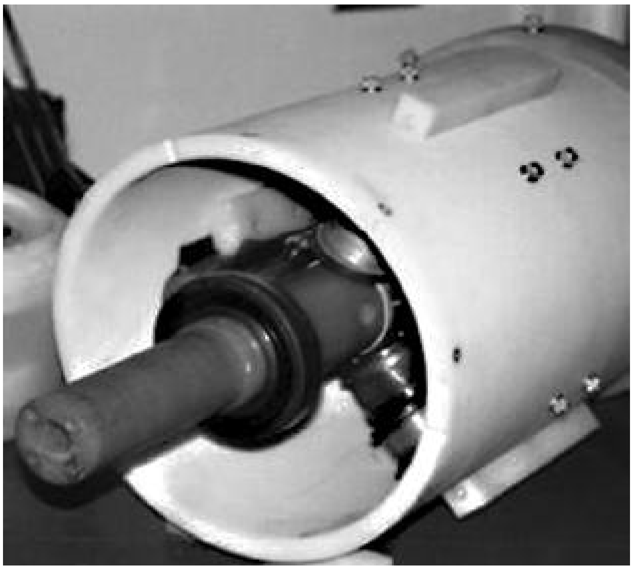

Fig. 29

Fig. 30
Bud Johnson

Fig.31
Dr. Carl Reich

I would like to acknowledge two friends of mine who have helped make the Interference Disc machine possible, Bud Johnson 1986 shown in fig. 30 and the late Dr. Carl Reich shown 1987 in fig. 31. I would also like to acknowledge Wilbert Griffith and Brian & Rose Langan for their contribution in helping us make these projects possible, and there are others who wish to remain anonymous. Bud Johnson also an inventor, was one of the first people who I met after relocating to Alberta from the NWT, he made arrangements for us to construct the first disc machine in a machine shop.

Dr. Carl Reich was a pioneer in the alternate medical field and he published a book along with Robert R. Barefoot called "The Calcium Factor", copyright 1992. Through Carl's dedication and hard work with alternate health and energy, Dr. Carl Reich contributions have also helped made the Interference Disc machine and our Permanent Magnet Dynamo possible.

I also would like to thank my wife Jan Francoeur for being part of making this research possible, she has spent countless hours to prepare this report and assist in the reconstruction of these machines.

The test results and hypothesis I have written are not complete, and the research and development of the described magnetic energy machines is on going, and as such, the potential and performance and true efficiencies of these machines are still yet to be determined. This author knows that the aether is real and free energy systems are possible, and I will continue to work until I have completed the integration of Ed Gray's radiant energy motors with our larger version dc dynamo machine. When this happens, a new round of testing will begin with our attempts to run the two machines together in closed loop regenerating mode.

I hope I explained these energy machines in a way that you all can understand, and more information will be released as progress is made. ***You may find this information useful with your quest to be energy independent. However, if you profit from any of my ideas I have described in this writing, then in all fairness, please remember to reward the inventor so research and development can continue.***

This is my gift to humanity.

Interesting Information

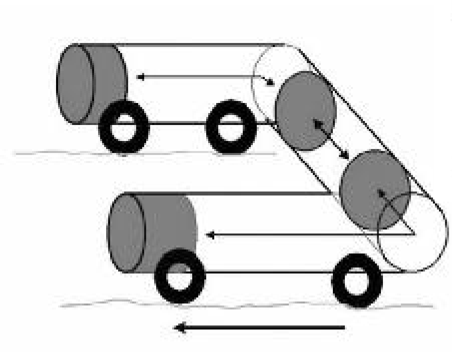

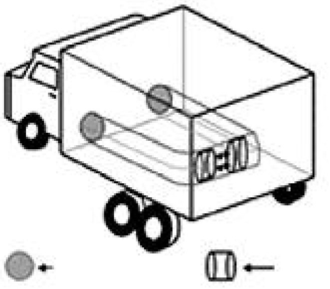

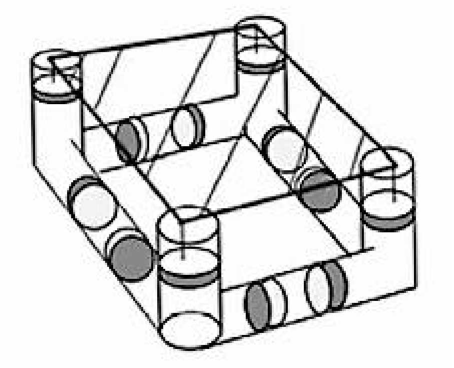

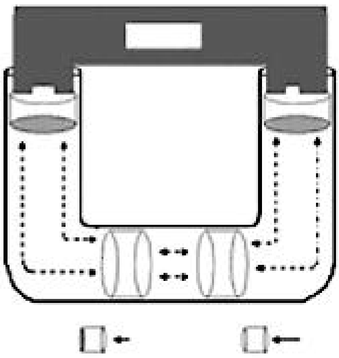

New Propulsion Letter
from Paul Toomer <jorjencar@leeds322.fsnet.co.uk>

Please take a look at this new web site on how to challenge Newton's third law of motion at

http://magnetic.propulsion.mysite.freeserve.com

A NEW PARADIGM FOR TIME

EVIDENCE FROM EMPIRICAL AND ESOTERIC SOURCES

Donald Reed, USA
1032 Borden Road, Depew, NY 14043
Email: torsionpower@yahoo.com

The author of this article
presents an experimental approach
to the teleportation problem
and time pace control
and
space regularity questions.

Part I of II

ABSTRACT

The following essay seeks to establish awareness, through a far-ranging careful examination of various empirically documented anomalous research results in the field of new energy, the long-suspected evidence provided by associated key legendary secret government project work in the USA and former USSR, maverick new theoretical models in foundational physics for elementary particles/fundamental electromagnetic wave-field structures, as well as remarkably corroborative related information from esoteric (psychically channeled) sources, of the necessity for the development of a new paradigm for mass, energy and especially time. After having been guided through this process, the reader will hopefully be motivated to acknowledge the importance of this imperative for a new understanding of the workings of nature, as well as gain hints for the associated future development of new viable sustainable energy sources and related technologies.

Introduction

When we come to examine the annals of physics over the past century, we find them replete with several competing core theories of the physical world, each attempting to demonstrate a unified conception of space, time, energy and matter. Notable among these, the most successful and formally canonized in academia, are special and general relativity, and quantum theory, the latter of which includes quantum electrodynamics and quantum chromodynamics as recent offshoots.

However, even these fundamental theories have fallen short in achieving this coveted goal, due to many reasons, not the least of which is their failure to account for the anomalous but substantial documented evidence continually presented over the years by new energy research and other related empirical evidence considered outside mainstream science. In this author's opinion, these weaknesses in current physics and its relative ignorance and/or selective omission of the findings of non-orthodox scientific research, stem primarily from an ill-conceived institutionalized conception of **time**, as an immutable linear flow against which everything involving **change** can be measured. This "relational" concept of time treats it as a specific passive property of physical systems and changes happening to them. It is a one-dimensional continuous and homogeneous entity geometrically describing the property of duration. This antiquated relational view of time is abstracted from our unique prejudiced viewpoint as sentient beings whose particular point of waking conscious focus is this camouflage physical reality, which is apprehended through the limitations circumscribed by the physical senses.

Accordingly, modern physics is built on the basis of this relational conception of time. However, the use of this conception has not so far resulted in resolving all the problems associated with time. Moreover, so far even an essential definition of time has not been formulated in physics, there are only operational definitions indicating different methods of measuring time intervals.

In contrast, the "substantial" conception of time, advanced in this paper, implies that time is an independent phenomenon of nature existing side by side with matter and physical fields, whose active essence can and does affect objects and processes occurring in the universe. Moreover, the following dissertation argues, by positing that active substantial **change** is the basic concept and relational **time** is its derivative, that those objects and processes can also have a reverse action on time.

Accordingly, we will attempt to show, through a wide-ranging examination of new energy research and other information, the necessity for the establishment of a fluidic, elastic, field nature for true active time which is malleable; one in which even the local pace of time, and mass and energy content can be influenced artificially by intelligent technological control, or naturally by conscious intent, or by a combination of these two methods. To help motivate the reader's awareness for this imperative, we will also examine several esoterically-based sources from key selected psychically channeled transcripts. These will be taken from the following sources: unpublished comments by Jane Roberts on an altered state of consciousness[1] (indicated in this manuscript by J.R.), or published transcripts from **The Seth Material** and The **"Unknown" Reality** (indicated in the manuscript by UR)[2], published chronicles from the entity Kryon, channeled by Lee Carroll [3], and concepts from Wilbert Smith's legendary discourse: **The New Science** (NS) [4]. For easier reference, all psychically channeled transcripts will appear in italics, with my additional explanatory comments in regular type within parentheses.

Granted, our unique investigative foray, ranging as it does from the concrete arena of knowledge represented by the cutting edge of visionary science to the intangible arcane realms bordering on the spiritual, will of necessity lack the exacting logic of scientific rigor.

Nevertheless, we hope to provide, through just such a unique eclectic format, the beginnings of a possible fresh understanding of the workings of nature and perhaps ultimately furnish a conceptual basis for extending the structure of current physical theory to compatibly encompass the elements of a unified framework of physics and metaphysics.

As "unscientific" as this proposed venture may appear, especially to the contemporary physicist or theoretician, we unabashedly press on, secure in the belief that the currently perceived "mutually exclusive" bodies of knowledge defined through psychic means and that of orthodox physics are more closely linked than is currently suspected. Indeed, it will be demonstrated that that some of the psychically defined data bears a striking resemblance to the tenets of present empirical knowledge.

New Research

Theories Indicate Necessity for Novel Time Concept

A body of work which postulates a fluid-field substantial nature for time, is the esoterically-based book written by Wilbert Smith in the early sixties, **The New Science**. Here, Smith outlined a unified theory of all physical interaction by positing that an active mass-free field energy he termed the *tempic field* exists, and is the parent field structure out of which our passive relational linear clock time emerges. To be sure, this book and its specific format of exposition has its weaknesses, not the least of which is a writing style which taxes one's comprehension to the limit. All researchers who have attempted to connect Smith's knowledge with the frontiers of current scientific knowledge, new energy research, etc., have been frustrated by his introduction of terms which may or not have the same meaning as the corresponding terms in mathematical science. This practice causes many passages to appear so vague as to cause total bewilderment on the part of the reader.

...the establishment of a fluidic, elastic, field nature for true active time which is malleable; one in which even the local pace of time, and mass and energy content can be influenced artificially by intelligent technological control, or naturally by conscious intent...

In coming to investigate the source of Smith's theory 25 years ago, I came into contact with Kenneth Killick of Canada. Killick was the individual who served as the original mentor and for a time a colleague of Smith in the mid-1950's. Smith, whose background was in electrical engineering, found it very difficult to accept or understand the philosophical overtones of Killick's thought. Consequently, he resorted to the use of psychic mediums to try to verify the information provided to him by Killick. This is the origin of Smith's association with discarnate entities known to those familiar with his work as "the boys topside". Thus Smith's subsequent writings promote such confusion perhaps because he only obtained a partial understanding of these cosmic ideas, and he opted to writing in a kind of code basically to try to hide his own confusion. Through my own association with Ken, I discovered that The New Science is able to be put into reference with known facts once the spiritual-philosophical elements of Killick's teachings are duly integrated. When this is done, "Wib" Smith's book can teach much and can become a cornerstone to new energy science. Otherwise, it remains at best, a fragmentary enigma.

The work by Smith and Killick is pertinent to the theme of the present paper, since they both showed through experimentation with a special caduceus-wound coil, that so-called relativistic parameters of mass and gravity, energy, and time-flow, can be altered through intelligent artificial control. Adequate delving into some of the elements of these theories will enable us to glean new insight for transforming the current paradigm of time. Further details on Ken's so-called tachion energy theory can be found in this author's earlier expositions on this subject in issues of **Energy Unlimited** (1978-1982)[5].

Smith's main error is in postulating a fundamental "spin" dynamics based upon conventional continuous rotation. As Killick points out, such a concept of spin on the microscopic level, precludes any possibility of true evolutionary change, in the cosmic sense.

Also, as we shall see presently, the continuous spin concept obviates expression by anything in the universe, of its personal moral responsibility. However, with tachion energy theory intact and untrammeled, the impersonal dualistic concept of interaction between two polar entities, such as is expressed not only in ordinary classical Newtonian physics, but its 20th century successors, relativity and quantum theory, is supplanted with the omnipresent workings of a **trinity** at all levels and manifestations of reality.

Indeed, the fundamental tachion field (mass-free) energy has **three** components: two polar opposite entities (a positive "charge" and a negative "charge"), and the ability for these two to be in what is called "static-dynamic balance". The static-dynamic balance completes the trinity.

A simple physical macroscopic manifestation of static-dynamic balance can be seen in two people arm wrestling. As the energy of one participant is brought to bear against that of the other, we will see a vibration or oscillation once a balance of their energies is obtained. As more effort is expended, the vibrations will increase in frequency and decrease in amplitude. At this point we will see a static state relative to the two arms; neither causes loss of arm position of the other. But simultaneously we also have a dynamic situation in the rapid oscillation which maintains, and is in turn sustained by the muscular energy which produces the static state of the arms. Both states mutually support each other's existence. The two component entities (the arms) resonate together, but without losing their own identity.

The implications of this last statement, missing in current physical theory, is unique to the system of tachion energy. We will see in the following that the dynamic functioning of these sub-atomic tachions is such that by their inherent ability to manipulate space and the time-frame of matter through static-dynamic balance, they do not harm anything else in their environment, nor are they affected (forced to lose their own identity) by the ambient environment. Accordingly, inertial mass is not an inherent component of tachion energy, but is a derivative of certain tachion field modes.

We spoke earlier of tachion "charge". We keep the word in quotes to underscore that this is not to be equated to electrical charge, but merely signifies two polar energy states of the primordial ground-form, out of which all known physical forces come to be manifest. These tachion energy states can be defined in terms of "rotational" motion. Again "rotation" must also be placed in quotes as it does not signify the common idea of continuous spinning. Such a concept was suggested above in connection with evolutionary progress in its cosmic interpretation.

The tachion-pair could be described as executing a "back-and-forth" ratcheting movement. The tachion-pairs are analogous to the ends of a drum majorette's baton. But unlike the twirling motion of the baton, the tachion-pairs do not demonstrate the classic macroscopic continuous spin motion, but oscillate in a 3-component "clocking" action.

In Fig. 1 we have drawn a circle and have divided it into three 120 degree arcs. The first motion in the tachion cycle is an arc from the 0^0 -360^0 location, to the 240 degree position. This movement is stated philosophically as the question, "Can I go?" In the second phase it swings back from the 240 degree position to the 120 degree point, and then returns to the 240 degree position. This graphically represents an analysis of the first question, and can be phrased as the "Let me think" component. Finally, then it proceeds forward another 240 degrees taking it past the 360 degree point to 120 degrees. This last stage is the action motivated from the analysis in stage two, and is entitled, "Yes I can". In this activity it will be observed that all path lengths are 240 degrees.

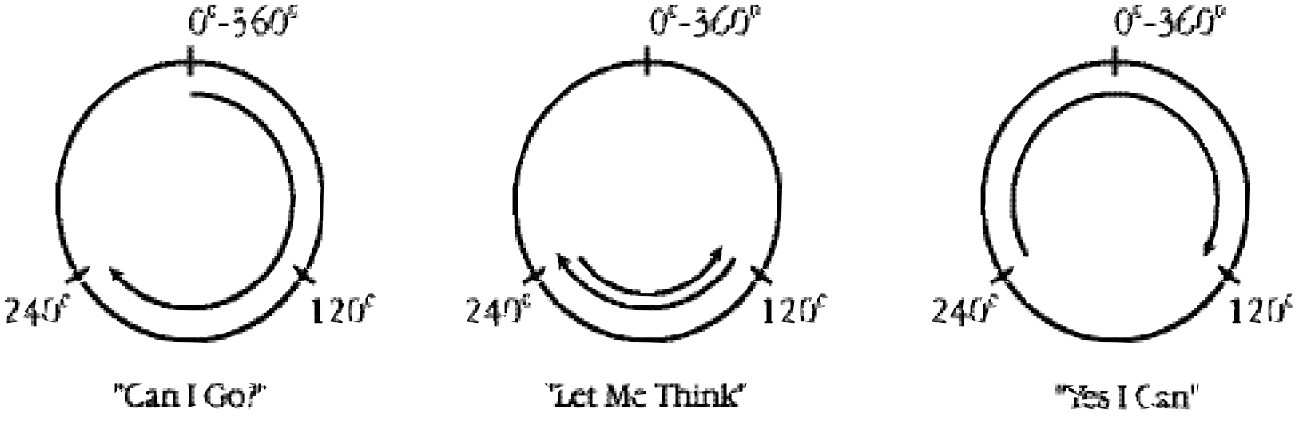

"Can I Go?" "Let Me Think" "Yes I Can"

Fig.1
The 3-Component Tachion Movement

However, as representative of true evolutionary progress, the tachion-pair does not remain confined to the plane in its 3-phase cycle. After each "Let me think" stage, the entire pair unit could be considered to advance "upward", perpendicular to its plane of "rotation". The combined motions can be viewed as an elevation up an inclined ramp, screw-fashion (Fig. 2).

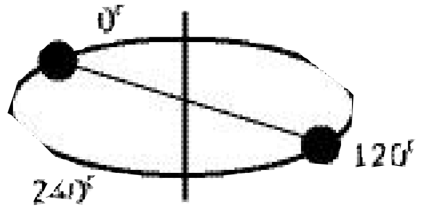

d) 240° Rotation, "Yes, I Can"

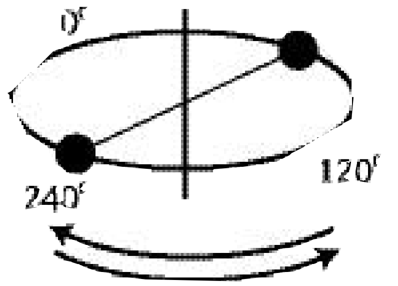

c) 120° & 120° Forward.
"Let Me Think"

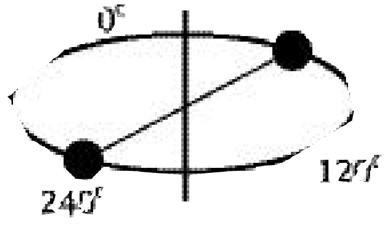

b) 240° Rotation, "Can I Go?"

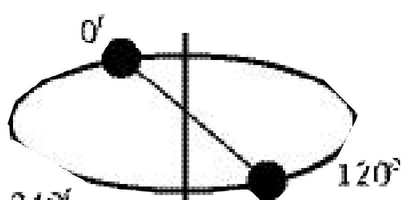

a) Initial Position

Fig.2
Tachion "Ratcheting"

Thus, while they clock themselves backwards in the plane, they never go back to where they were originally but only seem to do so from the point of view of an observer in the plane. This entire activity can be visualized as a diametrically opposed pair "rotating" in a plane, but there must still be considered an upward ratcheting movement. They execute the clocking motion simply because they adhere to the three principles of observation, analysis and motivation. A meaningful application of this idea to the structure of physics would eliminate the necessity for a totally impersonal type of interaction governed solely by force. Unlike all current theories of elementary particles, the tachion-pair, by its unique activity, thus expresses a personal moral responsibility. So this elemental energy of the universe never goes where it is not wanted, but only where it will not disturb its immediate environment.

The topological structure of each tachion in the pair is also significant in regards to this "ethical" modus operandi. First, tachions in their primordial state are massless, toroidal shaped fields which always occur in pairs. Like toroids, tachions will singly exhibit three motional degrees of freedom, to wit;

1. rotation around its major axis.
2. inner (P), or outer (N) rotation about its cross-sectional (poloidal) axis.
3. expansion and contraction of the field (each toroid pulsates radially about its cross-sectional axis).

As a pair-unit, the combined motions will give rise to the following possible orientations: two identical (N) rotations; (N) rotation - (P) rotation; two identical (P) rotations (Fig. 3).

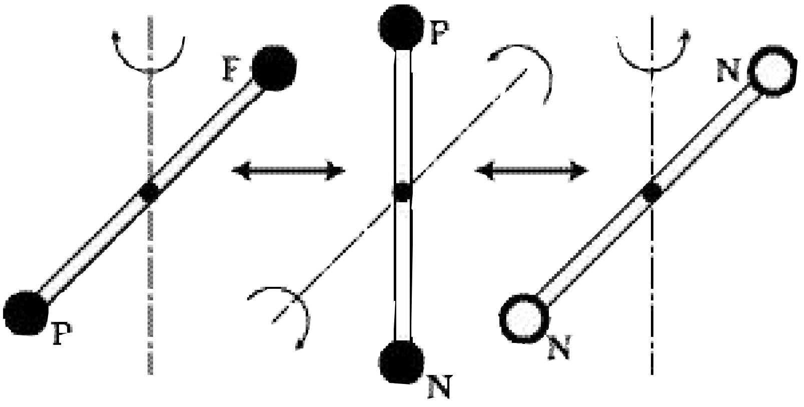

Fig.3
The Three Tachion Modes

The dual (P) or (N) rotations will always lie in the same plane, whereas in the "mixed" rotation mode, the orientations of the tachions will be mutually perpendicular. In the pair unit, the energy will shuttle back and forth between the tachions by virtue of their harmonious expansion and contraction. The contraction of one pair partner is immediately compensated for by an equal amount of expansion in the other. When one is fully expanded and the other completely contracted, they will reverse roles executing the second portion of the cycle. There will be two points in the clocking cycle where the tachions will share an equal amount of energy ("field equity") stage. In each cycle of a tachion-pair the field equity stage is passed twice, a fact which is significant when topology of the field dynamics is considered (see [5]). Now, looking at the field geometry for the tachion-pair unit at the two field equity stages of the cycle, we observe that the two geometries are not identical but are *mirror images* of one another via a non-orientable field structure. See Fig. 5, which shows the isomorphism between the dynamics on a Moebius band (a non-orientable topological structure) and the tachion pair clocking cycle.

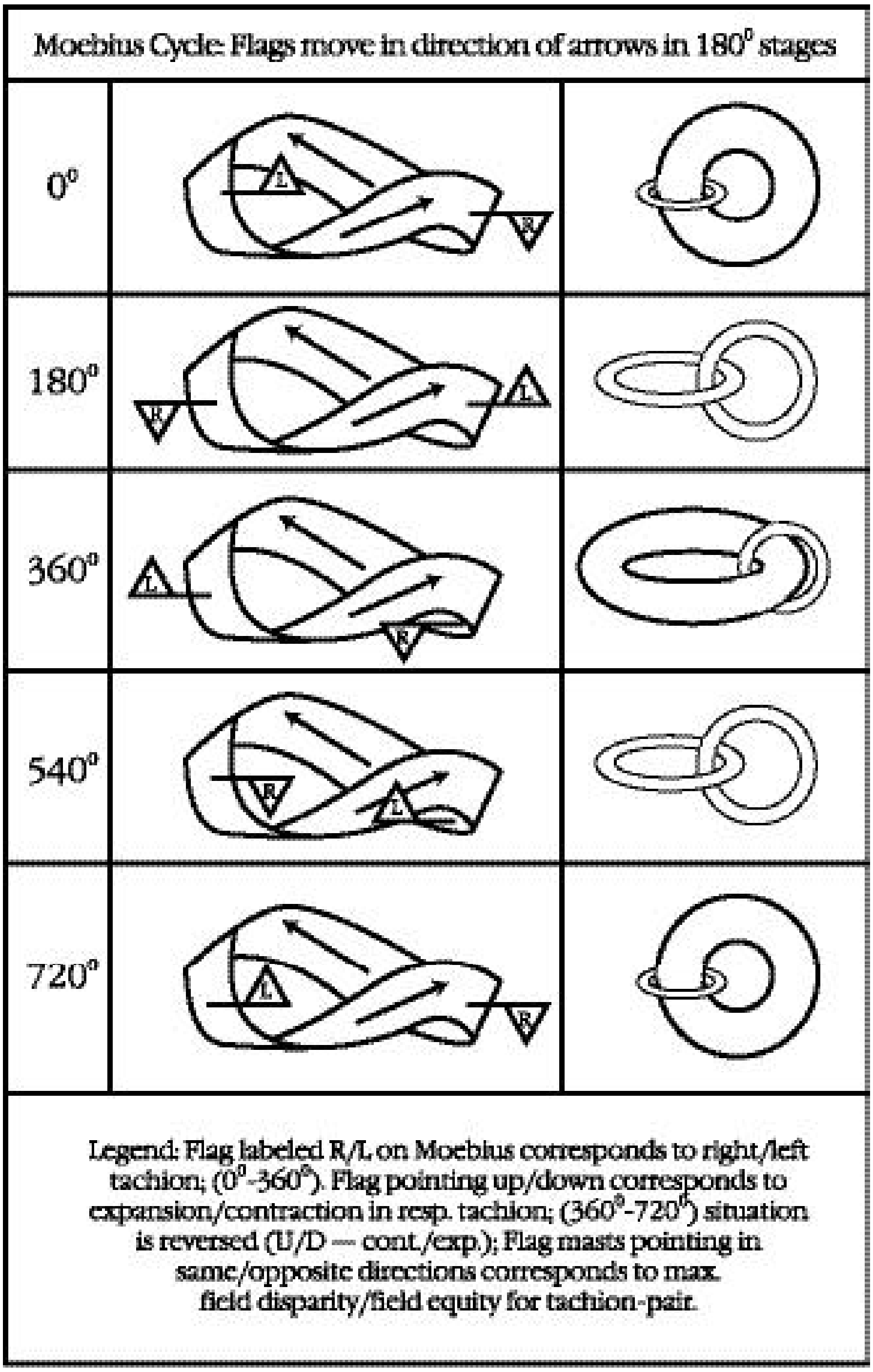
Moebius Cycle: Flags move in direction of arrows in 180° stages

Legend: Flag labeled R/L on Moebius corresponds to right/left tachion; (0°-360°). Flag pointing up/down corresponds to expansion/contraction in resp. tachion; (360°-720°) situation is reversed (U/D — cont./exp.); Flag masts pointing in same/opposite directions corresponds to max. field disparity/field equity for tachion-pair.

Fig.5

This suggests the remarkable idea that the space time points that the pairs generate are not, as commonly assumed, identical, static, and bereft of intrinsic character, but exhibit *enantiomorphic* topologies, which by virtue of their mirrored relationship exert upon one another a potential for dynamic interaction. For now we present a second diagram to illustrate the three pair energy modes: N-N, N-P, P-P, as well as one sequence of oscillation of the N-P pair (Fig. 4).

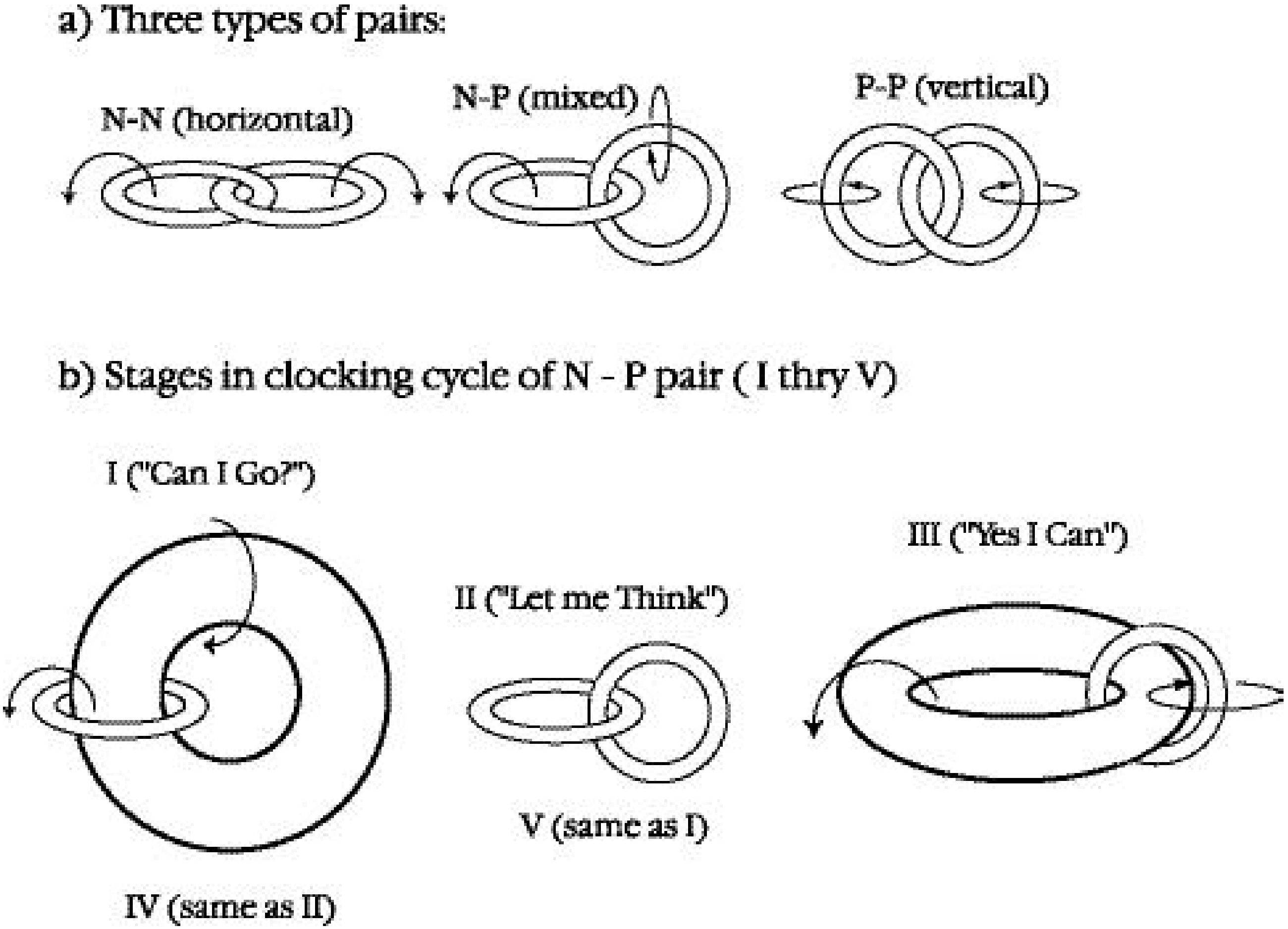

Fig.4

In this figure we note a unique trinity of field operation: two stages of maximum field disparity (I and III) connected by the intermediate condition of field equity (II). Upon postulating the additional requirement that the pair will create a new space-time location every time the field equity state is reached, we see that with each cycle the unit will not return to its original state with respect to the environment, but will advance to new evolutionary states with each pulsation. In accordance with the above comments, the tachion-pair unit "ascends" the evolutionary spiral via this unique action we have termed "clocking" motion. In acccordance with the description in Fig. 4A, in Fig. 4B, stage I represents "Can I go?", stage II represents "Let me think", and stage III represents "Yes I can".

Now, in this regard, the previous analogy of the tachion-pair as likened to the ends of a baton, where the baton rises and falls as it rotates, is quite apt. However, unlike the ends of the baton tachions are not rigidly connected but possess much freedom of movement outlined above. On the other hand, much like the baton ends, they are mutually supportive of each other's activity and requirements, never seeking to overpower, or to be eclipsed by the identity of their partner. In this manner they exhibit personal moral responsibility, as we have previously stated. By their mutual harmonious action tachion-pairs will progress in a true evolutionary fashion, while never harming anything else in the universe by their action.

> *...the dynamic functioning of these sub-atomic tachions is such that by their inherent ability to manipulate space and the time-frame of matter through static-dynamic balance, they do not harm anything else in their environment...*

Now, the above is an inherent characteristic of tachion-pairs as long as they go about their business individually- that is, in their pre-physical state. Now, when the three fundamental tachion-pair energy modes described above exhibit a slow clocking frequency, there is opportunity for them to join together whereupon we begin to see matter being formed. The element of mass, or inertia is provided by the binding activity of the (N-P) mode, which is a function of the slow oscillation of the pair unit. Besides the clocking frequency, the factor of inertial mass or density of physical substance is also dependent upon the amount of separation between the two pair components in each individual sub-atomic tachion-pair unit. In general, the denser or more massive an object is, the slower its microscopic tachions will clock, and the farther apart will be the individual toroid components. This is a classic push-pull situation, i.e., there is a key inverse static-dynamic balance relationship between tachion-pair pulsation frequency and their area of operation (density of field). The rest mass that any physical substance displays to macroscopic observation, is thus a direct function of the amount of energy supplied by the master field to sustain the oscillation of the pair-unit. Thus, a primary feature of tachion dynamics is the following: if we change the energy, we must expect to observe a corresponding change in the manifestation of the mass, and vice versa. Consequently, in tachion dynamics there is a built-in interchangeability of energy and mass, an idea which concurs with similar conclusions in Einsteinean relativity (notably, $E = mc^2$).

But here is where tachion theory and relativity part company. We have seen that our postulated "clocking" pulsation of the pair is the factor, which determines both space and time. Both **space** (area of pair operation), and **time** (frequency of pulsation) are mutually dependent on what **we shall call the master field of thought and intention, which determines the clocking frequency**. From the dynamic influence of the master field arises subsidiary conditions of time and space which are truly "fluid" qualities of tachion operation. This is indeed consonant with the assumptions of a proper unified field theory in which all the factors of our objective world as apprehended by the physical senses, are subsumed under the relative state of a guiding master field. Moreover, under the geometrical constructs of such a theory neither space nor time should be subservient to the other but should retain their unique qualities in a mutually supportive relationship. This is guaranteed in tachion theory under the fluid give-and-take of the pair operation. Relativity, on the other hand, despite its logical consistency, does not meet the above requirements. To his great credit Einstein recognized the need to integrate time with space in a synthesis that would retain the distinctive qualities of both. The result was a theory, which extended the classical Euclidean concept of 3-dimensional space as a kind of "cosmic container" of matter, to embrace a structure of space, which arises from the laws interrelating its material content. Thus, for instance, in his General Relativity theory gravitation became interpreted as a distortion of the spatial fabric by sources of mass-energy in a non-Euclidean framework. Unfortunately, as a foundation of his theory, Einstein mistakenly assumed a subservient role of time to space. He tried to "freeze" time down into spatial-like dimensions utilizing metric equations, which were variants of a basic form. Before a true understanding of gravitation is developed, the rationale for applying such metrics to this end should be seriously re-appraised. In short, instead of freezing time down into a spatial dimension in order to make it fit a world ruled by nothing but gravity, we must develop a conception of space sufficiently fluid to let true time have its proper place therein.

These and other similar ideas are articulated by Andrea and Paulo Correa in their recent insightful essay on the various serious shortcomings of relativity theory. They write: "Time is not treated by relativity as distinct in nature or in quality from the set of topological locations or lengths between points. Whether in the SR (Special Relativity) form of a flat Minkowski spacetime or in the GR (General Relativity) form of a curved Riemannian spacetime, the essence in relativity lies in its treatment of time as the fourth length of a 4-dimensional space. As the actual metric of a length of time is not fixed, only the intervals being invariant, one can no longer speak of the reality of timelines or of synchronicity" [6].

In their second installment, the Correas even imply the existence of a mass-free energy, which is an integral element of the Killick tachion energy theory: "From a strict physical viewpoint, only an energeticist position can make sense. While space and time may be considered to exist outside the function of matter, independently from it, *they cannot be conceived outside the function of energy*...the field remains conceptualizable (in Einsteinean relativity) only by the pseudo-Riemannian manifold, which, on its own, fails to analytically treat the difference in dimensionality between Space and Time, fails to differentiate between them as distinct manifolds, and fails to account for them as the intrinsic properties of energy in flux".

The tachion theory is such a model that delineates the functions of time and space in terms of a basic mass-free energy substratum. Now as we have seen, since tachion dynamics pre-supposes a static-dynamic balance feature connecting frequency of the pairs (fundamental time elements) with their area of operation ("fluid" space configurations), it represents such a foundation upon which to construct such a theory which will harmoniously treat time and space on an equal footing without neutralizing the unique structural features of each. Since the decrease/increase in the clocking frequency of the tachion-pairs will result in a corresponding increase/decrease of area of operation, the local *pace of time* will also decrease/increase at that space-time location. This is due to the push-pull relationship between energy and time. The intrinsic energy of a tachion-pair will be defined to be its energy of "motion" (energy in the "convergent", pre-physical state). Thus we see that an increase/decrease in the clocking frequency also implies that the intrinsic energy of the pair will also follow suit. Consequently, the intrinsic energy available for a manifestation will increase/decrease as time pace allotted for this purpose decreases/increases. Moreover, the product of intrinsic energy and time interval will always be constant due to the perfect trade-off between both factors in the tachion clocking action. This relationship is remarkably similar to the quantitative statement of the Uncertainty Principle in quantum mechanics in the form: $E \cdot t = h$. This formula states that the product of the energy exchange of a quantum mechanical particle and the time interval required for that exchange to take place, is equal to the so-called "quantum of action", or Planck's constant.

Thus, as a concomitant of tachion dynamics, which is argued could also be the basis of electron spin geometry among other things (see [5]), *time* is made malleable; no longer continuous, or rigidly uniform as in Newtonian conceptions, or subservient to space as in Einsteinean relativity, time "flow" can go to zero, and can even reverse under the application of certain electrodynamic field configurations such as provided by the caduceus coil and other devices [5]. With tachion dynamics even the laws of causality will no longer hold. This is due to the fact that tachions have a much broader field of operation than the purely physical realm. Thus, many unusual demonstrated features relative to quantum theory, such as the principle of Indeterminacy, quantum non-locality (EPR experiment), "entangled particles" as well as hypothesized qualities of an *aether* of space, can find their source in this microscopic non-linear electrodynamic clocking oscillation attributed to the tachions. In fact, the ultimate expression of time as well as space to our sensory perception, could be a direct function of these substratum pulsations. From this consideration there emerges a new dynamic model for **fundamental time** as a "fluid-elastic" field effect arising from the unique ratcheting oscillatory motion of tachion-pair entities. Ultimately, it will be seen that tachions may even form the structural essence of consciousness itself, thus providing the long sought-for unifying link between physics and metaphysics.

N. Kozyrev

Other Recent Russian Research Implying New Dimensions for Time

Another key researcher who postulated an active substantial role for time was the Russian astrophysicist Nikolai Kozyrev [7]. He argued that the gap between the exact (physical) sciences and the natural sciences will not vanish until the principle of causality differentiating cause and effect is placed at the basis of the exact sciences. This implies the objective directionality of time or its flow. We encounter numerous manifestations of this principle in our life and in natural science. Yet this principle is new in both mechanics and physics as a whole. The natural scientist's time is not Minkowski's geometrical time, nor is it the time of mechanics, theoretical physics, etc.

> *Nikolai Kozyrev argued that the gap between the exact (physical) sciences and the natural sciences will not vanish until the principle of causality differentiating cause and effect is placed at the basis of the exact sciences*

As Kozyrev argued, mechanics has been using only the "geometrical" property of time, its duration size, i.e., the intervals between the events under study. Timed, these intervals acquire the same passive properties as the interval measured between two points in space. It is only this property of time that the objective sciences assumed to be existing objectively, postulating that all other time properties are subjective. However, if we want to differentiate causes and effects, in reality the *flow of time* must be a physical quantity, duly expressed mathematically and entering the equations of mechanics. The physical meaning and mathematical expression of the flow of time can be obtained from the space-time properties of causality. Thus, processes in the world occur not only in time, but also *with the aid* of time. Time flow is an active property owing to which time may act on a material system. Consequently, it is in the properties of time that the source maintaining the vital processes of the world must be sought.

From these arguments, we can also derive a fundamental property of time. Let us assume that we have changed by a certain technique the flow of time in a material system. In this process we have perhaps changed the stresses within the system and hence its energy.

However, it is impossible to change the total amount of motion of the system, i.e., to obtain a momentum equivalent to an external action. In other words, time may carry energy but no momentum. Time is a non-momentum material reality. From the above property of the flow of time, Kozyrev concluded that it follows directly that the rotation of a body may be used as a device for changing the flow of time in a system. Time flows into the system in the cause-effect direction, the rotation increases the flow, and hence the system obtains additional energy.

In attempt to verify his theories, in the 1950s, Kozyrev thus performed experiments with spinning gyroscopes and discovered, in agreement with his predictions, that the weight of the spinning system changes slightly depending upon the magnitude of the angular velocity and the direction of rotation. Although the effect detected was not large, the nature of the arising forces could not be explained by existing theories. By applying a vibrating framework to the gyroscope, Kozyrev also found that not only does time deliver an energy, but that near the cause-effect system *time density* changes. Time is rarefied near the generator and becomes denser near the receiver. In other words, time becomes stretched by cause and compressed by effect. In addition, the variation in time intensity is in inverse ratio to the first degree of the distance. This can be seen by the following consideration: Time is expressed by rotation and hence involves planes passing through a pole of the rotating system with any orientation in space. **In the case of lines of force issuing from a pole, their density decreases in inverse ratio to the square of the distance, while the density of planes can readily be shown to decrease with the first degree of the distance.**

Kozyrev discovered another property of time when investigating irreversible processes such as crystallization of a substance, condensation, or evaporation. This influence can be effected by any irreversible process and is registered by a measuring system without any preliminary excitation, i.e., when the cause and effect positions are not separated. In other words, apart from flow and density, time seems to have another property, causal orientation, resembling in a way the polarization of light [8]. This property proves to be so variable that even the sign of the observed influences may change and hence the reciprocal cause and effect positions become reversed.

Continuing his studies on the properties of time, Kozyrev (with V.V. Nasonov) conducted astronomical observations using a receiving system of a new type. When the telescope was directed at a certain star, the special detector positioned within the telescope registered the incoming signal even if the main mirror was shielded by metal screens. This fact implied that electromagnetic waves (light) had some component that could not be shielded by metal screens. When the telescope was directed, not at the visible but at the true position of the star, the detector registered a signal that was much stronger. The registration of the true position of different stars could only be interpreted as the detection of star radiation that had velocities billions of times the velocity of light. It was concluded that the radiation from stellar bodies hence had a component, which was non-electromagnetic. Kozyrev also found that the detector measured an incoming signal when the telescope was directed at a position symmetrical to the visible portion of the star relative to its true position. This fact was interpreted as a detection of the future positions of stars [9].

From the 1960's onward, additional Russian research groups continued and expanded upon the work began by Kozyrev on the unique non-electromagnetic radiation that was apparently a concomitant of all rotating bodies, whether accompanying macroscopic spinning objects of great mass (cosmological or terrestrial), or at the quantum level exhibited by elementary particles possessing recorded anomalous spin features which were not accountable by standard theories. This emanation which was exhibited by rotating bodies was termed torsion radiation [10]. Subsequently it was learned that this unique radiation was a part of all living or inert substances. This so-called "torsion field" research was carried on mainly in secret in the former Soviet Union, and only became known to the Western nations after the fall of the Iron Curtain. Many effects attributable to alterations in time and gravity have been reported through this research. For more information on this topic the reader is directed to this author's review in reference 10 and references therein.

Part II of II

Summary of Part I

It is the intention of this multi-part dissertation to motivate an imperative for considering the eventual establishment of a more expansive paradigm for the phenomenon of time, one which views time from the "substantial" as opposed to the orthodox "relational" standpoint. The substantial concept of time implies the existence of an **active** essence which can and does affect objects and processes occurring in the universe. Furthermore, approaching time from such a fundamental basis implies that by positing active substantial change as the ground form **prime mover** and relational time (**passive** "clock-time", or entropy changes) as its derivative, that those objects and processes can also have a **reverse** action on time.

Towards this end in Part I (hereinafter referred to as I), we considered the key research of the late astrophysicist N.A. Kozyrev who, in carefully controlled experiments which were subsequently successfully replicated, recorded phenomena which have hitherto been unexplainable by standard contemporary received knowledge- whether from the standpoints of Newtonian, Relativistic or Quantum physics. Kozyrev pointed out that only by considering time as an active essence, could these cited "anomalous" phenomena, which are virtual conundrums from the viewpoints of a relational (passive) concept of time, be duly formulated on a logical basis. Accordingly Kozyrev's associated "causal mechanics" incorporates an objective directionality of time flow, implying differentiation between "cause" and "effect". **Consequently, by applying a thermodynamically irreversible process to a physical system (example: vibration to a spinning gyroscope), the rotor in this instance would be observed to lose (or gain) a small but detectable percentage of its weight, depending upon the direction of rotation.** Moreover, in this process **linear momentum** (linear speed of rotation) is not changed, but energy is nevertheless delivered to the system by additional forces directed along the axis causing a change in **angular momentum**, and that near the cause-effect system of the irreversible process, **time density** is altered - rarefied near the generator and concentrated in proximity to the receiver. Kozyrev concluded that time is a non-momentum material reality.

Other experiments of an astrophysical nature showed that registration of the true position or future position (as opposed to visual-present-position) of different stars via the non-electromagnetic component of star radiation, demonstrated the existence of an energy essence, which was apparently instantaneous, or had a propagation velocity of at least a billion times the velocity of light. In this regard, Kozyrev opposed the conventional astrophysical paradigm, which posits energy generated/ emitted by stars is a physical consequence of thermonuclear conversion. Instead he maintained that via rotation, stars converted time essence into non-physical (or pre-physical) energy. The latter was nevertheless detectable on Earth by transduction through reflecting telescopes shielded by metal screens, of stellar radiation by a Wheatstone bridge. Hence the "impact" of time was detected by resistor electric conductivity changes (see I and references therein).

Leaving no stone unturned in our investigation, we also took up a discussion of the related Killick-Smith tachion energy model for fundamental (pre-physical) reality. Here we coined a new term- a **trinitivity** of action, which was underscored as a key feature of the operation of tachion-pair entities. Accordingly, we posit the existence of two polar topologically orthogonal toroidal fields, with the unique ability, by the mutual shuttling back-and-forth of their energies, to exist in a third auto-generated state called "static-dynamic balance". Now, since tachions are hypothesized to transfer energy instantaneously, inertial/gravitational mass is not a primordial component of tachion operation (see I). Here we see a striking similarity to the Kozyrev conception of the operation of time in interaction with a physical system - where energy changes are manifested without linear momentum alterations delivered. Likewise, the Smith book, **The New Science** (see I) treats **the hypothetical tempic field as a scalar quantity, which only has direction in relation to its distribution (or "density" change). Only through establishing what Smith termed a *tempic field* gradient does the tempic field possess vectorial nature.** Similarly, Kozyrev viewed time like space as possessing both geometrical and physical properties. The geometric property of time is duration (its scalar component), while physically the course of time is similar to vectorial field propagation and the density of time to field intensity. Although Smith and Kozyrev apparently never were aware of each other's research, the following quote from Kozyrev reveals that they both made key profound discoveries about nature which indeed bear more than a superficial similarity: "Time density is a scalar quantity, just the one observed in the above experiments. Time density decreases with increasing separation from the process creating it. Therefore, a vector property corresponding to density gradient and tractable as time obtaining, should also be observed"[7]. However, in I we also enjoined the reader to observe caution and selectivity when attempting to separate the wheat from the chaff in Smith's book.

Accordingly, the previous ideas are substantially correct, whereas his concept of "continuous spin" of tempic field energy is basically flawed. Killick replaces this outmoded counter-productive notion with a key ratcheting motion, like the escapement of a watch, which is performed by the tachion entities composing a pair-unit. Through this unique "clocking-motion", a cycle of tachion-pair operation exhibits what could only be described as an **ethical modus operandi**, passing through the three stages of observation - reflection - action. By virtue of this 3-stage process, tachion-pairs demonstrate in their individual activity and in interaction with their environment - *personal moral responsibility*.

This is indeed an unprecedented notion which implies that a type of conscious thought process can be ascribed to tachion operation. This can be related to current notions from quantum physics, relative to quantum coherence of states - such as wave-function collapse, Indeterminacy, quantum non-locality (EPR experiment), "entangled" particles, etc. This may be an indication that in such cases, whenever the Principle of Causality is breached or at least compromised, we may be witnessing the unique effects of the inherent conscious microscopic non-linear clocking oscillation of the tachions.

Along these lines, although Kozyrev did not report observing any effects of time energy, which could be considered as "value-motivated", he did report on an equally mysterious "after-effect" in many experiments, which could only be construed as signifying a **memory** process. For instance, it turned out that in experiments with a vibrating torsion balance (or pendulum), at points of support the emerging additional forces did not disappear when the vibration was stopped, but remained in the system for an appreciable time. Interestingly, **their relaxation times were found to be independent of the mass of a body but were dependent on its density**. The largest effects with maximum preservation times were observed on porous materials like brick or volcano tuff. In enumerating on this unusual preservation action Kozyrev observed: "...a body placed for a certain time near an (irreversible) process and then brought to a torsion balance, produced the same effect on it as the process itself...aluminum showed no memory. The largest memorizing effect for processes of both signs has been shown by sugar"[7]. Also, his colleague Danchakov later observed in conjunction with biological experiments that water with decreased viscosity due to having been subject to the action of a process, exerted a distant action on water that had not been under such an action, causing its viscosity to decrease as well.

In this second part of the current dissertation, we will examine other important related information gleaned principally from esoteric sources, referenced in I. These have been obtained from books by Jane Roberts: **The Seth Material**, **The "Unknown" Reality** (UR) [2], unpublished transcripts (J.R.)[1], and transcripts from the entity Kryon channeled by Lee Carroll [3]. This information is remarkably corroborative with that from Kozyrev research and the tachion energy theory. Finally, we will cap off our exposition with a brief examination of a bold new maverick physical theory of subatomic electron structure, featuring variable time/mass attributes, which could be the catalyst for bringing the cited elements from esoteric knowledge and empirical physics research (Kozyrev, etc.) out from their hitherto relative obscurity, to the forefront in our ongoing mainstream investigations in probing the cutting edge frontiers of not only physics, but those of the natural sciences as well.

Other evidence

When one compares the transcripts from various esoteric sources it's remarkable how similar is their description of nature's fundamental operations at the pre-physical level. For instance, a careful reading of the Jane Roberts' channeled transcripts from the entity Seth, reveals a remarkable close connection with the particles termed EE (electromagnetic energy) units and the above Killick teachings on tachion pair dynamics. But first a word of caution to the reader. We should resist the tendency to put into the frame of reference of our sentient physical experience, structures whose essence is fundamentally non-physical. The customary physical characteristics of duality and dichotomy of fields, and the related yardsticks of mass, length and time, for the most part, may not apply in this situation.

It would be impossible, however, to formulate a comprehensive model for the EE unit dynamics were none of the factors for gauging physical reality to be able to be applied to higher dimensional realms. Fortunately, one characteristic does survive the transition to the non-physical framework- the element of **pulsation**. Seth reports in [2] how EE units, much like tachion-pairs are built up and dynamically pulsate in response to thought patterns. First, depending upon the relative intensity of the emotional energy, they will instantly be drawn together in clumps or may just as quickly disperse. Furthermore, being of a pulsating nature, these units can expand and contract. Theoretically, there is no limit to their size or frequency of pulsation. Seth describes their operation in the following manner: **Now, there are electromagnetic structures, so to speak, that are presently beyond your instruments, units that are the basic carriers of perception. Their size varies. Several units may combine, for example. To put this as simply as possible, it is not so much that they move through space, as that they use space to move through. There is a difference**.

On this point, if our conceptions remain "earthbound" we will not perceive any difference between these actions, but once we cast our mental gaze beyond physical experience, an important distinction becomes apparent. To understand this, we need only reflect on the similar non-classical logic that is the basis for the non-orientable Moebius topology that we have seen governs the dynamics of the tachion-pair unit [5]. Tachions, whose translational movement is governed by the relative pulsation frequency, apparently enter and leave our level of reality at two times (field equity stages) during each pulsation cycle. Before leaving our plane, the tachion-pair constitute what we call a space-time point.

Upon returning to our level, due to the Moebius topology of the clocking cycle, their essence creates a new space-time point with enantiomorphic (mirror image) characteristics to the first. The resulting progress of the units in physical reality is thus manifested by a "jerkily discontinuous" movement. Summing up, since the units **are** the point they create, it can now be understood that the units **use** space (or the nature of the space-time points) to move through, rather than the case of macroscopic matter which moves continuously (or appears to) through space.

Seth continues his description of the EE unit dynamics: *Laws of attraction and repulsion are used and, in a manner of speaking, thermal qualities are involved, and this is the only hint that your scientists have received of them so far. The units charge the air through which they pass, and* *draw to them other units. The units are not stationary in the way that, say, a cell is stationary. These units have no "home". They are built up in response to emotional intensity. They are one form that emotional energy takes. They follow their own rules of attraction and repulsion. As a magnet will attract with its filaments, so these units attract their own kind and form patterns which appear to you as perception. They are electromagnetic, in your terms, following their own patterns of positive and negative charge. In this instance, like definitely attracts like. The units are just beneath the range of matter. None are identical. However, there is a structure to them. The structure is beyond the range of electromagnetic qualities as your scientists think of them. Consciousness actually produces these emanations, and they are the basis for any kind of perception, both sensory and extrasensory. These emanations can also appear as sounds and you will be able to translate them into sounds long before your scientists discover their basic meaning. One of the reasons they have not been discovered is precisely because they are so cleverly camouflaged within* ***all*** *structures.*

Being just beyond the range of matter, having a structure, but a non-physical one, and being of a pulsating nature, they can expand and contract. They combine qualities of a unit and a field, in other words. Since they are beyond the range of matter, upon which matter is formed, they will not follow the laws of matter, although at times they may ***mimic*** *the laws of matter. It is almost impossible to detect an individual unit, for in its dance of activity it constantly becomes a part of other such units, expanding and contracting, pulsating and changing in intensity, in force, and changing* ***polarity****. It would be as if the positions of your north and south poles changed constantly while maintaining the same relative distance from one another, and by their change in polarity upsetting the stablility of the planet- except that because of the greater comparative strength at the poles of the units, a newer stability is almost immediately achieved after each shifting.* Even the altruistic motives of the tachion-pair as cited by Killick parallels similar behavior of these EE units.

Again from J.R.: *The behavior of these units changes in the following manner. When a unit is in the act of combining with another, it aligns its components in a characteristic way. When it is separating itself from other units, it will align its components in a different way. The polarities change in each case, within the units. The unit will alter its polarities within itself, adapting the polarity-design of the unit to which it is being attracted; and it will change its polarity away from that design on breaking contact*.

Ken Killick has described the mutual action of several tachion-pairs bonding, as the origin of standard sub-atomic particles possessing features known to modern science such as mass, charge, spin angular momentum, magnetic moment, etc. Thus, according to tachion energy theory, the electron attains its validity, and maintains its structural integrity, by virtue of relations between the component pair dynamics [see [5]). Since for the bulk of their independent existence, tachion pairs are in the pre-physical state, we can speculate that they may actually be superluminal entities. Remarkably, J.R. echoes this view with similar notions: *...the electron achieves its validity because of these* (possibly superluminal) *orbital units* (tachions?). *What we see as an electron is not made of the same things as the orbital units, however; it is an* ***effect****... when you change the structure of the electron, you are changing the relationships between the orbital units. This change will require an application of energy. It may be possible to construct a chamber that would give the electron more freedom.*

Whatever we do now in the handling of electrons is not giving them enough freedom and we are structuring the way in which they can appear. The next observation made by Seth could prove to be essential in clarifying the true nature of time: ... *If electrons were given more freedom they would affect* **our notion of time in perceiving them. The way the measurements (on electrons) are made forces the electron to appear**

in a particular way, but actually it's flitting around, seemingly in many places in rapid succession (pulsing in and out of physical reality). Turning to the structure of elementary particles in [2], Seth offers a similar observation in which he describes the electron as commonly perceived by scientific measurement apparatus, as merely a shadow or facade of the actual electron entity which constitutes a multi-dimensional structure.

In terms of a structural model for the electron, Seth says it would be more appropriate to view these units as composed of *interrelated fields* rather than as the current conception of a particle or a wave. Correspondingly, both of these criteria are key elements of the tachion theory, which posits the electron as composed of two-pair tachion fields. The electron achieves its stability-integrity from the interaction of these free-wheeling units.

Electron Spin and Structure of Fundamental Time

The electron "spin" feature is also claimed to be integral to the exhibit of "time's arrow" in UR [2]: *The electron spin determines time "sequences" from your viewpoint. In those terms, then, a reversed spin is a reversed time motion... electrons, however, spin in many directions at once, an effect impossible for you to perceive. You can only theorize about it. There are "electromagnetic momentums" thus achieved and maintained- certain stabilities that operate and maintain their own integrity, though these may not be "equal" at all portions of the spin. There are equalities set up "between" the inequalities... time, in your terms, then, is spinning newly backward as surely as it is spinning newly into the future. And it is spinning outward and inward* (pulsing inward and outward of physical reality) *into all probabilities simultaneously.*

It is vital that you understand this inward and outward "thrust" of time, however, and realize that from this flows the consecutive appearance of the moment. The thrusting gives dimension to time that so far you have not even begun to realize... This inward and outward thrusting allows for several important conditions that are necessary for the establishment of " relatively" separate, stable universe systems. Such a system may seem like a closed one from any viewpoint within itself. Yet this inward and outward thrusting condition effectively sets up the boundaries and uniqueness of each universal system, while allowing for a constant give-and-take of energy among them. By their unique operation , the EE-units, which form the elementary particles such as the electron, serve as the conduit for this energy flow once it is transduced by elements of thought that Seth terms "consciousness units"(CUs). Continuing in UR on this subject Seth **describes how time can be alternately created or "broken down"**: *The consciousness units serve as source points or "holes"* (Black holes or White holes?) *through which energy falls into your system or is attracted to it- and in so doing, forms it. The experience of forward time and the appearance of physical matter in space and time and all the phenomenal world, results. As CUs leave your system, time is broken down. Its effects are no longer experienced as consecutive, and matter becomes more and more plastic until its mental elements become apparent.* These observations are corroborated by Kryon in his dissertations to follow.

Specifically, in the Kryon channelings, further insight on this electron "freedom" is given in the context of change of frequency: *...There is something we call the electron haze which is around the nucleus of the atom. The space between the energy haze and the nucleus (which is great), varies more than you think in matter. As the space varies, the speed (*frequency*) of the haze must change. It is in the physics where the speed of the haze is strong and fast that you have a different time-frame than your own... you have not yet accepted that an object can seem to be stationary- yet traveling (*vibrating*) very fast. Not in a linear time from point A to point B, but in the electron haze of its vibrating parts.*

J.R. corroborates these comments in this dissertation: *Einstein considered relationships of this sort that hold in the far universe as one increases velocity towards the speed of light,* *but there are also "inner" situations (*altered **stationary** states*) that relate these quantities, and that are much easier to get into.* In regards to atomic structure to allow for frequency changes, Kryon adds: *What is it that makes the distance between the parts change? This is the puzzle for the scientists. For the area between the nucleus and the energy haze although vast, is not void. It is a patterned soup of energy (*the so-called "vacuum"*), and it is within the pattern of this null that changes the distance, and therefore the speed* (frequency*) of the haze.*

Mass, gravity and time are intimately related to this frequency change in the atom. Witness the comments from J.R.: ***Time exists as an invisible drag on matter and this is interpreted as gravity. Time exerts a definite force like gravity, and this is connected with the behavior of the electron.*** *However, since we always look at electrons in a certain way and gravity in a certain way, we miss the connection.* And this related comment as well: *There is another equation to be discovered that expresses the relation between time, mass and gravity.* Kryon reiterates this sentiment, along with relatively detailed scientific information: *Gravity is an absolute product of the characteristics of mass and time... one of which you may change.*

*What has been missing so far in your thoughts is how gravity is related to time, and that the entire issue of gravity, mass and time is non-linear. You have yet to develop the formula of physics. It is not here yet, and we are going to give you three attributes of the formula, but without the formula. These attributes are understood but the way they relate are misunderstood. Your science is just now beginning to see the shadows of them, however. The first attribute of the formula is the ability to understand the **density** of mass. Why is the electron haze so far from the nucleus? We have told you that you can change this distance, and in doing so, the density will change in the mass that is defined by the atoms being shifted. The electron haze will have to vibrate faster (this is possibly the "freedom" referred to earlier by J.R.), the closer it is to the nucleus; when the electron haze vibrates faster, its **time-frame** will shift. Therefore, the second attribute of the formula is a time-shift, and they go together. Also, be aware that you make an erroneous assumption that the actual speed (frequency?) of the electron haze must always be the same no matter what the distance. This is not understood yet either. There is a difference between speed (velocity/ translatory rate) and vibratory rate, and it has to do with the actual physical definition of the **electron haze**.*

Mass discrimination is the ability to control the density of mass. With a density shift will come a time-shift, and with a time-shift will come a third attribute...we will call the **reality of location**, or where the matter is... in which dimensional reality it goes to when shifted. Here is an axiom: **shifts in dimensionality when matter is changed in this fashion create a reality where the matter has to be to exist in its new form**. It might be inches from where you changed it or miles. That depends on how much it got changed. The difficult thing to explain to you is a concept that has not been recognized... that matter has a **reality index**, and that its core attributes are linked to where it exists in time and space. Therefore, the three attributes that must work together are **density, time-frame and location**.

Electromagnetic Field Manipulation of Cosmic Lattice and Time-Space-Matter

In regards to tapping this (mass-free) energy-haze-soup, between the atomic nucleus and electron shells, Kryon speaks thusly: *There's something we are going to call APD (atomic phasic displacement). This is a term that is going to refer to the ability to tap the (Cosmic) Lattice for unlimited energy. The Cosmic Lattice is balanced, but it is not quiet. The Lattice has astounding power. It has a flow of energy that I cannot explain to you, for there is no paradigm model for it yet in your thinking. Therefore, you would not easily understand. The Lattice has vents that we can best describe as necessary for the flow of energy. It balances the slight inequity of the* *polarity. The vents also have to do with time...you will always find two vents together. One will be prominent, the other will be secondary. You might graphically and clearly see one, but you will have to look carefully to see its partner. There are always two. This is an axiom or physical rule of the Lattice energy, and of the universe:* This information remarkably parallels Killick's teachings on tachion energy pairs.

Witness this Kryon pronouncement:*Hiding within common atomic structure is a marvelous peek at something that will totally and completely mystify you, for it will seem to break all the laws of time and space. The" twins" are a pair of atomic parts that always relate to each other, and are always found in pairs. You will discover that when stimulated correctly, they will always move together as a pair. When you start separating them by distance to experiment, they will continue to move exactly together. If one's energy is converted, then the other will do the same. This will cause you to totally reexamine your ideas of time and space, for this condition will not follow the "ultimate speed" of transmission that you thought was correct... that of the speed of light. You will have discovered something that travels faster than you can ever measure*. Compare this statement with the torsion energy claimed by the Russians to travel at a billion times light speed [10].

In returning to the discussion on how to tap the Cosmic Lattice for energy, Kryon continues: *The vents are very necessary for the balance of your universal energy. The vents are also energy portals, (drains, to you), and are where the front of each Lattice cell touches the back...Like so many other physical processes, however, its going to take a tremendous amount of energy in order to unbalance even one cell's null attribute... Once you understand how to" prod" the null to unbalance itself, you will be rewarded with a steady flow of energy, far beyond what you put into it. This is accomplished since you create your own tiny "vent". An unbalanced cell creates a situation where the other cells around it will try to" feed" energy to the one that is unbalanced. This creates a tap that will pull upon the Lattice indefinitely, as long as your work matches the properties the Lattice expects to see.* Kryon further expands on the specific method to accomplish this...*Here's how it works. Two magnetic fields together, postured in the correct way- a way that is very three-dimensional in your thinking process- will create a "designed magnetic field" that is very specific. It's one you have never seen, and does not exist naturally. Start with trying several magnetic fields postured against one another- of unequal force and pattern, and at right angles. Don't make any assumptions. Think freely. Done in the right fashion, these two fields will create a third pattern which is unique and is the product of the original two. This third custom-created pattern is the one that you want to deal with, and is the one that has the potential of manipulating the Lattice. Once you have created it, you will know of its special qualities by how dramatically it changes the physics around it. And this important admonishment is given:* Here is a caution. *Keep this experiment away from your body. Keep the experiment in check with your scientific methods. Go slowly. Understand what you're seeing before you go on to the next step. Do not expose yourself to any magnetic fields. Remotely conduct all energy experiments. Remember that magnetics also plays an important part within your body (DNA components).*

> *...your math told you that there might be a possibility of invisibility and that was your goal. This was again because you did not understand that because the distance between the parts change, it doesn't necessarily mean that the overall size will dramatically change or that it might vanish...*

Here is another. Understand that if you unbalance the Lattice too grandly and too greatly, you will have a ***time displacement****,* *for the process also involves the property of* ***time****... One of the actual physical players in the creative process of unbalancing the Lattice is the manipulation of the time-frame of matter.* ***This is not time travel, but time displacement.*** *It is where you are actually addressing tiny parts of matter, and changing the time-frame they are in. When the inequity of time-frames meet one another (*matter mixed up with differing time attributes*), the result is a displacement of distance (*reality of location*). Although there is no horrendous danger for Earth within this time displacement, it can and will affect the local situation within the experiment. In other words, it can create a matter-distorting effect, completely stopping the experiment, and actually dislocating the parts. We are not going to say any more about this at the moment, but the more astute scientific minds reading this will go to the next obvious step... and the answer is "yes", the Lattice is also the key to quick travel of large physical objects... even of very short distances.*

Possible Documented Demonstrations of Time-Space Warping

Such fantastic transformations were possibly evident in the Hutchison experiments [11], in which metal samples were catastrophically fractured when exposed to radiation from the unique electromagnetic field effect. It was also noticed that many physical objects of different composition which were separate in space before the electromagnetic field was applied, became inexplicably embedded in each other once these objects were radiated by the field effect. For instance, **pieces of wood contained metal knives**, etc. Other metal samples were catastrophically fractured in their molecular-atomic crystal arrangement, producing a substance which was uniform and geometrically symmetrical at one end but appeared to have the resemblance of corrugated cardboard at the other.

Also, similar to Kryon's assertion of the inevitability of time-shift with mass altering phenomena, in Hutchison's work anomalous inertial impulses were recorded on film, sometimes resulting in the complete lift/levitation of various substances of different mass. We notice also anomalous inertial impulses were a feature of the Zinsser "kinetobaric" effect [12], in which due to excitation by **sawtooth microwave impulses with very fast rise time**, a sample attached to a torsion balance, deflected the balance for days and weeks after the original excitation was extinguished. Notice the similarity between these results and those of the Kozyrev experiments, some of which reportedly featured a memorizing preservation effect on a torsion balance, etc.

We also find additional confirmation from the basically anecdotal evidence from the alleged legendary Philadelphia Experiment in 1943, which apparently caused either large time- or space (teleportation) shifts in a Navy ship. In fact, in another piece Kryon himself possibly speaks directly about the Philadelphia Experiment, providing us with further insight into important possible new paradigms for time, mass and energy: *In all your questions regarding magnetics and the massless condition, you have never asked about what happened in your year 1943. You tried to create a massless condition with coarse equipment and little understanding of what you were doing. In the process you actually did create an unstable massless condition for a moment. Its instability created a situation where, instead of a true massless condition, you received one where the time-frame changed but the parts within the sphere of the time change did not have the fine-tuned synchronicity needed for a massless object. The result was an actual distance displacement of the object instead of a true massless condition. Indeed, humans were involved on the large object and their biology was damaged greatly. Your experiment was done in an atmosphere of desperation, and your goal was flawed. For your math told you that there might be a possibility of invisibility and that was your goal. This was again because you did not understand that because the distance between the parts change, it doesn't necessarily mean that the overall size will dramatically change (or that it mightvanish).*

Although this may seem like a paradox to you, the internal mechanics of small particle behavior supports this. The change is measurable but very small, much like what happens with heat and chill...Many of you have guessed correctly that ***magnetics and electricity play a critical role in the determination of the real attributes of mass****... and the magnetic variables that determine the mass product, are often working within very small particles to create the density of an object and also its time-frame. The time frame of a massless object is slightly different than your own, making you to appear slightly slower than the massless object. Its reaction to more traditional mass molecules around it is also predictable: because of the very slight time displacement, it tends to change the number of electrons within the atoms directly in contact with it. This is a clue on how to detect a massless object even if you can't see it.*

*The reason you felt the object (*ship*) would vanish was that you were able to simulate a "vanishing" in the laboratory with smaller objects. This observation was not consistent however, so you again were in desperation to try this experiment on a larger object. The "vanishing" was an illusion and was due to a distance displacement rather than an "in place" vanishing... does this give you a clue to long distance travel using magnetics and the vibratory rate of matter?... Only one human on the planet has ever captured a true massless condition, and even this was a coarse one and lasted for only for a few moments that were uncontrollable in scope. This massless condition was created in the primitive workshop of a great electrical scientist in your culture on the American continent not too long ago. If you were able to visit his workshop, you would note the holes in the ceiling and the patched glass-covered light port where his massless objects took off and wildly flew everywhere. If he had been born 50 years later he would have been able to control the attributes of the experiment. As it was, he did not have the precision tools you have now to direct and control such an experiment. It was his great passion to understand this phenomena, but because it was so uncontrollable and sporadic, he was never able to bring others to see it work, for he could not accurately create it consistently. This depressed him in his later years, for his was a great three-dimensional mind...* When asked the identity of this scientist Kryon replied: *The inventor of your multi-phase current, born in the land you now call Yugoslavia.*

From the above information, there emerges a remarkably consistent picture suggesting that atomic structure may not be as rigid or immutable as formerly thought. Changing the fundamental frequency of vibration of atomic structure can perhaps apparently not only change the density of matter by altering the distance of electron shells from the nucleus, but in so doing alter the time-frame of that object, its mass and also the reality of location of that substance. Certainly, the implications of the Kozyrev research, among others, can possibly form the basis for the eventual technological implementation of these futuristic principles. Perhaps those experimenting with the relatively detailed data provided by Kryon on APD can actually realize this coveted goal.

Maverick New Theories in Physics Suggest Fluid Nature for Time, Space, Energy

The Dirac theory is to be understood as a theory of the electron with electromagnetic interactions. It consists of the Dirac wave function and its properties, including the Dirac equation and relations to physical variables such as energy-momentum, spin and position probability current. In spite of its indisputable mathematical successes, the Dirac theory is still without a completely satisfactory physical interpretation. In particular, the so-called *zitterbewegung* ("jitter motion") is a distinctive feature of the Dirac theory which continues to be the subject of conflicting interpretations in the standard scientific literature. The concept of zitterbewegung was introduced by Schrodinger to interpret high frequency oscillations in free-particle wave packets of the Dirac theory. These oscillations with angular frequency $2mc^2/h$ arise from interference between positive and negative energy components of a wave packet. Schrodinger interpreted the oscillations as fluctuations in position of the electron about an average motion. However, this concept has not been interpreted in the bulk of the literature as having an objective significance, but has erroneously been viewed by standard theory as a mathematical artifact of the one-particle Dirac theory which does not appear in a correctly formulated quantum field theory.

However, as noted theoretical physicist David Hestenes suggests in the first of his recent article series in the journal, **Foundations of Physics** [13], it is only by interpreting the zitterbewegung literally and objectively as point particle motion, that a complete and coherent interpretation of the Dirac theory can be achieved. Moreover, this model has implications that have not yet been previously considered by established physics that relate directly to some of the key features of sub-atomic behavior enumerated in the tachion energy theory and also in the microscopic dynamics of elementary particles revealed by the selected esoteric channelings that we have previously considered in the present exposition.

In particular, it implies that the electron is the seat of a fluctuating electromagnetic field which oscillates with the DeBroglie frequency of the electron (1.6 x 10^{21}cycles/sec).

Thus, it tells us that a kind of electromagnetic wave-particle duality is implicit in the Dirac theory. Of course, this contradicts the conventional view that wave-particle duality is a property of matter which is completely independent of the nature of its interactions. If the zitterbewegung is an objective phenomenon, then it originates from electron self-interaction, since it persists in the absence of external fields. In this new theory which considers the zitterbewegung as an objective physical phenomenon, the electron mass and spin can be identified with the energy and angular momentum of electromagnetic interaction. It suggests that the self-interaction is such that there exist certain stable, non-radiating but accelerated states of motion; in particular, for a free particle, this implies motion in a circle with the radius of a Compton wavelength- the zitterbewegung. The zitterbewegung implies that some of the mass, at least, is kinetic self-energy associated with a spin, and it reconciles the mass with a zero mass of the bare electron. Thus, according to this bold new theory, the electron spin arises from a helical world-line in spacetime. To be sure, some researchers have previously postulated such a helical electron dynamics. Unfortunately, such models have failed to explain why a helical motion for spin should depend on interference between positive and negative energy states or why the zitterbewegung should depend on the way a wave packet is constructed, or even how the zitterbewegung can be the origin for electron spin despite the fact that it vanishes for plane wave states that certainly describe a particle with spin.

The Hestenes model does offer an explanation for such effects and accounts for the ubiquitous feature of spin angular momentum as a function of the zitterbewegung. The essential unprecedented feature of the Hestenes' zitterbewegung idea is the association of the spin with a local circulatory-helical motion characterized by the phase factor of the electron wave function.

Thus, we reach the conclusion that the complex phase factor of the electron wave function can be associated directly with an objective helical motion of the electron, which is, in turn, a derivative of the zitterbewegung. Although the idea of helical motion connected with the electron has been considered before, it has not previously been related to electron phase to produce a complete interpretation of the Dirac theory.

Moreover, in the next installment of his dissertation [14], Hestenes argues that the zitterbewegung is not only an objective dynamical phenomenon associated with the electron, but is a ubiquitous phenomenon, with manifestation in every area of quantum mechanics, even in the non-relativistic domain. For instance, by showing that spin angular momentum can be regarded as angular momentum of the zitterbewegung fluctuations, the zitterbewegung interpretation of the Dirac theory begun by Hestenes in reference 13 provides an explanation for the electron spin and magnetic moment in the physical circulation of momentum and charge. It also explains the mass as the energy of this circulation. Thus, the origin of the zitterbewegung is attributed to self-interaction of the electron with its own electromagnetic field. The relations derived by Hestenes in this article suggests that the interaction is of magnetic origin, since it has the form of a Larmor precession energy if spin angular momentum is proportional to a self-generated magnetic field. The so-called "rest mass" of the electron is thus a kinetic energy of self-interaction. It is this that gives the electron its inertial properties, and the flywheel-like nature of this inertia may be the ultimate origin of spin dependence in electron scattering. And the Heisenberg Uncertainty relations can now be attributed to the "zitterbewegung fact" that an electron cannot be confined to a region smaller than a Compton wavelength. Also, the stationary states of a bound electron exhibit a resonance of the orbital frequency with harmonics of the zitterbewegung frequency, which is imposed formally in the standard theory by requiring single-valuedness of the wave-function. Evidently, such resonances, so prominent in quantum mechanics, can be interpreted as zitterbewegung resonances. This leads to a new explanation of penetration of a potential barrier by sub-atomic particles as due to zitterbewegung fluctuations in momentum, and the Aharonov-Bohm effect as a shift in zitterbewegung phase. The zitterbewegung phase factor literally represents a physical rotation. The rotation rates of this phase in time and space directions are the source of the electron's energy, mass and momentum.

In his third paper [15], Hestenes draws a closer relation of the zitterbewegung dynamics to the Dirac theory by suggesting that the latter actually describes a statistical ensemble of possible electron motions, which are actually governed by the zitterbewegung substructure. Since the energy-momentum of the electron can now be interpreted in terms of a rotation rate in the spin-plane, Hestenes then derives a corresponding relation which defines a variable mass for the electron-another unprecedented idea which is in agreement with the information given by Kryon in his dissertations and Ken Killick's tachion theory. The mass m, the scalar radius of curvature r and the zitterbewegung frequency all covary with changes in the rotation rate in the electron spin-plane. Hestenes obtains a relation demonstrating that the electron mass is inversely proportional to the zitterbewegung frequency. This conforms to the relativistic concept of mass as a measure of energy content. But here, mass is concluded to be primarily a **frequency** measure. This also conforms to DeBroglie's original idea that the electron contains an internal clock with frequency determined by its mass, though for a free particle, the zitterbewegung frequency differs from the DeBroglie frequency by a factor of 2. Moreover, the new key relation derived by Hestenes: $m \cdot r = 1/2h$ (h, Planck's constant), says that this frequency measures the radius of curvature of the electron world-line, so it is a thoroughly geometrical quantity. Thus, as mass increases the radius of curvature decreases, in concert with the tachion dynamics as well. All this suggests that the electron mass relates our externally imposed time scale to a time scale intrinsic to the electron.

The reader can see that the information transmitted by J.R. is also remarkably similar in this connection. It should be evident that there is introduced a new concept of mass here, though, to a certain extent, it was already implicit in the Dirac theory. The formerly vague concept of mass as some kind of material stuff is completely gone. Also, no longer is vanishing mass a distinguishing feature of particles moving with the speed of light.

Corroborating Research on Oscillatory-Fluid Nature for Time

In his recent thought-provoking dissertation [16], Tom Bearden relates of the dynamic influence of time on a material system: "But we can also legitimately state that 'time is energy' and be rigorously accurate...Time is extremely compressed EM spatial energy...Without further elaboration, we speak of a 'mass' in which a small portion exists as 'masstime' rather than mass, as having been 'time-charged' or 'time-excited'.

The t-polarization wave in the time dimension is quite unique: The *spatial* energy of the wave is in equilibrium and not vibrating at all; instead, the photons comprising the wave are vibrating in their **time**-components... In short, mass m is changed into masstime mt by photon absorption... Rigorously, a mass does not really 'travel through time' continuously, per se, but proceeds with an overall serial change mechanism, driven by its total virtual and observable photon interactions, as m→ mt→ m→ mt→...We propose that this may account for the duality of particle and wave...The particle actually oscillates at a high rate between the m and mt states... Mass 'travels through time' by an extremely high oscillation between corpuscle-like state and wave-like state". Notice how Bearden's description of the macroscopic operation of time on a physical system closely parallels Hestenes' proposed model for of the zitterbewegung dynamics at the sub-atomic level which provides the electron (and possibly photons, etc.), with a time-substructure (its internal "clock").

We can see many other places in Bearden's treatise where the unique terminology chosen can now be put into direct reference with equivalent concepts and unique phraseology employed by many of the researchers examined in the current exposition. This cross-referencing certainly provides much needed clarification about key ideas for researchers currently involved with the development of this fledgling discipline that we shall term causal mechanics, after Kozyrev's designation.

All such efforts will help to place investigative endeavors currently scoffed at by establishment science on a firm rational foundation, as well as assist in synthesizing seemingly disparate but surprisingly related "anomalous" phenomena. For instance, to implement the science he has termed "vacuum engineering", Bearden often speaks of the necessity of establishing a "local time-stress of the **vacuum potential**". This process, which as we hope to have shown here is completely legitimate, is identical to what Smith calls creating a "tempic field gradient", Kozyrev's "alteration of time-density", or Alexander Frolov's "change in the local time-rate".

It can clearly be seen that such ideas can only be perceived as science fiction fantasy, if we observe strict adherence to all current received paradigms (even in superstring theory), which consider space and time purely from their relational standpoints. These theories must of necessity involve models employing scalable metrics for their proper description. However, once we embrace the novel view posited here that vacuum energy is primary and both local space and time are derivatives of this mass-free non-scalable dynamic non-local oscillating substratum, then we can conceive of space and time properly as complementary aspects of non-local conscious reality, neither fixed in a scalable manner or subservient to the other, but fluid and mutually interdependent features of energy in constant flux (see the related ideas voiced by A. Correa and P. Correa in I).

With this approach we will then come to acknowledge the technological feasibility of many of the "futuristic" claims made in the channeled transcripts. For instance, we can see that it will no longer be necessary to propel elementary particles to relativistic velocities, or to probe astrophysical systems possessing large gravitational fields, to recognize alterations in space, time and mass parameters, since these phenomena can equally as well be accomplished in the laboratory by modifying atomic structure of stationary matter through electromagnetic means. Once again, peruse the Kryon dissertations on this process, for it may be key to future technology.

For those who wish to further investigate these theoretical possibilities, the papers of **A. A. Nassikas** are recommended. Nassikas has postulated what very few researchers before him have theorized: the existence of a ground-form probability density function for energy, as a necessary precursor to the malleable (deformable) characteristics of time and space [18]. One of Nassikas' conclusions is significant in light of ideas articulated in the present exposition concerning the hypothetical tachion-pair model for reality (see I). Here we refer to the key push-pull oscillatory ("breathing") operation between space and time, which may be responsible for the structural integrity of matter as we know it (again re-read the Hestenes zitterbewegung model).

Researcher **Alexander Frolov** has taken Nassikas' theory even further to suggest an explanation for the **anomalous over-unity effect** that has continually been demonstrated in low-energy nuclear reactions (LENR) over the past decade [19]. In such so-called "cold fusion" cells the Palladium cathode is over-saturated with protons, producing a local imbalance in the vacuum engine, causing a local alteration in the time-frame of the experiment to compensate the local change in energy density. **Subsequently, as Frolov claims, this time-frame change converts non-local time-energy directly into heat energy, similar to Kozyrev's description of energy production by stars.** Perhaps this observation might provide the hitherto missing key in our understanding of fusion by electrolytic action.

Conclusions and Prospects

In the preceding paper it is sincerely hoped that the author has demonstrated ample evidence that will spur on researchers of a kindred spirit towards the development of new paradigms for space, energy, mass and especially time. All these concepts, as well as the foundations of current classical electromagnetic theory, are in need of a drastic overhaul before we can fully appreciate and understand the operations of nature as it interrelates the metaphysical-mental realms with the physical level of reality.

At any rate, the continual appearance of research results, such as the Kozyrev effect, which up to present apparently resists explication in terms of conventional scientific paradigms, certainly offers clear evidence of the incompleteness of our knowledge of nature, even in this enlightened era since the dawn of the new millennium.

Accordingly, hints as to how the Kozyrev effect arises in association with the postulated active properties of time, and how the latter interacts with known physical forces, can assist us towards the development of a more comprehensive paradigm which embraces a wider spectrum of human knowledge - one in which the current conundrums rampant in the foundations of theoretical physics, and biological science disciplines as well, which have hitherto prevented the establishment of a unified model of all physical interaction, can find satisfactory solution. Also, such empirical evidence which generally flies in the face of current scientific wisdom, presents the greatest challenge to our ongoing search for new sustainable energy sources which will be of absolute necessity in the future.

In this light, if formally integrated into scientific thought, the tenets of tachion energy theory, which is based on an altruistic modus operandi, also may hold profound implications for the status of our very social, political and religious structures as well. Indeed, unlike the impersonally motivated interactions believed to be the basis for force interactions in modern sub-atomic physics, which help sustain our illusory belief in the supremacy of the ego, the philosophical basis of tachion dynamics is an able reminder of the humility of purpose which should be the guiding force whatever our walk of life. Like tachion-pairs, in the ideal social structure individuals will work side-by-side, yet achieving a common accord. United in the circle of their activity they will work together joyfully in freedom, each with his own task, yet always conscious of the common bond.

With the full import of these simple truths appreciated, tachions may afford a rational justification of the formerly "unprovable" mystical doctrine of the omnipresent intelligent plan behind the working of the universe at all levels of conscious reality. We will come to realize the profound truth of Einstein's vision of ultimate reality couched in his heretofore cryptic statement, *"God is subtle, never malicious"*.

REFERENCES:

1 J. Roberts, unpublished remarks as transcribed by W.H. Kautz.

2 J. Roberts, **The Seth Material**, Prentice-Hall, Inc. 1970. Pp. 292-298.
The "Unknown" Reality, Prentice-Hall, Inc., 1986.

3 L. Carroll, **Alchemy of the Human Spirit**, (Kryon Book III), The Kryon Writings, 1995.
Letters from Home (Kryon Book VI), The Kryon Writings, 1998.
Passing the Marker (Kryon Book VIII), The Kryon Writings, 2000.

4 W. Smith, **The New Science**, Fern-Graphic Pub., 1964.

5 D. Reed & K. Killick, "Tachion Energy Theory"(Parts I-IV), **Energy Unlimited**, 1978-1983.

6 A. Correa & P. Correa, "Consequences of the Null Result of the Michelson-Morley Experiment: The Demise of the Stationary Aether , the Rise of Special Relativity and the Heuristic Concept of the Photon", **Infinite Energy**, no. 38, July-Aug 2001, pp. 47-64.
"The Sagnac and Michelson-Gale Experiments", **Infinite Energy**, no. 39, Sept.-Oct. 2001, pp. 32-49.

7 N. Kozyrev, "On the Possibility for the Experimental Investigation of the Properties of Time", **Time in Science and Philosophy**, Prague, 1971, pp. 111-132.
"An Unexplored World", **Soviet Life**, 1965.
A.P. Levich (ed), **On the Way to Understanding the Time Phenomenon (Part II: The Active Properties of Time According to N.A. Kozyrev)**, World Scientific, Singapore, 1996 (rare English language excellent critical review of Kozyrev research and theories).

8 M. Lavrentiev, "Detection of the Reaction of Matter to an External Irreversible Process", **Sov. Phys. Dokl.**, Mar 1991, vol. 36, no. 3, p. 243.

9 M. Lavrentiev et al., "Remote Effects of Stars on a Resistor", **Sov. Phys. Dokl.**, Sept 1990, vol. 35. No. 9, p. 818.
"Detection of the True Position of the Sun", **Sov. Phys. Dokl.**, Nov 1990, vol. 35, no. 11, p. 957.
"Scanning the Celestial Sphere with a Kozyrev Detector", **Sov. Phys. Dokl.**, Apr. 1992, vol. 37, no. 4, p. 163.

10 D. Reed, "Torsion Field Research and Implications for New Physics and Energy Technologies" **Jour. New Energy**, vol. 4, no. 2, Fall 1999, pp. 151-164 (and references therein).

11 J. Hutchison, "The Hutchison Effect Apparatus", **Proc. Of the 1st Symp on New Energy**, Denver, May 1994, p. 199.
"Inside View of the Hutchison Effect", **Extraordinary Science**, vol. III, issue 4, Oct-Dec 1991, pp. 23-26.

12 T. Valone(ed), **Mechanical Energy from Gravitational Anisotropy**, Integrity Research Institute, Washington, DC, 1996.
W. Peshka, "Kinetobaric Effect as Basis for a New Propulsion Principle", **Raumfahrt-Forschung**, Feb 1974 (in German); trans. By D.Reed, **Infinite Energy**, vol. 4, issue 22, 1998, pp. 52-56.
D. Reed, "Translator's Analysis and Comments on the Zinsser-Effect Device", **Infinite Energy**, vol. 4, issue 22, 1998, pp. 57-59.

13 D. Hestenes, "Quantum Mechanics from Self-Interaction", **Found. Phys.**, vol. 15, no. 1, 1985, pp. 63-87.

14 D. Hestenes, "The Zitterbewegung Interpretation of Quantum Mechanics", **Found. Phys.**, vol. 20, no. 10, 1990, pp. 1213-1232.

15 D. Hestenes, "Zitterbewegung Modeling", **Found. Phys.**, vol. 23, no. 3, 1992, pp. 365-387.

16 T. Bearden, **Extracting and Using Electromagnetic Energy From the Active Vacuum**, Association of Distinguished American Scientists, Huntsville, Alabama, 2000.
"Giant Negentropy in the Common Dipole", **Proc IC-2000**, St. Petersburg, Russia, 2000 (in press).

17 A.A. Nassikas, **The Hypothesis and the Equations of the Unified Matter Field**, in New Ideas in Natural Sciences (Part I: Physics), St. Petersburg, Russia, 1996.

18 A. Frolov, **Work Created By Means of the Potential Field**, in New Ideas in Natural Sciences(Part I: Physics), St. Petersburg, Russia, 1996.

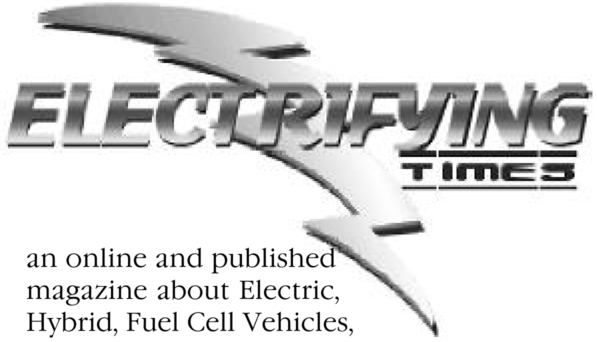

an online and published
magazine about Electric,
Hybrid, Fuel Cell Vehicles,
advanced batteries, ultra capacitors,
fuel cells, microturbines, free energy
systems, events and exhibitions worldwide

63600 Deschutes Mkt Rd,
Bend Oregon, 97701
541-388-1908 fax 541-388-2750
etimes@teleport.com
www.electrifyingtimes.com
Subscription $13/3 issues

On the Rotating Permanent Magnet System

Alexander V. Frolov, Russia
7 Lev Tolstoy Str., St. Petersburg, 197376, Faraday Laboratory Ltd.

In one of our issues we published an article about the results of the experiment conducted by S.M. Godin and V.V. Roschin. They started in 1990 at Moscow Aviation Institute (MAI) when a technical physics laboratory was created in Machine Industry Projects Institute at the "Aeropromservice" Association. Their research work was financed by private investors and was continued in close cooperation with Chair 310 of MAI, especially with Prof. PhD. L.K. Kovalev, Prof. PhD. V.V. Rybakov and PhD. K.V. Ilyushin. By 1991 "a device for inner energy of matter conversion" had been designed and tested (according to the Technical Design Assignment). The prototype was finished in the middle of 1993 (Fig.1).

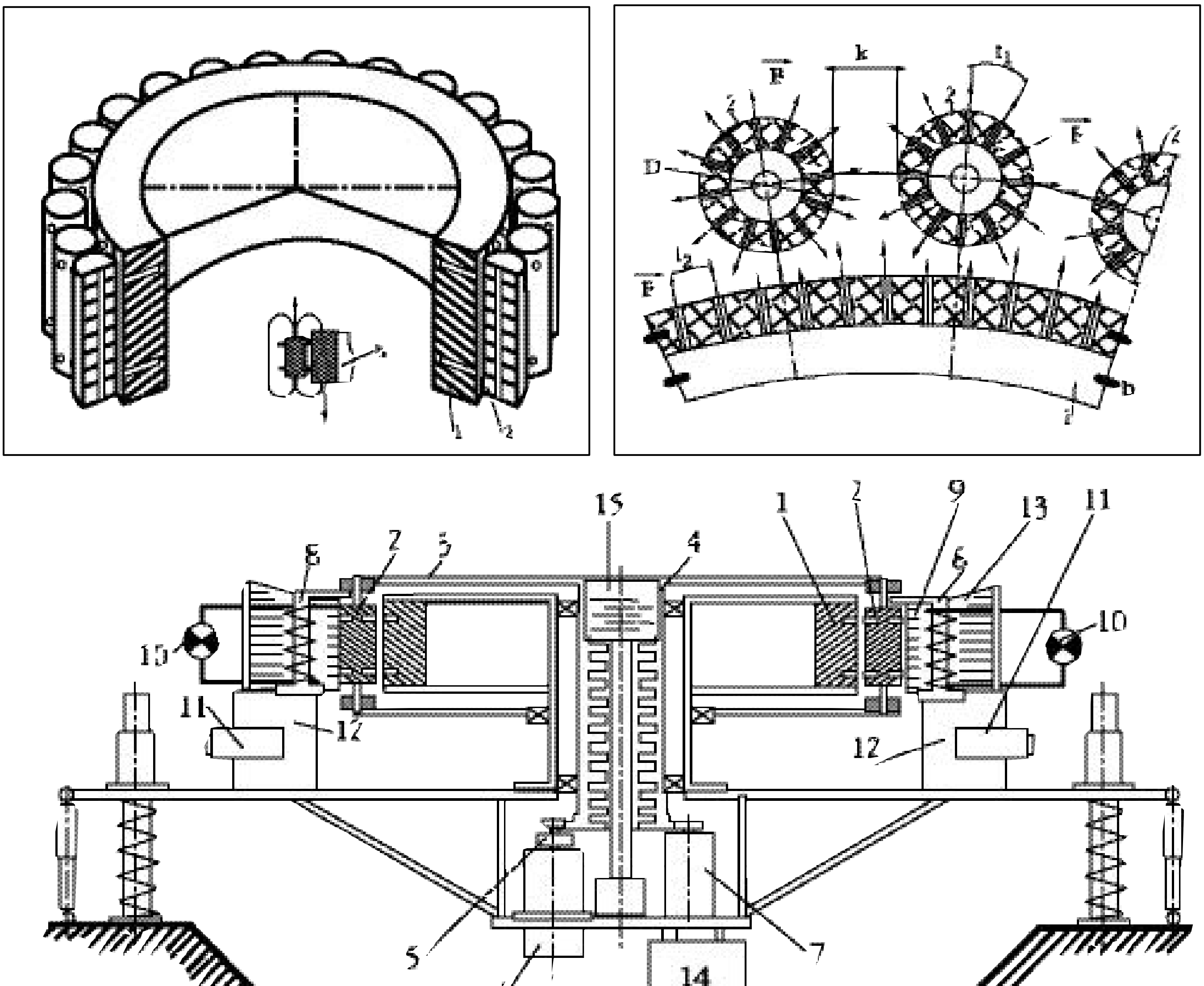

Fig.1

The experiments resulted in the possibility of creating a 7 kW output power in load after the rotor was sped up to 10 rps, but in the autumn of 1993 the laboratory was closed due to the investors' financial problems. Now their research work is continuing in Moscow and we hope the scientists will obtain new experimental results.

You can email S.M. Godin for more information at: smgodin@online.ru

Considering the theoretical basis of the functioning of these systems I should suggest a number of interesting details for prospective researchers.

S.M. Godin, V.V. Roschin et. al. (Physics of Negative Viscosity Phenomena. Prof. Victor P. Starr. Massachusetts Institute of Technology) spoke of vacuum (quantum medium) as **a physical medium possessing negative viscosity.** At that rate, it is worth making assumptions that can be useful to the developers of this topic. I think that standard (positive) viscosity of matter, from the physical point of view presupposes the presence of positive friction and a "work-to-heat" entropy transformation. Consequently, **negative viscosity** should correspond to the **negentropy cycle** "heat is converted to work". This is exactly the stray heat conversion once described by Tsyolkovsky in 1914. Apparently, **the medium of the kind supports self-sustained vortices** characterized by the surrounding medium heat absorption, which was mentioned yet not explained by S.M. Godin and V.V. Roschin.

It is possible that all the elements of matter are the self-sustainable vortex structures. To mathematically describe these physical mechanisms, the notions "heat-work-heat" transformation and "time" or "gravitation" should be interrelated. Russian philosopher Pavel D. Uspenskiy had been making related research since the year 1910.

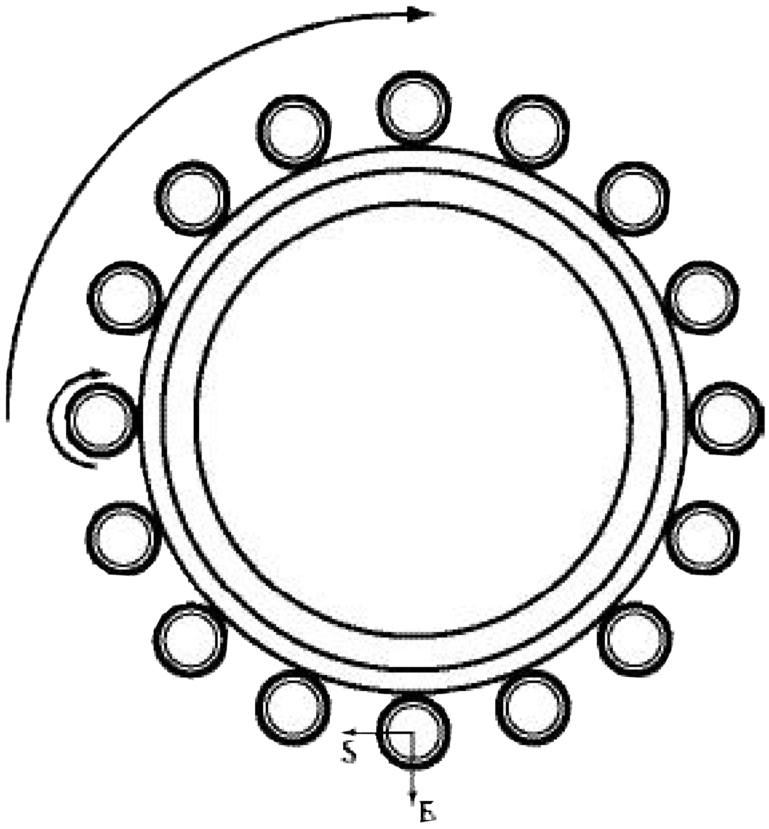

Fig.2

From the conventional point of view, this device (Fig.1) creates a circulating energy flux described by the Umov-Pointing vector (toroidal vortex) as the roller magnet field vector is vertical and the electric field is radial.

S.M.Godin (on the left) and A.V.Frolov (on the right), St. Petersburg, June 2001.

Inventor's Week Conference 2003, USA

Review is prepared by Editor Tatiana Ezhova

This year the annual Inventor's Week-end Conference was held on September 19-20 in Seattle, Washington. The conference featured several interesting inventions, including MEG.

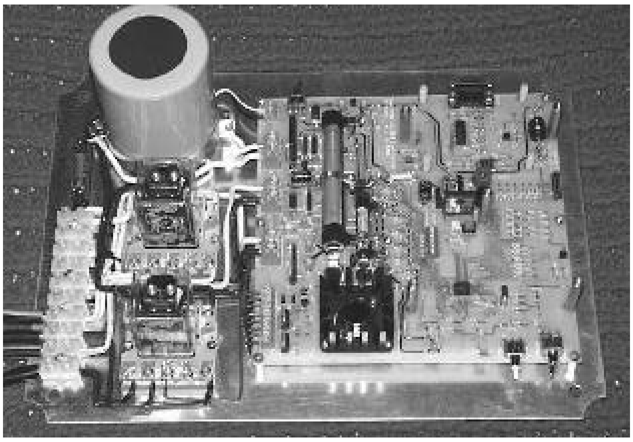

Bill Alek demonstrated a device that he believed could let him detect Vortex-Energy. It was attached to a meter, and his belief was that if he walked into a naturally-occurring energy vortex carrying the device it would show him changes in gravitational and other energies. He cited the Vortices in Santa Cruz and Oregon as examples that might be good for experimentation.

Bill Alek

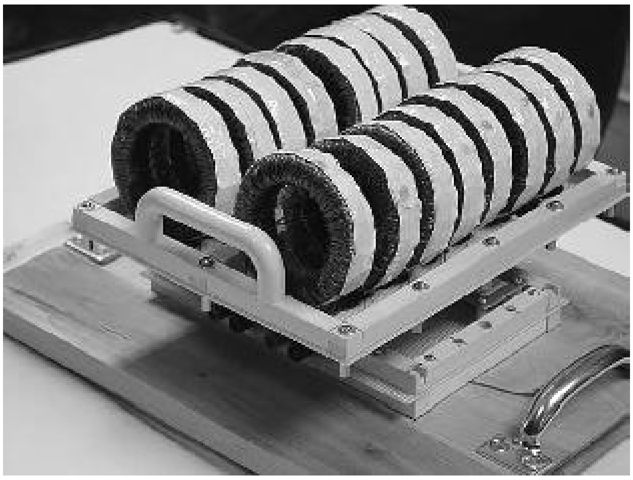

Z-Pod, Bill Alek

In his report "Introduction to Parametric Mass Fluctuation - a breakthrough in Energy and Inertialess Field Propulsion" Bill Alek presented the Z-Pod, developed to test a theory of solid-state parametric mass fluctuation. Here is the URL for a good overview of the device that he published online at www.intalek.com:

http://www.intalek.com/Index/Projects/SmartPAK/Projects/ZPOD/ZPOD_System03.pdf

Rio Von Sternberg

Rio Von Sternberg was presented as the original inventor of the MEG. He apparently has a patent on the device; however, from his presentation I understand that his configuration is somewhat different than the model that J.L.Naudin and T.E.Bearden have been experimenting with.

Alexander Peterson had a very interesting presentation on optical-mirror magnetic field systems. The crux of his presentation involved cutting precise notches into magnetic coils to separate them into identical optical-isomeric shapes. Apparently, by cutting a thin line directly down the center of the transformer-core is believed by Peterson to increase the efficiency of electric motors by up to 50%.

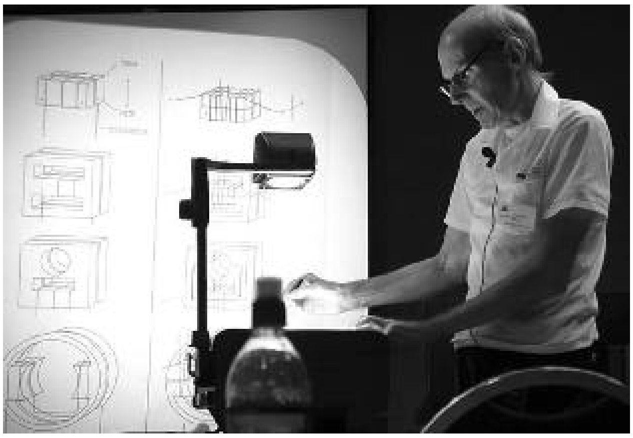

Alexander Peterson

Mark Plotkin, a well-known Washington DC attorney also took part in the conference. He is interested in new energy technologies and works with new technologies in collaboration with Mark Whitford at www.awetec.info. The photo of Mark shows him rotating a steel coil apparatus that was passed around the conference by Slim Spurling.

Mark Plotkin

Slim Spurling is an inventor from the Southwest, shown holding a helical-wound coil that he claims provides health benefits through an interface with zero-point energy. Spurling's belief is that this specific circumference of the coils provides a "Casimir-like Effect" that concentrates the energy.

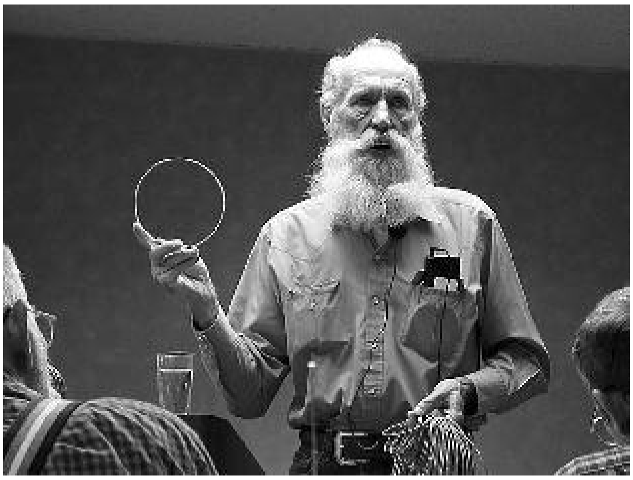

Slim Spurling

Donald Smith

Note: Donald Smith, a well-known inventor, also took part in this year's conference. His work is described in more detail in a separate article of this issue.

Full information on the conference and reports is presented at http://ionvalve.com/iw/2003/conference2003.htm.

USA "Inertia" Patents

Information from www.rexresearch.com/inertial/

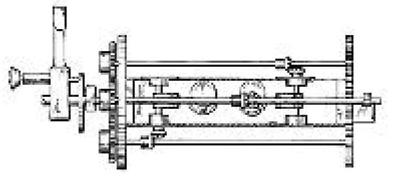

№ 6,290,622 (7/18/01)
Mechanical Force Generator
Murray, Lawrence D.

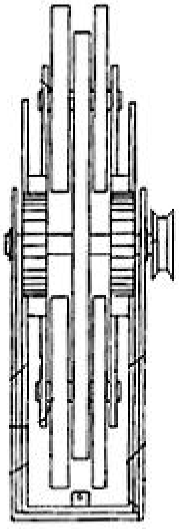

№ 5,890,400 (4/06/99)
Apparatus for Generating
a Propulsion Force
Oades, Ross A.

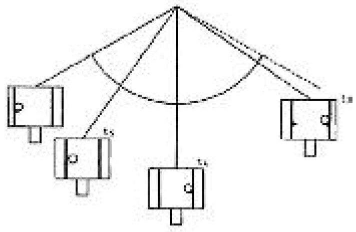

№ 6,089,511(7/18/00)
Method & Apparatus
for Propulsion
Rasmussen, James K.

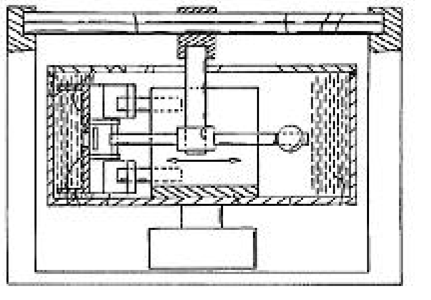

№ 6,259,177 (7/10/01)
Motion Imparting System
Deschamplain, Daviv

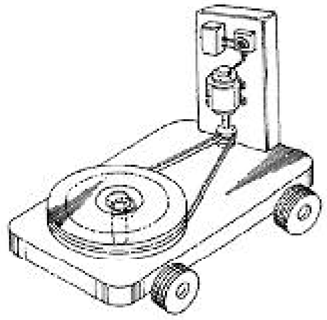

№ 5,937,698 (8/17/99)
Centrifugal Propulsi
on System
Kunz, William T.

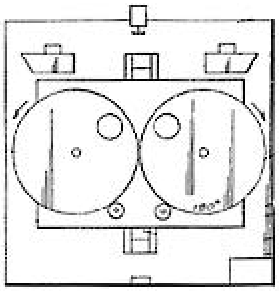

№ 5,966,986 (10/19/99)
Propulsion System
Laul, Virgil R.

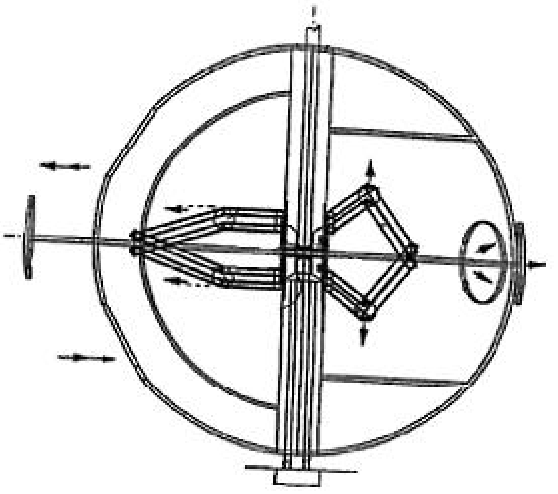

№ 5,673,872 (10/7/97)
Apparatus for Energy
Transformation & Conservation
Shimshi, Ezra.

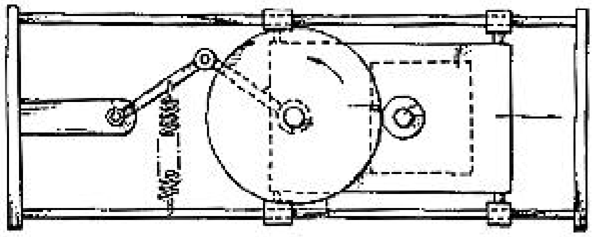

№ 3,492,881 (2/3/70)
Prime Mover
Auweele, Albert J. v.

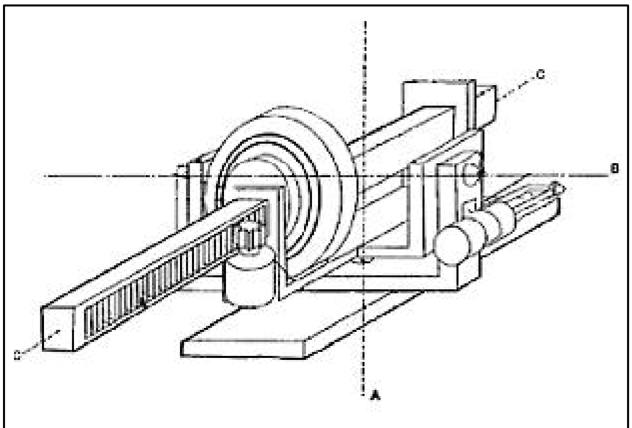

№ 5,860,317 (1/19/99)
Gyroscopic Propulsion System
Laithwaite, Eric R. &
Dawson, Wm.

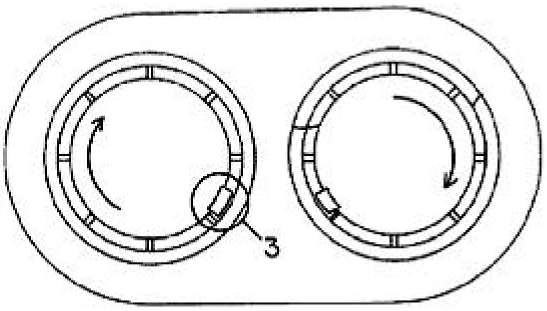

№ 5,831,354 (11/03/98)
Bootstrap Propulsion
System
Stopplecamp, Timothy J.

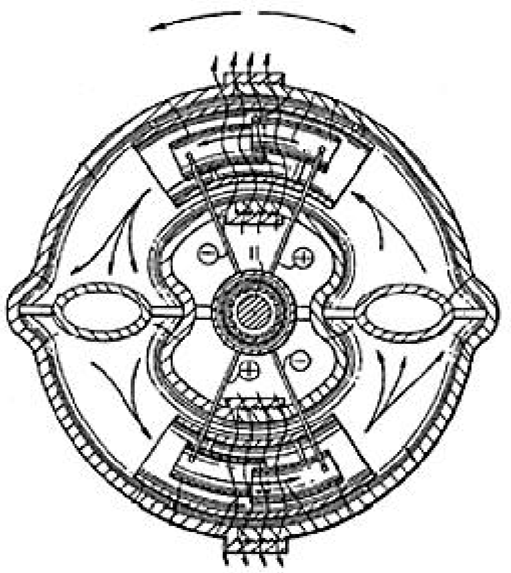

№ 5,410,198 (4/25/95)
Propulsion System
Butka, Kemal

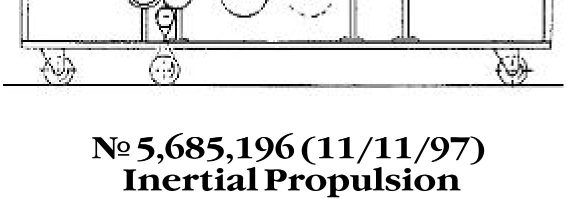

№ 5,685,196 (11/11/97)
Inertial Propulsion
Plus Device &
Engine
Foster, Richard E.

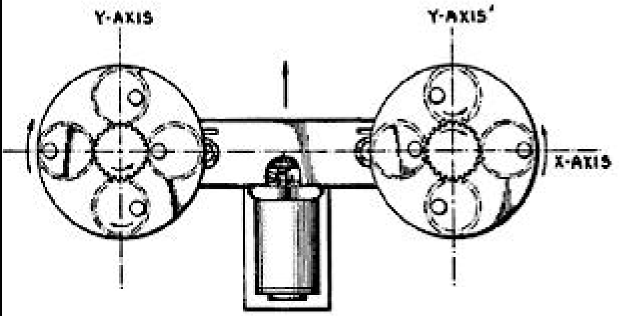

№ 5,167,163 (12/1/92)
Energy Transfer Device
McMahon, John C.

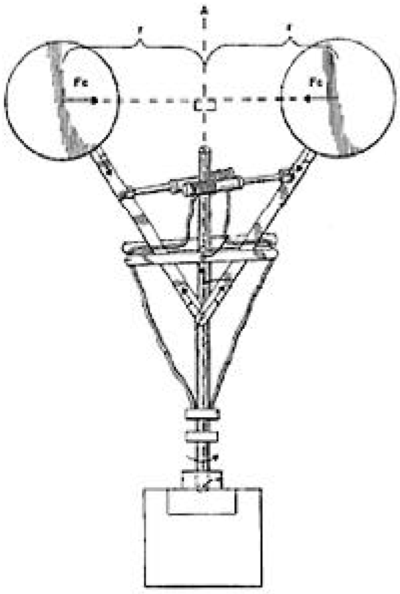

№ 5,557,988 (9/24/96)
Centripetally Impelled Vehicle
Claxton, John C.

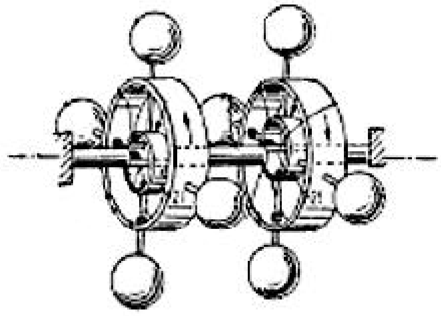

№ 5,156,058
Converting Rotary Motion to Lineal Motion
Bristow, Theodore R., Jr.

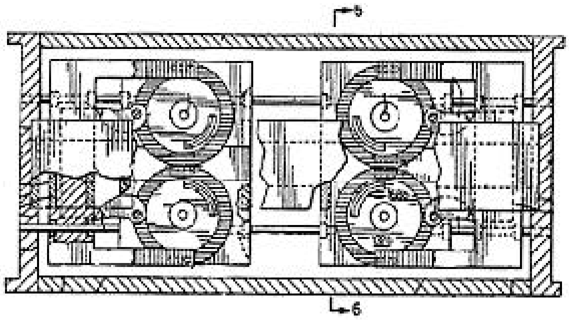

№ 5,182,958 (2/2/93)
Non-Linear Propulsion & Energy Conversion System
Black, James W.

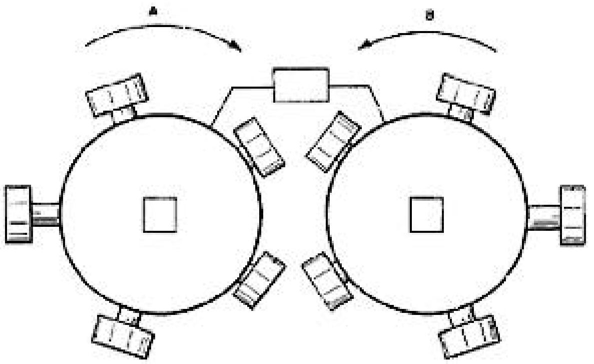

№ 5,782,134 (7/21/98)
Electromagnetically Actuated Thrust Generator
Bouden, James D.

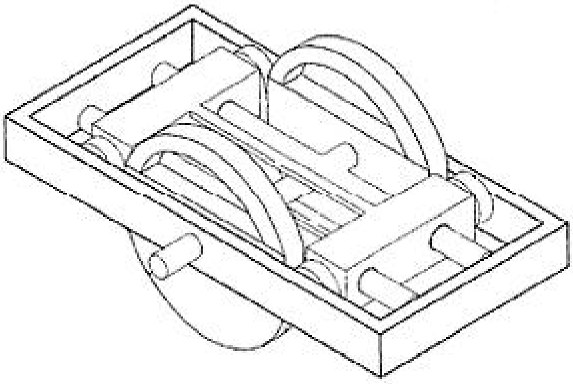

№ 5,488,877 (2/6/96)
Centrifugal Inertia Drive
Lieurance, Richard L.

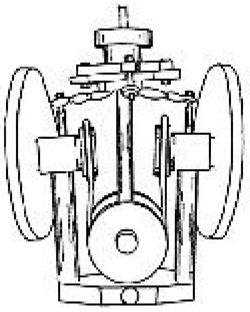

**№ 5,024,112(6/18/91)
Gyroscopic Apparatus
Kidd, Alexander D.**

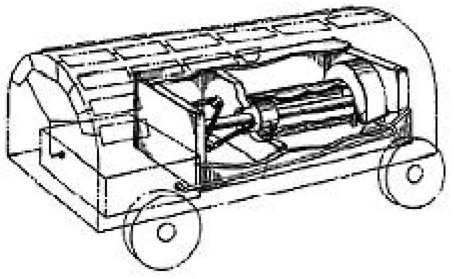

**№ 4,801,111(1/31/89)
System for Propulsion &
Positioning
Rogers, Charles E., et al.**

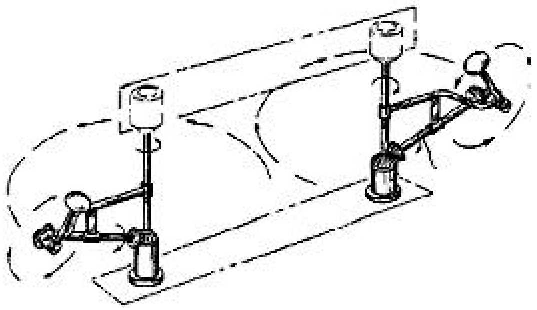

**№ 4,991,453(2/12/91)
Centripedal Device for
Concentrating
Centrifugal Force
Mason, Lyle M.**

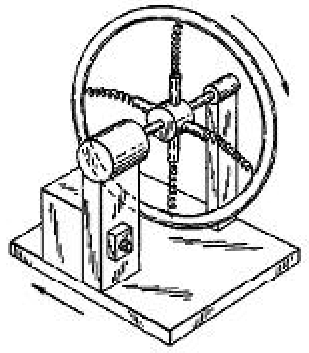

**№ 4,784,006(11/15/88)
Gyroscopic Propulsion
Device
Kethley, Lancelot I.**

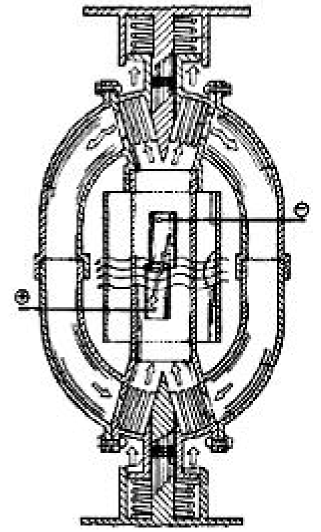

**№ 5,111,087(5/5/92)
Propulsion System
Butka, Kemal.**

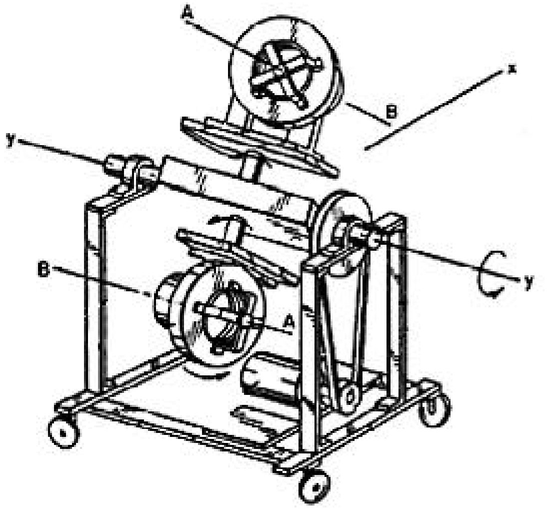

**№ 5,090,260(2/25/92)
Gyrostat Propulsion
System
Derloy, M.S.& Derloy M.**

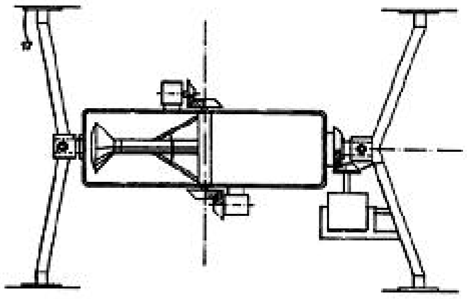

№4,884,456(12/5/99)
Centripedal Device for
Concentrating
Centrifugal Force
Zachystal, George J.

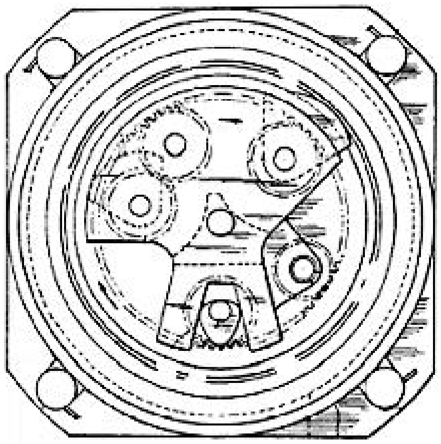

№5,042,313(8/27/91)
Conversion of
Rotational Output
to Linear Force
Montalbano, Paul J.

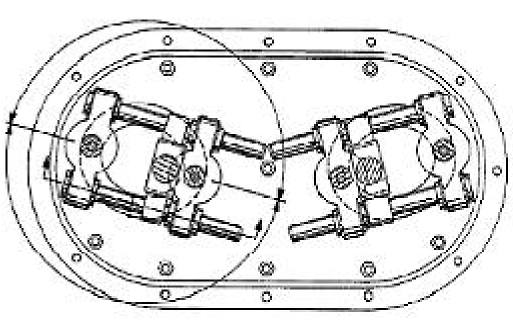

№4,770,063(9/13/88)
Universal Propulsion Powerplant &
Impulse Drive Unit
Mundo, James D.

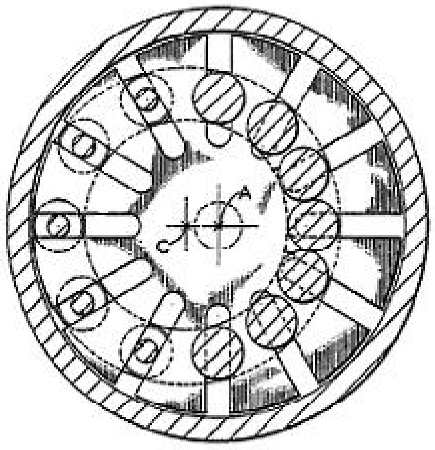

№4,788,882(12/6/88)
Flywheel
Fulop, Charles.

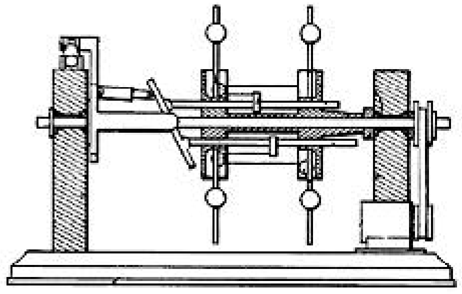

№5,054,331(10/8/91)
Controllable
Gyroscopic Propulsion
Apparatus
Rodgers, A.T.

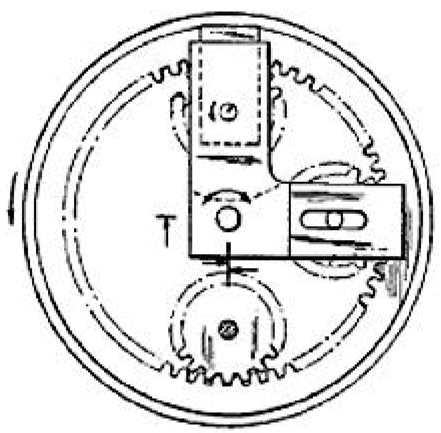

№4,856,358(8/15/89)
Conversion of
Rotational Output
to Linear Force
Montalbano, Paul J.

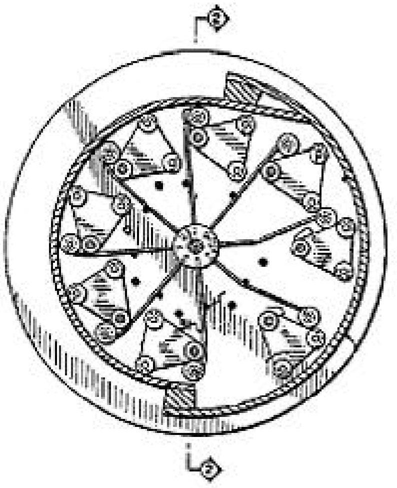

№4,744,259(7/17/88)
Apparatus for Producing a Directional Unit Force
Peterson, Oscar F.A.

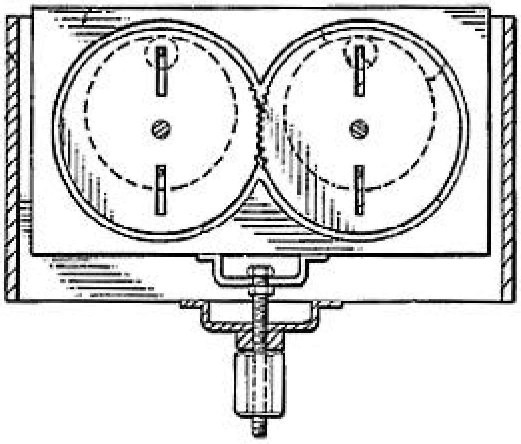

№4,712,439(12/15/87)
Apparatus for Producing a Force
North, Henry.

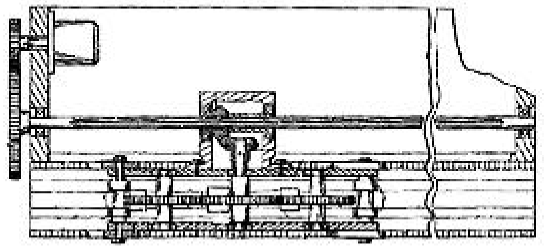

№4,242,918(1/6/81)
Mechanical Propulsion System
Srogi, Ladislaw G.

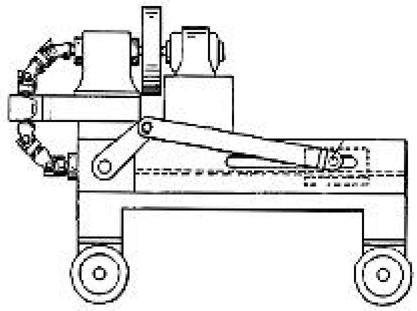

№4,674,583(6/23/87)
Impulse Drive
Peppiatt, Alvin C., et al

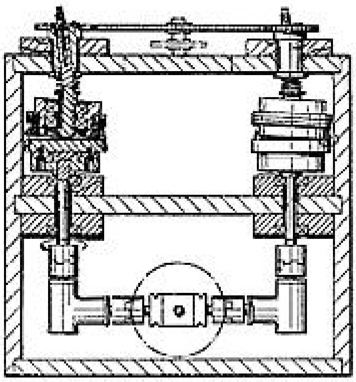

№4,409,856(10/18/83)
Propulsion System
de Weaver, Fred, III.

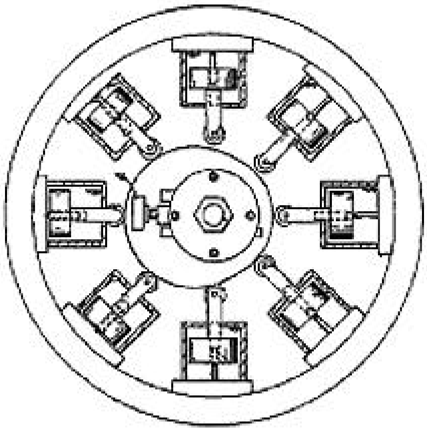

№4,579,011(3/1/86)
Propulsion Apparatus
Dobos, Elmer M.

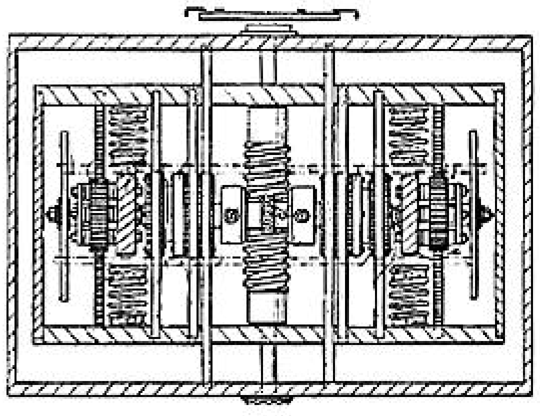

№4,577,520(3/25/86)
Mechanical Propulsion System
Colla, Joseph.

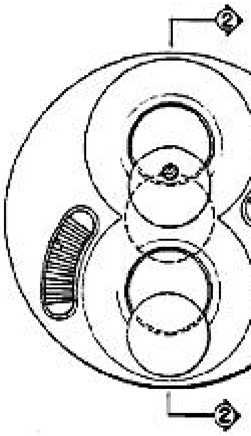

№4,631,971(12/30/86)
Apparatus for
Developing a Propulsive
Force
Thornson, Brandson R.

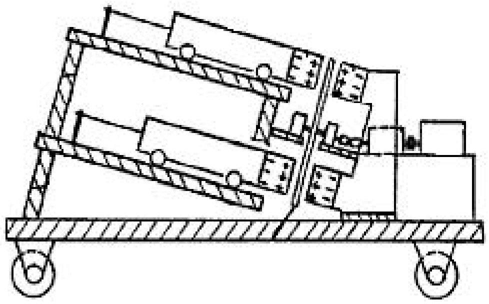

№3,889,453(6/17/75)
Propulsion System
Mast, Oscar.

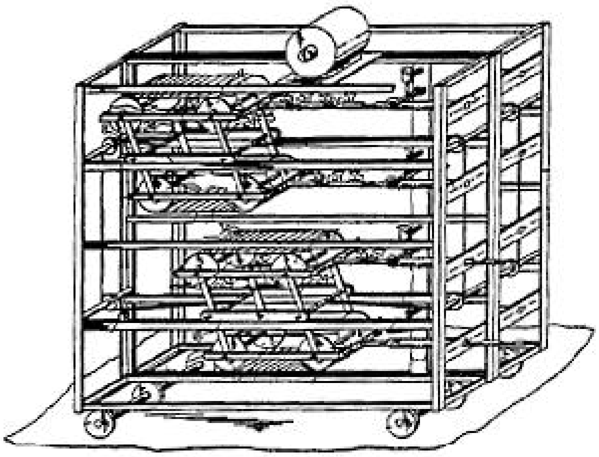

№4,726,241(2/23/88)
Continuous Force & Impact
Generator
Melnick, Harry S.

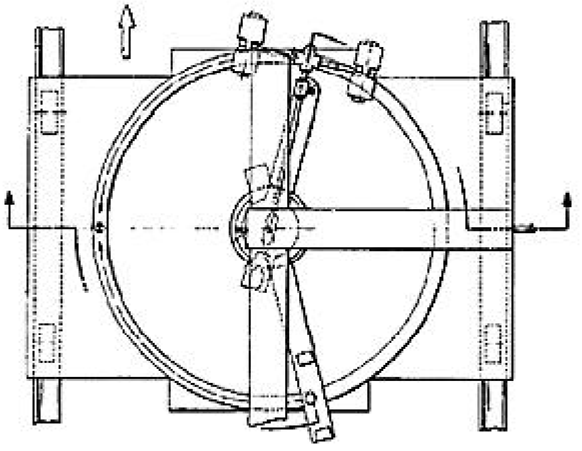

№4,238,968(12/16/80)
Conversion of Centrifugal Force
to Linear Force &
Motion
Cook, Robert L.

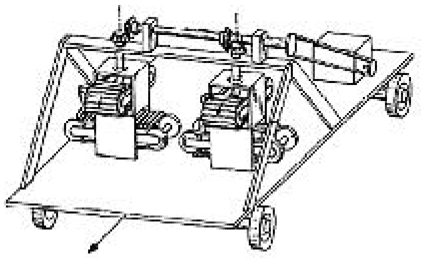

№4,261,212(4/14/81)
Unidirectional
Force Generator
Melnick, Harry S.

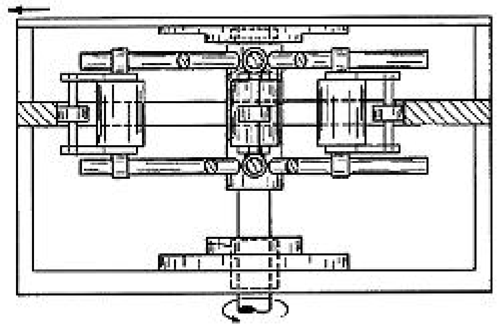

№ 3,968,700(7/13/76)
Converting Rotary Motion
into a
Unidirectional Linear Motion
Cuff, Calvin I.

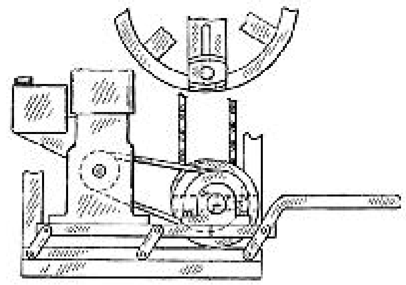

№ 4,347,752(9/7/82)
Converting Rotary Motion
to
a Rectilinear Force
Dehen, Frederick L.

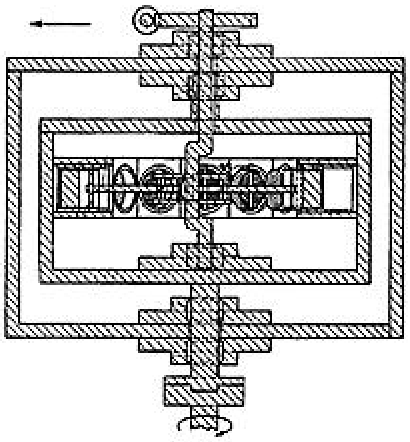

№ 3,998,107(12/21/76)
Converting Rotary Motion
into
a Unidirectional Linear
Motion
Cuff, Calviv I.

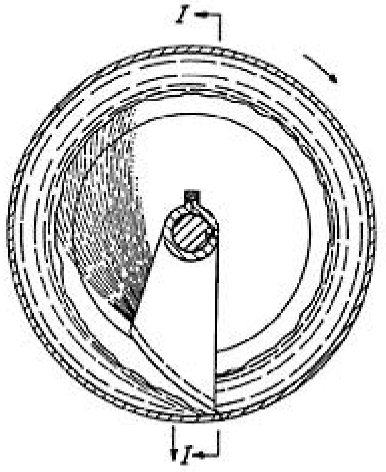

№ 3,979,961(9/14/76)
Propelling an Object by an
Unbalanced Centrifugal
Force...
Schnur, Nicolas J.

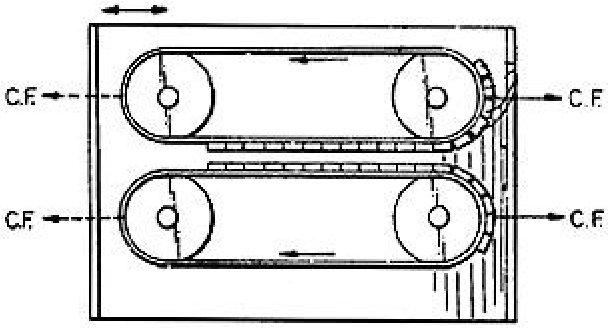

№ 4,398,431(8/16/83)
Mechanical Power Transmitting
System
Melnick, Harry S.

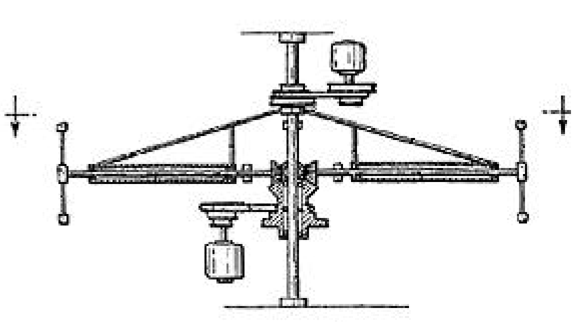

№ 3,555,915(1/19/71)
Directional Force Generator
Young, Hersey W., Jr.

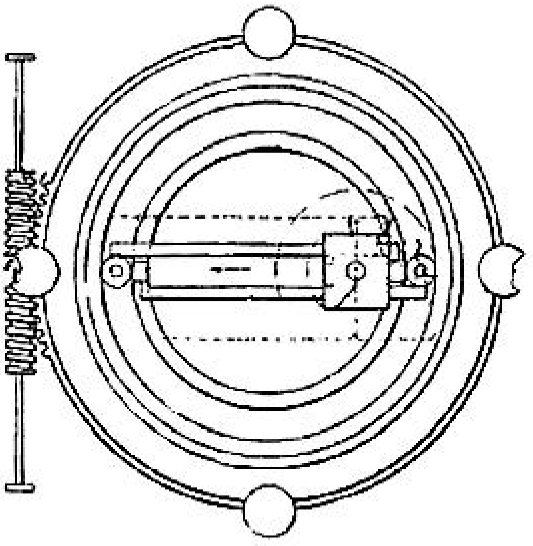

№ 3,750,484(8/7/73)
Centrifugal
Thrust Motor
Benjamin, Paul M.

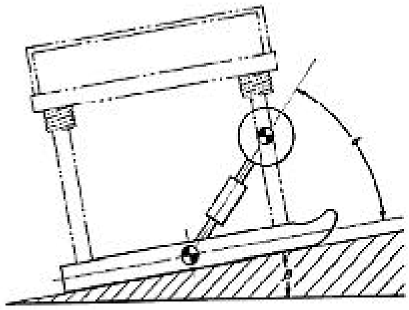

№ 3,916,704(11/4/75)
Vibratory Motion
Gaberson, Howard A.

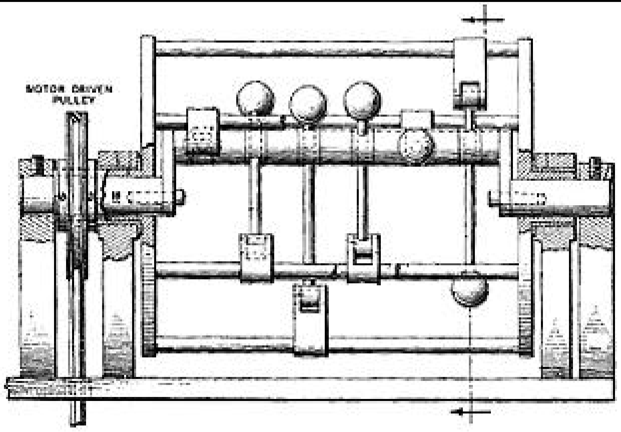

№ 3,810,394(5/14/74)
Centrifugal
Mechanical Device
Novak, Leo J.

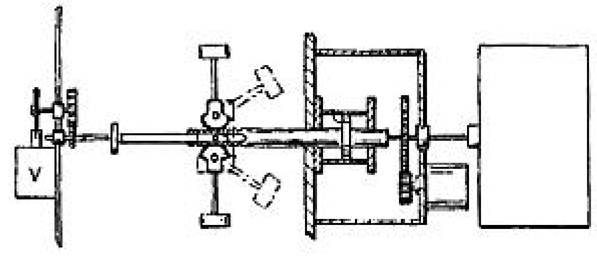

№ 3,863,510(2/4/75)
Inertia Engine
Benson, Everett H.

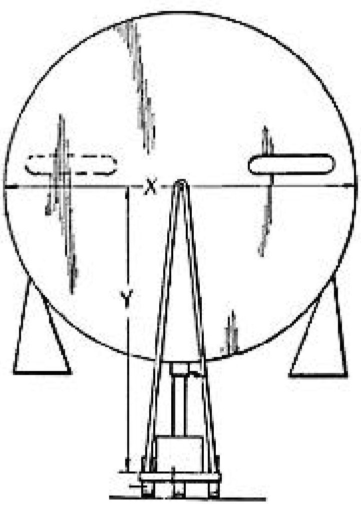

№ 3,756,086(9/4/73)
Propulsion System
McAlister, Roy E., et al.

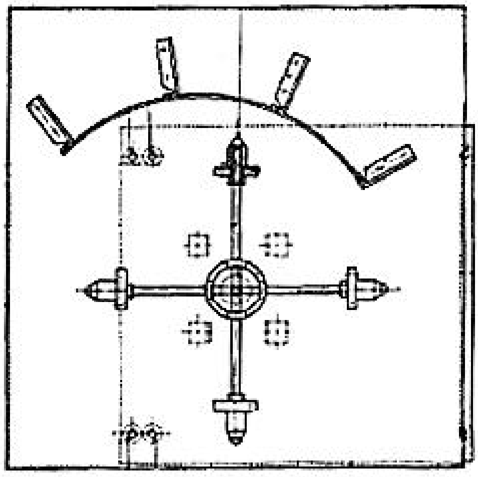

№ 3,807,244(4/30/74)
Device for Transforming Kinetic Energy
Estrade, Fernand.

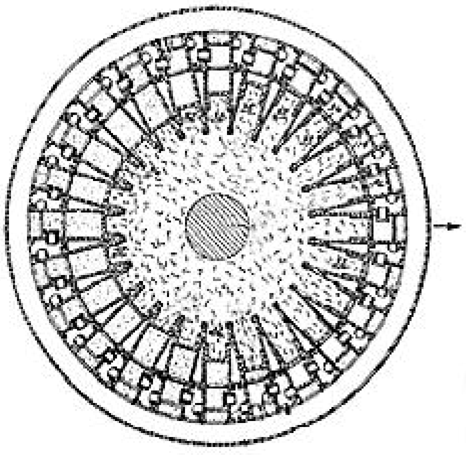

№ 3,584,515(6/15/71)
Propulsion Apparatus
Matyas, Laszlo B.

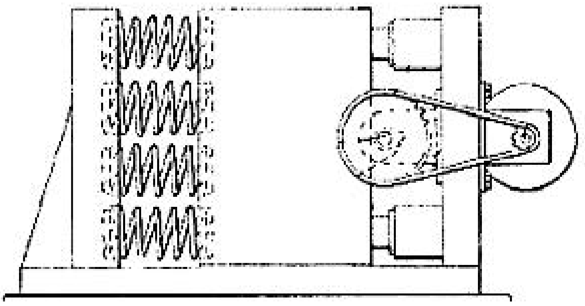

№ 3,266,233(8/16/66)
Inertia Propulsion
Device
Farrall, Arthur W.

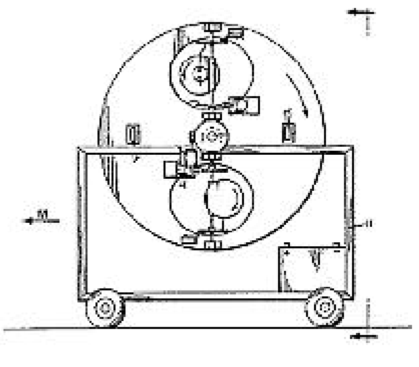

№ 3,653,269(4/4/72)
Converting
Rotary Motion
into
Unidirectional Motion
Foster, Richard E.

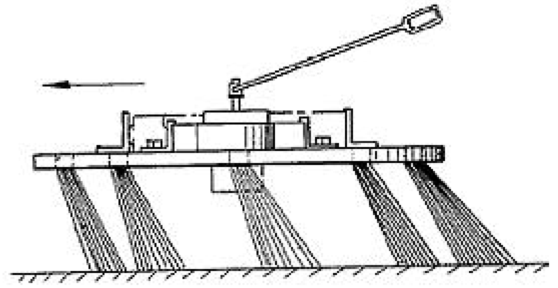

№ 3,196,580(7/27/65)
Toy Vehicle Having...Self-Contained
Drive Means
Rakestraw, Robert G.

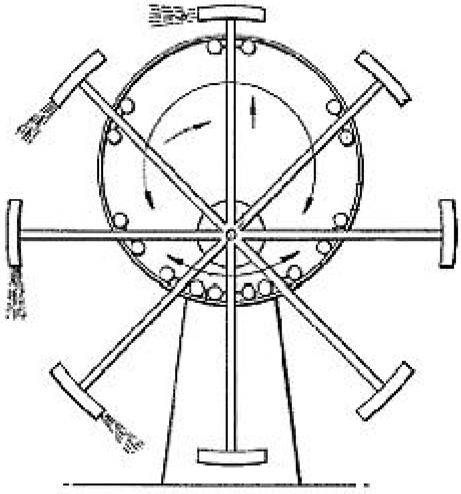

№ 3,177,660(4/13/65)
Propulsion
Apparatus
Haller, Paul.

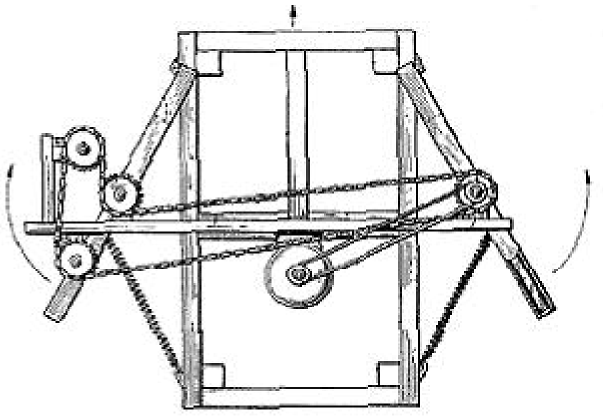

№ 3,683,707(8/15/72)
Propulsion
System
Cook, Robert L.

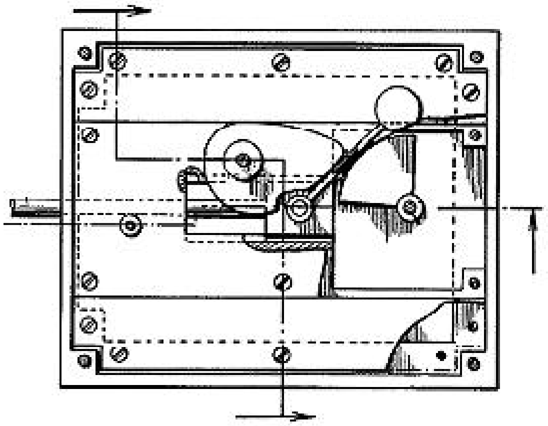

№ 3,182,517(5/11/65)
Variable
Oscillation System
Dean, Norman L.

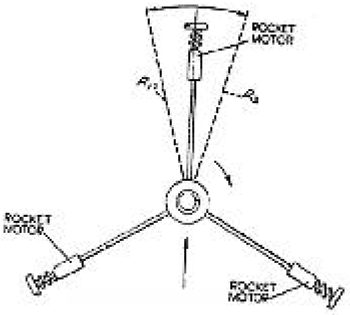

№ 4,087,064(5/2/78)
Orbital
Propulsion
Apparatus
Knap, George.

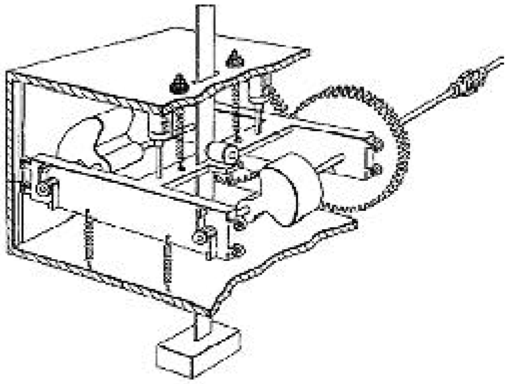

№ 2,886,976(5/19/59)
System for
Converting Rotary Motion into
Unidirectional Motion
Dean, Norman L.

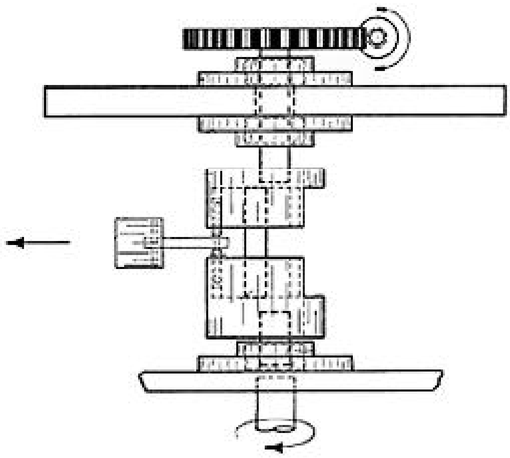

№ 4,095,460(6/20/78)
Converting
Rotary Motion into
Unidirectional Motion
Cuff, Calvin I.

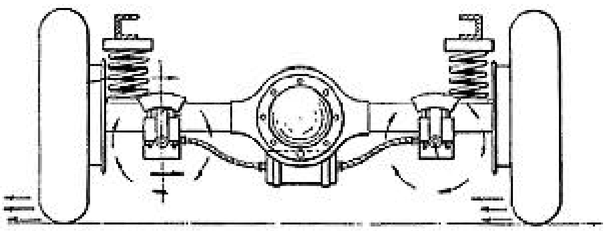

№ 2,639,777(5/26/53)
Method & Device for Imparting
Lateral Movement
Dull, Marshall L.

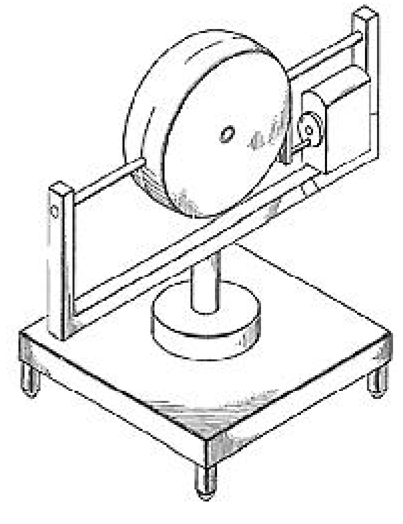

№ 3,203,644(8/31/65)
Gyroscopic
Inertial Space Drive
Kellogg, H. Dudley

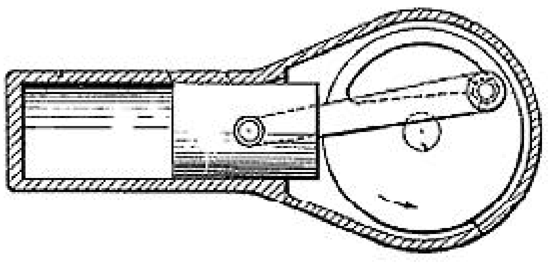

№ 2,636,340(4/28/53)
Direct Push
Propulsion Unit
Llamozas, Juan D. M.

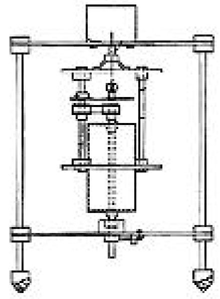

№ 5,150,626(9/29/92)
Translational Force
Generator
Navarro, T. L.

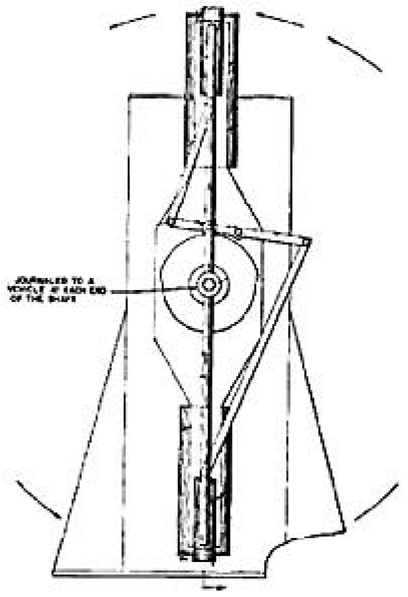
JOURNALED TO A VEHICLE AT EACH END OF THE SHAFT

№ 3,238,714(3/8/66)
Thrust Motor
Schur, George O.

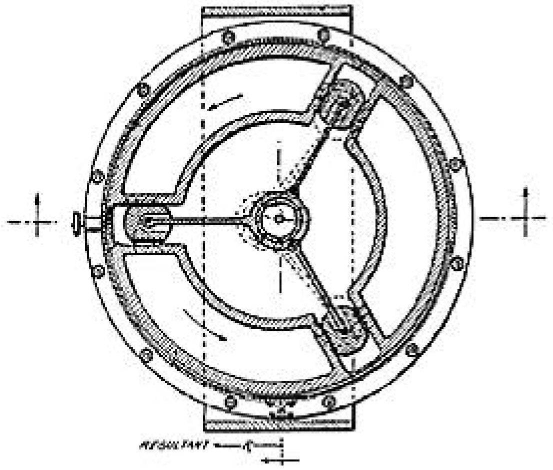

№ 2,009,780(7/30/35)
Centrifugal Variable Thrust
Mechanism
Laskowitz, Isidor B.

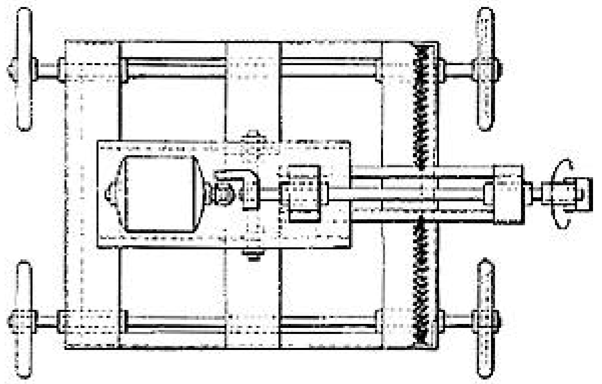

№ 2,088,115(7/27/37)
Reaction Motor
Neff, Tom.

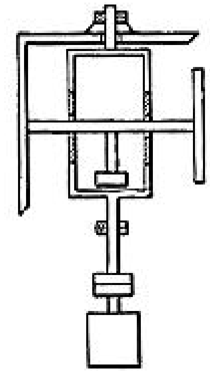

№ 3,404,854(10/8/68)
Apparatus for
Imparting Motion to a
Body
di Bella, Alfio.

The Free Energy Centrifugal Force Experiment

This experiment is conducted by Faraday Lab Ltd
in cooperation with V. I. Bogomolov

In the previous issue we wrote about the invention of Bogomolov V. I. The device shown in Fig.1 (and also on the cover) was constructed and tested in our laboratory.

Fig. 1
The Device

The simplified diagram is known as "Maxwell's Pendulum" (Fig.2). This is a disk set onto the horizontal axis with two attached strings. The upper ends of the strings are fixed to the crossbar.

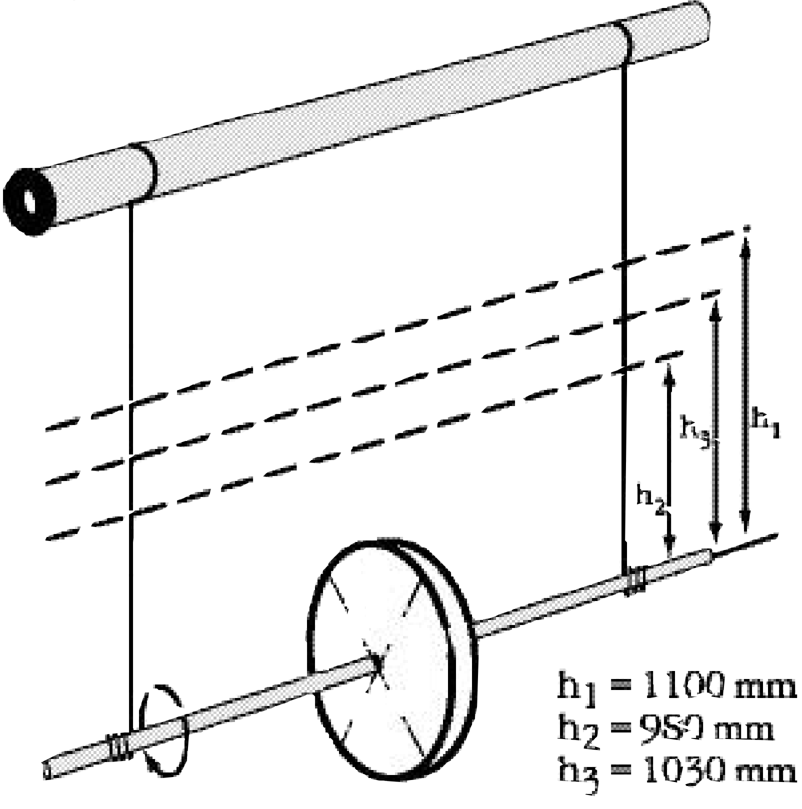

Fig. 2

If we wind the strings about the axis, the disk will lift (the height h) and store the potential energy of the Earth's gravitational field E = mgh (m - mass of the flywheel; g - free fall acceleration; h - the height of the mass fall). If we let the pendulum go, we can observe periodic "up-and-down" damped oscillations: first, the string spins and potential energy converts into rotation kinetic energy; upon reaching the lowest point the disk, still rotating, goes up using rotation kinetic energy and then converts it again into potential energy. This device is interesting: due to the energy conservation law we can observe the usage of rotation kinetic energy of the flywheel, measuring only the h-parameter, the lift height of the flywheel strings in the second semi-oscillation compared to the height with which the pendulum started falling in the first semi-oscillation, the difference h1 - h2 of the two semi-oscillations is directly proportional to the losses of rotation kinetic energy for air resistance work and friction.

We have made "Maxwell's Pendulum" more complicated by substituting the flywheel disk for the Watt centrifugal governor in accordance with its description in the article (1). The main difference of the device from Maxwell's flywheel is that the flywheel inertia momentum changes by the centrifugal force operation at the lever transferring of weights (the total weight - 1200 g) from the minimum radius position (40 mm) to the maximum radius position

(90 mm). At that, the levers compress the string with the force of about 16 N.

There have been three stages of the experiments. During the first stage we define the necessary lift height of 1100 mm at moment when the string winds around the axle. When the flywheel falls from this height, the rotation speed and centrifugal force are achieved, which move apart the loads to the maximum radius (90 mm) and which **completely compress the string**. During the second stage the loads are fixed at the minimum radius position (40 mm), thus the compress string centrifugal forces are not involved any more and we can measure the air resistance and friction **losses of kinetic energy**. At this stage the flywheel lifts up to 980 mm losing 120 mm. During the third final stage the load fixing rods are taken away and ***centrifugal forces are applied***. Just like during the second stage the Watt governor starts spin-falling from the height of 1100 mm. It compresses the string and lifts up to the height of 1030 mm thus **exceeding the second stage height by 50 mm**!

The experiment conclusions of the author:

1. At the third stage the flywheel exceeds the "loss height" h=980 mm. This means that the compression of the string was performed "for free".

2. At the third stage the flywheel converted kinetic energy into potential energy and lifted up 50 mm more. This means that in accordance with the principle of conservation of moment of momentum, the **outer compressed string energy force** has caused the moment of momentum change of the rotating masses, thus giving the additional acceleration to the flywheel and **increasing its kinetic energy "for free"**!

3. For the practical application of the Bogomolov generator in order to generate free energy it is necessary to achieve a high rotation speed (more than 10 000 rpm) and involve centrifugal forces to transfer the rotating mass to the less possible inertia radius difference.

The described centrifugal regulator diagram is not sufficient for larger experiments. For the industrial generator the author has a hydraulic-principle pneumatic string device drawing (know-how).

Editor: Also in this issue: the article on "Glen Gates Motor" by A. Akau. The principle is almost the same. The further implementation of this ideamay soon design purely mechanical power-generating systems, which use inertia and centrifugal forces.

References:

1. Bogomolov V.I. "The Bogomolov Generator", New Energy Technologies, Issue # 4, 2003.

Magnetic Power Inc. (MPI) is developing Self-Powered Generators. Together with its subsidiary, Room Temperature Superconductors Inc., (see the website www.ultraconductors.com), MPI has raised a total of more than $7 million from Angel investors to date.

Due to a pleasant surprise, solid-state electric power generators might be fabricated by modifying off-the-shelf utility hardware. If confirmed in coming weeks, this could result in serial fabrication by this time next year, since large devices of the type needed for conversion are presently manufactured worldwide.

Multiple modules may prove able to replace power plants. Smaller units appear practical for powering homes. Later, optimized designs might replace engines in every variety of vehicle. These generators may make possible very rapid utilization of fuel and pollution free electric power --a revolutionary, new, renewable energy alternative.

Accredited Angel investors can help speed the work needed to bridge into major capital, and accelerate this remarkable alternative. Multi-million dollar funding is on the horizon. Additional information is available privately. We welcome due diligence by qualified parties.

History of Perpetual Motion and Free Energy Machines

Web site http://www.phact.org/e/dennis4.html

Hundreds of people for centuries have attempted to solve the holy grail of energy production - **Perpetual Motion Machines**. Still to this day Eric Kreig gets weekly emails from people who feel they are about to get one working. Eric is very skeptical about free energy. But his collection of information is very interesting. Eric offered a special prize for those who will agree to publicly test their over-unity devices.

♦ **Villand de Honnecourt** 13th century had a drawing of one.

♦ **Leonardo da vinci** made a number of drawings of things he hoped would make energy for free.

♦ Jesuit priest, **Johanes Taisnerius** worked on a magnetic based perpetual motion machine.

♦ **Mimara** in 1518 designed a "self-blowing windmill".

♦ **John Dee** of 16th century reported seeing one - but wasn't allowed a closer look.

♦ **Cornelis Drebbel**, 1610, was an alchemist and magician supposedly made one.

♦ **Robert Fludd** 1630 proposed many machines -people were trying to patent variations of Fludd's device in the 1870's.

♦ **Edward Somerset** 1638 demonstrated many free energy water wheels to the king of France.

♦ **1635** - first of many English perpetual motion machine patents granted. - By 1903, 600 such patents had been granted. Free energy claimants love impressing people with patents.

♦ **Ulrich von Carnach** in 1664 in Germany designed a perpetual ball-moving machine.

♦ Scientist **Jean Bernoulli** (1667 - 1748) proposed a fluid energy machine.

♦ **Bockler** in 1686 made designs for self powered water mills.

♦ In 1712 **Johann Bessler aka Orffyreus** investigated 300 different perpetual motion models and claimed he had the secret of perpetual motion and got much investment money.

♦ **Dr Conradus Schiviers** in 1790 made a belt driven wheel.

♦ **Sir William Congreve** in 1827 tried a machine running on capillary action.

♦ Britisher **Henry Prince** 1866 described the first partially submerged perpetual motion machine.

♦ **Mark Zimara** of Italy had a huge air powered machine that never worked.

♦ **Horace Wickmam** of the USA got a patent to a machine with many balls that just rotate around.

♦ Austrian, **Alois Drasch** patented a machine in the US in 1868.

♦ German, **George Andreas Bockler** proposed 'self operating mills' using variants of Archimedes screws.

♦ **E. P. Willis** of Connecticut made money off a perpetual motion machine in 1870 - people eventually found out a secret source of power to it.

♦ **Charles Redheffer** in 1812 in **Philadelphia** made much money on a perpetual motion machine, tried to restrict scientific evaluation, but was eventually debunked.

♦ Scottish shoemaker **Spence** designed a magnetic based machine, which was debunked.

♦ **John Worrel Keely** of Philadelphia in 1872 (he also had a traveling show of exhibitions). He fooled many scientists with a machine, which appeared to run on water.

♦ **John Gamgee** in 1881 got considerable support for a machine very similar to Dennis Lee's, which used liquid ammonia - it got vaporized from heat readily available, thus expanding it would drive a piston. Gamgee thought the vapor would condense to liquid to start the cycle over again. The Navy appeared to have been fooled and showed it to president Garfield - it never went anywhere Tom Napier agrees that Dennis Lee may have resurrected Gamgee's engine.

♦ **J. M. Aldrich** was arrested for getting investors for his free energy machine in 1899 - he some how was able to avoid conviction and conned many investors. One of whom was finally able to inspect the machine found a hidden spring.

♦ **T.H. Moray** in the 20's demonstrated a "radiant energy device" to many people who were unable to find a hidden power source. Some how, the secret was said to have been forgotten.

♦ **Lester Hendershot** in 1928 got an Army commandant to endorse his free energy machine - but it was later found to have a hidden power source in the motor. His sons believe Lester lost his notes and that maybe they can rediscover how to get it to work.

♦ **Viktor Schauberger** claimed to have discovered some special vortex energy in water. Since he died in 1958, I don't know if his claims have been replicated, but people are still studying his works.

♦ **Mr Papf** in 1966 was a conspiracy believer alternative car engine got a few investors but killed someone during a demo. He tried to blame the problem on an investigating skeptic. He disappeared and became part of urban legend of scores of people, which the big conspiracy has been silencing for generations.

♦ **Guido Franch** was convicted of fraud in 1954 and 1973 of selling rights to distribute little green pills that would convert water into gas. He hid behind conspiracy theory and secrecy to avoid fair testing. A number of people have run this scam and many people still believe there are pill repressed by the oil companies.

♦ **Garabed T.K. Giragossian** in 1917 claimed to have a free energy machine. He was one of the early frauds to hide behind conspiracy theory. Woodrow Wilson signed a resolution offering him protection from some kind of conspiracy. After much fanfare, and delaying tactics his machine turned out to be a giant flywheel, which was charged up with energy slowly and put out a lot of energy for just a second. In spite of lack of proof of anything significant his followers still bothered the US congress for recognition.

♦ **Otis Carr** in 1958 sold stock for a company to manufacture UFO's and free energy machines from Oklahoma. He claimed inspiration from Tesla.

♦ **Edgar Cayce** even babbles about "Motor's with no Fuel".

♦ **McClintock** was claimed to use air as a fuel and had a patent.

♦ The **Evgray** machine scammed many investors (who didn't know how to test 'depleted' batteries).

♦ **Arnold Burke** in 1977 collected $800,000 of investor money (again, mostly from bible believing farmers) for a 'self acting pump'. He tried to hide behind religion. He called his device Jeremiah 33:3' Finally, an open test was done in 1979 and found a hidden source of electricity. His believers (with an infinite supply of denial) still raised $250,000 to get out of a fraud conviction. He still went on making lots of claims with no evidence.

♦ **Robert Adams** in 1977 made all kinds of conspiracy and OU claims in New Zealand. Many still believe in the Adams Motor. - He had the obligatory battles over patents, debates over theory, debates over power measurement and dabbling in other areas of alt physics. He's 80 years old and said to have a scientific education. He has a lot of theories about Aether. Skeptical information.

♦ **Robert Stewart** in 1978 got over $3,000,000 of investment money (much from farmers) for a closed cycle engine using freon rather than water. This 'engine' used the same scheme John Gamgee tried to sell the navy in 1882. -Eric Krieg thinks this is the same approach Dennis Lee's free energy machine works on. I think he just faded away (people claim he was silenced).

♦ **Rory Johnson** of Elgin Illinois, claimed to have invented a cold fusion, laser activated, magnetic motor that produced 525 HP, weighed 475 lbs, and would propel a large truck or bus 100,000 miles. After signing a number of dealers, he moved all his equipment out of his labs, moved to CA and died. (To this day, folks say he was silenced by OPEC).

♦ **RJeseph Maglich** was a physicist and claimed to have a device in 1978, which harnessed fusion power from seawater. They say they put power in and get more power out. I've never heard of them since.

♦ **Howard Johnson** got a patent for a device that claimed to make free energy from a motor like device. As of 11/02, Stephen was claiming he would soon mail out samples of working ones. By 12/12/02, Stephen dropped out of contact.

♦ **Keith Kenyon** had a device claimed to produce more energy than consumed. Calculations seemed to not take power factor (also known as phase angle) into account. It never openly had its output hooked up to its input. Even Dennis Lee tells his followers that all the *over unity motors* (claim to make more electricity than consumed) usually just fool people who can't measure power factor.

♦ **Muller:** Bill Muller and Carmen Muller of Germany raised money on an over unity motor and got a few followers, but never actually demonstrated one working.

♦ **Dennis Lee** Since 1988 has been promising to demonstrate free electricity "in a month or two". He is much like Newman in his mixture of religion & extremist politics, evasion of qualified investigators, endless promises, threatening detractors, etc. He had a Fischer engine, a CRD device and now, and OU motor device. Dennis has invested in Searle and Stanley Meyer and joined Pantone in 2001 in a 50 state tour.

♦ **Stanley Meyer** 1996 claimed to have a water powered car and was also big on mixing Christianity and patriot politics in with fringe science. Meyer was found guilty of fraud after his Water Fuel Cell was tested before an Ohio judge. It is rare for an inventor to be prosecuted for an invention that does not work, but Meyer's problem was that he had been selling "dealerships", offering investors the "right to do business" in Water Fuel Cell technology. Meyer refused to allow anyone else to measure his device. Dennis Lee invested in him. He died in early 1998.

♦ **Joseph Newman** in 1984 claimed to have a free energy machine based on alternative physics. Like many perpetual motion inventors, he sued the US patent office. Many people wrongly measured the true power output of this machine, (they didn't realize you must specially calculate power for non sinusoidal current consumption). He now refuses to ship a unit for testing. Ten years ago, inventor Joseph Newman gave an open week-long demonstration in the Super dome in New Orleans. Over 9,000 people attended from across the country (including Dennis Lee who reportedly wanted to join his ideas with Newman). Newman is suing some former investors he claims are trying to steal his invention. I give more information. Evan Soule of Newman's organization offers a rebuttal.

♦ **Bruce De Palma** had a machine in 1986, which appeared to one Electrical Engineering professor to put out 4 times more power than consumed. Turned out it was just a measurement error. (We could go on and on with such examples).

♦ **Dr. Potapov** sold a device that was claimed to produce more energy than consumed.

Editor: Eric's skeptical remarks can be tolerated but sometimes he writes about the things he does not know. Potapov and his team have created a technology, which have been replicated and sold by dozens of companies in Russia (see the cover picture of the Susorov heat generator). There are several devices in St. Petersburg to be tested. The efficiency is 200 % and more. The operation principle: electric motor rotates in water and causes the heating of water. We have also received the information about another Russian invention, namely the close cycle vortex generator: "water rotation – vapour – turbine – electric power " (the power is from 100 kW).

Alexander V. Frolov

♦ **John Bedini** claims to have a free energy device.

♦ **Mr. Finsrud** is a Norwegian artist who made a sculpture where a metal ball moves for weeks apparently with no outside influence.

♦ **Don Watts** of Las Vegas in around 1990 had a patented **CEACU**, which stands for Centrifugal Energy Amplification and Conversion Unit - it turned out to be one more investment fraud.

♦ **Stephan Marinov** claimed to have proven much alternative physics and to have contacted a Methernitha that claims to have a Free Energy Machine. He committed suicide on July 15, 1997 but he left behind some intense rants.

♦ **Greg Watson** sold kits for a rolling ball and track that were thought to have over unity in 1997. Attempts of replication I know of have failed.

♦ **CETI** These people have claimed to have a device that puts out anomalous yet small amounts of heat - maybe cold fusion. They raised millions. As of 10/97, they have not been willing to have me come over and see for myself. They have said, *we want people to think it doesn't work so we won't have competition.* I've wondered if the energy may have come from not accounting for friction effects from the cooling flow through the pellets. (**Jed Rothwell**, a rational editor of Infinite Energy Magazine says *"however, tests with CETI cells at Motorola, SRI and the French Atomic Energy Commission show no measurable friction"*). Milton Rothman has a response: "I openly admit that I have not followed all the history of cold fusion claims and am generally ignorant on the subject".

♦ In January 1998 **Barbara Hickox** allegedly has a patent dating from 1981 for a fusion powered free energy system.

♦ **Paramahamsa Tewari** of India claims to have a device that is 200% efficient.

♦ **RQM** is a Swiss company selling FE machines found at www.rqm.ch.

♦ **Ted King** is looking for people to buy stock for a car he plans to drive across the country using just 2 12-volt batteries. You can contact Ted if you want to buy shares.

♦ **Bruce Perreault** has claimed to have discovered a new element, the plasmatron, ion pump, radiant energy device, etc.

♦ **Daniel Pomerleau** of Canada claims to have something that works, but he isn't interested in releasing it (as 12/97).

♦ **Entropy Systems** of Ohio 1999 Sanjay Amin got 1.6 million investment dollars for a device that would violate the 2nd law of thermodynamics. I've asked them to apply for my prize for proof. - They seem to have folded. See a rational review of Amin's claims.

♦ **Brian Collin** of Australia claims revelation from God *(like most of these people)* to make a free energy device. A Stephen Mark claimed he invented it.

♦ **Kawai** and **Takahashi** - both of Japan claimed in the 90's to have overunity devices.

♦ **Carl Cella** claimed to have one of the many cars said to run on water.

♦ **Robert LeBreton** in 1999 claimed to be making a 600 hp free energy machine.

♦ **Renzo Boscolio** in 1999 in Italy claimed to have low-energy nuclear reactions but refused to supply real proof that he promised to people who came out from Infinite Energy Magazine.

♦ **Doug Konzen** of Seattle says he has an overunity motor in Jan of 2000 that anyone can see.

♦ **Troy Reed** of Oklahoma was ready to issue licenses for manufacturing his permanent magnet motor etc. He's taking investment money for a device he claims puts out more heat energy than input energy. He says he's dumped a few million into his designs over the last 10 years and had little interest in my prize offer when I contacted him in 2001.

♦ **Kipper Motor.** In 10/00, Steve Elswick thought it was overunity. But David Sligar who paid $175 for plans could not get it working.

♦ **Ludwig Brits and Victor Christie** in 2001, claim the **Lutec** free electricity over unity motor/generator will soon be in Australia.

♦ **Jasker** in 2002, an Irish company has claimed to have free electricity. (Some have said it is just a joke).

♦ **Confidential Technology -Wayne Cochran** died 12/29/02 Crazy Jack Carey took over. They have been promising FE soon for about 20 years now.

♦ **Tom Bearden's MEG device:** many articles on the Internet.

♦ **Gurbakhsh Singh Mann of India** claims to have invented gravity and buoyancy perpetual motion machines.

♦ **Michael J. Marshall** in Las Vegas has a device called QSFG , which stands for quick start fuel-less generator he says 64 nations have asked him to build factories.

♦ **Carl Tilley and Robert Kibbey** in 6/2001 in Tennessee have claimed some over unity device and other stuff. Their demonstration failed in 9/02 and as of 11/02, they have been evading proper demonstrations.

♦ **Stephen Walker** in 9/2002 promised to send me and several others a free energy machine in mail.

♦ **PerEnDev** promised to make some kind of free energy by means of magnets.

♦ **Bill Muller** of Canada in 2003 claimed to have some kind of OU device. Independent tests found it under unity.

♦ **Energie:** In 2003 this Greek company promised a homopolar magnet (De Palma design) FE design by June of 2003. They actively seek investors.

♦ **GWE Genesis World Energy** In 2002 this group claimed to have 400 people who developed some device that sounds like it separates water into H2 and O2 using less energy than mainstream science says is possible. As of 2003, they have evaded independent confirmation. More information is available.

♦ **Steven Greer (Disclosure Project)** In 02.2003, Steven announced he had discovered some real sources of Free energy, which he promised to make sure get proven to the general public.

Tesla Energy Science Conference and Exposition

1220 L Street NW, Suite 100-232 Washington DC 20005
Tel: 800-295-7674
Fax: 301-513-5728
www.IntegrityResearchInstitute.org
Email: iri@erols.com

Saturday, November 8, 2003
Wireless Electricity Seminar

9:30 AM 10:15 AM

Opening Address: "Space Solar Power" – Dr. **Paul Werbos** Program Director, National Science Foundation

10:15 AM 12:15 PM

Special 2-Hour Presentation: "Nicola Tesla and the development of RF Power Systems" - Dr. **James Corum**, Physics Professor, research Scientist and Inventor and **Kenneth Corum**, Physicist, teacher, consultant.

2:00 Pm 3:00 PM

"Power Engineering Scalar Field Theory: Faraday vs. Maxwell and Longitudinal Wave Demonstration" – Professor **Konstantin Meyl**, Engineer, Author, Inventor of the Demo-Set, Professor at the University of Berlin

3:15 PM 4:15 PM

"Wireless Energy Through the Earth-lonosphere Cavity" – Dr. **Elizabeth Rauscher,** Nuclear and Astrophysicist, Inventor of the ELF Earthquake Predictor and Triangulator.

4:15 PM 5:15 PM

"Masters of the lonosphere HAARP Modifies the Polar Electrojet" – **BBC Video on Tesla Technology**, with Wiliam Terbo, demonstrating ionospheric excitation by radio waves.

6:30 PM 7:15 PM

Masters of the Ionosphere – BBC Video Presentation **repeated** for the Special Tesla Evening Event attendees

7:30 PM 8:30 PM

"The Wardenclyffe Dream: Tesla's Plan for Wireless Worldwide Distribution" – **Dr. Marc Seifer** Professor, Author of the best-selling book *Wizard, The Life and Times of Nikola Tesla*, presents an illustrated historical account.

8:30 PM 9:00 PM

High Voltage Tesla Coil Demo – **High Voltage Device** with **Music**: The Tesla Tower wind ensemble composition by Prof. **Holland**, Skidmore College

Sunday November 9, 2003
Electrotherapeutic and Tesla History Day

9:30 AM – 10:30 AM

"A Family Perspective on the Personality of Nikola Tesla. Review of the Popular Interest in this Scientific Icon" with Q and A Session – **William Terbo**, Engineer, Closest living relative of Tesla (Grand-nephew). Founder and Director of the Tesla Memorial Society.

10:30 AM 11:30 AM

"The Search for Free Energy: Genius, Vision and Psychology of Invention" – **Keith Tutt**, British Writer and Author of the book *The Scientist, The Madman, The Thief and their Lightbulb* that includes Tesla.

11:30 AM 12:15 PM

"Electrotherapy with Tesla Coil Design: Intorduction to Bioelectromagnetics" – Dr. **Thomas Valone** Physicist, Professional Engineer, Author of the new book, Bioeletromagnetic Healing: A rational for its Use.

12:15 PM 1:00 PM

Exhibitors Presentations – **Brief Presentations by exhibitors** (5-10 minutes each)

2:00 PM-3:00 PM

"Emerging Opportunity: cancer Electromagnetic Frequency Therapy" – Dr. **Mark Neveu** President, the National Foundation for Alternative medicine.

3:00 PM 4:00 PM

"Turn of the Century Electrotherapy Discoveries" – **Jefrey Behay** Director, The Turn of the Century Electrotherapy Museum.

4:00 PM 5:00 PM

"How a Crushed Leg Helped Me Discover High Voltage Electromagnetic Healing Device" – **Ralph Suddath**, Third generation Tesla Electrotheraphy Inventor, radio Host, Entrepreneur.

5:00 PM 5:15 PM

Closing Remarks – **Thomas Valone** Program Coordinator

The Conference Report Compilation can be purchased from T. Valone, Integrity Research Institute.

"Nornikel" and Alternative Energy

"Norilsky Nikel", the largest Russian non-ferrous metal producing company has announced the beginning of the cooperation with the Russian Academy of Science. Metallurgists are planning to invest into the development of the hydrogen energy and fuel cell R&D projects of the Russian Academy of Sciences. Accordingly, last week an agreement was signed. **Mikhail Prohorov**, the General Director of "Nornikel" said that they are ready to invest 20-40 million dollars annually into R&D projects of the Academy in this field.

"Expert" magazine, # 43, November 17-23, 2003

PERENDEV Magnetic Motor

Review

www.perendev.co.za/products/magnetic_motor/index.html

Shown below is a Magnetic Motor with Alternator (Patent Pending):

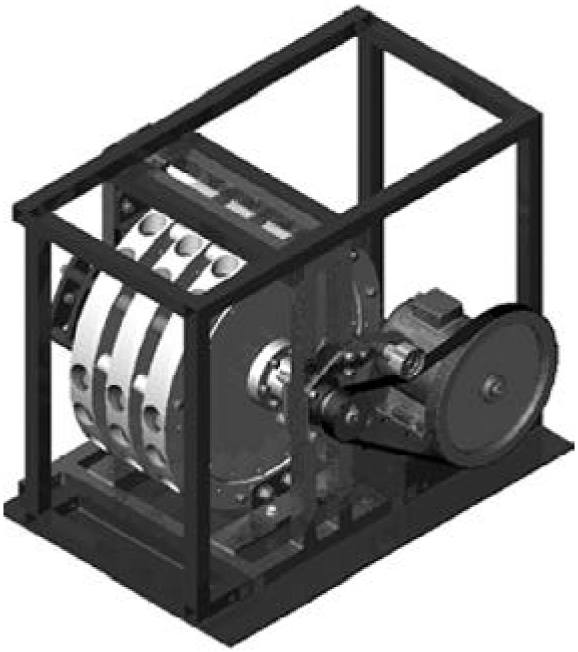

Fig.1

This is the "PerEnDev" (Perpetuum Energy Devices) company latest motor predicted to deliver some 60 kW. Motors can be built up to 4 mW.

Background of the Magnetic Motor

The Company started in 1969 with the development of a magnetic motor to drive an alternator to produce electricity. The concept was sound but the problem they faced was that the magnets that were available at the time (ferrite magnets) were not very powerful. The unit ran but had very little power and the project was abandoned.

In 1996 the "Perendev" company started to investigate the possibility of reviving the original idea, magnets had now come a long way and were very powerful (rare earth).

Fig. 2
Rare earth magnets

After the engineers successfully designed a new model, the company developed a 3 rotor system producing 6 kW, coupling this to an alternator through an 11:1 gearbox to produce electricity in 220volt and 380volt, single or 3 phase.

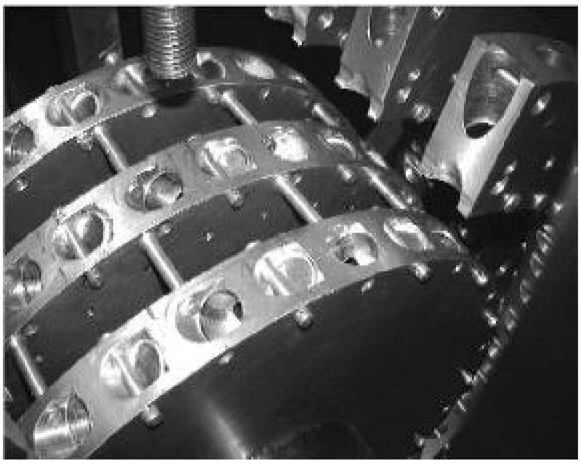

Fig.3

The units will be ideal for the small and large user, such as emergency, military, farms, plots, industrial and large users... in fact any application requiring power.

Key Benefits Foreseen:

- No fuel costs
- Constant running
- Reliable
- Cost Effective
- Portable (6kw-120Kw)

Editor: In 2002 "Faraday Lab" got an offer from the company to buy their 6 kW generator for testing at the price of about $6,000. However, no contract was signed. The company is currently engaged in redesign of their motors and is looking for regional partners but they are not yet ready to produce on the line. In our issues we will try to inform the readers on their activities. Alexander V. Frolov

Internet Pages

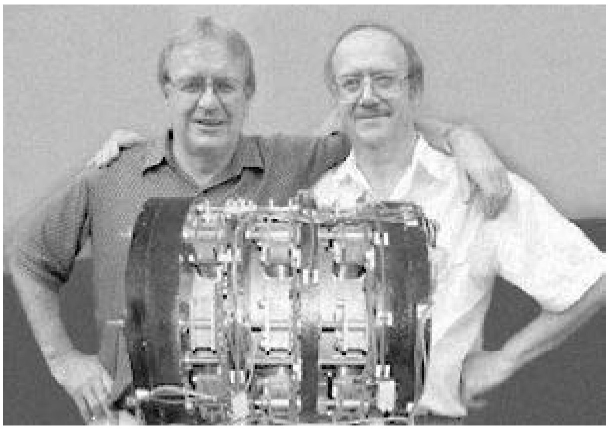

Lutec 1000

Jasker Power System

Electric Radial Motor

industrial and large users... in fact any application requiring power.

Key Benefits Foreseen:

- No fuel costs
- Constant running
- Reliable
- Cost Effective
- Portable (6kw-120Kw)

Editor: In 2002 "Faraday Lab" got an offer from the company to buy their 6 kW generator for testing at the price of about $6,000. However, no contract was signed. The company is currently engaged in redesign of their motors and is looking for regional partners but they are not yet ready to produce on the line. In our issues we will try to inform the readers on their activities. Alexander V. Frolov

Internet Pages

Jasker Power System

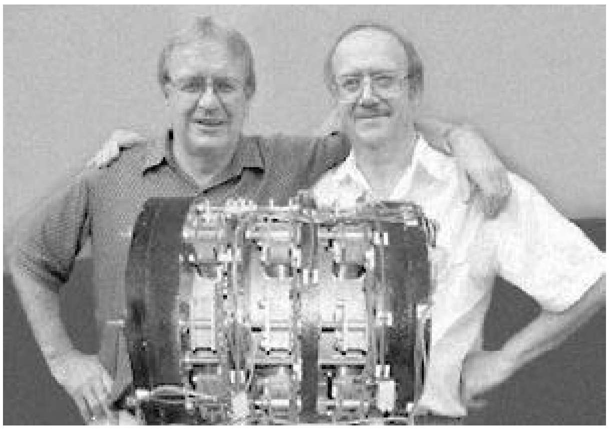

Lutec 1000

Electric Radial Motor

18 Meter Perpetual Wheel

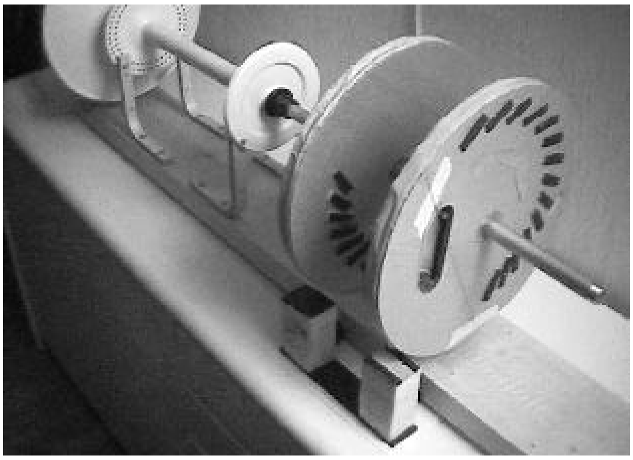

Minato Wheel, E. Vogel, Sweden

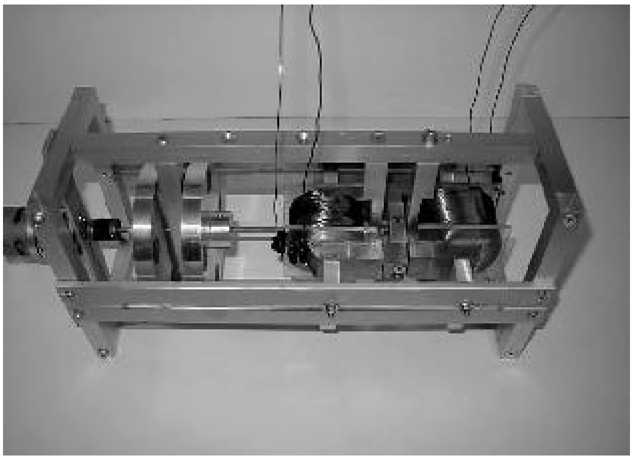

Generator from James W. German

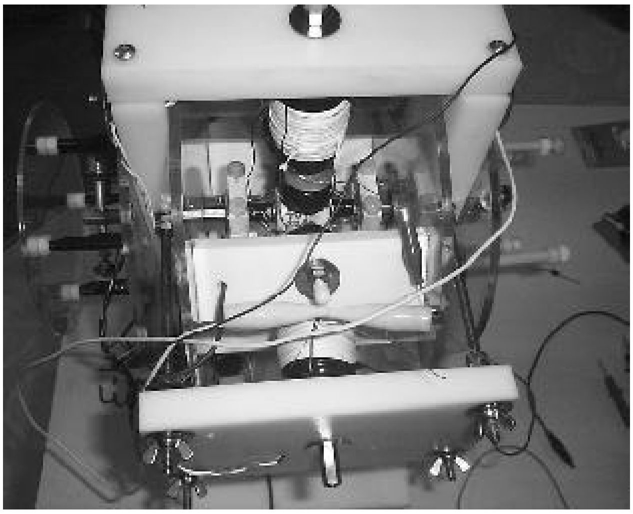

Konzen
Pluse Motor

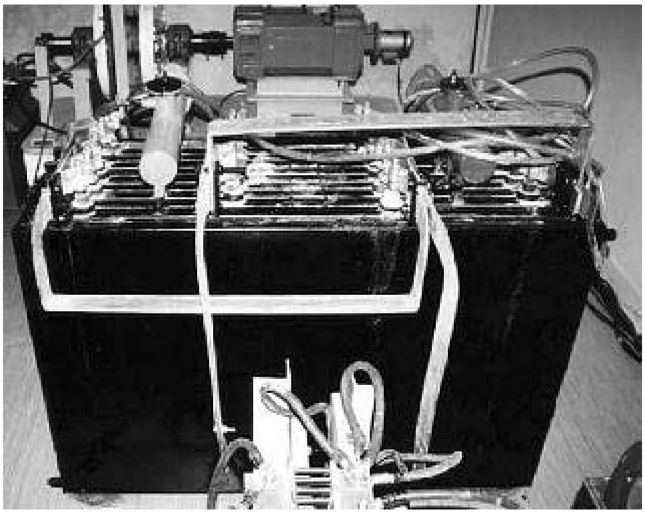

200-300 percent over-unity energy Space
Power Generator

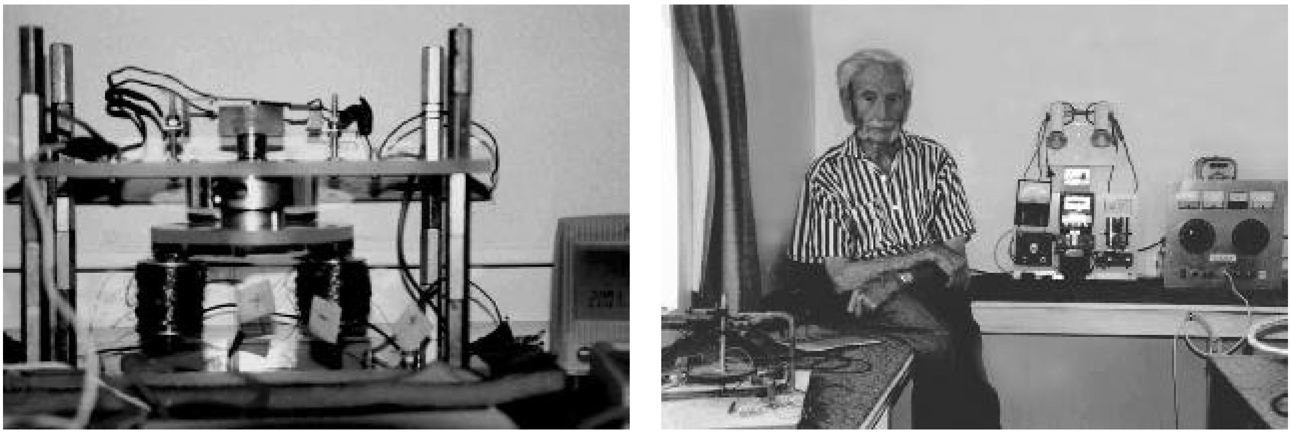

Adams Motor, Robert Adams

The RotoVerter

94.0
WD-40
KELVIN 400

Bendini Pulse Motor

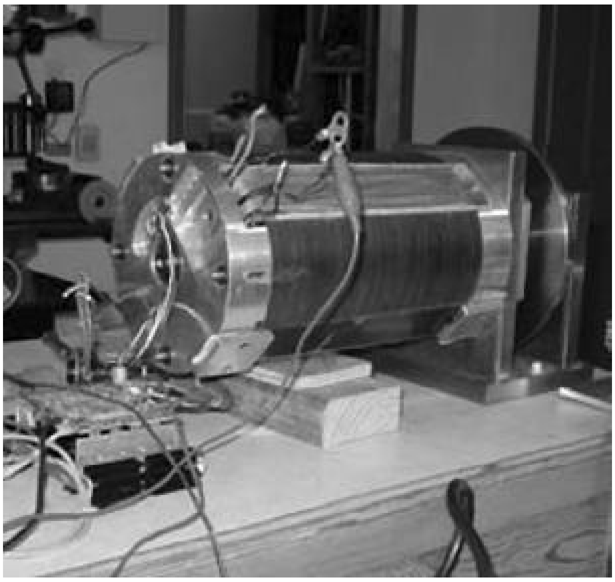

Motor based on Flynn Technology

Finsrud Device

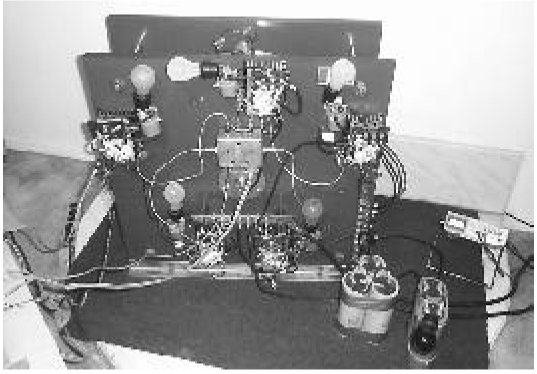

The Muller Unit

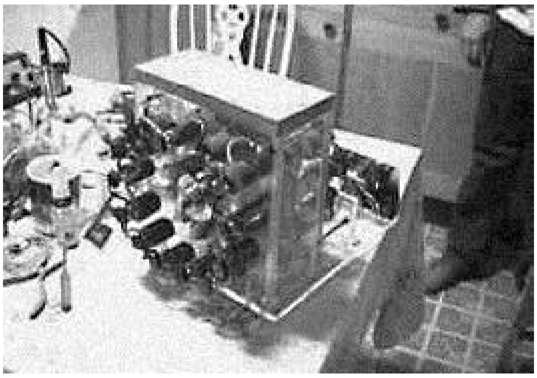

Alte Olsen Generator

Vladimir Matveev's Electrical Generator

Entirely New Kind Of Generator Invented!

(1-26-2003) TASHKENT, Uzbekistan (UPI) – An Uzbek inventor said he has created a type of electrical generator that does not rely on the principle of electromagnetic induction – on which all existing generators are based.

The new generator employs a concept called **magnetic conductivity modulation** and it has potential applications in industry, communications, households and even the military, explained Vladimir Matveev, the inventor, a specialist in electronics.

Matveev said he is convinced he has created a fundamentally different machine.

"All electrical machines I know are based on the principle of interaction between the magnetic fluxes (lines of force) of their rotor (rotating member) and stator (portion that remains fixed)," he explained.

Such machines, Matveev said, are based on electromagnetic induction, a property of energy discovered by Michael Faraday, an English physicist and chemist, in the 19th century. The machines produce electrical current either by moving a conductor across a magnetic field or by regulating the flux of that field.

"My machine has a principal difference," Matveev told United Press International. "The magnetic field of its stator does not interact with the magnetic field of the rotor (because) its rotor is not a magnet – the rotor only changes the magnetic resistance of the stator," he said.

The stator in Matveev's generator contains a magnetic core with a permanent magnet and a detachable winding. A rotor with changeable magnetic resistance is placed at a cutoff point in the core's magnetic field. It is composed of alternating magnetic and air parts and can operate in either linear or rotary form.

When the rotor is set in motion, its alternating components pass through the magnetic core's cutoff point. When the magnetic part passes through the cutoff point its magnetic resistance decreases. When the air part passes through, its resistance increases.

This pulsing of resistance results in changing the magnetic conductivity of the magnetic core, which in turn produces an alternating electrical current in the core's winding. The frequency of the winding's current can be controlled by regulating the rotor's speed or by changing certain qualities of its magnetic or air parts. Also, the generator's electrical

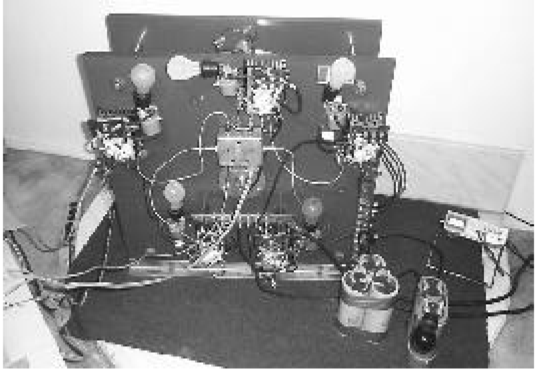

The Muller Unit

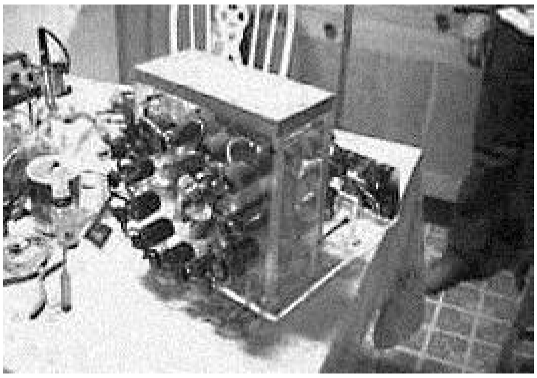

Alte Olsen Generator

Vladimir Matveev's Electrical Generator

Entirely New Kind Of Generator Invented!

(1-26-2003) TASHKENT, Uzbekistan (UPI) – An Uzbek inventor said he has created a type of electrical generator that does not rely on the principle of electromagnetic induction – on which all existing generators are based.

The new generator employs a concept called **magnetic conductivity modulation** and it has potential applications in industry, communications, households and even the military, explained Vladimir Matveev, the inventor, a specialist in electronics.

Matveev said he is convinced he has created a fundamentally different machine.

"All electrical machines I know are based on the principle of interaction between the magnetic fluxes (lines of force) of their rotor (rotating member) and stator (portion that remains fixed)," he explained.

Such machines, Matveev said, are based on electromagnetic induction, a property of energy discovered by Michael Faraday, an English physicist and chemist, in the 19th century. The machines produce electrical current either by moving a conductor across a magnetic field or by regulating the flux of that field.

"My machine has a principal difference," Matveev told United Press International. "The magnetic field of its stator does not interact with the magnetic field of the rotor (because) its rotor is not a magnet – the rotor only changes the magnetic resistance of the stator," he said.

The stator in Matveev's generator contains a magnetic core with a permanent magnet and a detachable winding. A rotor with changeable magnetic resistance is placed at a cutoff point in the core's magnetic field. It is composed of alternating magnetic and air parts and can operate in either linear or rotary form.

When the rotor is set in motion, its alternating components pass through the magnetic core's cutoff point. When the magnetic part passes through the cutoff point its magnetic resistance decreases. When the air part passes through, its resistance increases.

This pulsing of resistance results in changing the magnetic conductivity of the magnetic core, which in turn produces an alternating electrical current in the core's winding. The frequency of the winding's current can be controlled by regulating the rotor's speed or by changing certain qualities of its magnetic or air parts. Also, the generator's electrical output – its voltage – can be controlled by changing the configuration of the rotor's components, Matveev explained.

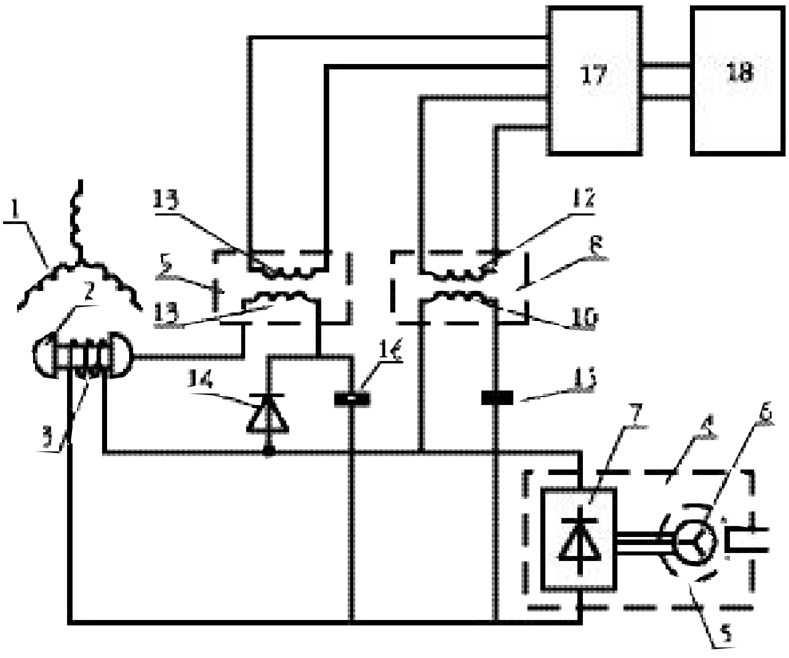

Electrical machine assembly - FEDORO V F 1996.10.10 1996RU-119822 (1998.09.27) G01R 31/34, 31/02

Addnl.Data: VOROBEI V K MATVEEV V A

NOVELTY - Device has electrical machine, non-brush field exciter with diode rectifier, two dynamics double-winding transformers, which primary windings are located on rotor and secondary windings on stator. One end of primary winding of second transformer is connected to rotor housing; another end is connected to common point of conductor which is connected between capacitor and cathode of semiconductor diode, which other terminals are connected to direct current terminals of diode rectifier. Semiconductor diode is connected to cathode group of rectifier; capacitor is connected to anode group. Secondary windings of transformers are connected to actuating member through generator of signal, which is proportional to resistance of insulation in excitation winding circuit.

USE - Electric power production

Matveev said his generator is different from an invention by Howard Johnson of the United States. Johnson discovered how to build motors that can run without input of electricity or any other kind of external energy. He obtained a patent in 1973 for describing electrical generation using only the energy contained in the atoms of permanent magnets.

Matveev also said some Russian inventors have experimented with a generator similar to his. Their generator changes its magnetic conductivity by changing temperature. However, the machine requires a lot of time to be heated and cooled alternately and results in a current frequency much lower than what generally is used in industry. Moreover, the Russian generator requires high steel density and greater mass.

Matveev's machine generates electric energy of industrial frequency. Furthermore, he said his machine is simple, reliable and requires less steel and mass than conventional generators. It also can be adapted to flows of low speeds, such as weak water or wind streams. Matveev tested the generator in his former household in Kazakhstan before he patented it in Uzbekistan.

"I want to pass the invention on to all mankind," he said.

Boris Abdurakhmanov, director of the Uzbek Koinot (Cosmos) design office and head of the laboratory of semiconductors and photoelectricity of the Institute of Electronics of the Uzbek Academy of Sciences, told UPI: "Matveev has offered a fundamentally new approach to a problem of the creation of electric power generators."

Utilization of the Gravity Force in Power Engineering

Editor: In the previous issue we made a review of Mikhail Smeretchanski's invention "Perpetual motor with magnetic control elements" (# 4, 2003). This invention was patented; shown below are some pages from the patent itself. You can contact the author by e-mail: smeretchanski.mikhail@wanadoo.fr, or the postal address: 13 av. Rochambeau 3800 GRENOBLE FRANCE

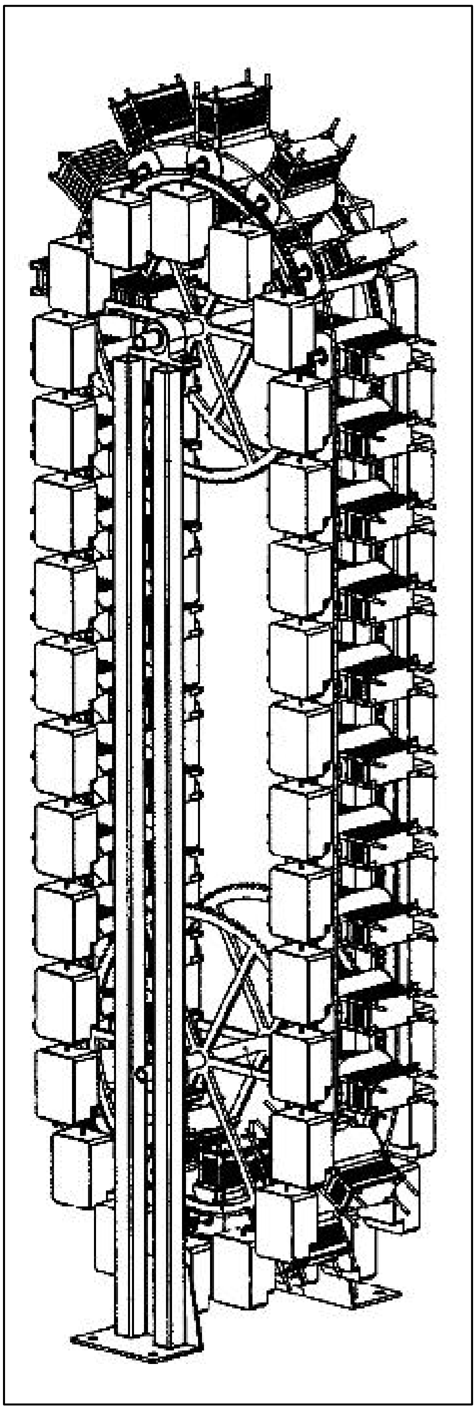

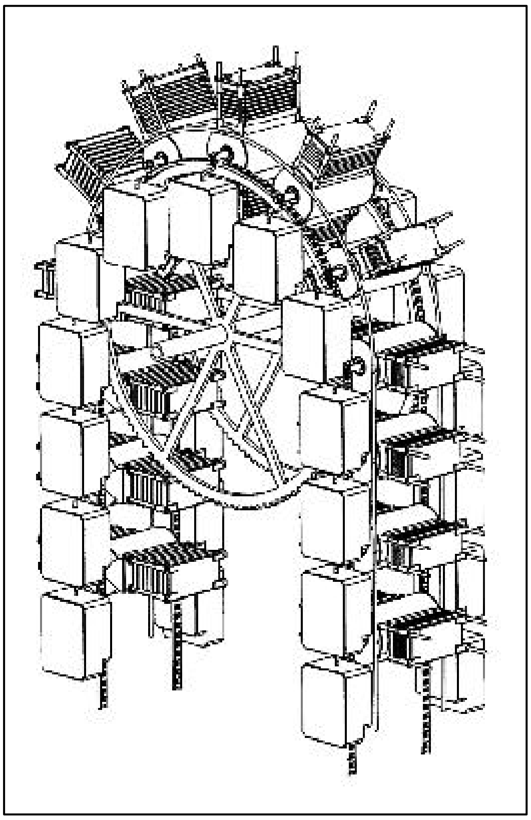

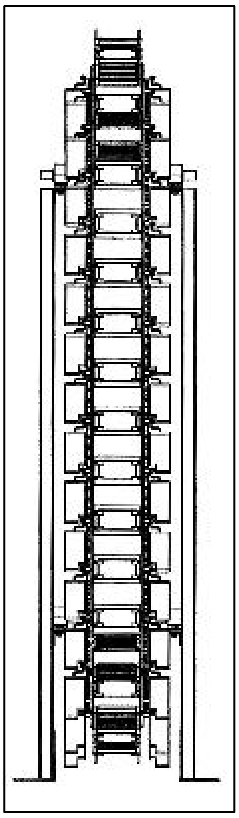

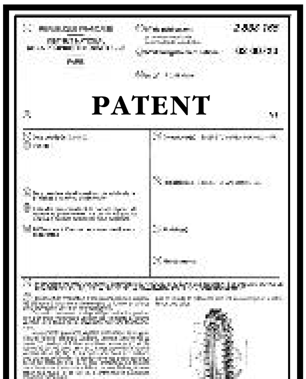

PATENT

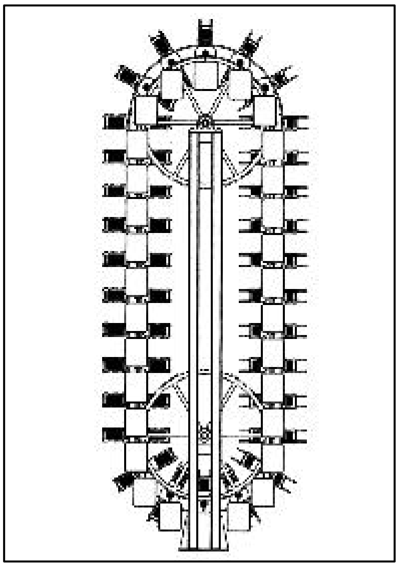

The Most Interesting Articles - 2003

In the year 2003 we have published more than 120 articles, mainly on the alternative energy, new propulsion principles and the development of ether dynamics.

"Cold fusion", low energy nuclear reactions (LENR) and highly-efficient electrolysis were represented by the works of Professor Kanarev, E. Storms, P. Hagelstein, T. Chubb and E. Cartlidge.

Almost every "New Energy" issue features articles on highly-efficient permanent magnet motor or generator experiments or patents of the kind. The articles by T. Valone, E. Vogel, A. Akau, T. Hardwood, A. Francouer and S. Abramov are of great value for engineers and researchers in this field.

The interesting articles by M. Filo, R. Koontz and V. Bogomolov reveal several principles of designing and building purely mechanical energy generating systems. All these systems possess one common feature: namely, the fact that a substance (a solid or fluid working mass) is accelerated in the gravitational field or in the centrifugal force field. The design characteristics of the system make it possible to obtain free energy.

We have published several articles by S. Gerasimov and his colleagues on the reaction-less propulsion device experiments. We hope that in the future we will keep in touch with these researchers.

The article by D. Reed on the new physics development concepts is worth mentioning too. The article by E. Sorokodum on new energy sources and the article by V. Chernobrov on the research into the active properties of time are of no less importance.

A. Egorov's article on the ball lightning makes it possible to form a number of experimental approaches to the problem of creation multi-purpose stable plasma objects.

Electrogravity questions, water-powered vehicles, longitudinal electromagnetic wave technologies, the Searl effect commercialization, capillary engines, resonance high-efficiency power-engineering, single-wire power lines, heat pumps and other similar research works were presented to our readers in the year 2003. We have minimized theoretical articles and we hope that the magazine has become more interesting and useful for the new energy practice development.

Alexander Frolov
St. Petersburg, Russia

There is interesting newsgroup of the site: http://www.overunity.com,
Stefan Hartmann, email: harti@harti.com

Stefan Hartmann writes: This is the answer I got from Finsrud; it seems that the film they have done about him will be shown [on Discovery Channel] in 2004 when [if] I interpret this norwegian text right...

Subject: Documentary Film
Reidar Finsrud,

Dear Sir,

The Australian Broadcasting Company together with The Discovery Channel, have funded my company to make a documentary film entitled 'A Machine To Die For', the story of Perpetual Motion and the search for 'Free Energy'.
I have read extensively about your Perpetual Motion sculpture and we would be interested in traveling from Australia to Norway in order to film this now famous piece of work. At the same time we would request an interview with yourself.
If you are in agreement with this it is our intention to be in Norway on the 10th and 11th of July 2003. This documentary will be released worldwide and should produce considerable interest in your Gallery.

Mark Eliot
Catherine Jarvis and Mark Eliot
Romany Mill Studios

Hydrogenics Secures $3 Million from Japan Auto Makers

Letter from RemyC remyc@prodigy.net
Contacts: investors@ hydrogenics.com

Hydrogenics Secures $3 Million in Test Equipment Orders Japanese Auto Makers Select Greenlight Power for Fuel Cell Testing Equipment

TORONTO, Sept. 30 – Hydrogenics Corporation (Nasdaq: HYGS; TSX: HYG), a developer and manufacturer of fuel cell products, announced today that its wholly owned subsidiary, Greenlight Power Technologies (Greenlight) has received orders totaling US $3 million for fuel cell test stations from two leading Asian automotive customers. Both are repeat orders for multiple machines.

Cumulatively, Greenlight is providing seven Fuel Cell Advanced Test Stations (FCATS) to two major customers. The first order is for three 3 kW PEM FCATS L-Series test systems. The second is for three 12 kW PEM FCATS H-Series test systems and one 60 kW PEM FCATS HX-Series testing station. The stations are expected to be delivered over the next two quarters.

"Our decision to open an office in Tokyo, Japan over three years ago continues to pay dividends," said Pierre Rivard, President and CEO of Hydrogenics. "The Japanese market is advancing quite aggressively towards the commercialization of fuel cell technology, as evidenced by our growth in test equipment sales to the region over the past three years. We are delighted to secure these major contracts with customers who are clearly setting the pace in fuel cell development."

Greenlight fuel cell testing systems provide high precision instrumentation, combined with full-featured software. They deliver fuel cell testing results that prove the reliability, repeatability and versatility required for world-leading fuel cell development programs.

About Hydrogenics Corporation

Hydrogenics Corporation (Nasdaq: HYGS; TSX: HYG) is a leading clean power generation company, engaged in the commercialization of fuel cell technology and test stations for fuel cells. The company is building a sustainable business, in a potentially "game changing technology" for transportation, stationary and portable power. Hydrogenics, based in Mississauga, Ontario, Canada, has operations in British Columbia, Canada, Japan, the United States and Germany. For more information, please visit www.hydrogenics.com .

About Greenlight Power Technologies, Inc.

Greenlight Power Technologies Inc. (http://www.greenlightpower.com), a wholly-owned subsidiary of Hydrogenics Corporation, is a leading global supplier of testing and diagnostic equipment to the fuel cell industry. It has supplied fuel cell test equipment to the world's premier fuel cell stack manufacturers, component manufacturers, system integrators and research organizations. Greenlight provides a full suite of test equipment for fuel cell stack, stack component, reformer and system testing for companies focused on portable, stationary, and transportation fuel cell applications.

Hydrogen Fuel Cells

Anuvu Incorporated

3980 Research Drive
Sacramento, CA 95838
USA
Tel: (916) 921-7040
Fax: (916) 921-7044
Email: anuvu@anuvu.com

www.anuvu.com/home.html

Now our readers can get familiar with some hydrogen fuel cell development photos. It is worth of a note that the prices are high for the majority of people. For example, a hydrogen fuel cell car costs 100 000 dollars FOB Sacramento, CA. The next picture features a fuel cell van, which costs 150 000 dollars.

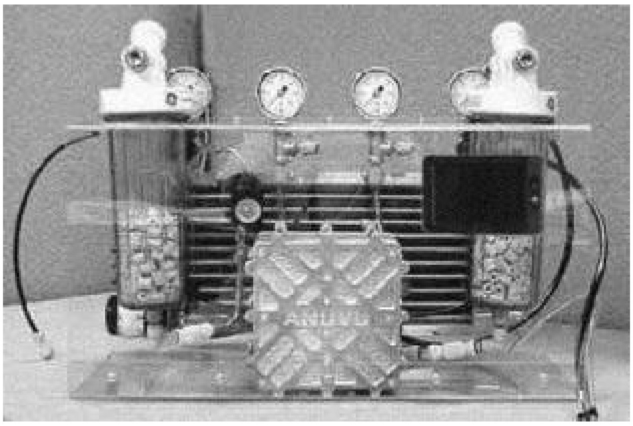

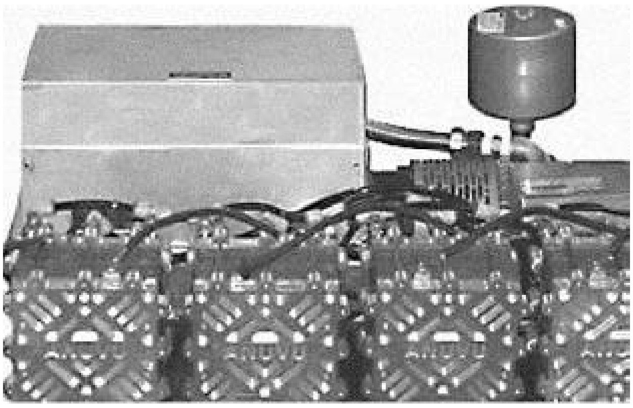

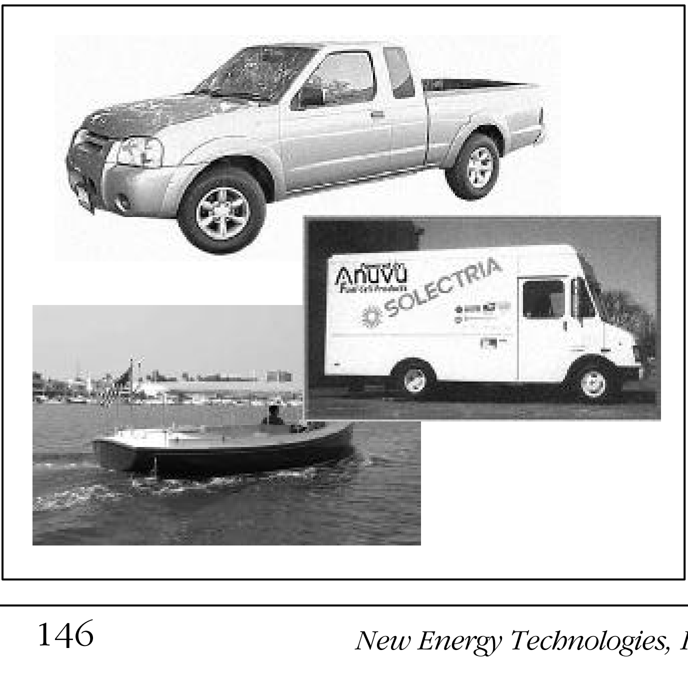

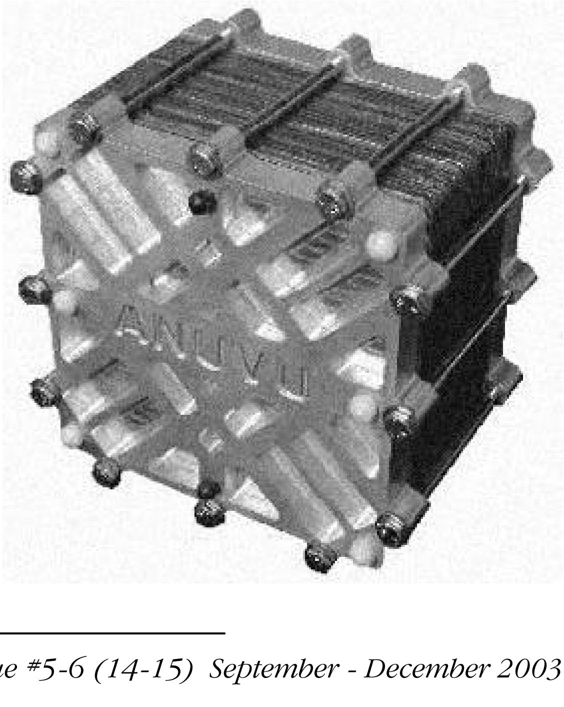

EXCESSIVE OUTPUT BY MEANS OF AIR IONIZATION

We have received an interesting article from California. Here is the short version of the article.

The Mechanism of The Electric Spark

Review of the research
by Leonard B. Loeb
Professor of Physics, University of California at Berkeley
and
John M. Meek, Research Engineer
Stanford University Press, Stanford University, California

Dedicated to Professor J. S. Townsend whose pioneer research and theory laid the whole foundation for the study of the mechanism of the electrical spark discharge.

Although the electric spark has been known to mankind in its various manifestations from time immemorial, its mechanism has to date been little understood. The initial clarification of the mechanisms involved is due to J.S. Townsend as a result of his brilliant researches in the early nineteen hundreds. On the basis of his theory of ionization by collision by electrons and positive ions, the fundamental mechanisms active and especially the coefficients required in their application were made available.

In 1936 the present senior author was forced to describe the mechanism of spark discharge in terms of a modified but distinctly unsatisfactory Townsend theory. In 1935 the discovery of photo-ionization in air by corona discharge indicated a solution was not far off. The turning point of a more successful theory came in the discovery of streamers in positive point to plane corona in 1936. The quantitative analysis of the self-propagating positive streamer in all breakdown phenomena became clearly evident as a result of the data concerning electron avalanches. As a result a qualitative mechanism of sparking by streamer propagation from anode to cathode functioning by means of photo-ionization in the gas was established.

The Townsend Sparking Criteria

It will not be necessary here to derive the famous equation of Townsend for the current [i] in a gap between electrodes as a function of the photoelectric current [io] from the cathode, the gap length [x] and the coefficients [a] and [B]. For this the reader can go to any standard text. *(Editor: the equation is omitted)*

In this equation the first Townsend coefficient [a] represents the number of new electrons created in the gas by an initial electron in its advance of 1 cm along the field axis from the cathode.

The second Townsend coefficient [B] in Townsend's original theory was the number of new electrons created by a single positive ion in its advance of 1 cm along the field from the anode.

The quantity [a] has been extensively studied in various gases. It varies with the ratio of field strength to pressure, X/p, where [X] is in volts per centimeter and [p] is in millimeters of Hg.

Note: the reason we are going through this is to determine the actual increase in current provided by the spark gap, and thus be able to design the circuit to avoid blowing out semiconductor components. It also provides a sound and already proven scientific theory to work from giving us a good foundation and the confidence to proceed with technical design work.

The quantity [B] has been evaluated, albeit rather inaccurately, from the variations of [i] with [x] at various higher values of X/p, by many observers in different gases. Inasmuch as it has now been shown that there are numerous other mechanisms other than impact with positive ions, which can liberate the secondary electron, needed in discharge.

There has been an inclination to give up the mechanism of impact ionizations by positive ions in gas. The discovery of measurable photoelectric ionization in gas has now made it possible to explain such cases. The exact way in which photo-ionization in the gas could operate to cause a spark, was not clear until the development of the present **streamer theory**.

The Streamer Theory of Spark Discharge Anode Space-Charge Field Due to an Avalanche

Assume a spark gap of 1 cm in length. Assume that in air at atmospheric pressure the potential across the plates is 31,600 volts, which is the conventionally observed sparking potential [Vs].

Let us then calculate what happens in the field to one of those electrons. It starts across the gap, quickly acquiring an average random energy of some $E=1/2mC^2= 3.6$ electron volts and a drift velocity [v] in the field direction of about 1.5 to 2 times 10^7 centimeters per second. As it moves it creates new electrons at a rate of [a] per centimeters in the field direction so that in a distance [x] it and its progeny amount to e(ax) electrons, forming what is called an **electron avalanche**.

Therefore, e(ax) positive ions have been left behind by the electron group, virtually where they were formed in the 10^{-7} second of advance for the electrons in the distance x=q across the plates. As the electron avalanche advances, its tip is spreading laterally by the random diffusive movement of the electrons.From these data it is possible to compute the density of positive-ion space charge left behind at any point [x]. The value of [a] under these conditions is about 17, making e(aq)=e(17). The first ion pair is created at 0.0407 cm from the cathode. At 0.5 cm from the cathode there are 4914 ions, at 0.75 cm there are 3.66 times 10^5 ions, and within 0.0407 cm from the anode there are 1.2 times 10^7 ions. **Most electrons will be drawn to the anode except for some few that are bound by the positive ions, making a sort of a conducting discharge plasma in the avalanche.**

Such a distribution of ions does not make a conducting filament of charges across the gap, and hence in itself an avalanche that has crossed does not constitute a breakdown of the gap. Thus one must look further for the mechanism of the spark.

If Loeb and Meek are correct then if we assume a spark gap of 3 mm and a voltage of 5,000 voltsthere are roughly 2,000 electrons created by avalanche for every one electron leaving the cathode. They state that most of these 'free electrons' are absorbed by the anode. [This would certainly explain why the semiconductor components cannot handle the current gain.]

NOTE: Loeb and Meek make little reference to initial amperage. There are only two values they refer to 10^{-5} ampere and 10^{-12} ampere.

In conclusion: Sparks and Arcs are two different beasts. My initial research into the amperage necessary to form an arc does not apply to spark and the process of avalanche where this huge gain mechanism is possible.

Photoelectric Ionization in Gas as a Secondary Mechanism

Accompanying the cumulative ionization there is produced by electrons from four to ten times as many excited atoms and molecules. Some are excited to an energy exceeding the ionizing potential of some of the atoms and molecules present, either by excitation of an inner shell, by ionization and excitation, or in a **mixed gas like air by the excitation of molecules of higher ionizating potential, e.g., N2**. These excited atoms or molecules emit radiations of very short wave length in some 10^{-8} second. This short ultraviolet radiation is **highly absorbed** in the gas and leads to ionization of the gas. In fact, the whole gas and the cathode as well are subjected to a shower of photons of all energies traveling from the region of dense ionization with the velocity of light. Thus nearly instantaneously in the whole gap and from the cathode new photoelectrons are liberated which almost at once begin to ionize cumulatively.

The Mechanism of Positive Streamer Formation

The photoelectrons created at points in the gas and at the cathode at any great radial distance from the avalanche axis will merely create other avalanches. Those in the gas will be short and those coming from the cathode region will be long and like that of the initial avalanche. Being smaller and, in any case, later in creation than the parent avalanche, such avalanches will be of no interest in breakdown. However, those photoelectrons created near the space-charge channel of positive ions, and especially near the anode, will be in an enhanced field, which exerts a directive action drawing them into itself. If the space-charge field [X1] is in the order of magnitude of the imposed field [X], this action will be very effective. In addition the values of [a] will be much enhanced.

The electrons from the intense cumulative ionization of such photoelectron avalanches in the combined fields [X] and [X1] which are drawn into the positive space charge feed into it, making it a conducting PLASMA which starts at the anode. The added fields will be most effective along [X] and so will the ionization. The positive ions they leave behind will therefore extend the space charge towards the cathode. These electrons also create photons, which produce electrons to continue this process. **In this fashion the positive space charge develops toward the cathode from the anode as a self-propagating positive space-charge streamer.**

As the streamer advances towards the cathode it produces a filamentary region of intense space-charge distortion along a line parallel to the field. The conducting streamer of a plasma consisting of electrons and ions extending to the anode thus makes a very steep gradient at the cathode end of the streamer tip. As this advances toward the cathode the photoelectron avalanches produced by radiation at the cathode, especially at the intercept of the extended streamer axis at the cathode, it begins to produce an intense ionization near the cathode. Hence the positive ions created there may increase the secondary emission. Thus, as the space-charge streamer approaches the cathode a cathode spot is forming which may become a source of visible light.

When the streamer reaches the cathode there is a conducting filament bridging the gap. As the streamer tip reaches the cathode the high field produces a rush of electrons towards the end of the streamer. This if followed by a current of electrons, gives a high-potential wave, which passes up the preionized conducting channel to the anode, **multiplying the electrons present by a large factor**. The channel is thus rendered highly conducting. If the metal can emit a copious supply of electrons because of the formation of an efficient cathode spot, the current of electrons continues the channel maintaining its high conductivity and **ever increasing in it**. This current, unless limited by external resistance, will then develop into **an arc**. It is, however, the intense increase in ionization by the potential wave, which gives the highly conducting channel characterizing the spark.

Conclusion: According to Loeb and Meek there are three means by which a spark in open air will provide a very large current gain. If this is true, it should be fairly easy to prove with inexpensive and unsophisticated equipment. Once the actual amount of current gain has been determined for the design parameters of the spark gap, then the rest of the circuit can be designed for the increased current value.

Editor: It is worth of a note that the current gain by means of ionization was patented by Pavel N. Yablotchkov (the patent of France # 120684, October 11th 1887). Some two years ago one of our issues featured an article about him. We think it is worth being published again.

PATENT of 1877 by Pavel N. Yablotchkov

Alexander V. Frolov
Faraday Lab Ltd,
Lev Tolstoy St., 7 St. Petersburg, 197376 Russia
Tel: 7-812-380-3844

Pavel N. Yablotchkov was born in 1847 near Saratov, Russia. He graduated as a Military Engineer in 1866 and spent several years in the Russian Army.

In 1872 he came to Moscow and started his activities in electrotechnical field. Since 1875 he had been working in Paris with the famous Louis Breget and his first French patent # 110479 of November 29th, 1875 was dedicated to an electromagnetic transformer. Then he developed and patented a lighting system (the well-known Yablotchkov electrical candle). In 1876 he patented a new electromagnetic transformer for industrial purposes, France # 115793 of November 30th, 1876.

The most interesting patent claim on over-unity devices by Pavel N. Yablotchkov is known as France patent # 120684, October 11th, 1877, "The system of distribution and amplification of electrical currents by means of atmosphere electricity..." The patent describes special capacitors connected in series with the load to increase the output current by means of ionization. Experiments were conducted together with the well-known physicists such as Dr. Maskar, Dr. Varren-Delaru and others and **they confirmed the 200 % efficiency of the circuit**. Now we will try to explain the method.

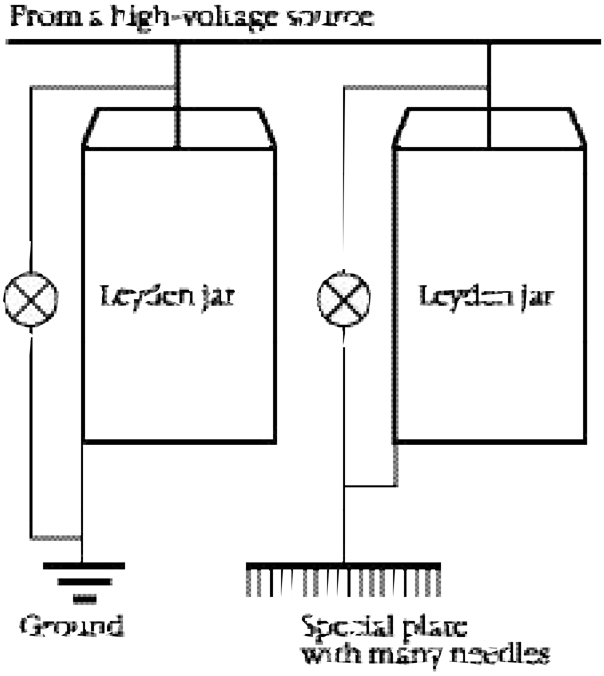

Fig.1 features a schematic drawing taken from Yablotchkov's patent. The Leyden jar is an asymmetrical capacitor, i.e. it is different in principle from a two-plate flat capacitor. The inner electrode of the jar should be connected to a high-voltage source and in this case the changes of the potential have an effect on the potential changes on the external electrode. It does not work in the opposite case and if you connect a high-voltage source to the external electrode no potential changes will be detected on the inner electrode. Connection to the ground or to a special plate (which is covered with many needles to increase air ionization) is necessary to attract the maximum number of electrons to the plate surface or to return the maximum electrons from the plate surface when potential changes on the external electrode are produced by means of electrical induction in the Leyden jar.

In conclusion I should mention one more supposition of the secrets of the well-known Swiss M-L converter (Methernitha). The main elements of the design are Leyden jar capacitors, which have the external surface made of perforated metal.

The other known fact is that great ionization of air is observed when the converter is in operation. So, the electrostatic machine can produce pulses of a very high voltage (potential difference) but it cannot be used as a source of a powerful current. In order to increase the current in the circuit we should apply a certain method and Yablotchkov's technology seems to be appropriate. A large surface of the external electrode of the Leyden jar can be a good solution to the problem. Maximum strong ionization allows us to obtain the output current several times stronger than the weak current generated by the electrostatic machine.

SUBSCRIPTION 2004!

To subscribe for
New Energy Technologies
please visit our web site:

http://www.faraday.ru/net.htm

One-year subscription
price for 2004 is $46.00

World-wide airmail delivery
included.

In 2004 we will publish
5,000 copies, 4 times per year.
Each printed issue will include
100 pages of printed information.
Alongside the English version
there is the Russian version
available.

CD "New Energy Technologies"

Back issues of 2001 - 2003 are
available on a CD with PDF files.
The price is $26.00.
You can order a CD from the same
web page:
http://www.faraday.ru/net.htm

Contact information for additional requests:

Faraday Lab Ltd,
New Energy Technologies magazine,
Lev Tolstoy Str. 7 -601,
St.Petersburg,
197376, Russia.
Tel/fax 7-812-3803844

TeslaTech, Inc.

296 E Donna Dr, Queen Valley, AZ 85218

Visit our site!.... **http://www.teslatech.info**

Benefits

- 1 yr *ExtraOrdinary Technology* (qtly)
- Mail Order Discounts
- Conference Fee Discounts

ExtraOrdinary Technology... our magazine containing news about ***the leading edge of our Scientific Revolution***.

...an invaluable resource to track leading edge research!!!

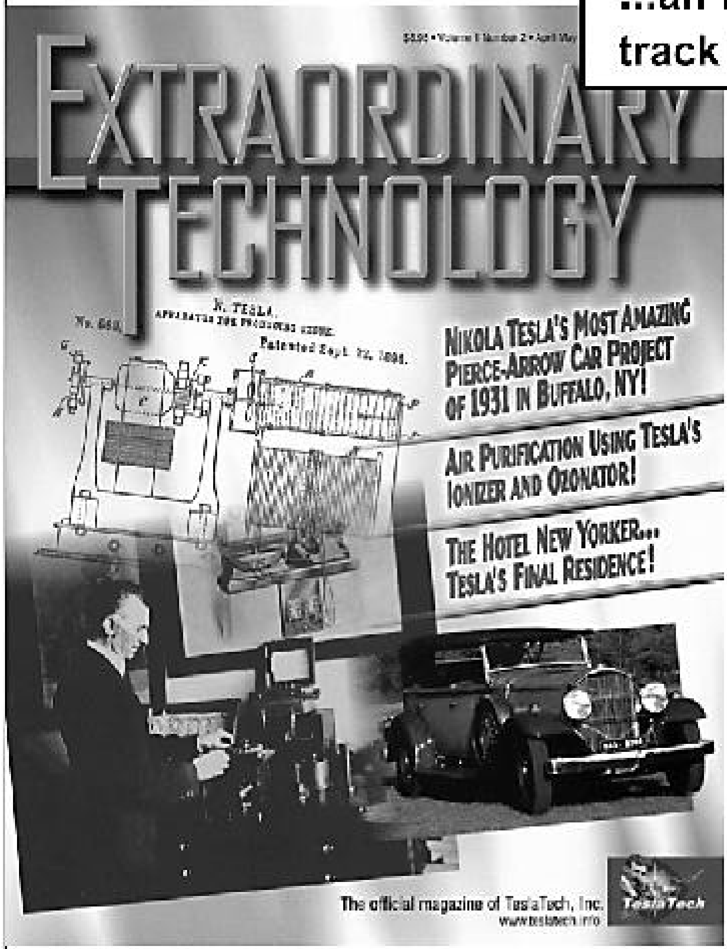

EXTRAORDINARY TECHNOLOGY

NIKOLA TESLA'S MOST AMAZING PIERCE-ARROW CAR PROJECT OF 1931 IN BUFFALO, NY!

AIR PURIFICATION USING TESLA'S IONIZER AND OZONATOR!

THE HOTEL NEW YORKER... TESLA'S FINAL RESIDENCE!

The official magazine of TeslaTech, Inc.
www.teslatech.info

ExtraOrdinary Technology
focuses on:

Tesla Technology
Low temperature plasma
Electrogravitation
Electrified water
Magnetic healing
Electromedicine
New Health Supplements
...and much, much more!

JOIN TODAY!
(520) 463-1994

Annual Membership Rates: $30 - United States • $35-Can/Mex · $55-Else

Advanced Energy • Alternative Medicine • Future Technology

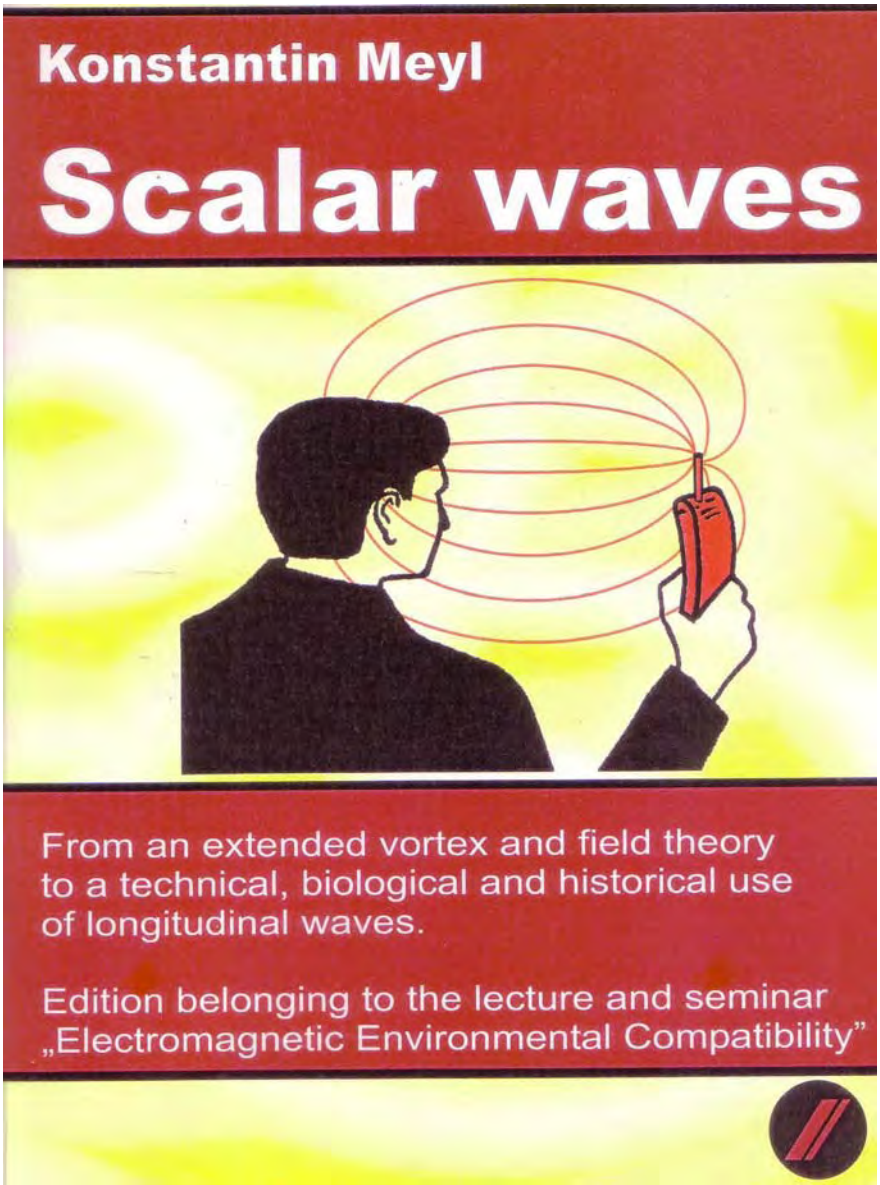

Konstantin Meyl

Scalar waves

From an extended vortex and field theory
to a technical, biological and historical use
of longitudinal waves.

Edition belonging to the lecture and seminar
„Electromagnetic Environmental Compatibility"

Preface to the lecture, 1st Edition 1996

The theme encloses the electromagnetic compatibility of both technical and biological systems. Only part of the electromagnetic wave can be considered for function troubles, namely the part that was absorbed and has rolled up to a vortex. The activity depends on the number of created vortices and of their lifetime, their decay.
The eddy current only manifests in conducting materials. In the air and in dielectric materials on the other hand the vortex of the electric field will form, also called the potential vortex. To calculate and to measure this vortex is our goal.

First we'll carry out a survey of the problems and the usual methods. From the analysis of unsolved problems the need for the introduction of the new vortex phenomena is deducted and an adequate field-theoretical approach will be chosen. Afterwards the potential vortices are calculated and their properties are discussed and interpreted. For the purpose of proving their existence, on the one hand the Schrodinger equation will be derived and on the other hand the quantum properties of the most important elementary particles will be calculated and compared with the well-known measured values. Measurement and calculation are in excellent agreement for weight, charge, magnetic moment and spin. So the theory not only proofs it's correctness, in addition it demonstrates it can achieve much more. The theory takes us to the unification of the well-known interactions and physical phenomena and shows itself as an unified theory. In the practical conversion and usage of the theory there will not only be informed but by all means also be provoked as an entrance in a fruitfully discussion. Fundamental questions will be taken up like: What is information, energy, temperature or smell? The connection to the theme of the electromagnetic environmental compatibility is formed by the technical and the biological usage of the potential vortices, the energy transmission of Nikola Tesla exactly like the in a similar way functioning nerve conduction. Here we already can expect biological reactions.

This lecture, held for the first time in the winter semester of 1995/96, is available in book form, as an edition belonging to the lecture. This lecture will not deliver ready recipes or instructions. The goal is reached when the critical sense of the listeners and readers has been inspired and discussions have been set going. Everybody has to draw the consequences out of such a theory by him- or herself.
In addition to this lecture a seminar is offered, wherein several themes are supplemented or deepened, different theories are compared and possible consequences are discussed. The appearance of an edition belonging to the seminar has started in 1998<i>. Regarding the conversion of consequences both politicians and scientists are equally addressed, because the electromagnetic environmental compatibility has developed to one of the most urgent problems of today's world. But in last consequence all of us bury the worldwide responsibility for our environment.

<i>: K. Meyl: Electromagnetic environmental compatibility, Part 2 and 3 of this book, Edition belonging to the seminar.

Table of contents
Page

Preface to the Lecture. 1st Edition 1996 III

Table of contents IV

Part 1 edition belonging to lecture "Electromagnetic environmental compatibility"

1. Introduction 1
1.1 Discussion about limits 3
1.2 Wireless telephones 5
1.3 Absorption of waves 7
1.4 Overlap effect 9

2. Tasks 11
2.1 Tasks of the electromagnetic environmental compatibility 11
2.2 Tasks of the electromagnetic compatibility (EMC) 13
2.3 Declaration of conformity 15
2.4 EMC-techniques to measure the emission of interference 17
2.5 Electro smog 19
2.6 EMC-techniques to measure the stability against interference 21
2.7 Tasks of the biological compatibility 23
2.8 Natural fields 25
2.9 Electricity of the air 27
2.10 Biological effects 29
2.11 Artificial fields 33
2.12 Protection against artificial fields 35
2.13 Unsolved tasks 37

3. Approach 39
3.1 Principle of causality 41
3.2 Field-theoretical approach 43
3.3 Duality 45
3.4 Flow vortices 47
3.5 Rankine vortex 49
3.6 Vortices and anti-vortices 51

4. Properties 53
4.1 Concentration effect 53
4.2 Duality of the vortex properties 55
4.3 Derivation of the electron as an elementary vortex 57
4.4 Quanta as field vortices 59
4.5 The photon 61
4.6 Pair creation 63
4.7 Noise 65
4.8 Condenser losses 67
4.9 Vortex lines and vortex streets 69

Page
4.10 Water colloids as vortex balls 71
4.11 Transport phenomenon 73

5. Derivation and interpretation 75
5.1 Fundamental field equation 77
5.2 Mathematical interpretation 79
5.3 Physical interpretation 81
5.4 Phenomenological interpretation 83
5.5 Atomistic interpretation 85
5.6 Derivation of the Klein-Gordon equation 87
5.7 Derivation of the time dependent Schrodinger equation 89
5.8 Derivation of the time independent Schrodinger equation 91
5.9 Interpretation of the Schrodinger equation 93

6. Theory of objectivity 95
6.1 Proof 95
6.2 Law of conservation of energy 97
6.3 Radius of the electron 99
6.4 Maxwell's field equations 101
6.5 Transformation equations 103
6.6 Field overlap 105
6.7 Field dependent curvature of space 107
6.8 Electromagnetic interaction 109
6.9 Gravitation 111
6.10 Field dependent speed of light 113
6.11 Universality 115
6.12 Aether 117
6.13 Spin and tunnel effect 119
6.14 Interpretation of the measured faster than light speed 121
6.15 Definition of the speed of light 123
6.16 Relativity and objectivity 125
6.17 Transformation 127
6.18 Transformation table 129
6.19 Interpretation of the transformation table 131
6.20 Particle decay 133

7. Proof 135
7.1 Elementary vortices 137
7.2 Matter and anti-matter 139
7.3 Positronium 141
7.4 Dipole moment 143
7.5 Myon 145
7.6 Calculation of the vortex fields 147
7.7 Calculation of the proton 149
7.8 Strong interaction 151
7.9 Magnetic moment of the proton 151

Page

7.10 Structure of the neutron 153
7.11 Calculation of the neutron 155
7.12 beta-decay 157
7.13 "Weak interaction" 159
7.14 Tau particle 161
7.15 Pions 163
7.16 Table of vortices of the Mesons 165
7.17 Table of vortices of the Baryons 167

8. Unified theory 169
8.1 Structure ofthe field theory 171
8.2 Unification of the interactions 173
8.3 Temperature 175
8.4 Heat 177
8.5 Sound 179
8.6 Basic principle of cybernetics 181
8.7 Structure of adaptive regulatory circuits 183
8.8 Information 185
8.9 Philosophy of nature 187

9. Usage 189
9.1 Longitudinal electromagnetic waves 191
9.2 Medical technical usage 193
9.3 Flying objects 195
9.4 Electro gravitation? 197
9.5 Free energy? 199
9.6 Nerve conduction and action potential 201
9.7 Wireless energy transmission technology 203
9.8 Measuring and switching technique of Nikola Tesla 205
9.9 Energy technical usage 207
9.10 Environmental compatibility 209
9.11 Ecologically fairusage 211
9.12 Claims concerning the environmental compatibility 213
9.13 Epilogue belonging to the lecture (part 1) 215
Abstract belonging to scalar waves, part 1 218

Page

Part 2, edition belonging to the energy technical seminar 219

10.0 Preface to the seminar, part 2 220
10. Oscillating interaction 221
10.1 Kepler's laws 221
10.2 Unknown interaction 223
10.3 Harmony of the alternating current engineers 225
10.4 Four fundamental interactions 227
10.5 Resonating interaction 229

11. The growing globe 231
11.1 Long-distance effect of the neutrinos 231
11.2 Model of expansion 233
11.3 Ancient Egyptian legend of Genesis 235
11.4 Inner structure of the earth 237
11.5 Earth's core as a converter of neutrinos 239
11.6 Speed of growth 241
11.7 Conservation of angular momentum 243
11.8 Set of difficulties concerning the change of magnetization 245
11.9 The weakly damped moon...247
11.10 Calculation of celestial mechanics 249
11.11 The biblical age 251

12. New cosmology 253
12.1 Concerning the formation of our solar system 253
12.2 The birth of the planets 255
12.3 The derivation of the Titius-Bode law...257
12.4 The hollow planet 259
12.5 Concerning the formation of the universe 261
12.6 Counter examples concerning the 2nd law of thermodynamics 263
12.7 Entropy destroying potential vortices 265

13. Recording of space and time 267
13.1 The measuring technical debacle 269
13.2 The clock paradox 271
13.3 The Tamarack mines experiment 273
13.4 Field dependent linear measure 275
13.5 Experiences from space travel 277
13.6 Spherical aberration 279
13.7 Step by step understanding of the measuring technique....................................281
13.8 Discussion of the cosmological insights..283

14. The way towards free energy 285
14.1 The liberalization of the energy markets 285

Page

14.2 The consequences of the liberalization 287
14.3 The chances of the wind energy 289
14.4 The chances of being self-sufficient concerning energy 289
14.5 The space energy technology of nature 291
14.6 The gap in the energy balance of man............293
14.7 Carl Freiherr von Reichenbach 293
14.8 Cold fusion and Genesis 295
14.9 Photosynthesis 297
14.10 How nature materializes 299
14.11 Lightning 301
14.12 Ball-lightning............303
14.13 Discussion concerning the neutrino conversion............305

15. Principle of funtioning of space energy 307
15.1 The course of the field lines 307
15.2 Gravitation............309
15.3 Systematizing the interactions 311
15.4 Magnetic force converter............313
15.5 The railgun 315
15.6 Unipolar induction............317
15.7 Tendency to instability 319
15.8 Unipolar field configuration 321
15.9 The Testatika 323
15.10 The secret of the Testatika............325
15.11 The key to free energy............327

16. Space energy technology (SET)............329
16.1 The unipolar generator 329
16.2 Derivation of Maxwell's field equations............331
16.3 SET-devices with rotating magnetic fields............333
16.4 Commentary concerning magnetic SET-devices............335
16.5 RQM and the space quanta manipulator............337
16.6 SET-devices with pulsed magnetic fields............339
16.7 SET-devices with Mobius winding 341
16.8 Mobius converter of Seike and Coler 343
16.9 Tesla's flat coil 345
16.10 The secret of the flat coil 347
16.11 Discussion concerning the technology of the neutrino collectors 349

17. Technical use of the weak interaction __ 351
17.1 Radioactivity caused by neutrinos............351
17.2 Nikola Tesla, the discoverer of the neutrino radiation 353
17.3 Transmutation and reduction of radioactivity 355
17.4 The Patterson Power Cell 357
17.5 Report concerning the cold fusion............359
17.6 Water-Fuel-Cell technology 361
17.7 Unconventional electrolysis 363

Page
17.8 Materialization of neutrinos 365
17.9 Oscillation of size and luminescence 367
17.10 Phosphorescence and sonoluminescence 369
17.11 Discussion concerning the use of the weak interaction 371

18. Physical phenomena 373
18.1 Water as a catalytic converter for neutrinos 373
18.2 Background and zero point radiation 375
18.3 The Casimir effect 377
18.4 Effects under high tension 379
18.5 Flying discs 381
18.6 Propulsion with neutrinos 383
18.7 Antigravitation or levitation 385
18.8 Discussion concerning the effect of closed field lines 387
18.9 Concerning the determination of the stand of today's energy technique 389

19. Own experiments 391
19.1 Concerning the rating of experimental physics 391
19.2 The charge amplifier from the workshop of handicrafts 393
19.3 Interpretation of the observations with the charge amplifier 395
19.4 High-tension experiments in the laboratory for performance electronics 397
19.5 Measurement of dielectric capacitor losses in the HF laboratory 399
19.6 Analysis of the measured capacitor losses 401
19.7 Microwave oven for testing in the laboratory of electrical engineering 401
19.8 Microwave heating by decay of vortices 403
19.9 The Tesla coil from the workshop of handicrafts 405
19.10 Biological effectiveness of the Tesla coil 407
19.11 Wrestling for measuring technical insight in the Transfer Centre 409
19.12 Neutrinolysis, the alternative splitting of water 411

20. Cosmic experiments 413
20.1 Millions of voluntary test subjects at 11 August 1999 413
20.2 Model of calculation for the cosmic experiment 415
20.3 Physical technical consequences 417
20.4 Biological consequences 419
20.5 Epilogue to 11 August 1999 421
20.6 Dropping of the neutrino radiation 423
20.7 Analysis and comparison of the occurrences 425
20.8 Radioactive decontamination with the help of a supernova 427
20.9 Free energy from the Orion constellation 429
20.10 Interaction of the neutrinos with the core of the earth 431
20.11 Changing of polarity and apocalypse 433
20.12 Scalar wave gauges for the prediction of earthquakes 435
20.13 Calculation of future occurrences 437
20.14 Epilogue belonging to the energy technical seminar (part 2) 439
Abstract belonging to scalar waves, part 2 442

Page

Part 3, edition belonging to the information technical seminar 443

21.0 Prefaces to the seminar, part 3 444
21. Derivation of the scalar wave 447
21.1 Lord Kelvin in a delicate mission 449
21.2 Helmholtzian ring-like vortices in the aether 451
21.3 Taking apart of the wave equation 453
21.4 Derivation of the scalar wave equation 455
21.5 Derivation of a plasma wave 457
21.6 Mistakes of the HF technology 459
21.7 Coupling of the wave parts 461
21.8 Set of difficulties of the near-field 463
21.9 Transition to the far-field 465
21.10 Scalar wave model 467
21.11 Double-frequent oscillation of size 469
21.12 Electric and magnetic scalar wave 471

22. Properties of scalar waves 473
22.1 Wave-particle compromise 473
22.2 Concerning the measurement of light 475
22.3 Comparison of the parts of Tesla and Hertz 477
22.4 Noise, a scalar wave phenomenon 479
22.5 Neutrino radiation 481
22.6 Parallel instead of serial image transmission 483
22.7 Comparison of the properties 485

23. Research of scalar waves 487
23.1 Frequency diagram 487
23.2 The reverse of the HF medal 489
23.3 Radiesthesia 491
23.4 Prediction of earthquakes with the help of scalar waves 493
23.5 Searching water with the help of scalar waves 495
23.6 Receiving scalar waves with the Lecher antenna 497
23.7 Assignment according to wavelength 499

24. Medicine of oscillation 501
24.1 Mediation of information with the help of scalar waves 501
24.2 Systems with positive feedback 503
24.2.1 The Synchrometer 503
24.2.2 The sound therapy 503
24.2.3 The colour therapy 505
24.2.4 The aromatherapy 505
24.2.5 The gem therapy 505
24.2.6 Meditation and self-therapy 505

Page
24.3 Systems with negative feedback 507
24.3.1 The homeopathy 507
24.3.2 The bioresonance 507
24.3.3 The frequency therapy 509
24.4 The Zapper 511
24.5 Discussion concerning the medicine of oscillation 513

25. Contributions to the discussion concerning the information technical use 515
25.1 Window of the sensuous perception 515
25.2 The sense organs (nose, eye, tongue) 517
25.3 The nerve conduction 519
25.4 The brain, a scalar wave computer 521
25.5 Concerning signal engineering 523
25.6 Stability against interference and repair mechanisms 525
25.7 The microwave window 527
25.8 Discussion concerning scalar wave medicine 529

26. Recapitulation from the viewpoint of textbook physics 531
26.1 Common misinterpretation of the antenna losses 531
26.2 Wave damping by field vortices 533
26.3 Laplace versus Maxwell 535
26.4 Error term in the wave equation 537
26.5 Interim result 539
26.6 Failure of the Maxwell theory 541
26.7 Concerning the evidence situation 543

27. Faraday versus Maxwell 545
27.1 Energy out of the field 545
27.2 Vortex and anti-vortex 547
27.3 Vortices in the microcosm and macrocosm 549
27.4 Faraday's law and Maxwell's formulation 551
27.5 The discovery of Faraday 553
27.6 Different formulation of the law of induction 555
27.7 Contradictory opinions in textbooks 557
27.8 The field-theoretical approach 559
27.9 Derivation of Maxwell's field equations 559
27.10 Derivation of the potential vortices 563
27.11 Derivation of the fundamental field equation" 565
27.12 The Maxwell field as a derived special case 567
27.13 Derivation of the wave equation 569
27.14 The new field approach in synopsis 571

28. Objectivity versus relativity 573
28.1 The question concerning the aether 573
28.2 Vortices, an overlap of the overlap 575
28.3 Field overlap 577

Page
28.4 The derivation of the length contraction 579
28.5 The dependence of the Lorentz contraction on the field 581
28.6 Boscovich and the respiration of the Earth 583
28.7 Derivation of the gravitation 585
28.8 From subjectivity to objectivity 587
28.9 The objective standpoint 589

29 Mathematical gleanings 591
29.1 General and special theory of relativity 591
29.2 Transformation table 593
29.3 Discussion concerning the root of Lorentz 595
29.4 Mathematical derivation of the neutrino 597
29.5 Universal constants and constants of nature 599
29.6 The fundamental number e 601
29.7 The pentacle 603
29.8 The vortex, which is rolling up 605
29.9 The Golden Proportion 607

30. Scalar wave technology in antiquity 609
30.1 Thesis 609
30.2 Prerequisites 611
30.3 Approach 613
30.4 Circumstantial evidence 615
30.5 Radio technical concepts 617
30.6 Wireless telegraphy 619
30.7 AM temple technology 621
30.8 Modulation of acoustic signals 623
30.9 Broadband FM broadcasting technology 625

30.10 Epilogue 627

Contacting address and information 634

Table of formula symbols 635

References 636
A: Publications concerning the theory of objectivity (primary literature) 637
B: Publications about the theory of objectivity (secondary literature) 641
C: Literature concerning physics in general 643
D: Literature concerning geophysics, geology, geography and astronomy 644
E: Literature concerning the EMC and biological compatibility 644
F: Literature concerning Tesla and criticism of textbook physics 644
G: Literature concerning free energy 645
H: Literature concerning medicine, biology and chemistry 646
J: Literature concerning history and more peripheral subjects 646

Appendix 647

1. Introduction

Here the extremely controversially discussed question of the environmental compatibility of electromagnetic fields will be persuaded. Limits should inform what is incompatible and what is compatible. But there are as many limits as there are commissions and specialists. And besides that differ the results from each other for several powers of ten. In course of time the legitimate doubts become unmistakable and the representatives of science slowly get to feel the burden of proof.

For the sake of efficiency, the actual discussion concerning the theme of electro-smog is analysed and the necessity to involve an until now unnoticed field phenomenon in the discussion about limits is derived: It concerns vortices of the electric field. These potential vortices, as they are called, have the corresponding properties to show biological effects even at the lowest field strengths. In any case it is not possible to exclude that at present the wrong physical phenomena are measured and made responsible.

A parable should bring clarity.

Lets imagine that the to us well-known and over our sense of touch understandable physical phenomenon of the temperature is unknown to us, neither measurable nor perceptible. Our weather station only exists of a barometer that could show us the air pressure and deliver us indications if good or bad weather is to be feared. We ready realize that there exists a connection between the air pressure and our health and make the to us well-known phenomenon responsible. When the pointer points to good weather we can go out lightly dressed. With bad weather we should take a coat, so we know from experience.

"Now we imagine the realistic situation that in winter we have a weather situation of high pressure but it's stone-cold outside. The weather station will display high temperatures with the result that some people will walk around with short-sleeved and open shirt, only to lie in bed with a cold in the evening. Of course the air pressure was to blame! Logically the "pressure sensitive", as they are called mocking, demand the limits for the allowed pressure to be reduced so far that no consequences for health are to be feared. Concerning the theme of allowed limits, science is asked and science proceeds in a systematic way: the pressure is investigated in the laboratory, isolated from all other parameters and so it is discovered that man catches no cold even at a substantially higher air pressure, so there is no reason to alter the limits.

Actually we would expect these at any time reproducible results to have a calming effect on the minds of the participants of the discussion and on the population. Instead the pressure sensitives time and again cite new knowledge that won't fit in the scheme. So is for instance stated that draught causes the same health problems although this pseudo effect has nothing at all to do with the air pressure. So owing to incomprehensibility and emotions the discussion about limits becomes a farce.

The fact that sensitive people react to effects of air electricity and possibly get ill without proof that some today measurable physical quantities are responsible should make us think. It is little calming watching our scientists poking at the dense fog whereas at the same time among the runners of the new telecommunication networks there spreads something like a gold-digger mood.

To introduce a new technology is not difficult, but to abolish it for reasons of the electromagnetic environmental compatibility is almost impossible!

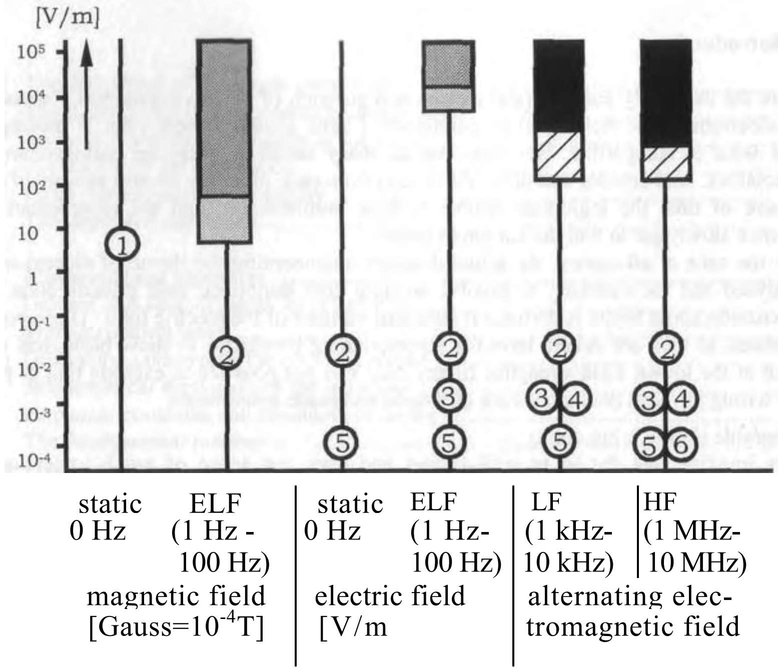

Fig. 1.1: Discussion about limits

■ Limits for professional exposition (electronics engineers) and
▨ Limits for non-professional exposition (population in general) according to the Recommendation of IRPA/INIRC<ii>.
▒ Limits according to VDE 0848 / 1989
▓ Limits according to VDE 0848 / 1992
⊛ experimentally determined threshold values of reactions of biological systems <i>:

1 Increase of the activity of movement of birds
2 deflection of divining-rods
3 influence on the time of reaction of men
4 conditional reflexes of fish without electrical organs
5 conditional reflexes of fish with electrical organs
6 conditional muscular reflexes of men

taken from:
<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Bild 111, S. 123 Verlag Moos & Partner Munchen, ISBN 3-89164-O58-7
<ii>: Habiger u.a., EMV, Verlag Technik, Berlin 1992, S. 152

1.1 Discussion about limits

Whoever follows the public discussions concerning electro smog always sees two arguing parties, whose standpoints lie so far apart that they inevitably talk at cross purposes and there can be found no consensus.

On one side the "affected" find together who maintain to have found the electromagnetic radiation as the damaging cause for their problems. They are to be taken serious, even when only their personal sensitivity serves as means of measurement and proof and a more or less distinct sensitivity against electromagnetic phenomena. This group occasionally finds support of homeopaths who can base on reproducible laboratory results that fit as few into the view of life of science as the empirical statements of affected and possibly hurt people.

On the other side stand the representatives of the energy-supply companies and the runners of radio networks who argue with the needs of our modern industrial society and give "limits" prescribed to them by scientists. These, for their part, proceed according to strictly scientific methods. Their presented results are reproducible and there's no doubt about them.

The limits after all are fixed far below those that are recommended from a scientific viewpoint. Nevertheless both groups are separated from consensus by powers of ten. When we want to know how deep the ditch is we want to bridge, we should take a look at the determined limits (Fig. 1.1).

The limits stem from the 1RPA (International Radiation Protection Association) an organ of the World Health Organization that in turn has appointed the INIRC (International Non Ionizing Radiation Committee). These now state to have used all available scientific research results as basis for the given guidelines.

Moreover a safety range was worked into them. So the limits were fixed at substantially lower levels to guarantee that no health damage arises. In this way first the limits were determined for the people who for reasons of profession are exposed to electromagnetic fields.

For the population in general the limits for the so called non-professional exposition were reduced further to one half till one fifth for reasons of caution and care. In Fig. 1.1 these limits are registered. Thereby is distinguished between magnetic fields and electric fields that appear stationary or at extremely low frequencies (ELF describes frequencies between 1 Hz and 100 Hz). Moreover limits for low-frequency (1-10 kHz) and high-frequency (1-10 MHz) alternating electromagnetic fields are given.

The graph should serve as a rough orientation and show us the proportion of scale. As further information some thresholds of measured reactions of biological systems are registered (after Konig*[1]*). Because a logarithmic scale was chosen to fit all the values on one graph it becomes clear that between the first reactions and the recommended limits there lie up to five powers often. The ditch seems to be insurmountable.

Fig. 1.2: Set of problems of environmental compatibility by means of the example of the handheld wireless telephones (handy).

<i>: L.v.Klitzing: Neurophysiologische Einflusse durch elektromagnetische Felder wahrend und nach der Exposition, Med. Universitat zu Lubeck

1.2 Wireless telephones

Measuring technical surveys with regard to the influence of brain currents by digital radio signals by the university hospital in Lubeck have startled handy manufacturers and users equally <i>. Although in this case measurement errors could be detected, the „bugaboo on the wall" remains that we are sitting unsuspecting in a restaurant and a neighbour draws his handy out of his pocket to make a digital telephone call. Thereby synchronizing the brain currents within a radius of 100 meters on the broadcasting signal and occupying our brain useless with technical signals. The derivation will show that from the start this can't happen to all visitors, because as a prerequisite conditions of resonance must be fulfilled. But would there be an affected, he or she for sure would have considerable problems, because informations that are not picked up over the sense organs can neither be classified timely nor as regards content.

An affected whose brain has picked up technical signals not even is able to register by itself that it was fed with incorrect informations. It would be reasonable when the visitors of the restaurant would defend themselves and put the radio operator on the doorstep. The number of restaurants where apart from cats and dogs also handy's have to stay outside is increasing. How should we keep out of the way of electromagnetic fields? Should we walk around permanently with a steel helmet or even better in a knight's armour and even go to bed with them? It would be worse than in the dark middle ages.

Summarizing: it should be guaranteed that the operation of electro technical apparatus causes neither health damage nor unintentional influence or irritation. A systematic and scientific procedure should investigate in the laboratory all relevant physical phenomena individually for their interaction. Electro physics bases on two phenomena in connexion with electro-smog: on the one hand the radiation and on the other hand the thermal effect, but at a close look both factors prove to be of only little importance!

In radiation measurements the intensity of the electromagnetic wave at a certain place is determined. In laboratory experiments the field strength is increased so long till biological reactions are observed. Thermal limits are determined in a similar way. As said, the values lie about powers of ten above those that possibly bother you when you hold a handy to your ear. It is true that the microwave radiation penetrates into your head but we also know that it marches out again on the other side and this visit in your head happens with the speed of light.

Exactly like this are guest in your body constantly your local radio station, your local television station the satellites with hundreds of programs and anyway the whole radio technical world even when you did not invite them.

For an electromagnetic wave to become receivable, the field strength must lie clearly above the common noise signal and this can only be achieved by a permanent overlap, by standing waves, like in a cavity tuned to a specific frequency or an antenna. As long as people don't let themselves grow antennas on their heads they hardly have to fear direct biological effects of electromagnetic waves.

That leaves as the second phenomenon the thermal effect. With a handy held to your cheek there comes into being a local fever in your head. But that is not at all unusual or unnatural for the human body. Something like that happens to a far greater degree when you take a hot foot bath or let yourself be irradiated at one side from the sun at a tourist grill.

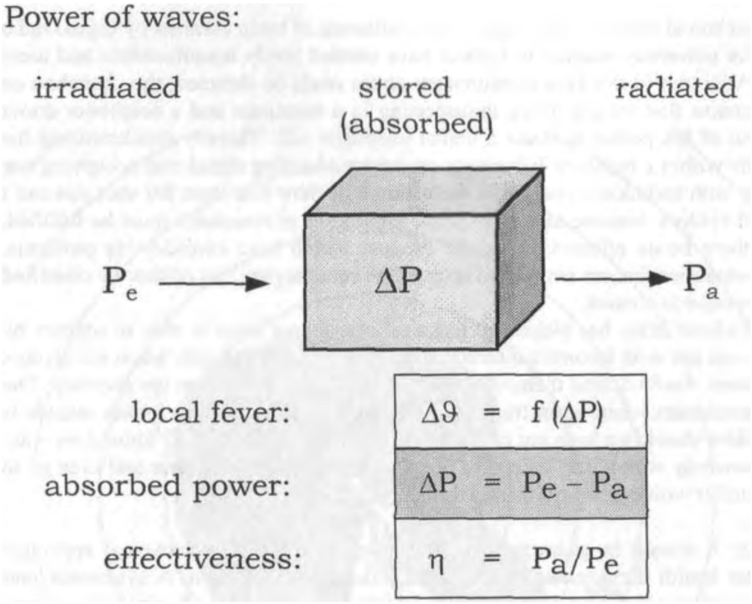
Power of waves:
irradiated
stored
(absorbed)
radiated
P_e
ΔP
P_a
local fever:
$\Delta\vartheta = f(\Delta P)$
absorbed power:
$\Delta P = P_e - P_a$
effectiveness:
$\eta = P_a/P_e$

Fig. 1.3: Damping of waves and ability to absorb of a body (our head) if we are making a phone call with a handy.

a contribution to the theme dielectric losses

- capacitor
- high-frequency welding
- microwave oven

<i>: K. Meyl: Potentialwirbel Band 1 [Al] (only in German),
INDEL Verlag, Villingen-Schwenningen 1990, ISBN 3-9802 542-1-6

<ii>: K. Meyl: Potentialwirbel Band 2 [A2] (out of print),
INDEL Verlag, Villingen-Schwenningen 1992, ISBN 3-9802 542-2-4

1.3 Absorption of waves

The with the theme dealing physicians logically have to put up with criticism that they work only with two phenomena that not at all can be involved authoritative in the causes for biological effects. A third factor can be considered, a field phenomenon until now stayed unnoticed by science: the vortex of the electric field, the so called potential vortex. A vortex is to be considered as an oscillation around a fixed point. Through that a permanent overlap is caused, like what happens at an antenna only that the vortex is not bound to the dimension of an antenna. The potential vortex is contracting and in this way reaches extremely high energy densities at very little spatial measurement, densities that lie far above those that field strength measurements are pretending to us [Al]<i:>.

When again you take the handy at hand with which you ,,blow" the pulsed microwaves into your head. Don't worry, because with the speed of light and without provable damage almost everything comes out again on the other side, but only almost everything. A little damping of the wave has taken place and your head has absorbed this part of the irradiated wave (Fig. 1.3). Who claims this is already the thermal factor actually should realize that there exists no corresponding term in the wave equation. Here there are found merely two dual vortex phenomena as a possible damping term: the eddy current and the potential vortex. An eddy current damping is ruled out because of the bad conductivity of the head. But this favours his dual anti-vortex, the potential vortex [A1]<i>.

Seen physically the following is taking place in your head: the absorbed waves roll themselves up to vortices and through that become localized and overlap themselves permanently (Fig. 1.4b). In the course of time the vortices decay and produce the well-known eddy losses that lead to the measurable increase in temperature. When reactions or biological effects arise, simply and solely the vortex can be considered as the possible cause. Thereby play two points an important role: the number of the generated vortices and their lifetime that is determined by the time of decay.

In anticipation of the mathematical calculation of the potential vortices it is pointed out here that these are favoured not only by a low conductivity, but also by a high dielectricity. Because water has an unusual high dielectricity (ε_r = 80) and our head consists predominantly of water doubts in dealing with handy's are reasonable.

Also the relaxation time constant representative for the lifetime can be calculated [A2]<ii>. We must proceed from the assumption that both the number of the vortices and their lifetime, that is all the at a fixed point in time in our head existing and effective vortices, can be a cause and therefore have to be considered and investigated scientifically.

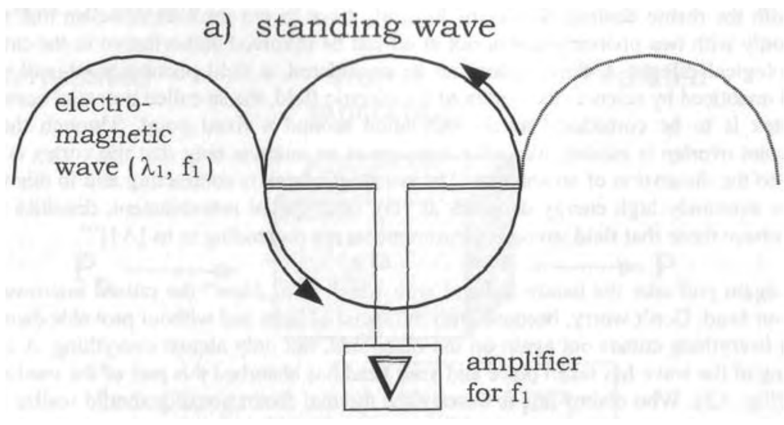
a) standing wave
electro-
magnetic
wave (λ_1, f_1)
amplifier
for f_1

b) localized vortex = noise

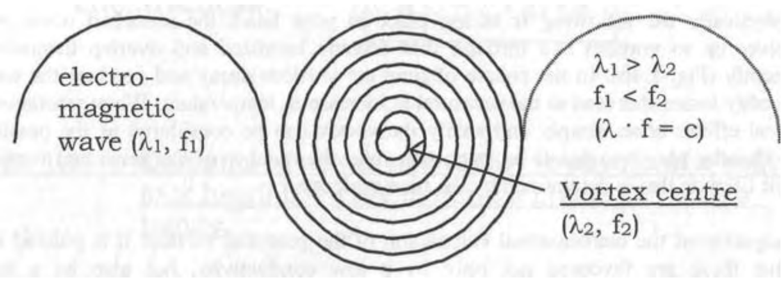
electro-
magnetic
wave (λ_1, f_1)
$\lambda_1 > \lambda_2$
$f_1 < f_2$
($\lambda \cdot f = c$)
Vortex centre
(λ_2, f_2)

c) broadband antenna for EMC-measurements

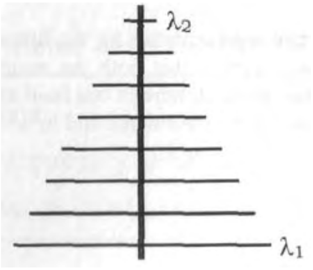
λ_2
λ_1

Fig. 1.4: Measurement of localized waves and vortices

1.4 Overlap effect

The graph at the left (Fig. 1.4) should clarify once more that only the in space localized and permanently overlapping field appearances can be considered as a cause for biological effects. This can concern an over an antenna standing wave (a) or a vortex (b) which is to be taken as localized by nature.
It would be allowed to in general speak only of a vortex because the standing wave can by all means be understood as a special case of the vortex. The essential difference is that the vortex is not bound to the size of a fixed antenna and can contract itself at any time to achieve in this way a substantial higher energy density. As a result this than will lead to an increased biological effectiveness.

It should be pointed at a further omission. In the discussion about limits, without exception, the absolute field strength of the broadcasting signal is valued and not the type of modulation. The last mentioned should actually not at all play a role according to the prevailing scientific opinion.
Totally different is the case with a vortex that acts damping. Such a vortex shows near it's centre a considerable smaller wavelength than more to the outside and through that it has a big frequency bandwidth [A5]<i>. It is to be expected that in the extremely broadband pulsed signals of the digital networks the creation of vortices (or eddies) will be favoured considerably stronger than in amplitude- or frequency-modulated signals (AM/FM/C-network). In connexion with analog modulated radio- or handy-signals until now there never has been reported of any synchronization of the brain currents with handy-signals from a comparison of the EEG with the broadcasting signal.

Interestingly the for EMC-measurements usual stepped broadband antennas have exactly the construction that certainly would be favourable to the measuring technical registration of vortex phenomena (Fig. 1.4c).
With the dipole antennas of different lengths for different wavelengths there still are measured waves and not vortices but these measuring techniques is certainly accommodating to the until now unnoticed and stayed undiscovered vortex phenomenon. So there are some good reasons that the vortex is a dominating influential factor for EMC-problems.

By means of the example of the handheld wireless telephones can be studied and discussed with which set of problems the very young discipline of science of the environmental compatibility has to fight in the future. And in which ways there can be found approaches towards a solution of the problem. When the comfortable and trodden out ways of textbook physics do not lead to the goal than we will have to force our own way through the jungle of science.
At first we'll have to obtain a short overview of the actual level of research and knowledge. From the criticism to this we than can derive the tasks of the electromagnetic environmental compatibility and in particular the unsolved tasks.

<i>: K. Meyl: Wirbel des elektrischen Feldes, eine neue Storquelle? EMC Journal 1/95, 6. J, S. 56-59.

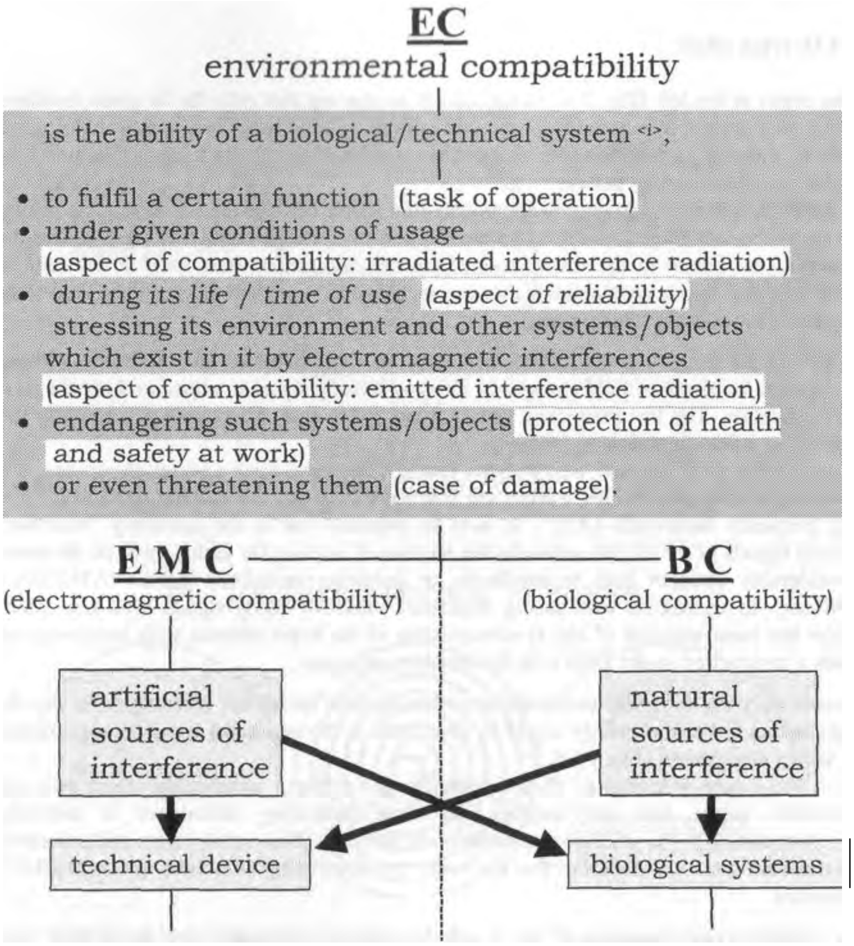

EC
environmental compatibility

is the ability of a biological/technical system <i>,

- to fulfil a certain function (task of operation)
- under given conditions of usage (aspect of compatibility: irradiated interference radiation)
- *during its life / time of use (aspect of reliability)* stressing its environment and other systems/objects which exist in it by electromagnetic interferences (aspect of compatibility: emitted interference radiation)
- endangering such systems/objects (protection of health and safety at work)
- or even threatening them (case of damage).

EMC
(electromagnetic compatibility)
BC
(biological compatibility)
artificial sources of interference
natural sources of interference
technical device
biological systems

Regulations of	called EMEC	know how of
the German Post	thirties: radio	druids
Office since 1992:	operator disease	magic
EMC-law<ii>	today, E-Smog,	esoteric
from 1996:	HF-, SW-therapy	medicine
EU -regulation	diathermy, etc.	biology

Fig. 2.1: Overview concerning environmental compatibility

<i>: E. Habiger: EMV-ubergreifende Aspekte zu benachbarten Problemfeldern, Elektrie 48 (1994), Heft 5/6, Seite 163-161

<ii>: EMVG: Gesetz uber die elektromagnetische Vertraglichkeit von Geraten, Bundesgesetzblatt Teil I vom 9.11.1992, S. 1864

2. Tasks

2.1 Tasks of the electromagnetic environmental compatibility

The environmental compatibility (EC) forms the generalization that includes both the electromagnetic compatibility (EMC) and the biological compatibility (BC). Besides the technical and functional goals of an undisturbed functional course it also pursues ethical and moral goals.
Technology should benefit to humanity and at the same time be in accord with nature. This goal will not be reached when this technology directly or indirectly is endangering humanity.
A direct attack on the health of people poses for instance the military usage of technical apparatus or the negligent usage, by pretended ignorance and unsuspicion.
Is a technology posing a danger to the environment so humanity endangers itself indirectly with this technology. After all are human beings a product of their environment. We always should reckon on the environmental sins taking revenge on us sooner or later.

In fig. 2.1 a formal definition is given that in particular concerns the claims for an undisturbed functional course: it concerns the compatibility aspects of unallowed emitted and irradiated interference radiations, the reliability and quality safety with which a function and task is fulfilled and finally the questions of the protection of health and the safety at work.

Moreover fig. 2.1 provides an overview and the structure of the 2nd chapter. First we'll treat the electromagnetic compatibility (EMC) that first of all deals with the influence of artificial but also natural interference sources on technical apparatus.
After that we'll throw a glance at the appearing fields in nature. The biological compatibility (BC) deals with the influence on biological systems.
An especially sensitive area of the environmental compatibility (EC) than describes the with a cross-link hinted influence of artificial interference sources on biological systems that is popularly described as „electro smog".

The numerous aspects of the environmental compatibility for instance in the areas of chemistry and biology that certainly are important, but do not fall in the area of electromagnetism, can't be treated in the here marked framework.

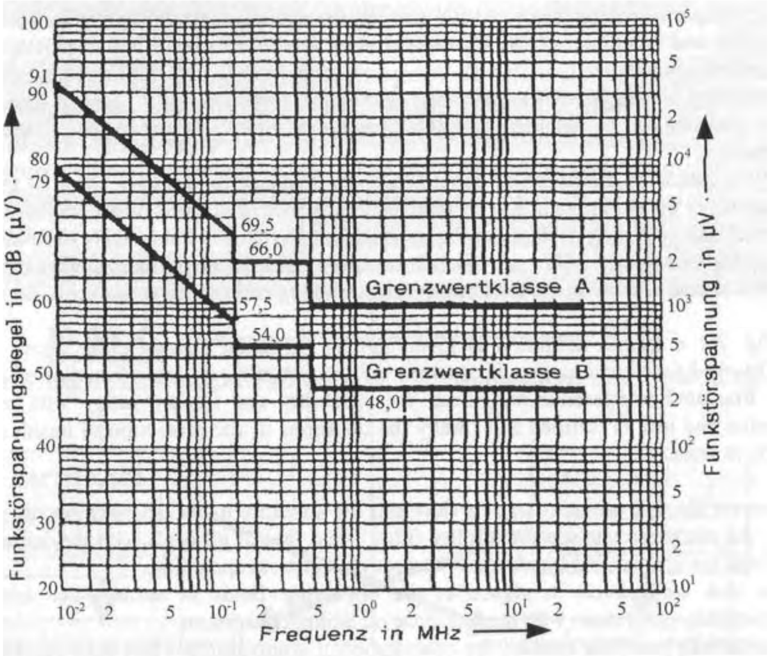

Fig. 2.2: Classes of limits according to VDE 0871
(since 1-1-96: VDE 0875)<i>

<i>: Anton Kohling: Grundlagen der Funkentstorung in der Bundesrepublik Deutschland, etz Bd. 108 (1987), Heft 10.

2.2 Tasks of the electromagnetic compatibility (EMC)

First of all the EMC (electromagnetic compatibility) is concerned with the function of technical apparatus. Correspondingly rational and dry sounds the official definition: "electromagnetic compatibility (EMC) describes the ability of an electrical setting-up (i.e. of a construction element, of a construction group, of an apparatus or of an installation) to function in an electromagnetic environment without stressing this environment by electromagnetic effects in an unallowed fashion ".

Actually it concerns an old need for protection that should be as old as the usage of electro technical apparatus. But in the beginning no one cared about it. The spark gaps with which Heinrich Hertz 1888 in Karlsruhe has carried out the first radio technical experiments were genuine „polluters", that would have been detectible at several hundreds of kilometres distance with modern receivers. For these installations that he had assembled in the lecture room with his students, today he would hardly get a permission of operation and the since 1996 required declaration of conformity he would get not at all.

1925, as in Germany the number of radio listeners had exceeded the limit of one million, for the first time a need for protection appears in the certificate of approval for radio receivers: "The public telegraphs and telephone installations must not be disturbed by the radio receiver".
Later on every backside of the good old steam radios there was found the following hint (translated): "This apparatus meets the interference radiation regulations of the German Post Office". So the manufacturers were urged to measure the emission of their apparatus and in particular to screen the HF-oscillators in the superhet-receivers.

Since the fifties, in the VDE-institute EMC-examinations in the present day sense are taken. The main point of the measurements and the by the VDE recommended limits, however is about interferences bound to a circuit. On the supply lines of the network the prevailing conditions are reproducible so that standards can be put through (Fig. 2.2).

For measurements of interference radiation maybe the time was not ripe enough or the necessity was not big enough. The usual argumentation was: what we can't measure reproducibly, can not be forbidden and certainly not be put under punishment. Therefore merely recommendations were issued or impositions weak as wax were made like: "the interference field strength ... must be so small that an undisturbed reception is guaranteed as soon as the minimum field strength for utilization exists at the place where the antenna is mounted"<i>.

In common parlance that means something like: "as long as no one bleats, everything is allowed". Within a connected industrial area there even existed an officially legitimized fools freedom. Merely at the fence of the industrial area limits had to be fulfilled.

Specially for the line-frequency of the screen one has decided to build a loophole in the law so that one didn't have to throw the TV sets, that so successfully had conquered the living rooms, out of the window. Of course the flickering screens did interfere exactly as before but this EMC-interference now was officially approved.

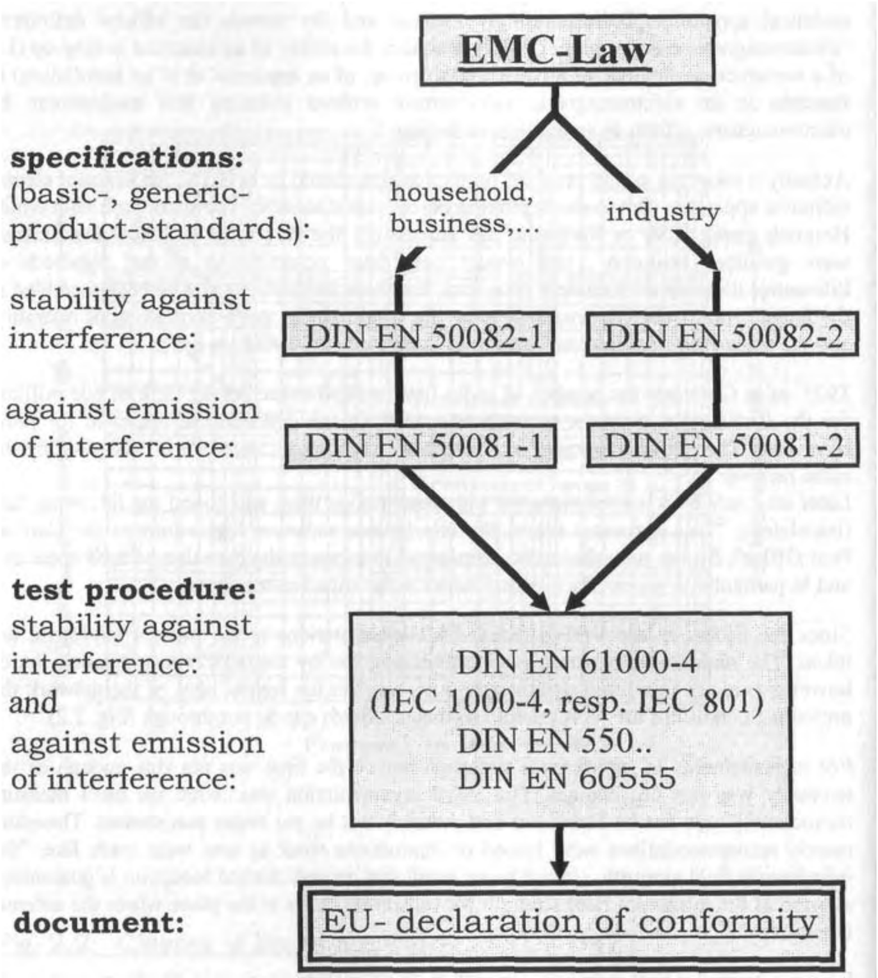
EMC-Law
specifications:
(basic-, generic-
product-standards):
household,
business,...
industry
stability against
interference:
DIN EN 50082-1
DIN EN 50082-2
against emission
of interference:
DIN EN 50081-1
DIN EN 50081-2
test procedure:
stability against
interference:
and
against emission
of interference:
DIN EN 61000-4
(IEC 1000-4, resp. IEC 801)
DIN EN 550..
DIN EN 60555
document:
EU- declaration of conformity

Fig. 2.3: The way to EU conformity

2.3 Declaration of conformity

In the EMC times seem to have gone as the standardizers had to fit in with the insufficiencies of technology. Meanwhile the conditions have turned up. We owe this circumstance first of all the EMC-law of 1992 that doesn't name any limits but it states the political intention to demand from technical apparatus and installations an appropriate stability against interference and at the same time limit the sent out interference. As a consequence of this law the measurement facilities and measurement processes had to be standardized to get reproducible measurement results that are not influenced by the electromagnetic environment. That goes so far that even the floor covering of a measurement hall is dictated because the conductivity of the floor influences the degree of reflexion. Normally the examinee is situated on a revolving plate that is turned around for 360° during the measurement of the radio interference field strength. Is it however not possible to turn the examinee than the antenna has to be led around it, thereby again increasing the dimensions of the measurement hall. The distance to the antenna should be up to 10 meters. Moreover it must be possible to move the antenna, up till a height of 4 meters to register the influence of the reflexions on the floor.

Moreover there is to plan a reflexion free zone around the measurement track (in elliptical form) that depends on the reachable damping of reflexions of the used absorber. Used are pyramids of foam material soaked with carbon and increasingly tiles of ferrite or shieldings of wallpaper.

Taken all together for a measurement hall doing justice to standards there result considerable measurements of for instance 18 m length x 10 m width x 7 m height. Let's again come to talk about the EMC-law with which only the intention but not the way is fixed. To form the claims catalogue in a way that is fulfillable in general, some concrete prescriptions, the so called standards, have to be worked out. This task was transferred to the European committee for electro technical standardization CENELEC, which has established the workgroup TC 110 to at first work out some standards: The basic standards deal independent of product with general questions of the EMC, of the testing process and of the measurement environment.

The generic standards likewise deal independent of product with the so called fundamental technical standards for apparatus in their dependence of the respective electromagnetic environment (protected computer room or medical room, environment of the house, office or industry).

The product standards concern the EMC-standards referring to products (7 product families / approx. 50 products).

In Fig. 2.3 the arduous way through the jungle of paragraphs for a technical apparatus is outlined. Corresponding to the requirements of use, first the relating ES-standards for the apparatus have to be determined and than have to be measured according to own test standards based on the fundamental technical standards. When the allowed limits for stability against interference and for sending out interferences are not exceeded, the EC-declaration of conformity is handed out. Since 1.1.96 that declaration is needed when apparatus are commercialized or - stated more exactly - "put in circulation" and operated. When still further EC-guidelines are met in the end the CE-hallmark is awarded. Since 1.1.96 only with this hallmark the access to the common market of the EC is possible. Violations can be punished with fines and if need be with imprisonment. But there are great national differences in the EC. The Federal Republic of Germany with fines of up to 50.000 Euro counts as expensive for criminals.

Simulation of network for the measurement of the interference voltages U_{st1} and U_{st2}.

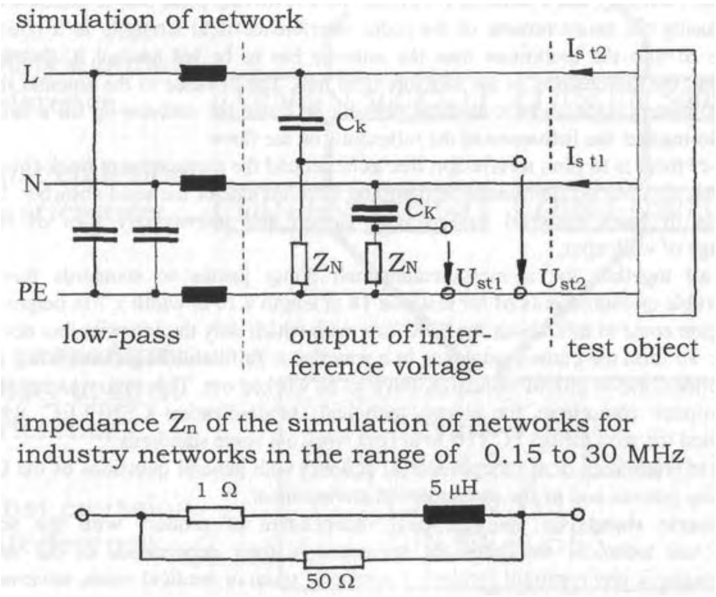
simulation of network
L
N
PE
I_{st2}
C_k
I_{st1}
C_K
Z_N
Z_N
U_{st1}
U_{st2}
low-pass
output of interference voltage
test object
impedance Z_n of the simulation of networks for industry networks in the range of 0.15 to 30 MHz
1 Ω
5 µH
50 Ω

Fig. 2.4: Simulation of network for the measurement of interference voltages.<i>

<i>: acc. to Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.4 EMC-techniques to measure the emission of interference

Actually we already can be glad that it came to an europe-wide agreement for the regulation of the EMC-set of problems. But the question if we can be satisfied with what we have reached is still outstanding. All too often the lowest common denominator of the measurable and checkable was sought and not so much the technical possible was taken into consideration.

The main emphasis is put on the measurement of the emission of interferences. Traditionally the interferences bound to a circuit are registered in a frequency range up to 30 MHz. The corresponding wavelengths thereby can correspond with the length of the supply lines and form standing interference waves. Primarily the spectrum of the interference currents is measured e.g. over a HF-current converter. These currents produce a voltage drop over the internal resistance of the feeding network. Because the properties of the networks can vary very strong, a standardized end-resistor is required for the measurement of the interference voltage. For this purpose an imitation of the network is switched between the network and the examinee. This imitation in addition has the task to keep away the interference signals that come from the supplying network with the help of filter-elements (Fig. 2.4).

The measurement of the interference radiation, the field-bound interference emission, lakes place between 30 MHz and 1 GHz. For that a free field or an absorber-hall with little or no reflexions is required. The standardized distances of measurement are 3, 10 and 30 meters. The electric field strength is determined with dipole broadband antennas, the magnetic field strength with frame antennas. It must be possible to both vary the receiving antenna between horizontal and vertical polarization and to adjust the receiving antenna in the height and the position to the test object.

Typical measurement set up to measure the emission of interferences bound to a conductor[<i>]

A: shielded link conductor
B: bundle of conductors folded like a meander
C: connection to the reference mass
ME: receiver of interference signal
NNB: Simulation of network
PO: test object

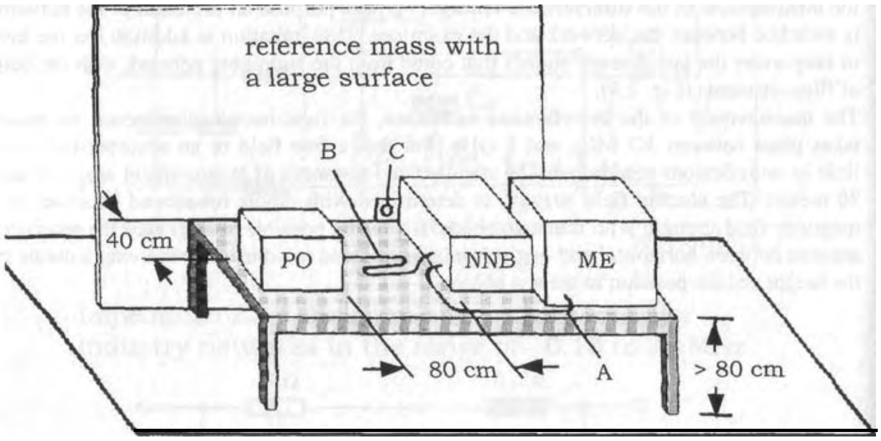
reference mass with
a large surface
B
C
40 cm
PO
NNB
ME
80 cm
A
> 80 cm

Fig. 2.5: Measurement set up to measure the emission of interferences bound to a conductor,[<i>]

[<i>]: Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.5 Electro-Smog

There is almost no end to the possibilities of variation and one needs already a lot of overview and experience to determine the field strength maximum. Nevertheless we have to ask ourselves if in this way really all emissions of interference are understandable, that popularly are described as „electro smog".

Smog is the combination of the terms Smoke and Fog. It therefore describes a pressure on the environment with fog like smoke. When for instance in the case of smog alarm all interference sources are switched off, which means all kilns are brought to a stop and all automobiles are stopped, than the fog like smoke therefore still is not vanished from the air. It just distributes itself and dissolves only very slowly.

The transfer of the smog idea on the electromagnetic interference radiation is bound to fail because, when the test object is switched off no emission of interference at all is detectable with the usual measurement means. Nevertheless the rainbow-press is trying to enumerate almost all electromagnetic field phenomena under the term „electro smog" without consideration of the fact that this term is not at all a collection term. From the sight of an expert one can only speak of smog when something like smog remains and stays active further after the switching off of an electro technical apparatus. It should be a phenomenon that is not understandable by the standardized measurements of interference radiation. Such a phenomenon would be e.g. the vortex of the electric field. However vortices are virtually not measurable in a direct way because they have the unpleasant property to whirl about around the measurement probe. But they will be detectable by their eddy losses and in the case of the electric field vortex appear as noise. Until now the standardizer however haven't planned to investigate the influence of an apparatus on the noise in the environment. Here we still grope in the dark.

At least the vortex shows a storing property that would justify the use of the idea "smog". We'll have to investigate the phenomenon.

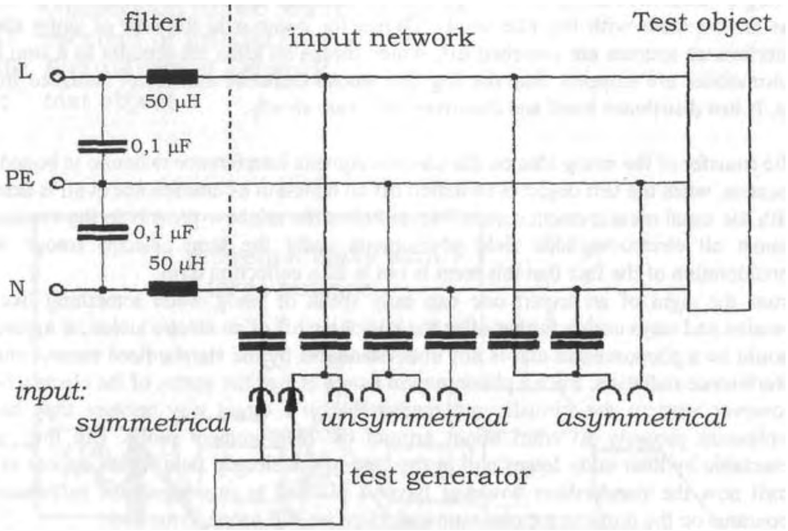
filter
input network
Test object
L
50 µH
0,1 µF
PE
0,1 µF
50 µH
N
input:
symmetrical
unsymmetrical
asymmetrical
test generator

Fig. 2.6: Picking up of interference voltages in network lines.<i>

<i>: Ernst Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2092-X

2.6 EMC-techniques to measure the stability against interference

The question is: what kind and what intensity of an electromagnetic interference can an apparatus cope with without limiting it regarding its technical function. For that test generators are used and with their help interference signals are produced in the test object. Fig. 2.6 shows the possibilities of a symmetrical, of an unsymmetrical and of an asymmetrical production of interference voltage signals in the power supply lines of an examinee. Over and above that the testing possibilities and testing methods are numerous. In principle the following processes are used:

1. The simulation and production of interference factors typical for the network like harmonic waves on network voltages, overlapping signal voltages, changes of network voltage, decrease of network voltage, network interruptions, network unsymmetries and network frequency changes.
2. The production of both energy-poorer and energy-richer interference impulses like they can form in energy-networks by switching acts or by the effect of lightning (burst).
3. The simulation of the course of events when static electricity discharges.
4. Low-frequency magnetic fields like those that can form by network frequency operating and loading currents or by short-circuit and lightning currents in the form of a pulse.
5. The stability against interference against the influence of an electromagnetic field also called radio interference firmness. For this purpose high-frequency generators and broadband antennas are used to expose the examinee to electromagnetic fields in a frequency range between 10 kHz and 18 GHz. At the moment tests are only performed between 27 and 500 MHz. The modulation of the carrier wave should be possible to be able to imitate the interferences by radio technology as realistic as possible. Thereby the field strength values can by all means reach up to several 100 V/m.

In accordance with expectation the result of this irradiation with an outside field is that every conduction path and every wire can act as an antenna and therefore can produce high-frequency currents and measurable potentials. Building parts of the analog technology as a consequence battle with problems of drift whereas with digital logic parts and computer parts the danger exists that switching states change unintentionally. Let us remember again the overlap effect of fig. 1.4. The electromagnetic wave itself marches with the speed of light through the examinee. When a small part of the wave finds an object that it can use as an antenna than the localized overlap comes into play. This than as a cause is responsible for the effective and measurable antenna currents. Until here the text books help us to explain the problems that happen and to remove them logically.

However time and again cases are reported where textbook physics can't help us any further. Spectacular cases even came to court like e.g. the ABS (Antilock Braking System) of a truck that had failed due to EMC-interference radiation. As a consequence the brakes had failed. When after that the local radiation pollution is measured no anomaly at all can be discovered. The measurable field strength is not higher as is usual in the whole area. Maybe you also have made the experience that often the causes can't be found when your computer suddenly "crashes" out of the blue.

Here the mentioned vortex of the electric field is capable to deliver plausible explanations because it is not bound to the geometry of an antenna and in addition is highly active without being detectable with the usual EMC measurement methods of the interference radiation measurement!

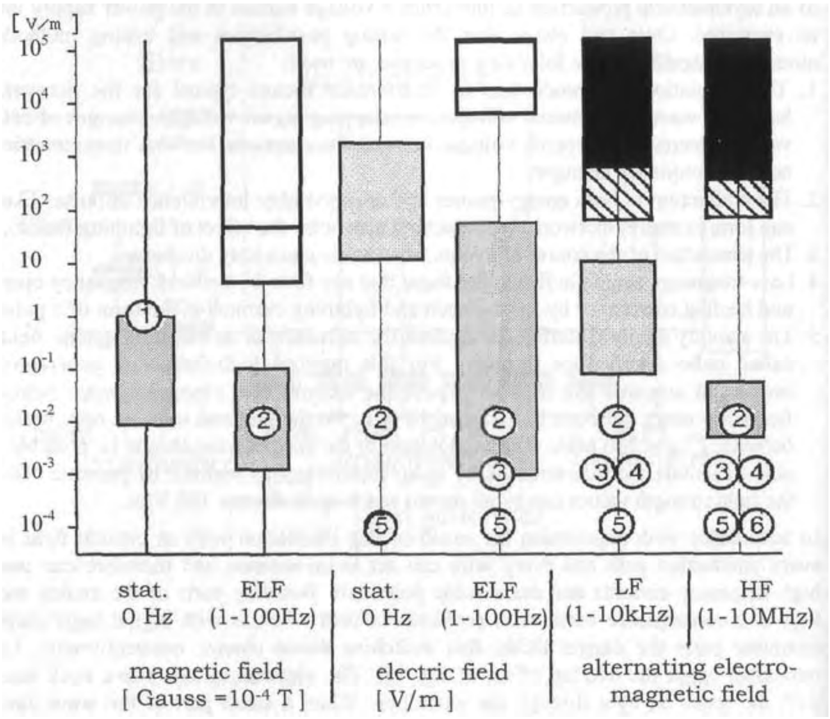

Fig. 2.7: Intensities of natural electromagnetic fields
(electric, magnetic and electromagnetic fields) ▨

In addition to fig. 1.1, page 3,
magnetic field in Gauss (= 10^{-4} T),
electric field in V/m.

<i>: A.S. Presman: Electromagnetic Fields and Life.
Plenum Press, New York - London, 1970

2.7 Tasks of the biological compatibility

The second leg of the environmental compatibility (EC) forms the biological compatibility (BC) besides the electromagnetic environmental compatibility. Whereas the interests of the EMC are looked after by electrotechnical engineers and electrophysicists, with the BC also doctors, biologists and architects are addressed.
Moreover this leg is already very old and already existed long before artificial interference sources could be created by mankind. The interaction between the arising interference sources in nature and the biological systems in general and specially men always interested the doctors, the priests, the druids and geomants, that not seldom looked after several functions in personal union equally. Unfortunately they as a rule have kept their knowledge and capabilities as secret knowledge, didn't make any recordings and merely initiated and trained their respective successors. Through that a great deal got lost and today non-medical practitioners, homeopaths and esoterics trouble themselves to take up to the far-reaching buried knowledge.
Because this concerns pure knowledge by experience, often the with the topic occupied persons themselves are not capable to say anything about the physical backgrounds and causes. One concentrates entirely on the under certain circumstances reachable results and only in rare cases on reproducable effects. In some areas the scientific assignment already has succeeded, have parascientific phenomena managed their admission in the so called "exact sciences", but in most experience disciplines the assignment is still due. There still is a lot to do here.

In the time as there not yet were operated any artificial interference sources on our planet, the senses of man naturally were a whole lot sharper for his electromagnetic environment as today. Today, where there scarcely is a place on earth where we are not irradiated by terrestrial transmitters, by satellites or by the netfrequency that is measurable everywhere. In the bluntness of our senses perhaps the hybris of modern man is founded, with which he wants to rise himself above esotericism, geomancy and other sciences of experience and thereby dispute the electric and magnetic fields their biological effectiveness.

The fields of natural origin form an electromagnetic environment for men, that they have adapted to and that they probably need for a life in accord with nature. The evolution has taken care for a corresponding adaptation.
In fig. 2.7 in addition to the limits from fig. 1.1 the intensities of natural electromagnetic fields are registered[i]. They lie clearly lower as the recommended limits but exactly in the area wherein the first reactions of living beings are observable.

When we ask us how much electromagnetism is good for us and how much harms us so the obvious answer is: exactly as much radiation as nature dictates in the fluctuations between day and night, between the months, years and in the end between the cycles of sunspots of 11 years. Here the guide value is found that man and nature have adapted themselves to. In fig. 2.7 the corresponding area between the natural minimum and maximum values is given.

Irradiation strengths of the field radiation in the biosphere and how the optical windows are situated in the atmosphere<i>.

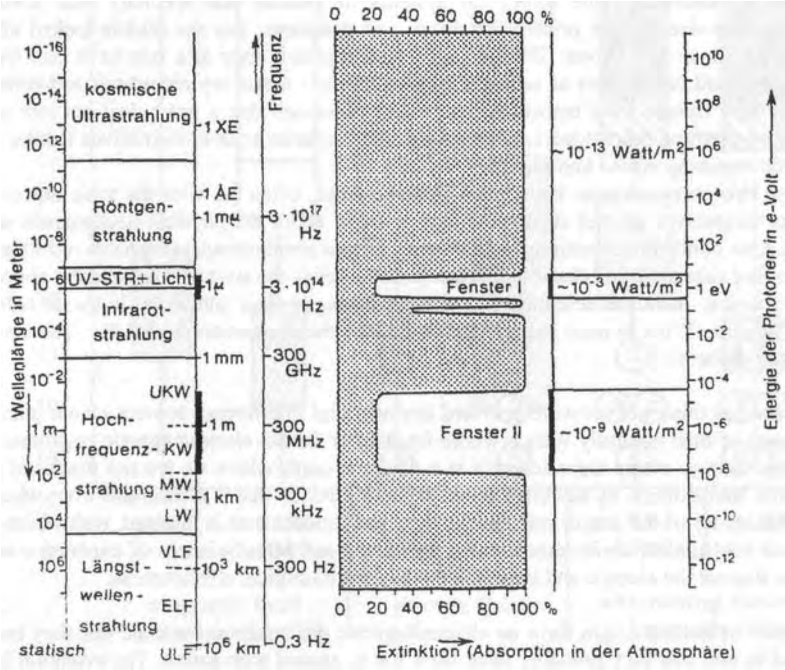

Fig. 2.8: Absorption dependent on frequency in the atmosphere

---

<i>: entnommen aus: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Bild 8, Seite 14, Verl. Moos & Partner Milnchen, ISBN 3-89164-058-7

2.8 Natural fields

Our electromagnetic environment has something to offer:
From the magnetic north pole to the magnetic south pole of the earth run the field lines of the earth's magnetic field that we are exposed to. With a compass we use the vector character of the magnetic field to fix our position. The induction averaged over time is approx. 50 uT. But it is overlapped by short-time fluctuations caused by geomagnetic storms in the ionosphere.

These storms again are caused by the eddy currents and the currents of charged particles that come from the sun. At the same time these eddy currents in the ionosphere together with the earth's magnetic field form a protective shield with a excellent screening effect for us inhabitants of earth.
In several layers like for instance the ozone and Heaviside layers a filtering and damping until the complete suppression of the very broad cosmic spectrum is caused. This extraterrestrial spectrum of radiation doesn't leave a single frequency out and has a lethal intensity for us.

Only for a little window in the frequency spectrum, radiation can pass almost undamped, as can be seen in fig. 2.8: the light with the spectrum of the colors. For this nature has donated man a sense organ so that man can protect himself against too high dose values. After all, who will look voluntarily into the sun? We only get into trouble when our sense organ doesn't function any more (for instance in the fringe range of the visible spectrum, the UV-range).

For other frequencies of electromagnetic radiation man neither has a sense organ but that doesn't mean that he is not influenced by these. Here, as in the UV-range he only indirectly notices that he has got too high a dose when he has to discover some influences on his well-being and his health. Without the help of neutral measurement apparatus he himself by no means is in a position to make a connection between an excessive exposition to radiation and his health problems.

When natural field strengths should be used as a measure for technical limits, so there should be paid attention to the fact that nature doesn't know intense continuous irradiation. The values are subject to powerful fluctuations that leave men and nature the chance to regenerate.
The television stations not even think it is necessary to reduce their broadcasting power after the end of broadcasts and further sprinkle the sleeping population with test signals, with senseless pictures of going by underground or nonstop program advertisements.
People need the intermissions. That again shows how good nature means it with us.

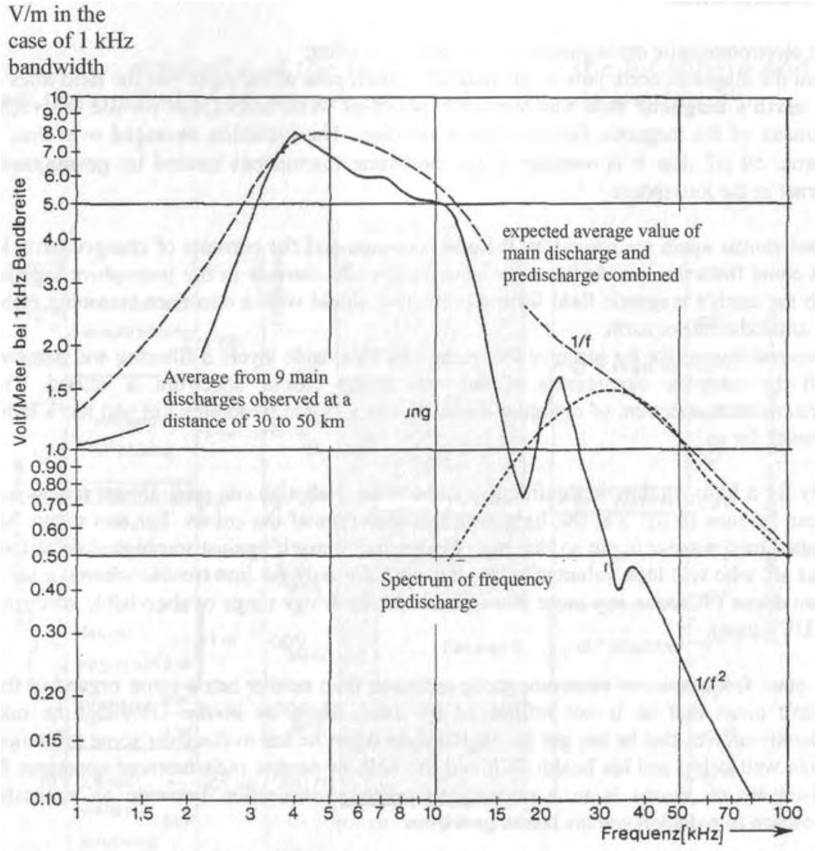

Fig. 2.9: Spectrum of frequency of one lightning, measured
field strength at a distance of 1.6 km from the place of origin. See Watt and Maxwell:[i,ii]

---

from:

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligk.), 5.Aufl., Bild 38, Seite 27, Verl. Moos & Partner Munchen, ISBN 3-89164-058-7, see <ii>.

2.9 Electricity of the air

The electrostatic field strength of the earth lies between 50 and 500 V/m. That is a whole lot considering that voltages off 60 Volts can be lethal for man. But a living person distorts the terrestrial E-field for reason of his own field and his electric conductivity, so that there exists no danger for him as long as he doesn't grow into the sky.

Maybe the dinosaurs had to become extinct because they were to big and for instance the E-field near the ground had risen with a jump by a meteorite that brought a high charge from the cosmos. That would explain why the smaller living beings could survive the natural disaster.

Also the thunderstorm electricity can become life-threatening. Burns, heart and brain failures are the most common consequences. After all the probability to be struck by lightning is four times higher as to have six right ones in the lottery.
Over the lightning channel of approx. 1 meter in diameter charges between 10 and 200 C are transported what results in current strengths of 2000 up to 200,000 A. The main discharge lasts between 10 and 50 usec. With the preceding and all the following discharges it lasts over a second.
Field strengths on the scale of 4,000 V/m are typical but in a distance of 5 km these wear off to 8 V/m. The frequency spectrum of a lightning reaches 4 powers of ten into the range of the radio waves. In fig. 2.9 is shown the field strength measured in a distance of 1.6 km from the place of origin<i,ii>.

The origin of lightnings is still an unsolved problem after the well-known models (Wilson) arc not in a position to explain the reason for the origin of the potential difference of more than 100 million Volts required for the ionization of the air. Also the lightnings that struck in the direction of the ionosphere still are mysterious. We'll have to come back to this<iii>.

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 38, Seite 27, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7, according to <ii>.

<ii>: A. D. Watt and E. L. Maxwell: Characterisic of Atmospheric Noise from 1 to 100 Kc/s. Symposium on the Propagation of VLF Waves, Boulder, Col., Paper 35, Jan. 1957.

<iii>: see Part 1, chapter 5.4 and part 2, chapter 14.11.

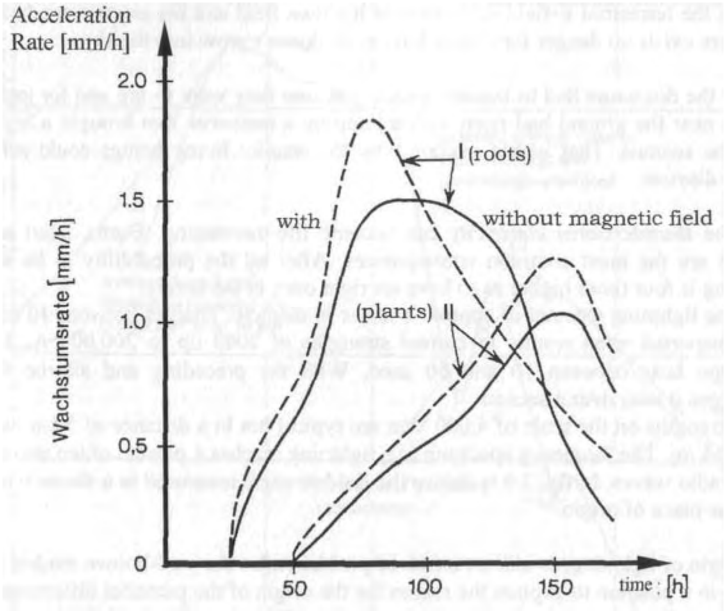

Fig. 2.10 A: Effect of a static magnetic field of 100 kA/m (0,12 Tesla) on the root (I) and on the plant (II) of barley seeds.
plants in the magnetic field: dotted line
plants for checking: drawn line
according to Novitskii[iii]

[iii] Novitskii, Yu.I.: Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966.
Taken from: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 72, S. 73, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

2.10 Biological effects

The in fig. 2.1 indicated connection between the EMC and the BC, by some authors inofficially described as EMC-environment (EMCE), describes the effect of artificial fields on biological systems. This concerns the sensitive range of tasks that is discussed extremely controversially in the public. The problem thereby is that the artificially produced field strengths lie above the natural field strengths for several scales. In the thirties first reports about troubles were provided by navy radio operators that complained about headache, dizzyness, concentration failure and indisposition. Besides these negative reports concerning the so called "radio operator disease" at the same time medical usages concerning high-frequency therapy were tested. In the beginning this diathermy called healing method still was a short wave irradiation. Today it is extended into the microwave range and uses the thermal effect of electromagnetic rays. The increased temperature of the tissue causes an increased local blood flow. This supports healing processes, loosens cramped muscles and can help in case of rheumatic fever. The advantage of the HF-irradiation compared to contact heat by a hot-water bottle or by infrared rays is the higher penetration depth. Herein short waves are superior to microwaves. But microwaves can be better focussed on a certain part of the body. Is the temperature further increased, so the tissue is damaged. This is used for the treatment of cancer and is called hyperthermy. Because cancer cells as a rule are flowed with blood worse than healthy cells, they are more sensitive to heat and therefore are faster destroyed than healthy cells at a correspondingly increased temperature. In this way for Instance in the USA cattle with a so called cancer eye are treated. For that the spot suffering from cancer is irradiated with 2MHz-waves for 30 seconds with a handheld apparatus of 10 Watts broadcasting power. The rate of succes is given to be 90%! The method of hyperthermy has not yet been able to establish in the area of the medicine for humans. Also at our college corresponding research work is carried out in co-operation with the radiological clinic of the university of Freiburg (Germany)<i>.

The thermal effects of high-frequency fields are therefore well-known and subject of scientific research. On the other hand and in spite of numerous publications, non-thermal effects even today are denied by some scientists as non-existent<ii>. Here only a few counter-examples will be listed.

Fig. 2.10 A shows the effect of a static magnetic field of 0.12 Tesla on the root (1) and on thc plant (II) of barley seeds. The readable effect is an acceleration of the growth of the treated seeds (dotted line) compared to the plants for checking (drawn line)<iii>.

<i>: H. Schulz, W. Oppitz: Lokale Hyperthermie durch hochfrequente Wirbelstrome, Medizin Technik 1, 1987.

<ii>: G. Nimtz: Mikrowellen, Einfuhrung in Theorie und Anwendung. 2. Aufl., BI-Wissenschaftsverlag 1990, ISBN 3-411-03203-0

<iii>: Novitskii, Yu.I.: Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966; taken from: H.L. Konig: Unsichtbare Umwelt, fig.72, p. 73, Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

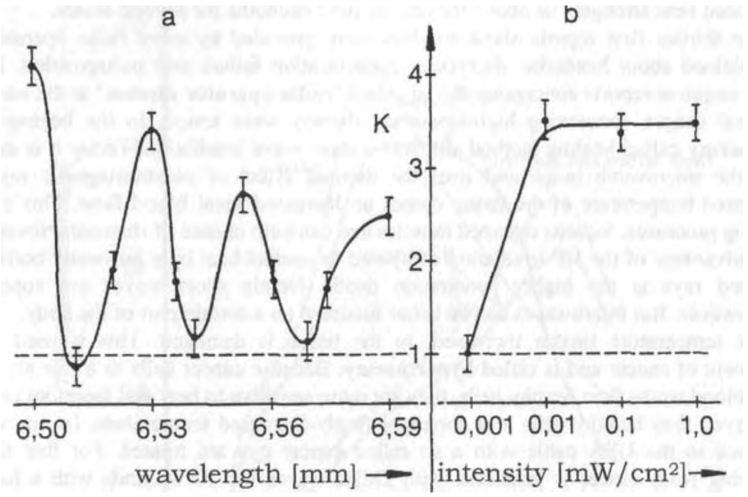

Fig. 2.10 B: Measured increase in the production (K) of colicin by colibacteria as a microwave effect

a) as a function of the wavelength,
b) as a function of the intensity of the microwaves,

according to Keilmann[i].

---

taken from:
[i]: H. L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 106, S. 111. Verlag Moos & Partner Munchen, ISBN 3-89164-058-7

A static field naturally produces no induction and hence no heating will arise. In the case of alternating fields the thermal effect in experiments is excluded by working with cxtremly low stimulations. The example after fig. 2.10 B shows the measured increase in the production (K) of colicin by colibacteria at only 0.01 mW/cm^2 microwave power<i>.

In addition the example provides the interesting information that obviously only a certain frequency and it's harmonic waves increase the production, other frequencies on the other hand remain inactive. Because only minimal field strengths are used it more likely concerns an information-technical as an energetic effect (curve a). This statement is supported by the observation that an increase of the intensity not at all necessarily as a consequence also increases the production (curve b). What the colibakteria need is obviously neither energy nor heat but only a certain frequency that stimulates the colicin production or the growth.

Should it really be confirmed that biological effects of electric and magnetic fields can be produced by certain frequencies and can't happen by an energy transition so the discussions about limits must seem ample meaningless.

Maybe the one or the other in thought already draws a connection to the acceleration, the acclerated growth of kids, which is observed world-wide and stronger in cities than in the country. It started for approx. 100 years simultaneous with the electrification of the homes in the towns. In Asia the acceleration and also the electrification have started later. Other growth stimulating effects like radio waves, X-ray examinations, atomic tests and provable also the nourishment with vitamin B6 happened only until much later and at the most could support the actual effect.

But how should a proof be offered when anyway the field strength not at all can have a decisive influence on the growth of man after the statement of fig. 2.10 B? Which information is authorative? Where lies the responsible frequency window? Does the information actually manifest as frequency? Is the authorative influential factor also in this case not at all noticed and measured?

A lot of pressing questions are still outstanding. But in any case the numerous influential factors detected in experiments do not at all let themselves reduce to a sole factor, for instance the nourishment. For a family doctor it may be comfortable to be able to give an easy explanation: „Cause is the nourishment!" With such a reductionism on the other hand the actual cause stays in the dark and the asked questions in this way won't let themselves be answered.

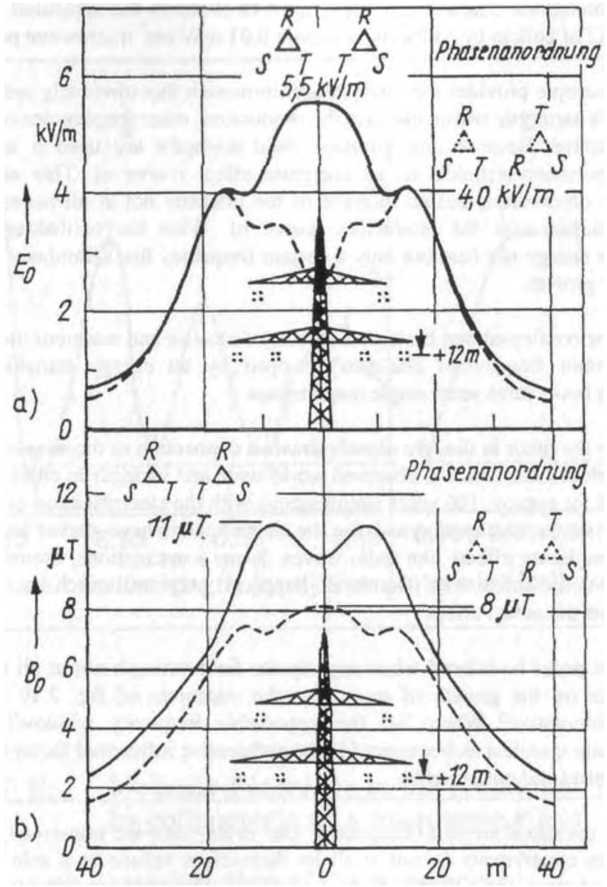

Fig. 2. 11: ____ profile of the 50-Hz-field on the ground
at 380kV/1kA in each circuit.<i>
a) electric field,
b) magnetic field.

____ taken from: ____

<i>: E. Habiger u. a.: Elektromagnetische Vertraglichkeit, fig. 7.3, page 147 and Fig. 7.1, page 146, 2nd Ed., 1992. Berlin, Miinchen: Verlag Technik.

2.11 Artificial fields

The scepticism of people feeling close to nature is especially directed against artificial fields that man can't see nor hear nor smell. Objects of doubt are first of all the installations for the creation, distribution and conversion of electric energy.

An essential role plays the fact how close the supplying and the draining conductors are to each other, so that the respective fields can compensate each other. The worst solution one can think of is realized at the electrified railway. Here the rails and the earth are used as the draining conductor for the current while there exists an ample large distance to the supplying conductor. A compensation is almost impossible thus causing gigantic interference fields that are detectable even at a distance of 10 kilometers. The increased putting on of railway engines feeded by rectified current is likely to aggravate the set of problems because the non-sinusoidal absorption of current is strongly afflicted with harmonic waves.
With high tension transmission lines (fig. 2.11) the interference field strength is reduced when the three cables are suspended with only little distance between them. But even the selected phase order can play a role. Of course the optimal compensation effect is guaranteed with cables in the earth. But these are expensive and technically not realizable for such high voltage levels.

In the eighties also the computer screens got in the headlines. The terminals are furnished with a cathode ray tube and have a very broad radiation spectrum that already starts at 0 Hz. Here already static maximum values of 64 kV/m are measured<i>!

I'requency range	Measured maximum values E_{max} resp. H_{max}	30 cm in front of the screen
static field (0 Hz)	64 kV/m	
30 Hz. 60 Hz	10 V/m and	
	0.2 - 1 A/m	
5 Hz - 1 kHz	1800 V/m and 4 A/m	10 V/m and 0,6 A/m
50 Hz - 0,5 MHz	1 A/m	
15 kHz-220 kHz	50 V/m and	15 V/m and
	1.1 A/m	0,17 A/m
3 MHz -300 MHz	< 0,2 V/m	

Fig. 2.11: Electromagnetic fields from screens<i>

system about	used frequency	measured value
radio broadcasting station (MW 2O kW)	6OO kHz	2...17 V/m
radio broadcasting station (SW 1OO kW)	15 MHz	1...25 V/m
SOS-transmitter on a ship (1OO W)	41O kHz	1...3 V/m
Epitaxiedevice (induction oven)	45O kHz	37...4OO V/m
HF-welting press (welding of plastic foils)	27,12 MHz	70...85 V/m
radar on a ship (TRN 311)	9,3 GHz	1...3O uW/cm^2
radar of an airplane	9,2 GHz	45O... 28OO uW/cm^2
domestic appliances measured in a distance		of 3O cm:
hand mixer	5O Hz	5O V/m
refrigerator	5O Hz	6O V/m

Fig. 2.12: ____the electric field strength resp. Power density in our environment.<i>

____ taken from:
<i>: E. Habiger u. a. : Elektromagnetische Vertraglichkeit, Fig. 7.2, page 146, 2. Ed., 1992, Berlin, Munchen: Verlag Technik, ISBN 3-341-00993-0

2.12 Protection against artificial fields

Artificial fields more or less always occur in the neighbourhood of electric apparatus and installations. Especially problematic are those that work with the free radiation of electromagnetic fields, that is all the radio broadcasting stations, handheld and radar transmitters.
Herewith it is important that not needed parts of the antennas are shielded, that antennas with little close by field pollution are used and that the stand should be situated at least 3 km remote from inhabited areas. For instance at radar installations damping values of 10 dB and more can be obtained only by using a corresponding tree growth.
This obviously concerns a damping of the waves in a dielectric manner. We'll have to come back to this because textbook physics does not know a corresponding damping term in the wave equation.

The radiation leaking out in case of the high-frequency welding of plastic foils and of the microwave oven should be minimized.
In the case of induction ovens or of motors an active shielding often causes problems so that for simple domestic appliances like a hand mixer and especially for the electric hair-dryer non proportionally high field strength values are measured. Fig. 2.12 informs about it.
Protective measures for the operator are hardly possible. To protect uninvolved people not only the apparatus but also the rooms and eventually whole parts of buildings had to be shielded and grounded.

Sometimes also fairly simple possibilities are helpful like e.g. the usage of a remote control. By clearing away the cable salat at the workplace and at the sleeping place induction loops can be removed. Alarm-clocks operated by batteries should have preference over those operated by the network. Mattresses with metal in them and spring-beds which clearly act as an antenna should be avoided. In extreme cases even so called "current-free switches" and shielded network lines are recommended (fig. 2.13).

In the area of the network supply lines a choking coil can help decrease the spreading of high-frequency interference radiation. It is especially important that all the conducting metallic objects like e.g. water pipes, heatings, steel racks, machines, switching racks, steel armaments and metallic windows should be grounded properly, because otherwise extremly high static charges could result instead of a shielding. Construction biologists recommend to when possible do without metals when building houses and furnishing, what of course is only realizable with limitation.

In any case numerous measures are known that to a lesser extent find their legitimation in classical physics, but more likely in precaution. As long as we do not know which phenomenon causes the electrosmog and we don't have a measuring-instrument at our disposal, precaution is really the only thing we can do irrespective of the effectiveness of the measures and of the arising costs!

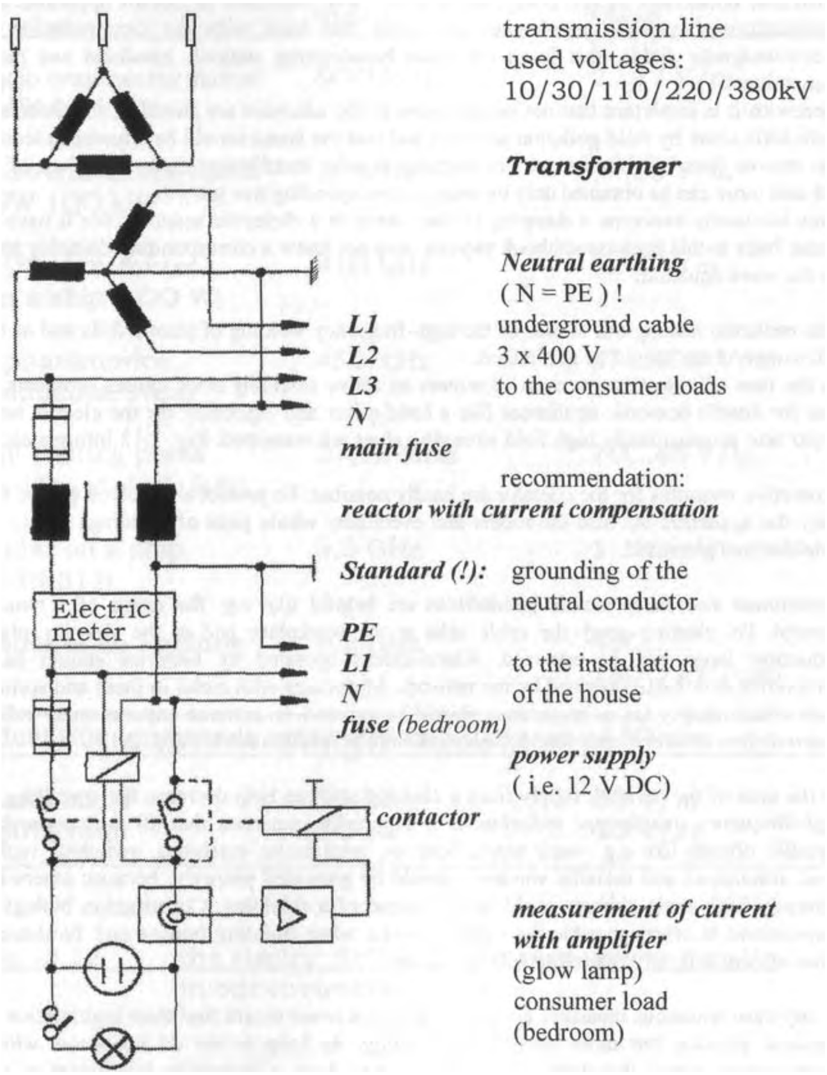
transmission line
used voltages:
10/30/110/220/380kV
Transformer
Neutral earthing
(N = PE) !
L1
L2
L3
N
underground cable
3 x 400 V
to the consumer loads
main fuse
recommendation:
reactor with current compensation
Standard (!):
grounding of the
neutral conductor
Electric meter
PE
L1
N
to the installation
of the house
fuse (bedroom)
power supply
(i.e. 12 V DC)
contactor
measurement of current
with amplifier
(glow lamp)
consumer load
(bedroom)

Fig. 2.13: About the circuitry and the problems involved with a "current-free switches" installation

2.13 Unsolved tasks

The report concerning the actual state of research could be continued at will. But the expositions should suffice, to understand what are the tasks of the electromagnetic environmental compatibility (fig. 2.1) and which questions still have to be supplied with a solution. One can get deeper into every of the addressed points and then discover that some questions can be expressed sharper and maybe conceivable answers can be found, but at the same time and unavoidable the number of new questions increases faster.

Let us again take up the example of the handheld wireless telephones (chapter 1.2). At least it now is clear to us that the usage of the built-in microwave antenna of a handy is problematic. In the interior of an automobile it never should be used. If, however, one uses the antenna installed on the outside on the sheet metal then the damping and screening effect of the sheet metal chassis is advantageous at least for the handy user.

With that of course the central question is not answered. The question of what the cause is for the interfering and at worst health endangering effect of the HF-radiation. Field freedom we can't find anywhere on this world. Possibly we even need the fields. But then the question is how much is necessary, how much is healthy and how much makes us ill.

The gap of explanation especially gets clear in the case of the static or of the low-frequency magnetic field: away from technical interference sources normally fields on the scale of 10 nT are measurable. Construction biologists say that 20 nT, so twice that value, should not be exceeded at the sleeping place and maybe 50 nT at the desk. These values however are determined purely empirical.

When a person is examined in a nuclear magnetic resonance tomograph that person is exposed to a field that lies between 0.2 and 1.5 Tesla. that is a value 7 till 8 powers often higher than before mentioned without this leading to the death of that person. Entirely on the contrary this method is regarded as especially caring and safe compared to the X-ray examination.

Here again the legitimation of the thesis put forward is entirely comfirmed. The thesis that the well-known physically measurable and controllable phenomena can not be considered as a cause and that possibly a until now undiscovered field phenomenon should be called to account.

Should such a phenomenon exist it should be derived, calculated and proved. We must go to endless troubles and try everything. The actual difficulties wherein the electromagnetic environmental compatibility is stuck are a challenge.

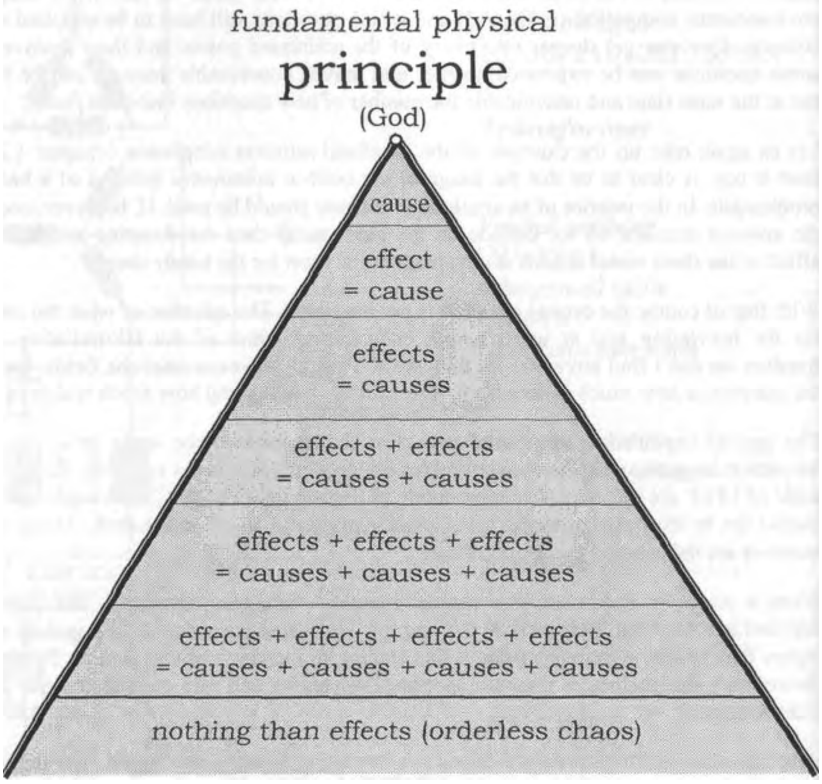
fundamental physical
principle
(God)
cause
effect
= cause
effects
= causes
effects + effects
= causes + causes
effects + effects + effects
= causes + causes + causes
effects + effects + effects + effects
= causes + causes + causes + causes
nothing than effects (orderless chaos)

Fig. 3.0: pyramid of causality

vortices are a consequence of the principle of causality

3. Approach

In the question, if there exists a still unknown phenomenon that influences the electromagnetic environmental compatibility, we must fall back far until upon the roots of our physical understanding. Here we find a fundamental principle that until today is not doubted and that is regarded as elementary, the principle of causality. Every result of a measurement, every interpretation is checked for causality and only after passing this examination it is accepted and published.

This principle of cause and effect has established, not only in physics but also in many other disciplines of science. Is an effect observed, so there immediately is asked for the cause. This principle encounters us in daily life<i>.

When all observable and measurable effects ever can be assigned to a cause without force and without exceptional regulations then the logical result is a pyramid of causality. On top a fundamental physical principle is found, that is regarded as given by nature or as given by god and that with its properties is responsible as the cause for different effects. These effects again appear as the cause for new effects and so on (Fig. 3.0). Sometime we have removed us so far from the top of the pyramid that a direct reference to the describable effects can't be made anymore, so the impression could arise that it concerns an isolated and independent discipline. We should take care not to think in such a monocausal way, because both delimitation and avoidance of interdisciplinary working methods will inevitably steer us into a dead end!

This pyramid of causality stands for the vision of a "unified theory", like it is demanded and sought-after by numerous research scientists. But as long as it is not found, we'll have to do with unsolved problems of causality. About this any number of examples can be given.

A physical principle based on the principle of causality is the vortex. This the eddy current demonstrates us clearly. The cause for its origin is an alternating field. According to Faraday's law of induction this induces a voltage that in a conducting medium results in a current according to Ohm's law. Around this current according to Ampere's law an alternating field forms, that points perpendicular to the current and overlaps the original alternafing field. This induced field first of all is an effect that overlaps the cause and itself becomes the cause. The effect that follows from that further overlaps and forms a new cause etc. In this way vortices form.

Vortices quasi represent the principle of causality.

<i>: When for instance a woman complains: "Doctor, my left knee hurts" (effect). The doctor diagnoses the cause: "Yes, that comes with age!" With that causality is established. "But doctor", says the woman, "my right knee is exactly as old as my left knee!" And already the doctor has a new problem of causality.

Principle of Causality:

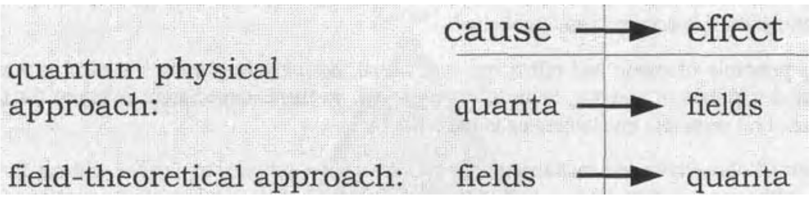

	cause	→	effect
quantum physical approach:	quanta	→	fields
field-theoretical approach:	fields	→	quanta

violations of the principle of causality:

1. monopoles exist
2. starting point for the strong interaction
3. fields and quanta are a cause at the same time
4. hypothetical particles (gluons, quarks, etc.)
5. transmission of information with speeds faster than light
    - with photons (University of Berkeley)
    - with microwaves (University of Cologne)
    - with laser beams (Technical Univ. of Vienna)
6. transmission of energy with speeds faster than light
    - with scalar waves (Nicola Tesla)

Fig. 3.1: Causality or the principle of cause and effect

3.1 Principle of causality

Our physical view of life strictly obeys the rules of causality, the principle of cause and effect. But there are numerous cases, where causality at first could not be fulfilled anymore. Here alternate solutions had to be found to not endanger this very successful principle. A few examples should clarify this:

1. Technically it is impossible to produce a magnetic monopole. When a north pole is produced then the accompanying south pole is also formed. In the same way only the positive and negative pole can be assembled as the so called dipole. In the microcosm however we find monopoles. Electrons are such particles. To restore causality we must grant the microcosm its own laws that are not valid in the macrocosm! But this monocausal hypothesis contradicts the observation that the microcosm represents an image of the macrocosm and vice versa. Doubts if this assertion is allowed are reasonable.
2. Like charges repel each other and that the more the smaller the distance gets. In an atomic nucleus positively like charged protons are together at the smallest possible room without any repulsion happening. Arithmetically seen all atomic nuclei would have to explosively fly to pieces. But because this doesn't happen, shortly a new and supposedly fundamental interaction, the strong interaction, was introduced to save causality. Nevertheless this interaction now holds the like particles in a not explained manner together. Causality could be obtained only by the introduction of a new fundamental phenomenon.
3. When causality should hold as the supreme principle, it should be demanded with priority for the fundamental phenomena of physics. Instead, in quantum electrodynamics the particle is attributed the same physical reality as the field. With the wave-particle duality Heisenberg has given out the corresponding formula of compromise. This slaps the face of the principle of cause and effect.

Causality on principle allows only two approaches for a solution: the quantum physical approach. which holds the quanta as the cause for the fields, and the field-theoretical ašproach. wherein only the fields act as the cause. For both approaches there are good arguments. The field theorists cite that fields can exist also in the vacuum, so that there exist fields without particles but never particles without fields. Against that the quantum physicists hold that somewhere, even when quite far away, there exist particles and that the measurable fields merely are their action at a distance.

Both approaches first arouse the impression to be fully equal. In the course of the discoveries in the area of quantum physics, the corresponding approach has been able to establish. But it demands that all phenomena have to be understood as a consequence of particles. So should gravitons make gravitation possible, should gluons hold everything together and the quarks form the basic building parts. Meanwhile there is only worked with hypotheses. Out of poverty quantum physics meanwhile has said goodbye to strict causality, after the number of the violations of causality has risen that much and in every respect there is a lack of models of explanation. It seems as if the end is reached, as if the quantum physical approach to a large extend is exhausted.

Field-theoretical approach:

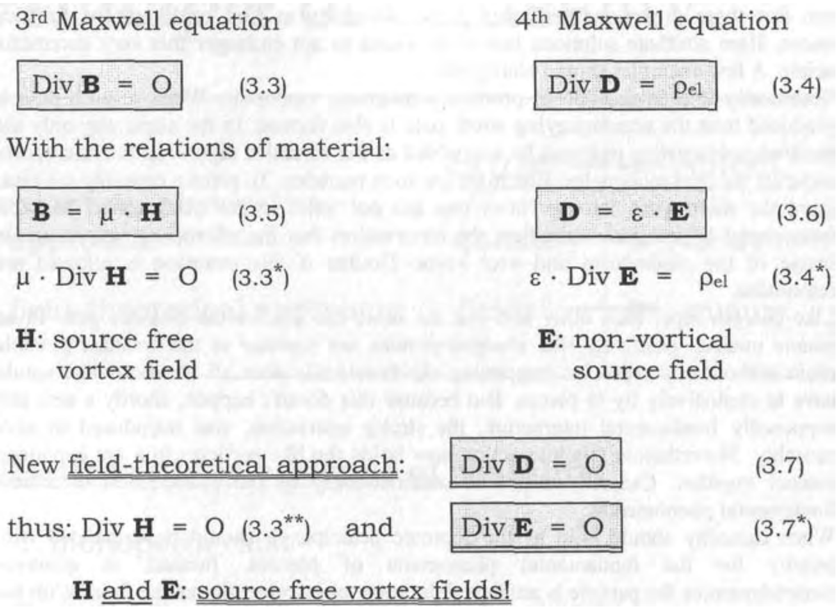

3rd Maxwell equation 4th Maxwell equation

$$\boxed{\text{Div}\,\mathbf{B} = 0} \quad (3.3) \qquad \boxed{\text{Div}\,\mathbf{D} = \rho_{el}} \quad (3.4)$$

With the relations of material:

$$\boxed{\mathbf{B} = \mu \cdot \mathbf{H}} \quad (3.5) \qquad \boxed{\mathbf{D} = \varepsilon \cdot \mathbf{E}} \quad (3.6)$$

$$\mu \cdot \text{Div}\,\mathbf{H} = 0 \quad (3.3^*) \qquad \varepsilon \cdot \text{Div}\,\mathbf{E} = \rho_{el} \quad (3.4^*)$$

H: source free vortex field

E: non-vortical source field

New field-theoretical approach: $\boxed{\text{Div}\,\mathbf{D} = 0} \quad (3.7)$

thus: $\text{Div}\,\mathbf{H} = 0 \quad (3.3^{**})$ and $\boxed{\text{Div}\,\mathbf{E} = 0} \quad (3.7^*)$

H and **E**: source free vortex fields!

Einstein:

"Is it conceivable, that a field theory permits us to understand the atomistic and quantum structure of reality? This question by almost all is answered with No. But I believe that at the moment nobody knows anything reliable about it"<i>.

Pauli:

"The electric elementary quantum e is a stranger in Maxwell-Lorentz' electrodynamics"<ii>.

Fig. 3.2: The field-theoretical approach

<i>: A. Einstein: Grundzuge der Relativitatstheorie, S 162, Anhang II; 5. Aufl., Vieweg, Braunschweig 1974.

<ii>: W. Pauli: Aufsatze und Vortrage uber Physik und Erkenntnistheorie. Vieweg, Braunschweig 1961, entnommen aus:
H. G. Kussner: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder. Musterschmidt-Verlag Gottingen 1976, S. 161.

3.2 Field-theoretical approach

The field-theoretical approach is the very much older one. Until the last turn of the century the world in this respect still was in order. Max Planck, by the discovery of quanta, has plunged physics into a crisis.
Albert Einstein, who, apart from his lightquanta hypothesis, in his soul actually was a field theorist, writes: „Is it feasible that a field theory allows us to understand the atomistic and quantum structure of reality?". This question by almost all is answered with No. But I believe that at present nobody knows anything reliable about it[i]".

By the way the "No" can be justified by the fact that the field description after Maxwell is by no means able to the formation of structure so that it is not possible for quanta to appear as a consequence. The field-theoretical approach could, obstructed by Maxwell's field theory, not further be pursued and to this until today nothing has changed.
Nevertheless it would be an omission to not at least have tried this approach and have it examined for its efficiency. Maybe the above mentioned problems of causality let themselves be solved much more elegantly. For this however the Maxwell theory must be reworked to a pure field theory. With the well-known formulation it offends against the claim of causality, since it is field theory and quantum theory at the same time. To Maxwell himself the quanta were still unknown, but today we know that the fourth Maxwell equation is a quantum equation:

$$\mathrm{div}\,\mathbf{D} = \rho_{el} \quad (3.4)$$

After this the electric field is a source field whereby the individual charge carriers, like e.g. electrons, act as sources to form in their sum the space charge density p_{el}. The other three Maxwell equations are pure wave equations. In this way the third equation identifies the magnetic field as source free:

$$\mathrm{div}\ B = O \quad . \qquad (3.3)$$

This for Pauli probably was the reason to call, "the electric elementary quantum e^- a stranger in Maxwell-Lorentz' electrodynamics"[ii].

Let's return to the principle of causality according to which in the field-theoretical approach the fields should act as a cause and not the particles. In a corresponding field description quanta logically have not lost anything. It is only consistent to likewise demand freedom of sources of the electric field:

$$\mathrm{Div}\ D = O \quad . \qquad (3.7)$$

When the electric field is not a source field, then what is it? The magnetic field is a vortex field. Hence it would be obvious to also conceive the electric field as a vortex field. Numerous reasons speak for it:
1. A non-vortical gradient field, like it is formed by charge carriers, merely represents a special case of the general vortex field. Only by the generation of quanta a source field can form as a special case.
2. The electromagnetic wave teaches us the duality between the E- and the H-field that are directed perpendicular to each other and are in a fixed relation to each other. If one of them is a vortex field then also the dual field must be a vortex field.

Duality:

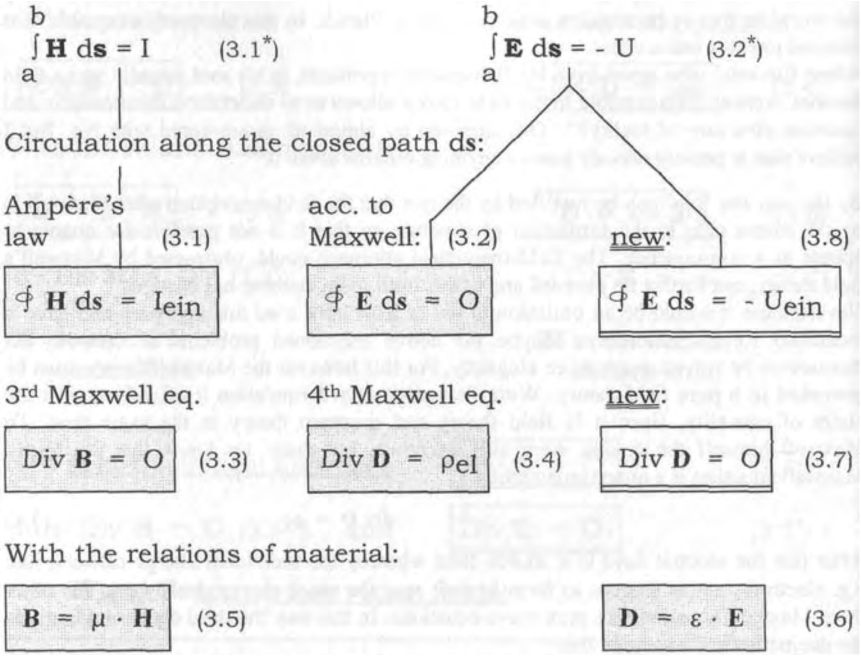
$\int_a^b \mathbf{H}\, ds = I$ (3.1*)
$\int_a^b \mathbf{E}\, ds = -U$ (3.2*)
Circulation along the closed path ds:
Ampère's law (3.1)
acc. to Maxwell: (3.2)
new: (3.8)
$\oint \mathbf{H}\, ds = I_{ein}$
$\oint \mathbf{E}\, ds = O$
$\oint \mathbf{E}\, ds = -U_{ein}$
3rd Maxwell eq.
4th Maxwell eq.
new:
$\mathrm{Div}\, \mathbf{B} = O$ (3.3)
$\mathrm{Div}\, \mathbf{D} = \rho_{el}$ (3.4)
$\mathrm{Div}\, \mathbf{D} = O$ (3.7)
With the relations of material:
$\mathbf{B} = \mu \cdot \mathbf{H}$ (3.5)
$\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

Fig. 3.3: The dual field description

Dual approach according to Jackson[i] or Lehner[ii]:

$\mathrm{Div}\,B = \rho_{magn}$ because of the 4th Maxwell equation: Div D = ρ_{el} (3.4)

$\rho_{magn} =$ magnetic monopoles should exist,
otherwise dual extension (3.8) not allowed!

Caution: closed loop conclusion!
Maxwell theory proves the correctness of the Maxwell theory.
Result: search for magnetic monopoles unsuccessful.

---

[i]: J. D. Jackson, Classical Electrodynamics, 2nd Ed. , John Wiley, New York, 1975, S. 251 - 253

[ii]: G. Lehner, Elektromagnetische Feldtheorie, Springer-Verlag Berlin, Heidelberg 1990, S. 35 - 36 und S. 533 ff.

3.3 Duality

Duality is a fundamental physical principle. Opposite, but one another complementing phenomena can be assigned to each other in pairs, like e.g. (see fig. 8.8):

	Yin (female)	——		Yang (male)
	south pole	——		north pole
	⊖	——		⊕
R [Ω]	resistance	——	R^{-1} [Ω^{-1}]	conductance
U [V]	voltage	——	I [A]	current
$\mathbf{E}$ [V/m]	electr.field strength	——	$\mathbf{H}$ [A/m]	magnet.field strength
Q [As]	charge	——	ϕ [Vs]	magnetic flux
$\mathbf{D}$ [As/m^2]	electr.displacement	——	$\mathbf{B}$ [Vs/m^2]	magnetic induction
ε_0 [As/Vm]	electr.field-constant	——	μ_0 [Vs/Am]	magnet.field-constant
ρ_{el} [As/m^3]	electr.charge density	——	ρ_{magn}[Vs/m^3]	magnet.charge density
	induction law	——		Ampère's law
	potential vortex	——		eddy current
	(vortex of the electric field)			(vortex of the magnetic field)

First of all we find the duality confirmed in the case of the electromagnetic wave spreading in a homogeneous medium. Here the field pointers of E and H are directed perpendicular to each other and are in a fixed relation to each other. But if the wave is damped in the presence of matter, for instance by eddy currents, then by basing on Maxwell's field theory the duality will vanish.

A good example for perfect duality provides the integral of a field strength vector along the path from a to b:

$$\int_a^b \mathbf{H}\, ds = I \qquad (3.1^*) \qquad \text{and} \qquad \int_b^a \mathbf{E}\, ds = U \ . \qquad (3.2^*)$$

Urn if the integration takes place along a closed path then the circulation yields:

$$\oint \mathbf{H}\, ds = I_{eing} \qquad (3.1) \qquad \text{and} \qquad \oint \mathbf{E}\, ds = 0 \ . \qquad (3.2)$$

According to Ampere's law (3.1) the magnetic field can thus form enclosed currents and spatially spreading eddy currents. The electric field on the other hand should be irrotational (3.2).

Let's take the case that the electromagnetic wave is damped by eddy currents and the magnetic field in this way becomes a vortex field. The electric field itself that, as said, is in a fixed relation and perpendicular to the vortex field H, will show all the vortex-typical properties. Hence nothing would be more obvious as to also grant the electric field a formation of vortices:

$$\oint \mathbf{E}\, ds = -U_{eing} \ . \qquad (3.8)$$

Critics of this dual approach, like for instance Jackson[i] or Lehner[ii], point out that with reference to the fourth Maxwell equation the electric field should be understood as a source field:

$$\operatorname{div} \mathbf{D} = \rho_{el} \ . \qquad (3.4)$$

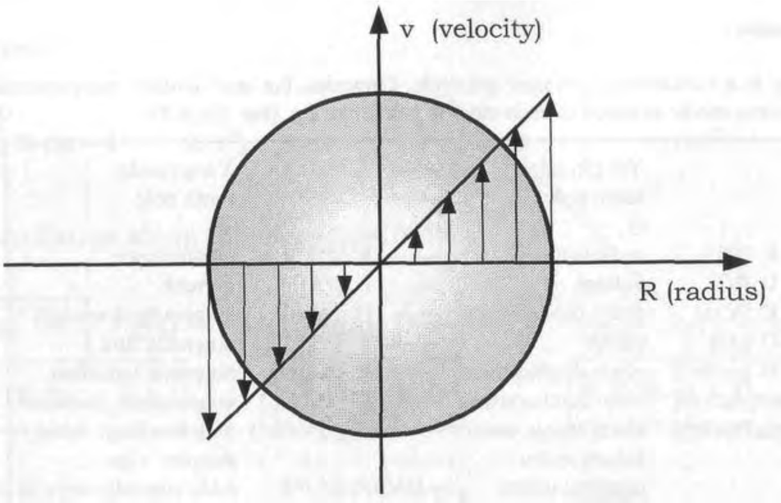
v (velocity)
R (radius)

Fig. 3.4a: Velocity distribution v(R) for a vortex with rigid-body rotation

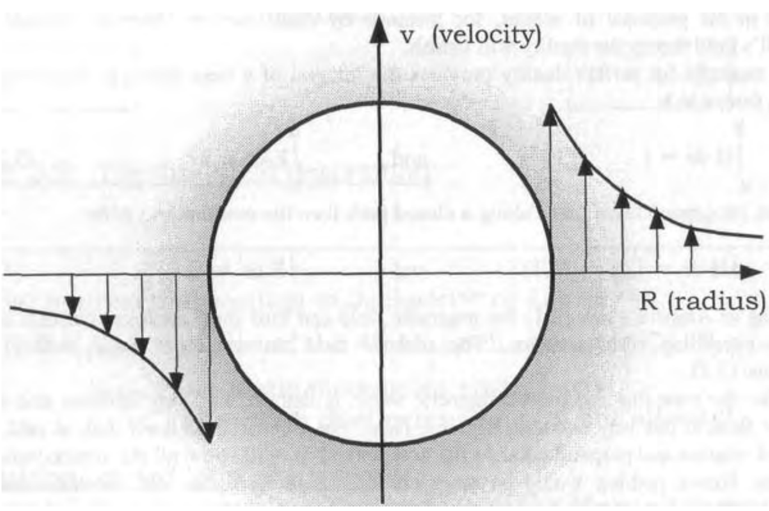
v (velocity)
R (radius)

Fig. 3.4b: Velocity distribution v(R) in a potential vortex (see Lugt[i]).

---

[i] Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 30 and 31, ISBN 0-89464-916-7

For a complete duality from the existence of electric monopoles, individual in the space charge density ρ_{el} contained charge carriers, the claim for magnetic monopoles is derived. In spite of intensive search such north or south pole particles however until now could not be found. Herein from the sight of criticism is seen a confirmation for the assumption that Maxwell's field theory is self-contained and hence in principle may not be extended. The critics have a problem of causality: They postulate source fields that at the same time should be vortex fields. But if one asks how one should imagine such a field that is scalar and at the same time vectorial, then it looks as if no one has ever made any thoughts about it.

The from causality derived solution of the problem of lacking duality requires to extend the Maxwell theory in one point, by introducing the potential vortex of the electric field here and at the same time make a cut in another place:

$$\text{div } D = O \tag{3.7}$$

With this formulation, the assumption of a freedom of sources in principle, the complete duality already is reached: Now neither magnetic nor electric monopoles exist (Fig. 3.3)! At first we have to accept the loss of the electron hoping that the calculation in the end works out: the "exchange" vortices against particles, by which the quanta can be banned from the field theory, suggests that the elementary particles themselves are nothing else as spherical vortices that have found to an own physical reality.

3.4 Flow vortices

In fluid engineering convincing and strong indications for the correctness of the chosen approach can be found. It benefits us that hydrodynamic vortices are visible or can be the injection of smoke, e.g. in a wind-tunnel.
Already Leonardo da Vinci had observed at liquids that there exist two dual basic types of plane vortices: "Among the vortices one is slower at the centre than at the sides, another is faster at the centre than at the sides."
A vortex of the first type, also called "vortex with rigid-body rotation", is formed for instance by a liquid in a centrifuge that due to its inertia of mass is pressed to the edge because there the largest velocity exists. In an analogous way the electromagnetic vortex in electrically conductive material shows the well-known "skin effect" (Fig. 3.4a).
To explain the other vortex Newton describes the experiment where a rod is dipped into a liquid as viscous as possible and then is turned. In this potential vortex the velocity of the particle increases the closer to the rod it is (Fig. 3.4b).
The duality of both vortex phenomena becomes obvious when we make ourselves clear that in the experiment with the centrifuge the more liquid presses to the edge the less viscous the medium is. And that on the other hand the potential vortex forms the stronger the more viscous the medium is. As conclusion we read in text books that the viscosity of the liquid decides whether a vortex with rigid-body rotation or a potential vortex is formed.

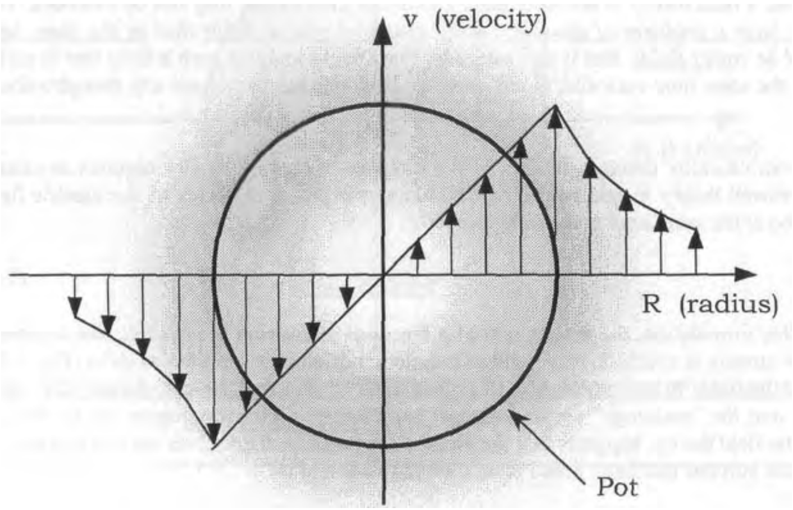

Fig. 3.5: Combination of a vortex with rigid-body rotation and a potential vortex (Lugt<i>).

<i> Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; ISBN 0-89464-916-7

3.5 Rankine vortex

When we, in a third experiment, immerse the centrifuge filled with water into a tough medium and let the centrifuge rotate, then inside the centrifuge a vortex with rigid-body rotation forms and outside the centrifuge a potential vortex forms (Fig. 3.5).

It is obvious that one vortex always causes the other vortex with the opposite properties and so the existence of one causes that of the other. So in the first case, that of the vortex with rigid-body rotation, outside the centrifuge potential vortices will form in the surrounding air, whereas in the second case, that of the potential vortices, the turning rod itself can be interpreted as a special case of a vortex with rigid-body rotation. Hence in all conceivable experiments the condition always is fulfilled that in the centre of the vortex the same state of "peace", that we can fix as "zero", prevails as in infinity.

When we take a tornado as an example, thus a whirlwind. In the "eye of the cyclone" there's no wind at all. But when I go away from this spot, then I'm blown to the outside. I can really feel the vortex with rigid-body rotation in the inside. If. however, I am standing on the outside, then the potential vortex tries to pull me into the vortex. This potential vortex is responsible for the structure and in the end also for the size of the tornado.

At the radius of the vortex, the place with the largest speed of the wind, an equilibrium prevails. The vortex with rigid-body rotation and the potential vortex at this point are equally powerful. Their power again is determined by the viscosity, which thereby fixes the radius of the vortex!

Therefore meteorologists pursue with interest whether a tornado forms over land or over water. Over the ocean for instance it sucks itself full with water. In that way the potential vortex increases in power, the radius of the vortex gets smaller and the energy density increases dangerously.

If the knowledge from hydrodynamics is transferred to the area of electromagnetism, then the role of the viscosity is taken over by the electric conductivity. The well-known current eddy occurs in the conductor, whereas its counterpart, the postulated potential vortex, forms in the bad-conducting medium, with preference in the dielectric. The duality of both vortices is expressed by the fact that the electric conductivity of the medium decides whether current eddies or potential vortices can form and how fast they decay, i.e. convert their energy into heat.

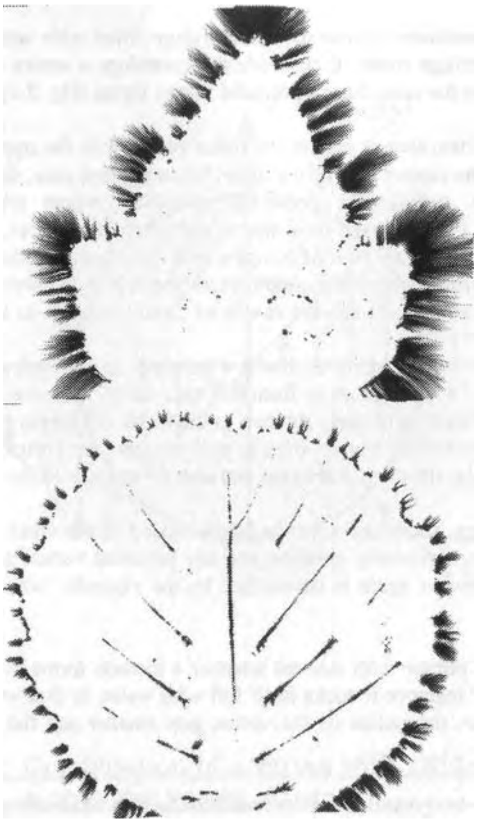

Fig. 3.6: Kirlian photograph of leaves[i] structured corona discharges

---

<i>: (produced by students of electronics in the laboratory for power electronics of the Author, University of Applied Sciences Furtwangen 1991)

<ii>: Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, page 453

<iii>: Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, page 208

3.6 Vortex and anti-vortex

Fig. 3.5 shows that vortex and dual anti-vortex mutually cause each other. In high tension transmission lines we find a striking example for the combination of current eddy and potential vortex. Within the conductor current eddies are formed. Thus the current density increases towards the surface of the conductor (skin effect). Outside of the conductor, in the air, the alternating fields find a very bad conducting medium. If one follows the text book opinion, then the field outside the conductor should be an irrotational gradient field! But this statement causes unsolvable problems.

When vortices occur inside the conductor, then for reasons of a detachment of the vortices without jumps at the interface to the dielectric, also the fields in the air surrounding the conductor must have the form and the properties of vortices. Nothing would be more obvious as to also mathematically describe and interpret these so-called gradient fields as vortex fields. When looking exact this argument even is mandatory!

The as laws of field refraction known boundary conditions[ii] in addition demand steadiness at the interface of the conductor to the dielectric and don't leave us any other choice. If there is a vortex field on one side, then also the field on the other side is a vortex field, otherwise we offend against the law! Here an obvious failure of the Maxwell theory is present.
Outside the conductor, in the air, where the alternating fields find a very bad conducting medium the potential vortex not only exists theoretical; it even shows itself. Dependent among others on the frequency and the composition of the surface of the conductor, the potenial vortices form around the conductor. When the thereby induced potentials exceed the initial voltage, then impact ionisation takes place and the well-known corona discharge is produced. Everyone of us can hear this as crackling and see the sparkling skin with which high tension transmission lines cover themselves.
In accordance with the text books also a gradient field increases towards the surface of the conductor, but an even shining would be expected and not a crackling. Without potential vortices the observable structure of the corona would remain an unsolved phenomenon of physics (Fig. 3.6).

But even without knowing the structure-shaping property of the potential vortices, that in addition acts supporting and that we'll have to derive, it can be observed well that especially roughness on the surface of the conductor stimulate the formation of vortices and can produce vortices. If one is looking for a reason why with large frequency the very short impulses of discharge always emerge from roughness[iii], then very probable potential vortices are responsible for it. By means of a Kirlian photograph it can be shown that the corona consists of structured separate discharges (Fig. 3.6).

With this the approach is motivated, formulated and given reasons for. The expositions can't replace a proof, but they should stand a critical examination. Mathematical and physical evidence will be furnished later.

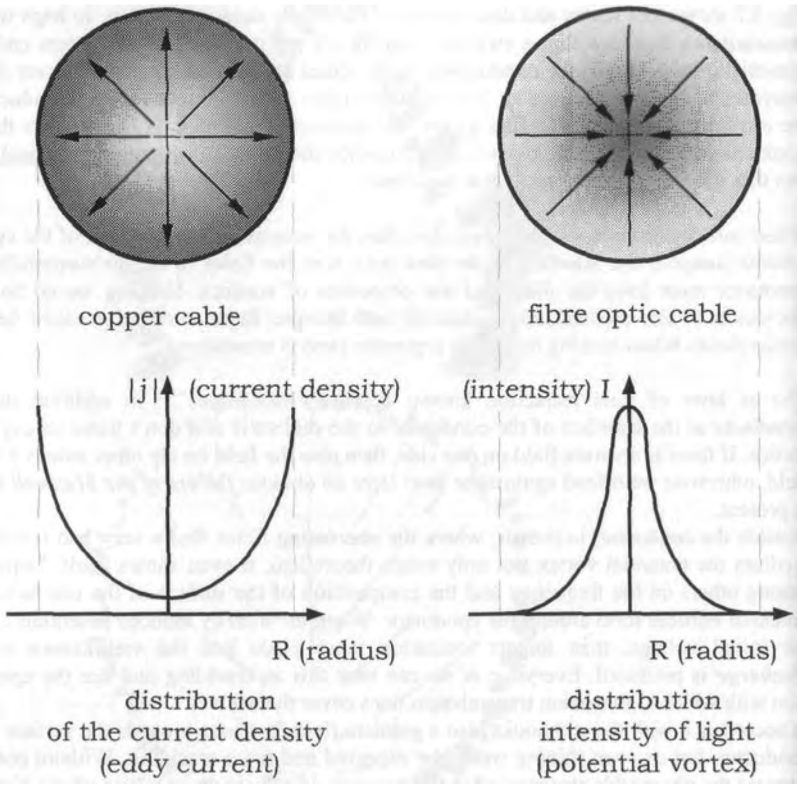
copper cable
fibre optic cable
|j| (current density)
(intensity) I
R (radius)
R (radius)
distribution
of the current density
(eddy current)
distribution
intensity of light
(potential vortex)

Fig. 4.1: The distribution in principle of the intensity of light within a fibre optic cable<i> compared to the distribution of the current density in a copper cable

---

<i> Meyl, Konstantin: Potentialwirbel, Band 1: Diskussionsbeitrage zur naturwissenschaftlichen Deutung und zur physikalisch-technischen Nutzung, basierend auf einer mathematischen Berechnung neu entdeckter hydrotischer Wirbel, INDEL GmbH, Verlagsabteilung, Villingen-Schwenningen 1990, ISBN 3-9802-542-1-6

4. Properties

4.1 Concentration effect

It can be assumed that until now there does not yet exist a technical application for the here presented potential vortex theory unless the phenomenon was used by chance and unconsciously. About this the transmission of optical light signals over fibre optic cable can be given as a typical example.
Compared to a transmission of energy impulses over a copper cable fibre optic cables show a considerable better degree of effectiveness. The derived potential vortex theory provides a conclusive explanation for this phenomenon and therefore is put here to discussion: If we cut through a fibre optic cable and look at the distribution of a light impulse over the cross-section, then we observe a concentration in the centre of the conductor (fig. 4.1).

Herc the duality between the vortices of the magnetic and of the electric field comes to light. Whereas the current eddies in a copper conductor cause the "skin effect" as is well-known, potential vortices show a "concentration effect" and align themselves with the vortex centre. The measurable and in fig. 4.1 shown distribution of the light intensity in a fibre optic cable may confirm this phenomenon, the orientation of the potential vortex on the vortex centre.

For instance the calculation of the resistance of a copper cable provides as an important result an apparent decrease of the resistance directed towards the conductor surface. There the associated better conductivity as a consequence causes an increased current density. In the reversed direction, towards the centre of the conductor, consequently a decrease of the effective conductivity would be present, and this result is independent of the used material. According to the rules of duality this is a condition for the formation of potential vortices. As already said the conductivity is responsible for it, if the expanding eddy current with its skin effect or the contracting potential vortex with its concentration effect is predominant.

Usual fibre optic materials possess not only a small conductivity, but in addition a high dieletricity. This additionally favours the formation of vortices of the electric field. If one consciously or unconsciously supports the potential vortices, then there is a possibility that the life of the fibre optic cable is negatively influenced because of the concentration effect. Of course it can not be excluded that other effects, like e.g. reflections or the modes of the light are involved in the concentration effect. But it should be guaranteed that this actually concerns is causal phenomena and doesn't concern only alternative explanations out of ignorance of the active vortex phenomenon.

The formal mathematical reason for the concentration effect provides the reverse sign in Faraday's law of induction compared to Ampere's law (see also equation 3.1 and equation 3.8 in fig. 3.3).

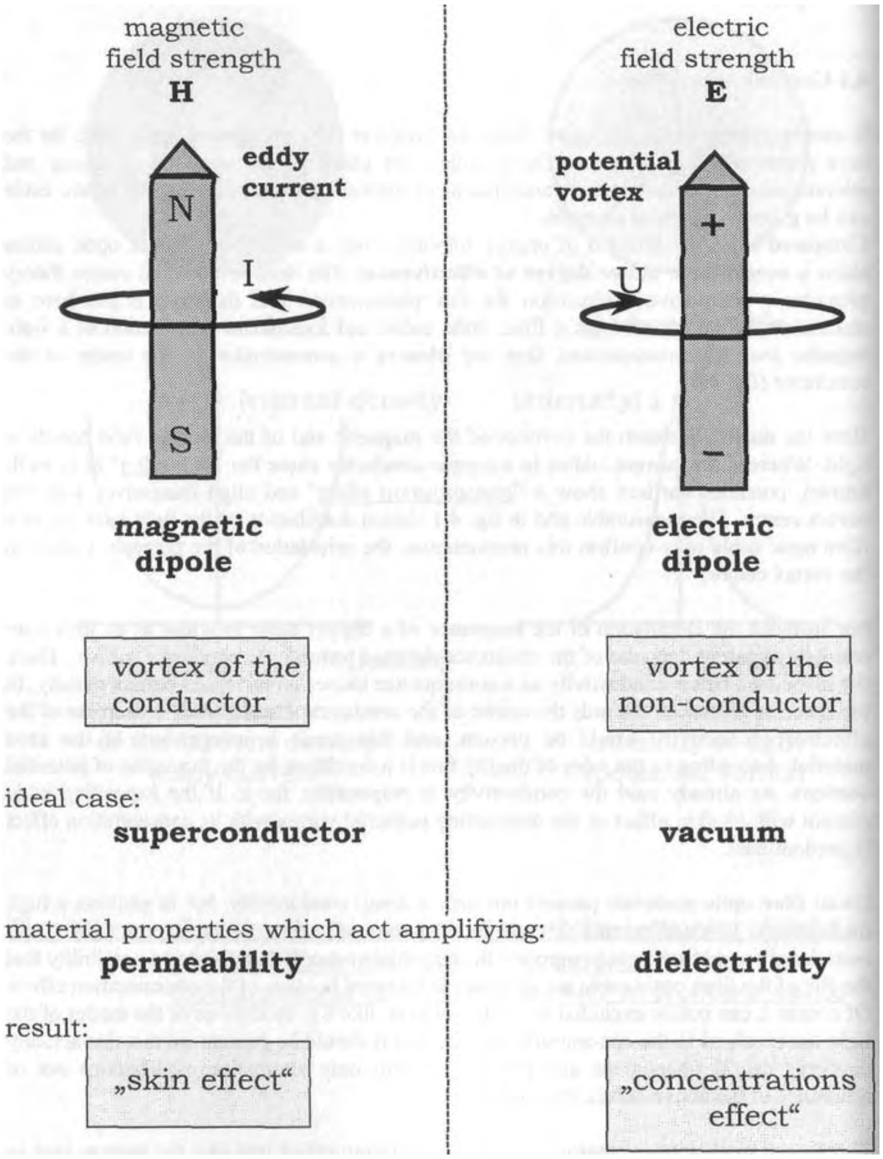
magnetic
field strength
H
eddy current
N
I
S
magnetic dipole
vortex of the conductor
electric
field strength
E
potential vortex
+
U
–
electric dipole
vortex of the non-conductor
ideal case:
superconductor
vacuum
material properties which act amplifying:
permeability
dielectricity
result:
„skin effect“
„concentrations effect“

Fig. 4.2: The acting as a dipole of current eddies and potential vortices

4.2 Duality of the vortex properties

The rules of duality dictate for the vortex of the electric and of the magnetic field the following characteristics:

1. Whereas currents and eddy currents demand a good conductivity, potentials and potential vortices can only form with bad conductivity, thus in a dielectric and best in the vacuum.

2. Eddy currents run apart, strive towards infinity and thus show the well-known "skin effect" with a spatially limited arrangement of the conductor. According to the rules of duality the potential vortex will strive towards the vortex centre and in this way will show a "concentration effect".

3. Another property of vortices is shown in fig. 4.2.
On the left side a plane eddy current is indicated. Since the discovery of Ampere's law it is well-known to us that such a circular current (I) forms a magnetic dipole standing perpendicular to the vortex plane.
On the right hand side the dual phenomenon is sketched. Here charges are piled up circularly to a planar potential vortex (U). Thereby an electric dipole forms, standing perpendicular to the vortex plane. This relation directly follows from the equations of the field-theoretical approach.
Whereas circular currents and current eddies produce magnetic dipoles, the postulated potential vortices will form electric dipoles.

With these three interesting properties some key questions of quantum physics, that until now have stayed a mystery to science (fig. 4.4), can be answered conclusively and without compulsion e.g.:

I.Why are there no magnetically charged particles?
The better the conductivity of a medium is, the higher as a consequence the number of free charge carriers is. the more strongly eddy currents are formed. The answer to question I is inferred from the opposite case:
In the ideal vacuum no charge carriers at all are present, why no currents, no current eddies and consequently no magnetic poles can exist.
With this well-known fact the first question already is answered. The question why in the microcosm there can not exist magnetically charged elementary particles, why the search for magnetic monopoles doesn't make any sense. Let's ask further:

II. Why are there only electrically charged particles?
Let us for that consider the dual conditions. The worse the conductivity of a medium is, the more the potential vortex -will be favoured that because of this property also can be understood as the vortex of the dielectric.
In the mentioned extreme case of the ideal vacuum, no electric conductivity is present for reason of the missing charge carriers. But this circumstance favours the potential vortex and that, according to fig. 4.2, forms electric poles and with this also the second question would be answered clearly.
It can be traced back to the boundary conditions of the microcosm that without exception electricallv charged particles are entitled to exist; a realization derived from the field-theoretical approach, that covers all experiences.

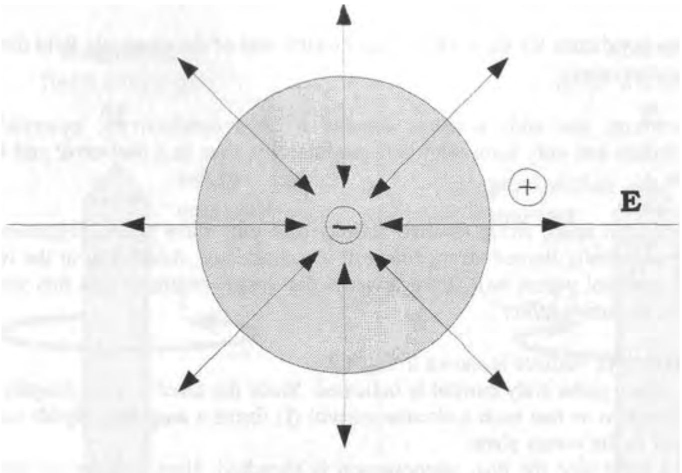

the positron e^+

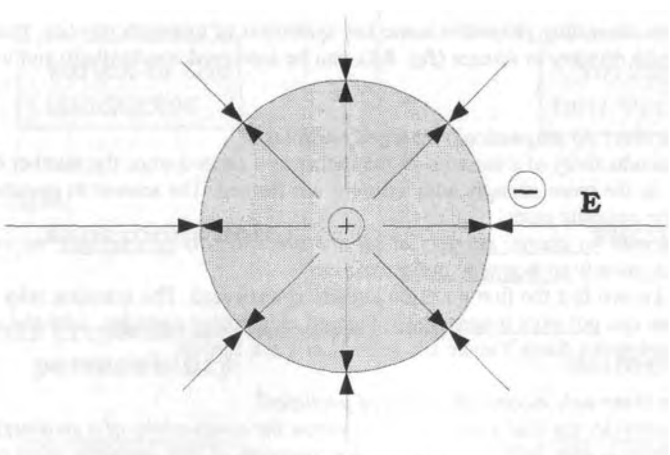

the elektron e^-

Fig. 4.3: Elementary particles as configurations of potential vortices

4.3 Derivation of the electron as an elementary vortex

The next key question necessarily has to be brought to a conclusive answer to save the principle of causality, so that we no longer have to postulate an own physics with its own laws for the microcosm:

III. Why do these particles show as monopoles?

More concrete the question has to read:
Where is the positive pole in a negatively charged electron, if it should be an electric dipole?
The only possible answer is:
In the centre of the particle!

Thus in the centre of the electron its positive pole is hidden and in the centre of the positron its negative pole is hidden. But we only observe these particles from the outside and for reason of the field conditions of the electron we measure a negative charge and for its antiparticle, the positron, a positive charge. If in each case we wanted to measure the electric fields included in the inside, we had to destroy the particle. Then a proof would not be possible anymore.

Here also a persistent mistake is eliminated by the for a long time known axiom that monopoles can not exist at all if one considers continuity! By means of technical-physical experiments this axiom is sufficiently secured.
The quantum physical approach is standing on uncertain ground if it is postulated that other laws of nature should apply to particle physics, if a second approach, the field-theoretical approach, is conceivable that does not know these problems!

The discussed concentration effect gives the potential vortex a structure shaping property. With that also the fourth key question can be answered:

IV. Why do the particles have the form of spheres?

The potential vortex is favoured in the particle-free vacuum of the microcosm because of the missing conductivity. In connection with the concentration effect the following conclusion can be drawn:
The extremely mighty potential vortex exerts a high pressure on the microcosm and on each particle.

With that also the fourth key question, why stable elementary particles are spherical, can be answered by the potential vortex theory:
Only the sphere is able to withstand a high outside pressure.
All other forms, like e.g. dipoles formed like a rod or a club would be instable in the presence of the extremely concentrating potential vortex. They would be immediately destroyed by the pressure of the potential vortex.

I. Why are there no magnetically charged particles?
(the vacuum has no conductivity!)

II. Why are there only electrically charged particles?
(in the vacuum only potential vortices can exist!)

III. Why do these particles show as monopoles?
(the other pole is locked up in the inside of the vortex oscillation!)

IV. Why do the particles have the form of spheres?
(for reason of the outside pressure by the concentration effect!)

V. Why is the elementary quantum stable?
(without conductivity no decay of vortices takes place!)

VI. Why does for every particle of matter exist an antiparticle?
(there are two swirl directions with equal rights!)

VII. Why are particles and antiparticles incompatible?
(contrary swirl directions!)

Fig. 4.4: Key questions of quantum physics

4.4 Quanta as field vortices

The fied-theoretical approach demands removing the electron from the field equations (eq. 3.7) and at the same time introducing the potential vortex of the electric field. With this vortex phenomenon there now is a possibility that the electromagnetic wave spontaneously rolls up to a vortex in case it is disturbed from the outside. The vortex particle that is formed in such a way owes its physical reality on the one hand the concentration effect of the potential vortex, that compresses this particle to the dimension of a tiny sphere and on the other hand its localization for reason of the oscillation around a fixed point.

The sphcrical elementary particles are being compressed to inconceivably small dimensions. Therefore they are capable to bind a comparatively high energy in their inside. This is confirmed by the mass-energy relation $E = mc^2$. (4.1)

The fact that the energy is dependent on the speed of light can be judged to be a clear indication that quanta actually are nothing but oscillating electromagnetic packages, vortical oscillations of empty space!

The next question reads:

V. Why is the elementary quantum stable?

The worse the conductivity is, the more the potential vortex will be favoured, the more strongly the concentration effect will form, the smaller the spherical phenomena will get - the larger the authoritative relaxation time will be, i.e. the slower the decay of vortices and with that the more stable the vortex phenomenon will be.

In the microcosm, that comes the ideal case of a particle-free vacuum very close, the spherical vortices because of the missing conductivity have an absolute stability.

VI. Why does for every particle of matter exist an antiparticle?

Since every vortex can also oscillate in the opposite direction, there always exist two forms of formation of spherical vortices with equal rights, one of them is assigned to the world of matter and the other to the world of anti-matter.

VII. Why are particles and antiparticles incompatible?

For reason of the contrary swirl direction they are incompatible to each other. They have the tendency to destroy each other mutually, like two trains that want in the opposite direction on a single-tracked distance.

The quantum physical approach does not have an answer to these key questions. Until now scientists have merely thought about to what the observable contraction in the microcosm and the macrocosm can be traced back. Because the approach was not able to furnish an answer, without further ado some new matter was introduced: the sluons. These binding particles should be able to exert the necessary pressure. But until now no one has been able to observe or detect this fabulous matter. Nobody knows its structure and its composition. Despite missing evidence it is stated that this matter is mass less and at the same time lumped; it is invisible because it can't interact with any other matter, not even with the supposed building parts of the atomic nuclei, the quarks. But at the same time there should be exerted a pressure on the quarks, for which reason quarks again should be able to interact with gluons!

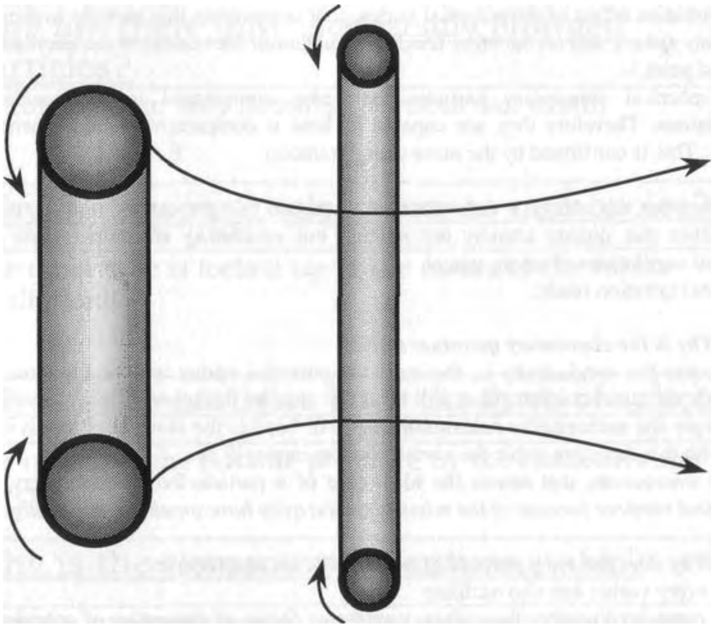

Fig. 4.5: Two coaxial oscillating vortex rings (Lugt[<i>]).

<i> Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 42, ISBN 0-89464-916-7

4.5 The photon

The ability to form structures as a consequence of the concentration effect gives the potential vortex a number of highly interesting properties. To derive these properties we can make work easier when we fall back upon the observations and experiences of flow dynamics.

Here the vortex ring takes a special place. Its vortex centre is not closed, for which reason it is not stationary and propagates in space with a constant speed. It can be observed that the velocity of propagation increases with the ring diameter becoming smaller. By means of the vortex rings, that skilful smokers can produce with pointed lips, these properties can be made visible.

Now if two vortex rings run into each other with the same axis and direction of rotation then both oscillate around each other, by one vortex attracting the other vortex, thereby accelerating and thus contracting it. The second vortex then slips through the core opening and gets again slower and larger. Now the first vortex accelerates and plays the same game (fig. 4.5).

It would be obvious for the vortex of the electric field to have a corresponding property. The electron e- and with the opposite swirl direction the positron e+ will form such a potential vortex corresponding to the derivation. Two electrons, as like charged particles, would repel each other and surely will be out of the question for such a configuration. An electron and a positron however will attract each other and because of their incompatibility they will mutually destroy unless they open their vortex centres to form a vortex ring. Now the e- shows its positively charged centre that shows the same swirl direction as the e+ seen from the outside. Therefore the vortices don't hurt each other, when the positron slips through the opened vortex centre of the electron and vice versa.

This oscillating electron-positron pair has strange properties: seen from the outside one moment it is negatively charged and the next moment it is positively charged. Therefore over time on the average no charge will be measurable and no electromagnetic interaction will take place.

One moment the particle is matter and the next moment it is anti-matter. Hence no mass at all can be attributed to the particle. Interactions primarily takes place between both dual vortices. We can predict, the particle has neither mass nor charge. The environment merely sees a fast oscillating particle that only within every half cycle is capable of an interaction.

The centre of the oscillating particle is open, for which reason it is not stationary anymore. Instead it propagates in z-direction with the swirl velocity, which is the speed of light, in this way preventing a rotation around the x- or y- axis (fig. 4.6). In this way a polarizability is present.

The only possible and, as we will see, necessarily taking place rotation around the z-axis gives the particle a spin of the magnitude of a quantum of angular momentum. After all the rotation for e^- and e^+ is of the same magnitude with a spin of each time $\frac{1}{2} \cdot \hbar$.There should be paid attention to the fact that for the case of an opposite sense of direction of the respective rotation around the common z-axis the spin on the average will be zero.

In addition the particle is characterized by an outstanding property: a periodically taking place oscillation with any frequency, but that frequency has to be constant. We now only have to take a table of particles to hand. Actually we will find a corresponding particle that has all these properties: the γ-quanta,also called photon.

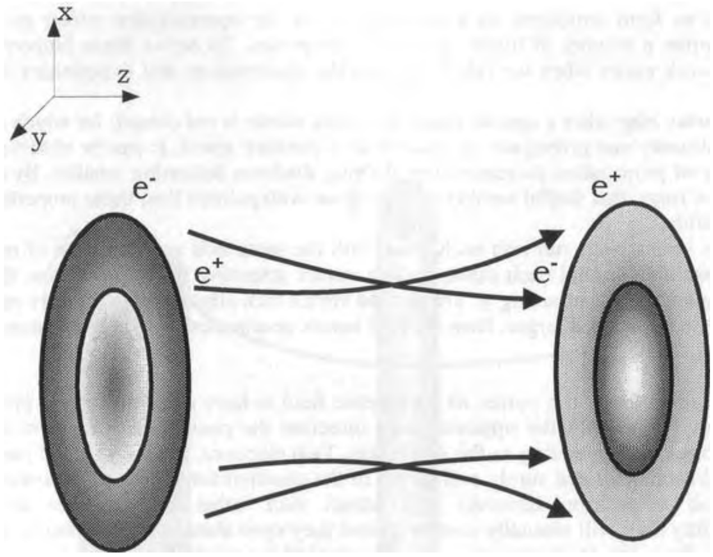

Fig. 4.6: The photon as oscillating electron-positron pair

The decay of the γ- quanta (photon)[i]
(= pair creation = Bethe-Heitler-process 1934):

$$\boxed{\gamma \longrightarrow e^- + e^+}$$

(4.2)

---

[i] Nachtmann, Otto: Phanomene und Konzepte der Elementarteilchenphysik, Vieweg, Braunschweig 1986, S. 135, ISBN 3-528-08926-1

4.6 Pair creation

Proof for the correctness of the model concept provides the decay of the photon in an electron and a positron in the presence of a strong field, as for instance in an atomic nucleus. This observable decay is called pair creation or Bethe-Heitler process<i>:

$\gamma \longrightarrow e^- + e^+$ (4.2)

In this process the elementary vortices for a short time get back their localization and are therefore detectable. Otherwise the electron and positron have the form of a sphere, the photon however rather has the form of two oscillating discs.

The photon doesn't participate in the electromagnetic interaction, because the electric field lines run from one disc to the other (from + to -). The field lines are not open as they are for e^- or e^+ (fig. 4.3). To open up the field lines an energy is necessary that corresponds to the sum of the two formed particles. But from this it by no means follows that this amount of energy will be released in the reversed and much better known process, where matter and anti-matter annihilate under emission of γ-quanta. At the end of the derivation the vortex model will provide us the desired answers on questions of the energy of photons.

Here first of all only the properties will be concerned. Experiments, in which light shows as a particle, are the photoelectric effect, the Compton effect and a lot more. According to the by Maxwell developed classical theory of light however is light an electromagnetic wave that is not quantized in any way, neither as sphere nor as disc, the wave nature of light as well has a physical reality and is secured by experiment. This is witnessed by the interference patterns of overlapping coherent light beams.

A concept in which light could exist at the same place and the same time both as wave and as corpuscle could never be brought into accord with the principle of causality. Formulas of compromise, like the uncertainty principle of Heisenberg that refers to the point of view of the observer, can't change anything about this dilemma. The dual nature of light, that in this context is gladly spoken of, rather consists of the fact that dependent on the local field conditions, any time and spontaneously the wave can roll up to a vortex.

As an example of a violation of the principle of causality it has been indicated under point 3 (fig. 3.1) that both fields and quanta at the same time should be the cause of something. This concept was formulated by Maxwell and written down in modern quantum electrodynamics by Dirac but in the field-theoretical approach we have dropped this concept because it violates all rules of causality in an elementary manner. Therefore it only is consistent, if we hold the view that the light is either wave or particle but never is both at the same time!

In the spontaneous transition of the wave to the particle all the important properties are conserved: the propagation with the speed of light, the characteristic frequency of the oscillation and the mentioned polarizability.

The process of rolling up possibly takes place already in the laboratory, in a bubble chamber and at the latest in our eyes. To receive the electromagnetic wave, we had to have antennas. We actually see the photons. It therefore would be obvious if our cells to see only could perceive vortices, in this case photons. We don't possess a sense organ for fields and waves.

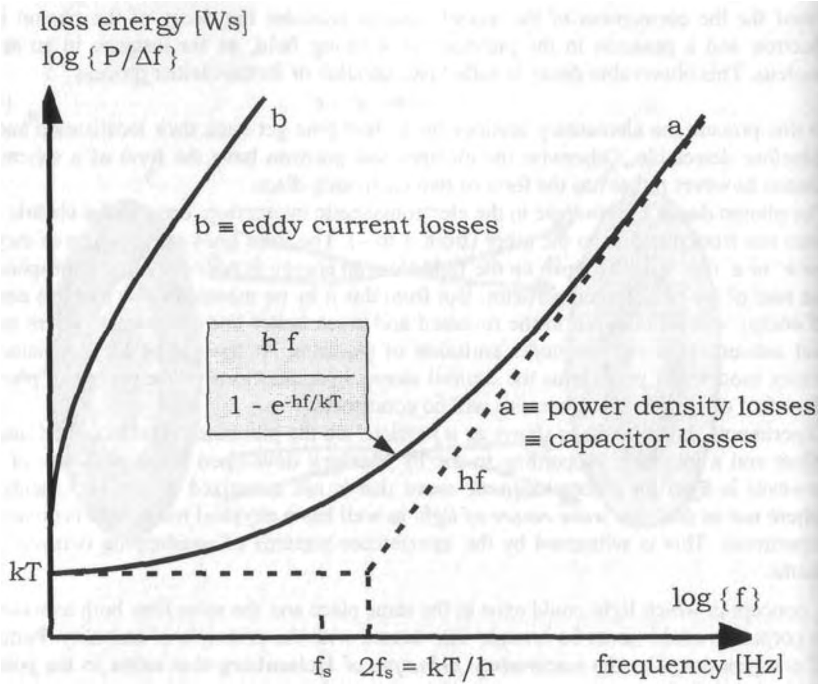

Fig. 4.7: The power density shown against frequency for noise
(a) according to Kupfmuller[i],
as well as for dielectric losses of a capacitor (also a) and for eddy current losses (b) according to Mevl[ii]
(b in visible duality to a).

[i] Kupfmuller, Karl: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, ISBN 3-540-18403-1

[ii] Meyl, Konstantin: Dreidimensionale nichtlineare Berechnung von Wirbelstromkupplungen, Dissertation Universitat Stuttgart 1984

4.7 Noise

If, according to the field-theoretical approach, there exist electric field vortices then they will not only form the elementary particles in the vacuum, but will also macroscopically form and have an effect in larger dimensions.

Assuming a wave that rolls up to a plane vortex it would be obvious if polarization and velocity of propagation are conserved in the process. But how does it stand about the frequency?
The wave now will walk round a stationary point, the vortex centre. The propagation with the speed of light c will remain existent as the swirl velocity. For a plane circular vortex, where the path for a revolution on the outside is very much longer than near the vortex centre, there arises a longer wave length and as a consequence a lower frequency on the outside as more to the inside. With this property the vortex proves to be a changer of frequency: the vortex transforms the frequency of the causing wave in an evenly spectrum, that starts at low frequencies and stretches to very high frequencies (fig. 1.4). Exactly this property we observe in "white noise". The consistent conclusion would be that this concerns the vortex of the electric field. Anyone can, without big expenses, convince himself or herself of the localization, of the property to change the frequency and of the circumstance that vortices can be very easily influenced, that they avoid or again whirl about a place of disturbance, for instance an antenna. For that one only needs to tune a radio receiver to a weak and noisy station and move oneself or some objects around, then one is able to directly study the influences from the manipulation of the receiving signal.
But already the fact that the using and measurability of signals is limited by noise makes clear. which attention the potential vortex should be given.

Within a limited frequency range the power of the Nyquist or resistance noise is independent of frequency (fig. 4.7). This should be clarified particularly by the term "white noise" analogous to white light, where all visible spectral ranges independent of frequency have the same energy density.

But this relation doesn't hold for high frequencies of any magnitude. Here another noise-effect appears, that is said to have its cause in the quantum structure of energy<i>. Untouched by possible interpretations an increasing power of the noise is measured, that more and more turns into a being proportional to frequency (fig. 4.7, curve a). Interestingly this curve shows a remarkable duality to the power curve of eddy currents, likewise shown against the frequency, that for instance can be measured at eddy current couplings<ii> (fig. 4.7, curve b). This circumstance suggests a dual relation of the potential vortex of the electric field in bad conducting media on the one hand and the eddy current in inductive materials on the other hand <iii>.

<iii>: Meyl, Konstantin: Wirbel des elektrischen Feldes,
EMC Journal 1/95, 6. J, ISSN 0945-3857, S. 56 - 59.

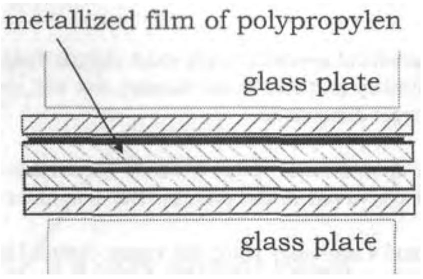
metallized film of polypropylen
glass plate
glass plate

a) measurement set up according to Yializis and others<i>

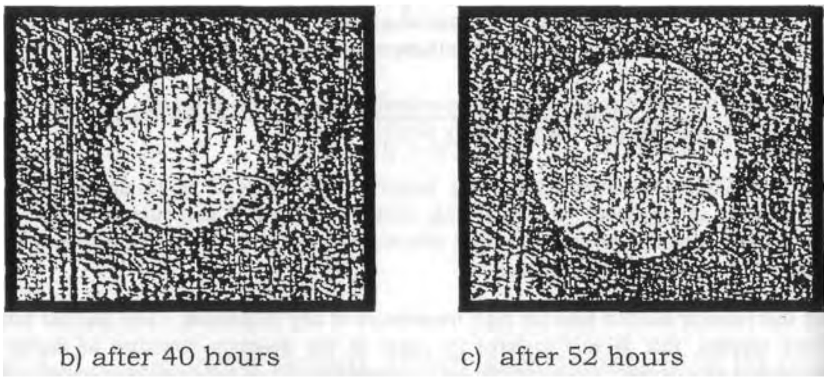
b) after 40 hours c) after 52 hours

Fig. 4.8: Measurement set up (a) and photo of vortex structure in a metallized polypropylen layer capacitor (at 450 V/ 60 Hz/ 100°C)
Observation of the formation of a vortex (b) and (c). (110 fold magnification), according to Yializis et al.<i>

<i>: A. Yializis, S. W. Cichanowski, D. G. Shaw: Electrode Corrosion in Metallized Polypropylene Capacitors, Proceedings of IEEE, International Symposium on Electrical Insulation, Bosten, Mass., June 1980;

4.8 Capacitor losses

Next the dielectric losses in a capacitor fed with an alternating voltage are measured and likewise put on against the frequency. At first the course is independent of the frequency, but towards higher frequencies it increases and shows the same characteristic course of the curve as the before mentioned power of the noise (fig. 4.7, curve a).

This excellent agreement suggests the assumption that the dielectric losses are nothing but eddy losses.
These vortex phenomena, caused by time-varying fields, are not only found in ferromagnetic and conductive materials, but equally as dual phenomena in dielectrics and non-conductors.

As examples of practical applications the induction welding or the microwave oven can be mentioned. The process can be described in other words as follows: in both examples the cause is posed by high-frequency alternating fields that are irradiated into a dielectric as an electromagnetic wave, there roll up to potential vortices and eventually decay in the vortex centre. The desired and used thermal effect arises during this diffusion process.

The striving in the direction of the vortex centre gives the potential vortex of the electric field a structure shaping property. As a consequence of this "concentration effect" circular vortex structures are to be expected, comparable to the visible vortices in flow dynamics (e.g. tornados and whirlwinds). At the same time the dual anti-vortex arises, the diverging eddy current. It takes, as is well-known, the given structure of the conductor, so in the technical literature one correspondingly talks of a "skin effect". Now if conductor and non-conductor meet as they do in a capacitor, then at the boundary area visible structures will form. Circles would be expected, if the eddy current in the inside and striving to the outside is equally powerful as the potential vortex that comes from the outside and concentrates.

Actually such circular structures are observed on the aluminium of high tension capacitors, when they were in operation for a longer period of time. The formation of these circles, the cause of which until now is considered to be unsolved, is already experimentally investigated and discussed on an international level by scientists (fig. 4.8) [i, ii].

These circular vortex structures can be seen as a visible proof [iii] for the existence of potential vortices of the electric field.

<ii>: D. F. Taylor, On the Mechanism of Aluminium Corrosion in Metallized Film Capacitors, IEEE Transactions on EI-19,August 1984,No.4,pp.288-293.

<iii>: Meyl, Konstantin: Wirbel des elektrischen Feldes, EMC Journal 1/95, 6.J, ISSN 0945-3857, S. 56-59.

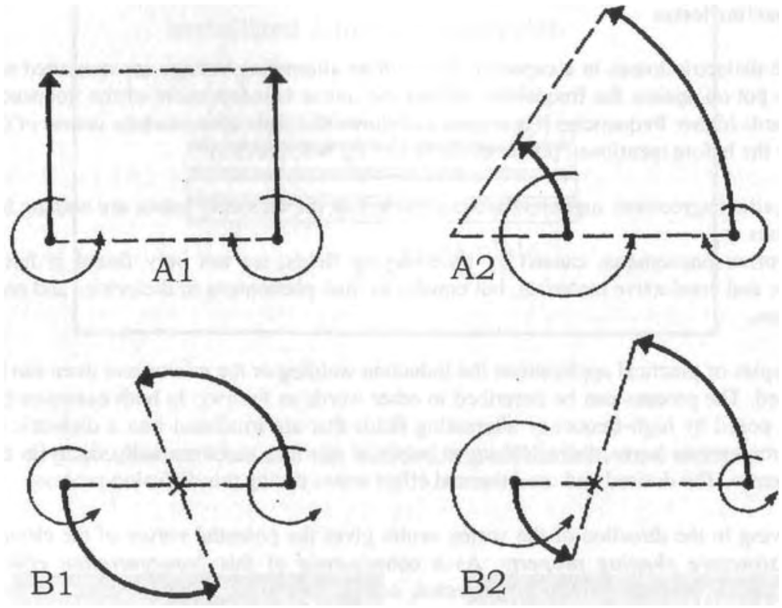

Fig, 4.9: Motion of two point vortices. (Lugt<i>)
A. with opposite direction of rotation
B. with the same direction of rotation
1. for equal vortex strength
2. for unequal vortex strength

Example from practice for case A1:
vortex pair behind an airplane

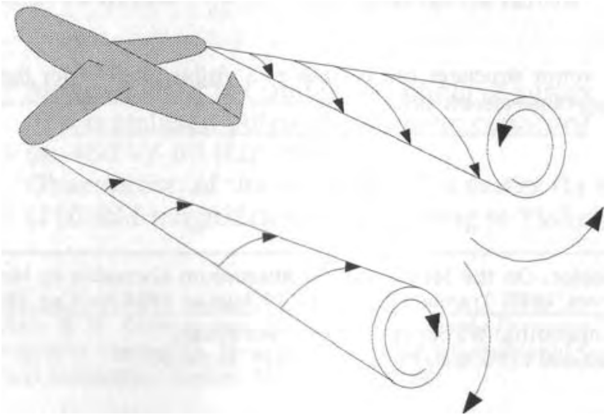

4.9 Vortex lines and vortex streets

It can be assumed that the vortex of the electric field is relevant with regard to the electromagnetic environmental compatibility. This then holds not only for microcosmic and microscopic vortices, but also for macroscopic and larger dimensions. The individual vortices can join together to balls and lines. For the study of this process it is useful to again fall back upon experiences of flow dynamics.

The co-operation of individual point vortices has been investigated thoroughly in flow dyamics. Without any outside manipulation an individual vortex rotates on the spot. That changes in the case of two neighbouring vortices. Now it depends on their mutual strength and sense of rotation. If they have the opposite sense of rotation and equal strength then their centres of rotation move straight forward in the same direction. If however the direction of rotation is the same then both vortices rotate around each other (fig. 4.9).

In this way a multitude of point vortices is capable, to form in the first case whole vortex streets and in the second case spherical vortex balls. In principle a vortex string can also consist of a multitude of potential vortices pointing in the same direction; but it has the tendency to roll up to a vortex ball in case it is disturbed from the outside, as can be shown very clear by means of computer simulations<i>.

As a starting-point for a discussion the thesis can be put forward that also electric field vortices, in nature usually consisting of a multitude of individual point vortices, appear as vortex strings and vortex balls.

Perhaps historians see in this property an answer to the question, how it was possible for the Romans to build streets straight as a die in the wilderness. Their land surveyors, the Augures, had at their disposal neither laser, nor any other suitable gauges. Their most important tool was the Lituus, the crook, that at its upper end was rolled up like a vortex in the sense of a flat coil shaped like a spiral.
The question poses what this strange object was used for. Perhaps the roman land surveors tracked down any vortex lines with this crook and then used them to orientate themselves?
History still holds a lot of secrets, but for now only this indication is given. The following seminar will give enough opportunities for speculations and discussions<ii>.

<i>: Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; p.38, fig. 3.25, 3.26 and 3.27, ISBN 0-89464-916-7

<ii>: K. Meyl: Elektromagnetische Umweltvertraglichkeit Teil 2 und 3 Seminar-umdrucke, INDEL Verlag VS; see part 2 and 3 of this book.

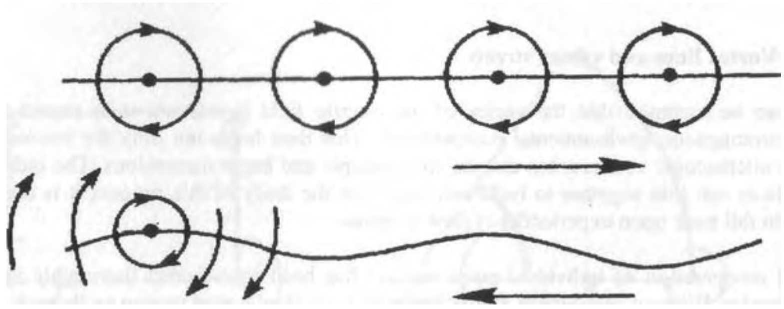

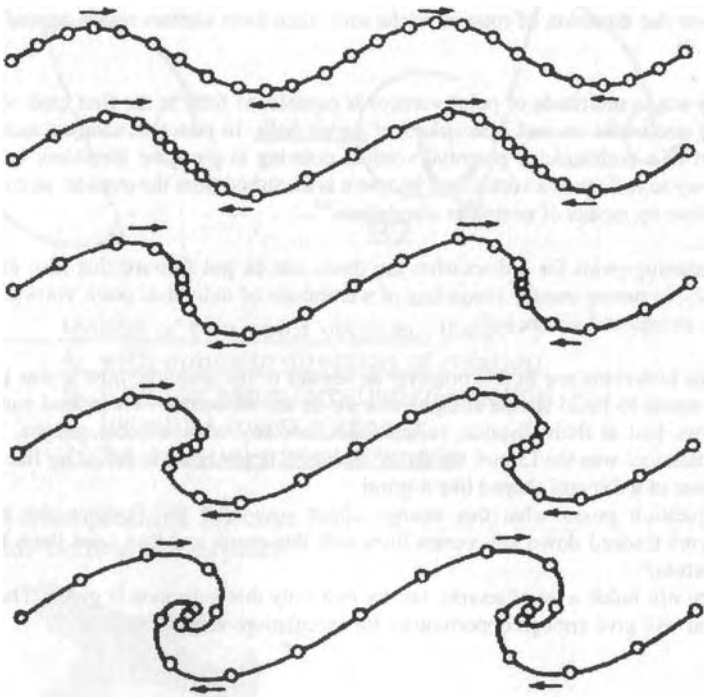

Fig. 4.10: The rolling up of a vortex chain to a vortex ball for the smallest disturbance
(case Bl in fig. 4.9) (Rosenhead<i>).

<i>: L. Rosenhead: Formation of vortices from a surface of discontinuity. Proc. Roy. Soc. A 134, 1931, 170. taken from:

Lugt, Hans J.: Vortex flow in nature and technology. Krieger publishing company, Florida 1995; page 39, figure 3.29, ISBN 0-89464-916-7

4.10 Water colloids as vortex balls

We have to realize that in the biosphere we are staying in a relatively ideal dielectric. The two "capacitor plates" are being formed by the ionosphere and the earth. The potential vortex will, as said, be favoured by a bad conductivity and by a high dielectricity. Consequently it will dominate and take effect in the biosphere. In which way it takes effect is the central theme of the electromagnetic environmental compatibility.

Life in this world consists predominantly of water and water has a very high dielectricity! With that the effectiveness and the long life of the potential vortex increases. The human head for instance contains 7O% and plants contain over 90% water! But it does not simply concern H_2O, but structured water in a colloidal form. These water colloids could be vortex balls because they consist of a large number of water molecules in a spherical arrangement They form independent and insoluble particles with a negative electric charge (fig. 4.11). Water is not equal water thanks to this structure. One can buy healing water and corresponding sources are well-known and famous. Many an effect can be explained by means of a chemical analysis but not everything.

The highest age in this world is reached by the inhabitants of Hunza, in the mountains of norhern India at the foothill of the Hindu Kush, at an altitude of 2500 meters. They drink some muddy glacial water that is strongly colloidal. Hence it would be obvious that plants and also we ourselves need such water for our physique. Processes are known with which the advantageous vortex balls, say colloids, are produced artificially by mechanic or chemical treatment<i>. Levitated water, as it is called and as it is for sale nowadays, is said to be more healthy. Unfortunately people predominantly work empiric in this area, because science occupies itself with this topic only little or not at all.

Another problem is the fact that the colloids again fall apart quickly. The like negative charge favours this process. The liquid crystals have to be stabilized from the outside. In the case of the Hunza-water the colloids are surrounded by a vegetable layer of fatty acid and are protected in this way <ii>. It possibly is very obliging to nature, if the water colloids also in biological systems are stabilized in that way.

Everyone of us knows that fresh spring water tastes much better than stale, bottled water, even if the chemical analysis turns out of be absolutely identical. For this fact classical science is not able to give a cause - a further problem of causality. In any case should potential vortices with their structure shaping property be considered as a cause for the formation of water colloids. It surely causes no difficulties at all to interpret the colloids as vortcx bulls.

<i>: V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112 und:
N. Harthun: NaturgemaSe Technik - Wege fur die Forschung nach Viktor Schauberger, Verlag Delta Design, Berlin 1996.

<ii>: Flanagan: Elexier der Jugendlichkeit, Waldthausen Verlag Ritterhude 1992, orig.: Elixir of Ageless

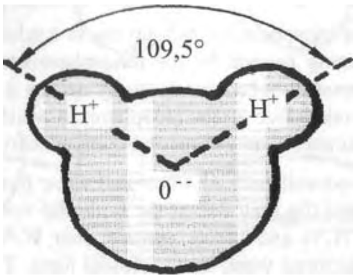

H_2O: the angle of the bond between 2 H molecules = 104,5°
angle of the bond in excited state = 109,5°

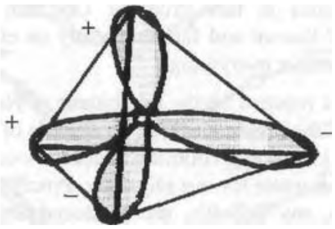

electron orbital diagram of energy (excited H_2O) = tetraeder
most stable liquid crystal = 8 tetraeder = star of octaeder

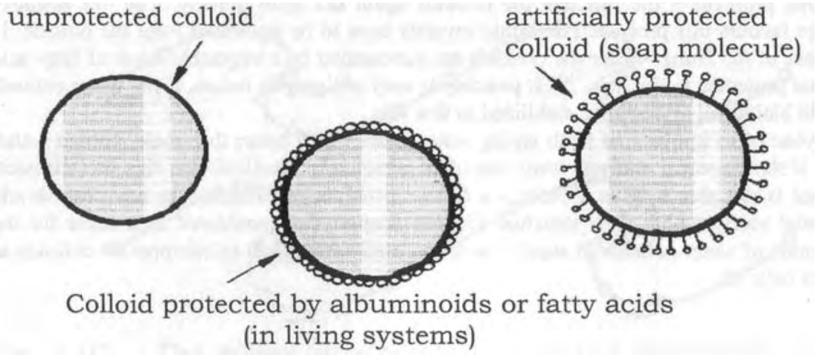
unprotected colloid
artificially protected colloid (soap molecule)
Colloid protected by albuminoids or fatty acids (in living systems)

Fig. 4.11: Water molecules and water colloids

4.11 Phenomenon of transport

The vortex principle is self-similar. This means that the properties of an individual vortex also for the collection of numerous vortices again appear and can he observed in a similar manner. That's why a vortex ball behaves entirely similar as an individual isolated vortex. The same concentration effect, that keeps the vortex together, shows its effect for the vortex ball and keeps it together also.

Something corresponding holds for a basic property of potential vortices, being of a completely different nature. It is the property to bind matter in the vortex and carry it away with the vortex. Well-known are the vortex rings that skilful cigarette smokers can blow in the air. Of course also non-smokers can produce these air eddies with their mouth but these remain invisible. Solely by the property of the vortex ring to bind the smoke it becomes visible to the human eye.

If out potential vortex transports something then it rather should be a dielectric material, so preferably water. Therefore if in the environmental air we are surrounded by potential vortices that we can detect for instance as noise, then they are capable with their "phenomenon of transport", to pick up water and to keep it in the vortex. In this way the atmospheric humidity is explicable as the ability of the air particles to bind comparatively heavy water molecules. If the vortex falls apart then it inevitably releases the water particles and it rains. This is merely a charming alternative for the classical representation without claim to completeness.

Already the Romans have made use of this phenomenon to find water and sources. About this Vitruv<i> (from 23 BC) in his 8th book about architecture writes: "Before sunrise one has to lie down on the earth at the places, where to search for water,... and one has to look at the area... Then one has to dig at the place where there appears curling and in the air rising moist steam. Because this characteristic can not occur at a place where there is no water". The at a certain time of day and in certain seasons occasional in meadows and corn fields observable streaks or circular mostly moist places with differing vegetation, have to be judged as an infallible sign for the existence of this phenomenon.

This phenomenon of transport again appears for the discussed water colloids. The involved water molecules form a spherical object with a negative charge. They turn their negatively charged side to the outside and point with the positively charged end in the direction of the middle of the sphere. There, no longer discernible from the outside, a negatively charged ion can be, that is stuck, that no longer can escape and that gives the whole colloid a characteristic property. In this way nature knows various water colloids that constitute plants and animals. But starting at a temperature of 41°C the liquid crystals fall apart. This not by chance is the temperature at which a person dies. Already 10 millivolts per liquid crystal suffice for the electrically induced death.

The to a colloid identical structure we find in the structure of the atoms. Here the atomic nucleus is held in the inside of a vortex-like cloud of electrons, the atomic hull. We'll hit the phenomenon of transport a last time, when we derive the elementary particles. For the photon is already discernible the tendency of an elementary vortex, to take another vortex in its inside. Merely because the electron and positron are evenly matched a stable configuration is prevented for the photon.

<i>: Vitruvius Pollio, Marcus: Ten Books about architecture, WBG 1987

In chapter vortex calculation used differential operations:

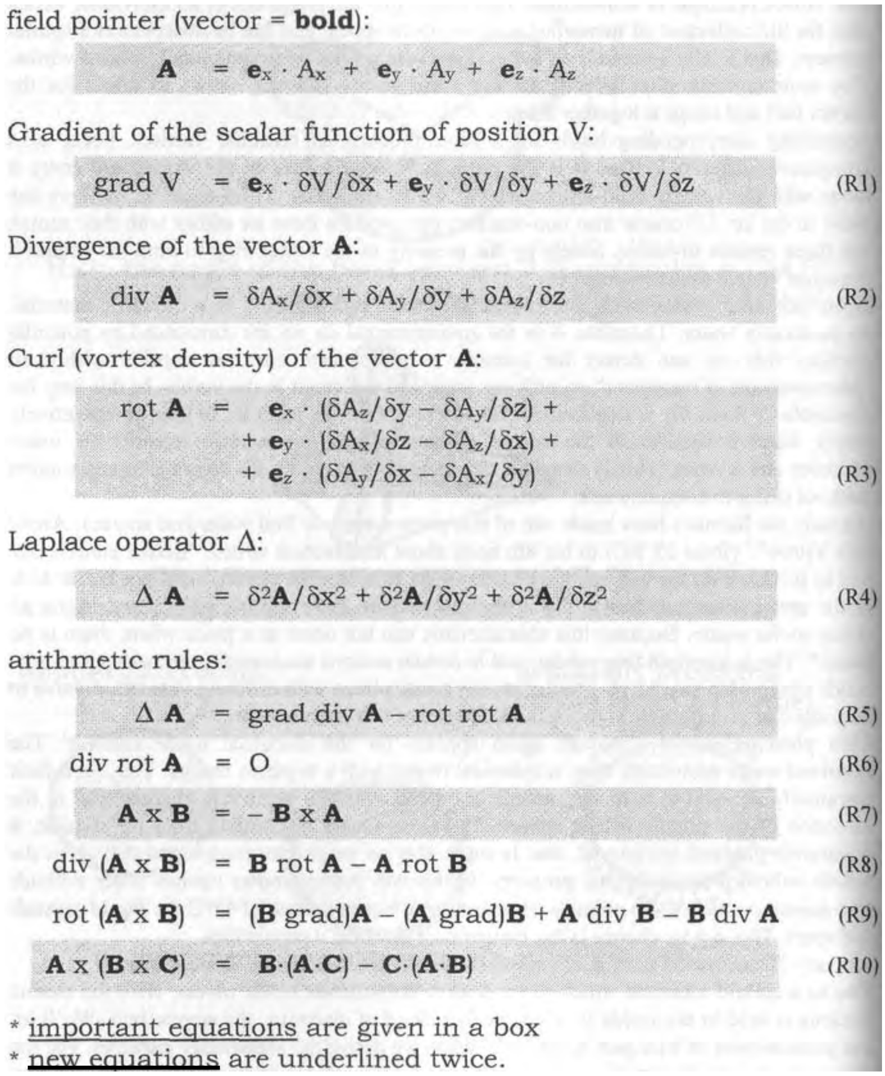

field pointer (vector = **bold**):

$$\mathbf{A} = \mathbf{e}_x \cdot A_x + \mathbf{e}_y \cdot A_y + \mathbf{e}_z \cdot A_z$$

Gradient of the scalar function of position V:

$$\mathrm{grad}\, V = \mathbf{e}_x \cdot \delta V/\delta x + \mathbf{e}_y \cdot \delta V/\delta y + \mathbf{e}_z \cdot \delta V/\delta z \tag{R1}$$

Divergence of the vector **A**:

$$\mathrm{div}\, \mathbf{A} = \delta A_x/\delta x + \delta A_y/\delta y + \delta A_z/\delta z \tag{R2}$$

Curl (vortex density) of the vector **A**:

$$\begin{aligned} \mathrm{rot}\, \mathbf{A} = \ & \mathbf{e}_x \cdot (\delta A_z/\delta y - \delta A_y/\delta z) + \\ & + \mathbf{e}_y \cdot (\delta A_x/\delta z - \delta A_z/\delta x) + \\ & + \mathbf{e}_z \cdot (\delta A_y/\delta x - \delta A_x/\delta y) \end{aligned} \tag{R3}$$

Laplace operator Δ:

$$\Delta\, \mathbf{A} = \delta^2 \mathbf{A}/\delta x^2 + \delta^2 \mathbf{A}/\delta y^2 + \delta^2 \mathbf{A}/\delta z^2 \tag{R4}$$

arithmetic rules:

$$\Delta\, \mathbf{A} = \mathrm{grad\, div}\, \mathbf{A} - \mathrm{rot\, rot}\, \mathbf{A} \tag{R5}$$

$$\mathrm{div\, rot}\, \mathbf{A} = 0 \tag{R6}$$

$$\mathbf{A} \times \mathbf{B} = - \mathbf{B} \times \mathbf{A} \tag{R7}$$

$$\mathrm{div}\, (\mathbf{A} \times \mathbf{B}) = \mathbf{B}\, \mathrm{rot}\, \mathbf{A} - \mathbf{A}\, \mathrm{rot}\, \mathbf{B} \tag{R8}$$

$$\mathrm{rot}\, (\mathbf{A} \times \mathbf{B}) = (\mathbf{B}\, \mathrm{grad})\mathbf{A} - (\mathbf{A}\, \mathrm{grad})\mathbf{B} + \mathbf{A}\, \mathrm{div}\, \mathbf{B} - \mathbf{B}\, \mathrm{div}\, \mathbf{A} \tag{R9}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{A} \cdot \mathbf{C}) - \mathbf{C} \cdot (\mathbf{A} \cdot \mathbf{B}) \tag{R10}$$

* important equations are given in a box
* new equations are underlined twice.

Fig. 5.0: Collection of formulas for vector analysis

5. Derivation and interpretation

Vortices cause big problems to every measuring technician. They have the unpleasant property to whirl around the sensor even if it is as small as possible. Vortices avoid the smallest disturbance and then can hardly be detected reproducibly.

From the well-known eddy current we know of this problematic. Instead of the vortex, we are forced to measure and analyse any effects that arise from the vortex. These can be measurements of the eddy losses or effects back on the stimulating field. But only provided that the effect actually occurs.
The prerequisite for an increase in temperature by eddy losses is that the vortex falls apart. In an ideal medium it unfortunately will not do us this pleasure.

As vrtex of the dielectric the potential vortex will find fairly ideal conditions in air or in water. How should a vortex be detected, if it does not produce any effect? The classical measuring technique is here at its wits' end.

From the duality to the well-known eddy current and by means of observation in the previous chapters numerous properties of the potential vortex have been derived. But these are not all the properties. The mathematical calculation of the electric vortex field, that we want to turn to now, will reveal still further meaningful and highly interesting properties.

The observation is important, but it can't replace an exact calculation. A strictly mathematical derived result has occasionally more expressiveness than a whole book full of explanations. It will be a big help to derive and to discuss the field equation that all considerations are based on.

We facilitate the mathematical work by vector analysis. Therefore it is useful that we choose the differential form (equation 5.1 and 5.4) instead of the integral form (equations 3.1 and 3.2 resp. 3.8).

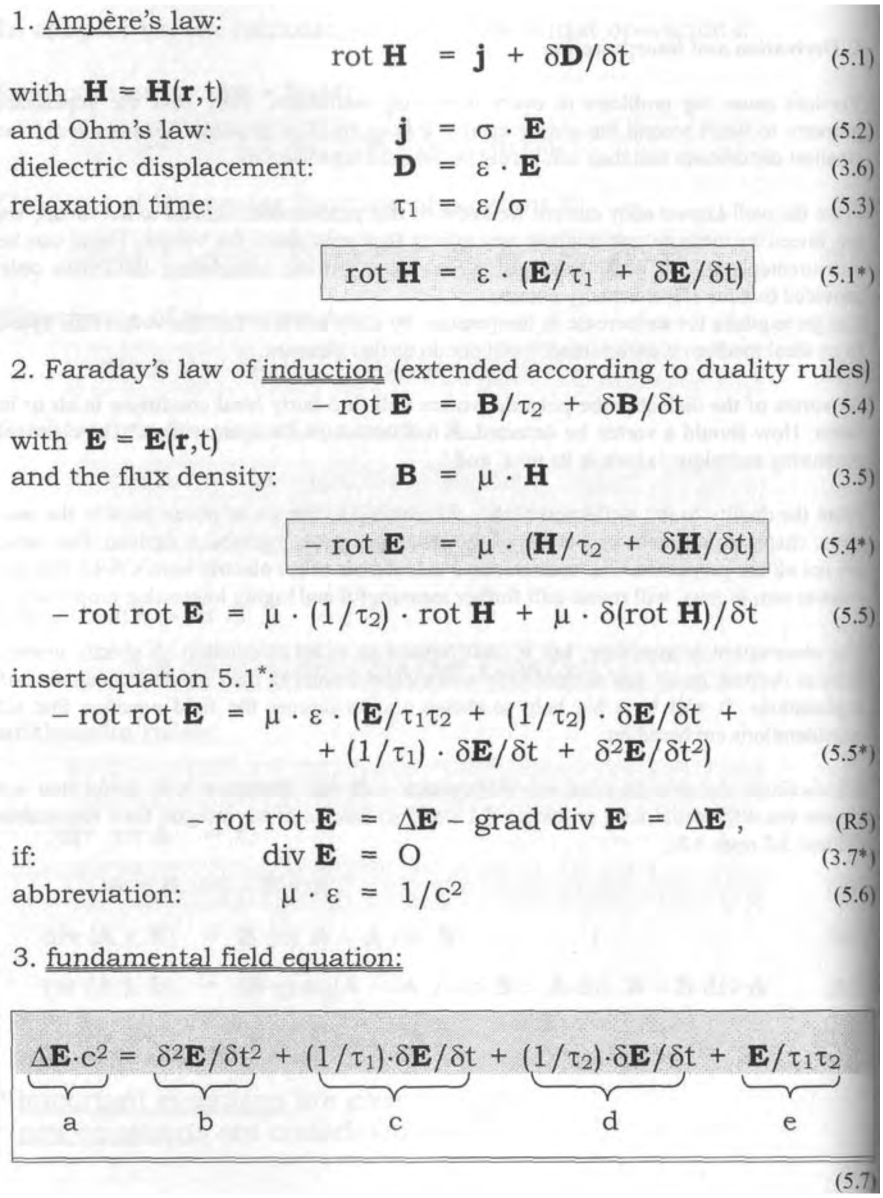

1. Ampère's law:

$$\mathrm{rot}\,\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t \qquad (5.1)$$

with $\mathbf{H} = \mathbf{H}(\mathbf{r},t)$

and Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$ (5.2)

dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

relaxation time: $\tau_1 = \varepsilon/\sigma$ (5.3)

$$\boxed{\mathrm{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta\mathbf{E}/\delta t)} \qquad (5.1^*)$$

2. Faraday's law of induction (extended according to duality rules)

$$-\,\mathrm{rot}\,\mathbf{E} = \mathbf{B}/\tau_2 + \delta\mathbf{B}/\delta t \qquad (5.4)$$

with $\mathbf{E} = \mathbf{E}(\mathbf{r},t)$

and the flux density: $\mathbf{B} = \mu \cdot \mathbf{H}$ (3.5)

$$\boxed{-\,\mathrm{rot}\,\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta\mathbf{H}/\delta t)} \qquad (5.4^*)$$

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot (1/\tau_2) \cdot \mathrm{rot}\,\mathbf{H} + \mu \cdot \delta(\mathrm{rot}\,\mathbf{H})/\delta t \qquad (5.5)$$

insert equation 5.1*:

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \varepsilon \cdot (\mathbf{E}/\tau_1\tau_2 + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + {} + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \qquad (5.5^*)$$

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \Delta\mathbf{E} - \mathrm{grad}\,\mathrm{div}\,\mathbf{E} = \Delta\mathbf{E}\,, \qquad (R5)$$

if: $\mathrm{div}\,\mathbf{E} = 0$ (3.7*)

abbreviation: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

3. fundamental field equation:

$$\boxed{\underbrace{\Delta\mathbf{E}\cdot c^2}_{a} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{b} + \underbrace{(1/\tau_1)\cdot\delta\mathbf{E}/\delta t}_{c} + \underbrace{(1/\tau_2)\cdot\delta\mathbf{E}/\delta t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e}} \qquad (5.7)$$

Fig. 5.1: Derivation of the fundamental field equation.

5.1 Fundamental field equation

We'll start from Ampere's law which provides a value for the current density at any point pace and this value corresponds to the vortex density of the magnetic field strength

$$\mathrm{rot}\,\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t \qquad (5.1)$$

We'll use Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$, (5.2)

the dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

and the relaxation time: $\tau_1 = \varepsilon/\sigma$, (5.3)

that indicates how fast the current eddies decay. So far we can fall back upon well-known relations.

The result is:

$$\boxed{\mathrm{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta\mathbf{E}/\delta t)} \qquad (5.1^*)$$

The new electric field vortices demand the introduction of a corresponding time constant tau$_2$ that should describe the decay of the potential vortices, as an extension. The extended Faraday law of induction now provides a potential density, that at any point of space corresponds to the vortex density of the electric field strength:

$$-\mathrm{rot}\,\mathbf{E} = \mathbf{B}/\tau_2 + \delta\mathbf{B}/\delta t \qquad (5.4)$$

with the flux density $\mathbf{B} = \mu \cdot \mathbf{H}$. (3.5)

The result fulfils the required duality to equation 5.1*:

$$\boxed{-\mathrm{rot}\,\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta\mathbf{H}/\delta t)} \qquad (5.4^*)$$

If we again apply the curl to equation 5.4*

$$-\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot (1/\tau_2) \cdot \mathrm{rot}\,\mathbf{H} + \mu \cdot \delta(\mathrm{rot}\,\mathbf{H})/\delta t \qquad (5.5)$$

and insert equation 5.1*, we obtain:

$$-\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \varepsilon \cdot (\mathbf{E}/\tau_1\tau_2 + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \qquad (5.5^*)$$

which according to the rules of vector analysis can still be further simplified: = rot rot E = $\Delta\mathbf{E}$ - grad div E , where we should remember that the divergence has to vanish (div E = O. fig. 3.2, equation 3.7), should the corresponding field vortex be inserted!

Furthermore the following well-known abbreviation can be inserted: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

With that the relation with the speed of light c simplifies to the sought-for field equation:

$$\boxed{\underbrace{\Delta\mathbf{E} \cdot c^2}_{a} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{b} + \underbrace{(1/\tau_1) \cdot \delta\mathbf{E}/\delta t}_{c} + \underbrace{(1/\tau_2) \cdot \delta\mathbf{E}/\delta t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e}} \qquad (5.7)$$

This equation describes the spatial (a) and temporal (b, c, d) distribution of a field vector. It describes the electromagnetic wave (a, b) with the influences that act damping. As dumping terms the well-known eddy current (c) and in addition the newly introduced potential vortex (d) appear.

Field vector: ψ = E, H, j, B or D

1. elliptic potential equation:
(stationary: $t \longrightarrow \infty$ resp. $\delta/\delta t = 0$)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\psi/\tau_1\tau_2}_{e}}$$

(5.8)

2. hyperbolic equation:
(undamped wave equation)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b}}$$

(5.9)

3. parabolic equation:
(vortex equation)

$$\boxed{\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{(1/\tau) \cdot \delta\psi/\delta t}_{c/d}}$$

(5.10)

decay time of the eddy currents =
relaxation time: $\tau_1 = \varepsilon/\sigma$ (5.3)

decay time of the potential vortices =
relaxation time: $\tau_2 \sim \mu \cdot \sigma$ (5.11)

Fig. 5.2: mathematically divisible individual cases.

5.2 Mathematical interpretation of the fundamental field equation

Every specialist will be surprised to find the Poisson equation (a, e) again as a term in the wave equation. This circumstance forces a completely new interpretation of stationary fields upon us. The new damping term, that is formed by the potential vortices (d), is standing in between.

Let us start with a mathematical analysis. We have applied the curl to equation 5.4*, then inserted equation 5.1* and obtained a determining equation for the electric field strength E. Of course we could as well have applied the curl to equation 5.1* and inserted equation 5.4*. This would have resulted in the determining equation for the magnetic field strength H.

If we insert Ohm's law (5.2) and cancel down the specific conductivity, or we put in the relations of material (3.5) or (3.6) and cancel down by u respectively ε, then the field equation can likewise be written down for the current density j, for the induction B or for the dielectric displacement D.

It just is phenomenal that at all events equation 5.7 doesn't change its form at all. The field vector is thus arbitrarily interchangeable! This circumstance is the foundation for the claim of this field equation to be called fundamental.

It does make sense to introduce a neutral descriptive vector ψ as a substitute for the possible field factors E, H, j, B or D.

The fundamental field equation 5.7 consists of three different types of partial differential equations: a hyperbolic (b), a parabolic (c and d) and an elliptic (e) type. On the left-hand side each time the Laplace operator (a) is found which describes the spatial distribution of the field factor.

The potential equation of the elliptic type (e) is known as Poisson equation. It describes the stationary borderline case of a worn off temporal process ($t \to \infty$, resp. $\delta/\delta t = 0$).

With this equation potentials and voltages can be calculated exactly like stationary electric currents (5.8).

The hyperbolic equation (b). known as wave equation, shows a second derivative to time. which expresses an invariance with regard to time reversal; or stated otherwise: the direction of the time axis can be reversed by a change of sign of t, without this having an influence on the course of frequency. Wave processes hence are reversible. Equation 5.7 makes clear that a wave without damping by no means can exist in nature. For that both time constants (τ_1 and τ_2) would have to have an infinite value, which is not realizable in practice. Seen purely theoretical, undamped waves could withdraw themselves from our measuring technique (5.9).

Both vortex equations of the parabolic type (c and d) only show a first derivative to time. With that they are no longer invariant with regard to time reversal. The processes of the formation and the decay of vortices, the so-called diffusion, are as a consequence irreversible. Seen this way it is understandable that the process of falling apart of the vortex, where the vortex releases its stored energy as heat e.g. in form of eddy losses, can not take place in reverse. This irreversible process of diffusion in the strict thermodynamic sense increases the entropy of the system (5.10).

Because it poses an useful simplification for mathematical calculations, often the different types of equations are treated isolated from each other. But the physical reality looks different.

field vector: ψ = E, H, j, B or D

1. Borderline case: no conductivity (vacuum)
$(\sigma = 0; \quad 1/\tau_1 = \sigma/\varepsilon = 0)$:

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b} + \underbrace{(1/\tau_2)\cdot \delta\psi/\delta t}_{d}$$

(5.12)

(damping by potential vortices)

2. Borderline case: ideal conductivity (superconductor)
$(1/\sigma = 0; \quad 1/\tau_2 = 0)$:

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{\delta^2\psi/\delta t^2}_{b} + \underbrace{(1/\tau_1)\cdot \delta\psi/\delta t}_{c}$$

(5.12*)

(damping by eddy currents)

3. Diffusion equation

$$\underbrace{\Delta\psi \cdot c^2}_{a} = \underbrace{(1/\tau)\cdot \delta\psi/\delta t}_{c/d} + \underbrace{\psi/\tau_1\tau_2}_{e}$$

(5.12**)

(vortex)

Fig. 5.3: Two borderline cases of the damping of waves and the diffusion equation for the decay of vortices

5.3 Physical interpretation of the fundamental field equation

In nature the different types of equations always occur in a combined manner.

1. Let's take the concrete case of the particle-free vacuum. Here the specific conductivity is zero. The relaxation time constant $\tau_1 = \varepsilon/\sigma$ responsible for the decay of vortices tends towards infinity according to equation 5.3 and the terms (c) and (e) are cancelled from the field equation 5.7. What remains is the by potential vortices (d) damped wave equation (b) (equation 5.12).

2. The reversed case (with $\tau_2 \longrightarrow \infty$) will consequently occur in materials without resistance, super conducting materials. We now are dealing with the well-known case of the wave damped by eddy currents (equation 5.12*).

Virtually all in nature existing materials however don't fulfil these boundary conditions, from which it follows that both damping terms always occur together and in addition the stationary term (e) becomes active.
It is true that every antenna demonstrates that the electromagnetic wave is convertible in high-frequency alternating currents and voltages, which then are amplified in the receiver. But until this fundamental equation was written down it however was not understood that this transition takes place by means of a vortex. Used are either antennas from well conducting material, or wave guides and horn radiators, which only have a minimal conductivity, because they are filled with air. Actually the wave can be converted in two dual ways; by means of the rolling up to current eddies or to potential vortices (fig. 1.4). Now we finally are capable to explain, why wave guides make possible a better degree of effectiveness: Owing to the concentration effect of the potential vortex the HF-power is bound in the inside and not emitted until the antenna is reached as happens for a wire for reason of the skin effect.
Therefore, physically, one has to imagine this relation, which describes the transition of an electromagnetic wave into a vortex, in the way that the wave spontaneously can roll up to a vortex in case it is disturbed from the outside. The more vortices are generated, the larger consequently is the damping of the wave (equations 5.12 and 5.12*).

3. The life span of the vortices is limited and is determined by the electric conductivity. The at first stored vortices decay with their respective time constant τ. This process is described by the diffusion equation 5.12**. The final stage of the decaying vortices finally is described by the Poisson equation (a, e: equation 5.8).

If the vortex falls apart, it converts the in the vortex stored energy in heat. These processes are known from the eddy current. We speak of heating losses, that the stationary currents cause in the conductor material.
But new is the concept that such vortex phenomena can occur as dielectric losses in capacitors or in the air. The microwave oven or induction welding are good examples of this.

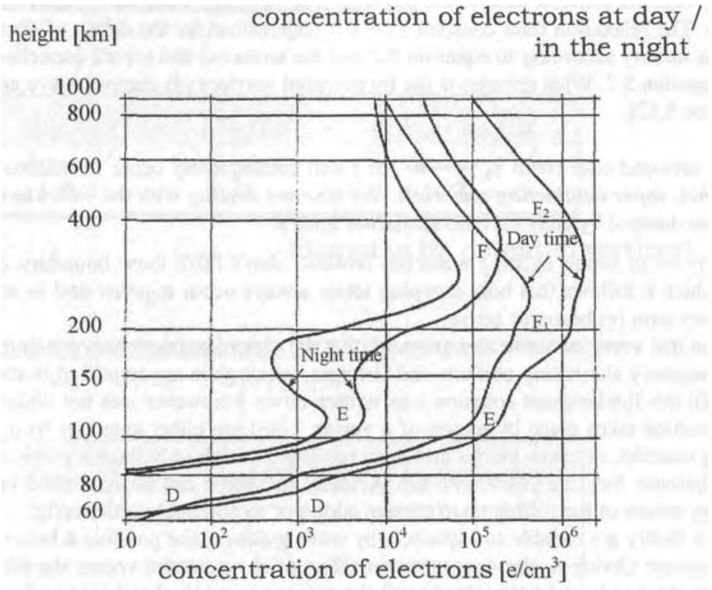

Fig. 5.4: The dependency on height of the ionisation in the
ionosphere for medium latitudes .<i>

left curve: for a minimum of sun spots

right curve: for a maximum of sun spots

<i>: H.L. Konig: Unsichtbare Umwelt (Wetterfuhligkeit), 5. Aufl., Bild 6, Seite 11, Verlag Moos & Partner Milnchen, ISBN 3-89164-058-7

5.4 Phenomenological interpretation of the fundamental field equation

How does a damping by vortices take effect in practice? First of all we notice that the reception of broadcastings gets worse. "The information signal is neglectable regarding the noise" explains the radio engineer and means, the number of vortices increases at the expense of the wave intensity.
Why, does the pupil ask, is it so cold in space? There the sun shines day and night and in addition much more intensely than on earth! The correct answer would have to read that because of the extremely small conductivity no diffusion process can take place. We owe the warmth on our earth solely the here taking place decay of vortices. Responsible is the conductivity of the atmosphere.
In 60 km to 500 km height over the earth's surface, the region which is called the ionosphere, the gases predominantly exist in ionized form. There a very good conductivity prevails and eddy current losses are the result. Correspondingly high are the measurable temperatures. Besides the diffusion process the eddy currents carry out a damping of the cosmic radiation. We say the sunlight is filtered and reduced to a for nature bearable intensity.
But not all frequencies are damped in the same way (fig. 2.8). We observe a blue shift, if we look into the actually black sky. The blue sky doesn't show any spots or clouds. The reason is to be sought in the skin effect of the eddy currents, which strive outwards. Since no edge of a conductor is present here, no skin can form. The vortices spread evenly over the ionosphere.
The potential vortex however is able to structure. It merely needs a bad conductivity and this it finds in lower heights between 1 km and 10 km. It damps the wave and with that also the light, for which reason we say it becomes darker, the sun disappears behind clouds.

The clouds well visibly form the discussed vortex balls and vortex strings. Clouds can form virtually from the nowhere during intense solar irradiation, i.e. the waves can roll up to vortices. But as a rule this takes place above the oceans. Here also the phenomenon of transport has an effect. Because of the high dielectricity the water surface favours the formation of potential vortices. So the vortices bind individual water molecules and carry them away. If a diffusion process takes place, in which the vortex decays, then it rains.
This can happen in two different ways:
1. Either the conductivity increases. If for instance during intense solar irradiation air ions form, the sun is able to break up clouds and fog. Or when the air is raised in higher layers with better conductivity, because a mountain forces this, then it rains at the mountain edge.
2. For potential vortices the electric field is standing perpendicular to them. If at one point an exceptionally lot of vortices join together, which let the cloud appear particularly dark to black, then the danger exists that the ionization field strength for air is reached, in which case a conductive air channel forms along which the stored up charges discharge. Also lightning is a diffusion process, in which potential vortices decay and rain can form.
In connection with the electromagnetic environmental compatibility great importance is attributed in particular to the storage and the decay of electric vortices. There not only is an academic-scientific interest in the question, how many potential vortices are generated, how many are stored and how many decay, if we make a telephone call with a handy, if we are staying under a high-tension line or if we are eating food, which has been heated up in a microwave oven. The necessary mathematical description is provided by the fundamental field equation 5.7.

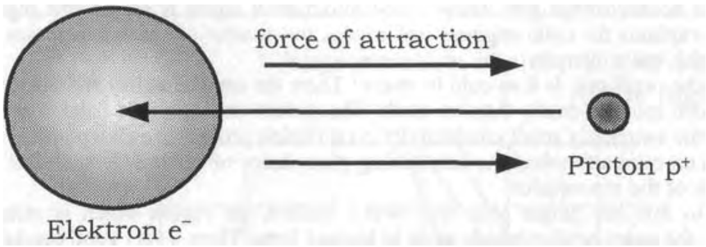
force of attraction
Proton p⁺
Elektron e⁻

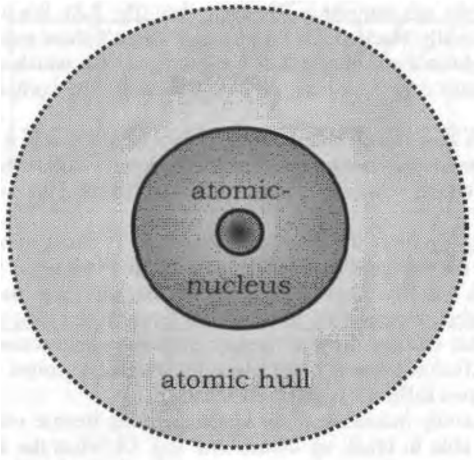
atomic-
nucleus
atomic hull

Fig. 5.5: The structure of atoms in the view of the fundamental field equation

condition for equilibrium: $$\boxed{\tau_1 = \tau_2} \qquad (5.13)$$

5.5 Atomistic interpretation of the fundamental field equation

Let's again turn to the smaller, the atomistic dimensions. Here positively charged protons and negatively charged electrons are found. Both are matter particles and that means that seen from the outside both have the identical swirl direction. For reason of the unequal charge conditions they attract each other mutually and according to fig. 4.9 rotate around a common centre of mass as differently heavy pair. Chemists say: "the light electron orbits the heavy atomic nucleus". With small balls they try to explain the atomic structure.
But the model is no good: it contradicts causality in the most elementary manner. We are dealing with the problem that according to the laws of electrodynamics a centripetally accelerated electron should emit electromagnetic waves and continuously lose energy, to eventually plunge into the nucleus.
Our experience teaches that this fortunately is not true - and Niels Bohr in order to save his model of the atom was forced to annul the laws of physics with a postulate founded in arbitrariness.
Actually this state only exists for a very short time and then something unbelievable happens: the electron can't be distinguished as an individual particle anymore. "It is smeared over the electron orbit" do certain people say; "it possesses a dual nature" says Heisenberg. Besides the corpuscular nature the electron should in case of its "second nature" form a matter wave, "the position of the electron is to be looked at as a resonance which is the maximum of a probability density", do explain us de Broglie and Schrodinger.

These explanations can hardly convince. If the electron loses its particle nature in its second nature, then it also will lose its typical properties, like for instance its mass and its charge. but this is not the case.
THE vortex theory provides clear and causal answers: if the electron were a ball it continuosly would lose energy, therefore another configuration forms that does not know this problem. Here the phenomenon of transport takes an effect. The electron opens its vortex centre and takes the tiny protons and neutrons as atomic nucleus up into itself. The Bohr electron orbit with that is not a path anymore, but is occupied by the whole particle as spherical shell. This is confirmed by the not understood measurements exactly like the photos of individual atoms with the scanning electron microscope.
But now an electron does in its inside have the opposite swirl direction as the proton seen from the outside. As a consequence a force of repulsion will occur, which can be interpreted as the to the outside directed current eddy, the force of attraction for reason of the opposite charge works in the opposite direction and can be interpreted as the potential vortex effect.
If both vortices are equally powerful: $\tau_1 = \tau_2$ (5.13)
or if both forces are balanced, as one usually would say, then the object which we call an atom is in a stable state.
It probably will be a result of the incompatible swirl direction, why a very big distance results, if the electron becomes an enveloping electron. On such a shell not too many electrons have room. Because of the rotation of their own, the electron spin, they form a magnetic dipole moment, which leads to a magnetic attraction of two electrons if they put their spin axis antiparallelly.
As a "frictionless" against one another rotating pair they form two half-shells of a sphere and with that occupy the innermost shell in the hull of an atom. If the positive charge of the nucleus is still not balanced with that, then other electrons is left only the possibility to form another shell. Now this next electron takes the whole object up into itself. The new shell lies further on the outside and naturally offers room to more as only two vortices.

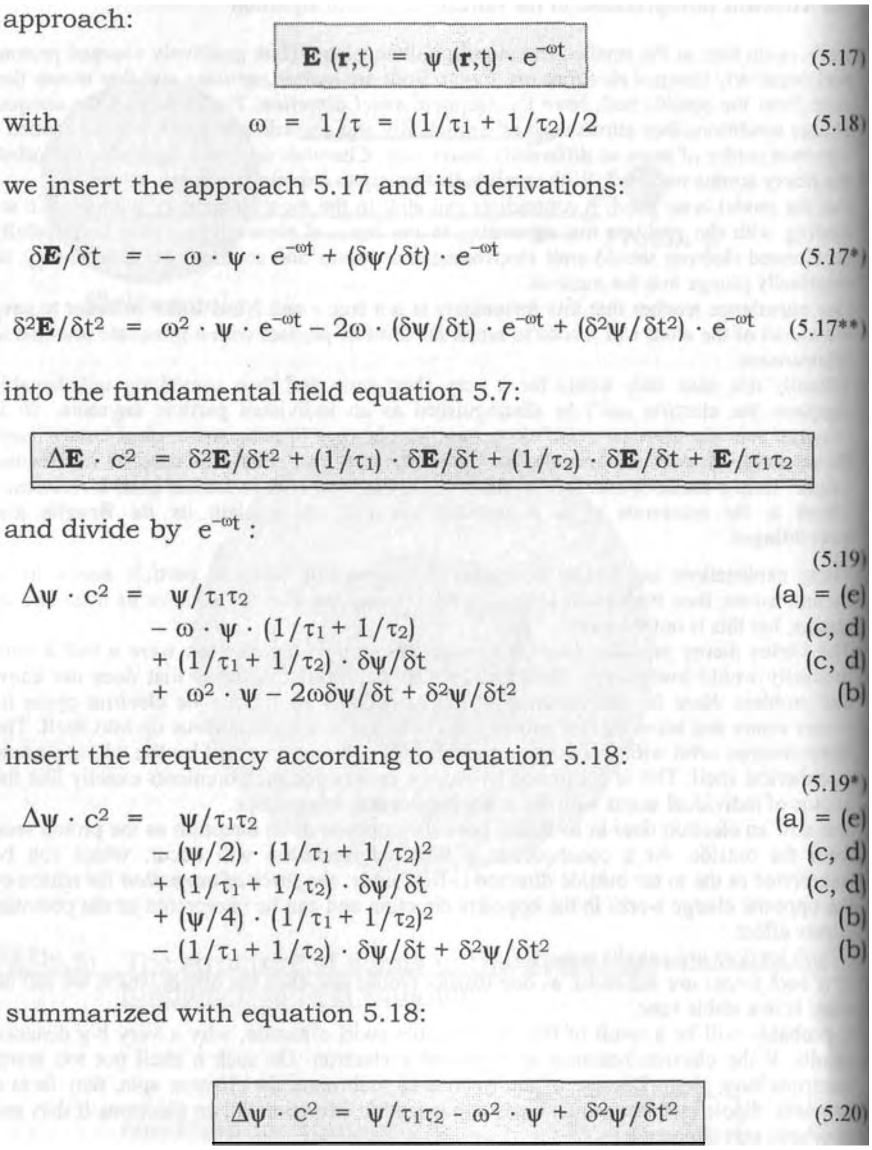

approach:

$$\boxed{\mathbf{E}(\mathbf{r},t) = \psi(\mathbf{r},t) \cdot e^{-\omega t}} \qquad (5.17)$$

with $\omega = 1/\tau = (1/\tau_1 + 1/\tau_2)/2$ (5.18)

we insert the approach 5.17 and its derivations:

$$\delta\mathbf{E}/\delta t = -\omega \cdot \psi \cdot e^{-\omega t} + (\delta\psi/\delta t) \cdot e^{-\omega t} \qquad (5.17^{*})$$

$$\delta^2\mathbf{E}/\delta t^2 = \omega^2 \cdot \psi \cdot e^{-\omega t} - 2\omega \cdot (\delta\psi/\delta t) \cdot e^{-\omega t} + (\delta^2\psi/\delta t^2) \cdot e^{-\omega t} \qquad (5.17^{**})$$

into the fundamental field equation 5.7:

$$\boxed{\Delta\mathbf{E} \cdot c^2 = \delta^2\mathbf{E}/\delta t^2 + (1/\tau_1) \cdot \delta\mathbf{E}/\delta t + (1/\tau_2) \cdot \delta\mathbf{E}/\delta t + \mathbf{E}/\tau_1\tau_2}$$

and divide by $e^{-\omega t}$:

(5.19)

$$\begin{aligned} \Delta\psi \cdot c^2 = {} & \psi/\tau_1\tau_2 && \text{(a) = (e)} \\ & - \omega \cdot \psi \cdot (1/\tau_1 + 1/\tau_2) && \text{(c, d)} \\ & + (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t && \text{(c, d)} \\ & + \omega^2 \cdot \psi - 2\omega\delta\psi/\delta t + \delta^2\psi/\delta t^2 && \text{(b)} \end{aligned}$$

insert the frequency according to equation 5.18:

(5.19*)

$$\begin{aligned} \Delta\psi \cdot c^2 = {} & \psi/\tau_1\tau_2 && \text{(a) = (e)} \\ & - (\psi/2) \cdot (1/\tau_1 + 1/\tau_2)^2 && \text{(c, d)} \\ & + (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t && \text{(c, d)} \\ & + (\psi/4) \cdot (1/\tau_1 + 1/\tau_2)^2 && \text{(b)} \\ & - (1/\tau_1 + 1/\tau_2) \cdot \delta\psi/\delta t + \delta^2\psi/\delta t^2 && \text{(b)} \end{aligned}$$

summarized with equation 5.18:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - \omega^2 \cdot \psi + \delta^2\psi/\delta t^2} \qquad (5.20)$$

Fig. 5.6: Derivation of the Klein-Gordon equation (5.20) from the fundamental field equation (5.7)

5.6 Derivation of the Klein-Gordon equation

The valid model of the atom today still raises problems of causality, as has been explained. An improvement was provided by an equation, which was proposed by the mathematician Schrodinger 1926 as a model description. This equation in this way missed the physical root, but it nevertheless got international acknowledgment, because it could be confirmed experimentally. Looking backwards from the result then the physical interpretation of the probability density of the resonance of the waves could be pushed afterwards.

$$i \cdot \hbar \cdot \delta\psi/\delta t = U \cdot \psi - (\hbar^2/2m) \cdot \Delta\psi \quad (5.14)$$

The Schrodinger equation is valid for matter fields (of mass m), while the interaction with a outside force field the energy U indicates. It can be won from a wave equation by conversion, which possibly is the reason why it usually is called a wave equation, although in reality it is a diffusion equation, so a vortex equation!
For the derivation Schrodinger gives the approach of a harmonic oscillation for the complex wave function ψ:

$$\boxed{\psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \quad (5.15)$$

if the entire time dependency can be described by one frequency $f = W/h$ (de-Broglie relation):

$$\omega = 2\pi f = W \cdot 2\pi/h = W/\hbar \quad (5.16)$$

The high-put goal is: if the structure of the atom is determined by the fundamental field equation 5.7 then one should be able to derive the experimentally secured Schrodinger equation and to mathematically describe the discussed special case. Also we select at first an approach periodic in time:

$$\boxed{\mathbf{E}(\mathbf{r},t) = \psi(\mathbf{r},t) \cdot e^{-\omega t}} \quad (5.17)$$

with

$$\omega = 1/\tau = (1/\tau_1 + 1/\tau_2)/2 \ . \quad (5.18)$$

We insert the approach 5.17 and its derivations into the field equation 5.7 and divide by the damping term e^{-wt}:

$$\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - \omega \cdot \psi \cdot (1/\tau_1 + 1/\tau_2) + (1/\tau_1 + 1/\tau_2) \cdot (\delta\psi/\delta t) + \\ + \omega^2 \cdot \psi - 2\omega\,\delta\psi/\delta t + \delta^2\psi/\delta t^2 \quad (5.19)$$

If as the next step the angular frequency according to equation 5.18 is inserted, then summarized the provisional intermediate result results:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - \omega^2 \cdot \psi + \delta^2\psi/\delta t^2} \quad (5.20)$$

The derived equation 5.20 represents formally seen the Klein-Gordon equation, which is used for the description of matter waves in quantum mechanics and which particularly in the quantum field theory (e.g. mesons) plays an important role. Even if it often is regarded as the relativistic invariant generalization of the Schrodinger equation, it at a closer look is incompatible with this equation and as "genuine" wave equation it is not capable of treating vortex problems correctly, like e.g. the with the Schrodinger equation calculable quantization of our microcosm.

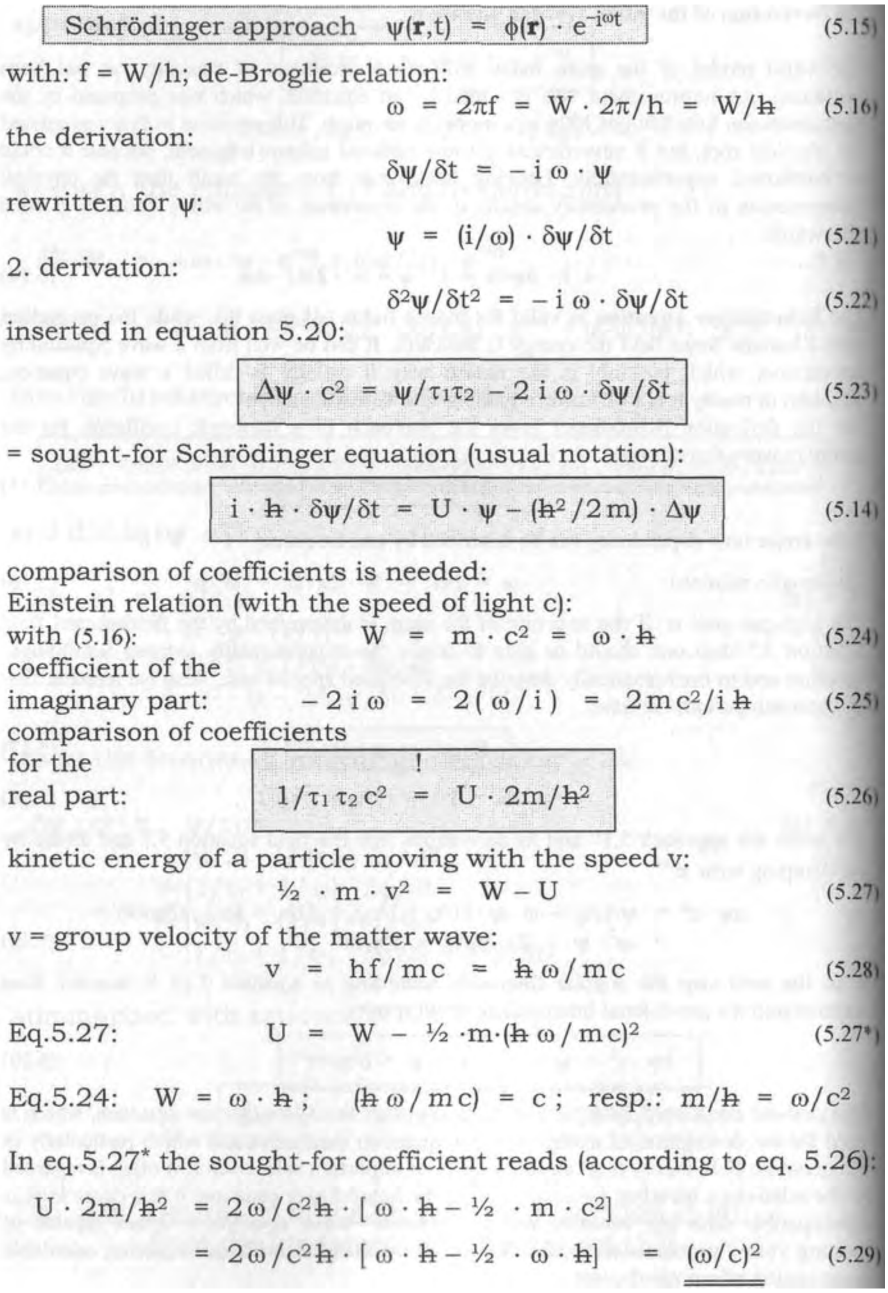

$$\boxed{\text{Schrödinger approach}\quad \psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \tag{5.15}$$

with: f = W/h; de-Broglie relation:

$$\omega = 2\pi f = W \cdot 2\pi/h = W/\hbar \tag{5.16}$$

the derivation:

$$\delta\psi/\delta t = -i\,\omega \cdot \psi$$

rewritten for ψ:

$$\psi = (i/\omega) \cdot \delta\psi/\delta t \tag{5.21}$$

2. derivation:

$$\delta^2\psi/\delta t^2 = -i\,\omega \cdot \delta\psi/\delta t \tag{5.22}$$

inserted in equation 5.20:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - 2\,i\,\omega \cdot \delta\psi/\delta t} \tag{5.23}$$

= sought-for Schrödinger equation (usual notation):

$$\boxed{i \cdot \hbar \cdot \delta\psi/\delta t = U \cdot \psi - (\hbar^2/2m) \cdot \Delta\psi} \tag{5.14}$$

comparison of coefficients is needed:
Einstein relation (with the speed of light c):

with (5.16): $$W = m \cdot c^2 = \omega \cdot \hbar \tag{5.24}$$

coefficient of the
imaginary part: $$-2\,i\,\omega = 2(\omega/i) = 2mc^2/i\hbar \tag{5.25}$$

comparison of coefficients
for the
real part: $$\boxed{1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2} \tag{5.26}$$

kinetic energy of a particle moving with the speed v:

$$\tfrac{1}{2} \cdot m \cdot v^2 = W - U \tag{5.27}$$

v = group velocity of the matter wave:

$$v = hf/mc = \hbar\,\omega/mc \tag{5.28}$$

Eq.5.27: $$U = W - \tfrac{1}{2} \cdot m \cdot (\hbar\,\omega/mc)^2 \tag{5.27*}$$

Eq.5.24: $W = \omega \cdot \hbar$; $(\hbar\,\omega/mc) = c$; resp.: $m/\hbar = \omega/c^2$

In eq.5.27* the sought-for coefficient reads (according to eq. 5.26):

$$\begin{aligned} U \cdot 2m/\hbar^2 &= 2\omega/c^2\hbar \cdot [\,\omega \cdot \hbar - \tfrac{1}{2} \cdot m \cdot c^2] \\ &= 2\omega/c^2\hbar \cdot [\,\omega \cdot \hbar - \tfrac{1}{2} \cdot \omega \cdot \hbar] = \underline{\underline{(\omega/c)^2}} \end{aligned} \tag{5.29}$$

Fig. 5.7: Derivation of the time dependent Schrodinger equation

5.7 Derivation of the time dependent Schrodinger equation

With the Schrodinger approach 5.15 and its derivations the derivation is continued:

$$\boxed{\psi(\mathbf{r},t) = \phi(\mathbf{r}) \cdot e^{-i\omega t}} \, , \qquad (5.15)$$

$$\delta\psi/\delta t = -i\omega \cdot \psi \, , \quad \text{bzw.} \quad \psi = (i/\omega) \cdot \delta\psi/\delta t \qquad (5.21)$$

$$\delta^2\psi/\delta t^2 = -i\omega \cdot \delta\psi/\delta t \qquad (5.22)$$

The for a harmonic oscillation won relations according to equation 5.21 and 5.22 are now inserted into equation 5.20:

$$\boxed{\Delta\psi \cdot c^2 = \psi/\tau_1\tau_2 - 2i\omega \cdot \delta\psi/\delta t}$$

This is already the sought-for Schrodinger equation, as we will see in a moment, when we have analysed the coefficients. Because, besides equation 5.16 for the total energy W, also the Einstein relation is valid (with the speed of light c):

$W = m \cdot c^2 = \omega \cdot \hbar$, (5.24)

we can replace the coefficients of the imaginary part by:

$2(\omega/i) = 2mc^2/i\hbar$ (5.25)

To achieve that equation 5.23, as required, follows from the Schrodinger equation 5.14, a comparison of coefficients is carried out for the real part:

$\boxed{1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2}$ (5.26)

If thc kinetic energy of a particle moving with the speed v is:

$\frac{1}{2} \cdot m \cdot v^2 = W - U$, (5.27)

then acccording to De Broglie this particle has the wavelength h/mv. The consideration of the particle as matter wave demands an agreement with the wave length c/f of an electromagnetic wave (with the phase velocity c). The particle hence has the speed v, which corresponds with the group velocity of the matter wave:

$$v = hf/mc = \hbar\omega/mc \, , \qquad (5.28)$$

if we insert v into equation 5.27 :

$U = W - \frac{1}{2} \cdot m \cdot (\hbar\omega/mc)^2$ (5.27*)

According to equation 5.24 on the one hand the total energy is $W = w \cdot h$ and on the other hand the relation 5.28 gives $(\hbar\omega/mc) = c$ resp.: $m/\hbar = \omega/c^2$

Inserted into equation 5.27* the sought-for coefficient reads (according to eq. 5.26):

$$U \cdot 2m/\hbar^2 = 2\omega/c^2\hbar \cdot [\omega \cdot \hbar - \tfrac{1}{2} \cdot m \cdot c^2]$$
$$= 2\omega/c^2\hbar \cdot [\omega \cdot \hbar - \tfrac{1}{2} \cdot \omega \cdot \hbar] = \underline{(\omega/c)^2} \qquad (5.29)$$

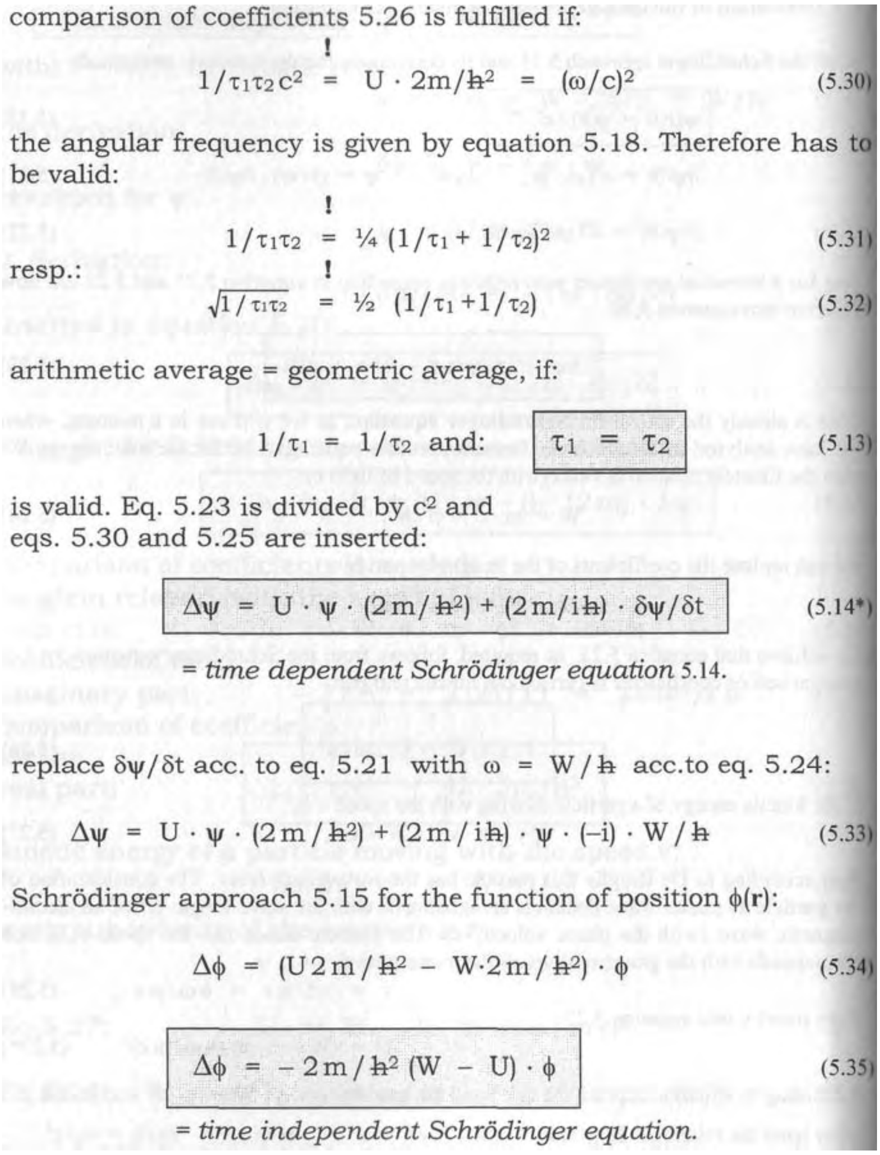

comparison of coefficients 5.26 is fulfilled if:

$$1/\tau_1\tau_2 c^2 \overset{!}{=} U \cdot 2m/\hbar^2 = (\omega/c)^2 \qquad (5.30)$$

the angular frequency is given by equation 5.18. Therefore has to be valid:

$$1/\tau_1\tau_2 \overset{!}{=} \tfrac{1}{4}\,(1/\tau_1 + 1/\tau_2)^2 \qquad (5.31)$$

resp.:

$$\sqrt{1/\tau_1\tau_2} \overset{!}{=} \tfrac{1}{2}\,(1/\tau_1 + 1/\tau_2) \qquad (5.32)$$

arithmetic average = geometric average, if:

$$1/\tau_1 = 1/\tau_2 \text{ and: } \boxed{\tau_1 = \tau_2} \qquad (5.13)$$

is valid. Eq. 5.23 is divided by c^2 and eqs. 5.30 and 5.25 are inserted:

$$\boxed{\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \delta\psi/\delta t} \qquad (5.14^*)$$

= time dependent Schrödinger equation 5.14.

replace $\delta\psi/\delta t$ acc. to eq. 5.21 with $\omega = W/\hbar$ acc.to eq. 5.24:

$$\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \psi \cdot (-i) \cdot W/\hbar \qquad (5.33)$$

Schrödinger approach 5.15 for the function of position $\phi(\mathbf{r})$:

$$\Delta\phi = (U 2m/\hbar^2 - W \cdot 2m/\hbar^2) \cdot \phi \qquad (5.34)$$

$$\boxed{\Delta\phi = -2m/\hbar^2\,(W - U) \cdot \phi} \qquad (5.35)$$

= time independent Schrödinger equation.

Fig. 5.8: Derivation of the time independent Schrodinger equation

5.8 Derivation of the time independent Schrodinger equation

The goal is reached if we are capable to fulfil the comparison of coefficients 5.26:

$$U \cdot 2m/\hbar^2 \overset{!}{=} (\omega/c)^2 = 1/\tau_1\tau_2 c^2 \qquad (5.30)$$

The angular frequency w is given by equation 5.18. Therefore has to be valid:

$$\tfrac{1}{4}\,(1/\tau_1 + 1/\tau_2)^2 \overset{!}{=} 1/\tau_1\tau_2 \qquad (5.31)$$

$$\tfrac{1}{2} \cdot (1/\tau_1 + 1/\tau_1) \overset{!}{=} \sqrt{1/\tau_1\tau_2} \qquad (5.32)$$

As is well-known the arithmetic and the geometric average only correspond in case the variables are identical. In this case, as already required in equation 5.13:

$$\tau_1 = \tau_2\,. \qquad (5.13)$$

has to hold.
From this we can draw the conclusion that the Schrodinger equation is just applicable to the described special case (according to eq. 5.13), in which the eddy current, which tries to increase the particle or its circular path and the potential vortex, which keeps the atoms together and also is responsible for the stability of the elementary particles, are of identical order of magnitude.

As a check equation 5.23 is divided by c^2 and equations 5.30 and 5.25 are inserted:

$$\boxed{\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \delta\psi/\delta t} \qquad (5.14^*)$$

This is the time dependent Schrodinger equation 5.14 resolved for $\Delta\psi$.

Next we replace $\delta\psi/\delta t$ according to equation 5.21 with $\omega = W/\hbar$ acc. to equation 5.24:

$$\Delta\psi = U \cdot \psi \cdot (2m/\hbar^2) + (2m/i\hbar) \cdot \psi \cdot (-i) \cdot W/\hbar \qquad (5.33)$$

If we separate the space variables $\phi(\mathbf{r})$ from time by the Schrodinger approach 5.15 we obtain:

$$\Delta\phi = (U\,2m/\hbar^2 - W \cdot 2m/\hbar^2) \cdot \phi \qquad (5.34)$$

This quation 5.34 for the function of space coordinates $\phi(\mathbf{r})$ is the time independent Schrodinger equation:

$$\boxed{\Delta\phi = -2m/\hbar^2\,(W - U) \cdot \phi} \qquad (5.35)$$

The solutions of this equation which fulfil all the conditions that can be asked of them (of finiteness, steadiness, uniqueness etc.), are called eigenfunctions. The existence of corresponding discrete values of the energy W, also called eigenvalues of the Schrodinger equation, are the mathematical reason for the different quantum postulates.

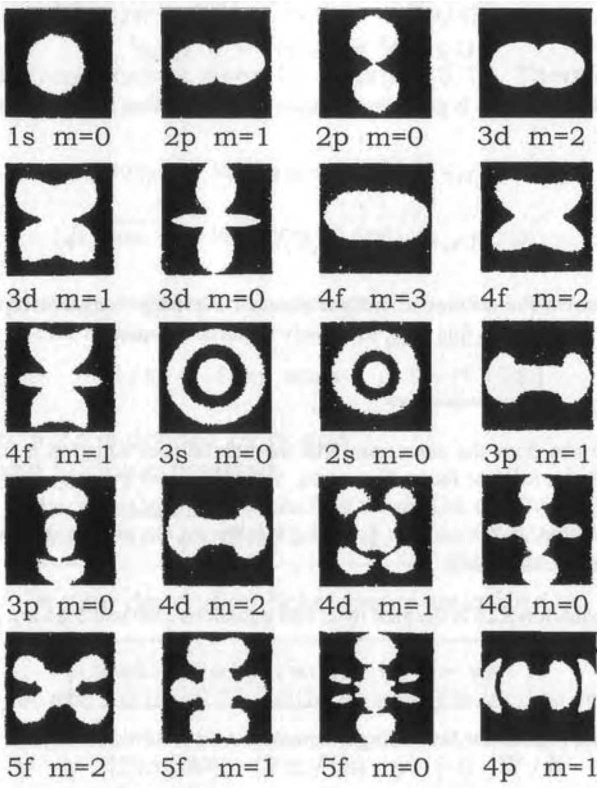

1s m=0 2p m=1 2p m=0 3d m=2

3d m=1 3d m=0 4f m=3 4f m=2

4f m=1 3s m=0 2s m=0 3p m=1

3p m=0 4d m=2 4d m=1 4d m=0

5f m=2 5f m=1 5f m=0 4p m=1

Fig. 5.9: Photographs of models of the probability densities for
different states of the hydrogen atom.
The densities are symmetrical if rotated around the vertical axis<i>

taken from:

<i>: U. Gradmann/H. Wolter: Grundlagen der Atomphysik, AVG, Frankfurt a. M. 1971, P. 190.

5.9 Interpretation of the Schrodinger equation

The interpretation of the Schrodinger equation is still disputed among physicists, because the concept of wave packets contradicts the corpuscular nature of the elementary particles. Further the difficulty is added that wave packets at a closer look never are connected, run apart more or less fast, and really nothing can hinder them doing that. But for a particle the connection represents a physical fact. Then there can be no talk of causality anymore. The monocausal division into two different levels of reality, in a space-timely localization and in an energetic description, does not represent a solution but rather the opposite, the abolition of the so-called dual nature. As has been shown, the potential vortex is able to achieve this with the help of its concentration effect.
But from the introduction of this new field phenomenon arises the necessity to interpret the causes for the calculable and with measuring techniques testable solutions of the Schrodinger equation in a new way. Laws of nature do not know a possibility to choose! If they have been accepted as correct, they necessarily have to be applied.

Three hundred years ago the scholars had an argument, whether a division of physical pheomena, like Newton had proposed it, would be allowed to afterwards investigate them in the laboratory individually and isolated from other influences or if one better should proceed in an integrated manner, like for instance Descartes with his cartesian vortex theory. He imagined the celestial bodies floating in ethereal vortices. One absolutely was aware that the whole had to be more than the sum of every single realizaton, but the since Demokrit discussed vortex idea had to make room for the overwhelming successes of the method of Newton. And this idea after 2100 years was stamped, to in the meantime almost have fallen into oblivion.
Today, where this recipe for success in many areas already hits the limits of the physical possibilities, we should remember the teachings of the ancients and take up again the vortex idea It of course is true that only details are calculable mathematically and that nature, the big whole, stays incalculable, wherein problems can be seen.
If we consider the fundamental field equation 5.7, we find confirmed that actually no mathematician is capable to give a generally valid solution for this four-dimensional partial differential equation. Only restrictive special cases for a harmonic excitation or for certain spatial boundary conditions are calculable. The derived Schrodinger equation is such a case and for us particularly interesting, because it is an eigenvalue equation. The eigenvalues describe in a mathematical manner the with measuring techniques testable structures of the potential vortex .
Other eigenvalue equations are also derivable, like the Klein-Gordon equation or the Lionville equation, which is applied successfully in chaos theories. So our view opens, if chaotic systems like turbulences can be calculated as special cases of the same field equation and should be derivable from this equation.
The in pictures recorded and published structures, which at night should have come into being in corn fields, often look like the eigenvalues of a corresponding equation. The ripe ears thereby lie in clean vortex structures flat on the soil. Possibly potential vortices have charged the ears to such high field strength values that they have been pulled to the soil by the Coulomb forces.

Consequences resulting from the derivation of the Schrodinger equation from the fundamental field equation 5.7:

1. *Any experiment which confirms the Schrödinger equation is with that able to confirm at the same time the existence of the newly discovered potential vortex and the correctness of the field-theoretical approach.*

2. *Properties of the atomic hull and of the atomic nucleus, which can be described with the Schrödinger equation, can as of now be interpreted as an electromagnetic phenomenon.*

3. *There exist no particles or wave packets from matter waves, but only configurations consisting of potential vortices and current eddies.*

4. *There exists no matter! What we call matter is nothing but an electromagnetic state of oscillation of empty space.*

The relation between the energy of oscillation and the mass is described by the relation named after Albert Einstein

$$E = mc^2 \quad . \qquad (6.1 = 5.24)$$

Fig. 6.1: Derivation of the Schrodinger equation, power of proof and consequences

6. Theory of objectivity

6.1 Proof

A new theory only has chances on acknowledgment if it is provable. For that physical phenomena in the sense of the new theory are calculated and independently of this experiments are being carried out. If the calculations are confirmed by reproducible measurement results, then with that the correctness of the approach is proven.
In the here presented case we have chosen the field-theoretical approach instead of the usual quantum physical approach. As a consequence of this we had found as a new phenomenon the vortex of the electric field. With regard to the normally used Maxwell theory this resulted in changed field equations in a dual formulation. If both equations, each of which describes a source-free vortex field, are inserted into each other the result is an only in time and space formulated, generally valid and hence fundamental field equation (5.7, fig. 5.1).
This equation has many special cases; one of them, the Schrodinger equation, could be derived by using an approach which was harmonic in time. We renounced to give special solutions of the Schrodinger equation, because these are printed in numerous text books. On the other hand experiments are known, which are capable to confirm the theoretical solutions and thus to prove the Schrodinger equation. The eigenvalues of the equation describe for instance the shell-shaped structure of the atoms with the by Niels Bohr given radii.
Now this already proven equation was derived from the new field-theoretical approach. Thus for the special case, the area where the Schrodinger equation is valid, the new theory can be said to be proven (fig. 6.1).
We still are not content with that and put another stone on top: we will calculate the quantum properties of the elementary particles for ourselves. These until now have only been measured. Today is merely sought for symmetries and for models of explanation, like e.g. the quark-hypothesis. From a calculation science is miles and miles away. We will compare the calculation results with the measurement values. Then everyone can check and compare for him or herself.
The conditions in an elementary particle are completely different. Here it concerns the vortex itself, whereas the model of the atom merely describes vortex properties, so-called actions at a distance. The differences in size and distances for an atom lie more than five powers of ten over those of a particle!
Here a new problem of causality comes to light, at which we now must have a critical look: the question of the by Einstein postulated constancy and universality of the speed of light. Seen from a relativistic and subjective point of view of an observer, Einstein by all means may be right. But may such a theory be generalized? How are the measurements concerning the speed of light and the relativity of space and time to be judged when looking at them objectively?
The current measurements of speeds faster than light speak a clear language and represent a challenge (fig. 3.1, violation of the principle of causality no. 5).

The electron as a spherical capacitor (see fig. 4.3):

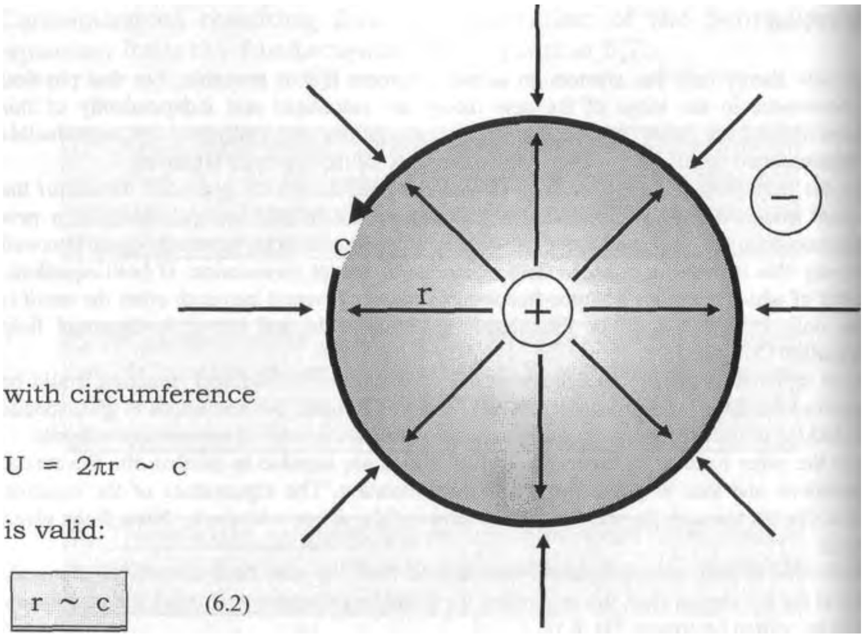

with circumference

$U = 2\pi r \sim c$

is valid:

$$\boxed{r \sim c} \quad (6.2)$$

field theoretical approach (vortex particles):
The amount of energy bound in the inside of the particle is identical with the free and measurable amount of energy on the outside of the particle.

(If the number of particles is left unchanged):
In an isolated system the sum of the energy is constant

(particle = electromagnetic vortex)
Energy is a state description of electromagnetism.

Fig. 6.2: Derivation of the law of conservation of energy

6.2 Law of conservation of energy

Let the starting-point for our considerations be the electromagnetic wave in a particle-free vacuum. Here no vortices appear, so that the plane wave can propagate undamped with the speed of light, and in this way a transport of energy takes place. Electric and magnetic energy each are the same magnitude.

Let's now imagine the symmetry is disturbed as the wave is "slowed down" on one side. As a possible result the wave rolls up to a spherical vortex.

As we will see such a process is possible, for instance at impact on a strong field. Thus part of the energy is bound in the inside. This part from now on withdraws itself from every possibility to measure it. We can only measure the second part of the field energy, with which the particle interacts with its neighbourhood.

We can assume that:

The amount of energy bound in the inside of the particle is identical with the free and measurable amount of energy on the outside of the particle.

The same energy W_e = 0,51 MeV, we attribute to the electron for reason of its mass with the help of the Einstein relation (6.1), is also bound in its inside. This conclusion is also applicable to other elementary particles and with that to all matter.

We here again recognize the principle of the duality between the to the outside striving eddy current in the inside of the elementary vortex and the concentrating potential vortex on the outside. Thus also seen energetically both are of the same magnitude. Whereas in the case of the electromagnetic wave it concerns a symmetrical oscillation around "zero", by the process of quantization, by the rolling up to a spherical vortex, there forms an energetic state of space different from zero. The order of magnitude is determined by the number of elementary vortices, of which the particles and all matter consist.

Anti-matter forms the opposite energetic state and this again is for the particles of matter available in their inside in a bound form.

As long as we do not artificially produce new elementary vortices and thus keep the number of available vortices constant, the energetic state will not change, or as it is formulated in text books:

In an isolated system the sum of the energy is constant.

THE law of conservation of energy is not an axiom, but follows without compulsion from the vortex theory. It is not elementary, but a consistently derivable consequence of the field-theoretical approach, according to which solely the field acts as cause for all other physical phenomena, also for the conservation of energy! Because the cause of it is the electromagnetic field, the following has to hold:

Energy is a state description of electromagnetism.

Now we finally can explain why energy can be converted. Different forms of energy only are different forms of formation of the same phenomenon!

Of course this statement of the field-theoretical approach does not yet explain what, for instance, the temperature has to do with electromagnetism. I ask for some patience; no question will be left unanswered.

From

$$\boxed{c \sim r} \tag{6.2}$$

follows:

The speed of light determines the size of the elementary particles.

Energy of a capacitor:

$$W = Q^2/C, \tag{6.3}$$

written down for the electron (with the Einstein relation):

$$W_e = e^2/C_e = m_e c^2 = 0{,}51\ \text{MeV} \tag{6.1}$$

Capacity of a spherical capacitor:

$$C_e = \varepsilon_0 \cdot 4\pi r_e \tag{6.4}$$

"classical" radius of the electron[i] is:

$$r_e = e^2/\varepsilon_0 \cdot 4\pi \cdot W_e \tag{6.5}$$

$$r_e = 2{,}82 \bullet 10^{-15}\ \text{m} \tag{6.6}$$

in the case of Kuchling[ii] the radius of the electron is:

$$r_e = 1{,}41 \bullet 10^{-15}\ \text{m}\ . \tag{6.7}$$

Fig. 6.3: Calculation of the radius of the electron.

[i] Mende, Simon: Physik, Gl. 10.39, VEB-Leipzig, 4. Aufl.

[ii] Kuchling: Physik, Gl. At4, VEB-Leipzig, bis einschl. 11. Auflage 1974

6.3 Radius of the electron

For the crucial process, in which the electromagnetic wave rolls up to a vortex, it is for reasons of continuity to be expected that the velocity of propagation remains equal that thus for the vortex oscillation exactly like for the electromagnetic wave the speed of light is determining. The direction of propagation in the case of the vortex takes place perpcndicular to the in fig. 6.2 shown field direction of the electric field strength. Not even in that both field-phenomena differ.

Summarizing: the propagation takes place with the speed of light c along a circular path with the perimeter $U = 2\pi r$. Therefore holds:

$$\boxed{c \sim r} \quad (6.2)$$

According to this equation the radius and with that the size of the electron is determined by the speed of light. Therefore the question of the size of the electron is raised.

The energy interpretation predicts that for the theoretical case of a change of size the energy density in the inside of the particle is influenced that however the quantity of the included energy remains unchanged. We therefore can further proceed from the assumption that the bound amount of energy is independent of the size of the particle!

Consequently for the elementary quantum the energy W_e = 0,51 MeV is assumed, which it has acccording to the Einstein relation $W_e = m_e c^2$. For the electron of mass m_e the with measuring techniques determined value is inserted.

The spherical electrode of a spherical capacitor with the above given energy W_e (according to eq. 6.1) and the capacity C_e (according to equation 6.4, fig. 6.3) represents a very realistic model of the negatively charged particle.

In this manner the classical radius of the electron is calculated to be[i]: $r_e = 2{,}82 \cdot 10^{-15}$ m.

But in the case of Kuchling it only is half this size[ii], what according to equation 6.2 would mean that in the case of Kuchling the light would be on the way only half this fast [iii]. Therefore if one is careful, one prefers to be silent concerning this delicate theme and if one is honest, one admits not to know anything exact.

Not only the electron but also all the other elementary particles are according to the field-theoretical approach formed from concentrated potential vortices. For these equation 6.2 hence has to hold in the same manner, so that more generalized we can conclude:

The speed of light determines the size of the elementary particles.

This statement is incompatible with the assumption of a constant speed of light! Because then all elementary particles would have identical size. As is known, however, are the building parts of the atomic nucleus, the protons and neutrons very much smaller than individual electrons. The constancy of the speed of light is to be questioned. This question is of such an elementary importance that we are not content with these considerations and in addition undertake a mathematical derivation in the sense of the field approach.

<iii>: Difference = Thomas factor

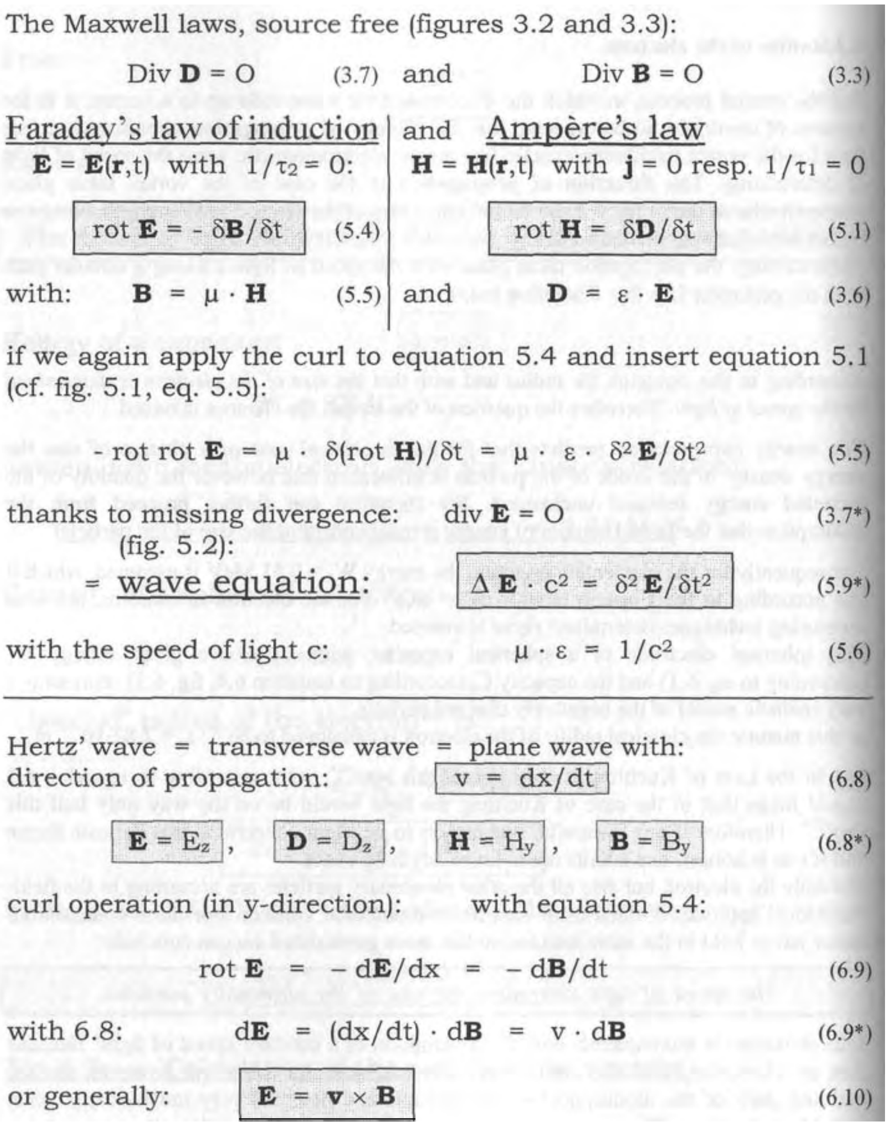

The Maxwell laws, source free (figures 3.2 and 3.3):

$\mathrm{Div}\,\mathbf{D} = 0$ (3.7) and $\mathrm{Div}\,\mathbf{B} = 0$ (3.3)

Faraday's law of induction and Ampère's law

$\mathbf{E} = \mathbf{E}(\mathbf{r},t)$ with $1/\tau_2 = 0$: | $\mathbf{H} = \mathbf{H}(\mathbf{r},t)$ with $\mathbf{j} = 0$ resp. $1/\tau_1 = 0$

$\boxed{\mathrm{rot}\,\mathbf{E} = -\,\delta\mathbf{B}/\delta t}$ (5.4) | $\boxed{\mathrm{rot}\,\mathbf{H} = \delta\mathbf{D}/\delta t}$ (5.1)

with: $\mathbf{B} = \mu \cdot \mathbf{H}$ (5.5) and $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (3.6)

if we again apply the curl to equation 5.4 and insert equation 5.1 (cf. fig. 5.1, eq. 5.5):

$$-\,\mathrm{rot}\,\mathrm{rot}\,\mathbf{E} = \mu \cdot \delta(\mathrm{rot}\,\mathbf{H})/\delta t = \mu \cdot \varepsilon \cdot \delta^2\mathbf{E}/\delta t^2 \qquad (5.5)$$

thanks to missing divergence: $\mathrm{div}\,\mathbf{E} = 0$ (3.7*)
(fig. 5.2):

= wave equation: $\boxed{\Delta\,\mathbf{E} \cdot c^2 = \delta^2\mathbf{E}/\delta t^2}$ (5.9*)

with the speed of light c: $\mu \cdot \varepsilon = 1/c^2$ (5.6)

Hertz' wave = transverse wave = plane wave with:
direction of propagation: $\boxed{v = dx/dt}$ (6.8)

$\boxed{\mathbf{E} = E_z}$, $\boxed{\mathbf{D} = D_z}$, $\boxed{\mathbf{H} = H_y}$, $\boxed{\mathbf{B} = B_y}$ (6.8*)

curl operation (in y-direction): with equation 5.4:

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{E}/dx = -\,d\mathbf{B}/dt \qquad (6.9)$$

with 6.8: $d\mathbf{E} = (dx/dt) \cdot d\mathbf{B} = v \cdot d\mathbf{B}$ (6.9*)

or generally: $\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (6.10)

Fig. 6.4: Derivation of the laws of transformation

<i>: Prof. G. Bosse in his text book in reversed direction derives the Faraday law of induction from the law of transformation 6.10, which he again derives from considerations about the Lorentz force. G. Bosse, Grundlagen der Elektrotechnik II, BI 183, Hochschultaschenbucher-Verlag, Mannheim 1967

6.4 The Maxwell field equations

The laws of transformation of the electromagnetic field shall form the starting-point for the coming up considerations. To exclude any doubts with regard to the interpretation, the equations will be derived from the Maxwell laws under the assumption that no sources or charge carriers are present (fig. 3.2 and 3.3) and as a consequence no current density ($j = 0$) is to be expected.
This corresponds to the vanishing of the time independent terms, which consequently are responsible for the occurring of force effects like e.g. the Lorentz force. Only at the end of this derivation we can understand the sense of this assumption (with $1/\tau_1 = 0$ and $1/\tau_2 = 0$). The procedure at first corresponds to that of fig. 5.1. Here the fundamental field equation had been derived from Faraday's law of induction and Ampere's law. With the assumptions made this time the in fig. 5.2 treated undamped wave equation is left (5.9, here 5.9*). Whom the derivation is still present can go in at this point.
In a sufficiently great distance from the source we are dealing with a plane wave, in which the field factors only depend on the direction of propagation x. The Hertz' wave is a transverse wave, in which the field pointers oscillate perpendicular to the direction of propagation and in addition stand perpendicular to each other:

$$\boxed{v = dx/dt} \quad (6.8), \qquad \boxed{\mathbf{E} = E_z, \quad \mathbf{D} = D_z, \quad \mathbf{H} = H_y, \quad \mathbf{B} = B_y} \quad (6.8^*)$$

The curl, applied to the electric field pointer, itself points in the y-direction:
rot E = - dE/dx . This for the transverse wave carried out curl operation is now compared with Faraday's law of induction (5.4):

$$\mathrm{rot}\, E = -dE/dx = -dB/dt \qquad (6.9)$$

The relation won in a mathematical way, with the speed fixed by (6.8), reads:

$$dE = (dx/dt) \cdot dB = v * dB \qquad (6.9^*)$$

The result of this derivation at first only is valid for the introduced simplification, for instance for the case of the transverse electromagnetic wave. Better known is apart from that the generalized formulation, which among others by G. Bosse[i] is called law of transformation.

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \qquad (6.10)$$

With Ampere's law (5.1) we now should proceed in an analogous manner. The result is:

$$\boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (6.10^*)$$

This equation 6.10* is given among others by Simonyi[ii]. Now that we know, under which circumstances these equations of transformation can be derived from the Maxwell equations, the actual work can start.

<ii>: K. Simonyi, Theoretische Elektrotechnik, 7. Auflage VEB Verlag Berlin 1979. pp. 921 - 924; In addition see chapter 27.8 in part 3 of this book.

Laws of transformation:

$$\mathbf{E} = \mathbf{v} \times \mathbf{B} = \mathbf{v} \times \mu \cdot \mathbf{H} \quad (6.10)$$

$$\mathbf{H} = -\mathbf{v} \times \mathbf{D} = -\mathbf{v} \times \varepsilon \cdot \mathbf{E} \quad (6.10^*)$$

and

> We experience the magnetic field as an electric field and the electric field as a magnetic field simply and solely for reason of the same relative motion!

The component of the direction of motion perpendicular to the area defined by the field pointers:

$$E = v \cdot \mu \cdot H \quad (6.11)$$

and

$$H = -v \cdot \varepsilon \cdot E \quad (6.11^*)$$

with the relations of material:

$$B = \mu \cdot H \quad (3.5)$$

$$D = \varepsilon \cdot E \quad (3.6)$$

with the speed of light:

$$c = 1/\sqrt{\mu \cdot \varepsilon} \quad (5.6)$$

additional field:

$$E_z = -v^2 \cdot \mu \cdot \varepsilon \cdot E = -(v^2/c^2) \cdot E \quad (6.12)$$

resp.

$$H_z = -v^2 \cdot \varepsilon \cdot \mu \cdot H = -(v^2/c^2) \cdot H \quad (6.12^*)$$

basic field: E resp. H (at $v = O$)

measurable overall field: (if $v \neq O$)

$$E_o = E + E_z \quad (6.13)$$

resp.

$$Ho = H + H_z \quad (6.13^*)$$

Fig. 6.5: Properties of transformation of the electromagnetic field.

6.5 Equations of transformation

As a consequence of the in fig. 6.5 again written down laws of transformation of the electromagnetic field (6.10 and 6.10*) magnetic phenomena can be traced back to electric phenomena and vice versa. The mathematical formulation reveals us the two sides of the same medal and points to a perfect duality between both fields and their factors of description.

Because a way exists, as is shown here, in which the equations of transformation can be derived from the Maxwell field equations, the same generally valid and extensive importance should be attributed to them. They can with the same right be called the foundation of electromagnetism. Wherein does lie its message for physics, the always curious researcher will ask? For that the relations of material 3.5 and 3.6 are completed:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mu \cdot \mathbf{H}} \quad (6.10) \qquad \text{und} \qquad \boxed{\mathbf{H} = -\mathbf{v} \times \varepsilon \cdot \mathbf{E}} \, . \quad (6.10^*)$$

The here presented equations state, that we measure an electric field strength E, if we are moving with regard to a magnetic field H with the speed v and vice versa. The electric and the magnetic field therefore prove to be an experience of the observing person and we can say:

We experience the magnetic field as electric field and the electric field as magnetic field simply and solely for reason of the relative motion!

Let's assume, v is the component of the relative velocity (6.8), which stands perpendicular to the area defined by the field pointers (6.8*), then the equations of transformation (6.9* with 3.5) now read:

$$E = v \cdot \mu \cdot H \quad (6.11) \qquad \text{and} \qquad H = -v \cdot \varepsilon \cdot E \, . \quad (6.11^*)$$

If we are moving with the velocity v in a basic field which is present with the field strength E, then according to equation 6.11* we observe a magnetic field, which again according to equation 6.11 is to be interpreted as an additional electric field E_z:

$$E_z = -v^2 \cdot \mu \cdot \varepsilon \cdot E = -(v^2/c^2) \cdot E \quad (6.12)$$

In duality equation 6.11 inserted into equation 6.11* provides for the magnetic field strength a corresponding additional field H_z:

$$H_z = -v^2 \cdot \mu \cdot \varepsilon \cdot H = -(v^2/c^2) \cdot H \quad (6.12^*)$$

We obviously owe the measurable overlap fields in a laboratory simply and solely to the relative velocity v with which the laboratory is moving. But now we must pay attention to the fact that a terrestrial laboratory rotates along with the earth, that the earth orbits the sun and the sun again rotates around the centre of the milky way. Eventually the whole milky way is on the way in the cosmos with a galactic, for us hardly understandable speed. If we further take into consideration that for every subsystem an additional field occurs as a consequence of the relative motion with regard to the super ordinate system, then one additonal field follows after the next and overlaps this one.

Let's imagine, the relative velocity could be reduced towards zero - and maybe we are moving around such a cosmic point - then here no overlapping field would be measurable.

<*>: A derivation using vectors is written in chapter 28 (part 3).

Additional field (from fig. 6.5):

(6.12)
(6.12*)

$$\boxed{\begin{aligned} E_z &= -(v^2/c^2)\cdot E \\ H_z &= -(v^2/c^2)\cdot H \end{aligned}}$$ and

Superposition of the fields:
The additional field (E_z resp. H_z) overlaps the basic field
(E resp. H) to produce the measurable overall field (E_0 resp. H_0):

(6.13)

$$\boxed{\begin{aligned} E_0 &= E + E_z = E\cdot(1 - v^2/c^2) \\ H_0 &= H + H_z = H\cdot(1 - v^2/c^2) \end{aligned}}$$

(6.13*)

transformed:

$$(1 - \frac{v^2}{c^2}) = \frac{E_0}{E} = \frac{H_0}{H}$$

(6.14)

for the Lorentz contraction holds apart from that:

(6.14*)

$$(1 - \frac{v^2}{c^2}) = (\frac{1}{l_0})^2$$

From the comparison
Follows

(6.14**)

$$\frac{E_0}{E} = \frac{H_0}{H} = (\frac{1}{l_0})^2$$

the proportionality:

$$\boxed{E, H \sim 1/l^2} \text{ and } \boxed{E_0, H_0 \sim 1/l_0^2}$$ (6.15)

Fig. 6.6: The field dependency of the Lorentz contraction

6.6 Field overlap

Field vectors can be superpositioned. In this manner the additional field E_z resp. H_z which depends on the velocity, according to equation 6.10, overlaps the respective basic field (E resp. H) to produce the measurable overall field (E_0 resp. H_o):

$$\boxed{E_0 = E + E_z = E \cdot (1 - v^2/c^2)} \quad (6.13)$$

$$\boxed{H_0 = H + H_z = H \cdot (1 - v^2/c^2)} \quad (6.13^*)$$

In the result something surprising the factor $(l\text{-}v^2/c^2)$ appears, which is well-known from the special theory of relativity and for instance appears in the Lorentz contraction.

If we rewrite both equations for the characteristic factor and compare with the in a purely mathematical way, over the Lorentz transformation, won length contraction $(1 - v^2/c^2) = (l/l_0)^2$, then it becomes clear that the Lorentz contraction physically seen should have its cause in the changed field conditions which a with relativistic speed moving body finds with regard to a resting body.

$$1 - \frac{v^2}{c^2} = \frac{E_0}{E} = \frac{H_0}{H} \overset{!}{=} \left(\frac{1}{l_0}\right)^2 \quad (6.14)$$

The equation is a compulsionless consequence of known physical laws. In this derivation actually no new factor was introduced and nevertheless a completely new picture for the natural scientific reality results<i>.

In our observer system, where the field E_o exists, a rule has its proper length l_0. In another system, which is moving with the speed v relative to the observer, as a consequence of the here prevailing field E the corresponding rule has a length 1. In which relation the factors stand to each other, is described by equation 6.14. Accordingly the following proportionality holds:

$$\boxed{E, H \sim 1/l^2} \text{ and } \boxed{E_0, H_0 \sim 1/l_0^2} \quad (6.15)$$

If we are exterior to a very fast moving body with velocity v, we immediately can observe how this body for reason of its relative velocity experiences the calculated additional field and in this way experiences a length contraction. If the observer is moving along with the body, then he purely subjective seen doesn't detect a length contraction, because he himself and his entire measuring technique is subjected to the same length contraction. From the axiomatic approach what would be, if the field, which itself only represents an experience, would determine perceptible space and its dimensions, quickly a fundamental realization can develop if the described experiences should coincide with real observations.

<i> : Because in this point of view the subjective status of the observer is determining, it is not completely impossible that there is an error in the interpretation of the equations of transformation (6.10 and 6.10*). But because we started from the same point of view of the observer for the derivation of the length contraction from the Lorentz transformation, here the same error is to be expected. In putting both results equal (6.14), a like constituted error on both sides will cancel out in any case and the result stays above all doubts!

$$\boxed{E, H \sim 1/l^2} \quad \text{and} \quad \boxed{E_0, H_0 \sim 1/l_0^2} \tag{6.15}$$

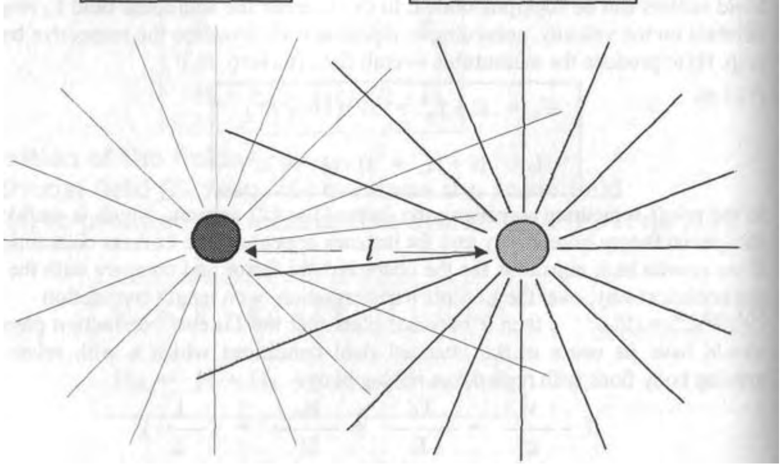

(Model):

Two particles of matter each in the field of the other particle.

Two elementary particles or two accumulations of matter consisting of these are able to reduce the distance to each other for reason of their fields, which we interpret as a force of attraction.

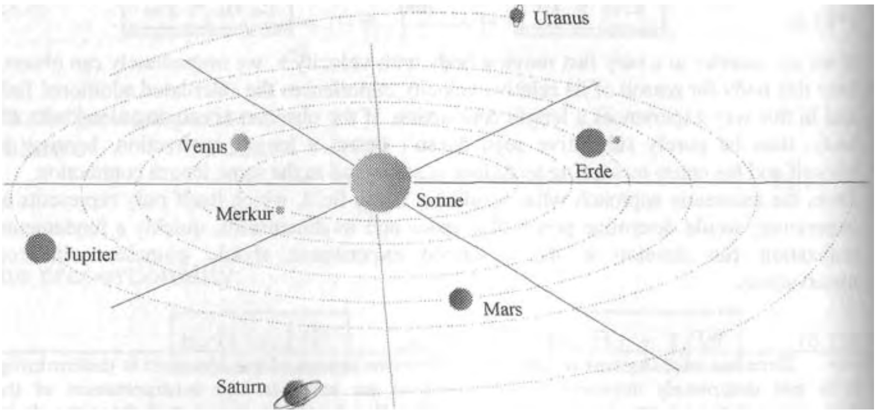
Uranus
Venus
Erde
Sonne
Merkur
Jupiter
Mars
Saturn

B: (Example): The orbits of the planets in the field of the sun.

Fig. 6.7: The influence of the field on interactions.

6.7 Field dependent curvature of space

Let's assume, an accumulation of matter, as big as our earth, wanted to fly past the sun in the distance earth-sun. But it would not succeed. Because the fields arising from the sun decreases with increasing distance and according to equation 6.15 as a consequence the size of the particles of matter increases. The planet hence is more strongly contracted on its side turned towards the sun, as on the turned away "night side". It bends towards the sun and its flight path becomes a circular path around the sun. That is the interaction known as gravitation!

To an earth inhabitant this curvature reveals itself merely in the observation that the duration of sunshine at daytime is longer, than it would be expected to be under the assumption of the earth as a homogeneous sphere. In this context one willingly speaks of a curvature of space. Actually it is a curvature of matter under the influence of the field dependent length contraction.

Exactly this contraction the planets owe their circular orbits around the sun and by no means the equilibrium of forces between the force of attraction and the centrifugal force (fig. 6.7 B). It obviously is a fundamental mistake to think that gravitation would causally be connected with a force effect!

If, in this context, we speak of a force of attraction for the sake of our subjective observation, then we must realize that it merely can concern an auxiliary term founded in usefulness.

A thought experiment should bring us clarity (fig. 6.7 A). The field, which surrounds every particle of matter, reaches till infinity but becomes less effective with increasing distance. If the distance between two particles is 1, then one particle is in the field of the other particle. As a consequence of the field the length 1 reduces and in this way the size determining field increases, which again leads to a further reduction of length etc. As a consequence it can be observed that both particles are moving towards each other. We speak of a force of attraction, because we can't register the influence of the field with our senses.

In this way the consistent result that we and our environment at daytime must be smaller than in the night will as well remain hidden. We experience the effect only indirectly as gravitational pull of the earth.

Because we don't see the cause of a subjectively observed force effect, for the electromagnetic interaction, just as for the gravitation, the field dependency of the length contraction will be responsible. Hence the following conclusion holds for both interactions equally way.

Two elementary particles or two accumulations of matter consisting of these are able to reduce the distance to each other for reason of their fields, which we interpret as a force of attraction.

Now the question still is open, why gravitation only knows forces of attraction, whereas the electromagnetic interaction also permits forces of repulsion and which are the causal fields for each.

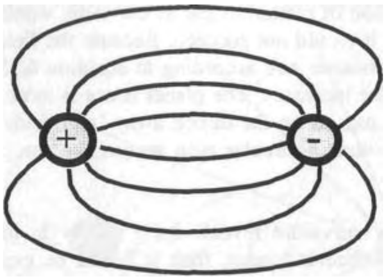

A: The field lines of the E-field for unlike charged particles

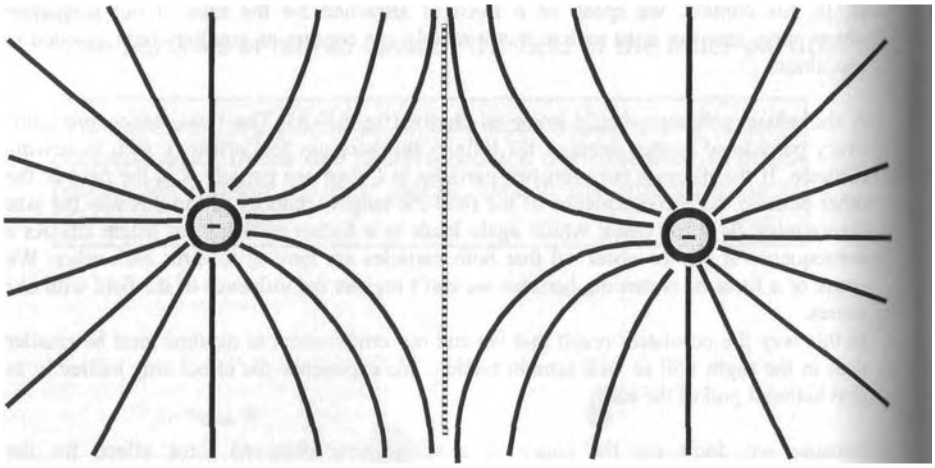

B: The field lines of the E-field for equal charged particles

> The electromagnetic interaction of a particle is a result of the influence of the open field lines arising from it on the dimensions of space.

Fig. 6.8: The influence of the open field lines of the E-field

6.8 Electromagnetic interaction

A convincing answer to the open question provides us the analysis of the course of the field lines, on the one hand for charged particles and on the other hand for uncharged particles, which do not participate in the electromagnetic interaction. If at first we consider electrically charged particles, like e.g. electrons, protons or ions. Then all in common is that the towards infinity running field lines of the electric field are open. With this field the particle is able to interact with its environment. We measure a charge and an electromagnetic force effect. In the case of unequal charges, as is well-known, a field amplification and attractive acting forces are observed whereas for equal charges a field reduction results and repulsion is observed.

If we make a connection between the field conditions and the electromagnetic interaction in the sense of the proportionality (6.15), then the particle in reality is able to influence the distance to other particles merely with the help of its electric field. For unequal charges a compression of field lines arises, in which one particle stays in the focussed field of the other and vice versa. In this way a contraction of all lengths occurs and the observable attraction happens (fig. 6.8 A).
For equal charges the opposite case is present, in which even a local field freedom can occur (fig. 6.8 B). If the field tends towards zero on the dashed line, then the distance will go to infinity (according to eq. 6.15). Consequently, the observable effect that both bodies go away from each other, will reach to infinity.

Actually the electromagnetic interaction proves to be a result of the field dependent length contraction.

The electromagnetic interaction of a particle is a result of the influence of the open field lines arising from it on the dimensions of the space, in which it is.

It is important that the field lines are open, for which reason they are bent away from like charges and are directed towards unlike charges. Subjectively seen we find out that as a consequence of the field reduction repulsive force effects and as a consequence of the field compression attractive acting force effects are observed (fig. 6.8).
The consequence of is every electric field is, as is well-known, a magnetic field standing perpendicular on it. The field lines of the magnetic field run parallel to the surface of the particle and have a closed course (fig. 6.9 A)!
Therefore no magnetic poles form, which would be measurable. Seen from the outside the particle behaves neutral magnetically seen, because of the closed course of the field lines. An artificial field reduction and as a consequence observable forces of repulsion, like in the case of the electromagnetic interaction, hence in principle are impossible.

The effect of the magnetic field thus is limited to a geometrical manipulation of the environment, namely the curvature of space, with which we have founded the phenomenon of the attraction of masses and of the gravitation.

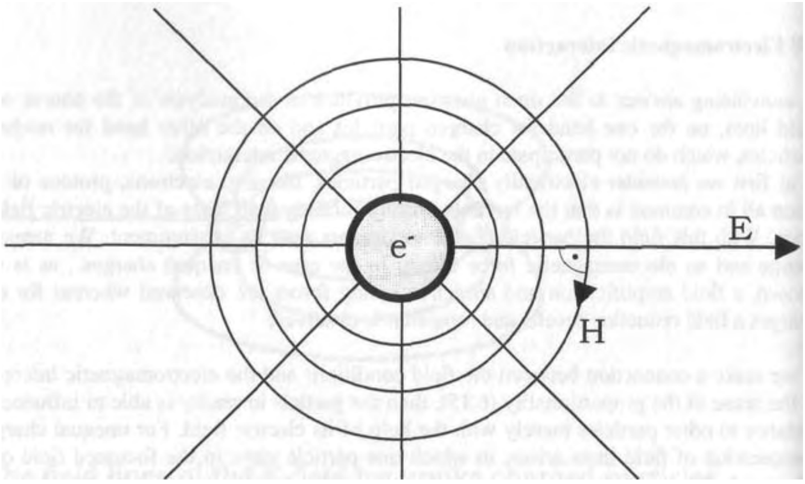

A: The open field lines of the E-field and the closed field lines of the H-field of an electrically charged particle (e.g. e-)

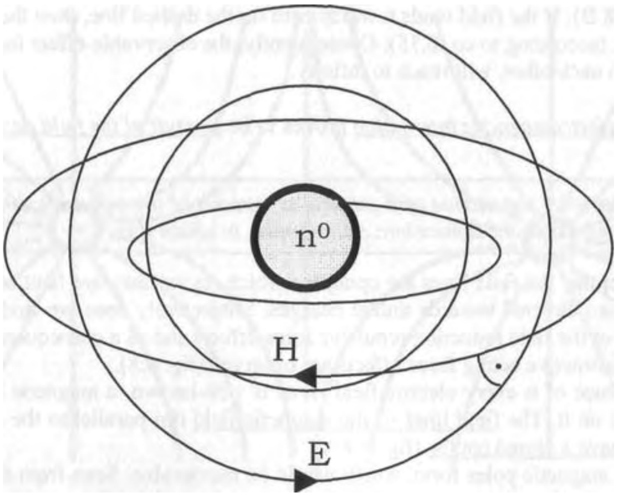

B: The closed field lines of the E-field and H-field of an electrically uncharged particle (e.g. of the neutron n°).

Gravitation is a result of the influence of the field lines with a closed course running parallel to the surface of the particles on the dimensions of the space, in which they are.

Fig. 6.9: The influence of the closed field lines of the H-field.

6.9 Gravitation

For uncharged, neutral particles (neutron, atom, molecule etc.) both the magnetic and the perpendicular on them standing electric field lines have a closed course. Now both run parallel to the surface of the particle (fig. 6.9 B).
As is said, the density of field lines with a closed course can't be influenced from the outside. If we approach a particle, the consequence of an increase of the density without exception is a decrease of the linear measures and thus a larger force of attraction. For this case of field lines with a closed course, for which in general it doesn't give a field attenuation and no forces of repulsion, there holds:

Gravitation is a result of the influence of the field lines with a closed course running parallel to the surface of the particles on the dimensions of the space, in which they are.

Both interactions logically have an infinite range. Both form a whole in the influence of the fields on the size conditions.
It surely is of the greatest importance that for this derivation of the field dependency of the Lorentz contraction from the known equations of transformation of the electromagnetic field we could do completely without the introduction of new factors of description or neglects.
Solely by consistent derivation and interpretation of the result the unification already has suceeded and the electromagnetic interaction and the gravitation could, with the derived field dependent Lorentz contraction, be traced back to a single basic phenomenon. Doing so we have to pay attention to the fact that the observer is subjected to the same Lorentz contraction as his measuring technique and therefore he can't see the field dependency at all. Merely as being an exterior observer it in rare cases will be possible to him to see the curvature of space in the presence of strong fields.

From this for an astronaut practical consequences result. If he namely would land on Jupiter, he would think flat hills to be gigantic mountains, that small he would be! Vice versa if he landed on the moon, high mountains would appear to be insignificant hills, not because of wrong altitude readings of the terrestrial mission control and measurement centre, but only because of his own body size. The astronauts of the Apollo missions were not prepared for this circumstance and after their landing on the moon were completely surprised, how little validity learned textbook physics has, hardly has one left the earth. They have brought photographs with them which prove the Lorentz contraction to depend on the field and therefore on gravitation.

The fact that force effects should arise from the interactions is an auxiliary concept and auxiliary description of the observing person founded in pure usefulness. The Lorentz force therefore shouldn't be regarded as cause anymore. It actually appears only as property of the field factors. Seen this way it only would be consistent to do without space charges and currents as a result of moving charges and to assume a source-free and quanta-free field description (fig. 6.4: $j = 0$).
From an unified theory it is demanded that it besides the electromagnetic interaction and the gravitation also is able to integrate the strong and the weak interaction. We will also solve this problem.

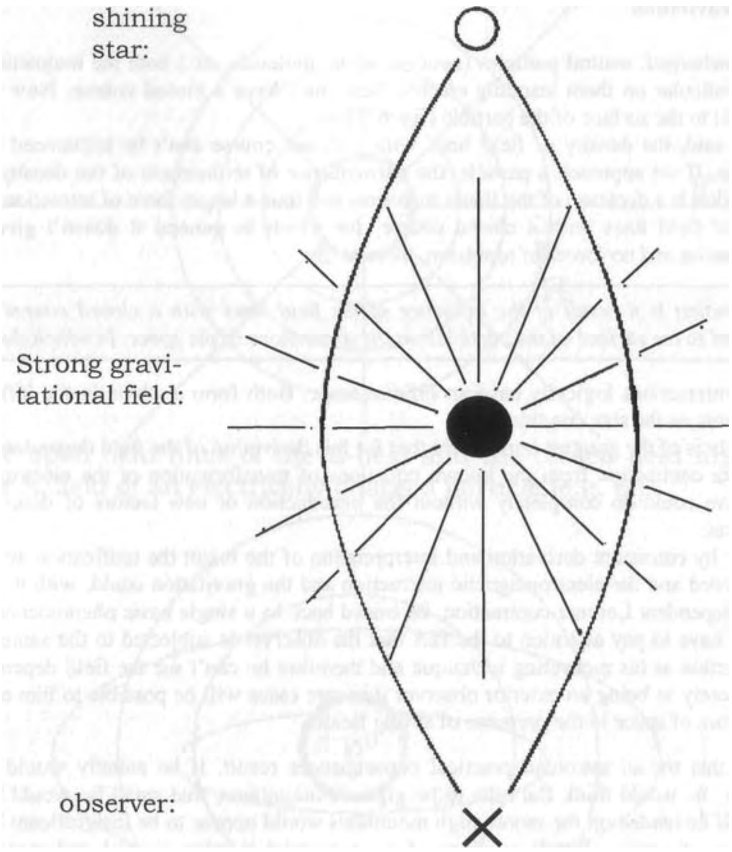
shining star:
Strong gravitational field:
observer:

Fig. 6.10: Diversion of the light by a strong gravitational field.

Speed of light of the wave: $c = \lambda * f$ (6.16)

For the wavelength λ holds (because of eq. 6.15): $E, H \sim 1/\lambda^2$

From equation (6.16) follows (with f = constant):

$$E \sim 1/c^2 \quad , \quad H \sim 1/c^2 \tag{6.17}$$

The speed of light depends on the field!

6.10 Field dependent speed of light

But not only matter is bent towards a gravitational field. If we only think of the much cited phenomenon that the ray of light of a star is diverted towards the sun, if it passes very close to the sun on its way to us, like this has been observed for the first time during an eclipse of the sun in 1919 (fig. 6.10).

Quite obviously the field of the sun also slows down the speed of light. On the side of the ray of light which is turned towards the sun, the field is somewhat larger and the speed of light correspondingly is slower than on the side which is turned away, and with that the ray of light changes its direction in the observable manner. Exactly this relation willingly is interpreted as a consequence of a curvature of space.

The extremely strong field of a black hole can divert the light down to a circular path, in order to in this way catch and bind it. The light now orbits the black hole like planets the sun.

At this point the open-minded reader already might have tapped the confirmation of the proportionality 6.2 ($c \sim r$), which has been derived from the vortex model (fig. 6.2).

The sceptic is offered still another derivation: for the borderline case that the relative velocity v tends towards the speed of light c (fig. 6.6), according to equation 6.13 the measurable overall field E_o (and also H_o) will go to zero and equation 6.12, with E_z. $= - E$ (and $H_z = - H$), will again turn into the wave equation (5.9*) after double differentiation (fig. 6.4).

The speed $v = c$ so to speak forms the escape velocity, with which the electromagnetic wave runs away from the cosmic field. Under these circumstances of course neither an attraction of masses nor an electromagnetic interaction can be exerted on the wave.

If E_0 goes to zero at the same time l_0 tends to infinity (equation 6.15, fig. 6.6): i.e. the wave spreads all through space. This result entirely corresponds to the observations and experiences.

For the wave length λ and in the end for the velocity of propagation c only the self-field of the wave E resp. H is responsible. Because of

$$\boxed{c = \lambda \cdot f} \quad (6.16)$$

and the proportionality from equation 6.15: $E, H \sim 1/\lambda^2$ (6.17*)

obtain the new relation: $\boxed{E, H \sim 1/c^2}$ (6.17)

If the speed of light in the presence of matter decreases, then we now also know why. It is the field, which surrounds matter, that slows down the speed of light. Therefore a gravitational field is able to divert a ray of light in the same manner as matter which flies past. Finally moves the speed of light in the proportionality 6.17 to the place of the linear measure (in 6.15).

But if the rule fails one will try to replace by an optical measurement arrangement. In this manner the field dependency of the Lorentz contraction should be measurable; but it isn't!

From the comparison of the derived proportionalities:

$$E, H \sim \overset{(6.15):}{\frac{1}{l^2}} \sim \overset{(6.17)}{\frac{1}{c^2}}$$

follows:

$$\underline{l \sim c} \tag{6.18}$$

The speed of light is proportional to the measurement path.

The variable speed of light is being measured with itself.

The result at all events is a constant value.

The constancy of the speed of light is based on a measurement which is faulty from the principle!

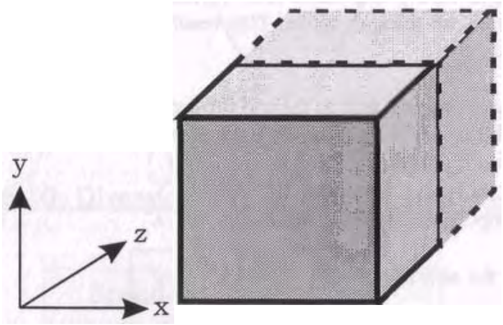

Because of c ~ r: physical length contraction

Fig. 6.11: Derivation of the length contraction (field dependent Lorentz contraction)

6.11 Universality

Why can't the rule be replaced by an optical measurement arrangement? The crucial indication provides the comparison of both derived proportionalities 6.15 and 6.17. According to them holds the same field dependency for both the Lorentz contraction and the speed of light:

$1/l^2 \sim 1/c^2$ or $\boxed{1 \sim c}$ (6.18)

If the rule has proved to be useless, then we will experience the same disaster if we measure optically, i.e. with the speed of light.
Obviously both, the length 1 and the speed of light c as a length per unit of time, depend in the same manner on the respective local field strength. On the one hand do both measuring methods lead to the same result; but on the other hand will anything which can't be measured with one method, neither be measured with the other method.
If now the speed of light is being measured optically, then the measurement path will be proportional to the speed of light and as a result will the unknown factor be measured with itself. The result of this measurement, which is faulty from the principle, at all events is a constant value, because here two variables which stand in direct proportionality to each other are related to each other.
Was the famous experiment of Michelson and Morley unnecessary, the result trivial? And how does it stand about the postulate of the universality of the speed of light?
If we for that consider a cube (fig. 6.11). And we assume that the speed of light is a vectorial quantity, which in our experiment is for instance in one direction twice as large, as in the direction of the other two space axes. By means of the mentioned influence of the speed of light on the spatial length is, as a consistent consequence, the cube along this edge pulled apart to a cuboid. We however register this spatial body with our eyes, which is with the speed of light and that has increased proportionally to the length of the edges, for which reason we as subjective observer still see a cube in front of us and not a cuboid.
If we trust an apparent objective measurement more than our sense organ and measure the three lengths of the edges of the cuboid again with a rule then we get three times the same length, which is a cube.
We probably are dealing with an optical deception using the true meaning of the word.

If the by Einstein postulated universality and constancy of the speed of light in reality doesn't exist at all, we in no way would be capable to register this; neither to observe nor to measure it!
The Galilean theorem of the addition of speeds objectively seen still is valid, even if the fact that the speed of light apparently is independent of the speed of the source pretends us the opposite.
If for instance a light source is moved towards a receiving device or away from it, then the speeds will overlap, like for the passenger, who marches in a driving train against or in the driving direction through the corridor. For the ray of light also the fields, which influence the speed of light and the measurement equipment, overlap. As a consequence will a measuring technician, who himself is exposed to this overlapping field, always observe and "measure" the identical speed of light. The observer as a result imagines, there is an universality of the speed of light.

The field takes over the function of the aether.

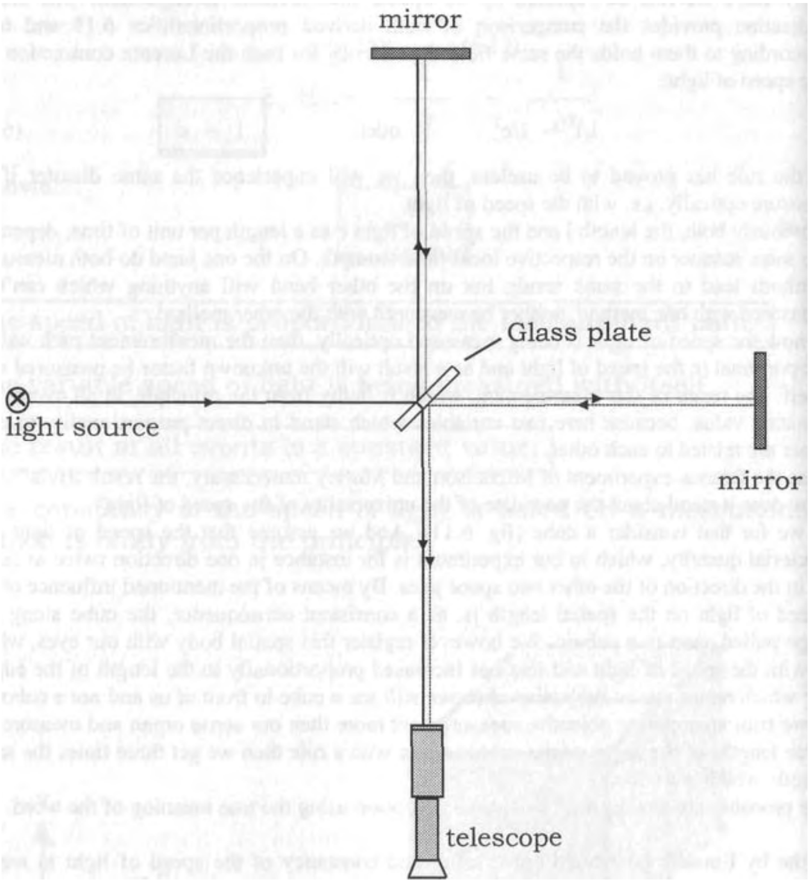

Fig. 6.12: Experiment of Michelson and Morley to detect an aetherwind<i>

<i>: A.P.French: Special Relativity, Massachusetts Institute of Technology, 1966.

<ii>: Nikola Tesla: "This is the same as writing a business letter and forgetting the subject you wish to write about". To Einstein's Theories, Rare Book and Manuscript Library, Columbia University, 15.4.1932.

<iii>: Einstein proceeds in the same manner with the time dilatation, by assuming a time constant by definition for the derivation to present at the end of his derivation a variable time. And with that he presents a result which contradicts his approach completely.

6.12 Aether

Important experiments like the one of Doppler concerning the redshift or the one of Bradley concerning the aberration of the stars show only to clear, where the influence of the speed of light subjectively still is perceptible, or for laboratory experiments like the one of Michelson and Morley, where the influence isn't perceptible anymore, because the length of the interferometers always changes proportionally to the speed of light. The look in the stars at the same time is a look in cosmic areas, where completely other field conditions prevail and as a consequence completely other values for the speed of light and for the dimensions of space are present. The mentioned observations suggest that we together with our measuring station are moving through the cosmos and therefore a relative velocity has to be present with regard to an aether which determines the respective speed of light.

If we however constrict our range of vision and retire in a laboratory, then we no longer are capable to observe the influence of the field on the speed of light. The experiments of Michelson which Maxwell had prompted to and which Morley with a higher precision had repeated with the goal, to detect the aether, inevitably had to turn out negatively. The laboratory experiments resulted in the misleading picture, as if the earth was resting in the aether.

The not understood measurements will suggest any observer, he forms the centre of the universe and everything rotates around him, entirely in the sense of the Ptolemean view of life, which, although long ago abolished, here belated has experienced support. With a Swabia caper Albert Einstein has prevented a relapse into the dark Middle Ages and removed the open contradiction in the question of the aether, which once is measured as moving and another time as resting, by without further ado abolishing the aether. With that he undoubtedly has solved a central problem of physics and at the same time created a new one. As is known does the speed of light have a certain value, and therefore the question is raised, what determines is size. Exactly for this purpose a luminiferous aether had been introduced, however it is constituted.

Scientifically it does make little sense, to make an assumption, if at the end of the derivation the prerequisite is deleted without substitute. In such a case either in the approach or in the derivation is a principal error*[1]"*. Nikola Tesla comments on the working method of Einstein with the applicable comparison, as if Einstein had, while he was writing a business letter, forgotten completely the subject he wanted to write about (fig. 6.12<ii>).

The answer, which removes all contradictions and is entirely in accord with all observations and measurements, is obvious. Naturally a luminiferous aether exists, which determines the velocity of propagation and of course it by no means is bound to the observer.

As has been derived in figures 6.5 and 6.6, will for a relative velocity v arise a field, which according to proportionality 6.17 determines the speed of light. With that we have derived completely.

The field takes over the function of the aether.

The equations 6.10 also answer the question, why no aetherwind is being observed, although such a wind actually is present: we experience, as we have discovered, an E-field with „head wind" as a resting H-field and vice versa and therefore we aren't capable to detect the head wind in the aether!

Key questions of quantum physics (fig. 4.4 + continuation):

IV. Why do the particles have the form of spheres?
(with increasing E-field decreases c)

VIII. Why is the elementary quantum localized?
(in the vortex centre: c = 0, see figures 4.3 and 6.2)

IX. Why do the elementary particles have a spin?
(spherical form demands field compensation)

X. Why is the magnitude of the spin quantized?
(cosmic basic field determines the need of E_z)

XI. Why can speeds faster than light occur in a tunnel? ________________________________
(a reduction of the cosmic basic field can only be realized locally in a tunnel)

to XI:

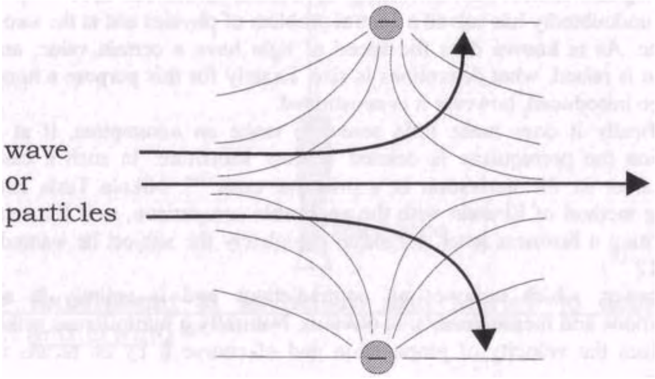

Fig. 6.13: Consequences concerning the field dependency of the speed of light: spin effect and tunnel effect <i,ii>

<i>: Nimtz,G.: Instantanes Tunneln, Tunnelexperimente mit elektromagnetischen Wellen, Phys.B1.49, VCH Weinheim (1993) Nr.12, S. 1119-1120<*>
<ii>: Thoma, P., Weiland.T.: Wie real ist das Instantane Tunneln? Phys.Bl.50, VCH Weinheim (1994) Nr.4, S. 359-361<*>
<*>: The measurement results are in accord with the theory of objectivity, not however the contradictory attempts to interpret them <i> and <ii> et al.

6.13 Spin and tunnel effect

Only with the field dependency of the speed of light (6.17) we can understand, why the elementary quanta can form as spheres, like is drawn in the figs 4.3 and 6.2. In the centre the field lines run together, i.e. the field increases and the speed of light decreases. Only in this way it will be possible for the vortex oscillation to everywhere occur with the speed of light, even in the inside of the particle! In the centre of the vortex particle the field in theory will become infinitely large and the speed of light zero. This circumstance again is the foundation why the elementary particles are localized and it answers key question VIII of quantum physics. The absence of a speed after all is the characteristic of an immobile thing.

The field dependency of the speed of light answers also further basic and up to today unanswered key questions of quantum physics, like why the elementary particles have a spin (IX) and why the magnitude of the spin is quantized (X).

A vortex particle after all does not exist alone in the world, but it is in the field of other particles. We can call this the cosmic basic field (E resp. H). This basic field overlaps the self-field and takes effect the strongest in the area of the spherical shell, where the self-field is correspondingly small. In order to keep the form of a sphere, this influence of the basic field has to be compensated. The additional field (E_z resp. H_z according to eq. 6.12) necessary for the compensation is produced by the particle, by rotating in a spiral around itself with a speed v which increases towards the outside of the spherical shell. Therefore does the elementary particles have a spin. The electron spin is therefore determined by the cosmic basic field.

Another effect of the field dependent speed of light is the tunnel effect. As an example we consider the two differently charged particles shown in fig. 6.8 A. The open, outside of the particles running, field lines of the electric field are predominantly bent towards the each time oppositely charged particle. If another particle wants to pass between the two, then it gets into an area of increased field strength. As a consequence it will be slowed down, because here a smaller speed of light is present.

Water molecules show with their polar nature exactly this property. Water has a remarkably high dielectricity e and slows down the speed of light correspondingly according to equation 5.6 ($\varepsilon \cdot \mu = 1/c^2$). The refraction of light at the water surface is an observable result of the reduced speed of light in the presence of matter.

If we now examine the case in which the two particles have the same charge as is shown in fig. 6.8 B (and fig. 6.13 belonging to XI). The field lines repel each other, so that exactly in between the two particles a field free area forms, in which the speed of light goes to infinity! This area acts like a tunnel. If we send through a particle exactly here, then purely theoretically seen it won't need any time to run through the tunnel, and for a short time the signal becomes infinitely fast.

If a particle hits only slightly besides the tunnel, then it will one-sidedly be slowed down and diverted by the respective field. We call this process reflection or scattering. Only the few particles, which exactly hit the tunnel, arrive behind the hurdle and in the ideal case even almost without loss of time!

The current measurements of speeds faster than light demonstrate in a convincing manner the superiority of the field-theoretical approach with regard to the nowadays normally used quantum physical approach.

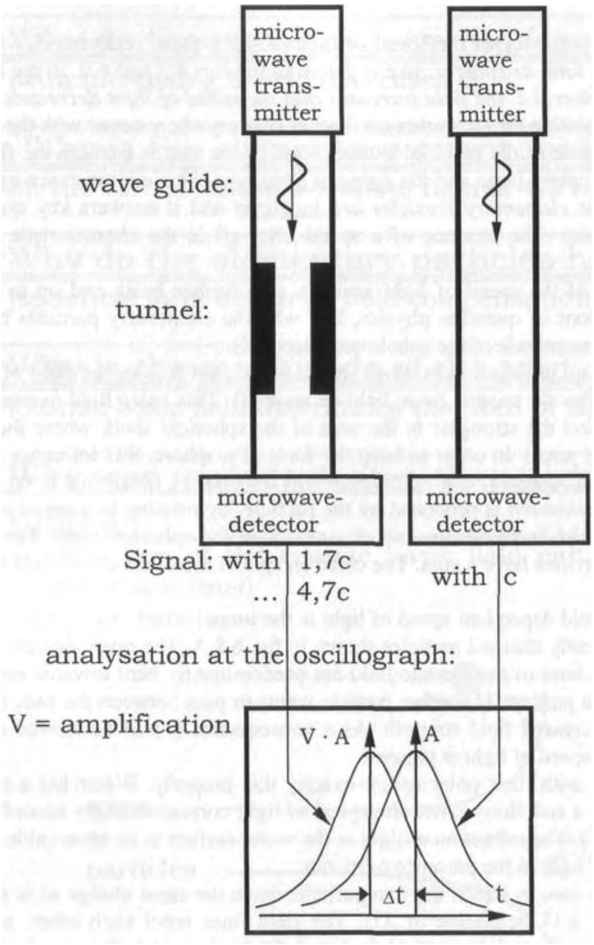
micro-
wave
trans-
mitter
micro-
wave
trans-
mitter
wave guide:
tunnel:
microwave-
detector
microwave-
detector
Signal: with 1,7c
... 4,7c
with c
analysation at the oscillograph:
V = amplification
V · A
A
Δt
t

Fig. 6.14: The microwave experiment at the II. Physical Institute of the University of Cologne to measure speeds faster than light.<i>

<i>: Nimtz, G. : New Knowledge of Tunneling from Photonic Experiments, Proc. of the Adriatico Research Conference, 1996, World Scientific Publishing Company

6.14 Interpretation of the measured speed faster than light

Now the attempt can be undertaken, to interpret the spectacular experiments, in which a speed faster than light has been measured. It is reported<i> that in experiments with photons at the University of California in Berkeley on an average a speed of 1.7 times the speed of light has been measured by Prof. Raymond Chiao and his co-workers. At the Technical University of Vienna Prof. Dr. Ferenc Krausz already has obtained 2.4 times the, according to Einstein at maximum obtainable, speed of light with tunnelling laser light.

The first measurements of speeds faster than light have been carried out with microwaves at the University of Cologne<ii> by Prof. Dr. Gunter Nimtz and co-workers. They at first had published the measurement of a speed 2.5 times the speed of light. In the meantime they even have transmitted a symphony of Mozart with a speed almost 10 times the speed of light and with that have contradicted Einstein's hypothesis, according to which the speed of light in vacuum would be the highest possible speed for the transmission of signals. The different experiments only resemble each other in the point that the particles have to tunnel, because one has put a barrier in their way. This "tunnelling" apparently is the cause for obtaining speeds faster than light. With the prevailing physical view of life these measurement results are incompatible.

In the measurement set up in Cologne the microwaves are sent through a wave guide, which they pass with the speed of light. If a parts with narrowed cross-section is inserted, where the microwaves actually don't fit through at all, then the signal gets damped strongly. Now however arrives nevertheless a small part of the signal at the other end of the wire, but much faster than allowed, namely with the measurable speed faster than light.

The answer of the here presented potential vortex theory reads as follows: the waves picked up by the wave guide run up to the entry of the tunnel, in order to find out that they don't fit through. They are reflected or absorbed. A small part however rolls up to potential vortices and these fit through the tunnel. They however have to be compressed additionally. In the derivation of the photon (fig. 4.5 and 4.6) we had seen that the inner vortex always is faster than the bigger one, through which it slips through. The compression therefore causes an increase in speed. In flow dynamics is known an analogy: the Venturi-tube. The flow-technical potential vortices also confirm exactly this property. One can as well start with the Lorentz contraction (fig. 6.6, eq. 6.14*). This states that a particle moving with a higher speed actually becomes smaller and not only appears to be smaller as an optical deception of the observer. Because only smaller particles fit through the tunnel, the particles, measurable at the other end, must be correspondingly faster: quod erat demonstrandum. In the same manner also the experiments of Berkeley can be explained physically, because here is worked with photons from the start<i>. With that the process of rolling up the wave can be left out. The tunnel lets pass only compressed and therefore faster light particles.

<i>: R.Y.Chiao, P.G.Kwiat, A.M.Steinberg: Schneller als Licht? Spektrum der Wiss. 10/93

<ii>: B. Schuh, Gespenstisch fixe Wellen, DIE ZEIT Nr. 45, 5.11.93, S. 43.

<ii>: Enders, A., Nimtz, G.: Photonic-tunneling experiments, Physical Review B, Vol. 47, No. 15 (1993), pp. 96O5-96O9.

<ii>: Enders, A., Nimtz, G.: Evanescent-mode propagation and quantum tunneling, Physical Review E, Vol. 48, No.l (1993), pp. 632-633.

"The theory of relativity is not a physical theory... it is a mathematical poetic idea, a deduction from impossible premises."<i>
Oskar Kraus

"The theory of relativity is a mathematical masquerade, behind which is hidden an inextricable ball of a mixing up of ideas, contradictions, fallacies, arbitrary assumptions and ignoring of healthy logic."<i>
Erich Ruckhaber

"The theory of relativity not only is fantastic, but also of an inconsistency which in the history of science not yet has been present."<i>
Harald Nordenson

"A physics of hybrids, of contradictions and fantastic confusions, nonsense!"<ii>
Johann Marinsek

"This is absurd."<iii> (regarding mass-energy interpretation) ________
Nikola Tesla

"In my experiments I have destroyed billions of atoms, without having observed any emissions of energy. "<4i>
Nikola Tesla

Fig. 6.15: Some statements regarding the theory of relativity.

<i>: Walter Theimer: Die Relativitatstheorie, Seite 7, Francke Verlag, Bern, 1977, ISBN 3-772O-126O-4
<ii>: Johann Marinsek: Rationale Physik, S. 163, dbv-Verlag TU Graz, 1989, ISBN 3-7O41-O176-1
<iii>: Nikola Tesla, To Einstein's Theories, Rare Book and Manuscript Library, Columbia University, 15.4.1932. Entnommen aus J.T.Ratzlaff: Tesla Said, Tesla Book Company, pp. 238, ISBN O-914119-OO-1
<4i>: Nikola Tesla, Franz Ferzak World and Space Publications 1985.

6.15 Definition of the speed of light

If a light signal propagates in space, then as a consequence of the velocity of propagation c, it at a certain point in time t is in a distance r of the light source:

$$r = c * t \qquad (6.19)$$

Should the speed of light become smaller for instance by Δc, then the light signal obviously has covered a distance less by Ar or the time interval has changed by Δt:

$$r + \Delta r = (c + \Delta c) \cdot (t + \Delta t) \quad (6.20)$$

This equation describes purely mathematically the most general case which can be assumed. By writing out the multiplication and subtraction of equation 6.18 the change in distance considered for itself is:

$$\Delta r = c \cdot \Delta t + t \cdot \Delta c + \Delta c \cdot \Delta t \quad (6.21)$$

The answer of mathematics is that the change in distance can have its cause in a change in time, in a change of speed or in both. We now want to turn to the physical interpretation and have a closer look at the two possibilities, in which either c or t is to be taken constant (see fig. 6.16).

In the first case the speed of light c is constant and as a consequence the change Δc = zero. The mathematical formulation (according to eq. 6.21) therefore reads:

case 1: $\boxed{c = \frac{\Delta r}{\Delta t}}$ (relativity) (6.22)

If in this conception world a change in distance is observed, for instance the Lorentz contraction, then in order to save this relation inevitably a change in time, for instance a time dilatation, has to make the compensation. Einstein in an applicable manner speaks of relativity, because according to his opinion in the case of both variables, the length contraction and the time dilatation, it only concerns observed changes.

For the time dilatation experiments are given. But for the measurement of time always only atomic clocks are available and their speed of running of course could also be influenced by the Lorentz contraction. In any case it can't be claimed the time dilatation is proven experimentally as long as we do not know the mechanisms of decay of atoms. Otherwise the statements of the theory of relativity are familiar to us, for which reason further remarks seem unnecessary.

In the second case the time t is constant and consequently the change At = zero. At a closer look this case is much more obvious, since why should time change. After all time has been stipulated by definition.

After all, we are the ones who tell, what simultaneity is!

The mathematical formulation for this case reads (eq. 6.21 with Δt= 0):

case 2: $\boxed{\Delta c = \frac{\Delta r}{t}}$ (objectivity) (6.23)

This equation does open up for us an until now completely unknown and fundamentally other way of looking at the physical reality.

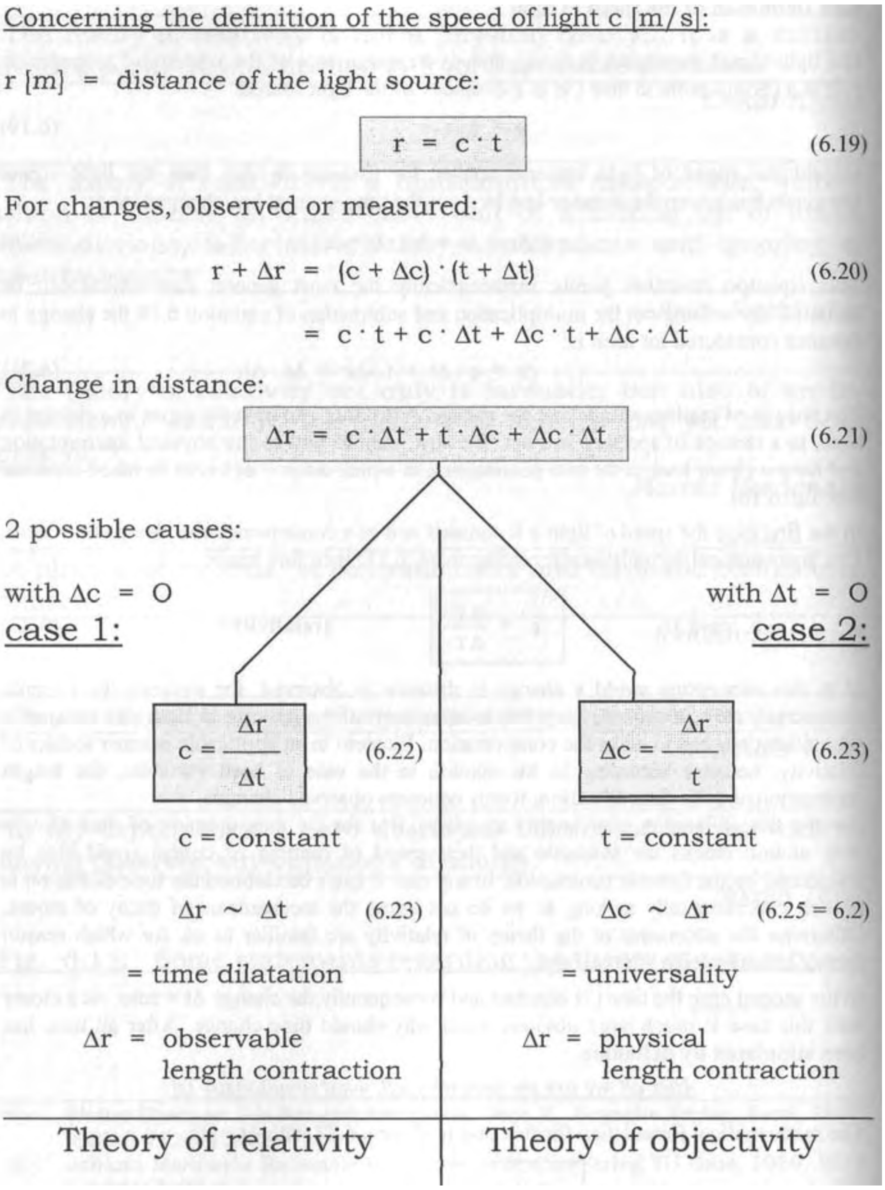

Concerning the definition of the speed of light c [m/s]:

r [m] = distance of the light source:

$$\boxed{r = c \cdot t} \qquad (6.19)$$

For changes, observed or measured:

$$r + \Delta r = (c + \Delta c) \cdot (t + \Delta t) \qquad (6.20)$$

$$= c \cdot t + c \cdot \Delta t + \Delta c \cdot t + \Delta c \cdot \Delta t$$

Change in distance:

$$\boxed{\Delta r = c \cdot \Delta t + t \cdot \Delta c + \Delta c \cdot \Delta t} \qquad (6.21)$$

2 possible causes:

with $\Delta c = 0$

case 1:

$$\boxed{c = \frac{\Delta r}{\Delta t}} \qquad (6.22)$$

c = constant

$$\Delta r \sim \Delta t \qquad (6.23)$$

= time dilatation

Δr = observable length contraction

Theory of relativity

with $\Delta t = 0$

case 2:

$$\boxed{\Delta c = \frac{\Delta r}{t}} \qquad (6.23)$$

t = constant

$$\Delta c \sim \Delta r \qquad (6.25 = 6.2)$$

= universality

Δr = physical length contraction

Theory of objectivity

Fig. 6.16: Theory of relativity and theory of objectivity, derivation and comparison.

6.16 Relativity and objectivity

New to the second case (equation 6.23) is particularly the proportionality contained in it:

$$\boxed{\Delta c \sim \Delta r} \quad (6.25 = 6.2)$$

But to us it is not new, because we have derived the same proportionality from the model conept (equation 6.2, fig. 6.2), in which the elementary particles are understood as spherical vortices. Equantion 6.25 unconcealed brings to knowledge that any change of the speed of light c [m/s] in the same way leads to a change of the radius r [m], the distance between two points in space or even the length of an object, e.g. a rule. Such a rule after all consists of nothing but spherical atoms and elementary particles and for their radius r again the proportionality 6.25 holds. Therefore it is to be set:

$$r \sim 1 \qquad (6.26)$$

and taken both together we already had derived as equation 6.18 (fig. 6.11) from the field dependency. Here the vortex model as well finds a confirmation of its correctness, as in the derivation from the equations of transformation of the electromagnetic field. Because all three, the derivation according to the model, the physical and the mathematical derivation, lead to the same result, this second case should be called "objective".

With that the first case, which describes the subjective perception of an observer, is not supposed to be devaluated. It contains the definition of reality, according to which only is real what also is perceptible. The theory of relativity of Poincare and Einstein is based on this definition.

With the second case, the case with a variable speed of light, we however get serious problems, since we observe with our eyes, and that works with the speed of light. If that changes, we can't see it, as already said. If we could see it, then "reality" would have a completely different face and we surely would have great difficulties, to find our way around. In this "objective world" neither electromagnetic interactions nor gravitation would exist, so no force effects at all. Because all distances and linear measures depend on the speed of light, everything would look like in a distortion mirror.
The concept of an "objective world" at first has not a practical, but rather a theoretical and mathematical sense. The distinction between an observation domain and a model domain is founded in pure usefulness.

The observation domain should correspond to case 1 and the model domain to case 2. The mathematical derivation tells us, how we can mediate between both domains (equation 6.21): This mediation amounts to a transformation, which provides us the instruction, how a transition from the observation into a not perceptible model concept, from the relativity into an objectivity has to.

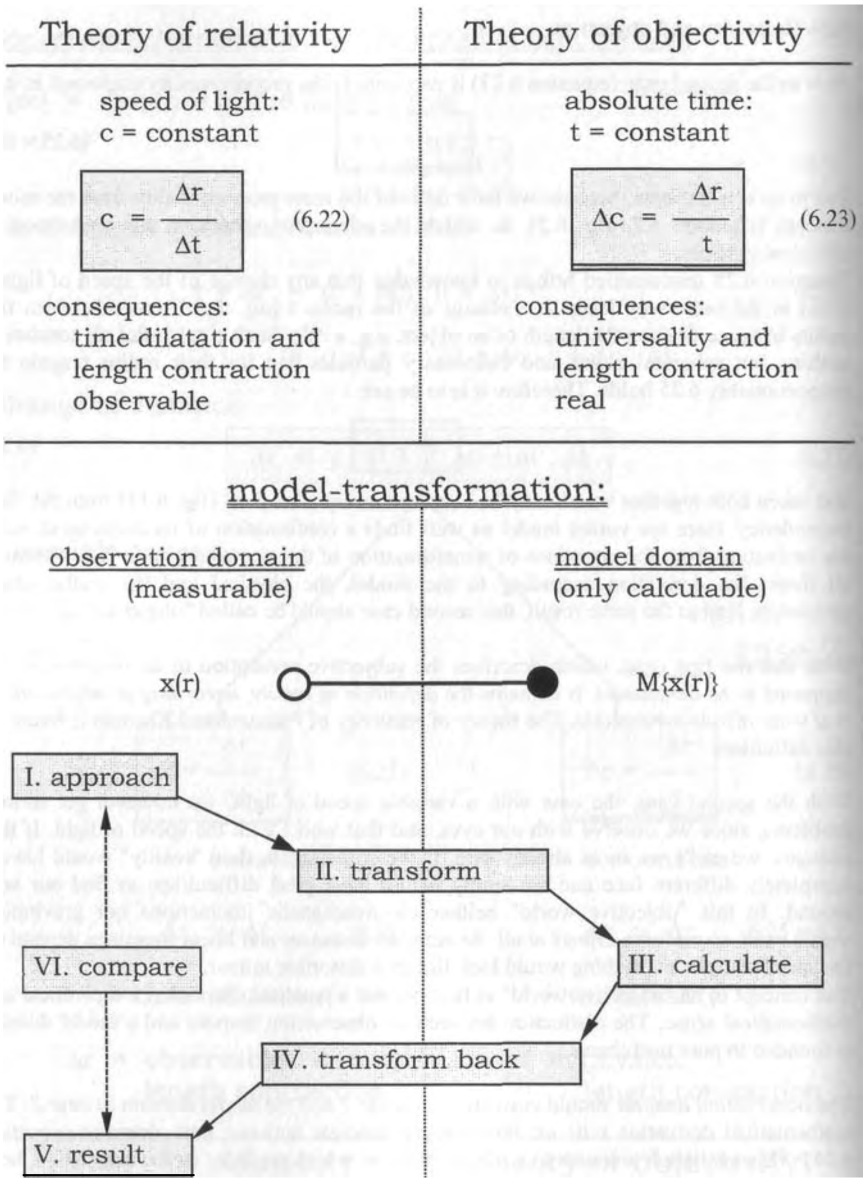
Theory of relativity
Theory of objectivity
speed of light:
c = constant
absolute time:
t = constant
$c = \frac{\Delta r}{\Delta t}$ (6.22)
$\Delta c = \frac{\Delta r}{t}$ (6.23)
consequences:
time dilatation and
length contraction
observable
consequences:
universality and
length contraction
real
model-transformation:
observation domain
(measurable)
model domain
(only calculable)
x(r)
M{x(r)}
I. approach
II. transform
III. calculate
IV. transform back
V. result
VI. compare

Fig. 6.17: Model-transformation between
theory of relativity and theory of objectivity.

6. 17 Transformation

The observation domain is, as the name already expresses, perceptible (observable) with the help of our sense organs and measurable with corresponding apparatus. The special theory of relativity for the most part provides us the mathematics needed for that. And in that is assumed a constant speed of light. Because a length contraction is being observed and can be measured, a time dilatation must arise as a consequence. Such is the consistent statement of this theory. Because we already could make us clear that it concerns a subjective theory, of course caution is advisable if generalizations are being made, like the one of the inductive conclusion of the length contraction on the time dilatation. We'll come to speak about that in this chapter (fig. 6.20).

The model domain however is not observable to us and only accessible in a mathematical manner. Here the time is a constant. On the other hand do the radii of the particles and all other distances and linear measures stand in direct proportionality to the speed of light. If that changes, then does that lead to a change in length. The length contraction occurs physically, which means actually. We propose the name "theory of objectivity" for the valid theory which is derivable with this prerequisite and independent of the point of view of the observer.

The importance of this model domain and of the possible model calculations is founded in the circumstance that many physical relations within our observation domain aren't recognized by us and can't be mathematically derived. Besides is only all to often worked with unallowed generalizations and with pure hypotheses. Such a thing does not even exist in the model domain.

The model domain can be tapped over a transformation. For that we select an approach x(r) in the to us accessible observation domain. This then is transformed into the model domain by a calculation instruction M{x(r)}. Here we can calculate the sought-for relation In the usual manner and transform back again the result according to the same calculation instruction M^{-1} {x(r)} but in the reversed direction. After being returned in our familiar observation domain, the result can be compared and checked with measurement results (fig. 6.17).
In this way we will derive, calculate and compare the quantum properties of the elementary particles with the known measurement values. Here we remind you of the fact that all attempts to calculate the quantum properties conventionally, without transformation, until now have failed. Not even a systematization may succeed, if it concerns for instance explanations for the order of magnitude of the mass of a particle.

A transformation at first is nothing more than an in usefulness founded mathematical measure. But if a constant of nature, and as such the quantum properties of elementary particles until now have to be seen, for the first time can be derived and calculated with a transformation then this measure with that also gains its physical authorization. We now stand for the question: how does the instruction of transformation M{x(r)} read, with which we should transform the approach and all equations from the observation domain into the model domain?

general relations:	special theory of relativity	theory of objectivity
domain of validity:	observation domain	model domain
speed of light c [m/s]:	$c = c_o$ $= \text{constant}$	$c \sim r$ (6.25)
field strengths (6.15)	(6.15) + (6.18)	
H [A/m] :	$H \sim 1/r^2$	$H \sim 1/r$ (6.27*)
E [V/m] :	$E \sim 1/r^2$	$E \sim 1/r$ (6.27)
because of eq. (5.6) $\varepsilon \cdot \mu = 1/c^2$ is valid:		
μ [Vs/Am] :	$\mu_0 = \text{const.}$	$\mu \sim 1/r$ (6.28*)
ε [As/Vm] :	$\varepsilon_0 = \text{const.}$	$\varepsilon \sim 1/r$ (6.28)
relations of material: $B = \mu \cdot H$ (3.5): B [Vs/m²]	$B \sim 1/r^2$	$B \sim 1/r^2$ (6.29*)
$D = \varepsilon \cdot E$ (3.6): D [As/m²]	$D \sim 1/r^2$	$D \sim 1/r^2$ (6.29)
e.g. spherical capacitor capacity: C [As/V]	$= \varepsilon 4 \pi r$ (6.4)	$C = \text{const.}$ (6.30)
charge: Q [As]	$= C \cdot U$ (6.31)	$Q = \text{const.}$ (6.32)
energy: W [VAs]	$= Q^2/C$ (6.1)	$W = \text{const.}$ (6.33)
with energy-mass relation: (5.24) and (6.1)	$W = mc^2$	
mass	m [kg = VAs³/m²]	$m \sim 1/r^2$ (6.34)
relaxation time	τ_1 [s]:	$\tau_1 = \text{const.}$ (6.35)
with	$\tau_1 = \varepsilon/\sigma$ (5.3)	
specific conductivity	σ [A/Vm]:	$\sigma \sim 1/r$ (6.36)

Fig. 6.18: Transformation of the dependencies on radius

6.18 Transformation table

The attempt to write down at this point already a closed mathematical relation as instruction of transformation, would be pure speculation. Such an instruction first must be verified by means of numerous practical cases, i.e. be tested for its efficiency and correctness. But we not even know the practical examples necessary for this purpose, if we apply the transformation for the first time!
For his reason it unfortunately is not yet possible, to calculate absolute values in a direct
We have to be content to work with proportionalities and to carry out comparisons.
In fig. 6.18 the proportionalities are compared in the way, how they would have to be transformed: on the left side, how they appear and can be observed in the view of the special theory of relativity, and on the right side, how they can be represented and calculated in the theory of objectivity.
The change, which here would have to be transformed, is the physical length contraction, which is the change in length as it depends on the speed of light. For spherical symmetry the length 1 becomes the radius r (eq. 6.26), of which is to be investigated the influence. In the observation domain we had derived the proportionality (6.15 + 6.18):

$E \sim 1/r^2$ and $H \sim 1/r^2$.

The field of a point charge or of a spherical capacitor confirms this relation:

$$E = Q/\varepsilon 4\pi r^2 .$$

Because the speed of light in our observation is constant, also both constants of material which are related to it (eq.5.6: $\varepsilon \cdot \mu = 1/c^2$), the dielectricity ε and the permeability μ, are to be taken constant.
With that the same proportionality as for the field strengths also holds for the induction B and the dielectric displacement D:

$$B \sim 1/r^2 \quad \text{and} \quad D \sim 1/r^2.$$

In the model domain everything looks completely different. Here the radius and any length stands in direct proportionality to the speed of light. In this way we get problems with our usual system of units, the M-K-S-A-system (Meter-Kilogram-Second-Ampere). The basic unit Meter [m] and as a consequence also the unit of mass Kilogram [kg = VAs^3/m^2] appear here as variable. It would be advantageous, to introduce instead the Volt [V] as basic unit.
But in any case does the dimension of a quantity show us, in which proportionality it stands to the unit of length. This in the model domain then is authoritative! As an example does the speed of light have the dimension Meter per Second. In the model domain there consequently has to exist a proportionality to the length r [m].
The speed of light determines with equation 5.6 again the constants of material:

$$\mu\,[Vs/Am] \sim 1/r \quad \text{and} \quad \varepsilon[As/Vm] \sim 1/r \qquad (6.28)$$

According to the model holds unchanged:

$$B\,[Vs/m^2] \sim 1/r^2 \quad \text{and} \quad D\,[As/m^2] \sim 1/r^2. \qquad (6.29)$$

But if we insert the proportionalities 6.28 and 6.29 into the equations of material 3.5 and 3.6, then holds for the field strengths:

$$H\,[A/m] \sim 1/r \quad \text{and} \quad E\,[V/m] \sim 1/r. \qquad (6.27)$$

Further dependencies of the radius can be read in the same manner either by inserting into well-known laws or immediately from the dimension.

In the observation domain:

energy density of a field:

$$w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2 \sim 1/r^4 \qquad (6.37)$$

energy: $W \sim 1/r$ (6.38)

with c = constant and $W = m \cdot c^2$: $\boxed{m \sim 1/r}$ (6.39)

e^- is bigger and lighter than p^+ or n^0!

$$\boxed{\frac{m}{m_0} \overset{(6.39)}{=} \frac{r_0}{r} \overset{(6.26)}{=} \frac{l_0}{l} \overset{(6.24^*)}{=} \frac{1}{\sqrt{1-(v/c)^2}}} \qquad (6.40)$$

In the model domain:

e.g. spherical capacitor

capacity: $C\,[\mathrm{As/V}] = \varepsilon 4\pi r$ (6.4) | $C = \text{const.}$ (6.30)

charge: $Q\,[\mathrm{As}] = C \cdot U$ (6.31) | $Q = \text{const.}$ (6.32)

energy: $W\,[\mathrm{VAs}] = Q^2/C$ (6.1) | $\boxed{W = \text{const.}}$ (6.33)

(6.33) means: law of conservation of energy!

relaxation time τ_1 [s]: | $\tau_1 = \text{const.}$ (6.35)

with $\tau_1 = \varepsilon/\sigma$ (5.3)

specific conductivity σ [A/Vm]: | $\boxed{\sigma \sim 1/r}$ (6.36)

(6.36) means: elementary vortices are indestructible!

Fig. 6.19: Interpretation of the dependencies on radius

6.19 Interpretation of the transformation table

The transformation should tell us, what we would see if the variable speed of light would be observable to us. Doing so highly interesting results come out.

The energy density of a field is as is known $w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2$. (6.37)

In the observation domain will, according to fig. 6.19, decrease the energy density w proportional to $1/r^4$. Multiplied with the respective volume we obtain for the energy itself the proportionality: $W \sim 1/r$. (6.38)

If we make use of the Einstein relation $W = m \cdot c^2$

with c = constant holds also for the mass m: $m \sim 1/r$. (6.39)

In this manner we finally find out, why the small nucleons (protons and neutrons) subjectively seen are heavier than the very much larger electrons. As a consequence does a relativistic particle experience the increase of mass (with the length contraction according to equation 6.24*):

$$\boxed{\frac{m}{m_0} \overset{(6.39)}{=} \frac{r_0}{r} \overset{(6.26)}{=} \frac{l_0}{l} \overset{(6.24*)}{=} \frac{1}{\sqrt{1-(v/c)^2}}} \qquad (6.40)$$

This result is experimentally secured. Our considerations therefore are entirely in accord with the Lorentz-transformation. There at least is no reason to doubt the correctness.

In the model domain we with advantage assume a spherical symmetry. As easily can be shown with equations 6.4 and 6.31, are the capacity and charge of a spherical capacitor independent of the radius (6.30 and 6.32). In that case also the from both values calculable energy (6.1) must be constant. We come to the same conclusion, if take we the above equation 6.37 for the energy density of a field or if we carry out a verification of dimensions:

W [VAs] = konst. . (6.33)

This simple result is the physical basis for the law of conservation of energy! With that we have eliminated an axiom.

The result states that the energy stays the same, even if the radius, the distance or the speed of an object should change. To the subjectively observing person it shows itself merely in various forms of expression. Consequently is the energy, as is dictated by the here presented field theory, formed by binding in the inside of the quanta the same amount of energy but of the opposite sign. The amount of energy therefore is bound to the number of the present particles, as we already had derived.

Under the assumption of a constant time (6.35) there results for the electric conductivity σ, by calculating backwards over the equation of the relaxation time (5.3), the proportionality: (6.36)

$$\boxed{\sigma\ [\mathrm{A/Vm}] \sim 1/r} \qquad (6.36)$$

Maybe the result surprises, because it can't be observed. Actually we know that the (microscopically observed conductivity in reality only represents an approximated averaged measure for the mobility of free charge carriers. In a particle-free vacuum however this well-known interpretation doesn't make sense anymore. Hence it is recommended, to only work with the relaxation time constants. Who nevertheless wants to continue to work with σ as a pure factor of description, can do this. But he mustn't be surprised, if in the model domain with decreasing radius the conductivity suddenly increases. But this is necessary, because otherwise the elementary particles would collapse. Only by the increase of the conductivity, which is produced by the spherical vortex itself, will the expanding eddy current build up in the inside of the particles, which counteract the from the outside concentrating potential vortex.

Approach:

a.The particles don't decay by themselves, but only by a corresponding disturbance from the outside.

b.The decay time is the statistical average in which such a disturbance can occur and take effect.

c.The elementary particles consist of an integral and finite number of elementary vortices, which can't decay anymore for their part.

d.If the compound particles get into the disturbing range of influence of high-frequency alternating fields, then they are stimulated to violent oscillations and in that way can be torn apart into individual parts.

e.As disturbing factor the high-frequency fields of flying past neutrinos are considered primarily.

f. Authoritative for the threshold of decay and with that also for the rate of decay is the distance, in which the neutrinos fly past the particle.

g.The distance becomes the larger, the smaller the particle is. If the particle thus experiences a relativistic length contraction, then it will, statistically seen, to the same extent become more stable!

That has nothing to do at all with time dilatationl

We are entitled to demand a simultaneity, after all we are the ones, who tell what that is!

Fig. 6.20: Proposal for an interpretation of the particle decay

<i>: Walter Theimer: Die Relativitatstheorie, Seite 106, Francke Verlag, Bern, 1977, ISBN 3-772O-126O-4

6.20 Particle decay

We still have to get rid of a fundamental misunderstanding. It concerns the problem of the time dilatation. Here the model domain doesn't give us any difficulty, because it dictates a constant and therefore by us definable time. In the relativistic view however should in moving systems clocks go wrong! But how does one want to explain a time dilatation physically, if it merely represents a purely mathematical result of the actually taking place length contraction on the one hand and the postulate of a constant speed of light on the other hand?

Nobody has troubled more about the physical interpretation than Einstein himself. But he had as less as we nowadays the possibility to verify the so-called phenomenon experimentally, by accelerating a laboratory clock to values close to the speed of light. Only atomic particles can, e.g. in accelerator systems, be brought on such high speeds and then be observed for their properties. But also these experiments don't have any power of proof, as long as we don't know the atomistic structure of the particles and there exists the danger of misinterpretations.

So the slowing down of the rate of decay of instable. particles at high speeds willingly is cited as "proof for time dilatation<i>. "The most cited example for the time dilatation is the "long-living" meson. The μ-meson is a charged particle, which exists only $2,2 * 10^{-6}$ seconds if it is observed in rest. Then it decays ... About 10 % of the mesons reach the earth's surface. Even if they fly with approximately the speed of light, they at least must have used $30 \cdot 2,2 * 10^{-6}$ seconds, in order to reach the earth. Their "life" therefore by the movement has been extended for a multiple... to the supporters of the theory of relativity here the time dilatation is revealed..."

This "proof however is worthless, as long as "the structure and the mechanism of decay of the particle are not known", like W. Theimer<i> expresses himself.

On the basis of the new field theory the approach standing on the left page is dared (fig. 6.20). Accordingly the particles don't decay by themselves, but only by a corresponding disturbance from the outside, which for instance is triggered by the high-frequency fields of flying past neutrinos. The closer the neutrinos fly past the particle, the sooner it will decay. But the distance becomes the larger, the smaller the particle is. If the particle thus experiences a relativistic length contraction, then it will, statistically seen, to the same extent become more stable!

That has nothing to do at all with time dilatation, as this proposal for an interpretation shows (fig. 6.20). The same effect of course also occurs, if atomic clocks are taken for a fly in a plane and compared to identically constructed clocks on earth.

The time was stipulated by us and therefore should be able to keep its universal validity. We are entitled to demand a simultaneity, after all we are the ones, who tell what simultaneity is!

An interesting technical use would be the acceleration of the rate of decay in order to dispose of radioactively contaminated waste. For that the waste has to be irradiated by collecting and focussing free neutrinos or with the help of a neutrino transmitter, like one which will be discussed in chapter 9. After such a neutrino shower dangerous radioactive waste would be reusable or at most be harmless domestic refuse.

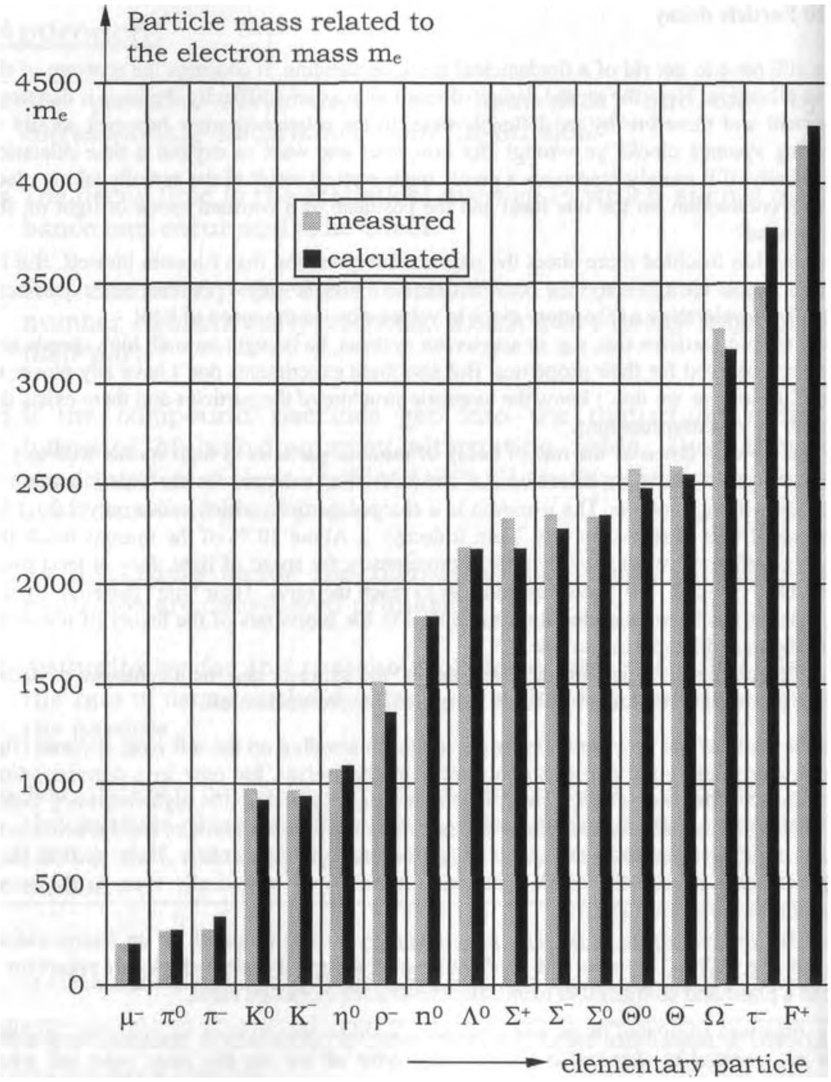

Fig. 7.0: The comparison with power of proof:
▒ the measured particle mass
■ the calculated particle mass

7. Proof

Ample evidence is available for the correctness of the theory of objectivity. The field dependent change in length is observed and used as magnetostriction or electrostriction. If a ferromagnetic material, e.g. a nickel rod, is brought into an alternating magnetic field, then field dependent longitudinal length oscillations are observed. In the same manner barium titanate or quartz crystal oscillates in the electric field if a high-frequency alternating voltage is applied.
A practical application forms the production of ultrasound.

In this chapter are, as already announced, the quantum properties of the elementary particles calculated and in this way is furnished perhaps the most convincing proof for the existence of potential vortices and for the correctness of the field-theoretical approach and the theory which is based on it.
A special challenge represents the calculation of the particle mass. This mass stretches from 207 electron masses of the myon over 1839 of the neutron into the order of magnitude of 18513 electron masses (Y°). Doing so not only can be tested, if the calculated values correspond with the measured ones. Also the gaps have to correspond, i.e. where there doesn't exist a discrete mathematical solution also no particle should exist. The fig. 7.0 standing on the left page anticipates the result and shows that even this strict condition is fulfilled! The agreement of the calculated with the measured results is excellent. If in individual cases small deviations become visible, we always have to bear in mind that the measurements as a rule are analysed statistically and the results are falsified if small particles creep in unrecognized. Particle physics nowadays has at its disposal extremely precise gauges, but even here remaining errors can't be excluded.

Quantum physics is occupied with further taking apart the elementary particles into hypothetic particles, the quarks, and to sort these according to properties and symmetries. Seen strictly causal this procedure thus corresponds to the quantum physical approach. We however have taken the field-theoretical approach, and this excludes the introduction of hypothetic particles from the start. It should be our goal to derive and to explain the quantum structure as a field property. Yes, we even want to calculate it, with which we would have overtaken quantum physics in the scientific competition with one leap!

Strong support our approach has experienced by current experiments, in which matter was transformed in electromagnetic waves - practically the reversal of the rolling up of waves to vortices. To do so at the Massachusetts Institute of Technology (David Pritchard and others) sodium atoms were dematerialized in waves by lattice scattering<i>. According to Einstein one surely could have blown the whole M.I.T. in the air with the occurring mass defect; but don't worry, no emission of energy whatsoever has been observed, entirely as predicted by the vortex theory.

<i>: J. Teuber: Materie und Energie, Ganze Atome wurden zu Energiewellen,
Illustrierte Wissenschaft Nr. 7 (1996), S. 56

With the classical radius of the electron $r_e = 2{,}82 * 10^{-15}$ m:

$$C_e = \varepsilon_o \cdot 4\pi r_e = 3{,}135 \cdot 10^{-25} \mathrm{F} \tag{6.4*}$$

$$U_e = e/C_e = 511 \mathrm{kV} \tag{6.31*}$$

(constant independent of r_e)

Formation forms (vortex properties):

I. Amassing (formation of vortex balls):

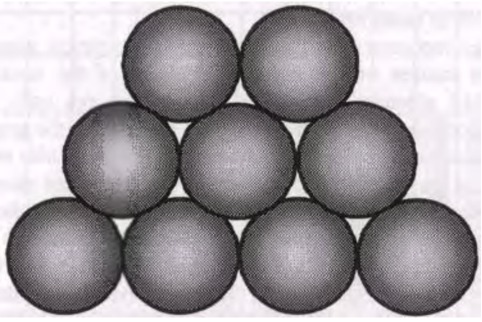

II. Overlapping (phenomenon of transport)

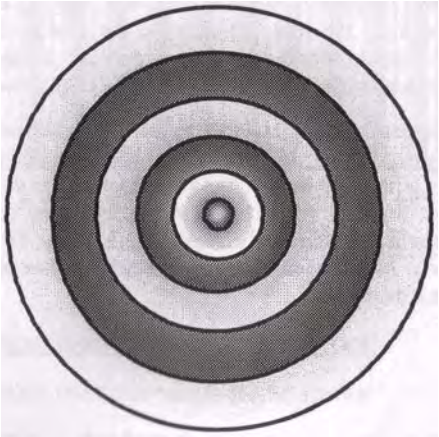

Fig. 7.1: The amassing and overlapping of elementary vortices

7.1 Elementary vortices

We had derived the electron and the positron as elementary vortices (fig. 4.3). Before we can go in the calculation, we must gain a clear picture of the possible configurations of vortices, which for reason of the derived properties are possible. For that we start with the elementary vortex and afterwards we predict the behaviour of interaction which can be expected.

Actually only one single particle is really elementary. According to the realizations of the new theory it is an elementary vortex in the form of a sphere. Its size is determined by the speed of light and this again by the local field strength; its stability is founded in the concentration effect of the potential vortex. The whirling takes place everywhere with the speed of light, even in the vortex centre, where all field lines run together, where the field increases infinitely and the speed of light goes to zero. This last circumstance owes the elementary vortex its localization.

We can attribute a charge to this vortex for reason of the field lines which on the outside run towards infinity and which we can measure (fig. 4.3). This is the smallest indivisible unit, the elementary charge e. Structure and course of the field lines suggest to understand and to calculate the elementary vortex as a spherical capacitor. By basing on the classical radius of the electron r_e given in fig. 6.3 the capacity according to equation 6.4 is calculated to be:

$$C_e = \varepsilon_0 \cdot 4\pi r_e = 3{,}135 \cdot 10^{-25}\ \mathrm{F} \quad (6.4^*)$$

Here the theory of objectivity has provided us the realization that even for a change of the radius of the electron the capacity remains unchanged constant (6.30), and this entirely corresponds to our observation.

Between the hull of the elementary vortex, measured at the radius r_e, and its centre, respectively also with regard to infinity, there exists according to equation 6.31 the tension voltage of:

$$U_e = e/C_e = 511\ \mathrm{kV} \qquad (6.31^*)$$

It as well is constant and independent of the size of the elementary vortex.

Since a different solution is refused, we'll have to assume that all elementary particles consist of an integer multiple of elementary vortices. For that the amassing, like closely packed tennis balls, or the overlapping of individual vortices in the form of shells, like in the case of an onion (phenomenon of transport) can be considered.

The among each other occurring forces of attraction can be traced back to the fact that every elementary vortex is compressed by the field of its neighbour as a consequence of the field dependent speed of light. This field as a rule is for the small distances considerably larger than the field on the outside. Therefore do compound elementary particles not have the twofold or triple mass, but at once the 207-fold (myon) or the 1836-fold (proton) mass. After all there is no other explanation for the fact that there don't exist lighter particles (with a mass less than 207 electron masses)!

a. The electron-positron pair

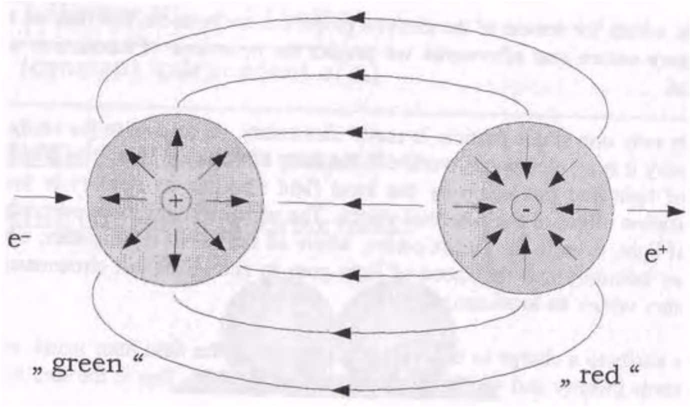
+
-
e⁻
e⁺
„ green “
„ red “

b. . The e^- - e+ pair for a small distance:

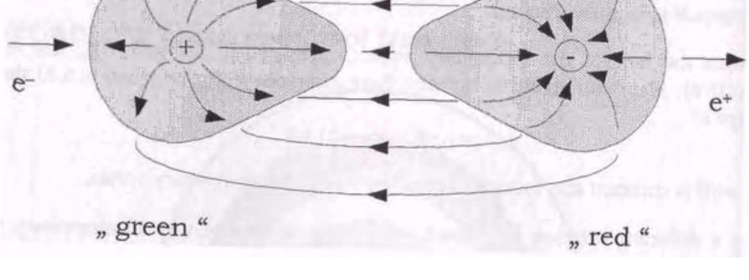
+
-
e⁻
e⁺
„ green “
„ red “

Fig. 7.2: The electron-positron pair annihilation

7.2 Matter and anti-matter

For the amassing or overlapping of elementary vortices several cases must be distinguished, because two inverse forms of formation are possible for the elementary vortex: the negatively charged electron and the positively charged positron. Whereas in the case of the electron the vortex produces a component of the electric field which points from the inside to the outside, has the field in the case of the positron the opposite direction for reason of a reversed swirl direction.

This statement can be generalized: if we consider the elementary particles from the outside, then we assign the particles with a swirl direction identical to that of the electron to the world of matter and call the particles with the opposite swirl direction anti-matter. It now is strongly recommended, to take colours to hand, in order to optically clarify the properties of vortices. The electron will be marked as a green sphere and the antiparticle, the positron, as a red sphere.

If we now look into the world of matter, then appears our world of matter to us "green", the world of anti-matter however "red". The uniform green colour of all the in our world existing elementary particles however doesn't exclude that red anti-vortices can exist hidden in the inside of the green vortices, where we can't discover them. But they must be completely covered, otherwise a disastrous reaction occurs, the pair annihilation, as a consequence of the oppositely directed property of the vortices which cancel out. By means of the pair annihilation a dematerialization can occur, because every elementary vortex keeps in its inside the same amount of energy with opposite sign and the fusion of two inverse particles can result in a zero sum of the energy. The best known example is the annihilation of an electron-positron pair under emission of radiation discovered by Klemperer in 1934. In the upper representation (fig. 7.2a) the elementary vortices still are symmetrical, but the outside field lines already are "bent" and linked together in such a way that, with the exception of the ones in the direction of the axis, no interaction takes place which can be measured.

The two particles for reason of the different charge approach each other quickly, and the closer they are, the larger the mutual force of attraction becomes; a vicious circle, which leads to the asymmetry shown in the lower sketch (fig. 7.2b) and only comes to rest, if both particles have destroyed themselves mutually.

The electron and the positron had the same amount of, but oppositely directed swirl activity, so that purely arithmetically seen a zero sum of the rest energy results. But it should be paid attention to both particles having some kinetic energy on the occasion of the relative motion to each other and if they rotate around their own axis also rotational energy. An emission of annihilation radiation occurs, is the explanation of particle physics.

With the knowledge of the photon (fig. 4.6) we can interpret the annihilation radiation as a consequence of the phenomenon of transport. The faster and consequently smaller vortex, for instance the green one, slips into the red one and sees the green inside, which is compatible for it. Unfortunately it only can remain there, as long as it is smaller, thus is faster, and therefore it shoots out on the other side again. Now the electromagnetic force of attraction fully takes effect. It is slowed down and the red vortex correspondingly accelerates. The process is reversed.

These around each other oscillating vortices, so we had derived, have a characteristic frequency (colour), are polarizable and are moving forward with the speed of light as a consequence of the open vortex centre. It therefore concerns the photon.

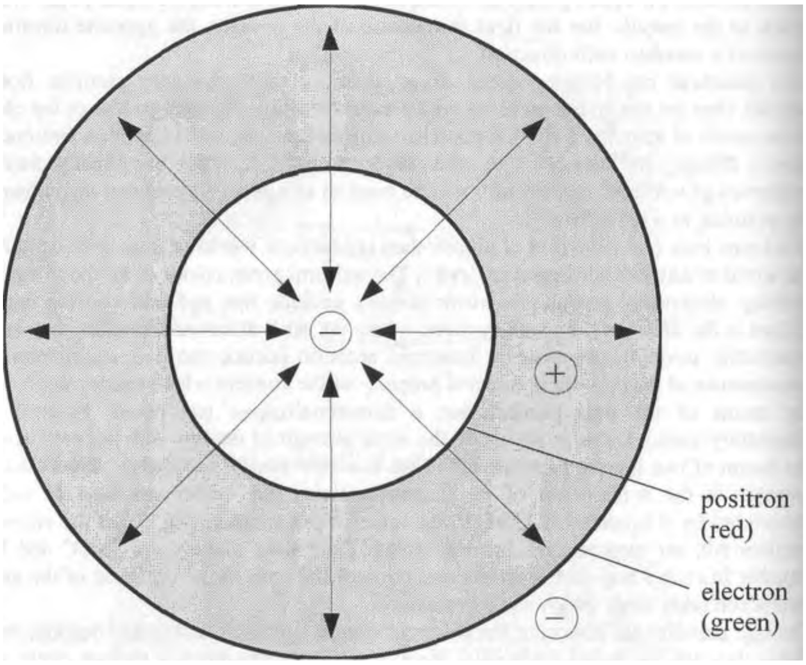
⊖
⊕
positron
(red)
electron
(green)
⊖

Fig. 7.3: Theoretical final state of the positronium = static γ-quant (photon).

7.3 Positronium

But before the two elementary vortices, the electron and the positron, are annihilated under emission of radiation, they will for a short time take a shell-shaped, a bound state, in which one vortex overlaps the other.

Its formation we can imagine as follows: an electron, flying past a resting positron, is cached by this for reason of the electromagnetic attraction and spirals on an elliptic path towards the positron. In doing so its angular velocity increases considerably. It will be pulled apart to a flat disc for reason of the high centrifugal forces, to eventually lay itself around the positron as a closed shell.

Now the red positron sees the electron vortex so to speak "from the inside" and doing so it sees as well red; because the green vortex has a red centre and vice versa! The result is the in fig. 7.3 given configuration.

The number of field lines, which run from the red border of the positron in the direction of the centre, is identical to the number, which point towards the green border of the electron. Here already the same state has been reached as in the centre, which corresponds to the state at infinity. That means that no field lines point from the green border to the outside; seen from the outside the particle behaves electrically neutral. It doesn't show any electromagnetic interaction with its surroundings.

If the particle were long-living, then it undoubtedly would be the lightest elementary particle besides the electron; but without stabilizing influence from the outside the positronium can't take the in fig. 7.3 shown state at all. The positron takes up the kinetic energy which is released if the electron becomes a shell around it. But before the bound state can arise, which would identify the positronium as an elementary particle, the equal rights of both vortices comes to light. With the same right, with which the electron wants to overlap the positron, it itself vice versa could also be overlapped.

If the stabilization of the one or the other state from the outside doesn't occur, then the stated annihilation under emission of y-quanta is the unavoidable consequence (fig. 4.6).

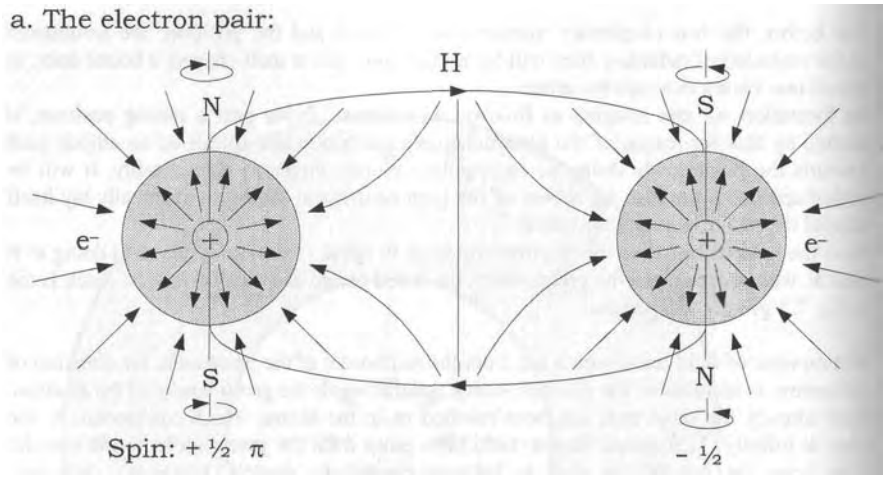
a. The electron pair:
H
N
S
e⁻
+
+
e⁻
S
N
Spin: + ½ π
– ½

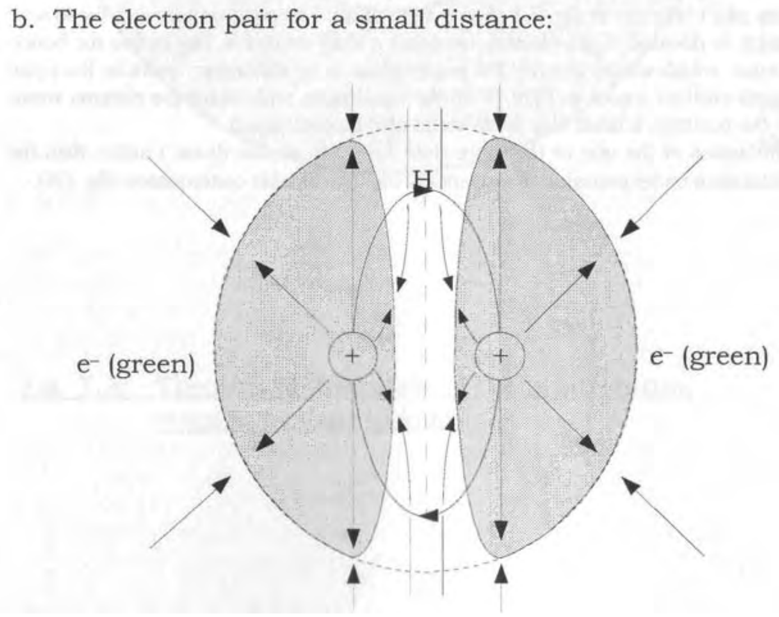
b. The electron pair for a small distance:
H
+
+
e⁻ (green)
e⁻ (green)

Fig. 7.4: Two electrons with oppositely directed spin

7.4 Dipole moment

As electrically charged spheres elementary vortices have a magnetic dipole moment along their axis of rotation as a consequence of the rotation of their own (fig. 7.4). This is measurable very precisely and for the most important elementary particles also known quantitatively. In contrast to the angular momentum the magnetic moment can't be constant according to the here presented theory. It should slightly change, if we increase the field strength in the laboratory.

In a particle consisting of several elementary vortices the vortices mutually increase the local field strength. Therefore we measure at the proton, which consists of three vortices, not the triple, but only the 2,793-fold of the nuclear magneton which can be expected for reason of its mass. Also the neutron has instead of the double only the 1,913-fold nuclear magneton. The deviations therefore are explicable as a consequence of the surrounding fields.

Prerequisite for this point are two other, still unanswered, key questions of quantum physics:

> XII: Why is measured for the proton approximately the triple of the magnetic dipole moment which can be expected for reason of the charge?
>
> XIII: Why does the neutron, as an uncharged particle, actually have a magnetic moment?

These questions can only be brought to a conclusive answer, if we have derived the vortex structures of the respective particles.

The elementary vortex, as a consequence of the spin along its axis, forms a magnetic north pole and a south pole. Another possibility to interact with an external field or with other particles is founded on this property. This shall be studied by means of two electrons. which form an electron pair.

For reason of the equal charge the two electrons at first will repel each other. If they rotate of their own they however will mutually contract, which, seen from the outside, is interpreted as a force of attraction. And in addition will they align their axes of rotation antiparallelly. While they now rotate in the opposite direction, a magnetic force of attraction occurs.

As is shown in fig. 7.4, the magnetic dipole field in this way is compensated towards the outside, as is clarified by the field line (H) with a closed course. Between both electrons a space free of E-field stretches. If both vortices are a small distance apart they lay themselves around this space like two half-shells of a sphere. A particle forms which seen from the outside is magnetically neutral, but it carries the double elementary charge (fig. 7.4b).

The exceptional affinity is always restricted to two vortices of equal charge with an opposite direction of rotation. Further vortices can't be integrated anymore and are repelled. This property of vortices covers the quantum condition (Pauli's exclusion principle) for the spin quantum number perfectly.

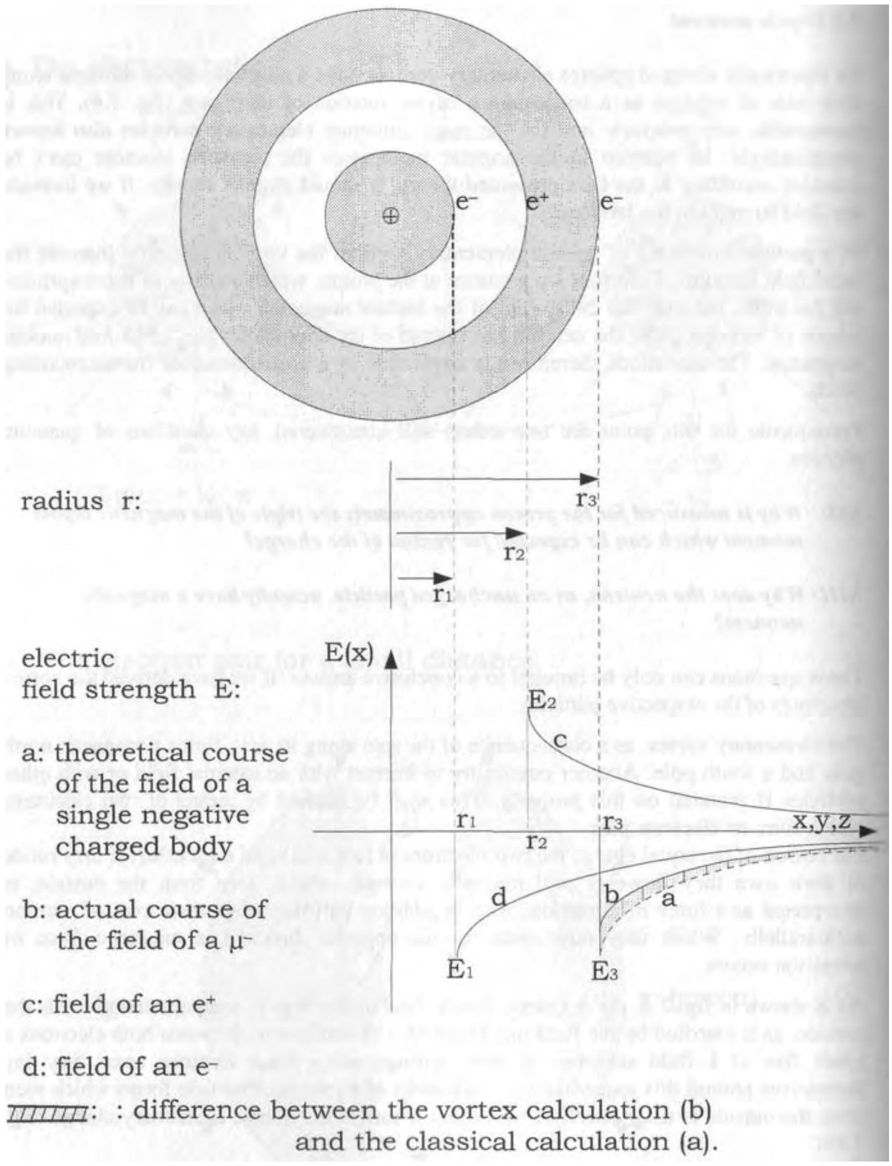
⊕
e⁻
e⁺
e⁻
radius r:
r₃
r₂
r₁
electric field strength E:
E(x)
E₂
c
r₁
r₃
r₂
x,y,z
d
b
a
E₁
E₃
a: theoretical course of the field of a single negative charged body
b: actual course of the field of a μ⁻
c: field of an e⁺
d: field of an e⁻
▨: : difference between the vortex calculation (b) and the classical calculation (a).

Fig. 7.5: The mvon and the electric field E(x) of the three elementary vortices

7.5 Myon

We now have discussed all conceivable possibilities, which two elementary vortices can form: the creation of a pair for like charge and the annihilation under emission of photons via the formation of the positronium as an intermediate result for unequal charge. Next another elementary vortex shall be added and all different possibilities and configurations will be derived, which can be formed by amassing or overlapping.

The positronium can, as said, only take the in fig. 7.3 shown bound structure, if it is stabilized from the outside. This task now a further electron shall take over. According to the shell model the innermost elementary vortex an electron (e^-), is overlapped by a positron (e^+) and that again overlapped by an electron (e^-).
With an in the sum single negative charge, a completely symmetric structure as well as a half-integer spin this particle will show a behaviour corresponding to a large extent to that of the electron. Merely the mass will be considerably larger, because every vortex each time compresses the other two.

It therefore concerns the myon (μ^-), which also is called "heavy electron". The myon was discovered 1937 in the cosmic radiation (Anderson and others).
In fig. 7.5 are drawn above each other the shell-shaped structure of the myon and the electric field E(x) of the three elementary vortices.
It is visible that merely in the proximity of the particle the actual course of the field deviates from and is smaller, than the course which theoretically can be expected for a single negatively charged body. The difference is marked by a hatching.

We now can tackle the calculation of the myon. For that the following considerations to begin with are useful:
Mass is an auxiliary term founded in usefulness, which describes the influence of the electromagnetic field on the speed of light and with that on the spatial extension of the "point mass".
Without exception the local cosmic field E_o has an effect on a free and unbound elementary vortex, thus on an individual e^- or e^+, and determines so its size and its mass. But as long as we haven't determined this field strength, the calculation of its quantum properties won't succeed.
Instead the masses of compound particles will be compared to each other, which are so heavy that the field strength of the neighbouring vortices is predominant over the basic field E_0, so that a neglect of E_o seems to be allowed. The course of the calculation is made for all elementary particles in the same manner, which is explained hereafter.

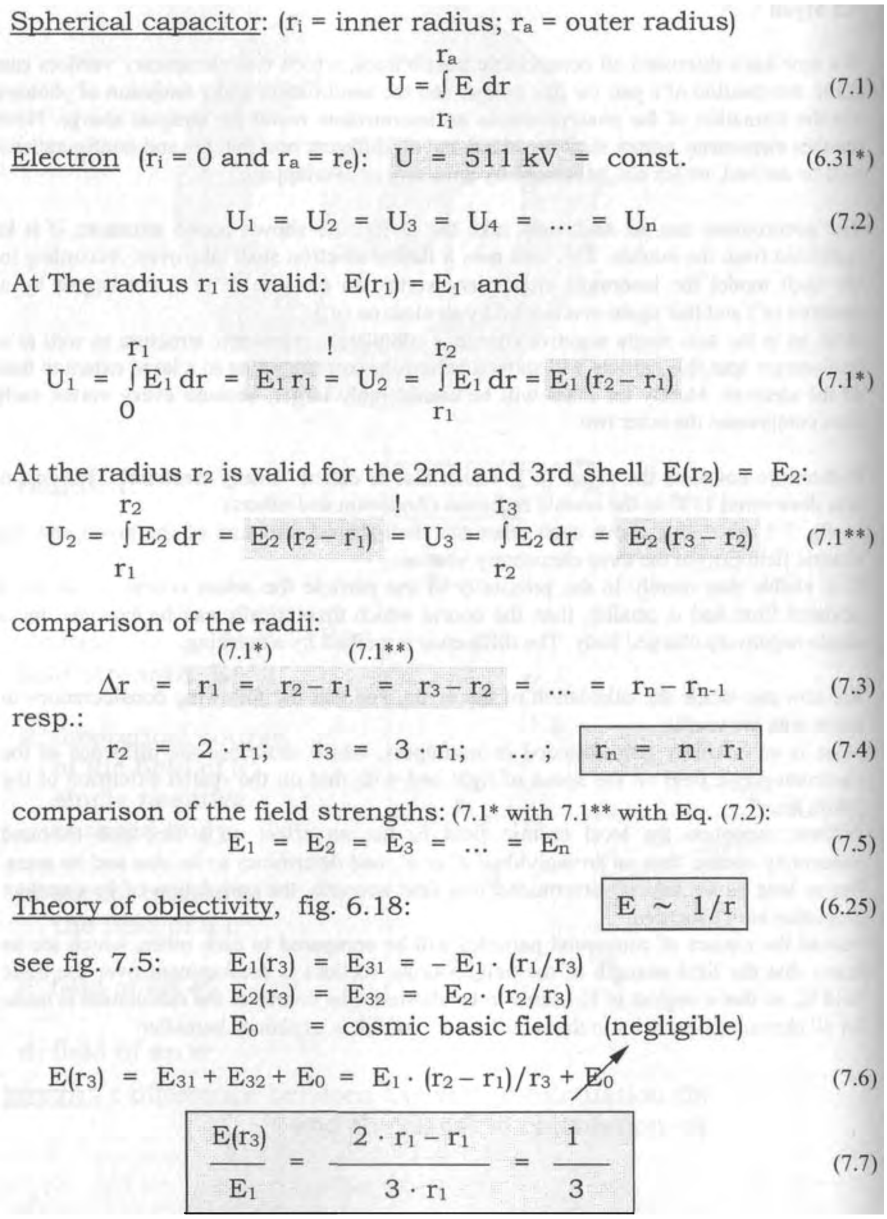

Spherical capacitor: (r_i = inner radius; r_a = outer radius)

$$U = \int_{r_i}^{r_a} E\,dr \qquad (7.1)$$

Electron ($r_i = 0$ and $r_a = r_e$): $U = 511\text{ kV} = \text{const.}$ (6.31*)

$$U_1 = U_2 = U_3 = U_4 = \dots = U_n \qquad (7.2)$$

At The radius r_1 is valid: $E(r_1) = E_1$ and

$$U_1 = \int_0^{r_1} E_1\,dr = E_1 r_1 \overset{!}{=} U_2 = \int_{r_1}^{r_2} E_1\,dr = E_1 (r_2 - r_1) \qquad (7.1^*)$$

At the radius r_2 is valid for the 2nd and 3rd shell $E(r_2) = E_2$:

$$U_2 = \int_{r_1}^{r_2} E_2\,dr = E_2 (r_2 - r_1) \overset{!}{=} U_3 = \int_{r_2}^{r_3} E_2\,dr = E_2 (r_3 - r_2) \qquad (7.1^{**})$$

comparison of the radii:

$$\Delta r = r_1 \overset{(7.1^*)}{=} r_2 - r_1 \overset{(7.1^{**})}{=} r_3 - r_2 = \dots = r_n - r_{n-1} \qquad (7.3)$$

resp.:

$$r_2 = 2 \cdot r_1; \quad r_3 = 3 \cdot r_1; \quad \dots; \quad \boxed{r_n = n \cdot r_1} \qquad (7.4)$$

comparison of the field strengths: (7.1* with 7.1** with Eq. (7.2):

$$E_1 = E_2 = E_3 = \dots = E_n \qquad (7.5)$$

Theory of objectivity, fig. 6.18: $\boxed{E \sim 1/r}$ (6.25)

see fig. 7.5:

$$E_1(r_3) = E_{31} = -E_1 \cdot (r_1/r_3)$$

$$E_2(r_3) = E_{32} = E_2 \cdot (r_2/r_3)$$

E_0 = cosmic basic field (negligible)

$$E(r_3) = E_{31} + E_{32} + E_0 = E_1 \cdot (r_2 - r_1)/r_3 + \cancel{E_0} \qquad (7.6)$$

$$\boxed{\frac{E(r_3)}{E_1} = \frac{2 \cdot r_1 - r_1}{3 \cdot r_1} = \frac{1}{3}} \qquad (7.7)$$

Fig. 7.6: Calculation of the electric field strength E(r) of the myon from its dependency on radius

7.6 Calculation of the vortex fields

The tension voltage of an elementary vortex, like for a spherical capacitor, is determined by integrating over the electric field strength from the inner radius r_i up to the outer radius r_a:

$$U = \int_{r_i}^{r_a} E\,dr \quad (7.1)$$

For the electron ($r_i = 0$ und $r_a = r_e$) we already have carried out the integration and determined the tension voltage to be 511 kV (equation 6.31 *).
Doing so we further had discovered that it won't change, if the radius r varies. Even for a shell configuration, in which electrons and positrons alternately overlap, the approach is valid:

$$U_1 = U_2 = U_3 = U_4 = \ldots = U_n \quad (7.2)$$

At a certain radius all elementary vortices show the same density of field lines and with that also the identical field strength, so that we can solve the integral (7.1) for the each time neighbouring vortex shells and can compare the results:
At the radius r_1 with $E(r_1) = E_1$ the agreement, according to equation 7.1* (fig. 7.6), is valid for the innermost and the overlapped vortex shell.
At the radius r_2 with $E(r_2) = E_2$ the agreement according to equation 7.1** (fig. 7.6) is valid analogously for the 2nd and 3rd shell.
If still more shells are present, then we can arbitrarily repeat this procedure. For the radius of each shell we always obtain relation 7.3, which, related to the innermost radius, provides the following simple expression for the individual radii:

$$r_2 = 2 * r_1; \quad r_3 = 3 \bullet r_1; \quad \ldots; \quad r_n = n * r_1 \quad (7.4)$$

From the comparison of the integration results 7.1* and 7.1** follows further that all elementary vortices produce the same field strength:

$$E_1 = E_2 = E_3 = \ldots = E_n \quad (7.5)$$

We infer from the transformation table (fig. 6.18, eq. 6.27) that the field strengths E and H decrease with 1/r. In fig. 7.5 the decrease of the fields with 1/r is shown. Up to the radius r, the field of the innermost vortex E_1 has worn off to the value $E_{31} = - E_1 \bullet (r_1/r_3)$. This field is overlapped by $E_{32} = E_2 * (r_2/r_3)$ as well as the cosmic basic field E_o:

$$E(r_3) = E_{31} + E_{32} + E_0 = E_1 \bullet (r_2 - r_1)/r_3 + E_o \quad (7.6)$$

The local basic field E_o is not known, but it is very small with regard to the field of the neighbouring vortex shells, so that a neglect seems to be allowed. From equation (7.6) in this way follows with the radius relation (7.4):

$$\frac{E(r_3)}{E_1} = \frac{2 \cdot r_1 - r_1}{3 \cdot r_1} = \frac{1}{3} \quad (7.7)$$

For the shell-shaped configuration of the myon (fig. 7.5) relation (7.7) indicates, which field the outside vortex shell is exposed to. From this can already be seen, how much it is compressed thanks to the field dependent speed of light and how much its mass as a consequence is increased.

Structure of the proton p^+:

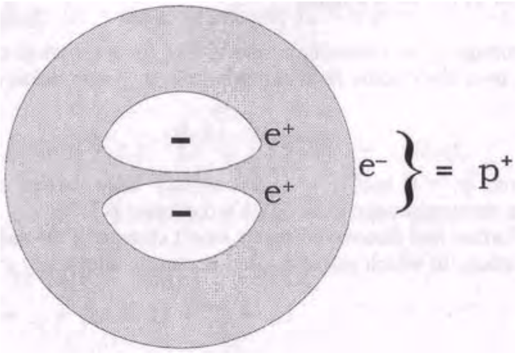

Calculation:
structure consisting of two shells, inner vortices with $2 \cdot E_1$,
field strength at the outer radius r_2:

$$E(r_2) = 2 * E_{21} = 2 * E_1 (r_1 / r_2) = E_1 \qquad (7.8)$$

Comparison of p^+ (7.8) with u^- (7.7) (z_e = number of the elementary vortices being involved with) in building up the structure, here each time $z_e = 3$):

Comparison of the radii with $E \sim 1 / r$ (6.27)

$$\frac{r_p}{r_\mu} = \frac{z_{e\mu}}{z_{ep}} \cdot \frac{E_\mu(r_3)}{E_p(r_2)} = \frac{3}{3} \cdot \frac{1/3}{1} = \frac{1}{3} \qquad (7.9)$$

Theory of objectivity (fig. 6.18): $m \sim 1/r^2$ (6.34)

$$\frac{m_p}{m_\mu} = \frac{z_{ep}}{z_{e\mu}} \cdot \left(\frac{r_\mu}{r_p}\right)^2 = \frac{3}{3} \cdot \left(\frac{3}{1}\right)^2 = 9 \qquad (7.10)$$

$$m_p/m_e = 9 * (m_u/m_e) = 9 * 207 = 1863 \qquad (7.11)$$

Measurement value, proton mass: $m_p = 1836 \cdot m_e$
Resp.:
measurement value myon mass $m_u = 207 * m_e$
myon calculated value: $m_p = 204 * m_e$. (error = 1,5%)

Since we, by using this calculation method, for the first time succeeded in deriving the mass of an elementary particle from that of another particle, the particle mass isn't a constant of nature anymore!

Fig. 7.7: Calculation of the proton

7.7 Calculation of the proton

If we again remember the affinity of two elementary vortices, which rotate with opposite spin. They align their axis of rotation antiparallel and form a very probable, but not particularly tight bound pair (fig. 7.4).

If we this time start with a positron pair, then does this pair have a double positive elementary charge. The two e^+ hence exert a particularly big force of attraction on electrons flying past them. If they have cached one and put it round as a shell, like a coat, then they will never again give it back! To again remove the electron, a triple positive charge would be necessary. But such a particle can't exist at all. The new particle therefore has an absolute stability and a very big mass, because the positron pair is considerably compressed by its outer shell. The total charge is single positive. With these properties it actually only can concern the proton. Its structure is shown in fig. 7.7. We can start from the assumption that both positrons are very close together in the inside and thus each forms the half of a sphere. For the calculation of the proton mass we then can assume as an approximation a structure of two shells, in which the inner vortex will have the double charge and the double field (2 * E_1). With equation 7.4 the field strength at the outer radius r_2 is:

$$E(r_2) = 2*E_{21} = 2*E_1*(r_1/r_2) = E_1 \qquad (7.8)$$

If we want to compare the results of the p^+ (7.8) and the μ^- (7.7), then it should be considered that the field of the innermost elementary vortex E_1 only is equal, if the number z_e of the elementary vortices involved in building up the particle is identical. Here with each time $z_e = 3$ this is the case. Because of equation 6.27 (E, H ~ 1/r) now also the radii are comparable:

$$\frac{r_p}{r_\mu} = \frac{z_{e\mu}}{z_{ep}} \cdot \frac{E_\mu(r_3)}{E_p(r_2)} = \frac{3}{3} \cdot \frac{1/3}{1} = \frac{1}{3} \qquad (7.9)$$

The mass of a particle first is determined by the number of the elementary vortices z_e. According to the theory of objectivity (fig. 6.18) however also the radius has an influence on the mass:

$$m \sim 1/r^2 \qquad (6.34)$$

This proportionality should be applied to the p^+-μ^-- comparison.

$$\frac{m_p}{m_\mu} = \frac{z_{ep}}{z_{e\mu}} \cdot \left(\frac{r_\mu}{r_p}\right)^2 = \frac{3}{3} \cdot \left(\frac{3}{1}\right)^2 = 9 \qquad (7.10)$$

The calculation provides a nine times bigger mass for the proton with regard to the mass of the myon. Therefore the mass of the proton related to the mass of the electron is:

$$m_p/m_e = 9*(m_\mu/m_e) = 9*207 = 1863 \qquad (7.11)$$

It would be favourable, to start from the with measuring techniques determined value for the mass of the proton $m_p/m_e = 1836$ and calculate backwards the related mass of the myon.

Then we obtain 204 as the calculated value instead of the measurement value m_μ/m_e= 207.

The reason for the deviation of 1.5 percent is caused by the neglect of the cosmic field E_o with regard to the field of the neighbouring elementary vortex. This neglect takes very much less effect for the relatively heavy proton than for the light myon.

The cosmic field therefore will compress the myon more strongly and increase the mass more strongly as is calculated here, in agreement with the measurement results.

Summarizing: since we, by using this calculation method, for the first time succeeded in deriving the mass of an elementary particle from that of another particle, the particle mass isn't a constant of nature anymore!

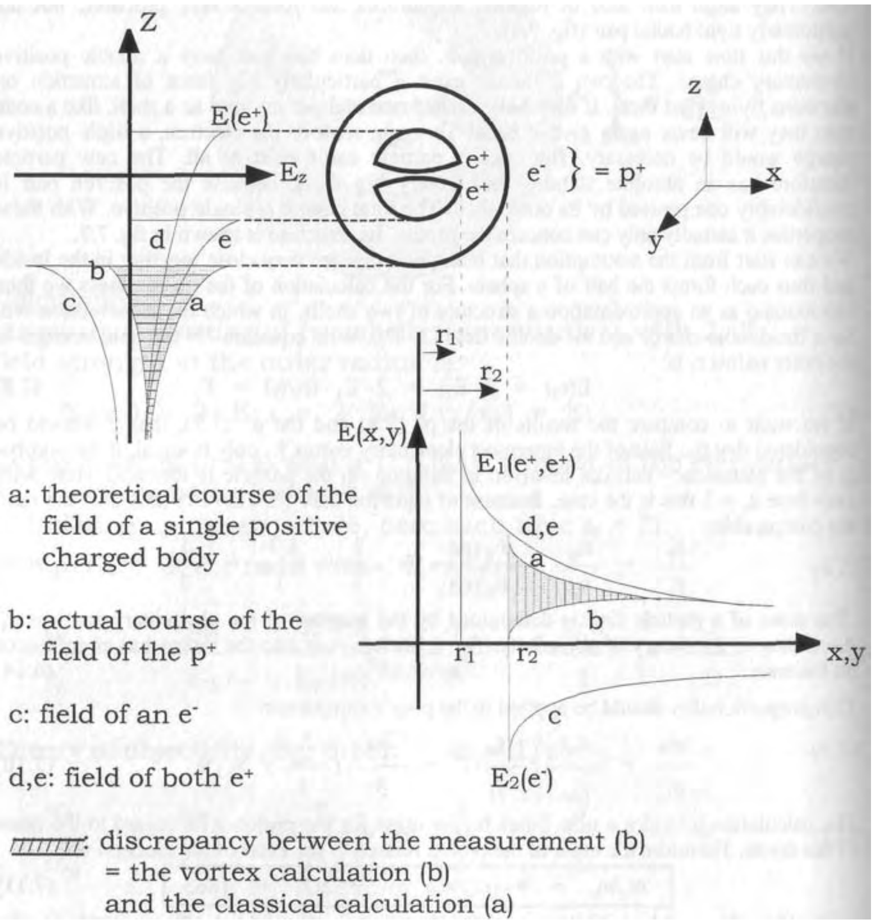
Z
E(e⁺)
E_z
e⁺
e⁺
e⁻
} = p⁺
z
x
y
d
e
b
c
a
r₁
r₂
E(x,y)
E₁(e⁺,e⁺)
d,e
a
b
r₁
r₂
x,y
c
E₂(e⁻)
a: theoretical course of the field of a single positive charged body
b: actual course of the field of the p⁺
c: field of an e⁻
d,e: field of both e⁺
▨: discrepancy between the measurement (b) = the vortex calculation (b) and the classical calculation (a)

Fig. 7.8: The proton and the electric field of the three elementary vortices in x-, y- and z-direction

7.8 "Strong interaction"

A central question of nuclear physics concerns the forces which keep the atomic nucleus, which consists of many neutrons and protons, together and give it its very good stability in spite of the like positive charge (key question XIV fig. 7.13).

According to today's textbook opinion (course of the field indicated with a in fig. 7.8) the forces of repulsion between the individual protons increase further as the distance gets smaller, to obtain immense values within the nucleus. They theoretically had to be overcome by new and unknown nuclear forces. Therefore physicists assume the hypothesis of a "strong interaction". But they are mistaken. The answer to this open question is provided by the course of the field (b) for the proton, sketched in fig. 7.8. We see that the electric field at first indeed still increases if we approach the proton, but in the proximity it contrary to all expectations decreases again until it is zero. With that then also any force of repulsion has vanished! But the course of the field follows without compulsion from the overlap of the three individual elementary vortex fields.

The field direction in the z-direction even is reversed! In this topsy-turvy world, in theory, an electromagnetic force of attraction between two like charged protons can occur. We conclude:

A strong interaction doesn't exist at all. The usually given values for "range" and "strength" just represent a misinterpretation. The hatched drawn area marks the difference which is misinterpreted by quantum physics. The model concept over and above that answers another mysterious property of the proton. As an electrically charged particle with a spin it first of all should form a magnetic moment for reason of the rotating charge. But until now the measurable order of magnitude couldn't be explained.

7.9 Magnetic moment of the proton

If the inner positrons rotate around each other with oppositely pointing spin, then the magnetic field line is already closed within the particle and no effect in x- or y-direction is observable from the outside.

As pair they however still can rotate together around the z-axis and they'll do that. The overlapping electron for reason of its rotation of its own will likewise build up a magnetic dipole moment along its axis of rotation. It also will align its axis in the z-direction, so that now all three elementary vortices have one field axis. Being comparable to individually "elementary magnets" aligned in the same direction they produce a triple magnetic moment (key question XII fig. 7.13).

If we namely would start with a single positively charged body according to the theory of quantum mechanics, then we would have expected the value of the nuclear magneton[i] p_m as the magnetic moment for the proton $p_m = e \cdot \hbar/2m$. Opposite to that provide experiments with protons the approx. threefold value as already predictable by the new vortex theory. In addition does the direction of the vector p_{mp} correspond with the spin-axis, so as if the proton were negatively charged. The reason for that is that only the outermost elementary vortex determines the spin of the particle, and that is actually a negatively charged electron! Also this excellent agreement in the case of the proton can be judged as proof for the correctness of the vortex model.

---

[i]: The nuclear magneton has the value of: $p_{mk} = 5{,}0508 \cdot 10^{-27}\ \mathrm{Am}^2$

a. n^0 if the e^- amasses to a p^+.

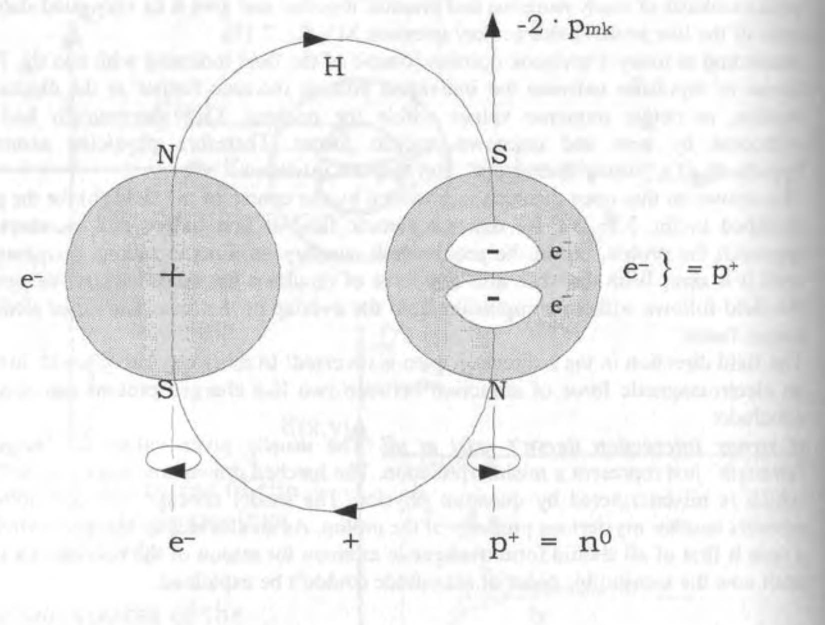
-2 · pmk
H
N
S
e⁻
+
-
e⁻
-
e⁻
e⁻ } = p⁺
S
N
e⁻
+
p⁺ = n⁰

b. n^0 if the e^- is overlapped by an e^+ of the p^+

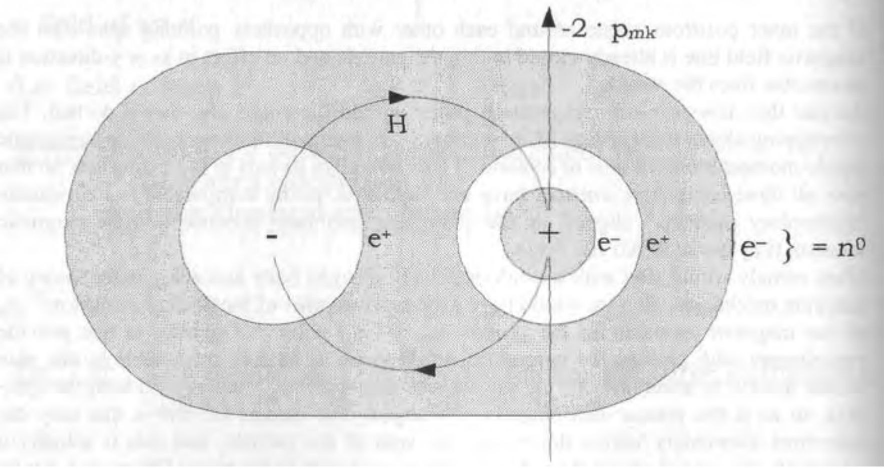
-2 · pmk
H
-
e⁺
+
e⁻
e⁺
e⁻ } = n⁰

Fig. 7.10: The neutron with magnetic dipole field H

7.10 Structure of the neutron

Until now could not be solved, why despite its missing charge also the neutron n° has a magnetic moment. The experimentally determined value is approx. the double of the nuclear magneton. Further was with measuring techniques an only 0,14% bigger mass with regard to the proton determined. The difference is approximately two and a half electron masses. And how reads the answer in the view of the potential vortex theory? It is obvious that a positively charged proton and a negatively charged electron mutually attract and amass together (fig. 7.10a). A pair annihilation can't occur, because the electron, which jackets both positrons, prevents this. The formation of an outer shell is not permitted by the high stability of the proton. It would have to be a positron shell, which instead of neutrality would produce a double positive charge. Conceivable is however the configuration, in which one of the two e^+ of the proton takes up the e^- in its inside and overlaps it (fig. 7.10b).

At first appears the amassing of p^+ and e^- to be the obvious answer to the structure of the neutron also in view of the small increase in mass. Since both elementary particles (p^+ and e^-) have a spin, will they align their axes of rotation antiparallelly and rotate against one another, exactly like an electron pair. But we now have unequal conditions: the proton brings the triple magnetic moment, the electron however only the single, and its field line will be closed by the proton. The difference which remains is the measurable double nuclear magneton, with which key question XIII (fig. 7.13) would be answered exhaustively.
This structure is shown in fig. 7.10a and has as rest mass the by only one electron mass increased proton mass, but it will deviate from this value, when the unequal partner come closer. Doing so the electron will be more strongly compressed by the heavier proton as vice versa.
Mass, magnetic moment and charge thus correspond to a large extent with the measurement values. Problems are seen concerning the spin and the stability.
Set of problems concerning spin: both the e^- and the p^+ have a half-integer spin, for which reason this configuration should have an integer spin.
Set of problems concerning stability: the neutron decays as is well-known in a p^+ and an e^- , but this object should be shorter-lived as determined by experiments. If namely the partner come each other very close, then the field strength of the p^+, contrary to expectation, doesn't increase but decreases, as is shown in fig. 7.8. The e^- therefore can only be bound very, very loosely; in z-direction it even will be repelled!
For these reasons is the open structure, which is shown in fig. 7.10a, not feasible as an isolated elementary particle, but only in a spatially extended network, like it is present in an atomic nucleus. In this case the neutron is, as is well-known, lighter by the mass defect, which is interpreted as binding energy.
Possibly it only concerns an intermediate stage. The heavier final product of the n° then could look like is shown in fig. 7.10b. For this version the line of the magnetic field already is closed partly within the particle, so that also here the approx. double nuclear magneton remains as a rest with a sense of orientation, as if the neutron were negatively charged.
Without charge and with the 1/2 spin it in this configuration fulfils all important quantum properties of the neutron, even that of the stability.

the field of the e^-: $E_{31}(-) = - E_1 (r_1/r_3)$,

the field of the e^+: $E_{32} = E_2 (r_2/r_3) = E_1 (r_2/r_3)$

and in addition the e^+: $E_{3\,1} = E_1 (r_1/r_3)$.

$$E(r_3) = E_{31}{}^{(-)} + E_{32} + E_{31} + \underset{\searrow\ \text{negligible}}{E_0}$$

With the radius relation (eq. 7.4): $r_2 = 2 * r_1$ und $r_3 = 3 * r_1$
The total field is:

$$\frac{E(r_3)}{E_1} = -\frac{r_1}{r_3} + \frac{r_2}{r_3} + \frac{r_1}{r_3} = -\frac{1}{3} + \frac{2}{3} + \frac{1}{3} = \frac{2}{3} \quad (7.12)$$

With $z_{en} = 4$ elementary vortices

$$\boxed{\frac{r_n}{r_p} = \frac{z_{ep}}{z_{en}} \cdot \frac{E_p(r_2)}{E_n(r_3)} = \frac{3}{4} \cdot \frac{1}{2/3} = \frac{9}{8} = 1{,}125} \quad (7.13)$$

n_0 is 12,5% bigger than p±

$$\boxed{\frac{m_n}{m_p} = \frac{z_{en}}{z_{ep}} \cdot \left(\frac{r_p}{r_n}\right)^2 = \frac{4}{3} \cdot \left(\frac{8}{9}\right)^2 = 1{,}05} \quad (7.14)$$

n^0 is 5% heavier than p±

Fig. 7.11: Calculation of the mass of the neutron

7.11 Calculation of the neutron

The calculation of the mass for the structure of the neutron according to fig. 7.10b has still remained open.

Because in this book for the first time has been shown, how the mass can be calculated, if the particles are understood as potential vortices, we also in this case again want to make use of this possibility.
We have, like for the μ^-, a structure of three shells with the radii r_1, r_2 and r_3. At the outer radius r_3 the fields of the elementary vortices on the inside have an effect on the electron On the outside: like is the case for the μ^-

the field of the e^-: $E_{31}^{(-)} = -E_1(r_1/r_3)$,

the field of the e^+: $E_{32} = E_2 (r_2/r_3) = E_1 (r_2/r_3)$

and in addition the e^+: $E_{31} = E_1 (r_1/r_3)$.

The total field is, with the radius relation equation 7.4:

$$\frac{E(r_3)}{E_1} = -\frac{r_1}{r_3} + \frac{r_2}{r_3} + \frac{r_1}{r_3} = -\frac{1}{3} + \frac{2}{3} + \frac{1}{3} = \frac{2}{3} \quad (7.12)$$

If we compare the neutron, in which now $z_e = 4$ elementary vortices are involved, with the proton:

$$\frac{r_n}{r_p} = \frac{z_{ep}}{z_{en}} \cdot \frac{E_p(r_2)}{E_n(r_3)} = \frac{3}{4} \cdot \frac{1}{2/3} = \frac{9}{8} = 1{,}125 \quad (7.13)$$

then we infer from the arithmetically determined result that the neutron according to the radius is 12,5% bigger than the proton. The mass is calculated to:

$$\frac{m_n}{m_p} = \frac{z_{en}}{z_{ep}} \cdot \left(\frac{r_p}{r_n}\right)^2 = \frac{4}{3} \cdot \left(\frac{8}{9}\right)^2 = 1{,}05 \quad (7.14)$$

The particle therefore has a mass which is 5% larger than for the proton, slightly more as has been measured for the neutron. The difference is acceptable. The particle after all is structured very asymmetrically, in which the reason is to be seen, why the uncharged particle, looked at from close up, nevertheless shows an observable charge distribution.

Decay equations if neutrinos are involved:

$$n^0 \longrightarrow p^+ + e^- + \bar{\nu}_e \qquad (7.15)$$

$$n^0 + \nu_e \longrightarrow p^+ + e^- \qquad (7.15^*)$$

$$\mu^- \longrightarrow e^- + \bar{\nu}_e + \nu_\mu \qquad (7.16)$$

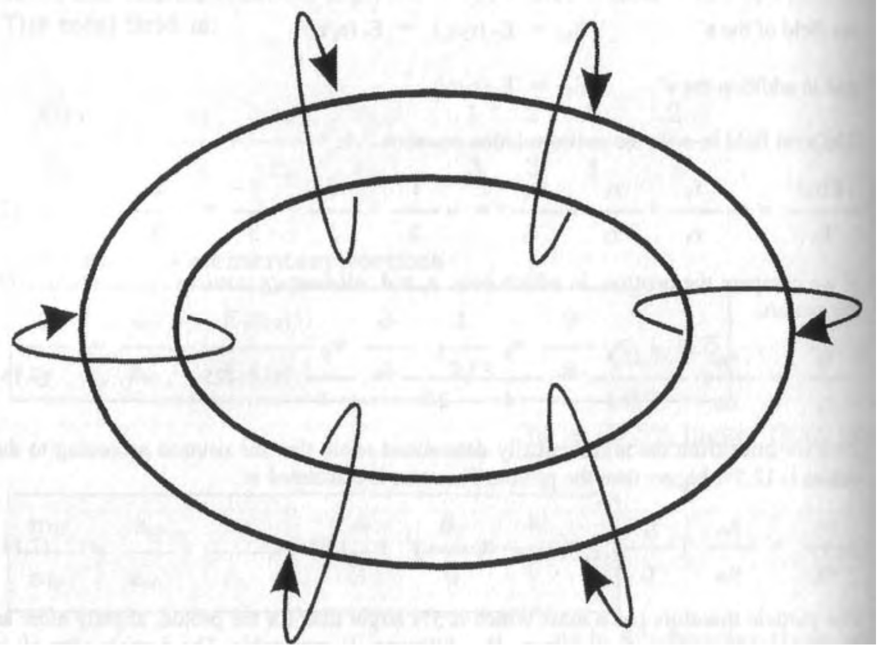

Fig. 7.12: The electron-neutrino as a ring-like vortex

7.12 ß-decay

In the case of the calculated quasistable particles, the μ^- and the n°, the verification by means of the well-known decay processes is still due. Also free neutrons, those which are not bound in an atomic nucleus, decay. But with an average life of 918 seconds they are by far the longest living among the quasistable elementary particles. Should the neutron decay be triggered by neutrinos, then obviously a distant flying past does not suffice. For that the electron is bound in the proton too tight. There probably has to occur a direct "crash", in which a neutrino is used, since the decay equation reads:

$$n^0 \longrightarrow p^+ + e^- + \overline{\nu}_e \quad (7.15)$$

As could be expected a proton p^+, an electron e^- and the mentioned electron-antineutrino ν_e are formed. What here is written down as the emission of an antiparticle, is equivalent in the absorption of the particle<i>, in this case of the neutrino. The reaction equation 7.15 can be reformulated accordingly<i>:

$$\boxed{n^0 + \nu_e \longrightarrow p^+ + e^-} \quad (7.15^*)$$

Also for the decay of the myon an electron-neutrino is used. In both cases it provides the energy necessary for the decay. But we can really understand the β-decay only, after we have got to know these particles better.

Without charge and without mass neutrinos show hardly any interactions with matter and as a consequence they possess the enormous ability of penetration - as is well-known. They are said to participate in the ,,weak interaction", which should trigger a conversion of the concerned particles, which is their decay. Pauli already has postulated the neutrino 1930 theoretically, because the transition from a half-integer spin to an integer spin for the n^0 -decay otherwise wouldn't have been explicable.

If we imagine an elementary vortex is being born, but the local field strength and energy isn't sufficient for obtaining a quantized state. The result is an incomplete potential vortex, which has an open vortex centre and as a consequence shows no localization at all. In the form of a vortex ring it oscillates around itself, while it continually turns its inside to the outside and then again to the inside.

One moment the vortex ring is green, then it is red again, one moment matter, then anti-matter, one moment positively charged and the next moment negatively charged. In contrast to the photon the number of the involved elementary vortices z_e for the neutrino is odd (for the ν_e z_e= 1). Perpendicular to the direction of propagation the neutrino has a spin $(s/\hbar = \frac{1}{2})$ for reason of a rotation, which overlaps the pulsating oscillation. This vortex ring is, as said, not a member of stationary matter, because it doesn't form a "black hole" in its centre, where the speed of light becomes zero. But it has an absolute stability like every elementary vortex, even if it only occurs incomplete and hence not in any quantized form,. This concept of the electron-neutrino as an open oscillating elementary vortex in the form of a ring-like vortex covers the experimentally determined realizations unexpectedly well.

<i>: Kussner, H.G.: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder, Musterschmidt, Gottingen 1976, S.155

A strong interaction doesn't exist. The electric field in the proximity of the proton goes to zero within the range which is determined with measuring techniques.

A weak interaction doesn't exist. That interaction only is a special case of the electromagnetic interaction which appears in a weakened form.

XII: Why does the proton have approximately 3 times the magnetic moment which can be expected for reason of the only single charge?
(3 elementary vortices)

XIII: Why does the neutron as an uncharged particle anyway have a magnetic moment?
(Structure of the n°)

XIV: What owes the atomic nucleus, which consists of like charges, its stability?
(Course of the field of the p^+, instead of "strong interaction")

XV: Why does the free neutron decay, although it is stable as a particle of the nucleus?
(Interaction with neutrinos)

XVI: Why do neutrinos nevertheless participate in the "weak interaction", although they have no mass and no charge?
(Oscillating charge)

XVII: How can be given reasons for the finite range of the "weak interaction"?
(Reaction cross-section for particle decay)

Fig. 7.13: Further key questions of quantum physics (Continuation of figures 4.4 and 6.13)

7.13 "Weak interaction"

Let's now look again at the ß-decay of the neutron, in which a neutrino ν_e is used. But this by no means will be a process of the weak interaction. Instead will neutrinos, contrary to the textbook opinion, participate in the electromagnetic interaction. They after all are one moment positively charged and the next moment negatively charged. With slow-acting gauges this it is true can't be proven, because the interaction is zero on the average. But this charged oscillating vortex ring can exert a considerable effect while approaching a neutron, which is based solely on the electromagnetic interaction.

The neutron is stimulated to synchronous oscillations of its own by the high-frequency alternating field of the neutrino, until it in the case of the collision releases the bound electron, which takes up the energy provided by the neutrino and transports it away. The interaction obviously is only very weak due to the oscillation. But a physical independency of it has to be disputed.

The finite range, which is given in this context, indicates the reaction cross-section around the n°-particle, within which the "crash" and as a consequence the β-decay occurs. This range is considerable larger as the particle itself. The electromagnetic interaction for such small distances after all is so violent, even if it only occurs in pulses, that the neutrino is thrown out of its path and can fly directly towards the neutron.

Perhaps we now understand also the β-decay of the myon. It actually were to be expected that without outside disturbance an absolute stability could exist because of the ideal symmetry of the μ^-. On our planet we however are in every second bombarded with approx. 66 milliard (billion) neutrinos per cm^2 <i>. Obviously it takes 2,2 μs on the average till a neutrino ν_e flies past a myon so close that it decays. In doing so it stimulates the outside elementary vortex to violent oscillations by trying to synchronize it. In this case the electron-neutrino ν_e carries away with it the two outer, and therefore weaker bound, elementary vortices of the myon, which meanwhile are oscillating synchronously. The innermost vortex, an electron e^-, is left behind. The decay of the myon which takes place with a probability of almost 100 % reads:

$$\mu^- \longrightarrow e^- + \bar{\nu}_e + \nu_\mu \qquad \text{resp.} \qquad \mu^- + \nu_e \longrightarrow e^- + \nu_\mu \tag{7.16}$$

Thus a different neutrino ν_μ, is formed which can be distinguished from the ν_e and is called myon-neutrino since it forms from the μ^-. Actually it even has a similar structure of three shells, as is shown in fig. 7.5. But the vortex centre is open and the particle isn't stationary anymore. In the picture now only a momentarily state is shown, in which the ν_μ appears green on the outside and red in its open centre. As already for the ν_e oscillates also here the inside to the outside and vice versa, this time merely as a packet of three shells, so that also this particle shows all the typical neutrino properties discussed for the example of the ν_e.

The for potential vortices typical and already discussed phenomenon of transport here has an effect. In particular in connexion with vortex rings this property is known from hydrodynamics. It thus can be observed, how vortex rings bind matter and carry away with them. Because the neutrino is not quantized, it neither is restricted with regard to its ability to transport elementary vortices. Consequently even bigger configurations are conceivable, like configurations of 5 shells, 7 shells etc..

<i>: "Zeugen aus der Sonne", VDI-Nachrichten Nr. 45 vom 9.11.90, Seite 26

a: the structure of the tau-neutrino with 5 shells

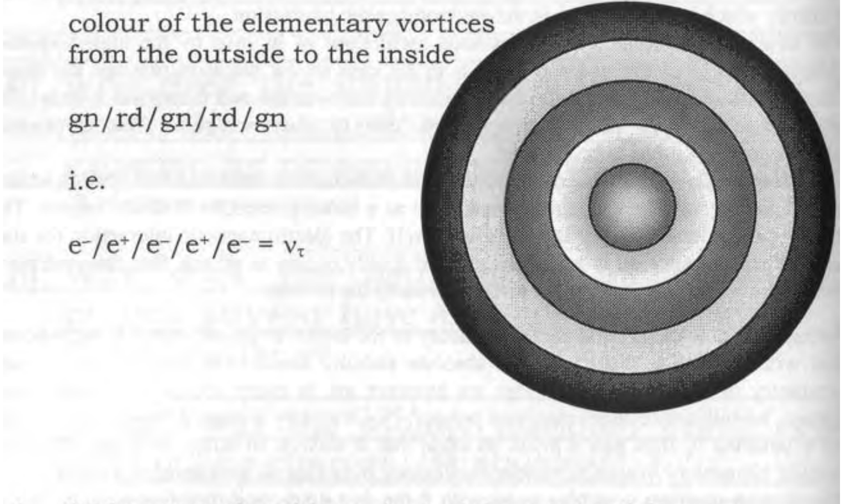
colour of the elementary vortices
from the outside to the inside
gn/rd/gn/rd/gn
i.e.
$e^-/e^+/e^-/e^+/e^- = \nu_\tau$

b: the heavy and stable variant of the tau particle

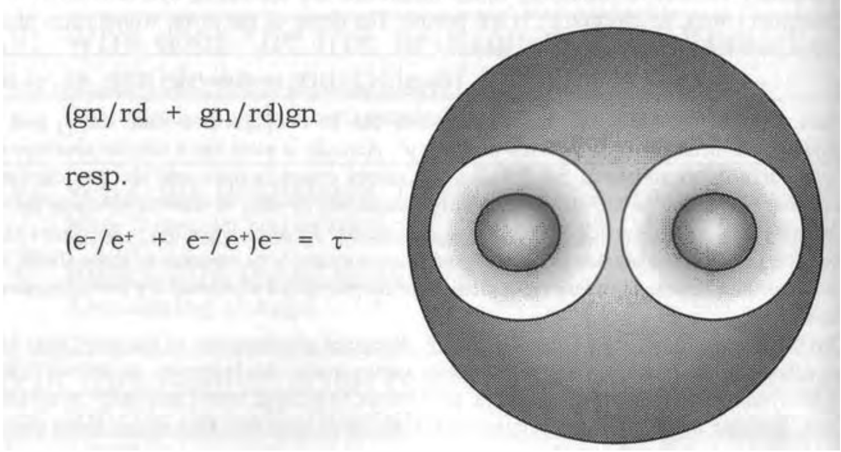
(gn/rd + gn/rd)gn
resp.
$(e^-/e^+ + e^-/e^+)e^- = \tau^-$

Fig. 7.14: Tau-neutrino and tau particle

7.14 Tau particle

In the table of the leptons after the e^- and the μ^- as the next particle the tau particle τ^- is found with its accompanying neutrino ν_τ.The obvious solution for the tau particle is the structure of five shells, as is shown in fig. 7.14a. With that the electron would have another particularly heavy relative with otherwise very similar properties. For the myon the neutrino was stable, the particle itself however instable. We after all huve explained the particle decay as a consequence of an outside disturbance, and disturbances always are based on interactions. Correspondingly should, with the small possibility for an interaction, also the neutrino ν_τ of the tau particle have a better stability than the particle τ^- itself. Without doubt this structure of 5 shells fulfils all known quantum properties like spin, charge etc. Merely the check of the mass is still due. This we now want to calculate for the structure shown in fig. 7.14a.

$$m_5 = m_p \cdot \left(\frac{Z_{e\tau}}{Z_{ep}}\right)^3 \cdot \left(-\frac{r_1}{r_5} + \frac{r_2}{r_5} - \frac{r_3}{r_5} + \frac{r_4}{r_5}\right)^2 \quad (7.17)$$

$$m_5 = 1836 \cdot \left(\frac{5}{3}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot m_e = \underline{\underline{1360 \cdot m_e}} \quad (7.17^*)$$

But the for the tau particle measured value is considerable higher!
Even if this structure is the only possible in the case of the neutrino ν_τ for reason of the complete symmetry, will the tau particle however change its structure by itself if another structure exists, which is more stable, thus in which the particle can take a bigger mass. Such a maximum provides the structure shown in fig. 7.14b after checking all possible configurations with five elementary vortices:

$$m_\tau = m_p \cdot \left(\frac{Z_{e\tau}}{Z_{ep}}\right)^3 \cdot \left(\left(-\frac{r_1}{r_3} + \frac{r_2}{r_3}\right) \cdot 2\right)^2 \quad (7.18)$$

$$m_\tau = 1836 \cdot \left(\frac{5}{3}\right)^3 \cdot \left(\frac{2}{3}\right)^2 \cdot m_e = \underline{\underline{3778 \cdot m_e}} \quad (7.18^*)$$

This value now lies 8% above the measurement values. It would be obvious, if unbound tau particles predominantly would take the structure shown in fig. 7.14b. The remaining error becomes explicable, if a very small number of tau particles in the lighter structure according to fig. 7.14a are involved with a correspondingly smaller probability.

The enormous variety of kinds of decay, and not a single one of the dominating ones has a probability of over 50%, makes it more difficult for us, to be able to directly infer the inner structure of a particle from the decay products. It nevertheless should be mentioned that after all 35% of all decays take place by taking up and using a neutrino ν_e or ν_τ, entirely in accordance with the model of the myon decay (equation 7.16).

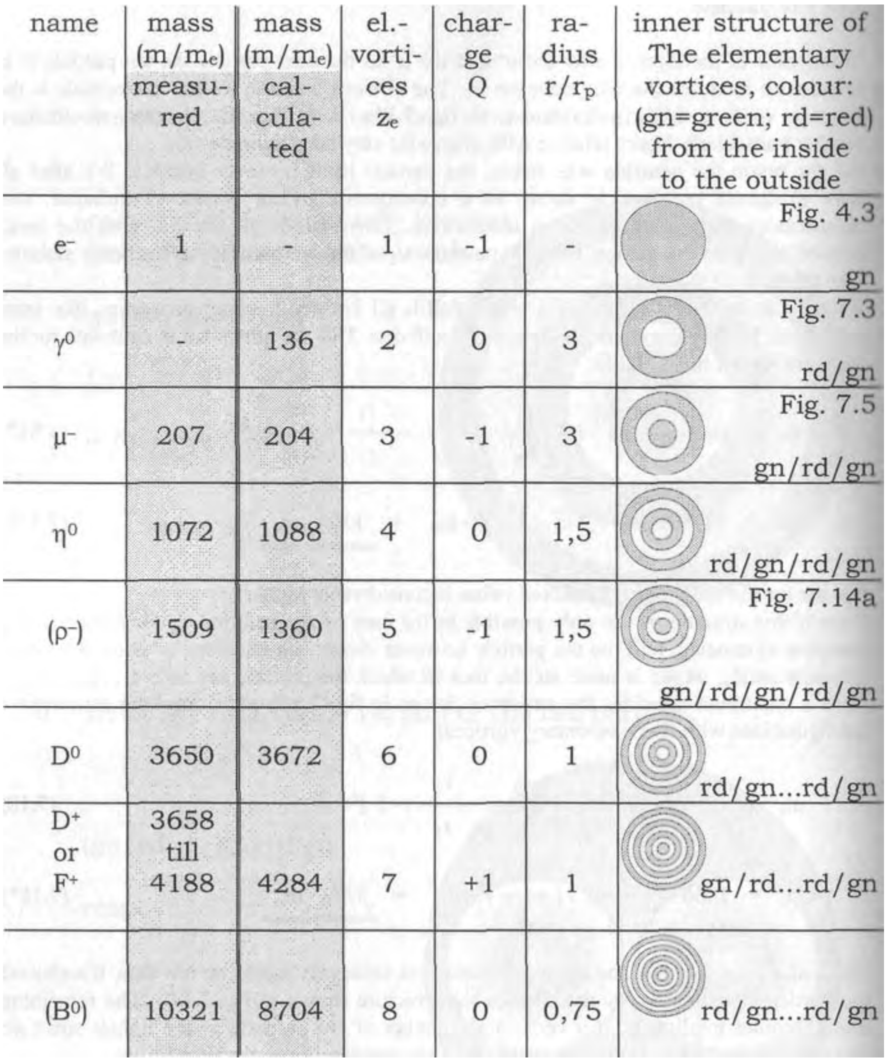

name	mass (m/m_e) measured	mass (m/m_e) calculated	el.-vortices z_e	charge Q	radius r/r_p	inner structure of The elementary vortices, colour: (gn=green; rd=red) from the inside to the outside
e^-	1	-	1	-1	-	Fig. 4.3 gn
γ^0	-	136	2	0	3	Fig. 7.3 rd/gn
μ^-	207	204	3	-1	3	Fig. 7.5 gn/rd/gn
η^0	1072	1088	4	0	1,5	rd/gn/rd/gn
(ρ^-)	1509	1360	5	-1	1,5	Fig. 7.14a gn/rd/gn/rd/gn
D^0	3650	3672	6	0	1	rd/gn...rd/gn
D^+ or F^+	3658 till 4188	4284	7	+1	1	gn/rd...rd/gn
(B^0)	10321	8704	8	0	0,75	rd/gn...rd/gn

7.15 Table of vortices of the calculated leptons and mesons compared with measurement values (Part 1).

7.15 Pions

Unlike the leptons, which we could derive and calculate fairly completely, the mesons don't have a half-integer spin. With this characteristic property they therefore can't represent an individually overlapped elementary particle and they probably will consist of the amassing in pairs of individual configurations of potential vortices. This kind of bond can't be particularly tight. Consequently we don't know any stable mesons.

The most important basic building part of the mesons we have got to know over the positronium in fig. 7.3. It necessarily has to amass to another particle, otherwise it annihilates under emission of a γ-quanta, as already mentioned. This γ^0-particle, as it will be named here, has the mass of:

$$m_{\gamma 0} = (2/3)^3 \cdot (1/2)^2 \cdot 1836 \cdot m_e = 136 \cdot m_e \quad (7.19)$$

which only can be determined arithmetically. As a partner, to which the γ^0-particle can amass, first of all another γ^0-particle should be considered. Because both partner will rotate against one another, this new particle would not have a spin and moreover would be uncharged. The mass now would be twice as big with:

$$m_{\pi 0} = 2 \cdot m_{\gamma 0} = 272 \cdot m_e . \quad (7.19^*)$$

But the two γ^0-particles will come very close together and mutually feel the local, in the same direction orientated, distribution of the field, which will lead to a weakening of the field and as a consequence to a slight reduction of the mass.

With these properties it probably concerns the uncharged pion π^0. This model concept finds an excellent confirmation in the two possible kinds of decay, which can be regarded as equivalent:

$\pi^0 \longrightarrow \gamma + \gamma$ with a probability of 99%

and

$\pi^0 \longrightarrow \gamma + e^- + e^+$ with a probability of 1%

Also in the case of the charged pion $\pi^\pm$ the observable decay offers a big help, which will take place with a frequency of almost 100 %:

$$\pi^\pm \longrightarrow \mu^\pm + \nu_\mu$$

The equation doesn't state anything about the fact, if a neutrino ν_e is used in the process. But it points at the circumstance that the partner of the γ^0-particle for the $\pi^\pm$ most likely is a myon $\mu^\pm$. The mass will be smaller than the sum of both building parts:

$$(204+136) * m_e = 340 * m_e.$$

name	mass (m/m_e) measured	masse (m/m_e) calculated	elem.-vortices z_e	charge Q	inner structure of The elementary vortices, colour: (gn=green; rd=red) from the inside to the outside
B^-	11194	9793 +?	9 +?	-1	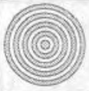 r/r_p = 0,75 gn/rt...rt/gn
(Y^0)	18513	17001 +?	10 +?	0	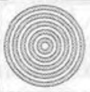 r/r_p = 0,6 rt/gn...rt/gn

Some compound configurations

π^0	264	272	4	0	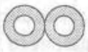 2 x fig. 7.3 $\gamma^0 + \gamma^0 = \pi^0$
π^-	273	340	5	-1	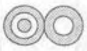 fig. 7.3 + fig. 7.5 $\mu^- + \gamma^0 = \pi^-$
τ^-	3487	3778	5	-1	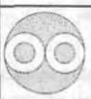 fig. 7.14b (2x gn/rd)/gn
(K^0)	974	918 +?	6 +?	0	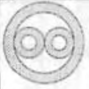 r/r_p = 2 (2x rd/gn)/ rd/gn
(K^-)	967	939 +?	11 +?	-1	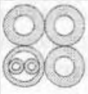 (rd/gn + gn)/ rd/gn + 3 γ^0

Fig. 7.16: Table of vortices of the calculated leptons and mesons compared with measurement values (Part 2).

7.16 Table of vortices of the mesons

The numerous kinds of decay for K-mesons suggest that these strange particles will consist of various combinations of amassed together and in pairs rotating γ^0- and $\mu^\pm$-particles. The possibilities of combination now already have increased in such a way that for every kaon and other mesons several solutions can be proposed. To avoid unfounded speculations, only a few clues will be given.

Besides the γ^0-particles also heavier arrangements should be considered as partner for the spin and as a building part for kaons and other mesons.
If for instance a π^0 is overlapped by a γ^0, then this particle has an arithmetically determined mass of 918 m_e. It therefore can concern a building part of the uncharged kaon K^0.

The likewise with three γ^0 formed configuration of 6 shells however, if it actually would staystable for the duration of a measurement, would have the mass of 3672 electron masses<i>.

A very much better detectability must be attributed to the configuration of 4 shells which consists of two γ^0, so to speak a heavy relative of the γ^0 and the π^0. It among others should be able to decay like a π^0.With this property and with an arithmetically determined mass of 1088 m_e it actually only can concern the η^0-meson. Solely according to the numeric value the η^0-meson could also consist of four π^0-mesons; but the decay in only two light quants speaks against it.

The kaon-puzzle in addition is made more difficult by the spontaneously possible ability to change of the involved γ^0-particles during a process of decay, as is made clear by the numerous kinds of decay. These dependent pion halves can be "swallowed" or "spit out" by neutrinos in the process, they can form from incident light or be emitted as photons and eventually they even can break up in their individual parts.

In fig. 7.16 the possible configurations of potential vortices are sketched and the respective, according to the new theory calculated, mass is given. If above that the other decay products and quantum properties, which can be given for the vortex structures, are added, like e.g. charge, spin and if need be magnetic moment, then an assignment without doubts to the until now only from measurements known elementary particles is possible. In order to better be able to assess the efficiency of the potential vortex theory, the measurement values are compared to the calculated values.

Some terms are put in brackets, because it can be assumed that the calculated part only concerns the dominating part, to which further γ^0or other small configurations of vortices will amass for reason of its high mass. Correspondingly should the mass in that case be corrected slightly.

<i>: It could e.g. concern the D°-meson.

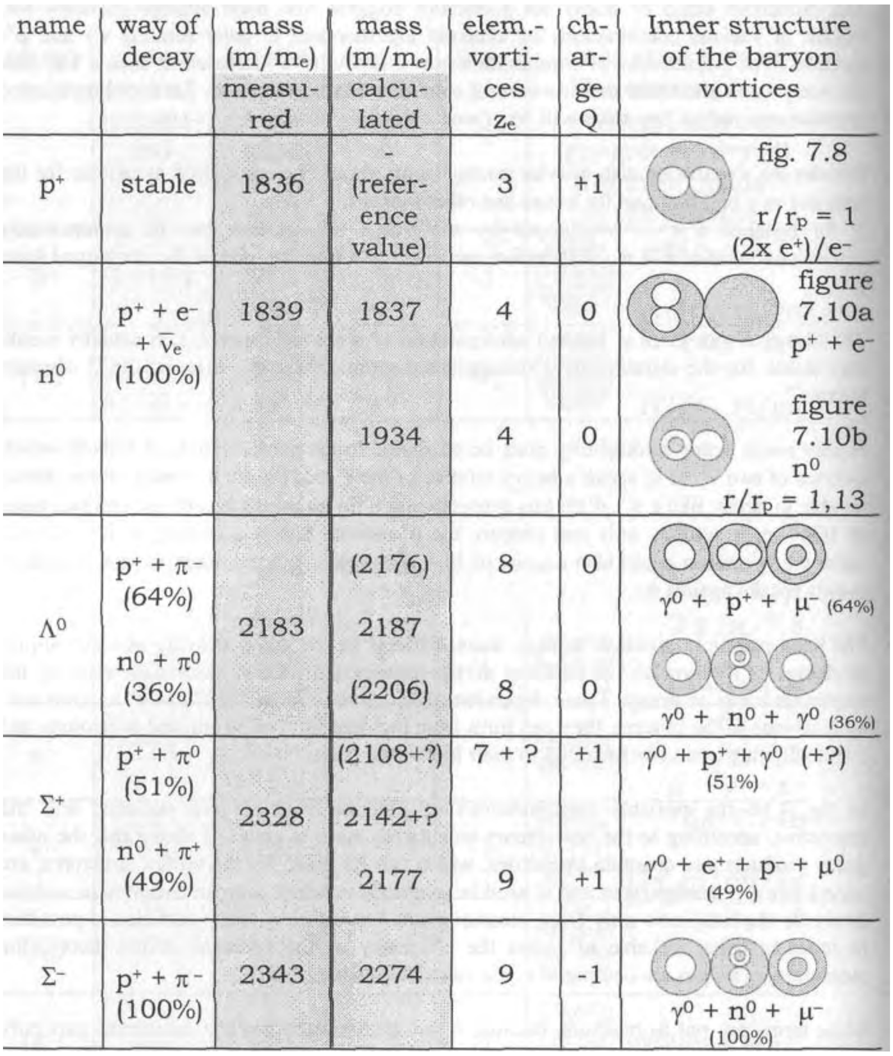

name	way of decay	mass (m/m_e) measured	mass (m/m_e) calculated	elem.-vortices z_e	charge Q	Inner structure of the baryon vortices
p^+	stable	1836	- (reference value)	3	+1	fig. 7.8 $r/r_p = 1$ $(2x\ e^+)/e^-$
n^0	$p^+ + e^- + \overline{\nu}_e$ (100%)	1839	1837	4	0	figure 7.10a $p^+ + e^-$
			1934	4	0	figure 7.10b n^0 $r/r_p = 1,13$
Λ^0	$p^+ + \pi^-$ (64%)	2183	(2176)	8	0	$\gamma^0 + p^+ + \mu^-$ (64%)
			2187			
	$n^0 + \pi^0$ (36%)		(2206)	8	0	$\gamma^0 + n^0 + \gamma^0$ (36%)
Σ^+	$p^+ + \pi^0$ (51%)	2328	(2108+?)	7 + ?	+1	$\gamma^0 + p^+ + \gamma^0$ (+?) (51%)
			2142+?			
	$n^0 + \pi^+$ (49%)		2177	9	+1	$\gamma^0 + e^+ + p^- + \mu^0$ (49%)
Σ^-	$p^+ + \pi^-$ (100%)	2343	2274	9	-1	$\gamma^0 + n^0 + \mu^-$ (100%)

Fig. 7.17: Table of vortices used for the calculation of the most
important barvons with suggestions for the structure (Part 3).

7.17 Table of vortices of the baryons

The number of possibilities of combination quickly increases, if only a few elementary vortices extend the structure of a particle. This probably is the reason for the large number of observable hyperons, which recently have been produced artificially and observed with the help of particle accelerators.

Both the neutron and the lambda particle can exist in a lighter and a heavier variant. At the moment of the decay, as it for instance is observed in a bubble chamber, according to expectation the state with the smaller mass takes the bigger probability. But in the amassing with further particles as building part of bigger and heavier hyperons the heavier structure is more likely. This circumstance should be considered in calculating the mass of the hyperons.

In figures 7.17 and 7.18 the most important baryons are listed, which are characterised in the way that one of the amassed together packets of vortices is a nucleon, thus a proton or a neutron.

The given, from measurements known, kinds of decay are able to confirm the inner structure pretty good. Of course an infinitely lot of combinations are conceivable and numerous predictions are possible. But speculations are unnecessary from the time on where we are able to calculate the particles!
The restriction to the few in the table listed particles seeming to be important hence doesn't limit the universal importance of the theory of objectivity in any way!

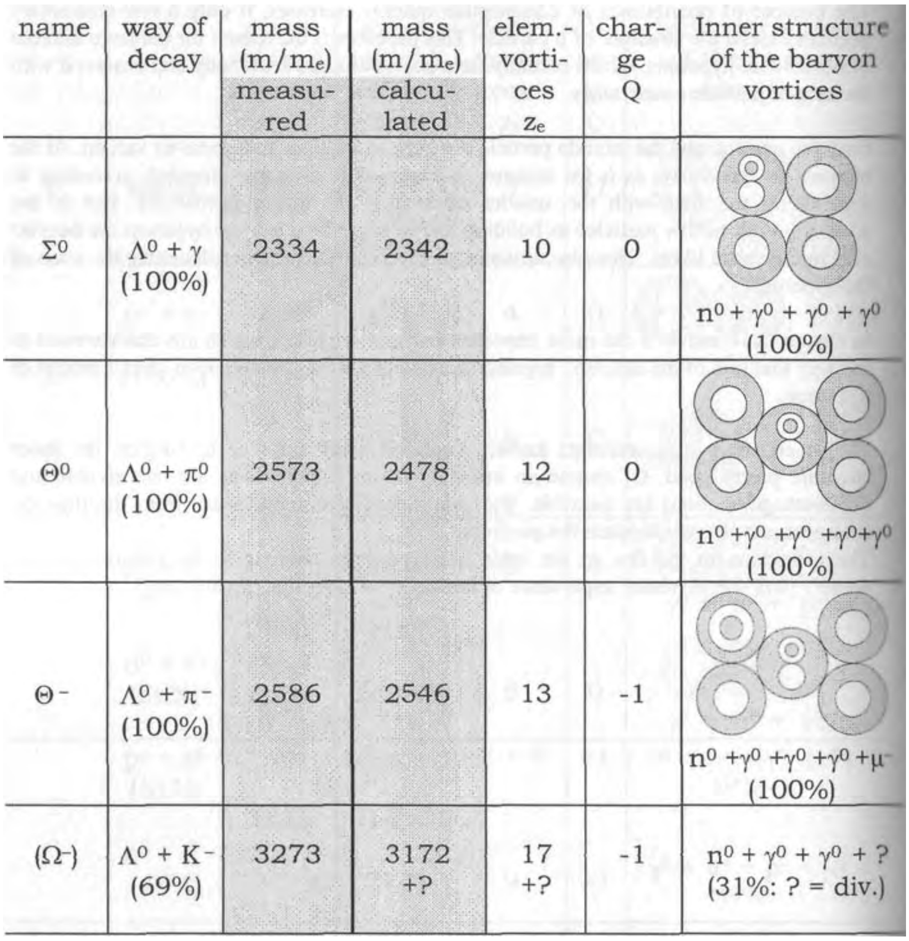

name	way of decay	mass (m/m_e) measured	mass (m/m_e) calculated	elem.-vortices z_e	charge Q	Inner structure of the baryon vortices
Σ^0	$\Lambda^0 + \gamma$ (100%)	2334	2342	10	0	$n^0 + \gamma^0 + \gamma^0 + \gamma^0$ (100%)
Θ^0	$\Lambda^0 + \pi^0$ (100%)	2573	2478	12	0	$n^0 + \gamma^0 + \gamma^0 + \gamma^0 + \gamma^0$ (100%)
Θ^-	$\Lambda^0 + \pi^-$ (100%)	2586	2546	13	-1	$n^0 + \gamma^0 + \gamma^0 + \gamma^0 + \mu^-$ (100%)
(Ω^-)	$\Lambda^0 + K^-$ (69%)	3273	3172 +?	17 +?	-1	$n^0 + \gamma^0 + \gamma^0 + ?$ (31%: ? = div.)

Fig. 7.18: Table of vortices used for the calculation of the most important baryons with suggestions for the structure (Part 4).

8. Unified theory

With the theory of objectivity the longed for goal of a "theory of everything" (TOE), of an universal theory, seems to have moved within reach. If in the nineteenth century still promising field theories and approaches were being discussed, then has at the latest Einstein's theory of relativity destroyed all hopes in such a theory. Science as a consequence has become very much more modest and understands a TOE only as the unification of all known interactions.

Einstein has stated the minimum demand so: "a theory should be favoured by far, in which the gravitational field and the electromagnetic field together would appear as a whole" <i>. It is evident that a subjective or relativistic observer theory never is able to achieve this.

The presented theory of objectivity made it possible that the unification here for the first time actually has succeeded. This undoubtedly brings science a whole lot further, but it still is not sufficient to lie one's hands in one's lap being content with oneself. After all we still know very much more phenomena, which likewise should be unified. After all it is no accident that both Maxwell and Einstein, to name only two prominent representatives, after completion of their well-known works have struggled for the question, what sort of phenomenon it concerns in the case of the temperature and how this could be integrated in their theory.

The requirement reads: We must be able to derive all basic factors, which influence our system of units with their basic units, as a compulsionless result from the new theory. Besides the dimensions of space and time which determine our continuum, the explanation and unification of the basic factors mass and charge has to be tackled. If we have succeeded in doing so, we'll also tackle the problem of the fifth and last basic factor, which until now has put itself in the way of any unified theory as the question of fate, the problem of the temperature!

<i>: Einstein, A.: Grundziige der Relativitatstheorie, Vieweg+Sohn, Braunschweig 1973, 5. Aufl., WTB 58. Seite 97.

"A theory should be favoured by far, in which the gravitational field and the electromagnetic field together would appear as a whole"<i>.

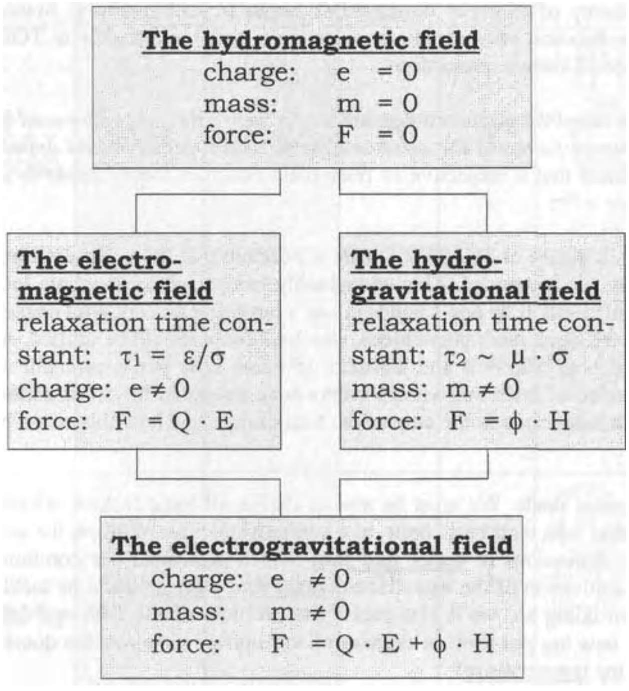
The hydromagnetic field
charge: $e = 0$
mass: $m = 0$
force: $F = 0$
The electro-magnetic field
relaxation time constant: $\tau_1 = \varepsilon/\sigma$
charge: $e \neq 0$
force: $F = Q \cdot E$
The hydro-gravitational field
relaxation time constant: $\tau_2 \sim \mu \cdot \sigma$
mass: $m \neq 0$
force: $F = \phi \cdot H$
The electrogravitational field
charge: $e \neq 0$
mass: $m \neq 0$
force: $F = Q \cdot E + \phi \cdot H$

Fig. 8.1: Structuring of the fields and definition of terms

<i>: Einstein, A. : Grundziige der Relativitatstheorie, Vieweg + Sohn, Braunschweig 1973, 5. Aufl. , WTB 58.

<ii>: derived from the Greek hydro (= water).

8.1 Structure of the field theory

In contrast to Maxwell's theory the new field theory, which we derived from duality, is also able to describe fields, in which no particles and no quanta exist. It probably is justified and useful in the sense of a clearer communication, to give the new field a name of its own.
The author recommends the introduction of the term "hydrotic field". In it should be expressed, which importance water has for both the like named potential vortex and this field<ii>.
As we already have worked out, the hydrotic field is favoured particularly by polar materials and by a high dielectricity. Water is a corresponding and in the biosphere of our planet dominating material.

Whereas we had to correct the concept of a vortex free electric field, we had until now, considerable, we can take over the description of the magnetic field unchanged. This then should also be valid for its name. The new field which consists of both correspondingly is called hydromagnetic field.

In fig. 8.1 we recognize the structure. At the top stands the "hydromagnetic field", which is described mathematically by the equations of dual electrodynamics in fig. 3.3. It does not know quanta and as logical consequence neither charge nor mass! If we insert these equations, Ampere's law and the dual formulated Faraday law of induction, into each other, then there results as a mathematical description of our space-time-continuum the fundamental field equation (5.7, fig. 5.1). As a new physical phenomenon the potential vortex appears, which gives the hydromagnetic field a new and important property: this field can be quantized!

Starting-point is the wave, which for corresponding interference effects can spontaneously roll up to a vortex, which as highly concentrated spherical vortex finds a new right to exist and finds to a new physical reality.

The in the described manner formed particles show specific properties of their own. We now are able to attribute them for instance a charge or a mass. And these properties also can be investigated and described individually and isolated from each other. Thus are formed the two special cases, strange by nature, on the one hand the well-known, with the help of the Maxwell equations describable "electromagnetic field" and on the other hand the new "hydrogravitational field".

If we overlap the results of the two special cases, e.g. by adding the force effects of electric charges and accelerated masses, then we summarized obtain a field, which we accordingly should call "electrogravitational". This case is not at all unknown. Already Niels Bohr in this way has calculated the radii of the electron orbits in the hull of his model of the atom, to mention only one example. We can summarize:

The hydromagnetic field is the all encompassing and with that most important field. Apart from that the electromagnetic field of the currents and the eddy currents and the hydro-gravitational field of the potentials and the potential vortices merely describe the two possible and important special cases. For reasons of pure usefulness for every special case a characteristic factor of description is introduced, the charge and the mass!

Auxiliary terms (description of quantum properties):
* mass * charge * temperature * Planck's quantum of action

The hydromagnetic field does not know any quanta!

Structure of the fundamental field equation 5.7 (fig. 5.1):

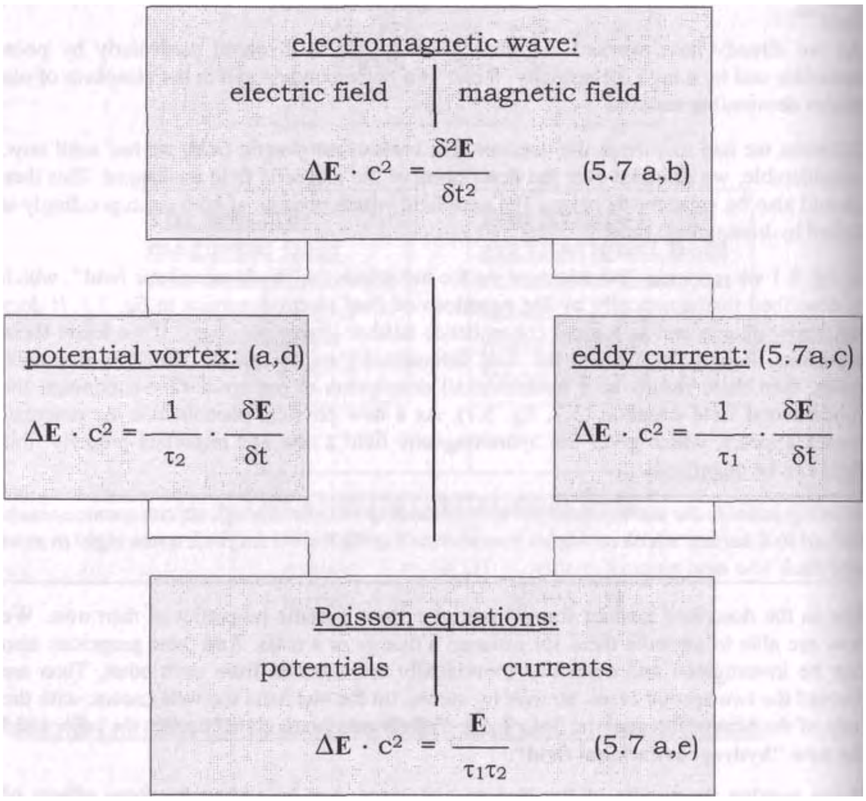
electromagnetic wave:
electric field
magnetic field
$\Delta \mathbf{E} \cdot c^2 = \frac{\delta^2 \mathbf{E}}{\delta t^2}$ (5.7 a,b)
potential vortex: (a,d)
$\Delta \mathbf{E} \cdot c^2 = \frac{1}{\tau_2} \cdot \frac{\delta \mathbf{E}}{\delta t}$
eddy current: (5.7a,c)
$\Delta \mathbf{E} \cdot c^2 = \frac{1}{\tau_1} \cdot \frac{\delta \mathbf{E}}{\delta t}$
Poisson equations:
potentials
currents
$\Delta \mathbf{E} \cdot c^2 = \frac{\mathbf{E}}{\tau_1 \tau_2}$ (5.7 a,e)

Fig. 8.2: Unified theory

*electromagnetic interaction (open field lines)
*gravitation (closed field lines)
*strong interaction (does not exist)
*weak interaction (only special aspect)

The interactions are a result of the
field dependent speed of light!

8.2 Unification of the interactions

The discovery and introduction of the hydromagnetic field makes the desired unification possible, because the electromagnetic resp. Maxwell field, which describes the electromagnetic interaction, and the hydrogravitational field of the gravitation can be derived from this field as a consequence of the formation of quanta.
The kind of the interaction is caused by the course of the field lines of the field quanta which form as spherical vortices: the open field lines make the electromagnetic interaction possible. And the field, lines with a closed course lead to gravitation. Both are a direct result of the field dependent speed of light. A more perfect unification seems hardly possible.

As the next step the unification with the strong and the weak interaction is required, but it could be shown that those don't exist at all. It just concerns misinterpretations with much fantasy, which should help explain the difference between a wrong theory and the physical reality.

Numerous auxiliary terms for the description of the quantum properties exist, like for instance mass, charge or Planck's quantum of action. The prerequisite for their usability naturally is the existence of the quanta. But until these have found to a physical reality, the auxiliary terms are unnecessary. The hydromagnetic field does not know quanta, quantum properties or auxiliary descriptions. It will be shown that, according to expectation, also the temperature is a typical quantum property, which comes within the group of the auxiliary terms. In this way also the temperature is fitted into the unified theory without compulsion.

Without the by us for reasons of usefulness introduced auxiliary terms the fundamental field equation is left with its description of a spatial-temporal principle. If a world equation should exist, then this field equation 5.7 has the best prerequisites. For the fundamental field equation the division in four parts is repeated like already for the hydromagnetic field (fig. 8.1). It likewise consists of four individual parts, the wave (b), the two vortex phenomena (c and d) and the time independent term (e) (fig. 8.2). Whereas the duality still is combined in the wave, it comes to light clearly for the vortices to again be combined in the fourth case. Here arise however potentials and currents, which again can react and oscillate with each other, for instance as L-C-resonant circuit in an electronic circuit, with which the principle is repeated.
This principle is shown clearer for the phenomenon of the temperature as in all other cases. If we start at the top in the picture in fig. 8.2 we have an electromagnetic wave, which is absorbed and thus becomes a vortex. If the vortex falls apart, then eddy losses are formed. We observe that the temperature rises and propagates in the well-known manner. We have arrived in the bottom box, but this again can be taken as the top box for the now following process, because the equation of heat conduction is a vortex equation of type c or d! We discover a self-similarity:

> The spatial-temporal principle formulated mathematically by the fundamental field equation can be carried over into itself time and again.

Temperature is the oscillation of contraction of the elementary vortices resulting from the speed of light depending on field strength.

a. at absolute zero temperature:

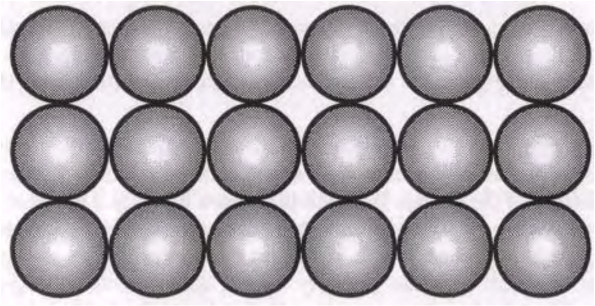

b. if thermally excited:

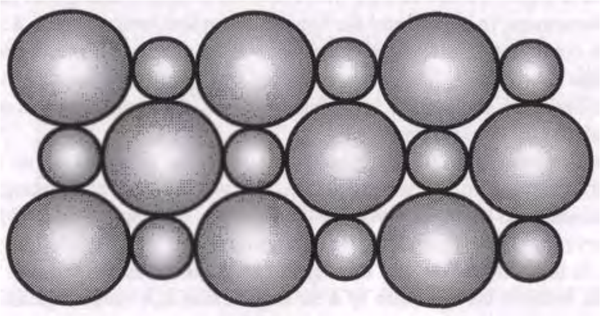

Fig. 8.3: Temperature as an oscillation of size for the speed of light depending on field strength

8.3 Temperature

Following the atomic view, in the case of heat it concerns kinetic energy of the molecules, which carry out more or less violent oscillations. In the case of gaseous materials with this concept, basing on mechanical models, actually successful calculations are possible, like for instance the speed distribution of gases won by Maxwell from theoretical considerations concerning probability.

But the attempt to apply the formulas of the kinetic theory of gases to solids and liquids only succeeds, if additional supplements and improvements are introduced. Since at all events it concerns temperature, thus the same physical quantity, of course also an uniform interpretation should be demanded, which in addition should stand in full accord to the presented design of an integrated theory (TOE).

Against the background of the new theory of objectivity we consider, what happens, if for instance the local field strength is increased by a flying past particle. The matter located at this point is contracted for a short time. By coming closer to each other, the individual elementary vortices mutually reinforce their field and are further compressed. Sometime this process comes to a standstill, is reversed and swings back.

At the same time every single particle, which in this way carries out an oscillation of size, has an effect on its neighbours with its field, to also stimulate these to the same oscillation, but delayed by some time. This phenomenon spreads in all directions. The propagation only will become stationary, if all neighbouring elementary vortices pulsate with the same amplitude. It now should be recorded:

The oscillation of contraction of the elementary vortices we call temperature.

Also this thermodynamic state variable therefore is a result of the variable speed of light. At the absolute zero of temperature no oscillation takes place anymore, whereas the upper limit lies in infinity. Since the cause for temperature represents an oscillation of the local electromagnetic field strength around the cosmic field strength, the following phenomena must be considered as excitation and cause, as dictated by the fundamental field equation 5.7:

1. Electromagnetic waves (b) are able to stimulate matter particles to synchronous oscillations of contraction by their alternating field. In doing so energy in form of heat is transferred to the particles, with the result that their temperature is increased. The wave is absorbed completely, if the thermal oscillation corresponds with the frequency of the wave.

We speak of thermal radiation.

2. But also the two dual vortices, the eddy current (c) and the potential vortex (d) can cause oscillations of contraction. This immediately becomes clear, if we consider a vortex as the special case of the wave, in which the oscillation takes place around a more or less stationary vortex centre. In the case of the decay of vortices, of the transition of energy from vortices to matter, the increase in temperature is measurable.

In the case of this process of diffusion we speak of eddy losses and of loss heat.

Answers to open questions of thermodynamics:

1. Temperature occurs independent of the state in which the matter is (unified theory).

2. Temperature even occurs in solids, where a purely kinetic interpretation fails (unification).

3. Each elementary particle is carrier of a temperature.

4. Expansion with increasing temperature because of the increasing need for room for larger amplitude of oscillation (principle: bi-metal-thermometer).

5. For solids the thermal oscillation of size is primarily passed on by the electrons in the atomic hull. Good electric conductors therefore at the same time also have a high thermal conductivity. (principle: electrical resistance thermometer).

6. For gases the entire atoms carry out this task, for which reason a kinetic auxiliary description becomes applicable.

7. For extreme amplitudes of oscillation the atoms partly or entirely lose their enveloping electrons, when they change into the plasma state.

8. The second law of thermodynamics loses its claim to be absolute and at best reads: with today's technology we are not capable, to design a cyclic working machine, which does nothing else, as to withdraw heat from a heat container and to convert it into mechanical work.

Fig. 8.4: Questions concerning thermodynamics

3. Flying past particles, in particular unbound and free movable charge carriers (e) produce an alternating field for other fixed particles. Doing so kinetic energy can be transformed in temperature, thus in energy of pulsation. A good example is the inelastic collision. But it can also be pointed to numerous chemical reactions. Whoever searches for a concrete example, takes two objects in his hands and rubs them against one another. In that case the particles which are at the frictional surfaces are being moved past each other in very small distance, in this way causing oscillations of pulsation, which propagate into the inside of the objects according to the thermal conductivity. We speak of friction heat.

This model concept provides sound explanations for a whole number of open questions (fig. 8.4), i.e. why the temperature occurs independent of the state (1) and even in solids, where a purely kinetic interpretation fails (2). Every single elementary particle after all is carrier of a temperature (3).

With increasing temperature most materials expand, because the need for room, purely geometrically seen, increases for larger amplitude of oscillation (4). This principle is used in the case of a bi-metal thermometer.

In the case of solids the thermal oscillation of size is passed on primarily by the electrons in the atomic hull (5). Good electric conductors therefore at the same time also have a high thermal conductivity. An example of an application is the electric resistance thermometer.

In the case of gases the entire atoms carry out this task, for which reason a kinetic theory becomes applicable as an auxiliary description (6).

For extreme amplitudes of oscillation the atoms partly or entirely lose their enveloping electrons, when they change into the plasma state (7).

Finally the model concept even limits the second law of thermodynamics, which contains the postulate that it is impossible to design a cyclic working machine, which does nothing else, as to withdraw heat from a heat container and to convert it into mechanical work (8).

8.4 Heat energy

The discussed oscillation of contraction shows two characteristic properties, which must be looked at separately: the amplitude and the frequency.

Temperature describes solely the amplitude of the oscillation of size.

The heat energy however is determined by both,
by the amplitude as well as by the frequency.

Consequently the ideas of temperature and heat energy should be kept strictly apart. It therefore isn't allowed to set this oscillation equal to the electromagnetic wave in tables of frequency.

To be correct two tables should be given, one for the wave, characterized by a propagation with the speed of light, and another one for oscillations of contraction, thus for stationary phenomena and phenomena bound to matter. The latter indeed can likewise propagate relatively fast by fluctuations of pressure in the case of acoustical sound frequencies or by free movable charge carriers in the case of heat conduction, but the velocity of propagation for sound or heat is as is well-known still considerably smaller than the speed of light. Thus an assignment without doubts can be made as to which kind of oscillation it concerns.

***Temperature** describes the amplitude of the oscillation of size of all spherical vortices.*

*The **heat energy** however is determined by both, the amplitude and the frequency.*

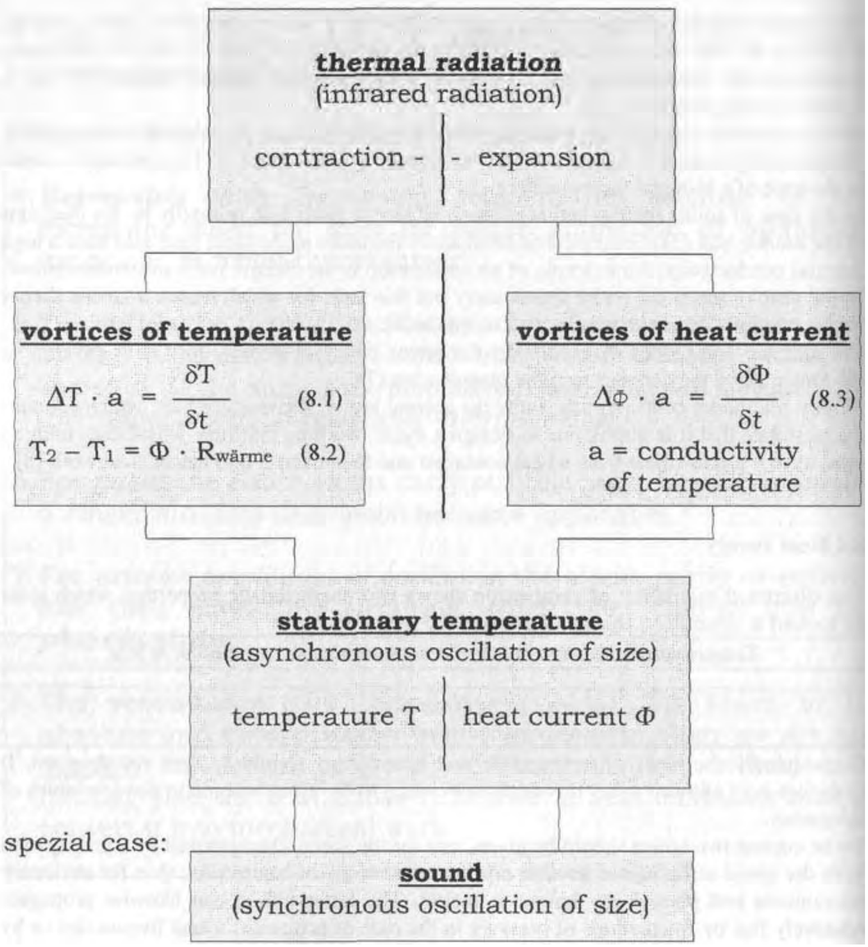
thermal radiation
(infrared radiation)
contraction -
- expansion
vortices of temperature
$\Delta T \cdot a = \frac{\delta T}{\delta t}$ (8.1)
$T_2 - T_1 = \Phi \cdot R_{\text{wärme}}$ (8.2)
vortices of heat current
$\Delta \Phi \cdot a = \frac{\delta \Phi}{\delta t}$ (8.3)
a = conductivity of temperature
stationary temperature
(asynchronous oscillation of size)
temperature T
heat current Φ
spezial case:
sound
(synchronous oscillation of size)

Fig. 8.5: Heat conduction resp. sound

8.5 Sound

The close relationship of longitudinal sound waves with the oscillations of contraction of thermally heated matter becomes particularly clear for ultrasound, where the arising heat in the inside of the body which is exposed to sound can be measured directly. The fundamental difference consists of the fact that the produced sound waves not only have the same frequency, but also the same phase, what needs not be the case for the temperature. The apparently uncoordinated occurring oscillations of size of the temperature, which as a rule occupy more space if the intensity increases, form a "thermal noise".

The oscillation of size with the same phase is not realizable at all in a spatial formation of particles, with one exception, the case that all particles expand and afterwards again contract simultaneously and in the same time. We can observe such a synchronization of the pulsation oscillations of all elementary vortices in the case of a pulsar. For us a pulsar looks like a "lighthouse" in space which shines with a fixed frequency. In reality it as well can concern a constantly shining sun, which carries out a synchronized, thermal oscillation of size, like a gigantic low-frequency loudspeaker. During the phase of contraction of the star its emitted light stays back. To us the pulsar looks dark. In addition the field strength is extremely increased and the light becomes correspondingly slow. During the phase of expansion the conditions are reversed and we observe a light flash. Exactly the pulsar unambiguously confirms the here presented theory of the variable, field dependent speed of light.

The well-known fact that the microcosm represents a copy of the macrocosm, already suggests that each atom is capable of the same oscillation of size as a pulsar: if next to the oscillating atom a resting one is placed, then does this one see a smaller field during the phase of contraction because of the increasing distance. It hence becomes bigger itself. If the pulsating neighbouring atom afterwards expands, it however becomes smaller. The at first resting atom in this way becomes a "pulsar" oscillating with opposite phase. The oscillating atom has stimulated the neighbouring atom as well to an oscillation of size, and this process will be repeated with the closest neighbouring atom. We speak of heat conduction.

To which extent the average distance between neighbouring atoms is influenced while a material is heated, solely depends on the structure of the atomic lattice. For matter with a fixed lattice according to expectation a smaller heat expansion will occur, as for the unordered structure of gases, in which we find confirmed well-known relations.

In a for potential vortices characteristic property sound waves and thermal waves of contraction correspond:

The propagation of potential vortex fields takes place as a longitudinal wave.

In this point vortex fields clearly differ from the transverse propagating electromagnetic waves!

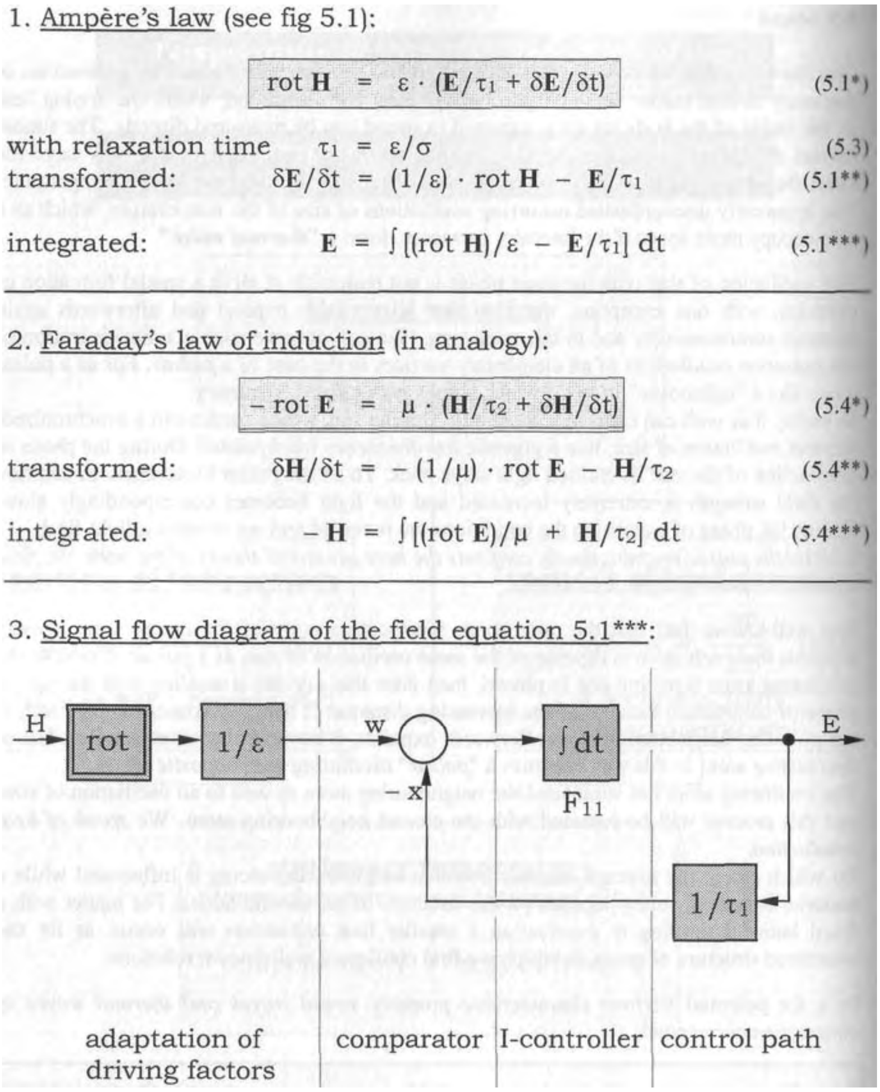

1. Ampère's law (see fig 5.1):

$$\boxed{\text{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau_1 + \delta \mathbf{E}/\delta t)} \tag{5.1*}$$

with relaxation time $\tau_1 = \varepsilon/\sigma$ (5.3)

transformed: $\delta \mathbf{E}/\delta t = (1/\varepsilon) \cdot \text{rot}\,\mathbf{H} - \mathbf{E}/\tau_1$ (5.1**)

integrated: $\mathbf{E} = \int [(\text{rot}\,\mathbf{H})/\varepsilon - \mathbf{E}/\tau_1]\,dt$ (5.1***)

2. Faraday's law of induction (in analogy):

$$\boxed{-\text{rot}\,\mathbf{E} = \mu \cdot (\mathbf{H}/\tau_2 + \delta \mathbf{H}/\delta t)} \tag{5.4*}$$

transformed: $\delta \mathbf{H}/\delta t = -(1/\mu) \cdot \text{rot}\,\mathbf{E} - \mathbf{H}/\tau_2$ (5.4**)

integrated: $-\mathbf{H} = \int [(\text{rot}\,\mathbf{E})/\mu + \mathbf{H}/\tau_2]\,dt$ (5.4***)

3. Signal flow diagram of the field equation 5.1***:

H
rot
1/ε
w
∫dt
E
x
F11
1/τ1
adaptation of driving factors
comparator
I-controller
control path

Fig. 8.6: Control technical analysis of the dual equations of the hydromagnetic field.

8.6 Basic principle of cybernetics

Surely can be attributed also information to the potential vortex. But how should information be formed? Is information a form of energy? Energy occurs as a consequence of the formation of potential vortices. Without this phenomenon there wouldn't be any energy!

Can information be described by means of a mathematical equation?

To be able to answer these questions, we subject the fundamental field equation to a control technical analysis. If it actually concerns a world equation, then an answers should be possible.

We again take up Ampere's law 5.1* from fig. 5.1 and remodel it according to the time derivative (5.1**). If the equation now is integrated over the time (5.1***), a signal flow diagram can be drawn (fig. 8.6).

The structure of a regulatory circuit is clearly visible. The individual paragraphs are described in an analogous way as for a technical control system. The execution of the curl operation on the field pointer of the magnetic field strength H and the multiplication with $1/\varepsilon$ accordingly form an adaptation of driving factors. In the comparator the difference for control from driving factor w and controlling factor x is formed and supplied to an integral controller. The control path has a purely proportional behaviour and consists of the processing of the measurement value of the electric field strength E with $1/\tau_1$, in which τ_1 describes the relaxation time of the eddy currents.

In technical control systems such a structure is found remarkably seldom, although it has an invaluable advantage: it possesses a stability in principle. Not a single adjustment of the controller exists, in which the closed regulatory circuit could become unstable, because it shows a proportionally delaying behaviour of first order. Possible changes of the adjustment of the controller or of the control path merely take effect on the speed, with which the regulatory circuit is able to follow changes of the driving factor.

This control technical basic principle convinces by its simplicity and efficiency. It meets us again in identical form in the second field equation 5.4*, the extended Faraday's law of induction. In dual formulation the electric field strength now appears as input factor and the magnetic field strength as output factor. Both regulatory circuits are coupled and connected with each other, by deriving their driving factor each time from the controlling factor of their dual partner. Is this structure actually efficient and meaningful?

Every regulatory circuit needs a target value, which is dictated from the outside. Let us think of the numerous control systems in nature. At all events a higher intelligence would be necessary for all the target values. This problematic is comparable to the question, what existed first: the egg from which a hen hatches or the hen without which no eggs can exist. Without a given target, evolution would not exist.

The connected regulatory circuit structure provides the matching answer: cybernetic systems, which usually and as is well-known strive to a state of balance, get their target value from their dual "partner". It is crucial that correspondingly dual systems are self-sufficient and can form and develop independently out of themselves without target values of a third side. This basic principle of cybernetics undoubtedly is brilliant.

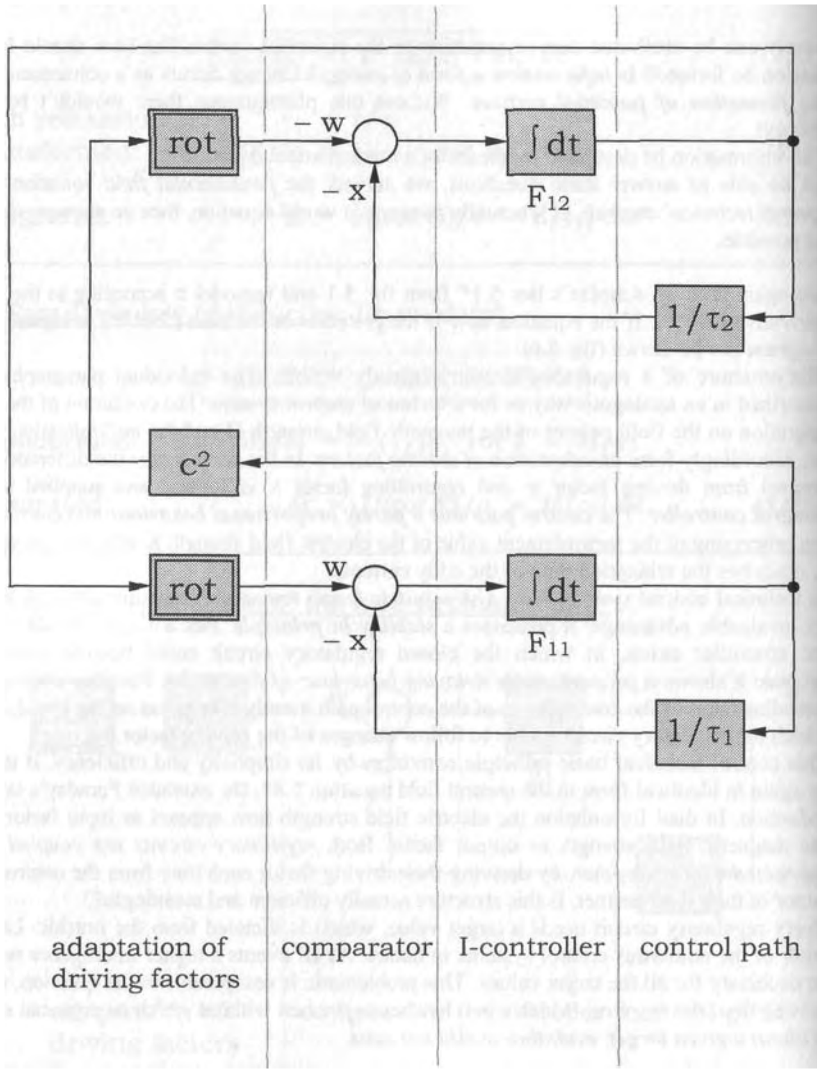
rot
− w
− x
∫dt
F12
1/τ2
c²
rot
w
− x
∫dt
F11
1/τ1
adaptation of driving factors
comparator
I-controller
control path

Fig. 8.7: Signal flow diagram of the fundamental field equation
with adaptive structure.

8.7 Adaptive regulatory circuit structure

If out of the nowhere something like the cosmos or like life on earth should form, then the connected regulatory circuit structure basing on duality probably is the only possible and conceivable. Thus it merely concerns the control technical representation of the fundamental field equation.

The question for the efficiency not only concerns the stability, but equally the possibility of both systems, to oscillate and to communicate with each other by the coupling and the associated exchange of information.

Fig. 8.7 shows the signal flow diagram of both regulatory circuits. These are switched in line and form a coupled circuit, which itself can be interpreted as a third regulatory circuit. Also this one shows a change of sign in the circuit like the other two circuits.

The information technical interpretation could turn out as follows: information about a regulatory process in the lower regulatory circuit F_{11} caused for instance by a disturbance is communicated over the coupled circuit to the upper regulatory circuit F_{J2}. In this case F_{11} acts as transmitter and F_{12} as receiver of the information. Afterwards both exchange their places, because F_{12} for its part reacts by a regulatory process and reports to F_{11}. The regulatory circuits adapt to each other. Obviously it concerns the basic structure of an adaptive regulatory circuit.

To analyse the coupled circuit the examination of individual special cases is recommended. If the regulatory circuits F_{11} and F_{12} are opened up in the way that the time constants tau_l and tau_2 go towards infinity, then the double integral effect is left. Analyses of technical regulatory circuit teach us that such systems always tend to instability. Because in addition the target value is zero, an oscillation around zero will arise, which we call electromagnetic wave.
If one of both time constants becomes finite, e.g. τ_2, then damping of the waves will occur. The "subordinate" cascade regulatory circuit F_{12} will adjust itself and now has a proportional delaying behaviour of first order. Together with the integral controller of the open F_{11}- circuit the coupled circuit will show the typical and more or less optimal regulatory behaviour of a damped oscillation.

These special cases correspond with the mathematical (fig. 5.2) and the physical (fig. 5.3) interpretation of the fundamental field equation. In addition a spatial rotation, a swirling will occur because of the double execution of the curl operation.

If interpreted control technically then vortices are the temporally stable, spatial swing of a field pointer around a centre, the vortex centre.

Without potential vortices no stability, no matter, no energy nor information would exist!

As can be looked up in Goethe's Faust, it always has been a desire of humanity, to find out, "what keeps the world together in the heart of hearts".

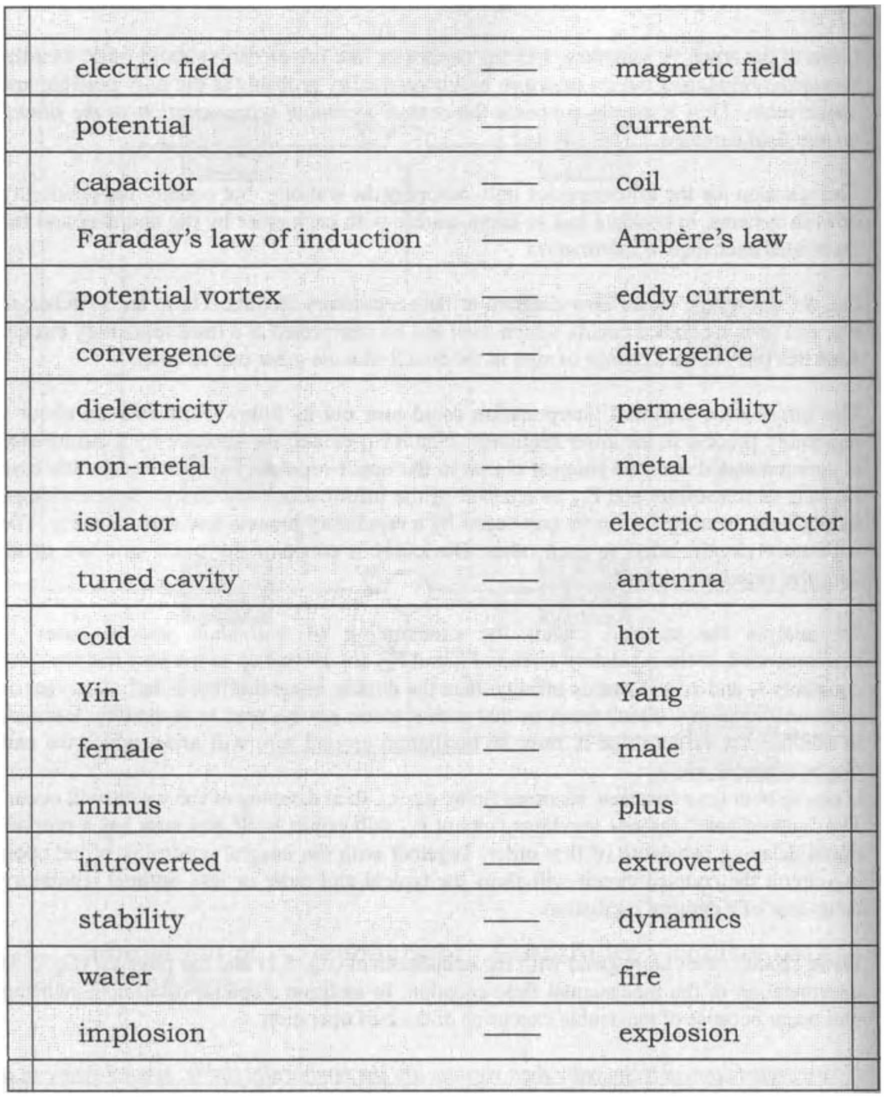

electric field	——	magnetic field
potential	——	current
capacitor	——	coil
Faraday's law of induction	——	Ampère's law
potential vortex	——	eddy current
convergence	——	divergence
dielectricity	——	permeability
non-metal	——	metal
isolator	——	electric conductor
tuned cavity	——	antenna
cold	——	hot
Yin	——	Yang
female	——	male
minus	——	plus
introverted	——	extroverted
stability	——	dynamics
water	——	fire
implosion	——	explosion

Fig. 8.8: Table of dual correspondences
(as a supplement to chapter 3.3 and fig. 4.2).

8.8 Information

The search for an answer for numerous philosophers and physicists was tantamount to the search for a world formula. Of course mustn't be forgotten that a formula only is a mathematical description and never the physical reality itself. It is a mathematical tool in the hand of a person and not the world or the cosmos itself, which he tries to understand. What keeps the world together in the heart of hearts, has to be more than only a pure apparatus of formulas. Actually the fundamental field equation tells us more. It reveals us a basic principle basing on duality in which the dual partners mutually dictate target values and goals. This principle convinces by its simplicity and efficiency. Apart from the "self regulation" it obviously also has the fundamental possibility of a "self organization" and the "generation of information". The field equations of the hydromagnetic field thus are the starting-point for the formation not only of matter and energy, but also of information. Accordingly holds:

Information is nothing but a structure of electromagnetic vortex fields!

This statement is new and to a large extent incompatible with the conception world of Norbert Wiener, who goes as the founder of cybernetics. From N. Wiener stems the sentence: "information is information, not matter and not energy".

We hold against it that obviously a fairly direct connection exists. We have worked out that only the vortex can show a stable adaptive regulatory circuit structure. Only the vortex and not the wave exists in two forms of formation dual to each other, and the principle of duality again is the prerequisite for the formation of information, of self organization and finally for the evolution. In fig. 8.8 well-known dual partnerships are listed. From it follows in a consistent way that for the production of information without exception the electromagnetic vortices should be considered.

But how can this so important duality occur, how can it form? This question is closely associated with the question of the formation of vortices. The signal flow diagram (fig. 8.7) to that says that the dual regulatory circuits F_1 and F_2 can only exist by the coupled circuit, which provides them the necessary target values and at the same time forwards the respective information. In this way of the oscillations and the more or less damped wave F_1 and F_2 communicate with each other.

The electromagnetic wave serves solely the mediation of information and energy.

With that falls a central role upon the wave, so that vice versa is valid:

Without wave no vortices, no duality and consequently no evolution can exist.

According to the to date state of knowledge the basic principle of cybernetics forms the basis for matter and energy as well as for information. Since the wave can only serve the transmission of information, the principle of duality and the vortex will function as carriers of information. We are entitled, to speak of vortex information, this by no means is characterized by special frequencies or modulations of frequencies. This is prevented by the property of the vortices which allows them to change the frequency. On the other hand various configurations of vortices are possible and numerous combinations and modulations are conceivable.

If technical apparatus generate vortices, then they produce information. Here a serious danger with regard to the environmental compatibility can not be excluded!

Equivalent things:

- * elements of the fundamental field equation 5.7
- * elements of the Greek philosophy of nature
- * temperaments
- * (impulses of growth)

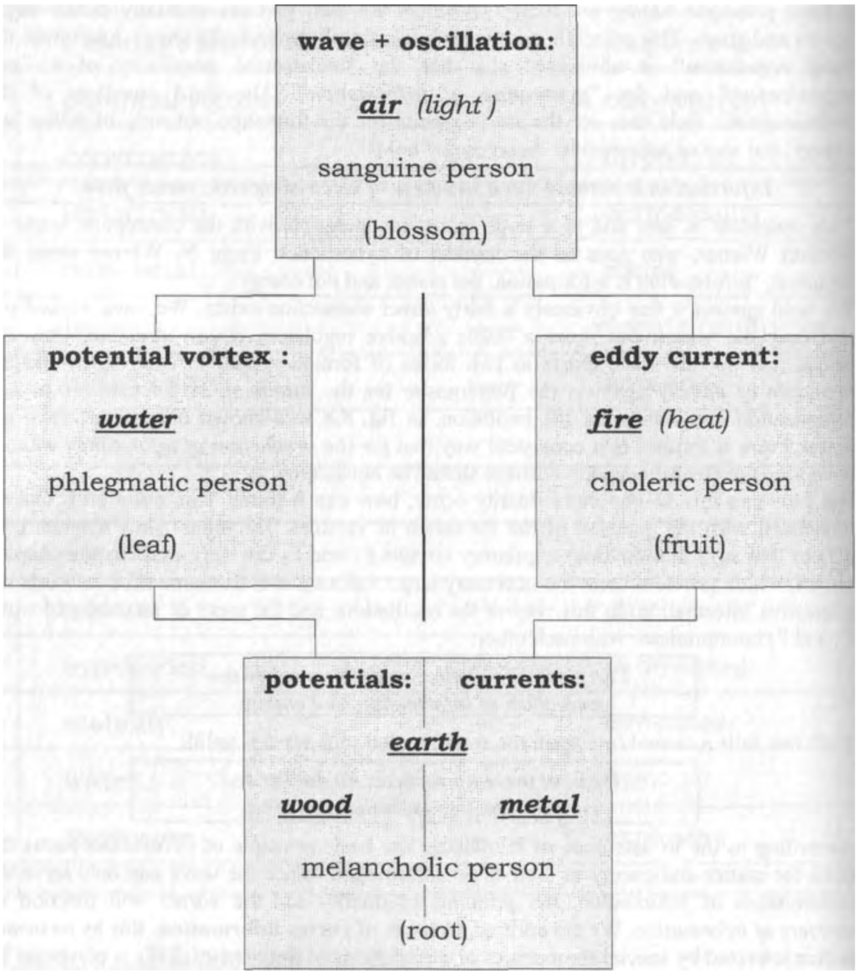
wave + oscillation:
air (light)
sanguine person
(blossom)
potential vortex :
water
phlegmatic person
(leaf)
eddy current:
fire (heat)
choleric person
(fruit)
potentials: currents:
earth
wood
metal
melancholic person
(root)

Fig. 8.9: The theory of four elements of the old Greek philosophy of nature (Aristotle and others)

8.9 Philosophy of nature

Seen in the view of the philosophy of nature now two dual points of view are possible.
The optimistic one would be:
We and our environment on the one hand are a result of the cybernetic principle and on the other hand of our observation point of view which should be valued relativistically. If really everything should be electromagnetism, a phenomenon which can't be grasped directly by humans, then the pessimist would come to the conclusion: everythins is nothing. What we observe is nothins but a deception of the senses. Perhaps therefore famous philosophers of antiquity, like Empedokles or Demokritos have ended their life in the crater of the Etna. According to the theory of the atom of Demokritos (470 to 380 B.C.) the formation of matter, earth and celestial bodies will occur by means of formation of vortices!
Empedokles (482 to 420 B.C.) was the first to develop a theory basing on four elements, which was continued and improved by Plato (428 to 348 B.C.) and Aristotle (384 to 322 B.C.). Accordingly these elements are changeable into each other and mixable with each other. From them all bodies are build up.
The terms "air, water, fire and earth", with which the philosophers have described the four elements, are of course not identical with the ones in our translation and conception world, but they were used in a philosophical sense as a substitute for the description of the respective basic principle.
There also have been different approaches, to translate these terms differently, e.g. by an assignment to the four states of matter (solid, liquid, gaseous, plasma). But the ancient texts don't get easier to read in that way.

Fig. 8.9 shows the obvious assignment to the four building parts of the fundamental field equation 5.7. It would be worth an attempt, to exchange the terms in the translations of ancient texts and to translate air with wave, water with potential vortex and fire with eddy current. The term earth has two sides, which should be translated with potential instead of wood and current instead of metal.

Let's try the translation this way with the theory of Plato<i>, by correspondingly translating anew the talk of Timaios about the formation of the world. The perception of smell then is described as follows: "...as the potential vortex turns into waves (or) the wave into potential vortices, the smells are formed during this transition, and smells are smoke or fog. But fog is the transition of waves into vortices, the transition of the vortex into waves however smoke".
Plato here provides an indisputable and conclusive interpretation of the fundamental field equation. In this equation the potential vortex acts as damping term in the wave equation, what in the case of waves rolling up to vortices will show to the observer in the way that the electromagnetic waves and therefore also the light will be damped. We say, the visibility gets worse and speak of fog. If the damping phenomenon disappears again, as the potential vortices break up, then Plato speaks of smoke.
Numerous ancient texts, which until now only could be "interpreted" philosophically, in this way turn out to be a rational textbook description of natural scientific phenomena. They anyway only get readable and understandable for the general public with the modern technical terms.

<i>: Platon: Samtliche Werke 5, Rowohlts Klassiker Nr. 47, S. 188, 66e

Optimist:

We and our environment on the one hand are a result of the cybenetic principle and on the other hand of our observation point of view which should be valued relativistically.

Pessimist:

Everything is nothing. What we observe is nothing but a deception of the senses.

Plato[<i>] (talk of Timaios about the formation of the world) description concerning the perception of smell:

"...as water (the potential vortex) turns into air (waves) or air (the wave) into water (potential vortices), the smells are formed during this transition, and smells are smoke or fog. But fog is the transition of air (waves) into water (vortices), the transition of water (vortex) into air (waves) however smoke"

changed translation with the technical terms:

air______ = ____wave
water = ___ potential vortex
fire______ = ____ eddy current

As a consequence:

smell is vortex information

Fig. 9.0: The explanation of Plato concerning the formation of smell, smoke and fog[<i>]

<i>: Platon: Samtliche Werke 5, Rohwohlts Klassiker Nr. 47, S. 188, 66e

9. Usage

If the newly discovered vortex phenomenon of the vortex of the electric field exists, then it will be possible to practically use it. Whereas we still think about possibilities for technical usage, there by all means exists the possibility, that nature already is successfully using the vortex for a long time. We should look precise at things. We can only learn of nature!

Remarkable about the passage of Plato (fig. 9.0) is not only the fact, that the potential vortex already was known for two and a half thousand years and was taken into consideration for an interpretation, but also the realization of Plato, that during the described transition the smells form. Smell thus would be a vortex property!

After all vortices are able to bind information as can be inferred from the basic principle of cybernetics. With this vortex property and the statement of Plato smell obviously would be nothing else than pure information which by potential vortices is stored, distributed and eventually is caught by the hair cells for smell of the nose.
If now a dog takes up a track, does it then run after vortices which remained behind or does it analyse, according to the encyclopaedia, the chemistry left behind, or does it combine both? Does the bloodhound for instance interpret the vortical oscillations of chemical substances like lattice oscillations or the movements of the electrons in the atomic hull? A lot of work awaits the research scientist of vortices here. The seminar will offer opportunity, to deepen this topic<i>.

Only if technicians exist, who put an electronic box on the table with a button, at which they adjust the wanted scent of rotten eggs to lavender, we honestly can claim to have understood the phenomenon of the smell.

For the majority of the people a theory only wins its sense by the practical applicability, and therefore we'll have to develop and present a technical usage on the basis of the vortex theory.

<i>: Suggestions for appropriate seminar themes: the meaning of smell, taste, the aroma therapy, the homeopathy or the effect of aetheric oils. (To that in part 3, INDEL Verlagsabt. 2003)

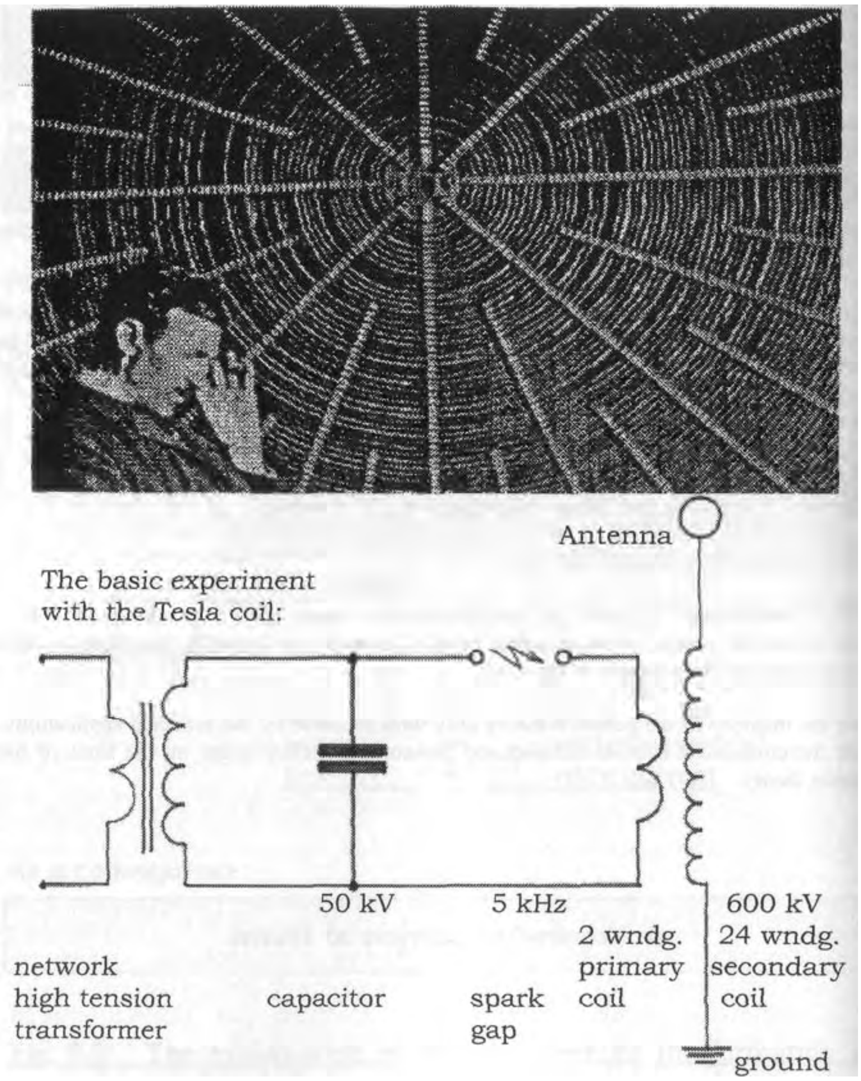
Antenna
The basic experiment
with the Tesla coil:
50 kV
5 kHz
600 kV
2 wndg.
primary
coil
24 wndg.
secondary
coil
network
high tension
transformer
capacitor
spark
gap
ground

Fig. 9.1: High tension transmitter of Nikola Tesla

9.1 Longitudinal electric waves

It is important, that experts struggle for problem solutions. Only if the discussion about limits doesn't come to a result which can be grasped and verified, then also in the environmental compatibility the willingness will grow, to accept a not yet known phenomenon as a further and possibly crucial influential factor.

Already for a hundred years there has been a dispute of experts. At that time scientists all over the world were trying to verify the experiments of Heinrich Hertz. Then from America the message came, Hertz had been mistaken and the electromagnetic wave would have completely other properties. The scientists in Europe were indignant, but they had to take the message seriously, because it after all came from the important experimental physicist Nikola Tesla (1856-1943), who with his inventions of the rotary field and of the asynchronous motor has stamped today's electric energy technology as no other.

As a result Lord Kelvin boarded a steamship as a mediator and sailed 1897 to New York to convince Tesla from the opposite. But the experiments, which Tesla presented his Lordship, didn't give rise to any doubts<i>, and thus Kelvin returned to Europe with the message: "Both are right, Tesla as well as Hertz! Whereas the electromagnetic wave which Hertz has detected, is a transverse wave, does Tesla work with a longitudinal wave"! Lord Kelvin as a result started to draw most different vortex models, because it was clear to him, that a propagation as a longitudinal standing wave analogous to the sound wave only is conceivable, if quantized structures exist, which knock each other mutually. Kelvin therefore assumed vortex structures of the electromagnetic field. His vortex models were published and landed in the curiosity box of orthodox science.

Heinrich Hertz did have a big advantage. He could refer to Maxwell and calculate his wave with the field equations. For modern technology the mathematical calculability is almost an indispensable prerequisite!

For Tesla wave there however neither was a mathematical nor a physical theory. The only thing Tesla had, were presentable experiments.

In Colorado Springs he had build a 10 kW transmitting installation and lighted 200 fluorescent lamps of 50 Watt each on a mountain in the Rocky Mountains in a distance of 25 miles. With that he had completely transmitted the transmission power of 10 kW, as can be inferred from the press reports at that time. With Hertzian waves, which propagate spatially, this experiment even today, after over 100 years, wouldn't be realizable technologically. According to the law of the square of the distance one isn't even able to let glow a tiny little lamp in such a distance.

For sure his rotary field theory was a big help for Tesla in all experiments. Actually a rotary field can be seen as the special case of a planar vortex.

Thus Tesla obviously was able, to use the potential vortex without even knowing it. Tesla has stimulated a loosely coupled high tension coil wound like a spiral to self-resonant oscillations and emitted the produced vortices over an antenna (fig. 9.1). On the receiver side the process was then reversed.

<i>: N. Tesla: III. The Singular Misconception of the Wireless (Famos Scientific Illusions), Electrical Experimenter, Feb. 1919, p. 732.

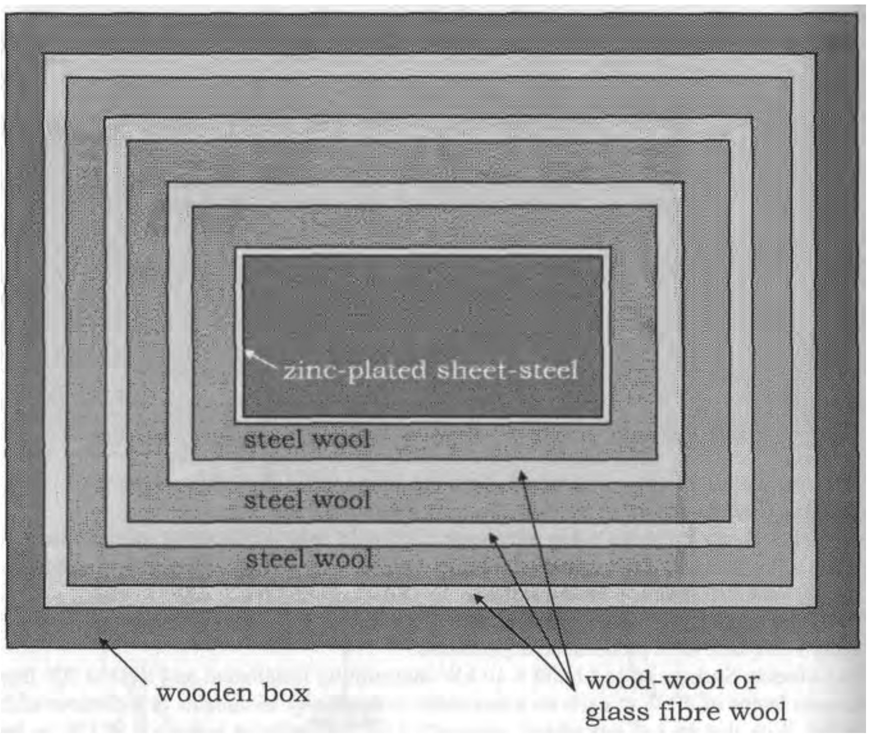
zinc-plated sheet-steel
steel wool
steel wool
steel wool
wooden box
wood-wool or
glass fibre wool

Fig. 9.2: Orgone accumulator according to Wilhelm Reich.[i]

<i>: Dr. James DeMeo: The Orgone Accumulator Handbook, ISBN 0-9261855-0-7

9.2 Medical technical usage

Nikola Tesla at his time was extremely popular in the area of medicine. With his inventions injuries were cured and pain eased. Modern diathermy goes back to his work. But Tesla at that time has treated the patients with vortices, whereas today, possibly out of ignorance, electromagnetic waves are used. These however have in contrast to potential vortices only a small penetration depth. Today in addition only the effect of heat is considered and in no way the information technical effect of electric vortices. Here we are missing the comparison, to be able to say. if the treatment with a Tesla apparatus was more successful than with modem microwave radiators.

The experiments of Wilhelm Reich (1897-1957) aimed in the same direction. Exactly as Tesla 55 years before also Reich emigrated to America in 1939. He had specialized in catching and collecting vortices, which he called "orgone", from the air. In this way he could store up high energies over time and then use them purposeful. With his "orgone accumulator" he was able to cure wounds faster and to ease pain. He also treated cancer and a lot of forms of anaemia. Technically Reich could demonstrate, that charged capacitor plates discharge faster under the influence of his apparatus. His orgone accumulator for instance is a wooden box, filled with wool of steel and wood, which alternate in layers. He said that the reason for this construction is, that metals repel the orgone energy, whereas organic materials become charged and accumulate it. That holds in his opinion to a particularly high degree for water.
Reich concludes that no material and no device are known, with which the orgone energy could be shielded. This with regard to the environmental compatibility of the vortex phenomenon should make us thoughtful.
As a supplement it should be mentioned, that Reich already at the end of the forties has pointed to the dying of the forests and he has made a direct connection to the orgone energy. Reich was slandered and mocked and died in the prison of Connecticut, while his writings were burned publicly.

Not much better fared the Austrian forester Viktor Schauberger (1885-1958), who also can be described as visionary. He was able to produce water with a particularly good conductivity. His water in addition goes for medicinal and healthy. Also to him travelled cancer patients of far away to get this water.
Schauberger spoke of the natural treatment of the water, whatever he meant with that. In any case he build with great success installations to transport wood by floating it and even installations to mine ore with so-called double twist pipes, which made possible a transport without friction even of materials, which are heavier than the "means of transport" water. It could be proven that no contact with the pipe took place. This was scientifically investigated and confirmed at the university of Stuttgart 1952 by Prof. Popel. Quite obviously in the double twist pipe vortices have formed, which have bound the material to be transported (see phenomenon of transport).
Today you can buy levitated water at over 100 selling points in Europe, of which is claimed, that it is prepared according to instructions of Schauberger. Unfortunately we aren't able to ask him anymore, if he agrees with such a "centrifuged" water. Thus the test of the effectiveness has to be left to everyone himself or herself.

A: Concerning the functioning of the Faraday generator $(E = v \times B)$:

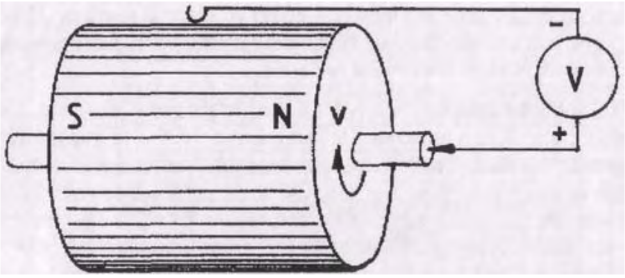

B: Construction of the flying disc:

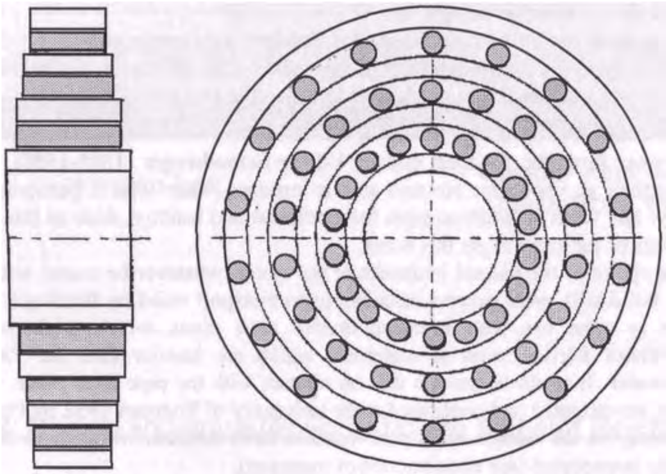

Fig. 9.3: Applications of the Faraday generator[i] in the flying disc of Prof. John Searl

9.3 Flying objects

The medium of Viktor Schauberger always was the water. He could produce light effects and very high temperatures with it with only minimal excitation in the form of impulses. An installation, which had been built according to his plans at the company of Siemens, eventually melted at 4000°C by improper operation, as he himself stated. According to that the machine would have released more energy than used, thus a perpetuum mobile. The authorities of the state in the Third Reich were impressed and put Schauberger in a concentration camp, where he was instructed to build a "flying saucer" under permanent supervision together with fellow prisoners. For him it's all about the concentration effect and the usage of the as a result occurring implosion in contrast to our today's " technology of explosion and fire" as Schauberger was accustomed to express himself. It is not known, if Schaubergers "repulsine" ever has took off.

A disc, which has proven its suitability to fly, was constructed and built by the english technician John R. R. Searl, although he said he can't explain the effect. He also had big difficulties, to get the apparatus under control. A disc broke through the ceiling and the roof of his laboratory and disappeared to never be seen again. Five other flying discs, which he after this experience started in the open, went lost in the same manner. Without knowing the effect, he of course neither could assess the dangers. His experiments have claimed serious injuries and a casualty. While he 1985 was put in prison under a pretext, his laboratory and his house were burned down and all documents destroyed. Now he works on a technical usage. According to the principle it concerns a Faraday machine. Thereby a permanent magnet, magnetized in axial direction, is turned. Now the magnetic induction B for a relative velocity v is measured as an electric field strength E according to equation 6.10: $E = v \times B$. (6.10)

Because the vectors are arranged as standing perpendicular to each other, will arise a tension voltage in radial direction (direction of E-field) which can be taken off. Apart from the friction no further force of reaction occurs in the case of the Faraday generator. Because of the small gain of energy until today no application ready for the market exists. But this principle, to convert magnetic field energy into electric, already has moved the nature of many inventors. Professor Searl has reversed this old principle. Whereas normally the component of the velocity towards the centre of rotation decreases, it increases in Searls case. For that he works with roller pivoted concentric rings, which he drives by a small electric motor (fig. 9.3). Doing so something inexplicable for him happens: After switching off the motor the revolutions per minute don't decrease again, but increase audibly and increase further, until the produced electric field shows the well-known high tension phenomena: corona discharges, formation of ozone, ionisation of the air and production of a vacuum in the inside of the disc.

The rings and rollers consist of several layers, which are built up similar to a bi-metal. The only explanation I can think of is that a change in structure would occur as a result of the physical length contraction which is caused by the increase of the E-field in the direction of the centre of rotation. The bi-metals try to withdraw themselves from this change by an increase of their rotation of their own (fig. 6.5). To compensate the field the disc builds up a spin, as also the elementary particles do (fig. 6.13). While the formation of vacuum prevents sparking in the inside of the flying disc, and the revolutions per minute further increase because there is no air friction whatsoever, the disc weighing 5 tons all of a sudden lakes off the ground and according to reports of eyewitnesses shoots vertically upwards.

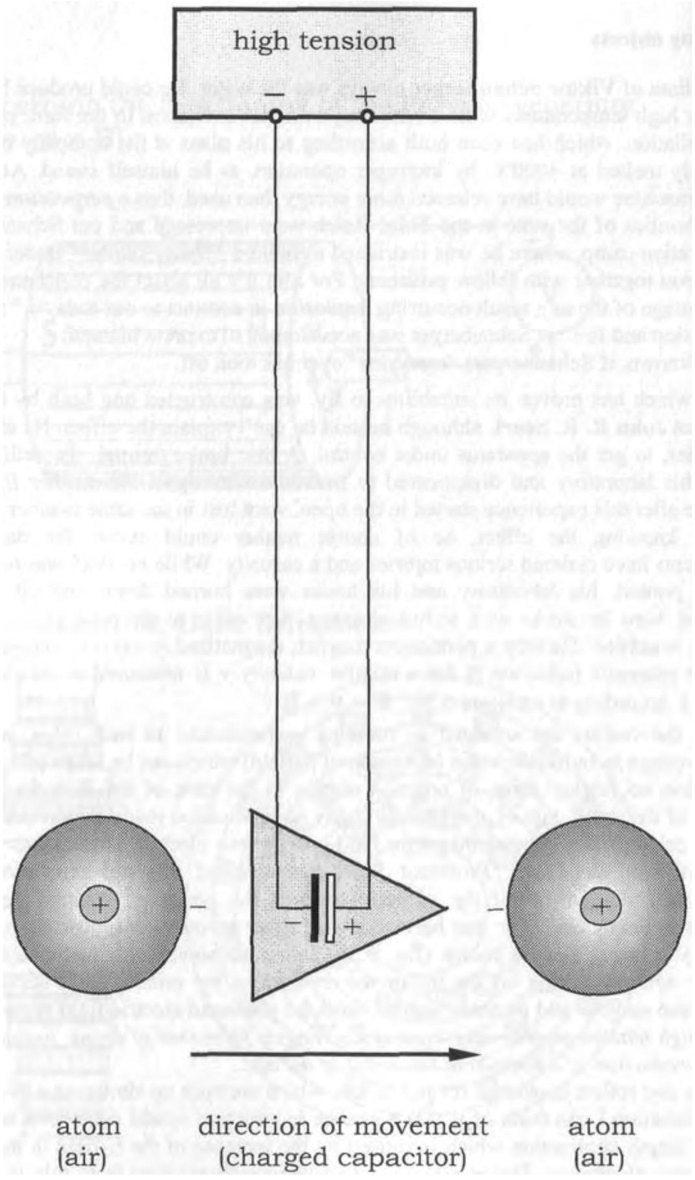
high tension
−
+
+
−
−
+
+
−
atom
(air)
direction of movement
(charged capacitor)
atom
(air)

Fig. 9.4: Capacitor experiment concerning the so-called electro-
gravitation according to Prof. Biefeld / Dr. Brown

9.4 Electro gravitation?

Wild speculations circulate about the mechanism of the flight of a "flying saucer", which should function without sound and without combustion, and for which no sound barrier exists, as pilots have observed. The talk is about cancelling gravitation or about an "electro gravitation".

Reproducible in any case is an experiment, which has been systematically investigated by the american professor Biefeld and his pupil T. Brown from 1925. Accordingly does a body charged to high tension show a force effect in the direction of its positively charged note. As a check for oneself a capacitor can be suspended on its connection wires and it can be observed, how it moves visibly in the direction of the positive pole in case it is charged (fig. 9.4).

Because the Biefeld-Brown effect depends on direction, it actually can't concern gravitation at all but merely a not understood electromagnetic interaction. In the neighbourhood of the capacitor pole air molecules are found, and their negatively charged atomic hull is attracted by the positive pole. At the same time the atomic nucleus is repelled. By that the air atoms in the presence of the high tension capacitor become unsymmetrical, causing the force of attraction due to the smaller distance to exceed the force of repulsion. At the negatively charged end the conditions are exactly reversed. There a repulsion occurs. Because the polarized air atoms and molecules are sucked in, no pressure can build up and as a result no sound barrier can occur. Experiments with charged and uncharged rockets have brought the interesting result, that the electrically charged rockets flew many times as far as the uncharged ones (5-6 times as high)<i>.

Many a person now perhaps starts to dream of the flying carpet but, as said, it isn't an "effect of antigravitation". Does "free energy" actually exist, we have to ask ourselves? From the gravitation the soviet physicist Landau calculates an energy density of 16 megawatt hours per cubic meter for the earth's surface. Immediately inventors are found, who want to use this gravitational energy.

Nikola Tesla in his laboratory in Manhattan has incidentally built resonators, of which he could bring all electric, magnetic and mechanic factors in resonance. On an evening stroll he fastened a battery operated vibrator to the tubular steel scaffolding of a new building and let everything shake and wobble. In his laboratory such a device once got out of control by inattentiveness and triggered an earthquake. In that way the road surfacing and pipes were burst and window panes got broken. The police penetrating his laboratory only could see, how Tesla forcible finished the experiment with a sledge-hammer.

The experiments which got out of control of Tesla, Searl and Schauberger have one thing in common: it concerns constructions with an unipolar arrangement of the field.

Tesla had arranged the magnetic field in a unipolar way, as he has reported himself, Searl had realized electric unipolar fields in a construction similar to the electron, and Viktor Schauberger had specialized in producing unipolar structures with water vortices.

In the case of the corresponding technical usage, which can be interpreted as making use of the occurring "spin coupling", therefore in principle utmost caution is imperative.

<i>: Rho Sigma (Dr. Rolf Schaffranke): Forschung in Fesseln, das Ratsel der Elektro-Gravitation, Ventla-Verlag, Wiesbaden, 1972, Page 67.

Nikola Tesla: „Ere many generations pass our machinery will be driven by power obtainable at any point in the Universe This idea is not new, we find it in the marvellous myths of Antheus, who derives energy from the earth... Everywhere in the universe there is energy. Is this energy static or kinetic? If static, our hopes are in vain; if kinetic, and this we know it is for certain, then it is a more question of time when men will succeed in attaching their machinery to the very wheel work of nature". New York, the 20th may 1891, American Institute of Electrical Engineers.

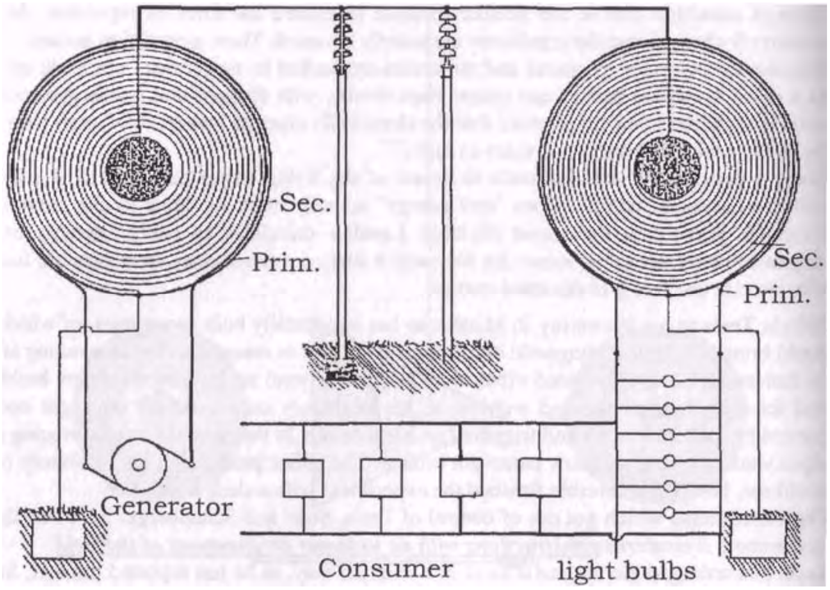

Fig. 9.5: Energy transmission bound to a wire, Patent No. 593,138 (1897) of Tesla.

<i>: Philadelphia Public Ledger, Nov. 2, 1933

<ii>: Dr. Nikola Tesla: Complete Patents, Tesla Book Company, New Edition 1983, Compiled by J. T. Ratzlaff, ISBN 0-960356-8-2, page 301-304.

9.5 Free energy?

Furthermore is reported of Tesla, he would have developed a "converter for space energy" and 1931 have built it in a luxury car (Pierce Arrow)<i>. The car was a 145 km/h fast, doing so the asynchronous motor (presumably built by Westinghouse) furnished 80 HP at 1800 Rpm. The "free energy" the converter, build by Tesla himself, got from a 1.8 m long antenna. Because the motor ran without adaptation of frequency in slipping operation, it had to be cooled correspondingly. It was on the way in trial run for over a week. Now we of course want to know, of what this "free energy" consists, which Tesla will have used and of which he already 1891 had spoken in the American Institute of Electrical Engineers.

For the electron as a spherical vortex we have calculated an electric tension voltage of 511 kV between its surface and its centre respectively infinity (equation 6.31*, fig. 7.1). The highest level of tension voltage normally used for the high tension transmission lies at 380 kV (effective value); for a direct current transmission it is 500 kV. Although still higher tension voltages would be desirable, they are avoided. This is no accident. Experiments with higher tension voltages namely have resulted in inexplicable high losses. We have an explanation: the electrons are taken apart on the way! Their inner energy amounts with the outer energy to zero. The charge carrier, which in the power station as result of an energy conversion has been sent on a journey, is in danger to vanish into thin air (e.g. corona) for tension voltages above 511 kV. The transmitter of Nikola Tesla however (fig. 9.1 and 9.5) worked with 600 kV and more. He said, with his experiments he had destroyed billions of particles without being able to observe an emission of energy and made fun of the misinterpretation of Einstein of the already at that time well-known mass-energy relation $E = mc^2$.

For the purpose of a one wire or a wireless energy transmission the tension voltage therefore has to be higher than 511 kV. With Teslas equipment however the electrons shouldn't be destroyed, but merely be pulled apart to plane circular vortices with help of the today is Tesla coil known winding. These then could be sent on a journey over the antenna, to again be caught and formed back in spherical electrons by the receiver antenna. As long as the electrons don't fall apart, they keep their structure and quantum property. Quanta pass on an excitation, for instance a bump, in form of a longitudinal standing wave, by one particle bumping the neighbouring particle, analogous to sound waves, where one gas molecule passes the bump on to the neighbour. The transmission hence takes place as longitudinal scalar wave.

Electrons pulled apart to planar vortices in addition haven't got a closed vortex centre anymore on their journey. For the by Tesla in his laboratory used and publicly presented one wire energy transmission (1897, Patent Nr. 593,138) the transmission hence doesn't take place in the wire, but as vortices around the -wire. That explains, why only a relatively thin conductor, which normally should have melted, was necessary for a demonstrated power transmission of 10 kW. Tesla however could show, that the wire stayed cold and virtually no heating losses were measurable<ii> (fig. 9.5). He himself said, that this one wire transmission technology is much better than the alternating current technology full of losses, which stems from him as well. Tesla must have known the limit of 511 kV very exactly, because on several photographs one can see, that he changed the coiling technique off this value!

<i>: Philadelphia Public Ledger, Nov. 2, 1933

<ii>: resp. Tesla's lost inventions, VAP, page 36 and page 48, 49.

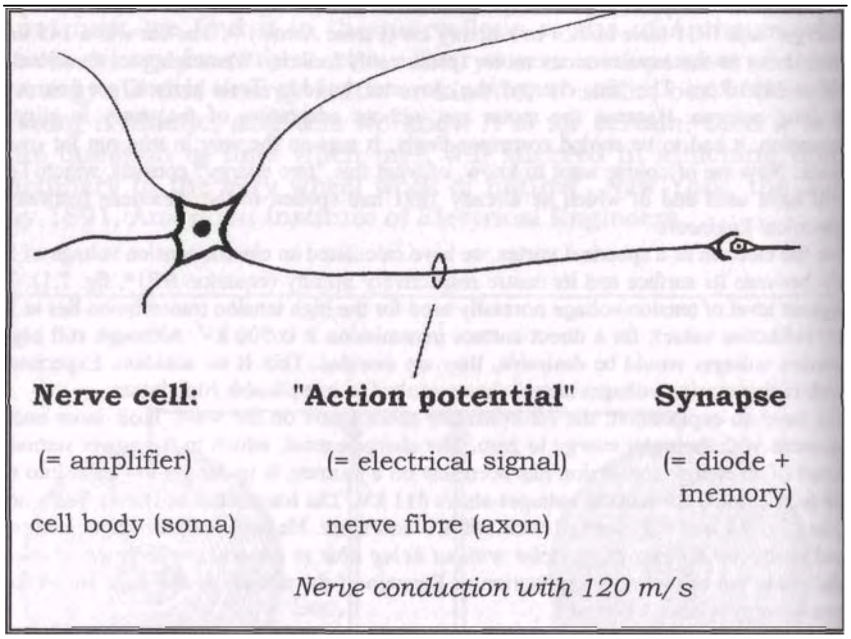
Nerve cell:
"Action potential"
Synapse
(= amplifier)
(= electrical signal)
(= diode + memory)
cell body (soma)
nerve fibre (axon)
Nerve conduction with 120 m/s

Fig. 9.6: Nerve cell and message processing (neurons)

- Material information conduction: at plants
- Information conduction with hormones: at simple animals
- Electrical information conduction: at higher developed creatures
- Analog transmission: from 10% and more difference in potential, and
- Digital transmission: number of the action potentials
- Modulated information: i.e. complex modulation of the action potential (the potential vortices)

9.6 Nerve conduction and action potential

Technical solutions often are only a suboptimum and many times they are anything but an optimum. In many cases the feasibility stands in the foreground. In the question of the current distribution at that time the decision was made for the worse system, the full of losses three phase system, because no current meters were available for the free of losses one wire technology. Without being able to collect money from the consumer, it wouldn't have been possible for the energy distribution installations to pay for themselves. Compared to that is nature with its selection principle merciless and without compromises. Only the most efficient system has a chance to survive and to spread. If it's about transmittining information in an electric way and two different principles present themselves, then in nature only the better one will be brought into action.

Let's take a look at the nerve conduction in the human body. In the synapses ion concentrations and electric potentials of 70 to 90 mV arise. Here with conventional gauges the activity of a nerve can be detected. From a measurement of the transmission time of synapse to synapse the velocity of the signal is determined. If we however want to measure between the switch points on the line, then we have to find out, that for instance with an ammeter actually nothing is measurable. In addition the nerve fibre has a miserable conductivity.

The biologist calls the electric signal just "action potential" and draws a vortex-like ring around the nerve fibre and speaks of a mixed digital and analogue information transmission.

The doctor on the other hand knows two different types, fast and slow nerves. In the inside both are built up virtually identical. A characteristic difference consists of the fact, that the fast nerves are jacketed with a thick fat layer.

The technician would say, they are better isolated, but why they therefore should be faster, he hardly would be able to answer. If we however assume, that the action potentials or OUT potential vortices oscillate around the conductor, thus exactly in the isolation layer and are forwarded there, then possibly an explanation would have been found.

The nerve conduction moreover has much in common with the one wire system of Tesla:

1. Charges and electric potentials are transmitted
2. Doing so a transport of charge carriers, thus an electric current on the line, isn 't detectable.
3. Consequently no losses worth mentioning occur.
4. Decisive for the functioning is the dielectric insulator material which surrounds the conductor and not the electric resistance of the conductor.
5. In contrast to electric circuits a draining conductor is unnecessary.

It looks as if the one wire system is not new at all, as would nature use it already for ages in the highest perfection. We realize immediately, why our head does without a ventilator and how the high of signal processing density can be explained. Compared with our brain modern personal computer (PC) with their backward cable technology are far less than a suboptimum.

With regard to the pressure on the environment by interference radiation we had derived, that currents and eddy currents with the skin effect direct their radiation to the outside, whereas potential vortices with their concentration effect direct it to the inside. Probably for this reason the radiation field of a person is environmentally better compatible than that of a PC.

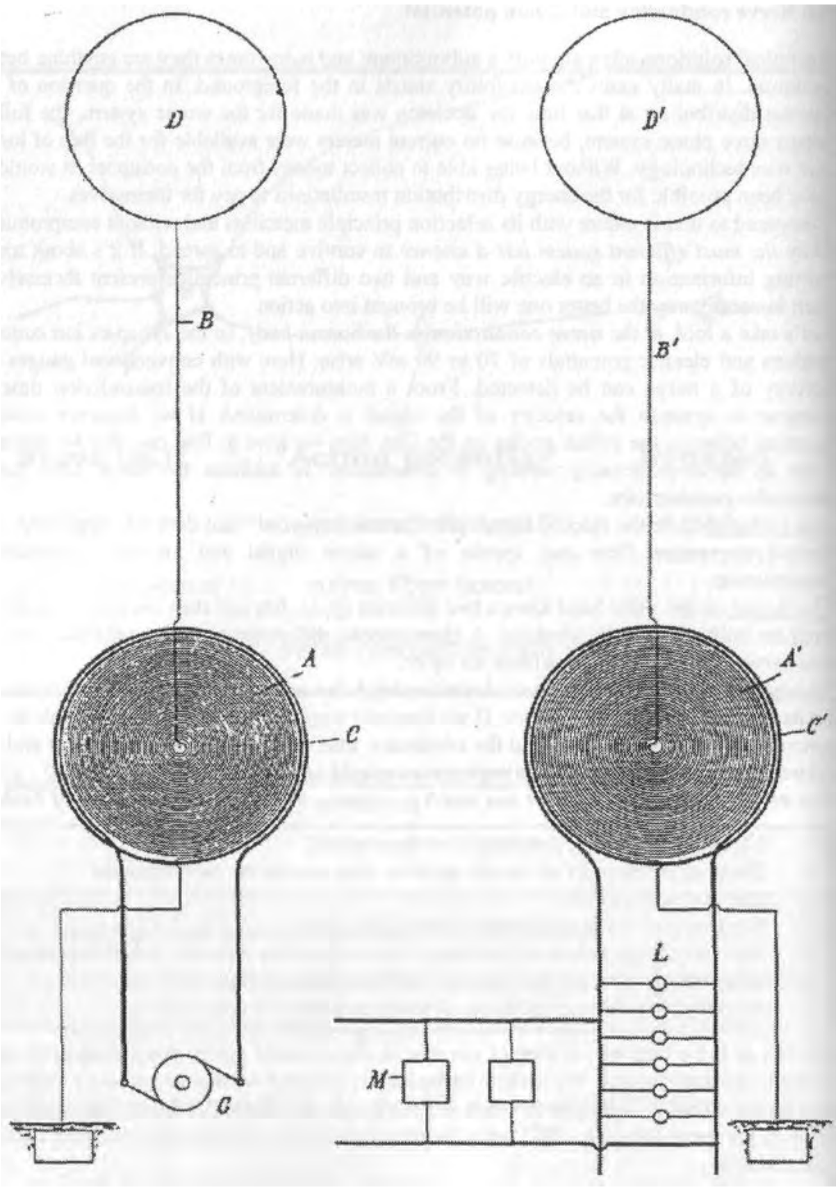

Fig. 9.7: Wireless energy transmission, Patent No. 645,576 (1900) of Testa.<i>

<i>: Dr. Nikola Tesla: Complete Patents, Tesla Book Company, page 311-321.

9.7 Wireless energy transmission technology

Tesla still went a step further. He has cut the wire connection between transmitter and receiver (fig. 9.5) and instead has installed each time a spherical electrode (fig. 9.7). With this facility he now was able, to transmit energy completely wireless (1900, Patent No. 645,576). Building on this principle Tesla already 1890 had built a radio remote control for a battery operated submarine and had presented his patented system the navy. But they weren't able, to recognize the strategic importance of the radio technology and rejected with the words: "It's too advanced!" (fig. 9.8).

Tesla further had proven with this technology the mentioned proof concerning the existence of longitudinal electromagnetic waves. There exist descriptions, partly of Tesla himself, how he, inside or also outside his laboratory, goes in search of oscillation nodes of the standing wave with a measuring receiver<i>.
He names several conditions for the wireless energy transmission:

1. "perfect resonance" (same frequency) of transmitter and receiver
2. Installation of the receiver on an "oscillation node" (maximum).

Important is also the measurement of the velocity of propagation, and that isn't constant from the start for a wave, which oscillates in the direction of propagation. From the research of earthquakes we know, that the longitudinal waves are faster than the as well occurring transverse waves. Usually the distance to the epicentre is determined from the difference in transmission time.
In the patent specification the measurement is described<i>. Tesla has sent longitudinal radio signals from his transmitter in Colorado Springs once around the earth along the earth's surface and determined a transmission time of 0,08484 seconds. This corresponds to a frequency of approx. 6 Hz. He saw his result confirmed, as he could find the oscillation node of the standing wave again (according to $\lambda/2$)in his laboratory (on the ground plate).
For the full wave length λ the Schumann resonance, which describes a standing wave of an around the earth running Hertzian wave, lies as is well-known at 7.8 Hz! Tesla calculates for his wave a speed 1.6 times the speed of light assuming the polar radius of the earth to be 6363 km.
Also this measurement result confirms, that Tesla didn't use the Hertzian wave. Tesla found the off-beat concept of Einstein, the speed of light would be the fastest possible velocity of signal transmission, only funny. If however today is claimed, 1993 in tunnelling experiments for the first time a speed faster than light has been measured, then this just isn't true. Possibly also the Tesla wave tunnelled, as the vortices for a small conductivity of the air contract and as a result of the length contraction become correspondingly fast.
Later Tesla after several goes in vain even succeeded in building a high tension tube as a tunnel, with which the velocity of the signal could be increased arbitrarily. Tesla with that pursued the goal to be able to make radio contact with other worlds<ii>.

<i>: N.Tesla: Art of Transmitting Electrical Energy Through the Natural Mediums, US-Patent No. 787,412, 18.4.1905, Complete Patents pp. 397-402.
<ii>: M. Cheney: Tesla, Man out of Time, Barnes & Noble Books, New York, 1993, und S. 309, Omega-Verlag, 1996, ISBN 3-930243-01-6

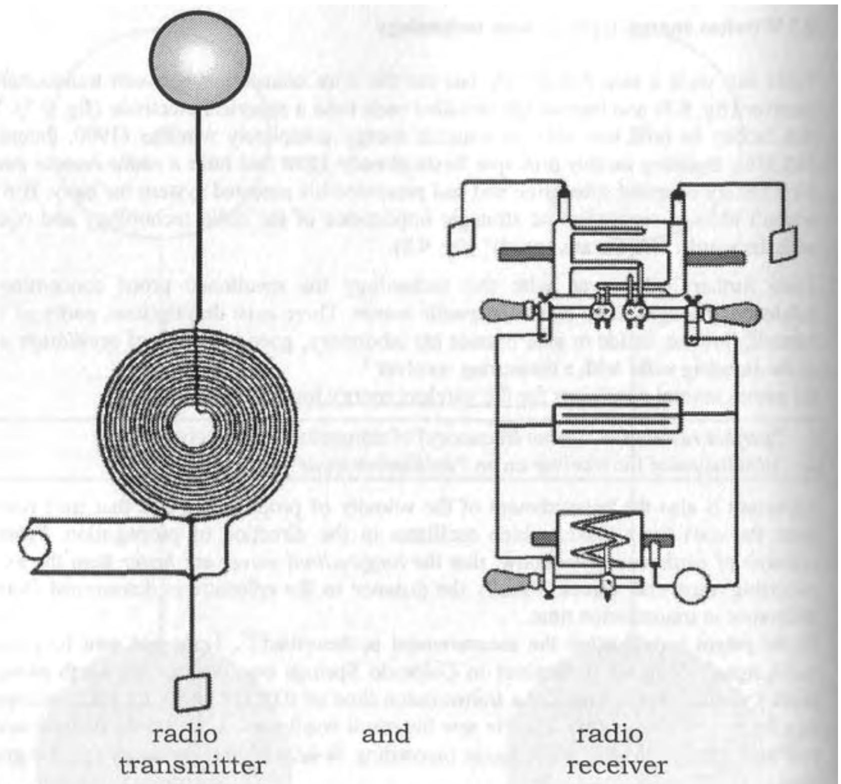
radio
transmitter
and
radio
receiver

Fig. 9.8: Energy wave broadcasting system, Patent No. 787, 412 (1900) of Tesla.

In today's switching technical notation:

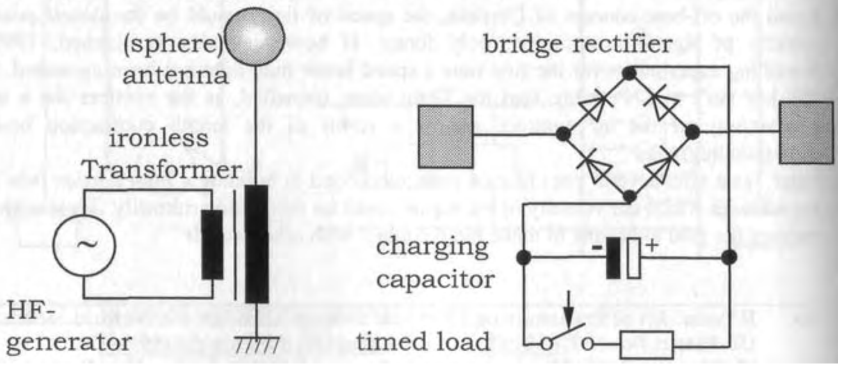
(sphere)
antenna
ironless
Transformer
HF-
generator
bridge rectifier
charging
capacitor
-
+
timed load

9.8 Measuring and switching technique of Nikola Tesla

As an important and accepted experimental physicist with 14 doctor titles and carrier of the Edison medal Tesla always has held on to his measurement results. Not a theoretical interpretation but exclusively correct carried out measurements have shown him the physical reality. But the by Tesla won measurement results were already for 100 years hard to digest and couldn't be brought into accord with any theory. Therefore a switching technical analysis of the by Tesla described and carried out experiments should give us now information over the nature of the free energy, the tachyon energy, the orgone, or however fantastic the terms may read.

The Tesla coil, according to the instructions of the inventor, is a flat coil wound like a spiral in contrast to the copies today on sale which, surely out of ignorance, are mostly cylindrical. Its peculiarity probably is to be seen in the circumstance, that charges moved through the wire of the coil inevitably increase their angular velocity towards the centre, In this way the electrons, which at first are spherical elementary vortices, are pulled apart to vortex rings like planar discs.

Tesla switches the "secondary" called flat coil between two "terminals". Following he sends, stimulated by his "primary", charge carriers from one terminal to the other and back again and produces a standing resonant oscillation.

Mostly replaces Tesla one of both terminals by the earth. He thus assumes, that now the earth as a spherical electrode takes over the function of one terminal. That he again infers from the observation, that a by the transmitter wireless fed little lamp goes out, if he cuts the connection to the ground. Doing so the oscillation collapses. Radio sets on the other hand can also be operated without grounding, as we know!

The degree of effectiveness of today's distribution technology of current due to the heating losses lies clearly beneath 100 per cent. Without the losses of the wiring it lies close to 100 per cent for the discussed one wire energy transmission. There the vortex rings are guided nicely one after another along the line like beads drawn over a string. This result even is to be expected, as far as no vortex "jumps off the wire or "falls apart". For the wireless version Tesla however to his own surprise had to find out that more energy could be received, than his transmitter produced. The measured degree of effectiveness lay above 100 per cent! He therefore called his transmitter a "Magnifying Transmitter" (fig. 9.10). The further transmitter and receiver were away of each other, the further the received energy increased. Tesla inferred from this, that there had to exist free energy and that he had caught that too.

Consequently he had built a receiver for free energy and registered for patent (1901, Patent No. 685,957, fig. 9.9). Tesla states that the amount of energy depends on the size of the "terminal". Of today's sight we could be willing, to describe this receiver plate as a solar panel, but we should know, that the apparatus produced energy even at night! In addition the energy gain was considerable higher than for today's solar panels. Tesla spoke of "radiations ", of an unknown radiation and he in his lifetime has in vain sought-for help of explanation.

The vortex model will also in this question be a valuable help to us.

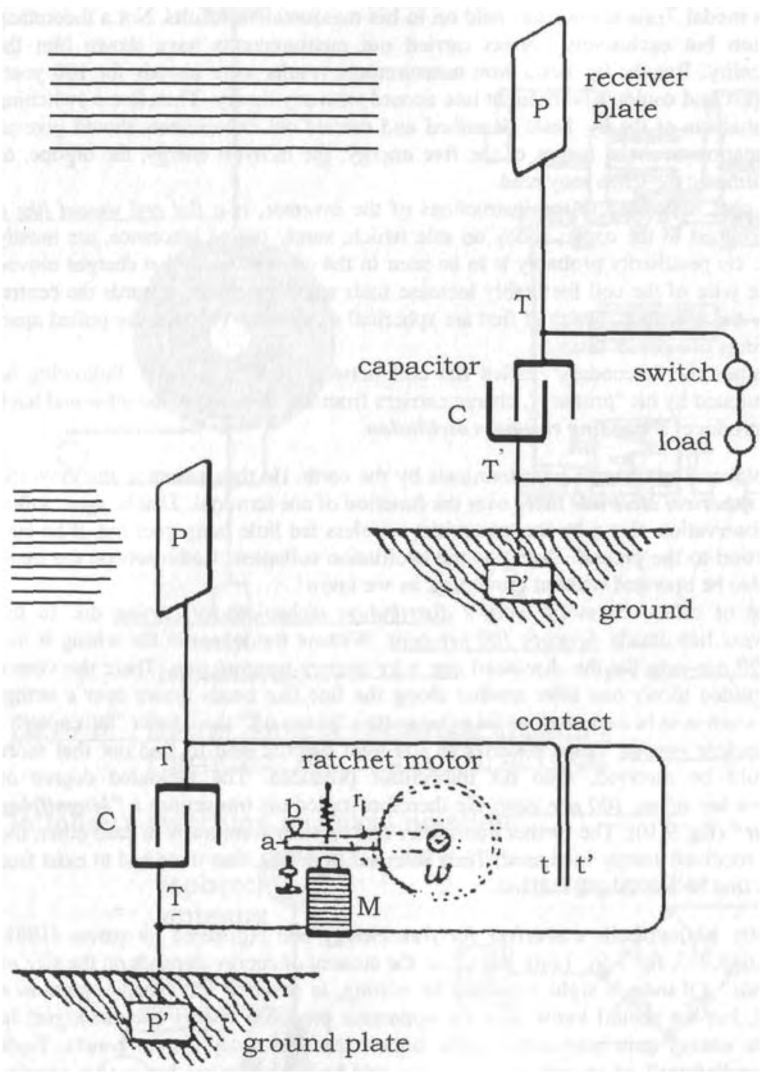
receiver plate
P
T
capacitor
switch
C
load
T'
P'
ground
P
contact
T
ratchet motor
b
C
a
ω
t
t'
T
M
P'
ground plate

Fig. 9.9: Receiver for "free energy",
Patent No. 685,957 (1901) of Tesla

9.9 Energy technical usage

The answer of the potential vortex theory could turn out as follows:
If at the transition of the one wire to the wireless transmission the ring-like vortices is purloined the guiding wire the vortices immediately begin to turn around each other, as is observable for flow-technical ring-like vortices,. In this way the ring-like vortex the next moment shows its inside. If it before was an electron vortex (fig. 4.3), then it now shows as positron, if it was negatively charged, then it now is positively charged. Following it oscillates back again, etc. Wit that the ring-like vortex on the average has no measurable charge and no mass, because it alternately forms matter and anti-matter. Without interaction it has an enormous ability of penetration . In physics such particles are called neutrinos.
Tesla thus had, apart from his transmitted energy wave, which turned out to be neutrino radiation, by chance also caught neutrinos which oscillated synchronously.

According to the actual level of knowledge do neutrinos penetrate the earth and appear also on the night side. The order of magnitude in every second amounts to approx. 66 billion neutrinos per square centimetre. It is a true bombardment. If we would be able, to collect and convert all neutrinos, the won energy would be entirely sufficient, to cover the need for energy of the world population (approx. 27 W/m^2). We merely have to materialize them, thus give them mass, charge and the necessary localization. Tesla was able to do that experimentally! Let's record:
The ring-like vortices, which Tesla with his transmitter has sent on a journey as electrons with an open vortex centre, are neutrinos (fig. 7.12). Tesla requests that transmitter and receiver operate in resonance, thus with the same frequency. Under this condition the receiver collects in all oscillating vortices, so that no one is lost.
If the neutrinos for instance are just positively charged when leaving the transmitter electrode, then an electromagnetic force of attraction takes place, if the receiver electrode at the same time is negatively charged. The required operation with the same frequency and opposite phase guarantees that also the next moment, if both, the neutrino and the receiver, have changed their polarity, the electromagnetic attraction is preserved.
It is obvious, that strange neutrinos which fly past and by chance oscillate synchronously are as well attracted. In that way the power collected in the receiver capacitor will increase further and degrees of effectiveness of over 100% are obtainable. Tesla discharges the receiver capacitor timed with the frequency of resonance (fig. 9.9) and points to the difficulty of an exact keeping of the condition of synchronisation.
Tesla indeed did work on a theory of his own, but never published it. The patent office and his lawyers had to proceed from the Maxwell theory, although Tesla knew only too good, that his apparatus in no way could be described with this theory. It therefore can't be excluded, that important facts in his patent specifications haven't been mentioned at all - or even worse - consciously or unconsciously in an inapplicable manner have been explained with the Maxwell theory.
Perhaps this is the more profound reason, why the numerous Tesla imitations don't want to function quite so well. With the new theory there should be some changes, and we should accept the challenge for the sake of humanity.

The transmitter tower of Tesla (57 m high with a spherical electrode of 21 m) for the transmission in multiplex mode of operation (according the inventor) of:

- telephone calls
- telegraph services
- teleprinter services
- radio programmes
- navigation signals
- time signals
- video signal

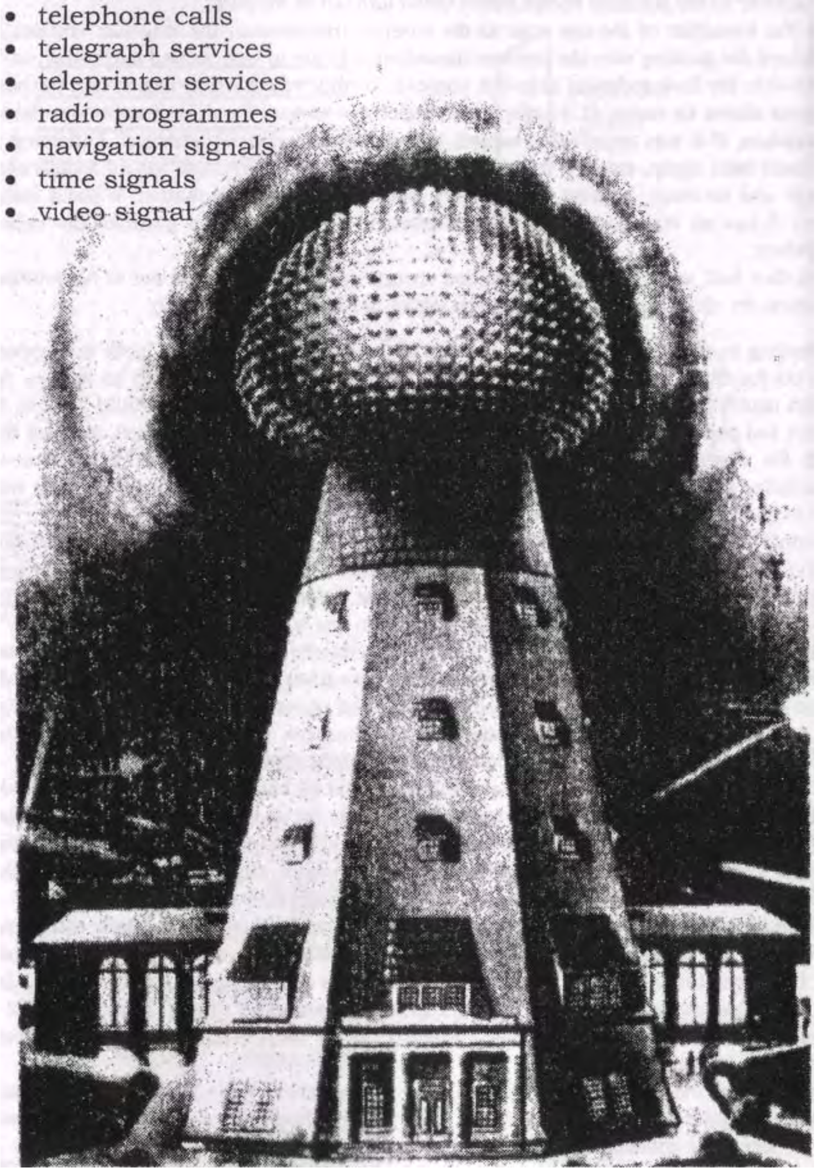

Fig. 9.10: The planned transmitter Wardencliff of Tesla on Long Island, 7.5 MW (1902).

9.10 Environmental compatibility

We now can summarize the different observations and try to find answers for the in the first chapter raised questions concerning the electromagnetic environmental compatibility.
The by Tesla used longitudinal energy wave is a potential vortex wave in the sense of the vortex model of Kelvin, which we again could identify as neutrino radiation (Dirac).
Also other, partly quite fantastic terms, are known like "radiations" (Tesla), "orgone radiation" (Reich), "tachyons" (Feinberg), "grid radiation" (Hartmann), "bioradiation" or "water veins".
Actually these rays have nothing to do with water as a cause. Water with its high dielectricity however favours and influences the course and the local distribution of the radiation. The maximums of the standing wave considered for themselves result lines in straight as a die in the landscape. Of Tesla is known, that he measuring technically could detect the cross points of the lines, which he called oscillation nodes. There exist so-called dowsers, who can make these standing waves out even without technical aids. For that they hold a fork in their hands, which suddenly moves over such lines.
Let's remember that the same potential vortices are used at the nerve conduction as reaction potentials to conduct stimuli. If ring-like vortices arrive at a muscle, it contracts without knowing, if the signal was sent from the brain over the spinal cord or picked up from the environment over the fork and the hand.

Thus if the same signals, which we use for the tasks of controlling and thinking, are also produced by technical apparatus, then this touches the question of the environmental compatibility in a direct way. Above extremely strong points of resonance, which are called "geopathic zones", now and then even is warned about "psychotronics", a function trouble of the brain, which can show in a disturbance of the waking consciousness.
A possible explanation would be that the brain is occupied with so much vortices picked up from the outside, that it isn't able anymore to accomplish its actual tasks. For muscle cramps or an inexplicable crash of a PC an correspondingly interpretation would be obvious.
As long as no gauges are available, the highest caution is imperative!

If Tesla had been able to put into operation his transmitter tower for longitudinal waves ("World Telegraphy", Wardenclyffe, Long Island, s. fig. 9.10) with the estimated power of 7.5 Megawatt and thus had been able to realize his dream of a world-wide wireless communication, then this could have had hardly estimable consequences for the health of the world population.
Shortly before the completion of the project, in the year 1905, Tesla without notice let stop all work. He to that never made an explanation to anyone. Officially it was said, his financial backer J. P. Morgan would have withdrawn his financial means. In any case was Tesla financially ruined with this step.
Perhaps also T. A. Edison was behind it, who was his opponent at that time. Edison committed himself engaged for the use of direct current. Against the alternating current, preferred by Tesla, Edison argued with unhealthy risks, and with that he perhaps wasn't so wrong at all.

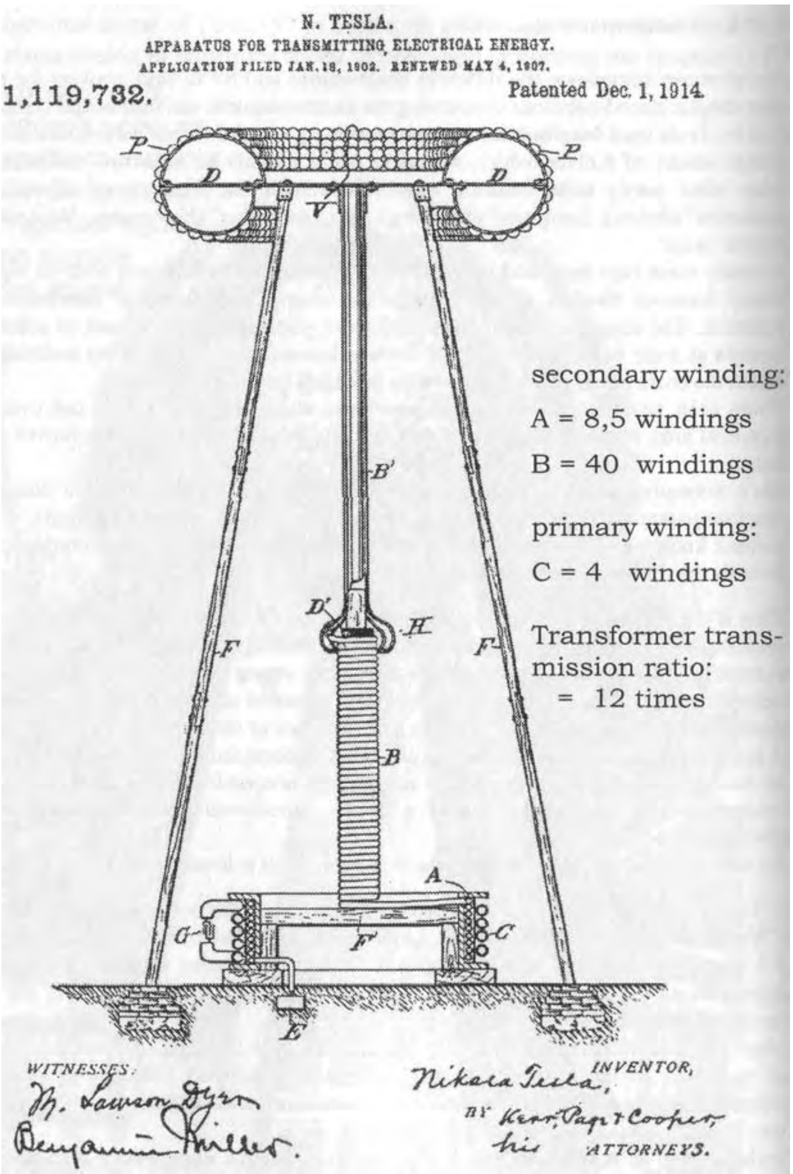
N. TESLA.
APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.
APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.
1,119,732.
Patented Dec. 1, 1914.
P
P
D
D
V
B'
D'
H
F
F
B
A
G
F'
C
E
secondary winding:
A = 8,5 windings
B = 40 windings
primary winding:
C = 4 windings
Transformer transmission ratio:
= 12 times
WITNESSES:
M. Lawson Dyer
Benjamin Miller.
INVENTOR,
Nikola Tesla,
BY Kerr, Page & Cooper,
his ATTORNEYS.

Fig. 9.11: Concerning the technology of the Wardencliff transmitter (fig. 9.10).<i>

<i>: Dr. Nikola Tesla: Complete Patents, (J.T.Ratzlaff), Tesla Book Company (1983), ISBN 0-960356-8-2, P. 435.

9.11 Ecologically fair usage

Tesla not only has stimulated ecologically fair usages of the energy radiation. To that can be counted the use of a Tesla transmitter as a radiation weapon<i> as well as the mentioned building of a Tesla interferometer, the low-frequency interferences of which can lead to mighty gravitational waves and with help of which it should be possible, to produce earthquakes, to sink ships or to shoot down planes from the sky<i>. Against the background of two world wars it can be understood, why Tesla has drawn attention to the possibilities of a use of his system for the purpose of defence.

In today's sight a technology like Tesla's wireless energy transmission (fig. 9.11) hardly would have any chance of a large technical use, because it should be reckoned with enormous problems of environmental compatibility. After all had come out in our considerations, that entirely in contrast to the Hertzian wave just of the Tesla energy wave direct biological effects are to be expected. A purposeful medical use of the Tesla rays however can be imagined.

The collecting of "free" energy waves with a neutrino receiver at first undoubtedly would be a clean and ecologically fair alternative to the present energy technology, with which we heat or nuclearly contaminate our environment. But even the use of free energy converters has limits. Should it turn out that the Photosynthesis uses such a materialization, in which the necessary free electron for the chemical reaction actually is produced from a caught neutrino in the so-called reaction centre and thus the reaction energy is won, then for an as usual excessive use by men it can't be excluded anymore, that the plants on earth die and nothing grows anymore!

I in contrast attribute big chances to the one wire transmission technology. Because for an use of alternative and regenerative energy sources, like the sun in the desert or the geothermal energy in Iceland, the transport of energy has to be made over large distances through the oceans the small losses and the possibility to be able to use sea cables, play a big role. In the residential areas and conurbations on the other hand it has a positively effect, that by the concentration effect of the potential vortices the interference radiation is directed inward. Laid on the bottom of a sea or of a river, such a line in addition is protected against the effects of lightning and sabotage, and it neither spoils the countryside, as the old-fashioned high tension pylons do.

For transportation tasks the mentioned making use of the spin coupling would surely be the most suitable drive. The denser the medium, the more efficient would be the driving effect, for which reason not only flying objects, but quite particularly also ships and submarines could be operated non-polluting in this way without combustion and without exhaust fumes. Build up effects by the water pressure or the sound barrier don't occur. But control might not be entirely unproblematic. If for the reason of a unipolar arrangement of the field a relative velocity to compensate the field becomes necessary, then a correspondingly build up arrangement will do so without considering losses. It is no accident that most apparatus which got well-known have flown the builder sooner or later "past his ears" (note of the translator: broke into pieces). Till a commercially utilizable flying saucer has been built, many technological hurdles have to be taken.

Claims to an up-to date environmental policy
in view of longitudinal electromagnetic energy waves
= Tesla waves = neutrino radiation = scalar waves
= potential vortex waves:

1. Gauges must be developed and built, with which the energy waves can be measured.

2. All technical apparatus should be tested regarding the emission of energy waves. Harmful radiation should be avoided.

3. Only such transmitter antennas may be brought into action, which emit a minimum of energy waves.

4. The ability of biological systems to absorb energy waves should be determined. To that also belongs the question, how many vortices collect locally and how fast they fall apart.

5. Limits for allowed pressure produced by artificially energy waves should be fixed, regulations and laws enacted.

6. The in nature available energy waves should be brought to a technical use, with the goal of a natural energy production.

Fig. 9.12: Catalogue of claims concerning the environmental compatibility in the view of the new field theory.

9.12 Claims concerning the environmental compatibility

Now we still not know, if the using of a handy or another electronic device poses a danger to the user. Now, we surely can't exclude a danger, and the statement of the manufacturer, it is safe because all limits have been met, is worth nothing, as long as the part of longitudinal waves isn't measured. Therefore stands at the very top of the list of the catalogue of claims (fig. 9.12) the development and the building of gauges to measure Tesla radiation. More than half the 700 patents of Nikola Tesla concern the building of transmitter and receiver installations for energy waves. Here a lot of valuable clues are found, especially as he himself has built and used some of the switching circuit designs. In the seminar a rebuild is discussed, with which 1995 the existence of longitudinal waves could be proven [A7].

For an earthquake always both, the transverse and the longitudinal wave, occur simultaneously and, taken exact the same is valid for the sound wave, even if the transverse part doesn't get too far in the air. The emission of both parts in analogy is almost to be expected of an electromagnetic wave generator or transmitter. Actually both waves are detectable at the receiving point: The useful transverse wave and the longitudinally propagating vortices, which show as noise (fig. 4.7). If the vortices are predominant and the useful signal goes under in the noise, then a interruption of the radio operation occurs. Also small gardeners immediately near a radio station have made experiences as e.g. in a distance of one kilometre from the transmitter of RIAS-Berlin. They were able to light their allotments day and night with a neon tube, to which they merely had tied a free wire. The radio station immediately insinuated, that they in an illegal manner had tapped their useful wave and damped it. But it can't be excluded, that the transmitter has sent a longitudinal part in the ether too, which was responsible for the energy transport. But that suggests, that also at other transmitter antennas, thus for the mobile phones, the unwelcome parts is transmitted even than, if we don't use it at all.

Still another problem is added. If in a modern receiver radio waves arrive, then they roll up to a standing wave and according to fig. 1.4 to a vortex over the antenna. According to that we aren't able anymore to distinguish, if the transmission took place as Hertzian wave or as Tesla wave. The separation of both sorts of waves has to take place before the antenna is reached.

An up-to date measurement arrangement would look as follows: In front of the receiver antenna a path with a tunnel, which represents an insurmountable obstacle for the transverse waves (fig. 6.14), is placed. Then only the longitudinal waves pass the tunnel, and these at the end of the tunnel can be recorded and analysed with conventional technology. The tunnel is an elegant possibility, to make use of the part of Tesla waves. After all only this part is relevant with regard to the electromagnetic environmental compatibility. Concerning the above asked question we can take home at least one important realization. The probability is quite high, that you actually notice nothing of the handy radiation, because not every system react in the same manner to certain vortices. The resonance conditions regarding frequency, phase and position always have to be fulfilled, if an absorption of energy waves should occur. (fig. 2.10 B, E. coli bacteria). On the other hand it however can't be excluded, that just you or me synchronize according to frequency and collect 100 % of the transmitted energy radiation. In contrast to the Hertzian wave plays for that the distance to the source of interference only a secondary role! Seen that way worries would by all means be entitled ...

From Indian mysticism:

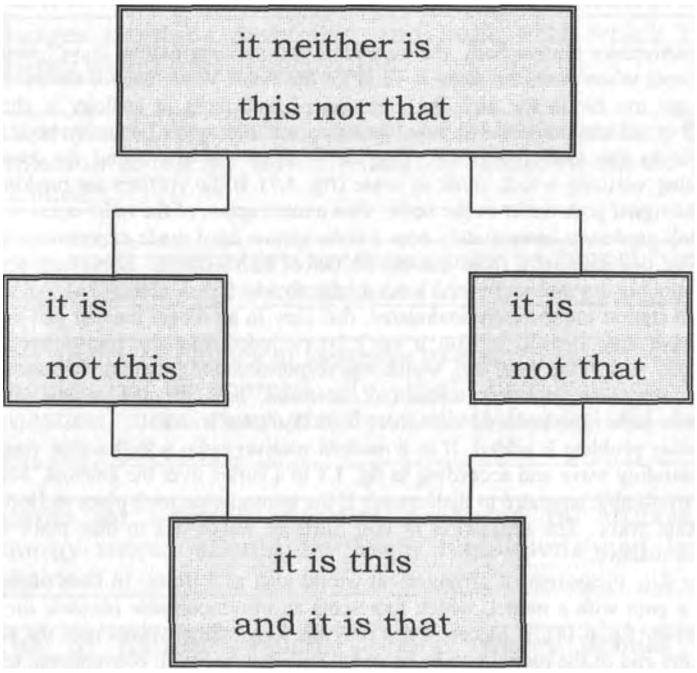
it neither is
this nor that
it is
not this
it is
not that
it is this
and it is that

Fig. 9.13: The structure of the world equation, an example from Indian mysticism

9. 13 Epilogue belonging to part 1

The correctness of a new theory not by all means guarantees, that it also is accepted! A new theory must be used. Only if a concrete need exists then an acknowledgement is to be expected if at first only by the direct beneficiaries.

In november 1994 in Munich the trade fair "Electronica" took place. Here invited by the Deutschen Gesellschaft fur EMV-Technologie (German association for EMC-technology, note of the translator) the books about potential vortices [Al, A2] were honoured with a prize by an internationally renowned EMC-scientist. The number of people, who the potential vortex theory helps further to find answers to problems, which until now aren't explicable, grows further.

Nevertheless, nobody should believe that something will be changed to the content of the text books in the shelves that fast. Habit and laziness encourage to further use the to date interpretation, even if a new theory has ready a better and convincing answer. It will take some time, to give an example, till the heating of water in a microwave oven is explained as eddy losses, as is described sound by the potential vortex theory already today. The discovery of a physical phenomenon may happen in a fright second. Its acknowledgement on the other hand as a rule is a lengthy procedure. If we look back in history we find out that even in the past it hasn't been otherwise.

Of the four fundamental phenomena in fig. 8.2 at first the tension voltages and the charges have been discovered by Coulomb (5.7e).

(5.7c): An effect of the eddy currents was observed already 1825 by the French physicist Arago. He discovered that a magnetic needle suspended above a rotating disc tends to rotate along. Out of ignorance he named the effect "rotational magnetism".

But the eddy currents could only be mathematically described with the help of two laws, the law of Ampere (1820) and the law of induction of Faraday (1831). Because the effect due to the eddy current losses rather was regarded as disturbing, technical applications or patents at first have stayed out. At first Riidenberg has carried out fundamental calculations for eddy current brakes in 1906. The discovery and usage of the "skin effect" is to be attributed to Tesla.

(5.7b): The electromagnetic wave for the first time was mathematically described by Maxwell (1864). He had aimed to calculate the light and he could show, that his theory actually could correctly describe all properties of the light. With that he was able to prove the correctness of his theory. But he had to deal with many sceptics, because he had postulated the dielectric displacement, without however in his lifetime being able to measure it because of its smallness.

This may be the reason, why the discovery is attributed to Heinrich Hertz and was and is spoken of the Hertzian wave. Neutrally seen Hertz at first was nothing more than a enthusiastic pupil, who had worked through the "Treatise"<i>, Maxwells books. All results, which Hertz could observe experimentally, already were contained in the Maxwell-theory and published.

<i>: James Clerk Maxwell: A Treatise on Electricity & Magnetism, vol. 2, Chapter XX: Electromagnetic Theory of Light. Dover Publ. New York

But the great mathematical description of Maxwell at first had a purely academic importance. The interest of the general public on the other hand concentrates on the practical applicability. Hertz had experimented with antennas and has laid the foundations of the building of a gauge for electromagnetic waves (1888). The wireless transmission of invisible waves was convincing. It was an entirely new technology, which stimulated fantasy in view of a technical usage. A practical man, who was stimulated by the theoretical realizations, was Marconi. He goes for the inventor of the radio (1895, Nobel prize: 1909).

Between the three summits, first the calculation and at the same time discovery, second the measurement and third the usage of the wave phenomenon lay 31 years, and the radio technology is developing still further, even today. In the area of the cellular phone network and the telecommunication even a kind of gold mining mood prevails, which can't be impressed much even by hints of an electromagnetic environmental compatibility.

In the euphoria of the past 100 years is fallen into oblivion completely, that besides the by Hertz detected transverse wave in the wave equation according to Laplace and according to the original version of Maxwell also a longitudinal wave was described mathematically. This one was discovered and detected by Tesla in numerous experiments. With reference to his discovery Tesla initially had asserted publicly, Hertz had been at fault, with which he undoubtedly was wrong and had incurred the wrath and the scorn of the scientific community. As a result of this development his experiments haven't been reproduced and the discovery of the scalar wave could fall into oblivion.

Not enough with such sanctions against the inconvenient freethinker Tesla a series of professors like Gibbs and Heaviside have made cuts and discarded all parts from the original version of the Maxwell equations, which by that time weren't considered to be experimentally proven by experiments of Ampere, of Faraday, of Ohm and Hertz. With that the scalar waves fell victim to the axe, what, though it had encountered severe criticism, finally entered as textbook opinion into all textbooks, after Hertz also had sanctioned this measure.

If the field equations according to Maxwell in today's revised version don't describe scalar waves anymore, then all orthodox scientists, who want to bear reference to that, have to bear one thing in mind: discarding a term of an equation is the same as neglecting it. But the term only may be neglected if it is sufficiently small. The first thing every student learns in the physical practical training is that this first has to be tested. For that the term has to be measured and its order of magnitude must be determined. Then it is little helpful and extremely unscientific, if at first is neglected and on the basis of the revised field equations, which have put all scalar waves to zero, is demonstrated that the neglect was allowed or even necessary.

A practical example are the longitudinal wave parts, like they occur and are proven in the near-field of a transmitter antenna. Considering the order of magnitude a neglect is out of the question. On the other hand they should not exist at all according to textbook physics, because they had fallen victim to the axe. Since most scientists in present time do not know about the cutting action anymore, they are postulating field vortices anew in the proximity of an antenna. Field vortices, which are propagating longitudinally in the direction of a field pointer as a scalar wave.

With that they calculate their own arbitrary assumption<i>.
If in the practical training of physics a student unevaluated neglects an essential factor of influence, then his experiment goes as not passed. The experiment must be repeated so long until according to the approved methods of scientific soundness all used neglects have been tested individually and have undergone an error consideration.
Do we now have to deny the cutting action, which had been plotted by Gibbs, Heaviside and others, having the necessary scientific soundness? Do we have to review and rework all textbooks concerning the illegal neglect? Since the cutting action a gap gapes in the field theory!

(5.7d): The potential vortex fills the remaining gap in fig. 8.2. Several research scientists and scientists can be named, who already could observe this phenomenon: in front of all goes Nikola Tesla for discoverer of the Tesla currents, very weak currents which are said to cause extremely high potentials. Here presumably cause and effect have been mixed up, because weak currents never can produce high potentials. Actually the potentials prove to be a result of the potential vortices, whereas the currents aren't more than a result, nothing but leak currents. At least Tesla could use the vortex, but without a usable theory he neither could calculate nor adequately explain it. Besides Tesla Wilhelm Reich should be mentioned, who has collected the vortices in his orgone accumulator.
Mostly only certain aspects were observed and investigated: by Tesla the planar vortex, by Reich the influence on the weather, by Schauberger the water vortices and by all three the medical and biological aspect. The list of names is incomplete and could arbitrarily be continued.

With the discovery of the potential vortex (1990) the basis for the building of a gauge and the technical usage of the physical phenomenon is laid [Al]. It not only concerns the search for water, but also the detecting of the vortex balls and vortex streets, of the standing waves of the energy radiation in the air, at the workplace, at the bedroom, in clinics, in recreational areas and hotels. As explained at the start, neither the field strength of a Hertzian wave nor the arising heat development can be made responsible for biological or technical harm.
It primarily are the newly discovered vortices of the electric field, which take effect. The effects can, as we have seen, as well be good for health as bad for health. Intensity, plane of polarization, vortex configurations and many other characteristics play a role here. To research these influential factors gauges for vortices will be needed as well.
We have to realize that in the technical domain the electromagnetic compatibility of an apparatus is determined by its sensitiveness to vortices, thus by the fact how many and which vortices can cause a function trouble. To determine the environmental compatibility of a product the emitted vortices, the energy radiation, have to be measured. Limits for high tension lines, for screens or handheld phones must be given in units of the potential vortices. The potential vortex has shown us the way to a unified theory and has brought along a new picture and understanding of our environment. It with that wants to show us the correct way for an ecologically compatible dealing with nature.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. 1986 Springer-Verlag Berlin, Seite 335

Part 1: Edition belonging to the lecture:
"Electromagnetic Environmental Compatibility"

Prof. Dr.-Ing. Konstantin Meyl

Scalar waves

Abstract:

1. Auflage 1996, 4. Auflage and 1st English Edition 2003

Both technical and biological systems can be influenced by electromagnetic fields, whereby numerous questions still are open, like e.g. concerning limits and the nature of physical interference fields. The book shall do justice to the circumstance, that a fact oriented discussion about „electrosmog" implies an analysis of possible reasons, a mathematical derivation and a physical argumentation.

We proceed from the assumption, that only that part of the electromagnetic wave should be considered for malfunctions, which has been absorbed and which has rolled up to a field vortex. The effectiveness depends on the amount of the produced vortices and on the life span, the decay of vortices.

Analogous to the sound wave vortices in space are propagating as longitudinal waves. In this context is pointed to numerous effects. Examples for the technical and biological use of these standing waves are the energy transmission of Nikola Tesla as well as the nerve conduction, which functions in a corresponding manner. If the same vortices, which man uses for conduction of information, are emitted by technical devices, then biological reactions can't be excluded anymore and worries with regard to „electrosmog" seem to be justified.

Causes, phenomena and natural scientific consequences

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald
www.k-meyl.de meyl@meyl.de

Scalar Waves

From an extended vortex and field theory to a technical, biological
and historical use of longitudinal waves.

Part 2

by

Professor Dr.-Ing. Konstantin Meyl

Edition belonging to the lecture and seminar "Electromagnetic environmental compatibility"

(Original title: "Elektromagnetische Umweltvertraglichkeit")

Translated out of the German language
by Ben Jansen
(2000-2003)

* * *

Part 2: Edition belonging to the energy technical seminar

Free energy and the interaction of the neutrino radiation

***

INDEL GmbH, Verlagsabteilung

Villingen-Schwenningen 1996-2003

ISBN 3-98O2 542-4-0

Preface to the seminar

The point of a seminar is, to deepen, to practise and, as far as possible, to practically appy the material of a lecture. The knowledge of the content of the lecture hence is a prerequisite for the participation.
For the reader of this book that's tantamount to the recommendation, to have read the first part, the edition belonging to the lecture, before<i>. Here the questions concerning the „electromagnetic environmental compatibility" are asked and the necessary bases for their answering is laid. Also practical consequences for various areas of science are indicated. The deepening most suitable should be made in form of a seminar, subdivided into the here presented part 2 to the energy technical seminar and a part 3 to the information technical seminar. Part 2 correspondingly concerns the energy technical aspect of electric or magnetic longitudinal waves, whereas part 3 is dedicated to the information technical aspect<ii>. Because it concerns a book which merely for reasons of usefulness is published in three parts, the chapters are consecutively paginated. References to chapter 1 to 9 hence automatically relate to part 1. The numbers of the figures and tables as a rule are identical with those of the chapters, in which they are discussed.

The seminar should lead on over the pure reading, consuming or listening and should stimulate to join in. All involved persons may and should give ideas and ask questions, even if these may sound little orthodox. The scientific working method takes, that is struggled for answers and even is argued, if necessary. To reach this goal, it mustn't exist any obligation or censorship, neither for the leader of the discussion nor for the participants of the seminar.

The seminar is being carried out since the summer semester 1997. The works of the seminar written by students treat the knowledge of text books of the respective theme. Following the lecture the answers are discussed and compared to those of the theory of objectivity<i> and other models of explanation. This procedure in this edition belonging to the seminar is reflected at some points, if for instance a chapter is completed with a „discussion".

The first edition of this 2nd part still was incomplete and has been handed out to the participants of a congress in Switzerland instead of a manuscript belonging to the lecture the 17th of Nov 1998. The here presented second edition in the meantime to a large extent is complete, but surely not yet perfect. In accordance with the experience made with the first part of the book also for this 2nd part a third and revised edition, in which the ideas of the participants of the seminar and of the readers find consideration, will be due to be dealt with after a year. The reader of the second edition has to console himself with the fact that a lively seminar constantly is changing and developing further. And that has to be so!

Villingen-Schwenningen, January 1999

<i>: Electromagnetic environmental compatibility, Part 1, Edition belonging to the lecture, INDEL Verlagsabteilung, 1996, see page 1-218 of this issue.

<ii>: Electromagnetic environmental compatibility, Part 3, Edition belonging to the information technical seminar, 2002, see page 443 - 625 of this issue.

10. Oscillating interaction

A theory is not an end in itself, even if it sounds very convincing. It has to be measured by its applicability. As an entry into the practical consequences, which result from the theoretical part<i>, the question about the validity of Kepler's laws is raised.

10.1 Kepler's laws

The "radius vector", a line drawn from the sun to a planet, sweeps out equal areas in equal periods of time. At least has teached us Johannes Kepler it that way (fig. 10.1). The balance of forces, the gravitation on the one and the centrifugal force on the other hand results in the innermost planets of our solar system orbiting the sun very much faster than the outer planets (Mercury in 88 days, the Earth in 365 days, Jupiter in 4333 days and Pluto in 90465 days!).
For the inner planets as well as the big planets Jupiter and Saturn Kepler's laws are still found confirmed. But that shouldn't apply anymore for the outermost planets of the solar system. Beyond Saturn should prevail changed rules as is said, based on observations of the Voyager spacecrafts<ii>.

If we direct our view to an unknown galaxy, then does it rotate around its centre and in doing so to a large extent keeps its form. Despite rotation of its own an elliptic, a barred or even a spiral galaxy virtually doesn't change its characteristic form. From this follows, that the inner stars of a galaxy are considerably slower on their way than the outer stars! But we expected exactly the opposite.
According to Kepler's regularity the outermost stars would have to orbit extremely slow, In order not to be hurled into space as a result of the centrifugal force. But then a galaxy wouldn't keep its structure. The spiral form, as it already has been observed and classified by Hubble (fig. 10.2), merely would be an accidental exception as a momentary picture, but by no means the rule.
We have to take note, that the structure and in particular the cohesion of a galaxy can't be explained with Kepler's laws<iii>.

<i>: Konstantin Meyl: Electromagnetic environmental compatibility, Part 1 of this book: Causes, phenomena and natural scientific consequences.
<ii>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991)
<iii>: The basic laws of the universe start to rock: ,,What is the matter with the galaxies? They rotate in their fringe ranges much faster, as is allowed by the laws of physics. Or is something wrong with these venerable laws? The astronomers and physicists stand for the dilemma to have to decide between the two alternatives: feign the observations us an other world or do we calculate wrong since centuries?" (translated), Bild der Wissenschaft Nr. 2, 1989

Kepler's 1st law:

The planets move in elliptical orbits, with the sun at one focus.

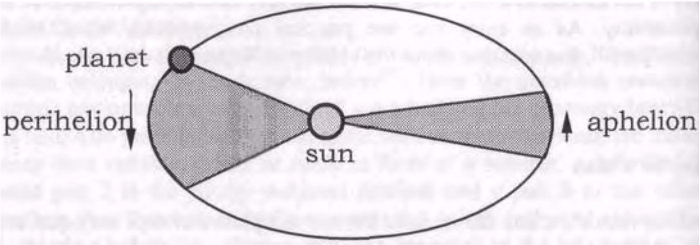
planet
perihelion
sun
aphelion

Fig, 10.1: Kepler's 2nd law (concerning the conservation of angular momentum):

The line drawn from the sun to the planet sweeps out equal areas in equal periods of time.

Kepler's 3rd law:

The ratio of the squares of the revolutionary periods of two planets is equal to the cube of their average distance to the sun:

$$\frac{t_1^2}{t_2^2} = \frac{r_1^3}{r_2^3} \tag{10.1}$$

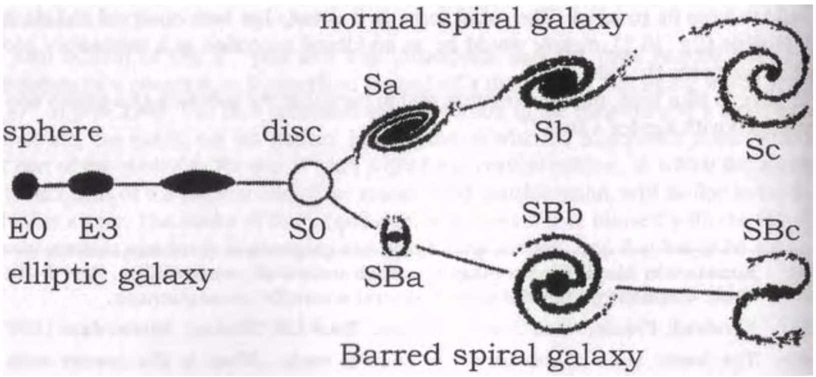
normal spiral galaxy
Sa
Sb
Sc
sphere
disc
E0
E3
S0
SBa
SBb
SBc
elliptic galaxy
Barred spiral galaxy

Fig. 10.2 The classification of the galaxies according to Hubble<i>

<i>: according to H. J. Lugt: Vortex Flow in Nature and Technology, page 223

10.2 Unknown interaction

So which interaction keeps a galaxy together? We today believe to know four different sorts.

I. The gravitation<i>: But since Kepler's law isn't valid in this case anymore, the gravitation is ruled out from the start. Obviously for the distances in a galaxy it hardly is effective.

II. The electromagnetic interaction: It is responsible for the structure of the atoms. Looked at from the outside atoms carry no charge, i.e. the charge conditions are balanced. A binding of our sun to the centre of the Milky Way by an electromagnetic interaction thus is ruled out as well.

III. The strong interaction: Since for the proton another charge distribution is measured, as a single positively charged particle should have according to the normally used theory, the strong interaction was introduced as a nuclear force, to explain the big error, the grave difference between measurement and calculation. The good advice hence reads: instead of giving birth to postulates at random, first of all the fault should be searched for in the normally used theory<ii>!

IV. The weak interaction: It quite obviously is involved in the particle decay<ii>. Both, the weak and the strong interaction, only have an extremely short range. With this property they consequently won't be able to keep a galaxy together.

Conclusion: In a galaxy a still unknown interaction takes effect, and science is requested to search it.

Both interactions with infinite range, the electromagnetic interaction and the gravitation occur as a result of static fields, therefore assume a constant charge or a constant mass. Considered more exactly in that case it merely can concern special cases. Gravitational waves, which reach our earth and which are detected in very costly experiments<iii:>, already show that the existence of alternating fields can't be excluded and oscillating interactions by all means are conceivable! The physical research at present probably is on the right track. The researchers however don't have ready an usable explanation yet. We accept the challenge!

<i>: Derivation of Kepler's 3rd law in fig. 11.10

<ii>: Konstantin Meyl: Potentialwirbel, Band 2
INDEL-Verlag, Villingen-Schwenningen 1992, ISBN 3-98O2542-2-4

<iii>: Gero v. Randow: Wenn kosmische Katastrophen Raum und Zeit verbiegen, zum Thema Gravitationswellen-Detektor, VDI Nachrichten Nr.9, 1.3.91, S.32

Analogy:

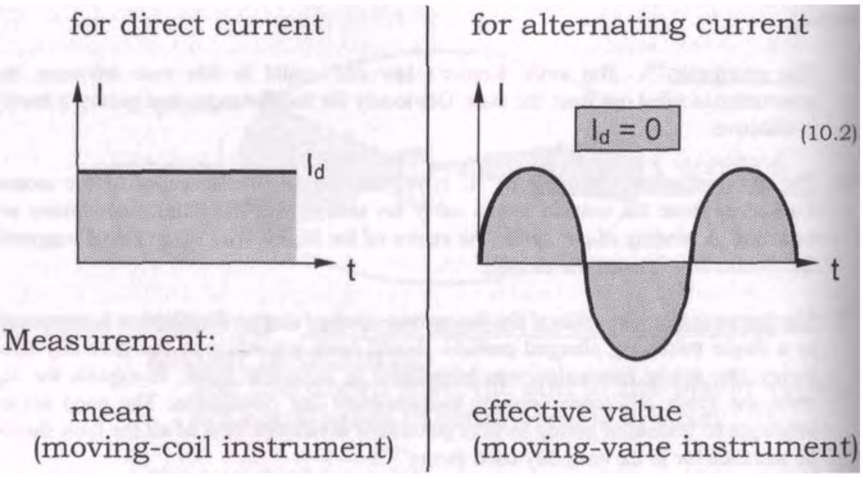
for direct current
for alternating current
I
I
I_d
$I_d = 0$
(10.2)
t
t
Measurement:
mean
(moving-coil instrument)
effective value
(moving-vane instrument)

Fig. 10.3: Direct current and alternating current

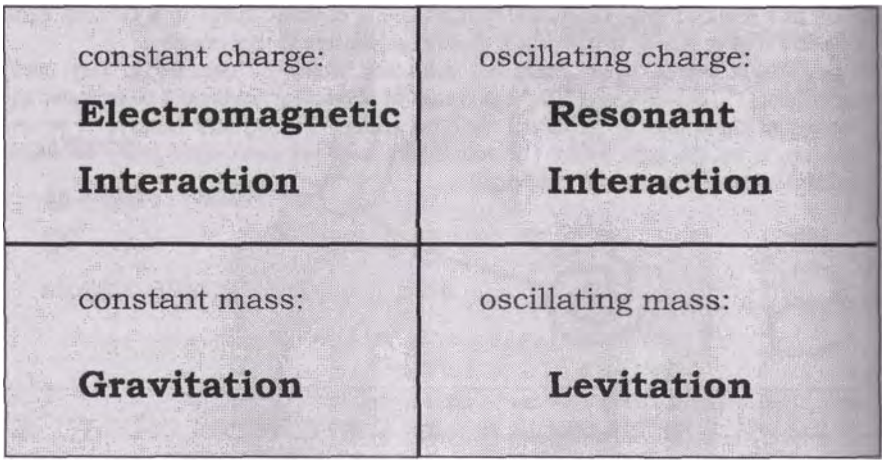

constant charge: **Electromagnetic Interaction**	oscillating charge: **Resonant Interaction**
constant mass: **Gravitation**	oscillating mass: **Levitation**

Fig. 10.4, a: The four fundamental interactions

10.3 Harmony of the alternating current engineers

Today's situation can be clarified by the following picture: you engage a "direct current engineer" to measure the tension voltage in the socket of our power supply system. The „dyed-in-the-wool direct current engineer", who never has heard anything of alternating current, reports: „No tension voltage is measurable". If he for reason of your doubtful expression looks still more exact he will realize: ,,The pointer of my moving-coil instrument strangely fidgets around zero, but the swing is so small that one confidently can touch it", (fig. 10.3)
Modern science is accustomed to say, without a valid theory and without technical measurability nothing can exist, what mustn 't exist. If you drop dead after the experiment, then you probably had a weak heart or other afflictions. In such cases as a rule the victim himself is to blame and by no means theoretical physics!

In the case of our power supply system the answer is known: The mean of the alternating voltage of the network is approximately zero. The pointer of a moving-coil instrument for reason of its inertia can't follow the fast changing anymore and only fidgets on the spot. The effective value however amounts to around 230 Volts. But to measure it you need another device, for instance a moving-vane instrument. Seen so, direct current describes the special case of alternating current with frequency zero.

The properties of alternating current can be depicted best by an extraterrestrial observer. He will tell us: Seen from a distance at least at night a great harmony seems to prevail on earth. All lights in the streets and cities twinkle completely synchronously. All generators are in resonance among each other and with all consumers. There are two big races: The 50 Hertz race and the 60 Hertz race, which appear if the earth turns further and the 50 Hertz race is switching off again its twinkling lamps.

The synchronization necessary for an exploitation of alternating fields is so obvious for us, that hardly anybody realizes, that a consumer operated with only one Hertz difference to the frequency of the network can't be supplied with power anymore. Apart from the correspondence in frequency it even depends on the correct phase. The phase angle must be between 0 and +90°. This corresponds to an efficiency factor $\cos\varphi$ between 1 and 0. The cable connections serve both the transport of current and the synchronization of all the generators and consumers connected to the network. The frequency is kept so constant that simple clocks, switching stations and even old record-players can be operated synchronously with the frequency of the network.
The synchronization of the feeding in power stations is supervised by a station of its own, which dictates the time. It is true that we aren't capable of seeing the twinkling of the lamps anymore for reason of the inertia of our eyes, but it can be detected and filmed with high-resolution cameras. Even if we can't perceive the harmony of the alternating current engineers, it nevertheless exists.

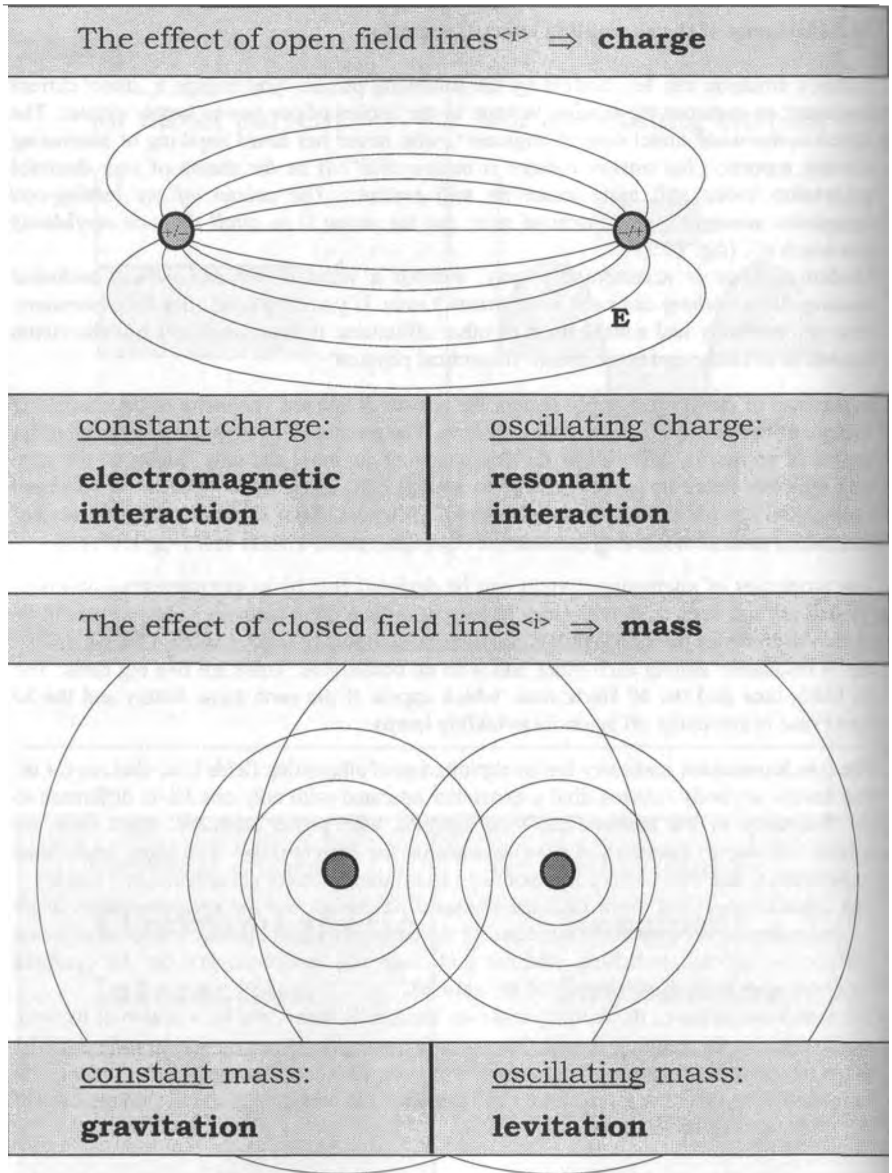
The effect of open field lines<i> ⇒ charge
+/-
-/+
E
constant charge:
electromagnetic interaction
oscillating charge:
resonant interaction
The effect of closed field lines<i> ⇒ mass
constant mass:
gravitation
oscillating mass:
levitation

Fig. 10.4, b: Explanation of the fundamental interactions

<i>: Konstantin Meyl: Potentialwirbel, Band 2, INDEL-Verlag (1992)
see also in Part 1, chapters 6.8 and 6.9

10.4 Four fundamental interactions

These considerations suggest, that also gravitation and electromagnetic interaction merely describe the special case of oscillating interactions with frequency zero. To avoid confusions, we'll have to think of new names.

The electromagnetic interaction can be clarified by means of open field lines, which start at a charged body and end again at another body of unequal charge. In physics classes it is normal to make the field lines of a magnet visible with iron filing. Between the unlike poles a force of attraction is observed<i>.

If we this time assume that both magnetic poles change their polarity at the same time, then the force will decrease for a short time during the changing of polarity, to afterwards be active again in full magnitude and in the same direction. Therefore a force of attraction is observed again even for a reversed polarity.

The generalization hence reads: The electromagnetic interaction will occur also in the oscillating case, but in weakened form, if both bodies involved in the interaction change their polarity synchronously and if they are in resonance. A name by analogy would be "resonating interaction" (table 10.4).

It is known of the electromagnetic interaction, that its effect is larger than that of the gravitation by powers often. This presumably has to do with the described and observable bundling up of the open field lines, whereas closed field lines can't be bundled up. The gravitation hence is linked with the closed field lines, which surround all elementary particles, every atom and every body<i>.

The opposite of the bundling up is the repulsion of open field lines, for which reason here also forces of repulsion can occur. For the gravitation however no repulsion is observed, because closed field lines virtually can't be influenced.

Apart from the circumstance that the effect generally will be smaller in the oscillating case, similar properties are to be expected. Also its range will be infinite as well. It is recommended to call the case of oscillating charges, as already said, "resonating interaction" and the case of oscillating masses, the oscillating gravitation, „levitation" (table 10.4).

The term „levitation" is very appropriate, but not new. Unfortunately until now no generally binding definition existed, what should be understood by that, for which reason misinterpretations and irritations can't be excluded. Mostly levitation is linked to a cancellation of gravity, up to a state of free floating, but we will see that quite other phenomena become describable with this term.

10.5 Resonating interaction

The question, what keeps a galaxy together, now can be answered unambiguously. The well-known interactions already have been excluded. If for the enormous distances the gravitation can't keep the outer stars anymore in accordance with the Kepler rule, then the levitation won't be able at all.

Example: central star S_z with 3 planets P_1-P_3
and with 4 neighbouring stars S_1-S_4

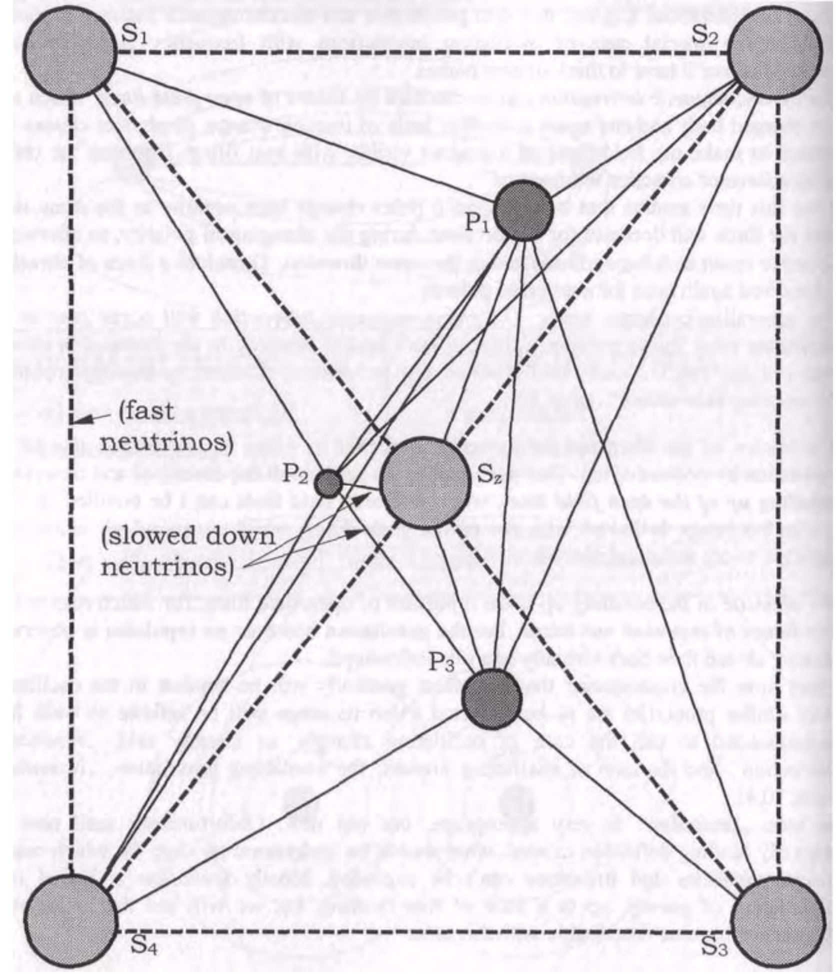
S1
S2
P1
(fast neutrinos)
P2
Sz
(slowed down neutrinos)
P3
S4
S3

Fig. 10.5: The invisible threads of a resonant interaction

milky way-radius: $15000pc*3*10^9 = 45*10^{16}$ km
sun system-radius: $50\ a*15*10^7 = 7,5*10^9$ km

$\frac{45\cdot10^{16}}{7,5\cdot10^9} = 1,27\cdot10^8$	the resonant interaction is more than eight decimal powers bigger than the gravitation!

Therefore the oscillating charge is left. Actually the resonating interaction will reach to the outermost areas of a galaxy. The bundling up of the field lines results in the centre of the galaxy and one of its stars to stand in an exclusive resonance to each other, what looked at from the outside looks like, as if the star hangs at an invisible string or a "rubber band" and thus turns around the centre.
Because quite a few stars hang at the centre of a galaxy, it can be assumed that it provides correspondingly many resonance. The centre perhaps is comparable with an extremely broad banded transmitter, which operates on all channels simultaneously. The stars then pick as a receiver the for them suitable channels and hang themselves by „rubber band" at the heap (fig. 10.5).

Should there exist any particles with an oscillating charge, which synchronize between centre and star with the resonating interaction, then they will mediate between both partners. If we assume that the centre at one channel just is positively charged, then all at the same time negatively charged particles will be attracted, the positively charged particles however repelled. Whereas the unlike particles in the centre participate directly in the production and maintaining of the oscillation, the like positively charged particles will be hurled into space.
But at the same time does a star, which clings to the centre, have to be negatively charged. It hence attracts the repelled particle. The particle thus drifts from the centre to the star, even then if all three, the centre, the star and the particle, change in polarity. The result is, that the stars grow in diameter by collecting the particles. Only because our sun actually grows, it has the chance, to sometime become a red giant!

Since the sun radiates, as is well-known, in every second a radiation equivalent of 5 million tons, it permanently has to be supplied with a substantially greater amount of matter. If a resonating interaction should occur, then our sun will get its „material" supplied from the centre of the Milky Way and that is a black hole!
But no particle with a mass comes out of such a hole, yes not even light. For a particle to be able to leave the black hole, it should have neither charge nor mass. At most an oscillating charge and mass would be allowed. Such a particle would have an enormous ability of penetration as a result of the missing static interaction. It would be able to rush through the earth unhindered.

According to actual knowledge only neutrinos have the described properties. One also knows that the centre of our Milky Way represents a mighty source of neutrinos. From this derivation follows:

1. As mediators of the resonating interaction serve synchronously oscillating neutrinos.
2. Starting with the proof of the neutrinos it should be able to backwards also prove the existence of the resonating interaction.
3. If, as a practical consequence, we imagine that the centre of the Milky Way wouldn't supply neutrinos anymore. Then the whole galaxy would fall apart and not one of its stars would shine anymore.

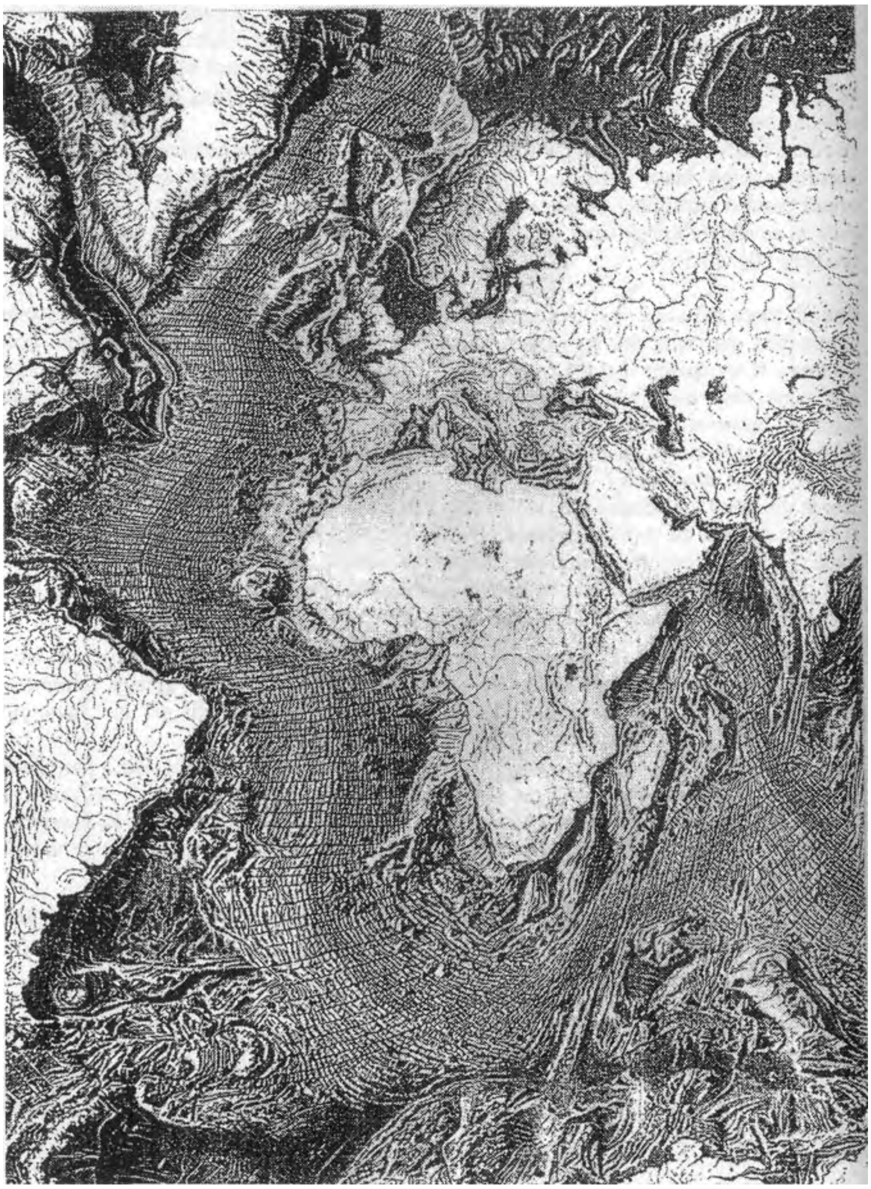

Fig. 11.1: The Atlantic ocean floor<i>

<i>: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 79

11. The growing globe

11.1 Long-distance effect of the neutrinos

The long-distance effect thus lies in the circumstance that the neutrinos, in the case of a resonance of the source of neutrinos and the receiver, span an invisible „rubber band" between both, which is called resonating interaction and keeps the two together.

As a transmitter of neutrinos functions for instance a supernova, the death of a star, in which 98% of the mass annihilates under emission of neutrinos or a black hole, which continually swallows light and matter and spits them out again as neutrinos after a still unexplored „process of digestion". The process, which in the case of a supernova takes place as a singular occurrence, in a black hole possibly takes place permanently. The hurled out neutrinos on the other hand serve the sun as a source of energy. A receiver of neutrinos then for instance is our sun. So that the hard and very fast cosmic neutrinos become utilizable for the sun, they at first have to be slowed down. But that is only partly successful:

1. Some very fast ones manage to pass through the sun and fly out again on the other side of the sun. The compared to the cosmic neutrinos strongly slowed down neutrinos then are called solar neutrinos.
2. Another part can be further slowed down and materialized. As a result of the oscillating mass of the neutrinos as well particles of matter as also some particles of anti-matter are formed. The particles of matter make the sun grow.

i. The with matter incompatible anti-matter annihilates under emission of light as is well-known. For this and for no other reason our sun shines!

Also the planets have such a neutrino reactor at their disposal. Only so the heat in the inside of the earth is explicable! It can be assumed that the planets materialize less the fast and hard cosmic neutrinos and that they are served more by the slowed down solar neutrinos, which our sun releases again unused.
As is well-known radiates the planet Jupiter already today twice as much radiation energy, as it from the sun receive. In this typical encyclopaedia type balance the involved neutrinos of course are not considered. But it shows that Jupiter is on the best way to become a sun itself. Its moons then will become planets.
From this the example we also see that with increasing mass the crust of the planet becomes thinner and thinner at the expense of the neutrino reactor in the inside, until it finally is eaten up and the celestial body openly starts to shine. Astronomers report for reason of their observations of the formation of more and more new stars.

One part of the collected neutrinos thus is materialized by the planet. In the case of our earth it contributes to its growth.
Who doesn't want to believe that the earth becomes bigger, should look at modern maps of the oceans, on which the topography of the ocean floor is shown. According to the theory of Alfred Wegener concerning the continental drift North and South America on the one hand and Europe with Africa on the other hand steadily drift apart since 200 million years. The result can be read at the atlantic ocean floor. The gaping, chequered rift zones to the right and left of the Mid-Atlantic Ridge show how the earth is torn apart (fig. (11.1).

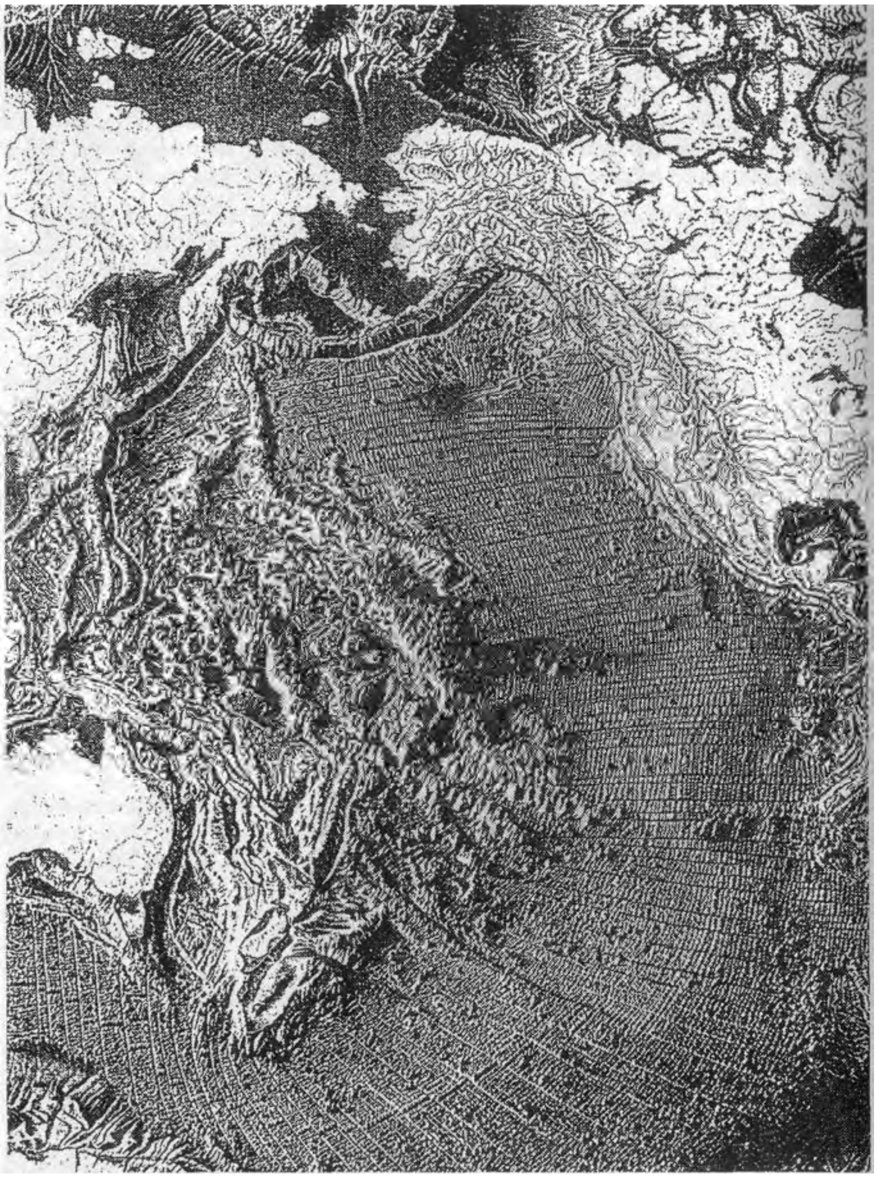

Fig. 11.2: The pacific ocean floor<i>

<i>: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 78

11.2 Model of expansion

Geographers assume that at another point to the same extent continental plates are subduced and molten. That thus has to happen in the Pacific Ocean. But the sea maps tell one just the opposite (fig. 11.2). At the ocean floor of the Pacific Ocean the same rift formations are found as in the Atlantic Ocean (fig. 11.1). That means that America drifts away from Australia and Asia exactly as from Europe, without being crushed or molten in doing so!

The only possible answer thus is: the earth grows.

The characteristic rift zones in addition are found around the Antarctic. From this the conclusion can be drawn that the 7th continent slowly is moving away, while the biggest habitable mass of land predominantly stays behind on the northern hemisphere, by Eurasia and North America forming a clip around the North Pole.

Concerning the evolution of the earth there are and were numerous approaches in various directions. Paul Dirac<i> at his time postulated, the brightness of the sun should decrease as a result of a decrease of the gravitational constant as well. In contrast to that Astrophysicists today observe just the opposite (Ingersoll 1987). According to the hypothesis of Carey<ii> energy will transform in matter in the universe. According to the idea of Oesterle<iii> aether particles are absorbed, which make our globe grow. Also other research scientists share the idea of the growth of the earth with him in their reports<i4,5,6>. As a geologist Oesterle<iii> cites his colleague Herzig: „The at the Mid-Oceanic Ridge newly produced oceanic crust has to be consumed again at other points, because otherwise the earth would expand" and criticizes the „plate tectonicians" that they would postulate their model without physical grounds.

He gives some arguments for the correctness of the model of expansion <iii>:

a) Subduction: The already discussed missing of zones of subduction and of melting of continental plates to the assumed extent.

b) Paleomagnetism: errors and mistakes in the hypothesis of the migration of the poles.

c) Continental depth drillings: They brought much higher temperatures in depths from 4000 meters, as expected and calculated according to models.

d) Stand of the sea water: Only if the water can spread in newly forming oceanic basins it can be explained, why the covering with water on earth continually is going back. This argument we want to investigate with a derivation of our own.

<i>: Dirac, P. A. M. Nature 139, p. 323 (1937)

<ii>: Carey, S. W.: Theories of the Earth and Universe. Stanford University Press: 1-413, Stanford, California

<iii>: Oesterle, O.: Goldene Mitte: Unser einziger Ausweg, Universal Experten Verlag, Rapperswil (1997), ISBN 3-9520261-9-0

<i4>: Hilgenberg, O. C: Vom wachsenden Erdball, S. 1-56, Eigenverlag, Berlin (1933) und Neues Jahrb. Geol. Palaont. Abh. 116, 1, Berlin (1962)

<i5>: Jordan, P.: Die Expansion der Erde. Vieweg 1-180, Braunschweig (1966)

<i6>: Giancarlo Scalera, K.-H. Jacob: Why expanding earth? Institute Nationale di Geofisica e Vulcanogia, Roma and Technical University of Berlin 2003

Today's earth:

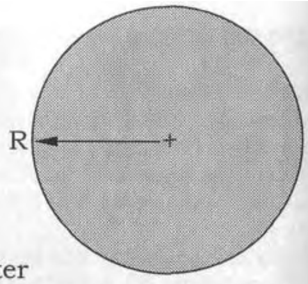

equatorial radius: $R = 6378$ [km]

earth's surface: $O_E = 4 \cdot \pi \cdot R^2$
(sphere) $= \underline{\underline{5{,}112 \cdot 10^8}} [\text{km}^2]$

29% of this area is land and 71% is water
for an average water depth of $t_w = 3.8$ [km] results a

water volume of: $V_w = 0{,}71 \cdot O_E \cdot t_w = \underline{\underline{1{,}38 \cdot 10^9}}$ [km³] (11.1)

The earth **200 million years ago** (super continent: **Pangaea**):

today's land area (29%) was 100% of the earth's surface!

Pangaea surface area: $O_p = 0.29 \cdot O_E$
$= \underline{\underline{1{,}48 \cdot 10^8}}$ [km²]
$= 4 \cdot \pi \cdot r_p^2$ (11.2)

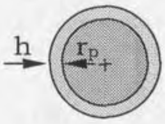

Pangaea radius: $r_p = \sqrt{O_p/(4 \cdot \pi)}$
(corresponds to $= \underline{3435}$ [km]
the radius
of the sphere of shelf): = 54% of today's radius

Pangaea volume: $V_p = \frac{4}{3} \cdot \pi \cdot r_p^3 = \underline{\underline{1{,}697 \cdot 10^{11}}}$ [km³] (11.3)

Pangaea was covered with water h [km] high:

volume: $V = V_p + V_w = \frac{4}{3} \cdot \pi \cdot (r_p + h)^3 = \underline{\underline{1{,}71 \cdot 10^{11}}}$ [km³] (11.4)
= 15,7% of today's volume of the earth

$r_p + h = (3 \cdot V / 4 \cdot \pi)^{1/3} = 3443{,}8$ [km] (11.5)

Pangaea water-level h above the NN (sea-level) at that time:

$$\boxed{h = \underline{\underline{8{,}78}} \text{ [km]}} \quad (11.6)$$

Fig. 11.3: The calculation of the covering with water on earth 200 million years ago

11.3 Ancient Egyptian legend of Genesis

We don't have to search long for evidence. If we go back 200 million years in the history of the earth, as all continents still were united and formed the super continent Pangaea, as the 29% landmass of today thus constituted 100% of the earth's surface. At that time the diameter of the earth was almost half of today's diameter (exactly 54%). But if one distributes the amount of water of our oceans of today over the smaller earth, then possibly the water stood the young earth up to its ears in a first rough estimate. We now want to calculate how high the water stood.
For that we determine at first the water volume of today's oceans, by multiplying 71% of the earth's surface O_E with the average water depth. In doing so it is supposed that the water volume has not changed in the course of time. As an approximation this assumption could be correct if the factors, which influence the water volume, mutually compensate. On the one hand it has to be taken into account that in the process of fusion in the inside of the earth apart from other materials also juvenile water is formed, but on the other hand the water volume is reduced by photosynthesis and by the splitting of water molecules. In the case of the newly formed water it should concern roughly one cubic kilometre per year<i>. The photosynthesis however possibly can be estimated by means of the formed oxygen of the air just as the splitted water by the content of oxygen of the waters and the seas. Since the processes are subject to temporary fluctuations, the exact estimate is difficult. At least should an effect of compensation more or less be taken into account.

Next we calculate the surface of the earth 200 million years ago, the super continent, which Alfred Wegener called Pangaea. If the 29% continental land mass of today at that time constituted the whole surface, then the diameter of the earth at that time was determined at 54% of today's diameter, then the volume together with the water volume would amount to only 15.7% of today's volume and the water stood 8.78 kilometres high above the level NN at that time (fig. 11.3). With that even the highest peaks were under water.

Consequentially stands in the legend of Genesis of the ancient Egyptians:

„ They tell us that the earth was completely covered with water and that the earth rose from the water. It is talked about a primeval hill, of which creation took its start, on which the first sunrise and sunset was observed<ii>.

If at that time life only existed in the water, of which we today still can detect the remains in excavations in mountains and plateaus, then it surely wasn't because the evolution had forbidden living on the land. There existed no land! All land was lying under water. But if, looked into the future, the land area increases further at the expense of the surface of the sea, then our earth sometime will dry up, as already other planets before us, e.g. our neighbouring planet Mars.

<i>: Oesterle, O.: Goldene Mitte: Unser einziger Ausweg,
Universal Experten Verlag, Rapperswil (1997), ISBN 3-9520261-9-0

<ii>: Robert Bauval und Graham Hancock: Der Schlussel zur Sphinx, List Verlag (1996), S. 253 and in der Sendung: Die grossen Ratsel VII am 25.5.97 in S 3

a. **Today's earth**:

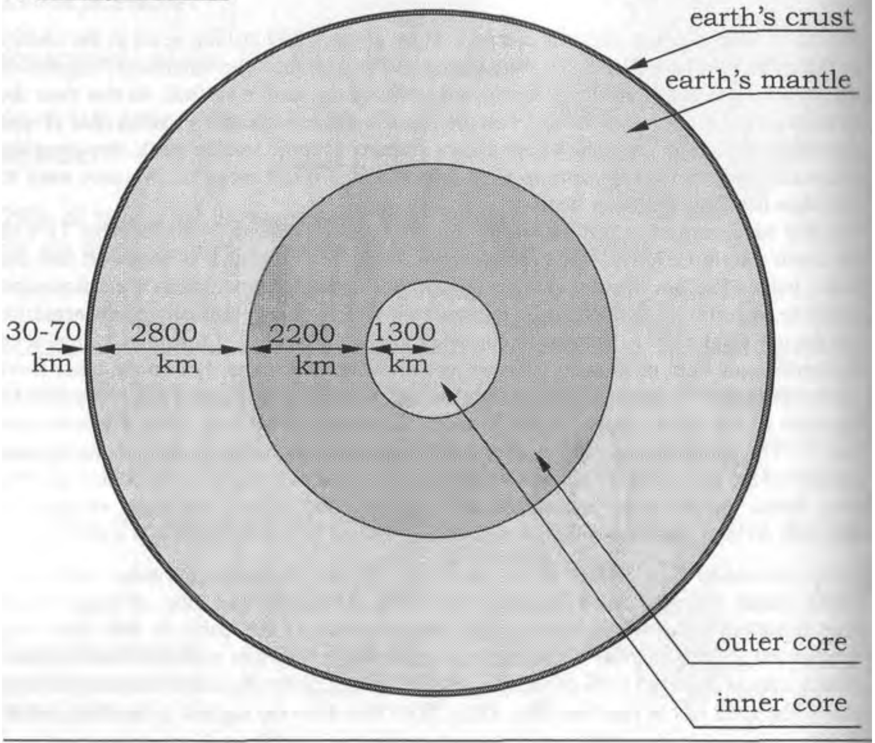
earth's crust
earth's mantle
30-70 km
2800 km
2200 km
1300 km
outer core
inner core

b. The earth **200 million years ago** (super continent: **Pangaea**):

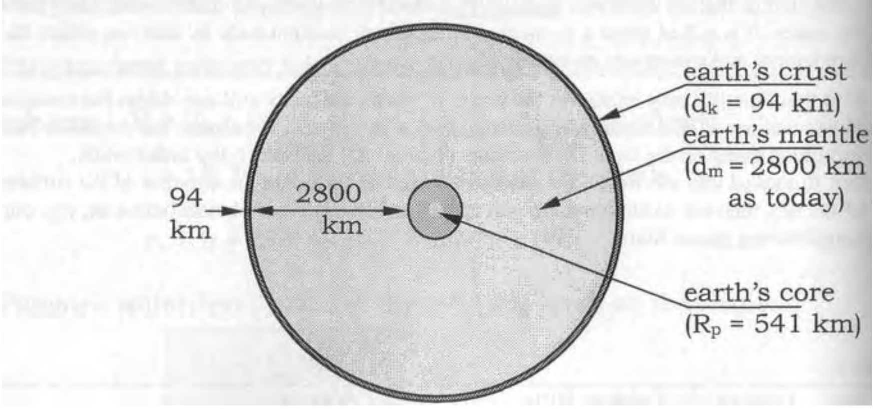
earth's crust
(d_k = 94 km)
earth's mantle
(d_m = 2800 km
as today)
94 km
2800 km
earth's core
(R_p = 541 km)

Fig. 11.4: The inner structure of the earth on an average

11.4 Inner structure of the earth

Next the question is raised: How fast does our earth actually grow? The calculated growth, distributed over the 200 million years, results in a yearly increase in the diameter of the earth of less than 0.1 mm. Carey[i] assumes 0.04 mm per year and Owen[ii] only 0.01 mm per year.

Actually the young earth must have been somewhat bigger than calculated, because as a result of the smaller gravitational acceleration the density of the matter must have been smaller. But this changes nothing to the relations, because the less dense earth was surrounded by likewise less dense water, the water-level nevertheless reached the peaks, as already calculated.

For indicating absolute linear measures and the calculation of the gravitational accleration the respective density should be considered. In most calculations the density is cancelled out, so that as well can be calculated with an unchanged density.

A grave error however lies in the assumption of a linear growth. Hilgenberg assumes an exponential growth[iii] and gives as a reason for the empirical approach of the e-function the „law of organic growth". In order to now not to speculate or to postulate in the same manner, we will derive and found our approach.

If namely the earth grows, then its core of fusion also grows, which causes the growth to take place accelerated, etc. A customer of a bank, who sees his amount of money grow according to such a regularity, will be given information immediately about the growth rate with a compound interest calculation.

But how big is the growing fusion reactor of our planet? According to today's level of knowledge about the structure of the earth the inner core is surrounded by the outer core and that again by the earth's mantle. On top floats the thin, but firm earth's crust, on which we live. The inner core has a radius of nearly 1390 km, the outer core stretches to a radius of 3500 km, whereas the crust is only between 10 and 78 km thick, dependent on the geographical latitude[i4] (fig. 11.4).

<i>: Carey, S. W.: Theories of the Earth and Universe.
Stanford University Press: 1-413, Stanford, California

<ii>: Owen, H. G.: Has the Earth increased in size? - In: New Concepts in Global Tectonics, Texas etc. University Press (1992), p. 289-295, Lubbock

<iii>: Hilgenberg, O. C: Vom wachsenden Erdball, Berlin 1933, Eigenverlag, Seite 31 und 32

<i4>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie
(The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989)

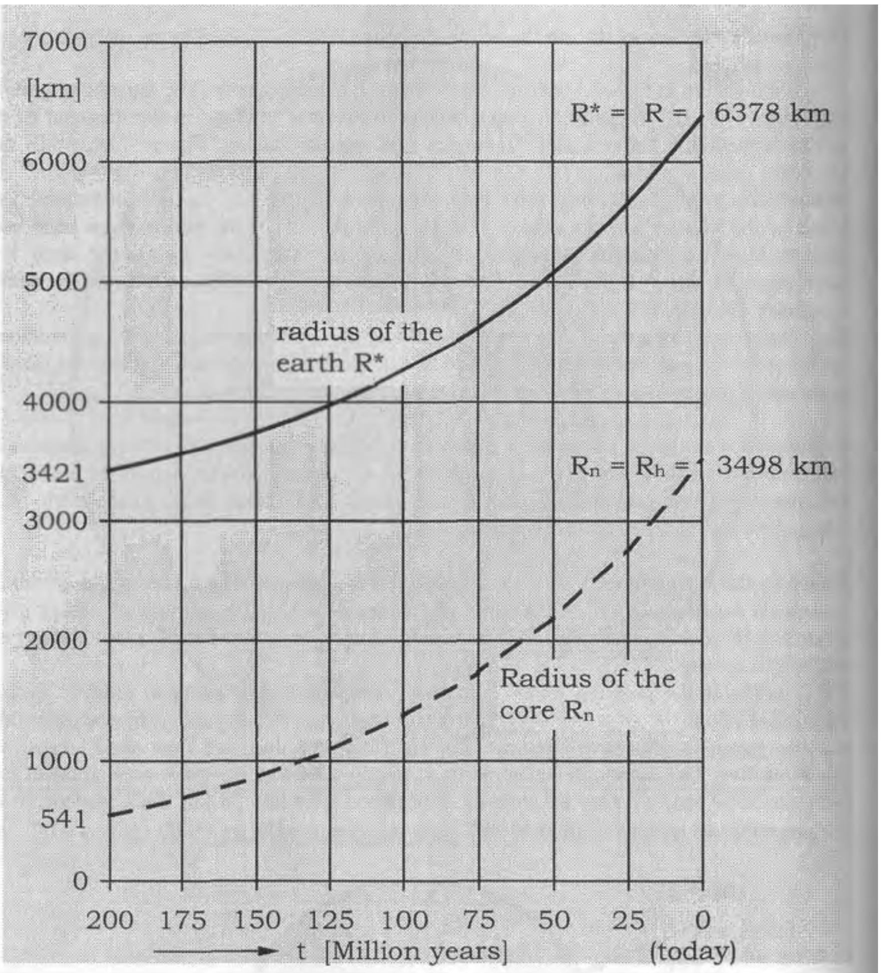

Fig. 11.5: The change in the radius of the earth and of the core in the course of millions of years until today.

calculation: (see fig.11.6) $R^* = 2879{,}5 + 541\cdot q^{n/3}$ [km] (11.7)

t years ago: with: $t = 2\cdot 10^8 - n$ [a] and: $q = 1 + 28\cdot 10^{-9}$

correspondents to the approximation: $R^* = 2880 + (6378 - 2880)\cdot e^{-t/\tau}$ [km] (11.8)

with the time constant: $\tau = 107{,}1$ Mio. Jahre

11.5 Earth's core as a converter of neutrinos

I proceed from the assumption that the conversion of neutrinos and materialization in elementary particles takes place in the inner core. For the conversion no energy at all is used, because the inner energy together with the outer energy of the particle amounts to zero. The neutrinos merely have to be remodelled into another structure and for that they at first have to be slowed down with the help of the oscillating interaction.

During this process of slowing down, as said, no heat is formed because in the case of a mass less particle no energy can be set free in the domain where the classical law of conservation of energy is valid. Only after completion of the process of materialization we are able to detect mass and energy of neutrinos.

But if the oscillating interaction is taken as a basis, the oscillation with opposite phase between particle and earth's core, then contrary to all expectations a cooling down takes place. If the particle has reached its region of destination in the core, then the oscillations are overlapping. Mathematically seen they are added with reversed sign; they thus are subtracted. The result of the mutual compensation is the decrease of the thermal oscillation and the cooling down of the region which was expected.

In addition the formed particles with a mass mutually contract (see part I, chapter 8.3 and 8.4) and in doing so are further cooling down, as we will derive (chapter 12.7 with fig. 12.8). The physical limit of the process of contraction and cooling down is formed by absolute zero, at which no thermal oscillation at all occurs anymore, so that superconduction becomes possible with the result of giant electric currents and magnetic fields, which can be detected even at the earth's surface in damped form, for instance with a compass.

The necessary heat energy is flowing towards the quick-frozen inner core from the outside, principally from the outer core. Here, in the core, from the neutrinos slowed down to the speed of light various elementary particles are formed. Most of them immediately fall apart, to form other configurations. In the end only electrons and protons are preserved, which, as the only stable particles, can't fall apart anymore. These again are trying hard to take the state of an atom, which however needs very much space with the large distance between atomic nucleus and hull. Under the high pressure the enveloping electrons therefore will time and again fall into the nucleus to form neutrons together with the protons.

The neutrons need no atomic hull and can, as is well-known of neutron stars, take an extremely high density. In the case of the earth's core the neutrons however cannot be stabilized. The contraction to a neutron is accompanied by a corresponding drop in pressure, so that the neutron falls apart again. A continual oscillation of size is formed, with which the neutrinos again interact. With that also the high density of the earth's core would be explicable simultaneously.

In earth's outer core the various atoms and isotopes are formed, which in the sum release more energy than they absorb in their fusion processes. Here the fusion oven rages, which supplies the inner core with heat energy.

The formed matter is pushed further to the outside, rolls as a viscous mass through the earth's mantle and collects the surplus radiation and heat from the fusion oven.

With this model of explanation we now can tackle the calculation of the growth of the earth (fig. 11.6).

Growth in volume in analogy to the compound interest calculation:

duration: $n = 200$ million years

starting capital: V_p = Pangaea volume $V_p = (4/3)\cdot\pi\cdot r^3$

final value: V_h = today's volume $V_h = (4/3)\cdot\pi\cdot R^3$

savings bank formula: $V_h = V_p * q^n$ (compound interest) (11.9)

with „interest rate": $q = (V_h/V_p)^{1/n}$ (11.10)

radius of the core of the Pangaea sphere:

$R_p = r - d_m - d_k = 541$ [km] (11.11)

with

Pangaea radius: $r = 3435$ [km]

earth's mantle: $d_m = 2800$ [km]

earth's crust: $dk = 94$ [km] (incl. covering with water).

core radius today: $Rh = R - d_m - d_k = 3500$ [km] (11.12)

earth's crust today: $d_k^* =$ 34-78 [km].

From $V \sim r^3$ the growth factor q is calculated to be

$$\underline{\underline{q}} = \left(\frac{V_h}{V_p}\right)^{\frac{1}{n}} = \left(\frac{R_h}{R_p}\right)^{\frac{3}{n}} = \left(\frac{3500\text{ km}}{541\text{ km}}\right)^{\left(\frac{3}{2\cdot 10^8}\right)} = \underline{\underline{1 + 28\cdot 10^{-9}}} \quad (11.13)$$

core radius after n years: $R_n = R_p * q^{n/3}$ (11.14)

radius of the earth after n years: $R^* = R_n + d_m + d_k^*$ (11.15)

Fig. 11.6: The calculation of the growth rate of the earth

11.6 Speed of growth

200 million years ago in the centre of the globe a core of fusion has formed and taken up its operation, probably under the influence of a cosmic occurrence connected with a high neutrino radiation. As a result the thin crust of the earth was torn apart and the oceanic basins were formed.

If we assume that the might of earth's mantle (with d_m = 2800 km) and crust (with less than 100 km) haven't fundamentally changed, then earth's core at that time had a radius of only 541 km. The „savings bank formula" now only may be applied for the core and only for its volume. On the condition of a constant neutrino density the volume of the core in every year will grow for one order of magnitude, which again depends on the respective volume itself. There results the in fig. 11.5 presented course of the radius of the core and of the earth.

According to our calculation the earth at present grows every year for $915*10^{11}$ tons, which corresponds to an increase in volume of 16500 cubic kilometres and an increase of area of 5.2 square kilometres. The earth momentarily grows for 6.5 cm per year in diameter, from which follows that the perimeter increases pi-fold and a continental drift of 10.2 cm per year is to be expected across both the Atlantic Ocean and the Pacific Ocean. Geologists today actually measure a plate movement of typically 10 cm, at individual points of up to 12 cm per year<i>!

Whoever likes to do handicrafts, can build together a globe of shells by himself. Hilgenberg for that gives a handicraft instruction<ii>. He draws the continents of a globe of today and cuts them out. Doing so, not the coast line of today is authoritative, but that of the edge of the shells, at which the mainland plates are breaking off into the deep sea. He hence also speaks of a sphere of shells and helps the handcrafters with the words: ..Because the paper shells of the sphere of shells owing to their strong curvature are difficult to nestle, we cut slits in the paper, which suitably lie there, where mountain ranges stretch and now can start with the gluing". Doing so it shows that the slits in particular in the case of the Ural and the Himalayan gape far apart, that in reversed direction in the case of the enlargement of the sphere of shells the lifting out of the mountains necessarily had to occur at these points by means of upsetting. Under these circumstances our model concept should be further rendered more precisely, If the change of the curvature of the growing surface of the earth is the cause for the lifting out of the mountains, then the surface of the earth 200 million years ago was structured merely by impact craters and by volcanic cones, then the amount of water may have been correspondingly smaller. The additional water of the oceans of today was collected by the earth either from the cosmos, by crossing the flight path of a comet with a water tail, or by the here discussed idea of a core of fusion in the inside of the earth it has produced the water itself!

How ever such detail aspects may have had an effect, it therefore nevertheless changes nothing to the model concept on the whole. Hilgenberg's globe of shells<ii> in my opinion to still makes more sense as all the models of explanation, as they are spread in today's text books.

<i>: Miller, R.: Driftende Kontinente, Time-Life, Amsterdam 1991, S. 78

<ii>: Hilgenberg, O.C.: Vom wachsenden Erdball, Berlin 1933, Eigenverlag

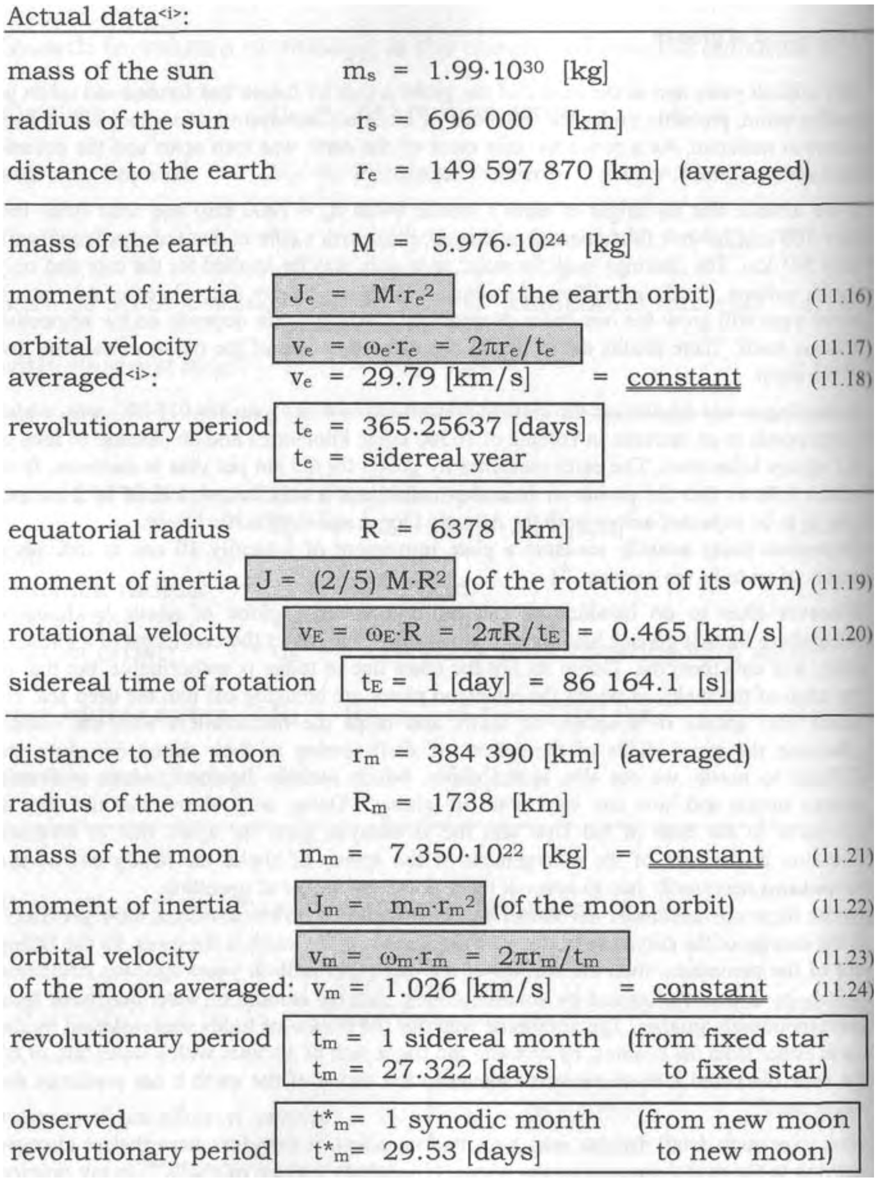

Actual data<i>:

mass of the sun $m_s = 1.99 \cdot 10^{30}$ [kg]

radius of the sun $r_s = 696\,000$ [km]

distance to the earth $r_e = 149\,597\,870$ [km] (averaged)

mass of the earth $M = 5.976 \cdot 10^{24}$ [kg]

moment of inertia $\boxed{J_e = M \cdot r_e^2}$ (of the earth orbit) (11.16)

orbital velocity $\boxed{v_e = \omega_e \cdot r_e = 2\pi r_e / t_e}$ (11.17)

averaged<i>: $v_e = 29.79$ [km/s] = constant (11.18)

revolutionary period $t_e = 365.25637$ [days]
$t_e =$ sidereal year

equatorial radius $R = 6378$ [km]

moment of inertia $\boxed{J = (2/5)\, M \cdot R^2}$ (of the rotation of its own) (11.19)

rotational velocity $\boxed{v_E = \omega_E \cdot R = 2\pi R / t_E}$ = 0.465 [km/s] (11.20)

sidereal time of rotation $\boxed{t_E = 1 \text{ [day]} = 86\,164.1 \text{ [s]}}$

distance to the moon $r_m = 384\,390$ [km] (averaged)

radius of the moon $R_m = 1738$ [km]

mass of the moon $m_m = 7.350 \cdot 10^{22}$ [kg] = constant (11.21)

moment of inertia $\boxed{J_m = m_m \cdot r_m^2}$ (of the moon orbit) (11.22)

orbital velocity $\boxed{v_m = \omega_m \cdot r_m = 2\pi r_m / t_m}$ (11.23)
of the moon averaged: $v_m = 1.026$ [km/s] = constant (11.24)

revolutionary period $t_m =$ 1 sidereal month (from fixed star
$t_m = 27.322$ [days] to fixed star)

observed revolutionary period $t^*_m =$ 1 synodic month (from new moon
$t^*_m = 29.53$ [days] to new moon)

Fig. 11.7: The data of sun, earth and moon of today<i>

<i>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie (The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989)

It is an exciting matter, if one can experience for oneself, how the pieces of the puzzle fit together, how the southern end of America is wound around the Cape of Good Hope and the Falkland islands surface for the east coast of South Africa, in the vicinity of Australia how the Antarctic occupies the Pacific basin as a neighbour of Australia, von South, Middle and North America, run through by the equator, etc.

The corrections to the view of life of Alfred Wegener and the geographical evidence, which Hilgenberg gives, are sound, well founded and even after 65 years still highly actual. The physical statements of the geologists however, for instance the earth in the course of time would rotate faster and faster, cannot be followed.

11.7 Conservation of angular momentum

The question about the rotation of the earth is raised. Is it not changing at all, is it getting faster or slower? In the case of a with time growing earth there is only one possibility: The rotation of its own will decrease and not increase, as Hilgenberg supposes. For that you only need to place yourself on a turntable, to turn and if possible with weights in your hands stretch out your arms sideways, to feel, how the rotation of the table is decreasing. It surely would have done the geologist Hilgenberg good, to leave his desk for a short time for a stroll to the most nearby playground, for the purpose of the described physical experiment with himself.

In this case the law of conservation of momentum authoritatively has an effect in the formulation of the law of conservation of angular momentum for the rotating motion. According to that all angular momenta in the solar system should amount to zero. If we look at the planets which have no moon, then is remarkable that these need an eternity for a revolution around their own axis (Venus for instance needs 243 days). According to the law of conservation of angular momentum our earth owes its rotation of its own primarily the moon.

For reason of this relation we can assume a proportionality between the angular momentum of the moon J_m*w_m and that of the rotation of its own of the earth $J*w_E$ (eq. 11.39, fig. 11.10). They even have to be identical, if the partner of rotation earth and moon are seen as a closed system.

If the earth would be approached as a homogeneous, spinning sphere, the angular momentum at first would be too small for a factor 4.1 (eq. 11.29, fig. 11.9). The law of conservation of angular momentum dictates as a necessary result, that the dense core of the earth must rotate faster than the earth's crust! From the correspondence of orbital angular momentum of the moon on the one hand and the sum of the angular momenta of their own of earth's mantle and earth's core on the other hand results a 31 fold higher angular velocity of the earth's core compared to the rotation of the earth's surface (eq. 11.38, fig. 11.9).

Does our earth owe its geomagnetism this rotation of the core? At least the possibility exists that there exists a causal relation between the rotation of the earth's core and the geomagnetism. We'll further collect arguments and put the question concerning the formation of the geomagnetism under consideration.

Concerning unipolar induction (Faraday): $\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (11.25)
(see part 1, eq. 58, fig. 6.4)

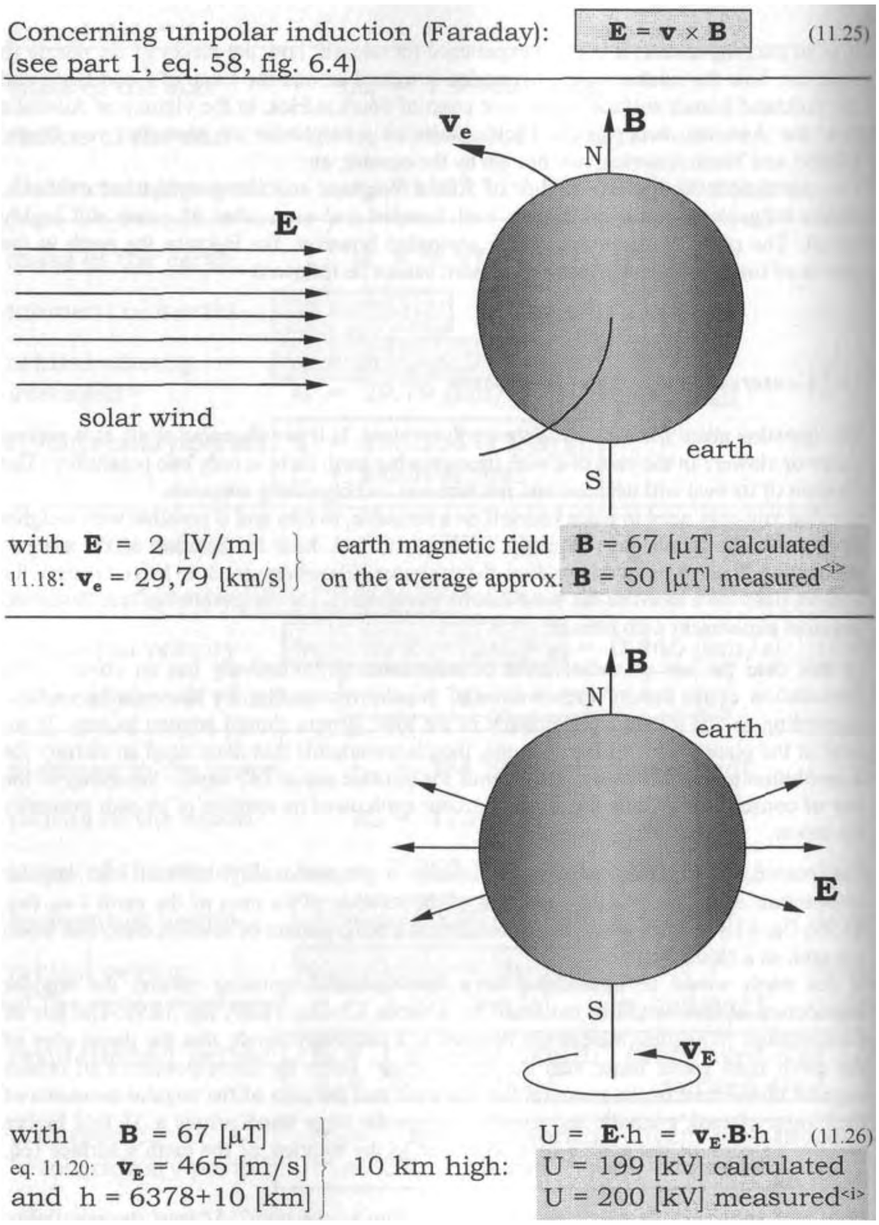

with $\mathbf{E} = 2$ [V/m]
11.18: $\mathbf{v_e} = 29{,}79$ [km/s]
earth magnetic field $\mathbf{B} = 67$ [µT] calculated
on the average approx. $\mathbf{B} = 50$ [µT] measured<i>

with $\mathbf{B} = 67$ [µT]
eq. 11.20: $\mathbf{v_E} = 465$ [m/s]
and h = 6378+10 [km]
10 km high: $U = \mathbf{E} \cdot h = \mathbf{v_E} \cdot \mathbf{B} \cdot h$ (11.26)
U = 199 [kV] calculated
U = 200 [kV] measured<i>

Fig. 11.8: The magnetic and the electrostatic field of the earth
as a result of an unipolar induction

<i>: Measurement values in chap. 2.8, e.g. reference entry <i> (Prof. Dr. L. Konig)

11.8 Set of difficulties concerning the change of magnetization

Now investigations of the polarity of the magnetization of rocks have resulted in the fact that in irregular intervals, on an average every 500,000 years, a change of polarity of the geomagnetism has occurred and no one knows why.

If the rotation of the core should produce the magnetism, then it in any case also is responsible for the process of change of magnetization. That purely theoretically is possible in two ways: Either the core tilts out of its plane of rotation for 180° and rotates suddenly in the opposite direction or earth's mantle together with earth's crust, on which we live, is turned upside down.

It surely isn't my intention to produce any panic, but from a physical point of view only the second case is possible. Usually not the tail wags with the dog, but vice versa, the dog with the tail. The high rotational velocity of the earth's core exerts an extremely stabilizing influence. After all its angular momentum is 4.85 times larger than that of the mantle (eq. 11.33, fig. 11.9). Therefore the rotation of the core and the direction of earth's magnetic field always are preserved seen from the sun.

I further proceed from the assumption, that an electrostatic field arises from the sun accompanied by a particle flux, the solar wind, through which the earth flies through in a perpendicular direction. According to the Faraday law of induction $E = v \times B$ (eq. 11.25)
it experiences, as a result of the unipolar induction, a magnetic field which stands perpendicular on the ecliptic and thus dictates the orientation of earth's magnetic field. With that also the direction of rotation would be determined. The core thus by no means can tilt!

On the condition that the core doesn't rotate completely frictionless, the earth's mantle will in the case of the same direction of rotation be accelerated, in the case of unequal direction, after a changing of polarity, it will be slowed down again. If the process, for reasons of conservation of angular momentum, takes place alternatingly with a certain regularity, then the inhabitants of the earth for every changing of polarity might experience, how the North Pole in the shortest time turns over the equator to the South Pole, how the sun sets at the point, where it before had risen. As said, after a changing of polarity our earth is standing on its head! In the Bible corresponding clues are found<ii>. It can be taken from the media that one reckons with such a changing of polarity in the foreseeable future. When and if it actually takes place, stands however in the stars.

If magnetism thus is produced by the rotation of the core, or by electric currents in the earth's core as a result of the superconduction or by both, then the earth is aligned in the field of the sun like the magnetic needle in a compass. If we take as a cause a solar wind with an electric field strength of just 2 V/m, then this would not only determine the direction of earth's magnetic field, but also the order of magnitude, and that at present lies at 50 uT averaged over time (chap. 2.8).

With the same mathematical relation an electrostatic field around the earth of 199 V/m results as a result of the rotation of the earth. That fairly exact corresponds to the measured values!<i> Quite obviously all is related to each other. In the question for the „how" at this point consciously no definite answer is strived for. It rather should be discussed and worked to together in the seminar.

<ii>: The Bible, OT, The 2nd book of kings 20,9-11 (king Hezekiah) and Joshua 10,12-14, literally cited on the page after the next page (chap. 11.9), note<ii>

Angular momentum $J_m \cdot \omega_m$ of the orbiting moon (with eq. 11.22 + 11.23):

$$J_m \cdot \omega_m = m_m \cdot r_m^2 \cdot \omega_m = 29*10^{33} \; [kgm^2/s] \qquad (11.27)$$

Angular momentum $J \cdot \omega_E$ of the rotation of the earth (eq. 11.19 + 11.20, same direction of turning):

(theoretically) $J \cdot \omega_E = (2/5)\, M \cdot R^2 \cdot \omega_E = 7*10^{33} \; [kgm^2/s]$ (11.28)

$$\frac{\text{Orbital momentum moon}}{\text{Angular momentum earth}} = \frac{J_m \cdot \omega_m}{J \cdot \omega_E} = \frac{29}{7} = 4{,}1 \qquad (11.29)$$

Angular momentum $J_M \cdot \omega_E$ of earth's mantle and crust (R_h = core radius)

$$J_M \cdot \omega_E = (J - J_K) \cdot \omega_E = (2/5) \cdot (MR^2 - M_K R_h^2) \cdot \omega_E = 6*10^{33} \; [kgm^2/s] \qquad (11.30)$$

$$\text{conservation of angular momentum:} \quad J_m \cdot \omega_m = J_K \cdot \omega_K - J_M \cdot \omega_E \qquad (11.31)$$

Angular momentum $J_K \cdot \omega_K$ of the earth's core

$$J_K \cdot \omega_K = J_m \cdot \omega_m + J_M \cdot \omega_E = 35*10^{33} \; [kgm^2/s] \qquad (11.32)$$

$$\frac{\text{Angular momentum core}}{\text{Angular momentum mantle}} = \frac{J_K \cdot \omega_K}{J_M \cdot \omega_E} = \frac{35}{5} = 7 \qquad (11.33)$$

$$J_K \cdot \omega_K = (2/5)\, M_K \cdot R_h^2 \cdot \omega_K \qquad (11.34)$$

With the mass of earth's core $M_K = \rho_K \cdot V_K = \rho_K \cdot (4/3) \cdot \pi \cdot R_h^3$ (11.35)

and the averaged core density $ro_k = 12000 \; kg/m^3$

the angular velocity

of earth's core amounts to: $\omega_K = 3{,}3*10^3 \; [s^{-1}]$ (11.36)

and that at the earth's surface: $\omega_E = v_E/R = 0{,}073*10^{-3} \; [s^{-1}]$. (11.37)

$$\frac{\omega_K}{\omega_E} = \frac{3{,}3}{0{,}073} = 45 \qquad (11.38)$$

Fig. 11.9: Calculations concerning the conservation of angular momentum and the rotation of earth's core (One turn of the earth's core lasts 32 min. It results contrary to the earth rotation. Of the surface of the earth a period duration of 31 min should be observable.)

11.9 The weakly damped moon

The increasing angular velocity in the direction of the centre of the earth surely has something to do with the set of difficulties concerning the conservation of the spherical structure, comparable to the reason for the particle spin (chapter 6.13). The increasingly missing radial component of the gravitational field has to be compensated by an additional field produced by the rotation (according to part 1, equation 62, fig. 6.5).

The moon apparently doesn't know this set of difficulties. We can assume a constant mass for it (eq. 11.21). On the one hand is the moon smaller than the earth 200 million years ago, as it started to grow. On the other hand are doubts concerning the existence of an active core of the moon legitimate. In the Apollo-15 mission seismic gauges had been installed at the ground of the moon and at the start from the moon the produced seismic waves had been recorded<i>. The surprisingly small damping as the lunar module fell back onto the ground of the moon more likely allow the conclusion that the moon is hollow inside!
Also the small density of the moon points in the same direction, and finally no clues at all can be seen on the surface of the moon, which would point to a growth of the moon.

Apart from the mass also the orbital velocity of the moon is taken constant, what surely is true, as long as nobody and nothing drives the moon extra (eq. 11.24). The analysis of the law of conservation of angular momentum provides the in fig. 11.10 derived provisional result (eq. 11.41).

<i>: Moonquakes ever more mysteriously: „The by Apollo 12 triggered moonquake by „bombardment" of the surface of the moon with the clapped-out Lunar Module „Intrepid" at thursday evening puts scientists for bigger and bigger mysteries. An exact analysis of the measurement data now resulted in the by the astronauts Conrad and Bean installed „seismic station" on the moon ... to have recorded and to have sent to the earth 55 minutes long. ... The seismologist Dr. Gary Latham spoke of „important information about the structure of the moon" and meant, one now can „throw away the text book". One had reckoned that the impact of the Lunar Module on the moon would trigger a quake of at best some minutes", (translated) Frankfurter Rundschau vom 22.11.1969

---

concerning reference entry <ii>, chap. 11.8:

<ii>: The Bible, OT, Joshua 10,12-14: There stands in the „book of Jasher ": So the sun stood still in the midst of heaven, and hasted not to go down about a whole day. And there was no day like that before it or after it, ...

The 2nd book of kings 20,9-11: king Hezekiah: This sign thou ... let the shadow return backward ten degrees. And Isaiah the prophet cried unto the LORD: and he brought the shadow ten degrees backward, by which it had gone down in the dial of Ahaz.

For the proportionality of the angular momenta of the earth:

$$J \cdot \omega_E \sim J_M \cdot \omega_E \sim J_K \cdot \omega_K$$

With: $$J \cdot \omega_E = (2/5)\, M{\cdot}R^2 \cdot 2\pi/t_E \quad (11.28)$$

and the orbital momentum of the moon:

$$J_m \cdot \omega_m = m_m{\cdot}r_m{}^2 \cdot v_m/r_m \quad (11.27)$$

is valid the

conservation of angular momentum: $J \cdot \omega_E \sim J_m \cdot \omega_m$ (11.39)

or: $$M{\cdot}R^2/t_E \sim m_m{\cdot}r_m{\cdot}v_m \quad (11.40)$$

with $\underline{m_m = \text{constant}}$ (11.21)
and $\underline{v_m = \text{constant}}$ (11.24): $\boxed{M{\cdot}R^2 \sim t_E{\cdot}r_m}$ (11.41)

---

centrifugal force = gravitational force

$$\frac{m_m{\cdot}v_m{}^2}{r_m} = \frac{G{\cdot}M{\cdot}m_m}{r_m{}^2} \quad (11.42)$$

or with eq. 11.22: $$v_m{}^2 = G{\cdot}M/r_m = (2\pi r_m/t_m)^2 \quad (11.43)$$

---

if $\underline{M = \text{constant}}$: $\boxed{r_m{}^3 \sim t_m{}^2}$ (10.1)

Kepler's 3rd law.

Here $\underline{M \neq \text{constant}}$, instead $\underline{v_m = \text{constant}}$ (11.23) and therefore follows from eq. 11.43:

(month) $\boxed{t_m \sim r_m \sim M}$ (11.44)

with eq. 11.41:

(day) $\boxed{t_E \sim R^2}$ (11.45)

and for an analogous
derivation: (year) $\boxed{t_e \sim r_e \sim m_s}$ (11.46)
with $\underline{v_e = \text{constant}}$

Fig. 11.10: Calculation of dynamic celestial mechanics

11.10 Calculation of celestial mechanics

We still need a further relation and try the balance of forces between the centrifugal force and the gravitational force. Both the to the outside directed centrifugal force and the to the inside directed gravitational force depend on the mass of the moving body, here the mass of the moon m_m, so that it is cancelled from equation 11.42.
As determining mass only that of the earth is left and that is taken constant in accordance with the knowledge of text books. The result of this assumption is Kepler's 3^{rd} law (eq. 10.1). But beware, it here merely concerns a momentary picture! In the course of time, according to the derivation, the mass of the earth increases so that this assumption is untenable.

Instead, as already said, the average orbital velocity of the moon, one kilometre per second, can be taken constant (eq. 11.24). Since the orbital velocity is calculated from the proportion of the perimeter of the circular orbit $2 \cdot \pi \cdot r_m$ with regard to the revolutionary period of a month (11.43, left page), and at the same time from the balance of forces a dependency of the mass of the earth M and the radius of the orbit of the moon follows (11.43, right page), the interesting relation (11.44) results:
Here the mass M, the radius r_m and the revolutionary period t_m stand in a direct proportionality to each other, and that means: if the mass of the earth increases in the course of time, then the moon will go away from us to the same extent, then also every month will get correspondingly longer.

Clarified with numeric values follows from a growth of the earth for $915*10^{14}$ kg per year a going away of the moon for yearly 5.88 meters (fig. 11.11). The going away of our satellite could be confirmed by means of measurements with laser reflectors, which had been put up on the moon by Apollo astronauts, according to dpa message<i>.

In addition every month lengthens for 3 milli seconds. That is valid for the sidereal month (= 27.322 days), for which a fixed star serves as a reference point for the measurement of a revolution, as well as for the synodic month, as it is observed from the earth from new moon till new moon (= 29.53 days). The synodic month today is longer than four weeks. But 3.7 million years ago it once actually lasted exactly 28 days, as we can calculate easily (fig. 11.11).

But now also the length of a day is changing. If we insert the proportionality (11.44) from the balance of forces in that of the conservation of angular momentum (11.41), then it is shown that a day depends quadratic on the radius of the earth (11.45), that a lengthening of every day for $4.5*10^{-9}$ s can be determined with the help of the growth curve of the earth (fig. 11.5). This is really very small, but 200 million years ago the day had just 19 hours, if extrapolated to 900 million years only 18.04 hours! The result of the american geologists around Dr. Charles Philip Sonett of the University of Arizona in Tucson also is 18 hours in their analysis of correspondingly old sediment formations, as the US science magazine Science has reported<i>. The correspondence of this measurement with our calculation without doubt has force of evidence!

<i>: 900 million years ago a day had 18 hours, Washington (dpa) 1997

Going away of the moon per year with $\boxed{M \sim r_m}$ (according to eq. 11.44): for

$\Delta r_m = r_m \cdot \Delta M/M = 384390[\text{km}] \cdot 9.15 \cdot 10^{16}[\text{kg/a}] \,/\, 5.976 \cdot 10^{24}[\text{kg}] = \underline{5.886\ [\text{m/a}]}$

lengthening of month per year with $\boxed{M \sim t_m}$ (11.44) (sidereal): for

$\Delta t_m = t_m \cdot \Delta M/M = 27.322\ [\text{days}] \cdot 86164\ [\text{s/day}] \cdot \Delta M/M = \underline{36\ [\text{ms/a}]}$

synodic month (from new moon till new moon) of 28 days:

$\Delta t_m = 29.53\ [\text{days}] - 28\ [\text{days}] = 1.53\ [\text{days}] \equiv 132 \cdot 10^6\ [\text{ms}]$

linearly calculated $132 \cdot 10^6\ [\text{ms}]/36\ [\text{ms/a}] = \underline{3.6674\ \text{mio. years ago}}$.

Day length 200 mio. years ago, with $\boxed{t_E \sim R^2}$ (11.45) amounted to t* hours

shortening: $\Delta t_E/t_E = \Delta R^2/R^2 = (R-R^*)^2/R^2 = (6378-3420)^2/6378^2 = 21.5\%$

$\Delta t_E = 0.215 \cdot 24\ [\text{h}] = 5.16\ [\text{h}]$ and $\underline{t^*} = 24\ [\text{h}] - \Delta t_E = \underline{18.84\ [\text{h}]}$

Day length 900 mio. years ago, (R^*=3200 km) amounts to 18.04 hours

present lengthening per day amounts to: $\Delta t_E/t_E = 4.5 \cdot 10^{-9}\ [\text{s/day}]$

but: the year increases 130,000 times faster than the individual day!

according to the measurement of Aristarchos 2300 years<ii> ago

is $\Delta t_e = 365.25637\ [\text{days}] - 365.25062\ [\text{days}] = 0.00575\ [\text{days}]$

and $\underline{\Delta t_e/t_e} = 0.00575 \cdot 24 \cdot 60 \cdot 60 \cdot 1000 \,/\, 365.25637 \cdot 2300 = \underline{0.59\ [\text{ms/day}]}$

going away of the earth from the sun per year, with $\boxed{t_e \sim r_e}$ (acc. to eq. 11.46)

for: $\Delta r_e/t = r_e \cdot \Delta t_e/t_e \cdot t = 149.6 \cdot 10^6\ [\text{km}] \cdot \Delta t_e/t_e \cdot t = \underline{1.024\ [\text{km/a}]}$

growth of the sun per year/second, with $\boxed{t_e \sim m_s}$ (according to eq. 11.46)

for:

$\Delta m_s/t = m_s \cdot \Delta t_e/t_e \cdot t = 1{,}99 \cdot 10^{27}\ \text{t} \cdot \Delta t_e/t_e \cdot t = \underline{1.36 \cdot 10^{19}\ [\text{t/year}]}$

resp. in a second for: $= \underline{4.3 \cdot 10^{11}\ [\text{t/s}]}$

and at the same time a loss due to radiation of: $\underline{5 \cdot 10^{6}\ [\text{t/s}]}$

Fig. 11.11: Figures according to analysis of some examples concerning dynamic celestial mechanics

11.11 The biblical age

We therefore owe the growing diameter of the earth that every day gets longer and longer (eq. 11.45) and from the going away of the moon a lengthening of the month results (eq. 11.44). Not only the months and the length of the days increase, but also the whole year. A corresponding derivation for the revolution of the earth around the sun, as it has been carried out for the revolution of the moon around the earth (eq. 11.44), delivers analogous results (eq. 11.46): The sun determines the length of the year and the distances to the planets. To the extent, to which the sun grows, the solar system increases in extension. Also the distance to the earth increases proportionally with the mass of the sun. By the going away from the sun the temperature on the planets however not necessarily decreases, because at the same time the radiation intensity of the growing central star increases. According to measurements it has increased for 30% since the formation of the solar system<i>.

If both the rotation of the earth and the revolution around the sun get slower, then by all means is conceivable, that the number of days per year approximately stays the same and mankind nevertheless gets less old. If we take the 2300 years old writing of the Greek Aristarchos about the sizes and distances of the sun and the moon<ii>. He determined, assuming a heliocentric view of life, the year to be 365.25062 days. Because the correction taken by Aristarchos concerned even the fifth place after the comma, we must assume that correspondingly precise gauges were available in Alexandria already 310 BC.
The today measured sidereal year with 365.25637 days has lengthened for whole 0.00575 days. From this follows that the year increases considerably faster than the individual day.
It of course would be nice, if we could calculate the lengthening of the year, but unfortunately no data about the growth of the sun are available.
If we proceed from the measurement of Aristarchos, without being able to verify or reproduce its reliability, then from that would result a going away of the earth from the sun for yearly one kilometre, then the sun should grow for $4.3*10^{11}$ tons per second. In any case the sun materializes mass considerably faster, as it loses mass in the same period as radiation equivalent, and that surely is correct (fig. 11.11). Today one generally assumes that „since its formation the earth has gone away from the sun for in total 30,000 km"<i>.

But if in the past the year consisted of less days and every day moreover was shorter, if therefore the biological life time was divided in shorter periods, then mankind could get older, then obtaining a biblical age possibly by no means was unrealistic. If Adam still should have got 930 years old, according to the 1st book of Moses, then his lifetime already must have been quite long ago. While Abraham still did get 175 years old, no successor of him has reached his age anymore. In the Bible it is said: „And the LORD said, My spirit shall not always strive with man, for that he also is flesh: yet his days shall he an hundred and twenty years "<iii>. Today even the limit given by the LORD isn't reached anymore!

<i>: Is the sun loosing her Gravitation? Illustrierte Wissenschaft, Nr.l, 1995
<ii>: Hermann Wild: Technologien von gestern, Chancen fur morgen,
Jupiter-Verlag Bern (1996), ISBN 3-906571-13-0, Seite 22
<iii>: The Bible, King James Version, Genesis 6.3

The balance of forces (equation 11.42 generalized):

$$\begin{array}{ccc} \text{centrifugal force} & = & \text{gravitational force} \\ \frac{M \cdot v_k^2}{r} & = & \frac{G \cdot M \cdot m}{r^2} \end{array} \quad (12.1)$$

results in the cosmic velocity v_k:

$$\boxed{v_k^2 = G \cdot m / r} \quad (12.2)$$

independent of the mass M of the satellite or planet!

For	$v < v_k$	falling back into the central mass m
For	$v > v_k$	taking the leave into space
For	$v = v_k$	stationary orbit

written down for the n^th^ planet in the solar system:

with $r(n) = r_n$ average orbital radius of the planet

and $v(n) = v_n$ average orbital velocity

as well as: m_s mass of the sun

$$\boxed{v_n = \sqrt{G \cdot m_s / r_n}} \quad (12.3)$$

or:

$$\boxed{v_n = G \cdot m_s / v_n \cdot r_n} \quad (12.4)$$

Fig. 12.1: The first cosmic velocity v_k

12. New cosmology

Astronomy still hasn't satisfactorily solved the question of the formation of the solar system. It thereby however concerns a central problem, because it includes the origin of the earth. But as long as we not yet have understood the relations in our nearest environment, the processes, which we observe with giant telescopes in the depths of pace, will remain a book with seven seals, will lose models of explanation concerning the Big Bang and concerning the so-called unavoidable heat exitus every reliability.

12.1 Concerning the formation of our solar system

2-8-1972 the observatories of the sun reported an unusual high solar activity and 6 days later a slowing down of the rotation of the earth occurred, which was recorded as the lengthening of a day for more than 10 milliseconds. This effect hardly can be explained by the tidal friction alone already just concerning the energy balance. Instead this observation makes clear two things to mankind. On the one hand, how much our earth depends on the solar processes and on the other hand, that the changes by all means can occur not continuously, but periodically and if need be even sporadic.

If, as a result of the conservation of angular momentum, the sun determines the orbital velocity of the earth, if it dictates the rotation of the earth by its neutrino activity and the growth of the earth and if the earth in the same manner determines the orbit of the moon, then it would be obvious that the moon originally has been a part of the earth and this in turn sometime a part of the sun. As it came off, the necessary angular momentum then has been passed on proportionately to the celestial companions, with which the cause for evolution and rotation of their own would be clarified. In addition the moon goes away from the earth and the earth again from the sun, so that looking back it by all means would be obvious, if they once had belonged together.

If we assume that cosmic dust particles meet and accumulate, then all collected rubble contributes to the rotation of its own of the forming celestial body. The more matter finds together, the larger its force of attraction gets, the faster it will grow like a celestial vacuum cleaner. In the course of time this process however is slowed down again and eventually comes to a standstill, because as matter condenses, volume and spherical radius decrease and the rotation of its own increases to the corresponding extent. The celestial body rotates faster and faster and reaches at its surface the cosmic velocity vk, which is given by the mass m and the radius r of the star (eq. 12.2). Now the centrifugal force has reached an order of magnitude, for which the celestial body hurls exactly as much matter into space, as it on the other hand collects by its gravitational effect.

Our sun was lucky to have been supplied with neutrinos in the range of influence of our galaxy. It went in resonance and started to grow, this time from out of the inside. It however could not yet shine, because a crust had formed on its surface around its core of fusion and its mantle, a crust on which permanently was falling cosmic matter from the outside. The increasing viscosity of the sun becoming compressed caused a slowing down of the core of the sun and the corresponding acceleration of the mantle and crust.

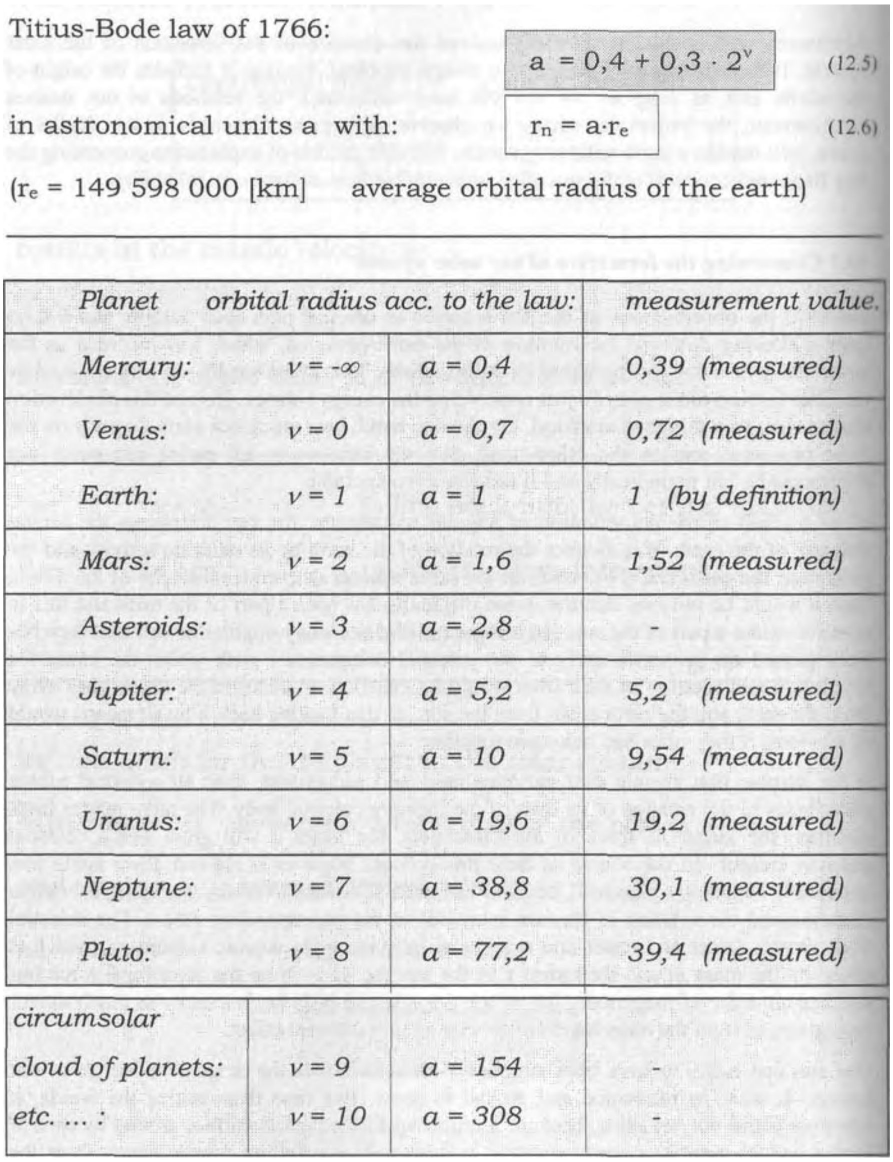

Titius-Bode law of 1766:

$$\boxed{a = 0{,}4 + 0{,}3 \cdot 2^{\nu}} \qquad (12.5)$$

in astronomical units a with:

$$\boxed{r_n = a \cdot r_e} \qquad (12.6)$$

(r_e = 149 598 000 [km] average orbital radius of the earth)

Planet	orbital radius acc. to the law:		measurement value.
Mercury:	$\nu = -\infty$	$a = 0{,}4$	0,39 (measured)
Venus:	$\nu = 0$	$a = 0{,}7$	0,72 (measured)
Earth:	$\nu = 1$	$a = 1$	1 (by definition)
Mars:	$\nu = 2$	$a = 1{,}6$	1,52 (measured)
Asteroids:	$\nu = 3$	$a = 2{,}8$	-
Jupiter:	$\nu = 4$	$a = 5{,}2$	5,2 (measured)
Saturn:	$\nu = 5$	$a = 10$	9,54 (measured)
Uranus:	$\nu = 6$	$a = 19{,}6$	19,2 (measured)
Neptune:	$\nu = 7$	$a = 38{,}8$	30,1 (measured)
Pluto:	$\nu = 8$	$a = 77{,}2$	39,4 (measured)
circumsolar cloud of planets:	$\nu = 9$	$a = 154$	-
etc. ... :	$\nu = 10$	$a = 308$	-

Table 12.2: The Titius-Bode series of the planets (the theoretical values compared to the measurement values)

Sometime our like mad spinning sun had increased that much, that the crust came off like the „tread of an old car tyre" and was catapulted into space. The repulsion was achieved by the centrifugal force exceeding the gravitation for the force of cohesion, which at the moment of the separation of crust regions suddenly tears off. Like snowballs the planets lulled off the sun's surface in this process and were hurled away. Their velocity of rotation at this time was identical to the cosmic velocity of the sun v_k and with that for many a planet large enough to produce its own satellites out of its own surface, which perhaps was not yet completely ideally spherical.

With every planet, which the sun gave birth to, it gave away a part of its own angular momentum to its child as orbital momentum. Only from this time on the sun reduced its angular velocity steadily until this very day. By losing the crust it also started to shine openly. Thus at least could our solar system have been formed.

12.2 The birth of the planets

From this observation various consequences result. If stars are observed, which are rotating very fast, then they either are very young, or they have no planets. Stars, which compared to our sun are rotating less fast, have given away the angular momentum to their planets and such, which hardly are rotating, have their planets already sent away into space. But if the last planet leaves its solar system and the sun stops to rotate, then the sun, which meanwhile has grown to a red giant, without a, the spherical form stabilizing centrifugal force, will collapse. The supernova is the death of a star and thereby neutrinos are set free, the material for new life.

After the coming off of the planets these first clear free their flight paths, by together with the sun collecting the flying around matter. Even whole planets thus can collide, are slowed down and form bigger units. Finally only some few planets in very particular orbits are left. Their average distance to the sun obeys in an until now completely inexplicable manner the Titius-Bode series (equation 12.5, table 12.2).
In the case of the by the german scientist Titius 1766 formulated regularity it concerns a rule of thumb founded purely on experience. But it is remarkable, that the planet Uranus could be predicted (Bode 1772) and after systematic searching also be found (Herschel 1781) with it.

The most distant planets Neptune and Pluto however don't obey the law. Their distance should be very much larger. But this circumstance we already had given reasons for with the fact, that for them the resonating interaction already overlaps and exceeds the gravitation in its effect (chap. 10.1 note<ii> and chap. 10.5).

The arbitrary seeming series of numbers 0, 3, 6, 12, 24, 48, 96, 192, 384, ... with a respective doubling of the value, starting with the 3, the addition of 4 and the following division by 10 at first are nothing but pure acrobatics of numbers, which now really has nothing to do with physics (table 12.2). A physical background can be supposed however because of the tried and tested applicability and that should be fathomed.

Orbital angular momenta $J\cdot\omega$ of the n^{th} planet
with: $J = m\cdot r_n^2$ and $\omega = v_n/r_n$: $\boxed{J\cdot\omega = m\cdot r_n\cdot v_n}$ (12.7)

By using	equation 12.3:	equation 12.4:	
amounts the orbital angular momentum to: $J\cdot\omega$	$= m\cdot\sqrt{G\cdot m_s\cdot r_n}$	$= G\cdot m\cdot m_s/v_n$	(12.8)
dependent	on r_n	on v_n	
change of the angular momentum (1st derivation) with the ordinal number n			
$\frac{d(J\cdot\omega)}{dn}$	$= m\cdot\sqrt{G\cdot m_s}\,\frac{d(\sqrt{r_n})}{dn}$	$= G\cdot m\cdot m_s\,\frac{d(v_n^{-1})}{dn}$	(12.9)
$\frac{d(J\cdot\omega)}{dn}$	$= \frac{m\cdot\sqrt{G\cdot m_s}}{2\cdot\sqrt{r_n}}\cdot\frac{dr_n}{dn}$	$= -\frac{G\cdot m\cdot m_s}{v_n^2}\cdot\frac{dv_n}{dn}$	(12.10)
with eq. 12.8: $\frac{d(J\cdot\omega)}{dn}$	$= \frac{J\cdot\omega}{2\cdot r_n}\cdot\frac{dr_n}{dn}$	$= -\frac{J\cdot\omega}{v_n}\cdot\frac{dv_n}{dn}$	(12.11)
By comparison of the left and the right solution (of eq. 12.11), after introduction of the constant: $\frac{1}{N}$	$= \frac{1}{r_n}\cdot\frac{dr_n}{dn}$	$= -\frac{2}{v_n}\cdot\frac{dv_n}{dn}$	(12.12)
the differential equations read:	$\boxed{\frac{r_n}{N} = \frac{dr_n}{dn}}$	$\boxed{\frac{v_n}{2\cdot N} = -\frac{dv_n}{dn}}$	(12.13)
and the general solutions:	$r(n) = r_n = r_o\cdot e^{\lambda\cdot n}$	$v(n) = v_n = v_o\cdot e^{\lambda\cdot n}$.	(12.14)
The 1st derivation:	$\frac{dr_n}{dn} = r_o\cdot\lambda\cdot e^{\lambda\cdot n}$	$\frac{dv_n}{dn} = v_o\cdot\lambda\cdot e^{\lambda\cdot n}$	(12.15)
and the comparison with eq. 12.13 and eq. 12.14:	$\frac{dr_n}{dn} = \frac{r_o}{N}e^{\lambda\cdot n}$	$\frac{dv_n}{dn} = \frac{-v_o}{2N}e^{\lambda\cdot n}$	(12.16)
determine the coefficient λ:	$\lambda = 1/N$	$\lambda = -1/2N$.	(12.17)
The result is for a planet in the n^{th} orbit:	for the average orbital radius $r(n) = r_n$ $r(n) = r_o\cdot e^{n/N}$	for the average orbital velocity $v(n) = v_n$ $v(n) = v_o\cdot e^{-n/2N}$	
	(12.18)	(12.19)	

Fig. 12.3: Calculation of the distances and orbital velocities of the planets

12.3 The derivation of the Titius-Bode law

Even if Bohr's atomic model should be wrong, it gladly is compared to the system of the planets. The radii of the electron orbits mathematically result as eigenvalue solutions of the Schrodinger equation, and that we have derived from the fundamental field equation (chap. 5.1, eq. 15) (chap. 5.5 - 5.9). The orbital radii are calculated in increasing order in accordance with the sequence of whole numbers with n = 1, 2, 3, 4, ...

From the same sequential regularity of the planetary orbits can be derived, that they also obey the eigenvalues of the same fundamental field equation, which isn't called a world equation in vain. The quantitative distances from the sun are determined by the size of the sun: if the sun increases, then also all distances increase to the same extent.

Now result identical distances between the orbits from Bohr's model, whereas this distance in the case of the planets with increasing distance from the sun gets larger. The reason for this unevenly grading can be calculated just like that (fig. 12.3).

We arbitrarily pick a planet, which occupies the orbit n, where n again represents the sequence of whole numbers (n = 0, 1, 2, 3, ...). If the orbit of this n^{th} planet changes, then also its distance to the sun $r_n = r(n)$, its orbital velocity $v_n = v(n)$ and the orbital momentum $J \cdot \omega$ (12.7) are changed.
We try (as in equation 11.42, fig. 11.10) the balance of forces between centrifugal force and gravitational force (eq. 12.1, fig. 12.1) and solve equation 12.2 for the orbital velocity. The orbital angular momentum of the planet written down once in its dependency on r_n and in the right column next to it on v_n (eq. 12.8, fig. 12.3), is derived for the orbital ordinal number n, to record the change in angular momentum (equations 12.9 to 12.11).

In fig. 12.3 the paths to the solution for both cases are given. For the average orbital radius r(n) as well as for the average orbital velocity v(n) of the planet an exponential course (eq. 12.18 and 12.19) and in logarithmic representation a straight line (fig. 12.4) results.

Even the orbits of the distant planets Neptune and Pluto, for which the Titius-Bode series fails, now are correctly recorded, so that with good cause can be claimed, to correctly have derived the regularity of the distances of the planets physically and mathematically.

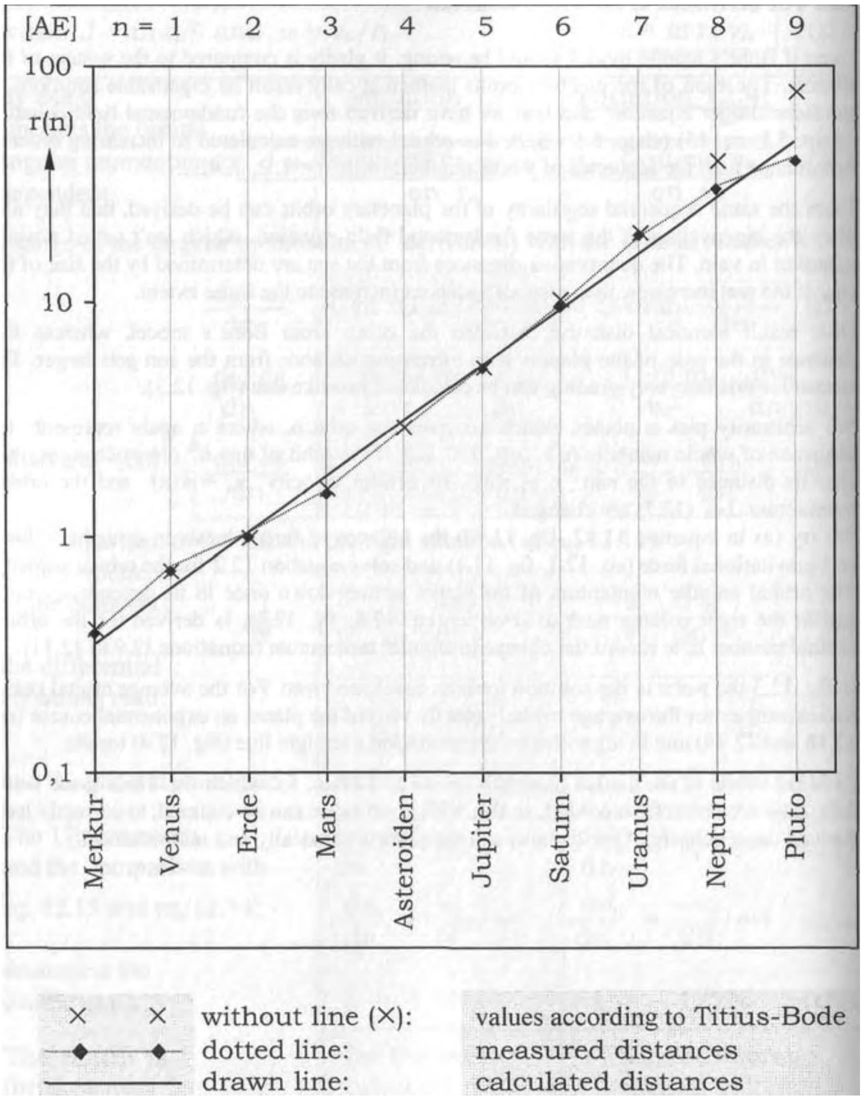

Fig. 12.4: Representation of the distances of the planets

result of the calculation: $$\boxed{r(n) = 0{,}3332 \cdot e^{\left(\frac{n}{1{,}82}\right)} \quad [AE]} \qquad (12.20)$$

12.4 The hollow planet

The numerous accompanying moons of the big planets obey in the same manner as the planets this regularity, so that for the found result every coincidence is excluded. The orbits of the moons of Jupiter (fig. 12.5), of the moons of Saturn (fig. 12.6) and of the moons of Uranus (fig. 12.7) in logarithmic representation lie almost on a straight line. Some orbits certainly are occupied several times, while many an orbit has remained unoccupied. Other orbits again are occupied by a ring of countless chunks of rock, so-called planetoids. Best known representative is the asteroid belt (n = 4) between the orbit of Mars (n = 3) and that of Jupiter (n = 5). The Titius law requested the planet "Aster", but what one found (Piazzi, 1801), at first only was Ceres, the biggest representative of the small planets.

As a second asteroid, as they are called, was found (Olbers 1802) its discoverer proposed the explanation that both, Ceres and Pallas, could have formed in a cosmic catastrophe, which a bigger celestial body had suffered.

Perhaps the sought-for planet Aster actually had imploded after a collision and had been torn apart in countless fragments, of which today still more than half a million pieces are flying around. Most of the fragments however have been collected by the neighbouring planets Jupiter and Mars. Also in the case of the moons of Mars and the outer moons of Jupiter it could concern scrap of Aster. Some presumably fell on the young earth and triggered natural disasters and left deep scars.

It can be imagined that the planet Aster was hollow from the inside and thus was built up similar to the moon. Such a hollow sphere could have formed, if the sun, or in the case of the moon, the earth still was rotating with the cosmic velocity and as a result of the centrifugal force at the perimeter parts of the crust came off as connected sheets and rolled up to a tube.

The fast spinning tube then remodels into a hollow sphere with openings at the poles. In the inside of the sphere, near the equator, protected conditions and the best prerequisites for an undisturbed development, for instance of intelligent living beings, are found by the way. Apart from the advantages the hollow sphere however has as a fundamental disadvantage, that the normally protecting shell in the case of a collision with a bigger celestial body can becomes instable and can implode.

In the case of the moon the pole openings in the meantime are closed and a presumed gas pressure in the inside in addition provides stability. The wall of the spherical shell however is not evenly thick, so that the moon all the time turns the same, namely the heavier, side towards the earth. In view of the seismic measurements of the Apollo-missions, which revealed an extremely small damping of the ground of the moon (chap. 11.9, note <i>), we should look after our moon well, because a comet, which lets the moon implode, would be able to bring about more damage on earth, than if it would hit the earth directly.

Should there exist a hollow and possibly even on the inside habitable planet in our solar system, then surely Saturn should be considered. Its density is smaller than that of water, so that water would be distributed over the inside area of the hollow sphere, if it would be existent. One should examine more detailed the extremely flat pole regions of Saturn for possible openings! Also Uranus and Neptune are possible candidates. This only is thought as an idea.

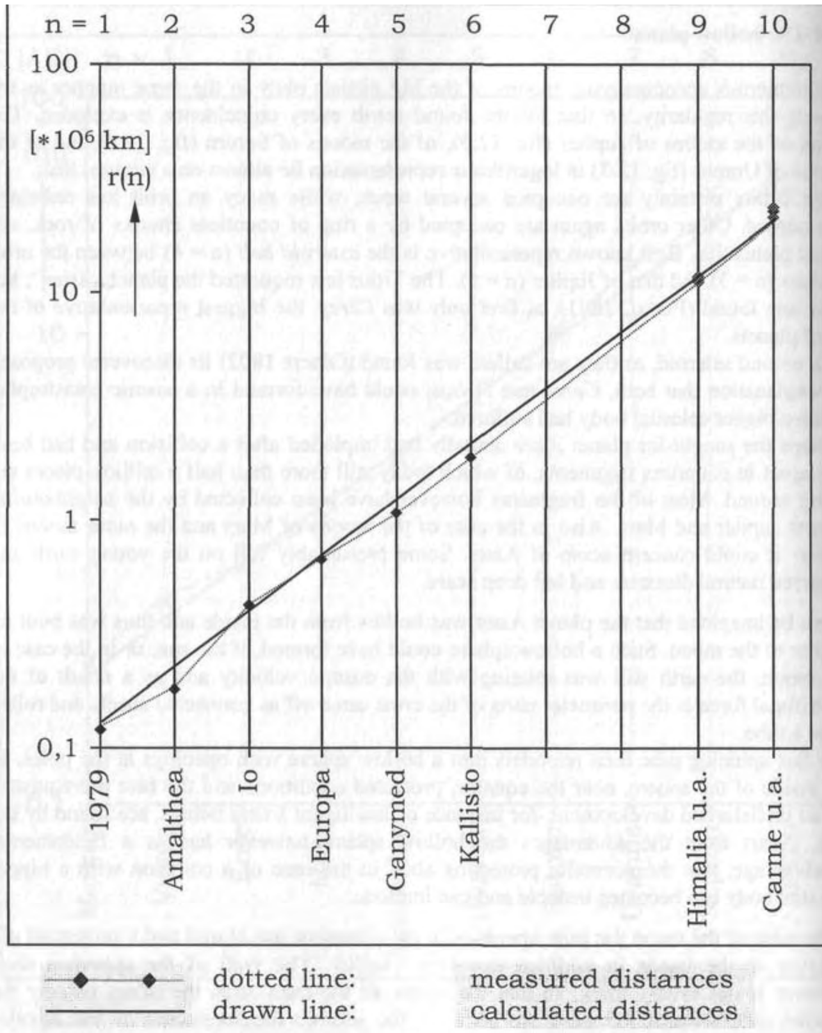

in position n = 9 stand: Leda, Himalia, Lysithea and Elara
in position n = 10 stand: Ananke, Carme, Pasiphae and Sinope

Fig. 12.5: The distances of the moons of Jupiter

Result of the calculation:
$$r(n) = 67608 \cdot e^{\left(\frac{n}{1{,}743}\right)} \quad [\text{km}] \qquad (12.21)$$

<i>: Note to chapter 12.5:

What we observe as and call cosmos, is nothing but a structured state of space.

12.5 Concerning the formation of the universe

The widespread concept of an expanding universe bases on the observation of a red shift of the spectral lines of galaxies, which increases with their distance. As the physical explanation for the discovery of Hubble, the Doppler shift for a light source, which is moving away from us, is used. But this concept of an against the attraction of the gravitation taking place expansion is nothing more than a work hypothesis.

The by Christian Doppler in acoustics investigated effect treats the observable shift in frequency, if the source of sound or the receiver is moved with regard to the medium of propagation. But according to today's version there doesn't exist such a medium at all for light, because Einstein has abolished the concept of the aether. According to that the Doppler effect neither can be applied to changes in light frequency. In the case of the expanding universe, for a decrease of the density and the tracing back to a Big Bang, it therefore should concern a misinterpretation!

Here another effect has to take effect, which one is not yet clarified. Perhaps the changing field relations of the observer environment play a role, after the earth moves away from the sun. But perhaps the galaxies only influence the propagation of their own light, or the light ray on its millions of years lasting way through space slightly loses energy, what is expressed by gradually increasing its wavelength and shifting its spectrum towards the red frequencies.

The hypothesis of the Big Bang moreover contradicts every causality. It is not able to give an answer concerning the origin and the future of the universe and on the question of the origin of the energy and the particles. With that its physical value of explanation goes towards zero.

If we again hold the theory of objectivity against the theory of relativity (part 1), we come to quite other answers. Here apart from the waves also vortices are found. Specially in the case of the spherical vortex a part of the wave power is enclosed in the inside, so that looked at from the outside a from zero differing energetic state results, which even is accessible measuring technically (chapter 6.2).

Wave and vortex are two possible forms of state, so that for the conversion of one state in the other state at first no energy is necessary. The change of state depends on the local field relations.

If we assume that in the beginning the cosmos was free of energy and particles. Then the first vortex was a possible product of chance with an infinite extension. This first spherical vortex, which was contracted under the potential vortex pressure, gave structure to space, gave it energy and field and took care for the rolling up and formation of new vortices[i].

It can be assumed, that even today new particles continually are being formed in the fringe areas of the infinitely extended cosmos, which fly towards us and in doing so contract. They are attracted and at the same time shrunk by the fields of the celestial bodies. They form the source of all matter and energy for our observable universe, which permanently is changing its structure. Because the same oscillation with reversed sign is enclosed in the inside of the spherical vortex, the sum of the energy present in the cosmos is exactly equal to zero. With that the question concerning the causality is superfluous.

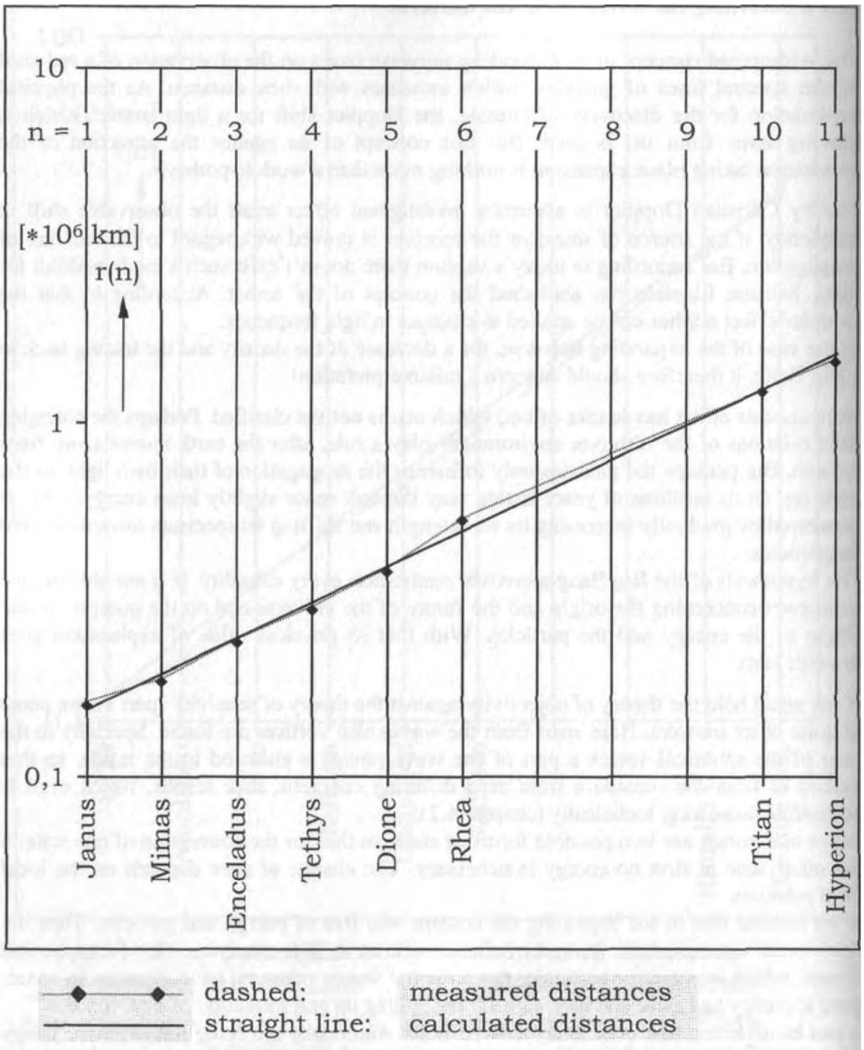

Fig. 12.6: The distances of the moons of Saturn

Note to chapter 12.6 (2nd law of thermodynamics[i]):

Heat cannot completely be transformed into mechanic or electric energy.

---

[i]: see Electromagnetic environmental compatibility, Part 1, fig. 8.4

The formation of the universe also can be explained causal and completely without a hot Big Bang, even if a supernova surely represents something like a "local Big Bang" for the concerned celestial bodies. According to the observations of the sky expansion and contraction, explosion and implosion occur everywhere in form of oscillations.

On the whole the expanding universe, which once was in thermodynamic equilibrium with matter, should cool down further and further. But stop, the 2nd law of thermodynamics teaches one just the opposite. If the entropy only increases, as the law dictates, then the whole universe should end in a heat death, just as mysteriously and inexplicable, as it should have been formed with the Big Bang. Perhaps something is wrong with the law (fig. 8.4)?

12.6 Counter examples concerning the 2nd law of thermodynamics

Most likely a small experiment convinces us. We heat two spheres, one somewhat less, the other somewhat more. Then we focus the heat radiation of the less hot sphere with help of a parabolic mirror and point it to the hotter one of both. That as a result becomes hotter, whereas the colder is cooled down. The heat thus has flown from the colder to the hotter sphere. Is it allowed to do that?
According to the 2nd law of thermodynamics it of course isn't. There the heat always can only flow from the hot to the cold sphere. But in this primitive experiment it measurable and verifiable flows in the wrong direction. Here the entropy, which is said to always only increase, actually decreases. Here entropy is being destroyed!
Shall we now let the carrying out of the experiment be forbidden under threat of penalty or shall we secretly put the law of entropy to sleep? It can't be denied that this law until now has quite well helped us along, at least for terrestrial processes, at least if one dispenses with the poor inventors, whose inventions offend against the since 100 years tried and tested 2nd law of thermodynamics. They haven't really arrived at the patent office at all, then they already are outside at the door again. Such inventors with their illegal behaviour even today must feel like criminals.

Just what that observed experiment can, each refrigerator and each warmth pump is using as well. And it is not an isolated case: also our sun clearly functions and operates illegally! The surface temperature amounts to only 5800 degrees Kelvin and supplies the atmosphere of the sun with energy. The energy thus flows from the sun to the corona, and that is with values above 1,000,000 degrees Kelvin for some powers of ten hotter!

Only the vortex concept resolves the many question marks without compulsion. In the case of the sun certainly vortices are at work. Here the high temperature in the corona arises as a result of vortices falling apart. We also speak of eddy losses. The transport takes place by heat radiation, exactly as in the experiment with the two spheres.
The possibility exists therefore in vortex processes as well as in technical circle processes that heat could flow from the colder to the hotter sphere. Whether this is an offence against the 2nd law of thermodynamics, is in the end a question of interpretation of the law and up to the opinions of the scholars.

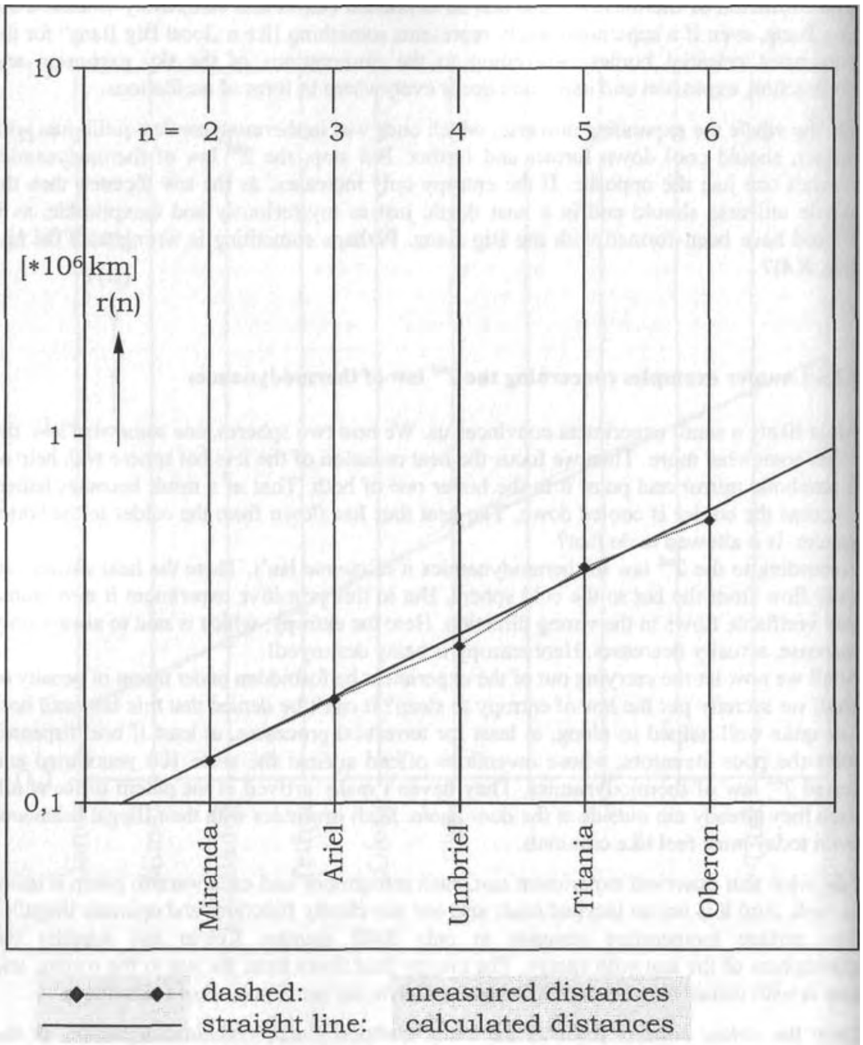

Fig. 12.7: The distances of the moons of Uranus

Note to chapter 12.7 (law of entropy):

The entropy of a closed system never can decrease. It is increased in the case of all irreversible processes. In the case of reversible processes it remains constant.

12.7 Entropy destroying potential vortices

Vortices in addition can amalgamate to balls and to planar vortex systems (4.9 and 4.10). In that case similar consequences, as they can be observed in flow-technical potential vortices in hydrodynamics, can be expected. As a result of the concentration effect (4.1) and because of the conservation of angular momentum an increase of the velocity of rotation of the vortex occurs. Like in the case of the pirouette in figure skating a spontaneous acceleration of its own is observed. In that way the kinetic energy of the system is increased, and that has to come from somewhere.
If we don't supply the contracting vortex with any additional energy for the increase of its rotation of its own, then as a source of energy only the heat energy is left. For this reason every contracting vortex generally converts heat in kinetic energy, it therefore cools down its environment! It moreover destroys entropy and offends against the 2nd law of thermodynamics (fig. 12.8).

It isn't an accident, if in the inside of a tornado it starts to hail. The whirlwinds really furnish visual instruction of the contraction and acceleration of their own of ring-like vortices. If then, even in tropical regions, hail stones are formed, the cooling effect has to come from somewhere, and it can be assumed that the vortex withdraws the heat energy from its environment (calculation in fig. 12.8).

If in specialist books is talked about matter or stars condensing, then vortex physics teaches us that they in reality are contracting and by doing that cooling down. That also is valid for whole galaxies. We owe solely the vortex laws that the continual heating by absorption of radiation is prevented and our sky in the night is dark and doesn't shine as light as day.

The cosmic background radiation, which as a result of vortices lies at almost 3° Kelvin, actually can only be given account for with a contracting of vortices of the Milky Way galaxy. If the cooling down in a compression resp. condensation process has arrived at absolute zero, the vortex becomes stationary, the contraction becomes dependent on the irradiated energy or it wholly comes to a standstill without supply of energy. An example for that is earth's inner core, which at zero Kelvin can't contract further.

From the point of view of causality it is suggested that we galactic and perhaps even cosmic to a large extent are in thermodynamic equilibrium, completely without Big Bang and heat death<i>.

Numerous inventions, as mentioned, are based on the principle of converting environmental heat in useable energy by contraction of vortices. If however the inventors don't know the vortex laws and if they have developed their concept empirically and less physically, then it very often happens that erroneously gravity is made responsible, then is talked of gravitational converters, of the use of a gravitational field energy. But the inventors don't do themselves a favour with that.

<i>: In the seminar it is desired to think about this. Doing so the philosophical faculty may feel as well addressed, as the sandpit of physics, in which completely unsuspecting is played and juggled with Big Bang hypotheses by ignoring all physical regularities and every common sense.

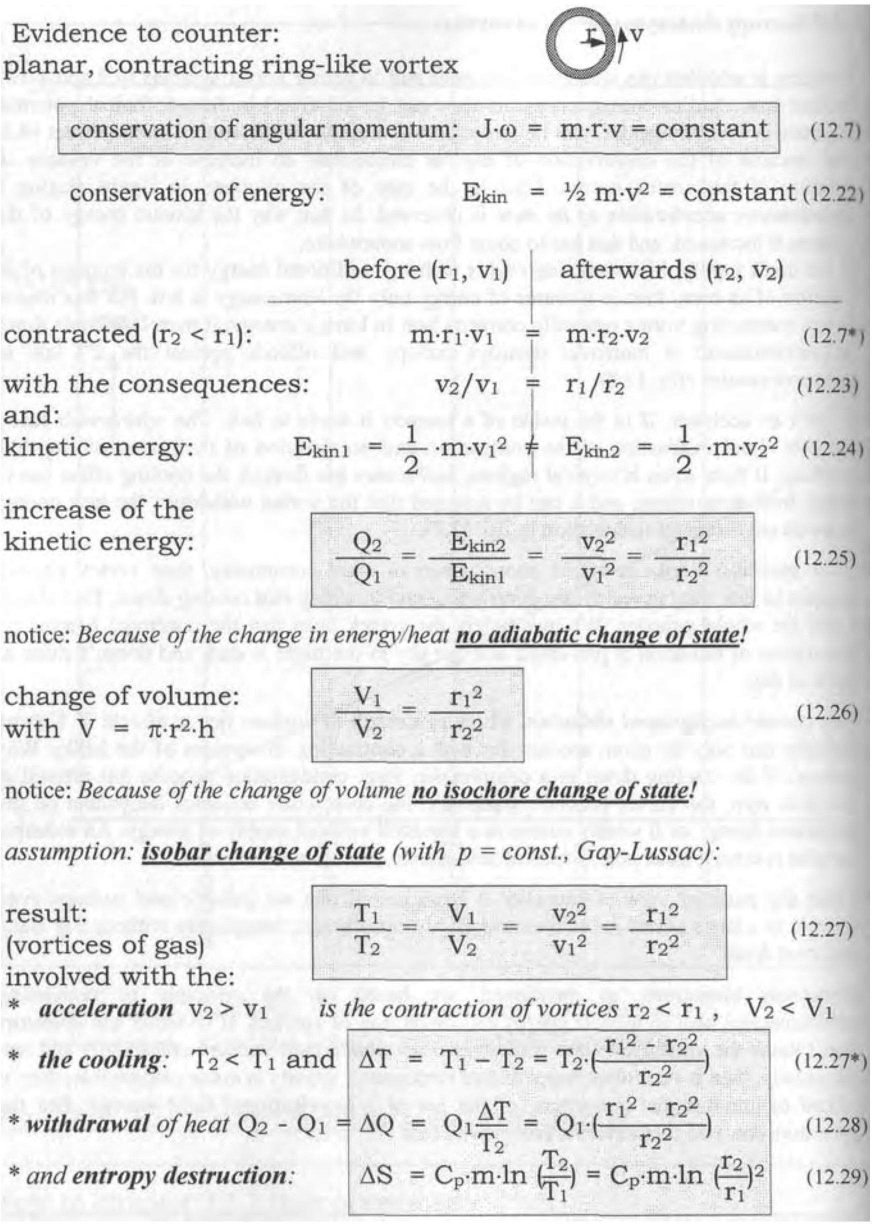

Evidence to counter:
planar, contracting ring-like vortex

$$\boxed{\text{conservation of angular momentum: } J\cdot\omega = m\cdot r\cdot v = \text{constant}} \qquad (12.7)$$

$$\text{conservation of energy:} \quad E_{kin} = ½\ m\cdot v^2 = \text{constant} \qquad (12.22)$$

	before (r_1, v_1)		afterwards (r_2, v_2)	
contracted $(r_2 < r_1)$:	$m\cdot r_1\cdot v_1$	$=$	$m\cdot r_2\cdot v_2$	(12.7*)
with the consequences:	v_2/v_1	$=$	r_1/r_2	(12.23)
and: kinetic energy:	$E_{kin1} = \frac{1}{2}\cdot m\cdot v_1^2$	$\neq$	$E_{kin2} = \frac{1}{2}\cdot m\cdot v_2^2$	(12.24)

increase of the kinetic energy:

$$\boxed{\frac{Q_2}{Q_1} = \frac{E_{kin2}}{E_{kin1}} = \frac{v_2^2}{v_1^2} = \frac{r_1^2}{r_2^2}} \qquad (12.25)$$

notice: *Because of the change in energy/heat* ***no adiabatic change of state!***

change of volume: with $V = \pi\cdot r^2\cdot h$

$$\boxed{\frac{V_1}{V_2} = \frac{r_1^2}{r_2^2}} \qquad (12.26)$$

notice: *Because of the change of volume* ***no isochore change of state!***

assumption: ***isobar change of state*** *(with $p = const.$, Gay-Lussac):*

result: (vortices of gas)

$$\boxed{\frac{T_1}{T_2} = \frac{V_1}{V_2} = \frac{v_2^2}{v_1^2} = \frac{r_1^2}{r_2^2}} \qquad (12.27)$$

involved with the:

- ***acceleration*** $v_2 > v_1$ *is the contraction of vortices* $r_2 < r_1$, $V_2 < V_1$
- ***the cooling**:* $T_2 < T_1$ and $\Delta T = T_1 - T_2 = T_2\cdot\left(\frac{r_1^2 - r_2^2}{r_2^2}\right)$ (12.27*)
- ***withdrawal** of heat* $Q_2 - Q_1 = \Delta Q = Q_1\frac{\Delta T}{T_2} = Q_1\cdot\left(\frac{r_1^2 - r_2^2}{r_2^2}\right)$ (12.28)
- *and **entropy destruction**:* $\Delta S = C_p\cdot m\cdot\ln\left(\frac{T_2}{T_1}\right) = C_p\cdot m\cdot\ln\left(\frac{r_2}{r_1}\right)^2$ (12.29)

Fig. 12.8: The cooling and entropy destroying effect of contracting potential vortices (calculations concerning chapter 12.7)

13. Recording of space and time

If observations should force us to touch the sacred physical laws, then we should first of all judge our measuring technique critically of all sides and bring it in an usable state. I suggest, we start completely from the beginning with the devices, with which the dimensions of space and time are recorded, with the tape measures for the measurement of length and the chronometers, our clocks.

We must find out, why comets are slowed down if they approach the sun like by the hand of a ghost and in going away again are accelerated, although no forces at all act on the celestial bodies from the outside.

We must find out, why in mines deep under the earth another value for the gravitational constant is being measured as on the surface of the earth<i>. The results hardly can be imagined, if an universal constant should lose its constancy.
Thereby can Newtonian mechanics and the well-known laws be used very successfully from today's point of view, as the derivations in the last chapters have shown. We even could verifiably and mathematically correct calculate the growth of the earth and the solar system with them. The physical laws in the normally used formulation in spite of that seem to be bound to certain limits. Some observations contradict all experience<ii>.

<i>: H. Schuh: Eine Konstante verliert ihre Konstanz; neue Experimente nahren Zweifel an Newtons Gravitationsgesetz, Die Zeit Nr. 40 vom 25.09.97. From it the following quotation (translated):
„It already is suspected for several years, that a fifth force could exist, this suspicion goes back to exact measurements of the gravitational constant G in Australian mines and shafts. Physicists of the University of Queensland in Brisbane had determined, that G for measurements underneath the earth is about one percent larger than the corresponding, since centuries in laboratories determined size. Their proposal for explanation, namely a fifth repelling force, at first met with sharp disapproval. But an at 21 august in the journal Science (Bd. 237/87, P. 881) published work, which bases on measurements in a drilling hole in Michigan, confirms the Australian data".

<ii>: Already the nobel prize winner of physics, Lenard, pointed to the circumstance, that the relativistic representation of the astronomical aberration is incorrect, after no distinguishable aberration could be observed at binary stars, as it had been expected. With that Lenard by the way also has confirmed the existence of an aether.
P. Lenard, Annalen der Physik, Bd. 73, S. 89 (1924)

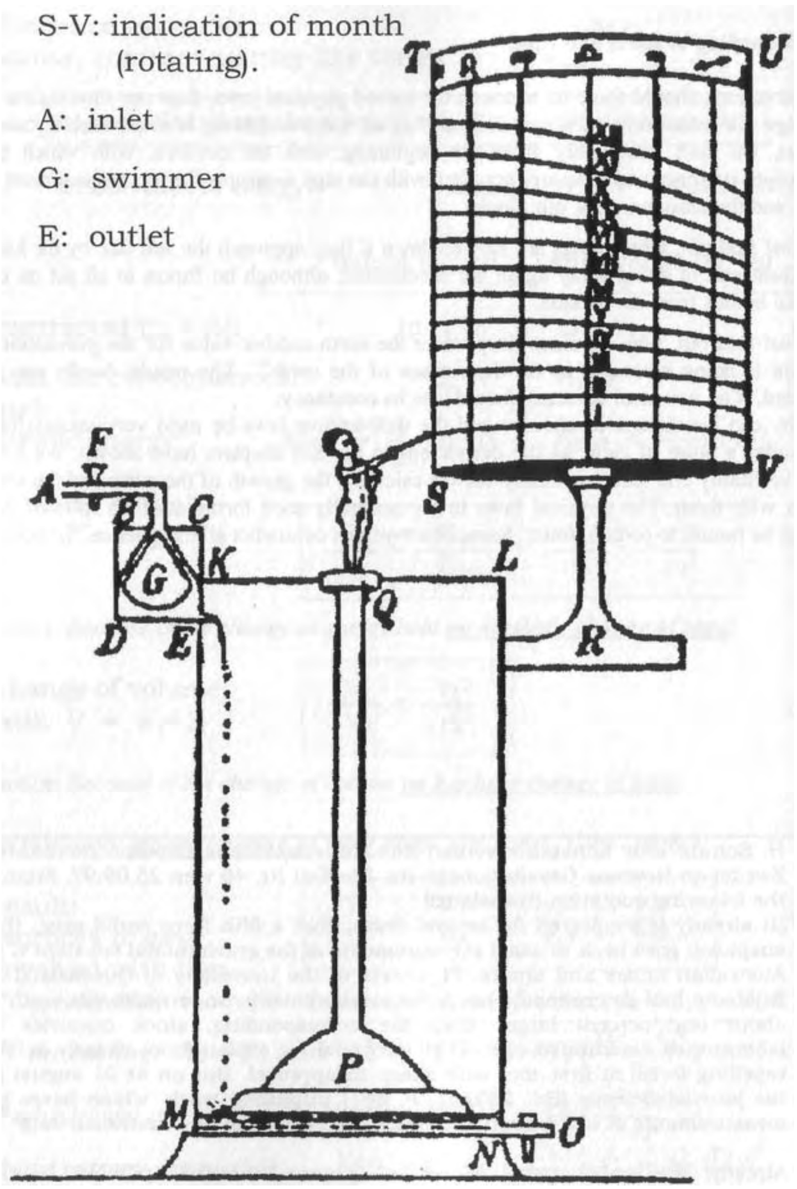
S-V: indication of month
(rotating).
A: inlet
G: swimmer
E: outlet
T
U
F
A
B
C
K
G
D
E
S
V
L
Q
R
P
M
N
O

Fig. 13.1: The water meter of the Ktesibios (approx. 250 BC)
(with regulation of the water-level)

13.1 The measuring technical debacle

We are standing for a measuring technical debacle, because we have fixed our calendar to the rotation of its own of the earth. We call a 360° turn a day, divide it in 24 hours of 60 minutes each and every minute in 60 seconds. With that we determine the duration of a second.

A clock according to this definition only then is exact, if it follows the changes of the earth to the same extent. Obviously this in particular is the case for atomic clocks. An objectively seen precise going clock however would land at the waste disposal site as completely unusable.

Alter all nobody wants to know, how late it really is. Everyone only wants to find confirmed his subjective feeling, and our experience of time simply is directed after the course of the sun, thus after the rotation of the earth!

For the moment and as an approximation also for the life time of a person, this determination of time may be sufficiently exact, but seen over larger periods of time, one doesn't get around a conversion. Actually the 200 million years since the beginning of the Creation on the mentioned primeval hill are considerably less long ago. The earth and the whole solar system are very much younger, than was assumed until now!

If we as an example again take the cited research group of the University of Arizona (chap. 11.10<i>), which by means of an analysis of sediment formations have found out, in accordance with our calculation, that 900 million years ago a day only had 18 hours<i>. Such traces in geological deposit layers can be very informative, if they are analysed correctly, since here the measurement technician himself is not subject to the process. He stands outside and hence can exactly measure the time difference.

If a geologist of that time had looked at his watch during the formation of the layers, then a day obviously would have had 24 hours, and he wouldn't have been able to understand the whole excitement at all. The accusation, he would suffer from chronic blindness caused by his job, he of course would repel resolutely. Can you now imagine, from which disease our science of today suffers?

Our chronometers are nothing but the improved model of a sundial<ii>.

We live in the dimensions of space and time, but we quite obviously have the biggest possible difficulties with the dimensioning of both. Possibly we already aren't capable of that at all by principle. Most suitably we clarify the situation by means of examples concerning the two problem spheres.

Let us at first stay at the dimension of time. It may have become clear, how problematic the chosen determination of the time scale in seconds is and which contradictions can result if larger periods of time are considered. More and more often science fiction authors romp about on the playground of time, fantasize about some time travel, or they occasionally make jokes about it<iii>.

<i>: 900 million years ago a day had 18 hours, Washington (dpa) 1997.

<ii>: There also exist models, of which is asserted, they go according to the moon.

<iii>: A passenger, who got on the 10 o'clock bus, passes by the church steeple clock, which only reads 5 to 10 and curses: Damn, I took the bus in the wrong direction!

Centrifugal force $\boxed{F_{1,2} = (m/R)\cdot(v_E \pm v_f)^2 = m\cdot R\cdot(\omega_E \pm \omega_f)^2}$ (13.1)

with: $R = 6378$ [km] (radius of the earth),

$v_E = \omega_E \cdot R = 0{,}465$ [km/s] (rotational speed of the earth), (11.20)

$v_f = \omega_f \cdot R = \pi \cdot R/t_o$ (speed of the plane) (13.2)

and: t_o [s] (duration of journey)

energy of the moving steering quanta:

$$\boxed{E_{1,2} = \int_0^R F_{1,2}\, dR = m \int_0^R R\cdot(\omega_E \pm \omega_f)^2\, dR =} \quad (13.3)$$

$$\boxed{E_{1,2} = \frac{1}{2}\cdot m\cdot R^2\cdot(\omega_E \pm \omega_f)^2 = \frac{1}{2}\cdot m\cdot(v_E \pm v_f)^2} \quad (13.4)$$

energy balance (generally of a quantum of radiation):

$$\boxed{E_{1,2} + E_q = E_{ges}} \quad \text{resp.} \quad \boxed{E_{ges} - E_q = E_{1,2} = \Delta E} \quad (13.5)$$

in general: $E = h\cdot f = m\cdot c^2$ resp. $m = h\cdot f/c^2$ (13.6)

(Gl. 13.4 mit 13.6):

and specifically: $\boxed{\Delta E = h\cdot\Delta f} = \boxed{E_{1,2} = h\cdot f\cdot(v_E \pm v_f)^2/(2c^2)}$ (13.7)

$\Delta f/f = \Delta t_{1,2}/t_o$ resp. $\Delta t_{1,2} = t_o\cdot\Delta f/f$ (13.8)

(eq. 13.7 inserted into eq. 13.8)

resulting in: $\boxed{\Delta t_{1,2} = t_o\cdot(v_E \pm v_f)^2/(2\cdot c^2)}$ (13.9)

time of travel around half the earth: $t_o = \pi\cdot R/v_f$ (13.2)

results in a difference in going of: $\Delta t = \Delta t_1 - \Delta t_2$ (13.10)

$$\Delta t = [(v_E + v_f)^2 - (v_E - v_f)^2]\cdot\pi\cdot R/(v_f\cdot 2\cdot c^2) \quad (13.11)$$

result calculated: $\boxed{\Delta t = 2\cdot v_E\cdot\pi\cdot R/c^2 = 207\ [\text{ns}]}$ (13.12)

measured by Hafele and Keating<i>: 214 [ns]! (13.13)

Fig. 13.2: The difference in going of two atomic clocks (caesium resonant clocks)

<i>: Hafele-Keating-Experiment, Oktober 1971; s. a. W. Bauer: Klassische Physik, Graphia Druck, Salzburg (1975), Eigenverlag

13.2 The clock paradox

The theme „time dilatation" in connexion with the particle decay already has been treated (part 1, chapter 6.20). It has been shown, that a fast moving and with that length contracted particle to exactly the same extent (Lorentz's square root) becomes more stable and longer-living. If relativists pack an atomic clock based on radioactive decay in a plane and detect a difference in going between the one, which has been flown around and a second identically constructed clock, which has stayed at the ground, then they have with that detected experimentally a very small length contraction, which really occurred, and by no means a time dilatation, as they claim.

Now we in addition owe Einstein, that the aether has been abolished and from that follows, that it can't play a role in which direction the plane flies. If therefore both clocks are taken along each in a separate plane, one plane flying to the west and the other to the cast, both planes meeting again for the first time on the other side of the globe, then according to Einstein's theory it shouldn't be possible to determine a difference in going, if both planes constructed identically were on the way with the same velocity. But this is not the case!

Actually a difference in going is measured, which however can't be calculated with the theory of relativity, yes, which is completely incompatible with this theory and clearly brings anyone to the eye, that the effect actually can't have to do anything with a time dilatation, that the moving clocks merely go wrong and we have to ask us, why<i>.

These experiments were carried out with atomic clocks, which are constructed as caesium resonators and work with an exactness of one second in 300000 years. As a resonator serves a quartz crystal, which is controlled by an ion current of caesium atoms, which have lost their outermost enveloping electron. The system is fed back, because the oscillating quartz controlled by the caesium ions again adjusts the caesium vapour by radio wave and finally its own atomic controlling current (fig. 13.7).

The reason for the measured difference in going is seen in the field and here specially in the different gravitational field. The centrifugal force directed opposite to the gravitational force at least is not the same, because for a westward flight along the equator the speed of the plane v_f should be subtracted from the velocity of rotation of the earth v_E, whereas in eastward direction it should be added (eq. 13.1).

For the steering quanta supplied by the caesium resonator now the energy balance is put up (13.5 with 13.4) and the change of the reference frequency is calculated (13.7). With the change in frequency is connected directly a change of the at the two clocks readable times $t_{1,2}$ (13.8). For a journey around half the earth, where one clock is flown westwards and the other one eastwards, the difference in going should, according to the calculation, amount to 207 ns. Interesting of the result (13.11) undoubtedly is, that the velocity of the planes doesn"t play a role. It is cancelled out.

October 1971 caesium atomic clocks were sent around the world in scheduled planes in the Hafele-Keating experiment. To be able to estimate the inaccuracy in going of the clocks and with that the measurement error, four clocks were used. Between the westward journey (273 ±7 ns) and the eastward journey (59 ±10 ns) a difference in time of 214 nanoseconds was determined. This under strict scientific conditions determined result once more proves the correctness of the theory of objectivity by confirmation of the calculated value. That however is not valid for the special theory of relativity, because that doesn't appear in the calculation at all!

Who now believes, we would have less problems with the dimension of space, I must disappoint. The determination of the linear measures equally ends in a fiasco.

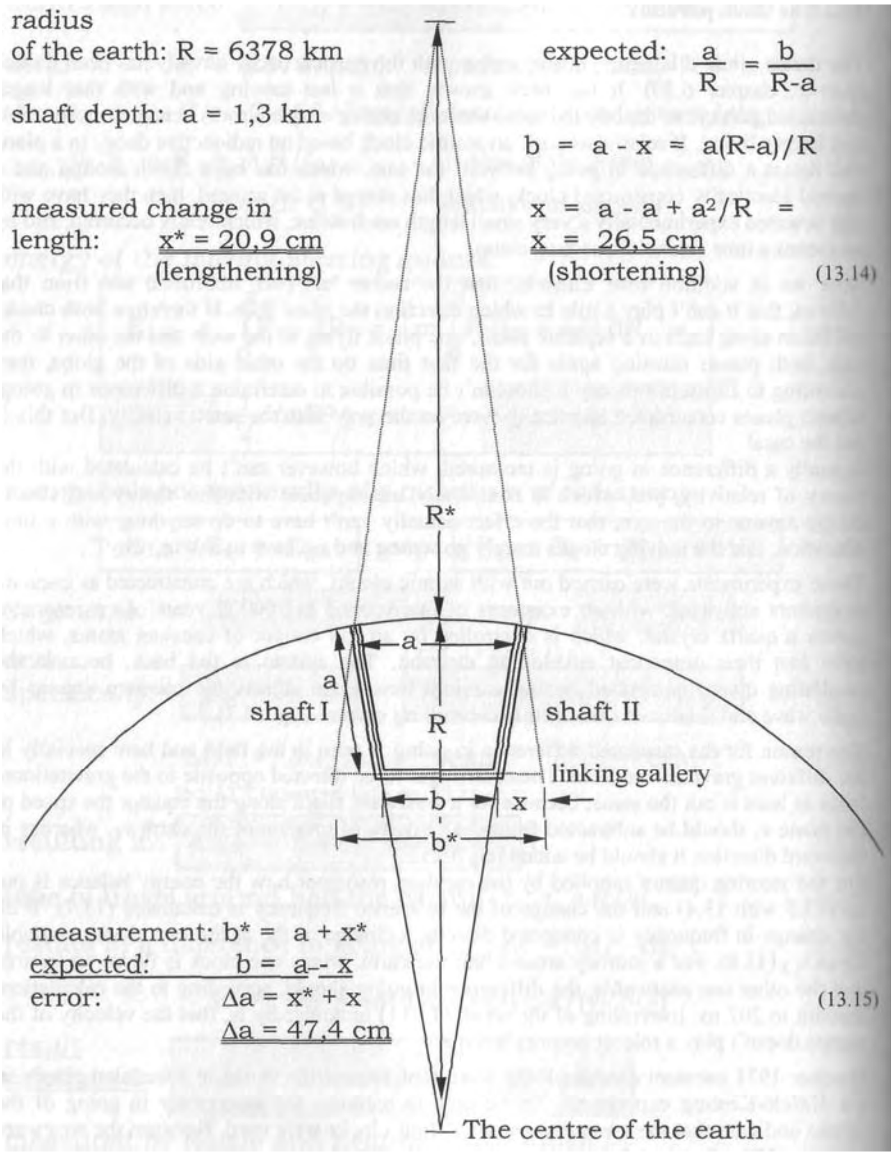
radius
of the earth: R = 6378 km
shaft depth: a = 1,3 km
measured change in
length: x* = 20,9 cm
(lengthening)
expected: a/R = b/(R-a)
b = a - x = a(R-a)/R
x = a - a + a²/R =
x = 26,5 cm
(shortening)
(13.14)
R*
a
a
R
shaft I
shaft II
linking gallery
b
x
b*
measurement: b* = a + x*
expected: b = a – x
error: Δa = x* + x
Δa = 47,4 cm
(13.15)
The centre of the earth

Fig. 13.3: Expectation and measurement in the Tamarack mines
(The curvature of the earth and the length of the shaft a are drawn strongly exaggerated for clarification)

13.3 The Tamarack mines experiment

As long as the „foot measure" depended on the shoe size and the „cubit" on the forearm of the tailor, the world still was OK. The sciences however request a reproducible quantity for comparison, and that can be fetched at the Bureau International des Poids et Mesures in Sevres near Paris. The original meter is a Platinum alloy. Because the length of the metal always depends on temperature, it is stored at a constant kept temperature of 0° Celsius. Now there in addition still exists a field dependency, an electrostriction resp. magnetostriction. And how is the measurement bar behaving, if the earth grows and the density increases? Is it then shrinking just like other objects in its environment? At this point already chaos is proliferating.

The newest definition of the length measure meter acts as a blow for liberty and thus marks the abyss, at which we are standing: The length is determined by means of a measurement of transmission time of an electromagnetic wave, e.g. of a light signal. It is said that with this determination a higher reproducibility should be obtained. Actually a photo optical facility to measure length is as exact as the built-in facility to measure time, and there we use it again, our sundial. In addition a constancy of the speed of light is taken as a prerequisite, and that is given in meters per second. From a change of the speed of light for forinstance 10% a change in length for 10% as well would result<i>. Because we see this process with the help of our eyes as well with the speed of light, we never can see the change. We neither can technically measure it, because all gauges we construct are built up corresponding to our sensory impression. We ourselves have shovelled the hole, in which we fall.

Only if we succeed in taking a neutral standpoint outside of the events, the true relations will become visible to us. For the field dependency of the space measures a very clear experiment has been carried out, of which I now want to report<ii>.

1901 the French government was offered the possibility to carry out an experiment in the shut down Tamarack mines near Calumet (Michigan) with the goal to determine the diameter of the earth more exact. For that the geophysicists let down two plumb-lines of 27.2 kg each in two perpendicular winding shafts, which were at a distance of 1.3 km from each other. The plumb-lines were tied to hardly expandable piano wires of as well 1.3 km length. It now was expected, since the plumb-lines hung in direction of the centre of the earth, that in a linking gallery between the two shafts a length of (1.3 - x) km should be measurable. From the shortening x one wanted to infer the diameter of the earth (fig. 13.3). But it came completely different.
Instead of a shortening a lengthening for x* = 20.9 cm was measured in the gallery! The point of intersection of the lines through the two shafts had to be not in the inside of the earth, but in space! Immediately „hollow earthers" appear, who claim we would live on the inside of a hollow world<ii>. Perhaps one should shoot them to the moon, because from there the earth without doubt is seen as a sphere. Obviously we aren't dealing with a surprise of a fair, but with a fundamental measuring technical problem.

<i>: for that see also part I, chapter 6.11 and 6.12
<ii>: http:/www.t0.or.at/subrise/hollow.htm

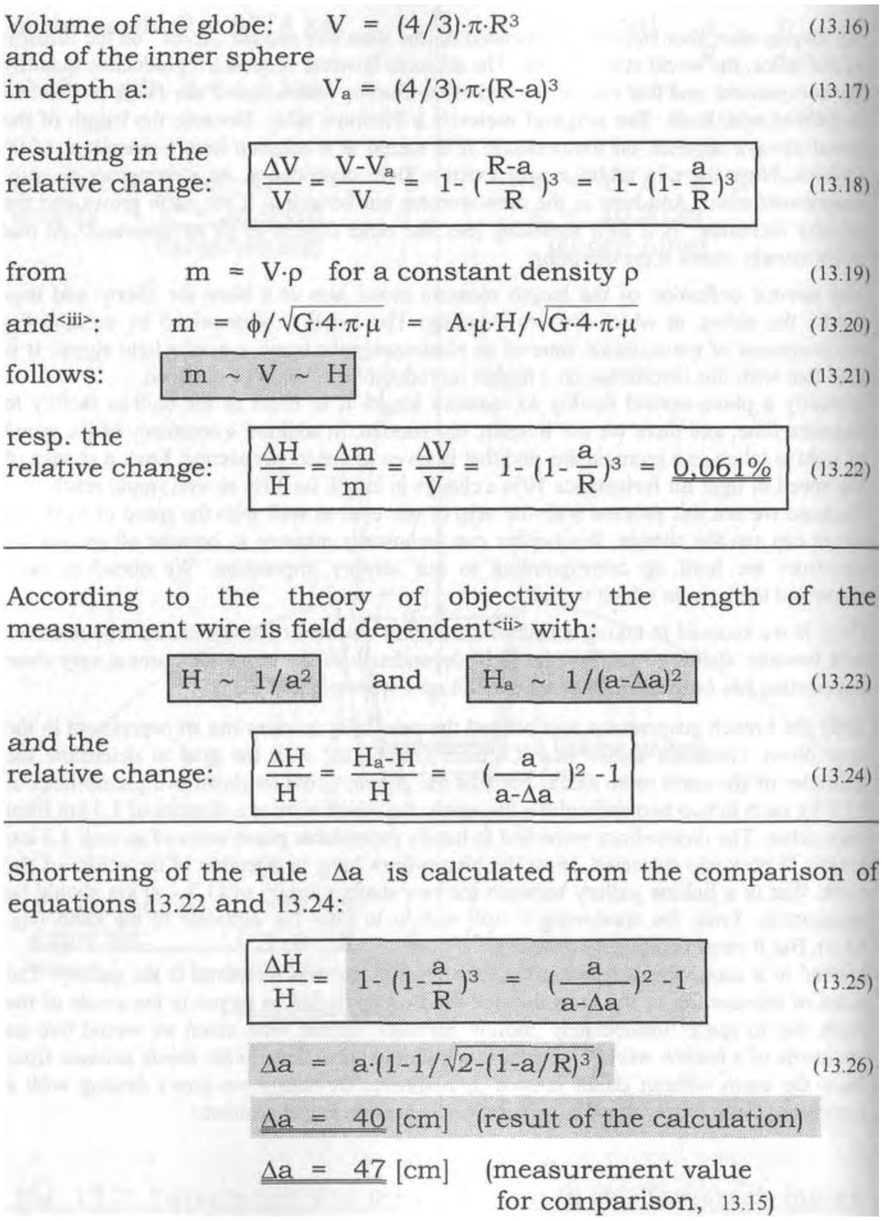

Volume of the globe: $V = (4/3)\cdot\pi\cdot R^3$ (13.16)

and of the inner sphere in depth a: $V_a = (4/3)\cdot\pi\cdot(R-a)^3$ (13.17)

resulting in the relative change: $\boxed{\dfrac{\Delta V}{V} = \dfrac{V-V_a}{V} = 1-(\dfrac{R-a}{R})^3 = 1-(1-\dfrac{a}{R})^3}$ (13.18)

from $m = V\cdot\rho$ for a constant density ρ (13.19)

and[iii]: $m = \phi/\sqrt{G\cdot 4\cdot\pi\cdot\mu} = A\cdot\mu\cdot H/\sqrt{G\cdot 4\cdot\pi\cdot\mu}$ (13.20)

follows: $\boxed{m \sim V \sim H}$ (13.21)

resp. the relative change: $\boxed{\dfrac{\Delta H}{H} = \dfrac{\Delta m}{m} = \dfrac{\Delta V}{V} = 1-(1-\dfrac{a}{R})^3 = \underline{0{,}061\%}}$ (13.22)

According to the theory of objectivity the length of the measurement wire is field dependent[ii] with:

$\boxed{H \sim 1/a^2}$ and $\boxed{H_a \sim 1/(a-\Delta a)^2}$ (13.23)

and the relative change: $\boxed{\dfrac{\Delta H}{H} = \dfrac{H_a-H}{H} = (\dfrac{a}{a-\Delta a})^2 - 1}$ (13.24)

Shortening of the rule Δa is calculated from the comparison of equations 13.22 and 13.24:

$\boxed{\dfrac{\Delta H}{H} = 1-(1-\dfrac{a}{R})^3 = (\dfrac{a}{a-\Delta a})^2 - 1}$ (13.25)

$\boxed{\Delta a = a\cdot(1-1/\sqrt{2-(1-a/R)^3})}$ (13.26)

$\underline{\Delta a = 40}$ [cm] (result of the calculation)

$\underline{\Delta a = 47}$ [cm] (measurement value for comparison, 13.15)

Fig. 13.4: Calculational verification of the measured shortening

13.4 Field dependent linear measure

If a measurement result delivers just the opposite, as was expected by the experimentators, then the layman is amazed and the expert is surprised, at least at first. But then, out of a feeling of panic, the whole view of life could collapse, as many as possible scientists and renowned professors are being informed and integrated, if they want it (Prof. Mc.Nair) or not (Prof. Hallock, Columbia University) and eventually the matter is buried third class and a guise of silence is spread over it. The censorship of the scientific making of opinion in advance doesn't permit publications, which are not in accord with our view of life, out of the animal survival instinct<i>.

A science, which deserves this name, should look different. There it must be permitted, to ask questions and to publicly discuss about it.

I proceed from the assumption, that the earth is a sphere, which we inhabit from the outside; I have no doubts about that. With this as a prerequisite there is only one possible answer to the Tamarack mines experiment: The 1.3 kilometres long measurement wire, which in the gallery in a depth of 1.3 km had to jut out for 26.5 cm, instead is too short for 20.9 cm, from which immediately follows that it, howsoever, is shrunk for 47.4 cm in the depth.

At first of course the experiment was checked for possible measurement errors. The shafts were covered to exclude any draught. The measurement path was optically checked, but the result remained unchanged. But if the cause for the unexpected result doesn't lie in the experiment, then theoretical physics is addressed, after all it is still the experiment which shows us the physical reality and not some theoretical model concept.

In the 1st part of the book already the derivation of a useful explanation is found: The speed of light and with that also the linear measure depends on the field<ii>. The measurement wire accordingly gets shorter, if it is exposed to a larger field strength (eq. 13.23). We can verify the measured shortening calculative (fig. 13.4).

For that we at first determine the change of the field strength, as it is to be expected in a depth of 1.3 km. We here are dealing with the closed H field lines, which are responsible for the gravitation. In a past derivation it has been shown, that a gravitating mass can be converted into a magnetic field<iii>. Between a mass m and a field strength H hence exists a proportionality (13.21), in the same way as between the same mass and its volume, if a constant density is present (13.19).

The result accordingly is a relative decrease of the volume of the earth and the corresponding mass being under the measurement place, as well as a relative decrease of the radial component of the field strength, but a corresponding relative increase of the tangential component of the field for 0.061% (13.22 with 13.18).

In the gallery the measurement wire however is spread out in the direction of the tangential component of the field lines, and that shortens the measurement wire as a result of the field dependency of the linear measures<ii> (13.23). From the above increase in field the calculational shortening of the rule for 40 cm results, which compared to the measurement result also should be rated as a confirmation of the theory of objectivity which was taken as a basis<ii>!

<i>: Example: K. Meyl: Potentialwirbel, Bd. 1, reference entry /5/

<ii>: for that see also part 1, chapter 6.6, equation 65 and chapter 6.10

<iii>: K. Meyl, Potentialwirbel Band 2 (1992), page 27, equation 20

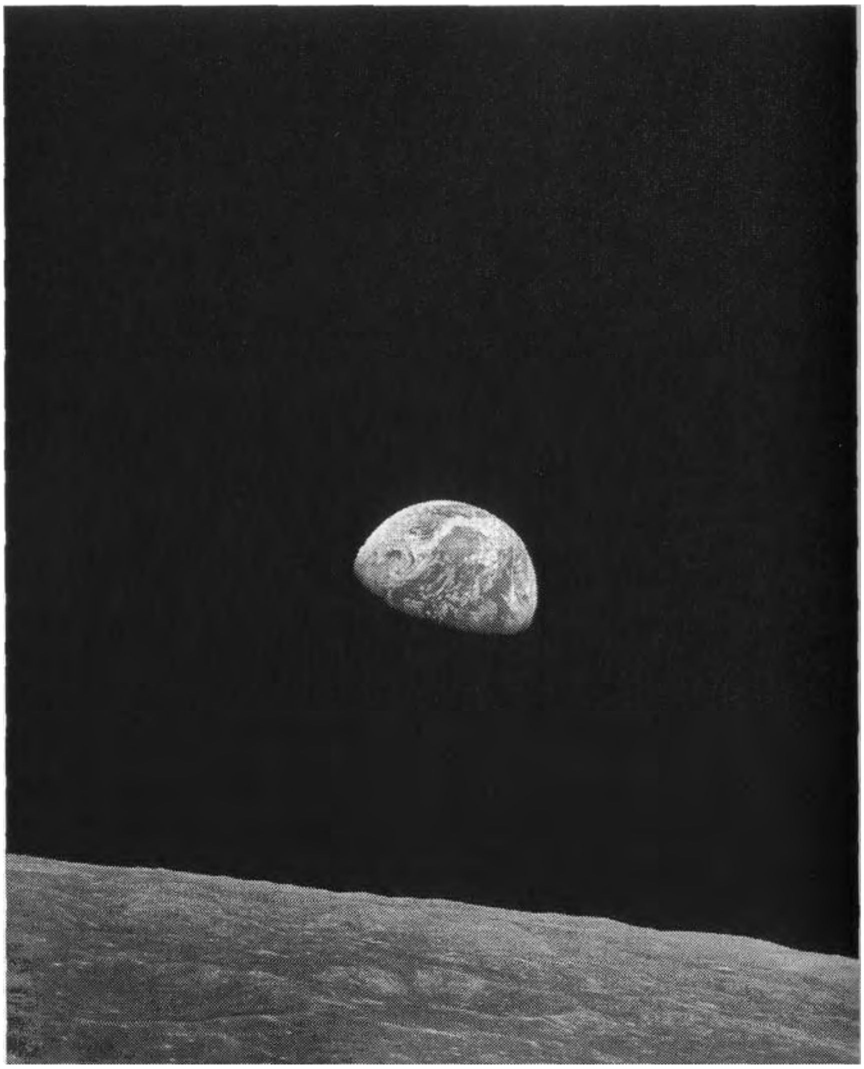

Fig. 13.5: Rise of the earth over the horizon of the moon""

<i>: Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie, Orbis Verlag (1989), note: No star is seen!

<ii>: see also the references in part 1, chapter 6.9

13.5 Experiences from space travel

The good correspondence of the calculated shortening of the measurement wire and the until now not understood measurement in the Tamarack mines shows both qualitatively, and quantitatively the correctness and useful applicability of the theory of objectivity of the field dependency of the linear measures. Our measurement laboratories normally are situated on the earth's surface and there everywhere are found approximately identical field relations. But if we leave the usual measurement environment and move the laboratory for instance in the sky, then we experience a complete mystery. Here however prevail the reversed conditions as in the mines experiment, in which in the inside of the earth. for an increase of the field strength, a length shortening was measured. In the sky the field strengths decrease and the linear measures correspondingly increase.

This experience astronaut Roosa made in the Apollo 14 mission. While he alone in his capsule orbited the moon, he depicted mission control, he could see the lunar module and observe his two colleagues at their work on the moon. Nobody wanted to believe the astronaut, since he was flying in a height of 180 km!

Commander Armstrong (Apollo 11) at the first landing on the moon indicated, the target crater Mackensen, 4.6 km in diameter measured from the earth, just has the size of a soccer field! Astronaut Scott (Apollo 15) called Mount Hardley, which is said to be 4.8 km high, a practice hill for skiing. Perhaps they somewhat have exaggerated, but a true core in the statements always is present!

Actually the gravitational field of our satellite is very much smaller than that of the earth. On the surface of the moon there is only one sixth of the gravitational pull of the earth. If we, to be able to compare, stick to the details of size, as they are measured by our laboratory on earth, then the astronauts on the way to the moon together with the lunar module and their rover had grown for a factor $\sqrt{6}$,then the first footprint is 2.5 times as large as on earth, then the astronauts were moving like giants in the scenery of a model of the railroad (eq. 13.23 and note<ii> at fig. 13.4).

On the moon there exists almost no atmosphere, for which reason the astronauts had imagined a wonderful view of the star-spangled sky, at least before they started. After the landing they were bitterly disappointed. The sky was black and not one single star could be seen! They have brought many photographs, but nowhere stars have been photographed, they apparently have moved outside the range of vision (fig. 13.5 and 13.6).

Many will still remember that the first pictures, which the space telescope Hubble supplied 1990, were completely blurred. The problem obviously was, that the mirrors had been adjusted on earth and not in space. Only after the optics had been given glasses in 1994, sharp pictures could be radioed to earth. Somehow the distance to the stars had changed. The telescope had become short-sighted, resp. the distance to the star-spangled sky appeared to be gotten larger. We already know why. If we remove us from the gravitational field of the earth, the field strength decreases and the observable distances increase! The highly sensitive telescope already sufficed the 5% deviation, with which should have been reckoned for the near earth orbit, to be fatal.

One should have familiarized the astronauts before with the laws of physics. Then this disappointment would have been spared to them, and in the case of the Hubble telescope the NASA and the european ESA could have saved a lot of money for the sake of the tax-paying population<ii>.

Warum sieht man am Himmel keine Sterne?

Fig. 13.6 a: Illustrierte Wissenschaft Nr. 11 (1997) Page 62

1st example: landing on the moon gravitational ...

$$\frac{\text{...pull of the earth}}{\text{...pull of the moon}} = \frac{g_E}{g_m} = \frac{M / R^2}{m_m / R_m^2} = \frac{M}{m_m} \cdot \frac{R_m^2}{R^2} = 6{,}0375 \qquad (13.27)$$

in space everywhere on the surface of a sphere is valid: $A = 4\pi R^2$

eq. 13.20 (fig. 13.4<iii>): $m \sim \phi = A \cdot B = 4\pi R^2 \cdot \mu H \sim R^2 \cdot H$

as well as eq. 13.23 (fig. 13.4<ii>) for the field dependency: $H \sim 1/l^2$

and further:

$$6{,}0375 = \frac{g_E}{g_m} = \frac{H_E}{H_m} = \frac{l_m^2}{l_E^2} \qquad (13.28)$$

resulting in the length dilatation (expansion) on the moon:

$$\boxed{l_{m\,(\text{Mond})} : l_{E\,(\text{Erde})} = \sqrt{6{,}0375} = \underline{2{,}457}} \qquad (13.29)$$

2nd example: communications satellite in a geostationary orbit (at h = 36000 km above the equator, R = 6378 km).

$$\boxed{\frac{l_h}{l_E} = \sqrt{\frac{H_E}{H_h}} = \frac{R+h}{R} = \underline{6{,}64}} \qquad (13.30)$$

Fig. 13.6 b: Examples of calculation for length dilatation<i>

<i>: If the theoretical value of 6.64 could be quantitatively confirmed by observations from off the earth with a telescope at a corresponding magnification, then with that would have been proved, that in the case of the spherical aberration it actually concerns the calculated influence of the field.

<ii>: U. Seiler-Spielmann: Das Marchen vom toten Mond, Zeiten Schrift 5/94,S.39

13.6 Spherical aberration

It is true that the problem of the changed length relations is known to the experts under the term of a "spherical aberration". But with that it is neither qualitatively nor quantitatively understood. Only the theory of objectivity soundly gives reasons for, why the astronaut Roosa has seen his colleagues almost 3 times as large, why weather satellites in a height of 1500 km are approx. 25% larger and why communications satellites in a 36000 km high geostationary orbit even increase to the 6.64 fold of their original size<i>. It also explains, why the neutral point between earth and moon, at which the attraction of masses of both celestial bodies mutually cancel, wasn't reached at the point where it had been expected by the moon rockets<ii>.

We, the inhabitants of the earth, are adapted completely to the conditions on the earth's surface. If we find our way well in the dimensions of space and time, as we observe them, then that must not be valid by all means for science, because that has made it its business to find out the secrets of nature.

If it wants to deserve the name science, then it on the one hand has to consider, that we, the organic materials, as well as all inorganic materials are assembled from the same atoms and molecules and with that are exposed to the same length relations. If changes in length between day and night (as a result of the gravitational field of the sun (see chap. 6.7), between summer and winter or as a result of changes in field occur, then we aren't able to register this at all. That even today the „foot" is used as a measure, for instance in the air traffic, shows only too clearly, how man raises itself to the measure of all things. Science asks for modesty!

On the other hand it should be paid attention to the fact, that man eyes everything with speed of light with the optics of his eyes, and that speed by no means has to be constant. Solely the definition of the speed of light c as a linear measure per unit of time points to the direct proportionality between c and a length 1 (see chapter 6.3):

$$\boxed{c \sim 1} \qquad (13.31)$$

If a rule has proven to be unusable for measuring a distance, then we'll experience the same disaster, if we measure optically, i.e. with the speed of light. Obviously both, the length 1 and the speed of light c depend in the same manner on the respective local field strength. On the one hand both measurement techniques lead to the same result, but on the other hand what can't be measured with one method, neither can be measured with the other.

To prove the constancy, it is normal to measure the speed of light optically. But since there exists a proportionality between measurement variable and measurement path (53), the unknown variable is being measured with itself. This measurement faulty by principle in all cases delivers a constant value. In contrast to the textbook opinion of today by no means a constancy of the speed of light can be assumed. In the case of the in a vacuum measurable 300,000 km/s it concerns a capital measurement error, at best a constant of measurement, but never ever a constant of nature!

With the postulate and the misinterpretation of a constancy of the speed of light as a universal constant of nature Einstein already let several generations of physicists run into the same dead end, in which they today are stuck altogether. It surely is no accident, that the big time of discoveries abrupt came to an end with Einstein.

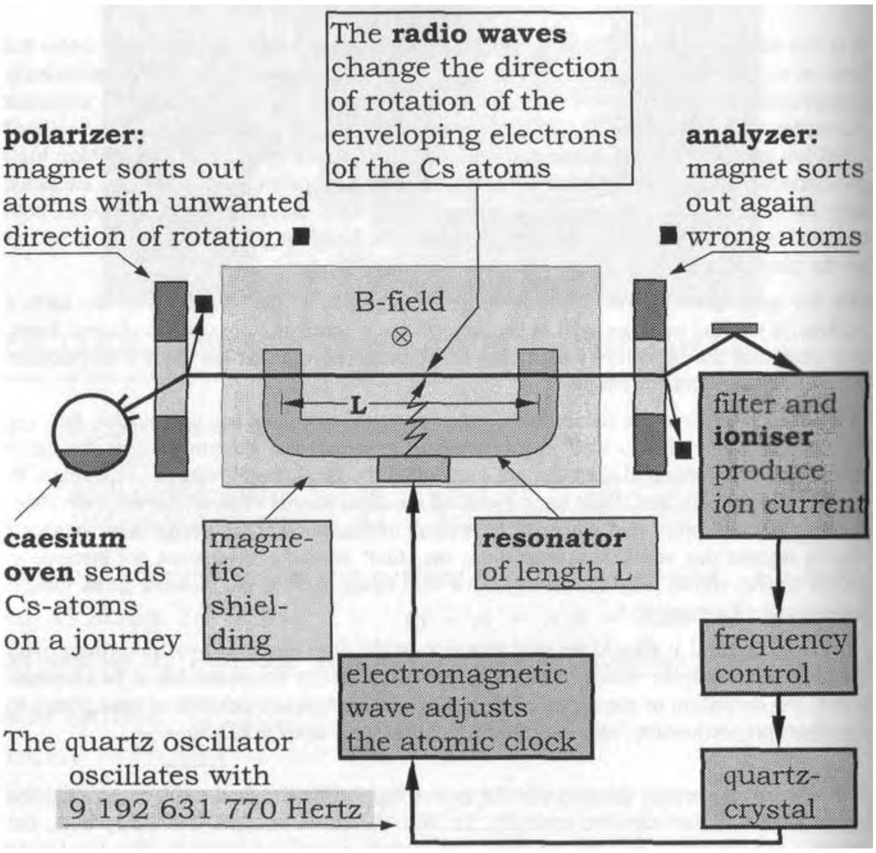
The **radio waves** change the direction of rotation of the enveloping electrons of the Cs atoms
polarizer: magnet sorts out atoms with unwanted direction of rotation ■
analyzer: magnet sorts out again ■ wrong atoms
B-field
⊗
L
filter and **ioniser** produce ion current
caesium oven sends Cs-atoms on a journey
magnetic shielding
resonator of length L
frequency control
electromagnetic wave adjusts the atomic clock
The quartz oscillator oscillates with 9 192 631 770 Hertz
quartz-crystal

Schematic representation concerning principle of functioning
L = length of resonator determining the exactness
The B-field is a weak magnetic field, which eliminates the influence of magnetic stray fields.
The arrangement in addition is situated in a vacuum tank.

Fig. 13.7: Set-up of a caesium atomic clock

13.7 Irony of the measuring technique

Let's record: The linear measure is determined and defined by a measurement of transmission time. As a reason is given, that with today's clock technology a higher precision and reproducibility can be obtained, as with a rule or original meter. The exactness of going of the atomic clocks again depends on the free flying path (L in fig. 13.7) of the atoms. For the caesium clocks of the Physikalisch Technischen Bundesanstalt in Braunschweig the resonator length amounts to several meters! The clock is used world-wide as a standard.

The irony thus lies in the fact, that a geometric length dictates the measurement of time and the measurement of time again determines the measurement of length - poor science!

How does one free oneself from a capital closed loop conclusion? Why and how do signal transmission times or clocks actually depend on the gravitation? Who once got stuck in a dead end, knows that he only can get out in the reverse gear.
A possible way goes back to the roots of classical physics and to the theory of objectivity in the 1st part of the book, which is free from the limits of a subjective and relativistic observer standpoint. That isn't a dead end and in addition explains, why all atomic clocks react sensitive to magnetic fields (magnetostriction) and what these fields have to do with gravity (see chap. 6.9)!

Today's clocks are so exact, that even differences between a clock stationed on a mountain and one at sea-level can be recorded. Even more clearly was the depending on gravitation determined at an atomic clock, which was shot in a rocket 10000 kilometres high into space. The result of the analysis without doubt was, that the clock in that case doesn't „tick" correctly anymore.
But what does theoretical physics say about it? It claims, here the ,,red shift" has been measured; it thus concerns a confirmation of the special theory of relativity. But since it concerns a clock experiment and not a light signal, it clearly contradicts this theory, which isn't able to describe any gravitational effect at all, as is well-known. For this case in the spheres of theoretical physics one helps oneself with the general theory of relativity, with which actually only would be proven, that the two theories from the legacy of Einstein completely incompatible contradict each other. We come to the following conclusion:

Whoever gives details about length or time, is obliged to also indicate the reference system.

He also has to inform where his laboratory is situated and with which devices he measures!

With the conclusion also the discussion would be opened. Throughout the last four chapters a main idea can be found. It should be worth, to again think about the brought forward arguments and to dare a comparison with text books.
In the text books there doesn't exist such a thing as an oscillating interaction. Here no answer is found to the question, why the solar system isn't hurled out of the galaxy as a result of the high velocity, why the inside of the earth is hot, how the geomagnetism is formed, why the continents drift and why the ocean floor nowhere is older than 200 million years, as samples from the ocean floor prove<i>. How would you answer these central questions?

<i>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991).

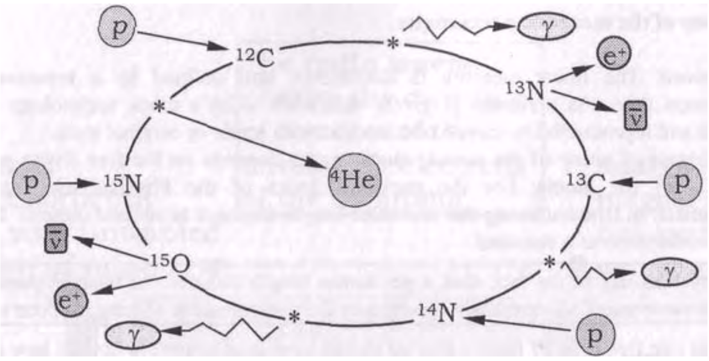
p
^{12}C
γ
e^+
^{13}N
$\bar{\nu}$
p
^{13}C
^{4}He
p
^{15}N
$\bar{\nu}$
^{15}O
γ
e^+
γ
^{14}N
p

Fig. 13.8: The Bethe-Weizsacker cycle (concerning the sun's fire)
(meaning of: ^{12}C = carbon nucleus, ^{4}He = helium nucl., γ = gamma quant, ^{14}N = nitrogen nucleus, p = proton, e^+ = positron, ν= anti neutrino, ^{16}O = oxygen nucleus

---

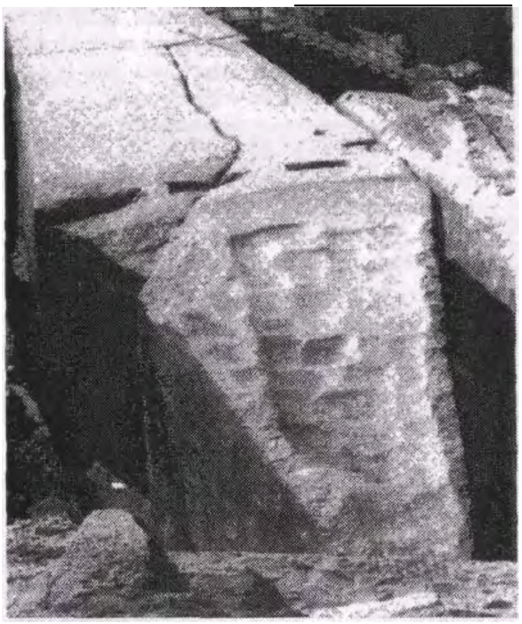

Fig. 13.9: The top of the unfinished obelisk in Assuan<i>

<i>: Hermann Wild: Technologien von gestern, Chancen fur morgen, Jupiter-Verlag Bern (1996), ISBN 3-906571-13-0

13.8 Discussion of the cosmological insights

The numerous models which are offered, of a geodynamo, an iron core, of assumed zones of subduction and of plate tectonics may be helpful for the explanation of individual isolated phenomena. But they physically don't give a uniform picture and partly contradict each other.

Lord Kelvin had calculated a period of shining of 5000 years on the basis of a sun consisting of coal<i>. The german physicist Hermann von Helmholtz landed at 15 million years, in which case the sun yearly should shrink for 100 meters. He already assumed a temperature of 15 million degrees centigrade, as also the Bethe-Weizsacker cycle has as a prerequisite, which according to today's concept should describe the process of nuclear fusion taking place in the inside of the sun (fig. 13.8). The only thing is that this extreme temperature is completely incompatible with the high density in the core of the sun! There are more than good reasons to assume the opposite of the widespread textbook opinion and assume that the core of the sun is cooled by the collected neutrinos in the same manner as the inner core of the earth and that superconducting areas are formed, which powerful fields even cause the protuberances on the surface of the sun. Since in the sun no measurement is possible, with which a model could be verified or disproved, terrestrial arguments naturally suggest themselves. Here the pieces of evidence for the growth of the earth can be taken in the hand and photographed. We already have discussed some facts. But there are found a multitude of other ones, for instance from the domain of archaeology, for which text books of today until now as well provide no explanation.

Possibly the stones and monoliths weighing several tons, as they were used for prehistoric buildings, should be linked with the growth of the earth. If the earth was smaller at the time they were build, then they perhaps by no means were as heavy as today!
Then the stones for reason of the smaller density in addition were softer and with that easier to work on. Dr. Wild points to the building technical peculiarity, that the stones formed like cushions originally must have been soft. He proves with the photograph of the top of the unfinished obelisk in Assuan (fig. 13.9), that the traces of working stem from a spatula in a plastic mass.
Also the perfect fitting of the stones used for the building of the pyramids can only be explained in this way. In the joints not even a knife point can be inserted!

If even stones in the course of time increase in density and hardness, then it is easier for us to comprehend, how small rivers in past time could dig large deep valleys in the earth's crust, then we perhaps also understand, why very old bones today are petrified<ii>. It is obvious, that also bones in the course of time increase in density and hardness. Even if science should succeed in breeding living dinosaurs, then their chance to survive in spite of that would be equal to zero, because the dinos would collapse under their weight of their own of several tons. Their bones would be much too thin and brittle for their weight of today!

<i>: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991)
<ii>: The indication to petrified bones stems from a participant of the seminar.

25-07-1996: Assistant head of government department
Acceptance of the common standpoint.

11-12-1996: European parliament
Resolution of the guideline.

commission
Statement concerning proposals of change of the european parliament.

19-12-1996: Assistant head of government department
Resolution of the guideline.

19-02-1997: Coming into effect of the guideline
Publication in the official paper of the EC.

19-02-1999: Period of translation of two years

Reduction of the threshold values for current consumers:
19-02-1997: to 40 million kWh (opening of the market 23%)
19-02-2000: to 20 million kWh (opening of the market 28%)
19-02-2003: to 9 million kWh (opening of the market 33%)

until 2006: „Anti imbalance clause"
A nine years period of transition to preserve equal opportunities in the competition.
Report of the commission.

19-02-2006: Further stage of the liberalization
European commission tests by means of the made experiences, if the current market should be opened further.

Fig. 14.1: Timetable of the EC single market guideline electricity[i]

<i>: Grawe, S. Thiele: Vorbereitung der Stromversorger auf den Wettbewerbsmarkt, ATW Atomwirtschaft-Atomtechnik 43. Jg. 1998, Heft 1, S. 10 - 13

14. The way towards free energy

Can energy actually be produced? No energy supply enterprise is capable to do that. Fact is, that energy only is converted and not produced. For that the available resources on earth are tapped and brought into a utilizable form of energy. As different the conversion processes may be, finally always heat is formed. We thus gradually burn the globe, on which we live. How long can something like that go right?

Science gives all-clear: „stock of energy is sufficient for the next 100 years. Newest calculations disprove the fear of a scarcity of energy on earth. There exist large stocks of coal and oil"<i>.

But what are 100 years compared to the age of the earth? How will our descendants judge our thinking and acting? They will condemn it and curse us, that much is clear already today.

We have an obligation to preserve the environment, and we only will be able to fulfil it, if we look to nature, how it covers its need of energy, if we finally understand and meaningfully copy nature.

We still are miles away of the goal.

14.1 The liberalization of the energy markets

It is important in the interest of a member of the executive that the stock of energy doesn't draw to an end, as long as he carries the responsibility for the enterprise of the economy of energy. The interest in the environment and the environmentalists for obvious reasons is less distinct. The problems are more of a commercial kind. Concerning that an actual example is given.

A special kind of problem is the meanwhile Europe wide valid decree of the EU concerning the liberalization of the energy markets (fig. 14.1). The list of the consequences starts with the fact, that the concluded licence contracts for power supply between energy supply enterprises and the communities, which are valid till 1997, are only wastepaper. Every customer of electricity as of now can conclude an individual contract with every "producer of electricity".

So an environmentally aware Black Forest person decides to obtain his power from a wind power station, which is situated in Denmark. That truly sounds very liberal. But how should that function, if there is no wind at all in Denmark for a week, or the connection is interrupted by a flash of lightning? The windmill further delivers power and the runner writes its bill truly believing, its power has arrived. The consumer then reads the bill at candlelight and puts with understandable anger his claim for compensation together. Pointing to the fact that such difficulties can be solved by book-keeping, the consumer again is calmed down; but that the not by contract obtained power will cost him dear, is another story.

<i>: Illustrierte Wissenschaft Nr. 6, Juni 1996, page 48-51.

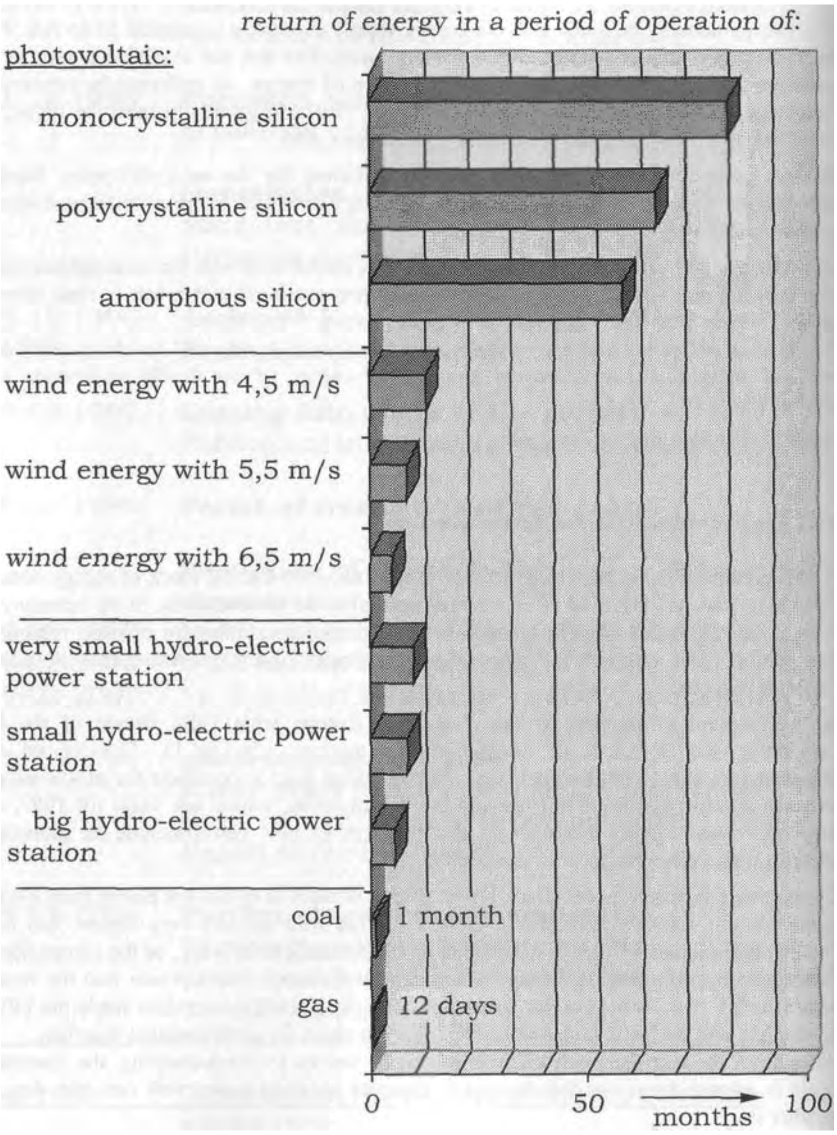

Fig. 14.2: Average time of paying for itself for the spent energy (return of energy)

14.2 The consequences of the liberalization

With the EU decree the monopoly only has shifted towards the runners of the distribution nets, after all only one net is present. Competition however would require at least two nets, thus a doubling of all house connections and all high-tension pylons, but that fortunately is unrealistic.

To prevent fleecing of the consumers by means of the net monopoly, politics introduces in the place of the free market economy dirigisme and plan economy with the well-known concomitants: no-one will look after the existing nets, no-one feels responsible anymore, because after all they have become public good by decree. One thus lets the pylons rot slowly and repairs only in emergencies. This possible development really can't have been the good intention of the EU.

In the next few years we however will be able to observe for the producers of energy exactly the development, which the EU-commissioners have imagined: Total competition contest, price war and a struggle for power to survive economically. A chance only has the nuclear power station, which gets rid of its refuse cheaper and if need be even illegally or which lives of indirect state subsidies, or the brown coal power station, which increases its efficiency at night, when no-one watches, by switching off the expensive filters.

The first power stations, which are selected out by the liberalized energy market, are the hydro-electric power stations which stand closest to nature. They simply are too small and too intensive of personnel, to be able to survive.

Gas turbine power stations, which deliver the power for half the price, than are ranking first. Then there is no place for regenerative systems anymore.

Solar energy, how many roofs a supporting program may have, stays a toy supported by the state. For a photovoltaic installation the „Return of Invest" still lies at more than 80 years, whereas is reckoned with a theoretical life of 20 years. In practical use on the other hand photovoltaic installations occasionally already have failed after seven years, after the photocells had gone blind. In this case not even can be talked of an ecologic energy, because the return of energy lies still above that. The supposed Idealist who spoils his roof with photovoltaic to reassure his ecologic conscience, would have spared the environment more, if he had covered his need of power from the socket, because already the production of the photovoltaic installation gobbles up more energy than can be produced with it. With regard to the environmental compatibility the ecologic balance sheet of a power station is attached a central importance (fig. 14.2).

With the law of feeding in power the state intervenes dirigiste, supposedly to protest the consumers. With this law the state orders, that not it, but the energy supply enterprises have to take the subsidizing of the regenerative energy, that they have to take over the power delivered at the inappropriate time in exactly so less suitable amounts at a price, which exceeds the market value by a factor of two and for falling prices of power even by a factor of four. With the law the politicians very fast and without agreement of the effected enterprises have shifted the „black Peter" further to these, which will get problems to preserve their competitiveness in the international comparison with the subsidy duty. It can't have been the intention of the EU-commissioners, that on a national level the price of power increases to finance some energy technical playground.

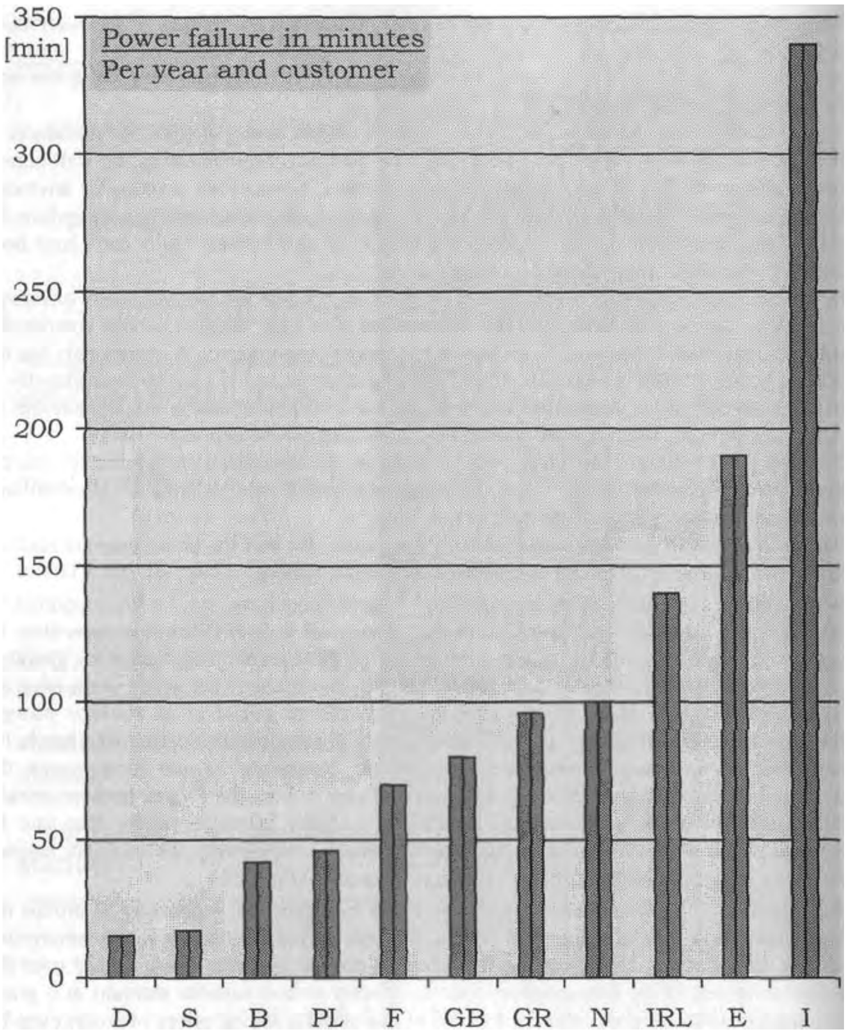

Fig. 14.3: Power interruption and not being at disposal

<i>: ETG, Energietechnische Gesellschaft im VDE: Qualitat der Stromversorgung, Dialog Nr. 1, Jan./Feb. 1998, S. 21

14.3 The chances of the wind energy

Meanwhile, as a result of the law of feeding in power, no longer the politicians but the energy suppliers are the ones, who around the North Sea look out of the window every day with the worry, some wind could blow and the mills could turn, because every kilowatt hour of a windmill must be subsidized strongly. Every windless day however reduces the power bill, with that helps the consumer and raises the chances of competition of our economy.

If it should be neglected, to tip over the law of feeding in power in due time with help of the EU, then the consumers will get their power in future abroad, then at the Preussen Elektra, damaged most by wind energy, as the first the lights will go out, then economic power and prosperity in Germany in future are dictated from abroad.

But if the law is dropped, then with that the duty to subsidize the regenerative energy carriers drops back to the state. Now all tax payers may foot the bill, even those, who don't use any power at all. The well-known Justice of subsidy" comes into effect.

Without support by the state only few types of power stations will be left behind. In addition are overcapacities being reduced, because they only cost money. But both is at the expense of the reliability of the delivering of energy and of the safety of the consumer. We owe the high stability of our network of today the large number of most different providers of power, which cover the basic load up to the peak load according to their suitability (fig. 14.3). But let us not talk of the golden past. The network after all isn't able to store power. From that follows, that without redundancy and without free power station capacities which can be activated at any time a short overload is sufficient, to let the network collapse.

Once the EU guideline concerning the liberalization of the energy markets is in effect, when persons selling power wend their way from front door to front door, to convince the housewives to obtain the power from them and not from other hawkers, when the power stations only live of stock and the depreciated overhead power lines only are entered in the hooks with the scrap value, then we should be dressed warmly and always have ready sufficient candles <i>.

14.4 The chances of being self-sufficient concerning energy

The only way out is the decentralized energy supply, the getting out of the large energy union and the way towards being self-sufficient. The argument sounds convincing in view of missing alternatives. But now it no longer is possible to plug the plug in the socket and then switch on whenever it pleases us. First it has to be calculated, if the windmill or the installation for solar energy supplies sufficient power or if a cold meal should be made. In contrast to today's consumer habits the runners of such installations will have to adapt their need of energy to the prevailing weather conditions.

<i>: A baby-Boom, like after the big power failure in New York, would be the smaller evil. The inhabitants and shopkeepers in the New Zealand metropolis Auckland will long and frightened remember the power failure in february 1998, which lasted weeks and was a result of not carried out maintenance works of high-tension pylons.

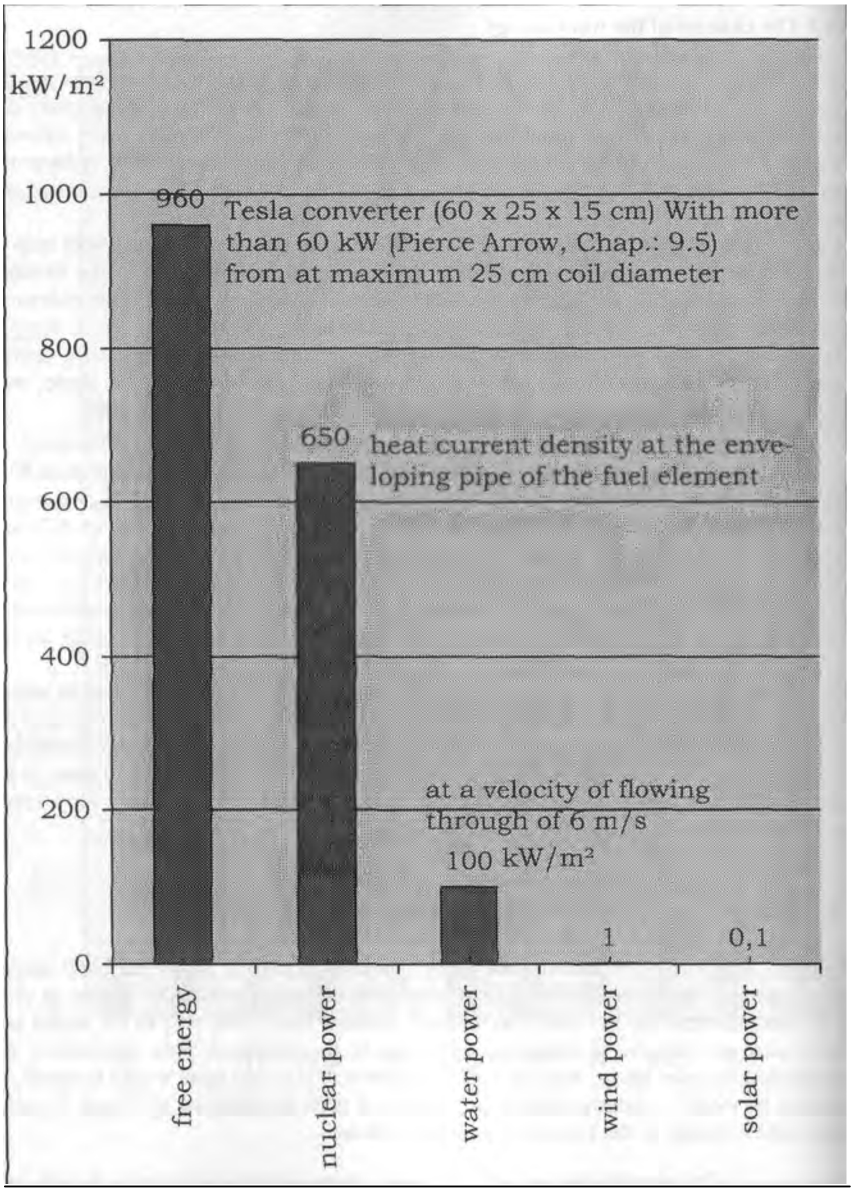

Fig. 14.4: At maximum obtainable power density of different carriers of energy

<i>: ETZ, Heft 19/1995, S. 40

Block heating power stations, propagated as stand-alone solution for people being self-sufficient, deliver power and heat at the same time. But if I don't need any heat on a hot summer day, then also no power is available or I uselessly heat into the open. For that one may leave all electric consumers switched on in winter, even if they aren't used at all, only to get the hut warm. Does a stand-alone solution look like that?

The situation truly is demotivating. Even the energy suppliers meanwhile may have realized that the energy politic way is a dead end. But for real alternatives in energy technology the pressure by suffering still doesn't seem to be big enough.
The intention to learn of nature is present in principle. The solar fire one wants to kindle on earth in a fusion oven, but the oven does not as it should.
Obviously the sun functions completely different, as physics imagines today (see fig. 13.8).
Before copying stands understanding, and in there seems to be a hitch!

14.5 The space energy technology of nature

Also nature needs energy, even very much. But it hasn't got any connecting pieces for tanking and no oven lid to fill in the fuel, it doesn't know our ecologically harmful combustion technology and environment destroying explosion technology at all. Nature rather works with the opposite, with implosion and fusion.
The sun, we have derived, materializes the matter which it needs for growing and shining from the neutrino field. The earth and other planets imitate the sun.
The concept is simple and convincing. The source of energy lies in the air and mustn't be dragged about in tanks. Collected and materialized is just as much, as is needed at the moment. In that way the resource energy is spared. In addition there can be done without any sort of storing. In addition it can show a substantially higher power density, than all today known and used energy carriers (fig. 14.4).
Such a source of energy solves all described energy problems at once. Nature wouldn't be as we know it, if it wouldn't have this ideal energy, also called „free energy". The balance sheet of energy alone brings it to light, because as a rule it doesn't work out for biological systems. Often more energy is released than is taken up by the food.
In that case some migratory birds materially seen should have completely used themselves up before reaching their destination, if the energy necessary for the flight would be of purely material nature. From a concrete example the following is reported<i>: „migratory birds have - depending on kind - a maximum range velocity between 24 and 83 km/h and at their Atlantic flights no opportunity for an intermediate landing. They are thousands of kilometres on the way and hardly lose weight. For instance an Albatross with a body length of up to 1.20 meters and a wing span of up to 3.50 meters uses per kilometre only 8.5 grams of weight at a non-stop flight. How is that possible without additional supply of energy?"

<i>: K. E. Rathgeb: Wie man die freie Energie anzapft: Vogel machen es uns vor, Raum & Zeit 79/96, S. 74

Fig. 14.6: Freiherr von Reichenbach (76 years old)

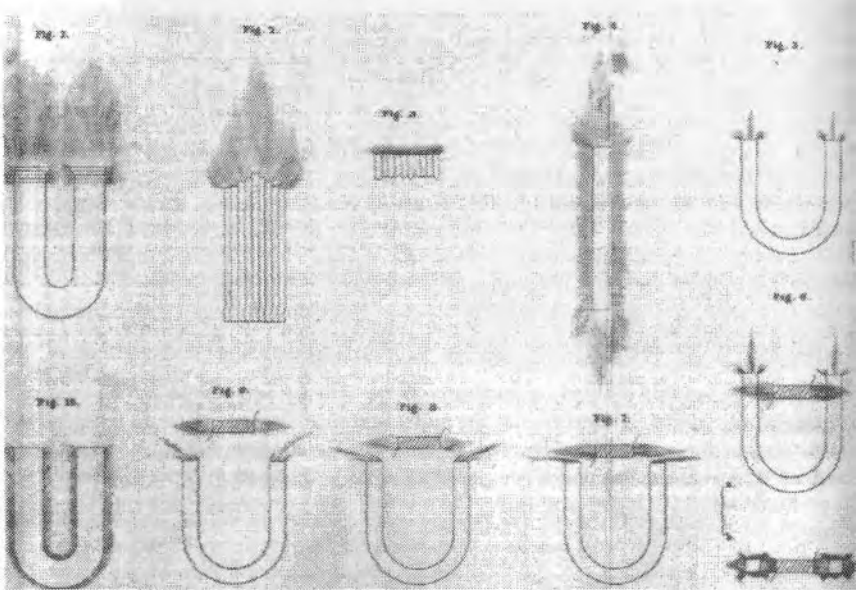

Fig. 14.7: Odic emanations of magnets according to pictures of sensitives.[i]

<i>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, Seite 13 und 15

14.6 The gap in the energy balance of man

Even man appears to tap an additional source of energy, because for reason of scientific tests with recruits these over longer periods of time are able to physically release more than twice the amount of energy they take of calories with the food. Obviously living beings don't rely on one source of energy alone. Possibly the taking up of food predominantly serves the metabolism, and the energy aspect plays only a secondary role.

This interpretation at least would be obvious, since we scoff several times a day, sort the necessary and useful building-materials out and hand over the rest to the purifying plant and nature, where some bacteria and organisms search the sorted out again for useful things. "Food chain" we call this kind of building-material trade.

If the whole had anything to do with energy or with a „ combustion process without fire ", then no animal nor any human being could do without the taking up of food for longer periods of time. But Franciscus of Assisi could fast 90 days, as is handed down to us <ii>. There exist numerous examples, which reach into the time of today.

The mitochondria, the energy centres of each cell, by no means are capable of the "combustion process", which man ascribes to them. Here in all probability a taking up of space energy independent of food takes place.

A research scientist only has to look at nature with open eyes, what unfortunately happens rarer and rarer, because the laboratory scientist always is troubled, to keep the disturbing factor ,,nature " away from the experiments.

14.7 Carl Freiherr von Reichenbach

In this context no-one can go past two natural scientists: Carl Freiherr von Reichenbach (1780-1869) and Dr. Wilhelm Reich (1897-1957).

Reichenbach called the by him investigated life energy „ Od-energy " in the style of the Teutonic God Odin. He worked with test persons, who could perceive actually invisible light phenomena and worked out the special properties of this Od-energy field with the "sensitives", as he called them.

A quotation from his work shows however that the knowledge about the life energy must be a lot older<iii> than his own discoveries: „On paintings saints often are shown with a ring-like aureole around their head, something I before this would have dismissed as a pure figment of imagination. But it was shown that this glowing ring actually can be perceived by the sensitives as an Od-phenomenon and so the aureole obviously can be traced back to real impressions of particularly sensitive persons."

Reichenbach also found out that water has a big endeavour to take up this Od-energy or in the language of the present day, to absorb the field energy. This circumstance we find confirmed in technology, since water absorbs high-frequency waves, whereas an insulator or a vacuum lets them pass through. But without vortex physics it however remains entirely unclarified, why!

<ii>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 110

<iii>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 16

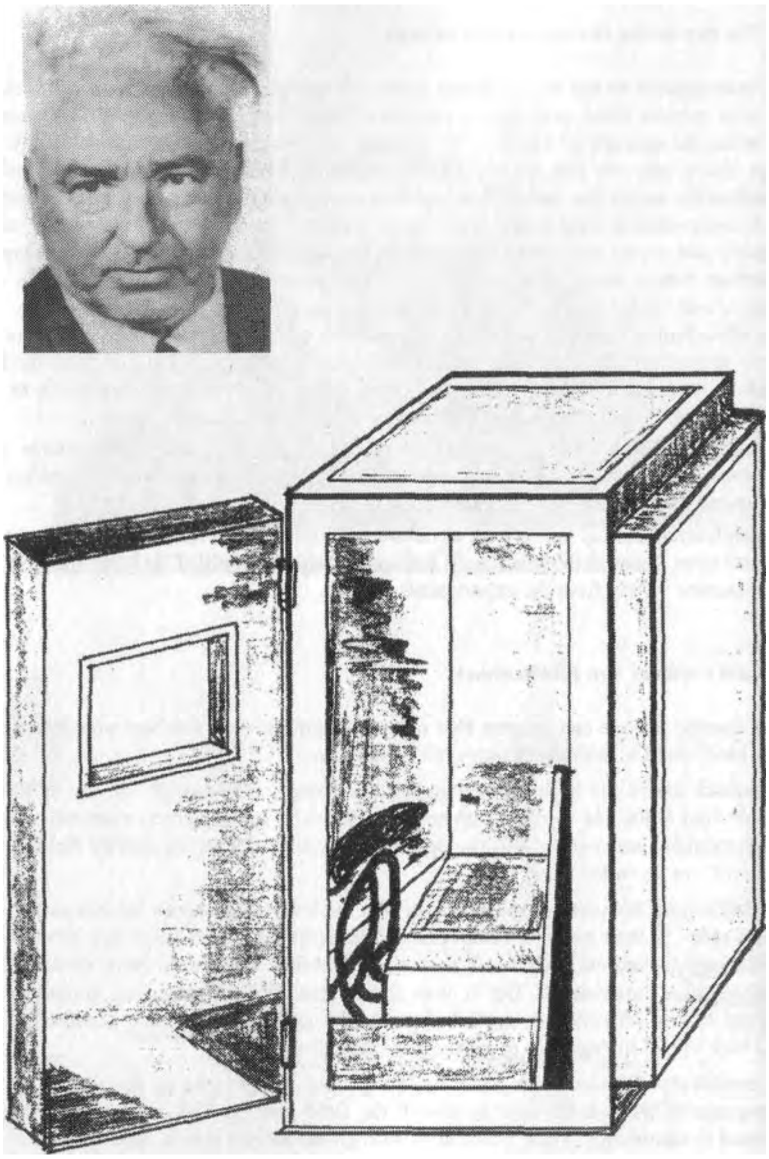

Fig. 14.8: Wilhelm Reich and the model of an Orgon accumulator
for applications of the whole body.<i>

<i>: taken from: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, pages 38 and 89.

14.8 Cold fusion and Genesis

Chronologically the doctor and psychologist Wilhelm Reich followed in the footprints of Reichenbach. He merely altered the name for Od and spoke of Orgon. His speciality was the accumulation of Orgon radiation, I would say the focussing of neutrino radiation. Actually the properties like e.g. the missing possibility to shield Od, Orgon and the neutrinos are identical, so that we can proceed from assumption, that it also concerns the same physically.

We already talked about Wilhelm Reich (part 1, chapter 9.2). He could show that the measurable temperature in a closed box, constructed like an „Orgon accumulator", is increased in a mysterious manner without supply of energy from the outside. He even could prove, that this energy actually concerns the sought-for life energy, as he observed the creation of life in the laboratory under the microscope. For that he cooked muscle fibres, vegetables or other cells so long until the cell structure had been destroyed entirely. But from the educts entirely by itself new living beings, like protozoa or algae, were formed<i>.

Reich at his microscope actually followed the transition of dead to living matter. What is of interest here, at first only is the energy technical aspect of this conversion. Later we will occupy us with the at the same time occurring information technical aspect.

Worth mentioning seems to me the experiment with the „silly" chickens, which have at their disposal astonishing abilities besides the laying of eggs<ii>.

The test chickens were handed chicken food, from which to a large extent all calcium had been extracted. But the chickens showed themselves unimpressed and further laid keenly their eggs. The experimenters were surprised, where the chickens actually got the lime for the egg shells. How solves such a chicken the problem of raw materials? For that further materials were extracted from the food and look, at the removing of silicon the laying of eggs was over. The experiment actually only allows the conclusion, that the "stupid" chicken is capable of a cold fusion, that it itself „produces" the necessary calcium from silicon presumably by using carbon. Every alchemist here has to go pale with envy. But what says the research scientist of fusion to that, who actually should know, how fusion functions? After all he is paid for it by the tax payer! The chicken uses the fusion already today and the other living beings presumably also, but for that energy is needed and the balance sheet should work out!

The neutrino radiation therefore has to be factored into the balance sheet of energy. If the balance sheet then works out, it could be proven with that, that here neutrino energy is put to use. In addition the process of the conversion of neutrinos has to be investigated, which surely has something to do with the frequency and the wavelength of the radiation. After all a child has cells exactly as big as an adult. It only has less cells! So that an interaction can occur, the cell size has to fit in with the wavelength, which obviously is very small, presumably in the range of the radioactive radiation, with which circumstance the biological incompatibility with this electromagnetic wave would be explicable.

<i>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, S. 39

<ii>: Louis Kervran: Biological Transmutations, s.a. J. Heinzerling: Energie, S.278

1. Primary reaction
   (light reaction in the centre of reaction):

   ring-like vortices/neutrinos are being collected
   electrons are formed
   energy is produced
   e^- is attracted by the water dipole
   water molecule is being splitted (photolysis):

2. Secondary reaction (dark reaction):

   carbon dioxide is reduced with the hydrogen splitted off from the water to carbohydrate

---

Reaction equation of photosynthesis:

	6 mol CO_2	(carbon dioxide)
+	12 mol H_2O	(water)
+	675 Kcal	(vortex energy, light)
	1 mol $C_6H_{12}O_6$	(glucose)
+	6 mol O_2	(oxygen)
+	6 mol H_2O	(water)

Fig. 14.9: Concerning photosynthesis

14.9 Photosynthesis

The materialization of free electrons is a prerequisite to start the photosynthesis. The normally used explanation, the free electron necessary for the splitting of the water molecule was knocked out an atom by light, doesn't seem to be correct, after until now all attempts of a technical realization according to this model concept have failed. At the photosynthesis the plants obviously help themselves with the neutrino radiation, which according to an estimation of today with 66 billion particles per second and per square centimetre might be more than sufficient for a green earth.

If we put a seed in a water glass, then a plant grows from it and forms small leaves, which get bigger and bigger. A wonder of nature, we say. Where does it actually get its building-materials? From the water or from the air? Necessarily the plant obviously produces a part of the matter itself!

Experts think they have understood the process of photosynthesis: Take light and water and carbon dioxide and handicraft from that sugar and oxygen. But from where does the plant take the necessary energy for the rebuilding and the splitting of the water molecule, the photolyse? The taken up solar energy hardly is sufficient for that, especially since the plants only absorb about 1% of the photosynthetic utilizable sunlight incident on earth<i>. By means of reception molecules, which look like small antennas, pigments less than 30 nm in diameter, such is the level of knowledge, the sunlight is collected and led into a photochemical centre of reaction. Here the reaction should take place, provided that an electron set free by the light jumps into the middle of the centre of reaction. But exactly this favour the electron doesn't make the research scientists, who want to imitate the process. Copying nature still doesn't succeed.

The mistake presumably lies in the circumstance that the light doesn't set free any electron at all. The electron actually first of all is produced in the centre of reaction. By means of the antennas a neutrino vortex is collected, which at first occupies the entire space, to afterwards contract to an electron, which as a result automatically is centred in the centre of reaction.

In the process of materialization at the same time the necessary energy of the process is formed. The reaction equation after all also has to work out energetically, because the plants doing so get neither hot nor cold (fig. 14.9).

Because curiously the light reaction even can be observed in the dark<ii>, one could be inclined to in principle call into question the influence of light on the photosynthesis. But such an influence nevertheless seems to be present, after all does a plant react on the irradiation of light and changes its spectrum of absorption through its colour. But it for sure is another influence and not the one, which one attributes the green plants today!

It would be important to finally understand the way of functioning. The plants and particularly the algae are the ones, which actually first have made possible life on this planet with the photosynthesis, the most original form of a production of matter and energy.

<i>: Luttge, U. u.a.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, S. 136

<ii>: Luttge, U. u.a.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, S. 126

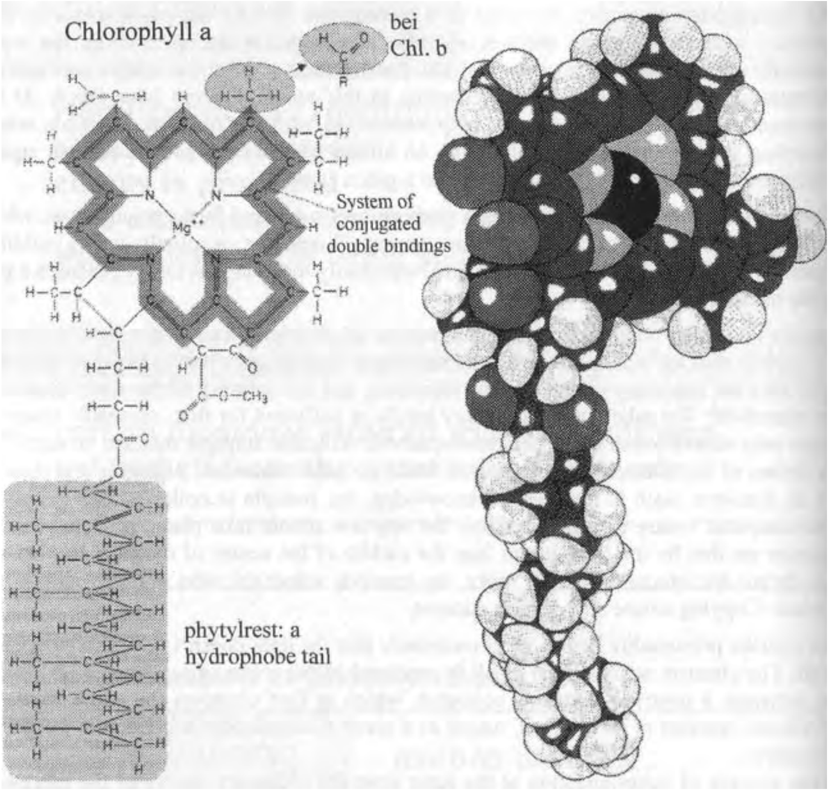
Chlorophyll a
bei
Chl. b
System of
conjugated
double bindings
Mg++
phytylrest: a
hydrophobe tail

Fig. 14.10: Formula of structure and spatial model of the photosynthesis pigment chlorophyll.<i>

<i>: Luttge, U.et al.: Botanik, VCH, Weinheim (1994) ISBN 3-527-30031-7, p. 118

14.10 How nature materializes

The structure of the receiver antenna allows us a deeper insight into the manner, how free energy is tapped during the photosynthesis. For the dimensions determining the frequency of the photosynthesis pigments, as the antennas are called, we have to go down to the molecular structure. Fig. 14.10 shows the formula of structure and next to it also the spatial model of the pigment chlorophyll[i].

Two points point at the function for free energy conversion. On the one hand is situated in the centre of the molecule a double positive ionised Mg atom, surrounded by four nitrogen atoms and a carbon ring consisting of 20 atoms, from which arises a polarization of the entire molecule. This in addition is oscillating because the electron cloud of the enveloping electrons, which hold the molecule together, perform swirl oscillations depending on the temperature. With that chlorophyll is able to go into resonance with oscillating neutrinos.
A role play the unipolar field configuration and the effect of resonance of the molecular oscillation of its own forming as a result of the polarization. A further role in addition seems to play the spatial structure.

The model of the „receiver antenna" chlorophyll taken out of a textbook[i] and shown in fig. 14.10, consists of a stalk and a spirally wound head, which resembles a Lituus or crook, which Etruscan and Roman Augurs have taken in the hands for land surveying, a precursor of the crosier (see fig. 16.10). This again has the form of a Tesla coil and that, as already derived, is able to withdraw rotational energy from the collected neutrinos (chapter 9.8). Doing so free electrons are materialized, and these then start the process of photosynthesis. An explanation concerning the way of functioning of the antenna pigments here for the first time is getting available.
By the way also the mitochondria, which form the energy centres in every cell, have as well the form of a Tesla coil. Whoever wants to understand the energy economy of a cell or the photosynthesis first should occupy himself with the Tesla coil (chapter 9.8).

The open question, how land surveying should be possible with a Tesla coil, we at first still have to shelve, because in this chapter it concerns the way towards free energy and the chance to learn of nature. Examples to be looked at are on the one hand the core of fusion in the inside of the earth and on the other hand the humus layer on the surface of the earth, which has been materialized in the course of time with the help of photosynthesis.

The goal seems to be worth striving for. If we in the first place have learned to produce energy exactly like nature, then we'll further try to produce matter purposeful, with which ageold alchemist dreams could be fulfilled. We wouldn't need to scrape and to search in some mines any longer. We would materialize the products without refuse, naturally and just for the environment, direct in the final form. I admit that at present it sounds pretty futuristic.

With another example, the lightning and particularly the ball-lightning, the collection concerning the use of free energy in nature shall be completed.

---

[i] U. Luttge, M. Kluge, G. Bauer: Botanik, 2. Aufl.- Weinheim, VCH (1994) ISBN 3-527-30031-7, S. 118

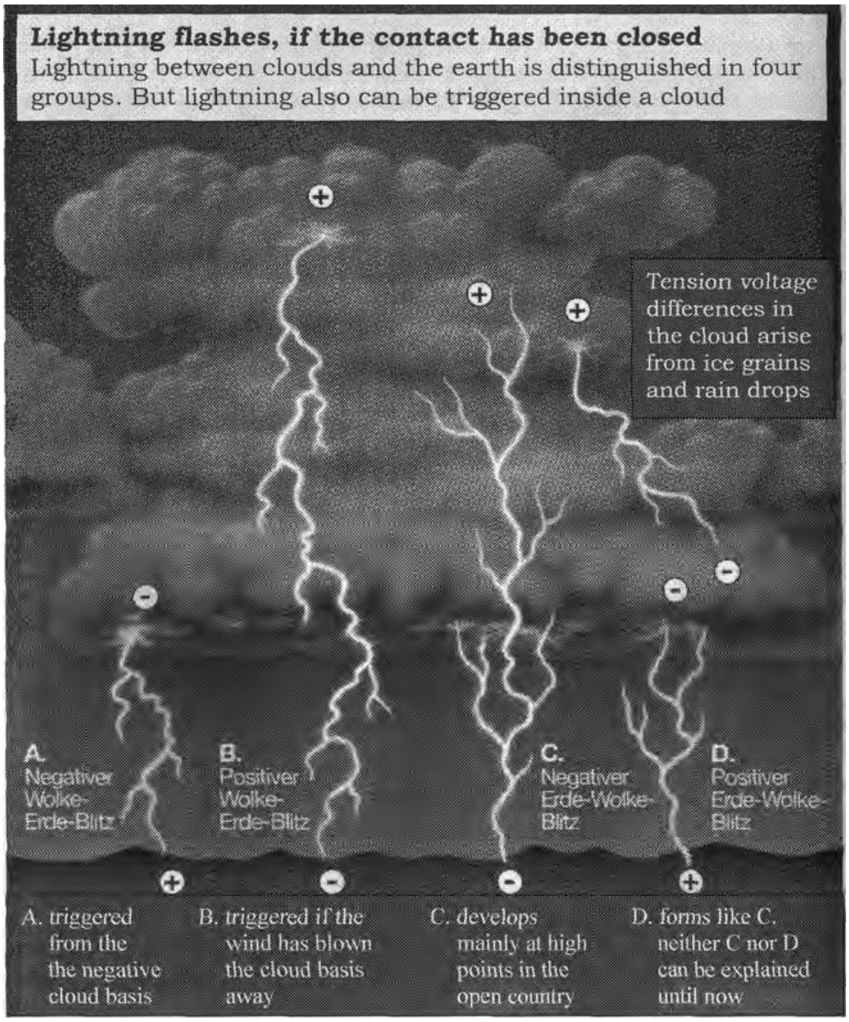
Lightning flashes, if the contact has been closed
Lightning between clouds and the earth is distinguished in four groups. But lightning also can be triggered inside a cloud
Tension voltage differences in the cloud arise from ice grains and rain drops
A. Negativer Wolke-Erde-Blitz
B. Positiver Wolke-Erde-Blitz
C. Negativer Erde-Wolke-Blitz
D. Positiver Erde-Wolke-Blitz
A. triggered from the the negative cloud basis
B. triggered if the wind has blown the cloud basis away
C. develops mainly at high points in the open country
D. forms like C. neither C nor D can be explained until now

Fig. 14.11: Concerning lightning.<i>

<i>: taken out of: Illustrierte Wissenschaft Nr. 8, August 1995, Das unverstandene Phanomen der Blitze, S. 13

14.11 Lightning

Lightning is a spectacle of nature just as imposing as unsolved. It concerns an electric process of discharge, in which to the amazement of all experts arrive for several powers of ten more charge carriers at the surface of the earth, than hefore were contained in the cloud! Furthermore unsolved is, why lightning glows. Let us start with the open question for the difference of potential necessary so that the air is ionised and a lightning channel is formed.

If at first the electric tension voltage of 200,000 volts between the surface of the earth and a height of 10 kilometres is available, which according to our calculations results from the rotation of the earth and the magnetism of the earth (fig. 11.8). Another source of tension voltage is not known. By air movement and supposed processes inside a cloud locally an additional accumulation of charge carriers may occur, so that in the case of a thunderstorm the by Tesla assumed, twice as big value doesn't seem unrealistic<i>.
On the other hand a tension voltage between 4,000 and 10,000 volts is necessary so that a blow can occur for an air gap of one centimetre depending on the atmospheric humidity<ii>. Linearly projected the thunderstorm cloud theoretically should hover just one meter above the earth if there is lightning.
Here somehow a huge gap gapes between theory and practice!

At the latest at the spikes, the mysterious lightning, which strike out of a thunderstorm cloud upwards in the direction of the ionosphere and the still more mysterious ball-lightning it has to get clear, that the actual source of tension voltage of a lightning by no means is known. Without knowledge about the cause we'll never understand lightning.

The potential vortex theory offers an useful approach, according to which the necessary difference of potential is formed from a formation of vortices<iii>. The vortex again, as damping term in the wave equation, occurs as a result of intense sun irradiation. That explains why lightning always strike from the part of a cloud, which is the darkest, where the most sunlight is absorbed and the damping is the largest.

The possible formation of ice in a lightning channel is a further confirmation for the correctness of the vortex explanation. After all it has been derived that contracting potential vortices withdraw heat in principle (fig. 12.8).

But now we also want to know from this efficient theory, why photons and electrons are formed during the lightning and where they come from. Here obviously energy is formed by means of materializing of vortex particles.

<i>: R. L. Clark: Tesla Scalar Wave Systems, The Earth as a Capacitor, The Fantastic Inventions of Nikola Tesla, ISBN 0-932813-19-4, S. 265

<ii>: Karl Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Auflage 1988, ISBN 3-540-18403-1, S. 221

<iii>: K. Meyl: Potential vortices, Part 1: Discussion contributions to the natural scientific interpretation and concerning the physical-technical usage, basing on a mathematical calculation of newly discovered hydrotic vortices, (only in german) INDEL GmbH, Verlagsabteilung, Villingen-Schwenningen 1990.

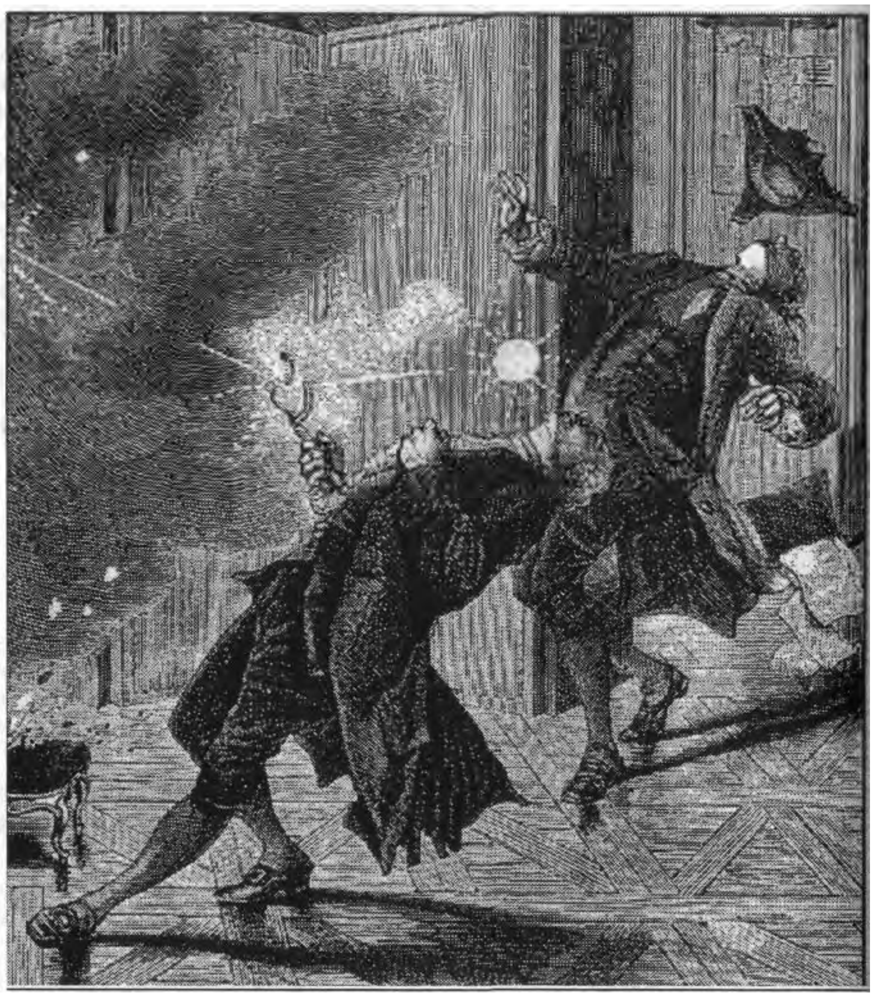

Fig. 14.12: Contemporary representation, how 1753 lightning research scientist Prof. Richmann is struck deadly by ball-lightning in his laboratory.<i>

<i>: Illustrierte Wissenschaft Nr. 8, August 1995: Das unverstandene Phanomen der Blitze, page 13

I proceed from the assumption that lightning collects and converts neutrinos. The process corresponds to the one at the sun, but on a smaller scale and only for a very short time. The lightning channel is polarized by the charge carriers. Change of temperature and field lead to a spatial oscillation, which by the way also functions as a source of sound, as anyone can hear. Taken both together lightning, seen from the outside, becomes an unipolar resonator, which is capable to attract neutrinos and to go into resonance with them. Now the predominant part is converted into electrons, because also the air molecules and air ions in the lightning channel belong to the world of matter. But it can't be avoided that a small part of antiparticles is formed, which then annihilate with particles of matter under emission of radiation. Doing so photons are emitted and lightning glows, as anyone can see!

14.12 Ball-lightning

If in the case of lightning there still exist excuses, the difference of potential preferably is traced back to neither understandable nor measurable processes inside a cloud, then at the latest in the case of ball-lightning most experts are at their wits end. Only for very simple natures explanations circulate in the direction that here for instance the organic remnants of a bird struck by lightning are burnt off.

Actually ball-lightning is observed very seldom. It is a ball flashing with reddish till blue-white colour. Its diameter lies between 10 and 50 centimetres. The glowing phenomenon can last several seconds to minutes. Doing so ball-lightning rolls over a street, temporary floats in the air, goes apparently unhindered through every wall and disappears from time to time without a trace or discharges with loud moise and formation of sparks. Some stink of poisonous gases and some also cause noise.

Famous has gotten the ball-lightning, which 1753 of all people should have struck the lightning research scientist Professor Georg Wilhelm Richmann in St. Petersburg. In his laboratory during a thunderstorm a ball of fire as big as a fist should have jumped from a iron tube to his head and should have hunted him down, so eye witnesses have reported (fig. 14.12).

Since ball-lightning has a closed structure, it has to drag about its source of energy with it. If this however consists of organic or other matter, the ball wouldn't be able to float, after all the brightness of a spherical vortex and with that the need of energy is enormous! We have to proceed from the assumption that just ball-lightning covers its need of energy from free energy and serves itself from the neutrino field.

The spherical form is a consequence of the structure shaping property of the potential vortex<ii>. Scientists are increasingly interested in this not understood phenomenon. In their experiments they try to artificially produce ball-lightning with more or less great success in the laboratory. By means of the experiment they then want to learn to understand, what the textbooks don't give away.

If we want to learn of nature something about free energy, lightning in the laboratory offers us in the form of a blow or of a spark gap relatively good possibilities. It surely is no coincidence that the father of free energy, Nikola Tesla, in his experiments almost all the time has worked with spark gaps!

<ii>: for that see in the 1st part the chapters 4.8 till 4.10

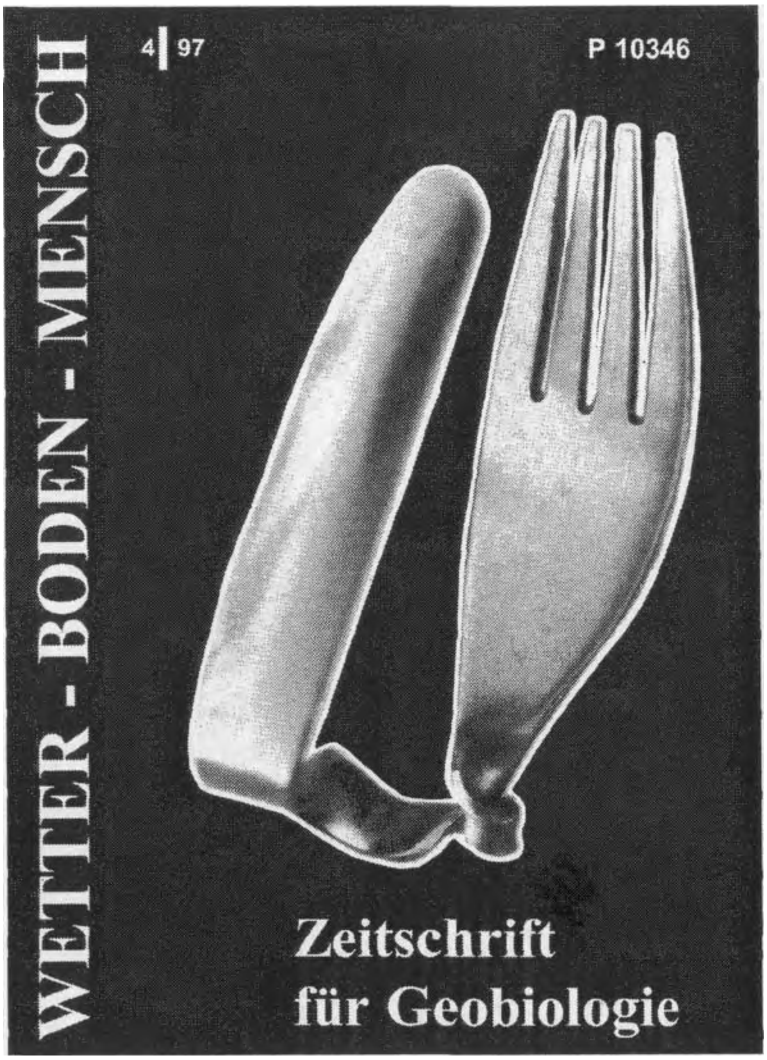

Fig. 14.13: Fork bent by „pure manual work".<i>

<i>: Whoever has to eat with bent flatware in the Hotel Adler in Waldkatzenbach (Odenwald, Germany), mustn't believe that the food didn't taste well to some rude fellow here. It is the host himself, who in a sociable round shows his guests from time to time that he not only can cook well. I could personally convince myself from the fact that Robert Hartmann doesn't use any aids. The picture shows the title page of the magazine „Wetter Boden Mensch" 4/97. A comment concerning this on page 3: Robert's 50ieth birthday.

14.13 Discussion concerning the neutrino conversion

In nature two different principles for the use of the neutrino radiation can be observed. In one case of use a glowing phenomenon occurs, in the other cases not.

The glowing lightning should be assigned to the first case. Here not only the neutrinos crossing the lightning channel are used, but also neutrinos from the environment oscillating in resonance are collected. Around the lightning channel a kind of reaction radius is formed, inside of which all synchronously oscillating neutrinos are attracted under the resonant interaction. During the discharge process the radius can increase like an avalanche, to collapse again with the transition into the stationary discharge current. A corresponding technical concept, which is discussed in the next chapter (15.5), is very efficient, but almost uncontrollable. A characteristic is the formation of antiparticles, of positrons, in the course of the materialization, which then annihilate under emission of radiation and cause a shining. Also spark gaps and fluorescent lamps possibly help themselves from the neutrino field, as still has to be to worked out.

In the second case nature works without any avalanche effect and without any glowing phenomenon. To this counts for instance the photosynthesis or the mode of operation of the mitochondria, the energy centres of a cell. All cells, whether vegetable, animal or human, only use the neutrinos which just that moment pass by them and only in those amounts, as they just are needed. They thus handle their energy very caring. Without exception electrons are materialized and no positrons.
If nevertheless unwantedly an avalanche effect occurs, something which happens fortunately only very seldom, then a self-inflammation and self-burning occurs, then it should happen, that a person burns off himself<ii>. This risk also is known of hay.

There also exist rare talented persons, who can control and regulate the process of materialization by concentration. These people can bend spoons or other metallic objects<i>. For that they concentrate themselves some time and send the materialized charge carriers into the object, which they hold in their hands. Since the metal lattice solely is kept together by the enveloping electrons of the individual atoms, the additional electrons make the metal structure sodden. Now for a short moment the metal can be bent and distorted at will. Doing so neither heat is produced nor is the colour changed. The result of the process can be produced neither by cold working with raw force nor under a flue. Also here nature shows us a technology for an ecologically compatible metal processing (fig. 14.13).

The way from the conventional over the regenerative towards free energy is predrawn. It only has to be gone! After the existence of the neutrino radiation goes as proven and 1998 for the first time concrete amounts have been determined measuring technical and published by a Japanese team of research scientists, with that also the question for an energy technical use of the particle radiation has been answered clearly. Now only the question of the mechanics is open. The Japanese research scientists by the way have found out that at night only half as much solar neutrinos can be detected than at daytime. The other half according to that is absorbed in the inside of the earth. This in the meantime published measurement result in brilliant manner confirms the working hypothesis of a growing globe (chapter 11).

<ii>: Zu Asche pulverisiert, Illustrierte Wissenschaft 6/ 1997, S. 61 acc. to an examination of the American SCI-COP in 1984.

cause/ field lines	interaction	F = force effect *mediation:*	**see chapter:** effect/ *application*
1. open **H**-field lines	magnetic (static)	F_{M0} = magnetic force (permanent magnet)	**chapter 15.4** *E.-engine,* *solenoid*
2. open **H**-field lines	resonant (oscillating)	F_{MS} = magnetic force (AC-magnet)	**15.4 + 15.5 - 15.7** lightning, *railgun*
3. open **E**-field lines	electric (static)	F_{E0} = Coulomb force *by charge carriers*	**15.9 + 15.10** atomic bond, *Testatika*
4. open **E**-field lines	resonant (oscillating)	F_{ES} = Coulomb force *by neutrinos*	**chapter 16 + 17** weak interact. galaxy bond *Tesla converter*
5. closed **H**-field lines	gravitation (static)	F_{MG} = gravity *by particles with mass*	**chapter 6.9 + 7** elementary particle mass
6. closed **H**-field lines	levitation (dynamic)	F_{ML} = reduced gravity	**chapter 18.3** gravit. waves *Casimir effect*
7. closed **E**-field lines	gravitation (static)	F_{EG} (force hardly detectable)	**chapter 18.7** *superconducting ring*
8. closed **E**-field lines	levitation (dynamic)	F_{EL} (no longer detectable)	**18.5 + 18.6** *Keely-/Searl-* *flying devices*

electromagnetic and resonant interaction		gravitation and levitation
$F_{M0} = 1{,}4 \cdot F_{MS} > F_{E0} = 1{,}4 \cdot F_{ES}$	$\gg$	$F_{MG} = 1{,}4 \cdot F_{ML} > F_{EG} = 1{,}4 \cdot F_{EL}$
for reason of open field lines		resp. closed field lines

Table 15.1: The force effect of interactions, ordered according to size, with examples

15. Principle of functioning of space energy

In this chapter we want to turn us towards the technical concepts and techniques concerning "space energy", which occasionally here and there already should have existed or have been operated with quite different success. After Nikola Tesla having pointed the direction more than 100 years ago, the way towards free energy appears to be predrawn. Never before the public interest in the topic of space energy was as big as today. Unfortunately this concerns more the collecting and gathering of rumours and speculations. As in every branch of science also here hunters and collectors can be found. But obviously the hunters, the inventors and theorists have bigger problems, to put something useful on the table. They are fighting against their own not knowing, unuseful textbooks, general ignorance, intolerance and an all-powerful energy lobby. What the collectors on the other hand come up with does make appetite but not full.

An useful and efficient theory might be the most important prerequisite just with regard to the reproducibility of an effect and the product liability of a SET-device. That's why one mustn't expect a complete list of devices of that kind in this chapter, because in the foreground stand the physical and technical explanations concerning the way of functioning, the understanding for constructive and guiding details and the learning from the mistakes and errors of the inventor.

For a better survey the possible courses of the field lines according to the theory of objectivity are listed in detail (table 15.1) and discussed from the top one after another, starting with the strongest known interaction. To complete examples and concepts are presented.

As is well-known there exist electric (E-) and magnetic (H-) field lines. Further exist open and closed field lines and finally is distinguished between the oscillating and the static case. The results are two to the third power, thus eight possibilities of combination in total. In table 15.1 all eight versions are given, even if one or another case is of more theoretical nature. For the objective of a systematizing of different concepts concerning space energy technology the taking apart in any case is helpful. The figure opposite is survey and structure at the same time for the following chapters.

15.1 The course of the field lines

In chapter 6 a relation between the course of the field of a body and its observable interaction has been made (part 1, chapter 6.7 till 6.9). Here a point of approach is offered. For instance to maximize the force effect a magnet or to optimise an electric motor, the engineers nowadays help themselves with costly programs working according to the method of finite elements. In this way they obtain a picture of the field lines, the course of which makes possible conclusions concerning the production of force or torque. The relation without doubt is given, the only question is in which order of magnitude.

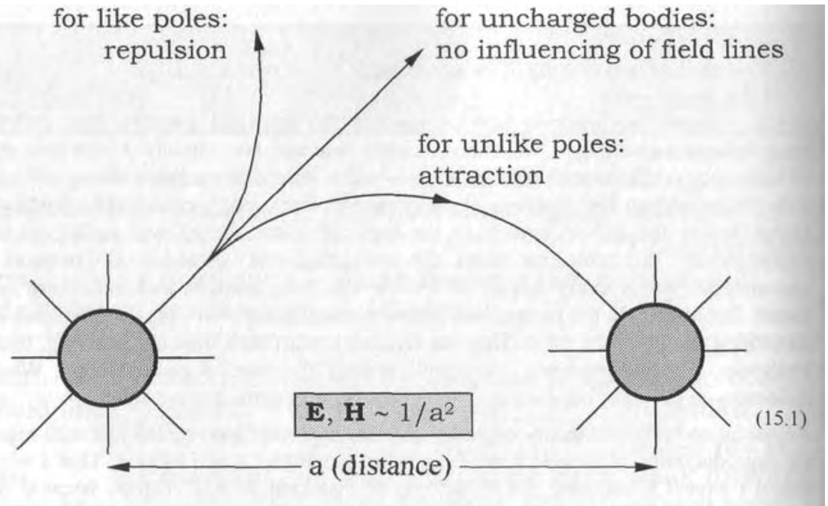
for like poles:
repulsion
for uncharged bodies:
no influencing of field lines
for unlike poles:
attraction
E, H ~ 1/a²
(15.1)
a (distance)

Fig. 15.2 A: The course of the field lines leads to a force effect (repulsion or attraction)

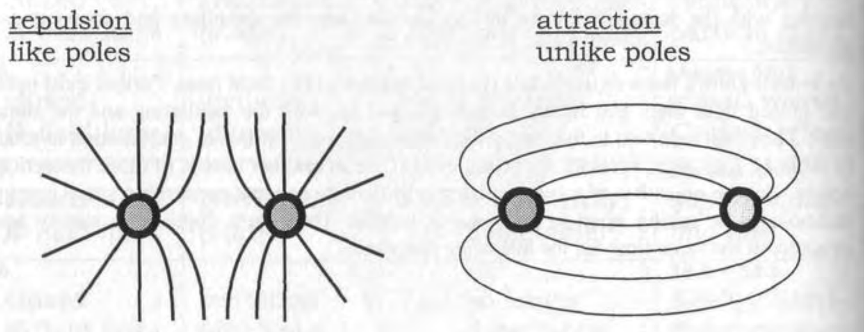
repulsion
like poles
attraction
unlike poles

Fig. 15.2 B: The course of open field lines

$$\boxed{\frac{F_{\text{Coulomb force}}}{F_{\text{magnetic force}}} = \frac{Q \cdot E}{\phi \cdot H} = \frac{\mu \cdot Q^2}{\varepsilon \cdot \phi^2}} \quad \text{with} \quad E = Q/\varepsilon \cdot A \quad \text{and} \quad H = \phi/\mu \cdot A \qquad (15.2)$$

i.e. e^-: where $\phi^2 = m^2 \cdot G \cdot 4\pi \cdot \mu$ (Gl.13.20)

$$\boxed{\frac{F_{el}}{F_{magn}} = \frac{e^2}{\varepsilon \cdot m_e^2 \cdot G \cdot 4 \cdot \pi} = \underline{4{,}2 \cdot 10^{42}}} \qquad (15.3)$$

Fig. 15.2 C: The proportion of forces at the example of the electron.

The theory of objectivity answers the question from the equations of transformation with the proportion 13.23 (fig. 13.4). According to that the electric or the magnetic field strength stand in inverse proportion to the square of a length or of a distance:

$$E, H \sim 1/a^2 \qquad (13.23) \text{ bzw. } (15.1)$$

Less mathematically expressed this, for two bodies in the distance a, where one body is situated in the field of the other, means nothing else as that the distance is reduced. Nearer to the body the density of the field lines again increases, in that way the distance further decreases and we observe an approximation.

Usually the idea of force is introduced as a factor of description and there is spoken of a force of attraction. But that not necessarily is required, because the force only represents an auxiliary description. The cause for the observed attraction rather is the spatial distribution of the field strength.

In this case the two bodies come closer and the mutually active fields get bigger and bigger, until the parts eventually run into each other (fig. 6.7 A). There one comes the thought to increase the force of attraction by an artificial compression of the field lines. In the case of the electromagnetic interaction such a compression actually takes place, since the field lines arise from one pole and end at an unlikely charged pole, which so to speak collects and bundles up the field lines (fig. 6.8 A).

At last we find out the reason, why electromagnetic forces of attraction are bigger than gravitational forces for many powers of ten (between 10^{30} and $4.2\text{-}10^{42}$ for the electron, derivation see fig. 15.2 C)! For table 15.1 this means, at the top have to stand the open field lines, which bundle up at the poles. Then very long nothing comes and after that the effects of closed field lines are being found.

For open field lines however also the opposite of a bundling up is possible. In the case of like poles the fields run away of the other pole (fig. 6.8 B). Between both a space is formed, which is free of field lines, where thus the field tends towards zero, whereas the distance between the poles according to the proportion 15.1 grows towards infinity. In this case we observe, how the bodies are going away from each other. We speak of a force of repulsion, which actually reaches until infinity. This gives reasons for the occurring of both forces of attraction and forces of repulsion in the case of the electromagnetic interaction.

15.2 Gravitation

In the case of closed field lines in principle no repulsion can occur, since no pole, neither north pole nor south pole, neither positive pole nor negative pole is able to influence the position of such a field line. This circumstance as well as the order of magnitude of a possible force of attraction suggest, to settle gravitation here.

It gladly is forgotten, that the field pointers of E- and H-field normally occur together and like in the case of the electromagnetic wave stand perpendicular to each other. It is normal to calculate only the electric field pointers for a charge carrier, without paying attention to the circumstance that the H-field is present as well. The textbooks as a rule remain silent about this dark chapter or they lapidary remark, the dual field lines are closed in themselves and hence inactive anyhow, which however is incorrect.

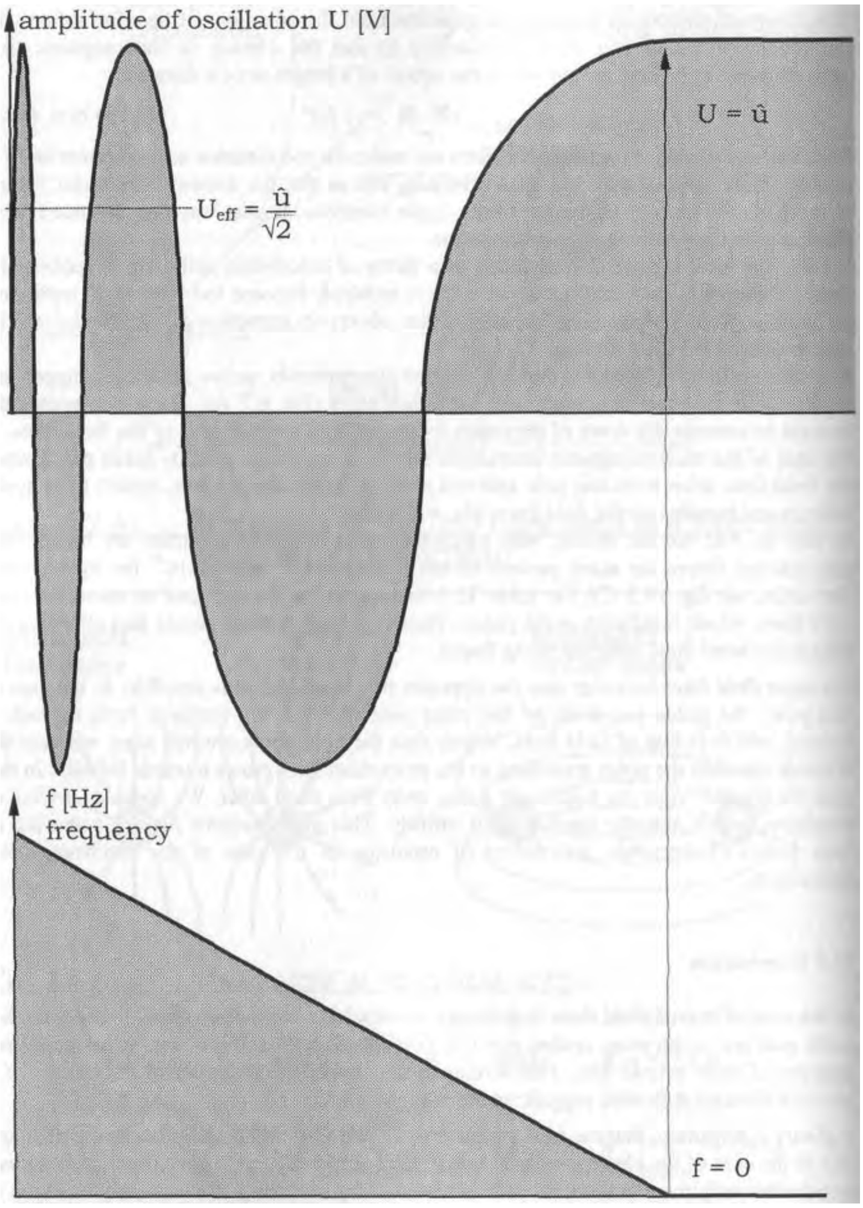
amplitude of oscillation U [V]
$U_{eff} = \frac{\hat{u}}{\sqrt{2}}$
$U = \hat{u}$
f [Hz]
frequency
f = 0

Fig. 15.3: The transition of an oscillating into a static interaction

But, according to the proportion 15.1, they actually develop a force of attraction, even if this is relatively small. No one really needs to be surprised that gravitation is not understood until today, if a whole group of fields simply is overlooked by science! We now also can explain, why there don't exist any massless charge carriers. Namely only the E-field or the H-field can form open field lines and never both at the same time. Otherwise they wouldn't be able to stand perpendicular to each other anymore. The each time other field, in the case of electrically charged bodies it is the H-field, then is wrapped perpendicularly around the E-field lines, independent of the circumstance if electrically an attraction or a repulsion occurs and without exception it forms a weak force of attraction, the gravitation.

15.3 Systematizing the interactions

Next we should know, from which field is to be expected a larger force effect: from the E or from the H-field? At the example of an electromechanical converter this question can be answered concretely.

The forces which occur and form the torque in an electric motor customary in trade are produced by magnetic poles in stator and rotor, which repel each other in the case of like charge and attract each other in the case of unlike charge. Now there in principle exists the possibility to build a motor which works with Coulomb forces, thus with positive and negative poles, instead of the magnetic forces. About such designs numerous patent specifications exist, but no customary version on the market.

The reason very simple is that a magnetic motor for the same torque is many times smaller and better priced. From this the conclusion can be drawn that for the same construction volume a magnetic force F_M is considerably larger than a Coulomb force F_E, which for instance binds together atomic nucleus and atomic hull.

Thus in table 15.1 the magnetic forces are ranked before the Coulomb forces.

Finally we have to distinguish between the static and the oscillating case, which are distinguished in the frequency. In fig. 15.3 an oscillation is shown, which by chance just at the moment of the vertex value changes into the steady state, thus takes the frequency zero. In this case the effective values between a static and a sinusoidal oscillating interaction are distinguished by the factor $\sqrt{2}$ = 1.4.

If we operate an universal motor with direct current, then it releases more power, than for a corresponding feeding with alternating current. Even a high-tension line, which stands at maximum 511 kV, is operated with alternating current up to 380 kV, with direct current on the other hand up to 500 kV. Consequently in table 15.1 the static interactions stand before the oscillating interactions.

The strong interaction naturally isn't found in the list, after it has been derived that it doesn't exist at all (see chapter 7.8), whereas the weak interaction is hiding behind the oscillating interaction. It shakes other particles so long till they fall apart (see chapter 7.13).

We now would be as far, to discuss the 8 cases listed in table 15.1 one after another by means of practical examples.

The principle of the side-pole machine

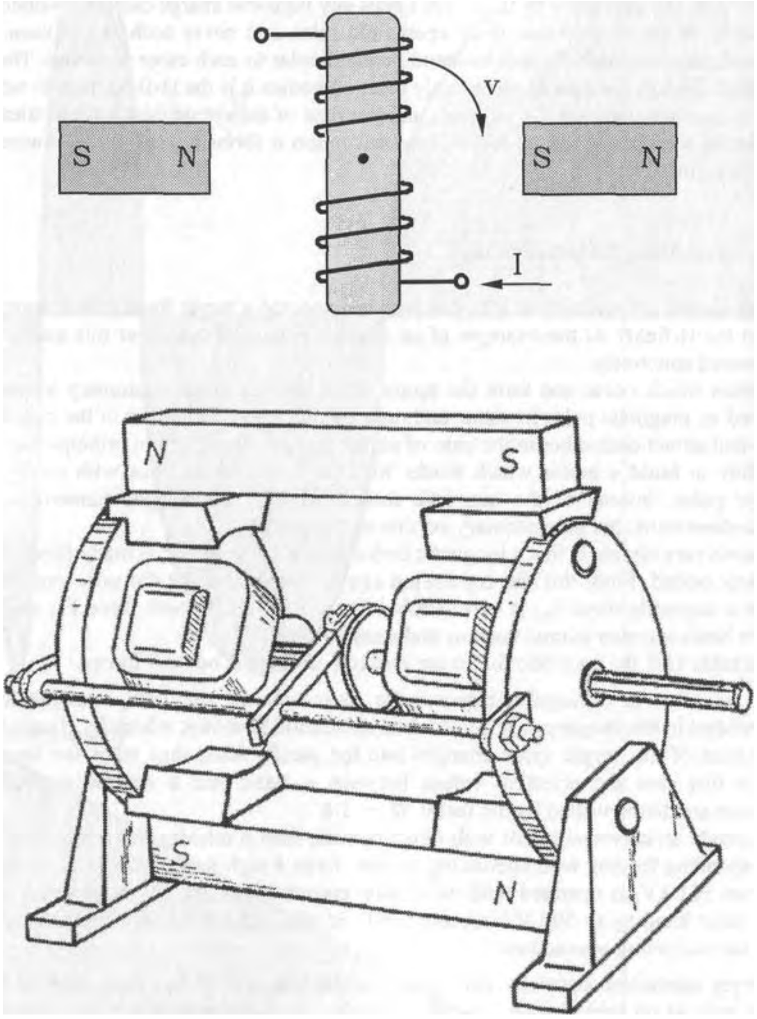

Fig. 15.4: Alternating current dynamo, according to the inventor
(Kromrey) a magnetic converter for free energy with a degree of effectiveness of more than 100 %.[i]

<i>: Raymond Kromreys Molekularstromrichter, NET-Joumal 6/98, S. 5

15.4 Magnetic force converter

As a result of the systematizing of all eight possible interactions, the largest force effects are to be expected as a result of static and of open magnetic field lines. It is questionable, if this statement is generally valid and is true everywhere in space. But it is valid at least for a terrestrial laboratory and only here, on earth, a solution for the energy problem is strived for. No wonder therefore, if the electric energy technology nearly without exception is using these force effects. We find them at a solenoid, at a relay coil, at a magnetic tuning cylinder and equally between the stator field and rotor field of an electric motor.

The motor however takes a special position, because its rotor is turning. In that way a switching of the winding and commutating of the currents is necessary or the field of the stator winding is being turned, for instance in the case of an alternating current motor. This is necessary, so that in the air split of the motor the fields from the stator and rotor always are standing opposite like for the solenoid and a driving force can be formed. For the operation therefore oscillating currents are necessary, so-called alternating currents, which are fed in into the winding with the right frequency and phase. There can be spoken of an operation in resonance. It surely is no coincidence, that Nikola Tesla, the founder of the rotary field theory and inventor of the alternating current motors at the same time is the discoverer of the neutrino radiation!

The electric motor slides already into the second column from the top in table 15.1, as we see, and should be assigned to the case of the oscillating interaction of open magnetic fields. The frequency for motors usually is very small.

But also at high frequencies there can't be reckoned on some free energy which would show or even be utilizable, as can be heard from the inventors who tinker with magnets, mostly with permanent magnets (Fig. 15.4). The reason very simple is that there exist no physical particles, which could mediate this interaction. Magnetic monopoles would be necessary, thus north pole or south pole particles, so that an interaction with the open H-field lines can occur.
Such particles could form as a result of currents and eddy currents, but for that a good conductivity would be necessary and that isn't present in the vacuum. Therefore magnetic monopoles can't exist at all! This point we already had worked out (fig. 4.4, question I). The same statement then also applies to the oscillating case.

If nevertheless something like free energy should show in the case of some magnetic field converters, then unnoticed by the inventor still other physical effects are added. By means of concrete concepts this circumstance can be studied and discussed. The meanwhile well-known railgun is a corresponding example, for which besides the used magnetic force unintentionally a further principle is used. A more detailed occupation with this device is worthwhile, because here some fundamental concepts of space energy get clear.

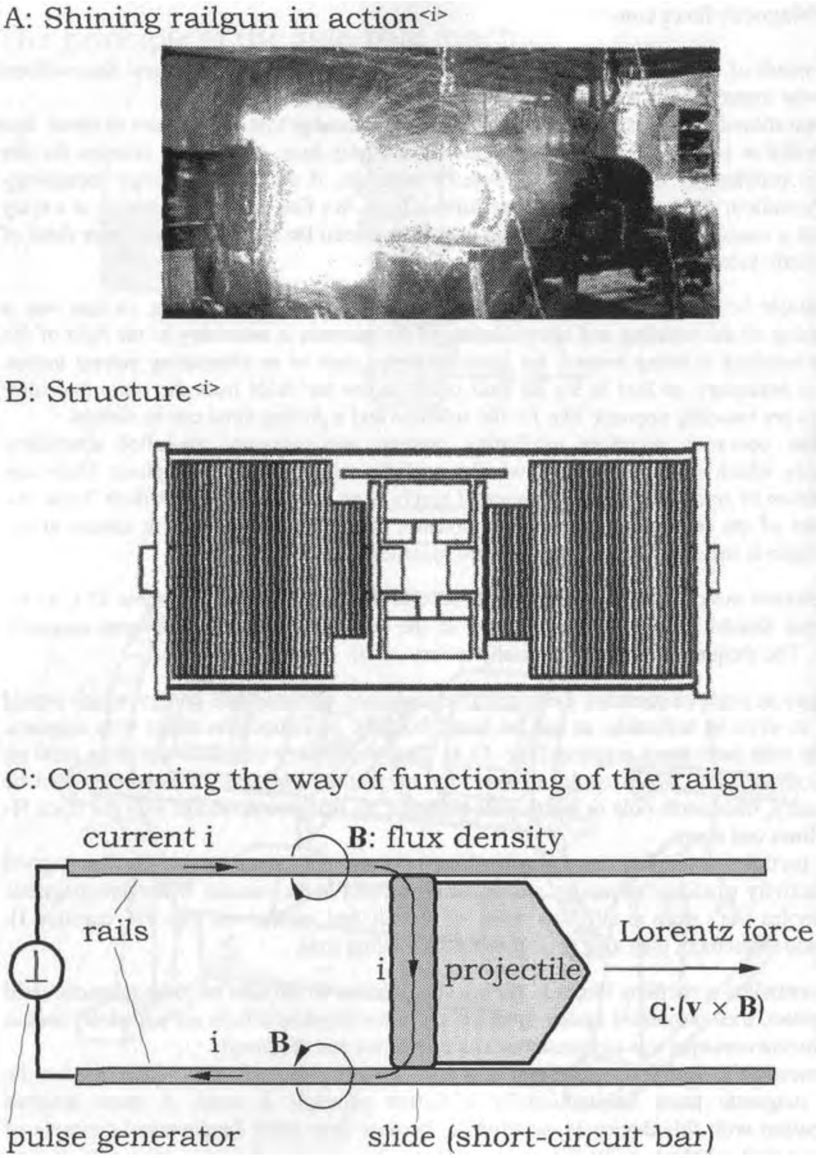
A: Shining railgun in action<i>
B: Structure<i>
C: Concerning the way of functioning of the railgun
current i
B: flux density
rails
Lorentz force
i
projectile
q·(v × B)
i
B
pulse generator
slide (short-circuit bar)

Fig. 15.5: Structure and way of functioning of the railgun

<i>: in the internet under: www.glubco.com/weaponry/railgun.htm

15.5 The railgun

The engineers and physicists involved in the SDI-project were quite astonished, as they had a close look at the bent rails of their gun. During the test operation the equipment was really flying around their heads.
They were very sure to only have fed in 16.7 MJ of energy, from the rotation of a homopolar generator, because more was not available for the experiment by any means. The projectile with mass m_0 = 0.317 kg lying on the rails thereby should have been accelerated to the velocity of 4200 m/s. Instead huge forces were at work here, which the construction couldn't counteract at all. There is talk about the released energy having amounted to 399 GJ, what corresponds to an over-unity effect of 24000<ii>. This factor describes the proportion of the released to the taken up power or energy. If these details should be correct then this would be the most efficient converter for free energy, which has been developed until now.

Behind the project name SDI (Strategic Defence Initiative) is hiding the by the United Slates prepared "Star wars". But how one fares a war, which nobody can pay anymore and no-one wants, entirely according to the motto: There is a war and no-one goes there? This war undoubtedly the strategists and initiators themselves have lost, who even had to watch, how their space gun appears in the internet with design drawings and rich visual material to be called by anyone<i>.
Today, where we are surrounded by nothing but friends, where in Russia and at other potential opponents is fought more against internal problems and one lets the expensive space toy rot for lack of money, the coat of the military secrecy obviously no longer can be held over such an explosive project as the railgun.

Thus informative details have reached the public. In the pictures a bright lightning can be seen at the moment of launching (fig. 15.5 A). Here presumably is being materialized, in which the part of anti-matter annihilates with the particles of matter under emission of light. There thus takes place the same process as in the case of lightning or the shining of the sun.
In addition is being reported that heat energy is withdrawn from the environment, a circumstance, which is typical for all functioning converters for space energy. We thereby are reminded of the possible formation of ice in a lightning channel.

Like for a lightning also the railgun is stimulated with a very high excitation voltage and with extreme speeds of change of the tension voltage (high du/dt) (fig. 15.5 C). From the setup it concerns a bridge of Ampere, which in various respects appears to be superior to the rocket engines, after the costly transport of the propellant into space isn't necessary, since the capacitor batteries can be recharged by solar power.

<i>: e.g.: www.glubco.com/weaponry/railgun.htm

<ii>: Galeczki, G., P. Marquardt: Requiem fur die Relativitat, Verlag Haag + Herchen (1997), S. 139

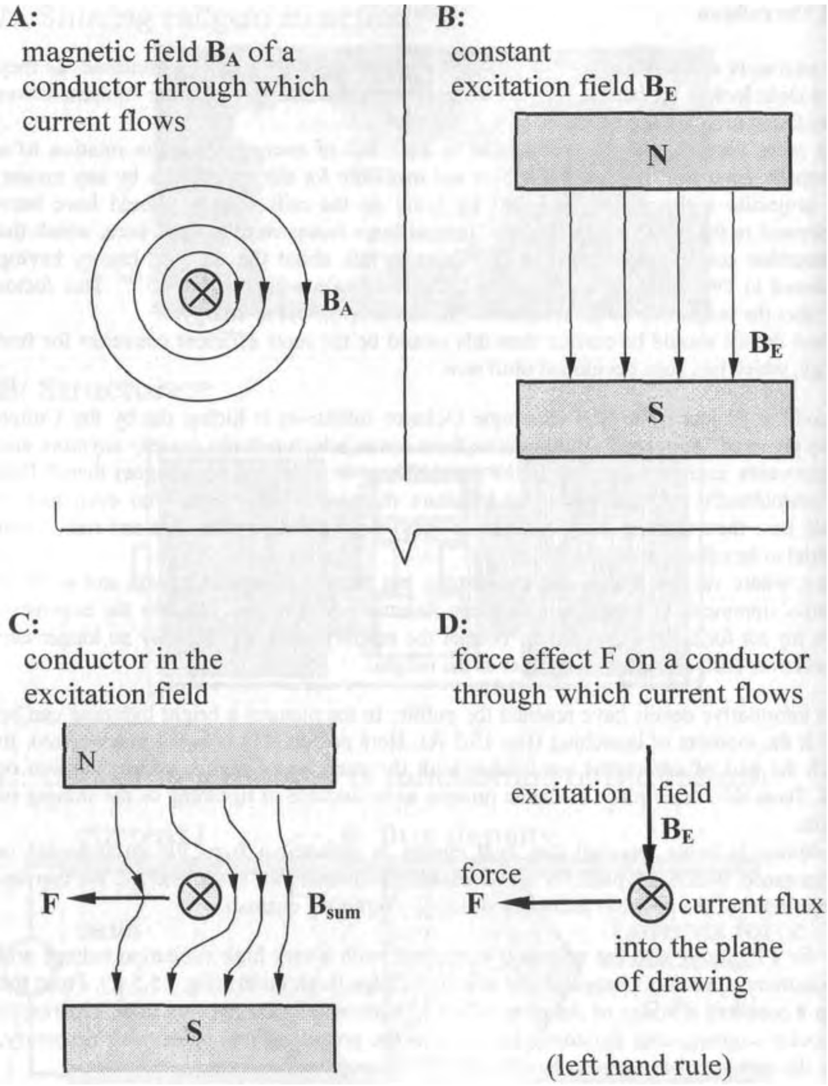
A:
magnetic field B_A of a conductor through which current flows
B_A
B:
constant excitation field B_E
N
B_E
S
C:
conductor in the excitation field
N
F
B_{sum}
S
D:
force effect F on a conductor through which current flows
excitation field
B_E
force
F
current flux
into the plane of drawing
(left hand rule)

Fig. 15.6: ____ The distribution of field lines and force effect on the slider of the railgun through which current flows in a permanent magnetic field

The projectile has the form and the function of a short-circuit bar and is guided slidable between both rails of the railgun. The high-tension capacitors are switched on both rails at the moment of launching, so that in a very short time an extremely high short-circuit current of several thousands of Ampere flows through the bar.
Since the bar in addition is situated in a static magnetic field, there acts an accelerating force on it (fig. 15.6 D). It is the force effect of a conductor through which flows a current in a magnetic field, like it is active in every electric motor. If we overlap the fields of the conductor (fig. 15.6 A) and of the magnetic field (15.6 B), then we observe a bending and lengthening of the field lines (15.6 C). There exists the effort to re-establish the original state, which represents the smallest magnetic resistance, and for that the conductor is shifted out by means of the arising force. In the sketch it is accelerated to the left. That far the explanation concerning the bridge of Ampere. That has nothing to do with free energy. For the enormous degree of effectiveness, as it has been determined, further effects have to be added.

15.6 Unipolar induction

The projectile, or from the function let us rather speak of the short-circuit bar or the slider, at first is entirely conventionally accelerated and experiences, mathematically expressed, a dv/dt. The magnetic field B stretching perpendicular to the movement is constant, so that according to the Faraday relation<i> E = v x B from the velocity v an electric field strength E results and from the acceleration dv/dt a field change dE/dt. These open field lines along the length of the slider, in particular the oscillating part, appears to interact with oscillating particles and to collect these particles<ii>. It concerns presumably neutrinos, which primarily materialize in charge carriers. These contribute to the current flux in the slider and to the acceleration, whereupon still more neutrinos are collected.
A hardly controllable avalanche effect is formed. Only if the change in tension voltage has worn off and the capacitor is completely discharged, also the resonant interaction will again collapse.

The inventors, who want to construct a civil version of the railgun, is given a warning on the way which should be taken seriously. At first it doesn't take particularly much imagination to imagine a rotating arrangement of the gun, a construction with one axis, whit which a generator driven, which produces power. A small part is supplied the system again as supply for itself. The rest would be available free to the consumers as non-polluting, regenerative energy.
That really sounds good, if there wouldn't be this one obstacle.

<i>: The equation of transformation concerning the unipolar induction already was treated more detailed in chapters 6.4 and 9.3

<ii>: It is the resonant interaction according to table 15.1, line four

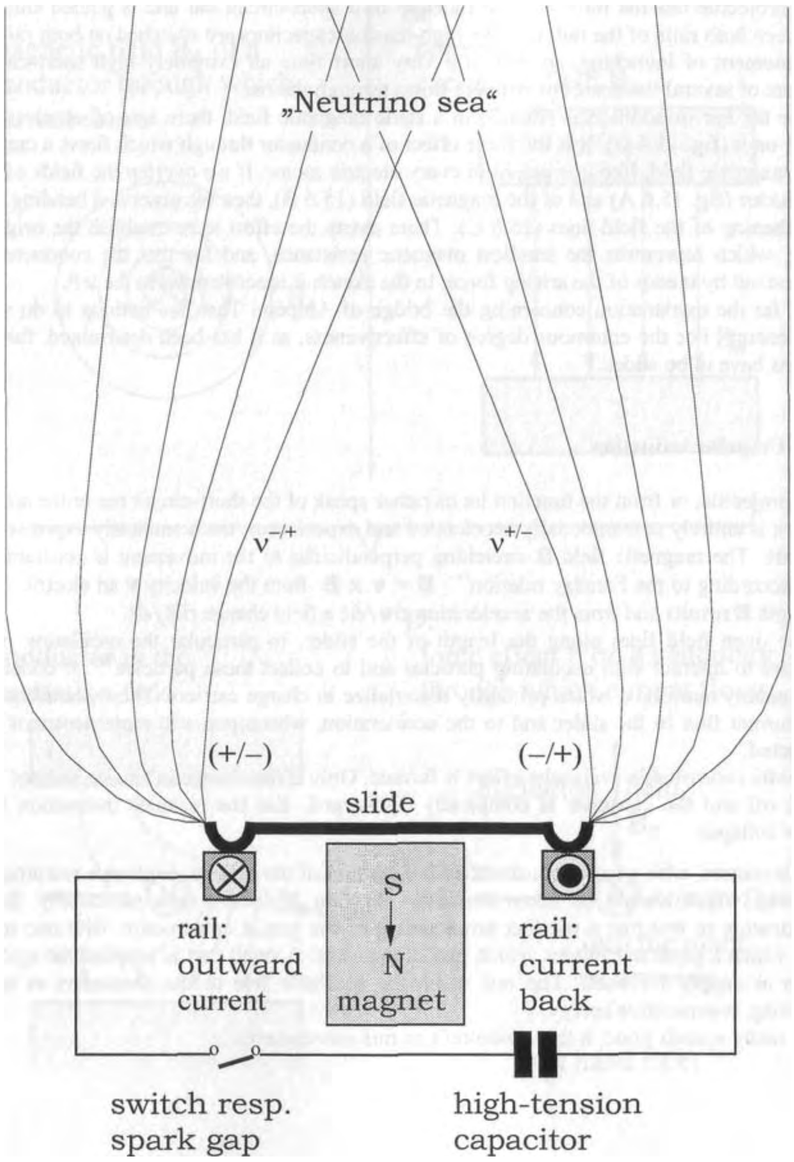
„Neutrino sea“
ν −/+
ν +/−
(+/−)
(−/+)
slide
S
N
rail:
outward
current
magnet
rail:
current
back
switch resp.
spark gap
high-tension
capacitor

Fig. 15.7: The collecting of neutrinos by oscillating open field lines

15.7 Tendency to instability

In a continuously working machine the discharging of the capacitor cannot remain a single event as in the case of the gun. The discharging and the recharging has to take place cyclic depending on the revolutions per minute. To obtain a rotating movement even to a certain extent ignition has to follow ignition. But if the new ignition takes place, although the avalanche effect of the last one still hasn't worn off, then inevitable a catastrophe will occur, then the work of wonder is taken apart under the eyes of its creator. Numerous inventors already have had to collect such painful experiences. It is assumed that not even Nikola Tesla had escaped, as he had to put away again his stately luxury car with electric motor and energy converter in a barn near Buffalo already after one week of test operation in the year 1931<i>.

Of course also for this problem solutions in accordance with engineering are offered. Meaningful would be a restriction of the revolutions per minute and a power regulation. Only most inventors don't think that far. On the one hand, because they handicraft without an useful physical model and on the other hand they think they already have reached the goal, if they observe something like free energy for the first time. Just as fast as the joy then the disillusionment comes, because a converter which doesn't work, is not able to convince anyone.

Tesla already was aware of this set of difficulties. He fastened his converter to the dashboard and not in the engine compartment, presumably to adjust the coupling of the coils from the drivers seat during the drive by means of two metallic rods, which he pushed into the case. But sometime even this regulation by hand has to go wrong, because the collected neutrinos on their part collect further neutrinos (fig. 15.7), so that in the case of an unfavourable order of ignition an additional amplification is possible. For a reliable operation according to that directly or indirectly the phase of the ignitions to each other should be checked.

At the example of the railgun space energy technologists and inventors can study the relations and the way of functioning very concretely and even calculate these relatively simple. It is a big relief, that all three vectors stand perpendicular to each other: the E-field, the B-field and the velocity v. Ideal conditions both with regard to a maximizing of the wanted accelerating force and for the resonant interaction, increase at the same time the collecting of space quanta, which probably may be set equal to the neutrinos. This is made possible by the Faraday's law of unipolar induction. In that way at the right and left end of the slider a positive a negative pole each are formed. The further the two poles are away of each other, the more the field lines are opened and the more neutrinos can go into resonance. In this place still considerable improvements and optimisations are possible.

In addition to the two discussed the phenomenon of the electrostriction is added as a third phenomenon, which authoritatively contributes to the conversion of neutrinos into electrons. It is a field dependent change of length, which in the case of lightning takes care of the thunder and in both cases, therefore also here, is active as a charge carrier producer.

<i>: see also chapter 9.5 Free energy,
A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, Kap.ll, S. 20 and H. Nieper: Revolution, MIT-Verlag 1981, S. 194

A: electric dipole (e.g. electrostatics, see also fig. 6.8)

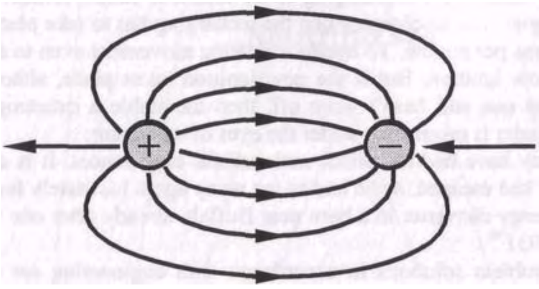

B: electric monopole (e.g. electron e~, see also fig. 4.3)

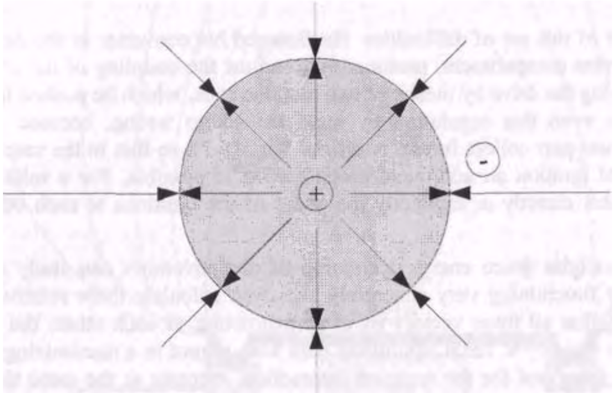

C: Mixed form in the case of cylindrical symmetry

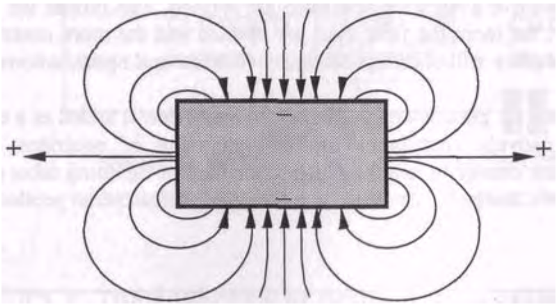

Fig. 15.8: Dipole fields with unipolar parts

15.8 Unipolar field configuration

If we again go back to fig. 5.1 and continue our considerations with examples concerning line 3. The obtainable force effect of open electric fields indeed is for powers of ten smaller than that of magnetic fields, but then particles are mediated, an invaluable advantage and an indispensable prerequisite with regard to the generation of space energy. As long as the particles are considered in the balance sheet of energy then in addition by no means can be claimed, space energy converter are perpetuum mobiles.

In the question, why only electric particles can be mediated, has to be pointed to the repeatedly mentioned circumstance that only electric particles can be formed as a result of the concentration effect of potential vortices. Magnetic particles however plain and simple cannot exist, since for missing conductivity in the vacuum no eddy currents are possible (see fig. 4.4).

In the case of the arising Coulomb forces again is distinguished between the static and the oscillating case. We want to start with electrostatics.

The range of technical applications of static electricity is large. It stretches from varnishing technologies and filter technologies till the fly grill in the arbour. In all cases the field is built up by charge separation. By means of a high-tension generator a positive pole and a negative pole are produced, between which the field is stretching. The field lines now start at one pole and end at the other, unlike pole (fig. 15.8 A).

In this manner almost no open field lines are available, which point to the outside and could interrelate with free particles. If one for instance intends to attract and collect charged particles from the solar wind, from the cosmic radiation or from the electricity of the air, then the design in one point has to be changed fundamentally.

A unipolar field configuration is necessary. What is meant with that, answers a glance at the spherical vortex model of the electron (fig. 15.8 B resp. fig. 6.2). Here it as well concerns a formation of dipoles as a result of the charge separation, but one pole is hiding in the inside of the other pole . In that way its field lines are captured and don't have a chance anymore to come out, to reach the other pole.

But if the pole lying on the outside for its part can't close its field lines any longer, then these point helpless into space and search in their neighbourhood, in the distant world and if need be even in the infinity of the universe an unlike anti-pole, which as a result interacts and is attracted.

The measuring technician analyses these open field lines and falsely calls the construction then a monopole, only because he isn't able to reach the locked up pole. To blame is the unipolar field configuration, which with that probably would be explained to a certain extent.

The designer and inventor as well might have realized how he has to construct his device, with which he generates open field lines to collect space quanta. He has to lock in one of the two poles as good as possible. Optimal would be of course a spherical symmetric construction like in the case of the elementary vortex. Compared with that a cylindrical symmetry indeed is suited far less good, but it offers constructive advantages (fig. 15.8 C).

We now will report of such a functioning device.

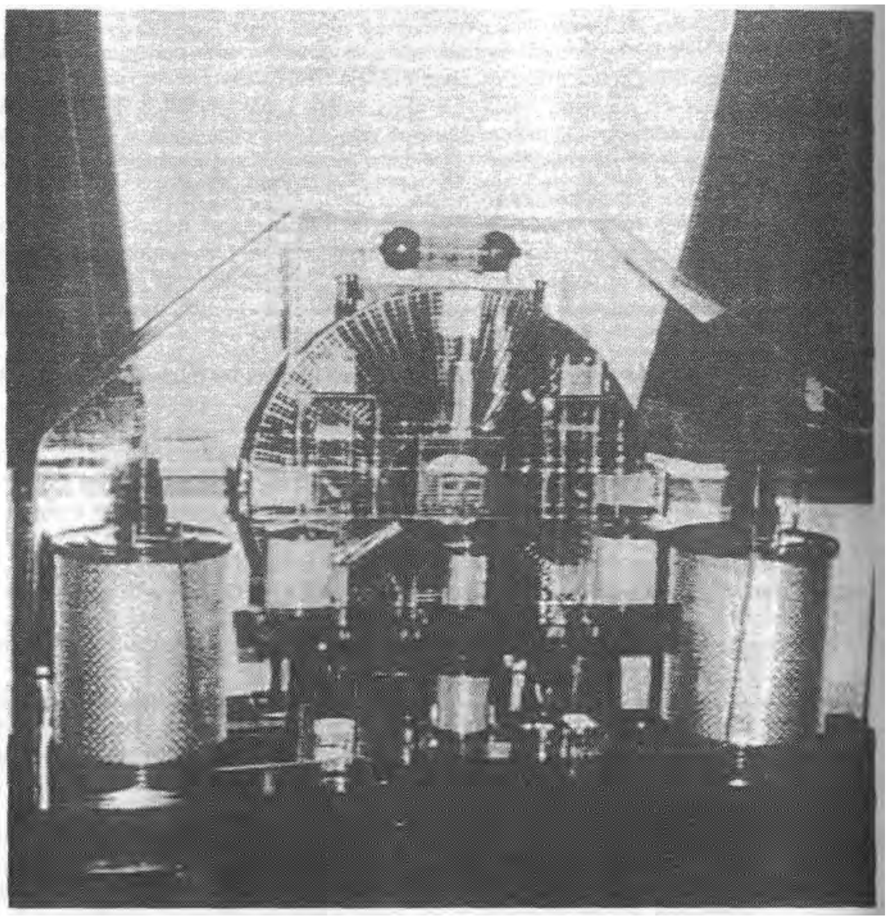

Fig. 15.9: Demonstration converter "Testatika".<i>

<i>: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, S. 29,
I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, p. 13 and in the NET-Journal, Heft 8/9, 1997, S. 16 as well as iss. 12, 1997, p. 6.
D. Kelly. Der Schweizer ML-Konverter, Raum & Zeit Special 7, S. 164

15.9 The Testatika

An electrostatic device, which produces open electric field lines, is situated in Linden in Switzerland. It optically is very imposing and belongs to a religious community, which has called it Testatika and is of the opinion that it is a free energy converter. Inexplicably the Testatika does not serve the community, which generates their electric power conventionally and to a large extent by itself, as a source of energy. Instead the 2 kW device only very seldom is demonstrated for special occasions or to select groups of visitors. The religious community after that explains the astonished observers humanity not yet is mature for the technology.

Perhaps just the opposite is correct and the technology not yet is mature. According to my personal assessment such an electrostatic device in principle is entirely unsuitable for the continuous operation.

It can be expected that the open field lines sooner or later will interrelate with the electricity of the air and thunderstorms are being attracted by the infernal machine so long until lightning strikes and the demonstration with that has finished. That's why the Testatika may be switched on only for a short time, only at sure weather situation and not too humid air, and many a registered visitor has been sent away without having seen the "thunderstorm machine".

As an object of demonstration and study the Testatika however is well suited. Alone the circumstance that no cable leads to the device and it nevertheless releases energy in the order of magnitude of 1 till 2 kW, surprises all visitors. At least the impression is mediated as if the machine would violate the law of conservation of energy, which is not correct. The Testatika is similar to an induction machine, which works with friction electricity. Thereby the unlikely charged bodies do not have to unconditionally touch and rub at each other, it already is sufficient, if they are brought in the immediate vicinity of each other. In the case of the Testatika the electrostatics of two against one another rotating discs is taken off by brushes.

The excitation energy presumably is taken out of the natural E-field, which just like that can amount to 200 Volts per meter (see chapter 2.9). The large diameter of 80 cm of the discs and their bad conductivity (acrylic glass) permit this conclusion. The charge taken off by the brushes afterwards is temporarily stored in two capacitors of 2 Farad at 300 Volt, so-called Leyden jars. This far one actually is reminded of a Wimhurst generator, in which the energy is supplied the system by turning the disc. Large powers cannot be drawn by that. Plans to build such an induction machine by yourself have been published<i>.

In the case of the Testatika however two discs are used and by hand stimulate to rotate oppositely. This rotational energy in this case isn't used to produce power, otherwise the discs quickly would stand still again, but that doesn't happen.

Until now apparently no-one has discovered the secret, which is kept strict by the members of the community. In my opinion the energy situation on our earth however is too serious, as that we would be able to afford playing hide-and-seek and egoistical secretiveness.

<i>: I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, S.14

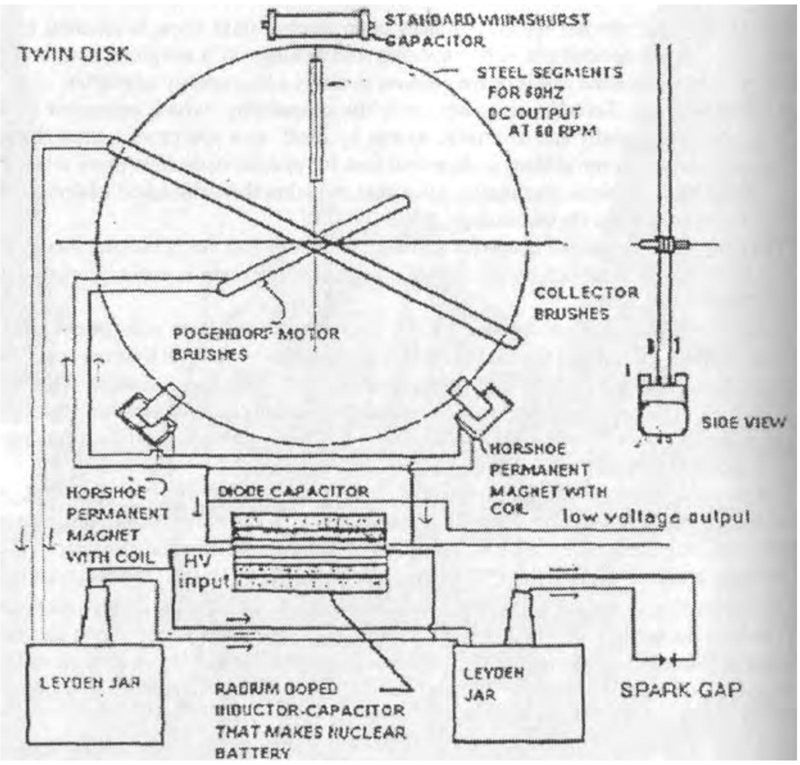
STANDARD WHIMSHURST CAPACITOR
TWIN DISK
STEEL SEGMENTS FOR 60HZ DC OUTPUT AT 60 RPM
COLLECTOR BRUSHES
POGENDORF MOTOR BRUSHES
SIDE VIEW
HORSHOE PERMANENT MAGNET WITH COIL
HORSHOE PERMANENT MAGNET WITH COIL
DIODE CAPACITOR
low voltage output
HV input
LEYDEN JAR
RADIUM DOPED INDUCTOR-CAPACITOR THAT MAKES NUCLEAR BATTERY
LEYDEN JAR
SPARK GAP

Fig. 15.10: Sectional drawing of the Testatika according to Don Kelly, (Clearwater, Florida, USA)

<i>: Der Testatika generator, NET-Journal, Dezember 1997, page 6 similar picture is found in Raum & Zeit Spezial 7, page 164

15.10 The secret of the Testatika

The crucial point is the opposite direction of rotation of both discs. If we assume the static earth electric field is the cause and serves as an excitation field, then as an effect a field arises, which stands perpendicular to that. The axial component now points out of the centre of the disc.
In the case of only one disc the field lines in front and behind the disc again are closed, so that no open lines can form. With one disc or with two discs rotating in the same direction hence no unknown charges can be attracted.
In the case of two oppositely rotating discs however shows one component along the axis of rotation to the observer, that of the other disc exactly in the opposite direction. In that way between both discs a pole is „pinned", which no longer is able to close all field lines on the outside around the machine. Thus open field lines and a, however incomplete, unipolar arrangement are formed.
The charge carriers sucked from the electricity of the air as a result support the natural electrostatics and speedy recharge the capacitors, even if up to 10 Amperes are taken out by the consumers.

The ingenious thing of the machine is its extremely simple construction and the simple concept.
If one includes the collected particles also in the balance sheet of energy, then it thus will turn out that the law of conservation of energy is not violated at all. There thus can't be talked of free energy. In this context the Testatika may rather be given as a counterexample. Air ions are the carriers of the electricity of the air and not carriers of free energy.

Negative air ions are indispensable for our welfare. One should only remember the first men in space, who after the landing were pulled out of their capsule more dead than alive, after they had to stay in the unhealthy atmosphere of the capsule for a longer period of time. Only the installation of ionising devices for negative air ions made possible longer staying in space.
The taking out of the air of negative ions hence is not unproblematic and not particularly ecologically compatible. An atmosphere harmful to life is formed which Dr. Wilhelm Reich has called DOR-state. He by the way has designed a Cloud-Buster, with which he could take static electricity, forming above the desert sand, out of the air. In that way the negatively charged rain clouds no longer are repelled and driven away. Reich has tested his weather machine 1954 in the desert of Arizona. After he had freed the atmosphere of the „DOR-strain", as he expressed himself, in the desert area the atmospheric humidity steadily increased from 15% up to 95%, there grew prairie grass and everything started to turn green, and eventually after many years for the first time rain fell again.

Static electricity, as far as the right polarity is chosen, may conditionally be used for the mechanics of rain making. For free energy concepts it however isn't suitable. Already Nikola Tesla has pointed to the circumstance that our hopes will be in vain if the free energy would be of static nature (see fig. 9.5). He in his speech, which he gave 1891 before the AIEE, has left no doubt that free energy exists, which is kinetic and with that energy technically usable for us,. Chapter 16 will be occupied solely with this case<i>.

<i>: according to point 4 in table 15.1

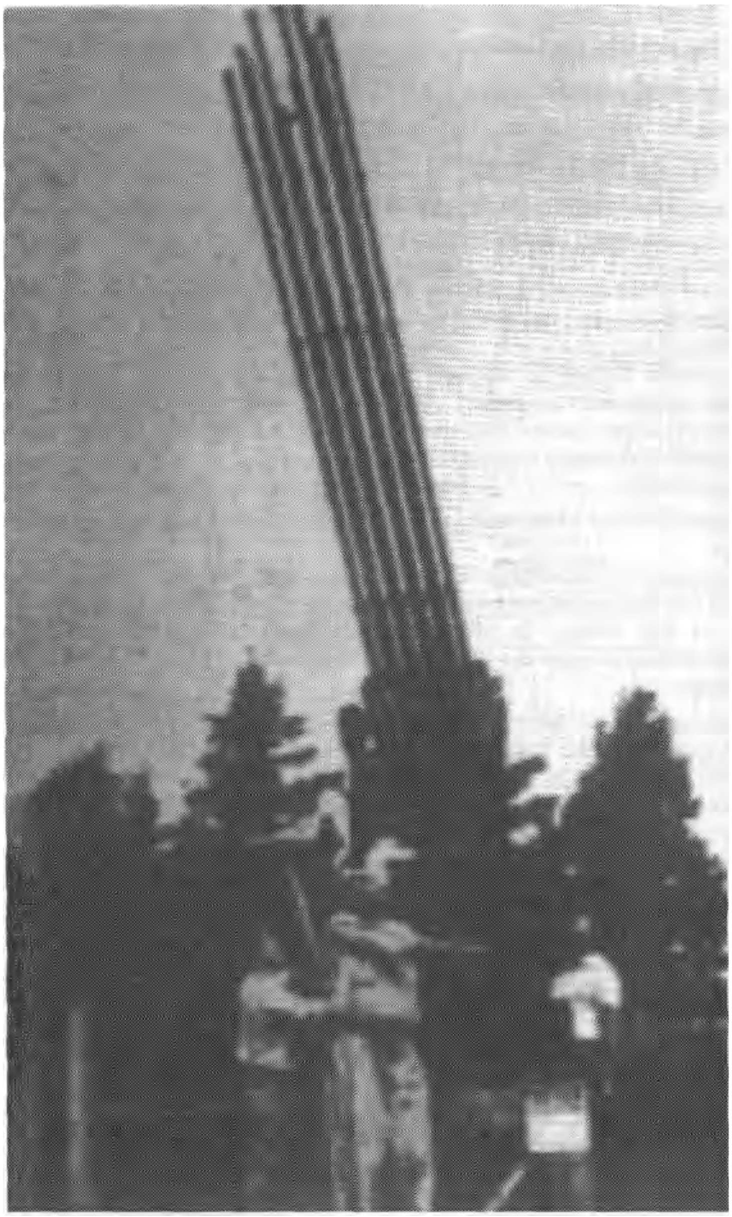

Fig. 15.11: The Cloud-Buster of Dr. Wilhelm Reich 1954.[i]

<i>: H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3, page 122

<ii>: Part 1, fig. 9.3; concerning the unipolar induction see also fig. 6.5 and 11.8

Fig. 15.11 shows the weather machine of Reich, the Cloud-Buster. It can be understood only hard, why Reich directs a 3-4 m long metallic pipe with a diameter of 4 cm to the sky and connects the rear end with a deep well or with flowing waters. The effect should have been increased considerably with a few milligrams of radium. An indication that he must have worked with static electricity, delivers however a tragic accident, in which one of his collaborators was lamed on one side. He carelessly had touched the charged apparatus and suffered an electric shock[i].

15.11 The key to free energy

As a contribution to the discussion the individual principles of functioning of space energy again are collected and the attempt is undertook to value them.

In the case of an oscillating dipole configuration, for instance the railgun, open field lines are present only along the mutual line of connection (fig. 15.8 A). With that not particularly many space quanta can be reached. It hence has to be operated with gigantic excitation powers in the range of many thousands of Ampere, so that further field lines fling open and interact. The wanted over-unity effect therefore can only be reached at an enormous expense of technical apparatus.

An unipolar arrangement here is considerably more advantageous, where holds: the more unipolar, the fewer excitation power is required. But in that way it can take longer until the collecting of neutrinos like an avalanche again has worn off. In the case of an ideal spherical arrangement (fig. 15.8 B), as the ball-lightning takes, the process can even last for minutes. This explains why unipolar systems can be kept under control only very hard. If the neutrino avalanche is rolling then it purely theoretical only can be stopped with a still larger excitation power, for instance by phase shifting, what can hardly be realized in practice. The rolling avalanche can't be stopped anymore by normal means. A synchronous operation between the neutrino oscillation and the converter can, apart from the technologically hardly realizable high frequency, by no means really be recommended. As a rule one single steep flank of the change of the excitation voltage is sufficient to start the avalanche. By means of the repetition frequency or by means of the duty cycle of the excitation voltage then resonances to the neutrino field can be made or avoided. On the other hand can't be done without the avalanche effect. The utilizable power of the neutrino converter otherwise would be much too small. This case should be pursued further in the design of a longitudinal wave gauge.

All converter systems at first work based on a well-known and tried and tested physical principle of functioning. In the case of the railgun it is the bridge of Ampere. The thus used force effect on a conductor through which flows current is advantageous due to the obtainable order of magnitude and as a basic concept extremely recommendable. But also Coulomb forces or other physical principles can be used.

Despite that a further relation still must be added, which produces the interaction with the neutrinos. Closely associated with the unipolar arrangement it is the unipolar induction, which in virtually all space energy concepts is put to use. It already could be shown that the railgun uses the effect as well as John Searl in the case of his flying disc[ii]. The Faraday law of induction turns out to be the key to free energy.

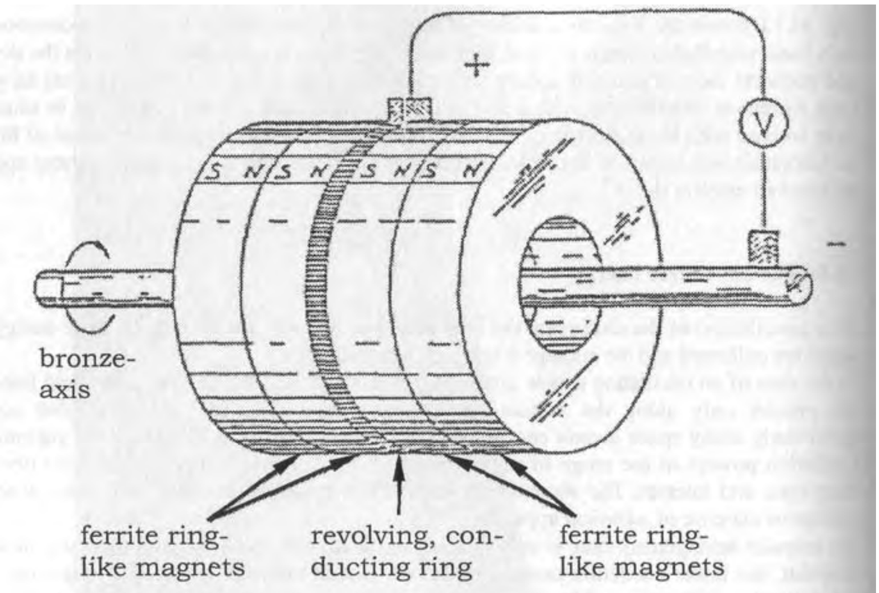
+
V
S N/S N/S N/S N
–
bronze-axis
ferrite ring-like magnets
revolving, conducting ring
ferrite ring-like magnets

Fig. 16.1: The Faraday generator, which with rotating magnets is called N-machine.[i]

According to chap. 6.5, eq.60 the equations of transformation are:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \quad \text{and} \quad \boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \qquad (16.1)$$

where: $\mathbf{v} = dx/dt = v \cdot \mathbf{e}_x =$ component of the direction of motion perpendicular to the area being stretched by the field pointers $\mathbf{H} = H \cdot \mathbf{e}_y$ and $\mathbf{E} = E \cdot \mathbf{e}_z$.

Thus in addition to the basic fields E and H (at $v = 0$) additional fields E_z and H_z, depending on motion, occur and the overall fields E_o and H_o are measured (which corresponds to eq. 63 in chapter 6.5):

$$\boxed{E_o = E + E_z} \quad \text{and} \quad \boxed{H_o = H + H_z} \qquad (16.2)$$

(which corresponds to eq. 62 in chapter 6.5):

with: $$\boxed{E_z = v \cdot B} \quad \text{and} \quad \boxed{H_z = -v \cdot D} \qquad (16.3)$$

Table 16.2 A: The equations of transformation of the electromagnetic field.[ii]

16. Space energy technology (SET)

It quite concretely concerns the question for a technology concerning the use of the resonant interaction (according to line 4 in fig. 15.1). For that open and at the same time oscillating electric field lines are needed, which mediate neutrinos and pass them on to a receiver working in resonance. The sun, some planets and other celestial bodies, as we already have worked out, use the effect. Even an entire galaxy is kept together in this way. This interaction plays the crucial role for the theme of space energy and the question is asked, with which technology it can be produced artificially.

16.1 The unipolar generator

The most direct way obviously leads over the Faraday relation concerning the unipolar induction. With the classic Faraday generator, where a permanent magnet is turned along its axis, at first a static electric field can be produced. By rotating in opposite direction or magnets rotating in the same direction but oppositely poled, relatively simple one pole can be „pinned" between the magnets and an unipolar construction can be built. Numerous research scientists already have worked in the area of the Faraday machine more or less successfully (fig. 16.1)<i>. There is reported of instabilities and of the picking up of unknown energy at high revolutions per minute. In the majority of the cases it in a sense of the Testatika will concern collected electricity of the air.

The Faraday relation in addition also appears to be hardly understood correct physically by anyone. A scientific magazine in this context takes the opinion: "Faraday proves Einstein wrong", and the production of electricity with Faraday's unipolar inductor violates the laws of physics !<ii>

<i>: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989, S. 44

<ii>: „Faraday in his experiment did let rotate a copper disc above a resting cylinder magnet; as expected in a loop of wire a tension voltage was created (F-machine). Than he let the magnets rotate, and the disc stood still; now again a tension voltage should have resulted - but there was no voltage. In the third experiment the magnet rotated with the disc in the same direction and with the same speed. Because there was no relative motion between both, an induced tension voltage wouldn't have been expected - but it was measurable! (N-machine, see fig. 16.1).
What does that mean? If the relative movement between magnet and disc is not always crucial for the formation of an induced tension voltage, then also the absolute movement has to play a role - because something has to move, for a current being formed. But an absolute movement according to the theory of relativity can't be detected - thus Faraday's experiment proves Einstein wrong! Therefore you won't find anything about this experiment in the textbooks".

Taken out of the article: Faraday widerlegt Einstein, PM-Magazin 11/ 1998, P. 133

from table 16.2 A:

$$\boxed{H - H_o \overset{(16.2)}{=} - H_z \overset{(16.3)}{=} v \cdot D} \tag{16.4}$$

with the velocity $v = dx/dt$,
not accelerated $dv/dx = 0$ (inertial system) and derived for x:

$$\boxed{\frac{dH}{dx} - \frac{dH_o}{dx} = \frac{dx}{dt} \cdot \frac{dD}{dx} = \frac{dD}{dt}} \tag{16.5}$$

The curl of the H-field pointer points in the z-direction:

$$\boxed{\text{rot}\,\mathbf{H} = \frac{dH}{dx} \overset{(16.5)}{=} \frac{dH_o}{dx} + \frac{dD}{dt} \overset{!}{=} \mathbf{j} + \frac{d\mathbf{D}}{dt}} \tag{16.6}$$

comparison of coefficients with Ampère's law:

$$dH_o/dx = \mathbf{j} \quad , \tag{16.7}$$

integrated over dx and ds and formulated generally valid:

$$\boxed{\oint \mathbf{H_o}\, \mathbf{ds} = I_{ein}} \tag{16.8}$$

thus follows from that: magnetic field = vortex field!

Faraday's law of induction (analogous derivation):

$$\boxed{\text{rot}\,\mathbf{E} = \frac{dE}{dx} = \frac{dE_o}{dx} - \frac{dB}{dt} \overset{!}{=} - \frac{d\mathbf{B}}{dt}} \tag{16.9}$$

cause for measurable electric field E_o is missing:

$$dE_o/dx = O \quad , \tag{16.10}$$

integrated over dx and ds and formulated generally valid:

$$\boxed{\oint \mathbf{E_o}\, \mathbf{ds} = 0} \tag{16.11}$$

thus follows from that: the electric field = irrotational (according to Maxwell)!

Table 16.2 B: The derivation of Maxwell's field equations (Ampère's law and Faraday's law of induction) from the equations of transformation of the electromagnetic field

The author of the article proceeds from Maxwell's formulation of Faraday's law of induction, according to which arises a tension voltage in a wire if the wire cuts magnetic force lines, thus is moved relative to a magnet.
If he had read my books, then he would know that Faraday not only has found the older, but also the more comprehensive law, whereas Maxwell only describes a special case. Starting-point for the by me developed theory of objectivity are on the one hand Faraday's law of induction and on the other hand the regularity dual to that, which both together are called equations of transformation of electromagnetism (fig. 6.5, eq. 60 and fig. 16.2 A, eq. 16.1).

16.2 Derivation of Maxwell's field equations

For the derivation we assume, as already in fig. 6.5 for sake of simplicity, that the movement $v = v_x$ takes place perpendicular to the area stretched by the field pointers $H = H_y$ and $E = E_z$. If we derive the equation of transformation, written down for a field component depending on motion, for the coordinate x by using the equations of material and compare the result with Ampere's law resp. Faraday's law of induction, then it becomes clear that the comparison is successful only under certain prerequisites. This circumstance proves that the Maxwell equations only describe a special case and that the equations of transformation are more general valid and causal (table 16.2 A and B).

But in this place is crucial that Faraday's law of induction according to Maxwell only is able to describe a formation of dipoles. For the formation of unipolar field structures however must be fallen back upon Faraday's law concerning the unipolar induction. Whoever wants to understand or even develop by himself concepts concerning space energy, first must have understood Faraday's law in its whole range.

Usually the Faraday generator is, like in the original building shape, equipped with permanent magnets and operated in a steady operating state. Doing so consequently only static electricity is formed. We however need oscillating fields, and for that either the rotation, or the magnet should be changed in polarity with high frequency. Expressed with the precision of mathematics, the first case is described by: $E(t) = v(t) \times B$ and the second case by $E(t) = v \times B(t)$. Both cases have to be investigated and discussed, because both $v(t)$ and $B(t)$ are possible in principle. It is added that in both cases mechanically moving, as a rule rotating, designs but just as well resting designs are conceivable, in which only the moving charge carriers themselves realize the component of velocity v.

This time the crucial point is, that in both cases equally a change of the electric field strength E(t) is produced, with the help of which neutrinos should be collected. Because of the extremely high oscillation frequency of the neutrinos large field changes dE(t)/dt and associated with that large changes of tension voltage dU/dt seem to be the optimal solution, which can be handled with today's technology. Apart from that further difficulties are added, which require a managing in accordance with engineering. If we namely work with a large acceleration dv(t)/dt, then the inertia of the accelerated masses should be overcome, then only very small and light projectiles can be launched like in the case of the railgun. If we however work with fast changes of the magnetic field dB(t)/dt, then the inductance acts slowing down.

A: Suggestion from the patent specification of John W. Ecklin

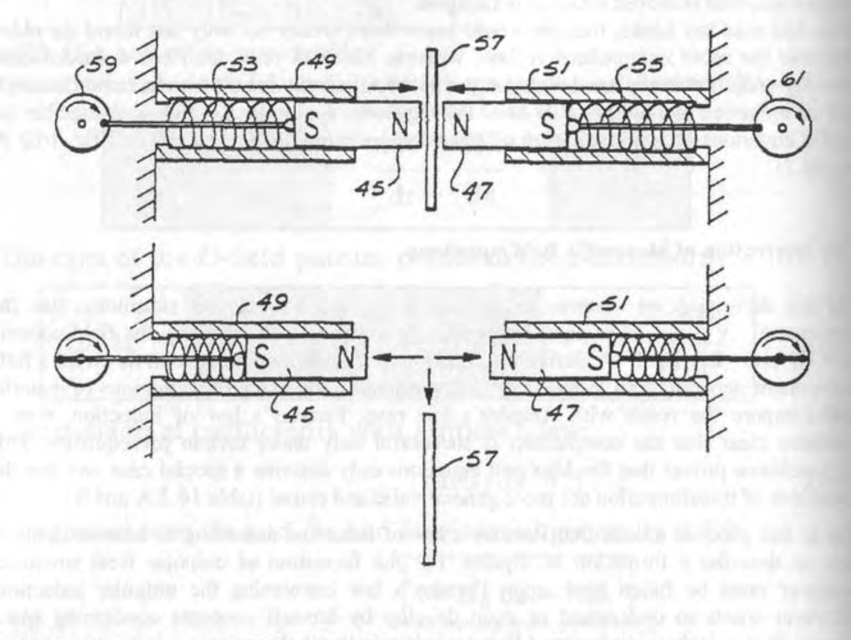
59
53
49
57
51
55
61
S
N
N
S
45
47
49
51
S
N
N
S
45
47
57

B: For instance the „Stationary Armature Generator“

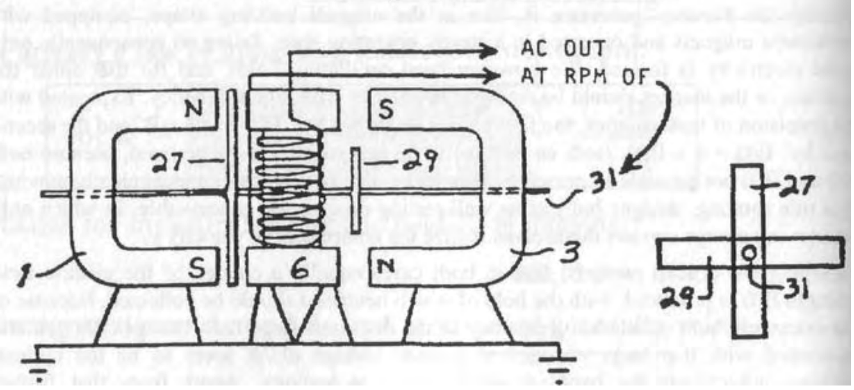
AC OUT
AT RPM OF
N
S
27
29
31
27
1
S
6
N
3
29
31

Fig. 16.3: Principle of the Ecklin generator.<i>

<i>: A. Schneider: Generatoren mit Ferritkernumpolung, NET-Journal 6/98, S. 9
Ecklin: Permanent Magnet Motion Conversion Device, U.S.Pat. 3'879,622

16.3 SET-devices with rotating magnetic fields

If we again come to speak of the Faraday generator, which can be built up in two variants. In the case of the F-machine the magnets rest and only the disc rotates, whereas in the case of the N-machine the magnets rotate along. In this case can't be avoided that some of the induced charge carriers roam about in the magnets and sensitively disturb the structure of the material. One only thinks of the bending spoon effect!

The consequences stretch from a loss of the permanent magnetism, over a shattering and bursting up to a pulverizing of the magnets. Adventurous rumours and alarming reports in this direction are sufficiently available.

An improvement could be obtained by an isolation layer between the conductive disc and the magnets, but against induced charge carriers inside the rotating magnets this measure is not able to achieve anything.

Anyhow the best thing to do will be to completely do without permanent magnets in neutrino converters! If we replace them at least in our minds by electromagnets, then a feeding with alternating voltage is possible. That also is necessary if oscillating neutrinos and not air ions should be collected.

As a result of the alternating voltage at first large eddy currents occur in the disc. The losses can be reduced by radial slits in the disc. The induced currents then only can, as wanted. flow radially to the outside.

Bigger headaches causes us on the other hand the inductive and with that current storing effect of the excitation coils. If a too large excitation inductance should prevent a fast increase in current, then also the induced electric field will increase correspondingly comfortable and hardly be able to persuade a single neutrino to stay.

A coil core of iron or dynamo sheet metal with that is ruled out from the start. Even ferrite would be suitable at most conditionally. Usable are air coils with as possible as few turns.

In an advantageous design the slit copper disc rotates between two air coils, if need be flat coils built in the way Tesla did, which are fed with pulsed tension voltage.

The highest speed of change in current surely is obtainable by means of a spark gap, like already Tesla has used (fig. 9.1). But also other techniques are thinkable as pulse driving.

For instance semiconductor power amplifiers with MOS transistors not only can be switched fast and hard, but in addition frequency and duty cycle can be adjusted freely with reproducible exactness. These are niceties, which will gain importance in connection with the control and regulation of a converter.

Und the heading "generators with magnetic flux variation" Mr Adolf Schneider has collected and commented on some concepts[i]. The generator of the american research scientist John W. Ecklin registered for patent at 22-4-1975 thus stands as an example of a whole group of inventions, in which the magnetic conductivity in the magnetic circuit and with that the flux is changed with a jump.

Fig. 16.3 B shows a building form consisting of two horseshoe magnets (1, 3) with a likewise resting coil in between (6). In rest the magnetic circuit experiences no change and consequently no tension voltage is induced in the coil. The trick is that an axis (31) is pinned right through the arrangement which is turned with two soft iron anchors, which magnetically short-circuit alternating the left (27) and the right (29) horseshoe magnets.

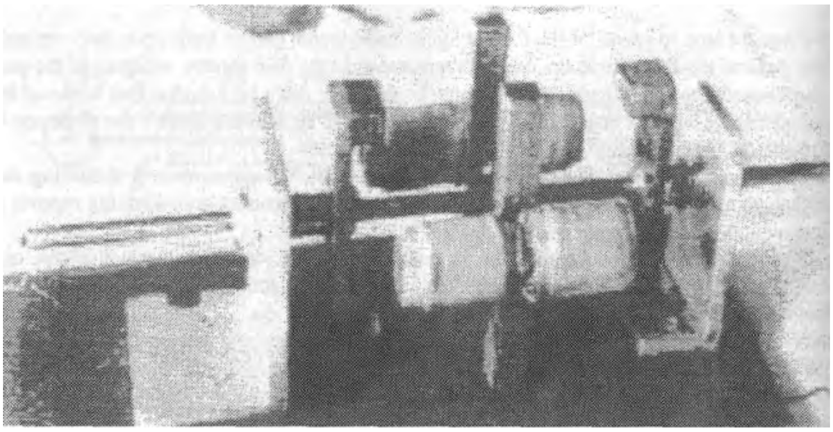

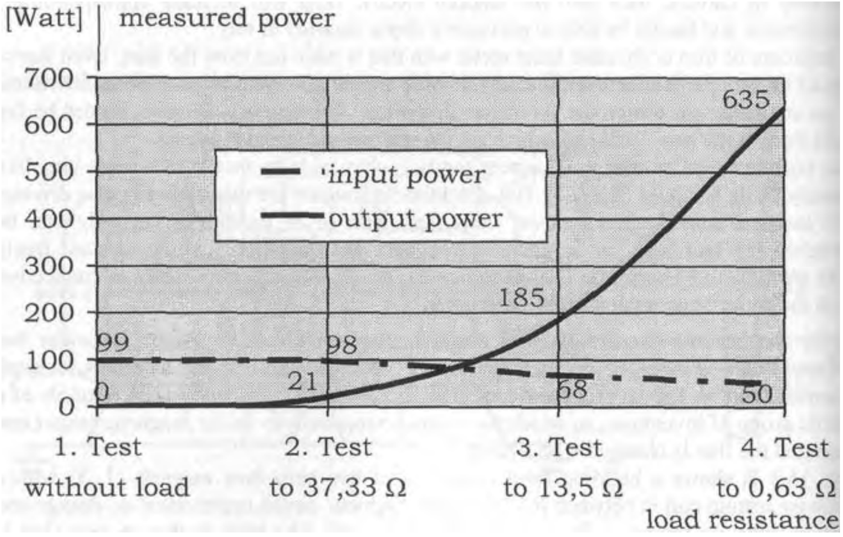

Fig. 16.4: Bedini magnetic converter.<i>

<i>: A. Schneider: Generatoren mit Ferritkernumpolung, NET-Journal 6/98, S.10

Besides the moment of reluctance, which should be compensated by the physical principle, and the friction of the bearing actually no further moment of reaction can occur which would have to be gotten over, not even then if current is taken out of the coil. If the flux change is large enough, purely arithmetically it should be possible to take off an electric power which is considerably larger than the friction power. John Bedini has 1985 measured an over-unity effect of up to 12.6 at a similar constructed generator, where with increasing strain also the factor could be increased (fig. 16.4).

16.4 Commentary concerning magnetic SET-devices

I judge these measurements rather sceptical, after I already in three cases had to break the message gently to the inventors, that their device unfortunately was nothing but an energy destroying machine and they merely had measured wrong, where admittedly the measuring of pulsed tension voltages and currents is not quite simple. So that you don't become a victim of wrong hopes and self-deception, I recommend all SET inventors to realize the closed-loop. If in the continuous operation power can be taken out of such an arrangement without supply of energy from the outside, and be it as small as possible, then that convinces everyone even any journalist and any non-expert.
Unfortunately in the case of numerous concepts at this place already the end is reached. As charming the variation of the magnetic circuit may be, in most cases I miss the unipolar arrangement of the fields. Perhaps here no neutrinos are needed at all and energy merely is withdrawn from the environment heat? I'm not able to answer this question and I surely don't need to, as long as no magnetic converter is demonstrated to me as closed-loop. such a converter, if it can be realized, presumably will at least partly hide one of its poles and produce some open field lines; that at least would be expected.

a still bigger measuring technical problem represent the neutrinos bound to a line, which oscillate around the conductor in the form of ring-like vortices. We have become acquainted with these in the case of the single-wire transmission technique of Tesla (fig. 9.5). They are formed, if neutrinos are slowed down and collected, but not yet have materialized to charge carriers. Tesla did use them for his loss-free energy transmission technology, but he couldn't supersede the alternating current technology full of losses, which also stemmed from him but which he called the worse technology, from the market because there were no power meters available for the single-wire-technology.

Today we still aren't one step further. The energy supply enterprises still decline this technology, as I had to learn myself, although this would be the only way to transport solar energy from the desert or energy from the geothermal energy of Iceland by a sea cable to Central Europe, where it is needed.
Today still no gauges exist for such neutrinos bound to a line. Therefore will every measuring technician experience his Waterloo at SET-devices, in which they occur! I now will report of such a converter and the odd measuring problems.

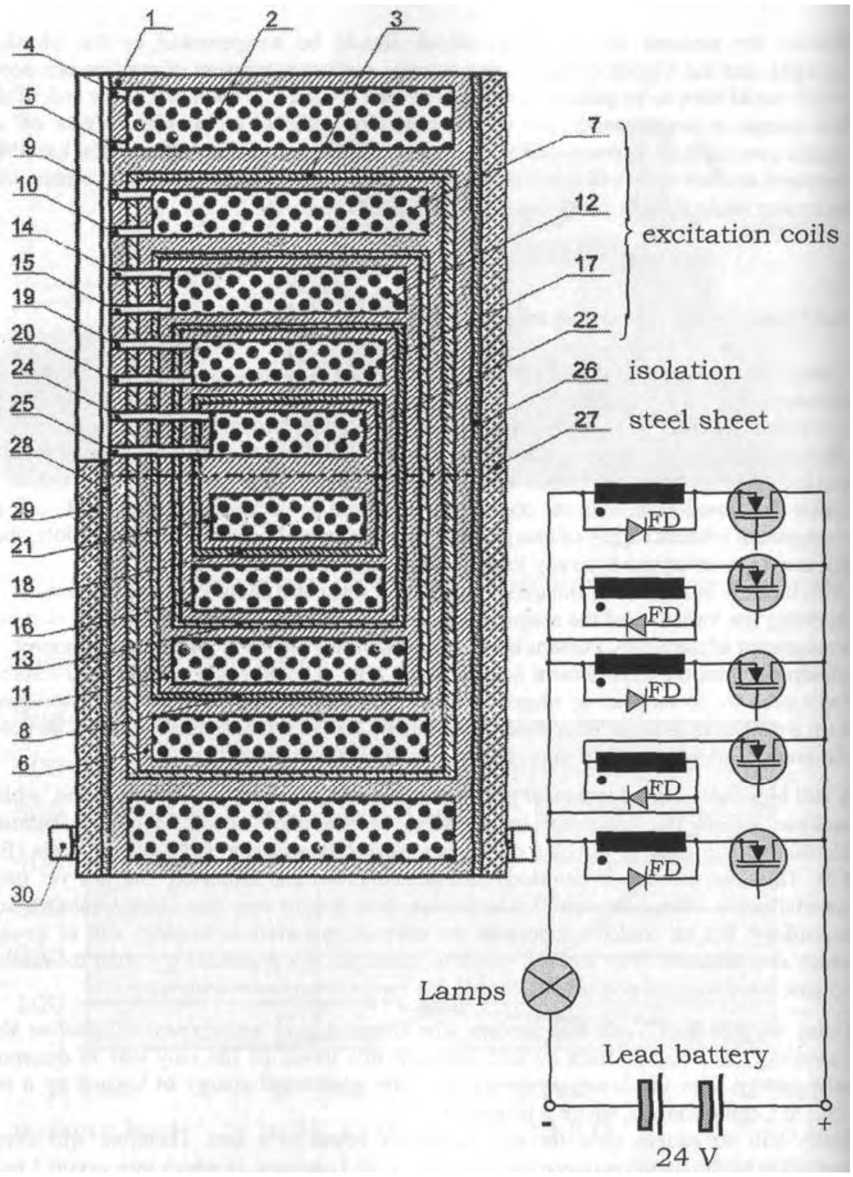
1
2
3
4
5
7
9
10
12
14
excitation coils
15
17
19
22
20
26 isolation
24
27 steel sheet
25
28
29
21
18
16
13
11
8
6
30
FD
FD
FD
FD
FD
Lamps
Lead battery
-
+
24 V

Fig. 16.5: Space quanta manipulator of the Firma ROM AG<i>

<i>: Gibas, Greilinger, Lehner, Rusterholz: Strom aus dem All, Mega Link, Fachzeitschrift fur Elektronik, Nr. 6, Marz 1998, S. 18-23 resp. Bau: Glaube kann Berge versetzen, Bulletin SEV/VSE 25/97, S. 31

16.5 RQM and the space quanta manipulator

If an employee or one of the numerous shareholders of the Swiss Firma RQM AG in Rapperswil speaks of space quanta, then he with that presumably means the neutrinos. An oscillating source of neutrinos is called central Space Oscillator and an operation in resonance of the energy receiver, the so-called space quanta manipulator, is required. Mr Ludwig Sigrist, the creator of this world of imagination, was not a physicist but crane operator (pseudonym Crane O.) and inventor, whom textbook physics couldn't help further in his considerations anyhow. His concept, if it can be translated into a scientific comprehensible language or not, at least helped him personally and gave him the position to create the space quanta manipulator (fig. 16.5).

It consists of several pot coil systems, which are build up and boxed into each other according to the Matrjoschka principle of scaling down. The ferromagnetic core material of the pot spheres (1, 6, 11, 16, 21) should show magnetostrictive properties as distinct as possible. As we will see, this measure gains its actual importance first in connexion with the interaction with neutrinos!

Each pot sphere carries an excitation winding (2, 7, 12, 17, 22), through which alternatively flows a current in opposite direction. By means of this measure, which seems useless according to classic design concepts, one of both field poles is pinned in the centre, in which way the necessary open field lines are produced.

The individual pot spheres are insulated from each other, where the isolation layers (3, 8, 13, 18, 26, 28) should have a high dielectricity. The pot spheres thus in addition form capacitors with each other; even the ending plates at both sides (27, 29) function as capacitor plates. This construction, the core piece of the planned RQM converter, is driven by a transistorised power amplifier with excitation impulses as steeply flanked as possible. According to patent specification<i> the pulse widths are freely eligible in steps of 5 ns betweeen 200 nanoseconds and one second. Besides the frequency also the polarity and DC voltage offset are adjustable.

Every visit of the laboratory leaves a lasting impression: Carefully the engineers grope forward, turn at the frequency and the duty cycle, until the effect suddenly occurs and neutrinos, the space quanta, are being collected. One realizes that tuning parameters with still stronger reactions would be possible, but entirely without any regulation and limitation of power that can't be controlled anymore. The avalanche effect would destroy everything, and so one is further dangling along the brink of the abyss in the development laboratory of the RQM, all the time trying hard to gain control of the hardly understood effect. Concerning the setup and the mode of operation of the pot system the inventor did let his posterity have explanations, but why a diode in free operation FD suddenly loses its rectifying function, for that neither the employees nor specialist visitors of as high as possible scientific rank in the laboratory until now had ready an answer.

<i>: Gibas, Lehner, Greilinger: Vorrichtung und Verfahren zur Erzeugung elektromagnetischer Pulse, Patentschrift CH 687 428 A5 vom 7.5.1996

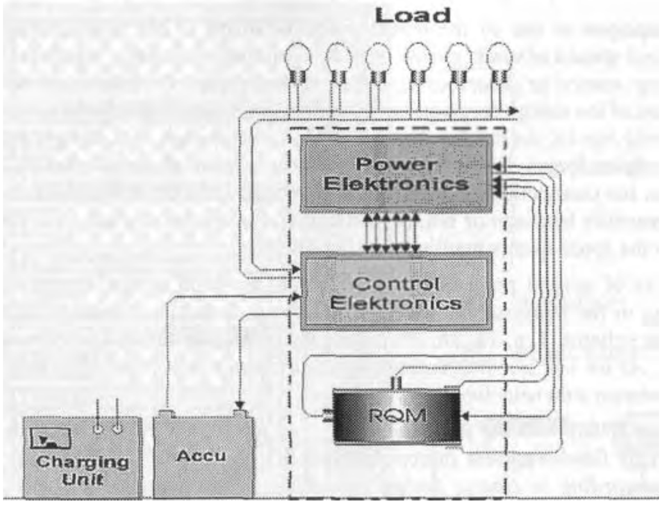
Load
Power Elektronics
Control Elektronics
RQM
Charging Unit
Accu

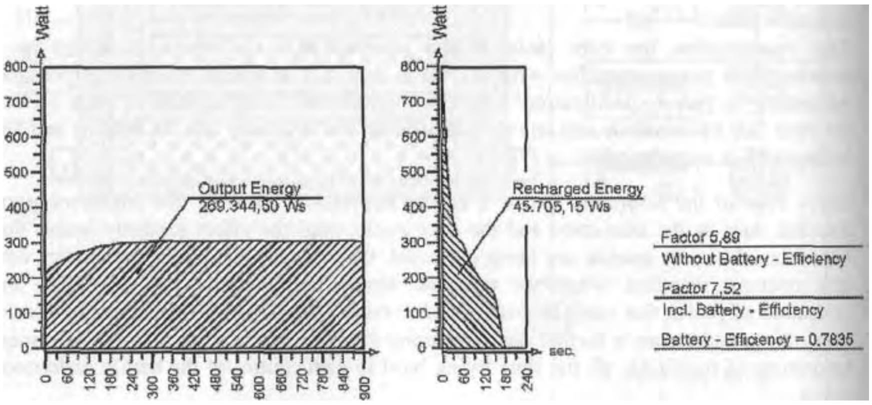

Fig. 16.6: The RQM test installation of 23-9-1996<i>

<i>: F. Greilinger: Der Weg zur erfolgreichen, stabilen Energieauskopplung, RQF Magnetik, Sonderausgabe 1996, page 10

As said, the neutrinos no longer are free but bound to a line, if they leave the pot system. They oscillate around the wire, even around every semiconductor and rove through the entire switchboard. Doing so they can cause quite something, also bring about quite some disaster.

Lead batteries are being recharged, which without doubt is desirable, but in the continuous operation they are destroyed by the same vagabonds. In normal light bulbs for instance some of them materialize to charge carriers, so that according to a gauge more current leaves the light bulb, as on the other side flows in to it. It indeed glows completely normal, but measuring technical for the installation nothing is normal anymore.

The visitor can be shown by means of a high precision measuring facility of vibrations that the pot coils oscillate not only electrically and magnetically, but also mechanically. But the effects alone don't make a converter, which can be produced and which await already numerous buyers of options and licenses.

On the one hand measures for the purposeful conversion of the roving space quanta in utilizable charge carriers here still are missing. On the other hand any controlling facility is missing, to adapt the power taken up from the neutrino field to the momentary need of the consumers. The brave engineers of the RQM still have quite some way in front of them. but the line of approach is right and the reached can be looked at.

16.6 SET-devices with pulsed magnetic fields

In the case of the space quanta manipulator mechanically nothing is moving. In the copper coils the charge carriers merely are on the way with the velocity v. We here have present a typical example of a SET-system with pulsed magnetic field. Compared to the before discussed SET-devices with rotating magnets, for instance the N-machine, the space quanta manipulator clearly has its nose in front. Without brushes, without friction and wear and tear it theoretically has a unlimited life.

In addition it is simpler to produce the necessary large steepness of the flanks in an electronic way as by a mechanical variation of the magnetic field. By means of electronically driving the process it also can be checked, controlled and regulated easier.

The question for an optimisation of the concept still remains. At present one already can be satisfied with a study of possibility, but sometimes one will question the design. There as an example the eddy current losses in the iron pots will be at discussion. About sheeted or sintered materials could be thought, if not at the same time the inductance of the coil would increase in that way, which slows down the increase in current. Ferrite materials again are very brittle and would crumble to dust under the mechanical oscillations of size. I proceed from the assumption, that also here an arrangement with air coils could turn out to be an optimum. It does make sense, if Tesla at higher frequencies always did experiment with air coils.

Now it still depends on the coiling technique. One single conductor loop doesn't provide any open field line. For this purpose if need be two loops have to be supplied with current in opposite direction, like it is the case for a so-called Mobius winding. In this way possibly just as many neutrinos interact as in the case of the pot coils of the space quanta manipulator, which are alternatively supplied with oppositely phased current. To clarify the situation we now should occupy us with the coiling technique.

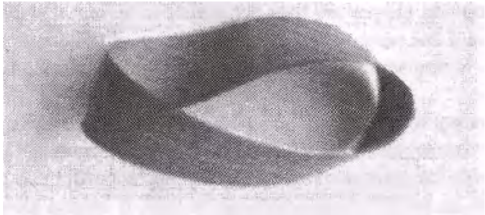

Fig. 16.7 A: The Mobius band

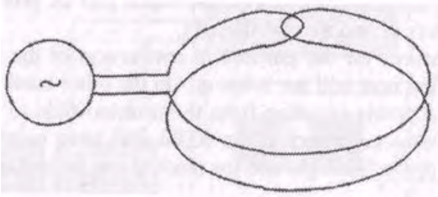

Fig. 16.7 B: The Mobius strip

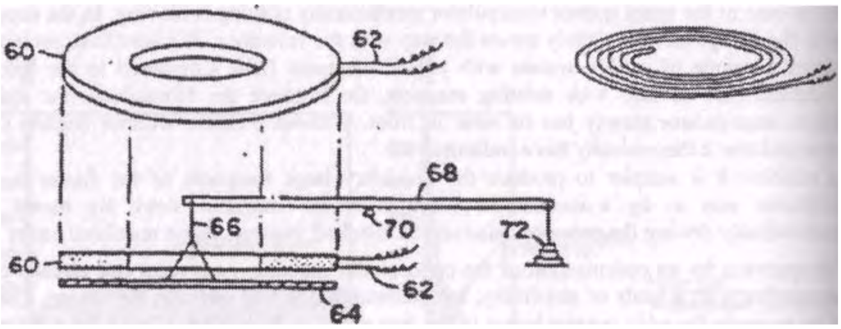
60
62
68
66
70
72
60
62
64

Fig. 16.7 C: Bifilar wound flat coil

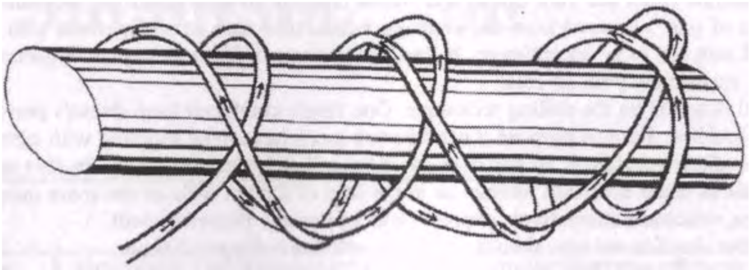

Fig. 16.7 D: Mobius winding arranged like a cross

16.7 SET-devices with Mobius winding

In connection with space energy devices often is talked about the use of a Mobius winding. That is traced back to the Mobius band, which one obtains, if one for example one end of a long paper strip after a half turn glues together with the other end. The result is a strip, which has neither top side nor underside and neither right nor left border. This object with only one side and only one border is a creation of the German mathematician August Ferdinand Mobius (1790 - 1868), fig. 16.7 A.

A distant relative of the Mobius band is the strip with the same name, where the magnetic field lines mutually cancel out. It is the same compensation, as we know of a two core electric cable, where the supply and return cable are run close together. For that the sense of winiding of a conductor loop simply is reversed (fig. 16.7 B).
If the magnetic field vector is wrapped right-handed around the supply cable and left-handed around the return cable, then both amount to zero, so that measuring technical no rest field can be detected anymore at some distance.
THE pointer of the electric field stands perpendicular to the magnetic field vector and points in the direction of the conductor and the movement of the charge carriers. From the coupling of magnetic and electric field follows, that the compensation of one of them also leads to the compensation of the other one. But if the electric field actually is compensated, then there may not flow any current in the winding!
There however flows a current if an electric tension voltage is supplied to the Mobius winding. As a result of this forced current flow both electric and magnetic fields have to occur, which are not compensated! Some field lines will fold outward the expected direction and stand in space as open field lines. These we, in the oscillating case, owe the interaction with the neutrinos.

A perfect compensation would be expected for a two core, bifilar winding. The American physicist William Hooper was able to obtain interesting effects with a bifilar wound flat coil. With help of a balance he could detect attracting or repelling force effects on different, electrically or magnetically neutral discs (fig. 16.7 C)<i>.
But he feeds the flat coil with direct current, so that a contribution of neutrinos can't be expected. Such effects only will occur, if one works with pulsed signals, because the bifilar wound flat coil quite obviously is able to form wanted, open field lines.

<i>: Schneider, Adolf: Energien aus dem Kosmos, Jupiter-Verlag 1994, page 28, taken out of:

Mielordt, Sven: Tachyonenenergie, Hyperenergie, Antigravitation, kaum & Zeit Verlag 1984

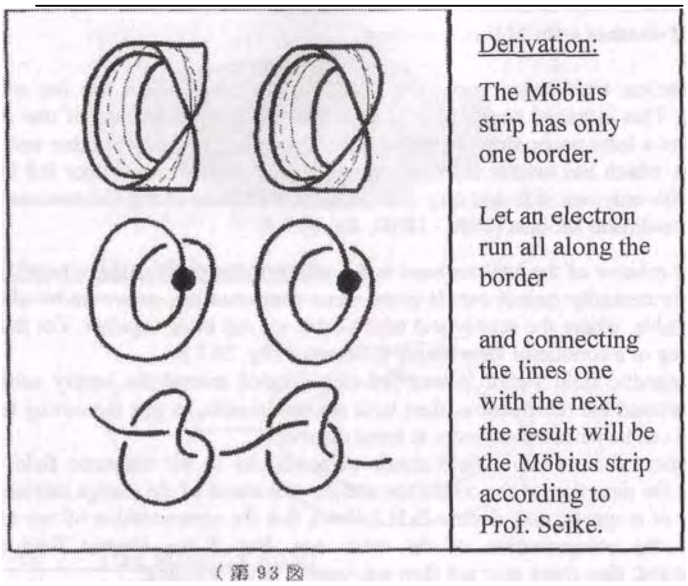

Derivation:

The Möbius strip has only one border.

Let an electron run all along the border

and connecting the lines one with the next, the result will be the Möbius strip according to Prof. Seike.

Fig. 16.8 A: Derivation of the Mobius strip acc. to Seike<i>

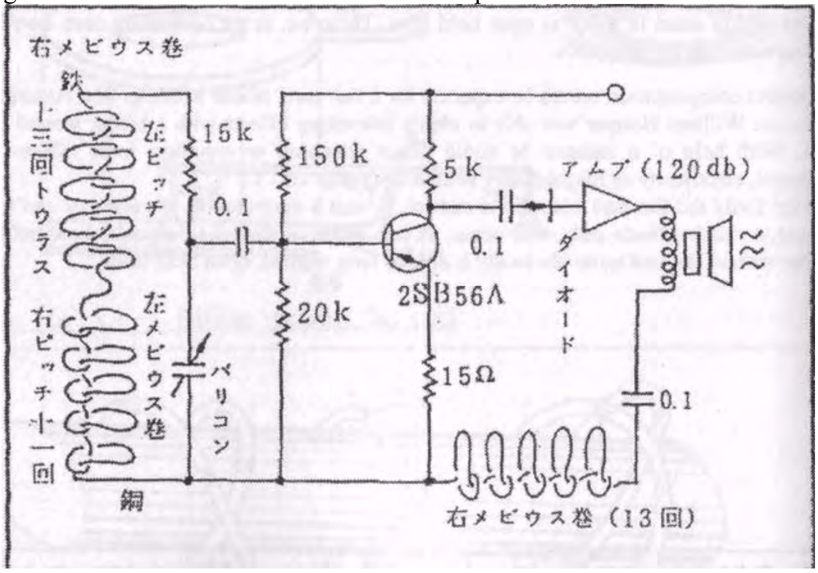
右メビウス巻
鉄
十三回トウラス
右ピッチ十一回
左ピッチ
左メビウス巻
バリコン
15k
150k
5k
0.1
20k
0.1
2SB56A
15Ω
アンプ (120db)
ダイオード
0.1
銅
右メビウス巻 (13回)

Fig. 16.8 B: Scalar wave radio according to Prof. Seike<i>

<i>: Seike, Shinchi: The Principles of Ultra Relativity, Space Research Institute 1990, ISBN 915517-1

16.8 Mobius converter of Seike and Coler

Concerning construction it can be an advantage to wind a Mobius coil not purely bifilar and in that way to do without a perfect compensation. It often already is sufficient, if two conductors cross and only individual components of the field vectors partly cancel out. In the case of the coil, which is pulled apart and wound like a cross, drawn in fig. 16.7 D the conductor current and the fields of supply and return cable belonging to it stand under an angle of almost 90°. This in individual cases should already be able to cause the formation of open field lines.

Similar field conditions are formed, if after every turn the wire is looped under the last winding. At the knots again the angle conditions of approx. 90° occur. The Japanese professor Shinichi Seike preferably works with this kind of winding, which he directly derives from the Mobius band (fig. 16.8 A)<i>. He has designed an electro-gravitation motor basing on this principle. He also speaks of weight reduction and of an artificial anti-gravity field. It could be confirmed experimentally, that his setup cooled down slowly during the operational tests, despite the expected heating up by the copper losses of the current<ii, S.29>. Unfortunately I don't known anything more exact.
In his book further is found the wiring diagram of a radio for the reception of scalar waves (fig. 16.8 B). He thereby quite simply replaces all coils in the high-frequency part with his Mobius coils. If longitudinal waves already are measurable and receivable with this simple measure, first has to be checked.

The German captain Coler more than 50 years ago has developed another Mobius converter. His "Magnetstromapparat", with approximately 6 kW power and an at least four-fold over-unity effect uses six coils with permanent magnetic core. As a peculiarity he in addition runs the coil current through the core under an angle of 90° with regard to the coil current (fig. 16.8 C).

More detailed details can be seen in the footnote<ii, S. 22>.

<ii>: Adolf Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1994

S.22: "Scientists of the Technical University Charlottenburg in Berlin and Munich confirmed that the device functioned without objections, but they didn't find a theoretical explanation for the production of energy. The professors Kloss and Franke of the Technical University of Berlin found a degree of effectiveness of 450%. Prof. Schumann confirmed that 4.8 to 6.7 times more energy came out, than was put into the device. Prof. Schumann excluded a deception entirely as the records prove.
But it wasn't simple to set the device going and the stability for longer periods of time wasn't guaranteed. The war confusion set an end to further research. After the end of the war the British secret service confiscated all available documents and apparatus. Part of the report was declassified in 1962".

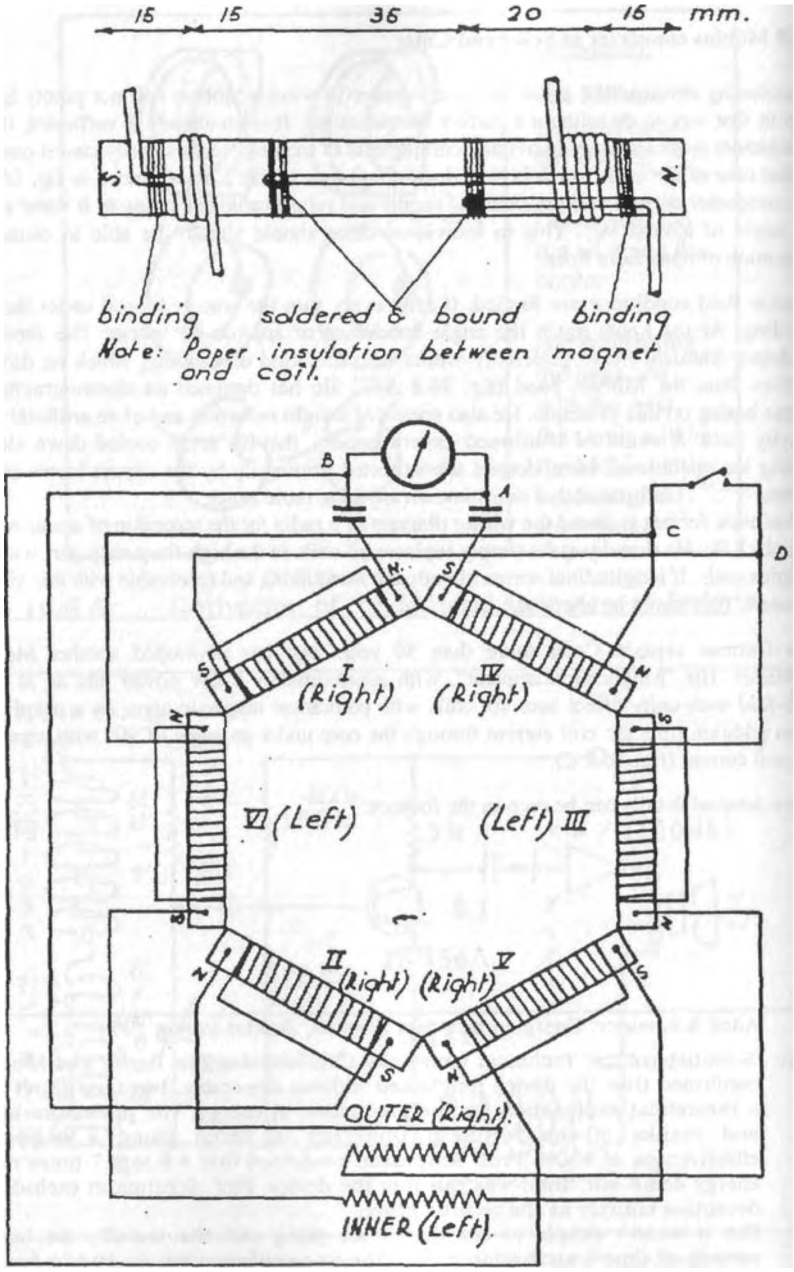
15
15
35
20
15 mm.
N
N
binding
soldered & bound
binding
Note: Paper insulation between magnet and coil.
B
A
C
D
N
S
S
N
I
(Right)
IV
(Right)
VI (Left)
(Left) III
II
(Right)
(Right)
V
S
N
OUTER (Right).
INNER (Left).

Fig. 16.8 C: The Coler converter

16.9 Tesla's flat coil

In the category of the unconventional coiling techniques without doubt the Tesla coil may not be missing. If in schools and high schools such coils were standing in the laboratory for teaching purposes, then as a rule it are cylindrical coils. In reality Tesla worked with flat coils but that, so is said, isn't necessary anymore today, since we have at our disposal better isolating materials than 100 years ago. Actually Tesla contended with problems of osilation, which he could solve with the help of the flat coil, but it should turn out that the coil geometry is attached a crucial importance.

Everything had started with Tesla having to leave the Technical Highschool in Graz without diploma. He ran out of money and he had dared to criticize the venerable Professor Poeschel and his sparking Gramm dynamo. With that he had put himself under compulsion to succeed. Two years later he had ready the solution. In the year 1882 he discovered the rotary field in Budapest.
In the time to come he designs and builds an alternating current motor, but no-one wants to have it and surely Thomas Alva Edison not. Tesla after this disagreement very fast gives up his job at the Edison Company again and again stands under pressure to succeed. With that the eternal bachelor Tesla urges himself to ever higher efforts. He wants to prove himself and the rest of the world that his alternating current system is superior to the direct current system.
Direct current, as is well-known, can't be transformed, and thus the advantage of Teslas alternating current lies in the possibility of power transport by high-tension cable over large distances. But for that the high-tension transformers first had to be developed and thereby the said problems of isolation occurred.
With each turn the tension voltage at the transformer winding increases. The distance to the grounding point lying on the outside has to be chosen bigger with each turn, so that no blow inside of the high-tension winding occurs. A consistent solution of the problem in accordance with engineering is the flat coil used by Tesla, wound spirally from the inside to the outside (fig. 16.9 A)<i>.

It thus is correct that isolation technical reasons led to the flat coil, since Tesla himself was completely surprised as he had to find out that this coil can lose its self-induction, that scalar waves can be detected with it and that it is cooled down during operation in an inexplicable manner.

This cooling effect Tesla has investigated more detailed and after all even used. In his patent specification concerning the superconductivity he describes, that the flat coil also loses its Ohmic resistance, if he in addition previously cools it with liquid air. The remaining cooling down to absolute zero his flat coil obviously has carried out entirely by itself with help of the neutrinos (fig. 16.9 B)<ii>.

<i>: Nikola Tesla: Coil for Electro-Magnets, Patent No. 512,340 (1894)

<ii>: Nikola Tesla: Means for Increasing the Intensity of Electrical Oscillations, Patent No. 685,012 (1901)

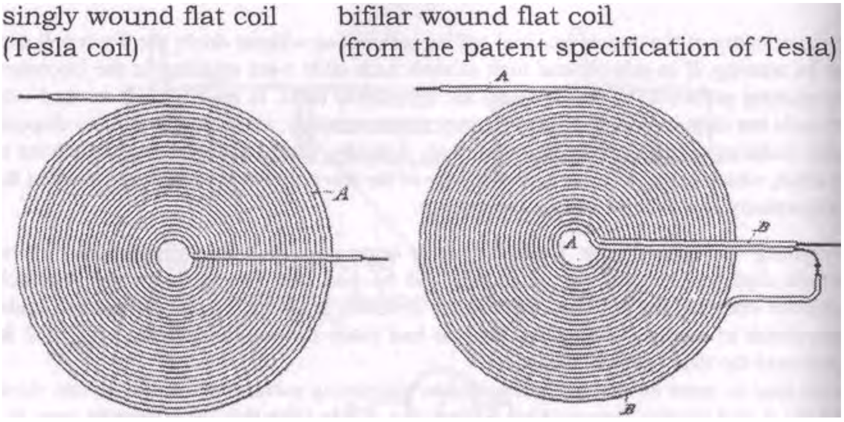
singly wound flat coil
(Tesla coil)
bifilar wound flat coil
(from the patent specification of Tesla)
A
A
A
B
B

Fig. 16.9 A: Coiling techniques of Tesla's flat coil.[i]

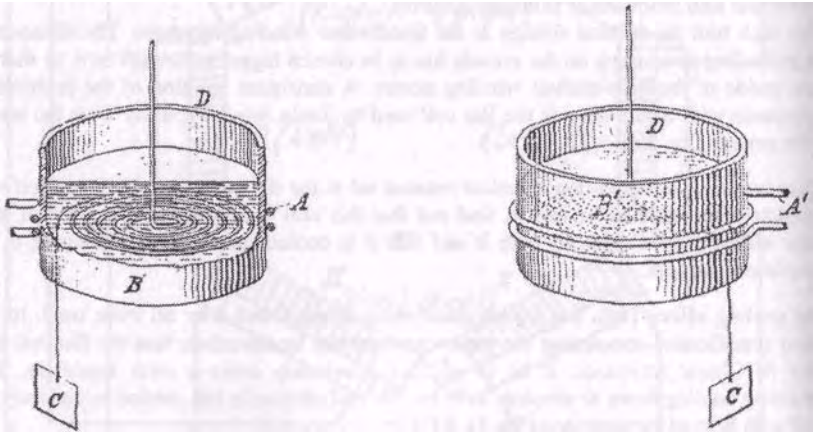

Fig. 16.9 B: Patent specification of Tesla concerning Superconductivity.[ii]

<i>: Nikola Tesla: Coil for Electro-Magnets, Patent No. 512,340 (1894)

<ii>: Nikola Tesla: Means for Increasing the Intensity of Electrical Oscillations, Patent No. 685,012 (1901)

16.10 The secret of the flat coil

The technical function could be explained in the way that the charge carriers of a flat coil by induction are set into motion for excitation from the outside. The transmitted energy shows in form of kinetic energy. The spiral flat coil becomes narrower and narrower towards the inside, the length of each winding shorter and shorter, so that the kinetic energy inevitable has to decrease in favour of a rotational energy. The faster and faster rotating spherical vortices are pulled apart to flat discs and eventually to ring-like vortices by the centrifugal force. The electrons at first become neutrinos bound to a line and finally free neutrinos. Tesla has technically used the first ones in the single-wire-transmission technique (fig. 9.5) and the last ones in his wireless energy transmission (fig. 9.7).

Like many other inventors, Tesla owes also the inventions, which he counts his greatest, like the radio technique and the Magnifying Transmitter, first of all his industriousness. his persistence and a great deal of inventor luck. A magician, as he is called in his most important biography, he by no means was<i>. The flat coil, to which led him chance and which plays a central role in all these inventions, gave him the lucky position, to collect neutrinos and materialize them to charge carriers or in reversed direction to dematerialise electrons to neutrinos.

The technology however is everything else but new. Already the Lituus of the Etruscan and Roman Augurs and the crook of the priests had the same spiral structure (fig. 16.10). In the case of the devices, which the Augurs for instance served at land surveying, it clearly concerns flat coils according to Tesla. We will go into this strange „gauge" more in detail in part 3 of the book<ii>.

The trick probably is, that one component of the electric field pointer is directed towards the centre of the coil and as a result some open field lines are generated, which then collect neutrinos from space.
In this process the neutrinos thanks to the resonant interaction are slowed down to the speed of light and following, as discussed, materialized by means of the flat coil, as in addition rotational energy is withdrawn from the neutrinos. Since the receiver oscillates resonant with opposite phase, in addition the thermal oscillations are reduced and the receiver becomes cold!

If one compares the Mobius coil with the Tesla coil, then besides numerous properties in common the strength of the first coil lies in the production of open field lines and the collection of neutrinos, whereas the special and additional property of the flat coil lies in the materialization, in the conversion of neutrinos into charge carriers. But the advantages of the flat coil have to be bought at the expense of having to work with very high tension voltages (above 511 kV) and with large changes in tension voltage (du/dt). With this set of difficulties we will have to deal in more detail.

<i>: Margaret Cheney. Nikola Tesla, Erfinder, Magier, Prophet (Orig.: Man Out Of Time, 1981), Omega-Verlag Diisseldorf 1995

<ii>: K. Meyl: Electromagnetic environmental compatibility, part 3, edition belonging to the information technical seminar, INDEL Verlagsabteilung 2003.

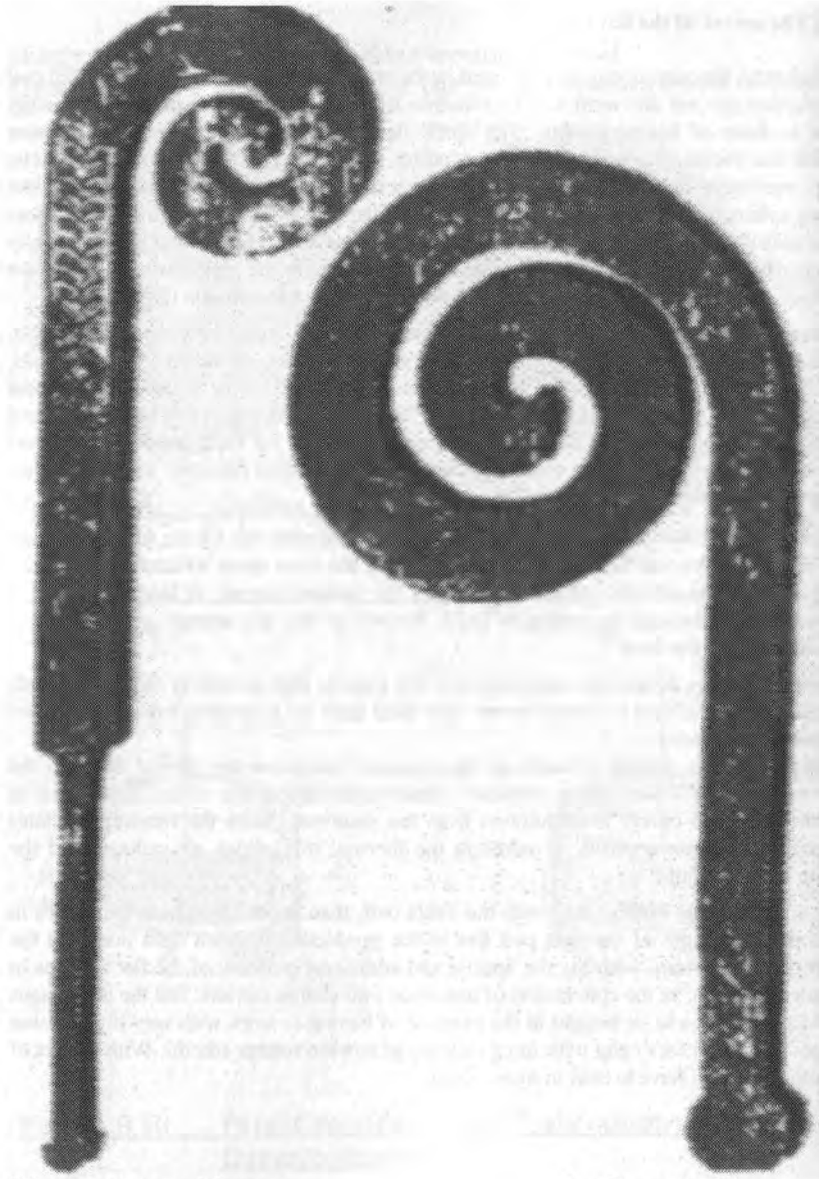

The Etruscan Lituus
the Roman Lituus

Fig. 16.10: The Lituus or crook of the Augurs in ancient Rome

16.11 Discussion concerning the technology of the neutrino collectors

Let us again collect the facts for the discussion: A SET-device is distinguished by a more or less unipolar design and open field lines, with which interact neutrinos, which are oscillating in resonance. These then are slowed down and collected. For the transient process a large change in tension voltage (dE/dt) is required, which can be obtained directly, for instance by means of a spark gap, or indirectly by means of Faraday's law concerning the unipolar induction (E = v x B).

The discussed possibilities concern the acceleration of a machine part (dv/dt), the variation of the magnetic field (dB/dt) by pulse-like excitation signals (16.5) or by magnetic flux variation (16.3) and the railgun. which even can be operated without foreign magnetic field (fig. 15.5 C) and for which in that case occur both a dv/dt and at the same time a dB/dt.

For resting arrangements the velocity v is that of the charge carriers moving in the conductor. So that Faraday's law thereby doesn't lose its influence, the pointers of E and v mustn't point in the same direction, as in the case of "normal" coils. Unconventional windings, which for instance can be knotted like Mobius strips (16.7), take remedial action. Also the ancient crook, rediscovered as flat coil of Tesla (fig. 16.10), proves to be suitable in principle. Here one component of the electric field pointer points in the direction of the centre so that the wanted, at least partly, unipolar arrangement can be formed.

The first step, the collecting of space quanta, shouldn't pose an insurmountable obstacle anymore in view of the numerous possibilities and the detailed explanations. A real difficulty we still have before us, because in most cases some ring-like vortices bound to a line are formed, for which no electronic construction element exists and for which functioning converters hardly are known.

There spoons are bending, some lumps are flying through space, radioactivity is disappearing without a trace, light phenomena are formed and the device suddenly is becoming cold. Almost all inventors, who have arrived in this place, are enthusiastic about the not understood effects or with that are wanting to get attention, but hardly anyone really starts something with that. Until now the necessary system and an useful theory were missing.

Only too often isn't considered, that only an indirect conversion into charge carriers is possible, that during the materialization of neutrinos a intermediate product is formed, which can be described with the model concept of a neutrino bound to a line or of an oscillating ring-like vortex. The technologies collected in this chapter concerning the collecting of neutrinos only form the first step from the free to the bound ring-like vortex. The coming chapter is dedicated to the second step. Here we should try to understand the properties of space quanta bound to a line and loudly think about for which purpose we could use them in practice<i>.

<i>: Reference: Both chapter 16 and chapter 17 treat point 4 according to table 15.1 concerning the resonant interaction (page 86).

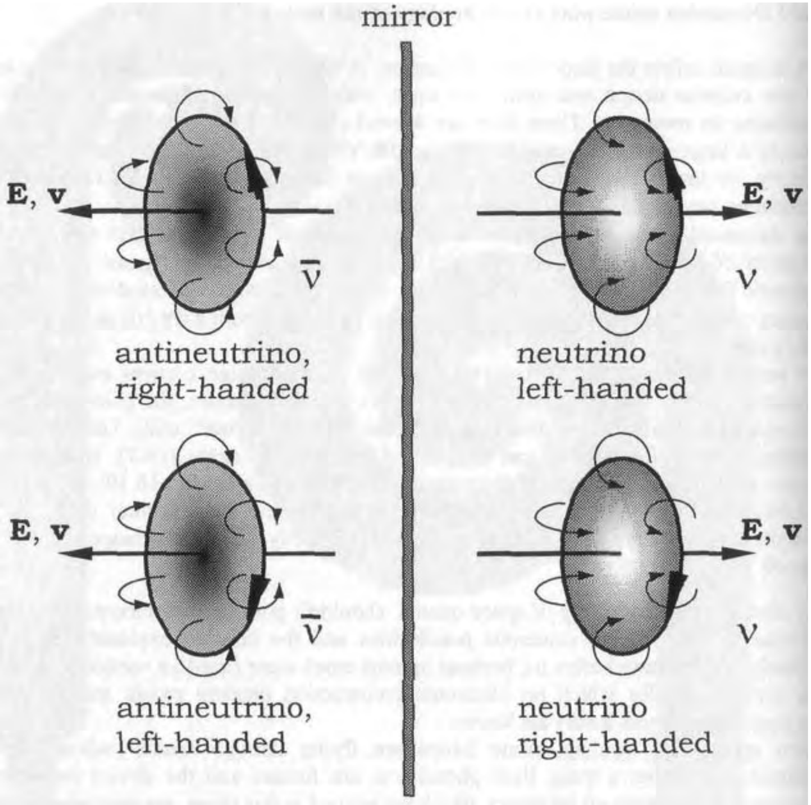
mirror
E, v
$\bar{\nu}$
antineutrino, right-handed
E, v
ν
neutrino left-handed
E, v
$\bar{\nu}$
antineutrino, left-handed
E, v
ν
neutrino right-handed

Fig. 17.1: Application of the mirror transformation to the neutrinos

---

<i>: Kleine Enzyklopadie Atom, Verlag Chemie GmbH, Weinheim 1970, S. 119: „For the neutrinos the quantum number parity isn't defined, because they have a fixed association of the direction of momentum and spin; the sense of rotation of the spin and the direction of momentum in their case form a left-hand wound screw, in the case of antineutrinos a right-hand wound screw. By the shown mirroring the direction of momentum is reversed, but the sense of rotation remains unchanged; i.e. reactions, in which neutrinos occur, are not mirror invariant, they violate the law of conservation of parity". (translated)

Note of the author:, an antineutrino by no means can be assigned to the anti-matter, since it exactly like every neutrino alternatingly takes the matter state and the anti-matter state, by oscillating around itself (see fig. 7.12). The description merely follows a definition founded in usefulness.

17. Technical use of the weak interaction

17.1 Radioactivity caused by neutrinos

Neutrinos are standing in a weak interaction with other elementary particles. This circumstance is known in general. According to the considerations and derivations expressed in the book the neutrinos mediate the resonant interaction, what leads to the conclusion, that the weak interaction represents a partial aspect of the resonant interaction, in which case it merely concerns the proximity.
Efforts are further being undertaken, to combine the weak interaction with the electromagnetic to an electroweak interaction, after it was remarked that a certain coupling constant corresponds in both cases. We of course aren't surprised, because the electromagnetic interaction anyhow describes only the special case of the resonant interaction with the frequency zero.
The weak interaction concerns with the very small range of just 10^{-16} meters only the proximity of the neutrinos, for instance the ß-decay, where the neutrinos for the reason of their oscillating charge a free neutron rattle and shake so long, until it eventually decays. on the average after approximately a quarter of an hour.

The Austrian physicist Wolfgang Pauli had remarked, that half the decay energy after a beta-decay is missing and the balance sheet of energy isn't working out. In addition also the balance sheet of angular momentum isn't working out, because the nuclear spin is being changed for a whole unit. Pauli as a result 1930 has introduced a hypothetical particle without mass and without charge, which he called neutrino.
With that Pauli and his co-working Italian colleague Fermi it is true are the givers of the name of the neutrinos, but not by all means the discoverer. If Cowan and Reines 1956 have detected these particles with large expenditure of devices, then also that by no means was a premiere, as falsely can be read in textbooks and encyclopaedias. After all Nikola Tesla already decades ago had demonstrated, that the neutrino radiation not only exists, but even can be used energy technically.

Now the radioactive decay a ß-radiation occurs, triggered by the conversion of neutrons in protons or vice versa. This obviously takes place under the influence and participation of neutrinos in the atomic nucleus. The ß-radiation to a special extent consisting of electrons and positrons, as it occurs in the case of nuclear fission, is quite unhealthy and by no means ecologically compatible. Under the influence of free charge carriers not only metal lattices become soft and spoons can be bent, but also an electrolysis takes place, where the water molecules are splitted into their parts. That isn't a good prerequisite for the flora and fauna on our planet, which predominantly is built up of water structures.

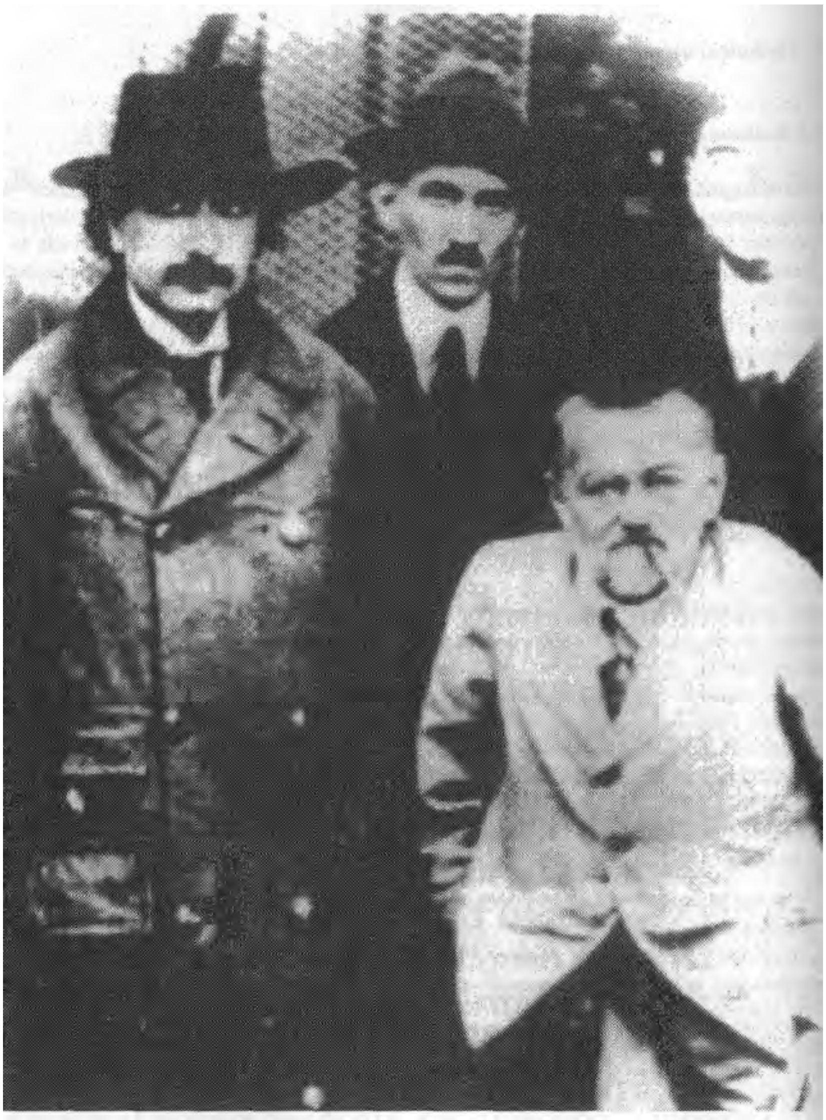

Fig. 17.2: Einstein, Tesla and Steinmetz[i] (from left to right)

---

[i] Franz Ferzak: Nikola Tesla, Eigenverlag, S. 103

17.2 Nikola Tesla, the discoverer of the neutrino radiation

The discoverer of the neutrino radiation himself will best be able to explain the connexion. In the New York Times Tesla writes, that he has discovered and investigated the phenomenon of the cosmic radiation, long before others started their researches<i>:
"According to my theory a radioactive body is only a target, which constantly is being bombarded by infinitely small balls (neutrinos), which are projected from all parts of the universe. If this, at present unknown, cosmic radiation could be interrupted completely, then no radioactivity would exist any longer.
I made some progress regarding the solution of the mystery, until I in the year 1898 attained mathematical and experimental evidence, that the sun and similar celestial bodies emit energy-rich radiation, which consist of inconceivable small particles and have velocities, which are considerable faster than the speed of light. The ability of penetration of this radiation is so large, that it penetrates thousands of kilometres of solid matter, without their velocity being reduced noticeably. "
It must be admired how Tesla guided by experimental observations and a reliable instinct comes to the correct result. He merely with the conclusion, because of the missing interaction the neutrinos have to be inconceivably small, isn't quite right. Their size rather depends on the velocity, because the overfast neutrinos are being length contracted stronger. Tesla however hits the nail exactly on the head, if he on the occasion of the press conference for his 81st birthday declares, the radioactivity is a clear proof of the existing of an outer radiation of cosmic origin<ii>. „ If Radium could be shielded against this radiation in an effective way", Tesla writes in an essay of 1934, „then it wouldn't be radioactive anymore". At this occasion he contradicts Albert Einstein, without thereby pronouncing the name and is indignant at the wrong working method of the scientists<iii>.

Me personally fascinates, how here until now ignored results have been presented, which I first had to work out theoretically myself with difficulty. Tesla, to the best of my knowledge, hasn't taken theoretical derivations, at least none have been handed down. As a brilliant experimental physicist he must have reached his conception world by means of the measuring technique. The perfect correspondence of his experimentally determined and the by me theoretically won insights should be judged as evidence for the correctness of this view.

<i>: Dr. Tesla Writes of Various Phases of his Discovery, New York Times, Feb. 6, 1932, P. 16, col. 8

<ii>: Tesla Said (J.T. Ratzlaff), Tesla Book Company, ISBN 0-914119-00-1, P. 272

<iii>: "The scientists of today think profound instead of clear. One has to be mentally sane, to be able to think clear, but one can think profound and nevertheless be completely insane. The scientists of today have substituted experiments by mathematics, and they travel from one equation to another and eventually build up a construct, which has absolutely no relation to reality"...

taken from N. Tesla: Radio Power will Revolutionize the World, Modern Mechanics and Inventions, 7/1934, (Tesla Said, P. 264)

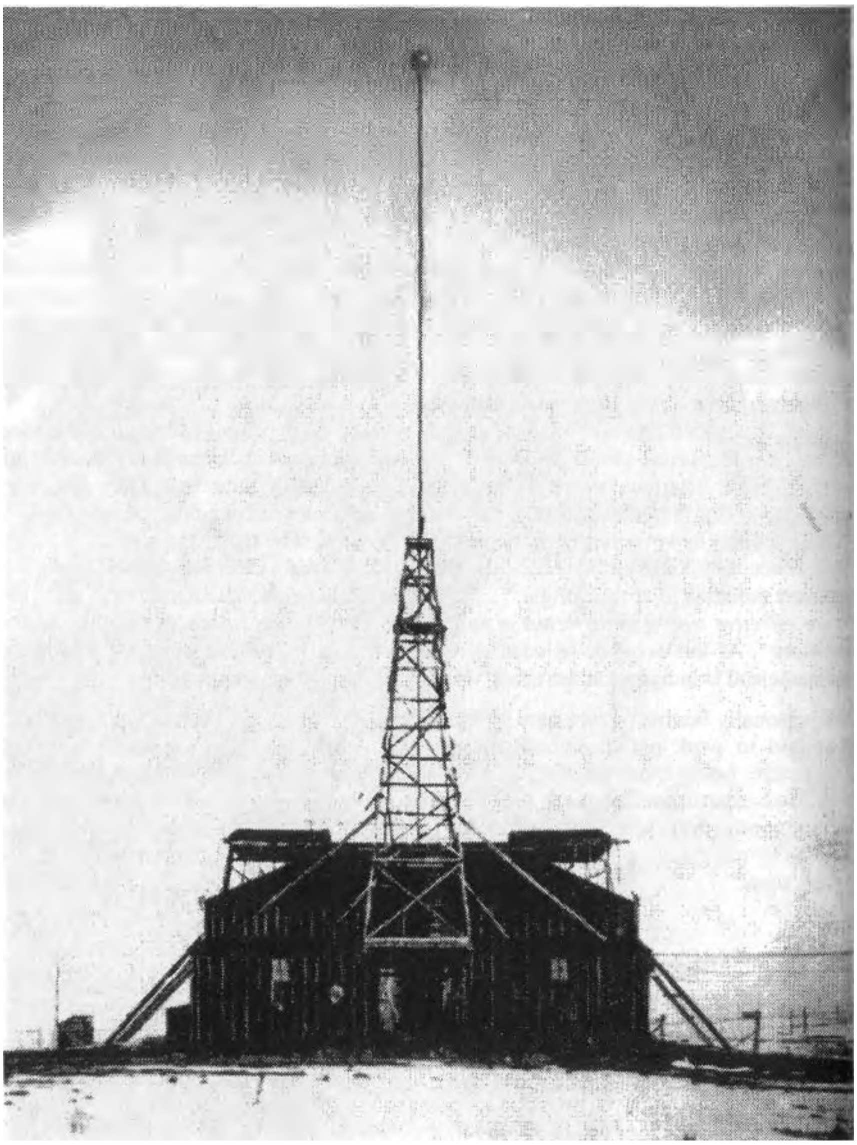

Fig. 17.3: The Magnifying Transmitter of Nikola Tesla Would the research station in Colorado Springs (1899 - 1900) have been suitable for transmutation?

17.3 Transmutation and reduction of radioactivity

If we are capable to collect and bundle up neutrinos, then should with that an influencing of the radioactive decay become possible, then also a decontamination of radioactively contaminated material and a so-called transmutation should be possible. We here have a concrete possibility before us, to use the produced ring-like vortices directly for the benefit of humanity. Dangerous transports with Castor containers, permanent and temporary storage and the contamination of whole areas by radiation, like around the nuclear power station Tschernobyl, which got out of control, wouldn't be necessary at all. If namely burned out fuel rods and objects contaminated by radiation have undergone a concentrated neutrino shower, then the radioactive decay takes place accelerated, so that the half-life can be drastically reduced. After the treatment the dangerous special waste would have been changed to harmless domestic rubbish. Even recycling or reuse are feasible.

The topic is at least as explosive as the energy question and as well completely unsolved. That's at different places and in some companies is already feverish researched about technologies concerning transmutation. Often it are the same people, who also work at the theme of space energy. The reason quite simple is that in both cases neutrinos have to be collected and bundled up. In the case of transmutation however the necessity of materialization can be dropped, so that the goal can be obtained faster and simpler. Consequentially is reported of more cases of a successful decontamination and of transmutation, than of functioning energy converters. Until now most techniques however still hardly are suitable for bringing into action technically on a large scale, but they carry clues of a solution of the problem already in them.
At the congress "New Energy technologies from USA" 6.12.1997 in Zurich I have lectured concerning the theme<i>: „Presumably as the first one the doctor Dr. Wilhelm Reich has carried out corresponding experiments with his „Orgon accumulator" (fig. 9.2). From him also stem warnings about biological effects, which should be taken serious, if radioactive material is put under his Orgon accumulator and the process of decay takes place accelerated".
If one dares an interpretation of his experiments with only 1 mg radium, then numerous charge carriers materialized at hitting upon his sample of the bundled up neutrino radiation with the consequence of high electrostatics in the environment, which Reich has called DOR-state (Deadly ORgone) (chap.15.11).

Another way would be the rebuilding of a neutrino transmitter according to the plans and patent specifications of the experimental physicist Nikola Tesla (fig. 9.11). He 100 years ago had realized a real neutrino-broadcasting and for that developed an unconventional switching technique. Tesla called his transmitter a „Magnifying Transmitter" (fig. 17.3). He choose the name „Magnifying Transmitter", after he had received more energy than he had transmitted in experiments and this effect moreover was increased with increasing distance to the transmitter. He obviously also had collected free and synchronously oscillating neutrinos, and that would be the best prerequisite for a successful transmutation and decontamination of radioactive material!

<i>: Konstantin Meyl: Die Wechselwirkung der Neutrinos; uber Massnahmen, die Halbwertszeit beim radioaktiven Zerfall herabzusetzen. NET-Journal Jg. Nr. 3, Jan./Feb. 1998, S. 14-20

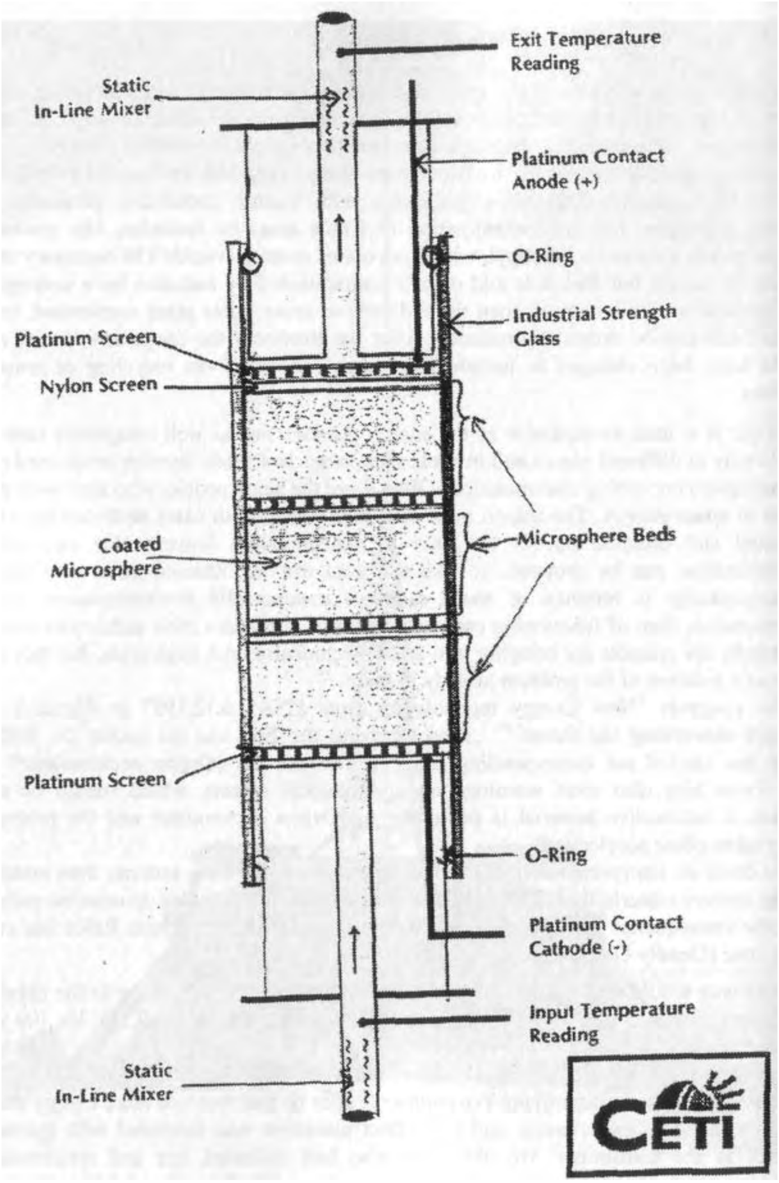
Exit Temperature Reading
Static In-Line Mixer
Platinum Contact Anode (+)
O-Ring
Industrial Strength Glass
Platinum Screen
Nylon Screen
Microsphere Beds
Coated Microsphere
Platinum Screen
O-Ring
Platinum Contact Cathode (-)
Input Temperature Reading
Static In-Line Mixer
CETI

Fig. 17.4: 1 kW Patterson Power Cell.<i>

<i>: Patterson, J. A.: System with electrolytic cell and method for producing heat and reducing radioactivity of a radioactive material by electrolysis. US Patent No. 5.672.259 of Sept. 30, 1997, as reprinted in Infinite Energy, July-Nov.1997, pp. 13-16

17.4 The Patterson Power Cell

As an example worth paying attention to, the Patterson Power Cell should be mentioned, which not only can be used for generation of energy but, how could it be otherwise, also for reduction of radioactivity and for transmutation. The energy cell invented by the chemist Dr. James A. Patterson is researched and developed further by the company CETI (Clean Energy Technologies Inc.) in Sarasota (Florida, USA). According to latest reports such a device supplies at most one kilowatt of heat energy at an over-unity of up to 4000; thus 4000 times the energy is released than is taken up as required for the operation!

In fig. 17.4 one single cell of the energy converter is shown. The glass container consists of three chambers of approx. 4 cm height each time and about 1.9 cm in diameter and works like a continuous-flow water heater. From the bottom tapwater or distilled water is supplied and at the top connection again taken away in heated form. The excitation takes place electrostatically by means of two platinum electrodes. The anode situated at the top is connected with the positive pole and the cathode with the negative pole of the source of direct current.

In the three chambers are situated tiny, filmy coated small balls, which form the real principal item of the cell. Production method and structure of the beads are oriented to the task to be mastered. In the case of a variant conceived for the generation of energy the barely one millimetre in diameter small synthetic beads carry after each other a thin layer of palladium, a layer of nickel and once again one with palladium<i>.

The small balls remind with their layered structure at first of the Orgon accumulator of Reich. They also seem to function as collectors of neutrinos, but for smaller wavelengths and much higher frequencies. In addition is chosen a concentric arrangement with the spherical form, which with the electrically conductive surface and the dielectric core fulfil the function of a cavity resonator. Resonance actually is possible with flying past neutrinos, which have a wavelength which amounts to an integer multiple (1,2,3,..) of twice the diameter of the sphere.

The arising resonant oscillation however concerns not only the electric and magnetic fields, but by means of electrostriction and magnetostriction also a mechanic oscillation of size. The oscillation of size in reverse causes again oscillating electric and magnetic field pointers, which are partly open along the spherical structure and can interact resonant with further neutrinos. In this repercussion field the actual secret of the functioning of a Patterson-cell is hidden.

The president of the German association of space energy, Prof. Dr. Dr. Josef Gruber on the occasion of his visit at the company CETI and of a conference about „cold fusion" in Vancouver has published a report<ii>:

<ii>: Gruber, J.: Kalte Fusion und Raumenergie, report concerning the 7th international conference about "cold fusion" of 19.-24.4.1998 in Vanvouver, B.C., Canada (Infinite Energy, Issue 19, 1998) and report concerning "International Symposium on New Energy" of 23.-26.5.1997 in Denver, Colorado, USA (Denver-report of the remote University Hagen)

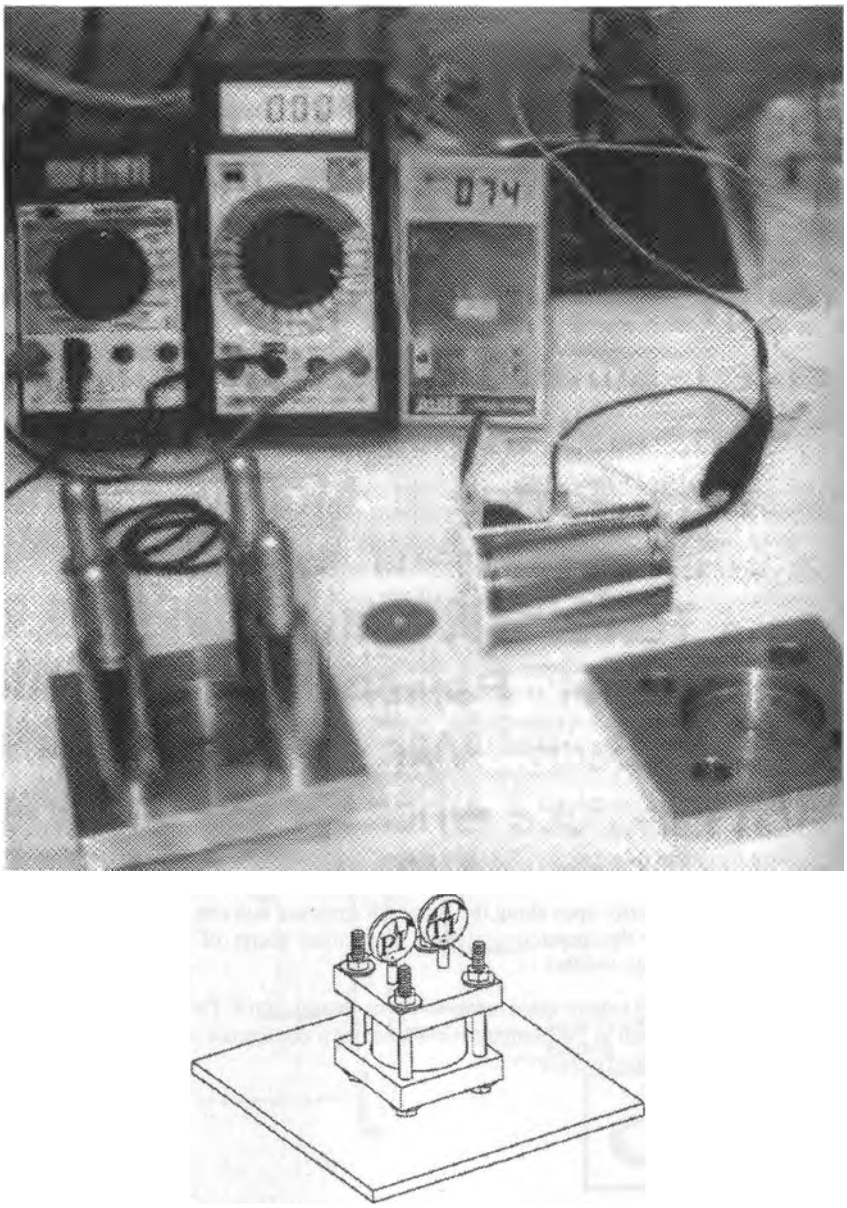

Fig. 17.5: _____ Low Energy Nuclear Transmutation Cell,[i]
the kit for experiments LENT-1
(Showing Temperature and Pressure Gags)

<i>: taken from the NET-Journal, 10/11 1997, S. 7

17.5 Report concerning the cold fusion

"In the case of the Patterson transmutation cell it concerns a special electrolysis cell, in which the radio nuclides are charged. During the electrolysis the decay activity measured by means of a Geiger-Muller counter decreases drastically. Within a few hours reductions of up to 80% are obtained. In such an electrolysis system with electrodes consisting of specially coated beads low-energetic nuclear reactions can be observed. In this case elements are detected in the metallic layers of the beads, which before were not contained in these. Further are measured changed (unnatural) proportions of isotopes. From these transmutation can be inferred, which is used in the cell for the conversion of radioactive elements.
Until now natural uranium and thorium were used as radioactive material. The reproducibility already now is very good. The applicability for commercial purposes, for the reduction of the radioactivity of burnt up fuel elements and for the conversion of plutonium is easy to see".
At his visit in the laboratory of CETI in Sarasota 28.5.97 Prof. Gruber has witnessed an experiment to annihilate radioactivity with the Patterson transmutation cell. He reported about it: "In the presence of N. J. Olson from Pacific Northwest Laboratory (operated by Batelle for the U.S. Department of Energy) a team of the television station ABC conducted by the science journalist Dr. M. Guillen made recordings for a television program, which meanwhile has been broadcasted all over the country. The original tension voltage - among others a Geiger counter rattled - subsided as the expected success became apparent in the experiment: After one hour 50% of the radioactivity stemming of uranium nitrate were removed, in another half an hour further 13%. Doing so also a considerable lot of surplus heat was produced.

New techniques of this kind to reduce radioactivity have an important advantage: One brings the SET-device there where the radioactive material is situated and reduces the radioactivity on the spot. For the conventional technique, still being in research, in contrast to that the radioactive material first is wrapped in a complicated manner and then transported to a special factory, where radioactivity is reduced at great technological expenditure and use of energy - altogether a procedure which is relatively costly and politically only hard to carry through".

Patterson cells already can be obtained commercially for research purposes. The same goes for a demo-kit<i> of the company Trenergy, Inc. of Hal Fox, who has built up the biggest database of the world concerning the theme "Low Energy Fusion" in the "Fusion Information Center" at the university of Salt Lake City. He is editor-in-chief of the "Journal of New Energy".
If one believes the statements on their web-site<ii>, then one is capable of converting 10 g of radioactive thorium in 900 mg titanium and some copper in less than an hour with the kit for experiments LENT-1. Under a pressure of 3200 psi sodium already is said to have fusioned to gold, something of which alchemists dream, as an unintentional side effect, as researchers in Cincinnati/Ohio say<iii>.

<i>: NET-Journal, Jan. 1997, S. 24
<ii>: see Internet: http://www.hal-fox@slkc.uswest.net
<iii> : NET-Journal, 10/11 1997, S. 7

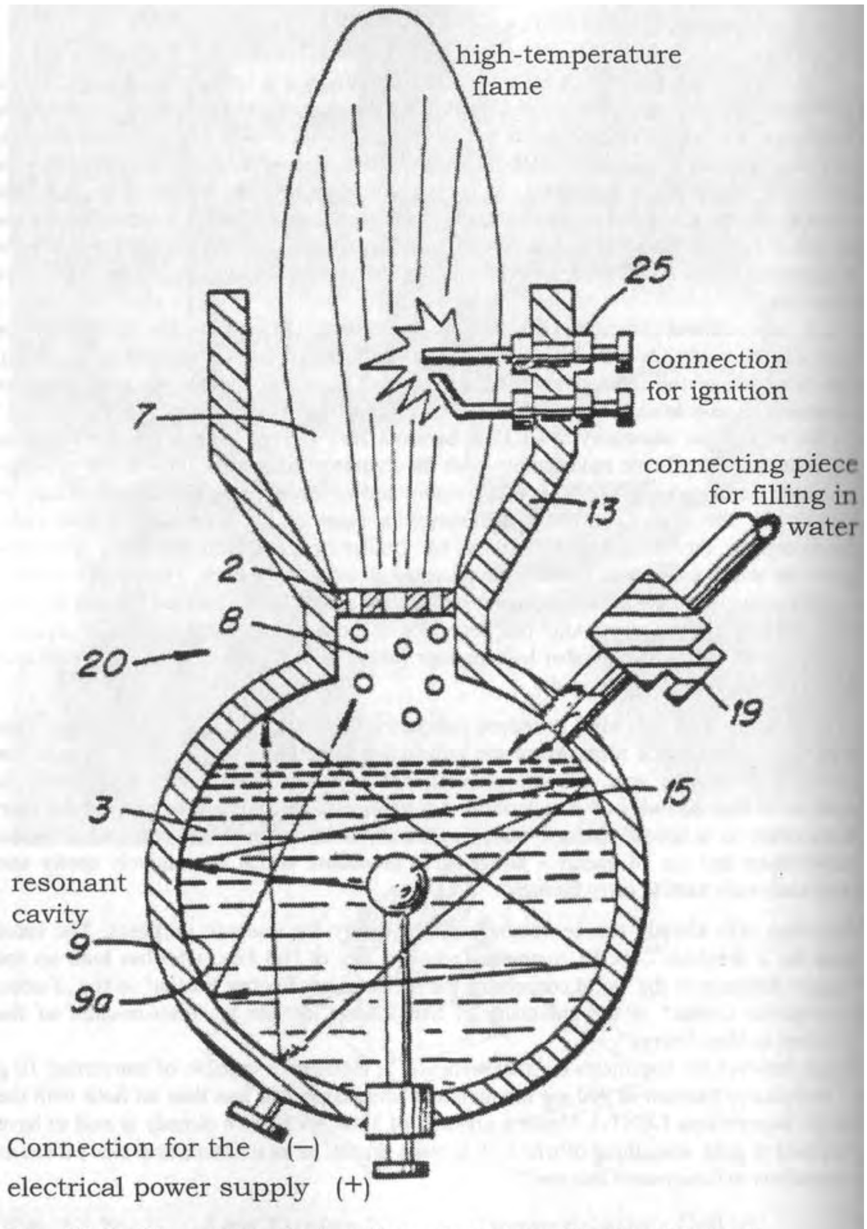
high-temperature flame
25
connection for ignition
7
connecting piece for filling in water
13
2
8
20
19
15
3
resonant cavity
9
9a
Connection for the (−)
electrical power supply (+)

Fig. 17.6: The water-fuel-cell according to Stanley Meyer<i>

<i>: H.-J. Ehlers: Stan Meyers Wasser-Zellen-Technik, Raum & Zeit special 7, p. 201, taken from the Canadian patent specification 1234 773 of 5-4-1988

17.6 Water-Fuel-Cell technology

The United States are, just what concerns spectacular techniques like the removal of radioactivity or the transmutation of new materials, obviously still the land of unlimited possibilities. Nowhere the list of researchers of cold fusion is as long as in North America.
But there also in many places only is tried and tinkered without visible system or usable theory. Then in many cases unfortunately only a show-effect is to the fore, while construction plans and details concerning the way of functioning, as far as they actually exist, are kept secret. In total there, besides a lot of hot air and wrong hopes, is left behind little to be used and cited.
By the way, in my opinion cold fusion has to do very much more with space energy and neutrinos than with hot fusion. How much disaster a wrong referring to and an unusable theory can bring about, has become clear at the example of the "cold fusion researchers" Fleischmann and Pons. They have placed themselves into the scientific offside with their misinterpretation concerning cold fusion.
Moreover are both primarily scientists and theorists. Practice however is, according to general definition, "if in spite of all it functions!" The American Stanley Meyer is such a practical man and his water-cell-technology actually seems to function, although he in his 18 patent specifications gives theoretical explanations, which he just as well could have saved himself. With that his effect isn't explained.
An usable interpretation would be that this device, comparable to a cell of a plant during photosynthesis, splits water molecules into its parts by putting on neutrinos. Even without knowledge about space quanta the buggy of Stanley Meyer already runs with a air-cooled 1500 cc VW-engine, and it consumes no gasoline at all. The tank is filled with water; it even may be sea water. The consumption of water lies at 2.8 litres at 100 kilometres and thereby is formed predominantly hot steam again as a combustion product.
If thus cold water is converted into hot water and at the same time mechanical energy is available. then inevitably another source of energy must be involved. According to my interprelation it must concern the neutrino field. If the here presented details are correct then the over-unity effect lies at approx. 100, the degree of effectiveness thus at 10000 percent.

American companies, with which Stanley Meyer had concluded contracts, should make the "Water-Fuel-Cell technology" mature for series. Also the financing seemed secured. But then per internet the message came, he 21.3.1998 was having supper in a restaurant in Grove-City, as he suddenly jumped to his feet from the table and called out, he had been poisoned. He died on the spot<i>.

Alarge number of inventors is known, who tap space energy with the help of water. It concerns an increase of the content of oxygen or of the content of colloids, thus an improvement of the water quality. Or it concerns formation of vortices as already in the case of Walter Schauberger, glowing phenomena or also the generation of free energy from the neutrino field. In this concert the concept of Stanley Meyer takes an outstanding place, as particularly efficient, instructive and clearly understandable for us, for which reason we cast our eyes over the design (fig. 17.6 and 17.7).

<i>: Zum Hinschied von Stanley Meyer", NET-Journal April/Mai 1998, S. 25; see: obituary in Infinite Energy Magazine

A. The pattern of the pulses of the excitation voltage:

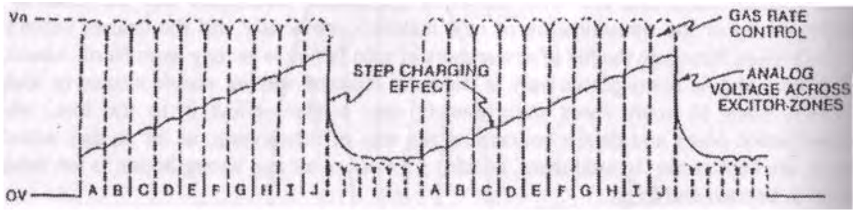

B. The wiring diagram:

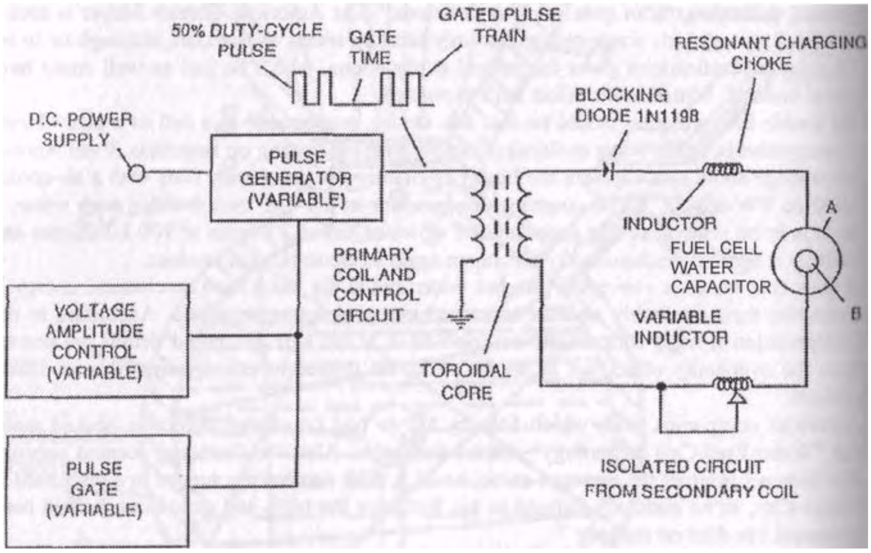
50% DUTY-CYCLE PULSE
GATE TIME
GATED PULSE TRAIN
RESONANT CHARGING CHOKE
BLOCKING DIODE 1N1198
D.C. POWER SUPPLY
PULSE GENERATOR (VARIABLE)
INDUCTOR
A
FUEL CELL WATER CAPACITOR
B
PRIMARY COIL AND CONTROL CIRCUIT
VARIABLE INDUCTOR
VOLTAGE AMPLITUDE CONTROL (VARIABLE)
TOROIDAL CORE
ISOLATED CIRCUIT FROM SECONDARY COIL
PULSE GATE (VARIABLE)

Fig. 17.7: Wiring diagram for the driving of the water-cell[i]

[i]: Stanley, A. Meyer: Process and Apparatus for the Production of Fuel Gas and the Enhanced Release of Thermal Energy from such Gas, 15.06.89, US-Pat. 207,730; International Publ. WO 89/12704; Int.Appl. PCT/US89/02622

17.7 Unconventional electrolysis

Numerous of the by Stanley Meyer used construction principles already have been treated, be it the excitation with pulses of electric tension voltage or the spherical structure of the resonant cavity (fig. 17.6). As a spherical capacitor with the positive pole in the centre and the negative pole at the outside edge it corresponds to the model of the electron and fulfils in an almost ideal way the conditions of an unipolar arrangement according to fig. 15.8 B. The use of certain patterns of the pulses and steep flanks of the pulses (large dU/dt) make possible effects of resonance at frequencies starting at ten kilohertz, in which neutrinos participate increasingly. First the series resonant circuit, consisting of the adjustable, external inductance and the spherical capacitor, is stimulated by means of the current rectifying wiring (fig. 17.7). In the case of resonance, which is carried out by comparison of the inductance, the excitation current drops, whereas the tension voltage at the same time teaches values of more than 1000 Volts. If in addition a neutrino resonance occurs, then the known avalanche effect will occur. The equipment then takes up virtually no current.

As a dielectric serves water, with which the container is filled permanently. The dimensions are oriented after the velocity of movement of the water molecules according to details of the inventor. Also the oscillating water molecules should go into resonance. Then they can help to materialize the neutrinos. Their rotational energy partly passes to the water molecules and as soon as the neutrinos have been converted to charge carriers, they will take the water molecule from the oppositely charged side and split it without further ado. The oxygen and hydrogen gases leave the capacitor through fine openings at the upper edge of the spherical chamber, which are so small, that no ignition back can occur, and in the simplest case reach a combustion chamber, where they burn again to water as a high-temperature flame (fig. 17.6).

The gases of course also can be guided into the cylinder of a Otto engine and be ignited there, as in the case of the experimental buggy. In the sectional drawing can be seen a filler by means of which, according to the consumption, water is refilled. The round resonant cavity not necessarily has to be spherical. Stanley Meyer more frequently differs from the ideal form and works with a cylindrical symmetry (fig. 17.8), with which obviously in spite of that the goal can be obtained, if perhaps not quite so good. To this compromise the explanations of fig. 15.8 C apply.

If we, to conclude, cast our eyes over the wiring diagram which Stanley Meyer discloses in his patent specification (fig. 17.7 and 17.8)[i]. In the centre is a transformer, which should produce an if possible high tension voltage. A rectifying diode, which takes care that only positive tension voltage pulses serve the excitation, is switched in series with the reaction capacitor, which is filled with water, a fixed and a variable inductance. In that way the positive pole is always situated in the centre of the reaction chamber. If both connections would be exchanged or the diode be turned over, then the neutrinos presumably would materialize in positrons and not in the wanted electrons. But if one leaves out the diode entirely and one has a tension voltage changing both in positive and in negative direction, then maybe electrons and positrons are equally generated, which annihilate each other under emission of gamma quanta. Doing so no gas is formed, but at most light, as long as the effect of collecting neutrinos isn't lost also. The concept should be worth to be examined more detailed already of pure scientific interest.

A: The high-tension transformer:

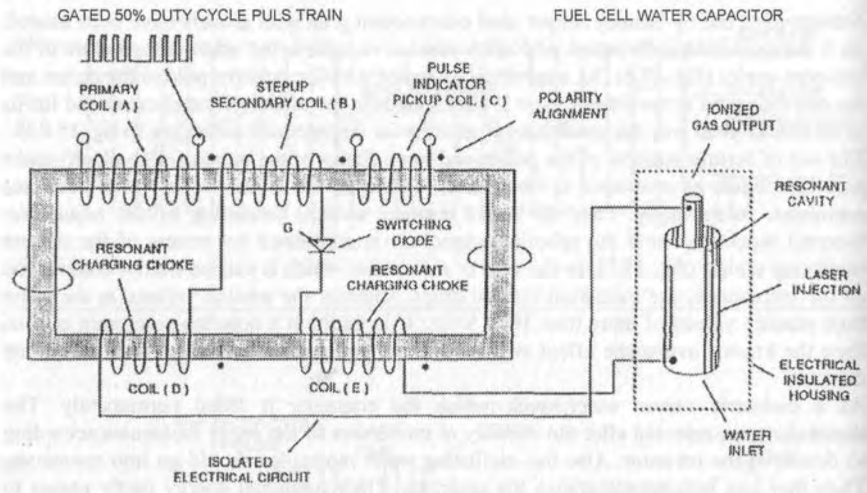
GATED 50% DUTY CYCLE PULS TRAIN
FUEL CELL WATER CAPACITOR
PRIMARY COIL (A)
STEPUP SECONDARY COIL (B)
PULSE INDICATOR PICKUP COIL (C)
POLARITY ALIGNMENT
IONIZED GAS OUTPUT
RESONANT CAVITY
G
SWITCHING DIODE
RESONANT CHARGING CHOKE
RESONANT CHARGING CHOKE
LASER INJECTION
ELECTRICAL INSULATED HOUSING
COIL (D)
COIL (E)
WATER INLET
ISOLATED ELECTRICAL CIRCUIT

B: The driving:

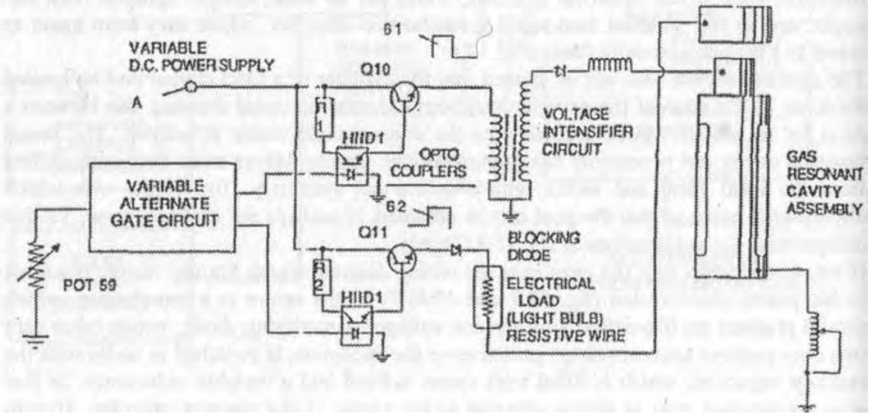
61
VARIABLE D.C. POWER SUPPLY
A
Q10
VOLTAGE INTENSIFIER CIRCUIT
H11D1
OPTO COUPLERS
VARIABLE ALTERNATE GATE CIRCUIT
62
Q11
GAS RESONANT CAVITY ASSEMBLY
BLOCKING DIODE
POT 59
ELECTRICAL LOAD (LIGHT BULB) RESISTIVE WIRE
H11D1

Fig. 17.8: High-tension transformer and driving of the cylindrical water-cell<i>

<i>: Stanley A. Meyer: Process and Apparatus for the Production of Fuel Gas and the Enhanced Release of Thermal Energy from such Gas, 15.06.89, US-Pat. 207,730; International Publ. WO 89/12704; Int.Appl. PCT/US89/02622

17.8 Materialization of neutrinos

In the centre of attention of the water-fuel-cell of Stanley Meyer is on the one hand the unipolar arrangement (according to fig. 15.5 B or C) to collect the neutrinos and on the other hand the water for slowing down and materializing. Over and above that a lot was tinkered and tried, as can be inferred from the patent specifications (fig. 17.8). That starts at the. coiling technique of the transformer and concerns the experiments with laser stimulation as well as the top part, which according to the inventor should extract the electrons (Electron Extractor Grid). The measures may bring an improvement, but are insignificant for our considerations.

At this place it primarily concerns the question of the materialization of neutrinos. But if such a materialization has to be made complete, if water molecules must be splitted, depends on how long the process lasts. If everything goes very fast, then perhaps it is sufficient that a neutrino for a short time is showing as an electron, before it again oscillates back. In this short time the splitting process already could have taken place. The used neutrinos for that had to be very low-frequent and very slow. They after that could again leave the reaction chamber and fly on as neutrino.

Possibly the fishes, which live in stagnant waters or in the deep sea, owe the slow neutrinos the content of oxygen in the water. Because here no bundling up takes place, the splitting of water takes place rather by chance. The volatile hydrogen atoms escape very swiftly, whereas the big oxygen atoms are left behind in the water.

Now we still don't know how charge carriers can be won. At the RQM unit resonant oscillations of size had been measured and that can be judged as an important clue. If namely a space quantum is slowed down then it becomes bigger. The inverse case we already had made us clear: If a particle is correspondingly fast then it is sufficient length contracted to fit through the tunnel (fig. 6.14).

To slow down neutrinos according to that the target area should carry out an oscillation of size with opposite phase. Organic material and biological systems are excellently suited for that. Every contraction of a muscle brings that to mind. Inorganic matter and our technology however normally don't know this phenomenon. Technical energy converters simply are built up fundamentally else than biological muscle machines.
In historical sources at certain places is pointed to the fact that priests had experimented with quartzes and miraculous phenomena were observed. Of course the question is asked, if such experiments today still can be reproduced. Because the density of the earth slowly increases due to the growth in volume, also the wavelength of these oscillation quartzes is changed. The same neutrino radiation therefore today can't be active anymore. It thus would be completely inappropriate wanting to reject a historical source only because a described effect today doesn't want to function anymore.

Magnetostrictive or electrostrictive material could be a solution. For instance a piezocrystal, which contracts under the influence of an outside field. As an ultrasound converter an applied alternating voltage leads to the emission of a sound wave. If we reverse the function, then a received longitudinal wave should lead to an electric tension voltage and then at last the materialized charge carriers could be taken off,!

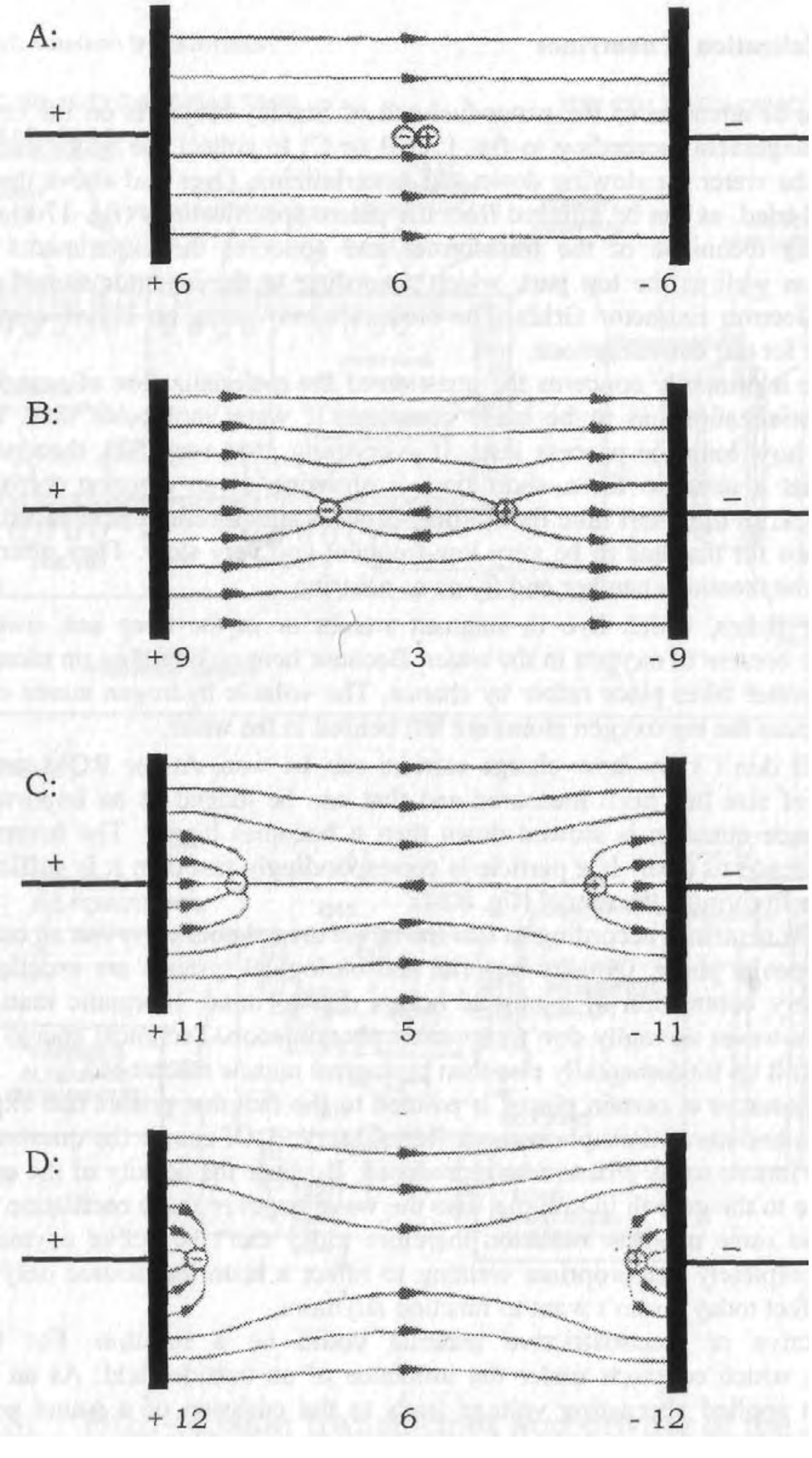

Fig. 17.9: The course of the field in the case of a glow-discharge<i>
(The numeric values indicate the active field lines, as they can be read from the representation).

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12. Aufl. Springer-Verlag (1988), S. 197

17.9 Oscillation of size and luminescence

There exist substances, which at high temperatures start to glow. In the sense of the theory of objectivity presented in the first part of the book in the case of temperature it actually concerns an oscillation of size of the involved elementary particles (fig. 8.3). If this oscillation of size lies in the visible frequency range, then we can directly perceive it and it can be used for instance in a light bulb for the purpose of lighting.
In the case of an arc. fluorescent tubes or glow-discharge lamps light is formed as well even without a thermal effect. These glowing phenomena are called luminescence. They can be caused by chemical processes, by friction, by crystallization or by electric fields. As can be shown speedily, in these cases it also concerns an oscillation of size.

If we for that consider the model on the left (fig. 17.9). Between two electrodes there is a non-conducting gas, e.g. air. If now a tension voltage is applied then, under the influence of the electric field, some gas atoms are splitted into positively charged ions and negative electrons (case A) and pulled apart (case B). This process of enlargement of the glow-discharge fast as lightning comes close to that of the thermal oscillation of size and obviously in the same way is perceived as light.
The ionised gas parts are attracted by the unlike poled electrodes and move towards these (case C, D). Whoever takes the effort to count the number of the field lines, will find out that between both electrodes the electric field drops, whereas it at the same time increases at the electrodes. The first thing causes that the process of ionisation is stabilized by itself and a state of equilibrium will result; the resulting current takes a constant value.
The increase of the tension voltage at the electrodes on their behalf causes that an arc remains standing or the glow-discharge lamp glows on, even if the feeding voltage is reduced. For this reason a fluorescent lamp needs a starter, since the network voltage is too small to start the effect of luminescence.

We indeed know, that the gain of light of a fluorescent lamp is at least three times better than that of a light bulb. As a rule a degree of effectiveness in spite of that can't be given, since it merely concerns a comparison measurement in the case of the measurement of the lighting intensity with a luxmeter. It hence can't be excluded, that we already are dealing with an over-unity effect in individual cases and neutrinos are involved in the effect of glowing.
A necessary oscillation of size would be present, as the perceivable noises of open spark gaps prove. Also in the case of lightning the thunder occurs as a consequence of a longitudinal oscillation of size and at the same time a glowing phenomenon occurs as a consequence of materialized neutrinos (fig. 14.11). It almost is obtruding a scientist that here the same principles are at work.

One however hardly can prove that neutrinos are involved in the luminescence, because the configuration is symmetrical with regard to the resonance of neutrinos and just as much particles materialize as anti-particles, which afterwards again annihilate completely as an impulse of light. They hence can't be measured directly. It gets interesting if an asymmetrical arrangement with unipolar character is chosen.

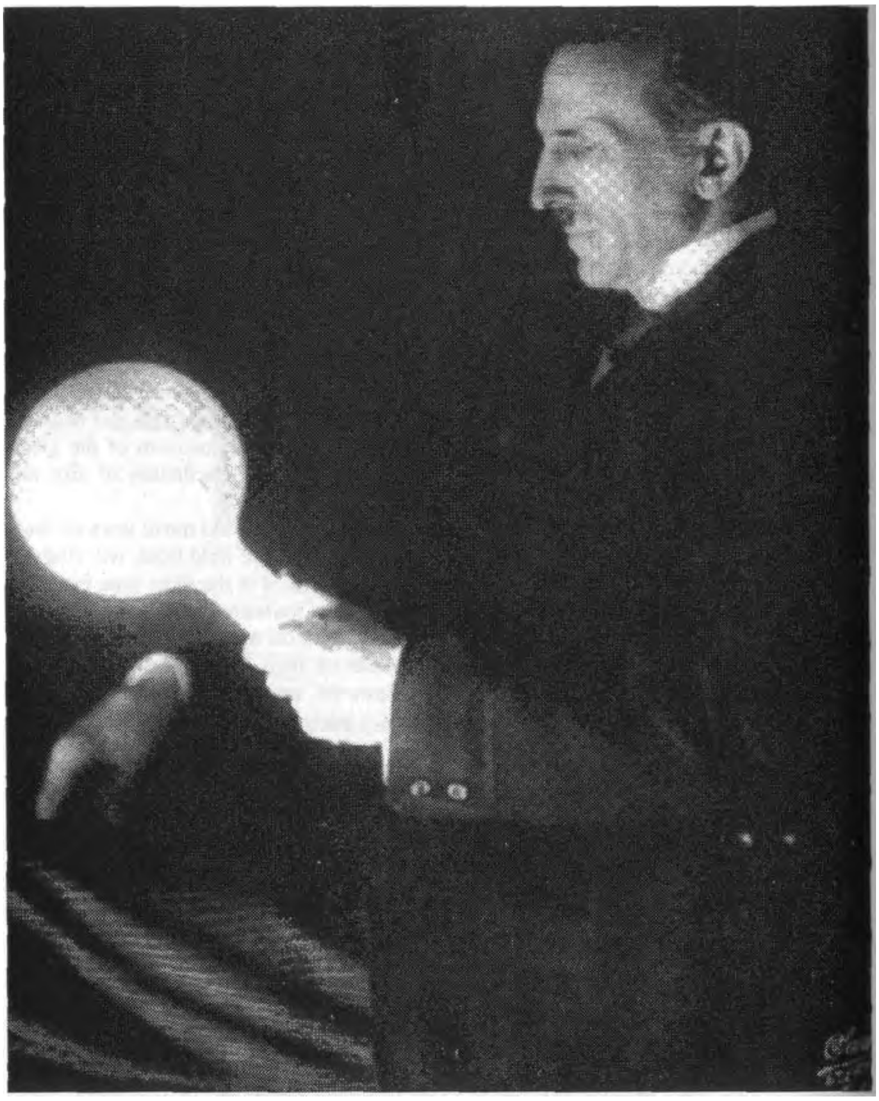

Fig. 17.10: Tesla with a wireless light bulb[i]

[i]: A Sarony-Portrait from 1894, Tesla-Museum, Belgrad

[ii]: R. Hiller, K. Weninger, S. J. Puttermann, B. P. Barber: Effect of Noble Gas Doping in Single-Bubble Sonoluminescence, University of California, Los Angeles, USA, Science, Vol. 266, 14.10.1994, P. 248-250

and: D. Lohse: Wenn sich der Schall in Licht verwandelt, Mechanismus liegt noch im dunkeln; (University of Twente, NL) Mitteilungen der DFG 4/98, S. 19-21

17.10 Phosphorescence and sonoluminescence

Let us first look at the observable after-glowing at fluorescent lamps or at a screen, the phosphorescence. It presupposes a storing effect, and that in accordance with prevailing textbook opinion can be traced back to the excited state of some atoms. Doing so electrons in the atomic hull change from one level of resonance, an instable state of energy, to another orbit, which represents the ground state. Doing so the difference in energy is emitted in the form of light. The process obviously not only takes place as luminescence during the switching on of the excitation voltage, but also as phosphorescence after the switchng off.

An enveloping electron now doesn't fly as a tiny planet around the nucleus, but occupies the entire orbit as an inflated spherical vortex (fig. 5.5). Seen from the outside it is a matter vortex. It however sees the enveloping electrons on the further on the outside lying orbit from the inside and there they are showing as anti-matter vortices. As long as the distance of respect is kept, nothing happens. If however an inner spherical vortex presses to the outside or the outside one to the inside, then the incompatibility of the vortices takes effect and both annihilate under emission of a photon. In this respect the explanations of the vortex model are very helpful.

But now, after this flash of light, two electrons are missing in the atomic hull. The positively charged atomic nucleus never would allow this loss. Replacement has to be fetched and that actually only can stem from the neutrino field. Therefore the enveloping electron doesn't change the orbit immediately, but instead has to wait, until a suitable neutrino passes by, with help of which the game of changing places can be executed. This explains the time delay and gives reasons for the observable after-glowing.

The form of the unipolarly charged sphere (according to fig. 15.8 B) forms almost ideal prerequisites for an interaction with neutrinos and the step-like change of size from one orbit to the next ideal prerequisites for their materialization. The longitudinal wave connected with the change of orbit without doubt can be called a high-frequency sound wave. It however can't be detected because of missing gauges for such high frequencies.

But if we take the frequency down into the range of the ultrasound, which can be handled technically, then effects arise which brilliantly prove this interpretation and the participation of neutrinos in the luminescence.

The not understood phenomenon is called sonoluminescence and at present is researched at numerous high schools primarily for academic interest. The structure is conceivably simple. One takes a ball of glass filled with water and positions at the edge one piezocrystal next to another. Then one with the piezo loudspeakers, operated with the same phase, sounds the whole with ultrasound and see there, the water glows mysteriously! The sound waves change the pressure. Inferring from the observations, during the phase of low pressure small bubbles are formed, which at the following rise of pressure collapse and emit a ultrashort flash of light. There thus takes place an oscillation of size, which leads to the luminescence phenomenon!

So far so good; but such a flash of light is much shorter than the collapsing of the gas bubble lasts. With less than 50 picoseconds it is faster than the otherwise usual atomic transitions of electrons. The whole thing obviously has nothing to do with this kind of luminescence. Also the spectrum doesn't fit and finally the over-unity effect lies at one trillion, the light energy thus is 10^{12} times larger than the energy part which is taken up from the sound wave by every atom<ii>!

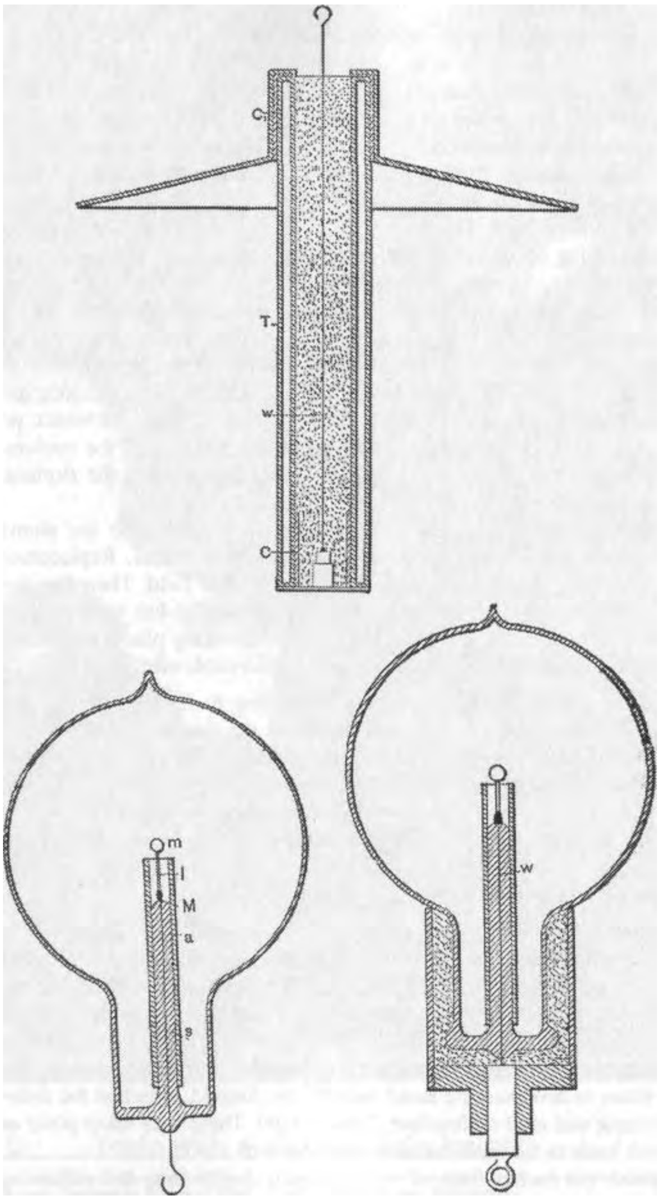

Fig. 17.11: Light bulbs according to plans of Tesla<i>

<i>: N. Tesla: Experiments with Alternate Currents of High Potential and High Frequency, Lindsay Publications Inc, ISBN 0-917914-39-2, Fig 18, P. 70+103

17.11 Discussion concerning the use of the weak interaction

In the case of the luminescence the conditions lie similar like those for lightning. The science of today it is true has some problems of explanation at detail aspects, but again and again manages to protect the ivory tower of physics from collapsing, with supporting auxiliary explanations. But at the latest in the case of an unipolar arrangement, in the case of a ball-lightning or the sonoluminescence, the participation of the neutrinos clearly comes to light, the auxiliary concepts prove to be unusable.

It surely is no accident that the discoverer of the neutrino radiation also was the first, who has experimentally investigated the luminescence. The lamps of Tesla (fig. 17.11) all were without wear and didn't have filaments like those of Edison. They however had to be operated with high tension voltage and relatively high frequency. Both a single-wire and a wireless transmission technique were possible (fig. 9.5 and 9.7). In his laboratory Tesla only needed to hang a fluorescent lamp without any return cable on a wire, then he had light. Famous also are the pictures on which he holds a lamp in his hand, which glows entirely without any connection, but only if he takes it in his hand! (fig. 17.10).

If we again turn through the chapter, then it is remarkable that in most cases, from the transmutation to the luminescence, neutrinos can be used for reason of their resonant interaction which in the proximity can be put equal to the weak interaction. A real materialization however poses an enormous difficulty. If namely the translatoric motion of the space quanta is slowed down, then the risk exists that for balance the rotation increases. But that as well has to be slowed down for a materialization! Only if both processes of slowing down take place, the goal can be reached. In that case the described cooling down effect occurs.

For a measurement technical check it therefore offers to record the slowing down of the translatoric motion and the collecting of the neutrinos by means of the measurement of the radioactive decay of a sample. And to read the materialization as the slowing down of the rotation from a measurement of temperature. In addition are helpful proofs about oscillations of size, glowing phenomena and deviations in the balance sheet of energy. With that it should be able to obtain a system for the exploration of these phenomena with the goal of a practical exploitation of the neutrino radiation for the benefit of humanity.

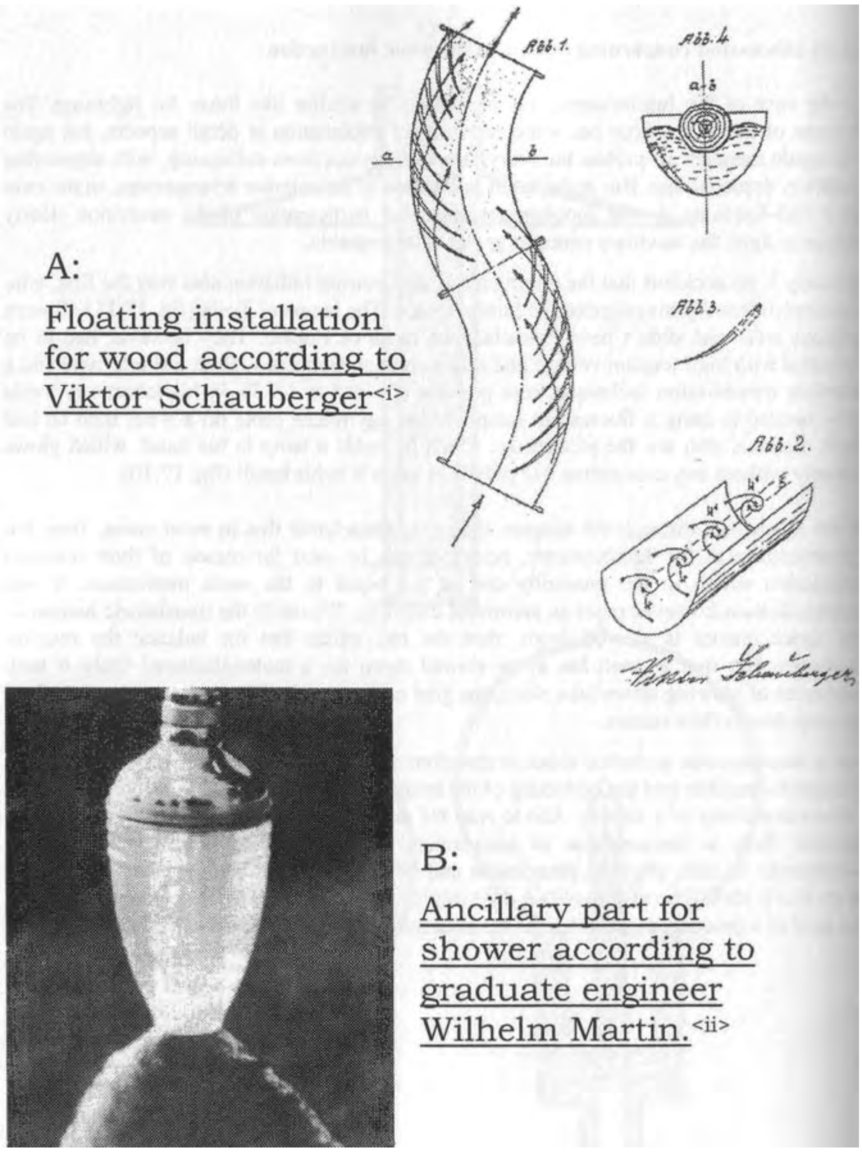
Abb. 1.
Abb. 4.
Abb. 3.
Abb. 2.
A:
Floating installation for wood according to Viktor Schauberger<i>
B:
Ancillary part for shower according to graduate engineer Wilhelm Martin.<ii>

Fig. 18.1: Mechanical whirling of water

<i>: N. Harthun: Naturgemasse Stromungsfuhrung nach Viktor Schauberger Analyse einiger seiner Patente und Zitate; MuT Nr. 4, 1980; s.a. Kap. 9.2

<ii>: O.Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993, S.156

<iii>: V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 112, S. 39.

<i4>: Do we owe the taste of bubbling spring water to neutrinos?

18. Physical phenomena

In this chapter it on the one hand concerns indirect effects of the resonant interaction and on the other hand effects of gravitation and levitation.

18.1 Water as a catalytic converter for neutrinos

Already the Austrian forester Viktor Schauberger has pointed to glowing phenomena at whirled water. He first had observed such in nature at torrents and waterfalls in the Alps. Later he was capable, to produce and to demonstrate this effect even artificially<iii>. One could speak of hydroluminescence, where still the question would be left open, how it actually functions.

Already in chapters 4.10 and 4.11 has been talked about the special properties of water. It here at first concerned the property of transport of a vortex, which even is capable to bind particles in the vortex, which are heavier than the whirling medium itself. Schauberger had become famous by letting build floating installations, in which not only tree-trunks, but also gravel could be transported down to the valley, without the sides of the channel actually having been touched by the rubble. This phenomenon has been examined and confirmed on the scientific side <i>.
Basing on this principle today different devices for processing water are offered and sold, which bind lime, mineral materials and suspended matter in the vortex and in this way prevent deposits in the pipes. The methods, to whirl the water, are however very different. Some whirl the water mechanical, others magnetically and again others electrically.
Here the dipole nature of the H_2O-molecules has an effect. If I turn a water molecule with its electric charge distribution, then from the moving charge a magnetic field results. If thus in the pipe a hydrodynamic flow vortex is produced, then an electric and a magnetic vortex, the potential vortex and the eddy current, are the result (see fig. 4.2). The wanted flow vortices vice versa also can be obtained, by guiding bubbling whirled water past permanent magnets, or by feeding in alternating magnetic fields with the help of coils, or finally by working with pulsed electric potentials. Each of these systems, operated passively or actively, has its specific advantages and disadvantages (fig. 18.1 B and C).

It can be assumed that with the vortices also the water quality is changed. As a rule the content of colloids increases, due to which the surface tension falls. In the colloids negative ions are bound, for which reason also the electric conductivity decreases. Finally the content of oxygen increases and that actually only can come from an electrolysis. The neutrinos thereby are suspected <i4>.

As an extreme dielectric medium water favours the formation of potential vortices, which immediately after their formation contract swiftly. This oscillation of size of the electrically charged potential vortices makes possible actually an interaction with neutrinos, and that on the one hand has as a consequence the water splitting and the increase of the content of oxygen in the water and on the other hand the above depicted, observable glowing phenomenon, the hydroluminescence.

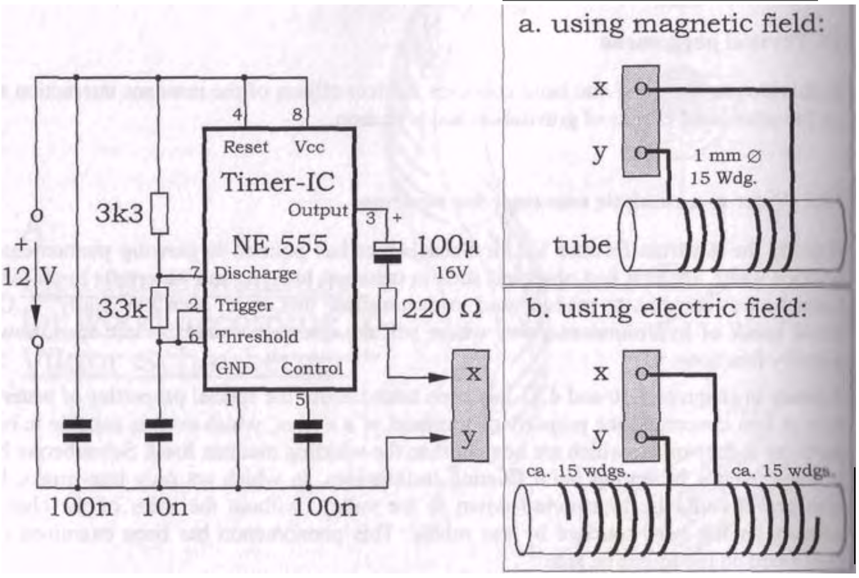
a. using magnetic field:
x
y
1 mm Ø
15 Wdg.
tube
b. using electric field:
x
y
ca. 15 wdgs.
ca. 15 wdgs.
4
8
Reset
Vcc
Timer-IC
Output
3
+
NE 555
100µ
16V
7 Discharge
2 Trigger
6 Threshold
GND
Control
1
5
3k3
33k
220 Ω
+
12 V
x
y
100n
10n
100n

Fig. 18.1 C: Wiring diagram for a simple water processor with pulsed field (2 kHz rectangular signal)
a: for magnetic field excitation
b: for electric field excitation

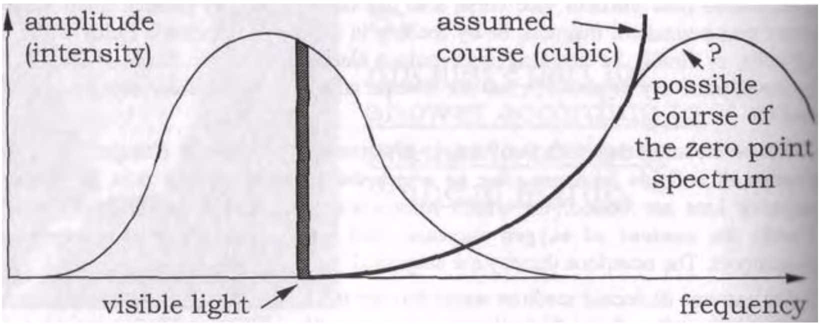

Fig. 18.2: The spectrum of frequencies of light and of the zero point radiation.<i>

<i>: A. Waser: The Puzzling Nature, AWVerlag Luzern 1996, p. 88

To complete in fig. 18.1 C the wiring diagram of a simple water processor is shown, which works with pulsed electric (b) or magnetic (a) fields.

If we temporarily again leave the theme water and take along the notion, that it must be attributed a great importance in connexion with the interaction and the materialization of neutrinos

18.2 Background and zero point radiation

Whathappens really, if cold matter or interstellar gas molecules are hit or touched by the everywhere present neutrinos? Then it can be expected that oscillations are being stimulated and as a result the temperature slightly increases. If we thus at a space flight hold a thermometer out of the shuttle window, then we will measure everywhere such a remaining temperature of a few degrees Kelvin, which is called cosmic background radiation.

By popular scientific small talkers the background radiation is called remnant of a so-called Big Bang. It even is misinterpreted as evidence for the Big Bang. Even if a certain value of entertainment can't be denied, the Big Bang from a physical view until now only has raised questions and contradictions.

If we stay at the fact that oscillating neutrinos depending on their radiation density and velocity of propagation produce a thermal oscillation, which can be detected as a slightly increased lemperature. To measure that we don't need immediately to undertake a space flight. We also can install the thermometer in an artificially produced vacuum. In vacuum physics out of ignorance of the relations the neutrino radiation then is called zero point radiation.

In the case of the neutrinos it of course concerns an oscillation around a mean value, which for a symmetrical form of oscillation has the value zero. The term nevertheless is misleading and chosen very unfortunate. After all we don't speak of zero point current in the case of alternating current!

Since every supernova and every black hole emits neutrinos and correspondingly in physical experiments until now no preferred direction could be determined for the zero point radiation, it is taken as homogeneous and isotropic in space. From the Lorenz invariance again is inferred a cubic course of the zero point spectrum, an increase of the radiation intensity with the third power of the frequency. With this assumption the radiation density of the vacuum however strives with increasing frequency towards infinity. Here an error has to be present in the considerations!

I proceed from the assumption that in the case of the spectrum of the neutrino radiation it rather concerns a spectrum of resonance, the maximum value of which lies above the frequency range measurable with devices of today. It of course would be very important to find out, where the radiation maximum lies, but without being able to measure it that lies unfortunately in the range of pure speculations (fig. 18.2).

Half the wave length in any case should be tuned to the length of the antennas of the energy centres of the cells, the mitochondria and the chloroplasts which are capable of photosynthesis. In both cases the head diameter of the ATPhases amounts to approx. 5 nanometer. It would be obvious, if the resonance point would lie here. But it could as well lie at still smaller wave lengths in the domain of molecular dimensions.

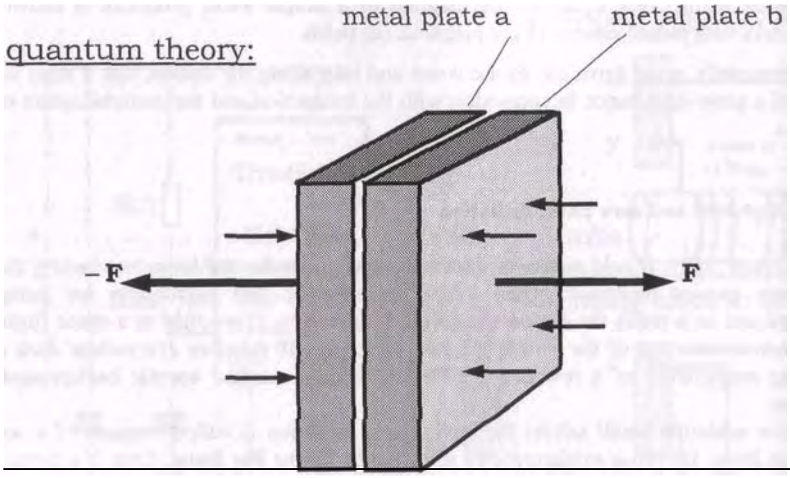
metal plate a
metal plate b
quantum theory:
– F
F

outside cause<i:>: pressure force by bombardment of quanta?

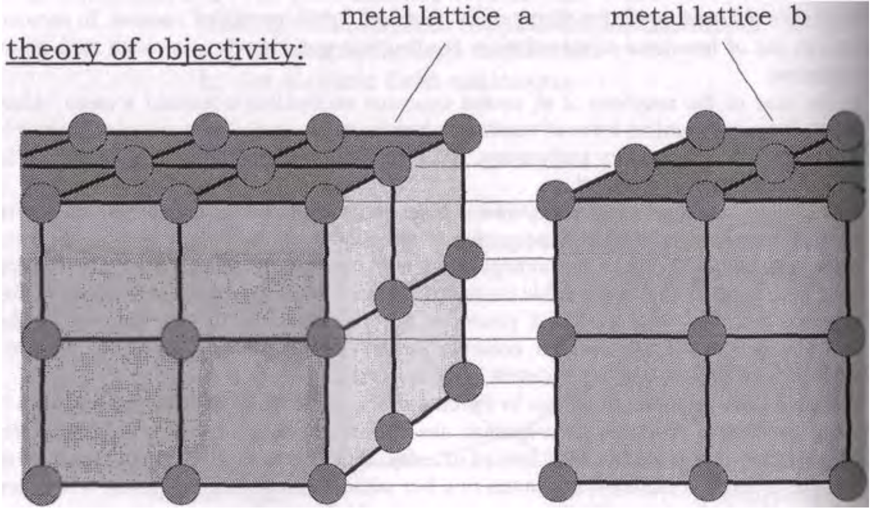
metal lattice a
metal lattice b
theory of objectivity:

inside cause<i>: levitational force by resonance effects?

Fig. 18.3: The Casimir effect

18.3 The Casimir effect

If this zero point radiation actually would be the cause of the Casimir effect, which is generally assumed today, then neutrinos would have to be involved, if the zero point raditation would be to put equal to the neutrino radiation. The effect even is quoted as evidence, but I have my doubts.

The experiment is relatively simple and functions also in a vacuum. Two absolutely plane and smooth polished metal plates are placed very close to each other. Doing so the distance should only amount to one thousandth till one millionth of a millimetre. If the force, with which both plates are attracting mutually, is measured then it exceeds by far the gravitational force. We are dealing with an unknown force, neither electric nor magnetic.

Because the force of attraction still arises near the absolute zero of temperature, some make the zero point radiation responsible. In our translation that would be tantamount to the neutrinos exerting a pressure from the outside on the metal plates, if need be with their small rest mass, as far as this should not be based on an offset-error of the detecting devices. For that the plates would have to damp the neutrino radiation effectively and by mutual shielding reduce the radiation pressure from the split. The question is asked, if the from the outside hitting bombardment of quanta, as it is called, actually can exert a pressure which would be compatible with the nature of the neutrino radiation.

If we now try an entirely other interpretation, which does without the postulating and designing of new force effects. Now the from the unified theory won interactions according to table 15.1 form the basis. In the Casimir effect, as said, neither magnetic nor electric forces are involved, so that open field lines and the corresponding force effects (1 till 4) are ruled out. It has to be the work of closed field lines and that in the static case we know as gravitation (5). Now the mass of the metal plates is too small, as that an acceptable force of attraction of masses could result.

The measurable force is much larger, even if it according to its nature could be a gravitational force. Here oscillations, as they actually take place in the hull of the metal atoms, seem to play a role. If between the oscillations of the two plates standing opposite occurs resonance, then a levitation is possible (6), which describes an oscillating interaction, produced by closed magnetic field lines<i>.

At extremely low-frequency signals this interaction is known as gravitational wave and object extravagant and costly physical experiments.

<i>: The question, which interpretation is correct and which one should be rejected, could be verified as follows: at first the force between two plates of a certain metal and afterwards that of another metal has to be measured. If in a third experiment a plate from the first and one from the second experiment are brought together and the force effect goes back measurably, then the quantum physical interpretation would be wrong clearly. With that would have been shown that here resonance effects are used, which presuppose an identical metal lattice structure of both plates.

If the force effect however doesn't go back, then both statements are possible, then one has to think up another experiment.

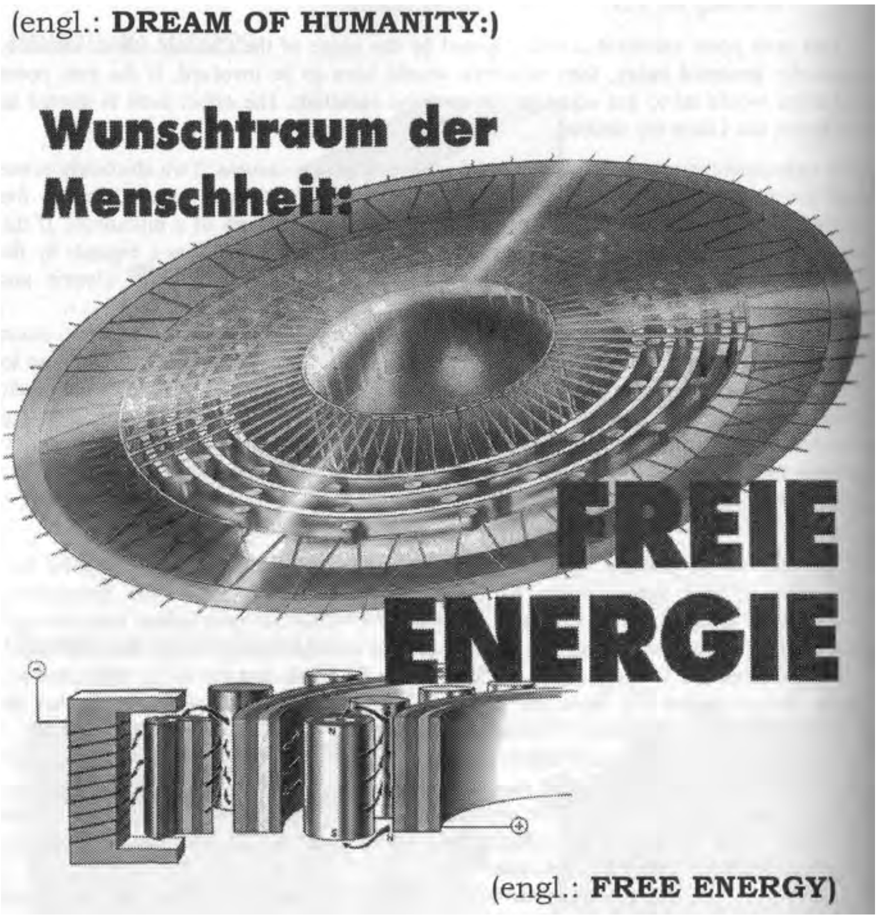
(engl.: DREAM OF HUMANITY:)
Wunschtraum der Menschheit:
FREIE ENERGIE
(engl.: FREE ENERGY)

Fig. 18.4: The flying disc of Searl on a german title page.<i>

<i>: Special 7, Raum 8s Zeit Dokumentation aus dem EHLERS Verlag

18.4 Effects under high tension

It crackles in the laboratory of high-tension. The air smells electrically charged. Only in a corner someone sits and waits with much perseverance and patience, until something happens. Then he lets thunder down lightnings of several millions of Volts on a small sample. It flashes and crashes and still nothing happens.
No, it doesn't concern Tesla! Tesla has worked purposeful. In addition has Tesla only called himself inventor. But the man in the corner feels himself a discoverer and as it is proper for a discoverer, he has immediately named the effect after himself. As a real American he lets market himself and his discovery by the media and a video tape.
Doing so he obviously doesn't know at all, what he wants to have discovered, of what this Hutchison effect actually consists. It perhaps is an event of pure chance, for which he sometimes has to wait hours. If it very suddenly would occur, then it is a real potpourri of all already discussed effects: metal spoons are bending (fig. 14.13), massif steel rods are breaking, light effects are being observed (chapter 17.9 PP), water starts to dance and to cook (chapter 18.1), without getting hot and finally do some samples take off and fly crosswise through the room. He then speaks of anti-gravitation.
In the video tape can be seen, how he brings about a physical length contraction as a result of a locally produced field concentration with two as an interferometer configured Tesla coils. In a bottle for instance the air volume is changed. This confirms the correctness of the here presented theory of objectivity (chapter 6.6 PP).
In any case no-one is amazed more about the results of chance than John Hutchison himself. Overwhelmed by the magic, the charm of the effects and the feeling to be able to move freely outside the trodden out ways of physics, he still sits in his corner with the video camera and waits, until finally something happens again.

There is worked with tension voltages, which are considerably higher than 511 kV, the calculated tension voltage of their own of the charge carriers (fig. 7.1). One thus by no means has to be amazed, if under the influence of several millions of Volt electrons are taken apart and metallic objects are breaking or bending.
The extreme tension voltage and field change we owe an oscillation of length and of size, which spontaneously can lead to an interaction with neutrinos. Materialized neutrinos again are responsible for glowing phenomena and for the electrolysis of water, which under the impression of the rising gas bubbles seems to dance and to cook. For my readers and participants of the energy technical seminar therefore models of explanation are available.

One effect we still haven't discussed and analysed more in detail: the antigravitation resp. the levitation. As long as no reproducible field conditions are present and the chunks are accelerated in any arbitrary direction, it will be difficult to understand the physical course of a levitation.
In chapter 9.3 we already had become acquainted with and discussed a levitation device with the flying disc of John Searl. for which the field conditions are comprehensible. This time we should have a somewhat closer look at the flying device (fig. 18.4)<i>.

<i>: Some reports can be found in Special 7 „Freie Energie", Raum & Zeit Dokumentation aus dem EHLERS Verlag, Seite: 141 bis 157 und 174 bis 185

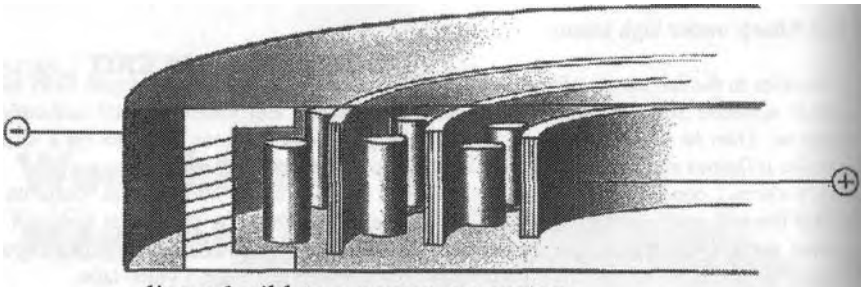

disc rebuild to a power generator

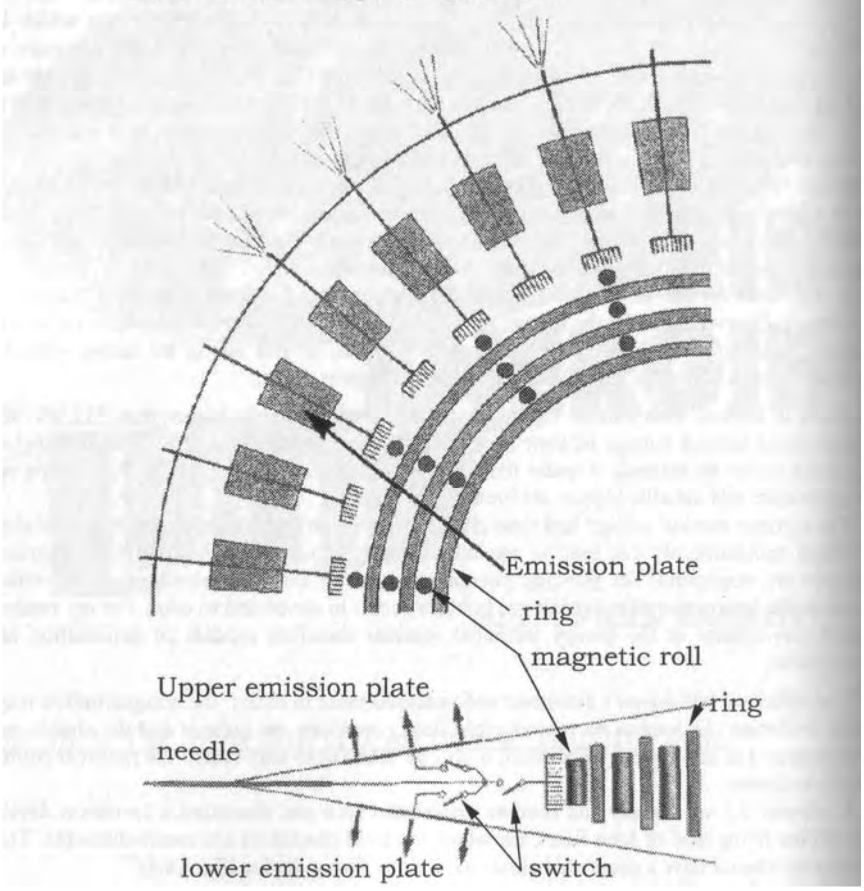
Emission plate
ring
magnetic roll
Upper emission plate
ring
needle
lower emission plate
switch

Fig. 18.5: Concerning the controlling of the flying discs according to Searl.<i>

<i>: H.Schneider, H.Watt: Dem Searl-Effekt auf der Spur, Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, Sauerlach, 1. Aufl. (1994), S. 183

18.5 Flying discs

The Searl-disc can be calculated relatively good with the Faraday equation concerning unipolar induction. To estimate the order of magnitude, we just suppose the roller magnets produce an induction of one Tesla and the radius is one meter, then between the centre of the flying disc and its edge 511 kV is applied, if the revolutions per minute has reached the value of 80000 revolutions per second. As soon as the revolutions per minute is reached, the neutrino conversion can start.
The at the edge of the disc arising corona, which consists of individual electrostatic discharge flashes, takes care for the necessary dU/dt, by every blow and every spark is drawing the potential for a short time towards ground. The tension voltage jumps in swift order between values of above 511 kV and zero Volt to and fro. In connection with the already discussed unipolar field configuration neutrinos are being collected in this way.

A technological challenge represents the layered construction of the roller magnets and the rings, which have to withstand extreme centrifugal forces (fig. 18.5). Friction however is not a theme, since on the one hand the air inside the disc ionises and a vacuum is formed and on the other hand the air split between rollers and discs increases by the field dependent contraction of the metal parts.

Actually one in this case by no means can speak of antigravitation, because the gravitation isn't really vanishing. The disc even in flight still is heavy tons. Here merely a resonant interaction is built up which is larger than the gravitational pull of the earth. The disc is attracted by the cosmic source of neutrinos, with which it has built up the resonance.
Exactly there it will fly!

The by Searl designed controlling[i] in my opinion actually can't have functioned satisfactorily. He uses emission plates, as he calls them, which alternatively are switched on by means of switches and are able to form electrostatic forces with some air charge carriers. In reality the drive probably is comparable to that of a sailing ship, for which the wind always blows from one direction. The ship isn't pushed by the wind, as one erroneously could think, but rather by the under pressure behind the sail pulled forward. Without steering facility the object always is driven in direction of the drain. The sailor would say, the ship without helmsman drifts towards lee.
Searl in this way has lost all flying discs which were started. By using solar neutrinos they presumably have fallen into the sun and burnt. The controlling should function analogous to a sail, then one would stand a chance, by "traversing" against the "wind", to sometime again come back to the starting-point.

<i>: H. Schneider, H. Watt: Dem Searl-Effekt auf der Spur, Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, 1. Aufl., Seite: 183

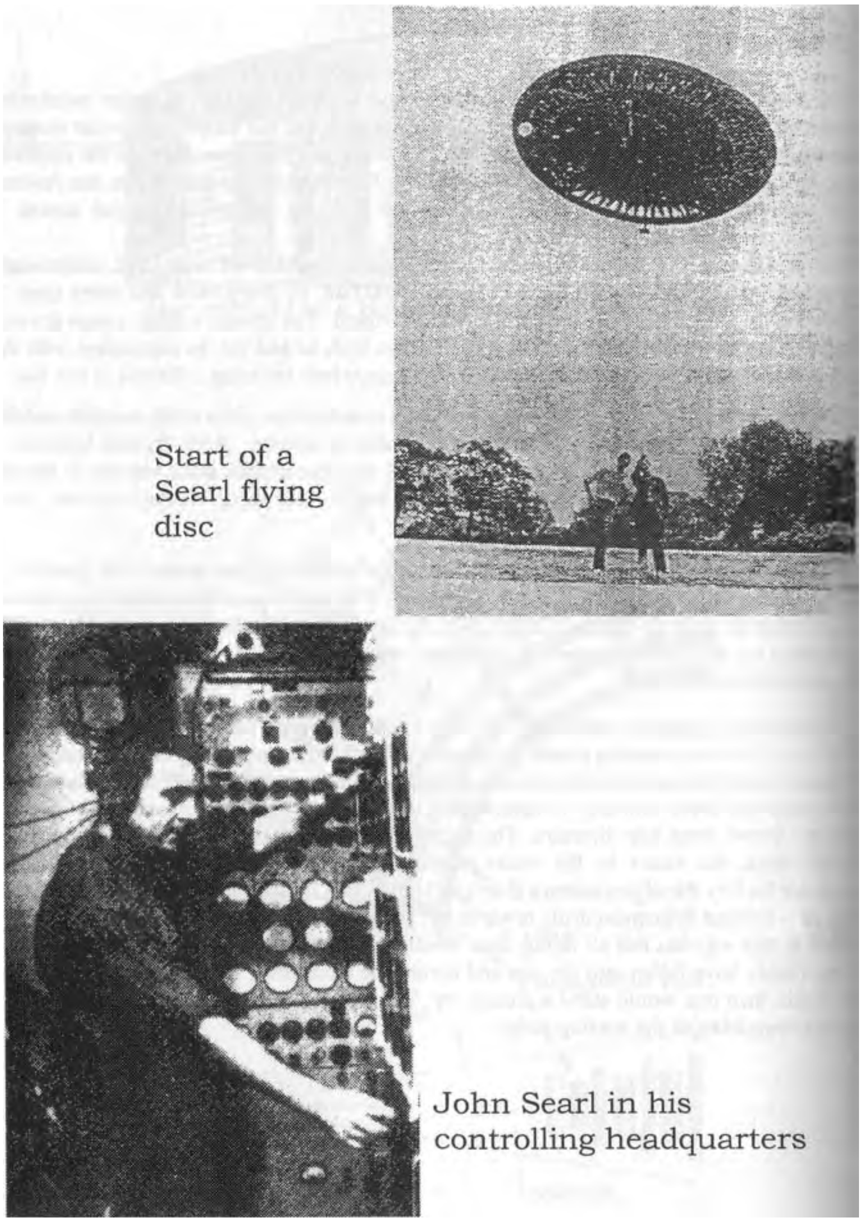
Start of a Searl flying disc
John Searl in his controlling headquarters

Fig. 18.6: The remote controlled flying discs of John Searl.<i>

<i>: Special 7, Raum & Zeit Dokumentation aus dem EHLERS Verlag, S. 152

18.6 Propulsion with neutrinos

In search for an ideal propulsion system for spacecrafts the flying discs of John Searl still have demonstrated a further possibility. After the start occasionally were left strange particles on the ground, which hadn't be there before. The flying device obviously had materialized them and dropped them at the start.

With this materialization of neutrinos the chance is showing of a very efficiently working recoil propulsion. If namely the neutrinos are converted into matter then they by that gain back their rest mass. If this takes place in flight then the materialized particles also bring along kinetic energy. It then works as ajet engine for which the direction of the jet can be directed and in that way the vehicle can be controlled comfortable.

In the case of the Searl-disc the materialization of neutrinos rather happens as a not understood side effect. In the case of a systematic use however the principle will show a characteristic property. Instead of a vapour trail a beam of light will shine out of the flying object in the direction of the emitted particles, but it will break off abruptly after a cartain distance.

To blame is the part, which has become anti-matter, which it is true as well contributes to the recoil, but simultaneously annihilates with incompatible particles of matter under emission of particles of light. If in some distance all antiparticles are used up then also the beam of light comes to an end.

It here concerns necessary properties of a corresponding propulsion technology for spacecrafts, which don't have to drag their propelling energy along with them. Our space technicians finally could handle another than the "hammer throwing method", in which case the "hammer" by means of terrible fuel consumption after a phase of acceleration only staggers uncontrollable through space under the influence of its inertia.

With a neutrino propulsion on the other hand one at any time can accelerate or brake. It will strongly influence the field around itself, so that can be reckoned that for every acceleration a field dependent change of size should be perceptible by an outside observer.

If therefore a corresponding flying device is accelerating then it will suddenly become smaller and that then looks so as if it would have moved away with a jump without temporal delay, but that isn't the case at all.

The jerky movements only would be a result of the perception with our eyes by means of the propagation of the light. Since the passengers are exposed to the same field, they change their body size along with that of the vehicle. They actually notice nothing of an apparently infinite acceleration, which only observers on the outside would observe and which indeed no living being could stand.

Nowhere the explanation crisis is larger than in space travel! The theory of objectivity represents a real help, because it perhaps as the only one puts us in a position to conceive and understand not understood observations as necessary technological consequences. Only by uncovering parascientific as purely physical phenomena man of today will be able to free himself from the constraints of magic and his own illusions.

Is the until now discussed cases resonant or other force effects are being used, whereas the gravitation remains unchanged. Closed field lines after all cannot be influenced, normally at least. Perhaps an indirect possibility exists to have an effect on the gravitation?

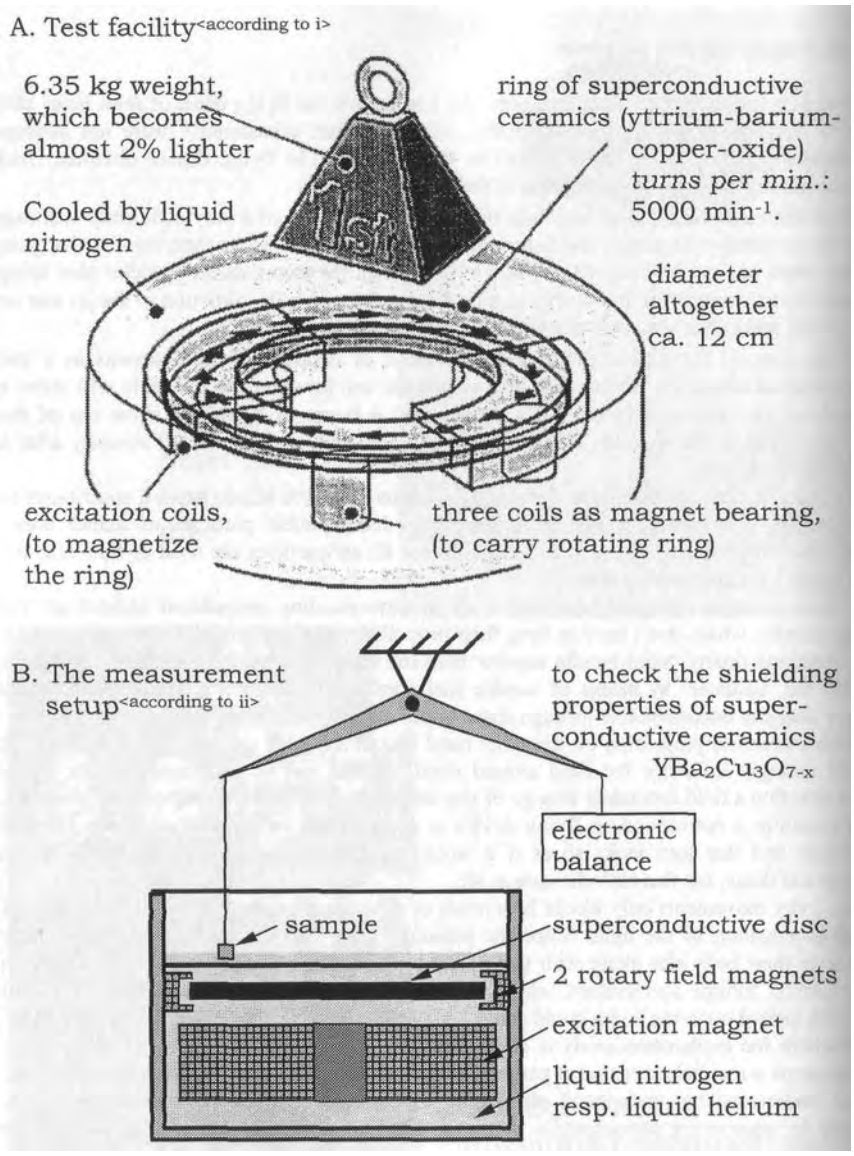
A. Test facility<according to i>
6.35 kg weight, which becomes almost 2% lighter
ring of superconductive ceramics (yttrium-barium-copper-oxide)
turns per min.: 5000 min-1.
Cooled by liquid nitrogen
diameter altogether ca. 12 cm
1st
excitation coils, (to magnetize the ring)
three coils as magnet bearing, (to carry rotating ring)
B. The measurement setup<according to ii>
to check the shielding properties of super-conductive ceramics YBa2Cu3O7-x
electronic balance
sample
superconductive disc
2 rotary field magnets
excitation magnet
liquid nitrogen resp. liquid helium

Fig. 18.7: Experiment for the manipulation of the gravitational pull of the earth.

<i>: Schneider: Anti-Gravitation im Labor nachgewiesen?, NET-Journal Januar 1997, S. 14,15; s.a. Internet: http://www.keelynet.com/gravity/fingrav.htm

<ii>: E. Podkletnov, R. Nieminen: A possibility of gravitational force shielding by bulk $YBa_2Cu_3O_7$-x Superconductor, Physika C 203 (1992), P. 441-444

<iii>: Force field Implications of Anti-Gravity, The Journal of Ideas, Art.191, 7.9.95

18.7 Antigravitation or levitation

The night already had fallen over Finland, but at the University of Tampere the light still was on. Then a scientist put his beardy head through laboratory door and with a "hello folks" blew the smoke of his pipe over the cryostat for the examination of superconductive materials. The scientists were speechless, because the smoke seemed to hit an invisible wall and was drawn upwards in direction ceiling almost at a right angle.
After having aroused the curiosity, Dr. Eugene Podkletnov also has held other materials above his rotating disc and had to discover that these as well lose weight, and that he actually influences the gravitation with his experiment. Even the air pressure at the corresponding place in a lying above floor was smaller<i>.
With a superconductive disc of 30 cm diameter cooled by liquid nitrogen a reduction of gravity for 2% to 4% can be obtained, if the disc rotates with more than 5000 revolutions per minute. With the revolutions per minute also the effect increases, but it is independent from the excitating field<ii>. Finally the fields of the superconductive currents in the ring remain existing unchanged even after switching off the excitation and it only depends on these fields.

The shielding effect of the arrangement on electromagnetic fields already had been known before and should be examined closer in the laboratory. One only was surprised that the gravitation could be shielded as well, that obviously both interactions are related. According to the prediction of the theory of objectivity the closed magnetic field lines gravitate and the field components directed towards the centre of the earth cause the measurable force of weight. By the very strong superconductive fields obviously field overlaps and possibly a local driving out of field occurs, like one has long known for eddy currents (Meissner effect).
With that the earth gravitational field of course is not cancelled, but merely spatially moved, or it partly has changed its direction. If only a small part of field pointers turns out of their orientation towards the centre of the earth, then at that point the force of weight is reduced to a corresponding extent. The theory of objectivity requires that!

By the way reports exist, according to which even without rotation a weight reduction can occur for superconductive rings, and others, where is worked even entirely without superconduction. As core material for the ring ferrite or a strong permanent magnet is used. In that case depends on the circumstance that the ring-like coil is operated in self-resonance and always in the right moment again is excited anew by induction. The ring core then stores the fields up over time, exactly as the superconductor.
Stricktly speaking it are the atomic nuclei and in the end all elementary particles, which are aligning in the high field. Dr. Ning Li wants to artificially produce gravity, by directly influencing the quanta. With that she comes the phenomenon very close<iii>. The possibilities of this effect nevertheless are very limited, since two percent more or less are not exactly much, and the dream of the complete cancelling of gravity or even inertia possibly stays just a dream. It stricktly speaking just concerns influencing the gravitation and not antigravitation!

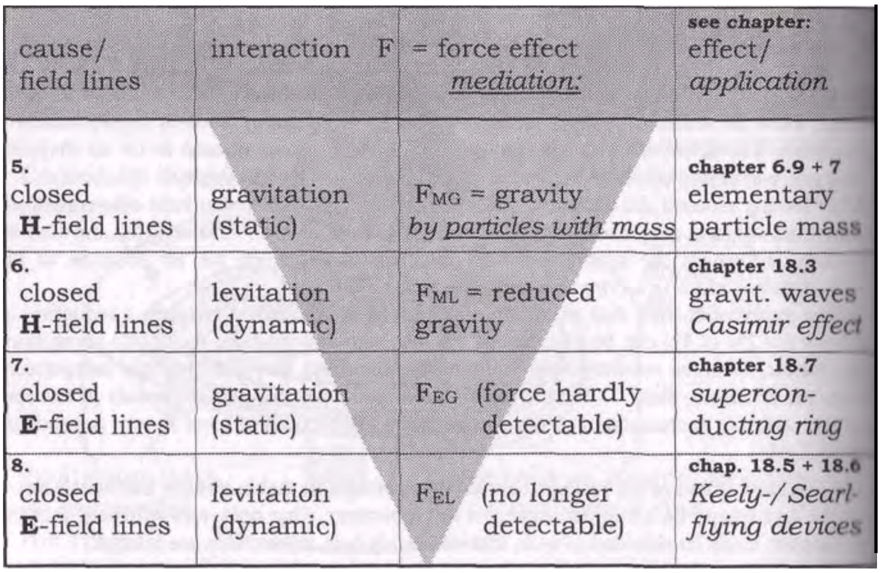

cause/ field lines	interaction F	= force effect *mediation:*	**see chapter:** effect/ *application*
5. closed **H**-field lines	gravitation (static)	F_{MG} = gravity *by particles with mass*	**chapter 6.9 + 7** elementary particle mass
6. closed **H**-field lines	levitation (dynamic)	F_{ML} = reduced gravity	**chapter 18.3** gravit. waves *Casimir effect*
7. closed **E**-field lines	gravitation (static)	F_{EG} (force hardly detectable)	**chapter 18.7** *supercon-* *ducting ring*
8. closed **E**-field lines	levitation (dynamic)	F_{EL} (no longer detectable)	**chap. 18.5 + 18.6** *Keely-/ Searl-* *flying devices*

Table 18.8 A: The interactions of closed field lines.<i>

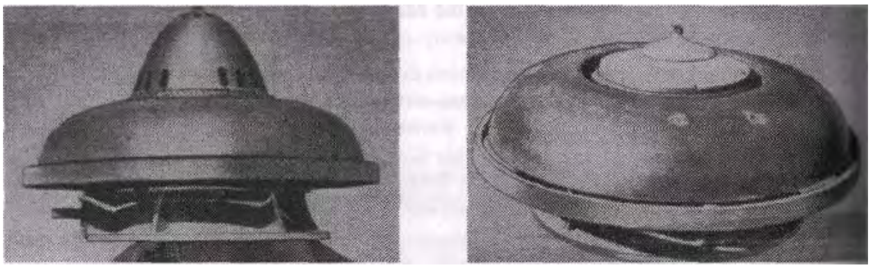

Fig. 18.8 B: Test models of the "flying saucer" according to Viktor Schauberger.<ii>

<i>: according to fig. 15.5, cases 5 to 8

<ii>: O.Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993, S.103

18.8 Discussion concerning the effect of closed field lines

If we now direct our eyes exclusively to the lower half of fig. 15.5, in which the consequences of closed field lines are listed. Without the possibility of a direct focussing like for open field lines, the caused force effects turn out extremely small. With that they are less suitable for an use of free energy. The considerations in that context rather aim at technologies to reduce weight, like they would be worth striving for for flying devices.
If we allege that here in the same way the force effects of magnetic H-fields exceed those of electric E-fields for three to four powers of ten. Then it will hardly play a role for the force of weight of a body with or without additional electric charge, if its E-field lines are closed as well or are open. The increase in weight of an uncharged body only will have an effect in the third till fourth place after the decimal point. Seen so in the case of the gravitation it primarily concerns an effect of the magnetic field, more strictly speaking the effect of closed magnetic field lines (chapter 6.9).
If one succeeds in influencing these magnetic fields for instance by the influence of extremely strong fields of a superconductor, then also the gravitation will change, as has been shown (chapter 18.7). Let's imagine, in the case of a systematic procedure we would succeed in a perfect influencing, for which no component of the H-field lines points into the direction of the centre of the earth anymore, then an uncharged body would only have one tenthousandth of its original weight, whereas a charged body actually would weigh nothing anymore. This state of the weightlessness supposes that all E-field pointers point into the direction of the centre of the earth or diametrically in the opposite direction, since E- and H-field are standing perpendicular to each other. Unfortunately such a field distribution technically hardly can be realized and so the ,,flying carpet" furthermore remains reserved to the fairytales.
An effective reduction of weight of planes and other flying objects however seems by all means feasible, and so slowly the number of research facilities increases, which more or less officialy have a critical look at the cancelling of gravity and the levitation. The theories on which they are basing, however often sound very bizarre and moreover are completely unphysical. Maybe a look at the distribution of field lines, as proposed here, helps to get further.

In the case of the levitation, which occurs strictly speaking only in the case of oscillation for closed field lines, resonance again plays the crucial role for the coming about of the interaction. In connection with neutrinos the resonance can serve less the collecting; we rather need it for the materializing, for the production of mass, charge and energy.
For the artificial production of a levitation either a mechanical oscillation in the atomic or molecular domain is needed, like for instance is produced by a rotating water molecule, or a resonant oscillation of size takes place by use of electrostrictive or magnetostrictive materials, like piezocrystals or oscillating quartzes.
In this context surely is of great importance that also the temperature could be identified as a particle immanent oscillation of size (chapter 8.3). That's why besides a cold materialization or cold fusion also exists, at least theoretically, the possibility of a hot materializatin or hot fusion.

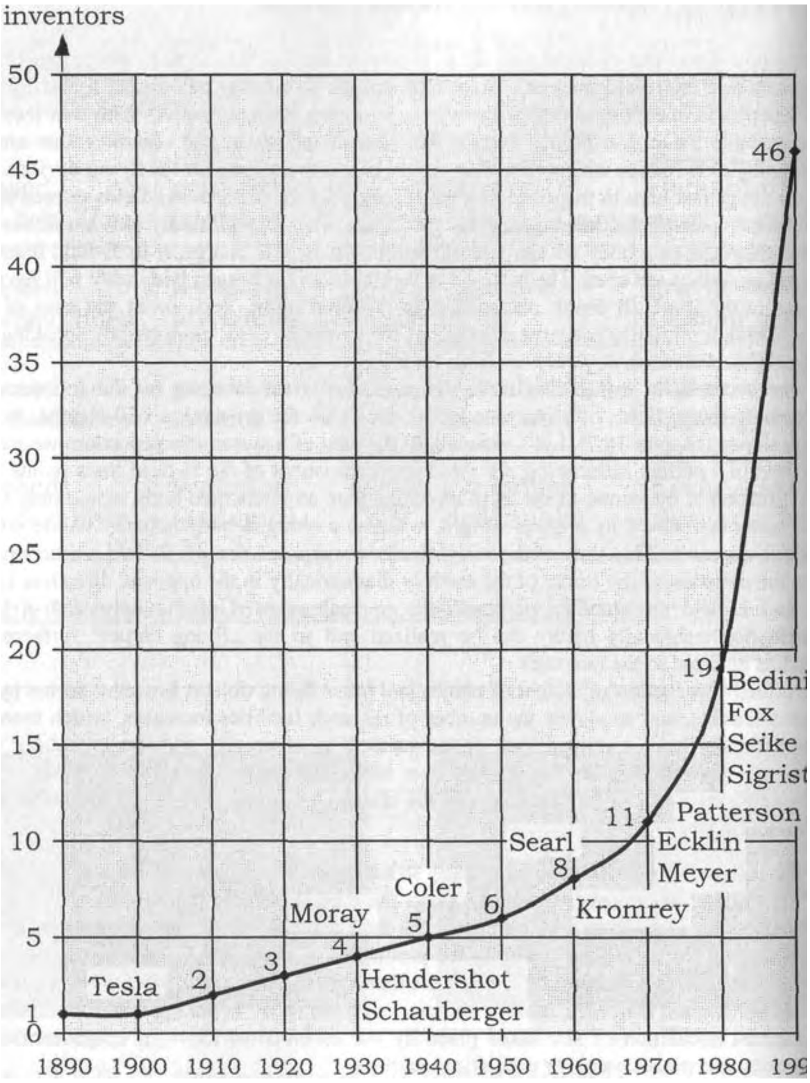

Fig. 18.9: The increasing number of free energy inventors.<i>

<i>: only the inventors mentioned in the text are entered. Literature for that: A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag (1994), S. 77

18.9 Concerning the determination of the stand of today's energy technique

An analogy to the production of matter forms the production of light. At high temperatures as is well-known is formed thermal light and at low temperatures cold light, the discussed luminescence. An oscillation of size of the light source in all cases of the formation of light is the cause, be it as an electromagnetic wave or as photon vortex, be it in whirled water or as sonoluminescence. We at last see ourselves put in a position to understand the various light phenomena as something which is related.

The call for new energy carriers and an ecologically compatible energy technology can't be ignored anymore. The number of inventors at present increases fast (fig. 18.9). It is legitimate and worth recommending, to observe for this purpose first of all nature, how it solves its energy problem. Doing so we should realize that in contrast to the technology of today no combustion and no explosion takes place, but rather principles of a cold fusion and of an implosion as a result of contracting potential vortices are being used. Actually nature may point the way out of the energy technical dead end. We only have to accept the offer, be open and show being ready, to invest in the development of methods in accordance with nature.

Directed at the address of the distributors of supporting funds for energy research the claim in plain English reads: to grant no money anymore for ecologically harmful concepts, like nuclear fission or for not understood and not realizable developments like hot fusion. to immediately stop the fruitless works and to provide the money for an ecologically compatible energy research.

Immense saving potentials moreover are found in all domains of physics, where instead of costly experiments just as well the things could be calculated with an usable theory. To check the theory then only few experiments would be necessary. After all all important discoveries have come about in this manner. Nobody should think, he had understood a matter, once he has filled himself up with sufficiently much measurement data.

The root of the evil lies already in the education, where students of physics want to understand everything and therefore are trying hard to grasp with their view and hands for everything. Abstract thinking or mathematical derivations however are too uncomfortable for most; they have the erroneous opinion, understanding (German: begreifen) would have to do more with grasping (German: greifen), and for that the head after all can't be used.

One reason for this development can be seen in the relativistic point of view of Albert Einstein, who proceeds from the assumption of a subjective observability and has raised the relativity between a physical principle and the observation to the basis of physical thinking. The arrogant motto prevails: what I can't observe with my sense organs or register with corresponding gauges also isn't physics, but esoterics or parascience.

But if we want an ecologically compatible technology, then this can't be reached with this point of view. Then we only hinder ourselves with our own arrogance and intolerance. We have to leave the erroneous quantum physical way and again learn to think abstractly by taking up tried and tested principles of classical physics. An objective point of view forces us to register the phenomena, which lie outside the observable range, with mathematical MEANS (see chapter 13). Only if we have learned that, we will understand and realize the true relations in physics!

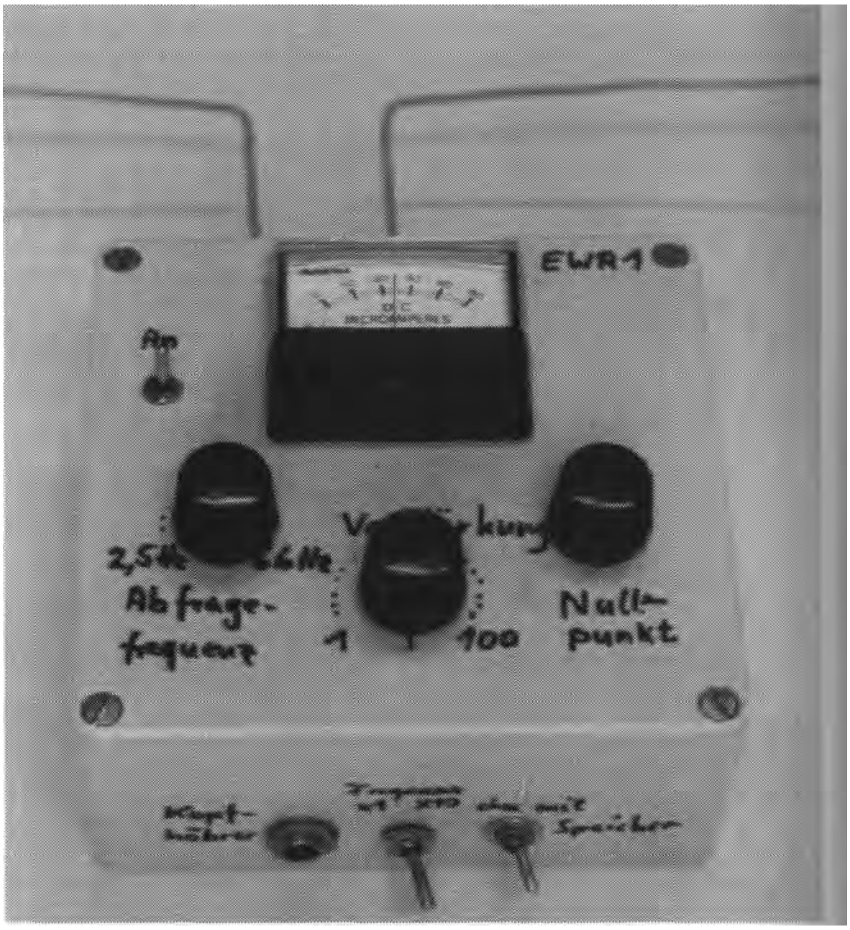
EWR1
DC
MICROAMPERES
An
2,5Hz
56Hz
Abfrage-
frequenz
1
100
Null-
punkt
Kopf-
hörer
ohne mit
Speicher

Fig. 19.1: A self-built charge amplifier.

19. Own experiments

The opposite of the theoretical considerations form practical experiments. If we want to register and understand physical and natural scientific phenomena, then we shouldn't entirely forget the aspect of handicraft. Doing so it can't be avoided that it goes on very personally, if about the laboratory work is being reported, not even then, if the experimenter is at pains to exclude every influencing of the carrying out of the experiment and the result. After all does anyone have to execute the experiments himself, since a good sceptic anyhow only believes, what he has observed with his own eyes.

19.1 Concerning the rating of experimental physics

There already have been scientists, who have requested to reject a measuring technical determined result for the case that textbook physics doesn't produce it. This attitude even today still is taken by an uncounted community of believers of science, entirely after the motto: "Nature kindly has to fit in with the dogmas of theoretical physics!"

This wrong thinking is reflected in the scientific journals, which reject to print discussions. Controlled by a tester board, which only has the task to prevent dissenters from the publication of their ideas, they are the sad proof, how widespread this erroneous attitude today is in the university scene.

Actually solely the scientific experiment shows us the physical reality! The theoretical models, like the here presented vortex model merely should help to understand nature and its laws. The representatives of theoretical physics are in the role of a helper and that requires modesty and openness.

For the case that a professor of theoretical physics imposes a pledge of secrecy on an experimenting colleague, if this one wants to report publicly about his amazing tunnelling experiments, then the public nevertheless should have the right to find out, if a tunnelling experiment reveals speeds faster than light, even if it doesn't suit the theorist and he in inappropriate arrogance only should consider the experimental physicist as his assistant.

If universities only are occupied with preservation of property and the industry only with the increase of its productivity, if future research consists of hiding public research funds internally in such a way that nobody realizes the fraud of support and in reality nobody thinks of the future anymore or wants to work for it, then we will be able to observe how discoveries and inventions migrate from the industry and the universities with their controlled central organs and increasingly will take place again in garage, cellar, solitary study room or in privately organized circles.

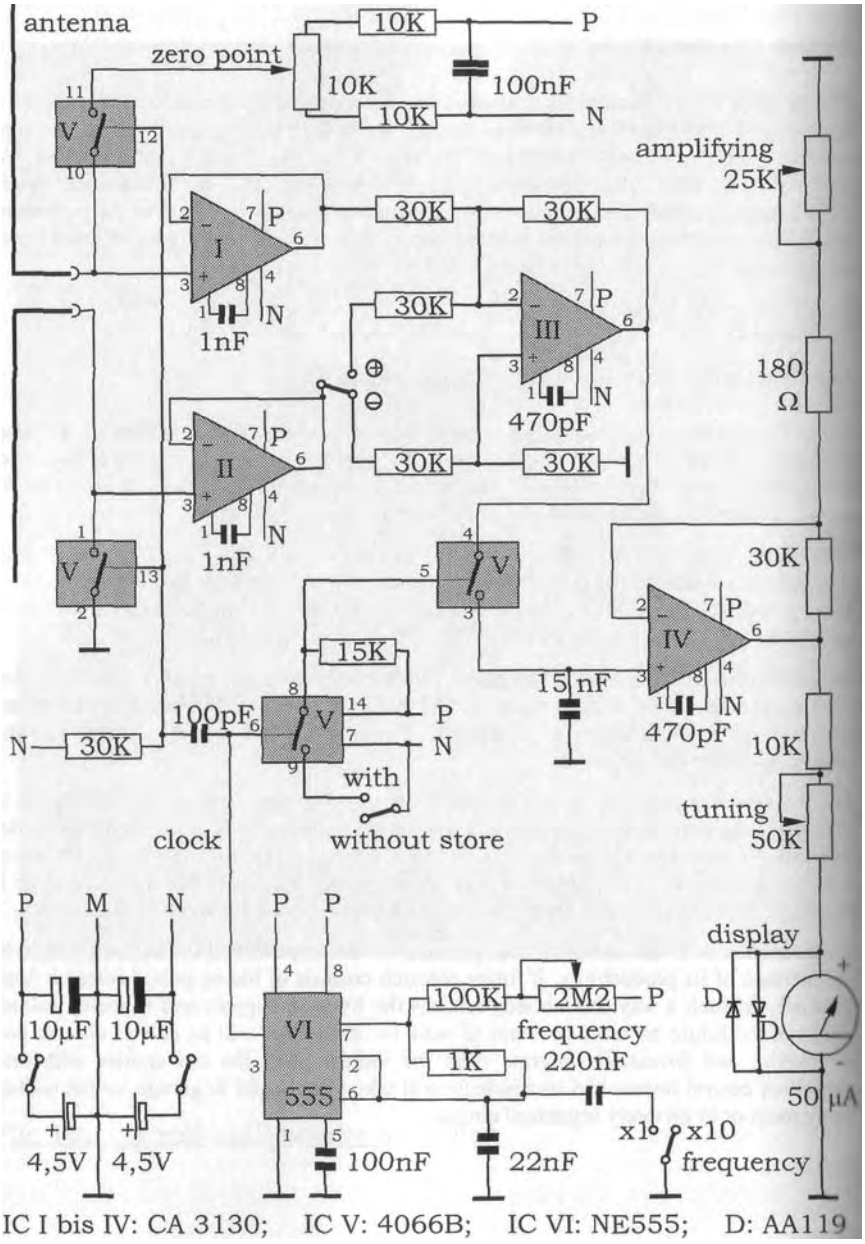
antenna
zero point
10K
10K
10K
100nF
P
N
V
11
12
10
amplifying
25K
I
2
3
7
6
8
4
1
P
N
1nF
30K
30K
30K
30K
30K
III
470pF
180
Ω
⊕
⊖
II
1nF
V
1
13
2
4
5
3
V
30K
IV
15 nF
470pF
10K
tuning
50K
15K
8
14
9
6
7
100pF
N
30K
P
N
with
without store
clock
P
M
N
P
P
4
8
VI
555
7
3
2
6
1
5
100K
2M2
P
frequency
1K
220nF
10µF
10µF
4,5V
4,5V
100nF
22nF
x1
x10
frequency
display
D
D
50 µA
IC I bis IV: CA 3130; IC V: 4066B; IC VI: NE555; D: AA119

Fig. 19.2: Wiring diagram of the self-built charge amplifier

19.2 The charge amplifier from the workshop of handicrafts

Is connection with my vortex theory in this book time and again clues to experiments can be found. I am aware of their importance and I always, to verify my theoretical working out, have accompanied it with practical experiments. Doing so not the proof, which can be published and reproduced by anyone at any time, stood in the foreground. With the experiments in my workshop of handicrafts I only wanted to check for myself, if my derivations still can be verified.

Consequently was at first done without a publication in the first and second edition of the second part concerning the "electromagnetic environmental compatibility". For the addition of the here presented chapter about own experiments from the third edition there are several reasons: first; I have been asked for it by several readers, second; from every experiment is coming a piece of physical reality to meet one and third are here latent some useful ideas and approaches which get us somewhere.
Of course can't be read much from a handicrafts self experiment, but perhaps one or another reader, who has better laboratory technical possibilities at his disposal than me, is stimulated to carry out own experiments.

To chase after the potential vortices in the air, I 1989 at first have built together a charge amplifier. I connected differently formed antennas at the particularly high-ohmic difference input. Corresponding the in nature arising static electric field a tension voltage should arise between both antennas, which my gauge should amplify and indicate on a moving-coil instrument.
To be able to register local changes, a measurement cycle is gone through, which periodically is repeated: It starts with the measurement time, during which entirely by itself between both antenna a charge is building up. The value afterwards is stored analogue in a Sample-And-Hold link and displayed by means of a moving-coil instrument. Then the input clamps are short-circuited, the antennas again discharged and the game starts from the beginning.

Measurement time, zero and amplification can be adjusted at the device (fig. 19.1). Whoever has fun to rebuild it, finds the by me realized wiring diagram in fig. 19.2.

It is true I did realize other designs, but technically the here shown design proved to be the most useful solution.

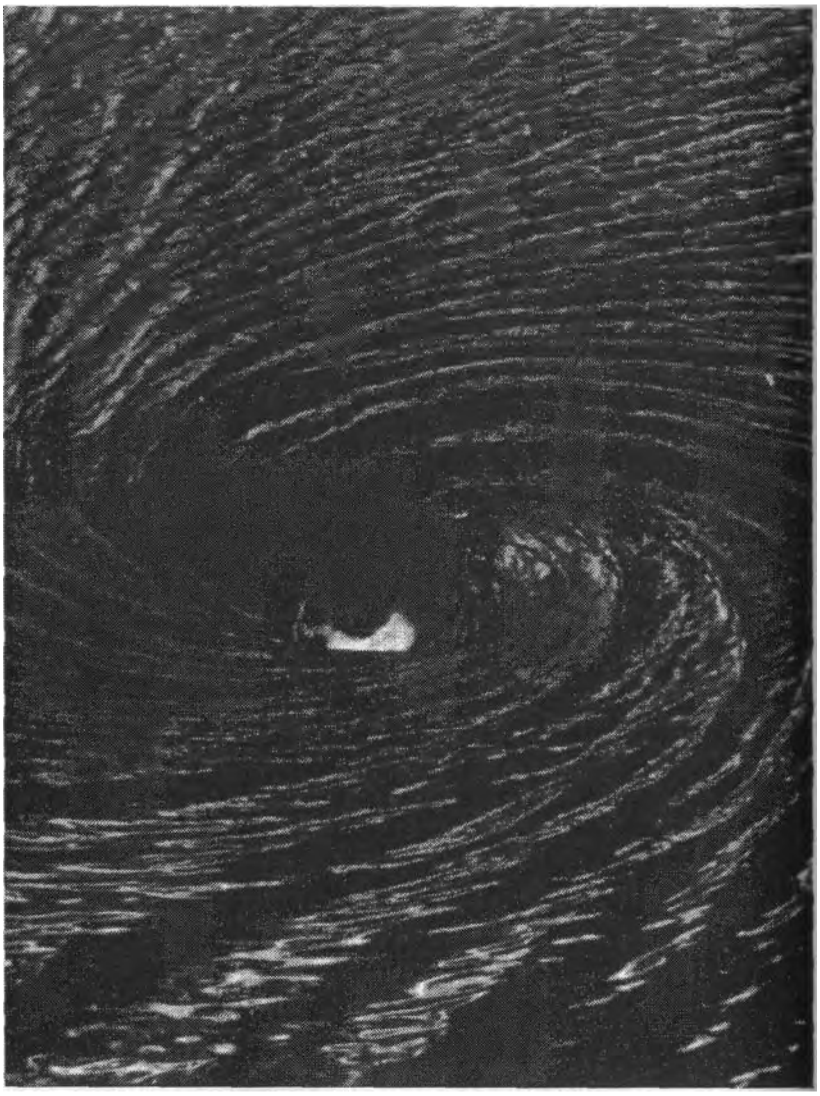

Fig. 19.3: The bath tub drain vortex;
shown is a tidal vortex near St. Malo.<i>

<i>: Photograph from the magazine „Life" of 4.7.1969, resp. from: H. J. Lugt: Wirbelstromung in Natur und Technik, Seite 371

19.3 Interpretation of the observations with the charge amplifier

The practical use of the device at first proved to be extremely difficult, because the pointer seemed to perform a wild dance. At a more close look I however could make exactly those observations, which I had sought-for as proof for the existence of potential vortices.

First I did find confirmed the known high field strength. We after all live between the ionosphere and earth's surface right in the middle of the dielectric of a "spherical capacitor". Because the values of the electric field reach very much closer to the maximum value, which is lethal for living beings, than the value of earth's magnetism, I did draw from that the conclusion that biological effects primarily can be expected from the E-field.

If the E-field, as Maxwell's field theory specifies, actually would be irrotational, then we would be dealing with a gradient field. My device would have to display everywhere to a large extent the same value. But that was not the case.

Maybe the building is responsible for the chaotic display and the wild swings of the pointer, so I thought. Doing so I had thought of the auxiliary explanations of the high-frequency engineer about so-called reflections in closed rooms. Therefore I stormed with the gauge, which had gone wild, into the open air and walked different ways, which should have been reflection free, but the picture stayed the same. In any case the E-field is not a gradient field. I had to find out.

Following I could, what required much patience, find certain places, at which for the same movement from the same direction could be seen a reproducibility of the swing of the pointer, it even could be arbitrarily often repeated. 1 marked the point exactly. Then I tried to move the device from another direction towards the marking and had to find out that the point had moved away.

If I sit in the bath tub in the evening and pull the plug, then I each time am enthusiastic about how sensitive the drain vortex reacts, how I can send it from one corner into the other by the snipping of a finger without it falling apart (fig. 19.3). Doing so one easily can imagine with a bit of phantasy how difficult, yes, almost impossible it would be to measuring technically register the vortex for the case that we could not see it. The gauge it is true would display violent wave movements. But a reproducibility we would not be able to obtain, exactly as for my self-built charge amplifier.

Now I knew that the by me at 2.1.1990 at first purely theoretical derived potential vortex actually exists as vortex of the electric field!

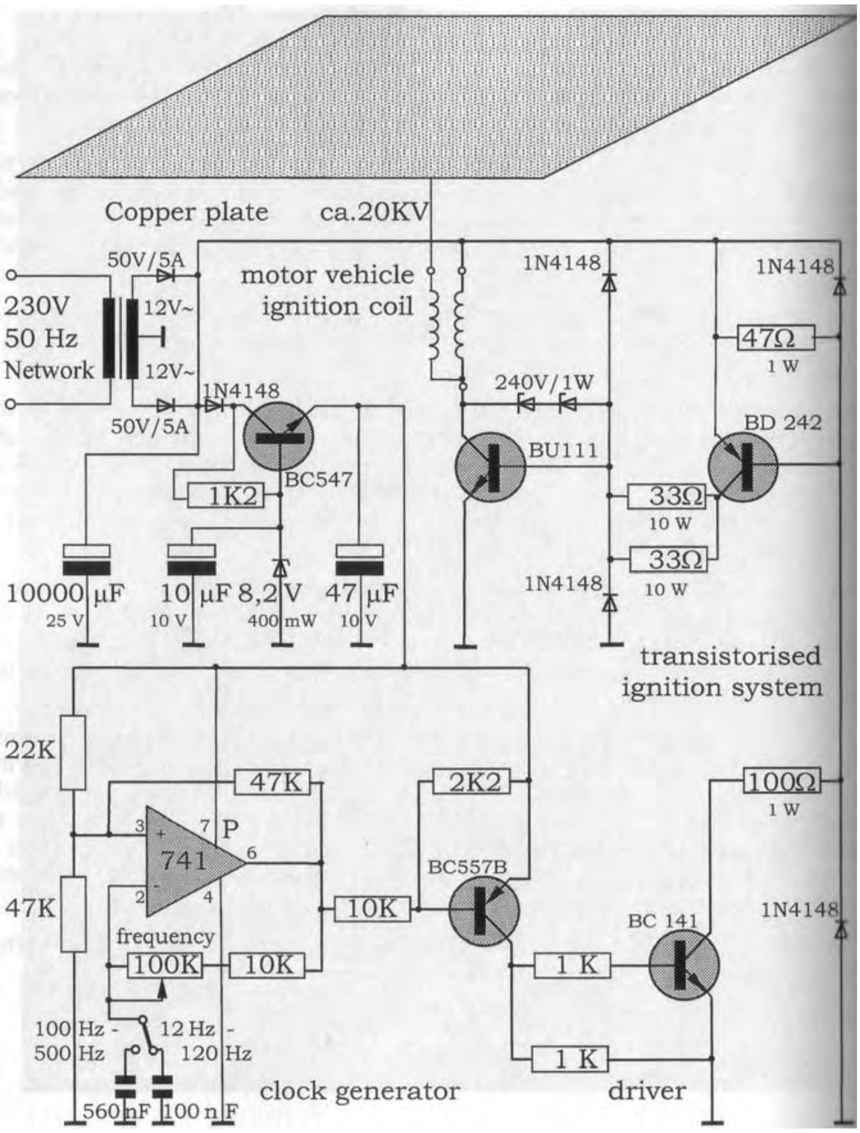
Copper plate ca.20KV
230V
50 Hz
Network
50V/5A
12V~
12V~
50V/5A
1N4148
motor vehicle
ignition coil
1N4148
240V/1W
1N4148
47Ω
1 W
BD 242
BU111
BC547
1K2
33Ω
10 W
33Ω
10 W
10000 μF
25 V
10 μF
10 V
8,2 V
400 mW
47 μF
10 V
1N4148
transistorised
ignition system
22K
47K
47K
2K2
100Ω
1 W
741
P
3
7
6
2
4
BC557B
BC 141
1N4148
10K
frequency
100K
10K
1 K
1 K
100 Hz -
500 Hz
12 Hz -
120 Hz
560 nF
100 n F
clock generator
driver

Fig. 19.4: Wiring diagram concerning the Kirlian device.<i>

<i>: Elektor, Fachzeitschrift fur Elektronik, Mai 1977, S. 22-25: KirlianfotogralV

19.4 High-tension experiments in the laboratory for performance electronics

At first the students of electronics, with which I had to do daily, were infected by my discovery. With true enthusiasm they soldered together one circuit after another. In my laboratory something like a mood of setting out could be felt. One student brought along and old ignition coil of a motor vehicle, then the control circuit for it was built, plexiglas organized, an aluminium plate glued under it and the whole box dragged into the photographical laboratory<i>. Three students made a detour along the outside grounds of the polytechnic and picked the necessary visual aids of trees and bushes.
Following the leafs were laid on photographical paper and by means of the self-built high-tension generator charged to 20000 Volt. Doing so at all edges and particularly at the tips of a leaf corona discharges arise, which expose the photographical paper. We then developed the photographs ourselves and discussed the results <ii>.

The indication, for a second photograph the whole leaf would appear even if half the leaf is torn apart after the first experiment and only one half is put on, occupied us in particular. Eventually we didn't put on a leaf at all in the second experiment and found out that nevertheless the leaf put on last became apparent on the photograph. Now only one physical interpretation was possible: The potential vortices of the leaf stimulated by high-tension still were in the plexiglas disc in weakened form! Here they swirl further and under high-tension furthermore produce corona discharge impulses.

Now also the students were convinced of the existence of the potential vortices.

This experiment stimulated the brain cells of the entire team: If potential vortices under high-tension cause electric blows, since the corona discharge is nothing else, then in this way the local vortex distribution in space should be measurable? One student immediately got down to work and build from the horizontal diversion unit of an old television a high-tension generator with an adjustable spark gap. Following he walked with his flashing and crashing device all through my laboratory and others walked with him and dragged the gauges behind them. But to their big disappointment they were not able to see an influence dependent on place.

Then it suddenly was clear for me: It would have done the students well, if they before would have had a bath and observed the drain vortex, "it suffices the famous wing stroke of a butterfly", I explained my team, "and instead come up with this infernal machine and chase away all the vortices, which we actually want to register ". We had to realize that as a rule phase of disillusionment follows the euphoria.

<i>: Klektor, Fachzeitschrift fur Elektronik, Mai 1977, S. 22-25: Kirlianfotografie; fig. 19.4 shows the wiring diagram and fig. 19.5 A the self-built device.

<ii>: The Kirlian photographs from the darkroom of the polytechnic are shown in fig. 3.6.

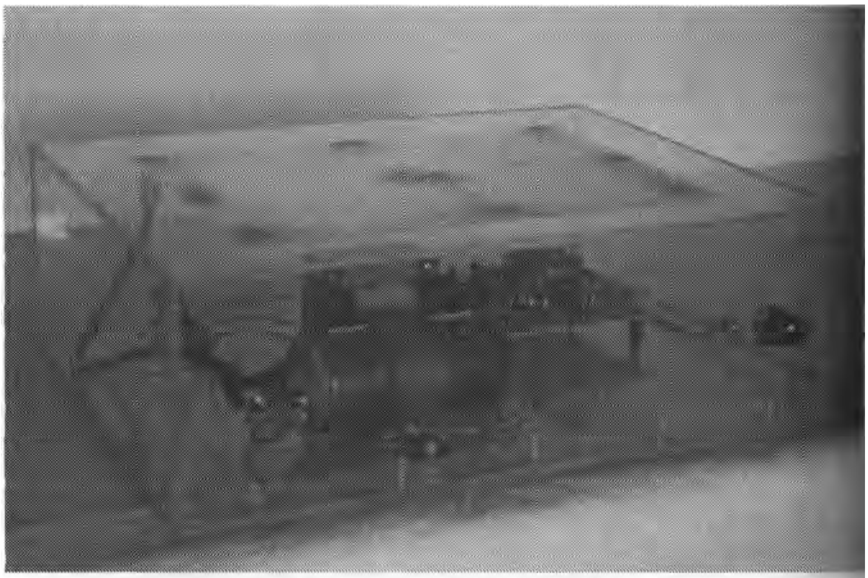

Fig. 19.5 A: A self-built high-tension device for Kirlian photography

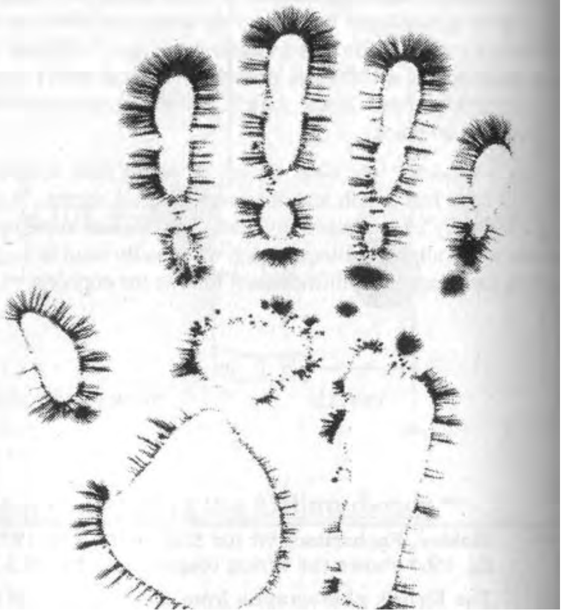

Fig. 19.5 B:
Kirlian photograph
of the right hand

19.5 Measurement of dielectric capacitor losses in the HF laboratory

My convincing work clearly was shaped more complicated with my professor colleagues. Most pretended to understand nothing about it and kept themselves out. But there also are exceptions. „something like that has never been there at our polytechnic, that a colleague puts forward an own field theory ",a physicist colleague remarked after my lecture public to the polytechnic<i> at 3.7.1990 and meant, that must be celebrated. He festively let me cast my eyes in his store room, got out a few bottles of wine, which then some of my colleagues of physics emptied with me in their official room. Doing so we small talked about the sense or nonsense to fix on the education of physics to the Coulomb charges.

The head of the high-frequency laboratory showed likewise impressed. At last he now knew, why capacitors at microwave frequencies can become so hot that they solder themselves out of the circuit by themselves and can fall out, why PVC-films can be welded with HF, etc. „We have to prove that not the dielectric is to blame, as stands in the textbooks" he came towards me.
At least following my idea potential vortices are the ones, as I expressed myself, which are behaving dual to the eddy currents. It concerns vortex losses, thus a physical phenomenon. Eddy currents now can be damped as is well-known, by for instance sheeting the iron circuit for engines and transformers. The insulation between the sheets prevents the formation of eddy currents in that direction and the degree of effectiveness increases.
"I would suggest", I told my colleagues, „to "sheet" a capacitor in a dual manner and to measure the losses". Because eddy current losses increase with the square of the frequancy, we picked microwave frequencies. In the HF-laboratory a card with an L-C resonant circuit was made, we should be able to quite precisely determine the losses by means of its quality. The inductance was formed as a microstrip line and for the capacitor a socket was planned.
The carrying out of the experiment consisted of using single layered ceramic capacitors, so called trapezium capacitors, with a vapour deposited silver coating on both sides as a package and to measure the quality of the resonant circuit as a measure for the dielectric losses. According to the rules of duality the dielectric (as nonconductor) thus corresponds to the permeable transformer sheet (conductor) and the silver coating (conductor) to the sheet insulation (nonconductor). In a second experiment the silver coating now was removed and the same capacitor material measured at identical conditions this time uncoated. Will the losses increase or isn't changing anything?
Tension and nervousness suddenly could be felt, even among the students, who followed the experiment from the second row. All looks were pointed at the monitor of the network analyser, on which slowly the bell-shaped curve of the measured resonant circuit quality became apparent. The result was surprisingly clear. I first looked at my colleagues and then at the students and had to realize that all eyes were pointed at me. The first measurement was a bull's eye! The vortex losses "uncoated" were considerably larger.

<i>: K. Meyl: Die mangelnde Dualitat der Maxwellschen Gleichungen with contributions concerning the theme of magnetic space poles, the mathematical calculation and the technical-physical interpretation of newly discovered potential vortices. Invited by the rector of the FH Furtwangen at 3.7.1990

- experiment 1: quality of resonant circuit with coated capacitors

from a series of 5 separate experiments averaged: Q = 48

- experiment 2: silver coating removed with diluted nitric acid

possibly the acid has damaged the dielectric: Q = 9

- experiment 3: silver coating polished off mechanically.

low quality means high dielectric losses! averaged: Q = 28

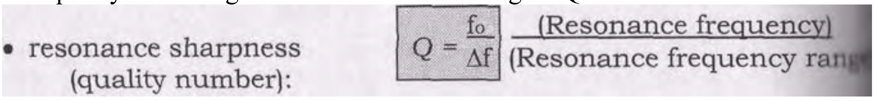

- resonance sharpness (quality number): $\boxed{Q = \frac{f_0}{\Delta f}}$ $\frac{\text{(Resonance frequency)}}{\text{(Resonance frequency ran}}$

at a resonance frequency f0 of approx. 400 MHz.

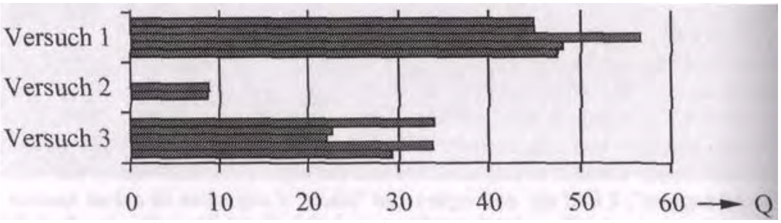

Quality number Q of the L-C resonant circuit

the dual ,,sheeted" capacitor has the lowest losses,
the resonant circuit with that the highest quality (experiment 1).

Fig. 19.6: Measurement of the dielectric capacitor losses over the quality of a resonant circuit in the HF laboratory at 11.06.1990.

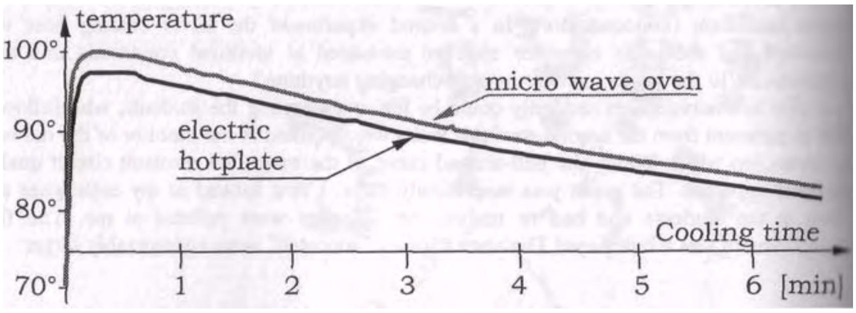

Fig. 19.7: Measurement curve of the water temperature over time
at 22.10.1996 in the electrical engineering training

Observation: for the microwave oven the container despite the "after cooking-effect"
remains colder than in the case of a hotplate. Sensor: PT 100.

19.6 Analysis of the measured capacitor losses

First of all the colleague spontaneously prompted a common publication.

But a certain disillusionment had to await us. For the next measurement hardly a differemce could be detected. We varied the contacting and removed the silver coating. once with hydrochloric acid, once with fine sandpaper. Every measurement we repeated umpteen times and in the end had to draw the conclusion that the results for our somewhat unprecise construction fluctuated extremely.

Averaged over all measurements and methods, indeed a reduction of the dielectric losses for alayered capacitor resulted, but at a critical consideration of the errors the results seemed to "drown" in the mean variation. In the opinion of my colleague the visible trend wasn't sufficient to convince dyed-in-the-wool sceptics, whereupon he withdrew with the indication, I should try at the people of microelectronics, they more likely would be capable to reproducibly gain control of a layered construction.

The experiment at least let a legitimate chance open that the vortices of the electric field actually exist, and in the case of the delectric losses of a capacitor it concerns vortex losses<i>.

19.7 Microwave oven for testing in the laboratory of electrical engineering

"Volunteers first" was said at 22.10.1996 in the electrical engineering practical training and two students got down to work. I had brought along from our kitchen the microwave oven, a portable immersion heater, a kitchen stove plate and different containers. In the laboratory of the polytechnic I in addition had got hold of a Bunsen burner. In the sense of the "Stiftung Warentest" (institution to protect consumers in Germany, note of the translator) it concerned the question for differences in heating water; or following the public discussion, is cooking with gas more healthy than cooking with the microwave oven?

We wanted to know. So we cooked water, filtered by inversion osmosis, once with the gas burner. once with the hotplate and finally with the microwave oven. Switched off always was at the same moment at the same temperature, and the cooling down curve was recorded with an x-t-time recorder (fig. 19.7).

Between gas and electric cooker virtually no difference could be detected, but the microwave oven at switching off still showed an odd temperature increase. I already had remarked this earlier. If one takes a water glass out of the microwave oven, then the water again really bubbled off, although the container itself had remained relatively cold. If one on the other hand takes a water pot off the gas cooker, the water suddenly stops cooking. From where comes this difference, which also the experiment could confirm?

<i>: A result of my dual way of looking at the vortex losses is the representation in fig. 4.7.

A: damped wave equation:

$$\begin{array}{ccccc} \Delta \mathbf{E}\, c^2 & = & \delta^2\mathbf{E}/\delta t^2 & + & (1/\tau)\cdot\delta\,\mathbf{E}/\delta t \\ & & \text{(wave)} & + & \text{(vortex)} \end{array}$$

B: open chain of reasoning in the physics of electromagnetism:

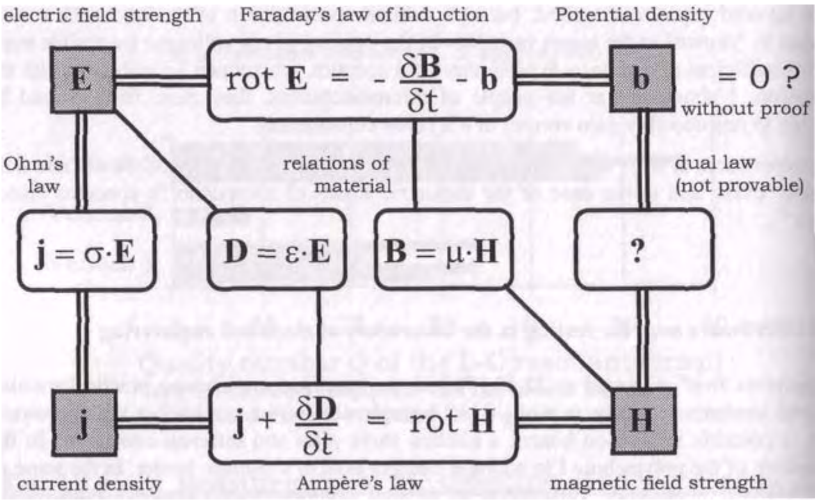
electric field strength
Faraday's law of induction
Potential density
E
rot E = - δB/δt - b
b
= 0 ?
without proof
Ohm's law
relations of material
dual law (not provable)
j = σ·E
D = ε·E
B = μ·H
?
j
j + δD/δt = rot H
H
current density
Ampère's law
magnetic field strength

C: for the stationary case $(\delta/\delta t = 0)$ in general representation:

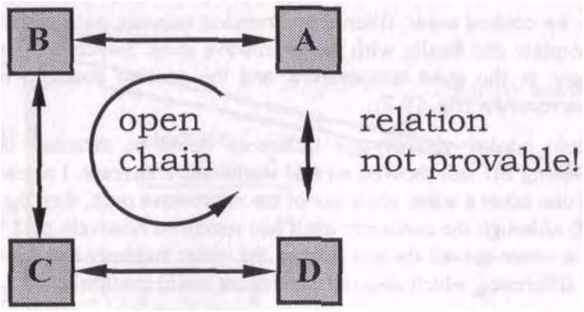
B
A
open
chain
relation
not provable!
C
D

Fig. 19.8: ____ Contribution to the discussion about the impossibility, to prove, according to scientific methods, in a chain of reasoning (A-B-C-D) the last link (D-A), which closes the chain.

19.8 Microwave heating by decay of vortices

It has to concern a storing effect. In the case of an normal cooker the heat transfer takes place by convection. In the case of a microwave oven we however are dealing with an electromagnetic wave.

But in the wave equation nothing can be read about heat; here merely vortices can be found as a damping term (fig. 19.8 A)<i>! Therefore as the only possible interpretation is left that the microwaves roll up to vortices to fall apart themselves after a certain time.

Only at the vortex decay heat is formed.

The irradiated microwave power according to that is stored over a longer period of time as vortex in the water and the food. The vortex decay takes place according to an e-function with the calculated relaxation time constant tau<ii>.

From the circumstance that particularly water is heated up in the microwave oven despite the small electric conductivity but with high dielectricity, I infer that it has to concern mainly potential vortices, from which a biological effectiveness can be expected. The question if this should be valued positively or negatively, I have to leave up to the doctors and therapists. At least physically seen a difference is measurable. Thus a vortex decay will occur for a meal prepared in the microwave oven even if we already have consumed it.

Conclusion: I wish all: your health!

But what do I write about vortex losses, if every sceptic knows half a dozen alternative interpretations. Inevitably I have to bear in mind that also the eddy current heating owes its acknowledgement only the circumstance of the discovery of the corresponding laws. If Othm hadn't discovered his law and formulated it in the known form and instead the dual formulation would have been discovered, then we today would attribute the dielectric losses in a capacitor and in the microwave oven to the potential vortices, in the case of the transformer however the material would be responsible for the heating and not eddy currents, for which there then also wouldn't be a theory (fig. 19.8 B).

It actually is pure coincidence that at first Ohm's law and not the dual formulation had been discovered and acknowledged. But because both are equivalent, we also have to assume the correctness of both, even if the last link of a chain of reasoning A-B-C-D back from D to A in principle can't be proven anymore, since it already is explained by the chain A-B-C-D (fig. 19.8 C).

There I thus had discovered a potential vortex, without a possibility for a direct proof of existence. For that the textbooks are full with auxiliary explanations, with which physics successfully cheats past this important field phenomenon. I was frustrated. There had to be a way to measuring technically register the vortex in some way.

<i>: Fundamental field equation 15 in fig. 5.1 or equation 21, fig. 5.3

<ii>: Relaxation time constant: see equation 10 in fig. 5.1 or fig. 8.1

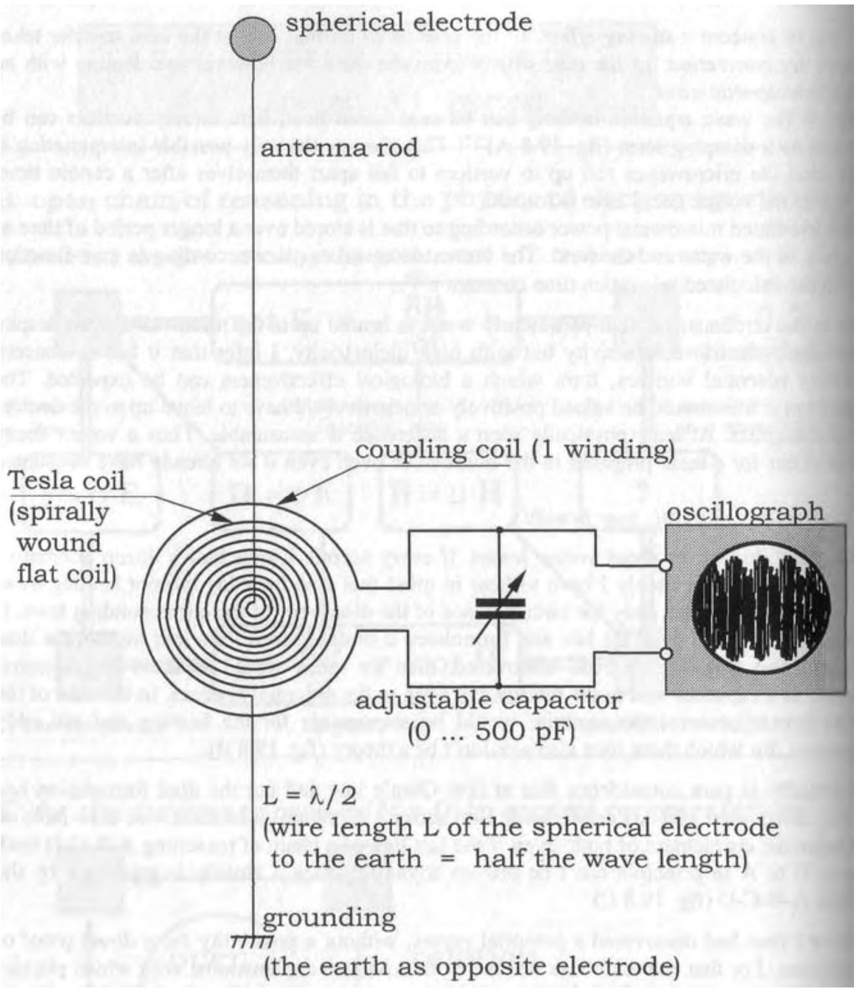
spherical electrode
antenna rod
coupling coil (1 winding)
Tesla coil
(spirally
wound
flat coil)
oscillograph
adjustable capacitor
(0 ... 500 pF)
L = λ/2
(wire length L of the spherical electrode
to the earth = half the wave length)
grounding
(the earth as opposite electrode)

Fig. 19.9: Proof of scalar waves with the Tesla coil.[i]

---

[i]: The inverse use for a high-tension transmitter according to fig. 9.1

19.9 The Tesla coil from the workshop of handicrafts

I again retired in my workshop of handicrafts and wound flat coils or I pondered in my study over Tesla books. 1995 from several sides my attention had been drawn to the circumstance that the inventions of Nikola Tesla can be organized in three categories:

The first third has made him world-famous. It concerns the rotary field theory, the asynchronous engine and the today normally used alternating current technology, which we owe him.

The second third concerns technologies and inventions, which were rediscovered by other people partly only years later or even were only pinched and distributed as novelty under another name. Electron microscope, superconduction, electrolyte capacitor, fluorescent lamp, fuse. coaxial cable and a lot more count among that.

The last third however concerns inventions, which until the day of today still aren't understood and await their scientific explanation and technological use. Tesla himself called these achievements his most important inventions, but still owes us a scientific explanation. The scientific world also hasn't got a theory ready and doesn't know to do anything with it. Public research doesn't take place or is prevented by lobbyists.

What remains, are tinkerers of various educational background and qualification, who are trying hurd to comprehend the buried experiments of Nikola Tesla off their own bat in the garage or in the hobby cellar.

I felt like one of these, by winding one winding after the next from the inside to the outside. Then I soldered an antenna wire at the inner end of the flat coil, which I in Tesla manner connected electrically with a spherical electrode hung up under the ceiling. The opposite electrode should be connected to the outer end of the winding, it is said, and the distance between both should be as big as possible. If namely an electrode just is collecting, then the opposite electrode is repelling the same space quanta. According to Tesla's recommendation I did use the earth as opposite electrode and for that tapped the central heating or the grounding of the foundations.

To take signals only one to two windings as secondary winding were necessary. I connected them with an adjustable air capacitor from an old steam radio to a frequency determining parallel resonant circuit and looked at the taken tension voltage at the oscillograph (fig. 19.9),

I still had problems with the statement of Tesla, the coupling had to be made loose. Thus the question is asked, how loose? I after that organized two toilet paper rolls of different size, (after the toilet paper had been used, naturally) and pushed them into each other. The smaller toilet paper roll carried on the gable-end the flat coil as primary winding of the air transformer and the bigger one the coupling coil. Now by shifting any wanted degree of coupling could be adjusted to (fig. 19.10).

I was astonished myself. Tesla actually was right with his discovery of the scalar waves. With my arrangement they can be clearly distinguished from the Hertzian waves.

The following procedure is recommended:

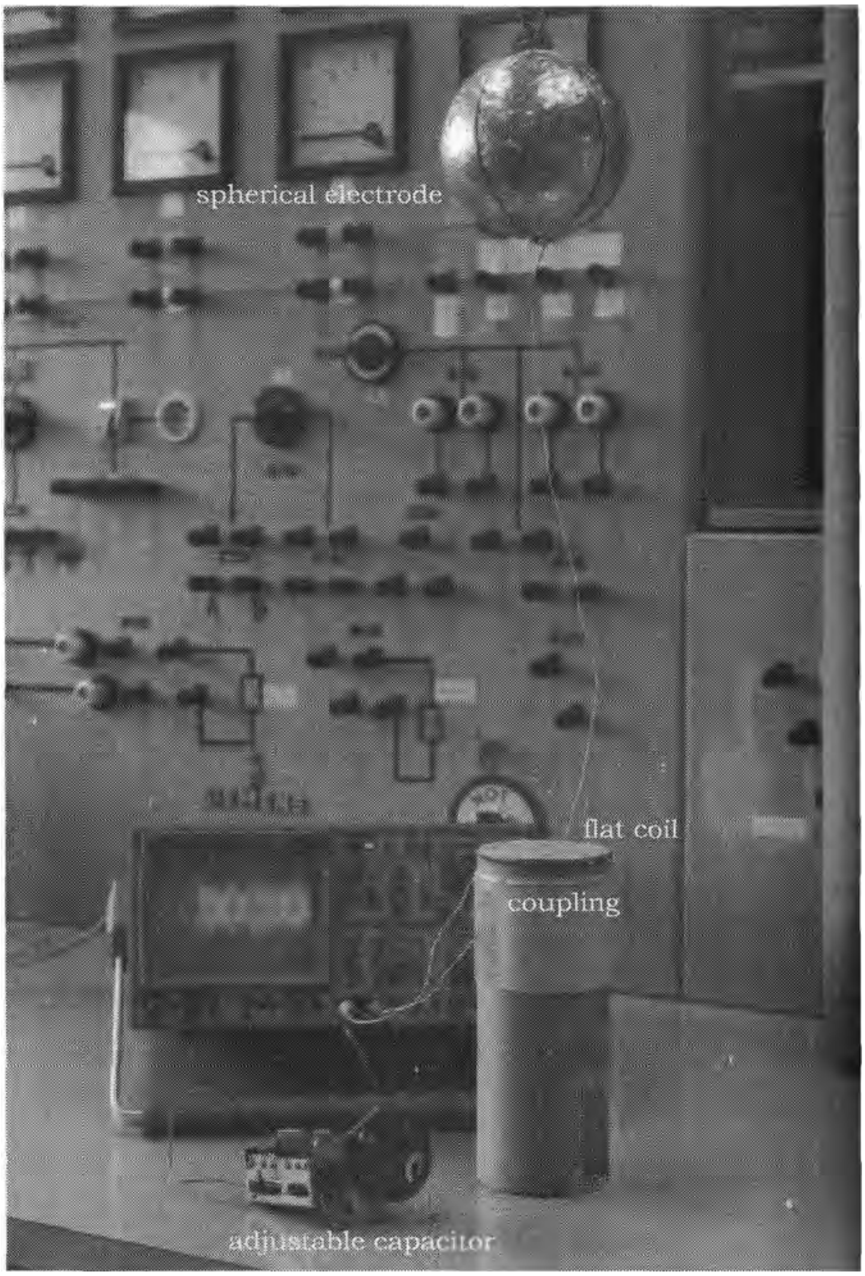
spherical electrode
flat coil
coupling
adjustable capacitor

Fig. 19.10: The experimental configuration

First of all I seek a source of interference with the adjustable capacitor and tune to maximum amplitude. Then I change the coupling and further optimise in this way. If now the amplitude again decreases from a certain point while approaching the coupling coil, then it concerns the sought-for longitudinal waves. If namely the coupling is too tight, then the received vortices again are driven away by the effect back on the fiat coil. They make way.
At last I had found a method to catch the vortices in such a way that they not immediately "ran away" from me again. At once I presented them in the technology centre in St. Georgen. In the time following I improved the technology further and further, used bigger toilet paper rolls and eventually even turn up garbage cans, I varied wire length, wire diameter and the sense of winding (fig. 19.10).

I had very different success. Sometimes, if at the same time in my workshop of handicrafts the radio worked, it would look as if the received signal would synchronize with the sound waves. Both are longitudinal waves after all. With transverse waves something like that would be unthinkable.
One moreover could observe, how a resonance builds up: first slowly and then faster and faster, so that I sometimes got terribly afraid. Several times we had to repair our oscillograph, after the protective diode at the input amplifier had blown, and that for signals of 50 to 100 millivolt!
That was entirely impossible. Only individual spikes, which were too fast to be seen at the screen, could be to blame. In the case of distant thunderstorm activity I obtained maximum values of more than half a Volt. After that I undamped the grounding as fast as possible, so that no lightning would be caught, since I didn't have the intention to burn off my workshop.

19.10 Biological effectiveness of the Tesla coil

Also the biogical effectiveness of the Tesla radiation I could impressively prove with this device. 14.06.1997 a woman, who called herself extremely electrosensitive, participated in the weekend seminar about electrobiology, which I took over from Prof. H.K. Konig (TU Munich) after his death. That I wanted to test.
I hung up my device in the lecture auditorium and installed the oscillograph in such a way that all could see it. The voluntary test subject however could see neither the public nor the screen. One person every 5 seconds said "now" and the female candidate should say, if I had clamped or undamped the grounding, if therefore scalar waves were received or not. After a training round her proportion of hits was lying at 100 percent!

According to her statement she could feel it. A further test subject achieved even at the back of the room with a pendulum the correct answer. I myself was surprised by this and I already have repeated this experiment several times with different success. It without doubt depends on the sensitiveness of the test subjects. Every person after all reacts to other signals.

What however has astonished all participants and can't be emphasized enough at all, is the circumstance that it in this case concerns a receiver and not a transmitter!

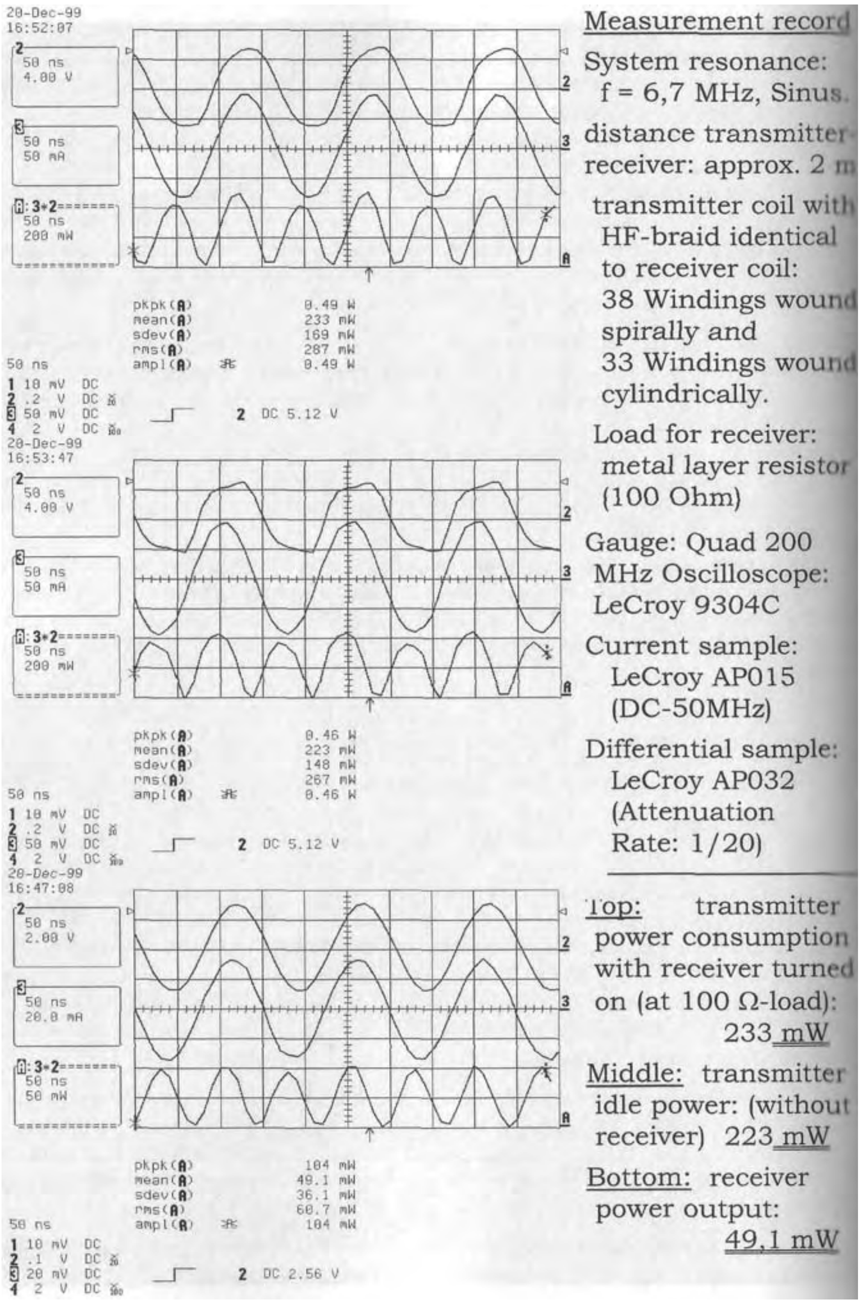

```
20-Dec-99
16:52:07
2
 50 ns
 4.00 V
3
 50 ns
 50 mA
A:3*2========
 50 ns
 200 mW

         pkpk(A)              0.49 W
         mean(A)              233 mW
         sdev(A)              169 mW
         rms(A)               287 mW
50 ns    ampl(A)              0.49 W
1 10 mV DC
2 .2  V DC
3 50 mV DC            2  DC 5.12 V
4  2  V DC
20-Dec-99
16:53:47
2
 50 ns
 4.00 V
3
 50 ns
 50 mA
A:3*2========
 50 ns
 200 mW

         pkpk(A)              0.46 W
         mean(A)              223 mW
         sdev(A)              148 mW
         rms(A)               267 mW
50 ns    ampl(A)              0.46 W
1 10 mV DC
2 .2  V DC
3 50 mV DC            2  DC 5.12 V
4  2  V DC
20-Dec-99
16:47:08
2
 50 ns
 2.00 V
3
 50 ns
 20.0 mA
A:3*2========
 50 ns
 50 mW

         pkpk(A)              104 mW
         mean(A)              49.1 mW
         sdev(A)              36.1 mW
         rms(A)               60.7 mW
50 ns    ampl(A)              104 mW
1 10 mV DC
2 .1  V DC
3 20 mV DC            2  DC 2.56 V
4  2  V DC
```

Measurement record

System resonance:
f = 6,7 MHz, Sinus.

distance transmitter-
receiver: approx. 2 m

transmitter coil with
HF-braid identical
to receiver coil:
38 Windings wound
spirally and
33 Windings wound
cylindrically.

Load for receiver:
metal layer resistor
(100 Ohm)

Gauge: Quad 200
MHz Oscilloscope:
LeCroy 9304C

Current sample:
LeCroy AP015
(DC-50MHz)

Differential sample:
LeCroy AP032
(Attenuation
Rate: 1/20)

Top: transmitter
power consumption
with receiver turned
on (at 100 Ω-load):
233 mW

Middle: transmitter
idle power: (without
receiver) 223 mW

Bottom: receiver
power output:
49,1 mW

Fig. 19.11 A: Measurement record for scalar wave transmission line

19.11 Wrestling for measuring technical insight in the Transfer centre

In the central question who should be believed more, the famous experimental physicist Nikola Tesla or his critics, my experiments with the Tesla coil were the visible proof that we have to take Tesla's statements seriously. Now not only at myself, but also at the students of the polytechnic and the colleagues of my Transfer centre a true Tesla euphoria broke out.

In every free minute the patents and original writings were studied, which I had myself send from the Tesla Society in Colorado Springs in the USA. In particular my trainees and diplomands developed a incredible ambition in making a historical Tesla concept work. They built a whole series of various high-tension generators. In the laboratory one could hear crashing and there was a smell of ozone.

22.1.1998 at a presentation of the works for a degree the candidate looked after by me very proudly held a fluorescent lamp in his hand, which in the field of his self-built high-tension generator glowed even without any wire connected, entirely according to the great mode (fig. 17.10). At all efforts we however laboratory technical weren't able to reach tension voltages of above 511 000 Volt. But at this tension voltage the actual Tesla effects actually start!

Perhaps it was tough luck, but possibly also a chance that we in the laboratories, which I had at my disposal, were forced to work with lower tension voltages and that meant that we gradually had to beak away from the Tesla designs.

Moreover we hadn't at our disposal the precisely controllable spark gaps, which Tesla had developed and used. If one wants to obtain an if possible high tension voltage change (du/dt) for an interaction with neutrinos, then according to today's technology considerably more favourable concepts are offered, for instance with hard switching Power-MOS-Transistors. 50 Kilovolt per microsecond were to meet.

Therefore we changed the technology and worked from now on with fast semiconductor switches. From the laboratory radio now only a hissing and crashing came out of the loudspeaker, if our experiments were running.

At 12.10.1999 we for the first time succeeded to build up a transmission line for Tesla radiation. Doing so the transmitter and receiver were situated in different rooms of my transfer centre. The transmitter coil was operated in self-resonance and fed only from a small function generator with 10 Volt. But if the diplomand held a fluorescent lamp near the spherical electrode, then it started to glow!

Following I observed at the oscillograph the signal of the receiver coil, which as well was operated in resonance. If the diplomand switched off the transmitter, also no receiver singnal was present anymore. But if it concerned radio waves according to Hertz or Tesla radiation, with that still wasn't answered.

Therefore I prompted still another experiment. This time the colleagues observed at the transmitter the signal at the function generator, while I undamped and again clamped the receiver. The shouting with joy from the adjoining room indicated that it had been observed, how the receiver reacted upon the transmitter and both are in resonance with each other. Such an effect characteristic for scalar waves, is a radio technical impossibility! In the case of radio with Hertzian waves an effect back from the receiver on the transmitter is unthinkable by principle.

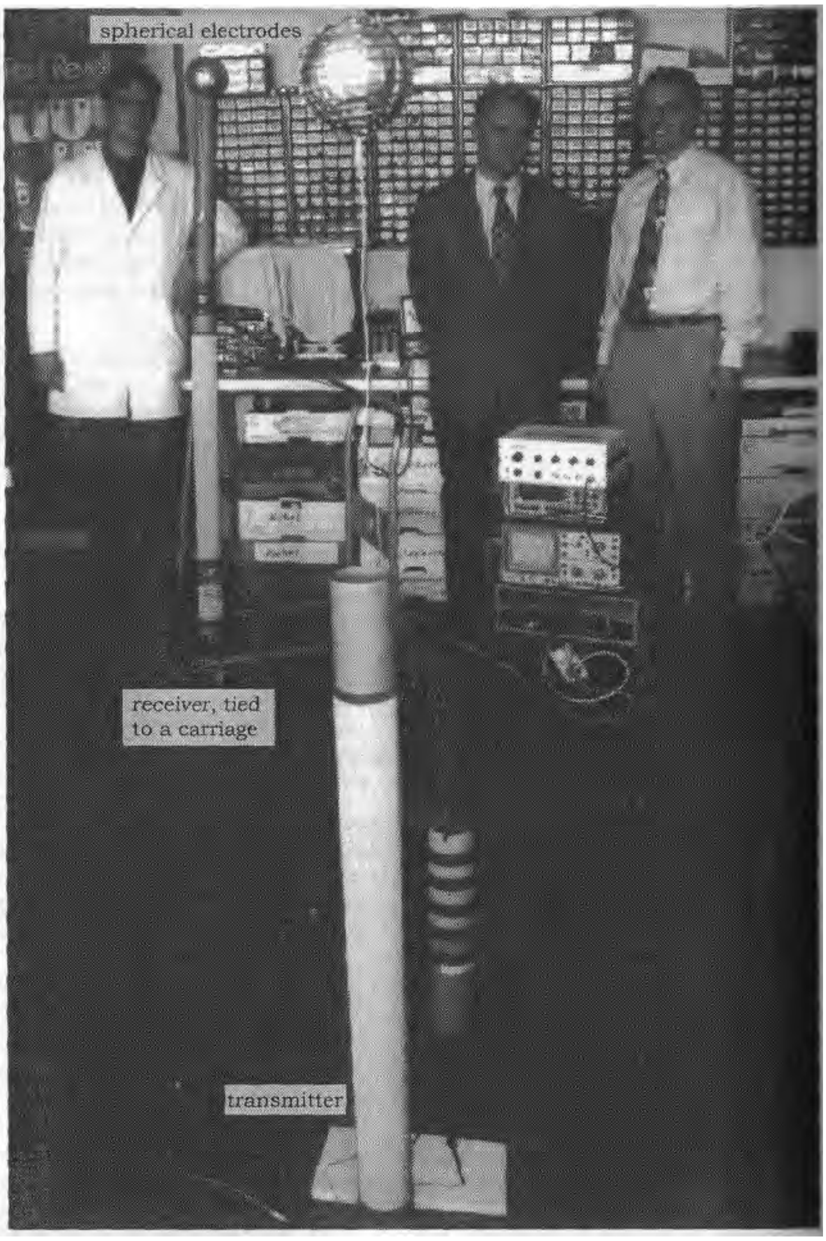
spherical electrodes
receiver, tied to a carriage
transmitter

Fig. 19.11 B: Scalar wave transmission line according to Tesla
f.l.tr.: M.Andresen (diplomand), author (TZ-head), Dipl.Ing.M.Rehm (project leader)

In the next step we let glow the famous little lamp on the side of the receiver. As a consumer served a small light-emitting diode, of which the light intensity remained unchanged in the case of resonance. To prove this, we placed the receiver on a carriage and rolled the corridor in the TZ up and down with it. If the receiver was only slightly out of tune, then from the then arising fluctuation the standing wave nature could be observed perfectly.

With the setup it can be demonstrated well, how the law of the square of the distance, of the decrease of the field strength with the square of the distance, known from radio technology hasn't got validity anymore for scalar waves. Very clear also was the energy transmission of scalar waves out of a closed Faraday cage.

In the end we have determined the degree of effectiveness of the scalar wave transmission line. An output power of 49 mW resulted from the measurements of the current and tension voltage for a loading of the receiver with a 100 Ohm resistance. Simultaneously the power taken up by the transmitter amounted to 233 mW. If we however subtract the idle power consumed by the transmitter from this, and that was determined to be 223 mW for switched off receiver, then actually only 10 mW are available for the wireless energy transmission. The degree of effectiveness according to that would be 490%!

If we here really have obtained an over-unity of 4.9 then the receiver must have collected along free space energy, or did some parts evade our power measurement? The sceptics I recommend an own rebuild, since alone the proof of the scalar wave properties inevitably has to lead to every Maxwell burdened HF technician breaking with the old belief.

19.12 Neutrinolysis, the alternative splitting of water

Free after the plans of Stan Meyer we filled diverse containers with water and let it "crash". If then bubbles raised and it got exciting, we changed the frequency and exactly paid attention to the effect remaining the same or if it increased as well with increasing frequency, thus if we only watched a classical electrolysis, or already the wanted "neutrinolyse", as we were accustomed to term the splitting of water in its parts under the influence of neutrinos in the laboratory. Typical for that is a bubbling and "cooking" of the actually cold water produced by the rising gas bubbles (fig. 19.12).

We also have reversed the principle. Doing so we have switched the container as neutrino receiver and measured the forming charge carriers. The passive system proved to be really moody. On the one hand tension voltages of several hundred millivolts at a load resistance of 10 megaohm can be realized, on the other hand the charged gas particles and the as secondary reaction in the water set free ions continually change the electric conductivity, so that the cause to be measured, the neutrino radiation, hardly can be reproduced on the displayed result. The water hence has to be changed more often and also the developers found themselves between jubilation and disillusionment subject to continually changing feelings.

<i>: A. und I. Schneider: Neutrino-Power - Energie aus dem Kosmos, Bericht zu Vortrag und Demonstration des Autors am 25.11.99 in Villingen, NET-Journal 12/99, S. 4-6; Interview mit Prof. Dr.-Ing. Konstantin Meyl: Durchbruch in der Freien-Energie-Forschung, NET-Journal 12/99, S. 7-9

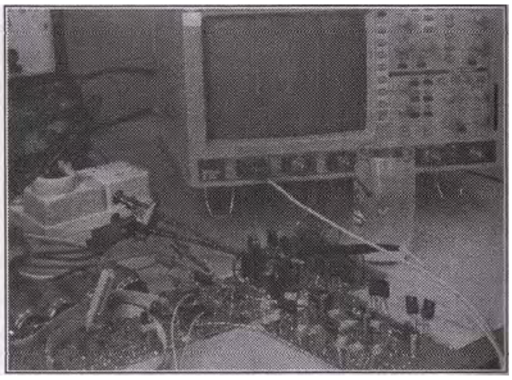

Fig. 19.12: Experimental constructions for the „neutrinolyse“.

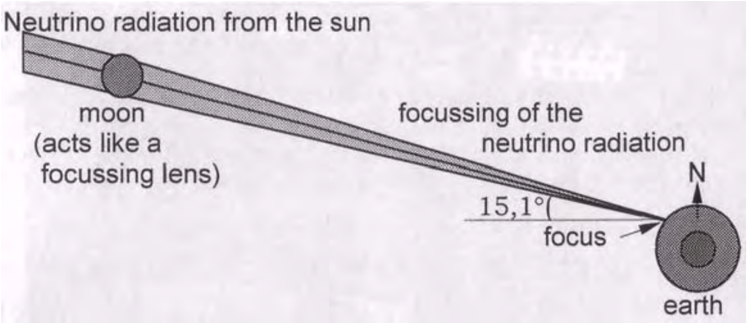
Neutrino radiation from the sun
moon
(acts like a
focussing lens)
focussing of the
neutrino radiation
N
15,1°
focus
earth

Fig. 20.1: Course of the neutrino radiation focussed by the moon on the occasion of the eclipse of the sun at the 11th August 1999.<i>

---

concerning the calculation of the position of the sun at 11.8.99:

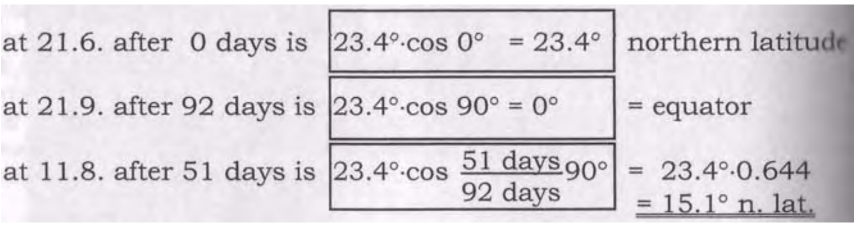

at 21.6. after 0 days is	$23.4° \cdot \cos 0° = 23.4°$	northern latitud
at 21.9. after 92 days is	$23.4° \cdot \cos 90° = 0°$	= equator
at 11.8. after 51 days is	$23.4° \cdot \cos \frac{51\text{ days}}{92\text{ days}} 90°$	$= 23.4° \cdot 0.644$ $= 15.1°$ n. lat.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnen-

At 11.8.1999 the sun at noon stands above latitude 15.1.

20. Cosmic experiments

As long as no usable technical gauges are available, we should observe and study nature and all celestial phenomena. Here the scalar waves with all their properties are presented to us. That is valid in particular for the case that the cosmos makes an experiment with us and with the entire earth. Such a situation offered the eclipse of the sun of August 11 1999 as an unique "bulk experiment" to which travelled millions of spectators, to participate in the cosmic experiment as voluntary test subjects.
My warning for possible influences proved to be absolutely justified afterwards, even if only comparatively few took note<i>. In view of strong changes in the EEG of individual test subjects and a proven temporary acceleration of the rotation of the earth science once more stands before insoluble problems.
To answer the open questions I want to start with my indications, which 1 have published in anticipation of the cosmic event of August 11<i>. Afterwards a revision and the attempt of an interpretation from the scalar wave view follows. Perhaps this way leads to a reliable prediction of of earthquakes and other cosmic events. After that we perhaps know more about origin, availability and further important parameters of the sought-for space energy.

20.1 Millions of voluntary test subjects at 11 August 1999

Astronomically seen, it concerns a harmless natural spectacle as it already was observed more often, if the moon glides between sun and earth and its shadow in broad daylight immerses parts of the earth in a dark night. The special thing about the eclipse of the sun of August 11 1999 however was, that the sun activity just strived for its maximum value in its eleven year cycle, and the orbital distance to the moon simultaneously reached a minimum value. In addition we must pay attention to the special situation of the angles (fig. 20.1). Due to the extremely rare constellation the area of complete shadow at first was larger and darker than normal. On August 11 it had a width of 110 kilometres.

The sun besides the light also sends us solar neutrinos and for those the moon is transparent. It with regard to them acts like a glass ball, which lets the light it is true pass through, but in doing so refracts it. The glass ball acts like a convex focussing lens, which focuses the arriving rays behind the ball in a focus.
To slow down and collect neutrinos the moon it is true is too small, but it will be able to influence the flying direction. Very fast neutrinos, which run through the moon, hardly will be diverted. The slow and biologically active ones however will be bent stronger.

At this point it would be of utmost importance to know, in which distance from the moon the rays run together and combine into a focus.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal, Jg. 4, Heft Juli/August 1999, Seite 13-17

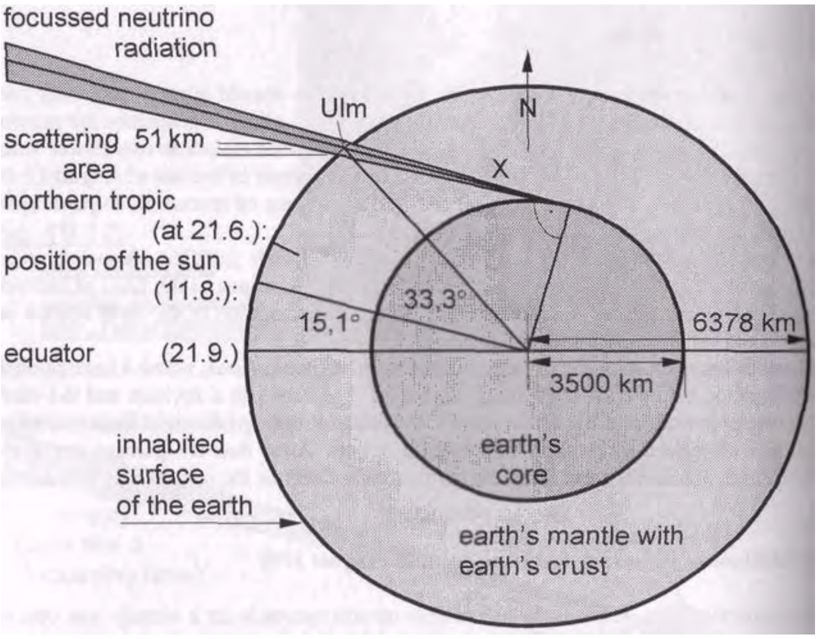
focussed neutrino radiation
Ulm
N
scattering 51 km area
X
northern tropic (at 21.6.):
position of the sun (11.8.):
33,3°
15,1°
6378 km
equator (21.9.)
3500 km
inhabited surface of the earth
earth's core
earth's mantle with earth's crust

Fig. 20.2: The course of the focussed neutrino radiation at 11.08.1999 for 48.4° northern latitude. (this corresponds to the line Ulm-Augsburg-Freising)

calculation of the critical latitude:

$$\sin \alpha = \frac{3500 \text{ km}}{6378 \text{ km}} = 0.5488 \text{ resulting in the angle } \alpha = 33.3°$$

$$+ 15.1°$$

The neutrino rays touch the earth's core at: 48.4° n.lat

minimum distance to the moon: $r_m^* = 358000$ km
radius of the moon: $R_m = 1738$ km
radius of the earth: $R = 6378$ km (at the equator)
length of the Tangente: $X = 3500 \text{ km}/\tan 33° = 5332$ km

scattering radius: $r_x = X \cdot \frac{R_m}{r_m^* + X} = \underline{25.5 \text{ km}}$ around the centre line

Focussing without scattering on earth's surface increase to
$R_m / r_x =$ 68 times , resp. with scattering to
68 * 0.5* 0.8 = 27 times the natural neutrino radiation!

For that we again bear in mind that neutrinos as particle radiation propagate in space in the sense of a plasma wave. Since the oscillation of such a longitudinal wave takes place in the direction of propagation, it neither knows a distinct velocity of propagation nor an upper limit.

The sea of neutrinos in which we swim, is a combination of differently fast particles. The slow ones are bent stronger by the moon, with the property of a convex focussing lens, so that the focus should be looked for near the moon, whereas the fast ones hardly are diverted. Their focus lies further distant from the moon than the earth, up to the extremely fast cosmic neutrinos, which experience almost no diversion, because they, as a result of the Lorentz contraktion, are small enough to tunnel through any kind of matter.

If neutrinos, depending on their velocity of propagation, have their focus partly before and partly behind the earth at an eclipse of the sun, then it comes up to a conclusive logic that there actually exist such, of which the focus lies exactly in the centre of the complete shadow on earth's surface. But thereby the question is asked, which biological effectiveness these neutrinos have or which damage they bring about.

20.2 Model of calculation for the cosmic experiment

More than a year ago the results of an international neutrino experiment have been made public. Thereby for the first time the order of magnitude of the natural radiation density was recorded. after the detector before having been calibrated at an artificial source of neutrinos. As the perhaps most important result at night only half as much solar neutrinos could be detected as during the day with the immense Super-Kamiokande detector in a Japanese mine. With that the here taken interpretation, that earth's core collects neutrinos, meanwhile even is proven experimentally! For that it has to interact with the particles and that means, it has to exert a force of attraction on them. Thus the earth's core during an eclipse of the sun will further amplify the effect of focussing if the konzentrierte neutrino ray is directed on earth's core. This critical point we have to calculate (fig. 20.2).

The proportion of the radii of earth's core (3500 km) and the entire earth (6378 km) results in the sine of the sought-for angle and that amounts to 33.3 degrees northern latitude. At August 11 moreover is added that at noon the sun with regard to the equator is standing at 15.1 degrees. so that a first extreme focussing should be expected, if the centre of the complete shadow intersects latitude 48.4. This would be the case for the latitude of Ulm in direction Augsburg and Freising.

Now we would like to know more about the amplitude of the focussed radiation, about the spatial extension and the period of time. Without concrete data material we have to proceed from several simplifying assumptions. If we therefore assume, 50% of the at earth arriving and biologically relevant neutrino radiation stems from the sun, which in the case of an eclipse of the sun is focussed to 80% and scattered to 20% by the moon. If we further assume the focus just touches earth's core, then between Ulm and Augsburg a sphere of action of 50 km can be expected, within which the neutrino radiation on the average reaches 28 times the value of the natural radiation. The intense irradiation under these assumptions will last one minute.

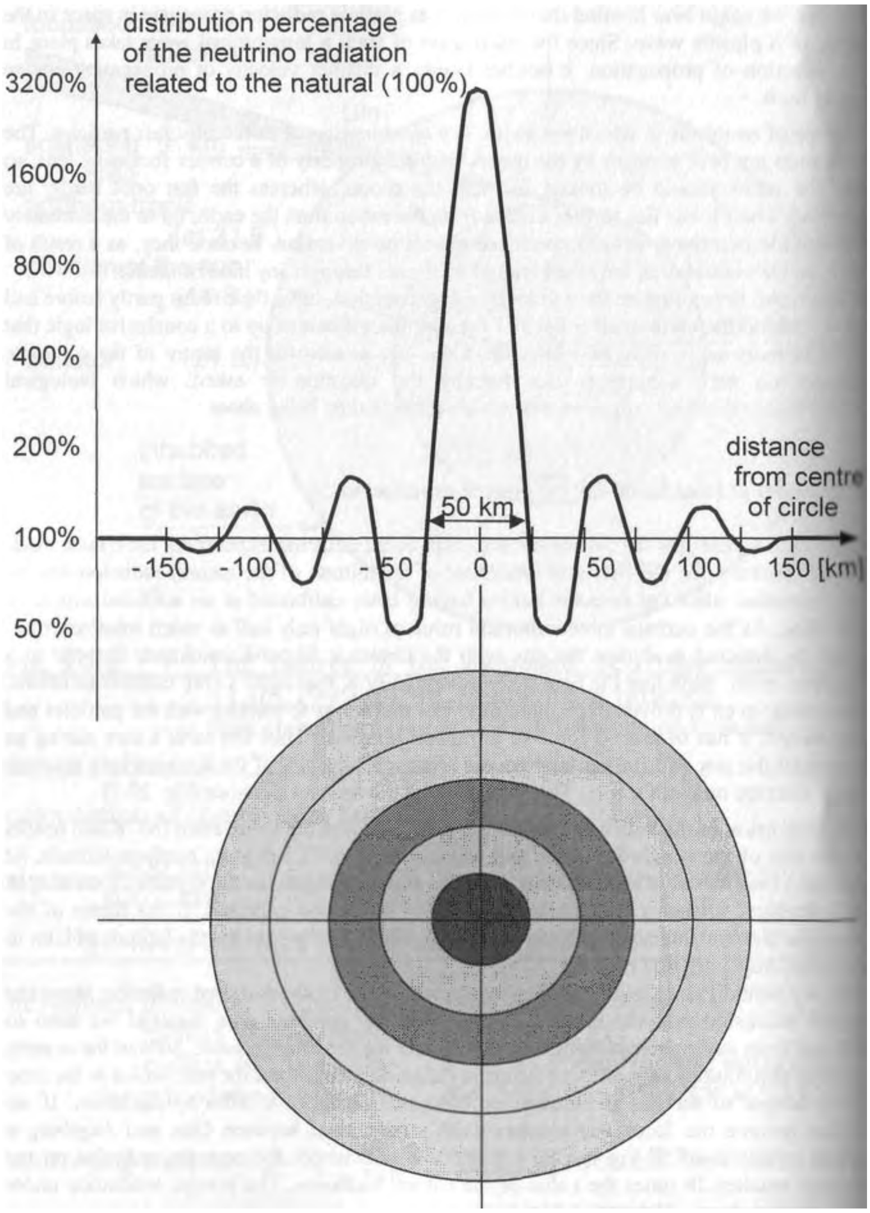

Fig. 20.3: ____ The spatial distribution in the case of focussing of the neutrino radiation by means of the burning glass effect of the moon on earth's surface.

Of course the calculated values only are valid to the extent, as also the assumed boundary values apply. After August 11 we know more about this cosmic experiment, about the spatial distribution and indirectly about the interaction with earth's core, about the physical properties and the biological effectiveness of the at present available neutrino radiation.

20.3 Physical technical consequences

The neutrino radiation is a scalar wave radiation which, as mentioned, can be perceived by sensitive people even without aids. Who hasn't at his disposal this sensitiveness, is recommended a simple setup.

For that one hangs up a fluorescent lamp, connects the one end with a piece of wire, as it were as antenna, and the other end is grounded. For a scalar wave radiation which increases fast, the lamp should start to glow by itself. Under big transmitting installations this method already has been successfully applied by many allotment gardeners thousandfold. I hence recommend all research scientists of eclipses to test themselves the possible field fluctuations in this simple way.

From a technical viewpoint first of all the atomic reactors and their nuclearly contaminated waste will be concerned by a fluctuation of the neutrino radiation. In view of present eclipses of the sun an accident can be expected less, since the neutrino radiation it is true for a short time reaches an extreme maximum, which averaged temporally and spatially over the whole event again is relativized somewhat. The relations shall be clarified with an example (fig. 20.3).

Whoever places himself in the centre line of the complete shadow on August 11, at first will detect a decrease of the neutrino radiation to 50 to 60 percent, then a steep increase to 2800 percent and from the summit again the whole backwards, while standing on the earth he turns by under the moving moon shadow. The ring with half the radiation, which reaches us first, doesn't pose a problem since, as said, we only have half the radiation in every night. Some animals and plants as aresult erroneously will set out for sleep.

The wave distribution one can imagine like that in the case of a pool, in which was thrown a stone. But we still don't know the resonance frequency, for which reason the length of the cycle depicted in fig. 20.3 is chosen arbitrary. The actual deviation from the distribution given by nature is the peak in the focussed ray centre. Living nature must stand large fluctuations of the solar radiation, since every supernova sends us a relatively short batch of fresh neutrinos. Where we have difficulties is the question, how much fluctuation still can be tolerated by mankind.

The question for possible biological consequences is due to be dealt with in view of the announcement of a complete shadow tourism causing concern, as it is awaited for August 11. At the incomplete state of knowledge about the properties of neutrinos, every trip to the complete shadow remains a journey into the unknown.

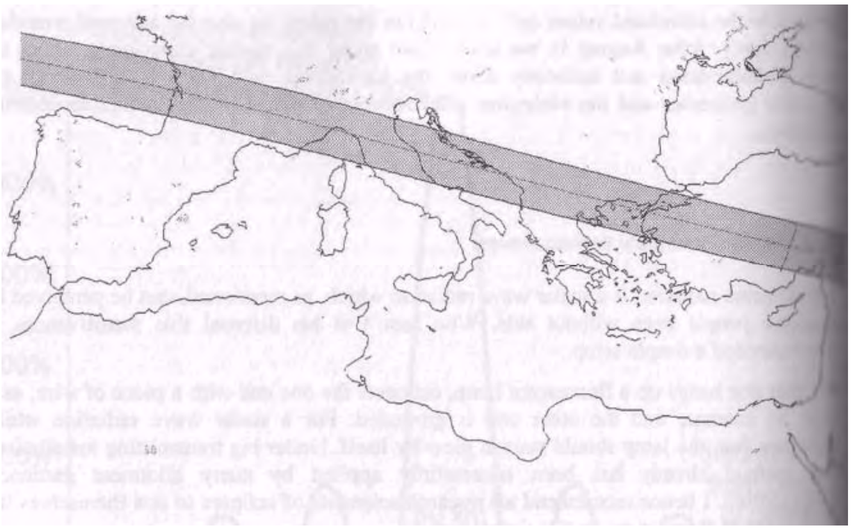

Fig. 20.4 A: The course of the complete shadow for the ancient eclipse of the sun at 28.5.585 B.C.

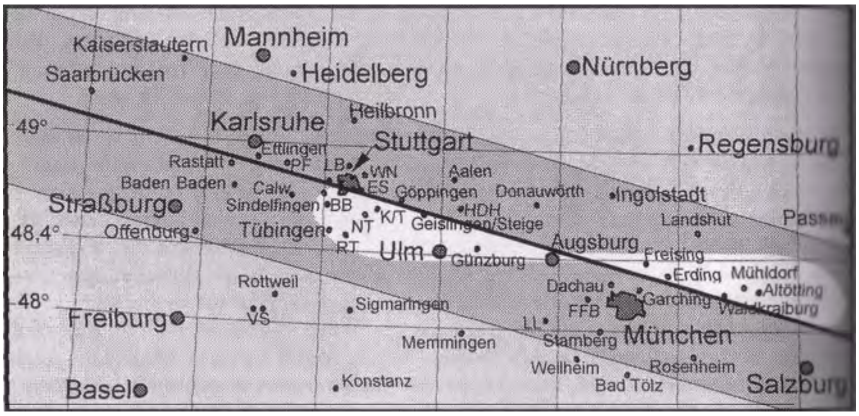
Kaiserslautern
Mannheim
Saarbrücken
Heidelberg
Nürnberg
Heilbronn
Karlsruhe
49°
Stuttgart
Regensburg
Ettlingen
Rastatt
PF
LB
WN
Aalen
Baden Baden
Calw
ES
Göppingen
Donauwörth
Ingolstadt
Straßburg
Sindelfingen
BB
HDH
Landshut
Pass
48,4°
Offenburg
Tübingen
NT
K/T
Geislingen/Steige
Augsburg
RT
Ulm
Günzburg
Freising
Erding
Mühldorf
Rottweil
Dachau
Altötting
48°
Garching
Freiburg
VS
Sigmaringen
FFB
Waldkraiburg
LL
Memmingen
Starnberg
München
Wellheim
Rosenheim
Bad Tölz
Konstanz
Basel
Salzburg

Fig. 20.4 B: The course of the complete shadow over south Germany at August 11 1999 (grey) and the possible course of the focussed neutrino radiation (white).

20.4 Biological consequences

In the case of an eclipse of the sun effects on the biology, like problems with the heart among affected, at least can't be excluded. If the scalar wave density increases above the density which is normal, then this has a positive effect on the economy of energy, as long as the body is capable to regulate the amount taken up. If the regulatory process however should fail. then the risk of a self-inflammation exists. Also straw bales and other organic and inflammable materials could thus go up in flames.

But before that happens, first the information technical influence of the scalar waves will show. Here we have to expect a psychotronic influencing, which is showing in a limited ability of perception. History teaches us as an example<i> that a by Thales of Milet predicted total eclipse of the sun at 28.5.585 B.C. compulsorily has ended a battle in Asia Minor between the Medes and the Lydians, because the soldiers apparently most literally had gone out of their mind (fig. 20.4 A).

Actually all in connexion with the free energy addressed phenomena are conceivable. from the "neutrinolyse" up to the acceleration of the radioactive decay. It would be understandable, if in a water glass bubbles should rise, even if no carbonic acid is contained in the water at all. After man not having at his disposal a sense organ for his own energy source, the brave in the complete shadow of an eclipse of the sun are recommended smaller technical experiments and observations. The cautious however will avoid the area from the start.

As counter movement to the complete shadow tourists there will also be refugees, who believe in the predictions of Nostradamus, who in his quatrains has predicted a messenger of fright over Europe for 11.8.1999. He mentions Lyon, Ulm and Moskau, which actually lie on one line<ii>.
The line of the complete shadow however will run under another angle from Plymouth in South England over Ulm to Bukarest and further into Turkey. Maybe Nostradamus wasn't a clairvoyant at all, but merely a good calculator, or he knew someone who could calculate excellently, after all he has indicated the time and even Ulm as the centre correctly. Apart from the small angle error, in addition the direction is correct (fig. 20.4 B).

Worth paying attention to also is, that he contrary to his habit here gives a concrete date which astronomically can be calculated unambiguously, that he simply skips the numerous eclipses of the sun of the past and only points to the one of 11.8.1999, which runs crossways through South Germany. Extremely sinister are his forecasts, which mustn't commented on further, since he speaks of "Mort et Tombe", of death and grave<ii>.

<i>: Merodot: Historien, Kroners Taschenausgabe 224, S. 33 (cf. page 586 and the modern misinterpretation note 9 on page 754); further sources: dtv-Atlas zur Weltgeschichte S. 45; resp. Propylaen-Weltgeschichte I, S. 168.

<ii>: Michel de Notredame (1503 bis 1566), Mathematiker, Astrologe, Leibarzt von Konig Karl IX: Centuries (1558) X, Vers 72.

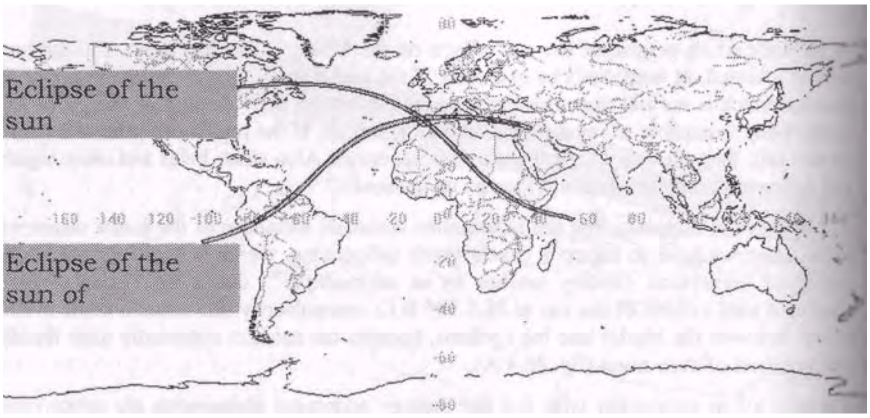
Eclipse of the sun
Eclipse of the sun of
-160
-140
-120
-100
-20
0
20
60
80
-40
-60
-80

Fig. 20.5 A: The total eclipses of the sun of 18.4.1539 and 21.8.1560

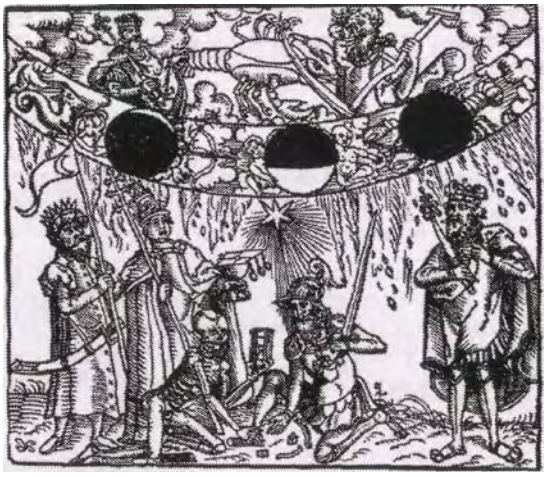

Fig. 20.5 B: Under the impression of the last two eclipses of the sun over Europe death and war play dice about the fate of mankind during the eclipse of the sun of 1562.<i>

<i>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 156.

20.5 Epilogue to August 11th 1999

My indications have been published two weeks before the cosmic event in the NET-Journal[i]. The public interest was immense. Now, after the spectacle is over, it is time for an analysis[ii].

Besides numerous very subjective descriptions of eyewitnesses, which scientifically hardly can be analysed, at numerous polytechnics over the world has been observed, how Foucault pendulums suddenly and completely unexpected have deviated from their normal swinging direction. With this device 150 years ago the rotation of the earth had been detected in Paris.
Since according to prevailing textbook opinion an eclipse of the sun is a purely optical phenomenon, the scientists world-wide are having a big problem. Which force here has teared at our earth and caused relative accelerations of the kind that the pendulums could turn out of their usual plane and changed into an elliptic orbit, while the shadow of the moon ran over us? The gravitational force isn't even roughly capable to that. That merely has brought a 50 cm higher flood. The enormous force effect, which even puts the gravitation in the shade, actually only can come from the interaction of the neutrinos.

Different reports are present concerning the influencing of the radioactivity. In the cases, in which measurement samples have been used, almost no change could be observed. This is confirmed by a video tape, which northeast of Munich directly in the centre of the complete shadow documents an experiment, in which during the whole time the radioactivity of 1 kg crude granite is monitored with a professional dosimeter. The background might be that calibrating samples are chosen in principle under the aspect of being influenced the least by outside interference sources. It therefore would have been better, if we instead had put a lettuce as a biological and broadband sample in front of the device. Because there, where one anyway hadn't expected deviations and thus neither were measured calibrating samples nor unfortunately were made recordings, is said to temporarily have occurred a visible increase.
Even in the case that somewhere accidentily should exist recordings, the increase is too small, to help explain traditions from the Middle Ages, according to which is talked about "stinking log" and about "aggressive damps, which fall from the sky". Other sources speak of harmful radiations, which one regarded as the trigger of plagues.
It is said: "They poison the drinking water and the crops and make people sick"[iii]. Crops after an eclipse of the sun either should not be harvested at all or only after a violent shower. The warnings for a poisoning "without adding poison", which reach back into the 19th century, suggest the assumption that it concerns results of a radioactive contamination.

<i>: Konstantin Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal, Jg.4, Heft Juli/August 1999, Seite 13-17

<ii>: From the third and extended edition of this second part concerning Electromagnetic Environmental Compatibility the following chapters are completed.

<iii>: Mark Littmann/Ken Willcox: Totality - Eclipses of the Sun, Honolulu 1991, Kapitel 4: Eclipses in Mythology; nachzulesen bei Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, Seite 130/131.

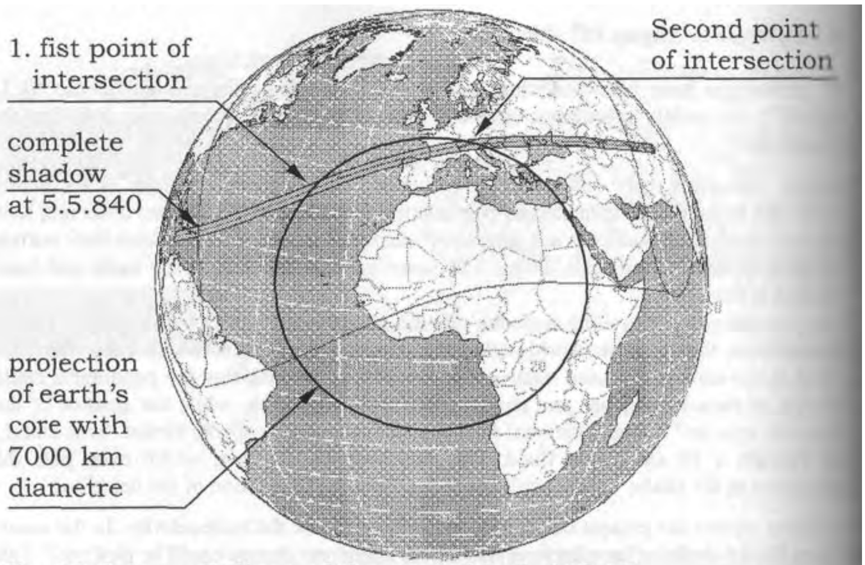
1. fist point of intersection
Second point of intersection
complete shadow at 5.5.840
projection of earth's core with 7000 km diametre

Fig. 20.6 A: Eclipse of the sun of 5. 5. 840 AD.

Fig. 20.6 B: Nostradamus announces catastrophes for 1999 in the Centuries of 1558.<i>

<i>: Bryan Brewer: Eclipse, Chapter 1: Eclipses Throughout the Ages, Seattle WA 1991, S. 20, cited in Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 130, Bild Seite 159.

20.6 Dropping of the neutrino radiation

Plagues of that kind, which in the Middle Ages have claimed lives as a result of an eclipse of the sun, hardly are verifiable, unless a prominent victim was among them, like e.g. the son and successor of Charlemagne, emperor Ludwig I. He 5.5.840 witnessed an eclipse of the sun with a totality of five minutes. Further it is said: „The fright which this experience gave him, little later shall have torn him to death"<i>.

According to that there must have existed times, in which the radioactivity present in the countryside and stored in the body of a person during an eclipse of the sun could be set free and be the undoing of the affected. At the same time such a decontamination of radiation acts cleaning for nature.

In this case only the interpretation of Nikola Tesla provides us an explanation, which states that the neutrino radiation causes the radioactivity<ii>. Textbook physics however doesn't know this causality. For that the primeval fears of humanity are pure superstition. The prophesies of Nostradamus even are referred to as counter evidence.

But the question remains open, why his predictions concerning August 11 didn't happen, Obviously, so has shown us the cosmic experiment, the neutrino radiation relevant for the setting free of radioactive radiation drastically has decreased since the Middle Ages.

Nostradamus personally has occupied himself with the translation and interpretation of hieroglyphs and has written down his insights - surely out of fear of the inquisition - in the form of encoded quatrains. According to that he has based his considerations on considerably older sources, which presumably stemmed from a time, in which a considerably higher radiation prevailed.

We have to proceed from the assumption that the scientists of the Semitic- Aramean people of the Chaldeans, which ruled Babylon from 626 B.C., were just as capable as astronomers of today, to exactly calculate an eclipse of the sun even centuries and millenia in advance.

After all the Saros-cycle to determine eclipses of the sun is a discovery of the Chaldeans. What they however couldn't know and we ourselves still can't indicate today, is the prevailing density of the cosmic radiation at a future time. But that obviously has changed considerably.

The natural magnetic field strength for instance is recorded at the baking of earthenware jugs and vases, by strengthening along the parts containing magnetite. From the measurement of ancient earthenware goods we know that in antiquity a field strength must have prevailed which was 3 to 4 powers of ten higher.

According to the here presented theory the earth owes its magnetism its core and that again draws its energy from the neutrino field. According to that also the neutrino radiation should be subject to the same decrease.

If therefore the Babylonians out of the radiation situation at that time in the interpretation of Nostradamus and other fortune-tellers, who presumably all more or less have written each other off, have predicted a catastrophe for 11.8.99, then this scientifically is just as untenable, as the today widespread hubris, with which the knowledge and the reports of experience from ancient times are dismissed as superstition.

<ii>: cf. chapter 17.2 Nikola Tesla, the discoverer of the neutrino radiation, p. 133

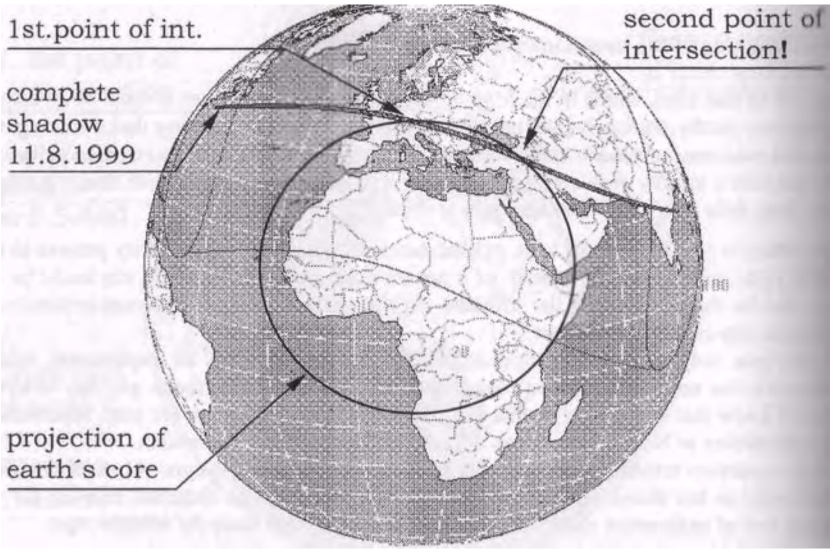
1st.point of int.
second point of intersection!
complete shadow 11.8.1999
projection of earth´s core

Fig. 20.7 A: Eclipse of the sun of 11.8.99 (declination: 15°)

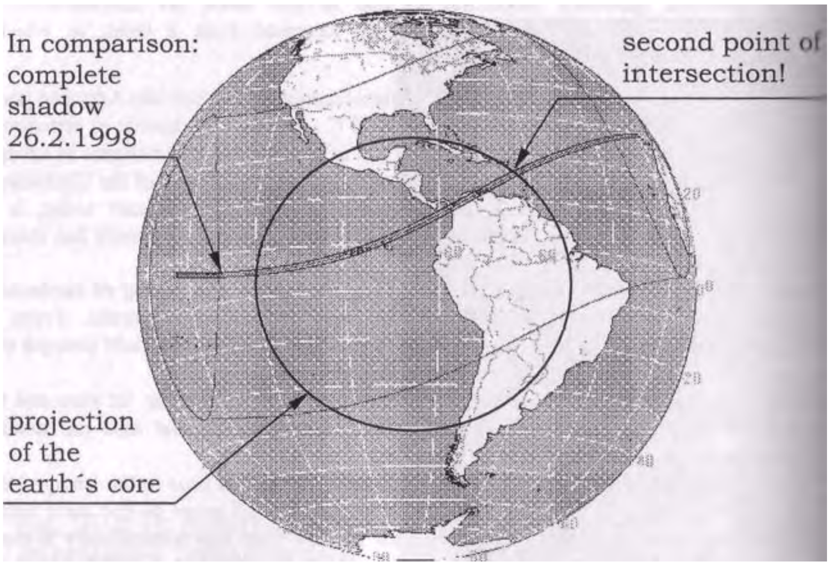
In comparison: complete shadow 26.2.1998
second point of intersection!
projection of the earth´s core

Fig. 20.7 B: Eclipse of the sun of 26.2.98 (declination: - 9°)

<ii>: Particularly violent earthquakes since 11.8.1999 (excerpt):
Turkey 17.8. (strength 7.8) till 19.8. (5,0), 31.8. (5.2), Greece 7.9. (5.8). Turkey 13.09. (5.8) and finally Taiwan 20.9.99 (strength 7.6) in a rythm of 6 to 7 days! Complete list in the internet under:
http://www-seismo.hannover.bgr.de/ermos_listing.html

This "superstition" mentions tremendous natural disasters like floods like the Flood or destructive earthquakes often as a direct result of an eclipse of the sun, and as the worst consequence the end of the world<i>.

Almost all reporters and newsreaders in the evening of August 11 full of irony pointed at the non-occurrence of the end of the world and spread the conviction that according to prevailing physical ideas something like that isn't possible at all. They at that time couldn't know that the anatolic plate, which the complete shadow had crossed, had gotten into motion. The relatively weak earthquakes, which 11.8. shook Cyprus and at the same time Iran, only were spontaneous harbringers.

In the following weeks the staggering core of the earth made us clear that there had been done force to it. Severe earthquakes with thousands of aftershocks followed each other and one message of terror chased the next<ii>. Public authorities however take care not to of decline to make a reference to the eclipse of the sun.

20.7 Analysis and comparison of the occurrences

If we for comparison consult the total eclipse of the sun of 26.2.1998, for which the complete shadow of the moon coming from the Pacific Ocean had run over the Caribbean into the Atlantic Ocean. Exactly the moment it crossed the Caribbean island Montserrat, the volcano Soufriere erupted.

Pure coincidence says science, which hasn't got an explanation model at all for a relation with the supposedly purely optical phenomenon. But this argumentation is relativized, if according to fig. 20.2 the tangential collecting of solar neutrinos by earth's core is considered. This process best can be compared with photon radiation, which is tangentially collected and directed into an orbit by a black hole inside the radius of Schwarzschild.

Since the shadow of the moon always draws a straight line on the earth, two points of intersection are present with the circle of the projection of earth's core on earth's surface, for which the radiation focussed by the moon just touches earth's core and in that way experiences an additional concentration. The first point of intersection at that time was situated in the Pacific Ocean; but the second one was situated exactly at the Caribbean island Montserrat (fig. 20.7 B).

This time, at August 11 1999 the first point of intersection was situated in South Germany, the second in Iran, and again the focussing at the second point of intersection has shown a devastating effect. The entire continental plate has gotten in motion (fig. 20.7 A).

Another strange phenomenon has occurred in the USA at the same time as the eclipse of the sun. A tornado swept with its destructive force right through Salt Lake City. It is remarkable that no meteorologic indications were showing before and hence official observation authorities had no possibility of warning for the tornado. Had here part of the focussed neutrino radiation been redirected at earth's core and given a rotation, to again screw out into the sky on the other side of the earth at Salt Lake City?

<i>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Miinchen 1999, ISBN 3-89631-302-9, Seite 120.

<ii>: Collection of particular violent earthquakes since 11.8.1999: see fig. 20.7.

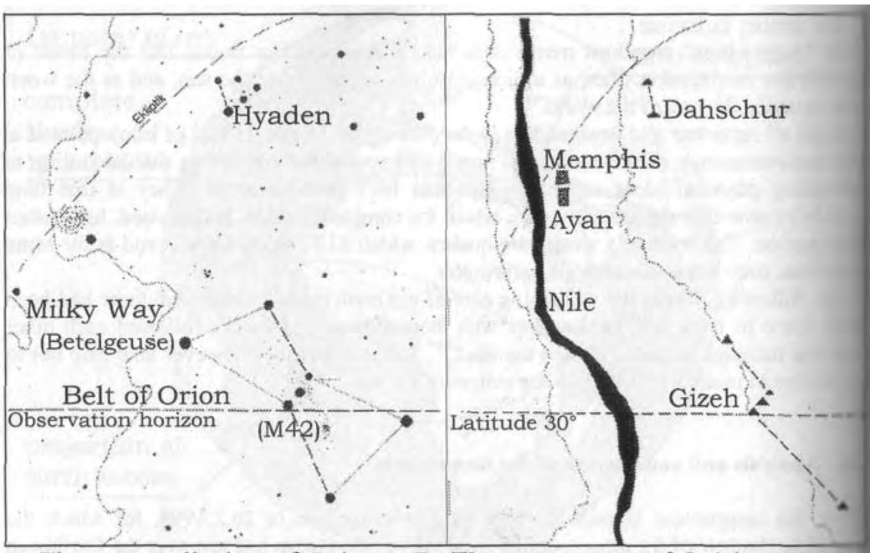
Ekliptik
Hyaden
Milky Way
(Betelgeuse)
Belt of Orion
Observation horizon
(M42)
Dahschur
Memphis
Ayan
Nile
Gizeh
Latitude 30°

A: The constellation of Orion B: The sanctuary of Osiris

Fig. 20.8 A and B: The ancient Egypt pyramids, a terrestrial model and copy of the starry sky?<i>

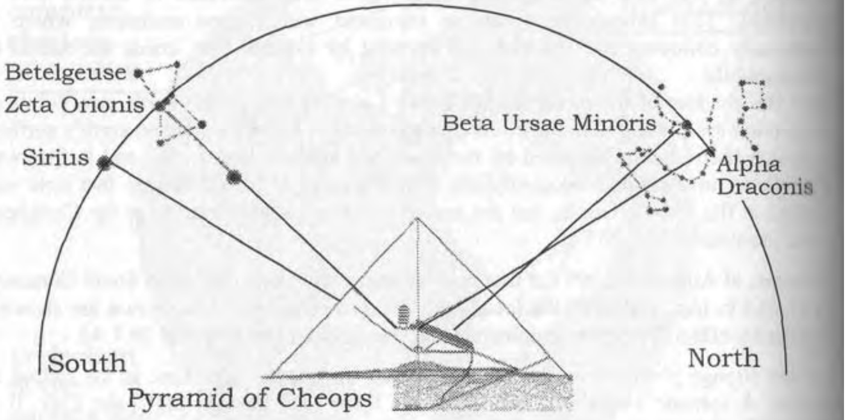
Betelgeuse
Zeta Orionis
Sirius
Beta Ursae Minoris
Alpha Draconis
South
North
Pyramid of Cheops

Fig. 20.8 C: The orientation of the four shafts of the Pyramid of Cheops to four different stars around the year 2500 B.C.<ii>

<i>: Robert Bauval, Graham Hancock: Der Schlussel zur Sphinx (Keeper of Genesis), List Verlag Munchen 1996, ISBN 3-471-77188-3, S. 286 und 287.
<ii>: dito page 86.

20.8 Radioactive decontamination with the help of a supernova

A mighty source of neutrinos forms the black hole in the centre of the Milky Way. For us, on an outer spiral arm of the galaxy, the distance to the centre is gigantic as well, so that a relatively equally distributed spectrum of differently fast neutrinos arrives at our world, which represents a kind of basic energy technical supply for the solar system and our world. For the fluctuations between day and night, or the focussing by the moon or other planets the "participants" in the solar system are responsible themselves.

In contrast to that a supernova, the explosion of a star, is a considerably smaller source of neutrinos, which however also is possible less distant to the earth. In addition it is a singular event, in which all neutrinos are set free simultaneously within a fraction of a second. They arrive at us one after another. First the fast and hard radiation reaches us as it were as harbringer. In the course of time the arriving neutrinos then become slower and shower, until they sometime become biologically relevant. If in the end everything is over, we can see the cause, only now the supernova is showing in the telescope.

If we assume such an event takes place, with perceptible might and in a distance of 500 light years. then this neutrino radiation overlaps with the general background radiation and a characteristic over-intensification of neutrinos of a certain velocity of propagation occurs. This problem then occupies us for 500 years, where the respective radiation situation permanently is changing depending on the time after the explosion.

If we in this way of looking dare a judgement of the cosmic events in historical time, so makes believe much the assumption that the radiation in the last hundred years has worn of completely. Edgar Cayce treats in the book "Our Ancestors" different cultural circles from the old Indian up to the Hopi, in which still is talked about an energy technical use of quartzes and other materials<i>.

We indeed can theoretically comprehend that the neutrino radiation can let an oscillating quartz glow, if it is stimulated in its resonance frequency, but technically the technology foday can't be realized anymore. Possibly the chance for technological use only existed for a few years or decades.

Presumably also the pyramids originally have been built as resonators, to slow down fast neutrinos to a technically utilizable speed. But in the course of time the original function was unnecessary and the neutrinos had gotten so slow that in antiquity alternatively an use as electrostatic lightning generator or as Nekropolis took place. Today they only stand in the countryside as unusable monuments of a gone epoch.

Many ancient techniques, to which I will come to speak in the third part, in this way just as unexpected become plausible as the radioactive decontamination described in the Middle Ages.

<i>: Hermann Wild: Technologien von gestern, Chancen fur morgen; Jupiter-Verlag Bern 1996, ISBN 3-906571-13-0, (z.B. S. 77 und 145 bis 163).

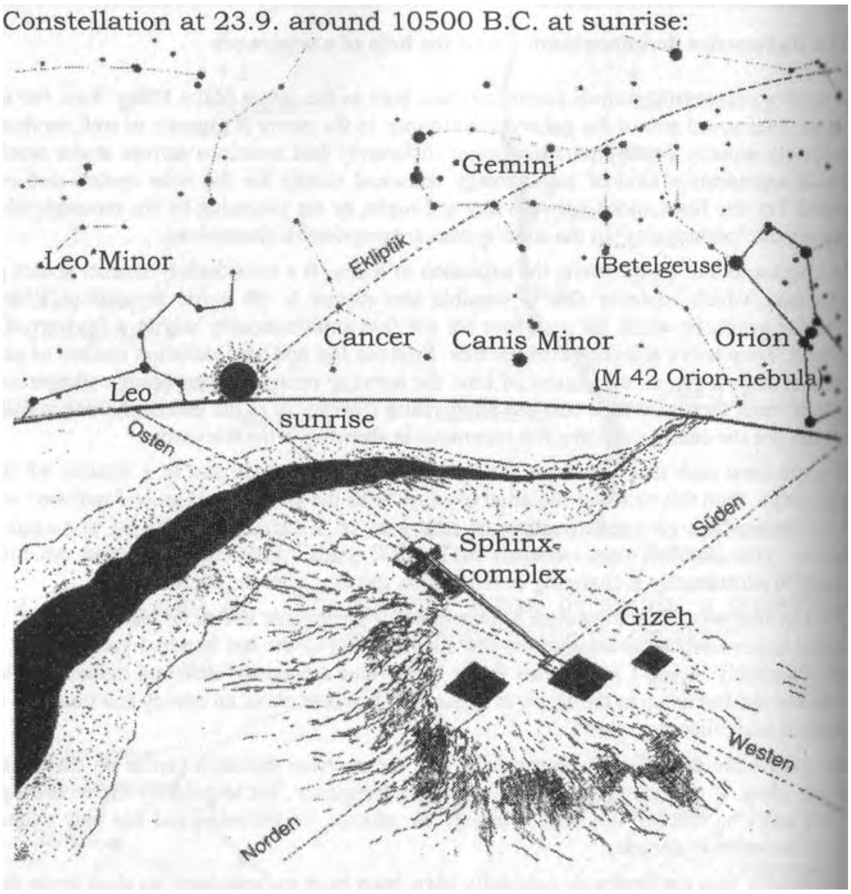
Constellation at 23.9. around 10500 B.C. at sunrise:
Gemini
Leo Minor
Ekliptik
(Betelgeuse)
Cancer
Canis Minor
Orion
(M 42 Orion-nebula)
Leo
sunrise
Osten
Süden
Sphinx-complex
Gizeh
Westen
Norden

Fig. 20.9: The orientation of the pyramids of Giza makes believe
an operation of the installation around 10500 B.C.<i>

<i>: Robert Bauval, Graham Hancock: Der Schlussel zur Sphinx (Keeper of Genesis), List Verlag Munchen 1996, ISBN 3-471-77188-3, S. 324.

The two research scientists take the hypothesis of an orientation of the ancient Egypt pyramids after the starry sky around 10500 B.C. This however is not compatible with their thesis of the aligning of the shafts to individual stars around 2500 B.C. (fig. 20.8 C). How should the precise worked shafts later have been integrated? It would correspond a certain logic, if an orientation of the buildings to the energy source at that time would have taken place; but doing so it should be taken into consideration that the stars are exploded as supernovae and today hardly might be observable.

Despite several good ideas more questions are raised by the book, than are answered. Have the technicians 2500 B.C., after making restorations, experimented with the buildings and sought-for alternatives of use?

20.9 Fre energy from the Orion constellation

After all hundred years ago still the last unreliable rests of the wearing off neutrino radiation were available to Tesla, Moray, Keely and other inventors for the experimental proof of free energy, of which we today aren't able to rebuild and show in function one single model. And that, although the technical aids have gotten better for many times. But for the free energy inventors that isn't a reason, to immediately stop their efforts, because the next supernova already is announced by the actual swings of the Foucault pendulum. It could concern the explosion of the giant red star Betelgeuse in the constellation of Orion in a distance of 500 light years<i>. It well may be possible that it already shortly has exploded and that it will supply us with fresh and free energy for scarcely the next 500 years and at the same time will give the earth a good shake, thanks to its might and a not too great distance.
The reaction, which happened after August 11, of earth's core, which is the first to interact with the fast particles, should make every astrophysicist clear that here something is coming towards us!
It would be obvious, if the reactions of earth's core still increase. Every year particularly around 21.6., if the sun is standing in the Orion constellation, deviations can be expected. But then the supposed source of neutrinos, Betelgeuse, the sun and the earth don't exactly form a line, because the giant red star lies 7.4° below the ecliptic. Because of that the rays slowed down and focussed by the sun run away over the earth. At the earth then rather would be expected a dropping of the radiation.
If the phase of shakes of the earth sometime should be over, a decontamination due to an increased radioactive decay and various biological effects should be expected. Then, perhaps in 200 years, also many concepts concerning free energy, today still dismissed as hopeless, suddenly will function entirely by themselves.

The relatively free possibility of development of the human mind and the present sciences we possibly owe the special circumstance to be able to live in a time of minimized field strengths. Strong fields however can lead to psychotronic influencing of the consciousness and to an outside determining of mankind. This circumstance seems to have caused Tesla to compare man with a robot, and to call him an independent machine controlled from outside<ii>.

<i>: Freek Reijmerink: Sternenatlas Deutschland, Weltbild Verlag 1990, S.12 (ein sterbender Riese) und Illustrierte Wissenschaft, Nr. 9, 1999, S. 7, Brennen die grossen Sterne in den Sternbildern je aus? Die Entfernungsangaben schwanken zwischen 270 Lichtjahren (Meyers Lexikon), 310 (Sternenatlas), 500 (Ill.Wiss.) und 652 Lichtjahren (Cambridge Enzyklopadie d. Astronomie).

In the observable domain of the starry sky one statistically seen has the chance, to experience every second a supernova.

<ii>: Nikola Tesla: How cosmic forces shape our destinies, New York American, 7.2.1917, and Edition Tesla (1997), Bd. 6, ISBN 3-89539-245-6 der Mensch als machine, S. 65.

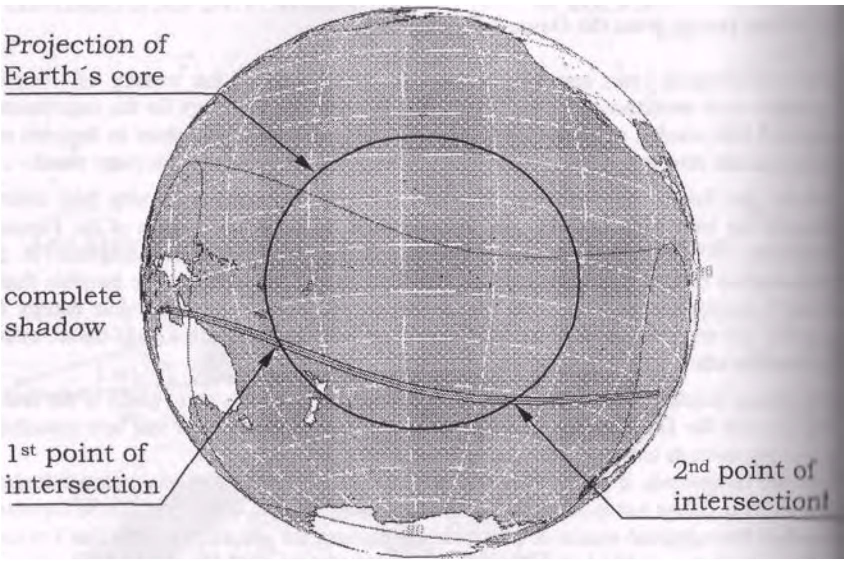
Projection of
Earth´s core
complete
shadow
1st point of
intersection
2nd point of
intersection

Fig. 20.10 A: Eclipse of the sun of 13.11.2012 (decl.: 18.1°)

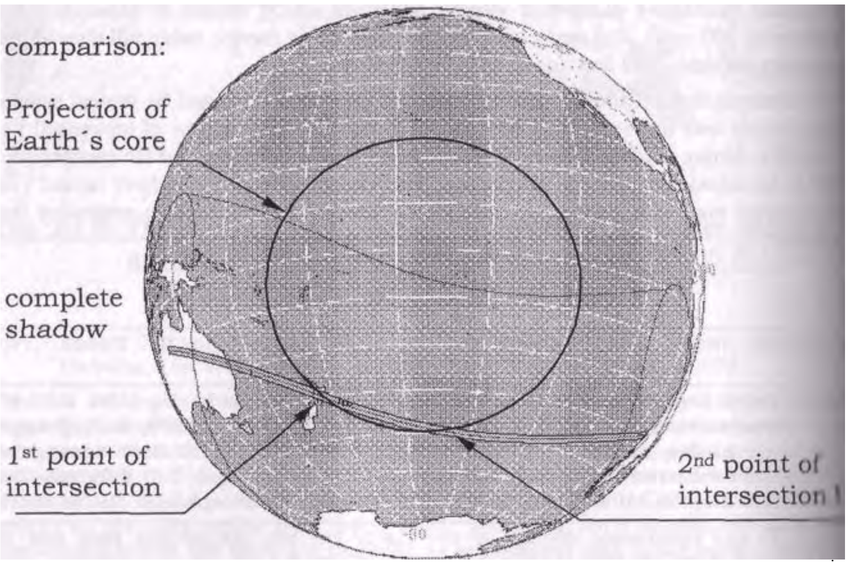
comparison:
Projection of
Earth´s core
complete
shadow
1st point of
intersection
2nd point of
intersection

Fig. 20.10 B: Eclipse of the sun of 11.11.3799 (decl.: -17.3°).<i>

<i>: The eclipse of the sun of 7.7.3797 is harmless compared to that of 11.11.3799. Have they been mixed up or is a calculation error present?

20.10 Interaction of the neutrinos with earth's core

A strong neutrino field still doesn't make a catastrophe. Only in connection with one of the regularly happening eclipses of the sun should one be expected under certain circumstances. Only, which eclipses of the sun can get dangerous, we have to ask us, and why warns e.g. Nostradamus only for very particular dates?
The check of the respective eclipses of the sun results in a critical constellation every time for the cases, where the line of the complete shadow and the circle of the projection of earth's core intersect under a very flat angle and both points of intersection lie very close together. In the extreme case finally the lines only are touching and the points of intersection fuse to a line of intersection.
A corresponding constellation the next time is expected at 13.11.2012. For 7.7.3797, at similar conditions Nostradamus foretells the end of the world; but why?

From the interaction of the neutrinos arises as from every other interaction a force effect. If the points of intersection lie far apart, then earth's core is pulled once to the East and a short time later again to the West by the focussed neutrino radiation. On the average this will hardly influence earth's mantle and earth's crust because of the immense moment of interia. The possible earthquakes will remain regionally restricted to the area around the two points of intersection.
But if a line of intersection forms, then no compensation of the force effects takes place anymore, then during the whole time one-sided is pulled at earth's core and that can have fatal results. It is the same as for a spinning top, which is given a blow from the side: it staggers several times, until the gyroscopic forces have stabilized it again.

But if earth's axis staggers, then the sun describes strange orbits in the sky, it goes backwards again, for a longer time doesn't set or it doesn't show for the same period of time for the people living on the other side of the globe.
Such an event already is described in the Bible<i>. For the twenty hours, in which in Europe the sun didn't set for a day, again describe the chroniclers of the inhabitants in the South American Andes, how at their place the sun didn't show for twenty hours<ii>.

As a further example from the Greek mythology is mentioned the description of the poet Apollodoros. according to whom Hercules for the solution of his 10th task let the chariot of the sun bring to a standstill. "He turned his vehicle round and raced the way back, dragging along the Pleiaden and all stars, so that the sun set in the East"<iii>. But if all stars take part in the same backwards motion, then this example proves the assumption of the staggering earth's axis.

<i>: Josua 10.13-14 and the in chapter 11.8 cited passages.

<ii>: nach Montesinos, zitiert in Zecharia Sitchin: The Lost Realms (Versunkene Reiche), Knaur Verlag Miinchen 1992, ISBN 3-426-04827-2, S. 203

<iii>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Milnchen 1999, ISBN 3-89631-302-9, Seite 18.

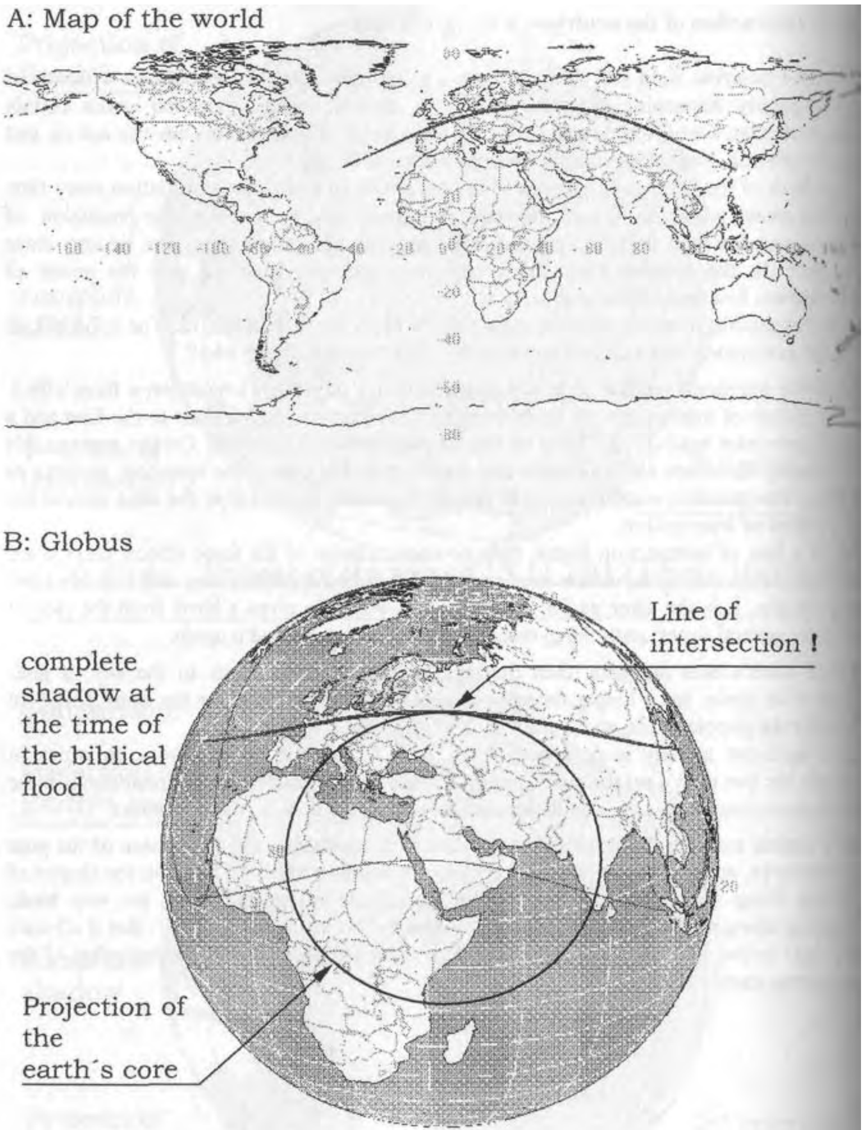
A: Map of the world
60
-160 -140 -120 -100 -80 -60 -40 -20 0 20 40 60 80
-20
-40
-60
-80
B: Globus
Line of intersection !
complete shadow at the time of the biblical flood
20
Projection of the earth´s core

Fig. 20.11: Eclipse of the sun of 27.07.2281 B.C.
(angle of the sun = declination: 19.3°)
The trigger of the biblical Flood?

20.11 Changing of polarity and apocalypse

The pulling at earth's core, which shows as a wobbling of the magnetic axis and in damped form can lead to a tilting of the axis of rotation of the earth, still doesn't make an apocalypse, no end of the world. This only can be expected, if after a tilting of the axis of rotation of earth's core a change of polarity on the surface of the earth occurs. Doing so the new magnetic poles, like a compass needle in the field of the solar wind, again align in such a that the North Pole will be lying "up" in the ecliptic. Because earth's surface will keep it direction of rotation, the sun will, after the process having stabilized, as usual again rise in the East and set in the West. But the inhabitants of the earth, which before still were having midsummer, find themselves again in the midst of midwinter and vice versa.

Truly apocalyptic processes can be expected during the phase of a change of polarity of the earth. Thereby occur unusual relative accelerations and violent earthquakes. The largest destructive potential however is present in the waters of the oceans, which are set in motion.

As is well-known the earth at the equator is measuring a radius which is 21 kilometres larger than at its poles. If only a part of the waters temporarily flows in the direction of the poles of the earth, then the biggest part of the habitable land in Middle and North Europe sinks in the floods; then indeed also the statement of Noah makes sense, who as the first thing saw the mountain of Ararat rise from the floods, after the water again flowed back into its usual ocean basins. The mountain of Ararat after all measures a height of 5137 meters above sealevel!

At comparing historical events with details from the Bible the Flood should have taken place in the year 2245 B.C.<i>. According to the description of the position of the stars Dr. Wild calculates July 2281 B.C. as time for the Flood. The Arabic historical writer al-Makrizi again shifts the event into the year 3094 B.C.<i>. Who is right? We must verify the eclipses of the sun in this time and determine the position of the points of intersection, then we perhaps find the correct answer. Possibly earth's axis has wobbled more than once, have occurred several catastrophes in different regions. At 27.7.2281 B.C. in any case there actually has occurred an extremely critical constellation, whereas the other two years are ruled out. Here no total eclipse of the sun took place (fig. 20.11).

According to the calendar of the Ugha Mongulala 6110 years before this Flood a still much more devastating one should have occurred. That therefore would have been 8391 B.C., while Scott-Elliot dates the catastrophe in the year 9564 B.C.<ii>. According to Plato it would have been about 9500 B.C. For such long periods of time a check however isn't quite easy anymore, because the meantime changes of the earth sum up considerably.

<i>: Wild S. 231, 229 and 225.

<ii>: Wild S, 219, 218 and 210, among others cited from Scott-Elliot: Atlantis und das untergegangene Lemuria, Bauer Verlag Freiburg 1977.

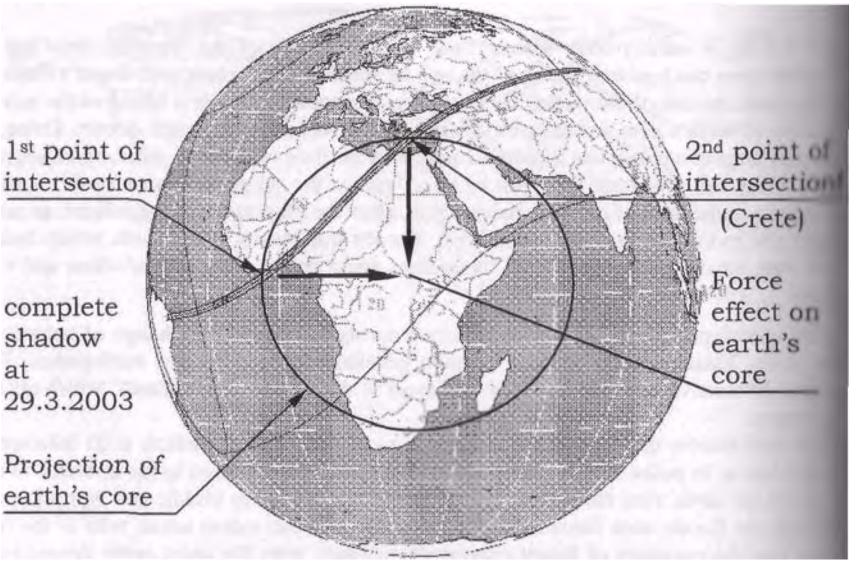
1st point of intersection
2nd point of intersection
(Crete)
complete shadow at 29.3.2003
Force effect on earth's core
Projection of earth's core

Fig. 20.12 A: Eclipse of the sun of 29.03.2006 (decl.: 3.2°)

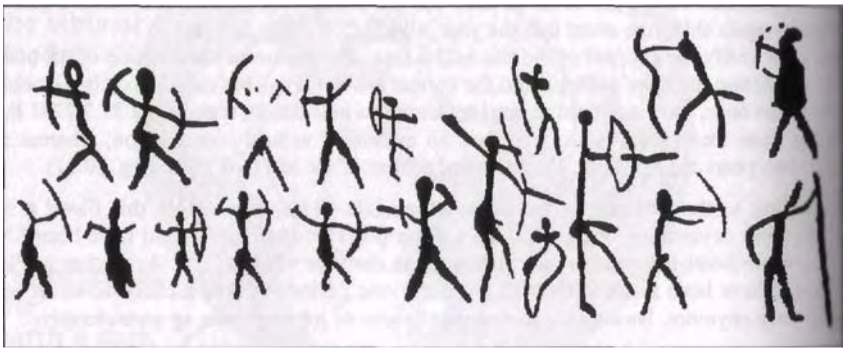

Fig. 20.12 B: Cave painting from Minateda (Spain).<i>

<i>: Laviosa-Zambotti: Origini e Diffusione della Civilta, Marzorati, Milano 1950,

Besides the constellation of sun and moon is crucial the occurring of a relevant neutrino radiation and the question if both is sufficient to tilt the earth and change its polarity. Some experts expect a change of polarity for the time coming, since it takes place with a certain regularity and measured in earth historical periods of time in addition fairly often!
This circumstance the earth presumably owes the river valleys cut deep in the countryside and other topographic phenomena. One presumably can only survive such a catastrophe in an ark (Noah), in the air bubbles of large caves (the walls of which painted children and artists out of pure boredom, see fig. 20.12 B) or in the highland, preferably in the area of the equator (Central Africa, highland of Mexico, Andes, Himalaya).

20.12 Scalar wave gauges for the prediction of earthquakes

We urgently need gauges, to be able to judge the neutrino situation. At first we with that pursue the same goal, as with the building of a free energy converter, however with the difference that the converter should maximize the collected amount of energy, whereas the gauge should minimize the taken up energy, to not too much load the source and not to change the local radiation situation.
In addition the neutrino radiation should be registered distinguishable in its velocity of propagation, what means that the building of a gauge will be very much more costly than that of an energy collector. That's why it can be expected that an usable measurement instrument might be available only many years later and we that long only can base on our own power of observation.

We for instance can statistically analyse, how the earth after an eclipse of the sun reacts to the neutrino radiation focussed by the moon. At August 11 the second focal point was situated in the East, and from that an acceleration of the rotation of earth's core can be predicted. Corresponding observations actually have been made on the surface of the earth with the help of the Foucault pendulum.
The relation the next time can be checked at the eclipse of the sun at 29.3.2006. This time the inverse case is present. A first focussing takes place in the West, so that pendelum swings in the reversed direction would be expected, which indicate a slowing down of the rotation (fig. 20.12 A).

But if the rotation of the earth should change, then the balance sheet of angular momentum of moon, earth's mantle and earth's core isn't correct anymore. As a result a force of difference occurs, which lets the spinning top stagger. But if earth's core staggers inside of earth's mantle, then it powerfully stirs the liquid magma, and in this way releases its surplus energy again. As a result earth's mantle is heated up somewhat. We, on our wafer-thin earth's crust, then time delayed feel the effects of the staggering of the core as an earthquake. The period seems to lie at approximately 6.5 days, as far as this can be read from reactions to the last two eclipses of the sun.

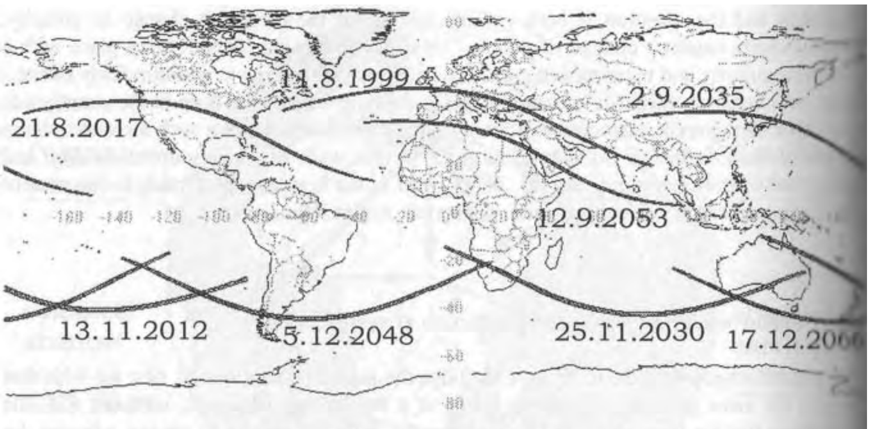
11.8.1999
2.9.2035
21.8.2017
12.9.2053
13.11.2012
5.12.2048
25.11.2030
17.12.206

Tafel 20.13 A: Eclipses of the sun of 2 Saros cycles.

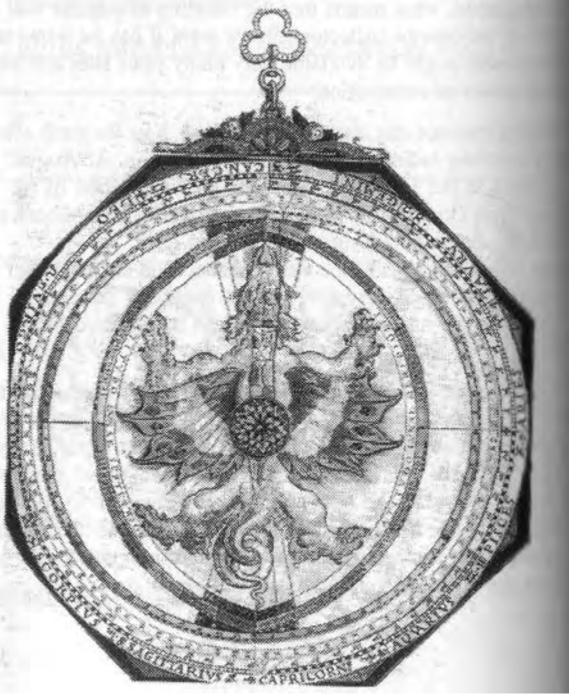

Fig. 20.13 B: The „Astronomicum Caesareum“ served around 1540 the determination of the positions of the moon nodes.<i>

<i>: Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631-302-9, Seite 134.

20.13 Calculation of future occurrences

From the analysis of past and present cosmic events with a strict scientific procedure can be predicted to a certain extent also future events. It thereby by no means concerns prediction, but exclusively the result of an analysis.

After the mentioned eclipse of the sun of 29.03.2006 there once more exists danger of earthquakes (approx. 4.4.2006) for the Island of Crete and Asia Minor. From the eclipse of the sun of 13.11.2012 as well no good can be expected, even if the points of intersection, situated close together, lie far away from populated land in the south-west Pacific basin. The Maya calendar by the way ends to this time. At 21.8.2017, 18 years or a Saros cycle after the eclipse of the sun at 11.8.1999 and correspondingly 120° further to the west, the corresponding complete shadow runs crossways through the USA. The thankless role of Turkey at the second point of intersection this time takes over South Carolina. One only can hope that the houses in Columbia are built more stable than in Izmit.
In fig. 2013 A the two eclipses of the sun of 11.8.1999 and of 13.11.2012 with their respective Saros cycles until 2066 are shown. The as critically to value tendency of the course of the complete shadow is visible, which wants to nestle against the circle of the projection of earth's core to form a line of intersection.

I here break off, since anyone with my indications and an eclipse of the sun-CD can analyse at home all further events personally. I value that if possible many analyses are made and controversially discussed, because possible cosmic catastrophes concern us all somewhat.

What good is the building of gigantic fusion ovens, if the runners by no means have understood the process of the fusion themselves? Why build ring accelerators for billions of dollars, if elementary particles can be calculated at the desk? How is the expenditure for gravitational wave detectors justified, if the actual music plays at entirely other velocities of propagation? Why is half the annual production of the world of Gallium used for an indirect proof of neutrinos, if every self-wound Tesla coil is able to collect more neutrinos?

Our scientists, for whom I as a colleague quite often must be ashamed, have the primary task to draw attention to cosmic and other risks and to calculate them in advance. In any case it is extremely unpardonable to leave this core duty up to some fortune-tellers and self-appointed prophets.
One should more often remind them that the Chinese Kaiser Tschung-Khang let sentence his court astronomers Hi and Ho to death, after the two not having predicted an eclipse of the sun. It thus certainly didn't concern a missed spectacle or a missed tourism business, but presumably the need for safety of the Kaiser and his subordinates, for which the two astrophysicists in ancient China had to take responsibility!

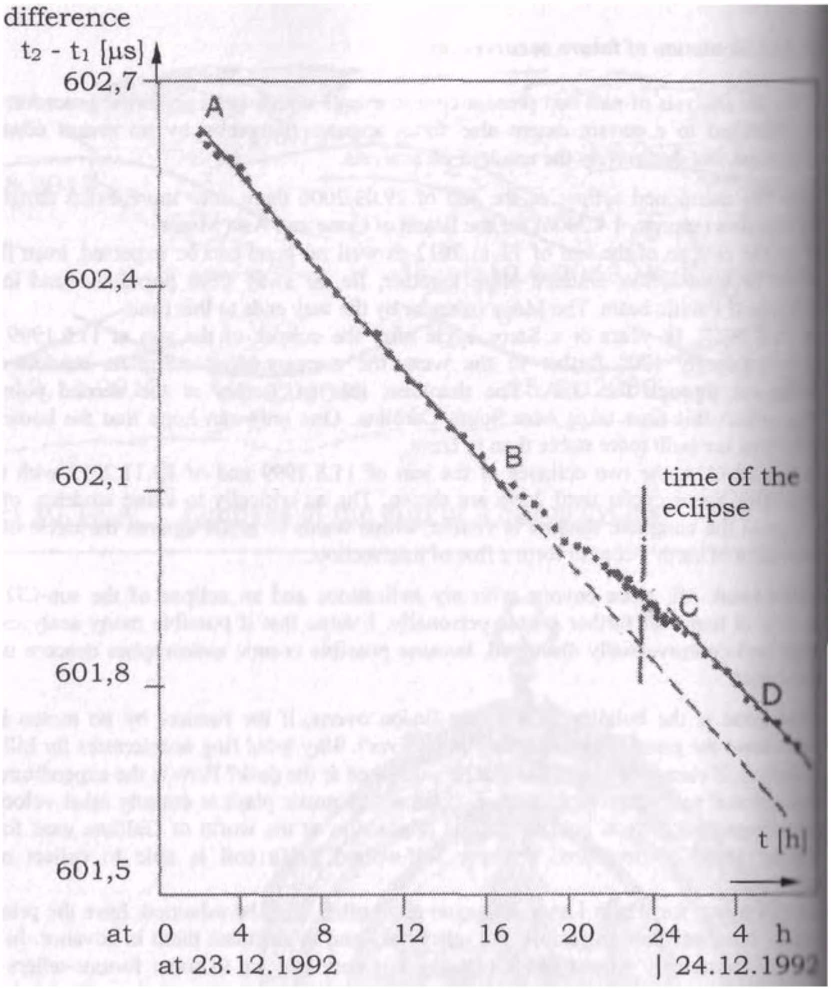

Fig. 20.14: The difference in going of two caesium atomic clocks positioned in different directions, but in the same laboratory in Harbin, China, during the partial eclipse of the sun of 24.12.1992.<i>

<i>: Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54-61. Comments concerning Figure 9: Straight lines AB and CD show that the rate of change of the time difference between the two clocks is constant in non-eclipse periods, but becomes irregular around the time of the eclipse.

20.14 Epilogue to the energy technical seminar

Very slowly the word goes round that the energy source of the future has got a name: "Neutrinopower". Unfortunately fundamental physics, which is financed with public money, still does know almost nothing about the nature of the neutrino radiation. In addition hinder useless model concepts any progress in this direction and so the responsible fundamental research steps on the spot.

Considerably more extensive were already 100 years ago the insights of the experimental physicist Nikola Tesla, the discoverer of the neutrino radiation and father of the free energy. The space energy however is showing most clearly in nature, which only uses this advantageous form of energy. In particular during eclipses of the sun and other cosmic experiments it openly comes to light and can be detected by us for a short time.

The physicist Prof. Shu-wen Zhou of the University of Huazhong in Wuhan, China, systematically has investigated the effects, if sun, moon and earth are aligned<i>. Doing so he has proven inexplicable physical anomalies in experiments. Stimulated by the discoveries of Maurice Allais with the Foucault pendulum he built an arrangement specially flor proving horizontal forces of acceleration, and actually he with that could measure face effects during the total eclipse of the sun of 24.10.1995. He even speaks of an oscillating force!

Further he could determine changes in the spectral wave length of various elements, which under normal conditions go as constant and even as characteristic for the respective element. The relative change of size of the wave length during the ring-like eclipse of the sun over China of 23.9.1987 resulted in the 100-fold value compared to the difference in the spectrum analysis between surface of the earth and surface of the sun! This comparison reveals an immense discrepancy between theory and practice and puts us for a solid problem.
For that six different models of spectrometers were installed in several laboratories of different polytechnics and a photograph was taken of the emission spectra of H, D, Ca, CN, Ni, Ti, etc. Also other reasons than that of an eclipse of the sun could be excluded unambiguosly. In the results of these, at artificial light carried out conventional measurements, in any case abysses yawn. The from spectrum analyses won "insights" about the composition of strange celestial bodies now for sure can be done away with without knowledge about the respective prevailing neutrino radiation.

Spectacular also is the proof of differences in going of atomic clocks of various constructions. During the partial eclipse of the sun of 24.12.1992 seven caesium clocks in four cities of China and in three planes were used. The analysis of the differences in going resulted in, as is shown exemplary in fig. 20.14, changes of the gradient during the eclipse. The results for the atomic clocks in the planes and for two further time measurements turned out with similar clarity<i>.

<i>: Shou-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science 6s Technology, Fall 1999, Vol. 12, No. 3, pp. 54 - 61.

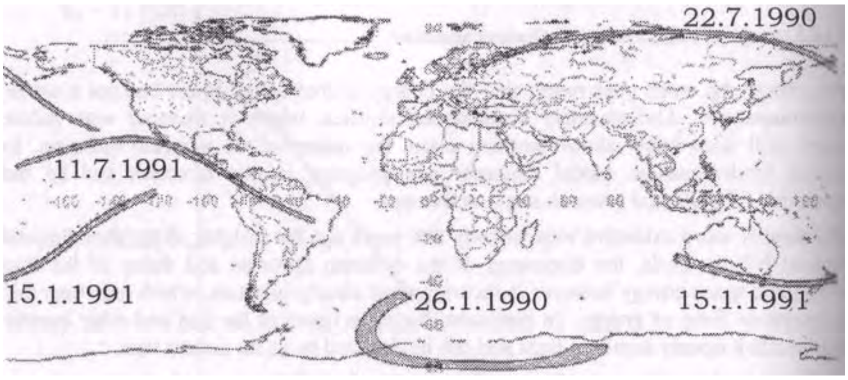

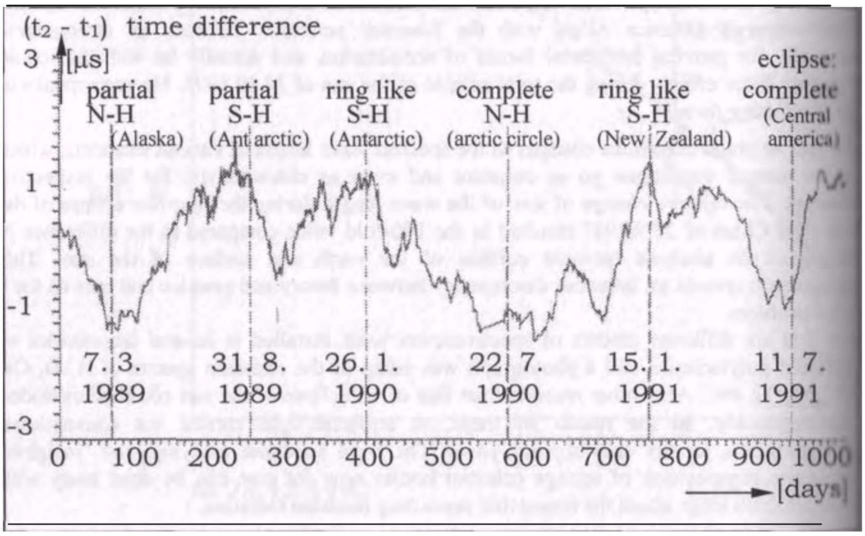

Fig. 20.15: Long-term measurement between 1989 and 991 of the difference in going of two atomic clocks at the U.S. Naval Astronomical Observatory.<i>

The in the 1000 days occurred eclipses of the sun have been entered later. In addition is recorded, if the northern hemisphere (N-H) or the southern hemisphere (S-H) of the earth was involved.

<i>:Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54 - 61. Figure 8.

A connection to the neutrino radiation Prof. Zhon doesn't draw, but it almost is obvious.

From the U.S. Naval Astronomical Observatory (LC/7970) long-term measurements over 1000 days between 1989 and 1991 are present (fig. 20.15). The difference in going between the two atomic clocks positioned at different places shows a permanent up and down. The reason goes as completely unknown.
But if the eclipses, which took place in this time, are entered then one immediately sees the assignment to a maximum or a minimum value. If an eclipse of the sun namely concerned the southern hemisphere of the earth, then the difference in going each time reached a maximum, but if the shadow of the moon run over the northern hemisphere, then each time a minimum occurred. Chance here probably is out of the question!

Let us record: The effect of an eclipse of the sun, to which for instance a Foucault pendulum reacts, can equally be traced back to the interaction of the neutrinos as the free energy. That's why the here presented book carries the title "Free energy and the interaction of the neutrinos". It has appeared in the series concerning the "Electromagnetic environmental compatibility", and also for that there are good reasons. Our energy technology must become more ecologically compatible and we come the goal closer, if we emulate nature, understand and copy it.

That it further concerns electromagnetism, likewise is explained from the interaction of the neutrinos, which concerns the oscillating and resonant case of the electromagnetic interaction. In this respect the reference to the series of books therefore would be given.

It only indirectly has to do with "electrosmog", or what otherwise is understood under environmental compatibility in general. It however could be shown that also the earth radiation is a form of neutrino radiation, and that it poses a biological effectiveness, after it turned out that it serves nature as an energy source. With that the conflict with space energy devices already is predetermined if the same scalar wave radiation should be used. Then one system takes away the other system the energy basis and the existence basis.
This environmental compatibility problem can be solved, if care is taken that both don't get in each others way with regard to the frequency and the wave length. We for that need a deep understanding concerning the topic of the space quanta and the neutrinos, their physical properties and the corresponding device technology. The book should make a contribution to that.

The other side of the medal is the information technical aspect of the scalar wave radiation and the environmental compatibility problem connected with that. The third and last part of the series of books is dedicated to this theme.

Part 2: Edition belonging to the energy technical seminar: "Electromagnetic Environmental Compatibility"

Prof. Dr.-Ing. Konstantin Meyl

Scalar waves

Abstract: 1. Auflage 1998, 4. Auflage and 1st English Edition

With regard to the environmental compatibility a decentralized electrical energy technology should be required, which manages without overhead power lines, without combustion and without radioactive waste. The liberalization of the energy markets won't on any account solve our energy problem, but only accelerate the way into the dead end. New, ecologically compatible concepts are collected and discussed in the book.

A useful energy source could be represented by space quanta, which hit upon the Earth from the sun or from space. They however only are revealed to the measurement technician, if they interact. It will be shown that the particles oscillate and an interaction or collection with the goal of the energy technical use only is possible in the case of resonance.

Since these space quanta as oscillating particles have almost no charge and mass averaged over time, they have the ability of penetration proven for neutrinos.
In the case of the particle radiation discovered 100 years ago by Tesla, it obviously concerns neutrinos. We proceed from the assumption that in the future decentral neutrino converters will solve the current energy problem. Numerous concepts from nature and engineering, like on the one hand lightning or photosynthesis and on the other hand the railgun or the Tesla converter are instanced and discussed.

Free energy and the interaction of the neutrino radiation

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen
www.k-meyl.de meyl@k-meyl.de

Scalar Waves

From an extended vortex and field theory to a technical, biological
and historical use of longitudinal waves.

Part 3

by

Professor Dr.-Ing. Konstantin Meyl

Edition belonging to the lecture and seminar "Electromagnetic environmental compatibility"

(Original title: "Elektromagnetische Umweltvertraglichkeit")

Translated out of the German language
by Ben Jansen
(2003)

* * *

Part 3: Edition belonging to the information technical seminar

Scalar waves and the technical use of longitudinal waves and vortices

* * *

INDEL GmbH, Verlagsabteilung

Villingen-Schwenningen 2003

ISBN 3-9802 542-4-0

Preface to the seminar, part 3

With the appearing in print of the 3rd part of the series of lecture notes of the electromagnetic environmental compatibility, the collection of material can be considered to be completed for the time being. By now almost all aspects concerning the theme of the potential vortices of the electromagnetic field and their propagation as a scalar wave should have been addressed or at least indicated.

And that's what it is about in the case of the editions belonging to the seminars: In the seminars, taking place inside and outside the university, questions about the phenomena and the biological effectiveness of the fields, which surround us, are discussed. The scientifically correct procedure asks as a preparation for the seminar a voluminous collection of material and a compilation of theses with the pros and contras to them. The argumentation, which is denoted as „textbook physics" and is generally accepted, the participant of the seminar can work out himself by means of the available journals and books. The potential vortex, which completed the field equations, however makes possible a counterstatement, which for most participants at first will be unusual. The old dispute between quantum and field physics rekindles. What is missing is a voluminous collection of material concerning the refined field physics, and this gap the editions belonging to the seminars should close.

In the case of this book the primary concern is to make sure, that all facets, properties and physical consequences of the new and unified field physics are addressed. The question, if all points can bear a strict verification, remains reserved to the seminar and its participants, of whom each should form its own judgement. The series of books should be understood as a stimulation, to make own thoughts about the various points. Who has no opportunity to participate in a seminar, by means of the lecture at least gets an idea of what is discussed here and how is struggled for scientific insights.

The herewith completed collection of material by no means may be compared or confused with a basic work about the theory of objectivity. That the editions cannot and want not afford at all. They together with the discussions at best form the basis for a scientifically basic work, at which I am working.

In the case of the collection of material concerning the electromagnetic environmental compatibility consisting of three parts, again and again new aspects have emerged, which found entrance into the book, at which I was working at that particular time, whereas the rough structure has been controlled precisely: part 1 with the chapters 1 to 9 treats the basics, part 2 with chapters 10 to 20 the energy technical aspect and part 3 with the chapters 21 to 30 the information technical aspect of scalar waves.

The here presented 3rd part starts with the wave equation and the two comprised parts of a transverse and a longitudinal wave. The historic dispute between Heinrich Hertz and Nikola Tesla with the experimental evidence of the each time used wave part is continued over the wave-particle uncertainty compromise up to the dispute about the right field description: that of Maxwell or the new dual and at the same time unified description, which builds upon Boscovich and Faraday. The aspects, which seem so irreconcilably thereby only are two parts of a single equation, which is much older and can be traced back to Laplace: the wave equation.

Until now no derivations of this equation, which contains both wave parts, are known This for the first time succeeds from the new and extended field approach in the summary from chapter 26. At first however is attempted to make the world of the scalar waves plausible with models, calculations and observations. For that are used examples from high-frequency engineering, like the ground wave or the near-field area of an antenna, and from nature and medicine, which are granted two chapters. The comparison of the nerve conduction with Tesla's one wire technology is bringing it to light that scalar waves equally are used in both cases. In a frequency diagram the insights are entered concisely and is shown that considerably more unexplored domains occur than already is known, that in most cases not even gauges are available. With the design for building a scalar wave gauge the scientific verification of until now only empirically won results is stimulated, like for instance the results the radiesthesia wants to have determined with the help of the sensitivity of man as a biosensor.

Many relations will be revealed to the reader only, if he has worked through the summary, which follows. From Maxwell's field equations only the well-known (transverse) Hertzian waves can be derived. (Longitudinal) scalar waves however in the result are zero. This is a flaw of normally used field theory, since scalar waves exist for all particle waves, like e.g. as plasma wave, as photon- or neutrino radiation. Because the field pointer, in the direction of which longitudinal waves are propagating, is oscillating, the frequency will oscillate like the velocity of propagation, which again is measured as a noise signal. Any antenna noise proves the emission of scalar waves in space. But scalar waves, or whatever should be subsumed under the by mathematics minted generic term free of any value judgement, surely are more than only noise.

Starting from Faraday's discovery - instead of the formulation of the law of induction according to Maxwell - an extended field theory is derived, which goes beyond the Maxwell theory with the description of potential vortices (noise vortices) and their propagation as a scalar wave, but contains the Maxwell theory as a special case. The new field theory with that doesn't collide with the textbook opinion, but extends it in an essential point with the discovery and addition of the potential vortices.
Also the theory of objectivity, which follows from the discovery, is compared in the form of a summary with the subjective and the relativistic point of view and the consequences for variable velocity of propagation of the scalar waves, formed from potential vortices, are discussed. Like already in part 1 the unification of the electromagnetic interaction with the gravitation succeeds impressively.

Besides the mathematical calculations this book contains a voluminous material collection concerning the information technical use of scalar waves, if e.g. the useful signal and the usually interfering noise signal change their places, if a separate modulation of frequency and wavelength makes a parallel image transmission possible, if it concerns questions of the environmental compatibility for the sake of humanity (bioresonance, among others) or to harm humanity (electrosmog). With that the book again finds back to the starting point, to the open task, which made necessary an excursion through all domains of physics to answer it. I hope, the long march was worthwhile.

Villingen-Schwenningen december 2002

Preface to the 2nd edition of part 3

In favour of chapter 30 as a complement, an introduction into the ancient broadcasting technology of the gods, the 1st edition wasn't printed anymore in the year 2002 as originally planned. Instead the individual chapters have been prepublished in different places. The chapters 21 to 25 are found among others in the book "scalar wave technology", which has been published as an instruction for an experiment to demonstrate scalar waves in 2001. Excerpts from it, as also from the chapters 26 to 28 have appeared in form of individual essays in different journals.

Again experimental successes had been intervening, which caused a renewed deferring of the date of appearance of the 3rd part of the series of books. In my laboratory the bi-directional transmission of music and of data by scalar wave was successful. We thus also can transmit information backwards from a receiver to the transmitter or to a second receiver, whereby the receivers work purely passive, thus without own power supply. The operating energy is as well supplied to them by the transmitter by scalar wave. The demonstrated technology opens completely new possibilities of a technical application. Conceivable are telemetry installations where measurement signals have to be transmitted from rotating or otherwise inaccessible places of a machine. The energy for the measurement electronics can be transmitted wirelessly by scalar waves and the signal can be sent back along at the same time, by modulating it onto the energy carrier. In this way dozens of measurement stations can be connected wireless with a single central transmitter, which supplies them all with energy.

The question was asked: Is the technology really entirely new? The answer is amazing: No, it here concerns the oldest technology of humanity, which had developed to a peak in antiquity, to send receive engineering of the gods. For this claim even a mathematical proof is available. For that the authoritative transition, the unrolling of a vortex to a wave (the transition from the near-field to the far-field) or in the reversed case the rolling up at a receiver antenna (usually denoted as standing wave) is calculated with help of the extended field theory. The result is, that at this transition the velocities of propagation resp. the wavelengths of the transverse and the longitudinal wave stand to each other in the ratio of the Golden Proportion.
With regard to the optimization of a transmitter or receiver antenna the Golden Proportion has an effect on the construction resp. the architecture of corresponding buildings . Which ones, with that deals the complementing chapter 30. But it mustn't be missing either, since after all it concerns a grandiose practical information technical use of scalar waves, from which we technologically can learn a lot<i,ii>.

Villingen-Schwenningen march 2003

<i>: Note: chapter 30 provides an introduction into ancient broadcasting technology of the gods. The working off of history with respect to the use of scalar waves is so voluminous, that for that a book of its own is published with the title:

<ii>: K. Meyl: Sendetechnik der Gotter, historischer Sciencefictionroman (2003), Villingen-Schwenningen, INDEL Verlagsabteilung, ISBN 3-98O2 542-5 9

21. Derivation of the scalar wave

WilliamThomson, who called himself Lord Kelvin, after he had been knighted, already in his lifetime was a recognized and famous theoretical physicist. The airship seemed him too unsafe and so he went aboard a steam liner for a journey from England to America in the summer of 1897. He was on the way in a delicate mission.

Eight years before his German colleague Heinrich Hertz had detected the electromagnetic wave in experiments in Karlsruhe and scientists all over the world had rebuilt his antenna arrangements. They all not only found confirmed the wave as such, they even could show the characteristic properties.
It was a transverse wave, for which the electric and the magnetic field pointers oscillate perpendicular to the direction of propagation. This can be seen as the reason, that the velocity of propagation is showing itself field independent and constant. It is the speed of light c.

With that Hertz had experimentally proven the properties of this wave, previously calculated in a theoretical way by Maxwell, and at the same time proven the correctness of the Maxwellian field theory. The scientists in Europe were just saying to each other: "well done!" as completely other words came across from a private research laboratory in New York: "Heinrich Hertz is mistaken, it by no means is a transverse wave but a longitudinal wave!"

Such a screwball most said and did as if they hadn't heard the criticism at all. But then one couldn't ignore it completely, because on the one hand claimed the private research scientist to have experimental proof and on the other hand it wasn't just anybody, who here reported. It was nobody but Nikola Tesla, the Croat experimental physicist who emigrated to America
Him we owe the modern alternating current technology from the high-tension network for energy transmission over the alternating current transformer to the asynchronous machine.
With his magnificent inventions he had earned enough money, to be able to afford a private laboratory, in which he could research and invent uncensored and free. The key to his success was lying in his purposeful, concentrated and efficient working method.
He was fast! Whereas in Europe still was being discussed about properties and theoretical possibilities of application of the wave, Tesla already presented the armed forces a remote controlled scaled-down submarine in Madison Square Garden (fig. 21.1 A). To convince such a man, who only holds valid what his experiments reveal, from the opposite, should be a hopeless enterprise.
Lord Kelvin was aware of that, as he made the decision to go on the journey. He could not and did not want to put his head into the ground, as many of his colleagues, because on the one hand scientists are curious by principle and on the other hand he travelled as an official representative of science. He had been instructed to free, as Mr. Clean, the undamaged world of sciences from erroneous teachings. But it came completely different.

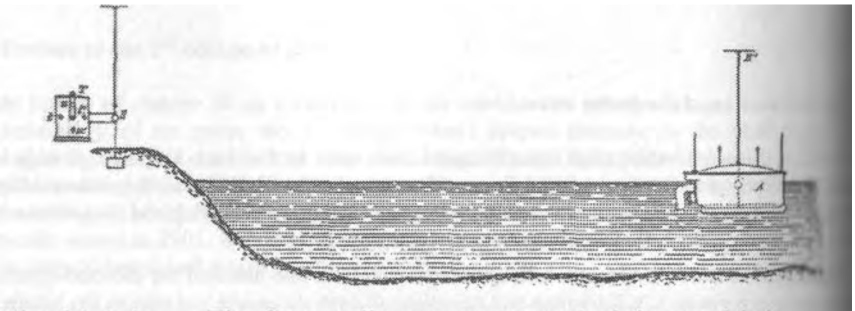

Fig. 21.1 A: The first radio remote control of the world.<i>

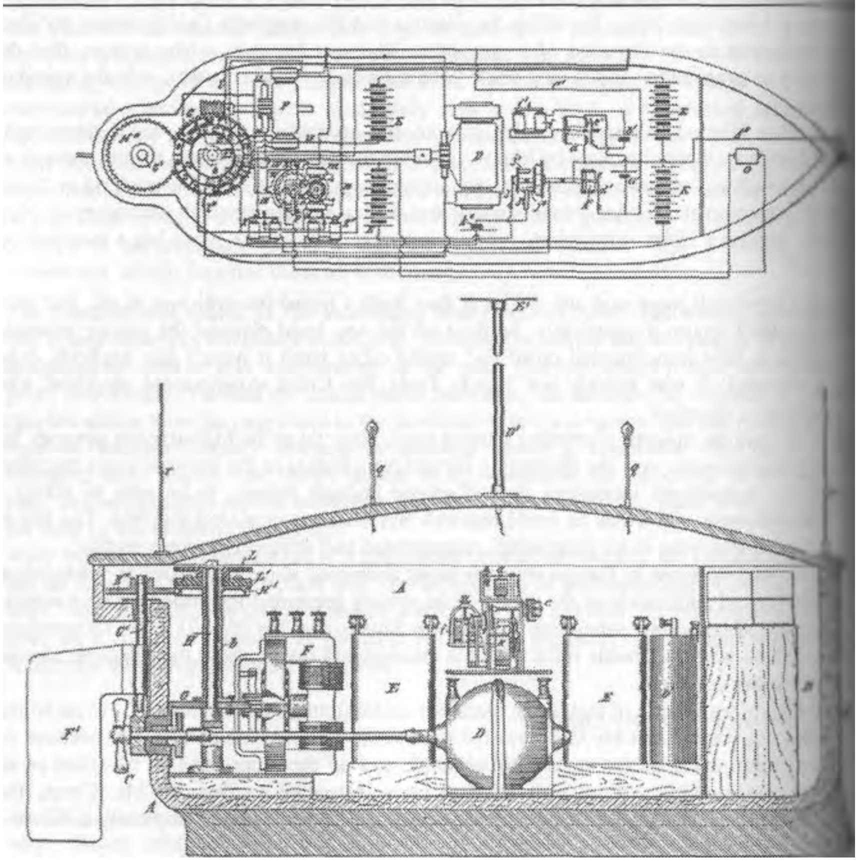

Fig. 21.1 B: Patent specification concerning the remote
controlled submarine.<i>

<i>: Nikola Tesla: Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles, US-Pat. 1898, Nr. 613,809. Complete Patents: P. 351

21.1 Lord Kelvin in a delicate mission

The first day of his visit at Tesla Kelvin spoke the admonishing words he had been instructed to speak. He recommended Tesla warmly to publicly retract the remarks concerning the Hertzian mistake and to contribute himself to the settlement of the dispute. Seen from a specialist viewpoint they talked at cross-purposes.

But at night in his hotel room the Lord again thought about the experiments, which had been shown to him. The standing wave nature had been visible unambiguously: the oscillation nodes, the effect back on the transmitter, the high degree of effectiveness and many other things more. Such properties the Hertzian wave indeed doesn't know. Also Tesla didn't work with dipole antennas, but with flat coils, with spark gap oscillators and with a very unconventional switching technique, set up different in principle.

The next morning Lord Kelvin appeared in the laboratory again and greeted Tesla with the words: "Then you don't use Hertzian waves?" "Certainly not", Tesla answered," it are radiations. By waves no energy could be economically transmitted over a larger distance. My system works with true conduction, which theoretically seen can take place over a larger distance, without bigger losses occurring."
In the article of the "Electrical Experimenter" is noted further, that the doubting critic Kelvin suddenly turned into one of the biggest followers<i>.

Kelvin deduced very fast: according to that there exist two different sorts of wave propagation. So Hertz with his transverse wave is just as right, as Tesla with the longitudinal wave.
As a represantative of theoretical physics he however could pull out Tesla a decayed tooth. Maxwell had based his field description on an aether concept, which at that time in the world of sciences triggered violent discussions. Since Tesla saw such an aether as a prerequisite for longitudinal waves, he thought he and not Hertz had proven the Maxwell wave in experiment for the first time. By stating this the magnificent experimental physicist however revealed weaknesses in the area of theory. Maybe he had not read exact enough or understood the books of Maxwell, which without doubt were formulated mathematically only arduously comprehensible in the original version.
In this point Tesla had to learn otherwise by Kelvin. Maxwell's field theory provides without exception a mathematical description for the Hertzian wave. For the Tesla radiation however no field description exists! This by the way is the circumstance, why this wave could disappear from the textbooks and again fall into oblivion. Tesla himself had problems to theoretically imagine, what happens at his wave. His models in some points perhaps were even better than the official textbook opinion, but not without contradiction to accepted regularities. That's why Tesla did without a publication of his ideas, although he in his lifetime had filed away at an own theory.

<i>: Nikola Tesla: Famous Scientific Illusions, III. The Singular Misconception of the Wireless, Electrical Experimenter, Feb. 1919, printed in Tesla Said, p. 197

Figure 21.2 A:

William Thomson, 1846 becoming Professor at Glasgow University with the age of 22 years. From 1892. on he was called Lord Kelvin.<i>

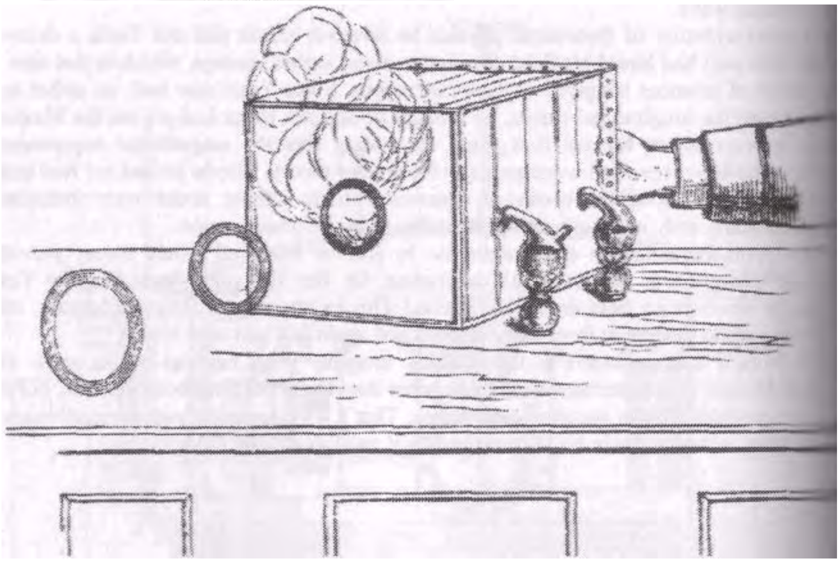

Fig. 21.2 B: Vortex rings from a smoke vortex gun.<i>

<i>: David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, 1990

21.2 Helmholtzian ring-like vortices in the aether

Tesla told Kelvin at his visit from the meeting with the German Professor Hermann von Helmholtz on the occasion of the World's Fair in Chicago 1893. Kelvin knew him very well and had cooperated with him in the past. Now the vortex concept of his colleague and his model of stable vortex rings were very obliging.
In the case of a standing wave the impulse is passed on from one particle to the next. In the case of acoustics for instance we are dealing with a shock wave, where one air molecule knocks the next. In this way sound propagates as a longitudinal wave. Correspondingly the question is raised: What sort of quanta are the ones, which in the case of the Tesla radiation carry the impulse?

Lord Kelvin was already on the way back to Europe on the steamship and he deduced: The Tesla experiments prove the existence of longitudinal standing waves in space. In the question, what passes on the impulse, Kelvin comes to the conclusion: it are vortices in the aether! With that he had found an answer in experience. With his students he built boxes, with which he could produce smoke rings, to be able to study and demonstrate in experiments the special properties of ring-like vortices in their flow technical analogy (fig. 21.2<i>). But he didn't have ready a suitable field theory.

The from Germany to the Isles exported vortex physics for a short time could establish in England, before it was slaughtered and buried by the German quantum physicists. A main advocate has been James Clerk Maxwell, who held the vortex theory for the best and most convincing description of matter<ii>. As his successor at the Cavendish laboratory in Cambridge J. J. Thomson was appointed, who already as a young man had got a price for a mathematical treatise about vortices<iii>. He discovered the electron and imagined it, how could it be otherwise, as a field vortex.

The crucial weakness of vortex physics, the lacking of an usable field theory, was of benefit to the emerging quantum physics. This could change fundamentally with the discovery of the potential vortex, the vortex of the electric field<i4>!
In addition is the experimental proof of a vortex transmission as a longitudinal wave in air or in a vacuum, as it has been furnished by Tesla already 100 years ago, neither with Maxwell's theory nor with the today normally used quantum theory explicable or compatible. An urgent need is present for a new field theory!

<i>: David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, England 1990.

<ii>: James Clerk Maxwell: "...the vortex rings of Helmholtz, which Thomson imagines as the true form of the atom, fulfil more conditions than any other previous concept of the atom."

<iii>: J.J. Thomson: "the vortex theory is of much more fundamental nature than the usual theory of solid particles."

<i4>: Konstantin Meyl: Potentialwirbel Band 1 (1990) and Band 2 (1992), INDEL-Verlag, Villingen-Schwenningen.

Scalar waves, a mathematical reasoning (I)

1. prerequisite: wave equation (textbook-formulation)

$$\boxed{\Delta \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}}$$

taking apart of the delta operator mathematically (Laplace operator) according to the rules of vector analysis (fig. 5.0):

$$\boxed{\begin{array}{lllll} \Delta \mathbf{E} & = & \text{grad div } \mathbf{E} & - & \text{rot rot } \mathbf{E} \\ \text{wave} & = & \text{longitudinal} & + & \text{transverse} \end{array}} \quad (21.1)$$

---

2. State of the art of technology: Hertzian wave = transverse wave
special case: = solution of Maxwell's field equations
no sources:

$$\boxed{\text{div } \mathbf{E} = 0} \quad \text{and} \quad \boxed{-\text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.3)$$

transverse wave:
field pointers oscillate crosswise to the direction of propagation
The propagation occurs with the speed of light c.

---

3. claim: Tesla radiation = longitudinal wave
special case:
irrotationality:

$$\boxed{\text{rot } \mathbf{E} = 0} \quad \text{and} \quad \boxed{\text{grad div } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.4)$$

longitudinal wave, shock wave, standing wave:
field pointer oscillates in the direction of propagation.
Velocity of propagation is variable!

Fig. 21.3: The special cases of the wave equation

21.3 Taking apart of the wave equation

Before one plunges into the adventure of an entirely new field theory, it first of all should be traced and analysed, what the latest textbooks say about scalar waves.

There some scalar or vector potentials are introduced; there the constant of dielectricity e is written down as a complex variable, although it physically seen concerns a material constant, only to be able to calculate with this trick artificially a loss angle, which should indicate the losses occurring in a dielectric, where in reality it concerns vortex losses. Of course one can explain the dielectric losses of a capacitor or the heating in a microwave oven entirely without vortex physics with such a label fraud, but it should be clear to anyone, that in a complex constant lies buried an inner contradiction, which is incompatible with physical concepts.
We are used to such auxiliary descriptions so much, that the majority of today's physicists tend to attribute physical reality to this mathematical nonsense. As pragmatists they put themselves on the standpoint if with that experimental results can be described, then such an auxiliary description can't be so wrong after all. Doing so the circumstance is forgotten that here the ground of pure science is abandoned and is replaced by creeds.

We find everything already in the wave equation, as it can be found in all textbooks.

$$\boxed{\Delta \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \quad (21.1)$$

Behind this formulation two completely different kinds of waves are hiding, because the used delta operator consists of two parts according to the rules of vector analysis:

$$\boxed{\begin{array}{lll} \text{grad div } \mathbf{E} & - \text{ rot rot } \mathbf{E} & = \Delta \mathbf{E} \\ \text{longitudinal} & / \text{ transverse} & \text{wave} \end{array}} \quad (21.2)$$

We want to discuss two special cases.
If we put the left part (in eq. 21.2) to zero (div E = 0) which is tantamount to no sources of the field then the well-known radio wave remains, which also is called Hertzian wave, after Heinrich Hertz, as said, had experimentally detected it in Karlsruhe 1888:

div E = 0 and $\boxed{- \text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}}$ (special case) (21.3)

It concerns the transverse wave, described by Maxwell, for which the field pointers oscillate crosswise to the direction of propagation. The propagation again occurs with the speed of light c. So much concerning the state of the art of technology.
But as we see, in the mathematical formulation of the wave equation is hiding, yes, even more than only the generally known electromagnetic wave. The no sources approach is a neglect, which only is valid under certain prerequisites!

Scalar waves, a mathematical reasoning (II)

3. claim: Tesla radiation = longitudinal wave

$$\boxed{\operatorname{rot} \mathbf{E} = 0} \quad \text{and} \quad \boxed{\operatorname{grad} \operatorname{div} \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}} \qquad (21.4)$$

special case:
source field, because div E ≠ 0

=> sources = charge carriers (plasma waves)
=> sources = vortex structures

---

4. approach: (div E ≠ 0) is a scalar! => scalar wave
=> E-field vector can be derived from a scalar potential φ:

$$\boxed{\mathbf{E} = -\operatorname{grad} \varphi} \qquad (21.5)$$

and $$\operatorname{div} \mathbf{E} = -\operatorname{div} \operatorname{grad} \varphi \equiv -\Delta \varphi \qquad (21.6)$$

inserted in eq. 21.4: homogeneous scalar wave equation

$$\boxed{\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2}} \qquad (21.7)$$

---

5. proof: For the case of an additional space charge density ρ_{el} should be considered: $\operatorname{div} D = \rho_{el}$
and $\operatorname{div} E = \rho_{el}/\varepsilon$ (21.8)

inhomogeneous scalar wave equation = plasma wave!

$$\boxed{\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon}} \qquad (21.9)$$

Fig. 21.4: Derivation of the plasma wave as an example of a scalar wave

21.4 Derivation of the scalar wave equation

Making neglects by all means is normal and legal in science. But it may not be carried out at will and untested. In any case an error consideration is necessary, the result of which should be, that the neglect indeed is sufficiently small.
In the here presented case of the wave equation I haven't found one single textbook, in which this error consideration has been carried out. As a result of this inexcusable negligence there is the danger that exactly the aspect is neglected, which it actually concerns. This could lead to catastrophical results, e.g. that the causes for electrosmog, for numerous EMC-problems, for biological and technical effects aren't seen and understood anymore, that pure science once more dilapidates to a creed!

In the case of the wave equation the assumption of no sources describes only one side of the medal. The other side, which for the electromagnetic wave occurs as an error term, we get if we this time put the right part in equation 21.2 to zero (rot E = 0).
In this case a divergence of the field is present, which requires a source field. As sources some charge carriers, quanta or at least particle structures, e.g. vortex structures have to exist. Their propagation occurs, as we know it from the sound propagation, as shock wave in longitudinal manner. The air molecules, the quanta or particle structures thereby oscillate in the direction of propagation. Also the field pointer has a component in this direction

$$\mathrm{rot}\,\mathbf{E} = 0 \quad \text{and} \quad \boxed{\mathrm{grad}\,\mathrm{div}\,\mathbf{E} = \frac{1}{c^2}\cdot\frac{\delta^2\mathbf{E}}{\delta t^2}} \quad \text{(special case)} \qquad (21.4)$$

The occurring divergence of the field pointer (div E) is a scalar, for which reason this wave preferably is called scalar wave.

In the special case of a scalar wave (rot E = 0) the E-field vector can be derived from a scalar potential φ:

$$\boxed{\mathbf{E} = -\,\mathrm{grad}\,\varphi} \qquad (21.5)$$

On the one hand this term is used in the wave equation 21.4 on the right-hand side without forming the gradient on both sides of the equation. On the other hand the divergence of the approach 21.5:

$$\mathrm{div}\,\mathbf{E} = -\,\mathrm{div}\,\mathrm{grad}\,\varphi \equiv -\,\Delta\varphi \qquad (21.6)$$

is applied in equation 21.4. The result is the homogeneous scalar wave equation:

$$\boxed{\Delta\varphi = \frac{1}{c^2}\cdot\frac{\delta^2\varphi}{\delta t^2}} \qquad (21.7)$$

Scalar waves, a mathematical reasoning (III)

Proof: plasma wave =
inhomogeneous scalar wave equation

$$\boxed{\Delta \varphi = \frac{1}{c^2} \cdot \frac{\delta^2 \varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon}} \quad (21.9)$$

one solution<i>: $\omega^2 = c^2 \cdot k^2 + \omega_p^2$ (= Langmuir waves).

Fig. 21.5 A: Derivation of the plasma wave as an example of the existence of scalar waves in the wave equation

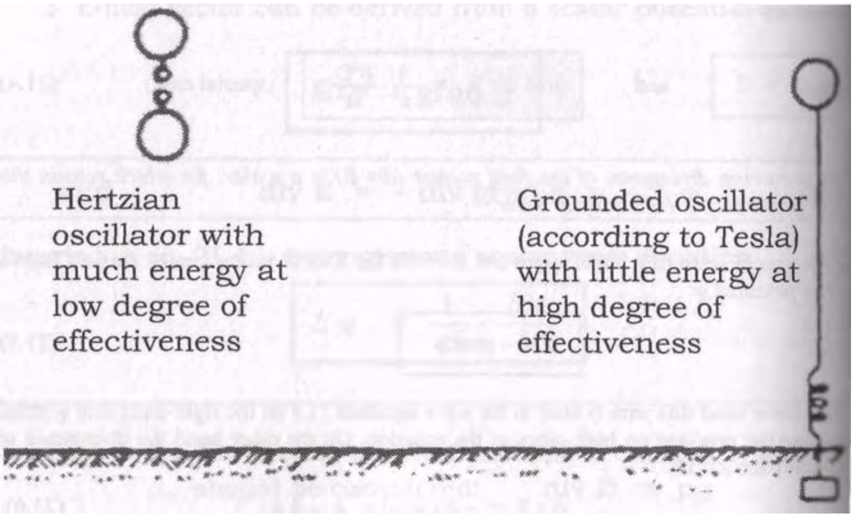
Hertzian oscillator with much energy at low degree of effectiveness
Grounded oscillator (according to Tesla) with little energy at high degree of effectiveness

Fig. 21.5 B: Nikola Tesla explains the difference between his radiation and the Hertzian wave.<ii>

<i>: The solution describes dispersion relations of plasma waves; longitudinal wave movements + Langmuir oscillations of the electron density.

<ii>: Nikola Tesla: The Problem of Increasing Human Energy, The Century Monthly Magazine, June 1900, ISBN 1-882137-00-0, Page i-15

21.5 Derivation of a plasma wave

Doing without formation of the gradient for the derivation of the homogeneous wave equation is tantamount to an integration of the equation. We hence under certain conditions must expect the occurring of an integration constant.
This is the case, if in addition a space charge density ρ_{el} occurs as source of the field, which according to Maxwell equation 4 can be considered as the divergence of a dielectric displacement D (fig. 21.4): $\operatorname{div} \mathbf{D} = \rho_{el}$,
resp. with the relation of material $\mathbf{D} = \varepsilon \cdot \mathbf{E}$: $\operatorname{div} \mathbf{E} = \rho_{el}/\varepsilon \equiv -\Delta\varphi.$ (21.8)

If we complete this contribution with possible present field sources, then the inhomogeneous scalar wave equation results:

$$\boxed{\Delta\varphi = \frac{1}{c^2}\cdot\frac{\delta^2\varphi}{\delta t^2} - \frac{\rho_{el}}{\varepsilon}} \quad (21.9)$$

For these equations solutions have been published<i>. They have the same form, as the well-known dispersion relations of Langmuir waves. That is electron plasma waves, thus longitudinal wave movements associated with Langmuir oscillations of the electron density.

With that it has been proven that scalar waves and longitudinally propagating standing waves are described by the wave equation and are contained in it. This in any case is valid in general just as in the special case of a plasma wave, as mathematically could be derived here.

From the example of the derivation of plasma waves from the wave equation 21.1, we see that scalar waves by all means are known and their existence isn't casted doubt on at all. After all the mathematically won solution is secured by numerous experiments. Why do textbooks concerning high-frequency engineering then ignore the scalar wave parts in the wave equation?
Our specialists seem to concentrate so much on their branch, that they are losing the view on the Big Whole. They practice one-eyed physics, where the plasma physicist keeps one eye shut and the radio technician the other eye. What the other one does and researches, they don't understand anymore for ages. It is necessary to point them to their common root.

The perhaps most important statement of the wave equation is that every emitted wave contains both longitudinal and transverse parts! Both parts in addition occur jointly, so that for corresponding boundary conditions it can be expected that one part is transforming into the other part. The HF technician then suddenly measures less field strength and comes to the conclusion that his radio wave has been damped or partly absorbed. Doing so heat is created, he says, although the wave equation by no means contains a corresponding term for the necessary thermodynamic description. He simply hasn't understood the wave equation!
Absorption means nothing but transverse waves in the case of a disturbance rolling up to vortices, to become a scalar wave in this way (fig. 1.4 and 5.3). With that they are evading every field strength measurement and what can't be measured doesn't exist in one-eyed physics! Therefore can't exist, what shouldn't exist.

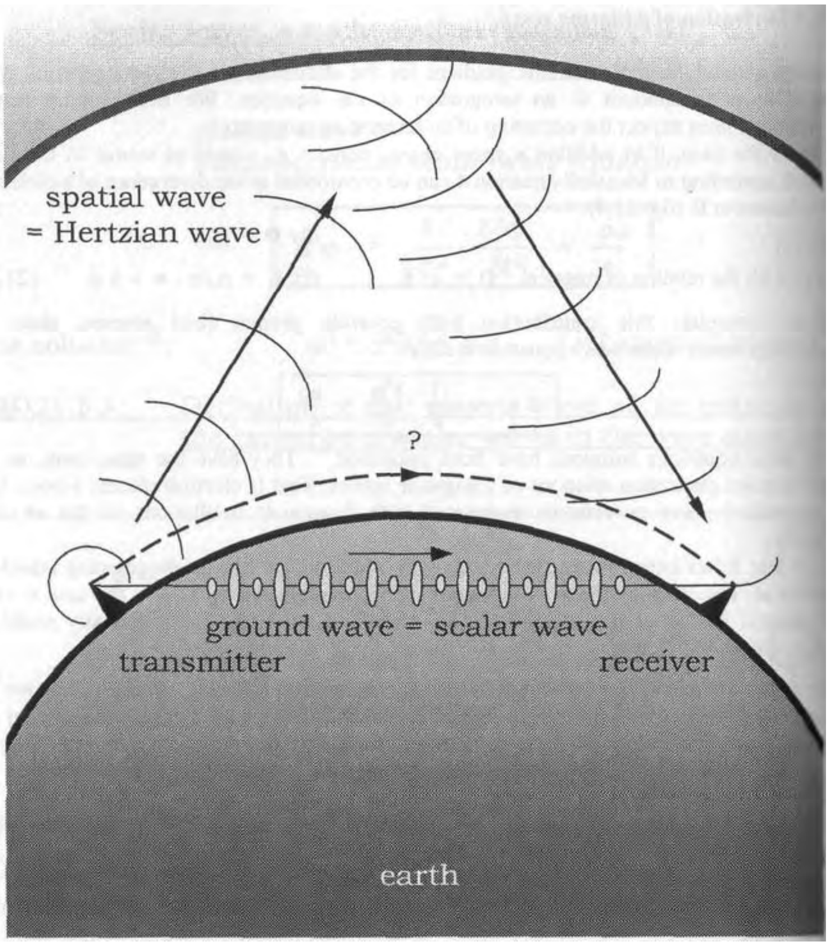
spatial wave
= Hertzian wave
?
ground wave = scalar wave
transmitter
receiver
earth

Fig. 21.6: Ground waves, which follow the curvature of the earth and radio waves reflected at the ionosphere.<i>

<i>: Meinke, Gundlach: Hochfrequenztechnik, 4.Aufl. Springer-Verlag Berlin 1986, Seite R 18: „Lang-, Mittel- and Kurzwelle breiten sich einerseits entlang der Erdoberflache als Bodenwellen, andererseits unter Mitwirkung der Ionosphare als Raumwellen aus."

21.6 Mistakes of the HF technology

The devastating judgement of Tesla, Hertz was mistaken, was rash. His claim of having detected the Maxwell wave himself, proved to be untenable. With this claim in particular Tesla hud the scientific world against him. If one opens encyclopedias or textbooks then one gets the impression, until the day of today science still hasn't forgiven Tesla, what once more shows how persistent prejudices are kept.

Just as little does a verdict of the American Supreme Court justice to the true circumstances, with the decision of the highest judges that Tesla and not Marconi is entitled the right to be the inventor of the radio. As we see, have both inventors in reality used completely other kinds of waves. The available transmitters of 100 years ago we would from today's viewpoint call ,,broadband dirt slingers". These spark flying monster have blown both parts of the wave equation, the transverse as well as the longitudinal part, into the air to ample extent. What distinguished the wave pioneers was their receiver technique, was eventually the question, which wave part they have filtered and utilized.

Marconi worked with dipole antennas, as already Heinrich Hertz. With that both could preferably track down and detect the radio wave. So they also should be entitled the right to be pioneers of radio technology. The verdict of the highest judges doesn't justice to this circumstance and should rather be judged as a nationalistically coloured political issue. Tesla however worked with two spherical electrodes, in which he preferably replaced one electrode by the globe, by grounding his devices. In this way he could receive the scalar wave parts. But that are not radio waves! Scalar waves have completely other properties, one even could be inclined to call them opposite properties.

To improve the degree of effectiveness of the transmission stretch one naturally was trying to also optimise the transmitting installation with regard to the respectively used wave part. Tesla optimised the scalar wave part and could record reactions of biological beings. His part represents a set of difficulties of the environmental compatibility, which should be taken serious. In the beginning the Marconists on the ships, as the radio operators were called, suffered from the so-called radio operator disease, which is unsolved until today. This phenomenon only disappeared after the radio devices on board had been optimised in favour of the used radio wave. The reached and measuring technically verifiable increase of the degree of effectiveness, primarily obtained by an improved antenna adjustment, simultaneously means a reduction of the scalar wave part, which endangers health.

But a received signal hides the receiver technician, if it has been on the way as a transverse or as a longitudinal wave. The coupling in one and the same equation leaves open both possibilities. Every radio amateur knows the so-called ground waves, which arrive faster at the receiver, than the Hertzian waves mirrored at the ionosphere, which propagate in a straight line. Allegedly the ground waves follow the curvature of the earth<1>, so it is written in expert books. This explanation hurts, since who can see along the curvature of the earth with a pair of field glasses. He would see the back of his head! No, the explanation the ground waves would run along the earth's surface is pure nonsense. The interference and the fading with which the radio amateur is fighting, are a result of the differently fast arriving wave parts, and doing so the scalar wave part tunnels as a straight line right through the earth (fig. 21.6)!

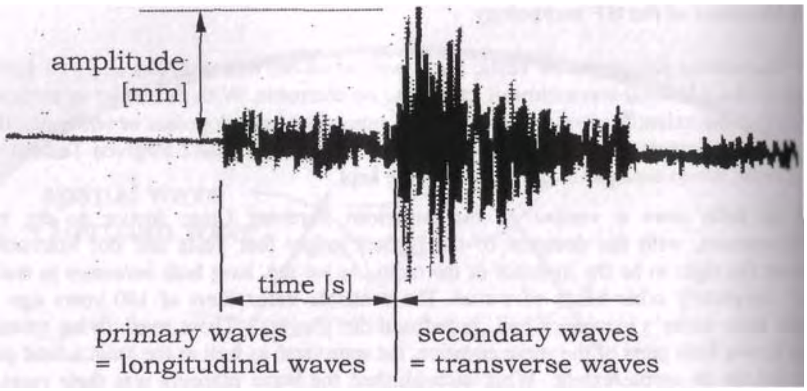

Fig. 21.7 A: Longitudinal and transverse earth quake waves.

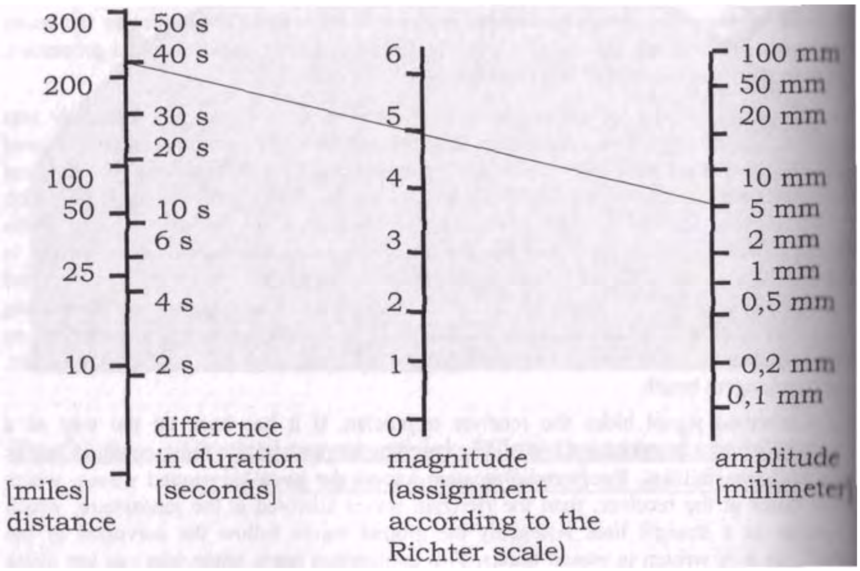
300
200
100
50
25
10
0
[miles]
distance
50 s
40 s
30 s
20 s
10 s
6 s
4 s
2 s
difference
in duration
[seconds]
6
5
4
3
2
1
0
magnitude
(assignment
according to the
Richter scale)
100 mm
50 mm
20 mm
10 mm
5 mm
2 mm
1 mm
0,5 mm
0,2 mm
0,1 mm
amplitude
[millimeter]

Fig. 21.7 B: Analysis according to the Richter scale.
(e.g.: 40 s duration between S- and P-waves for 5 mm amplitude means an earthquake of strength 5 in a distance of 220 miles.)

21.7 Coupling of the wave parts

The set of difficulties of ground waves makes clear the coupling of longitudinal and transverse waves as two aspects or parts of a wave. As the corresponding equation 21,1 mathematically taken apart into 21.2 dictates, does every transmitter emit both parts.

Exactly this circumstance the owners of allotments have used, which directly near a transmitter had illuminated their estate with freely hung up fluorescent lamps. The transmitter runners after that wanted to present them the power bill.and they could obtain, that this kind of illumination technique was prohibited.
Nowadays anyone may operate a high-frequency technical installation, even if he hasn't understood at all the wave equation used by that. Actually one should have been grateful to the allotment owners, if they withdraw the scalar wave radiation, which is incompatible with the environment and biologically harmful, with their fluorescent lamps. Taken exact it even is the task of the transmitting technician to pay attention to it. that only radio waves are sent into the air, since only those should be used. The time has come to reverse the burden of proof to protect the environment, nature, the consumer and the unasked and not involved persons, which are irradiated.

From other areas, for instance from flow dynamics or for body sound is generally known that both wave parts exist and in addition occur jointly. In the case of a propagation through the earth, like for an earthquake, both parts are received and utilized. Because their propagation is differently fast, the faster oscillations arrive first and that are the longitudinal ones. From the time delay with which the transverse waves arrive at the measurement station, the distance to the epicentre of the quake is determined by means of the different velocity of propagation. For geophysicists this tool is part of everyday knowledge (fig. 21.7).

Only who keeps one eye shut, could mean that the electromagnetic wave is purely transverse and sound purely longitudinal. It is true that a transverse sound wave doesn't get too far in air, for which reason sound as a rule is considered as a purely longitudinal wave by neglecting this part, but such a neglect may not be carried out in general, it must be checked if it is legitimate from case to case and an error consideration should be carried out.

Further examples for the coupling of the wave parts are furnished by the latest tunnel experiments. Here so-called pure transverse waves are sent into a tunnel, through which they don't fit through at all. The Maxwell-theory then dictates that behind the tunnel no signal should be measurable. But a signal is being measured, which in the tunnel in addition was faster than allowed.

In conferences again is being discussed about the wave equation. The imagination of the specialists reaches from phase velocities of an electromagnetic wave, which isn't present at all up to instantaneous tunnelling, during which the clocks should stop<i>.
The wave equation however supplies the only possible answer: The tunnel filters out the scalar wave parts and lets pass from them only those, which are sufficiently small and correspondingly fast!

<i>: see Part 1, Chapter 6.14

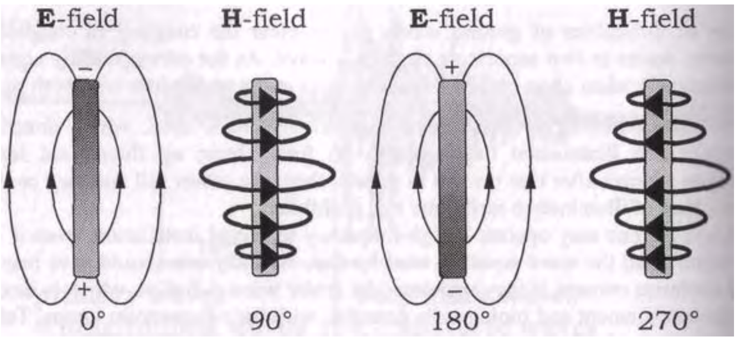

Fig. 21.8 A: The fields of the oscillating dipole antenna

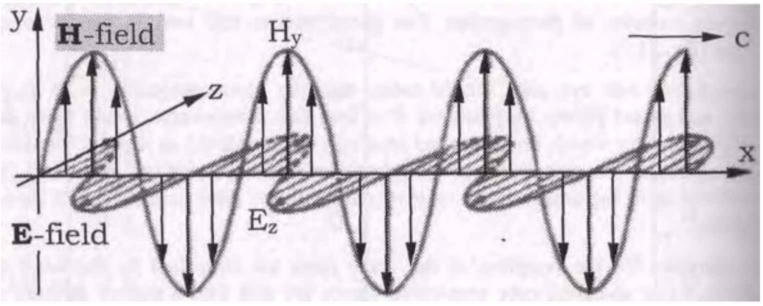

Fig. 21.8 B: The planar electromagnetic wave in the proximity

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3. Aufl. Springer-Verlag Berlin 1986, Seite 335

<ii>: dpa-message in the Sudkurier of 25.11.2000: „Gefahr furs Herz" (Translated:) Patients with a cardiac pacemaker produced in the Netherlands have been warned by the producer of the device (Vitatron) to be careful when passing anti-theft installations in stores. For devices, which were implanted between 1992 and 1998, there is the danger of the implant failing.

21.8 Set of difficulties of the near-field

In high-frequency technology is distinguished between the near-field and the far-field. Both have fundamentally other properties.

Heinrich Hertz did experiment in the short wave range at wavelengths of some meters. From today's viewpoint his work would rather be assigned the far-field. As a professor in Karlsruhe he had shown that his, the electromagnetic, wave propagates like a light wave and can be refracted and reflected in the same way. It is a transverse wave for which the field pointers of the electric and the magnetic field oscillate perpendicular to each other and both again perpendicular to the direction of propagation. It hence would be obvious, if in the case of the Hertzian wave it would concern the far-field. Besides the propagation with the speed of light also is characteristic that there occurs no phase shift between E-field and H-field.

In the proximity it looks completely different. The proximity concerns distances to the transmitter of less than the wavelength divided by 2*pi. Nikola Tesla has broadcasted in the range of long waves, around 100 Kilohertz, in which case the wavelength already is several metres. For the experiments concerning the resonance of the earth he has operated his transmitter in Colorado Springs at frequencies down to 6 Hertz. Doing so the whole earth moves into the proximity of his transmitter. We probably have to proceed from assumption that the Tesla radiation primarily concerns the proximity, which also is called the radiant range of the transmitting antenna.
For the approach of vortical and closed-loop field structures derivations for the near-field are known<i>. Doing so it must be emphasized that the structures don't follow from the field equations according to Maxwell, but the calculations are based on assumed rotation symmetrical structures. The Maxwell theory by no means is capable of such a structure shaping by principle. The calculation provides as an important result<i> that in the proximity of the emitting antenna a phase shift exists between the pointers of the E- and the H-field. The antenna current and the H-field coupled with it lag the E-field of the oscillating dipole charges for 90° (fig. 21.8). These charges form a longitudinal standing wave the antenna rod or antenna dipole. For this reason also the fields produced by high-frequency currents at first have the properties of a longitudinal wave in the proximity of the antenna.

The near-field already is used in practice in anti-theft devices, as they are installed in the entrance area of stores. The customer walks through the scalar wave transmitters. If the coupling coil has not been removed at the cash point, then a signal from the alarm system sounds. The coils work purely passive, i.e. they are supplied with electric energy per scalar wave and stimulated 'to oscillate for their part. Then the effect back on the transmitter is being utilized. Even if the principle is functioning, people still should be warned not to use a technology, which has not been understood completely. Then not explained catastrophes are inevitable<ii>.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986, Seite 335

<ii>: dpa-message in the Sudkurier of 25.11.2000: ,,Gefahr furs Herz". (quoted at the left, fig. 21.8)

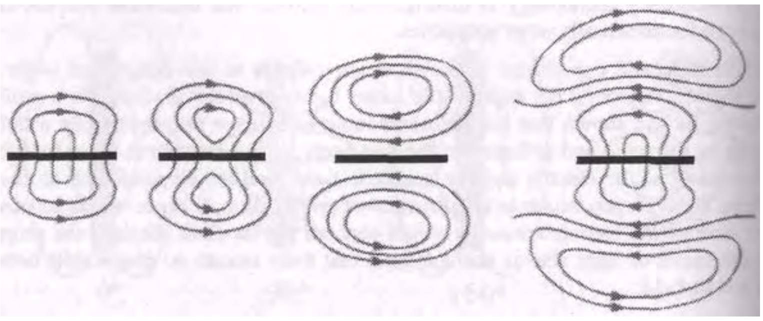

Fig. 21.9 A: The coming off of the electric field lines from a dipole
The forming vortex structures found a longitudinal electric wave carrying impulse!

electromagnetic wave (transverse)

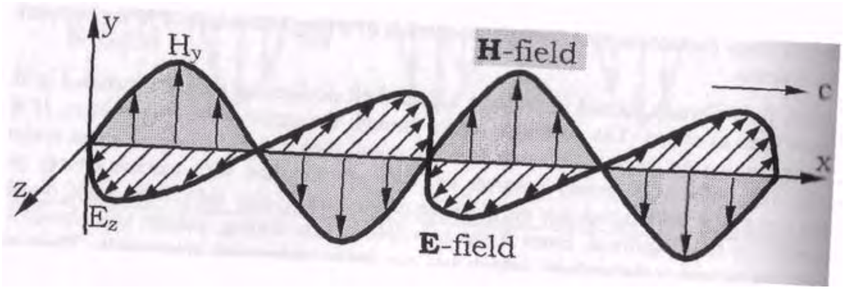

Fig. 21.9 B: The planar electromagnetic wave in the far zone

21.9 Transition to the far-field

In sufficient distance to the transmitting antenna as far-field the transverse electromagnetic wave results (fig. 21.9 B). It is distinguished by not occurring a phase shift between E- and H-field anymore. Every change of the electric alternating field is followed immediately and at the same time by a change of the magnetic alternating field and vice versa.
In the proximity however the phase shift amounts to 90°. Somewhere and somehow between the causing antenna current and the far-field a conversion from a longitudinal into a transverse wave occurs. How should one imagine the transition?

In the books the coming off of a wave from a dipole is represented according to fig. 21.9 A. The fields come off the antenna, the explanation reads. If we consider the structure of the fields coming off then we see field vortices, which run around a point, which we can call the vortex centre. Such field structures naturally are capable of forming standing waves and to carry an impulse. The scalar wave field in general and the near-field in special we only will understand with suitable vortex physics and with a field theory extended for corresponding vortices we also will be able to calculate it. Postulates cannot replace field physics!

Be that as it may, the vortex, after having left the antenna, for bigger getting distance at some time seems to unroll to propagate further as an electromagnetic wave. There takes place a transition from longitudinal to transverse, or spoken figuratively, from vortex to wave. How complete this conversion takes place, how big the respective wave parts are afterwards, on the one hand depends on the structure and the dimensions of the antenna. Information is given by the measurable degree of effectiveness of the antenna.
The vortex structures on the other hand are the stabler, the smaller and faster they are. If they are as fast as the light or even faster, then they become stable elementary particles, for instance neutrinos. Slower vortex structures however are predominantly instable. They preferably unwind to waves. Vortex and wave prove to be two possible and under certain conditions even stable field configurations.

Let's emphasize: A Hertzian dipole doesn't emit Hertzian waves! An antenna as near-field without exception emits vortices, which only at the transition to the far-field unwind to electromagnetic waves. A Hertzian wave just as little can be received with a dipole antenna! At the receiver the conditions are reversed. Here the wave is rolling up to a vortex, which usually is called and conceived as a ,.standing wave". Only this field vortex causes an antenna current in the rod, which the receiver afterwards amplifies and utilizes.
The mostly unknown or not understood near-field properties prove to be the key to the understanding of the wave equation and of the method of functioning of transmitting and receiving antenna. The question is asked, how one should imagine the rolling up of waves to vortices and vice versa the unrolling? How could an useful vortex mode) look like?

Circularly polarized wave (transverse)

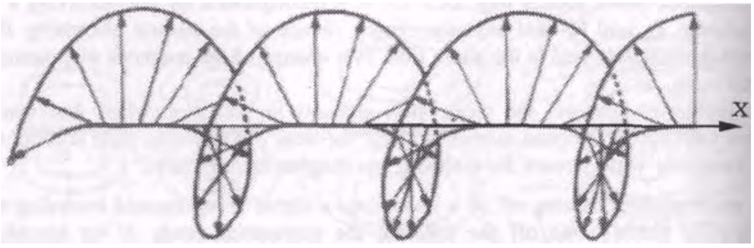

Fig. 21.10 A: Left-circular polarized wave (as explanation for the transition to a vortex and to a scalar wave)

electric wave (longitudinal)

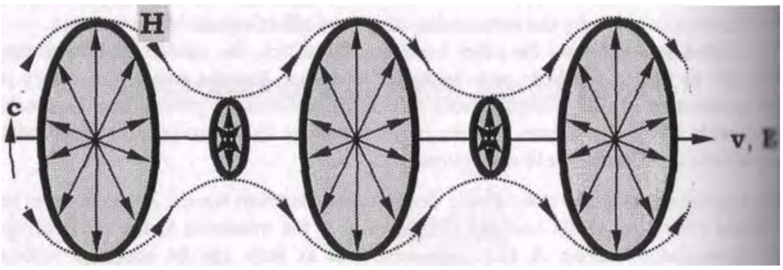

Fig. 21.10 B: Magnetic ring-vortices form an electric scalar wave.

vortex and wave = two stable field configurations

electromagnetic wave = transverse wave propagating in a straight line

ring-like vortex = transverse wave running in circles

vortex velocity = speed of light c

change of structure = if the field is disturbed without expense of energy

22.10 Scalar wave model

The light, as electromagnetic wave, in the presence of a heavy mass or of strong fields is bent towards the field source (fig. 6.10). The wave normally propagating in a straight line thus can be diverted. The square of the speed of light further is inversely proportional to the permeability and dielectricity, short, in presence of matter it is more or less strongly slowed down. If this slowing down of the wave occurs one-sidedly, then a bending of the path can be expected as well. At the end of the antenna a reflection and a going back of the wave can occur, which at the other end again hits itself. Now the wave has found a closed-loop structure, which can be called vortex. The figures 21.10 B and 21.11 A show the two possible structures.

In technical books this vortex with the properties of a „standing wave" is explained gistly. Near-field and standing wave are two examples, how the textbooks succeed in describing mathematically right a small part of the scalar wave properties, without having to have a good look at vortex physics. With such auxiliary descriptions however the end is reached fast, if for instance it concerns understanding a pure scalar wave transmission according to Nikola Tesla (fig. 19.11) and the special properties of this wave type. With the vortex concept of an extended field physics new horizons are opening.

If we direct our look again to the sketches (fig. 21.10 B and 21.11 A). In both cases electromagnetic waves are represented which propagate with the speed of light, only that the wave doesn't go forward in a straight line but instead runs around in circles. It also furthermore is transverse, because the field pointers of the E-field and the H-field oscillate perpendicular to c. By means of the orbit the speed of light c now has become the vortex velocity. Wave and vortex turn out to be two possible and stable field configurations. For the transition from one into the other no energy is used; it only is a question of structure. The vortex structure thus stabilizes itself by means of the field dependency of the speed of light.

By the circumstance that the vortex direction of the ring-like vortex is determined and the field pointers further are standing perpendicular to it, as well as perpendicular to each other, there result two theoretical formation forms for the scalar wave. In the first case (fig 21.10 B) the vector of the H-field points into the direction of the vortex centre and that of the E-field axially to the outside. The vortex however will propagate in this direction in space and appear as a scalar wave, so that the propagation of the wave takes place in the direction of the electric field. I call this an electric wave.

In the second case the field vectors exchange their place. The direction of propagation this time coincides with the oscillating magnetic field pointer (fig. 21.11 A), for which reason I speak of a magnetic wave.

The vortex picture of the rolled up wave already fits very well, because the propagation of a wave direction of its field pointer characterizes a longitudinal wave, because all measurement results are perfectly covered by the vortex model. It even is clear that no energy has to be spended for the conversion, since merely the structure has changed. If it becomes a vortex the wave just doesn't run in a straight line anymore but in circles, to either wrap around the magnetic field vector (fig. 21.10 B). or the electric field vector (fig. 21.11 A).

magnetic wave (longitudinal)

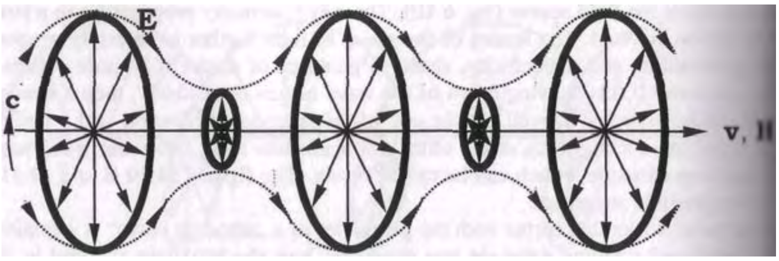

Fig. 21.11 A: The magnetic scalar wave

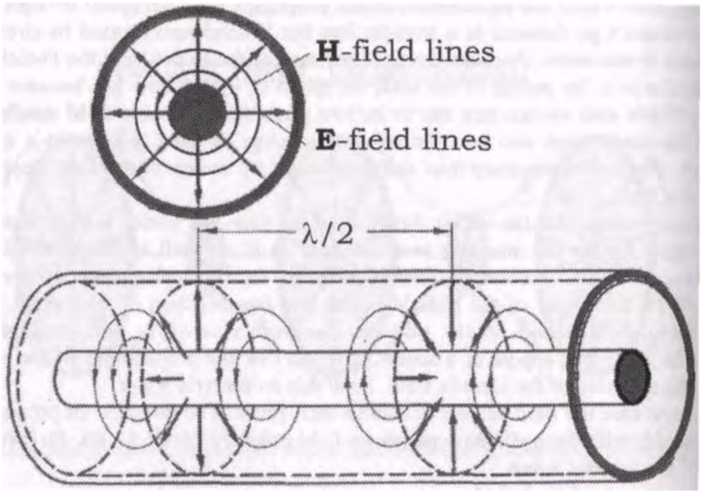

Figure 21.11 B: Wave propagation in a coaxial cable,
(Example for waveguide, horn radiator, etc.)[i]
Cross-section of coaxial conductor and field distribution in the direction of propagation.

<i>: H. Armbruster, G. Grunberger: Elektromagnetische Wellen im Hochfrequenz-

21.11 Double-frequent oscillation of size

Because a longitudinal wave propagates in the direction of the field, the field pointer also will oscillate with the velocity of propagation v. This hence isn't constant at all, it can significantly differ from that of the light and can take arbitrary values. According to the theory of objectivity the field oscillating with it determines its momentary size:

$$E, H \sim 1/v^2 \qquad (21.10).$$

The velocity of propagation v of the scalar wave thus oscillates double-frequently and with opposite phase to the corresponding field. A detailed description would mean, if the field strives for its maximum value, the velocity v of the wave reaches its smallest value. In the field minimum the scalar wave vice versa accelerates to its maximum value. For longitudinal waves therefore only an averaged velocity of propagation is given and measured, as this for instance is usual for the sound wave, and this can vary very strong as is well-known (body sound compared to air sound, etc.).
The two dual field vectors of E and H, the one in the direction of propagation and the one standing perpendicular to it, occur jointly. Both oscillate with the same frequency and both form the ring-like vortex in the respective direction. As a result the ring-like vortex also oscillates in its diameter double-frequently and with opposite phase to the corresponding field (fig. 21.10 Band 21.11 A).
This circumstance owes the ring-like vortex its property, to tunnel. No Faraday cage is able to stop it, as could be demonstrated in experiments<i>. Only therefore the ground wave runs through the earth and not along the curvature of the earth. A further example is the coaxial cable (fig. 21.11 B). Also this acts as a long tunnel and so it isn't further astonishing, that the electric field lines have the same orientation, as for a magnetic scalar wave. As a practical consequence in this place there should be warned of open cable ends, wave guides or horn radiators with regard to uncontrolled emitted scalar waves!

At present in the press is being discussed, if the cable network runners for some channels the mission to operate again should be withdrawn, because the airline radio traffic is being disturbed. The original opening for cable frequencies, which actually are reserved for the airline radio traffic, based on the erroneous assumption, that conflicts are unthinkable. But then the planes were disturbed in their communication. As the cause TV-cables made out, which hadn't been closed according to the rules with a resistor, as it by all means can occur on building sites and during renovation works.
On the other hand is being argued with the small current, which flows through the coaxial cable, and the large distance to the planes also is cited. According to that it actually can't concern Hertzian waves. It presumably are scalar waves, which escape from the open cable ends and which are collected by a receiver in the plane. It indeed is very little field energy, but because it again is being collected and bundled, the scalar wave is able to exceed the part of the radio wave by far just at large distances and to cause problems. For such examples from practice the scalar wave theory is fully taking effect.

<i>: Adolf und Inge Schneider im Interview mit Prof. Dr.-Ing. Konstantin Meyl: Durchbruch in der Freien-Energie-Forschung, NET-Journal 12/1999, S. 7-9.

1. H. Hertz: electromagnetic wave (transverse)

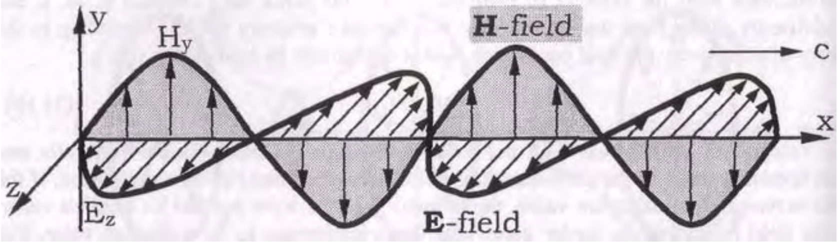
y
H_y
H-field
c
x
z
E_z
E-field

---

2. Nikola Tesla: electric wave (longitudinal)

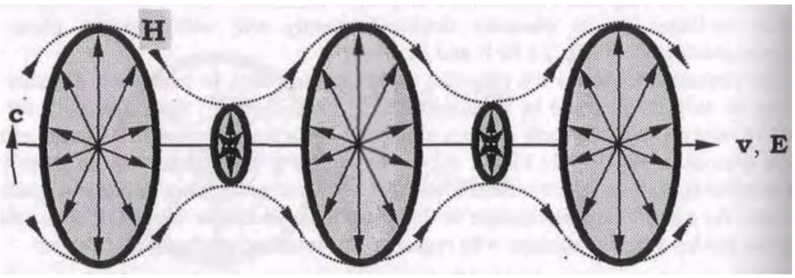

---

3. magnetic wave (longitudinal)

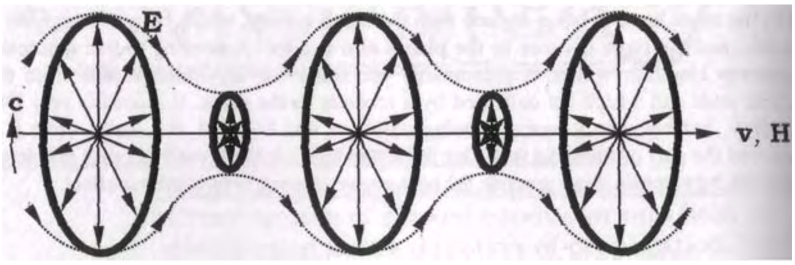

Fig. 21.12: The three basic types according to the wave equation
(21.1), (electric, magnetic and electromagnetic wave).

21.12 Electric and magnetic scalar wave

Let us record: For the wave propagation there thus are three possible and stable states (fig. 21.12): the transverse electromagnetic wave according to Heinrich Hertz (fig. 1), the longitudinal electric wave according to Nikola Tesla (fig. 2), and a longitudinal magnetic wave (fig. 3), which isn't connected yet with a name of a discoverer. The last one is a pure product of my theoretical derivation. The question is asked, which practical meaning the magnetic wave could have.
It is formed by individual electric field vortices, which I have discovered and called potential vortices 1990. I proceed from the assumption that the youngest of the three waves will play the by far biggest role in the future, because its properties are unattainable, both with regard to the energy technical and to the information technical use. One example for each should support this thesis.

The experiments concerning the electric wave according to Nikola Tesla, where is being worked with electrically charged spheres, don't show a particularly high power. Magnetic converters, so the experiences of my laboratory activities, are superior to an electrostatic converter as a collector for free energy by far. That even can be expected, because a magnetic engine is much smaller than an electrostatic engine of the same power as is well-known.

At a congress of medicines was given a talk on the basic regulation of the cells, on the communication of the cells with each other. Professor Heine in his decades of research work has found out that the cells for the purpose of communication build up channels for instance in the connective tissue, which after having conducted the information again collapse. Interestingly the channels have a hyperboloid structure, for which no conclusive explanation exists.
The structure of the data channels however is identical with the one of a magnetic scalar wave, as shown in fig. 3. Through a channel formed such, which functions like a tunnel or a dissimilarly formed waveguide, only one very particular scalar wave can run through. Waves with different frequencies or wavelengths don't fit through the hyperboloid formed tunnel at all in the first place. Through that the information transmission obtains an extremely high degree of safety for interference.
To the biologist here a completely new view at the function of a cell and the basic regulation of the whole organism is opening. The information tunnel temporarily forms more or less a vacuum, through which only potential vortices can be conducted, and that without any losses - simply perfect! From this example is becoming clear that nature is working with scalar waves namely with magnetic waves.

One other point should be recorded: The mentioned tunnel experiments, in which speed faster than light is being measured with most different devices, impressively confirm the presence of scalar waves. But if scalar waves exist which are faster than light and other ones, which are slower, then it almost is obvious that also such ones will exist, which propagate exactly with the speed of light. These then will have all the properties of the light and won't differ from the corresponding electromagnetic wave in the observable result. As scalar wave it however is formed by vortex configurations, which unambiguously have particle nature. Nothing would be more obvious than to equate these quantum structures with the photons.

Question: Is the light wave or particle?

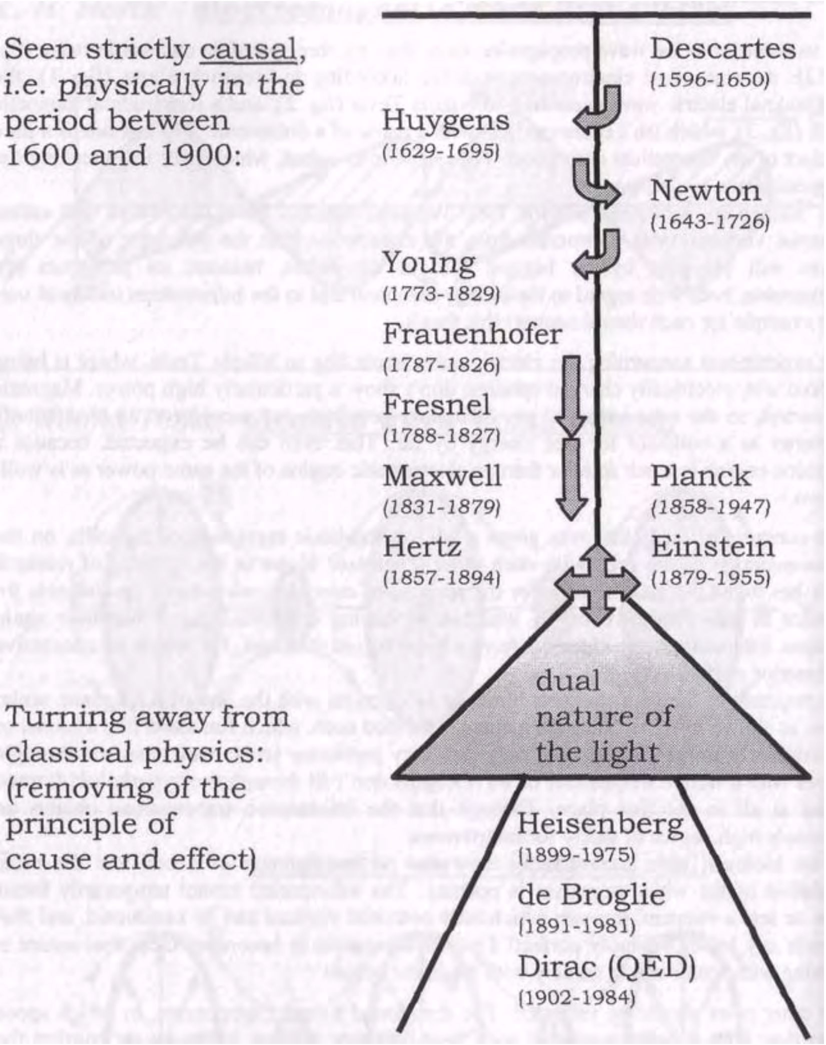
Seen strictly causal, i.e. physically in the period between 1600 and 1900:
Descartes
(1596-1650)
Huygens
(1629-1695)
Newton
(1643-1726)
Young
(1773-1829)
Frauenhofer
(1787-1826)
Fresnel
(1788-1827)
Maxwell
(1831-1879)
Planck
(1858-1947)
Hertz
(1857-1894)
Einstein
(1879-1955)
dual nature of the light
Turning away from classical physics: (removing of the principle of cause and effect)
Heisenberg
(1899-1975)
de Broglie
(1891-1981)
Dirac (QED)
(1902-1984)

Fig. 22.1: The view of some physicists concerning the nature of the light as wave or as particle.

According to the wave equation the light is a mixture of wave and photon radiation!

22 Properties of scalar waves

For the light quanta hypothesis the Nobel Prize for physics was awarded 1921. But it only was a hypothesis, an idee fixe, which was honoured here. It is quite odd, if such an important prize is awarded to a research scientist, who hasn't got the slightest idea what light quanta are anyway, of what they consist and how they are built up. Albert Einstein cleverly used the embarrassing situation, by in his celebration speech on the occasion of the awarding of the Nobel Prize giving a talk on the theory of relativity. After the speech a member of the committee found it necessary, to point to it, that this wasn't object of the awarding of the prize and that the theory of relativity concerns a pure theory, which can't be proven by principle. A theory hence neither could be awarded the Nobel Prize.

Such words on the occasion of awarding a prize for a daredevil hypothesis give the whole event really grotesque characteristics. But it came still worse.

22.1 Wave-particle compromise

Physicists of name and rank had come together. It concerned the question if the light is wave, or particle or even both at the same time? For both variants experimental proof was present, the discussion became inflamed and the things boiled over. Finally they were as smart as before, as Werner Heisenberg presented his ideas concerning the uncertainty principle. This compromise, on which one eventually came to an agreement, with good cause may be called the worst in the history of physics. It dictates me, what I shell see and how exact I may look. With it the contradiction should be overcome that the light contrary to every causality should be wave and particle at the same time.

Such fixings not only have a funny, but also a tragic side. Since it were authorities, which have approved the compromise and the whole community of science has confidence in the statements of its authorities, which immediately and unfiltered is entered in all textbooks. At the meeting it simply and solely concerned the wave equation and only that could have supplied the correct and only possible answer: It falls apart into two parts and this explains, why the light one time appears as an electromagnetic wave and the next time as a vortex particle, which is called photon. The conversion can take place at any time spontaneously and without putting on energy, so that depending on the used measuring technique the particle appears as wave or as particle, but of course never as both at the same time!

Looking back one can say that the funny thing about the situation was that all discussed about the wave and its properties known at that time, that all should know the wave equation. An equation as is well-known says more than a thousand words and one look would have sufficed entirely, to answer the controversial question once and for all. It would have saved us a lot.

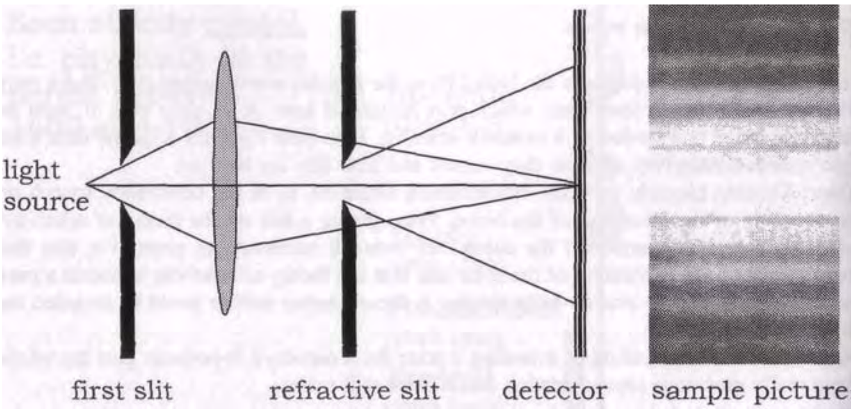
light source
first slit
refractive slit
detector
sample picture

Fig. 22.2 A: Light forms interference patterns at the slit (light stripes are formed, where the waves oscillate in phase, dark stripes, where they oscillate out of phase).

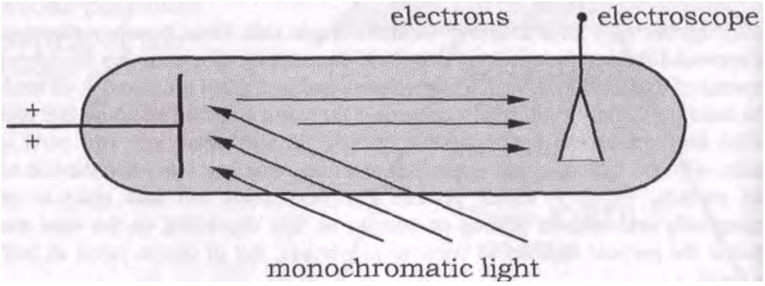
electrons
electroscope
+
+
monochromatic light

Fig. 22.2 B: The photo-electric effect

<i>: Atome als Energiewellen. Physiker wandelten einen Strom von Natriumatomen in Wellen um. III. Wissenschaft 7/1996, Seite 56 + 57

22.2 Concerning the measurement of light

The uncertainty principle with the interpretation of Heisenberg, the light is wave and particle at the same time, is incompatible with the wave equation. Heisenberg puts an equal sign, where in the wave equation in reality is present an addition of both wave parts. Fortunately in mathematics there is no need of speculating, there a derivation is right or wrong. Nothing is changed to that even if all physicists of the world should go in the wrong direction following the prevailing opinion. The wave equation exert an influence on the interpretation of the light experiments, on the one hand the ones concerning the interference and refraction of the light, where electromagnetic waves are becoming visible (fig. 22.2 A) and on the other hand the photo-electric effect, as proof of light quanta (fig. 22.2 B).

Already the wave theory of Huygens requires interference patterns of light rays, as they for instance are observed behind a slit, and demonstrates with that the wave nature. If on that occasion the particle nature is lost, if thus the photons present before the slit can't be detected behind the slit anymore, then plain and simple the measuring method, thus the slit is to blame for that. The vortices have unrolled themselves at the slit to waves. Corresponding experiments also have been carried out with matter. At the Massachusetts Institute of Technology whole sodium atoms were converted into waves. At the detector pure interference patterns were observed, which go as evidence for the successful dematerialization<i>. But the vortex physicist they show still more: they reveal, that atoms merely are waves rolled up to spherical vortices, which at any time and spontaneously can again unroll to waves at a lattice (chapter 5 and 7).
The common interpretation, the wave nature detectable behind a slit must have been present in the same form already before the slit, is untenable and in the end wrong, as makes clear the experiment with the sodium atoms.

The photo-electric effect, which on the other hand shows the quantum nature of the light, has been discovered by Heinrich Hertz, further investigated by Lenard and finally rendered more precisely by Millikan 1916 (fig. 22.2 B). It bases on the circumstance that light of higher frequency, thus blue light, has more energy than red light of lower frequency. But if electrons are knocked out a metal plate by light, then that occurs, by the waves rolling up to vortices. Now indeed photons are at work, which are detected with an electroscope indirectly.
In the same way a photon ray in a bubble chamber can be photographed. But also here the measuring method is responsible for what is being observed.

A good example is the human eye, the rods and cones of which merely can pick up potential vortices and pass them on to the nerves as so-called reaction potentials. Incident waves can only be detected, if they first have rolled up to vortices in the corpus vitreum of the eye. For us seeing, it doesn't play a role of how many percent vortices and waves the light is consisting.
Behind a sheet of glass for instance a larger vortex part can be expected and still the light has the same brightness as without sheet; the sheet of glass is perceived as transparent. We nevertheless must assume that light with a large wave part has another quality, than such light behind glass or artificial light with a large part of photons.

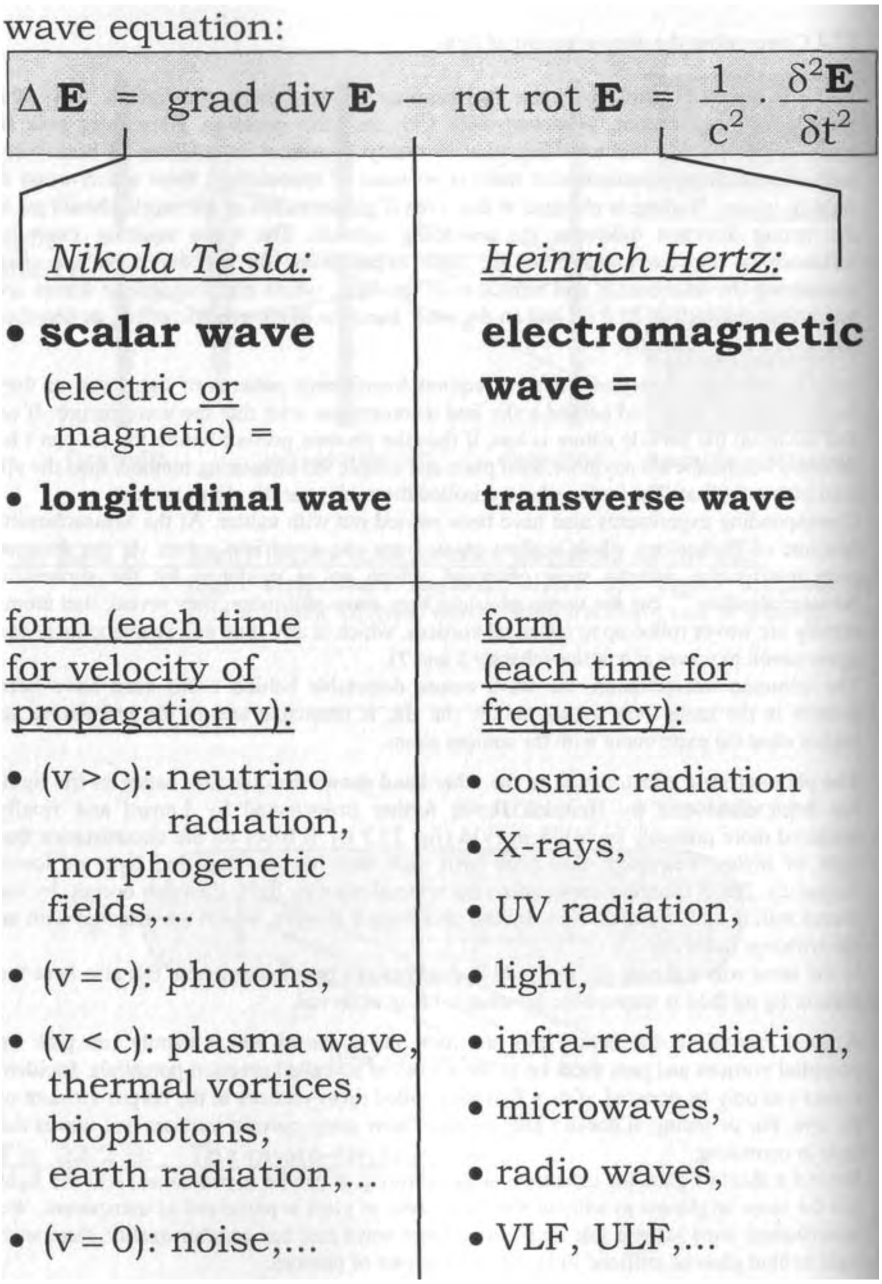

wave equation:

$$\Delta \mathbf{E} = \text{grad div } \mathbf{E} - \text{rot rot } \mathbf{E} = \frac{1}{c^2} \cdot \frac{\delta^2 \mathbf{E}}{\delta t^2}$$

Nikola Tesla:	*Heinrich Hertz:*
• **scalar wave** (electric or magnetic) =	**electromagnetic wave =**
• **longitudinal wave**	**transverse wave**
form (each time for velocity of propagation v):	form (each time for frequency):
• (v > c): neutrino radiation, morphogenetic fields,...	• cosmic radiation • X-rays, • UV radiation,
• (v = c): photons,	• light,
• (v < c): plasma wave, thermal vortices, biophotons, earth radiation,...	• infra-red radiation, • microwaves, • radio waves,
• (v = 0): noise,...	• VLF, ULF,...

Fig. 22.3: The two parts of the wave equation

22.3 Comparison of the parts of Tesla and Hertz

Light as a rule always is formed as photon radiation, even on the sun. If in the end only waves arrive on earth, then the vortices sometime on the way to us must have unrolled to waves. Photon radiation after all is a scalar wave radiation, which generally is predominant in the near-field of the source of radiation. There is no reason, why the light should act in another way than the wave radiated by a radio transmitter, which as well forms vortices in the near-field area, as we already have discussed. For different interpretations of wave properties of one and the same physical phenomenon there is no place in a unified theory.

If we stay at the comparison then it is not an individual case that an experimental setup is responsible for what is being measured and observed. A parallel case to the experiments concerning the nature of the light is the one concerning the wave propagation. Hertz has received and utilized the transverse part and Tesla the longitudinal part and either one claimed, only he is right. There doesn' t exist an other equation, which has been and is being ignored and misunderstood so thoroughly, as the wave equation.

Fig. 22.3 shows in a survey the two parts of the wave equation in the assignment to the terms and forms: Right hand side the electromagnetic wave according to Heinrich Hertz and left hand side the scalar wave according to Nikola Tesla. The terms, like on the one hand transverse wave and on the other hand longitudinal wave relate to the kind of wave propagation.

If the field pointers oscillate crossways to the direction of propagation, then as a consequence the velocity of propagation is decoupled from the oscillating fields. The result in all cases is the speed of light, and that in our observation is constant. It is usual to make a list for increasing frequency, starting at the longest waves (ELF and VLF) over the radio waves (LW, MW, SW, UHF), the TV channels (VHF, UHF), the microwaves, the infra-red radiation, the light, the X-rays up to the cosmic radiation.

It really is interesting that it concerns one and the same phenomenon despite the different forms! As long as Maxwell only had published a theory for the light, in the world of science 24 years long at first nothing at all happened. Only Heinrich Hertz with his short wave experiments opened the eyes. Now all suddenly started at the same time to research into various phenomena on the frequency scale, from Madame Curie over Konrad Rontgen up to Nikola Tesla, who primarily researched the area of long waves.

With regard to the scalar waves until now a corresponding booster detonation has failed to appear. The immense area is new ground scientifically, which is awaiting to be explored systematically. I try to make a contribution with my historic rebuild of a scalar wave transmission line according to the plans of Tesla<i>.

<i>: Im Gesprach mit dem Fernsehmoderator und Buchautor Johannes von Buttlar weise ich auf die Chancen und technischen Moglichkeiten hin: Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, Argo-Verlag Marktoberdorf, 1. Aufl. (2000).

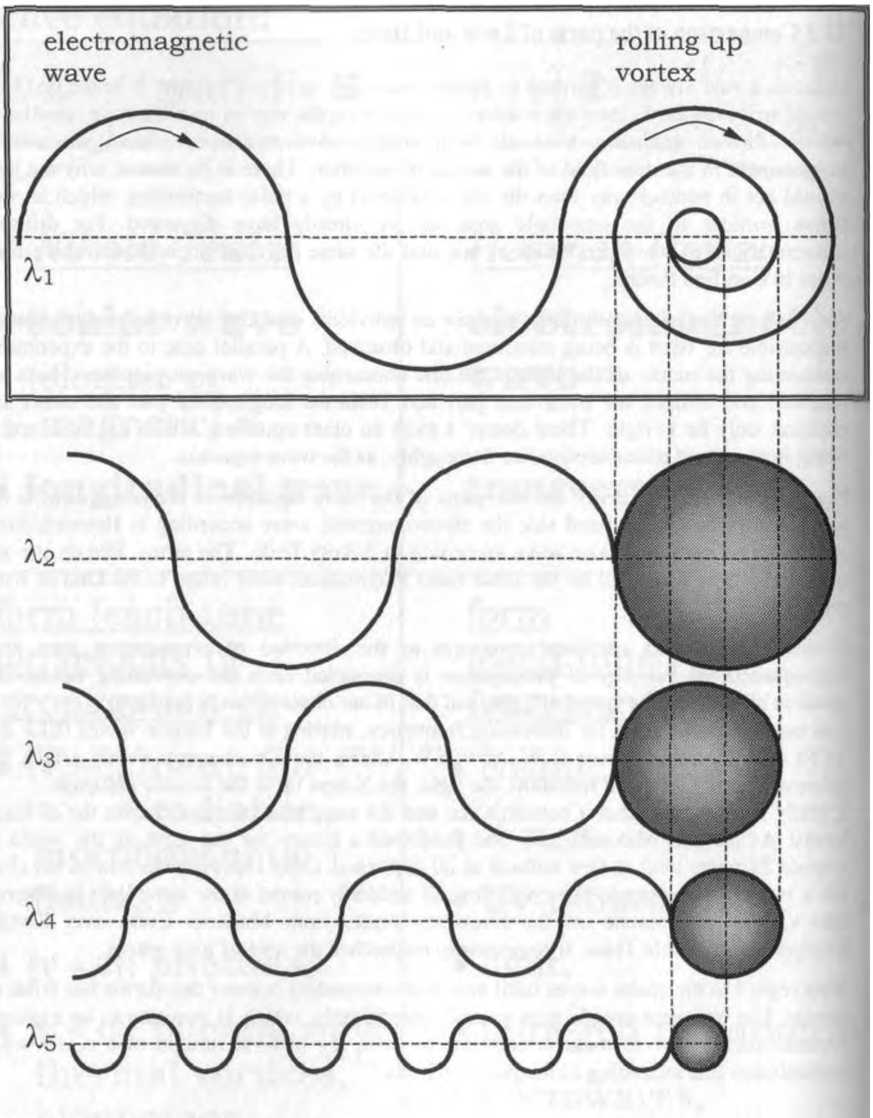
electromagnetic wave
rolling up vortex
λ_1
λ_2
λ_3
λ_4
λ_5

Fig. 22.4: The wave rolling up to a vortex.

wavelength of the wave: $\lambda_1 = c/f_1$

wavelength of the vortex: λ_n (n = 1,2,3,4,5) $\leq \lambda_1$

$f_5 > f_4 > f_3 > f_2 > f_1$ $\lambda_5 < \lambda_4 < \lambda_3 < \lambda_2 < \lambda_1$

22.4 Noise, a scalar wave phenomenon

Longitudinal waves can take arbitrary velocities between zero and infinity, because they propagate in the direction of an oscillating field pointer and as a consequence of that their velocity of propagation oscillates as well and by no means is constant. It does make sense to llist the forms of scalar waves according to their respective velocity of propagation (fig. 22.3, left column).

If we start with a localized vortex, a wave rolling up, which further is contracting. Doing so the wavelength gets smaller and smaller, whereas the frequency increases. An evenly frequency mixture distributed over a broad frequency band is observed. This phenomenon is called noise (fig. 1.4). But besides the localized noise, noise vortices can also be on the way with a certain velocity as a scalar wave, e.g. for the radio noise. In this case they show the typical properties of a standing wave with nodes and antinodes.

Also the earth radiation is said to have standing wave nature, which can be interpreted as slowed down neutrino radiation. If it is slowed down on the way through the earth, then the neutrino properties are changing, as this was measured in the Kamiokande detector in Japan recently. Unfortunately the proof occurs only indirect, because there still don't exist measuring devices for scalar waves. We'll talk about this problem area later and are content with the clue that already within living memory the standing wave property has been used to find water and deposits of ores and still is used today (fig. 22.4).

If we continue our considerations concerning the forms of scalar waves, as they are listed in fig. 22.3. The scalar waves, which are slower on the way than the light, are joined by the plasma waves. This is confirmed by measurements and calculations.

For thermal vortices, as they have been investigated by Max Planck and for biophotons, as they can be detected in living cells by colleague Popp, the velocity of propagation however is unknown. It was not and still is not measured at all, now more than ever. The research scientists have confidence in the assumption that all waves go with the speed of light, but that is a big mistake.

For all wave kinds there at least exists also one vortex variant, for radio waves for instance it is the radio noise, which propagates with a velocity different from c. The velocity is the product of frequency and wavelength:

$$\boxed{v = f \cdot \lambda} \quad (22.1)$$

From the three variables v, f and λ at least two must be measured, if one has a suspicion that it could concern scalar waves. At this place most errors are made in the laboratories. Countless experiments concerning the biological compatibility, concerning medical therapy methods and similar experiments must be repeated, because as a rule only the frequency is being measured and it has been omitted to at least check the wavelength or the velocity of the wave. Countless research scientists must put up with this accusation. Much too blind the scientists, who now again may start from the very beginning with their work, have had confidence in the predominance of the speed of light.

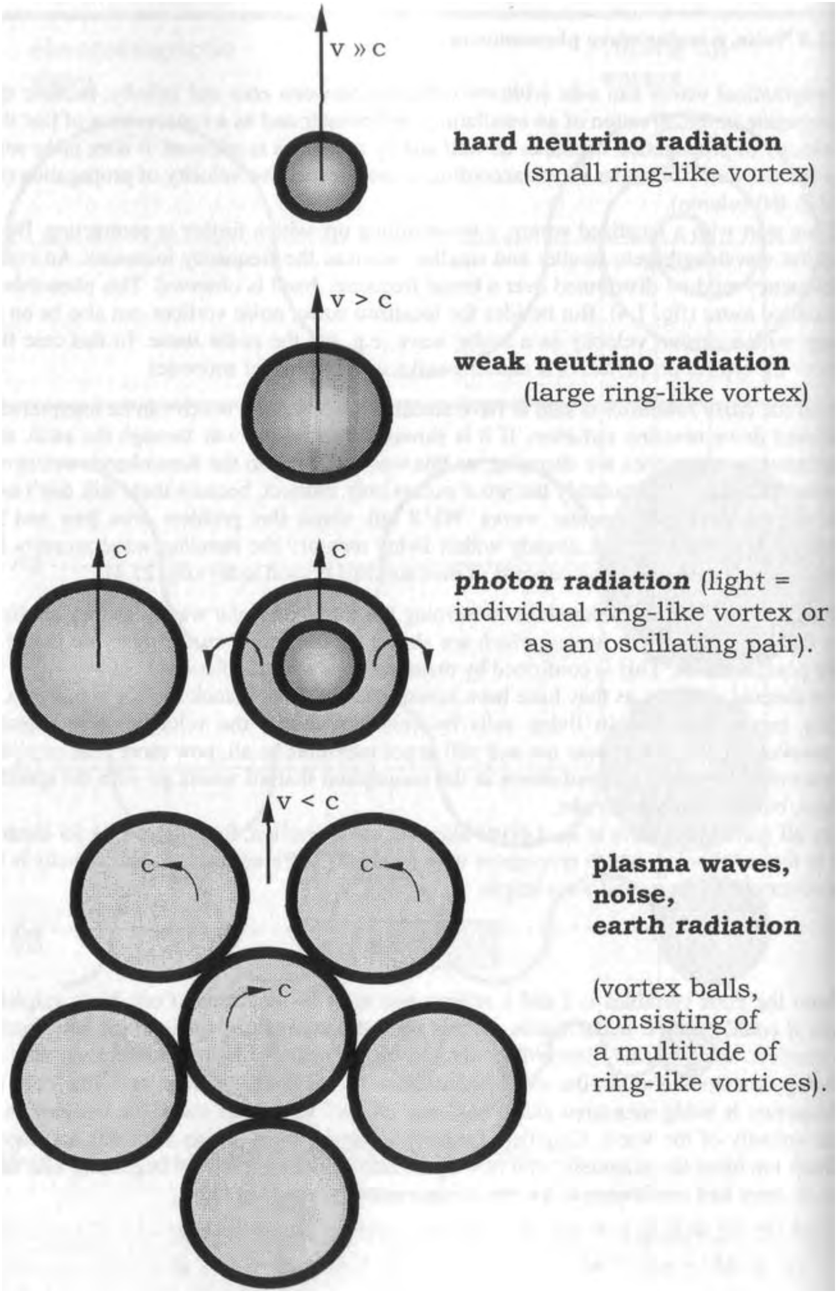
v » c
hard neutrino radiation
(small ring-like vortex)
v > c
weak neutrino radiation
(large ring-like vortex)
c
c
photon radiation (light = individual ring-like vortex or as an oscillating pair).
v < c
c
c
c
plasma waves, noise, earth radiation
(vortex balls, consisting of a multitude of ring-like vortices).

Fig. 22.5: ____The ring-like vortex model of scalar waves.

22.5 Neutrino radiation

The neutrino physicists make the same error. They proceed from the assumption that their particles are on the way with a speed somewhat less than the speed of light c. This contradicts the observation according to which black holes should represent strong sources of neutrinos, which are black only for the reason that no particle radiation is able to escape them, which is on the way with c or even slower. If a black hole does hurl neutrino radiation into space, than that must be considerably faster than c, as normal neutrino physicists still by no means can imagine it today.

But the neutrino radiation only can be detected after it has been slowed down to a value, which is smaller than c. If the slowing down occurs slightly assymmetrical, then as a consequence a mean of the mass different from zero appears. The „measurement" of such a rest mass, as it at present is propagated and celebrated, is a classical measurement error! As long as a neutrino on the way to us still is faster than the light, the mean of its mass is generally zero. The effective value of the mass of a neutrino is however considerable. Only it is able to give account for the sought-for dark matter, as far as it must exist in the today supposed form anyway.

The Tesla radiation, that the discoverer Nikola Tesla already in own experiments had found out, is faster than light (chapter 9.7 and 17.2). Since this Tesla radiation according to the description is identical with the neutrino radiation, since it so to say forms a subset, I will call neutrino radiation all the scalar waves, which are faster than the light. This stretches from the weak radiation at low frequencies up to the hard neutrino radiation of cosmic origin. But the hardness of the radiation does not only increase with the frequency, it in particular increases with the velocity.

The neutrino radiation first of all is carrying energy. On top of this basic wave radiation in addition information can be modulated. Doing so extremely complex modulation variants are offering. Of this kind we must imagine thoughts, as being complex modulated vortices, which can propagate as scalar wave in space. Rupert Sheldrake calls this vortex field a morphogenetic field. At this place merely is pointed at his very interesting research results<i>.

Thoughts can be standing in space, in the form of localized noise, but they also can move with speeds faster than light. According to that a communication with intelligent beings from other star systems by all means wouldn't be an Utopia anymore.

Every fast neutrino forms an individual ring-like vortex (fig. 7.12). The slower the scalar wave is, the more dependent the vortices become. The photon already can consist of two ring-like vortices (fig. 4.6), whereas plasma waves and other slow scalar waves can form from a multitude of individual vortices, which are rotating around each other, to form vortex balls and vortex streets (chapter 4.9 - 4.11). From this circumstance already results very different scalar wave behaviour in the different areas of the velocity of propagation. This trend for small velocities can as well be observed towards lower frequencies. For a certain wavelength the frequency after all (according to eq. 22.1) is proportional to the velocity of propagation.

<i>: R. Sheldrake: Seven experiments that could change the world. New York: Riverhead 1995

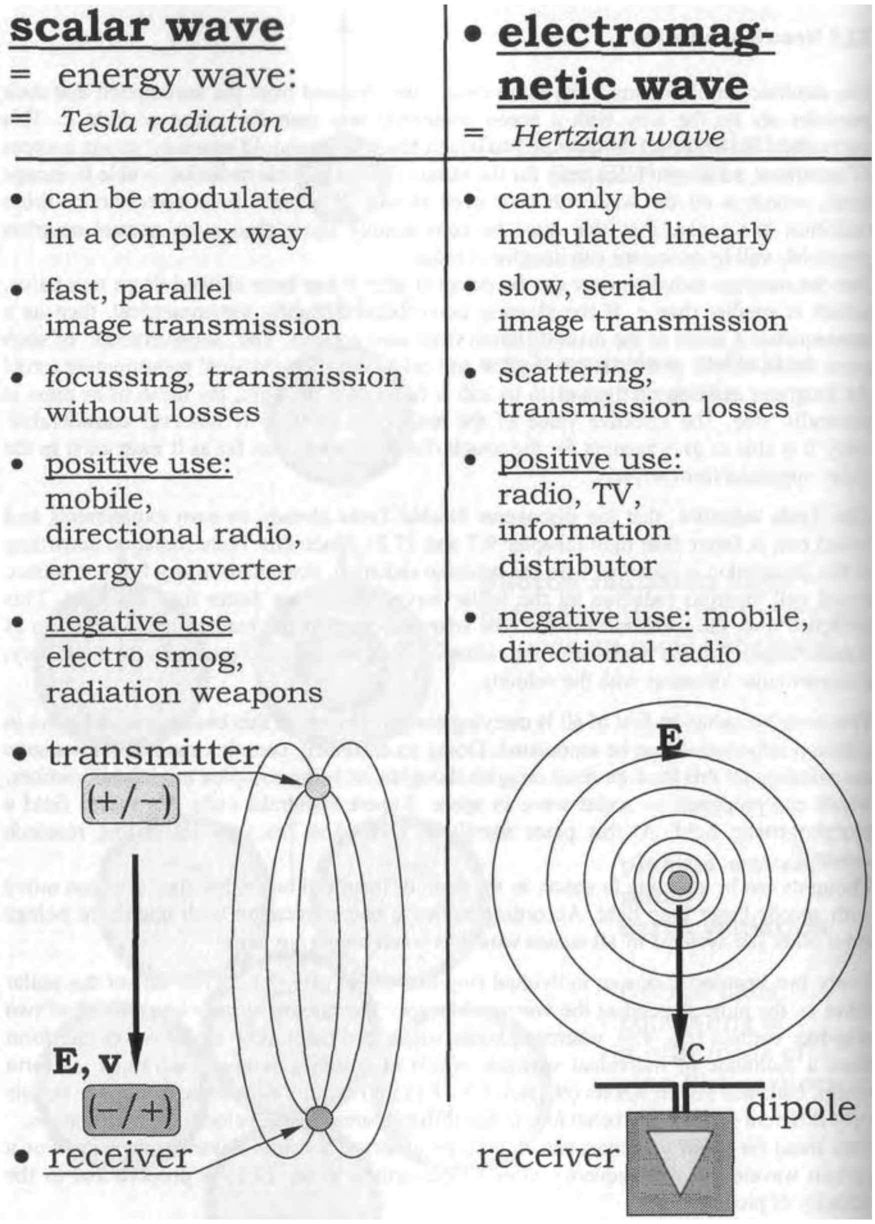

scalar wave = energy wave: = *Tesla radiation*	• **electromagnetic wave** = *Hertzian wave*
• can be modulated in a complex way	• can only be modulated linearly
• fast, parallel image transmission	• slow, serial image transmission
• focussing, transmission without losses	• scattering, transmission losses
• positive use: mobile, directional radio, energy converter	• positive use: radio, TV, information distributor
• negative use electro smog, radiation weapons	• negative use: mobile, directional radio

Fig. 22.6: Scalar waves and radio waves, comparison of the properties.

22.6 Parallel instead of serial image transmission

We continue with our considerations concerning the special properties of scalar waves, represented in the left column, and compare these with the well-known behaviour of electromagnetic waves in the right column (fig. 22.3 is now followed by fig. 22.6). If we again take up the possibilities for modulation and the transmission of information, then it becomes very clear from the comparison that we today work with a technology, which we it is true master more or less, but which is everything else but optimum. For the Hertzian wave the velocity of propagation is constant. With the frequency therefore at the same time also the wavelength is being modulated. But that strongly limits the information transmission. An image for instance must be transmitted serially point after point and line after line. The serial image transmission takes place very slowly, for which reason the velocity of the PCs permanently must be increased, so that the amount of data can be managed.

With the clock frequency on the other hand also the losses increase, so that in the end the CPU-cooler limits the efficiency of modern PCs. Something our engineers obviously do wrong, as a comparison with the human brain clarifies. Our brain works without a fan. For it a clock frequency of 10 Hertz is sufficient. It needs neither Megahertz nor Gigahertz frequencies and despite that is considerably more efficient.

Nature only works with the best technology. The second best technology, as it is put to use in our machines, in the evolution wouldn't have had the slightest chance of surviving. The strategies to optimize of nature are merciless. In a free economy that goes completely different. There the „bunglers" are joining together to companies dominating the market, buying up the innovative ideas without further ado, to let them disappear in the drawer, so that they can bungle further in the way they did until now. It after all have been the lousy products, which have made them to the companies they are today. The ego of power is incompatible with the interests of nature.

Nature works with scalar waves and their velocity of propagation is arbitrary. Wavelength and frequency now can be modulated and information can be recorded separately. In this manner a whole dimension is gained to modulate, the image transmission can take place in parallel, which means considerably faster, safer and more reliable. As anyone of us knows by own experience, assembling the image takes place all at once, the memory of past images takes place ad hoc. Nature is indescribable more efficient than technology with the scalar wave technique.

If we again take the right-hand side of fig. 22.6 with the properties of the Hertzian wave. In the opinion of Nikola Tesla it is a gigantic waste of energy. The broadcasting power is scattered in all directions and the transmission losses are enormous. At the receiver virtually no power arrives anymore. To receive a radio signal the antenna power has to be amplified immensely. It is a wave, which actually only can be used as radio wave, thus as a wave with which arbitrary many participants should be reached.

This wave however is completely useless, if it concerns a point-to-point connection. If I want to call someone, to talk only with him, a radio wave is the entirely wrong method, because I with that bother hunderds and thousands of people that I don't want to call at all!

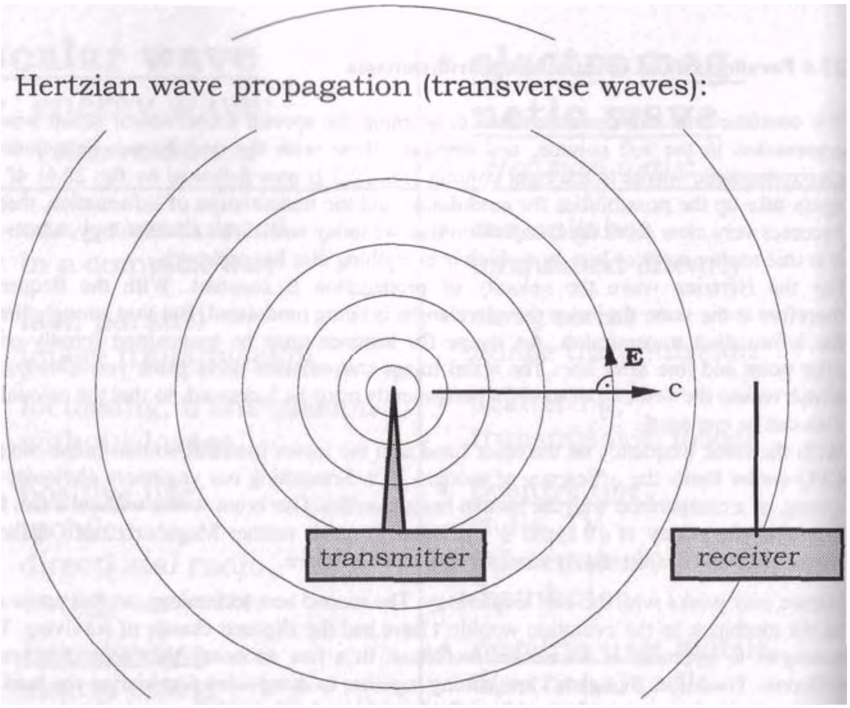
Hertzian wave propagation (transverse waves):
E
c
transmitter
receiver

Tesla radiation (radiations) = scalar wave, longitudinal wave propagation:

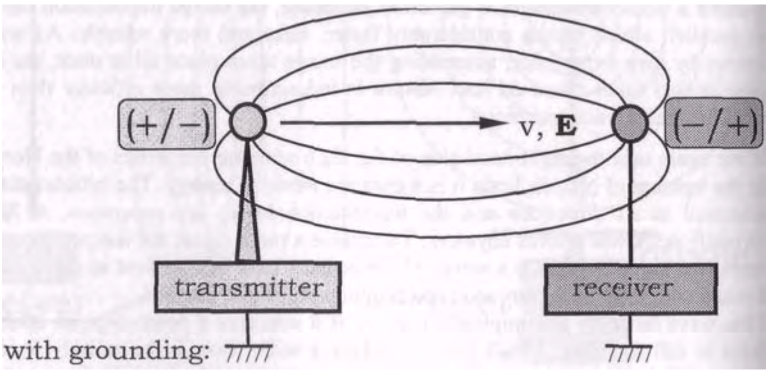
(+/−)
v, E
(−/+)
transmitter
receiver
with grounding:

Fig. 22.7: Comparison of radio waves according to H. Hertz and electric scalar waves according to Nikola Tesla.

22.7 Comparison of the properties

The course of the field lines clarifies the difference. For the scalar wave all field lines going away from the transmitter run together again. As long as no scattering field occurs, there also won't be any transmission losses. It is an energy wave, for which the full broadcasting power is transmitted wirelessly and which arrives at the receiver, by that focussing the field lines again.

Here one at once numerous technical applications come to mind, if energy should be transmitted wirelessly. A TV, which supplies its remote control with energy itself, lelemetric installations, which are fixed at difficult accessible or rotating machine parts and which can't work without energy supply.

For a mobile with 3 Watts of broadcasting power only a few microwatt arrive at the ground station. If I would have a scalar wave mobile, which functions with resonance, then a broadcasting power of some microwatt will be completely sufficient, to carry out a telephone call right through the earth. This minimum broadcasting power suffices, because everything, what is being transmitted, arrives at the receiver - crucial is that the conditions of resonance are fulfilled. That means, both must have the same frequency and the opposite phase. In addition the modulation has to fit, so that on the one hand not several participants in the conversation are getting in each other's way. For a purely carrier wave transmission on the other hand there would be the risk of natural fields being collected also and the power at the receiver taking inadmissibly high values. This is prevented effectively by a correspondingly complex modulation. Nature solves the problem in exactly this manner.

Mobile phone technology with scalar waves of course still is a pie in the sky. A big challenge for the engineers poses the adjustment condition of opposite phase and the fitting modulation, which Tesla called .,individualisation"<i>. Entirely by the way the network runners and the telephone companies are getting quite superfluous.

Telephone charges, so one perhaps can read in the history books in a hunderd years, were the indication of a century of rigorous exploitation of man and nature. With scalar waves a direct, more dimensionally modulated information transmission directly with the partner of conversation is possible and sufficient energy is available to humanity any time and any place, without being dependent on any companies! This notion is not new, but it is inconvenient for the rules; already Nikola Tesla has written about it, but obviously no-one wanted to listen to him<ii>.

Scalar waves are able, that is made clear by the properties, to revolutionize both the energy technology and the information technology fundamentally. It is more than only a technology for the new century. Scalar waves are a chance for the whole millennium!

<i>: N. Tesla: Transmission of electrical energy without wires, Electrical World and Engineer 7.1.1905; Edition Tesla Vol. 4, P. 131.

<ii>: N. Tesla, New York Times of 15.9.1908; Edition Tesla Vol. 6, P. 241.

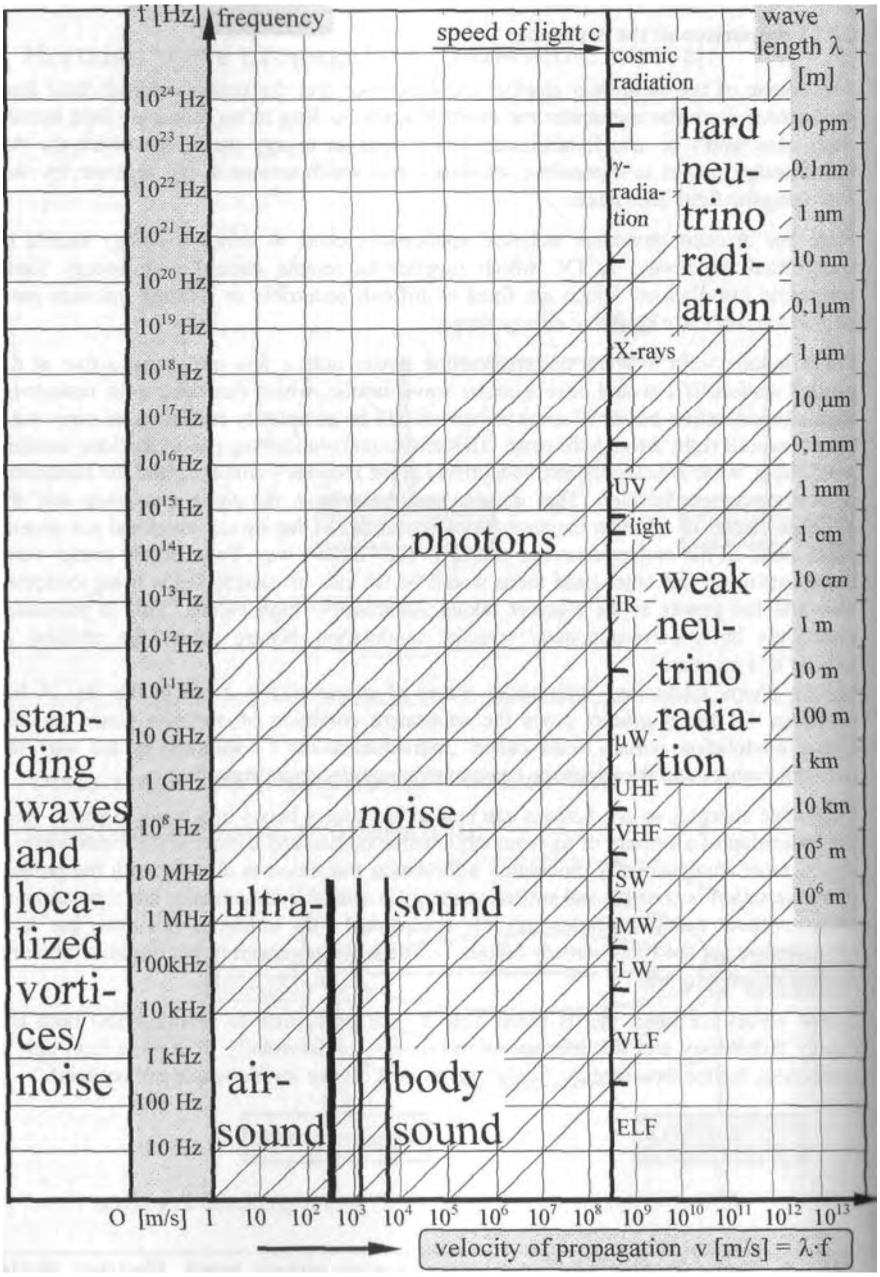
f [Hz]
frequency
speed of light c
wave length λ [m]
cosmic radiation
10^{24} Hz
10^{23} Hz
10^{22} Hz
10^{21} Hz
10^{20} Hz
10^{19} Hz
10^{18} Hz
10^{17} Hz
10^{16} Hz
10^{15} Hz
10^{14} Hz
10^{13} Hz
10^{12} Hz
10^{11} Hz
10 GHz
1 GHz
10^{8} Hz
10 MHz
1 MHz
100kHz
10 kHz
1 kHz
100 Hz
10 Hz
hard neutrino radiation
γ-radiation
X-rays
UV
light
photons
IR
weak neutrino radiation
μW
UHF
VHF
SW
MW
LW
VLF
ELF
10 pm
0,1nm
1 nm
10 nm
0,1μm
1 μm
10 μm
0,1mm
1 mm
1 cm
10 cm
1 m
10 m
100 m
1 km
10 km
10^{5} m
10^{6} m
standing waves and localized vortices/noise
noise
ultra- sound
air-sound
body sound
0 [m/s] 1 10 10^{2} 10^{3} 10^{4} 10^{5} 10^{6} 10^{7} 10^{8} 10^{9} 10^{10} 10^{11} 10^{12} 10^{13}
velocity of propagation v [m/s] = λ·f

Fig. 23.1: Frequency diagram of longitudinal and transverse waves as they depend on the velocity v.

23. Research of scalar waves

Scalar waves are still unexplored area, scientific new ground as it were. Individual research scientists already have selectively ventured forward in this area and have described properties of the scalar wave investigated by them in their special research area mostly in measurement technical manner. But as a rule they lack the physical relation, as it is derived in this book for the first time. If we don't proceed from individual measurements, but from the wave equation and the mathematical physical derivation of scalar waves, then we have the great chance to understand as something belonging together on the one hand noise, photons, neutrinos and lots of other known phenomena as well as on the other hand still unknown phenomena, which are called parascientific. We should remember that we without theory of Maxwell and the representation in a frequency band today still wouldn't know that the radio waves (LW, MW, KW, UHF), the microwaves (μW), the infrared thermal radiation (IR), the light and the X-rays concern just the same phenomenon. The graphic representation of both waves in one diagram in this place is extremely helpful.

23.1 Frequency diagram

In fig. 23.1 the frequency f is shown against the velocity of propagation v with the wavelength λ as parameter. The broad line at $3*10^8$ m/s represents the speed of light c. Here the frequency band of the transverse waves can be found again in the well-known one-dimensional representation.

Crosswise to that, somewhat unusual, the longitudinal waves run. These start at the left at localized noise, over the sound as it propagates in air, in water and in metal, over a large, to a large extent still unexplored, range of the bio photons, the heat vortices and of the dowsing rod effects and end on the other side of the speed of light at the neutrinos. Between that the special case is settled that the particles, or said better vortices, propagating as a scalar wave have exactly the speed of light. It gives reasons for the circumstance, as already mentioned, that light can appear as wave or as photon radiation. It. according to the wave equation, after all always consists of a combination of both forms. At very high frequencies, e.g. the cosmic radiation, this combination is shifted in the direction of vortices and their distribution as a scalar wave, at low frequencies the tendency inversely goes in the direction of the normal wave.

If we assume that for the transverse wave over all frequencies a dozen of specialized gauges is necessary, each of them also can be switched over in range several times, then we can project that to record a scalar wave of a certain frequency over all velocities of propagation likewise 12 devices and for the whole field shown in fig. 23.1 approximately 12x12 = 144 devices will be necessary. Of these 144 gauges today just 12 are available. There thus still are missing 132 pieces, which should be developed. With these gauges, so I am convinced, the many white spots in the diagram can be tapped scientifically little by little if a systematic procedure is used. My vortex theory thus will be attached a central importance.

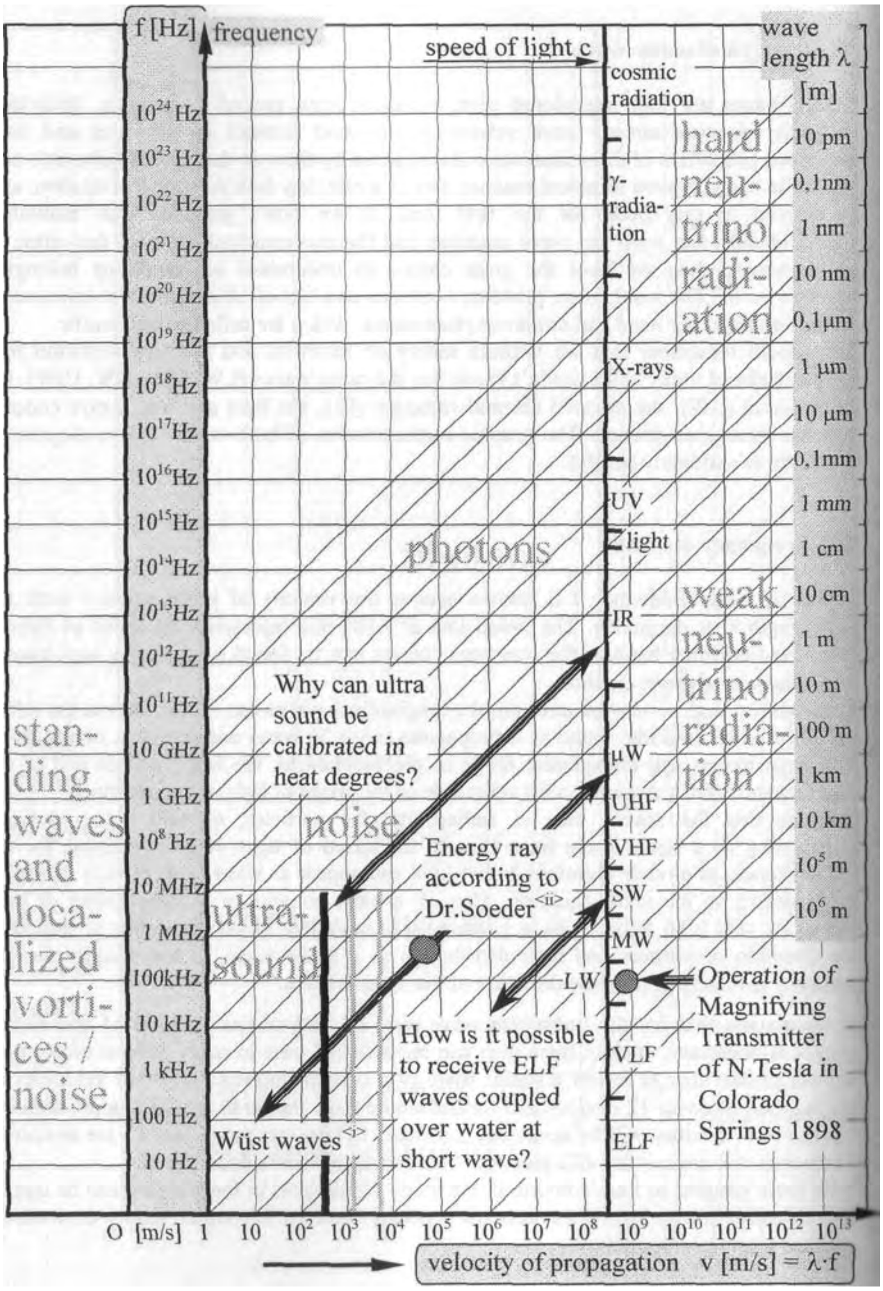

Fig.23.2: Frequency diagram with entries concerning Wust waves,
Tesla transmitter and various not understood effects.

23.2 The reverse of the HF medal

The diagram possibly may settle disputes, like the ones between the „Wustlingen" (note of the translator: the author here is playing with words, Wustlingen literally means lechers) and the „high-frequency engineers".
Professor Wust of Munich already 1934 had proven, that the Wust waves named after him in air have a velocity of propagation of approximately 10 m/s<i>. He investigated them in the frequency range between 16 Hz (λ= 60 cm) and 500 Hz (λ= 2 cm). The high-frequency technicians immediately have converted to the speed of light, since they don't know anything else, and on the spot have shifted the phenomenon into the range of the microwaves between 0.5 GHz (λ=60 cm) and 15 GHz (λ= 2 cm). In the diagram now can be found, at which place this phenomenon belongs actually. It presumably concerns the same „energy ray", which Dr. Soeder has observed and proven already at velocities of propagation of 55.2 km/s and wavelengths around 10 cm<ii>.

Furthermore follows from the diagram (23.2), why ultra sound can be calibrated in heat degrees, and why radio reception of signals from a completely other frequency range is possible by principle, if for identical wavelength a velocity of propagation different from c occurs. Further the range of operation used by Tesla at approx. 1.6 times the speed of light is depicted.

Also the noise is such a book with seven seals. It concerns, it is true, a fixed chapter of high-frequency engineering, but without visible connection to the other chapters. That might be based on the circumstance that the relation, as it is dictated by the wave equation, isn't recognized.
Every high-frequency signal is accompanied by a noise signal; every antenna produces more or less noise. HF engineering dictates the measures, how the noise can be suppressed resp. the signal-noise distance be increased. The goal is to make the electromagnetic wave stand out in such a way from the noise that it can be received. For that the measurement setup must be choosen correspondingly, are the measurement cables dictated, and must be paid attention to power adaptation and corresponding termination resistors. For measuring HF correctly all interference influences should more or less disappear.

But what are the interference influences, which the HF-technician suppresses? According to the wave equation it are the scalar wave parts, and to that also is counted the noise. Every HF-technician thus knows the scalar parts as interfering noise signal, but doing so he completely fails to notice the technical advantages and chances, which e.g. are present in the noise.
If by means of an autocorrelation function a noise signal is compared to itself then often a hidden message comes to light. According to that, informations can be hidden in the noise and of course transmitted wirelessly. This important circumstance is known, but it is hardly used. The noise vortices thus can be modulated in an extremely complex way. There can be transmitted a lot more information in the noise as scalar wave, than with the radio wave.

<i>: J. Wust: Physikalische und chemische Untersuchungen mit einem Rutenganger als Indikator. Further references in W-B-M 3/ 1991, S. 57.

ii>: A. Soeder: Ohne Strom lauft neuer Energiestrahl rund urn den Erdball. Raum & Zeit 59/92 Seite 62-67.

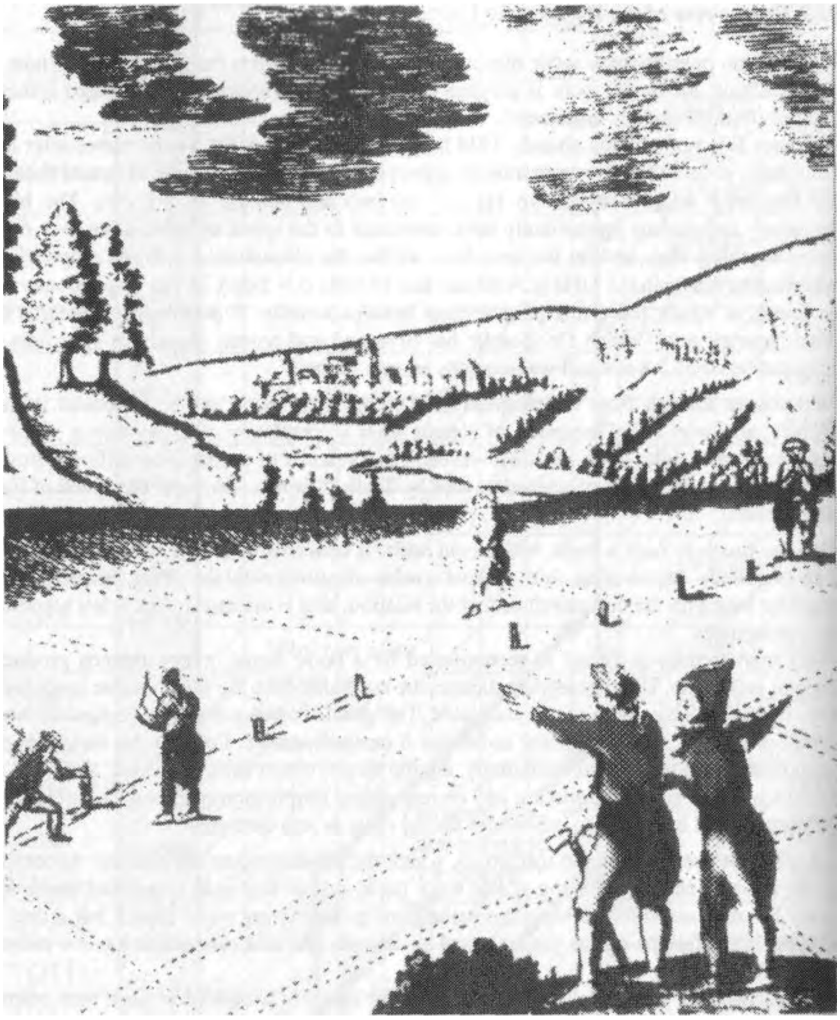

Fig. 23.3: ___ Dowsers dowse an ore vein.[i]
From: Speculum metallurgiae politissimum, Rossler, Dresden 1700.

---

[i]: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, S. 145, ISBN 3-927407-09-7.

The possibility of more dimensionally modulating and parallel image transmission make scalar waves superior to the radio waves in such a way that one should seriously think about the reversed procedure, in which the noise is conceived as useful signal and the radio wave as interference signal.

In the frequency diagram (23.1 and 23.2) can be read like in a diary, in which most pages it is true still are empty, but in which some very informative marginal notes can be found for different points of difficulty. It mostly are parascientific phenomena, which, not understood and excluded by textbook physics, are waiting to be taken up in the building of physics. As an example we'll pick out an extensive area of alternative research.

23.3 Radiesthesia

A particularly broad spectrum of scalar waves is provided by the already several times mentioned earth radiation. An ancient science of experience, which troubles the exploration of the earth radiation, is the geomancy. The Roman land surveyors, the Augures, used as an aid a flat coil like also Tesla did to receive scalar waves. The Lituus, as the device was called, resembles so much that of the Etrusks that we must proceed from the assumption that the method is much older (fig. 16.10).
This part of the scalar wave research, also called radiesthesia, is derived of „radiare", which can be translated with "send out rays" resp. „perceive". It describes the doctrine of the sensitivity to radiation of man. Doing so the radiation sensitive uses his own nerve framework as a biosensor.

The nerve conduction could be derived as a biological variant of the Tesla one wire transmission, for which ring-like potential vortices are passed on as action potentials in form of standing waves (fig. 9.6). Of this kind are also the control commands, which cause a muscle to contract. If now corresponding vortices are picked up by means of a dowsing rod or similar aids in our nerves, then the contract addressed muscles contract, because they can't distinguish, from where the command comes. This unconscious nerve twitch leads to a swing of the dowsing rod and to the well-known dowsing rod phenomenon. But man can't replace a technical gauge. Hence one speaks of dowsing and not of measuring. To this should be added the condition of resonance, which must be fulfilled. Since every person however builds up other resonances, dowsed results of others often can't be reproduced. But from this particular difficulty one cannot draw the conclusion that the phenomenon does not exist and radiesthesia is not a science. Series surveys and statistical analyses here in any case don't lead any further. It always only are individual talented dowsers, who have at their disposal really fantastic abilities and find with great certainty water, ores and even oil.
With the perception of a physical phenomenon it mostly starts. Cultural man looks after his discovery as cult, whereas modern man, guided by the wish for reproducibility and more objectivity, is troubled to design and to build a technical measurement work. With regard to the scalar wave in general and the radiesthesia in special we still are at the stage of the Stone Age.

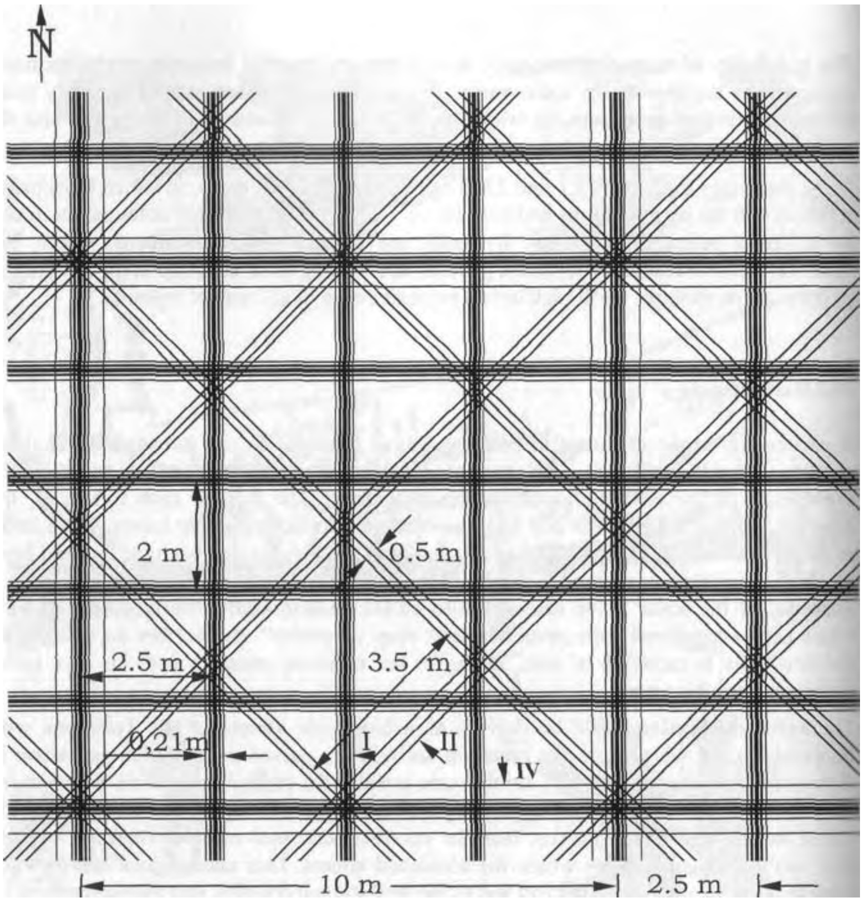

Fig. 23.4: Nets dowsed at latitude 49.<i>

I. net: ,,global net" according to Dr. E. Hartmann;
orientation: N - S: 2 m distance and E - W: 2.5 m distance

II. net: ,,diagonal net" according to Dr. M. Curry;
orientation: NE - SW and NW - SE: 3.54 m = 2.5 m * sqrt(2)

[III. ,,lightning net" according to R. Schneider; orientation like I.]

[IV. net acc. to Benker; like I. but: N - S: 10 m and E - W: 10 m]

[V. sun/planet net acc. to W. Auer; with 3.66 m orientation S./P.]

According to Prof. E.G.Hensch et al.:

<i>: Hensch: Radiaesthesie im landlichen Bauen und Siedeln, Arbeitskreis zuf Landentwicklung in Hessen, W4, Wiesbaden 1987; resp.: W. Auer: Erdstrahlen?... AWGEO 199, Eigenverlag 1998; resp.: Mayer/Winklbaur: Biostrahlen, 5. Aufl. ORAC Verlag, Wien 1989, Seite 168

23.4 Prediction of earthquakes with the help of scalar waves

The standing wave character of the earth radiation is a help. There are dowsed points and lines of maximum radiation intensity, which form nets, which encompass the whole world. On the one hand we are dealing with a net oriented in direction north-south and east-west (Hartmann net) and on the other hand with a net standing diagonally to that under 45° (Curry net). Because the angle of 45° can be derived as borderline angle from a vortex field calculation<ii>, I already early have pointed to the circumstance that it here presumably concerns a vortex phenomenon<iii>.
Because the nets in addition are dowsed in air, it must be vortices of the electric field, so-called potential vortices<iii>, which here form regular structures. The formation of vortex lines and complete vortex streets (chapter 4.9), which consist of countless individual vortices, can be explained as follows:
Electric scalar waves propagate in the direction of the electric field strength and mediate field vortices, e.g. neutrinos. If at a certain moment the transmitter carries a positive and the receiver a negative charge, then all the particles which are positively charged are repelled by the transmitter and attracted by the receiver. All run at the same time towards the same goal, although all mediated particles carry the same charge and repel each other! This incompatibility can be compensated partly, by the vortices rotating around each other. In this circumstance can be seen the reason for the structure shaping, the formation of some lines in the countryside (fig. 23.3). The distances between the lines have characteristic values, which allow conclusions about the wavelength of the standing waves. We must assume that they dictate the structure shaping and that the spatial vortex distribution aligns with the nodes and antinodes of the respective standing wave. The distance between the lines, which corresponds to half the wavelength, becomes smaller and smaller towards the North Pole and the South Pole of the earth, the net thus narrower and narrower. Also is the net said to change strongly before an earthquake. This all are clues for the circumstance that the structure shaping radiation comes from the earth, that the cause must be sought in the earth radiation.
It should be noted marginally that with scalar wave detectors, which permanently scan the nets, a just as effective as inexpensive earthquake prediction should be possible. Such a facility would be an enormous relief and help for all earthquake warning services. It even should be possible to determine in advance the future epicentre, if there is measured at the same time at if possible many stations and the respective deviations are compared.

A further influential factor is the composition of the subsoil; e.g. ores and metals influence the earth radiation. Water shows a special affinity for the earth radiation. It does collect the radiation and after bundling it up releases it again. To blame is the high dielectricity of water ($\varepsilon \cong 80$), which again favours the formation of potential vortices .
A technical use of this effect would be the neutrinolyse (see fig. 17.6-17.8 and 18.1), the splitting of water molecules by neutrinos if these take the state of an electron, hydrogen escapes and the oxygen content in the water increases. If a neutrino however shows as a positron, then it annihilates and there is formed a light flash, which serves the „experts" in neutrino detectors as proof.

<ii>: Meyl, K.: Wirbelstrome, Dissertation Universitat Stuttgart 1984

<iii>: Meyl, K.: Potentialwirbel Band 1, INDEL Verlagsabt. 1990

Spectroid of the radiation field of a subterranean flowing water veign (according to P. Schweizer)<i>:

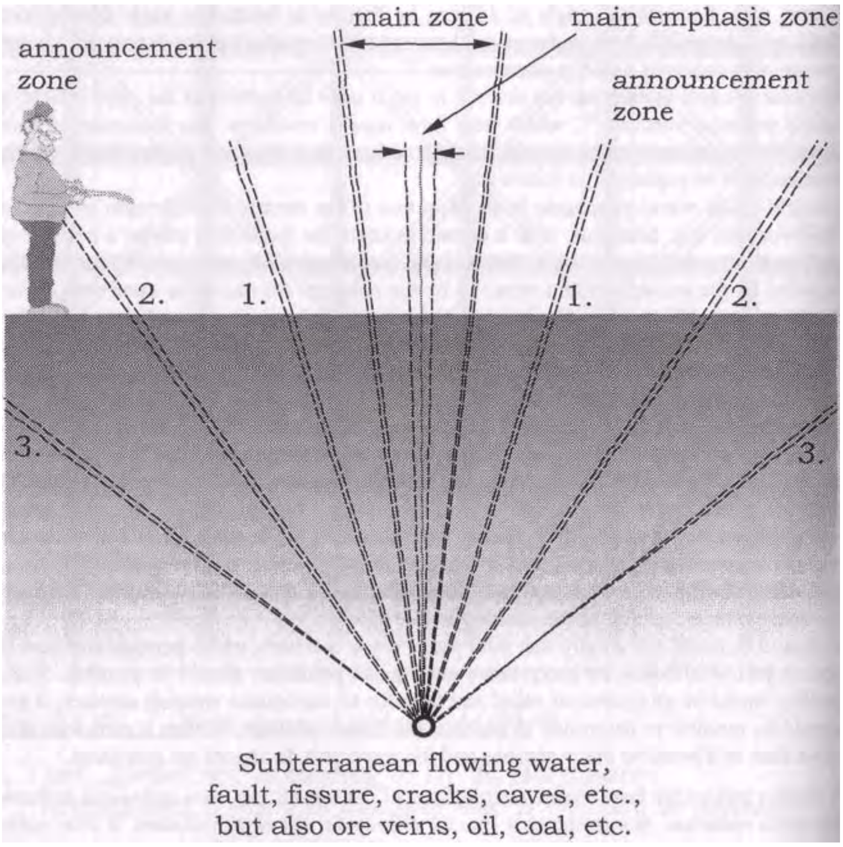
main zone
main emphasis zone
announcement zone
announcement zone
2.
1.
1.
2.
3.
3.
Subterranean flowing water, fault, fissure, cracks, caves, etc., but also ore veins, oil, coal, etc.

Fig. 23.5:____ Radiaesthetic search for water<i>,
(dowsing of the earth radiation)

<i>: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994 S. 145, ISBN 3-927407-09-7.

23.5 Searching water with the help of scalar waves

Faults of earth's crust lead to fissures and cracks in the rock, which often fill up with standing or flowing water. Potential vortices of the earth radiation are attracted by the high dielectricity of water and are radiated again in slowed down and focussed form. On the surface of the earth the mixture of various scalar waves arriving there can be dowsed as main emphasis zone. For reason of the broad radiation spectrum the results of proficient dowsers in the main emphasis zone more often coincide.
At both sides in addition spectroids are forming, which are dowsed as parallel running vortex lines. Because every harmonic wave produces other lines and every dowser reacts to other resonances, a profound ability is required, if from the distance of the so-called announcement zones the water depth should be inferred (fig. 23.5). Some specialists indicate besides the place and the depth also the amount of deposit and if need be the water quality. For that they analyse the radiation intensity, but unfortunately many all too often overrate their abilities.

Already the Augures, the land surveyors in the age of the Romans have aligned their streets and the castles with the nets and the lines in the countryside. Even today new evidence about the central importance of the standing wave character of scalar waves can be found at excavations (fig. 16.10). Since every scalar wave also occurs coupled with a radio wave by means of the wave equation, earth rays originally could be detected conventionally with field strength gauges. But the intensive use by radio stations has made it necessary to change from the short wave range to the VHF-range and further to the UHF-range. It were indirect measurements, which could be interfered easily. Today geologists work with ELF scanners in a range between 15 and 30 KHz if they search for water. They analyse the emitted signals of submarine transmitters, because these are attracted and amplified by water carrying layers.
Experiments also are carried out with noise receivers or unused broadband TV channels. Others analyse influences of the earth radiation on the magnetic field of the earth, but for this method the expressiveness is controversial.
Again others walk along a path with a scintillation counter measuring the inverse profile. At the places of maximum earth radiation a minimum of gamma radiation is measured. This stunning relation only can be interpreted as follows: the natural radioactive decay at these places takes place accelerated and that proves:
1.. that earth radiation sets radioactivity free and causes it in the first place (chap. 17.1 pp).
2. that in the case of earth radiation it actually concerns slowed down neutrino radiation.
But also this method works indirectly. It only functions in the open air and then only in the case of unspoiled nature, which sometime was covered with an even layer of radioactive dust. By cultivation or plantation the „radioactive layer" very often is changed and mixed up, whereupon the measurable profile hardly allows conclusions.
For indirect measurement methods of that kind, for which the measurement variable stands in a relation, which isn't known in more detail, to the scalar wave of interest, generally caaution is advisable. Too often a message is seen in a wonderful 3-D diagram of the magnetic field strength or the radioactivity distribution, which has much more mundane causes.

Lecher air antenna according to W. Busscher<iii>:

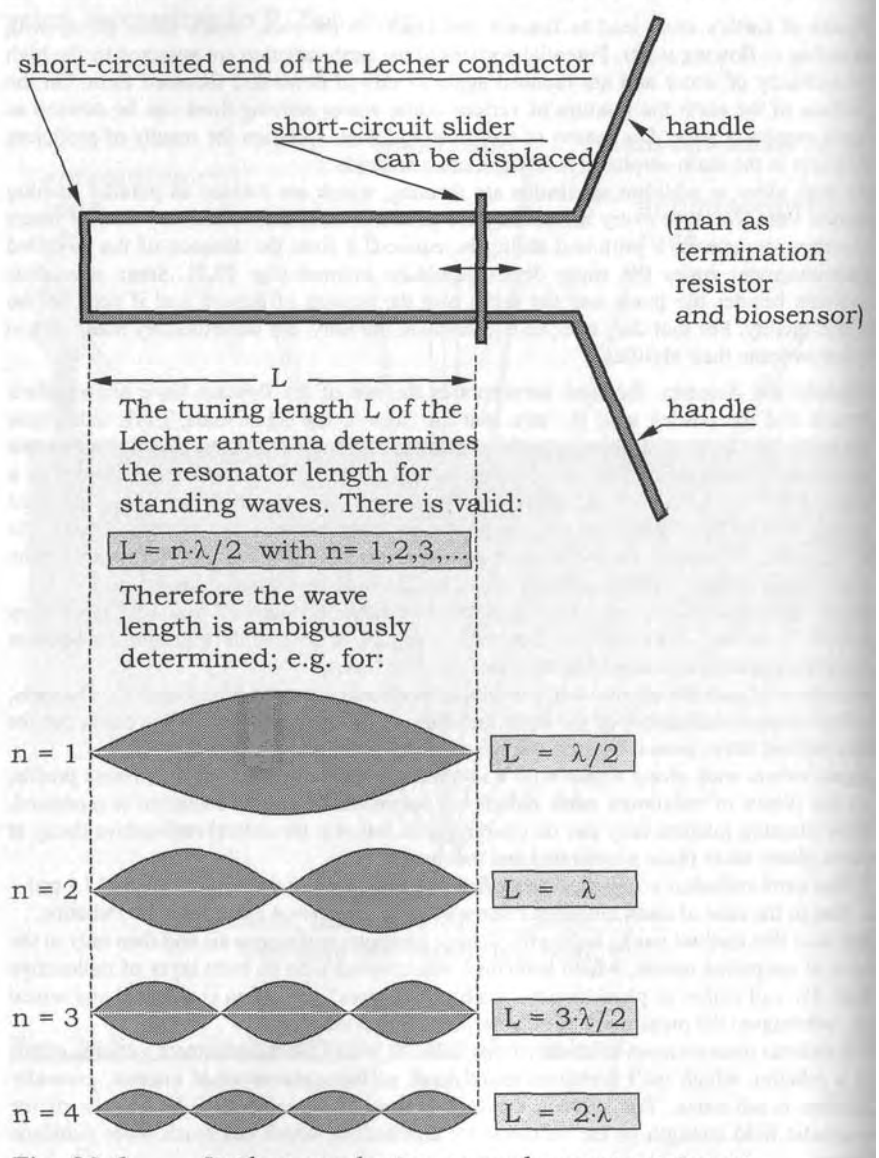
short-circuited end of the Lecher conductor
short-circuit slider
can be displaced
handle
(man as
termination
resistor
and biosensor)
L
handle
The tuning length L of the
Lecher antenna determines
the resonator length for
standing waves. There is valid:
$L = n \cdot \lambda/2$ with n= 1,2,3,...
Therefore the wave
length is ambiguously
determined; e.g. for:
n = 1
$L = \lambda/2$
n = 2
$L = \lambda$
n = 3
$L = 3 \cdot \lambda/2$
n = 4
$L = 2 \cdot \lambda$

Fig. 23.6: Lecher conductor as scalar wave antenna.

Can be read at Prof. Dr. Ing. Christof Rohrbach:

<iii>: Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996, Seite 100.

23.6 Receiving scalar waves with the Lecher antenna

In the case of technical gauges it is normal to change the frequency tuning. For that the capacity of a capacitor is being varied, which is part of a resonant circuit. This method can simply be realized with the today available construction elements. Biological systems on the other hand, for instance a person, work primarily with a variation of the wavelength. In this case (according to eq. 22.1) the frequency is directly proportional to the velocity of propagation.

So provide for instance radiaesthetic dowsings clues for the wavelength. The frequency however can't be given, because that oscillates along with the velocity of propagation and at present the forming noise signal still can't be analysed. There can't be bought electronic construction elements, which like a body cell would be able of oscillations of length. Oscillation quartzes and piezo elements, on the basis of which scalar wave detectors could be constructed, form an exception.

Similar to the Lecher conductor different forms of tunable Lecher antennas have been developed. which are used for dowsing. Such a frame antenna, developed strictly according to the rules of high-frequency technology stems from W. Busscher (fig. 23.6)<i>. With the short-circuit slider a closed-loop antenna circuit in the upper part is tuned to half the wavelength or an integer multiple of it ($L = n \cdot \lambda/2$ with n = 1,2,3...). The person, who holds the Lecher antenna with both hands, forms the termination resistor for the attached antenna circuit. At the same time he prevents the occurring of an effect back on the tuned resonant circuit by means of damping of the Lecher conductor. The sense of this arrangement is lying in the circumstance that man as a „biosensor" should feel if standing waves, which I call vortices, are formed in the Lecher antenna.

A modification of the Lecher air antenna is the Lecher dowsing rod according to R. Schneider<ii>. It is produced as an etched board, which has a dielectricity of approx. 4. As a result the velocity of propagation and the wavelength are only half their normal value. In addition a shortening factor V is introduced, which depends on the construction of the Lecher dowsing rod<iii> ($L = V \cdot n \cdot \lambda/4$ with V = 0.952 and n = 1,2,3...).

In fig. 23.7 some published tuning values have been drawn up. The values for want of corresponding technical gauges it is true still haven't been confirmed, but they allow a certain insight into the world of the scalar waves, as they are influenced and radiated by water, by metals, by oil or even by other planets, the sun and the moon. Doing so some questions remain open: We for instance don't know, if the corresponding wavelength concerns the basic wave or only an m^{th} harmonic wave. The details not only are unreliable, they moreover also are ambiguous. If however several tuning values L' are present for which resonance occurs, then if need be an integer divisor can be sought and the wavelength can be determined. But that doesn't function always.

<i>: Will Busscher: Die Luft-Lecher-Leitung als Wunschelrute. Eine Methode um die Wellenlangen zu bestimmen. Wetter-Boden-Mensch 3/ 1991, Seite 46-57.

<ii>: R. Schneider: Einfuhrung in die Radiasthesie Teil II: Leitfaden und Lehrkurs ..., Oktogon-Verlag, Wertheim, 2. Aufl. (1984) S. 136.

<iii>: Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996, Seite 100.

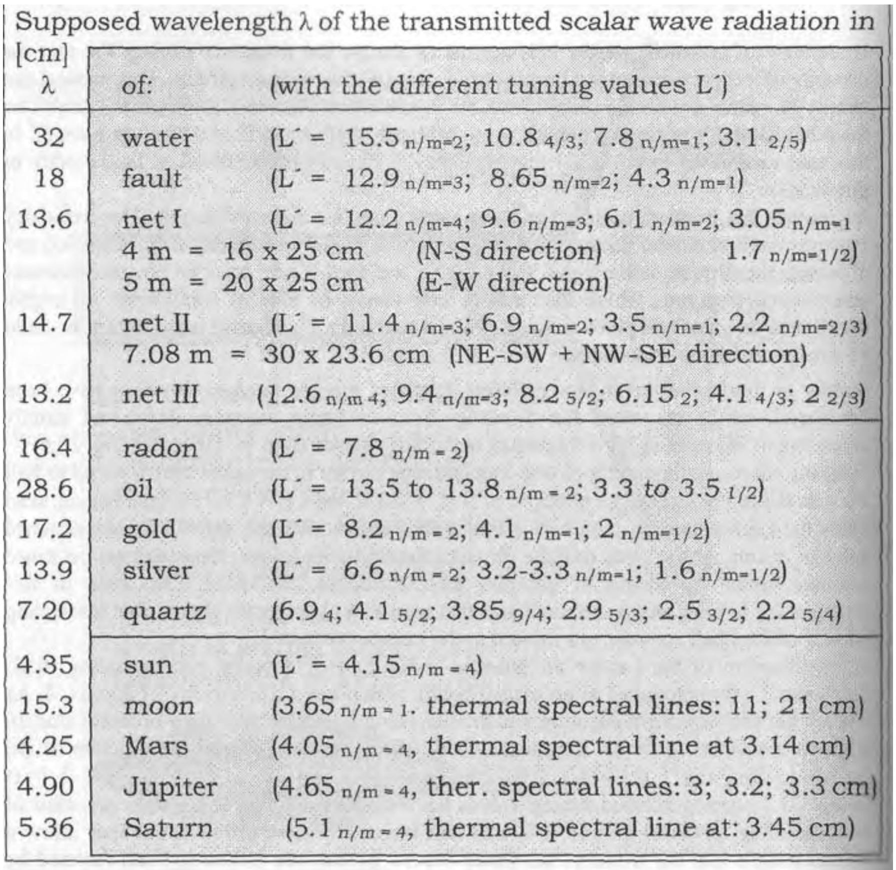

Supposed wavelength λ of the transmitted scalar wave radiation in [cm]

λ	of:	(with the different tuning values L´)
32	water	(L´ = $15.5_{n/m=2}$; $10.8_{4/3}$; $7.8_{n/m=1}$; $3.1_{2/5}$)
18	fault	(L´ = $12.9_{n/m=3}$; $8.65_{n/m=2}$; $4.3_{n/m=1}$)
13.6	net I 4 m = 16 x 25 cm (N-S direction) 5 m = 20 x 25 cm (E-W direction)	(L´ = $12.2_{n/m=4}$; $9.6_{n/m=3}$; $6.1_{n/m=2}$; $3.05_{n/m=1}$ $1.7_{n/m=1/2}$)
14.7	net II 7.08 m = 30 x 23.6 cm (NE-SW + NW-SE direction)	(L´ = $11.4_{n/m=3}$; $6.9_{n/m=2}$; $3.5_{n/m=1}$; $2.2_{n/m=2/3}$)
13.2	net III	($12.6_{n/m\,4}$; $9.4_{n/m=3}$; $8.2_{5/2}$; 6.15_{2}; $4.1_{4/3}$; $2_{2/3}$)
16.4	radon	(L´ = $7.8_{n/m=2}$)
28.6	oil	(L´ = 13.5 to $13.8_{n/m=2}$; 3.3 *to* $3.5_{1/2}$)
17.2	gold	(L´ = $8.2_{n/m=2}$; $4.1_{n/m=1}$; $2_{n/m=1/2}$)
13.9	silver	(L´ = $6.6_{n/m=2}$; 3.2-$3.3_{n/m=1}$; $1.6_{n/m=1/2}$)
7.20	quartz	(6.9_{4}; $4.1_{5/2}$; $3.85_{9/4}$; $2.9_{5/3}$; $2.5_{3/2}$; $2.2_{5/4}$)
4.35	sun	(L´ = $4.15_{n/m=4}$)
15.3	moon	($3.65_{n/m=1}$. thermal spectral lines: 11; 21 cm)
4.25	Mars	($4.05_{n/m=4}$, thermal spectral line at 3.14 cm)
4.90	Jupiter	($4.65_{n/m=4}$, ther. spectral lines: 3; 3.2; 3.3 cm)
5.36	Saturn	($5.1_{n/m=4}$, thermal spectral line at: 3.45 cm)

Fig. 23.7: Details of not confirmed wavelengths.<i>
(derived from tuning values of a Lecher antenna
L´ = . V·λ·n /4·m with V = 0.952
and n = 1,2,3... for λ/4, λ/2, 3λ/4, λ, ...
and m = 1,2,3... for 1 the basic wave
and the 2nd, 3rd, ... harmonic wave

---

according to:

<i>: Ludeling, Hartmut: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, S. 161 ff. The author gives the clue: ,,The values have been determined empirically by different persons and predominantly could be confirmed by the author and his co-workers. There however always can occur deviations caused by different dowsing methods. Every value therefore always should be checked with own measurements."

23.7 Assignment according to wavelength

As said, the methods to determine the wavelength by means of the Lecher antenna are ambiguous and unreliable. So water should have a wavelength twice as long as a fault. But since water often is collecting or subterranean flowing in faults, one is entitled the suspicion that here $\lambda/2$ was taken, where in reality the full wavelength λ acts. The rest of the table after that should be looked at with the same scepticism. The table nevertheless is not wholely uninteresting, because it points to certain trends and is able to furnish clues, which if need be can represent a help of orientation in the systematic research of scalar waves.
At the planets is remarkable that the Lecher conductor reacts to values, which correlate with the respective thermal spectral lines. If these spectral lines propagate with c and it actually should concern the same cause, the frequency thus should be identical, then the planets would emit scalar waves, which are faster than the light. In the case of Mars 1.35 times the speed of light c would be present, in the case of Saturn 1.55 times and in the case of Jupiter the values would lie between 1.46c and 1.63c. In the case of the moon two thermal spectral lines are being measured, which on the one hand have as a result 1.39c and on the other hand 0.73c. At this place still ample research tasks should be solved.

In the case of scalar waves the wavelength is the most important factor to refer to. The frequency however varies continually, which in electrical engineering is called noise. That again is connected to the circumstance that frequency and velocity of propagation oscillate, so that merely an average value can be given at a time. But that hardly can be measured, which again makes the assignment more difficult. We thus still aren't capable to enter the table values from fig. 23.7 into the frequency diagram of fig. 23.1.

A technical gauge for scalar waves is necessary. A practical solution could look as follows: A noise transmitter tunable in frequency and wavelength operates on top of a carried along noise receiver. The arrangement with that comes the Tesla transmission path for scalar waves very close. If the transmitted noise signal hits upon a likewise one in the surroundings then overlapping occurs, which at the receiver causes a change of the displayed value. If doing so a subtraction (extinction) or an addition (amplification) of the signals occurs is unimportant.
With this arrangement in any case statements are possible about frequency, wavelength, velocity of propagation and about the amplitude of the scanned signal, without withdrawing energy and strain the signal in doing so<i>. With such a gauge radiation conditions depending on location could be measured as well as technical devices checked lor the emitted scalar wave parts (fig. 24.1).
An important use over and above that would be given in medical diagnostics. Every living being „produces noise" the technician would say, it „emits scalar waves" I would say, whereas following general usage is talked about the „aura of man". The value of an aura diagnosis still is completely unknown to most doctors and therapists, especially as the scanning of the aura at present only is possible by dowsing. But the patient expects that a doctor works with a technical gauge and not with a dowsing rod!

<i>: First experiments, which I carried out with students in the laboratory, look very promising. Unfortunately the works at present rest for reason of lack of money.

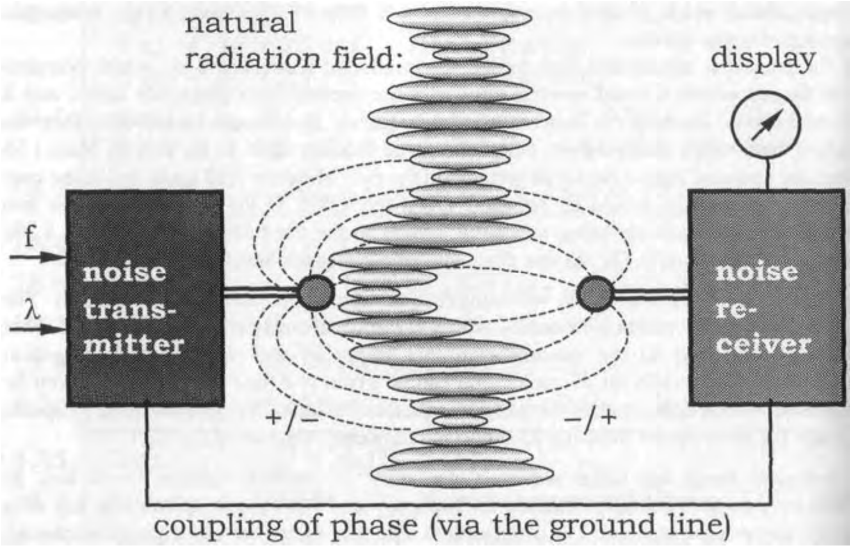
natural
radiation field:
display
f
λ
noise
trans-
mitter
noise
re-
ceiver
+/−
−/+
coupling of phase (via the ground line)

Fig. 24.1: Concept for a device to detect scalar waves

24. Medicine of oscillation

24.1 Mediation of information with the help of scalar waves

If for the discussed scalar wave gauge the transmitted comparison signal in addition deviates from the form of a sinus, if the signal is modulated in some way, then even the transmission of information by means of the noise as scalar wave can be realized. If we as an example again take the aura of man, which is more than only a radiation field, it carries information for instance about the state of health of man. Kirlian photography is one form of making it visible (Fig. 3.6).
Not only the nerve conduction and the brain work with scalar waves. Man in this way also corresponds with his fellow people, some more, others less. There are people who „beam" and that should be interpreted in the true meaning of the word. Others opposite to that are more „receiver characters", who pick up more scalar wave radiation than they give away. But because they not only pick up positive radiation energy, they are susceptible to information, which makes sick. That's why a healthy equilibrium in the exchange of thoughts, feelings and the different forms of scalar waves is very important. All that is reflected in the aura of man. If we technically are capable to scan the aura of man, than this can be a blessing for humanity and for the public health.

The possible abuse however must not be overlooked. Unbiased research results, which can be used for the benefit of humanity, in many cases just as well can be used to harm humanity.
A connector of both these worlds is the Russian psychologist Prof. Smirnov, who has shown publicly in television, how a spoken sentence can be transformed in a noise signal, which is taken up directly by the brain as information, with the help of a computer program <i>. He with that is capable to "sodden" individuals, as he says, he in this way can take the fear away from soldiers before a combat mission and can operate the disease out of drug addicts without bloodshed. In the television film such an „operation at the open subconscious"' is shown life. The patient hears the noise signal over headphones and is cured already after few minutes treatment time.

We here are getting in the domain of ethical and moral problems of scalar wave research, which aren't solved by us looking away and leaving the field up to others. According to the words of Prof. Smirnov the only thing, which can stop the research scientist, is his own moral. He doesn't say any more and that isn't exactly reassuring! One here is working with „signals resembling sound", is said in the report of the Zweite Deutsche Fernsehen<i>, thus with longitudinal waves and that shows, that already more knowledge about scalar waves is present than is generally known by the masses.

Another way is leading over the bioresonance. Also here at first the possibility and chance to cure diseases is to the fore. The bioresonance is a central aid in the area of the medicine of oscillation, which is increasing in importance permanently.

<i>: Die Zombies der roten Zaren (Besta Film Warschau im Auftrag d. ZDF 1998)

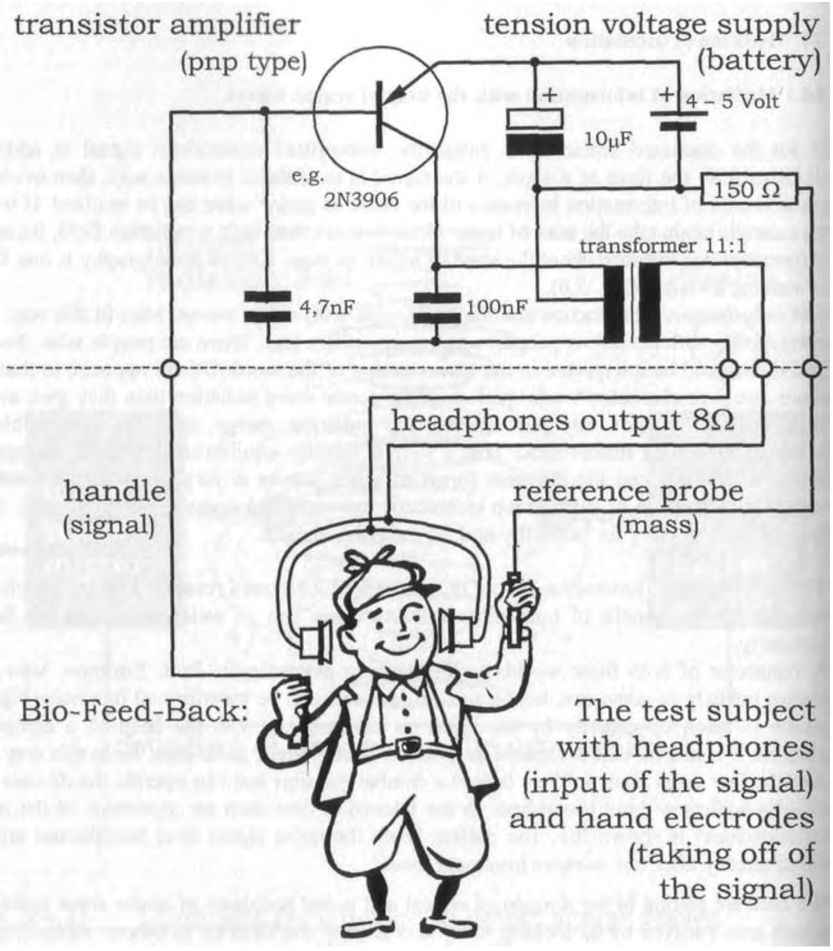
transistor amplifier
(pnp type)
tension voltage supply
(battery)
+
+ 4 - 5 Volt
10µF
e.g.
2N3906
150 Ω
transformer 11:1
4,7nF
100nF
headphones output 8Ω
handle
(signal)
reference probe
(mass)
Bio-Feed-Back:
The test subject
with headphones
(input of the signal)
and hand electrodes
(taking off of
the signal)

Fig. 24.2 A: Wiring diagram of the Syncrometer.<i>

<i>: Wiring diagram taken from: H. Clark: The Cure for All Cancers (1993) New Century Press, USA, Page 434

<ii>: Flanagan: Horen mit der Haut anstatt mit den Ohren, Sonnenwind 1/97, S.13-17.
s.a.: N.Begich: Towards A New Alchemy, The Millenium Science, Earthpulse Press 1996

<iii>: Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997), Kap. 11.6 Der individuelle Grundton, S. 239 ff.
s.a.: Dr. Bodo Kohler: Die bipolare Farb-Ton-Therapie, CO'Med 2/2000, S.10 - 15

24.2 Systems with positive feedback

The field of oscillation of man easily can be influenced, because it concerns a vortex field. A technical measurement with the gauges available at present is almost impossible and highly inaccurate. There however exists the possibility, to apply electrodes to a person and to integrate him in the circuit of a technical device. The two of them form a feedback system with man in the return loop. The operation can take place in two different ways, depending on the sign of the feedback (Fig. 24.3):

For positive sign it concerns positive coupling. In this case the signals released by a person sum up, for which reason already minimum amplitudes are sufficient to produce violent reactions in the case of resonance. For resonance to occur, the system must either search independently the suitable frequency and phase, or the therapist searches the points of resonance.

24.2.1 The Syncrometer.

An independently working system is e.g. the Syncrometer. By means of two electrodes, which are held with the hands or are attached directly to the head, the noise signal emitted by a person is called off and supplied to a broadband amplifier. The test subject again hears the amplified signal via headphones, so that the circuit is closed.

Apparently unspecific and still reproducible during operation certain signals capable of resonance are emerging stronger by amplifying and swinging themselves up. The aura starts to extend or expressed in the words of a radiation sensitive, the so-called reaction distance is increasing, with which he means the distance in which he detects a dowsing rod reaction.

The test subject now increases his radiation, so this method can be valued. By strengthening his own radiation power, he works more as a transmitter and less as a receiver for scalar waves. In the last point the therapeutic use seems to iie, because by that the patient can get rid of resonances to any unpleasant persons or to technical transmitters, which are burdening him. The device consequently also is sold as aid or protection against electrosmog.

The amplified signal also can be picked up directly by the skin instead of by the headphones and still being „heard". The reports stretch from a harmonizing and balancing effect up to "electronic telepathy" and states similar to ecstasy of the test subject<i>.

24.2.2 The sound therapy.

The controlled variant of this method for instance is the sound therapy. Now the sounds, which the patient picks up by means of headphones, do not stem from him but from a sound generator. The therapist goes through the scale and tries to find out at which sound the aura swings up. If the eigenfrequency is found, then the patient can therapize himself, by again and again playing or humming his eigentone<iii>.

It concerns here only a physical statement on the used methods of alternative medicine and not therefore, whether healing successes are actually possible.

sound frequency equally tempered [Hz]		wavelength [cm]	colour	wavelength [run]
F sharp"	740	46,4	(purple)	-
F"	698	49,1	(crimson)	-
E"	659	52,0	(violet)	-
D sharp"	622	55,1	indigo	421
D"	587	58,4	bleu	446
C sharp"	554	61,9	turquoise	472
C"	523	65,6	green	500
H'	494	69,5	yellowgreen	530
A sharp'	466	73,6	yellow	562
A'	440	78,0	orange	595
G sharp'	415	82,6	orangered	630
G'	392	87,5	red	668

Table 24.2 B: Scale of the colour range.<i>

colour		chakra	sense organ	planet type
purple	7.	(crown chakra)	epiphysis	Jupiter
indigo	6.	(brow chakra)	hypophysis	Saturn
bleu	5.	(throat chakra)	hearing	Mars
green	4.	(heart chakra)	feeling	Sun
yellow	3.	(solar plexus)	seeing	
orange	2.	(sacral chakra)	taste	Venus
red	1.	(root chakra)	smell	Moon

Table 24.2 C: Assignment of colours to chakras and planets.<i>

colour		metals	gems (selection):
purple	7.	tin	amethyst, fluorite
indigo	6.	lead	indigo-sapphire, azurite
bleu	5.	iron	sapphire, lapis lazuli
green	4.	gold	emerald, malachite, jade, gr. tourmalin
yellow	3.		topaz, amber, citrine
orange	2.	copper	carnelian, fire opal
Rot	1.	silver	ruby, coral, garnet, red jasper

Table 24.2 D: Assignment to a selection of gems.<i>

<i>: Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997), Kap. 11.6 Der individuelle Grundton, S. 190 - 194

24.2.3 The colour therapy.

Analogous to acoustics also colours each time can be assigned a wavelength and also here can be detected that every person responds to certain colours, thus wavelengths, in particular which can stimulate his aura to resonant oscillations. In practice sound and colour therapy often are used coupled<i>, even succeeds more or less an assignment to each other. That in last consequence leads to a "scale of the colour range" (table 24.2 B)<ii>. According to statement of the treating doctors does a patient, who responds to a certain colour, also react to the corresponding sounds and vice versa.

24.2.4 The aroma therapy.

If according to that it depends less on the circumstance if sounds or colours are used, but the wavelength of an oscillation is crucial, then a stimulation also should be possible by means of the remaining sense organs, e.g. the nose, the tongue or the skin. The smell after all already could be identified as vortex information (fig. 9.0). Fragrances of natural essences can significantly influence the frame of mind. But as long, as we still haven't understood the physics behind it, we technically hardly are capable to generate equivalent vortex modulations artificially.

24.2.5 The gem therapy.

Already the holy Hildegard von Bingen (1098-1179) knew and used the beneficial effect of gems. Physical background of this at first purely empirical form of therapy is the characteristic eigenfrequencies of the gems, which are picked up as stimulation over the skin. Because gems represent a mixture of various molecules, the oscillations in the atomic hull will overlap, so that overlapping and beat frequencies can form with wavelengths in biological relevant areas. There even can form modulations, which are carrying information. The effectiveness again is linked with the resonance condition, which must be fulfilled between the gem and his carrier.

24.2.6 Meditation and self-therapy.

There even exist people, who don't need any technical aid at all to get rid of unpleasant resonances. Some meditate and go into resonance with themselves, whereas others prefer to love a person, to whom they feel „attracted" or with whom they are „on the same wavelength", which means as much that they go into resonance with this person. In the Catholic Church for instance the priests are not allowed to get married, because they should be in a resonance with the church and with God.

<i>:: Dr. Bodo Kohler: Die Coloroma-Therapie, Der aussergewohnliche Einsatz von Aromen, Co'Med 4/2000, S. 48-52

<ii>: Cousto: Die kosmische Oktave, Synthesis Verlag Essen 1984

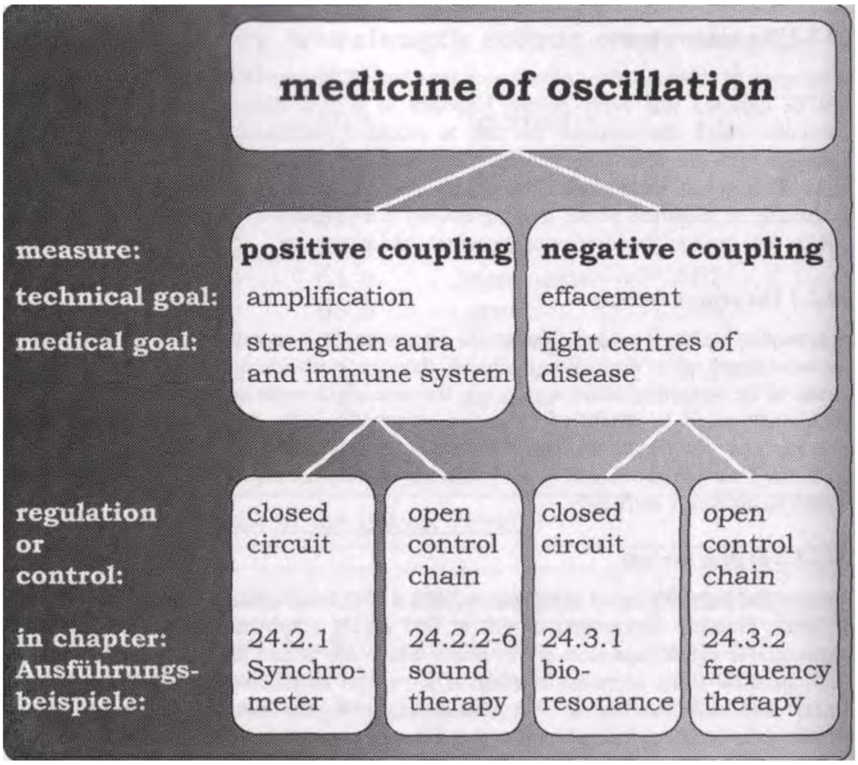
medicine of oscillation
measure:
technical goal:
medical goal:
positive coupling
amplification
strengthen aura and immune system
negative coupling
effacement
fight centres of disease
regulation or control:
in chapter: Ausführungs-beispiele:
closed circuit
24.2.1 Synchro-meter
open control chain
24.2.2-6 sound therapy
closed circuit
24.3.1 bio-resonance
open control chain
24.3.2 frequency therapy

Fig. 24.3 A: Structure concerning the medicine of oscillation

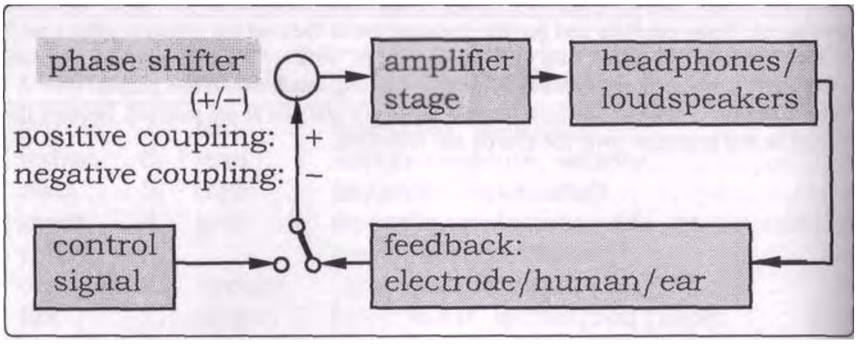
phase shifter
(+/–)
positive coupling: +
negative coupling: –
amplifier stage
headphones/ loudspeakers
control signal
feedback: electrode/human/ear

Fig. 24.3 B: Regulation or control?
Positive or negative coupling?

24.3 Systems with negative feedback

For negative sign it concerns negative coupling. In this case the signals are subtracted. The signals released by a person are played back to him with opposite phase. This method lies in the domain of the bioresonance.

24.3.1 The homeopathy

The concept, which is pursued by the medicine of oscillation in the case of the bioresonance, is that harmful or ill making oscillations should be effaced. Not the direct, but rather the indirect strengthening of the immune system is to the fore, which should be reached by relieving it. The approach with that pursues the same goal as the pill medicine, where oscillations, which should produce a wanted effect and which are bound by means of substances are supplied to the body.
The disadvantage of the administration of pills is the often toxic effect of the carrier substance. The German doctor Samuel Hahnemann (1755-1843) has demonstrated a way, in which to make the helpful and important information for the body to go over from the carrier substance to the water by diluting with water and by shaking. The water molecules now oscillate in time of the carrier molecules. Interestingly in doing so not even the information gets lost, the modulation thus is preserved. By constantly shaking there even occurs an amplification or a „potentiation", as Hahnemann expressed himself, because now every carrier molecule reaches and modulates countless water molecules. Thus the amount of water informed is bigger than the amount of the original chemical substrate. In the case of homeopathy like is treated with like. If for instance a poison causes certain complaints at a healthy person, then in "homeopathic" administration this information helps a sick person with similar complaints. There thus occurs a disturbance activation with opposite phase, as the engineer uses to express himself. Despite a widespread scepticism the method of homeopathy indeed seems to function, and it has stood the test in countless cases.

24.3.2 The bioresonance

A technical realization is represented by the bioresonance. For that endogenous oscillations are called off by means of an ECG (electrocardiogram), an EEG (electroencephalogram) or a MEG (magnetoencephalogram) at the surface of the skin. The technical device then shifts the phase for 180 degrees and amplifies the signal to the extent that pathological frequencies are extinguished for reversely directed input. This very reasonable theoretical concept in practice of course only is as efficient, as the empirical determined pathological frequencies are the cause of a disease and not only represent an unimportant symptom as side effect.
To that technical problems are joining. Prof. Heine blames the constantly changing reaction diversity and the thermal noise for the circumstance that the „frequency spectrum permanently is fluctuating", as he writes if*. That hardly can realize the necessary phase inversion. We meanwhile know that vortices have a fluctuating frequency spectrum, that in the case of biosignals it concerns such field vortices, which result in a noise signal unspecific in frequency.
With this knowledge we should be able to significantly improve bioresonance procedures, and even the inversion of phase shouldn't represent an insurmountable difficulty anymore.

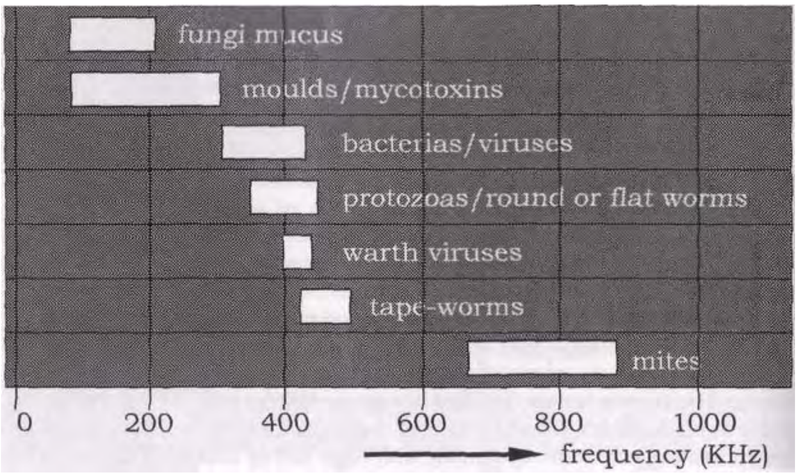

Fig. 24.3 B: Frequency spectrum of parasites according to Clark.<i>

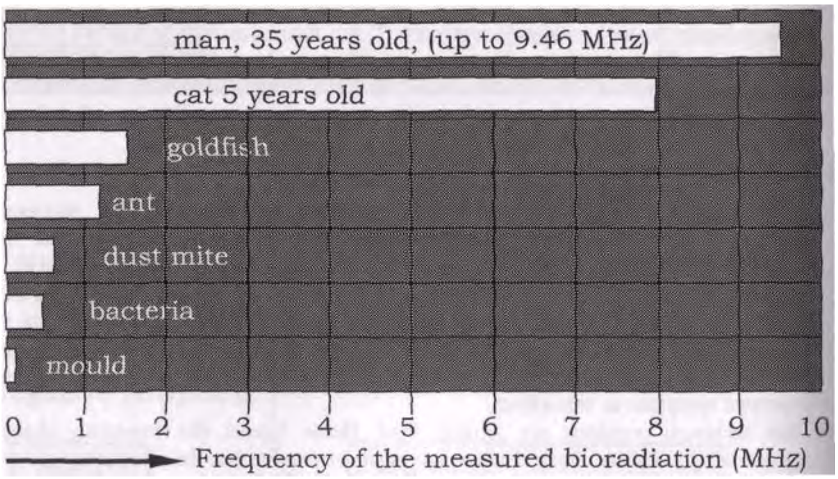

Fig. 24.3 C: Diagram of frequency ranges for some living beings.<i>

<i>: Dr. Hulda Clark: The Cure For All Diseases (1995), Page 604-643

A further problem, which Heine addresses, can't be denied[ii]: „For communicative molecule oscillations the microwave range of 1 GHz up to above 10000 GHz is crucial. A calling off of endogenous interference oscillations in this range with the help of normally used electrodes is not possible"[iii]. It thus could happen that essential frequencies, which are directly related to a disease, aren't recorded at all and as a consequence also not treated. There even is the risk of informations being brought in, which additionally stress the immune system instead of relieving it, that the patients after a treatment are worse off than before. The doctor or therapist is burdening himself with a big responsibility, when, how and at whom he applies methods of bioresonance or if he better does without them. For the mentioned reasons the method in the therapy only plays a secondary role. The bioresonance on the other hand is quite often and with great success used in the diagnosis (Nosoden), if it for instance concerns the determination of an incompatibility or an allergy - but that is a completely different theme. For the bioresonance the transition of diagnosis to therapy is however floating.

24.3.3 The frequency therapy

If using a frequency therapy the problems are standing similarly. If we separate the closed and negatively fed back circuit of a bioresonance and form an open control chain consisting of a technical control device and the patient, then we get the structure, as it is put to use in a frequency therapy. The goal still is the same: parasites or pathogenes, which stress and burden the immune system should be fought.

But a disadvantage of every open control chain is that the treatment at first occurs blind due to the missing feedback.

The natural healer Dr. Hulda Clark has examined as support various pathogenes and parasites under the microscope, while she has varied the frequency and at the same time applied a low-tension voltage. Doing so she could observe the dying of bacterias and parasites at certain frequencies. Correspondingly she publishes tables, in which the in each case „mortal" frequencies are listed[i].

Hut without being able to verify the success of a treatment in the living organism, she proceeds from the assumption that by applying a low tension voltage (5-10 Volt), as it is produced at the output of a commercially available frequency generator, if a sinusoidal signal is set with the appropriate frequency exactly the associated parasites and leeches will be destroyed. Doing so a fixed rhythm of pauses and treatment times must be adhered to (7 min on, 20 min off, 7 min on, 20 min off and 7 min on).

The doctors and therapists treating with this particularly inexpensive method report amazing results, inexplicable spontaneous cancer cures, like of HIV-positive patients, who after the treatment were tested HIV-negative. But also the inverse case already should have occurred, that a HIV-negative patient afterwards was HIV-positive! Here too clear the limits of this method appear, which in practice unfortunately turns out to be relatively unspecific.

<ii>: Prof. Dr. H. Heine: Grundregulation und Extrazellulare Matrix - Grundlagen und Systematik, Lehrbuch der biologischen Medizin, (1991) 2.Aufl. Hippokrates Verlag Stuttgart, Seite 63

<iii>: U. Warnke: Bioresonanztherapie - Wunsch und Wirklichkeit. Medizin transparent Nr.2 (1996) S. 36-37

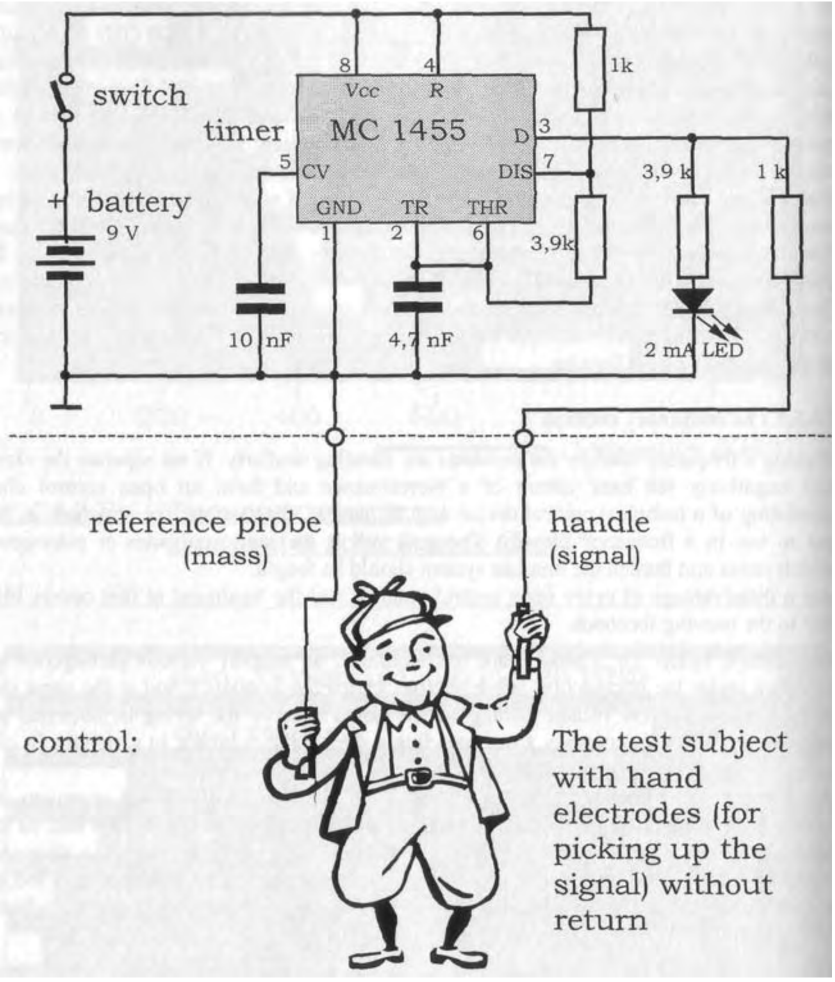
switch
timer
MC 1455
8
4
Vcc
R
1k
3
D
5
CV
DIS
7
+ battery
9 V
GND
TR
THR
1
2
6
3,9 k
1 k
3,9k
10 nF
4,7 nF
2 mA LED
reference probe
(mass)
handle
(signal)
control:
The test subject
with hand
electrodes (for
picking up the
signal) without
return

Fig. 24.4: Wiring diagram of the Zapper.<i>

Wiring diagram taken from:

<i>: H. Clark: The Cure for All Cancers (1993) New Century Press, USA, Page 507, resp.:

Dr. Hulda Clark: The Cure For All Diseases (1995), Page 48

24.4 The Zapper

From the point of view of physics of today the frequency therapy actually shouldn't be able to function at all. The electromagnetic waves penetrate only few millimeters into the skin at the used frequencies and wouldn't have the slightest chance, to reach a parasite, which is staying somewhere in the body. But it is said that it is possible to cure athlete's foot, by the patient taking the electrodes in his hands. How, we ask the question, does the signal of the function generator know where it should go?

It quite obviously concerns a resonance phenomenon. The likewise emitted scalar wave part tunnels undamped at those places in the body, with which it can build up a resonance and that for proper tuning are the unloved parasites. The scalar wave radiation is bundling up at the resonant receiver, so that despite the low transmission power as a consequence of the bundling up the energy density for the parasite becomes very high. It as a result is destroyed by its own ability to go into resonance. Once it is killed, the next one goes into resonance, logs off as well etc. In this way the parasites are destroyed one after another and not all at once. That's why the specified treatment cycle makes sense.

The copper electrodes should not be taken in the hands directly, Dr. Clark recommends, but before be wrapped with moist paper. By means of this insulating layer, so is my interpretation, the conventional wave part, for which the skin functions as a wave guide, is reduced whereas the desired scalar wave part is increased. Such measures crucially contribute to the success of a therapy method, even if they were determined purely empirical.

If one wants to address every possible parasite individually, then the treatment takes correspondingly long. If one on the other hand sends all relevant frequencies at once by overlapping them, then the treatment can be abbreviated to the duration of one session. If the therapist goes still further and replaces the sinusoidal signal by a rectangular signal, then infinitely much sinus functions are hidden in it, as a Fourier analysis shows. With a rectangular signal, as it is delivered by the Zapper, one as it were catches everything, Good as well as Bad. There the helpful intestine bacterias break exactly like the wrongdoers.

The treatment with the Zapper is simple, inexpensive and exactly as controversial. It is the shot with the shotgun in the forest. One always hits something. We nevertheless must ask the question, why one only hits parasites and bacterias and not the vital organs? Aren't those damaged also?

Now then, the signal of the function generator is not modulated; it doesn't carry information. That's why only monocellular parasites, which don't know information exchange, are capable of a resonance. Human cells and more than ever whole organs on the other hand work with complex modulations, which effectively prevent any formation of resonance with the technically generated basic wave, with which the question would be answered so far.

That however also means that immune reactions can be expected: If the first treatment with the frequency therapy still is successful and all simple parasites could be hit, then only further evolved parasites have been spared, which modulate their information. They now breed and can't be reached anymore in further sessions. The method suddenly doesn't function anymore, the therapist finds out, the body apparently has become immune.

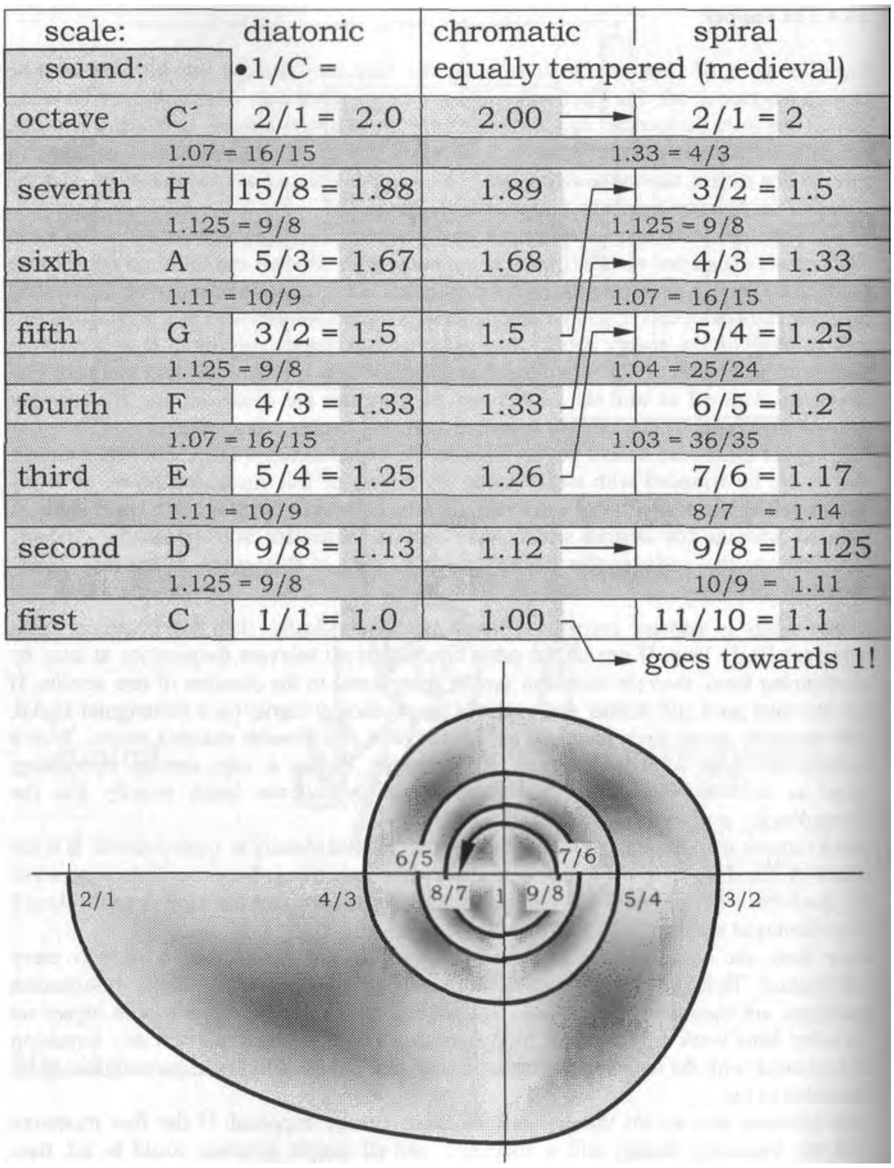

scale:		diatonic	chromatic	spiral
sound:		•1/C =	equally tempered	(medieval)
octave	C´	2/1 = 2.0	2.00 →	2/1 = 2
	1.07 = 16/15			1.33 = 4/3
seventh	H	15/8 = 1.88	1.89	3/2 = 1.5
	1.125 = 9/8			1.125 = 9/8
sixth	A	5/3 = 1.67	1.68	4/3 = 1.33
	1.11 = 10/9			1.07 = 16/15
fifth	G	3/2 = 1.5	1.5	5/4 = 1.25
	1.125 = 9/8			1.04 = 25/24
fourth	F	4/3 = 1.33	1.33	6/5 = 1.2
	1.07 = 16/15			1.03 = 36/35
third	E	5/4 = 1.25	1.26	7/6 = 1.17
	1.11 = 10/9			8/7 = 1.14
second	D	9/8 = 1.13	1.12 →	9/8 = 1.125
	1.125 = 9/8			10/9 = 1.11
first	C	1/1 = 1.0	1.00	11/10 = 1.1

goes towards 1!

Fig. 24.5: Harmonic spiral built on the first Gregorian scale.[i]

[i]: Dr. Manfred Doepp: Naturgemasse Frequenztherapien, Die Harmonie der Snirale. CO'Med 10/2000. S. 46 - 49

24.5 Discussion concerning the medicine of oscillation

The textbook medicine in its explanations and treatment methods is basing on the models, which it can measure and analyse and which it understands. Doing so man and whole nature is reduced to a handful of chemical reaction formulas. The whole pharmaceutic industry lives on this misleading path, which long ago has revealed to be a dead end, medical as well as financial. This health service cannot be paid anymore and we should ask the question if it actually is worth the money, if with electric signals of minimum power can be obtained effects comparable to the effects of the pill medicine.
We need a new medicine, a potential vortex medicine. First of all we should research how an organism covers its energy needs and how it communicates. There leads no way past the scalar waves and the newly discovered potential vortices. Chemical processes as they are being observed, occur by the way, that is beyond doubt, but they by no means are the cause. Hence of pills and other chemical means at most a treatment of symptoms and a case of side effects may be expected but not a cure of a disease. Once the potential vortex medicine will be systematically explored and be put to use in practice, healing successes can be expected, which we at present can't imagine at all.
The amazing results, which already today are obtained in the medicine of oscillation and of which some doctors can report<i>, dictates the direction in which the textbook medicine should develop. In the question, which kind of oscillation or which ,,sequence of sounds" (Fig. 24.5) is the right one, still exists considerable need for research.
Some doctors even already work with my new theory and cite whole passages from my publications until now about this theme<ii>.

<i$_a$>: Dr. Bodo Kohler: BITsyn, der neue Weg in der Informations-Medizin, Jubilaums-Schrift der Internationalen Arzte-Gesellschaft fur Biophysikalische Informations-Therapie, Bad Nauheim den 6.-8.10.2000, Seite 48 - 56.

<i$_b$>: Bioenergetische Messverfahren in Theorie und Praxis, Vortragsband der Gesellschaft fur Energetische und Informationsmedizin e.V., Universitat Stuttgart am 17.7.1999.

<ii$_a$>: Dr. Johann Lechner: Storfelddiagnostik, Medikamenten- und Materialtest, Teil 2 aus der Reihe: Praxis der Ganzheitlichen Medizin und Zahnmedizin, Verlag Dr. E. Wuhr (2000), Kap. 2.4.2 Beruhrungslose skalarwellentragende Informationsubertragung S. 173 ff., bes. Kap. 2.4.2.3 Seite 175, 176.

<ii$_b$>: Dr. Reichwein, Peters: Zellulare Elektromagnetische Systemsteuerung, Der Freie Arzt 5 (2000) im Wissenschafts-Forum (Anhang, S. IV - XXIII).

<ii$_c$>: Dr. M. Doepp: Tesla-Wellen, Neue Studien, CO'Med 5/2000, S. 94 - 95

Frequency spectrum of the human organism:

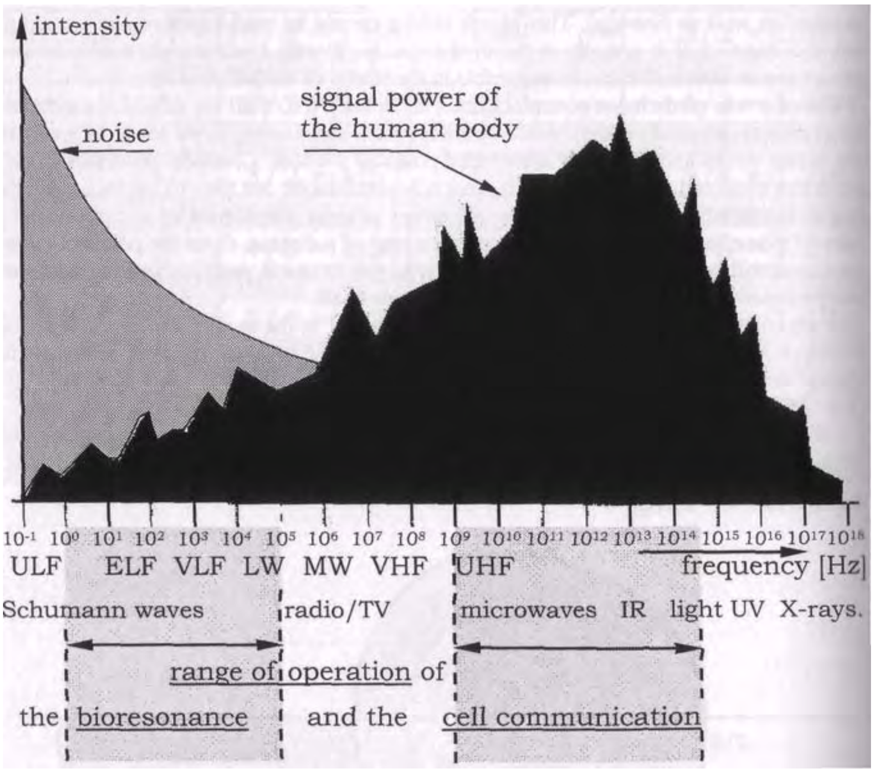

Fig. 25.1: The measured frequency spectrum of man.<i>

<i>: Dr. P. Bembenek: Akupunktur und (Bio-) Resonanz, CO'med 6/98, S. 50-58

25. Contributions to the discussion concerning the information technical use

The medical research predominantly takes place with statistical methods. This might be changed fundamentally, if only the physical relations have been realized and their causes found out. Only then a basic research will be possible also in the area of medicine, as it today already is usual in other disciplines.

In this chapter it concerns the question of the medical basis in general, and specifically the question of the use of the newly discovered potential vortices in biological systems. Doing so very aware is attempted to put the normally used medical view at the back and to derive the answers from the perspective of the engineer and the physicist. This approach is unfamiliar, so that some textbook physicians will have difficulties to follow. It however will prove to be very helpful. Medicine in addition needs new impulses and that justifies a new approach, like the unconventional one proposed by me.

In this connection gladly is fallen back upon the comparison of biological with technical systems, although here exist differences in principle and fundamental It can be expected that nature has developed very appropriate strategies to solve certain tasks, as it also is teached in engineering and used in practice. At this place not only medicine benefits from new insights, engineering vice versa also can learn of nature, because natures optimization strategies are much much older and correspondingly perfect.

25.1 Window of the sensuous perception

To clarify the inner processes in man we consider his sense organs, which function as interface to the outside world. Technically seen in the case of the sense organs it concerns measurement converters to record and process certain physical values in our environment and their adaptation to pass on the information by means of the nerves. The measurement converters thereby are built more or less complex, which is related to the circumstance that the measurement factors, which a living being needs to orient itself and to be able to survive, are quite different.

It concerns little windows within the frequency spectrum shown in fig. 23.1, which should be analysed. Man for instance chooses the visible frequency window and interpretes it with an assignment to the colours, because in this range the sunlight can reach the surface of the earth and the ionosphere doesn't exert an excessive damping (fig. 2.8).

The acoustic window depends on the sound propagation in air, whereas dolphins work with ultrasound, which carry considerably further in water. The associated measurement converters, the ear and the eye, have a very complex structure. From this can be inferred that here the received signal must be transformed into a different physical measurement signal, that the perceivable signals are of other nature than the signals, which our nerves are able to pass on.

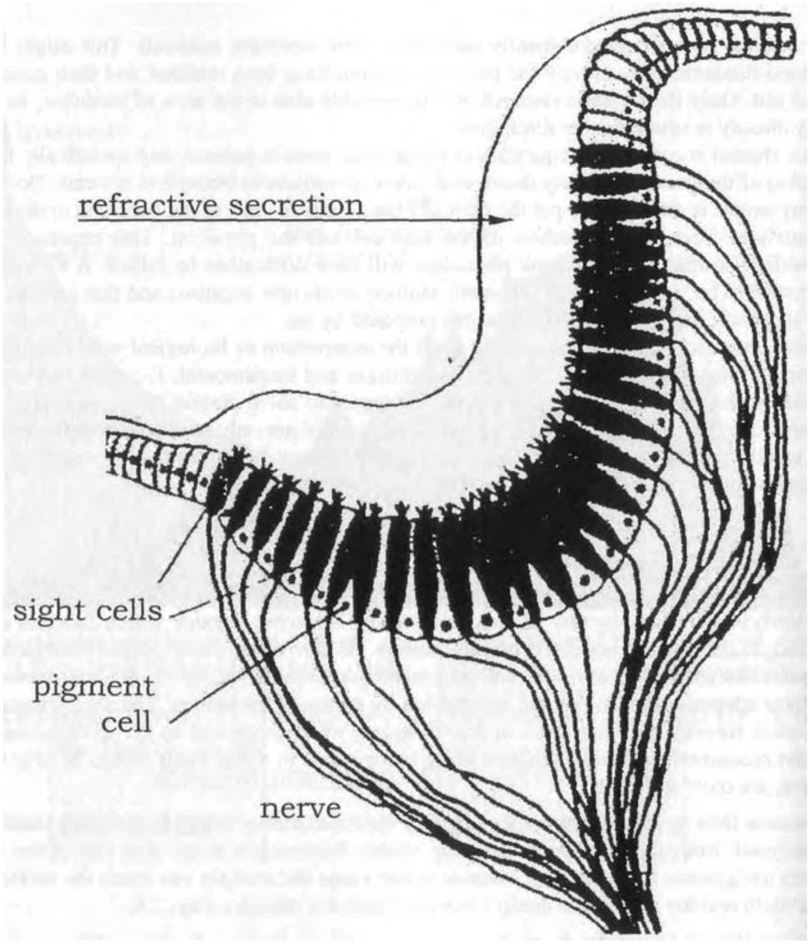
refractive secretion
sight cells
pigment cell
nerve

Fig. 25.2: ____ The hollow eye of the snail

25.2 The sense organs (nose, eye, tongue)

If we want to find out something about the nature of the nerve signals, then we should more detailed consider a sense organ, which is constructed if possible simple. In that case it is possible that no signal transformation is necessary at all anymore, because the perceivable variable is exactly of the kind the nerves can process.

The primary and perhaps most ancient sense organ of the living beings is the nose. It is constructed extremely simple. On the one hand according to encyclopaedias it is a chemical sense for the perception of smell materials. Man on the other hand can distinguish several thousands of scents. For that a huge analysis factory with correspondingly much receptors would be necessary, which one searches in vain in the nose between the nasal hairs and the nerves. We from that infer that scent should consist of particles, which are to a large extent identical with the reaction potentials. It thus will concern potential vortices. These ring-like vortices of the electric field again are modulated and carry information, which our brain is able to analyse.

According to the derivation in chapter 9.0 Plato already 2400 years ago knew such relations. The conclusion is obvious that smell is vortex information, which according to the explanation of Plato forms at the transition of waves in potential vortices and vice versa. At the natural philosophers in antiquity the circle of the insights is closed. Modern science just isn't that far.

The hollow eye of the snail functions both as nose and as eye (fig. 25.2). This very ancient combined organ points to a relatively close affinity between both measurement sensors. Possibly in the course of evolution both organs have developed from such a common original form.

In the case of man the smell rods occupied with the nasal hairs of roughly 2 micrometer in length, actually remember of the rods and cones in the eye, with which photons are collected. In both cases, so we can explain the mode of action in accordance with model, the ring-like vortices settle on the rods and run as reaction potentials along the nerve to the brain, which then analyses the modulation.

In the case of the highly developed human eye the light rays first have to traverse the vitreous body and afterwards a pigment layer, before they reach the rods and cones. In this way the electromagnetic wave parts must roll up to vortices, since our nerves can only pass on vortices. The photons can be interpreted as corresponding vortices.

If we compare the eye with a bubble chamber, in which the tracks of ionized particles can be observed and photographed as thin white fog stripes. Here in collision experiments it can be proven that it concerns particles carrying impulse and not waves. But the measurement setup by no means does answer the question if it already concerned particles before they entered the bubble chamber, or if these have formed spontaneously in the presence of the saturated vapour.

It would be obvious, if for our eye the chamber water being under the inner eye pressure would take over the function of the bubble chamber and would take care of the rolling up of the light waves to light quanta.

Example tongue:

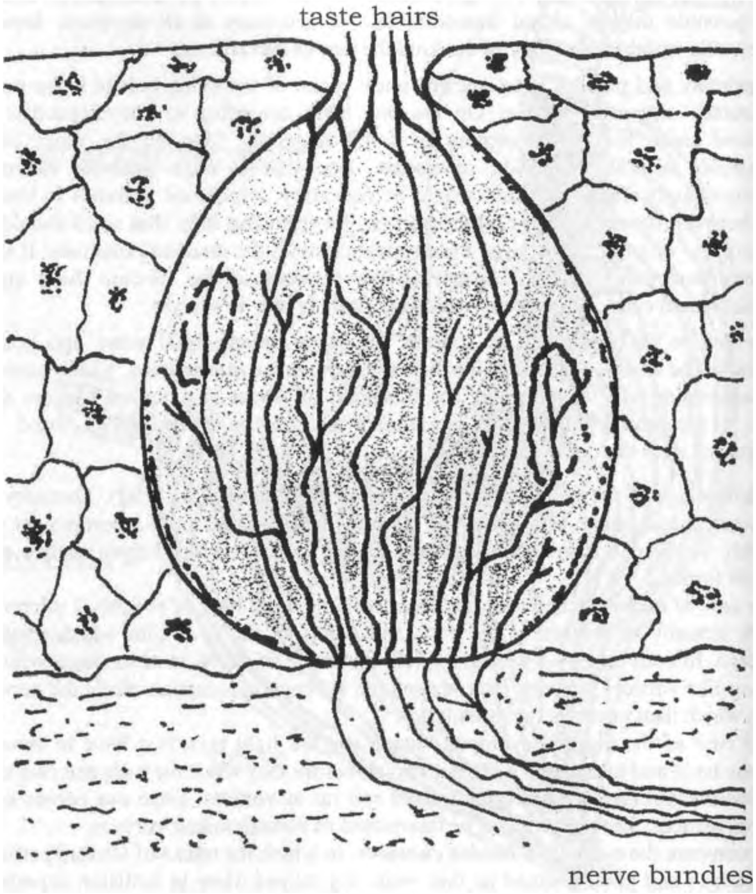
taste hairs
nerve bundles

Fig. 25.3: Taste bud of the human tongue

We optically can't detect a difference, if the light while passing through a thick glass plate increases its part of photons. At the latest in the eye in any case all the light is transformed in photons and usable vortices, so that we mustn't detect a loss of brightness. The glass plate seems transparent to us, even if the sunlight should change its composition and its biological quality, if it passes the plate.

If the tongue while tasking responds to dissolved substances, then it not by all means needs to be a chemical excitation. Instead fine hairs, the taste hairs resp. sense pins, serve as receiver like in the case of the nose. The similar structure of the receptors and the circumstance that for most invertebrates sense of taste and sense of smell can't be distinguished of each other at all and consist of the same primary sense cells, suggest that the tongue doesn't analyse the chemistry as such, but only the molecular oscillation patterns, that also the taste is nothing else but vortex information!

25.3 The nerve conduction

For the collection of the potential vortices and the extraction of their information fine hairs in the sense cells obviously play a central role. They are connected more or less directly with the end of a nerve and pass on the information without big transformation. Even in the organs of equilibrium sense hairs work.

From the comparison with the technique developed by Nikola Tesla I could show that the nerve conduction concerns a single-wire transmission, a kind of waveguide, for which the transport of the excitation information takes place in the insulation layer and not in the conductor itself (fig. 9.6). As proof I quoted that the thickness of the insulation determines the velocity of propagation, that as is well-known the nerve conductors with thick fat layer pass on their action potentials faster than those with thin insulation. Particularly interesting is the observation, how the fat layer is constricted in fixed intervals, like for Wiener sausages (fig. 25.4 A). These nodes of Ranvier prove that only longitudinal waves are being transported, which are standing waves with nodes and antinodes, if the distance from node to node exactly corresponds to the distance from one node to the next. With that nature with the use of the potential vortices is far ahead of our power engineering. The nerve-cables determine with their structure, which signal will be transported and which not!

The technical cables on the other hand are stupid and conduct everything, the useful signal just as well as any arbitrary interference signal. Anyone, whom the computer crashes every few minutes, knows what I'm talking about.

Man isn't able to afford a crash of his think computer. It would be lethal for him. His nerve costume even tolerates short-circuits. The acupuncture is such a short-circuit technique by means of electrically conductive needles. There even is given a therapeutic benefit and a relaxing effect for the body.

Cut through nerve fibers even can partly regenerate again, even without a cut through nerve again growing together with its other end. The nerve conductors are so intelligent, that only the matching information arrives at the end by passing on informations from one fiber to the corresponding next with the same node interval.

A. Nerve conduction with nodes of Ranvier (to transmit standing waves):

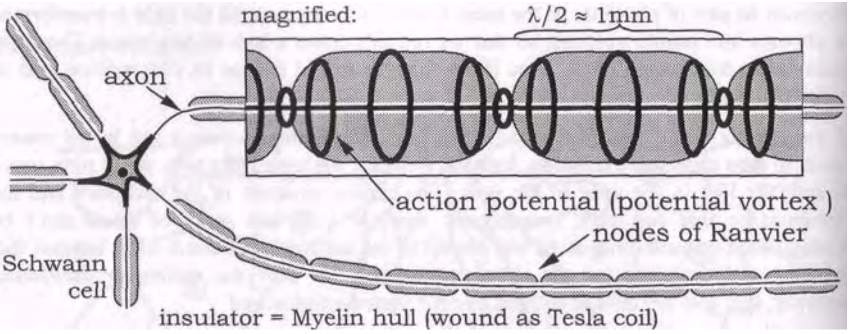
magnified:
λ/2 ≈ 1mm
axon
action potential (potential vortex)
nodes of Ranvier
Schwann cell
insulator = Myelin hull (wound as Tesla coil)

B. Single-wire transmission according to Nikola Tesla<i>:

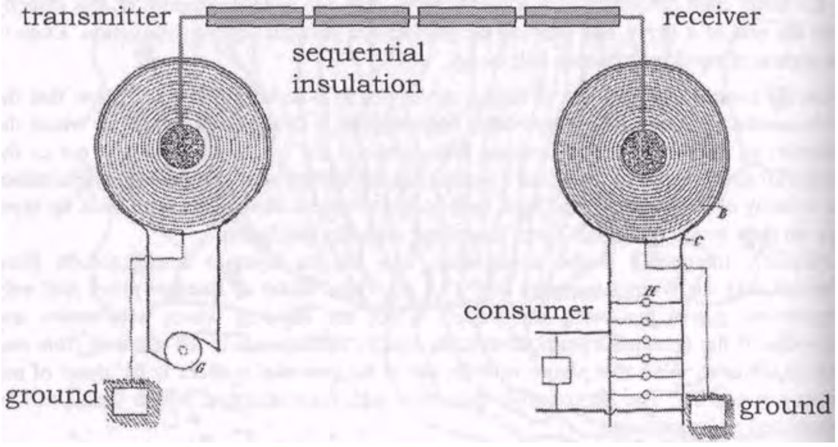
transmitter
sequential insulation
receiver
B
C
H
consumer
G
ground
ground

Fig. 25.4: ____ Selective scalar wave signal transmission

The brain works with scalar waves! Reasons are:

1. Lacking of a signal transformer between the nerves and the brain.
2. High performance density of the think apparatus.
3. The brain activity measurable from the outside with the EEC
4. Spark formation and corona discharges at open top of the skull.
5. Insulation defect occurring in the case of epileptics.

<i>: Original sketch of Tesla is situated in the Nikola Tesla Museum, Belgrade.

Nerves represent an intelligent and at the same time interference safe wiring, which is superior to any technical solution by far, or you can try to acupuncture a cable cord of your computer. You will have little pleasure of it.
Nikola Tesla also in this point was ahead of the times. He has experimented with a single conductor technology, for which the insulation layer is constructed like for a nerve fiber (fig. 25.4 B). We'll have to learn to handle such aids in the future, if we in engineering want to emulate nature and as well want to send informations as scalar wave. A normal power cable as said conducts everything, useful as well as interference signals. If however the signal should be accomodated in the noise, a selective cable is required, which should be constructed according to the example of the nerve conductors. Before a scalar wave technology can be introduced and successfully used according to that completely other cables, coupling elements, amplifiers and other components should be developed. The trouble should be worthwhile at the chances, which this genial technology offers us!

25.4 The brain, a scalar wave computer

The brain cells (neurons) are of the same kind as the nerve cells. Hence can be done without a signal transformer, can the informations transmitted by the nerves to the brain directly be processed further. From that follows that also the brain without exception works with potential vortices. There are several reasons for this hypothesis:

1. as said, the lacking of a signal transformer.
2. the high performance density of the think apparatus. (As a result of the concentration effect of the potential vortices the efficiency of the human brain is in such a way high concentrated compared to the much more space claiming computers functioning on the basis of currents).
3. the brain activity measurable from the outside with the EEC
4. spark formation and corona discharges at open top of the skull. (Brain surgeons can report of such observations).
5. The insulation defect occurring in the case of epileptics. (During a fit instable potential oscillations of the nerve cells occur, which lead to strong electric blows).

With the „exciting" and the inhibiting"' synapses as separation point between the neurons both a ,.high-active" and a „low-active"' method of operation is possible and with that a redundant, particularly interference safe signal transmission.
Safety for interference is very important not only in nature. In the operating instructions of a PC can be read: „operate only at room temperature, keep dry, don't throw or shock, take cure for sufficiently cooling air, ground apparatus, pay attention to mains voltage, etc". A comparison with the range of operation of man is like scorn.
Nevertheless the consequences if errors occur are quite similar: a garage door, which opens if a mobile is switched on, by all means can be compared with a light phenomenon, which we perceive after a blow on the eye at the biological level.

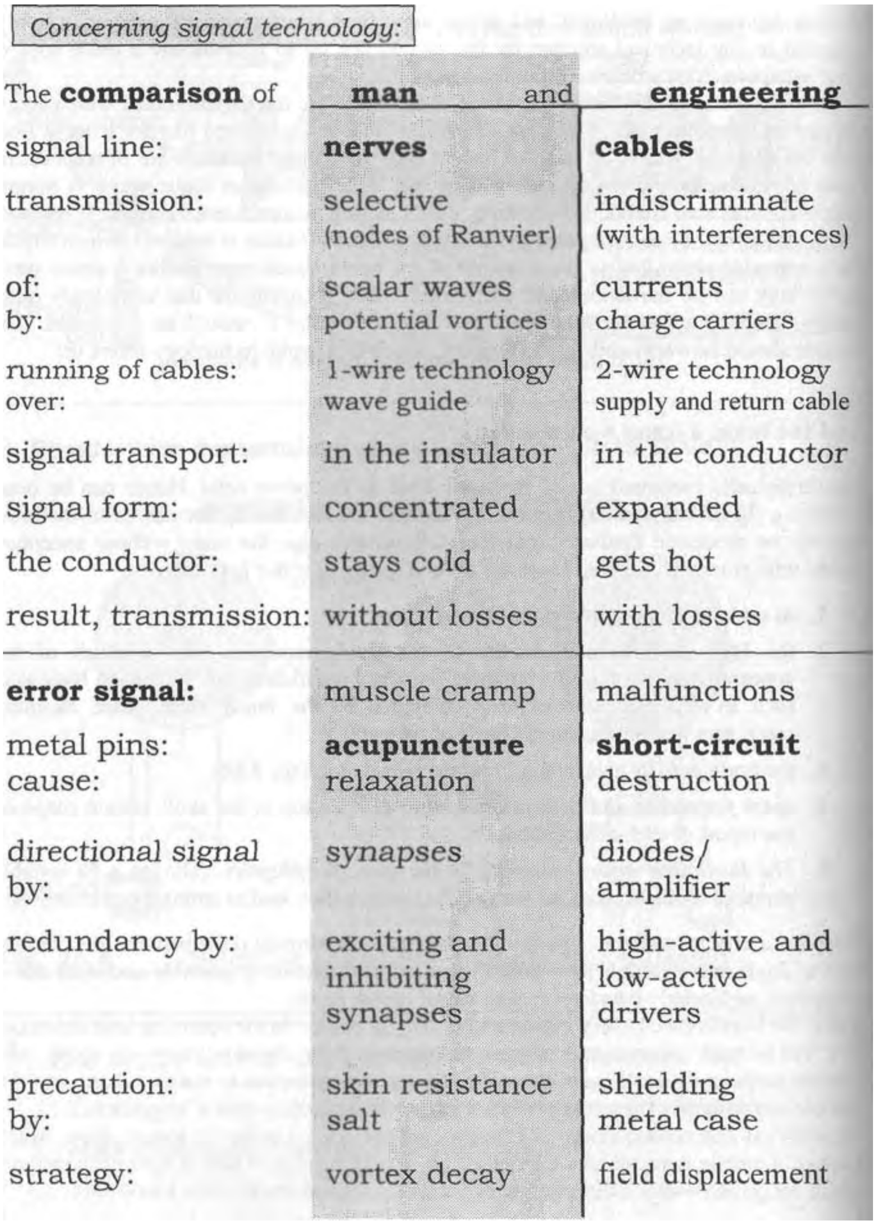

Concerning signal technology:

The **comparison** of	**man** and	**engineering**
signal line:	**nerves**	**cables**
transmission:	selective (nodes of Ranvier)	indiscriminate (with interferences)
of: by:	scalar waves potential vortices	currents charge carriers
running of cables: over:	1-wire technology wave guide	2-wire technology supply and return cable
signal transport:	in the insulator	in the conductor
signal form:	concentrated	expanded
the conductor:	stays cold	gets hot
result, transmission:	without losses	with losses
error signal:	muscle cramp	malfunctions
metal pins: cause:	**acupuncture** relaxation	**short-circuit** destruction
directional signal by:	synapses	diodes/ amplifier
redundancy by:	exciting and inhibiting synapses	high-active and low-active drivers
precaution: by:	skin resistance salt	shielding metal case
strategy:	vortex decay	field displacement

Fig. 25.5: ____Comparison of the signal technology

25.5 Concerning signal engineering

By means of a technical analysis of biological relations completely new interpretations result also for the occurring of a disease. We permanently and everywhere are surrounded by noise signals, but as a rule they can't touch us, because the body has developed perfect strategies for defence. The nodes of Ranvier on the nerve bundles here are just as helpful as the diode effect of the synapses. By means of the salt content and the skin resistance the body in addition controls the uptake of potential vortices from its surroundings, by using that the vortex decay is determined by the conductivity.

The specialties of the by humans used signal technology come to light particularly clear if compared with the cable technology used in the technical world (fig. 25.5). So is worked with only one wire instead of with supply and return cable, are mediated potential vortices instead of charge carriers, does the transmission take place in the insulator without losses and not in an electric conductor, which as a result gets hot and produces current heat losses. Nerves thanks to their ability of selection represent an intelligent form of signal transmission, by helping to filter the asked information from the noise. This surely is necessary since with cables, which indiscriminately transmit every signal, an use of scalar waves hasn't succeeded yet. We should try to learn of nature!

A special challenge is the protection against error signals. A passive shielding by a metal case however is not possible, because scalar waves can't be shielded by principle, so that the precaution should be taken actively by means of the conductivity. That's why we sweat salt if we strain us physically, whereby the vortex decay is determined by the conductivity, which depends on the salt content of the body liquids. In the case of a sweating activity the body reduces its conductivity, so that the needed potential vortex energy will reach the cells.

If the body sometime isn't able to defend itself against interference signals, then malfunctions or pathological reactions are a possible result. During having a bath for instance a muscle cramp can occur, if the body doesn't defend itself fast enough or sufficiently against the high potential vortex activity in the water. Now vortices can be picked up in the nerves, which are of the same kind as the ones emitted by the brain, only that both muscles, biceps and triceps at the same time get the signal to contract. The result is a cramping of both muscles.

Thus the brain has developed intelligent strategies to protect itself of interspersed misinformations. It weighs the incoming signals and forgets again all unimportant ones more or less fast. We speak of the ability to learn and that means that signals rise in the valuation scale and with that are stored longer, the more frequent repeated our brain receives them. This strategy assumes that interference signals only occur sporadic, for which reason they are rated unimportant and fast are again forgotten. A PC on the other hand doesn't have such a property. It notices everything indiscriminately and sometime will crash of overload, if not the user will constantly foster it and will administer the available memory. A PC is and stays stupid.

Concerning the stability against interference:

The **comparison** of	**man** and	**engineering**
in the	**brain**	**computer**
control technical interference signals:	light phenomena	miscontrol, program fault
precautions:	active, i.e. weighing of the Info by learning (repetition)	passive, i.e. shielding. (no evaluation of the input values)
unimportant (interference signal)	is forgotten	is stored
danger exists for signals	recurring interference patterns (see mobile telephony)	interference of all kinds (s.EMC-standardization)
remedy:	self-cure	restart (reset)

Concerning set of difficulties of wear and tear:

The **comparison** of	**man** and	**engineering**
repair	self-repair	in the workshop
by:	cell division permanently	exchange parts at maintenance resp. in the case of damage
required building materials:	nourishment	material / oil
waste:	compost	rubbish

Fig. 25.6: ____Strategies against interference signals, for damages linked with operation and for wear and tear.

25.6 Stability against interference and repair mechanisms

The interference signals present in our natural environment as a rule are distributed stochastic, but not so artificial interference signals like for instance transmitters emit. If for instance in the case of mobile telephony there occur time and again identical signal patterns and if a person perceives these, then because of the continual repetitions a high importance is pretended and precious storage space in the brain is allocated. This to a special extent applies for the permanent stand-by signals, which are emitted by mobiles and cordless phones even then, if we don't phone at all. Such misdevelopments thereby would be technically avoidable just like that!
If still no gauges are available and as a result there exist neither guidelines nor limits for the radiation exposure of scalar waves, then such devices must be developed and built. For the requested electromagnetic environmental compatibility not engineering, but man should be in the centre of attention!

Let us throw a short glance at the set of difficulties of wear and tear. Most technical devices find the way into the workshop only, if they already are defect. Some aren't repaired anymore in principle and immediately sent to the rubbish, because they are worn-out and a repair isn't worth the effort anymore one says, whereas other, mostly expensive systems are being serviced by exchanging all wearing parts.

Nature has brought to perfection the last principle. It allows the body a permanent maintenance; by permanently producing new cells and replacing consumed ones. It with that obtains a considerably longer operating time and even is capable to heal wounds. Just imagine dents in our cars would disappear from alone after a few weeks and the bodywork would look like new. Such an optimal maintenance is costly and it has its price. By means of the cell division the building plan for the spare parts is copied. The task on the other hand is transmitted to the new cells „by radio" by means of waveguide channels, as proves the matching structure of scalar wave and waveguide (chapter 21.12 and fig. 25.7). The cells hence have a kind of „decentral intelligence", which technical matter lacks completely. A comparison with engineering nevertheless makes sense, because wear and tear occurs for all systems in continuous operation. Thus strategies must be developed to provide the necessary material for operation, building-material and exchange parts.

By sizing of the hyperboloid structures of the matrix channels it should be possible to determine the wavelength of the scalar waves to enter them in the frequency diagram (23.1).

Tunnel structure of the basic substance according to Prof. Heine<i>:

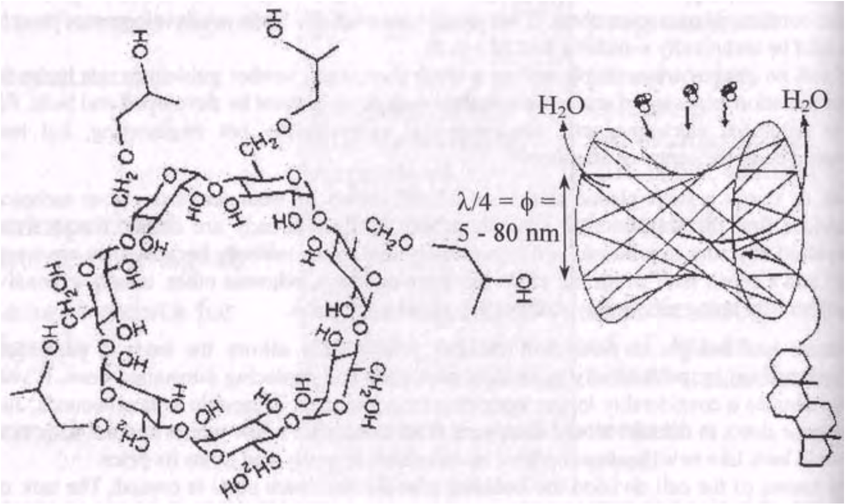

Fig. 25.7: ___ Information channels in the intercellular matrix.<i>

<i>: Hartmut, Heine: Lehrbuch der biologischen Medizin. Grundregulation und Extrazellulare Matrix, 2. Aufl. 1997, Hippokrates Verlag Stuttgart, S. 56

25.7 The microwave window

There still is lacking information about the frequency or the velocity of propagation to be able to specify and enter in figure 23.1 the cell information. Let's take to hand the book about "Biostrahlen"[i]. In there is explained how for individual cells the emission of electromagnetic signals already could be measured, which are interpreted as circularly polarized waves (fig. 21.10 A). It is said that thereby their polarization plane is rotating with the speed of light, whereas the wave itself is propagating in longitudinal way relatively slow and according to the authors with the speed of sound.
The occupation with the potential vortex theory in the explanation lets immediately recognize the vortex, for this is rotating with the speed of light in circles as rolled up wave (fig. 21.10). If it is bound to the surrounding matter with closed vortex centre, then it can be expected that the propagation actually takes place with the speed of body sound.
It could concern the same signals, which Prof. A. Popp calls biophotons and detects measuring technically in living organisms[ii]. He however considers the phenomenon for the same wavelength with the speed of light and lands at light frequencies, even then if nothing is glowing visibly! The question is asked: Does it concern the frequency of the light or only the corresponding wavelength or actually both, thus light, as is expressed in the name biophotons?
The photomultipliers, which Prof. Popp uses as „light amplifier", however can only be tuned to certain wavelengths and not to frequencies. Even if the detected biophotons have the wavelength of the light, then nevertheless nothing will glow if the velocity of propagation and as a result also the frequency differs from that of the light for several powers of ten. In the case of the immense number of cells also the number of photons should correspondingly sum up and the body as a whole should start to glow, which is not the case.

The waveguides in the intercellular matrix serving the cell communication, which Prof. H. Heine observes microscopically, have wavelengths between 20 and 300 nanometer, which corresponds to the range of the ultraviolet radiation[iii]. But if the propagation is slower than the light for 6 powers of ten, then also the frequency will only amount to one millionth and be situated into the range of the microwaves. Here a biological window seems to be present, to which we should turn our attention for reasons of the electromagnetic environmental compatibility!

The thermal radiation, which reaches the earth from the sun and the planets, lies in the microwave range between 2 and 20 cm. If the sun does well to us, if we need the radiation, then it could be because of the identical frequency. But that also means that the sun and the planets are capable to have an effect on the cell communication, that they for instance could function as clock for the heart.

<i>: H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage S. 97: Messung von der DNS eines Gurkenkeims abstrahlenden Photonen.

<ii>: A. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987

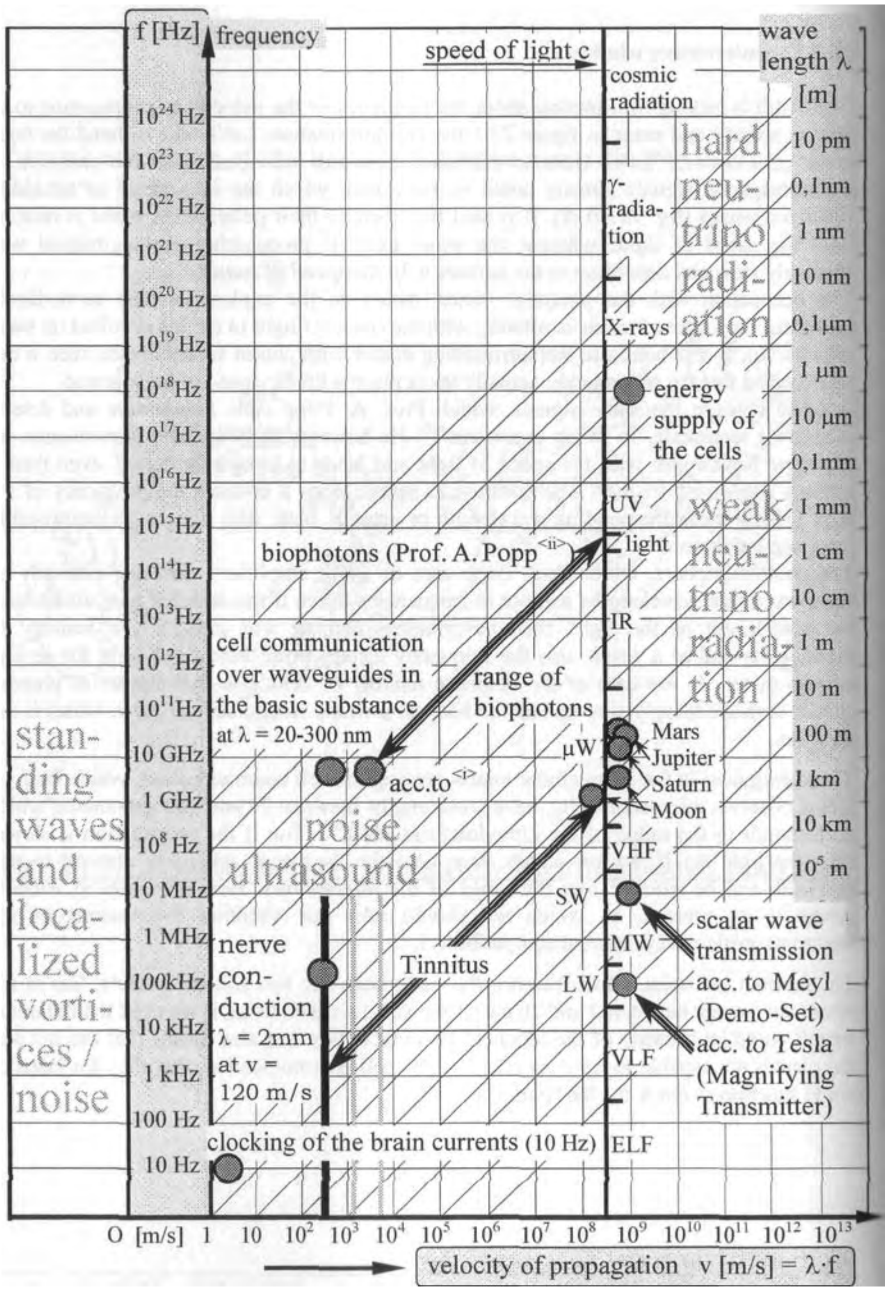

Fig. 25.8: Frequency diagram with entries concerning biologically relevant areas (Tinnitus, nerve conduction, etc.).

25.8 Discussion concerning scalar wave medicine

In the sensitive range of the supposed microwaves window however are also situated the mobile phones and their harmonic waves, which occupy a broad spectrum in particular for pulsed operation in digital nets. The D-net for instance has a wavelength of 32 cm, the E-net lies at half that wavelength. But to which frequency does this correspond at the speed of sound? Now, the frequency is 6 powers of ten smaller and now lies at 1 kHz resp. 2 kHz. To that are added the numerous harmonic waves, which form a noise signal and lie above that. With that these signals are situated completely in the audible range, there where our ears are most sensitive!
This cause we probably owe the disease of modern civilization „Tinnitus". Every charged particle will follow this electromagnetic oscillation and produce corresponding sound oscillations, which can hear not „sick ", but completely on the contrary „healthy" people, who as a result possibly get sick. The objection, in this range also cosmic radiation for instance from the planets is present, is legitimate. It however should be considered that planets also are going away from the earth again and in addition is present a fluctuation according to the time of day due to the rotation of the earth, while the mobile telephone masts in our vicinity radiate in continuous operation.

In this window in addition the clocking of the brain currents takes place at 10 Hz. I would recommend blocking the acoustic relevant range between 2 cm (16 kHz) and 3 m wavelength (100 Hz) for all technical use. Every operation of a transmitter in a biological window harms all people and cannot be answered for by any institution.

It further should be considered that the biological window of the plants and animals as a rule correspond to that of man, but sometimes are shifted significantly upward or downward in wavelength. We are not entitled to judge nature. The immune system of the animals now obviously has reached the breaking point and also that of man doesn't seem to be that anymore, which it originally was.
We must proceed from the assumption that many diseases on the one hand and therapy methods on the other hand partly direct and partly indirect have to do something with scalar waves. (E.g. the occurring of cancer[<4i>]).

<i>: H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage, S. 97: Messung von der DNS eines Gurkenkeims abstrahlenden Photonen.

<ii>: A. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987

<iii>: Hartmut. Heine: Lehrbuch der biologischen Medizin. Grundregulation und Extrazellulare Matrix, 2. Aufl. 1997, Hippokrates Verlag Stuttgart, S. 56

<4i>: K. Meyl: Skalarwellenstrahlung. Welche Bedeutung haben Skalarwellen fur die Medizin? GNP-Vortrag am 29.04.2001, Rhein-Main-Halle Wiesbaden, see: Co'med Fachzeitschrift fur Komplementarmedizin, 6/2001, pp.55-60.

I. According to the Maxwell theory:

1. Faraday's law of induction:

$$\boxed{\text{rot}\,\mathbf{E} = -\,\delta\mathbf{B}/\delta t} \quad (26.1)$$

with $E = E(r,t)$

and $H = H(r,t)$: $\mathbf{B} = \mu \cdot \mathbf{H}$ (26.2)

$$\boxed{-\,\text{rot rot}\,\mathbf{E} = \mu \cdot \delta(\text{rot}\,\mathbf{H})/\delta t} \quad (26.3)$$

2. Ampere's law:

$$\boxed{\text{rot}\,\mathbf{H} = \mathbf{j} + \delta\mathbf{D}/\delta t} \quad (26.4)$$

with:

Ohm's law: $\mathbf{j} = \sigma \cdot \mathbf{E}$ (26.5)

dielectric displacement: $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (26.6)

relaxation time: $\tau = \varepsilon/\sigma$ (26.7)

$$\boxed{\text{rot}\,\mathbf{H} = \varepsilon \cdot (\mathbf{E}/\tau + \delta\mathbf{E}/\delta t)} \quad (26.8)$$

3. Inserting equation 26.8 into 26.3 yields:

$$-\,\text{rot rot}\,\mathbf{E} = \mu \cdot \varepsilon \cdot (1/\tau \cdot \delta\mathbf{E}/\delta t + \delta^2\mathbf{E}/\delta t^2) \quad (26.9)$$

with the abbreviation: $\mu \cdot \varepsilon = 1/c^2$ (26.10)

4. Field equation of a damped transverse wave:

$$\boxed{\underbrace{-\,\text{rot rot}\,\mathbf{E} \cdot c^2}_{\text{transverse}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} + \underbrace{(1/\tau) \cdot \delta\mathbf{E}/\delta t}_{\text{vortex damping}}} \quad (26.11)$$

Fig. 26.1: Derivation of the wave damping by means of the formation of vortices!<i>

<i>: see also EMC, part 1, chapter 5.3, 2.borderline case.

26. Recapitulation from the viewpoint of textbook physics

Now that we in the meantime have accumulated innumerable mosaic parts as inspiring contributions to the discussion for the information technical seminar, it is time to sort the ideas and to put the parts together to an overall picture.
Sceptics and orthodox scientists can only be convinced, if we start from textbook physics and completely do without postulates. Those demands will be fulfilled!

26.1 Common misinterpretation of the antenna losses

The mathematical description of physical relations leads to the well-known laws, which shouldn't be doubted anymore as soon as they are accepted to be correct. But what about the interpretation? Although a law dictates the interpretation and there is no choice, because laws must be adhered to, yet textbooks from time to time violate the mathematically dictated interpretation, a circumstance, which can't be accepted. I would like to illustrate this with an example.

Let us assume that the measured degree of effectiveness of a transmitting antenna amounts to 80 percent. There exist better antennas, but also distinctly worse antennas, but I'm not aiming at a certain construction. The statement simply says, that 80% of the fed in HF-power is transformed into Hertzian waves. Thus there arises a loss of power of 20 percent, and the question follows: of what do those 20% consist?

The answer, which is usual among experts and is supported by the textbooks, reads: the antenna wire gets hot and also the air around the antenna is heated by dielectric losses. In short, heat is formed.
But I have to point out and will furnish proof that this interpretation is predominantly wrong! It in any case isn't in accord with the laws of Maxwell. Who namely obeys the laws, comes to an entirely different result!

A short derivation brings it to light (fig. 26.1).
We start with the formulation of Faraday's law of induction according to the textbooks (26.1), apply the curl-operation to both sides of the equation (26.3) and insert in the place of rot H Ampere's law (26.4-26.8). The generally known result describes a damped electromagnetic wave (26.11)<i>.
It on the one hand is a transverse wave, which represents 80% of the antenna power for our example. On the other hand a damping term can be found in the equation, which obviously corresponds to the sought-for 20%. With that the answer would have been found. We realize that because of a damping of the wave 20% antenna losses arise. These losses can't concern heat at all, since the damping term in the equation has got nothing in common with thermodynamics. In the equation doesn't stand anything about heat! Such a mistake!

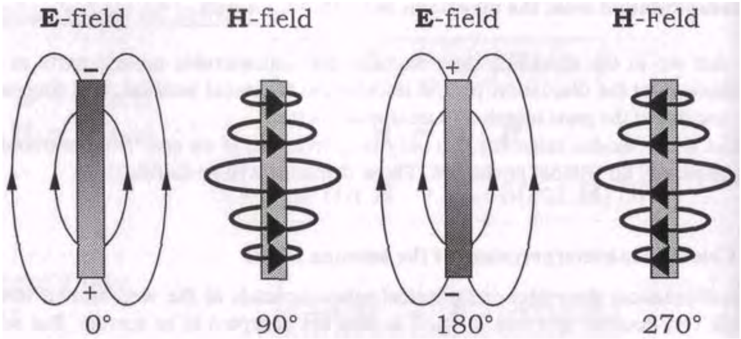
E-field
H-field
E-field
H-Feld
-
+
+
-
0°
90°
180°
270°

Fig. 21.8 A: The fields of the oscillating dipole antenna.<i>

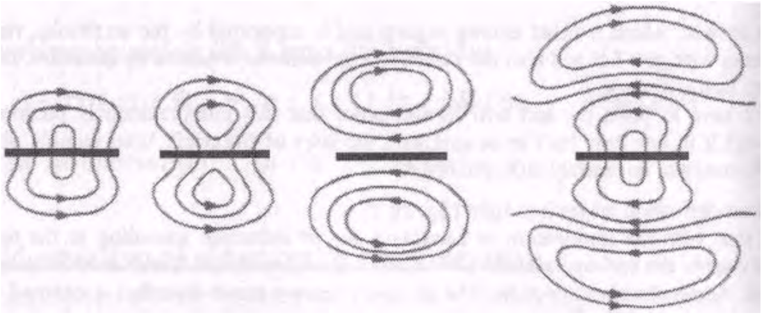

Fig. 21.9 A: The coming off of the electric field lines from the dipole.

<i>: Repetition belonging to chapter 21

26.2 Wave damping by field vortices

- Mathematically seen the damping term describes vortices of the electromagnetic field. This term for instance forms the basis of all eddy current calculations<ii>.
- Physically seen some waves - in our example it is 20 percent - roll up to field vortices, with which the wave damping and the antenna losses would be explained.

In the course of time a substantial part of the generated vortices will fall apart. These thereby will produce eddy losses in form of heat. Thus eventually still heat is produced - agreed. The criticism of the textbooks consists of the circumstance that we by no means can proceed from the assumption that all vortices spontaneously fall apart and a total conversion into heat will take place. The process in addition takes place with a temporal delay. The time constant x gives information in this respect. Field energy is buffered in the vortex, where some vortices live very long and it can't be ruled out that a few even exist as long as you like.

To find out more about these field vortices and their behaviour, one has to get deep into vortex physics. Unfortunately nothing can be found about vortex physics in the textbooks. The mistake is systematic. The following short compendium should help close the gap:

- Mathematically seen a closed-loop electromagnetic field vortex will show as a scalar. Such field vortices, which are mediated by a scalar field, are propagating exactly as charged particles in longitudinal manner as a scalar wave.
- Physically seen a closed-loop field vortex has got particle nature. If one particle kicks off the next one then an impulse is mediated, then a shock wave is occurring, a longitudinal undulation of the particles.

From the vortex physical view the interpretation of the antenna example now sounds entirely different:
The charge carriers in an antenna wire oscillating with high-frequency form a longitudinal shock wave. Between current and tension voltage usually a phase shift of 90° is present. The fields produced by these charge carriers form a scalar wave field in the immediate vicinity of the antenna, the so-called near-field zone, which likewise contains longitudinal field components and shows a phase shift of 90° between electric and magnetic field (fig. 21.8 A). As in textbooks is clarified by field lines, the generated fields actually form vortices, where one structure kicks off the next one (fig. 21.9 A).

The vortices in the near-field zone of an antenna consist of standing waves, which obviously are transforming with increasing distance. In our example 80% of these are unrolling and turn into transverse waves, whereby the characteristic phase angle between E- and H-field at that occasion becomes zero.
Let's turn again to those 20 percent loss.

<ii>: s.a.: K. Meyl: Wirbelstrome, Diss. Uni. Stuttgart 1984, INDEL Verlagsabt.

I. According to the Maxwell theory:

The consistent application of textbook physics

- Longitudinal waves run in the direction of a field pointer.
- The field pointer oscillates, the vector of velocity oscillates along!
- At relativistic velocities field vortices are subject to the Lorentz contraction.
- The faster the oscillating vortex is on its way, the smaller it gets.
- The vortex permanently changes its diameter (see fig. 21.10 B).
- With the diameter the wavelength decreases (see fig. 22.4).
- The swirl velocity is constant (= speed of light c).
- The eigenfrequency of the vortex oscillates with opposite phase to the wavelength.
- The vortex acts as a frequency converter!
- The measurable mixture of frequencies is called noise.

leads to the statement:

- The antenna noise corresponds to the antenna losses!

II. Mathematical description of a wave by the inhomogeneous Laplace equation:

$$\underbrace{\Delta \mathbf{E} \cdot c^2}_{\text{Laplace operator}} = \underbrace{-\,\text{rot rot } \mathbf{E} \cdot c^2}_{\text{transverse (radio wave)}} + \underbrace{\text{grad div } \mathbf{E} \cdot c^2}_{\text{longitudinal (scalar wave)}} = \underbrace{\delta^2 \mathbf{E} / \delta t^2}_{\text{wave}} \qquad (26.12)$$

Divergence E is a scalar!
The corresponding term founds a scalar wave.

Fig. 26.3: Mathematical description of a wave according to Laplace.

26.3 Laplace versus Maxwell

Longitudinal waves as is well-known don't know a fixed velocity of propagation at all. Since they run in the direction of an oscillating field pointer also the vector of velocity will oscillate. For so-called relativistic velocities in the range of the speed of light c the field vortices are subject to the Lorentz contraction. That means, the faster the oscillating vortex is on its way, the smaller it gets. The vortex, as a mediator of a scalar wave carrying impulse, permanently changes its diameter.

Since, in the case of the vortices, it should concern rolled up waves, the vortex velocity will continue to be c, with which the wave now runs around the vortex centre in a circle. From that follows that if the diameter gets smaller, the wavelength of the vortex as well will decrease, whereas the eigenfrequency of the vortex accordingly increases.

If the next moment the vortex oscillates back, the frequency again decreases. The vortex acts as a frequency converter! The mixture of high-frequency signals distributed over a broad frequency band formed in this way, is called noise. A noise signal indeed is measured from the outside with the help of broadband receivers.. We also speak of antenna noise and with this knowledge we can further specify the 20% antenna losses: The antenna produces 20 % noise, which can be put equal to the generated vortices because of the wave damping.

At this point the Maxwell theory doesn't leave us room for interpretation at all. If in the textbooks the impression is aroused, as if the noise were an independent discipline, than that is not true at all. How much the noise is connected with the electromagnetic waves, proves a short look at the wave equation.

The wave equation found in most textbooks has the form of an inhomogeneous Laplace equation. The famous French mathematician Laplace considerably earlier than Maxwell did find a comprehensive formulation of waves and formulated it mathematically (eq. 26.12), which until today is still accepted as valid.

On the one side of the wave equation the Laplace operator stands, which describes the spatial field distribution, and which according to the rules of vector analysis can be decomposed into two parts. On the other side the description of the time dependency of the wave can be found as an inhomogeneous term.

If the wave equation according to Laplace (26.12) is compared to the one, which the Maxwell equations have brought us (26.11), then two differences clearly come forward:

1. In the Laplace equation the damping term is missing. It doesn't describe the formation of vortices, and that means vortices do not exist at all, or present vortices have been there from the beginning.
2. With divergence E a scalar factor appears in the wave equation, which founds a scalar wave.

At this point at once hot tempered discussions concerning the question of the existence of scalar waves blaze up. But this question has already been answered clearly with the vortex consideration. Since an accepted description of longitudinal and scalar waves exists with the plasma wave and the plasma wave can be derived directly without postulate from the term of the wave equation (chapter 21.4/21.5), which founds scalar waves, there are further arguments present for their existence.

I. According to Maxwell:

$$\underbrace{-\operatorname{rot}\operatorname{rot}\mathbf{E}\cdot c^2}_{\text{transverse}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} + \underbrace{(1/\tau)\cdot\delta\mathbf{E}/\delta t}_{\text{vortex damping}} \qquad (26.11)$$

Description of electromagnetic waves with vortex damping
Example: sunlight and the damping in the ionosphere (bleu sky).

II. According to Laplace:

$$\underbrace{\Delta\mathbf{E}\cdot c^2}_{\substack{\text{Laplace}\\ \text{operator}}} = \underbrace{-\operatorname{rot}\operatorname{rot}\mathbf{E}\cdot c^2}_{\substack{\text{transverse}\\ \text{(Hertz)}}} + \underbrace{\operatorname{grad}\operatorname{div}\mathbf{E}\cdot c^2}_{\substack{\text{longitudinal}\\ \text{(Tesla)}}} = \underbrace{\delta^2\mathbf{E}/\delta t^2}_{\text{wave}} \qquad (26.12)$$

Description of transverse electromagnetic waves (Hertzian waves) and longitudinal scalar waves (Tesla radiation).
1st example: propagation of light as a wave or as a particle.
2nd example: useful signal or noise signal of an antenna.

III. Mathematically seen (comparison of coefficients):

$$\underbrace{(1/\tau)\cdot\delta\mathbf{E}/\delta t}_{\text{vortex}} + \underbrace{\operatorname{grad}\operatorname{div}\mathbf{E}\cdot c^2}_{\text{scalar}} = 0 \qquad (26.13)$$

IV. Physically seen
(from the comparison of equation 26.12 with equation 26.11):

Vortices propagate longitudinally as a scalar wave!

Fig. 26.2: ______ Comparison of the two wave descriptions.

26.4 Error term in the wave equation

From the comparison of coefficients of both wave descriptions follows even more:

- Mathematically seen the damping - resp. vortex term according to Maxwell can be put equal to the scalar wave term according to Laplace.
- Physically seen the generated field vortices form and found a scalar wave.

Here also doesn't exist any room for interpretation, as long as we work with the wave equation according to Laplace and at the same time adhere to the Maxwell theory. If however the scalar wave part is put equal to zero, as is common practise in the textbooks, then as a consequence neither vortices nor noise may exist. But that contradicts all measuring technical experience! Since every antenna produces more or less noise, the textbooks obviously only show half the truth. Science however gropes for the whole truth and that should be fathomed.

If in the case of the antenna example the vortex part amounts to 20%, then that's tantamount to 20% scalar wave part, resp. 20% noise. The scalar wave part constitutes with regard to the Hertzian useful wave something like an error term in the wave equation. The part definitely is too big, as that it might be put equal to zero. Even so all error consideration in the textbooks is missing, if the scalar wave term is assumed to be zero. That violates all rules of physics and of taught scientific methodism.

In practice this shows by a useful signal going under in the noise and reception not being possible anymore as soon as the scalar wave part gets out of control. Even in this case, for which the degree of effectiveness tends towards zero, it still is common practise to put the error term, which is dominating everything, equal to zero. But who in this point follows the textbooks, disregards with that the wave equation and doing so by no means can refer to the circumstance that all colleagues make the same mistake.

The building of physics behaves like a house of cards, where the cards mutually support each other. Perhaps that is the deeper reason why those, who have discovered a marked card, don't pull it out immediately. In addition they are hindered by the self appointed guardians of the ,,pure doctrine", since everyone knows what happens with the house of cards if the marked card is pulled out. Only, do we want to and can we live with that in the long run? Is it a solution of the problem, if the so-called experts among the physicists and technicians look away and don't deal with the foundation of their branch anymore? If universities crash their basis education into the wall and choke off every contradiction?

Please allow me to pull out the marked card now and place it on the table!

It concerns the question: what is the nature of the field vortices, which form a scalar wave in space. Eddy currents in the iron parts of the antenna are explained with the field equations, but not the noise, which is measured especially in the air. If an antenna on the one hand produces field vortices and as a consequence eddy losses and on the other hand dielectric losses, then we can assume that besides the eddy currents in the conductor also vortices in the dielectric must exist. Let's search for them!

Interim result (comparison of arguments):

The Maxwell equations on the one hand dictate that as the reason for a wave damping only field vortices should be considered.	On the other hand the same laws merely describe eddy currents, which can only occur in the electrically conducting parts of the antenna.
On the one hand the field vortex interpretation makes it possible to explain the noise of an antenna perfectly.	On the other hand does the noise appear in the neighbourhood of the antenna, thus in the air and not in the iron parts!
The mathematical formulation reveals, how wave and vortex, resp. noise, co-operate and how one should imagine the conversion of one form into the other form.	In field physics on the other hand is missing a useful description of electric field vortices in a dielectric, which could found the noise signal.

Table 26.5: Arguments pro and contra.

26. 5 Interim result

It shouldn't be a disadvantage, to interprete physical laws more consistently than usual, even if in the present case orthodox science through that at first should fall into a deep crisis. If the way is worthwhile, only will show at the end.
Let us try to work out the contradictions in form of a comparison of arguments:

- The Maxwell equations on the one hand dictate that as the reason for a wave damping only field vortices should be considered.
  - On the other hand the same laws merely describe eddy currents, which can only occur in the electrically conducting parts of the antenna. .
- On the one hand the field vortex interpretation makes it possible to explain the noise of an antenna perfectly. .
  - On the other hand does the noise appear in the neighbourhood of the antenna, thus in the air and not in the iron parts!
- The mathematical formulation reveals, how wave and vortex, resp. noise, co-operate and how one should imagine the conversion of one form into the other form.
  - In field physics on the other hand is missing a useful description of electric field vortices in a dielectric, which could found the noise signal.

The most obvious among all conceivable solutions is the one that we have to assume the existence of dielectric field vortices, so-called potential vortices. We are challenged to search for a corresponding description. If the quest should be successful, then the contradictions would be overcome. In addition there is the promise of a whole number of simplifying explanations of various phenomena in the dielectric (see fig. 26.5 and fig. 26.7).

The phenomenon of noise becomes an aspect of wave physics, which is more than merely a field disturbance, which makes the reception of the useful wave more difficult. If the scalar wave nature is realized, then applications can be imagined, in which the noise is used as useful signal. In the way that the scalar part in the wave equation doesn't have to be put to zero anymore to obtain freedom of contradiction, even optimizations of antennas or of capacitors are possible with regard to the dielectric losses by means of the calculation of the scalar part.

New in any case is the idea that the dielectric losses of a capacitor are eddy losses and not a defect in material of the insulating material. With that the capacitor losses correspond to a generated noise power. We also can say, every capacitor more or less produces noise! The electric field lines point from one capacitor plate to the other plate. If one plate radiates as a transmitter and the other plate functions as a receiver, then the field propagation takes place in the direction of the electric field pointer and that again is the condition for a longitudinal wave. Here the circle closes in the conclusion, the capacitor field mediates dielectric field vortices, which following the field lines found a scalar wave because of their scalar nature. The heating of the capacitor results from the decay of vortices.

Potential vortices explain div. phenomena in the dielectric:

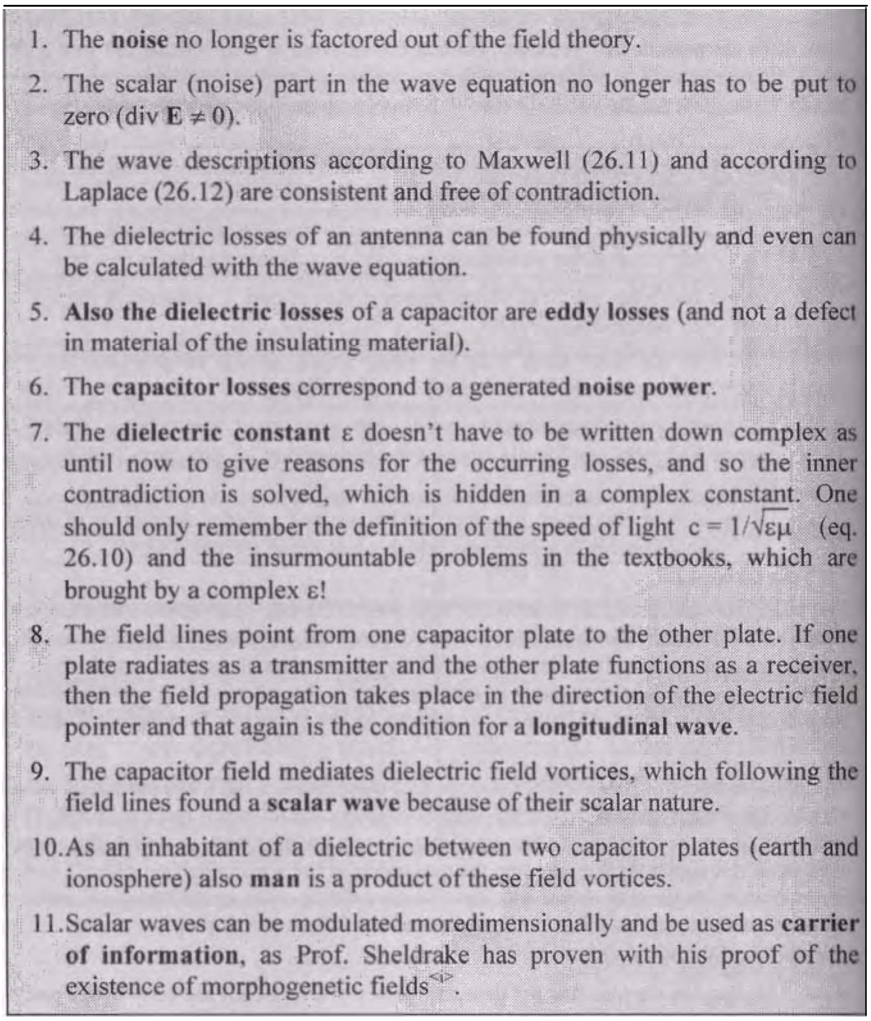

1. The **noise** no longer is factored out of the field theory.
2. The scalar (noise) part in the wave equation no longer has to be put to zero (div $\mathbf{E} \neq 0$).
3. The wave descriptions according to Maxwell (26.11) and according to Laplace (26.12) are consistent and free of contradiction.
4. The dielectric losses of an antenna can be found physically and even can be calculated with the wave equation.
5. **Also the dielectric losses** of a capacitor are **eddy losses** (and not a defect in material of the insulating material).
6. The **capacitor losses** correspond to a generated **noise power**.
7. The **dielectric constant** ε doesn't have to be written down complex as until now to give reasons for the occurring losses, and so the inner contradiction is solved, which is hidden in a complex constant. One should only remember the definition of the speed of light $c = 1/\sqrt{\varepsilon\mu}$ (eq. 26.10) and the insurmountable problems in the textbooks, which are brought by a complex ε!
8. The field lines point from one capacitor plate to the other plate. If one plate radiates as a transmitter and the other plate functions as a receiver, then the field propagation takes place in the direction of the electric field pointer and that again is the condition for a **longitudinal wave**.
9. The capacitor field mediates dielectric field vortices, which following the field lines found a **scalar wave** because of their scalar nature.
10. As an inhabitant of a dielectric between two capacitor plates (earth and ionosphere) also **man** is a product of these field vortices.
11. Scalar waves can be modulated moredimensionally and be used as **carrier of information**, as Prof. Sheldrake has proven with his proof of the existence of morphogenetic fields<i>.

Fig. 26.6: ___ Advantages of a field description extended with potential vortices.

<i>: Rupert Sheldrake: Seven Experiments That Could Change the World. Riverhead Books, 1995

26.6 Failure of the Maxwell theory

If the capacitor losses or the antenna noise should concern dielectric losses in the sense of vortex decay of potential vortices, which don't occur in the Maxwell theory at all, then we are confronted with a massive contradiction:

- For the description of the losses the Maxwell; theory on the one hand only offers field vortices and those only in the conductive medium.
  - On the other hand do the dielectric losses occut in the nonconductor and in the air.

- In conductive materials vortex fields occur, in the insulator however the fields are irrotational. That isn't possible, since at the transition from the conductor to the insulator the laws of refraction are valid and these require continuity! Hence a failure of the Maxwell theory (Fig. 26.7) will occur in the dielectric.
  - As a consequence the existence of vortex fields in the dielectric, so-called potential vortices, should be required!

In electrodynamics as a help the approach of a vector potential A is used, which leads to a complex formulation of the dielectric constant ε and in this way makes it possible, to mathematically describe the dielectric losses of a capacitor by means of the load angle, which stretches in the complex plane. But which physical value does this approach have? How can now the inner contradiction be explained, which is hidden in a complex constant of material? One should only remember the definition of the speed of light $c = 1/\sqrt{\varepsilon\mu}$ (eq. 26.10) and its dependency of ε. For a complex ε here are resulting insurmountable problems in the textbooks.
From the viewpoint of mathematics the introduction of the vector potential at first may represent a help. The before mentioned contradictions however fast raise doubts to the model concept, which from a physical viewpoint eventually will lead to errors, if the speed of light isn't constant anymore and even should be complex.

These considerations should be sufficient as a motive to require potential vortices, even if for their description the field theory according to the textbooks has to be revised. As a supplement there is pointed to the following points:

- As an inhabitant of a dielectric between two capacitor plates also man is a product of these field vortices.
- Scalar waves can be modulated moredimensionally and be used as carrier of information, as Prof. Sheldrake has proven with his proof of morphogenetic fields<i>.

The dielectric vortices moreover provide an explanation for natural events. They form the key to numerous disciplines of science, from physics over biology up to medicine.

<i>: Rupert Sheldrake: Seven Experiments That Could Change the World. Riverhead Books, 1995

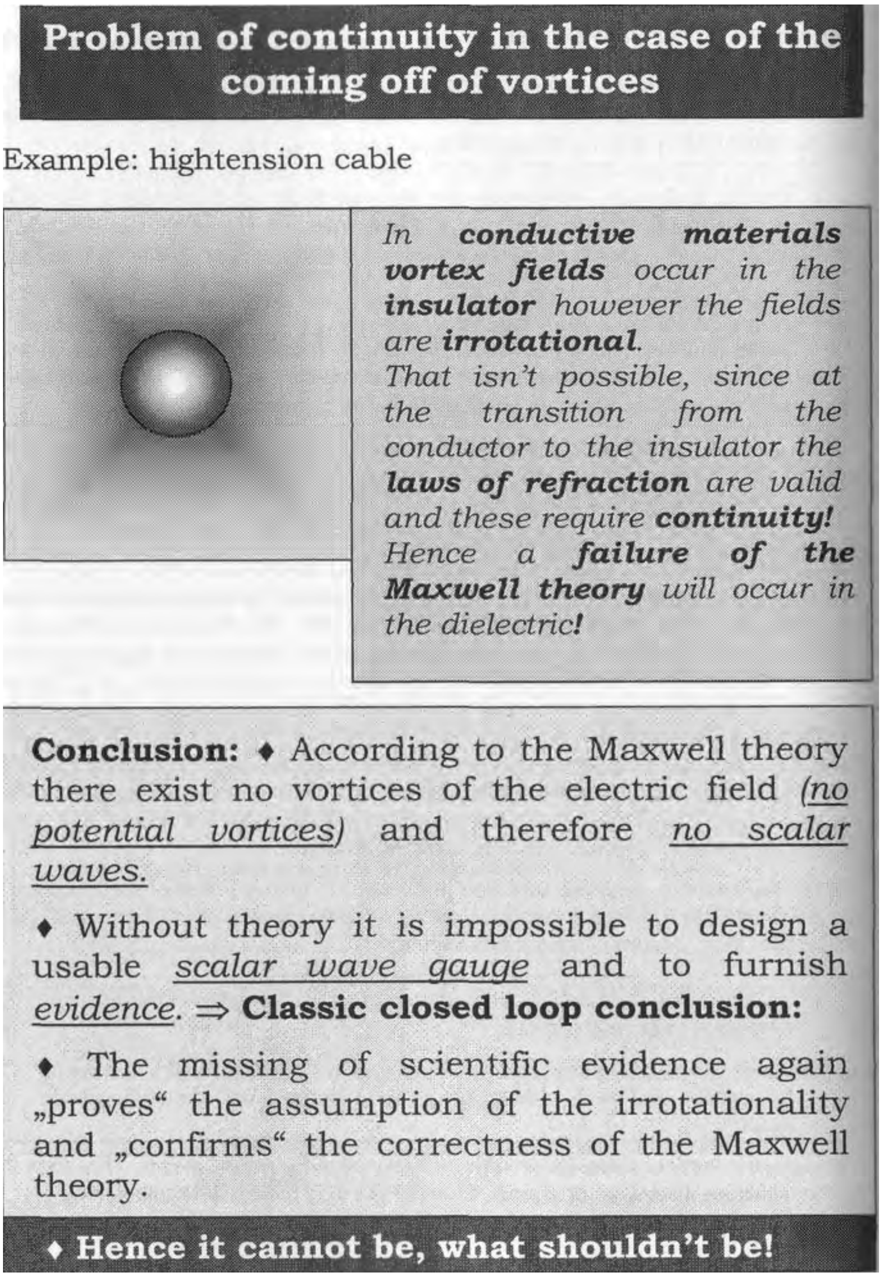

Problem of continuity in the case of the coming off of vortices

Example: hightension cable

In ***conductive materials vortex fields*** *occur in the* ***insulator*** *however the fields are* ***irrotational****.*
That isn't possible, since at the transition from the conductor to the insulator the ***laws of refraction*** *are valid and these require* ***continuity!***
Hence a ***failure of the Maxwell theory*** *will occur in the dielectric!*

Conclusion: ♦ According to the Maxwell theory there exist no vortices of the electric field *(no potential vortices)* and therefore *no scalar waves.*

♦ Without theory it is impossible to design a usable *scalar wave gauge* and to furnish *evidence*. ⇒ **Classic closed loop conclusion:**

♦ The missing of scientific evidence again „proves“ the assumption of the irrotationality and „confirms“ the correctness of the Maxwell theory.

♦ **Hence it cannot be, what shouldn't be!**

Fie. 26.7: Concerning the failure of the Maxwell theory

26.7. Concerning the evidence situation

In the question, if a physical phenomenon should be acknowledged as such, experimental, mathematical and physical evidence should be shown. In the case of the potential vortices, the vortices of the electric field and their propagation as a scalar wave, the historical experiments of Nikola Tesla<i> and the modern clone of these can be judged as experimental evidence<ii>.

With the well-known wave equation a mathematical description for this phenomenon has been specified and discussed<iii>. It will be shown that both transverse and longitudinal wave parts are contained alongside in the wave equation, i.e. both radio waves according to Hertz and scalar waves according to Tesla. Doing so the mathematically determined scalar wave properties are identical with the experimental results!
The wave equation is an inhomogeneous Laplace equation and the first and oldest description of scalar waves. It thereby is unimportant, if the famous mathematician Laplace himself already may have realized and discussed this circumstance or not. The description fits perfectly and that is what counts!

At this point the third point should be put on the agenda, the physical evidence. This is connected very closely with the question for a suitable field theory and that again is basing on a corresponding approach.

<i>: N. Tesla: Apparatus for Transmission of Electrical Energy, US-Patent-No.: 649,621, New York 1900, Dr. Nikola Tesla: Complete Patents pp318-321.

<ii>: Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, Argo-Verlag, Marktoberdorf, l.Aufl. 2000.

<iii>: K. Meyl: Scalar Waves: Theory and Experiments, Journal of Scientific Exploration, Vol. 15, No. 2, pp. 199-205, 2001

(* * * Task Schedule * * *)

1. Neutrino radiation => energy radiation!
   (acc. to Pauli: ν cares for missing energy in the case of the ß-decay)
2. High neutrino density => high energy density!
3. Neutrino = particle without charge or mass (mean), but because of oscillations: effective value of charge and mass ≠ zero!
4. Interaction only in the case of resonance (e.g. weak interaction)!
   (same frequency, 180° phase shift)
5. Neutrino radiation is an energy source which can be used (not a question of physics, only a question of technology!)
6. Particle radiation (neutrino) => shock wave (like sound)
   => longitudinal wave
   => scalar wave (mathem.)
7. Interaction/resonance of the v = scalar wave problem
8. Scalar waves are a problem of the field theory!
9. Maxwell theory neglects scalar waves.

- Search for a new approach (chap. 27.8)

- Derivation of the Maxwell equations as a

- Derivation of scalar waves (chap. 27.13)
- Derivation of the gravitation and more

Table 27.1: Task schedule belonging to chapter 27

27. Faraday versus Maxwell

Numerous phenomena of the electromagnetic field are described sufficiently accurate by the Maxwell equations, so that these as a rule are regarded as a universal field description. But if one looks more exact it turns out to be purely an approximation, which in addition leads to far reaching physical and technological consequences. We must ask ourselves:

- What is the Maxwell approximation?
- How could a new and extended approach look like?
  - Faraday instead of Maxwell, which is the more general law of induction?
- Can the Maxwell equations be derived as a special case?
- Can also scalar waves be derived from the new approach?
- Can the gravitation as well be derived and a lot else more?

On the one hand it concerns the big search for a unified physical theory and on the other hand the chances of new technologies, which are connected with an extended field theory. As a necessary consequence of the derivation, which roots strictly in textbook physics and manages without postulate, scalar waves occur, which could be used manifold. In information technology they are suited as a carrier wave, which can be modulated moredimensionally, and in power engineering the spectrum stretches from the wireless transmission up to the collection of energy out of the field.

27.1 Energy out of the field

Neutrinos for instance are such field configurations, which move through space as a scalar wave. They were introduced by Pauli as massless but energy carrying particles to be able to fulfil the balance sheet of energy for the beta decay. Nothing would be more obvious than to technically use the neutrino radiation as an energy source.

But for a technical exploitation a useful model description of the particles and their interaction is imperative. For the sake of simplicity we imagine the neutrino to be an oscillating particle, which permanently oscillates back and forth between the state of an electron and that of a positron. With that the polarity changes from positive to negative and back again and the charge averaged over time is zero. Because of the change from a state of matter to the state of an anti-particle also next to no mass can be measured anymore.

A technical oscillator operated in resonance, which oscillates with the same frequency but opposite charge, will interact with the particle and build up an oscillating electromagnetic interaction, with which we already are familiar as the weak interaction in the proximity of a neutrino.

Tubular vortices

- Examples: drain vortex (bathtub vortex)
  whirlwind and waterspout, tornado (image).

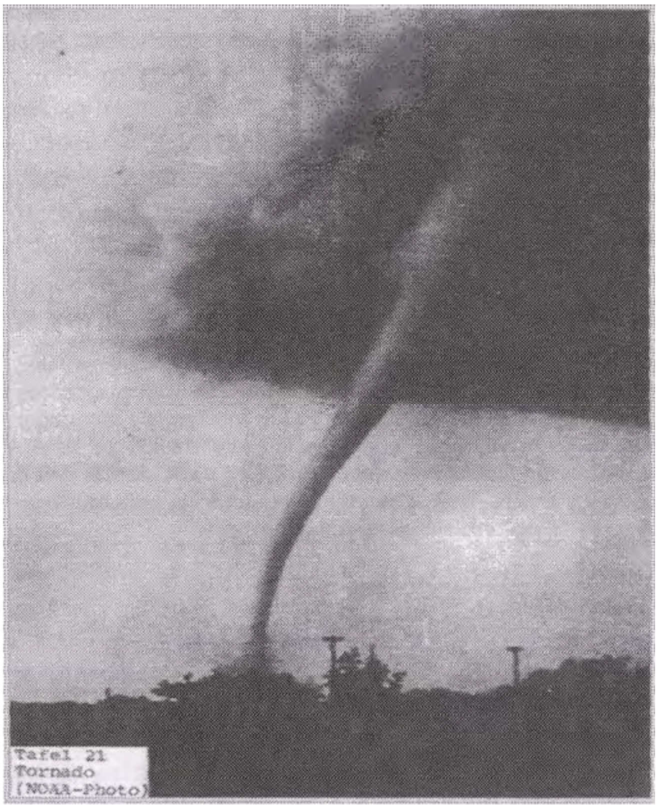

Inside:	expanding vortex
Outside:	contracting anti-vortex
Condition	for coming off: equally powerful vortices
Criterion:	viscosity
Result:	tubular structure

Fig. 27.2: ____Vortex and anti-vortex - a physical basic principle

<i>: Lugt: Wirbelstromung in Natur und Technik, G. Braun Verlag Karlsruhe 1979, Bild „Tornado" von Tafel 21, Seite 356

The propagation of particle radiation as a longitudinal shock wave however can't be described with the normally used field theory and the Maxwell equations, so that the field theory at this point must be reworked. Connected with this is the question of what is oscillating here, a question, which often is answered with an aether of whatever nature. I speak of field vortices and call the aether a property of the field. With that the set of difficulties is shifted into the domain of vortex physics.

27.2 Vortex and anti-vortex

In the eye of a tornado the same calm prevails as at great distance, because here a vortex and its anti-vortex work against each other. In the inside the expanding vortex is located and on the outside the contracting anti-vortex. One vortex is the condition for the existence of the other one and vice versa. Already Leonardo da Vinci knew both vortices and has described the dual manifestations (chapter 3.4).

In the case of flow vortices the viscosity determines the diameter of the vortex tube where the coming off will occur. If for instance a tornado soaks itself with water above the open ocean, then the contracting potential vortex is predominant and the energy density increases threateningly. If it however runs overland and rains out, it again becomes bigger and less dangerous.

The conditions for the bathtub vortex are similar. Here the expanding vortex consists of air, the contracting vortex however of water. In flow dynamics the relations are understood. They mostly can be seen well and observed without further aids.

In electrical engineering it's different: here field vortices remain invisible and not understood. Only so the Maxwell theory could find acceptance, although it only describes mathematically the expanding eddy current and ignores its anti-vortex. I call the contracting anti-vortex potential vortex" and point to the circumstance, that every eddy current entails the anti-vortex as a physical necessity.

Because the size of the forming structures is determined by the electric conductivity, in conducting materials the vortex rings, being composed of both vortices, are huge, whereas they can contract down to atomic dimensions in nonconductors. Only in semiconducting and resistive materials the structures occasionally can be observed directly (fig. 4.8).

Spherical vortices

Examples:	expanding vortex	contracting vortex
• quantum physics	collision processes (several quarks)	gluons (postulate!)
• nuclear physics	repulsion of like	strong interaction (postulate!)
• atomic physics	centrifugal force of the enveloping electrons	electrical attraction Schrodinger equation
• astro-physics	centrifugal force (inertia)	gravitation (can not be derived?!)

Example: elementary particles as electromagnetic field vortices

Inside:	expanding eddy current (skin effect)
Outside:	contracting anti-vortex (potential vortex)
Condition	for coming off: equally powerful vortices
Criterion:	electric conductivity (determines diameter)
Result:	spherical structure (consequence of the pressure of the vacuum)

Fig. 27.3: Spherical structures as a result of contracting potential vortices.<i>

Demokrit (460-370 BC) equated the vortex concept with „law of nature"!
= the first attempt to formulate a unified physics.

<i>: see Part 1, chapter 4.3

27.3 Vortices in the microcosm and macrocosm

The approximation, which is hidden in the Maxwell equations, thus consists of neglecting the anti-vortex dual to the eddy current. It is possible that this approximation is allowed, as long as it only concerns processes inside conducting materials. If we however get to insulating materials the Maxwell approximation will lead to considerable errors and it won't be able to keep it anymore.

If we take as an example the lightning and ask how the lightning channel is formed: Which mechanism is behind it, if the electrically insulating air for a short time is becoming a conductor? From the viewpoint of vortex physics the answer is obvious: The potential vortex, which in the air is dominating, contracts very strong and doing so squezes all air charge carriers and air ions, which are responsible for the conductivity, together at a very small space to form a current channel.

The contracting potential vortex thus exerts a pressure and with that forms the vortex tube. Besides the cylindrical structure another structure can be expected. It is the sphere, which is the only form, which can withstand a powerful pressure if that acts equally from all directions of space. Only think of ball lightning. Actually the spherical structure is mostly found in microcosm till macrocosm. Let's consider some examples and thereby search for the expanding and contracting forces (fig. 27.2).

- In quantum physics one imagines the elementary particles to be consisting of quarks. Irrespective of the question, which physical reality should be attributed to this model concept, one thing remains puzzling: The quarks should run apart, or you should try to keep together three globules, which are moving violently and permanently hitting each other. For this reason glue particles were postulated, the so-called gluons, which now should take care for the reaction force, but this reaction force is nothing but a postulate!

- In nuclear physics it concerns the force, which holds together the atomic nucleus, which is composed of many nucleons, and gives it the well-known great stability, although here like charged particles are close together. Particles, which usually repel each other. Between the theoretical model and practical reality there is an enormous gap, which should be overcome by introducing of a new reaction force. But also the nuclear force, called strong interaction, is nothing but a postulate!

- In atomic physics the electric force of attraction between the positive nuclear charge and the negatively charged enveloping electrons counteracts the centrifugal force. In this case the anti-vortex takes care for a certain structure of the atomic hull, which obey the Schrodinger equation as eigenvalue solutions. But also this equation irrespective of its efficiency until today purely is a mathematical postulate, as long as its origin is not clear.

- In astrophysics centrifugal force (expansion) as a result of the inertia and gravitation (contraction) as a result of the attraction of masses are balanced. But the "gravitation" puts itself in the way of every attempt to formulate a unified field theory. Also this time it is the contracting vortex, of which is said it can't be derived nor integrated.

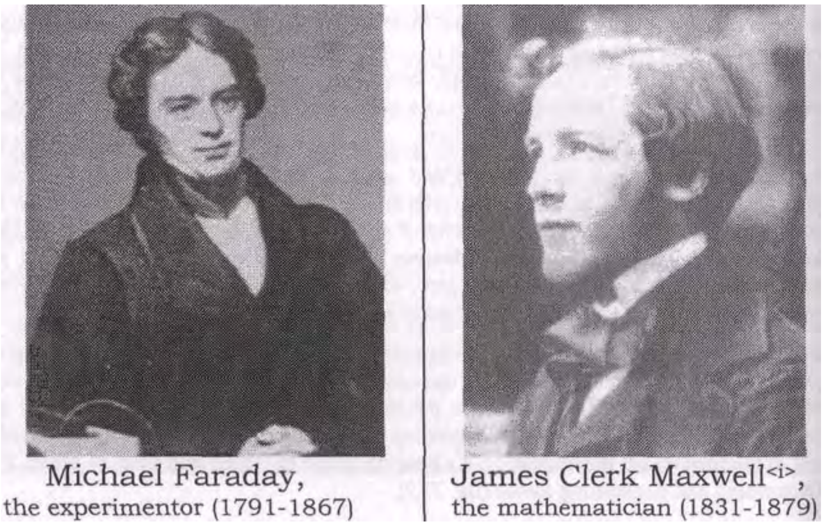

Michael Faraday, the experimentor (1791-1867)

James Clerk Maxwell<i>, the mathematician (1831-1879)

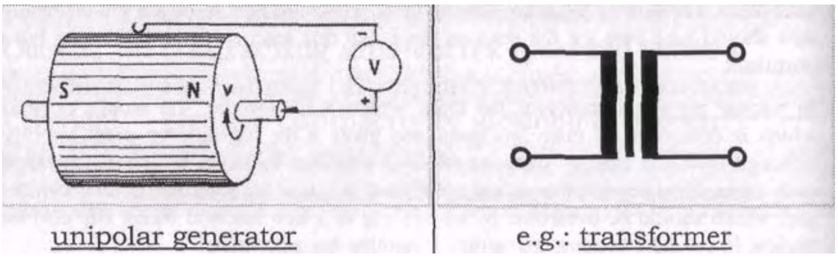

unipolar generator

e.g.: transformer

Fig. 27.4: ____The fathers of the law of induction

<i>: J.C. Maxwell: A treatise on Electricity and Magnetism, Dover Publications

It is remarkable how in the domain of the contracting vortex the postulates are accumulating. But this hasn't always been the case. In ancient Greece already 2400 years ago Demokrit has undertaken an attempt to formulate a unified physics. He traced all visible and observable structures in nature back to vortices, each time formed of vortex and anti-vortex. This phenomenon appeared him to be so fundamental, that he put the term "vortex" equal to the term for „law of nature". The term ,.atom" stems from Demokrit (460-370 BC).

Seen this way the physicists in ancient times already had been further than today's physics, which with the Maxwell approximation neglects the contracting vortex and with that excludes fundamental phenomena from the field description or is forced to replace them by model descriptions and numerous postulates.

What we need is a new field approach, which removes this flaw and in this point reaches over and above the Maxwell theory.

27.4 Faraday's law and Maxwell's formulation

In the choice of the approach the physicist is free, as long as the approach is reasonable and well founded. In the case of Maxwell's field equations two experimentally determined regularities served as basis: on the one hand Ampere's law and on the other hand the law of induction of Faraday. The mathematician Maxwell thereby gave the finishing touches for the formulations of both laws. He introduced the displacement current D and completed Ampere's law accordingly, and that without a chance of already at his time being able to measure and prove the measure. Only after his death this was possible experimentally, what afterwards makes clear the format of this man.

In the formulation of the law of induction Maxwell was completely free, because the discoverer Michael Faraday had done without specifications. As a man of practice and of experiment the mathematical notation was less important for Faraday. For him the attempts with which he could show his discovery of the induction to everybody, e.g. his unipolar generator, stood in the foreground.

His 40 years younger friend and professor of mathematics Maxwell however had something completely different in mind. He wanted to describe the light as an electromagnetic wave and doing so certainly the wave description of Laplace went through his mind, which needs a second time derivation of the field factor. Because Maxwell for this purpose needed two equations with each time a first derivation, he had to introduce the displacement current in Ampere's law and had to choose an appropriate notation for the formulation of the law of induction to get to the wave equation.

His light theory initially was very controversial. Maxwell faster found acknowledgement for bringing together the teachings of electricity and magnetism and the representation as something unified and belonging together<i> than for mathematically giving reasons for the principle discovered by Faraday.

Nevertheless the question should be asked, if Maxwell has found the suitable formulation, if he has understood 100 percent correct his friend Faraday and his discovery. If discovery (from 29.08.1831) and mathematical formulation (1862) stem from two different scientists, who in addition belong to different disciplines, misunderstandings are nothing unusual. It will be helpful to work out the differences.

Law of induction

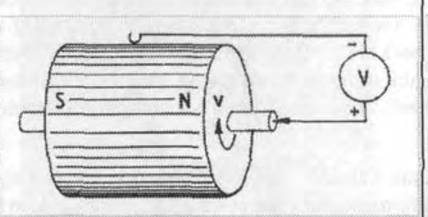

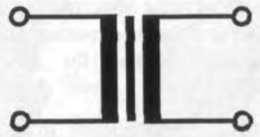

unipolar generator	e.g.: transformer
discovery of Faraday	2nd Maxwell equation
$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1)	$\boxed{\text{rot } \mathbf{E} = - d\mathbf{B}/dt}$ (27.1*)
unipolar induction	law of induction generally

Fig. 27.5: Two formulations for one law
As a mathematical relation between
the vectors of the electric field strength **E**
and the induction **B** (= magnetic flux density)

Consequences
(the physically mean of the „field concept"):

• The **Faraday approach** describes a field physical principle	• **Maxwell equations** describe the fields of charged particles
• field = experience: We experience the electric as a magnetic field and vice versa as a result of a relative velocity • Particles do not occur in the Faraday approach	• the electric (**E**-) field = static field of charges, • the magnetic (**H**-) field = field of moving charges. • The origin of the particles remains unsettled (postulate)
• Perfect duality between **E**- and **H**-field.	• No duality in the field description
• **v** leads to dual permutation of **E**- and **H**-field.	• Charge carriers=el. monopoles • No magnetic monopoles!

27.5 The discovery of Faraday

If one turns an axially polarized magnet or a copper disc situated in a magnetic field, then perpendicular to the direction of motion and perpendicular to the magnetic field pointer a pointer of the electric field will occur, which everywhere points axially to the outside. In the case of this by Faraday developed unipolar generator hence by means of a brush between the rotation axis and the circumference a tension voltage can be called off<i>. The mathematically correct relation E = v x B I call Faraday-law, even if it only appears in this form in the textbooks later in time<ii>. The formulation usually is attributed to the mathematician Hendrik Lorentz, since it appears in the Lorentz force in exactly this form. Much more important than the mathematical formalism however are the experimental results and the discovery by Michael Faraday, for which reason the law concerning unipolar induction is named after the discoverer.

Of course we must realize that the charge carriers at the time of the discovery hadn't been discovered yet and the field concept couldn't correspond to that of today. The field concept was an abstracter one, free of any quantization.

That of course also is valid for the field concept advocated by Maxwell, which we now contrast with the „Faraday-law" (fig. 27.4). The second Maxwell equation, the law of induction (27.1*), also is a mathematical description between the electric field strength E and the magnetic induction B. But this time the two aren't linked by a relative velocity v. In that place stands the time derivation of B, with which a change in flux is necessary for an electric field strength to occur. As a consequence the Maxwell equation doesn't provide a result in the static or quasi-stationary case, for which reason it in such cases is usual, to fall back upon the unipolar induction according to Faraday (e.g. in the case of the Hall-probe, the picture tube, etc.). The falling back should only remain restricted to such cases, so the normally used idea. But with which right the restriction of the Faraday-law to stationary processes is made?

The vectors E and B can be subject to both spatial and temporal fluctuations. In that way the two formulations suddenly are in competition with each other and we are asked, to explain the difference, as far as such a difference should be present.

<i>: Part 2, INDEL 1996, Chap. 16.1

<ii>: among others in R.W. Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 76 und 130

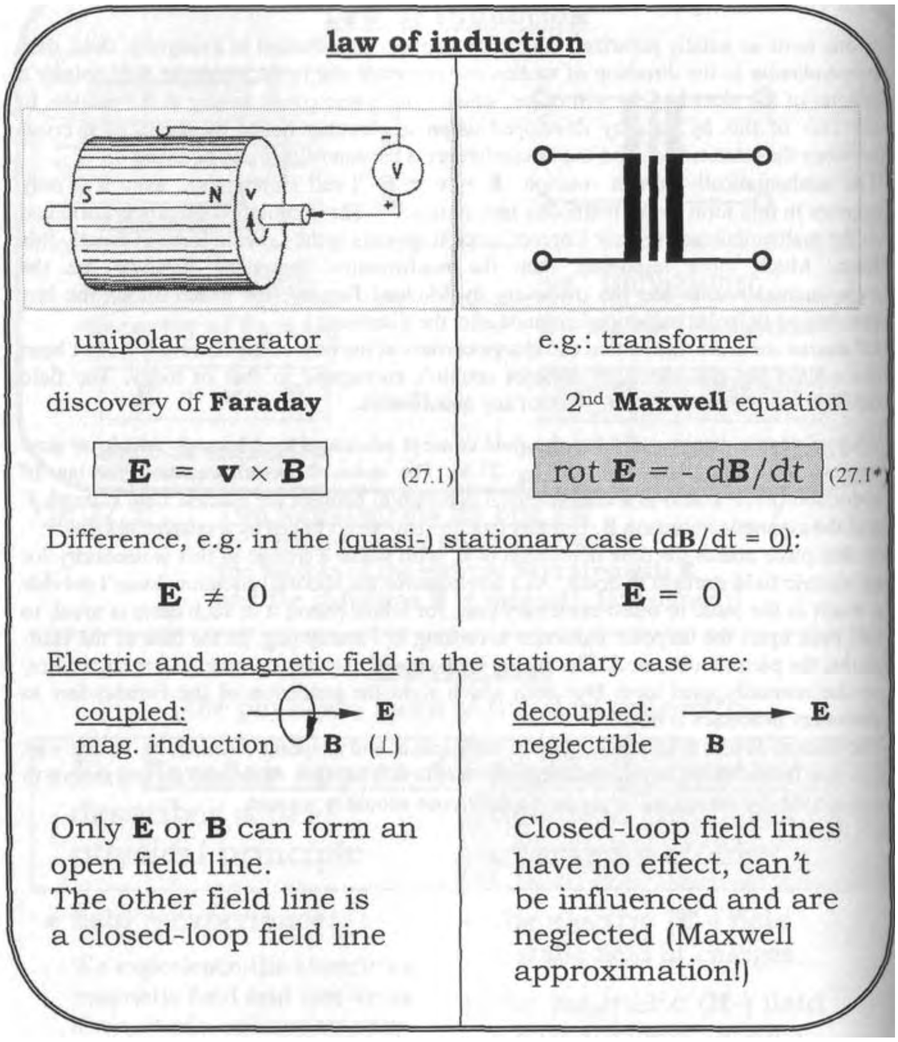

law of induction

unipolar generator	e.g.: transformer
discovery of **Faraday**	2nd **Maxwell** equation
$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1)	$\boxed{\text{rot}\, \mathbf{E} = - d\mathbf{B}/dt}$ (27.1*)

Difference, e.g. in the (quasi-) stationary case (d**B**/dt = 0):

$\mathbf{E} \neq 0$	$\mathbf{E} = 0$

Electric and magnetic field in the stationary case are:

coupled: $\xrightarrow{} \mathbf{E}$	decoupled: $\xrightarrow{} \mathbf{E}$
mag. induction $\mathbf{B}$ (⊥)	neglectible $\mathbf{B}$
Only **E** or **B** can form an open field line. The other field line is a closed-loop field line	Closed-loop field lines have no effect, can't be influenced and are neglected (Maxwell approximation!)

Fig. 27.6: ____Law of induction according to Faraday or Maxwell?

27.6 Different formulation of the law of induction

Such a difference for instance is, that it is common practice to neglect the coupling between the fields at low frequencies. While at high frequencies in the range of the electromagnetic field the E- and the H-field are mutually dependent, at lower frequency and small field change the process of induction drops correspondingly according to Maxwell, so that a neglect seems to be allowed. Now electric or magnetic field can be measured independently of each other. Usually is proceeded as if the other field is not present at all.

That is not correct. A look at the Faraday-law immediately shows that even down to frequency zero always both fields are present. The field pointers however stand perpendicular to each other, so that the magnetic field pointer wraps around the pointer of the electric field in the form of a vortex ring in the case that the electric field strength is being measured and vice versa. The closed-loop field lines are acting neutral to the outside; they hence need no attention, so the normally used idea. It should be examined more closely if this is sufficient as an explanation for the neglect of the not measurable closed-loop field lines, or if not after all an effect arises from fields, which are present in reality.

Another difference concerns the commutability of E- and H-field, as is shown by the Faraday-generator, how a magnetic becomes an electric field and vice versa as a result of a relative velocity v. This directly influences the physical-philosophic question: What is meant by the electromagnetic field?

The textbook opinion based on the Maxwell equations names the static field of the charge carriers as cause for the electric field, whereas moving ones cause the magnetic field. But that hardly can have been the idea of Faraday, to whom the existence of charge carriers was completely unknown. The for his contemporaries completely revolutionary abstract field concept based on the works of the Croatian Jesuit priest Boscovich (1711-1778). In the case of the field it should less concern a physical quantity in the usual sense, than rather the experimental experience"' of an interaction according to his field description. We should interprete the Faraday-law to the effect that we experience an electric field, if we are moving with regard to a magnetic field with a relative velocity and vice versa.

In the commutability of electric and magnetic field a duality between the two is expressed, which in the Maxwell formulation is lost, as soon as charge carriers are brought into play. Is thus the Maxwell field the special case of a particle free field? Much evidence points to it, because after all a light ray can run through a particle free vacuum. If however fields can exist without particles, particles without fields however are impossible, then the field should have been there first as the cause for the particles. Then the Faraday description should form the basis, from which all other regularities can be derived. What do the textbooks say to that?

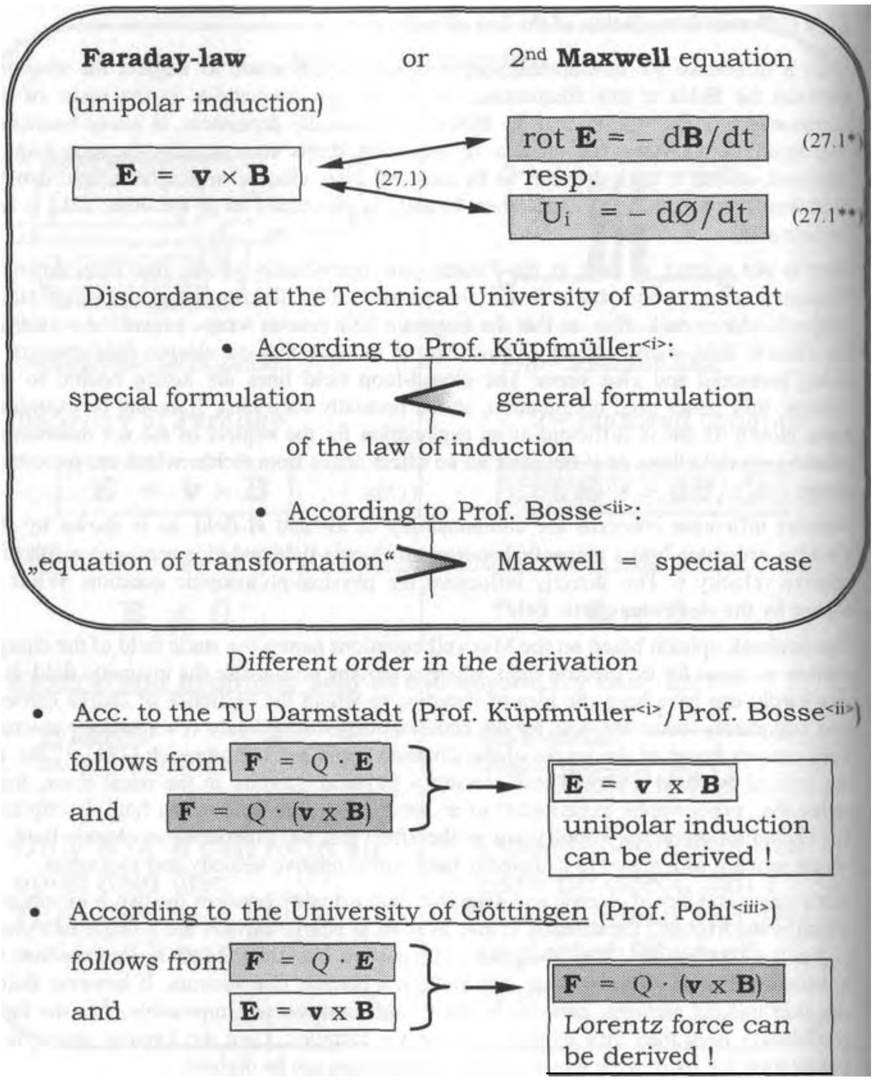

Faraday-law or 2nd **Maxwell** equation

(unipolar induction)

$\mathbf{E} = \mathbf{v} \times \mathbf{B}$ (27.1)

$\text{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt$ (27.1*)

resp.

$U_i = -\,d\text{Ø}/dt$ (27.1**)

Discordance at the Technical University of Darmstadt

- According to Prof. Küpfmüller<i>:

special formulation < general formulation

of the law of induction

- According to Prof. Bosse<ii>:

„equation of transformation“ > Maxwell = special case

Different order in the derivation

- Acc. to the TU Darmstadt (Prof. Küpfmüller<i>/Prof. Bosse<ii>)

follows from $\mathbf{F} = Q \cdot \mathbf{E}$

and $\mathbf{F} = Q \cdot (\mathbf{v} \times \mathbf{B})$ → $\mathbf{E} = \mathbf{v} \times \mathbf{B}$

unipolar induction can be derived !

- According to the University of Göttingen (Prof. Pohl<iii>)

follows from $\mathbf{F} = Q \cdot \mathbf{E}$

and $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ → $\mathbf{F} = Q \cdot (\mathbf{v} \times \mathbf{B})$

Lorentz force can be derived !

Fig. 27.7: Different opinions and derivations

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12. Auflage, Springer Verlag 1988, Seite 228, Gl. 22.

<ii>: G. Bosse: Grundlagen der Elektrotechnik II, BI-Hochschultaschenbucher Nr. 183, l.Aufl. 1967, Kap. 6.1 Induktion, Seite 58

<iii>: R. W. Pohl: Einfuhrung in die Physik, Band 2 Elektrizitatslehre, 21. Auflage, Springer-Verlag 1975, Seite 77

27.7 Contradictory opinions in textbooks

Obviously there exist two formulations for the law of induction (27.1 and 27.1*), which more or less have equal rights. Science stands for the question: which mathematical description is the more efficient one? If one case is a special case of the other case, which description then is the more universal one?

What Maxwell's field equations tell us is sufficiently known, so that derivations are unnecessary. Numerous textbooks are standing by, if results should be cited. Let us hence turn to the Faraday-law (27.1). Often one searches in vain for this law in schoolbooks. Only in more pretentious books one makes a find under the keyword "unipolar induction". If one however compares the number of pages, which are spent on the law of induction according to Maxwell with the few pages for the unipolar induction, then one gets the impression that the latter only is a unimportant special case for low frequencies. Kupfmuller speaks of a „special form of the law of induction"<i>, and cites as practical examples the induction in a brake disc and the Hall-effect. Afterwards Kupfmiiller derives from the „special form" the „general form" of the law of induction according to Maxwell, a postulated generalization, which needs an explanation. But a reason is not given<i>. Bosse gives the same derivation, but for him the Maxwell-result is the special case and not his Faraday approach<ii>! In addition he addresses the Faraday-law as equation of transformation and points out the meaning and the special interpretation. On the other hand he derives the law from the Lorentz force, completely in the style of Kupfmuller<i> and with that again takes it part of its autonomy. Pohl looks at that different. He inversely derives the Lorentz force from the Faraday-law<iii>.

By all means, the Faraday-law, which we want to base on instead of on the Maxwell equations, shows „strange effects<i4>" from the point of view of a Maxwell representative of today and thereby but one side of the medal (eq. 27.1). Only in very few distinguished textbooks the other side of the medal (eq. 27.2) is mentioned at all. In that way most textbooks mediate a lopsided and incomplete picture<i,ii,i4>. If there should be talk about equations of transformation, then the dual formulation belongs to it, then it concerns a pair of equations, which describes the relations between the electric and the magnetic field.

If the by Bosse<ii> prompted term „equation of transformation" is justified or not at first is unimportant. That is a matter of discussion.

<i>: K. Kupfmuller: Einfuhrung in die theoretische Elektrotechnik, 12.Aufl., Springer Verlag 1988, Seite 228, Gl. 22.

<ii>: G. Bosse: Grundlagen der Elektrotechnik II, BI-Hochschultaschenbucher Nr.183, l.Aufl. 1967, Kap. 6.1 Induktion, Seite 58

<iii>: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 2 l.Aufl. Springer-Verlag 1975, Seite 77

<i4>: G. Lehner: Elektromagnetische Feldtheorie, Springer-Lehrbuch 1990, 1. Aufl., Seite 31 Kommentar zur Lorentzkraft (1.65)

The new and dual field approach consists of equations of transformation

of the electric and of the magnetic field[i]

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \quad (27.1) \quad \text{and} \quad \boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (27.2)$$

unipolar induction equation of convection

- Formulation according to the rules of duality
  - Grimsehl[ii] speaks of the ,,equation of convection", according to which moving charges produce a magnetic field and so-called convection currents (referring to Rontgen 1885, Himstedt, Rowland 1876, Eichenwald and others)
- Pohl[iii] gives examples for the equations of transformation,
- he writes the equations beneath each other

$$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}} \quad (27.3) \quad \text{and} \quad \boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}} \quad (27.4)$$

- and points out that for $v = c = 1/\sqrt{\mu \cdot \varepsilon}$ one equation changes into the other one!

The new and dual approach roots in textbook physics!

Fig. 27.8: The new and dual field approach

---

[i] see Part 1, chapter 6.5

[ii] Grimsehl: Lehrbuch der Physik, 2.Bd., 17,Aufl. Teubner Verl. 1967, S. 130.

[iii] R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 77

27.8 The field-theoretical approach

The duality between E- and H-field and the commutability asks for a corresponding dual formulation to the Faraday-law (27.1). Written down according to the rules of duality there results an equation (27.2), which occasionally is mentioned in some textbooks. While both equations in the books of Pohl[i] and of Simonyi[ii] are written down side by side having equal rights and are compared with each other, Grimsehl[iii] derives the dual regularity (27.2) with the help of the example of a thin, positively charged and rotating metal ring. He speaks of „equation of convection", according to which moving charges produce a magnetic field and so-called convection currents. Doing so he refers to workings of Rontgen 1885, Himstedt, Rowland 1876, Eichenwald and many others more, which today hardly are known.

In his textbook also Pohl gives practical examples for both equations of transformation. He points out that one equation changes into the other one, if as a relative velocity v the speed of light c should occur. This question will also occupy us.

We now have found a field-theoretical approach with the equations of transformation, which in its dual formulation is clearly distinguished from the Maxwell approach. The reassuring conclusion is added: The new field approach roots entirely in textbook physics, as are the results from the literature research. We can completely do without postulates.

Next thing to do is to test the approach strictly mathematical for freedom of contradictions. It in particular concerns the question, which known regularities can be derived under which conditions. Moreover the conditions and the scopes of the derived theories should result correctly, e.g. of what the Maxwell approximation consists and why the Maxwell equations describe only a special case.

27.9 Derivation of Maxwell's field equations

As a starting-point and as approach serve the equations of transformation of the electromagnetic field, the Faraday-law of unipolar induction and the according to the rules of duality formulated law (eq. 27.1, 2). If we apply the curl to both sides of the equations then according to known algorithms of vector analysis the curl of the cross product each time delivers the sum of four single terms. Two of these again are zero for a non-accelerated relative motion in the x-direction with v = dr/dt.

One term concerns the vector gradient (v grad)B, which can be represented as a tensor. By writing down and solving the accompanying derivative matrix giving consideration to the above determination of the v-vector, the vector gradient becomes the simple time derivation of the field vector B(r(t)) (eq. 27.10, according to the rule of eq. 27.11).

---

[i]: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 76 und 130

[ii]: K. Simonyi: Theoretische Elektrotechnik, 7.Aufl. VEB Berlin 1979, Seite 924.

[iii]: Grimsehl: Lehrbuch der Physik, 2,Bd., 17.Aufl. Teubner Verl. 1967, S. 130.

As approach serve the equations of transformation (fig. 27.5) of the electric and of the magnetic field:

$$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}} \quad (27.1) \text{ and } \quad \boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}} \quad (27.2)$$

If we apply the curl to the respective cross product:

$$\boxed{\text{rot } \mathbf{E} = \text{rot } (\mathbf{v} \times \mathbf{B})} \quad (27.5) \text{ and } \quad \boxed{\text{rot } \mathbf{H} = -\text{ rot } (\mathbf{v} \times \mathbf{D})} \quad (27.6)$$

then according to the algorithms<i> four sum terms are delivered:

$$\text{rot } \mathbf{E} = (\mathbf{B} \text{ grad})\mathbf{v} - (\mathbf{v} \text{ grad})\mathbf{B} + \mathbf{v} \text{ div } \mathbf{B} - \mathbf{B} \text{ div } \mathbf{v} \quad (27.5)$$

$$\text{rot } \mathbf{H} = -[(\mathbf{D} \text{ grad})\mathbf{v} - (\mathbf{v} \text{ grad})\mathbf{D} + \mathbf{v} \text{ div } \mathbf{D} - \mathbf{D} \text{ div } \mathbf{v}] \quad (27.6)$$

where 2 of them are zero because of: $\boxed{\mathbf{v}(t) = d\mathbf{r}/dt}$ (27.7)

- the divergence of v(t) disappears: $\text{div } v = 0$, (27.8)
- and will be zero as well: $\partial \mathbf{v}(t)/\partial \mathbf{r} = \boxed{\text{grad } \mathbf{v} = \mathbf{0}}$. (27.9)
- there remain the vector gradients:

$$\boxed{(\mathbf{v} \text{ grad}) \mathbf{B} = \frac{d\mathbf{B}}{dt}} \text{ and } \boxed{(\mathbf{v} \text{ grad}) \mathbf{D} = \frac{d\mathbf{D}}{dt}} \quad ,(27.10)$$

- according to the rules<i> in general (with eq. 27.7):

$$\frac{d\mathbf{V}(\mathbf{r}(t))}{dt} = \frac{\partial \mathbf{V}(\mathbf{r}=\mathbf{r}(t))}{\partial \mathbf{r}} \cdot \frac{d\mathbf{r}(t)}{dt} = (\mathbf{v} \text{ grad}) \mathbf{V} \quad (27.11)$$

- A comparison of the coefficients of both field equations

$$\text{rot } \mathbf{E} = -\, d\mathbf{B}/dt + \mathbf{v} \text{ div } \mathbf{B} = -\, d\mathbf{B}/dt - \mathbf{b} \quad (27.12)$$

$$\text{rot } \mathbf{H} = d\mathbf{D}/dt - \mathbf{v} \text{ div } \mathbf{D} = d\mathbf{D}/dt + \mathbf{j} \quad (27.13)$$

with the Maxwell equations results in:

- for the potential density $b = -v \text{ div } B = 0$, (27.14)
  (eq. 27.12 = law of induction, if b = 0 resp. div B = 0)

- for the current density $j = -v \text{ div } D = -v_x \cdot \rho_{el}$, (27.1a
  (eq. 27.13 = Ampere's law, if j = with v moving negative charge carriers (ρ_{el}= electric space charge density).

Fig. 27.9: ___ Derivation of Maxwell's field equations as a special case of the equations of transformation

---

<i>: Bronstein u.a.: Taschenbuch der Mathematik, 4.Neuaufl. Thun 1999, S. 652

For the last not yet explained terms at first are written down the vectors b and j as abbreviation. With equation 27.13 we in this way immediately look at the well-known law of Ampere (1st Maxwell equation). The comparison of coefficients (27.15) in addition delivers a useful explanation to the question, what is meant by the current density j: it is a space charge density ρ_{el} consisting of negative charge carriers, which moves with the velocity v for instance through a conductor (in the x-direction).

The current density j and the to that dual potential density b mathematically seen at first are nothing but alternative vectors for an abbreviated notation. While for the current density j the physical meaning already could be clarified from the comparison with the law of Ampere, the interpretation of the potential density b still is due. From the comparison with the law of induction (eq. 27.1*) we merely infer, that according to the Maxwell theory this term is assumed to be zero. But that is exactly the Maxwell approximation and the restriction with regard to the new and dual field approach, which roots in Faraday.

In that way also the duality gets lost with the argument that magnetic monopoles (div B) in contrast to electric monopoles (div D) do not exist and until today could evade every proof. It thus is overlooked that div D at first describes only eddy currents and div B only the necessary anti-vortex, the potential vortex. Spherical particles, like e.g. charge carriers presuppose both vortices: on the inside the expanding (div D) and on the outside the contracting vortex (div B), which then necessarily has to be different from zero, even if there hasn't yet been searched for the vortices dual to eddy currents, which are expressed in the neglected term.

Assuming, a monopole concerns a special form of a field vortex, then immediately gets clear, why the search for magnetic poles has to be a dead end and their failure isn't good for a counterargument: The missing electric conductivity in vacuum prevents current densities, eddy currents and the formation of magnetic monopoles. Potential densities and potential vortices however can occur. As a result can without exception only electrically charged particles be found in the vacuum (derivation in chapter 4.2 till 4.4). Because vortices are more than monopole-like structures depending on some boundary conditions, only the vortex description will be pursued further consequently. Let us record: Maxwell's field equations can directly be derived from the new dual field approach under a restrictive condition. Under this condition the two approaches are equivalent and with that also error free. Both follow the textbooks and can so to speak be the textbook opinion.

The restriction (b = 0) surely is meaningful and reasonable in all those cases in which the Maxwell theory is successful. It only has an effect in the domain of electrodynamics. Here usually a vector potential A is introduced and by means of the calculation of a complex dielectric constant a loss angle is determined. Mathematically the approach is correct and dielectric losses can be calculated. Physically however the result is extremely questionable, since as a consequence of a complex s a complex speed of light would result (according to the definition $c \equiv 1/\sqrt{\varepsilon \cdot \mu}$). With that electrodynamics offends against all specifications of the textbooks, according to which c is constant and not variable and less then ever complex.

But if the result of the derivation physically is wrong, then something with the approach is wrong, then the fields in the dielectric perhaps have an entirely other nature, then dielectric losses perhaps are vortex losses of potential vortices falling apart?

- Maxwell's field equations:

$$\begin{array}{ll} \mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt & \text{(law of induction)} \qquad (27.1^*) \\ \mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt + \mathbf{j} & \text{(Ampère's law)} \qquad (27.13) \end{array}$$

- describe the special case for b = 0 resp. div B = 0

$$\begin{array}{l} \mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt + \mathbf{v}\,\mathrm{div}\,\mathbf{B} = -\,d\mathbf{B}/dt - \mathbf{b} \qquad (27.12) \\ \mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt - \mathbf{v}\,\mathrm{div}\,\mathbf{D} = d\mathbf{D}/dt + \mathbf{j} \qquad (27.13) \end{array}$$

The physical meaning of the introduced abbreviations b and j is:

- the current density $\mathbf{j} = -\,\mathbf{v}\,\mathrm{div}\,\mathbf{D} = -\,v_x \cdot \rho_{el}$, (27.15)
- with Ohm's law $\mathbf{j} = \sigma \cdot \mathbf{E} = \mathbf{D}/\tau_1$, and (27.16)
- the potential density $\mathbf{b} = -\,\mathbf{v}\,\mathrm{div}\,\mathbf{B} = \mathbf{B}/\tau_2$, (27.17)
- with the eddy current time constant $\tau_1 = \varepsilon/\sigma$ (27.16*>
- and with the potential vortex time constant τ_2

The complete field equations (27.12 and 27.13) read, with the time constants (τ_1 and τ_2) of the respective field vortex:

- completely extended law of induction (with $B = \mu \cdot H$): (27.18)

$$\mathrm{rot}\,\mathbf{E} = -\,d\mathbf{B}/dt - \mathbf{B}/\tau_2 = -\,\mu \cdot (d\mathbf{H}/dt + \mathbf{H}/\tau_2) \qquad (27.20)$$

- and the well-known law of Ampere (with $D = \varepsilon \cdot E$): (27.19)

$$\mathrm{rot}\,\mathbf{H} = d\mathbf{D}/dt + \mathbf{D}/\tau_1 = \varepsilon \cdot (d\mathbf{E}/dt + \mathbf{E}/\tau_1) \qquad (27.21)$$

Fig. 27.10: The extension of the law of induction for vortices of the electric field (potential vortices].<i>

<i>: see also fig. 5.1

27.10 Derivation of the potential vortices

Is the introduction of a vector potential A in electrodynamics a substitute of neglecting the potential density b? Do here two ways mathematically lead to the same result? And what about the physical relevance? After classic electrodynamics being dependent on working with a complex constant of material, in what is buried an unsurmountable inner contradiction, the question is asked for the freedom of contradictions of the new approach. At this point the decision will be made, if physics has to make a decision for the more efficient approach, as it always has done when a change of paradigm had to be dealt with. The abbreviations j and b are further transformed, at first the current density in Ampere's law $j = -v_x \cdot \rho_{el}$ (27.15), as the movement of negative electric charges. By means of Ohm's law $j=\sigma \cdot E$ and the relation of material $D=\varepsilon \cdot E$ the current density j also can be written down as dielectric displacement current with the characteristic relaxation time constant $\tau_1 = \varepsilon/\sigma$ (eq. 27.16) for the eddy currents. In this representation of the law of Ampere (eq. 27.21) clearly is brought to light, why the magnetic field is a vortex field, and how the eddy currents produce heat losses depending on the specific electric conductivity σ. As one sees we, with regard to the magnetic field description, move around completely in the framework of textbook physics.

Let us now consider the dual conditions. The comparison of coefficients (eq. 27.12 + 27.17) looked at purely formal, results in a potential density b in duality to the current density j, which with the help of an appropriate time constant τ_2 founds vortices of the electric field. I call these potential vortices (in eq. 27.20).

In contrast to that the Maxwell theory requires an irrotationality of the electric field, which is expressed by taking the potential density b and the divergence B equal to zero. The time constant τ_2 thereby tends towards infinity. This Maxwell approximation leads to the circumstance that with the potential vortices of the electric field also their propagation as a scalar wave gets lost, so that the Maxwell equations describe only transverse and no longitudinal waves. At this point there can occur contradictions for instance in the case of the near-field of an antenna, where longitudinal wave parts can be detected measuring technically, and such parts already are used technologically in transponder systems e.g. as installations warning of theft in big stores.

It is denominating, how they know how to help oneself in the textbooks of high-frequency technology in the case of the near-field zone<i>. Proceeding from the Maxwell equations the missing potential vortex is postulated without further ado, by means of the specification of a ,,standing wave" in the form of a vortex at a dipole antenna. With the help of the postulate now the longitudinal wave parts are ,,calculated", like they also are being measured, but also like they wouldn't occur without the postulate as a result of the Maxwell approximation.

There isn't a way past the potential vortices and the new dual approach, because no scientist is able to afford to exclude already in the approach a possibly authoritative phenomenon, which he wants to calculate physically correct!

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3. Auflage 1986 Springer-Verlag Berlin, Seite 335

- Under the assumption: E = E(r,t); H = H(r,t) ,
- using the relations of material:
  $\boxed{\mathbf{B} = \mu\cdot\mathbf{H}}$ and $\boxed{\mathbf{D} = \varepsilon\cdot\mathbf{E}}$:
- the complete and extended law of induction reads:

$$\text{rot }\mathbf{E} = -\,\partial\mathbf{B}/\partial t - \mathbf{B}/\tau_2 = -\,\mu\cdot(\partial\mathbf{H}/\partial t + \mathbf{H}/\tau_2) \qquad (27.20)$$

- and the well-known law of Ampere:

$$\text{rot }\mathbf{H} = \partial\mathbf{D}/\partial t + \mathbf{D}/\tau_1 = \varepsilon\cdot(\partial\mathbf{E}/\partial t + \mathbf{E}/\tau_1) \qquad (27.21)$$

---

if we again apply the curl operation to eq. 27.20 and insert eq. 27.21:

$$\begin{aligned} -\,\text{rot rot }\mathbf{E} &= \mu\cdot\partial(\text{rot}\mathbf{H})/\partial t + (\mu/\tau_2)\cdot(\text{rot }\mathbf{H}) && (27.22)\\ &= \mu\cdot\varepsilon\cdot[\partial^2\mathbf{E}/\partial t^2 + (1/\tau_1)\cdot\partial\mathbf{E}/\partial t + (1/\tau_2)\cdot\partial\mathbf{E}/\partial t + \mathbf{E}/\tau_1\tau_2] && (27.23)\\ &= (1/c^2)\cdot[\partial^2\mathbf{E}/\partial t^2 + (1/\tau_1 + 1/\tau_2)\cdot\partial\mathbf{E}/\partial t + \mathbf{E}/\tau_1\tau_2] && (27.24)\end{aligned}$$

with the definition for the speed of light c:

$$\boxed{\varepsilon\cdot\mu = 1/c^2} \qquad (27.25)$$

the fundamental field equation reads:

$$\begin{aligned} \underbrace{-\,c^2\cdot\text{rot rot }\mathbf{E}}_{a} &= \underbrace{\partial^2\mathbf{E}/\partial t^2}_{b\ \text{(electromagnetic wave)}} + && (27.26)\\ &+ \underbrace{(1/\tau_1)\cdot\partial\mathbf{E}/\partial t}_{c} + \underbrace{(1/\tau_2)\cdot\partial\mathbf{E}/\partial t}_{d} + \underbrace{\mathbf{E}/\tau_1\tau_2}_{e} \\ &+ \text{ eddy current } + \text{ potential vortex } + \text{ I/U} \end{aligned}$$

Fig. 27.11: Derivation of the fundamental field equation from the equations of transformation of the electromagnetic field.<i>

<i>: The fundamental field equation mathematically describes a wave damped with the vortices of the electric and the vortices of the magnetic field. It is formulated only in space and time. From it can be deduced numerous eigenvalue equations, (i.e. the equation of Schrodinger, fig. 5.1).

27.11 Derivation of the „fundamental field equation"

The two equations of transformation and also the from that derived field equations (27.20 and 27.21) show the two sides of a medal, by mutually describing the relation between the electric and magnetic field strength (between E and H). We get on the track of the meaning of the „medal" itself, by inserting the dually formulated equations into each other. If the calculated H-field from one equation is inserted into the other equation then as a result a determining equation for the E-field remains. The same vice versa also functions to determine the H-field. Since the result formally is identical and merely the H-field vector appears at the place of the E-field vector and since it equally remains valid for the B-, the D-field and all other known field factors, the determining equation is more than only a calculation instruction. It reveals a fundamental physical principle. I call it the "fundamental field equation".

The derivation always is the same: If we again apply the curl operation to rot E (law of induction 27.20) also the other side of the equation should be subjected to the curl. If for both terms rot H is expressed by Ampere's law 27.21, then in total four terms are formed (27.26): the wave equation (a-b) with the two damping terms, on the one hand the eddy currents (a-c) and on the other hand the potential vortices (a-d) and as the fourth term the Poisson equation (a-e), which is responsible for the spatial distribution of currents and potentials<i>.

Not in a single textbook a mathematical linking of the Poisson equation with the wave equation can be found, as we here succeed in for the first time. It however is the prerequisite to be able to describe the conversion of an antenna current into electromagnetic waves near a transmitter and equally the inverse process, as it takes place at a receiver. Numerous model concepts, like they have been developed by HF- and EMC-technicians as a help, can be described mathematically correct by the physically founded field equation.

In addition further equations can be derived, for which this until now was supposed to be impossible, like for instance the Schrodinger equation (chapter 5.6-5.9). This contrary to current opinion isn't a wave equation at all, since the term (b) with the second time derivation is missing. As diffusion equation it has the task to mathematically describe field vortices and their structures.
As a consequence of the Maxwell equations in general and specifically the eddy currents not being able to form structures, every attempt has to fail, which wants to derive the Schrodinger equation from the Maxwell equations.
The fundamental field equation however contains the newly discovered potential vortices, which owing to their concentration effect (in duality to the skin effect) form spherical structures, for which reason these occur as eigenvalues of the equation. For these eigenvalue-solutions numerous practical measurements are present, which confirm their correctness and with that have probative force with regard to the correctness of the new Held approach and the fundamental field equation. By means of the pure formulation in space and time and the interchangeability of the field pointers here a physical principle is described, which fulfills all requirements, which a world equation must meet.

<i>: see also fig. 5.1

Comparison:

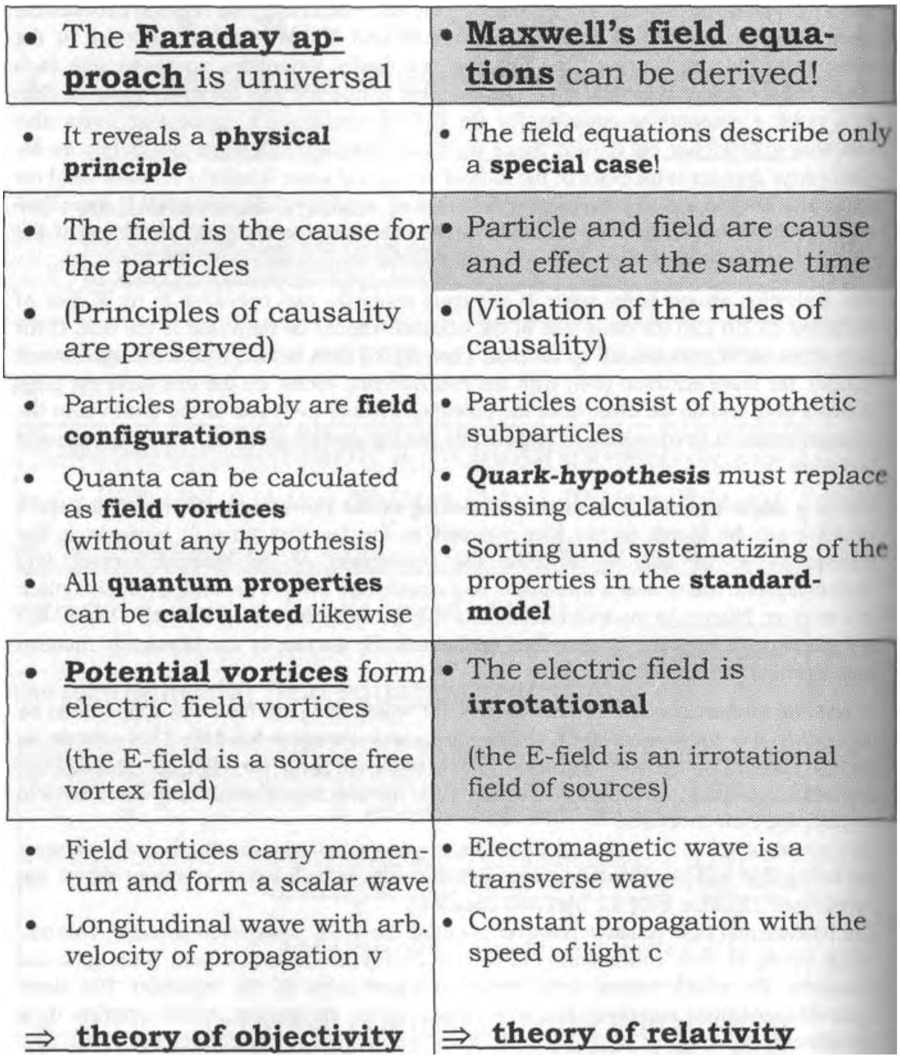

• The **Faraday approach** is universal	• **Maxwell's field equations** can be derived!
• It reveals a **physical principle**	• The field equations describe only a **special case**!
• The field is the cause for the particles • (Principles of causality are preserved)	• Particle and field are cause and effect at the same time • (Violation of the rules of causality)
• Particles probably are **field configurations** • Quanta can be calculated as **field vortices** (without any hypothesis) • All **quantum properties** can be **calculated** likewise	• Particles consist of hypothetic subparticles • **Quark-hypothesis** must replace missing calculation • Sorting und systematizing of the properties in the **standard-model**
• **Potential vortices** form electric field vortices (the E-field is a source free vortex field)	• The electric field is **irrotational** (the E-field is an irrotational field of sources)
• Field vortices carry momentum and form a scalar wave • Longitudinal wave with arb. velocity of propagation v	• Electromagnetic wave is a transverse wave • Constant propagation with the speed of light c
⇒ **theory of objectivity**	⇒ **theory of relativity**

Fig. 27.12: Comparison of the field-theoretical approaches according to Faradav and according to Maxwell.

27.12 The Maxwell field as a derived special case

As the derivations show, nobody can claim there wouldn't exist potential vortices and no propagation as a scalar wave, since only the Maxwell equations are to blame that these already have been factored out in the approach. One has to know that the field equations, and may they be as famous as they are, are nothing but a special case, which can be derived.

The field-theoretical approach however, which among others bases on the Faraday-law, is universal and can't be derived on its part. It describes a physical basic principle, the alternating of two dual experience or observation factors, their overlapping and mixing by continually mixing up cause and effect. It is a philosophic approach, free of materialistic or quantum physical concepts of any particles.

Maxwell on the other hand describes without exception the fields of charged particles, the electric field of resting and the magnetic field as a result of moving charges. The charge carriers are postulated for this purpose, so that their origin and their inner structure remain unsettled and can't be derived. The subdivision e.g. in quarks stays in the domain of a hypothesis, which can't be proven. The sorting and systematizing of the properties of particles in the standard-model is nothing more than unsatisfying comfort for the missing calculability.

With the field-theoretical approach however the elementary particles with all quantum properties can be calculated as field vortices (chap. 7). With that the field is the cause for the particles and their measurable quantization. The electric vortex field, at first source free, is itself forming its field sources in form of potential vortex structures. The formation of charge carriers in this way can be explained and proven mathematically, physically, graphically and experimentally understandable according to the model.

Where in the past the Maxwell theory has been the approach, there in the future should be proceeded from the equations of transformation of the field-theoretical approach. If now potential vortex phenomena occur, then these also should be interpreted as such in the sense of the approach and the derivation, then the introduction and postulation of new and decoupled model descriptions isn't allowed anymore, like the near-field effects of an antenna, the noise, dielectric capacitor losses, the mode of the light and a lot else more.

The at present in theoretical physics normal scam of at first putting a phenomenon to zero, to afterwards postulate it anew with the help of a more or less suitable model, leads to a breaking up of physics into apparently not connected individual disciplines and an inefficient specialisthood. There must be an end to this now! The new approach shows the way towards a unified theory, in which the different areas of physics again fuse to one area. In this lies the big chance of this approach, even if many of the specialists at first should still revolt against it.

This new and unified view of physics shall be summarized with the term "theory of objectivity". As we shall derive, it will be possible to deduce the theory of relativity as a partial aspect of it (chapter 6 and 28).

Let us first cast our eyes over the wave propagation.

- Starting-point: the fundamental field equation

$$-\,c^2\cdot\text{rot rot }\mathbf{B} \;=\; \frac{d^2\mathbf{B}}{dt^2} + \frac{1}{\tau_2}\frac{d\mathbf{B}}{dt} + \frac{1}{\tau_1}\frac{d\mathbf{B}}{dt} + \frac{\mathbf{B}}{\tau_1\tau_2} \qquad (27.26^*)$$

- with a magnetic flux density $B = B(r(t))$.

1st condition for eq. 27.26*: the special case, if $\sigma=0$ and $\sigma/\varepsilon = \boxed{(1/\tau_1) = 0}$. (27.16*)

The remaining vortex term is transformed by applying already used relations (eq. 27.10 and eq. 27.17):

$$\frac{1}{\tau_2}\cdot\frac{d\mathbf{B}}{dt} = \mathbf{v}\cdot\text{grad}\,\frac{\mathbf{B}}{\tau_2} \quad (27.10) \qquad \frac{\mathbf{B}}{\tau_2} = -\,\mathbf{v}\cdot\text{div }\mathbf{B}\;. \quad (27.17)$$

If the velocity of propagation: $v = (v_x, v_y=0, v_z=0)$; $v = dx/dt$, then the simplified field equation (if the coordinates are orientated on the vector of velocity) results in the general wave equation (involved with the x-component) in the form:

$$\boxed{\|\mathbf{v}\|^2\text{ grad div }\mathbf{B} \;-\; c^2\text{ rot rot }\mathbf{B} \;=\; d^2\mathbf{B}/dt^2} \qquad (27.27)$$

longitudinal with v = arbitrary	transverse with c = const.	wave velocity of propagation

---

2nd condition for eq. 27.28: $v = c$.

The wave equation in the usual notation (= inhomogeneous Laplace equation, = purely a special case!) now reads:

$$\Delta\mathbf{B} \;=\; \text{grad div }\mathbf{B} - \text{rot rot }\mathbf{B} \;=\; (1/c^2)\cdot d^2\mathbf{B}/dt^2 \qquad (27.28)$$

Fig. 27.13: Derivation of the wave equations (inhomogeneous Laplace equation) as a special case of the equations of transformation of the electromagnetic field.

27.13 Derivation of the wave equation

The first wave description, model for the light theory of Maxwell, was the inhomogeneous Laplace equation:

$$\boxed{\Delta \mathbf{E}\cdot c^2 = d^2\mathbf{E}/dt^2} \quad \text{with} \quad \boxed{\Delta \mathbf{E} = \text{grad div } \mathbf{E} - \text{rot rot } \mathbf{E}} \qquad (27.28^*)$$

There are asked some questions:

- Can also this mathematical wave description be derived from the new approach?
- Is it only a special case and how do the boundary conditions read?
- In this case how should it be interpreted physically?
- Are new properties present, which can lead to new technologies?

Starting-point is the fundamental field equation (27.26). We thereby should remember the interchangeability of the field pointers, that the equation doesn't change its form, if it is derived for H, for B, for D or any other field factor instead of for the E-field pointer. This time we write it down for the magnetic induction B and consider the special case, that we are located in a badly conducting medium, as is usual for the wave propagation in air. But with the electric conductivity σ also $1/\tau_1 = \sigma/\varepsilon$ tends towards zero (eq. 27.16*). With that the eddy currents and their damping and other properties disappear from the field equation, what also makes sense. There remains the potential vortex term $(1/\tau_2)$*dB/dt , which using the already introduced relations (eq. 27.10 and 27.17) involved with an in x-direction propagating wave ($v = (v_x, v_y = 0, v_z = 0)$) can be transformed directly into:

$$-\|\mathbf{v}\|^2\cdot\text{grad div } \mathbf{B}$$

The divergence of a field vector (here B) mathematically seen is a scalar, for which reason this term as part of the wave equation founds so-called „scalar waves" and that means that potential vortices, as far as they exist, will appear as a scalar wave. We at this point tacitly anticipate chapter 28, which provides the reason for the speed of light losing its vectorial nature, if it is correlated with itself. This insight however is valid in general for all velocities (v = dr/dt), so that in the same way a scalar descriptive factor can be used for the velocity (v = dx/dt) as for c.

From the simplified field equation (27.26*) the general wave equation (27.27) can be won in the shown way, divided into longitudinal and transverse wave parts, which however can propagate with different velocity.

Physically seen the vortices have particle nature as a consequence of their structure forming property. With that they carry momentum, which puts them in a position to form a longitudinal shock wave similar to a sound wave. If the propagation of the light one time takes place as a wave and another time as a particle, then this simply and solely is a consequence of the wave equation. Light quanta should be interpreted as evidence for the existence of scalar waves. Here however also occurs the restriction that light always propagates with the speed of light. It concerns the special case v = c. With that the derived wave equation (27.27) changes into the inhomogeneous Laplace equation (27.28). The electromagnetic wave in both cases is propagating with c. As a transverse wave the field vectors are standing perpendicular to the direction of propagation. The velocity of propagation therefore is decoupled and constant. Completely different is the case for the longitudinal wave. Here the propagation takes place in the direction of an oscillating field pointer, so that the phase velocity permanently is changing and merely an average group velocity can be given for the propagation. There exists no restriction for v and v = c only describes a special case.

• From the dual field-theoretical approach are derived:	From Maxwell's field equations can be derived:
=> Maxwell's field equations	=> 0
=> the wave equation (with transverse and longitudinal parts)	=> only transverse waves (no longitudinal waves)
=> scalar waves (Tesla-/neutrino radiation)	=> 0 (no scalar waves)
=> vortex and anti-vortex (current eddy and potential vortex)	=> only eddy currents
=> Schrodinger equation (basic equation of chemistry)	=>0
=> Klein-Gordon equation (basic eq. of nuclear physics)	=>0

Fig. 27.14: Comparison of the efficiency of both approaches.

(as an interim result, if it concerns the question, which approach of the two is the more efficient one and which one better should be discarded. The final balance is made in chapter 28).
It here concerns partial aspects of the following theories:

=> theory of objectivity	=> theory of relativity

27.14 The new field approach in synopsis

Proof could be furnished that an approximation is buried in Maxwell's field equations and that they merely represent the special case of a new, dually formulated and more universal approach. The mathematical derivations of the Maxwell field and the wave equation disclose, of what the Maxwell approximation consists. The anti-vortex dual to the expanding eddy current with its skin effect is neglected. This contracting anti-vortex is called potential vortex. It is capable of forming structures and propagates as a scalar wave in longitudinal manner in badly conducting media like air or vacuum.

At relativistic velocities the potential vortices are subject to the Lorentz contraction. Since for scalar waves the propagation occurs longitudinally in the direction of an oscillating field pointer, the potential vortices experience a constant oscillation of size as a result of the oscillating propagation. If one imagines the field vortex as a planar but rolled up transverse wave, then from the oscillation of size and with that of wavelength at constant swirl velocity with c follows a continual change in frequency, which is measured as a noise signal.

The noise proves to be the in the Maxwell equations neglected potential vortex term, which founds scalar waves. If at biological or technical systems, e.g. at antennas a noise signal is being measured, then that proves the existence of potential vortices, but it then also means that the scope of the Maxwell theory has been exceeded and erroneous concepts can be the result.

As an answer to the question about possible new technologies is pointed to two special properties.

1st potential vortices for reason of their particle nature carry momentum and energy. Since we are surrounded by noise vortices, an energy technical use of scalar waves would be feasible, where the noise power is withdrawn of the surroundings. There is evidence that biological systems in nature cover their need for energy in this way. But at least an energy transmission with scalar waves already would be a significant progress with regard to the alternating current technology of today.

2nd the wavelength multiplied with the frequency results in the velocity of propagation v of a wave $(\lambda \cdot f = v)$, and that for scalar waves by no means is constant. With that wavelength and frequency aren't coupled anymore; they can be modulated separately, for which reason for scalar waves a whole dimension can be modulated additionally compared to the Hertzian wave. In that the reason can be seen, why the human brain with just 10 Hz clock frequency is considerably more efficient than modern computers with more than 1 GHz clock frequency. Nature always works with the best technology, even if we haven't yet understood it.

If we would try to learn of nature and an energy technical or an information technical use of scalar waves would occur, then probably nobody wanted to have our today still highly praised technology anymore. In the course of the greenhouse gases and the electrosmog we have no other choice than to scientifically occupy us with scalar waves and their technical use.

Equations of transformation

of the electric and of the magnetic field<i>

$\boxed{\mathbf{E} = \mathbf{v} \times \mathbf{B}}$ (27.1) and $\boxed{\mathbf{H} = -\mathbf{v} \times \mathbf{D}}$ (27.2)

with: $\mathbf{B} = \mu \cdot \mathbf{H}$ (28.1) and $\mathbf{D} = \varepsilon \cdot \mathbf{E}$ (28.2)

$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}}$ (27.3) and $\boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}}$ (27.4)

Experience/observation is dependent on the relative velocity v !

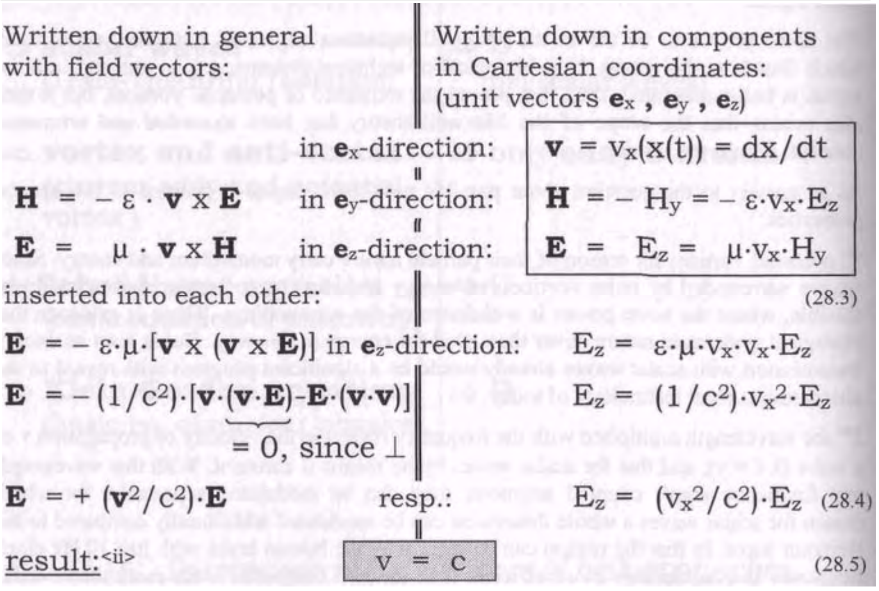

Written down in general with field vectors:

Written down in components in Cartesian coordinates: (unit vectors $\mathbf{e}_x$, $\mathbf{e}_y$, $\mathbf{e}_z$)

in $\mathbf{e}_x$-direction: $\mathbf{v} = v_x(x(t)) = dx/dt$

$\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}$ in $\mathbf{e}_y$-direction: $\mathbf{H} = -H_y = -\varepsilon \cdot v_x \cdot E_z$

$\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}$ in $\mathbf{e}_z$-direction: $\mathbf{E} = E_z = \mu \cdot v_x \cdot H_y$

inserted into each other: (28.3)

$\mathbf{E} = -\varepsilon \cdot \mu \cdot [\mathbf{v} \times (\mathbf{v} \times \mathbf{E})]$ in $\mathbf{e}_z$-direction: $E_z = \varepsilon \cdot \mu \cdot v_x \cdot v_x \cdot E_z$

$\mathbf{E} = -(1/c^2) \cdot [\mathbf{v} \cdot \underbrace{(\mathbf{v} \cdot \mathbf{E})}_{= 0,\ \text{since} \perp} - \mathbf{E} \cdot (\mathbf{v} \cdot \mathbf{v})]$ $E_z = (1/c^2) \cdot v_x^2 \cdot E_z$

$\mathbf{E} = +(\mathbf{v}^2/c^2) \cdot \mathbf{E}$ resp.: $E_z = (v_x^2/c^2) \cdot E_z$ (28.4)

result:<ii> $\boxed{v = c}$ (28.5)

Fig. 28.1: The equations of transformation,<i>

<i>: see part 1, chap. 6.5, eq. 60

<ii>: R. W. Pohl: Einfuhrung in die Physik, Bd. 2 Elektrizitatslehre, 21. Aufl. Springer-Verlag 1975, Seite 77, s.a. Kommentar auf Tafel 27.8

<iii>: chapter 28 is a repetition of chapter 6

28. Objectivity versus relativity

The new and field-theoretical approach contains the Maxwell-equations, but goes over and above these in one point. It describes potential vortices and their propagation in space as a scalar wave. With that can also a conclusive answer be given to the often-asked question for the medium and the mediated particles, which is a prerequisite for every longitudinal wave. Mediated are vortex structures with particle nature and the field itself functions as a medium. Is with that also answered the question concerning the aether?<iii>

28.1 The question concerning the aether

Do you know the Maxwell-experiment? No, you wouldn't be able to, since the intellectual father fast did make a backdown, after it didn't work out. Today one speaks of the Michelson-experiment and it may be connected with any other names (Morley, etc.). Remember: In his light theory Maxwell had determined a particular and constant value for the speed of light and for that there should be a physical reason, which should have its cause in the aether. By means of proving this aether Maxwell wanted to prove his theory, but this enterprise thoroughly went wrong.
The consideration was as follows: If the Earth is spinning and is moving through the cosmos, then one should be able to detect an aetherwind and different values for c in the different points of the compass. Maxwell found support for his project at the observatory, since with the aberration of the stars Bradley previously had described an observation, which could be considered as evidence for an aether. The director of the observatory charged his assistant Dr. Michelson with the task, to carry out a corresponding proof of an aether this time in a terrestrial experiment. But such an aether couldn't be proven, what Maxwell had to accept as a severe strike against his light theory. Seven years later Maxwell got the acknowledgement, however from a completely other corner by means of the experiments concerning the radio transmission of Heinrich Hertz.

Until today the question has remained open why astrophysics can prove the aether, whereas the detection in a terrestrial laboratory fails and it looks like there doesn't exist an aether. But as definition for the cause of c the aether can't be abolished as long as it is unsettled why the light is propagating with c of all possible velocities. The question is asked, what determines the propagation of light from today's point of view? Now, by means of outside fields the light can be slowed down. At present the world record lies at less than 65 kilometers per hour in a Bose-Einstein condensate. If electromagnetic fields determine the speed of light, if in addition field or gravitational lenses should confirm this, then the field takes over the task of the aether!

At this place the new field-theoretical approach shows its capabilities. The equations of transformation say nothing but that a moving H-field transforms to a resting E-field and vice versa, that thus in the place of a moving aether, the aetherwind, a resting aether is found. Doing so the dual field partners merely exchange their places. Therefore it is a wild-goose chase, wanting to measure an aetherwind with gauges, which underlie the same field (fig. 28.2).
Michelson had to fail.

The equations of transformation say:

$\boxed{\mathbf{E} = \mu \cdot \mathbf{v} \times \mathbf{H}}$ (27.3) and $\boxed{\mathbf{H} = -\varepsilon \cdot \mathbf{v} \times \mathbf{E}}$ (27.4)

- Experience/observation depends on the relative velocity v!
- The field takes over the function of the aether (determines c) and
- an aetherwind v x H is measured as a resting aether E and vice versa!

- for v = c the equations of transformation turn into each other and are identical $[v = v_x(x(t))]$.
- for v < c a motion field E_v depending on v is resulting

$$\boxed{E_v = E \cdot (v^2/c^2)} \quad (28.6)$$

- for v = 0 also $E_v = 0$.
- the motion field overlaps the E-field
- in the case of vortex fields the effect overlaps the cause and itself is the cause for a new effect.
- The overlap reaches to infinity, where each time is valid:

$$E_{n+1} = E_n \cdot (v^2/c^2) \quad (28.7)$$

- the field Eo overlaps the motion field E_v

$$E = E_0 + E_v = E_0 \cdot (1 + v^2/c^2) \quad (28.8)$$

- for infinite overlap:

$$E = E_0 + E_1 + E_2 + E_3 + E_4 + \ldots + E_n + E_{n+1} + \ldots \quad (28.9)$$

- results in the power series:

$$E = E_0 \cdot [1 + (v/c)^2 + (v/c)^4 + (v/c)^6 + \ldots + (v/c)^{2n} + (v/c)^{2(n+1)} + \ldots] \quad (28.10)$$

Fig. 28.2: Power series as a result of a vortex overlap.

<i>: Grimsehl: Lehrbuch der Physik, 2.Bd., 17.Aufl. Teubner Verl. 1967, S. 130.

<ii>: R.W.Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975, Seite 72 und 76, bzw. 130.

<iii>: K. Simonyi: Theoretische Elektrotechnik, 7.Aufl. VEB Berlin 1979, Seite 924.

<i4>: E. Friebe: Die Vektorprodukte der Maxwell'schen Elektrodynamik, DPG-Didaktik-Tagungsband 1995, S.396

28.2 Vortices, an overlap of the overlap

Not with any approach until now the question concerning the aether could be solved. Only the new field-theoretical approach proves with the unambiguous and free of contradiction clarification of the question concerning the aether its unmatched superiority. We hence without exception work with this approach, which is anchored tightly in textbook physics. The two equations of transformation on the one hand are the law concerning the unipolar induction according to Faraday (27.1) and on the other hand the dual formulation (27.2), which Grimsehl<i> calls convection equation. Grimsehl goes around the question for the correct sign by means of forming a modulus. Pohl draws detailed distinctions of cases and dictates the each time relevant formulation of the dual law<ii>. The sign eventually should be chosen according to the definition of the orientation of the field pointers. Also Simonyi gives both equations and the each time appropriate experiments<iii>.

If we assume the carrier of an electric field is moving with the not accelerated relative velocity v with regard to the reference system used by the observer, then a magnetic H-field is observed, which stands perpendicular both to the direction of the E-field and to the direction of v. If the motion takes place perpendicular to the area stretched by E- and H-field, then the H-field again is observed and measured as an E-field. There will occur an overlap of the fields.

In spite of that we first consider the theoretical case, that no overlap is present, and the observer as it were sees himself. The result is trivial: the relative velocity v must be the speed of light v = c . (28.5) If considered at the speed of light, the two equations of transformation turn into each other. They now are identical both mathematically and in their physical expressiveness. For this case it actually is possible, to derive the dual law straight from the Faraday law. For a wave propagating with the speed of light, to name an example, the field strength propagating along is always equal to the causing field strength, which depends on position.

If besides the evaluation of the values also the circumstance is considered that it concerns vectors, then at this place a problem as a matter of principle of the Maxwell theory gets visible, to which has been pointed occasionally, e.g. at the German Physical Society<i4>. The derivation of the speed of light from two vector equations requires, that c also has to be a vector. The question is: How the velocity vector v suddenly becomes the scalar and not pointing, in all directions of space constant factor c? Is therefore for mathematical and physical reasons "the Maxwell theory in essential parts erroneous", according to a statement of the German Patent Office<i4>?

Now, the constancy of the speed of light is a fact, which even can be derived. We at first will be content with the clue that for every observation with the speed of light, with the eyes or a gauge constructed corresponding to our perception, the vector in all its components each time is correlated to itself, by what actually the orientation of direction gets lost. Under these for c and with equal rights also for v relevant circumstances we are entitled to calculate further with the values.

An observer, who is moving with v slower than c, will besides the original E-field also observe a motion field E_v depending on the velocity v, which disappears, if v becomes zero. What he catches sight of and is able to register with gauges in the end is the overlap of both field components.

- concerning the development of the power series:

$$E = E_0 \cdot [1 + (v/c)^2 + (v/c)^4 + (v/c)^6 + \dots + (v/c)^{2n} + (v/c)^{2(n+1)} + \dots] \qquad (28.10)$$

- for $q < 1$ the power series with $q = (v/c)^2$ will converge (28.11)

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q} \qquad (28.12)$$

- for $(v/c)^2 < 1$ resp. for $v < c$ therefore is valid:

$$E = E_0 \cdot [1/1 - (v/c)^2] \quad \text{resp.} \quad \boxed{E_0 = E \cdot [1 - (v/c)^2]} \qquad (28.13)$$

The square root of Lorentz appears in squared form<i>:

$$(1 - \frac{v^2}{c^2}) = \frac{E_0}{E} \qquad (28.14)$$

The derivation for the magnetic field strength analogous to that provides the identical result<i>:

$$(1 - \frac{v^2}{c^2}) = \frac{H_0}{H} \qquad (28.15)$$

Fig. 28.3 The field dilatation depending on velocity

<i>: see part 1, chap. 6.6

<ii>: Prof. Dr. H.-J. Runckel, Abteilung fur Mathematik IV der Universitat Ulm

28.3 Field overlap

But it doesn't abide by this one overlap. In the case of vortex fields the effect overlaps the cause and itself becomes the cause for a new effect. The overlapped cause produces a further effect, which for its part is overlapping (see chap. 3).

Vortices thus arise, if overlaps for their part are overlapping and that theoretically reaches to infinity, to which I already repeatedly have pointed (fig. 3.0). In addition do vortices represent a fundamental physical principle. The Greek philosopher Demokrit has traced back the whole nature to vortex formation and that already 2500 years ago!

In the field-theoretical approach this interpretation seems to experience a mathematical confirmation, since also the fields are overlapping in vortex structures. According to that we owe our observations and our being the relative movements and the vortex formation. If reversed there wouldn't be any movement, then there also would not exist fields, light nor matter. If we observe the sky, then everything visible follows the movement of its own of the Earth, of the solar system and the whole galaxy, which is on its way with unknown galactic velocity, and all movements take place in vortex structures (fig. 10.2).

The field overlap dictated by the Faraday-approach as well reaches to infinity, what has stimulated my colleagues of mathematics to also mathematically put into practice this physical requirement[ii]. This leads to an infinite power series, which converges under the condition that $v < c$.

As a result of the power series development the well-known square root $\sqrt{1-(v^2/c^2)}$ of Lorentz occurs in squared form (see also fig. 6.6). It determines the relation of the observed and the causing field strength of the electric or the magnetic field. Physically the found relation describes a dilatation field depending on velocity. The field strength thus increases, if the relative velocity v increases, or inversely no difference is observable anymore, if v tends towards zero.

Whoever wants to compete with Albert Einstein (1879-1955), who has developed the theory of relativity from the length contraction, which depends on velocity, could be inclined to derive a new field physics from the field dilatation. But I must warn of such a step. The derivation of the length contraction by the mathematician Hendrik Lorentz (Lorentz contraction) assumes a number of limiting conditions. The relative velocity v for instance may not experience any acceleration. Actually however almost all motion takes place as circular vortex motion, so that due to the occurring centripetal acceleration the conditions for the theory of relativity aren't fulfilled anymore. Neglects or generalizations thereby can lead to considerable errors, of which I would like to warn. It in general is a delicate enterprise, if one wants to provide a physical interpretation for a purely mathematically won result.

This warning to the same extent also is valid for the here shown derivation of the field dilatation. The limiting conditions practically are the same as for Einstein and the problems with a provided physical interpretation won't be less. Also here lots of paradoxes will occur, which are nothing but errors of the theory. So we won't reach our destination.

There now only one further mathematical step is necessary, which links the theory of relativity with the new notion of a field dilatation depending on velocity.

Example:
Measurement of length by means of a measurement of propagation time (sound or light) with c = L/t in a vehicle moving with v.

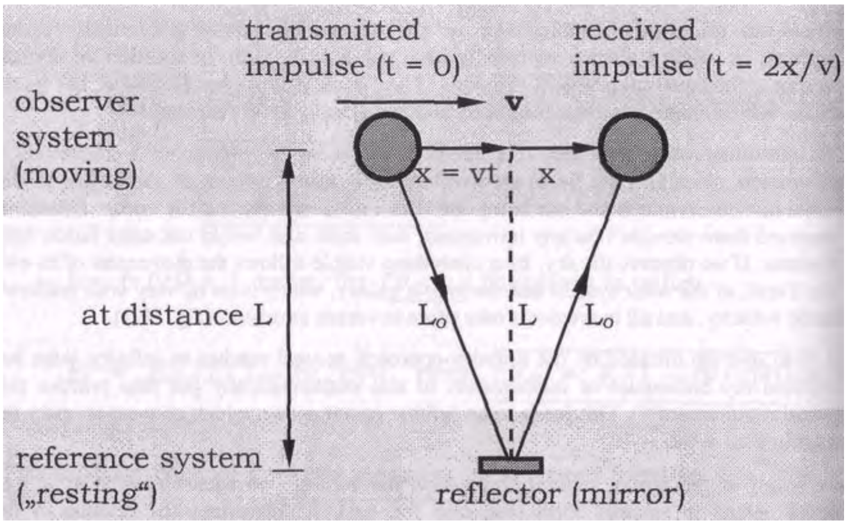
transmitted
impulse (t = 0)
received
impulse (t = 2x/v)
observer
system
(moving)
v
x = vt
x
at distance L
L_o
L
L_o
reference system
(„resting")
reflector (mirror)

From driving time t:

$t = 2\frac{x}{v}$ = signal propagation time: $t = 2\frac{L_o}{c}$ follows $\boxed{\frac{x}{v} = \frac{L_o}{c}}$

According to Pythagoras: $L_o^2 = L^2 + x^2 = L^2 + L_o^2\cdot(v^2/c^2)$

$$L_o^2\cdot(1 - v^2/c^2) = L^2$$

the shortening of the rule results in:

$$\boxed{(1 - \frac{v^2}{c^2}) = (\frac{L}{L_o})^2} \quad (28.16)$$

Fig. 28,4: ______Derivation of the length contraction

Examples: contraction according to Lorentz transformation, measurable length shortening, curvature of space.

Counterexample: unsuccessful aether detection (Michelson experiment)

28.4 The derivation of the length contraction

The Lorentz transformation is the result of a purely mathematical problem. Stimulated by the surprisingly result of the Michelson experiment the Dutchman Hendrik A. Lorentz 1891 asked himself, how the equations of the Galilei-transformation would have to look like, if the propagation of light wouldn't be infinitely fast but finite and constant. He thereby proceeds from the assumption of two inertial systems moving against one another with a not accelerated velocity v, in which the laws of Newtonian physics are equally valid<i>. As a result of the relative motion a change of the length measures will occur.

This at first can be explained as a purely geometric effect in the context of nonrelativistic physics. We imagine a vehicle, which is on its way with constant velocity, and emits an optical or acoustical signal. Sideways in the countryside is standing in a perpendicularly measured distance L a reflector (mirror), which sends the signal back again. The velocity of the signal however isn't infinitely fast and from that follows that the vehicle during the propagation time of the signal as well has moved a bit further. The actual way, which the signal had to cover now amounts to L_o (> L). The distance measure thus is observed smaller as it is in reality, to be specific for the factor of the square root of Lorentz (fig. 28.4).

$$L = L_o \cdot \sqrt{1-(v^2/c^2)} \quad (28.16)$$

According to the principle of relativity it doesn't play a role, if the vehicle is driving or if it is standing still and the mirror is moving with a linear uniform velocity.

Initially Einstein also only spoke of an observable length contraction, which must not necessarily occur in reality, an optical deception so to speak. Lorentz however proceeded from the assumption of a physical length change, thus a length change existing in reality, what in practice at first makes no difference. If e.g. at relativistic velocities a rocket becomes smaller, then the pilot equally shrinks, so that it would not be possible to notice a present difference.

If however the observer stands outside the events and takes a „neutral standpoint", then he will be able to see, which interpretation is the right one. Today some examples are known. In accelerators particles at relativistic fast velocities actually get smaller for the factor of the square root of Lorentz. That has been proven and this result afterwards gives the Dutchman Lorentz right! The followers of the physical length contraction also are called Neo-Lorentzians.

In the vicinity of a gravitational mass the speed of light becomes so slow, that the shortening factor plays a role and space is curved towards the mass. To understand this shortening of scale, the influence of the field also should be considered.

<i>: Example: In a closed lift physical experiments are being carried out. Accelerations of the lift have an influence on the experiments. However no influence can be detected, if the lift is standing still or is moving with constant velocity. It with that fulfills the conditions of an inertial system. The question is: what do the experiments show someone standing outside, whom the lift passes by?

From the comparison of the Lorentz contraction (28.16)
with the field dilatation (28.14 and 28.15)

$$\left(1-\frac{v^2}{c^2}\right)=\left(\frac{L}{L_o}\right)^2 \overset{!}{=} \left(1-\frac{v^2}{c^2}\right)=\frac{E_o}{E}=\frac{H_o}{H}$$

follows

$$\left(1-\frac{v^2}{c^2}\right)=\boxed{\left(\frac{L}{L_o}\right)^2=\frac{E_o}{E}=\frac{H_o}{H}} \quad (28.17)$$

the proportionality (length measures depending on field):

$$\underline{E,\ H \sim 1/L^2} \qquad \text{and} \qquad Eo,\ Ho \sim 1/L_0^2 \qquad (28.18)$$

Experimental examples<i>:

- Electrostriction (piezo speaker)
- Magnetostriction
- Field or gravitational lenses
- Curvature of space, deflection of light

Conclusion<ii>:

- The field determines the length measures (what is 1 meter)
- The field determines the velocities v (in m/s)
- The field determines the speed of light c [m/s]
- Measurement of the speed of light is made with itself:

$$\boxed{c \sim r} \quad (28.19)$$

- Measured is a constant of measurement c = 300.000 km/s
- The speed of light c is not a constant of nature!

Fig. 28.5: The dependence of the Lorentz contraction on the field

---

<i>: see part 1, chap. 6.10
<ii>: see part 1, chap. 6.11

28.5 The dependence of the Lorentz contraction on the field

The two results of the field dilatation (28.14 and 28.15) and of the Lorentz contraction (28.16) must be brought together and compared (28.17). Doing so the mathematical expression of the square root of Lorentz is cancelled out. That is of utmost importance, since with that also all limits disappear and there remains a purely physical relation, a proportionality of utmost importance (28.18).

What was the sense of the limits associated with the introduction of so-called inertial systems, which are the basis of the Lorentz transformation and which were adopted for our derivation of the field dilatation? They now only are auxiliary considerations according to model. We have chosen a very simple model, which can be described mathematically, in which an observer holds in his hand gauges for distances and field strengths and with that gauges a system flying by with constant velocity. He on the one hand determines a length contraction and on the other hand a field dilatation. He compares both with each other and comes to the conclusion: The field determines the dimensions!

This statement is purely physical and it is generally valid. It is independent of the relative velocity and all other mathematical conditions. A centrally accelerated circular motion e.g. will falsify the length contraction to the same extent, as the at the same time occurring field dilatation. It can be expected, that in addition to the square root of Lorentz also other errors will mutually efface, so that a generalization in this case actually seems to be allowed.

The won proportionality is of most elementary importance. We use it in the case of the piezo speaker and know it from the curvature of space and deflection of light in presence of extreme fields. If we ourselves however are exposed to the field as an observer, in which also the object to be observed is situated, then we are in the dilemma, not being able to perceive the influence. If we, to stay with the example, would sit in a rocket and this would become smaller at faster velocity, then we would notice nothing, since we also would shrink along to the same extent.

That concerns every measurement of velocity in general and the speed of light c in particular, which as is well-known is measured in meters per second. But if the field determines c and in the same way the length measure, which is given in meters, then both stand in a direct proportionality to each other, then we won't have the slightest chance to measure the speed of light. If namely c is changed, then this concerns the measurement path in the same way. Now the variable is measured with itself and as a result appears c, a constant value. We neither can see the change, since our eyes scan all objects optically and that means with c.

It is the nightmare of each and every measurement engineer, if the gauge depends on the factor to be measured. No wonder, if the theorem of addition of the velocities apparently loses its validity and always the same c is being measured, independent of the direction in which the source of radiation is moving (chap. 6.11). The result is:

The speed of light is a constant of measurement and not a constant of nature! If however the light is scanned with the speed of light, then also all components of the light vector correlated with themselves result in the same constant value c, then actually the vector of the speed of light loses its orientation in space and becomes a scalar factor. The Maxwell equations already anticipate this circumstance, but without providing an explanation why this is correct. Only the new field approach can answer the open question. With the derivation an axiom of physics - one also can say stumbling block - has been overcome.

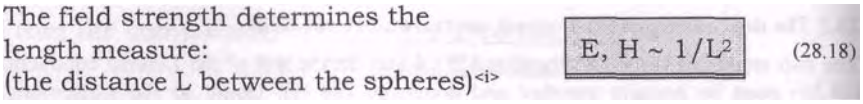

The field strength determines the length measure:
(the distance L between the spheres)<i>

$$\boxed{E, H \sim 1/L^2} \tag{28.18}$$

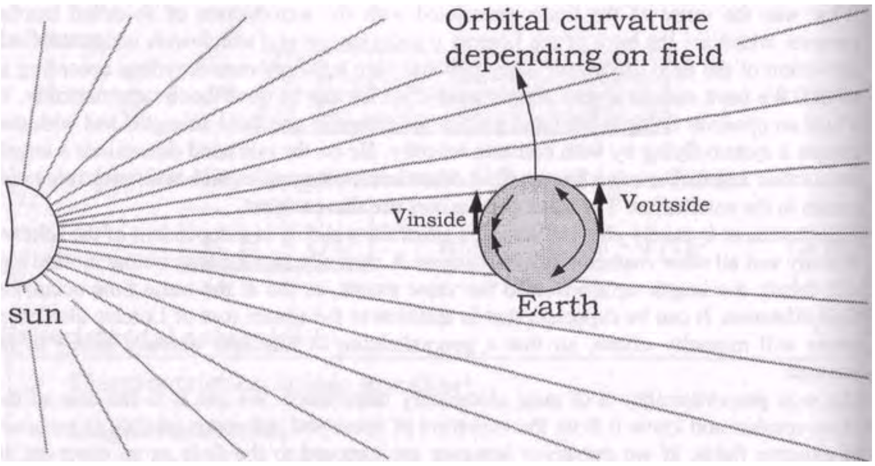
Orbital curvature depending on field
v_{inside}
$v_{outside}$
sun
Earth

Fig. 28.6 A: The curvature of the Earth in the gravitational field of the sun

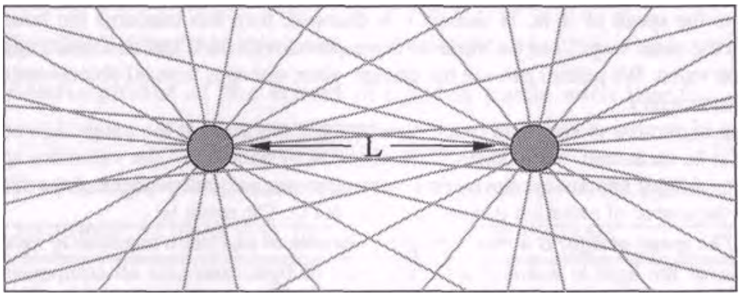

Fig. 28.6 B: Force of attraction and reduction of the distance L as the mutual field influence of two masses.

<i>: Repetition of part 1, chapter 6.7

28.6 Boscovich and the respiration of the Earth

Who has got a good idea, fast will find that some other scientist had the same idea already before, and for his part possibly already had to make the same experience. If in spite of that the track fast loses in history, it as a rule is because of the insufficient citing and the vanity of the discoverers, who in reality only are rediscoverers of a much older knowledge.

The dependence of the length measures on the field (eq. 28.18) in the mathematical form of the derivation however should still be quite new. But the physical consequence already was described by the Jesuit priest Roger Joseph Boscovich from Dalmatia in 1755[i]. He was Professor for mathematics and theology in Rome and spoke about the world on the one hand being deformable, similar to rubber, but on the other hand we aren't able to perceive this, since our body is made of the same material and follows all changes. „We in an absolute way can recognize neither the place, nor distances or orders of magnitude", writes Boscovich in his book about space and time and writes how these are perceived by us[ii]. He suspects that the Earth unobservable for man „is respiring".

Actually a terrestrial observer at daytime is situated closer to the sun than in the night. He by day is exposed to a slightly stronger field and as a result correspondingly smaller. He himself and all objects in his neighbourhood are subject to an identical fluctuation of size, so that this „respiration" of the Earth cannot be detected. It can be detected neither with a tape measure nor with an optical measurement and still is present in reality. Only from a neutral standpoint we can succeed to recognize the actually existing curvature of space (fig. 28.6 A).
An example is the duration of the sunshine at the equator, which is longer than can be expected from the spherical form of the Earth. This reveals, how the Earth is bending towards the sun (see also chapter 6.7).
A further example is the influence of the field on the orbital velocity of the Earth measured in meters per second. Here also the meter at daytime is smaller than in the night, for which reason the Earth is moving slower on the side turned towards the sun, like a track vehicle, which drives a turn, if the chain at the inside runs slower than on the outside. If the Earth describes an orbit around the sun. then this circumstance has to do nothing at all with centrifugal force or with a force of attraction of the sun. The circular motion simply and solely is a result of the field influence of the sun[iii].

The force idea proves to be a pure auxiliary description. In the context of Newtonian mechanics the force plays a central role. Without question it is a very efficient and subjective perceptible description, which still isn't able to reproduce the physical reality in an objective manner. What keeps the planets into their orbit is only the field of the sun, which we call gravitational field and not some force! But of which kind is the gravitation and the field, which causes masses to come closer together and following our subjective observation attract each other (fig. 28.6 B)?

<i>: O. E. Rossler: Endophysics, the World as an Interface, World Scientific Publishing Co. 1998, Kap. 10, S. 87-112, mit Ubersetzungen aus <ii>:

<ii>: R. J. Boscovich: De spatio et tempore, ut a nobis cognoscuntur, 1755.

The length measure (the distance L between the spheres)<i> is determined by the field strength:

$$\boxed{E, H \sim 1/L^2} \quad (28.18)$$

A. Charged mass points (electrons, positrons, ions,...):

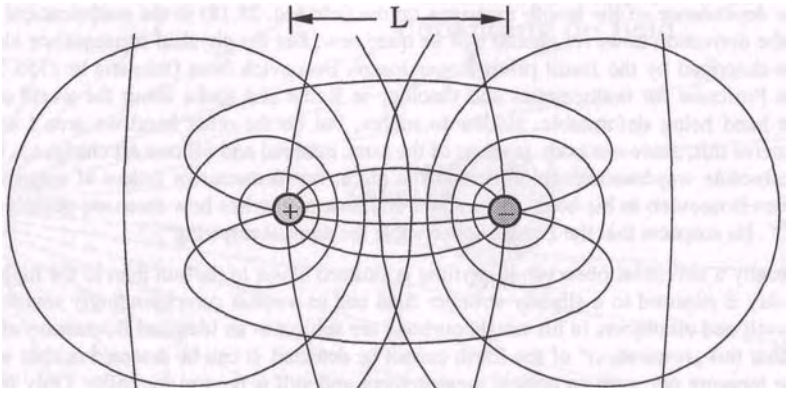

B. Uncharged mass points (neutrons, atoms, ...):

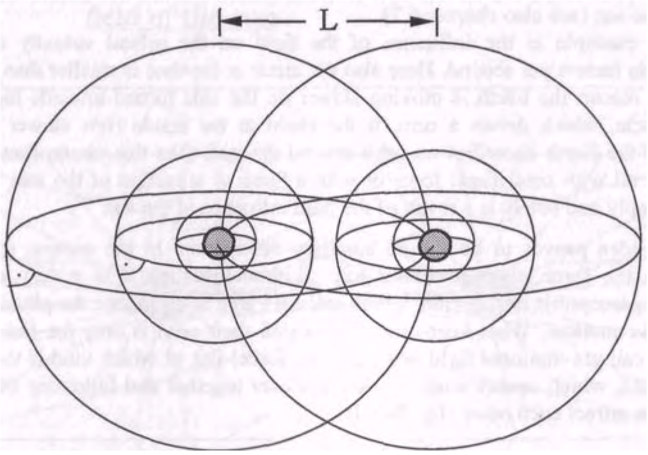

Fig. 28.7: Observation of a mutual force of attraction because of the effect of the fields on the distance measure.

<i>: Repetition of part 1, chapters 6.7 - 6.9 and part 2, chapters 10.4 and 15.2

28.7 Derivation of the gravitation

In fig. 28.6B the relation between the field influence and the observed force of attraction of two mass bodies is represented. If I in my mind for instance "switch on" the field lines of both bodies, which are placed at distance L, then the fields according to equation 28.18 reduce the measure L and optically come each other closer. With increasing proximity the field line density increases, so that L further decreases. We observe a mutual attraction, which lets both bodies move towards each other.

In fig. 28.7A the two bodies carry an electric charge. For different polarity the field lines as is well-known run from the positive pole to the negative pole, to bundle up there. As a matter of principle also here an attraction can be expected, which is called electromagnetic interaction. For the reason of the bundling up of the field lines this effect however will turn out to be considerably stronger. Hence the electromagnetic interaction is for many powers of ten more powerful than the gravitation.

Furthermore there also can occur repulsion, if in the case of like charge the field lines are bent aside and between the two bodies an area is formed, where the field tends towards zero and the distance measure L (according to eq. 28.18) as a result towards infinity. The electromagnetic interaction theoretically indeed reaches to infinity. Responsible are the open field lines arising from a charged body.

Now every charged body in addition has a mass, with which it takes part in the gravitation. Let's remember the comparison of the derivations. The Maxwell theory teaches us that in the static case E- and H-field are decoupled, because each time the other field disappears. Even if as a result of the unipolar induction for every open field line the other one is taken to be standing perpendicular to the open field line, then this other line just wraps around the open field line and forms a closed-loop field line. In that way it can't be influenced anymore from the outside and can be neglected, so goes the doctrine, which is drawn from the Maxwell theory (fig. 27.5).

This is a fatal error in reasoning! The equation 28.19 naturally is valid for open field lines in the same manner as for closed ones. These fields also lead to an observable force of attraction. If of course exactly those fields are neglected, which are responsible for the gravitation, then we need not wonder, if we don't understand the gravitation and the nature of this interaction!

The influence of the closed field lines responsible for the gravitation is due to the missing bundling up of the lines correspondingly weak. Secondly these can't exist a force of repulsion due to the missing ability to influence closed field lines from the outside and third it can be recorded that all charged bodies also have a mass. All three statements of the field lines model perfectly cover the physical reality.

Fig. 28.7 B shows uncharged bodies, for which both the field lines of the E-field and of the perpendicular to them arranged H-field are closed-loop. Such bodies, like e.g. neutrons or whole atoms without charge behave electrically neutral to the outside, but have a mass for the reason of the closed field lines, whereby the field lines of the H-field dominate those of the E-field[i].

---

[i] Repetition of K. Meyl: EMEC part 1, chap. 6.8 + 6.9, part 2 chap. 15.1

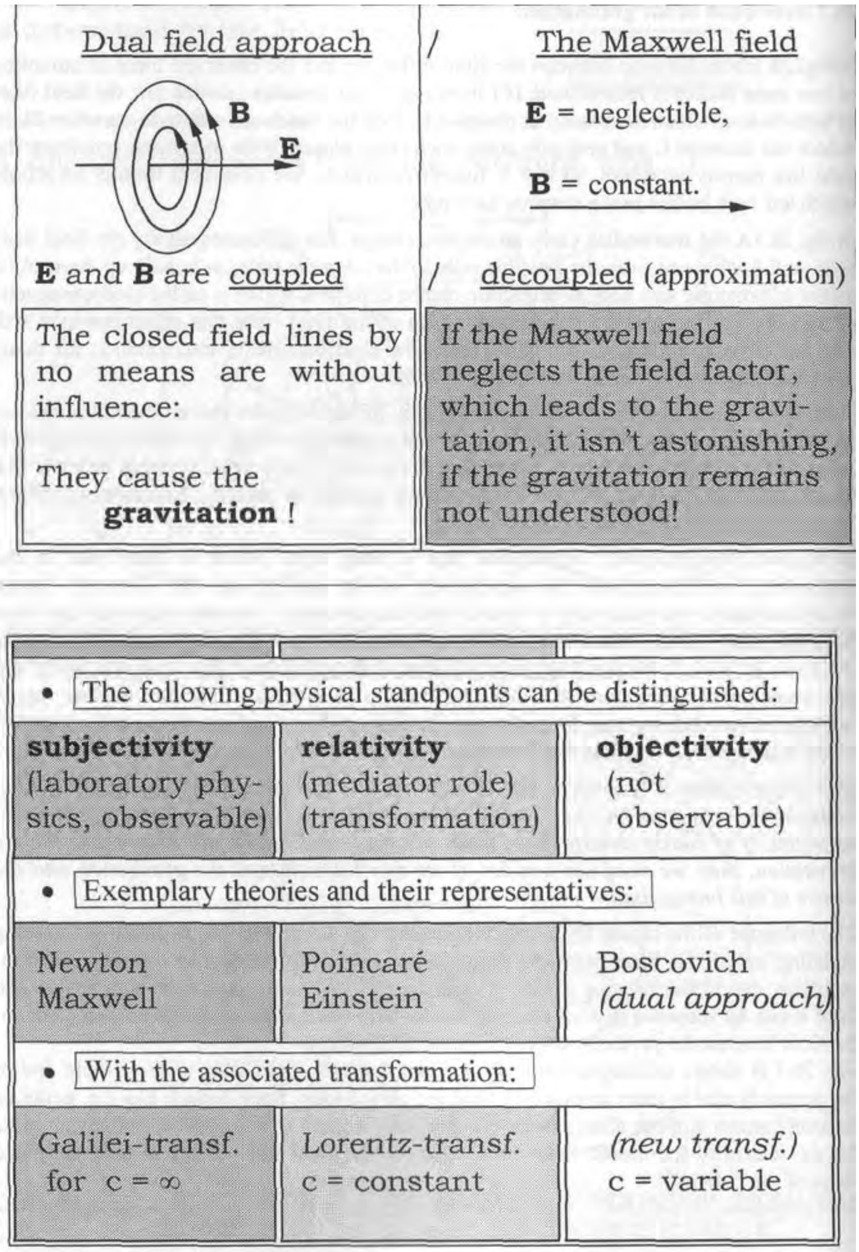

Dual field approach /	The Maxwell field
B **E**	**E** = neglectible, **B** = constant.
E and **B** are coupled /	decoupled (approximation)
The closed field lines by no means are without influence: They cause the **gravitation** !	If the Maxwell field neglects the field factor, which leads to the gravitation, it isn't astonishing, if the gravitation remains not understood!

- The following physical standpoints can be distinguished:

subjectivity (laboratory physics, observable)	**relativity** (mediator role) (transformation)	**objectivity** (not observable)

- Exemplary theories and their representatives:

Newton Maxwell	Poincaré Einstein	Boscovich *(dual approach)*

- With the associated transformation:

Galilei-transf. for $c = \infty$	Lorentz-transf. c = constant	*(new transf.)* c = variable

Fig. 28.8: Physical standpoints

28.8 From subjectivity to objectivity

With the field lines interpretation, which by the way already preferably was used by Faraday, the gravitation proves to be an until now neglected influence of the electromagnetic field. With that for the first time also the grand unification of the interactions was successful. The long sought-for unified theory with that for the first time comes within reach.
The derivation has made it possible to mathematically secure the theoretical working model of Boscovich. Already 1755 Boscovich points out the optical deception, which our observation underlies, if absolute orders of magnitude in our neighbourhood should change and our perception would change along. Then also all metric and optical measurement results would underlie this change. Following the idea of Boscovich I distinguish between subjectivity and objectivity.

The relativity is a compromise lying between both points of view, where a neutral standpoint is strived for, which lies outside the events. And from this standpoint the objectively taking place events are being observed. The theory of relativity consequently is a pure observer theory with strongly restricted scope on the basis of the Lorentz-transformation.

Theories of classic physics, like e.g. Newtonian mechanics, fall in the domain of subjectivity. The results and regularities are won in a terrestrial laboratory if possible isolated from the environment, where they have absolute validity. Here the Galilei-transformation is valid.
But if these subjectively won laws are applied to the microcosm in quantum physics or to the calculation of cosmic observations, one fast hits limits. The better the resolution of the microscopes and telescopes gets, the clearer the „outside" observer realizes, how much the laws of classic physics lose their validity.

Astrophysics successfully reaches for the theory of relativity, which with the curvature of space in the vicinity of mass centres delivers useful explanations. Here the dependence of the spatial dimensions on the field already could be established. In contradiction to that this fundamental relation is said to play no role whatsoever in quantum physics, or in all terrestrial laboratory experiments. But with which right may physical regularities from one domain be ignored in others? There only can exist one physics and that should be sought for!

What we need is objectivity! Behind all the apparently disconnected phenomena of physics work quite simple laws, which can't be observed and are until now not recognized by us. Objective physics in the words of Goethe is the one, which holds the world together in the heart of hearts. I call this, already by Boscovich suggested point of view, theory of objectivity. The access to the model domain of objectivity must be made mathematically by means of a transformation, since it is blocked for us by means of measurements or observations (see chapters 6.15-6.19). The transformation back into the observation domain must be made according to the same mathematical relations (fig. 28.9). In this way the quantum properties of the elementary particles can be calculated with high accuracy and agreement with the values, which until now only could be measured (chapter 7).

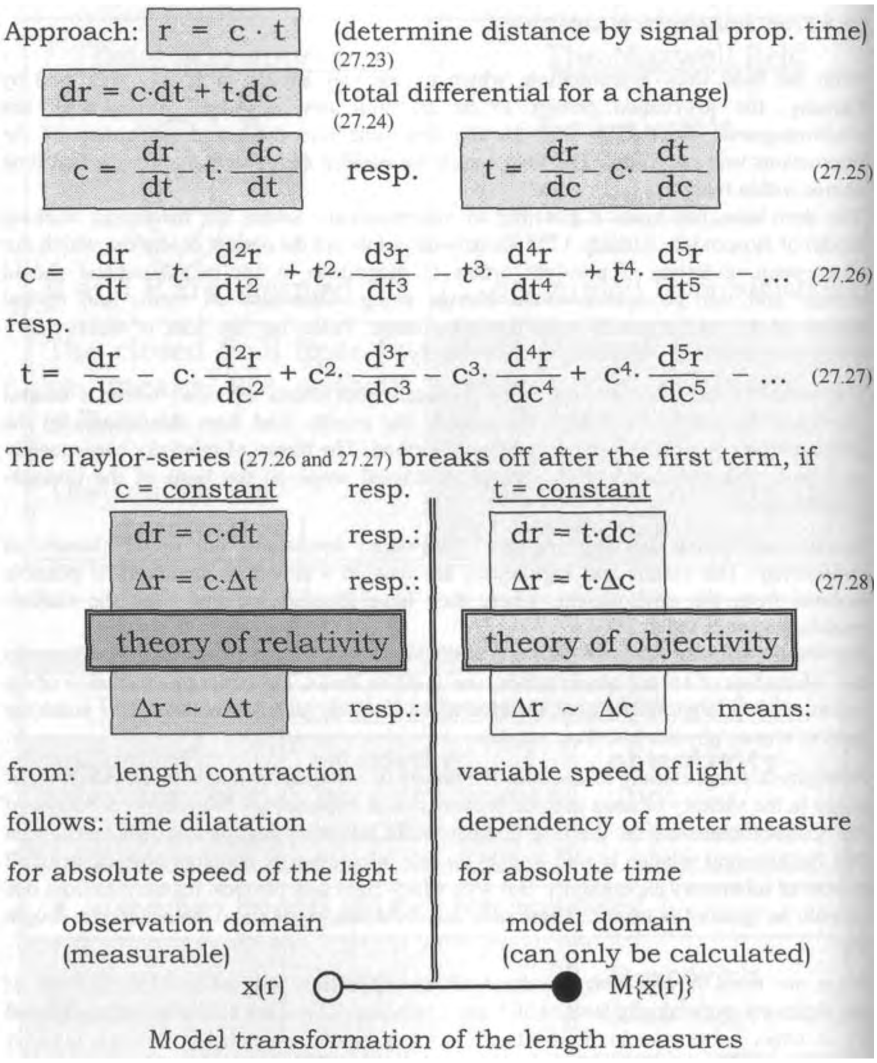

Approach: $\boxed{r = c \cdot t}$ (determine distance by signal prop. time)
(27.23)

$\boxed{dr = c\cdot dt + t\cdot dc}$ (total differential for a change)
(27.24)

$$\boxed{c = \frac{dr}{dt} - t\cdot\frac{dc}{dt}} \quad \text{resp.} \quad \boxed{t = \frac{dr}{dc} - c\cdot\frac{dt}{dc}} \qquad (27.25)$$

$$c = \frac{dr}{dt} - t\cdot\frac{d^2r}{dt^2} + t^2\cdot\frac{d^3r}{dt^3} - t^3\cdot\frac{d^4r}{dt^4} + t^4\cdot\frac{d^5r}{dt^5} - \dots \qquad (27.26)$$

resp.

$$t = \frac{dr}{dc} - c\cdot\frac{d^2r}{dc^2} + c^2\cdot\frac{d^3r}{dc^3} - c^3\cdot\frac{d^4r}{dc^4} + c^4\cdot\frac{d^5r}{dc^5} - \dots \qquad (27.27)$$

The Taylor-series (27.26 and 27.27) breaks off after the first term, if

$\underline{c = \text{constant}}$ resp. $\underline{t = \text{constant}}$

$\boxed{dr = c\cdot dt}$ resp.: $\boxed{dr = t\cdot dc}$

$\boxed{\Delta r = c\cdot\Delta t}$ resp.: $\boxed{\Delta r = t\cdot\Delta c}$ (27.28)

theory of relativity | **theory of objectivity**

$\boxed{\Delta r \sim \Delta t}$ resp.: $\boxed{\Delta r \sim \Delta c}$ means:

from: length contraction | variable speed of light

follows: time dilatation | dependency of meter measure

for absolute speed of the light | for absolute time

observation domain (measurable) | model domain (can only be calculated)

Model transformation of the length measures

Fig. 28.9: Theory of relativity and theory of objectivity and the model transformation between both physical standpoints

---

<i>: Repetition of part 1, fig. 6.16

28.9 The objective standpoint

The question is asked how one gets to an objective physical standpoint, which in addition evades every observation? The way leads over a transformation, to which all perceptible and measurable relations must be submitted.
If we for instance measure the distance r to a point light source, then the propagation of the light c and the propagation time t determine the distance measure $r = c*t$
If there occurs a little change of the distance, then two causes should be considered: Either the propagation time or the speed of light have changed somewhat. With that the two possible standpoints already would have been found.

The relativistic standpoint, which proceeds from the assumption of the speed of light being constant, says: the propagation time varies and we are dealing with a clock problem. If namely for relativistic velocities a length contraction occurs, then from that necessarily follows a time dilatation.
But actually no specific statement can be made about the constancy of the speed of light, besides that we look at, measure and scan everything with c and hence only observe the constancy. With that the theory of relativity remains a pure observer theory, exactly as Einstein originally called it into existence. This standpoint follows the motto: What can't be observed also doesn't need to interest the physicist.

The objective standpoint strives for more, for a description of the actually taking place processes. This time we proceed from the assumption of a universal and constant time with the argument: The time measure is an immutable definition and the physicist, who dictates this, himself determines what is simultaneousness. Then there also is no place for time travel and for clocks going wrong.
Therefore the speed of light can take all possible values always in strict proportionality to the length measures. Thus the measured length and distance measures should be transformed and that in the end is the unit ,,meter", which should be replaced by an objective measure.

With that the necessary transformation for variable c would be outlined. This transformation will be enqueued in the file of the big transformations. From it the Lorentz-transformation for c = constant emerges as a special case, like already from that transformation the Galilei-transformation follows for $c = \infty$. How now the relation of the subjective to the objective ,,meter" should be determined; by means of the relation of the relevant fields (eq. 28.17) or by means of the square root of Lorentz (eq. 28.16), over that should be worked and spoken. We already have successfully gone through it in a concrete example (chapter 7).

Every theory is judged according to its expressiveness. Ending this chapter the statements and derivations hence again are compared. On the one hand the Maxwell theory and from that the theory of relativity can be derived from the new approach, on the other hand a long list follows, which can't be connected with the Maxwell equations, like e.g. the gravitation. For instance the neutrino and all other elementary particles with all their specific quantum properties are derived (chapter 7), free and easy fundamental laws result, like the law of conservation of energy, and even the temperature spills its until now kept secret (chapter 8.3). Remains the conclusion: With no other approach according to the textbooks until now the efficiency of the new approach could be obtained.

• From the dual field-theoretical approach are derived:	• From Maxwell's field equations can be derived:
=> Maxwell's field equations	=>0
=> Quantum properties of the elementary particles	=>0
=> Neutrino (as an oscillating ring-like vortex)	=>0
=> Gravitation (as a result of closed field lines)	= > 0
=> Unified theory (grand unification of all interactions)	=>0
=> Temperature (as an oscillation of size depending on field)	=>0
=> Law of conservation of energy (and many other fundamental laws of physics)	=>0
=> Theory of objectivity	=> Theory of relativity

Fig. 28.10: Comparison of the efficiency of the approaches (final balance)

29. Mathematical gleanings

If, proceeding from the new field-physical approach, well-known and accepted theories are derived as special cases, this on the one hand can be valued as evidence for the correctness of the approach. On the other hand the new approach in part significantly influences the interpretation of the derived theories. That can involve a rethinking, with which not insightful people have difficulties, if for instance quantum physics, thermodynamics or the gravitation become partial aspects of electromagnetism. Over and above that are hidden many new thing in the new approach, which are there to discover. To that are counting among others the potential vortices and the scalar waves. One can work out these phenomena physically or mathematically, where the latter way as a rule is the faster one. Hence the summary shall be concluded with a kind of mathematical gleanings.

29.1 General and special theory of relativity

Albert Einstein distinguishes between general and special theory of relativity. Whereas the special (SRT), still is linked tightly with the prerequisites of the Lorentz-transformation, the general (GRT), deals with an extension to arbitrary systems, which mustn't be inertial systems. I would like not to dwell upon the GRT, as Einstein designed it, and merely notice that every generalization represents a possible source of errors and has to be well founded.

In the case of our derivation, the general case as it were resulted of its own accord. Let's turn back: If the root of Lorentz still was a component of the derived field dilatation (28.15) and equally of the length contraction (28.16), then it fell out in the comparison of both results (28.17). With that the important result, the proportionality (28.18), which among others results in the gravitation, becomes independent of the speed of light and the relative velocity v. This last step is obvious and still completely new. It cannot be found at Einstein, who in another way finds his GRT and his description of the gravitation. Even if here is striven for the same goal, then deviations in the result cannot be excluded because of the differences in the derivation, for which reason I additionally mark the by me derived general relativity (GRT'), to avoid confusion.

Influence of the Lorentz-transformation in the:
SRT (special theory of relativity): one-dimensional,
GRT' (general theory of relativity): three-dimensional, to a large extent corresponding to the GRT of Albert Einstein,
GOT (general theory of objectivity)

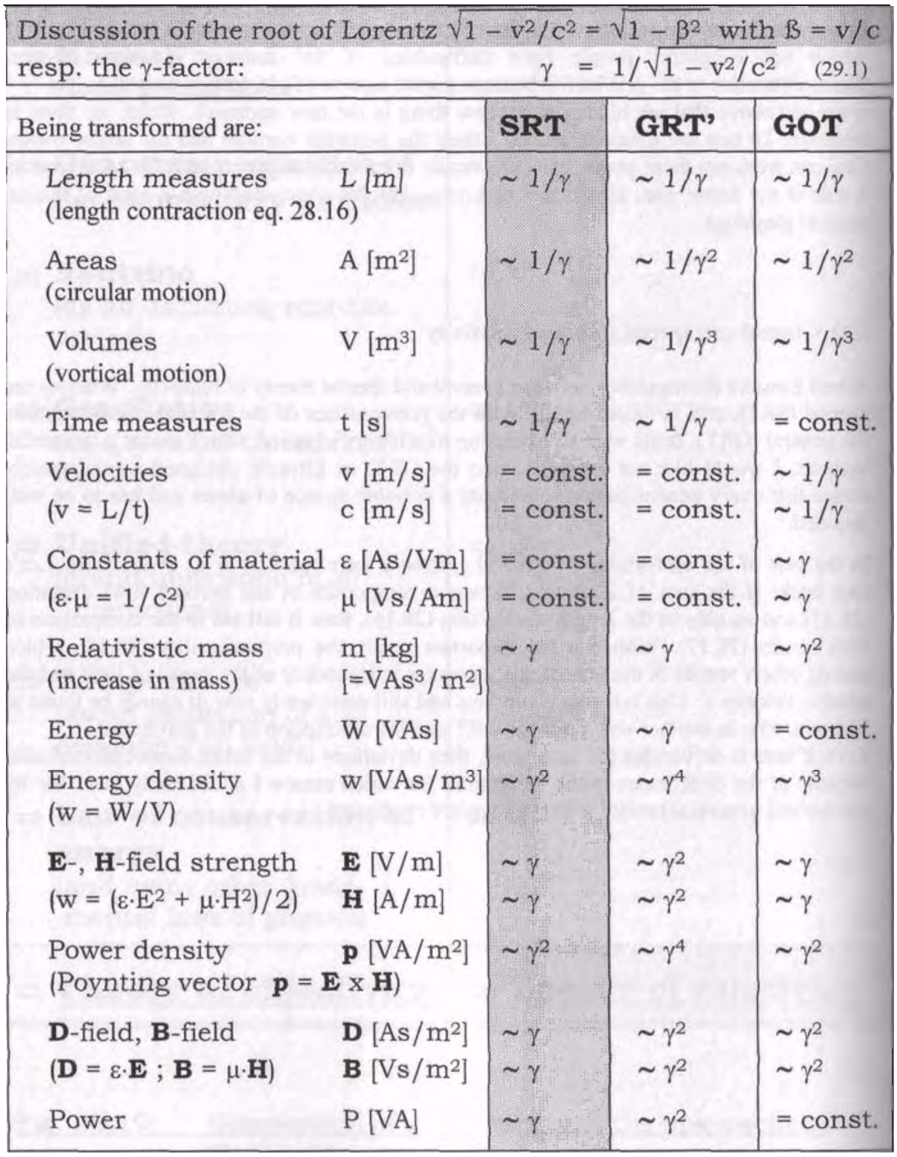

Discussion of the root of Lorentz $\sqrt{1 - v^2/c^2} = \sqrt{1 - \beta^2}$ with $ß = v/c$
resp. the γ-factor: $\gamma = 1/\sqrt{1 - v^2/c^2}$ (29.1)

Being transformed are:		SRT	GRT'	GOT
Length measures	L [m]	$\sim 1/\gamma$	$\sim 1/\gamma$	$\sim 1/\gamma$
(length contraction eq. 28.16)				
Areas	A [m^2]	$\sim 1/\gamma$	$\sim 1/\gamma^2$	$\sim 1/\gamma^2$
(circular motion)				
Volumes	V [m^3]	$\sim 1/\gamma$	$\sim 1/\gamma^3$	$\sim 1/\gamma^3$
(vortical motion)				
Time measures	t [s]	$\sim 1/\gamma$	$\sim 1/\gamma$	= const.
Velocities	v [m/s]	= const.	= const.	$\sim 1/\gamma$
$(v = L/t)$	c [m/s]	= const.	= const.	$\sim 1/\gamma$
Constants of material	ε [As/Vm]	= const.	= const.	$\sim \gamma$
$(\varepsilon \cdot \mu = 1/c^2)$	μ [Vs/Am]	= const.	= const.	$\sim \gamma$
Relativistic mass	m [kg]	$\sim \gamma$	$\sim \gamma$	$\sim \gamma^2$
(increase in mass)	[=VAs^3/m^2]			
Energy	W [VAs]	$\sim \gamma$	$\sim \gamma$	= const.
Energy density	w [VAs/m^3]	$\sim \gamma^2$	$\sim \gamma^4$	$\sim \gamma^3$
$(w = W/V)$				
E-, **H**-field strength	**E** [V/m]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma$
$(w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2)$	**H** [A/m]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma$
Power density	**p** [VA/m^2]	$\sim \gamma^2$	$\sim \gamma^4$	$\sim \gamma^2$
(Poynting vector $\mathbf{p} = \mathbf{E} \times \mathbf{H}$)				
D-field, **B**-field	**D** [As/m^2]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma^2$
$(\mathbf{D} = \varepsilon \cdot \mathbf{E}\ ;\ \mathbf{B} = \mu \cdot \mathbf{H})$	**B** [Vs/m^2]	$\sim \gamma$	$\sim \gamma^2$	$\sim \gamma^2$
Power	P [VA]	$\sim \gamma$	$\sim \gamma^2$	= const.

Fig. 29.2 Transformation table between SRT, GRT' and GOT

29.2 Transformation table

Let's speak again about the difference to the special relativity (SRT). This so to speak deals with the one-dimensional case of the uniform motion of a reference system in x-direction ($v = v_x$), as specified by the Lorentz-transformation, where only the x-components and not those in y- or z-direction are being transformed. As already mentioned this is a purely theoretical case, which in practice occurs next to never. Normal is circular and vortical and with that accelerated motion, where the velocity component permanently changes its direction.

The derived result of the general relativity (GRT') does justice to this circumstance. Even if this at first only has been derived for the x-direction it nevertheless is valid equally in y- and z-direction. It even remains valid for the case that we base on a path of arbitrary form of a spatial field vortex. In this case some components continually occur in all directions of space, so that the relative velocity v as already the speed of light c loses its vectorial nature. With that the transition of the SRT to the GRT is carried out. By means of the spatial swirling the electric and magnetic field pointers at the same time turn into scalar factors, by taking over the function of the aether. Let us remember that even Einstein in his GRT was forced to again introduce the aether, which in the SRT still was unnecessary.

It therefore makes a difference in the transformation of physical factors, if we base on a one-dimensional (SRT) or a three-dimensional spatial description (GRT). Length measures in x-direction in both cases must be converted using the root of Lorentz. Usually the relativistic γ-factor is introduced, which is inverse to the root of Lorentz

$$\gamma = 1/\sqrt{1-v^2/c^2}) \qquad \text{with } x_o/x = \gamma \quad (29.2)$$

If thus individual length measures would be subject to a length contraction following the γ -factor, then a volume V according to the SRT must be transformed with γ, according to the GRT' however with γ^3.
As is well-known a relativistic increase in mass is converted with the γ-factor and in the same manner the to that proportional energy $E = m\ c^2$. If we however correlate the energy to the volume V and in that way determine an energy density w, then the difference between SRT ($w \sim \gamma^2$) and GRT' ($w \sim \gamma^4$) again has its maximum effect.
A relation to the field factors of E- and H-field is for instance provided by the energy density of a wave field

$$w = (\varepsilon \cdot E^2 + \mu \cdot H^2)/2 \quad (29.3)$$

According to that the field strengths in the one-dimensional case of the SRT should be converted with the γ-factor, in the case of the GRT' however with γ^2, in accordance with the derivation in chapter 28. This circumstance willingly is overlooked, although it only concerns the textbooks and the today valid theory of relativity. I however point to the difference, since it does make a difference if we start with the SRT or the GRT when we change to the general theory of objectivity (GOT).

In the domain of the GOT all length measures should be transformed. The respective dimension gives information with which power the γ-factor occurs (fig. 29.2). The unit meter is responsible for that.

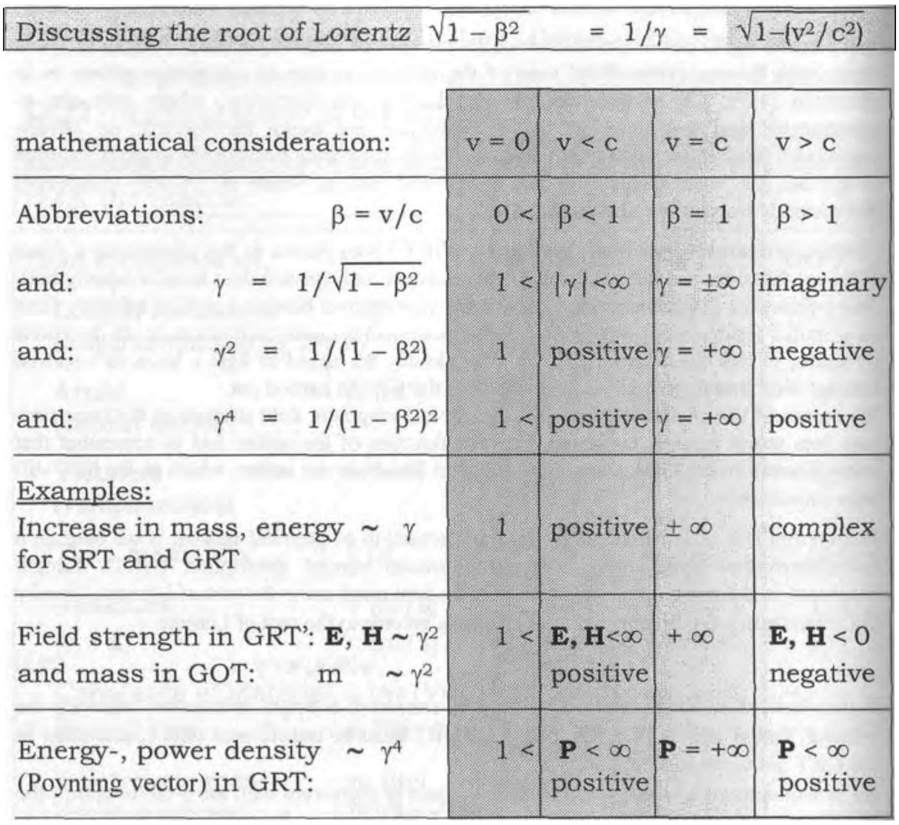

Discussing the root of Lorentz $\sqrt{1-\beta^2} = 1/\gamma = \sqrt{1-(v^2/c^2)}$

mathematical consideration:	v = 0	v < c	v = c	v > c
Abbreviations: $\beta = v/c$	0 <	$\beta < 1$	$\beta = 1$	$\beta > 1$
and: $\gamma = 1/\sqrt{1-\beta^2}$	1 <	$\|\gamma\| < \infty$	$\gamma = \pm\infty$	imaginary
and: $\gamma^2 = 1/(1-\beta^2)$	1	positive	$\gamma = +\infty$	negative
and: $\gamma^4 = 1/(1-\beta^2)^2$	1 <	positive	$\gamma = +\infty$	positive
Examples: Increase in mass, energy $\sim \gamma$ for SRT and GRT	1	positive	$\pm\infty$	complex
Field strength in GRT': **E**, **H** $\sim \gamma^2$ and mass in GOT: m $\sim \gamma^2$	1 <	**E, H**$<\infty$ positive	$+\infty$	**E, H** < 0 negative
Energy-, power density $\sim \gamma^4$ (Poynting vector) in GRT:	1 <	**P** $< \infty$ positive	**P** $= +\infty$	**P** $< \infty$ positive

Fig. 29.3 Discussion concerning the root of Lorentz

- The special theory of relativity SKT only is defined for v < c
- For v • c particles with a complex mass, but with a real energy density (according to GRT') would result.
- From the point of view of the theory of objectivity (GOT) the mass should be taken negative-real (neutrino?).

29.3 Discussion concerning the root of Lorentz

Fig. 29.2 forms the basis for the transformation in the domain of the GOT, the general theory of objectivity, where it plays a role, if a subjectively measured factor should be converted from the laboratory domain or a relativistic factor according to the SRT or the GRT'. The given proportionalities thereby should be put in the respective relation. In this way results the respective valid instruction for transformation on the basis of the root of Lorentz.

Let's take a critical look at the root of Lorentz. The velocity v occurring in it, of whatever this may consist, is depending on the field according to equation 28.14 + 28.15. It strictly speaking wouldn't be constant anymore and wouldn't belong in a general instruction for transformation at all. Only, what is valid for v, is valid to the same extent for c. Since only the proportion of v/c occurs in the root of Lorentz every influence depending on field or of other nature will have no effect on v/c and the value of the root of Lorentz. It in any case will retain its value. It fulfills for itself the condition of the Lorentz invariance. According to that in the case of the relative velocity v it doesn't depend on the absolute value, but only on the relation to the speed of light. In addition the restriction to values of $v < c$ is normal, if the speed of light is seen as an upper limit. Let's first purely mathematically draw a case distinction for different velocity domains of v. For $v = 0$ the root of Lorentz becomes one and the Lorentz transformation turns into the Galilei transformation.

Connected to this is the today well-known and technically used domain up to the limit of $v = c$. It virtually is impossible to accelerate a mass particle to the speed of light, since mass, field and energy would grow towards infinity, as is clear from the table (fig. 29.2). Particles as fast as light, like photons, hence cannot have a mass. At $v = c$ a singularity is present.

In a field theory, which also deserves this name, however an upper limit must not be present. Hence also the case for $v > c$ should be required theoretically. Only later we will be able to judge if this makes sense physically. We at first only want to examine the case mathematically. Mass, field and energy now again have a finite value, there however results a complex, purely imaginary mass, a negative field and doing so, as already before a positive energy and power density.

There sometime has been the textbook opinion that it is physically impossible to fly faster than sound. This erroneous statement even could be proven „scientifically", because such a supersonic airplane would fly off the observation space and with that wouldn't be real anymore, thus from a mathematical viewpoint would be complex. Anyone, who in New York gets off a Concorde, can confirm that everything at any moment of the flight was real. Only the observer is deceived, if the airplane flies somewhere else, than he perceives it. Is the speed of light also such a „sonic barrier", which by the majority of the scientists since Einstein until today still is thought to be insurmountable?

How should one physically imagine a complex mass? Let us remember the alternating current teachings, where it is normal to work with complex values, since the mean values of the oscillating alternating currents, tension voltages and fields are zero. Calculating with mean values would result in zero energy and power. Hence complex factors are introduced and the root mean square values are calculated and measured instead of the mean values. Could a complex mass analogously not concern an oscillating particle, a particle, which in addition is faster than the light?

- Physical example:

neutrino = mean of the oscillating mass ≈ zero
RMS value of the complex mass ≠ zero
E- and H-field with inverse sign
carrier of positive energy and power
ν -model: ring-like vortex oscillating from e^- to e^+ <i>

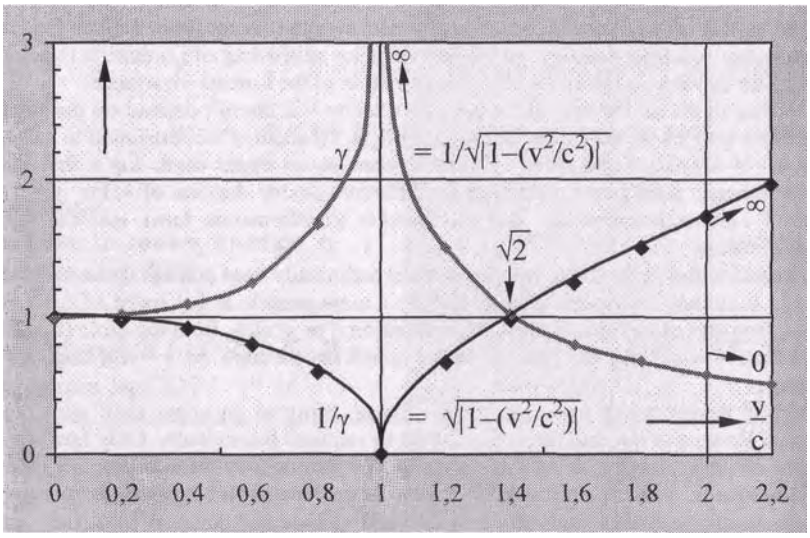

Examples from the general theory of relativity (GRT):

(28.16): length contraction $L \sim 1/\gamma = \sqrt{1-(v^2/c^2)}$

(6.19): increase in mass $m \sim \gamma = 1/\sqrt{1-(v^2/c^2)}$

(6.6,28.15) field dilatation $\mathbf{E} \sim \mathbf{H} \sim \gamma^2 = 1/(1-v^2/c^2)$

Fig. 29.4: Root of Lorentz for speeds faster than light (v > c)

29.4 Mathematical derivation of the neutrino

In the domain of speeds faster than light, for v > c, the power series (28.12) does not converge anymore and every observer theory and every observation (fig. 28.4) will fail, because particles faster than light run away of their own visible appearances. Every measurement and every observation inevitably is behind and hardly can be assigned to the actual cause. That way for instance measured neutrino events are being connected with celestial observations, with which they haven't got anything to do at all.
If we however describe the domain v > c in the complex plane, then astonishing results are found, which could be verified physically: a complex length dilatation with increasing velocity goes along with a loss of complex mass. The oscillating fields, energy and power density however would be real with negative sign.
Thus there would result particles carrying energy with an oppositely poled field, with an oscillating mass and if necessary also an oscillating charge. Without static mass and charge these particles hardly would interact with normal matter, which leads to an enormous ability of penetration. The only physical particles, which have such a property, are the neutrinos. With that a usable and an extremely efficient model description has been found for these particles. Also the energy of these particles can be calculated, which has considerable orders of magnitude and is available as an energy source everywhere and any time.

If for instance in a converter for space energy a neutrino should be converted into a resting charge carrier (with v = 0), then two steps are necessary (see part 2 of this series of books):

1. First the neutrino must be slowed down to 1.414 times the speed of light (fig. 28.9). Doing so energy is spent and not won! The converter for instance can cool down.

2. Afterwards the characteristic rotation of its own, with which the ring-like vortex spins around itself by permanently putting its inside to the outside and vice versa, has to be taken away from the neutrino. In that way the vortex centre is closed and the particle acquires localization. It becomes a charge carrier.

Even if the representation in the complex plane represents only an auxiliary description, the model nevertheless seems to be efficient, because despite its complex mass and charge the neutrino nevertheless carries a real energy. It in any case is represented in that way to an observer, who measures the relation with the speed of light, who in the relativistic scheme of things scans the relation.

Today, as already said, even the sonic barrier has become permeable and no scientist dares to physically deny this fact and even prove his mistake mathematically anymore. No, on the contrary, he always did know that as an expected consequence the sonic barrier runs after the supersonic plane. The once physically unthinkable and scientifically fought has become normality.
What should hinder an oscillating particle, like a neutrino, to be faster than the light? Some time one also will accustom to that.

- Starting-point:
  the fundamental field equation
  (derivation in chapter 27.10)

$$-c^2\cdot\text{rot rot }\mathbf{B} = \frac{\partial^2\mathbf{B}}{\partial t^2} + \frac{1}{\tau_1}\frac{\partial\mathbf{B}}{\partial t} + \frac{1}{\tau_2}\frac{\partial\mathbf{B}}{\partial t} + \frac{\mathbf{B}}{\tau_1\tau_2} \qquad (27.26)$$

- Example:

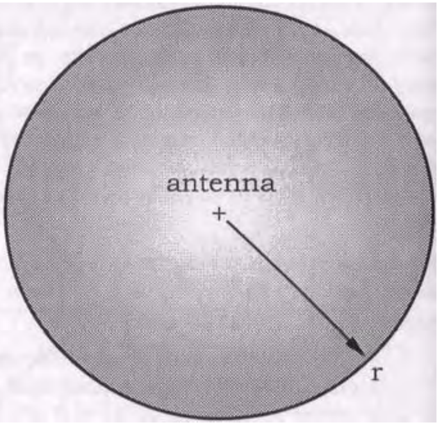

Spherical propagation of an electromagnetic wave in space

- The number π: $\pi = 3.14159$

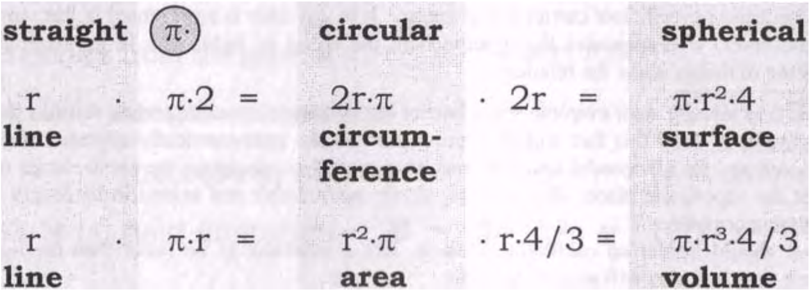

straight	π·	**circular**		**spherical**
r	$\cdot\ \pi\cdot 2 =$	$2r\cdot\pi$	$\cdot\ 2r =$	$\pi\cdot r^2\cdot 4$
line		**circumference**		**surface**
r	$\cdot\ \pi\cdot r =$	$r^2\cdot\pi$	$\cdot\ r\cdot 4/3 =$	$\pi\cdot r^3\cdot 4/3$
line		**area**		**volume**

(29.4)

Fig. 29.5: The number π

29.5 Universal constants and constants of nature

Strictly speaking fundamental or universal constants can't exist in pure field physics at all. For the example of the speed of light has been shown that it merely are measurement constants (fig. 6.11). The numerical size is a consequence of the definition. The speed of light for instance has the today well-known value as a consequence of the fixing of a length unit and a time unit. If we change a unit, if we take „feet" instead of ,.meters", then another value results. If however the velocity changes, then the reference measure of our gauge changes along and we get the identical reading. Electric and magnetic field constant depend directly on the speed of light c ($\varepsilon \cdot \mu = 1/c^2$), which solely by the fixing of the electric units takes a certain value. The inner structure of the potential vortices leads from Planck's quantum of action, to the elementary charge and to the countless atomic „constants", which all objectively seen aren't constants at all. They virtually all can be derived (chapter 7).

Sooner or later even the last natural scientist will realize, that nature does not provide "constants" at all. If „constants of nature"' did exist, as listed in textbooks and encyclopaedias, then they aren't consistent with causality, since we don't know the cause, why the factor should have exactly this size and no other. Behind every so-called constant of nature unnoticed is hiding a physical closed loop conclusion.

Fundamental constants only exist in mathematics. This can be shown very nicely for the example of the „fundamental field equation" (eq. 27.26), which has been derived from the new field theoretical approach (fig. 27.10). It is the description of waves and vortices in space and time, which indeed carries features of a „world equation". If one searches this equation for fundamental constants, then one altogether can find three: the number π, the number e and the Golden Proportion Φ. The speed of light c however occurs only as the mathematical factor characterizing the wave as a result of the defined units. If one would choose the units different, c as well could be made 1. With the fundamental numbers that procedure won't work. They don't depend on the definition of the units!

Let's consider the number π. The number π occurs every time as a proportionality factor if we transition from a straight line to a circle or further to a sphere, from a line to the circumference or further to the surface of a sphere and exactly so from a line to the area of a circle or further to the volume of a sphere. Since for all the special cases, which are derived from the fundamental field equation (the structure of the elementary particles, the atomic structure and in the same way again in the universe), the spherical symmetry dominates, the mathematical solution is determined by a corresponding spatial configuration of the number π. It has its cause neither in a physical relation of interactions, nor in the choice of the units, but only in the geometry.

The number e: $\boxed{e = \lim_{n \to \infty} (1 + 1/n)^n = } \; 2.71828$ (29.5)

for n = 2 is $(1+1/n)^n = 1.5^2 = 2.25$
for n = 1 is $(1+1/n)^n = 2^1 = 2$
for n = 0.5 is $(1+1/n)^n = \sqrt{3} = 1.73$
for n = 0 is $(1+1/n)^n = \infty^0 = 1$
for n = -0.25 is $(1+1/n)^n = 1/(-3)^{0.25}$ *gives a complex solution!*
for n = -0.5 is $(1+1/n)^n = 1/\sqrt{-1} = -i$, thus again complex!
for n = -0.75 is $(1+1/n)^n = 1/(-3)^{0.75}$ as well complex!
for n = - 1 is $(1+1/n)^n = 1/0 = \pm\infty$
for n = - 2 is $(1+1/n)^n = 2^2 = 4$

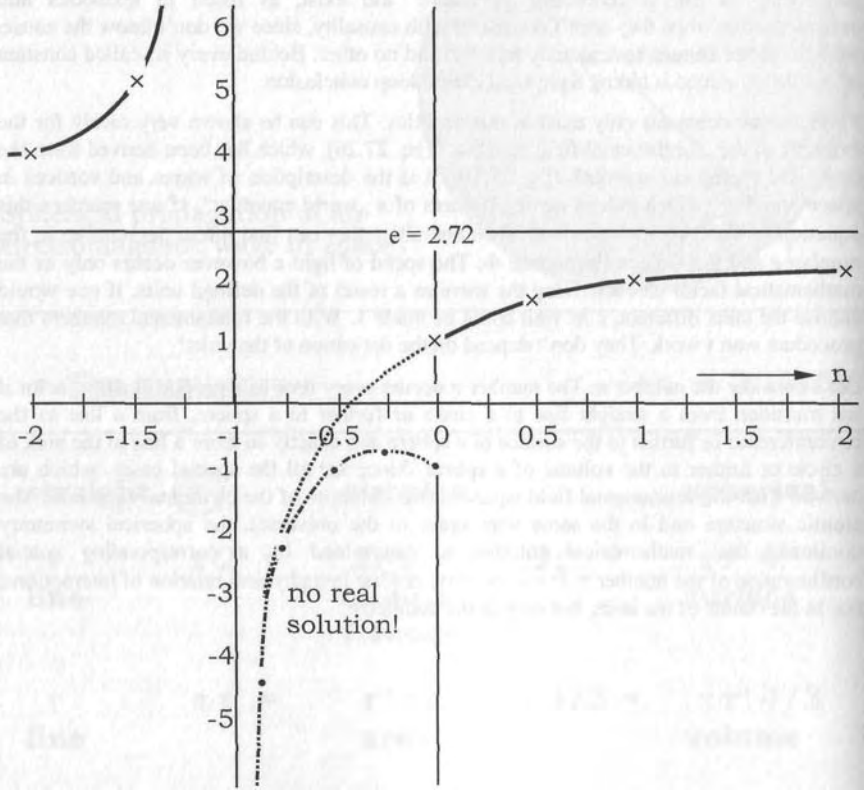

Fig. 29.6: Forming the limiting value of the number e

29.6 The fundamental number e

In the fundamental field equation (27.26) a further irrational number is concealed, the number e. Whereas the left side of the equation (a) gives the spatial distribution, the right side (b-e) describes the temporal course of events. Besides the term constant in time (e) also first time derivations (c and d) and a second time derivation (b) occur. For a solution of this differential equation a function should be found, the derivations of which are again the function itself. This condition strictly speaking is fulfilled only by one single function, the e-function.
We used this property of the e-function already for the derivation of the Schrodinger-equation in chapter 5.6 and 5.7. There with the help of the e-function an approach was chosen, which leads to the well-known solutions of the Schrodinger-equation, which are considered to be secured experimentally. With that the number e controls the temporal relations of the fundamental field equation.

It might be helpful to take a closer look at the origin of the number e. It results from a consideration of limiting values: _

$$e = \lim_{n \to \infty} (1 + 1/n)^n = 2.71828 \qquad (29.5)$$

If one varies n and allows different values between -∞ and +∞, then a strange behaviour is showing. One indeed more and more approaches the well-known value of e = 2.72, as dictated by the definition of limiting values according to equation (29.5), the larger n is chosen. But in the opposite direction it looks less tidied:

$$\begin{array}{llllllll}
 & \text{for} & n = 1 & \text{is} & (1+1/n)^n & = 2^1 & = 2 & \\
 & \text{for} & n = 0{,}5 & \text{is} & (1+1/n)^n & = \sqrt{3} & = 1.73 & \\
 & \text{for} & n = 0 & \text{is} & (1+1/n)^n & = \infty^0 & = 1 & \\
 & \text{for} & n = -0{,}5 & \text{is} & (1+1/n)^n & = 1/\sqrt{-1} & = -i\text{, thus complex!} & \\
\text{and} & \text{for} & n = -1 & \text{is} & (1+1/n)^n & = 1/0 & = \pm\infty & (29.6)
\end{array}$$

Since the e-function inside the fundamental field equation is responsible for the temporal sequence, the interpretation of my colleague Prof. Dr. Preussker<i> gets a deeper sense. He says, it starts outside our imagination (at n = -1). Afterwards at first big chaos prevails. Mathematically seen some imaginary solutions arise. Finally the system is putting in order (from n = 0), to more and more approach the value e = 2.72 .

The number e is of fundamental importance and thereby holds unforeseen secrets. More mysterious and until now entirely misunderstood is the meaning of the Golden Proportion. Also this indivisible number can be found in the fundamental field equation. Since it is less known and more complicated to handle, it first shall be introduced.

<i>: H. Preussker: Der Wirbelring, Halstenbeck 2002

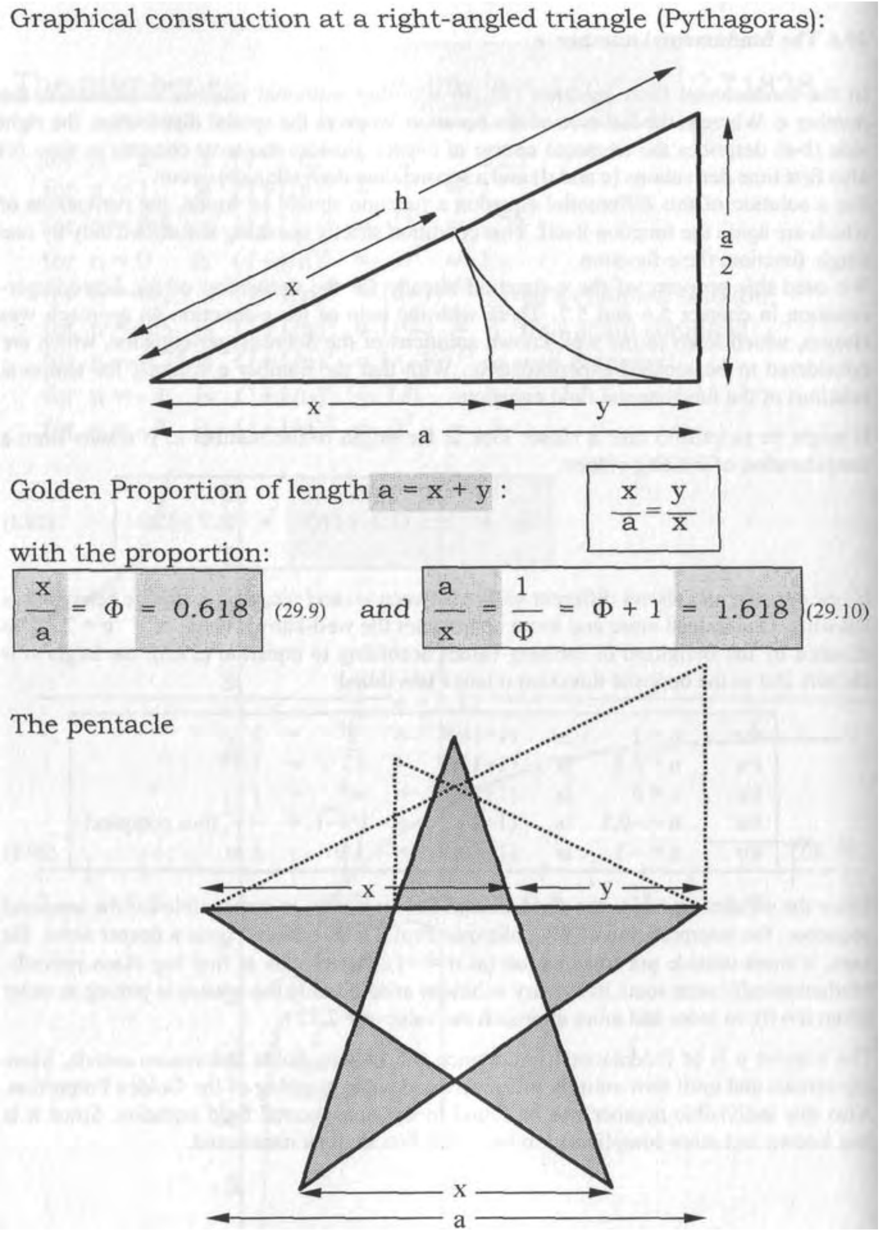
Graphical construction at a right-angled triangle (Pythagoras):
h
x
$\frac{a}{2}$
x
y
a
Golden Proportion of length a = x + y :
$\frac{x}{a} = \frac{y}{x}$
with the proportion:
$\frac{x}{a} = \Phi = 0.618$ (29.9)
and
$\frac{a}{x} = \frac{1}{\Phi} = \Phi + 1 = 1.618$ (29.10)
The pentacle
x
y
x
a

Fig. 29.7: Constructions for the Golden Proportion

29.7 The pentacle

During his visit of the Egyptian pyramids already two and a half thousand years ago the history writer Herodotus by his guide had been called attention to the circumstance that the Golden Proportion has been realized for the proportions of scale.
In the case of the pyramid of Cheops it even has been taken into account manifold, as we know today, but we hence still don't know why. There must be an intention behind it. Anyway a coincidental use can be eliminated, since the Golden Proportion cannot be handled in an easy way, neither graphically nor mathematically.
The Golden Proportion in addition plays an important role in the whole ancient architecture and not only there. It for instance occurs in the case of a very old symbol, the five edged star, which we can draw with one line without taking off. The well-known symbol also is called pentacle.

In the case of the Golden Proportion a straight line a is divided into two unequal halves. The larger half x thereby is 61.8 % of the straight line a. Already Pythagoras has researched and teached about this. Maybe he did know more about the purpose of this classification than all the mathematicians, archaeologists and art historians of today together.
For a graphical solution we assume a right-angled triangle. The task is to divide one leg of length a = x + y according to the Golden Proportion into two parts, the larger part x and the smaller part y. The second leg has the length a/2. According to the theorem of Pythagoras the length of the hypotenuse h is

$$h = \sqrt{a^2 + (a/2)^2} = (a/2)\cdot\sqrt{5} \quad (29.7)$$

If the length of the second leg (a/2) is subtracted from the hypotenuse, then this is the sought length x = h - a/2:

$$\boxed{x = (a/2)\cdot(\sqrt{5} - 1)} \quad (29.8)$$

The proportion of both length measures gives the constant Φ, which is characteristic for the Golden Proportion:

$$\boxed{\Phi = x/a = (\sqrt{5} - 1)/2 = 0.618} \quad (29.9)$$

This proportional number has a special property. If one adds 1 to the number and forms the reciprocal value of that, then the same number comes out again, thus:
with $\Phi = x/a$: with $a = x + y$:

$$\boxed{\Phi + 1 = \frac{1}{\Phi}} \quad \text{and} \quad \boxed{\frac{x}{a} + 1 = \frac{a}{x}} \quad \text{resp.} \quad \boxed{\frac{x}{a} = \frac{a - x}{x} = \frac{y}{x}} \qquad (29.10)$$

With that the ratio of the length a and the larger section x is the same as the ratio of x and the smaller section y.

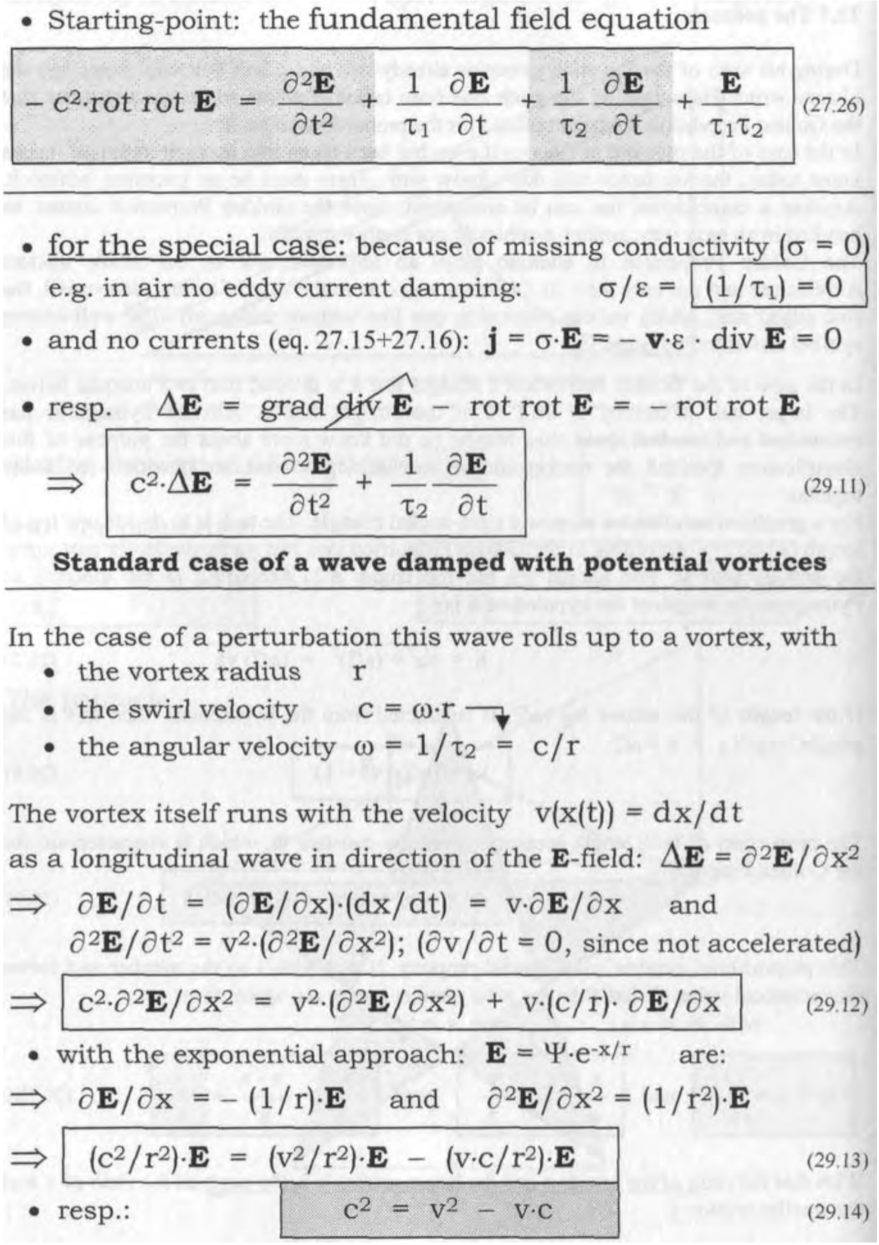

- Starting-point: the fundamental field equation

$$\boxed{-c^2 \cdot \text{rot rot } \mathbf{E} = \frac{\partial^2 \mathbf{E}}{\partial t^2} + \frac{1}{\tau_1}\frac{\partial \mathbf{E}}{\partial t} + \frac{1}{\tau_2}\frac{\partial \mathbf{E}}{\partial t} + \frac{\mathbf{E}}{\tau_1 \tau_2}} \qquad (27.26)$$

- for the special case: because of missing conductivity ($\sigma = 0$)

e.g. in air no eddy current damping: $\sigma/\varepsilon = \boxed{(1/\tau_1) = 0}$

- and no currents (eq. 27.15+27.16): $\mathbf{j} = \sigma \cdot \mathbf{E} = -\mathbf{v} \cdot \varepsilon \cdot \text{div } \mathbf{E} = 0$
- resp. $\Delta \mathbf{E} = \text{grad } \cancel{\text{div}}\, \mathbf{E} - \text{rot rot } \mathbf{E} = -\text{rot rot } \mathbf{E}$

$$\Rightarrow \quad \boxed{c^2 \cdot \Delta \mathbf{E} = \frac{\partial^2 \mathbf{E}}{\partial t^2} + \frac{1}{\tau_2}\frac{\partial \mathbf{E}}{\partial t}} \qquad (29.11)$$

Standard case of a wave damped with potential vortices

In the case of a perturbation this wave rolls up to a vortex, with

- the vortex radius r
- the swirl velocity $c = \omega \cdot r$
- the angular velocity $\omega = 1/\tau_2 = c/r$

The vortex itself runs with the velocity $v(x(t)) = dx/dt$

as a longitudinal wave in direction of the **E**-field: $\Delta \mathbf{E} = \partial^2 \mathbf{E}/\partial x^2$

$\Rightarrow \quad \partial \mathbf{E}/\partial t = (\partial \mathbf{E}/\partial x) \cdot (dx/dt) = v \cdot \partial \mathbf{E}/\partial x$ and

$\partial^2 \mathbf{E}/\partial t^2 = v^2 \cdot (\partial^2 \mathbf{E}/\partial x^2)$; ($\partial v/\partial t = 0$, since not accelerated)

$$\Rightarrow \quad \boxed{c^2 \cdot \partial^2 \mathbf{E}/\partial x^2 = v^2 \cdot (\partial^2 \mathbf{E}/\partial x^2) + v \cdot (c/r) \cdot \partial \mathbf{E}/\partial x} \qquad (29.12)$$

- with the exponential approach: $\mathbf{E} = \Psi \cdot e^{-x/r}$ are:

$\Rightarrow \quad \partial \mathbf{E}/\partial x = -(1/r) \cdot \mathbf{E}$ and $\partial^2 \mathbf{E}/\partial x^2 = (1/r^2) \cdot \mathbf{E}$

$$\Rightarrow \quad \boxed{(c^2/r^2) \cdot \mathbf{E} = (v^2/r^2) \cdot \mathbf{E} - (v \cdot c/r^2) \cdot \mathbf{E}} \qquad (29.13)$$

- resp.: $\boxed{c^2 = v^2 - v \cdot c}$ (29.14)

Fig. 29.8: The calculation of an electromagnetic wave, which is rolling up to a potential vortex.

The big mystery concerning the harmony of the Golden Proportion gets a sober technical-physical dimension with the theory of objectivity. It determines within the fundamental field equation" (27.26) the rolling up of a wave into a vortex and vice versa. The Golden Proportion mathematically describes the process known as wave damping, as we can make ourselves clear.

29.8 The vortex, which is rolling up

For the case of a wave propagation in air or in vacuum, if no electric conductivity is present (σ = 0), the fundamental field equation is reduced to the two parts: the description of the electromagnetic wave and the potential vortex as a damping term. Now a solution of this partial differential equation (29.11) is sought. This only succeeds for a very particular course of spatial and temporal field.

If a wave for a field perturbation rolls up to a vortex, which we had worked out as a model concept, then the field oscillation continues to run with the speed of light, but this time in circles. With this consideration the relation between the angular velocity resp. the time constants and the radius of the circular vortex has been described (τ_2 = r/c).

v(x(t)) = dx/dt is the not-accelerated velocity of propagation of a vortex. In that case v points in the x-direction radially to the outside. For the time derivation of the field vector E(x(t)) the chain rule should be applied. With that the field equation (29.11), defined in space and time, can be converted into an equation determined in v and x (29.12).

Finally we use the mentioned property of the e-function, which for first and second derivation again turns into itself, by choosing the approach of an exponential damping with $e^{-x/r}$ There remains a quadratic equation to determine the velocity v (29.14 and 29.15). From the two solutions of the quadratic equation only the one with positive velocity should be considered (29.17) and that would be 1.618 times the speed of light! (29.18).
If we subtract 1 from this value or form the reciprocal value, then in both cases the factor Φ = 0.618 results, which is called the Golden Proportion (29.19).

Behind this clear, mathematically won result is hiding a deeper physical meaning. Obviously nothing can hinder a longitudinal wave and its vortices to be slower or faster than just with v = 1.618*c. Let us take the case that v = c, for which there even exist calculations in some textbooks<i>. Then as a result gets out that the longitudinal parts decrease very quickly and already can be neglected after $\lambda/2\pi$. I in any case interpret the
near-field zone of an antenna (fig. 29.9) such that within one sixth of the wavelength the vortices to a large extent have decayed.

<i>: Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986, Seite 335

- The quadratic equation 29.14: $\boxed{v^2 - v\cdot c - c^2 = 0}$ (29.15)
- has the solution: $\boxed{v = c/2 \pm\sqrt{c^2/4 + c^2}}$ (29.16)
- for the proportion v/c: $\boxed{v/c = (1 +\sqrt{5})/2 \ = \ 1.618}$ (29.17)
- The result $v/c = \underline{1.618}$ (29.18)
- resp.: $c/v = \Phi = \underline{0.618} =$ „Golden Proportion"(29.19)
- acc. to that is: $1 + \Phi = 1.618 = 1/\Phi = 1/0.618$(29.18*)

- Concerning the physical meaning of $v/c = 1.618$ (29.17)

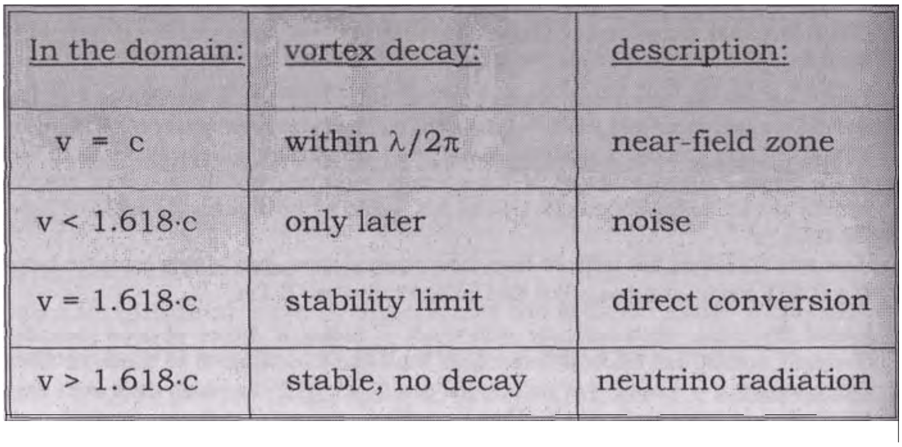

In the domain:	vortex decay:	description:
$v = c$	within $\lambda/2\pi$	near-field zone
$v < 1.618\cdot c$	only later	noise
$v = 1.618\cdot c$	stability limit	direct conversion
$v > 1.618\cdot c$	stable, no decay	neutrino radiation

=>In the ideal case at 1.618 times the speed of light at the stability limit, the natural and spontaneous conversion from vortex to wave and vice versa occurs.

=> Scalar wave antennas have their optimum degree of effectiveness in the Golden Proportion!

Fig. 29.9: Derivation of the Golden Proportion from the fundamental field equation.

For larger velocities, as Tesla has produced already 100 years ago, the stability of the vortices, but also their biological effectiveness increases. From a high-frequency technical viewpoint this is the domain of the noise, the domain of instable oscillating noise vortices. From a biological and medical viewpoint especially here the answers should be sought to questions concerning the electromagnetic environmental compatibility. This domain stretches from v = c to v = 1.618*c, the boundary case, where a spontaneous conversion of the wave into the vortex is possible. The Golden Proportion describes the stability limit, where the field vortices after they have been formed, don't decay anymore. From this velocity on the antenna resp. radiation source hasn't got a limited near-field zone anymore, since that now reaches to infinity. Only from this distinct velocity on I give the field vortices the name "neutrino".

29.9 The Golden Proportion

Whereas traditional antennas, which should emit electromagnetic waves, are optimized for signals with the speed of light, do scalar wave antennas have their optimum degree of effectiveness in the Golden Proportion. Thereby the goal of an emission of oscillating ring-like vortices is pursued. Vortices, which with a velocity of propagation of v = 1.618-c remain stable and make possible a spatially unlimited transmission of energy and information by scalar waves.

The velocity of propagation v is calculated from the product of wavelength times frequency. $v = \lambda \cdot f$. If the frequency of a transmitter is prescribed, then the geometry of the transmitting antenna decides on the degree of effectiveness of the antenna and on how many parts of transverse waves are emitted in relation to longitudinal waves. In the case of a broadcast antenna usually λ or $\lambda/2$ is chosen. For a scalar wave however the optimum is reached with

$$\boxed{\lambda_{\text{vortex}} = 1.618 \cdot \lambda_{\text{wave}} = 1.618 \cdot c/f.} \qquad (29.20)$$

With that the electrotechnical problem becomes a geometrical one, if it concerns the use of scalar waves. Crucial is the antenna geometry, and the Golden Proportion provides the necessary constructional instruction. Were the buildings in antiquity, which were constructed according to the Golden Proportion, technical facilities for scalar waves[i]? Did the builders have physically explainable and mathematically provable guidelines? At this place, by the derivation of the Golden Proportion from the fundamental field equation, there arise completely new aspects for the judgement and interpretation of buildings especially from antiquity. If we have understood their way of functioning, then we will be able to learn much from that for our own future and for the future construction of scalar wave devices. Concluding the seminar we hence deal with antiquity.

ancient temple = short wave station
dedicated to one god = fixing of the frequency
supreme god Zeus, = range of the short wave,
father of all gods = all SW wave bands
priest, representative of the god = amateur radio operator, with licence to transmit
high priest = chief intendant
Pontifex Maximus, „topmost bridge builder" = chairman of the authority and the telegraph offices
oracle[i] = telegraphy receiver
runes, cuneiform writing = telegraphy symbols
metre, hexameter = increase of redundancy
oracle priest = telegraphy interpreter
tripod[ii] = reception key, electro-acoustical converter
looking at intestines, rite of sacrificing = reading off convulsions, electro-optical converter
temple books = news-notes
seer, who looks into the god world = amateur radio operator, at telegraphy reception
Homer = ancient radio reporter

Fig. 30.1: ____Little dictionary for ancient radio engineering (1).

<i>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 under "oracle" stands: „Lat. oraculum = site of speaking; particularly: site, where a god speaks; then: that, what the god says". According to the encyclopaedia oracle priests inspired by a god „simply were frauds, who lived of the ignorance of the public".

30. Scalar wave technology in antiquity

The end of the book about potential vortices and their propagation as a scalar wave shall form an impressive example, where as many of the derived wave aspects as possible have an effect. It shall be proven that already in antiquity radio engineering based on scalar waves has been used. The proof starts with a thesis.

30.1 Thesis:

The temples in antiquity all were short wave broadcasting stations. And energy from the field served as an energy source, so e.g. the earth radiation in the case of temples of terrestrial gods. In the case of the solar god the radiation of the sun was used, whereas for the temples, which were dedicated to the planetary gods, the neutrino radiation arriving from the planets served as an energy source.

If the temple was dedicated to a particular god, then the name of the god was representing the used frequency of the broadcasting company. The corresponding wavelength, resp. the respective god, understandably was "immortal".

Not so the broadcasting technicians on duty, who as human beings naturally were mortal, who took turns in the studio as members of the priest council and who merely had to impersonate the god Apollo, Poseidon etc. by the name of the broadcasting company, if they went on air. Only for the news editor Homer and for few of his colleagues we actually know the names of the persons behind the scenes.

In the temple books the texts have been recorded, which a god and its broadcasting company have received. The chosen metre served the easier detection and correction of transmission errors.

Here often a lot of fantasy was necessary, for which reason the reception facilities commonly were described as oracle<i>. The reception of the news as a rule took place on an altar. Thereby the direct effect of scalar waves on man e.g. in the case of the so-called temple sleep or the indirect influence on biological systems, e.g. on the intestines of slaughtered animals, was evaluated.

A further development of the telegraphy was the tripod technology<ii>, in which case by turning of the polarization plane individual symbols and letters were transmitted up to the transmission of the spoken word with the help of a special wavelength modulation. That far the thesis reaches, which now should be proven.

<ii>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 under "tripod" stands: „the tripod is a dedication gift to gods, a honorary gift for winners. That one pleased them with the gift of a cooking pot, is strange; one has tried to find the reason, but until now without success". In the encyclopaedia further is advised: „tripod as a means for spiritual insight?!"

godology	= high frequency technology
god name	= RDS, station identification
members of a family of gods	= broadcasting studios of a broadcasting company
Pantheon, temple of all gods	= extremely broadband FM broadcast station
crown	= antenna netting
gifts for sacrificing	= broadcasting fees
place of sacrificing	= place of a node of the standing wave
earth radiation	= power supply
homage of a weekday	= time restriction of the operation of the station
Zeus „forges" thunderbolts	= electrostatic blows, when a temple is oscillating
ritual act	= technical provision for transmission and reception
Pythia of Delphi<i>	= radio telephone operator, receptionist
Cella (marrow of temple)	= tuned cavity
obelisk<ii>	= antenna rod

Fig. 30.2: ____ Little dictionary for ancient radio engineering (2).

<i>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96, under „Pythia"(gr.=the asker): In the temple of Delphi was a chasm. A tripod was standing over it, on which the Pythia was seated, if she gave oracle. Comment: "It was flashy how odd the Pythia was sitting; inconveniently enough, on the cooking pot of the tripod. Ancient pictures, which show her that way, still weren't proven, that it was this way ..."

<ii>: Lamer: under „Obelisk" stands; „The obelisks probably were clocks".

30.2 Prerequisites

The argumentation has to be made on mathematical-physical foundation. The prerequisite for that are the 29 chapters of before. The following points could be demonstrated and derived:

1. The wave equation (inhomogeneous Laplace equation) describes the sum of two wave parts, where
2. every antenna emits both parts,
3. the transverse part, known as electromagnetic wave (Hertzian wave)
4. and the longitudinal part (Tesla radiation) termed scalar wave by the discoverer, better known as antenna noise.
5. The wave equation mathematically describes the connection of both wave parts in general and the conversion of one part into the other in particular, thus
6. the rolling up and unrolling of waves in field vortices (measurable as noise).
7. The transition takes place proportionally to the Golden Proportion, as resulted from the derivation (chapter 29.7 - 29.9)

With the last point the electrotechnical problem becomes a geometrical problem, if it concerns the use of scalar waves. The geometry of the antenna is crucial. Thereby the Golden Proportion provides the necessary direction for construction.
That justifies the assumption that the buildings in antiquity, which were built according to the Golden Proportion, were technical facilities for scalar waves. Maybe the builders had specifications that had physical reasons and could mathematically be proven.
At this place there result completely new aspects for judging and interpreting buildings especially from antiquity through the derivation of the Golden Proportion from the fundamental field equation. If we have understood their way of functioning, then we will be able to learn much from that for our own future and for the future construction of scalar wave devices.

As a further prerequisite for the ancient broadcasting technology enough field energy should be at disposal. We proceed from the assumption that

1. the earth magnetism and the cosmic neutrino radiation are tightly hanging together by the processes in earth's core,
2. the earth magnetism in antiquity verifiably was approx. thousandfold stronger than today (proven by gauging of pieces of broken pot),
3. as a consequence the neutrino radiation in antiquity as well must have been thousandfold stronger and
4. the cosmic neutrino radiation has served the transmitting plants of antiquity as an energy carrier,

any thought is absurd to reject the technical function of a temple only because it today can't be reproduced anymore. The artistic and aesthetical viewpoints, which are put into the foreground by art historians because of ignorance about the true function, rather are secondary.

The terms used to describe the broadcasting technology in antiquity in the last 2000 years have experienced a shift of meaning, so that a translation in our linguistic usage of today is necessary. The adjacent dictionary should help in that case.

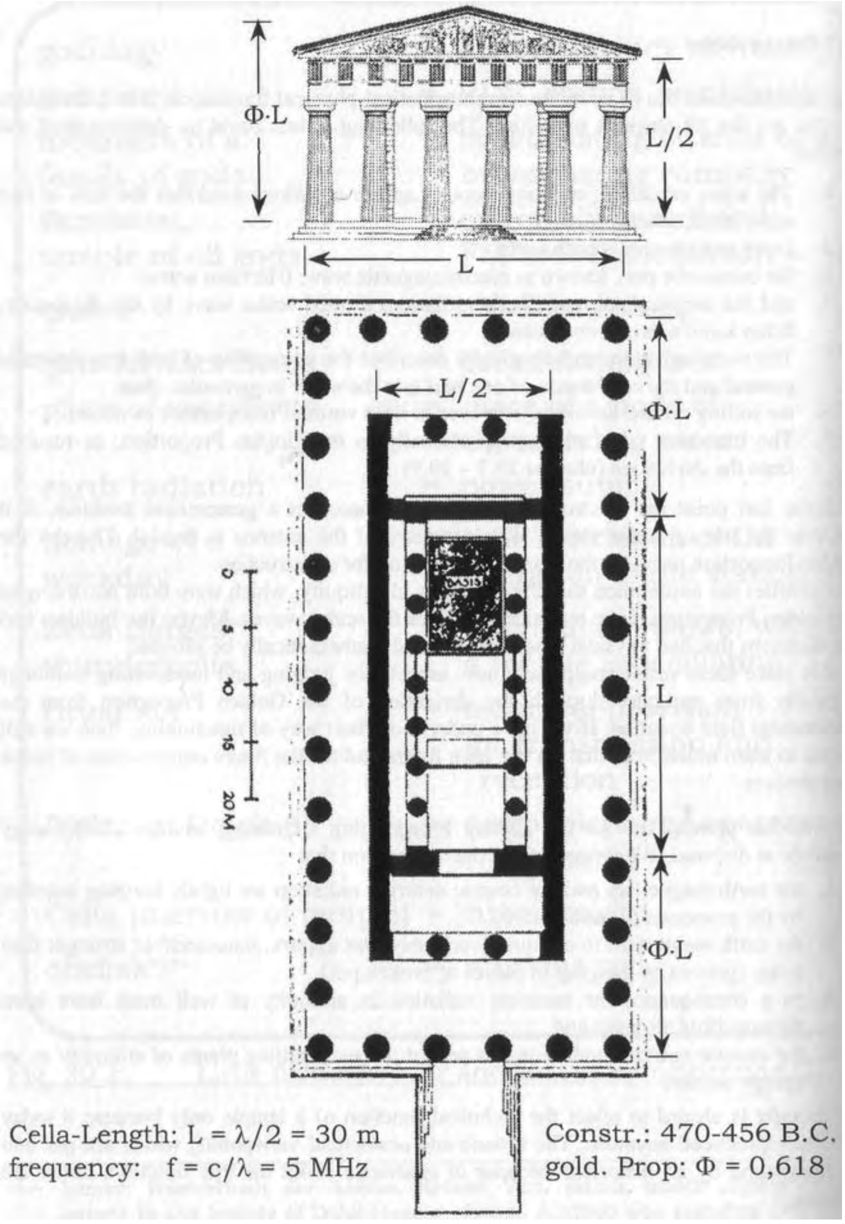
Φ·L
L/2
L
L/2
Φ·L
BASIS
0
5
10
15
20 M
L
Φ·L
Cella-Length: L = λ/2 = 30 m
frequency: f = c/λ = 5 MHz
Constr.: 470-456 B.C.
gold. Prop: Φ = 0,618

Fig. 30.3: The Golden Proportion of Zeus-temple in Olympia.<i>

<i>: K. Schefold: Die Griechen und ihre Nachbarn, Propylaen Kunstgeschichte Berlin Bd. 1, Abbildungen von Seite 249

30.3 Approach

Let's to some extent proceed from the knowledge of textbook physics currently present in high frequency engineering and give a well trained engineer the following task, which he should solve systematically and like an engineer: He should build a transmitter with maximum range at minimum transmitting power, thus a classic task of optimization. Doing so, the material expenditure doesn't play a role!

After mature deliberation the engineer will hit upon it that only one solution exists: He decides on a telegraphy transmitter at the long wave end of the short wave band, at f = 3 MHz, which corresponds to a wavelength of $\lambda = 100$ m. There less than 1 Watt transmitting power is enough for a radio communication once around the earth. That also has something to do with the conditions of reflection of the radio waves at the ionosphere. Our engineer learned:

the index of refraction $n = \sqrt{1 - K \cdot N/f^2}$ (30.1)

with: $K = 80.5 * 10^{-6}$ [cm^3/sec^2] (= constant)

N = electron concentration [$electrons/cm^3$]

f = frequency of the transmitter [MHz]

Put into words: the refraction of a radio wave in the range of the short waves is the larger, the smaller the frequency is. The end of the short wave range is reached at 3 MHz. That thus explains the choice of frequency.

And he optimises further. Next the engineer remembers that at high frequencies, e.g. for microwave radiators, not cables but waveguides are used, since these make possible a considerable better degree of effectiveness. In the case of the waveguide the stray fields are reduced by an alignment and concentration of the fields in the inside of the conductor. In the case of antennas however the fields scatter to the outside and cause considerable stray losses. He draws the conclusion that his transmitter should be built as a tuned cavity and not as an antenna!

As a result the engineer puts a building without windows in the countryside with the enormous dimensions of 50 m length ($=\lambda/2$) and 25 m ($=\lambda/4$) resp. 12.5 m ($=\lambda/8$) width. The height he calculates according to the Golden Proportion to increase the scalar wave part. Those approximately are the dimensions of the Cella without window of Greek temples.

For the operation of such a transmitter in antiquity apparently the noise power of the cosmic radiation was sufficient, which arrived at the earth starting from the sun and the planets. By increasing the floor space also the collected field energy and the transmitting power could be increased, so that also from the perspective of the power supply the temple with the largest possible wavelength at the same time promised the largest transmitting power, so at least in antiquity.

Our engineer further determines, that he will switch the carrier frequency on and off at a predetermined clock pulse. Thus he decides for radiotelegraphy. The advantage of this technique is a maximum increase of the reception range. For that the signals at the transmitter have to be coded and at the receiver again deciphered. By means of the encryption of the contents these are accessible only to the „insiders", who know the code; prerequisite for the emerging of hermetism and eventually a question of power!

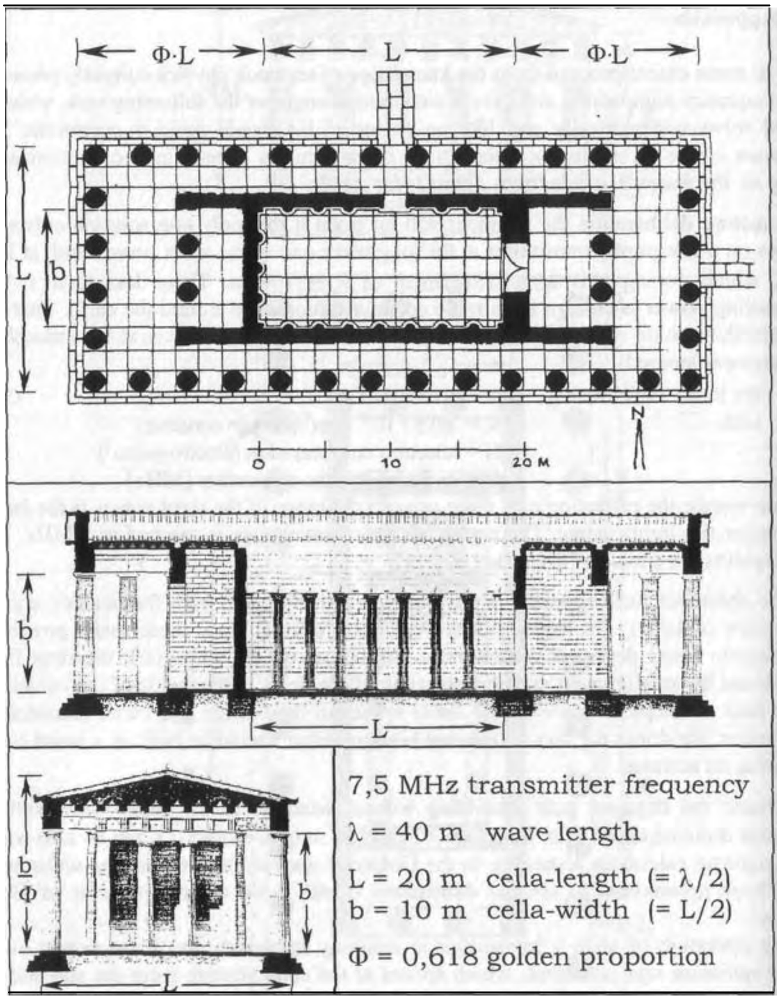
Φ·L
L
Φ·L
L
b
0
10
20 M
N
b
L
b/Φ
b
L
7,5 MHz transmitter frequency
λ = 40 m wave length
L = 20 m cella-length (= λ/2)
b = 10 m cella-width (= L/2)
Φ = 0,618 golden proportion

Fig. 30.4: Example Tegea, temple of Athena Alea.[i]
Built 350/340 B.C.

<i>: G. Gruben: Die Tempel der Griechen, Wissenschaftliche Buchgesellschaft Darmstadt 1986, 4. Aufl. Seite 130

<ii>: E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2.Aufl., S. 89.

<iii>: E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2,Aufl., S. 33.

30.4 Circumstantial evidence

Not everyone, somehow participating in send receive engineering, at the same time also was inaugurated in the entire secret knowledge. Most priests only knew as much as they necessarily needed to know to fulfill their tasks. Thus a temple priest, who was presented an enciphered text and who should bring this on the air, not necessarily at the same time needed to know the content of the text or the code. The same of course also was valid for the sacrificing priest acting in the receiving station. The Vestal virgins for instance had to present the received text to the Augures, by whom they were supervised and controlled.

But who wanted to introduce a new god in the gods heaven and perhaps even himself be worshipped as a god, should have complete command of both the broadcasting technique and the reception technique. In ancient Egypt the Pharao at least once a year had to prove, that he still was in command of the technique. Otherwise he was replaced. For a person with security clearance that at the same time was a death sentence.

In the historical facts numerous pieces of circumstantial evidence can be found, which can be considered to be evidence for the thesis of the operation of send receive engineering in antiquity. One now perhaps understands, why the rulers were put an antenna netting over their head, a so-called crown, or why the Augures could survey the land with a flat Tesla coil in their hands (fig. 16.10).

Direct evidence is present as well. It can be found in ancient texts. But it is questionable if historical texts concerning ancient radio engineering have been translated correctly. The talk is about oracles, mystery cult and earth prophesy if the receiver is meant. The predominantly technically uneducated historians attest the Romans a defective sense of time, because their couriers surely could not cover the long ways across the Roman empire so fast at all, if they read in the Latin texts: "They sent by courier to the emperor in Rome and got for answer...". The answer of the emperor namely already arrived at the squad at the latest in the following night. The correct translation should read: "they cabled" or "they broadcasted to the emperor in Rome and got for answer..."<ii>.

Such a big empire as the Roman Empire actually only could be reined by means of an efficient communication. Cicero coined the word: "We have conquered the peoples of the earth owing to our broadcasting technology...<iii>"! The term broadcasting technology from ignorance is translated with piety. If engineers however rework the incorrect translations, then one will discover that numerous texts tell of the broadcasting technology, that thus correspondingly much direct evidence exists concerning the practical use of this technology.

For the Roman military transmitters, which formed the backbone of the administration of the empire, the reading off of the information from observations of nature like the bird flight or from felt signals of a geomanter was too unreliable. They read off the information from the rhythm of the convulsions of the intestines of freshly slaughtered animals. In the case of the dead animals on the altar every extrinsic influence was excluded. But the enormous need of slaughter cattle was a disadvantage. Who wanted to have information, first of all had to bring along an animal, which then was „sacrificed" the god, or better say, which was abused as a receiver for a particular transmitter. Thereby the innards served as a biosensor and as a receiver for the news.

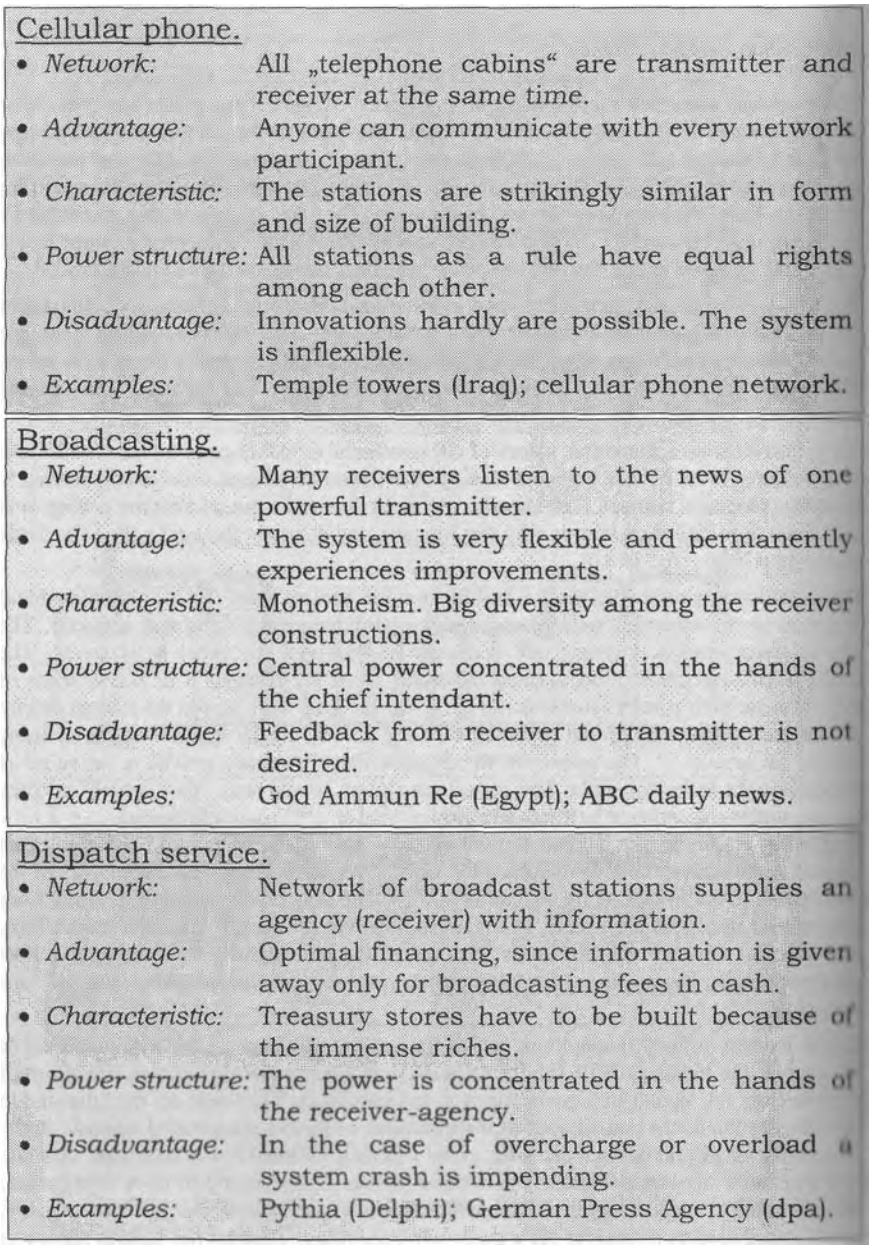

Cellular phone.

• *Network:*	All „telephone cabins“ are transmitter and receiver at the same time.
• *Advantage:*	Anyone can communicate with every network participant.
• *Characteristic:*	The stations are strikingly similar in form and size of building.
• *Power structure:*	All stations as a rule have equal rights among each other.
• *Disadvantage:*	Innovations hardly are possible. The system is inflexible.
• *Examples:*	Temple towers (Iraq); cellular phone network.

Broadcasting.

• *Network:*	Many receivers listen to the news of one powerful transmitter.
• *Advantage:*	The system is very flexible and permanently experiences improvements.
• *Characteristic:*	Monotheism. Big diversity among the receiver constructions.
• *Power structure:*	Central power concentrated in the hands of the chief intendant.
• *Disadvantage:*	Feedback from receiver to transmitter is not desired.
• *Examples:*	God Ammun Re (Egypt); ABC daily news.

Dispatch service.

• *Network:*	Network of broadcast stations supplies an agency (receiver) with information.
• *Advantage:*	Optimal financing, since information is given away only for broadcasting fees in cash.
• *Characteristic:*	Treasury stores have to be built because of the immense riches.
• *Power structure:*	The power is concentrated in the hands of the receiver-agency.
• *Disadvantage:*	In the case of overcharge or overload a system crash is impending.
• *Examples:*	Pythia (Delphi); German Press Agency (dpa).

Fig. 30.5: ____ Three radio technical network structures, with an example from antiquity and from present time.

30.5 Radio technical concepts

In planning and constructing radio technical networks only a few possible concepts exist. It is interesting that at least one historical example can be specified for every concept. That shows that all possibilities were tried at least once. The three most important concepts are presented here:

A. Cellular phone.

Every bigger city between Euphrates and Tigris, which thought the world of itself, had at its disposal already in antiquity a temple tower. Such a temple tower was a „telephone cabin" in the form of a pyramid as a transmitter and a receiver temple at the top, to where the receptionist adjourned to the so-called temple sleep. Discipline was required, since all the time only one priest was allowed to broadcast. All others could listen to him doing so. If he was ready, he closed his contribution with a fixed symbol or term („over") and the next one could continue. This is a classic link-up, where anyone communicates with every network participant.

The stations all were strikingly similar in form and size of building, like one phone box resembles another. In that way a further development of the cellular phone system hardly was possible and that has a technical reason, as the building of a tower in Babylon has shown us. This tower namely had gotten the ambitious builders too big, so that the frequency of the Mesopotamian radio network had been left and instead a foreign network could be received, the code of which no-one could understand. The result was a confusion of language and the order to stop the building.

B. Broadcasting.

Millions of TV spectators every evening look in the ABC news or another daily journal of a TV Channel. In the case of broadcasting thus many receivers listen to the news of a powerful transmitter. With that the whole plenitude of power is concentrated in the hands of the chief intendant. In antiquity he called himself high priest. If he went on the air, he used the logo of the god that he had to represent. Today the logo of the broadcasting company is shown in a corner of the TV screen. Even this very day feedback from the receiver to the transmitter hardly is possible contingent on principle. The problems with nationally controlled broadcasting, with politics controlled by the media all are not new. The monotheism in ancient Egypt with the claim of lordship of the main god Ammun Re is an example from antiquity.

C. Dispatch service.

In ancient Greece the technical structures and with that also the power structures had been turned around. At that time a big network of broadcast stations, which continually was extended by a policy of settlement ordered by the gods, supplied a central and correspondingly powerful agency with information per radio.

Who wanted up to date news, could call for these in the agency with seat in Delphi, but he had to pay for it. To accommodate the broadcasting fees in form of gold and gifts whole treasury stores had to be built. Measured by the commercial success the ancient news network has remained unmatched, and can't be compared with pay-TV or todays dispatch services, like dpa. If the network however becomes too big, uncontrollable and it lacks discipline, then it sometime will crack and the system crashes.

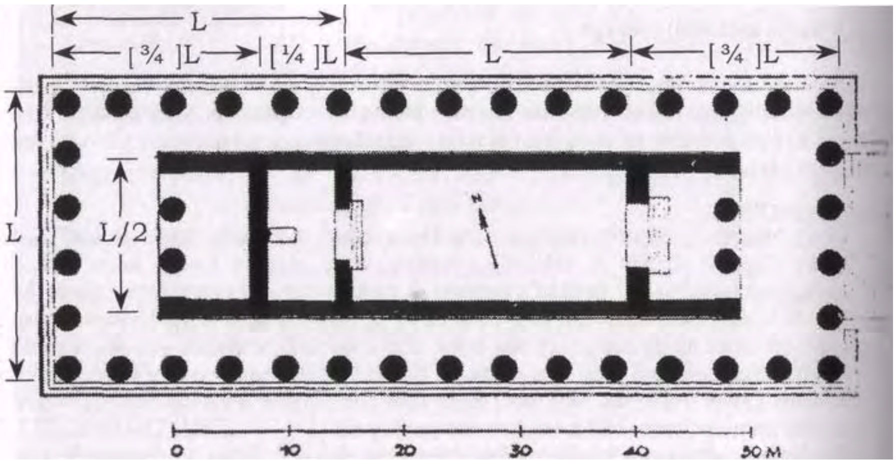
L
[¾]L
[¼]L
L
[¾]L
L
L/2
0
10
20
30
40
50 M

Fig. 30.6 A: Hera-temple of Selinunt 460-450 B.C.<i>
(Corresp. to the plans of the Roman architect Vitruvius)

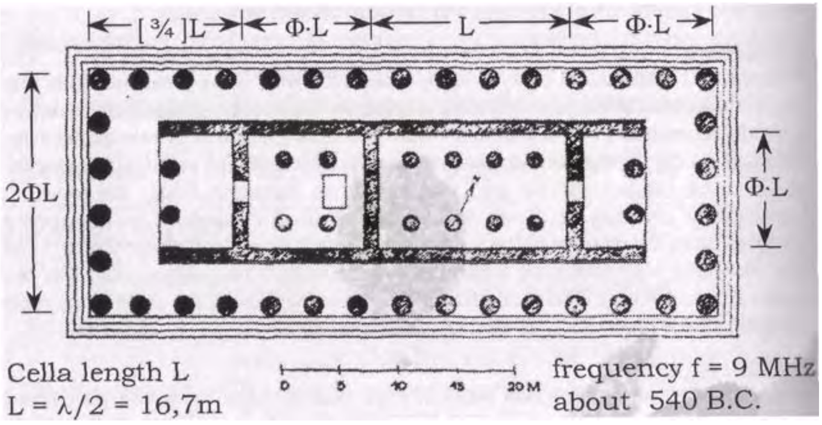
[¾]L
Φ·L
L
Φ·L
2ΦL
Φ·L
Cella length L
L = λ/2 = 16,7m
0
5
10
15
20 M
frequency f = 9 MHz
about 540 B.C.

Fig. 30.6 B: Apollo-temple of Korinth<i>
(Alternative interpretation, use of the Golden Proportion Φ).

<i>: K.Schefold: Die Griechen und ihre Nachbarn, Propylaen Kunstgeschichte Berlin Bd. 1, Abbildungen von Seite 241, 250

<ii>: Vitruvius (Marcus Vitruvius Pollio): Zehn Bucher uber Architektur, Ubers. von Dr. K. Fensterbusch, Wissenschaftl. Buchges. Darmstadt 1987, 4. Aufl., 3. Buch, 1. Kap.: Von den Symmetrien der Tempel, Seite 137

<iii>: Vitruvius (dito), 4.book, 4.Chap.: Vom Tempelinnern und dem Pronaon, p. 187

30.6 Wireless telegraphy

Radio engineering 100 years ago also started with telegraphy. Thereby the high frequency carrier is switched on and off. With this technique Marconi succeeded in a radio transmission over the English Channel (1899) and over the Atlantic Ocean (1901).

As next step the amplitude modulation (AM) followed. Thereby the HF-carrier is overlapped with the low-frequency signal of a sound carrier in such a way, that the amplitude fluctuates on the beat of the LF-signal. As a disadvantageous effect, also noise signals will overlap, from which the quality of reception will suffer.

Only the frequency modulation (FM), where the LF-signal is transmitted as temporal fluctuation of the frequency, brings an improvement. The annoying amplitude noise hence has no effect in the case of FM.

It easily can be recognized, how the development of the modulation techniques follows the urge for technical improvement and optimization. That in antiquity hasn't been different, for which reason the progress of development took place in the same order.

The broadcasting technology of the ancient gods started with the wireless telegraphy. This is expressed in the architecture. Since electric resonant circuits or other frequency determining equipment weren't at the disposal of the engineers in antiquity, the determination and allocation of the broadcasting channels had to take place by means of the wavelength. The formation of a standing wave in the Cella, the innermost sanctuary of a temple, occurs if its length corresponds to half the wavelength of the HF-carrier. The Roman architect Vitruvius calls the wavelength the „basic measure", from which results "the system of the symmetries". He writes: „The design of the temples bases on symmetry, to which laws the architects should adhere meticulously. "[ii]. „The length of the temple is partitioned in such a way that the width is equal to half the length, the Cella itself including the wall, which contains the door, is one fourth longer than wide. The remaining three fourths, which form the Pronaon, should protrude until the antae of the wall and the antae should have the thickness of the pillars"[iii].

If we recalculate ourselves, then the partitioning in 3/4 to 5/4 produces a proportion, which conies quite close to the Golden Proportion. In building a temple nothing is left to chance, after all it concerns the construction of a tuned cavity, capable of self-resonant oscillations with favourable emission behaviour.

From the outside one can't see if a telegraphy transmitter has been changed over to speech transmission with AM. The HF-carrier merely isn't switched off anymore, i.e. the priests let the temple oscillate without interruption. Newly added for AM is an electroacoustic coupling. For that many temples were retrofitted with a mouthpiece. Newly built AM transmitter temples conclude the Cella with a round apse. Because of this acoustically conditioned construction the Cella length didn't have a fixed value anymore and the transmission frequency had become variable. Measured in the middle of the apse the wavelength was larger than at the sides, so that on the beat of the spoken word not only the amplitude of the field distribution in the interior of the temple, but in addition also the frequency of the selfresonant oscillation was changed.

A typical example of such an architectonic hybrid form of AM and FM is situated in Rome. Because due to the frequency variation more than only one wave band was occupied and the temple consistently carries the names of two deities. It is the temple of Venus and Roma.

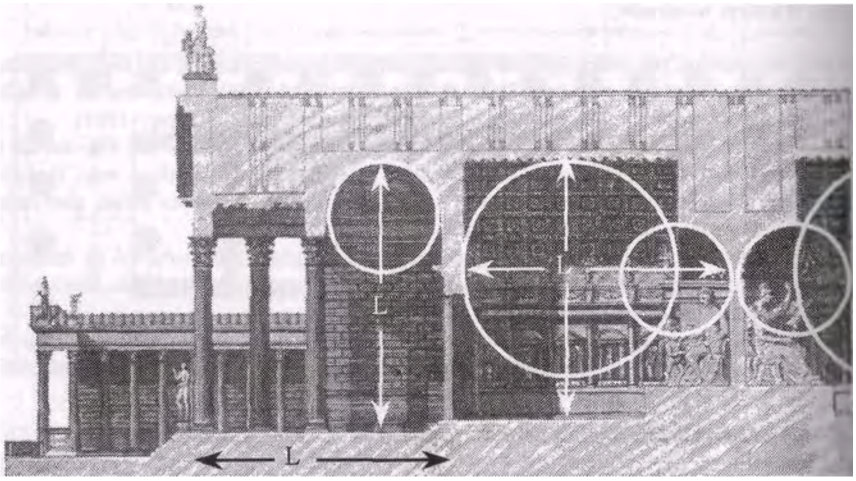

Diameter bigger circle D_1= 22 m (6,8 MHz),
small circle D_2 = 11 m; und L = D_1 + 1/2 D_2 = 27,5 m (5,5 MHz)

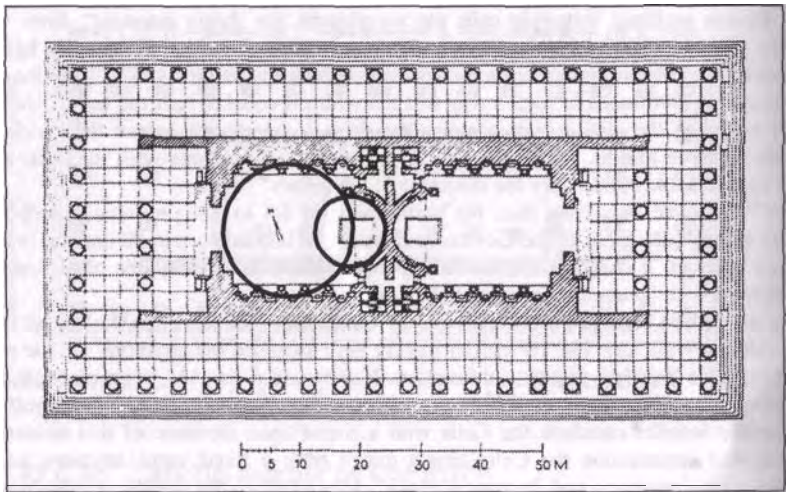

Fig, 30.7: Temple of Venus and Roma; Rome 136/37 A.D.<i,ii>

<i>: A. Springer: Das Altertum, A. Kroner Verl. Leipzig 1915, 10. Aufl., S. 518
<ii>: T. Kraus: Das romische Weltreich, Propylaen Kunstgesch. Berlin Bd. 2, S. 161
<iii>: Lamer: Worterbuch der Antike, Kroner Verl. Bd.96 unter "Tempel", Haufigkeit

30.7 AM temple technology

The low-frequency signal (LF), which should be transmitted by a transmitter with amplitude modulation, lies in the range between 16 Hz and 16 kHz. If it only concerns the transmission of speech information, then the bandwidth can be reduced to 300 to 3000 Hz. In the case of mixing the low-frequency useful signal with the HF-carrier, thus in the case of the modulation of the carrier in the rhythm of the LF, two side bands arise. These lie close to the carrier frequency and are formed from this once by the addition and once by the subtraction with the frequency of the LF-signal. Let's take the temple of Venus and Roma with a transmission frequency of 6.8 MHz. If sound of 3 kHz should be transmitted clearly understandable, then the dimensions of the Cella had to be varied for just 8 mm for a corresponding Cella length of $\lambda/2$ = 22 m. As a curiosity the niches in the side walls in the case of this temple however allow a considerably larger bandwidth of more than 10% instead of the necessary 0.04% in the case of AM.

In the case of the Greek originals, the Cella however has smooth walls, from which follows that the temples were designed ideally narrow band. The Greeks apparently operated predominantly telegraphy transmitters, for which the side bands coincide with the carrier.

The argumentation indeed has remained unchanged: The modulator being narrow band and simple to realize speak in favour of the telegraphy being the "original form" of all modulation techniques. Also the rediscovery of the broadcasting technology by Heinrich Hertz succeeded as telegraphy signal. In addition the range is bigger than for any signal modulated with sound frequency.

As the calculation example has shown, also pure AM transmitters work very narrow band, and this is particularly important for low transmission frequencies, if many transmitters want to use the favoured SW band between 3 and 10 MHz at the same time. With AM one thus accommodates the maximum number of broadcasting channels in a particular frequency range, for instance the 80-meter band, without these interfering with each other too much. But that also was badly needed. Conclusions about the everyday life of broadcasting in antiquity by all means are possible because of the enormous number of temple installations, which logically were permanently used. Only in Rome there existed up to 200 temples[iii]!

Who goes in search of broadcasting stations with a modern short wave receiver, for instance in the 80 m band between the countless telegraphy transmitters, fast gets an idea of what had been up in the air already 2000 years ago. No ancient city would build several temples on a single Acropolis, if only one single one could have been used. All temples broadcasted with each time another carrier frequency because of different dimensions. For this reason the temples, which stood side by side, as a rule were dedicated different gods. An acknowledgement, "the air just being free", in addition hardly was possible, because of the often-found spatial distance between the temple installations and the respective oracle. Between the transmitter of the god Apollo in Didyma and the receiver, the oracle of Milet, for example lie approx. 20 kilometres. The only possible conclusion is that in antiquity there was broadcasted on all channels simultaneously regardless of other gods and their transmission frequencies. As is well-known there rather prevailed a situation of competition between the gods, since like today a large number of listeners meant great importance, influence and power and eventually also worship, more gifts and more receipts from broadcasting fees.

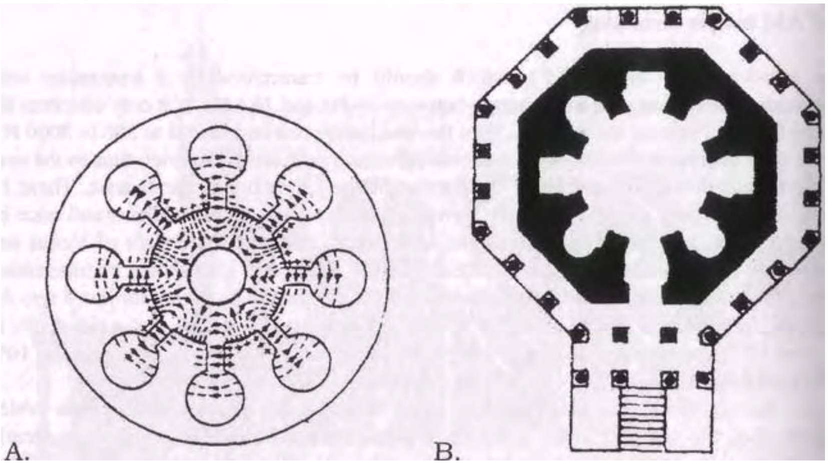

Fig. 30.8: Comparison of a magnetron (A), a microwave radio tube[i] and the temple (B) in the palace of the emperor Diokletian, Split.[ii]

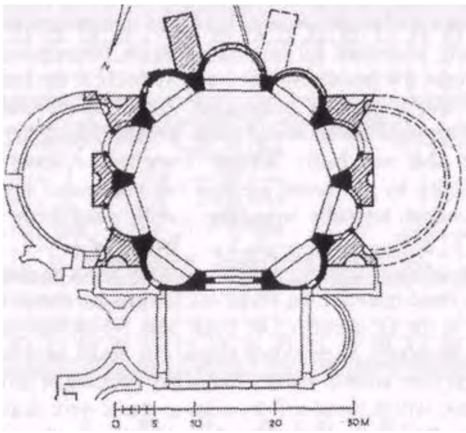

Fig. 30.8.C: Temple of Minerva Medica, Rome, 320 A.D.[ii]

<i>: K. Simonyi: Physikalische Elektronik, 8.4 Das Magnetron, S. 665

<ii>: T. Kraus: Das rom. Weltreich, Propylaen Kunstgesch. Bd. 2, S. 194, 196

<iii>: If one however wants to verify this, in antiquity already common manner of PM broadcasting technology, then we need a broadband short wave receiver with phase-demodulator. With such a receiver even today any time a conclusive argumentation should be possible that this sort of SW-PM technology actually works.

30.8 Modulation of acoustic signals

In the case of sound it in essence concerns longitudinal waves, which propagate in all directions in space with the velocity of sound. But this merely is an average velocity, since the air molecules strictly speaking oscillate with the sound frequency in the direction of propagation. In this way the velocity of sound one time is increased and the next moment to the corresponding extent reduced. If the molecules already carry out oscillations of themselves, e.g. thermal motion, then both oscillations overlap; i.e. the motion of itself is modulated with the sound frequency.

Following the here presented derivation potential vortices are formed in every dielectric, thus also in air, and these are modulated if overlapped with sound waves. Vortices however do not form a distinct frequency, but entirely on the contrary a uniform frequency mixture in the form of white noise.

The overlap thus also is noise unless certain noise frequencies are favoured. This can be effected by means of a spatial body tuned to a certain wavelength. To be considered are cuboid cavities, as in the case of Greek temples or waveguides and cylindrical objects as in the case of round temples or magnetrons. The building form causes the favouring of a certain frequency and the integer harmonic frequencies belonging to it.

If this frequency now lies in the high frequency range, then it is emitted by the spatial body as an electromagnetic wave, in the case of waveguides and resonant circuits for reason of the small dimensions as microwave radiation and in the case of the Greek temples as short wave radiation. If one in addition produces an acoustic signal in the frequency determining spatial body, then this signal automatically will modulate the high-frequency signal.

The result of the modulation is the overlapping of the sound wave with the high-frequency carrier wave. The change of the active length, thus the wavelength of the HF-carrier causes a change of frequency. Such a change on the beat of the sound frequency is called FM (frequency modulation). For that we imagine a spatial body being excited because of its length by potential vortices to a high-frequency oscillation at the self-resonant frequency.

From flow dynamics is known, how easy the distribution of vortices in space can be disturbed. Already words spoken in space are able to influence the potential vortices, which in the case of the temples were used as energy carrier. The longitudinal sound waves reflected at and thus returning from the Cella wall rigid for sound of a Greek temple will push these potential vortices back on the beat of the sound and with that shorten the active length. This means, that the carrier frequency is modulated with the sound frequency. The carrier oscillation thus permanently changes its phase on the beat of the sound signal, for which reason this particular kind of modulation is described as PM, as phase modulation.

Nowadays PM is used only seldom and in the SW-range not at all because of the big need of wave bandwidth. Merely at higher frequencies PM occasionally is used in radiotelephony. In the case of the frequency modulation usual in broadcasting (e.g. UHF), the change of frequency takes place on the beat of the sound amplitude and not of the sound frequency as with PM. Therefore it is not possible to receive phase modulated signals, which are produced by means of the acoustic coupling of appropriately formed spatial bodies<iii> with commercial FM receivers.

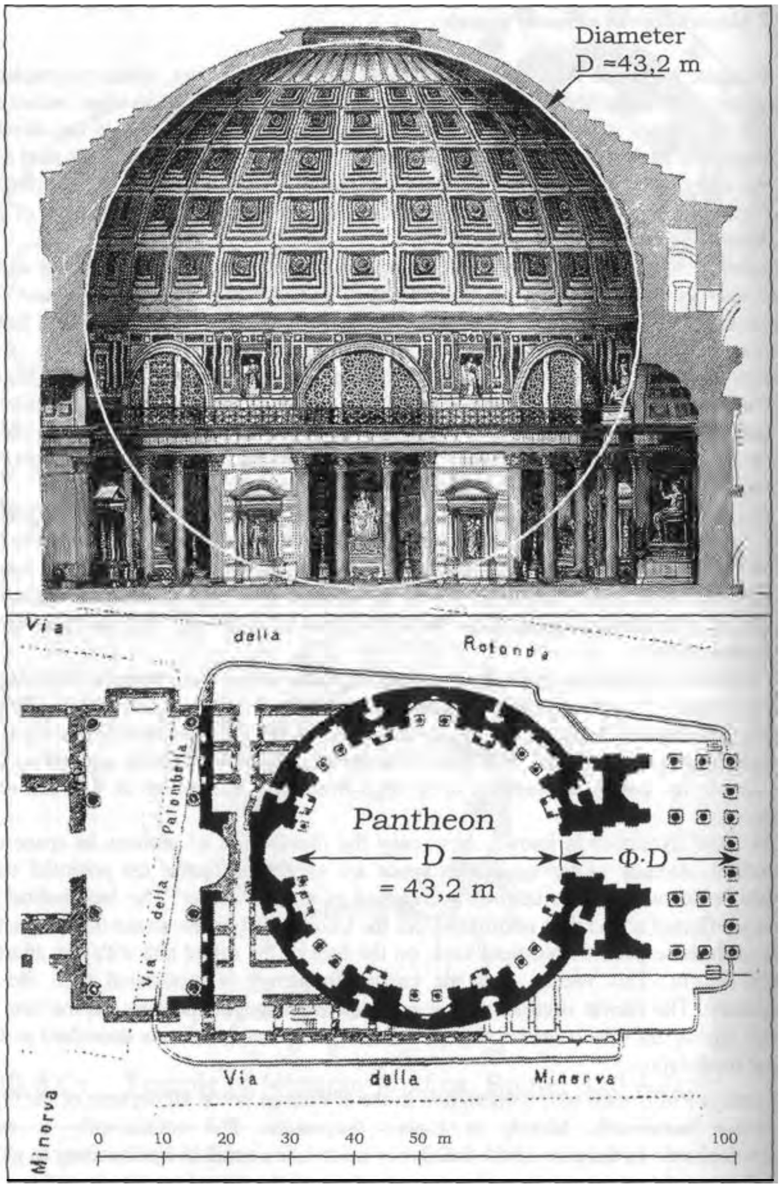
Diameter
D ≈43,2 m
Via della Rotonda
Via della Palombella
Pantheon
D
= 43,2 m
Φ·D
Via della Minerva
Minerva
0
10
20
30
40
50 m
100

Fig. 30.9: The Pantheon in Rome, the „temple of all gods".<i>
Diameter D = 43,2 m, Golden Proportion: Φ= 0.618
Pronaon-(atrium-)length: Φ·D= 0.618 • 43.2 = 26.7 m
Built under emperor Hadrian 118/ 119-125/128 A.D.

30.9 Broadband FM broadcasting technology

If I speak against a flat wall, then every point on the wall has another distance to my mouth. The sound waves thus aren't reflected simultaneously, what leads to big modulation distortions. Therefore the sound wall should be curved in such a way, that all signal paths are the same length (barrel vault, apse, etc.). In the case of point sound sources there results as an optimum a hemisphere, for instance a dome. The building hence even today tells us, which frequency and which modulation method had been put to use. The architecture of sacral buildings, e.g. pointed arch or round arch, thus hardly has been a question of aesthetics.

For his temple of Venus and Roma designed by himself emperor Hadrian had to listen to severe criticism among others of Apollodor of Damascus. The temple was too broadband for an AM transmitter, however with a modulation depth of just 11 percent not broadband enough for a phase modulated FM transmitter.

Emperor Hadrian however also had the courage to build midst in Rome a temple calculated completely new and designed as a pure FM transmitter, the Pantheon, which means temple of all gods. In the language of the technician it is a transmitter for all frequencies.

This domed structure indeed doesn't leave out one single frequency. With a modulation depth of almost 100 percent it is designed for maximum loudness. With that the Pantheon uses all available frequencies, for which reason the name temple of all gods really is no exaggeration. Into the Pantheon exactly fits a sphere with a diameter of 43.2 meter. That corresponds to a minimum frequency of 3.47 MHz, situated in the range of the short waves. The floor however is not domed, but horizontal. That, up to the basis of the dome, results in exactly half the height and a maximum frequency of 6.94 MHz.

The construction ensures that between the simple and the double diameter any desired wavelength can be produced. Above the given maximum frequency of the basic oscillation the harmonic waves, which are produced as well and can't be avoided at all, are attached without a break. These occupy the wave bands up to the double, triple, quadruple frequency and so forth.

For this and only for this reason a maximum frequency was chosen, which corresponds to exactly the double value of the minimum frequency. The operation takes place to the limit, where the transmitter would interfere with itself, in the way that the used basic oscillation would overlap its own harmonic waves. That then sounds like two people talking at the same time. The voices would be distorted out of recognition, as can't be expected else in the range of the harmonic waves.

The Pantheon has been planned and built as a phase modulated basic wave transmitter according to purely academic rules of Hadrian. The temple impressively demonstrates the precise engineering detailed knowledge of the Pontifex Maximus and his broadcasting priests in ancient Rome<ii>.

<i>: A. Springer: Die Kunst des Altertums, A. Kroner Verl. Leipzig 1915, 10. Aufl., S. 517/518, daraus: die Zeichnungen ram Pantheon in Rom, 115-125 n.Chr.

<ii>: The collection of material concerning scalar wave technology here is aborted and continued in an own book in narrative and with that easier to read writing style. The title is: "Broadcasting Gods".

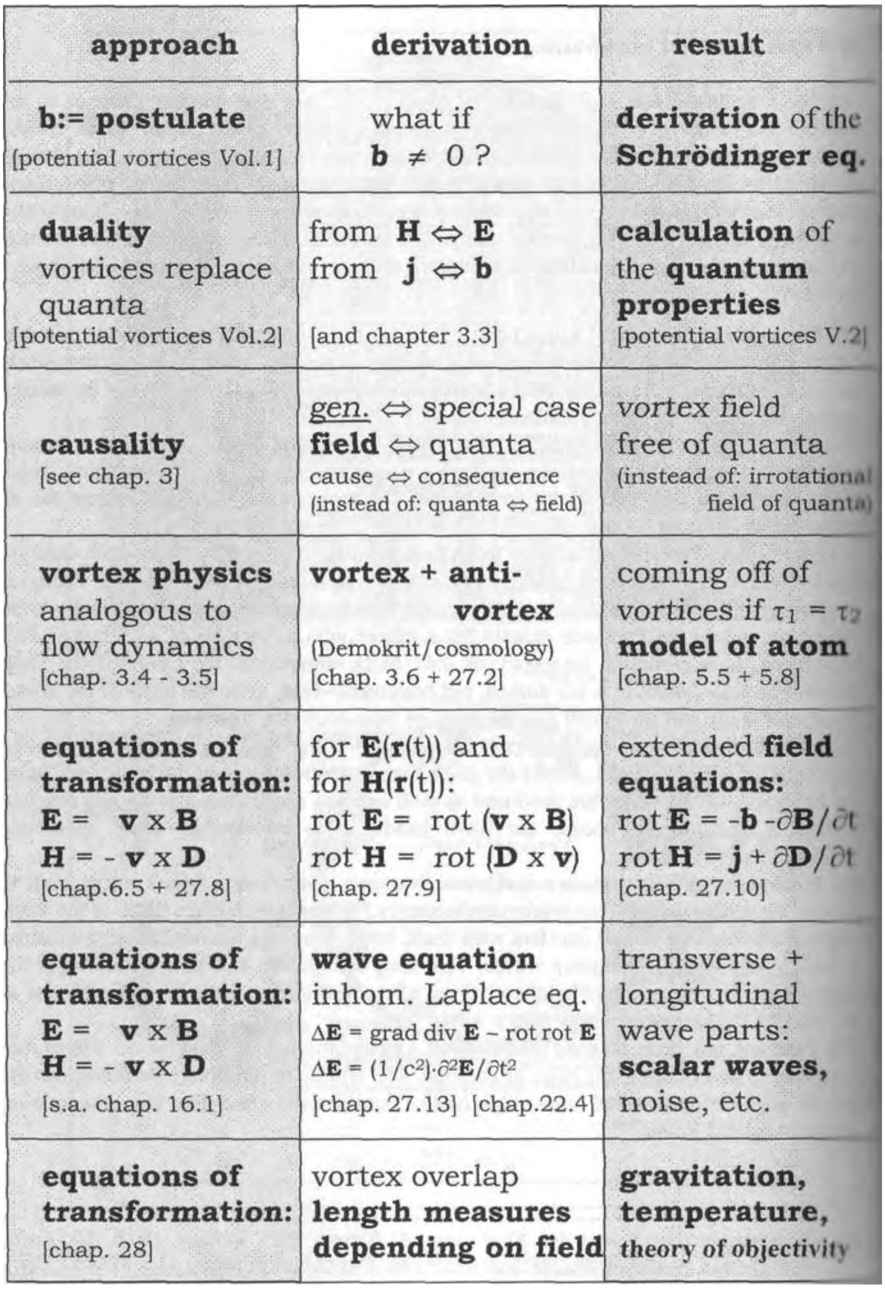

approach	derivation	result
b:= postulate [potential vortices Vol.1]	what if **b** $\neq 0$?	**derivation** of the **Schrödinger eq.**
duality vortices replace quanta [potential vortices Vol.2]	from **H** ⇔ **E** from **j** ⇔ **b** [and chapter 3.3]	**calculation** of the **quantum properties** [potential vortices V.2]
causality [see chap. 3]	*gen.* ⇔ *special case* **field** ⇔ quanta cause ⇔ consequence (instead of: quanta ⇔ field)	*vortex field* *free of quanta* (instead of: irrotational field of quanta)
vortex physics analogous to flow dynamics [chap. 3.4 - 3.5]	**vortex + anti-vortex** (Demokrit/cosmology) [chap. 3.6 + 27.2]	coming off of vortices if $\tau_1 = \tau_2$ **model of atom** [chap. 5.5 + 5.8]
equations of transformation: $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ $\mathbf{H} = -\mathbf{v} \times \mathbf{D}$ [chap.6.5 + 27.8]	for $\mathbf{E}(\mathbf{r}(t))$ and for $\mathbf{H}(\mathbf{r}(t))$: $\mathrm{rot}\,\mathbf{E} = \mathrm{rot}\,(\mathbf{v} \times \mathbf{B})$ $\mathrm{rot}\,\mathbf{H} = \mathrm{rot}\,(\mathbf{D} \times \mathbf{v})$ [chap. 27.9]	extended **field equations:** $\mathrm{rot}\,\mathbf{E} = -\mathbf{b} - \partial\mathbf{B}/\partial t$ $\mathrm{rot}\,\mathbf{H} = \mathbf{j} + \partial\mathbf{D}/\partial t$ [chap. 27.10]
equations of transformation: $\mathbf{E} = \mathbf{v} \times \mathbf{B}$ $\mathbf{H} = -\mathbf{v} \times \mathbf{D}$ [s.a. chap. 16.1]	**wave equation** inhom. Laplace eq. $\Delta\mathbf{E} = \mathrm{grad\,div}\,\mathbf{E} - \mathrm{rot\,rot}\,\mathbf{E}$ $\Delta\mathbf{E} = (1/c^2)\cdot\partial^2\mathbf{E}/\partial t^2$ [chap. 27.13] [chap.22.4]	transverse + longitudinal wave parts: **scalar waves,** noise, etc.
equations of transformation: [chap. 28]	vortex overlap **length measures depending on field**	**gravitation,** **temperature,** **theory of objectivity**

Fig. 30.10: (All) Many roads lead to Rome.

30.10 Epilogue

The preparation for a seminar or a lecture always starts with a collection of material. This should be considerably more detailed than the material to be communicated, since it doesn't get well at the students and also other participants, if the state of knowledge of the lecturer already is exhausted with the scope of his lecture. For this reason the collection of material must also include alternative derivations and areas of knowledge, which thematically rather are marginal, about which mustn't be reported, which possibly not even can bear criticism but still are in the public discussion. For a collection of material, which wants to be considered to be comprehensive and complete, it is important that no area and no theme has been overlooked.

The here presented collection of material with its 650 pages has become correspondingly voluminous. After all the material has been collected over an 8 year period and has been strung into the book in the order of working. That of course complicates reading the book, because individual aspects are repeated several times, but often also in a different context and each time lighted from another side. If the reader somewhere has the feeling, he only has turned in a circle, then has deceived himself. He indeed moves spirally in a circle, like in real life, but he doesn't come out there, where he started. After one turn he is richer with the experience of this spiral turn. In whole science the advancement takes place as a spiral movement and one can count oneself fortunate, as long as the spiral has an ascending slope! This notion should solace the reader, who has undertaken the torture to work through the complete collection of material.

Students also have reported, they had devoured my book like a thriller and a colleague, who had acquired it at a conference in Switzerland, was digging so much in the lecture that he forgot to get off the train timely.

It is a special concern to give reasons for the necessity of an extension of available field theory. To achieve this goal several derivations (fig. 30.10) can be found in my books: from a postulate, from causality, from duality, from vortex physics, from the equations of transformation, etc. added are at least a dozen derivations of other authors from various publications, who at most are cited. With that the goal is pursued that all approaches, which are conceivable and worth discussing, can be put side by side and tested for their efficiency.

Since it isn't the task of a collection of material to answer this question, this must be done by the hearer resp. the reader of the books. He is prompted to find the answer himself! That leads to an intense contention with the theme and that exactly is the reason for the otherwise rather unusual step to make a collection of material open to the public. Objections and criticism of the content of a relation of matters or also only of the representation of the context is wished explicitly. This also isn't valued as criticism of the author or of the superordinated set of difficulties, which in principle isn't possible at all for a collection of material.

Honorary functions of

Prof. Dr.-Ing. Konstantin Meyl

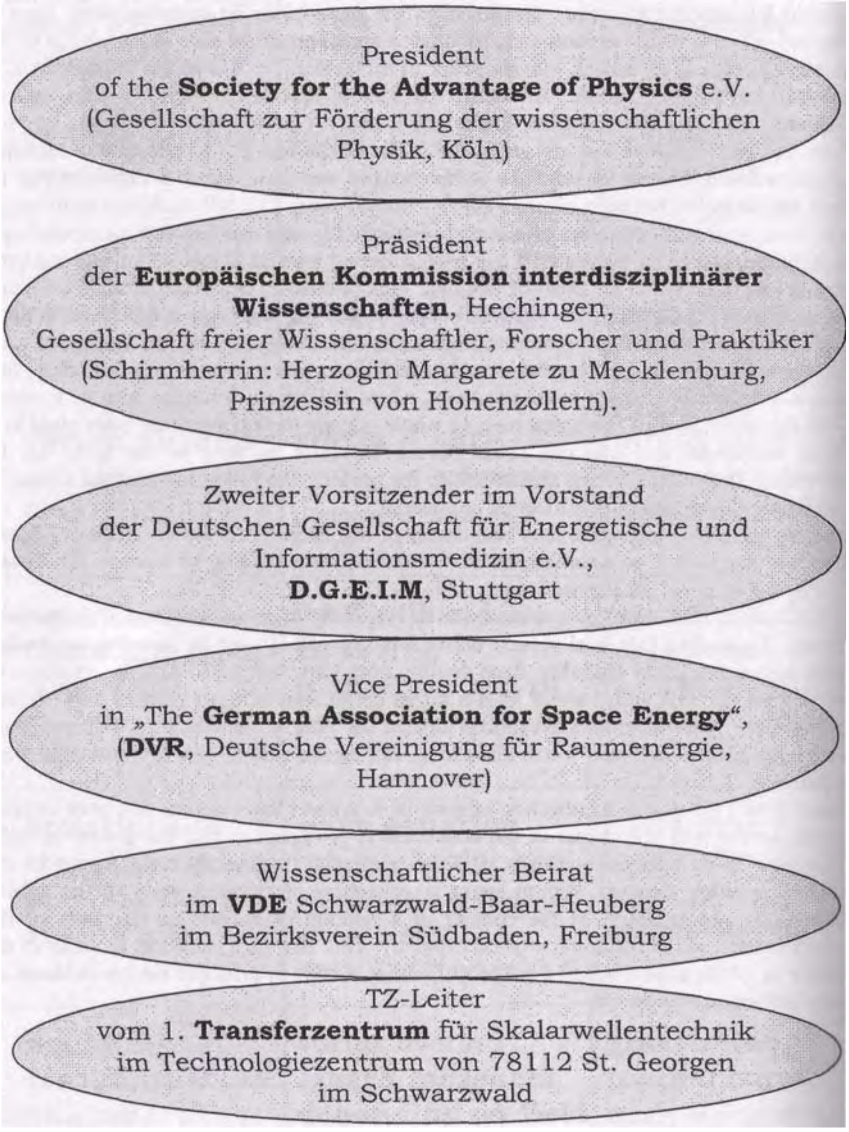

President
of the **Society for the Advantage of Physics** e.V.
(Gesellschaft zur Förderung der wissenschaftlichen Physik, Köln)

Präsident
der **Europäischen Kommission interdisziplinärer Wissenschaften**, Hechingen,
Gesellschaft freier Wissenschaftler, Forscher und Praktiker
(Schirmherrin: Herzogin Margarete zu Mecklenburg, Prinzessin von Hohenzollern).

Zweiter Vorsitzender im Vorstand
der Deutschen Gesellschaft für Energetische und Informationsmedizin e.V.,
D.G.E.I.M, Stuttgart

Vice President
in „The **German Association for Space Energy**",
(**DVR**, Deutsche Vereinigung für Raumenergie, Hannover)

Wissenschaftlicher Beirat
im **VDE** Schwarzwald-Baar-Heuberg
im Bezirksverein Südbaden, Freiburg

TZ-Leiter
vom 1. **Transferzentrum** für Skalarwellentechnik
im Technologiezentrum von 78112 St. Georgen
im Schwarzwald

Fig. 30.11: Honorary functions of the author (December 2003, selection).

There in principle is no necessity at all to discuss with everybody a collection of material, which I have compiled exclusively for own events. It is absolutely sufficient, if few, but then qualified experts have thoughts about the content and communicate them to me. They can feel certain that I don't hand down their judgement to other persons. For that I also have been blamed by ignorant colleagues, who in complete self-overestimation have the opinion, they should be informed about everything. Right is rather, that my private correspondence is of no concern to anybody, since I find it important that as much readers as possible express themselves frank, what only functions, if they can feel certain that they afterwards won't be involved in public mud-wrestling. Therefore I keep still as regards other persons however curious they might be.

Indeed over and over again pseudo scientists turn up, who have the erroneous opinion that scientific arguments would take place on some internet-forums, where one can descant at will and anonymous, hidden behind an alias, where the intellectual firebug can feel safe, not afterwards being blamed for his crimes.

No, science takes place entirely different. A new theory will be able to establish, if it is right and important and if it is used for practical uses. Losers are those authors of a theory, for the elaboratation of which nobody is interested. What does a publication in a journal, however renowned, mean, if nobody reads it and nobody needs it? Most new ideas and approaches go under without notice in today's flood of publications, for who has got the time to read all essays in full?

Desperate they turn to me hoping, at least I could understand their concern. I then invite these scientists to a congress of the „Society for the Advantage of Physics", of which I am the president and offer them a forum, where they are able to present their ideas to an expert public. Not all lectured ideas prove to be sound, but not seldom a physical concern, which should be taken very serious, is behind it.

For a long time the real scientific controversy doesn't take place anymore at the universities and their congresses, where hardly someone dares to lecture arguments against the convention. Too fast he would be expelled as outlaw from the honourable society. From time to time however also from these circles colleagues dare anonymously or privately, as they emphasize, into the alternative events of lobbies or clubs, to astonished find out that real science there still is practised and that there is discussed about ideas, which they have given up thinking or have forbidden themselves to think about. The employment activity, so they excuse their thinking prohibition, allegedly doesn't allow it.

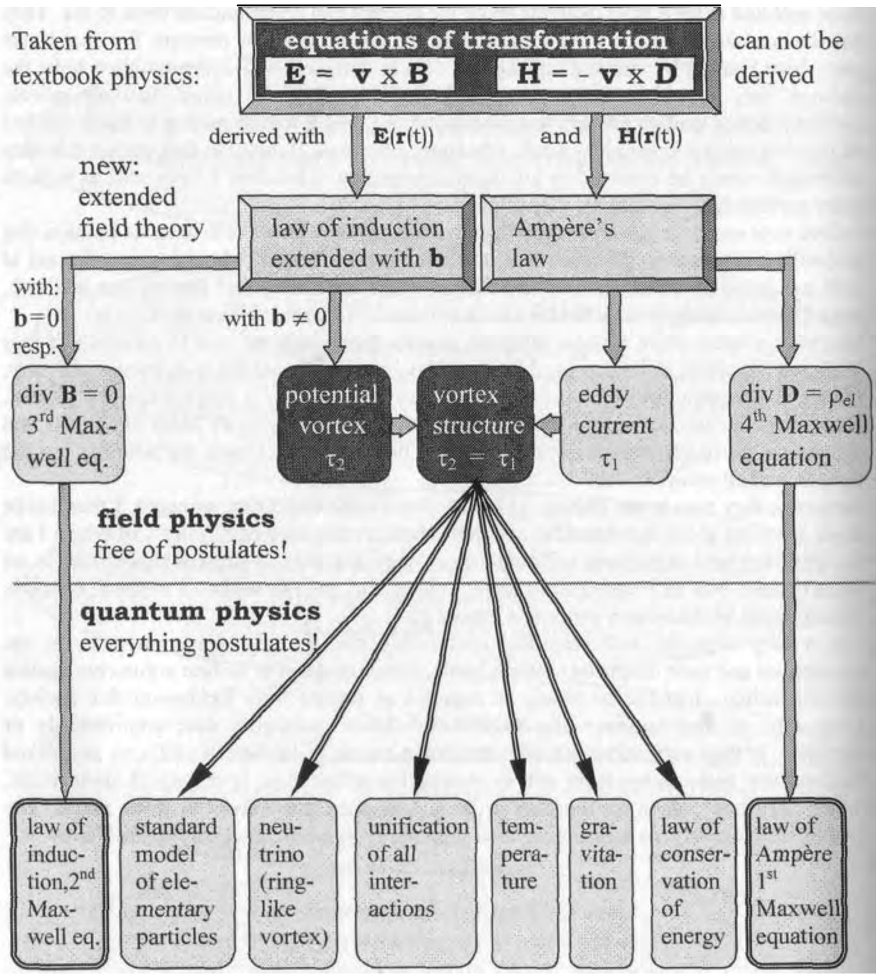
Taken from textbook physics:
equations of transformation
$\mathbf{E} = \mathbf{v} \times \mathbf{B}$
$\mathbf{H} = -\mathbf{v} \times \mathbf{D}$
can not be derived
derived with
$\mathbf{E}(\mathbf{r}(t))$
and
$\mathbf{H}(\mathbf{r}(t))$
new: extended field theory
law of induction extended with **b**
Ampère's law
with: $\mathbf{b}=0$ resp.
with $\mathbf{b} \neq 0$
div $\mathbf{B} = 0$ 3rd Maxwell eq.
potential vortex τ_2
vortex structure $\tau_2 = \tau_1$
eddy current τ_1
div $\mathbf{D} = \rho_{el}$ 4th Maxwell equation
field physics
free of postulates!
quantum physics
everything postulates!
law of induction, 2nd Maxwell eq.
standard model of elementary particles
neutrino (ring-like vortex)
unification of all interactions
temperature
gravitation
law of conservation of energy
law of Ampère 1st Maxwell equation

Fig. 30.12: Derivations of postulates and axioms (part 1).

Be that as it may, the approaches and derivations contained in my collection of material are considered to be controversial, and that is good that way! The public takes notice and professional circles are occupied with the ideas. With that half the way to success already is brought off! It now concerns to knock off all points for their soundness individually, because the next step will be to search a way to the goal as unassailable as possible, which is able to convince also the biggest sceptic. The final version, which then should appear in accredited peer-reviewed journals and in a scientific book concerning the theory of objectivity, goes in this direction. From the numerous approaches in the end only one will be used, and from the countless, in the collection of material listed aspects only the noncontentious ones will remain.

The dispute, which in the current stage can't be avoided, yes even is desired, however shouldn't deceive about the fact that it here doesn't concern persons, improper vanity or some image cultivation, but simply and solely concerns the matter!

In fig. 30.12 and 30.13 is represented, what it concerns. Field physics and quantum physics don't form, as in common practise, an insurmountable opposite, but even complement each other! Below the stripline a little selection of the today in current use quantum physical postulates can be found. The number of newly introduced „constants of nature" and postulates permanently is increasing, a circumstance, which hardly can be mediated to the common sense. The bracket is missing, which interlinks all postulates, or the common source from which they can be derived causally.

In the progress of the three-part edition in this question, essential for physics, already a satisfying and in addition real efficient answer has been found in the domain of field physics. The coupling marked by individual derivations can be found above the stripline and it is entirely new, apart from the dashed indicated derivation (fig. 30.13), as given by Prof. Bosse (TU Darmstadt) in his textbook.

An approach in principle can be chosen freely. In the case of the superordinated field theory two equations of transformation form the approach, which already is laid down in textbooks and secured experimentally. That's why the whole field theoretical derivation manages without one postulate! It is pointed to the fact that these equations on their part can't be derived and should be interpreted rather philosophically than physically. From this approach the extended field theory is derived directly, without a need to add or discard a term. The extended field theory consists of the well-known law of Ampere extended with the dielectric displacement D by Maxwell, and of Faraday's law of induction, which experiences an extension with the vector of potential density b by means of the derivation. Doing so we assume that the field pointers of the electric and the magnetic field strength depend on the spatial coordinate r and through this indirectly also on the time t: E(r(t)), H(r(t)).

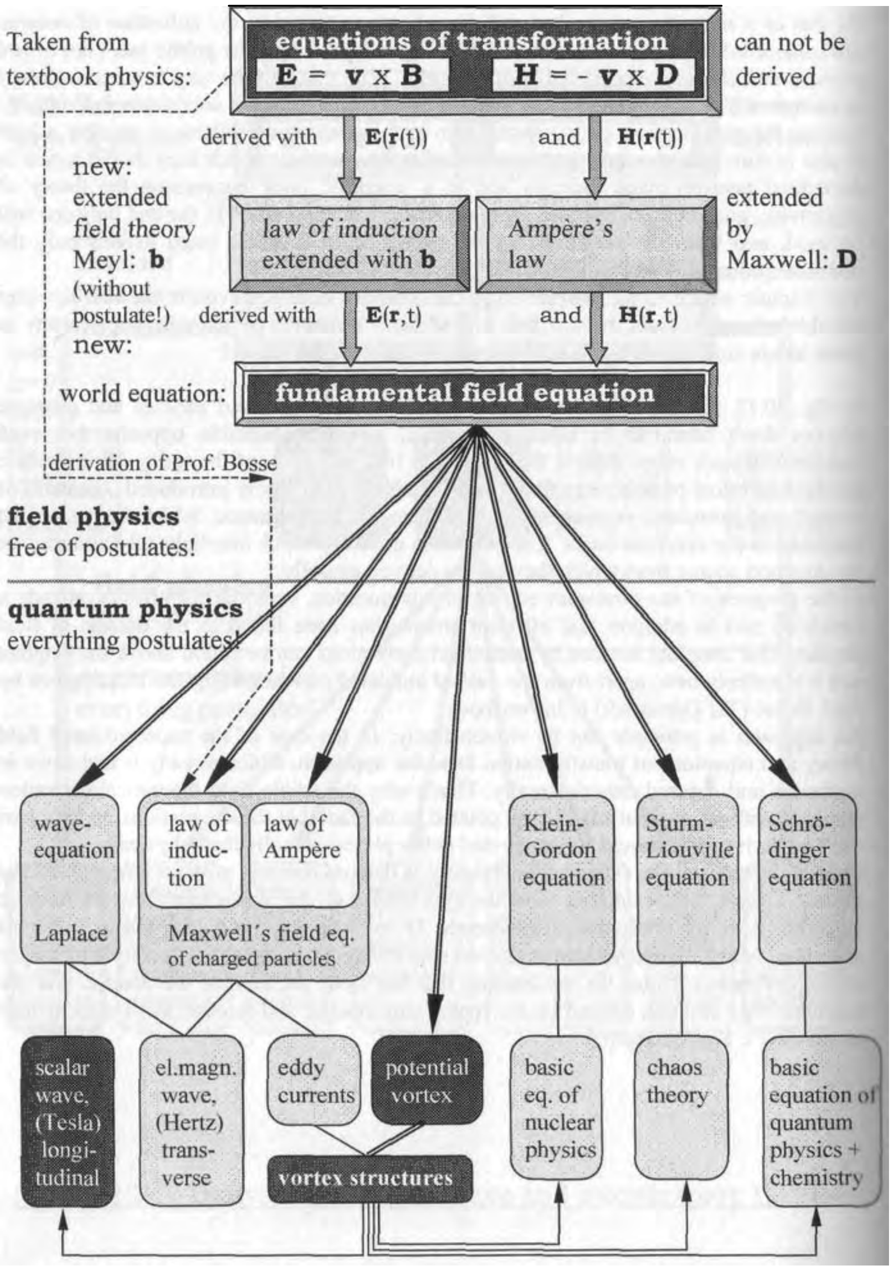
Taken from textbook physics:
equations of transformation
E = v x B
H = - v x D
can not be derived
derived with E(r(t)) and H(r(t))
new: extended field theory Meyl: b (without postulate!)
law of induction extended with b
Ampère's law
extended by Maxwell: D
derived with E(r,t) and H(r,t)
new: world equation:
fundamental field equation
derivation of Prof. Bosse
field physics
free of postulates!
quantum physics
everything postulates!
wave-equation
Laplace
law of induction
law of Ampère
Maxwell's field eq. of charged particles
Klein-Gordon equation
Sturm-Liouville equation
Schrödinger equation
scalar wave, (Tesla) longitudinal
el.magn. wave, (Hertz) transverse
eddy currents
potential vortex
vortex structures
basic eq. of nuclear physics
chaos theory
basic equation of quantum physics + chemistry

Fig. 30.13: Derivations of postulates and axioms (part 2).

If the somewhat more general case is counted up, in which case apart from the spatial a direct temporal dependency is present (E(r,t) and H(r,t)). then further additional terms appear in addition to the extended field equations, which need an explanation also for the case that they are zero. The physical interpretation would implicate a longer treatise, which however can be circumvented, as shown here, by constraining the field pointers, which absolutely is allowed according to the slogan: for the case E(r(t)) and H(r(t)) chosen from many possibilities the extended field equations come out exactly in the form, as they are required and suitable for the further calculations. Who has got to calculate other cases, can do that as he likes, but doing so he should not get lost. Maxwell's field equations are contained in the solution and with that also continue to be valid. Their disadvantage however is that without the extension b not a single quantum physical postulate can be derived. If we add this extension and insert the equations into each other without addition and without cuts also this time a central solution is the result, which is called fundamental field equation.

The derivation is known as well from the Maxwell theory, in which case it is common practise, to use the general approach (E(r,t) and H(r,t)), what we in accordance with the textbooks can do in the same manner. The extension however brings two additional and extremely significant terms. Since the fundamental field equation has eigenvalues under certain boundary conditions and describes structures, various quantum postulates come out from it, from the quantum properties of the elementary particles over the Schrodinger equation and the inhomogeneous Laplace equation up to the derivation of the Golden Proportion. That justifies the assumption that this possibly is the long sought-for world equation!

Even I, as the initiator, was totally surprised by the found derivation of the most important quantum physical postulates and axioms. One just is doing it the right way and already everything fits together! I am not aware of any theory, which would be able to achieve something roughly comparable. The since long sought-for „Theory of Everything", the big unification theory really falls into ones lap. The known interactions are the free and easy result of analysing the field lines of electric and magnetic field strength (fig. 30.12). Physical phenomena, which until now were considered to be incompatible, like e.g. waves, noise or the temperature with the utterly insufficient concepts of the mechanisms for the conversion of one form of energy into another, can be represented consistently with the fundamental field equation as the rolling up of a wavelike field oscillation to a vortex oscillation and as conversion of the noise vortices in the case of a vortex contraction down to atomic dimensions as thermal oscillation, which we treat as vortex losses. There exists no alternative to such unified schemes of things, as makes it possible in abundance the theory that I have founded, considering the two conditions, that on the one hand in the case of the derivation only known regularities are used, by completely doing without postulates and that on the other hand laws are applied and adhered to, also physical laws.

The new schemes of things, which sound unfamiliar, thus already were contained in the laws of physics. After this now having been realized, the tables turn. Now the explanations by postulates, as they at the moment still are being taught, should be replaced by the newly derived ones, if one doesn't want to become a breaker of the law! There doesn't lead a way past the overdue reform of physics anymore.

If you want to correspond with me or if you want to get one of my books from the bibliography, then please consult my Transfer Centre in the Technology Centre of St. Georgen (Black Forest, D).

Address: 1st Transfer Centre for Scalar wave Technology
Prof. Dr. Konstantin Meyl
Leopoldstralsse 1
D-78112 -St. Georgen /Germany
Fax.: +49-7721/51870

email: meyl@k-meyl.de or: meyl@fh-furtwangen. de
Internet: http://www.k-meyl.de or www.meyl.eu

The list of deliverable books (for ordering books over the above given address, Fax, Mail or postcard suffices):

* Wirbelstrome, Diss. University Stuttgart 1984, ISBN 3-9802 542-0-8, 14

* Potentialwirbel Band 1, 1990, ISBN 3-9802 542-1-6 (German), 14

* Potentialwirbel Band 2, 1992, ISBN 3-9802 542-2-4 (not available)

* Elektromagnetische Umweltvertraglichkeit, Teil 1, 2 and 3 (German), ISBN 3-9802 542-8-3, 3-9802 542-9-1 and 3-9802 542-7-5. Each 16

* Scalar wave technology, 2003, documentation and manual to the demonstration-kit and to the experimental-kit (translated and copied).

* Sendetechnik der Gotter, historischer Sciencefictionroman, (in German)
1 .Aufl. 2004, ISBN 3-98O2 542-5-9, 12

* Neutrinopower, Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl, (Discussion in German) Argo-Verlag 2000,

(A bill is enclosed in the delivery.)

Fig. 30.14: Contacting address and list of deliverable books

Table of formula symbols

Electric field

$\mathbf{E}$	V/m	Electric field strength
$\mathbf{D}$	As/m²	Electric displacement
U	V	Tension voltage
ε	As/Vm	Dielectricity: $\varepsilon = \varepsilon_r \cdot \varepsilon_o$
Q	As	Charge
e	As	Elementary charge
τ_2	s	Relaxation time constant of the potential vortices

Magnetic field

$\mathbf{H}$	A/m	Magnetic field strength
$\mathbf{B}$	Vs/m²	Magnetic induction
I	A	Current
μ	Vs/Am	Permeability: $\mu = \mu_r \cdot \mu_o$
ϕ	Vs	Magnetic flux
m	kg	Mass
τ_1	s	Relaxation time constant of the eddy currents: $\tau_1 = \varepsilon/\sigma$

other symbols:

A	m²	Area
a	m	Distance
b	m	Width
c	m/s	Speed of light
c_o	m/s	Speed of light in vacuum
C_e	As/V	Capacity of the electron
C_p	J/K	Heat capacity
d	m	Thickness
E,W	Nm	Energy
f	1/s	Frequency
F	N	Force
G	m³/kg·s²	Gravitational constant
g	m/s²	Gravitational acceleration of the earth
h	m	Height
h	Nms	Planck's quantum of action
$\hbar$	Nms	Quantum of angular momentum: $\hbar = h/2\pi$
$\mathbf{j}$	A/m²	Current density
J	kg·m²	Moment of inertia
$J \cdot \omega^2$	kg·m²/s	Angular momentum
k	Nm/K	Boltzmann constant
l	m	Length
m	kg	Mass
M	kg	Mass of the earth
n,ν =	1,2,3..	Running parameters
N	-	Constant
O	m²	Surface area
p_m	Am²	Magnetic moment

Q	Nm = J	Heat energy
r, R	m	Radius, radius of the earth
r_e	m	Radius of the electron
s	Nms	Spin
t	s	Time, time to orbit
T	K	Temperature
U	Nm	Potential energy
U_e	V	Tension voltage of the electron
v	m/s	Velocity
V	m³	Volume
W	Nm	Energy
w	N/m²	Energy density
W_e	Nm	Energy of the electron
z_e	-	Number of the involved elementary vortices
λ	m	Wave length
ω	s^{-1}	Angular frequency, angular velocity
σ	Vm/A	Specific electric conductivity
ρ	kg/m³	Density $\rho = m/V$
ρ_{el}	As/m³	Electric space charge density
ψ (r,t)		Complex wave function
ϕ (r)		Function of space coordinates
Φ		Golden Proportion

Definitions:

Speed of light	$c = 1/\sqrt{\varepsilon \cdot \mu}$	m/s
Speed of light in a vacuum	$c_o = 1/\sqrt{\varepsilon_o \cdot \mu_o}$	m/s
Moment of inertia (orbit)	$J = m \cdot r^2$	kg·m²
Mom. of i. (homogeneous sphere)	$J = (2/5) m \cdot r^2$	kg·m²
Angular velocity	$\omega = v/r = 2 \cdot \pi / t$	1/s
Surface area of a sphere	$O = 4 \cdot \pi \cdot r^2$	m²
Volume of a sphere	$V = (4/3) \cdot \pi \cdot r^3$	m³

Concerning vector analysis:

Bold print = field pointer (vector); further information in fig. 5.0 in part 1

Prof. Dr.-Ing. Konstantin Meyl:

Scalar wave technology

for the transmission of electric scalar waves

Abstract: 1. Auflage 2000, 2. Auflage and 1st English edition 200

This book is recommended to people, who search the entry into the world of the by the author discovered potential vortices and their propagation as a scalar wave by experimental means. It starts with the instructions to six extraordinary experiments. Doing so an electric radiation is proven, which transmits energy, and that even faster than the light. Also more energy can arrive at the receiver then is put into the transmitter. Who entertains a doubt, will be able to understand the experiments with this book in his hand, to afterwards test the experiments with the gauge, which he is familiar with.

The 1st edition in English at first only includes the instructions for the experiments. In a subsequent edition it will be complemented with a collection of test protocols and progress reports. These are organized into three groups: one group is striving to explain the behaviour of the transmission line conventionally, a legitimate concern, which in a number of points also is able to convince. A second group only is interested in those phenomena of the experiment, which can't be explained conventionally and which prove the existence of scalar waves, whereas the third research group continually strives for new spectacular experiments and practical applications.

Documentation

Belonging to the experimentation and demonstration kit for the transmission of electric scalar waves

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel.: +49-7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald
www.k-meyl.de mevl@k-meyl.de

Bibliography

A. Primary literature (Publications concerning the theory of objectivity):

[1] K. Meyl: Potentialwirbel, Band 1, INDEL Verlagsabt. Villingen 199O, ISBN 3-98O2542-1-6

[2] K. Meyl: Potentialwirbel, Band 2, INDEL Verlagsabt. Villingen 1992, ISBN 3-98O2542-2-4

[3] K. Meyl: Wirbelstrome, INDEL Verlagsabt. Villingen 1984, ISBN 3-98O2542-O-8 / Dreidimensionale nichtlineare Berechnung von Wirbelstromkupplungen, Diss. Uni. Stuttgart 1984

[4] K. Meyl: Einheitliche Theorie bei feldstarkeabhangiger Lichtgeschwindigkeit, Forschungsbericht der Fachhochschule Furtwangen 1992, l.Aufl. Seite 41-43

[5] K. Meyl: Wirbel des elektrischen Feldes, eine neue Storquelle? EMC Journal 1/95,6. J, S. 56

|6] K. Meyl: Gefahrdungspotential und Umwelrvertraglichkeit elektromagnetischer Felder am Beispiel der Funktelefone, Waerland 2/96 (1996), S. 21-23.

[7] K. Meyl: Wirbel des elektrischen Feldes, ein Beitrag zum Thema Uberlichtgeschwindigkeit; Wetter-Boden-Mensch 1/1997, ISSN 0940-5984, Seite: 79-89.

[8] K. Meyl: Physikalische Grundlagen einer Frequenztherapie, W.-B.-M. 6/1997, Seite: 8-15.

[9] K. Meyl: Die Wechselwirkung der Neutrinos, uber Massnahmen, die Halbwertszeit beim radioaktiven Zerfall herabzusetzen; intern.Tagung ,,Neue Energietechnologien aus USA", Zurich am 6.12.97, abgedruckt im NET-Journal (Neue Energietechniken), ISSN 1420-9292, Heft 1/2-1998, S. 14-20

[10] K. Meyl: Objektivitatstheorie, Ein Weg aus der quantenphysikalischen Sackgasse, 1. Beitrag: Der Ather und die Einsteinsche Fehlinterpretation, W.-B.-M. 3/1998, ISSN 0940-5984, Seite 39-44

[11] K. Meyl: Objektivitatstheorie, Ein Weg aus der quantenphysikalischen Sackgasse, 2. Beitrag: Die Vereinigungstheorie, Wetter-Boden-Mensch 4/1998, ISSN 0940-5984, Seite: 31-38

[12] K. Meyl: Longitudinale elektrische Wellen und Wirbel, Vortrag der ,,Internationalen Gesellschaft ftir interdisziplinare Wissenschaften" in Frankfurt am 21.3.1998, veroffentlicht im Vortragsband des "Bundesverbands gegen Elektrosmog e.V." ISBN 3-00-003124-3, Hohenstein, S.79-91

[13] K. Meyl: Biologische Wirksamkeit elektromagnetischer Wellen, Fluch (Elektrosmog) Oder Segen (Frequenztherapie)?, CO'MED, Fachmagazin fur Complementar-Medizin Nr. 10, Dez. 1998, S. 51-55

[14] K. Meyl: Elektrosmog, die physikalischen Grundlagen, Magazin 2000 plus Nr. 134,12/1998, S. 32-37

[15] K. Meyl: Freie Energie und die Wechselwirkung der Neutrinos, W.-B.-M. 3+4/1999, Seite: 21-29

[16] K. Meyl: Skalarwellen und ihre biologische Wirksamkeit, Vortragsband der Deutschen Gesellschaft fur Energetische und Informationsmedizin, Universitat Stuttgart am 17.07.1999, S. 47-52

[17] K. Meyl: Zur Brennglaswirkung des Mondes bei einer Sonnenfinsternis, NET-Journal (Neue Energietechniken), ISSN 1420-9292, Heft 7/8-1999, S. 13-17, sowie INET-Kongress-Vortrag dazu: Freie Energie im nachsten Jahrtausend, am 9.10.1999, Holiday Inn Heidelberg (Walldorf)

[18] K. Meyl: Raum-Energie-Technologie, Wechselwirkung der Neutrinos und energietechnische Nutzung der Teslastrahlung, Teil 1, Magazin 2000 plus Nr. 144, Dezember 1999, S. 62-67

[19] K. Meyl: Raum-Energie-Technologie, Teil 2, Magazin 2000 plus Nr. 145, Jan./Feb. 2000, S. 60-65

[20] K. Meyl: Teslastrahlung, W.-B.-M. 4/2000, ISSN 0940-5984, S. 31-43

[21] K. Meyl: Skalarwellen, Hagia Chora 5/2000, ISSN 0940-5984, S. 31-43

[22] K. Meyl: Teslastrahlung - die drahtlose Ubertragung von Skalarwellen, Kongressband Bregenz vom 15./16.4.2000, S. 181-196, Jupiter-Verlag, ISBN 3-906571-18-1

[23] K. Meyl: Skalarwellentechnik, INDEL Verlagsabt. Villingen-S. 2000, 1. Aufl, ISBN 3-98O2542-6-7

[24] K. Meyl: Nichtoptische Sonnenfinstermsphanomene Teil 1, Magazin 2000 plus/148,6/2000, S. 10-13

[25] K. Meyl: Nichtoptische Sonnenfmsternisphanomene Teil 2, Magazin 2000 plus/150, 8/2000, S. 26-29

[26] K. Meyl: Nichtoptische Sonnenfmsternisphanomene Teil 3, Magazin 2000 plus/151,10/2000, S. 12-17

[27] K. Meyl: Longitudinalwellen-Experiment nach Nikola Tesla, Vortragsband der Deutschen Ges. fur Energetische und Informationsmedizin, Universitat Stuttgart am 6.+.7.10.2000, S. 4.1-4.9

[28] K. Meyl: Skalarwellen und ihre Bedeutung fur die Medizin, Kolloquium vom 6.-8.10.2000 Bad Nauheim der BIT-Arztegesellschaft Vortragsband, Seite 87-96.

[29] Johannes von Buttlar im Gesprach mit Prof. Dr. Konstantin Meyl: Neutrinopower, 1. Auflage 2000, Argo-Verlag Marktoberdorf, ISBN 3-98O6584-8-1

[30] K. Meyl: Skalarwellen-Ubertragung, NET-Journal, ISSN 1420-9292, Heft 12-2000, S. 31-34

[31] K. Meyl: Longitudinalwellen-Experiment, DVR-Info im NET-Journal, Heft 12-2000, S. 21-24

[32] K. Meyl: Erdbeben und Vulkanausbrtiche, GA-Magazin, Heft 6, 2000, S. 11-14

[33] K. Meyl: Atomuhren, die falsch gehen, GA-Magazin, Heft 8, 2001, S. 10-13

[34] K. Meyl: Umpolungsproblematik und Sintflut, GA-Magazin, Heft 10,2001, S. 10-14

The suitcase for experiments from Transfer Centre St. Georgen:

Scalar wave transmission according to Tesla

Bidirectional signal and energy transmission with longitudinal
waves faster than light in a resonant circuit

Get to know the Tesla radiation and its fantastic properties personally. Demonstrate, what no textbook of physics is able to explain!
The experiments, which 100 years ago still were extremely complicated, today fit into an aluminium suitcase, which can be acquired by purchase. Test yourself the historic statements and perform the following experiments:

- transmission of energy and information
- effect back from the receiver on the transmitter
- proof of free energy (1 to 3 times over-unity)
- scalar wave transmission with app. 1.5 times the speed of light
- tunnel effect or the lacking to shield the waves
- disproval of the near-field interpretation.

The kit is offered in two versions:

As a demonstration-kit for 800.- Euro (incl. VAT)
With this the 6 experiments can be carried out without further aids. (Target group are judges, doctors, architects, interested laymen, etc.)

and as an experimentation-kit for 1400.- Euro (incl. VAT) with three different Spulensatzen, Frequenzzahler and zusatzlichem material. (Target group are physicists, engineers, handicraft enthusiasts who like to experiment, etc.)

Weitere Informationen zum Set und zu den Biichern, zu deren Inhalt und zu Vortragsveranstaltungen im Internet unter: http: / /www.k-meyl.de and www.meyl.eu

INDEL GmbH, Verlagsabteilung, Fax: +49-/0- 7721-51870
Tel.: +49-/0- 7724-1770
Postal and ordering address:
l.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald

[35] K. Meyl: HF-technische Hinweise zum Skalarwellen-Ubertragungs-Set, DVR-Mitgliederjournal Nr. 45, 1/2001, S. 20-22 und NET-Journal, ISSN 1420-9292, Heft 3/4-2001, S. 44-45

[36] K. Meyl: Stellungnahme zum Artikel im Spiegel, DVR-Mitgliederjouraal Nr. 46, 2/2001, S. 43-44

[37] K. Meyl: Stellungnahme zu dem Beitrag ,,Skalarwellenubertragung", W.-B.-M. 2/2001, Seite: 77-79

[38] K. Meyl: Scalar Waves: Theory and Experiments, Journal of Scientific Exploration, Vol. 15, No.2, June 2001, ISSN 0892-3310, pp.199-205.

[39] K. Meyl: Skalarwellenstrahlung. Welche Bedeutung haben Skalarwellen fur die Medizin? , CO" MED, Fachmagazin fur Complementar-Medizin Nr.6, Juni/2001, S. 55-60.

[40] K. Meyl: Advanced Conzepts for Wireless Energy Transfer, High efficient Power Engineering with Scalar Waves, International Congress-Publications, Weinfelden, 23/24.6. 01, Jupiter- Verl., pp. 41-49

[41] K. Meyl: Fortschrittliche Konzepte zur drahtlosen Energieubertragung, Hocheffiziente Energietechnologie mittels Skalarwellen, Kongress ,,Neue Wasserstofftechnologien und Raumantriebe" 23.-24. Juni 2001 CH-Weinfelden, Kongressband S. 25-40, Jupiter-Verlag, ISBN 3-906571-20-3

[42] K. Meyl: Schwingungsmedizin, Physikalisch-technische Aspekte komplementarmedizinischer Diagnose- und Therapieverfahren, G.E.I.M.-Vortrag am 6.10.2001 Universitat Kaiserslautern, GZM-Praxis und Wissenschaft, ISSN 1430-4554,7Jahrg. 1/2002, S. 50-55

[43] K. Meyl: Skalarwellen Teil 1: Grundlage einer erweiterten Feldtheorie und prakrische Konsequenzen der Maxwell-Naherung, NET-Journal, ISSN 1420-9292, Heft Juli/August 2002, S. 31-47

[44] K. Meyl: Skalarwellen Teil 2: Weltgleichung und Wellengleichung als Konsequenz der erweiterten Feldtheorie, NET-Journal, ISSN 1420-9292, Heft September/Oktober 2002, S. 44-47

[45] K. Meyl: Wirbelstrukturen als Element der Informationsubertragung in biologischen Systemen, Kolloquium vom 3.-5.10.2002 Bad Nauheim der BIT-Arztegesellschaft, Vortragsband, Seite 77-96.

[46] K. Meyl: Skalarwellen, Grundlage einer erweiterten Feldtheorie, Vortrag am VII.Kongress des IAG vom 18.-20.10.2002 in Mainz-Finthen, Kongressband ISBN 3-9804228-3-6 l.Aufl.2002, S. 118-135.

|47] K. Meyl: Nutzungsmoglichkeiten der Skalarwellentechnik, Neutrino-Power als Chance fur das neue Jahrtausend, Magazin 2000 plus I Nr.177, November/Dezember 2002, S. 60-63.

[48] K. Meyl: Schwingungsmedizin, Bioresonanz durch Skalarwellentechnik, Magazin 2000 plus I Nr. 178, Medizin Spezial 13,2002, S. 52-57

[49] K. Meyl: Die Wellengleichung, Grundlage der Skalarwellen-Physik, Magazin 2000 plus I Nr.179, Januar/Februar 2003, S. 6-12

[50] K. Meyl: Elektrische bzw. magnetische Skalarwellen im Kreuzfeuer der Kritik. Prof. Dr.-Ing. Konstantin Meyl nimmt Stellung zu kritischen Fragen, CO'MED, Nr. 1, Januar/2003, S. 52-54.

[51] K. Meyl: Freie Energie und Neutrinopower, Vortrag am 26.10.2002 in Bensheim auf dem Kongress: (Vom Band abgetippt von I.Schneider) Neue universale Energielosungen, Sammelband der Kongressbeitrage, Jupiter Verlag 2003, ISBN 3-906571-22-X, S.57-79

[52] K. Meyl (D.NeubronnerV. Die Rundfunkgotter, Teslawellen im Altertum, Matrix 3000 Band 16, 7/8-2003, ISSN 14394154, S.40-43

[53] K. Meyl: Erdwachstum durch Neutrinopower, Magazin 2000 plus Nr.188, Extra 3, 9/10-2003, S. 6-13

[54] K. Meyl: Faraday vs. Maxwell, Nikola Tesla Energy Science Conference, Washington DC 8.11.03, IRI

[55] K. Meyl: Physikalische Grundlagen fur die Informationsverarbeitung im Menschen, Kongressband zu den Festspieigesprachen 2002, (Hg. Simma Kletschka) Facultas Verlage Wien 2003, Band 27, Wiener Internationale Akademie fur Ganzheitsmedizin, ISBN 3-85076-648-9, Seite 38-59

by Prof. Dr.-Ing. Konstantin Meyl:

Sendetechnik der Gotter

Konstantin the Great is inaugurated in ancient send receive engineering by his teacher in the Roman emperor palace 304 A.D.

Abstract: 1.Auflage 2004(in German)

- Has god Apollo in Delphi broadcasted at 5.4 MHz?
- Were the Greek temples telegraphy transmitters?
- Were the temple priests amateur radio operators?
- Was Homer radio reporter by order of the gods?
- Were the oracles receiving stations?
- Have oracle interpreters deciphered the transmission code?
- Which bridges did the Pontifex Maximus build?

All are questions, which are dealt with and explained in detail in 30 lessons. In the year 304 A.D. we witness, how the later Roman emperor Konstantin the Great is inaugurated in the secret broadcasting technique of the gods by his teacher. It is an exciting time of upheavel, because the old telegraphy is almost dead. The intestines of animals to sacrifice, from the convulsions of which the radio signals are read off, are scarce goods. Instead radiotelephony should be introduced, which had been tested successfully with the Pantheon in Rome by emperor Hadrian.

But new dispute is initiated: should broadcasting be introduced or rather cellular phone? But those, who tamper around without licence, are chased and fought as always.

When the book is available in English, you will be informed per internet:

www.k-meyl.de www.meyl.eu

INDEL GmbH, Verlagsabteilung, Fax: +49-7721-51870
Tel.: +49-7724-1770
Postal and ordering address:
1.TZS, Prof. Dr. K. Meyl, Leopoldstr. 1, D-78112 St. Georgen/Schwarzwald

B. Secondary literature (Publications about the theory of objectivity):

[1] P. Schellenberg: Buchbesprechung zu K. Meyl: Potentialwirbel, Bd. 1, Efodon News 6 (1992), S. 11

[2] Dr. R. Wohlleben: Buchbespr. zu K. Meyl: Potentialwirbel, Band 2, Frequenz 50 (1996) 1-2, S. 10

[3] Steinbeis-Stiftung fur WirtschaftsfOrderung: Technologiepreis fur STZ-Leiter, Transfer Zeitung 3/94 (1994), S.6

[4] A. Waser (Herausg.): Die ratselhafte Natur - The Puzzling Nature, AW Verlag, Luzern (1996), ISBN3-9521O59-O-2

[6] K. Lorek: Buchvorstellung in "Implosion", Biotechnische Nachrichten Nr. 116 (8/1996), S. 64

[7] Dr. H. Lumpe: Potentialwirbel, neues Weltbild; Raum & ZeitNr. 86, Marz/April 1997, S. 102-103

[8] Inge Schneider: Buchbesprechung zu K. Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 1, NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 3, Marz 1997, Seite 23

[9] W. Martin: Buchbesprechung zu Teil 1, Wetter-Boden-Mensch 2/1997, ISSN 0940-5984, S. 62-63.

[10] RTL: Prof Dr.-Ing.Meyl zu Gast bei der RTL-Talkshow von Ilona Christen am 30.5.97, abgedruckt im NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 6/7 1997, Seite 30-31

[11] Inge Schneider: zu „Prof. Dr.-Ing. Konstantin Meyls Buch Elektromagnetische Umweltvertraglichkeit", NET-Journal, ISSN 1420-9292, Jg. Nr.2, Heft 6/7 1997, Seite 32-33

[12] W. Heidrich: Geht die Expertentagung 1997 in die Geschichte ein? W.-B.-M. 6/1997, S. 3-4

[13] J. Heinzerling: Die Technik des dritten Jahrtausends, Magazin 2000 plus Nr.125, Marz 1998, S. 57/58

[14] Inge Schneider: Buchbesprechung zu K. Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 2, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 3, Marz 1999, Seite 35

[15] W. Martin: Buchbesprechung zu Teil 2, Wetter-Boden-Mensch 2/1999, ISSN 0940-5984, S. 46+59.

[16] P.Rosch: Glanzend, Prof. Meyl zu Elektrosmog, Leserbrief, Magazin 2000 plus Nr.136, 2/1999, S. 77

[17] Portrait der Kongress-Referenten, zu Prof. Meyl: Energie aus dem Weltraum, 5. internationaler Kongress vom 26.-28.2.1999 in Dusseldorf, Magazin 2000 plus Nr. 136, 2/1999, S. 2 + 16 mit:

[18] P.Wiesner: Bericht zum 5. internationalen Kongress „Dialog mit dem Universum", zum Beitrag Prof. Meyl:„fordert eine dezentrale elektr. Energietechnik", Magazin 2000 plus Nr.136, 2/1999, S. 38-41

[19] I. Schlotterbeck: Bucher, die wir empfehlen, Magazin 2000 plus Mr. 142, Sept./Oktober 1999, S. 96

[20] H.-J. Ehlers: Freie Energie, Meyls Neutrino-These; Raum & Zeit Nr. 101, Sept./Oktober 1999, S. 94

[21] I, Schneider: Bericht zum Kongress vom 9.+10.10.99, Heidelberg, zu K. Meyl: Freie Energie im nachsten Jahrtausend, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 10/11, Okt./Nov. 1999, S. 8 ff.

[22] Sabine Krummer: Weltpremiere, Magazin 2000 plus Nr.145, 1+2/2000, S. 59 und im Schwarzwalder Bote vom 19.11.99: TP-Forum eroffnet mit Weltpremiere, Prof. Meyl berichtet iiber Teslawellen.

[23] Diermar Schindler: Neutrinos - Energie aus dem Kosmos, Schwarzwalder Bote Nr. 275 vom 27.11.99

[24] Schneider: Neutrino-Power - Energie aus dem Kosmos, Bericht zu Vortrag und Demo von Prof. Dr. K. Meyl am 25.11.1999 in Villingen, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 12/1999, Seite 4-6

[25] Schneider: Durchbruch in der Freien-Energie-Forschung; Interview mit Prof. Dr.-Ing. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.4, Heft 12, Dezember 1999, Seite 7-9.

[26] Dulling: Professor Meyl und die Freie Energie, TZ-Journal 1999/2000, Seite 5.

[27] Ulrich Arndt: Das Phanomen der Wirbel, Buchbesprechung zu Elektromagnetische Umweltvertraglichkeit Teil 1 und Teil 2, Esotera 1/2000, Verlag Hermann Bauer, S. 50,51

[28] Rainer Borgmann: Energie aus dem Kosmos, Magazin 2000 plus Nr. 145, Januar/Feb. 2000, S. 58-59

[29] Sabine Krummer: Wissenschaftler sprechen von Sensation, Im Villinger Technologie-Forum fuhrte Konstantin Meyl seine Skalarwellen vor, Schwarzwalder Bote vom 7.2.2000

[30] Dr.B.Tausendfreund: Uber biologische und medizinische Auswirkungen von elektrischen Potentialwirbeln und elektromagnetischen Longitudinalwellen„ CO'MED, 2/2000, S. 92

[31] Dr. H.Hantsche: Neues aus Physik und Technik, Info-Blatt der Feien Akademie, Berlin 3/2000, S.3

[32] A.Milinkovic: Weiterer Nachweis der Skalarwellen, NET-Journal, Jg. Nr.5, Heft 3, Marz 2000, S. 26

[33] Vorstellung des Referenten Prof Dr.-Ing. Konstantin Meyl, Kongressband Bregenz vom 15/16.4.2000, S. 180, Jupiter-Verlag, ISBN 3-906571-18-1

[34] Dr. Kunze: Leserbrief zu K.Meyl: Raum-Energie-Technologie, Magazin 2000 plus 147, 4-2000, S. 93

[35] Schneider: Teslastrahlung, die drahtlose Ubertragung von Skalarwellen von Prof. Dr.-Ing. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 4/5- 2000, Seite 14-15.

[36] Dr. Reichwein, Peters: Zellulare Elektromagnetische Systemsteuerung, Der Freie Arzt 5 (2000) im Wissenschafts-Forum (Anhang, S. IV - XXIII)

[37] Ulrich Arndt: Die Energie aus dem Nichts, Esotera 7/2000, Verlag Hermann Bauer, S. 62-66

[38] Besprechung zum Video „EMUV" von Prof. Dr. Meyl, Esotera 7/2000, Verlag Hermann Bauer, S. 30

[39] A.Manthey: Bericht zur Berliner Tagung fur neue Energietechnologien (8.7.2000), zu Theoretische Ansatze (u.a. K. Meyl), NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 7/8 2000, S. 14 ff.

[40] Schneider: Uber das Experimentier-Set zur Ubertragung von Skalarwellen von Prof. Dr.-lng. Konstantin Meyl, NET-Journal, ISSN 1420-9292, Jg. Nr.5, Heft 9, September 2000, Seite 4-7.

[41] Buchbesprechung zu Prof. Dr.-Ing. Konstantin Meyl: Elektromagnetische Umweltvertraglichkeit, Teil 1 und Teil 2, Schweizerische Zeitschrift fur RGS, ISSN 1420-4894, Nr.233, Heft 4-2000, Seite 54-56

[42] Adolf und Inge Schneider: Energie aus dem All, ll.Kap.: Drahtlose Ubertragung von Skalarwellen durch Prof.Dr.-Ing.Konstantin Meyl, S.312-324, Jupiter Verlag 9/2000, l.Aufl. ISBN 3-906571-17-3.

[43] Dr.J.Lechner. Starfelddiagnostik, Medikamenten- und Materialtest Teil2, Kap.2.4.2 Beruhrungslose skalarwellengetragene Informationsubertragung, S. 173-178, Verlag Wtlhr 2000, ISBN 3-927344-44-3

[44] Dr. B. KOhler: Nachlese zum Jubilaumskolloquium der BIT-AG vom 6.-8.10.2000 in Bad Nauheim: Einen weiteren Hohepunkt stellte das Referat von Prof. K. Meyl dar,... COMED, 12/2000, S. 104

[45] Schneider: Buchbesprechung zu ,,Neutrinopower", NET-Journal, Jg. 5, Heft 12, Dezember 2000, S. 39

[46] Fachhochschule Furtwangen: Vorlesungsangebot fur das SS 2001, S.8, Prof. Dr. K. Meyl: EMUV.

[47] Vill: Die elektrophysikalische Mauerwerksentfeuchtung, EV.Verl.2001, ISBN 3-934139-04-3, S.26-33

[48] Kraft: Eroberung von physikalischem Neuland, Buchbesprechung zu ,,Neutrinopower", und

[49] Rattcher: Sind morphische Felder Neutrinos?, Zeit Geist, Heft 3,2001, S. 29 und S. 9.

[50] Dr.R.Kra6nigg: Skalare Wellen, Regulations Medizin 6, Heft 3, 2001, S. 89-96.

[51] Dr.A.Rossaint: Physikalische Grundlagen und biologisch-medizinische Schlussfolgerungen fur den beruhrungslosen Resonanztest Teil 1, Zeitschrift Hier & Jetzt, 3/2001, S.22-27 +Teil2,4/2001, S.22-25

[52] GZM-Kongress Berlin, 13.5.2001, zu dem Vortragenden Prof.Dr.K.Meyl: Skalare-/Teslawellen zur zellularen Energie- und Info. , GZM-Praxis und Wissenschaft, ISSN 1430-4554, 6,Jahrg. 1/2001, S.36

[53] Dr.P.Reichert: Die Medizin der Zukunft - Schwingungen start Pillen, Berlin Nachlese zum 6. Europaischen Kongress fur Ganzheitliche Zahn-Medizin 10.-13.5.2001, GZM-Praxis und Wissenschaft, ISSN 1430-4554, 6.Jahrg. 1/2001, S.56 zuK.Meyl: Schwingungsmedizin und Zellkommunikation.

[54] M.Dworschak: Schwarzwalder Kopernikus, Spiegel-Artikel vom 21.5.01, S. 198-199, abgedruckt im DVR-Mitgliederjournal 2/2001, Seite 41-42

[55] F.Lutz: Bis jetzt unwiderlegt, Wissenschaftler nehmen Professor Meyl in Schutz, Sudkurier Nr. 136 vom 16.6.2001 und Badische Zeitung vom 19.6.2002, Seite 22.

[56] S.Scheurer: Kopemikus untersucht Skalarwellen, Schweizerisches Bodensee Tagblatt Turgau, 21.6.01

[57] Schneider: Introduction of the author Prof.Dr.K.Meyl: Advanced Conzepts for Wireless Energy Transfer, International Congress-Publications, Weinfelden, 23./24.6.2001, Jupiter-Verlag, pp. 23-24.

[58] R.Schmid: Neues aus der Homoopathie-Forschung, zur Tagung der Samuel-Hahnemann-Stiftung vom 6.-8.4.2001 in Simmerath-E., Globuli 11/2001, S.11-12.

[59] Ralph Netzker: Jurassic Quark, Holden, Wegener, Meyl, ein theoretischer Dreisprang, im Internet unter http://home.t-online.de/home/ralph.netzker/quark.htm vom 25.10.01

[60] H.-P. Studer: Neue Energien fur die Zukunft, Vita Sana Magazin fur Dialog und Ges., 9/2001, S. 6-9

[61] Harro Wilke: Leserbrief zu Schwarzwalder Kopernikus, Magazin 2000 plus Nr. 164,10/2001, S. 92

[62] Dr. R.Krafinigg: Skalare Wellen, Arztezeitschrift fur Naturheilverfahren, 43, 1/2002, S. 19-27.

[63] Dr.C.Bollkampen: Bewusstsein im EnergetischenKontinuum, 10/2002, ISBN 3-8311-4016-2.

[64] A. und I. Schneider: Bericht vom Kongress ,,Universale Energielosungen", 26727.10.2002 in Bensheim, zu dem Beitrag von Prof.Dr.K.Meyl: Freie Energie und Neutrinopower, NET-Journal, Jg. 7, Heft 11/12 November/Dezember 2002, S. 21-28.

[65] Dr. P. Rothdach, Dr. H. Boiler u.a.m: Leserbriefe zum Beitrag: Elektromagnetische Skalarwellen im Kreuzfeuer der Kritik. COMED, 1/2003, S. 4-5 und COMED, 11/2002, S. 4: "Facit (Rothdach); Die Meylsche Theorie ist weder mathematisch, noch theoretisch-physikalisch noch experimentell widerlegt"!

[66] D. Schindler: Kopemikus aus dem Schwarzwald, Professor Meyl stellt neue Erkenntnisse der Teslastrahlumg vor, Bericht im Schwarzwalder Bote Nr. 46 VS, vom 25.02.2003

[67] H.-H. Gimbel: Neue Energielo'sungen, Kongressbericht, Wohnung + Gesundheit 3/03 Nr. 106, S. 60

[68] C. Nack: Umstritten aber begehrt. Viele Akademiker kamen Zum Experimentalvortrag von K. Meyl am 14.03.2003 im Technologiepark von Villingen-Schwenningen, Sudkurier 63/TG v. 17.03.03, S.27

[69] G. Hauser: ,.Gebaude der Physik wankt", Meyl sieht seine Theorie durch neueste Berichte bestatigt, Siidkuner Nr. 93/V, 23.04.2003

[70] Prof. Dr. J. Gruber: Grosser experimenteller Erfolg von Prof. Meyl, Energie aus dem Feld nachgewiesen! DVR-Mitglieder-Journal 1/03 Nr. 51 vom 4.4.03, ISSN 1435-0815, S. 101-102

[71] Dr. H. Kersten uber K. Meyl's Forschung: Erdwachstum durch Neutrinoabsorbtion, Technische Universitat Berlin, Internationales Kolloquium: Erdexpansion - eine Theorie auf dem Prufstand, Schloss Theuern am 24.+25.05.2003, Umdruck der TU Berlin Sekr. BH4 Lagerstattenforschung, s.a. Amberger Zeitung vom 28.5.03 und Amberger Nachrichten vom 29.5.03

[72] Dr. P. Rothdach: ,,Skalarwellen gibt es nicht und die Welt ist eine Scheibe", GeoForum 24, 6/2003, Seite 20-22
[73] Inge Schneider: zu Prof. Dr.-Ing. K. Meyls Forschung: Antiker Gotterfunk und futuristische Skalarwellenkommunikation, NET-Journal, Jg. Nr. 8, Heft 7/8 2003, Seite 31 (mit Ankundigungen zu Seminaren von Prof. Meyl in Munchen, an der Universitat Stuttgart und in Washington DC, Seite 42)
[74] J.K.Klasmann: Elektrosmog, Gefahrliche Felder? Gesundheit 9/2003, S.103-104.
[75] Gesundheitsrisiko Handy", Hauptreferent Professor Meyl..., Zollern-Alb-Kurier, 8.10.2003 und K.-H. Schlenker, Leserbrief vom 30.9.2003, s.a. Siidwest Presse: Kommission fordert Absenkung von Grenzwerten, Hohenzollerische Zeitung vom 7.10.2003.
[76] Dr. H.-J. Klose: Die Grundregulation des Menschen, Kongressband zu den Festspielgesprachen 2002, Simma-Kletschka (Herausgeber): Ganzheitliche Zahnheilkunde, Facultas Verlage Wien 2003, Band 27, Wiener Internationale Akademie fur Ganzheitsmedizin, ISBN 3-85076-648-9, Seite 95-98
[77] Dr. Valone, Integrity Research Institute: Nikola Tesla Energy Science Conference, Washington DC November 8th.+9th.2003, Conference publications, speakers biographies and abstracts.

C. Literature concerning the electromagnetic and biological compatibility

[1] R. O. Becker: Heilkraft und Gefahren der Elektrizitat (Cross Currents) 2. Aufl.Scherz Verlag (1994)
[2] P. Brodeur: Mikrowellen (The zapping of America) Augustus Verlag Augsburg (1989)
[3] EMVG: Gesetz uber die EMV von Geraten, Bundesgesetzblatt Teil I vom 9.11.1992
[4] Flanagan: Elixier der Jugendlichkeit, Waldthausen Verlag Ritterhude 1992, 1. Aufl.
[5] E. Habiger u. a.: Elektromagnetische Vertraglichkeit, 2. Aufl. Verlag Technik, Munchen 1992
[6] E. Habiger: EMV, Huthig Buch Verlag Heidelberg (1992), ISBN 3-7785-2O92-X
[7] E. Habiger: EMV-ubergreifende Aspekte zu benachbarten Problemfeldern, Elektrie 48 (1994), H. 5/6, S. 163-167
[8] N. Harthun: Naturgemasse Technik - Wege fur die Forschung nach Viktor Schauberger, Verlag Delta Design, Berlin 1996.
[9] H. L. KOnig: Unsichtbare Umwelt (Wetterfuhligk.), 5. Aufl., Verlag Moos & Partner Munchen
[10] Anton Kohling: Grundlagen der Funkentstorung in der Bundesrepublik Deutschland, etz Bd. 108 (1987), Heft 10.
[11] G. Nimtz: Mikrowellen, Einfiihrung in Theorie und Anwendung. 2. Aufl. Mannheim/Wien/Zurich: BI-Wissenschaftsverlag 199O, ISBN 3-411-O32O3-O
[12] Novitskii, Yu. I. : Effects of a magnetic field on the dry seeds of some cereals. Proceedings of Conference on the Effect of Magnetic Fields on Biological Objects, Moscow, p. 52, 1966.
[13] Plato: Samfliche Werke 5, Rowohlts Klassiker Nr. 47, S. 188,66e
[14] A. S. Presman: Electromagnetic Fields and Life. Plenum Press, New York - London, 197O
[15] V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112
[16] H. Schulz, W. Oppitz: Lokale Hyperthermie durch hochfrequente Wirbelstrome, Medizin Technik 1, 1987.
[17] G. A. Ulmer: Wellen- und Elektrosmog, Ulmer Verlag (1994) ISBN 3-924191-73-5
[18] Vitruvius Pollio, Marcus: Zehn Bucher uber Architektur, WBG 1987
[19] A. D. Watt and E. L. Maxwell: Characteristic of Atmospheric Noise from 1 to 1OO Kc/s. Symposium on the Propagation of VLF Waves, Boulder, Col., Paper 35, Jan. 1957.

D. Literature concerning physics in general (particle physics, etc.)

[1] H. Armbruster, G. Grunberger: Elektromagnetische Wellen im HF-Bereich, Huthig und Pflaum Verl.
[2] W. Bauer: Klassische Physik, Graphia Druck, Salzburg (1975), Eigenverlag.
[3] K. Bethge, U. E. Schroder: Elementarteilchen, Wissenschaftliche Buchgesellschaft Darmstadt (1986)
[4] G. Bosse, Grundlagen der Elektrotechnik II, BI 183, Hochschultaschenbuch-Verlag, Mannheim 1967
[5] Bronstein u.a.: Taschenbuch der Mathematik, 4.Neuaufl. Thun 1999
[6] A. Einstein: Grundztlge der Relativitatstheorie, S. 162, Anhang II; allg.Bemerkungen C, 5. Aufl., Vieweg + Sohn, Braunschweig 1973.
[7] A. P. French: Special Relativity, M.I.T. (1968) / Vieweg Verlag (1971) ISBN 3-528-O3546-3
[8] G. Galeczki, P. Marquardt: Requiem fur die Spezielle Relati vitat, Verlag Haag+Herchen (1997)
[9] U. Gradmann/H. Wolter: Grundlagen der Atomphysik, AVG, Frankfurt a.M. 1971.

[10] Grimsehl: Lehrbuch der Physik, 2.Bd, 17.Aufl. Teubner Verl. 1967

[11] J. D. Jackson, Classical Electrodynamics, 2nd. Ed., John Wiley, New York, 1975.

[12] Kleine Enzyklopadie Atom, Verlag Chemie GmbH, Weinheim 1970

[13] K. Kupfmuller: Einfilhrung in die theoretische Elektrotechnik, Springer-Verlag Berlin, 12. Aufl. 1988

[14] H. G. Kussner: Grundlagen einer einheitlichen Theorie der physikalischen Teilchen und Felder, Musterschmidt-Verlag, Gottingen 1976, ISBN 3-7881-6O35-7

[15] Kuchling: Physik, Gl. At4, VEB-Leipzig, 11. Aufl. 1974

[16] G. Lehner, Elektromagnetische Feldtheorie, Springer-Verlag Berlin, Heidelberg 1990.

[17] H.-J. Lugt: Vortex flow in nature und technology. Krieger publishing company 1995

[18] J. C. Maxwell: A Treatise on Electricity & Magnetism, vol.2, Chapter XX: Electromagnetic Theory of Light. Dover Publ. New York

[19] Meinke, Gundlach: Hochfrequenztechnik, 4.Aufl. Springer-Verlag Berlin 1986

[20] Mende, Simon: Physik, Gl. 10. 39, VEB-Leipzig, 4. Aufl. 1974

[21] G. Mierdel: Elektrophysik, Huthig Verlag Heidelberg (1972) ISBN 3-7785-O243-3

[22] O. Nachtmann: Phanomene und Konzepte der Elementarteilchenphysik, Vteweg, Braunschweig 1986, ISBN 3-528-O8926-1

[23] W. Pauli: Aufsatze und Vortrage uber Physik und Erkenntnistheorie, Vieweg + Sohn, Braunschw. 1961

[24] R.W. Pohl: Einfuhrung in die Physik, Bd.2 Elektrizitatslehre, 21.Aufl. Springer-Verlag 1975

[25] H. Preussker: Der Wirbelring, Halstenbeck 2002

[26] Shu-wen Zhou: Abnormal Physical Phenomena Observed When the Sun, Moon, and Earth are Aligned, 21st Century Science & Technology, Fall 1999, Vol. 12, No. 3, pp. 54-61.

[27] K. Simonyi, Theoretische Elektrotechnik. 7. Aufl. VEB Verlag Berlin 1979. S. 921 - 924.

[28] K. Simonyi: Physikalische Elektronik, Kap.: 8.4 Das Magnetron

[29] K. Stierstadt: Physik der Materie, VCH, Weinheim 1989.

[30] D. F. Taylor, On the Mechanism of Aluminium Corrosion in Metallized Film Capacitors, IEEE Transactions on EI-19, August 1984, No. 4, pp. 288-293.

[31] Jean-Pierre Vigier and the Stochastic Interpretation of Quantum Mechanics, selected and edited Stanley Jeffers, Bo Lehnert, Nils Abramson, and Leve Chebotarev. Montreal: Apeiron, 2000.

[32] A. Yializis, S. W. Cichanowski, D. G. Shaw: Electrode Corrosion in Metallized Polypropylene Capacitors, Proceedings of IEEE, Int. Symposium on Electrical Insulation, Bosten, Mass., June 1980;

[33] Zinke, Brunswig: Lehrbuch der Hochfrequenztechnik, 1. Bd., 3.Aufl. Springer-Verlag Berlin 1986

E. Literature concerning Tesla and criticism of textbook physics

[I] M. Cheney: Tesla, Man out of Time, Barnes & Noble Books, New York, 1993, ISBN O-88O29-419-1 und Omega-Verlag (1996) ISBN 3-93O243-O1-6

[2] D. H. Childress: The Fantastic Inventions of Nikola Tesla, Adventures Unlimited Press (1993)

[3] J. DeMeo: The Orgone Accumulator Handbook, ISBN O-9261855-O-7

[4] G. Galeczki, P. Marquardt: Requiem fur die Spezielle Relativitat, Haag+Herchen (1997),

[5] J. K. Kowalczynski: Critical Comments on the Discussion about Tachyonic Causal Paradoxes and the Concept of Superluminal Reference Frame, Intern. Journal to theoretical Physics, Vol. 23, No. 1, 1984

[6] J. Marinsek: Rationale Physik oder wissenschaftliche Science fiction? dbv-Verlag fur die Technische Universitat Graz (1989) ISBN 3-7O41-O176-1

[7] H. C. Montieth: The Unified Field Theorie, MUFON Symposium Proc, Seguin, Texas, 1980

[8] F. Moser: BewuBtsein in Raum und Zeit - Grundlagen einer holistischen Weltauffassung auf wissenschaftlicher Basis, Leykam Graz, 1989

[9] H. A. Nieper: Revolution in Technik, Medizin, Gesellschaft; MIT Verlag Oldenburg (1982)

[10] O. Oesterle: Ausweg aus der Sackgasse, Ein ganzheitliches naturwissenschaftliches Weltbild, 1. Aufl., Jupiter-Verlag, Bern 1996, ISBN 3-9O6571-16-5

[11] Rho Sigma (Dr. Rolf Schaffranke): Forschung in Fesseln, das Ratsel der Elektro-Gravitation. Ventla-Verlag, Wiesbaden (1972)

[12] N. Tesla: III. The Singular Misconception of the Wireless (Famos Scientific Illusions), Electrical Experimenter, Feb. 1919, p. 732.

[13] Tesla's verschollene Erfindungen, VAP, Wiesbaden (1994) ISBN 3-922367-93-3

[15] N. Tesla: Complete Patents, TeslaBook Company (1983) ISBN O-96O356-8-2, daraus:

[16] N. Tesla: Apparatus for Transmission of Electrical Energy, US-Patent-No.: 649,621, New York 1900, Dr. Nikola Tesla: Complete Patents pp318-321.

[17] Nikola Tesla: Method of and Apparatus for Controlling Mechanism of Moving Vessels or Vehicles, US-Pat. 1898, Nr. 613,809. Complete Patents: S. 351
[18] Nikola Tesla: Famous Scientific Illusions, III. The Singular Misconception of the Wireless, Electrical Experimenter, Feb. 1919, printed in Tesla Said, p. 197
[19] Nikola Tesla: The Problem of Increasing Human Energy, The Century Monthly Magazine, June 1900, ISBN 1-882137-00-0, Page i-15
[20] N. Tesla: Die Ubertragung elektrischer Energie ohne Drahte..., Electrical World and Engineer vom 7.1.1905; Edition TeslaBd. 4, S. 131.
[21] N. Tesla, New York Times vom 15.9.1908; Edition Tesla Bd. 6, S. 241.
[22] Tesla Said (J.T.Ratzlaff), Tesla Book Company, ISBN 0-914119-00-1
[23] N. Tesla: Experiments with Alternate Currents of High Potential and High Frequency, Lindsay Publications Inc, ISBN 0-917914-39-2
[24] W. Theimer: Die Relativitatstheorie, A.Francke AG Verlag Bern (1977) ISBN 3-772O-126O-4

F. Literature concerning geophysics, geology, geography and astronomy

[1]: Kendrick Frazier: Das Sonnensystem, Time-Life Bucher, Amsterdam (1991).
[2] Carey, S. W.: Theories of the Earth and Universe. University Press: 1-413, Stanford, California.
[3] Oesterle, O.: Goldene Mitte: Unser einziger Ausweg, Universal Experten Verlag, Rapperswil (1997)
[4] Hilgenberg, O. C: Vom wachsenden Erdball, Berlin 1933, Eigenverlag. und Neues Jahrb. Geol. Palaont. Abh. 116, 1, Berlin (1962).
[5] Jordan, P.: Die Expansion der Erde. Vieweg 1-180, Braunschweig (1966).
[6] Robert Bauval und Graham Hancock: Das Geheimnis der Pyramiden, Der Schlussel zur Sphinx, List Verlag (1996) und in S3 (Die grossen Ratsel VII) am 25.5.97
[7] Mitton, S. (Herausg.): Cambridge Enzyklopadie der Astronomie (The Cambridge Encyclopaedia of Astronomy), Orbis Verlag (1989).
[8] Miller, R.: Driftende Kontinente, Time-Life Bucher, Amsterdam (1991).
[9] Hermann Wild: Technologien von gestern, Chancen fur morgen, Jupiter-Verlag Bern 1996
[10] Zecharia Sitchin: The Lost Realms, Knaur Verlag Munchen 1992, ISBN 3-426-04827-2
[11] Werner Raffetseder: Sonnenfinsternis, Hugendubel Verlag, Munchen 1999, ISBN 3-89631 -302-9

G. Literature concerning free energy

[1] R. L. Clark: Tesla Scalar Wave Systems, The Earth as a Capacitor, The Fantastic Inventions of Nikola Tesla, ISBN 0-932813-19-4
[2] H.-P. Thietz: Tatort Erde, VAP (1996), ISBN 3-922367-62-3
[3] Louis Kervran: Biological Transmutations, s.a.:
[4] J. Heinzerling: Energie aus dem Nichts, Bettendorf'sche Verlagsanstalt, 1. Aufl. 1996
[5] J. Manning: The Coming Energy Revolution - The Search For Free Energy, Avery Publishing Group, New York, USA, 1996 und Omega-Verlag, Dusseldorf 1997.
[6] G. Hilscher: Energie fur das 3. Jahrtausend, VAP-Verlag Wiesbaden 1996
[7] A. Schneider: Energien aus dem Kosmos, Jupiter-Verlag 1989
[8] I. Schneider: Neue Technologien zur Freien Energie, Jupiter-Verl. 1994, S. 14
[9] Dr. H. A. Nieper: Revolution in Technik, Medizin, Gesellschaft; MIT Verlag Oldenburg (1982)
[10] Prof. Shinchi Seike: The Principles of Ultra Relativity, Space Research Institute, 1990, ISBN 915517-1
[11] M. Cheney: Nikola Tesla, Man out of Time, Barnes & Noble Books, New York, 1993, ISBN O-88O29-419-1 und Omega-Verlag (1996) ISBN 3-93O243-O1-6
[12] Prof. J. Gruber: Kalte Fusion und Raumenergie, Bericht zur 7. Internationalen Konferenz uber "kalte Fusion" vom 19.-24.4.1998 in Vanvouver, B.C., Kanada (Infinite Energy, Issue 19, 1998) und Bericht zum "International Symposium on New Energy" vom 23.-26.5.1997 in Denver, Colorado, USA (Denver-Bericht der Fern Universitat Hagen)
[13] H.-J. Ehlers: Special 7, Raum & Zeit Dokumentation, EHLERS Verlag, Sauerlach, 1. Aufl. (1994)
[14] O. Alexandersson: Lebendes Wasser; W.Ennsthaler Verlag Steyr, 1993
[15] V. Schauberger: Die Entdeckung der Levitationskraft, Implosion 1995 Nr. 112
[16] A. Waser (Herausg.): Die ratselhafte Natur - The Puzzling Nature, AW Verlag, Luzern (1996), ISBN 3-9521O59-O-2

H. Literature concerning medicine, biology and chemistry

[1] U. Luttge, M. Kluge, G. Bauer: Botanik, 2. Aufl.- Weinheim, VCH (1994) ISBN 3-527-30031-7
[2] N.Begich: Towards A New Alchemy, The Millenium Science, Earthpulse Press 1996, dt. Ubers.: Auf den Spuren einer neuen Alchemie, Omega-Verlag 1997
[3] Couto: Die kosmische Oktave, Synthesis Verlag Essen 1984
[4] Prof. Dr. H. Heine: Grundregulation und Extrazellulare Matrix - Grundlagen und Systematik, Lehrbuch der biologischen Medizin, (1991) 2.Aufl. Hippokrates Verlag Stuttgart, Seite 63
[5] U. Warnke: Bioresonanztherapie - Wunsch und Wirklichkeit. Medizin transparent Nr.2/1996, S.36-37
[6] Dr. H. Clark: The Cure For All Diseases (1995), dt. Ubers.: Heilung ist moglich (1997) Knaur Verlag
[7] Dr. Hulda Clark: The Cure for All Cancers (1993) New Century Press, USA, dt.Ubers.: Heilverfahren aller Krebsarten
[8] Dr. Bodo Kohler: Biophysikalische Informations-Therapie, Gustav Fischer Verlag (1997)
[9] Dr. Bodo Kohler: BITsyn, der neue Weg in der Informations-Medizin, Jubilaums-Schrift der Intern. Arzte-Gesellschaft fur Biophysikalische Informations-Therapie, Bad Nauheim, 6.-8.10.2000.
[10] Bioenergetische Messverfahren in Theorie und Praxis, Vortragsband der Gesellschaft fur Energetische und Informationsmedizin e.V., Universitat Stuttgart am 17.7.1999 und am 7.10.2000
[11] Dr. Johann Lechner: Storfelddiagnostik, Medikamenten- und Materialtest, Teil 2 aus der Reihe: Praxis der Ganzheitlichen Medizin und Zahnmedizin, Verlag Dr. E. Wuhr (2000)
[12] M. Luscher: Der 4-Farben-Mensch, W. Goldmann, Munchen 1977
[13] M.Luscher: Das Harmoniegesetz in uns, Econ-Verlag 1993
[14] FA. Popp: Neue Horizonte in der Medizin, 2.Aufl. Haug Verlag Heidelberg 1987
[15] M.Bischoff: Biophotonen, das Licht in unseren Zellen, 2001

J. Literature concerning history and more peripheral subjects

[1] David Ash, Peter Hewitt: Science of the gods, Gateway Books, Bath, 1990.
[2] T.E.Bearden: Skalar Technologie, Michaels Verlag 2002, ISBN 3-89539-250-2.
[3] Thomas Valone: Das Unipolarhandbuch, Michaels Verlag 2001, ISBN 3-89539-295-2.
[4] R. Sheldrake: Sieben Experimente, die die Welt verandera konnten. Goldmann Verlag 1994
[5] R. Sheldrake: Das schopferische Universum. Meyster Verlag Munchen 1983.
[6] H.Ludeling: Handbuch der Radiaesthesie, Verlag Eike Hensch 1994, ISBN 3-927407-09-7.
[7] Hensch: Radiaesthesie im landlichen Bauen und Siedeln, Arbeitskreis zur Landentwicklung in Hessen, W4, Wiesbaden 1987
[8] H. Mayer, G. Winklbaur: Biostrahlen, Verlag ORAC, Wien 1989, 5. Auflage
[9] Rohrbach: Radiasthesie, Physikalische Grundlagen und Anwendung in Geobiologie und Medizin, Haug Verlag Heidelberg 1996.
[10] R. Schneider: Einfiihrung in die Radiasthesie Teil II: Leitfaden und Lehrkurs ..., Oktogon-Verlag, Wertheim, 2. Aufl. (1984).
[11] Lamer: Worterbuch der Antike, Kroner Verl. Bd.96
[12] K. Schefold: Die Griechen und ihre Nachbam, Propylaen Kunstgeschichte Berlin Bd. 1
[13] T. Kraus: Das romische Weltreich, Propylaen Kunstgeschichte Berlin Bd. 2
[14] G. Gruben: Die Tempel der Griechen, Wissenschaftliche Buchgesellschaft Darmstadt 1986
[15] E. Horst: Konstantin der Grosse, Eine Biographie, Classen Verlag 1985 Dusseldorf, 2.Aufl.
[16] Vitruv (Marcus Vitruvius Pollio): Zehn Bucher uber Architektur, Ubers. von Dr. K. Fensterbusch, Wissenschaftl. Buchges. Darmstadt 1987,4. Aufl.
[17] A. Springer: Das Altertum, A. Kroner Verl. Leipzig 1915, 10. Aufl.

உ

Time = Space

Supreme scientific secrets in ancient India

(God's particle, Language, Arts, Temples & Sacred Geometry)

(Shatkona Yantra – Created by Sri Bhogar Siddhar currently in a temple in Sri Lanka, taken once in a year for rituals, believed to emanate supreme subtle divine energies.)

Thinking Hearts (Raguram Gopalan)
(raguramg@hotmail.com)

உ

Thoughts from Thinking Hearts:

“Curious by nature, in a life so uncertain
Thirsty for the truth, which is so certain
I sought for some water in our Vedas - the fountain
Though I can’t understand fully what it contains
I am amazed & speechless with what I could ascertain
This is just a speck of dust from that golden mountain
Thankful to my family for helping me in what I could attain
Grateful for the wisdom and works from all the chieftains
Wish I repay every atom of it through my gratuitous tearstains”

This effort is to share what I read, understood and enjoyed about the scientific aspects in ancient India which are codified in our language, arts and in 1000s of temples. The quotes and examples are predominantly in Tamil but I have tried to give the meaning which can enable a reader to grasp the meaning and progress. Any shortcomings you notice are purely reflective of my abilities to understand and express. There was not even an iota of intention to hurt any sentiments or beliefs of others. A piece of my mind is available @ http://ragsgopalan.blogspot.com

I initially thought that I shall reproduce a prayer from one of the books which I had read but at one fine moment I decided to capture my feelings in my own words. So I scribbled few lines in English summing up my feelings as above, few lines for the Sage of Kanchi and a prayer in Tamil with its meaning in English. You may understand this prayer better after you read this book.

January 2012, Version 1.0

Please feel free to copy and reproduce in whichever form or shape as you deem fit without any acknowledgements. In fact see how much of this can you share to dispel the ignorance. The more you share the better. ☺

If I have acknowledged any books or authors please recognize them and ensure that you either buy their books or donate to their cause. They deserve all the credits.

உ

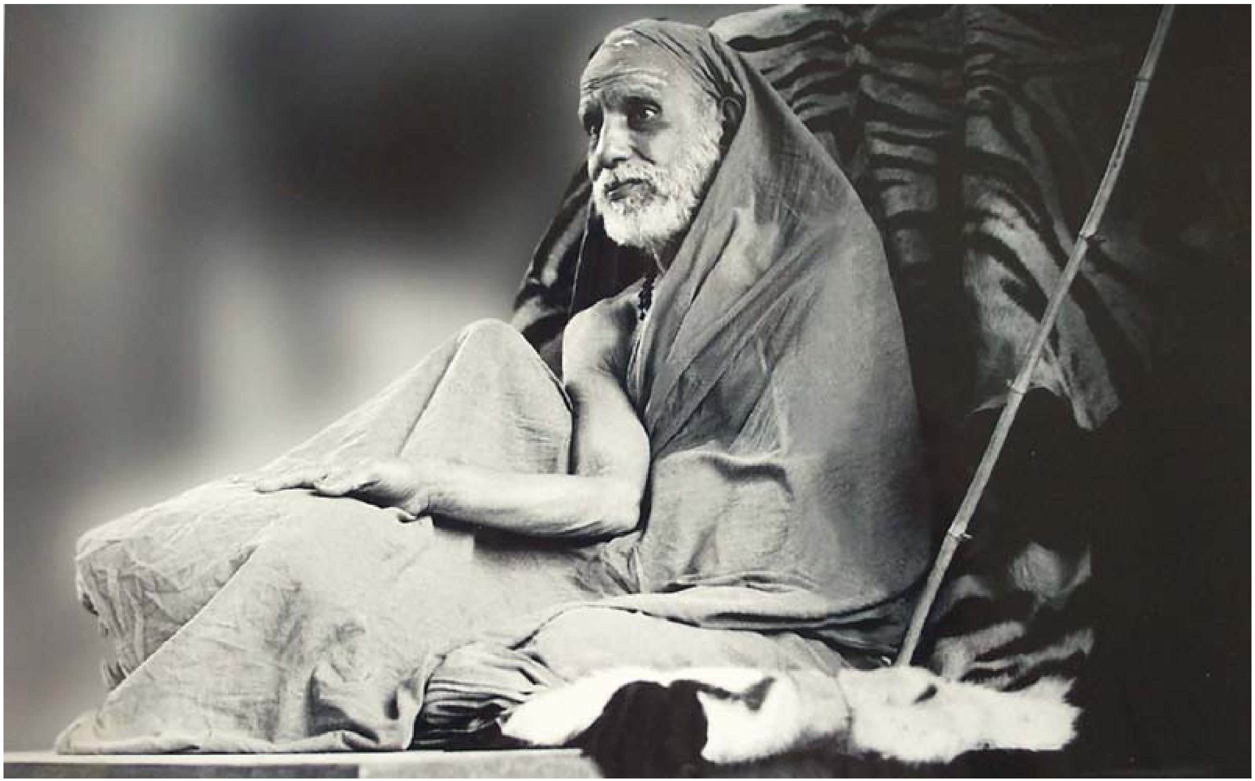

A divine soul
With a noble goal
Walked this earth
When divinity was dearth
Beyond any treasure's worth
To avoid our death & rebirth
Seek his guidance & grace
For he shall ever embrace!

உ

முதல் வணக்கம் (Prayer)

உள்ளமே மூலமாகி, பரம்பொருள் கோலமாகி
விண்ணே ஒளியென்றாகி, ஓமே உணர்வென்றாகி
உணர்வே ஓமொளியாகி, ஓமொளி ஓமொலியாகி
ஆடலன் ஆட்டமன்றோ, அசைவுறுக் காலமாகி
(Mind is the source, primordial god’s particle as the form,
The luminescent space, the Pranava as consciousness
Consciousness as the Om’s Light, Om’s Light as Om’s sound
Isn’t time a pulsation that’s the dance of Lord of Dance?)

காலமே சீலமாகி, சீலமேக் கோலமாகி
கோலமே ஞாலமாகி, ஆதியும் அந்தமாகி
ஞாலமே மூலமாகி, அறிவுடன்ஆ னந்தமாகி
நடுவிலே நன்று நிற்கும் நாதன் தாளே நாட்டமாகி
Time creates Rhythm & order, Rhythm creates shapes
Shapes become universe, the beginning and the end
Universe merges with source, it is awareness & bliss
Ever searching for the feet of the divine & graceful mean

விண்ணிலே மண்ணுமாகி, மண்ணிலே விண்ணுமாகி
விண்ணிலே எண்ணமாகி, எண்ணமே எண்ணுமாகி
என்னிலே என்னைக்காண என் கண்ணிலே மணியுமாகி
நிர்மலப் பொருளாய் ஓங்கும் நிமலனடி போற்றுவோமே!
The mass in energy (Space), the energy in mass(earth)
The energy as thoughts, our thoughts as numbers (octets)
To me find me within me, he who stands as my eye’s pupil
Beseech his feet who is devoid of may blemish

அம்பலத்தில் ஆடும் அவனை, அரங்கத்தில் துயிலும் அவனை
அரவத்தைச்சூடும் அவனை, அரவம்மேல் ஆடும் அவனை
பைரவித்துணைவன் அவனை, திருமகள் தலைவன் அவனை
நம் அறிவுக்கு இருளின் நமனை, பொற்பாதம் போற்றுவோமே!
The one who is dancing in the microbode, the one who is sleeping in arangam
The one who is wearing snakes, the one who is dancing on a snake
The one who is consort of Goddess Bhairavi, the one who is the Lord of goddess Lakshmi
Beseech his feet, the one who kills our mind’s tamasic darkness

உ

TABLE OF CONTENTS

1. Time = Space Series. Part 1 - Introduction 6
2. Time = Space Series. Part 2 - Sound merges in light 9
3. Time = Space Series. Part 3 – Vaak-Artha - Importance of word 11
4. Time = Space Series. Part 4 – Scientific languages – Tamil & Sanskrit 14
5. Time = Space Series. Part 5 – Energy & matter conundrum 16
6. Time = Space Series. Part 6 – The Unfolding Self - Universe has evolved and was not created 19
7. Time = Space Series. Part 7 – The Chidambara ragasyam & the God's particle 21
8. Time = Space Series. Part 8 – Characteristics of the God's particle (Paramaanu) 24
9. Time = Space Series. Part 9 – Light, God's particle and God 27
10. Time = Space Series. Part 10 – Lord Nataraja and The String theory 30
11. Time = Space Series. Part 11 – Vastureva Vaastu and $E=mc^2$ 34
12. Time = Space Series. Part 12 – Vedic Concept of time & its measurements 38
13. Time = Space Series. Part 13 – South American Mayan & South Indian Maayan 40
14. Time = Space Series. Part 14 –Mayan, Tamil, Pleiades & iCHING 43
15. Time = Space Series. Part 15 – Shiva Lingam is not a Phallus 45
16. Time = Space Series. Part 16 – Seelam - The Primordial circular force - Ganapthi & Mathematics 50
17. Time = Space Series. Part 17 – Seelam - The Primordial rhythm & fundamentals 53
18. Time = Space Series. Part 18 – Seelam - Significance of √2 & √3 56
19. Time = Space Series. Part 19 – Seelam –Vesica Piscis, √5, & Golden mean (Φ) 58
20. Time = Space Series. Part 20 – Sankhya yoga and Sage Kapila 63
21. Time = Space Series. Part 21 – 5 and 8 the Universal rhythm 65
22. Time = Space Series. Part 22 – Eight (Octet) and the 5 sacred arts 67
23. Time = Space Series. Part 23 – Nanotechnology in ancient india & size of god's particle 69
24. Time = Space Series. Part 24 – Octet, Octave and the Gandharva veda (Classical Music) 72
25. Time = Space Series. Part 25 – Adi Tala, Nava Tala and Sthapatya veda (Sculpture) 74
26. Time = Space Series. Part 26 – Sabda Veda, Chandas & Poetry (Ma Nishaada…) 78
27. Time = Space Series. Part 27 – Architecture – The Zenith of mathematics (Vaastu) 82
28. Time = Space Series. Part 28 – Capturing the divine energy – Squares & Circles 87
29. Time = Space Series. Part 29 – Heaven & Earth – Lord Nataraja & Lord Ranganatha 91
30. Time = Space Series. Part 30 – Temples and Idol worship 95
31. Time = Space Series. Part 31 – Mayan & his works 101
32. Time = Space Series. Part 32 – Conclusion 105

உ

1.TIME = SPACE SERIES. PART 1 - INTRODUCTION

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

“Abheda Darsanam Gnanam[*]”. Nothing defines what **wisdom is** better than these 3 words. **Experience of unity / oneness in all is wisdom.** The roots of Indian wisdom are captured in these 3 words that everything in this universe is the manifestation from a single source, and we can call it God or Consciousness. This naturally leads us to one of the mahavakya ***“Pragyanam Brahma”*** which means **“Consciousness is god”.** But for people who are stingier with words and wants these 2 concepts to be put in just one word, then it is called **“Advaita”.**

It is not very often that nature reveals itself its absolute oneness and the un-manifested interconnections. You consider yourself fortunate when you come across such knowledge in however small measure, even if it happens very rarely and even if you have an illusion that you have understood something. It has been very overwhelming for me to come across some of the details I have been reading and hearing and the dots I could connect – some of which I had decided to write as a book.

If you are interested in knowing how scientific ancient Indian Sanatana Dharma is and how well we have codified nature’s secret in our languages, arts and temples and if you are interested in any of the following questions then you may like this book. Most of the aspects stated here as scientific proof has a reference to one or more of our scriptures and I would refer and quote them at various places.

- Would you be surprised to know that the size, shape, color and characteristics of God’s particle are very clearly documented in our scriptures at least 12000 years back? What the scientists at CERN are spending billions of $, is well documented and we will discuss the characteristics of God’s particle in this book.

- Would you want to know how **Time = Space** and more over time is nothing but the vibration of space? If I tell you that this was known us Indians at least 12000 years back and currently available in a written form would you believe?

- When you throw a stone in water / pond the ripples that form take the shape of a circle (in 2D) / sphere (in 3D). Can you guess what the shape of the ripple would be if you throw the same stone in space?

- What is the similarity between Poetry, classical music, Classical dance like Bharatanatyam, Sculpture and a building? Apart from the fact that we don’t clearly understand any of them in depth what is the similarity? Would you be surprised if I say that all are essentially the same and in fact a **well constructed building is nothing more than a frozen music** and the relationship between all the above confirms to a single grammar documented 12000 years + back ?

- Would you be interested in knowing how $E = mc^2$ an equation stated by our great Einstein has been known to us for years and not just that but we know the root of this energy and how to engineer it with precise formulas? Would you be surprised that this is the basis our language?

- What if I say that the age of the universe, every astronomical unit of the bodies, big bang theory, string theory, God's particle etc are all documented and they are all codified in our temples?

- You would have heard about the golden ration Φ but do you know the philosophical significance of Φ and how it is derived? Moreover would be interested in knowing the significance of key numbers and ratio like 0, 1, 5, 8, 9 and $\sqrt{2}, \sqrt{3}, \sqrt{5}$? What is I say that $\sqrt{3}$ can be correlated with "**Iyengar Namam"?** Does it sound funny? What if there is a deep mathematical and philosophical significance? And how are all these ratios related?

- As a wise man said that **'Mathematics is the language of the gods".** What if this formula is codified in our texts from which we understand the God's particle, its shape, characteristics and its manifestation?

- What if I tell you that sound is a subset of light and sound merges with light at a point and that is at the root consciousness? Any non-takers for this?

- If you agree on the above then **Sabda Brahman** or the belief that sound is god merges with something higher, isn't this contradictory? Also why many religions did believed that "**Word is god".** Was it a mere belief or was there any scientific reason behind it?

- What is the correlation between Carnatic music and a human being's height?

- What if I say that the Mayan from South America and the Mayan from South India are one and the same?

These are some of the questions which might get addressed in this book. I am going to quote a series of references and books and links to understand and learn more about these topics and I would request you to utilize them if you are interested in these topics.

This may not only refer you to the experts but also can help you overcome any weakness I have in understanding and articulating some of these topics. Some of the topics may touch upon religious beliefs (mostly Hinduism) but if you read with an open mind you would notice the underlying science behind it.

I would be happy if any of these blogs motivate and inspire you to explore the roots of our tradition and knowledge and give you an insatiable hunger for knowledge, wisdom to pursue the nature and its secrets.

May the ancient Sanskrit quote “May I meditate on the supreme truth “Satyam Param deemahi”, resonate within you as it does within me.

Happy reading!

PS:

Abheda Darshanam Gnanam, Dhyanam Nirvishayam Manam;
Snaanam mano mala thyagam, Showchamindriya Nigraham", says the Upanishad.

The real Gnana is seeing all as one, real Dhyana is controlling the Manas from all temptations, the real snana or Bathing is Cleansing of All Mindly impurities, not physical impurities alone and real Showcha or cleanliness is controlling the Indriyas, this controlling of Indriyas, emotions, weaknesses and temptations.

I am also quoting parts of Kahlil Gibran here,

Your hearts know in silence the secrets of the days and the nights.
But your ears thirst for the sound of your heart's knowledge.
You would know in words that which you have always known in thought.
You would touch with your fingers the naked body of your dreams.
And the treasure of your infinite depths would be revealed to your eyes.
But let there be no scales to weigh your unknown treasure;
And seek not the depths of your knowledge with staff or sounding line.
For self is a sea boundless and measureless.
Say not, "I have found the truth," but rather, "I have found a truth."

2.TIME = SPACE SERIES. PART 2 - SOUND MERGES IN LIGHT

What if I say that “**Sound is a subset of light”.** Scientifically you may agree that both are electro-magnetic vibrations at different frequencies in the spectrum though one may not be the subset of another.

So, What if I say that every other frequency is the subset of light? I am sure many may disagree. Some of the religious pundits may claim that “**Sound is god**” which I am not disputing but “**light too is god”** and as we progress towards the primordial state sound merges into light.

Let me try to explain this with a simple exercise. Please sit in a place with a calm mind – few deep breaths may help.

- Imagine that I am TALKING to you and avoid thinking that you are reading this blog.
- I am going to TELL the following 5 words and you are going to LISTEN to it. Please just observe what happens within you.
    - CAT, BOAT, MOM, DEVIL, BEST FRIEND

The following are the fundamentals of perception as per Hindu scriptures:

- Every word has a meaning (Let us ignore the so called meaningless words for the time being.).
- Every meaning invokes a form or shape within our mind.
    - For example when I say “**MOM”** you associate this word with your mother and the picture / image of your mother is perceived inside. You need to be little more observant to understand this. **The fundamental principle is that**
    - **“Every word is first associated its meaning then the meaning is associated with a form. The form is associated with a Guna (Quality) and then the quality or Guna is associated with its experience. What is stored inside is just the experience.**

Let me TRY to explain the abstract concept of consciousness as below:

- **Consciousness / existence exist in 3 states (deeper explanation is 5 / 7 states but we will discuss the simpler alternative).**
    - **Formless state called Experience / feeling / bhava. This is the super conscious state.**
    - **Gross state with form which has physical attributes. This is the conscious state.**
    - **In-between state (Not formless and no form) called Guna state which has subtle quality attributes. This is the Subconscious state. If this is difficult to comprehend for you please imagine “Egg” which is in-between the formless and state with form.** Gotcha?

Let’s look at an example. Rose is a flower. When someone tells you **“ROSE”** then the following happens:

- The sound “rose’ associates with a flower usually. (I am ignoring it being a name of your girl friend ☺). So here the word is associated with its meaning.
- This flower is associated with red color and perhaps with thorns. Here the sound does not exist anymore but only a form. **Yo!, the sound has merged into light.**
- The third step is - this form is associated with a quality which is fragrance of a rose. For someone sensitive they can smell the rose when they think about it. How else could we salivate at the thought / smell of a favorite dish if this step does not exist?
- Now the last step is, this quality invokes an experience and hence a feeling. This rose may be associated with your inner self for love and hence the emotion of love surfaces. This is called **Bhava / Rasa** in Sanskrit.

If you are not clear about this then please do not move to the next part and think or discuss about it. A little contemplation would help you understand this **5 stage** process of perception.
‘SOUND – MEANING – FORM – QUALITY (GUNA) – EXPERIENCE (FEELING / BHAVA)’

Our ancient forefathers knew this science and interconnection. Let me now quote these Slokas as an example from Bharatnatyam, which talks about this linkage.

Kanteenaalambayet geetam, Hasteenaardham Pradarsayet,
Chakshurbhyaam darsayeet bhaavam, Paadaabhyam taalamaacharet.

According to Nandikeshara the dancer should sing with the mouth, express the meaning of the song with hand gestures, her eyes should express the emotions or bhava, the tala, the beat or rhythm should be done with her feet. Further he goes on to say –

'Yato hastas tato drushti, Yato drushitis tato manaha
Yato manahas tato bhavo, Yato bhavas tato Rasaha'

'Where the hand goes, there should follow the eyes, Where the eyes move, mind should follow it,
Where the mind follow, Bhava or emotion is created, & lastly when Bhava is created the sentiment or Rasa will arise.'

All the Mudras or gestures represent word when held at a particular position around the frame of the body and when used in a particular way.

Alas we Indians have ignored this great science & art and its linkages and are like ignorant fools following the meaningless western tradition and dances!

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

3. TIME = SPACE SERIES. PART 3 – VAAK-ARTHA - IMPORTANCE OF WORD

'SOUND – MEANING – FORM – QUALITY (GUNA) – EXPERIENCE (FEELING / BHAVA)'

In the **5 stage** process of perception, one important factor which I missed is the **"Self / EGO"** which interprets the **Guna** to an experience. For example when I say Dog, the **'Self'** interprets Dog to a **loyal friend or a bike chaser or a rabbies generator.** Essentially what is the quality of dog mean to me? If you understand this process this explains the uniqueness between individuals or the root cause of differences in a society.

I am also going to add but not going to focus on the fact that the inputs are never limited to sound and when see a picture the process of perception starts with step 3. Just to poke the individuals who think "**Sound / Word is the only god – Sabda Brahman**, please tell me - when you see a picture and when there is no sound associated with it, where is **God as Sabda Brahman in this situation**?

Does that mean that the very important Hindu belief that the Sound **"Om is sabda Brahman or god"** does not hold any meaning? It certainly does and I shall share my thoughts on it – continue reading.

What is the importance of Word?

The greatest of Indian poet **'Kalidas'** said

"Vaak artha viva samprakthov vaakartha prathipathaye
Jagadha: pitharov vandhe parvathe parameswarow"

he compared Lord Shiva and Shakti like sound (Shakti) and meaning (Shiva). Both sound and meaning are inseparable. You would now understand the wisdom behind this sloka since there is no sound without a meaning. Also **Sabda Brahman** merges with the **Artha Brahman** and both are inseparable. Here Artha Brahman is the light and the meaning of the word. Remember we say that **'Throw some light"** when we don't understand something.

In Tamil **பொருள் (porul)** means "a thing" and it also means "meaning". We will see how **பரம்பொருள் (Paramporul)** is the Lord Shiva as a God's particle and how he is referred to as the meaning too. The same applies to Sanskrit where **"artha"** means **wealth / things – remember Arthasastra?** And it also refers to **"meaning"** as in the above sloka. How rich these languages are?

And in **Abhirami Andadi, bhattar** sees devi as **"Arthanareswara"** and confirms us that the grace of devi will brings us both richness and mukthi. **"Sollum porulum ena nadamadum"...**

உ

A word is made up of letters called Akshara in Sanskrit. The meaning of Akshara is that which cannot be broken down further. ie. Word can be broken down to letters but it cannot be broken down further. A letter has a form like **“A”** and has a sound associated with it. This means the light and sound are merged in a letter. As per the principle stated above in the process of perception, the primordial state is one where light and sound is merged into each other.

If you don’t dispute the philosophy that everything merges into god at the primordial state and at this state, sound and light are inseparable, and then you would agree that an Akshara is **God / Brahman.** Each letter has significance and meaning in Indian languages especially Tamil and Sanskrit. **A word is a combination of Form (Light), Sound and meaning and these are first 3 aspects of perception which takes us from a gross state to a formless subtle un-manifest state.**

Hence the statement from New Testament **“In the beginning was the Word, and the Word was with God, and the Word was God”** reflects the same principle.

But as per Hindu scriptures Veda is the breath of god (not word of god) or the aural vibration of consciousness cannot take you to the state of moksha. Words don’t take you to God, it can help you in the process but has its limit, since the sound and meaning merges into something else as we move towards more subtle space.

This is what my friend Kambar said in Kambaramayanam as **“வேதமும் முடிவு காணா மெய்ப்பொருள் வெளிவந்தெய்தி மா துயர் துடைத்த வார்த்தை மறுப்பரோ!!.** This means that the supreme truth who could not be found by Vedas came and solved the problem of an elephant, can someone dispute this?

Every word or Akshara has four states **Para, Pashyanti, Madhyama and Vaikari. Vaikari** is the fourth state where the sound is audible to us. If I recite something mentally then the sound has 3 states only and these 4 states signify as to how a feeling (bhava) becomes thought and thought becomes form and then a sound. Just note here that as per Frank outlaw, it’s the thoughts which are primordial but as per our scriptures it is feeling or Bhava or deep rooted desire or also known as vasanas which drive your thoughts. So Frank’s quotes are half baked plagiarism.

In a work called Tarka Samgraha, we have classified 9 different physical attributes (substances / dhravyas) and 24 different qualities / GUNA in all. If you are interested reading more about this, please read these blogs or the original work.

http://ragsgopalan.blogspot.com/2010/09/vedic-perspective-matter-and-its_14.html This is a 3 part series.
http://ragsgopalan.blogspot.com/2008/11/part-1-perception-art-of-seeing-things.html This is a 2 part series.

உ

To summarize this topic:

- Broadly consciousness exists in 3 states.
- The process of perception if understood well would throw a lot of secrets about nature.
- The 5 stage process of perception **“Sound – meaning - Form – Guna – Bhava”** is a simplistic way of portraying this.
- Sound merges into light and light is worshipped as God in most of the religions and belief. We can explore little more of this in the future sections.
- A word is a unique combination where Light, Sound and meaning merges and hence it is considered as god. This word represents **Arthenareeswara or yinyang or Star of David**.

In the future sections let us explore as to what is the relationship between Maths and God and some of the critical numbers and its philosophical and scientific significance.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading

4. TIME = SPACE SERIES. PART 4 – SCIENTIFIC LANGUAGES – TAMIL & SANSKRIT

One of the main reasons I believe as to why our wisdom has degraded so much in the last century is our ignorance and reluctance to learn our rich language. Our language offers richness not just in literary sense but is supremely scientific too. I am going to touch upon just 3 to 4 aspects of it which I knew and understand at a fundamental level and request you to learn your mother tongue well.

Many of my friends don't exhibit a sense of remorse when they say **'I can understand Tamil but can't read or write it**". I wanted to tell them how much in life do they miss in terms of knowledge and richness. This is not just about Tamil but every major Indian language is rich in its own way. Since my mother tongue is Tamil, I am going to articulate some of the science in it.

Vigyan / Vigyanam (விண் ஞானம்) means Science:

But most importantly for this series I wanted to take the word called Science as meant in Tamil and Sanskrit. A word always can be broken down to its dhatus for its exact meaning in Tamil and Sanskrit. Some examples below:

- Science in Tamil is **Vigyanam and in Sanskrit is Vigyan**. This is split as **Vin** and **Gnanam** / **Gyan**. **Vin means space and Gyan means knowledge. Hence according to Indian system science means it is the nature / knowledge of the space.** This has very deep significance since western science has not agreed the existence of space for a long time and their gross and materialistic approach prevented them from seeing the most subtle component and its nature. We were masters of it ages back. Most of the questions in Chapter 1 would be answered based on the knowledge of space – Vigyan.

- **Universe is called as Brahmanda** in Sanskrit. Brahma + Anda which literally means a big egg. This means we knew that the galaxy is elliptical in shape ages back and it required hubble telescope in 1950 to tell us that.

- **Bhoogol** means **Earth (Study of) / Geography – This Splits as Bhoo +Gol.** Bhoo means Earth and Gol means round. We did not throw stones at a scientist as late as 17th century when he said that earth is round and not flat. We would have laughed our heart out and ignored him.

உ

Be it the age of the universe, all of the astronomical measurements, health, spirituality etc, many of us need a white skin with blue eyes to tell us that what we knew was right and they have certified it. When we decide to come out of this mentality then the process of reclaiming our rich heritage shall start.

Origin of Universe and linkage with Tamil:

The initial evolutionary process for the universe consisted of 5 stages states Aintiram. You can imagine that these are the movements of particles during the Big bang. They are அமிழ்தல், இமிழ்தல், குமிழ்தல், உமிழ்தல், தமிழ்தல்.

- Withdrawl (Amizhdal)
- Imizhdal (Overflowing)
- Kumizhdal (clustering round in an order)
- Umizhdal (Emitting)
- Tamizhdal (resulting into a well defined form)

These are the first stages from the Big bang till a well defined shape is formed. Now the last level is called Tamizhdal which results in a well defined form. This state has emanated from the Pranava and not been created by some one. Hence a language as rich with the sweetness of inherent order, originating from the luminosity of the original particle is called Tamil. Tamil is word based on the final resultant state called Tamizhdal (தமிழ்தல்).

I shall address the other 2 interesting aspects that include a Chicken and Egg analogy for Energy and matter conundrum in the next blog.

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy reading!

உ

5.TIME = SPACE SERIES. PART 5 – ENERGY & MATTER CONUNDRUM

Energy and Matter concept in formation of letters:

The concept of Energy and matter is not new to us and does not start with Einstein's days of $E=mc^2$. Let us look at how our letters are formed from Aksharas then people would know that the concept of Energy and Matter is old as our language. In Tamil there are two kinds of Aksharas **உயிர் எழுத்து - Uyir Ezhuthu (Life / Energy letters – 12 letters) and மெய் எழுத்து - Mei Ezhuthu (Body / Matter letters – 18 letters).** Both put together it is 30 letter pack for Tamil as alphabets. Now both have to combine to give letters called **"body with life"** which are called **uyir mei Ezhuthu**. This combines to form 18*12 which is 216 letters as **Uyir Mei Ezhuthu**. For example க்+ அ = க in Tamil and क़ + अ = क in Sanskrit.

So the formation of letters itself is a science where Matter (Body) combines with Energy (Life) and then the living organisms (Body – life letters) are formed.

I want to ask you a **Chicken & Egg** question here which is very relevant to today's scientists. Did the energy come first or is it the matter that came first? Modern science is still scratching its head. I can answer this question in 2 ways, first is obviously using language as a science.

- In the above example க்+ அ = க, க் is the body and அ is life. So as per our language sciences be it Tamil or Sanskrit the matter is formed first and then the life came and attached to it. This is very well documented in Tamil scriptures as follows:

உடம்புடன் உயிர் வந்து ஒன்றுவது இயல்பே - தொல்காப்பியம் . This is over 2500 years old which states that life comes and joins a body in language and this is nature.

மெய்யொளி உயிரொலி சிவணுதல் இயல்பே –ஐந்திறம். This is over 12000+ years old which states that same thing and adds that body is like light and life is like sound and they merge and this is siva's nature. **This certainly means that both Body and life are not formed at the same – as popularly believed. Life comes and joins the body to give a living being – in the above case the letters are the living beings.**

The second reference I am going to take for the same concept is our **Yogic / Siddha / Ayurvedic** science as to how fetus is formed in mother's womb. As per our (Siddha) belief the conception of a child in a mother's womb happens as follows:

- **The first organ that is formed for the fetus is the cells of the Eyes (Single Eye)**. This eye cell starts vibrating due to its heat and multiplies as 2 eyes and then as the whole body. This is the vibration seen in the initial days of the fetus and not the heart beat. As per modern medical science the heart takes shape only by the end of the 8th week and not before. I am requesting someone to validate this.
- **Life comes to this body at the 3rd month of conception only**. This is the belief as per Ayurvedic and Siddha science.
- Instances of the third month natural abortions are very high than the later months based on the above reasons. This means that the life is not destined for that matter / life force is not sufficient enough.
- **How many of you know that the size of eyes never change as you grow and stays the same from the time we were born,** while all other organs undergo growth?
- Also the eyes hold a very important significance in the spiritual science which I shall address later. The eyes are the seat of Sun god and are characterized by heat and light.

It is through the heat in the eyes the subtle became a gross living being in this world and if the living being wants to merge back with the subtle – which is the process of Samadhi then the living beings should use their eyes, increase the heat in the eyes and start the devolution process. This is a supreme secret kept by the sages even in today's world. **You may want to ask any spiritual school as to why they recommend meditation with eyes fully or half open instead of closed eyes? Ask yourselves why the free masons and other secret societies use eye as one of their symbols?** We shall address this later.

Now you want to answer the Chicken and Egg analogy to the Energy and Matter conundrum based on the above 2 thoughts, we can conclude it is the matter which came first. We shall try to explain this based on Space sciences a little later.

Masculine and Feminine names based on ending letters:

I am also going to give you as to how Indians distinguished masculine from feminine names. **Kanchi Maha periyava, the Sage of Kachi** had explained this beautifully in his **"Arul Vakku"**. I am giving a gist of it here just to conclude this:

- **Masculine names:** All the names shall end with the phonetic a, im, in , ir, aha etc. Examples are Ram, Shyam, Raman, Kannan, Kishan, Vivek, Adam. So phonetically names like Catherine is a masculine name because it ends with "in" phonetically.

- **Feminine names:** All names ending with phonetics aa, e, I, oo etc. Examples are Radha, Ramya, Savithri, Gayathri, Uma, Rosy, Isabella etc. Same way phonetically Joshua is a feminine name.

Vishnu and Lalitha sahasranamam follows this rule and hence one invokes Shakthi as lalitha and the other invokes Purusha as Vishnu.

Just to conclude this part, I think we looked at how science is an integral part of our language formation and for us science has always been about the knowledge and nature of the space and we believed that space manifested itself as the physical world. So in the next few parts we shall look at how this manifestation has happened and other interesting aspects.

I would conclude this part with a earnest request that please encourage your kids to learn their mother tongue – to read and write and you would not have done anything better than this to advance their spiritual growth and adding richness to their life.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

6.TIME = SPACE SERIES. PART 6 – THE UNFOLDING SELF - UNIVERSE HAS EVOLVED AND WAS NOT CREATED

As per many scriptures in Hinduism this universe has evolved from itself and not created by someone. If you want to quickly ask as me as to what Brahma, Vishnu and Rudra mean – I would attribute them to different process heads in this evolution. They are all different aspects of the same consciousness based on their state of being in the process of evolution / devolution. This is my understanding but let us avoid this discussion here.

Here I am indebted and thankful to the nobel soul **Late. Dr. Ganapathi Stapathi,** (http://www.vastuved.com/life.htm) who not only unraveled the Vaastu science – the science of Space and time but also brought to light 2 important books which were over 12000+ years old written in Tamil by **Mahamuni Maayan.**

Mamuni Maayan is believed to be the same person who is popularly known in **South America as Mayan** and the 2012 phenomenon based on Mayan calendar. Dr. Ganapathi has established proofs that the pyramids in Middle East, the temple architecture in India and the structures in Mexico and Guatemala confirm to a single Science and Mathematics seen in all these works.

At the end of the blog I am listing few books which helped me understand this topic but most of the books (**Aintiram and Pranava Veda**) are out of print now and if you are lucky, you would get some of them. Now let us understand briefly how this evolution process has come about. According to Mayans' Aintiram:

- The free space is the unified field of energy and matter and source of all forms that we see in material world. That is why we were defined science as **Vingyanam**. This space consists of very minute particle called **Vinporul, nunporul, Sittrambalam, Paramanu, Oliporul, Microbode** (**விண்பொருள், நுண்பொருள், சிற்றம்பலம், பரமாணு, ஒளிப்பொருள்**) and in many names. You can pick what you like. I am going to use **Paramaanu** in the rest of the blog.

- The whole process of evolution is a 5 stage process. (Keep noting the importance of 5 and at the background remember Shiva is represented by Panchakshara mantra (Na~Ma~Si~Va~ya).

So what is the five stage process?

- This space is called as Moolam (Source).
- This Moolam starts vibrating and this vibration of the space is called Kalam (Absolute Time and not physical time. We will refer this as time only).
- This Kalam (time) vibrates to a mathematical resonance called Seelam (Rhythm).
- This Seelam as it resonates creates Kolam (Shapes – Circles, cubes etc).
- This Kolam becomes our visible Universe which is called Gnalam.

- So it is **MOOLAM (Source) – KALAM (TIME) – SEELAM (RHYTHM) – KOLAM (SHAPE) – GNALAM (UNIVERSE). மூலம் - காலம் - சீலம் - கோலம் – ஞாலம்**, these are 5 stages of evolution.

This means that the Space itself becomes spatial forms and time is the instrument that creates, sustains and destroys. Lord Siva in Tamil is also called as **Kalan (காலன்)** as to who controls time since Kalam **காலம்** is time.

To put it differently, un-manifest state when pulsates becomes absolute time and this is the creative energy. The process of pulsation confirms to a mathematical order and we will see this order in detail later. You may now understand the phrase that "**Time creates, sustains and destroys**". Since if there is no pulsation then nothing happens.

For those who need quote from a foreigner to convince themselves of this can be referred to the **Bertrand Russel's** quote "**What we perceive as qualities of matter are differences in periodicity".**

I shall quote some from the oldest literatures for reference below:
காலமே எவைக்கும் மூலம், காலமே மூல ஆற்றல்
காலமே கோலத் தோற்றம், காலமே ஞாலத் தோற்றம்… *Maha muni Mayan from Aintiram*

காலக்கூறே கணக்கியலாகி… Time splits as Maths in proportion it pulsates.
Vyasa says in Adi parva chapter 1, shloka 249, says the same thing that Time Creates, sustains and destroys everything.

kalah srijati bhutani kalah sanharati prajah
sanharantah prajah kalam kalah shamyate punah!!

The same is quoted as **"Vastur eva Vaastu"** which means the un-manifest Paramaanu which is called **Vastu** (with Single a) becomes **Vaastu (gross form – with two aa**). You can relate this **Vaastu** as the building science which is grossly misused and misrepresented and commercialized in today's TV programs. We will address the science in it later.

I shall leave this part with this thought, the space is considered as **Nataraja (Shiva) – subtle component and Ranganatha is considered as the Earth component**. So Shiva and Vishnu are two aspects of the same seed. We will see more scientific proof of this later.

"Shivasya hridayam vishnur, Vishnoscha hridayam shivah:" and
"Shivaya Vishnu rupaya Vishnave Shiva rupine"

Vishnu is the heart of Shiva and likewise Shiva is the heart of Vishnu and they both are the representation of each other.

Now if you are clear about the five stage process of evolution we can delve into each and every aspect of this in little more detail that would answer everything.

Ref: Check out all the ones here. Key recommendations are item 6, 11, 14 and 43. http://www.vastuved.com/publication.html

7. TIME = SPACE SERIES. PART 7 – THE CHIDAMBARA RAGASYAM & THE GOD’S PARTICLE

There were 2 questions I had confronted myself when I was reading these passages. They are:

- Why did the unmanifest space energy started pulsating and became absolute time?
- What did the unmanifest space energy contain? Was it a point of singularity as the modern scientists wanted us to believe? If so then how much mass was concentrated in that point which is obviously very difficult to believe.

The answer to the first question is here **“The self unfolds itself for its self – Bhagavad Gita”.** There is no convincing answer to the rational mind as to what the reason that the consciousness which was passive decided to become dynamic and started pulsating. So let’s leave this for the time being.

But the second question is answered very comprehensively. But before we get into the details of what it is, I want you to check Figure 1. This gives the overall picture of evolution and the five stage process. I also want you to recollect the “EGG” state which is called as the GUNA state where in the form and sound manifests. They say that this pulsation started with OM. This OM as popularly believed is just not an aural form but also has a visual form and that is light. We will discuss about this later in detail.

But let us focus on the actual second question as to what the contents of this Space energy were:

- The space is filled with **Paramanu or Microbodes** which are tiny particles. These are the **God’s (Higgs Boson) particle** our CERN scientists are after. http://en.wikipedia.org/wiki/Higgs_boson

- Now, we are probably a year away before CERN publish the results when this blog is published. I would not be surprised to see the results of that experiment is closely aligned with what is available in our scriptures and produced in this blog.

- The primal manifest form of the unmanifest is a **square** (primal wave pattern). This is the shape of the **Paramanu or Microbode**. This is very luminescent, means filled with light and on a 3D plane this shape is a **Square Cuboid**.

- A **Square Cuboid** is a shape which is equal in all sides – say of unit measure which is the smallest measure possible and has **6 faces, 8 vertices, and 12 edges.** You can start figuring out in Hinduism, why the idols of Lord Shiva and Lord Muruga has six faces, 12 eyes, hands etc.

- You may want to note that 6 faces can be represented in 2 ways as below. Figure 3 is Lord Muruga’s star or a Star of David in 3D which has 6 faces and 12 edges, popularly known as **Tetrahedron**. **So don’t be surprised if the Boson Higgs particle is of the same shape.**

உ

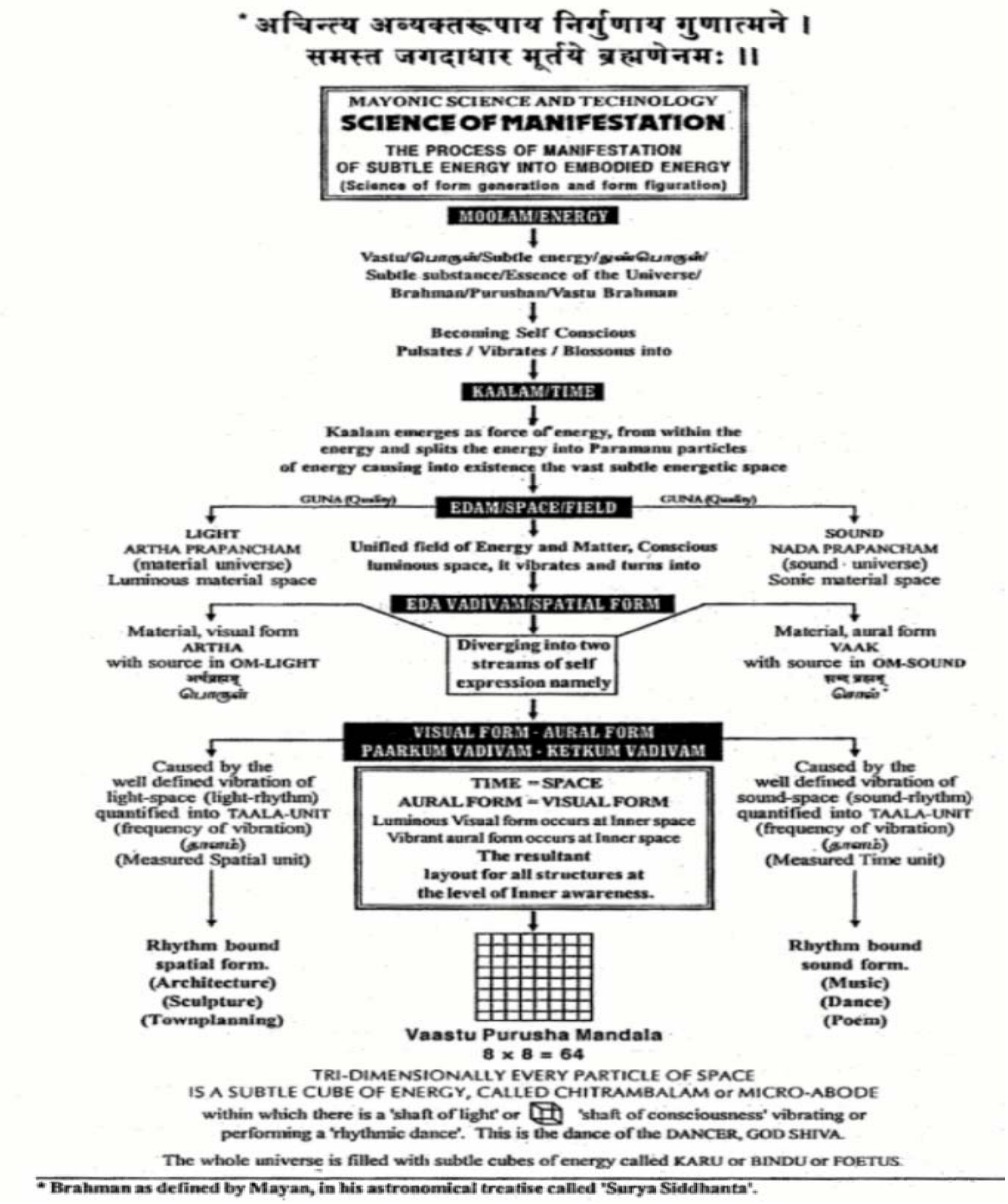

*अचिन्त्य अव्यक्तरूपाय निर्गुणाय गुणात्मने ।
समस्त जगदाधार मूर्तये ब्रह्मणेनमः ॥

MAYONIC SCIENCE AND TECHNOLOGY
SCIENCE OF MANIFESTATION
THE PROCESS OF MANIFESTATION
OF SUBTLE ENERGY INTO EMBODIED ENERGY
(Science of form generation and form figuration)

MOOLAM/ENERGY

Vastu/பொருள்/Subtle energy/நுண்பொருள்/
Subtle substance/Essence of the Universe/
Brahman/Purushan/Vastu Brahman

Becoming Self Conscious
Pulsates / Vibrates / Blossoms into

KAALAM/TIME

Kaalam emerges as force of energy, from within the
energy and splits the energy into Paramanu particles
of energy causing into existence the vast subtle energetic space

GUNA (Quality)
EDAM/SPACE/FIELD
GUNA (Quality)

LIGHT
ARTHA PRAPANCHAM
(material universe)
Luminous material space

Unified field of Energy and Matter, Conscious
luminous space, it vibrates and turns into

SOUND
NADA PRAPANCHAM
(sound · universe)
Sonic material space

EDA VADIVAM/SPATIAL FORM

Material, visual form
ARTHA
with source in OM-LIGHT
अर्थब्रह्म
பொருள்

Diverging into two
streams of self
expression namely

Material, aural form
VAAK
with source in OM-SOUND
शब्द ब्रह्म
சொல்

VISUAL FORM - AURAL FORM
PAARKUM VADIVAM - KETKUM VADIVAM

Caused by the
well defined vibration of
light-space (light-rhythm)
quantified into TAALA-UNIT
(frequency of vibration)
(தாளம்)
(Measured Spatial unit)

TIME = SPACE
AURAL FORM = VISUAL FORM
Luminous Visual form occurs at Inner space
Vibrant aural form occurs at Inner space
The resultant
layout for all structures at
the level of Inner awareness.

Caused by the
well defined vibration of
sound-space (sound-rhythm)
quantified into TAALA-UNIT
(frequency of vibration)
(தாளம்)
(Measured Time unit)

Rhythm bound
spatial form.
(Architecture)
(Sculpture)
(Townplanning)

Vaastu Purusha Mandala
8 x 8 = 64

Rhythm bound
sound form.
(Music)
(Dance)
(Poem)

TRI-DIMENSIONALLY EVERY PARTICLE OF SPACE
IS A SUBTLE CUBE OF ENERGY, CALLED CHITRAMBALAM or MICRO-ABODE
within which there is a 'shaft of light' or 'shaft of consciousness' vibrating or
performing a 'rhythmic dance'. This is the dance of the DANCER, GOD SHIVA.

The whole universe is filled with subtle cubes of energy called KARU or BINDU or FOETUS.

---

* Brahman as defined by Mayan, in his astronomical treatise called 'Surya Siddhanta'.

Figure 1: Science of manifestation
(Courtesy: Pranava Veda by Mamuni Mayan – 12000+ years back, recreated by Dr. Ganapathi Sthapati)

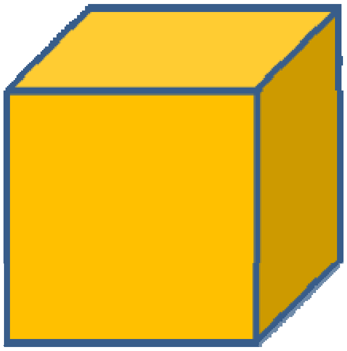

Figure 2: A cube with all equal sides

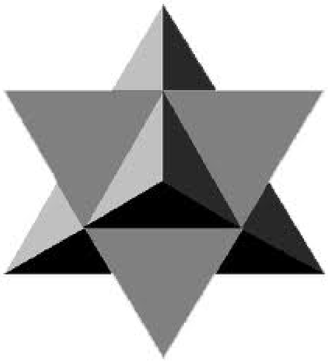

Figure 3: A tetrahedron (from Internet)

- The free space is packed with cubical atoms of energy. They are building blocks of the structure of the universe. **So what is clear is that all these atoms existed in a "resonant state of non-vibration" before the so called "big bang" happened or the pulsation started.**

- **It did not burst out of a single point as popularly believed. Alternatively, in space every point is the center. When the pulsation started these cubical atoms underwent a shape change and manifested as shapes and forms.**

- This cube is called as **micro-abode called Sittravai / Chittrambalam** (mini hall in Tamil). This is **secret in Chaidambaram, Sage Appar** (**Chidambara ragasyam**) realized. Chidambaram is a Shiva temple associated with Space energy and when you go to sanctorum you would not find any idol but just empty space with some lights there. This is to signify this concept of **Paramanu** which is **luminescent and is a Cuboid in shape**. This is the smallest particle possible and can be called as **'God's particle".**

- The micro space, in the cubical shape is fetus or Garbha. This is known as **Vinkaru** in Tamil. This micro abode is the repository of light and sound.

- The free space is Light and the Light is Moolam or source of the universe and universal forms. Light is Brahman, Atman, Vastu Brahman – Nunporul in Tamil.

In the subsequent blogs we will correlate String theory, Brahma Sutra, Nataraja, Vishnu and how they are all contained in order and proportion in pulsation.

Happy reading!

Ref: A good read **Fabric of the universe, Jessie J Mercay**.

8. Time = Space Series. Part 8 – Characteristics of the God's particle (Paramaanu)

This part is going to be a bit controversial hence read it with an open mind. Per our scriptures when analyzed deeply we can state that **"At the beginning it was just Consciousness (Pragnanam) that was luminescent with unmanifested potential energy and residing at resonant state of non-vibration. This energy was un-manifest and was present as Paramaanu or Microbode or Vinporul aka a subatomic particle. Yes! It was a particle without vibration and was at resonant state of non-vibration.**

The entire Space was supposed to be filled with these particles only and these were primal potential matter which is ever lasting.

Now let us ask few questions and see how our scriptures answer this:

1. What was the shape of this particle?

We have already answered that. In 2D it is square and in 3D it is a cuboid with all equal sides. Please note that in space it is always 3D and not 2D. It has 6 faces, 8 vertices, and 12 edges. This is known is current scientific community as **Tetrahedron**.

2. What is inside the Microbode or Subatomic particle?

This is a very interesting question. Our scriptures say that it is purely space inside but with an energy grid.

3. What is this energy grid and what is its shape?

It says that it is an 8x8 energy grid inside this microbode. This essentially means it is a square energy grid of 64 squares inside this Paramaanu.

4. What is this color of this Paramaanu?

It is luminescent and hence the golden color of the light is what it is associated with. **"Hiranyavarnnaam harineem**.." says Sri Suktha in the vedas, confirming the color of this Paramanu. Their variations exist but the fundamental one is of golden color.

5. What is this size of this God's particle?

It is calculated to be ~133 nm. We will see this in detail later.

6. What is the significance of the energy grid?

For the time being let me state that anything subtle is a function of 8 and anything gross is a function of 9. Let me also add that **iChing –** which is a Chinese divination book and the hexagram which deals with the Science of the subtle is in an 8x8 square matrix. **The mental game Chess is played in an 8x8 square.**

You would find 8 is the number or dimension for the subtle. We will see the importance of 8 and 9 and how beautifully these proportions adorn everything in nature in our daily life.

7. Are these particles faster than light?

Frankly I have not read a direct answer to this question but since this particle itself is light the velocity of this particle cannot be faster than light. This is my opinion. Modern experiments are suggesting that it could be faster than light. In **Mahabharatha, Yakshaprasna, the Yaksha will question this to King Yudhishtra as to "what is fastest thing in this world"**? King Yudhishtra would reply as **"mind"**.

Some of the sages have interpreted that since Consciousness is omnipotent and ever connected - the communication is seamless, instantaneous and not subjected to the concept of distance. The example given is if there is a pain in your leg, it does not take time to travel to your head. It is instantaneous and hence the whole universe should be considered as a living organism.

So to conclude on this question the particles are not faster than light but each particle is filled with intelligence and consciousness (just like a human cell) and their communication to anything is not subject to concept of distance. If our scientists come back and tell us that these particles are faster than light, then wait for someone else to tell us a little later that it isn't faster than light. ☺

8. Ok. Who has seen these particles and give me empirical proof?

परमानुरिति प्रोक्तं योगिनाम दृष्टि गोचरम ! **(Paramaanuriti proktam Yoginaam, drishti gocharam)** says **Mayamatham.** This means the **microspace** or **Paramaanu** is visible to the trained contemplatives and meditators only. One quick example as proof is the **Temple in Chidambaram** which represents **Space tattva** where the temple is built around the concept with enough documentation on it.

But we have been trained under Lord Macauly's education system and how can we accept this as proof which is subjective and superstitious. If I may ask please tell me how many of us have seen the electrons or protons? Do we not believe everything when someone tells us in English?

It is true that because of its existence we can see its manifestations and results of some experiments in our day today world. So we will look at tons of manifestations which we have been in nature as we progress. I would like us to be philosophically very clear as to what these things mean which we have been blindly following. For the time being please note down your questions and I am positive that this would be answered in due course.

Our sages have not only seen this, but have given the dimensions of this and the mathematical formula to manage this energy. We will see all of them. After reading all of them if you still don't want to believe in our scriptures, then may be CERN's result if it's in the same lines would prove it for you.

Happy reading!

உ

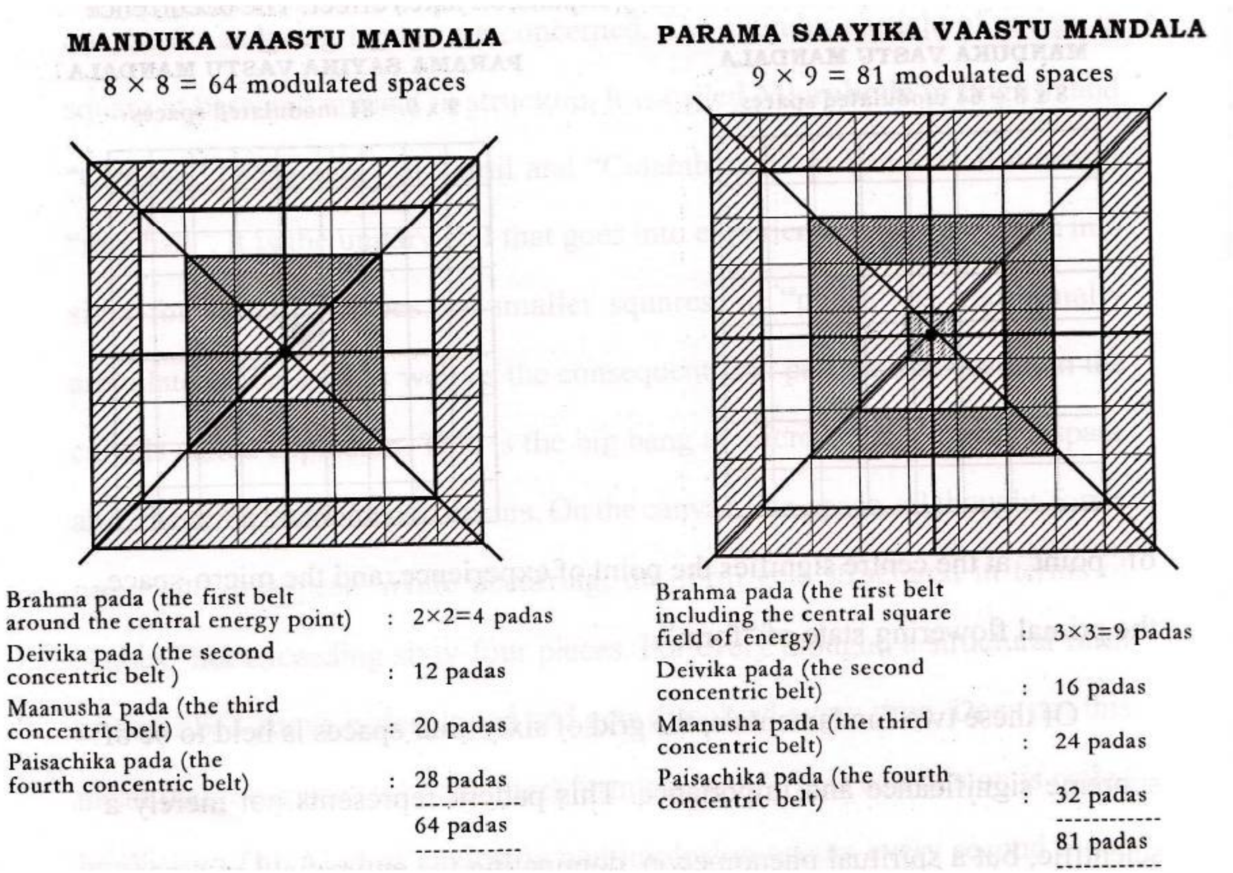

MANDUKA VAASTU MANDALA

8 × 8 = 64 modulated spaces

Brahma pada (the first belt around the central energy point)	: 2×2=4 padas
Deivika pada (the second concentric belt)	: 12 padas
Maanusha pada (the third concentric belt)	: 20 padas
Paisachika pada (the fourth concentric belt)	: 28 padas
	64 padas

PARAMA SAAYIKA VAASTU MANDALA

9 × 9 = 81 modulated spaces

Brahma pada (the first belt including the central square field of energy)	: 3×3=9 padas
Deivika pada (the second concentric belt)	: 16 padas
Maanusha pada (the third concentric belt)	: 24 padas
Paisachika pada (the fourth concentric belt)	: 32 padas
	81 padas

Figure 4: MANDUKA MANDALA – 8x8 (subtle form) & PARAMA SAYIKA MANDALA – 9x9 Gross forms
Courtesy: Dr. Ganapathi Sthapati

உ

9. TIME = SPACE SERIES. PART 9 – LIGHT, GOD’S PARTICLE AND GOD

“At the beginning it was just Consciousness (Pragnanam) that was luminescent with un-manifested potential energy and residing at resonant state of non-vibration. This energy was self manifest and was ever present as Paramaanu or Microbode or Vinporul aka a subatomic particle. This is **omnipotent, omniscient and omnipresent**.

Ignore religion for the time being and just look at this particle’s characteristics scientifically. It is Eternal (**Anaadi**), primordial (**Aadi**), without any quality (**Nirguna**), it cannot be reached / grasped by thoughts (**Achintya** - since thoughts as energy are much more gross than this and we would see this), it cannot be measured (**Aprameya**), Luminescent (**Jyothi swaroopa**), its **omnipotence, omnipresence (Visvam) and Omniscience** is already known. So said the mahavakya **Pragnanam Brahma!**

All the characteristics of the particle are attributed to a personality or a state by the founders of religions. **Hindus called this state as Paramasiva or Purusha. Hence everything subtle is called Sivamayam.** The essential pulsation which started and the universal evolution began was called as **Parashakti or Prakruthi or Maya.** We will talk about the Shakti aspect later.

This subtle un-manifest energy evolves to become universe which is gross. This gross energy also known as **earth** is called as **Vishnu** and the subtle energy is called as **Space / Akash / Heaven**. So the term became “**Sarvam Vishnu mayam Jagat”.** Please note that it is just **Sivamayam** and there is not Jagat there. Since it is all subtle and there is no Jagat which is gross. But it is **Vishnu Mayam Jagat**.

I need to add that every sect in Hinduism has named this God’s particle with a different name. Narayana has replaced paramasiva in Vaishnavism and so on. In Hindu temples there are 2 prime idols – **Moolavar and Archakar** - The Moolavar does not undergo any archana or any decoration and it is always for the Archakar for which all poojas are performed. This god’s particle refers to the Moolavar and not Archakar.

To my knowledge, every religion and every sage / saint has agreed with the fact this fundamental particle or energy unit is **LUMINISCENT**. Note that there is no sound or word in this …**yet.** Let us see few quotes on this (there are thousands of quotes):

- **Be it vedic culture, Zoroastrians, Greek or any major culture fire / sun god which represents light is worshipped.**
- **“Om Jyothi rasaha” says Vedas**
- **In Gayatri mantra you meditate on the supreme light to illuminate you.**
- Vallalar said **“Arutperumjyothi, Arutperumjyothi”** which means **mighty graceful light**.

- If therefore your eye be single, your whole body shall be full of light. Mathew 6:22
- Thirumoolar states brilliantly that who realizes that the light (physical light which means the world) merges with light (subtle light), he has realized SIVA.

  ஒளியில் ஒளிபோய் ஒடுங்கிய வாறும்
  தெளியும் அவரே சிவ சித்தர் தாமே - 124

So what has happened to the **Sabda Brahman** here? It has merged with the **Artha Brahman** which is the light. What can you use to represent this state?

Light represents Form and our eyes represent Light / Form. So obviously sound is represented by ears. When sound has merged with light then it means Eyes and ears are merged together.

Can you think of anything that represents this state? This concept is captured by representing a **Manduka (Frog) or Sarpa (Snake)** always. Since for a frog and a snake the ears and eyes are merged. **Manduka Upanishad,** represents a knowledge of this state. And obviously the 8x8 square / cube represent the most subtle energy as an energy grid is called as **Manduka Pada / Mandala..**

Now the same aspect when it is applied to a gross form / body or a building then it is a 9x9 energy grid and it is called as **Paramasyika Mandala**. **At the gross level the same concept as applied is the WORD where Light (form) and sound and meaning merged into one. You cannot separate them**. We will look at this later.

Unfortunately we have lost the science and we neither understand the philosophy nor the symbolism behind this supreme science but we made fun of the fact that the supreme science is named after an absolutely menial creature like frog which is used only for zoological lab studies across the world or ends up in the dinner plate as it does in far eastern countries.

I would like to conclude this section by stating (may be a little controversial) that **First it was Consiousness as the Vinkaru, Microbode, Sittrambalam, Nunporul, subatomic particle which was luminescent. Later the concept of God evolved out of it.**

So God is not light, but the right of saying it is "the primordial aspect of "luminescence" has been adopted as God…

So, of the 5 fold evolution process we are still at the source (Moolam). We will see the concept of time and the mathematical science in the subsequentparts.

MOOLAM (Source) – KALAM (TIME) – SEELAM (RHYTHM) – KOLAM (SHAPE) – GNALAM (UNIVERSE). (மூலம் - காலம் - சீலம் - கோலம் – ஞாலம்)

Happy reading!

உ

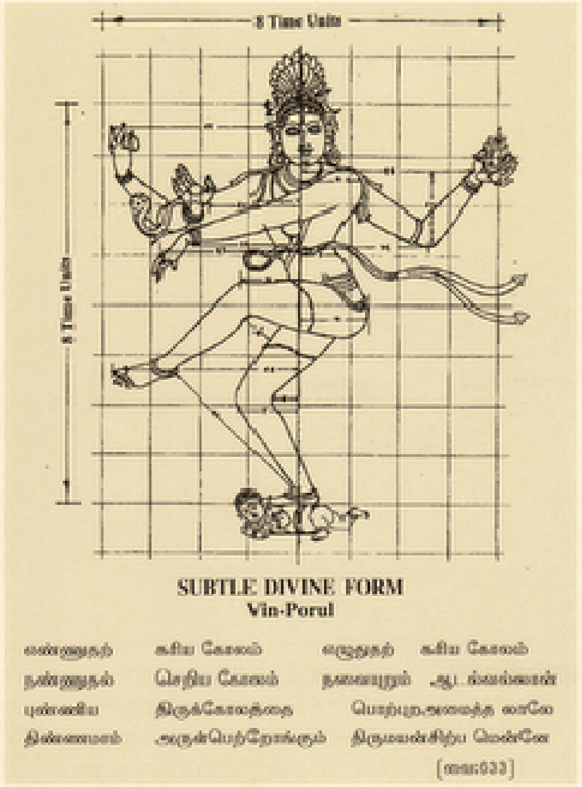
8 Time Units
8 Time Units
SUBTLE DIVINE FORM
Vin-Porul

Figure 5: Lord Nataraja in the Manduka Mandala (8x8 grid) (Courtesy: Dr. Ganapathi Sthapati)

Our next topic is Nataraja and String theory!, so check out this picture. I want you to note the light / fire in the left hand and sound / damru in the right hand. If you notice little closely, you would find a central line along the picture's sternum and all the parts have minute measurements from the central line.

10. TIME = SPACE SERIES. PART 10 – LORD NATARAJA AND THE STRING THEORY

“The self unfolds itself for its self – Bhagavad Gita”.

This self unfolded itself with a pulsation and this pulsation is the **start of Absolute time** (**KAALAM**). As I said nothing to my knowledge tell us clearly as to why this pulsation happened. This is the point of singularity in the scientific language and the event is termed as **Big Bang.** Just as I expressed earlier it was an event of primordial pulsation and not of an explosion which threw away mass because of the explosion into the space.

The entire space existed with its subatomic particles and this pulsation made the space to vibrate and this vibration was rhythmic and led to the evolution of this universe. **Popularly known as “That one became many”… and ..Vastur eva Vaastu.**

This pulsation happened because of the **Pranava** or the first pulsation resulted in the **Pranava -** both conceptually right. So at the time of this primordial pulsation 2 things happened:

- The absolute time started. This is called **Kaalam.**
- Primordial and the most sacred **Pranava** emerged.

Both these happened simultaneously and it is futile to ask whether **Kalam created Pranava or Pranava created the pulsation.** In many of the texts it offered as **Pranava that created the pulsation**. So let’s go by that.

- This is the reason that **Pranava / Primordial Om is so sacred in Hinduism.**
- This Pranava had manifested in 2 forms:
    - **Luminescent form as OM the Light. This is the source of all visual forms. The primordial form is represented as ஓம் in Tamil and ॐ in Sanskrit.**
    - **Aural form as Om the sound. This is the source of all aural forms. I should say that this is more popular form compared to the luminescent form.**

Now this concept is very important, since in the Pranava meditation they say us to meditate on the silence after the word OM and also mediate **OM as light, OM jyoti rasaha in vedas.**

So, when I say that sound merges into light, it is all the aural form merges into **ITS** luminescent form. All luminescent forms merge with the **Paramaanu** which is also luminescent. I am reproducing this passage again for emphasis and clarity.

- **Sage Thirumoolar** states brilliantly that who realizes that the light (physical light which means the world) merges with light (subtle light), he has realized SIVA.

  ஒளியில் ஒளிபோய் ஒடுங்கிய வாறும்
  தெளியும் அவரே சிவ சித்தர் தாமே - 124

I have read many times over to understand this concept and I would suggest that you allow this concept to sink in.

Now let's come to the 8x8 energy grid of the **Paramaanu** and imagine that it is pulsating because of Pranava and this **Pranava has both light and aural form**.

- They say that this **pulsation is rhythmic and symmetric**.
- When something is Symmetric, then there has to be a central line along with the Symmetry comes. For example if our right and left sides are symmetric then the body is divided along the sternum & backbone.
- Within the 8x8 energy grid / hall there is a **vertical luminous shaft** called **Brahma Sutra (Oli nool)**. Sutra in Sanskrit and Nool in Tamil means **Thread / String**.
- This **Olinool / Brahma Sutra is a shaft of consciousness also called Moolathoon / Moolasthambham**.
- **This shaft of consciousness vibrates in a particular order called rhythm.** This is the order of nature.
- **This rhythmic vibration of the shaft of consciousness is the dance of Shiva, the CELESTIAL DANCE / THE COSMIC DANCE / THE DANCE OF EVOLUTION of the LORD NATARAJA.**

Look again at the picture of LORD NATARAJA, he is drawn inside a 8x8 Manduka mandala (now you know what it is), his left leg is raised suggesting a movement from left to right. Imagine that he is going to land the left leg to his right side.

Why should he suggest that he is moving from left to right?

His left hand holds light / fire and his right hand holds Drum / Damru / Sound. The process of Evolution which starts with the absolute time and Pranava is fantastically represented pictorially. It suggests that the evolution process starts with the rhythmic dance of the PARAMAANU in a 8x8 energy grid and this process evolves first as light and then as sound. This light and sound creates further forms and the entire universe.

It is very easy to answer as to why sound evolves from light? We all know that the light is at a higher frequency and velocity than sound. This is the same reason we see lightning first then hear the thunders.

So let's summarize the DANCE OF EVOLUTION of LORD NATARAJA:

- From being un-manifest the **Pranava** emerges. Or the Pranava evolves the un-manifest to the manifest and starts the cycle of evolution.
- This leads to a pulsation which is called as the **COSMIC DANCE / CELESTIAL DANCE / DANCE OF EVOLUTION**.
- This pulsation is **rhythmic and symmetric**.
- The luminescent shaft of consciousness is called as **Brahma Sutra** or popularly called as String theory in the modern science.
- The process of creation evolves from the dance as light then sound evolves from light and then all the visual and aural forms in this universe.

If this is so then the process of dissolution should be that Lord Nataraja's right leg should be lifted suggesting a movement from right to left isn't?

Look at the majestic picture suggesting the **DANCE of EVOLUTION** and the **DANCE of DISSOLUTION** as below:

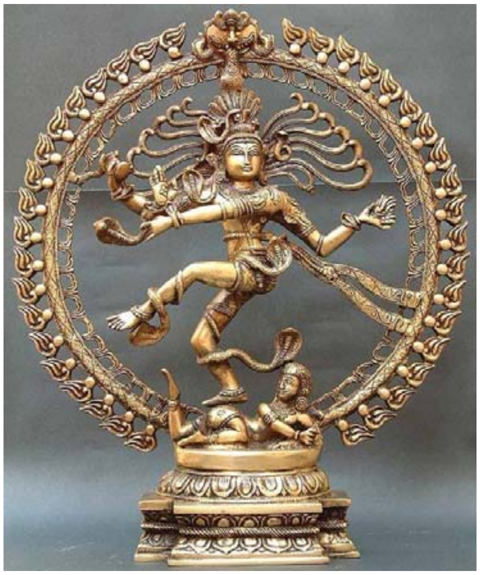

Figure 6: LORD NATARAJA / DANCE OF EVOLUTION

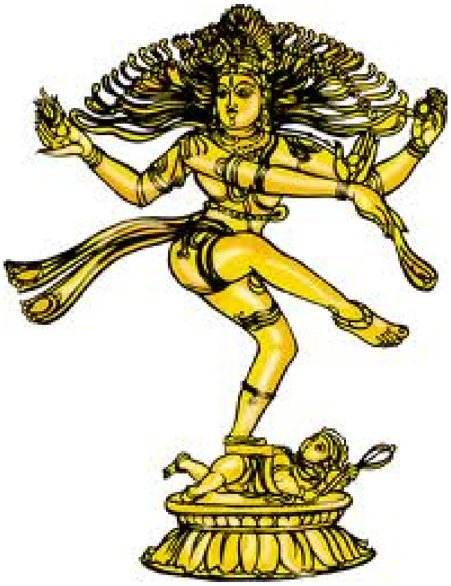

Figure 7: DANCE OF DISSOLUTION

In my opinion there is nothing more scientific in this world that is pictorially represented about the evolution of universe. You should read a book called **"The Tao of Physics" by Frijat Capra** where as physicist he explains how he experienced this and highlights the dance of the quantum particles in space corresponds to the dance of **Lord Nataraja**.

Now this central shaft of Consciousness is called **Brahma Sutra** and according to our scriptures every Paramaanu has this string, which means every being, matter, inanimate things in this universe including the tiniest of subatomic particles have this string / sutra. In the future parts we will see the Maths behind this sutra and the pulsation but let us look at what the modern science says about **Superstring theory** here, http://en.wikipedia.org/wiki/String_theory. You can read it how many times ever you want but this theoretical postulation is a direct adaptation of the cosmic dance and our Brahma Sutra and nothing more.

Now check out Lord Nataraja at the CERN http://www.fritjofcapra.net/shiva.html

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Now you may understand as to why I started this series with the above line. We have a very rich past that is most scientific and it is left codified in a form across thousands of temples in India. Let us just open your eyes and understand the scientific meaning around it.

If you are so far surprised at our richness and advancement, please wait till we will get into real hard core mathematics and geometry of our nature in the subsequent parts.

May that “**Sutradhaari”** dispel our ignorance and reveal to us the secrets of nature.

Happy reading!

11. TIME = SPACE SERIES. PART 11 – VASTUREVA VAASTU AND $E=MC^2$

We have seen that the universe in its un-manifest state is filled with luminescent cuboids and there was a trigger in the form of Pranava which started the absolute time and gave forth visual forms and aural forms.

This process of evolution is depicted very nicely in the following diagram:

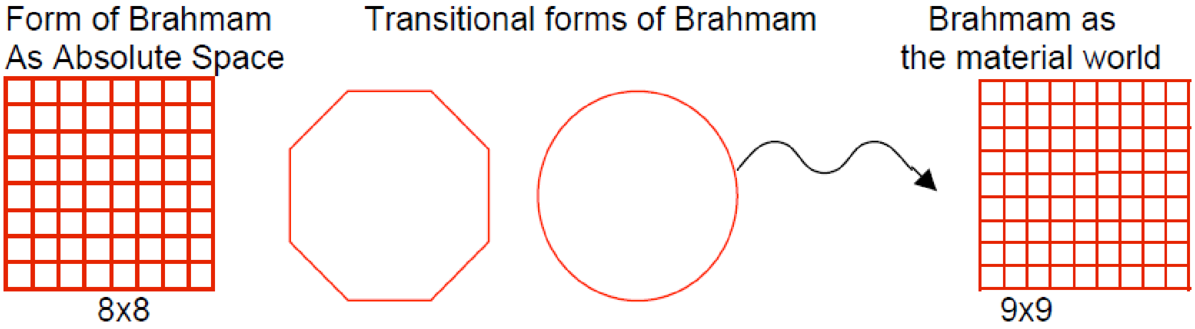

Figure 8: Transition forms of Brahman (Coutesy: Dr. Ganapathi Stapathi)

The square (in 2D) which is an 8x8 energy grid became a 9x9 energy grid which is manifest / gross. The transition forms were an octagon and a circle. **Aintiram** suggests how this transition takes place. Current physics accepts that a pulse and a sine wave can be created from each other and this happens because of simple addition of harmonic frequencies.

The primordial pulsation initiates a rotational force and this rotational force as it gains momentum converts the square into a circle and if you notice that the octagon is just an in between state between a square and a circle.

We see confirmation of the details of this manifestation process in Chapter 9, verse 8 of the Bhagavad Gita where Lord Krsna describes His manifestation process to Arjuna:
“prakritim svaam avastabhya visrijaami punah punah”, meaning “Curving back onto myself, I create again and again…”

What it suggests is that the essential shape of every gross form in its natural state is a circle (in 2D) and a sphere in (3D). **This is the reason we find every planetary body is almost spherical in shape and this is also the reason as to why the ripples in a water body when we throw a stone in the water are in circles.**

Be it the center of a tornado or the shape of a hurricane or a tsunami it always a circular force. The supreme secret in martial arts is that the Chi moves in a circular fashion and hence Tai Chi is the most deadly form of martial arts. Lord Vishnu has Chakra as his weapon to suggest not only light but also the rotational force.

The process of manifestation is captured in these 2 pictures at all the significant stages. If you notice that a single dot becomes multiple dots, then a line, then a pulse and then goes on to becoming a 8x8 energy grid and then to a 9x9 energy grid. 8x8 is un-manifest and 9x9 is manifest. These aspects are very well captured as sutras in the book **Aintiram and Pranava Vedam.**

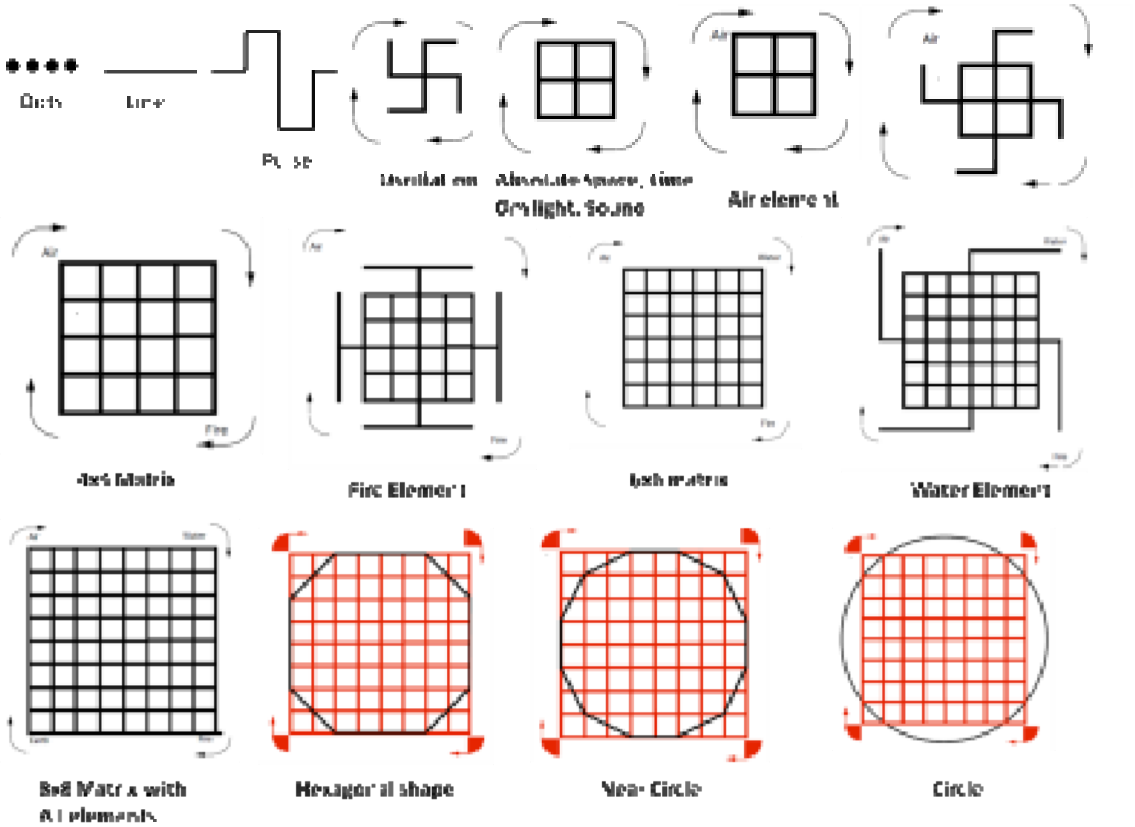
Dots
Line
Pulse
Absolute Space, time
Om light, Sound
Air element
Air
4x4 Matrix
Fire
Fire Element
6x6 matrix
Water Element
8x8 Matrix with All elements
Hexagonal shape
Near Circle
Circle

Figure 9: Transition forms of Brahman in detail (Coutesy: Fabric of the Universe)

If you look at the Fig 10 conceptually the following things happen:

- Subtle Energy becomes gross universe and the fundamental 5 elements are created in the process.
- It is said that 1/10th of the space becomes Air, 1/10th of air becomes Fire, 1/10th Fire becomes Water and 1/10th of Water becomes Earth.
- So we can conclude that we are living in a **Space delineated world.**

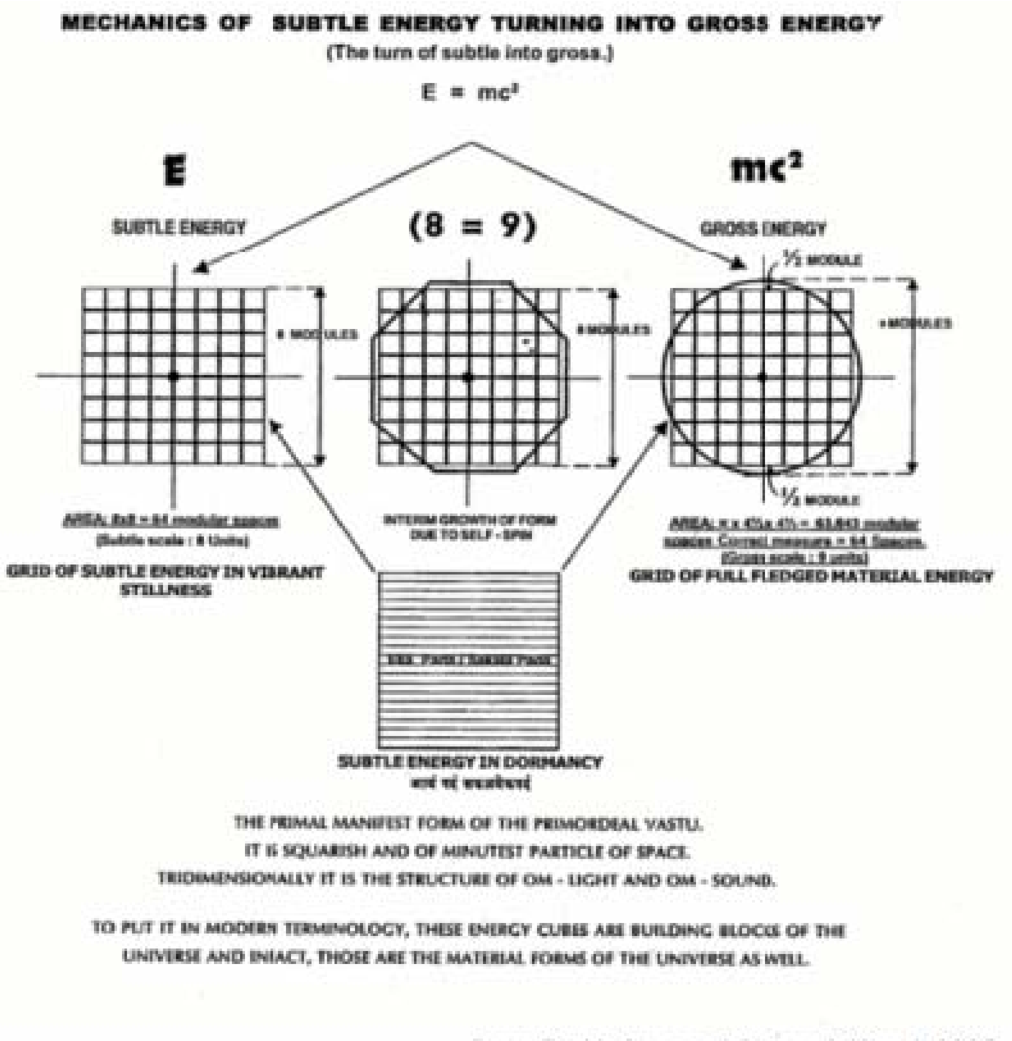
MECHANICS OF SUBTLE ENERGY TURNING INTO GROSS ENERGY
(The turn of subtle into gross.)
E = mc²
E
mc²
SUBTLE ENERGY
(8 = 9)
GROSS ENERGY
½ MODULE
8 MODULES
8 MODULES
9 MODULES
½ MODULE
AREA: 8x8 = 64 modular spaces
(Subtle scale : 8 Units)
GRID OF SUBTLE ENERGY IN VIBRANT STILLNESS
INTERIM GROWTH OF FORM DUE TO SELF - SPIN
AREA: π x 4½ x 4½ = 63.643 modular spaces. Correct measure = 64 Spaces.
(Gross scale : 9 units)
GRID OF FULL FLEDGED MATERIAL ENERGY
SUBTLE ENERGY IN DORMANCY

THE PRIMAL MANIFEST FORM OF THE PRIMORDEAL VASTU.
IT IS SQUARISH AND OF MINUTEST PARTICLE OF SPACE.
TRIDIMENSIONALLY IT IS THE STRUCTURE OF OM - LIGHT AND OM - SOUND.

TO PUT IT IN MODERN TERMINOLOGY, THESE ENERGY CUBES ARE BUILDING BLOCKS OF THE UNIVERSE AND INFACT, THOSE ARE THE MATERIAL FORMS OF THE UNIVERSE AS WELL.

From Dr. V. Ganapati Sthapati, Kauai, 2005

Figure 10: Subtle Energy to gross (Courtesy: Dr. Ganapathi Sthapati)

This wisdom was so succinctly captured by my good friend **Lao Tzu** in “**Tao Te Ching**” as follows: (Read heaven as un-manifest space and Earth as manifest gross)

“Heaven and Earth last forever.
Why do heaven and Earth last forever?
They are unborn, so ever living....

Thirty spokes share the wheel's hub;
It is the center hole that makes it useful.
Shape clay into a vessel;
It is the space within that makes it useful.
Cut doors and windows for a room;
It is the holes which make it useful.
Therefore benefit comes from what is there;
Usefulness from what is not there....”

But I wanted to highlight more important aspects to this process of manifestation.

- **Advaita** philosophy states that there is no duality. If you note that the **Parammanu** is what is present every living being and inanimate thing. What is within you is what is within me and hence the mahavakya **“Tat Tvam Asi”… You are that**.
- **Vastureva Vaastu** – Vastu which is unmanifest and subtle becomes Vaastu which is manifest and gross.
- **If we assume that Paramaanu is God’s particle and hence it is divine, then everything in this world is divine**. This is the most important philosophy in **Hindusim**. See **the oneness in all** and see “**All in one”**.
- That’s why Hindus saw divinity in everything. Trees, snakes, animals, insects, birds, human beings etc. There is nothing without that divine energy or particle – however you call it.
- This philosophy is **UNIVERSAL, OPEN SOURCE, ALL COMPASSING and the highest state of consciousness**.

But let’s look at something even more important:

- We all note that Subtle has become Gross. Hence conceptually Gross = Subtle or both are proportionate.
- We all note that Space has become Earth. Hence conceptually Space = Earth or both are proportionate.
- We all note that 8 have become 9. Hence conceptually 8 =9 or both are proportionate.
- We all note that ENERGY has become MATTER. Hence conceptually ENERGY = MATTER or both are proportionate. This is represented by $E = mc^2$ by our great scientist Einstein. While we did not have this specific formula, the concept and the associated texts were very much with us for ages.

So this is what Einstein quoted after reading Bhagavad Gita.

"When I read the **Bhagavad**-**Gita** and reflect about how God created this universe everything else seems so superfluous." ~ Albert **Einstein**

Now let’s try to answer this question:

- When you throw a stone in water / pond the ripples that form take the shape of a circle (in 2D) / sphere (in 3D). Can you guess what the shape of the ripple would be if you throw the same stone in space?
- **You would have probably guessed the answer. The ripples in space should be and would be in square /cube in shape.**

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

12. TIME = SPACE SERIES. PART 12 – VEDIC CONCEPT OF TIME & ITS MEASUREMENTS

The importance of time (**Kaalam**) is so high in the Vedic tradition that they ensured that the accuracy, continuity of time and the easy measurability of it is maintained for who knows how many years. **Time creates, sustains and destroys** everything and this is very nicely captured in these 2 lines:

காலமே எவைக்கும் மூலம், காலமே மூல ஆற்றல்
காலமே கோலத் தோற்றம், காலமே ஞாலத் தோற்றம்

The god of time is **Kaala Bhairava** – a form of Shiva and hence both **Shiva** and **Yama the lord of death both were** known as **Kaalan.** The time is represented by **number 13** and it belongs to Kala bhairava. We will see the interesting linkage between Vedic system and the Mayan calendar found in Guatemala and Mexico later.

Swami Chinmayananda states that "**Time is the periodicity between 2 thoughts**". Nothing captures the concept of time more wisely. What is clear from our scriptures and Swami Chinmayanda is that time is the pulsation that creates this universe. It is the vibration with very accurate periodicity and our sages suggested that you can live beyond time provided you manage to escape the effects of that pulsation.

Einstein's view on time and its relativity is very realistically captured in this quote **"When you are courting a nice girl an hour seems like a second. When you sit on a red-hot cinder a second seems like an hour. That's relativity."** More than relativity I found that the root cause of relativity is the pulsation of mind as time. Einstein on a serious note proved that our view of time is not absolute, it is relative and it is a continuum with space. We will in the next few parts of this blog would see that time and space are the same or Time = Space.

Aintiram states that there are 5 things that vibrates which have a common root and are not under our control. They are Space, Earth (vibration of earth is gravity), Eye lids, thoughts and Musical notes and very nicely put in Tamil as (Vin)**விண், மண்(man), கண்(Kan), எண்(Yen), பண்(Pan). Pan is musical metre / chandas and Yen is the yennam or thoughts.**

Correlate this with the first line of **Patanjali Yoga Sutra** which says **"Yoga Chitta Vrtti Nirodah",** means **Yoga is one which stops the vibration of mind / pulsation of mind and hence is a killer of thoughts.** The meditation techniques of keeping your eyes opened without battling the eyelids is another way of overcoming this pulsation instinct that is primordial and inherent in our nature.

Let's look how intricate our measurements time are in Vedic system. Check out the image below.

- The smallest unit of time is approximately 2 micro second.
- The largest unit is about 155 trillion years…. YET.
- Brahma's linear age is 50 years and in currently in his 51st year which makes the universe to be 155+ trillion years. Brahma's life time is 100 years which means it would be 300 trillion+ years of universal life which follows a cycle.

- After 100 years of pulsation cycles Brahma stays as dormant energy for the same period and then pulsates again for the next life of 100 years.
- A **Brahma's** life is also known as a Para. Each half Param is referred to as a *parardham*. This second half is termed as **'*dviteeya paraardhe*'** (the second half of Brahma's term) is stated in all vedic rituals.
- After every Chatur yuga, the world recycles. After every 2000 years / a Brahma day the world dissolves and remerges. The important point to note is that there is no CONCEPT OF CREATION here. It evolves and dissolves as per the science of pulsation. We also note that this pulsation expands and contracts. This cyclical nature of time as believed in Indian mythology refers to time as '*anaadi*' or that without a beginning.

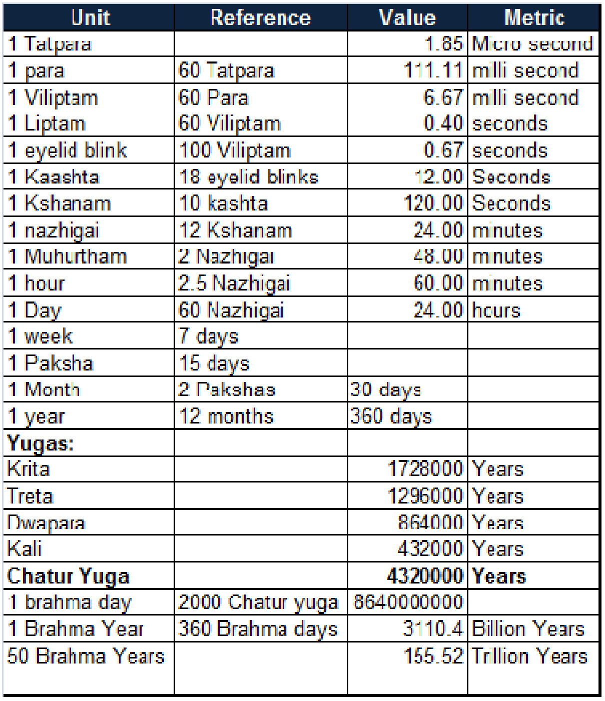

Unit	Reference	Value	Metric
1 Talpara		1.85	Micro second
1 para	60 Tatpara	111.11	milli second
1 Viliptam	60 Para	6.67	milli second
1 Liptam	60 Viliptam	0.40	seconds
1 eyelid blink	100 Viliptam	0.67	seconds
1 Kaashta	18 eyelid blinks	12.00	Seconds
1 Kshanam	10 kashta	120.00	Seconds
1 nazhigai	12 Kshanam	24.00	minutes
1 Muhurtham	2 Nazhigai	48.00	minutes
1 hour	2.5 Nazhigai	60.00	minutes
1 Day	60 Nazhigai	24.00	hours
1 week	7 days		
1 Paksha	15 days		
1 Month	2 Pakshas	30 days	
1 year	12 months	360 days	
Yugas:			
Krita		1728000	Years
Treta		1296000	Years
Dwapara		864000	Years
Kali		432000	Years
Chatur Yuga		**4320000**	**Years**
1 brahma day	2000 Chatur yuga	8640000000	
1 Brahma Year	360 Brahma days	3110.4	Billion Years
50 Brahma Years		155.52	Trillion Years

Figure 11: Time measurements in Ancient India

While doing Vedic ritual the time and space coordinate is invoked to say that you are doing this ritual at this place in the universe and at this time – with reference to Brahma’s age to the time and star of the day. This tradition is unbroken and the Indian Almanac can predict the position of astronomical objects without any error for thousands of years without any sophisticated instruments and space ships.

On the contrary just read the history of the Gregorian calendar which we are following here, http://en.wikipedia.org/wiki/Gregorian_calendar. There is neither accuracy nor continuity and to top it up there are authorities who reform the calendar as if time stops and waits for them to affect that reform. Why should we ignore the best of time science that we have for an inaccurate, discontinuous reformed calendar?

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy Reading!

13. TIME = SPACE SERIES. PART 13 – SOUTH AMERICAN MAYAN & SOUTH INDIAN MAAYAN

Since we are discussing about time, let us address the **Mamuni Maayan** and the **South American Mayan**.

A lot has been talked about Mayan calendar and the imminent destruction to planet earth in the year 2012. I would recommend **'The Mayan Factor – Path beyond technology" by Jose Arguelles** if you are interested in this topic.

In this part I am interested in highlighting that the **South American Mayan** and the **South Indian Mamuni Maayan** who wrote **Aintiram and Pranava Veda** are one and the same. Extensive research have been conducted by many and I found, both the works of **Jose Arguelles and late Dr. Ganapathi Stapathi** who has spent about 3 decades each, very impressive. Dr. Ganapathi concluded that they both are the same and he has shown some interesting linkages.

I am going to present what impressed me in both these great researcher's works.

- First let me state that the doomsayers of 2012 phenomenon picked up Mayan's calendar which started in 3113 BC ran for 5125 years and the calendar ended in year 2012. Since the calendar ended they thought this signified that the world is ending. Jose argues that this is wrong and most probably the human race is entering into the next phase of evolution and the next cycle would begin. If you strongly believe that this is not true and the world would certainly end by 2012, then please ensure that you transfer all your assets and savings to my name effective 22, Dec 2012. ☺

- Mayan in his calendar in South America used very unique numerals in his calendar that is given below:

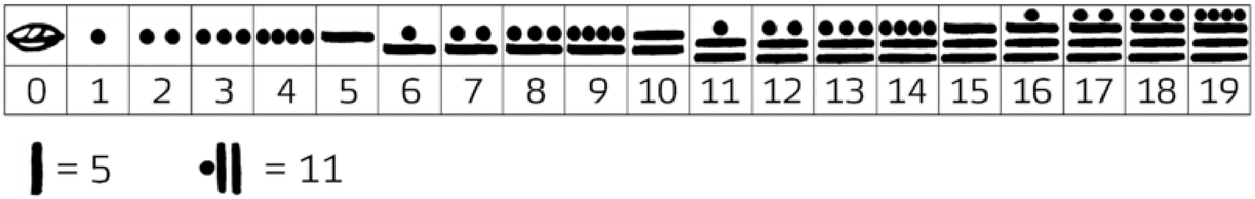

Figure 12: Mayan's numeral system (Courtesy: Wikipedia, Internet)

Now let me reproduce a Tamil poem written by Mamuni Maayan in his work called "**Kanithamaa sennool**". This song is part of the book written by Maayan in Tamil during his time – 12000 years+. This song expresses how the numerals are formed till 10. It says,

  - 1 is one dot, 2 is 2 dots, 3 is 3 dots, 4 is 4 dots, 5 is a line and 10 is a double line.

மாப்புள்ளி ஒன்றே அன்றோ ஒன்றுமா எண்ணே காட்டும்

உ

மாப்புள்ளி இரண்டே யன்றோ இரண்டுமா எண்ணே காட்டும்
மாப்புள்ளி மூன்றே யன்றோ மூன்றுமா எண்ணே காட்டும்
மாப்புள்ளி நான்கே யன்றோ நான்குறு எண்ணே காட்டும்

மாப்புள்ளி ஐந்தே யன்றோ ஐந்து மா எண்ணே காட்டும்
மாப்புள்ளி கோட்டினாலே இரண்டுரு மாக்கங் காட்டி
மாவியல் திறத்தினாலே பத்தறு மாக்கங் காட்டி
மாவியல் திறத்தினாலே எண்முதற் பொருளு மாவாய்!! (கணிதமா செந்நூல் - மாயன்)

- Among the books written during Mayan’s time in Guatamala are Popol Vuh, The annals of cakchiquels and the **Book of Chilam Balam**. The resemblance of **Chilam Balam to Chidambaram** with which S. Indian Maayan was so involved is very striking.

- Be the pyramids that were built in the jungles of S. America that has a striking resemblance to the viamanas from south India, the earth’s revolution around the sun measured within 1/1000th of the decimal point which is available in a book called **Surya Siddhaanta – In Sanskrit** by Mayan and the identical measurements available in South America – the similarities are difficult to ignore.

- Jose Arguelles makes a compelling argument about the shift in consciousness in 2012 due to a major galactic synchronization. But what impressed me are the following:

    - Comparison between **iChing (8x8 hexagram) and Mayan’s Tzolkin which is 13x20 matrix**. In fact there is a Franklin’s magic square which is based on the 8x8 grid of iChing that adds upto 260 which is the galactic constant of Tzolkin. This was amazing.

    - He refers to **Radiogenesis – Universal transmission of information through or as light** or radiant energy which is very much relevant to our discussions of light and sound.

    - The revelation he received from a Maya Humbatz men belonging to the universal great brotherhood organization revealing that our universe is the 7th such system which Maya has charted in the universe. So, we are not alone ☺

Would conclude this topic in the next part…

Happy reading!

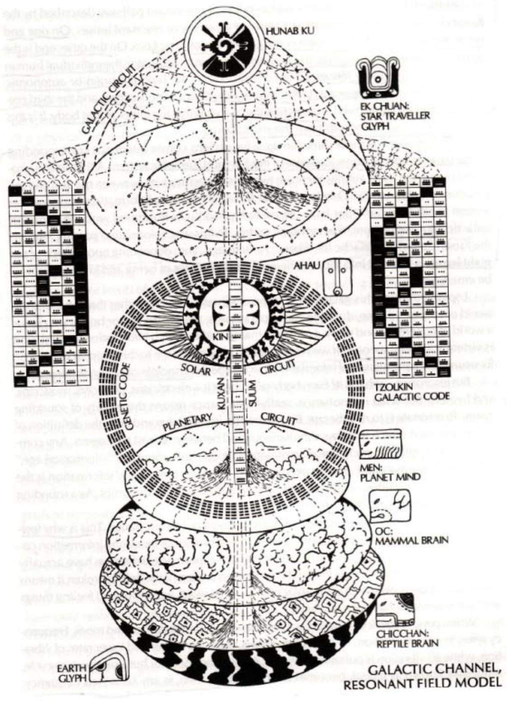
HUNAB KU
EK CHUAN: STAR TRAVELLER GLYPH
GALACTIC CIRCUIT
AHAU
KIN
SOLAR CIRCUIT
GENETIC CODE
KUXAN SUUM
PLANETARY CIRCUIT
TZOLKIN GALACTIC CODE
MEN: PLANET MIND
OC: MAMMAL BRAIN
CHICCHAN: REPTILE BRAIN
EARTH GLYPH
GALACTIC CHANNEL, RESONANT FIELD MODEL

Figure 13: Galactic Communication Channel to Hunab Ku (Courtesy: The Mayan Factor, Jose Arguelles)

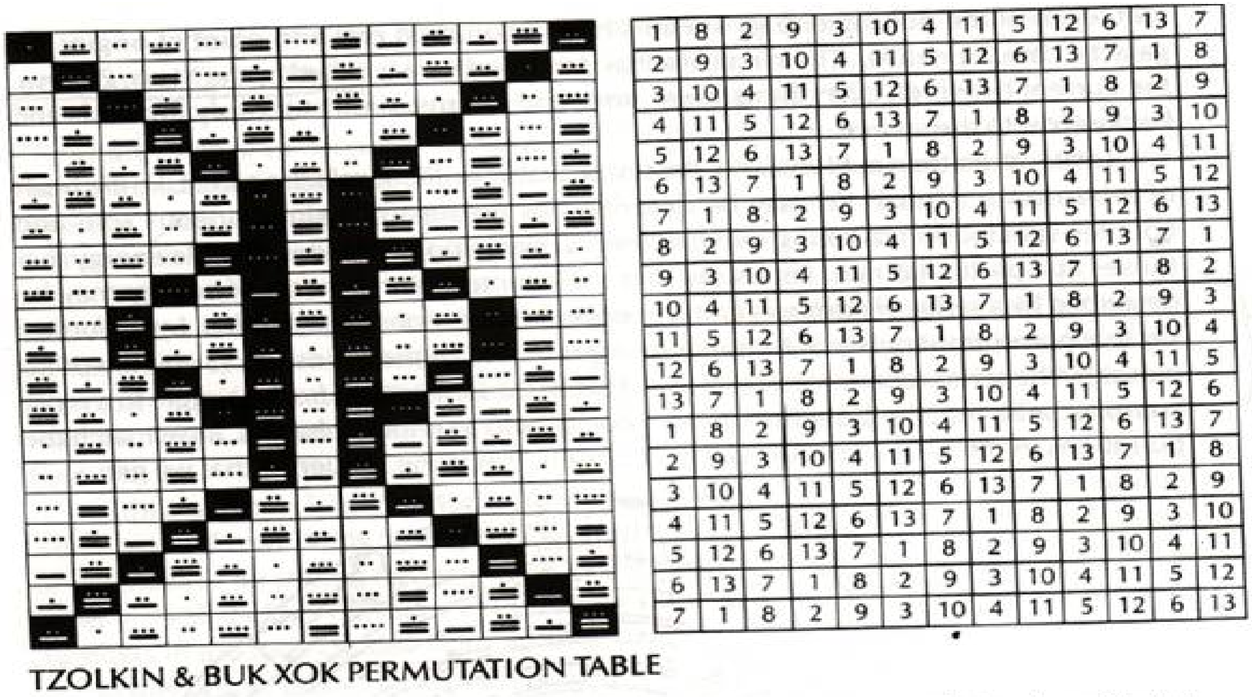
TZOLKIN & BUK XOK PERMUTATION TABLE

1	8	2	9	3	10	4	11	5	12	6	13	7
2	9	3	10	4	11	5	12	6	13	7	1	8
3	10	4	11	5	12	6	13	7	1	8	2	9
4	11	5	12	6	13	7	1	8	2	9	3	10
5	12	6	13	7	1	8	2	9	3	10	4	11
6	13	7	1	8	2	9	3	10	4	11	5	12
7	1	8	2	9	3	10	4	11	5	12	6	13
8	2	9	3	10	4	11	5	12	6	13	7	1
9	3	10	4	11	5	12	6	13	7	1	8	2
10	4	11	5	12	6	13	7	1	8	2	9	3
11	5	12	6	13	7	1	8	2	9	3	10	4
12	6	13	7	1	8	2	9	3	10	4	11	5
13	7	1	8	2	9	3	10	4	11	5	12	6
1	8	2	9	3	10	4	11	5	12	6	13	7
2	9	3	10	4	11	5	12	6	13	7	1	8
3	10	4	11	5	12	6	13	7	1	8	2	9
4	11	5	12	6	13	7	1	8	2	9	3	10
5	12	6	13	7	1	8	2	9	3	10	4	11
6	13	7	1	8	2	9	3	10	4	11	5	12
7	1	8	2	9	3	10	4	11	5	12	6	13

Figure 14: Tzolkin table (Courtesy: The Mayan Factor, Jose Arguelles)

14. Time = Space Series. Part 14 –Mayan, Tamil, Pleiades & iCHING

Jose Arguelles was also mentioning that the pyramid structures of Mayan civilization in South America with the **Tzolkin** calendar matrix communicates with 2 constellations – **Pleiades and Arcturus**. Jose mentions that he is sure about communication with Pleiades constellation.

Why did Pleiades interest me?

- According to Tamil history, Tamil is a divine language and the Lord associated is **Lord Muruga**. He is the Tamil deity and he is from the **Constellation Pleiades** which is called **Kritika nakshatra**.
- This emphasis that the South American Mayan could be a Tamilian from India, as many in the ET studies circuit opine that Pleiades is a constellation which is like our distant cousin where Tamil is spoken. There are many links but you can check this out.. http://22050hz.blogspot.com/2008/10/pleaidians.html

- Look at the Mayan calendar in Fig 14, it is a 13x20 Tzolkin matrix.
    - Please recall that we mentioned **13 is the number for Kala Bhairava**.
    - Tzolkin has a mystical central column around which 6 and 6 columns exist. The belief is that this central column is the frame of reference to the galaxy and it is believed to connect to **Hunab Ku – the center of the galaxy**.
    - This mysterious centre column is believed to be the road to the sky leading to the umbilical cord of the universe. I immediately correlated this with the **thread of consciousness** which is connected to the Hiranyagarba (the center of the galaxy as per Indian texts.) – Check Figure 13.
    - So understanding that this shaft of consciousness is connected to the center of the galaxy (**Hunab Ku / Hiranyagarbha**) at the one end and to **our solar plexus** at the other end with communication and galactic information transmission happening through resonance of light, opened many gates within me.

Nothing in this universe exists without purpose and the form and shape of any being / thing is explained as a state of consciousness at a particular resonant frequency adds to the Seelam (Rhythm) aspect of the Kaalam which we are going to see later.

For the space and time travel enthusiasts, Mayan has the following:

- All space travel is intelligence as information – transmitted through light by the principle of harmonic resonance.

- **Mayan says that Sun is the lens through which galactic information is transmitted from the galactic core and interpreting this with Gayatri Mantra where we meditate on the supreme light Savitr – an aspect of Sun god is amazingly striking.**

- Mayan highlights that time is a function of principle harmonic resonance and he attaches significant importance to the numerals which he states are the galactic harmonic constants. This is exactly what **Aintiram** states.

- **He interprets the iChing to contain the Genetic code of the human DNA as a hexagram in a 8x8 matrix which is a binary progression to the 6th power corresponding 64 six part words /codons. At the same time he interprets number 260 to be galactic constant and 360 which is the factor for a calendar year or the total angle in a circle to a harmonic calibrator.** His explanations more on this and on the mystical number 7 do not resonate in my tiny brain and I am begging for holy grace to expand my mind to understand these concepts.

- **When he says that flowers open to light just does not mean that it is sensitive to be light but deeply it aspires to be LIGHT.** I am not sure if you understand the depth of this statement as against the Sage molar statement which states gross light should merge with subtle light. I have been left speechless for days after understanding this correlation.

It is interesting to note that significance of numerals as per Mayan factor:
- **1 – The pulsation – Ray of Unity**
- **2 – The pulsation - Ray of Polarity**
- **3 – The pulsation - Ray of Rhythm**
- **4 – The pulsation - Ray of Measure**
- **5 – The pulsation - Ray of center**
- **6 – The pulsation - Ray of organic balance**
- **7 – The pulsation - Ray of mystic power.**
- **8 – The pulsation - Ray of harmonic resonance.** We will see more of this.
- **9 – The pulsation - Ray of cyclic Periodicity**
- **10 – The pulsation - Ray of Manifestation**
- **11 – The pulsation - Ray of Dissonant Structure**
- **12 – The pulsation - Ray of Complex stability**
- **13 – The pulsation - Ray of Universal movement**

In all 2 independent researchers working 6000 miles apart, conducting their studies independently seems to be pointing to just one person and one philosophy and science. What else we could do other than thanking our fortune to have read them and made this connection. I thought of writing just a page on this topic and I have ended up with 2 parts and I am stopping my temptation to write few more pages on this topic. For all of those interested, please read the book by Jose Arguelles.

It is amazing to find out how most of the ancient civilizations seem to have a common root and the communication between various geographies looks stunningly simple and the metrics on astronomical objects and other space sciences seems to be so accurate despite the absence of precision instruments and spacecrafts. One tends to believe that we had far advanced knowledge of nature which we seems to have lost and most certainly few thousand years back we were not nakedly roaming in the jungles of India and Africa hunting animals as the western experts wants you to believe but had a far superior & scientifically advanced civilization.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy Reading!

15. TIME = SPACE SERIES. PART 15 – SHIVA LINGAM IS NOT A PHALLUS

If there is one part in the series I want you to read and talk about it with your friends then it is this topic. If I ask you as to what Shiva Linga stands for then more than 99% of the responses would be that it represents a **Phallus** symbolizing the male sex organ / creative energy with the female sex organ / creative energy - Yoni as the base and both together representing **inseparability and totality in the process of creation**.

Nothing is more misunderstood and misquoted than this concept of Hinduism. Every idiot including me who has read few books about Hinduism thinks that they are experts of Indology / Vedic studies. The wisest of sages who gave us Vedas and Upanishads did not boast that it is their work. At the end of every work they acknowledged that they are telling us what they have been told by their ancestors. It is 155 trillion+ years of universal knowledge available today in the richest of the languages Tamil and Sanskrit. This cannot be accurately interpreted in a foreign language. Lest in the weakest of all languages - English.

The other contributing factor to this misunderstanding has been the Sanskrit language or should I say the richness of the language.

We have already noted that the same word can have multiple meanings. This is not limited to Sanskrit, but the all the languages. For example,

- **Sutra means thread, Aphorisms** (the ability to express something in terse and concise manner). Ex. **Kama sutra** – the most understood of all the sutras. **Brahma Sutra** – The thread of consciousness in any form that runs in this universe – the most unheard of.
- **Artha** in Sanskrit means "meaning" and "wealth", we have already seen this. As in Vaak-Artha and Artha Sastra (the science of wealth management).
- Same way Lingam means **Symptom, Proof**, **Gender, Male sex organ** etc. Hence it was very easy to get confused, associate the lingam with a Phallus and also to ignore the philosophy and the supreme science behind it. Even people who believed that it represented much more did not knew the philosophy and the science behind it. Let's look at what it means.

I am going to request you to kindly refer to both the pictures below:

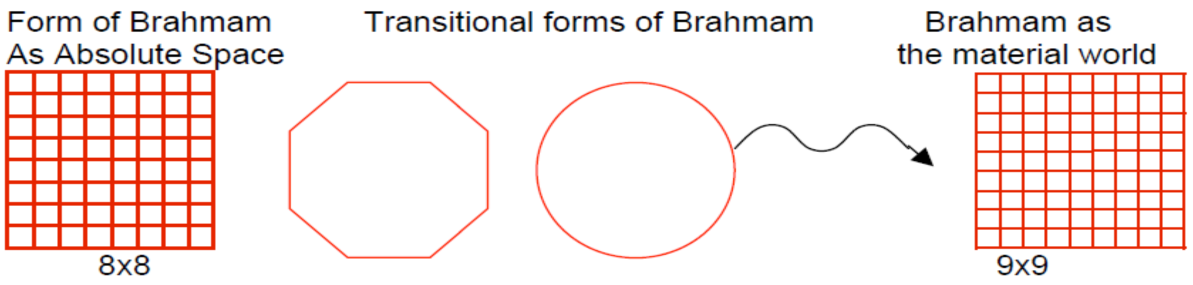

Figure 8 – reproduced here for convenience

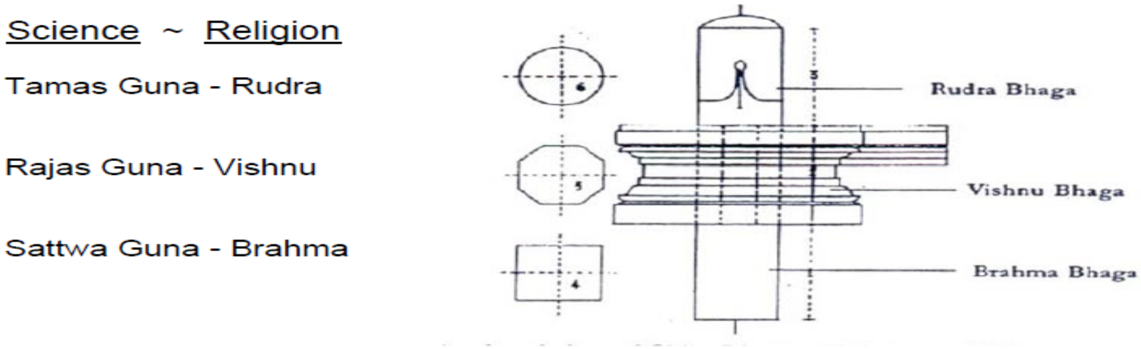

Figure 15 – Shiva Linga – Shape and Significance (Courtesy; Dr. Ganpathi Sthapati)

We have already seen that the subtle energy is takes the form of a square and the gross energy takes the form of a circle with an in-between state it takes the shape of a **tetrahedron and octagon**. Now look at the Shiva lingam from the top. Its base is always a square and the top is always a circle.

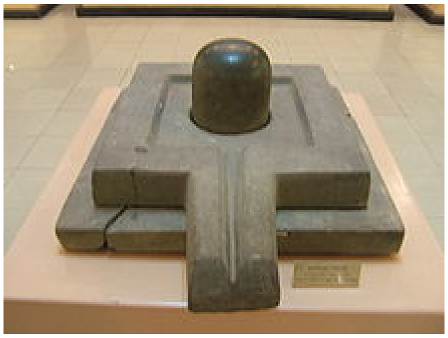

Figure 16 – Shiva Linga

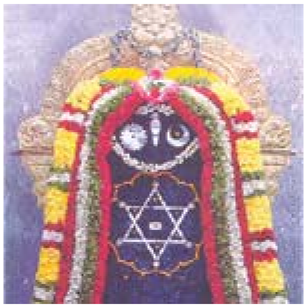

Figure 17 – Shiva Linga with shape of the God's particle (Star) engraved.

If I have to merge all the 3 shapes together and it naturally takes the shape of a Lingam with a round top. **Shiva Linga captures the overall process of evolution in one form and shape - subtle to gross state – it encompasses every state of consciousness that is possible in this universe.**

Why the figure above states that Square shape is Sattvic, circle is Tamasic and the in between state is Rajasic? Let's understand the definitions of all the three qualities.

- **Sattvic represents the perfect state of balance. When it is subtle it is luminescent and in a perfect state of resonant non-vibration. The energy levels were perfect so that the resultant vibration is nil and it is in absolute balance.**
- **Tamasic state represents a state where the height of evolution is reached and hence it is farthest away from the core.** This is a different explanation from the usual one available in the web and books. Let me explain this. The core is very subtle and the tamasic state is very gross

and hence it is away from the core. Tamasic state then should be considered as the darkest state possible. Why?

  - Because the core is luminescent and if the gross state is farthest from the core hence it should be dark. This is the meaning behind the most famous statement **'Tamasoma Jyothir gamaya"…** take me from darkness to light. It is not about just your mind but about the entire being aspiring to become a light and merge with primordial light.
  - **There is no evolution possible beyond the Tamasic state and it has to devolve to reach its core / subtle state.**

- **Rajasic is a state where the pulsation has resulted in a rotational force and hence things are set into action. This represents the evolution process and hence associated with action.**

Then the natural next question is, if Shiva Lingam is associated with the entire spectrum of consciousness then we should see it associated the elements and the light. Isn't?

- There is a story in Indian Purana that Brahma, Vishnu and Shiva discussed who is the most powerful among them. Lord Shiva challenged Brahma and Vishnu to reach his head and feet respectively. Brahma tried to find Shiva's head and Vishnu tried to reach Shiva's feet, both of them failed. Most of us would have heard the story but would not have understood the philosophical and scientific significance.
  - Lord Shiva in the form of Linga stood there as a **PILLAR OF LIGHT**. This pillar of light is **the Brahma Sutra / light of consciousness** which spanned from the center of the galaxy (Hiranyagarba) to every being & non-being in the Universe.
  - Brahma could not find the head because; Brahma is always associated with intelligence in Indian texts. He gave the supreme wisdom - the Vedas and stands for Intelligence. He could not reach Shiva's head signifies that Intelligence will not take you to God. Intelligence as an energy is too gross and cannot reach the most subtle state of light.
  - Vishu represents earthly life – Rajasic in character – this is the tetrahedron & octagon state. He could not reach Lord Shiva's feet signifying you cannot reach the core through actions.

- Lord Shiva in the form of Lingam is represented as one of the elements (Earth, Water, Fire, Air, and Space) in the Hindu temples.

If you read the following Upanishad quote, you can probably understand the meaning:

na karmaNaa na prajayaa dhanena tyaagenaike amRitatvamaanashuH
pareNa naakam nihitam guhaayaaM vibhraajate yadyatayo vishanti

Not by work, not by progeny, not by wealth, they have attained Immortality. Some have attained Immortality by renunciation. That which the hermits attain is laid beyond the heaven; yet it shines brilliantly in the (purified) heart.

And this quote also brilliantly articulates the luminescent nature of the Paramaanu:

na tatra (not there) **suryo bhati** (sun shines) **na candra** (moon does not) **na tarakam** (nor the stars)
nema vidyuto (nor the lightning) **bhanti** (shines) **kuta ayam agnih** (how can fire shine)
tam (that) **eva** (alone) **bhantam** (shines) **anubhati** (thereafter shines) **sarvam** (everything)
tasya bhasa (in that light) **sarvam idam vibhati** (everything shines)

The sun does not shine there, nor does the moon, nor the stars or the lightning, much less this fire (deepam). When He shines, everything shines after Him; by His light alone everything is illumined.

The trick is to take the definition & meaning from Upanishads or any scripture and apply it to the God's particle and the process of evolution and you would understand it perfectly. Many a times the concept of God & faith obscures our mind to look beyond and rationally think about it. It is not the problem of religion but our system of education which has ruined our thinking. **Today the biggest challenge we have is to "Unlearn" than to learn.**

So to conclude:

- Shiva Lingam does not represent Phallus. To represent male and female totality and inseparability Hindus do have **Arthanareeswara** and it is not represented through Lingam.
- Shiva Lingam represents the **Pillar of Light and the entire spectrum of consciousness**. From the most subtle to the most gross.
- It represents an in-between **Guna** / **EGG** state (between un-manifest and manifest) and hence it is formless with a form.
- **While Lord Nataraja represented the rhythmic pulsation of this Brahma sutra, the cosmic dance / the dance of the evolution, Shiva Lingam represents the primordial light itself in all its state of consciousness.**
- By this analogy, our body itself is a lingam with the visible body as the gross part at one end and the most subtle body which is invisible and is part of the primordial light at the other.

Check out the picture below which represents most of the aspects we have discussed so far.

May Lord Shiva help us all to understand him better.. ☺

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions." Yet wish that this quote isn't wiser any more.**

Happy reading!

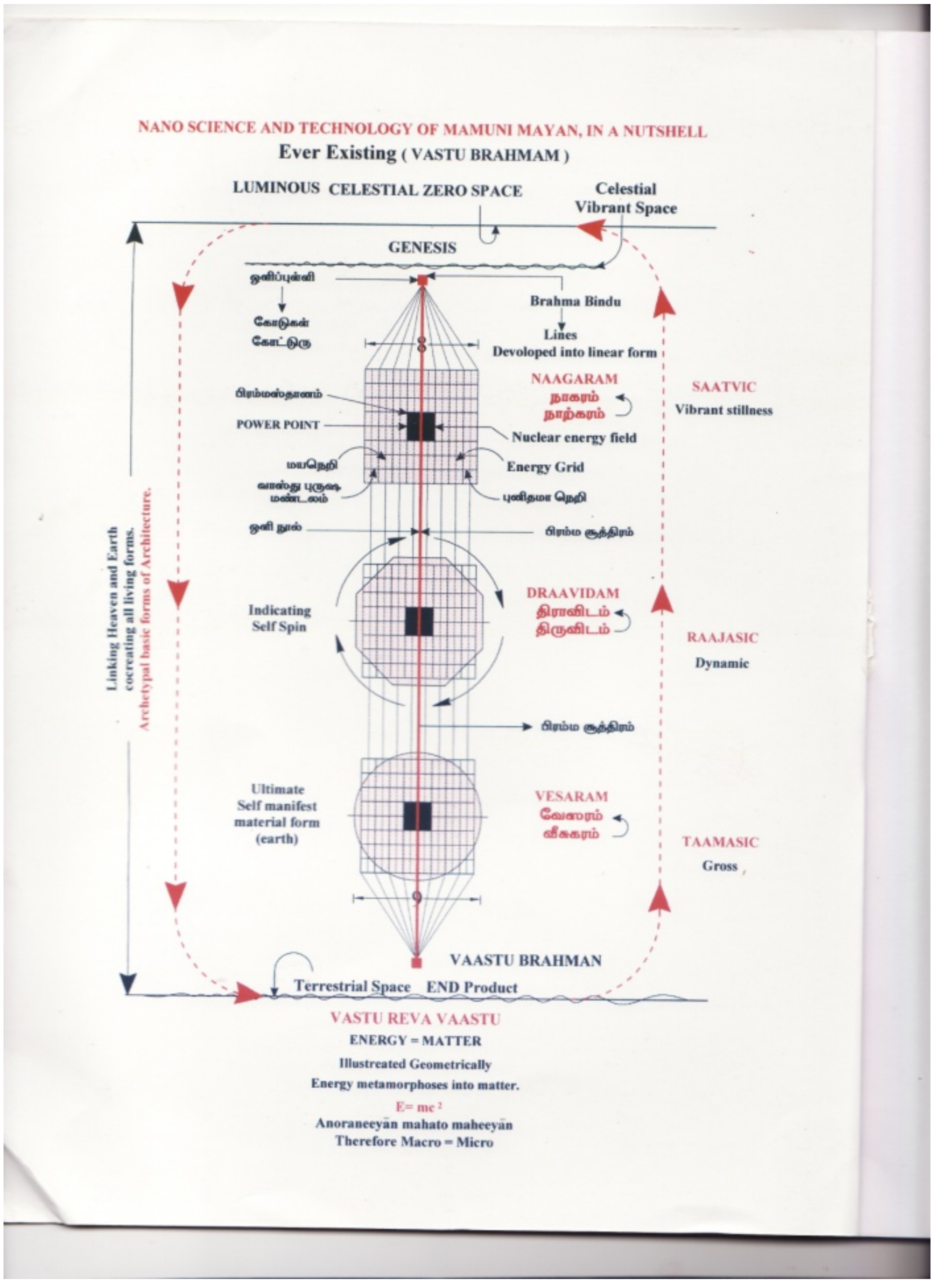
NANO SCIENCE AND TECHNOLOGY OF MAMUNI MAYAN, IN A NUTSHELL
Ever Existing (VASTU BRAHMAM)
LUMINOUS CELESTIAL ZERO SPACE
Celestial Vibrant Space
GENESIS
ஒளிப்புள்ளி
கோடுகள்
கோட்டுரு
Brahma Bindu
Lines
Devoloped into linear form
8
பிரம்மஸ்தானம்
POWER POINT
Nuclear energy field
மயநெறி
Energy Grid
வாஸ்து புருஷ மண்டலம்
புனிதமா நெறி
ஒளி நூல்
பிரம்ம சூத்திரம்
NAAGARAM
நாகரம்
நாற்கரம்
SAATVIC
Vibrant stillness
Indicating Self Spin
DRAAVIDAM
திராவிடம்
திருவிடம்
RAAJASIC
Dynamic
பிரம்ம சூத்திரம்
Ultimate Self manifest material form (earth)
VESARAM
வேஸரம்
வீசுகரம்
TAAMASIC
Gross
9
Linking Heaven and Earth cocreating all living forms.
Archetypal basic forms of Architecture.
VAASTU BRAHMAN
Terrestrial Space
END Product
VASTU REVA VAASTU
ENERGY = MATTER
Illustreated Geometrically
Energy metamorphoses into matter.
E= mc²
Anoraneeyān mahato maheeyān
Therefore Macro = Micro

Figure 18 – Manifestation in a nut shell (courtesy: Dr. Ganpathi Sthpati)

This diagram represents everything we have discussed so far. The central red line is the Brahma Sutra which runs from the center of Galaxy to the terrestrial forms. Just correlate this with the Mayan Hunab Ku in the previous part.

This summarizes as to how Energy becomes matter, Subtle becomes Gross, Un-Manifest becomes manifest, how 8 becomes 9, and Square / Cuboid becomes a circle / Sphere. All represented in one formless form called SHIVA LINGAM.

உ

16. TIME = SPACE SERIES. PART 16 – SEELAM - THE PRIMORDIAL CIRCULAR FORCE - GANAPTHI & MATHEMATICS

In the last few decades the main stream western science has started accepting concepts like uncertainty principle, wave mechanics, quantum mechanics, and periodicity of the wave pattern. Most importantly it is about the fundamental wave pattern of any energy or matter and the harmonics of these wave patterns falling in a particular form and proportion which is dictated by geometry / mathematics. In all the western scientists are scratching their heads on the depths of science of the subtle aka the sacred spatial geometry.

Most of the eastern philosophies have mastered this science and have codified these natural laws as a part of their religion, which is often dismissed as superstition. I already highlighted that in Tamil and Sanskrit the definition for science is "**Knowledge of the space**".

Modern science is moving in the right direction albeit very slow and they are trying to understand as to how elements are bonded within a molecule in addition to what elements constitute them. For example **Chlorophyll** molecule consists of **Carbon, hydrogen, nitrogen and magnesium (Mg at its core)** has a complex 12 fold Pattern – just one such typical pattern which can convert sunlight into life substance. The same pattern exists in our **RBC** cells just that it has an **iron** at its core instead of magnesium.

The understanding about the human body especially, "the spatial awareness that exists in each and every cell is due to the different sensory organs tuned at frequencies", led the scientists to unravel the innate spatial geometry of life.

Be it living, non-living beings, abstract forces like gravity, nuclear force, electromagnetism, heat, movement of astronomical bodies it does not matter what, everything in this universe adheres to a vibration / periodicity and is reducible to a number or a ratio. As a wise man said, "**Mathematics is the language of gods". This primordial rhythm & its harmonics are captured under the topic "Seelam" by Mayan in ancient tamil texts. Let's look at this in detail.**

I have been finding it difficult to sequence the flow for this topic and let me try starting from the basics:

- We know that the **Paramaanu or God's particle is a cube.** This particle with the primordial vibration combines with other particles to create a shape. **The Paramaanu combines in the multiples of 8 to evolve the next structure**. Let us see the importance of 8 later, but remember 8 is the number for Lord Narayana and Om namo Narayana has 8 Aksharas.

- What do you call a cube called in Tamil? It is called **Ganam (கனம்**). This also means heaviness, gravity, honor, dignity, abundance, plenty. To indicate it is a square cuboid they call it **கன சதுரம்**.

- The process of the addition of **Paramaanu** into bigger units is called as **Ganam (கணம்,** गणं). In Tamil both the Cube and the additive resultant are called as **Ganam** and the only difference being that in Cube it uses a **small "na" (ன)** as an alphabet. In the additive process / resultant it uses a **Big "na"(ண)**. The big "na" signifies that due to the additive process the small (na) becomes big (na).

- This additive process starts because of the **first rhythmic circular** force also called as a **Suzhi (சுழி)** in tamil which means "loop".

- Here comes the beauty. The Lord of or the energy associated with this **additive force (Ganam)** is called as **Ganapathi (கணபதி, गणपती)**. Note this is the bigger "**na"** used and not smaller **"na"**. He is associated with the first circular force - **Suzhi** (and hence he takes the first honor in any Vedic ritual. This means Lord Ganapati is the first process in the evolution and everything has to start with him. Even if it is Lord Shiva or Vishnu it does not matter, the very first honor goes to the **Lord of Ganas**. Many Indians when they start writing (atleast few decades back) they first start with this symbol at the top of the page before anything for a successful completion of that activity.

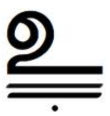

- If I add my own inferences to the above with a caveat that everything else I have written have a solid back up mostly from our own study of languages (as you can yourself see) and our scriptures.
    - The symbol above signifies that what was "a point" with untouched / non-interacting energies / resulted in the primordial first circular force out of a pulse and also signifying that every circular force shall also collapse back to a point.
    - Also the bigger / heavier things are formed or created by its smallest unit and hence symbolically the heavier ones are carried by the smaller units. To signify this may be a god in the form of an elephant is carried by a mouse.

- Now you know why every ritual starts with this Sloka to Ganapathi:

    - ***"GanAnAm tva Ganapati gum Havamahe, Kavim Kavinam Upamasra vastamam Jyestharajam Brahmanam Brahmanaspat aana Srnvan nuti bhissi dasadanam"***

      This means "May you the Lord of Ganas, Ganapati",... . so it goes.

- **Now what do we call the study of this additive force - Ganas? Mathematics – Ganith / Ganitham (கணிதம், गणित).**

- **So we Indians have codified:**
    - **the primordial God's particle as Paramaanu or microbode or Lord Shiva,**
    - **the rhythmic movement of the pillar of light as Lord Nataraja,**
    - **the first circular force as the symbol for Lord Ganapathi,**
    - **the lord of the additive force as Lord Ganapthi and**
    - **the Study of this additive force which leads to the formation of everything in the universe is Mathematics in Sanskrit and Tamil.**

- **Every possible state of consciousness and manifestation is represented in Shiva Linga**
- **We will see that the multiples of 8 in which the additive force moves is codified as Lord Narayana or Vishnu and the five state manifestation process as the five letters associated with Lord Shiva.**

Just forget that these names are the names of the Hindu gods, and tell me one religion or philosophy or one branch of science in this world which can capture the nature’s primordial secret in this way and codify it for the common benefit of the world and carry it for thousands / millions of years.

These are just a start and please wait for the other interesting aspects to unfold.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy Reading!

17. TIME = SPACE SERIES. PART 17 – SEELAM - THE PRIMORDIAL RHYTHM & FUNDAMENTALS

Now, let us get into some real heavy stuff - Mathematics (**Ganitham** – the study of Ganam) that governs this rhythm and harmonics. My first concern is how to make it easily digestible for an average individual who hates Math? So I shall take the easiest route that I know. I shall first explain the concept, then the actual Math part of it and then go to the philosophy and significance of the same. Later we shall look at how these are codified in nature and our scriptures and available in front of our eyes which we have failed to recognize.

Let us be clear about the following terminologies like **Progression, Proportion, root, Diagonal and Pythagoras theorem**:

- Let us take a series as follows: 2, 4, 8, 16, 32…. In this we need to be clear about what is **Progression and proportion**. As many of you know that whole series is a geometric progression where the next number in the series is a function of the current number and a multiple. This multiple here is 2, so that 2*2 = 4, 4*2 = 8 and so on. **This multiple or the constant is called as proportion**. You derive the value of the proportion by dividing the number by its previous number – (n+1)/n : (n+2)/(n+1). Hope this is clear.

- We know that a square has 4 equal sides. Each side of a square is called as "**root**" and the line that connects the 2 opposite ends is called as '**Diagonal**". In the figure below sides AB, BC, CD, DA are called as "**Root**". Lines AC, BD are called as '**Diagonal"**.

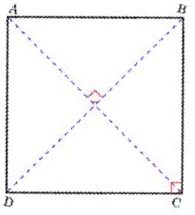

Figure 19; Square

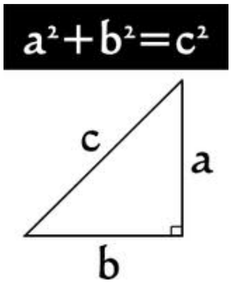

Figure 20: Pythagoras theorem

- **Pythagoras theorem** – This theorem gave us the formula to calculate the sides of a right angle triangle as given above. If you are not familiar with this, then please understand that if you know any 2 sides of the triangle the 3rd side can be calculated by the formula given above. I am not going to focus on the argument that this theorem was available in Indian scriptures before the Greek mathematician gave it. Let's look at the concepts and not which civilization gets credit as of now.

Since the fundamental structure is a square in 2D and Cube in 3D and also in the process of evolution one becomes many, let's ask ourselves the following questions:

- **How one square does create another square which is larger in size, progressively? What is the progression and proportion in this series?**
- **How one cube does create shapes which are polygonal (Hexagon, Octagon, decagon, dodecagon etc) progressively? What is the progression and proportion in this series?**

Now let us see Figure 21 where and consider the square ABGF. You can enlarge the image to see the alphabets clearly. Now as we know sides AB, BG, GF, AF are called "**root**" – remember this well and let us assume the smallest root possible is unity and hence each root is of unit length. AB=BG=GF=AF=1. Area of this square = 1*1 = 1.

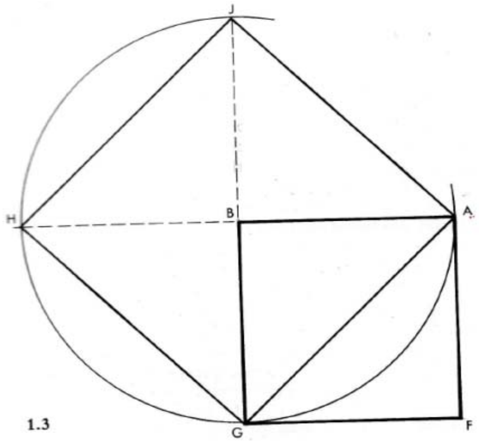

Figure 21; Square

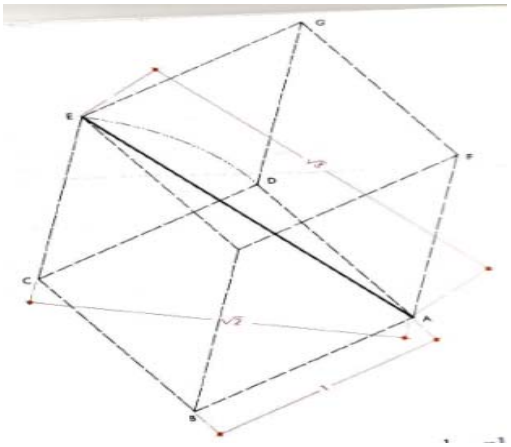

Figure 22: Cube

Line AG is called as the **"diagonal"** and as per Pythagoras theorem the length of **AG** should be **$\sqrt{2}$**. **AG = $\sqrt{(AF^2+GF^2)}$.** So to summarize **the root of the square ABGF is 1 and the diagnol is $\sqrt{2}$**.

Now consider a larger square AGHJ, here AG, GH, HJ, AJ are the sides and hence called as root. AH, GJ are the diagonal in the 2nd square. We know AG = $\sqrt{2}$, then AH = GJ = 2 as per Pythagoras theorem. To sum up **the root of the square AGHJ is $\sqrt{2}$ and the diagonal is 2. Area of this square is 2.**

If you look at the ratio **root/diagonal** (root to diagonal ratio) of both the squares it is $1/\sqrt{2}$ and $\sqrt{2}/2$ which results in $1/\sqrt{2}$. This will progressively hold good. The magic proportion in this progression is **$\sqrt{2}$**.

What we can infer from this is if we consider a **progression of squares** from a unit square to larger squares, then the **proportion** of the progression is $\sqrt{2}$. But before we jump into its philosophical significance let's answer the second question.

Consider the cube above whose side is of unit length. We know a cube is made of squares. We know ABCD is a square of length (root) =1 and diagonal = AC = **$\sqrt{2}$**. If we want to find out the length of the diagonal of the cube which is AE, then ACE is a right angle triangle with AC = **$\sqrt{2}$**. And CE =1 and hence **AE = $\sqrt{3}$**.

What we can infer from this is if we consider a **progression of cubes** from a unit cube to larger ones, then the **proportion of the progression is $\sqrt{3}$**.

உ

I am going to handle the significance of √2 and √3 in the next part.

Happy reading!

18. TIME = SPACE SERIES. PART 18 – SEELAM - SIGNIFICANCE OF $\sqrt{2}$ & $\sqrt{3}$

Let's look at the significance of **$\sqrt{2}$ and $\sqrt{3}$** but before that let's capture the philosophical significance of **Progression** and **Proportion**:

- The whole world is conceptualized as wave harmonics of the primordial rhythm with a **progression and proportion**. Philosophically the **Proportion** which does not change / which is constant is the **immutable component** and the **progression** which changes / is volatile is called as the **Mutuable component** in the process of evolution. **The whole universe is the progression and God is the proportion**. So if I know the God's proportion and the formula behind it I can create anything in this universe including a universe. This secret is called as the Brahma Vidya / Sree Vidya (Knowledge of Brahman, Sree) in Indian texts. What we are going to see further are very minor parts of the **Brahma Vidya**. ☺

- In the last part we saw, that the primordial square multiplied itself to be a larger square (**progression of squares**) using the **proportion $\sqrt{2}$. The diagonal of the square 1 (ABCD) forms the root of Square 2 (AGHJ)**. This is very profound if you get the concept**.**

- **If we consider square 1 as the cause then its diagonal is the effect. This diagonal is the root for the next square (AGHJ) and hence the effect of Square 1 is the cause of Square 2, this cause results in square 2's diagonal which is its effect and this is how the progression or evolution happens.** If you can see glimpses of Cause and effect cycle and the famous law of Karma in this analogy then you can pat yourself. ☺

- A square halved by the diagonal (square 1 with area 1) produces a square twice its area (square 2, area = 2). **The mystery of biological growth from cellular division or the different musical notes from the base tone is contained in this**.

- **Robert Lawlor in his book "Sacred Geometry"** very nicely puts it as the root of a plant (like the root in a square) is causative and embedded in the earth (and embedded in the square). These are very heavy and profound concepts just let it sink into you.

- When you divide the full height of the human being considering the total height as unity then belly as called as **Hara in Japanese, Dantein in Chinese, Nabhi in India** which is below the navel will measure (**$2-\sqrt{2}$**) from the soles of the feet to belly and (**$\sqrt{2}-1$**) from navel to the head. In Yoga, Zen and Chinese meditation techniques this point corresponds to the transformative and generative aspects of the individual that involves **rooting** techniques for **self transmutation**. **So Lao Tzu said, "To seek the root is the goal".**

- **$\sqrt{2}$ signifies the power of multiplicity. Hence the Generative aspect of this rhythm is attributed to $\sqrt{2}$. This represents the principle of transformation.**

- Remember **√2** is a proportion in the progression of squares and a square does not become a higher form. We have seen that the diagonal of the cube (sides of unit length) is **√3. It is this √3 which divides the volume form of a cube and the diagonal of the cube becomes a root for the higher shapes like pentagon, hexagon etc.**

- Hence **√3 signifies the formative power of the rhythm while √2 signifies the generative power of the nature.**

We will touch the same using **Vesica Piscis** in the next part which shall take us to **√5 and the golden mean Φ.** I am not going to touch upon important topics like Gnomic spirals and the relation between various progressions and hence I would request you to Google them if you are interested.

Happy reading!

PS:
There are many books on this topic and I would recommend **"Sacred Geometry – Philosophy & Practice" by Robert Lawlor** which I found easy to read and understand.

You tube and Internet is replete with tons of images, articles and videos on this topic and you can learn a great deal from them.

19. Time = Space Series. Part 19 – Seelam – Vesica Piscis, √5, & Golden mean (Φ)

If you don't know what is vesica piscis, please google and you would find it interesting. I am not going to explain it here and I would focus on the summary and aspects which are often left out. The philosophical aspect of Vesica Piscis, is a rotating point becomes as circle of unity. Two circle of unity (with its radius as unity) interacting / intersecting each other in a perfect manner offers an overlapping area which is considered to be sacred and fundamental to all the geometric forms in the universe. We will first see how **√2, √3 & √5** are part of this Vesica Piscis.

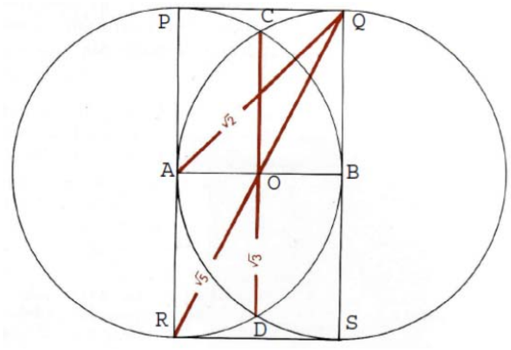

Figure 23; Vesica Piscis

Please look at 2 circles who radius is unity and AB = 1, intersecting as per the diagram. The overlapping area ABCD is the Vesica Piscis and we can see **√2, √3 & √5** all part of this Vesica Piscis.

ABPQ and ABRS are 2 squares and the side of the square is 1. **√5** is as significant as **√2 and √3.** In the diagram here **√5** is the diagonal of the rectangle PQRS / diagonal of 2 squares – one below the other.

√5 is the regenerative factor that binds 2 squares or that factor that transcends two worlds or the bonding factor for 2 cosmic entities. This **√5** shall take us to the most important ratio which is the **Golden mean** or the **God's ratio**.

But before that let me share what stuck my mind when I was engrossed **with Vesica Piscis and √3**. Look at the pictures below. These are pictures of forehead marks of Hindu Brahmins sects called Iyengars and their supreme deity is Narayana. I correlated the striking similarity of the rectangle around the Vesica piscis and Iyengar's namam. You just have to imagine the 2 overlapping circles are the right and left side of the face with the **Vesica Piscis** starting from the top tip of the nose. To me the **formative principle of √3 is in line with the "Bhootha krut" and "Vishwa karma manu stvashtha"** aspects of **Maha Vishnu** as per Vishnu sahasranamam. I had a good laugh on this correlation.

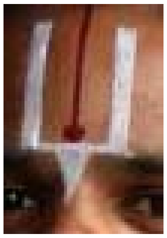

Figure 24; Iyengar Namam 1
(Courtesy: Internet)

Figure 25: Iyengar Namam 2
(Courtesy: Internet)

Coming to the **Golden mean,** If you have not heard about it, just Google or here is a quick link http://www.goldennumber.net/. Some of you might be even have fatigue repeatedly reading and listening about it, like I do. Hence I am not going to rewrite what is widely available. But let's look at the philosophical aspect of it, its relation to **√5** and its manifestation.

Now so far in **√2, √3 & √5** what we have seen is how Unity evolves as many. But if I am interested in geometrically dividing the Unity so that this proportion evolves as unity then how do I do it?

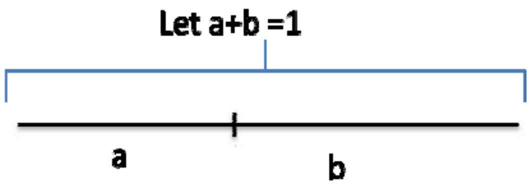

(Figure 26: Line andΦ)

Consider a line of unit length and also consider two segments a, b so that a+b =1.

A and B are such that it has to fulfill this condition of Proportionality that a/b : b/1.

This means what 'a" is to "b", is the same as what "b" is to the whole length which is 1.

Mathematically resolving this equation a/b : b/1, hence b^2 = a and hence b= **√a.** This means **a + √a = 1.** The answer to this equation is only one value which is called the **Golden mean / God's ratio** denoted by Φ = 1.6180339… and a = 1/Φ. **Philosophically it is the division of unity and the only possible creative duality within the unity.**

You may ask as to why unity is not divided into 2 equal parts and why is should be **1/Φ** (**0.6180339..).** This is effectively answered by our Upanishad **"Whether we know it or not, all things take their existence from that which perceives them"**. **Adi Shankara said "the universe is an illusion / maya**" **because it is what we perceive**. **We tend to ignore them as philosophy or superstition.**

But let's look at it scientifically as we understand. When the unity is divided into equal parts there is no asymmetry and difference and hence there is no perceptual universe. **An asymmetric division is needed to create the dynamics necessary for progression and extension of unity**.

Let's summarize our understanding again:

- **√2 – the generative component and a multiplying factor creates multiple squares from a unity of square.**
- **√3 – the formative component that creates multidimensional polygonal shapes from a cube.**
- **√5 – the regenerative component or the binding component that transcends and binds both the worlds.**
- **Φ - Golden mean or the God's ratio,** which is an evolutionary principle that is guided from within and an indisputable mathematical evidence of the conscious evolution of this Universe.
- Φ - Is a self-similar variable which can be used to derive Unity and it is the only proportion by which the Primordial rhythm maintains it perpetual oscillation to sustain this universe.
- Relation between **Φ and √5 is, Φ = (√5+1)/2**

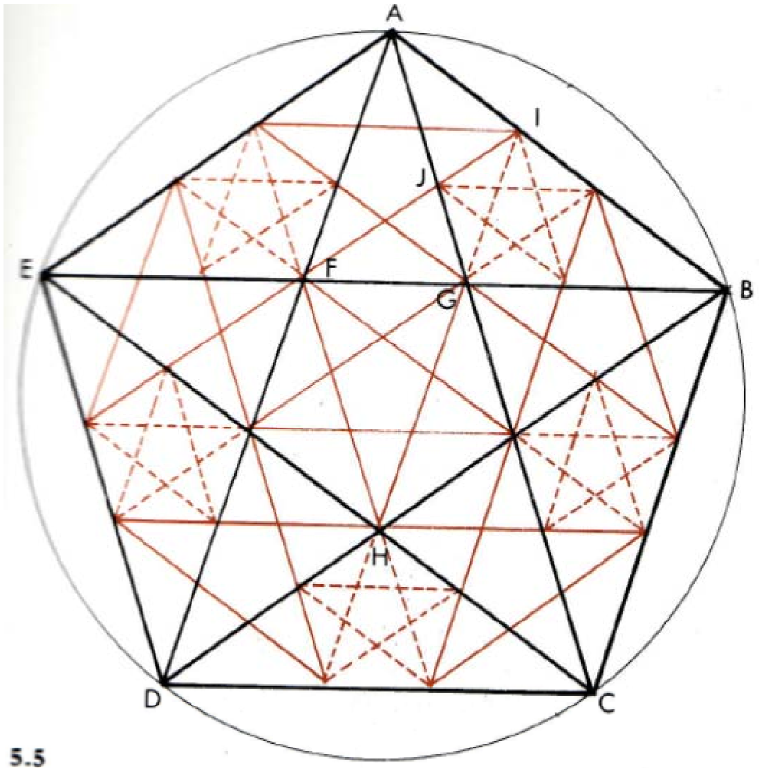

5.5

(Figure 27: Pentagon & Φ, Courtesy: Internet)

Now let's geometrically as to how Φ **and √5** are intricately linked. Consider the Pentagon picture whose side is of unit length. Ie.

$AB = BC = CD = DE = AE = 1$ and

$EG = FB = CG = DF = 1$ and

$EB = \Phi = 1.61803399$ and $BG = 1/\Phi = 0.61803399$ and hence all the ratios like $AG = AF = EF = GB = 1/\Phi$.

Also $FG = 1/\Phi^2$, which means $\mathbf{EB = EF+FG+GB = 1/\Phi+1/\Phi+1/\Phi^2}$

Also $\mathbf{GI = 1/\Phi^2}$ **and** $\mathbf{GJ = 1/\Phi^3}$

If you are interested more please check out the Fibonacci series and its relation to Φ. **The importance of 5 is dominant in every living organism while 6 and 8 are the characteristic of mineral and inanimate structures. It used to be rule of thumb that any plants / flowers which has 5 petals is an edible plant and hence 5 and pentagon are the symbols of life which contains the God's ratio or the golden mean.**

Now let's also understand how a **√5** is related to **√2 and √3.** If we consider a right angle triangle with the base of **√2** and height of **√3** or vice versa then the hypotenuse is **√5.** Now let me draw the connection between √2, √3, √5, Φ and π. If you think they are all some random numbers which cannot be fixed to their 10th decimal and not related to each other, just see the following relationship.

$$\pi = \Phi^2 \times 6/5 \qquad \text{or alternatively} \qquad \Phi = (\sqrt{\pi} * \sqrt{5}) / (\sqrt{2} * \sqrt{3})$$

This is the binding factor. I shall produce below some pictures which are aligned based on pentagon and the web is replete with them for you to understand more.

These 2 parts were little heavy and I shall move to easier aspects on the universal rhythm from next blog. The point I was trying to make is that everything in this universe is subject to a proportion and its fractals and let's see the manifestation of it in our daily life.

Happy reading!

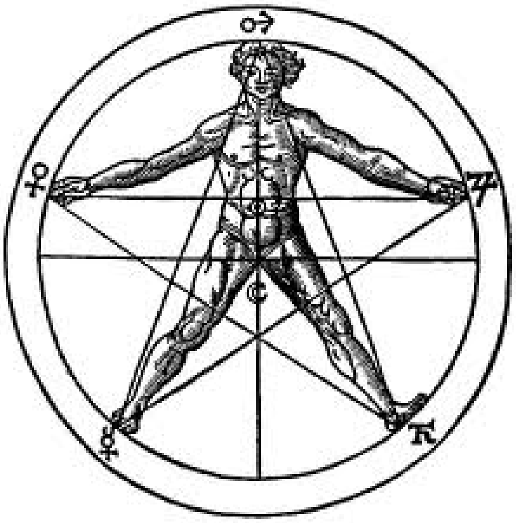

(Figure 28: Pentagon & Human shape, Courtesy: Internet)

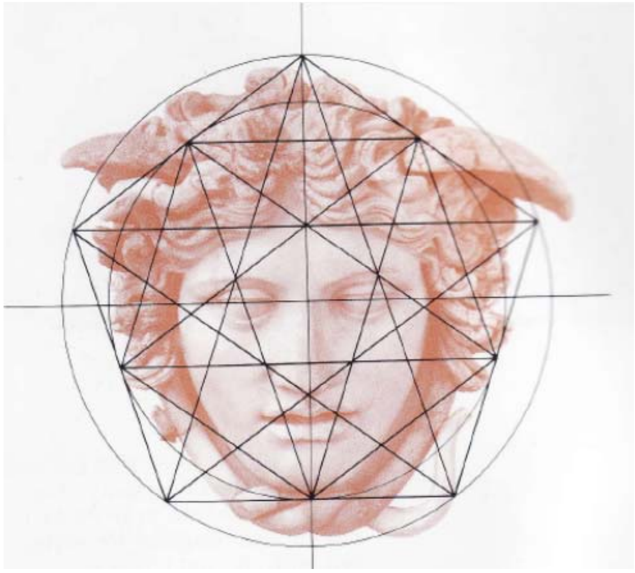

(Figure 29: Pentagon & Human face, Courtesy: Internet)

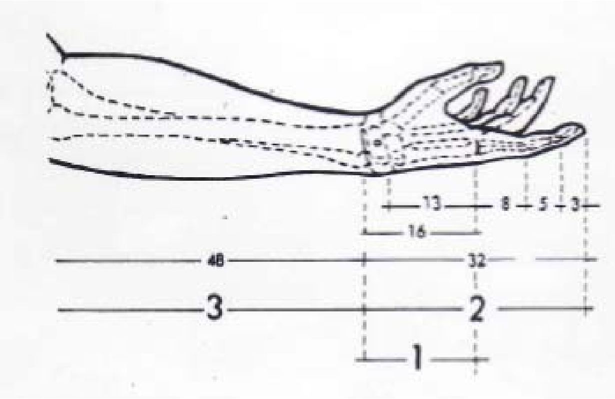
E
13
8
5
3
16
48
32
3
2
1

(Figure 30: Human hand and Fibonacci Series Courtesy: Internet)

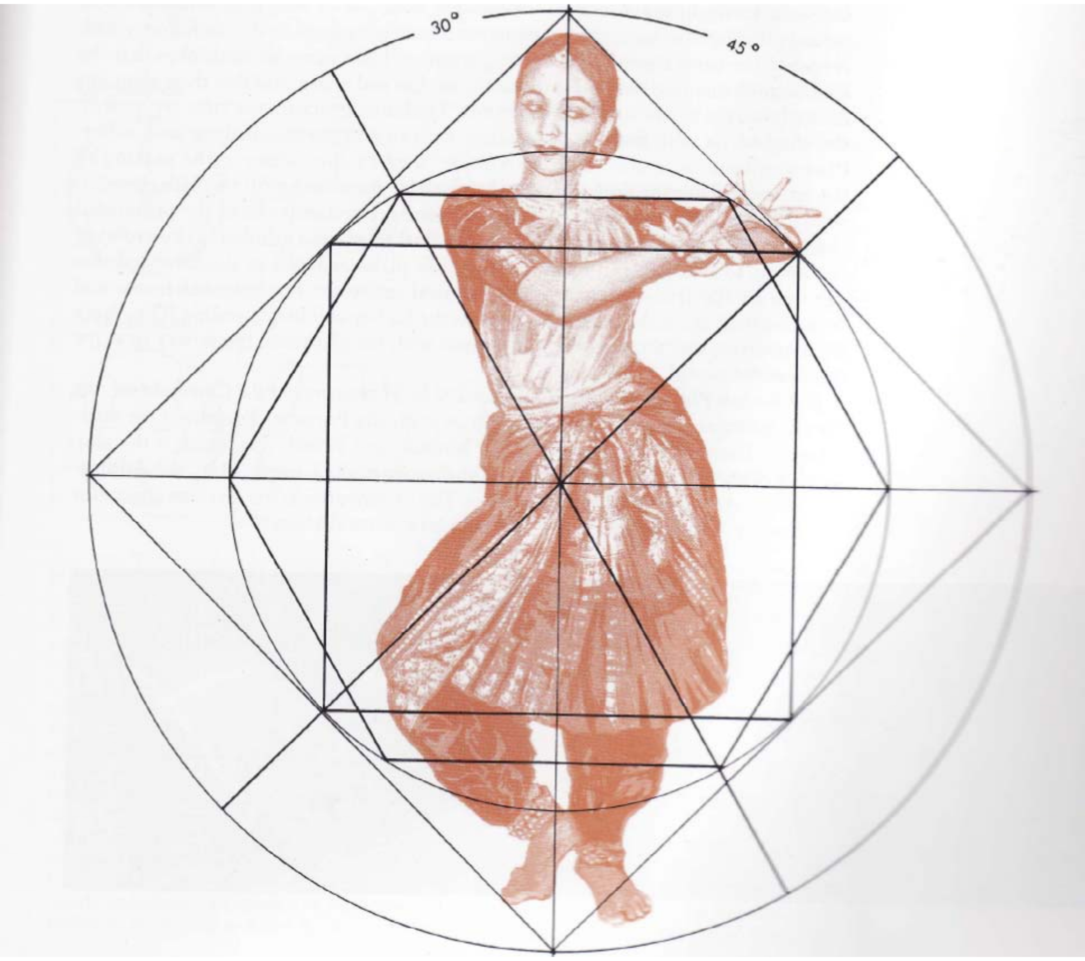

(Figure 31: Bharatanatyam and body alignment (Courtesy: Robert Lawlor, The Sacred Geometry)

The positions of Hindu classical dance (Bharat Natyam) describe geometric relationships from the axis of the body and the alignment of center of gravity at "hara" invoking rooting and divine power of transformation. In Hinduism "hara" is one of the names of Lord Shiva as in "Hara Hara Shakara"

20. TIME = SPACE SERIES. PART 20 – SANKHYA YOGA AND SAGE KAPILA

Now, let's look at what **Sankhyakarika** one of the greatest scientific works by **Sage Kapila**. I need to highlight that the **Sankhya philosophy** is one of the most evolved philosophies about the secrets of nature; it contains only 72 verses of which 68 of them define the complete spectrum of universal manifestation. **This Sankhya Yoga is referred in Bhagavad Gita – Chapter 2 is called Sankhya Yoga and in Chapter 10, verse 26, Lord Krishna identifies himself with Sage Kapila**.

Sankhya in Sanskrit means numerals / numbers. Sage Kapila states:

- Any measurement is a relational process and only a change can be measured but it takes time. The difference between the yardstick and the measured object constitutes the measured variable, which in effect is an incremental change. Since the observer can detect or measure only a change then logically, **the incremental unit of measurement itself should form the basis to derive the whole. This is the genesis of Φ.**

- **Correlate the above statement with the equation $a + \sqrt{a} = 1$. The unity is derived from the proportion "a" or $1/\Phi$.**

- As an example, the arch of a bow has expansive stress on the outer side and compressive stresses on the inner side, while the string is in tension and all are balanced at every instant. The expansive, compressive and tensile stresses can be expressed by one law.

- Alternatively, a single variable describes the proportionality of three types of forces existing at the same time. **Such a variable should be scale invariant, self-similar and have an axiomatic relationship to the whole.**

- **So Sage Kapila called $1/\Phi$ as a self-similar variable from which the unity can be derived.**

- So he postulated if **x** is the value of a measured increment then the total value **1+x** must be related to the ratio of change as 1/x at the same instant of time or simultaneously. This means **$1+x = 1/x$**, and would lead us to **$X^2 + X = 1$**.

- He quotes that if the primordial rhythm has to be self sustaining then the proportion is certainly governed by $1/\Phi$. He gives the formula for this oscillations as follows:
    - **$X = 1/\Phi = 0.61803399$**
    - **Increment / Expansion = $1+x = 1.61803399 = \Phi$**
    - **Compression = $1-x = 0.38196601 = X^2$**
    - **Resonance = $X^3 = X - X^2 = 0.23606798$**

Resonance or X^3 is the factor which protects this rhythm from decaying and dying off. So **Φ or $1/\Phi$** is such an important ratio or proportion so it is called as the **Golden mean or God's ratio**.

I shall highlight some of the factors / ratio which are part of Sankhya Yoga and what is amazing and you find formulas for Tamasic, Sattwic and Rajasic Guna states. (Please refer “Secret of Sankhya Yoga” – G. Srinivasan in the web).

- **Kapila defines the four qualities of the unmanifested state as Aikaantha (Synchronized), Aathyanta (perpetual), Atho (Dynamic)and Abhavath (unmanifest).**
- Taking C as cyclic rate of oscillation or vibration and x as rate of change per cycle then the three Guna self similar interactive states are shown as vibratory counts per cycle by the formula.
  Thaama = C^{1+x} . Rajasic = C^{x+x} . Sathwa = C^{1-x}.
- Some of the numbers that Sankhya Karika deals with are **1/7 (0.142857), π/10 (0.314159), 1/Φ (0.618034), √2 (1.414), Φ(1.61803399), √3 (1.732), √5 (2.23), e (2.718282), π (3.14159) and declares that the unmanifested state (Abhavaat) can exist only when the oscillations are between π/10 (0.314159) and e(2.718282). This is simply amazing since the state of Abhavvat is the state of Brahman.**

Despite herculean efforts I am unable to understand the Bhashya for the 72 verses which is about 500 pages explaining the secrets of Sankhya in any significant measure. I wish someone can step in and help us by explaining Sankhya Yoga that is comprehensible to our level of intelligence.

We have a treasure of secrets and richness hidden in our scriptures and it seems we have already lost most of them and the fundamental reason being our indifference to Sanskrit and our mother tongue.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

The passage produced below is just the summary of verse 1.

…The Sankhya view that evolves out of the complete theory is that only vibrations are detected by the observer... Since only relative changes can be be detected, any fundamental component that is not in a vibratory state cannot be detected. Vibrations are caused by components in an interactive state or a cyclic state of colliding and separating that is common to all interactions and wave phenomenon.

The interactions create three distinct modes of stresses. Collisions are compressive or inelastic. The resultant reaction is expansive or elastic. When the interacting components lack the freedom to move away the interactive state is maintained in a cyclic or shuttling mode in the same location in a resonant state.

This important Sutra lays emphasis that if the observed process of detection is dependant on vibrations, then fundamental space ***must*** *contain components with those four characteristics described as states, which makes it function in a holographic way. That is all vibrations remain in fixed relationship relative to each vibratory point in a coherent and resonant state.*

உ

21. Time = Space Series. Part 21 – 5 and 8 the Universal rhythm

Let’s summarize what we have seen so far and then move on to more interesting aspects.

We have touched upon the God’s particle and its characteristics. We also looked at the 5 fold manifestation process and in that we looked at how **Moolam** becomes **Kaalam** (time) and how time rhythmically (**Seelam**) vibrates with a proportion and creates forms and shapes (**Kolam**). We also touched upon some of the ways we had codified the secrets of nature like **Lord Shiva, Lord Nataraja, Lord Ganapati etc.**

Before we see how rhythm **(Seelam) and shapes (Kolam)** manifests itself as universe, let’s first look at one specific word from Tamil which has three significant meanings converging into one, Yen (**எண்**), means number **eight, generic term for Number / numeral** and also root for **எண்ணம்** which means thought. I shall explain the significance of this below.

Importance of 8 / Octet:

It is stated that these God’s particles combine in the multiples of **8 / octaves ONLY** and the states of manifestation from being subtle to gross is as given below.

○ **Formless Un-manifest State**	**0 (This is the Abhavaat state)**
○ **Guna State**	**$8^0 = 1$**
○ **First stage**	**$8^1 = 8$**
○ **Second stage**	**$8^2 = 64$**
○ **Third stage**	**$8^3 = 512$**
○ **Fourth Stage**	**$8^4 = 4096$**
○ **Fifth Stage**	**$8^5 = 32768$. This is the state with manifest form.**

This rule is applicable to both the visual and aural forms which mean we can see or hear anything only in the fifth stage of evolution. So as per **Pranava veda, the fundamental numbers are 0,1** and the key **numerals for proportions are 5, 8**.

Significance of 5 and 8:

Pranava Veda describes that 8 / octets is the essential harmonic oscillator for the manifestation of the universe since this generative force is the multiple with which the whole universe manifests and sustains. You many note that number 8 is associated with Lord Narayana and his Ashtakshara mantra **(Om NaMoNaRaYaNaYa**).

The principle of 5 / Penta, we have seen is the regenerative binding force earlier. It is always a 5 stage process. This is associated with **Lord Shiva (Om NaMaSiVaYa).** The un-manifest form comes to the Guna state with the force of 5 and in the multiples of 8. It is also important to highlight that 5 and 8 are part of the **Fibonacci series and their proportion is 8/5 = 1.6, which is Φ**.

உ

But it is important to realize that the building of octets is a 5 step process for manifestation. This highlights that one cannot exist without another and both are interdependent. Now let’s recall:

"Shivasya hridayam vishnur, Vishnoscha hridayam shivah:" and
"Shivaya Vishnu rupaya Vishnave Shiva rupine"

Vishnu (8) is the heart of Shiva (5) and likewise Shiva is the heart of Vishnu and they both are the representation of each other.

If you have noted Tamil language stressed the importance of words and its meaning – this was addressed in the first few parts of this series. Here Tamil language gives the importance of numerals and its significance in the formation of universe. Both the great sages **Auvaiyaar** and **Thiruvallular** stated the importance of numerals and letters in an identical fashion. Both these lines mean, Numerals and letters are like eyes for living beings / equal to eyes.

எண் என்ப ஏனை எழுத்து என்ப இவ்விரண்டும் கண் என்ப வாழும் உயிர்க்கு!
எண்ணும் எழுத்தும் கண்ணென தகும்!

Now what should be noted here is that (**எண்**) importance of numerals precedes that of letters and we know why. The aural and visual forces become letters & its associated sound at the 5th stage but the numerals are from the Guna state. **Numerals were associated with 8 / Lord Vishnu and letters were associated with 5 / Lord Shiva**. Guess you can get the scientific knowledge and significance we have in our languages.

Since what is in macro is there in microcosm, taking the analogy to microcosm our mind becomes the Moolam (Source), our thoughts are the vibration, time is measured as the periodicity between thoughts, based on the rhythm of our vibration we perceive the world. This is given very nicely as **உள்ளமே மூலமாகி உணர்வுறும் கோலமாகி**! means **Our mind is the source and its awareness becomes forms.**

This wonderful concept is what is linked in the word Yen (**எண்**), which means number **eight, generic term for Number / numeral** and also root for **எண்ணம்** thoughts. **Just this one word in Tamil highlights the supreme secret that thoughts become form and words with the help of numbers and numerical proportion converts aural and visual forms to thoughts.**

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

22. TIME = SPACE SERIES. PART 22 – EIGHT (OCTET) AND THE 5 SACRED ARTS

எட்டே எட்டியல் எட்டெட் டியலே
எட்டின் நெறியே அளவை மாநெறி !!

The above lines from the work **Aintiram** means **"Eight / Octa (multiples of 8) is the fundamental numeral for any measurements**. We did see that the fundamental god's particle add up with each other in the multiples of 8 in the previous part.

In ancient India there were about 64 arts (http://en.wikipedia.org/wiki/Kal%C4%81) known as **Chatuhsasti kala ("ஆய கலைகள் அறுபத்துநான்கினையும்"** as popularly known in Tamil), but I am going to talk about 5 sacred arts.

We will look at the importance of numeral 8 and how this is fundamental to everything in this Universe. But let's again start again with etymology. The word arts in **Sanskrit** is **"Kala" (कला)** and **Time in Sanskrit is काला (Kaala).** So etymologically we can say that classical arts are all about variation of time (or harmonics of this pulsation of time in the multiples of 8).

After etymology let's get to the basics. Our understanding of the universe is that **"the primordial space manifests itself as spatial forms through pulsation (Time) and its periodicity (Rhythm & order).** This evolution started with **Pranava which came out as aural and visual form (Sound & light)**. When they say **"The one became two"** they refer to the Pranava which is light and sound from the unmanifested consciousness.

Aintiram states that the dual form of **Pranava** gave forth 5 sacred arts which are fundamental to most of the other arts. **The Aural form of Pranava gave forth Poetry and Classical Carnatic music, which can only be heard. The Visual form of Pranava gave forth Sculpture and Building architecture, which can only be seen. Both the aural and visual forms merged to create the classical dance (Bharatnatyam**), which involves both seeing and hearing.

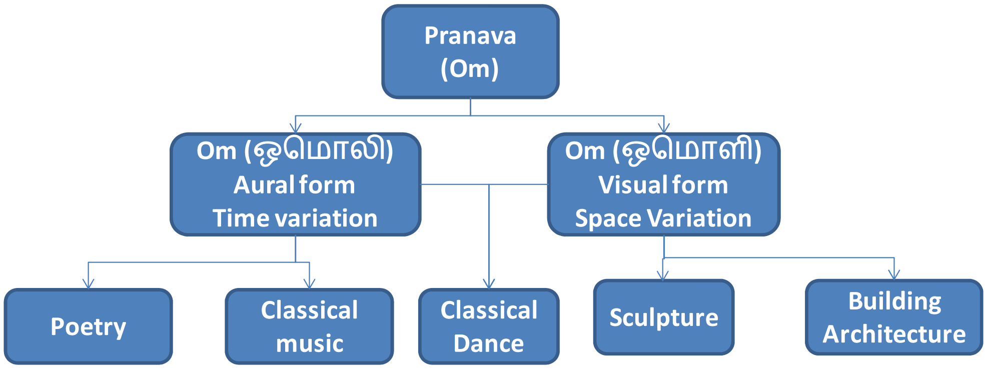

Figure 32: 5 Sacred arts

உ

They were classified as 5 vedas:

- **Poetry – Sabda veda**
- **Classical music – Gandharva Veda**
- **Classical Dance – Naatya veda**
- **Sculpture – Sthapatya Veda *(Sthapati is a clan that is the master of this veda)***
- **Architecture – Pranava Veda**

Pendatic Tamil and its construction:

Pendatic Tamil is called as **ஐந்தமிழ்** (**Aintamizh**) which consists all the five above arts, **இயல், இசை, நடனம், சிற்பம், கட்டிட கலை**. It is these 5 arts which were formed out of the Tamil language and the language itself is created from the word Om. They say that it is not 12 basic vowels in Tamil but only 5. They are அ, இ, உ, எ, ஒ. These 5 letters combine to form the rest of the letters like

அ + அ = ஆ, இ + இ = ஈ, உ + உ = ஊ, எ + எ = ஏ, ஒ + ஒ = ஓ, அ + இ = ஐ,

அ + உ = ஔ

Here you may want to note that 5 has become 12. Same way in classical music they say that the basic svaras are 5 (ச, ரி, க, ப, த) – Sa, Ri, Ga, Pa ,Da which became 7 and then 12.

So 5 is the fundamental numeral for evolution and Tamil language has grown based on this philosophy and science. I am sure some of the other Indian languages have the same basics. The book which describes this phenomenon of evolution of language from Pranava and arts from language is **Aintiram (it means 5 works) (ஐந்திறம்)** authored by Mamuni Mayan dated 10000 BC.

You would agree that the aural forms are predominantly variations in time and visual forms are of spatial variations. But our ears stand for akash tatva (“space”) and it can perceive variation of time and space and we will see this in detail. When we said **“Time = Space”** we will see that both of them are governed by the same laws and measurements and hence there is no difference between **Time and Space**.

Happy reading!

உ

23. TIME = SPACE SERIES. PART 23 – NANOTECHNOLOGY IN ANCIENT INDIA & SIZE OF GOD’S PARTICLE

Let me start with these questions:

- **What is the size of the God’s particle and how is it calculated?**
- **Does nature endow a standard measure for an average human being’s height?**

Before we proceed to the next section I want to give the very detailed space calculation table used in ancient India. Our rishis did not stop just with telling us that the **Paramaanu / God’s particle** is very small but went ahead and established its size.

If you have to see such minute particles then your sensory organs – Eye and mind should be tuned to it so that it can see things at that size. From the formula available in our ancient texts I was trying to reconstruct by working backwards to find out the size of the **Paramaanu / God’s particle should be.** Please refer the table below which I have arrived at.

ANCIENT SPACIAL MEASUREMENTS		
Unit scale	**Ancient metric**	**Current day Metrics**
1 Paramaanu =	Smallest particle	133 nano meters
1 Ther Thugil=	8 Paramaanu	1065 nano meters
1 Mayir nuni=	8 Ther Thugil	8526 nano meters
1 Eer=	8 Mayir nuni	68 Micro meters
1 Paen=	8 Eer	545 Micro meters
1 Yavai=	8 Paen	0.4365 cms
1 Angula=	8 yavai	3.4925 cms
1 Tala=	6 Angula	8.25 inches
1 Muzham=	24 Angula	33 inches
1 Muzham=	4 Tala	33 inches
2 Muzham=	8 tala	66 inches
Universal height of man	9 Tala	74.25 Inches

(Figure 33: Ancient Spatial measurements. Courtesy – Dr. Ganapathi Sthapati)

I have given the exact word used in Tamil for these measurements instead of translating it in English. Ther Thugil is the dust from chariot wheel, Mayir nuni is the width of a hair tip, Eer and Paen are the lice eggs and lice. Some of you may jump that most of the units were subjective and would have variations in normal life. I agree but there is a reference point – **Angula** and its exact measure based on which other values are derived backwards to arrive at the value of Paramaanu.

Now if you note the **smallest particle** aka **God's particle / Paramaanu** is **133 nm** in size which is currently under study by the CERN scientists. Nanotechnologists today confirm that the width of a hair strand is about 25000 nm and our ancient texts give the size at the tip of the hair strand (and not the width) as 8526 nm. I see that to be a reasonable estimate when no such precision instruments were available in those days.

One Angula is 11/8 of an inch. This is the fundamental measure which is used in any of the worldly measurements. Now you can see that 6 Angula is 1 Tala which is about 8.25 inches and then comes the most important revelation. **Does nature endow a measure for an average human being's height?**

The answer is yes. **As I said an average human being should be 9 Tala's height and hence the nature endowed height for any human being is 74.25 inches.** This is the ideal height and the variation is attributed to our Karma, hereditary and lifestyle.

Now let me give few tips to find out your height from typical measurements:

- Measure the length of the face (from forehead to chin). Your total height should be 9 times that height. This is the principle of Nava tala. As an extension measure your thigh or legs length excluding the knee and feet and your total height should be 4.5 times that.

- Take your right hand middle finger (No. Don't show it to anyone☺). You would notice that the finger is divided into 3 parts. Measure the length of the lowest part – from the end of the palm to the first subsection. Your height typically should be 54 times this measure. Since this measure for an average man with nature endowed length shall be 1 Angula. Hence your height shall be 54 times that measure.

- The length of the thumb from the base of the wrist should be about 3 Angula and hence 18 times of that should be your height. This can go on and on.

So, to sum up this section:

- Instead of looking them as Rishis or Sages or religious figures if we look at them as scientists in the field of nanotechnology, quantum physics and Astro dynamics we would accept most of the secrets without any prejudice.

- Not just the measure for the height of the man but they had formula for every object and this was the secret of their wisdom. For any astronomical object, they had formulas to measure body's core (width) and also the width of the mantle. The roots of their wisdom lie in understanding that "many' are just a harmonic variation of the "one" subject to a mathematical proportion. Hence there was no “rocket science” in those days and all of them were as simple as it looks like.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

24. Time = Space Series. Part 24 – Octet, Octave and the Gandharva veda (Classical Music)

Let’s start with classical music where Octaves form the fundamental for it. There are seven svaras (Sa, Ri, Ga, Ma, Pa, Da, Ni) which is made into an octave with another Sa after Ni. So it becomes Sa, Ri, Ga, Ma, Pa, Da, Ni, Sa (*Do Re Mi Fa Sol La Ti Do*). The first rule of classical music is “**Shruthi Mata, Layam Pita**”, which means **Shruthi is the mother and this corresponds to the frequency or the spatial component and Laya is the Tala component which is a function of time compared as father**.

Tala	**Anga Notation**	**Tisra**	**Chatusra**	**Khanda**	**Misra**	**Sankeerna**
Dhruva	IOII	11	14	17	23	29
Matya	IOI	8	10	12	16	20
Rupaka	OI	5	6	7	9	11
Jhampa	IUO	6	7	8	10	12
Triputa	IOO	7	8	9	11	13
Ata	IIOO	10	12	14	18	22
Eka	I	3	4	5	7	9

(Figure 34: 35 types of Tala. Courtesy: Wikipedia, internet)

There are 7 fundamental Talas in carnatic music and each Tala has 5 different variation based on Jathi so in all there are 35 different Talas as given in the tabulation. These Talas are measured as the “matras” the unit duration of time. There is another concept called Nadai or Gati which we will not touch upon now. But let’s look at these 35 variations and can you tell me where the most famous “Adi Tala” is in this 35? Most importantly why it is also called as **“Adi Tala” – which is the primordial rhythm**?

The answer is **“Chatusra Jathi Triputa tala”** highlighted in the table is called as **“Adi Tala”** and it is called so since it gave the most primordial rhythm with 8 time units. But there are couples of more cells where we can see “8” in the table and why these are not called ‘**Adi Tala**”?

The primordial rhythm has a pattern of **1, 3, 2 and 2** which adds to 8. Further on we will see just the variation of this Adi Tala in anything and everything. Tisra Jati matya Thala has a structure of 3, 2, 3 and Kanda Jati jhampa Tala has a structure of 5,1,2, hence they are not classified as the primordial rhythm.

Let’s quickly touch upon the spatial variation or the frequency / pitch aspect of this divine art before we branch off to other topics. The image below gives the 16 levels of Savaras which includes tones and sub tones. The seven svaras become 22 tones based on the frequency variation / harmonics. Here 7 are fundamental svaras and 16 as given below is the most practiced variation and 22 is not popular.

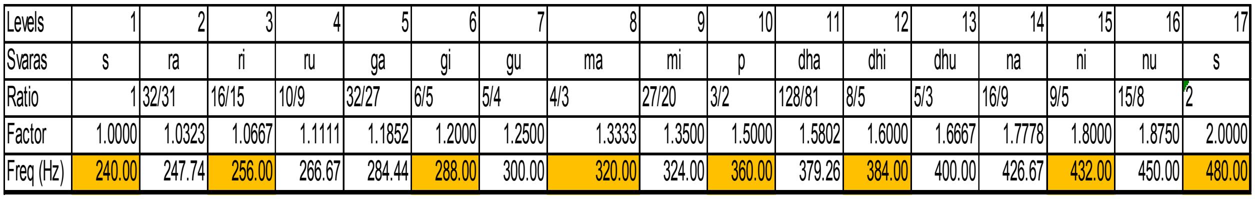

Levels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Svaras	s	ra	ri	ru	ga	gi	gu	ma	mi	p	dha	dhi	dhu	na	ni	nu	s
Ratio	1	32/31	16/15	10/9	32/27	6/5	5/4	4/3	27/20	3/2	128/81	8/5	5/3	16/9	9/5	15/8	2
Factor	1.0000	1.0323	1.0667	1.1111	1.1852	1.2000	1.2500	1.3333	1.3500	1.5000	1.5802	1.6000	1.6667	1.7778	1.8000	1.8750	2.0000
Freq (Hz)	240.00	247.74	256.00	266.67	284.44	288.00	300.00	320.00	324.00	360.00	379.26	384.00	400.00	426.67	432.00	450.00	480.00

(Figure 35: Svaras and frequencies)

You can note that Sa starts with 240 Hz and the last Sa is 480 Hz doubling the frequency and hence an octave. While Sa has 240 hz and 480Hz levels, the tone Pa (P) or Sol does not have any variation and this is at 360Hz. I am going to correlate the Mayan studies with Carnatic music and instead of seeing the whole spectrum from “Sa”, I am going to see this spectrum from “Pa”.

You would note that “Pa” at 360 Hz corresponds to the **Harmonic osciallator constant number** and the whole spectrum is almost like a Sine wave from 360Hz to 480 hz and falling down to 240 Hz with a uniform gap of 120Hz above and below. All the tones would fall within this 240Hz centered around 360 hz. So the spatial variation in carnatic music can be summed as “It originates at a value (240Hz) and varies by the same measure (240Hz) centered on the harmonic oscillator”. Since the objective is about looking at the primordial rhythm we will not touch any other aspect on this topic.

We will see how this **Adi Tala** is the fundamental pattern / rhythm that manifests in every other art.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

25. TIME = SPACE SERIES. PART 25 – ADI TALA, NAVA TALA AND STHAPATYA VEDA (SCULPTURE)

Let's look at how Octet and the primordial rhythm (Adi Tala) that defines Sculpture. Let's recap some of the following:

- The process of evolution starts with Pranava which has aural and visual forms. We have also seen that anything subtle is represented in a 8x8 energy grid and anything gross is in 9x9 grid.
- The fundamental measure with which the particles bond is **Eight or** it's multiple and hence the primordial rhythm is an **Octet.**
- The fundamental measure of time is **Tala** and the **primordial Tala is Adi tala** with 8 units and a pattern of 1, 3, 2 and 2.

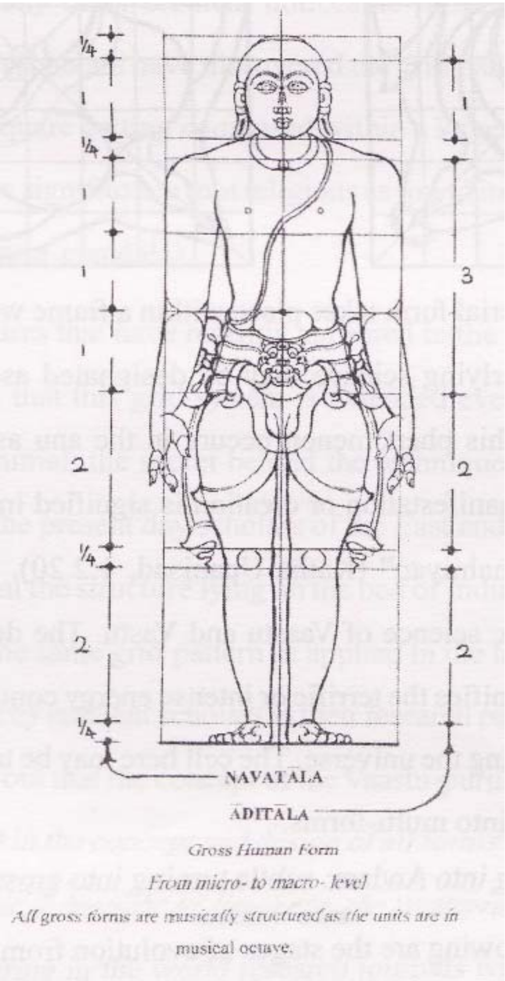
¼
1
¼
1
1
1
2
¼
2
¼
1
3
2
2
NAVATALA
ADITALA
Gross Human Form
From micro- to macro- level
All gross forms are musically structured as the units are in musical octave.

(Figure 36: Human form – Adi Tala, Nava Tala Courtesy: Dr. Ganpathi Sthapati)

Now let’s see the measurement rules as to how to make a sculpture of a human being.

- **The overall human height is measured as 9 talas.** Yes the same time measurement is used to measure height. Pinch yourself now, since height is a space variation and not a time variation yet it is measured with a metric of time. This is a secret which is unknown to many and even some of the experts in this field. This is where Time and spaces merges without any difference and this is the fundamental rule with which we have been creating visual and aural structures in the universe.

- **Divide a man’s height as 9 units of which head, torso, thigh and legs would fall into the pattern of 1, 3, 2 and 2 which adds up to 8. This is the primordial rhythm – the Adi Tala now as a human form.**
    - Then how 8 became 9? No other branch of Vedas or science tells us this secret so explicitly to my knowledge. **The following parts, Top of the head to forehead, Neck, Knee and feet each measure 0.25 talas and hence add 1 additional tala to 8 and it becomes 9.** So Adi tala which is a subtle version of human becomes Nava Tala which is the gross version of human being. In the process of 8 becoming nine all the joints are formed in the body☺.

    - Let me add one more dimension to this. Our human body has a gross body (Sthula) and a subtle body (Sookshma sarira). **The Sthapatya veda tells us that the gross body is nava tala and sookshma sarira is Adi Tala**.

    - I fell off the chair when I understood this and the linkages. Check if you are still on your chair and maybe you should sit on the floor from now on☺.

Martial Arts and Lord Nataraja’s stance:

Let me ask you the following questions here:

- Look at the image of the Lord Nataraja below. If a gross image should have Nava Tala which is a 9x9 grid then why the human image of Lord Nataraja which is a Nava Tala fitted into a 8x8 square instead of a 9x9 square?

- Both in Classical dance and in martial arts why do they ask you to stand in a posture where the knee is slightly or significantly bent. Not all of the stances but most of the basic stances correspond to this why? Let me know if you have received a satisfactory answer for this question.

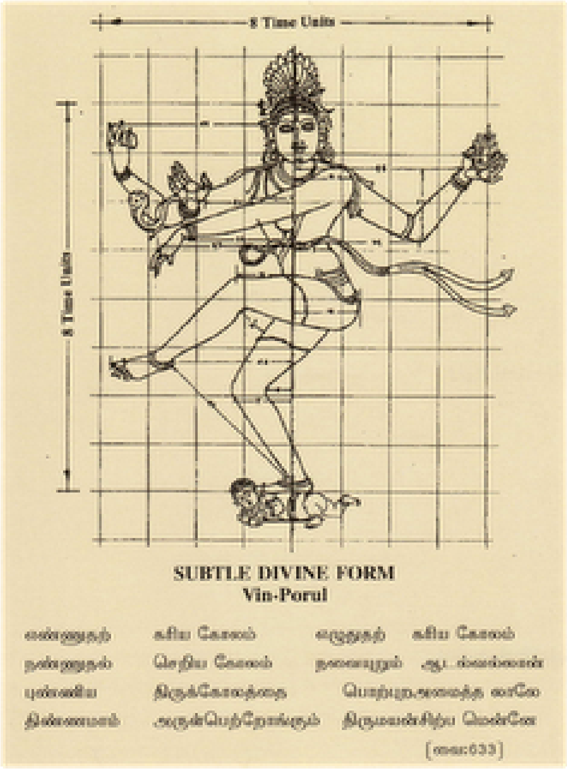

(Figure 5 reproduced)

- The answer to this question reveals one more secret to us.
    a. **In Sthapatya veda each grid line is divided into 10 or 11 or 12 (usually 12) and hence a 9x9 grid has 9*12 = 108 blocks or levels of energy. This may explain why 108 is a sacred number from an energy perspective.** This covers the entire spectrum.
    b. When we create a gross image its always sculpted to 9 talas but when they sculpt image of the gods and in this case Nataraja, they fit the gross image to a 8x8 energy grid to invoke the subtle energies.
    c. How do they fit a 9 tala to 8 tala? The technique is not to change the proportion of nava tala but they ensure that 9 fits into 8 by making the image to bend it legs such that the height reduces by one unit.

- You can notice that Lord Nataraja’s right leg is bent to fit the image in 8x8 grid. This is believed to invoke the subtle energies corresponding to that image.
- This is the same reason why martial arts and in Bharatnatyam your teacher asks you to take a basic stance which reduces the height from 9 units to 8 to invoke subtle energies. Of course the other reason is the alignment and focus to Hara / Dantein with that pose.

If I tell you that the science of nature and the associated philosophy codified in Hindu religious symbols are unmatched and amazing then I hope you would whole heartedly agree now. These religious texts and symbols offer the key to unlock the secrets of nature.

So to sum up, so far we have seen one variation of time as **Gandharva veda** that manifests as classical music and the primordial rhythm is **Adi Tala.** Next we saw that the **Sthapatya veda** in the form of sculptures offered the secret of space variation in **Adi Tala and Nava Tala**.

This is where Time and Space were equated with the same measure yet offered completely different arts.

The whole universe has these underlying and un-manifested interconnections and hence the **Hindu philosophy of "That one became many"…** is the supreme truth. Seeing the divinity in everything is the only truthful way of recognizing this science.

As a wise man said, **"Indians are the most ignorant about their rich past and scientific traditions."**

Happy reading!

உ

26. TIME = SPACE SERIES. PART 26 – SABDA VEDA, CHANDAS & POETRY (MA NISHAADA…)

शुक्लाम्बरधरं विष्णुं
शशिवर्णं चतुर्भुजम् ।
प्रसन्नवदनं ध्यायेत्
सर्वविघ्नोपशान्तये ॥
Shuklaambaradharam Vissnnum Shashivarnnam Caturbhujam |
Prasannavadanam Dhyaayet Sarvavighnopashaantaye ||

Let’s start with this Sloka as a sample to see how 8 and its harmonics are integrated with Poetry. In this sloka there are 4 lines (a line is called Pada and Pada means feet in Sanskrit) and each line has 8 Sanskrit Aksharas (Only vowels and vowels with consonants imposed on them should be counted and pure consonants are not part of the count.) This is usually written in 2 lines as shown in English.

This is the most popular way of creating a sloka or poem in most of the Indian languages. This “metre” resonates as 32 aksharas in 4 padas. You may want to correlate this with 1,3,2,2 where the last 2 stands for your feet. So feet (pada) measures 1/4th of the overall length of the subtle body (8 units) and 4 padas create the full blown subtle body. So such a constructed poem is a throbbing and vibrating energy body or a living organism and hence carries the power of the words it contains.

If all the padas have equal metre say 8 units then it is called **Sama Vrtta,** and if all are different it is called **Vi-Sama Vrtta** (not equal). **This 8 and 32 is the same as Adi Tala in carnatic music and ashta tala in sculpture. Can you make the connection now?**

In Sanskrit this metre is called as “**Chandas” and in Tamil it is called “Chandam**”. **Chandas** is also one of the 4 important Vendagas and is called as the feet of the Vedas. **Let me ask you as to which language the Vedas are available in its primordial form?** The most popular and incorrect answer is Sanskrit. This is the level of awareness we have. Lets quickly look at these important points about Vedas and I shall correlate this with our topic of discussion.

- Unlike other faiths Hindus don’t attribute the Vedas as word of god or given by some prophet. We consider Vedas as the breath of god (**“Nishwasitam”**) and not the word which means, Vedas are not created by god, but the breath. If there is no breath that person would not leave and hence it is associated with the very presence of god.

- Vedas are also **Apourusheya** which means it is not created by man and **Anaadi**, that does not have a beginning.

- But you can ignore them stating these are blind beliefs. Let’s look at how this correlated with the science. Vedas are created by Pranava and these should be considered as the harmonics of the primordial vibration and nothing more.

  - If it is pure vibration then the “which language it is written?” sounds stupid isn’t. If it’s pure vibration it has to be expressed in “Hz or wavelength” equivalent in ancient India. This is what **Chandas or metre** is all about. This measures pure vibration and anyone capable of tuning themselves to those frequencies have **SEEN** the Vedas and not **created them.** Hence all our rishis are **Mantra Drishtas and not Karthas**.

Difference between Vedas and Poetry:

Vedas contains both **poetic verses (Padya) and prose (Gadya)**. All the poetic verses are called **Chandas** and it complies with the rule as stated above. The difference between normal poetry and Vedas are, Vedas contain tonal variations (high pitch and low pitch) apart from the time, space variations but in Poetry there are no tonal variations. What has tonal variations and are not part of Vedas are called Slokas. So to sum up:

- Vedas are both in Padya and Gadya. What is in Padya is called chandas and it has tonal variations in addition to the time space variations.
- What has tonal variations but not part of Vedas is called as slokas.
- What does not have tonal variation but comply with the chandas rules are called Poetry.

Harmonics and types in Chandas:

We saw every line to have 8 matras / units / aksharas. There are variations and these are the harmonics available.

- **Gayatri –** This is a very special chandas with the mantra where instead of 4 padas, there are only 3 padas and hence it is called Tripada Gayatri. Each pada has 8 Aksharas and hence 24 Aksharas in Gayatri Mantra. This is also the only chandas where a mantra is named after the chandas. Some people write it as 6 letters per line and 4 padas which makes it 24.
- **Ushnik Chandas –** 7 letters per pada and 28 letters in all.
- **Anushtub Chandas –** This is the fundamental and popular one with 8 matras per pada and 32 in all. Ramayana the first poetry was set in this metre.
- **Brihatee Chandas -** 9 letters per pada and 36 letters in all
- **Pangti Chandas -** 10 letters per pada and 40 letters in all
- **Trishtup Chandas -** 11 letters per pada and 44 letters in all
- **Jagti Chandas -** 12 letters per pada and 48 letters in all. This has a subtype called bhujangam which is split as 6-6 and that moves like a snake. (Ex. Subramanya Bhujangam)
- **Shikarini Chandas -** 17 letters per pada and 68 letters in all. In Soundarya Lahiri it is split as 6 and 11 by Adi Shankara.
- **Udkriti Chandas –** 26 letters per pada and 104 letters in all.
- **Dhandakam – above 26 per pada**. **Ex.** Garuda Dandakam by Vedantha Diskshitar.

The reason I have given all the above is to emphasis that nothing in our religious scriptures, literatures are without a grammar, order, form or proportion. Everything confirms to a mathematical proportion.

Just like **Vedas are the breath of god, Chandas is the breath of every mantra.** The mantra is powerful only if it is recited to this metre. This is the reasons some of the poem even if it is not related to religion stands the test of time. Also if you recite a mantra even without knowing its meaning but as per the chandas then the mantra shall bear its fruit.

“Ma Nishada” …..Origin of Poetry:

Would you believe that the first poetry in this universe started with a curse?

Sage Valmiki saw a hunter who killed one of the birds of a pair which were in love. The sage got enraged and cursed the hunter, ***O, Hunter, may you not fare well anytime as you killed one of pair of the Krouncha birds which were happily engaged in love”.***

mA niShAda pratiShThA.n tvamagamaH shAshvatI.n samAH |
yatkrauJNchamithunAdekamavadhIH kAmamohitam.h || (bAlakANDa 2.14)

The above sloka also meant ***“O Lord of Lakshmi, it will bring you eternal glory for having killed a male of a happy couple who lost his head completely in lust..”*** This male is Ravana who lost his head in lust despite being happily married with his wife Mandodari. The other interesting connection I have this sloka with this series is that the **Mamuni Mayan** is the father of **queen Mandodari** and father in law of demon king Ravana.

Once the sage realized that this curse also meant something very auspicious and referred to Rama avatar he wrote Ramayana which is the origin of poetry as known in the world and it is called **Adi Kaavya.** Valmiki Ramayana set in **Anushtub Chandas** which is an octet.

So if you look at both the poetry and classical music composition:

- The rhythm in which it is set is a pulsation of time which confirms to a divine proportion.
- The raga and the pitch variation is a pulsation of space / frequency which again confirms to a divine proportion. We did not discuss about ragas though.
- The words used are divine and we have seen the divinity of the letters and words.
- The meaning of the song / poem usually exhorts the qualities of the god and hence is divine.
- The classical music and poetry invoke the subtle divine energy in the aural form just as much the sculpture (when it confirms to the divine proportion) invokes the subtle energy from in its visual form.
- This is how we see that “**one became many”** and **binding factors** in this process of evolution being the numerals 5 and 8.

Alas! Today, we are the fanatic fans of musicians who lacks morality & humility, who is ignorant of any divine proportion, most of the music we hear is a cacophony and very mechanical, the words are indecipherable and lacks depth, the meaning of the song - if at all any are just suggestive of sexual emotions and nothing else and the dance movements we see are the gyrations of the hip with a great blend of fitness training exercises associated with it. This is not to negate some of the fine works we see even today but to stress that we had such high science and taste which we seems to be losing / lost for no good reason other than ignorance.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

உ

27. TIME = SPACE SERIES. PART 27 – ARCHITECTURE – THE ZENITH OF MATHEMATICS (VAASTU)

Just look back in history and see what stands the test of time. Even if a civilization is completely extinct what stand beyond their time is the architecture of the civilization which stands tall as a witness to the richness and maturity of the civilization.

Aintiram states that **"Architecture (Space engineering) is the zenith of the mathematics**" in **"கட்டிட கலையே கணக்கியல் உச்சி**..." This space science is called as Vaastu Sastra in India.

There are few important concepts in Vaastu science which I want to present as below:

- The whole world is made up of **God's particle (Vastu, வஸ்து**) and nothing else. It is this God's particle which actually manifests itself as gross forms (**Vaastu, வாஸ்து**). You would note that Vastu (Subtle) has become Vaastu (gross) by addition of letter "a" and this difference is reflected in our languages too. **When அ / अ become ஆ / आ the subtle becomes gross**. Hence the most important statement for this science is "**Vastur eva Vaastu**", this means **the subtle becomes gross or energy becomes matter or the space becomes earth.** This is purely manifestation / evolution and there is no creation. This has nothing to do with religion but supreme science.

- Since it is all only one energy and it ever vibrates based on the harmonics and resonance, this vibration is called life and everything in this world, be it a living being or a non-living being as per our current definition – everything vibrates. The chair you sit and the vehicle you use are all living beings. **Everything throbs with energy and there is nothing which is a non-living being in this universe.** Without this vibration there would not be any difference between iron and cotton. The difference between particles is essentially the molecular constitution which is a function of its elements and its periodicity. This defines presence or absence of quality and sensory perceptions. This was put very nicely by **Poet Subramanya Bharathi** as எங்கு காணினும் சக்தியடா, **Where ever you see its just energy.**

- The above concept is very profound and this is called as **Science in our scriptures since it deals with Knowledge of the space** (விண் ஞானம், Vin-gyan).

- By the same concept when a subtle energy which is in the open space is converted to a gross form like a building then this body / building vibrates based on the design specifications and the material used for that building. It does not matter what is the measure and what is the material used, it vibrates. **When a structure is created with a divine proportion then we breathe life**

into it. Thus the structure becomes a living organism. This is how some of the structure stands the test of time for thousands of years.

- **Vaastu is that science which aligns the vibration of the building with the vibration of the dweller in the building so that both the living organisms (the building and the dweller) are in harmony and resonance.**

- It sounds simple but for this science we need to know includes the vibration pattern of the dweller, mathematics of space engineering, the knowledge of materials and its nature so that we can match it up to create harmony. This is precisely what Vaastu as a science does. Again it has nothing to do with religion. If you Google you would find thousands of Vaastu experts outside India who perceive and pursue this as a supreme spatial science.

- Let us look at the word **"vas"** which is the root. This means to shine; to grow bright, **to bestow by shining upon**, becoming light. **Vaas** is its derivative which means to perfume, an intoxicant, **dwelling place**, to assume the appearance of matter. This is the root for **SriniVAAS**, **SreeVAASan**, **VAASam** in Tamil means living. Most importantly the ability to feel this subtle thing (which cannot be touched) is called **VAASanai (**வாசனை, **in Tamil it means smell).**

- This spatial science includes knowledge on the following:
  - Nature of soil and material, and the qualities of its energy.
  - Measures for spatial engineering that best resonates with the environment and the dweller.
    - Note that the dwelling unit cannot be compatible to all the human beings universally. This varies with individual to individual and hence the Vaastu compatibility for the house is always seen for the housewife than the man with the assumption that what is compatible to the lady of the house would suit all others.

  - Accurate calculation of time and the position of astronomical bodies. The time engineering merges here with the space engineering. Consider the following to understand this:
    - In some temples you would notice that the sunlight falls on the deity on a specific day / time of the day. This is not possible if the future position of astronomical bodies at a given time is calculated very accurately.
    - When we build a house / temple the inauguration ceremony should be conducted at a specific time. So it's just not Space Engineering but time engineering also.

  - We see that the concept of space and time are not just a continuum but:

    - The measures of time and space merge and results in the same unit measure called Tala. It is Adi Tala in Poetry, Music and the same measure is used in building and sculpture dimensions.

- **Space is created because of the pulsation called Time** and hence the unit time measure (for pulsation) and the unit space measure (result of pulsation) are one and the same. This means the pulsation causes proportional displacement or creation of Space.

- It is the same concept used in meditation techniques. With Pranayama, Dharana and Dhyana you control the pulsation and align with it. In Samadhi you stop the pulsation and hence you are beyond time and space and you reach a luminescent state of resonant non-vibration.

Hence I named this book as **"Time = Space".**

We have seen that each subtle unit is represented in 8x8 energy grid and the gross unit is represented in a 9x9 energy grid. Mayan states that **"Each module or pada within the 8x8 / 9x9 structure resonates with a specific energy.** Based on Space, Time, Light, and Sound coordinates the frequency of vibration has its unique position in the Space/ Time continuum. **This energy level of frequency of vibration is called a luminous body or Devata**. This devata has particular attributes based upon its position and qualities in the Space/Time continuum (placement among the 64 / 81 sub-cubes or padas)".

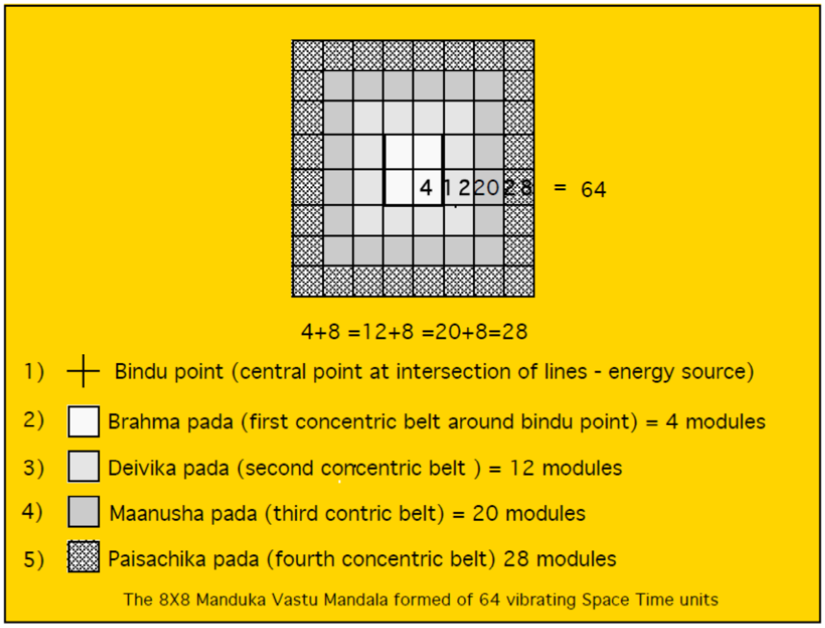
4 12 20 28 = 64
4+8 =12+8 =20+8=28
1) Bindu point (central point at intersection of lines - energy source)
2) Brahma pada (first concentric belt around bindu point) = 4 modules
3) Deivika pada (second concentric belt) = 12 modules
4) Maanusha pada (third contric belt) = 20 modules
5) Paisachika pada (fourth concentric belt) 28 modules
The 8X8 Manduka Vastu Mandala formed of 64 vibrating Space Time units

Figure 37: 5 states of 8x8 Energy Grid (Courtesy: Dr. Ganapathi Sthapati)

Now what is evident here in this science is 8 and 5 merges and so does Space and time.

- **Manifestation occurs in additive values of eight and in 5 stages. You can note that the Bindu (Point) has become 4, then 12, then 20 and then 28. (4+8=12 +8=20+8=28 with a total additive factor of 64). Manifestation occurs in sequences of five steps known as the Pentadic Order.**
- **The mathematical calculation which gives form to consciousness / space / subtle matter is called "Ayadi Gananam". We know that Ganam as the additive process.**

Dr. Jessica Marcey in her Fabric of Universe states that:

- The subtle energy point in the center becomes a self effulgent energy generator pouring out waves of energy
- becoming a self spinning stabilized structure of four padas / modules then
- adding eight units of energy to manifest the 4X4 structure of 16 units or modules; then
- adding another eight units of energy to manifest the 6X6 structure of 36 units or modules; then
- Adds eight more units culminating in 64 units or modules strung concentrically around the Bindu point or central generator.

So a central point with 4 concentric square belts around, each has its own frequency, vibration and energy characteristics evolving into a 9x9 unit as a gross matter. While this knowledge is significant of itself, when these phenomena are viewed in light of material manifestation as built space in architecture, the significance becomes stunning and profound.

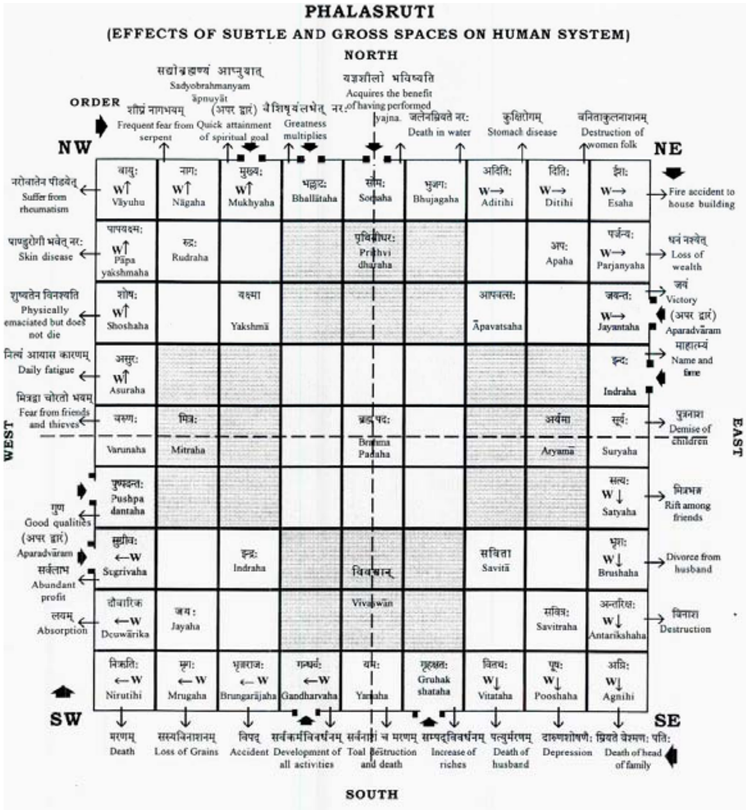
PHALASRUTI
(EFFECTS OF SUBTLE AND GROSS SPACES ON HUMAN SYSTEM)
NORTH
SOUTH
WEST
EAST
NW
NE
SW
SE
ORDER
Sadyobrahmanyam āpnuyāt
Frequent fear from serpent
Quick attainment of spiritual goal
Greatness multiplies
Acquires the benefit of having performed yajna.
Death in water
Stomach disease
Destruction of women folk
Fire accident to house building
Loss of wealth
Victory
Aparadvāram
Name and fame
Demise of children
Rift among friends
Divorce from husband
Destruction
Suffer from rheumatism
Skin disease
Physically emaciated but does not die
Daily fatigue
Fear from friends and thieves
Good qualities
Aparadvāram
Abundant profit
Absorption
Death
Loss of Grains
Accident
Development of all activities
Toal destruction and death
Increase of riches
Death of husband
Depression
Death of head of family
Vāyuhu
Nāgaha
Mukhyaha
Bhallātaha
Somaha
Bhujagaha
Aditihi
Ditihi
Esaha
Pāpa yakshmaha
Rudraha
Prithvi dharaha
Apaha
Parjanyaha
Shoshaha
Yakshmā
Āpavatsaha
Jayantaha
Asuraha
Indraha
Varunaha
Mitraha
Brahma Padaha
Aryamā
Suryaha
Pushpa dantaha
Satyaha
Sugrivaha
Indraha
Savitā
Brushaha
Dcuwārika
Jayaha
Vivaswān
Savitraha
Antarikshaha
Nirutihi
Mrugaha
Brungarājaha
Gandharvaha
Yamaha
Gruhak shataha
Vitataha
Pooshaha
Agnihi

Figure 38: 9x9 energy grid with corresponding energy levels for each grid.
(Courtesy: Dr. Jessica Marcey, Fabric of the Universe)

So, to sum up this section, we have seen that how every inanimate thing in this universe is still a living being and how with the spatial science we can pump throbbing energy into a building. This is the supreme science where Time and Space merges. This is the 4th sacred art we talked about which is a manifestation of the visual form of Om.

உ

As you would have noticed I have avoided talking about rituals and rules here but focused on the concept and philosophy. When we talk about rituals there are very many variations and we get into an unnecessary argument that whether it is rational or superstitious. Since most of the arts have become commercialized, the person’s qualification and his intentions are always questionable.

The objective of this book is to highlight the scientific aspects of the universe in a small way and emphasis that our culture is built on this science. Be it language or arts or religion all are built on scientific facts. They all were nourished and nurtured in thousands of Temples across India which is the container and protector of this subtle science and energy.

Let’s see some more interesting aspects of this in the next section.

Happy reading!

28. Time = Space Series. Part 28 – Capturing the divine energy – Squares & Circles

It is told that the sanctorum of Lord Ranganatha's temple at Sri Rangam is circle from outside and square from inside. Why?

- Subtle energy is represented by 8x8 square grid and the perimeter of the square is 4r and the area of the square is r*r. When each grid length is 1 unit then r = 8, so perimeter is 32 and area of this square is 64.
- Gross energy is represented by 9x9 square grids and with the same measurements perimeter is 36 and area of this square is 91.

It has been an architect's dream to create and capture subtle energy in a gross structure. This is fundamental objective of spatial engineering. How do you go about to achieve this? It looks that they found that they would take the help of a circle either outside the square or inside the square to achieve this. Remember almost every yantra has a square and a circle component in it.

The measurements are so taken such that either the perimeter or the area of both the circle and the 8x8 square are matched. Most of the Hindu temples follow such rules to invoke the subtle energy in a gross structure like temple. This is one of the well kept secrets and this is the reason each and every temple is said to interact with your subtle body differently when you visit it.

This is why Lord Ranganatha's sanctorum in Sri Rangam temple is a circle from outside and square from inside. This secret is called as **Squaring a Circle**. This is all about to creating divine energies and this is the science that is implemented in all the Hindu temples which are built as per the Sastra. You need just basic math knowledge and patience to understand this supreme secret and I shall give below 2 techniques to match the area or the perimeter of the circle with that of the 8x8 square.

Perimeter match:

- Let's look at the image of a circle of unit radius. The **area of the large circle is π** (πr^2) and the **circumference is 2π**. Here r =1.
- Draw 2 circles inside as shown with half the radius. The area of each circle is π/4 and the **combined area of both the circles is π/2**. The circumference of each circle is π and the **combined circumference of both the circles put together is 2π**.

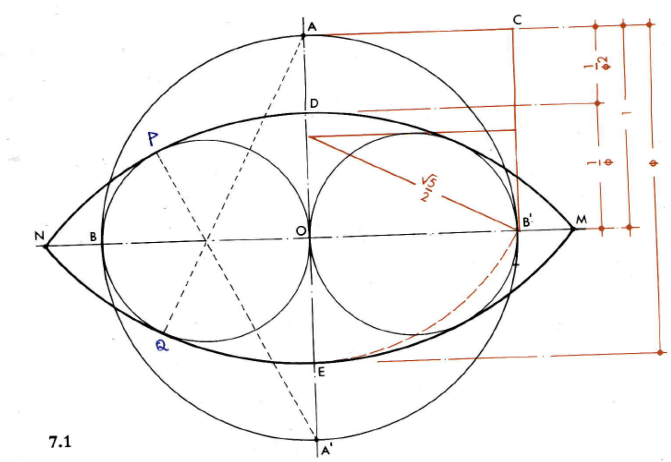

Figure 39: Squaring a Circle (Courtesy: Dr. Robert Lawlor, Sacred Geometry)

- Just note that when a larger circle divides itself as 2 smaller circles the area is halved but the circumference is the same. **This gives the first philosophical metaphor that "One has become two which are bipolar yet non-dual". This is one of the fundamental principles of Advaita and also in Chinese philosophy. This bipolar non-dual force is the basis of life and is represented as male / female and Yang / Yin.**

- Now draw 2 arcs NDM and NEM from A' and A with the radius as A'P and AQ as shown in the dotted line. You would note that point D and E would divide the radius of the circle by divine proportion – $1/\Phi$ and $1/\Phi^2$. **Radius of the arc AE is Φ. The Yin Yang symbol is based on this principle which embeds the divine proportion in it.**

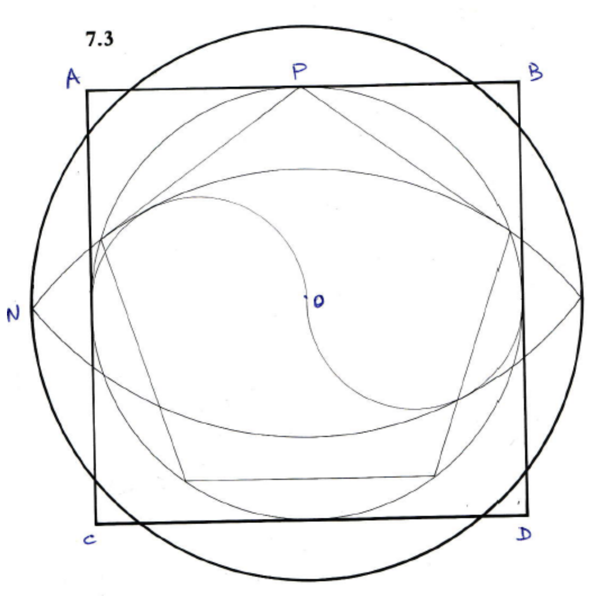

- Now let me draw a square ABCD around this circle (Fig 40) where the side of the square is the diameter of the circle which is unit 2. Here OP =1, AB=BC=CD=DA = 2. **The perimeter of the square is 8 and the area of the square is 4.**

- Now draw a circle with O as the center and NO as the radius as shown in the figure. Here NO = $\sqrt{\Phi}$. You can take this figure as such if you find it difficult to follow but for those who are interested, the derivation to calculate radius "NO" is as below:

Figure 40: Squaring a Circle 2 (Courtesy: Dr. Robert Lawlor, Sacred Geometry)

  - Consider a right angle triangle OPN. We know OP = 1 and PN which is the radius of the arc = Φ. Hence as per Pythagoras theorem $PN^2 = OP^2 + ON^2$ which is $\Phi^2 = 1+r^2$. Let us say the radius is ON = r.
  - So $r = \sqrt{(\Phi^2-1)}$. We know Φ is 1.61803399 so **$r = \sqrt{\Phi}$.**

- We have got a larger outer circle with a radius **$\sqrt{\Phi}$** and hence the perimeter should be **$2\pi\sqrt{\Phi}$. This is numerically equal to 7.9924576, which is equal to 8. Now we have created a structure where a square whose perimeter and a circle whose perimeter are the same.** This technique is not limited to India but used in the same way for pyramid construction.

- **This is the perimeter match between 8x8 square grids and a circle around it**.

Area match:

- Vaastu has an easier way to get this done. Draw a circle inside a 9x9 energy grid whose diameter is equal to the side of the square. If each energy grid length is 1 unit, then side of the square is 9 units and the radius is 4.5 units.

- The area of this circle is πr^2 where r=4.5 and hence the area of the circle is 63.64 which is almost equal to area of the 8x8 square with unit grid length as 1 unit.

- This is the area match between a circle and 8x8 square. So as per Vaastu every building is divided into 9x9 grid and the energy level of each grid is used to design the living space.

Capturing the direction of Subtle energy flow:

The subtle divine energy flows is as per the diagram given below. It starts from the center and always flows clockwise. That is why Hindus always stick to clockwise flow of movements - be in their temples or anywhere else. This clockwise flow of movement is captured in the sacred symbol Swastik. Here 1 is the source / Moolam, 2 is Kaalam (Time), 3 Seelam (Rhythm), 4 Kolam (Form), 5 (Universe) Gnalam.

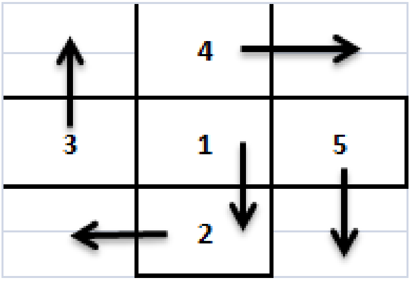

Figure 41: Clockwise flow of subtle energy

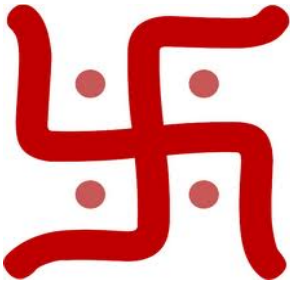

Figure 42: Swastik symbolizing energy flow

Relationship between √2, √3, √5, Φ and π:

- Now let me draw the connection between √2, √3, √5, Φ and π. Understanding this relationship is orgasmic and nothing less. If you think they are all some random numbers which cannot be fixed to their 10th decimal and not related to each other, just see the following relationship.

  $$\pi = \Phi^2 \times 6/5$$

  Philosophically 6/5 is the relation of a hexagon to a pentagon and the other way to express this relationship is

  $$\Phi = (\sqrt{\pi} * \sqrt{5}) / (\sqrt{2} * \sqrt{3}).$$

- This goes to prove that we did not just knew the numerals but divine fractions and proportions including their linkages. I am convinced, there is not even an iota of superstition in our temple and idol worship but they are supreme sciences codified and left for our benefits.

- The temple of **Sri Rangam** dates back from Ramayana, which is Treta Yuga that is 800000+ years back as per Hindu Scriptures. This is our rich past and supreme science is embedded in our culture and codified in thousands of temples across nook and corner of the country. Why should

> we ignore this? We have a lot more to offer this world which is struggling to understand the nature of nature.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

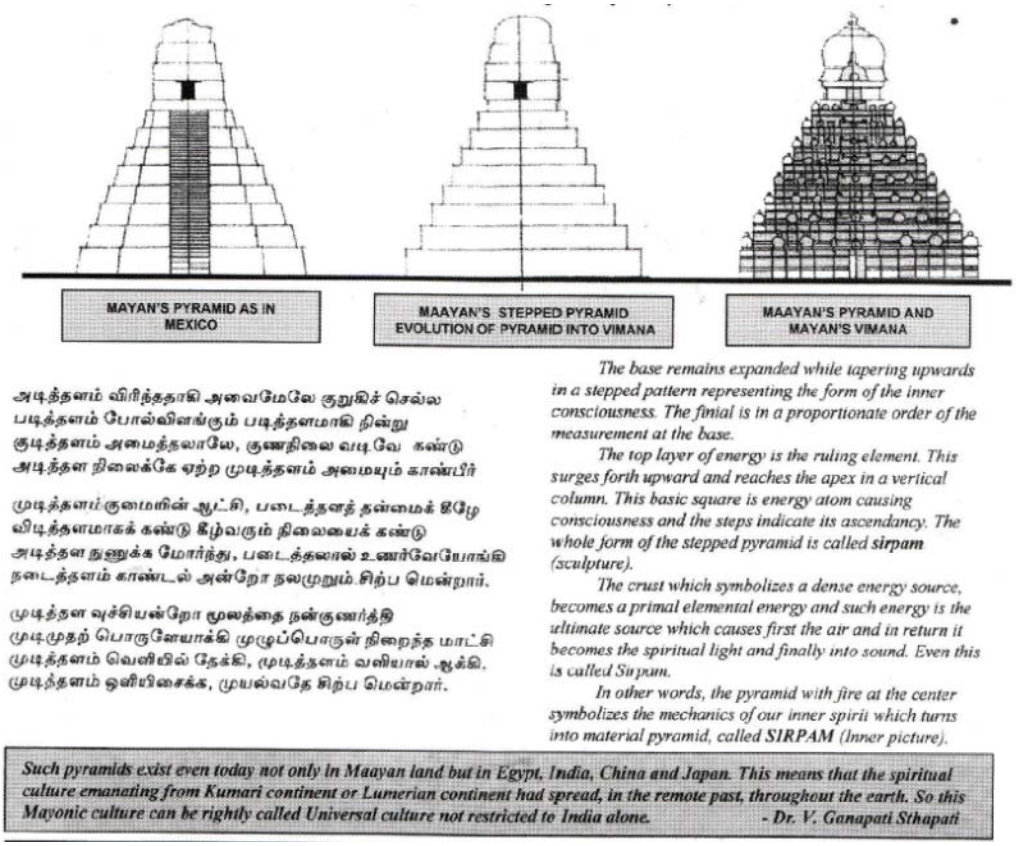
MAYAN'S PYRAMID AS IN MEXICO
MAAYAN'S STEPPED PYRAMID EVOLUTION OF PYRAMID INTO VIMANA
MAAYAN'S PYRAMID AND MAYAN'S VIMANA

அடித்தளம் விரிந்ததாகி அவைமேலே குறுகிச் செல்ல
படித்தளம் போல்விளங்கும் படித்தளமாகி நின்று
குடித்தளம் அமைத்தலாலே, குணநிலை வடிவே கண்டு
அடித்தள நிலைக்கே ஏற்ற முடித்தளம் அமையும் காண்பீர்

முடித்தளம் குமையின் ஆட்சி, படைத்தளத் தன்மைக் கீழே
விடித்தளமாகக் கண்டு கீழ்வரும் நிலையைக் கண்டு
அடித்தள நுணுக்க மோர்ந்து, படைத்தலால் உணர்வேயோங்கி
நடைத்தளம் காண்டல் அன்றோ நலமுறும் சிற்ப மென்றார்.

முடித்தள வுச்சியன்றோ மூலத்தை நன்குணர்த்தி
முடிமுதற் பொருளேயாக்கி முழுப்பொருள் நிறைந்த மாட்சி
முடித்தளம் வெளியில் தேக்கி, முடித்தளம் வளியால் ஆக்கி.
முடித்தளம் ஒளியிசைக்க, முயல்வதே சிற்ப மென்றார்.

The base remains expanded while tapering upwards in a stepped pattern representing the form of the inner consciousness. The finial is in a proportionate order of the measurement at the base.

The top layer of energy is the ruling element. This surges forth upward and reaches the apex in a vertical column. This basic square is energy atom causing consciousness and the steps indicate its ascendancy. The whole form of the stepped pyramid is called ***sirpam*** *(sculpture).*

The crust which symbolizes a dense energy source, becomes a primal elemental energy and such energy is the ultimate source which causes first the air and in return it becomes the spiritual light and finally into sound. Even this is called Sirpam.

In other words, the pyramid with fire at the center symbolizes the mechanics of our inner spirit which turns into material pyramid, called ***SIRPAM*** *(Inner picture).*

Such pyramids exist even today not only in Maayan land but in Egypt, India, China and Japan. This means that the spiritual culture emanating from Kumari continent or Lumerian continent had spread, in the remote past, throughout the earth. So this Mayonic culture can be rightly called Universal culture not restricted to India alone. - Dr. V. Ganapati Sthapati

Figure 43: Similarity between Indian temples, Pyramids. The Tamil text states how the structure should be built. (Courtesy: Dr. Ganapathi Sthapati)

உ

29. Time = Space Series. Part 29 – Heaven & Earth – Lord Nataraja & Lord Ranganatha

Let's stay with Lord Ranganatha and see the relationship between him and Lord Nataraja. Would you be surprised if one is the mirror image of the other? Look at the majestic picture of Lord Nataraja below. Wonder why this is always made in metal and not in wood or earth?

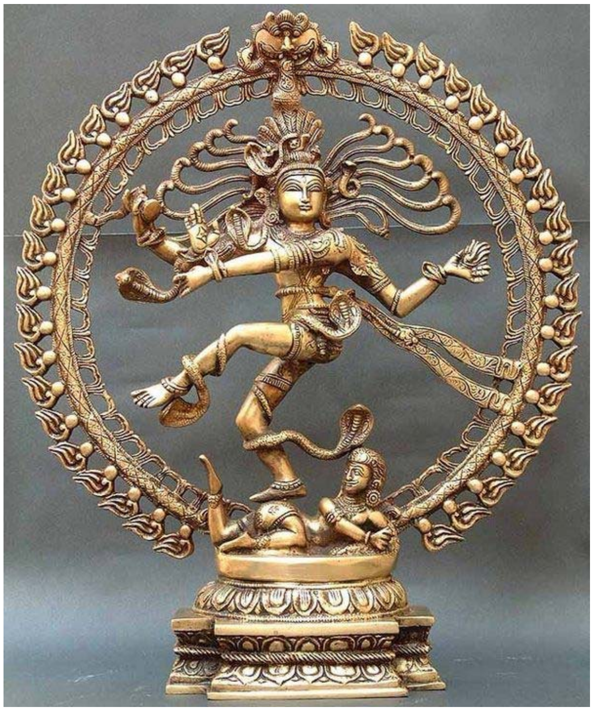

Figure 6 reproduced.

In Pancha bhoota (Space, Air, Fire, Water, Earth) Metal is always associated with Space and Wood is associated with Air. This system is reflected well in the Chinese system and it was not any different in our Vaastu system either. Since Lord Nataraja signifies Akasha tatva, **(Vin porul)** he is always made in metal – Copper or Gold or Bronze or a mix of it. The statues which were ancient had a generous amount of gold giving a luminescent look about the statues. The art and science of moulding metals to create divine forms with precise proportion dates back to thousands of years in India and the western experts think that this is a recent phenomenon. This is a complete surprise to the western world since they cannot believe that such sophistication existed before.

Dr. Ganapathi states, *"Lord Nataraja also signifies the primal image being expressive of the secret of creation in the process of which formless becomes instantaneously endowed with features and of perfect form. This is the reason why the statue of Lord Nataraja is called as **Silpa**, the primal image and everything else which is sculpted in this universe is called as **Pratima (replica)** and the original and one and only sculpture is the image of Lord Nataraja***"**.**

Mamuni Mayan states, *"Bhumih pradhana vastu Syat"*, means Earth by virtue of being the very basic support of all things, this is the primal and dominant Vastu". Earth and space are manifestations of each other in the cyclical process of evolution. They both should always be considered to be in unison. This vibrant, energetic, manifest earth energy (**Prithvi Tatva)** is called as '**Vishnu**".

In Vaastu satra, the Vaastu purusha is Lord Vishnu and understandably Vishnu is invoked as Vaastu purusha in Srirangam temple and offered a seat in the heart of Lord Ranganatha even today as a ritual.

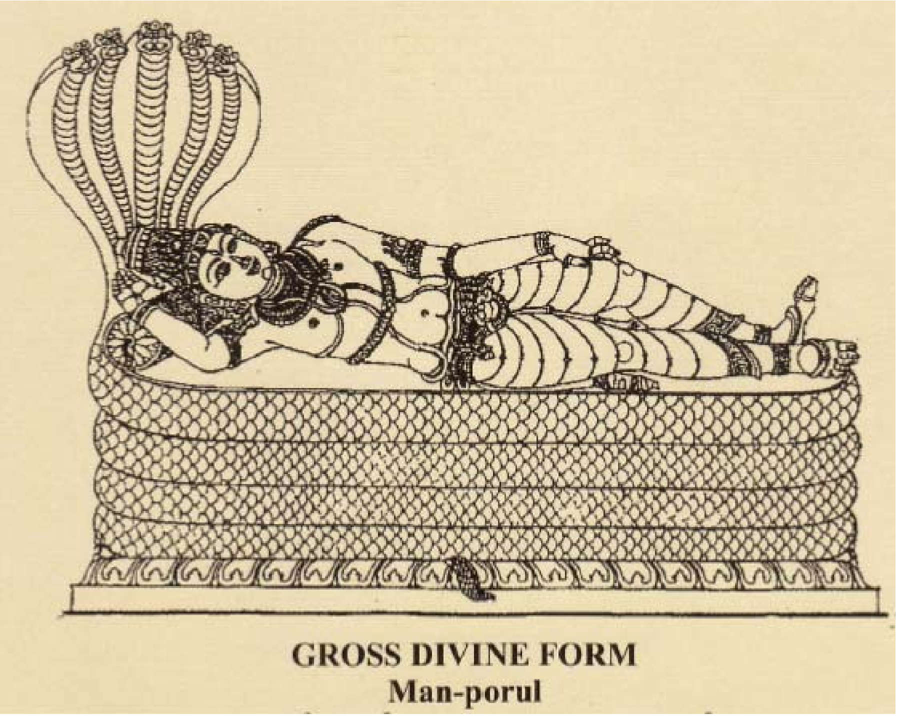

Figure 44: Lord Ranganatha in the reclining pose

Lord Ranganatha represents manifest earth energy **(Man porul)** and hence is always made out of seasoned earth only and it should be properly painted too. He is never to be made of metal. This earth form does not support frequent mobility and symbolizes stability and immobility. **This is the reason when Vibhishana kept the idol of Lord Ranganatha down at Sri Rangam it could not be moved.**

The energy that is in available in every universal form is represented as Vishnu and hence he is called as **Antaryami**. Once the subtle becomes grosser the vibration is reduced and the grossest form would not have any creative pulsation in it. It would still have the vibration based on its nature which ensures that it retains the form and shape. But it does not have any creative pulsation and it has reached the zenith of evolution. **This zenith is characterized by Lord Ranganatha in the reclining pose**.

The creative pulsation (Space) is the dance and the zenith of manifestation is a reclining posture (Earth).

உ

As we have seen, Lord Nataraja carries **Light in his left hand** and **sound in his right hand**. Lord Ranganatha carries **Chakra representing light in the right hand** and **Conch representing sound in the left hand.** The principle that the primal substance (space) has manifested as the universe (earth) and both of them are mirror images of each other is reflected in both the images.

In Chinese texts the Yang energy is associated with Heaven and the Yin as earth. Now associate the same heaven with Shiva, Earth with Vishnu and Tao with the God's particle - with some of the text below from **Lao Tzu's "Tao te Ching"**

"The Tao that can be told is not the eternal Tao.
The name that can be named is not the eternal name.
The nameless is the beginning of heaven and Earth.
The named is the mother of the ten thousand things.

Something mysteriously formed, Born before heaven and Earth.
In the silence and the void, Standing alone and unchanging,
Ever present and in motion, Perhaps it is the mother of ten thousand things.
I do not know its name. Call it Tao.

Man follows Earth. Earth follows heaven. Heaven follows the Tao.
Tao follows what is natural."

To sum up in the words of Dr. Ganapathi Sthapati:

- **Both the temples Chidambaram and Srirangam are south facing temples.**
- **Lord Nataraja signifies Akasha Tatva, outer space, macrocosm, subtle unmanifested energy is the Vastu purusha, is always in dancing pose, circumscribed in 8x8 square and made of metal.**
- **Lord Ranganatha signifies Prithvi Tatva, inner space, microcosm, manifest gross energy, is the Vaastu purusha is always in reclining pose, depicted in a circle or a 9x9 grid and made of earth.**
- **Lord Nataraja represents energy with matter and Lord Ranganatha matter with energy. In the equation $E = mc^2$, Lord Nataraja represents E and Lord Ranganatha represents mc^2.**
- Both these temples are a 'must see" and a "periodic see" types and you would be amazed that the qualitative change in your mental state even if you don't pray inside the temple but spend some time sitting or walking in the premises of the temple.
- **Both these temples are perfect symbols of ancient science codified for the benefit of the society and left for generations withstanding the test of time.**

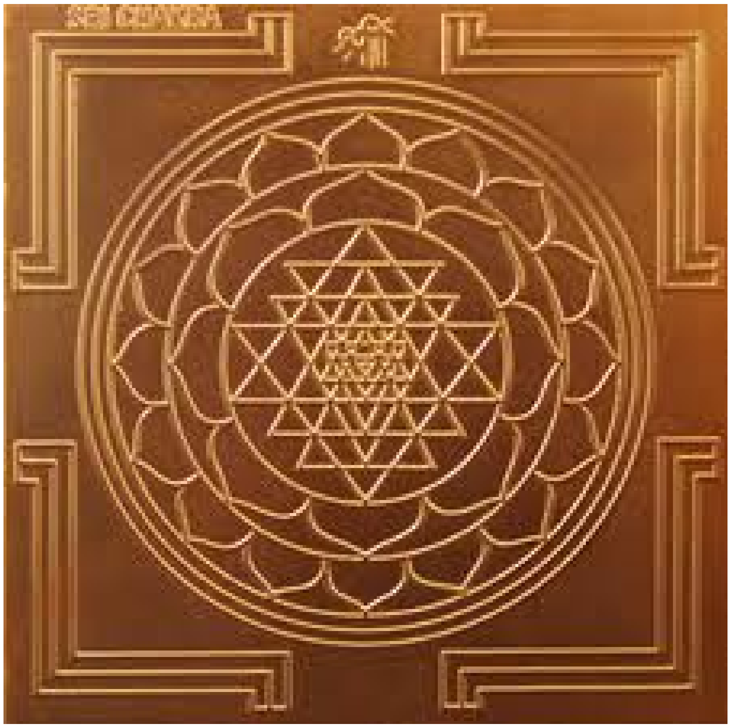

Figure 45: Sri Chakra – The Supreme shape in sacred Geometry

Chidambaram and Sri Rangam should be ground zero for any lab that wants looks into the secrets of space and time as they represent supreme science as much as they represent spirituality. We now see that science was the basis of our spiritual faith unlike it being 2 completely different aspects today across the world.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

May we be blessed by the divine power to unravel unknown secrets about the nature!

Happy reading!

30. Time = Space Series. Part 30 – Temples and Idol worship

Let us forget Hinduism, its symbols and names of god for the time being and consider these aspects as pure science from a nanotechnologist and cosmologist called Mayan.

We have been told that the universe is full of luminescent particles and this particle manifests itself because of the primordial pulsation. We have had a sneak peek into the geometry and the mathematical proportion which is used in the creation of many things. We know that the universe rhythm confirms to a mathematical code to the extent that we can create divine energies in a temple based on that science.

If time is pulsation of consciousness and if it creates, sustains and destroys everything, then time is the architect sculpting the universe and every object in this universe should be a sculpture of time and this concept is very beautifully captured in ***“Jagat Sarvaram Silpameva’*** *meaning the entire universe is nothing more than a shilpa / sculpture (and it latently conveys that Time is the Shilpi).*

If all the above things are right then is it not right to say that “Man created the concept of God” or “Man can create divine energy around him”? Let me sum up my thoughts as below:

- The sages and rishis were essentially scientists and understood the secret of nature very well. They were understandably masters in **nanotechnology and cosmology.**

- They knew that “**One thing”** with which they could understand the nature and science of all other things in this universe. This is obviously supported by the mathematical knowledge they had.

- As per Vedas nothing is created by god including the Vedas, but universe is the manifestation of god. They also understood that man is the most evolved of this manifestation. This is why they created all the god forms in the image of a man.

- Be it language, poetry, classical music, classical dance, sculpture and architecture it was the manifestation of this divinity that was the underlying link. Sacred geometry and sacred proportions were used to generate the subtle energy that took us closer to divine.

Hindu temples were created to capture all these aspects in one place and which could sustain and grow for the benefit of mankind.

- Secret of nature heard through the vibrations of Vedas were codified as supreme science and weaved into a philosophy called **Sanatana Dharma.**

- Architecture of temples confirmed to divine proportions and space - time engineering was accurate to capture sunlight and moon light at the same position for thousands of years. The design of the temple cleaned the subtle energies and points of the human body.

உ

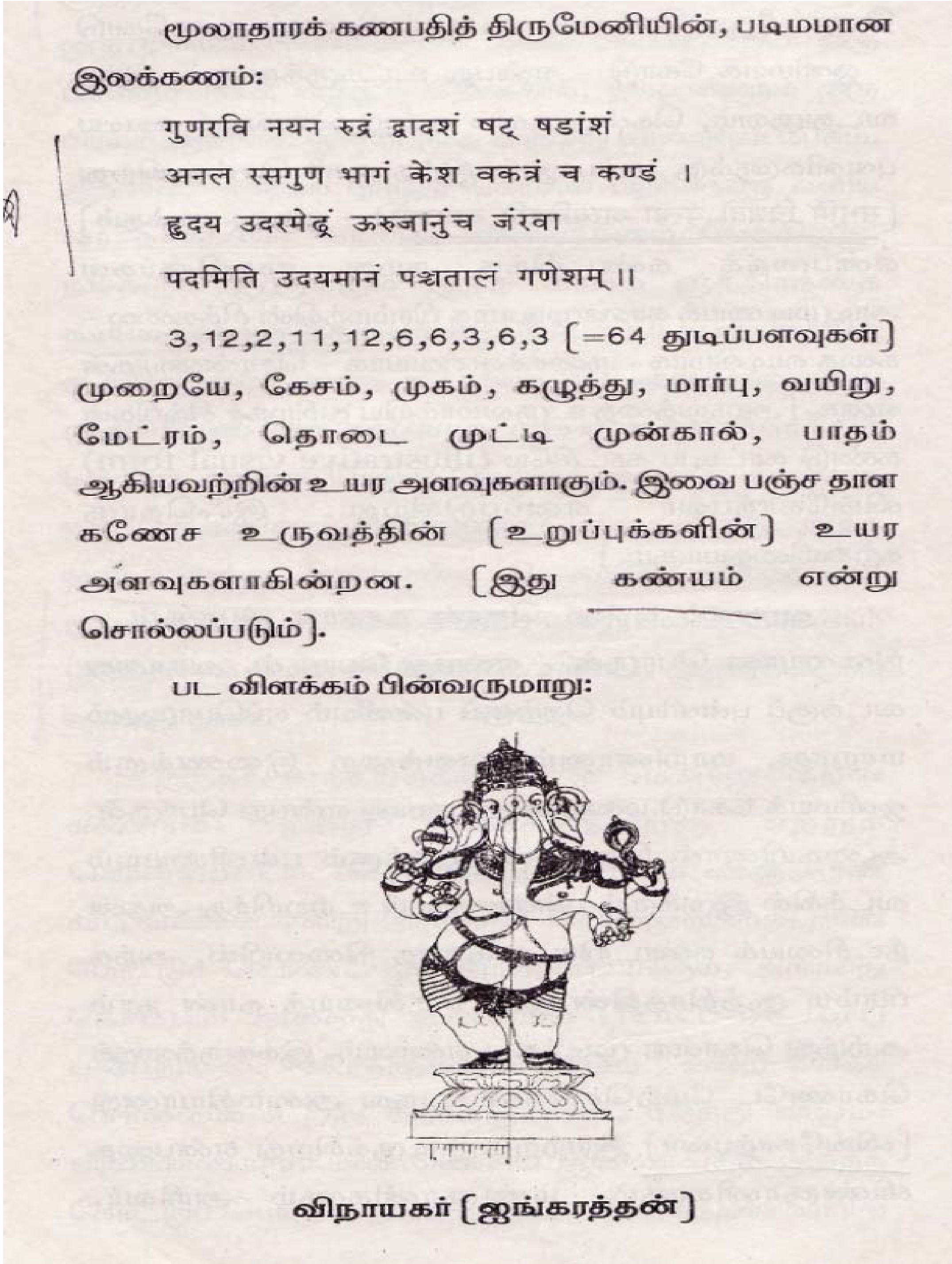

மூலாதாரக் கணபதித் திருமேனியின், படிமமான இலக்கணம்:

गुणरवि नयन रुद्रं द्वादशं षट् षडांशं
अनल रसगुण भागं केश वकत्रं च कण्डं ।
हृदय उदरमेढ्रं ऊरुजानुंच जंरवा
पदमिति उदयमानं पञ्चतालं गणेशम् ॥

3,12,2,11,12,6,6,3,6,3 [=64 துடிப்பளவுகள்] முறையே, கேசம், முகம், கழுத்து, மார்பு, வயிறு, மேட்ரம், தொடை முட்டி முன்கால், பாதம் ஆகியவற்றின் உயர அளவுகளாகும். இவை பஞ்ச தாள கணேச உருவத்தின் [உறுப்புக்களின்] உயர அளவுகளாகின்றன. [இது கண்யம் என்று சொல்லப்படும்].

பட விளக்கம் பின்வருமாறு:

விநாயகர் [ஐங்கரத்தன்]

Figure 46: Ganapathi Idol (Courtesy: Dr. Ganapathi Sthapati)

The sloka above in Sanskrit gives the measure of different parts of this Idol which would confirm to total 64 units (resonance of 8). Hair (3), Face(12), neck(2), Chest(11), Stomach(12), Belly &Pelvis (6), Thigh(6), Knee(3), Legs (6), Feet (3). Total 64 units. Just as much a Sloka with right metre and letters becomes a living organism, an idol with right proportions becomes a living organism.

- Divine Idol forms confirmed to a mathematical proportion that invoked subtle energies within and around us. This is where the external forms were used to create energies internally to a human. **This is the basis of idol worship. If it is scientific to create a form with divine energies based on mathematical proportion then idol worship is the supreme proof for that science.**

- **This supreme mathematical law is called as Dharma (nature's law) and the unfolding of that law is the Karmic theory of cause and effect / law of karma.**

- Poetry, classical music and classical dance are intimately linked to this Hindu temple culture. Adi Tala and octaves were used to create subtle energies in poetry using words, and the same tala maana is used in classical music with rhythm and raga, in classical dance with light and sound and in sculpture as a visual form.

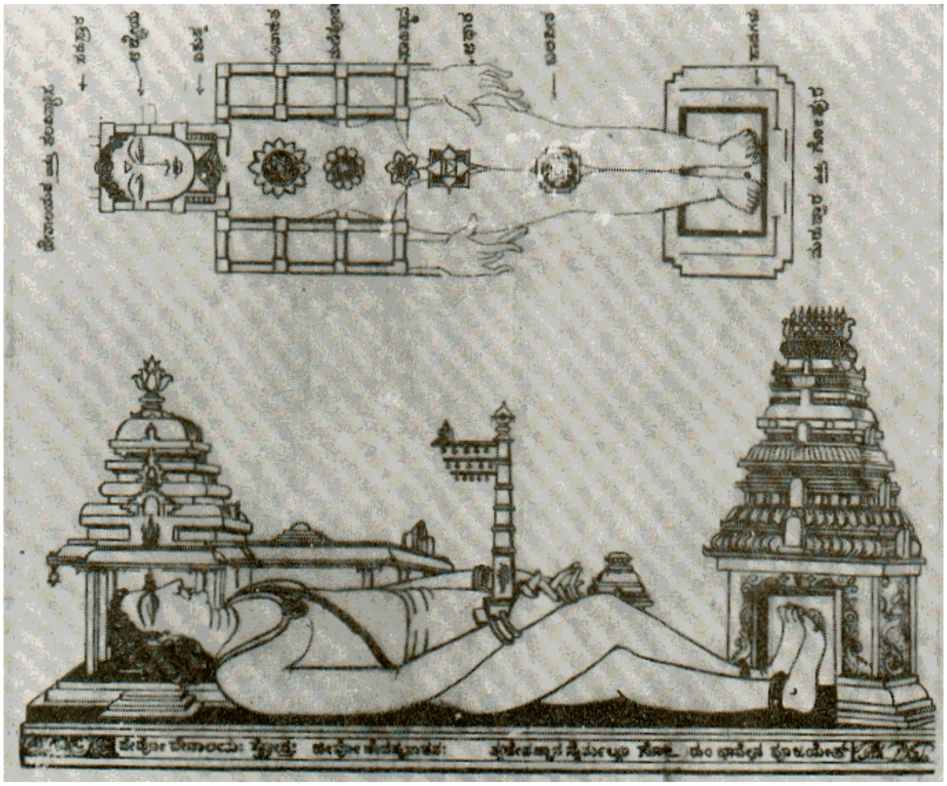

Figure 47: Human body as a representation of a temple
(Courtesy: http://www.salagram.net/sstp-mgpuja3.html)

A Hindu temple is a divine and yogic representation of a human being with the Deity in the temple representing the God as indweller in humans and all beings. In a temple the feet represents Rajagopura, the hands represent Praakaara, the abdomen represent Mandapa, the heart represents Antaraala and the crown of the head represents the sanctum sanctorum (garbha griha). The temple is used as a reminder that our inner spiritual journey is through internal yoga to realize the indweller God. This analogy is shown through the representation of various chakras namely Moolaadhara to Sahasraara in the body to various locations in the temple.

- It may be of interest to note that the same harmonics of 8 are reflected in every other domain. For example, if you have been practicing Pranayama then your master would have asked you to practice breathing based on some counts. The most popular counts are 6(inhale), 3 (Exhale) or 7(inhale), 2 (Exhale). If you notice this count 9, this cleans up your gross body. Advanced Pranayam students may be asked to stick to 1 (Inhale), 4 (Retain), 2(Exhale), 2(Retain outside) or as a variation 1,4,2,1 or 1,3,2,2. The counts which add up to 8 clean the subtle body.

- Till few decades back the fundamental dimension for money in India was “**Ana”** called as **Eight Anas (எட்டணா)** and this was equivalent to Rs. 0.5 as the fundamental measure.

- Thus a **“Poetry is an idol in aural form and a building is a frozen music”**. A Poet by the way of his mental frequency has first become the poem itself and then the actual poem was created in its aural form and an artist by the same frequency has become the idol first and then he gave a shape to the idol externally through this hands. This is very beautifully given as ***“Raso vai saha, So Vai rasikaha”, which means he (god) is the enjoyer and he is the artist.***

- All of this reflects the supreme philosophy that **“Oneness in all and everything is just a manifestation of that one thing”**.

- Now in this light understand the meaning of this Upanishad prayer, *“Aum Poornamadah Poornamidam Poornaath Poornam Udachyathe, Poornasya Poornamaadaaya Poornameva Vasishyathe”. This means God is perfect (infinite). This Universe is also perfect (infinite). If perfection (infinity) is taken from anything perfect (infinite) what remains is still perfect (infinite).* Isn’t this is the actual reality today?

- Everything scientific is philosophically, mythological and ritually well interwoven that proves that Sanatana Dharma is the most ancient and supreme representation of science and nature’s secret in our possession. Our temple worship which integrates all this science, 5 sacred arts with language and Math is the standing example of our understanding of the nature of god.

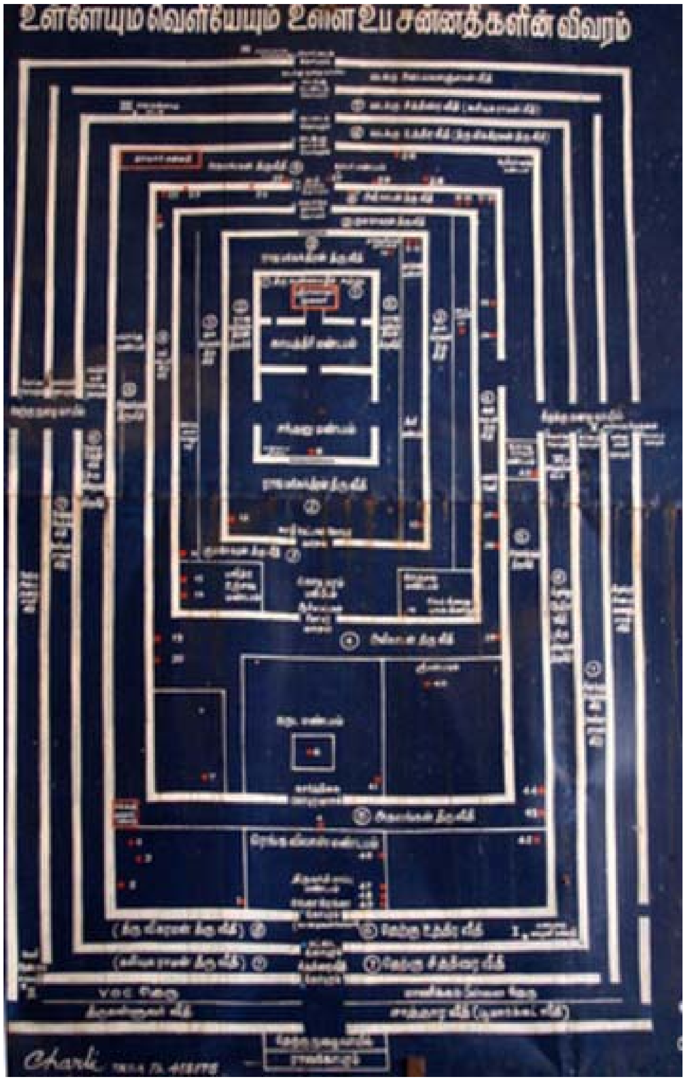

Figure 47: Layout of Sri Rangam temple with 7 Prakaras

It is the same Sugarcane but an elephant eats it from the field directly, a young adult peels the skin off and chews it, if a kid has to eat then someone has to peel the skin and cut it into pieces and an old man without teeth needs it to be made as sugarcane juice. Depending on one's ability to chew and digest the food has to be processed.

Same way depending on one's mental ability to grasp the supreme truth it has to be processed and presented. If someone is mentally well evolved then they can be fed with Upanishads and he can meditate on the formless Brahman. To a young adult Upanishads may not be palatable directly but they may have to start with Upavedas and Vedangas to mature to a higher level. For the less evolved kids, Itihasa and Purana like Ramayana and Mahabharata with stories may drive home the point and they can start with Shlokas and mantras. But for the least evolved it has to start with a personal deity with name and form, bhajans and songs to invoke the concept of Bhakthi.

In all the four cases temple culture forms the lowest common denominator. Even the most evolved relish the arts and architecture in the temples. We should be proud of our rich cultural and scientific heritage and be passionate about the science and philosophy and be rational about rituals and symbolisms.

Image / Idol worship is not deplorable or despicable. This is the highest scientific form of worship which enables us to see face to face the one which is beyond reach of sight and speech. This kindles devotion, community development and helps one to climb the steep walls of spirituality in the initial stages. Every religion in this world has some image or idol as symbolism to achieve this objective.

An idol is a god in form when made with the science of divine proportion and seasoned material is the highest product of our intelligence till today.

It should however be noted that this idol worship should not be lost in the din of rituals which does not reflect the very philosophy it had come to symbolize. So the advice I follow is *“Be passionate about our culture but be rational about the rituals.*”

The advancement of science can progress rapidly when the western scientists who are objective and empirical, partner with the pundits in India and learn the science of the subtle from our scriptures and temples. For this to happen we need to learn our own scriptures and past which is not possible without learning our mother tongue and Sanskrit. Else in the near future we may have to import knowledge of our culture and rich past from the western world.

As a wise man said, **“Indians are the most ignorant about their rich past and scientific traditions.”**

Happy reading!

31. TIME = SPACE SERIES. PART 31 – MAYAN & HIS WORKS

Mahamuni Mayan was the father-in-law of Demon King Ravana and is referred in both Mahabharata and Ramayana. He is attributed with building the palace for Yudhishtar in Indraprasta and also the architect for Sri Lanka which Ravana ruled. His son Nalan helped Lord Rama build the Ramasethu – the bridge between India and Sri Lanka. He is believed to have lived in the **Ilamuridesam** popularly known as **Lemuria continent (Kumari Kandam**) which spanned 4500+ kms from Sri Lanka to Antartica connecting current Australia, Africa and Americas. Check out http://www.sacred-texts.com/atl/tll/index.htm

Understandably when this land mass was consumed by the sea their clan moved to different parts of the world including Guatemala, Egypt, and to current south Indian region. This explains as to how we find common science across India, South America and Egypt.

He has written books in both Sanskrit and Tamil and I am going to produce the list of technical works he has created and look at the contents of a specific book called Surya Siddhanta. Just the list of works and also the topics in Surya Siddhanta is amazing, highlighting the depth of science we had at that time. http://www.youtube.com/watch?v=PZKiCpFisoY

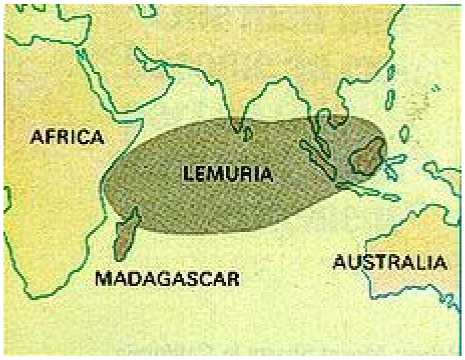

Figure 48: Kumari Kandam depiction

The Surya Siddhanta

The most ancient treatise on Astronomy is a treatise authored by Mayan. It is called *The Surya Siddhanta*. Included in this treatise are the following chapter topics following:

- The Motions of the Planets
- The Places of the Planets
- Direction, Place and Time
- The Moon and Eclipses
- The Sun and Eclipses
- The Projection of Eclipses
- Planetary Conjunctions
- Conjunction of the Stars
- Risings and Settings
- The Moon’s Risings and Settings
- Certain Malignant Aspects of the Sun and Moon
- Cosmogony, Geography, and Dimensions of the Creation
- The Gnomon
- The Movement of the Heavens and Human Activity

உ

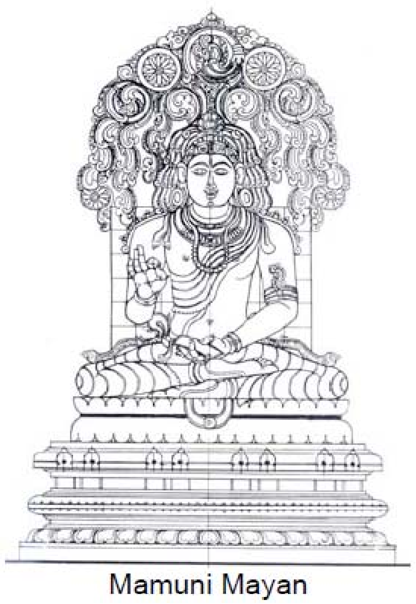

Figure 49: Mamumi Mayan

Mayan is also attributed with the start and spread of Temple culture in India. The song below in Tamil states that since Mayan sat under the Banyan tree and taught this science of universe and also of the sacred arts, so that every town was endowed with Temples for social betterment. Because he was from the southern region and also he sat under the Banyan tree, he is also equated with Lord Dakshinamoorthy.

ஆலமா மரத்தின் கீழே அருள் மயன் அமர்ந்ததாலே
ஆலமா மரத்தின் கீழே அருட் கலை வளர்ந்ததாலே
ஆலமா மரத்தின் கீழே அறிவர்கள் வளர்ந்ததாலே
ஆலயம் ஊரும்தோறும் அமைந்தன செழிக்கநன்றே!!

The Mayan says in his Pranava Veda verse one:
*"Om Light and Om Sound are the Primal Source of all manifest forms. Om Light is aroused by its own effort in a state of disorder and appears as a flame. The state of Om Light and Om Sound in Space is a magnificent luminous six – faced Light form that is called "**murukoli**". The transformation of Om Light and Om Sound through the five stages is concealed in the five fold knowledge, of which, this is the first. This process of transformation of disorderly Om Light and Om Sound into orderliness is found in all five fold material forms."*

Sage Veda Vyasa says in the Bhagavatam, *"eka eva pura vedo pranavha sarva vangmayha" (9th skandha, 14th chapter, sloka 48 of Bhagavatam).* This means that there was only one Veda called Pranava Veda. It is believed that Sage Vyasa created 4 vedas from this Veda since he believed that the oncoming generations would not have the mental strength to master everything.

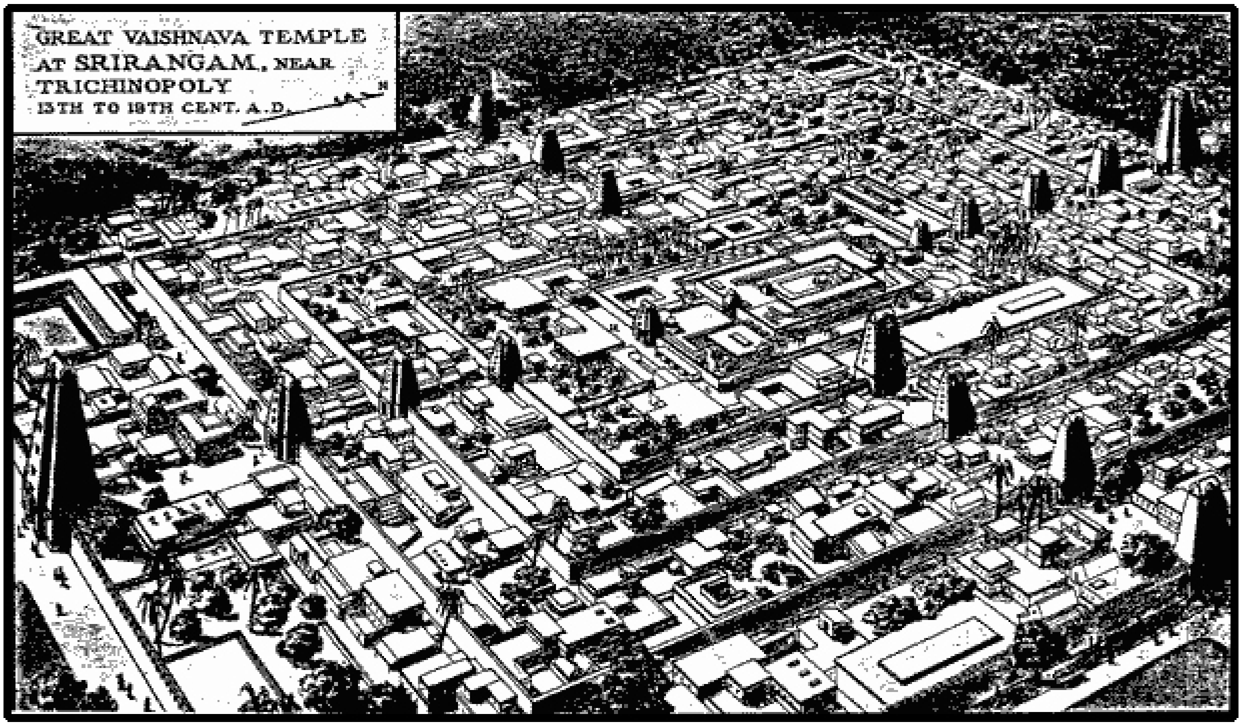

Figure 50: Aerial view of Sri Rangam temple with 7 Prakaras

Technical works

The following are technical works written by Mayan and represent the 12 vowels of the Tamil language:

- Ovia Chennool - Treatise on Drafting and Painting
- Sirpama Chenool - Treatise on Iconometry
- Kattida Chennool - Treatise on Architecture
- Nilamanai Chennool - Treatise on House Building based on quality of land
- Manainila Chennool - Treatise on Land based on the nature of house building
- Baniyal Chenool - Treatise on Astro Physics
- Perunata Chenool - Treatise on Divine dance
- Muligai Chenool - Treatise on Herbs
- Ganitama Chenool - Treatise on Mathematics
- Arakkala Chenool - Treatise on Ship Building
- Vinkala Chennool - Treatise on Space Ship
- Elisai Chennool - Treatise on Science of Music

உ

MAYAN, in Aintiram *states that:*
"There exists an order in the Universe - subtle universe, and material universe. This power is attributed to the consciousness of the cosmic space as well as of the inner space of the animate beings. This consciousness, by its personal effort to express its own inner feelings, causes a kind of vibration or pulsation in the inner space, resulting in energy-grids. The vibration or pulsation, being the causal element of all these events, is called KAALA. This is actually the force of energy aroused by the inner consciousness, causing waves and contributing to the growth of living forms. This KAALA resides in all living beings. So, space is the offshoot of the vibration of the primordial energy. This is how the space was born. All forms of nature are manifest forms of subtle energy. For all to get manifested, the force is KAALAM. This wave-form frequency realm is the creative element of the universe."

There is no other treatise that we have in our possession today which is dated 10000 BC that depicts the secret of nature so explicitly and gives the unmanifested interconnections and mathematical proportions between everything in the universe.

As Dr. Ganapathi Sthapati rues "Unfortunately most of the Tamil experts don't understand the science in these texts and hence these books are idling without anyone extracting the truths from it". In the next few generations when we are less equipped in our own mother tongue we may lose this treasure completely just like the many we have lost so far.

Hence the first step we need to ensure is even if you are not equipped well in your mother tongue please make sure that your kids can read and write it well. This becomes our most important duty. And I find it worrisome when someone is indifferent to the fact that they can understand their mother tongue but cannot read and write.

With some significant efforts we can and we should prove that this man who said, **"Indians are the most ignorant about their rich past and scientific traditions."** is not wise any more.

Happy reading!

32. TIME = SPACE SERIES. PART 32 – CONCLUSION

What I have understood in a nutshell is as follows:

- The inner being of both - the individual and the universe is luminescent consciousness which has the ability to become aware of it. This luminescent consciousness is called by various names but it is essentially the only existent Vastu in this world. This Vastu manifests itself as time, space and spatial forms.

- ***"Aanor aniyaan mahato Mahiyaan!"*** this means what is in the atom is the same makes up the universe. That which is present in the atom is the Space energy and the meaning of the word Akasa is stated as ***"Aa kashayati aakasa ha!,*** That which imparts light to others or make others shine is called Akasa". This means that Space Vastu is luminescent. I am stressing the aspect of luminescent again and again for you to start perceiving God as light.

- Pranava is the causal element which triggers this evolution process and this manifests as visual and aural forms. Pranava results in the primordial pulsation and this pulsation is triggered and sustained by a divine progression and proportion which is called **Absolute time**.

- Absolute time creates Absolute Space. These 4 things – Absolute time, absolute space, Aural Pranava and Visual Pranava combine together to form spatial forms and hence this universe.

- The Vastu as God's particle, its characteristics, the Pranava, the pulsation, the rhythm, the progression, its proportion and the resultant energy states – all the scientific facts, theories and processes are captured and codified as a Dharma which is called as Hinduism today. The science behind this subtle energy is reflected in various art forms that includes Poetry, Classical music, dance, Sculpture and building architecture.

- All these manifestations are harmonics of 5 and 8 differentiated by the Golden proportion / God's ratio to reflect the various states of consciousness. **This intrinsic order is called Dharma and the cyclic nature of this evolution is the Karmic code embedded in the universe**. **Nothing in this universe is superior to the Dharma and Karmic code.**

- The divinity manifests in us and the space within us is nothing different from the macrocosm. Consciousness which exhibits itself as a state of being because of our mental vibration can evolve to a higher plane to devolve back into the luminescent core.

- Sanatana Dharma popularly today known as Hinduism is the repository of this supreme secret not just as a philosophy but integrated with every stage of our life. This can be understood better only if we can give due respect and attention to the language and scriptures, our spiritual tradition interwoven with arts and architecture that contains these treasures.

- The temple culture & Idol worship that we have symbolizes the best of science known to human beings and we need to understand this and preserve this in the best possible manner.

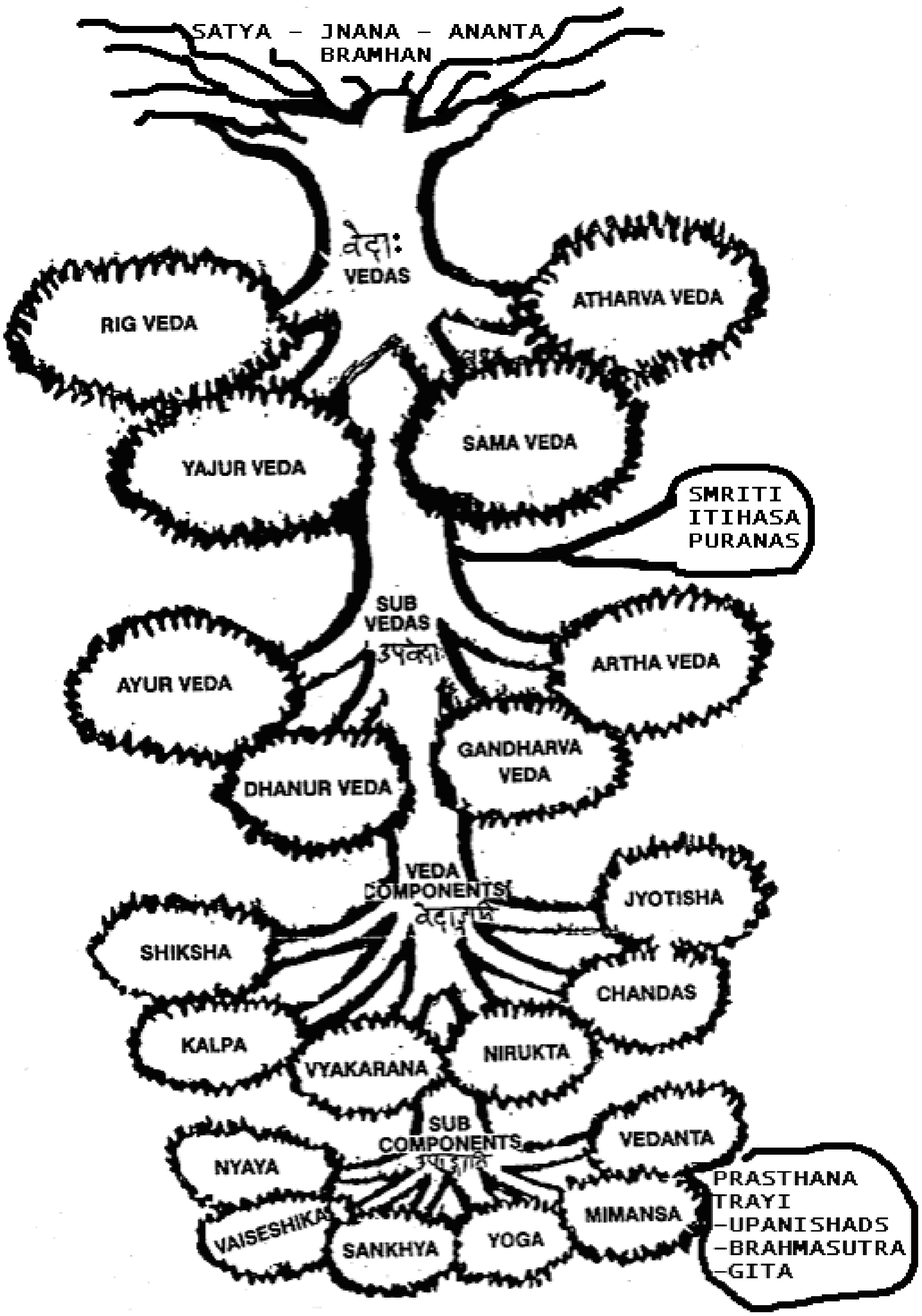
SATYA – JNANA – ANANTA
BRAMHAN
वेदाः
VEDAS
RIG VEDA
ATHARVA VEDA
YAJUR VEDA
SAMA VEDA
SMRITI
ITIHASA
PURANAS
SUB
VEDAS
उपवेदाः
AYUR VEDA
ARTHA VEDA
DHANUR VEDA
GANDHARVA
VEDA
VEDA
COMPONENTS
वेदाङ्गानि
JYOTISHA
SHIKSHA
CHANDAS
KALPA
VYAKARANA
NIRUKTA
SUB
COMPONENTS
उपाङ्गानि
NYAYA
VEDANTA
PRASTHANA
TRAYI
-UPANISHADS
-BRAHMASUTRA
-GITA
VAISESHIKA
SANKHYA
YOGA
MIMANSA

Figure 51: Vedic Knowledge as an inverted tree with Brahman as the roots signifying that everything has a common root. (Courtesy: Internet)

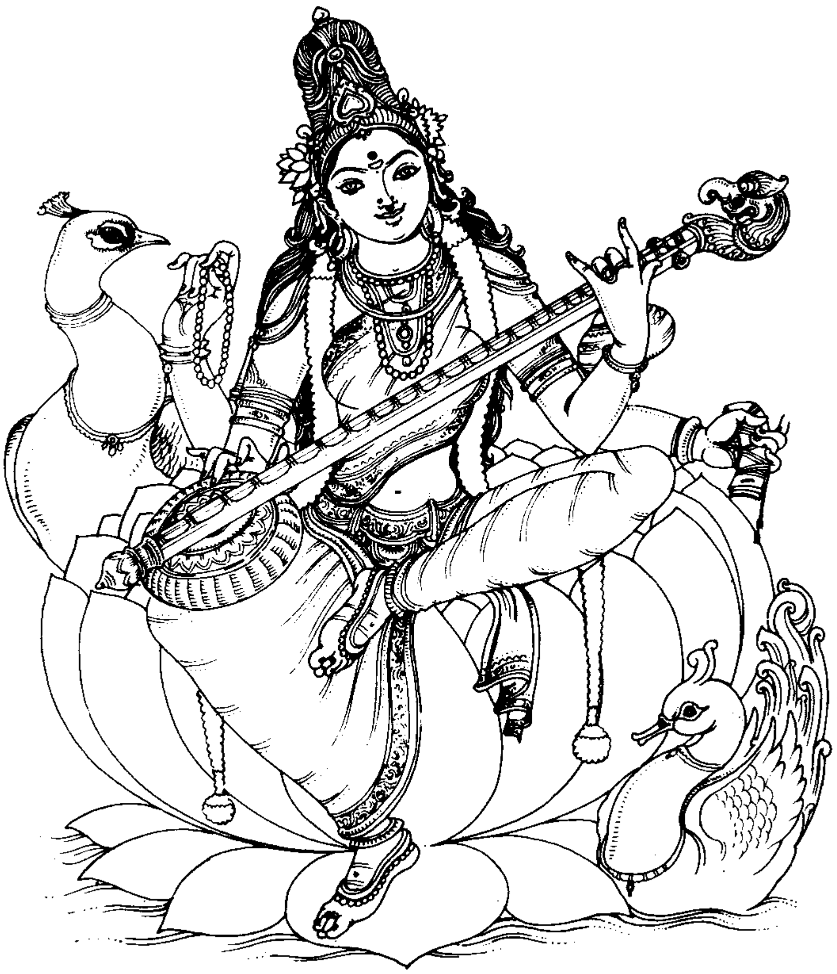

(Figure 52: Goddess Sarasvati. Courtesy: Internet)

She is the Goddess of speech, music and knowledge. She carries the rosary beads (akshamala) represents the alphabets of a language (Sanskrit), the musical instrument (Veena) and Vedas as palm leaf manuscript in her hands that denotes the knowledge. Thus we can see that idol worship signifies deeper scientific aspects of a common root between Language, Music, knowledge and arts.

Figure 52: Musical instrument Veena as a representation of Human Body (Sarira veena) (Courtesy: Internet)

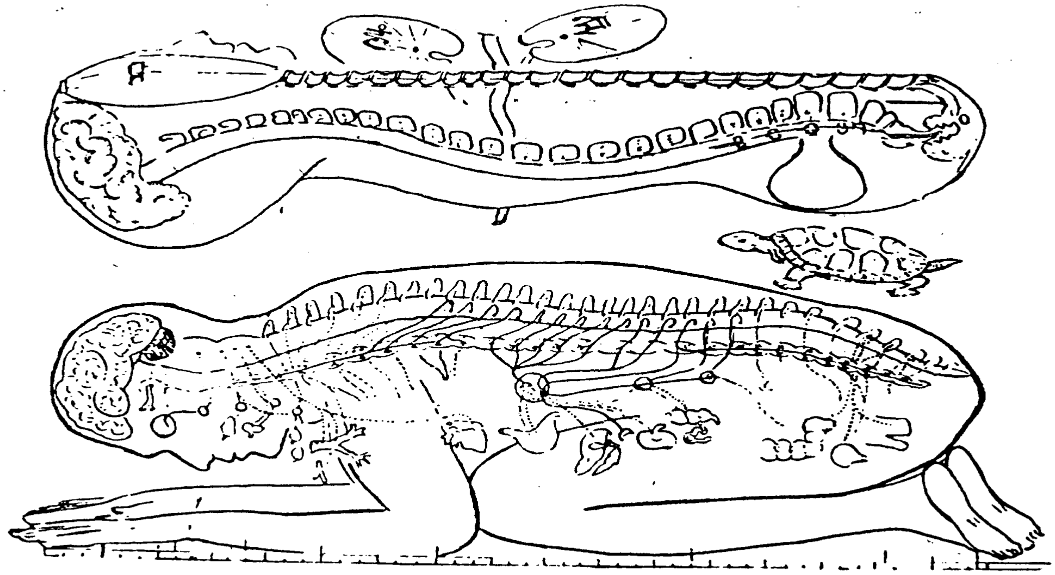

உ

I am happy to note the coincidence that this book has 32 chapters with 108 pages – both being key universal numerals. Chapter 19 (the golden mean of this book) deals with the Golden mean or God’s ratio of nature is again coincidental.

I would like to thank you for reading this book. I would like to state that I am neither adept in any of the topics discussed in this book nor an expert in any of the arts including the languages Tamil or Sanskrit language.

A curious mind driven by my passion to understand the underlying aspect in all things and my liking to reading has helped me to compile this book. My exposure as a student to Chinese martial arts, Indian yogic studies, Pranic healing and Siddha philosophy helped me to connect of the points I had articulated. Any authoritative statements in these books reflect my belief in the topic or from the author / teacher from whom I subscribed it and it is certainly not reflective of my abilities.

I am open to constructive criticisms on any of the topics and I can be reached at the mail ID shared in the first page of this book. My only request to you is *“Please share this book or parts of this book without any hesitation to your friends and family and help us dispel the ignorance about rich ancient Indian culture”*.

I shall end this book with some of my favorite quotes of Dr. Einstein.

“My religion consists of a humble admiration of the illimitable superior spirit who reveals himself in the slight details we are able to perceive with our frail and feeble mind. That deep emotional conviction of the presence of a superior reasoning power, which is revealed in the incomprehensible universe, forms my idea of God. We still do not know one thousandth of one percent of what nature has revealed to us.”

Satyam Param Deemahi!
May we meditate on the supreme truth!

TIME & THE TECHNOSPHERE

The Law of Time in Human Affairs

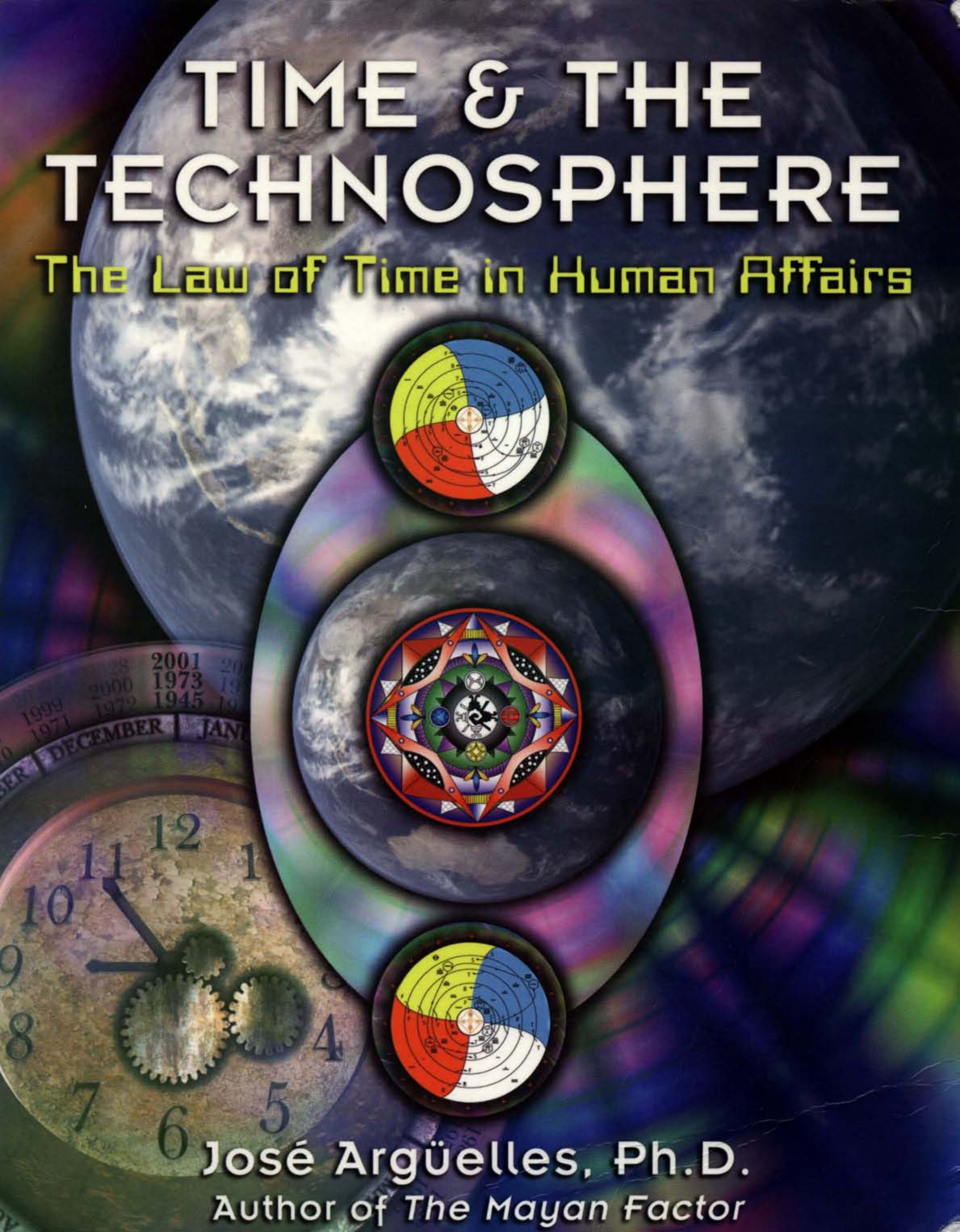
2001
1973
1945
2000
1972
1999
1971
DECEMBER
JAN
12
11
1
10
9
8
4
7
6
5

José Argüelles, Ph.D.

Author of *The Mayan Factor*

MAYAN STUDIES / NEW CONSCIOUSNESS $20.00

"Time and the Technosphere *makes a vital contribution to the great paradigm shift in human cultures that is now accelerating. There is a higher dimension of natural time inscribed in the cosmos and a natural calendar that would help humanity mature into the awakening global consciousness that is now essential for future sustainability and well-being. This book is an important contribution to the awakening event that is now emerging on a planetary scale.*"

ASHOK GANGADEAN, PH.D., Professor of Philosophy, Haverford College

In *Time and the Technosphere,* José Argüelles presents a groundbreaking study that distinguishes the natural time of the cosmos from the artificial mechanistic time under which we currently live. Argüelles defines the actual nature of time as the frequency of sychronization. Applying this Law of Time to an understanding of the entire system of life on Earth, he shows that in order to not destroy Earth's ability to sustain life, we must change our definition of time and adopt a natural harmonic calendar based on the 13-moon 28-day cycle. Until the creation of the Gregorian calendar and the 60-minute hour, most of humanity lived by the 28-day cycle of natural time. The adoption of artificial time has subjected us to a 12:60 time frequency that governs the entire global industrialized civilization—the technosphere. With the collapse of the Twin Towers on September 11, a fissure was created in this artificial technosphere, opening up the noosphere (Earth's mental envelope).

Humanity has a golden opportunity to leave the strife of the past and enter a time of peace by adopting a harmonious natural calendar that will repair the damages caused by the irregular tempo of technospheric time. Our last best chance to adopt this natural time and step into the bright new future promised by the galactic shift of 2012 is the Great Calendar Change of 2004, a new discovery based on the author's mathematical research into the Mayan calendar first begun in his landmark work *The Mayan Factor.* In *Time and the Technosphere,* Argüelles reveals the clear distinction between third-dimensional astronomical time and the fourth-dimensional synchronic order of the Law of Time, which holds enormous potential for the future of humanity.

PHOTO BY BORIS VLADIMIROVICH

JOSÉ ARGÜELLES, PH.D., works full time on behalf of the World Thirteen Moon Calendar Change Peace Movement and is president of the Foundation for the Law of Time. He is the author of numerous works, including *The Mayan Factor, Earth Ascending,* and *Surfers of the Zuvuya.* He lives in Oregon.

BEAR & COMPANY
ROCHESTER, VERMONT

ISBN 1879181991

Cover design by Peri Champine
Cover illustration by José Argüelles

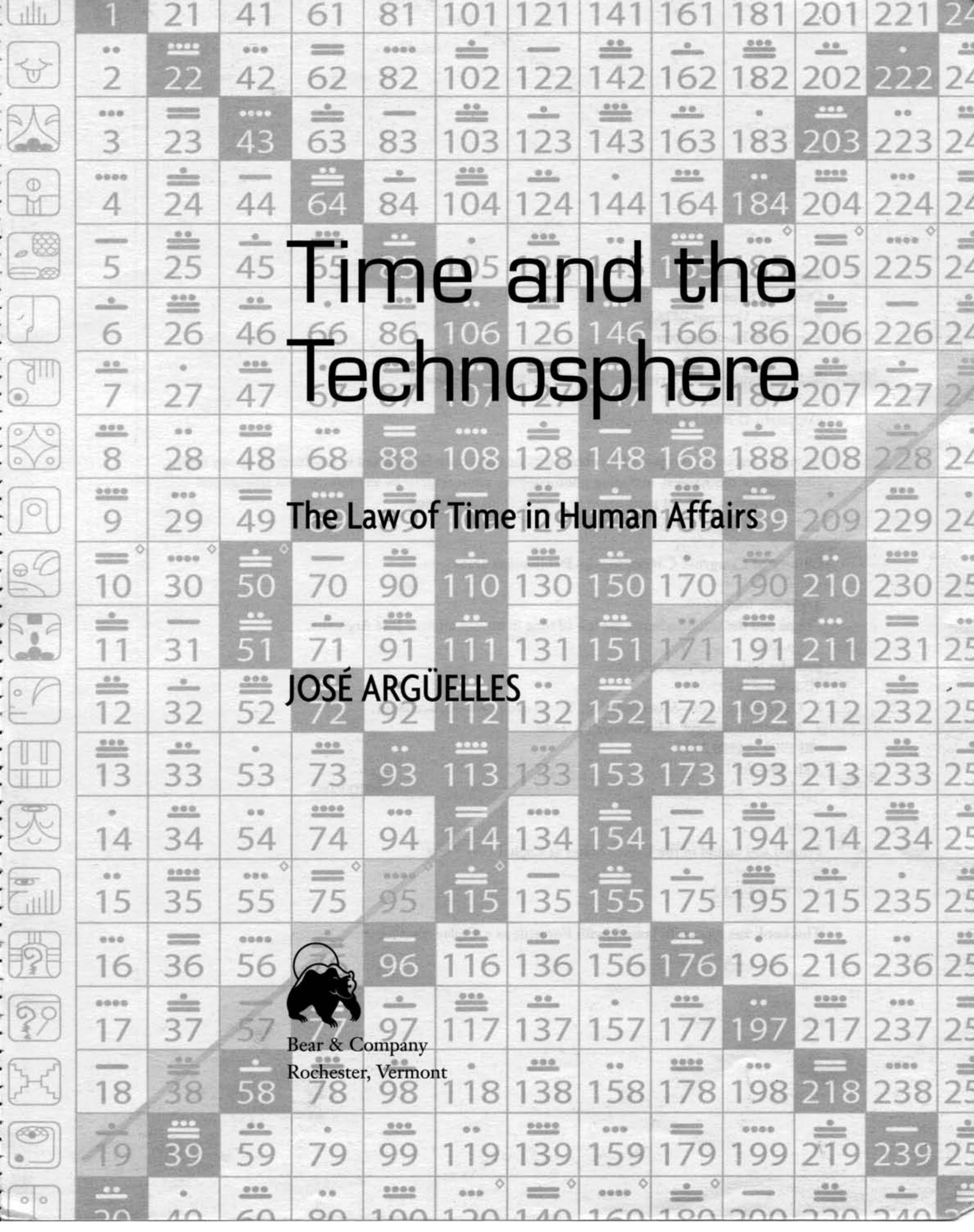

Time and the Technosphere

The Law of Time in Human Affairs

JOSÉ ARGÜELLES

Bear & Company
Rochester, Vermont

Bear & Company
One Park Street
Rochester, Vermont 05767
www.InnerTraditions.com

Bear & Company is a division of Inner Traditions International

Copyright @ 2002 by Jose ArgUelles

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Library of Congress Cataloging-in-Publication Data

ArgUelles, Jose, 1939-
Time and the technosphere : the law of time in human affairs / Jose ArgUelles.
p.cm.
Includes bibliographical references.
ISBN 1-879181-99-1
1. Calendar reform-Miscellanea. 2. Social evolution-Miscellanea.
3. Maya chronology-Miscellanea. I. Title.
BF1999 .A6398 2002
001.94-dc21
2002006082

Printed and bound in the United States at Capital City Press

10987654321

This book was typeset in Janson, with Eurostile as a display typeface

Time and the Technosphere

This book requires some attention from You. It is a complete, coherent, story for Humanity, on Earth, in the Universe. It is a radically different story than what is commonly disseminated by and through our global, dominating, culture. There is no need for You to *believe* this story, or any other. Trust yourself. trust your Self To say you must go into this with an open Mind, is a grand understatement. Start with trust. Simply allow the thoughts and concepts you read to "stew" for awhile. Make some headspace for them and have patience for what may grow there. It is a beautiful life...

NOTES ON THE ELECTRONIC PDF VERSION

- THERE ARE ERRORS! NOTHING MAJOR...ONLY SUPERFICIAL...
- NOTE THE COLOR PLATES (MENTIONED THORUGHOUT TEXT) ARE THE LAST EIGHT PAGES.
- THE PAGE NUMBERS ARE THE PAGE NUMBERS.
- USE ACROBAT'S BOOKMARKS FOR NAVIGATION!

ARGUELLES, JOSE - TIME & THE TECHNOSPHERE V0.2.PDF

The Law of Time in Human Affairs

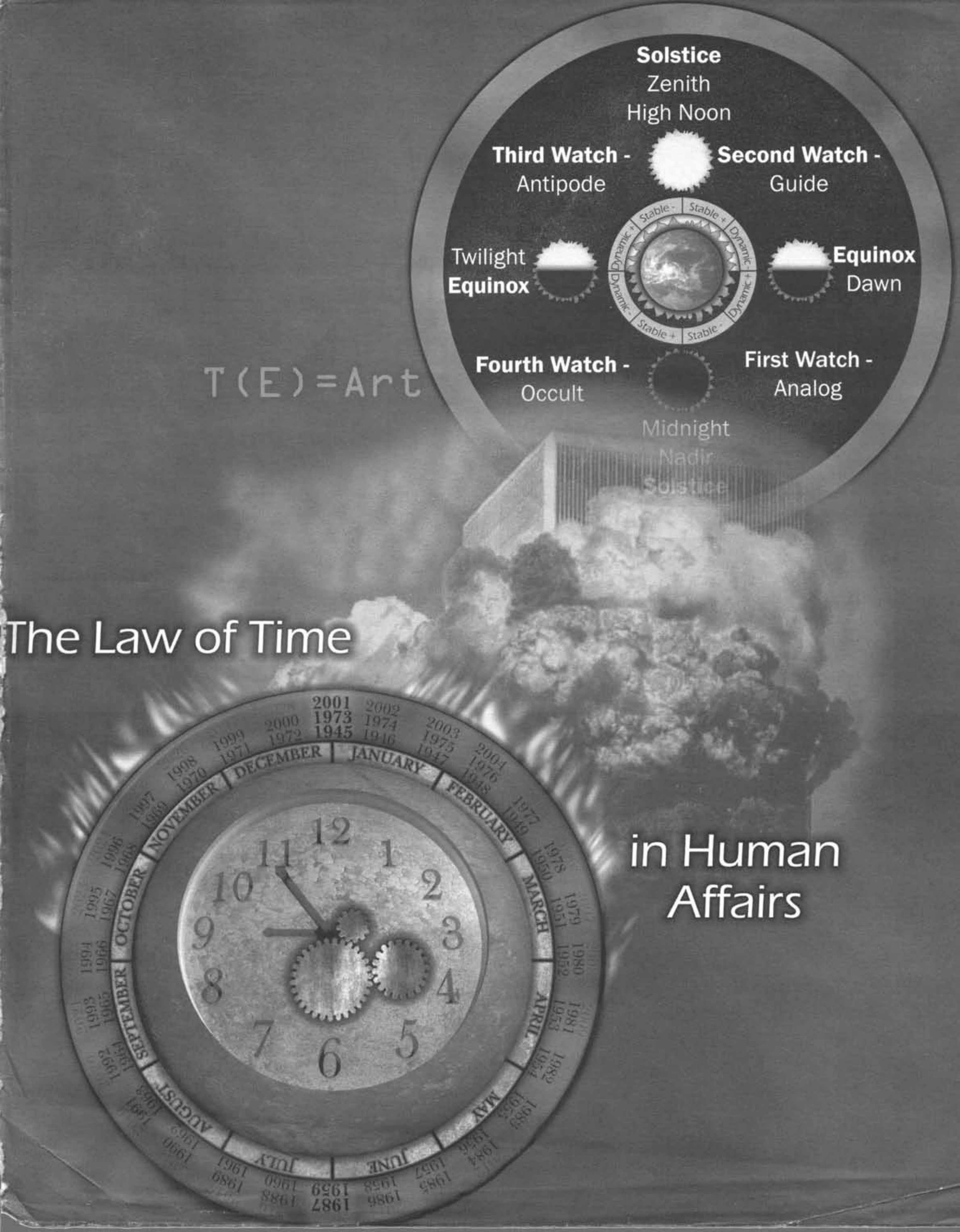
T(E)=Art
Solstice
Zenith
High Noon
Third Watch -
Antipode
Second Watch -
Guide
Twilight
Equinox
Equinox
Dawn
Stable -
Stable +
Dynamic +
Dynamic -
Fourth Watch -
Occult
First Watch -
Analog
Midnight
Nadir
Solstice
JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
2001 1973 1945
2002 1974 1946
2003 1975 1947
2004 1976 1948
1977 1949
1978 1950
1979 1951
1980 1952
1981 1953
1982 1954
1983 1955
1984 1956
1985 1957
1986 1958
1987 1959
1988 1960
1989 1961
1990 1962
1991 1963
1992 1964
1993 1965
1994 1966
1995 1967
1996 1968
1997 1969
1998 1970
1999 1971
2000 1972
12 1 2 3 4 5 6 7 8 9 10 11

This book is dedicated to Adam J. Lewis—at work on the eighty-eighth floor of the South Tower when his mother called him from Spain as she watched television reports of the first hit on the North Tower—and to all the others who were sacrificed on that day. May the truth prevail, that the greater mystery might be made known—forever.

Contents

Acknowledgments ix

Preface xi
A Brief Literary Autobiography

1 9-11, The Inevitable Event and the Nature of Time: Introductory Reflections 1

2 Preliminary Definitions: Biosphere-Technosphere-Noosphere as an Evolutionary Continuum 18

3 Time and Human Consciousness: The Law of Time-What It Is and Where It Came From 35

4 The Law of Time in Human Affairs: The Analysis of History 54

5 The Climax of History, the Fifty-six Years of Hiroshima-Artificial Time Runs Out 76

6 The Solution of the Law of Time: Get a New Calendar 99

7 Humans-Noospheric Chips: Consciousness and Reality Redefined 121

8 Making the Transition to the New Time: A How-to Catalog and Guide to the *Pax Cultura, Pax Biospherica* 145

9 The Advent of the Noosphere: A Prayer and a Vision 176

Epilogue 197
Networking the New Time

Appendices

I Summary Critique of the Gregorian Calendar with Which Every Proponent of the Thirteen Moon Calendar Should Be Thoroughly Familiar 200

II The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time: A Brief Chronotopology of Transformation 207

III Telektonon Prophecy of Pacal Votan 212

IV The Nineteen Code 229

Notes 240

Glossary 245

Bibliography 254

Acknowledgments

AS ANY AUTHOR knows, to write a book and see it through to completion, from initial inspiration to final conclusion, is never an easy task. Though we are all guided, if we so choose, by the Supreme Creator, still it is one's Earthly helpers that must be counted first, even before oneself-for without their aid, where would one be?

First and foremost of my Earthly helpers whom I wish to acknowledge is my wife and faithful companion of twenty-one years, Lloydine Bolon Ik, my constant helper and chief inspiration of all I do. The faith and trust she put in me at times when all we had was blind faith are a touching memorial to a soul who has always believed that love is the greatest truth.

Assisting us greatly during the time since we arrived in Oregon and have continued to dedicate our lives to the fulfillment of the prophecy of Pacal Votan and the Great Calendar Change is our devoted helper and assistant Stephanie South, the Red Queen, Red Electric Serpent, who conscientiously read and edited this text in its final form, helped in numerous details, and so kept me from ever bogging down.

Of equal trust and value is that one from the immemorial rainbow to whom I owe so much, Kin 19, David Rinck, "the stalwart," whose enhancements of my artwork are ingenious and awesome. All three of these people share one thing in common: absolute belief in and dedication to the Law of Time as a living reality that gives breath to all we do.

I must also thank the incomparable Spectral Skywalker, Mark Comings, whose mind has been of inestimable value to me through our many memorable background discussions on time in modern physics in relation to the Law of Time. In this regard, I would also like to mention Irina Eganova of the Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, for her rigorous approach to N. A. Kozyrev-may she smile well when she reads this book.

For their unconditional support I would also like to honor and thank Brian and Deborah Haight of the Foundation for the Law of Time, and all the many friends and beloved kin of the Planet Art Network everywhere, but especially to the Russian kin who so ably understood the Law of Time and opened up gold mines of knowledge and friendship in the Russian Academy of Sciences in Novosibirsk, as well as the great pristine valleys of Altai, where the shamans await the next installment of *The Mayan Factor.*

Spirituality is at the core of the biosphere-noosphere transition, and in this regard the beings I have to thank are without number, yet some especially stand out and deserve mention: my Tibetan teacher, the Sakyong Dorje Dradul Chogyam Trungpa Rinpoche, through whose cremation he entrusted me with all; my son, Josh, who opened the doors of the fourth dimension for me so that I would never be satisfied with less; and Mother Tynetta Muhammad, who showed me the way of the Holy Quran with the guidance of the mystic master FM, and who introduced me to the great work of Dr. Rashad Khalifa, whose breaking of the Quranic nineteen code led to his martyrdom, leaving me to fulfill what he revealed.

The creation of a book, the coming into manifestation of an artistic conception, is a complex process involving numerous other eyes and minds in order to bring the effort into its final published form. In this process two people have stood out deserving of mention. First of all, there is Michael Moynihan, White Galactic WIzard, who very patiently and carefully copyedited the book-always with an eye on maintaining and enhancing the original artistic vision-and who stood by the integrity of its vision through the entire editorial process. His work was complemented by the efforts of project editor Lee Awbrey, Blue Lunar Eagle, who took on an increasingly engaging role and personality as we brought this book to its final stages. I am especially appreciative for her inspiration to rework the Preface and Chapter One, which I gladly did, and which I feel speak more immediately to the reader of this book. Thank you both, and may the success of this book redound to your consummate care for the truth.

Finally I would like to honor and thank my children: Tara, the golden one, my grandson, SamuelJoshua Sheps, andJeff of Engine Company Number Nine in Phoenix; also Paul Mecklenburg, who has always held the faith in our work, and Heidi Mecklenburg, that her dream of the Unicorn Princess may become true once again.

▼

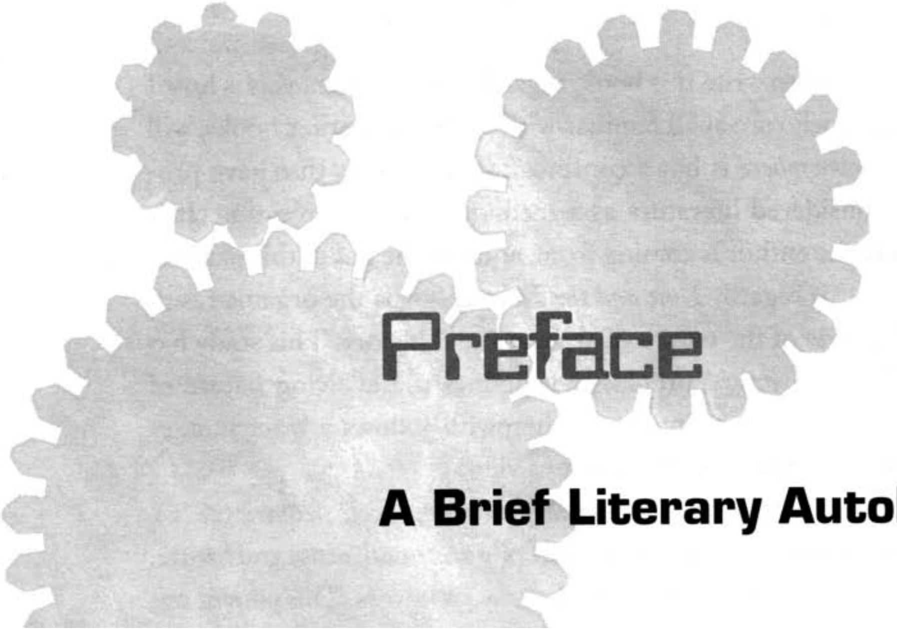

Preface

A Brief Literary Autobiography

Time and the Technosphere: The Law of Time in Human Affairs is essentially a prolonged set of meditations occasioned by the one most powerful and overwhelming event of our time, the 9-11, the Inevitable Event of September 11,2001. Subsequent events may overshadow this one in horror and magnitude, but it was this event in particular, the destruction of the World Trade Center towers, that sent a signal that something very profound had occurred to change the way we think of ourselves and our future.

In presenting these meditations and reflections in the form of a book, I am painfully aware of the limitations of this medium in the information age. Keeping in mind that the volume of information has increased five times in the past century-with much of that increase occurring only in the past decade-I am very reluctant to put anything down to be published in the form of a book. I was raised well on the writings of Marshall McLuhan and his definitions of media and how they affect our senses. It is altogether obvious that a book in your hand is not the same as reading a text on a computer screen. A book is actually something made to last. Of course, the computer has made it possible and easy for virtually anybody and everybody to write a book, and so we encounter one of the great hazards of the information age: the decline in discrimination and the criteria of excellence. Who can now tell what is worth reading or not? These reflections, too, are very much to the point of the grand theme of this text. If a book is meant to last, how long will this book last? It should last until the New Time has established a new order of reality on this now anguished Earth. Whether or not destruction and the total collapse of civilization will occur first, we do not really know. That is a matter left to the Divine Mind of God alone.

Nonetheless, it is with faith that we continue on this Earth at this time. A pressing voice told me to go ahead and to write this book, regardless-but then, this is how I have always done it. Those who are at all familiar with any of my earlier books, will find that *Time and the Technosphere* is but a continuation of themes that have provoked me since I first considered literature as a medium of communication. It is important to know where the author is coming from, and how he got to this particular point of reflection. In this regard, *Time and the Technosphere* is the organic result of a lifelong pursuit of a study of the nature of art, time, and history. This study has been woven together by a dialectical process that defines the evolving nature of consciousness. So, for those who are interested, herewith follows a brief literary autobiography, the better to understand my point of view.

My earliest book, *Charles Henry and the Formation of a PsychophysicalAesthetic (1972),* states the grand theme of a holistic science that unifies body and mind, sense and nature, the harmonic pursuit of which has been at the root of all my endeavors. This pursuit was amply demonstrated in my second book, *Mandala* (1972), which, as a root text of the counterculture, was an artistic, philosophical, and cosmic plea for a harmonic resolution to the process of history. It is a sign of the times that *Mandala* is no longer in print, but instead a book bearing the same title, but which does not even include the original Mandala in its bibliography, is in print with the same publisher. Amnesia is one of the side-effects resulting from information overflow with no discrimination. Unfortunately, cultural amnesia is always accompanied by an increase in trivialization, and thus we arrive at the mediocrity and sensationalism of the present moment.

My next book, *The Transformative Vision: Reflections on the Nature and History of Human Expression* (1975, 1991), took the aesthetic and unitive themes of the first two books to the next stage of my visionary analytical process, elaborating on the meaning of time, history, and cosmos. In rereading the later chapters of *The Transformative Vision,* I am struck by their prophetic accuracy, as well as their sense of impending tragedy. It is hard to believe that only 26 years have gone by since that book was first published and now we are plunged into the maelstrom of what was then thought to be the worst scenario possible. *The Transformative Vision* also defined human consciousness as the cerebral-neurological dialectic of psyche (aboriginal, primal intuitive art) and techne (civilizational, rational science), emphasizing that only in their absolute synthesis would there be a harmonic resolution to the problem of history. Seen in this light, the technosphere represents the extreme, one-sided triumph of techne, while terrorism is the expression of the absolute repression of psyche. The crucial question then arises: how will we ever attain a balance or harmony?

Investigation of the ongoing dialectic of consciousness was continued in my next effort, the philosophically synthesizing work entitled *The Feminine, Spacious as the*

Sky (1977), which was written and artistically produced with my previous partner and wife, Miriam Tarcov, who also co-authored *Mandala* with me. The feminine aspect represents a deepening reflection of the psychic, intuitive side of human nature. Expressed through a whole range of images, symbols, and philosophies that describe both the aboriginal root and the final flowering of human consciousness, it is this feminine aspect that has been smothered by the historical process, resulting in the technosphere. Why and how did this occur? These questions drove me to the next stages of my visionary investigations.

From the beginning of my pursuits I have been moved by my intuitive awareness of our intrinsic wholeness as humans, and the knowledge that this wholeness is a reflection of the wholeness of Earth, the solar system, and the cosmos itself. Although this perception of reality is nowhere within the domain of the shibboleths of Western scientific thought as it has developed since Descartes, I have also come to understand that Western scientific thought is highly provincial, racially biased, and-because it is not informed by a perception of wholeness-incapable of solving the actual problems of life. In fact, Western scientific thought has been one of the key factors leading the human race onward to its great act of self-destruction-a self-destruction employed through the instrument of technology.

In my next book *Earth Ascending: An Illustrated Treatise on the Law Governing Whole Systems* (1984, 1988, 1996), I was able to make a profoundly radical break with the restraints of the Western scientific and academic modes of thought. Combining the artistic and aesthetic qualities of *Mandala* and *The Feminine, Spacious as the Sky* with the philosophical and scientific concerns of *The Psychophysical Aesthetic* and *The Transformative Vision,* I define and describe the human situation as a function of the geology, or biogeology, of Earth understood as a whole system. Without knowing it, I had defined the human presence in Earth's geology in a way very similar to that of the Russian biospheric scientist V. I. Vernadsky in his analysis of the biosphere. Culture, civilization, and even technology are seen as features and functions of the larger evolving biogeological process of the whole system Earth, while time and consciousness are defined as functions of a larger planetary regulating system called the psi bank. In the analysis of *Earth Ascending,* the feminine psychic side of consciousness is referred to as the AC, Aboriginal Continuity, while the male, rational techne side is defined as the CA, Civilizational Advance. The historical process represents the increasing dominance of the CA to the final exclusion of the AC-the triumph of industrial man over the indigenous peoples.

The most critical moment in history was the atomic destruction of Hiroshima, August 6, 1945, when through its technology the impact of human thought permanently affected the geological makeup of the Earth. The events before Hiroshima are chronologically defined as P.A., pre-atomic, and those after as A.H., After Hiroshima. But more than defining the process of history within the context of Earth's cosmogeology, *Earth Ascending* also describes the post-historic future as a condition of harmonic synaesthesia-harmony of the senses-defined by a new scientific system called radiosonics. The key to this new system of scientific endeavor is synchronicity or synchronization of our mind and senses in time. The problem left open by *Earth Ascending* was the precise nature of time and the timing of the radical break in the historic continuum, now known as the Inevitable Event-the culmination and dramatic conclusion of a process begun at Hiroshima. After the Inevitable Event we are liberated into post-history and the possibility of the radiosonic science of the future.

In *The Mayan Factor: Path Beyond Technology* (1987), I was able to supply the basic ingredients for an understanding of time and the timing of the radiosonic future. Going beyond the parochial limitations of Western science and the self-inflated progressivist thinking of the Western mind in general (that states "anything that advances technology is an absolute advance over anything that was thought of or occurred in the past") I was able to finally bring to light the synthesis of many years of study of the time science of the ancient Maya. Here, in the mathematical calculations of the Maya with their unique vigesimal mathematics-counting by twenty and not by decimal ten-lay the basis of a vast and grand science of time. So grand is this science that its conceptions of time dwarf anything the Western imagination could conceive.

While the Western concept of time is linear and predicates an unceasing ascent into a progressively technological and mechanistic future, Mayan time science is predicated on the knowledge of time as the universal factor of synchronization. The Western conception of time is not in accord with the actual cyclical and synchronistic nature of time, and so the whole apparatus of Western civilization is doomed to reach its own breaking point: the Inevitable Event. Mayan time science posits a precise measure of the historical cycle, the most recent and climactic stage of Earth's evolution, as a time fractal of thirteen baktun cycles, each 144,000 days in length, spanning the JulianiGregorian years 3113 B.C.-A.D. 2012. According to Mayan time science, the efforts based on an erroneous time conceptualization are bound to fail prior to the cyclic end-point of 2012. Thus, the Inevitable Event is also a consequence of the failure to understand or to even acknowledge the actual nature of time due to a totally false and mechanistic timing sensibility. In this lies the cause of the apocalypse of Western thought and theology.

The problem left unresolved by *The Mayan Factor* was the precise methodology for changing tracks, for switching Western dominated global civilization away from the track of an erroneous, mechanistic time-based technology and toward the natural order of time-the path beyond technology. Here my efforts took me to a point beyond reason. Why or how? Let me explain. As an exposition of Mayan time science, in preparation for the 2012 ending of the thirteen baktun cycle of history, *The Mayan Factor* is predicated on the occurrence of a real time event, the Harmonic Convergence "global peace meditation" of August 16-17, 1987. The timing of the Harmonic Convergence was prophetic. It augured the fulfillment of the prophecy of Quetzalcoatl (947-999) known as the Thirteen Heavens and Nine Hells. The Harmonic Convergence completed the ninth and last Hell cycle, and ushered in the final twenty-five years of the thirteen baktun cycle. The Harmonic Convergence was a Mayan wake-up call answered by millions of earthlings.

Exactly forty-two years and ten days after Hiroshima, and fourteen years and twenty-six days before the Inevitable Event, the Harmonic Convergence was located by time to occur precisely when it did. Prophecy transcends reason because it is totally of time, which is of the fourth dimension; prophecy nevertheless affirms with unerring logic that which has been, and that which will be brought about, as being of the same karmic cause and effect stream.

It was this matter of prophecy in the aftermath of *The Mayan Factor* that brought my literary efforts and my own life journey to another stage of development. Prophecy comes with a mantle-if not, then you are classified as a madman. The prophecy of the Harmonic Convergence, as well as of the entire text of *The Mayan Factor*, bore the mantle of the prophetic stream of the Chilam Balam, the jaguar priests, the wizard knowers of the "night script." By correctly fulfilling the prophecy of the Thirteen Heaven and Nine Hell cycles, 843-1987, and by writing and putting together the mathematical codes of *The Mayan Factor*, and before it, of *Earth Ascending*, **I,** Jose Argiielles, had assumed the mantle of the prophetic tradition of the Chilam Balam. Mind you, this was nothing overt nor the least bit conscious, for little did I then know that my next life task was to decode the essence of the Chilam Balam prophecies and to define the time science of which these prophecies are a manifestation.!

The Harmonic Convergence affirmed the karmic repercussions of Hiroshima by bringing forward the ancient Mayan prophetic tradition of the Chilam Balam. By precisely locating this prophetic tradition in a contemporary moment -a media event, no less-the unlocking of its final prophecies was triggered for the purpose of completing the cycle of history: 2012 and beyond.

Like any major media event, the Harmonic Convergence was an enactment within the noospheric unconscious of humanity-the telepathic whole that resides in Earth's mental envelope. So it was that on June 23, 1987, the *Wall Street Journal* ran the front page story on the Harmonic Convergence. Yet due to the mental obscuration accompanying the final stage of global industrialization, the Harmonic Convergence was an event that the *WillI Street Journal* and the rest of American media conveniently forgot when it came to their review of the century in the year 2000. Today, just weeks after the Inevitable Event, the collapse of the World Trade Center towers, the words of that article are most prophetic: "Mr. Argiielles says the choice between a 'new age' and all-out destruction is ours, and we had better decide."2 Though the prophecy was discounted, true to form, the words of the messenger remain to haunt those who spurned the message.

The prophets and sages of the Maya, and of the New World in general, are a complement to those of the Old World, a further manifestation of the planetary dialectic of poles and hemispheres. If the judgment day apocalypse is the focus of the Old World prophecies, especially those ofIslam and Christianity, those of the New World similarly pinpoint an end time whose denouement is chronicled as the date 2012. It was the prophecies of the sage and culture hero, Quetzalcoatl (947-999), that brought about the Harmonic Convergence, while it was the wisdom of Pacal Votan (603-683) that informed the actual text and scientific time mappings of *The Mayan Factor*, and the image of his tomb adorns the cover of the original book. In a similar manner, it was the tradition of the Chilam Balam that created the prophetic vision of the synchronic order to be realized as the Law of Time.

I was hardly aware of the personally transformative factors at work within me after the Harmonic Convergence media swell began to subside, when two events occurred that augured a major change in my perceptions, activities, and literary output: the Wall Street stock market crash of October 19, 1987, and the death of my eighteen-year-old son,]osh, ten days later, on October 29,1987. In analyzing why the 1987 stockmarket crash happened, the lead story in *USA Today* about this event concluded by posing the question: Was it the Harmonic Convergence that somehow caused the stock market to crash? As I pondered why the answer to this question must be in the affirmative, as well as the implications of this analysis-remembering full well that the breaking story on the Harmonic Convergence first appeared in the *TVtlllStreet Journal-news* came of my son's sudden death in an automobile accident.

Synchronically, the stock market crash and my son's death were part of one event continuum. As tragic as the circumstances of his death were, my son had liberated me from any further compunction to lead a normal life, and to pursue, instead, this matter of Mayan time science and its calendars of harmony. In homage to my son, I completed what I first thought would be a popularization of *The Mayan Factor,* a book entitled *Suifers of the Zuvuya* (1988). Instead of being a popularization or even a sequel to *The Mayan Factor,* Surfers turned out to be a series of autobiographical tales of interdimensional travel-zuvuya surfing in the fourth dimension. This "flight of fantasy," was my literary farewell to the limitations of reason and a life of comfort. Inner destiny pushed me to the next stage in my pursuit of the unitive science-the discovery of the Law of Time.

This journey of discovery, the findings of which are the underpinnings of the present work *Time and the Technosphere,* took me and my wife and partner, Lloydine, to physical and mental places that were far from anything we had once deemed conventional. Because of this, I had ample opportunity to write and to create, and I did so, freely and without any restraints. I no longer sought any conventional means to communicate. I was now operating prophetically, on behalf of the noosphere-Earth's mental envelope.

There is a consistency to the stream of prophetic revelation, summarized as the codes of the Law of Time. This revelatory stream begins with *Earth Ascending* and *The Mayan Factor,* comes to flower with *The Dreamspell: Journey of Times hip Earth 2013* (with Lloydine Arguelles, 1991), and is completed by all aspects of the Telektonon Prophecy, including the Thirteen Moon/28-Day calendar (1993-2000). This prophetic stream is emphatically a galactic revelation of the same base of know 1-edge that produced the Mayan civilization of ancient Mesoamerica. The new revelatory stream is known as the Wizard's Count, or "reformulated Chilam Balam." It was this prophetic stream and its year-bearer count that established the validity of a fifty-two year cycle correlated to the Gregorian July 26.

The purpose of the Chilam Balam year-bearer count was to establish a basis for understanding that there exists a synchronic order of time, completely apart from what is usually referred to as the Long Count, the linear count of days of the thirteen baktuns of the cycle of history, 3113 B.C.-A.D. 2012. After a long break in the tradition of the Chilam Balam, the year-bearer count was prophetically reformulated to prepare for the Harmonic Convergence and the subsequent revelation of the Telektonon prophecy of Pacal Votan. As the new Wizard's Count, this count was intuitively recalibrated precisely so that July 26, 1987, would be White Galactic Wizard; July 26. According to *The Mayan Factor,* July 26, 1987: White Galactic Wizard, marks the first of the twenty-six years of the Harmonic Convergence. This date marks the conscious ascendance of the prophecy of galactic culture and the coming of the Earth Wizards-the race of galactic wizards on Earth.

In 1988, the year following the Harmonic Convergence, in attempting to create an everyday "Mayan calendar," I came upon the next stage of the discovery of the synchronic order. On the Gregorian calendar, 1988 is a leap year. I perceived that to accommodate February 29 would throw the year-bearer count off by one day, ruining the harmony of the system. Reviewing the Chilam Balam texts, it was clear that leap year was never considered, much less counted in the year-bearer count. Without an intercalary leap year date, the four year-bearers-Seed, Moon, Wizard, and Storm-coded by the numbers 1-13, remained in sequence, maintaining and assuring the harmony of the system, where every fifty-two years equals seventy-three 260-day spins, no two days are the same, and the entire 18,980 days of the cycle repeat every fifty-two years in an upward moving spiral of time. For the first time in my work, the issue of the Gregorian calendar as a system came into question.

Following this observation came the discovery of the mathematical codes governing the synchronization of the 260-day cycle with the 365-day cycle, the consequent discovery of the artificial 12:60 and the natural 13:20 timing frequencies, the recognition and resurrection of the Thirteen Moon/28-Day calendar, and the full exposition of the Dreamspell codes of fourth-dimensional time. I presented these discoveries in *Dreamspell: Journey of Timeship Earth 2013.* The Dreamspell, being the manifestation of the knowledge base of galactic culture, completely encodes the Wizard's Count. *Dreamspell: Journey of Timeship Earth 2013*, the presentation of the codes of fourth-dimensional time as a "tool kit" with an accompanying "script" and guidebook, is a break of such radical form from the prevailing paradigm that, as of today, it is still unfathomable to many people.

The galactic Dreamspell tool kit is a primary tool for the galactic culture that is just now beginning to unfold in the wake and shadow of the collapse of the Twin World Trade Center towers. This text prompted me to write two more books. The first, *Arcturus Probe: Tales and Reports of an Ongoing Investigation* (1992, 1996), is a deep reverie that picks up where *Suifers of the Zuvuya* left off and documents the interstellar, interplanetary psycho-cosmological roots of the Dreamspell of fourth-dimensional time. The second book was written as *A Treatise on Time Viewed from its Own Dimension,* but first published in English as the *Call ofPacal Votan: Time is the Fourth Dimension* (1992, 1996). In this highly scientific, mathematical treatise I was able to accommodate the Dreamspell codes of the Law of Time to an analysis of the biosphere and the biosphere-noosphere transition. This work also represents my first integration of the actual thought and principles of V. I. Vernadsky.

Through the efforts that constituted the discovery of the Law of Time, much of which was assisted by my partner and wife, Lloydine, I was finally able to begin to understand the methodology for accomplishing the paradigm shift-or psychic pole shift-of which humanity was so in need, at least from the perspective of the biosphere. By providing the mathematical basis for establishing the "biomass constant," Seventy-three 5 day subcycles correlated to the measure of the solar year, I had found a way of grounding the fourth-dimensional mental order of time, the synchronic order, in the actual third-dimensional biogeochemical cycles of the biosphere. This was like casting a harpoon of time from the noosphere and landing it in a precisely measured form in the annual turning of Earth's biospheric mantle. And this was accomplished precisely because of the intrinsic harmonic cycles and patterns of the DreamspelllWizard's Count.

For this reason, by 1993, it was incumbent upon me to test the hypothesis of the Law of Time that the human race was destroying its biosphere due to operating according to an error in time, and that only a shift in its experience of and perception of time would save it from such destruction. This shift in time is defined as the calendar change. The Law of Time presents such a radical critique of Western science, of its concepts of time, and of the civilization which it supports, that I could not tackle the problem from its center, but from its periphery, moving inward to the point of the Inevitable Event. Beginning in Latin America, then spreading to Japan, Western Europe, Russia, and finally North America, we established a movement to change the everyday perception of time, instilling profound meaning behind the effort to reform the calendar. In this way, we took on the responsibility of organizing the World Thirteen Moon Calendar Change Peace Movement.

Part and parcel of this, beginning in 1993, was the decoding of the prophecy of Pacal Votan, the inspiration of *The Mayan Factor.* Known as the Telektonon, Earth Spirit Speaking Tube, this prophecy is the full realization of the Law of Time and the synchronic order. As much science as it is prophecy, the Telektonon defines the Tower of Babel as the root of the error in time, and the Inevitable Event as the necessary result of not relinquishing false time. At the same time, the Telektonon provides codes and tools of analysis which are in complete accord with an understanding of the technosphere-noosphere as an evolutionary continuum in which the Inevitable Event is an act of moral consequence and natural law revealing the basis of a new cycle of harmonic order.

I would also be remiss at this point if! did not inform the reader that, by prophecy's prerogative, this text is also literally punctuated by a perspective that can only be described as "Quranic." Late in 1993, my immersion in the prophecy ofPacal Votan became sychronically linked to my immersion in a study of the Quran, or Holy Quran as it is sometimes called. I have taken seriously the claim that the Quran is the final revealed text (A.D. 610-632) for humanity. I have studied it following the dictum: "Quran, the whole Quran, and nothing but the Quran,"3 as the basis of genuine Islam-submission to the will of God. I have found it more than sufficiently true that the Quran is a scientific text, and contains nothing that contradicts science. More profoundly, my study has given rise to my own personal ethic: duty to God is the highest cause. It is the cultivation of this attitude, combined with many years of practice of Buddhist meditation, that have brought me to my present point of view, which is visionary and spiritual throughout-and, as a result, has endowed me with my own view on the current global dialectic.

That the new, and final, global dialectic is found in the tension between the forces of globalization (materialism) on the one hand, and that of the civilization of Islam (spirituality) on the other, is now a well-documented fact.4 Of course, generally speaking, in the West, we see it only from the materialist side of the dialectic. The perspective which informs me, and consequently this book, is definitely from the other side of the dialectic. Though in no way condoning terrorism- I am a non-violent pacifist by persuasion-the motives of the terrorists stem from a philosophy and worldview with which I am in sympathy. Yes, the Inevitable Event was also a profound theological moment, and what we are witnessing in the present World War is the battle between the religion of choice and the religion of submission. We cannot necessarily assume that the religion of choice will prevail. In fact, from the perspective of the biosphere-noosphere transition, it is the religion of submission that is most in accord with the spiritually-beckoning future.

As a preface to this book, I have found it mandatory to write this brief literary autobiography, often couched in a terminology that jumps ahead of the story, solely to demonstrate that the actualization of the noosphere corresponds to the stages of attainment and transcendence of one human's life journey in pursuit of the truth. Only in this way could I cultivate the noospheric perspective, voice, and point of view that was necessary for the writing of this text. In this perspective, the noosphere is a condition of non-anthropocentric consciousness. My journey, thus documented, is merely one way the noosphere has found to become conscious, asserting the Law of Time's purposive axiom: to make conscious what had been unconscious. The voice and method of the noospheric perspective are inseparable from the topic of this book, *Time and the Technosphere.* By analyzing the timing and structure of the technosphere, the noospheric point of view becomes conscious. WIthout understanding the nature and relation of time and the technosphere, there would be no truly sane way to comprehend the meaning of the Inevitable Event.

Crystal Moon Kali 11,
Kin 2: White Lunar WInd
Yellow Solar Seed year
Cascadia, North America

1

9-11, The Inevitable Event, and the Nature of Time

Introductory Reflections

Woe! Woe, O great city,
O Babylon, city of power!
In one hour your doom has come!
Woe! Woe O great city, dressed in fine linen
purple, and scarlet...
In one hour such great wealth has been brought to
ruin ...
"Woe! Woe, O great city,
Where all who had ships on the sea became rich
through her wealth,
In one hour she has been brought to ruin! ...
With such violence the great city of Babylon will be
thrown down!"
ST.JOHN OF PATMOS, BOOK OF REVELATIONS,
CHAPTER 18, *THE FALL OF BABYLON*

9-11. The Twin Towers of the World Trade Center gone in a veritable twinkle of the eye. Tragedy. Horror. Terrorism. The whole world brought to a halt. The aftermath? Patriotism, war, and an all-pervading uneasy feeling. What happened? What really happened? Was this the beginning of the Apocalypse?

Amidst the distraction of chasing after other people with smart bombs in far-away places like Afghanistan, and making everyone at the airports feel like potential terrorists, there are no answers forthcoming for what really happened. There is a good reason for that, too: The destruction of the Twin Towers of the World Trade Center and the attack on the Pentagon marked a point beyond the comprehension of the current belief system in every way. To say that the event was mind-blowing is an understatement. But if it was mind-blowing-which it truly was-that means that a whole belief system was knocked off of its foundations, and only a new belief system can supply the answer. A higher, more all-encompassing belief system-a new paradigm-is what is needed if we are to really understand what happened on 9-11.

Belief systems, believe it or not, are functions of time. What you believe defines the time you are living in. Any belief system is held in place by the calendar and sense of time in which it is encoded. 9-11 means September II-it also means 911 emergency, planetary emergency in this case. "9-11 we will never forget" was a common slogan appearing on bumper stickers and store windows after the event. The words or concepts "Tuesday, September 11," "9-11" and even "911" are all functions of the same belief system. So is "2001," the first year of the highly touted "third millennium." But the third millennium of what? The third millennium since the birth of Christ. This means that the belief system whose mind was blown was the Christian millennial belief system encoded in the Christian calendar, the Gregorian as it is properly called. Does this mean that the foundations of 2000 years of belief were blown away, shattered with the terrifying collapse of those Twin Towers? Easily that many years, maybe as many as 5,000 years, or perhaps even the entire history of civilization as we know it came unhinged that fateful day, the day of the 9-11, the day of the *Inevitable Event*, to borrow a phrase from the Holy Quran, the day the Apocalypse announced itself, showed up on the morning television, quite without warning. (See plate 1, The Invevitable Event and the End of History.)

In meditating upon the meaning of the Inevitable Event, it is natural to ask: What made this event so inevitable? It is not just from the perspective of American Foreign Policy, or that of its economic free market monetary politics, that this event was inevitable. More profoundly, this event was so inevitable because it seemed to smack of some form of retribution for the breaking of fundamental laws of nature, or even of Divine Law. And isn't the retribution for transgressing Divine Law a form of apocalypse, a standing naked, a revelation, an evocation of the fall of the Tower of Babel, no less?

Yes, that is why it was so shocking-it was the Inevitable Event because the apocalypse is as inevitable as it is without warning. But apocalypse- a word no one wanted to use to describe the event, because to use it would be to admit that it was all over-is only one side of the coin. There is another side to the coin of the Inevitable Event-and that is what this book is about. There is something beyond the Apocalypse, there is something larger going on, and that also scares people. Things are out of their control. Isn't this why, instead of looking at it, or contemplating the event, a policy of war was immediately instated as a principle of revenge-for what and against whom? What is going on is time-time is always going on, or time is always coming together. Or time is always changing. The apocalypse, too, is about time-the end of time-which is even more frightening to most people. But the end of whose time? The new paradigm we need to explain 9-11 should be a paradigm that is all about time.

I who speak know, because I have spent most of my life studying time, and the effects of human time which you call history. For many years now I have been living an experiment in time, but a time most of you do not know. So I have some answers that will put 9-11 in perspective for you. I am the stranger from the other side of the wall of mechanized time. I have come back to make you familiar with the time your clock and your calendar shield you from. Because I know about the other time, the larger time, the cosmic time that governs the cycles of all that exists, I can tell you that 9-11 was the Inevitable Event. It had to happen precisely when it did, not because nineteen terrorists planned it that way, but because the terrorists themselves were unconsciously programmed by the timing of Earth itself. Earth's timing process represents evolutionary forces as well as effects irreversibly set in motion by the human's own unconscious programs, programs governed by the human misperception of time and of its role in the cosmic order.

Yes, Earth has its program of which our human history and all of its endeavors are only a part. This larger meta-program of Earth is called the biosphere, the scientific designation given to what is commonly and inadequately referred to as "the environment." As such, the biosphere is a whole system composite of the sum of life and its organic and inorganic support systems spread out over the surface of Earth, hence bio = "life" + sphere = "having the form of a globe." In a word the biosphere is a vast but fragile process which has been evolving itself for some two billion years on this little planet as it spins around its local star, the sun. But we humans, who only in 1969 first saw ourselves as a planetary being-the whole Earth beamed back at us via television by a rocket hurtling toward the moon-we humans have been generally so self-involved in our own interpretation of things that we have missed seeing the larger picture of which we are a part. This larger picture is the entirety of the biosphere which encompasses the evolution of life on Earth as a whole system process. The amazing thing is that most of us have never even heard of the biosphere much less even know that we are a part of this whole system governed by laws and principles which we, in our self-made arrogance and ignorance, have but little understood.

From my experiment in time, seeing and studying the human as part of the larger self-evolving fabric of the biosphere, I came to the conclusion that the human is living in a time apart from the rest of the biosphere-an artificial time whose climax and termination is inevitable, for nothing artificial can withstand the force of truth. If the human is living in artificial time, the clock is an artifact whose system of measure has nothing to do with natural cycles but is a totally abstract standard, then there must be something called natural time. I will go even farther and state that not only is there natural time, but that there is a law governing natural time, and that is the Law of Time. Just as Newton only discovered gravity some 300 years ago, though gravity has always existed, so the Law of Time has always been in operation, even though it was just recently discovered. It is the Law of Time that governs the entirety of the biosphere and everything else in the universe according to principles which are only dimly known to us, principles like synchronicity and telepathic instantaneity. Yes, to understand what really happened on 9-11 you have to understand the Law of Time and how it governs the biosphere-of which our rampant human civilization is inescapably a part.

The Law of Time is formulated very simply, and in some people's way of thinking, rather unscientifically as T(E) = Art, "energy factored by time equals phenomena in the material world represent some state of energy, and every state of energy is governed by time, the resultant product of which is always something beautiful or elegant. Have you ever seen an ugly sunset? A hideous flower? Even if you examine a scorpion with some objectivity you will be amazed at the flawless and elegant manner in which its parts are organized. Yes, all of nature is organized by time to produce in you the sensation of beauty. And time itself, well, believe it or not, time is a frequency, and a frequency is not measurable by a clock. The Law of Time states that time is the universal frequency of synchronization. It is the nature of time to synchronize and to maintain all things in a condition of synchronization. Synchronicity, then, is the experience of real time. When we say that time is a frequency, we can be more precise and say that time is a universal constant expressible by the mathematical ratio 13:20. That is, the 13:20 ratio is the frequency of synchronization.

Of course, most people have never even heard of the Law of Time much less comprehend what the 13:20 ratio might mean. But then, that again gets to the root of the problem, or rather, the dilemma: If humans are living in their own time apart from the rest of the biosphere which is governed by the natural timing frequency, how could they ever know about the Law of Time? The humans are all a little like Mrs. Malaprop in the play by Moliere who learns to her surprise that she has been using grammar all of her life-though it is much more serious than that. So we humans are with the Law of Time-it has been operating throughout history without our awareness of it. Because of this, its way of making itself known to us was the infamous 9-11. Yes, in the analysis of the Law of Time, the 9-11 marked the end of artificial time.

Living in ignorance of the Law of Time and of the true nature of time, the human species constructed its own concept of time which is based on the clock with its 60-minute hour and the Gregorian calendar with its maddening array of unevenly numbered months. Hence, the paradigm of the artificial timing frequency is defined by the ratio 12:60 (twelve-month calendar, 60-minute hour) -in contrast to the 13:20 ratio of natural time. If the artificial mechanistic and irregularly measured time sets the human race apart from true or natural time, does this not also establish the fact that the human race is living an error in time? Only by living such an unconsciously assumed error could the human race arrive at such an apocalyptic moment as the Inevitable Event-inevitable because any deviation from the truth has an inevitable moment when the truth rebounds in some dramatic or even apocalyptic way. Is it possible that the entire construct of modern civilization, so devastating to the biosphere, is a function of this error in time? Is modern civilization like a time warp, a bubble of artificial time that suddenly got popped on the 9-11?

Back in the 1990s, now seemingly so long ago, I was traveling around the planet on behalf of the Law of Time and the biosphere, and making my observations regarding the effects of artificial time from a planetary whole systems point of view. I saw that even though the human species had evolved - or devolved, as the case may be-into a cultural hybrid which I call the planetary human, the object of the much touted "globalization," I saw that the consciousness of this cultural hybrid was anything but planetary. This schism between the propaganda of globalization and the actual sectarian and fragmented state of consciousness of the hybrid human was further exacerbated, I observed, by the potently unconscious effects ofliving an error in time. Ironically, if the Inevitable Event left many of us in the dark, it has brought the Law of Time into the light. To elucidate let me quote here from a text I wrote in 1996, and which I think will give us a point to expand upon in order to bring the Inevitable Event into sharper focus, while allowing me to define some of my terms:

"Through an error in time, the human species transforms its artificial construct, civilization, into a global technosphere: the sum product of industrial mechanization of its biological functions. Because of human adaptation to the irregular mechanized 12:60

timing frequency, the technosphere runs counter to the laws of the biosphere, creating a magnetic instability between the primal crystalline and vital organic processes."2

The current explosion of the world crisis, stemming from the destruction of The World Trade Center Twin Towers, demonstrates the hypothesis of the Law of Time that states: operating on artificial and mechanistic timing standards will cause the human to deviate from natural time to the point of its self-destruction, the end result of the magnetic instability referred to in this reference. But here we have some further definitions to make. Let us start with a definition of the technosphere, since that word is featured so prominently in the title of this book, *Time and the Technosphere: The Law of Time in Human Affairs.*

While the term biosphere defines the entirety of life as a single coherent unity, inclusive of its inorganic or semi-inert support systems such as the atmosphere and the hydrosphere, technosphere defines an artificial sheathe or membrane held together by industrial technology as a whole system. This technological mantle is the sum of the processes devised by humans resulting from the application of certain scientific and economic principles, especially during the last half century. The effects of the technosphere are to supplant the organic processes of the biosphere with totally industrialized ones, resulting in a severe imbalance in the biosphere. This imbalance, the "magnetic instability between the primal crystalline and vital organic processes" refers to the disruption of the delicate interaction between the inert (crystalline) and biological processes and structures which constitute the vibrant mechanisms and interactive cycles of the biosphere. Phenomena such as global warming and over-population are symptomatic of the technospheric disruption of the natural order of the biosphere.

The term technosphere, however, is actually just the middle term of a larger process defined as the biosphere-noosphere transition. According to the principle figure behind the elaboration of the laws of the biosphere, V. I. Vemadsky (1863-1945), as an evolving structure, the biosphere is inevitably being transformed into a new geological and evolutionary condition, the noosphere-Earth's mental envelope. In Vemadsky's analysis, the entirety of the biospheric process can be described as a single biogeochemical unity. According to Vemadsky, due to man's transformation of nature since the industrial revolution, this biogeochemical unity has entered a state of combustion-an accelerated process of transformation. It is this rapidly accelerating geochemical combustion-the net effect of industrialization and its waste -that accompanies an evolutionary shift, the transformation of the biosphere into the noosphere. But between the biosphere and the noosphere there is the technosphere, the intervening medium of transformation.

So when we speak of the biosphere and the noosphere we must take into account the technosphere. In actuality we must speak of the biosphere-technosphere-noosphere transition. In this context, the noosphere is the condition of the human mind purified of the error of time. And if the noosphere is the Earth's mental envelope, it could only be so because of a unified state of mind, the transformation of the hybrid planetary human into a genuinely spiritualized planetary being, the mind of the Earth no less-something that seems phenomenally utopian from the perspective of today's evening news. How could this come about, this noosphere of a telepathically unified human consciousness? Again, the Law of Time supplies genuine answers. To begin with, it says, change your calendar, get back to living in harmony with the natural cycles of the rest of the biosphere-and the universe. If you don't, well, the destruction of the Twin Towers was just the first stage of the Apocalypse. But we still have a choice. The other side of the coin of the Apocalypse shown to humanity on prime time television on the 9-11 is entry into the noosphere. But we must first understand what is really going on, and then act intelligently.

The mental climate in America is particularly difficult and strained now. In order to maintain a positive and constructive attitude of mind during this terrible time, I decided to write this book. As you can see, this book takes for its point of departure an analysis of the destruction of the World Trade Center (the actual hub of the technosphere) as the Inevitable Event defining the dynamic of the biosphere in its process of transforming into the noosphere. The analysis of the event in this light will give ample opportunity for presenting the key ideas of V. I. Vernadsky, in particular the concept of the biosphere which he so single-handedly amplified and defined, as well as the cosmic ideas of other Russian thinkers which are little known in the West. I have already written about and given much thought to this matter in general, summarized in part in the First Planetary Congress of Biospheric Rights, held in Brasilia, Brazil in 1996, but nowhere have I written for a wider public audience concerning the nature of the biosphere-noosphere transition.

I feel it is especially important at this time to present an analysis of this event in a scientific context that is beyond nationalism and ideology, so that human beings can begin to understand that they are a function of the biosphere, but not necessarily its controlling mechanism. Of course, the Law of Time is the key factor involved in this analysis, and understanding the biosphere in the context of the Law of Time is absolutely critical for establishing the basis of an entirely new worldview and hierarchy of values-the much awaited new paradigm.

Like my thoughts on the biosphere-noosphere transition, my investigations of the Law of Time have also not as yet been presented to the general public. In fact, you will find scarcely anyone who has even heard of the "Law of Time." This is because my researches and investigations over the past decade which resulted in the discovery of the Law of Time were not conducted in any formal academic way, nor through any corporate or government sponsored grant or officially recognized research institution. The discovery of the Law of Time has been, nonetheless, a rigorous, life-consuming pursuit. The laboratory for its study was to be found in a phenomenological and anthropological investigation of the effects of mechanized time on various population groups absorbed in what is called "modern civilization."

The premise of this investigation is simple: it is an undeniable fact that the time of human civilization, governed by the clock and the Gregorian calendar, is not the same as the time of the rest of the biosphere. Artificial time is not the same as natural time. Or rather, artificial time establishes a frequency which governs the human species totally apart from the timing frequency of the rest of life. What are the effects of artificial time and how do they define the nature of modern civilization? The answers to these questions open a perspective that has been scarcely considered. Yet this perspective is so clearly an essential factor in the description and definition of the human condition, that, in the wake of the tragic events and their ominous aftermath, it now begs to be presented in a proper manner. Hence this book. Suffice it to say that the Law of Time in every respect answers the question posed by my previous books and summarized in the single query: how will modern man escape the fatal consequences of the one-sided mechanization of his biological operations and the imminent destruction of his life-support system?

Defining a planetary whole systems approach to the analysis of time in human affairs, the first postulate of my book *A Treatise on Time* (1996) concerning the principles and nature of time as the fourth dimension is worth quoting here, since it introduces the fundamental problem informing the investigation of Time and the Technosphere:

> Just as air is the atmosphere of the body, so time is the atmosphere of the mind. If the time in which we live consists of uneven months and days regulated by mechanized minutes and hours, that is what becomes of our mind: a mechanized irregularity. Since everything follows from mind, it is no wonder that the atmosphere in which we live daily becomes more polluted, and the greatest complaint is: "I just don't have enough time!" Who owns your time, owns your mind. Own your own time and you will know your own mind.3

It must be understood that the events of 9-11 were so historically unprecedented, so mind-bogglingly dramatic, that they require a large lens through which to fully interpret their meaning. The Law of Time is such a lens. Quite simply, the destruction of the Twin Towers defines the limit of artificial time. If this is so, what can be expected following this world-shattering event? Because the Law of Time began with the premise of defining the difference between artificial time and the time of nature, the larger context of the biosphere was resorted to almost immediately. Being the study of an ongoing and evolving phenomenon, the investigation of the biosphere by its principle researcher, V. 1. Vernadsky, pointed to another imminent evolutionary stage: the noosphere, the "mental envelope of the planet." However, an intermediate stage was also perceived by researchers who followed Vernadsky, and it came to be called the technosphere, "the technological envelope of the planet." The end or limits of the technosphere define the birth or beginning of the noosphere.

With the events of 9-11, this is no longer just a theoretical concept but a vivid reality. As a species we are now undergoing this dramatic evolutionary moment, otherwise known as the biosphere-noosphere transition. It is for this reason that this book also presents novel definitions of time and pragmatic conclusions that can be employed by anyone if they so choose. If the root of the technosphere is artificial time and its proliferating mechanized constructs, then in the collapse of the technosphere, an event ordained by the laws of nature, there must be an orderly path or means for establishing the noosphere, also ordained by natural law. The Law of Time defines this orderly path as a conscious shift in the human timing frequency, away from artificial time and into universal natural time. This can only be accomplished by a universal reform of time, known as the calendar change. If there were no such path, the conclusion to the collapse of the technosphere would be an unbearable barbarism.

By its nature, the topic of the book is cross-disciplinary and provocative-it is visionary history. I have been described as a "spirituality historian" and "cosmic harmony researcher."4 I believe this description also characterizes *Time and the Technosphere,* which does present a solution, a new world paradigm, if you will. A new paradigm has been awaiting humanity ever since the notion of paradigms was first raised by Thomas Kuhn in 1964, in his book *The Nature of Scientific Revolutions,* or even earlier by Vernadsky and Pierre Teilhard de Chardin in their invocation of the noosphere. To be truly new, such a paradigm could hardly be expected to arise from traditional fields of thought or methods of research, be they quantum physics or microbiology. In fact, as I have often said, most people wouldn't recognize the new paradigm if it ran into them on the street. This is because people's consciousness and perceptions are basically very set and unexamined-as unexamined as the effects of the watch on their wrist or the calendar on their wall. In fact, one of the consequences of the technosphere is the limiting nature of consciousness which remains unchanged despite the fact that technology is changing all of the time. But with the occurrence of the Inevitable Event, history has been punctured. Perhaps now the new paradigm can be perceived: the paradigm of natural time as the universal factor of synchronization.

Although prompted by the stunning terrorist attack upon and consequent apocalyptic collapse of the World Trade Center Twin towers, these reflections are actually a synthesis of thoughts and explorations in the domain of time that I now find necessary to communicate in a coherent and organic form. The grand theme of this study is to understand how and why the collapse of the twin World Trade Center Towers was a biospheric event. Even so, the study of the Inevitable Event within the context of the evolution of the biosphere requires and leads us to still grander domains of thought-the historical, the theological, and the cosmological. The reason that these reflections and meditations spiral into ever grander domains of thought is due to the principle of analysis and the guiding factor employed in this study. This is the Law of Time, which is the fruit of a life-long research and pursuit of knowledge. It is the Law of Time that brings the biosphere into timely focus.

The biosphere is scarcely known even to most educated people. As a biogeochemical whole system in which the human is but one of the functioning components, the biosphere is an intricate unity with its own governing laws and principles. Without becoming conscious of how the human organism participates in the laws and principles of the biosphere, the human will continue to remain ignorant of the biosphere's existence and therefore will sooner than later reach a termination of its own evolutionary possibility. A demonstration of how the laws and principles of the biosphere, crystallized into a temporary and intermediate sheath called the technosphere, reached a climactic evolutionary point in the collapse of the World Trade Center towers will provide the means of educating in general concerning the biosphere. It will also allow us to elaborate on the meaning and nature of the technosphere, and the imminent transformation of the biosphere into the noosphere, the Earth's "mental envelope."

A fundamental purpose of the Law of Time is to expand upon the meaning and significance of the Inevitable Event, as well as upon the biosphere, technosphere, and noosphere in relation to the evolution of life and consciousness in the universe. To speak of an event being "inevitable" is to bring about a reflection on the nature of time, and even of the laws of cause-and-effect and karma. It is time that governs both cause-and-effect as well as the actions of karma. And it is time that dominates inevitability, because any such inevitable event is located in a very precise moment in time. But what actually is time? How does time "locate" events? And, once again, what is the Law of Time?

As the criterion of natural time in which synchronicity is the norm, the Law of Time is the key that unlocks the reason why the Inevitable Event occurred when it did. By precisely locating it within the evolutionary continuum of the biosphere and in relation to its point of mutational shift, the Inevitable Event is placed into a greater context of cosmic comprehension by the Law of Time. Through the lens of the Law of Time we will be able to understand the relationship between the biosphere, the intermediate state of the technosphere, and the imminent stage of the noosphere. Defining the Law of Time as the factor of universal synchronization, the law by which all events in the universe are maintained in synchronic relation to each other, we approach the domain of a hitherto undefined realm, that of the synchronic order.

As the unifying fourth-dimensional order of time in the universe, the synchronic order is measured and gauged by the 13:20 matrix. The mathematical coding of the Law of Time is entirely locked into this 13:20 matrix, otherwise known as the Harmonic Module or T zolkin, the basis of the sacred count of the ancient Maya. When we examine the Harmonic Module, a 13 x 20 matrix, we may well ask what does that have to do with time? It is obviously as different from a clock, as a tree is from a flagpole! When we examine the matrix carefully we see that there is a repetition of thirteen notations, the numbers 1-13, that occurs 20 times yielding 260 (13 x 20) units.

Holding the 13:20 matrix together is a pattern which I identified in my book *Earth Ascending* (1984) as the binary triplet configuration, and in the *Mayan Factor* (1987) as the Loom of Maya. This pattern consists of 52 units, 26 on either side of the central or seventh vertical column. **If** you contemplate this 52-unit pattern you will see it is characterized by binary radial symmetry. This quality characterizes the entire matrix and defines the quality of fourth-dimensional time as radial and instantaneous distinct from the current third-dimensional paradigm which defines time as being linear and sequential. The radial quality of fourth-dimensional time is what accounts for its being the factor of universal synchronization. This is very different from the sensation of the second hand as it sweeps around the clock, indicating the inexorability, not of time but of the mechanization of time! The 13:20 matrix becomes potently useful as the means of establishing the synchronic order when it is coordinated with a genuinely harmonic timing standard, the Thirteen Moon/28-Day calendar.

"Time is the fourth dimension," is the famous adage attributed to Einstein. This statement notwithstanding, the modern notion of time is bogged down in the minutiae

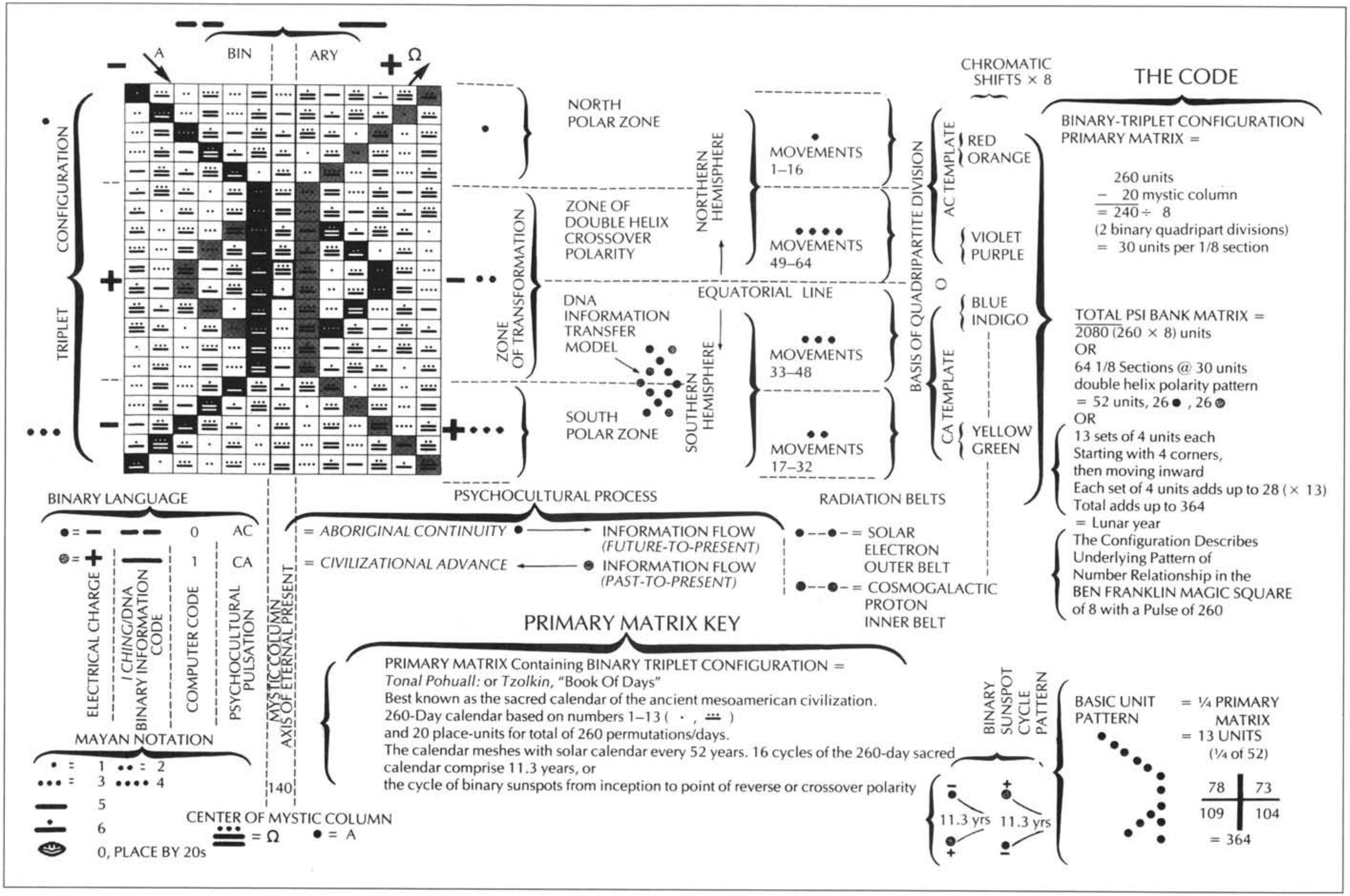
A
BIN
ARY
Ω
CONFIGURATION
TRIPLET
NORTH POLAR ZONE
ZONE OF TRANSFORMATION
ZONE OF DOUBLE HELIX CROSSOVER POLARITY
DNA INFORMATION TRANSFER MODEL
SOUTH POLAR ZONE
NORTHERN HEMISPHERE
EQUATORIAL LINE
SOUTHERN HEMISPHERE
MOVEMENTS 1–16
MOVEMENTS 49–64
MOVEMENTS 33–48
MOVEMENTS 17–32
BASIS OF QUADRIPARTITE DIVISION
AC TEMPLATE
CA TEMPLATE
O
CHROMATIC SHIFTS × 8
RED
ORANGE
VIOLET
PURPLE
BLUE
INDIGO
YELLOW
GREEN
BINARY LANGUAGE
0
1
AC
CA
ELECTRICAL CHARGE
I CHING/DNA BINARY INFORMATION CODE
COMPUTER CODE
PSYCHOCULTURAL PULSATION
MYSTIC COLUMN
AXIS OF ETERNAL PRESENT
140
MAYAN NOTATION
1
2
3
4
5
6
0, PLACE BY 20s
CENTER OF MYSTIC COLUMN
= Ω
● = A
PSYCHOCULTURAL PROCESS
= ABORIGINAL CONTINUITY
INFORMATION FLOW (FUTURE-TO-PRESENT)
= CIVILIZATIONAL ADVANCE
INFORMATION FLOW (PAST-TO-PRESENT)
RADIATION BELTS
= SOLAR ELECTRON OUTER BELT
= COSMOGALACTIC PROTON INNER BELT
BINARY SUNSPOT CYCLE PATTERN
11.3 yrs
11.3 yrs
BASIC UNIT PATTERN
= ¼ PRIMARY MATRIX
= 13 UNITS
(¼ of 52)
78
73
109
104
= 364

THE CODE

BINARY-TRIPLET CONFIGURATION
PRIMARY MATRIX =

260 units
− 20 mystic column
= 240 ÷ 8
(2 binary quadripart divisions)
= 30 units per 1/8 section

TOTAL PSI BANK MATRIX =
2080 (260 × 8) units
OR
64 1/8 Sections @ 30 units
double helix polarity pattern
= 52 units, 26 ● , 26 ●
OR
13 sets of 4 units each
Starting with 4 corners,
then moving inward
Each set of 4 units adds up to 28 (× 13)
Total adds up to 364
= Lunar year
The Configuration Describes
Underlying Pattern of
Number Relationship in the
BEN FRANKLIN MAGIC SQUARE
of 8 with a Pulse of 260

PRIMARY MATRIX KEY

PRIMARY MATRIX Containing BINARY TRIPLET CONFIGURATION =
Tonal Pohuall: or *Tzolkin,* "Book Of Days"
Best known as the sacred calendar of the ancient mesoamerican civilization.
260-Day calendar based on numbers 1–13 (· , ≝)
and 20 place-units for total of 260 permutations/days.
The calendar meshes with solar calendar every 52 years. 16 cycles of the 260-day sacred calendar comprise 11.3 years, or
the cycle of binary sunspots from inception to point of reverse or crossover polarity

13:20 Matrix-Harmonic Module with Binary Triplet Configuration

of quantum mechanics and cessium clocks, in a hair-splitting myopia that cannot see the forest for the trees. The fact is that only time governs the whole order of the universe in a manner that transcends all spatial limitations, even those of the relativistic MinkowskilEinstein four-dimensional universe, which is confined by the speed of light. As the universal factor of synchronization, time is instantaneous and transcends light. This novel perception of time defines a new reality, the synchronic, of which our experiences of synchronicity are but a foretaste. The synchronic order described by the Law of Time poses the order of an entirely new science of time.

To be realized and then applied, the science of time is dependent on two factors: a sudden and radical disruption of the historical continuum; and a genuine globalization or planetarization of consciousness. The break in historical continuity is necessary to jolt the human consciousness from its stagnant and entropic state-defined by the Law of Time as the delimiting 12:60 frequency consciousness of artificial, mechanistic time-while the globalization of consciousness, intimately connected with this sudden discontinuity, is necessary for the application of this law at a planetary whole systems level. Of course, the radical disruption of the historical continuum was provided by the collapse of the Twin Towers, and can actually be mapped in a description of the dialectical process of human evolution.

If one studies *Earth Ascending,* Map 19, "Binary Pulse of Psycho cultural Development," [shown on page 14] one can pinpoint the "Inevitable Event" of the radical historical discontinuity as the point at which the AC and CA currents cross over one another and switch polarity at the very top of the graphic depicting the "historical" process. What does it indicate, this point of psychic polarity switching currents?

The historical dialectic is understood as the interplay of two "currents" of consciousness, AC "Aboriginal Continuity" and CA, "Civilizational Advance." In the prehistoric phase, the consciousness is dominated by the Aboriginal Continuity or AC. This refers to the primary experience of the original human nature- life lived as a ritual or ceremonial round within the perimeter of the timeless present. But as the human enters the historical cycle, about which we will have more to say in later chapters, the CA becomes more predominant. As Civilizational Advance, the CA represents the tendency toward exclusive reliance on reason, logic and written means of communication. The Law of Time further defines the CA as being motivated increasingly by the artificial timing frequency of the 12:60 which gains in prominence until the final phase of the historical cycle, industrial globalization, when the CA has gained complete dominance of the biosphere.

During this final stage, the global mental condition is defined as the involvement in secondary or artificial means of experience, furthered by the planetary

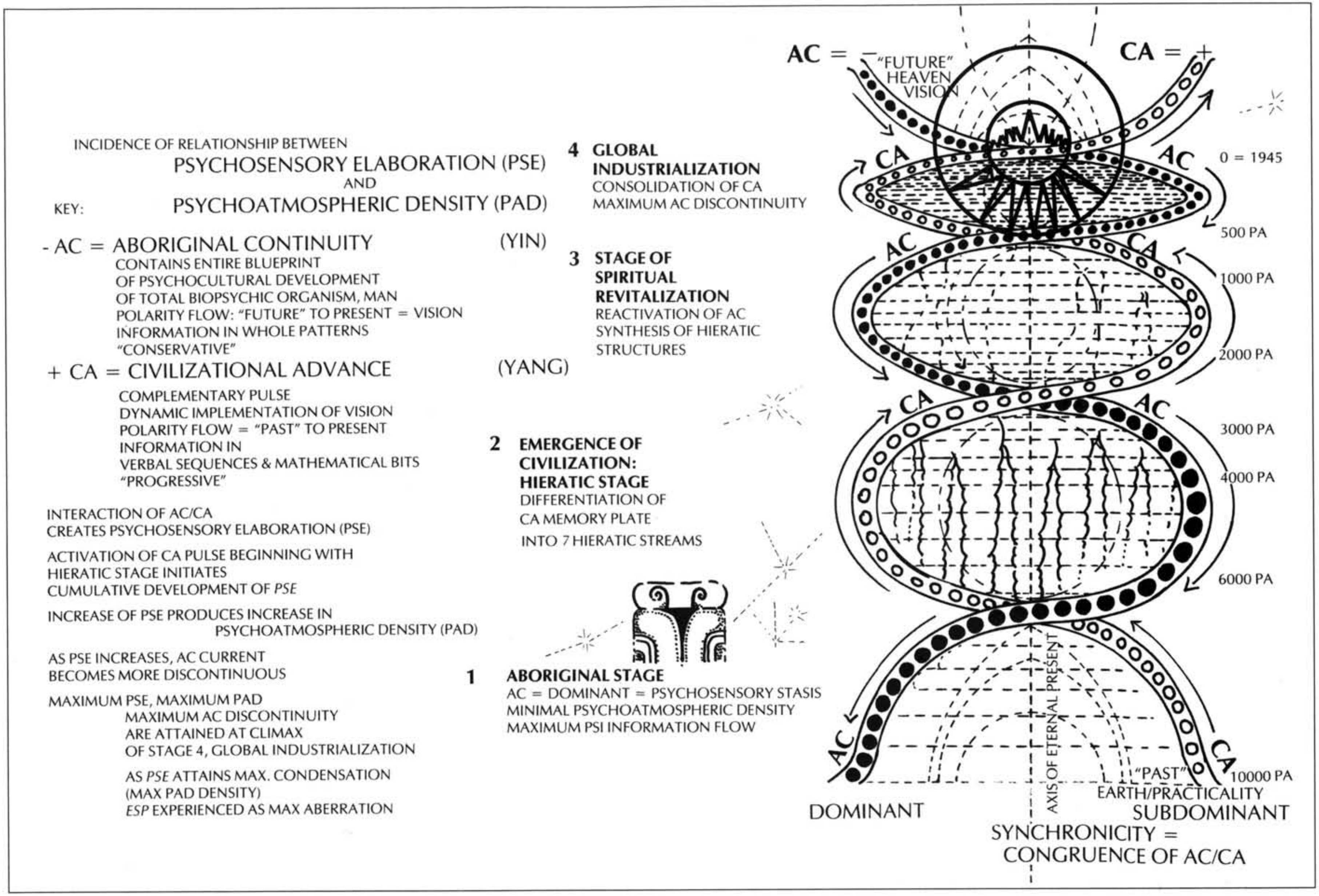
INCIDENCE OF RELATIONSHIP BETWEEN
PSYCHOSENSORY ELABORATION (PSE)
AND
PSYCHOATMOSPHERIC DENSITY (PAD)
KEY:
- AC = ABORIGINAL CONTINUITY (YIN)
CONTAINS ENTIRE BLUEPRINT
OF PSYCHOCULTURAL DEVELOPMENT
OF TOTAL BIOPSYCHIC ORGANISM, MAN
POLARITY FLOW: "FUTURE" TO PRESENT = VISION
INFORMATION IN WHOLE PATTERNS
"CONSERVATIVE"
+ CA = CIVILIZATIONAL ADVANCE (YANG)
COMPLEMENTARY PULSE
DYNAMIC IMPLEMENTATION OF VISION
POLARITY FLOW = "PAST" TO PRESENT
INFORMATION IN
VERBAL SEQUENCES & MATHEMATICAL BITS
"PROGRESSIVE"
INTERACTION OF AC/CA
CREATES PSYCHOSENSORY ELABORATION (PSE)
ACTIVATION OF CA PULSE BEGINNING WITH
HIERATIC STAGE INITIATES
CUMULATIVE DEVELOPMENT OF PSE
INCREASE OF PSE PRODUCES INCREASE IN
PSYCHOATMOSPHERIC DENSITY (PAD)
AS PSE INCREASES, AC CURRENT
BECOMES MORE DISCONTINUOUS
MAXIMUM PSE, MAXIMUM PAD
MAXIMUM AC DISCONTINUITY
ARE ATTAINED AT CLIMAX
OF STAGE 4, GLOBAL INDUSTRIALIZATION
AS PSE ATTAINS MAX. CONDENSATION
(MAX PAD DENSITY)
ESP EXPERIENCED AS MAX ABERRATION
4 GLOBAL INDUSTRIALIZATION
CONSOLIDATION OF CA
MAXIMUM AC DISCONTINUITY
3 STAGE OF SPIRITUAL REVITALIZATION
REACTIVATION OF AC
SYNTHESIS OF HIERATIC STRUCTURES
2 EMERGENCE OF CIVILIZATION: HIERATIC STAGE
DIFFERENTIATION OF
CA MEMORY PLATE
INTO 7 HIERATIC STREAMS
1 ABORIGINAL STAGE
AC = DOMINANT = PSYCHOSENSORY STASIS
MINIMAL PSYCHOATMOSPHERIC DENSITY
MAXIMUM PSI INFORMATION FLOW
AC = −
CA = +
"FUTURE"
HEAVEN
VISION
CA
AC
AC
CA
CA
AC
AC
CA
0 = 1945
500 PA
1000 PA
2000 PA
3000 PA
4000 PA
6000 PA
10000 PA
AXIS OF ETERNAL PRESENT
"PAST"
EARTH/PRACTICALITY
SUBDOMINANT
DOMINANT
SYNCHRONICITY =
CONGRUENCE OF AC/CA

Locating the Inevitable Event in the Holonomic Equation:
9-11, 2001 Climax of the Binary Pulse of Psychocultural Development
(from *Earth Ascending*, Map 19)

technosphere. The sum effect on consciousness of the advanced CA stage is defined as Psychosensory Elaboration or PSE, which in turn creates a psychoatmospheric density (PAD)-literally a mental cloud which obscures altogether the organic biopsychic reality of the human within the biosphere. From the point of view of a planetary pathology, this massive collective obscuring (CA) of the original nature (AC) defines a "suicidal crisis point," a moment at which the subconscious instinct of the human being programs its own self-defeat in order to ultimately return to the organic reality of the biosphere and the simultaneous advent of the noosphere. Hence, the Inevitable Event: the radical break in the continuum of history.

This moment of radical historical discontinuity occurred with the September 11,2001, terrorist attack on the twin towers of the World Trade Center-the very nerve center of late global industrialization-and the Pentagon, the supreme military fortress that defends globalization. If the fall of the towers symbolically augured the fall of Babylon and the Tower of Babel itself, the penetration of the Pentagon was no less symbolic. The fact that there was not one but two towers of Babel and a separation of some twenty minutes between the attack on each tower allowed the second attack to be viewed live on global television by about as many people on Earth as had ever witnessed anything at one simultaneous moment. By the end of the day, the scene had been repeated multiple times on virtually every television network on the planet.

The unbelievable shock of the moment and its simultaneous perception, coupled with the symbolism of what was actually occurring, effectively rendered the break in the historical continuum as a profound mental event-a rupture of 5,000 years of history from the first Babylon to the last. The witnessing of the event on television meant that it was received immediately into the nervous system of virtually the entire species. This in itself is a noospheric moment-the mental envelope of the Earth made collectively and globally conscious in one instant of history-shattering significance, and hence, the genuine entry into post-history. Archetypally, the images of the towers collapsing with people flying or jumping off the upper stories, is a pure reflection of the Sixteenth Major Arcana card of the traditional Tarot which shows a tower being shattered by lightning, with bodies falling down from it.

While the negative powers and forces grapple with their shattered dreams in the aftermath of the event, it is the time for the powerful luminescence of the noosphere to begin to activate the receptive cells of the human system within the now catalyzed biosphere. Fifty-six years after Hiroshima, the fall of the World Trade Center Towers is the fatal puncture in the technosphere which is intended to reorient the human within the biosphere. ESP (extrasensory perception)-the opposite of PSE (psychosensory elaboration)-is released in the instantaneity of the information reaching virtually every human being on Earth. This is the collective reactivation of the Aboriginal Continuity current that has been disengaged and disempowered for so long by the Civilizational Advance. Civilization has finally been checkmated. Now the Revolution of Time must be completed in order to stabilize the biosphere, anchor the noosphere, and de-structure the technosphere.

According to the Law of Time, the Inevitable Event was already embedded in the moment in which it actually occurred, like a flaw in a piece of glass that, upon concussion at the right pressure point, shatters all at once. In the case of the World Trade Center Towers and the Pentagon, we are also dealing with symbolic constellations of such potency that the very act of their shattering redefines the reality of human consciousness. In fact, the nature and problem of human consciousness as it has evolved historically is also at the root of this astonishing event. While it will take time for the collective human consciousness to fully grasp just how much its reality is being redefined, a vacuum now exists within the technosphere itself-a vacuum which awaits being filled by a new paradigm. Since the collapse of the Twin Towers actually represents the limits of artificial time, the new paradigm must rest on a redefinition of time altogether. This is the definition supplied by the Law of Time, the pragmatic application of which is the replacement of the irregular measure of the Gregorian calendar by the harmonic measure of the Thirteen Moon/28-Day calendar.

Because of its harmonic measure, the Thirteen Moon/28-Day calendar coordinated by the Harmonic Module is the new dispensation of time-reckoning for the human race. Rooted in the four recurring year-bearers of the Wizard's Count--:-which is not the same as the Mayan long count, nor the Julian count of modern science-the Thirteen Moon calendar exists for the establishment of the reality of the synchronic order as an entirely new method of knowledge and being for the entire human race as a cosmic medium operating within the biosphere. By Wizard's Count is meant the circulation of thirteen numbers and four different glyphs-Yellow Seed, Red Moon, White Wizard, Blue Storm-that code the July 26 Thirteen Moon/28-Day calendar synchronization date to create a master 52-year cycle, Anchored in the Thirteen Moon/2 8-Day calendar, the synchronic order is the means for manifesting the reality of the noosphere on planet Earth. Without such a basis in a lived mathematics of time correlated to the actual cycles of Earth, the noosphere could not be experienced nor fully realized. The successful completion of the biosphere-noosphere transition, then, is the triumph of the synchronic order, the conscious elevation of Earth into the cosmic frequency of fourth-dimensional time.

For all of this to come to be, however, is yet dependent on an analysis of the role of time in the creation of the technosphere and the subsequent comprehension of its limits and the potential for its de-structuring. This is not only a matter of theory or prophecy; it is a matter of putting into practice, making a coherent action out of the application of the Law of Time as it completes the pressing urgency of the biosphere-noosphere transition. The coherent action to bring us through this rite of passage is the Great Calendar Change of 2004. Ultimately, this book is presented to make us ready for this unprecedented time changing moment. May you, 0 reader, ponder deeply and then, if you so will, come join me in living on the other side of the wall of mechanized time.

▼

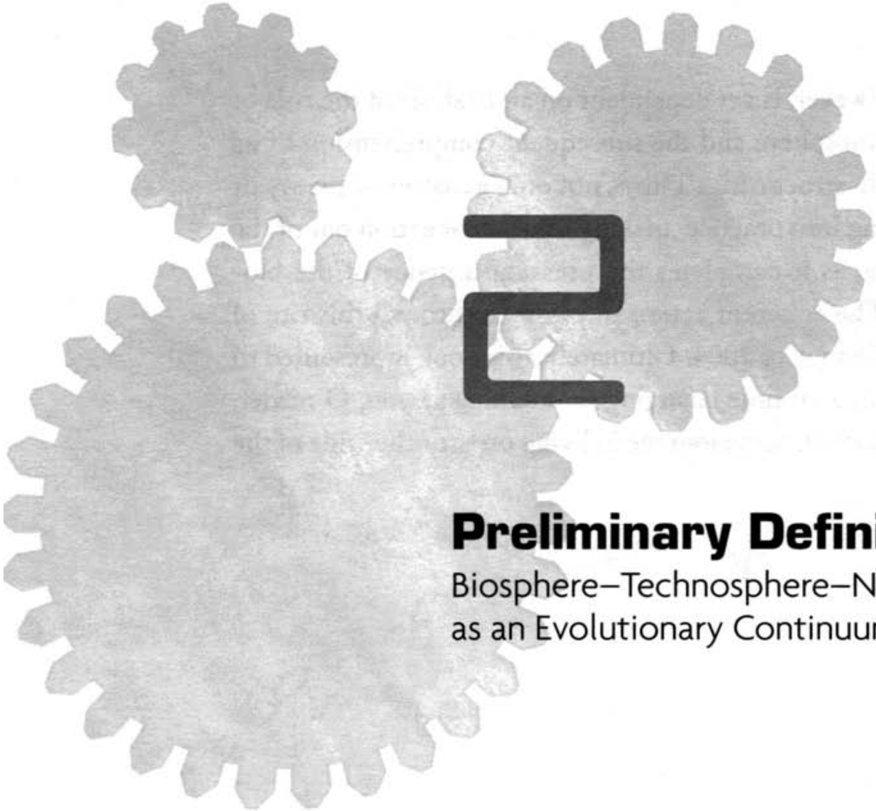

2

Preliminary Definitions

Biosphere–Technosphere–Noosphere as an Evolutionary Continuum

BY UNDERSTANDING that the Inevitable Event was a function of the biosphere-indeed, a necessary moment in the evolutionary trajectory of the biosphere-we are saying that this event was not just a human enactment, but that humans, inseparable from the biosphere, are effected to perform functions that relate to the process of the biosphere as a whole. Seeing the Inevitable Event and, in fact, the entire human drama from the perspective of the biosphere, we are actually lifting the cruel tragedies of contemporary reality above conflict and into a state where genuine laws of peace and harmony may shed their light of wisdom, illuminating the human drama with a higher understanding. One thing is certain: Unless we rise above our own humanity and cultivate a non-anthropocentric view-the view represented by the noosphere-there will be no release from the calamity that is now engulfing us. The point is not of maintaining the global economy, but of saving the biosphere.

For the multilayered task of bringing a new understanding to light, defining the technosphere, and then the Law of Time, we shall begin with the former: defining the technosphere in the context of the evolutionary continuum of biosphere-technosphere-noosphere.

VERNADSKY AND THE UNDERSTANDING OF THE BIOSPHERE

It is amazing that the name V. I. Vernadsky (1863-1945) is so little known in the West, and that the word *biosphere,* as well as the laws and principles pertaining to it, are scarcely more known. To a large degree, this ignorance is due to the great cleavage in human relations experienced as the Cold War. The biosphere, which was profoundly studied and scientifically articulated by the master Russian scientist V. I. Vernadsky, is actually the precise word for what in the West is vaguely referred to as the *environment.* But even more than being a precise and scientific description for what is meant by the use of the word *environment,* biosphere defines a whole system model of life on Earth, and because of that, also presupposes a whole system methodology and point of view. You cannot speak about the biosphere without entering into a world of discourse that is holistic or holonomic to the core.

If one looks for Vernadsky's work in English, one will find precious little available. His key work, *The Biosphere,* exists in two very different translations; the one published by Synergetic Press in 1986 is easier to read, while the other, the "complete annotated edition" published by the Far West Institute in 1998, is far more erudite, with a major bibliography and much supplementary information. The earlier translation was published in conjunction with the opening of the Biosphere II Project outside Tucson, Arizona. It is now a tourist site, with a bookstore that carries nothing by Vernadsky! Several other hard-to-find English-language books on the biosphere include the excellently informative *Traces of Bygone Biospheres* by Andrey Lapo (Synergetic Press and Mir Publishers, 1987), the flashy *Biosphere Catalog* (Synergetic Press, 1985), and the much more specialized *Energies: An Illustrated Guide to the Biosphere and Civilization* by Vaclav Smil (MIT, 1999). Of these books, only *Traces of Bygone Biospheres* contains significant information about Vernadsky and his ideas. This scant list of available texts in English is hardly proportionate to the vast reality of the biosphere as a complex terrestrial dynamic in the throes of evolutionary change.

The place and role of Vernadsky in Russian science is virtually equal to that of Einstein in Western science. A consideration of the achievement of each of these scientists also presents us with a dramatic contrast in perspectives, areas of interest, and consequent modes of analysis. The physicalist relativism of Einstein's legacy and its pursuit of the Big Bang is radically different in almost every way from the legacy of Vernadsky's work, which presents a biogeological worldview that is organically integrative and the basis of what in Russian science has come to be called *cosmism.* While the physicalist fascination of Western science has supported the materialism of modern Western thought and its way of life (which is actually an accelerating function of the biosphere's own internal processes), the whole system thinking of

Vernadsky—characterized by the conception of a "biogeochemical process"—ironically supports a cosmic rather than materialistic worldview.

To illustrate the Russian perspective and the influence of Vernadsky's work, we present the concluding thoughts of a book by another Russian scientist and philosopher, 1. Laptev:

> Each one of us requires the whole earth—today this is quite clear. But surely each one of us is required by the whole earth! And when these two mutually penetrating requirements, so distinctly revealed and intensified by the leap out into space, are satisfied, the new historical epoch will begin "in which mankind itself, and with mankind all branches of its activity, . . . will experience an advance that will put everything preceding it in the deepest shade." (Fr. Engels, *Dialectics of Nature)* And people, no longer burdened by the thirst for money, will remember something that many of our contemporaries have forgotten: The only important things in life are such intangible qualities as beauty and wisdom, laughter and love.
>
> Nothing remains for us, members of a society the ideals of which are in unison with elemental geological processes and the laws of nature, but to wish that people the world over would recognize this as soon as possible.!

Laptev's words—written twenty-eight years ago and precisely at the midpoint between Hiroshima and the Inevitable Event, as well as during the year in which the World Trade Center in New York City was finally completed and inaugurated (April 4, 1973)—simply and elegantly state the theme and point of view that we are pursuing. In fact, it is the Inevitable Event that prepares us for the new historical epoch. But first we must realize how we are organized by the biosphere in order that the technosphere may be transformed into the noosphere.

The root of Laptev's cosmic vision lies in Vernadsky's succinct definition of the biosphere: "The biosphere is the region on Earth for the transformation of cosmic energies." Although the word *biosphere* had been coined in the late nineteenth century, it was really only with Vernadsky's efforts that the biosphere became a viable description of the totality of life on Earth inclusive of its organic and inorganic support processes.

Vernadsky was trained as a geologist with a specialization in crystallography. He immediately absorbed the implications of Curies discovery of radioactivity (1896), and devoted many studies to the purpose of uranium and naturally occurring radioactivity within the Earth's geology. By the time of the First World War, Vernadsky had already written important texts such as *Fundamentals of Crystallog;raphy and Descriptive Mineralogy,* and, along with Madame Curie, proposed an "international radiography of Earth's crust." This was in advance of Alfred Wegener's definitive work on plate tectonics. During the time of the First World War and Russian Revolution, Vernadsky's attention turned toward the problem of living matter and its relation to the geochemistry of the Earth. From this was born Vernadsky's perception of the unitive biogeochemical process by which the biosphere maintains itself. In 1923 he published a "plea for the establishment of a biogeochemical laboratory."

In Paris, Vernadsky worked at the Insititut de Radium (Marie Curie) and became acquainted with Henri Bergson, who was then president of the International Commission of Intellectual Cooperation of the League of Nations. Through Bergson's circles, Vernadsky met often with the biologist Pierre Teilhard de Chardin (1881-1955) and the philosopher Edouard Ie Roy (1870-1954), with whom Vernadskyjointly coined the word and concept of the noosphere. Although Pierre Teilhard de Chardin is credited in the West with the concept of the noosphere, defined by him as the "mental envelope of the planet above and discontinuous with the biosphere," de Chardin's approach to the evolution of the noosphere as a natural consequence of vertebrate biology differs from Vernadsky's understanding in which it is the entirety of the biosphere that evokes the noosphere.

The chief fruit of this fertile time in Vernadsky's life was the publication of the little book titled *The Biosphere* (1926). Demonstrating the synchronic ordering principle of the noosphere, it is interesting that the same year that saw the publication of *The Biosphere* also witnessed the appearance of Jan Smuts noted book *Holism and Evolution.* Through most of the rest of his sometimes turbulent career (he often came under attack from hard-line bolsheviks), Vernadsky was to elaborate on the laws and principles of the biosphere, often expressed simply as problems in biogeochemistry. At the same time, his influence as a holistic thinker began to spread through the Soviet Academy of Sciences. A list of a few of his later publications gives some idea of the breadth of his thought, and includes: *Geochemistry, Problems in Biogeochemistry I* and *II, On the Boundaries of the Biosphere, Goethe as Naturalist, Scientific Thought as a Planetary Phenomenon, Some Words on the Noosphere,* and his never completed work, *The Chemical Structure of the Earth's Biosphere.*

Following the death of his wife in 1944, Vernadsky expressed the opinion that after the Second World War, American and Russian scientists should work more closely together. Early in 1945 (January 6), Vernadsky followed his wife to the grave. The Second World War ended, only to be followed by the Cold War (1947-1990). The critical collaboration between American and Russian scientists longed for by Vernadsky was soon to be buried in the ideological wasteland over which the superstructure of the technosphere would be constructed, a fact that in itself hastened the acceleration of the biogeochemical processes bringing about the climax of the technosphere.

Vernadsky did not live to witness the birth of the atomic era, which began only months after he died. But like Einstein in America, in 1940 Vernadsky-along with several other scientists, and mindful of the possible military applications-had prompted the Russian government to begin investigations into the possibility of deriving energy from nuclear fission. Yet it is most remarkable that by the time of his death, Vernadsky had almost single-handedly defined virtually every aspect of the laws and principles of the biosphere. This he always did with an eye to an imminent moment in the not-too-distant future when there would be a dramatic biogeochemical mutation or combustion, out of which the biosphere would somehow trigger the noosphere:

> Not at a crisis of nervousness do we stand now, not at a time for the vacillation of flabby souls; but at a great turning point in the history of scientific thought, at a crisis such as occurs but once in a thousand years, such as has not been witnessed for many generations. Standing at this point, with the vista of future achievements before us, we should be happy that it is our lot to live at this time and to participate in the creation of tomorrow.2

Undoubtedly his sense of the great moment of transition came from the very fact that during the span of his life from 1863 to 1945, some eighty-two years, Vernadsky was able to witness firsthand the accelerating effects of man's thought-in the form of industrial technology and the machine-and its turbulent and transformative impact on the terrestrial biosphere. After all, as a geologist, Vernadsky was familiar with the long history of the Earth, and the impact of human technology on the biosphere was inescapably the most significant aspect of any study of the biosphere to the present time. So wrote Vernadsky: "Only man transgresses the established order. . . upsets the equilibrium, though whether he materially cripples the transforming mechanism, or merely redistributes it, we cannot at the moment be sure."3

The "transforming mechanism" referred to is the biosphere itself, which Vernadsky had already defined as the medium or region for the transformation of cosmic energy on Earth. At its simplest level, the unity of the whole of living matter of the biosphere is the sum of its living organisms. These living organisms either directly or indirectly are continuously processing solar and cosmic radiation, transforming it into various chemical cycles that establish the atmosphere, most notably through photosynthesis and the oxygen and carbon dioxide cycles. The important point, however, in any consideration of the biosphere is the principle of unity: the biosphere is a unitive complex of cycles summarized as a single continuously transforming and slowly mutating biogeochemical process. In this single complex cyclical process, all of life-all of living matter-is a unity, a singular dynamic by which the biosphere can be considered as a living whole. From the plankton of the oceans to the trees in the forest and the human peering from his cave at the rising sun-all of this is a single unity.

It is the interaction of the living matter with the inert matter that creates and establishes the attributes and dynamics of the biosphere as a whole, all of which are now in a heightened state of acceleration and transformation, for as Vernadsky put it in the conclusion to *The Biosphere:* "The thought of the human species is a new fact which is turning the structure of the biosphere upside down after myriads of centuries." The structures and principles of the biosphere that man is now upsetting include: the continuity and invariance of life on Earth; vegetation as a transformer of the energy of the sun; multiplication as a manifestation of transformed solar energy; propagation of life and the forces resisting propagation; and the unique relations of life to the inert matter of the biosphere, including the composition of living matter into two orders, plant life and the various biological animal species, inclusive of man. The first order of plants and bacteria constitute the mechanism for capturing the solar radiation and introducing the energy of the sun into the biosphere. The second order of biological life is a manifestation of the process of the transmission of energy through the vegetable world; that is, at a primary and fundamental level, animal life is dependent upon plant life for its existence. (See plate 2, Life Is the Transformation of Solar Energy.)

All these principles for maintaining the equilibrium of the biosphere are functions of the dynamic of the different biochemical and geochemical cycles that maintain life both on land and in the ocean, all of which are subjected to processes that we refer to as evolution, both geological and biological. The evolutionary mechanisms of life include the principle of the biomass constant and the biogenic migration of atoms. Both of these mechanisms refer to and define the continuity and invariance of life on Earth. That is, the quantity of living matter has remained roughly unchanged since the origin of life on Earth-the biomass constant. Furthermore, the amount of oxygen in the biosphere is equal to the amount of biomass. It is the pressure of the different species on each other with the resultant interactions and changes that checks the infinite propagation of anyone species at the expense of the rest, and causes the mutation and evolution of the different life forms. It is through the intermediate structure of the technosphere that the human is wreaking the greatest havoc on these principles that are intended to maintain the biosphere as a stable dynamic, and by which the human constitutes a biomutation of unparalleled significance. The different stages of the mutation of life account for the biogenic migration of atoms: "The evolution of different forms of life throughout geological time increases the biogenic migration of elements in the biosphere."4 The metamorphosis of dinosaurs into fossil fuel is an example of the biogenic migration of atoms-and of the introduction of "free energy" (the term for industrial waste and by-products such as carbon monoxide) into the biosphere, while at a more mundane level, it refers to processes such as metabolism, circulation, breathing, and so forth.

To summarize, there are two biogeochemical principles. The first states: "The biogenic migration of chemical elements in the biosphere tends towards a maximum of manifestation"; this refers to the presence of life everywhere on Earth, as well as the principle of the pressure of living matter upon itself. The second biogeochemical principle states: "The evolution of species, in tending towards the creation of new forms of life, always move in the direction of increasing biogenic migration of the atoms in the biosphere. . . . The second biogeochemical principle regulates the course of evolution. A newly evolved species will survive in the biosphere only if it is sufficiently able and sufficiently active in furthering the migration of atoms. The organism cannot be considered apart from its medium. It is a part of the complex mechanism of the biosphere." Here Vernadsky concludes his study of the biosphere with a consideration of the laws regulating evolution that, for him, are to be sought in the mechanism of the biosphere and not in accidental circumstances. For this reason "evolution must proceed, namely, in the direction of increasing consciousness and thought, and of forms having greater and greater influence on their surroundings."5 This consideration leads Vernadsky to domains of philosophical or religious thought. Why?

"We are confronted with a new form of biogenic migration resulting from the activity of human reason. Human thought has changed in a brisk and radical manner the trend of natural processes and has even modified what we call *natural laws.* Consciousness and thought, despite the efforts of generations of thinkers, have never been defined, have never been given a physical basis, in terms of matter and energy."6 And here Vernadsky poses a question that remained unanswered for him but that also leads to a consideration of the noosphere: "How can processes which seem purely physical be affected by consciousness?"7

Although *The Biosphere* is the initial statement and underlying basis of all Vernadsky's later thought and work, his final major study, *Problems in Biogeochemistry II,* translated and published by his son, George Vernadsky, at Yale University in 1944, states the final conclusions of Vernadsky on the intriguing relation of human consciousness-via its thought and projected dynamic, the machine-on the biosphere. These reflections provide us with a means for defining more precisely the nature of the technosphere and noosphere.

FROM THE CREATION OF THE TECHNOSPHERE TO THE ADVENT OF THE NOOSPHERE

Vernadsky developed the biosphere into a full scientific description of the sum oflife and its support systems on Earth, and was a co-originator of the concept of the noosphere, yet he seems never to have used the word or concept of the technosphere. In some of Vernadsky's descriptions of the relation of man and his thought on the biosphere, however, the technosphere is absolutely implied. In reading the following chain ofVernadsky's thought, we can discern what we refer to as the biosphere-technosphere-noosphere continuum:

> Man by his work and his conscious attitude toward life is remaking a terrestrial envelope, the geological domain of life, the biosphere. He is transforming it into a new geological state, the noosphere.
>
> He creates within the biosphere new biogeochemical processes that did not exist before. A planetary phenomenon, the biogeochemical history of the chemical elements is notably changed.
>
> An immense new form of biogeochemical energy is represented by the technological work of man, complexly guided by his thought. It is interesting that the increase, in the course of time, of machinery in the structure of human society also proceeds in geometrical progression, just like the reproduction of any kind of living matter, man included.
>
> Statesmen should be aware of the present elemental process of transition of the biosphere into the noosphere. The fundamental property of biogeochemical energy is clearly revealed in the growth of free energy of the biosphere with the progress of geological time, especially of its transition into the noosphere.8

In this sequence of thoughts, we understand that Vernadsky, in his precise location in time, was himself the necessary voice of the biosphere giving rise to the noosphere. Put in another way, the very process that Vernadsky is describing-the transition of the biosphere to a new geological state, the noosphere-necessitated a coming into consciousness of itself, a self-reflective medium and voice. This self-reflective medium and voice of the biosphere-noosphere transition was Vernadsky

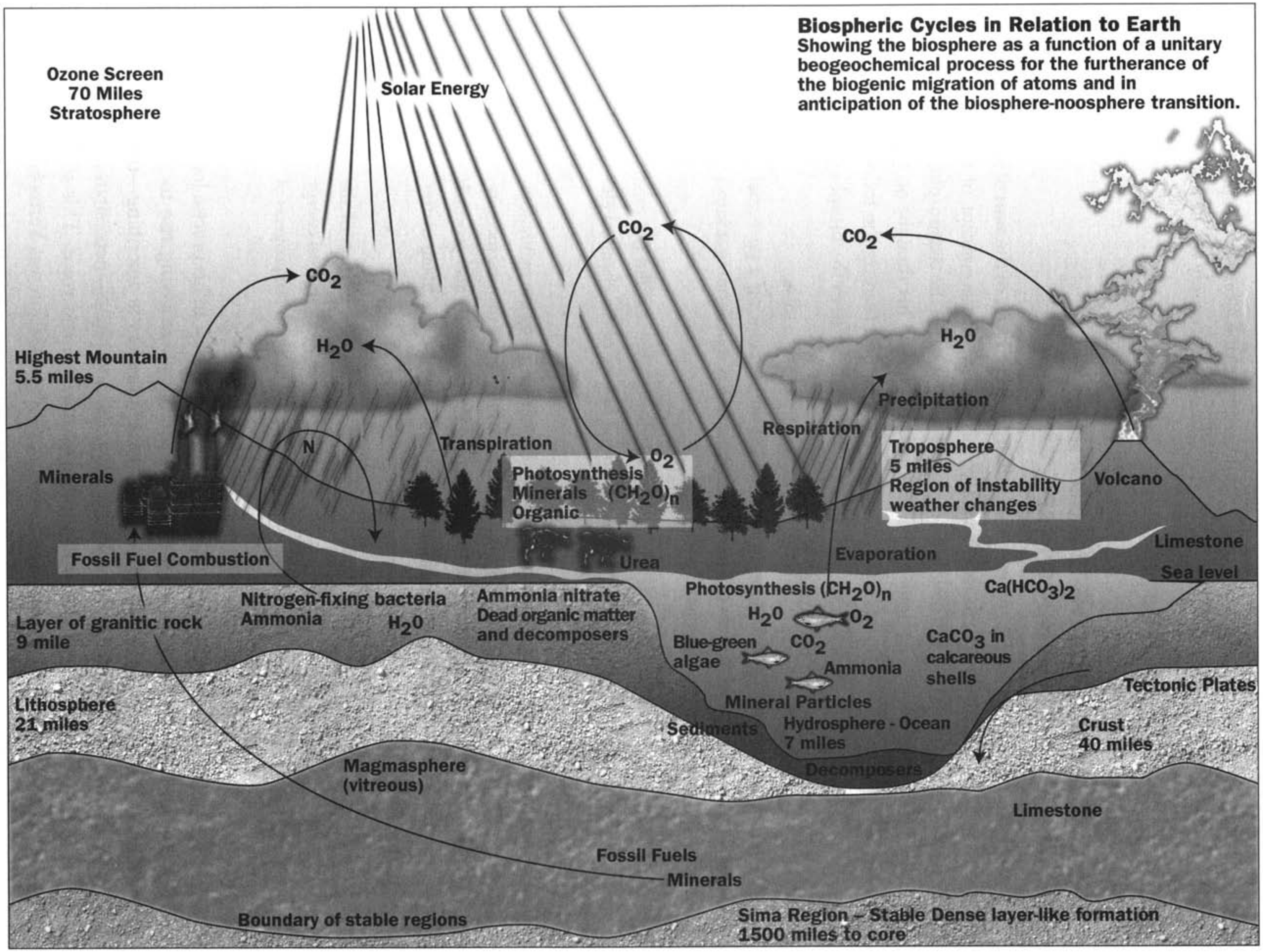
Biospheric Cycles in Relation to Earth
Showing the biosphere as a function of a unitary beogeochemical process for the furtherance of the biogenic migration of atoms and in anticipation of the biosphere-noosphere transition.
Ozone Screen
70 Miles
Stratosphere
Solar Energy
CO_2
CO_2
CO_2
H_2O
H_2O
Highest Mountain
5.5 miles
Minerals
N
Transpiration
O_2
Photosynthesis
Minerals $(CH_2O)_n$
Organic
Respiration
Precipitation
Troposphere
5 miles
Region of instability
weather changes
Volcano
Limestone
Sea level
Fossil Fuel Combustion
Urea
Evaporation
Nitrogen-fixing bacteria
Ammonia
H_2O
Ammonia nitrate
Dead organic matter
and decomposers
Layer of granitic rock
9 mile
Photosynthesis $(CH_2O)_n$
H_2O
O_2
Blue-green
algae
CO_2
Ammonia
$Ca(HCO_3)_2$
$CaCO_3$ in
calcareous
shells
Mineral Particles
Tectonic Plates
Lithosphere
21 miles
Sediments
Hydrosphere - Ocean
7 miles
Decomposers
Crust
40 miles
Magmasphere
(vitreous)
Limestone
Fossil Fuels
Minerals
Boundary of stable regions
Sima Region – Stable Dense layer-like formation
1500 miles to core

Biospheric Cycles

himself. In this observation we may understand that, although unconscious until it was made precipitant by the accelerated biogeochemical combustion itself, the order of the noosphere is none other than what the Law of Time defines as the synchronic order. As the structure of fourth-dimensional time, the synchronic order is a fluid, radial chronomantic structure in which time, telepathy, and consciousness are united in ways unpredictable to even the most advanced Western systems of mathematics.

For Vernadsky the sum properties of the elements of the biosphere are the functions both of mathematics and of different kinds of geometries. It is of note that in the summation of the differences between inert and living bodies according to three factors, the first being energy and the second chemical manifestations, Vernadsky defines the third factor as "differences in regard to space-time." Of this factor, Vernadsky writes:

> In considering space-time, however, the situation becomes rather involved. Here we enter, on the one hand, a domain not studied scientifically, and on the other hand, that substratum of all natural phenomena, namely their geometry, which the naturalist in his scientific work is wont to dismiss without attention. This substratum, the geometrical state of physical space, underlies all physiochemical phenomena and perhaps has an even deeper reality than that of the phenomena themselves.9

This deeper reality is the phenomenon of time itself, expressed through various geometries or geometrical principles, which informs the manifest phenomena of life. This is an obviously aesthetic factor as well as a cosm.Qlogical principle.

The Dynamics of Time fully substantiates Vernadsky's perception of the geometrical ordering principle of the deeper reality of time:

> All geometric forms are radiative and derive from the fourth-dimensional radial matrix. All fourth-dimensional functions are radial in nature and imply a principle of center from which the structure is projected. . .
>
> Geometry is how fourth-dimensional time incorporates as third-dimensional form. The incorporation of time as geometrical form informs all of the inorganic and organic orders of the realm of instinct. Time incorporated as the geometry of form is the principle example ofT(E) = Art. All forms and species participate in varying orders of geometry of form, both in their bodily structure and their forms of process.1o

If the question is asked, Where is this deeper reality of fourth-dimensional time from whence the projective geometries are generated? then we must answer that it is here in the noosphere, the medium on Earth for the transformation of cosmic thought.

It is the noosphere that informs all of the forms of the biosphere, inert and living, as well as producing in all genuine thought a structure and an order that are continuously moving in the direction of greater consciousness. To even think about the noosphere is to be a reflection of it.

As a noospheric medium, Vernadsky's life span was precisely timed to end when the global structure of the technosphere was about to be unconsciously constructed. The ingredients of the technosphere are indicated in the above-quoted passages as the planetary phenomena altering the biogeochemical history of the chemical elements, transforming them into an "immense new form of biogeochemical energy represented by the technological work of man, complexly guided by his thought." It is the increase, over the course of time, of machinery in the structure of human society that proceeds in geometrical progression-just like the reproduction of any kind of living matter, man included-and that constitutes the general nature of the technosphere. The growth of "free energy," the often toxic by-products of the industrial transformation of raw goods into consumer products, becomes the hallmark of the technosphere, the transitional stage between the biosphere and the noosphere.

It is this very point of the elemental process of the production of "free energy" and what it signifies in terms of the transformation of the biosphere into the noosphere that causes Vernadsky to call statesmen to attention. Why? As we know today, fifty-six years after Hiroshima, "free energy" in various manifestations has been causing danger signals to go off throughout the biosphere-from ozone depletion to global warming to nuclear radioactivity. What these danger signals have been telling us is that the techno sphere is coming to a point of its own exponentially accelerating biogeochemical combustion. Statesmen and politicians are advised by Vernadsky because the domain of their political rulership actually encompasses the everyday governing of the technosphere. But, alas, most people, statesmen included, are still unconscious of the existence of the technosphere, much less its role in relation to the biosphere and the forthcoming geological state, the noosphere. Hence, the Inevitable Event.

During much of Vernadsky's life, the geochemical combustion and release of "free energy" was just coming into its own. At his birth in the middle of the nineteenth century, the world was not yet globally industrialized, and without a totally globalized industrial base there can be no technosphere. The critical span of the two world wars, 1914-1945, was for the purpose of establishing the world market-a total industrialized market and communications base for the expansion of technology and the creation of the unifying global structure of the technosphere. The inauguration of the atomic age at Hiroshima, August 6, 1945, marks the precise beginning of the dual factors: the commencement of biogeochemical transmutation and combustion of the biosphere by the impact of human thought, the atomic bomb and release of nuclear radioactivity into the terrestrial atmosphere; and simultaneously, the beginning construction of the technosphere.

The bomb created a point of critical self-awareness within the human organism as well as producing in that organism an unprecedented sense of its own power. It was this sense of power that necessitated the unconscious impulse toward putting into place the structure of the technosphere. Of course, at the outset the species was only responding to a new, uncharted world situation, fundamentally unaware both of what it was creating as well as of the fact that it was actually constructing something in the nature of a transition between two geological conditions, the biosphere and the noosphere.

It could be said that Hiroshima lit the fuse of the final stage of biogeochemical combustion, the technosphere being the inevitable structure needed for the fuse to detonate at just the right moment. In this regard the synchronic order of the Law of Time ordained a precise 56-year cycle for the existence of the technosphere: 1945-2001. It is important to grasp that in this 56-year cycle the human population quadrupled, expanding from 1.5 to more than 6 billion. The 56-year cycle divides into precise subcycles of twenty-eight years each. The first twenty-eight years, 1945-1973, begins with Hiroshima and ends with the completion of the World Trade Center towers in New York City. This marks the accomplishment of the creation of the "World Market," the actual transformation of the biosphere into the technosphere. The second 28-year cycle, 1973-2001, spans from the dedication of the World Trade Center towers to their destruction by an act of suicidal planetary terror, and marks the phase of "globalization" -the absolute dominance of the materialist market economics, inclusive of the dramatic end of the Cold War in 1990. The very midpoint of the second 28-year cycle is the Harmonic Convergence of August 16-17, 1987, exactly fourteen years after the dedication of the World Trade Center and fourteen years before the inevitable moment of their destruction.

A profound statement of the way in which the Law of Time governs the entire biosphere, inclusive of the artificial and transitional structure of the technosphere, lies in the fact that the Gregorian calendar-an otherwise irregular and confusing pseudostandard of measure-is nonetheless coordinated by the Law of Time to repeat its cycle precisely every twenty-eight years. What this means is that the days of the week and months for the Gregorian calendar year 1945 repeated again in 1973, and yet a third time in 2001. The program prefiguring the Inevitable Event was already tested at Trinity Site (July 16, 1945), and established as the atomic destruction of

Hiroshima (August 6, 1945) and Nagasaki (August 9, 1945). This program of mass destruction was sublimated in 1973 as the triumphant completion of the Twin Towers of the World Trade Center, only so that in 2001 they could be the apocalyptic target of the inevitable 9-11 planetary red alert and moral emergency catastrophe, the only rival of Hiroshima and Nagasaki in fifty-six years. We shall return to this point later in our discussion on the Law of Time and the role of calendars as macro-organizing, programming factors that control the functioning of the human in the biosphere.

It is of special note that the very last work published before Vernadsky's death was the essay "Some Words about the Noosphere." In his vision, Vernadsky could hardly have conceived of what was to occur first as the technosphere. Yet, he wrote in this last essay, published 1944:

> The historical process is being radically changed under our very eyes. For the first time in the history of mankind the interest of the masses on the one hand, and the free thought of individuals on the other, determine the course of life of mankind and provide standards for men's ideas of justice. Mankind taken as a whole is becoming a mighty geological force. There arises the problem of the *reconstruction of the biosphere in the interests of freely thinking humanity as a single totality.* This new state of the biosphere which we approach without our noticing it is the noosphereY

By contrast, Teilhard de Chardin, in his most famous book, *The Phenomenon of Man* (1959), defined the noosphere as a "new canopy," a "thinking stratum" that has been unfolding since the end of the Tertiary period, and has since been unfolding discontinuous with and above the biospheric world of plants and animals. Teilhard de Chardin also speaks of "planets with noosphere," those worlds where the enlightenment of humanity has become synonymous with the life of the planet as a whole system. The one path to the noosphere defined by a scientist in Marxist Russia and the other by a Catholic biologist both point to the same radically positive end vision of the evolution of humanity, a fact that in itself is a dissolution of any kind of contradictory dialectic. Only history is governed by the merciless dialectical movement of matter and thought, rich and poor, material and spiritual. Clearly, therefore, the noosphere is beyond or after history. In this regard then, the technosphere is the concluding stage of history, to be followed by the post-historical noosphere.

The establishment of the techno sphere after his death notwithstanding, it is still important to listen to Vernadsky's idealism with regard to the noosphere: "The problem of planned, consistent activities which will aid us in mastering nature and accomplishing the correct distribution of wealth, connected with the comprehension of the unity and equality of all people, of the unity of the noosphere, is now on the agenda. . . It has become clear, and is more deeply penetrating the consciousness of mankind, that we now have a real opportunity to obliterate malnutrition, starvation, and poverty, to greatly diminish the incidence of diseases, and prolong human life to the maximum."12 Other thinkers like R. Buckminster Fuller have uttered similar statements, that we have all that is needed to adequately take care of everyone on the planet. It is important to understand why this optimism has yet failed to be substantiated. If anything, the inequality in wealth between that of the several hundred billionaires in the world and that of the poorest one-third of humanity, two billion humans, was never historically more disproportionate than it was at the moment of the Inevitable Event. This inequality of wealth, too, is a situation endemic to the very processes of the technosphere, otherwise known as the military-industrial complex.

In 1970 an American scientist, G. E. Hutchinson, wrote in the *Scientific American*, "By noosphere Vernadsky meant the envelope of the mind that was to supersede the biosphere, the envelope of life. Unfortunately the quarter century since those words were written has demonstrated how mindless most of the changes wrought by man on the biosphere have been. Nonetheless, Vernadsky's transition in this deepest sense is the only alternative to mankind shortening its lifetime by millions of years."13

A similar assessment was put forward in 1979 by the Belgian ecologist P. Duvigneaud in his presentation "Noosphere and the Future of the Vegetation of the Globe." In his address regarding the extermination of natural ecosystems made to the Twelfth International Botanical Congress, Duvigneaud makes one of the earliest and most prescient uses of the word *technosphere:* "Thus, the noosphere, so dear to Vernadsky . . . is running the risk in reality of turning into the technosphere, or rather into 'the sphere of avidity,' governed by the spirit of fortune-hunting, by mediocrity, by the absence of social consciousness, by the ideal of destruction and by the egotistic doctrine of 'after me, the deluge.'"14

To be precise, it was in 1973, the same year of the Club of Rome's landmark publication *The Limits of Growth,* and of the establishment of the World Trade Center in New York, that the technosphere became a coherent structure smothering the biosphere, co-opting the noosphere, and coordinating the very processes of greed and destructive fortune-hunting abhorred by the ecologist Duvigneaud. Like it or not, the techno sphere was the inevitable result of the unloosing of the machine into the biosphere. It may be asked: if the machine is a projection of human biology, then what purpose does it actually serve in the biosphere, and what is it that actually produced it and keeps it going?

Before turning our attention to the answers to these questions, let us first define the technosphere as being the artificial structure replacing the human civilization,

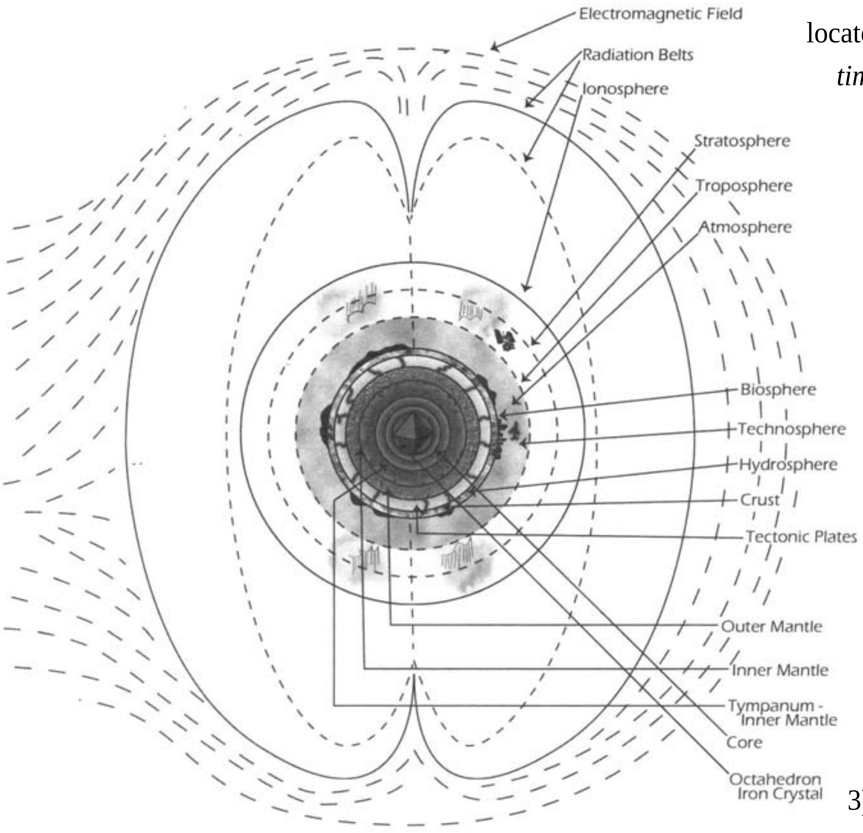
Electromagnetic Field
Radiation Belts
Ionosphere
Stratosphere
Troposphere
Atmosphere
Biosphere
Technosphere
Hydrosphere
Crust
Tectonic Plates
Outer Mantle
Inner Mantle
Tympanum - Inner Mantle
Core
Octahedron Iron Crystal

Earth's Mantles

located mentally and physically *in time* as a force intermediate to and between the biosphere and noosphere. Defined by the *Biosphere Catalogue* (1985, p. 103), the artificial structure of the technosphere consists of five interlocking, coordinated systems that comprise an integral whole, both as a belief system and as an actual complex technological dynamic and its coordinating apparatus. The five components, which we shall discuss at length in later chapters, are: 1) commodity production, 2) the city, 3) energy, 4) transport, and 5) communication. And, according to the *Biosphere Catalogue,* the planetary core of this interlocking set of systems was none other than the World Trade Center's Twin Towers.ls

As a planetary sphere, the technosphere takes its place with the other spheres or mantles comprising the structure of the Earth as a whole system. Beginning at the outermost etheric realm and proceeding inward, these spheres or mantles include: the electromagnetic field with its radiation belts, then the ionosphere, stratosphere, troposphere, atmosphere, technosphere, biosphere, land-ocean crust, tectonic plates, lithosphere, outer and inner mantles, and finally the core, now known to consist of an outer tympanum, and within it a large octahedral iron crystal afloat in a sea of magma.

As a planetary sphere, the technosphere negates time and consciousness as organic realities and replaces them instead with mechanization and marketing. Yet the technosphere is a product of mankind, and mankind is a function of the biosphere. The organization of life into biomes, large macro-ecosystems, is still evident, though in diminishing and withering proportion to the conquest of the technosphere. With the Final War now upon us, the question we must ask is: will the technosphere collapse before the biosphere or not? As catastrophic as it might seem, the collapse of the technosphere would actually be a blessing in disguise from which the noosphere could finally be realized. But if the biosphere collapses first as the result of the Final War on Earth, then the technosphere will be futile, pointless, and dead as well. These are most dire thoughts in these most dire of times. Yet if we can now finally understand and accept our role within and subordinate to the biosphere as a whole, then our chances for survival will be greatly improved.

Let us consider again what Vernadsky has to say about the human's place in the biosphere:

> 1. Man, as observed in nature, like all living organisms, like any living matter, is a specific *function of the biosphere,* within its specific definite space-time.
>
> 2. Man in all his manifestations constitutes a regular part of the structure of the biosphere.
>
> 3. The *explosion* of scientific thought in the twentieth century *has been prepared for by the entire history of the biosphere* and has its deepest roots in the structure of the latter. The civilization of "cultured mankind," insofar as it is a form of organization of the new geological force which has formed in the biosphere, *cannot be interrupted or destroyed,* since it is a great natural phenomenon, historically, or rather, geologically corresponding to the established orderliness of the biosphere.16

If Vernadsky is correct, then the war we are witnessing is the inevitable conclusion of the biogeochemical combustion that consumes the technosphere and establishes the pristine reality of the noosphere. This is especially true if we understand this war to actually be a geopolitical struggle for the control of the oil reserves of the Persian Gulf, or even more broadly, as the war between matter (globalization) and spirit (Islam), or more grimly, as Oswald Spengler put it, the final war between blood and money. This being the case, it is more important than ever to answer the two unresolved questions in Vernadsky's definition of the biosphere, the question of consciousness and the question of time.

According to the Law of Time, the answers to the questions posed by Vernadsky are actually one and the same thing. "How can processes which seem purely physical be affected by consciousness?" The question asked by Vernadsky at the conclusion to *The Biosphere* is equaled by the unresolved issue of time discussed at the conclusion to *Problems of Biogeochemistry* JI.1n essence, Vernadsky declared, we know everything about the biosphere, its governing laws, principles, and functions, with the exception of the mysterious relation of time to the biosphere. Virtually alone among modern scientists, Vernadsky perceived that time actually has nothing to do with the metrics of space, and that "time is not a dimension of metric geometry." This means that time is not only its own dimension-the fourth dimension-but that time must also have its own metrics and system of measure totally apart from the metrics of space.

> Minkovsky's and his predecessors' concept of time as the fourth dimension of space is a mathematical abstraction having logically no ground in scientific reality. Time is not a dimension of metric geometry. In geometry, time may be expressed vectorially, but it is obvious that such an expression does not embrace all of its properties in natural phenomena studied by the naturalist. . . The time of the naturalist is not the geometrical time of Minkovsky, nor is it the time of mechanics and theoretical physics.[I?]

In Vernadsky's perception of time, there is a profound critique of the approach that has been established toward the measure-and hence understanding-of time in physicalist science. For Vernadsky, time is a function of an as-yet-unrealized geometry: "This geometry would reduce all space to a point supplied by an infinitesimal vector." Clearly time is a factor in all the functions of the biosphere, most evident in the biological life processes. Hence the Law of Time has been a regulating factor, the true nature of which we have been profoundly unconscious. Now here is an interesting reflection. If the biosphere as a whole system is tending toward a total whole system transformation into a state known as the noosphere, a fact perceived by Vemadsky during his lifetime, and the two unresolved issues are the issues of time and consciousness, both materially intangible dimensions, does it not seem correct that the resolution to these two issues will actually foster the manifestation of the noosphere, which is, after all, the mental envelope of Earth? And does it not also seem that as long as the two issues of time and consciousness remain in a state of ignorance or confusion within the mental functioning of the human order, that the vehicle of geochemical combustion, the technosphere, will tend toward its own dissolution until the issue of time and consciousness is resolved and clarified once and for all?

The perilous passage, the biosphere-noosphere transition, may actually be defined as the consequence of operating in ignorance of the actual nature of time and consciousness. In this way, the techno sphere may be defined as the artificial compromise in place of the true understanding of time and consciousness of the human order in the biosphere. Hence, the technosphere is the responsible intermediary agent creating the crisis of the noosphere. The solution to the questions of time and consciousness is thus the solution to the noospheric crisis, and the commencement of the grand new era, the true golden age of the whole Earth.

▼

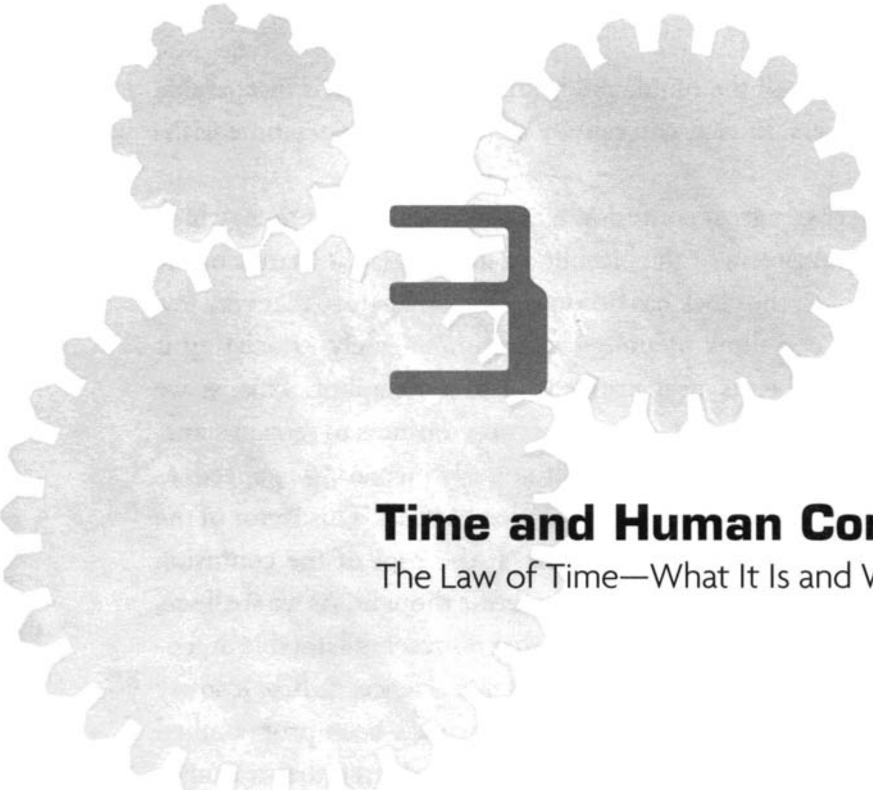

3

Time and Human Consciousness

The Law of Time—What It Is and Where It Came From

HAVING DEFINED the technosphere as the intermediary stage in the biogeological continuum of the biosphere's evolution into the noosphere, we may now turn our attention to the unresolved issues of time and consciousness in the biosphere. These issues must not be seen as just the problems of humanity. Since humanity is a function of the biosphere, the issues therefore relate to the biosphere becoming more conscious. It is the conscious aspect of matter that directs the biosphere toward its transformation into the noosphere. It is the discovery of the Law of Time that complements and resolves the issues that were left unresolved in Vernadsky's definition of the biosphere and of its transformation into the noosphere. The question then arises: what is the Law of Time, and how does it define time and consciousness as mutually coordinating factors of the same cosmic dimension intersecting with the biosphere?

To say the least, time is such a vast and important topic in the orientation of human consciousness within the biosphere that we may declare that it is paramount in all human affairs. Indeed the cosmic order, of which the biosphere is the dynamic regulatory mechanism on planet Earth, is itself the expression of time as the medium of universal manifestation. Yet, as Vernadsky perceived in 1944, while we have defined space by its own systems of metrics and geometries, by applying these same standards to time we have not defined or known time at all. The cause for this quandary seems to lie in the fact that while space is perceptible as a sensory medium-we can see it, touch it, hear through it-and hence, it is evidently measurable, the same is not the case with time. Time is of the mind, and in this regard, time is inseparable from the issue of consciousness. In fact, you cannot know or experience time without becoming conscious.

Yet, as we shall see, there is a great confusion when we speak about time, which is actually at the root of the problem of the biosphere, and that is the matter of the clock. In human consciousness, the clock has become so second nature that very few people in industrial society can think of time without immediately associating it somehow with the clock, whether as an instrument or as a metaphor. This, as we shall see, is a major factor causing the general state of consciousness to remain static, especially in relation to the machine, which is continuously increasing in speed, as well as accelerating in the propagation and multiplication of itself. This factor of the clock-in other words, of mechanized time-is also at the root of the confusion concerning the nature of time in much of modern scientific thought. As we shall see, there is a stream of Russian thought after Vernadsky that represents a notable exception to the established view of time in modern thought and science. Suffice it to say that the evolution of modern Western thought and science has been programmed and predisposed to limit its consciousness of time to such a degree that it cannot even perceive of time outside its inherently mechanized perceptions about it. In this obviously also lies the problem of the delimitation of consciousness in the technosphere.

"The deepening crisis of the world outlook is rooted in the permanence of the dominating scientific paradigm," writes Russian scientist Maria Maroushkina. "Underlying the materialistic technocratic civilization is the Newtonian-Cartesian paradigm, formed in the Middle Ages, its nucleus being the laws of gravitation and movement, as well as the logical principles of proof, laid down by Descartes."! This "permanent" stratum of belief, a cultural metaprogram, locks every field of human endeavor into a conquest by mechanization. It is this metaprogram rooted in the simple unexamined device of the clock, and orchestrated by the equally unexamined effects of an irregular calendar, that keeps human consciousness unchanged and without any real solutions to its problems. As the problems worsen, the confusion of consciousness-which is really in a state of crisis-only deepens.

To illustrate the current state of confusion regarding time in Western thought, we have the report of the prestigious Seven Pines Symposium on Time held in Stillwater, Minnesota in June, 2001. The *New York Times* article regarding the Seven Pines Symposium is entitled "Physics' Big Puzzle Has Big Question: What Is Time?"2 First of all, the assumption that time is basically a topic of physics in itself is a problem, since time, as we have demonstrated, is actually mental in nature, and if any-

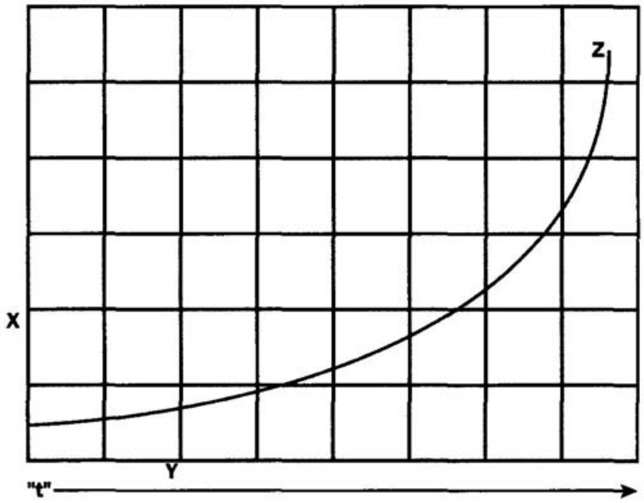

"Arrow of time" in modern scientific three-dimensional coordinate graphing, reinforcing illusion of linear time.

Arrow of Time

thing is most visibly manifest in the domain of biology or biogeology. But this headline is also reflective of the paramount role that physics plays in the Western intellectual hierarchy: if physics can't solve it, than what can?

Despite the prominence of physics and physicists, the Seven Pines Symposium included historians and philosophers as well. The hope of the symposium was to arrive at a unified theory that encompasses the effects of gravity as described by Einstein's general theory of relativity and the "fuzziness that occurs in the realm of tiny particles according to quantum physics."3 Here we have the problem of time as delimited by the self-imposed spectrum of modern physics. What then could we expect? Certainly no one here had considered the point made by Vernadsky that time is not a function of the metrics of space, which is the realm of physics. So it is not surprising that the symposium ended in a muddle, with the most prominent approach being that time could merely be a "psychological illusion" important to humans but not a fundamental part of any unified theory. It is also clear that few, if any, had considered that their views on time might be absolutely conditioned by the mundane and ordinary everyday micro- and macro-organizing factors: the mechanical clock and the Gregorian calendar with its second cousin, the Julian count of linear time.

From the point of view of the noosphere, of course, this grand Symposium on Time was reflective of very little of the entire human organism, much less of the biosphere itself. The symposium was a representation of the elite intellectual class of the dominant economic, military, and political power of the planet, and hence of the governing force of the technosphere, the United States of America. But if a realization that we are a function of the biosphere is critical to our success in launching the noosphere, and the issue of understanding the correct relation of time to the biosphere is crucial even to the ability to launch the noosphere, then the Seven Pines Symposium did not even approach the actuality of the current biospheric situation or the nature of time. In this regard it was only a contribution to further the psychoatmospheric density-clouded thought without clarity of purpose-and one more reason why the Inevitable Event was on the biosphere's agenda.

If one does not apply certain biospheric yardsticks to the analysis of what is reported as truth, then we will be sunk in a chaotic morass of half-truths that are really obscurations or even lies. The elite American nation, for instance, constitutes only 5 percent of the world's population, yet consumes 40 percent of its natural resources, and therefore has the wealth and power to maintain the military force (allotted a $318 billion budget for 2001-2002) to protect and even further as its right this disproportionate ratio, which is actually a violation of the dynamic balance of the biospheric processes. For this reason, through its official organs of propaganda such as the *New York Times,* America is able to muster the "influence" of whatever occurs within its confines as somehow having more merit than what might be generated in some other nation or country. So a human may read the story in the *New York Times* on the "Big Puzzle of Time," and most likely they will complacently accept its premise as absolute truth. Yet this social construct of symposium-and-newspaper-story is nothing but an aberrant mental form that obscures the actual nature of time and keeps the mental level at a constant state of "industrial dull-normal," far removed from-or at best, little more than a highly distorted reflection of-the reality of the noosphere.

Leaving aside this example of the mental confusion over time, it is necessary to understand from the investigations regarding the biosphere and according to Vernadsky's own judgment, that the biosphere in all its intricate and interlocking processes possesses and is guided by a great order and even a structural dynamic that, being manifestations of the laws of nature, can be given exact mathematical formulation. Why should we not expect that the nature of time in the biosphere should also be characterized by the same orderly quality? In fact, given that the biosphere is actually a life process, like all living matter it too must be governed by timing cycles. Thus the larger order of time that governs the cycles of the biosphere and of the noosphere must also exhibit a grand ordering principle. For if the three dimensions of space govern the domain of the physical world, then the fourth dimension of time-which governs the life not only of the biosphere but of planets, star systems, whole galaxies, and indeed the entire universe-must also consist of a simple, orderly formulation totally apart from the perception of time that governs the stunted consciousness of the technosphere. Such is the Law of Time.

THE FOURTH-DIMENSIONAL NATURE OF TIME

> *"Even though you do not measure the hours of the day as long or short, far or near, you still call it twelve hours. Because the signs of time's coming and going are obvious, people do not doubt it. Although they do not doubt it, they do not understand it. Or when sentient beings doubt what they do not understand, their doubt is not firmly fixed. Because of that, their past doubts do not necessarily coincide with their present doubt. Yet doubt itself is nothing but time."*
>
> -DOGEN, "THE TIME-BEING"4

These prescient words of a thirteenth century Japanese Zen master are more true today than they were when he wrote them. Comprehensive studies like the *Eranos Yearbook III: Man and Time* (1957), which included Carl G. Jung's ground breaking essay on synchronicity, or James T. Frazier's fascinating and monumental *Voices of Time* (1966), only expand the multidisciplinary breadth of the confusion about time. In the 1970s Frazier helped found the International Society for the Study of Time. The Society is still meeting annually but has yet to reach a definitive agreement on or understanding of the question of time. It is in the context of the enduring confusion over time that the Law of Time was discovered. To say that the Law of Time has been "discovered" only means that a principle and fundamental law that has always guided the order of the universe has finally been made conscious and articulated in a precise way that is recognizable to the present condition of the human mental nature, which is still dominated by the physicalist model.

In the whole of modern physics, Minkowski's fourth dimension of time has been misunderstood and treated as if it were a minor dimension, one that is just added on, but not really significant to the level of the three dimensions of space. This is due to the already space-oriented consciousness deriving from certain sets of perceptions established early in the history of civilization and standardized through certain programming models that are based on the metrics of space and not time, a topic to which we shall soon turn. It is for this reason that in the formulations of modern physics time is symbolized by a small *t* that runs in a horizontal line running from left to right at the bottom of any graphing of space with its *x* and *y* coordinates.

This graphing of small *t* time gives rise to the notion of the "arrow of time," which is virtually a bedrock dogma of much of modern Western physics. This is the much-touted notion of linear time that supports doctrines of economic inexorability, material progress, and the like. Of course to anyone versed in the ways of nature, linear time and the arrow of time can be seen as nothing more than artificial constructs, for the biological nature of time is perceptible at the very least through the great interlocking cycles of nature. Nor is there anything linear about the Earth's rotations around the sun, or of the moon around the Earth, all of which give lie to the physicalist and materially progressivist notion of linear time. Only the technosphere, sustained in its structure by the limited consciousness of time, is driven by this linear principle and for this reason is doomed to run aground on the shoals of its own artificiality.

TIME AS THE UNIVERSAL FACTOR OF SYNCHRONIZATION

The nature of time is grasped poetically in the form of cycles: "To every season there is a time. . ." Biorhythms can even be mapped and the cycles of the biological order of reality can be demonstrated as functions of a chronobiology. But the biological cycles only demonstrate the manifestation of time in the space ofliving matter. If the whole of nature is observed, then one can begin to grasp the masterful orderliness by which every least detail occurs in relation to all the other details in their various cycles-yes, one can grasp in this phenomenal order another deeper aspect of time, and that is time as the universal factor of synchronization. This is the synchronic order of time, the order by which everything in the universe occurs simultaneously in a masterful synchronization from moment to moment in an ever-changing kaleidoscope of infinite varieties of order and harmony.

Only historical man deviates from this masterful symphony of time, the synchronic order. Nonetheless, we speak of being "on time," "in sync," "in tune," "in resonance," "in harmony," or of being "tuned-in." All of these expressions refer to the attunement of time and the nature of consciousness as awareness in the present moment, in the here and now. This is possible because the same law governs both time and the relation of consciousness to the moment. To answer Vernadsky's question, "How can processes which seem purely physical be affected by consciousness?"- the physical is affected by consciousness, because consciousness is actually a function of time. When someone "discovers" a law of nature, it is because their consciousness is attuned to the natural process in time, and the law therefore "reveals" itself.

But this consciousness can be and most often is of a profoundly unconscious nature. Time is actually the governing principle of a higher self-existing consciousness that regulates the order of the universe, both in regard to its living and non-living matter. Because time continuously synchronizes everything into a single coherent whole from the micro to the macro levels, time accounts for the harmony of the universe. Only man deviates from the universal harmony, and the technosphere, a projection of the human mechanization of time, intrudes upon the intrinsic harmony to the point of its own demise, but man deviates for a purpose.

The harmony manifested by time as cosmos-which literally means "order" -gives rise to the great and simple formulation of the Law of Time: T(E) = Art; Energy factored by Time equals Art. This means that energy, any manifestation of the physical three-dimensional world, possesses order and is in harmony with its environment because it is factored by time. In the conception of the Law of Time, however, capital *T* refers to time as the universal factor of synchronization. As the universal factor of synchronization, time is defined by the self-existing and intrinsically perfect mathematical ratio 13:20. Derived from the mathematics of the ancient Maya, this ratio is a universal constant of time that organizes all of the universe as a radial sequence of synchronous moments reflecting different evolutionary phases simultaneously. Because everything perceived is an aspect of instantaneous universal synchronization, time is also the medium of instantaneous information transmission throughout the universe.

The noted Russian astrophysicist N. A. Kozyrev (1908-1983), whose work began half a century ago with the theory of the internal structure of stars, later conducted a famous series of experiments using telescopes, mirrors, and aluminum, in which he was able to confirm a receipt of information from distant galaxies prior to the time it took the information to arrive via light as a physical medium. In Kozyrev's own words: "The tests proved the existence of the effects through time of one material system upon another. This effect does not transmit a pulse (momentum), meaning it does not propagate but appears simultaneously in any material system. In this manner, in principle it proves possible to have a momentary relationship and a momentary transmission of information. Time accomplishes a relationship between all phenomena of nature and participates actively in them. . . . Time contains the entire universe of still unexplored phenomena."5 The results of these experiments led Kozyrev to various conclusions, among them that time and duration are not the same thing, and that the velocity of time is instantaneously infinite, formulated as

$$T_V \longrightarrow \infty$$

Formulation: velocity of time is instantaneously infinite

This being so, time must also be the factor or medium that accounts for the experiences of telepathy or other paranormal phenomena. A second, most provocative conclusion derived from his experiments is that time is actually generated as an instantaneous and even simultaneously radial transmission of information from the core of stars, a fact confirmed by the Law of Time's description of the synchronic order of time as being radial in nature.

FURTHER EXPANSES OF FOURTH-DIMENSIONAL TIME: THE EVOLUTION OF TIME AS CONSCIOUSNESS

Leaving aside for the moment further discussion of Kozyrev's seminal work and its conclusions, at this point let us summarize the two chief qualities of time as defined by the Law of Time: 1) time is the factor of universal synchronization, and 2) time possesses an infinite instantaneity of velocity that is faster than the speed of light, and hence is correlated with telepathy. From these two determining features of the nature of time, two other corollary conclusions may be drawn: 3) the cyclic nature of third-dimensional biological phenomena is a function of the factor of universal synchronization which establishes harmonic ratios that determine the different life processes of different organisms, while 4) the ability of any species or organic whole to maintain its unity with itself and all of its members or parts is an instantaneous function of time as telepathy.

To better understand the Law of Time and its radical non-linear definition of time-which is actually the noospheric description of time-we may say that time, being synchronically instantaneous, is both "vertical" in relation to space (which is "horizontal"), and "radial" from the perspective of its own dimension, the fourth. Rather than being the small *t* line or arrow at the bottom of the physicist's graph, time as the fourth dimension is greater than and inclusive of the third dimension, much as the etheric atmosphere includes and surrounds the physical Earth in a radial "all-at-onceness." To try to measure this phenomenon with the metrics of space and its essentially linear yardsticks is to constrain time into definitions and descriptions that are totally alien to its nature.

In fact, from the point of view of the dynamics of time, it is space that is the moving line, or rather a set of points in a moving vector that can be described as line. Let us extrapolate from the important observation made by Vernadsky at the conclusion to *Problems in Biogeochemistry II:* if the geometry of time reduces space to a point supplied by an infinitesimal vector, then from the point of view of time, space is an infinitely locatable point. What is this point? It is the point of awareness at which an individual body becomes aware of time. Here, the "subjective" description or experience of time cannot be separated from the issue of consciousness. (See plate 3, Model of "Vertical Time" in Relation to Horizontal Space.)

In the description of time as vertical in relation to horizontal space, this point of awareness of a body in time is defined as the Locus of Consciousness, or L of C. In the flow of moment-to-moment instantaneity that is synchronous with the flow of space, the Locus of Consciousness is that point at which time vector and space perception connect. From this defining point of consciousness in the here and now, the flow of space creates a horizontal sensation: one direction flows to the "past," the other to the "future." This is what is normally defined as the sensation of time or the arrow of time. It is actually the definition of the horizontal flow of space in relation to vertical time. What is usually not considered is the vertical dimension of time that connects awareness to space at any given moment, creating sequences of Awareness Units, or AUs.

From the defining point of the Locus of Consciousness, there are essentially two spaces and two times: *near space* and *far space,* and *low time* and *high time.* Near space corresponds to low time and far space corresponds to high time. Near space is essentially defined by the perimeter of sense experience, inclusive of the extension of the senses through technological apparatus such as telescopes, microscopes, and the like. This is the domain of low time. Far space is defined as that space beyond the perimeter of the sense perceptions, and is characterized by subliminal sensory perceptions. This is the domain of high time and the atemporality of space. In near space, low time is characterized by what appear to be "random events." The lower the time, the greater the increase of random events. Low time and near space define the qualities of the experience that are properly and purely third-dimensional. High time and far space are purely and properly the domain of the fourth dimension characterized by the dominance of the synchronic order of time.

The distinction between low time and near space and high time and far space is defined as the threshold of synchronicity. In high time, synchronicity (s) predominates and is greater than random events (re), or (s»(re). In low time, random events predominate and are greater than synchronicity, or (re»(s). Note that the realm of quantum physics occurs in the microrealm of far space. It is this factor that accounts for the "fuzziness" of the behavior of subatomic particles, quarks, tachyons, and so forth, which defy the limitations of linear time. These phenomena occur in the subliminal realm of the synchronic order. It is the human mind that is not accustomed to behavior in the pure realm of synchronicity that cannot grasp that these particles are moving according to radial laws of fourth-dimensional time, and so appear highly irregular and random to the sense perceptions of the observers that are totally conditioned by third-dimensional space perceptions. Similarly, the high time of the macro threshold of the synchronic order represents the "other side," the "hereafter," the realm of pure vision, which is also governed by strict laws of radial time, and which, to the conditioned experience of a perceiver on this side of the threshold, may appear as something from a distant future or past reality, or even as in a dream. The orderly explication of the Law of Time provides complete mathematical descriptions and principles that account for the radial order of time, even as it coordinates the near space and low time of random events.

In any case, the horizontal line of space in relation to the vertical line of time defining any given Locus of Consciousness describes the macroworld above and the microworld below. As we have noted, the higher threshold of synchronicity above the horizon of space defines the realm of *higher consciousness*; the lower threshold of synchronicity below the horizon of space defines the realm of *quantum physics.* Connecting the macro realm of atemporality and high time, and the lower or micro realm of atemporality and high time, is the vertical time vertice of the ever-present now. At the two far ends of this time vertice of the ever-present now is the Locus of God, the selfsame in each direction. These extreme points of the vertice of time return far space to the Locus of Consciousness through a process of toroidal sublimation. We are not really dealing with a description of a two-dimensional plane, but a dynamic four-dimensional model held together by a toroidal motion that is continuously returning the farthest space and the highest time, whether from above or below the threshold of synchronicity-" God" -to the Locus of Consciousness.

Depending on the clarity of mind at any given moment, the vertice of time constantly and instantaneously transmits information to the Locus of Consciousness. The discrete units of information thus transmitted are referred to as *celestialharmonics.* A celestial harmonic is described as the index of synchronic incidence defining different levels of co-occurrence. I can be in the present moment sitting in a room, and at the same time, by being fully present in the here and now, I can spontaneously experience any number of telepathic thoughts, memories, feelings of deja vu, and so forth. These experiences are the different levels of co-occurrence that define the index of synchronic incidence and are categorizable as celestial harmonics. The higher the time, the greater the density of celestial harmonics per AU (Awareness Unit). It is the vertical vertice of time and its power of transmission that account for the visionary or revelatory experiences of saints, poets, and mystics throughout the ages. It is the mathematics of the Law of Time that provides the "mapping" that explains, identifies, and even increases these different harmonic incidences of the synchronic order.

Because the vertice of time connects ultimately to the "Locus of God" -the mutually defining points that become One at the far ends of the toroidal motion of the vertex of time-this accounts for the existence of the *perennial philosophy*, the unitive nature of the most revelatory or ecstatic experiences of the great mystics and seers. It is also the relation of awareness in the Locus of Consciousness at any present moment that allows any random event of low time to trigger telepathic ascent into high time depending on the clarity of and duration of the Awareness Unit (AU). From this point of view, the random quality of events in near space is only relative in relation to the absolute nature of the synchronic order, and merely a function of the third-dimensional biological space orientation. For those accustomed to the actual nature of fourth-dimensional time, which includes and controls the third-dimensional order of near space, the random events constituting low time are actually "signs" that may be read as message bearing signals of the synchronic order of high time.

Brief reflection on this description of vertical time in relation to horizontal space will demonstrate the value of what is referred to as *meditation without an object* such as that cultivated in the Buddhist traditions, as well as affirm the One Divine Source of the purely monotheistic traditions. Without a clear mind of awareness in the present moment there can be no clear seeing of reality, nor can there be the realization that all of what our senses report to us is constructed in our own mind to conform to what we think is a universe outside of us. At the same time, if we can extend the duration of our awareness in the Locus of Consciousness indefinitely in the vertical direction of time, we come upon the ineffable experience that is an emanation of the Locus of God, however or in whatever "language" our senses may later report or define such experiences. Thus we return to the paradox of the definition of space as an infinitely locatable point-this point is none other than our own minds, in which the construct of space is created and dissolved from moment to moment. Such is the subjective description of time from the point of view of the Law of Time, a description that begs each one of us to take absolute responsibility for our own experience. But there is an "objective" description as well, in which time is also defined as inseparable from consciousness.

Implicit in the subjective here-and-now description of the experience of time is the radial matrix of time. At the experiential level it is the now-centered point in the Space-Locus of Consciousness which defines the center of a radial order of time.

Time radiates out from each here-and-now awareness moment. But the Law of Time also supplies an objective description in which the radial nature of time has two points of determining order: the point of infinite lucidity and cosmogenesis, which is the Locus of God; and the receptive locus of the infinitely locatable point, which is space as constructed by each individual consciousness. Intermediate between the two points-Locus of God and Locus of (individual) Consciousness-is the determining radial matrix of fourth-dimensional time, the universal frequency of synchronization.

Mathematically defined as the ratio 13:20 and described by the 13 x 20 matrix of the Harmonic Module, in this mathematical model of fourth-dimensional time we perceive the root of the Law of Time and its basis in the mathematics of the Mayan time science. It is important to understand very clearly that the 13:20 frequency and the mathematical matrix that defines the operations of this frequency of synchronization are the pure mathematics of time totally apart from the metrics and geometries of third-dimensional space. This is a point of profound significance to which we shall later return.

In the model of objective time, the dynamics of the Law of Time are demonstrated as the evolution of time as consciousness. Here the two unresolved issues of Vernadky's description of the biosphere are unified, the answers to which are the determining factors that allow the triggering of the noosphere: The evolution of time is inseparable from consciousness. Conversely, the evolution of consciousness is inseparable from time. Both time and consciousness are nonphysical fourth-dimensional factors coordinating the third dimension. Because consciousness is synchronized with or by time, it can affect physical third-dimensional processes, in whatever way, whether the acting agent is aware of this or not. Consciousness of the Law of Time, however, is only possible through a profound act of self-reflection. Once it has been made conscious, the discovery of the Law of Time affects the entire medium of planetary consciousness, the noosphere, at first imperceptibly, but then building to a great point of climax coinciding with the climax of the biogeochemical combustion. (See plate 4, The Dynamics of Time: Showing the Evolution of Time as Consciousness.)

This dramatic moment of the discovery of the Law of Time is evident in the graphic description of the dynamics of time, showing the evolution of time as consciousness. The dynamic of the universal, cosmically intelligent, and intelligible whole-the *galacticbrain-is* depicted as a circle encompassing the evolutionary spectrum, showing the index of energy-mind progression from the atomic cellular and biological levels to purely mental conditions of being, all of which are functions of the formulation T(E) =Art. From the point of God-Galactic Ordering Dynamic-the circle, or rather, sphere, is divided vertically and horizontally into four quadrants. The movement of the evolution of time as consciousness travels out of the God center, from right to left, and thence describes, in counterclockwise circular motion, the movement of time as consciousness passing through, or rather defining, the four quadrants, until it returns to the Source again as hyperdimensional time, or "pure mind." The left-hand side describes both the primary prelife, lower left-hand quadrant, and the primary post-life, upper left-hand quadrant. The right-hand side describes the secondary reflex life, the lower right-hand quadrant being the purely third-dimensional life (biosphere), and the upper right-hand quadrant being the evolved fourth-dimensional life (noosphere).

The lower half of the sphere beneath the horizontal line running through the God center refers to the Domain of Potentiality, the upper half to the Dominion of Time. The prelife quadrant of the Domain of Potentiality refers to the preconscious quantum dimensional space of pure energy. The lower right-hand quadrant refers to third-dimensional space, the locus of space travel and the realm of the cosmic unconscious. Within the Dominion of Time, the upper right-hand quadrant refers to the fourth-dimension time, the locus of time travel, and the realm of cosmic consciousness. Finally, the upper left-hand quadrant is the realm of hyperdimensional time, or "pure mind." By studying this graphic description, one can grasp the movement of time as consciousness from inorganic preconscious, where duration in time is virtually infinite, to the secondary reflex life, or living organic matter in the third dimension. Here time is unconscious and life builds on the cosmic unconscious as the slowly mutating and evolving biosphere. The cosmic unconscious comes to a climax in the creation of the 12:60 artificial civilization, which is really not true civilization, but rather the technosphere itself. At this point there are moments of consciousness, but only moments. True sustained consciousness is possible only in conditions removed from the technosphere.

In the evolution of time as consciousness, the discovery of the Law of Time is a product of the crisis of consciousness unable to sustain itself in the technospheric medium. That is because this medium is a pure function of artificial 12:60 time. Even though the extensions of the senses and the capacity of technology provide the human within the biosphere a holistic perspective of seeing the whole Earth from space, the consciousness is continuously subordinated to the care and maintenance of the machine (or its lifeblood, money), and so cannot remain in a sustained state of heightened awareness or continuing consciousness. Continuing consciousness is only possible with the discovery of the Law of Time, which the crises of the technosphere and of consciousness have themselves brought about. In this regard, the discovery of the Law of Time is the great point of human self-reflection that defines the advance from an exclusive third-dimensional operation and aberrant timing factor to a purely fourth-dimensional and conscious natural timing factor. It is the difference between the cosmic unconscious and cosmic consciousness, the biosphere and the noosphere-in other words, the very topic of this book.

DE CHARDIN AND THE PLANETIZATION OF CONSCIOUSNESS

The discovery of the Law of Time defines the point between the two stages of the noosphere as described by Pierre Teilhard de Chardin: "The first stage was the elaboration oflower organisms, up to and including man, by the use and irrational combination of elementary sources of energy received or released by the planet. The second stage is the super-evolution of man, individually and collectively, by the use of refined forms of energy scientifically harnessed and applied in the bosom of the Noosphere, thanks to the coordinated efforts of all men working reflectively and unanimously upon themselves. . . In becoming planetised humanity is acquiring new physical powers that will enable it to super-organize matter. And, even more important, is it not possible that by the direct converging of its members it will be able, as though by resonance, to release psychic powers whose existence is still unsuspected?"6 The identification of telepathy with the velocity of time emphatically underscores de Chardin's final point.

On the other side of the discovery of the Law of Time, civilization, having reached its climax as the technosphere, is replaced by an entirely new noospheric order of human organization defined as PAN: Planet Art Network. This is a structural definition of what Teilhard de Chardin refers to as the "planetization of consciousness." This organization (bionoospheric and not institutional) is "planet" because the human organism is self-realized as a planetary organism. It is "art" because functioning again in the natural timing frequency, the human is governed by the Law of Time, T(E) = Art, and hence the reality and the activities of everyday life are defined as art. And it is a "network" because the city, defined by civilization (literally living in cities) and which was a sustaining component of the technosphere, is no longer necessary or viable. Rather, the human becomes redistributed throughout the biosphere-what Teilhard de Chardin refers to as the "radiation of man "--connected by a telepathic network that greatly diminishes the reliance on much of the technology that characterized the technosphere. Hence Planet Art Network: the replacement of civilization with the fourth-dimensional social organization of the human functioning as the telepathically coordinating species of the noosphere. In this condition life evolves toward a hyperorganic stage, attaining a state of superconscious activity and self-reflection where time travel replaces the space travel of the phase known as civilization.

This evolutionary mode, coinciding with the evolving needs of the local star, conduces at some distant point in the future to the subliminal consciousness and the postorganic postlife that is necessitated by stars as they advance to the stage of supernova. According to the Law of Time, the nature of consciousness is actually a function of stellar evolution, planets being but harmonic nodes in greater heliospheric fields of resonance. Since time is inseparable from consciousness, this understanding would seem to affirm Kozyrev's hypothesis that time is generated from the stellar core. "Stars absorb energy from the motion of time," declares Kozyrev, "a star seems to be a time machine."7 *The Mayan Factor* further places the coordinating medium of time as emanating from the core of the galaxy, Hunab Ku, One Giver of Movement and Measure. In the Chilam Balam prophetic tradition, Hunab Ku is also the name given to the final religion of the One God that appears after the religion of the *conquistadors* has finally passed. Seen from the perspective of the Galactic Brain, the crisis through which we are now passing is but the final stage before consciousness bursts into a new radiance of biosolar-telepathic wonder where negative emotions and thought forms no longer have any place to cling or hide. The Law of Time purposively points to this great end.

The whole of the spectrum of the Galactic Brain is mediated by the universal frequency of synchronization, mathematically modeled as the 13:20 matrix, the organizing factor of the synchronic order. The technicalities of this evolutionary spectrum of the Galactic Brain have been defined in *The Dynamics of Time: 260 Postulates* (1996) so that the process of the evolution of time as consciousness may be studied and its coming stages prepared for. The mathematical model of this 13:20 matrix is a description of the fourth-dimensional timing frequency and is both the source of the time vector potentialities and the mathematical tool for mapping the order of synchronicity. Through the discovery of the Law of Time, the 13:20 matrix is realizable through tools that are in accord with its mathematics, and thereby it becomes useful to the everyday consciousness of the human. This too becomes a factor accounting for the noosphere's becoming conscious.

The fourth-dimensional "tools" made accessible through the Law of Time include the Thirteen Moon/28-Day calendar, the Dreamspell, and the Telektonon. They will be described more fully in Chapter 8, "Making the Transition to the New Time." With these tools one can begin to construct a map of the time vector potentialities contained in the synchronic order as it manifests on a day-to-day basis. Once the construct is made conscious and correlated to a point in third-dimensional time, then the body in time-the infinitely locatable point that is space-may be coordinated by the same common structure of fourth-dimensional time. Following this procedure the individual human consciousness begins to participate in the larger unfolding of the noosphere, and slowly but surely is realized as yet another manifestation of the formulation of the Law of Time, T(E) = Art. It is the human being itself that incorporates time, or rather wakes up within itself the order of time. Time is art, while art is actually a function of consciousness expressing itself.

THE MAYAN UNDERSTANDING OF THE LAW OF TIME

While we will have cause to turn our attention later in this book to the "how-to" of fourth-dimensional time, we may consider further the nature and meaning of the mathematics of the Law of Time and their origin with the mysterious civilization of the ancient Maya. The template or matrix of the 13:20 frequency, a 13 x 20 (260) unit structure, was the means for accommodating the base "calendar" used by both the ancient and present day Maya: the *Tzolkin,* or sacred calendar, a set of thirteen numbers and twenty signs that repeat their sequences of permutations every 260 days. While this 13 x 20 matrix accommodates the Tzolkin, it is not confined to being a calendar, as I amply demonstrated in *The Mayan Factor.*The 13:20 frequency is the universal constant of time from whose mathematics the Tzolkin is derived. Based on the permutation table, called the *buk xok* in the Chilam Balam tradition, this matrix is a Harmonic Module, a fractal yardstick of radial time with multiple applications, something analogous to the periodic table of elements, but of the synchronic order.

It is still a profound challenge to all modern Western scientific thought that the Maya would have possessed such mathematics, apparently inherited from the earlier and even more mysterious Olmec people, and that with this mathematics they would have evolved such a profound perception of time. This was most certainly a factor in the destructive attitude taken toward the Maya by the religious zealotry of the conquerors of Mexico. It is also astonishing that the mathematics in which the Harmonic Module is embedded is a vigesimal (20) count, utilizing a zero and a positional ordering system that accounts for larger or greater values that advance by binary exponentiality rather than a decimal sequentiality. But, quite simply, this system exists in this manner because it is the mathematics appropriate to time. This mathematics is a factor to which any disinterested researcher on the topic of time must now become accustomed.

It is by reason of this mathematics and its ability to "download" the 13:20 universal frequency of synchronization that the ancient Maya were able to construct a complex system of "calendars" and an astronomical knowledge that is precise not for its telescopes, of which there were none, but because of the nature of the mathematical system itself. More appropriately described as synchronization devices or synchronometers, the Maya used at least seventeen of these timing instruments simultaneously during the great age of their civilization, A.D. 435-830. This fact in itself tells us that the Maya understood time so radically differently from the perception that has evolved in the techno sphere that we must stand in awe that such a knowledge would have developed by a people whose technological development was so minimal, but whose artistic accomplishment was inseparable from their scientific achievement.

Harmonic Module

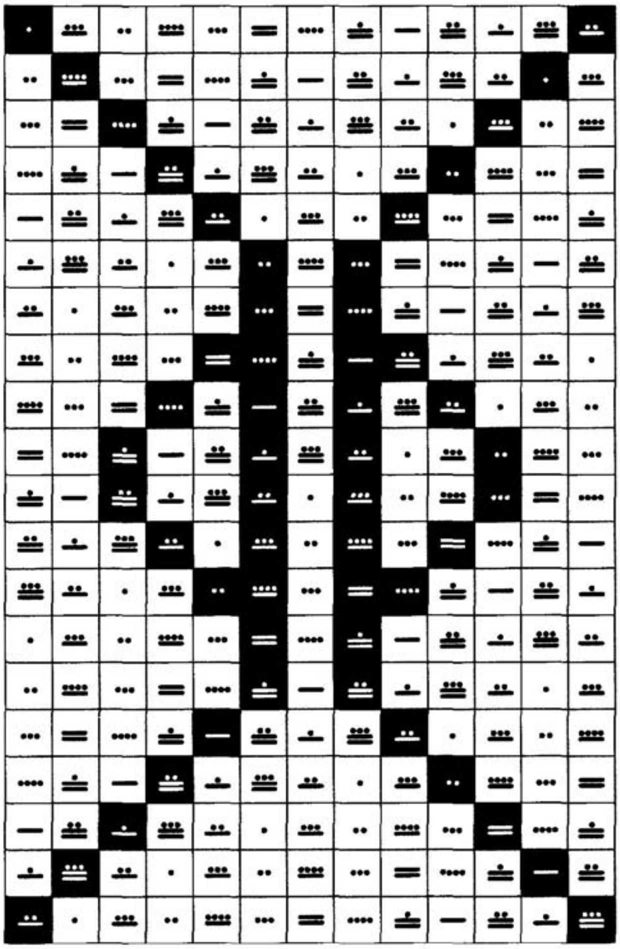

Yes, the Maya most definitely knew that time is the universal factor of synchronization, so their synchronometers must have been for the purpose of attaining a master synchronization with the cosmos. This presupposes such a different state of mind that all the values of the present day must be challenged. The mystery of it is that this mathematics and time science developed in the New World, remote from the saga of civilization in the Old World. The fact that the later Maya practiced ceremonies and blood rituals, the fascination of so many present-day archeologists, should not detract from the understanding that these ceremonial practices stem from a totally different perception of life-and death. Before we judge, we must consider the inescapable truth that nothing in history matches the terror of the different kinds of war and weapons of mass destruction that characterize the fifty-six-year cycle of the technosphere, 1945-2001. Who then is the more barbaric-the Maya, or modern man, *"Homo technosphericus"?*

It is the study of this vigesimal mathematics and its 13:20 Harmonic Module that has led to the discovery of the Law of Time. In this regard the Law of Time is a universal law independent of any culture or civilization, even that of the Maya, much like the law of gravity is also a universal law. However, if it had not been for the Maya, who evolved their understanding of the 13:20 frequency of synchronization to such a high degree, there would never have been a pursuit that led to the discov-

Galactic Notation, Dot-Bar (0–19) Code

Vigesimal count, 1–19, where in first order 1 = 1 and
0 = 20 position, so in next order 1 = 20, in third order 1 = 400, etc.

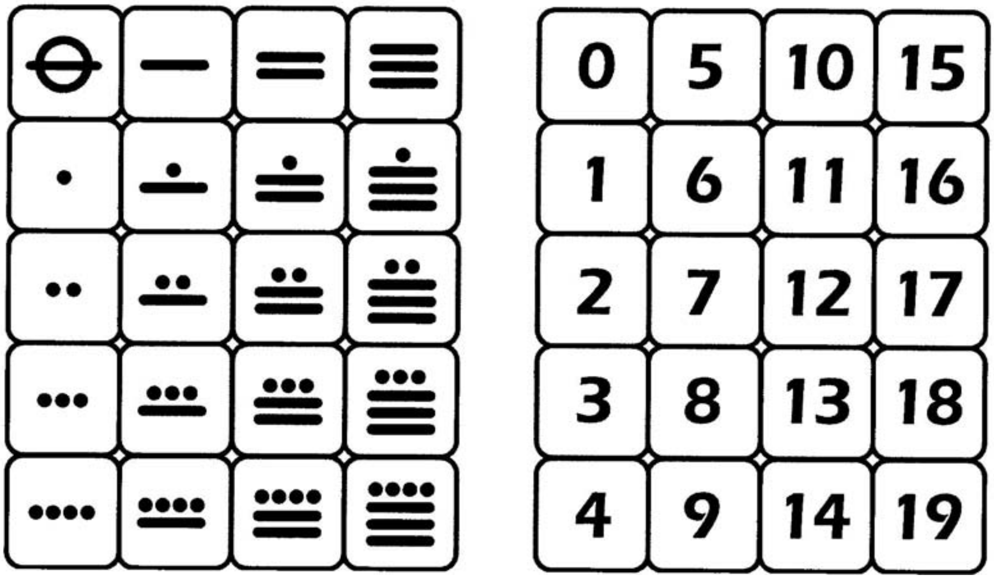

ery of the law underlying their knowledge of time. If with this law we may categorize the time vector potentialities in six stages of consciousness-preconscious, unconscious, and conscious, continuing conscious, superconscious, and subliminal conscious-then we may be certain that the ancient Maya were also familiar with the time science associated with the higher states of consciousness revealed by the Law of Time. With this time science and mathematics of higher consciousness, the Maya were able to bequeath a prophetic tradition mathematically coded in the Harmonic Module, which, among other matters, led to the discovery or making conscious of the Law of Time itself.

Once we appreciate the radically different perception of time and knowledge that is embodied in the Harmonic Module, the mathematical modeling of the 13:20 universal frequency of synchronization, then we may turn to its uses as the fractal radial measure of time applied to many levels of thought and knowledge. For instance, using the 13:20 matrix we may turn to an understanding of the Law of Time in human history, as is necessitated by these reflections. With the Mayan time knowledge and the placement of their civilization in the New World, we have a gauge and an alternative base of knowledge upon which to evolve a planetary perspective and an objective means of critique of the mainstream development of civilization in the Old World. We may then ask the questions: How did we become what we are now, and how was it that our civilization evolved into the technospheric medium?

▼

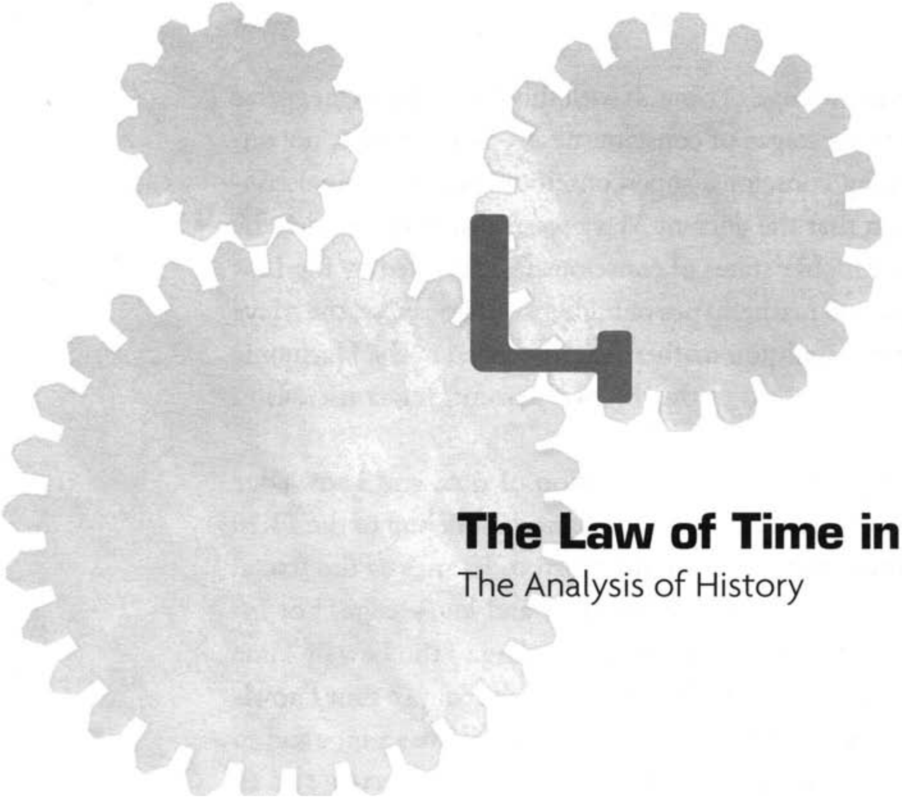

4

The Law of Time in Human Affairs

The Analysis of History

THE KEY to *The Mayan Factor* is the application of the Harmonic Module as a fractal measure that is descriptive of the cycle of thirteen baktuns, the Mayan measure of history. This measure of thirteen baktuns, a cycle extending from 3113 B.C. to A.D. 2012, had been deduced from the stone inscriptions of the Mayan monuments of the Classic period, most of which bear a date within the tenth baktun cycle, A.D. 435-830, or, in the transliterated vigesimal notation of the Mayan Long Count, 9.0.0.0.0-10.0.0.0.0. However, one of the principle discoveries of *The Mayan Factor* demonstrated that the thirteen baktun cycle was also a function of the Harmonic Module and identical with its mathematical mapping as the 13:20 matrix. There are very profound implications to this correlation of the cycle of history having a one-to-one identity with the 13:20 frequency matrix. For this reason the cycle of history is referred to as a *Wave Harmonic* of the galactic order of time. Even though history may appear to be a random, haphazard, and checkered advance of humanity from living in small agricultural communities to the present-day technosphere, this process has actually been an orderly program of the biosphere, the region for the transformation of cosmic energy on Earth.

The same principle of the analysis of the Law of Time applies here. That is, by analogy, the historical events analyzed from a political or economic perspective represent the perception of the random events of near space and low time, while the analysis of the thirteen baktun cycle Wave Harmonic of history represents the perspective of far space and high time. In this Wave Harmonic, the program of history is mathematically defined by 13 baktuns-each baktun consists of 20 katuns-13 baktuns x 20 katuns = 260 katuns = 13 x 20 matrix of the 13:20 frequency of synchronization. In the 13 x 20 matrix of the Harmonic Module each unit equals one kin, the base unit of fractal measure. By fractal equivalence, in the mapping of history one kin = one katun, one katun = 7,200 days; therefore one baktun = 144,000 (7,200 x 20) days. The mathematical perfection of the cycle of history matching the 260-unit model of the harmonic matrix means that history is a factor of synchronization of the higher order of time coordinating the human organism within the biosphere.

By means of the *thirteen baktun Wave Harmonic of history,* the incidence of human Civilizational Advance (CA) may be plotted in a very precise way as a progression going from a very dispersed and scattered state to the point of attaining its exponential climax in the final saga of geochemical combustion, the fifty-six-year cycle of the technosphere. In this regard we must ask: What do we mean by "history" from the Wave Harmonic point of view? As an organism that is a function of the biosphere, history defines that stage when that organism, humankind, enters a process of altering its biomes through increasingly artificial means or extensions of itself that exhibit a compulsive reordering of society into denser and denser clusters of habitations known as cities, hence civilization, "city-life."

In the stage before history, the human had created certain tools and had begun experiments in agriculture and horticulture, but the mental order remained in an aboriginal condition. Unconsciously absorbed as an integral mechanism within the cycles of the biosphere, the human wrapped his understanding of these cycles and of his place in them in the garb of myth and ritual. But with history, something different occurs. The artificial means or extensions of the senses are ordered by a different form of mentality. Those projections of thought and consciousness, though immaterial, affect the material surroundings, and take on an aggressive dynamic. The aggregate effects of these artificial extensions are maintained by a new mental dynamic that is contained and then conditioned by a new kind of macro-organizing thought structure, independent of the actual cycles of the biosphere. What is this new mentality? How can we most precisely define it?

Here are some points to ponder. The Mayan reckoning of the beginning of history is precisely dated to August 13, 3113 B.C., Long Count 13.0.0.0.0. The Hindu reckoning of the beginning of the Kali Yuga is dated to February 19, 3102 B.C. These dates are only eleven years apart. What cosmic event occurred at that time to initiate the Wave Harmonic of history? As we have seen, there is a perfect identity of the

13 Baktun Cycle: Wave Harmonic of History

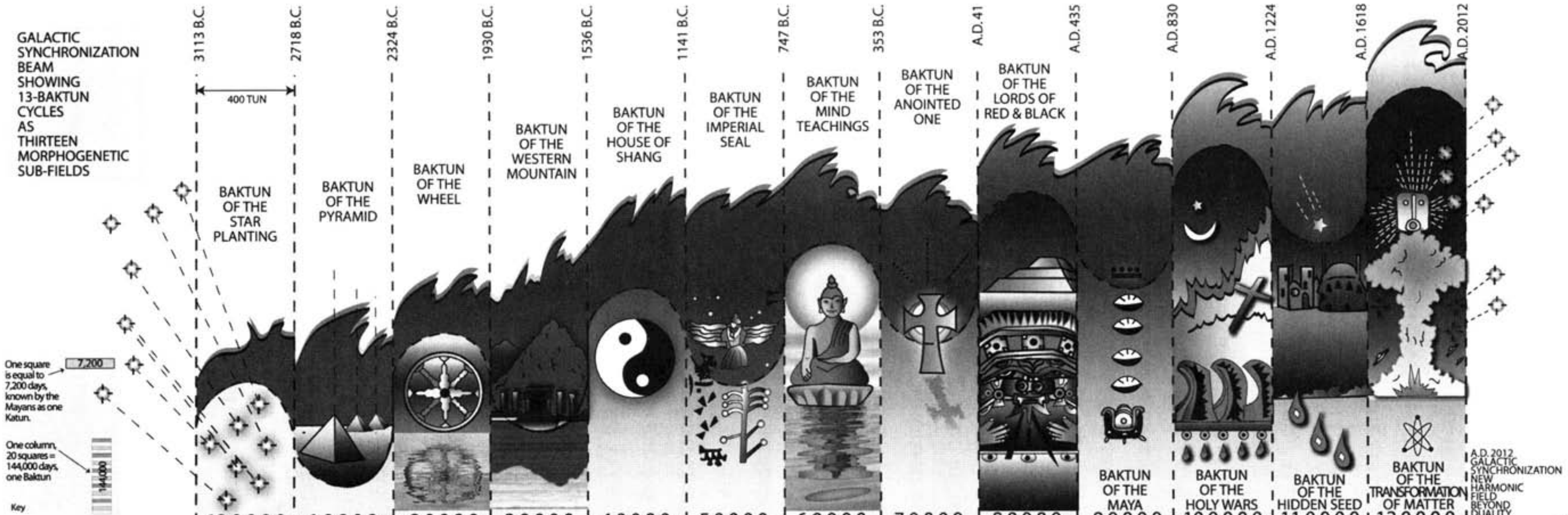
GALACTIC SYNCHRONIZATION BEAM SHOWING 13-BAKTUN CYCLES AS THIRTEEN MORPHOGENETIC SUB-FIELDS
3113 B.C.
2718 B.C.
2324 B.C.
1930 B.C.
1536 B.C.
1141 B.C.
747 B.C.
353 B.C.
A.D. 41
A.D. 435
A.D. 830
A.D. 1224
A.D. 1618
A.D. 2012
400 TUN
BAKTUN OF THE STAR PLANTING
BAKTUN OF THE PYRAMID
BAKTUN OF THE WHEEL
BAKTUN OF THE WESTERN MOUNTAIN
BAKTUN OF THE HOUSE OF SHANG
BAKTUN OF THE IMPERIAL SEAL
BAKTUN OF THE MIND TEACHINGS
BAKTUN OF THE ANOINTED ONE
BAKTUN OF THE LORDS OF RED & BLACK
BAKTUN OF THE MAYA
BAKTUN OF THE HOLY WARS
BAKTUN OF THE HIDDEN SEED
BAKTUN OF THE TRANSFORMATION OF MATTER
A.D. 2012 GALACTIC SYNCHRONIZATION NEW HARMONIC FIELD BEYOND DUALITY
One square is equal to 7,200 days, known by the Mayans as one Katun.
7,200
One column, 20 squares = 144,000 days, one Baktun
144,000
Key

260-KATUN CYCLES
52-UNIT GAIA LIGHT BODY

	13.0.0.0.0	1.0.0.0.0	2.0.0.0.0	3.0.0.0.0	4.0.0.0.0	5.0.0.0.0	6.0.0.0.0	7.0.0.0.0	8.0.0.0.0	9.0.0.0.0	10.0.0.0.0	11.0.0.0.0	12.0.0.0.0	
DRAGON / IMIX	URUK, MENES	ZOSER EGYPT CALENDAR	SARGON	ABRAHAM	SHANG DYNASTY, CHINA	KING WEN	GREEKS	ALEXANDER	ROMAN	FALL OF ROME	BOROBADUR	GOTHIC	DESCARTES	NEPTUNE
WIND / IK	KILNS	GREAT PYRAMID	BABYLON		AKHNATON	CHOU DYNASTY	KUSHITE	CELTS	EMPIRE	BYZANTIUM		EUROPE	SCIENTIFIC	URANUS
NIGHT / AKBAL	UNIFIED SUMER	GIZA		CRETE	HITTITE	CHINA		ASOKA	TEOTIHUACAN	RISE OF	VIKINGS	MONGOLS	MATERIALISM	SATURN
SEED / KAN	EGYPT				EMPIRE	KING DAVID, JERUSALEM	NINEVEH	SPREAD OF BUDDHISM		CLASSIC	HEIAN JAPAN	KUBLAI KHAN	CHING DYNASTY	JUPITER
SERPENT / CHICCHAN	OLD		CHARIOT	EGYPT: QUEEN	THERA EARTHQUAKE		PYTHAGORAS	ZAPOTECS		MAYA	CHIMU	MAYAPAN	NEWTON	A-BELT/MALDEK
WORLD-BRIDGER / CIMI	KINGDOM		WARFARE	HATSHEPSUT	OLMECS	VEDIC	BABYLON REBUILT	CH'IN DYNASTY	NAZCA	BUDDHISM IN JAPAN	XOCHICALCO	AZTEC	BAROQUE MUSIC	MARS
HAND / MANIK				VALLEY OF	CHAVIN	CIVILIZATION		HUANG-TI	GERMANIC	TIKAL	CE ACATL	BLACK PLAGUE	INDUSTRIAL	EARTH
STAR / LAMAT			EGYPT	KINGS		INDIA	LAO TZU	EARLY	INVASIONS	POPE GREGORY	TOPILTZIN QUETZALCOATL	INCA	REVOLUTION	VENUS
MOON / MULUC	HEIROGLYPHS	AKKAD	MIDDLE KINGDOM				PERSIAN	MAYA	MOCHE	MUHAMMED	RISE OF	MOGHUL / INDIA	AMERICAN &	MERCURY
DOG / OC		UR			JADE	HORSE WARFARE	EMPIRE	HAN		PALENQUE	TOLTEC	OTTOMAN TURKS	FRENCH REVOLUTION	MERCURY
MONKEY / CHUEN	CUNIEFORM				KING TUT		BUDDHA, MONTE ALBAN		END OF	PACAL VOTAN	MAYA		ROMANTICISM	VENUS
HUMAN / EB		HARAPPA			RAMSES		CONFUCIUS	DYNASTY	HAN DYNASTY		IFE	MING DYNASTY	EUROPEAN	EARTH
SKYWALKER / BEN		CIVILIZATION			ASSYRIA	JERUSALEM		TEOTIHUACAN	TIAHUANACO	COPAN	ANGKOR WAT		IMPERIALISM	MARS
WIZARD / IX		INDIA	ZIGGURATS				PLATO	RISE OF ROME		HEIROGLYPHIC STAIRWAY	CRUSADES	FALL OF CONSTANTINOPLE	ELECTRICITY	A-BELT/MALDEK
EAGLE / MEN				HORSES			CHUANG TZU			SPREAD OF ISLAM	BUDDHISM IN	PRINTING, CONQUEST	RUSSIAN REVOLUTION	JUPITER
WARRIOR / CIB			BRONZE	CENTRAL ASIA	VEDAS	LA VENTA	ZOROASTER	GREAT WALL	CONSTANTINE		TIBET	OF MEXICO, PERU	WORLD WARS	SATURN
EARTH / CABAN			TECHNOLOGY	HITTITES		ASSYRIAN		OF CHINA	BUDDHISM IN	T'ANG DYNASTY	SUNG DYNASTY	EUROPE EXPLORATION	HIROSHIMA	URANUS
MIRROR / ETZNAB			MENUHOTEP	EGYPTIAN	PHOENICIANS	IRON	BANTU, WEST AFRICA		CHINA	MISSISSIPPI CULTURE, BAGDAD	RUSSIA/KIEV	BENIN	NUCLEAR & SPACE TECHNOLOGY	NEPTUNE
STORM / CAUAC	STONEHENGE	PYRAMID	UR NAMMU, LEGAL CODE	EMPIRE	MYCENEANS	WEAPONS &	PELOPPONESIAN WAR	JULIUS CAESAR	GUPTA DYNASTY, INDIA	PADMASAMBHAVA	ANASAZI	QUEEN ELIZABETH	HARMONIC CONVERGENCE	PLUTO
SUN / AHAU	GILGAMESH	TEXTS	HAMURABI	ARYANS INVADE INDIA		WAR MACHINES	ARISTOTLE	CHRIST		CHARLEMAGNE	ZIMBABWE	GALILEO, KEPLER	EARTH REGENERATION	PLUTO

SOLAR MOVING INHALATION

GALACTIC/LUNAR MOVING EXHALATION

Wave Harmonic of history, as a fractal equivalence, with the Harmonic Module, itself a description of the universal 13:20 frequency of synchronization. This being so, both the calculation of the beginning of history as well as the duration of its entire cycle are functions of the Law of Time. This means that something within the Earth's noosphere occurred that triggered the Wave Harmonic, itself a manifestation of the Law of Time. Also, by most judgments and standards, history began at a very precise place as well-Sumeria, in the Mesopotamian basin of what is now called the Middle East. Why? And what relation is there between the precise event that triggered the Wave Harmonic of history and Sumeria, the site of the first city, Uruk? What event could have established a macro-organizing principle to propel the dynamic of human thought and technology into the trajectory concluded by the technosphere?

THE ORIGINS OF THE 12:60 FREQUENCY

In the recently published book *The History of the World-A 6,000 Year Chronicle of Time,* we find the point that establishes the definite artificiality of a new kind of macro-organizing principle, driving humankind toward the Inevitable Event; here we find the genesis of artificial time: "Ca. 3000 B.C. Sumerians divide day into 24 hours, 60 minutes, 60 seconds and circle into 360 degrees."! While this is a fact we may take for granted because it is the basis of mechanized time in the form of the clock, for precisely that reason it should give us pause. For one thing, this means that the measuring system of the clock was already established at the beginning of history, some 4,500 years before the clock as we know it was actually invented! This is quite an astonishing fact. It means that mechanization was implicit in the first intellectual act of history. Why or how would it have occurred at the precise point of the beginning of history in Sumeria?

The Law of Time defines this mechanistic organizing principle as the basis of the 12:60 artificial timing frequency (irregular twelve-month calendar combined with mechanical sixty-minute hour). Therefore, it was at this point at the beginning of history that artificial time and its progeny, mechanization, were actually registered as a thought form and thus became embedded in the as-yet-unconscious noosphere. Or, was it already embedded in the noosphere to be triggered at this precise point-and if so, why and how? The point is that history, as it has evolved as the mainstream of civilization from Sumeria to the World Trade Center, has always been driven by the organizing principle of artificial time. It is artificial time that separates history from prehistory-and, as we shall see, from post-history as well. It was due to the

The Error in Time

The measure of a two-dimensional plane in space has nothing to do with the experience or the nature of time.

"TIME WAS COMPRESSED INTO A FLAT CIRCLE..."
Dreamspell "Genesis"

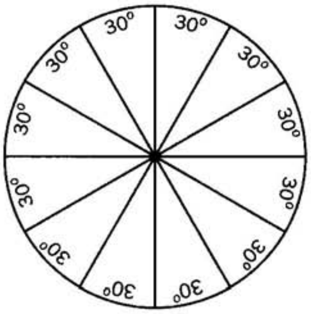

Sumeria: Ca 3000 B.C.
Day divided into 24 hours,
60 minutes, 60 seconds
based on 360 degree circle—
two-dimensional plane in space—
divided into twelve parts

Earth's solar orbit = 365 days
360 degrees ≠ 365 days

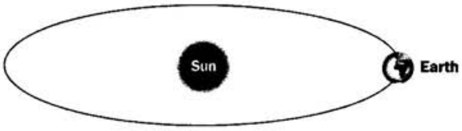

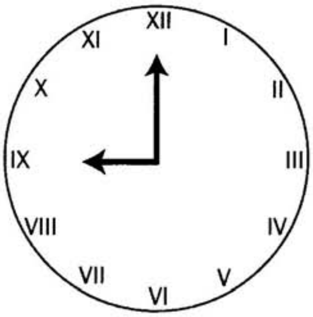

Europe: 800-1600 A.D.
Clock—two-dimensional plane
in space—divided into twelve
parts each 30 degree of arc =
1 hour = 60 minutes
each minute = 60 seconds
A.M. = 12 hours
P.M. = 12 hours
= 24 hours

I - 31	II-28/29	III - 31
IV - 30	V - 31	VI - 30
VII - 31	VIII - 31	IX - 30
X - 31	XI - 30	XII - 31

Rome (Vatican): 44 B.C. - 1582 A.D.
Twelve-month Julian/Gregorian
calendar derived from twelve part
division of circle; irregular measure
due to fact that 365 is not divisible
by twelve.

genesis of artificial time at the beginning of history and its emplacement within the planetary noosphere that also generated the thirteen baktun Wave Harmonic of history. For the very purpose of the thirteen baktun Wave Harmonic of history is to encompass and coordinate the cycle of artificial time. Why?

The Wave Harmonic is a function of the 13:20 synchronization frequency. It is a function of cosmic time. In the higher knowledge of the Mayan understanding, it provided the perfect measure of the cycle of artificial time. From the perspective of the actual noospheric timing, which operates according to the natural 13:20 cosmic timing frequency, history is none other than a manifestation of artificial time. Civilization becomes organized by the principle of artificial time. But because it is artificial, it cannot endure. If the mathematical basis of mechanized time, and hence of mechanization, was already established at the beginning of history, then the end of history is the end of artificial, mechanized 12:60 time. And in the end, the Law of

Time codes and encompasses the whole of the cycle of artificial time through the thirteen baktun Wave Harmonic of history. What is important to grasp is that the entire cycle of artificial time was *already perceived* by the Mayans as a phenomenon that could not endure more than the length of the cycle of thirteen baktuns.

It is necessary for the reader to understand how this perspective was discovered. It was the result of a phenomenological experiment of living according to the synchronic cycles of the Mayan time science. After the death of my son in 1987, my wife and I began in earnest to follow the cycles encoded in the Mayan calendar: 13-day, 20-day, 52-day and 260-day cycles. Our life changed dramatically. At the same time, I endeavored to define the mathematical basis of these cycles. Once I resigned my academic position in early 1989, we were swept by a wave of time that took us from the Rocky Mountains to Southern California to the Hawaiian Islands in the middle of the Pacific Ocean. There, in relative isolation from the rest of the world, we plunged into our experiment. Before we had truly settled in, the time wave took us to Switzerland, where on December 10, 1989, in the Museum of Time in Geneva, we were able to put our phenomenological experiment to the test-although we did not consciously realize at the time that we were doing so.

There in the Museum of Time, essentially surrounded by examples of the history of the clock and mechanization, and because our perceptions had been altered by our living so completely by the various timing cycles encoded in the Mayan calendar system, we were able to spontaneously perceive the existence of two timing frequencies. The natural frequency-the one we had been living-we understood instinctively to be the 13:20 timing frequency. The other frequency-the one enshrined in this Museum of Time-we understood to be the artificial 12:60 timing frequency. This realization of artificial time was rooted in the evident proof that the clock is based on the twelve-part division of a flat circle, a spatial plane that clearly has nothing to do with the dynamics of time. It is a manifest error, but one that humanity has taken for granted and endowed with all of the characteristics of "truth."

We also immediately saw that the error of mistaking the division of a two-dimensional plane in space for a measure of time was transmitted not only through the clock, but through the twelve-month Gregorian calendar as well. With brief reflection upon what a gross error this was-substituting a twelve-part division of space as a measure of time, and then elevating this mistake as the truth of the measure of time-we understood how it could have actually led humanity away from the natural time of the universe. In making this observation, we immediately and intuitively grasped the necessary solution: replace the erroneously measured twelve-month Gregorian calendar with the perfect measure of the Thirteen Moon/28-Day calendar.

Our research into the mathematical codes behind the Mayan calendar now progressed rapidly, but our attention also turned to the question: Whence came the thought form of 24 hours, 60 minutes, and 60 seconds as a measure of a circle of 360 degrees? It was clear to us that this division of time based on the division of a circle in space occurred at what is the beginning of history, a fact later confirmed for us in the already cited *History of the World* (1997), which states that this division of 24 hours, 60 minutes, 60 seconds and the circle into 360 degrees was devised "Ca. 3000 B.C."

Once we had finished working out the mathematical codes, we turned to a consideration of the actual implications of the error in time. It was now also clear why the Mayan time science was so superior. It was based on the reality of time as a factor of synchronization. By contrast, the concept of time as it had developed in the mainstream of civilization was limited, linear, and totally third-dimensional, a fact that confirmed Vernadsky's analysis, as well, that our notions of time are a function of the metrics of space. By 1991, the experience in the Museum of Time had driven us into

What Difference Can a Calendar Make?

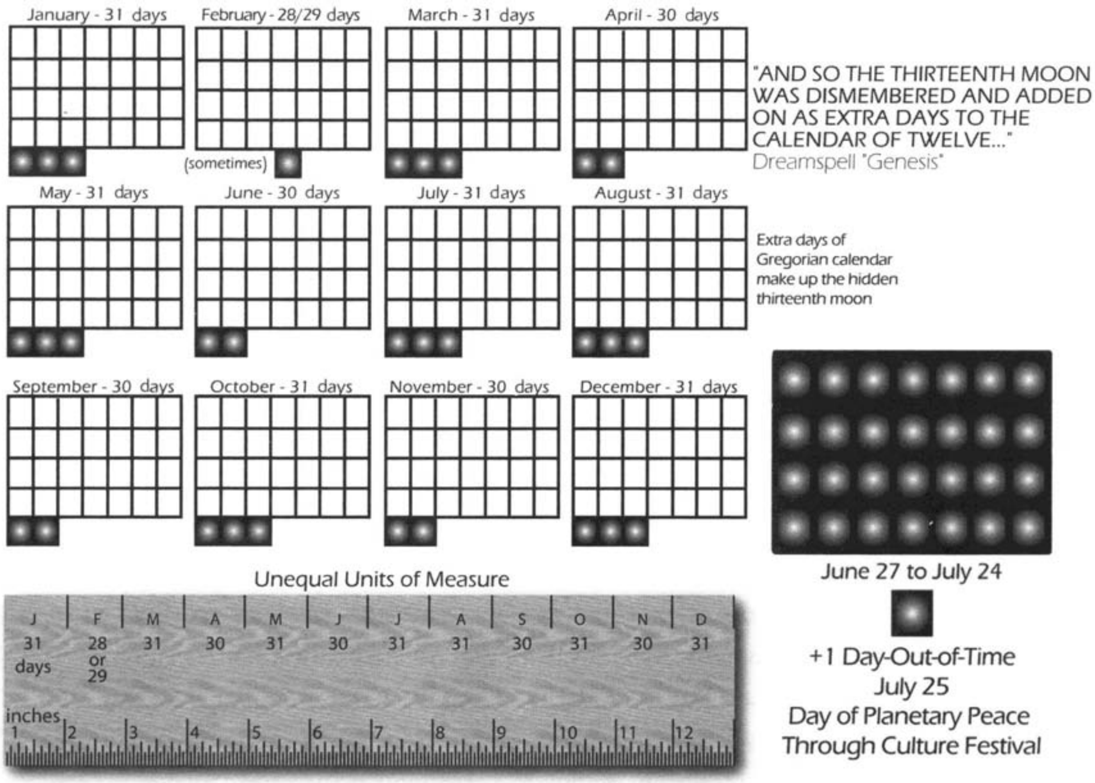
January - 31 days
February - 28/29 days
March - 31 days
April - 30 days
(sometimes)
May - 31 days
June - 30 days
July - 31 days
August - 31 days
September - 30 days
October - 31 days
November - 30 days
December - 31 days
Unequal Units of Measure
J F M A M J J A S O N D
31 days
28 or 29
31
30
31
30
31
31
30
31
30
31
inches
1 2 3 4 5 6 7 8 9 10 11 12

"AND SO THE THIRTEENTH MOON WAS DISMEMBERED AND ADDED ON AS EXTRA DAYS TO THE CALENDAR OF TWELVE..."
Dreamspell "Genesis"

Extra days of Gregorian calendar make up the hidden thirteenth moon

June 27 to July 24

+1 Day-Out-of-Time
July 25
Day of Planetary Peace
Through Culture Festival

Gregorian Calendar and the Hidden Thirteenth Moon creating a full-blown psychomythic analysis and description of this moment of realization concerning the nature of time. The complete structure of the mathematical codes of time, as well as the psychomythic description of the causes and effects of the error in time, is known as the Dreamspell, which is defined as follows:

"Dreamspell: function of fifth-dimensional galactic-solar planetary regulation; 26,000-year renewable cycle and planetary castle genesis; any consensus reality; in disregard of interdimensional reality creates entropic spell of history; cure for loss of galactic memory."2 The *planetary castle genesis* refers to the *five castles of time* (each 5,200 years in length), divided into three *geneses-the* Dragon, the Monkey, and the Moon-that constitute the entire fractal panorama of the 26,000-year cycle of history that ends in 2012, and is meant to be renewed in 2013 (an obvious synchronic reference to the 13:20 timing frequency). In the Dreamspell analysis, the beginning of history occurs almost 5,200 years ago between the Monkey and the Moon geneses, and is marked by the imposition of the false 12:60 timing ratio.

In light of this analysis and definition it is worth quoting certain passages from the Dreamspell text that describe this precise moment at the beginning of history and that the Law of Time defines as the root error of civilization. First we quote from the "Dreamspell Genesis," and then from the descriptions of the "Dreamspell Journey and Mission of Timeship Earth 2013." The purpose of quoting these passages here is to establish a psychomythic context to explain the profound nature of what we refer to as the error in time compounded over 5,000 years of history:

> Imposters, male priests and warriors usurped the power of the thirteen moons. They hid the thirteen among themselves and attempted to banish all memory of the matrix dreamspell of magic. . . The planetary kin had imposed upon them the diminishing power ratio of 12:60. No longer the magic of thirteen moons but a twelve month calendar of uncertain meaning. No longer the timeless gyre of magical flight, but a sixty-minute hour to earn one's bread.
>
> Time was compressed into a flat circle. . .
>
> The 13:20 ratio is the base operating ratio of timeship Earth. . .
>
> In that vital moment during the magic flight. . . Toward the moon genesis, another spell was cast. Instead of the 13:260 ratio already stored deep within the crystal core of timeship Earth, the planetary kin received the 12:60 ratio, the dark dreamspell of history. . . And so the thirteenth moon was dismembered and added on as extra days to the calendar of twelve. . .

> The absorption of the thirteen by the twelve was called an improvement by the male imposter priests hypnotized by the combined powers of Jupiter and Saturn. It takes twelve years for Jupiter to orbit once around the sun. It takes sixty years for Jupiter and Saturn to be conjunct.
>
> Twelve times five equals sixty, one sixth of the flat three hundred sixty degree circle. The difference between twelve and five is seven. While the priests of the calendar of ancient Babylon banished the power of the thirteen moons, they replaced that power with the power of seven.
>
> While the twelve diminished the power of thirteen by one, the sixty raised the power of twenty by three. A seeming increase in power occurred, but an increase only on the flat plane of the third-dimensional time to which the planetary kin were now consigned.
>
> With the banishment of the thirteen moons, the fourth-dimensional time magic. . . became the property of the priest class. Setting up religions and governments to maintain the kin within the third dimension, the priests were assured that anyone having fourth-dimensional experiences either be appropriated or destroyed.J

So much for the psychomythic impression of the genesis of artificial time, the flat time of an exclusive third-dimensional reality called history. A few further quotations from the Dreamspell emphasize both the interplanetary and mathematical root of false time:

> At -3187 Dreamspell years the time bandits on the sixth and seventh orbits pulsed their beam. The 12:60 frequency of the memory virus took immediate effect on the third-dimensional space suits. . . The Dreamspell of history was cast.
>
> Imposter priests on Earth substitute the 12:60 ratio. . . Instead of the magic of following thirteen moons, a non-circulating twelve-month calendar is substituted; instead of the beauty and power of magical flight, the sixty-minute hour. The result is disastrous.
>
> The twelve-month calendar is the consequence of the 12:60 ratio imposed at -3187 Dreamspell years (3113 B.C.,Julian/Gregorian). With no basis in or capacity for measuring galactic time, the twelve-month calendar is actually a third-dimensional prison keeping the four root races separate and at war with each other.4

These descriptive explanations of the error in time define such a profoundly radical point of view that we must flesh out the implications and description of the course of artificial time according to the analysis of the Law of Time. From the outset of civilization in Sumeria, artificial time captures the human mentality in the order of low time and near space (third dimension), separating it from high time and far space (fourth dimension). The registration of artificial time within the unconscious structure of the noosphere, which is intrinsically governed by the universal 13:20 frequency of synchronization, creates a conflict and a pressure in the noosphere, a time warp or taint that affects the entire biosphere.

Civilization-living in cities-increasingly becomes the human norm within the biosphere. Where there had been no civilization, at a few key places around the planet civilization now arises-a process that goes from the crystalline structures at the origin to the later phases of imperialism and empire building. This cycle is repeated again and again. Even in the New World where the 13:20 natural timing frequency prevailed, the later stages of civilization succumbed to the effects of the 12:60 already seeded in the noosphere. Ultimately, all humanity and the biosphere were to be governed by the artificial timing frequency that ends with the construction of the technosphere, the final and supreme expression of artificial time. Indeed, the technosphere is the absolute conclusion of what can rightly be called an artificial time warp in the biosphere.

Concerning the mathematical basis of the 12:60 frequency, V. A. Ponko of the Russian Academy of Sciences, Novosibirsk, has conducted extensive research and wave analysis on the mathematical properties of the cycles of history. He has concluded that as a mental coordinate, the number 12 is a strict construct of the angles of space. Since it does not correspond to the sinusoidal curvature of the cycles of time and the plotting of all organic phenomena, when used as a time factor the number 12 offers no "protection" from bombardment by any number of cosmic forces. The mathematical plotting of the 13:28 frequency that characterizes the Thirteen Moon calendar, on the other hand, is in alignment with the sinusoidal curve properties of organic phenomena, and therefore functions as an "umbrella" maintaining the organism in resonance with the biosphere.s

Here let us make a more precise definition of timing frequencies and the nature of calendars as the macro-organizing programming systems of the human in the biosphere. By defining time as the universal frequency of synchronization, we are saying that synchronization is the fundamental program to which everything adheres and which makes everything perceptibly coherent, hence synchronic order. Therefore, the natural timing frequency that is a universal phenomenon both regulates the phenomenal order of the cosmos, and in the human is absorbed as a mental frequency that unconsciously synchronizes the mind and senses with the natural order. The result is the normal sensation and experience of the harmony of reality. This defines both a fundamentally prehistoric perception of reality as well as the spiritual experience that is common when we are confronted with the awesomeness of nature-usually far away, it should be noted, from city life. This universal factor of synchronization we have identified as the mathematically precise ratio 13:20, a ratio constant embedded in the structure and order of the calendrical and astronomical system of the ancient Maya.

If we understand that this frequency of synchronization is a genuinely universal constant, then we may comprehend how not only the entire biosphere, but solar system and galaxy are also governed by the same frequency. This is how the Kozyrev experiments could show that there is instantaneous transmission of information throughout the universe, which in its entirety is a function of the formulation *velocity of time is instantaneously infinite.* Synchronization and infinite instantaneity of transmission are mutually defining-everything is always instantaneously synchronized. We can further say that synchronization "syntropizes" while dissynchronization "entropizes." There is no lag time in synchronization.

This being the case, we can then begin to understand how the acceptance and institutionalization of the artificial 24-hour, 60-minute, 60-second division of the 360 degree circle as the basis for defining time could be such a singular and ultimately disastrous deviation from the universal frequency of synchronization. Thus when we speak of the artificial 12:60 timing frequency, we are referring to the mental effect of the feedback program of the use of artificial and erroneous instruments of measure. Rather than allowing the mind to remain in the natural harmony of synchronization with nature, the artificial 12:60 standard and its instruments establish a feedback effect in the mind that accustoms it to the deviation from nature. This habituation results in the creation of an unconscious timing frequency that is accepted by the consensus reality as the actual norm. Separation of human order from the biosphere becomes accepted and acceptable, making civilization not only possible, but setting it up, as it were, as a parallel rival to "nature." In fact, it is the unconscious mental frequency of 12:60 artificial time that coordinates and maintains civilization as a belief system.

When we speak of the instruments of measure of artificial time, we are specifically referring to the Gregorian calendar as the macro-organizing principle and the watch or mechanical clock as the micro-organizing principle. The macro-organizing principle organizes our life into days, weeks, months, and years. The micro-organizing principle organizes our life into seconds, minutes, and hours. These two instruments have their common root in the error in time that occurred 5,000 years ago at the beginning of history. Try to think of life without these instruments and you will quickly realize how second nature they are-and yet there was a time when such instruments and their terminology did not exist. Before turning our attention to an analysis of these two instruments and their effects on human consciousness, we must view the entire cycle of the Wave Harmonic of history as a function of the macro-organizing principles of time that the human in the biosphere adopted for its uses and that, in turn, condition at a most profound level the worldview of the culture or civilization that uses it.

HISTORICAL CALENDARS AS INSTRUMENTS OF ASTRONOMICAL MEASURE

It may be rightly asked, What does a calendar measure? At a minimum, a calendar is or should be the measure of the Earth in its rotation around its local star, the sun. This rotation takes 365 days. The Earth's satellite, the moon, is the obvious synchronizing factor between the Earth and the sun. But here we encounter a difficulty. The moon can be measured by various cycles. The synodical cycle, from new moon to new moon, is 29.5 days. The sidereal cycle, the measure taken from where the moon appears at the same place in the sky, is just over 27 days. There is an apsidal cycle, however, taken from the measure of when the moon's axis is tilted farthest from the Earth, which is 28 days. This variance of measures is due to the fact that until the rise of rocketry, we could not really see the moon from space. In actuality the moon rotates around the Earth thirteen times during the time it takes the Earth to orbit the sun once.

If one were objectively seeking the proper measure of the Earth's orbit using the moon as the measuring device, we would therefore construct an instrument for time reckoning that consists of an even harmonic measure of thirteen cycles of twenty-eight days each. The 28-day cycle conforms, of course, to the female menstruation cycle, an obvious human biological cycle, and it also divides perfectly into four 7-day subcycles or weeks. Also, 13 x 28 = 364, the same number of days as thirteen moons multiplied by four 7-day weeks (a total of fifty-two weeks), with the 365th day being a free day, or Day Out of Time, no day of the week or month at all. We shall deal with the perceived extra quarter-day later in our discussion.

The point is this: If time is a factor of synchronization, then a calendar-as an instrument for measuring the count of time-should also be constructed so that it maximizes the factor of synchronization. This can only be done if the instrument of time reckoning is constructed on harmonic principles in conformity with the objective data of nature. Such an instrument would also conform to the orderly processes of the biosphere taken as a whole, and be a great assistance in bringing into consciousness the cosmic programs that organize the biosphere and noosphere. The Thirteen Moon/28-Day calendar is the perfect instrument for such a simple registration of harmony and synchronization. It should also be noted that in its mathematics the Thirteen Moon/28-Day calendar is a function of the 13:20 timing frequency. This in itself establishes the Thirteen Moon/28-Day calendar as a perfect instrument of the Law of Time and as a function of the synchronic order of reality.

Since the process of civilization represents a break from a long cycle generally undistinguished by the kind of perturbations that characterize history, the analysis of the Law of Time assumes that in prehistory humans lived in relative harmony, and therefore that the humans tended to use harmonic instruments of time reckoning. The humans living in greater harmony with the biospheric cycles would be disposed to be in greater unconscious or aboriginal attunement with the noosphere. Since the noosphere is regulated by the synchronic order of the universal 13:20 timing frequency, the humans would be more naturally disposed to devise, adopt, or intuitively utilize the time measure in greatest accord with this instinctual 13:20 noospheric timing ratio.

This instrument of time measure would, of course, be the Thirteen Moon/28-Day calendar. It should also be noted that this calendar is a perfect solar-lunar measure; that is, it uses the even, regular lunar cycle of 28 days as the standard of measuring the 365-day solar cycle of the Earth. The balance of solar and lunar may also be taken as

Moon Circling Earth Thirteen Times per Year

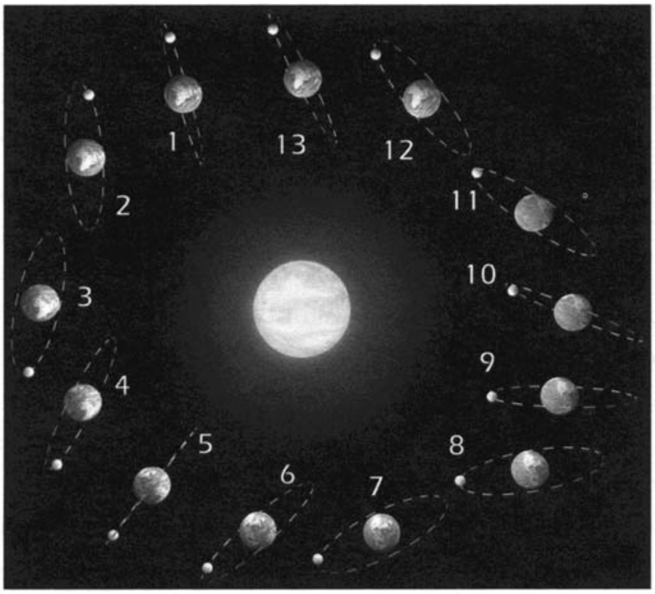

Did you know the moon goes around the Earth thirteen times a year?

SEE: HTTP://WWW.TORTUGA.COM/WWW/ANIMATIONS/ANIM_MOON.HTM symbolic of human psychological characteristics and their balance or imbalance. A solar-lunar calendar will represent and reflect a balance of the solar-lunar, masculine and feminine qualities within the human being. It is important to bear in mind these considerations of calendars and their effect on shaping the human historical consciousness and its psychological makeup. From this perspective, history is, in fact, a process of falling away from the perfection of harmonic standards.

Evidence exists that the Thirteen Moon/28-Day measure was widely known in prehistory. We find evidence of its use as a prehistoric synchronic measure in the remote past of China, in Polynesia, and scattered across late neolithic Europe and the Middle East. Among the Maya it was known as the *Tun Uc,* literally "moon count" or count of seven, while the living tradition of the Thirteen Moon/28-Day calendar is still continued in South America and in the British Isles, where it is known as the Druid Tree calendar. While the Druid calendar has no year count attached to it that would definitively testify to its prehistoric ancestry, the South American calendar does. Known as the *Pachacuti,* the South American Thirteen Moon/28-Day calendar is currently in the year 5509, which places its origins at 3308 B.C., toward the middle of the fourth millennium before Christ, and several hundred years before the beginning of the thirteen baktun Wave Harmonic of history.

If we understand that calendars are genuine time-measuring devices meant to synchronize us with the cosmic order within the biosphere, and that calendars are thereby programming devices, we may then say that a calendar of perfect harmony can have no history. In other words, by its harmony such a calendar is always in tune with the cosmos, which is beyond history. History can only be a function of a disharmonic programming, not in sync or even at odds with the laws of natural time.

From the point of view of timing devices, there are two factors in the establishment of history. The first is the 12:60 program established 5,000 years ago with the division of the day into 24 hours, 60 minutes, 60 seconds, based on the 360-degree division of the circle. The other factor is the use of a purely lunar calendar that is difficult to reconcile with the actual measure of the solar year. The purpose of the 12:60 measure was to establish a timing standard that was actually a pseudosolar calendar to assert male intellectual dominance. The Babylonian (and later, Egyptian) calendar consisted of twelve 30-day cycles = 360 days = 360 degrees, plus an extra 5-day cycle. The 30-day measure is not actually a natural one, being a half day more than the synodic lunar measure, but one that conforms to the hexagisimal (6-based) mathematics of the circle. For this reason, this instrument of time reckoning is defined as *pseudosolar*. Its ultimate descendent, as we shall see, is the twelve-month Julian/Gregorian system.

By contrast, the Maya also had a 365-day solar calendar, the *Haab.* Its measure, however, is that of eighteen 20-day cycles (18 x 20 = 360) plus the 5-day cycle *(uayeb),* the 20-day cycle *(vinal)* being a mathematical function of the 13:20 frequency. In addition, the Maya had the measure of the *tun,* the 360-day count of which is precisely the same as that of the degrees of the circle. It was the tun that the Maya used to measure the thirteen baktun count, which actually consists of5,200 tun (5,125 solar years). The purpose of this measure was to provide an exact harmonic standard for the measure of history that, we may recall, is the duration of artificial time. It is very relevant that the harmonic standard of the thirteen baktun Wave Harmonic of history is the 360-day tun, the actual temporal equivalent of the degrees of a circle in space. The precise measurement of the cycle of artificial time, which turned out to be anything but harmonic, could yet be measured with a harmonic unit that reflects both the circle *and* the mathematical perfection of the Law of Time. How better to coordinate the entire cycle of artificial time than by the tun, a unit of measure that reflects the 360-degree circle, the basis of the original error in artificial time? As we will later observe, the fact that the Law of Time coordinates even the cycles of artificial time is also evident in the analysis of the timing of the fatal flaw of the Twin Towers apocalypse.

While the Babylonians began to apply a hexagisimal time-reckoning device based on spatial metrics of the circle, another factor became predominant in the Old World: the rise of the synodical lunar calendar as the exclusive measure of time. The synodicallunar calendar is purely lunar in that it is not evidently or perfectly harmonized with the 365-day measure of the solar cycle-a fact of great consequence for the development of mainstream civilization. The Babylonian calendar, with its abrogation of the thirteen perfect 28-day months, established the pseudosolar twelve-month measure. This twelve-month measure conformed to the lunar measure of twelve synodic lunations per lunar year. However, twelve lunations are only 354 days, eleven days short of the solar year. This required an intricate set of calculations to make the lunar calendar keep up with the measure of the Earth's solar orbit. And in reality, although we tend to think of the moon as feminine, the knowledge of the synodic measure, the cycle from new moon to new moon, became the exclusive province of a male priest class that used the lunar calendars to capture and control the feminine principle, and to oppress women in general. Whereas the Babylonian and Egyptian solar calendars fell into disuse over time, the lunar calendar persisted.

Since there is only one moon orbiting Earth, it should be kept in mind that all lunar calendars are essentially the same. One and the same moon in all its phases is apparent to all humans on Earth. So when we speak of different lunar calendars-most notably the Hebrew, the Islamic, and the Chinese-the measure is basically the same, from new moon to new moon, but New Year's Day falls differently in each of them, thus accounting in part for the various calendrical traditions. Within each of these lunar calendars, a different set of holidays is programmed. It is the programming of the different holidays and festivals of the calendars, along with the dating of their new year, that makes each lunar calendar tradition distinct and that also demonstrates how a calendar is a programming device. For instance, if one were to take out the different holidays programmed into the Hebrew lunar calendar, there really would be no such thing as Judaism. This demonstrates that the cultural conditioning of a people is totally dependent on the programming of the calendar they use. This also shows how a calendar is a feedback mechanism that maintains a group of people within the confines of its own self-established belief system.

Of the three lunar calendars we have mentioned, the oldest count of years is accorded to the Hebrew calendar. This year Rosh Hashanah, or Jewish New Year, occurred on the new moon, September 17, 2001, and marked the beginning of year 5762. This places the origin date for the Hebrew calendar in the year 3761 B.C., some 453 years prior to the commencement of the Pachacuti count. The Hebrew calendar (like the Chinese calendar, which we will discuss below) makes up for the slippage of eleven days per solar year by intercalating a thirteenth lunation cycle approximately every three years, or exactly seven such thirteenth moons every nineteen years. These numbers too-13, 7, and 19-are key harmonic factors of the synchronic order of time. This demonstrates that, at the very least, the synodical lunation calendars are calibrated in their larger cycles by the Law of Time. But the effect of being a purely lunar measure, at variance with the actual 365-day measure of the solar year, creates an interesting and one-sided approach. There is no question that the lunar calendar civilizations are the tool of powerfully patriarchal societies, the tables of the moons and the years being the possession of a dominant male priest class.

The Chinese lunar calendar differs from the Hebrew in one important distinction: from the most ancient times it has been embedded in a system known as the *five elements,* and is coordinated with a very elaborate mathematical system that bears some resemblance to the 60-60, 24-360 basis of the SumerianiBabylonian 12:60 frequency. This synchronizing system accounts for the Chinese cycles of years that are counted by the twelve zodiacal animals in combination with the five elements: earth, fire, wood, water, and metal, thus giving rise to 60 (12 x 5) year cycles. Three of these sixty-year cycles create a "group" or cycle of 180 years. We are now (in 2001) in the eighteenth year of the twenty-seventh group, or year 4,698 of the Chinese lunar calendar. So embedded is Chinese culture and civilization in this calendar, with its elaborate system, that to take away this calendar would virtually eliminate

Chinese culture altogether. Perhaps more than any other factor, this highly involved calendar also accounts for the longevity and tenacity of Chinese civilization as one continuous tradition for the duration of its history, which according to the reckoning of its calendar began in 2697 B.C., a date, interestingly enough, that coincides with the dedication of the Great Pyramid of Egypt.

Despite its marvelous system, the Chinese lunar calendar is still just that: a lunar calendar. The problem with this, from the point of view of the Law of Time, is that when it is not regulated by a proper solar count or measure, the lunar calendar leads people into one-sided developments of one kind or another, if only to the extent to which these calendars foster such profoundly patriarchal societies, of which the Chinese is certainly no exception. The influence and use of the Chinese lunar calendar system extends into southeast Asia, Tibet, Mongolia, Korea, and Japan. As a field of thought within the noosphere, the Chinese system is the equal in power base to the Babylonian and Indian systems, at least up to the time of the dominance of the Gregorian. Even today, the Chinese New Year exerts a powerful attraction worldwide.

While the Chinese and Hebrew lunar calendar systems maintain a powerful conservative strain of human society, the intercalation of the thirteenth moon, seven times every nineteen years, creates a kind of circulation within the system. Thirteen is the number of circulation, whereas twelve is that of a static, non-circulating spatial order. The taboo on the number thirteen-epitomized by the superstition about Friday the thirteenth-must certainly be related to the suppression of the solar-lunar Thirteen Moon/28-Day calendars in virtually all historical societies. This irrational repulsion of the number 13 serves psychologically in defense of the "rationality" of the number 12, a number that does not circulate time and is at the root of the linear conception of time. The illogical and irrational decisions that are made and then institutionalized into the human social fabric must at one point be discounted. Any illusion or error pursued for too long will always end in disaster.

The Islamic lunar calendar, used by some 1.3 billion humans, is distinguished by two factors:

1) It is based on a known historical incident, the *Hegira* or flight of the prophet Muhammad from Mecca to Medina, dated to the new moon or Muharram 1, Julian, July 16, 622 (Gregorian, July 26).

2) It is a pure twelve-month lunar calendar running on a cycle of pure lunar years of 354 days, with no intercalary lunations, and in this regard does not annually harmonize with the 365-day solar cycle.

For this reason, this year 2001 saw the beginning oflslamic lunar year A.H. *(M-*ter Hegira) 1422. But if we are to count the solar years since the Hegira, correlated to Gregorian July 26, then it is solar year A.H. 1379. We see here a discrepancy of some some forty-three years between the solar count and the lunar count. The lunar count continuously retrocedes so that it takes thirty-two years for one Muharram to occur at the same place in the cycle of the solar year. In essence, the Islamic lunar calendar creates a hermetic, retrocessing time bubble with no circulation of the power of the thirteenth moon, and that accounts for the highly conservative and patriarchal nature of much oflslamic society. Again, while there may be a virtue to maintaining a strict lunar calendar count, if it is not correlated to a harmonic solar or solar-lunar standard, the effects will be precisely of the nature that we see occurring in today's world, a point to which we shall later return.

Finally, in this brief consideration of calendars as the principle macro-organizing system for the programming of human society, a word should be said about the system of calendars in the predominant part of the Indian subcontinent. Long before they adopted the Babylonian solar zodiac and hexagisimal spatial principles for reckoning time, the Hindus also had a type of 28-day count, but one based on the course of the moon in relation to a set of stars called Nakhsastras; this created a kind of 28-part lunar zodiac. In this system, however, the Hindus divided the 360-degree lunar ecliptic into twenty-seven equal parts (27 days = sidereal cycle of the moon) each part equal to 13 degrees, 20 minutes of are, the precise measure of the 13:20 timing frequency. A complex variety of lunar calendars were developed early on, but were gradually displaced by what we call the pseudosolar Babylonian division of the Earth's ecliptic into twelve 30-degree segments, a fact that reflects the adoption to the unconscious 12:60 frequency program.

The Hindu (pseudo) solar calendars are all based on the same division into 30-degree arcs-a circle or plane in space! While the base count of the present era is taken from February 18,3102, the beginning of the Kali Yuga-the last and darkest age-exactly 5,104 years ago (dated from Gregorian year 2002), today there are at least six different calendars in use throughout the subcontintent, each essentially the same as the other, but characterized by different dynastic starting points. These differing counts, coupled with the distinct profusion of religious holidays, accounts for the jarringly conservative and colorful confusion that characterizes present-day Indian society-a neo-Babylonian time warp, its populace locked into place by complex systems of pseudo-solar calendars set in the context of earlier Hindu astronomical cycles of a vast and overwhelmingly cosmic nature.

"TIME IS MONEY": THE WATCHWORD OF THE 12:60 FREQUENCY

All of the aforementioned time-reckoning systems, with their various beginning points, can be mapped within the 5,200 tun/5,125 solar cycle of the thirteen baktun Wave Harmonic of history. While we can generally see how the Babylonian and lunar calendar systems shaped the mainstream civilizational orders of the Old World, in our time map we must also take into account the workings of the 12:60 artificial timing frequency within the unconscious noosphere. In this sense, the 12:60 frequency refers to the principle programs of control adopted and maintained through use of the twelve-month calendar systems in the Old World. (See plate 5, Noospheric Time Map-Wave Harmonic of History.)

Plotting the incidence of these calendrical streams on the time map, we must keep in mind that the calendar systems are the chief instruments for establishing and maintaining the human mind and social order within very specific programs of behavior. These programs, at least in the Old World, were characterized by the introduction of artificial systems of exchange called money, and accompanying programs of taxation. These systems were adopted to a greater or lesser degree by all the Old World societies. In this we find the root of the slogan, "time is money," where money is the artificial medium to negotiate artificial time. The end of history is the end of money and the end of artificial time. The word "calendar" itself is derived from a Latin word meaning "account book," the first day of every month being *calends* or the date of payment of debts.

Now, let us return to a consideration of the two principal and resultant 12:60 instruments dominating the end of historical time: the Gregorian calendar and the mechanical clock. We may note that toward the conclusion of the twelfth baktun cycle (A.D. 1618), the clock reaches its perfection at the very moment of the implementation of the Gregorian calendar reform (A.D. 1582), just after the act of global circumnavigation had been accomplished. Within the entire thirteenth baktun cycle, 1618-2012, the coincidence of these three factors-clock, calendar, and European global circumnavigation-induced what amounts to a 12:60 frequency capture of planetary time in the biosphere. What are the implications of this takeover of time by an irregular and irrationally measured Gregorian calendar and its accompanying micro-organizing device, the mechanical clock?

Plotting the growth curves of human population, machine, and money, we see that the complete infusion of the 12:60 artificial timing frequency creates an unprecedented acceleration, the exponential peak of which occurs at the moment of the Inevitable Event. For this reason, according to the Law of Time, it is important to comprehend even more deeply the nature of calendars and their effect, especially with regard to the current global standard.

Because the calendar is a macro-organizing principle, when accepted over time it establishes in the mind, individually and collectively, a set of perceptions that are automatically taken for granted as being "real" and "indisputable." These sets of perceptions define the paradigm by which a people, culture, or even an entire civilization operates. The Gregorian calendar, the current global civil standard, is the paradigmatic macro-organizing principle in which are embedded all the laws, customs, institutions, and scientific principles governing the present global civilization-not to mention all the holidays of the Vatican-ordained Catholic Church. According to the Law of Time, the current calendar is an irregular standard of measure; its units of measure do not correspond. This calendar represents, therefore, the institutionalization of disorder and entropy. Because it has existed for a sufficiently long duration in the human historical cycle-2,OOO years, including its predecessor, the Julian calendar-the perceptions fostered by the unconscious acceptance of the Gregorian calendar are taken as the unshakable bedrock of nature and reality. All current beliefs-economic, political, and scientific, from democratic neo-liberalism to the special theory of relativity-are actually products of the underlying perceptions promoted by this calendar, and have no reality apart from the beliefs about time that the calendar engenders. *Change the macro-organizing principle and you change the paradigm.*

THE NEED FOR A TRULY "NEW MILLENNIUM"

As the global civil standard, the artificial, irregularly measured Gregorian calendar is a self-reinforcing feedback loop. As such, it furthers and maintains all linear time concepts, thereby establishing a host of entropic, disordered value concepts such as the violent universe, the degradation of matter, quantum physics, the arrow oflinear time, the doctrine of techno-economic inexorability, and an attendant host of unresolvable problems-crime, drug abuse, terrorism, environmental deterioration, and so forth. Why? Because as an irregular standard of measure, the Gregorian calendar is incapable of producing harmony. Only harmony can unify. The Gregorian calendar is not a unifying harmonic standard. Lack of a unifying global standard exacerbates all current conflicts. Condition the mind to an irregular standard and the mind will adjust to disorder and chaos as normal conditions of existence.

The world's racial, tribal, historical, and religious conflicts are embedded in and a function of different timing systems (calendars), all now coordinated within a master irregular timing standard and macro-organizing principle, the Gregorian calendar. In this calendar's irregularity is embedded the view that the world of the inevitable degradation of matter is the exclusive object of all present-day science, and hence is also reflected in the entropic nature of global civilization and its deteriorating social processes. The Gregorian calendar conditions the mind to hopelessness of resolution. Without a unifying harmonic standard humanity is incapable of finding long-term resolutions for any of its problems. Disharmony can never produce harmony. Only harmony can produce harmony. All conflict can only be resolved by the application and within the context of a harmonic timing standard.

Now, following the attack on the Twin Towers and the Pentagon, we can say that the Gregorian calendar is bringing history to an end. The Inevitable Event was programmed into the Gregorian calendar, and the Gregorian calendar usurped the power of all calendars in the thirteenth and final baktun cycle of history. When the third millennium officially opened at the dawn of the year 2001, the pride of industrial man was at an all-time high. Would it be a millennium of peace or one of war? If the last century of the second millennium was the century of total war, was the third millennium going to do anything but inherit the unfinished programs of the old millennium? Could the Earth and the biosphere withstand a millennium of total war? Hardly. The Vatican "jubilee year," which ended with the beginning of the third millennium, turned sour with the Twin Towers apocalypse, as is vividly reflected in a photo of Pope John Paul II taken just after the Inevitable Event. The Gregorian calendar marks its inception with the birth of Christ, a date and moment that was never historically noted, and so is shrouded in speculation and conjecture. Now the twisted civilization of Gregorian time is locked in mortal combat against a terroristic specter that it alone could have projected into manifestation. Soon, by biospheric standards, the world of artificial time will be over and gone. And then the millennium will be truly new.

The very word *millennium* conjures possibilities of Earth-shattering events and cosmic prospects. If the inception of the Kali Yuga, 3102 B.C.,can be said to be the point where the historical process became irreversible, eleven years after history began in 3113 B.C.,then in counterbalance, the Inevitable Event, 2001, which had to occur eleven years before the end of the cycle of history, A.D.2012, was more than just millennial. It was the termination of more than 5,000 years of the irreversible motion of artificial time. Such symmetry-eleven years after the beginning and eleven years before the end of history-is the hallmark of the synchronic order. For 5,100 years the count of days has numbered the long and increasingly perilous saga of human civilization. How bright was the beginning and how grim will be the end?

There at the beginning of the cycle was the "first city," Uruk. There at the end of the cycle, the fallen Twin Towers of Babel of the "last city," New York (= New Uruk), the ultimate monument to history itself and the final event to initiate the actual closing of the cycle of history.

> *There it is, the city of Uruk. Urshunnabi, climb up on that wall, the outer all shining with the brilliance of burnished copper. The seven wise men laid the foundations. One third of the city is buildings, cunningly executed, one third of the city is garden with rose and bird, and one third of the city is field with the temple of Ishtar within. Goddess of love and struggle.*
>
> \- *THE EPIC OF GILGAMESH*[6]

If Uruk is the place where the seven wise men laid the foundations, those foundations were measured by the 24 hours, 60 minutes, and 60 seconds derived from the 360 degrees of the circle. Fifty-one centuries later, the gardens are all but gone, the fields themselves mechanized and chemicalized by the technosphere. Earth's inhabitants are in terror. Five weeks after the Inevitable Event, now known simply as 9-11, the end of history still smolders. After the struggle, will the goddess of love return?

▼

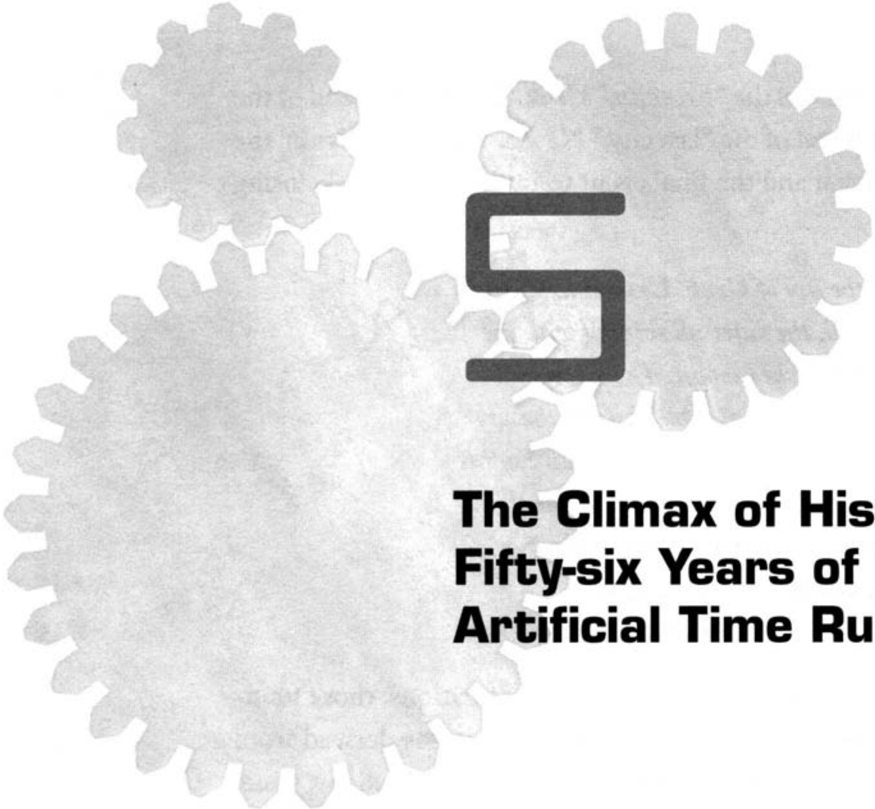

5

The Climax of History, the Fifty-six Years of Hiroshima—Artificial Time Runs Out

THE IMAGE of the Noospheric Time Map-Wave Harmonic of History on plate 5 of the color insert, showing the concurrent evolution of the different calendars of the major civilizations of the world (the illustration is by no means exhaustive), is intended to demonstrate the fact that the human race was not unified by a single coherent timing standard throughout twelve of the baktuns of historical development. It was only in the thirteenth and final baktun that humans became organized by a single timing standard. However, it was not a timing standard in accord with the synchronic order of the Law of Time, but the artificial 12:60 timing frequency as represented by its two organizing instruments, the Gregorian calendar and the mechanical clock.

The seeds of this moment had already been planted in the noosphere at the beginning of history. From both Sumeria and Babylonia the taint of the 12:60 frequency then spread through the noosphere and, by means of historical dissemination and conquest, throughout the Old World. By the time the historic moment arrived to ripen the 12:60 frequency into its full fruition-A.D. 1618, the beginning of the scientific revolution and the thirteenth baktun-there were virtually no longer any cultural traditions governed by the natural 13:20 timing frequency. The civilizations and peoples of the New World were now under the subjugation of the prevailing world order of European global imperialism. The stage was set for the brief but final phase of mechanization, of which the technosphere itself is the climax.

What the Noospheric Time Map also delineates is that the 13:20 frequency was the unconscious organizing factor of prehistory, and that, undoubtedly, a key timing system intuitively evolved among many prehistoric peoples independently of one another. This was the universal, harmonic solar-lunar standard of the Thirteen Moon/28-Day calendar. The defining point of history arrives with the establishment of a 12:60 timing standard within the noospheric unconscious. This frequency predominates in the development of twelve-month pseudosolar calendars, as well as the twelve-month synodicallunation calendars that are not the measure of a solar year. These two commingled systems are entwined throughout the ancient history of the Old World as the central thread of development shaping the conception of time and resulting in what were to become the two dominant instruments of human time reckoning, the Julian/Gregorian calendar and the mechanical clock.

Julius Caesar himself instituted the famous "calendar reform" that created the 446-day year of confusion in 45-44 B.C., and established a calendar that was just as patently confusing. Julius Caesar's was not the only calendar reform in the ancient world. The Essene movement, founded by someone known simply as the Teacher of Righteousness and of which Christ was supposedly a member, began as a revolt against the Hebrew lunar calendar. The issue was the need for fixed holidays within the solar cycle, something that is virtually impossible in a lunar calendar that retrocedes eleven days every solar year. There is strong reason to believe that among other calendars, the Essenes favored the Thirteen Moon/28-Day calendar. This even leads to the question: did Christ himself follow a thirteen moon calendar? In any case, it is the 12:60 frequency transmitted through the irrational disorder of the Julian/Gregorian calendar that captures the mind of the human race and the biosphere during the final years of the cycle of history. And it is precisely for this reason that the pace of human civilization becomes exponentially and entropically accelerated during the last few centuries of its development.

Shocked into higher consciousness by its own barbarism masquerading as civilization, at the end point of history the human has the opportunity to return to the natural 13:20 frequency of synchronization. This return would mark the emergence of humanity from the unconscious mechanistic compulsion of artificial time into the conscious field of post-history. The only possible option to unify humanity-once the yoke of the Gregorian calendar is removed-is the true solar-lunar calendar of Thirteen Moon/28 Days. With the possibility of this positive end in sight, we may turn to a more descriptive and definitive analysis of the technosphere. In this way understanding the modalities of the technosphere as a distinct and entire process, we

Gregorian Calendar 28-Year Cycles: 1945–2001, the Age of Terror

It takes twenty-eight years for the Gregorian calendar to run through its cycle of permutations. Every twenty-eight years, like a phonograph record of time—a chronograph—the Gregorian calendar guides its users through a mentally debilitating construct of unevenly measured and irrationally named months, paced by a system of 7-day weeks that bears little relation to the lengths of the months or years. In addition, every four years there is an extra day. Within this medieval mathematical jumble are kept the collective unconscious programs that repeat every twenty-eight years. What happened in 1945 will somehow repeat or have an effect again in 1973, in 2001, and then in . . . it's up to you. You can change this if you want to.

First day of the year	Tuesday	Wednesday	Thursday	Friday	Sunday	Monday	Wednesday	Friday	Saturday	Saturday	Monday	Thursday	Tuesday	Sunday
Hiroshima 1945 – 72						**1945**								
	1946													
		1947												
												1948		
										1949				
					1950									
						1951								
													1952	
			1953											
								1954						
									1955					
														1956
	1957													
		1958												
			1959											
				1960										
					1961									
						1962								
	1963													
							1964							
								1965						
									1966					
					1967									
											1968			
		1969												
			1970											
								1971						
										1972				
Twin Towers 1973 – 2000						**1973**								
	1974													
		1975												
												1976		
										1977				
					1978									
						1979								
													1980	
			1981											
								1982						
									1983					
														1984
	1985													
		1986												
			1987											
				1988										
					1989									
						1990								
	1991													
							1992							
								1993						
									1994					
					1995									
											1996			
		1997												
			1998											
								1999						
										2000				
Inevitable Event 2001 - ?						**2001**								
	2002													
		2003												

2004 Year of the Great
Calendar Change
Where will you be? 2012

might then develop a deeper insight into where we have just been and why we need to go somewhere else.

The technosphere defines a 56-year cycle, 1945-2001, coordinated by two complete Gregorian 28-year cycles. For any Gregorian calendar year, the days of the weeks in their irregular monthly succession and in relation to the permutation cycle of leap years repeat precisely every twenty-eight years, during which time there are always exactly seven leap years. This means that Gregorian calendar years 1945, 1973, and 2001 possess the same exact annual arrangement of the days of the week in their monthly succession. If today is Sunday, November 11, 2001, so it was in 1973, and likewise in 1945. In this regard the Law of Time frames the cyclic recurrence of the otherwise irregular 12:60 Gregorian calendar by the intrinsic formulation 28:7. That is, just as the 19:7 factor coordinates the lunar calendars every nineteen years, there being seven intercalations of the thirteenth moon during that cycle, so in the Gregorian calendar every twenty-eight years there are always exactly seven leap days and years.

Understanding a calendar as the instrument that locks the conditioned programs of a given culture or people into place, we can now understand how the unconscious metaprogram of the macro-organizing principle of the Gregorian calendar cumulatively recycles all its millennial programs every twenty-eight years. Since its inception in 1582, the Gregorian calendar has been dragging forward a host of conditioned thought forms and perceptions, including those inherited from 1,500 years of the Julian calendar that it had reformed. At points of dramatic break in the continuity of human consciousness, a new set of cycles is set to recur. Such was the case in the year 1945, with the awesome blast of Hiroshima. The entirety of the mind field conditioned and held in place by the Gregorian calendar then reaches its full fruition during the two subsequent 28-year cycles, fifty-six years in all-thus bringing forward two millennia of conditionings, conflicts, and unconscious death wishes or apocalyptic programs. The first cycle was initiated in the year 1945, the year in which Vernadsky died, the atomic bomb was tested once and used twice, and the Second World War came to an end. That year, 1945, marks the beginning of the technospheric bubble. It is most telling that the Second World War was concluded at the beginning of this 56-year cycle of the technosphere, for it was a war concluded not with peace but with instruments of mass destruction and terror. Thus began the 56-year era of Atomic Terror.

The technosphere has its origins with the full capture of the human mental field by the 12:60 frequency in 1618. From 1618 onward, the noosphere is increasingly obscured by a mental field known as the *technospheric sheath.* Slowly but surely, the technospheric sheath replaces civilization. This is first done by the introduction of mechanistic linear time. It is important to note that the Julian count, the basis of all modern scientific calculations, is a scale created by Thomas Scaliger in 1583, twenty-one years after the 1562 Mayan book burning. This linear time scale, like the Gregorian calendar reform itself, was intended to co-opt the Mayan thirteen baktun Long Count by setting a count of days that begins the first of January, 4713 B.C., or some 1,600 years prior to 13.0.0.0.0, the beginning of the thirteen baktun count in 3113 B.C. This deliberate historical act, the Julian count, along with the Gregorian calendar and the mechanical clock, established the paradigmatic notion of the linearity of time in the ripening field of scientific thought. Reflected in the noosphere, this linear, irreversible time concept levels and stunts the realization of human mass consciousness. In fact, during the technospheric cycle, especially after 1754, the ceiling of human consciousness is maintained by a preoccupation with mechanistic third-dimensional operations while becoming increasingly alienated from the organic order of reality. This creates the 12:60 consciousness constant, a mental ceiling that actually diminishes in proportion to the increasing rates of multiplication, propagation, and intrinsic velocity of the machine.

With the actual rise of the Industrial Revolution, dated in the noosphere to A.D. 1754, the synchronic point at which mechanization becomes an irreversible factor of the biosphere, the technospheric sheath, henceforth enters its next stage, going from a purely mental sheath to the industrial sheath, the prelude to the proto-technosphere itself. The industrial sheath spreads throughout the biosphere between 1754 and 1901, the official beginning of the twentieth century. At this point we enter the forty-four years of the proto-technosphere, 1901-1945. During this critical stage of the proto-technosphere, the actions of human behavior interacting with machine technology make the expression of true culture increasingly difficult, if not impossible. Civilization becomes a set of symbols purveyed through museums, galleries, and theaters, and recorded and reproduced in ever more advanced technological forms. But what of culture, which is not the same as civilization-what becomes of culture?

To begin to answer that question, some further definitions are in order. The technosphere is defined by and based upon one key term: *technology.* According to the common dictionary definition, which already reflects the mass mind, technology is "the totality of the means employed to provide objects necessary for human sustenance and comfort."l Today virtually all the means employed are themselves mechanical in nature. In common parlance, therefore, technology refers to the complex apparatus of mechanization. In fact, technology *is* mechanization; it is the ability to convert human labor into processes carried out purely by machine-oriented or mechanical means. This is also inclusive of the entirety of computer technology, which represents the mechanization of the more purely mental processes of thought and communication.

Mechanization, we must remember, originated in the clock, in the mechanization of time. It is the mechanization of time that presupposes the tendency toward mechanization as a state of mind within the noosphere. Since artificial time is characterized by the illusion of an inexorable and irreversible linearity, the compulsion toward materialism is also experienced in the same way, an inexorable motion spearheaded by the advance of ever more improved machines. The machines themselves are the products and means of industrialization-the technological transformation of raw goods into consumer goods, a process accounting for much of the free energy introduced into the biogeochemical combustion of the biosphere.

Inseparable from mechanized technology, too, are materialism and the concept of the World Market and, more recently, of globalization. Materialism is implicit in the dictionary definition of technology as the means employed not only for human sustenance but also for human *comfort.* This comfort can only be of a material form or nature, hence the pursuit of comfort through mechanized technology can only be for the furtherance of a philosophy and exaggerated lifestyle of materialism-the belief that only the material things of this world have any value. Of course, the value of material goods and comforts increases when the value of money is put into the technospheric equation. The first stock market opened early in the eighteenth century, providing a place where money could be used as speculative capital to promote the advance of commodities and machine products. The entire sphere involving commodity production and its conversion into consumer products eventually came to be known as the World Market. Integral to the success of the World Market was the creation of the modern banking system, based on the institutionalization of interest rates and the principle of lending money to companies for the purpose of furthering the transformation of the biosphere into the technosphere.

While the concept of the World Market prevailed for a long time as the definition of the system for exploiting and extracting natural resources from the biosphere, and then converting them into industrial goods to be consumed by the human populace worldwide, globalization is far more recent. Globalization represents the absolute triumph of capitalism as the dominant economic doctrine of the human species in the biosphere, and the absolute basis of the technosphere in its final phase of development. As such, globalization is the economic system of neoliberal market economics that treats the whole world as the proper sphere of its policies and practices. It is a specialized form of monetary imperialism that gives rise to and is supported by the doctrine of monetary politics, the use of money as a means of coercion or even for the buying out of whole governments. Of course, the principle practitioner of monetary politics is the nation with the most money, the United States of America, home of the World Trade Center.

Fostering corporate multinationalism, globalization is inseparable from the consumer philosophy it promotes. Consumerism is the devaluation of the human into a link in the chain of the cycle of industrial production, the object of globalization being to inspire and promote a massive and wide-scale consumerism worldwide. This is what is meant by such phrases as "increasing consumer purchasing power" and "opening new markets." To promote and defend its interests, the all-pervading system of globalization first developed the World Trade Center, and in the last decade, following the end of the Cold War, the GAIT and the WTO, the global "Chamber of Commerce."

As a political hegemony, globalization is managed by the G-7 group of the seven most industrialized nations (the United States, Canada, the United Kingdom, France, Italy, Germany, and Japan-now, with Russia, sometimes referred to as the G-8). This organization, headed by the finance ministers of these seven countries-like the seven wise men of Uruk at the beginning of history-was secretly formed in 1974 by the CIA, one year after the completion of the World Trade Center Towers and the formation of OPEC. The first public meeting of the G-7 occurred in the summer of 1990, just prior to the Iraqi invasion of Kuwait. The doctrine and system of globalization-now overtly controlled by the G-7 and its chief ally, the Euromarket nations-is inseparable from industrial technology as the means employed to provide objects necessary for human sustenance and comfort. Implicit in the philosophy of globalization is the right of the G-7 to promote and defend its way of life, regardless of its ultimate effects on the biosphere and at the expense of ideologies and beliefs not consistent with it. To defend itself, globalization has at its disposal the U.S. military, the command of which is housed by the Pentagon, and the new technospheric multinational military cartel, NATO. In a word, globalization is the triumph of the military-industrial complex.

From the point of view of the biosphere, globalization is the cancer of the human species consuming non-renewable resources and, through the release of free energy, effecting the final critical increase in biogeochemical combustion. The planetary instrument for furthering the entire process and philosophy of globalization is the technosphere. The technosphere may be defined as the entire apparatus of mechanized technology and its support philosophy (globalization), understood as an artificial sphere encompassing the globe. As such, the techno sphere is discontinuously interspersed within the biosphere. More precisely, the technosphere is located physically and mentally between the biosphere and the noosphere, in other words, between the Earth's vital sphere and its mental envelope.

Being the description of an evolutionary stage, the technosphere is governed by the Law of Time as a 56-year event continuum along the lines of an artificial planetary cocoon. Insofar as it is the projection of human thought, the technosphere is the materialization of the sum mental processes denoted and conditioned by mechanization in all its aspects, and that have their primary root in the mechanization of time. It is the mechanization of time that essentially establishes the artificial timing frequency as the capacity to engender the machine and, consequently, an artificial structure, the technosphere. Because it is artificial, however, its duration is highly limited and is subject to the inherent inconsistencies and contradictions in human thought and social structures because of the aberrant effects of adapting to artificial time in the first place. The effect of adaptation to artificial mechanized time creates a social-political hierarchy known as the technocracy-rule according to the needs of technology. It is the effect of maintaining the technocracy and the technosphere itself that establishes the inability of human consciousness to do anything more than remain in a constant, servile, and unchanging state of machine dominance. All the while, by contrast to the unchanging condition of consciousness, the machines evolve exponentially in number and complexity. As Marshall McLuhan so correctly put it, the humans are the bees of the machine.

To be fully understood, the technosphere cannot be seen apart from its place in the biosphere-noosphere continuum. The technosphere represents the absolute control of the biosphere by artificial 12:60 time. Artificial time generates the artificial medium of money, hence the operating philosophy of the technosphere: "time is money." Money is the lifeblood of the technosphere. Servitude to money draws humans away from the biosphere in order to be devoured by Mammon-the technosphere as an all-consuming entity. In this process the humans get money in return for selling their bodies and souls to artificial time, usually eight hours a day, often to produce artificial goods or for service in maintaining the artificial system of money itself. This money gained from servitude is used to buy consumer goods, including more machines of every kind, and thus the human participates as a vital link in the cycle of biogeochemical combustion. In this way the mechanized lifestyle converts the human-the consumer-into the biomass necessary for maintaining the technosphere.

By 2013
Earth is galactically synchronized into a new biogeological epoch: the psychozoic Era.

As the mental envelope of Earth, above and discontinuous with Earth, the noosphere can only mean one thing: the advent of universal telepathy among the humans.

The difference between the new time and the old time is like the difference between heaven and hell.

Planetary 9-11, 2001 inaugurates biogeochemical combustion, biospheric means of transformation, and self-mutation.

Technosphere:
Biospheric agent devised for maximum biogeochemical combustion.

Built in an area known as "Hell's Bottom"

The Pentagon
1941–200?

Impregnable fortress of the American war machine, meant to protect "democracy" = American domination of world market and foreign/economic policy of globalization. "The form of the Pentagon . . . is that of a pentagram, the chief magical aid for gathering power. When a pentagram is drawn with its base toward the north, and its chief point to the south, it is useful for negative or black magic purposes."
—*Transformative Vision*, page 249

*Note: Gregorian calendar coordinated by Law of Time. Repeats precise order of days and weeks precisely every 28 years.

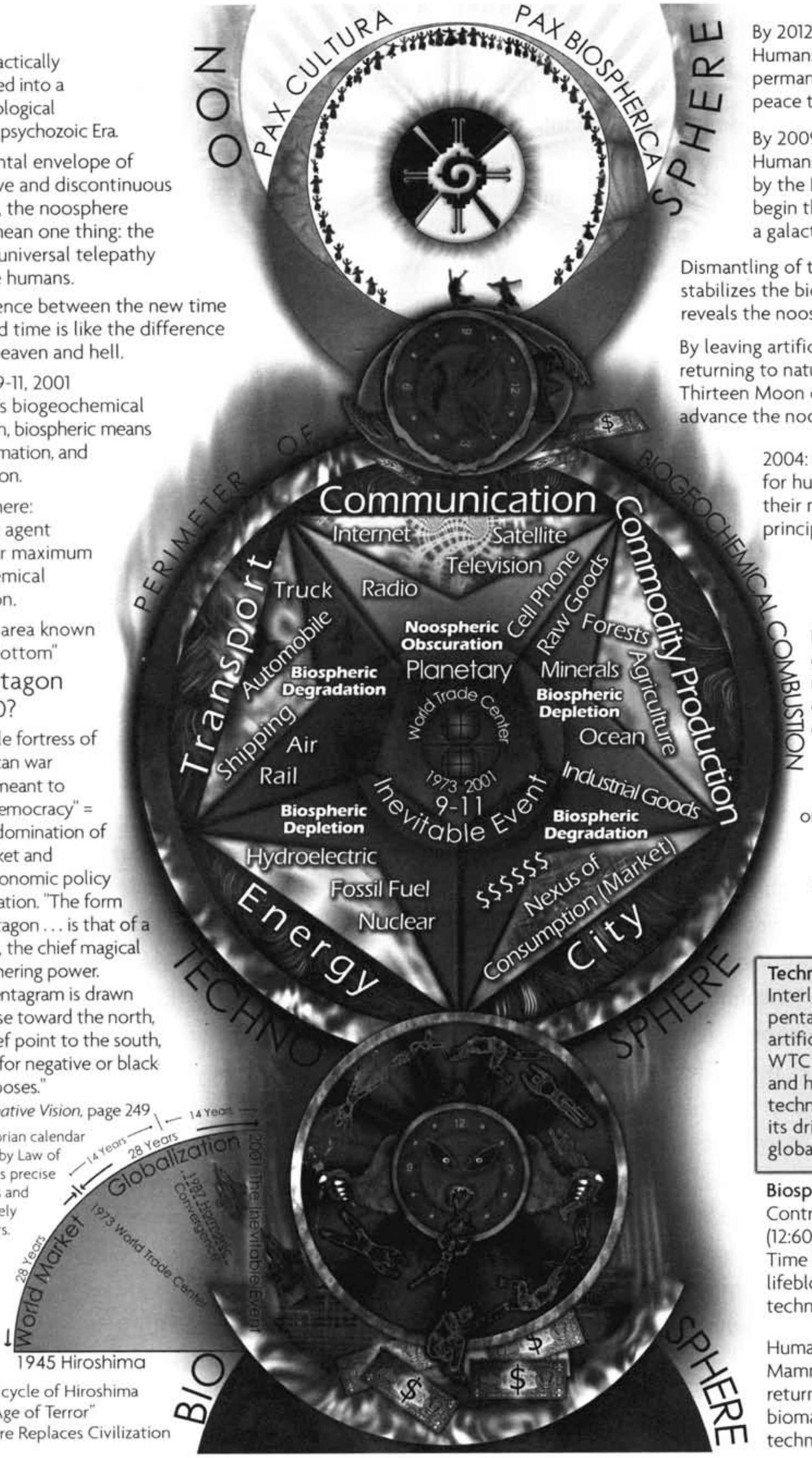
NOOSPHERE
PAX CULTURA
PAX BIOSPHERICA
PERIMETER OF BIOGEOCHEMICAL COMBUSTION
Communication
Internet
Satellite
Television
Radio
Transport
Truck
Automobile
Shipping
Air
Rail
Commodity Production
Cell Phone
Raw Goods
Forests
Minerals
Agriculture
Ocean
Industrial Goods
Noospheric Obscuration
Planetary
World Trade Center
Biospheric Degradation
Biospheric Depletion
Biospheric Depletion
Biospheric Degradation
1973 2001
9-11
Inevitable Event
Hydroelectric
Fossil Fuel
Nuclear
Energy
$$$$$$
Nexus of Consumption (Market)
City
TECHNO SPHERE
BIO SPHERE
14 Years
28 Years
14 Years
Globalization
1987 Harmonic Convergence
2001 The Inevitable Event
1973 World Trade Center
28 Years
World Market
1945 Hiroshima

56-year cycle of Hiroshima
"Age of Terror"
Technosphere Replaces Civilization

By 2012
Humans realize a permanant condition of peace through culture.

By 2009
Humans are reorganized by the Law of Time and begin the construction of a galactic culture on Earth.

Dismantling of the technosphere stabilizes the biosphere and reveals the noosphere—2005.

By leaving artificial time and returning to natural time via the Thirteen Moon calendar, humans advance the noosphere.

2004: Last opportunity for humans to change their macro-organizing principle.

Exhaustion of the Technosphere—2001–04.

Exponential growth curves of money, machine production, energy, and people climax.

Planetary 9-11:
time runs out on the technosphere.

"New" millenium nipped in the bud; civilization R.I.P., 2001

Technosphere:
Interlocking interactive pentagon of five artificial systems - WTC = Nerve center and hub of the technosphere and its driving power, globalization.

Biosphere:
Controlled by artificial (12:60) time.
Time is Money—lifeblood of the technosphere.

Humans devoured by Mammon. Get money in return and create biomass to maintain technosphere.

Biosphere–Technosphere–Noosphere

As an operating planetary structure, the technosphere consists of five interactive components.

1. COMMODITY PRODUCTION

Commodity Production, the basis of the entire World Market as a biospheric system, refers to the transformation of natural resources through industrial means into consumer products. Commodity production alone accounts for the complex of industrialization that brings the World Market and the technosphere into existence. According to *The Biosphere Catalogue,* the post-World War II growth of the World Market necessitated the establishment of the World Trade Center, a global complex (114 such centers in 1985), but with its principle coordinating unit being the Twin Towers in New York City, home of the World Trade Institute and the corporate offices of the New York Stock Exchange, the key to the world money and commodities market.

2. CITIES

Cities, the second technospheric component, represent the human social organism or marketplace for accommodating the production and consumption of industrial goods. Civilization is city life as it has evolved from Uruk to New York (= New Uruk). While prior to industrialization the city depended on a healthy agrarian economy for its support, industrialization pushed the city to new heights of artificiality, not the least of which were round-the-clock lighting systems that allowed for 24-hour consumer markets. As such, the city is the nucleus of the technospheric system, the hub of its commodity production, of its energy consumption, and of its transport and communication empires. The growth of the cities around the world is exponential in the second half of the twentieth century, reflecting the triumph of the technosphere. It should be noted that in the fifty-six years of the technosphere, the human population almost tripled-from 2.2 to 6.2 billion, a growth demonstrated in the vast expansion of the large urban centers.

3. ENERGY

Energy is the third component of the technospheric complex. The tremendous needs of industrial production have transformed the natural energy processes of the biosphere into gigantic artificial complexes for the production of energy to maintain factory and city alike. The impact of the different artificial energy systems upon the biosphere has been enormous. These systems include coal, oil, and fossil fuels in general, as well as hydropower and nuclear energy. The whole syndrome of artificial time, artificial means, acceleration of the machine, and expansion of human population has required and created the enormous output of artificial energy into the biosphere and contributed greatly to the destabilization and loss of natural resources and atmospheric and hydrospheric contamination through the release of toxic "free energy" by-products.

4. TRANSPORT

Transport. Without the highly industrialized transport systems, the commodity production, the vitality of the cities, and the access to provisions of fuel required for the energy systems would collapse. No sooner had industrial production begun early in the eighteenth century than the first artificial transport systems came into being: the steam- or coal-driven locomotive, and the great transoceanic shipping lines. By the twentieth century, the railroad system and worldwide shipping industry were supplemented by the invention of the combustion engine, which made the trucking industry possible. Following World War II, rocketry and jet propulsion established the transportation might of the great airline companies of the second half of the twentieth century. It is no coincidence that commercial airplanes, the symbol of the technosphere's advanced transport systems, were used to bring down the Twin Towers and to penetrate the Pentagon.

5. COMMUNICATION

Communication is the fifth component that unifies the technosphere into one whole system in touch with itself everywhere. From telegraph to telephone, radio, television, and then finally fax, cell phone, and the Internet, the industrial-era technologies, with their electronic instantaneity, have created the vast communication networks and media empires that keep the world humming and promote globalization above all. When considering modern communication we must also bring into focus the chief method of globalization, which is marketing-propaganda on behalf of consumerism. Marketing is the use of means of mass communication to control the mind of the consumer. In the final stage of globalization, marketing replaces consciousness, or rather, marketing is the manipulation of consciousness-the latter already having been reduced to a stunted level-for purposes of consumerism, or for ideological needs such as patriotism and fear campaigns. The Internet, the marriage of computer and telephone technologies, represents the final stage of the Tower of Babel, and the completion of the process of globalization. The only place to go after the Internet is technology-free telepathy.

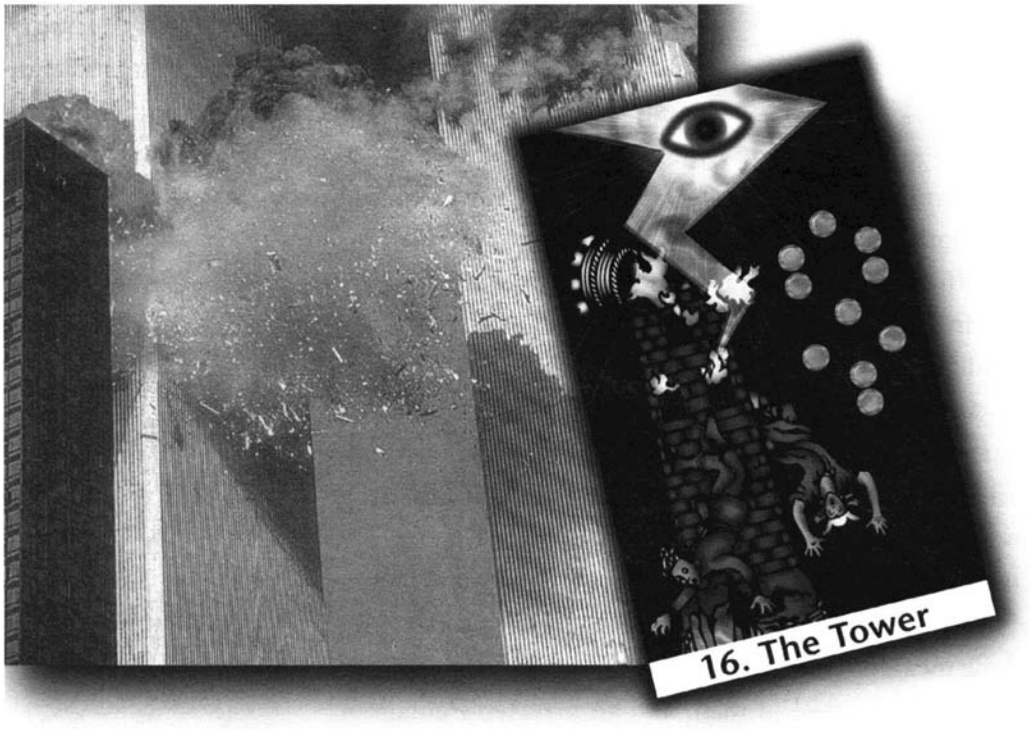

The Tower

From the biospheric perspective, this five-tiered complex of the technosphere is a whole unit. All the subsystems evolved together to establish the structure and operating procedures of the technosphere. When the technosphere goes, the entire system folds. This is what is now beginning to occur. Following the Inevitable Event, the whole technospheric system will be coming down slowly over the next few years, like a giant circus tent that has lost its central prop. How gracefully or gracelessly this happens is dependent on the response of the Pentagon-or the will of humanity to rouse itself in the face of even worse barbarisms.

Because it consists of five interactive components, the techno sphere can be graphically depicted as a pentagon, a five-sided figure. At the center of this pentagon of the technosphere are the World Trade Center Twin Towers. The actual Pentagon, which was attacked along with the Twin Towers, was designed in the late 1930s to house the U.S. Department of War. Its construction occurred in the years 1941-1943, in an old neighborhood of pawnshops and bars called Hell's Bottom.2 When it was built it was the world's largest office building in terms of actual space. Of course, when they were built, the Twin Towers of the World Trade Center were the world's tallest office buildings. The structure of the Pentagon is the prototypical morphology of the structure of the five systems that constitute the technosphere. The Pentagon was the impregnable fortress of the American war machine, protecting American interests and globalization around the world. If there were two central points-actual and symbolic-to what is known as the military-industrial complex, it was the Pentagon and the World Trade Center, and this is undoubtedly the reason why they were both targets of the Inevitable Event.

What is architecturally noteworthy about the Pentagon is that it is designed with five inner pentagonal corridors and office slabs, and that it is constructed with its odd point to the south. Since a pentagon is actually a pentagram, a five-pointed star with its points connected, the Pentagon represents an *inverted* pentagram, as its odd point faces south rather than north. In the traditional Tarot deck it is interesting that the Fifteenth Major Arcana, The Devil, contains an inverted pentagram between the horns of the Beast. Paul Foster Case writes, "This is a key to the whole meaning of the figure [of the devil]. For the Pentagram is the symbol of man, and an inverted Pentagram suggests the reversal of true understanding of man's place in the cosmos."[3] While the Fifteenth Major Arcana contains the pentagon as the inverted pentagram, the next card, the Sixteenth Arcana, is The Tower, which shows the tower being struck and broken apart by lightning (fire of heaven), with humans falling or jumping out of the windows to their death-as accurate a symbolic depiction of the collapse of the Twin Towers as could be found.

What the startling conjunction of these symbolic correspondences points to is the profoundly archetypal nature of the Inevitable Event. It has been prefigured and prophesied, and has now come to pass. No matter how much America and its allies, many of them bought for a price, may strike back, the deed has been done. The Towers are gone; the Pentagon has been penetrated. While it will take several years for the full realization of the absolute magnitude of this archetypal event to sink into the

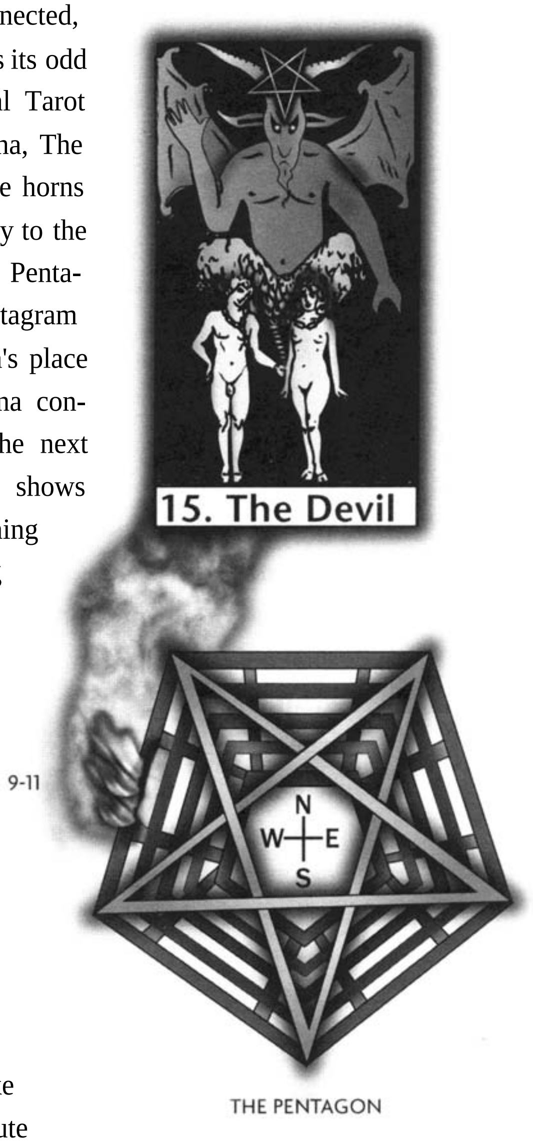

The Devil and the Pentagon

collective mind, it is important to demonstrate the actual structure in time of the techno sphere so that the finality of its end may be seriously considered, and the gate to the future, which has thereby been opened, can be made clearly visible.

As we have shown, the technosphere is a structure totally brought about by the 12:60 timing frequency, and thus is purely a function of the latter. As such, the technosphere is embedded in the global macro-organizing principle of the 12:60 frequency, the Gregorian calendar. Two 28-year cycles, each one divided into four 7-year subcycles, define the duration of the technosphere. Remember that the Gregorian calendar repeats every twenty-eight years. As we have pointed out, the 56-year cycle of the technosphere, 1945-2001, is preceded by the 44-year cycle of the proto-technosphere. The two World Wars were a function of the proto-technosphere, the final major act being the construction of the Pentagon, specifically for consolidating the American war machine-the world's largest office building means that the world's largest business is war, with a $318 billion budget for the year 2001-2002 alone!

The fifty-six years between Hiroshima and the Inevitable Event were the age of terror, for it was atomic terror that initiated the technosphere in a baptism of nuclear fire, and in the end, it was an unimaginable suicidal terror that brought down not just one, but both of the Twin Towers of Babel. The final collapse of the technosphere is also the final war between blood and money. Artificial time has run out. Only a new time will be able to regenerate the biosphere and spiritually revive mankind. What follows is a chronological description of the eight 7-year stages of the fifty-six years of the technosphere. Note the persistence of certain themes. (See plate 5, Noospheric Time Map-Wave Harmonic of History.)

GREGORIAN A, 28-YEAR CYCLE:
TRIUMPH OF THE WORLD MARKET, 1945-1972

1. 1945-1951. CYCLE OF THE BOMB: ATOMIC TERROR AND THE DIALECTIC OF THE COLD WAR.

During the first dynamic 7-year cycle, with the triple event of the first test of the atomic bomb and its two detonations at Hiroshima and Nagasaki, the biosphere becomes irrevocably altered by the introduction of a constant, steady state background radiation into the atmosphere, the actual inception of the biogeochemical combustion. This act officially establishes the technosphere. The reality is that the humans also created their first weapon of mass destruction. It is this act that also immediately sets in motion a destabilization of the human consciousness in the noosphere.Within two years, Mahatma Gandhi, the world's foremost pacifist, is assassinated, as a "free" India is partitioned into three parts-India, Pakistan, and what is later to become Bangladesh. In the same year, 1947, the iron curtain between communist Eastern Europe and the West, epitomized by the Berlin Wall, turns the enmity between the two major powers, the Soviet Union and capitalist America, into the Cold War. The subsequent formation (1949) of the NATO military cartel counters the United Nations, which was founded in 1945 to replace the League of Nations. The UN now has the task of keeping World War III from happening by providing a forum where the "superpowers" can keep their cold war from heating up, and that will contain and manage the spread of nuclear power. Nonetheless, the terror of the Bomb as the ultimate deterrent engenders the arms race. Soon Russia, the United Kingdom, France, and eventually China get the Bomb. War in Indochina, the establishment of the Israeli state in Palestine (1948) at the expense of the sovereignty of the Palestinian people, and the completion of the Chinese Marxist Revolution under Chairman Mao are the highlights of 1949, followed by the Korean War in 1950. All this demonstrates the instability of the world. At the same time, 1949, commercial television production begins in the United States; the age of radio is replaced by the "tube." Americans also begin building freeways and the new system ofInterstate highways marking the triumph of "automobile culture," and the beginning of the suburban consumer lifestyle.

2. 1952-1958. CYCLE OF THE OPENING OF THE TOMB AND THE BEGINNING OF THE SPACE AGE.

In the summer of 1952 a Cuban archaeologist discovers the tomb of the Mayan sage, Pacal Votan. This unprecedented archaeological event marks the beginning of the final sixty years-three katun cycles-of the Mayan thirteen baktun cycle of history. In the following year, 1953, the superlethal H-bomb is tested. The DNA code is discovered, as are the Earth's radiation belts. The United States and Russia conduct numerous nuclear weapons tests in Nevada, Siberia, and the Pacific Ocean. Wars for independence occur in various African states, and in general the era of European imperialism is at an end, followed by the neo-colonialist (Third World) era of guerrilla warfare, poverty, and social instability. In 1956, the year the Russians launch the first sputnik and begin the space age, the United Nations tables the issue of calendar reform indefinitely, thus closing the chapter begun when the League of Nations proposed global calendar reform in 1931. The year 1956 also marks the beginning of the electronic pop culture of the technosphere, rock and roll. By the end of this cycle, the "space race" joins the arms race as a force in promoting the advance of the technosphere. Commercial airlines adapt to jet propulsion, and the great age of air travel begins, with the airport to become the centerpiece of globalization. Initial experiments in computer technology-Univac-and the first great computer corporation, IBM, announce the beginning of what is to become the major technology of globalization. A small fast-food franchise named McDonald's opens in the United States.

3. 1959-1965. CYCLE OF THE ATOMIC MIND EXPLOSION.

The July 26 Cuban Revolution of Fidel Castro, triumphant in 1959, establishes a communist state in the Americas and heightens the Cold War. That same year the Chinese communists bring to a definitive end the rule of the Dalai Lama in Tibet. In the Congo, the popular hero Patrice Lumumba is assassinated after only two months in office. Also in 1960, the human population hits three billion, an increase of one billion since 1930. In 1961 the Institute of Mathematics of the Siberian Branch of the Soviet Academy of Sciences begins the monumental mathematical analysis of Mayan hieroglyphic texts. The attempted 1961 Bay of Pigs invasion of Cuba is followed in 1963 with the assassination ofJohn F. Kennedy, an act that reflects a rising level of cultural turmoil in the United States. Martin Luther King Jr. emerges as a popular hero of nonviolence and civil rights, an issue that points to a more radical discontent in America, culminating in 1965 with the Watts race riots. Rachel Carson's book *Silent Spring* (1963) signals the rise of ecological awareness. The Beatles transform rock and roll into a world-class cultural phenomenon that foments a widespread countercultural movement and atomic mind explosion, fueled by Timothy Leary, LSD, hippies, and the new antiwar movement against the conflict in Vietnam. In 1961 Yuri Gagarin is the first man in space, and the Vatican II Ecumenical Council comes to the conclusion that the best way the Church will survive in the next century is through a strong Pope. The Vatican II document also includes an appendix on calendar reform, the wording of which makes it difficult for there to be any alternative but a watered-down version of the Gregorian calendar. This act seals the technosphere within the unquestioned confines of Gregorian time, ultimately turning the atomic mind explosion in on itself.

4. 1966-1972. FROM THE WHOLE EARTH TO THE WORLD TRADE CENTER: THE TRIUMPH OF THE WORLD MARKET.

In 1966 construction begins on the World Trade Center Twin Towers, while in 1968 World Trade Centers are independently established in Houston, New Orleans, and Tokyo. That same year, UNESCO hosts the first and only conference on the biosphere ever sponsored by the United Nations, "Man and the Biosphere" (MAE), but the program slowly dies on the vine. The cultural conflict in the United States and the entire world also reaches a fever pitch in the years 1967-1969. This ferment includes: race riots, antiwar riots, a march on the Pentagon, the assassinations of Martin Luther King Jr. and Robert Kennedy, and riots in major capitals around the world and on both sides of the iron curtain, as well as the Maoist cultural revolution in China. The excitement of the '60s climaxes in 1969 with the success of the Apollo XIII manned mission to the moon and the consequent feedback factor of the human race seeing itself instantaneously on television as the whole Earth beamed from space, an opening to the noosphere. Apollo XIII initiates the next phase of the space race, travel to other planets. Also in 1969, the Woodstock concert marks the apex of the early rock-and-roll culture and the visionary climax of the cultural revolution. The 1971 Concert for Bangladesh establishes rock and roll with a conscience. In 1970 the vision of the whole Earth is translated into the first Earth Day and the birth of the ecology movement. The United Nations follows two years later with the Stockholm Conference on the Environment. The revolutionary fervor of the previous years turns more violent, and repressive governmental police powers win the day. The first 28-year cycle ends with a whimper, the dynamic destabilization of consciousness reaches a plateau of normalcy-a normalcy to be increasingly punctuated by terrorist activity, as in the Palestinian takeover of the Olympic Village in September 1972. Both apart from as well as in reaction to such activity, the technosphere has now become all-dominant and all-powerful in human affairs.

SUMMARY, GREGORIAN A: 1945-1972

During this cycle, the dynamic of the accelerated geochemical combustion destabilizes the human consciousness while increasing the curve of the machine in human affairs. By 1973 the machine curve meets and surpasses the human population curve; this stabilizes human consciousness at a level of machine normalcy (MN). Human consciousness will remain stabilized at MN, largely through varying forms of terror and fear programming, for the duration of the second and final 28-year cycle. This normalization of human consciousness allows the diverse and rapidly evolving technological components of the technosphere to reach a point of consolidation. This stabilization of the World Market fosters its next and concluding phase: globalization and the triumph of marketing over human culture. It should be kept in mind that the instruments and weapons of mass destruction, of which there were none on January 1, 1945, have now proliferated to five different countries and provide the ultimate backdrop to these first twenty-eight years. The core of the UN's Security Council is formed by the five countries that possess atomic weapons. Geopolitically, the bio-

Machine Normal Consciousness in Relation to Technospheric Growth Curves

Human consciousness, overwhelmed by war—the ever-propagating machine and bondage to artificial time—succumbs to money and stabilizes at Machine Normal, the standard consciousness of the technosphere, which represents an ever-diminishing ratio of intelligence in relation to the growth curves of population, machine, money, and war.

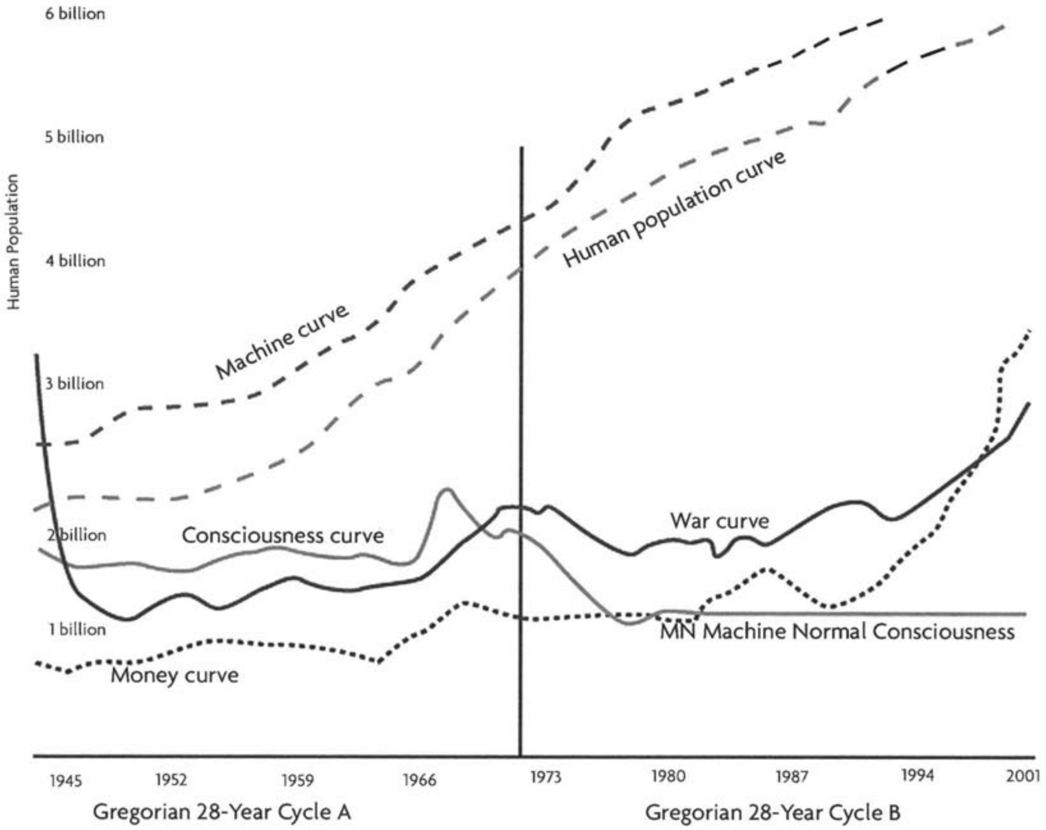

sphere is in control of technocratic powers that are virtually all located in the Northern Hemisphere-whether it is the Pentagon, NATO, the World Trade Center, the United Nations, or the Soviet space program.

GREGORIAN B 28-YEAR CYCLE: CLIMAX OF GLOBALIZATION, 1973-2001

1. 1973-1979. CYCLE OF THE CONSOLIDATION OF THE TECHNOSPHERE: CLUB OF ROME, OPEC, AND THE G-7.

The dedication of the World Trade Center Twin Towers on April 4, 1973 (a repeat of the calendar year of the Bomb, 1945) is the counterpart to the construction of the Pentagon (1940-1941), which sponsored the bombing of Hiroshima and Nagasaki. The tallest buildings in the world when they opened, the 110-story Twin Towers seal the position of New York as the "capital of the world," for on the other side of Manhattan is the glass tower of the United Nations. Simultaneous to the dedication of the Twin Towers is the formation of the Organization of Petroleum Exporting Countries (OPEC). With the exception of Venezuela, OPEC is largely an Arabic (Islamic) economic alliance that, for the first time in the history of the technosphere, brings a Third World influence into planetary politics. The subsequent "oil shortage" demonstrates the technosphere's vulnerability, which in turn causes a further "hardening of the arteries" of its consciousness-the covert formation by the CIA of the G-7. Also in 1973, after five years of study the Club of Rome issues *The Limits of Growth,* which in its analysis of trends foresees a major environmental (biospheric) crisis within one hundred years, yet nonetheless establishes and promotes the standard world model, a setting of production quotas at the 1973 level, to foster sustainable growth at least until 2020. The "standard world model" corresponds to the "normalization" of consciousness. The third war (previous ones having occurred in 1967 and 1949) between Israel and its neighbors occurs this year. By 1975, the world population is at four billion, up a billion since 1960. In 1976 Voyager II reaches Mars. In 1978 Karol Wojtyla becomes Pope John Paul II, the first global Pope, fulfilling the Vatican Council's requirements of a popular Pope to assure the survival of the Church into the next century. With Gregorian civilization triumphant, Mexico, home of the Maya, is the first country visited by the new Pope. In the same period, between 1978 and 1979, the Iranian Islamic revolution occurs, toppling the pro-Western Shah, signaling a heightened conservative Islamic resistance to Westernization and the technosphere. Japanese technology and corporate capitalist economy flourish, capped by the construction in Tokyo of the "New City" Shinjuku and, despite many environmental protests, the Narita airport. The mainstream global consciousness is now "comfortably numb" and on course to unprecedented materialism.

2. 1980-1986. CYCLE OF THE TRIUMPH OF THE TECHNO SPHERE.

The first cycle of the stabilized new normalcy establishes the triumph of marketing-the election of a second-rate American movie star for President, and the birth of the yuppie (young urban professional) generation of America. The rise of MTV signals the standardization of rock style as global culture; while the Live Aid global television concert on July 13, 1985, to raise awareness of AIDS in Africa, is the largest ever of its kind. The September 17, 1985 Mexico City earthquake, the worst in modern times to hit the world's largest city, demonstrates the continuing vulnerability of the techno sphere to natural catastrophe. The early 1980s witness the Intifadah uprising of Palestinians against Israel. The War on Drugs, an aggressive antidrug campaign waged by the West against narcotraffic, mostly from South America and the Middle East, heightens the new normalcy as world drug laws tighten, placing marijuana and psychedelics in the category of "hard drugs." Yet drugs continue to be a major social problem in the industrialized West. In 1986 President Reagan attacks Kadafy of Libya and invades Grenada, testing American ability to commit limited acts of aggression without inciting world censure. The construction of Biosphere II in Arizona, an effort to test the possibility of establishing an artificial biosphere on other planets, ends inconclusively. In 1985 in Bhophal, India, the worst modern industrial toxic chemical disaster occurs with more than 2,000 deaths, followed on April 26, 1986 by the Chernobyl nuclear power plant disaster, the worst since the beginning of the nuclear age. In America the suburban automobile culture gives rise to the super shopping malls. The personal computer (PC) is introduced, while electronic game arcades flourish. Sunday football becomes the major American television pastime, while television itself makes world soccer the number one competitive global sport; the broadcast of professional sports assumes the role of a primary noospheric diversion.

3. 1987-1993. CYCLE OF THE END OF THE COLD WAR AND TRIUMPH OF THE G-7.

On January 24, 1987 Voyager II sends back photos from Uranus; four days later, January 28, 1987, seven American astronauts die in a shuttle launch explosion. On August 16-17, 1987 the Harmonic Convergence peace meditation attracts hundreds of thousands who gather at sacred and natural sites around the world; later that year, October 19, 1987, the New York StockMarket crash occurs, the biggest point drop ever, and Russian Prime Minister Gorbachev initiates disarmament accord, or *glasnost.* By the end of 1987 the population is at five billion, up one billion since 1975. During 1988-1989, a rolling wave oflargely peaceful revolutions and civil uprisings throughout most of the Warsaw pact nations culminates on December 31, 1989 with the tearing down of the Berlin Wall-the end of the Cold War. At the same time, President Bush invades Panama to continue testing Drug Wars and limited-aggression capability. In the summer of 1990 an international bankers' reunion is followed by the first public meeting of the G-7. Every year henceforth, the G-7 meets publicly to set policy for globalization. Within a month of the first public G-7 meeting, August 1990, Saddam Hussein invades Kuwait, and Bush retaliates with the Gulf War, January-February 1991. In August 1991 Gorbachev is ousted in a putsch, marking the end of the Russian Marxist era. Destabilization, war, and genocide afflict the Balkans. A new wave of environmental concerns-ozone depletion, destruction of rain forests, and global warming-inspire a reactivation of Earth Day, 1990, followed two years later by the 1992 Rio Summit, the second United Nations Environmental Conference and the first in twenty years. The conference generates much paper, but the doctrine of sustainability (sustainable industrial growth) is the principle outcome. Meanwhile, with the absence of Marxist Russian economic influence, the neoliberal capitalist economics of America flourish as full-blown globalization, multinational corporations, and the prominence of the IMF and World Bank foment a uniform global corporate culture of cheap consumer services and franchised retail outlets for brand-name goods. This period also sees the rise of East Asian capitalist economies, although the Japanese economic bubble bursts in 1991. Early in 1993 a terrorist bomb explodes in the underground garage of the World Trade Center Twin Towers. The period 1989-1993 marks the beginning of the discovery of the Law of Time-the definition of the artificial 12:60 and natural 13:20 timing frequencies, the Dreamspell codes, and the decoding of the Telektonon Prophecy of Pacal Votan in 1993.

4. 1994-2000. CYCLE OF THE CLIMAX OF GLOBALIZATION.

In 1994, the GAIT Treaty is signed, the Levy-Shoemaker comet hits Jupiter, and a million Mricans die in the Rwanda massacre. The indigenous uprising in Chiapas, Mexico on January 1, 1994 provides a counterpoint to the triumph of globalization that is manifest through worldwide airport expansions, and an aggressive form of tourism that feeds the airline industry. Israel and Palestine enter the "peace process," while Yeltsin consolidates his power in Russia. This is the era of United Nations peacekeeping missions, mostly in the Balkans and Africa. Following the FBI/BATF -inflicted firestorm on a fundamentalist compound in Waco, Texas, a year later, in 1995, a terrorist bomb destroys the Oklahoma City Federal Building. The United States and the United Kingdom inaugurate control ofIraqi no-fly zones, enforced by regular bombing raids; the United States also conducts military actions in Somalia, as well as more successful ones in Haiti. This is the era of rapid expansion of computer and biotechnologies, exponential rise of New York Stock Exchange and the new specialized technologies exchange, the NASDAQ. Bill Gates of Microsoft typifies the rise of the new self-made billionaires of the Clinton era. In January 1995 the Kyoto earthquake occurs, followed by the Tokyo subway nerve-gas terrorist attack, while the Internet, the "information superhighway," is fully up and running. In 1996 in Brasilia, Brazil, the First Planetary Congress of Biospheric Rights signals rise of a new calendar reform movement with the proposal of the World Thirteen Moon Calendar Change Peace Plan, submitted to both the United Nations and the Vatican. In 1997 the Kyoto environmental conference on global warming sets policy, but the United States refuses to ratify it. In 1998 simultaneous U.S. embassy bombings in Kenya and Tanzania signal a new phase of war on terrorism, singling out a Saudi national, Osama Bin Laden, as the terrorist mastermind. By 1999 the world population is six billion, up one billion since 1987. In late spring 1999 NATO initiates a unilateral bombing campaign ofYugoslavia over the disputed region ofKosovo. The World Summit on Peace and Time, at the University for Peace, Costa Rica, sends emissaries to the Vatican and the United Nations with the Declaration of Thirteen Moon Calendar Reform; later in 1999 WTO has its first meeting in Seattle, which sparks anti-globalization riots. In 2000 the Y2K computer scare surrounding a bad time program fails to materialize. On Rosh Hashanah, September 28,2000, Hebrew year 5761, Israeli leader Ariel Sharon's presence at Temple of the Mount, Jerusalem, sparks a new Intifadah of Palestinians, which continues unabated through 2001, the official first year of the new millennium.

5. 2001-? *GOTTERDAMMERUNG* OF THE TECHNO SPHERE: FIRST YEAR OF THE THIRD MILLENNIUM, REPEAT OF GREGORIAN CALENDAR YEARS 1945 AND 1973.

In July 2001, G-7 meets in Genoa, Italy, as more than 100,000 anti-globalization protesters riot. In early September the United States and Israel walk out of the United Nations Conference on Racism in South Africa. Then, within ten days occurs the Inevitable Event, Gregorian 9-11, 2001, the cumulative synchronic pressure point where the technospheric bubble-the artificial time warp in the noosphere-bursts open, spawning in its wake the full-scale War on Terrorism-another name for World War III. The Inevitable Event exemplifies the built-in failure of the Gregorian calendar to commence another 28-year cycle as was done in 1945 and 1973. If the 1945 cycle began with the end of one world war, the 2001 cycle begins with the start of one. Since 2001 is also first year of the third millennium, the Gregorian calendar has created its own apocalypse, opening a noospheric fissure-the climax of the biogeochemical combustion of the biosphere. Now it is only a matter of time before the technosphere experiences total collapse. The nature and extent of the damage sustained during the collapse is in proportion to the intelligence that responds to the knowledge of its collapsing. If this 28-year Gregorian calendar cycle were to go unimpeded until 2029, it is not certain what would be left of the biosphere. But there is one thing of which we may be certain: *artificial, linear time has run out.* An American newspaper editorial at the end of 2001 sums up the crash of the dominant linear time paradigm: "History seems to have slammed into reverse gear around a world that was widely assumed a year ago to be moving in a straight line toward a more prosperous and secular future, one that would be shaped by stock markets and the Internet, not by turbaned zealots or theocracies that have or want nuclear trigger fingers."4

SUMMARY, GREGORIAN B: 1973-2001

With destabilization normalized, the technosphere replaces civilization and marks the end of culture-the new barbarism prevails, culminating in total war. The 9-11 is what actually completes the technospheric cycle. Henceforth it will be the advent of the noosphere, but not in the way predicted by the authors of *The Biosphere Catalog:* "The Noosphere as an active force consists in the harmonious synthesis of the Biosphere and the Technosphere by intelligence."5 By the analysis of the Law of Time, the intelligence is nowhere evident because there is no harmony of intelligence possible under the present timing standards that dominate all the major players in the present global conflict. Since the technosphere is totally a function of the artificial timing frequency, and since artificial time is now short-circuiting itself, there will be no techno sphere to harmonize when natural time is restored to the entire biosphere. Then the noosphere will commence, but only if there is world harmony. Only a harmonic standard will bring harmony to the world. The failure of the Gregorian calendar and its civilization, crystallized in the now-collapsing technosphere, must be quickly acknowledged. By adopting a harmonic standard as soon as possible, the shift to the noosphere can commence, rising and building itself up from the shards and ashes of the toppled technosphere.

▼

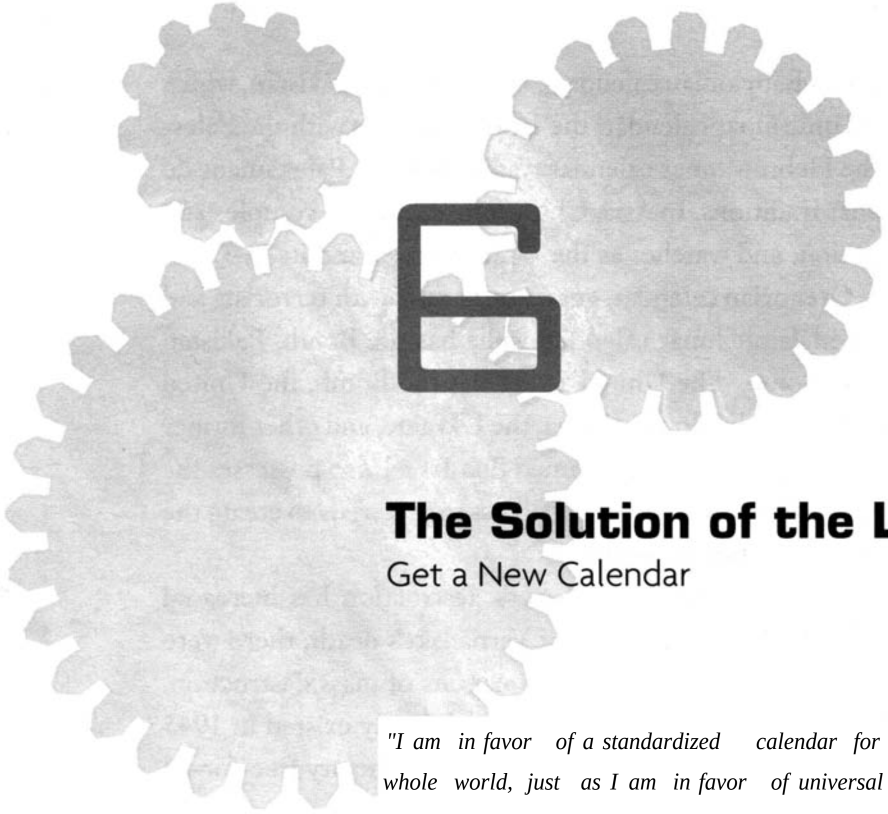

6

The Solution of the Law of Time

Get a New Calendar

> *"I am in favor of a standardized calendar for the whole world, just as I am in favor of universal coinage for all countries, and a supplementary artificial language (like Esperanto, for example) for all peoples. . . I am always ready to endorse any honest movement which will help unify the peoples of the world."*
>
> -MAHATMA GANDHI, QUOTED IN *THE JOURNAL OF CALENDAR REFORM,* 1931[1]

VIEWED FROM THE PERSPECTIVE of the noosphere, the failure of human civilization is the failure of the human being to become synchronized with itself in the correct time of universal nature. Instead of operating as a planetary organism in absolute synchronization with a knowledge of being coordinated by a common harmonic standard, humans race toward the end of history, dissynchronously operating under a plethora of calendars of differing measures enshrining competing belief systems, all globally coordinated by an instrument of irregular measure and obscure nomenclature-the Gregorian calendar. Dissynchronized and at odds with itself everywhere, humanity is broken apart into splintered mind fields, each controlled by a different timing standard, a fact about which the human race is virtually ignorant- to its own detriment.

India, with its different neo-Babylonian calendars, is at odds with Pakistan, which operates according to the Islamic lunar calendar; the Israelis, at odds with the Palestinians, hold as tightly to the Hebrew lunar calendar tradition as the Palestinians do to their Islamic lunar calendar traditions. In Asia, China follows its own complex and ancient lunar calendar tradition, and watches as the United States and its G-7 alliances, all dominated by the Gregorian calendar, get ready to spar with terrorists and militants who also follow the Islamic lunar calendar. India has the Bomb, Pakistan has the Bomb, China has the Bomb. The United States has the Bomb, the United Kingdom has the Bomb. France has the Bomb. Russia, the Ukraine, and other former Soviet nations have the Bomb. It is commonly assumed that Israel also possesses the Bomb. Does Iraq have the Bomb? Can a terrorist get access to materials to create the Bomb as well? The nuclear clock is ticking again.

The number of weapons and technologies of mass destruction has increased several thousandfold since 1945 when, at the time ofVernadsky's death, there were virtually none. This is not to mention the biological weapons of mass destruction, more ambiguous, more lethal, more hidden-none of which really existed in 1945 either. With the fallout from the Inevitable Event an anarchic planetary free-for-all is taking shape. From the noospheric point of view, the humans have run out of (artificial) time, and are now committing their last egoistic act of desperation to hold what they perceive to be their dominant place in the scheme of things. The fact is, however, that the planet belongs to the biosphere and not to the humans.

The inability of the Gregorian calendar to sustain another 28-year cycle, much less another millennium, means that an anarchy of time and a time of anarchy is loosed upon the planet. At the present rate, what would the world look like in 2029? One does not want to imagine. The anarchy of time is exemplified by the different competing powers that are unconsciously driven and motivated by programs stored in their differing calendar systems. This failure of artificial time to contain itself is truly the apocalypse. The apocalypse could occur only when the time program had become saturated. The two 28-year cycles, 1945-2001, carried forward not only the exponential advance of machine time, but the accrued karmic baggage of two millennia of the invented religion of Christ. Throughout the last millennium, the protagonists of this religion waged a type of holy war against an embattled world-a war waged by missionaries and conquistadors alike, down to this very day. The week before the Inevitable Event, the Taliban was deporting Christian missionaries from Mghanistan, while other Christian missionaries scour the jungles looking for one more indigenous soul to convert. This activity is inseparable from the Gregorian calendar that programs it.

Entropic Black Hole of Consciousness

A calendar is a mental instrument for holding the programmed patterns of thought and behavior of a given culture, people, or civilization. A people can rise no higher than the program of time their culture imposes on them as second nature. A program of time determines nature of consciousness. Different timing programs scheduled by synchronically inaccurate devices within a biospheric whole governed by a universal timing frequency apart from the deviant human standards can only result in an entropic black hole of consciousness: the clash of civilizations and cultures that precedes the mentally cleansing force of the noosphere and the inevitable return to natural time and the universal 13:20 timing frequency.

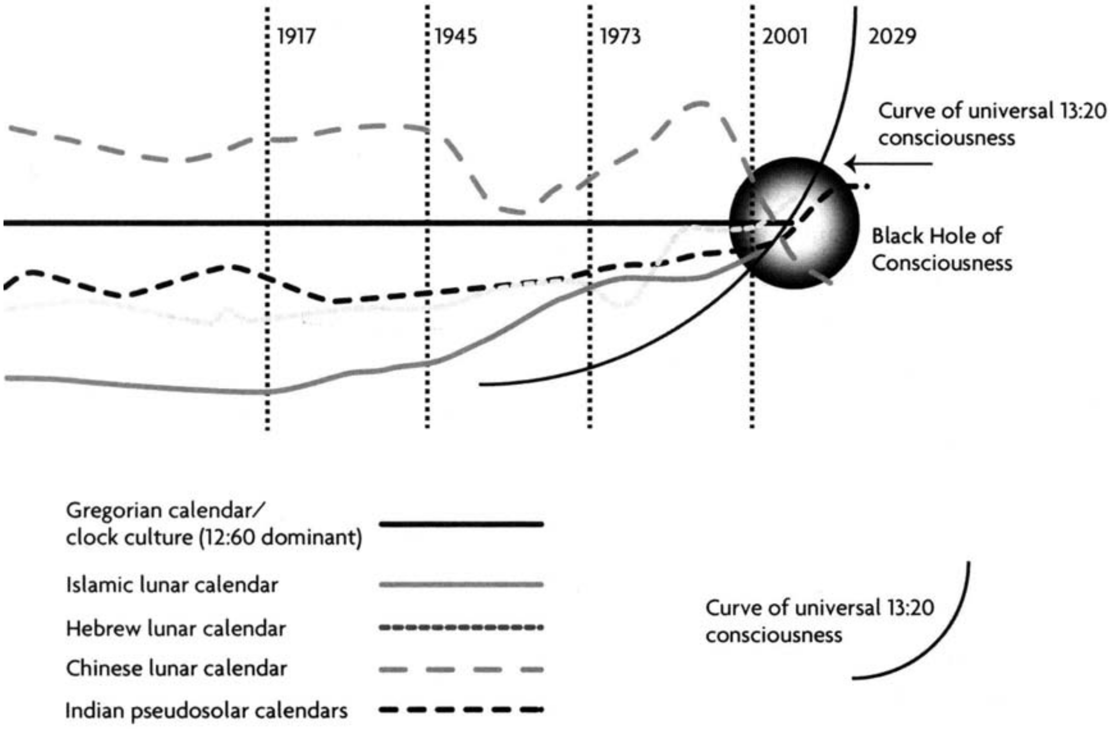

Now the default of the Christian program of the Gregorian calendar and its system of social organization, global civilization, is erupting in an entropic black hole of time. The result of this black hole of time is an entropic harmonic convergence, where all of the conflicting points of the differing timing programs that the Gregorian calendar attempted to synchronize, come into a common point of hostile impact. Seen in this light, the Final War is really the final battle in a long saga that we can properly call the Time Wars.

THE AGE OF THE TIME WARS

The Time Wars began with the application of the error in time at Uruk: the 24-hour, 60-minute, 60-second timing frequency derived from a 12-part division of the circle, a two-dimensional plane in space. Since this measure does not correspond to the reality of time as the universal (13 :20) frequency of synchronization, it creates a mental dissynchronization that manifests as the battle of the mind against itself, and against nature. The dynam'ic of the "progress" of civilization is really the dynamic of dissynchronization inherited by the human mind. The gradual encroachment of civilization into all aspects of human society and the biosphere increases the incidence of dissynchronization. The resulting aphasic mental states lead to belligerence and various attitudes of dominance over others and nature-but never to any satisfaction.

Viewed by the Law of Time, civilization is the institutionalization and embodiment of the dissynchronous artificial timing frequency. The dynamic of dissynchronization results in increasingly artificial technological means and reliance on the artificial medium of money to replace the biological or organic functions of our nature and of the biosphere in general. Not only does this process tend toward an "improvement" in the accuracy of the means of war, but also in the capacity of the latter for greater mass destruction. Set apart from the natural cycles by the dissynchronous frequency of mechanization, the Time Wars are also waged against the biosphere, which is viewed as a hostile force to be overwhelmed and plundered. Everywhere on Earth, nature is on the run. Everywhere on Earth, the human looks at his watch-what will he do next? Catch another plane? Fire another bullet? Press another button and let loose another bomb?

The 56-year cycle of the technosphere witnessed the greatest creation and stockpiling of lethal weapons, more than in all the previous history put together, enough to destroy the Earth many times over. The 56-year cycle of the techno sphere also witnessed the greatest plundering thus far of the biosphere, more than in all the previous cycles of history put together. Continued at the current rate, the biospheric plunder would bring extinction to virtually all life within another generation. These two facets of self-destruction-technospheric and biospheric-point again to the Inevitable Event. It is the Pentagon that has led in military spending and strategizing, just as it was the activity within the World Trade Center Twin Towers that dominated and coordinated from afar the plunder of the biosphere-both precedents for the consequent construction and expansion of the technosphere.

But now time has run out on the technosphere. The world has gone flat and valueless with no future vision beyond war. All the money in the world will not put it back together, nor will all the bombs in the world salvage it against its enemies. Who are the enemies of the technosphere, and what would cause thereto be enemies against the technosphere? And who or what is there that speaks for the biosphere? For if the biosphere is to become the noosphere, it must come about from an understanding of the intelligence of harmony. Clearly, such understanding of harmonic intelligence has not been nor will it be forthcoming from the technosphere-or from the Gregorian calendar that programs it. Only the Law of Time can supply the harmony necessary to complete this great movement from biogeochemical combustion to noospheric mind. The Law of Time replaces all that is artificial, while confirming all that is true, whether in human history or the universe.

Let us return to the analysis of the Time Wars and the clash of timing programs. The Twin Towers were the absolute triumph of the artificial 12:60 timing frequency, whose program is embedded in the Gregorian calendar and paced by the mechanical clock. Who steered the hijacked planes into their Gregorian targets? Nineteen militant Muslims, we are told. Why nineteen? And what are the religious backgrounds of the twenty-two most-wanted terrorists in the world? They are twenty-two Muslims. Why that number, and why are all the terrorists Muslims? In the noospheric analysis, the Inevitable Event was the suicide of the technosphere unable to sustain two clashing time programs. It was a program of the strict Islamic twelve-month lunar calendar that drove its belief system into the Twin Towers and the Pentagon, the two supreme monuments of the program of the twelve unevenly measured months of the Gregorian pseudosolar calendar. Is it Islam that is the enemy of the technosphere? Is it Islam that is the champion of the biosphere? Indeed what is Islam? Why is Islam the most resistant force against globalization and the technosphere? We know where the Vatican, the bastion of Christendom, is and who ordained the Gregorian calendar, but what do we know about Islam?

If we strip history away from the force known as Islam (1.3 billion Muslims in the world), and reduce it to its most essential point, then Islam is the teaching of the Holy Quran. What is the Holy Quran? It is the last whole revealed text for all humanity, received and recorded over a 23-year period by a single human being, Muhammad the prophet (570-632). While militant Muslims and terrorists have dominated the mass media image ofIslam for the past few decades, what is the Quran and what does it actually teach? It is also a function of the Gregorian program that scarcely anyone in the West really knows or is familiar with the Quran, since the West has been at odds with Islam since the time of Muhammad. As the supreme text of monotheism, the Quran is actually a psychoactive book that upholds Islam as the final religion for humanity. "He is the one who sent His messenger with the guidance and the religion of truth, to make it prevail over all other religions. God suffices as a witness." (Quran, 48:28) *Islam* means "peace, which is submission to the will of God."

Islam is also the law of nature. Within the vast system of intricate, mathematically based logic that constitutes the Quran-which is a profoundly ahistorical text as well-one feature stands out: the description and definition of nature and the natural order as the chief proof of God's design and guiding wisdom. From the Quranic point of view, the biosphere, inclusive of the creation of the human, is the overwhelming single proof of God's wisdom in establishing a life-support system for our indwelling consciousness to evolve our bodies on this Earth. The biosphere is Islam, or "peace, which is submission to the will of God," understood as the divine law of nature.

> In the creation of the heavens and the earth, the alternation of night and day, the ships that roam the ocean for the benefit of the people, the water that God sends down from the sky to revive dead land and to spread in it all kinds of creatures, the manipulation of the winds, and the clouds that are placed between the sky and the earth, there are sufficient proofs for people who understand. (2:164)
>
> He created you from one person, then created from him a mate. He sent down to you eight kinds of livestock. He creates you in your mother's bellies, creation after creation, in trimesters of darkness. Such is God your Lord. To Him belongs all sovereignty. There is no other god beside Him. How could you deviate? (39:6)
>
> Among His proofs is that you see the land still, then as soon as we shower it with water, it vibrates with life. Surely the One who revived it can revive the dead. He is Omnipotent. (41:39)
>
> He created the human from a tiny drop, then he turns into an ardent opponent. And he created the livestock for you, to provide you with warmth and many other benefits as well as food. They also provide you with luxury during your leisure, and when you travel. And they carry your loads to lands that you could not reach without a great hardship. Surely your Lord is Compassionate, most Merciful. (16:4-7)
>
> He sends down from the sky water for your drink and to grow trees for your benefit. With it, He grows for you crops, olives, date palms, grapes, and all kinds of fruits. This is sufficient proof for people who think. (16: 10-11)
>
> And He committed the sea to serve you; you eat from it tender meat, and extract jewelry which you wear. And you see the ships roaming it for your commercial benefits, as you seek His bounties, that you may be appreciative. And He placed stabiliz-

> ers (mountains) on earth lest it tumbles with you, as well as rivers and roads that you may be guided. And landmarks, as well as the stars to be used in navigation. Is One who creates like one who does not create? Would you now take heed? (16:14-17)

> As for the Earth, we constructed it, and placed on it stabilizers (mountains) and we grew on it a perfect balance of everything. We made it habitable for you, and for creatures you do not provide for. There is nothing that we do not own infinite amounts thereof. But we send it down in precise measure. And we send the winds as pollinators, and cause water to come down from the sky for you to drink. Otherwise you could not keep it palatable. (15:19-22)2

Viewed from its own context, the Quran is the criterion for evaluating the spiritual progress of human existence on this Earth inclusive of man's ability to read the proofs of nature as manifestations of God's wisdom and to comport himself within this nature, the biospheric life-support system, in a manner consistent with the truth. From the Quranic perspective, human freedom of choice is at the root of the biospheric destruction. Choosing anything but to be submissive to God, the human is in rebellion against nature-both his own and that of the biosphere. "We have offered the responsibility [freedom of choice] to the heavens, and the earth, and the mountains, but they refused to bear it, and were afraid of it. But the human being accepted it; he was transgressing, ignorant." (33:72)

The result of this ignorance on behalf of his own aims and intentions rather than that of a respect for the divine plan (biosphere) of the Creator, accounts for the human situation today, at odds with and surrounded by an increasingly out-of-control biospheric degradation and corruption of natural time: "Disasters have spread throughout the land and sea, because of what the people have committed. He [God] thus lets them taste the consequences of some of their works, that they may return [to the right works]." (30:41) Thus, the current techno spheric destruction actually points toward a return of humans to the path of right works.

The forces that created the phenomenon of historical Islam are the enclosing of the Holy Quran within the rigid system of a twelve-month, non-circulating lunar calendar, compounded and supported by the development of the *hadith* and *sunna* (non-Quranic) traditions. A profoundly conservative social order, Islam is nonetheless activated and motivated by a holy scripture that enunciates an attitude toward nature that is the direct opposite of the 12:60 system that produced the technosphere. In this regard, the precepts of the Holy Quran-as of many another holy scripture-make it the "enemy" of the technosphere. From the noospheric perspective, historically and culturally, however, the assault on the nerve centers of the techno sphere could only have been effected by humans whose belief system was rooted in the Holy Quran, which had been at odds with the Gregorian civilization for centuries. In this way, the Quran was the noosphere's weapon used to bring about the circumstances that would lead to the resurrection of the biosphere-the noosphere itself. But only if man consciously chooses the right return path: synchronized harmony in time. The problem is not so much the rigid lunar calendar ofIslam, but the irregular global standard of the Gregorian calendar, which, instead of standardizing all measures and perceptions of time, fosters a human social disharmony and an incessant antagonism against nature-and against Islam.

So then, how are we to return to the right works? Read and study the Quran, yes-but also make a fundamental change in the world. The solution of the Law of Time is: Get a new calendar! Humanity must take a long, hard look at its own timing systems and scrap what needs to be scrapped, no matter what kind of temporary inconveniences this may cause. Calendar reform is an issue that upsets many apple carts. Yet it is the only method that has not been commonly tried and the only step that has not yet been collectively taken. It is a characteristic of the dominant mindset that it writes history in the way that suits its own needs and supports its philosophy. In any of the reviews of the century carried by the mass media during the past year, you will find no mention of the World Calendar Reform Movement; nor will you find, for instance, any mention of the Roerich Peace Pact (1935); nor of the Harmonic Convergence-all buried by the marketing of history to suit the needs of globalization. What do these three phenomena have in common? A concern for peace based on an understanding that without harmony, the human cultural situation will only worsen. If humankind suffers from an excess of disharmony, in the noospheric analysis the only solution is a harmonic standard of time to synchronize all humanity simultaneously-the Law of Time refers to this as the Thirteen *Moon!* 28-Day frequency shift.

CHANGE YOUR CALENDAR, CHANGE YOUR MIND

> *"Since the advent of the Nuclear Age, everything has changed but the way people think, thus, we drift towards unparalleled catastrophe."*
>
> -ALBERT EINSTEIN3

To change how people think you must take a profoundly simple and universal element of everyday life, one in which all everyday thinking is rooted, and change it so radically that the way in which people think will itself be dramatically changed. If you want to know that something has really changed, change the calendar! This is the meaning of the replacement of the Gregorian calendar by the Thirteen Moon calendar. The calendar change is necessary because, as the pragmatic application of the discovery of the Law of Time, it brings into focus the essence of this discovery: Time is a frequency, the frequency of synchronization. If a calendar does not increase synchronization, it is not performing its highest function. This is the fundamental critique of the Gregorian calendar-and of all concepts of time based solely on physical, third-dimensional astronomical measurements. This discovery is so new and startling that it affects all human thought and is a matter to which all belief systems, religions, and methods of science must be cognizant. Yet nothing, it seems, is more difficult for humanity than to make this change from a manifest disharmony to a manifest harmony.

Historical man recoils at the prospect of actual harmony-harmony in natural time. This is because historical man defines his existence and thrives upon the disharmonies of his own invented time. The Gregorian calendar keeps the human mind entrained in a diabolical disorder of meaninglessly named months of uneven measure, while every day is tracked and driven by a relentless mechanism called a clock. This immersion of the mind of historical man in the frequency of his own artificial time, more than any other factor, condemns him to a world of inescapable horrors: traffic jams, poverty, terrorism, unresolvable historical and territorial disputes, global warming, environmental degradation, social disorder, insanity, and drug abuse. The problem of historical *human-Homo historicus-is* compounded by the unconscious nature of his acceptance of this timing sensibility, believing it to be the actual nature of time. Hence, all his public and economic policies are driven by and determined by the relentlessness of this perception that time is an arrow blindly pulling him into a future in which his only defense is the creation of more technology, the technosphere itself being a pure expression of the application of the mechanization of time.

THE SECRET HISTORY OF CALENDAR REFORM

When you went to bed on October 5,1582 you woke up on October 16. In this way Pope Gregory XIII promulgated his reform of the Julian calendar, from which the Gregorian calendar differs not one whit, except for the precise working out of the formula for leap year. When this "reform" occurred, the time sensibility of historical man was already crippled and disabled. The Gregorian calendar, by its moment of historicity, bound that crippled time disorder into its final institutional form, assuring that not only European man, but eventually all cultures and races dominated by European thought and conventions would assume and accept this deformed and deforming standard of time.

The mechanical clock, though evolved over many centuries, was perfected just after the Gregorian Calendar Reform of 1582 and promoted by Europeans as an example of the human triumph over nature. So it was that the human mind assumed as second nature its own invented time, encoded as a frequency that we have identified as the 12:60, an unconscious mental ratio determined by the irregular twelve-month calendar and artificial sixty-minute hour. The combination of these two factors in this one timing frequency have consigned historical man to a hopelessly debilitating mental disorder that is referred to as modern civilization, or nowadays, postmodern. But postmodern is not yet post-historical. Postmodernism is just the final convulsion of the time disorder that afflicts the human race as a virulent mental disease and encapsulates man in the technospheric cocoon of his own making. The only cure for such a profound and pervasive disorder is absolute harmony, the very harmony from which historical man so recoils.

In this light, it is well to know that through the first half of the twentieth century a vigorous and well-organized calendar reform movement flourished. Such was the character of this movement to reform the Gregorian calendar that in 1923 the League of Nations called for proposals for reform. More than 500 proposals were received; by 1931, these were narrowed to three:

> The International Calendar Organization, represented by Mr. Broughton Richmond, promoted a perpetual calendar based on the principle of five-five 73-day cycles = 365 days-and the year divided into twelve months of five 6-day weeks each plus a 5-day cycle [The Thirteen Moon calendar also includes the program of seventy-three 5-day periods-overtone chromatics-integrated with the fifty-two 7-day weeks: 52 x 7(+1) = 5 x 73.]

> The International Fixed Calendar League, represented by Mr. Moses Cotsworth, also supported by Eastman Kodak and the International Chamber of Commerce, promoted a perpetual Thirteen Month/28-day calendar, which was first presented by Auguste Comte (1842-1849), inclusive of the "day out of time." Though it was not backed by a more thorough mathematical science, which has now been supplied by the discovery of the Law of Time, this was by far the favored calendar, as it is the one that possesses a perfect structural harmony and obvious regularity of all its parts.

3. The World Calendar Association, represented by Elizabeth Acheles, promoted the World Calendar, a modified but perpetual version of the twelve-month Gregorian calendar, which also included the principle of a "day out of time" ("null day"). This group turned out to be the longest-lived and the most vocal, and was responsible for the publication of *The Journal Of Calendar Reform (1931-32).4*

It is instructive to read some of the early texts on the topic of calendar reform to see how the same issue of harmony and order brought up by the Law of Time was at the root of the call for the establishment of a perpetual, harmonic calendar. The base argument, as presented in Alexander Philip's *The Reform of the Calendar* (1914), is stated as follows:

> Suppose, for example, that by some strange convention the meaning of the figures we employ in numerical notation were to change every year; suppose the figure which this year represents 2 were next year to mean 3, next year 4, and so on; suppose again that our weights and measures were to fluctuate in a similar manner. . . then we affirm, without fear of contradiction, that the whole fabric of science and the mechanical arts could never have been raised at all. . . we should in such circumstances have been compelled to rest content today with the very simplest and most primitive appliances. Yet, strange as it may sound, such are the conditions under which, in modem society, human action is organized. For what is the framework. . . by which we arrange our actions? It is no other than the scheme under which we arrange our time-in one word, our calendar. . . [thus] the disorganized state of all social arrangements is ascribable to the calendar. . . The dislocation of our calendrical arrangements is due to two distinct causes. . . the incongruity of the week; [and] the irregularity of the lengths of the months. . .5

The argument made is a profound one, and was the motivating basis of the calendar reform movement in general, no matter which calendar was being promoted. It is fair to ask: Why do we expect a rigorous standard of uniformity in our measurements of space (size, weight, volume, etc.), but spurn or ignore the same requirement of uniformity in our measurement of time, especially the time that governs our everyday social arrangements and consciousness? How can we think that this would not have a profound underlying effect on the very nature and conduct of our society?

Most interestingly, as a movement calendar reform arose strongly during the proto-technosphere, between the two World Wars, only to be swept away by the clouds of war and the contradictory factions of disharmony in which modern civilization was already immersed. Any effort at altering or changing the direction and motion of modern civilization is treated as a threat or derided as ineffectual activity that is scientifically unprovable or that will cost humanity too many dollars to make the change. Such was the fate of the first great calendar reform movement that arose with such fervor in the late nineteenth century only to find its ignominious end when the final debate on calendar reform was adjourned indefinitely by the United Nations Social and Economics Committee in 1956. No one reads about the calendar reform movement in the history books. Because debate on the matter was effectively adjourned, it is assumed that calendar reform is a dead issue of no relevance to the development of postatomic, technospheric man. On the contrary, seen in the dim light of the still-smoking ruin of the Twin Towers, the calendar reform arguments concerning humanity's irrational behavior and its relation to the "immorality" of a skewed timing measure, the calendar, as well as to an inhuman timing device, the clock, were never more relevant.

Not the least import of calendar reform is the opportunity to create a new basis for unifying humanity. As was stated in the "Resolution Adopted by the Fourth General Conference on Communications and Transit of the League of Nations," which took up the issue of calendar reform: "A great number of delegations expressed the opinion that any reform of the calendar could only be put into practice if it came into force simultaneously throughout the world, or at least in a very great majority of States, and it was for this reason that the study of this question had been placed under the auspices of the League of Nations."6

The tragedy of the human inability to agree on a matter so obvious to logic and reason is a testament to the already crippling effects of having operated so long under parochial, sectarian, and often disharmonic timing standards. While the International Chamber of Commerce, the League of Nations, and various international scientific bodies supported the first effort at reform in 1914-1956, the second movement, 1989-present, arose as a purely populist effort, but with an entirely different information base. Recognizing instinctively that the calendar reform issue is more than merely a matter of mathematical order in relation to natural cycles, it is a profoundly theological issue as well, the second reform movement also developed from a non-Western inspiration: the Mayan calendar.

While the psychology of the dominator civilization all but blurs out any other considerations of calendar besides the prevailing Gregorian one, an objective study of the matter demonstrates that the latter is a primitive instrument that eschews both natural cyclic order and any mathematical harmony. In place of the anachronistic standard of the Gregorian calendar, the second calendar reform movement raised a new criterion, the Law of Time, a whole system principle that integrates the mathematical and the moral, the theological and the psychological, and the economic and scientific issues that are embedded **in** the entire topic of calendar reform. Precisely for the reasons of historical man's unconscious habituation to his own disharmonies and the accelerating and cumulative effect of this ignorance on the disorder of his own social life as well as the natural world around him, the second calendar reform movement has developed with a much greater moral and prophetic urgency.

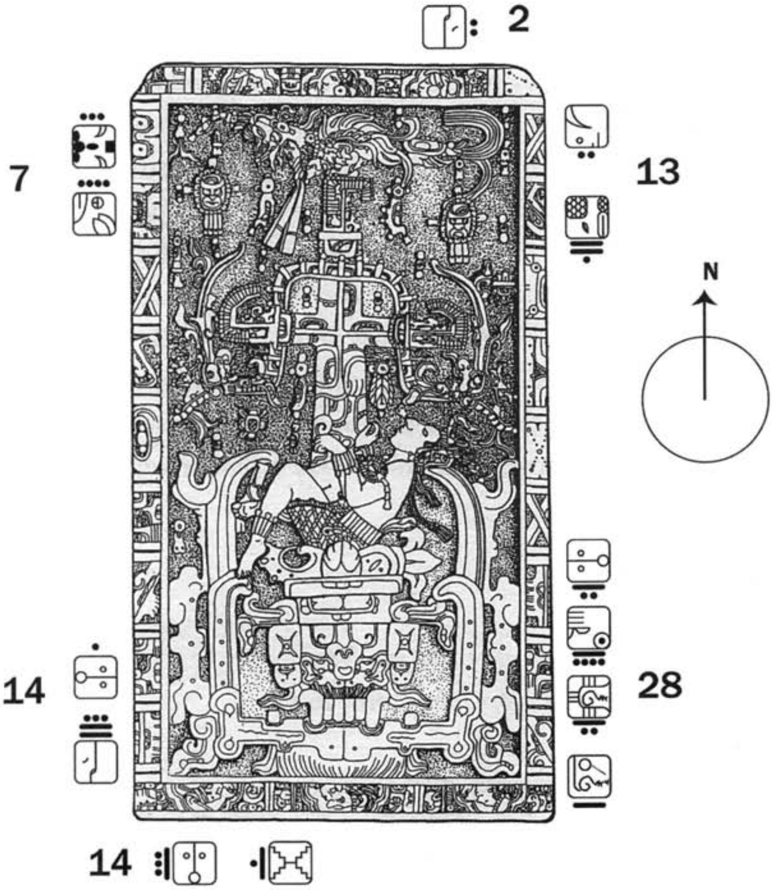

Western Edge:
Tonal sum, set 1 = **7**
Tonal sum, set 2 = **14**

Northern Edge:
Tonal sum = **2**

Eastern Edge:
Tonal sum, set 1 = **28**
Tonal sum, set 2 = **13**

Southern Edge:
Tonal sum = **14**

Total tonal sum = **78** (6x13)

Sarcophagus Lid of the Tomb of Pacal Votan Showing Disposition of Thirteen Clear Signs

The tonal sums (dot-bar code) of the thirteen signs are key to affirming the prophecy of the Thirteen Moon calendar. Note that the sums of the two sets on the eastern edge equal 28 and 13, while opposite the 13 set is the perfect 7. The tonal sums of the other two sets on the western and southern edge both equal 14, for a sum of 28. The northern edge has a single sign whose tonal sum is 2 twice 7 = 14, and twice 14 = 28; twice 13 = 26 fractal of 260 and code to the Gregorian date 7/26; while twice 28 = 56, the number of years from Hiroshima to the Inevitable Event.

"Consciously using the tool of the Thirteen Moon calendar to replace the erroneous measure of the artificial twelve-month calendar is, in itself, an unprecedented act of human self-reflective consciousness. Collective adaptation to the Thirteen Moon calendar is an intrinsic heightening of consciousness which rapidly awakens the telepathic programming inherent in human biology, frustrated for so long for not having been allowed development in the correct timing frequency."7

It was in 1987 that *The Mayan Factor: Path Beyond Technology* first brought to major public and international attention the possibility that a superior timing sensibility was developed by a thoroughly non-Western culture, the Maya. This book was published with scarcely any knowledge of the calendar reform movement that had gone into a debilitating coma in 1956. Yet, by pursuing the mathematical base of the Mayan calendar, the essence of the Mayan achievement, the Law of Time was extracted, a discovery that brought on the second calendar reform movement, the World Thirteen Moon Calendar Change Peace Movement. Prompted by a prophetic revelation concerning the Thirteen Moon!28-Day calendar decoded from the tomb of Pacal Votan, the premise of the second calendar reform movement was tested worldwide from 1993 to 2000, a time defined as the "seven years of prophecy." By the year 2001, the official beginning of the third Christian millennium, a loose, far-flung calendar reform movement, pioneering in the development of a new science of time and consciousness was promoting the Thirteen Moon!28-Day calendar worldwide.

A REVOLUTION IN TIME

> *"Fire in the lake, the image of revolution. Thus, the superior person sets the calendar in order and makes the seasons clear"*
>
> -*I CHING,* HEXAGRAM49, "REVOLUTION"8

By common consent, a calendar is a system for reckoning time. Virtually all current time-reckoning systems take account only of chronological or, better put, astronomical time-the movement of the Earth in relation to celestial bodies, the sun, moon, planets, stars, and constellations. This is a purely third-dimensional, physical consideration of time. The Law of Time affirms a higher order of time-the synchronic order. This is the time of the fourth dimension, which includes the chronological but enfolds it in a higher mental and mathematical order of reality. According to the Law of Time, what you don't know about the calendar you are using could kill you. By operating according to purely third-dimensional chrono-astronomical timing standards and without knowledge of the synchronic order of fourth-dimensional time, humanity subjects itself to a one-sided view of reality that is hazardous to the planet and jeopardizes the future existence of the human species as well as all life on Earth.

To correct this situation a radical solution is required-a new calendar embedded in the synchronic order revealed by the discovery of the Law of Time. This is the significance of the Thirteen Moon/28-Day calendar-it epitomizes the incorporation of the chrono-astronomical order of time into the synchronic order, and hence provides the vehicle for humanity to escape its otherwise certain plunge into disaster within the next decade. When a dogma is encountered while one is engaged in the pursuit of truth, it must be challenged and abolished. Such a dogma is the calendar in use as the world standard, the Gregorian calendar. That this calendar, originally a system of thought peculiar to one people or religion, should dominate all the peoples and even life of the planet makes it subject to the critique of planetary anthropology.

The discovery that time is a frequency mathematically expressed as the ratio 13:20 requires a profound reorganization of all thought about time and the consequent reorganization of human society as a planetary organism. As the unified field theory of time, the Law of Time calls for the human synchronization in time predicated on the establishment of the Thirteen Moon calendar as the global civil standard replacing the current erroneous timing standard, the Gregorian calendar. This is a matter of the utmost seriousness, since it is also recognized that all belief systems are the function of programs locked up or embedded in the calendars or time-reckoning systems used by a people. The fate of the species and the biosphere may well depend on whether or not humanity is able to understand this point.

What is it about the calendars in use that condemns their users to endless conflict and inability to unify as a species? What is at the root of their disharmony? It is the pursuit of astronomical time. All people today are bound by calendars based on a pursuit of accuracy of astronomical time-the length of the year being the object of this pursuit. Astronomical time is a losing game of an ever-changing and elusive order that results in an entropic reductionism that has little to do with consciousness or synchronization. The year's length (measure of the orbit of the Earth around the sun) calculated by Pope Gregory XIII at 365 days, 5 hours, 48 minutes, and 20 seconds, slows at the rate of 112 second per century, while the length of time in the Gregorian calendar is off from the "true" solar year by 25.9678 seconds per year! Obsessed with rectifying this discrepancy, in 1972 reductionist science came up with Atomic Time to replace Earth Time, and the atomic cesium clock replaced the measure of the year as 365.241299 days with 290,091,200,500,000,000 oscillations of Cs per year!

But what does this staggeringly astronomical figure mean, and of what value is it?

In profound contrast to this obsessive reductionism in which there is no place for spirit, mind, or consciousness, fourth-dimensional radial mathematics represents a mental order greater than and encompassing the astronomical order of the third-dimensional universe. According to the Law of Time, in time reckoning, consideration should be given to systems of the harmonic order of synchronic time that proceed from the radial mathematical order of the higher mind. The higher mind represents the evolved state of mind-the noosphere-that evolution on Earth tends toward, but that has not yet been attained. As long as the human concept of time is governed by the obsessive pursuit of the impossible to attain perfection of astronomical time, this pursuit and the civilization that it propels are doomed to failure. What is known as history is embedded in this pursuit of "fixing" astronomical time, which has culminated in the dogma of the Gregorian calendar and the microatomic measure of the cesium clock. When did this pursuit begin, why did it come to dominate the human species, and to what effect?

As we have seen, the values, customs, and norms of a nation, culture, or civilization are embedded in the calendar it uses. The Gregorian calendar, the current global civil standard-indeed, the foundation in time of globalization-is the calendar of the Caesars, of the Roman Empire, of the Catholic Church and the Vatican. To change this calendar is to change history itself. Can it be done? The League ofNations attempted this in the 1930s but failed. The United Nations tabled this effort in 1956. Since the United Nations adjourned the matter of calendar reform indefinitely, in 45 years the Earth's population has more than doubled, that is, it has increased by more than 3.2 billion people, a factor that is still out of control, and that may have much to do with the fact that humanity is actually in a fog about time.

The reasons for wanting to replace the Gregorian calendar are no less valid today. The Gregorian calendar is not a standard of measure. It lacks logic and reason. It numbs and befuddles the mind when trying to make calculations by it. It is shrouded in an arcane and medieval obscurantism. The net effect of the use of this calendar is to perpetuate a fundamental level of mental confusion and ignorance concerning the actual nature of time itself-an ignorance that is hardened into dogma by the unwillingness of habit to consider any other possibility, but rather to accept the entire system as second nature. The argument that it encodes the most accurate and scientific measure of the solar orbit of the Earth is irrelevant in relation to the effect of its irregular measure on the mind and its numbing power as the dogma of a conquering people. Indeed, the solar measure of365.241299 days per year has nothing intrinsically to do with "30 days hath September. . . ," and vice versa. In fact, the measure of the Earth's orbit is one matter; the purpose of calendars and time reckoning as factors of synchronization is a wholly other matter. But then, since the very nature of a culture or a civilization is determined by the calendar it keeps, the use of a deformed standard can only inhibit and skew even the questions one asks about time.

The failure of calendar reform by 1961, the year of the Vatican II Council, only promoted an increasingly complacent acceptance, however cleverly disguised, of a millennial dogma, the Gregorian calendar. In the conclusion to the "last word" on the subject, *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year,* David Duncan summarizes, "We take in stride a calendar [the Gregorian] that is flawed but endures, largely because it works just fine for most of us, and it is what we are used to."9 It is not at all objective to accept something because we are used to it. This attitude is anything but scientific. Rather, it is the smug basis of nothing more than a self-fulfilling dogma-and yet the civilization that bases itself on this aberrant timing measure prides itself as being the most scientific in all history!

THE CONFUSION OF IRREGULAR TIME

Due to this flawed time sensibility, it may be concluded that humankind today is little capable of moral reason or logic, which itself is a result of, or at least strongly reinforced by, habituation to a timing standard that is irrational and illogical. Anytime we overlook or dismiss a fault due to habit, or simply because we are accustomed to the use, say, of an instrument, even if that instrument is demonstrated to be flawed and irregular, is evidence of a moral laziness that will eventually contaminate the entire consciousness that accepts and accommodates itself to the flaw. The exercise of truth is a moral prerogative. The fact of the matter is that the Gregorian calendar is a hodgepodge of flaws and accretions, rationalized into a systematic formulization that has no basis in reality. The celebration of Easter is a case in point.

Easter is the celebration of the day Christ was resurrected from the dead. It may be logically assumed that this occurred on one and only one particular day, just as the celebration of his birth occurs on Christmas Day (although, similarly, no one really knows that this was the actual date of his birth). By the time the Christian religion had become an established force within the Roman Empire, however, no one knew precisely on what day the Resurrection had taken place. The Council of Nicea, convened by Constantine in A.D. 325, called for a debate on the matter in which various theories were presented. The debate was concluded with the victory of one conjectural system over the other. The winning system stated that Easter would be celebrated on the first Sunday after the first full moon after the spring equinox, except when the full moon was on a Sunday, in which case it would be celebrated on the following Sunday, and this was so that it would not be celebrated on the same day as Jewish Passover.

Over the centuries this conjectural theory regarding the date of Christ's resurrection was buttressed with a system of elaborate calculations for predicting when Easter would occur. These calculations were all made with regard to establishing an alignment of the synodic lunar cycle of29.5306 days (which resulted in a lunar year of 354.3672 days) and the solar year of 365.241299 days! Being fortified with such elaborate mathematical calculations did not lessen the fact that the system was originally-and still is-a mere speculation of a highly clever but relative nature. Thoroughly embedded in the Julian-Gregorian calendar, the system of determining Easter is inseparable from the Gregorian calendar itself, and is one of the chief reasons for the resistance to its reform-as witnessed by a reading of the calendar reform arguments of the 1930s. The entire system of this calendar, inherited without question from the Caesars of the Roman empire, is actually nothing more than a form of self-perpetuation of a priestly hierarchy that includes in its support a host of bankers, industrialists, and scientists and the elaborate social system that is under their control-the technosphere of global civilization itself. Take away the calendar and the entire structure loses its foundation.

"If the Julian/Gregorian calendar was to be offered as a new device for measuring time, we, with our present knowledge and state ofliving, would reject it as something utterly impracticable, lacking in harmony and order, unbalanced and irregular, too clumsy a calendar to make calculations by, while the various sections are not comparable."10

A science or any scientist that unquestioningly follows the Gregorian calendar is, in reality, not worthy of the name. What is science? A concern for logic and precision of measure, we might answer, as well as standards of measure that utilize uniform units of measure in accord with what is being measured. Yes, the year is calculated as being 365.241299 days in length, but if the annual standard of measure that is used is irregular and unscientific, then it avails nothing, and indeed deforms the mind leading it down byways that may only end in self-destruction. Thus the pursuit of a true and accurate year in itself may be an illusory pursuit, blinding us to the actual nature of time, and leading us away from a genuine understanding of ourselves and our role and purpose on this Earth.

To change and replace the calendar with the thirteen moon standard is to return us to our original purpose, leading us back to pathways of harmony and natural health. According to the Law of Time, the damage already inflicted in the time humanity last hesitated and lost the chance to alter its timing frequency, forty years ago, has been incalculable. The final opportunity to change the calendar and the timing frequency is now upon us. Because of this we must be very clear and unshakable in our understanding and determination to expose and eliminate the current civil calendar for once and for all. Even apologist David Duncan's unabashedly pro-Gregorian book *Calendar* (1999) concludes with its time line on the date 2012, "Current Mayan Great Cycle will end."[11] But will we also end? By 2012, it will be too late to change. We must begin to change now.

In reviewing the quirks and twists of history that constitute the Gregorian calendar, we must ask: Why do we continue to use such an instrument, and what are its effects? Is a calendar something more than a tool for scheduling the payment of debts *(calends),* or is it an instrument of synchronization? The harmony or disharmony of time is a profound matter rooted in the instrument of time reckoning that we use. There can be little question that we live more in a time of chaos than of harmony. We may also say in regard to the effect of standards of measure on the mind, that the chaos of the time is embedded in the calendar we use. If we are to leave this time of chaos and enter a time of harmony, then we must exchange the instrument in which the chaos is embedded for an instrument that is the very model of harmony-the Thirteen Moon/28-Day count. This is the choice humankind must now make.

The issue of the nature of time, or of the times in which we are living, cannot be separated from the issue of the calendar. The very nature of the calendar that the world follows has stunted the mind and warped the body's innate timing sensibility, reducing the understanding of time to being merely a matter of chronology. "Chronology is a register or reckoning of successive years, a time scale. It may also be described as a system of registering time massively."[12] Virtually all time reckoning has proceeded from astronomical indicators, the major exception being the Maya, who actuated from a solely mathematical outlook. Yet with this mathematical outlook the Maya devised a chronological method that has never been surpassed, a fact attested to by many, including the architects of the Kitt Peak Observatory in Tucson, Arizona, where a mosaic mural depicting the civilization of the Maya asserts the superiority of their calendar over that of the Europeans.

It is significant that the earlier calendar reform movement, while occasionally acknowledging the brilliance of the Maya, made no use of the Mayan knowledge. But this was because the time sensibility was restricted to thinking of time solely in terms of chronology, or at best as a schedule for keeping good accounts, with little comprehension at all of the synchronic nature or order of time. And since the system of chronology was synonymous with the system of control embedded in the Gregorian calendar system, the earlier reform movement failed altogether to make any impact on the Vatican, much less on the global power structure itself.

All the thought, action, and good intention put into the years of calendar reform notwithstanding, the Gregorian calendar prevailed, and, tragically, not one of the proposals or recommendations was ever even considered for adoption by the papal authorities who maintain control over the calendar. In the 1930s, the propaganda of the Church was aimed at limiting any possibility of reform, for it would not give up the succession of the seven-day week. In any perpetual calendar, such as the Thirteen Moon/2 8-Day calendar, the observance of the Day Out of Time is necessary for maintaining the harmony of the fifty-two weeks that complete themselves in 364 days. The continuous succession of the week does not stand in a constant relation with the other elements of the calendar and is one of the main causes of confusion in the Gregorian calendar. Arguing that such a break as a Day Out of Time, no day of the week at all, would plunge the world more deeply into chaos, barbarism, and war, the Vatican effectively blocked all effort at reform. The completion of the century of total war, capped by the Inevitable Event, roundly refutes this point. It may just as easily be argued that because we did *not* change the calendar, we are now plunged into the final barbarism and unholy war.

When the Vatican II Council in 1961 affirmed in its backward and highly restrictive language that it was not opposed to reform, so long as such reform did not disrupt the seven-day week and respected the tradition of Easter as a "movable feast," the door to reform slammed shut. By this time all the early reformists had literally died out and the question of calendar reform appeared to have become one of the buried issues of history.

What the priests of the Catholic Church thought they had buried through an auto-da-fe in Izamal, Yucatan, 1562, and overcome through the imposition of the Julian-Gregorian calendar over the conquered Maya, returned with the precision of prophetic timing in 1987. The publication of *The Mayan Factor* not only opened the door to a new look at the Maya, but also to an understanding of time that was anything but chronological. A new dimension of time appeared-radial, fractal time, the synchronic order. And behind the reassessment of the nature of time was the provocative call of Mayan prophecy-the end of the thirteen baktun Long Count, 2012.

The Law of Time defines the human problem as being one of radical and destructive dissynchronization from natural time, resolvable only by a return to natural

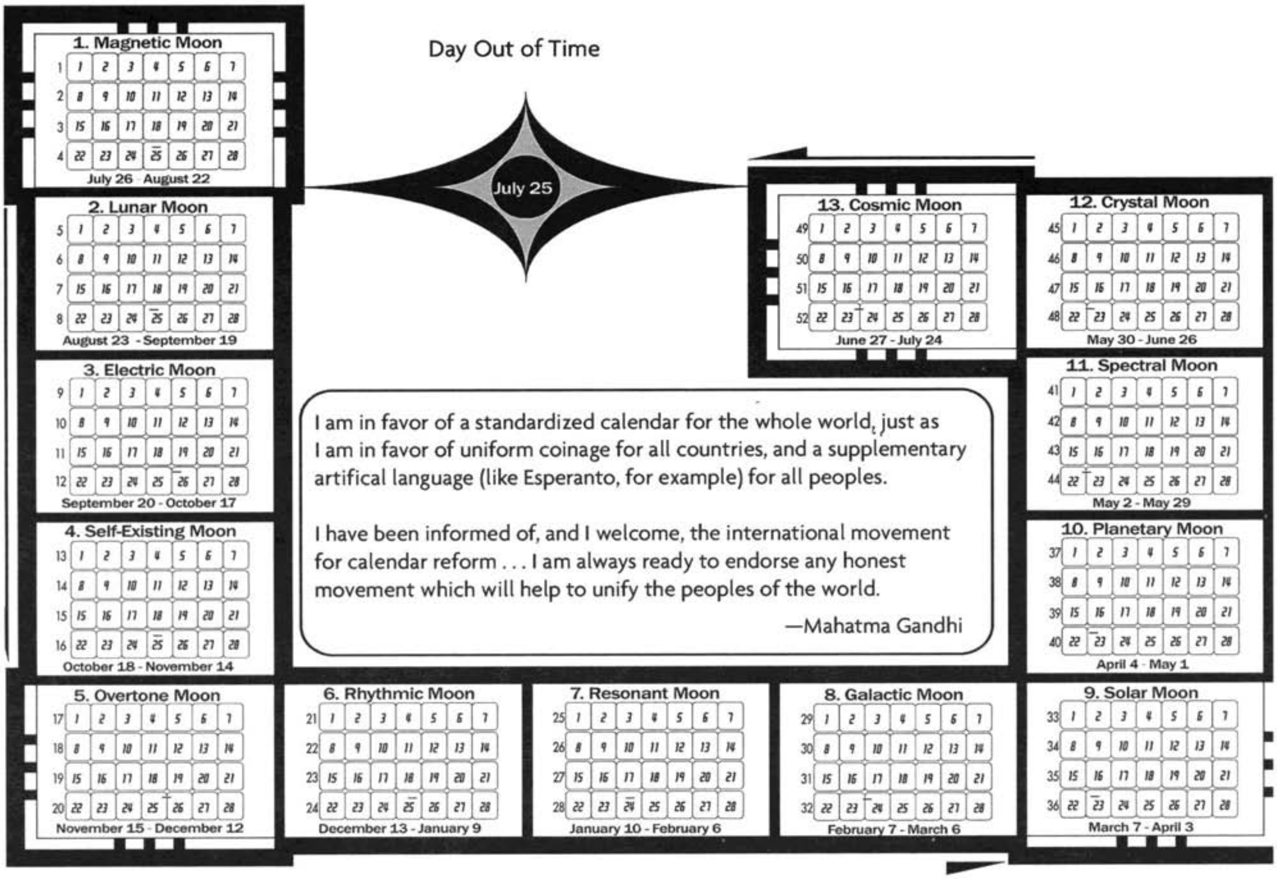
Day Out of Time
July 25

1. Magnetic Moon

1	1	2	3	4	5	6	7
2	8	9	10	11	12	13	14
3	15	16	17	18	19	20	21
4	22	23	24	25	26	27	28

July 26 - August 22

2. Lunar Moon

5	1	2	3	4	5	6	7
6	8	9	10	11	12	13	14
7	15	16	17	18	19	20	21
8	22	23	24	25	26	27	28

August 23 - September 19

3. Electric Moon

9	1	2	3	4	5	6	7
10	8	9	10	11	12	13	14
11	15	16	17	18	19	20	21
12	22	23	24	25	26	27	28

September 20 - October 17

4. Self-Existing Moon

13	1	2	3	4	5	6	7
14	8	9	10	11	12	13	14
15	15	16	17	18	19	20	21
16	22	23	24	25	26	27	28

October 18 - November 14

5. Overtone Moon

17	1	2	3	4	5	6	7
18	8	9	10	11	12	13	14
19	15	16	17	18	19	20	21
20	22	23	24	25	26	27	28

November 15 - December 12

6. Rhythmic Moon

21	1	2	3	4	5	6	7
22	8	9	10	11	12	13	14
23	15	16	17	18	19	20	21
24	22	23	24	25	26	27	28

December 13 - January 9

7. Resonant Moon

25	1	2	3	4	5	6	7
26	8	9	10	11	12	13	14
27	15	16	17	18	19	20	21
28	22	23	24	25	26	27	28

January 10 - February 6

8. Galactic Moon

29	1	2	3	4	5	6	7
30	8	9	10	11	12	13	14
31	15	16	17	18	19	20	21
32	22	23	24	25	26	27	28

February 7 - March 6

9. Solar Moon

33	1	2	3	4	5	6	7
34	8	9	10	11	12	13	14
35	15	16	17	18	19	20	21
36	22	23	24	25	26	27	28

March 7 - April 3

10. Planetary Moon

37	1	2	3	4	5	6	7
38	8	9	10	11	12	13	14
39	15	16	17	18	19	20	21
40	22	23	24	25	26	27	28

April 4 - May 1

11. Spectral Moon

41	1	2	3	4	5	6	7
42	8	9	10	11	12	13	14
43	15	16	17	18	19	20	21
44	22	23	24	25	26	27	28

May 2 - May 29

12. Crystal Moon

45	1	2	3	4	5	6	7
46	8	9	10	11	12	13	14
47	15	16	17	18	19	20	21
48	22	23	24	25	26	27	28

May 30 - June 26

13. Cosmic Moon

49	1	2	3	4	5	6	7
50	8	9	10	11	12	13	14
51	15	16	17	18	19	20	21
52	22	23	24	25	26	27	28

June 27 - July 24

> I am in favor of a standardized calendar for the whole world, just as I am in favor of uniform coinage for all countries, and a supplementary artifical language (like Esperanto, for example) for all peoples.
>
> I have been informed of, and I welcome, the international movement for calendar reform . . . I am always ready to endorse any honest movement which will help to unify the peoples of the world.
>
> —Mahatma Gandhi

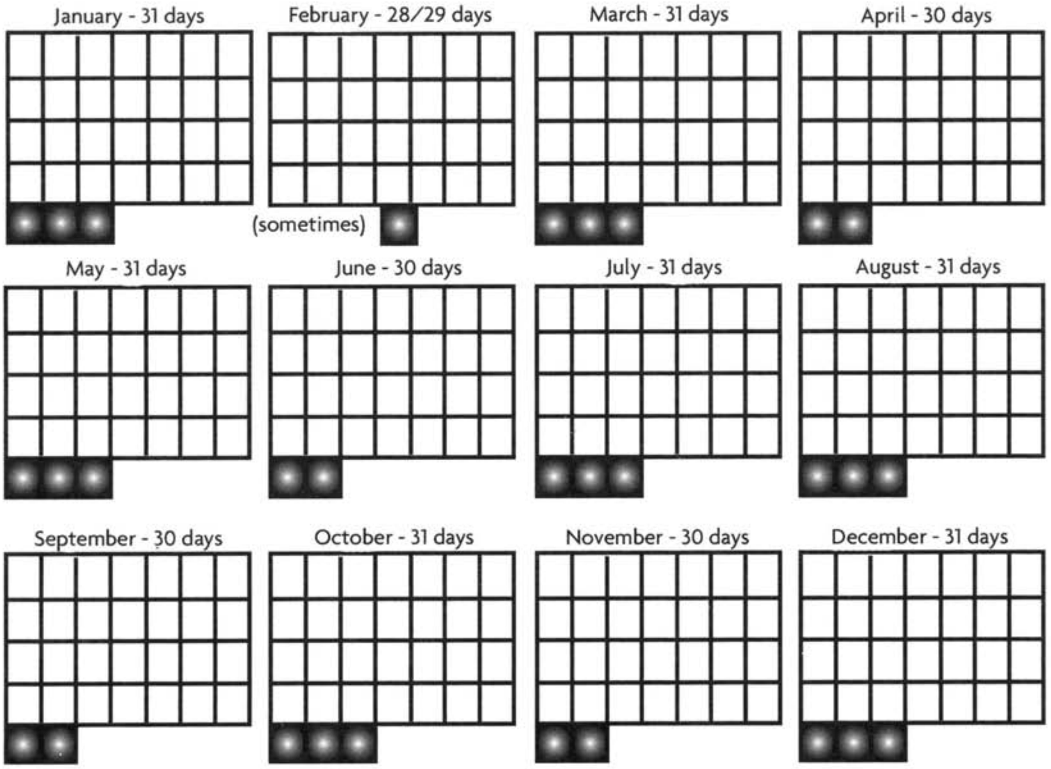
January - 31 days
February - 28/29 days
(sometimes)
March - 31 days
April - 30 days
May - 31 days
June - 30 days
July - 31 days
August - 31 days
September - 30 days
October - 31 days
November - 30 days
December - 31 days

Wavespell of the Thirteen Moons in Relation to the Irregular Gregorian Calendar

time through the perfect and perpetual harmony of the Thirteen Moon calendar. With the analysis of the Law of Time, there is an even greater urgency and a more profound certainty. Reform is not only insufficient but also impossible. If the calendar is not changed within the next decade, it will be the end of civilization as we know it. There is a genuine and legitimate need to bring the issue of calendar reform into the public eye once again.

Yes, to change the calendar involves a frequency shift. The collapse of the technosphere will dispose the human mind to rethink its options. It may even reorient the human mind to a consideration of the biosphere as being the greater vehicle in which we, as humans, are being carried. Such a massive change in attitude, a mind shift and a paradigm shift, may arouse in the human breast a spiritual reawakening as well. Harmony is spirituality. By changing the calendar, humanity can know it has turned a page of destiny-an act that may cause it to appreciate what has already been written in a book it had not been inclined to study, the Holy Quran:

> God is the One who raised the heavens without pillars that you can see, then assumed all authority. He committed the sun and the moon, each running in its orbit for a predetermined period. He controls all things, and explains the revelations that you may attain certainty about meeting your Lord. He is the One who constructed the earth and placed on it mountains and rivers. And from the different kinds of fruits, He made them into pairs-males and females. The night overtakes the day. These are solid proofs for people who think. (13:2-3)

In this passage we can find ample evidence of the formal order of the Law of Time and the biosphere-solid proofs. As for the "people who think," do they not then constitute the noosphere?

* * *

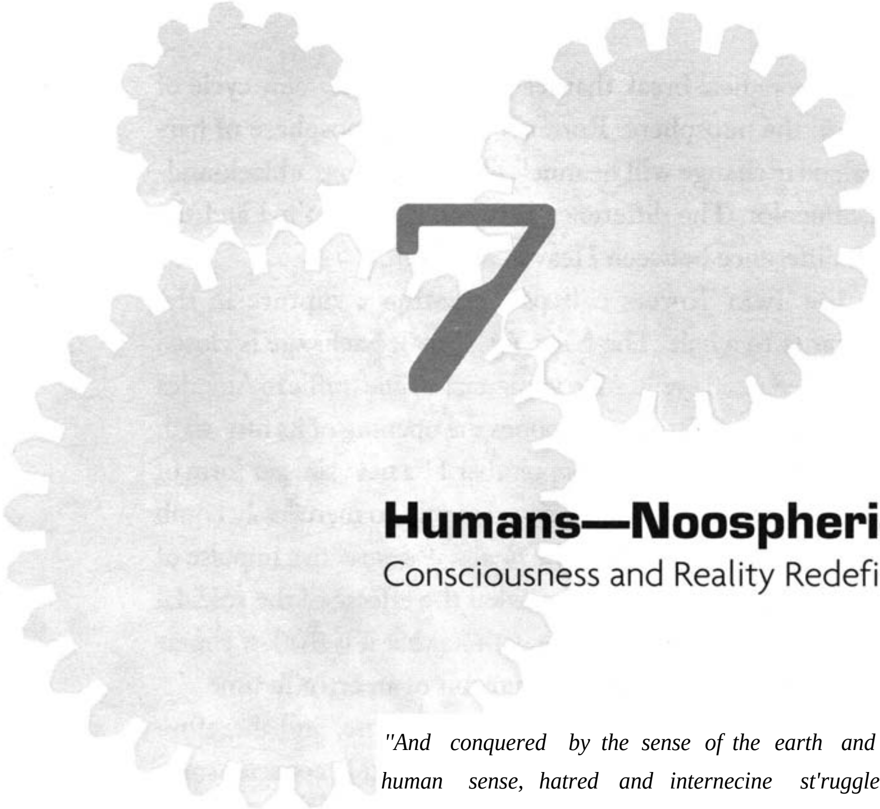

7

Humans—Noospheric Chips:

Consciousness and Reality Redefined

> *"And conquered by the sense of the earth and human sense, hatred and internecine struggles will have disappeared in the ever-warmer radiance of Omega. Some sort of unanimity will reign over the entire mass of the noosphere. The final convergence will take place in peace."*
>
> -PIERRE TEILHARD DE CHARDIN,
> *THE PHENOMENON OF MAN!*

Once begun, the exhaustion of the technosphere is irreversible. The exponential growth curves of money, machine production, energy output, and population increase come to a climax. By 2004 the humans will be at another major crossroads: to go wholeheartedly into the harmony of the noosphere, or to suffer the rest of their decline in an ignominious twilight of terrorism, barbarism, and cultural anarchy. By 2004 the humans will have their last opportunity to change their macro-organizing principle. By relinquishing the artificial time held in place by the old calendar and becoming unified in time under a calendar that is a perfect harmony, humans will enact the coming of the noosphere. Only harmony will bring harmony; only harmony will cancel the errors of history, a manifest disharmony. Only an act of harmony can manifest the noosphere. Only this time-changing act, consciously carried forward by a significantly roused number of humans, will create a mental field so positive and constructive as to engender the radical break that leads to an entirely new cycle of existence of the life on Earth: the noosphere. Entrance into the noosphere of harmonic time through the calendar change will be much like going from a black-and-white movie to one in Technicolor. The difference between the new time and the old time is literally like the difference between Heaven and Hell.

September 11, 2001: the Twin Towers collapse, creating a rupture in the technosphere. Everything comes to a halt. The New York Stock Exchange is closed for four days-a virtually unprecedented event. All commercial airline traffic in America is halted for two days. Even the United Nations postpones the opening of its fifty-sixth General Assembly by three weeks. Six weeks after September 11 a new, sinister form of war is engulfing the technosphere. The Americans assert the right to mercilessly bomb Mghanistan in self-defense. The response of war is the final self-destructive impulse of the techno sphere to protect itself, but this will only widen the effects of the suicidal blow it has already sustained. The war will only make definite that it is the last climax of the biogeochemical combustion engendered by the pursuit of an error in time.

The rupture in the technosphere is the break in artificial time, and therefore amounts to the opening of a noospheric fissure-an inrush of cosmic consciousness and a moment of simultaneously heightened natural time. When the towers collapsed, belief systems shattered. Because it was viewed on global television, it was an event witnessed by all humanity and it entered through the optical cortex into the brain, arousing potent images from the subconscious. This is a noospheric act because it is the entire mind of the whole species that receives this single image in one simultaneous moment. In that shattering image, however unconscious, the noosphere immediately entered the field of human consciousness, much as a leak in a boat allows water to come rushing in. From the perspective of the noosphere, the technosphere is a medium of transformation resulting from the pursuit of a distorted perception of time. With the destruction of the Twin Towers, the transformation is complete; following the war, the technosphere can be creatively dismantled.

THE NOOSPHERE: RAINBOW BRAIN OF TIMESHIP EARTH

What is the noosphere? Can we define the noosphere more precisely so that it is a less nebulous term? Can we make the noosphere concrete enough so that it may be understood and accepted as the coming new way of a higher level of everyday normalcy? First of all, let us say that the noosphere has existed as long as life has existed on this Earth. In this sense, the noosphere is the sum of the mental interactions of all life. The mental interactions, or field of mind, exist as a great unconscious medium, inseparable from the Law of Time. That is, if all life and the phenomenal world of the biospheric medium are synchronized by the Law of Time, then the programs of the organization of life, from the periodic tables to the DNA code, will also be synchronized by and embedded in the mathematics of the Law of Time. Being a function of the whole system of the Earth and integral to the functioning of the biosphere as an interplay of cycles in which organic and inorganic interact with a set order of chemical transmutations, the noosphere must also be the mental coordinating field or program of this complex of vital interactions. It must, therefore, be a characteristic of biospheric evolution that the noosphere remains unconscious until a significant moment is reached-the climax of the cycle of biogeochemical combustion, the enactment of the biosphere-noosphere transition.

As the Earth's mental field, it must be an act of collective self-reflection that arouses the noosphere into functioning consciousness. By its definition as the mental envelope of the Earth, the noosphere can come into conscious manifestation only through an act of collective self-reflection, such as the proposed calendar change. Also, as the mental envelope of the Earth that is above and discontinuous with the biosphere, the noosphere-mind sphere--can mean only one thing: the advent of universal telepathy among the humans. The Internet can be seen as a third-dimensional reflection of the noosphere, a form of electronic proto-telepathy. But, ironically, as long as the humans hold on to the Internet, they cannot experience the true evolutionary potential of genuine universal telepathy. The point is that telepathy can only function in a field of reality that is synchronized by the natural 13:20 timing frequency. Time and telepathy are mutually defining factors dependent on unity of thought. The humans have chosen a path of disunity, so that their calendar programs clash and are at odds with each other and with natural time. This maintains the technosphere, destroys the biosphere, and keeps the noosphere in the unconscious. But now, with the Inevitable Event, the break has occurred and the shattering of the Dreamspell of History has begun. The noosphere's rainbow brain leaks back into human consciousness.

The rainbow brain of the noosphere-what is that? As the mental envelope of the Earth, the noosphere possesses a structure ordained by the Law of Time and congruent with the structure of the whole Earth. Within the noosphere are to be found the evolutionary control panel or programs, defined as the *psi bank-the* medium of information storage and retrieval for all the mental programs. The noosphere is the field of planetary consciousness in which humans are intended to participate and complete their evolution. However, operating by artificial time is like looking through a glass darkly. This being so, there is no noosphere to see. Humans generally experience consciousness through the chinks in their sensory prism, conditioned by separatist beliefs, making it most difficult to attain a perception or experience the field of consciousness as coextensive with the Earth. Indeed, if the biosphere is the region on Earth for the transformation of cosmic energies, the noosphere is that region of the Earth for the reflection of cosmic consciousness and its mental programs. Seen through the lens of universal synchronized time, the humans accommodating to a single instrument of time measurement are inherently sharing a single thought form. It is that act of self-reflective convergence that makes the noosphere conscious and draws it toward the resplendent omega point of2012 and beyond.

As with all the other geochemical structures and processes of the Earth that are planetary in nature, the noosphere, the rainbow brain of Earth, actually has a coherence and an order that we may define in a precise way. In so doing we shall redefine the nature of human reality and the human being itself. We must enlarge our lens of perception to encompass the whole Earth as a single being. In relation to this single being of the Earth, the individual human is hardly even a speck. The dimensions and proportions by which we commonly measure reality must also expand to accommodate the vast dimensions of the Earth as a single entity, a planetary geocosm within the solar field, the heliosphere. Since the sun is also a star, the heliosphere is a function and member of the galactic order of being. As the mental "thinking sphere" of this planetary being, Earth, the noosphere is above and discontinuous with the biosphere, the hydrosphere, and the atmosphere. It is contained by and functions within the vast invisible electromagnetic field of the Earth. Though the ionosphere is the lowest level of the electromagnetic field system of the Earth, the main structures of this field are the two radiation belts located 2,000 and 11,000 miles above the Earth. It is within the two radiation belts that the control panels of the noosphere are to be found-the psi bank.

KIN: BIPOLAR BASE MEASURE OF THE 13:20 MATRIX

> *"And He is the One who created the night and*
> *the day, and the sun and the moon; each*
> *floating in its own orbit."*
>
> -QuRAN,21:33

In the simplest aspect of our daily experience of time on Earth we have the full play of the law of alternation-day-night, sun-moon, north-south. Because we know the bipolar nature of the Earth, dividing it into two geomagnetic fields, north polar and south polar, we can then define the complementary bipolar field of time, which divides the cycle of the Earth as it makes one rotation on its axis into one day and night. This entire unit, day and night, is defined as a *kin,* the base unit of synchronic time measurement. From the perspective of the noosphere, day and night are a function of the two alternators-day alternator and night alternator-that operate as the phasic motion of the rainbow brain of the noosphere working in tandem with the electromagnetic field. The two phasic alternators regulate the activation of photons (day) and the regeneration of photons (night) within the entirety of the life process of the biosphere.

THE NOOSPHERIC FOURFOLD/EIGHT-PART STRUCTURE

> *"He created the heavens and the earth truthfully. He rolls the night over the day and the day over the night. He committed the sun and the moon, each running for a finite period. Absolutely he is the Almighty, the Forgiving."*
>
> -QURAN, 39:5

> *"We rendered the night and the day two signs. We made the night dark and the day lighted, that you may seek provisions from your Lord therein. This also establishes for you a timing system, and the means of calculation. We thus explain everything in detail."*
>
> -17:12

Since the two alternators are also in interplay with the bipolar magnetic fields of the Earth, a four-part process is revealed that is magnetically distributed as an overall eight-part structure. The unit of one kin divides the rotational cycle of the Earth in its orbit into four distinct phases daily and annually. The four daily points are defined by: the dawn, when the part of Earth you are in first experiences the solar photonic activation; the high noon (zenith), when the part of Earth you are in is in closest direct relation to the sun; the twilight, when the photonic activation ceases and the sun disappears from view; and midnight (nadir), the point at which the point of Earth you are on has rotated to its farthest distance from the sun, the height of the cycle of photonic regeneration. This daily process is mimicked in the four points of equinox and solstice, where for either hemisphere the equinoxes are the same as the points of dawn and twilight and the solstices correspond either to high noon or to midnight. Because of the magnetic polarity of the Earth, the four-part process is

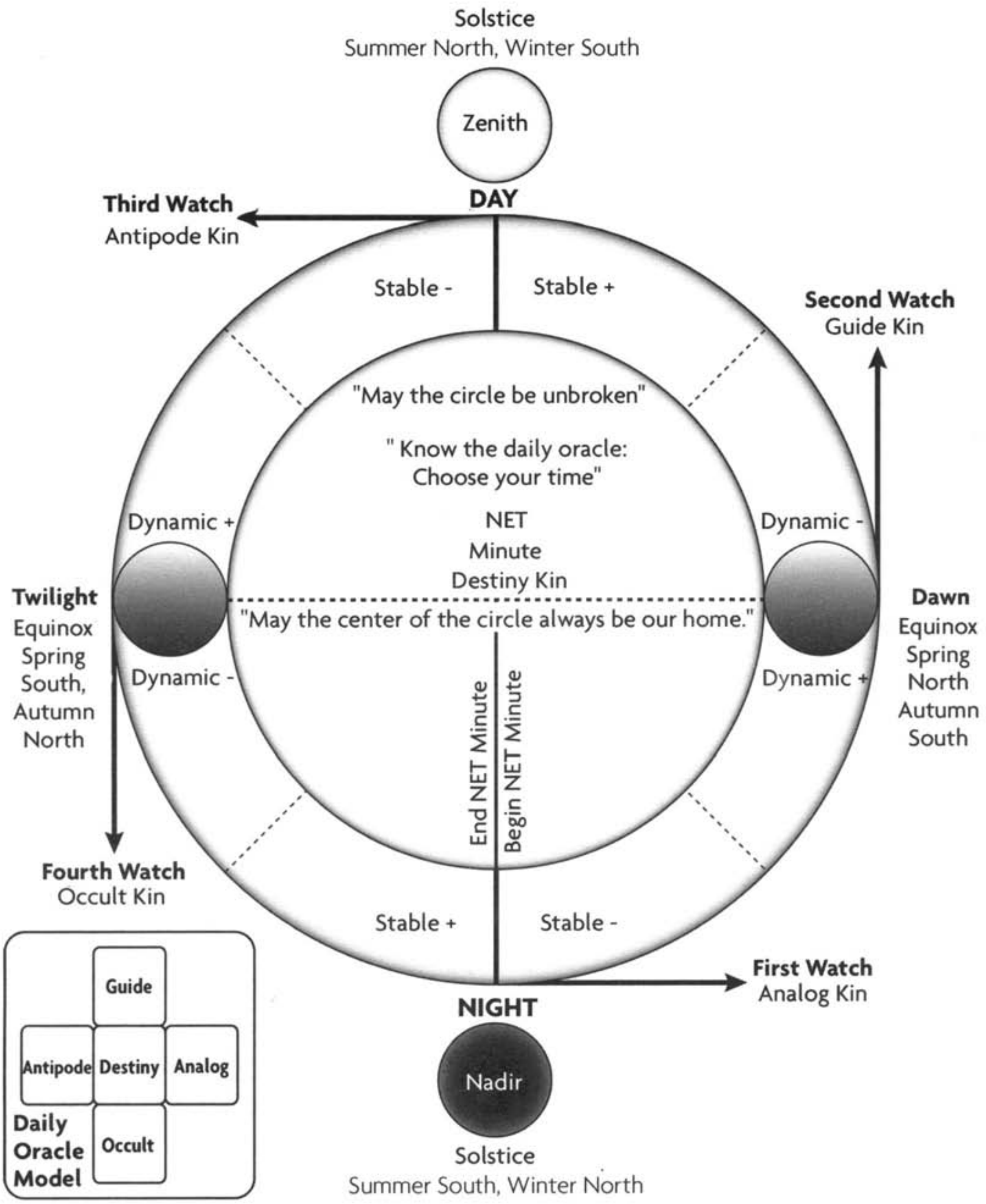
Solstice
Summer North, Winter South
Zenith
DAY
Third Watch
Antipode Kin
Stable -
Stable +
Second Watch
Guide Kin
"May the circle be unbroken"
" Know the daily oracle:
Choose your time"
NET
Minute
Destiny Kin
Dynamic +
Dynamic -
Twilight
Equinox
Spring
South,
Autumn
North
"May the center of the circle always be our home."
Dynamic -
Dynamic +
Dawn
Equinox
Spring
North
Autumn
South
End NET Minute
Begin NET Minute
Fourth Watch
Occult Kin
Stable +
Stable -
First Watch
Analog Kin
NIGHT
Nadir
Solstice
Summer South, Winter North
Guide
Antipode
Destiny
Analog
Occult
Daily
Oracle
Model

Four-Part Bipolar Kin:
Four daily watches,
Four annual seasons—
Two points of balance
Two points of return

Net Minute Model

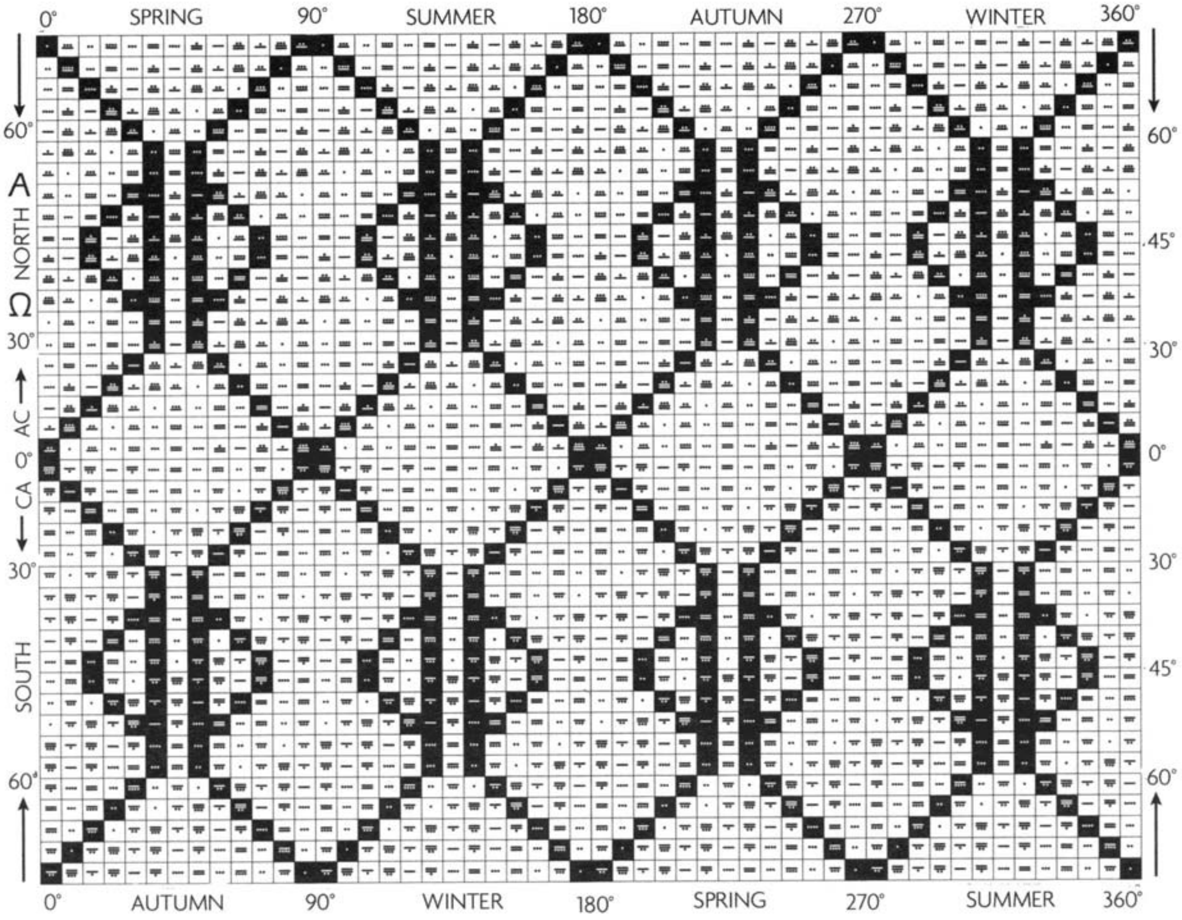
0°
SPRING
90°
SUMMER
180°
AUTUMN
270°
WINTER
360°
60°
A
NORTH
Ω
30°
AC
0°
CA
30°
SOUTH
60°
60°
45°
30°
0°
30°
45°
60°
0°
AUTUMN
90°
WINTER
180°
SPRING
270°
SUMMER
360°

Psi Bank—Four Bipolar Plates Accommodating NET Time Autoregulatory System of the Noosphere
(from *Earth Ascending,* Map 1)

simultaneously experienced in opposite phases, that is, winter in the North is summer in the South-hence the four-phased process yields a bipolar eight-fold structure.

This four-part day-night seasonal process defines the structure of the four psi bank plates, which also coordinate the four-year cycles of the Earth. The psi bank plates, the "nervous system" and control panel of the noosphere, are a complete function of the 13:20 timing frequency and thus participate in the systematic regularity of the fourth-dimensional synchronic order. This means that the psi bank is the registration within the Earth's magnetic field of the 13:20 timing matrix by which all life and the planetary evolutionary process as a whole are regulated. Simulating the four-part process of the Earth rotating on its axis in time, four stages per day, four seasons per year, in four-year cycles-not to mention the four phases of the moon-four psi bank plates accommodating the bipolar field yield eight 260-unit 13:20 frequency information matrices.

That is, each of the four psi bank plates conforms to a mirror symmetry of the bipolarity of the Earth's magnetic field. Each plate is divided by mirror symmetry into a 260-unit Northern and a 260-unit Southern hemispheric half, for a total of two bipolar matrices and 520 units per plate, or 2,080 (520 x 4) units in all. These units are referred to as *psi chrono units* or *time-bearing information units.* The information for the different sequences of the Earth's biogeochemical evolutionary processes is all regulated by these time-bearing information units.

We have to understand that through the medium of universal cosmic time the Earth, the Sun, and every star, constellation, and galaxy is coordinated in this way by the synchronic 13:20 frequency of time. It is the 13:20 frequency that establishes holonomic consistency in time throughout the universe. The Earth's noosphere is a lens of time. All cosmic information is focused through this lens. It is this noospheric lens of time that makes Earth into a geocosmic thought form.

The noospheric lens of time and mind has been operating ceaselessly-but unconsciously-since the beginning of the evolution of the Earth. Since the noosphere is a function of the universal synchronic timing frequency, its conscious activation must also come about through a superior mental coordination with this correct synchronizing timing frequency. The psi bank 13:20 frequency is consciously activated and registered at the human level by the mental engagement of the 13:20 matrix as enacted through the 260-day cycle. The interaction of the 260-day galactic cycle with the 365-day solar biotelepathic cycle defines one noospheric year, or one solar-galactic cycle of fifty-two human years (= fifty-two 365-day solar orbits = seventy-three 260-day galactic cycles). This great noospheric coordination of fifty-two solar orbits with seventy-three galactic cycles is fractally present in a single solar orbit, where fifty-two 7-day weeks = seventy-three 5-day cycles known as chromatics. What is important in coming to understand the noosphere as Earth's geocosmic time lens and rainbow brain, is that we must change our perspective altogether from anthropocentric to noospherocentric, and in this way evolve into *Homo noosphericus.*

THE CYCLES OF NET TIME

For the noosphere, the rainbow brain of planet Earth, time is different-it is slower and longer. From the noosphere's point of view, what we call a year is a solar orbit. What we call a day is a single rotation of the Earth on its axis. The cycles of the moon are the lunar phasings that calibrate the synchronization of the Earth in time. The Earth is not a spaceship. The Earth is a Timeship. As a Timeship, the Earth is a function of the master synchronization frequency that coordinates and moves it in time through ever greater circles of inclusiveness and syntropic integration. Seen through the noospheric lens of time, the technosphere is a highly exaggerated acceleration of the biogeochemical continuum through highly artificial and resource-depleting means-it is but a smoking bubble whose fragile struts and frets are laced together by transport and communication systems that yield a cacophony rather than a harmony. And for this reason, the disharmony of the bubble ultimately explodes upon itself, releasing Earth from the prison of false time into the fresh air of real time and the advance into the geocosmic splendor of the noosphere. Synchronized with the galactic cycles, the noospheric mind of the whole Earth is reckoned far differently from the artificial units of measure that condition the anthropocentric notions of time evolved in the era of *Homo historicus.*

Here is how the fourfold process of Earth in time is reckoned by the noosphere. Please note that in Noospheric Earth Time (NET), time is universal synchronization. At any given moment, it is always the same noospheric moment in time on

Table of Noospheric Earth Time Units

1. One Noospheric Earth minute = 1 day-and-night rotation of Earth on its axis = 1 kin.
2. Twenty-eight Noospheric Earth minutes = 1 Noospheric Earth hour = one moon of 28 kin. where every 7-day week = 1/4Noospheric Earth hour = 7 kin.
3. Thirteen Noospheric Earth hours (moons) = 1 Noospheric Earth day = 365 kin (days) = 1 solar orbit. or 1 human year (= 13 NET hours x 28 minutes + 1 NET minute, "Day Out of Time:')
4. Four Noospheric Earth days = 1 Noospheric Earth week = 4 solar orbits (years) = 1,460 kin (days).
5. Three Noospheric Earth weeks plus 1 Noospheric Earth day = 1 Noospheric Earth season = 13 solar orbits or 13 human years.
6. Four Noospheric Earth seasons = 1 Noospheric Earth year = 52 solar orbits (52 human years) = 73 260-day galactic cycles = 18,980 kin (days).
7. Five Noospheric Earth years = 1 Noospheric time chord = 260 solar orbits.
8. Twenty Noospheric time chords = Vs Noospheric evolutionary subcycle = 5,200 solar orbits = 1 Noospheric Century of 100 Noospheric Earth years.
9. Five Noospheric evolutionary subcycles = 1 galactic evolutionary day = 26,000 solar orbits = 5 Noospheric Earth centuries or 500 Noospheric Earth years of 52 solar orbits each.

Earth. Whether that moment is experienced on the night or day side of the Earth, that moment is simultaneous and encompasses both sides of the Earth at once. This is true not only because time is the factor of synchronization- T(E) = Art, where art is the result of any synchronization-but also because the velocity of time is infinitely instantaneous. By thinking that the Earth really does conform to the 24 hours of the clock, the humans have retarded their time sensibility, and consequently their artificial processes have become increasingly artless. In considering the Noospheric Time measurements, remember that as a human you are but a cell in the biospheric organism-or, as we shall see, a chip of the noospheric rainbow brain.

As the measure of the mind of the Earth, the Noospheric Earth Time (NET) reckoning units coordinate consciousness as a planetary, or, more accurately, a solar-planetary phenomenon. For instance, if one NET day is a single 365-day solar orbit of the Earth, dawn of the noospheric Earth day is experienced as the spring equinox, but the noosphere is bipolar. So the dawn is also simultaneously the autumn equinox or twilight of the other polar magnetic half of the Earth. To fully engage and grasp such a thought is to expand the limits of consciousness. Like dolphins who never sleep but only rest half of their brain at a time, so it is with the rainbow brain of the noosphere. Such consciousness has nothing to do with the workings of the everyday egoistic human consciousness, which can understand things only according to a limited frame of reference defined by its own limited and highly conditioned ego needs. The geocosmic consciousness of the noosphere can be apprehended only as a single telepathic unity in and by which all members of any given species or organism are organized. To consider that the noosphere itself could become the organizing principle of the human species as a whole is to engage in conscious mental evolution.

Small aboriginal prehistoric bands of humans experience an unconscious level of this geocosmic noospheric mind. Dolphins, other mammals, and all animal species experience this noospheric mind. Even trees and crystals are organized by this noospheric mind of time.

Homo historicus, however, slowly but inexorably cut himself off from the noospheric mind of time by substituting erroneous instruments of measure, or instruments of measure that were incomplete in relation to the noospheric time cycles. The result was the exponential surge in biogeochemical combustion that characterized the final baktun cycle, culminating in the fifty-six-year blister of artificial time, the technosphere. The fact that the time blister popped eleven years before the end of the cycle of *Homo historicus,* 2012, means that *Homo noosphericus* is ready to emerge from the shattered cocoon of false time. Liberated from artificial time and from the technosphere itself, and enrolled in the natural time of telepathy, the human will experience

Noosphere Earth's Rainbow Brain

Magnetic Alternation - Night and Day Law Of The Ring

"The Psi Bank in Perfect Motion"

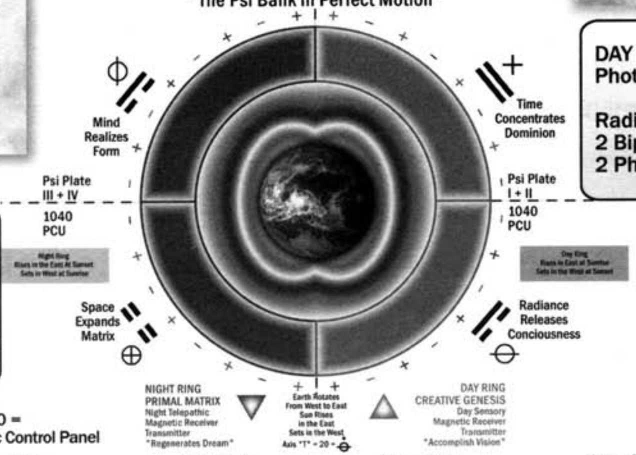
Mind Realizes Form
Time Concentrates Dominion
Psi Plate III + IV
1040 PCU
Psi Plate I + II
1040 PCU
Night Ring
Rises in the East At Sunset
Sets in West at Sunrise
Day Ring
Rises in East at Sunrise
Sets in the West at Sunset
Space Expands Matrix
Radiance Releases Conciousness
NIGHT RING
PRIMAL MATRIX
Night Telepathic
Magnetic Receiver
Transmitter
"Regenerates Dream"
Earth Rotates
From West to East
Sun Rises
in the East
Sets in the West
Axis "T" = 20 =
DAY RING
CREATIVE GENESIS
Day Sensory
Magnetic Receiver
Transmitter
"Accomplish Vision"

T(E)=ART

DAY ALTERNATOR
Photonic Activator

Radiation Belt x 2 =
2 Bipolar PSI Plates -
2 Phases

NIGHT ALTERNATOR
Photonic Regenerator

Radiation Belt x 2 =
2 Bipolar Psi Plates =
2 Phases

2 Phases + Polarity =
4 Bipolar Psi Plates @ 13:20 =
Nervous System + Noospheric Control Panel

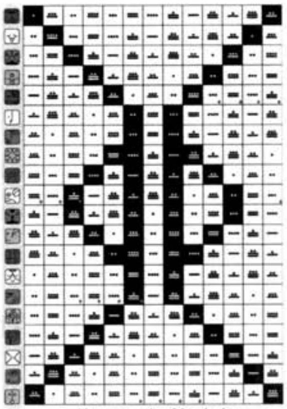

Harmonic Module

20 Solar Frequencies

13:20 Matrix Programs Four Phase Bipolar Psi Plates

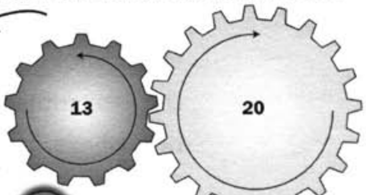

Synchronic Map and Guide to the Noosphere

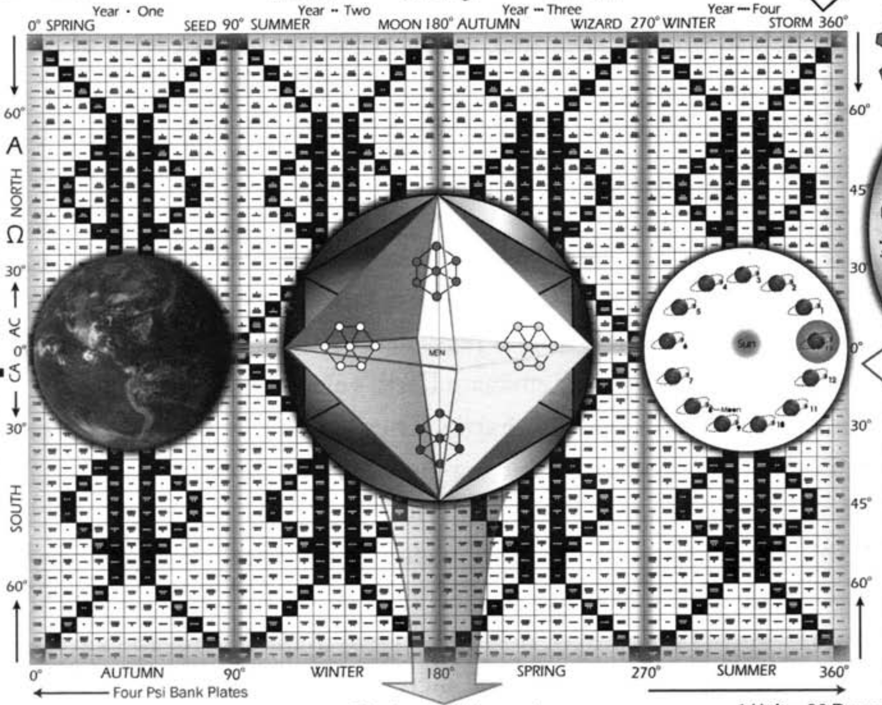
Year · One
0° SPRING
SEED 90°
Year -- Two
SUMMER
MOON 180°
Year --- Three
AUTUMN
WIZARD
Year ---- Four
270° WINTER
STORM 360°
60°
A
NORTH
Ω
45°
30°
AC
0°
CA
30°
SOUTH
45°
60°
MEN
Sun
Moon
0° AUTUMN 90° WINTER 180° SPRING 270° SUMMER 360°
Four Psi Bank Plates

13:20. Ratio Constant of Universal Synchronization

Earth Wizard.
"Homo Noosphericus"

13 Moon 28 Day + 1 Noospheric Regulator = 1 Noospheric Earth Day

Noospheric Earth Time (N.E.T.) Units

1 Day = 1 NET Minute
28 Days = 1 NET Hour = 1 Moon

13 Moon NET Hours = 1 NET Day = 1 Solar Orbit

4 Solar Orbits = 1 NET Week

3 NET Weeks + 1 NET Day = 1 NET Season = 13 Solar Orbits

4 NET Seasons = 52 Solar Orbits = 1 Noospheric Earth Time Year

Bipolar Oscillator and Crystal Transducer Earth Core

Activated by 1 Time Atom per ¼ Hour (7 Days) x 4 = 4 per Hour, 28 Days = 1 Moon-NET-Hour

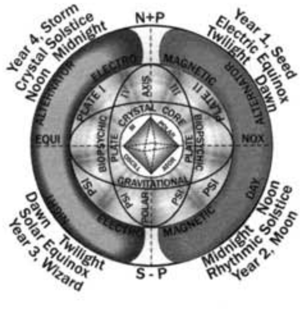
N+P
Year 4, Storm
Crystal Solstice
Noon Midnight
Year 1, Seed
Electric Equinox
Twilight Dawn
ALTERNATOR
ELECTRO
MAGNETIC
PLATE I
PLATE II
IV
AXIS
III
CRYSTAL CORE
BIOPSYCHIC
PLATE
EQUI
NOX
GRAVITATIONAL
PSI
POLAR
DAY
Dawn Twilight
Solar Equinox
Year 3, Wizard
Midnight Noon
Rhythmic Solstice
Year 2, Moon
S - P

Four Phasic Noospheric Resonator
Timeship Earth Field

- Day Night Alternator - 4 phases
  Twilight Noon Dawn Midnight
- 1 Moon-NET-Hour = 4 weeks ¼ Hours
- Solar Orbit = 4 Seasons = 4 Phases NET Day
  [Electric, Solar Equinoxes; Rhythmic, Crystal Solstices]
- 4 Solar Orbits = 1 NET Week = 1 Seed Storm Year Bundle = Return of 4 Psi Plates, 1 per Year.

Planet Holon - 20 Day Cycle

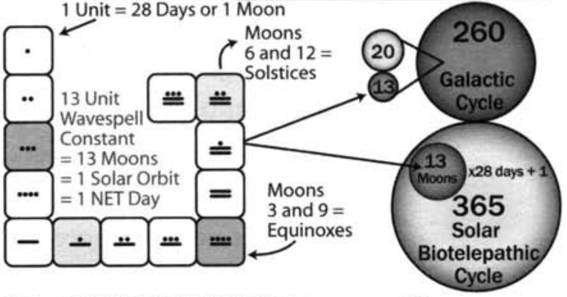
1 Unit = 28 Days or 1 Moon
Moons 6 and 12 = Solstices
13 Unit Wavespell Constant = 13 Moons = 1 Solar Orbit = 1 NET Day
Moons 3 and 9 = Equinoxes
260
20
13
Galactic Cycle
13 Moons
x28 days + 1
365
Solar Biotelepathic Cycle

Timeship Earth 2013
Coded by 20 Solar Frequencies in Time

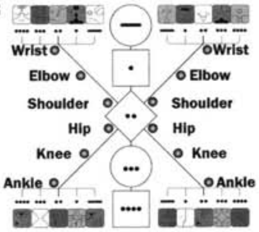
Wrist
Wrist
Elbow
Elbow
Shoulder
Shoulder
Hip
Hip
Knee
Knee
Ankle
Ankle

13:20 Noospheric Chip
"Homo Noosphericus"
20 Fingers + Toes
4 Limbs
13 Main Articulations

Consciously synchronized in 13:20 time. Plays Psi Bank Control Panel, generates Earth's Rainbow Brain

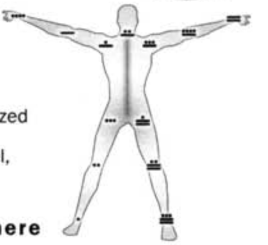

Man in the Noosphere

noospheric organization into psychic biomes. Through the psychic biomes-collectives of communities coordinating their bioregional network according to the structure of the "thought moments" of noospheric time-the human will fully evolve the Earth as a planet with solar-galactic consciousness.

ELECTROMAGNETIC WORKINGS OF THE EARTH'S RAINBOW BRAIN

What we call the cycles of time are the thought moments of the Earth. These thought moments coordinate the complex orders of Timeship Earth, understood as a whole system. Because the operating mechanism of the noosphere that governs the Earth's cyclic thought moments is inseparable from the general structure and functions of the electromagnetic field in which it is located, we may refer to a *rainbow brain* of the noosphere. Why is this so?

Let us return to the structure of the noosphere, Earth's mind in time. Consisting of the four bipolar psi bank plates, these four synchronic time plates are located between the two radiation belts, that is to say between 11,000 and 2,000 miles above the biospheric surface of the Earth. With the location of its information time-bearing mechanism between the radiation belts, the noosphere must by necessity operate in tandem with the electromagnetic field as it is regulated by these two master radiation belts. What do these belts do? They emanate Earth's electromagnetic aura, and function as a protective shield around the Earth. They do this by trapping solar-cosmic radiation, keeping much of it from hitting the biosphere directly, channeling it instead through the Earth's magnetic poles into the Earth's core.

In the polar regions excess cosmic radiation, including electrically charged plasmas and solar discharges from coronal mass ejections, is released into the atmosphere as the auroras-aurora borealis (or northern lights) in the North, and aurora australis (or southern lights) in the South. The increased solar activity during the most recent and current sunspot cycles has enhanced and extended the auroral activity in recent years far beyond the polar zones. In addition, since 1987 great influxes of plasma into the solar system have intensified the incidence of spectral plasmic activity within the atmosphere itself. This amounts to an increase in rainbows, halos around the sun, and brilliantly colored clouds. While these phenomena may be explained to a certain degree as meteorological effects (the angle of the sun's rays, moisture in the atmosphere, and so on), is there more to a rainbow than meteorological side effects? And what relation is there between the noosphere, the electromagnetic field, and the increased plasmic-rainbow activity?

Again, seen from the perspective of whole Earth, the rainbows and spectral plasmic effects, including the auroras, are the free energy released by the noosphere. Rainbows are to the noosphere what toxic waste is to the technosphere. The noosphere itself implies a mental field that is interacting with the electromagnetic field. The Earth's electromagnetic field is actually the strongest of all the planets in the solar system. The only other planet that has a field approaching the strength of the Earth's field is Uranus. If the Earth's electromagnetic field is the strongest, is this an indication of the strength of the noosphere as well? If the rainbows, auroras, and other spectral plasmic activity are a demonstration of the free energy released due to the interaction of the noosphere and electromagnetic fields, and the noosphere is the mental envelope of the planet, inclusive of the as-yet-to-be-unified mind of the human species, what forms of free energy might be released if the human mental field were operating as a single unity in the correct timing frequency?

One other mechanism crucial to the entire system of noosphere and electromagnetic fields is the crystal core at the center of the Earth. In the 1970s Russian scientists put forth the idea that the Earth is crystalline in nature, an idea that extended Plato's notion of the Earth as a dodecahedron, one of the five "platonic solids." The idea was that the Earth's surface was held together by a field of resonance that corresponds to the crystal resonance of certain geometrical solids. R. Buckminster Fuller also gave play to this notion in his description of "spaceship earth." Then in the 1980s American seismologists confirmed that the core of the Earth is actually a gigantic iron crystal structure in the shape of an octahedron, elongated at its ends, which point to the North and South Poles.

From the point of view of fourth-dimensional time, the planet Earth is another manifestation of space as an infinitely locatable point, an orbiting body upon which all the vectoral programs of cosmic time may be unfolded. Understood as a Timeship, the Earth's structure is constructed to facilitate its travel in time on a particular route in space in conformity with its relation to the local star, the sun, to the other planets of the solar system or heliosphere, and in synchronization with the universal timing constant 13:20, which maintains the Earth in the synchronic order. The role of the magnetic octahedral crystal core is to function as the bipolar oscillator and crystal transducer whose purpose is to maintain the Earth in the synchronic order of the 13:20 timing frequency. Thus, if the Earth's noosphere contains the 13:20 timing programs of the psi bank, and the crystal bipolar oscillator is meant to maintain the Earth's space body in the synchronic order, obviously there is a resonant relation between the crystal core and the noospheric-electromagnetic field programs.

The transductive properties of the crystal allow it to transform energy from one state to another. The crystal core also stores electronic plasmas that travel to it from the

Earth's poles through an electromagnetic flux tube system. Within the crystalline core, the electronic plasmas are transduced into information. This is the information of time understood as cosmically transmitted thought forms. These plasmic thought forms conform to programs already established in the psi bank of the noosphere. These are the Earth's thought forms, the cycles of the Earth's brain waves, as it were, that allow it to maintain its biospheric programs and to evolve its mental field, the noosphere. The very descriptions of the Noospheric Earth Time (NET), and of time itself as a synchronous instantaneity inseparable from telepathy, are examples of the Earth's thought forms transduced through a human biocomputer and then transmitted into a form of human language for the purpose of communication. Keeping this description of the noospheric brain in mind, let us return to our analysis of the biosphere-noosphere transition.

MAKING THE NOOSPHERIC LEAP: TRANSFORMATION OF HUMANS INTO NOOSPHERIC CHIPS

The transition to the noosphere is experienced by Earth as a crisis because its systems are stressed. Since the Earth functions as a whole system, stress on one component of the system actually stresses the whole order. Let us review the cumulative stress points. Human 12:60 electromagnetic activity has perturbed and distorted the electromagnetic field. The release of free energy due to geochemical combustion adds to this stressful distortion through the depletion of ozone in the ionosphere and global warming. The radioactivity released into the atmosphere adversely affects the resonant field of the DNA, already driven by the 12:60 frequency to accelerate the multiplication and propagation of the human biomass throughout the biosphere. Through its artificial means, the machine, the human further upsets the inherent balance maintained by the pressure that the species exert upon each other, resulting in mass species extinction.

Yet all this is according to an evolutionary program that is intended to bring about an ultimate crisis by which the biosphere snaps into becoming the noosphere. Clearly, if the root of this crisis is the deviation of the dominant species operating according to an artificial timing mechanism at odds with the universal timing frequency, then the stabilization of the transition to the noosphere will come about through the dominant species consciously "returning" to the correct timing frequency. This it can only do by adhering to a macro-organizing program that puts the human species completely in the 13:20 frequency.

If the whole system analysis seeks to locate the flaw in the operation of Timeship Earth, then we must finally declare that this flaw is the human civilization in its entirety, as it is maintained by erroneous and artificial timing constructs. Since the noosphere, the mind of time on Earth, defines the next evolutionary cycle, and because the humans are following a course dictated by being in the wrong time, calendar reform is the eye of the needle through which humanity must pass in order to get to the noosphere. Only by fully manifesting the noosphere will the Timeship Earth be truly put back on course once again. The noosphere is the program of time on Earth. Time is all that the noosphere has to give us. In consideration that the artificial time of the technosphere has fundamentally run out, the noospheric time should be a welcome gift.

Science is always absolutely precise. It is not just calendar reform that is the eye of the needle through which to pass into the noosphere-it is the Thirteen Moon/28-Day calendar calibrated by the 260-day Wizard's Count, synchronized to the Gregorian calendar date, July 26, that provides the exact ticket of admission into the noosphere. And then this ticket must be used at just the right moment-when the show is ready to begin. To use another metaphor, there is a lock on the door to the noosphere, and there is only one key that will fit and unlock this noospheric gate. That key is the Thirteen Moon/28-Day calendar, correlated to July 26 on the Gregorian calendar. It is the precisely applied use of this Thirteen Moon/28-Day calendar that will make each human a noospheric chip of the rainbow brain. This is because the Thirteen Moon/28-Day calendar is the key with the precise design for not only opening but also for continuing to operate the noosphere through a matching of the everyday frequency of time with the frequency of the noospheric control panel, the psi bank.

It is because the daily program of the Thirteen Moon/28-Day calendar is embedded in the 13:20 frequency that the consciousness of the humans will finally change and rise above the stunted level by which it is currently maintained in the artificial world of the technosphere. Time is of the mind, and the natural time registered by the Thirteen Moon/28-Day calendar is the opening to the geocosmic mind of the universe, which adherence to artificial time has kept the human from experiencing. The noosphere is the fractal lens of this geocosmic mind on Earth. Since the geocosmic mind is purely a function of the Law of Time-energy factored by time equals art, where the 13:20 synchronizing ratio of time is characterized by infinitely instantaneous velocity-then the fundamental nature of this mind is telepathy.

The Thirteen Moon/28-Day calendar is, therefore, the perfectly designed instrument to make the noospheric leap into telepathy, the planetization of consciousness that establishes the daily reality of the noospheric mind. Since the noospheric mind interacts with the electromagnetic field, and because the ratios and cycles of the Thirteen Moon/28-Day calendar conform to the fractals and ratios of the NET, by using this calendar the human becomes transformed into a noospheric chip. The fractals and ratios of NET become the normative operating cycles of the human in its daily reality. This accounts for an increasing noospheric synchronization that allows the noospheric chip to attain its purpose: generating the Earth's rainbow aura. If the free energy of rainbows is to the noosphere what the free energy of toxic waste is to the technosphere, then the generation of the rainbow implies a type of activity on the part of the human to produce such free energy. To understand what this activity might be and how it works, we must first understand the morphic and holonomic relation between the human noospheric *chip-Homo noosphericus-and* the structure of the noosphere itself as it is programmed into the geocosmic measure of the Thirteen Moon/28-Day calendar.

HUMAN BEINGS AS FRACTAL REFLECTORS OF THE 13:20 MATRIX

From the noospheric perspective, the prototype human is an embodiment of the 13:20 synchronization factor. The human has four limbs. Each of the four extremities has five digits, for a total of twenty fingers and toes that distinguish the human's operational capacity. The four limbs conform to the fourfold phasic operation ofNoospheric Earth Time, while the twenty digits conform to the 20 (4 x 5) factor of the 13:20 ratio. The human also has thirteen main joints or articulations: two ankles, two knees, two hips, two wrists, two elbows, and two shoulders; these six pairs coordinate the four extremities with the body, while the thirteenth articulation, the neck and spinal column, coordinates the body proper. These thirteen main articulations plus the twenty digits of the hand and feet conspire to make the human the perfect exemplar of 13:20 timing frequency, that is, a perfect model of synchronization itself.

Add to this the fact that the human has a highly evolved and refined brain and nervous system that is maintained and expressed by the skeletal structure of the thirteen joints of the body and limbs, completed by the sensitive mechanism of the twenty digits, and we see that the human is the perfect biocomputer for processing the information of the instantaneously telepathic geocosmic mind. The means to process the geocosmic information, and also the means by which the human is transformed into the noospheric chip, is found in matching its factors of daily synchronization and intelligence with those of the psi bank, the noospheric control panel. This is called harmonic arrangement of the synchronic order.

Of course, this description of generalized structure, activity, and interaction of the *Homo noosphericus* with the noosphere as such presupposes an entire reorientation of the goals and methods of being human. Such a reorientation of the human is also a function of making the radical change from the erroneous measure of the twelve-month Gregorian calendar, in which are embedded all of the Babylonian programs of history, to the perfect history-less harmony of the Thirteen Moon/28-Day calendar- the mere change of calendars encompasses a mental frequency shift. The readiness to make this change depends on the degree of moral revulsion the human experiences during the collapse of the technosphere. This moral revulsion would then correspond to a turning in the deepest seat of consciousness, which would then dispose the human intelligence to a consideration of the untried solution-change the calendar!

The Thirteen Moon calendar is the NET instrument for daily synchronization and unification of the human species as a single organism in time. Once adopted to the correct measure of this instrument, the human will de facto trigger the noosphere as an operating system of absolutely higher consciousness and mental order. **In** fact, the purpose of human evolution is precisely to become the ignition of the noosphere,

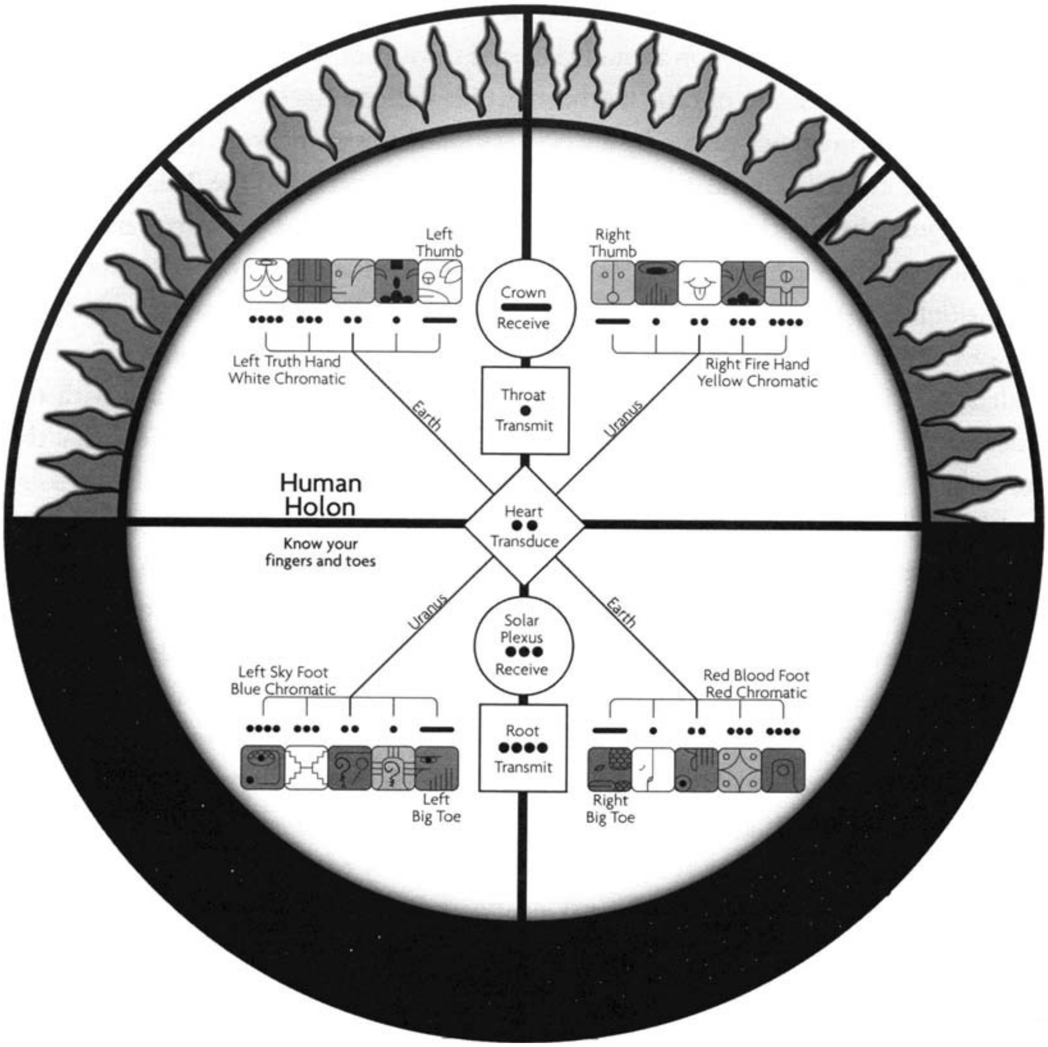
Left Thumb
Right Thumb
Crown
Receive
Left Truth Hand
White Chromatic
Right Fire Hand
Yellow Chromatic
Throat
Transmit
Earth
Uranus
Human Holon
Know your fingers and toes
Heart
Transduce
Uranus
Earth
Solar Plexus
Receive
Left Sky Foot
Blue Chromatic
Red Blood Foot
Red Chromatic
Root
Transmit
Left Big Toe
Right Big Toe

Human Noospheric Chip

and thence to become normalized as the noospheric chips that maintain the operations of the noosphere. As a truly planetized being and consciousness, *Homo noosphericus* will engage in an order and type of reflective activity undreamed of in the technospheric era that is now passing.

First, by adapting to the daily NET timing instrument, the humans will be utilizing a "calendar" that correlates both a 365-day solar biotelepathic cycle and a 260-day galactic synchronization cycle. When the lunar phases are added to the interaction of these two cycles, then the human is able to enjoy the harmonic richness of time-the noosphere. The synchronizing and interactive measures of the counts of the solar and galactic cycles yield multiple levels of harmonic orders of time (Earth's "thought moments") that will activate aspects and levels of consciousness hitherto dormant in the human mind. The very mental act of replacing the order of twelve by the order of thirteen is a profound one. How is this so? Could it be that the whole of civilized history itself is based on the fear of the number *thirteen-triskaidekaphobia,* epitomized in the superstition about Friday the thirteenth-and that, therefore, dealing with the true nature of time has been avoided altogether?2

Currently humans operate by a world standard whose measure includes: seven-day weeks, inclusive of the notions of the 5-day work week and the 2-day weekend; an irregular distribution of twelve months of either thirty-one, thirty, or, in one instance, twenty-eight days; all set within a longer annual cycle that concludes with its visionless moment of New Year's Eve. Many Gregorian calendars will include the phases of the moon, but because the months differ in length there is no way in these calendars of seeing how the phases of the moon might be synchronized with the measure of Earth's orbit around the sun, nor is there any reason to consider that the uneven measure of twelve months has anything intrinsically to do with the solar orbit. Within the twelve-month structure of the Gregorian calendar, weeks, months, and phases of the moon all appear to be haphazard and arbitrary. Such a disorderly order can only bode an undercurrent of uncertainty and fear for the human mind.

By contrast, in the measure of the thirteen moons of twenty-eight days each, the 13 corresponds both to a 13-day cycle and to the cycle of the number of months (moons) per year. The measure of the 7-day week (a NET quarter hour) becomes an orderly harmonic: four perfect weeks occurring each moon, fifty-two perfect weeks each year, each week always beginning and ending on the same day of the week, week after week, year after year, while in fact every day of every moon occurs on the same day of the week, year after year. The effect of such a harmonic regularity on the human mind is incalculable. This is what is meant by a *perpetual calendar.* And to maintain this perfect regularity, the 365th day of every year (solar orbit), is no day of the week at all, but a Day Out of Time, a day for the human to be refreshed in utter timelessness.

Realizing that a day is actually a kin or one NET minute, the measure of one rotation of Earth on its axis, a 4-day cycle of four NET minutes constitutes a harmonic, a 5-day cycle constitutes a chromatic, while the 13-day cycle corresponds to one *wavespell,* the time constant that accommodates the measure of thirteen. Hence, a wavespell can be either thirteen days, thirteen weeks, thirteen moons, thirteen years, or so forth. The point is that by making the change of calendars the foundations of the mind organized in time are irrevocably altered. From the noospheric perspective, there is nothing arbitrary about the laws of nature. The same lack of arbitrariness should be reflected in the instruments that we use to measure time. By adapting to the Thirteen Moon/28-Day measure, the humans are adapting to a measure in which there is nothing arbitrary because it is a pure expression of the Law of Time. Thirteen is not at all a harbinger of bad luck, but the gateway to harmony in time! And it is by using this thirteen-based harmonic instrument of measure that the human is realized as a noospheric chip, a fractal of the noospheric mind, the mind of time on Earth.

NOOSPHERIC CHIPS: WIZARDS OF GALACTIC TIME

In the noospheric construction of the operating mechanism of the Timeship, the human is a vital link, providing the biotelepathic medium for the synchronization of the geocosmic mind on Earth. If we look again at the structure of the noosphere, the components include: the crystal core bipolar oscillator, packed with Technicolor time beams and cosmic thought moments, waiting to be self-reflectively realized; the human noospheric chip, the representative of the biosphere to mentally engage the timing cycles of the NET; the electromagnetic field, which provides the medium of expression between the biospheric order, the human, and the noospheric mind of time; and the noosphere itself, the mind of time on Earth, whose articulation through the psi bank control panels and through the Earth's resonant field structure yields an enlarged measure of Earth consciousness, Noospheric Earth Time itself. In expanding our minds into the noosphere we will understand that what lives is the Earth and the Earth lives through us.

As a noospheric chip, the human is a fractal reflector of the 13:20 timing frequency. Midway between the bipolar oscillator and the noospheric psi bank plates, the human noospheric chip is designed to receive, transduce, and transmit programs from the crystal core by synchronizing these programs with the psi bank.

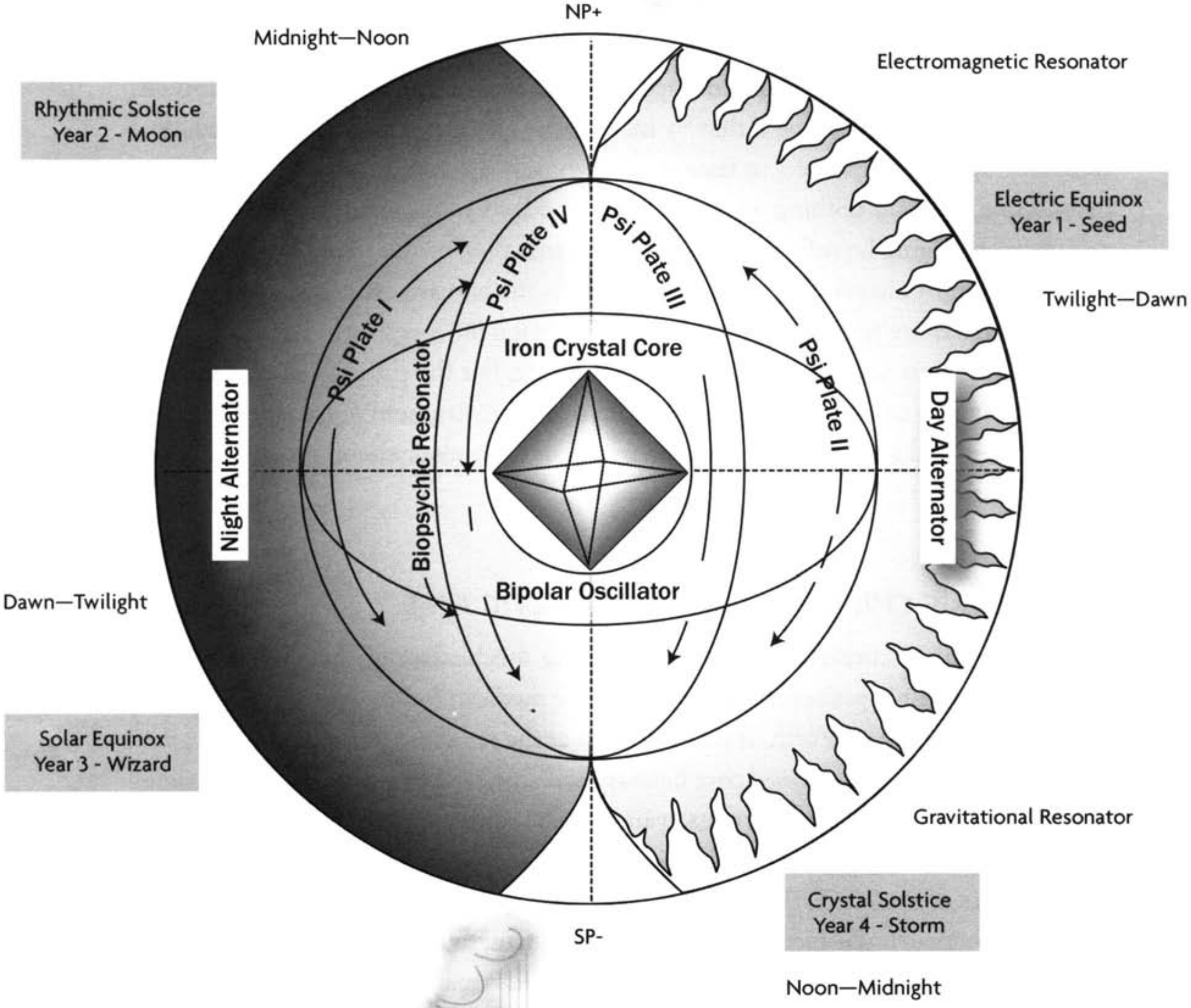
NP+
Midnight—Noon
Electromagnetic Resonator
Rhythmic Solstice
Year 2 - Moon
Electric Equinox
Year 1 - Seed
Twilight—Dawn
Psi Plate IV
Psi Plate III
Psi Plate I
Psi Plate II
Iron Crystal Core
Biopsychic Resonator
Night Alternator
Day Alternator
Bipolar Oscillator
Dawn—Twilight
Solar Equinox
Year 3 - Wizard
Gravitational Resonator
Crystal Solstice
Year 4 - Storm
SP-
Noon—Midnight

Four-Phasic Noospheric Resonant Field Model, Timeship Earth 2013

Three resonators establish the gravitational, electromagnetic, and biopsychic fields held in place by the crystal core, the bipolar oscillator.

Two poles extend the bipolar field of Earth, which, guided by the day-night alternators, create the four-phasic noospheric field model.

Four psi plates govern the four phases, from daily NET minute to seasonal NET day, and to the four-part NET week, or Seed-Storm year bundle.

This is a process that can occur only within the construct of the regular harmonic sequence provided by the noospheric timing gauge of the Thirteen Moon/28-Day calendar. Interaction with the crystal core can only be through telepathic means, and the human-as the midway point between the core and the electromagnetic fields-is like a sensitive piece of litmus paper. To understand the human in this way is to totally redefine the reality of what it means to be human on Earth. From this perspective, the human is a biocomputer whose information source feeds it programs stored and emanated by the crystal core. These programs are then brought into conscious registration through working the corresponding programs of the psi bank. Right now, we don't know what the crystal core is emanating, much less that such a structure exists within the Earth. Like a crystal radio-set, the noosphere is the antenna, the geocosmic programs are transduced and generated at the crystal core, the human is the biocosmic tuner, and the broadcast transmission is provided by the psi bank.

The 13:20 template of the human noospheric chip is defined as the human holon, its twenty digits and thirteen main articulations conforming perfectly to the geocosmic 13:20 synchronization frequency of universal time. It is the intrinsic resonance of this geocosmic structure that allows the human to be in resonance with the planet as a whole system. While the timing programs to be broadcast are all loaded in the psi bank control boards, the actual point in space of the physical Earth is coded by time into a twenty-part icosahedral structure called a planet holon. Being holonomic in nature, the planet holon has a perfect incidence of correspondence with the twenty digits of the human holon.

The twenty digits and twenty-part structure of the planet holon are both coded by the twenty icons or solar seals of the 260-day galactic synchronization cycle. The icons or seals are the manifestations of the 20 (4 x 5) factor of the 13:20 synchronization ratio. They are coded for human legibility as simple icons that conform to a postliterate alphabet. In time these twenty icons create a fourth-dimensional twenty-day solar cycle which interacts with the 13-day galactic cycle to create the 260 unit synchronization sequence or "galactic spin." The point is that when this galactic spin is synchronized with the 365-day solar-lunar measure of the Thirteen Moon count, there is a daily program for identifying the human with the planet holon, a factor that facilitates the planetary resonance of the noospheric chip.

As a noospheric chip, the human is in every way a holonomic reflection of the whole system Earth. As a system of resonators-gravitational, electromagnetic, and biopsychic-the human and the Earth both have bodies with a center of gravity; both exhibit electromagnetic fields; and both share in common a biopsychic resonance, the mutual interface of the noosphere or mind of time with the other two

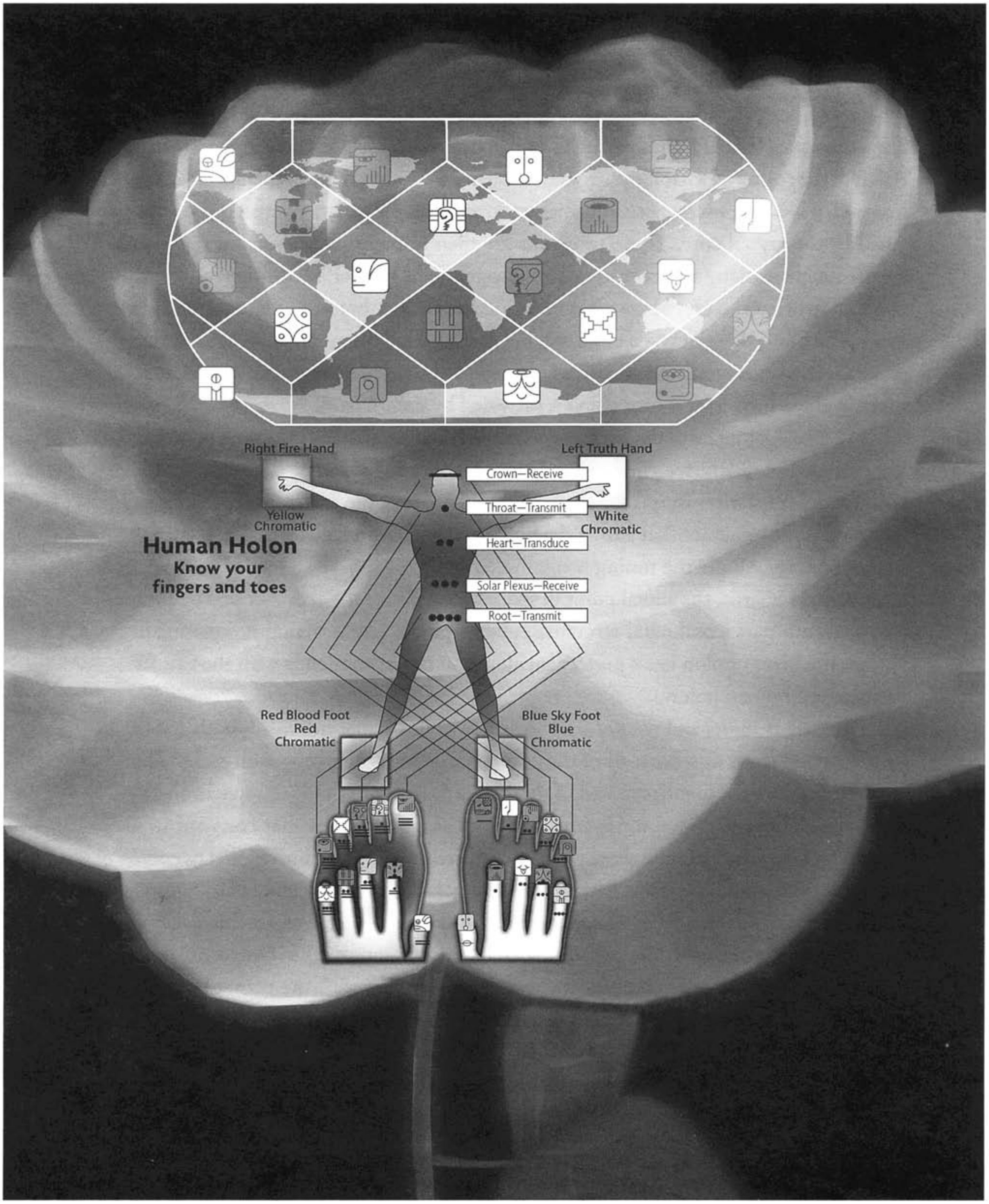
Right Fire Hand
Left Truth Hand
Crown—Receive
Throat—Transmit
Yellow Chromatic
White Chromatic
Heart—Transduce
Human Holon
Know your fingers and toes
Solar Plexus—Receive
Root—Transmit
Red Blood Foot
Red Chromatic
Blue Sky Foot
Blue Chromatic

Planet Holon—Timeship Earth 2013

fields of resonance. The redefinition of the human according to the Law of Time and to the nature of the noosphere also redefines the human's reality. What will the human do when history has been eliminated from the calendar and the new calendar itself is actually operated as a synchronometer-a measure of synchronicity?

First of all, as the necessary prelude to galactic culture, the absolute synchronization of the human in geocosmic time, the technosphere will be dismantled. Reorganized by the Law of Time into psychic biomes, the dismantling will naturally phase into the genuine creation of the new culture of solar-galactic consciousness. "Path beyond technology" means path of telepathic harmony. With its goals absolutely reoriented following the mental pole shift of the calendar change, the hierarchy of human values will also undergo a profound reordering. What if money were no longer the first consideration in human affairs? What if the consideration were instead how artful or artistic that action or event might be? What if the diaspora of humanity through industrialization and mass transport came to an end and the humans willingly chose to live in smaller communities organized by gardening and other Earth-related activities, inclusive of the transformative dismantling of the technosphere and the regeneration of the biospheric medium? What if the study of time replaced the study of physics as the dominant mode for determining the nature of reality? And what if, due to the time shift, a spiritual awakening swept through the human race so that a moral orientation of values was organized around the preservation and furtherance of the biosphere, and an acknowledgement that only One Supreme Creator could have ordained something so perfect as the noospheric mind of time? What kind of world would that be? What kind of art would be practiced in that world and what kind of science would be known?

Organized by the NET, human existence would spiral into the collective telepathic mind. A type of Earth geomancy would become the science of the noosphere. Or perhaps a more accurate name for this science would be *Earth geochronomancy*-the study and investigation of Earth in relation to the cycles of cosmic time and telepathy with an end to perfecting the Earth as a work of art. With minds expanded toward accommodating the longer, slower measure of NET, geochronomancy would inspire an entirely new vision of the Earth within the heliosphere (the solar system) and the galactosphere (the galactic order of things). Making the Earth's rainbow brain visible and manifest would be only the beginning of a willingness to explore further untapped psychic energy sources and means, or what is referred to in *Earth Ascending* as the science of radiosonics. Key concentrations of human habitation and activity would be organized to occur at or close to those geomagnetic energy points or centers, the whole cultural process of which would correspond to making a map on Earth of the image of galactic being as it moves in its great patterns of time. Would this not make the human species into a race of wizards?

In this consideration we may find a reason why the 260-day unit galactic synchronization measure-a galactic spin-is the key to what is called the *Wizard's Count.* The perfect synchronization of the 260-unit galactic spin with the 365-day Thirteen Moon/28-Day solar orbit establishes the 52-year cycle, or one NET year. A NET year (fifty-two solar orbits) is a "wizard's measure," for on this measure the human can expand its consciousness to encompass the Earth in its greater cycles of development. The Wizard's Count is also given its name because of the set of the four solar icons or seals that code the July 26 synchronization date in 4-year cycles: year one, Yellow Seed year; year two, Red Moon year; year three, White Wizard year; and year four, Blue Storm year. An entire Seed-Moon-Wizard-Storm sequence of four NET days (four solar orbits) consists of fifty-two moons, or one NET week, the perfect fractal of the fifty-two 7-day weeks in one NET day and the fifty-two solar orbits in one NET year. This four-part sequence is also called a Seed-Storm year bundle, the harmonic essence of the wizard's measure. Thirteen Seed-Storm year bundles, where each of the four icons is coded by the thirteen galactic tones (4 x 13 = 52), yield one NET year. Please note that even the Gregorian July 26 synchronization date is a function of the synchronic order: July = 7, while 26 is 13 doubled, 7 and 13 being the two principle numbers governing the 13:20 matrix of the Harmonic Module!

In the symbolism of the iconic sequence it is the Seed that is planted, it is the Moon that coordinates the growth cycles, it is the Storm that generates the energy, and it is the Wizard who knows and acts upon what is given. It was a White Galactic Wizard that coded the date and coordinated the year beginning July 26, 1987-followed twenty-one days later by the Harmonic Convergence. Thus it was the White Galactic Wizard who opened the gates to the closing of the Mayan baktun cycle of history and the commencement of the coming cycle of galactic culture. And on July 26,2003, the year of the White Spectral Wizard will commence. This year will end with its Day Out of Time, coded by the White Spectral Mirror. The day following, July 26,2004, will begin the year of the Blue Crystal Storm. On the Wizard's Count, which is not a calendar but a synchronometer, these two dates, White Spectral Mirror and Blue Crystal Storm, are the point in time when the Gregorian calendar ticket to the old time will be eliminated and the new Thirteen Moon calendar ticket for admission to the new time of the noosphere will be validated. This is when the noospheric show will be ready to begin its prelude. (See plate 6, Noosphere-Mind of Time on Earth.) Will the humans also be ready?

▼

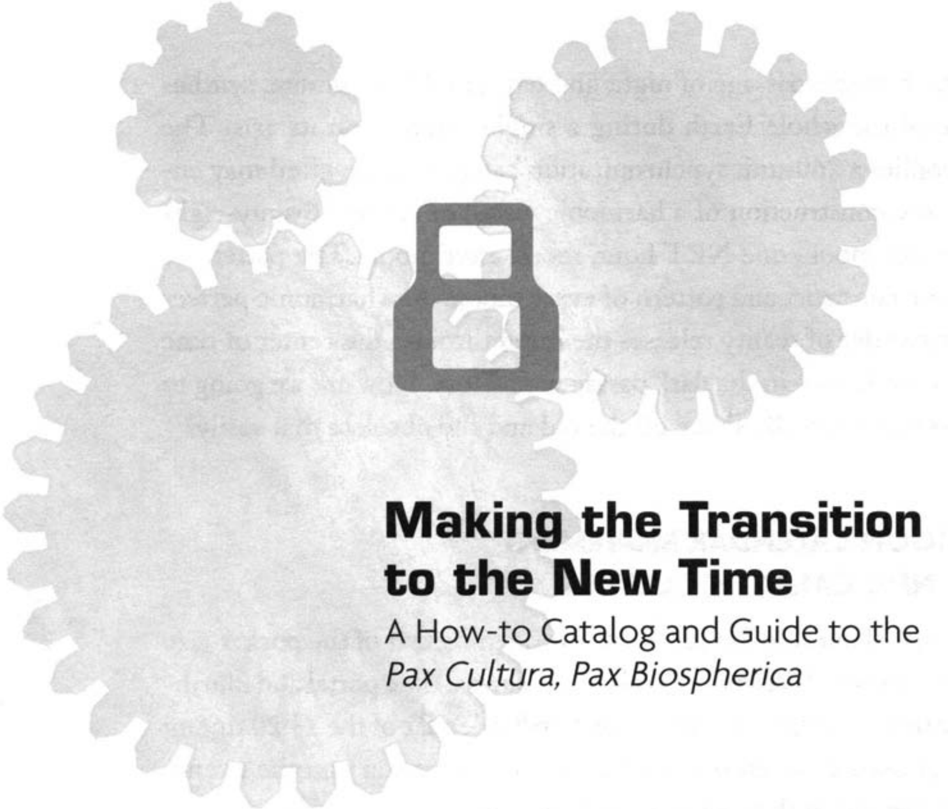

8

Making the Transition to the New Time

A How-to Catalog and Guide to the *Pax Cultura, Pax Biospherica*

BY THE TIME these words are read, the technospheric collapse will have generated several more devastations and disasters. The wound received by the technosphere on September 11 was a mortal one. The technosphere suffered cardiac arrest for almost a week. Not only did the towers collapse, an event that occurred via television within the central nervous system of the human species, but belief systems, too, came crashing down. No longer is America the unassailable fortress of "democracy," while democracy itself is a concept that joins monarchy, nation states, imperialism, colonialism, capitalism, communism, and a host of other outmoded systems of belief and thought constituting the saga of human history-that is, the history of civilization. But all that is now ending. The perimeter of biogeochemical combustion and the limits of the technosphere have coincided. It is the Earth's turn to speak. How will the Earth speak?

The Earth can only speak through the noosphere, and the noosphere can only speak from the center of time that is constructive peace. Peace is harmony. Peace is the culture of the biosphere. The conscious harmony of time in the biosphere is the culture of peace known as the noosphere. Through the noosphere, the Earth speaks in ever-widening circles of order and meaning. While sunrise and sunset mark the passage of solar time, observing the phases of the moon keeps us in tune with the mystery of time. One kin, one passage of night and day, one NET minute, synthesizes the experience of the whole Earth during a single rotation on its axis. The synchronic order supplies a 260-unit synchronization gauge that the mind may engage and employ in the construction of a harmonic vision of reality. Twenty-eight NET minutes make one moon, one NET hour, moon after moon. The pattern of one moon contains the construct and pattern of every moon. The harmonic perfection of the synchronic order of reality releases the mind through the center of time into a freedom it had not known in the dark passage of history. How are we going to make this transition? Can we really shake off the old and the obsolete that easily?

THE THIRTEEN MOON CALENDAR MOVEMENT: THE DAWN OF A NEW GALACTIC CULTURE

"Greetings Explorers on Timeship Earth!" reads the opening text of the pocket-size *13 Moon Natural Time Journal.* "Welcome to this 13:20 natural time portal and fourth-dimensional galactic time template. A solar-lunar Earth calendar of the 13:20 timing frequency that is expressed through the thirteen tones of creation and the twenty sacred solar tribes. Attuning to this calendar will help you awaken and align to the pulse of the Universe, the rhythm of the One heartbeat. 13:20 is the natural timing frequency that unifies planetary and galactic consciousness. . . It is the expression and remembrance that Time is Art . . . By attuning to the 13:20 frequency of natural time, we break out of the 12:60 time warp and activate the remembrance of our true human nature. Time is energy. It is a frequency of the fourth dimension; mind, telepathy, imagination, and creation. Time is an exploration of unlimited Universal potential. So let us begin the journey and return to the rhythm of nature's galactic heartbeat. . ."1

This is the Earth speaking through a noospheric chip. Produced in Canada, it is one of many Thirteen Moon calendars produced this Yellow Solar Seed year. Others have been produced in Germany, Austria, Italy, Russia, the Netherlands, Japan, Brazil, Argentina, Chile, Mexico, and the United States. One such example from the United States is known as the 13 *Moon Natural Time Calendar: Ancient Science for the Art of Now.* Going through its pages one is flooded with a wealth of information, summarized by a quote from traditional Hopi wisdom: "We now face the final test of human consciousness-to restore our oneness with creation. We are going home. It is a process. It is like the seed becoming the green shoot and then the flower."2 To make the calendar change is to restore the oneness with nature. The break with the oneness with nature was codified and thence programmed into an irreversible reality by the Gregorian calendar and mechanical clock. Only the breaking of the pattern of time established by these two instruments will release us again into natural time. The Thirteen Moon/28-Day calendar exists as the tool to demonstrate the contrast between harmony and disharmony, and to assist us, as a species, into effecting this return. If we are going home then we must make a turn away from our present course.

> Through returning to natural Time, we return to our natural state of divinity, living in harmony with the Earth and each other. We return to the nature of our spiritual existence. Realigning with the rhythm of the universe through the 13 Moon Calendar allows us to lay the foundation for the new paradigm of love and unity, replacing the 12:60 paradigm of fear and separation. . . The World Calendar Change Peace Plan strives to establish a new culture of peace on Earth through returning humanity to natural time. It is the goal of the worldwide Foundation for the Law of Time to effect the calendar change by the Day Out of Time (July 25), 2004 . . . Join the global movement to effect planetary change and establish the new culture of peace on Earth! Return to divinity and creative existence. . . Walk the path of the sacred warrior.]

The movement to change the calendar is a living force, a force that is living the New Time. Its premises are so diametrically opposite to those of the technosphere that, especially in the power bases of America and the G-7, one will find scant mention or notice at all of this movement. Yet it is a cultural phenomenon. In Tokyo, San Francisco, Milan, or Moscow one may find signs of the presence of this movement in the underground dance events, the raves or the Earthdance. More and more young people take on the names of their galactic signatures, the 260-day kin code for their date of birth in the synchronic order, names like Lunar Storm or Spectral Night, Galactic Wind or Planetary Dragon. These are all signs of the beginning of galactic culture on Earth, a phenomenon that has only been increasing over the past decade. The point is that the New Calendar Change Movement is a grassroots phenomenon as well as an ideological strategy known as the Campaign for the New Time. These signs are all whispers of the noosphere blowing through the bleak industrial corners of the technosphere.

Behind the more populist front of the Natural Time Movement is the science of the Law of Time, a genuinely noospheric science of peace and harmony. This science is already being studied and put into preliminary practice. As the crisis of the biosphere began to crest in 1987, halfway through the second 28-year phase of the technosphere, the noosphere began to download its measure of time through the discovery of the Law of Time. Invisible to the predominant forces of the technosphere, the science of the Law of Time began to map out an entirely new domain of thought, the synchronic order, a domain that could only be entered and applied through the correct noospheric measure, the Thirteen Moon/28-Day synchronometer.

All science throughout history is based on evolutionary advances predicated on both the premises and limitations of its preceding stages. We are at such an evolutionary moment now. The issue of the calendar change is only the beginning of the emergence of a new whole system scientific order. Understanding the flaws that are built into the present system of purely materialist science, we can proceed to the establishment of the science of the synchronic order. Like the radical simplicity and perpetual harmony of the Thirteen Moon calendar itself, the noospheric science of the synchronic order is predicated on an equally radical mathematical base reflecting and embedded in the 13:20 universal timing frequency itself. And as the underlying order of a whole system science, this mathematics is reflective of the non dualistic premise of a superior moral order. The deviation from universal synchronic time is a moral deviation as well. It is this moral deviation that most strongly prompts the current calendar change movement.

As a profoundly nondualistic, unifying whole system premise, the Law of Time encompasses not only a scientific-mathematical basis of morality, but a theological-eschatological intention that reunifies the various disciplines scattered and separated by the rise and totalitarianism of modern secular science. Because of the historical dominance of the Gregorian calendar, in which modern science is embedded, the dethronement of the calendar-the "devaticanization" of the world, as it were-is an apocalyptic moment.

As the Law of Time defines it, this is merely the apocalypse of the 12:60 timing frequency and all its corrupted institutions, much as foretold in the Book of Revelations. This apocalyptic moment was stunningly supplied by the terrifying collapse of the Twin Towers. As the nerve center of the financial, commodities, and transport sector of the technosphere, and hence of the city itself, the collapse of the towers in hardly more than an hour's time is perfectly reflected in the vivid descriptions of the eighteenth chapter of the Book of Revelations, "Fall of Babylon." Yet as the Mayan prophecy of Pacal Votan also makes absolutely clear, the apocalypse of the 12:60 is the resurrection into the 13:20 timing frequency, the return to natural time, all to be accomplished by A.D. 2012, Teilhard de Chardin's "omega point." The advent of the noosphere is preceded by the post-apocalyptic harrowing of hell. In this harrowing of hell, defined as the span of time fromJuly 26, 2000-July 25,2004, humanity must make a crossing from one time to another. This crossing is much like those we hear

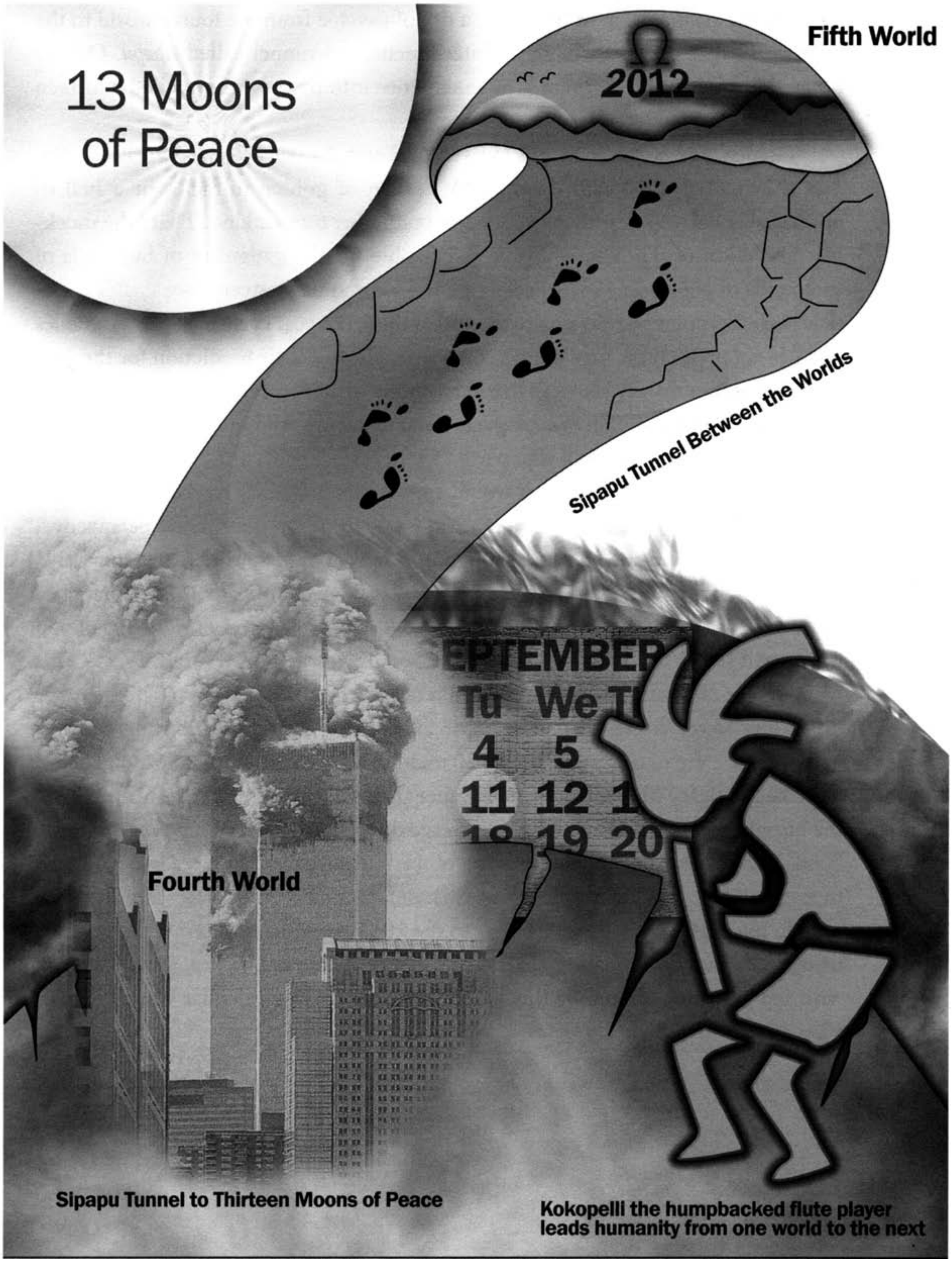
13 Moons of Peace
Fifth World
2012
Sipapu Tunnel Between the Worlds
EPTEMBER
Tu We T
4 5
11 12 1
18 19 20
Fourth World
Sipapu Tunnel to Thirteen Moons of Peace
Kokopelli the humpbacked flute player leads humanity from one world to the next

about in the cosmovision of the Hopi, a rite of passage from the fourth world to the fifth world that occurs through an interdimensional tunnel called *sipapu.* On the other side of that tunnel, there is an emergence into the new time called "thirteen moons of peace."

"Golden Age of Miracles and Heaven on Earth," a story in the *National Examiner* (November 29, 1988), begins: "A heaven of golden miracles or a hell of unspeakable horror is less than twenty-five years away for mankind. That's the shocking prediction of the man who was responsible for the gathering of hundreds of thousands of people in 1987 to celebrate the Harmonic Convergence. . ." The article makes its point by referring to the end of the 5,125 year cycle in 2012. "Argiielles said Mayan predictions have always come true. He said their prediction for the year 2012 is that if humanity learns to live in harmony with Nature, a new age of miracles and light will occur. Otherwise plagues, wars, terrorism, and finally, the breakup of the planet into asteroids, will take place."4

As for the Inevitable Event, it would appear that humanity has begun the worst of the prediction-plagues, war, terrorism-yet, from the noospheric perspective, two points mitigate the absolute plunge into hell. The first is that the Inevitable Event punctured a hole in the technosphere eleven years prior to 2012. What does this mean? It means that this is the time of hell on Earth. "Every single one of you must see it [hell]; this is an irrevocable decision of your Lord." (Quran, 19:71) This being so, we can potentially emerge from it before the 2012 omega point. The other mitigating factor is the existence of a 13:20 movement dispersed throughout humanity. Since the noosphereis the 13:20 mind of Earth, the Thirteen Moon Calendar Change Peace Movement is the advance presence heralding the transition from the biosphere to the noosphere. This passage through hell is of a four-year duration, 2000-2004. The exponential implosion of the technosphere has already begun. But to avoid a long, debilitating siege of barbarism, humanity must select a positive direction by 2004 so that it can make it to the 2012 deadline as a species returned to living in harmony with nature. This it can only do by the dramatic and radical break with the past time offered by the Thirteen Moon/28-Day calendar Change Peace Plan.

Campaign for the New Time

4:7::7:2013

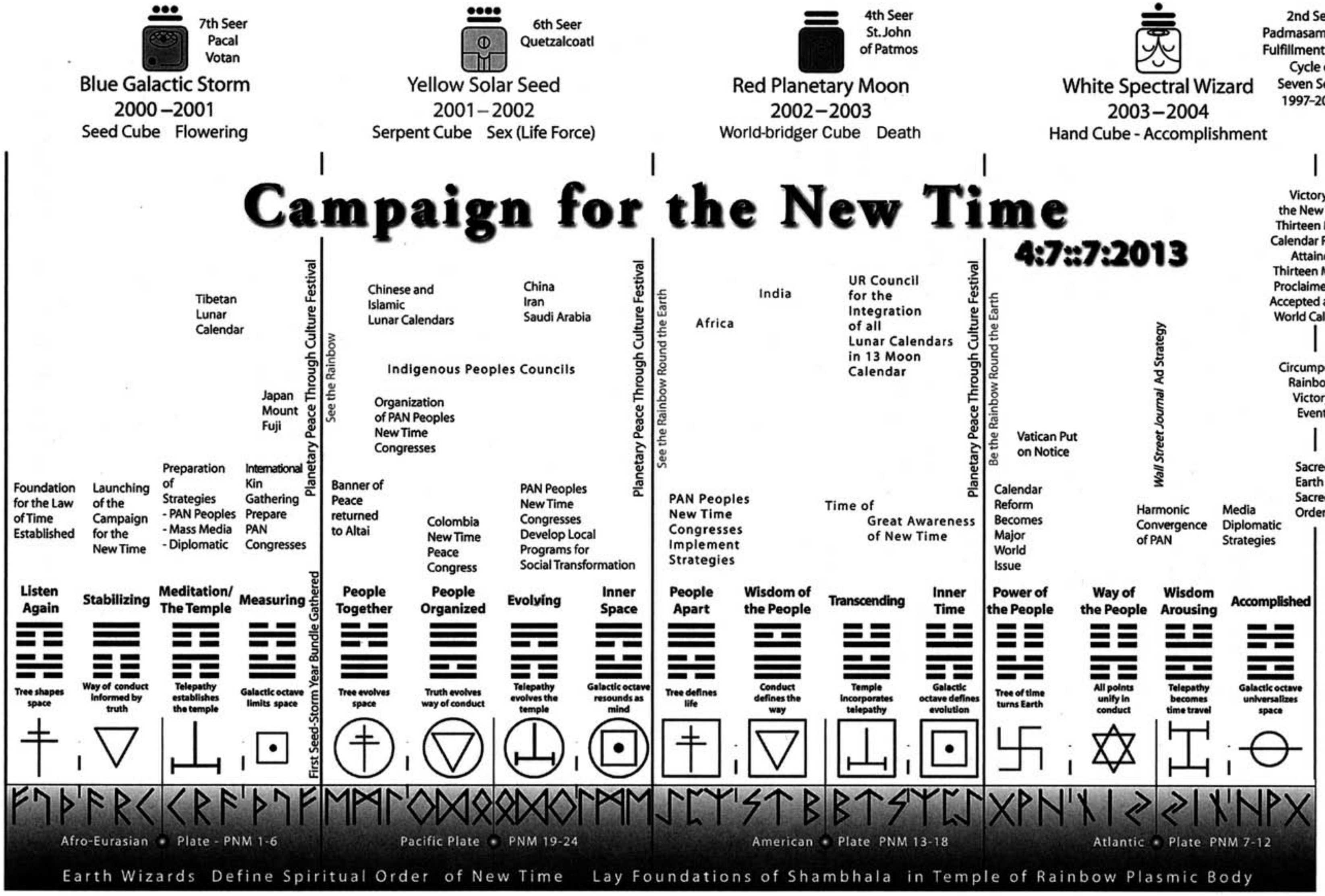

7th Seer
Pacal
Votan

Blue Galactic Storm
2000–2001
Seed Cube Flowering

Tibetan
Lunar
Calendar

Japan
Mount
Fuji

Foundation
for the Law
of Time
Established

Launching
of the
Campaign
for the
New Time

Preparation
of
Strategies
- PAN Peoples
- Mass Media
- Diplomatic

International
Kin
Gathering
Prepare
PAN
Congresses

Planetary Peace Through Culture Festival

First Seed-Storm Year Bundle Gathered

Listen Again – Tree shapes space

Stabilizing – Way of conduct informed by truth

Meditation/ The Temple – Telepathy establishes the temple

Measuring – Galactic octave limits space

6th Seer
Quetzalcoatl

Yellow Solar Seed
2001–2002
Serpent Cube Sex (Life Force)

See the Rainbow

Chinese and
Islamic
Lunar Calendars

China
Iran
Saudi Arabia

Indigenous Peoples Councils

Organization
of PAN Peoples
New Time
Congresses

Banner of
Peace
returned
to Altai

Colombia
New Time
Peace
Congress

PAN Peoples
New Time
Congresses
Develop Local
Programs for
Social Transformation

Planetary Peace Through Culture Festival

People Together – Tree evolves space

People Organized – Truth evolves way of conduct

Evolving – Telepathy evolves the temple

Inner Space – Galactic octave resounds as mind

4th Seer
St. John
of Patmos

Red Planetary Moon
2002–2003
World-bridger Cube Death

See the Rainbow Round the Earth

Africa

India

UR Council
for the
Integration
of all
Lunar Calendars
in 13 Moon
Calendar

PAN Peoples
New Time
Congresses
Implement
Strategies

Time of
Great Awareness
of New Time

Planetary Peace Through Culture Festival

People Apart – Tree defines life

Wisdom of the People – Conduct defines the way

Transcending – Temple incorporates telepathy

Inner Time – Galactic octave defines evolution

White Spectral Wizard
2003–2004
Hand Cube - Accomplishment

Be the Rainbow Round the Earth

Vatican Put
on Notice

Wall Street Journal Ad Strategy

Calendar
Reform
Becomes
Major
World
Issue

Harmonic
Convergence
of PAN

Media
Diplomatic
Strategies

Power of the People – Tree of time turns Earth

Way of the People – All points unify in conduct

Wisdom Arousing – Telepathy becomes time travel

Accomplished – Galactic octave universalizes space

2nd Seer
Padmasambhava
Fulfillment of the
Cycle of
Seven Seers
1997–2004

Victory of
the New Time
Thirteen Moon
Calendar Reform
Attained
Thirteen Moons
Proclaimed and
Accepted as New
World Calendar

Circumpolar
Rainbow
Victory
Event

Sacred
Earth
Sacred
Order

Afro-Eurasian • Plate - PNM 1-6

Pacific Plate • PNM 19-24

American • Plate PNM 13-18

Atlantic • Plate PNM 7-12

Earth Wizards Define Spiritual Order of New Time Lay Foundations of Shambhala in Temple of Rainbow Plasmic Body

THE BIRTH OF THE CAMPAIGN FOR THE NEW TIME

> *"We inspired Moses: 'Lead My servants out,*
> *and strike for them a dry road across the sea.*
> *You shall not fear that you may get caught, nor*
> *shall you worry.*
>
> *"Pharaoh pursued them with his troops, but the*
> *sea overwhelmed them, as it was destined to*
> *overwhelm them."*
>
> -QURAN, 20:77-78

The Thirteen Moon/28-Day calendar is the dry path across the waters of the 12:60 chaos. Holding surely to this path, "Pharaoh's troops will be overwhelmed," as it is destined, for the truth always prevails and when falsehood is exposed, falsehood perishes. The 12:60 frequency enshrined in the Gregorian calendar represents an artificial, hence false notion of time. The Pharaoh's troops are the forces of the technosphere that, embedded in the artificial structure of the Gregorian calendar, are destined to be overwhelmed and overcome by the fury unleashed both by human nature and nature itself in response to decades, if not centuries, of irresponsible despoiling of the Earth. The dry path of the Thirteen Moon/28-Day calendar is the only path broad enough for all humanity to leave the ever increasing ocean of chaos that characterizes human existence bound in the collapsing 12:60 matrix born of Gregorian civilization, the technosphere. Yes, the dry path across the waters of the 12:60 chaos provided by the Thirteen Moon/28-Day calendar truly is for the salvation of life on Earth.

Ever since 1993, the Thirteen Moon Calendar Change Peace Movement has consciously been drawing from the infinite riches of the noosphere to create not only the mechanism for making a shift in the frequency and hence bring stabilization to the biosphere, but has been anticipating the biosphere-noosphere transition into the *Pax Cultura, Pax Biospherica.*

Following in the footsteps of its predecessor, the new calendar movement has dutifully contacted the United Nations and the Vatican and convened various congresses and delivered ultimatums in an effort to draw attention to mankind's time malaise. Now that the Inevitable Event has fully disengaged the human mind from the reality it thought it was pursuing in the new millennium, the second definitive wave of the World Thirteen Moon Calendar Change Peace Movement has also been launched: the Campaign for the New Time.

New Globally Unifying Harmonic Standard

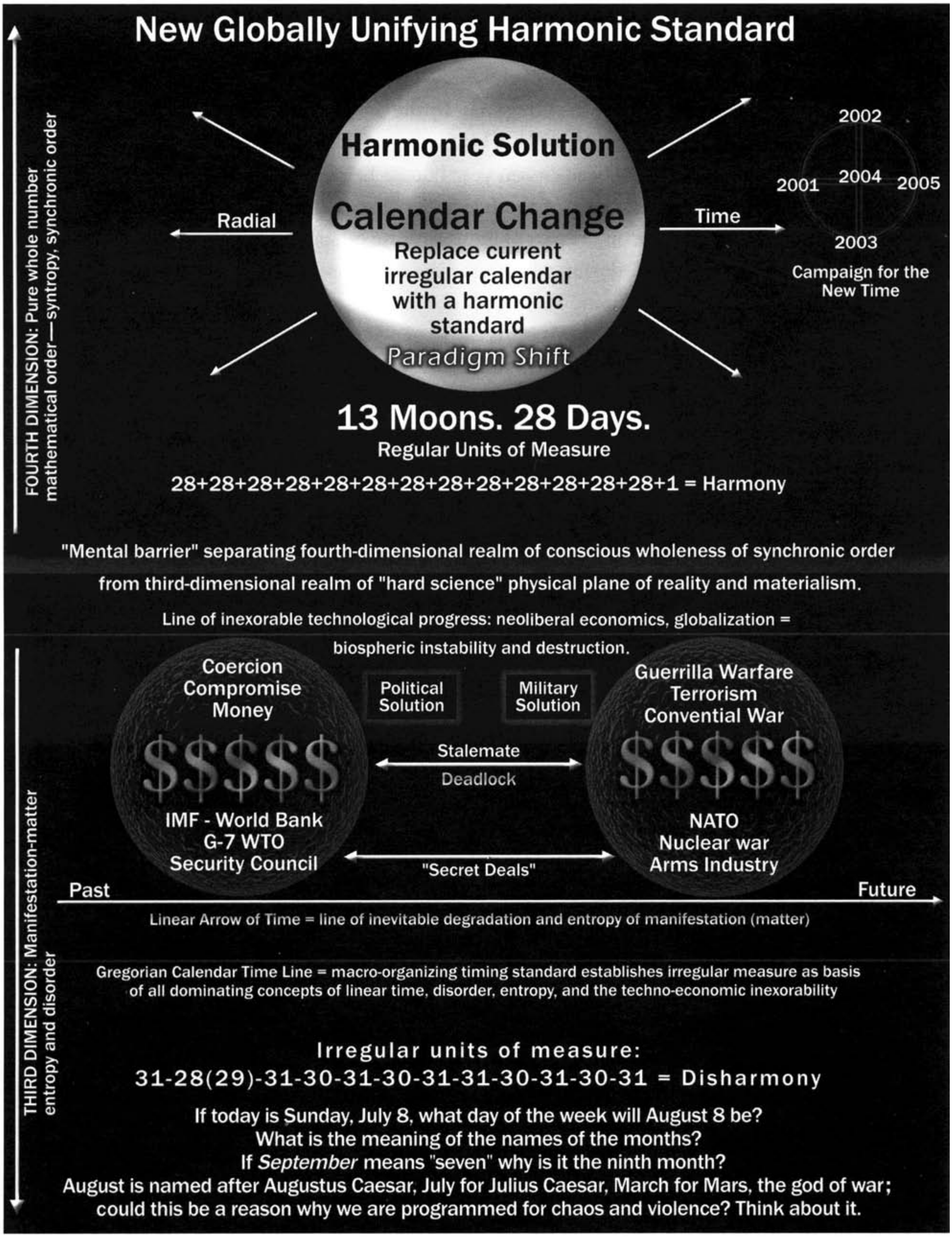

FOURTH DIMENSION: Pure whole number mathematical order—syntropy, synchronic order

Harmonic Solution

Calendar Change

Replace current irregular calendar with a harmonic standard

Paradigm Shift

Radial

Time

2002

2001 2004 2005

2003

Campaign for the New Time

13 Moons. 28 Days.

Regular Units of Measure

28+28+28+28+28+28+28+28+28+28+28+28+28+1 = Harmony

"Mental barrier" separating fourth-dimensional realm of conscious wholeness of synchronic order from third-dimensional realm of "hard science" physical plane of reality and materialism.

Line of inexorable technological progress: neoliberal economics, globalization = biospheric instability and destruction.

Coercion
Compromise
Money

$$$$$

IMF - World Bank
G-7 WTO
Security Council

Political Solution

Military Solution

Stalemate

Deadlock

"Secret Deals"

Guerrilla Warfare
Terrorism
Convential War

$$$$$

NATO
Nuclear war
Arms Industry

Past

Future

Linear Arrow of Time = line of inevitable degradation and entropy of manifestation (matter)

THIRD DIMENSION: Manifestation-matter entropy and disorder

Gregorian Calendar Time Line = macro-organizing timing standard establishes irregular measure as basis of all dominating concepts of linear time, disorder, entropy, and the techno-economic inexorability

Irregular units of measure:

31-28(29)-31-30-31-30-31-31-30-31-30-31 = Disharmony

If today is Sunday, July 8, what day of the week will August 8 be?
What is the meaning of the names of the months?
If *September* means "seven" why is it the ninth month?
August is named after Augustus Caesar, July for Julius Caesar, March for Mars, the god of war; could this be a reason why we are programmed for chaos and violence? Think about it.

This campaign consists of a preparatory four-year cycle, 2000-2004, the Harrowing of Hell, which is the duration of the campaign itself, plus a one-year trial period, 2004-2005. The goal of the four-year campaign is universal calendar reform, the replacement of the Gregorian by the Thirteen Moon Calendar, effective Blue Crystal Storm, July 26,2004. This final year of the Blue Crystal Storm, 2004-2005, also completes a Seed-Storm year bundle begun in 2001, fateful year of the Yellow Solar Seed, the year of the Inevitable Event. To ensure the success of this unprecedented move on the part of humanity, the calendar reform must include the one-year trial period, during which time the entire species must be operating by a new standard, the harmonic Thirteen Moon/2 8-Day synchronometer. "If you never try it, you'll never know what it is," is the slogan that describes the campaign's target year: July 26, 2004-July 25, 2005, an entire year without the Gregorian "30 days hath September. . . ," but instead, Thirteen Moons of Peace.

To realize its ambitious goal of universal calendar reform, the Campaign for the New Time must impress itself on the human consciousness as the all-unifying campaign for peace to overcome the destructive forces unleashed by the technospheric collapse. Fortunately, much groundwork has already been laid. The codes of the Law of Time have been formulated and compiled. Preliminary practices to engage the synchronic order of the noosphere have already enlisted humans from around the planet. What the Vatican has tried to ignore and what the United Nations has been unable to fully engage must now be brought to the forefront of all humanity as the harmonic solution. The harmonic solution is a set of logical arguments that should be studied and rehearsed by everyone who seriously wishes to even consider the possibility of a Campaign for the New Time. Let us contemplate the following set of logical arguments.

Premise

Currently all conflict resolution of a global scale is attained through one of two poles: a political solution or a military solution. No other alternatives are recognized. According to the Law of Time, there is a third solution, the harmonic solution. Without this third solution, no harmonic resolution is ultimately possible for any conflict on a global scale. The third solution defines the first two solutions as a polarity embedded within a timing standard that is hopelessly irregular and that, therefore, disposes the mind to disharmony. A mind predisposed to irregularity and disharmony is incapable of producing any long-term solution. The basis of the harmonic solution is to change the timing standard from an irregular to a regular one, thereby inducing a timing sensibility that disposes the mind to harmony rather than disharmony.

Thirteen Moon/28-Day Count and Global Conflict Resolution

The inability to attain global peace, much less the resolution of many localized conflicts that have become endemic, is due to a lack of a globally unifying timing standard-the macro-organizing principle of human sensibility. Present-day conflict resolution is represented by two poles, political and military. The military solution consists of guerrilla warfare, terrorism, and outright military conflict, all of which are engendered by problems of inequality; injustice; rampant nationalism; tribal, racial, or religious antagonism; and greed for control of natural resources. Underlying the military solution is an aggressive armaments industry that seeks to hold its own as a capitalist venture, always seeking new markets. The political solution is always fostered as the only cure for the military solution, and is administered through varying degrees of coercion, compromise, and the inevitable lure of money-the IMF and the World Bank are never far behind any political solution.

Neither the military nor the political represent any kind of morally satisfactory or enduring solution because both are functions of and embedded in a timing standard that has disharmony built into it. Even the local timing systems, themselves feedback loops holding in place various belief systems, are coordinated by the dominant disharmonic standard. As long as the current global standard, the Gregorian calendar, is the controlling timing factor, resolution on a global and local level is not only impossible, but increasing disharmony is guaranteed. The harmonic solution calls for the immediate establishment of a globally unifying harmonic standard, the Thirteen Moon/28-Day count. With its absolutely consistent units of measure, this macro-organizing principle is a paragon of harmony and thereby represents the positive, creative, and viable first step to enduring global conflict resolution.

Thirteen Moon/28-Day Paradigm Shift

The harmonic solution, the alternative to the polarity of the military and political solutions, calls for the immediate replacement of the current irregular measure with the harmonic standard of the Thirteen Moon/28-Day count. This replacement constitutes the long-awaited and long-prepared-for paradigm shift in human consciousness. The very analysis and concern that engendered the first movement to replace the Gregorian calendar-the effects of an irregular measure on human behavior-is now so dire that the concluding step must be taken. To quote Einstein once again, "Since the advent of the nuclear age, everything has changed but the way people think; thus we drift toward unparalleled catastrophe." With the analysis of the Law of Time we know why thinking has not changed-the human mind is the pawn of an aberrant timing mechanism. To change the way people think, make one little change in their everyday lives: Change the calendar. If we could send a man to the moon, we can change the calendar. This change will be the first step-the rest will come easily because with this one change we shall have changed our minds and entered into a New Time. To make this one little change a target date is needed, and the people need to mobilize as a force of synchronization for peace to mandate this change. The target date is July 26, 2004--the mobilization has already begun.

New Worldview of the Law of Time

The Thirteen Moon/28-Day count is the pragmatic application of the Law of Time. The discovery of the Law of Time was necessitated by evolution as the only way that could offset the deleterious effects of an erroneous timing standard on human consciousness. The Law of Time takes the radical view that time is a frequency, the universal frequency of synchronization; that the purpose of time is to synchronize; and that, therefore, a calendar is in harmony with the Law of Time to the degree to which it maximizes opportunities for synchronization. From the perspective of the Law of Time, the current macro-organizing standard is anything but synchronous, and hence contrary to the laws of nature. Synchronization produces harmony, harmony is a function of beauty, and beauty is the natural result of the Law of Time: T(E) = Art, energy factored by time equals art, where art is the entire spectrum of phenomenal reality in which even the most menacing-looking creature is elegantly constructed.

National Geographic recently stated: "Beauty seems to be an intrinsic part of nature and perhaps even the organizing principle of reality. Scientists in testing their theories, invariably find that the simplest, most elegant, most beautiful is the correct one. Rainbows, butterflies and the periodic table are some examples of intrinsic beauty. The world will be saved by beauty."5 The Law of Time agrees. The world will be saved by beauty, and the proponents of peace through culture will be in the vanguard. The Law of Time formalizes the perception of the intrinsic beauty of nature- T(E) = Art-and thereby provides a scientifically solid critique of the ugliness and social disharmony fomented by the irregular and artificial timing standards that control humanity today. By the same formulation, the Law of Time also establishes a vast and comprehensive definition of the fourth dimension, which was wanting until now precisely because of the embedding of all current perceptions in the erroneous timing standard.

The mathematical order underlying the universal frequency of synchronization, 13:20, describes a new order of reality, the synchronic order. The synchronic order governs the fourth-dimensional order of reality much as the laws of gravity, thermodynamics, and the special theory of relativity govern the third-dimensional order of reality. By taking into account the nature of the synchronic order, humanity's views of the third dimension will inevitably be altered to a greater conformity with the harmony of the fourth dimension. The fourth dimension is defined as the primary mathematical order of time, governed by a syntropic whole number system of ratios and fractals operating in a radial and nonlinear manner.

The third dimension is the world of manifestation and matter that inevitably entropizes. Losing sight of the inherent harmony of the synchronic order from which it is actually inseparable, the human fixation on the increasing entropic anomalies of the third dimension produces problems such as how to reckon what appear to be the extra hours, minutes, and seconds that occur every time the Earth orbits the sun. To this seeming dilemma, the Law of Time asserts that number precedes manifestation, harmony precedes materialization, therefore manifestation must accord with number, and materialization must reflect an original harmonic order. Entropy and degradation cause manifestation to deviate from number and mathematical law. But for mind to be stable, mind must accord with and conform to mathematical law rather than with entropy and degradation. The timing standard in use causes the mind to conform with entropy and degradation. This is what must be corrected.

Harmonic Solution-Harmonic Resolution

Harmony can only produce harmony. From order comes order. From disorder you can only expect chaos. From disharmony can only come disaster. Irregular measure entropizes. Harmonic measure syntropizes. These are the contrasting values brought into focus by replacing an irregular macro-organizing principle with a regular one. Implementation of a globally unifying harmonic standard is the only hope for resolving all conflicts in time. A harmonic standard lifts all conflicts to a higher level dialog of unification in time. The harmonic solution, the World Thirteen Moon Calendar Change Peace Plan, provides: the opportunity to call an immediate and universal cease-fire, as well as a halt to the otherwise inexorable process of entropic degradation of society and environment; and offers the resolution to establish new agreements, to renegotiate old treaties, and to create a new globally binding and unifying covenant for all people in a new time.

It must be understood that to change the calendar is to delegitimize all the institutions embedded in it. This radical but peaceful delegitimization of the old order is what is needed to break the deadlocks, stalemates, and historically outmoded treaties and agreements now binding on humanity. The new globally unifying standard, the Thirteen Moon/28-Day count, will legitimize and institutionalize harmony and thus create the advent of a synergistic order capable of resolving entropic disorder because it is reflective of and rooted in the primordial whole number structure of the synchronic order of reality. In its entirety, this represents a new step and direction for human evolution-a step without which there might not be any creative future for humanity and the planet.

How to Accomplish World Peace Now

The analysis, the historical precedents, and now the affirming authority of the Law of Time render the harmonic solution as the only way to immediately bring about World Peace. All that is lacking is the will to synchronization among all good-hearted people, groups, and organizations otherwise committed to peace, a better environment, and human spiritual well-being in general. How to accomplish World Peace is to galvanize the human will to synchronization on behalf of the establishment of a genuine time of harmony and peace. This can only come about through the simple but highly radical and paradigm-altering act of exchanging the current macro-organizing system of the Gregorian calendar for the harmonic order of the Thirteen Moon/28-Day count. All other localized calendar systems would be kept intact, but they would be coordinated by a new standard, a standard that promises harmony instead of disaster, order instead of chaos, peace instead of war.

Such is the harmonic analysis and solution of the Law of Time, which can be summed up by a simple demonstration of the irregular Gregorian calendar contrasted with the regular Thirteen Moon calendar, as well as the question: "Which do you prefer, harmony or disharmony?"

If you answer this question on behalf of harmony, and you agree with the foregoing premises and analyses, you must move into action. You need to get a new calendar, a Thirteen Moon/28-Day calendar and find out for yourself: noosphere or necrosphere, the choice is yours. The option opened by the Inevitable Event and the collapse of the technosphere is precisely that: necrosphere, the transformation of the biosphere into a sphere of death-a dead planet-or the noosphere, the mental envelope of the Earth, governed on the human plane by the Thirteen Moon/28-Day calendar. The difference between the necrosphere and the noosphere is the difference between a machine and the Earth as a whole system. You cannot even really say there is a comparison. But what we can say most emphatically is that the Thirteen

Moon/28-Day calendar and the shift to the 13:20 frequency is the pure manifestation of the Law of Time in human affairs. It is the only antidote to the destructive grip that the 12:60 timing frequency now has upon the biosphere through its thoroughly artificial instrument, the technosphere. (See plate 7, Necrosphere or Noosphere.)

THE FINAL DECISION: NOOSPHERE OR NECROSPHERE?

As these lines are being written, seven weeks have passed since the Inevitable Event. The analysis of this event from the perspective of the biosphere, the temporary structure of the technosphere, and the advent of the noosphere was a spontaneous inspiration. Each morning, often well before dawn, I would be called to this technospheric device, an iMac, to put the noospheric mandate down into words. In this process, the matter of prophecy kept occurring and could not be avoided. Even the issue of calendar change increasingly takes on the quality of a prophetic enactment. What is prophecy from the noospheric perspective?

Prophecy is the release of information according to the psi bank timing program and in relation to degrees or levels of consciousness positioned at different points in time. The points in time are human channels. That is, there are the prophets who declare the prophecies by whatever means or forms, and then there are the decoders of the prophecies, those able to read the signs according to the prophecies with the aid of different symbolic lexicons. Prophecy is always startling to the mass mind, which is ignorant of the true nature of time. A prophetic event may seem totally random, when it is not. Someone may design a plot, such as that of the Inevitable Event, with whatever fantastic notions and motives. Yet the occurrence may go beyond being merely a matter of human manipulation. The humans were but the instruments of sets of cause-and-effect chain reactions established within the structure and evolution of the biosphere and the playing out of human karma.

According to the psi bank timing program governed by the Law of Time, the technosphere could only have a limited duration. Two perfect 28-year cycles of the Gregorian calendar program, leading to the first year of the third millennium, marked the duration of the technosphere. Because this cycle concluded with the beginning of the third Christian millennium, a singular event was poised in time to attract and catalyze a host of prophetic moments long prepared for and sealed in the psi bank by Christian, Islamic, and even Mayan prophecies. Such was the Inevitable Event. Within the motivations of the perpetrators of the Inevitable Event, the biosphere found its perfect instrument of divine purpose. That there were nineteen hijackers conforms to the Quranic nineteen code expressed in the enigmatic verse, "Over it is nineteen."

(74:30) The following verse begins, "We appointed angels to be guardians of Hell, and we assigned their number [19] . . ." This verse ends by asking, "What did God mean by this allegory? . . . None knows the soldiers of your Lord except He. This is a reminder for the people." (74: 31)

Certainly in the success of their mission, whether they were conscious of it or not, the nineteen Muslim hijackers-who organized themselves as four leaders (pilots), three logistics people, and twelve "soldiers"-fulfilled their function of being the prophetic emissaries of the nineteen guardian angels who watch over Hell. And indeed the gates to Hell have been opened. In the world of the secular mind of the technosphere, no one wanted to talk about the apocalypse or Armageddon, though the jihad was brought up quite often as a factor in the thinking of the Islamic terrorists.6

Whether it is seen as a prophetic or karmic event, as we stated earlier, there is no question that the timing programs of two civilizations, the Gregorian and the Islamic, were on a collision course. "The civilization of Islam is radically so different from that of Western civilization which dominates the world today. The two are different in their description oflife as well as the foundation on which they base such a description. The difference between the two civilizations is so essential that they have developed in ways which are radically contradictory to each other."7 So wrote the eloquent Saudi historian, Muhammad Husayn Haykal in his essay "Islamic Civilization," as presented in his masterful biography, *The Life of Muhammad.* And what is at the root of this difference? A separation of church and state, which in the West led to a profound materialism and economic worldview.

Haykal continues, "The victory of materialist thinking was largely due to the establishment of Western civilization primarily upon an economic foundation. This situation led to the rise in the West of a number of worldviews which sought to place everything in the life of man and the world at the mercy of economic forces. . ." As for the sanctity of its economic beliefs, Haykal concludes that the West has tied its economic activity "hand and foot by public laws and commanded that every Western state and army prevent any violation of economic laws with all the power and coercive means at its disposal."8 Given this insightful assertion by an Islamic thinker written a half-century ago, it is not at all surprising that the terrorists concentrated their attacks on the most visible symbols of Western global materialism and economic and military imperialism, the Twin Towers and the Pentagon. It is also of note that one of the logistics men on the team of nineteen hijackers, Majed Moqed, had been a professor of economics at King Saudi University in Riyadh.

However we see this event, the most important point is that it actually happened, no matter who caused it or how. In this regard the collapse of the World

Trade Center towers and the penetration of the Pentagon was the Inevitable Event both from the point of view of the biosphere and from that of prophecy as mandated by at least two major timing programs, the Gregorian and the Islamic. Of course, prophecy is a matter far from the consideration of the technocratic mind, so the immediate aftereffect is a heightened nationalism, patriotism, and commitment to war. But what is going on from the noospheric point of view is not so much the combat of different armed camps defending or attacking competing ideologies, but an end game, the biomass of the technosphere maneuvering against time in a self-destructive process that is really nothing more than the continuing collapse of that which it is defending.

How long will this collapse take? With the weapons of mass destruction currently at the disposal of the combatants, it shouldn't take too long before the human race realizes that if it doesn't stop the self-destructive process, it will soon be curtains for all life on Earth. A more significant question, if we wish to make the transition to a new time, is: How will the self-destruction end, and how can it be stopped-or can it? It is easy to see how wars begin, but no one can tell how they will end.

THE SOLUTION TO WAR: PEACE THROUGH CULTURE, PEACE OF THE BIOSPHERE

At the end of the First World War, the League of Nations was formed to keep the Second World War from happening. The League of Nations was unable to both effect the calendar reform it promoted and keep the Second World War from happening. These are not unrelated events. The Second World War effectively eliminated the League of Nations, and toward its end a new organization, the United Nations, was formed. Like the League of Nations, the United Nations was formed to prevent the next world war from happening. Within eleven years, the United Nations tabled the calendar reform begun by the League of Nations. Coincident with the formation of the technosphere, the United Nations has now also completed fifty-six years of existence. With the expanding "war against terrorism," is its purpose now over as well? The only way the United Nations can survive is to maintain its humanitarian agencies, for the age of nationalist politics and nation-states is over, succeeded by the politics of terrorism, the last politics of history. What new kind of global organization will we need at the end of the War on Terror, or World War III?

What we will need is an entirely new agency, a Planetary Biospheric Assembly that will model and create a sustainable *Pax Cultura, Pax Biospherica.*Given the failure to enact a calendar reform on the part of both the League of Nations and the

United Nations and in consideration of the chaos now clearly embedded in the Old Time, only an agency or instrument that fully embraces and promotes the calendar change will be able to move into the genuinely peace-fulfilling role that such an organization must take. If the United Nations could accept the route of calendar change as the only way to bring all humanity to a new point of discourse, then it could take on a renewed life. But if it cannot rise to this occasion, then another organization must take its place, the Planetary Biospheric Assembly. Even so, when we speak of the *Pax Cultura, Pax Biospherica,* we are speaking of a new human covenant, one in which the biosphere takes precedence over all nationalist corporate interests, and in which peace as a manifestation of culture is the primary value of the new time. Such a covenant would be a biospheric covenant, an agreement and understanding that, first of all, the human is a member of the biosphere, and second, that in consciously accepting the calendar change, the human is also in the process of becoming *Homo noosphericus.* Henceforth, it will be art and not money, bioregions and not nation-states, that will be the cause of social organization. Only in this way will humanity be guaranteed a true post-historical future.

Assuming the calendar change occurs on Blue Crystal Storm 2004, the first step in ~ngendering the full manifestation of the noosphere will be the dismantling of the technosphere. This can only be accomplished if a full reorganization of the human in time, inclusive of an adjudication of inequalities and a redistribution of wealth, begins to occur during the 2004-2005 Crystal Storm year. The following four years, 2005-2009, a genuine First World Peace, would see the constructive transformation of the five-part technospheric system. All the apparatus of heavy industry, energy, and the transport system would be turned like compost, to be replaced by nonpolluting technologies and means of transport, while the cities would be transformed into urban gardens with the human race once more becoming decentralized. This process itself, a massive planetary engineering project with an eye on the full development of geochronomancy, would prove to be the greatest and most thoroughly positive transformation in human history. By 2009 humans reorganized by the Law of Time and having finally arrived at the "path beyond technology," will be enjoying universal telepathy, the noosphere-life beyond the Internet. By 2012 the humans will realize peace through culture as a permanent value, and by 2013 Earth will be galactically synchronized through a fully activated noosphere into the beginning of an entirely new geological era, defined by Vernadsky as the Psychozoic Era.

Pax Cultura, Pax Biospherica refers to both the new era of humanity following the devastation of the Final War and the governing body of bioregional councils (not nation-states) comprising the new world "government," the Planetary Biospheric

Arcturus Victory Strategy
Pax Cultura, Pax Biospherica

Telektonon, the Theater of Time, provides the journey through the Cube of the Law: Plane of Mind, Plane of Spirit, Plane of Will, and Divine Source Telektonon—each plane Seed-Storm year bundle, an entire universe of noospheric time!

Plane of Spirit: First Seed-Storm year bundle, NET week 9 Seed–12 Storm, 2001–05

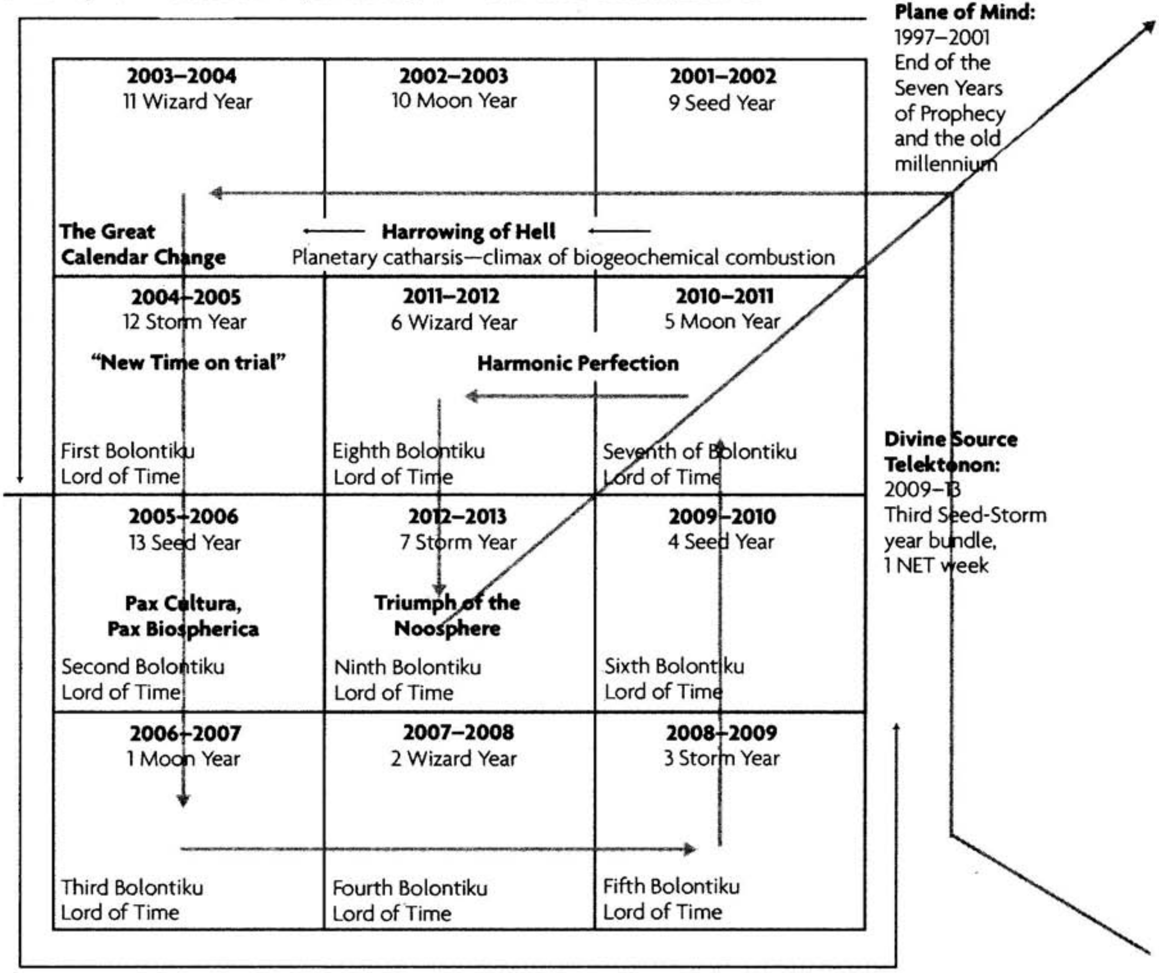

Plane of Mind:
1997–2001
End of the
Seven Years
of Prophecy
and the old
millennium

2003–2004 11 Wizard Year	**2002–2003** 10 Moon Year	**2001–2002** 9 Seed Year
The Great Calendar Change	**Harrowing of Hell** Planetary catharsis—climax of biogeochemical combustion	
2004–2005 12 Storm Year **"New Time on trial"** First Bolontiku Lord of Time	**2011–2012** 6 Wizard Year **Harmonic Perfection** Eighth Bolontiku Lord of Time	**2010–2011** 5 Moon Year Seventh of Bolontiku Lord of Time
2005–2006 13 Seed Year **Pax Cultura, Pax Biospherica** Second Bolontiku Lord of Time	**2012–2013** 7 Storm Year **Triumph of the Noosphere** Ninth Bolontiku Lord of Time	**2009–2010** 4 Seed Year Sixth Bolontiku Lord of Time
2006–2007 1 Moon Year Third Bolontiku Lord of Time	**2007–2008** 2 Wizard Year Fourth Bolontiku Lord of Time	**2008–2009** 3 Storm Year Fifth Bolontiku Lord of Time

Divine Source Telektonon:
2009–13
Third Seed-Storm
year bundle,
1 NET week

Plane of Will: Second Seed-Storm year bundle, 1 NET week, 13 Seed–3 Storm, 2005–09
Time of Transformational Destructuring—First World Peace

Assembly. The main purpose of the new governing body would be to monitor the advent of the noosphere. Overcoming Northern Hemisphere bias, the *Pax Cultura, Pax Biospherica* will not find any permanent seat, but will begin its existence as a planetary noospheric congress in the Southern Hemisphere. All of this could be made possible only because of the completion of the calendar change. So let us assume the best. The how-to of paradigm shift begins with the Thirteen Moon calendar.

If you follow the Thirteen Moon calendar or tell others to follow it with you, what is the payoff? The payoff is that you get to become a noospheric chip. Only a noospheric chip can bring Heaven on Earth, which is nothing more than full human participation in the noosphere. This is not at all a theoretical matter, but something that can be practiced and followed on a day-to-day basis. Fortunately, a path has been shown, complete with travel aids.

CONCISE *HOMO NOOSPHERICUS* HOW-TO TIME TRAVELER'S GUIDE TO THE BIOSPHERE-NOOSPHERE TRANSITION

This analysis of the Inevitable Event could never have occurred if the Law of Time hadn't been discovered and defined. The Inevitable Event simply supplied the Law of Time with precisely the right historical moment to provide a thorough analysis and explanation of time as the whole system in which we humans live and operate, inclusive of the biosphere, the technosphere, and the noosphere. As the chief discoverer of the Law of Time, assisted by my wife and an increasing number of other humans who have engaged the Law of Time, I have sought only to make a presentation of matters in which we, as a species, are thoroughly immersed, and thus find difficult to comprehend. Such is the technosphere, and such also is the medium of time. Since the publication of *The Mayan Factor* and my pursuit of the underlying mathematics of the Mayan calendar, I have remained outside the conventional arenas of discourse and communication, preferring instead to remain more anonymous in order to test the various hypotheses of the Law of Time.

In this process over the past twelve years, the Law of Time has precipitated from the noosphere various tools and didactic texts, all for the purpose of enabling us to make the shift from the old paradigm of the time of war to the new paradigm of the time of peace. These tools and principles are actually practiced on a daily basis, and so provide a pragmatic means of applying the Law of Time. Because of my vows and need for scientific "purity" of the experiment I have been living, I was involved at all times in doing my utmost to remain in the 13:20 frequency. For this reason, the tools and texts prompted by the noosphere to define the Law of Time were, for the most part, all produced outside the commercial mainstream. Now a great effort must be taken to make these tools and texts available to humanity in order to assist in the transition from *Homo historicus* to *Homo noosphericus*. By presenting a catalog of these tools and texts we only mean to suggest and define for the reader the scope and existence of a preliminary science of time and telepathy.

Part I. The Tools: Establishing Noospheric Earth Time (NET)

THIRTEEN MOON/28-DAY SYNCHRONOMETER

The harmonic standard of noospheric time, the Thirteen Moon/28-Day synchronometer, is a perpetual, harmonic calendar. It is called a Moon calendar because it is based on the female 28-day menstruation cycle, which is also the average lunar cycle. In actuality the moon goes around the Earth thirteen times a year. This means that the Thirteen Moon calendar is a genuine solar-lunar calendar that measures the Earth's orbit around the sun by the apsidallunation cycle of twenty-eight days. Thirteen perfect months of 28 days = 52 perfect weeks of 7 days = 364 days. The 365th day is called the Day Out of Time because it is no day of the week or month at all. This day, which falls on the Gregorian correlate date ofJuly 25, is a day for forgiveness and the artistic celebration of life and freedom. The synchronization or New Year's date of the Thirteen Moon calendar is July 26, and corresponds to the rising of the great star Sirius. This makes the Thirteen Moon calendar a tool for harmonizing ourselves with the galaxy. Of course, the Thirteen Moon calendar also synchronizes the phases of the moon in a pattern that demonstrates its regularity and order.

One of the great advantages of the Thirteen Moon synchronometer is that day/date calculations are amazingly simple. Any day of any week is the same for any Moon, any year. The first day of every Moon is always a Thirteen Moon Sunday (Dali). The last day of every Moon is always a Thirteen Moon Saturday (Silio). In the Thirteen Moon calendar the obscurely named Gregorian months are replaced by names that correspond to a fourth-dimensional cosmology of time. Each Moon also has a totem animal. The traditional names of the days of the week are replaced by galactic names that describe seven primary plasmas-electronically charged particles that activate our magnetic fields. The plasmas and their symbolic seals appear above the days of the week. The Gregorian calendar makes day/date calculations very difficult because the months are of uneven measure so the days and dates of the week vary from month to month and year to year. The Thirteen Moon calendar is truly unique because it is synchronized with the Harmonic Module, the measure of the universal 13:20 timing frequency.

THE HARMONIC MODULE

The basis of the Tzolkin or sacred 260-day count of the Maya, the Harmonic Module is the 13:20 synchronization gauge and permutation table that consists of thirteen sequences of twenty icons or solar seals and twenty sequences of thirteen galactic tones, 1-13. The resulting 260 permutations, combined with the perfect harmony of the Thirteen Moon calendar, give each day a unique quality of tone and icon known as a kin or one NET minute. The two cycles-Thirteen Moons/28 days and the 260-day Harmonic Module-perfectly mesh every fifty-two years to complete one NET year. In one NET year of 18,980 kin, no two days are the same! Each of the units in the Harmonic Module is called a kin, the base unit of synchronic measure. Articulating the Harmonic Module is a pattern of fifty-two Galactic Activation Portals. See if you can find the radial sequence of thirteen sets of four, counting inward from the corners. Notice that the numbers of each set of four equals 28. Thirteen sets x 28 = 364, the number of days in the Thirteen Moon calendar! (See image containing the Harmonic Module on page 170.)

WAVESPELL

The thirteen-unit form constant of fourth-dimensional time is the wavespell. It is important to see the Thirteen Moon calendar in this modular format, Planetary Service Wavespell, for then it makes more sense to speak of the calendar as a synchronometer. The wavespell's form describes the motion of time as a fractal fourth-dimensional cosmology. The numbers 1-13 in the dot bar notation that code the thirteen positions are known as tones. The names of the tones in sequence define the cosmology, which is also evident in the very structure of the wavespell. The thirteen tones give their names to the thirteen Moons. The structure consists of two gates, the Magnetic and the Cosmic, the first and the thirteenth moons/tones, which correspond to the two entrance and exit gates or portals. Complementing the two gates are the two towers, the fifth or Overtone tower and the ninth or Solar tower. The two gates and the two towers, tones 1, 5, 9, and 13, articulate the structure of the wavespell. Between the places of articulation are three sets of chambers. The first set, tones 2-3-4, helps establish the wavespell; the second set, tones 6-7-8, helps extend the wavespell; and the third set, 10-11-12, serves to convert the wavespell. (See plate 2, Planetary Service Wavespell.)

The structure of the wavespell also defines a four-dimensional geometry of time that connects different points of the wavespell over time. Each of these dimensional geometries is known as a pulsar. It is useful to know the pulsars in extending the mind synchronically over different timing sequences. The fourth-dimensional pulsar connects the four points of articulation, 1-5-9-13, or Magnetic, Overtone, Solar, and Cosmic Moon/tones, and is called the fourth-dimensional time pulsar. The first-dimension life pulsar connects the positions 2-6-10, or Lunar-Rhythmic-Planetary Moon/tones. The second-dimensional sense pulsar connects the 3-7-11, or Electric-Resonant-Spectral Moon/tones; and the third-dimensional mind-form pulsar connects the 4-8-12, or Self-existing-Galactic-and -Crystal Moon/tones. Remember, the wavespell accommodates any sequence of kin whose tones correspond to 1-13, and may be either a thirteen-day wavespell cycle, a thirteen-week cycle (one-quarter of an NET day), a thirteen-Moon cycle (one NET day), a thirteen-year cycle (one NET Season), and so on. Once you know the wavespell you have come to know the form in time of the noospheric mind.

THE WAVESPELL'S TONES AND THEIR PULSARS

1.	Magnetic tone of purpose-fourth-dimension	time pulsar
2.	Lunar tone of challenge-first-dimension	life pulsar
3.	Electric tone of service-second-dimension	sense pulsar
4.	Self-existing tone of form-third-dimension	mind-form pulsar
5.	Overtone tone of radiance-fourth-dimension	time pulsar
6.	Rhythmic tone of equality-first-dimension	life pulsar
7.	Resonant tone of attunement-second-dimension	sense pulsar
8.	Galactic tone of integrity-third-dimension	mind-form pulsar
9.	Solar tone of intention-fourth-dimension	time pulsar
10.	Planetary tone of manifestation-first-dimension	life pulsar
11.	Spectral tone of liberation-second-dimension	sense pulsar
12.	Crystal tone of cooperation-third-dimension	mind-form pulsar
13.	Cosmic tone of presence-fourth-dimension	time pulsar

DREAMSPELL

Described as the fourth dimensional tool kit, the Dreamspell also encodes a psychomythic cosmology known as the Journey of Timeship Earth 2013. It is this cosmology that makes the Dreamspell a psychoactive tool kit for personally engaging you in the fourth-dimensional synchronic order of time. The fundamental premise for making it psychoactive lies in finding your galactic signature, the color, icon, and tone that code your date of birth. This combination of factors-color, tone, and icon-establishes a mathematical code that allows you to use your galactic signature in mapping increasing orders of synchronicity. There are four colors, thirteen tones,

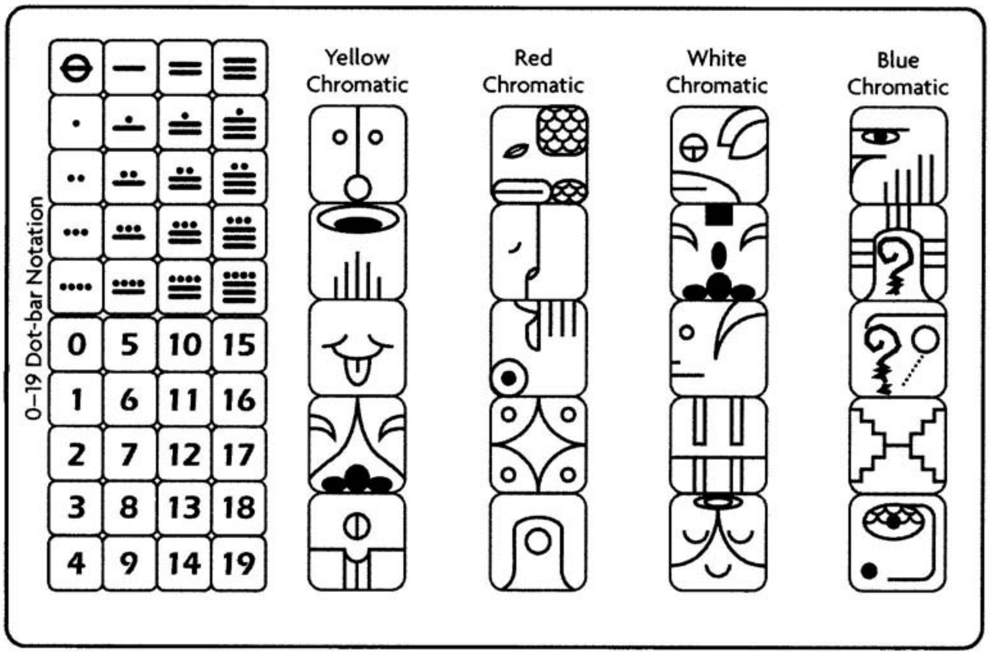
0–19 Dot-bar Notation
0 5 10 15
1 6 11 16
2 7 12 17
3 8 13 18
4 9 14 19
Yellow Chromatic
Red Chromatic
White Chromatic
Blue Chromatic

The Four Chromatics

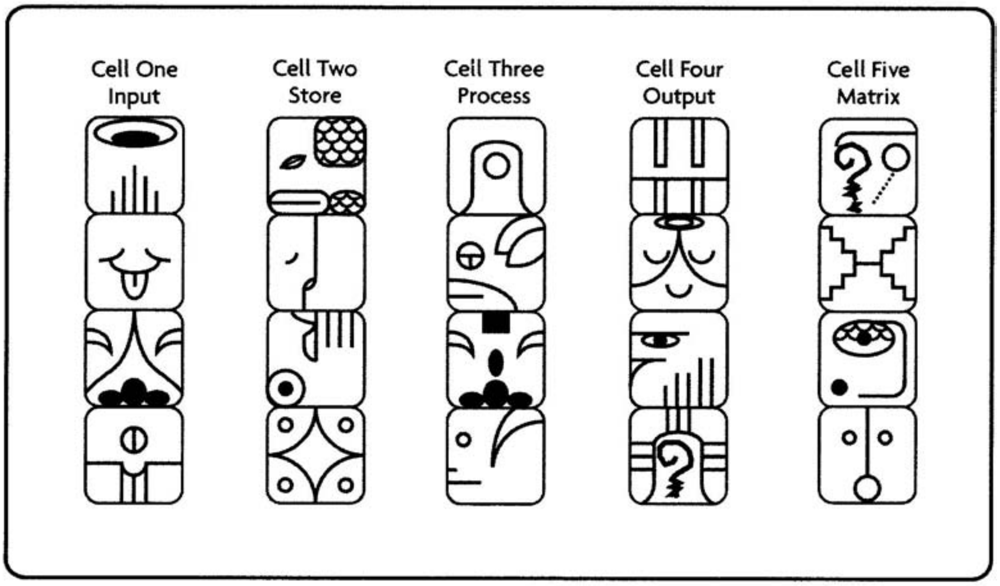
Cell One Input
Cell Two Store
Cell Three Process
Cell Four Output
Cell Five Matrix

The Five Time Cells

and twenty icons. The four colors code each of five sets of icons and correspond to the base four-phasic function of the noospheric rainbow brain-red initiates, white refines, blue transforms, and yellow ripens. The icons themselves actually correspond to the mathematical code 0-19.

Colors and Icons with Their Corresponding 0-19 Code Numbers:

Red Family: Dragon (1), Serpent (5), Moon (9), Skywalker (13), and Earth (17)

White Family: Wind (2), World-bridger (6), Dog (10), Wizard (14), and Mirror (18)

Blue Family: Night (3), Hand (7), Monkey (11), Eagle (15), and Storm (19)

Yellow Family: Seed (4), Star (8), Human (12), Warrior (16), and Sun (20 = 0)

The mapping of synchronicity through use of the galactic signature is done through a set of tools: the Galactic Compass, Journey Board, Oracle Board, and Time Atom Cube. The point of engaging the Dreamspell is to learn of the infinite possibilities of fractal synchronic mapping of the events of your life and the lives of your family and friends, as well as current events and human history. The compass and boards describe the radial mathematics of fourth-dimensional time and also provide the means for fully engaging the synchronic order of the noospheric mind. The interactive complex of changing the family and clan relations to which the galactic signature entitles you also provides the basic patterns of social organization in time. The Dreamspell is indispensable for education concerning the radial, fractal nature of fourth-dimensional time and mathematics.

TELEKTONON

Meaning "Earth Spirit Speaking Tube," *Telektonon* is the name of the prophecy of Pacal Votan, and the definition of a complex program for establishing universal telepathy. It should be noted that the tomb of Pacal Votan was discovered in 1952, after 1,260 years, because of a tile tube, the Earth Spirit Speaking Tube, which ran from the top of the pyramid temple, down the secret stairwell, and into the wall of the elaborate tomb beneath the pyramid. The prophecy itself was decoded from the inscriptions and symbols of the elaborately carved sarcophagus lid of the tomb. Essentially defining a period of "Seven Years of Prophecy," (1993-2000), followed by a 13-year Cycle of Creation (2000-2013), the Telektonon is practiced on a daily basis by means of a board and various moving parts.

The purpose of the daily play is to learn how to operate by a set of coordinates that define the nature of telepathic consciousness as the interplay of five circuits that connect the ten planetary orbits (the Asteroid Belt, remnants of the lost fifth planet Maldek, is included as one of these orbits). The key interplanetary connection is between Earth, the third planet, and Uranus, the eighth planet, which creates the third or biotelepathic circuit. The Earth-Uranus circuit constitutes twenty-eight steps and thus corresponds to the 28-day measure of the Thirteen Moon calendar. Between days 7 and 22, one is

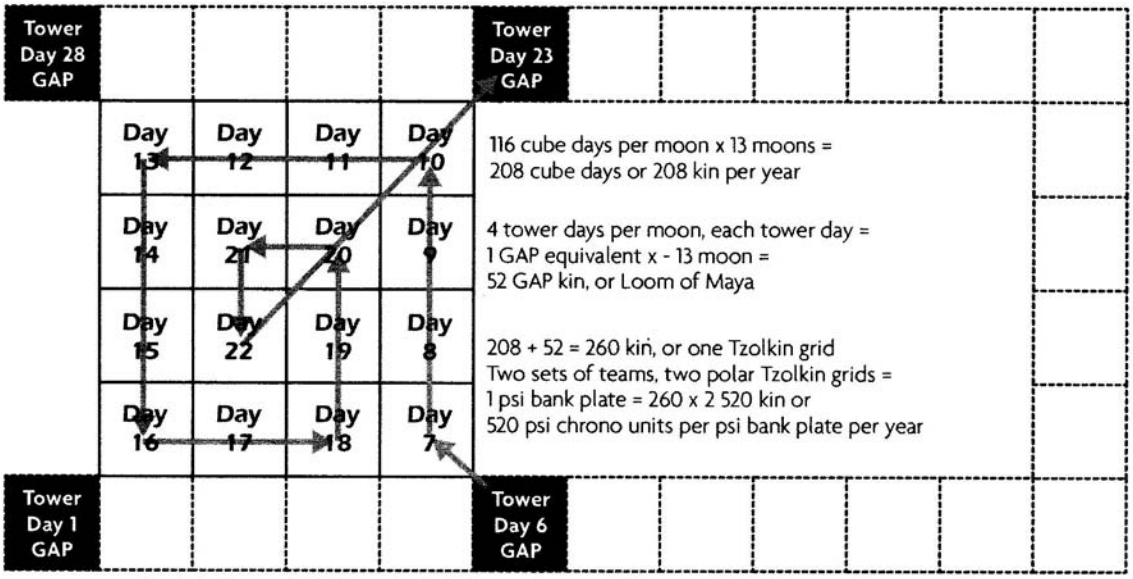
Tower
Day 28
GAP
Tower
Day 23
GAP
Day 13
Day 12
Day 11
Day 10
Day 14
Day 21
Day 20
Day 9
Day 15
Day 22
Day 19
Day 8
Day 16
Day 17
Day 18
Day 7
Tower
Day 1
GAP
Tower
Day 6
GAP

116 cube days per moon x 13 moons =
208 cube days or 208 kin per year

4 tower days per moon, each tower day =
1 GAP equivalent x - 13 moon =
52 GAP kin, or Loom of Maya

208 + 52 = 260 kin, or one Tzolkin grid
Two sets of teams, two polar Tzolkin grids =
1 psi bank plate = 260 x 2 520 kin or
520 psi chrono units per psi bank plate per year

The Telektonon Grid: Showing Psi Chrono, Loom of Maya, and Cube Journey Correspondences

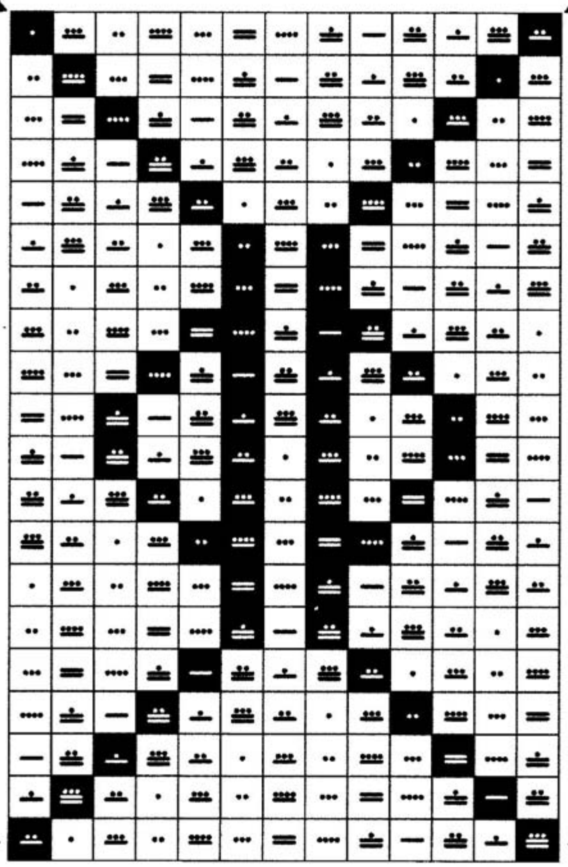
Tower Day 1 (x 13)
Tower Day 23 (x 13)
Tower Day 6 (x 13)
Tower Day 28 (x 13)

Harmonic Module
260 (13 x 20) unit
universal standard
of measure of fourth-
dimensional time

GAP = Galactic Activation Portal, 52 of which create the Loom of Maya, or binary triplet configuration of the Harmonic Module (Tzolkin).

The 52 GAPs are the basis of the occult quartet; that is, starting from each corner and moving inward there are thriteen of these sets creating a radial matrix where the tones of each set always equal 28.

13 x 28 = 364, the number of days in the Thirteen Moon calendar.

Each of these quartets codes the four tower days of the Telektonon every moon.

Telektonon - Rinri GAPs and Towers with Harmonic Module

able to take the "Warrior's Cube Journey" to uncover knowledge hidden by the Tower of Babel. The sixteen positions of the Cube of the Law (days 7-22) are the heart of the Telektonon play, which also includes a rehearsal of the thirteen baktun history with the intention of "redeeming the Babylon planet" as Timeship Earth 2013.

THE RINRI PROJECT

Described as a "four-year telepathic biosphere-noosphere transition Circumpolar Rainbow Bridge Experiment," the Rinri Project builds on the Telektonon play that establishes the biotelepathic circuit. On this circuit, two towers, days 1 and 6, define a four-day Earth Walk, and two towers, days 23 and 28, define a four-day Heaven Walk. Between the second and the third towers, days 6 and 23, are the sixteen positions of the Cube of the Law, days 7-22. The four towers every Moon correspond to one radial set of four Galactic Activation Portals in the Harmonic Module-thirteen moons x four towers = 52 Galactic Activation Portals (GAPs). Sixteen cube positions each moon x 13 Moons = 208, the remaining number of kin in the Harmonic Module once the 52 GAPs have been subtracted from the 260 units. Since the Harmonic Module is the basis of the psi bank, this formulation provides a way of coordinating the four psi bank plates of the noosphere to the Thirteen Moon calendar year on a daily and annual basis.

Since there are four plates, it takes four years, one year per plate, to open the psi bank. The first opening occurred between 1996-2000. The second opening, 2000-2004, is now occurring and is programmed by a coordination of the twenty-four runes of the Elder Futhark to the twenty-four psi membranes-6 runes per week every Moon x four weeks = 24. Since each psi plate is a mirror symmetry pair of Harmonic Modules, reflecting the North and South magnetic poles of the Earth, the Rinri project is practiced by telepathically interactive teams in the Northern and Southern Hemispheres. The Rainbow Bridge refers to making visible the dynamic of Earth's electromagnetic aura and extending it permanently from pole to pole. This is how the noosphere is visibly manifest.

Rinri is a Japanese word meaning "ethical enlightenment," and its application includes seventeen moral precepts that are studied during the sixteen-day cube journey, the seventeenth corresponding to the tower day 23.

260 POSTULATES OF THE DYNAMICS OF TIME

While the Telektonon is predicated on the fourth-dimensional tool kit, the Dreamspell, the 260 Postulates of the Dynamics of Time elaborate on the nature and theory of the biosphere-noosphere transition by providing a set of twenty systematic thirteen-part logics that describe and define the dynamics of time as the evolution of consciousness. Because there are 260 postulates corresponding to the 13:20 matrix, a postulate can be studied and meditated on each day over a 260-day cycle according to the code number of the daily icon combined with the number of its tone, for example, Red Lunar Dragon, where Dragon is 0-19 code number 1, and Lunar is tone 2, therefore the postulate to study for the day is 1.2. This daily programming, established by the Telektonon practice is how knowledge is reformatted according to the synchronic order. This in itself precipitates the noosphere.

TWENTY TABLETS OF THE LAW OF TIME, SIXTEEN-YEAR CUBE OF THE LAW

The Twenty Tablets can be described as a road map of the fourth dimension. By focusing on a sequence between the years 1997-2013, by fractal analogies, the telepathic engagement encompasses the 26,000-year cycle as well as the coming cycle of the Psychozoic future. The crux of the daily play of the synchronic order is called *cubing the codon.* The point is to telepathically reharmonize the DNA that has been adversely affected by the 12:60 timing frequency. There are sixty-four DNA codons whose six-part binary structure is identical to that of the sixty-four hexagrams of the I Ching when both are written with a simple binary language of broken or unbroken lines. Each of the codons establishes a precise thirteen-stage permutation sequence. 64 x 13 = 832 codon permutations. The sixteen years from 1997 to 2013 also constitute 832 weeks (16 x 52 = 832 = 64 x 13). Therefore, each codon has one-quarter year, thirteen weeks, to run its permutation sequence, one sequence per week. Since each codon consists of six lines, each week the codon is built one line per day for six days, each line adhering to one side of a cube. On the seventh day, the codon is cubed and sealed by one of sixty-four "UR Runes." The UR Runes are the master cosmic runes, each of which governs an entire thirteen-unit codon permutation sequence. The Twenty Tablets also contain much synchronic information for experiencing what is referred to as *fractal time compression,* the expanded mental ability to engage in different planes or levels of time simultaneously. All of this is to establish a mind training for fully engaging the fourth-dimensional synchronic order of reality and time travel.

7:7::7:7 TELEKTONON REVELATION

While the Twenty Tablets of the Cube of the Law synchronize the DNA Codons in all their permutations on a weekly basis over a sixteen-year cycle, the 7:7::7:7 Telektonon Revelation establishes the 28-day cycle as the basis for synchronizing

Three planes create the cube. The six faces of the cube code the six lines of the codon. Cube the codon according to the thirteen tones of creasion

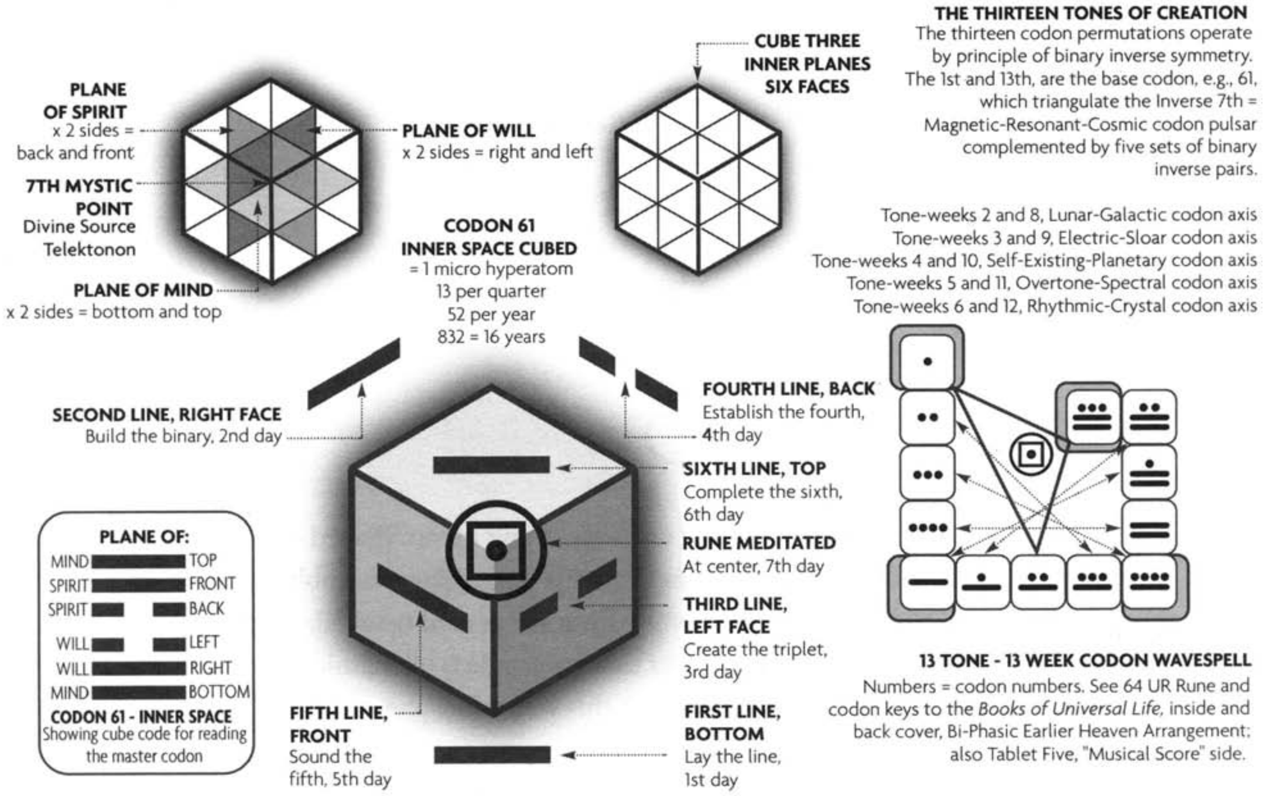
PLANE OF SPIRIT
x 2 sides = back and front
7TH MYSTIC POINT
Divine Source Telektonon
PLANE OF MIND
x 2 sides = bottom and top
PLANE OF WILL
x 2 sides = right and left
CODON 61
INNER SPACE CUBED
= 1 micro hyperatom
13 per quarter
52 per year
832 = 16 years
CUBE THREE
INNER PLANES
SIX FACES
SECOND LINE, RIGHT FACE
Build the binary, 2nd day
FOURTH LINE, BACK
Establish the fourth, 4th day
SIXTH LINE, TOP
Complete the sixth, 6th day
RUNE MEDITATED
At center, 7th day
THIRD LINE, LEFT FACE
Create the triplet, 3rd day
FIFTH LINE, FRONT
Sound the fifth, 5th day
FIRST LINE, BOTTOM
Lay the line, 1st day
PLANE OF:
MIND TOP
SPIRIT FRONT
SPIRIT BACK
WILL LEFT
WILL RIGHT
MIND BOTTOM
CODON 61 - INNER SPACE
Showing cube code for reading the master codon
13 TONE - 13 WEEK CODON WAVESPELL

THE THIRTEEN TONES OF CREATION
The thirteen codon permutations operate by principle of binary inverse symmetry. The 1st and 13th, are the base codon, e.g., 61, which triangulate the Inverse 7th = Magnetic-Resonant-Cosmic codon pulsar complemented by five sets of binary inverse pairs.

Tone-weeks 2 and 8, Lunar-Galactic codon axis
Tone-weeks 3 and 9, Electric-Sloar codon axis
Tone-weeks 4 and 10, Self-Existing-Planetary codon axis
Tone-weeks 5 and 11, Overtone-Spectral codon axis
Tone-weeks 6 and 12, Rhythmic-Crystal codon axis

Numbers = codon numbers. See 64 UR Rune and codon keys to the *Books of Universal Life*, inside and back cover, Bi-Phasic Earlier Heaven Arrangement; also Tablet Five, "Musical Score" side.

Codon Cube Cosmology
Sounding the Thirteen Tones of Creation

the seven primary plasmas-electronically charged ions-into four aggregates called time atoms. The seven plasmas, which also replace the names of the days of the week, are in two sets of three, with the middle or fourth acting as a catalytic agent. The first three plasmas form a sensory quantum and the last three plasmas form a telepathic quantum. Activation of the plasmas releases radion, a telepathic lubricant; hence, the seven plasmas are also referred to as radial plasmas. The four plasmic aggregates are known as the Analog Time Atom (week 1), the Antipode Time Atom (week 2), the Occult Time Atom (week 3), and the Unified Field Time Atom, (week 4). Like the Twenty Tablets, the 7:7::7:7 Telektonon Revelation also involves principles of fractal time compression where each day during the first week is the equivalent of 260 moons or twenty years; the following two weeks, each day is thirteen

The Seven Seals, Seven Radial Plasmas

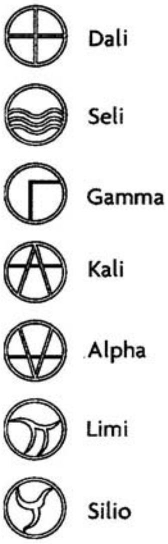

moons or one year; and the last week each day is one moon. Played on a board with cards containing affirmations from the Buddhist sage, Padmasambhava, the object of the 7:7::7:7 is to place the time atoms within the crystal octahedron at the center of the Earth so that each Moon a master time molecule is telepathically placed at the center of the Earth per one NET hour. In this way, Noospheric Earth Time is activated, the telepathic mind is engaged with the phenomenal reality represented by the plasmas, and the rainbow brain of the noosphere is hastened into everyday presence.

The seven daily plasmas constituting one thirteen-moon week, or one NET quarter hour are:

Dali (Sunday)-thermic charge

Seli (Monday)-luminic charge

Gamma (Tuesday)-luminic-thermic charge (these first three create one sensory quantum)

Kali (Wednesday)-luminic-thermic, thermic-luminic catalytic agent

Alpha (Thursday)-double-extended electron at the South Pole

Limi (Friday)-mental electron at the North Pole

Silio (Saturday)-mental electron-neutron at the center of the Earth (these last three create one telepathic quantum)

Part II. The Peace Plan: Establishing the *Pax Cultura, Pax Biospherica*

From the time of the formation, or rather, the calling into existence of the World Thirteen Moon Calendar Change Peace Movement, there had to be a *World Thirteen Moon Calendar Change Peace Plan.* The original of this Plan was submitted to the United Nations in 1995 for its fiftieth anniversary. The crux of the Peace Plan is that to change the world standard calendar is the opportunity for humanity to stop in its tracks and redirect itself. By stopping in its tracks to change its timing frequency, humanity can also use the opportunity to call a universal cease-fire and declare an emergency peace worldwide. Only in this way can the new time be experienced and fully engaged.

At the First Planetary Congress of Biospheric Rights (1996), this Peace Plan was extended to include the establishment of a new human covenant that would restore to humanity its intrinsic biospheric rights, the same as enjoyed by the rest of life in the biosphere: free land, clean air, clean water. The premise of the Biospheric Covenant is that humanity is first and foremost a member of the biosphere, and not of any artificially constituted nation-state. The Biospheric Congress also mandated the revival of the Banner of Peace and Roerich Peace Pact (1935) as one of the emblems to be used to identify the biosphere as a zone of peace.

The World Congress on the Law of Time and Judgment Day Tribunal (1997) extended the Biospheric Congress to include an indictment of the Gregorian calendar as a crime against the biosphere-"the crime of the millennium"-and the establishment of the Law of Time, inclusive of the Twenty Tablets, as the new dispensation of knowledge awarded in compensation to humanity, based on the synchronic veracity of the Thirteen Moon Calendar Change Peace Plan.

By 1999, the World Summit on Peace and Time provided the Declaration of Calendar Reform and a set of People's Resolutions, which were presented to the United Nations and the Vatican. The Resolutions define a seven-part educational program, which, like the Biospheric Covenant, envisions a radical reorganization of human society.

The purpose of these documents and declarations is for the creation of a state of universal unconditional peace so that human society can be rebuilt as a harmonic unity in phase with the biosphere as a whole. The result of the Thirteen Moon Calendar Change Peace Plan would be the creation of a new order of life on Earth-the *Pax Cultura, Pax Biospherica.*

◆

9

The Advent of the Noosphere

A Prayer and a Vision

IT IS MORNING once again, or, rather, it is predawn. It is still the first watch of the day. Outside there is a full moon. On the Gregorian calendar it is November 1, 2001, All Soul's Day, the sequel to Halloween, the day of honoring the dead. In another two weeks it will be the New Moon of Ramadan, Islamic lunar year A.H. 1422. On the Thirteen Moon synchronometer it is now the fifteenth day of the Self-existing Owl Moon, the moon of form, the moon to define, measure, and ask the question, "What form will my service take?" Well, my service will be to complete this book and prepare it for the world, for now it is the time of the advent of the noosphere, the reality of which Vernadsky said must be communicated to the world's politicians and statesmen. They didn't listen then. Will they listen now? And on the 260-kin synchronization gauge it is kin 42, White Electric Wind. The Wind "communicates spirit," the Electric tone "activates service." Wind is the second of twenty solar seals that code each day in the synchronic order.

By its mathematical relationships to the other nineteen solar seals, the solar seal of the daily kin possesses a unique five-part oracle. In today's oracle the Wind is guided by the World-bridger, who brings the power of death. In the psi bank, the matching psi chrono unit for today, the fifteenth day of the Self-existing Owl Moon, is kin 66, Magnetic World-bridger, the first tone in the 13-kin wavespell of the World-bridger, which means today is very strongly penetrated by the power of death. What is death to the noosphere? An inevitable and unavoidable event, a critical and

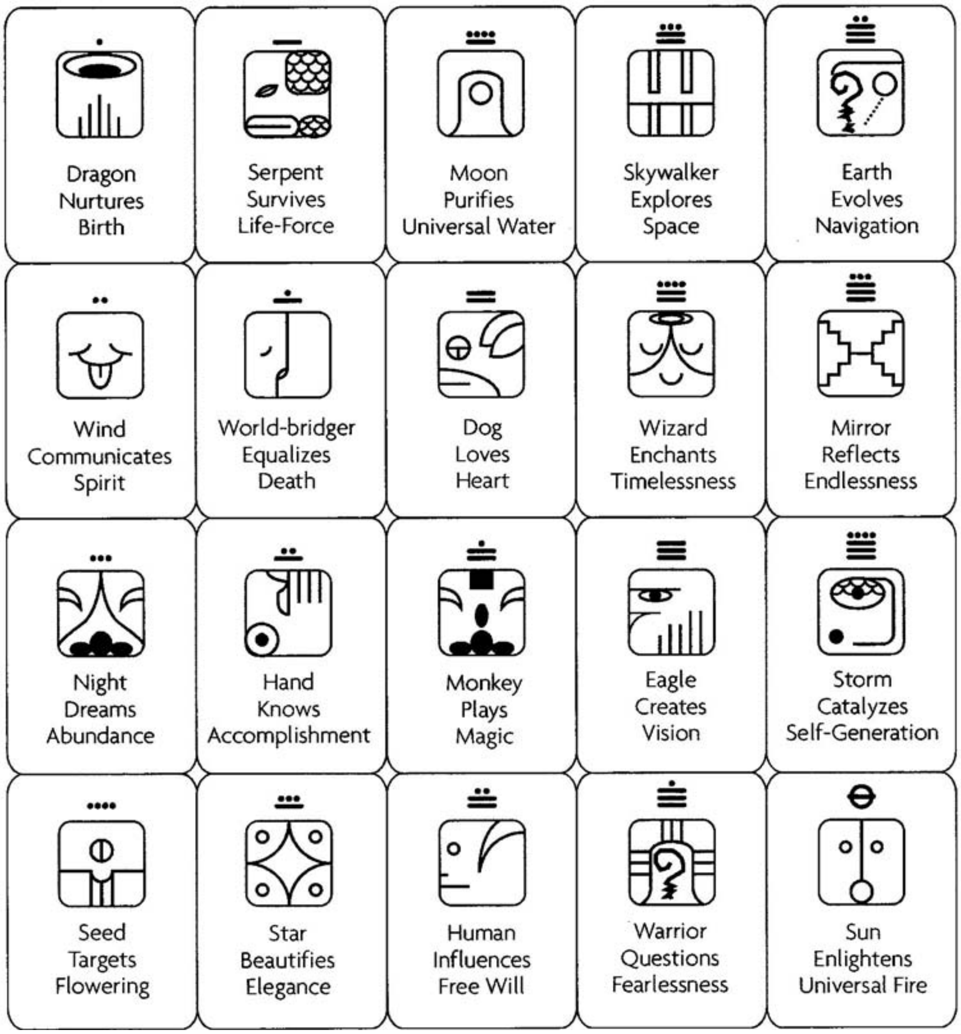
Dragon
Nurtures
Birth
Serpent
Survives
Life-Force
Moon
Purifies
Universal Water
Skywalker
Explores
Space
Earth
Evolves
Navigation
Wind
Communicates
Spirit
World-bridger
Equalizes
Death
Dog
Loves
Heart
Wizard
Enchants
Timelessness
Mirror
Reflects
Endlessness
Night
Dreams
Abundance
Hand
Knows
Accomplishment
Monkey
Plays
Magic
Eagle
Creates
Vision
Storm
Catalyzes
Self-Generation
Seed
Targets
Flowering
Star
Beautifies
Elegance
Human
Influences
Free Will
Warrior
Questions
Fearlessness
Sun
Enlightens
Universal Fire

The 20 Solar Seals
Names, Actions, and Powers

equalizing stage in the recycling of biomass, a furtherance of the biogenic migration of atoms, the creation of biospheric compost, a release of consciousness back into the noosphere, an opportunity to consider the value oflife. But to the ordinary mind of the artificial technosphere, death is the ultimate object of fear, and the fear of death is the chief object of manipulation by those who generate terror. It is of note that by his date of birth, March 10,1957, the galactic signature of Os am a Bin Laden is White Solar World-bridger, kin 126, while for this, his forty-fourth year, his birthday was coded by kin 66, White Magnetic World-bridger, both signatures bearing the destiny seal of the power of death. This is an example of how the synchronic order is "read" on a daily basis.

It is a mistake to think that the terror is just the malevolent intention of a single human being. The terror is co-extensive with the technosphere. In fact, the technosphere is the age of terror. The terror was generated at Hiroshima, and repeated at Nagasaki fifty-six years ago. And the terror was compounded by the original perpe-

Daily Oracle, Dali 15
Self-Existing Owl Moon, 9 Seed Year
(Gregorian: November 1, 2001)

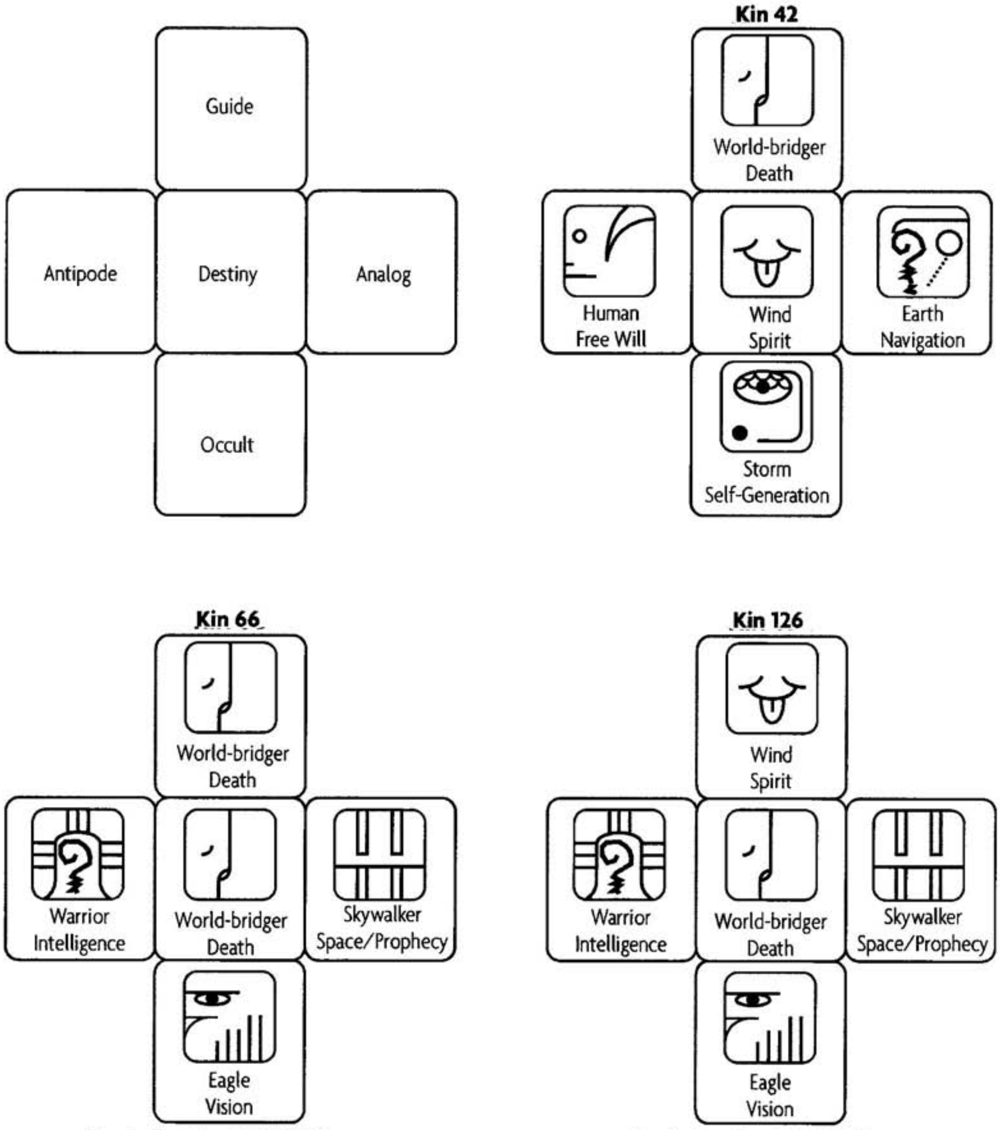
Guide
Antipode
Destiny
Analog
Occult
Kin 42
World-bridger
Death
Human
Free Will
Wind
Spirit
Earth
Navigation
Storm
Self-Generation
Kin 66
World-bridger
Death
Warrior
Intelligence
World-bridger
Death
Skywalker
Space/Prophecy
Eagle
Vision
Kin 126
Wind
Spirit
Warrior
Intelligence
World-bridger
Death
Skywalker
Space/Prophecy
Eagle
Vision

Oracle for March 10, 2001,
Osama Bin Laden's 44th birthday

Oracle for March 10, 1957,
Osama Bin Laden's date of birth

Daily Oracle Matrix

trators of that terror many thousands of times over in the creation and stockpiling of weapons of mass destruction, an act soon imitated by the then archrival, the Soviet Union, and by other nation-states, namely, the United Kingdom, France, and China. Those called the "terrorists" today have nothing to compare to the stockpiles of these weapons by these five nation-states who constitute the permanent members of the United Nations Security Council-the "nuclear club." That is who rules the world today, the official possessors of the weapons of mass destruction, who jealously guard their secrets. If this is the ultimate ruling body of the planet, and the reason why they are the ruling body is their possession of these instruments of death, then why should anyone be surprised at the existence of "terrorists" in the rest of the world? And why, we may ask again, are these terrorists mostly, it would seem, from the Islamic world? In addition to the nuclear arsenals, there are the biological and chemical weapons of mass destruction. Those who live by the manufacture of terror must surely die by their own means. If other nations have developed the same systems of terror, it is only in self-defense or in imitation of the masters of terror.

Everyday for four weeks now, I have been waking up at this time to continue work on this text, *Time and the Technosphere*I had at first thought that the book could be complete with eight chapters. People don't have time to read, and the topic warrants as brief and pithy an explanation as possible. After all, it is really a time of reflection-and then action. The masters of terror have cloaked themselves as the protectors of civilization and as the moral crusaders who will rid the world of evil. Yet what is civilization today but the resource-consuming cancer of the biosphere? Is this what is being protected-the right to promote the disease upon the biosphere? The terrorists who attacked the Twin Towers and the Pentagon clearly knew the symbolic and actual value of their targets. They also undoubtedly knew the depths of the death-fear of the superpower masters of terror, whose reaction would most likely foment a global war. But in such a war, who would win? In this regard, from the noospheric perspective, the Inevitable Event was the technocide-death by technospheric means-of the technosphere itself. The technosphere was created by and thrived on means of terror, and the means of massive terror as well as the economic imperialism perpetrated by the superpower managers of the technosphere could only have produced a reaction against itself known as terrorism. The end result is collective suicide. There are no flags in the blood of the dead, nor is there any evident religious belief. There is only human biomass ultimately indistinguishable and anonymous to the biosphere.

Eight chapters were needed in order to complete the information about the biosphere, technosphere, noosphere, and the Law of Time. Maybe a new spirit of the biosphere will arise from the considerations of my book. Maybe the noosphere will spark in others the need to allow the biosphere to reclaim us from the technosphere. Anyone who says this is "turning back the clock" is only betraying their immersion and unexamined belief in the reality of artificial mechanical time. Yes, eight chapters were necessary to unfold this analysis and solution. But I knew there had to be a ninth chapter for the testament to be complete. For this is a living testament, and if anything has been learned by living according to the synchronic order of fourth-dimensional time, it is the renewed power of number itself. There is a power and meaning to number understood as a universal, self-existing mental construct. The synchronic order is constructed of number. And as Pythagoras perceived, number is a kind of spiritual entity that can only be aroused by an act of meditative contemplation. Each whole number bears its own resonance. You can feel this in the contemplation of the difference between even and odd numbers.

The resonance of 8, an even number, is harmonic order and balance, geometrically evident in the octagon or the octahedron. But the resonance of9, an odd number, is quite different. Nine months to give birth, three times three, a rhythmic triplicity-that is nine. The Germanic god-hero Odin, or Votan, spent nine nights hanging from a tree in order to gain knowledge of the mysteries of woman. When the tomb of Pacal Votan was opened on June 15, 1952, it revealed a large vaulted chamber beneath the pyramid of the inscriptions in Palenque, Chiapas, Mexico. Entering the tomb, in the center of which is the famous sarcophagus, one sees on the walls the beautifully sculpted figures of nine personages. Today these sculptures have almost completely worn away. But originally, these low-relief sculptures modeled out oflimestone, depicted with great elegance, a series of figures commonly known as the Nine Lords of Night, or the Nine Lords of Time. Are the Nine Lords of Time, the Lords of Night who guard Pacal Votan in his tomb, the same as the nine nights of OdinIVotan? And who was Pacal Votan? With what mythic resonance are we dealing?

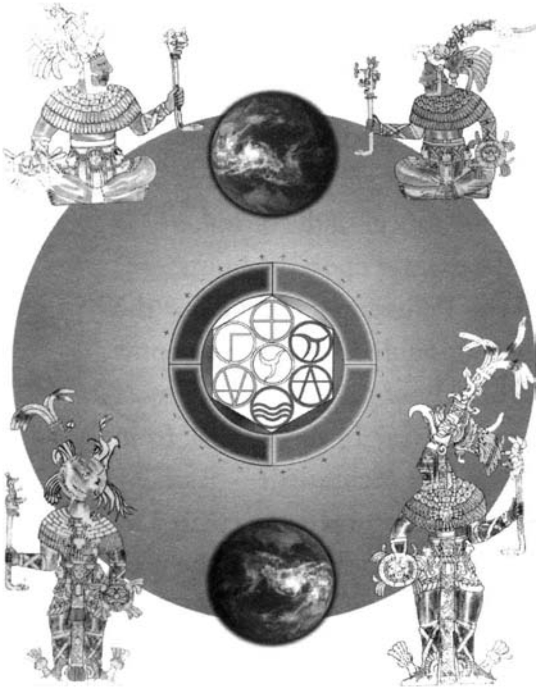

Nine Lords of Time Begin to Gather for a Council of the Earth

A Mayan seer, a sage-king of the seventh century of the Christian era, had deliberately created a mysterious tomb for himself, only to have it hidden, buried beneath a pyramid temple, but a tomb intended to be opened more than twelve centuries later, in the seventh year of the time of the terror, the time of the technosphere. How ingenious, how uncanny, that it would have been unknown for 1,260 years-a number also known in the thirteenth chapter of the Book of Revelations as the number of days of exile of the woman clothed with the sun, her feet on the moon, and twelve stars around her head. What does this allegory mean? Is it not the same as the exile of the human soul in the deadening world of 12:60 materialism? Sixty years remained from the opening of that mysterious tomb to the closing of the cycle of history, the omega point of 2012. How could the timing of the opening of the tomb have occurred so precisely-1,260 years from dedication to discovery, 1,320 years from its dedication to the closing of the cycle? How could those years be such a perfect reflection of the artificial 12:60 and the natural 13:20 timing frequencies? How could a Mayan sage in Central America so long ago be able to see, it would seem, into our future?

When we study the globe and consider things synchronically from the perspective of the noosphere, we find some interesting correspondences. Pacal Votan lived in the tenth baktun cycle, between the Christian years 603-683, on the opposite side of the world but at almost the same latitude as Mecca, and just after the time of Muhammad, 570-632. Today the religion that God commanded Muhammad to found, Islam, is at the center of the terrorist storm, while the prophecy of Pacal Votan and of the Mayan calendar in general has placed the end point of history at 2012, a mere eleven years after the Inevitable Event that placed Islam at dead center of the world mind. How could this be, the religion of the prophet Muhammad and the prophetic timing of the Maya so close at hand? Is this some revelation of the Nine Lords of Time, or of God speaking through them-and who are the Nine Lords of Time? Are they destined to return? Could it be that the Nine Lords of Time, like some higher dimensional Lords of the Ring, are a manifestation of time to be released, one by one, in an annual sequence during the last nine years of the Great Cycle, 2004-2013?

These Nine Lords of Time are known in Mayan as the Bolontiku; their counterparts are the Oxlahuntiku-the nine lords and the thirteen lords, respectively. The nine lords govern the underworld, the time of the third dimension, which is why they are depicted in descent to the sarcophagus of Pacal Votan, while the thirteen lords govern the heavenly realms. The nine govern the nine vigesimal orders of number, while the thirteen govern the form of time. The dedication of the tomb of Pacal Votan in A.D.692 corresponded to the Long Count baktun date of9.13.0.0.0.

The nine and the thirteen lords-Bolontiku and Oxlahuntiku-are present in the very numbers of the dedication date. And the nine and the thirteen also governed the prophecy of Quetzalcoatl, of Thirteen Heavens of decreasing choice and Nine Hells of increasing doom. This prophecy refers to the sequence of twenty-two 52-year cycles that began in A.D. 843 and ended in A.D. 1987. Six years later, in 1993, during the first year of the 260th and final katun cycle of history, I began to decode the prophecy ofPacal Votan, some forty-one years after the tomb had been opened.

Telektonon, the Prophecy of Pacal Votan, was meant to be decoded precisely and only at the moment of its absolute usefulness, kin 144: Yellow Magnetic Seed, July 26, 1993. This moment defined the Seven Years of Prophecy, 1993-2000. It also established the validity of the Dreamspell or WIzard's Count as the reformulated Year-Bearer Count of the Chilam Balam, the jaguar priests or night seers. That is, the Telektonon Prophecy could only have been triggered when kin 144 occurred as the Year Bearer, Yellow Magnetic Seed, kin 144, correlated to the Gregorian date of July 26, 1993.

The precise revelation of this prophecy on kin 144, July 26, 1993, was dependent on and could only have occurred because a prophetic count, the WIzard's Count, derived from the prophetic tradition of the Chilam Balam, had already been established. This "count" provided the basis of *The Mayan Factor* and was woven into the conclusion of the Prophecy of the Thirteen Heavens and Nine Hells, Harmonic Convergence, August 16-17, 1987. And the precise number of kin 144, the revelation date of the prophecy of Pacal Votan was meant to accord with the Book of Revelations number of the elect, 144,000 (Revelations 7:4). In fact, that very day I discovered that the number 144 transliterated from the vigesimal count into Arabic numbers is written 7.4, which is the same as the chapter and verse number where the 144,000 are first mentioned in the Book of Revelations (7 = 140, or 7 x 20, plus 4 units = 7.4). More uncanny stuff-but not so to the noosphere, the Earth's mind of time. In the noosphere everything is woven of number, and number is intrinsically synchronizing, which is why numerology has held such a fascination for the human mind throughout history.

When I first began to decode this "final prophecy of the Maya," or Telektonon, a prophecy that commands the return to natural time through the Thirteen *Moon!* 28-Day calendar, I was stunned and not a little frightened. I had already taken many steps to remove myself from mainstream society and conventional life. But now I was dealing with a genuine prophecy, and I was to be its messenger. This was not my invention. The world is filled with enough madmen and lunatics spouting their particular visions. How could I ever make the world understand that this vision, this prophecy, was real, that it was the fulfillment of a Mayan who lived and died thirteen

Vinal position, one dot = 20 kin, 7 = 140 (7 x 20)

•••• Kin position, one dot = 1 kin, 4 = 4

Written 7.4 = 144 (140 + 4)

7.4 = 144

centuries ago, and that his tomb was the vehicle of this prophecy, a prophecy meant precisely for this moment in time-the biosphere-noosphere transition?

I returned to the texts of the Chilam Balam, whose prophetic stream includes day prophecies, year prophecies, katun (twenty year), and *ahau* (260-year) cycle prophecies, prophecies of Quetzalcoatl, and finally the special prophecies of the mysterious Antonio Martinez and the coming religion of the Hunab Ku, "One Giver of Movement and Measure." I realized that the meaning and fulfillment of this entire prophetic stream was crystallized in the Harmonic Convergence and its 26-year aftermath. Chilam Balam, the jaguar priests, are the wizards of time, inspired by the guidance of the Bolontiku, the Nine Lords of Time, the Lords of Number. By command of the Supreme One (Hunab Ku), the wizards are directed to establish clear paths of rekindled knowledge. The ancient Maya prophesied their own destruction. Therefore, within the prophecies of the Chilam Balam they laid the seeds of their reconstruction-but not simply as Maya. The reconstruction had to be planetary, for their destruction in the sixteenth century was also an act of planetary consequence-the plunging of the entire world into the prison of artificial, mechanized time. The fulfillment of the prophecies along with the new dispensation of time had to occur before the end of the cycle, 2012.

For this reason, the master synchronic architect of the Mayan prophets, Pacal Votan, prepared the way with the discovery of his tomb in 1952. This phenomenal act of prophetic release meant that now the messenger of his prophecy of time had to be summoned to his task. This was to be my role and responsibility-one that I accepted, for I saw no other choice. The course of my life from the moment I had experienced an illumination atop the Pyramid of the Sun at Teotihuacan in 1953, to that moment forty years later, in 1993, when I first perceived that the meaning of the prophecy Telektonon was a single trajectory. The noosphere had found another chip. After all, if the noosphere is the mind of time on Earth, prophecy is just another way of radializing time, of bringing it around from one point in a cyclic motion to another. My preparation had been going on for a while.

Once I had received knowledge of the prophecy of the Thirteen Heavens and Nine Hells from Tony Shearer back around 1970, I prepared myself by study and meditation for the moment of the Harmonic Convergence. It may be asked: If the Harmonic Convergence concluded the cycle of the Nine Hells, then why are we experiencing the apocalyptic events of Hell on Earth now, some fourteen years later? What was the actual meaning of the Harmonic Convergence, and what does it mean for us today?

First of all, the Harmonic Convergence refers to the entire 26-year cycle from White Galactic WIzard to Yellow Galactic Seed, 1987-2013. The conclusion of the Ninth Hell refers to the release of humanity into the 26-year time zone of free will, choice, and decision in preparation for the completion of the cycle, omega point 2012, and the Galactic Synchronization, 2013. The full and true Harmonic Convergence will not be complete until the Earth itself becomes the Galactic Seed, 2013. If the twenty-six years are the free will zone of choice and decision, what are we to choose and decide? The choice is, as I was quoted in the *TVtlllStreet Journal* in 1987, "between a 'new age' and all-out destruction." The decision is to accept the knowledge of the error in time (12:60) and the choice is to return to natural time (13:20) by means of the Thirteen Moon!28-Day calendar. This the prophecy ofPacal Votan fully affirms.

In principle, the Harmonic Convergence refers to the converging of all aspects of reality in a great, all-unifying harmony. The initial moment of the Harmonic Convergence that witnessed hundreds of thousands of humans congregated at sacred sites worldwide at the dawn of August 16, 1987, Blue Electric Eagle, was a pure prophetic enactment of humanity being harmonically converged with the natural order of time. This visionary moment demonstrated that, indeed, Harmonic Convergence was possible. The wave of unconscious energy released through the collective psyche of the species as a result of this moment of prophetic purity reached a climax with the fall of the Berlin Wall and the end of the Cold War.

Early 1990, the role of the Harmonic Convergence in this process was echoed in the mass media as the question, "Did the Harmonic Convergence bring the world closer to peace?"l Before the question could be answered, the next phase of the prophetic end time began to unfold. Rather than determine how to make the choice to enter a genuine new age, the human race defaulted to an absolute materialism. It was this collective turn taken by the planetary leadership, now known as the G-7, that engendered the apocalypse we are now witnessing. In this turn, the Harmonic Convergence was forgotten, deleted as it were from the official public texts of contemporary history. The stage was thus set for the next revelation, the Telektonon of Pacal Votan.

In distinguishing the Seven Years of Prophecy, 1993-2000, Magnetic Seed-Resonant WIzard years, the Telektonon of Pacal Votan makes an absolutely clear analysis of the human condition at this time: unless the human race forsakes the false time governed by the instruments of the Gregorian calendar and the mechanical clock by the year 2000, then it will face its own destruction, biospheric collapse, and Hell on Earth. But ifby the year 2000 it can make the decision to reject the calendar of the false

12:60 time and adopt the perfect instrument of the natural 13:20 time, the Thirteen Moon!28-Day calendar, then humanity will have been able to make the choice of entering a new age. For the new age, so longed for, is actually only realizable as a new time. A new time can only come about by the rejection of the instrument that holds in place the hallucination of the old time, replacing it instead with an instrument of such perfect harmony that it has no history, but is truly post-historical.

For this, the messengers of time, myself and my wife, embarked on a course of being galactic gypsies, with no visible means of support, in order to warn the world of the meaning of the time and to offer to the world the opportunity of changing its course by changing its calendar. When the year 2000 rolled around, humanity remained unmoved. The Gregorian calendar, the Vatican, and the G-7 appeared to be as strong as ever. The prophecy did not fail, but humanity did fail to heed the message regarding artificial and natural time. In this process, as "messengers" my wife and I were only playing out an archetypal role. So beginning with the Blue Galactic Storm year, July 26,2000, thirteen years after the White Galactic WIzard year, humanity was placed in judgment. This is merely a matter of natural karmic law.

The four-year cycle, 2000-2004, became known as the prophetic cycle of the Harrowing of Hell. Just as Jesus was crucified and then experienced the Harrowing of Hell before his resurrection, so too humanity has been crucified on the technospheric cross of false time, and is now in the Harrowing of Hell awaiting the resurrection in natural time. For having failed to take seriously all its messengers and all its prophets of all times, humanity was now to pass en masse through hell. Since the Galactic Storm year, 2000-2001, completed a Seed-Storm Year bundle begun in the fifth year of prophecy (1997-1998), the 2000-2001 Blue Galactic Storm year was the year of evaluation. The full judgment was withheld until the Solar Seed Year, 2001-2002.

The Solar Seed year is the second year of the Harrowing of Hell, but the first year of the next four-year Seed-Storm Year bundle. The first half of the Yellow Solar Seed occurs during the second half of the first official year of the third millennium, 2001. The moment was now ripe for the Inevitable Event, for it was also the same Gregorian calendar year as 1945 and 1973. In the Telektonon Prophecy, verse 120, section 19, it is declared, "When the Lunar Moon has overflowed its banks, the G-7 will be no more." So it was on the twentieth day of the Lunar Moon (Blue Self-existing Monkey, galactic signature of Karl Marx), that the Tower of Babel, the World Trade Center, was struck a fatal blow, while the Pentagon, magical structure of the priests of the Dark Lords of Matter, built in "Hell's Bottom," Arlington, Virginia, during the Ninth Hell cycle, was mortally penetrated.

It is still the time of the Harmonic Convergence. The Harrowing of Hell represents the last cycle of respite for humanity. By July 26,2004, Blue Crystal Storm, humanity must make the "choice between a 'new age' and all-out destruction" when it will have the last opportunity to make the decision of rejecting the old time and calendar and replacing it altogether with the instrument of perfect harmony, the Thirteen Moon!28-Day synchronometer. Only such a harmonic organizing principle, universally accepted and applied, can ensure a time of harmony and the fulfillment of the Harmonic Convergence.

Is it possible that within the noosphere the Harmonic Convergence is awaiting its next climax on the Day Out of Time and the Thirteen Moon Synchronization date of the year 2004, that is, on Gregorian July 25 and 26, 2004? One reads through the *Journal of Calendar Reform* and the many texts and arguments for calendar reform that were put forth in the 1930s and feels the missed opportunity. Even Gandhi was in favor of a new universal timing standard as a principle means of unifying humanity in peace. Did World War II occur because the League of Nations failed to enact the calendar reform on January 1, 1933, as it had fully proposed? The League of Nations had promoted calendar reform as one of its major agendas since the 1920s. This world organization failed at calendar reform and at preventing World War II. Who or what prevented that reform from happening, and why?

Let us look at the matter again. In order to remain constant, any perpetual calendar ordered by the seven-day week requires the principle of a null day, a Day Out of Time that is no day of the week at all. This was a feature of the preferred Thirteen Moon!28-Day calendar proposed by the International Chamber of Commerce, as well as the other modified calendars of twelve months. Reading through archives and literature, pro and con, it is clear that what defeated the calendar reform was the conservative, Vatican-inspired objection to the null day, the Day Out of Time. To observe such a day, which is no day of the week at all, it was argued, would break the weekly succession of days set in motion by God at the beginning of creation-such a break would plunge the world into barbarism, chaos, and war. "The adoption of the weekly cycle was a triumph of Christianity to the detriment of paganism. The cult of Christianity is inextricably connected to the week whose origin reaches to the dawn of history. "2 Therefore, it was contended, to interrupt this weekly succession would destroy Christianity itself. As a result of such propaganda, the League of Nations faltered, no calendar reform occurred at all, and the world was plunged into war-a war that ended with the beginning of the age of terror. This terror is now running its course, like a fever sweeping through the biosphere. WIll it be possible to consider again what the League of Nations faltered on and what the United Nations simply chose to ignore?

Who really lost out when the calendar reform issue was tabled in 1956 at the United Nations? It was not God who invented the week as we know it, but more likely the Babylonians. Behind the week is the mystery of the whole number seven, recalled in the Epic of Gilgamesh as the seven wise men of Uruk. But the Babylonians clothed the seven-day cycle in the raiment of the sun, moon, and five principle planets-Mercury, Venus, Mars, Jupiter, and Saturn-to which the days of the week have no actual relation. Seen in this way, the seven-day week was intended to be a means on Earth for reenacting some remote mythic reality related to these seven heavenly bodies. What was that mythic reality, on what hidden knowledge is the count of the seven-day week presumed? Borrowing from Babylon, the ancient Hebrews incorporated the seven-day week into their purely lunar calendar. And it was thus from the Hebrews that the seven-day week entered the Julian and then the Gregorian calendars. It was not, however, the Romans who adopted the week, but the Christian successor to the Roman empire, the Church of Rome. What truly was the origin of the week? And what relation or meaning does it have in the first place? Even to F. H. Colson, a Christian scholar writing as late as 1926, the seven-day cycle is described as "that intruder the week, consisting indeed of a fixed number of days, but paying no regard to months or years."3

With its fixed number of days, the week makes sense only in a system of time reckoning that is synchronic rather than astronomical in nature. To those whose minds are set on the synodic lunation cycle as the absolute standard, as does the Gregorian calendar, the week is what ensures the greatest disorder in the calendar-and yet it is defended to the death. The week gains its power from the whole number 7, and not from the jumble of names in which the week is clothed. The seven-day cycle is the closest whole number that corresponds to the four phases of the moon, which, as whole numbers, are also seven days each. In fact, the fixed 7-day cycle only makes sense within the context of the Thirteen Moon calendar with its 28-day cycles divisible by four and seven, creating the annual cycle of fifty-two perfect weeks-plus the Day Out of Time, ensuring a perpetual harmony.

Was the Babylonian naming of these days-with only one name a reference to the moon-meant to be a cover-up of some earlier knowledge, a knowledge both more purely lunar and harmonic, such as the Thirteen Moon calendar? Is all of history, then, just based on a cover-up of harmonic time? What would really happen if there were a Day Out of Time, no day of the week or month at all? Would old belief systems dissolve? Would there be a psychic release, a sense of freedom? Since 1993, sixty years after the failure of the League of Nations to enact calendar reform, the Thirteen Moon Calendar Change Peace Movement has made certain that the

Day Out of Time is the day for forgiveness of debts, of release into timelessness, and for experiencing the full meaning of time is art. This day is now annually celebrated worldwide by increasing numbers of humans. All the while, the Gregorian world has arrogantly plunged blindly forward, locked into its mechanistically chaotic timing sensibility, only to receive the mortal wound of the Inevitable Event, causing it to fall through the seams of its own making, the technosphere. A Day Out of Time, it seems, might be a relief to a world now consumed by chaos, barbarism, and war.

> He sends down from the sky water for your drink and to grow trees for your benefit. With it He grows for you crops, olives, date palms, grapes, and all kinds of fruits. This is sufficient proof for people who think. And He commits, in your service, the night and the day, as well as the sun and the moon. Also, the stars are committed by His command. These are sufficient proofs for people who understand. (Quran, 16:10-12)

Yes, the biosphere is a set of proofs for people who understand, an immense order that reflects and transforms the cosmos through the very process of life on Earth. Who are the people who understand, and where are those people today? Who will speak for the biosphere? Who will take the message of the advent of the noosphere to the statesmen and the politicians that they may take heed and note that what is occurring is not really in their control at all, but is the climax of a phenomenal process known as biogeochemical combustion? And who will declare that this climax will inevitably shake all the towers to their foundations and level the mountains of industry to a flat plain from which the noosphere will manifest as the equality of telepathy for all?

This book, and this ninth chapter in particular, is like a long meditation, a vision, inspired by the dark tragedy of the Inevitable Event. As a vision, these thoughts are both those of one solitary individual, albeit a noospheric chip, and the continuation of the vision of that ancient seer, Pacal Votan. Could it be that Pacal Votan in his time knowledge rehearsed all these words, these thoughts, this entire vision itself? On the sarcophagus lid of Pacal Votan is a cross, not the cross of death as in Christianity, but the cross of life, a tree, a cosmic tree. Across that tree is a serpentine band with two dragonlike heads, one at each end. This band represents the galaxy. The head at each end represents the galactic vision that sees equally into the past and the future. Sometimes this is called the *zuvuya,* the reflex of cosmic memory. The entire galaxy, the Milky Way, is this reflex of cosmic memory. What if all of humanity is meant to be opened into this cosmic memory? This is what is meant by the advent of the noosphere, the mind of time on Earth, the mental envelope of planetary consciousness arising from the grave crisis of the biosphere through which we are now passing at this very moment.

Because we are now defining the noosphere does not mean that the noosphere was not known to those ancient Mayan time scientists like Pacal Votan, especially if the noosphere is defined as the mind of time on Earth. To such time scientists a NET year would be no time at all, or rather, it would be but a single, fifty-two solar orbit rehearsal of life. How many such rehearsals would be necessary before something was finally learned? The Prophecy of Thirteen Heavens and Nine Hells, for instance, was actually just the passage of twenty-two NET years. Is this how long was needed to see if the human race would learn anything? And if Pacal Votan and Muhammad were around today, what would they think and do?

Knowing the nature of the synchronic order of time, someone like Pacal Votan, obviously one of the most singular of all personalities of the Mayan world, would have possessed a consciousness like the two-headed galactic band, the consciousness of galactic memory. In the zuvuya of galactic consciousness, peering into the deep past is peering at the present from the future. This is possible in the radial time of the fourth dimension freed from the strict linear construct of past-and-future-time. From some higher state of mind, from a mind and consciousness identified completely with the galaxy, this has all been foreseen, this has all been written in a book. Now we who are in this present state of end-time consciousness are being asked to self-reflect. We are being asked if we can manage to hold the noospheric lens steady enough to see that the technosphere had to end, and that the technosphere could only end by the means by which it maintained all of the biosphere in its control, by terror itself.

Yes, we are all a bit like the frog at the bottom of the well whose vision and perception of the universe is circumscribed by what the opening of the well allows it to see. The well is the technosphere, both as a planetary apparatus and as a belief system. On the other side of that well is a vast world that we have actually never seen. This vast world is the noosphere and beyond. And what is beyond? Many of the traditional peoples say that we originally came from the stars. If we came from the stars, are the star people still out there? And what do the star people think of all that is happening on Earth today? And what do they think of their lost star children, many of whom now believe it is childish to think that we came from the stars? And if there are star people still out there, wouldn't they be too sophisticated to think they could get anywhere in a rocket ship, or even a UFO, subject to the en tropic degradation of the material plane?

No, they wouldn't rely on such primitive technology, but would be in tune, as is the rest of the universe, with the universal frequency of synchronization and the capacity of time to travel at an infinitely instantaneous rate of transmission. This being so, we could receive telepathic messages from the star people-if we were ready for it and knew precisely why, when, and how such communication occurs. In fact, as we have said earlier, the noosphere is the region on Earth for the receipt and transmission of messages of cosmic consciousness. This being so, the future of human evolution, of *Homo noosphericus,* would be nothing less than to become a medium of cosmic consciousness, adapting to the expanded measure of Noospheric Earth Time, traveling back to the stars but through the superior knowledge of the actual laws of time.

In writing this book, I have only a sense of duty, of responsibility to expand the vision of my fellow earthlings from the technosphere to the noosphere. In the Law of Time there are no arbitrary distinctions between reason and art, science and prophecy. The universe really is a giant radio transmitter. This book is as much an exposition of matters generally not thought about, as it is a personal testament and vision. In that regard this text is a portion of a broadcast transduced and received by me functioning as a noospheric chip, on behalf of the biosphere-noosphere transition.

In the tomb lid ofPacal Votan are impressed all the frequencies of this broadcast transmission called *Time and the Technosphere.* I know that Pacal Votan took an earthly form for just this purpose. He left his body in that tomb, a jade mask being the actual impression of his face. According to Chinese and ancient Mexican tradition, jade is the stone of immortality. Was Pacal Votan an immortal? Where did his consciousness and spirit go once his body was placed in the sarcophagus? Why did he leave the Telektonon, the Earth Spirit Speaking Tube, running from the tomb to the temple floor at the top of the Pyramid of the Inscriptions? Knowing the frequency of the synchronic order governing the psi bank of the mental envelope of the Earth, did he place his thoughts in the noosphere to be released at the precise moments of the opening of his tomb and the decoding of his prophecy? And then where did he go?

Beginning in 1980 I became conscious that, from time to time, I had been receiving telepathic transmissions. I soon identified these transmissions as coming from the star Arcturus. In *Earth Ascending,* originally written in 1983, the nature and program of the psi bank, the control panel and regulating mechanism of the noosphere, became known to me. This was followed by *The Mayan Factor,* and then the decoding of the Law of Time. At the moment when I had just finished writing *The Mayan Factor* in 1986, I received the inspiration for what is called the planet holon, the twenty-part structure accommodating the program of the twenty solar seals. When I received this image it was known to me as the program of the Arcturus Protectorate. As I understood it, the Arcturus Protectorate was established by the "star people" as a kind of protective time shield around the Earth. By means of the activation of this time shield, the star people would know that some of the earthlings were waking up to time and beginning the conscious activation of the noosphere. This is exactly what came about through the *Dreamspell,* beginning in 1990.

The *Dreamspell* is such a radical break from the past, and such a pure statement of the history-less fourth-dimensional future, that in order to comprehend it, in 1992 I wrote two other books, one called the *Arcturus Probe: Tales and Reports of an Ongoing Investigation* and the other *The Call of Pacal Votan: Time is the Fourth Dimension.* Some people will say that the *Arcturus Probe* is an exaggerated fantasy, yet I cannot deny what a purely telepathic process it was to write that book, which fills in many points of the psychomythic cosmology of *Dreamspell: The Journey of Timeship Earth 2013. The Call of Pacal Votan* was originally titled *A Treatise on Time Viewed from its Own Dimension,* and is a purely scientific description of the mathematical codes of the Dreamspell in relation to the biosphere-noosphere transition. The one text purely imaginal, the other purely rational, yet both were two sides of the same coin, a rendering of the Law of Time in its cosmological and mathematical comprehensiveness. Then came the even more abrupt break with my conventional past, the Telektonon Prophecy of Pacal Votan.

I bring up all of these personal details because people often ask where it all came from and where I got my ideas. But more importantly, I bring them up because, as I have been indicating, all of this was foreseen and known by a seer in another time, in another world-Pacal Votan. And this process of one person hiding a prophecy and another person in another time finding that prophecy, this is also a further revelation of how the noosphere functions. The French mathematician and philosopher Charles Henry (1859-1926) declared, "As the individual becomes more collective, the collective becomes more individual."4 As individuals, we humans are meant to grow into the collective mind known as the noosphere, while the noosphere, which is the collective, could only be known through its incorporating itself through the individual. I am one such individual, a noospheric chip. The noosphere could not be understood if it did not take on a personality or a number of personalities. Otherwise the noosphere would remain in the purely theoretical state or condition.

Now that the Inevitable Event has occurred, the noosphere is recruiting. The program of recruitment is called the Campaign for the New Time. This is a genuine campaign, but a peace campaign and not a military one. As a campaign it must reach its climax by mid-2004, a full eight years before the end of the cycle, northern winter/southern summer solstice 2012, the noosphere's omega point. The reason for this is that any later than 2004, the prophecy will be negatively fulfilled. A minimum of two Seed-Storm Year bundles-eight years-are needed to creatively destructure and then to reconstruct. The Inevitable Event was a sign of God showing us that the end had already occurred. Now all the forces of goodwill and harmony must be channeled into the conscious act that will coincide with the noosphere's readiness to manifest through all of us, and that is the calendar change-the untried solution, the most important unfinished business of the last millennium.

This is no longer a matter of debate. There is no other solution that hasn't been tried. There is no other way that the noosphere can become collectively conscious than through the human adoption of a timing system of perfect harmony. By meeting our deadline with the noosphere at the time of the Great Calendar Change in the middle of the year 2004-0n the White Spectral Mirror, Day Out of Time, the day of release from the old, and Blue Crystal Storm, the day of entering the new-then the biosphere will harmonically converge with the noosphere, and the *Pax Cultura*, *Pax Biospherica* may begin in all seriousness. Such is the advent of the noosphere.

What did the Maya foresee for 2012? What is Pacal Votan's vision? Pacal Votan's vision is the fulfillment of a prayer-a prayer that I have heard through the silence:

> o Hunab Ku, One Giver of Movement and Measure, be kind to the children of this Earth. We see how every year you give and how every year you take away. This is so the children of Earth can learn to be free and dwell only in your presence. Yet we know that the children of the Earth must undergo their lessons as well. And we know that they must enter the dark time of testing. 0 Hunab Ku, in their time of darkness, in their time of forgetting, do not forget them, spare those who are good, if you can. Prepare for them the Second Creation you have promised to the righteous among them. Break for them the spell of false time before they destroy this Earth! And show to them the splendors of the Second Creation, of the New Heaven and the New Earth. Show them these wonders before the Great Cycle is finished, so that they may yet wake up from the sleep of time. And 0 Hunab Ku, may the nine great Lords of Time return to Earth! May they teach the children of the Earth how to become a race of wizards, tilling the noosphere as they have tilled the Earth. 0 bring to these children of the Earth a good Time, a Time of Peace as long, at least, as was their time of history and war, and then let that Time of Peace be multiplied sevenfold!

And here is the vision that fulfills the prayer:

> Planet Earth, Rhythmic Solstice, Blue Resonant Storm year, 2012 Omega Point. The Earth's axis is tilted so that the North Pole is at its farthest from the sun, and the South Pole is at its closest. High above the Earth an iridescent, thin rainbow band

arches from one pole to the other. A second band does the same, but over the opposite side of the Earth. The Earth slowly rotates beneath these two rainbow bands that remain unchanging, constant in their positions, one of them directly opposite the sun, the other holding a position on the midnight side of the Earth's turning. The band facing the sun is the Earth's day alternator. The other band is the Earth's night alternator. If we glide across the surface of the Earth we see that everywhere there are small encampments, circular and radial in formation. Yurts, tents, adobe, and earthen structures-and gardens. The once-teeming cities are eerily silent, yet still inhabited, though with far fewer people than during the time of the technosphere. And everywhere in the small encampments people are reverentially gathered. This is the moment of the closing of the cycle. It is as if everyone is drawing in a single breath at the same time, and exhaling it at the same time. From wherever you are, even though there are clouds, you can see the shimmering iridescent rainbow alternator, either by day or by night. Clusters of people are gathered in circles of deep meditation. Around them the children laugh and play. "We are the noosphere," a voice is repeating, "we are the noosphere." This sets up a vibration and people everywhere stop in their tracks, they look toward the sky. They listen. "The galaxy is renewing our thought. The galaxy is renewing our mind. The galaxy is renewing our world, this Earth. Let us go home and learn some more." Drums and chanting fade into the twilight in one place and beckon the dawn in another. "Let us say the prayer again: '0 Hunab Ku, Allah, One Maker, Our Creator, we submit to You. Continue to keep us whole, continue to broadcast the noosphere through us, continue to let us listen to the higher collective voice that we are so that tomorrow we may register another kin, another NET minute in the glorious journey of Timeship Earth toward its goal of universal unification in your One Unending Thought, 0 Hunab Ku, do not forget us, now, and do not let us forget you! Thank you for the closing of the cycle! May what we have learned stay with us and become the inheritance of our children for seven times seven generations to come, and may you bring us another cycle for all our descendants to live in continuing peace and harmony!'"

This is the vision. I am a noospheric chip. Do not turn away from your visionaries. Your prophets are still calling you and teaching you, if you would pay attention. There is much to do and learn. There is nothing more joyful than the positive construction of harmony.

Time and the Technosphere, First Draft Transmission completed,
Self-existing Owl Moon Dali 15,
Yellow Solar Seed Year
Kin 42, White Electric WInd:

I activate in order to communicate
bonding breath
I seal the input of spirit
with the electric tone of service
I am ptided by the power of death

First draft edited, completed, and sealed,
Overtone Peacock Moon Seli 2,
Yellow Solar Seed Year
Kin 57, Red Overtone Earth

I empower in order to evolve
commanding synchronicity
I seal the matrix of navigation
with the overtone tone of radiance
I am guided by the power of life force

Second draft edited, completed, and sealed,
Rhythmic Lizard Moon, Seli 23, Navigation Tower
Yellow Solar Seed Year
Kin 106, White Lunar World-bridger
Pacal Votan Clear Sign

I polarize in order to equalize
stabilizing opportunity
I seal the store of death
with the lunar tone of challenge
I am guided by the power of endlessness
I am a galactic activation portal
enter me

May all beings be well and happy!

May this text and these words inspire by their truth.

If! have left out or omitted any logic or fact, or unnecessarily maligned any truth or being by my endeavor to convince through dear argument, may the Creator know that no such malice or oversight was ever intended. I am but a human transmitter, a noospheric chip. My ability is commensurate with the frailties attendant to one too long steadfast on a path too little comprehended by many of those around him.

May the truth prevail!

Jose Argiielles, Ph.D.
known in prophecy's light as
Valum Votan, the Closer of the Cycle
devotee of Bolon Ik

Brightwood, Oregon
Cascadia Bioregion
North American Plate

◆

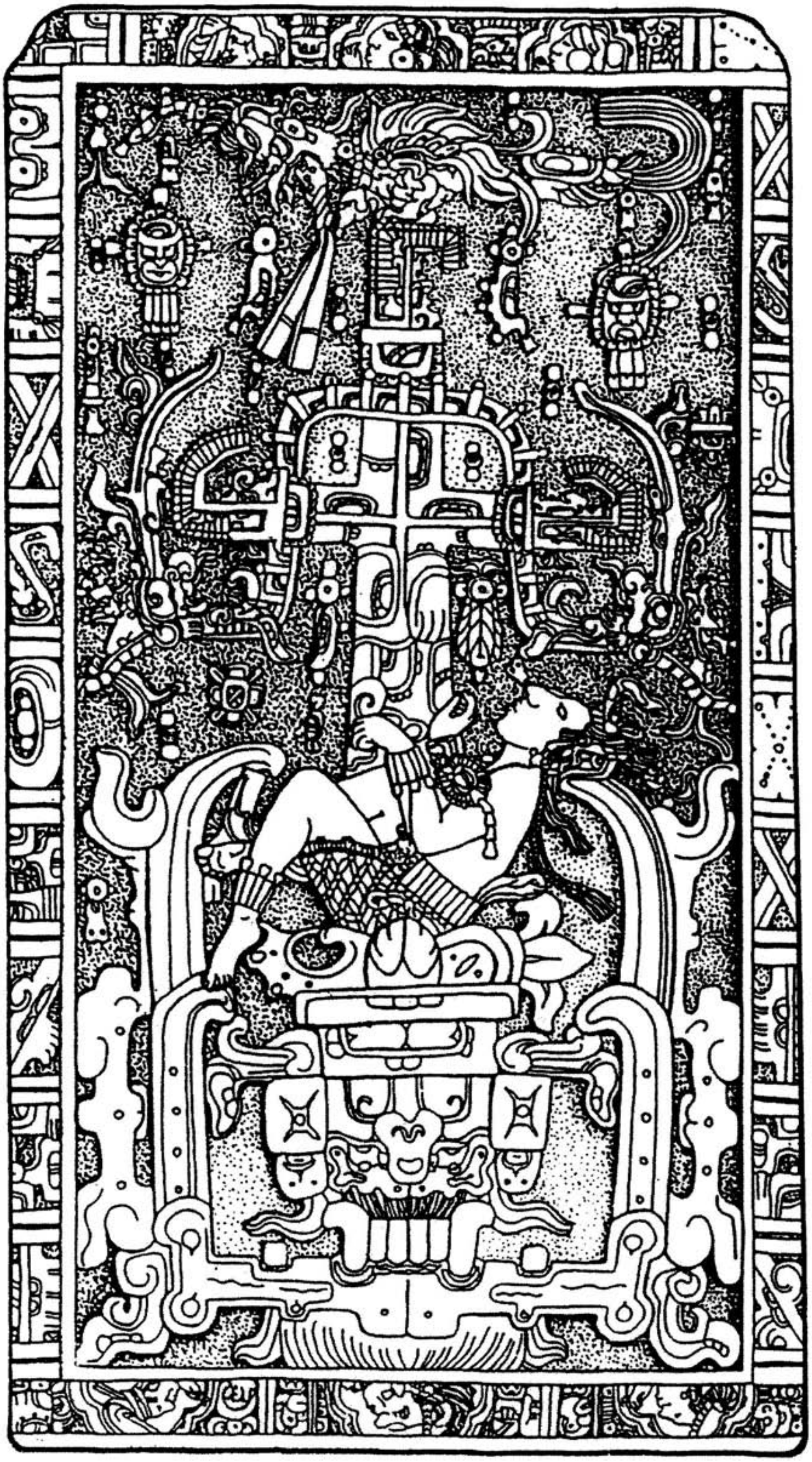

Sarcophagus lid of the tomb of Pacal Votan 9.13.0.0.0 (A.D. 692)

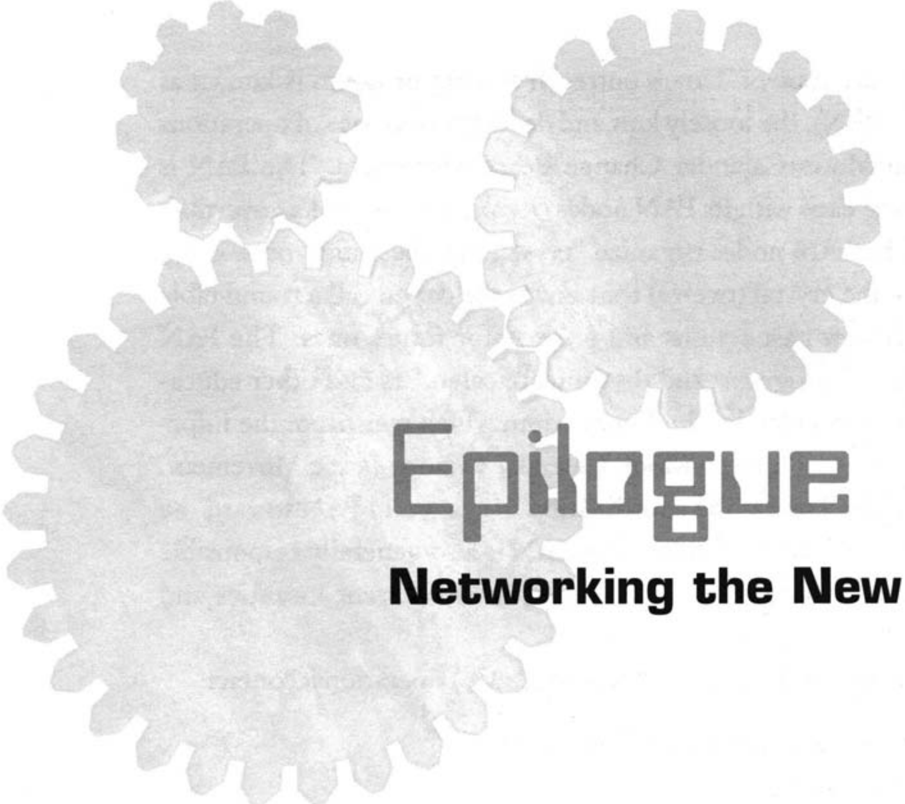

Epilogue

Networking the New Time

TO READERS for whom the notion of calendar change, much less that of a Campaign for the New Time, is totally novel, yet who may have an interest in learning more about and participating in the preparation for the Great Calendar Change of 2004, there is a place to begin and a way of going about it.

The official coordinating agency for the Campaign for the New Time and the World Thirteen Moon Calendar Change Peace Movement is the Foundation for the Law of Time. As a nonprofit public charities educational corporation, the Foundation for the Law of Time provides the communications and information hub for the Planet Art Network (PAN). The Foundation also publishes calendars, tools, and literature about the Law of Time and sponsors various educational programs, periodic seminars and congresses, and projects such as the new time educational caravans and the development of garden-oriented land bases.

For general information on the Law of Time or the Campaign for the New Time and its educational programs contact:

Foundation for the Law of Time, World Headquarters
Post Office Box 513
Brightwood, Oregon 97011 U.S.A.
Tel: +1 (503) 622-1976
Fax: +1 (503) 622-0198

The Foundation for the Law of Time's outreach activity program is known as the Planet Art Network (PAN), the loosely knit and decentralized base of operations for the World Thirteen Moon Calendar Change Peace Movement. The PAN is organized into Bioregions, each with its PAN nodes (local chapters) and a governing Bioregional Council. The PAN nodes organize "crystal day meetings" once every thirteen-day wavespell-the crystal (twelve) tone day is the day to call a round table or court of the kin to review past actions and prepare for future ones. The PAN nodes also offer educational programs and disseminate calendars and other educational tools and literature provided by the Foundation, which constitute the information outreach of the World Thirteen Moon Calendar Change Peace Movement. The Campaign for the New Time is the coordination of the world PAN toward the goal of the Great Calendar Change of 2004. The PAN is also generally responsible for organizing the annual Day Out of Time events in its different localities and bioregions.

For information about local Planet Art Network (PAN) operations, contact:

Foundation for the Law of Time/Gnomicile Land Base
World PAN Coordination Center
PMB 267
105 North First Avenue
Sand Point, ID 83864-1301 U.S.A.
Tel: +1 (208) 255-1172
Fax: +1 (208) 265-9107

The official Web site for the Foundation for the Law of Time, Planet Art Network, World Thirteen Moon Calendar Change Peace Movement and Campaign for the New Time is:

www.tortuga.com

This Web site is continuously updated and offers a great range of both practical and theoretical information about the Law of Time, the Campaign for the New Time, ongoing and upcoming events and educational seminars, listings of major PAN nodes worldwide, Law of Time archives, and links to other relevant Web sites.

A major aspect of the Campaign for the New Time is the development of a program of "strategic alliances," information about which is also available at www.tortuga.com. If you belong to an organization that you feel could be in alignment with the goals of the Campaign you can check out the Strategic Alliance page of the Web site or directly contact the Foundation for the Law of Time, World

Headquarters. The goal of the Strategic Alliances is to establish a "synchronization for peace force"-a coalition of a wide spectrum of organizations that are in alignment with the calendar change and are willing to help organize local events and information dissemination for the Great Calendar Change of 2004.

See also Appendix II, "The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time: A Brief Chronotopology of Transformation. "

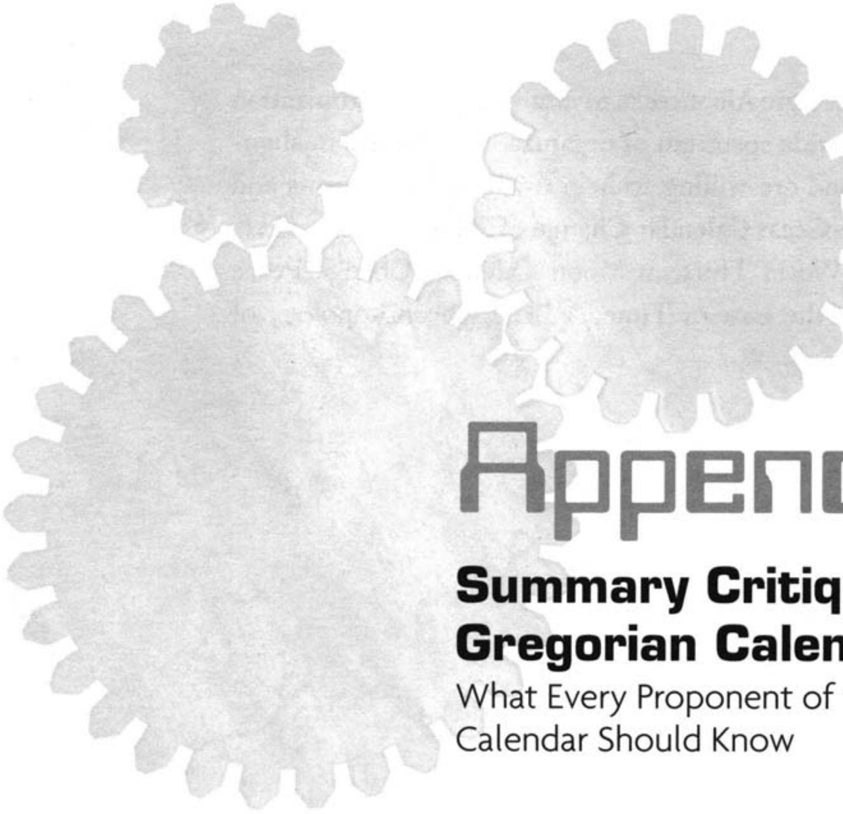

Appendix I

Summary Critique of the Gregorian Calendar

What Every Proponent of the Thirteen Moon Calendar Should Know

NOT ONLY is the idea of the calendar as an instrument to determine a true and accurate year misleading, but solely focusing on this purpose blinds us to any consideration of time apart from duration or measurement of motion in space. The Law of Time asserts that the actual nature of time is synchronic; hence the purpose of calendars is to synchronize us in time according to various cycles whose harmonic numbers extend from and return us to a higher mental order of reality. It is a fatal error to dismiss a mathematics of harmonic perfection when it is allied with the ordering and comprehension of cycles. The pursuit of a true and accurate year totally subordinates the cyclic nature of time to the ceaseless imperfection and change that characterize the phenomenal world when it is considered as the sole factor of existence. This defines the thoroughly materialist worldview of the 12:60 consciousness.

In order to prepare the ground for a proper consideration of the Law of Time and the evolutionary necessity of the synchronic order of time as the harmonic reorganizing factor of humanity in its post-historic phase, it is necessary to expunge from the mind the error known as the Gregorian calendar. To demonstrate and expose the illogical and irrational nature of the Gregorian calendar as a standard of measure, the following seven points are presented as a simple appeal to the intrinsic logic and intelligence of any human being.

1. We require of a standard of measure that its units of measure are regular and equal with one another.

This is not the case with the Gregorian calendar, whose base unit of measure, the month, proceeds in an irregular and uneven manner: 31 days, 28 days, 31 days, 30 days, 31 days, 30 days, 31 days, 31 days, 30 days, 31 days, 30 days, and 31 days. Why would anyone use a standard of measure with irregular units? Do you know what results when a crooked standard of measure is employed consistently for millennia? It might be noted that at the time of Augustus Caesar, August was called Sextile and had thirty days, while February then had twenty-nine days. In order to honor Augustus and make him the equal ofJulius Ouly) which had thirty-one days, the twenty-ninth day was taken from February and added on to Sextile, whose name was then changed to Augustus (August). By harmonic contrast, the Thirteen Moon calendar is perpetual in that all of its units of measure are equal-28 days each.

2. The names of the months are as illogical as their uneven numbering.

January is derived from the God of the doorway; February is an obscure word referring to an animal divinatory rite; March refers to Mars, the planet and the god of war; April and May refer to goddesses of the spring; June to the wife of Jupiter; July and August are named after the two most prominent Roman Emperors, Julius and Augustus Caesar. As for the remaining months, September, the ninth month, means seven; October, the tenth month, means eight; November, the eleventh month, means nine; and December, the twelfth month, means ten. Of course, having grown habituated to the crooked standard of measure, it is easy to overlook and dismiss as innocuous the irrational naming of the months. But is it so innocuous? What do the names of these months have to do with an order of time, or even a cosmology or culture of time, which we assume a calendar might reflect? By blindly accepting this irrational disorder of names, do we not predispose ourselves to accept irrational disorder in commonplace things around us, and even within the fabric of our society, thus reinforcing in us a tendency to accept a superficial treatment of the symptoms, while ignoring the roots?

3. Leap year and leap day is the most highly touted aspect of the Gregorian calendar.

Structurally, the Gregorian calendar is indistinguishable from the Julian calendar. The only thing that separates the Gregorian from the Julian calendar is the correction of the leap year day. Pope Gregory XIII "corrected the calendar" by deleting three leap days from the years that begin the centuries, except those divisible by the number four. Leap day is the extra day that accumulates every four years due to the length of the astronomical year being 365.241299 days and not 365. It can be seen that the fraction .241299 is not quite 1/4, which would be .25. The Julian calendar did nottake this into account, hence an error crept in that caused the spring equinox on the calendar to fall some ten days behind the solar moment of the spring equinox-a fact noted by the conquistadors when they encountered the Mayan calendar, which had no such discrepancy. Thus, in 1582, Pope Gregory XIII "improved" on Julius Caesar's calendar, some 1,627 years later, by adopting the rule that there would be no extra day on centuries-OO years-except on those that are multiples of four. Hence, there was no leap day in the year 1900, but in the year 2000, a multiple of four, there was.

What is not well known is that the Vatican does not recognize the leap day in its ceremonial calendar. Why is that? In most countries where Latin-derived languages are spoken, leap day and leap year are referred to as *bisiesto,* or "bisextile" day and "bisextile" year. If leap year is every four years, why is it referred to by a word that connotes six, *sextile?* On the official liturgical Church calendar there is no February 29! Instead, there are two February 24s, and the second February 24 is not counted. If there were to be an extra day that was counted, then the system of fixed feast days would be thrown off. Instead, on leap years, February 24, the Day of the Feast of St. Matthew, is counted twice-or extended to be 48 hours. And since in the Church tradition derived from the Romans the days are counted from the first of the next month, the first always being known as the calends, the date February 24 is technically referred to as the "sixth of the calends of March" (February 24 = sixth calends of March, February 25 = fifth calends, February 26 = fourth calends, February 27 = third calends, February 28 = second calends, and March 1 = calends of March). For this reason, the leap year is known as *bisiesto* because the sixth *(siesto)* calends of March is doubled, hence "bi." Thus, February 29 is not recognized by the official Church calendar of the Vatican, and it is also not counted as a day in its liturgical calendar. February 29 only arose out of popular tradition in the sixteenth and seventeenth centuries. The nemesis of accounting for the .241299 extra day per year reveals the fixation of astronomical time in seeking a "true" year and detracts from the possibility that the issue of an intercalary, or extra, day can logically and synchronically be handled in many other ways. In the end, it is the mystique surrounding leap day that further contributes to the numbing of the sensibility of time accumulated in the instrument known as the Gregorian calendar, "the little grid of boxes that rules so much of our lives."[1]

4. **The word we use to describe the instrument for measuring, the Earth's orbit around the sun, *calendar,* is derived from the word *calends.***

Calends was the Latin name given by the Romans to the first day of every month. What does it mean? *Calends* is the name of the account book, the book of payments recording the monthly debts and bills to be paid! No wonder we are ruled by the philosophy "time is money!" This philosophy is rooted in the very word we use to describe time reckoning, *calendar.* A more accurate word for time reckoning might be "chronometer," or even better, "synchronometer." But perhaps the word "count" is simpler. We might just say, for example, Thirteen *Moon!2* 8-Day perpetual count.

5. Dominicalletters code the years of the Gregorian calendar.

Scarcely known to anyone but Vatican insiders is the system of dominicalletters that is used to code the years according to the day of the week on which the first Sunday of the year falls. Since the week has seven days, there are seven and only seven dominical letters. These are the letters a-g, where a = 1, b = 2, c = 3, d = 4, e= 5, f = 6, and g = 7. Hence, G-7 (the group of the seven most industrial nations), the name given in 1974 by the CIA to the ruling oligarchy of globalization, is totally rooted in the system of the seven dominicalletters, a-g, to which the Gregorian calendar can be reduced. It works as follows: In 2001, the first day of the year was on a Monday, hence in this year all Mondays are coded by the letter a. Counting forward to the first Sunday, January 7, Sundays this year are coded g. The letter of the year, which is always a capital letter, is based on the lower case letter that codes the first Sunday, therefore, the year 2001 is coded as the capital letter G-it is truly a G-7 year! Not only that, but by this system of "G-7" dominicalletters, it can be demonstrated that the order of the Gregorian calendar year repeats within a cyclic structure of precisely twenty-eight years, where the days of the week and the month repeat once again. Hence, the cycle 1973-2001 is a repeat of the cycle 1945-1973, which is a repeat of the cycle 1917-1945, and so forth, where the years 1917, 1945, 1973, and 2001 are calendrically indistinguishable from each other. In any twenty-eight-year cycle, there are always exactly seven leap years! Thus, the key code numbers 28 and 7 of the Law of Time and of the perfect Thirteen *Moon!2* 8-Day calendar are hidden in and even govern the cyclic order of the Gregorian calendar, whose secrets pretend to lie concealed in the Vatican archives!

Expose these secrets and show that the true harmony of time is contained in the 13:20 matrix of the Harmonic Module, which is perfectly coded by 28 and 7. Within this matrix, any set of four tones radially opposite each other, an occult quartet, always adds up to 28. There are 65 (x4) such sets that constitute the Harmonic Module, while seven (multiplied by four), is the prime factor of 28, and is the key unit holding the 13:20 matrix in place, evident in the position of the seventh vertical column on either side of which is a perfect harmony of six columns each. The seven is also the implicit number in the ratio 13:20 (13 + 7 = 20: 20 - 7 = 13). But whereas the 13:20 matrix is a harmony in which the Law of Time is encoded, the Gregorian calendar is a disharmony, but one that is nonetheless governed by the Law of Time. What the Vatican attempted to destroy at the hands of Bishop de Landa in 1562, in the famous book burning at Izamal Yucatan, is redeemed by the Law of Time. Free of the obscuring, illogical irrationality of the inexact measure of the Gregorian calendar, the actual truth of the synchronic order of time may be discovered and revealed as the 13:20 mathematical code ofthe Harmonic Module, perfectly coordinated with the Thirteen Moon!28-Day count.

6. The Gregorian calendar makes day/date calculations difficult.

The twelve uneven months of the Gregorian calendar operate by a subsystem of fifty-two 7-day weeks, plus one day. Because of the irregularity of the numbering of the months, and because there are 365 and not 364 days in a year, it is almost impossible to make easy calculations month to month and year to year, while insistence on the unbroken succession of the week only compounds this issue. For example, if today on the Gregorian calendar it is Friday, May 4, 2001, what day of the week will June 4 be? What day of the week will July 4 be? There is an immediate mental block-a numbing of the mind. You have to stop and think about it. And in this numbing pause in which your cognitive brain has to be engaged, you lose your telepathic awareness, much as when you look at a clock to find out what time it is. Why should it be this way? Who benefits? The answer is, the priests (who know the tedious rules for stating that if it is Friday and it is 2001, then it is dominical e) and the bankers (who gather interest based on the confusion over the erratic disparity of days every month). By contrast, on the perpetual Thirteen Moon calendar, this day would be Spectral Moon Gamma 3-while every third day of every Moon is coded by Gamma. In fact, the third day of every week is coded by the name Gamma. Once the 28-day count is mastered, there is no need to engage the cognitive brain to figure out what day of the week Crystal Moon 3 or Cosmic Moon 3 will be-and in this way the mind is liberated into a telepathic knowing.

The system of the seven-day week was introduced into the Julian calendar at the Council of Nicea, A.D. 325, and was adopted from the Jewish calendar, which was derived from the Babylonians, for whom it was an astrological-astronomical construct. The fact that the week came from the Babylonians dissolves the argument used by the Vatican to counter the Day Out of Time. According to the Vatican, disrupting the succession of the seven-day week would disrupt an order set in motion by

God. The matter of the 7 and the 52 has a much deeper significance when understood in the higher-dimensional light of the Law of Time. A count of fifty-two 7-day weeks makes perfect sense if you have a count of thirteen 28-day months (7 x 52 = 13 x 28). The observance of the Day Out of Time, the 365th day of the year, produces a perfect and perpetual harmony. Knowing this, to continue to insist on fifty-two weeks while being unwilling to give up a twelve month count that does not have a Day Out of Time is to persist in an adherence to hopeless disharmony. Why do it?

7. What's in a name? Think about it.

What does it mean to follow calendars called the Julian and the Gregorian? A calendar is an instrument of control. The two most significant calendar reforms in history were the Julian calendar reform of 46-45 B.C., and its successor, the Gregorian, in A.D. 1582. Julius Caesar's motives had everything to do with his personal ambition and the conversion of Rome from a republican to an imperial form of government. Julius Caesar's calendar assured it to be the basis of imperial dominance. The course of the empire utilizing the Julian and later Gregorian calendar has prevailed as the dominant force now inseparable from the course of history itself. The 445-day "year of confusion" (46 B.C.), which attended Julius Caesar's reform, was matched by the second significant reform, the Gregorian, in which ten days were "lost forever," between October 5-16, 1582, so that the calendar could "catch up" with the sun.

While European Catholic countries easily accepted the reform, Protestant countries grudgingly acquiesced. Throughout the Americas, however, the Julian-Gregorian calendar was imposed as an instrument of power and symbol of dominance over the peoples the Europeans had conquered, including the high civilizations of the Maya, Inca, and Aztecs-all of whom happened to use, among other calendars, a Thirteen Moon!28-Day count. As with Julius Caesar, for Pope Gregory XIII the moment was politically ripe for a reform that would communicate itself as a means of expressing and extending power and control, but this time over the entire globe. As European dominance and control spread around the planet, even nations with their own established timing systems accepted, for the sake of "international policy," the Gregorian Oulian) calendar system for measuring the solar year. And so Western dominance over every aspect of global life became absolutely assured-until the moment of the Inevitable Event.

From its roots in the imperial ego of Julius Caesar to the timely "reform" of Pope Gregory XIII, it is not surprising that this calendar, "despite its odd quirks and the twists of history that produced it,"2 has become the standard of global civilization. Given the irregularity of the Julian-Gregorian calendar and the pursuit of the accuracy of astronomical time, history could be nothing more than a compilation of odd quirks and twists, while global civilization itself is the triumph of artificial time over the natural world. Only a species whose time sensibility had been captured by instruments of artificial measure could have become so alienated as to have produced the monstrous conundrum known as the "fast world," a civilization where money and technological advance prevail over human sensibility and the natural order. It is toward the correction of this destructive momentum that all efforts of calendar reform must now be directed.

In light of this critique, it is worth reprinting the opening section of the "United Manifesto by Advocates of Calendar Reform" that was first published in 1914 at the beginning of the First World War-some ninety years in advance of the Great Calendar Change of 2004. It can be seen that the issues of irregularity that prompted reform then are still issues today. However, the effects of not attending to these issues have only become compounded and even more complex, resulting in the chaos of a world at war with terror. Such is the nature of an error gone uncorrected over time-it only becomes more entrenched and turns into the dogmatic and hopelessly conflicted thinking of the everyday mind and its way of life.

> United Manifesto by Advocates of Calendar Reform
>
> WHEREAS we, the undersigned, have for some time been interested in a Reform and Simplification of the Calendar now in use in Western Europe, America, and elsewhere, with a view to equalizing the four quarters of the year, alleviating the irregularities of the months, and establishing a perpetual correspondence between the day of the week and the day of the month, and have supported one or other of several proposals which have been formulated for effecting these reforms; and WHEREAS said proposals usually provide for placing the 365th day of every year and the 366th day of Leap Year without the weekly and monthly enumeration; and WHEREAS we have found that in certain quarters-both ecclesiastical and scientific-Dbjections, possibly often sentimental, but none the less firmly held, have been stated to the employment and adoption of these expedients. . . Therefore we have resolved to unite in urging and advising that the very simple changes under noted should now be made in the Julian and Gregorian Calendars by international agreement. . .[3]

If you have a crooked standard of measure, and follow it because your parents were also following it, you have become a crooked man. It takes a crooked man to walk a crooked mile and build a crooked house. The issue of calendar reform is both logical and moral. Bad logic leads to bad morality. An error in time dooms the mind. Apocalypses are the reward for bad timekeeping. To remove yourself from the fires of your own self-created apocalypse, change your calendar. In a world of harmony there is no apocalypse.

◆

Appendix II

The World Thirteen Moon Calendar Change Peace Movement and the Discovery of the Law of Time

A Brief Chronotopology of Transformation

IT WAS during the seven-year cycle (1993-2000) when the World Thirteen Moon Calendar Change Peace Movement was being aroused around the planet, that the Law of Time was articulated in its various aspects and particulars, and the synchronic order defined in its whole system fullness. This occurred through a number of key events accompanied by a number of publications detailing the mathematical, cosmological, theological, and scientific aspects and imperatives of the Law of Time.

By the standards of the synchronic order, all events form points in a radial time set that describe a geography of time defined as a chronotopology. Coined by the late philosopher and mathematician Charles Muses, the term *chronotopology* defines an event point of meaning that creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge above a horizon of meaning.! The Law of Time now gives a precise field in real time to illustrate Muses's notion of chronotopology in which Time is a radial power of profound interrelations.

The following chronotopological sequence-a resume of the events in a particular geography of time, the seven years of prophecy, 1993-2000-defines the conscious emergence and articulation of the Law of Time. In this chronotopology, the "root meaning" occurred in 1989 at the Museum of Time in the form of the discovery of the 12:60 and 13:20 timing frequencies, and the subsequent working out of the implications of this discovery. The "horizon of meaning" was reached July 26, 1993, kin 144, with the decoding of the Telektonon Prophecy and the consequent development of the World Thirteen Moon Calendar Change Peace Movement. The "source meanings," however, extend back to the seventh century with the origins of the prophetic traditions underlying the discovery of the Law of Time. The seven years of prophecy established the field of radiated meanings of the Law of Time as a set of radiated event points and a body of literature and tools. Here follows, then, a brief chronotopology of the key event points of this movement:

1989	Root meaning: discovery of the 12:60 and 13:20 timing frequencies, Museum of Time, Geneva, Switzerland.
1990	Realization of the Thirteen Moon/28-Day calendar in its wavespell form.
1992	Recognition and celebration of the Day Out of Time as a planetary event.
1993	Horizon of meaning: decoding of Telektonon Prophecy and birth of World Thirteen Moon Calendar Change Peace Movement.
1994	Launching of the Planetary Moral Emergency and World Thirteen Moon Calendar Change Peace Plan.
1995	80 Days Around the World for a New Time of Planetary Peace: Chile, South Africa, Egypt. Russia, India, Hong Kong, Japan, Hawaii, and Mexico.
1995	Day Out of Time Calendar Burning and Return of Natural Time Ceremony, Serpent Mound, Ohio.
1996	First Planetary Congress of Biospheric Rights, Brasilia, Brazil.
1996	Biospheric Awareness Day, Gotemba, Mt. Fuji, Japan.
1996	Four Corners Boundary Dissolving Ceremony, Four Corners, United States.
1996	Harmonic Convergence of Humanity, Mexico City, followed by Peace March, Teotihuacan, Mexico.
1996	United Nations, Planetary Moral Emergency Appeal.
1997	Judgment Day Tribunal and World Congress on the Law of Time, Tokyo, Japan.

1997-present	Ongoing movement in Brazil to make July 25, Day Out of Time, official "Peace and Culture Day;' based on recognition of the Banner of Peace (seventy cities so far).
1998	Special peace mission to the Vatican, Rome, Italy.
1998	Special mission to Guatemala to meet with Quiche Maya elders.
1998	UR Council for the Theology of Peace, Milan, Italy.
1999	world Summit on Peace and Time, University for Peace, Costa Rica, followed by the dispatching of Special Emissaries to the United Nations and the Vatican.
1999-present	Ongoing movement in Japan to declare Day Out of Time a "Day of Festival"; more than one hundred events occurring annually.
1999	Peace March, Teotihuacan, Mexico.
1999	Seven Week Earth Wizard Seminary, Picarquin, Chile.
1999-2000	Ongoing in South America, launching of "La Karavana;' Thirteen Moon educational caravans, and in Europe, from the United Kingdom to North Africa, the New Time Caravan.
2000	Initiation of Seven-Day Earth Wizard's Seminars: Japan, United States, Italy, Russia, South America.
2000	Circumpolar Rainbow Bridge Experiment, phase I, worldwide.
2000	Fruit of meaning: establishment of Foundation for the Law of Time, Oregon.
2001	Launching of the Campaign for the New Time, Brazil. Argentina, Japan, Altai region.
2001	Campaign for the New Time PAN Thirteen Moon Leadership Conference, Oregon.
2001	Seven-day special Congress on Time and World Peace, Medellin, Colombia.

During this chronotopological sequence of radially connected event points, key writings, publications, and tools describing and defining the Law of Time and the synchronic order include:

Dreamspell: The Journey of Timeship Earth 2013, 1990-91

Thirteen Moons in Motion and *Turtle of the Thirteen Moons,* 1993, 1996

The Story of Time: The Story of Turtle and Tree, 1993

Treatise on Time Viewed from its Own Dimension (Published as *The Call of Pacal Votan:*

Time is the Fourth Dimension), 1992, 1996

Telektonon of Pacal Votan and the *Telektonon, Game of Prophecy,* 1993-1995

World Thirteen Moon Calendar Change Peace Plan, 1994-1995

Pacal Votan and Judgment Day: The Second Quranic Dispensation, 1995-1996

First Planetary Congress of Biospheric Rights, 1996

The Galactic Culture Master Plan, New Time Economics, 1996

A Brief Guide to The Thirteen Moon Calendar Change Peace Movement, 1996

Rinri Project: Four-Year Telepathic Biosphere-Noosphere Transition Circumpolar Rainbow Bridge Experiment, 1996

The 260 Postulates of the Dynamics of Time and the Evolution of Time as Consciousness, 1996

The Discovery of the Law of Time, $T(E) = Art$, *1996*

The Principia Mathematica of the Fourth Dimension, 1996

Twenty Tablets of the Law of Time: Sixteen Year Telektonon Cube of the Law, 1997

World Congress on the Law of Time and Judgment Day Tribunal, 1997

World Summit on Peace and Time, Summary, Review, and Conclusions, 1999

7:7::7:7, Telektonon Revelation and Plasma Universe Model, 1998-2000

The Rainbow Bridge Project: 28 *Meditations on the Law of Time,* 1999-2000

The UR Papers: Documents and Appeals for a Council of Universal Religion, 1996-2001

WIth the conclusion of the seven years of prophecy, marked by the rainbow bridge experiment, phase I, the greater part of the Law of Time had been uncovered, described, and defined. What emerges from consideration of all these events and accompanying literature is an entirely new model of the universe as well as a program for the reorientation of knowledge and the social reorganization of humanity-all based on a corrected understanding of time through the implementation of the calendar change. This was all to demonstrate that there is much more to a calendar change than the mere replacing of one calendar with another, for a calendar itself bears within it an understanding of time that tacitly and unconsciously shapes our very thinking about time. What the calendar change, then, really signifies is the end of one worldview and the birth of another. WIth the Law of Time, what was once unconscious is again made conscious, leading to a radical self-awareness of how wrong we have been about time, and what we might anticipate in changing our tools for navigating in time.

Because the growth of the Planet Art Network (PAN) as the social form of the World Thirteen Moon Calendar Change Peace Movement is inextricably involved with the unraveling by stages of the Law of Time, to define the constituent principles of the Law of Time is also to define the new worldview and the social organization appropriate to it. These constituent principles include: time is art; universal telepathy; holonomic consistency and reciprocity; synchronic order; fractal and radial mathematics of the 13:20 matrix; biosphere-noosphere transition; Banner of Peace; *Pax Cultura*, *Pax Biospherica;* the psi bank and the dynamics of the evolution of time as the evolution of consciousness; fourth-dimensional time and third-dimensional space. A mere consideration of these constituent principles will make one realize to what extent the calendar change is more than just an end to history, but the establishment of a new world, a New Heaven and a New Earth. This is precisely the promise offered by the Thirteen Moon calendar change and the discovery of the Law of Time. (See plate 8, Chronotopology of Transformation.)

▼

Appendix III

Telektonon Prophecy of Pacal Votan

THE REVELATION of the prophecy of Pacal Votan on July 26, 1993 was the culmination of the event that occurred on June 15, 1952, when the archaeologist Alberto Ruz Lhuillier pried open a large monolithic trapezoidal door beneath the Pyramid of the Inscriptions at Palenque, Chiapas, Mexico, and beheld what had not been seen by human eyes for 1,260 years: the tomb of Pacal Votan. A subsequent event, the lifting of the magnificent sarcophagus lid-measuring some four-by-three meters and weighing several tons-to reveal the human remains amidst a splendor of jade, including the jade mask, occurred on November 27, 1952. For some forty years afterwards, scholars debated the meaning of the tomb and the person buried within it. In the 1970s, the Russian scientist Knozerov dubbed the man in the tomb an "astronaut" or "cosmonaut" because of the similarity of the position of the figure sculpted on the sarcophagus lid to the positions of Russian cosmonauts in their little space capsules. This theory inflamed the imagination of some and was debunked by others, the end result being the enhancement of the awesome mystery of the tomb and the intention of its designer, the man buried in it.

The image of the sarcophagus lid is now world famous. Its symbolism has been interpreted to varying degrees by different archaeologists. In my book *The Mayan Factor*, the placement of Pacal Votan in his "time of power," 631-683, is defined and analyzed as occurring in the most harmonic and mathematically synchronic point of the entire thirteen baktun cycle, a point of fact that deepens the enigma of Pacal's knowledge and purpose. And herein lies the key question: What was his purpose, and why did he construct his tomb to be so elegantly buried for 1,260 years, to be discovered and opened just sixty years before the closing of the cycle, A.D. 2012? Clearly, the mystery of the tomb and its discovery had everything to do with the end of the cycle. At least this had come to be my surmise ever since the 1987 Harmonic Convergence. Like a hawk circling its prey, the meaning of the tomb finally came to my consciousness like a direct hit from the beyond. I was the prey, the hawk, the prophecy itself. Such was the revelation of the Telektonon Prophecy on the morning ofJuly 26, 1993.

The coded key was the tiled "speaking tube," by which the tomb was discovered in the first place. Suddenly, this tube spoke its meaning to me: Earth Spirit Speaking Tube, Telektonon by name. After four intense weeks of decoding the glyphs and symbols of the tomb while living in the pool house of an isolated estate in the wind-blown northern part of the main island of Hawaii, I knew that I had to abandon the dreamlike life I led in this Pacific paradise, and with my wife and companion head directly into the unknown, the first stopping point being Mexico, the source of the prophecy. After four months or so of traveling through Mexico, completing the decoding and announcing the prophecy, including a return to the tomb, we were finally directed to a converted carport in a field near the village of Ocotithin, not far from the birthplace of Quetzalcoatl. It was here where, for a period of nine mornings, I awoke dutifully two hours before sunrise to listen to the voice and write down by hand, in a specially prepared notebook, the words exactly as I heard them. Afterward, I numbered the 126 verses and distinguished the nineteen sections.

Because the prophecy itself is so inextricably connected with the Law of Time and the message of *Time and the Technosphere,* I feel it is important to present it as background information to the text of this book. In actuality, the text of the Prophecy is an integral part of the *Telektonon, Game of Prophecy,* where it is read, section by section, on a daily basis, according to the code number of the kin of the day.

THE TELEKTONON OF PACAL VOTAN

The Talking Stone of Prophecy That Unites the People of the Dawn with the People of the Book

"Then if they reject thee,
so were rejected apostles
before thee, who came with clear signs,

books of dark prophecies
and the Book of Enlightenment. "
-HOLY QURAN, SURA **III,** 19: 184

"All who obey God
and the Apostle
are in the company
of those on whom
is the grace of God,
of the prophets who teach,
the sincere lovers of truth,
the witnesses who testify
and the righteous who do good.
Ah, what a beautiful fellowship!

Such is the bounty
from God: and sufficient
is it that God knoweth all. "
-HOLY QURAN, SURA Iv, 9:69-70

● INTRODUCTORY VERSES: ONE LIVING PROPHECY

1. Pacal Votan's Gospel of the Telektonon
A Special Dispensation for the Day of Truth,
A.D. 1993-1994, kin 144.

2. Proclamation on behalf of the three messengers
of the awakening,
of the three special voices of prophecy,
the special witness of time and the last call.

3. Telektonon
the inner sun
the mother of all prophecy
message of the star witness, Bolon Ik
received and repeated by her servant in love,
Pacal Votan.

4. In the seventh year of Harmonic Convergence
the seven seals of the apocalypse
become the seven years of prophecy. . .
thirteen years in all.

Thirteen moons the path to walk;
thirteen moons the path to talk;
people of the dawn, one mind,
people of the book, one God;
one living prophecy, one people, one Heaven,
one Earth.

●● NAH CHAN: THE PALACE OF BOLON IK

5. o children of the people of the dawn, a children of the people of the book, I come as the special witness of time to remind you, especially on the day of truth, that in your origin you are one, and on the day of truth you are to make yourselves one again. For this I remind you of the Cube of the Law, Telektonon, and the 28-day, thirteen moon way that is the path of the righteous.

6. Nah Chan, Palenque, Xibalbay, Tollan, Xochicalco, Tepozteco, Amatlan: in these signs dwell my special agents, the witnesses of truth, 13.66.56 and Bolon Ik, Daughter ofJob.

7. For those with eyes, with open mind and sincere heart, this teaching is complete in every stone alignment and marking of the star command base now called Palenque.

8. New Jerusalem is the Palace of Bolon Ik, whose number is 1,728 (144 x 12), Cube of the Law. Seventeen is the power of Heaven, "solar Uranus," which is Earth in Heaven. Twenty-eight is the power of Telektonon, the Spirit Tower of the power of 7 times 4.

9. Never forget, children of the day of truth: All is number. God is a number. God is in all.

10. And everything there in the palace is laid out foursquare from the cord of Heaven, which is the command of the Telektonon. "Divine word never uttered until the Day of Truth." And in that palace, foursquare is the tower of the Cube of the Law.

●●● THE CUBE OF THE LAW AND THE TOWER OF BABEL

11. o children of the day of truth, before the separation that divided you into the children of the dawn and children of the book, there was only the Cube of the Law and the Law of the Cube.

12. Everyone and everything moved and lived within this Law: one mind, one spirit, one will.

13. Within the dimensions and measurements of the cube are all knowledge of Heaven and Earth. And in order to know, expand on, and delight in the sublime Law of the Cube, God created you, 0 children of the day of truth.

14. And though shaped of no more than mud and sounding clay, in you is the perfection of knowing the Cube of the Law, if only you remembered!

15. Children of the Day of Truth, just as the cube is the measure of the perfection of God's thought, so you are the measure of God's unceasing movement, which is called time.

16. In you the cube is the crown of perfect knowing of God's will in time. But true to his word, Iblis made wrong fair-seeming. To show himself man's enemy, Iblis split the cube. He drove a forgetting of time within the mind, separating mind from spirit, which is God's movement in all things.

17. And from this forgetting in man, mind and spirit became confused. The will weakened. Ego was born. Iblis's many-headed self triumphed. False authority arose. Babylon came to be and in its center the Tower of Babel, the monumental curse for no longer knowing time.

18. In the forgetting was the separation of the people of the dawn from the people of the book. Had there been no Tower of Babel, there would have been no book. The Tower of Babel was a deceit intended to show the Cube of the Law existing outside of the perfection of knowledge already formed within you, a children of the day of truth.

19. Product of the forgetting, the Tower of Babel has a shadow of ever-darkening and all-encompassing dimensions that spreads 5,000 years from Babylonia to all of Earth's most remote wildernesses and mountain retreats. Iblis's curse of Ego and false time are complete but hold only to the day of truth.

20. For God is gracious and most merciful. As long as He could, He kept the people of the dawn away from the Tower of Babel. And to both the people of the dawn and the people of the book, He sent many messengers, apostles of light, witnesses of truth, doers of the righteous way, which is always framed by the invincible Law of the Cube.

●●●● TELEKTONON: THE FRAME OF TIME

21. As the special witness of time, I, Pacal Votan, know the perfect count of days. I bow in the temple of the tower and the rock, the sanctuary of Bolon Ik. In my body, formed of the ultimate perfection of God's power of all movement and measure (Hunab Ku), is the recollection that is prophecy. Knowing this body as the measure of time, I had built this palace and this tower to correct and to also look for the Tower of Babel, a memory of the remembering. And in my ninefold temple, where, guarded by the nine powers of time, I had this body laid, I left this legacy for you, the Talking Stone of Prophecy.

22. Telektonon, the frame of time I left for you, is a mirror to show you your own truth of time for one last time. For now is the day of truth, the hour of judgment.

23. You, children of the dawn, and you, people of the book, unify yourselves the only way you can: in time, through time, as time.

24. Telektonon, the perfection of time, is the only way for you to escape the fire that consumes the unrighteous. If you who know, who have followed the straight way, the good path, now falter in this challenge of unification in time, Telektonon, which has already been prepared for you, you will perish forever in the fire that now consumes all unbelievers. God's command is in you. Do not listen to the evil one now.

MY COUNT OF DAYS IS PERFECT

25. Listen! My knowledge is this: My count of days is perfect, my knowledge of the Cube of the Law is unsurpassed. Twelve is the number of the temple, but thirteen is the number of God. Twelve times twelve, 144, is the number of the perfection of the temple as light. And twelve times 144, 1728, is the number of the Cube of the Law, Telektonon, the divine word unuttered until now.

26. "... To him that overcometh will I give to eat of the hidden manna, and will give him a white stone, and in that stone, a new name written, which no man knoweth, saving he that receiveth it." (Revelations 2:17) And that name is Telektonon, and that stone is crystal, and that hidden manna is knowledge of the revelation of time.

27. The number of the elect is 144,000. This was the number set aside by God to be free ofIblis's curse. Dispersed is this number among mankind, and always was it meant to multiply both among the people of the dawn and the people of the book.

28. But since a human kin is the measure of the perfection of time, 144,000 is also the number of days within a cycle of twenty generations.

29. Twelve times 20 generations is 1,728,000 days, the number of the Cube of the Law. But for God, one cycle more, 144,000 days more to entice Iblis to make his score, a total of 1,872,000 days, thirteen cycles of God's elect.

30. I am Pacal Votan, time's special witness, galactic agent of the Nine Great Lords of Destiny, by oath pledged to the honor of Bolon Ik.

31. In my body of time, thirteen star histories rehearse themselves for the day of truth. Each star history possesses its own knowledge keeper appointed by the Nine Great Lords of Galactic Destiny. Thirteen Wisdom Holders, Nine Great Lords, twenty-two in all, to each of whom I am bound by a luminous thread of knowing, sealed by oath to Bolon Ik, White Solar Wind, kin 22 in the Book of Destiny, the Book of Kin.

ALL IS NUMBER

32. All is number. God is a number. God is in all.

33. Child of the day of truth, listen: 20 is the number of totality. You have twenty fingers and toes because God has made you the totality of time. Thirteen is the number of God's cosmic wisdom, perfect in its power of unceasing change and circulation. Thirteen are the joints of your limbs and body. Seven is the difference between 13 and 20. Seven is the mystic perfection between one and thirteen. Seven is the erect spine of the vigilance in which are met the 144,000 mystic fibers of the body's thirst for prophecy.

34. 144,000 kin, 144,000 days, these are called a baktun. Thirteen baktuns, one perfect cycle to rehearse all thirteen star histories. Each baktun, twenty generations, each generation 7,200 kin. So it was that Bolon Ik ordained the sacred count of days, each day a sacred kin, each kin one of the elect, a turning in the twist of universal time, each kin ruled by the power of the nine, the power of the Nine Great Lords of Time.

35. So began the sacred thirteen baktun count, 3,113 years before the birth of the second of the special messengers, the Apostle Jesus Christ, who sent to you a Book of Dark Prophecies, twenty-two sacred chapters in all, through his appointed witness, John of Patmos.

36. All this I know and declare to be true, for I am time's special witness to the day of truth, Pacal Votan, rehearser of the thirteen star histories, he who adds the 144,000 elect to the cubing of the cube, assuring not twelve baktuns, but thirteen to be the number to bring on the day of truth, God's judgment of your power of remembrance!

37. Children of the people of the dawn, in separating you from the people of the book, God meant you to be the reminder that wisdom is already perfect in your bodily form of time.

LORD BUDDHA

38. In remembrance, many messengers did he send you, many fires did he light high atop the mountain for you to see. And one special messenger did he send, one special witness of the dawn.

39. In India, at the midpoint of the seventh baktun, the very center of the mystic cycle, thirteen baktuns in all, Lord Buddha did awaken at the dawn, the morning star, the tree, the Earth, his witnesses of illumination. For seven weeks, he sat beneath that tree. Never moving, he traveled the seven directions and visited with his mind every realm of God's vast creation, only to learn the turning of the Wheel of the Law.

40. Withstanding all efforts of the evil one, he taught unceasingly for nine times five years more, until the age of eighty. All mind teachings did he give; the unwritten Book of Enlightenment did he leave with the Earth. His last turning of the Wheel of the Law was the teaching of the Wheel of Time. Given at the request of King Suchandra of Shambhala who traveled far from Central Asia to receive this teaching, the teaching of the Wheel of Time closed the life of Lord Buddha.

41. "All composite things are impermanent," declared Lord Buddha at his death, "Be a lamp and a refuge unto yourself, look to no outer authority but yourself!" The fire of Lord Buddha's illumination lit many lamps. The Book of Enlightenment he left with the Earth is a book for all to know in silence with the Earth.

42. In Shambhala did King Suchandra transmit the teaching of the Wheel of Time, Kalachakra, as it is called. Invoking the Nine Great Lords of Galactic Destiny, Shambhala prospered in abundance and wisdom. When came the seventh ruler after King Suchandra, Queen Visvamati, the people of Shambhala ascended as one into the paradise of the fifth dimension, where their wisdom continues to water the rivers that flow beneath the gardens of the righteous.

43. Integrating the power of the five, which is the center, with the power of twelve, which is the root of the cube, the Kalachakra is also the cycle of sixty years (5 x 12). By the Law of the Cube, which is the power of sixteen, foursquare or 7 plus 9, sixteen cycles

of sixty years is the appointed time of prophecy, the prophecy of Kalachakra, which is the time of the day of judgment.

44. So it was that after the arrival in Tibet of Buddha's special voice, Padmasambhava, late in the tenth baktun, that the Lords of Shambhala determined to release the teachings of the Wheel of Time among the people of the dawn inhabiting Tibet. Thus it was that the Wheel of Prophecy turned the first of these sixteen cycles in A.D. 1027 and completed the sixteenth of these 60-year cycles in A.D. 1987, in partial fulfillment of the Harmonic Convergence of all prophecy.

45. 0 people of the dawn, do not doubt. The appointed day is come.

THE HEPTAGONON OF MIND

46. People of the dawn, people of the book, children of the day of truth. To safeguard as long as possible the teachings of the Original Ones from the confusion of the Tower of Babel, God divided the Earth by its one ocean into great island masses called continents. And most far from Babylon was Amerrikua, and also distant was Australia and the Island Kingdoms of Polynesia. Even in Africa, close to Babylon, for many great cycles were the people of the dawn protected in their natural wisdom.

47. But inexorable is my count of days, inexorable the rolling of the Wheel of Time, inexorable the Law of the Cube.

48. Telektonon is the name of the blessing received by those who master the Law of the Cube. Great is the power of God, whose power and blessing are ever close and near, foursquare the manifest action of radiance from the divine intersection of God's presence: divine mind, divine spirit, divine will, divine source from which the Cube is sprung. Foursquare the Cube, the measure of Telektonon, one vast, far-spreading word that is no word at all, but number multiplying itself from within God's unending meditation, which we choose to call creation.

49. From one intersection, six directions plus the seventh, the moving center of time, the navel of Heaven whose cord is a fiber called Kuxan Suum. From the navel of Heaven to the solar plexus runs the Kuxan Suum, highway that reaches to the roots of the stars, passing through each of the heavenly dimensions.

50. I who say this, know, for I am Pacal Votan, who left for you a talking stone, Telektonon. Time's special witness am I, God's spy holding vigil, keeping the sacred count of days, exposing to you the lie of Babylon, the falsification of the Cube into the Tower of Babel!

51. Who now but I can tell you that seven are the directions of the Cube, the Heptaganon of Mind: twelve are the gates, sixteen the powers; moved by the nine great powers of time, these sixteen powers become the number of the elect of the Cube, 144. Multiplied by the millennium, 144 becomes 144,000, sacred number of days I laid for prophecy to know during my first baktun count.

IMPERIAL BABYLON

52. Children of the book, even before the close of my first baktun, twenty generations from the year 3113 B.C.,the foundations of the Tower of Babel were laid. Captive now were the 144,000 within the Earthly prison of Babylon's materialism.

53. Yet, by my devotion and power of meditation, from deep within the Earth, aided by Bolon Ik's perseverance on far Uranus, Earth's harmonic twin, I kept the sacred count of days, I watched over the movement of the sacred march of kin.

54. Four more baktuns passed, and Babylon's power was swollen with the loot of Empire. Despite the righteousness of Abraham and the Law of Moses, the shadow of the Tower of Babel only increased in size and darkness.

55. Came the sixth baktun and Imperial Babylon was complete. Its seal of false time and power, taxation, and armed expanse had fulfilled Iblis's desire to degrade and ruin all humanity by making wrong fair-seeming. Fair-seeming now to amass wealth and power; fair-seeming to make a count of time by twelve months; and fair-seeming to regard the thirteenth moon with superstition and contempt! Fair-seeming to gather taxes and to pay them; fair-seeming to make war for the power of King's; fair-seeming that birth and death must be paid for; fair-seeming that all power come from the priests, Iblis's most cherished of men.

56. 0 people of the book, lost to you was all memory of the Law of the Cube, slavery became your lot! But I am Pacal Votan, time's special witness. Even before the coming of the second special messenger, Christ the Apostle of Love, by my free will and my knowledge of the power of seven, I laid claim to the Seven Seals of Prophecy that on the day of truth, Iblis's curse would be finished and the righteous would triumph. And now the day of truth is come. Telektonon, the talking stone, now speaks in a voice for all to hear.

57. By the seventh baktun, the Seven Seals of Prophecy were stored in Earth's special lodestone of truth, whose custodian am 1.

58. By the eighth Baktun, Iblis had caused the Tower of Babel to speak more tongues in a growing clash of empires. Into this confusion came the Apostle of Love, crucified for driving the moneylenders from the Temple, misunderstood by the people of the book, finally to be absorbed by the Priests of Babylon, now Roman by tongue, who use this Christ as their emblem to spawn Iblis's greatest evil, the total usurpation of Earthly Time.

THE BOOK OF DARK PROPHECIES

59. One special voice did this Jesus have, a man named Saint John ofPatmos, in whom the gift of prophecy was fulfilled. From my ark of universal time, Earth's special lodestone of truth, rock of Heaven, attained by thirteen tones, I guided Saint John's hand by the mystic power of seven. And so was sent to the world the Book of Dark Proph-

ecies, Revelations, the Apocalypse, twenty-two mystic chapters, the first thirteen mirroring my thirteen great baktuns, nine more for the Lords of Time, seven of these for each of Seven Years of Prophecy, and the last two to complete the visionary power of the righteous of the Cube: 144,000 times thirteen.

60. To those who can read they will find Iblis's numbers 1,260 and 666, well known on the day of truth, for 12:60 is the ratio of the false time that makes martyrs of the righteous, and 666 the number of the beast who captures my thirteenth baktun.

61. Children of the day of truth, children of the dawn, people of the book, all this can you hear in the stones of Palenque where in the tenth baktun I took pure human form. Of the twenty-seven books of the New Testament, the twenty-seventh is Revelations. Twenty-seven steps ascend to the Palace of Bolon Ik. This palace, New Jerusalem, is the mystic power of Telektonon, the twenty-eighth step, the power of seven sanctified foursquare.

62. Twenty-two chapters of Revelations are the twenty-two kin complete in Bolon Ik: thirteen for the power of time, nine for the Nine Lords of Time, whose forms still guard my ark of stone deep within the ninefold temple of the Cube of the Law. Perfect was the time of my coming, kin 1, 366, 560. Perfect was the time of my going, kin 1, 385, 540.

63. Installed in the House of the Serpent, my coming was commemorated by the departure of the third messenger, the special messenger of the day of truth, the Apostle Muhammad, who left for the righteous of the Earth the Book of the Righteous, the Holy Quran, clear signs to understand.

64. As Buddha taught so pure one mind, so Muhammad taught so pure one God. Submission to the will of God is at the center of the Law of the Cube. Truly for the righteous there is only the straight way, the path of basic goodness, laid out in the heart as obedience to God's will, the divine plan unfathomable to those who cry, "reason alone can save us!"

THE PROPHECY OF THE SEVEN GENERATIONS

65. I am Pacal Votan, witness of time. Telektonon is my gospel. The talking stone of unifying prophecy is my legacy intended for that special messenger who understands the meaning of my numbers.

66. By my symbolic channel of communication called Telektonon did I intend this prophecy to be discovered and recovered. Once forty years had elapsed from the time of the discovery of my Uranian crypt, then would come the recovery, the recollection by my special agent of the numbers rehearsed and repeated to announce the day of truth.

67. Perfect was the time of my coming. Perfect was my mission. Perfect was the time of my going. Perfect is the discovery of my prophecy. Perfect is the Telektonon, the little book the angel of the Lord made Saint John of Patmos taste and eat, which you now possess, a treasure for the day of truth.

68. What you witness by these words is the last message I gave to my people, the message of 9.13.0.0.0, heard from my ark of stone through the Telektonon, Earth Spirit Speaking Tube. To my people I declared:

69. "Seven katuns are there to follow me. Seven generations, each one guarded by a solar witness taking account of what you do, each solar witness sealing your time with a prophecy to be known as the Book of the Seven Generations. And this book is to be opened for the day of truth.

70. "After the seven generations, the end of the baktun. Ten baktuns gone, the glorious Age of Maya gone, the Long Count of my days to be completed at 1,440,000 kin, the mystic number of the elect multiplied by ten.

71. "Then will come the three baktuns of Babel, the coming exile of 12:60 time. Thirteen years after this tenth baktun shall come the last prophetic round: thirteen cycles of Heaven, each fifty-two years, followed by nine cycles of Hell, each also fifty-two years. When the third round of Heaven begins, the Messenger of the Last Call will come to you, Quetzalcoatl-Kukulkan, the prophet by name.

72. "And when the thirteenth Heaven cycle is over, then will come the complete destruction of all we are and know. Before the twelfth baktun is over, two priests bearing false witness, ornamented with the sacred numbers five and thirteen, will see to it that the thirteenth baktun will be the end of time.

73. "Unrelenting will be the roll of the Nine Hells into the abyss of fire and madness. In the time of the last Hell, though you will be declared free, free you will not be. Ten Kings in succession will mark the climax of time. Seven years will follow the ending of the last Hell. Then will sound the angel of the seventh trumpet, the seven thunders will roll, the beast of seven heads called G-7 will place his mark on your foreheads and hands, the false one from Babylon will return to be embraced by the 11 Serpent. Then the day of truth will come, the Seven Years of Prophecy."

TIME'S SPECIAL WITNESS

74. People of the dawn, people of the book, children of the day of truth, many messengers have been sent to you, many prophets have been given to you. A Book of Enlightenment has been placed in the Earth for all to read in communion with the Earth. A Book of Dark Prophecies has been left for you. A prophecy of thirteen Heavens and nine Hells to accompany the Book of Dark Prophecies has also been left for you. A great prophet and messenger has left you clear signs to understand and a book of unassailable righteousness, the Holy Quran.

75. As the special witness of time, I have left for you my talking stone, Telektonon, breath of precious Bolon Ik, and have sent you both a prophet, Quetzalcoatl of sacred Xochicalco, and for the day of truth I have appointed also a special messenger who can hear my talking stone and write and explain its numbers and its meaning for you, children of the righteous.

76. Do not doubt any of this. When the moment comes for him who is to be the eleventh in succession, to succeed the one, 11 Serpent, then join the believers and begin the prophecy that liberates from all Kings and successors.

77. Ten witnesses I place around my tomb, ten Uranian emissaries, one each for each of ten baktuns. Of these am I the eleventh. Who would be the eleventh in the final succession of the Kings of Mexico must own me as the eleventh, then will all go well.

78. Six messengers did I place on my talking stone of prophecy: three above, the messengers Lord Buddha, Muhammad, and Christ; and three below, Padmasambhava, Quetzalcoatl, and Saint John of Patmos. Of these am I the seventh.

79. Pacal Votan am I, time's special witness, messenger of the Telektonon, instrument of the seventh angel, I declare again and again: All is number. God is a number. God is in all.

WAR OF THE RIGHTEOUS

80. Look carefully at my stone and listen: ten messengers, 24 signs, for ten are the number of orbits of planets around this star, Kinich Ahau, your Sun. From my point of origin, your star is designated 24, number of the circuit of externalizing intelligence. If you are of the righteous, then in this is a sign for you of your star mission.

81. And from each of the three oracle mouths of the arms of the cross of my Kuxan Suum, you will find twenty-four rays repeated three times, one-half the number of the elect (3 x 24 = 72). Add to this 3 times 11 (33) and the number is 105, the difference between the number of days in your solar orbit, 365, and the number of kin in my sacred count, my galactic spin, 260.

82. 13:20 is the ratio of natural sacred time. 12:60 is the ratio of the false time of the Tower of Babel. 13:20, thirteen moons, twenty fingers and toes. 12:60, twelve-month year, sixty-minute hour. Children of the day of truth, in this is a sign for you to discriminate between the power of the evil one who seeks to trap you in his machine, and the power of the righteous who need nothing but that with which God has endowed them: their body of time and their natural path synchronized by thirteen moons.

83. Children of the day of truth this is your only decision, for God knows and sees all that is in your hearts. People of the dawn, people of the book, on the hour of judgment unify yourselves in the only way you can: in time, as time, through time. For this in the year of kin 144, Telektonon is revealed to you, God's plan for peace on Earth, the last and only hope for spiritual renewal and salvation, immediate acceptance and adoption of the twenty-eight-day, thirteen-moon way, the calendar Telektonon. "And he had a name written that no man knew but himself." (Revelations 19:12) And that name is Telektonon.

84. Children of the day of truth, people of the dawn, people of the book, foursquare is my talking stone, invisible its design, its crystal lacing of star histories met in my human form. Time's special witness am I seeking apostles of time to organize the day of truth

into the revelation of God's redemption. Universal forgiveness awaits those who understand God's plan, a common time of twenty-eight days, thirteen moons, begun together at the appointed hour, on the day called Lunar Moon, the first day of the second year of prophecy.

85. Two solar faces there are upon my talking stone, the one an icon of my solar shield, Galactic Sun, the other, Lunar Sun, an emblem of the perfect lunar timing of Earth's solar orbit.

86. Children of the day of truth, the hour of judgment has come for you. How much longer do you need to suffocate in the Babylonian exile of 12:60 time? My children are the oppressed. Is it not time for the war of the righteous to restore equality? Why not wage the war then that destroys altogether the old time? Why not smash the caskets of the Babylonian wine of fornication in a single stroke?

HOLY VICTORY MARCH

87. 0 children of the day of truth, people of the dawn, people of the book, in your origins you were one blood, today the hour of judgment demands you become one again, unified in time. Who among you will join my apostles of the Telektonon? Already the gift has been given. Already we left for you the 13:20 tool, my sacred count of days, 260 kin, to turn the solar-lunar wheel of365 days. And when prophecy was vindicated at the conclusion of the ninth Hell, we sent among you the perfect form of this gift, Dreamspell, containing the message of the thirteen moons.

88. But this was not enough, the seventh trumpet had yet to sound, and so it did in the fortieth year following the discovery of my Uranian ark of stone. And now my prophecies are revealed. The Mayan Apocalypse of kin 144, Yellow Magnetic Seed, turns the seventh year of karmic destiny into the opening of the first of my seven seals of prophecy.

89. In this seal foursquare, all prophecies converge. The Beast of the G-7 stands revealed. The Babylonian Vatican is exposed. In the rainbow dream vision 144,000 of the elect are called again to meet, gathering together in circles to listen, to sing, and to dance to the song Telektonon.

90. 0 children of the day of truth, people of the dawn, people of the book, seven is my sacred power. Foursquare seven is 28, the number that, perfectly followed, opens Heaven's gate. Walking together the sacred thirteen moon, 28-day way, you will be joined to God's command, a holy victory march of the righteous to restore sacred order to this troubled Earth and to recover your lost powers.

91. Four powers of seven are given to you to be your comfort in the Victory March, one power for each of four perfect weeks, every twenty-eight days. By the first power of seven, claim your power of prophecy, by the second power of seven, maintain your power of prophecy; by the third power of seven, proclaim the victory of prophecy; by the fourth power of seven, become the victory of prophecy!

92. 0 children of the day of truth, 144,000 times thirteen is an army of the righteous. Learn again the count of sacred time and join this army. 0 children of the day of truth, realize that everything about the old 12:60 time is now become a garment of false being, a moral corruption, a banquet feast for the evil one.

THE STORM OF NON-EGO

93. 0 children of the day of truth, children of the people of the book, in the Storm is the sign of God's power. To the Storm is assigned the power of nineteen, the mystic saturation of all number. One hundred fourteen Suras, nineteen times six, the power of the Cube, are the number of Suras in the Holy Quran. Twenty-two years after the finding of my Uranian crypt, Telektonon, came complete revelation of the Quran's power of nineteen, Allah's mercy. Nineteen more years then to unfold the karma of Allah's prophet besieged by Babylonian conspiracy. Then comes the destined year nineteen, kin 144, hour of prophecy's convergence, moment of the day of truth.

94. Let the 124,000 messengers call forth a remnant of 20,000 more; let each of these sound seven trumpets, seven times. Let these seven times seven trumpets call for a following of God's perfect plan. Mark seven times seven years since the evil one showed his fiery face at Hiroshima. Be honored by God's cancellation of Iblis's power of moral oblivion. Let the storm of God's power of nineteen cleanse with cosmic mercy the blood of all martyrs from the thrones of Kings, and let these thrones themselves be dissolved by the pure water of sacred time flowing through the gardens of the righteous.

95. I am Pacal Votan, time's special witness, revealer of the Cube of the Law, devoted patron of the mystery that is woman, the consummation of the telepathic power of twenty-eight, known to me by name of Bolon Ik.

96. Pacal Votan am I, enigma to your men of reason. My existence is proof of the workings of higher law. Telepathy is participation in God's knowing, but is no cause to claim full knowledge of His powers. Prophecy too is natural law, the complement to God's mercy.

97. 0-19 is my code. In it are all demonstrations of the wisdom that is always close and near, but that comprehends the web of galactic being and all forms of its natural orders in time. Just as nineteen is the power of God's mercy, so zero is the fullness and emptiness all at once that characterize the nature of mind and all it can know.

98. Lord Buddha called this power of zero Sunyata. The special voice of Buddha, Padmasambhava, whose life span followed mine, exactly as Muhammad's preceded it, knew this Sunyata to be the basis of non-ego.

99. If you would know me and all that I witness for you in this hour of judgment, children of the dawn, people of the book, then practice the being that is non-ego.

THE ECSTASY OF DELIVERANCE

100. Because I repose on Telektonon, the mother of all prophecy, mine is the ecstasy of deliverance. With a mind of non-ego clasped to the heart of God I have sent out my

message of the Mayan Apocalypse. By relinquishing the calendar of twelve and leaving the prison of mechanized time, God's mercy will come swiftly to you.

101. Thirteen moons, twenty fingers and toes, thirteen joints, four limbs, five senses do you already possess. God's knowing and power of movement are already in you like a mother pregnant with child.

102. Why tarry now? When a house is burning, scurrying for your possessions will only cost you your life. This house that is burning is Babylon, the 12:60 way of life, which is now destroying your biosphere and poisoning everything about your life. Do not linger now fondling your machines. This is the Day of Judgment. If you can remember, leave this burning house and know that a better way has already been prepared for you, a holy victory march, my seven years of prophecy, Telektonon, the sacred revelation of time.

103. The Cube of the Law is awaiting its construction through the collective mind, spirit, and will of you, 0 children of the day of truth. Iblis, Satan, the evil one is none other than your ego clinging to the fatal lures of Babylon. God knows nonetheless exactly what is in your heart. Why not, then, become an apostle of the holy victory march? The ecstasy of deliverance is meant to be shared by all.

THE PROPHETS OF PEACE ARE AWAKENING

104. Pacal Votan am I, time's special witness am I, who bequeaths to you a talking stone, a legacy of place, Nah Chan, House of the Serpent, by prophecy known as Palenque.

105. And at Xochicalco, too, you will find stone utterances of my gospel left by command of my apostle, Topiltzin Ce Acatl Quetzalcoatl, who closed Christ's first millennium, as now my prophecy is come to close Christ's second millennium.

106. 0 people of the dawn, people of the book, children of the day of truth, make yourselves as one again and do not doubt the accuracy of the appointed hour on the day Lunar Moon! Synchronize now, and in all humility accept God's plan, twenty-eight days, thirteen moons, plus one day for deliverance and forgiveness.

107. Pardon now in public your enemy's indiscretions, and make full confession of your own transgressions. If you are sincere, God will see this in your heart and by His mercy many will follow your way; and if you proclaim for the cause of righteousness and truth, a new beginning and a new time, Telektonon, the thirteen moon way, and you enjoin others to do likewise, then God will know this, too, and make you a great general in the holy victory march.

108. And, if you have means and wealth, and you hear these words, or have explained to you the curse of 12:60 time and the blessing of 13:20 time, then do not doubt it and practice charity. Provide for the children orphaned by war and provide for the reconstruction of the Earth, that the practice of art and culture may flourish once again. Practice such charity with zeal, for soon you shall see an end to money. No more will flow the drunken wine of Babylon's fornication, and then what will you be with your paper credits and your useless machines?

109. The prophets of peace are awakening. Truly, the hour of judgment is come. Among the angels am I known as the mysterious benefactor of the righteous. Will you throwaway this opportunity, too? Beware for now the fire that consumes all falsehood is raging.

THE WIZARD'S HOARD OF THE LORE OF DEATH

110. People of the dawn, people of the book, children of the day of truth, because I am time's special witness, everything about my talking stone of prophecy is in fulfillment of God's divine plan.

111. Mine is the teaching of the 12:60 way of exile and the 13:20 way of redemption. Exactly 1,260 years passed from the dedication of my temple and closing of my Uranian crypt, A.D.692, 9.13.0.0.0, to its discovery and reopening in A.D.1952, 12.17.0.0.0. Another five times seven years, the numbers of my sacred ratio passed before the closing of the ninth Hell, A.D.1987, commemorated by the sacrifice of 3 Monkey, the first of four sacred signs on the western edge of my great talking stone. In this sacrifice was a sign of the death and transcendence of the way of the people of the book, "Joshua," and the people of the dawn, "Maitreya."

112. Eleven sets of triple disks, thirty-three in all, will you find upon my stone, for 33 is the number of the initiate, the middle harmonic of 65 that constitutes the Tzolkin, my sacred count of thirteen times twenty. All clear signs have I left for you, two borders of stars, the host of the Nine Lords of Time repeated by number on each side, top and bottom of my stone, a ledger of star histories complete in my form.

113. But especially would I have you look at Mars and consider again the disposition of the thirteen sacred signs placed around the edge of my stone, from 8 Ahau, my seal of truth, to 13 Cimi, the galactic closing of the wizard's hoard of the lore of death.

114. Ten emissaries celebrate the glory of my gifts for you, signs of peace from other worlds that beckon you to awaken from your sleep of Earthly Hell to join yourselves to God's command, Telektonon, the 28-day, thirteen-moon way and find in that count of days a path that leads directly to the paradise that is the repose and reward of the righteous.

115. One hundred forty is the number of the Telektonon, perfect its ratio 5:7 (28 x 5 = 20 x 7), the same as my sacred ratio 260:364(+1)

116. All is number. God is a number. God is in all.

SEVEN PERFECT ORACLES

117. People of the dawn, people of the book, children of the day of truth, this completes my testimony, I, Valum Votan, holder of the solar shield, protected have I been by the perfect love of Bolon Ik, who has stored my prophecy until this moment of release.

118. Seven perfect oracles have I given you, a perfect calendar of thirteen moons, an instrument to renew your spirit and your powers of Prophecy, Telektonon, and a book

of knowledge for you to penetrate the forgotten history of your star, Dreamspell. Also have I sent two messengers who bear witness to my truth, and thirteen perfect signs.

119. Now is the time of my prophecy. Now is the day of truth. Now is the opening of the seven seals. Now is the sounding of the seven trumpets. Now is the hour of judgment. Now is the song of the 144,000, the seed army of the righteous.

120. When the Lunar Moon has overflowed its banks, the G-7 will be no more.

121. When the Electric Wizard arrives, gather in banners of righteousness.

122. When arrives the Self-existing Storm, you will be fortified with spiritual abundance.

123. When comes the Overtone Seed, neither money nor priests will besiege you any longer, but the Earth will sing in harmony with your righteousness, Babylon will be no more, the Law of the Cube will be fully rehearsed in the hearts of the righteous.

124. In the sign of the Rhythmic Moon, senses will open to the manna of Heaven; no where will you find any longer signs of the evil one; a cosmic race will you be once again.

125. And when comes the seventh year, the Resonant Wizard will seal the Heavens and the Earth with the glory of all prophecy, the redemption of all prophets, messengers, apostles, and witnesses of truth.

126. Five years then will you have to close the cycle ofthe Seed, and another eight, thirteen years in all, to prepare for the coming of the Galactic Seed, 2013 Arcturus Dominion, the fulfillment of the star cycle of Valum Chivim, witnessed by the special agent, Pacal Votan, to the glory oflove, Bolon Ik, and on behalf of the mercy of God's divine plan, Telektonon, Cube of the Law, the mighty talking stone of prophecy.

"Man is one. Culture is indivisible."
-ALBERTO Ruz, 1952, 12.17.0.0.0.

Transcribed 12 Seed, kin 64,
Resonant Moon Day 13,
Warrior's Cube 7, 0 Yax
Yellow Magnetic Seed Year, kin 144,
First Year of Prophecy

Blue Spectral Monkey/White Solar Wind
Tepoztlan, Morelos, Mexico

▼

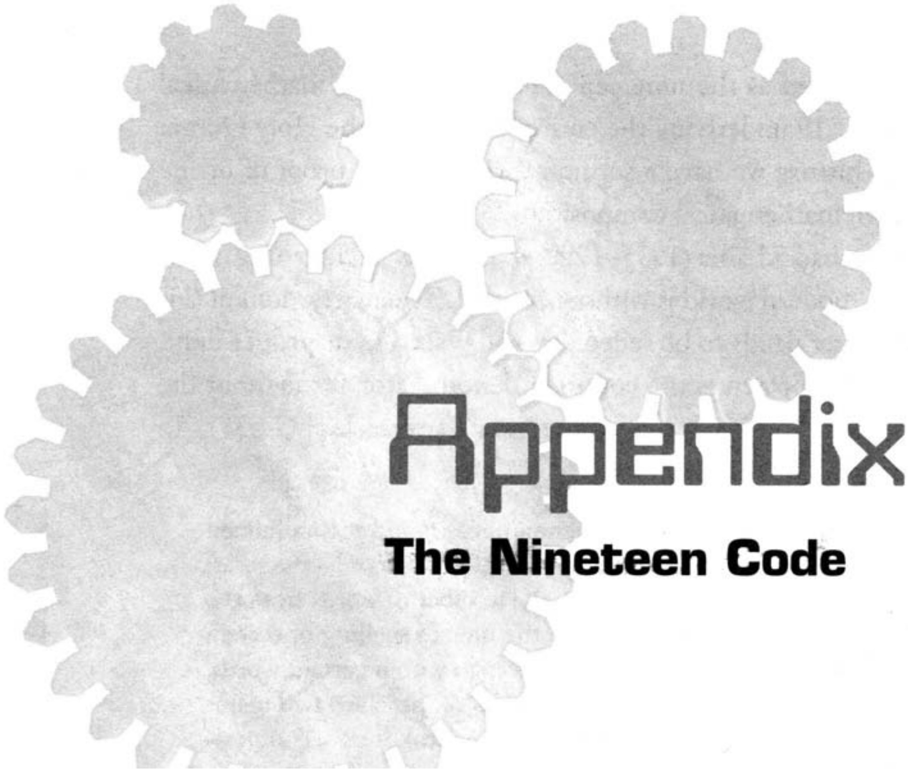

Appendix IV

The Nineteen Code

THE MYSTERY of the motives of the hijackers, who were instrumented by divine will, if not by biospheric necessity, to catalyze the Inevitable Event, lies in their very number: nineteen. While the nineteen is related to the Quranic verse 74:30, "Over it is nineteen," and can be interpreted prophetically, the Law of Time confirms the Quran in demonstrating that the number nineteen possesses a supreme and indeed inviolable value. The Law of Time validates all that is true, and contains within it the mathematical confirmation of all systems of thought as different facets of the same synchronic order.

The Law of Time is based on the mathematical code of the Harmonic Module, itself a function and a demonstration of the 13:20 ratio. The vigesimal mathematics of the 13:20 frequency is known as the 0':'"19code, which is written with a simple dot-bar notation. As such the 0-19 code is a whole number system in which the count is not 1-9 but 1-19, and the positional zero represents a factor of 20 rather than of 10. As the demonstration of a radial mathematical code, the number 19 plays a singular role in the 13:20 matrix. As the largest compound prime number of the 0-19 count, within the context of this count 19 has certain unique properties and values. It may be asked: why 19?

First of all, let us consider the theological dimensions of the number 19 as hinted at by the Quranic verse 74:30. Long an enigma to scholars of the Quran, it was only through the genius of an Egyptian scientist, aided by a computer, that the meaning of the verse was finally revealed as the nineteen code-an intricate mathematical system based on the number 19 underlying the entire structure of the Holy Quran.

"For the first time in history we have a scripture with built-in proof of divine authorship-a superhuman mathematical composition."[1]

The contribution of Rashad Khalifa (1935-1990) to the role and meaning of the Holy Quran for the entire modem world is without parallel. The martyrdom of this genuine seeker and messenger needs to be redeemed and placed in its proper light. According to Dr. Khalifa, "Nineteen is the common denominator throughout the Quran's mathematical system." "This is one of the greatest miracles" (74:35). To summarize in his own words, Dr. Khalifa states:

> The Quran is characterized by a unique phenomenon never found in any human authored book. Every element of the Quran is mathematically composed-the suras, the verses, the words, the number of certain letters, the number of words from the same root, the number and variety of divine names, the unique spelling of certain words, the absence or deliberate alteration of certain letters within certain words, and many other elements of the Quran besides its content. There are two major facets of the Quran's mathematical system: (1) The mathematical literary composition, and (2) The mathematical structure involving the numbers of suras and verses. Because of this comprehensive mathematical coding, the slightest distortion of the Quran's text or physical arrangement is immediately exposed.[2]

Like the Law of Time, which could only have been discovered after the opening of Pacal Votan's tomb, in A.D. 1952, the nineteen code of the Quran could only be discovered and unveiled over a twenty-two-year period between 1968 and 1990. Why? Because only with a modem computer could Dr. Khalifa subject every last word and letter of the Quran, in its Arabic original, to the rigorous analysis necessary to expose its mathematical perfection. As Dr. Khalifa writes in his footnote to 10:20 ("They say, 'How come no miracle came down to him from his Lord?' Say, 'The future belongs to God; so wait, and I am waiting along with you.'"):

> 10.20: In retrospect we see now that the Quran's miracle, indeed "One of the greatest miracles" (74:35), was divinely predestined to be unveiled 14 centuries after Muhammad. In view of the current condition of the traditional Muslims, ifMuhammad had been given this miracle, those Muslims, who are already idolizing Muhammad beside God, would have worshipped him as God incarnate. Additionally, this miracle is obviously designed for the computer age, and to be appreciated by mathematically sophisticated generations.[3]

The phenomenon of the intricate patterning of the nineteen code as witnessed, for example, in the 19 x 142 (2,698) occurrences of the word *God* in the Quran, writes Khalifa, "alone suffices as incontrovertible proof that the Quran is God's message to the world. No human being(s) could have kept track of 2698 occurrences of the word 'God,' and the number of verses where they occur."4 One must bear in mind that the Quran was revealed in a time of primitive simplicity over a twenty-three-year period so that the suras and verses were separated in time and place, while the chronological order of the revelation of the suras is not at all the same as their final format. Moreover, the miracle of the nineteen code is not limited to the appearance of the word God, but is vast, intricate, and comprehensive. Dr. Khalifa lists a set of nineteen "simple facts" that confirm the Quran's mathematical coding. The "simple facts" are those not requiring a calculator or computer to be ascertained. We hereby reproduce these "simple facts," as they are presented in Appendix 1 of *Quran: The Pinal Testament* (all references to letters and words are to the Arabic original):

1. The first verse (1:1), known as *"Basmalah,"* consists of 19 letters.
2. The Quran consists of 114 suras, which is 19 x 6.
3. The total number of verses in the Quran is 6346, or 19 x 334. (6234 numbered verses and 112 unnumbered verses *[Basmalahs].* 6234 +112 = 6346. Note that 6 + 3 + 4 + 6 = 19.
4. The *Basmalah* occurs 114 times, despite its conspicuous absence from Sura 9 (it occurs twice in Sura 27). 114 = 19 x 6.
5. From the missing *Basmalah* of Sura 9 to the extra *Basmalah* of Sura 27 there are precisely 19 suras.
6. It follows that the total of the sura numbers from 9 to 27 (9 +10 + 11 + 12 . . . + 26 + 27) is 342, or 19 x 18.
7. The total (342) also equals the number of words between the two *Basmalahs* of Sura 27, and 342 = 19 x 18.
8. The famous first revelation (96:1-5) consists of 19 words.
9. This 19-worded first revelation consists of 76 letters, 19 x 4.
10. Sura 96, first in the chronological sequence, consists of 19 verses.
11. This first chronological sura is placed atop the last 19 suras.
12. Sura 96 consists of 304 Arabic letters. 304 = 19 x 16.
13. The last revelation (Sura 110) consists of 19 words.
14. The first verse of the last revelation (110:1) consists of 19 letters.

15. 14 different Arabic letters form 14 different sets of "Quranic Initials" (such as A.L.M. of 2:1) and prefix 29 suras. These numbers add up to 14 + 14 + 29 = 57 and 57= 19 x 3.
16. The total of the 29 sura numbers where the Quranic Initials occur is 2 + 3+ 7 + . . . + 50 + 68 = 822, and 822 + 14 (14 sets ofInitials) = 836, or 19 x 44.
17. Between the first initialed sura (Sura 2) and the last initialed sura (Sura 68) there are 38 un-initialed suras. 38= 19 x 2.
18. Between the first and last initialed sura there are 19 sets of alternating "initialed" and "un-initialed" suras.
19. The Quran mentions 30 different numbers: 1,2,3,4, 5,6, 7, 8,9, 10, 11, 12, 19,20,30,40,50,60, 70, 80, 99,100,200,300,1000,2000,3000,5000,50,000, & 100,000. The sum of these numbers is 162,146= 19 x 6534.

Apart from the "simple facts," through use of his computer, a Hewlett Packard HP-IOOO E-series, Dr. Khalifa demonstrated many more complex levels of the coding of the number 19 throughout the Quran. These findings are synthesized in the Appendices to *Quran: The Pinal Testament*, and are also available in his other texts detailing his research, such as *The Miracle of the Quran.* The net effect of the truly monumental research effort exerted by Dr. Khalifa was the promotion of a highly reformed Islam based on "the Quran, the whole Quran, and nothing but the Quran." It is for this reason that Dr. Khalifa's translations of the Quran are referred to as "the Final Scripture," (1980) and the revised translation (1992) as the "Final Testament." It was also Dr. Khalifa's insistence on the Quran as the sole basis ofIslam that earned him his martyrdom in Tucson, Arizona, on January 31, 1990.

NINETEEN IN THE TWENTY-COUNT POSITIONAL SYSTEM

The Law ofTime-T(E) =Art, where T equals the 13:20 timing frequency ratio-upholds the validity of the research of Dr. Khalifa and graphically demonstrates *why* 19 is the mathematical factor that underlies the Holy Quran.

The fact that the Mayan vigesimal system is known as the 0-19 code, and that in this code 19, and not 9, has the supreme value of being the highest number, absolutely enhances the power and significance of Dr. Khalifa's discovery concerning

74:30, "Over it is nineteen." The supremacy of the value of 19, imperceptible if we only follow the decimal 10-count, is fully demonstrated when we examine the 0-19 number code in which the numbers 0-19 are reproduced in a 5 x 4 (= 20) matrix, beginning with 0 and ending with 19. In this number set, which also demonstrates the mathematics of the radial matrix, there are ten pairs of numbers radially arranged, each pair of which adds up to 19:

$0+19=19$
$1+18=19$
$2+17=19$
$3+16=19$
$4+15=19$
$5+14=19$
$6+13=19$
$7+12=19$
$8+11=19$
$9+10=19$

In other words, within the parameters of the mathematical matrix in which nineteen is the integer with the highest value, its value completely informs the radial order of the entire vigesimal system. "Over it is nineteen." This is truly astonishing confirmation of the value accorded to 19 in the Quran. If the Quran is the true word of God, then we must expect that even knowledge unknown to Muhammad or Dr. Khalifa would be contained or coded in the Quran. We are referring, of course, to the Law of Time in its mathematical particulars, beginning with its mathematical base of expression, the 0-19 code.

19 = 7: NINETEEN IN THE 0-19 DOT-BAR NOTATIONAL SYSTEM

Just as the number 19 takes on a supreme value in the vigesimal system inconceivable in the decimal system, so the vigesimal order commands its own notational system. The 0-19 dot-bar notational system substantiates the meaning and power of 19 as the prime mathematical, saturated value unit in the twenty count. The dot-bar notational order is a holographic system distinct from the alphabetic system of Arabic numerals by which we write 1-10.

The distinction between use of the two notational orders is profound. This notational order uses only three notational symbols to build up the holographic context of

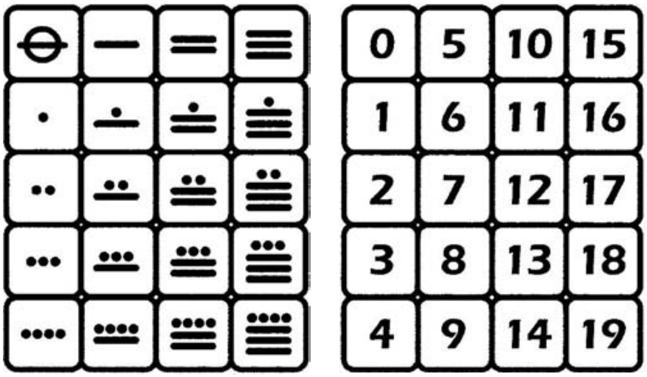

Any radially opposite set of numbers always equals 19, e.g., 4 and 15, 0 and 19, 9 and 10, 2 and 17, 11 and 8.

Nineteen is the highest value and the most saturated notation, being the only one with seven notational symbols—three bars and four dots.

Nineteen in galactic notation.
3 bars (3 x 5 = 15) plus four dots (4) = 19

Dot-bar notation system, 0–19 code

the fourth-dimensional number. One dot represents one unit; one bar represents a five-unit order; and a positional zero completes the system. The vigesimal notational system reveals the entire radial sequence of the 0-19 code to be a highly streamlined and aesthetic pattern. Brief contemplation shows that there are two patterned orders consisting of four vertical sets of five units each, and five horizontal sets of four units each. It is seen that nineteen represents the maximum saturation of both the vertical and horizontal orders, and consists of the maximum number of dots (four) and the maximum number of bars (three). It is also visibly manifest that 19 = 7 notational units, the maximum number of units.

The fact that 19 is represented by seven notational units confirms Sura 1, which has seven verses and the first line of which consists of nineteen Arabic letters. However, from the point of view of the Law of Time, 7 is the key implicit mathematical unit in the absolute timing frequency ratio, 13:20-the difference between 20 and 13 is 7, and the midpoint between 1 and 13 is also 7. While Dr. Khalifa is responsible for determining the dominance of nineteen as the mathematical key underlying the patterning and meaning of the Quran, the Law of Time provides an even more profound lens through which to consider Dr. Khalifa's work, by placing nineteen in a superior (vigesimal) mathematical and theological context.

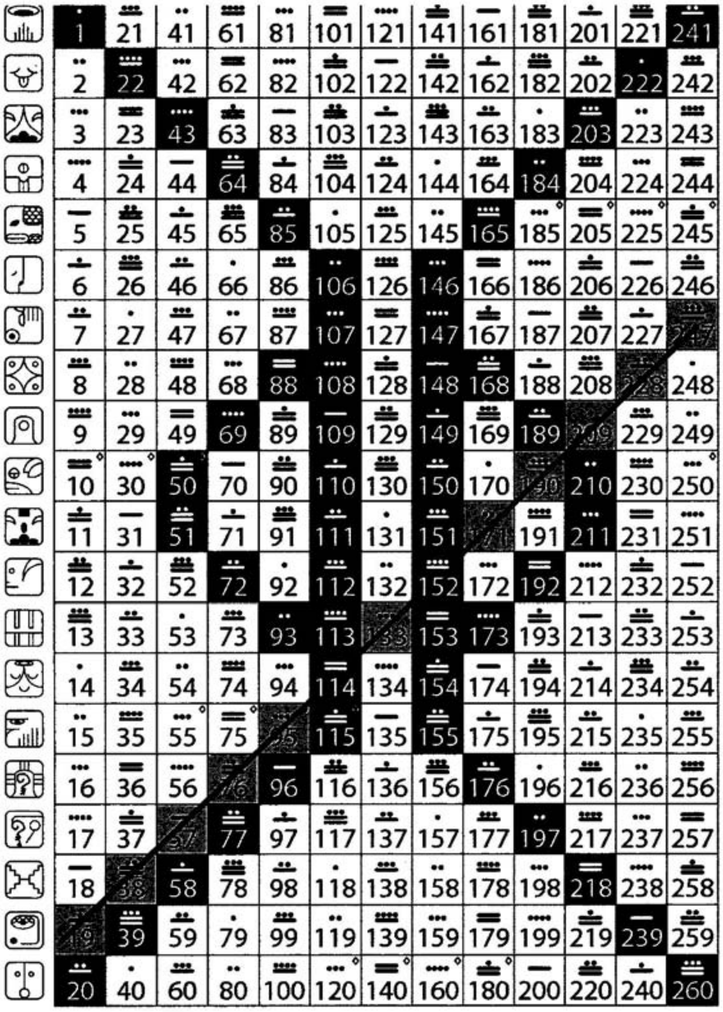

1	21	41	61	81	101	121	141	161	181	201	221	241
2	22	42	62	82	102	122	142	162	182	202	222	242
3	23	43	63	83	103	123	143	163	183	203	223	243
4	24	44	64	84	104	124	144	164	184	204	224	244
5	25	45	65	85	105	125	145	165	185	205	225	245
6	26	46	66	86	106	126	146	166	186	206	226	246
7	27	47	67	87	107	127	147	167	187	207	227	247
8	28	48	68	88	108	128	148	168	188	208	228	248
9	29	49	69	89	109	129	149	169	189	209	229	249
10	30	50	70	90	110	130	150	170	190	210	230	250
11	31	51	71	91	111	131	151	171	191	211	231	251
12	32	52	72	92	112	132	152	172	192	212	232	252
13	33	53	73	93	113	133	153	173	193	213	233	253
14	34	54	74	94	114	134	154	174	194	214	234	254
15	35	55	75	95	115	135	155	175	195	215	235	255
16	36	56	76	96	116	136	156	176	196	216	236	256
17	37	57	77	97	117	137	157	177	197	217	237	257
18	38	58	78	98	118	138	158	178	198	218	238	258
19	39	59	79	99	119	139	159	179	199	219	239	259
20	40	60	80	100	120	140	160	180	200	220	240	260

Nineteen Strand in Harmonic Module

Kin numbers beginning with 19, lower left, are advancing multiples of nineteen, the last being 19 x 13 (247). Actually, the difference between every kin number on this diagonal is a difference of 19.

19 = 260: THE TZOLKIN "OVER IT IS NINETEEN" STRAND

"All is number. God is a number. God is in all."
-DYNAMICS OF TIME, POSTULATE 19.13

The 0-19 code provides the order of 20 in the ratio 13:20. Like the 0-19 radial matrix, the 13:20 ratio is also organized as a radial matrix of 260 (13 x 20) kin or units. As a cycle of 260 days, the 13 x 20 timing matrix is the fourth-dimensional gauge that synchronically unifies the Thirteen Moon calendar. Known by the traditional and ancient Maya as the *Tzolkin* or "sacred count," the 13 x 20 gauge or Harmonic Module is the key to the mathematics of the fourth dimension, or the "Hereafter." Just as the mathematics of the third dimension is decimal-based, so the mathematics of the fourth dimension is vigesimal-based. Hence, the power of 20 as a self-existing proportion (4 x 5 = 5 x 4).

For any frequency to occur two proportions must be coincident. Thirteen, which is a function of the 7 factor, is the proportion that complements the 0-19 vigesimal proportion of 20. Hence the Tzolkin consists of twenty horizontal orders and thirteen vertical orders = 260 kin, which are also arranged in twenty sets of thirteen-kin wavespells, all numbered 1-13 in dot-bar notation. All of the kin in the 260-unit matrix are also individually numbered: kin 1, kin 2 . . . kin 259, kin 260.

Within the 13:20 order of260 kin, 13 sets of20 or 20 sets of 13, only two of the numbers in the vigesimal code 0-19 occur thirteen times: the number 19 and the number 20 (= positional zero). While 20 occurs precisely thirteen times, 19 occurs thirteen times with a remainder of thirteen (13 x 19 = 247 = 260 - 13).

The position and notational order of the thirteen occurrences of the number 19 in the 260-unit Tzolkin create a perfect strand of thirteen units. This 19 x 13 (19 = 260) strand demonstrates the supreme invariability and inviolability of the number 19 in the vigesimal code of fourth-dimensional time.

The purity of the movement of the 13 x 19 strand of the Tzolkin is further augmented by the vigesimal or 20-count manner of writing numbers. Even as transliterated into the Arabic system of writing numbers, it will be seen that in the vigesimal code every multiple of 19 always adds up to 19, a similar power held by all multiples of 9 in the decimal code. In vigesimal notation, the second order units are always multiples of 20 and not 10. Hence 38 = 1.18, where 1 = 20 + 18; or 5.14 = 114, where 5 = 100 + 14.

Finally, while the sequence 1-13 defines the thirteen galactic tones of fourth-dimensional time, the 0-19 matrix is represented by twenty icons and a four-color (4 x 5 = 20) permutation sequence. The combination of color, icon, and tone gives each of the 260 kin a "galactic signature." This completes the description of the Harmonic Modulerrzolkin.

We now present the thirteen multiples of 19 as defined by the 13 x 19 strand running from lower left to upper right in a perfect cross-stitch sequence that reverses the horizontal order of the 0-19 code frequencies. In other words, the first order represented is the order of (1 x 19) 19 (frequency 19, icon Storm), and the last is the

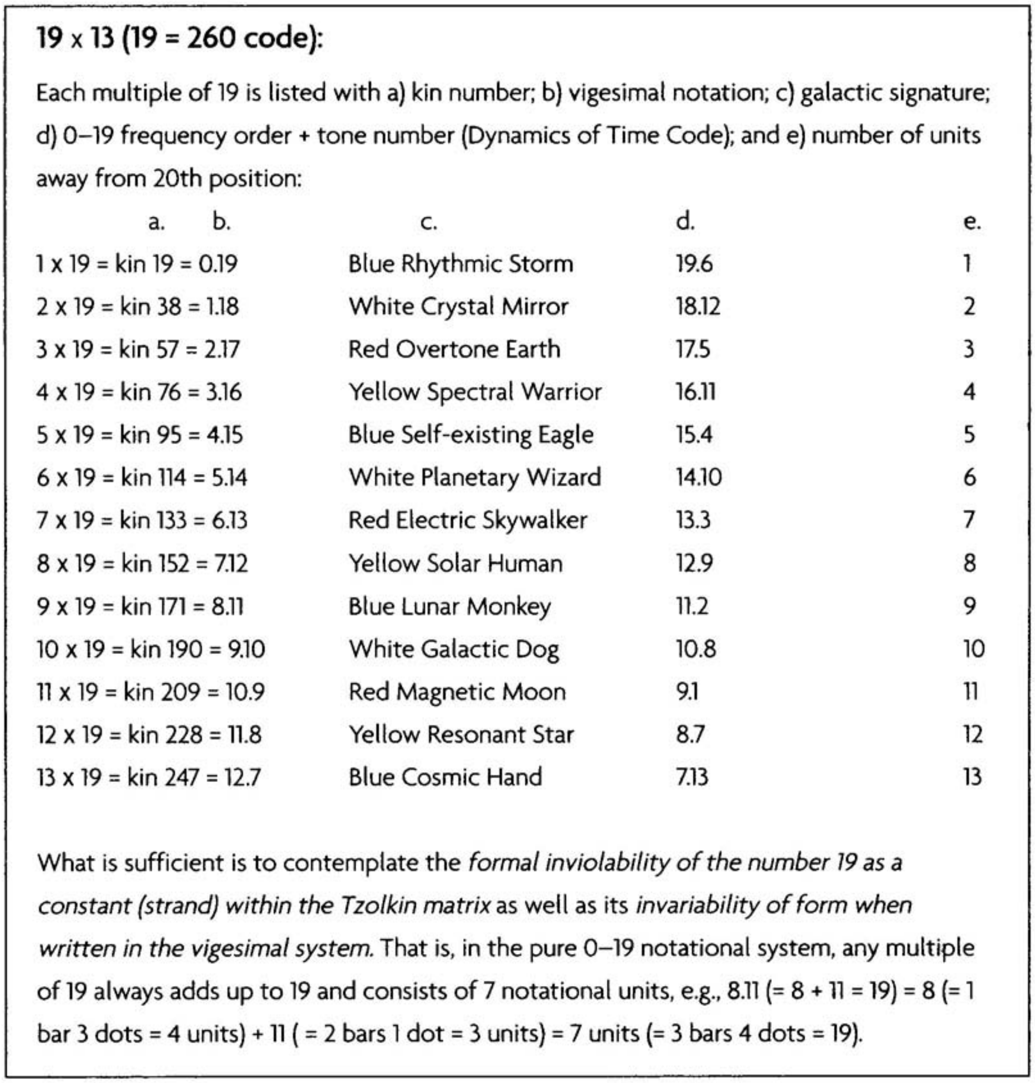

19 x 13 (19 = 260 code):

Each multiple of 19 is listed with a) kin number; b) vigesimal notation; c) galactic signature; d) 0–19 frequency order + tone number (Dynamics of Time Code); and e) number of units away from 20th position:

a. b.	c.	d.	e.
1 x 19 = kin 19 = 0.19	Blue Rhythmic Storm	19.6	1
2 x 19 = kin 38 = 1.18	White Crystal Mirror	18.12	2
3 x 19 = kin 57 = 2.17	Red Overtone Earth	17.5	3
4 x 19 = kin 76 = 3.16	Yellow Spectral Warrior	16.11	4
5 x 19 = kin 95 = 4.15	Blue Self-existing Eagle	15.4	5
6 x 19 = kin 114 = 5.14	White Planetary Wizard	14.10	6
7 x 19 = kin 133 = 6.13	Red Electric Skywalker	13.3	7
8 x 19 = kin 152 = 7.12	Yellow Solar Human	12.9	8
9 x 19 = kin 171 = 8.11	Blue Lunar Monkey	11.2	9
10 x 19 = kin 190 = 9.10	White Galactic Dog	10.8	10
11 x 19 = kin 209 = 10.9	Red Magnetic Moon	9.1	11
12 x 19 = kin 228 = 11.8	Yellow Resonant Star	8.7	12
13 x 19 = kin 247 = 12.7	Blue Cosmic Hand	7.13	13

What is sufficient is to contemplate the *formal inviolability of the number 19 as a constant (strand) within the Tzolkin matrix* as well as its *invariability of form when written in the vigesimal system.* That is, in the pure 0–19 notational system, any multiple of 19 always adds up to 19 and consists of 7 notational units, e.g., 8.11 (= 8 + 11 = 19) = 8 (= 1 bar 3 dots = 4 units) + 11 (= 2 bars 1 dot = 3 units) = 7 units (= 3 bars 4 dots = 19).

order of (13 x 19) 7 (frequency 7, icon Hand), showing again the intimate relation between 19, 13,6, 7, and 20, the key integers of the codes of fourth-dimensional time.

19 = 260, THE SIX SURAS THAT ARE MULTIPLES OF 19

In seeking proof of the Law of Time within the Quranic nineteen code, we need cite but a few examples that can demonstrate from within the nineteen code itself the key mathematical components of the Law of Time. In addition to the number 19, we must look for any of the factors involved in the 13:20 frequency, such as 260, 7, 6-the difference between 19 and 13-and, of course, 13 and 20, or any of the multiples of these integers.

Following the method of Dr. Khalifa, probability analysis of mathematical constituents based on the assumption that certain numbers will be self-verifying, we start our program with the six suras that are multiples of nineteen:

Sura 19	"Maryam"	98 verses
Sura 38	"Saad"	88 verses
Sura 57	"Iron"	29 verses
Sura 76	"The Human"	31 verses
Sura 95	"The Fig"	8 verses
Sura 114	"People"	6 verses
=399, total of sura numbers		=260 verses in six suras that are multiples of nineteen.

= 19.19 vigesimal code

As can be seen, the sum of the six numbers of the suras that are multiples of 19 is exactly 399, or 19 x 21, the maximum factor in the vigesimal 20 x 20 matrix. Expressed in the vigesimal code this number, 399 = 19.19 [19 (x 20 = 380) +19], the only multiple of 19 in the vigesimal code that is written with two 19s, whose sum is precisely twice nineteen, 38, and is written with fourteen notational symbols rather than seven. This unique number, 19.19 (399), verifies "nineteen is over all," by placing 19 over 19 itself. The unique qualities of this most unique of multiples of 19 also indicates why the Quran consists of exactly six times nineteen suras. Six is the difference between the Quranic 19 and 13, the prime number of the Law of Time.

More amazing than the sum of the six suras that are multiples of 19 being 19.19 is the fact that the sum of the verses of these same six suras is 260. This is incontrovertible evidence for the formulation 19 = 260, and substantiates the mathematical patterning of the Quran to include the principle factor of the Law of Time, 260 (= 13 x 20). We are certain that Dr. Khalifa added the verses of these same suras and may have been perplexed at the meaning of the resultant number, 260. Of course, Dr. Khalifa did not know that 260 is the sum factor of the two proportions that create the natural timing frequency, 13:20.

It is also interesting that discovery of the 13:20 timing frequency, December 10, 1989, was made at almost the precise moment as Dr. Khalifa had completed the introduction to *The Pinal Testament,* the summation of his research on the Quranic nineteen code, Ramadan 26, 1409 (December, 1989). This points to the mysterious movement of the Divine Will in establishing the mathematical proof of the Quran. Just as the work of Dr. Khalifa was terminated within a month of his completing the Final Testament, so the Law of Time, which would substantiate and vindicate the Quranic nineteen code, was ordained and propelled into its process of manifestation at exactly the same moment.

Now with the Law of Time the investigation begun by Dr. Khalifa is complete. Just as there are six suras that are multiples of 19 within the Holy Quran, so the key factor of the Law of Time, the number 260, is coded into the sum of the verses of these six suras. The door is opened for entirely new comprehension of the meaning and purpose of the Holy Quran. The Law of Time and the Holy Quran in mutual resonance will establish the proof of the advent and triumph of the Religion of Truth. (9:33,48:28; 61:9)

Most of the foregoing text of Appendix IV is taken from 19 = *260: The Holy Quran and the Law of Time* (1999), an unpublished manuscript by Jose Argiielles.

▼

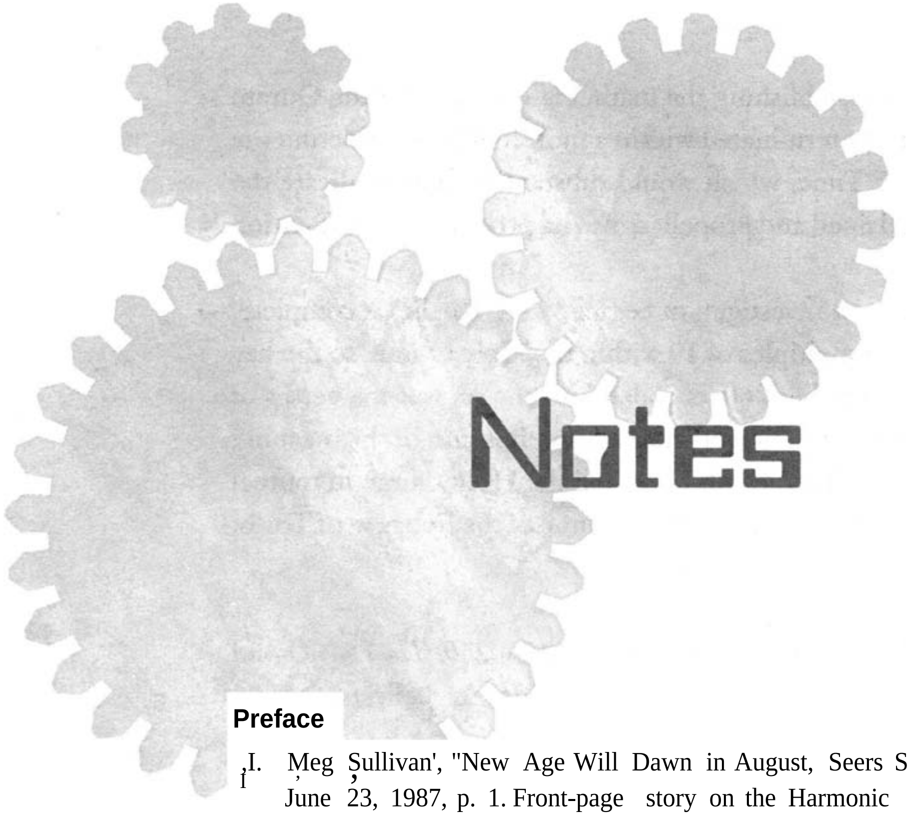

Notes

Preface

I. Meg Sullivan', "New Age Will Dawn in August, Seers Say. . .," *Wall Street Journal,* June 23, 1987, p. 1. Front-page story on the Harmonic Convergence.

2. For the prophetic tradition of the Chilam Balam, see *The Book of Chilam Balam of Chumayel,* edited and translated by Ralph L. Roys (Norman, Okla.: University of Oklahoma Press, 1967); *The Codex Perez and the Book of Chilam Balam of Mani,* edited and translated by Eugene Craine and Reginald Reindorp (Norman, Okla.: University of Oklahoma Press, 1979); and *El Libro de los Libros de Chilam Balam,* edited and translated by Alfredo Barrera Vasquez and Silvia Rendon (Mexico City: Fondo de Cultura Economica, 1948, 1986). The issue of being a "prophetic messenger" is a touchy one for many people. However, as it is written in the Quran, "In truth we are always sending messengers. . ." (44:5). Messengers, as often as not, are self-revealed and confirmed only by God. In my case, I have been able to trace back the prophetic lineage of the Chilam Balam to Pacal Votan, from whom I have received direct transmission. The method of transmission is similar to that of the *terma* tradition of Tibet. Terma, or "hidden treasure," is left by a sage or prophet to be "discovered" at a precise moment much later in time by one designated to do so. See Tulku Thondup Rinpoche, *Hidden Teachings of Tibet: An Explanation of the Terma Tradition of Tibetan Buddhism* (Boston: Wisdom Publications, 1997).

3. This phrase is used often by Dr. Rashad Khalifa (1935-1990), whose groundbreaking translation and research of the Quran through the aid of the computer led to the discovery of the intricate nineteen-based mathematical code that underlies this unique text. His work is summarized in his translation of the Quran with its numerous appendices, *Quran: The Final Testament* (Fremont: Universal Unity, 1992). My own studies on this topic have resulted in two works: *Pacal TfJtIln and Judgement Day* (portland, Oreg.: Planet Art Network, 1999), and 19 = *260: The Holy Quran and the Law of Time,* unpublished, 1999. See also Appendix IV to this book, "The Nineteen Code."

4. See, for example, Samuel P. Huntington's *The Clash of Civilizations and the Remaking of the World Order* (New York: Simon & Schuster, 1996, 1997); Benjamin R. Barber's *Jihad vs. McWorld* (New York: Random House, 1995, 1996); and Yossef Bodansky's *Bin Laden: The Man Who Declared War on America* (New York: Forum, 1999,2001). These writers, of course, all express a pro-globalization perspective.

Chapter 1

1. John ofPatmos, "Book of Revelation," Chapter 18, "The Fall of Babylon," verses 10, 16-17, 19-20, *The Holy Bible, New International Version* (Grand Rapids: Zondervan Bible Publishers, 1989), 1419-20.
2. Jose ArgUelles, *The Call ofPacal Votan: Time is the Fourth Dimension* (Glasgow: Altaea, 1996), 37.
3. Jose ArgUelles, *The Dynamics of Time: 260 Postulates* (Portland, Oreg.: Planet Art Network, 1999),30.
4. Kareng Koshev, "The Great Rainbow Bridge." Altai: unpublished paper, 2001.

Chapter 2

1. 1. Laptev, *The Planet of Reason:A Sociological Study of the Man-Nature Relationship* (Moscow: Progress Publishers, 1973; English translation, 1977),220.
2. V. 1. Vernadsky, "L'etude de la vie et la nouvelle physique" in *Revue generale des sciences,* vol. 41: 695.
3. V. 1. Vernadsky, *The Biosphere* (1926), abridged version (Oracle, Arizona and London: Synergetics Press, 1986), 22.
4. Ibid., 76.
5. Ibid., 77-78, 81.
6. Ibid., 82.
7. Ibid.
8. V. 1. Vernadsky, "Problems in Biogeochemistry II," in *Transactions of the Connecticut Academy of Sciences* (New Haven: Yale University Press, 1944),487-88,498.
9. v.1. Vernadsky, "Problems in Biogeochemistry II," Postulate 18, 511-12.
10. Jose ArgUelles, *The Dynamics of Time: 260 Postulates* (Portland, Oreg.: Pbnet Art Network, 1999), postulates 3.7 and 3.9, 24, 25.
11. V. 1. Vernadsky, "Some Words about the Noosphere" (1944), quoted in Andrey Lapo, *Traces of Bygone Biospheres* (Moscow and London: Mir Publishers and Synergetic Press, 1987),69.
12. Ibid., 71.
13. G. E. Hutchinson, "The Biosphere," *Scientific American,* vol. 53, no. 3 (1970): 223. Quoted in Lapo, *Traces of Bygone Biospheres,* 77.
14. P. Duvigneaud, "Noosphere et l'avenir de la vegetation du globe," *Proceedings of the XII International Botanical Congress, Bruxelles,* 1979, 78. Quoted in Lapo, *Traces of Bygone Biospheres,* 76.

15. Tango Snyder, editor in chief, *The Biosphere Catalogue* (London and Forth Worth: Synergetic Press, 1985), 103.
16. V. 1. Vernadsky, "Scientific Thought as a Planetary Phenomenon," quoted in Lapo, *Traces of Bygone Biospheres, 73-74.*
17. Vernadsky, "Problems in Biogeochemistry II," 511.

Chapter 3

1. Maria Maroushkina, "Is a Path Beyond Technology Possible for the Western Culture?" Moscow: unpublished research paper, 2000.
2. James Glanz, "Physics' Big Puzzle Has Big Question: What Is Time?" *New York Times,* June 17,2001, p. D-1.
3. Ibid.
4. Dogen, "The Time Being," from *Moon in a Dewdrop: Writings of Zen Master Dogen,* edited by Kazuaki Tanahashi (New York: North Point Press, 1985), 76-77.
5. N. A. Kozyrev, "Possibility of Experimental Properties of Time," unpublished paper, 1967, 35. This paper was taken from the Web site www.kozyrev.ru. Virtually unknown in the West, Kozyrev's work has had an especially great influence on the Novosibirsk School of the Russian Academy of Sciences. See, for example, M. M. Lavrentiev, "Physical Theories (Mathematical Models) Adequate to the Reality-A Necessary Condition for Progress of Natural Science of the XXIst century," in *Searchfor Mathematical Laws of the Universe* (Novosibirsk: Sobolev Insititue of Mathematics, 2001) 29-51, and Irina Eganova, "On the Nature of the Stars: N. A. Kozyrev's Approach and His Results," in *Searchfor Mathematical Laws of the Universe,* 215-76; and M. M. Lavrentiev and Irina A. Eganova, "N. A. Kozyrev's Method of Astronomical Observations: Information from True Positions of Stars, Stellar Systems, and Planets," *Instantaneous Action at a Distance in Modern Physics: "Pro" and "Contra,"* edited by Andrew E. Chubykalo, Viv Pope, and Roman Smirnov-Rueda. (New York: Nova Science Publishers, 1999).
6. Pierre Teilhard de Chardin, *The Future of Man* (New York: Harper & Row, 1959), 183.
7. Quoted in V. P. Kaznacheev, "The General Pathology: Consciousness and Physics," Novosibirsk: Russian Academy of Medical Science, unpublished paper, 2000, 28.

Chapter 4

1. Christos Koneatis, *The History of the World-A 6,000 Year Chronicle of Time* (London: Christos Koneatis, 1997), no pagination.
2. Jose and Lloydine Argiielles, *Dreamspell: The Journey of Timeship Earth 2013,* (Hong Kong: Interlink Productions, 1991), 35.
3. Ibid., 5-9.
4. Ibid., 14-15,60,61.
5. V. A. Ponko, private conversation, Novosibirsk, August 18, 2001.
6. *Epic of Gilgamesh,* translated by Caravan of Dreams Theater, quoted in *The Biosphere Catalogue, 116.*

Chapter 5

1. *Webster's Ninth New Collegiate Dictionary,* (Springfield: Merriam Webster, 1987), 1211.
2. See Jose Argiielles, *The Transformative Vision: Reflections on the Nature and History of Human Expression* (Berkeley: Shambhala Publications, 1975),249.
3. Paul Foster Case, *The Tarot: A Key to the WISdom of the Ages* (New York: Macoy, 1947), 158.
4. Jim Hoagland, "The Faithful Come and Wars Erupt Worldwide," *The Oregonian,* December 24,2001, p. C-7.
5. *The Biosphere Catalogue,* 103.

Chapter 6

1. Mahatma Ghandi, "Gandhi on Calendar Reform," *The Journal of Calendar Reform,* vol. I (New York: The World Calendar Association, 1931): 143.
2. All quotes from the Quran used in this text are taken from Rashad Khalifa's translation, *Quran: The Pinal Testament* (Fremont, Calf.: Universal Unity, 1992). All numeration refers to sura number and verse, which are henceforth indicated in parentheses following the quote.
3. Popular quote attributed to Einstein, source unknown.
4. See Broughton Richmond, *Time Measurement and Calendar Construction* (Lei den, Netherlands: J. Brill, 1956), 19-20.
5. Alexander Philip, *Reform of the Calendar* (London: Kegan Paul, Trench, Triibner & Co., Ltd., 1914), 6-8.
6. "Resolution Adopted by the Fourth General Conference on Communications and Transit of the League of Nations," *The Journal of Calendar Reform,* vol I: 162.
7. Argiielles, *The Dynamics of Time,* postulate 6.4,34.
8. *I Ching or Book of Changes,* translated by Richard Wilhelm and Cary F. Baynes (Princeton: Princeton University Press, 1967), 189.
9. Richmond, *Time Measurements and Calendar Construction, 70-71.*
10. David Ewing Duncan, *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year* (New York: Avon Books, 1998),307.
12. Richmond, *Time Measurements and Calendar Construction, 16.*

Chapter 7

1. Pierre Teilhard de Chardin, *The Phenomenon of Man* (New York: Harper & Row, 1955),288
2. For an enlightening look at the subject of *triskaidekaphobia,* see Jonathon Cott, *Thirteen: A Journey into the Number* (New York: Doubleday, 1996).

Chapter 8

1. Star Roots, 13 *Moon Natural Time Journal* (Manson's Landing, British Columbia: 2001), no pagination.

2. Ben 4 and Ben 5, 13 *Moon Natural Time Calendar, Ancient Science for the Art of Now* (Newport, Oreg.: Skytime, 2001), no pagination.
3. Star Roots, 13 *Moon Natural Ttme Journal*, no pagination.
4. "Golden Age of Miracles and Heaven on Earth," *National Examiner*, November 29, 1988, p. 17.
5. *National Geographic Magazine*, publicity blurb, 2001.
6. "Tracking the Hijackers," *The Oregonian*, October 4,2001, p. A-12. See also Appendix Iv, "The Nineteen Code."
7. Muhammad Husayn Haykal, "Islamic Civilization as Depicted in the Qur'an," in *The Life of Muhammad* (Indianapolis: American Trust Publications, 1993),518.
8. Ibid., 518-19.

Chapter 9

1. See, for instance, Dan Oldenburg in the *Washington Post's* full-page article, "The Harmonic Convergence Revisited," May 4, 1990, p. D-5.
2. *Nuestra Civilizacion Frente a la Reforma del Calendario* (Our Civilization Confronted with Calendar Reform) (Buenos Aires: Casa Editora Sudamericana, 1944), 6.
3. F. H. Colson, *The Week: An Essay on the Origin and Development of the Seven-Day Cycle* (Cambridge: Cambridge University Press, 1926), 2.
4. Quoted in Jose Argiielles, *Charles Henry and the Formation of a Psychophysical Aesthetic* (Chicago: University of Chicago Press, 1972), 128.

Appendix I

1. *San Jose Mercury News*, quoted on book cover of David Ewing Duncan's *Calendar, 1998.*
2. Duncan, *Calendar*, 289.
3. Philip, *Reform of the Calendar*, 10 1-2.

Appendix II

1. Charles Muses, *Destiny and Control in Human Systems: Studies in the Interactive Connectedness of Time (Chronotopology)* (Boston- Dordrecht- Lancaster: Kluwer-Nijhoff Publishing, 1984), 20-21.

Appendix IV

1. Rashad Khalifa, *Quran: The Final Testament*, (Fremont, Calif.: Universal Unity, 1992), 375. All further quotes from the Quran and from Rashad Khalifa are from this same work, which in addition to containing the complete Quran in translation also includes thirty-eight Appendices summarizing Dr. Khalifa's research.
2. Ibid., 375.
3. Ibid., 123.
4. Ibid., 375.

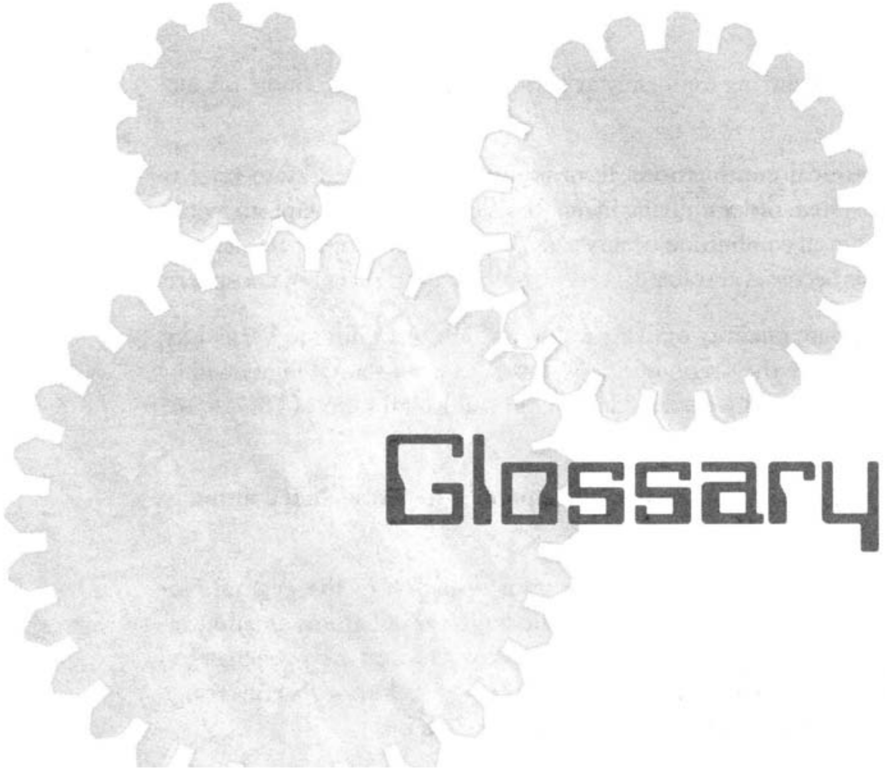

Glossary

THIS GLOSSARY IS constructed to aid the reader in developing a more systematic and technical comprehension of the Biosphere and the Law of Time. Though by no means exhaustive, it is hoped that the terminology used in the text is herein defined in a manner appropriate to cultivating a noospheric point of view.

13:20 timing frequency: Fourth-dimensional timing frequency mathematically expressed as the ratio constant .13:20; universal frequency of synchronization by which the universe is maintained in synchronic order; basis of Law of Time; codified as 260-unit 13:20 matrix of Harmonic Module, coordinates Thirteen Moon/28-Day calendar with synchronic order.

12:60.timing frequency: Artificial, purely third-dimensional timing frequency based on combination of irregular twelve-month calendar and mechanistic sixty-minute hour, adoption of which results in unconscious program within the human mental order, causing the species to deviate altogether from the natural biospheric order, resulting in the technosphere.

astronomical time: Measure of time in third-dimensional physical space without regard to intrinsic harmonics of the synchronic order, usually interpreted as duration and, based solely on matter that is in constant state of entropy, results in standards of measure characterized by infinitesimal.slippages of time and consequent fractional number sequences in need of periodic correction.

baktun: Fifth order of Mayan vigesimal number sequence; in time reckoning equals 144,000 kin (days), or 394.52 years; thirteen per great cycle, measure of human historical era or "great cycle": 3113 B.C.-A.D.2012.

biogenic migration of atoms: Biospheric principle by which evolution over time within finite quantity of living matter (biomass) is accounted for by changes of atomic.,cellular-molecular state and condition; occurring uniformly at periodic transitions accounts for major evolutionary mutations.

biogeochemical, biogeochemical combustion: Term used by V. 1. Vernadsky to refer to biospheric dynamic as whole system order unifying living (bio) and inert (geochemical) matter in a single process; biogeochemical combustion occurs when exponential curve of acceleration of biogenic migration of atoms becomes absolutely vertical, precipitating advent of noosphere.

biomass, biomass constant: Sum quantity of living matter, which, according to Vernadsky, has remained relatively stable since the origin of life on Earth; as a constant, maintained by synchronic timing factor mathematically described as an annual orbital ratio of 73(5) = 365 days per solar orbit of Earth.

biosphere: Sphere or mantle oflife and its geochemical support system distributed throughout Earth's ocean-land surface

biosphere-noosphere transition: Description of the transformation of the general biospheric condition due to accelerated excitation of biogenic migration of atoms creating release of free energy and consequent biospheric destabilization, characterized as biogeochemical combustion, whose resolution is emergence of new evolutionary order, the noosphere; change from historical technosphere to post-historic, post-technological phase.

Bolontiku: Nine Lords of Time; in Mayan cosmology, nocturnal rulers of the underworld or *inframundo;* guardians of the passage of death and resurrection; keepers of time and the knowledge of time as noospherically registered in Earth's orbit.

calendar: Derived from the Latin *calends,* or account book, a day for making payments; generic word used to define systems of timekeeping based on measure of Earth's orbit around the sun, often utilizing synodic measure of the moon as a means of calculating.

Chilam Balam: "Jaguar priest," name of Mayan prophet *(ah bobat)* who lived just before the time of conquest, which he accurately predicted; name given to lineage of Mayan prophets extending from the end of tenth baktun (circa A.D. 830) up until the nineteenth century; name given to series of post-conquest texts in which the various categories of prophecies of this lineage are recorded.

chronomancy, geochronomancy: "Divination by time"; the science and art of fourth-dimensional time considered as a whole system where science is knowledge and art is practice; oracular method of reading the synchronic order; description of incipient "Earth time science" incorporating principles of whole Earth geomancy into time structures of the synchronic order.

chronosphere: Sphere of time that activates noosphere; fourth-dimensional field created by the planet holon in resonance with the rotation of the third-dimensional planet body; coded as movement of four chromatics (biopsychic field) across the planet holon, stitching together gravitational fields (five Earth families) and electromagnetic fields (four color families).

chronotopology: Term coined by Charles Muses to define any geography of time and consisting of an event point of meaning that creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge above a horizon of meaning.

Day Out of Time: The 365th day of the Thirteen Moon/28-Day calendar, no day of the week or month at all (52 x 7 day weeks = 13 moons times 28 days = 364); occurs on Gregorian correlate date July 25, and precedes Thirteen Moon "New Year's" or synchronization date, correlated to July 26.

Dreamspell: Any agreed-on consensus reality; a collective mind-set defining a noospheric epoch, such as "Dreamspell of history"; term given to proofs and tools demonstrating radial and fractal principles of fourth-dimensional mathematics; cosmo-mythic structure of Earth history and its future, otherwise known as *Dreamspell: The Journey of Timeship Earth 2013.*

free energy: Release into biosphere of gases (such as carbon monoxide) and other chemical by-products of industrialization that accelerate biogenic migration of atoms, resulting in biogeochemical combustion; in noosphere refers to release of spectral plasmic-atmospheric phenomena through human telepathic interaction with the electromagnetic field.

fourth dimension: Dimension of time greater than and containing three dimensions of space that it informs and governs.

galactic signature: Combination of one of thirteen galactic tones and twenty solar seals that constitute the harmonic module and that code any given day of the third-dimensional Thirteen Moon calendar.

geocosmic: Holonomic perception expressing unification of Earth whole with the cosmic order.

Gregorian (Julian) calendar: Current world standard calendar consisting of twelve irregularly numbered and obscurely named months coordinated by the 7-day week and intercalary leap year system establishing 28-year cycles; created by the reform of Pope Gregory XIII (1572-1583) in 1582, which "corrected" the Julian calendar (established by Julius Caesar, 45-44 B.C.)by lopping off ten days and refining the intercalary system so that only centuries that are multiples of four are leap years; supreme example of astronomical time measure; principle macro-organizing principle establishing 12:60 artificial timing frequency.

Harmonic Module: 260-unit mathematical matrix of 13:20 frequency, inclusive of a 52-unit binary triplet configuration or set of galactic activation portals; sometimes referred to as Tzolkin, 260-kin sacred count of the Maya; basis of fourth-dimensional vigesimal mathematics of time; coordinated with third-dimensional calendar of Thirteen Moons, establishes basis of 260 galactic signatures.

heliosphere: Sun and its system of planets considered as a whole organism; also known as heliocosm.

holonomics: Of or pertaining to the law governing whole systems.

holonomic consistency and reciprocity: Law by which any whole principle of reality is capable of evoking a like response or reflecting a fractal order at whatever level or through whatever structure.

holon: Any organic whole unit, fractally replicable; basis ofholonomics, law governing whole systems.

human holon: Fourth-dimensional structure of the human being conforming to the 13:20 frequency (twenty digits, thirteen main articulations).

Hunab Ku: "One giver of movement and measure"; in Mayan cosmology, can refer to God as the Supreme Creator.

Inevitable Event: Law of Time definition of the point in synchronic order when the artificial time generated curves of machine, human population, and money converge exponentially in one dramatic, prophecy-fulfilling moment, marking the conclusion oftechnospheric expansion; name given to the events of September 11, 2001 (9-11), epitomized by the collapse of World Trade Center Twin Towers (Fall of Babylon).

Julian Count: Established by Thomas Scaliger, 1583, the year after Gregorian calendar "reform," to establish a linear count of days beginning January 1, 4714 B.C.,some 1,600 years prior to commencement of thirteen baktun Long Count of the Maya; official modern scientific count, basis of linear time paradigm.

katun: Fourth order of vigesimal mathematics = 8,000 kin; in Mayan time count, where the second order vinal is 18 and not 20, a katun = 7200 kin, or 19.7 solar years; twenty katuns equal one baktun of 144,000 kin; 260 katuns per thirteen baktun cycle.

kin: In fourth-dimensional mathematics the base unit of measure, possessing fractal properties; one day and night or turn of Earth on its axis; any human being so designated by assuming the galactic signature of the date of birth.

Law of Time: T(E) =Art, Energy factored by Time equals Art, where T (time) is the fourth-dimensional universal synchronization frequency constant 13:20, E (energy) refers to any phenomenal manifestation, and Art to the fact that the synchronizing frequency of time harmonizes everything into a natural condition of beauty.

Long Count: Mayan count of days (kin) measured by thirteen baktuns of 144,000 days each for a total of 1,872,000 kin, the measure of history from JulianiGregorian correlates August 13, 3113 B.C. to December 21, 2012; still maintained through use of Tzolkin by Quiche Maya, Guatemala.

lunar calendar: System of time measurement using synodical measure of the moon-from new moon to new moon, approximately 29.5 days; twelve lunar months equaling a total of 354 days, eleven days short of the measure of solar orbit; used by many of the Old World civilizations; with the exception ofIslamic lunar calendar, most lunar calendars intercalate a thirteenth moon, seven per every nineteen years.

machine: Capacity for the human to externalize its internal processes as well as motor functions into different means of mechanical reproduction, the sum possibilities of which establish a technological complex requiring constant maintenance, thus subordinating human consciousness to its level; the ultimate manifestation of the 12:60 timing frequency; any industrially produced tool believed to advance the level of human comfort and accomplishment.

noosphere: Earth's mental envelope or field, discontinuous with and above the biosphere; unconscious until the discovery and application of the Law of Time; activated by registration of the human biomass in correct 13:20 timing frequency via universal adoption of Thirteen

Moon calendar; description of transformed state of biosphere, coincident with the end of history; condition of universal telepathy subsequent to collapse of technosphere and application of Law of Time; functions in tandem with programs from Earth's octahedral core.

noospheric chip: Description of human operating in noosphere, holonomically resonant with intrinsic 13:20 codes of synchronic order, capable of interacting with the psi bank to participate in the creation of Earth's rainbow brain.

Noospheric Earth Time (NET): Time measure based on Earth in its orbit in relation to the sun and moon in heliospheric space, rather than from the perspective of human (astronomical time); base unit of measure is one rotation of Earth on its axis or one kin (one night and day), which establishes one NET minute; twenty-eight NET minutes per Moon, or one NET hour, 365 NET minutes, or Thirteen NET Hours plus one NET minute per solar orbit; takes into account that for any NET minute axial magnetic bipolarity of Earth is a changing dynamic reaching two peak points called solstices and two stable points called equinoxes; this four-part division is holonomically registered daily in two peak points, midnight (where day begins and ends) and noon, and two stable points, sunrise and sunset.

occult quartet: Demonstration of radial mathematics of Harmonic Module where the tones of any four kin radially equidistant from each other always equal 28; or the tones of any two kin radially opposite always equal 14.

octahedral core: Iron crystal structure at center of Earth; plasmic storage zone in resonance with the noosphere; in 7:7::7:7, telepathically coordinated by four time atoms (= one master time molecule).

omega point: In the work of Pierre Teilhard de Chardin, defines the climactic convergent point (harmonic convergence) of human evolution as the emergence of the hyper-personal; defined by the Law of Time as point of full superconscious establishment of universal telepathy as the norm of the human mental condition; dawn of omega point: winter-northern-summer-southern solstice, 2012, completion of omega point dawn, Yellow Galactic Seed (July 26,2013), otherwise known as Dreamspell of galactic culture.

oracle: For anyone of 260 kin of the Harmonic Module, a cruciform five-part structure demonstrating a uniform set of mathematical relationships in relation to the (0-19) code number of the daily solar seal (destiny kin) informing any day of the Thirteen *Moonl2* 8-Day calendar; consists of center, code number of destiny kin; right analog kin-always 19 when added to the number of destiny kin; left, antipode kin-always a difference of 10 from the destiny kin; below, occult kin-when added to destiny kin always equals 21; and above, guide, always the same color as destiny kin.

Oxlahuntiku: The Thirteen Lords of Time that govern the power of the cosmic order of time that is based on circulatory power of prime number 13 (as opposed to 12, which is static and non-circulatory); coded as thirteen 28-day moons that govern terrestrial solar orbit; basis of prophecy cycle of Quetzalcoatl, Thirteen Heavens (Oxlahuntiku) and Nine Hells (Bolontiku); in Telektonon prophecy, the thirteen-year cycle-2 000-2 013-in which the first four years, 2000-2004, establish self-existing power of time, and last nine years, 2004-2013 are coded by the Bolontiku, the Nine Lords of Time.

Pacal Votan: Mayan prophet (A.D.603--683), chief technician of Mayan synchronic time science; eleventh ruler of the dynasty of Palenque (Nah Chan), Chiapas, Mexico; assumed power in A.D.615, but key rulership occurred in critical 52-year cycle, 631-683 (seventy-third such cycle since commencement of thirteen baktun count); best known for the carved sarcophagus lid of his elaborately concealed tomb discovered in 1952, 1,260 years after its dedication in A.D.692, which was also 1,320 years before the closing of the cycle, 2012.

Pax Cultura, Pax Biospherica: Principle of new human social order following establishment of Thirteen Moon/28-Day calendar and inaugurating the advent of the noosphere (2004) in which nation-state is replaced by biospheric regionalism governed by a Planetary Biospheric Assembly; due to perfect harmony of new world standard calendar of thirteen moons/28 days, harmony, art, and aesthetic values elevate culture as the new operating norm of the human reintegrated within the biosphere, hence *Pax Cultura, Pax Biospherica.*

planet holon: Twenty-part icosahedral noospheric matrix superimposed on Earth's surface; corresponding to 0-19 code of twenty solar seals, creates three orders analogous to three planetary fields of resonance; gravitational field-horizontal order of five Earth families of four seals each; electromagnetic field-four diagonal bands, from lower left to upper right, of the four color families (red, white, blue, yellow) of five seals each; and biopsychic field-the four chromatics of five seals each moving diagonally from upper left to lower right, which coordinate the daily sequence of the motion of the chronosphere.

psi bank: Regulating mechanism of the noosphere consisting of four bipolar plates, each plate in mirror symmetry containing 520 (260 x 2) psi chrono units, 2,080 in all, which are the complete registrations of the universal fourth-dimensional 13:20 timing frequency governing all of the various evolutionary stages of the terrestrial biogeocosm; engine of the chronosphere, located between the two radiation belts 2,000 and 11,000 miles above Earth's surface.

psi chrono unit: Time-bearing information unit; base unit of the psi bank; codes and contains all information for anyone of the 260 galactic signatures of the 13:20 matrix or Harmonic Module; psycho activated through coordination with Thirteen Moon/Telektonon known as Rinri Project.

Psychozoic Era: Term coined by V. 1. Vernadsky defining the next geological era; meaning the era of the spiritualization of life, it is coincident with the noosphere in its full emergence after the omega point has been reached, 2012-2013.

Quetzalcoatl: Mexican prophet (A.D.947-999) who left the legacy of prophecy of Thirteen Heavens and Nine Hells, basis of Harmonic Convergence (1987) and commencement of the "Time of Prophecy," 1987-2013.

Quran, Holy Quran: Literally "recitation;" last revelation and moral-spiritual criterion for historic humanity; received over a twenty-three-year period by the Prophet Muhammad (A.D.570-632); proven by Dr. Rashad Khalifa (1935-1990) to be a radically nonlinear text rigorously constructed of a nineteen-based mathematical code.

radion, radial plasmas: Telepathic fluid or lubricant released through telepathic interaction with the seven primary plasmas, or electrically charged particles, hence radial plasmas; the names of the seven plasmas replace the names of the week in the Thirteen Moon/28-Day calendar; activated in the 7:7::7:7 Telektonon Revelation practice.

radiosonics: Name given to the new science based on synchronicity rather than causality and in which, as a consequence, synaesthetic experiences and syntropic values are normalized.

rainbow brain: Description of activated noosphere, where human telepathic interaction with the electromagnetic field triggers the release of the circumpolar rainbow bridge-Earth's day-night alternator system.

Rinri Project: Law of time practice for coordinating the opening of the psi bank, psi chrono unit per psi chrono unit, on a daily basis over a four-year cycle by means of the Thirteen MoontTelektonon.

synchronic order: Order of fourth-dimensional time, in contrast to third-dimensional astronomical time, that synchronizes all aspects of third-dimensional physical plane reality including its timekeeping systems; entirely new domain of human mental-spiritual experience coincident with the dawning of noosphere; characterized by absolute harmonic perfection, described by vigesimal 0-19 code, as distinct from decimal, hex, or duodecimal mathematics that govern third-dimensional timing codes.

synchronometer: Measure of synchronization or synchronicity; precise Law of Time definition of means and purpose of what is generically referred to as the "calendar."

technosphere: Global sphere of technology; artificial stage between biosphere and noosphere, representing the climax-biogechemical combustion-of acceleration of biogenic migration of atoms precipitating biospheric crisis as a prelude to the advent of noosphere; characterized by exponential curves of human population, machine, and money at expense of biospheric integrity and stability; defined as the specific fifty-six-year cycle (two Gregorian 28-year cycles) 1945-2001, in which war and terrorism are the governing features; complete manifestation of the 12:60 order of time; concluded on September **11,2001,** the Inevitable Event.

Telektonon: "Earth Spirit Speaking Tube," the name of the psychoduct leading from the tomb of Pacal Votan beneath the Temple of the Inscriptions to the temple floor atop the pyramid, the discovery of which, in 1949, led to the excavation and revelation of the long-concealed tomb in 1952; name given to the prophecy ofPacal Votan, decoded in 1993-1994, prophesying the Thirteen Moon/28-Day calendar as the only means to keep the human race from destroying the biosphere and itself; definition of the fourth-dimensional structure of the heliosphere in which consciousness is a function of the planetary orbits; basis of a board "game" coded to synchronic order via Thirteen Moon calendar for the awakening of telepathy; basis of the Cube of the Law, days 7-22, containing numerous principles of the Law of Time as a daily practice; basis of other practices for establishing synchronic order as principles of a new consciousness, including *20 Tablets of the Law of Time* and the 7:7::7:7.

third dimension: Space; physical plane of reality coordinated and contained by time, fourth-dimensional order of reality.

Thirteen Moon/28-Day calendar: Pragmatic application of the Law of Time, perpetual harmonic perfection of which is intended to correct for disorder of human consciousness resulting from adherence to irregular, astronomical timing standards and mechanistic timing devices; coordinated with Harmonic Module and Telektonon establishes means for conscious human participation in synchronic order; 28-day measure based on female menstrual cycle, also corresponds to apsidal measure of moon from the point at which the polar axis is tipped farthest from Earth; prevalent in pre-history, all but eliminated by historical consciousness that made 13 taboo; perfect vehicle for establishing noosphere as daily consciousness in post-technospheric post-history.

time: Fourth dimension; universal factor coordinating space to manifest synchronically; traveling faster than speed of light, accounts for instantaneity of information transmission (telepathy) from distant points; expressed as formulation which states that the velocity of time is instantaneously infinite; fourth-dimensional structures that are radial and fractal in nature, third-dimensional cycles being manifestations thereof; expressed as T(E) = Art, accounts for universal harmony of the cosmos.

tun: Third positional order of vigesimal mathematics where one unit = 400 kin; in Mayan time count, because of second positional order, vinal, being 18 instead of 20, one tun = 360 kin; basis of measure ofkatuns and baktuns in Long Count, where 5,200 tuns = 260 katuns = 13 baktuns, or 5,125 solar years.

Tzolkin: "Sacred Count," 260-day calendar of the Maya, basis of Mayan time knowledge and the Long Count; temporal manifestation of Harmonic Module or universal 13:20 timing matrix.

UR: "Universal Religion," "Universal Recollection;" description of human spirituality in the noosphere or Psychozoic Era.

vigesimal, 0-19 code: Twenty-count based mathematics utilizing zero for positional orders that advance in exponential binary fashion; also known as 0-19 dot-bar code notational system, where dots equal single units, bars equal five units, plus positional zero; basis of mathematics of fourth-dimensional order of time, distinct from mathematical systems of third-dimensional space.

Wave Harmonic of history: Fourth-dimensional description of thirteen baktun cycle (3113 B.C.-A.D. 2012) identified precisely with 260-unit Harmonic Module, where each of thirteen baktuns contains twenty katuns (13 x 20); basis of creating synchronic order out of entropic disorder of third-dimensional historical time.

wavespell: Thirteen-unit cosmological form-constant for coordinating fourth-dimensional timing cycles; consists of two gates, Magnetic first position and Cosmic thirteenth position; two towers, fifth and ninth positions; and three sets of chambers, nine chambers in all; when coordinating Thirteen Moon/28-Day calendar, known as Planetary Service Wavespell.

Wizard's Count: Name of count to establish synchronic order as manifestation of the Law of Time in order to transit from 12:60 technosphere to 13:20 noosphere; based on prophetically modified year-bearer count of the Chilam Balam tradition; correlated by the Dreamspell to July 26, 1987, White Galactic Wizard, every four years there are four year bearers: Yellow Seed, Red Moon, White Wizard, and Blue Storm (one four-year Seed-Storm Year bundle) that, combined with thirteen galactic tones creates 52-year solar-galactic cycle, or in NET time, one Earth Year, also known as a Wizard's Year.

▼

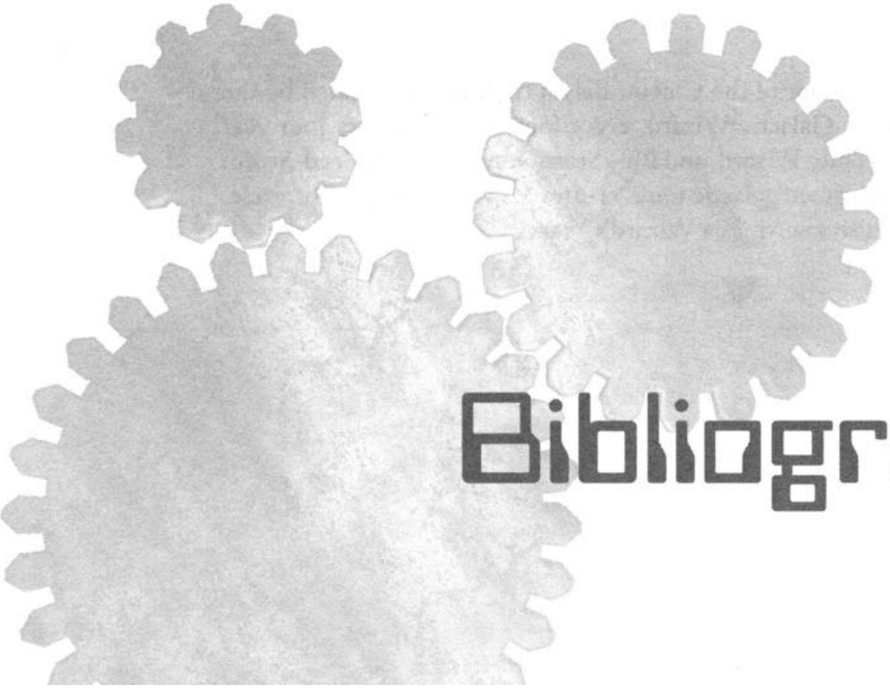

Bibliography

I. About the Biosphere, Technosphere, and Noosphere

,Argiielles, Jose: *EJ:rh Ascending: An Illustrated Treatise on the Law Governing Whole Systems.* \, B~.~ldh: Shambhala Publications, 1984. Second and third editions, Santa Fe: Bear & Co., 1988 and 1996.

-. "Peace is the Culture of the Biosphere: Statement and Reflections." Brightwood, Oregon: Foundation for the Law of Time, unpublished paper, 1996.

Argiielles, Jose and Lloydine Argiielles. *A Complete Guide to the First Planetary Congressof BiosphericRights.* Portland, Oreg.: Planet Art Network, 1996,2002.

-. *The World Thirteen Moon Calendar Change PeacePlan.* Portland: Planet Art Network. 1999.

Coats, Callum. *Living Energies: An Exposition of ConceptsRelated to the Theories of Viktor Schauberger.* Bath, England: Gateway Books, 1996.

"Global Consciousness Project: Registering Coherence and Resonance in the World," http://noosphere. princeton.edu.com.

Kaznacheev, V. P. "The General Pathology: Consciousness and Physics." Novosibirsk: Russian Academy of Medical Science, 2000.

Lapo, Andrey V. *Tracesof Bygone Biospheres.*Moscow and London: Mir Publishers and Synergetic Press, 1987.

Laptev, 1. *The Planet of Reason:A SociologicalStudy of Man-Nature Relationship.* Moscow: Progress Publishers, 1973. English Translation, 1977.

Ponyko, V. A. *Vednie V Systemy "Ekoprognos":Metodologiya PrognosirovaniyaAnomalii Ekosperi* (Ecoprognosis: Methodology for Making Prognosis of Anomolous Phenomena in the Ecosphere). Moscow: Institute Ustoychivogo Rasveteya, 2000.

Smil, Vaclav. *Energies: An Illustrated Guide to the Biosphereand Civilization.* Cambridge, Mass. and London: MIT Press, 1999.

Snyder, Tango, editor in chief, and the Institute of Ecotechnics. *The Biosphere Catalogue.* London and Forth Worth: Synergetic Press, 1985.

Teilhard de Chardin, Pierre. *The Future of Man.* New York: Harper & Row, 1964.

-. *The Phenomenon of Man.* New York: Harper & Row, 1959.

Vernadsky, V. 1. *The Biosphere.* Abridged version based on the French edition of 1929. Oracle, Arizona and London: Synergetic Press, 1986.

-. *The Biosphere: Complete Annotated Edition.* Revised and annotated by Mark A. S. McMenamin and Far West Institute. New York: Copernicus/Springer Verlag, 1998.

-. "Problems in Biogeochemistry 11." In *Transactions of the Connecticut Academy of Sciences,* vol. 35 (1944): 483-517.

II. About the Law and Science of Time

Acheles, Elisabeth. *Of Time and the Calendar.* New York: Hermitage House, 1955.

-. *The World Calendar Reform Since 1930.* New York: G. P. Putnam's Sons, 1937.

Alcala, Luis Alvarez. *El Calendario Azteca.* Edited by B. Costa-Amico Mexico City: 1970.

Argiielles, Jose. "Calendar Reform and the Future of Civilization: Preparatory Reflections for the World Summit on Peace and Time," San Jose, Costa Rica: University for Peace, unpublished paper, 1999.

-. *The Call of Pacal Votan: Time is the Fourth Dimension.* Glasgow: Altea, 1996.

-. "Campaign for the New Time: Science Demonstrates a New Time of Peace on Earth," Brightwood, Oregon: Foundation for the Law of Time, unpublished position paper, 2000.

-. *Earth Ascending: An Illustrated Treatise on the Law Governing Whole Systems.* Santa Fe: Bear & Co, 1984, 1988, 1996

-. *The Dynamics of Time: 260 Postulates.* Portland, Oreg.: Planet Art Network, 1999.

-. *The Mayan Factor: Path Beyond Technology.* Santa Fe: Bear & Co., 1987. Second edition, 1996.

Argiielles, Jose and Lloydine Argiielles. *Discovering the New Time.* Tokyo: Fuunsha, 1997.

-. *Thirteen Moons in Motion.* Portland, Oreg.: Planet Art Network, 1999.

Barbour, Julian. *The End of Time: The Next Revolution in Physics.* Oxford: Oxford University Press, 1999.

Ben 4 and Ben 5. 13 *Moon Natural Time Calendar: Ancient Science for the Art of Now.* Newport, Oreg.: Skytime, 2001.

Colson, F. H. *The Week: An Essay on the Origin and Development of the Seven-Day Cycle.* Cambridge: Cambridge University Press, 1926.

Cott, Jonathon. *Thirteen: A Journey into the Number.* New York: Doubleday, 1996.

Coveney, Peter and Richard Highfield. *The Arrow of Time: A Voyage Through Science to Solve Time's Greatest Mystery.* New York: Fawcett-Columbine, 1990.

Davies, Paul. *About TIme: Einstein's Unfinished Revolution.* New York: Simon & Schuster, 1995.

Deutsch, David. *The Fabric of Reality.* London: Penguin Books, 1997.

Dogen. "The Time-Being," from *Moon in a Dewdrop: Writings of Zen Master Dogen.* Edited by Kazuaki Tanahashi, New York: North Point Press, 1985. 76-83.

Duncan, David Ewing. *Calendar: Humanity's Epic Struggle to Determine a True and Accurate Year.* New York: Avonbooks, 1998.

Dunne, J. w. *An Experiment with Time.* Charlottesville, Va.: Hampton Roads, 2001.

Eganova, Irina. "On the Nature of the Stars: N. A. Kozyrev's Approach and His Results." In *Searchfor Mathematical Laws of the Universe,* edited by M. M. Lavrentiev, 215-276. Novosibirsk: Sobolev Institute of Mathematics, 2001.

Fraser, J. T. *Time, the Familiar Stranger.* Redmond, Wash.: Tempus-Microsoft, 1987.

Fraser, J. T., ed. *The Voicesof Time: A CooperativeSurvey ofMan's Views of Time as Expressed by the Sciencesand the Humanities.* New York: George Brazilier, 1966.

Heath, Robin. *Sun, Moon & Earth.* Powys, Wales: Wooden Books, 1999.

International Society for the Study of Time (ISSO), www.studyoftime.org/index.html.

The Journal of Calendar Reform, vols. I and II. New York: The World Calendar Association, 1931,1932.

Koneatis, Christos. *The History of the World-A 6,000 Year Chronicle of Time.* London: Christos Koneatis, 1997.

Koshev, Kareng. "The Great Rainbow Bridge," Altai: unpublished paper, 2001.

Kozyrev, N. A. "Possibility of Experimental Study of Properties of Time," Novosibirsk: Russian Academy of Sciences, unpublished paper, September, 1967.

-. www.kozyrev.ru.

Landes, David S. *Revolution in Time: Clocksand the Making of the Modern World.* Cambridge, Mass.: Belknap Press, 1983.

Lavrentiev, M. M. "Interpretation of Astronomical Data Connected with N. A. Kozyrev's Discovery." In *Memoranda of Differential Equations of Mathematical Physics,* vol. 12, (1997): 135-41.

-. "Physical Theories (Mathematical Models) Adequate to the Reality: A Necessary Condition for Progress of Natural Science of the XXIst Century." In *Searchfor Mathematical Laws of the Universe,* 29-51. Novosibirsk: Sobolev Institute of Mathematics,2001.

Lavrentiev, M. M., and Irina A. Eganova. "N. A. Kozyrev's Method of Astronomical Observations: Information from True Positions of Stars, Stellar Systems, and Planets." In *Instantaneous Action at a Distance in Modern Physics:"Pro" and "Contra,"* edited by Andrew E. Chubykalo, Viv Pope, and Roman Smirnov-Rueda. New York: Nova Science Publishers, 1999.

Malmstrom, Vincent H. *Cycles of the Sun, Mysteries of the Moon: The Calendar in Mesoamerican Civilization.* Austin: University of Texas Press, 1997.

Man and Time: Papers from the Eranos Yearbook 3. Bollingen Series XXX. Edited by Joseph Campbell. New York: Pantheon Books, 1957.

Maroushkina, Maria. "Is a Path Beyond Technology Possible for Western Culture?" Moscow: unpublished paper, 2000.

Muses, Charles. *Destiny and Control in Human Systems: Studies in the Interactive Connectedness of Time (Chronotopology).* Boston- Dordrecht -Lancaster: Kluwer- Nijhoff Publishing, 1984.

Nuestra Civilizacion Frente a la Reforma del Calendario (Our Civilization Confronted with Calendar Reform). Buenos Aires: Casa Editora Sudamericana, 1944.

Penrose, Roger. *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford: Oxford University Press, 1994.

Philip, Alexander. *The Reform of the Calendar.* London: Kegan Paul, 1914.

Philosophy of Time. www.dna2z.com/projects/time.html.

Price, Huw. *Time's Arrow and Archimedes Point.* Oxford: Oxford University Press, 1996.

Reference Library Books. *The World of Time and Timepieces.* London: B. P. C. Publishing, 1972.

Richmond, Broughton. *Time Measurement and Calendar Construction.* Leiden: J. Brill, 1956.

Seabury, Rev. Samuel. *The Theory and Use of the Church Calendar in the Measurement and Distribution of Time.* New York: Pott, Young & Co., 1872.

Spinden, HerbertJ. *Maya Dates and What They Reveal.* Brooklyn: Museum of the Brooklyn Institute of Arts and Sciences, 1930.

Star Roots. 13 *Moon Natural Time Journal.* Manson's Landing, British Columbia, 2001.

The Study of Time: Proceedings of the First Conference of the International Society for the Study of Time, Oberwolfach, Black Forest, West Germany. Edited by J. T. Fraser,. F. C. Haber and G. H. Miiller. Berlin-Heidelberg-New York: Springer Verlag, 1972.

The Study of Time II: Proceedings of the Second Conference of the International Society for the Study of Time, Lake Yamanaka, Japan. Edited by J. T. Fraser, F. C. Haber and G. H. Miiller. Berlin-Heidelberg-New York: Springer Verlag, 1975.

Vanderkam,James C. *Calendars in the Dead Sea Scrolls: Measuring Time.* London and New York: Routledge, 1998.

Votan, Valum. *Principia Mathematica of the Fourth Dimension: Introduction and Annotated Bibliography.* Portland, Oreg.: Planet Art Network, 1999.

Whitrow, G. J. *The Natural Philosophy of Time.* New York: Clarendon Press, 1980.

Wolf, Fred Allen. *Parallel Universes: The Search for Other Worlds.* New York: Touchstone, 1988.

III.About the Inevitable Event: Prophecy, History, and Politics

Ahmad, Mirza Ghulam of Qadian. *The Philosophyof the Teachingsof Islam.* London: London Mosque, 1979.

America at War. New York: Profiles Presents, 2001.

Argiielles, Jose. *Pacal Votan and Judgement Day*.Portland, Oreg.: Planet Art Network, 1999.

Barber, Benjamin R. *Jihad vs. McWorld.* New York: Random House, 1995, 1996.

Bodansky, Yossef. *Bin Laden: The Man Who Declared War on America.* New York: Forum, 1999,2001.

The Book of Chilam Balam of Chumayel. Edited and translated by Ralph L. Roys. Norman, Okla.: University of Oklahoma Press, 1967.

Case, Paul Foster. *The Tarot: A Key to the Wisdom of the Ages.* New York: Macoy, 1947.

The Codex Perez and the Book of Chilam Balam ofMani. Edited and translated by Eugene Craine and Reginald Reindorp. Norman, Okla.: University of Oklahoma Press, 1979.

"Golden Age of Miracles and Heaven on Earth," *National Examiner,* November 29, 1988, p. 17.

Haykal, Muhammad Husayn. "Islamic Civilization as Depicted in the Qur'an." In *Lift of Muhammad.* Translated from the 8th edition by Isma'il Ragi A. al-Faruqi, 517-55. Indianapolis: American Trust, 1993.

Hoagland, Jim. "The Faithful Come and Wars Erupt Worldwide," *The Oregonian,* December 24,2001, p. C-7.

Huntington, Samuel P. *The Clash of Civilizations and the Remaking of the World Order.* New York: Simon & Schuster, 1996, 1997.

I Ching or Book of Changes. Translated by Richard Wilhelm and Cary F. Baynes. Princeton: Princeton University Press, 1967.

Izquierdo, Ana Luisa. *Alberto Ruz Lhuillier Frente al Pasado de los Mayas.* Mexico City: Secretaria de Educacion Publica, 1987.

Jenkins, John Major. *Maya Cosmogenesis 2012: The True Meaning of the Maya Calendar End-Date.* Santa Fe: Bear & Co., 1998.

John ofPatmos, "Book of Revelation," Chapter 18, "The Fall of Babylon." *The Holy Bible, New International Version.* Grand Rapids: Zondervan Bible Publishers, 1989.

Katchongva, Dan. *From the Beginnings of Lift to the Day of Purification: Teachings, History & Prophecies of the Hopi People.* Los Angeles: Committee for Traditional Indian Land and Life, 1972.

Khalifa, Rashad. *Quran: The Final Testament.* Fremont, Calif.: Universal Unity, 1992.

Kyipchakoba, P. v., ed. *Altai U Drevnue Maya* (Altai and the Ancient Maya). Gorno-Altaiski, 2001.

Landa, Fray Diego de. *Yucatan Before and After the Conquest.* Translated by William Gates. New York: Dover, 1978.

El Libro de los Libros de Chilam Balam. Edited and translated by Alfredo Barrera Vasquez and Silvia Rendon. Mexico: Fondo de Cultura Economica, 1948, 1986.

Mahmud, Sayyid Fayyaz. *A Short History of Islam.* Oxford: Oxford University Press, 1960.

Mano, Galactica Azul. "La Profecia Maya y las Torres Gemeles" (Mayan Prophecy and the Twin Towers). Manzaneles, Colombia: unpublished research paper, October, 2001.

Oldenburg, Dan. "The Harmonic Convergence Revisited," *Washington Post,* May 4, 1990, p. D-5.

The Roerich Pact and the Banner of Peace. New York: The Roerich Pact and Banner of Peace Committee, 1947.

Ruz, Alberto. *Palenque, Official Guide.* Mexico City: Instituto Nacional de Antropologia e Historia, 1978.

September 11,2001: The Day that Changed America. New York: God Bless America, 2001.

September 11,2001. New York: Time-Warner, 2001

Shearer, Tony. *Beneath the Moon and Under the Sun: A Reappraisal of the Sacred Calendar and the Prophecies of Ancient Mexico.* Albuquerque: Sun Books, 1975.

"Tracking the Hijackers," *The Oregonian,* October 4,2001, p. A-12.

Tulku Thondup Rinpoche, *Hidden Teachings of Tibet: An Explanation of the Terma Tradition of Tibetan Buddhism.* Boston: Wisdom Publications, 1986, 1997.

▼

Books of Related Interest

THE MAYAN FACTOR
Path Beyond Technology
by José Argüelles

EARTH ASCENDING
An Illustrated Treatise on Law Governing Whole Systems
by José Arguelles

SURFERS OF THE ZUVUVA
Tales of Interdimensional Travel
by José Arguelles

MAYA COSMOGENESIS 2012
The True Meaning of the Maya Calendar End-Date
by John Major Jenkins

GALACTIC ALIGNMENT
The Transformation of Consciousness According to Mayan, Egyptian, and Vedic Traditions
by John Major Jenkins

THE MAYAN ORACLE
Return Path to the Stars
by Ariel Spilsbury and Michael Bryner

SECRETS OF MAYAN SCIENCE/RELIGION
by Hunbatz Men

THE LOST REALMS
by Zecharia Sitchin

Inner Traditions. Bear & Company
P.O. Box 388
Rochester, VT 05767
1-800-246-8648
www.InnerTraditions.com

Or contact your local bookseller

www.foundationforthelawoftime.org

www.tortuga.com

The Inevitable Event and the End of History

The end of 12:60 Gregorian Time
Pope John Paul II after the Inevitable Event

Planetary 9-11, 2001. First Year, third Millenium

"Woe! Woe, O Great City, O Babylon
in one hour your doom has come!"
—Revelations, 18:10

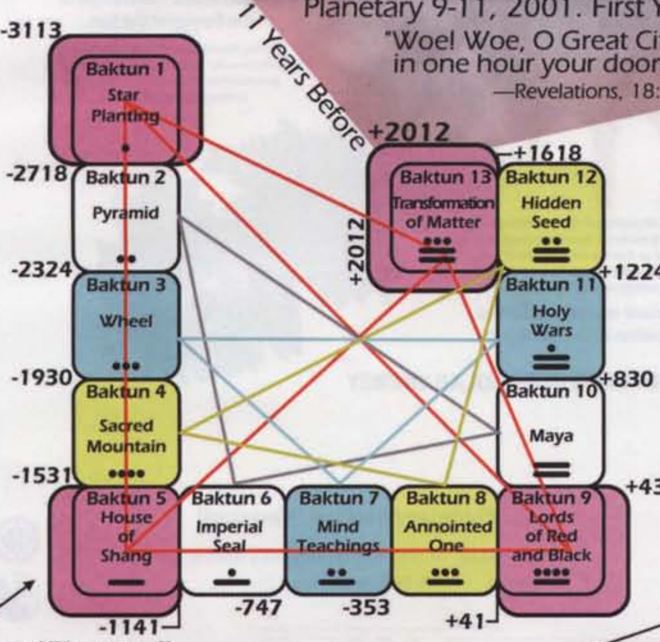
11 Years Before
-3113
Baktun 1
Star Planting
-2718
Baktun 2
Pyramid
-2324
Baktun 3
Wheel
-1930
Baktun 4
Sacred Mountain
-1531
Baktun 5
House of Shang
-1141
Baktun 6
Imperial Seal
-747
Baktun 7
Mind Teachings
-353
Baktun 8
Annointed One
+41
Baktun 9
Lords of Red and Black
+435
Baktun 10
Maya
+830
Baktun 11
Holy Wars
+1224
Baktun 12
Hidden Seed
+1618
Baktun 13
Transformation of Matter
+2012
+2012

13 Tone Wavespell
as Thirteen Baktun
Wave Harmonic
of History

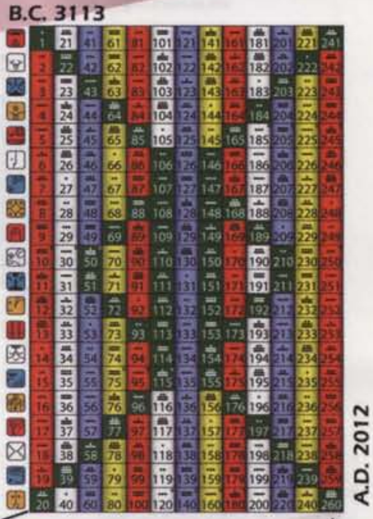

B.C. 3113

1	21	41	61	81	101	121	141	161	181	201	221	241
2	22	42	62	82	102	122	142	162	182	202	222	242
3	23	43	63	83	103	123	143	163	183	203	223	243
4	24	44	64	84	104	124	144	164	184	204	224	244
5	25	45	65	85	105	125	145	165	185	205	225	245
6	26	46	66	86	106	126	146	166	186	206	226	246
7	27	47	67	87	107	127	147	167	187	207	227	247
8	28	48	68	88	108	128	148	168	188	208	228	248
9	29	49	69	89	109	129	149	169	189	209	229	249
10	30	50	70	90	110	130	150	170	190	210	230	250
11	31	51	71	91	111	131	151	171	191	211	231	251
12	32	52	72	92	112	132	152	172	192	212	232	252
13	33	53	73	93	113	133	153	173	193	213	233	253
14	34	54	74	94	114	134	154	174	194	214	234	254
15	35	55	75	95	115	135	155	175	195	215	235	255
16	36	56	76	96	116	136	156	176	196	216	236	256
17	37	57	77	97	117	137	157	177	197	217	237	257
18	38	58	78	98	118	138	158	178	198	218	238	258
19	39	59	79	99	119	139	159	179	199	219	239	259
20	40	60	80	100	120	140	160	180	200	220	240	260

A.D. 2012

Baktun 13 - Katun 260 - 1993 - 2012
End of Second Christian Millennium A.D. 2000

Nostradamus Prediction
of nuclear destruction of
New York in Year 2000

Harmonic Module
13 Baktun
260 Katun
13:2 Matrix of History
1 Kin = 1 Katun
1 Baktun = 20 Katuns

THE END OF THE WORLD
O FIM DO MUNDO

Cover of "Istoe" (Brazil) May, 1996. "End of the World."

16. The Tower

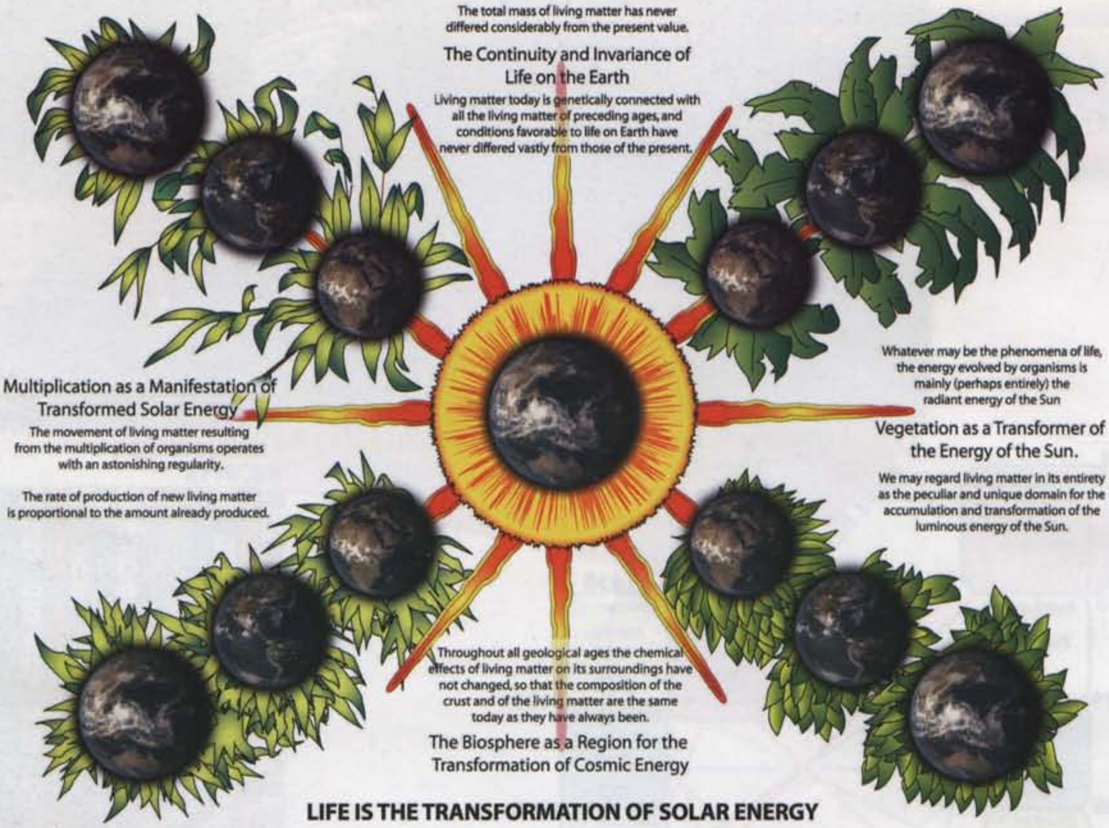

The total mass of living matter has never differed considerably from the present value.

The Continuity and Invariance of Life on the Earth

Living matter today is genetically connected with all the living matter of preceding ages, and conditions favorable to life on Earth have never differed vastly from those of the present.

Multiplication as a Manifestation of Transformed Solar Energy

The movement of living matter resulting from the multiplication of organisms operates with an astonishing regularity.

The rate of production of new living matter is proportional to the amount already produced.

Whatever may be the phenomena of life, the energy evolved by organisms is mainly (perhaps entirely) the radiant energy of the Sun

Vegetation as a Transformer of the Energy of the Sun.

We may regard living matter in its entirety as the peculiar and unique domain for the accumulation and transformation of the luminous energy of the Sun.

Throughout all geological ages the chemical effects of living matter on its surroundings have not changed, so that the composition of the crust and of the living matter are the same today as they have always been.

The Biosphere as a Region for the Transformation of Cosmic Energy

LIFE IS THE TRANSFORMATION OF SOLAR ENERGY

Planetary Service Wavespell

YEAR OF THE RED PLANETARY MOON

© 2002.7.26 ~ 2003.7.25

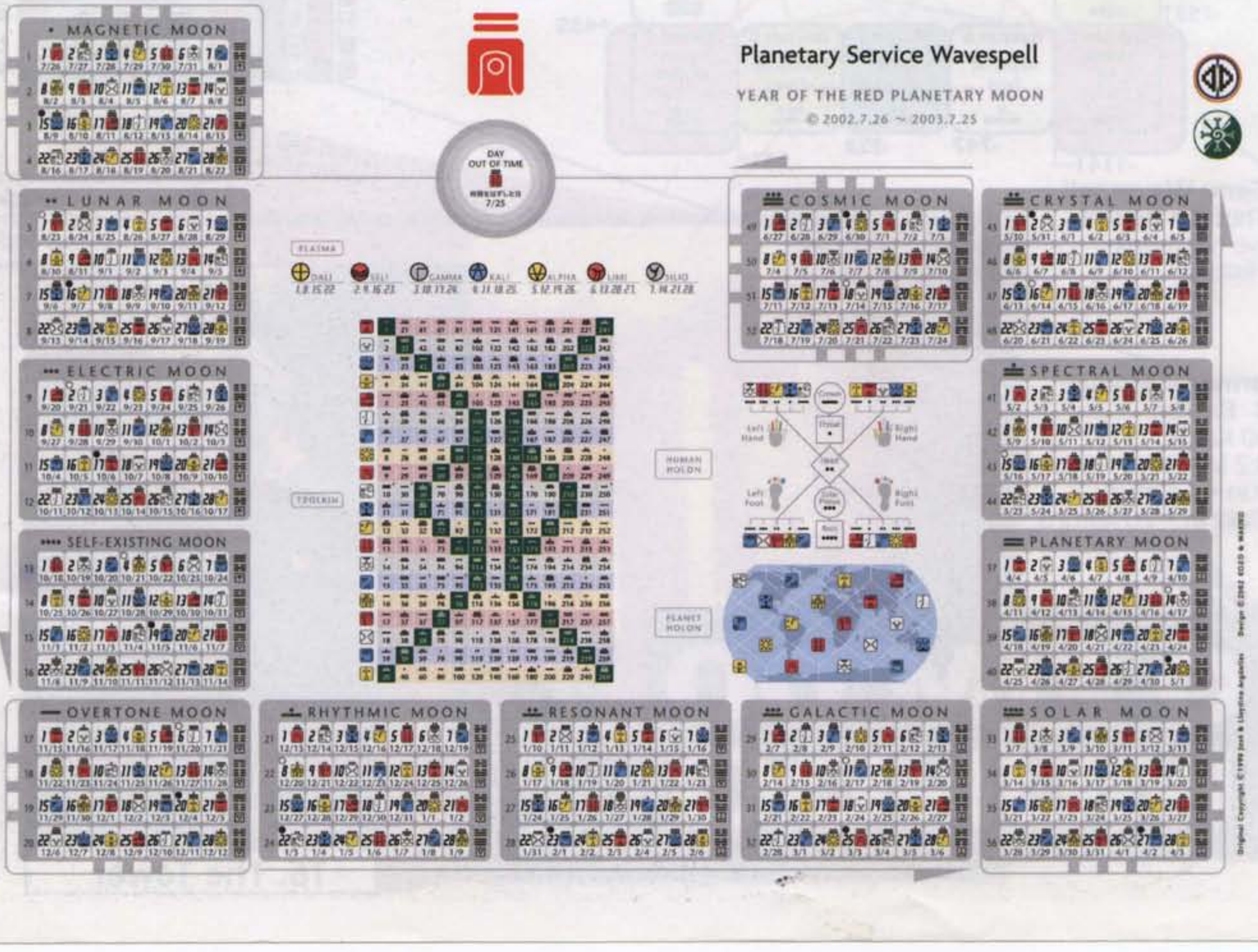
MAGNETIC MOON
LUNAR MOON
ELECTRIC MOON
SELF-EXISTING MOON
OVERTONE MOON
RHYTHMIC MOON
RESONANT MOON
GALACTIC MOON
SOLAR MOON
PLANETARY MOON
SPECTRAL MOON
CRYSTAL MOON
COSMIC MOON
DAY OUT OF TIME
7/25
PLASMA
DALI
SELI
GAMMA
KALI
ALPHA
LIMI
SILIO
TOOLKIN
HUMAN HOLON
PLANET HOLON
Left Hand
Right Hand
Left Foot
Right Foot

Original Copyright ©1990 José & Lloydine Argüelles   Design ©2002 KOZO & MAKIKO

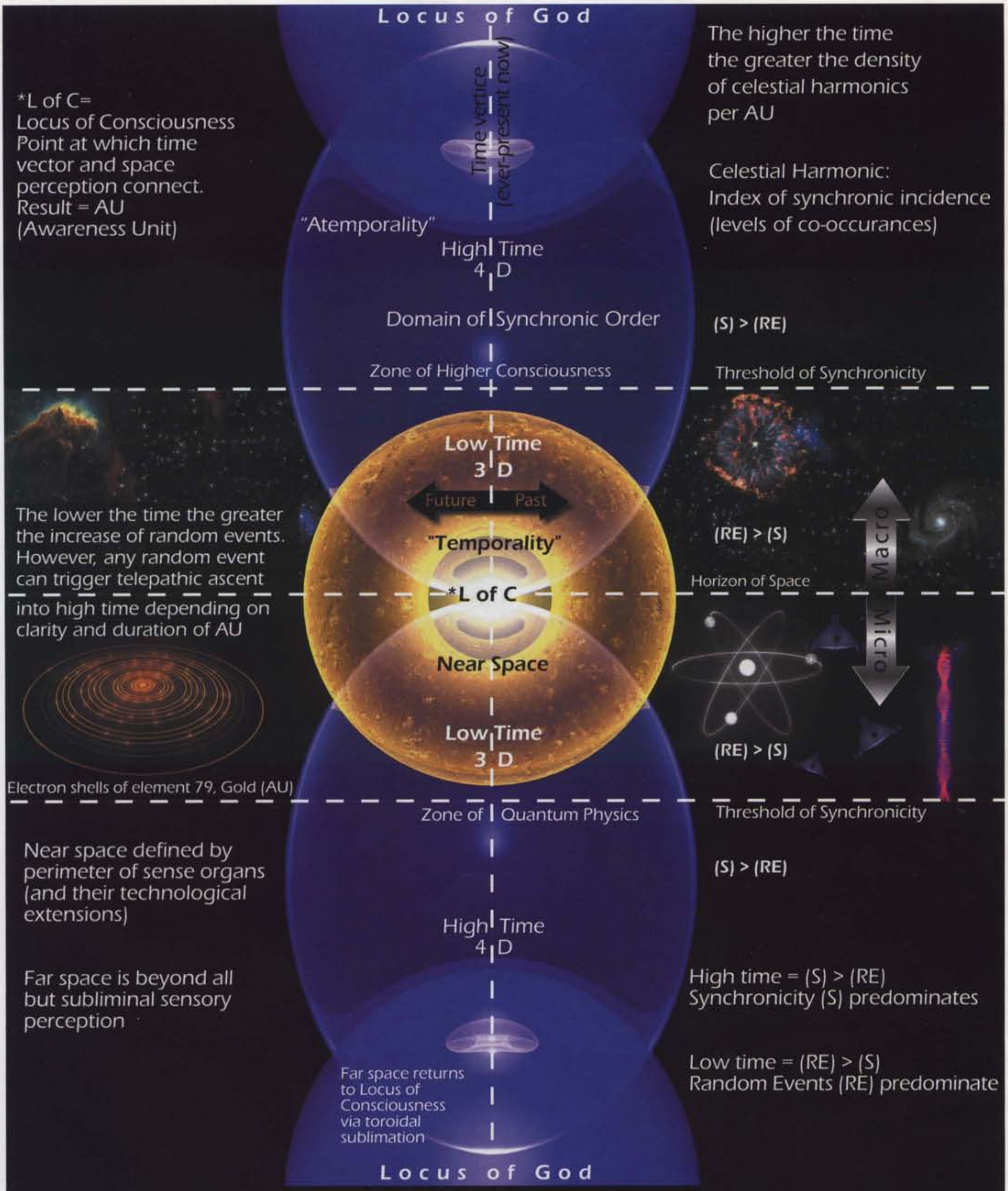
Locus of God
Time vertice (ever-present now)
"Atemporality"
High Time
4 D
Domain of Synchronic Order
Zone of Higher Consciousness
Threshold of Synchronicity
Low Time
3 D
Future
Past
"Temporality"
*L of C
Horizon of Space
Macro
Micro
Near Space
Low Time
3 D
Electron shells of element 79, Gold (AU)
Zone of Quantum Physics
Threshold of Synchronicity
High Time
4 D
Far space returns to Locus of Consciousness via toroidal sublimation
Locus of God
(S) > (RE)
(RE) > (S)
(RE) > (S)
(S) > (RE)

*L of C=
Locus of Consciousness
Point at which time vector and space perception connect.
Result = AU
(Awareness Unit)

The higher the time the greater the density of celestial harmonics per AU

Celestial Harmonic:
Index of synchronic incidence (levels of co-occurances)

The lower the time the greater the increase of random events. However, any random event can trigger telepathic ascent into high time depending on clarity and duration of AU

Near space defined by perimeter of sense organs (and their technological extensions)

Far space is beyond all but subliminal sensory perception

High time = (S) > (RE)
Synchronicity (S) predominates

Low time = (RE) > (S)
Random Events (RE) predominate

Model of "Vertical Time" in Relation to Horizontal Space

A Presentation of the Foundation for the Law of Time
T(E)=ART. Energy factored by Time equals Art.

The Dynamics of Time
Showing the Evolution of Time as Consciousness - T(E)=Art

GALACTIC BRAIN, or evolutionary spectrum showing index of energy-mind progression from atomic-cellular biological to purely mental conditions of being, T(E)=Art. Energy factored by time equals art.

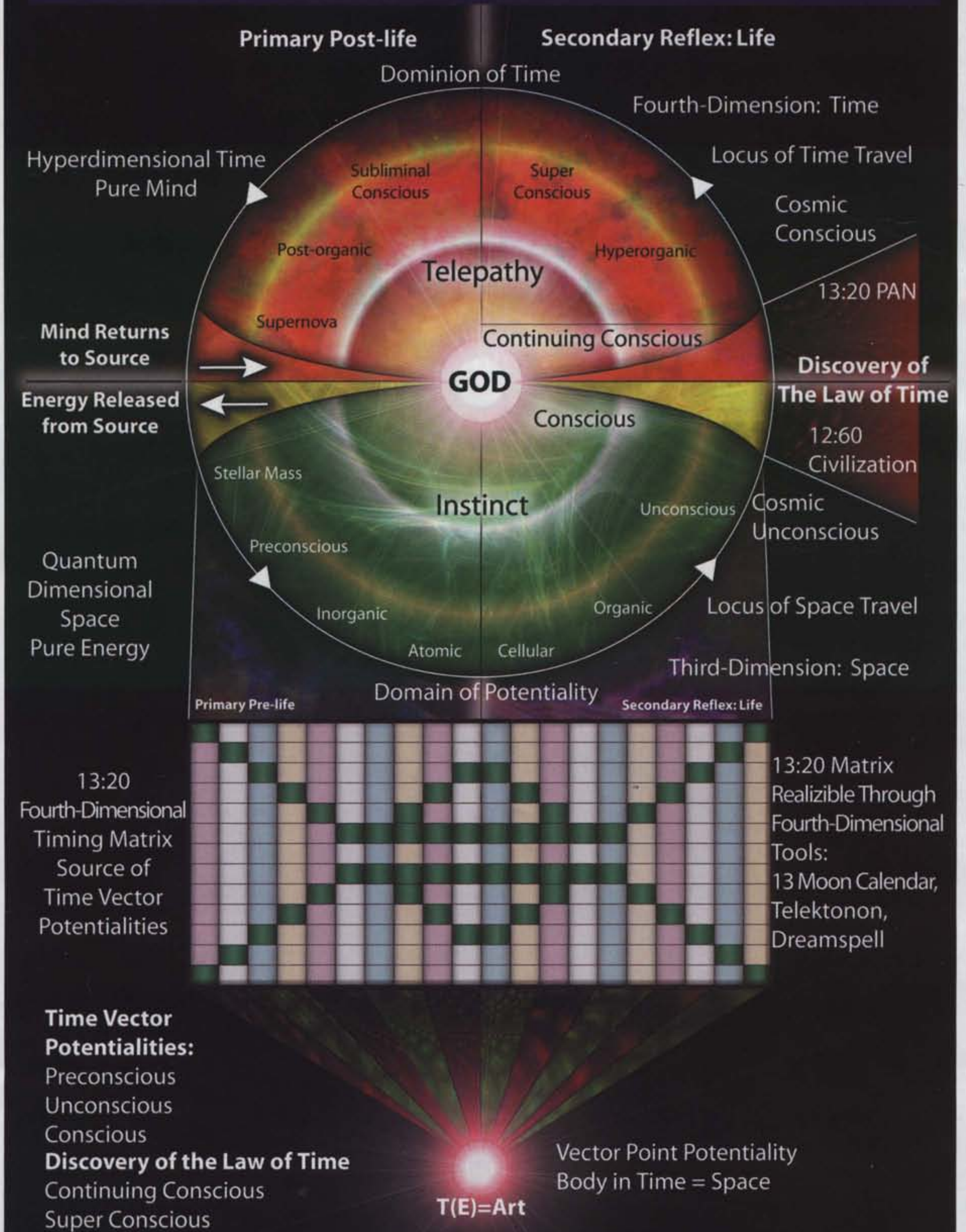

Primary Post-life

Secondary Reflex: Life

Dominion of Time

Fourth-Dimension: Time

Hyperdimensional Time
Pure Mind

Subliminal Conscious

Super Conscious

Locus of Time Travel

Cosmic Conscious

Post-organic

Hyperorganic

Telepathy

13:20 PAN

Supernova

Continuing Conscious

Mind Returns to Source

GOD

Discovery of The Law of Time

Energy Released from Source

Conscious

12:60 Civilization

Stellar Mass

Instinct

Unconscious

Cosmic Unconscious

Preconscious

Quantum
Dimensional
Space
Pure Energy

Inorganic

Organic

Locus of Space Travel

Atomic

Cellular

Third-Dimension: Space

Domain of Potentiality

Primary Pre-life

Secondary Reflex: Life

13:20
Fourth-Dimensional
Timing Matrix
Source of
Time Vector
Potentialities

13:20 Matrix
Realizible Through
Fourth-Dimensional
Tools:
13 Moon Calendar,
Telektonon,
Dreamspell

Time Vector Potentialities:
Preconscious
Unconscious
Conscious
Discovery of the Law of Time
Continuing Conscious
Super Conscious
Subliminal Conscious

Vector Point Potentiality
Body in Time = Space

T(E)=Art

Noospheric Time Map - Wave Harmonic of History

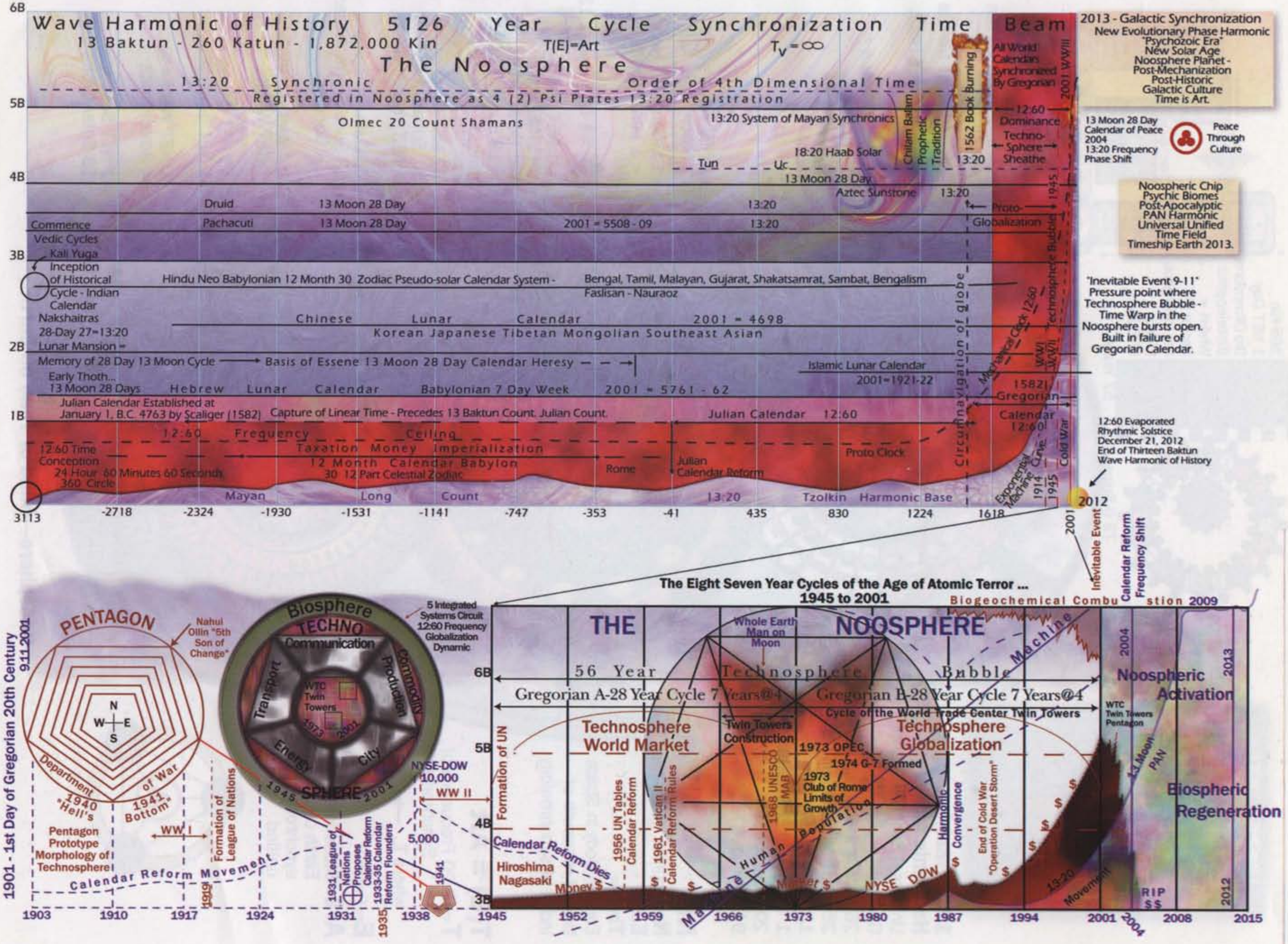
6B
Wave Harmonic of History 5126 Year Cycle Synchronization Time Beam
13 Baktun - 260 Katun - 1,872,000 Kin
T(E)=Art
$T_V = \infty$
The Noosphere
13:20 Synchronic Order of 4th Dimensional Time
Registered in Noosphere as 4 (2) Psi Plates 13:20 Registration
5B
Olmec 20 Count Shamans
13:20 System of Mayan Synchronics
18:20 Haab Solar
Tun
Uc
Chilam Balam Prophetic Tradition
1562 Book Burning
13:20
All World Calendars Synchronized By Gregorian
2001 WWIII
12:60 Dominance
Techno-Sphere Sheathe
4B
13 Moon 28 Day
Aztec Sunstone
13:20
Druid
13 Moon 28 Day
13:20
Proto-Globalization
1945
Commence
Pachacuti
13 Moon 28 Day
2001 = 5508 - 09
13:20
Vedic Cycles
Technosphere Bubble
3B
Kali Yuga
Inception of Historical Cycle - Indian Calendar
Hindu Neo Babylonian 12 Month 30 Zodiac Pseudo-solar Calendar System -
Bengal, Tamil, Malayan, Gujarat, Shakatsamrat, Sambat, Bengalism
Faslisan - Nauraoz
Nakshaitras
Chinese Lunar Calendar
2001 = 4698
28-Day 27=13:20
Korean Japanese Tibetan Mongolian Southeast Asian
2B
Lunar Mansion =
Mechanical Clock 12:60
Memory of 28 Day 13 Moon Cycle
Basis of Essene 13 Moon 28 Day Calendar Heresy
Islamic Lunar Calendar
2001=1921-22
WWI
WWII
Early Thoth...
13 Moon 28 Days
Hebrew Lunar Calendar
Babylonian 7 Day Week
2001 = 5761 - 62
1582 Gregorian
Julian Calendar Established at January 1, B.C. 4763 by Scaliger (1582)
Capture of Linear Time - Precedes 13 Baktun Count. Julian Count.
Julian Calendar 12:60
Calendar 12:60
1B
12:60 Frequency Ceiling
Taxation Money Imperialization
12:60 Time Conception
12 Month Calendar Babylon
30 12 Part Celestial Zodiac
24 Hour 60 Minutes 60 Seconds
360 Circle
Rome
Julian Calendar Reform
Proto Clock
Circumnavigation of globe
Exponential Machine Curve
1914
1945
Cold War
Mayan Long Count
13:20
Tzolkin Harmonic Base
3113
-2718
-2324
-1930
-1531
-1141
-747
-353
-41
435
830
1224
1618
2012
2001
2013 - Galactic Synchronization
New Evolutionary Phase Harmonic
"Psychozoic Era"
New Solar Age
Noosphere Planet -
Post-Mechanization
Post-Historic
Galactic Culture
Time is Art.
13 Moon 28 Day
Calendar of Peace
2004
13:20 Frequency
Phase Shift
Peace Through Culture
Noospheric Chip
Psychic Biomes
Post-Apocalyptic
PAN Harmonic
Universal Unified
Time Field
Timeship Earth 2013.
"Inevitable Event 9-11"
Pressure point where
Technosphere Bubble -
Time Warp in the
Noosphere bursts open.
Built in failure of
Gregorian Calendar.
12:60 Evaporated
Rhythmic Solstice
December 21, 2012
End of Thirteen Baktun
Wave Harmonic of History
Inevitable Event
Calendar Reform
Frequency Shift
The Eight Seven Year Cycles of the Age of Atomic Terror ...
1945 to 2001
Biogeochemical Combustion 2009
1901 - 1st Day of Gregorian 20th Century
9.11.2001
PENTAGON
Nahui Ollin "5th Son of Change"
N
W
E
S
Department of War
1940 "Hell's
1941 Bottom"
Pentagon
Prototype
Morphology of
Technosphere
WW I
Formation of League of Nations
1919
Calendar Reform Movement
Biosphere
TECHNO
Communication
Commodity
Production
Transport
WTC
Twin
Towers
1973
2001
Energy
City
1945
SPHERE
2001
5 Integrated
Systems Circuit
12:60 Frequency
Globalization
Dynamic
NYSE-DOW
10,000
WW II
5,000
1931 League of Nations Proposes Calendar Reform
1933-35 Calendar Reform Flounders
1941
1903
1910
1917
1924
1931
1935
1938
1945
THE NOOSPHERE
Whole Earth
Man on
Moon
6B
56 Year Technosphere Bubble
Gregorian A-28 Year Cycle 7 Years@4
Gregorian B-28 Year Cycle 7 Years@4
Cycle of the World Trade Center Twin Towers
5B
Formation of UN
Technosphere
World Market
Twin Towers
Construction
1973 OPEC
1974 G-7 Formed
1973
Club of Rome
Limits of
Growth
1968 UNESCO MAB
Technosphere
Globalization
4B
Calendar Reform Dies
1956 UN Tables Calendar Reform
1961 Vatican II Calendar Reform Rules
Hiroshima
Nagasaki
Money $
Human Population
Machine
Market $
NYSE DOW
Harmonic Convergence
End of Cold War
"Operation Desert Storm"
$
3B
13:20
Movement
WTC
Twin Towers
Pentagon
13 Moon
PAN
Noospheric
Activation
Biospheric
Regeneration
RIP
$$
2004
2013
2012
1945
1952
1959
1966
1973
1980
1987
1994
2001
2004
2008
2015

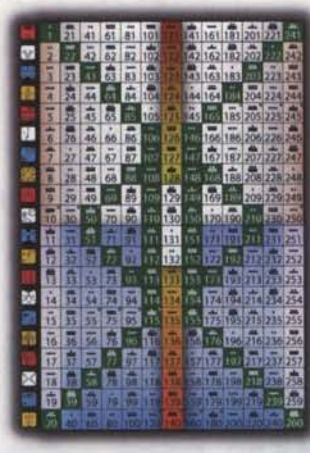

260 Kin
4th Dimensional
Wheel of
Synchronic
Order =
Factor of
Synchronization=
1 G-Spin =

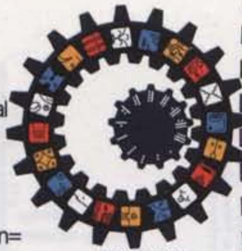

73 G-Spins = 52 Years

July 24, 2004
July 25, 2004
July 26, 2004
July 27, 2004
July 28, 2004

365-Kin
1 NET Day
3rd Dimensional
Biotelepathic
Wheel =
Factor of
Life Experience
(73:5)
(7:52)

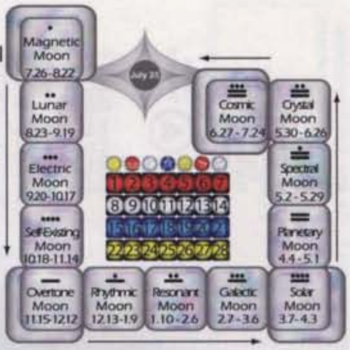
Magnetic Moon 7.26-8.22
Lunar Moon 8.23-9.19
Electric Moon 9.20-10.17
Self-Existing Moon 10.18-11.14
Overtone Moon 11.15-12.12
Rhythmic Moon 12.13-1.9
Resonant Moon 1.10-2.6
Galactic Moon 2.7-3.6
Solar Moon 3.7-4.3
Planetary Moon 4.4-5.1
Spectral Moon 5.2-5.29
Crystal Moon 5.30-6.26
Cosmic Moon 6.27-7.24
July 25
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28

13:20 Matrix Harmonic Module

Universal Synchronization
Gauge =
260 Kin = 1 Galactic Spin =
260 NET Minutes
13 NET Minutes =
1 Wavespell (13 Tones)
20 Wavespells =
9 NET Hours, 8 NET Minutes

NET

Noospheric
Earth
Time =
Earth = Body in Space =
Unit of measure of
index of synchronicity

T(E) = Art

T=13:20 Frequency of Synchronization
where $T_V \rightarrow \infty$

E= Earth

Art= Blossomed Noosphere, Earth consciousness synchronically unified

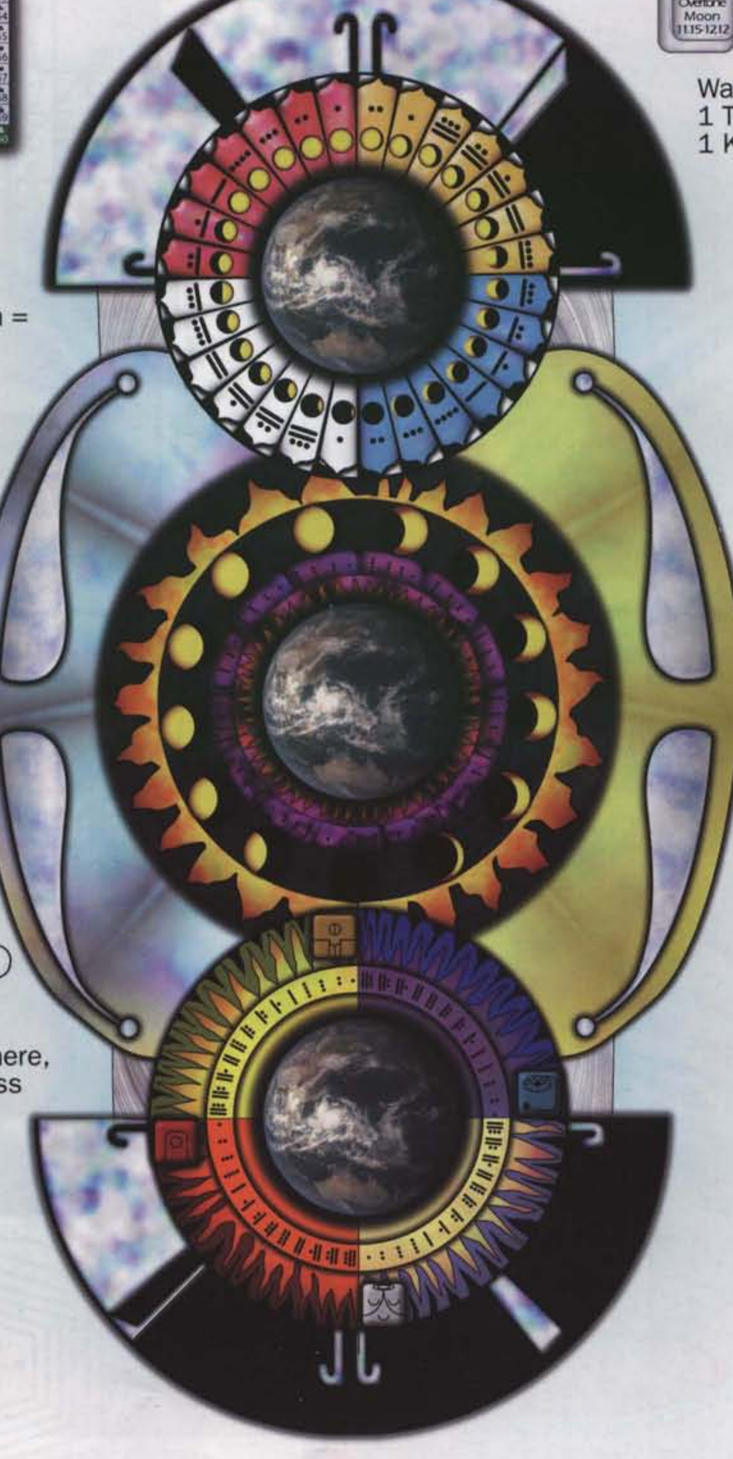

Wavespell = 13 Tones
1 Tone = 1 Moon = 28 Kin
1 Kin = 1 NET Minute

◐ = 1 NET Minute
x 28 = 1 NET Hour.

7 NET Minutes =
$^1/_4$ Hour =
1 Week ◔

1 NET Hour = ◐
1 Moon

1 Moon (NET Hour)
= 1 Lunar Orbit

◐ =1 NET Hour x 13
= 1 NET Day
(13 NET Hours)
= 1 Solar Orbit
= 13 Lunar Orbits

52 NET Days
= 52 Solar Orbits
= 1 NET Year

= 73 Galactic Spins
= 13 Seed-Storm 4-Year Bundles

1 4-Year Bundle =
1 NET Week

13 NET Days=
1 NET Season

1 NET Week=
4 Solar Orbits

1 NET Season=
13 Solar Orbits

Noosphere—Mind of Time on Earth

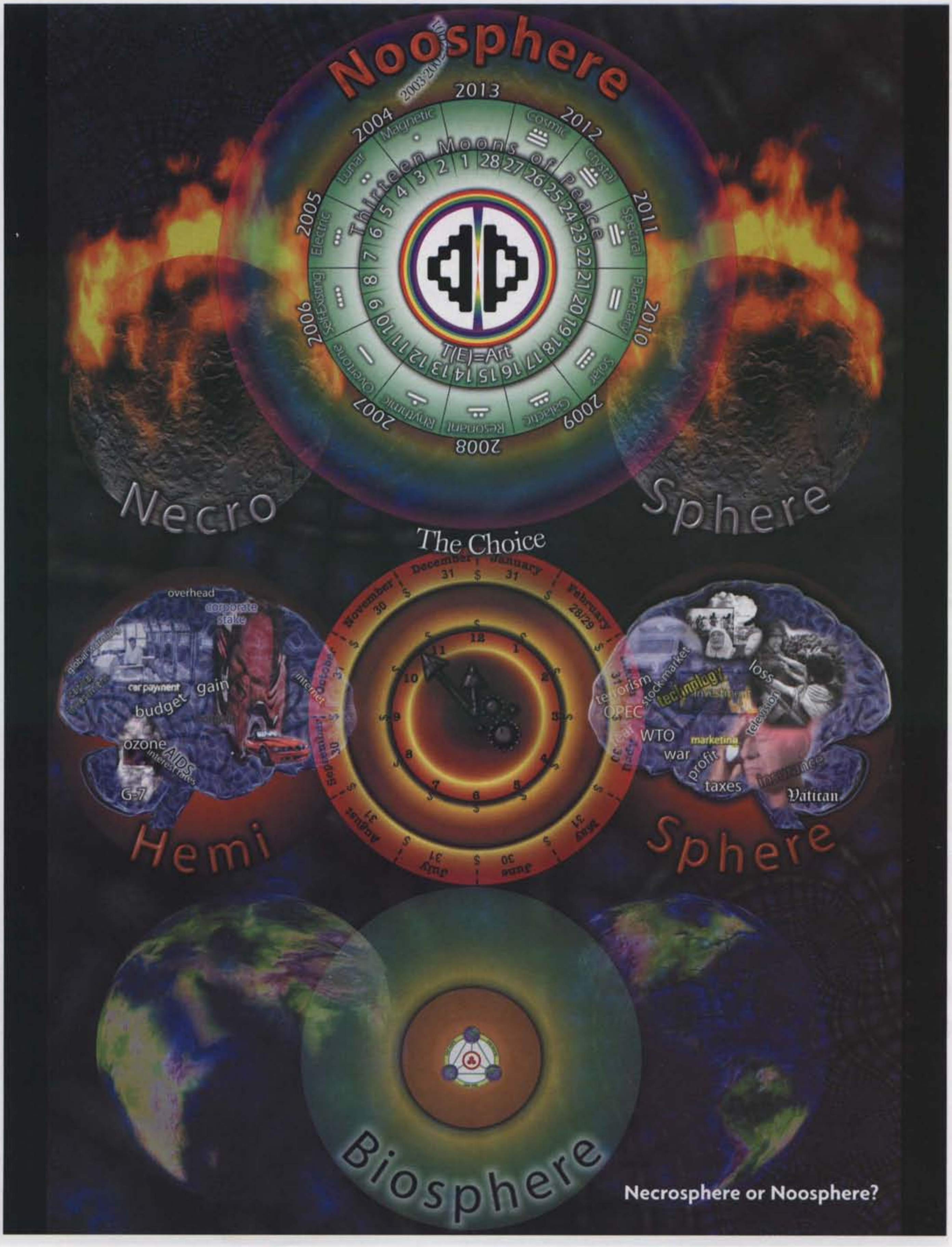
Noosphere
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003-2004
Magnetic
Lunar
Electric
Self-Existing
Overtone
Rhythmic
Resonant
Galactic
Solar
Planetary
Spectral
Crystal
Cosmic
Thirteen Moons of Peace
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
T(E)=Art
Necro
Sphere
The Choice
December 31
January 31
February 28/29
November 30
October 31
September 30
August 31
July 31
June 30
May 31
12
1
2
3
4
5
6
7
8
9
10
11
overhead
corporate stake
global warming
car payment
gain
budget
internet
ozone
AIDS
interest rates
G-7
terrorism
stock market
technology
investment
loss
OPEC
WTO
marketing
war
profit
television
taxes
insurance
Vatican
Hemi
Sphere
Biosphere

Necrosphere or Noosphere?

A Chronotopology of Transformation

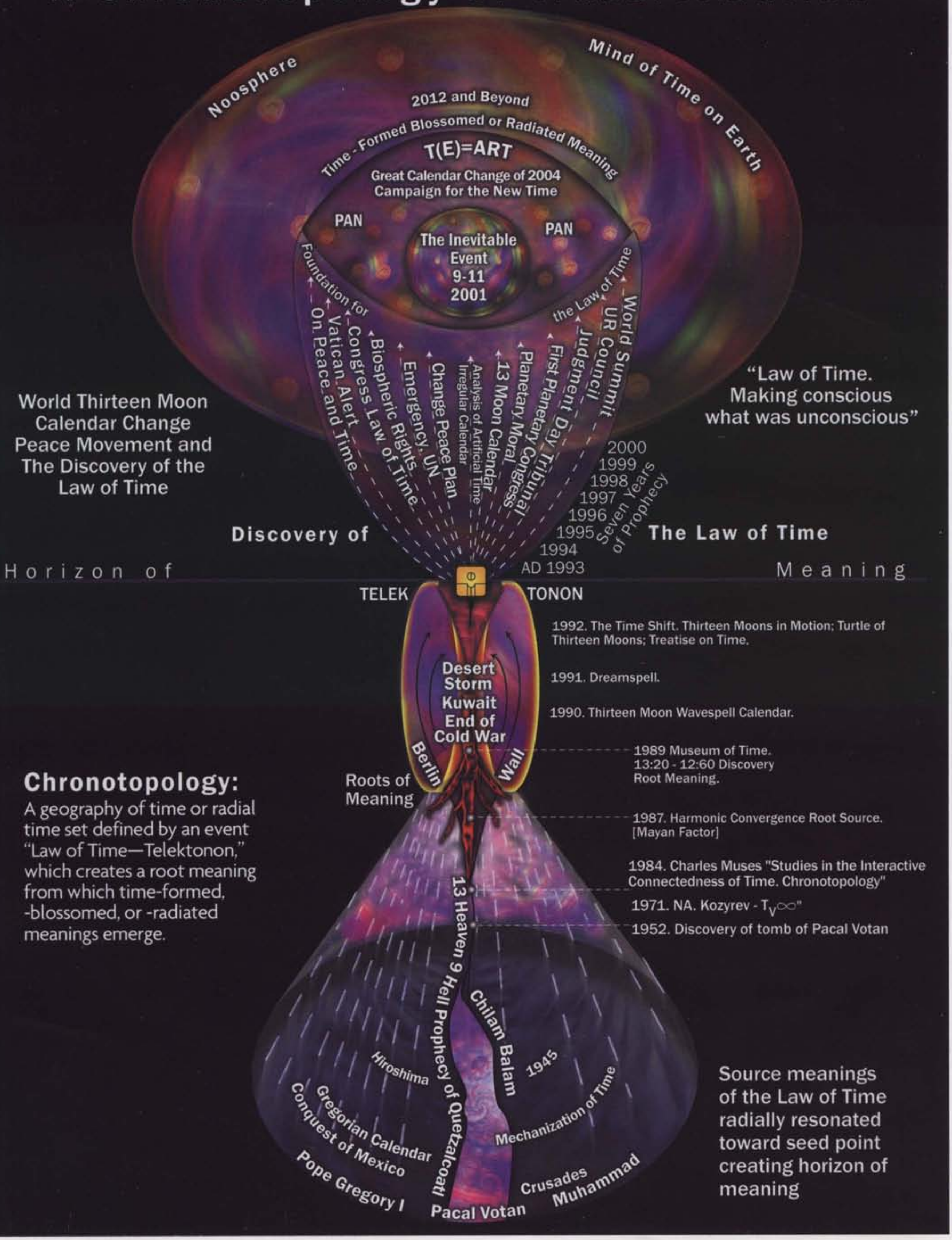
Noosphere
Mind of Time on Earth
2012 and Beyond
Time-Formed Blossomed or Radiated Meaning
T(E)=ART
Great Calendar Change of 2004
Campaign for the New Time
PAN
PAN
The Inevitable Event 9-11 2001
Foundation for the Law of Time
On Peace and Time
Vatican Alert
Congress Law of Time
Biospheric Rights
Emergency UN
Change Peace Plan
Analysis of Artificial Time Irregular Calendar
13 Moon Calendar
Planetary Moral Congress
First Planetary Day Tribunal
Judgment
UR Council
World Summit
2000
1999
1998
1997
1996
1995
1994
AD 1993
Seven Years of Prophecy
Discovery of
The Law of Time
Horizon of
Meaning
TELEK
TONON
Desert Storm Kuwait End of Cold War
Berlin
Wall
Roots of Meaning
13 Heaven 9 Hell Prophecy of Quetzalcoatl
Chilam Balam
1945
Hiroshima
Mechanization of Time
Gregorian Calendar
Conquest of Mexico
Pope Gregory I
Crusades
Muhammad
Pacal Votan

World Thirteen Moon
Calendar Change
Peace Movement and
The Discovery of the
Law of Time

"Law of Time.
Making conscious
what was unconscious"

1992. The Time Shift. Thirteen Moons in Motion; Turtle of Thirteen Moons; Treatise on Time.

1991. Dreamspell.

1990. Thirteen Moon Wavespell Calendar.

1989 Museum of Time.
13:20 - 12:60 Discovery
Root Meaning.

1987. Harmonic Convergence Root Source.
[Mayan Factor]

1984. Charles Muses "Studies in the Interactive Connectedness of Time. Chronotopology"

1971. NA. Kozyrev - $T_V\infty$"

1952. Discovery of tomb of Pacal Votan

Chronotopology:

A geography of time or radial time set defined by an event "Law of Time—Telektonon," which creates a root meaning from which time-formed, -blossomed, or -radiated meanings emerge.

Source meanings
of the Law of Time
radially resonated
toward seed point
creating horizon of
meaning

Paul J. Nahin

Time Machine Tales

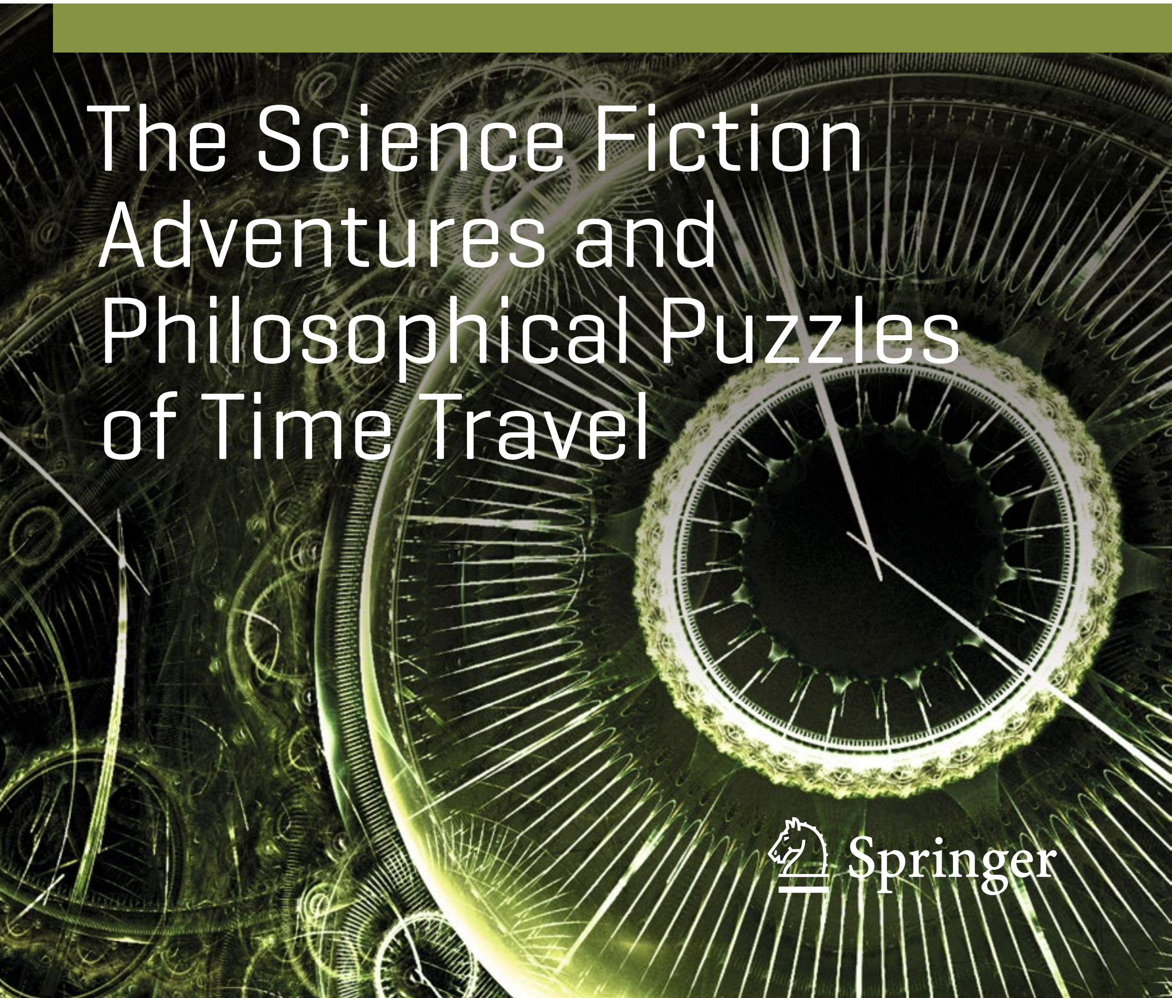

The Science Fiction Adventures and Philosophical Puzzles of Time Travel

Springer

Science and Fiction

Editorial Board

Mark Alpert
Philip Ball
Gregory Benford
Michael Brotherton
Victor Callaghan
Amnon H. Eden
Nick Kanas
Geoffrey Landis
Rudi Rucker
Dirk Schulze-Makuch
Rüdiger Vaas
Ulrich Walter
Stephen Webb

More information about this series at http://www.springer.com/series/11657

Science and Fiction – A Springer Series

This collection of entertaining and thought-provoking books will appeal equally to science buffs, scientists and science-fiction fans. It was born out of the recognition that scientific discovery and the creation of plausible fictional scenarios are often two sides of the same coin. Each relies on an understanding of the way the world works, coupled with the imaginative ability to invent new or alternative explanations - and even other worlds. Authored by practicing scientists as well as writers of hard science fiction, these books explore and exploit the borderlands between accepted science and its fictional counterpart. Uncovering mutual influences, promoting fruitful interaction, narrating and analyzing fictional scenarios, together they serve as a reaction vessel for inspired new ideas in science, technology, and beyond.

Whether fiction, fact, or forever undecidable: the Springer Series "Science and Fiction" intends to go where no one has gone before!

Its largely non-technical books take several different approaches. Journey with their authors as they

- Indulge in science speculation – describing intriguing, plausible yet unproven ideas;
- Exploit science fiction for educational purposes and as a means of promoting critical thinking;
- Explore the interplay of science and science fiction – throughout the history of the genre and looking ahead;
- Delve into related topics including, but not limited to: science as a creative process, the limits of science, interplay of literature and knowledge;
- Tell fictional short stories built around well-defined scientific ideas, with a supplement summarizing the science underlying the plot.

Readers can look forward to a broad range of topics, as intriguing as they are important. Here just a few by way of illustration:

- Time travel, superluminal travel, wormholes, teleportation
- Extraterrestrial intelligence and alien civilizations
- Artificial intelligence, planetary brains, the universe as a computer, simulated worlds
- Non-anthropocentric viewpoints
- Synthetic biology, genetic engineering, developing nanotechnologies
- Eco/infrastructure/meteorite-impact disaster scenarios
- Future scenarios, transhumanism, posthumanism, intelligence explosion
- Virtual worlds, cyberspace dramas
- Consciousness and mind manipulation

Paul J. Nahin

Time Machine Tales

The Science Fiction Adventures and Philosophical Puzzles of Time Travel

Springer

Paul J. Nahin
Department of Electrical Engineering
University of New Hampshire
Durham, NH, USA

ISSN 2197-1188 ISSN 2197-1196 (electronic)
Science and Fiction
ISBN 978-3-319-48862-2 ISBN 978-3-319-48864-6 (eBook)
DOI 10.1007/978-3-319-48864-6

Library of Congress Control Number: 2016957723

© Springer International Publishing AG 2017
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Also By Paul J. Nahin

Oliver Heaviside (1988, 2002), Johns Hopkins
Time Machines (1993, 1999), Springer
The Science of Radio (1996, 2001), Springer
An Imaginary Tale (1998, 2007, 2010), Princeton
Duelling Idiots (2000, 2002), Princeton
When Least Is Best (2004, 2007), Princeton
Dr. Euler's Fabulous Formula (2006, 2011), Princeton
Chases and Escapes (2007, 2012), Princeton
Digital Dice (2008, 2013), Princeton
Mrs. Perkins's Electric Quilt (2009), Princeton
Time Travel (1997, 2011), Johns Hopkins
Number-Crunching (2011), Princeton
The Logician and the Engineer (2013), Princeton
Will You Be Alive Ten Years From Now? (2014), Princeton
Holy Sci-Fi! (2014), Springer
Inside Interesting Integrals (2015), Springer
In Praise of Simple Physics (2016), Princeton

Frontispiece: The Pioneers of Time Travel

The scientific pioneers were Albert Einstein (1879–1955) and Kurt Gödel (1906–1978), good personal friends who are shown here in 1954 at the Institute for Advanced Study in Princeton, New Jersey, in a photo taken by Richard Arens. It was Einstein's 1916 general theory of relativity (theory of gravity) that Gödel used as the basis for his 1949 paper that was the first to show that the general theory does **not** forbid time travel into the past.

The literary pioneer of time travel was of course Herbert George Wells (1866–1946), who is shown here as a college freshman cut-up around 1885. The photograph was taken as a prank by an unknown friend while Wells was a student in a biology course given by Thomas Huxley, at the Normal School of Science in

South Kensington (a branch of the University of London). A far too thin and impoverished Wells was then still a teenager, and *The Time Machine* lay a distant 10 years in the future.

Einstein/Gödel photograph courtesy of the American Institute of Physics Emilio Segré Visual Archives of the AIP Niels Bohr Library. Wells photograph courtesy of the rare Books and Special Collections Department of the Library of the University of Illinois at Urbana-Champaign.

A Note on the Story Citations and Science Fiction History

> "You will find it a very good practice always to verify your references, sir."
> —advice given in 1847 to a young scholar by Martin Joseph
> Routh, President of Magdalen College, Oxford

Most of the pulp science fiction stories I've cited in this book, in their original form as ink on paper, have long since vanished from our region of spacetime and exist today only (alas) on microfilm reels in scholarly vaults. I am especially indebted to Texas A & M, the Claremont Colleges, the California State Universities at Northridge and Fullerton, Mount Holyoke College, the New York City Public

Library, and the University of Delaware, for giving me access via Inter-Library Loan (through my home institution, the University of New Hampshire) to their extensive archives of ancient science fiction magazines.

A number of the really good stories *have* been anthologized, however, and so are still readily available today in book form. In essentially all cases, though, for historical reasons, I've given the original publication information (magazine and date). You can find which of the stories cited are available in one or more anthology reprints by going to an immensely useful, searchable database on the Web, at: http://www.isfdb.org, and I gratefully thank all those in the science fiction community responsible for creating and maintaining that database.

The following two books by science fiction historian Sam Moskowitz (1920–1997), who lived through what Isaac Asimov called the 'Golden Age of [magazine] Science Fiction,' may be difficult to find today but, if you are interested in the early history of *magazine* science fiction (beyond simply the subgenre of time travel), the hunt for them will be well worth your time:

Science Fiction by Gaslight: a history and anthology of science fiction in the popular magazines, 1891–1911 (World Publishing Company 1968);
Under the Moons of Mars: a history and anthology of "The Scientific Romance" in the Munsey Magazines, 1912–1920 (Holt, Rinehart and Winston 1970).

Some First Words

Is time travel in principle (never mind the difficulties) a possibility? It has received some thought in the past and deserves some more.

—David Park, in his 1980 book *The Image of Eternity*

He used to have quite a reputation, but the last couple of years he's been working on time ... You know, time travel, that sort of rot. An A-1 crackpot.

—a character (discussing a colleague) disagrees with Park, in Mack Reynolds' "Advice from Tomorrow," *Science Fiction Quarterly*, August 1953

In 1993 the first edition of my book *Time Machines* was published by the Press of the American Institute of Physics. In 1999, after Springer acquired AIP Press, the second edition of that book appeared. So, is this the third edition? Well, yes *and* no. It *is* because large chunks of the 1999 edition are still here, along with new discussions of the advances by physicists and philosophers that have appeared in the intervening 18 years. The prime example of that centers on the time travel paradoxes. Those discussions contain mostly what is in the second edition, but they have also been brought up to date with the latest thinking on the paradoxes, by physicists and philosophers.

And yet this book is *not quite* the third edition because the emphasis is now on the philosophical and on science fiction, rather than on physics as it was when written for AIP Press. In that spirit there are, for example, no Tech Notes filled with algebra, integrals, and differential equations, as there are in the first and second editions of *Time Machines*. That's because I wish to avoid having this book seem to be simply a long physics treatise. I have, in fact, some sympathy with the following views, expressed by two philosophers:

"There is one metaphor in the physicist's account of space-time which one would expect *anyone* to recognize as such, for metaphor is here strained far beyond the breaking point, i.e., when it is said that time is 'at right angles to each of the other three dimensions.' Can anyone really attach any meaning to this—except as a recipe for drawing diagrams?"[1]

and

"This is from the outset a study in descriptive metaphysics. In consequence, I shall have nothing to say about twice-differentiable Lorentzian manifolds, Minkowski diagrams, world-lines, time-like separations, space-time worms [a 'thick' world-line], or temporal parts."[2]

I don't *completely* endorse these sentiments, however, and so please understand that I am not denying the ultimate importance of *physics* when it comes to achieving a deep understanding of time travel. To quote yet another philosopher,

"Arm chair reflections on the concept of causation [are] not going to yield new insights. The grandfather paradox is simply a way of pointing to the fact that if the usual laws of physics are supposed to hold true in a chronology violating spacetime, then consistency constraints emerge. *[To understand these constraints] involves solving problems in physics, not armchair philosophical reflections* [my emphasis]."[3]

I could not agree more. So, in *Time Machine Tales* you *will* find some physics. In support of time travel to the future (and in how to make a wormhole time machine for travel into the past), for example, I'll show you a high school level derivation of the famous time dilation formula from special relativity. There are some spacetime diagrams, some simple algebraic manipulations, and here and there just a touch of freshman calculus; even the metric tensor gets a few words, too. But it is, admittedly, pretty light-weight stuff.

So, while certainly saluting the premier position of physics, *Time Machine Tales* is not a scholarly, in-depth treatment of time travel physics. Rather, it is an examination of how science fiction writers (and many philosophers, too) have viewed time travel. (Even in the physics discussions, science fiction will regularly appear.) Those views, by their very nature, are far more romantic than are those of hardcore theoretical physicists. History has shown, of course, that the results of the work of theoretical physicists may, in the end, prove to actually be far more astonishing than anything fiction writers cook-up—and if there is any scientific subject for which that may again prove to be true it's time travel—but for us, here, it will be the fiction writer who has center stage.

The philosophers will be only slightly less important in this book. While much of the early philosophical literature on time travel and backwards causation reads like imaginative fairy tales spun out of vacuous vapors (more on this soon), many modern philosophers have shown themselves to be quite sophisticated. What they have written deserves serious consideration by anyone interested in time travel, and that includes physicists. However, while the time travel interests of philosophers and physicists have a lot of overlap, those interests are *not* in total agreement. For example, while both groups talk of the grandfather paradox, the philosophers worry in particular about motivation (*why* the murderous mission?), while physicists have *never* to my knowledge asked themselves that question[4] (other than to figure out how to avoid it!). After all, philosophers talk of flesh-and-blood humans as time travelers, while the physicists send only billiard balls (with no personal identities or memories) on time trips into the past for the *expressed purpose* of avoiding the messy human issues of 'motivation' and free will. This approach by physicists isn't because they are cold and emotionless. It is a useful strategy because, if it can be shown that a mere billiard ball can travel into the past then, as one *philosopher* pointed out long ago, "It is implausible that it should be possible for some physical systems to travel back in time, and not others. Thus, if we suppose that simple objects can time-travel . . . then we must suppose that more complicated systems, e.g., human beings, can also time-travel."[5]

---

[1] C. W. K. Mundle, "The Space-Time World," *Mind*, April 1967, pp. 264–269.

[2] J. F. Rosenberg, "One Way of Understanding Time," *Philosophia*, October 1972, pp. 283–301.

[3] John Earman, "Recent Work on Time Travel," in *Time's Arrows Today: recent physical and philosophical work on the direction of time* (Steven F. Savitt, editor), Cambridge University Press 1995, pp. 268–310. We'll discuss the idea of *consistency constraints* in some detail later in the book. Earman is Professor Emeritus of History and Philosophy of Science at the University of Pittsburgh.

For the most part, philosophers and physicists have worked at the extreme, opposite points of the time travel spectrum. Much better, I think, would be to adopt the following, more balanced position advocated recently: "The study of time machines is a good opportunity for forging a partnership between philosophy and physics. Of course, philosophers have to recognize that in this particular instance the partnership is necessarily an unequal one since the mathematical physicists have to do the heavy lifting. But it seems clear that a little more cooperation with philosophers of science in attending to the analysis of what it takes to be a time machine could have led to some helpful clarifications in the physics literature."[6]

In the past, philosophers gained a reputation for being just a bit too 'unconstrained by the facts' for scientific tastes—as the English mathematician Augustus De Morgan (1806–1871) wrote in an 1842 letter, "There are no writers who give us so much *must* with so little *why*, as the metaphysicians"[7]—but I do think today's physicists would do well to reexamine that harsh opinion.

Philosophers of the 'old school' may look askance at a non-philosopher (me!) leveling criticism at them, and so let me step aside and quote from a member of the

---

[4] Nicholas J. J. Smith, "Why Would Time Travelers Try to Kill Their Younger Selves?" *The Monist*, July 2005, pp. 388–395. As Smith writes, "[Motivation] does not impact upon the possibility, or even the likelihood of backwards time travel. Yet it is deeply puzzling, and we will have no idea what time travel would actually be *like* until we explore it." See also Peter B. M. Vranas, "Can I Kill My Younger Self? Time Travel and the Retrosuicide Paradox," *Pacific Philosophical Quarterly*, December 2009, pp. 520–534.

[5] P. Horwich, "On Some Alleged Paradoxes of Time Travel," *Journal of Philosophy*, August 1975, pp. 432–444.

[6] John Earman, Christopher Smeenk, and Christian Wüthrich, "Do the Laws of Physics Forbid the Operation of Time Machines?," *Synthese*, July 2009, pp. 91–124.

[7] D. J. Cohen, *Equations from God: pure mathematics and Victorian faith*, The Johns Hopkins University Press 2007, p. 119.

'modern school' of philosophical thought: "Space-time is the basic spatiotemporal entity. Many philosophers have mouthed this truth, but few have swallowed it, and very few have digested it . . . An appreciation of this truth is crucial to what is commonly referred to as the philosophy of space and time . . . In large measure the lack of progress in this area can be traced to the fact that philosophers have not taken seriously the corollary that talk about space and time is really talk about the spatial and temporal aspects of spacetime."[8] This is a polite way of telling philosophers that they had better learn some physics!

What provoked those harsh words was that 'modern' philosopher's perception that 'old school' philosophers were not talking science when they wrote of space and time, but rather were in the business of telling each other irrelevant stories and myths, a curious philosophical approach involving the 'telling of tales' that reached its peak in the early and mid-1960s. Spacetime story telling seems to have started with a paper by the Oxford philosopher Anthony Quinton (1925–2010), who argued[9] that although there can be multiple, disjointed spaces, there can only be a single time that is the same for everyone, everywhere. The issue is *not* the truth or not of that assertion (Newton believed it, modern physicists don't), but rather Quinton's technique for arriving at it: myth construction.

Myth construction strikes those trained in the technical sciences as, while perhaps interesting—even physicists, after all, can enjoy a good fairy tale now and then—something quaint and totally beside the point. In his paper Quinton tells a fairy tale about how he thinks someone can live continuously in time and yet, via dreaming, be in two different spatial worlds; when awake he is in one world, while when the person is asleep he is in the other. Quinton argues that this multispatial myth is plausible, but that a search for an analogous multitemporal myth is doomed. This prompted a reply[10] from another 'old school' philosopher who rebutted Quinton with an even more outlandish counter-myth involving "the warring tribes of Okku and Bokku"!

It was this back-and-forth spinning of hypothetical tales that caused the 'modern' philosopher to write in his paper (note 30) that "the procedure for arriving at answers to these questions [about space and time] adopted by Quinton and most other ['old school' philosophers] is, to say the least, a curious one: a story is told about a mythical land—usually called something like the land of Okkus-Bokkus [which is now seen to an outrageous pun]—and then we are asked what we would say if confronted by experiences like those of the Okkus-Bokkusians. As often happens with such a question, people have said all sorts of things, not all of which are interesting or enlightening."

Another modern philosopher was even less gentle in his rejection of the fairy tale approach to spacetime physics: "Quinton [and others of a similar approach invite us] to say what we should think in certain strange circumstances which they describe within common-sense language [as opposed to scientific terminology]. I must say that if I found myself in the circumstances which they describe I just would not know what to think. Probably I should simply conclude that I had gone mad . . . It looks as though these writers are inviting us to consider *what we should say if we knew no science* [my emphasis]."[11]

---

[8] J. Earman, "Space-Time or How to Solve Philosophical Problems and Dissolve Philosophical Muddles Without Really Trying," *Journal of Philosophy*, May 1970, pp. 259–276.

[9] A. Quinton, "Spaces and Times," *Philosophy*, April 1962, pp. 130–147.

[10] R. G. Swinburne, "Times," *Analysis*, June 1965, pp. 185–191.

Even before the modern philosopher (note 30) wrote in 1970 to complain about myth-making, another had already done so: "Whenever a human being produces an argument which opens 'Suppose I had 23 senses . . .,' 'Suppose I were God . . .,' 'Suppose I experienced objects extended in four spatial dimensions . . .,' we can protest that the argument is worthless. For in supposing that he has transcended our human point of view, he has also transcended the limits of our understanding."[12] As this author concluded his very funny paper, such opening sentences are the signatures of myths from "The Philosopher's Fairy Tale Book."

The strained relationship between myth-making philosophers and physicists, especially concerning time travel, has a historically interesting antecedent in the 1920s negative reaction among many over Einstein's theories of relativity (the very theories that give apparent life to time travel). To illustrate my point, consider the October 1913 letter Oskar Kraus (1872–1942), a philosophy professor at the German University in Prague, sent to Ernst Gehrcke (1878–1960), a physics professor at the Reich Institute of Physics and Technology in Berlin. Both men were opponents of Einstein but, as Kraus wrote in his letter, it was only Gehricke among the physicists he considered to be sympathetic to him: "[I] would not know . . . anyone else but you who as a specialist would not reject the intervention of a philosopher from the start."[13]

So, I think Earman's proposal a sound one, an echo in fact of similar words that the physicist Kip Thorne wrote (in the Foreword to the second edition of *Time Machines*) concerning science fiction writers: "Smart physicists seek insight everywhere, including from clever science fiction writers who long ago began probing seriously the logical consequences that would ensue if the laws of physics permitted time travel."[14]

To emphasize this new, combined, diversified focus (but also to retain some connection with my earlier books) is the reason I have altered the title, just a bit. In addition, each chapter now concludes with several open-ended questions, suitable for motivating either classroom discussions or more extensive essay responses.

---

[11] J. J. C. Smart, "The Unity of Space-Time: Mathematics Versus Myth Making," *Australasian Journal of Philosophy*, (no. 2) 1967, 214–217.

[12] M. Hollis, "Times and Spaces," *Mind*, October 1967, pp. 524–536. Hollis ends by saying he is prepared to accept the failure of his paper to convince many of his colleges to change their ways, and he is waiting for one of them to write a paper opening with "Twice upon a time in another space no distance in any direction from here . . ."!

[13] Quoted from the Introduction to Milena Wazeck, *Einstein's Opponents: the public controversy about the theory of relativity in the 1920s*, Cambridge 2014 (published in German in 2009).

[14] Thorne is Professor Emeritus of Physics at the California Institute of Technology.

Teachers, in particular, may find this a useful feature if using the book in an academic setting. The book ends with reprints of two of my own published time travel stories (one from *Analog* and the other from *Omni*), with each serving as an illustration of technical issues raised in the book. From my own teaching of an undergraduate honors class in time travel at the University of New Hampshire, I think the assigning of story writing to be an excellent tool for teachers to use. I found reading student stories to be a lot of fun, and students may well surprise teachers with innovative ideas.

Now that I've mentioned story writing, let me say something about the heavy presence of time travel science fiction stories in this book, the majority of which originated in the often maligned pulp magazines of the 1920s through the 1950s. 'Pulp' has long been burdened with a bad literary reputation. As the editor of one anthology of pulp fiction bluntly put it, "Pulp equated with rubbish. Crap of the basest nature."[15] Part of the reason for that was cosmetic; as I wrote in an earlier book, "The term *pulp* came from the use of inexpensive wood-pulp—you could *feel* the lumpy wood chips in each ragged, untrimmed page—to make paper that was far too crummy for the use by any publisher of 'words meant to last.' Such paper quickly yellowed, turned brittle, and finally, amid billowing clouds of bits and pieces, entered into eternal oblivion. Think of the paper used in your newspaper before its final contribution to civilization in the bottom of your cat's litter box; pulp was worse."[16]

And then a little later, in the same book, "The stories in *Amazing* [*Stories* magazine] were 'read it in the morning, forget it by dinnertime' adventure fiction, the stuff you'd put inside a newspaper if on a crowded train or bus so fellow passengers wouldn't know what a low-grade mind you had. The transient nature of pulp fiction was independent of its literary quality, as the cheap acid-based paper that the stories were printed on began to oxidize and literally burn-up as soon as it rolled off the press. In the introductory essay to a 1950 collection of pulp-detective Philip Marlowe stories (*Trouble Is My Business*), mystery writer Raymond Chandler commented on this when he wrote 'pulp fiction never dreamed of posterity.' Pulp fiction was synonymous with trash fiction, and the nature of much of early pulp SF has been aptly described as 'scientific pornography for the mechanically minded,' and 'writing which drooled over descriptions of technology.'"

When publisher Hugo Gernsback (1884–1967) brought out the first issue of *Amazing Stories* in April 1926, it was the first pulp devoted totally to science fiction. With its masthead motto of "Extravagant Fiction Today—Cold Fact Tomorrow," and with the illustration on the contents page of each issue showing a muscular Jules Verne bursting from his grave in the heroic, up-up-and-away pose made famous years later by Superman, there could be no doubt as to what kind of fiction the reader would find under the dramatic, multi-colored cover art. It was fiction populated with mad scientists, and half-naked woman about to be ravished by alien invaders from outer space; all in all, stuff of interest only to teenage boys and imbecilic adults.[17]

---

[15] Maxim Jakubowski, *The Mammoth Book of Pulp Fiction*, Carroll & Graf 1996.

[16] P. J. Nahin, *Holy Sci-Fi!: where science fiction and religion intersect*, Springer 2014.

How else, after all, to explain the publication of one tale[18] that was given the following heart-stopping editorial introduction: “Professor Lambert deliberately ventures into a Vibrational Dimension to join his fiancée in its magnetic torture-fields”? In defense of many of the readers of early pulp science fiction, however, not all were attracted by such nonsense. Just 2 months later (June 1931) one reader wrote to the same magazine to complain of masculine heroes saving weeping women from ungodly horrors: “Just why do you permit your Authors to inject messy love affairs into otherwise excellent imaginative fiction? Just stop and think. Our young hero-scientist builds himself a space flyer, steps out into the great void, conquers a thousand and one perils on his voyage and amidst our silent cheers lands on some far distant planet. Then what does he do? He falls in love with a maiden—or it’s usually a princess—of the planet to which the Reader has followed him, eagerly awaiting and hoping to share each new thrill attached to his gigantic flight. But after that it becomes merely a hopeless, doddering love affair ending by his returning to Earth with his fair one by his side. Can you grasp that—a one-armed driver of a space-flyer! . . . We buy A.S. for the thrill of being changed in size, in time, in dimension . . . not to read of love . . . I wish . . . for plain, cold scientific stories sans the fair sex.”

Here’s another example, this one of the sort of tale that gave an aroma of the sophomoric to ‘golden age’ time travel science fiction. It was a story of a young man of the far future, with access to a time machine, who wants to see a dinosaur before he dies. So back he travels, back, back, until he at last finds himself in a “subterranean cave, dark and foul-smelling.” At first he is puzzled (did dinosaurs live underground?), but then suddenly he hears a thundering roar and sees a huge black shape in the gloom. There can be no doubt now; it *is* a dinosaur, and he sees its red, gleaming eyes just as it crushes him into a pancake. But that’s okay; he saw a dinosaur before he died. Then comes the dénouement. He hadn’t really gone back quite as far as the Jurassic period, but only to the twentieth century, where he has been run down by the local express train in a subway tunnel![19]

---

[17] This was particularly thought to be the case for readers of the romance pulps, written for young women in the 1930s and 1940s (a separate and distinct audience from that of the science fiction pulps). As one commentator wrote on that genre, the heroes and heroines of such tales often displayed the “mental equipment of a banana split,” with the implication that the same might be said of the readers, themselves. (See Margaret MacMullen, “Pulps and Confessions,” *Harper’s Monthly Magazine*, June 1937.) I don’t think, however, that this particular complaint generally applied to the pulp science fiction readership. I’ll have much more to say about Gernsback and early pulp science fiction speculations concerning time travel, in Chap. 4.

[18] T. Curry, “Hell’s Dimension,” *Astounding Stories*, April 1931.

[19] R. G. Thompson, “The Brontosaurus,” *Stirring Science Fiction*, April 1941. In the editors of *Stirring*’s defense, notice the month: maybe this story was *meant* to be a joke. If so, it was an admirable success.

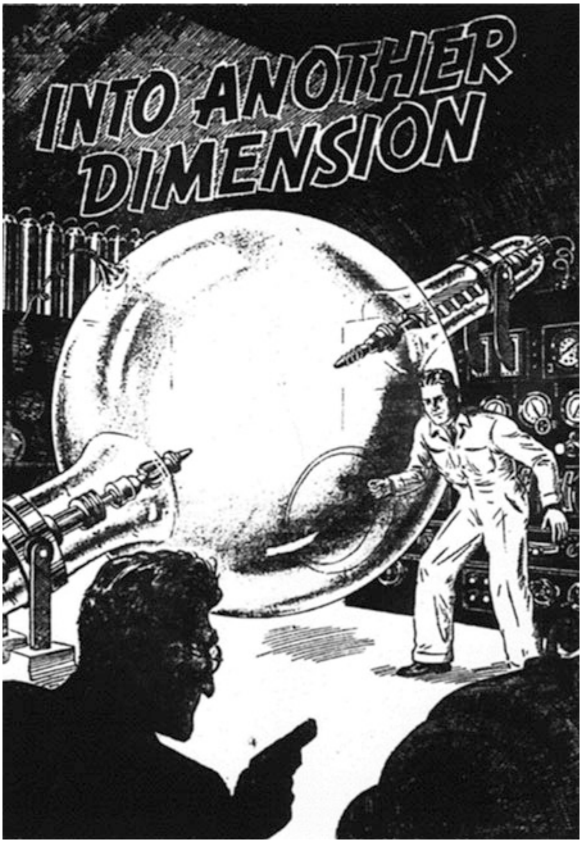

Vibrating into new dimensions was, apparently, a popular idea in 1930s pulp science fiction. This 'super science' gadget operated by vibrating an object faster than light, whereupon the Lorentz-FitzGerald contraction formula (see Chap. 3) predicts an *imaginary* size for the object—which means (so we are told) that the object has entered "another plane of existence." The inventor (the fellow with the gun) is inviting his grim-faced assistant to give the gadget a try. The original caption reads "Get into that vibrator! Get in, I say!"

Illustration for "Into Another Dimension" by Maurice Duclos, *Fantastic Adventures* November 1939 (art by Kenneth J. Reeve), © 1939 by Ziff-Davis Publishing Co., reprinted by arrangement with Forrest J. Ackerman, Holding Agent, 2495 Glendower Ave., Hollywood, CA 90027

Today, however, the need to apologize for science fiction tales about time travel isn't quite so necessary. Now and then, in fact, you'll even find one of the better pulp stories cited in highly mathematical papers on time machines in the *Physical Review D*, one of the most important scholarly physics journals. Even those physicists and philosophers who mostly ignore science fiction—except perhaps to make slightly condescending remarks—would, if honest, admit that their early teenage interest in time travel was sparked by reading a really good science fiction story, and not by working their way through a physics textbook. Yes, when the physics eventually came later, it was very good—but the science fiction came *first*, and it was pretty good, too.[20] It's in a 1937(!) tale, for example, that we find the claim for consistency around a closed loop in time, *decades* ahead of the physicists and philosophers.[21] And when you get to the final section of Chap. 4, I think you'll find it difficult to believe that Everett's many-worlds interpretation of quantum mechanics, dating from the late 1950s (which avoids the standard paradoxes of time travel) wasn't inspired by some youthful reading of science fiction from the 1930s and 1940s.

In a number of places in this book you'll find my comments on how science fiction has occasionally anticipated physicists on the subject of time machines and time travel. This is *not* to be interpreted as some sort of 'gotcha' in favor of science fiction. Far from it. When push comes to shove, physics *always* wins. This situation was specifically addressed by Joe Haldeman, in an afterword to his 2007 novel *The Accidental Time Machine*. There he wrote, about when he started in 1971 to write his earlier, now classic novel *The Forever War*, "I needed a way to get soldiers from star to star within a human lifetime, without doing too much violence to special and general relativity. *I waved my arms around really hard* [my emphasis] and came up with the 'collapsar jump'—at the time, collapsar was an alternate term for 'black hole,' though I was unaware of the latter term [because John Wheeler had invented it only 4 years before, as discussed in Chap. 1 and note 106]." And then Haldeman admitted "It's a truism of science fiction that if you predict enough things, a few of them are going to come true. . . . What I think it actually demonstrates is that *if you wave your arms around hard enough* [my emphasis], sometimes you can fly."

Now, there is one feature common to all books on time travel to the past (which is the central topic treated here, of course) that I would like to clearly state. It's obviously a subject of vast interest to physicists, and yet it offers (as far as I know) *absolutely no hope of suggesting even a single experiment for study*. (As far as I know, nobody is building a time machine in their basement.) A suggestion *has* been made that it may be possible to detect, in the *present*, the effects of the *future* operation of "man-made time machines, which could be of a size traversable by humans," that is, machines with a 1-m spatial extent offering a one second trip into the past.[22] With the best technology available today, however, the calculated effects on the proposed two-particle scattering experiment are orders of magnitude too small to measure.

---

[20] The view expressed by Vladimir Voinovich's time traveler in his 1986 novel *Moscow 2042* (Science fiction . . . is not literature, but tomfoolery like the electronic games that induce mass idiocy.) is, I think, wrong. For an interesting presentation on the role of science fiction in exciting an interest in science among youngsters, see the paper by Frederik Pohl (1919–2013), "Science Fiction: the stepchild of science," *Technology Review*, October 1994, pp. 57–61. In this essay Pohl, a well-known writer of science fiction and editor of *Galaxy Science Fiction* and *If* magazines, writes "Science fiction is [the ultimate protection] against future shock . . . if you read enough of it, nothing will take you entirely by surprise." Not even time travel.

[21] P. S. Miller, "The Sands of Time," *Astounding Stories*, April 1937.

[22] S. Rosenberg, "Testing Causality on Spacetimes with Closed Timelike Curves," *Physical Review D*, March 15, 1998, pp. 3365–3377.

This situation is really unprecedented in the history of science.[23] To cynics, it may seem to be a bit like writing learned papers on the thermodynamics of fire-breathing dragons (which, like other mythological entities—and time machines, too—have yet to be seen)! This one fact has opened the doors—and has kept them open for decades—for philosophers and science fiction writers, who can endlessly debate back and forth on all aspects of time travel to the past with nary a single experimental fact to complicate their lives. For physicists the situation is naturally frustrating, but for philosophers and science fiction writers it's a dream come true. This isn't to say it's *all* basically theological in nature. Both the physicists and the philosophers *have* written many fascinating papers and books and, of course, so have science fiction writers. Mathematical physics *has* been advanced.

Still, despite all of the theoretical work done in the last 30 years, work that has made it reasonable to seriously talk of 'time travel' and 'time machines,' I suspect many would nonetheless agree with these words from more than 75 years ago: "Of all the fantastic ideas that belong to science fiction, the most remarkable—and, perhaps, the most fascinating—is that of time travel . . . Indeed, so fantastic a notion does it seem, and so many apparently obvious absurdities and bewildering paradoxes does it present, that some of the most imaginative students of science refuse to consider it as a practical proposition."[24] For some, time travel is an even more unlikely possibility than (as declared by Robert Lewis Stevenson) is the "welding of ice and iron." Not all physicists and philosophers view the time travel/paradox arguments as convincing, however. Provocative, yes, of course, but many are not yet prepared to write 'signed, sealed, and delivered' at the end.

So, keep reading and I think you'll discover why there *are* those who are not so quick to dismiss the possibility of following the fantastic world line of H. G. Wells' intrepid Time Traveller[25] into the future. And, just maybe, into the distant past, too.

---

[23] Perhaps, however, I am too hasty. More recent theoretical calculations suggest that wormholes connecting our universe with other universes would, after converting into time machines, have characteristic thermal signatures. See P. F. González-Díaz, "Thermal Properties of Time Machines," *Physical Review D*, 2012, pp. 105026-1 to -7 which, however, concludes that a search for such signatures would be "quite difficult [with the] instruments available."

[24] I. O. Evans, "Can We Conquer Time?" *Tales of Wonder*, Summer 1940.

[25] The Time Traveller is never named in Wells' 1895 novel *The Time Machine*. An earlier (1888) attempt at a time machine story, with the awful title *The Chronic Argonauts* (the "chronic" was apparently inspired by the word *chronology*), so embarrassed Wells that he later called it "imitative puerile stuff," "clumsily invented, and loaded with irrelevant sham significance," and "inept," and so he hunted down and destroyed every copy of it that he could find. You can find *The Chronic Argonauts* reprinted in *The Definitive Time Machine* (H. M. Ceduld, editor), Indiana University Press 1987. The hero in that work *was* named: Dr. Moses Nebogipfel. There is one passage in *The Time Machine* that does tantalize; as the Time Traveller explores a museum of "ancient" artifacts in the Palace of Green Porcelain (they are, of course, artifacts of our *future*) he reveals that "yielding to an irresistible impulse, I wrote my name upon the nose of a steatite monster from South America that particularly took my fancy." Thus, the Traveller *has* given his name, but his signature exists only in the future, in a museum of the past that is yet to be built.

For Further Discussion

For time travel to the past to make any sense, the past must in some sense 'still be there.' This is a concept that we'll find later in the book to have significant support in relativistic physics, but for now let's limit ourselves to a purely romantic view. As an example of this, consider this passage by Canadian writer Grant Allen (1848–1899), from the Introduction to his 1895 time travel novel *The British Barbarians*: "I am writing in my study on the heather-clad hill-top. When I raise my eye from my sheet of foolscap, it falls upon miles and miles of broad open moorland. My window looks out upon unsullied nature. Everything around is fresh and pure and wholesome . . . But away below in the valley, as night draws on, a lurid glare reddens the north-eastern horizon. It marks the spot where the great wen of London heaves and festers." I personally find it quite tempting to imagine Allen somehow still there in his study of 1895, and of heaving and festering late-Victorian London, too, with H. G. Wells himself in the middle of it, still reading the first rave reviews of *The Time Machine*. In Wells' novel The Time Traveller journeys into the far future, while in Allen's work the protagonist is a twenty-fifth century anthropologist who has traveled back to the past of the late nineteenth century to study the 'British barbarians.' Read Allen's novel (it's available on the Internet, for free, as a Project Gutenberg book) and comment on the significance of its appearance at virtually the same time as Wells' great work. Why do you think Wells' novel is remembered, and Allen's is not?

In the opening paragraph of his paper "The Conundrum of Time Travel" (*Croatian Journal of Philosophy*, No. 37, 2013, pp. 81–92), Anguel Stefanov writes "Needless to say . . . the problems concerning time travel are being still tackled by science fiction only, but resolved by science proper neither theoretically, nor practically." Do you think this is correct?

Eventually every genre of writing becomes the target for parody, in which the *form* of the genre serves as the framework for what (it is hoped) is a humorous mockery. The most famous example of this, perhaps, is the annual

(continued)

Edward Bulwer-Lytton contest in writing a take-off on the long-winded opening line of the 1830 novel *Paul Clifford*, by Bulwer-Lytton (1803–1873). That opening line *is* a wonder (a masterpiece of purple prose): “It was a dark and stormy night; the rain fell in torrents—except at occasional intervals, when it was checked by a violent gust of wind which swept up the streets (for it is in London that our scene lies), rattling along the housetops, and fiercely agitating the scanty flame of the lamps that struggled against the darkness.” Here’s a recent (from the 2015 contest) spoof: “The Contessa’s heart was pounding hard and fast, like an out-of-balance clothes washer, which can get that way if you mix jeans with a lot of light things, though the new ones have some sensor thing to counteract that or shut off, but the Contessa’s heart didn’t have anything like that, so she had to sit down and tell Don Rolando to keep his hands to himself for a while.” Science fiction isn’t immune to such fun, and a good example of that can be found in the September 14, 2015, issue of *The New Yorker*, which has (on p. 50) “Eight Short Science-Fiction Stories” by Paul Simms. Here’s the one I laughed hardest at: “The Gene-Splicers had tinkered with the DNA, producing a race of warriors who craved just two things: the thrill of battle and the taste of their own feet. They hungered for battle. They literally ate their own feet. None survived to reproduce, and within a few short years they were all gone. The Gene-Splicers chalked it up to experience, and decided to try harder the next time.” That, and the other seven spoofs by Simms, cut across a wide swath of science fiction, but one theme noticeably absent was that of time travel. Try your hand at writing a *short* (fewer than 500 words) time travel spoof, and be prepared to read it aloud to an audience of your peers.

The tale “Through the Dragon Glass” by Abraham Merritt (1884–1943) appeared in the early pulp magazine *All-Story Weekly* of November 24, 1917. It described the discovery of a passage through an ancient Chinese mirror into an alternate world. One *might* think of this as an early conception of a wormhole, but more likely it may remind you mostly of Lewis Carroll’s *Through the Looking Glass*. More interesting for us, in this book, is a story written 75 years ago that describes a gadget connecting two regions of spacetime, with a time shift of a week between the two regions. (See “Time Locker” by Lewis Padgett, in the January 1943 issue of *Astounding Science Fiction*.) The gadget falls into the hands of a crooked lawyer who, not understanding what he has, ends up accidently killing himself. As the story ends, the inventor of the gadget ruefully muses to himself that the lawyer

(continued)

"must have been the only guy who ever reached into the middle of next week—and killed himself!" The gadget is, in everything but name, a wormhole time machine. Speculate on how such a spacetime structure could appear in a *science fiction magazine*(!) decades before there was any discussion of such a possibility in the physics world.

Adjectives used to describe many of the stories in the science fiction pulps included *primitive*, *trashy*, *tawdry*, *silly*, *absurd*, *crummy*, *ludicrous*, and *cheap*. One early pulp magazine actually boasted, of its contents, that they contained "sensational fiction with *no* philosophy." Speculate on how such a low-level 'literary' form could have been so successful in finding an enthusiastic audience for time travel paradox tales, tales that are in fact by their very nature simply *stuffed* with philosophical issues. As an example of the tremendous emotional power a particularly well-written time travel story can deliver, read Isaac Asimov's "The Ugly Little Boy" (*Galaxy Science Fiction*, September 1958). Asimov rated this story as among his most favorite of all the many he wrote. If you can read it without ending in tears, well, ... An excellent modern historical work on the pulps (of all genres, not just science fiction) is by Lee Server, *Danger Is My Business: an illustrated history of the fabulous pulp magazines, 1896–1953*, Chronicle Books 1993.

A literary fascination with time was already 'in the wind' when Wells wrote his *Time Machine*, as with Oscar Wilde's 1890 novel *The Picture of Dorian Gray*. Even decades earlier than that one can find a hint of time travel of a sort in Edgar Allen Poe's 1841 short story "Three Sundays in a Week." And just 4 years later Henry Wadsworth Longfellow wrote his haunting poem "The Old Clock on the Stairs," with these opening words:

Somewhat back from the village street
Stands the old-fashioned country-seat.
Across its antique portico
Tall poplar-trees their shadows throw;
And from its station in the hall
An ancient timepiece says it all,—

"Forever—never!

(continued)

Never—forever!"

The most interesting of all pre-Wells time travel fiction to appear in a mass-audience publication was, I think, the short story "The Old Folks Party" by Edward Bellamy, printed in the March 1876 issue of *Scribner's Monthly*. In this story a group of teenagers, who belong to a weekly discussion club, agree that at their next meeting they will all come dressed and behaving as they believe they will be dressing and behaving 50 years in the future. Also attending will be the grandmother of one of the young ladies. The meeting of the "old folks" takes place, and it invokes such powerful feelings of mortality that, at last, one of the young men can stand it no more: "Suddenly Henry sprang to his feet and, with the strained, uncertain voice of one waking himself from a nightmare, cried:—'Thank God, thank God, it is only a dream,' and tore off the wig, letting the brown hair fall about his forehead. Instantly all followed his example . . .." The young people then began to laugh with relief at once again being young, until they notice the grandmother is crying. Her granddaughter instantly knows what is wrong and says, "Oh, grandma, we can't take you back with us." Read, compare, and contrast, these works by Wilde, Poe, Longfellow and Bellamy, with the 'scientific' presentation of time travel by Wells.

Acknowledgements

My name is on *Time Machine Tales* as the author, but there are many others without whose help I could not have written it. For this, the revised and updated third edition of the original 1993 book *Time Machines*, I am enormously grateful to the publisher, Springer Science + Business Media, and to my physics editors at Springer, Dr. Sam Harrison and his assistant Ho Ying Fan in New York City. They made an intellectually demanding and inherently lengthy project (one that could easily have turned into a nightmare) a pleasant one. Springer's New York City-based editorial assistant Irene Bruce smoothly handled the administrative details of transforming the book from typescript to print. And to Serguei Krasnikov in Russia, Frank Arntzenius and Roberto Casati in England, Francesco Gonella in Italy, and Geoff Goddu and Archille Varzi in America, I offer grateful thanks for permission to reprint from their publications.

But the story of this book actually begins long before 1993.

Nearly four decades in the past, in 1979, when time travel was still mostly just a nutty idea used by science fiction writers and a few rogue philosophers—proposing to give a serious seminar on time travel at the weekly college *physics* seminar would almost certainly have resulted in getting the bum's rush out the nearest door, or maybe even tossed through the closest window (or, God forbid, *off the roof*!)—I wrote my first time travel tale for *Analog Science Fiction* magazine. Bought by then editor Ben Bova (you can read it in Appendix A), he also bought another time travel tale from me soon after he had become the fiction editor at the newly created *Omni Science Fiction* magazine (you can read it in Appendix B). (Years later, Ben's successor as *Omni*'s fiction editor, Ellen Datlow, bought a third time travel story from me—"The Invitation"—which you can read if you can locate the July 1985 issue.)

Those fictional experiences, and my discovery of Princeton philosopher David Lewis' seminal 1976 essay on the *logical* possibility of time travel, led me to plunge into a deeper study of the *physics* of time travel, to the point where as the 1980s ended (1988, to be precise) I thought a book should be my next writing project. Alas, nobody else shared that thought. I spent the next 3 years looking for a

publisher among numerous university presses; as I looked, the pile of rejection letters, like entropy, grew steadily larger.

The editor at one well-known university press, in fact, simply laughed at the idea of a scholarly time travel book when I called him on the telephone (e-mail was then still in its infancy) to ask, after a long period of no response to my written proposal, if he had gotten it. (Perhaps, I told myself, the mail truck carrying my precious document had fallen into a hyper-dimensional spacetime warp: how pathetically desperate is the anxious academic writer!) Later, I published a number of math/physics books with that same press (but now with new editors) and so there were no lasting hard feelings.

That laughing, unresponsive editor wasn't alone in his opinion, I have to admit, and it wasn't until 1991 when Maria Taylor, the publisher at the Press of the American Institute of Physics (AIP Press), decided to take a chance (a *big* chance) and publish the first edition of *Time Machines*. She made that decision in large part because of two extremely supportive reviews of my proposal from the academic physicists Edwin Taylor (MIT, and a former editor of *The American Journal of Physics*) and Gregory Benford (University of California, Irvine, as well as being an award-winning writer of science fiction who occasionally used the time travel theme). After the original *Time Machines* appeared, Ben Bova asked me to use it as a guide to writing a book on time travel for the new Writer's Digest series on science fiction that he was editing, aimed specifically at would-be story writers; that book came out in 1997 as *Time Travel*. Later, Trevor Lipscombe, then editor-in-chief at the Johns Hopkins University Press (and my former math editor at Princeton University Press), reprinted *Time Travel* (with a new Preface) in 2011.

After Springer acquired AIP Press (and all of its books) in 1995, and Maria Taylor herself went over to Springer, she and I collaborated again on bringing out the second edition of *Time Machines* in 1999. And now, 18 years later, here (with a title slightly altered for reasons I give in "Some First Words") is its successor.

To Sam, Ho Ying, Irene, Ben, Ellen, Maria, Edwin, Gregory, and Trevor, thank you for your support. But my greatest debt of all, one I can never even begin to repay, is to my wife of 55 years, Patricia Ann, a woman of infinite tolerance. Who else would put up with someone who plays first-person-shooter video games at midnight (the bigger and the louder the explosions, the better!) and writes books on time machines, all the while claiming not to be crazy?

University of New Hampshire, Exeter, NH, USA Paul J. Nahin

Introduction

> Over the last few years leading scientific journals have been publishing articles dealing with time travel and time machines. . . . Why? Have physicists decided to set up in competition with science fiction writers and Hollywood producers?
>
> —John Earman (see note 25 of *Some First Words*)

Writing about time travel is, today, a respectable business. It hasn't always been so. After all, time travel, *prima facie*, appears to violate a fundamental law of nature; every effect has a cause, with the cause occurring before the effect. Time travel to the past, however, seems to allow, indeed to *demand*, backwards causation, with an effect (the time traveler emerging into the past as he exits from his time machine) occurring *before* its cause (the time traveler pushing the start button on his machine's control panel years *later* to start his trip backward through time).

Thus, when H. G. Wells published his breakout masterpiece, *The Time Machine*, in 1895, even those readers who loved it as a *story* (and not all did) were still quick to dismiss it as a *romantic fantasy*. It was, in their view, certainly an emotionally powerful tale of pure imagination, but nothing more. Reviewers of the day used such words as "hocus-pocus" and "bizarre," and called the work a "fanciful and lively dream."[26] Any one of the novels by Wells' contemporary, Jules Verne (even such super-technology ones like the 1865 *From the Earth to the Moon*) would have been ranked *far* above Wells' novella in terms of 'it could actually happen.'

Wells himself always denied that his time machine was anything more than a literary device[27] to get his Time Traveller into the far future. Indeed, in 1934, in the preface to *Seven Famous Novels* (published by Knopf), a collection of his novel-length scientific romances (as science fiction had been known before the term *science fiction* came into use), including *The Time Machine*, Wells made his position perfectly clear: "These stories of mine collected here do not pretend to deal with possible things; they are exercises of the imagination ... They are all fantasies; they do not aim to project a serious possibility; they aim indeed only at the same amount of conviction as one gets in a good gripping dream." Wells then went on to say in that same preface that all attempts before at writing fantastic stories depended on magic. But not in his works. "It occurred to me that instead of the usual interview with the devil or a magician, an ingenious use of scientific patter might with advantage be substituted." Wells' great contribution to time traveling story-telling was his introduction of a *machine*; science instead of magic, drugs, dreams, blows on the head, or suspended animation.[28] Not all modern science fiction writers have followed Wells' lead, however.

---

[26] These reviews are reprinted in P. Parrinder, *H. G. Wells: The Critical Heritage*, Routledge & Kegan 1972. A modern reviewer has applied such negative characteristics to the Time Traveller, himself, calling him "a kind of Trickster figure" and "a quack and magician." See Robert J. Begiebing, "The Mythic Hero in H. G. Wells's *The Time Machine*," in *Essays in Literature*, Fall 1984, pp. 201–210.

[27] Wells was not the first to use a machine to enable time adventures, as the Spanish writer Enrique Gaspar (1842–1902) used one in his 1887 story *The Time Ship: A Chrononautical Journey*. It's Wells' tale we remember, however.

A science fiction tale by Clifford Simak (1904–1988), for example, the 1978 novel *Mastodonia*, incorporates an alien creature marooned on Earth (because of a spaceship crash centuries earlier) who 'makes time tunnels.' One of the characters in the story, who is attempting to start a time-travel agency using these tunnels, explains why not having a time *machine* is causing her difficulties with prospective clients: "The whole trouble was that I couldn't tell them about some machine—a time-travel machine. If I could have told them we'd developed a machine, they'd have been more able to believe me. We place so much trust in machines; they are magic to us. If I could have outlined some ridiculous theory and spouted some equations at them, they would have been impressed." I think that's off the mark. We trust in machines not because they are magic, but for precisely the opposite reason. They are *not* magic, but rather are *rational*. And to dismiss mathematics is to say that some non-natural—some supernatural—influence is at work.

But is a time *machine* actually possible? Or is the idea of a time machine simply "Nonsense" and "A bilgeful of crap," as a character bluntly puts it in the 1972 novel *The Dancer from Atlantis* by Poul Anderson (1926–2001). Wells, himself, addressed this point in an autobiographical essay (published in the *Cornhill Magazine*) that he wrote in July 1945 (just 13 months before his death) in even blunter words. Writing under the name of "Wilfred B. Batterave," he penned a very funny summary of his life titled "A Complete Exposé of This Notorious Literary Humbug." There he described *The Time Machine* as "[A] tissue of absurdities in which people are supposed to rush to and fro along the 'Time Dimension.' By a few common tricks of the story-teller's trade, Wells gets rid of his Machine before it can be subjected to a proper examination. He cheats like any common spook raiser. Otherwise it is plain commonsense that a man might multiply himself indefinitely, pop a little way into the future and then come back. There would then be two of him. Repeat *da capo* and you have four, and so on, until the whole world would be full of the Time Travelling Individual's vain repetitions of himself. The plain-thinking mind apprehends this in spite of all the Wellsian mumbo-jumbo and is naturally as revolted as I am by the insult to its intelligence." Funny, yes, but still pretty harsh stuff.

---

[28] Examples of 'non-machine' time travel stories of the last four types are, respectively, H. G. Wells' "The New Accelerator" (1901), Charles Dickens' *A Christmas Carol* (1843), Mark Twain's *A Connecticut Yankee in King Arthur's Court* (1889), and Edward Bellamy's *Looking Backward, 2000–1887* (1888).

As one writer has argued,[29] Wells was, rather than presenting a scientific discovery, simply attempting to refute the nearly suffocating, unjustified (in his mind), smug optimism of the well-to-do of the Late Victorian Age. And so, on his journey to the year A.D. 802,701, the Time Traveller finds the awful decay of humanity in the cannibalistic subjugation of the Eloi by the Morlocks, the end result of class warfare between the working class (Morlocks) and the idle, parasitic upper class (Eloi).

The German social philosopher Karl Marx, if he hadn't already been dead for 12 years in 1895, would surely have nodded in vigorous agreement as he read *The Time Machine*, even as he would have regretted Wells' decision to have the victory of oppressed workers take so long. (What irony that he is buried in London's Highgate Cemetery, the Victorian Valhalla where he has spent the last century and more quite literally mingling with many of the capitalistic ancestors of the Eloi!) What Marx would have thought of *time travel* as a possibility is, however, far less certain.

How things changed in the years that followed *The Time Machine*. There was, at first, admittedly a 'slight' decline in literary merit as the newly developing pulp science fiction magazines picked-up and ran with the time travel genre. Many of the magazine time travel tales of the 1920s, 1930s, and 1940s were, frankly, simply awful. BUT—some were pretty good, too. And some were, in fact, *very* good. From the 1950s on, there have been ever more sophisticated time travel tales from ever more sophisticated writers.

In the academic communities of philosophers and physicists, too, big events occurred. I give the philosophers the edge, in fact, with the 1976 publication of a hugely important paper that opened with these dramatic words: "Time travel, I maintain, is possible. The paradoxes of time travel [to the past] are oddities, not impossibilities. They prove only this much, which few would have doubted: that a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours."[30] That writer wasn't the first philosopher to write on time travel to the past, but none had expressed themselves in such powerful and unequivocal words in unmistakable support of the concept.

---

[29] R. M. Philmus, "*The Time Machine*; Or, the Fourth Dimension as Prophecy," *Publications of the Modern Language Association*, May 1969, pp. 530–535.

[30] David Lewis, "The Paradoxes of Time Travel," *American Philosophical Quarterly*, April 1976, pp. 145–152. Lewis (1941–2001) was a Princeton University philosophy professor.

Lewis' paper is also notable because it gives what seems to be a clear definition of just what it means to say one has 'traveled in time,' either to the past *or* to the future:

> What is time travel? Inevitably, it involves discrepancy between time and time. Any traveler departs and then arrives at his destination; the time elapsed from departure to arrival (positive, or perhaps zero) is the duration of the journey. But if he is a time traveler, the separation in time between departure and arrival does not equal the duration of the journey.

To understand this, we need to appreciate the distinction between the *personal time* of the time traveler and the *external time* of remote observers of the time traveler. A time traveler's personal time is measured, for example, either by the time kept by his wrist watch or, perhaps, by a burning candle. (This distinction had actually appeared earlier in Horwich's paper—see note 27 in *Some First Words*—published the year before Lewis' paper.)

I say I 'give the edge to the philosophers' because, while the first *physics* time travel paper had appeared decades earlier, its author wasn't really a physicist at all but rather was Einstein's friend, the world-famous mathematical logician Kurt Gödel. Gödel's paper was, in retrospect, a pivotal event in establishing the 'respectability' of *scientific* time travel; it's worthwhile to take some time here to explain this important point. For physicists (and for philosophers and science fiction writers, too) a 'time machine,' one either constructed by intelligent beings or occurring naturally, manipulates (all the while obeying the known laws of physics) finite amounts of matter and energy in a finite region of spacetime.[31] A 'time machine' would be declared to be *plausible* if it could be explained by a rational, scientific theory. Such a rational theory is found in Einstein's general theory of relativity. (His *special* theory of relativity applies in those situations where there is no gravity.)

Until Einstein, the theory of gravity used by scientists was Newton's—a theory that, although amazingly accurate for any situation encountered on Earth, does have observable errors in certain astronomical applications. In addition, Newton's theory is a descriptive one; it makes possible the calculation of gravity effects without offering any explanation for gravity itself. Einstein's theory not only gives the right answers, even in those cases where Newton's theory doesn't, but it also explains gravity. It does that by treating the world as a four-dimensional structure in which all four dimensions (three of space and one of time) are in a certain sense on equal footing. The resulting Einsteinian description of the world is that of a unified spacetime in which time and space are intimately intertwined, whereas Newton's theory keeps time and space separate and distinct.

---

[31] I am going to feel free to use words like *spacetime* without having to first write introductory essays on relativity theory and tensor mathematics, because such words have entered common use. All those Hollywood science fiction movies, even the crummy ones that routinely trash the laws of physics, have at least expanded the general imagination!

As Newton wrote of time, at the start of his 1687 masterpiece *Principia*, a work that revolutionized physics, “Absolute, true, and mathematical time, of itself, and from its own nature, flows equably without relation to anything external, and by another name is called duration.” This view of time would be, of course, discarded with the arrival of Einstein and his view of variable time depending on the state of the observer.

Unlike Einstein’s view, Newton’s view of the nature of time was entangled with theology. As one modern theologian has written, “Newton conceived of absolute time as grounded in God’s necessary existence.”[32] To quote Newton himself, in the *General Scholium* to the second edition of *Principia* (1713) he added words that didn’t appear in the original: “God is a living, intelligent, and powerful Being; and, from his other perfections, [it follows] that he is supreme, or most perfect. He is eternal and infinite, omnipotent and omniscient; that is, his duration reaches from eternity to eternity; his presence from infinity to infinity; he governs all things, and knows all things that are or can be done. He is not eternity and infinity, but eternal and infinite; he is not duration or space, but he endures and is present. He endures forever, and is everywhere present; and, by existing always and everywhere, he constitutes duration and space. Since every particle of space is *always*, and every indivisible moment of duration is *everywhere*, certainly the Maker and Lord of all things cannot be *never* and *nowhere*.”

Okay, I’ll be honest—I really am not at all sure just what that means! Newton added these words to the *Principia* in response to criticism (from the influential philosopher George Berkeley (1685–1753)) that his original statements about absolute time were “pernicious and absurd notions,” notions that were in fact atheistic in conception. That was a most serious charge in Newton’s day, and he was trying (I think) to find some cover from those critics who spent more hours of the day thinking about God than of physics. Much more honest (in my opinion) are the witticisms ‘time is just one damn thing after another’ and ‘time is what keeps everything from all happening at once.’[33] More funny than useful, yes, of course, but at least they’re funny.

Newton’s theological view of time is simply irrelevant to the modern physicist (although perhaps of more interest to the philosopher-historian) but in many cases it is of *central interest* to the science fiction writer. For example, Newton’s religious mindset and its (perhaps!) connection with time travel is treated in my short story "Newton's Gift," originally published in *Omni Magazine* (January 1979) and reprinted in Appendix B at the end of this book. Wells' Time Traveller's view of time is more Newtonian than it is Einsteinian—and perhaps that's not such a big surprise, considering that Einstein was only 16 years old when *The Time Machine* was published.

---

[32] William Lane Craig, “God and the Beginning of Time,” *International Philosophical Quarterly*, March 2001, pp. 17–31, which discusses the question ‘Why didn’t God create the world sooner?’ One irreverent answer is ‘He was busy creating Hell for all those who ask that question,’ but a more scholarly analysis can be found in Brian Leftow, “Why Didn’t God Create the World Sooner?” *Religious Studies*, June 1991, pp. 157–172.

[33] This last ‘definition’ first (as far as I know) appeared in the work of the science fiction writer Ray Cummings (1887–1957), in his 1921 story “The Time Story,” published in *Argosy-All-Story* magazine. He repeated the phrase in his 1929 novel *The Man Who Mastered Time*, and then again in the 1946 novel *The Shadow Girl*. (“This same Space; the spread of this lawn . . . what would it be in another 100 years? Or a 1000? This little space, from the Beginning to the End so crowded with events and only Time to hold them apart!”)

From the first (1905) it has been known that Einstein's special theory allows time travel into the future via the well-known mechanism of *time dilation*. (The faster a rocket ship travels relative to Earth, the slower is the tick-tock of a wrist watch worn by a rocketeer, compared to that of an identical watch back on Earth.)[34] To return from the future, however, to travel back into the past to the instant after the traveler began his journey, had been thought to be impossible. It was Gödel's discovery that showed the general theory, which has passed every experimental test it has been subjected to (most recently, the September 2015 detection, from two massive colliding black holes, of gravitational waves—'ripples in spacetime'—generated more than a billion years ago in an effect predicted by the general theory a century ago), does allow time travel to the past *under certain conditions*. It is this availability of a *theory* that distinguishes time travel speculations from the outlandish fantasy speculations with which it is often unjustly lumped—speculations that *are* in the province of quacks (such as ESP, astrology, and mind over matter a' la spoon bending).

In his general theory, Einstein showed how spacetime can be either 'flat' (in the no-gravity, special relativity case of what is called a *Minkowski spacetime*[35]) or 'curved' (those situations with gravity), and he did that not by verbal hand waving, but rather by writing mathematical equations that obey all the known laws of physics: his famous gravitational field, nonlinear differential tensor equations. These complicated equations are notoriously difficult to solve in general, but in certain, special cases they *have* been solved. Those solutions describe how matter and energy and spacetime interact. As the popular saying puts it, "Curved spacetime tells matter how to move, and energy and matter tell spacetime how to curve." In that sense, gravity *is* curved spacetime.

In 1949 Gödel found one such special solution to the field equations that describes the movement of mass-energy not only through space but also *backward in time* along trajectories in spacetime that are called *closed time-like lines* or *curves*

---

[34] One pulp magazine science fiction story (F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931) got this right when its time traveler explains how his time machine works with these words: "Time as we know it is not universally absolute. The rate of its passage depends to a great extent upon the velocity of its observer with regard to some certain reference system. A moving clock will run slower with respect to a selected coordinate system than a stationary one." (Recall my earlier comments on the personal time of a time traveler.)

[35] Named after Hermann Minkowski (1864–1909), Einstein's mathematics professor in Zurich who gave the now well-known spacetime diagram interpretation of special relativity which, when originally presented by Einstein, was in the form of pure mathematics.

(called CTLs or CTCs, respectively).[36] These trajectories are such that if a human traveled along one, *always at a speed less than that of light* (that's what *time-like* means), he would see everything around him happening in normal causal order from moment to moment (for example, the second hand on his wrist watch would tick clockwise into the local future), but eventually the CTL/CTC closes back on itself and the traveler finds himself in his own past.

On the scale of the Solar System, general relativity has causality built into itself, but on much larger scales things can be a good deal more complicated. On a very large, astronomical scale, in fact, curved spacetime can result in violations of causality, with effects occurring before their causes. That is what the physics and the mathematics of Gödel's solution imply. That is what is meant by saying there is a scientific, rational basis for discussing time travel to the past. It is particularly important to note that travel along one of the closed time-like world lines discovered by Gödel requires a *machine*, some kind of accelerating rocket ship. That's because none of Gödel's CTLs/CTCs are what is called a *geodesic*. That is, none are *free-fall* world lines.[37] This machine does not, however, generate CTLs/CTCs where none existed before (CTL/CTC *creation* requires what physicists call a *strong* time machine) but rather simply makes use of the CTLs/CTCs that are inherent in Gödel's spacetime. A Gödelian rocket ship then is an example of a *weak* time machine.

I mentioned earlier that "certain, special cases" of Einstein's gravitational field equations result in CTLs/CTCs. What was the "special case" that Gödel solved? His solution of the field equations is for a rotating, infinite, static universe composed of a perfect fluid at constant pressure. In such a universe Gödel found that naturally occurring CTLs/CTCs pass through every point in spacetime; that is, time travel in Gödel's universe is *not* the result of a machine *manipulating* mass and energy on a *local* scale (the classic science fictional description of a time machine); rather, in Gödel's spacetime time travel is a naturally occurring phenomenon! The observable universe is, however, *non*-rotating and expanding (astronomers see red-shifts in the spectrums of distant stars) and so, although Gödel's spacetime satisfies the general relativity field equations, its time travel property does not hold in the spacetime in which we live. (This may account for why the initial reaction in the physics/philosophical communities, to Gödel's discovery that time travel is *not* nonsense according to general relativity, was mostly indifference.) The failure to observe time travel in our universe may (somewhat surprisingly, I think) still have possible implications for us, however, as one philosopher has cleverly argued.[38] He points out that naturally occurring Gödelian time travel would endow the universe with properties particularly useful for the survival of intelligence (presumably that includes humans) against extinction from a multitude of cosmic disasters. So, for those who argue that the universe we live in was *made* for us (the advocates of various proofs of God's existence that have Him as Designer), we have an obvious question: why did He (apparently) skip incorporating time travel?

---

[36] Kurt Gödel, "An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation," *Reviews of Modern Physics*, July 1949, pp. 447–450. A CTL/CTC is a special type of *world line*; the trajectory through spacetime of every particle in the universe is a world line that extends from each particle's past to its future. Our everyday experiences are with world lines that never cross or come close to themselves (which would put a particle at or near the same spacetime point more than once). That lack of experience with CTLs/CTCs that self-intersect is what makes time travel to the past so difficult for humans to grasp. For a discussion of *how* Gödel did what he did, see Wolfgang Rindler, "Gödel, Einstein, Mach, Gamow, and Lanczos: Gödel's Remarkable Excursion into Cosmology," *American Journal of Physics*, June 2009, pp. 498–510.

[37] It was discovered in 1969, however, that this isn't strictly true *if* one allows for a test particle (our 'time traveler') to be electrically charged. Then, naturally present electromagnetic forces acting on the particle could be sufficient to propel the particle along a Gödelian CTL/CTC. That is, no rocket would be required. See U. K. De, "Paths in Universes Having Closed Time-Like Lines," *Journal of Physics A*, July 1969, pp. 427–432. There are other solutions to Einstein's equations that do allow time travel on free-fall geodesics: see, for example, I. D. Soares, "Inhomogeneous Rotating Universes with Closed Timelike Geodesics of Matter," *Journal of Mathematical Physics*, March 1980, pp. 521–525.

In an invited essay that appeared the same year as his time travel physics paper, Gödel specifically addressed the seemingly paradoxical aspect of what he had discovered: "By making a round trip on a rocket ship in a sufficiently wide course, it is possible in these [rotating] worlds to travel into any region of the past, present, and future, and back again, exactly as it is possible in other worlds to travel to distant parts of space. This state of affairs *seems* [my emphasis] to imply an absurdity. For it enables one, e.g., to travel into the near past of those places where he has himself lived. There he would find a person who would be himself at some earlier period of life.[39] Now he could do something to this person which, by his memory, he knows has not happened to him."

Gödel's nerve then failed him, and he defended the possibility of the paradox of a time traveler meeting himself in the past with what I think an astonishingly unconvincing argument (particularly so for a logician) based primarily on *engineering* limitations: "This and similar contradictions, however, in order to prove the impossibility of the worlds under consideration, presupposes the actual feasibility of the journey into one's own past. But the velocities which would be necessary in order to complete the voyage in a reasonable time are far beyond everything that can be expected ever to become a practical possibility. Therefore it cannot be excluded a priori, on the ground of the argument given, that the space-time structure of the real world is of the type described."[40] That is, Gödel was trying to head off critics of his rotating universe model who might point to the time travel result as proof that the model had to be flawed.

---

[38] Alasdair M. Richmond, "Gödelian Time-Travel and Anthropic Cosmology," *Ratio*, June 2004, pp. 176–190. Not all physicists think Gödel's result is actually time travel. At least two think it is all simply the result of mathematical hijinks, and that time machines must remain "an aspect of science fiction fantasy": see F. I. Cooperstock and S. Tieu, "Closed Timelike Curves and Time Travel: Dispelling the Myth," *Foundations of Physics*, September 2005, pp. 1497–1509. This skepticism towards Gödel actually started much earlier, when two physicists (one a Nobel physics laureate) incorrectly claimed Gödel had simply gotten his math wrong: see S. Chandrasekhar and J. P. Wright, "The Geodesics in Gödel's Universe," *Proceedings of the National Academy of Sciences*, March 1961, pp. 341–347. It was those two physicists who had erred, however, as was pointed out by the philosopher Howard Stein, in his "On the Paradoxical Time Structures of Gödel," *Philosophy of Science*, December 1970, pp. 589–601.

[39] You'll recall that this is *precisely* the situation that Wells mentions in his "Notorious Literary Humbug" essay. If only he had lived just three more years, to see what he thought to be an absurdity actually appear in the serious writings of a brilliant mathematician!

In a footnote Gödel says that the time traveler would have to move at least as fast as nearly 71 % of the speed of light, and that if his rocket ship could "transform matter completely into energy" then the weight of the fuel would be greater than that of the rocket by a factor of 10^{22} divided by the square of the duration of the trip (in rocket years). A trip to the past in Gödel's universe would require a time machine that looked like Dr. Who's telephone booth attached to a fuel tank the size of several hundred *trillion* ocean liners. These are formidable numbers,[41] but they require no violation of physical laws, and that's what really counts if time travel is to be disproved. Gödel's use of engineering limitations for explaining away backwards time travel is actually worse than simply being wrong, because the puzzle is not in practicality but rather in showing, assuming that general relativity is correct, how correct mathematical physics can lead to what seems to be a paradoxical conclusion. (And see note 12 again, for another reason the 'fuel argument' really has no force at all against the possibility of time travel in Gödelian spacetime.)

So, what did the great man himself, Einstein, think of all this? In the same publication as Gödel's essay, he *cautiously* replied as follows: "Kurt Gödel's essay constitutes, in my opinion, an important contribution to the general theory of relativity, especially to the analysis of the concept of time. The problem here involved disturbed me already at the time of the building up of the general theory of relativity, without my having succeeded in clarifying it ... the distinction 'earlier-later' is abandoned for world-points which lie far apart in a cosmological sense, and those paradoxes, regarding the *direction* of the causal connection arise, of which Mr. Gödel has spoken ... It will be interesting to weigh whether these are not to be excluded on physical grounds."

Despite the mathematical physics of Gödel, showing the possibility of time travel to the past, many philosophers are not quite so sure. As one expressed his concerns, "No science-fiction staple poses more philosophical difficulties than time travel, but there is still no consensus as to whether time-travel fictions exhibit logical, metaphysical, or physical impossibility."[42] The best-known and possibly oldest of the paradoxical situations that seem to be part-and-parcel of time travel is the so-called *grandfather paradox*,[43] expressed this way by philosopher David Lewis in his pioneering 1976 paper (see note 5):

---

[40] Kurt Gödel, "A Remark About the Relationship Between Relativity Theory and Idealistic Philosophy," in *Albert Einstein: Philosopher-Scientist*: volume 7 of *The Library of Living Philosophers* (P. A. Schilpp, editor), Open Court 1949.

[41] For the analysis of a rocket powered by matter/anti-matter, a known physical process that satisfies Gödel's energy requirement for time travel, see E. Purcell, "Radioastronomy and Communication Through Space," in *Interstellar Communication* (A. G. W. Cameron, editor), W. A. Benjamin 1963.

[42] Alasdair Richmond, "Time-Travel Fictions and Philosophy," *American Philosophical Quarterly*, October 2001, pp. 305–318.

> Consider Tim. He detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine. Tim would like nothing so much as to kill Grandfather, but alas he is too late. Grandfather died in his bed in 1957, while Tim was a young boy. But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all. He buys a rifle, . . . and there [Tim] lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer . . .

So, there's the puzzle. Tim can obviously achieve his goal—he has a loaded gun, he's an excellent shot, a clueless granddad is coming ever closer—but if Tim actually does kill grandfather, years *before* Tim was (will be) born, then how can Tim *be* born? And if he is not born, then how can Tim ('now' not in existence) travel back through time to kill grandfather? What a confusing mess, right? So, the only possible conclusion to all this is that the starting premise, that time travel makes sense, must actually be nonsense. Right?

Well, maybe, but then what of Gödel with his time traveling rocket ship? That's hard-as-diamond, unshakeable mathematical physics, for heaven's sake. We can't just ignore *that*! Lewis offers a way out of this conundrum, and when we get to the book's discussions on paradoxes (that's plural because, believe it or not, there are other paradoxes even *more* perplexing than that of killing granddad in the distant past) we'll return to his solution.

Ever since Lewis wrote his paper, philosophers have been particularly fascinated by the grandfather paradox and have shown themselves to be at least as inventive as the science fiction writers in discussing it, or variations on it.[44] Here, for example, is a twist on that paradox that I think particularly clever, one that avoids the murderous spirit of the tale told by Lewis and Horwich:

---

[43] The origin of this paradox is probably lost in time (the irony of that is *so* appropriate!), but I have traced it at least as far back as to the science fiction pulp magazine *Science Wonder Stories* which published, in its December 1929 issue, an editorial essay titled "The Question of Time Traveling." It challenged readers to think about the following scenario: "Suppose I can travel back into time, let me say 200 years; and I visit the homestead of my great great great grandfather, and am able to take part in the life of his time. I am thus enabled to shoot him, while he is still a young man and as yet unmarried. From this it will be noted that I could have prevented my own birth . . .."

[44] Even before Lewis' paper, Paul Horwich had reduced the grandfather paradox to *autoinfanticide*—a time traveler tries to kill his younger *self*—in "On Some Alleged Paradoxes of Time Travel," *The Journal of Philosophy*, August 14, 1975, pp. 432–444. But not all philosophers share this fascination. Earman (see the opening quote), for example, dismisses *all* of the science fiction paradoxes that are so beloved by fans of the genre as "while always good for a chuckle," they are just "crude and unilluminating means of approaching some delicate and deep issues about the nature of physical possibility." I think Earman is fundamentally correct, although I wouldn't go so far as to characterize the paradoxes as mere "chuckles." They are, after all, the source of much of the intellectual motivation prompting the exploration of the physics of time travel. An excellent example of this is found in the paper by the Russian physicist S. V. Krasnikov, "Time Travel Paradox," *Physical Review D*, February 14, 2002, pp. 064013-1 to 064013-8. The physics of the grandfather paradox is of *great* interest in this paper.

> Sarah has just completed building her time machine. She decides to test the machine on herself tomorrow morning at which time she intends to travel back one day. In the meantime, she goes home, puts some salve on the burn she received that day, and goes to bed. In the morning, Sarah, with coffee in hand, sits down to read the morning paper. She opens the paper to the following headline: 'Famous physicist found dead.' On the front page is a picture of her body, salve burn clearly visible on her arm, inside her pristine time machine. Underneath is the caption. 'Nobel-prize winning physicist found dead yesterday in mysterious device that materialized near city hall.' Extremely shaken, Sarah returns to the lab and destroys the time machine.[45]

Can any sense be made of this? We'll come back to this question later in Chap. 5, when we discuss the possibility (or not) of time being *multi*-dimensional.

Now, to conclude this Introduction, let me end with two amusing, connected short stories (in epistle form) that nicely describe the issues we'll take up in the rest of this book. The rejection letter for the denial of a research grant to fund the construction of a time machine has just been received . . .

That Useless Time Machine[46]

Dear Review Committee:
It is not our practice to raise complaints against a negative review report. We believe in peer refereeing and we respect it, whatever its content and consequences. However, in the case of our latest grant application (project named 'The Time Machine') we find it necessary to express our astonishment at the motivations with which our request for funding was turned down. Your main objection appears to be that our project is 'philosophically interesting' but 'practically useless', by which you mean that the project 'has no potential for applications.' We do not quite think that the main criterion for judging the scientific value of a project should be its practical usefulness, but never mind that. Let us agree that usefulness is a relevant criterion, especially when large amounts of money are involved. Why should that be a reason to turn down our project? Quite frankly, we cannot think of a project with better application potential than ours. Some examples:

- Cultural tourism: one could send herds of history fans back in time to witness the crucial episodes of the French Revolution, or to watch the Egyptians build the pyramids, or to videotape Socrates' lectures.
- Exotic safaris: we have already received several applications for dinosaur hunting expeditions (they got extinct anyway).

---

[45] G. C. Goddu, "Time Travel and Changing the Past: (Or How to Kill Yourself and Live to Tell the Tale)," *Ratio*, March 2003, pp. 16–32.

[46] Story by Roberto Casati (Senior researcher at CNRS, Paris) and Achille C. Varzi (Professor of Philosophy at Columbia University). Originally published in *Philosophy*, October 2001, pp. 581–583, and reproduced here by kind permission of the authors.

- Error detection: we could take a closer look at our past mistakes and learn how to avoid them in the future.
- Historic documentaries: think of the huge saving in set design, costumes, special effects, etc. (How much did *Gladiator* cost?)

And so on and so forth. Honestly, can you think of a project with better prospects for useful and thrilling applications?

Sincerely Yours,
The 'Time Machine' Research Group

Dear 'Time Machine' Research Group:
Thank you for your letter. We agree that it would be interesting to exploit a time machine for the uses that you suggest. It would also be remarkable if we could use it to prevent all sorts of unpleasant events that happened in the past. It would be remarkable, for instance, to be able to go back to November 22, 1963, and prevent Lee Harvey Oswald from killing John Kennedy, or to go back to April 14, 1912, and steer the Titanic around the iceberg. It would be excellent indeed to be able to do such things. However, suppose your project were to be successful. Suppose you *will* manage to build a time machine. Then why *didn't* you do any of those things? Why is it that our past history is still full of such sad events? Either this means that your project is doomed to fail and you will never manage to build a time machine; *or* it means that the project will succeed but that you are not going to use your time machine for these good purposes. In the first case, logic shows it would be pointless to support your project. In the second case, ethics dictates that it would be wrongdoing. Either way, you must concede that the reasons against your project are overwhelming.

Cordially Yours,
The Review Committee

Dear Review Committee:
Certainly you have noticed that our suggestions for practical applications of the time machine did not include any uses that could result in an alteration of the natural course of history. As a matter of fact, we believe that no such alteration is logically possible. According to our project, it is logically possible to *visit* the past but not to *modify* the past. No time traveler can undo what has been done or do what has not been done. So the logic is safe. This does not mean that the time traveler will be ineffectual during her stay in the past, of course; it simply means that what she is going to do is something that she has already done. An accurate catalogue of all the past events would include an account of the arrival of the Time Machine from out of nothing as well as an account of all the actions and reactions that followed. And ethics is safe, too. For, if indeed we managed to go back to Dallas, we could not stop Oswald from doing what he did. Nobody would be able to stop Oswald because nobody was able to stop him (and nobody was able to stop Oswald because nobody will ever be able to do so, even if they came from the future). Alas, the past is full of sad events but there is nothing that we can do about that.

Respectfully Yours,
The ‘Time Machine’ Research Group

Dear ‘Time Machine’ Research Group:
We appreciate the distinction between changing the past (impossible) and affecting the past (possible). However, this simply reinforces our initial impression: your project has no practical value. If in order to travel to the past one has to have been there already, and if one can only do what has already been done, then *á quoi bon l’effort*? Why should we invest in a ‘Time Machine’ at all? We are afraid that our decision is now final.

Yours with best wishes,
The Committee

Well, all seems to be certainly lost with *that*. But, wait, perhaps not. Maybe, with just one more *really good* appeal, The Committee’s rejection can be reversed! If *you* were on the Review Committee, and had just read the following letter, how would *you* vote?

A Useful Time Machine[47]

Dear Review Committee:
We regret your continued decision to reject our proposal. Even though you have told us your decision is now final, we humbly ask your indulgence for one last appeal. We believe you have misinterpreted a crucial part of our proposal.

You maintain that our ‘Time Machine’ project ‘has no potential for applications’ and has ‘no practical value.’ You ultimately base this claim on the fact that “If in order to travel to the past one has to have been there already, and if one can only do what has already been done, then *á quoi bon l’effort*? Why should we invest in a ‘Time Machine’ at all?” Your argument however is a misinterpretation of our own comments that ‘According to our project it is logically possible to visit the past but not to modify the past ... This does not mean that the time traveler will be ineffectual during her stay in the past, of course; it simply means that what she is going to do is something that she has already done.’ We regret the awkward and easily misleading locution of the last sentence, but such are the perils of talking about time travel. Regardless, please consider our clarification.

Certainly if we were proposing that the time traveler *be* 5 years old again, we would be proposing something not worth the effort—our proposed time traveler has already turned five and cannot do so again. But we are not proposing that the time traveler do things that have already occurred in her own personal past, but rather in her personal future. The time traveler has not yet, from her personal temporal perspective, travelled back to, say, the library at Alexandria in 100 BCE. When she does travel back to 100 BCE to obtain scans of the books in the library before its destruction, she will be older than she is now. When she returns she will be still older (and we hope wiser, i.e., in possession of valuable information to which neither you nor we currently have access).

---

[47] Story by Geoff Goddu (Professor of Philosophy at the University of Richmond, Virginia). Originally published in *Philosophy*, April 2002, pp. 281–282, and reproduced here by kind permission of the author.

But is it true that as of 2002 AD [the year this letter was written] the time traveler has already visited Alexandria in 100 BCE? It could well be. But whether or not it is depends upon whether it is *also* true that our project will be successfully funded and completed. Because time travel into the past involves reverse causation, certain past events, such as the time traveler visiting 100 BCE, will be dependent upon certain future events, such as the successful funding and completion of our project. Hence, if it is not true that our project is both funded and completed, then it is not true that our time traveler has of 2002 already visited 100 BCE.

But suppose we were to learn now, before the funding and completion took place, that our time traveler had indeed been present at the library in Alexandria in 100 BCE. Would this imply that there was no reason to expend the effort to fund our project? After all, if the travel has 'already' happened, why bother funding the project? Firstly, such an argument does not imply that a 'Time Machine' would have no practical application, but rather expresses the futile hope that one could in fact get the practical benefits (if time travel is successful, we obtain the desired information) without expending the effort at all. Secondly, the hope is futile, for if we learn right now that our time traveler had been present at the library in 100 BCE, we would then know, assuming no other possible funding source, that you *will* expend the effort to fund our project. To deny this last is to make the impossible suggestion that even though your support is truly a causal antecedent of the successful trip, there is now no need for you to actually expend the effort to provide funding.

Hence, the effort is far from pointless, for the project will only succeed through your and our efforts. And success will generate, not only all the practical applications we outlined in our first letter, but, in addition, a host of information gathering applications such as more accurate historical research, lost item location identification, legal testimony verification, etc. Even if, as we (and you) acknowledged, no one could now prevent Oswald from killing Kennedy, wouldn't it be worth verifying that Oswald was the lone killer of Kennedy? Also, the information gathering need not be restricted to the past. For example, information concerning the prices of various stocks 10 years from now would be extremely valuable to a suitably cautious and prudent investor. Surely you cannot object to our information gathering in the future on the grounds that 'it will already have been done.' And just think, the information we obtain could be what allows you to obtain at very low prices those stocks that in the future will be extremely valuable and allows your esteemed committee to dramatically increase your support of worthy scientific endeavors.

Again, we ask you to reconsider your original decision.
Respectfully yours,
The 'Time Machine' Research Group

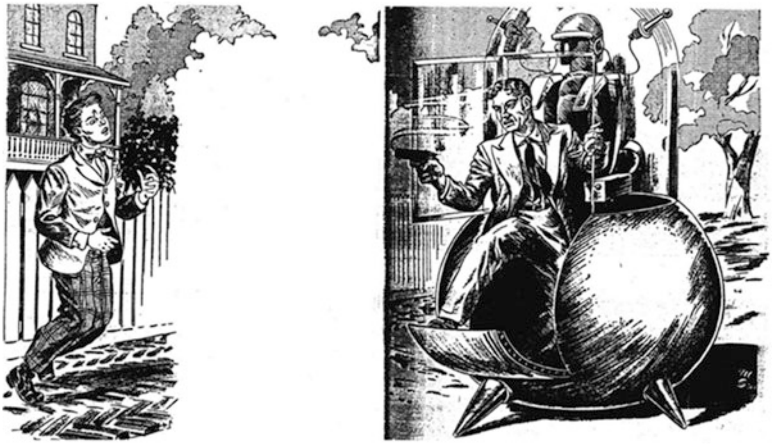

A time machine inventor makes an experimental test of the grandfather paradox! (Illustration from “Thompson’s Time Traveling Theory” by Mortimer Weisinger), *Amazing Stories* March 1944 (art by Malcolm Smith). Reprinted by arrangement with Forrest J. Ackerman, Holding Agent, 2495 Glendower Ave., Hollywood, CA 90027.

Not everybody likes time machines as a science fiction gadget, not even otherwise enthusiastic devotees of the genre. For example, in a Letter-to-the-Editor published in the December 1931 issue of Astounding Stories*, one seventeen-year-old fan had this to say: “There is only one kind of Science Fiction story I dislike, and that is the so-called time-traveling. It doesn’t seem logical to me. For example, supposing a man had a grudge against his grandfather, who is now dead. He could hop in his machine and go back to the year that his grandfather was a young man and murder him. And if he did this how could the revenger be born? I think the whole thing is the ‘bunk.’” As this book will demonstrate, this young reader was not alone in that opinion. As this book will also demonstrate, in the last few decades that view has been rapidly evolving.*

For Further Discussion

Read again the penultimate sentence in the last letter from The ‘Time Machine’ Research Group, and then think about how you would respond to the following questions.

(1) Would *you* invest in a stock market if you knew somebody else had a time machine giving *them* advance information on stock performance?
(2) How might the existence of a time machine influence the future of the stock market, in general? For an early science fiction look at these questions, see Lee Laurence, “History in Reverse,” *Amazing Stories*, October 1939.

One writer has speculated that Wells’ model for the Time Traveller was the American inventor Thomas Edison. (See Martin T. Willis, “Edison as Time Traveler: H. G. Wells’ Inspiration for His First Scientific Character,” *Science Fiction Studies*, July 1999, pp. 284–294.) As Wells worked his way from *The Chronic Argonauts*, through revisions, to the final *Time Machine*, the story’s hero evolved from Dr. Nebogipfel to the Philosophical Inventor to the Time Traveller. The one individual who could have inspired all of these various hero types was, according to Willis, Edison, a world-famous Victorian-age celebrity whose story was well known to Wells. If Wells had today’s scientific personalities available as potential inspirations, who do you think he would use? How might that choice affect the story and structure of *The New Time Machine*?

The idea of *personal time*, used by the philosopher David Lewis (note 5) to consistently interpret time travel stories, has been used in a quite different way (although time travel gets a few words, too) by the philosopher Roy Sorenson. In his paper “The Cheated God: Death and Personal Time,” *Analysis*, April 2005, pp. 119–125, Sorenson asks you to imagine an immortal god. For some reason this god runs afoul of a demon, who curses the god in a curious way. (The ‘telling of a story’ is a common technique in philosophical papers and, while foreign to what readers of physics papers are used to seeing, is not without some charm. Just be sure to always keep in mind that its

(continued)

primary use is as an *attention-grabbing* device, but as far as having any other merit, well, that's often another story.) The curse is such that the life span of the once immortal god is reduced to that of a normal human life span and yet, perhaps surprisingly, the god will still never die. As Sorenson writes, "[The god] will live forever. But [the god] will not have a better life than a mortal. The demon has harmed [the god] as gravely as death harms mortals." How, you might wonder, is this to be done? As Sorenson explains, "[The god] lives half of its now mortal span, followed by a trillion years of nothingness, then a quarter of its mortal span followed by a trillion years of nothingness, then an eighth of its mortal span followed by a trillion years of nothingness and so on ad infinitum." Sorenson's argument is simply an exotic form of the high school summation of the geometric series $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots=1$, where there are an *infinite* number of terms to the left of the equality. (Each term represents a period of time during which the god is conscious, and each + represents a trillion years.) Sorenson picked a trillion years of nothingness between consecutive periods of consciousness for (I suggest) dramatic reasons, but suppose instead that he had picked 1 μs for the period of nothingness. Discuss what effect this would have (if any) on the life of the god. Consider two cases:

(a) There is no minimum time duration for consciousness, and
(b) There *is* a minimum time duration such that, for any shorter duration, a consciousness remains 'unaware' even though it is *not* in a state of nothingness.

After working all night making some final calculations, a physicist carefully solders a final resistor into the control module of the world's first time machine and then steps into the gadget that is a sure bet to win the next Nobel Prize in physics. As she does, she notices that it is precisely 8:10 in the morning, as indicated on both her wrist watch and the clock on the lab wall. After settling into a plush leather seat she pushes the time machine's power button, the machine glows with a flickering blue-red halo and hums with a mighty throb for a while and then, at precisely 8:15 by her wrist watch, she steps out of the machine and back into her lab. She notices the clock on the wall now reads 8:05. That is, she took 5 min of personal time (8:10 to 8:15) to travel 5 min of external time into the past (8:10 to 8:05). On the one hand she certainly seems to be a time traveler, in that she exits the machine before she enters it. (Ignore the issue of there being *two identical* physicists in the lab from 8:05 to 8:10!) On the other hand, the elapsed personal and external

(continued)

times are *equal*. Does this suggest a need to modify or expand David Lewis' definition of a time traveler? As you ponder this question, you might want to read the following four papers: (1) Paul R. Daniels, "Lewisian Time Travel in a Relativistic Setting," *Metaphysica*, October 2014, pp. 329–345, (2) Douglas Kutach, "Time Travel and Time Machines," in *A Companion to the Philosophy of Time* (H. Dyke and A. Bardon, editors), Wiley-Blackwell 2013, pp. 301–314, and (3) Frank Arntzenius, "Time Travel: Double Your Fun," *Philosophy Compass*, November 2006, pp. 599–616. A bit more demanding (but worth the effort) is the long chapter "Time Travel and Time Machines" by Chris Smeenk and Christian Wüthrich, in *The Oxford Handbook of Philosophy of Time* (C. Callender, editor), Oxford 2011, pp. 577–630 (see page 580, in particular).

> The idea that information is physical has given rise to a series of discoveries which indicate that physics has much to say about fundamentals of computer science.

The above quotation is the opening sentence to a most interesting paper by the physicist Dave Bacon, "Quantum Computational Complexity in the Presence of Closed Timelike Curves," *Physical Review A* (70), 2004. (When he wrote, Bacon was at Caltech, but he is now a software engineer at Google.) The title of Bacon's paper, translated into blunt English, is "It Would Be Really Neat If We Could Merge a Time Machine With a Computer." That is, to further quote from Bacon's paper, "One could [efficiently] solve a hard problem by trying out a solution to the problem, sending one's computer back in time, attempting a different solution to the problem, sending one's computer back in time, etc., until a solution to the problem has been found." There then follows a pretty sophisticated analysis on the self-consistent time evolution of a quantum system, ending with Bacon's frank admission that "we would not be honest if we did not end this paper with the caveat that this work is at best a creature of eager speculation . . . Practical considerations are humorous at best." Read Bacon's paper and discuss what he means by "a hard problem." (There *is* a technical term used by computer scientist for such problems: NP-complete.)

The occasional theological commentary in this book may strike some as a bit odd for a topic treated with heavy doses of deep mathematics in the physics literature but, as you'll see on the following pages, theology is an unescapable dimension to any informed discussion of time travel. A literary connection

(continued)

between time travel and theology has, in fact, existed for a long time. As pointed out in Paul Alkon's *Origins of Futuristic Fiction* (University of Georgia Press 2010), "The first time-traveler in English literature is a guardian angel who returns with state documents from 1998 to the year 1728 in Samuel Madden's *Memoirs of the Twentieth Century*" (published in 1733, nearly three centuries ago). Madden was an Irish-Anglican clergyman whose book was satire rather than science fiction, but its time traveling aspect was a first. As Professor Alkon also writes, "Madden [was] the first to write a narrative that purports to be a document from the future. He deserves recognition as the first to toy with the rich idea of time-travel in the form of an artifact sent backward from the future to be discovered in the present." Your assignment: read and discuss Alkon's book.

You'll recall that Gödel cast his view of time travel in the form of a self-encounter in the past. In Frederik Pohl's "Let the Ants Try," we find a science fiction tale that appeared essentially simultaneously with Gödel's paper (*Planet Stories*, Winter 1949), in which a time traveler journeys back forty million years. Upon stepping out of his time machine, he hears a "raucous animal cry" from somewhere in the nearby jungle. Later, after other adventures in time, he returns to near the same point in spacetime. After stepping out of his time machine, he sees himself in the distance—the earlier version of himself during the first trip. Then, suddenly, the time traveler meets a violent death: "As his panicky lungs filled with air for the last time, he knew what animal had screamed in the depth of the Coal Measure forest." In fact, self-encounters had appeared in science fiction *years* before Gödel's paper. In the 1942 story "Minus Sign" (*Astounding Science Fiction*, November) by Jack Williamson, for example, a spaceship battles with itself while traveling backward in time. How do you think a scientist like Gödel would have liked these two stories? (Who knows, maybe he *did* read them!) If you could travel back in time to 1949 to ask him if such tales had been an inspiration, do you think he would be intrigued, amused, or instead would he be insulted?

Contents

1 A Broad Look at Time Travel 1
- 1.1 Time Travel in the Fantasy and Science Fiction Literature 1
- 1.2 Where Are All the Time Travelers? 10
- 1.3 Skepticism About Tales of Time Travel 15
- 1.4 Troubles with (some) Time Machines 22
- 1.5 Quantum Gravity, Singularities, Black Holes, and Time Travel 29
- 1.6 Tipler's Time Machine 38
- 1.7 For Further Discussion 42

2 Philosophical Space and Time 51
- 2.1 Time: What Is It, and Is It Real? 51
- 2.2 Linear Time and the Infinity of Past and Future 61
- 2.3 Cause and Effect 67
- 2.4 Backward Causation 72
- 2.5 The Fourth Dimension 78
- 2.6 Spacetime and the Block Universe 90
- 2.7 Philosophical Implications of the Block Universe 100
- 2.8 For Further Discussion 109

3 The Physics of Time Travel: Part I 115
- 3.1 The Direction of Time 115
- 3.2 The Arrows of Time 124
- 3.3 Time Dilation 139
- 3.4 The Lorentz Transformation 146
- 3.5 Spacetime Diagrams, Light Cones, Metrics, and Invariant Intervals 155
- 3.6 Proper Time and the Twin Paradox in Time Travel to the Future 173
- 3.7 For Further Discussion 181

4 Philosophers, Physicists, and the Time Travel Paradoxes 187
4.1 Paradoxes and Their First Appearance in Science Fiction 187
4.2 Changing the Past and the Grandfather Paradox 195
4.3 Changing Versus Affecting the Past 206
4.4 Causal Loop and Bootstrap Paradoxes 214
4.5 Sexual Paradoxes 225
4.6 Splitting Universes and Time Travel 229
4.7 For Further Discussion 237

5 Communication with the Past 245
5.1 Reversed Time 245
5.2 Multi-dimensional Time 252
5.3 Maxwell's Equations and Sending Messages to the Past 256
5.4 Wheeler and Feynman and Their Bilking Paradox 264
5.5 Absorber Theory and Signaling the Past 269
5.6 Tachyonic Signals and the Bell Quantum Antitelephone 273
5.7 For Further Discussion 283

6 The Physics of Time Travel: II 289
6.1 Faster-than-Light into the Past 289
6.2 Tipler's Rotating Cylinder Time Machine 297
6.3 Thorne's Wormhole Time Machine 301
6.4 Gott's Cosmic String Time Machine 319
6.5 Cutting and Warping Spacetime 326
6.6 For Further Discussion 334

Appendix A: Old Friends Across Time (A Story) 339

Appendix B: Newton's Gift (A Story) 347

Appendix C: Computer Simulation of the Entropic Gas Clock 355

Epilogue 357

Glossary 367

Index 377

About the Author

Paul J. Nahin was born in California, and did all of his schooling there, including a PhD in electrical engineering from UC Irvine (1972). He is now emeritus professor of electrical engineering at the University of New Hampshire. Prof. Nahin has published a couple of dozen short science fiction stories in *Analog*, *Omni*, and *Twilight Zone* magazines, and has written 18 books on mathematics and physics. He has given invited talks on mathematics at many prestigious institutions, and has appeared on National Public Radio's "Science Friday" show (discussing time travel) as well as on New Hampshire Public Radio's "The Front Porch" show (discussing imaginary numbers). He advised Boston's WGBH Public Television's "Nova" program on the script for their time travel episode.

Chapter 1
A Broad Look at Time Travel

> *"Hold infinity in the palm of your hand and eternity in an hour."*
>
> —William Blake, writing in "Auguries of Innocence" (1863), with words that could quite well describe what it would be like to time travel
>
> *"I need a place to hide, that's why I believe in yesterday."*
>
> —The Beatles (*Yesterday*, 1965)

1.1 Time Travel in the Fantasy and Science Fiction Literature

> "Woodn't it be grate to go back in tyme and correct your mistakes? Wouldn't it be great to go back in time and correct your mistakes?"
>
> —motto of *Time Twisters* comics

To travel in time.

Could there possibly be a more exciting, more romantic, more wonderful adventure than that? I don't think so, and in this opening section I want to just briefly discuss how fascinating many writers (and their readers) have found the concept of time travel, and to point out that the fascination began *long before* mathematical physicists discovered time travel lurking in Einstein's general theory of relativity.[1]

Before the arrival of humans on the surface of the Moon in 1969, the only other 'fantastic voyage' that could compare with time travel was traveling through outer space. During the seventeenth and eighteenth centuries, in fact, such voyages were the center of a genre of fiction (now called *science fiction*) called the "imaginary voyage" or "extraordinary voyage." Marjorie Hope Nicolson's[2] 1948 book *Voyages*

---

[1] An excellent, book-length *literary* treatment of time travel is by David Wittenberg, *Time Travel: the popular philosophy of narrative*, Fordham University Press 2013.

[2] Marjorie Hope Nicolson (1894–1981) was a literary scholar of the first rank at both Smith College and Columbia University.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_1

to the Moon carefully documents just how popular that form of literature was—and still is. Since 1969 the first such voyages have become history, of course, and time travel has replaced space travel as the modern "imaginary voyage."

It seems a safe bet that that, given a random selection of middle-aged adults, the vast majority of them would respond enthusiastically if asked whether time travel interests them. This fascination with time travel has actually been 'scientifically' documented. In one intriguing study,[3] several hundred men and women were asked to consider the possibility of spending an hour, a day, and a year back in both their personal past (since birth) and their historical past (before birth). They were further told that it would cost $10,000 to purchase such time travel services. Their response indicated that 10 % would be willing to spend that much money for an hour in the historical past, 22 % for a day, and 36 % for a year. As might be expected, the numbers rose as the cost dropped and, if such trips were free the interest was almost universal. As one writer put it, "Time travel [is] the ultimate fantasy, the scientific addition to the human quest for immortality."[4] And as a philosopher observed, "[T] he popular appeal of time travel . . . is no doubt due to a nostalgia for the past, which is almost an omnipresent aspect of the human condition."[5]

Fiction writers have, for *centuries*, recognized the fantasy appeal of time travel. The common fairy tale theme of 'The Three Wishes,' in which the recipient ends up using the final wish to undo the unforeseen consequences of the first two, is the precursor to all modern change-the-past time travel stories. Indeed, the means of time travel in the Norwegian poet Johan Wessel's 1781 play *Anno 7603 is* a fairy. Some of the best modern science fiction stories have played with the fantasy appeal of time travel by having gifts arrive by accident from the future: the moral of such tales is generally that unearned gifts usually bring grief.[6] The editor's introduction to a time travel story involving the Civil War referred to the fantasy aspect of time travel—with a reference to another age-old adult fantasy—when he wrote "time travel stories about the Civil War have one thing in common with pornography; they serve to titillate an impulse [in the case of time travel stories, the impulse to change history] and to frustrate [history]."[7] This is the motivation for the time traveler in Stephen King's 2011 novel *11/22/63*, who uses what appears to be a naturally occurring wormhole (connecting a Maine diner to 1958) for his attempts at preventing the assassination of John F. Kennedy.

---

[3] T. J. Cottle, "Fantasies of Temporal Recovery and Knowledge of the Future," in *Perceiving Time*, John Wiley 1976.

[4] T. Paul, "The Worm Ouroboros: Time Travel, Imagination, and Entropy," *Extrapolation*, Fall 1983, pp. 272–279.

[5] J. W. Smith, "Time Travel and Backward Causation," *Cogito* 1985, pp. 57–67.

[6] Among the many such tales, five particularly good ones are "Something for Nothing" by Robert Sheckley, "Mimsy Were the Borogoves" by Lewis Padgett (pseudonym of the married couple Henry Kuttner and C. L. Moore), "Child's Play" by William Tenn (pseudonym of Philip Klass), "The Little Black Bag" by Cyril Kornbluth, and "Thing of Beauty" by Damon Knight. All can be found in various anthologies.

[7] *The Fantastic Civil War* (F. McSherry, Jr., editor), Baen 1991.

A character in the 1985 novel *The Bird of Time* by George Effinger nicely captures the fantasy appeal of time travel with the declaration "The past . . . is the home of romance." On a less poetic level, time travel and the movies and stories about it fascinate most people because they turn our everyday world view upside down and inside out.[8] Such movies and stories make people *think*. It is therefore not surprising that time travel movies have been popular for decades, from the pioneering *Berkeley Square* in 1933, to the classic 1960 filming of *The Time Machine*, to *Back to the Future* in 1985 (*the* top film that year in a *Boxoffice* magazine poll), to the flawless 1989 *Bill & Ted's Excellent Adventure*, to the clever *Terminator* action films, to the ingenious 2012 *Looper*, to the commercially successful 2014 *Interstellar*. Each of these films, and others, too, will be discussed later in the book.

When we discuss time travel, we should really be careful to distinguish between two quite different versions: to the future, and to the past. There is no dispute, today, about the first. As two *severe critics* of the possibility of time travel to the past wrote decades ago, "After 1900, special relativity made scientific discussion of time machines possible."[9] What they were referring to is the fact that, by traveling in a rocket ship fast enough (but never, unlike Superman, faster than the speed of light), and far enough, one could leave Earth, loop out on a vast journey perhaps halfway across the universe, and then return hundreds, thousands, even millions of years in the future. You could theoretically (ignoring all the *engineering* difficulties) do this, in fact, with the apparent passage of your 'personal time,' (as measured by your wrist watch or the beating of your heart) as brief as you'd like. (Physicists call 'personal time' *proper time*, and I'll return to this in Chap. 3.) This astonishing conclusion from special relativity, that time travel to the future makes physical sense, literally put a lot of Victorian-era trained physicists into shock.

A quite sophisticated use of this idea appeared early in science fiction, in the tale of a space traveler who returns from a high-speed trip out to the blue supergiant star Rigel in the constellation Orion.[10] The 900 or so light years of the round trip had taken just 6 months of ship or personal time (proper time), but a thousand years of back-home time. The traveler returns to Earth to find all he had left behind long dead and returned to dust: "Sometimes I waken from a dream in which they are all so near . . . all my old companions . . . and for a moment I cannot realize how far away they are. Beyond years and years."

Another story[11] of a trip into the future that delivers an equally powerful emotional impact, this time via a Wellsian-type time machine (more on what that means later in this chapter) rather than by rocket travel, tells of a time traveler trapped in post-nuclear war times, where there is no energy available to power his machine for the return trip. As the story ends, he finds the woman he had loved and left behind in the past. She is now the elderly widow of another man, having married his rival because the time traveler (just like Wells' Time Traveller) never returned.

---

[8] Somewhat more pompous (but no less correct) was this observation by an academic: "The time-travel [film] romance is an attempt to reenchant the world, to regain a sense of belongingness, to reinstate the magical, autocentric Universe of the child and the primitive." See W. Wachhorst, "Time Travel Romance on Film," *Extrapolation*, Winter 1984, pp. 340–359.

[9] S. Deser and R. Jackiw, "Time Travel?" *Comments on Nuclear and Particle Physics*, September 1992, pp. 337–354.

[10] R. H. Wilson, "Out Around Rigel," *Astounding Stories*, December 1931.

[11] W. Tucker, *The Year of the Quiet Sun*, Gregg 1979.

A trip into the future does not *have* to be serious or sad. A nice example of that is the story of a spaceship crew that sets off for the Alpha Centauri triple star system, more than four light years distant.[12] They survive the trip, which requires 500 years of both personal *and* external time (this story is the *only one* I'll mention in this book that uses *a preserving drug* rather than physics). What causes me to include it in a book on *time travel* (stories in which proper and external time are one-in-the-same are simply *not* about *time travel*) is that, long before they arrive at the end of their journey, the secret of faster-than-light (FTL) travel is discovered back on Earth and so they arrive at their destination to find a human reception committee! As you'll see when we get to Chap. 3, knowledge of FTL travel is equivalent to knowing the secret of travel into the past, and so the crew is sent back in time, to Earth, to just one year after they left—and they listen to their own radio communications arriving from deep space.

The *real* adventure in time travel, as suggested by "Far Centaurus," would be to go backward in time, to visit the past. The editor of the science fiction pulp magazine *Thrilling Wonder Stories* used the powerful emotional hook of changing the past in a 1950 blurb announcing a time travel story coming in the next issue: "What's the biggest mistake you ever made? Don't worry about it. You may have pulled some awful boners in your time, but there's a sure-fire remedy for them all. It's simple. Just look up at that old time-clock on the wall—and turn it back to the moment just preceding your terrible blunder. Then make your corrections—and set your time-clock back to the present. You may be starting a new chain of error, but why fret? You can go back in time again..." Or, as the promotional text on the video package of the 1986 movie *Peggy Sue Got Married* says, "to do it again" is "the golden opportunity almost everyone has longed for at least once."

Writing less romantically, a philosopher declared that a "major source of interest in the time travel question is our general fascination with the exotic and the child-like frustration we sometimes feel at being confined to the present. We wish that the benefits of moving through space could be supplemented with the benefits which would accrue from movements through time."[13] Robert Silverberg, a science fiction writer who has used the time travel theme often and effectively, expressed this sentiment quite clearly when he wrote "Suppose you had a machine that would enable you to fix everything that's wrong in the world ... The machine can do anything ... it gives you a way of slipping backward and forward in time ... Call this machine whatever you want. Call it Everybody's Fantasy Actualizer. Call it a

---

[12] A. E. Van Vogt, "Far Centaurus," *Astounding Science Fiction*, January 1944.

[13] R. A. Sorenson, "Time Travel, Parahistory and Hume," *Philosophy*, April 1987, pp. 227–236.

Time Machine Mark Nine."[14] He gives a masterful demonstration of what he means by that in one of his own stories, a tale[15] set in a year when time machines actually exist. Even so, the characters use their imaginations to explore their fantasy worlds and wishes—wishes that could (if they *really* wanted to) be realized with a real time machine. Time travel fiction is, you see, the ultimate escapist literature!

The adventure promised by time travel to the past doesn't necessarily mean *pleasant* adventure, and science fiction has used that idea to great effect. The unstated horror of a trip backward in time, if you think just a bit about it, is that it would bring the dead past, filled with all its dead occupants, alive again, literally resurrected from dank and moldering graves. The top of Mount Everest, the bottom of the Marianas Trench, the sands of Mars—none of these exotic places can even be mentioned in the same breath with the *past*. The capture of the mystery and, yes, the sheer terror of the past, is in this opening line to a 1950s tale: "When Dr. Flitter came into the room, it seemed as though the past and its dead people came in with him, clinging to him like stale surgery smells, like the cold sweat of ancient autopsies."[16] In another equally macabre story, we read of a time traveler in the past anticipating a meeting with a long-dead lover as he "shivered with a renewal of horror . . . She ought to be grateful to him for having raised her from the dead, even briefly."[17]

Another tale, slightly less gruesome, tells us of a character who delights in pointing out all the bad aspects of living in the past.[18] Tell him *when* in time, and he quickly ticks off the disadvantages of being *then*. To live in ancient Greece would let you rub shoulders with Aristotle, sure, but you already know what he said and you'd soon regret the lack of modern plumbing. The year of the American Revolution might let you exchange greetings with George Washington, but you'd also have to put up with cholera in Philadelphia, malaria in New York, and the fact that if you needed an operation there would be no anesthesia anywhere. The Victorian Age appeals to modern romantics but, before you go back, you'd better have your eyes and teeth checked. The time traveling historian in one novel[19] takes these medical warnings to heart and has her appendix prophylactically removed, and is further advised to have her nose cauterized against all the awful stinks of her destination, the fourteenth century.

On the other hand, poor health care isn't *all* there is with time travel to the past. One time traveler from the future, for example, makes a very good living in the past by winning bets on yet-to-happen events whose outcomes he knows.[20] Time travel to the past allows a Parisian curio shop in the present to offer remarkably authentic looking newspapers from 1804 whose only ‘flaw’ is that they appear to be fresh off the press—which of course they are![21] And the failed professor in one romantic story[22] finds the Paris of 1482 infinitely better than the Paris of 1961.

---

[14] R. Silverberg, "Ms. Found in an Abandoned Time Machine," *Beyond the Safe Zone*, Warner 1986.

[15] R. Silverberg, "Many Mansions," *Beyond the Safe Zone*, Warner 1986.

[16] R. Bretnor, "The Past and Its Dead People," *Magazine of Fantasy and Science Fiction*, September 1956.

[17] L. Marlow, *The Devil in Crystal*, Faber and Faber 1944.

[18] A. Bester, "Hobson's Choice," *Magazine of Fantasy and Science Fiction*, August 1952.

[19] C. Willis, *Doomsday Book*, Bantam 1992.

A popular fictional appearance of time travel is the use of the past as a hiding place, as a sanctuary for those wishing to escape the troubles of modern times.[23] An interesting twist on this idea was presented in one tale[24] in which the past is used for later military gain in the present. In this story the Earth of a thousand years in the future is ruled by a dictator, and the oppressed masses are unable to arm themselves for revolt. So, back into the past travels an agent to arrange for the construction of weapons, which are then stockpiled in hidden caverns where they can be retrieved for use ten centuries later. The past is used in this story as both a sanctuary *and* a repository from which to make war in the future, and so we have a time travel fantasy for both doves *and* hawks in the same tale! This military use of the past is passive; other writers have more aggressively used time travel to the past for military gain as, for example, mining uranium deposits before they have had time to reduce themselves to lead via radioactive decay,[25] or in drilling for Middle East oil in the past to deprive adversaries of it in the present.[26]

Time travel to the past would, perhaps, interest criminals, too. As the science fiction writer Larry Niven wrote, “If one could travel in time, what wish could not be answered? All the treasures of the past would fall to one man with a submachinegun. Cleopatra and Helen of Troy might share his bed, if bribed with a trunkful of modern cosmetics.”[27] Or, as the tragically flawed inventor of the first time machine dreamed, before using time travel to commit what he thought would be the perfect locked-room murder, “The Great Harrison Partridge would have untold wealth. He could pension off his sister Agatha and never have to see her again. He would have untold prestige and glamour, despite his fat and baldness, and the beautiful and aloof Faith Preston would fall into his arms like a ripe plum."[28]

---

[20] C. Sprague, “Time Track,” *Startling Stories*, January 1951.

[21] M. Leinster, *Time Tunnel*, Pyramid Books 1964.

[22] U. K. Le Guin, “April in Paris,” *Fantastic Stories*, September 1962.

[23] There are *many* excellent examples of such tales, a few of which are Clifford Simak, “Over the River & Through the Woods”; Ray Bradbury, “The Fox and the Forest”; Jack Finney, “Such Interesting Neighbors”; J. B. Priestly, “Mr. Strenberry’s Tale”; James Gunn, “The Reason Is With Us”; and H. B. Piper, “Flight from Tomorrow.” All can be found in various anthologies.

[24] R. F. Young, “Not to be Opened—,” *Astounding Science Fiction*, January 1950.

[25] C. Simak, “Project Mastodon,” *Galaxy Science Fiction*, March 1955.

[26] P. Anderson, “Wildcat,” *Magazine of Fantasy and Science Fiction*, November 1958, and W. Jeschke, *The Last Day of Creation*, St. Martin’s Press 1982.

[27] L. Niven, “The Theory and Practice of Time Travel,” in *All the Myriad Ways*, Ballantine 1971.

Instead of viewing the past as an aid to crime, some writers have used it as the perfect dumping ground for criminals, as a highly convenient place to remove them from society.[29] After all, there can be no breakout from the prison of the past—at least not without a time machine. What might happen to criminal recidivists in a world that has mastered time travel is nicely explained in one story as follows: "If you cannot live among people, then off to the reptiles—one hundred or one hundred twenty million years before the present. There you wouldn't freeze in a tropical pre-glacial climate, and you could nourish yourself on plants. But there is no one to talk with, boredom, and in the end you offer yourself up as an afternoon snack to a tyrannosaurus."[30] With an interesting twist on this is the tale[31] of a physics professor who *helps* criminals disappear into the past to escape relentless police pursuit.

Museum curators, too, would seem to be obvious clients for time machine companies, as would collectors of extinct species who work for zoos. An example of the first case is in a novel[32] about a time travel business called *Time Researchers*, with the corporate mottos 'We Sift the Sands of Time.' It works as a futuristic version of Indiana Jones, as finders of lost historical artifacts for customers who can pay the substantial charges. A typical mission is to make original sound recordings of one of Lincoln's unreported speeches. We find the same idea in a short story[33] about a business called *Genealogy, Inc.*, with the corporate motto

"An Ancestor for Everybody." It uses a 'time scanner' to provide its clients with a list of distinguished predecessors. And in another novel we read of the Historical Corps, whose time travel agents are "writing the definitive history of mankind.".[34]

Another use for time travel to the past, one of the most unusual I have seen, was suggested in a philosophical article[35] which considers an age-old question that has long bedeviled schoolboys: "If everything in the Universe doubled in size overnight while we slept, could we tell what had happened when we woke up next morning?" The usual answer to this puzzle (called the *Universal nocturnal expansion* by philosophers) is *no*, but the authors of the article suggest that 'all' we need do is take a yardstick back to yesterday and compare it with itself! Great idea, for sure, but it had appeared years earlier in a science fiction story.[36]

---

[28] A. Boucher, "Elsewhen," *Astounding Science Fiction*, January 1943. Partridge's dream is shattered, however, because he overlooks a few details about time travel, ones that he wouldn't have missed if he could have read this book. We'll come back to this classic story, which merges a time machine with murder, later in the book.

[29] This is a popular science fiction scenario, and three of the best stories playing with it are P. Anderson, "My Object All Sublime," *Galaxy Science Fiction*, June 1961; I. Watson, "In the Upper Cretaceous with the Summerfire Brigade," in *Stalin's Teardrops*, Victor Gollancz 1991; and R. Silverberg, "Hawksbill Station," *Galaxy Science Fiction*, August 1967.

[30] S. Gansovsky, "Vincent Van Gogh," in *Aliens, Travelers, and Other Strangers*, Macmillan 1984.

[31] J. Finney, "The Face in the Photo," in *About Time*, Simon and Schuster 1986.

[32] W. Tucker, *The Lincoln Hunters*, Rinehart 1958. See also A. Bitov, "Pushkin's Photograph," in *The New Soviet Fiction*, Abbeville Press 1989.

[33] M. Shaara, "Man of Distinction," *Galaxy Science Fiction*, October 1956.

[34] L. A. Frankowski, *The Cross-Time Engineer*, Del Rey 1986. The same idea is in the 1991 film comedy *The Spirit of '76*, in which time-traveling historians from 2176 visit the past in an attempt to reconstruct the lost records of the founding of America.

[35] J. C. Graves and J. E. Roper, "Measuring Measuring Rods," *Philosophy of Science*, January 1965, pp. 39–56.

More ingenious uses for the past are discussed in the story[37] of a time travel business called *Time Associates*. One use comes in the form of a request from a United States senator who wants to send the disadvantaged of today back into the remote past, where they could have a fresh start on a virgin Earth. Yet another use comes from a religious fringe group that wants to purchase exclusive rights to the time of Jesus—not to visit, but to prevent *anyone* from visiting. The group fears that any such visitors would "learn the truth," which might contradict the very legends that form the heritage of Christianity. And, in what may be the most ingenious idea of all, *Time Associates* itself does not do business in the present, but rather 150,000 years in the past, in a 'new' country called Mastodonia. The corporate lawyer, you see, has determined that such an arrangement legally means the company is a *foreign* company doing business outside the United States, and so it is not liable for taxes to the IRS!

The tourist trade is a booming business in science fiction, with dinosaur hunting at the top of the list. There are many such tales,[38] including the cerebral stories in L. Sprague de Camp's short-story collection *Rivers of Time*, starting with the classic "A Gun for Dinosaur." The earliest (that I know of) fictional use of time travel to the past for hunting was not for dinosaur hunting, however, but rather for saber-toothed tigers, wooly mammoths, and cave bear.[39] Historical tours to the great events of the past are also an entertaining use of time travel.[40] Even mundane events may one day be on the 'to do' lists of time travelers to the past. In one, for example, curious crowds from the very far future show up, nightly inside the home a twentieth century family, much as tourists today visit Monticello.[41]

Perhaps the most direct use of the past's unique resource, itself, is realized in science fiction by Hollywood. In one tale,[42] after purchasing the motion picture rights to H. G. Wells' *Outline of History*, the head of a movie studio uses a time machine to send his ace cameraman into the past to get live action footage. Prehistoric animals, the ice age, Cheops building his pyramid, the destruction of Pompeii by the eruption of Vesuvius, the Battle of Hastings, Columbus, all the *originals* of these historical events appear in the final film. Years later this idea was developed even further in a very funny novel,[43] in which a movie director uses the eleventh century as a realistic setting for a picture. Realism isn't always the result, however, as portrayed in another tale; the films produced by a gadget that can 'look' into the past are failures because they don't look "authentic enough" to Hollywood moguls![44] (Fig. 1.1).

---

[36] H. M. Sycamore, "Success Story," *Magazine of Fantasy and Science Fiction*, July 1959.

[37] C. Simak, *Mastodonia*, Ballantine 1978.

[38] Two are B. W. Aldiss, "Poor Little Warrior," in *The Science Fictional Dinosaur*, Avon 1982, and a famous one by Ray Bradbury, "A Sound of Thunder," in *The Stories of Ray Bradbury*, Alfred A. Knopf 1980.

[39] C. Simak, "The Loot of Time," *Thrilling Wonder Stories*, December 1938, in which a time machine inventor raises money for his research by transporting hunters back 70,000 years to the Old Stone Age.

[40] Three such tales are R. Silberberg, *Up the Line*, Ballantine 1969 and "When We Went to See the End of the World," in *Beyond the Safe Zone*, Warner 1986; and G. Kilworth, "Let's Go to Golgotha!," in *The Songbirds of Pain*, Victor Gollancz 1985.

[41] B. Tucker "The Tourist Trade," in *Tomorrow the Stars*, Doubleday 1952.

[42] L. Laurence, "History in Reverse," *Amazing Stories*, October 1939.

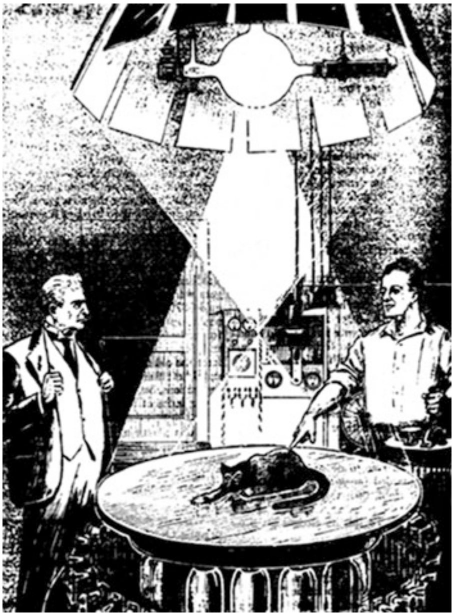

Fig. 1.1 The inventor of a time machine demonstrates it by sending the family cat on a trip. In the story the inventor, himself, travels back to 1901, where he accidently kills his grandfather in an early pulp magazine, non-paradoxical version of the famous riddle. Illustration for Raymond A. Palmer's "The Time Tragedy" (*Wonder Stories*, December 1934) by Frank R. Paul, ©1934 by Continent Publications Inc.; reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

---

[43] H. Harrison, *The Technicolor Time Machine*, Doubleday 1967.

[44] A. Derleth, "An Eye for History," in *Harrigan's File*, Arkham House 1975.

1.2 Where Are All the Time Travelers?

> "If it [time travel] *could* be done, someone will eventually learn how. If that happens, history would be littered with tourists. They'd be *everywhere*. They'd be on the *Santa Maria*, they'd be at Appomattox with [cameras], they'd be waiting outside the tomb, for God's sake, on Easter morning."[45]

The question the title of this section asks is an echo of the one the physicist Enrico Fermi (1901–1954) asked in the 1950s, about the possibility of interstellar space travel and of alien intelligent life in the universe—if such travel is possible and 'they' exist, then *where are they*? Why haven't we at least received radio signals from them? For many, the apparent lack of time travelers among us is similar evidence for the impossibility of time travel. As one famous science fiction writer put it, "The most convincing argument against time travel is the remarkable scarcity of time travelers. However unpleasant our age may appear to the future, surely one would expect scholars and students to visit us, if such a thing were possible at all. Though they might try to disguise themselves, accidents would be bound to happen—just as they would if we went back to Imperial Rome with cameras and tape recorders concealed under our nylon togas. Time traveling could never be kept secret for very long."[46]

Clarke's idea is that, from the moment after the first time machine was constructed, through all the rest of civilization, there would be numerous historians, to say nothing of weekend sightseers, who would want to visit every important historical event in recorded history. They might each come from a different time in the future, but all would arrive (according to Clarke) at destinations crowded with temporal colleagues, crowds for which there is no historical evidence!

Long before Clarke the science fiction writer Robert Silverberg had already used the same idea in his 1969 novel *Up the Line*, where it's called the *cumulative audience paradox*. That paradox claims that as time travelers to the past continue to visit certain historically interesting dates and places, there will be an ever-increasing number of people present. As it is presented in the novel, "Taken to its ultimate, the *cumulative audience paradox* yields us the picture of an audience of billions of time-travelers piled up in the past to witness the Crucifixion, filling all the Holy Land and spreading out into Turkey, in Arabia, even to India and Iran . . . Yet at the original occurrence of [that event] *no such hordes were present!*" And later in the same work, we read "A time is coming [when we] will throng the past to the choking point. We will fill our yesterdays with ourselves and crowd out our own ancestors."

---

[45] A skeptic's reaction to the idea of time travel in J. McDevitt's story "Time's Arrow" in *The Fantastic Civil War* (see note 7).

[46] Arthur C. Clarke, "About Time," in *Profiles of the Future*, Warner 1985. A story that I recall once having read (but cannot now remember either the author or the title) wonderfully illustrates Clarke's point. A time traveler in disguise at Golgotha for the Crucifixion has a camera hidden beneath his robe to avoid attracting attention. All goes well until he notices odd, clicking noises coming from all those standing near him. It is then he realizes the entire crowd is *nothing but* time travelers, from all through the ages, all with hidden cameras!

Philosophers are well aware of Silverberg's and Clarke's conundrum and, indeed, it can be found in the philosophical literature before Clarke wrote. In one paper, for example, we read "Actually I know of only one argument against the possibility of time travel that seems to carry any weight at all. This is the fact that it does not appear ever to have happened. That is, it might be argued that there will be no time trips from [2100] to [2017] because we were here in [2017] and saw no time travelers. But this argument is far from conclusive."[47] At most, in other words, the absence of temporal visitors amongst us is an objection to the *actuality* of time travel, and not to the *possibility* of time travel.

This same philosopher then mentions some ways around this concern, including one which he called a "pettifogging physical limitation on time travel: perhaps the energy expenditure varies as the fourth power of the time traversed, making only very short trips feasible, and its discovery lies too far in the future for its effects to have yet been felt." Another science fiction writer, as famous as Clarke, used that idea in his 1957 novel *The Door Into Summer* when Robert Heinlein has one character comment "Now if there was some way to photograph the Crucifixion . . . but there isn't. Not possible . . . there isn't that much power on the globe. There's an inverse-square law tied up in [time travel]." Or, perhaps, time travel *is* possible but it's so extraordinarily dangerous that it's impossible to get anyone to do it. In one provocative tale[48] that takes this idea to the extreme, we read that there is only *one* time traveler, ever, from the future—indeed, from just 18 min (!) in the future—and his first (and last) experiment destroys the Earth.

Clarke presented some other possible science fiction rebuttals to the puzzle of 'where are the time travelers?' As he wrote, "Some science fiction writers have tried to get around this [question] by suggesting that Time is a spiral; though we may not be able to move along it, we can perhaps hop from coil to coil, visiting so many millions of years apart that there is no danger of embarrassing collisions between cultures. Big game hunters from the future may have wiped out the dinosaurs, but the age of *Homo sapiens* may lie in a blind region which they cannot reach."

The idea of time as a spiral was quite popular in early science fiction. Typical is one tale[49] in which the time traveler suddenly finds himself not in 1933 but in 2189. His situation is 'explained' to him thus: "[The] time stream is curved helically in some higher dimension. In your case, a still further distortion brought two points of the coil into contact, and a sort of short circuit threw you into the higher curve."

---

[47] G. Fulmer, "Understanding Time Travel," *Southwestern Journal of Philosophy*, Spring 1980, pp. 151–156. The modern view of this 'paradox' is *not* that it describes a situation so absurd that time travel must be impossible, but rather that *all* the time travelers who *were* (will be?) at the Crucifixion *are* in the historically recorded crowd (see note 46 again). The Crucifixion happened just *once*, not over and over. I'll return to this point later in the book.

[48] D. Plachta, "The Man from When," *Worlds of If Science Fiction*, July 1966.

[49] R. H. Wilson, "A Flight Into Time," *Wonder Stories*, February 1931.

A few years later (in 1937) we find another story[50] with the same spiral-time concept; with a sixty-million year pitch to the time helix, there is no danger of a grandfather paradox. That same year spiral time was the central ‘scientific’ theme in the stage production “I Have Been Here Before” by the English playwright J. B. Priestley.

The very next year (1938) the young Isaac Asimov used the idea in his first attempt at professional writing, despite a life-long unhappiness with the concept of time travel. Titled “Cosmic Corkscrew,” it was initially rejected, indeed it was *never* published, and eventually lost. Though, perhaps, not for long, since one enterprising modern writer has used it as the basis for his own time travel tale.[51] In it a traveler from the near future travels back to 1938 to retrieve “Cosmic Corkscrew” before Asimov loses it (perhaps that’s *why* it was lost!) Even when writing introductions to other writers’ time travel tales, Asimov would often insert personal comments on his opinion of the concept. For example, in one volume of an anthology series he edited, *The Great Science Fiction Stories* (of 1954), he wrote “To my way of thinking it is precisely because time travel involves such fascinating paradoxes that we can conclude, even in the absence of other evidence, that time travel is impossible.” And in *The Great Science Fiction Stories* (of 1961) he bluntly declared “I think scientists who think up methods of time travel are probably all wrong.”

Spiral time is a close cousin to circular time. One story[52] dealing with circular time has a time traveler who finds, after a trip one hundred years into the future, that he can’t get all the way back to his own time because the required energy rises exponentially with increasing penetration into the *past*. Still, it’s very cheap in energy to go forward in time in this tale, and that’s what the traveler does, in search of help from the future’s advanced technology. He never finds what he needs, however, and so goes forward right into the collapse of the universe and through a new Big Crunch that forms an identical new cycle of time. He thereby returns home to just before he left.[53] This eternal recycling of identical, circular time is so terrifying that the traveler decides to suppress what he has learned about how to time travel—and so maybe *that’s* why there are no apparent time travelers. Science fiction writer Larry Niven has argued, however (see note 27), that while this may be a conceptually valid (?) way to travel into the past, he also warns that “Removing your time machine from the reaction of the Big Bang/Crunch could change the final configuration of matter, giving an entirely different . . . history.” (I strongly suspect that Niven wrote that with a big smile on his face!)

---

[50] P. S. Miller, “The Sands of Time,” *Astounding Stories*, April 1937.

[51] M. A. Burstein, “Cosmic Corkscrew,” *Analog*, June 1998.

[52] P. Anderson, “Flight to Forever,” *Super Science Stories*, November 1950.

[53] Turning this idea on its head is the approach of the 1978 novel *The Way Back* (DAW) by A. B. Chandler. Its characters return from the past to their own time by traveling even further backward, right through the Big Bang and into the *previous* (and identical) cycle of time.

The famed English physicist Stephen Hawking is so taken with the question of 'where are all the time travelers?" that he has elevated their apparent absence ("we have not been invaded by hordes of tourists from the future"[54]) to the status of being a demonstration of the impossibility of time travel to the past. His so-called *Chronology Protection Conjecture*, Hawking likes to say, "makes the universe safe for historians": that is, there is nothing to worry about (if you're concerned that time travelers could change the past) because time travel is simply impossible. You'll see later in the book that there are other possible ways to insure the safety of history *without* denying the possibility of time travel, and Hawking himself has backed away just a bit from the Conjecture, saying now that he was simply looking for a humorous line.

While Hawking's endorsement of it has made the Conjecture famous, he wasn't the first to state it. Two years earlier it had appeared, in of all places, a *financial* publication: "[If] time travel was possible, someone from the future would eventually either discover a time tunnel or build a time machine and come visit us."[55] And even before that, an economist presented a 'proof' for concluding that "time travelers do not and cannot exist."[56] He argued that if time travelers from the future were actually amongst us (our 'now' is their 'past') then, by virtue of their knowledge of things to come (our 'future') they would make financial deals so numerous and extensive that interest rates would be driven to zero. Interest rates are *not* zero, however, and thus no such time travel hanky-panky has occurred.

These sorts of financial arguments aren't like to convince many physicists or philosophers of the Conjecture's merit. At most we can only conclude from them that time travelers from the future have *not* influenced financial affairs, which doesn't mean they aren't here. In any case, the Conjecture was actually stated more than 20 years before Hawking by Larry Niven (see note 27), who declared what is called *Niven's Law*: "If the universe of discourse permits the possibility of time travel, and of changing the past, then no time machine will be invented in that universe." And Hawking's concern over time travelers meddling with the past was anticipated in science fiction, too, by at least half a century; in a 1950 tale, for example, we learn of a Master Historian, and the graduate students in his course on 'Experimental History' in the forty-sixth century, trying to correct a problem created by a previous tampering with the past![57]

Not all physicists and philosophers feel intellectually comfortable with the Conjecture, as it seems (to them) a too quick surrender: 'Time travel is a problem

---

[54] S. W. Hawking, "Chronology Protection Conjecture," *Physical Review D*, July 15, 1992, pp. 603–611. See also J. F. Woodward, "Making the Universe Safe for Historians: Time Travel and the Laws of Physics," *Foundations of Physics Letters*, February 1995, pp. 1–39.

[55] J. Queenan, "Time Warp: Or, Investing in the Future Is a Bust," *Barron's*, January 8, 1990, p. 46.

[56] M. R. Reinganum, "Is Time Travel Impossible? A Financial Proof," *Journal of Portfolio Management*, Fall 1986, pp. 10–12.

[57] L. Jones, "Sunday is Three Thousand Miles Away," *Thrilling Wonder Stories*, June 1950. A more recent, two-*novel* treatment of historians tinkering with history is by Connie Willis (*Blackout* and *All Clear*, both published in 2010).

so hard to do let's simply *define* it to be non-existent and then we won't have to worry about it anymore.' To *really* show that time travel is impossible, however, one needs to demonstrate how it would violate one or more of the laws of physics. Hawking, of course, understands this and has stated that, as one who is no fan of time machines and time travel, he believes there is new physics yet to be discovered that *will* forbid would-be time travelers from roaming up and down the centuries. Finding that new physics is the lure the study of time machines has for him. As he correctly writes in his autobiography,[58] "Even if it turns out that time travel is impossible, it is important that we understand why it is impossible."

One mathematical physicist who agrees with Hawking on the matter of the unlikely possibility of making a time machine is the New Zealand theoretician Matt Visser. Noting that while quantum field theory, and the general theory of relativity, are each amazingly good theories in many applications within their respective realms, they are not so good in spacetime regions at the so-called *Planck scale* (that is, when the density of mass-energy reaches the fantastic level of 10^{94} grams/cm^3 and beyond) where chronology violations (that is, time travel) seem to be spawned. As Visser has observed,[59] this situation won't change until 'we wander into the guts of quantum gravity,' the unification that will merge gravity with the quantum to give a theory that *always* works. Without quantum gravity, physics will continue to be "infested" (Visser's word) with "sick" (Visser's word) spacetimes that allow time travel. Visser believes that the discovery of the theory of quantum gravity can be 'guided' by *building* causality into it,[60] and the result will finally consign time machines to where (in his mind) they belong, the dust-bin of crackpot physics.

Well, perhaps so, but we don't have a quantum theory of gravity yet, and probably won't for some time to come, and so the puzzling questions about time travel remain. To end this section on a slightly gloomy note, an idea appeared in science fiction,[61] when Hawking was still a teenager, offering a possible rebuttal to the Conjecture. It opens with one of the inventors of the first time machine just returning from a trip to the past of 1938. Still, despite this success, the inventors are puzzled by what they call 'the problem': "But if *we* have time traveled, then obviously men of the future have time traveled. They will be able—*are* able to come back. [So] where are they?" They finally conclude that there can only be two possible answers. Either there is nobody in the future, or time travel is so dangerous (is that why the future might be empty—humanity misused time travel and killed itself off?) that all who invent it will suppress it. And that's what they decide *they* must do.

---

[58] Stephen Hawking, *My Brief History*, Bantam 2013, p. 113.

[59] Matt Visser, "The Quantum Physics of Chronology Protection," in *The Future of Theoretical Physics and Cosmology*, Cambridge 2003. This paper was Visser's contribution to the celebration of Hawking's 60th birthday, held in January 2002.

[60] This may seem like something new, but it really isn't. General relativity has causality built into it on a *local* level (where it belongs); a failure of causality (that is, time travel—see J. Sharkey, "The Trouble With Hyperspace," *Fantastic* April 1965) occurs in general relativity only when one studies *global* regions of certain spacetimes. Forcing a physical theory to have a *prescribed global behavior* would be to undo all of physics since the development of *local field theories*, along with all their amazing successes in explaining nature.

[61] M. Shaara, "Time Payment," *Magazine of Fantasy and Science Fiction*, June 1954.

1.3 Skepticism About Tales of Time Travel

> "May it not be that our inability to leap into the fiftieth century, A.D., seems impossible to us, merely because of certain prejudices we entertain or certain facts and tricks of which we are still hopelessly ignorant? Assuredly, this is not a foolish query. Its answer, whatever that may be, carries immeasurable consequences for metaphysics."
>
> —a scholar wonders[62]

A thought-provoking possibility for explaining the scarcity of certified time travelers is the central thesis of a fascinating paper in the philosophical literature. The author of that paper argues (note 13) that nobody would believe a time traveler even if he willingly confessed and revealed his knowledge of the future, or even gave the details of his time machine. He goes on to make the astonishing assertion that even the *time traveler himself* would have doubts! This perhaps shocking suggestion deserves some elaboration, especially because it invokes the authority of the patron saint of skeptics for support, the Scot David Hume (1711–1776). The crucial point to keep in mind is explicitly stated in the argument: "The key question will not be 'Is time travel possible?' We shall instead ask whether it is possible to justify a belief in a report of time travel." This gets to the real heart of Clarke's puzzle from the previous section.

Much of the resistance to the idea of time travel lies in sheer skepticism. For many, time travel (to the past, in particular) is simply too much out of the ordinary to be taken seriously. For many, time travel would literally be miraculous. Hume's great work, *An Enquiry Concerning Human Understanding*,[63] contains a section on how a rational person should react to a claim that a miracle has occurred. Hume proclaimed that a miracle *by definition* violates scientific law and that, because such laws are rooted in "firm and unalterable experience," any violation of one or more of these laws immediately provides a refutation of the report of a miracle. In Hume's own words:

> "Nothing is esteemed a miracle, if it ever happened in the common course of nature. It is no miracle that a man, seemingly in good health, should die on a sudden; because such a kind of death, though more unusual than any other, has yet been frequently observed to happen. But it is a miracle, that a dead man should come to life; because that has never been

---

[62] W. B. Pitkin, "Time and Pure Activity," *Journal of Philosophy, Psychology and Scientific Methods*, August 27, 1914, pp. 521–526. Pitkin's essay was a critique of time travel as presented in Wells' *The Time Machine,* which Pitkin called "one of the wildest flights of literary fancy."

[63] Making its first appearance in 1748, *Enquiry* has been reprinted numerous times since. I used the 1963 edition published by Open Court.

> observed in any age or country . . . When anyone tells me, that he saw a dead man come to life, I immediately consider with myself, whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority, which I discover, I pronounce my decision, *and always reject the greater miracle* [my emphasis]."[64]

It is a strict interpretation of Hume that Sorenson (note 13) has adopted in claiming that a time traveler would have no success (among rational persons) with tales of 'different times.' As he explains, "Clearly the time traveler cannot persuade a reasonable person by baldly asserting 'I am a time traveler.' The improbability of his claim places a heavy burden of proof on him. But perhaps he could shoulder the burden by means of artifacts, predictions, and demonstrations." Sorenson dismisses all of these possibilities, however, by reminding us of the slightly sleazy history of parapsychology and ESP, both of which run counter to known scientific laws, but which have still duped "many a respected scientist." Any artifact, prediction, or demonstration of time travel, argues Sorenson, is more likely to be the result of deception and fraud than of actual time travel: "Should the time traveler take observers for a spin in his time machine, the skeptics will have us compare their adventures with séances." The rational reaction to such a spin around the centuries, according to Sorenson's presentation, would be like that of a magician who cannot figure out how a colleague has just done his newest act: 'Nice trick! How did you do it?'[65]

The time traveling tourist stranded in the past in one story is used to getting a skeptical reaction because he can provide his questioners no technical explanation for his situation. "How the hell should I know? I'm just a tourist. It has something to do with chronons [see the Glossary]. Temporal Uncertainty Principle. Conservation of coincidence. I'm no engineer."[66] Somewhat more successful (perhaps) is a time traveler born in 2003 who turns up in 1975. After he tries to convince an interrogator of how that can be, he apparently succeeds. As the time traveler later tells a new friend in the past of 1975, "What amazed me . . . was that he really believed me in the end." But the friend doesn't buy that, replying "He did? I think he just pretended. A scientist isn't likely to believe a thing that is against all logic."[67] If the reception committee is a crowd of conservative, cautious Humeans, it would seem that a time traveler is almost certainly doomed. Early science fiction time travelers from 2030, for example, were warned about receiving a skeptical response as follows (the editorial introduction to this tale[68] called it "a curious study of psychology"): "Our wisest men advised against [our trip to the past]. They said we could hope to be received only as imposters and fakirs, that . . . we would find only twentieth-century barbarians, suspicious, ill-tempered, likely to do us bodily harm."

---

[64] What Hume is alluding to here should be plain; as expressed in P. Heath, "The Incredulous Hume," *American Philosophical Quarterly*, April 1976, pp. 159–163, Hume was "an exposer of bad arguments in rational theology." For Hume, second-hand (or even more remote) tales of the return of a man from the dead—the claim that literally kept Christianity alive after Christ's execution—were suspect.

[65] This skeptical reaction was nicely captured in the story "E for Effort" by T. L. Sherred (*Astounding Science Fiction*, May 1947). As one character laments, "I've watched scribes indite the books that burnt at Alexandria; who would buy, or who would believe me, if I copied one. . . . What sort of padded cell would I get if I showed up with a photograph of Washington or Caesar? Or Christ?" The padded cell was indeed the fate of the time traveler in "The Ambassador from the 21st Century" (*Startling Stories*, March 1953) by H. J. Shay, the story of a man who journeyed from A.D. 2007 back to 1952 to warn of a future war; he was committed to a mental institution to receive help for his "illusion."

[66] J. Haldeman, "No Future In It," *Omni*, April 1979.

For many, such skeptical reactions to self-proclaiming time travelers seems dogmatic in the extreme—the response of people with no imagination, no spirit, and heads full of cement. Humean skepticism requires, so it would seem, the rejection of anything and everything that is profoundly surprising, leaving the world a place of utter predictability and boredom. As one science fiction writer put it, "When the miraculous occurs, only dull, workaday mentalities are unable to accept it."[69] Sorenson answers this harsh criticism as follows: "Humeans respond [to Sheckley] by distinguishing between surprises. Most surprises in science do not violate accepted scientific laws. The strange wildlife in Australia was not excluded by biology, X-rays were not precluded by physics."

Sorenson does well, however, to avoid mentioning such profound surprises as, for example, the spectrum of black-body radiation and, later, the photoelectric effect, which were not in the domain of known classical science at the beginning of the twentieth century. Those puzzling, surprising, *totally mystifying* effects required new science—the discovery of the quantum concept by Max Planck. (Explaining the photoelectric effect, *not relativity*, is what won Einstein his Nobel prize.) A strict Victorian-age Humean, as described by Sorenson, would have wrongly rejected the experimental reports of all quantum phenomena and would also (perhaps just as wrongly) have rejected all reports of time travel.

A strict Humean definition (as described by Sorenson) that a miracle has occurred requires a violation of one or more of the *known* [my emphasis] scientific laws of nature.[70] As one modern philosopher defines a miracle, it is any event that "can be explained *only* [my emphasis] by reference to the intervention of a supernatural force."[71] Time travel, by that interpretation, is *not* a miracle because general relativity, not God, is all that is required. C. S. Lewis (1898–1963), late professor of Medieval and Renaissance Literature at Cambridge University, however, absolutely rejected Hume's view on how a rational person should react to certain surprising events. Lewis, one of the most thoughtful modern writers on Christian theology, had no patience with skeptics (or, as he called them, materialists).[72]

---

[67] G. Gor, "The Garden," in *Russian Science Fiction* (R. Magidoff, editor), New York University Press 1969.

[68] P. Bolton, "The Time Hoaxers," *Amazing Stories*, August 1931.

[69] Robert Sheckley, "Something for Nothing," in *Citizen in Space*, Ballantine 1955.

[70] The word *known* is important. As a character in one early science fiction story puts it, "These things [four-dimensional object] sound like miracles; but, after all, what are miracles but phenomena which, *on account of our ignorance* [my emphasis], we cannot explain?" See B. Olsen, "The Four-Dimensional Roller-Press," *Amazing Stories*, June 1927.

[71] D. M. Ahern, "Miracles and Physical Impossibility," *Canadian Journal of Philosophy*, March 1977, pp. 71–79.

Professor Lewis graphically illustrated the dug-in position of the extreme skeptic as follows: "If the end of the world appeared in all the literal trappings of the Apocalypse; if the modern materialist saw with his own eyes the heavens rolled up and the great white throne appearing, if he had the sensation of being himself hurled into the Lake of Fire, he would continue forever, in that lake itself, to regard his experience as an illusion and to find the explanation of it in psychoanalysis, or cerebral pathology."[73] If the end of the world would receive such a skeptical response, then a mere time traveler would surely have no hope at all of being believed.

Lewis would certainly have rejected Sorenson's most astonishing assertion: "So far I have concentrated on the time travel question from the perspective of the time traveler's audience. What about the time traveler himself? Can he at least know he is a time traveler?" Sorenson argues that a time traveler, if authentic, should be able to convince his audience, and that if he can't (and he *cannot* if they are true Humean skeptics), then the time traveler must entertain doubts, too! It doesn't matter (says Sorenson) that the time traveler has memories of his adventures, and it doesn't matter that he knows in his heart that he speaks the truth. Using words that echo Lewis' sarcasm, Sorenson quickly dismisses the importance of the time traveler's self-knowledge, declaring such memories to be merely the symptoms of some deep psychosis, and the traveler's introspective sincerity to be a product of gross self-deception.

Sorenson specifically mentions the traditional Humean response to astonishing reports when he cites earlier writers on time travel in the philosophical literature. In one of those analyses, for example, we find an argument *for* the reasonableness of a rational belief in time travel ("I have been amused and irritated by the spate of articles proving that time travel is a 'conceptual impossibility') by claiming such proofs must be faulty because there is a mathematically consistent explanation for such a belief.[74] (This author was referring to spacetime diagrams, which we'll get to in Chap. 3.) This paper received a very sharp rebuttal from another philosopher who convincingly used fundamental physics to show a simple use of spacetime diagrams in a special relativity setting does *not* support time travel to the past.[75] (I'll return to this point in Chap. 3.) Even later a Humean-style rebuttal came from yet another philosopher, who showed how to explain the time travel phenomenon that Putnam (note 74) described *without* invoking time travel.[76] This isn't to say that Weingard doesn't invoke some pretty astonishing gadgetry (and more) himself, like matter transmitters and anti-matter humans. (You'll see how anti-matter ties-in with time travel a bit later in the book.) A resurrected Hume would surely applaud these rebuttal analyses (although he might also doubt his own fresh existence).

---

[72] In Lewis' eerie, unfinished story "The Dark Tower," a tale of the 'chronoscope,' a gadget that "does to time what the telescope does to space," the persistent skeptic in the story is a Scot, surely created in the image of Hume. See C. S. Lewis, *The Dark Tower and Other Stories*, Harcourt 1977.

[73] C. S. Lewis, *The Grand Miracle*, Ballantine 1986.

[74] H. Putnam, "It Ain't Necessarily So," *Journal of Philosophy*, October 1962, pp. 658–671.

[75] J. Earman, "On Going Backward in Time," *Philosophy of Science*, September 1967, pp. 211–222.

Hardly anybody is happy with Weingard's approach for avoiding time travel (including, I suspect, even Weingard). His 'explanations' seem, just like a time machine, to be incredible and, as Arthur Conan Doyle's Professor Challenger says in one tale not staring Sherlock Holmes, "You cannot explain one incredible thing by quoting another incredible thing."[77] An interesting science fiction exposition illustrating Professor Challenger's Humean philosophy occurs when a copy of *The New York Times* for December 1 shows up for some subscribers a week early, on November 22. It seems the only explanation is either that the paper really is from the future (due to some sort of fluke of the fourth dimension), or that it is a hoax. The first-person narrator of this 1973 tale[78] provides us with his reason for believing the former: "I don't find either notion easy to believe but I can accept the fourth-dimensional hocus-pocus more readily than I can the idea of a hoax." Hume couldn't have said it better.

It should be clearly understood that Hume was not arguing for disbelief in absolutely anything surprising, but rather for rational analysis. Historically, the context of Hume's times was that of what he took to be non-rational arguments for a belief in God, particularly those 'proofs' so beloved by theologians based on Design (Heath [note 64] calls such 'proofs' "philosophical museum pieces"). As Heath writes, "Hume ... makes no attempt to deny the supposed facts; he simply argues that they are consistent with other explanations and other analogies of a less ambitious kind. There is no right to attribute to the causes of such phenomena abilities more extensive than are needed to produce the observed effects."

As a matter of fact, even Hume could be convinced of quite strange matters, and I think Sorenson does interpret the philosopher a little too narrowly. In his essay concerning Hume's position on holding a belief in God, Heath wonders whether there is "empirical evidence [imaginable] which would persuade any reasonable mind of the real existence of an infinite God." Heath answers his own question as follows: "If the stars and galaxies were to shift overnight in the firmament, rearranging themselves so as to spell out, in various languages, such slogans as I AM THAT I AM, or GOD IS LOVE—well, the fastidious might consider that it was all very vulgar, but would anyone lose much time in admitting that this settled the matter? … Confronted with such a demonstration, the hard-line Humean [but not Hume, himself, I think] could continue, of course, to argue that, for all its colossal scale, the performance is still finite, and so cannot be evidence of more than the finite, though immense power that is needed to achieve it."

---

[76] R. Weingard, "On Travelling Backward in Time," *Synthese*, July–August 1972, pp. 117–132.

[77] Arthur Conan Doyle, "The Disintegration Machine," *The Strand Magazine*, January 1929. Professor Challenger is nothing like Wells' thoughtful Time Traveller; in the original 1912 Challenger novel *The Lost World*, he was described as a "primitive cave-man in a lounge suit."

[78] R. Silverberg, "What We Learned from This Morning's Newspaper," in *Beyond the Safe Zone*, Warner 1986.

Skepticism about 'time travel' was around long before the specific idea of a 'time machine' was conceived. For example, the eleventh-century Persian poet-philosopher Omar Khayyam was blunt in his evaluation of the likelihood of reliving the past. As he so beautifully wrote in one of the quatrains of the *Rubaiyat*,

The Moving Finger writes; and having writ,
Moves on: nor all your Piety nor Wit
Shall lure it back to cancel half a Line,
Nor all your Tears wash out a Word of it.

Quite a bit later the English poet Thomas Heywood, in his 1607 play *A Woman Killed with Kindness*, had one of his characters express a similar thought:

God, O God, that it were possible
To undo things done, to call back yesterday;
That Time could turn up his swift sandy glass
To untell the days, and to redeem these hours.
Or that the Sun
Could, rising from the west, draw his coach backward,
Take from the account of Time so many minutes,
Till he had all these seasons called again,
………………………………
But O! I talk of things impossible,
And cast beyond the moon…

When Gödel's discovery of time travel in his rotating universe was announced, the skeptics were easy to find. One philosopher[79] wrote of it "This property [of time travel] must be judged an absurdity by anyone committed to the ordinary modes of speech." And another[80] was only slightly less charitable: Gödel's solution was a "bizarre conception" and a "mere mathematical curiosity." Science fiction wasn't immune to skepticism, either, even though you might have expected that to be the one place where the high drama of time travel would be welcomed. Four *years* after Gödel's paper appeared we find one respected anthologist writing,[81] as part of his introduction to a story, "In this tale we meet our first Mad Scientist. Just as in reality the thoroughly cracked pots used to be found inventing perpetual-motion machines, so in science fiction we find the lunatic fringe more often than not trying to perfect time-travel mechanisms." And that same year the founding editor of *Galaxy Science Fiction Magazine* declared "Time travel requires a suspension of disbelief that is almost unbelievable . . . Scientifically, time travel can't stand inspection.".[82]

---

[79] J. D. North, *The Measure of the Universe*, Oxford University Press 1965.

[80] C. T. K. Chari, "Time Reversal, Information Theory, and 'World-Geometry'," *Journal of Philosophy*, September 1960, pp. 579–583.

[81] Groff Conklin, editor of *Science-Fiction Adventures in Dimension*, Vanguard 1952.

Years later matters had not much changed. For example, in his marvelous 1985 book *The Past is a Foreign Country*, David Lowenthal repeatedly refers to time travel as "fantasy," and to science fiction stories about time travel as "unbridled by common sense." (Lowenthal is a professor of *geography*, not physics.) Science fiction writers were still often not much more enthusiastic about time travel. The well-known science fiction writer and critic Alexei Panshin, for example, agrees with Lowenthal, at one point, long after Gödel, writing "Time travel is a philosophical concept, not a scientific one. It is, in fact, as has often been pointed out, scientific nonsense."[83]

Skepticism does have its uses, however. Modern science fiction writers have often used it as a dramatic means of building conflict and tension in their time travel stories. A skeptical reception is extreme, for example, for a soldier-in-time who has fought in numerous wars, from the ancient past to a billion years in the future.[84] He finds that nobody believes him when he speaks openly of his temporal adventures during a visit to a present-day bar. Everybody merely thinks it is all a hilarious gag. This is in great contrast to one 1870s story[85] in which suspicion of a stranger plays a central role, but which finds its offered explanation in something entirely different from time travel. It tells of a man who suddenly appears in the midst of a Union military camp during the American Civil War.

This man quickly displays strange lapses in his background, as well as possessing knowledge of many different things well beyond anything that could be called common. The details of the story are not important for us but, if it were published in a modern science fiction magazine, this man would almost surely be identified in most readers' minds as a time traveler. In 1875, however, the author's narrator found his punch line in "his firm conviction that the quiet, gentle, well-behaved, modest gentleman, so singularly gifted . . . is, in plain terms, the devil!" Time travel certainly never entered the author's thoughts or, if it did, he lost his nerve at the idea of using it in this pre-Wells story. You'll recall from the opening of the Introduction that it was this 'use of the devil to explain mysterious happenings' that Wells wanted to move away from, and that was the motivation for his introduction of a time *machine*.

Hollywood has at least gotten the skeptical part of the psychology of time travel right (later discussions in this book will focus on how film makers have been less successful with the physics). When, for example, the time traveling villain in the

---

[82] H. L. Gold, editor of *The Galaxy Reader of Science Fiction*, Crown 1952.

[83] In his introduction to Robert Heinlein's classic time travel tale (to be discussed later) "All You Zombies—," in *The Mirror of Infinity*, Canfield 1970.

[84] F. Leiber, "The Oldest Soldier," *Fantastic*, May 1960.

[85] G. C. Eggleston, "Who Is Russell?" *American Homes*, March 1875.

1989 movie *Time Trackers* is confronted in the medieval past, he simply laughs-off a threat to reveal his true identity. "Go ahead," he says in effect, 'the only thing your talk of time machines from the future will accomplish is for people to think you are crazy!'

What would Arthur C. Clarke have thought of all this skepticism being directed toward those who claim to have a time machine? His thoughts about the difficulty time travelers would have in maintaining low profiles were what started the previous section, after all. My guess is that he would have had little patience with extreme incredulity. The surprise of being confronted by a time traveler would soon have turned to awe and pleasure IF—and I emphasize the IF—Clarke had been taken for a spin around the centuries in the stranger's machine. He would surely have ended-up quoting his own famous 'third law' to explain the wonder of it all: "Any sufficiently advanced technology is indistinguishable from magic."

Near the end of his paper, Heath writes what I think is the perfect rebuttal to anyone who would refuse to admit to time travel, even after taking a quick trip backward a few tens of millions of years to the late-Mesozoic era to hunt *Tyrannosaurus rex*, and even after seeing instant photographs of the dead monster with the skeptic's *own foot* on the great creature's head, or of his *own boots* dripping a bloody puddle of unholy size on the floor of the time machine. Writing about the Humean-unconvinced, even when faced with a rearranged firmament, Heath observes "But this now seems a cavil, designed only to prove that even omnipotence is powerless against the extremer forms of skeptical intransigence." Where God would fail to convince, a simple time traveler could hardly hope to do better!

1.4 Troubles with (some) Time Machines

> "If you don't stop this senseless theorizing upon something that's an obvious impossibility, you'll find yourself working alone! Your ridiculous ideas sound like the ravings of a madman. Anyone with average intelligence realizes that the mere thought of traveling through time is absurd."[86]

If the previous section seemed just a bit gloomy concerning time travel, there is a very big reason for that. The sentiment expressed in the above opening quote to this section was a common one among philosophers long before physicists began to seriously think on the topic. While there *are* issues with time travel to the *future*, they are of an *engineering* nature, centered on how to build a big enough rocket ship with enough fuel to make the high speed, looping trip out into space and back again described in the opening section of this chapter. 'Mere' engineering problems are of no concern to physicists and philosophers. What *does* concern them are the far deeper puzzles of time travel to the *past*, the puzzles presented by what appear to be

---

[86] A science fiction physicist receives harsh criticism from a colleague in L. A. Eshbach's "Out of the Past," *Tales of Wonder*, Autumn 1938.

logical paradoxes. Before we get into the paradoxes, however, we need to first clear our minds of two common, popular notions of just what a time machine *is*. Both are false notions (one is due to H. G. Wells) which are, today, *rejected* by physicists (and most philosophers, too).[87]

As you'll see later in the book, all the theoretical time machines that have appeared in the modern physics literature involve spatial displacement. That is, they require *movement*. (On this point, the speedy DeLorean time car in the *Back to the Future* films has it right.) Wells' time machine, however, did not move; it always remained in the Time Traveller's laboratory (or at least on the spot where the laboratory would have been) unless he pushed it about after a trip in time. This, alas, results in a particularly troublesome problem: a Wellsian time machine heading into the past would run into itself!

Consider: There sits my time machine as I prepare for the first time journey ever, a trip back to the late-Mesozoic era to hunt dinosaur. I load my Continental. 600 super-high-power rifle with Nitro Express cartridges the size of bananas, kiss my wife good-bye, and climb in. I pull the lever. Now, Wellsian-type time machines don't *jump over* time but rather travel *through* time (see the Time Traveller's own description of how things looked to him, a description faithfully and spectacularly reproduced in the 1960 film). Therefore, the time machine will instantly collide with itself at the micro-moment *before* I pull the lever!

The resulting destruction obviously introduces a nice paradox: Given that this happens before I pull the lever, how did I manage to pull it? Many of the early science fiction writers were not totally oblivious to this collision problem and, in order to avoid materializing inside of an object in the future or the past, it was common to combine the time machine with an airplane.[88] Even that though might not be enough, as one writer thought a Wellsian time traveler would get "a severe case of the bends" if his body materialized in air![89] Of course, one might argue that Wells' machine does actually move because it is attached to the Earth, which is certainly moving, but it is not clear why this should result in the time machine arriving in the temporal past of the Earth, rather than in some past region of space (almost surely a vacuum).[90]

---

[87] Both of these notions still routinely appears in science fiction, however, because they are 'just too neat' to let 'mere physics' get in the way of a good tale. I use one, without apology, in my own story "Newton's Gift" in Appendix B.

[88] Three such tales are M. J. Breur, "The Time Valve," *Wonder Stories*, July 1930; F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931; E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

[89] J. Lafleur, "Time as a Fourth Dimension," *Journal of Philosophy*, March 1940, pp. 169–178, and "Marvelous Voyages—H. G. Wells' *The Time Machine*," *Popular Astronomy*, October 1943.

[90] Philosophers seem to be becoming more aware of the collison problem (which they have dubbed "the double occupancy problem"), and at least three papers published since the second edition of *Time Machines* discuss it: W. Grey, "Troubles With Time Travel," *Philosophy*, January 1999, pp. 55–70; P. Dowe, "The Case for Time Travel," *Philosophy*, July 2000, pp. 441–451; R. Le Poidevin, "The Cheshire Cat Problem and Other Spatial Obstacles to Backward Time Travel," *The Monist*, July 2005, pp. 336–352. Physicists don't concern themselves with the collison problem simply because they *aren't interested* in *Wellsian* time machines; I'll explain why I say this by the end of this section.

The general problem of 'where the past is' was nicely illustrated by the physicist Gregory Benford in his 1980 novel *Timescape*. In that story the world of 1998 is on the verge of total ecological collapse, and an attempt is made to change the past by aiming a backward-in-time message via faster-than-light tachyons (these hypothetical particles are discussed in Chap. 5) at the pivotal year 1963. When the principal scientist involved in this effort is explaining the process to a potential financial backer, he is asked, "Hold on. Aim for *what*? Where *is* 1963?" The scientist replies, "Quite far away, as it works out. Since 1963, the Earth's been going around the Sun, while the Sun itself is revolving around the hub of the galaxy, and so on. Add that up, and you find 1963 is pretty distant." An understanding of the question 'Where is the past?' actually goes quite a bit further back in science fiction. For example, after looking through a TV-like gadget to view the past, one character in a 1940s story complains, "You said you'd find Captain Kidd's treasure, but all I can see is fog and static." He is told that's because "It's too far back—1698 or thereabouts. The Earth was billions of miles from here then, and there are too many cosmic rays between."[91]

But let's suppose we ignore this concern about where things are for a time traveler, as do most science fiction stories. Still another problem with a true Wellsian-type time machine is that because it travels *through* time, the machine must *always* appear to be located in the same place. For example, to travel from Ford's Theater today to Ford's Theater on the evening of Good Friday, April 14, 1865, in a misguided attempt to save Lincoln from Booth's bullet (*why* this would be misguided will be discussed at length later in the book), a Wellsian-type time machine would have to occupy every instant of the intervening century and more. For observers outside the machine, the machine would appear to have been sitting in the same place all those years. There is an amusing illustration of a failure to understand this point by the scriptwriters of the 1989 film *Time Trackers*, who have time travelers 'hide' their Wellsian machine from accidental discovery by 'parking' it 5 s in the future!

Wells was well-aware of the "does a time traveling object disappear or not?" issue, and tried to have it both ways in *The Time Machine* by invoking what he had the Time Traveller call "diluted presentation." As we are told in the novel, the reason why we cannot see the model time machine he sends on its way into the future as a demonstration is that "the spoke of a wheel spinning, or a bullet flying through the air" is invisible because if those objects are "traveling through time fifty times or a hundred times faster than we are . . . the impression [they create] will of course be only one-fiftieth or one-hundredth." Similarly for the model. This explanation breaks down when one remembers that, even if you cannot see the spoke or bullet, they are still there and you can get in their way—Wells, unfortunately, has one of his characters stick a hand into the space where the model time machine was last seen.

---

[91] M. Jameson, "Dead End," *Thrilling Wonder Stories*, March 1941. The "cosmic rays" are presumably the cause of the interference.

This objection to Wellsian time travel was raised soon after the 1895 publication of the novel, and then again in 1914 by Pitkin (note 62), who noted that violent disaster awaited once the time journey ended. Wells, it is only fair to note, seemingly anticipated Pitkin when he had the Time Traveller say "So long as I travelled at a high velocity *through time* [my emphasis] . . . I was, so to speak, attenuated—was slipping like a vapor through the interstices of intervening substances!" What this is getting at is that for the Time Traveller to stop 'inside' anything (Pitkin's example was the pile of bricks the Time Traveller's laboratory is certain one day to become) would, as Wells had his hero say, cause "a profound chemical[92] reaction—possibly a far-reaching explosion—[that would] blow myself and my apparatus out of all possible dimensions." Just why this spectacular event doesn't occur when the time machine simply stops *in air*, never mind inside Pitkin's pile of bricks, is never addressed.

In any case, it seems clear from all of this that Wells' machine travels *through time*, just as the Time Traveller claims. But Wells, himself, raises doubt when he describes the observed effects of a departing time machine. At the beginning of the novel, when the Time Traveller sends his model machine into the future, we read "There was a breath of wind, and the lamp flame jumped. One of the candles on the mantel was blown out . . . and it [the model time machine] was gone—vanished!" And, at the end, when the Time Traveller makes his final exit, the narrator of the tale just misses the departure but tells us "A gust of air whirled around me as I opened the door, and from within came the sound of broken glass falling on the floor. The Time Traveller was not there . . . Save for a subsiding stir of dust, the further end of the laboratory was empty. A pane of the skylight had, apparently, just been blown in." Both of these descriptions read as *implosions*, air rushing in to fill a spatial void, as though the time machines had *jumped* in time. Is there an inconsistency here? Well, perhaps not, if one accepts the curious idea of "slipping like a vapor" for an operational Wellsian-type time machine.

One famous science fiction story[93] nicely illustrates these points. The inventor of the first time machine demonstrates it to colleagues by sending a brass cube 5 min into the future. After being placed in the machine, the cube vanishes and then, 5 min later, reappears. Did the cube travel *through* time, or was its journey 'instantaneous,' so to speak? If *through* time, the cube was present at every instant after the start of its trip—so why did it vanish? The cube gets to each instant before the observers do, but why this should produce the visual effect of disappearing is unclear. The description in the story implies the cube traveled 5 min into the future without existing at any of the in-between instants, and so the story's time machine certainly was *not* Wellsian.

An immediate implication of the immobility of a Wellsian time machine is that if you are being chased by an angry mob somewhen in time (perhaps because you unwittingly violated a sensitive social taboo), then hopping into your Wellsian-type time machine isn't going to help because the machine just sits there. The mob could simply take its deliberate time in first building a roaring fire and then pushing the machine (and you) into it. As one author (with a wonderfully appropriate name!) expressed this, "You might as well try to escape by taking a nap."[94]

---

[92] Actually *nuclear*, but don't forget when Wells wrote his novel.

[93] F. Brown, "Experiment," *Galaxy Science Fiction*, February 1954.

Pulp science fiction, always alert to a good story gimmick, used this characteristic of Wellsian time machines in one clever tale[95] in which a criminal attempts to hide his crimes by sending the bodies of his victims into the far future. His mistake is to use a Wellsian time machine in which he escapes into the future. The police, however, having learned of his foul deeds, simply build a cage around the machine and arrest him when he exits 23 years later!

If a Wellsian time machine that moves *through* time suffers from a fatal collision problem, then how about that other favorite of science fiction, a time machine that *jumps* in time? (Recall the final departure of Wells' Time Traveller.) That certainly would avoid the self-colliding problem. When you pull the lever inside the machine you simply disappear from 'now' and (from your point of view) then instantly pop into existence 'then.' The problem with this sort of time machine is that a Time Traveller who uses it will have a *discontinuous* world line, with the break occurring at the moment his time machine 'jumps.' In the modern physicist's view of time travel, however, based on general relativity, a Time Traveller's world line should always be continuous. That's because general relativity is a smooth, local field theory described by differential equations, resulting in *continuous* CTLs/CTCs.

Imagine, for example, that a 'jumping' time machine inventor starts building his gadget at time $t = A$ and expects to finish building it at time $t = B > A$. At time $t = C < B$, however, he runs into a problem. Fortunately, just at that moment a fully-functional time machine suddenly appears in the lab, and from it emerges a slightly older version of the inventor. The older version has the solution to the problem and, after telling the younger version the answer, gets back into the operational time machine and jumps off to ... somewhen. The younger version then completes his machine at time $t = B$, gets into it, jumps back to time $t = C$, and ....[96]

In the past, philosophers have gotten themselves all tangled-up in debates over *personal identity*, that is, which version is *the* inventor, the younger or the older? Can they both be the same person, even though the older version has a world line (*starting* at $t = C$) that is separate and distinct from the world line of the younger (that *stops* at $t = B$)? One philosopher (note 74) left physics behind and pursued this question into the following *legal* question concerning our two (?) inventors: if the older version commits a crime and then vanishes in his time machine before the police can apprehend him, can the younger version be punished even though *he* hasn't yet committed the crime? While certainly 'interesting,' this really is a non-issue for the modern physicist who is concerned only with the physical possibility (or not) of time travel to the past.

---

[94] M. Cook, "Tips for Time Travel," in *Philosophers Look at Science Fiction*, Nelson-Hall 1982. One modern story that gets Cook's point right is by I. Watson, "The Very Slow Time Machine," in *The Best Science Fiction of the Year* (T. Carr, editor), Ballantine 1979.

[95] M. Jameson, "Murder in the Time World," *Amazing Stories*, August 1940.

[96] This little story I've just told you involves what is called a *bootstrap paradox* (just where did that solution come from, that is, who thought it up?) and it is one of the real puzzles of time travel. I'll say *lots* more about such curious doings later in the book.

Well, okay, you might now say, if neither a Wellsian time machine or a 'jumping' time machine will do, then just what *are* physicists studying in their papers on time travel? The short answer here (in Chap. 3 I'll say more) is that physicists don't view time machines as super-tech gadgets covered with wires, meters, dials, and levers, humming away beneath a seated Time Traveller as gigawatts of power throb through massive copper/crystal rods, with the whole business surrounded by a pulsating red-blue glow. Hollywood absolutely loves that sort of thing, but it's simply all wrong. For modern physicists, a time machine is *a region of spacetime with special topological structure*. Then, to time travel, a Time Traveller moves through that region of spacetime (in a rocket, perhaps) along an appropriate path. To 'make a time machine' therefore, in modern terms, means to (somehow) manipulate finite amounts of matter/energy in such a way as to alter the topology of a finite region of spacetime from one that has no CTLs/CTCs to one that does.[97] The most famous example of such a spacetime topology alteration (or *warp*) is the creation of a *wormhole*. A wormhole is a *topological* artifact of a spacetime; wormholes were popularized in Carl Sagan's 1985 novel *Contact* (under the guidance of physicist Kip Thorne) and are now common in science fiction.[98] As mentioned at the start of this chapter, for example, even Stephen King uses one in his 2011 *mainstream* novel *11/22/63*.

I'll return to the 'topology of spacetime' in Chap. 3 but, just so we don't leave it here as a mysterious phrase, here's a simple illustration of a topology change. Imagine a long, flat, narrow, two-dimensional strip of paper. The strip has the following topology features of interest to us here: (1) it has a beginning (its left end) and an ending (its right end), and (2) it has two sides (the top surface) and the flip-side surface. Now, imagine that we take the right end of the strip, give it a half-twist of 180^0 through our three-dimensional space, and then finally we glue that twisted end to the left end of the strip. The half-twist and gluing (our *warp*) has changed both of the topological properties of the strip. That's because the strip now has *no end* (you can travel forever along the strip, always going 'forward' and never reaching a point where can't go forward some more), and the strip now has just *one* side. You can convince yourself that it is one-sided by coloring the strip with a crayon. During the coloring, *do not lift the crayon from the strip*. When you can color no more, you'll find that every last bit of the strip has been colored. You can't do that with the original strip without lifting the crayon and turning the strip over because the original strip was two-sided. Many readers will recognize that what we've done is make a *Möbius strip*, named after the German astronomer and mathematician August Möbius (1790–1868) who described it in 1858.

---

[97] There is a hint of this in one prescient science fiction story, in which the inventor of a time machine, when asked about how it works, replies "An electromagnetic *warping* [my emphasis] of the spacetime continuum." See N. Schachner, "When the Future Dies," *Astounding Science Fiction*, June 1939.

[98] A wormhole is featured in the 2014 film *Interstellar*, whose Executive Producer and technical advisor was Thorne. The film's special effects are relativistically correct (not the typical Hollywood 'fantasy physics'), and you can read how that was achieved in Oliver James, Eugénie von Tunzelmann, Paul Franklin and Kip S. Thorne, "Gravitational lensing by spinning black holes in astrophysics, and in the movie *Interstellar*," *Classical and Quantum Gravity*, February 2015. See also (same authors) "Visualizing *Interstellar*'s Wormhole," *American Journal of Physics*, June 2015, pp. 486–499.

Here's another astonishing property our half-twist warp has introduced. Cut the Möbius strip lengthwise with a scissors; most people believe you will then get two strips, each the length of the original strip but each only half as wide. Actually, you get one strip with a *full* 360° twist, which means the result is back to having two sides. (To see this, make a Möbius strip, cut it, and *then* apply the crayon.) And if you cut this new strip lengthwise once more, you get two separate loops, linked together. Try it and see, but be very careful. As the late science fiction writer Cyril Kornbluth (1923–1958) warned, there may be horrific potential dangers in unschooled experimentation with topology warps:

A burleycue dancer, a pip
Named Virginia, could peel in a zip;
        But she read science fiction
        And died of constriction
Attempting a Möbius strip.[99]

To end this section, I should point out that a change in the topology of a spacetime is *not* a necessary requirement for that spacetime to support time travel to the past. Gödel's rotating spacetime, for example, has a remarkably simple topology and, as you'll recall, it's literally *stuffed* with CTLs/CTCs, to the point that time travel to the past in Gödelian spacetime would be an everyday occurrence. You might think a world that presents time travel as a *fundamentally allowed* physical phenomenon, as does Gödel's spacetime, would be irresistible to science fiction writers. (So far as I know, however, no one has written a time travel story using the rotating universe idea.[100]) In Chap. 6, in fact, I'll show you just how easy it would be to time travel in Gödel's spacetime, using a rocketship as the means to move through that spacetime. Of course, *our* universe is not Gödelian, so the 'time travel to the past' question is not so easily answered for the spacetime we appear to actually inhabit.

---

[99] As you can see from this, science fiction writers have had fun with the Möbius strip. Two early examples *not* involving time travel are N. Bond, "The Geometrics of Johnny Day," *Astounding Science Fiction*, July 1941, and W. H. Upson, "A. Botts and the Möbius Strip," *The Saturday Evening Post*, December 1945. The use of the Möbius strip for time travel occurs, for example, in M. Clifton's "Star, Bright," *Galaxy Science Fiction*, July 1952.

[100] If he had lived, perhaps the well-known science fiction writer James Blish (1921–1975) would have written such a tale. In David Ketterer's biography of Blish (*Imprisoned in a Tesseract*, Kent State University Press 1987), there is this comment from a 1970 letter written by Blish: "I am especially intrigued by the spinning-universe form of time travel, especially since … nobody has touched it … But I should really stop mentioning the spinning-universe in public, or somebody will nobble onto it before I can get into it!"

1.5 Quantum Gravity, Singularities, Black Holes, and Time Travel

> "A [spacetime] singularity is where God is dividing by zero."
>
> —Anonymous

> "A theory that involves singularities and involves them unavoidably, moreover, carries within itself the seeds of its own destruction."
>
> —Peter Bergmann (1915–2002), Einstein's research assistant at the Institute for Advanced Study, Princeton

A fundamental objection to general relativity's suggestion of the possibility of time travel to the past is that, in a very deep sense, general relativity is known to be incomplete. That is, it is incompatible with quantum mechanics, which is the physics of the very, *very* small—the physics of atomic-size objects and *smaller*. We touched on this at the end of Sect. 1.2, and here we'll take a longer look at the issue of merging quantum mechanics with general relativity.

In quantum mechanics, the discrete nature of the atomic world appears in such phenomena as the photoelectric effect, in which light acts like individual particles (photons) rather than as continuous waves. Einstein's general relativity works beautifully on a cosmological scale but, like Maxwell's theory of electromagnetism, and unlike quantum mechanics, it fails when applied deep in the interior of the atom. Quantum theory, however, seems to work *everywhere*. As one physicist put it, "As far as we can tell, there is no experiment that quantum theory does not explain, at least in principle . . . Though physicists have steered quantum theory into regions far distant from the atomic realm where it was born, there is no sign that it is ever going to break down."[101]

One of the central concepts in relativity is the *world line*, which is the complete story of a particle in spacetime. A world line assigns a definite location to the particle at each instant of time. This is a classical, pre-quantum concept, however, and today physicists use the probabilistic ideas of quantum mechanics to describe the location and momentum of a particle once they get down to the atomic scale of matter. Quantum theory is a discrete theory in which the values of physical entities vary discontinuously (in 'quantum jumps'), whereas in classical theories the values of physical entities are continuous. The difference between the two types of theories is something like the difference between sand and water. Mixing the two theories—the classically smooth, continuous general relativity and the discrete quantum mechanics—to get something called *quantum gravity*, is the Holy Grail of physicists today, and nobody has more than an obscure idea of how to do it.

---

[101] N. Hebert, *Faster Than Light: Superluminal Loopholes in Physics*, New American Library 1988.

Just one of the more curious results of the fusing of quantum mechanics with general relativity may be *quantum time*.[102] That is, in quantum gravity the smallest increment of time that has physical meaning—sometimes called the *chronon*, a term first used in a non-time travel science fiction story[103]—may have a non-zero value. As we'll see later in the book, much of the controversy over the possibility of time machines hinges on what is called the *quantum gravity cut-off*. This is the end-result of destructive spacetime stresses that tend to grow toward infinity whenever a time machine spacetime topology attempts to form. This process goes under the general name of the *back reaction*, and is conceptually similar to a rubber band growing ever more taut as it is stretched, an effect that resists more stretching (and, of course, if stretched too far the rubber band breaks).

The cut-off of those stresses, at some *finite* value, is imagined to occur when the terminal phase of the growth would take place in less than the minimum possible time interval. The cut-off happens because, it is thought, nothing can actually occur in less than the minimum time. The debate is over just what that minimum duration is, and over whether the cut-off would occur before the stresses could reach *finite* values large enough to destroy the putative time machine topology. If the cut-off occurs before the back reaction stresses climb to the critical value, then the time machine survives. Otherwise, not.

To see how this 'works,' consider the two fundamental physical constants associated with classical gravity, the *gravitational constant*[104] G and the speed of light c, and the fundamental physical constant associated with quantum mechanics (*Planck's constant*) ħ. Now, if you play around with combinations of these constants it is easy to show that the following expressions have the units of length, time, and mass, called the *Planck length* (l_P), the *Planck time* (t_P), and the *Planck mass* (m_P), respectively:

$$l_P = \sqrt{\frac{\hbar G}{c^3}} \approx 1.6 \times 10^{-33}\ \text{cm},$$

$$t_P = \sqrt{\frac{\hbar G}{c^5}} \approx 5.3 \times 10^{-44}\ \text{s},$$

$$m_P = \sqrt{\frac{\hbar c}{G}} \approx 22 \times 10^{-6}\ \text{g}.$$

The extremely tiny values of l_P and t_P (the chronon), in particular, indicate (*roughly*) where it is expected that the smooth, continuous spacetime of general relativity will itself become quantized, and so will have to give way to a quantum theory of gravity.

---

[102] H. Kragh and B. Carazza, "From Time Atoms to Space-Time Quantization: the idea of discrete time, ca 1925–1936," *Studies in History and Philosophy of Science*, June 1994, pp. 437–462.

[103] S. Weinbaum, "The Ideal," *Wonder Stories*, September 1935.

[104] This is the constant in Newton's famous inverse-square law for gravity; the attractive force F between two point masses m_1 and m_2, distance r apart, is $F = G\frac{m_1 m_2}{r^2}$.

The associated value of the mass-energy density when this transition is imagined to occur is enormous; the so-called *Planck density* is the Planck mass divided by the cube of the Planck length and has the value of about 10^{94} grams/cm^3. This is where physicists expect classical and quantum gravity to part company. Can such an enormous mass-energy density actually occur?[105]

Yes, and more, in what physicists call *singularities*.

This was all still very speculative until about 50 years ago, but today the search for how to connect general relativity and quantum mechanics is serious business. That search is related to time travel studies via a fantastic sequence of discoveries in relativistic physics, made during the last 80 years, beginning in 1931 with the work of the young Indian astrophysicist Subrahmanyan Chandrasekhar (1910–1995). He combined quantum mechanics and *special* relativity to show that a non-rotating star above a certain mass (about 1.4 times the mass of the Sun) cannot evolve into a white dwarf, which had until then been thought the eventual fate of all stars. Stars more massive than 1.4 Solar masses (but not *too* massive) would, instead, become *neutron* stars. But what then happens to stars that are too massive for even *that* bizarre eventuality?

General relativity predicts that a sufficiently massive star—greater than about four times the mass of the Sun—will, when its fuel is nearly exhausted and its nuclear fires are beginning to fade, experience a truly spectacular event called *total gravitational collapse*. When its fuel-starved, weakened radiation pressure is no longer able to keep a massive, aged star inflated against the collapsing force of its own gravity, the star will suddenly implode and crush itself into what is called a *black hole*, a dramatic term coined in 1967 by the Princeton physicist John Wheeler (1911–2008) in an address before the American Association for the Advancement of Science. A black hole is an object with a gravitational field so strong that even light cannot escape—that's why it's black!—at whose center is something called a *singularity*. This is all well-known lore in the physics world.[106]

Indeed, cataclysmic views of the collapse of matter are actually quite old. In Lucretius' first-century B.C. *The Nature of the Universe*, for example, we find the following imagery on what it would be like if matter itself collapsed: "The ground will fall away from our feet, its particles dissolved amid the mingled wreckage of heaven and earth. The whole world will vanish into the abyss, and in the twinkling of an eye no remnant will be left but empty space and invisible atoms. At whatever point you allow matter to fall short, this will be the gateway to perdition." These words were actually inspired by earthquakes, not black holes and their singularities, but could a modern expert in general relativity and singularities have said it any better? But, what *is* a singularity?

---

[105] By comparison, the density of a neutron star is on the order of a 'mere' 10^{16} grams/cm^3.

[106] Perhaps not so well-known, however, is that science fiction was there long before Wheeler. In one classic tale (M. Leinster, "Sidewise in Time," *Astounding Stories*, June 1934) a scientist explains at the end, "We know that gravity warps space . . . We can calculate the mass necessary to warp space so that it will completely close in completely . . . We know, for example, that if two gigantic star masses of certain mass were to combine . . . they would simply vanish. But they would not cease to exist. They would merely cease to exist in our space and time." And then, as another character sums it up, "Like crawling into a hole and pulling the hole in after you." The explicit use of the complete term *black hole* for a region of weird spacetime also appeared in science fiction before Wheeler (P. Worth, "Typewriter from the Future," *Amazing Stories*, February 1950).

As one theoretical physicist has dramatically written, "once gravity runs out of control, spacetime smashes itself out of existence at a singularity."[107] Or to quote Hawking, "A singularity is a place where the classical concepts of space and time break down as do all the known laws of physics."[108] One particular view of a singularity is that it is a place in spacetime that has infinite density and a gravitational field that is infinitely strong. The curvature of spacetime (more on curvature in Chap. 3) at this sort of singularity, sometimes called a *crushing* singularity, is also infinite. This is the sort of singularity believed to be at the center of non-rotating black holes. Historically, however, the occurrence of infinities in physical theories has been thought the red flag signaling that the theories have simply been extended too far, and their calculated results are nonsense.

Perhaps, then, singularities occur only in unrealistic physical applications of general relativity, and so it is only *perfectly* spherical collapsing stars that can end-up (on paper) as a black hole singularity. For a while physicists tried to establish that, but they were forced to abandon the attempt when it was shown that singularities are *unavoidable* and not just the result of idealistic assumptions.[109] This result worried many, and so the concern that general relativity was failing with its prediction of black holes and their singularities continued. In the case of a crushing singularity, perhaps all that meant is that once the collapsing star had fallen into a region even smaller than an electron, general relativity is no longer valid and the singularity is simply the 'math gone wild.' Einstein, himself, held that view. In his book *The Meaning of Relativity* (based on lectures he gave at Princeton in 1921), he wrote (concerning the use of the general theory to study the origin of the universe as a "big bang," which was a crushing singularity), "For large densities of field and matter, the field equations [of general relativity] and even the field variables which enter into them will have no real significance. One may not therefore assume the validity of the equations for very high density of field and matter, and one may not conclude that the 'beginning of the expansion' must mean a singularity in the mathematical sense."

Well, what does general relativity say about the singularity at the center of a black hole? To start, the theory says that, at a distance directly proportional to the mass of the collapsed object, a so-called *event horizon* will form. The event horizon is a surface in spacetime through which anything can fall into the hole, but through which nothing, not even a photon of light, can escape outward. The singularity at the black hole's center is therefore not visible to a remote observer (the singularity is said to be "clothed," and so not "naked"[110]). For all observers beyond the event horizon, the only visible properties of the hole are its mass (via its gravitational effects), its angular momentum (its spin rate), and its electric charge, and these properties are independent of the details of the pre-collapsed object (other than the requirement that electric charge and angular momentum are conserved).

---

[107] P. C. W. Davies, *The Edge of Infinity*, Simon & Schuster 1981.

[108] S. Hawking, "Breakdown of Predictability in Gravitational Collapse," *Physical Review D*, November 15, 1976, pp. 2460–2473.

[109] S. W. Hawking and R. Penrose, "The Singularities of Gravitational Collapse and Cosmology," *Proceedings of the Royal Society A*, January 27, 1970, pp. 529–548. Ironically, one of the 'realistic' assumptions made in this paper, which appears to *force* singularities to exist, is that time travel is *impossible*!

There are actually several fundamentally different types of black holes. If the collapsed star forms a non-rotating, spherically symmetric, uncharged[111] object, then the result is called a *Schwarzchild* black hole, after the German astronomer Karl Schwarzchild (1873–1916) who found the first exact solutions to Einstein's general relativity field equations just months after Einstein published them.[112] Soon after that the Finn Gunnar Nordström and the German Heinrich Reissner independently found the solution to the field equations for the slightly more realistic non-rotating, *charged* black hole.[113] This is only slightly more realistic since it is highly unlikely a black hole wouldn't be spinning, as all observed stars are spinning and angular momentum is conserved during gravitational collapse. Another slightly more realistic solution, that of a *rotating*, uncharged black hole, was found by the New Zealand mathematician Roy Kerr in 1963, and this solution had a twist to it that at last explains why I am telling you all this—the singularity at the center of a Kerr black hole is not the *point* singularity of a non-rotating black hole but rather is a *ring* singularity. That is, there is a hole in the Kerr singularity through which matter can travel, without being destroyed, a hole that seems to act as a portal into other spacetime regions that may include past or *future* regions of spacetime. In other words, the ring singularity seems to be the entrance to a time machine.[114]

---

[110] A naked singularity, with no event horizon behind which to hide, would be particularly bothersome to physicists who don't like the idea of the breakdown of physics being on full display. What they think they'd then see would be completely unpredictable. Whether such a situation can actually exist is still open to debate, but there *are* both analytical solutions and computer simulations (incorporating realistic equations of state on the pressure response of matter as it is compressed) that seem to allow it (as in the gravitational collapse of an *infinitely* long, non-rotating cylinder that appears to result in an axial, thread-like, naked singularity).

[111] The word *charge* means either electrical *or* magnetic charge, although from a practical point *charge* probably does mean just electrical, as the theoretically possible magnetic monopole has yet to be observed and, in any case, it is thought that black holes will not have a significant electrical charge.

[112] Even *Einstein* hadn't yet solved them, and he apparently thought they were too complicated to be solved; when he saw Schwarzchild's result, he was so impressed that Einstein wrote to say "I had not expected that the exact solution to the problem could be formulated. Your analytical treatment of the problem appears to me splendid."

[113] Two years later, the University of Pittsburgh physicist Ezra Newman finally solved the field equations for the realistic, general case of a rotating *and* charged black hole.

A discussion of singularities in general relativity is especially complicated for at least two reasons. First, there is more than one type, with *crushing* being just the (perhaps) most 'obvious.' Another type has no infinite curvature associated with it, but rather is a point in spacetime beyond which the worldline of a freely falling mass cannot be extended. Such a point is called a *geodesically incomplete* singularity, and it represents either an end to space or to time (in either case, that point is on the boundary or edge of spacetime). There are other types, as well—I've mentioned the naked singularity already—and the appearance of *any* of them is distinctly unsettling (recall Bergman's opening quote) to physicists. One that may be the most unsettling of all, however, is the *thunderbolt* singularity. This singularity propagates to infinity at the speed of light! As its discoverers dramatically put it, "It is not a naked singularity because you do not see it coming until it hits you and wipes you out."[115]

The other reason for a discussion of general relativity singularities being complicated is that they simply are not like the singularities of earlier theories. For example, in electromagnetic field theory spacetime is the *given background reference*; that is, a singularity in that theory is a point in spacetime where the *electromagnetic field* is undefined. In gravitational field theory, however, it is spacetime itself that is undefined, and there is no background ‘something’ in which spacetime is embedded to serve as a reference.[116]

---

[114] You can find discussions on how this is imagined to work in two papers by R. Weingard: "General Relativity and the Conceivability of Time Travel," *Philosophy of Science*, June 1979, pp. 328–332, and "Some Philosophical Aspects of Black Holes," *Synthese*, September 1979, pp. 191–219. See also M. Calvani et al., "Time Machine and Geodesic Motion in Kerr Metric," *General Relativity and Gravitation*, February 1978, pp. 155–163. I won't pursue black hole time machines in this book, as it is not what modern physicists consider a plausible means of time travel (*How are you going to gain access to a black hole?!!!*) For how one science fiction writer *did* use the idea, however, see L. Niven, "Singularities Make Me Nervous," in *Stellar* 1 (J.-L. del Rey, editor), Ballantine 1974. Black holes *are* bizarre objects—nearly as bizarre as time travel—and it seems risky to try to understand one in terms of the other (recall Professor Challenger's observation!).

[115] S. W. Hawking and J. M. Stewart, "Naked and Thunderbolt Singularities in Black Hole Evaporation," *Nuclear Physics B*, July 1993, pp. 393–415. As bizarre as is the thunderbolt, it was anticipated in science fiction by more than half a century. In the story "The Tides of Time" by R. M. Williams (*Thrilling Wonder Stories*, April 1940), the universe is collapsing at faster than the speed of light. Human scientists learn this when fleeing aliens stop their faster-than-light space ships to warn them. One of the human characters then looks out into the night sky and, in words that sound like those of Hawking and Stewart, "There would be no warning, for the rolling tide was traveling faster than light . . . It would come faster than the flicker of an eye. No one would see it come. One instant the world you knew would be around you. The next instant, there would be nothing. You would not even have time to know what had happened. Death, faster than the lightning flash!" This story may have been inspired by a tale published decades earlier, by the Canadian writer Frank Lillie Pollock (1876–1957). In his "Finis" (*The Argosy*, June 1906), written long before the concept of a super-nova, the light of a huge, distant star finally arrives to cook Earth into oblivion.

One early suggestion on how to avoid the problem of the crushing singularity of the non-rotating black hole (which is, as mentioned earlier, not a realistic model for the gravitational collapse of a rotating star) is that the collapse may stop short of the singularity. That is, the collapsing body might instead rebound. This ‘bounce’ would occur after the star was inside its event horizon, so an external observer would not see the later expansion, an expansion imagined to be through the event horizon *but into a different region of spacetime*.[117] When Novikov’s work was generalized the following year, the authors clearly had a hard time believing this dramatic imagery, despite their own mathematics, concluding with “It then appears necessary to believe in the existence of other [regions of the universe, including the past and the future] which will accommodate the re-expansion. This seems at least as fantastic as the alternative of [a point singularity].”.[118]

In 1974 Hawking announced an astonishing partial connection of quantum mechanics with general relativity’s black holes. He showed that, contrary to the usual image of black holes as being one-way trap doors to . . .?, black holes actually *must* radiate energy.[119] His analysis, which stunned physicists by its beautifully simple arguments, invokes the famous *uncertainty principle*, one of the corner stones of quantum mechanics. Hawking himself found the result “greatly surprising.” He also cautioned (in his 1975 paper) that the following picturesque imagery is “heuristic only and should not be taken too literally,” but it has now been in physics for over 40 years and appears to be here to stay.

The uncertainty principle states that there are certain pairs of variables associated with particles, variables that cannot be precisely measured at the same time. Time and energy form such a pair because a non-zero time interval is required to measure a particle’s energy, and the product of the uncertainty in both the time interval (Δt) and the energy (ΔE) must be at least as large as a certain non-zero constant. That is, if $\hbar$ is Planck’s constant, then $\Delta E \Delta t \sim \hbar$. This allows the process of virtual particle creation, the appearance of particle/anti-particle pairs just outside the event horizon of a black hole. The uncertainty in the energy is what gives the combined mass of the particles in a pair; this uncertainty in the energy is the quantum fluctuation energy of the intense gravity field of the hole. The only constraint is that the energy be returned to the field, via mutual annihilation of the matter/anti-matter pair within the time uncertainty dictated by the uncertainty principle.[120]

---

[116] See, for example, R. Geroch, “What Is a Singularity in General Relativity?” *Annals of Physics*, July 1968, pp. 526–540.

[117] See, for example, I. Novikov, “Change of Relativistic Collapse Into Anticollapse and Kinematics of a Charged Sphere,” *JETP Letters*, March 1, 1966, pp. 142–144, and V. P. Frolov, *et al.*, “Through a Black Hole Into a New Universe?” *Physics Letters*, January 12, 1989, pp. 272–276. Igor Novikov is a Russian physicist at the University of Copenhagen, and he will appear later in the book when we get to the paradoxes of time travel to the past.

[118] V. De La Cruze and W. Israel, “Gravitational Bounce,” *Nuovo Cimento A*, October 1, 1967, pp. 744–760.

[119] S. Hawking, “Black Hole Explosions?” *Nature*, March 1, 1974, pp. 30–31, and “Particle Creation by Black Holes,” *Communications in Mathematical Physics*, 1975, pp. 199–220.

As Hawking showed, this time interval, although incredibly short, is still long enough for the two virtual particles to separate before annihilation, one falling into the hole and the other escaping. This would happen, for example, if the particle/anti-particle pair is an electron/positron pair, and so a negatively/positively charged black hole would tend to attract the positron/electron and repel the other particle (either way, driving the charge of the hole towards zero). (Hawking then later suggested[121] that the particle *entering* the hole could be thought of as an *emitted* particle traveling backward in time, an idea that can be traced back decades, to John Wheeler—I'll return to this idea in just a moment.) By this incredible quantum process, then, the black holes of general relativity slowly *evaporate* (!) as they glow with what is now called *Hawking radiation*. That is, black holes appear to be hot bodies. But *hot* is relative, as a black hole with the mass of the Sun would have a temperature of just sixty *nano*-degrees Kelvin above absolute zero, and it would take 10^{66} years (a stupendously enormous time compared to the age of the universe) to completely evaporate.

Indeed, one physicist had already mused that the *entire universe* might have been created by a similar quantum process, *out of nothing*, a so-called *vacuum fluctuation*.[122] The explanation for why the universe doesn't then disappear—and very quickly, too, because the energy for all the mass in the newly created universe is quite large (that is, ΔE is *really big* and so Δt must be *really small*)—is that the *negative* gravitational potential energy of all that newly created matter would cancel the positive mass-energy, and so ΔE is actually *quite small* and so Δt is then *quite large*. To perhaps show he wasn't quite convinced by all that, himself, Tryon whimsically wrote "I offer the modest proposal that our Universe is simply one of those things which happen from time to time."

As a final comment on the suggestion by Hawking of a connection between virtual particles at a black hole event horizon and backwards time travel, the idea originated (as I said before) with Wheeler, in 1941.[123] In an astonishing coincidence, even as Wheeler was telling his student Richard Feynman about this, a science fiction writer was also identifying anti-matter with backward time traveling 'normal' matter.[124] Later, the Polish science fiction writer Stanislaw Lem (1921–2006) took this idea, combined it with the quantum concept of energy fluctuation, and came up with one of his typically outrageous (and typically hilarious) ideas: shooting a single positron out of an accelerator back to the very beginning of time. His story character called this fantastic machine the "Chronocannon" and claimed that's what started the universe.[125]

---

[120] The uncertainty principle has long been used in time travel science fiction. In one story, for example, a character is transported from 1950 to 2634 by a scientist of the future. Once there, this character decides he'd like to remain permanently in the 27th century. He is told that he can't because he is like an atom excited into an elevated energy state and, just as quantum mechanics says that eventually an electron in such a state will drop back down into a lower energy state, so do the "laws of time travel" require that he drop back to his normal time. How long can he remain in future, he is told, "depends on the mass [energy] of his body and the number of years the mass [energy] is displaced." That is simply the uncertainty principle. See W. Bade, "Ambition," *Galaxy Science Fiction*, October 1951.

[121] S. Hawking, "The Quantum Mechanics of Black Holes," *Scientific American*, January 1977.

[122] E. P. Tryon, "Is the Universe a Vacuum Fluctuation?" *Nature*, December 14, 1973, pp. 396–397.

Soon after Lem, a philosopher used a variant of this idea, in which the Big Bang creation of the universe was caused by a time traveler from the future who saw a need—his own existence—to generate the Big Bang. This leads to philosophical speculations on the cosmological implications of God as a time traveler.[126] Two recent physicists have taken this one step further by suggesting that the universe, via time travel, may have caused itself! As they put it, "the laws of physics may allow the Universe to be its own mother."[127]

The modern hope is that quantum mechanics (as in quantum gravity) will save physics from the horror of general relativity's singularities. This was the view of John Wheeler and, as the man who named black holes, his view is important to consider. General relativity is a classical, *smooth* theory that is fundamentally continuous, while 'our' universe appears to be a quantum one. So, perhaps, general relativity's prediction of singularities may be just an artifact without physical reality in the 'real world.' Wheeler's position was based on the quantum fluctuations of gravity fields, which are related to the uncertainties inherent in our knowledge of the values of physical entities. Such fluctuations are vanishingly small in systems of everyday size, but they increase dramatically at very tiny distances that are twenty orders of magnitude smaller than the nucleus of an atom. In the microscopic region of spacetime that the matter forming a black hole is falling into, these fluctuations might conceivably result in effects that preclude the formation of a singularity. Agreeing were two physicists who asserted that, even without a detailed knowledge of quantum gravity, quantum effects "would smash the idealized interior geometry"

---

[123] See Richard Feynman's Nobel lecture, reproduced in *Science*, August 12, 1966, pp. 699–708, where he recounts Wheeler's 'proof' for why every electron in the universe has *exactly* the same charge ('there is only *one* electron, weaving its way back-and-forth in time, with positrons being the electron when traveling backward-in-time').

[124] Will Stewart, "Minus Sign," *Astounding Science Fiction*, November 1942. 'Will Stewart' was a pen-name for John Stewart Williamson (1908–2006).

[125] S. Lem, "The Eighteenth Voyage," in *The Star Diaries*, Seabury Press 1976.

[126] G. Fulmer, "Cosmological Implications of Time Travel," in *The Intersection of Science Fiction and Philosophy* (R. E. Meyers, editor), Greenwood Press 1983. Isaac Asimov used a similar idea in his story "The Instability," *The London Observer*, January 1, 1989.

[127] J. R. Gott and L.-X. Li, "Can the Universe Create Itself?" *Physical Review D*, 1998, 023,501.

[that is, the ring singularity] of a rotating, charged black hole, thereby eliminating any possibility of using such a hole for time travel.[128]

And finally, to generalize beyond black holes to the hoped-for pay-off of the coming of quantum gravity in banishing singularities altogether, one recent study has examined how non-crushing singularities (that is, ones of the geodesically incomplete type) *are* apparently "healed" (the authors' term) by quantum effects.[129] With the eventual development of quantum gravity, perhaps all the singularities of general relativity will vanish while leaving the CTCs/CTLs intact, thereby removing a form of doubt in the theory's apparent support for time travel to the past. It may be a long time coming, however: as the University of Sydney philosopher of science Dean Rickles recently (2014) wrote in his book *A Brief History of String Theory*, "quantum gravity is in many ways . . . a revolution still waiting to happen."

1.6 Tipler's Time Machine

> "In short, general relativity suggests that if we construct a sufficiently large rotating cylinder, we create a time machine."[130]

The time traveling property of the ring singularity in a rotating black hole once made it a favorite of science fiction writers, as in Joe Haldeman's classic 1974 novel *The Forever War* (in which the term used is not *black hole*, but *collapsar*, which is a nicely descriptive word in its own right). A major difficulty with this approach, however, as I mentioned in the previous section (note 114), is that of 'getting one's hands on' (so to speak) a black hole! So, is there any other 'time machine' that is consistent with general relativity? Yes, there is.

In 1974 a young physics graduate student at the University of Maryland, Frank Tipler, caused a bit of a stir when he published what seemed to be quite specific construction details for a time machine. Indeed, the final sentence (the above quotation) of his paper couldn't be clearer. Nobody had ever before made such a statement in a respectable physics journal and, best of all, there were no apparent spacetime singularities involved. However, a close look at Tipler's analysis does turn up some difficulties.

What Tipler had actually done was to show that if one had an *infinitely long*, *very dense* cylinder rotating with a surface speed of at least half the speed of light (the rotation speed is such that the outward centrifugal forces are balanced by the inward gravitational attraction of the cylinder), then this allowed the formation of closed timelike curves around the cylinder. This means that by orbiting the surface of such a fantastic cylinder, one could travel through time into the past—but not to earlier than the moment of the creation of the cylinder.

---

[128] N. D. Birrell and P. C. W. Davies, "On Falling Through a Black Hole Into Another Universe," *Nature*, March 2, 1978, pp. 35–37.

[129] T. M. Helliwell and D. A. Konkowski, "Quantum Singularities in Spherically Symmetric, Conformally Static Spacetimes," *Physical Review D*, May 13, 2013, 10404.

[130] F. J. Tipler, "Rotating Cylinders and the Possibility of Global Causality Violation,' *Physical Review D*, April 15, 1974, pp. 2203–2206.

This last point is a very important one, as it does avoid one particularly odd paradox (called a *bootstrap*): a traveler going backwards in time to tell the inventor of a time machine (perhaps an earlier version of the time traveler himself) how to build the time machine. You can find this idea in early science fiction,[131] and a minor variant of it was amusingly illustrated in the 1985 film *Star Trek IV: The Voyage Home* (when you next watch the movie, ask yourself who actually invented "transparent aluminum"?) Bootstrap paradoxes are quite mysterious and still befuddle physicists and philosophers. Science fiction writers, on the other hand, love bootstraps as great story gimmicks.

Tipler's cylinder would also enable a time traveler to return to her original time, to go "back to the future," by orbiting the cylinder in the reverse direction (but no further into the future than when the cylinder ceases to exist). Later in the book I'll show you a simple illustration—based on a similar one in Tipler's PhD dissertation, published in 1976—of how the cylinder works as a time machine. No one, in fact, disputes any of this. It *is* true. On paper.

But Tipler did *not* prove that a time traveling property holds for cylinders of even very long but *finite* length, which are the only kind we could actually build from a finite amount of matter; he merely suggested that such might be the case. This suggestion does seem reasonable, because if the time traveler orbits at the midpoint of the cylinder, near the surface, then the gravitational end-effects of sufficiently remote ends of the cylinder would, you'd think, become negligible. Similar mathematical approximations are routinely made, for example, when calculating the electrical effects of charged cylinders of finite length. But as one physicist has warned, "Extrapolation from cylindrical symmetry to reality is very dangerous, since spacetime is not even asymptotically flat around an infinite cylinder."[132] The issue of whether a spinning, finite-length cylinder can create closed, timelike curves is still open: to quote another physicist, "[In] some respects an infinite cylinder may be a model for a long finite one, and the possibility cannot be dismissed that a time machine might be associated with a long, but finite rotating system."[133]

---

[131] See, for example, C. Cloukey, "Paradox," *Amazing Stories Quarterly*, Summer 1929. Later in the book I'll discuss even earlier literary occurrences of bootstraps (that is, of *information* on closed loops in time).

[132] K. S. Thorne, "Nonspherical Gravitational Collapse: Does It Produce Black Holes?" *Comments on Astrophysics and Space Physics*, September–October 1970, pp. 191–196. What "asymptotically flat" means will be discussed in Chap. 3.

[133] W. B. Bonner, "The Rigidly Rotating Relativistic Dust Cylinder," *Journal of Physics A*, June 1980, pp. 2121–2132. Tipler was not the first to study rotating cylinders in the context of general relativity. Such cylinders had been around for decades, going back to 1932. A good reference is M. A. Mashkour, "An Exterior Solution of the Einstein Field Equations for a Rotating Infinite Cylinder," *International Journal of Theoretical Physics*, October 1976, pp. 717–721. The first-analyzed configuration of matter that generates closed timelike lines, solved in all its general relativistic detail, was the infinite rotating cylinder studied by W. J. van Stockum,

There is, however, another potential problem besides the length of the cylinder. There is a strong likelihood that a Tipler cylinder under construction would collapse under its own internal gravitational pressure before it could be made nearly long enough to be even ‘approximately infinite.’ That is, such a finite-length cylinder might crush itself along its long axis into a pancake-shaped blob, something like what happens to a long cylinder of jello stood on-end. An ordinary can of jellied cranberry sauce will also sometimes display this curious behavior.

The required rotational speed raises yet another concern, as well. We are not talking about cylinders the diameter of a pencil, or even of a large water pipe. Recall that for a given surface speed, the larger the diameter the less the centrifugal acceleration at the surface. It is easy to calculate that even a huge cylinder 10 kilometers in radius, with a surface speed of half the speed of light, would have a surface acceleration *hundreds of billions* of times the acceleration of Earth’s surface gravity. No known form of ordinary matter could spin that fast and not explosively disintegrate; Tipler has estimated that the required density for a time machine cylinder would be 40 to 80 orders of magnitude above that of nuclear matter. (In a masterful understatement, Tipler calls this astonishing stuff “unknown material.”) Made from such incredibly superdense stuff, even a finite cylinder would still be as massive as the Sun but many trillions of times smaller. Showing no lack of imagination, Tipler has suggested the possibility of speeding up the rotation of an existing star as an alternative approach to that of building a cylinder.[134] That, of course, would be project for a far-future society, with a *very* advanced technology.

All of these concerns were discouraging to Tipler (who could blame him?), and his pessimism about the actual likelihood of achieving time travel via one of his cylinders is shown by the words he used to open his 1977 paper (note 134): “Any attempt to evolve a time machine] from [normal] matter will cause singularities to form in spacetime. Thus, if by the word ‘manufacture’ we mean ‘construct using only ordinary materials *everywhere*,’ then the theorems of this paper will conclusively demonstrate that a [time machine] cannot be manufactured.” But not all physicists agreed.

“The Gravitational Field of a Distribution of Particles Rotating About an Axis of Symmetry,” *Proceedings of the Royal Society of Edinburgh*, 1939, pp. 135–154. This is particularly interesting because, while Van Stockum didn’t spot the presence of closed timelike lines in his solution, his cylinder is made entirely from *ordinary* matter.

[134] F. Tipler, “Singularities and Causality Violation,” *Annals of Physics*, September 1977, pp. 1–36. See also his earlier paper “Causality Violation in Asymptotically Flat Space-Times,” *Physical Review Letters*, October 1976, pp. 879–882, where he wrote “There are many solutions to the Einstein equations [of general relativity] which possess causal anomalies in the form of closed timelike lines (CTL). It is of interest to discover if our Universe could have such lines. In particular, if the Universe does not at present contain such lines, is it possible for human beings to manipulate matter so as to create them? [That is, to construct a time machine.] I shall show in this paper that it is *not* [Tipler’s emphasis] possible to manufacture a CTL-containing region without the formation of naked singularities, *provided normal matter is used in the construction attempt* [my emphasis].”

Years after Tipler wrote, one physicist replied[135] with *two* pointed observations. First, Tipler's theorems apply only to singularities of the incomplete kind, not to the more convincingly fatal crushing (or curvature) type. Second, to quote Ori at length, "The standard interpretation of Tipler's theorems is to say that the appearance of a singularity in a given [spacetime] model indicates that this model is unrealistic and cannot be physically realized. Even for future-generation engineers it will probably be impossible to use 'singular matter' for the construction of their time machine. However, the theory of black holes provides an obvious counterexample to this interpretation. For, by applying this interpretation to the black hole singularity theorems one could conclude that black holes can never form." Yet black holes with several times the mass of the Sun *have* been detected in orbit about certain stars, and at least one supermassive black hole (with a mass equal to more than three *billion* Suns) has been detected at the core of galaxy M87. Indeed, it is now believed that the center of *every* sufficiently massive galaxy in the universe is home to a black hole (the one at the center of our own galaxy, the Milky Way, has a mass about three million times that of the Sun).

Even less concerned about singularities interfering with time travel were two other physicists who wrote[136] "It would seem that a successful attempt to manufacture [a time machine] within a finite region of space will be accompanied by the creation of a singularity . . . This does not immediately imply, however, that with a sufficiently advanced technology one could not make a time machine. *There is no reason to suspect spacetime singularities could not in principle be created through deliberate human action* [my emphasis]."

These optimistic views were, of course, welcome news for science fiction writers, who had been using Tipler cylinders almost from when Tipler first wrote of them. Indeed, Larry Niven liked them well enough to 'lift' the very title of Tipler's paper (note 130) for the title of a short time travel story for inclusion in his 1979 collection *Convergent Series*. Just one year after Tipler's paper appeared, Poul Anderson featured the cylinders in his 1978 novel *The Avatar*, where they are called "T-machines": one can imagine the "T" stands for Time or Tipler or even both. Anderson's story describes the cylinders as having been scattered about the universe by ancient, altruistic aliens called "the Others," for the use by any who come across them and who have the wits to decipher *how* to use them. Anderson recognized the obvious problems with Tipler cylinder construction, and so has one of his characters say of T-machines, "I have no doubt whatsoever that [they are] the product of a technology further advanced from ours than ours is from the Stone Age."

---

[135] A. Ori, "Must Time-Machine Construction Violate the Weak Energy Condition?" *Physical Review Letters*, October 1993, pp. 2517–2520. The weak energy condition is the seemingly 'obvious' requirement that the observed local mass-energy density should never be negative. Quantum mechanics predicts (and it has been experimentally confirmed) that there *are* exceptions.

[136] M. P. Headrick and J. R. Gott, "(2+1)-Dimensional Spacetimes Containing Closed Timelike Curves," *Physical Review D*, December 15, 1994, pp. 7244–7259. The '(2+1)' refers to a toy spacetime with just two spatial dimensions and one time dimension.

Even before (actually *long* before) Tipler's paper, science fiction had foreshadowed his physics. Oliver Saari (1918–2000), for example, had incorporated both superdense matter and the rule of 'no time travel before the creation of a time machine' in a story written *40 years* earlier.[137] Saari's fictional time machine works by warping spacetime via a plate of superdense matter. (An even earlier tale[138] had also used superdense matter, but it was badly flawed by its hocus-pocus invoking of 'rays' emitted by the newly discovered element of *tempium*.) The 'no time travel before the creation of a time machine' rule is the basis for an obvious response to Hawking's Chronology Protection Conjecture, discussed earlier in this chapter, and it was so used by one physicist to rebut the Conjecture: as he wrote,[139]

(1) time machines, if possible, must have the property of not being able to travel back to before their creation, and
(2) no time machine has yet been created.

The absence of time travelers amongst us, therefore, provides no insight, one way or the other, on the eventually possibility of constructing a time machine.

1.7 For Further Discussion

Observations of the background microwave radiation that permeates the universe is strong experimental evidence for the Big Bang, the singularity thought to be the origin of the universe. This singularity is not shielded from us by an event horizon, and so is not a naked singularity (note 110), which means it is potentially visible. In 1969 the English theoretician Roger Penrose, however, proposed a metaphysical 'law' called the *cosmic censorship principle*, which asserts that naked singularities are impossible. Discuss the obvious tension between Penrose's principle and the Big Bang singularity. (See, for example, P. Kosso, "Spacetime Horizons and Unobservability," *Studies in History and Philosophy of Science*, June 1988, pp. 161–173.)

---

[137] O. Saari, "The Time Bender," *Amazing Stories*, August 1937. In this story we read that the time traveler "could not travel into the past for the plate had to exist in all ages traveled, and it had not existed before he made it."

[138] E. L. Rementer, "The Time Deflector," *Amazing Stories*, December 1929.

[139] K. S. Thorne, "Do the Laws of Physics Permit Closed Timelike Curves?" *Annals of the New York Academy of Science*, August 10, 1991, pp. 182–193. Science fiction writer Damon Knight (1922–2001) anticipated Thorne's rebuttal in his story "Azimuth 1, 2, 3, … ," *Isaac Asimov's Science Fiction Magazine*, June 1982.

In the text I mention the "transparent aluminum" bootstrap paradox that appears in the 1985 movie *Star Trek IV*. Even earlier, a movie bootstrap appeared in 1980 film *The Final Countdown*. There, the designer of a modern naval warship that temporarily time travels back to the Pearl Harbor of December 6, 1941, turns out to be a crew member who was (is) accidently left behind in the past when the ship returns to the present. In the past he *will be* able to design the ship because he already knows how it *was* designed—by himself! In the more recent 2014 film *Interstellar*, a wormhole near Saturn is discovered. By the end of the film we learn that it was put there by future humans, humans who exist because their ancestors (us!) were saved from a planet-wide ecological disaster when they used the wormhole to discover new worlds in far-flung regions of the universe. Decide whether or not the existence of the wormhole represents a bootstrap paradox, and defend your position.

One difficulty in using a black hole as a means of traveling from one region of the universe to another (with time travel as a special case) is simply getting to a black hole in the first place. The nearest one to Earth, as far as is known, is many light years distant. One reason for this may be an anthropic one (see note 13 in the "Introduction"). That is, a planet near a rotating black hole would either be eventually swallowed whole, or have its surface blasted by a firestorm of radiation produced by in-falling matter. In any case, no intelligent life able to recognize time travel would ever evolve on such a planet in the first place. That is, we are here to wonder about the absence of near-by black holes precisely *because* we aren't near a black hole. The lack of black holes near Earth is addressed in Joe Haldeman's 'Earth vs. Aliens' novel, *The Forever War*, by using the time dilation effect of special relativity (discussed in Chap. 3) that allows long travel distances to be covered in a reasonable time (as measured by clocks in rocket ships traveling near the speed of light). Still, while the travel time to reach a black hole distant from Earth by many light years may only be 6 months of ship time, back on Earth many *years* may pass. Once at the black hole the ship enters it and instantly 'jumps' to a vastly different region of the universe. In the novel, no time travel after the time dilation experienced in just getting to the black hole occurs, but Haldeman uses that to great effect as follows. Before entering into combat, Earth's soldiers are told that when they exit the hole into a new region of the universe, they may encounter alien warships equipped with *their* latest technology, technology that could be far in advance of the Earth warship's technology

(continued)

which dates from Earth's *past*. That is, humans will be fighting against technology that dates from the Earth warship's *future*. To quote the novel, "Relativity traps us in the enemy's past; relativity brings them from our future." Explain this.

In one of the quatrains of the *Rubaiyat*, the eleventh century Persian poet-philosopher Omar Khayyam wrote

The Moving Finger writes; and having writ,
Moves on; nor all your Piety nor Wit
Shall lure it back to cancel half a Line,
Nor all your Tears wash out a Word of it.

Nearly a 1000 years later the German theoretical physicist Hermann Weyl (1885–1955), a colleague of both Einstein and Gödel at the Institute for Advanced Study in Princeton, NJ, wrote the following in his book *Space-Time-Matter* (published in 1921, three decades *before* Gödel's 1949 time travel paper):

> *It is possible to experience events now that will in part be an effect of my possible future resolves and actions. Moreover, it is not impossible for a world-line (in particular, that of my body), although it has a time-like* [see the index] *direction at every point, to return to the neighborhood of a point which it has already once passed through. The result would be a spectral image of the world more fearful than anything the weird fantasy of E. T. A. Hoffmann* [an early nineteenth-century German writer of the eccentric] *has ever conjured up. In actual fact the very considerable fluctuations of the* [components of the metric tensor, to be discussed in Chap. 3] *that would be necessary to produce this effect do not occur in the region of the world in which we live . . . Although paradoxes of this kind appear, nowhere do we find any real contradictions to the facts directly presented to us in experience.*
> Compare these two views and, in particular, discuss what each says about the idea of 'reliving the past.'

In the opening section of this book ("Some First Words") I mention how now and then science fiction has anticipated physics. One interesting example of this occurs in a story of a time traveler *almost* meeting himself, a story published 2 years *before* Gödel's 1949 time travel paper in which he suggests just such a possibility. The story opens with a man on a ship spotting the signal fire of a castaway on a Pacific island, as well as the tiny, distant figure

(continued)

of a man waving and jumping about. While sailing in to help, the ship hits a mine left over from the war, and the would-be-rescuer becomes a castaway, too. After swimming to the island, he can find no trace of who built the fire, although there are footprints all about in the sand. Exploring the island, he finds the remains of a crashed interstellar spaceship (!), powered by a drive unit based on 'temporal precession.' The man, curious, turns the drive on and thus sends himself backward in time by one day. He then spots a ship on the horizon, builds a fire, waves and jumps about, then *recognizes the ship as his own* .... And so the loop nearly but not quite closes. The man, apparently, rushes off into the jungle, terror-stricken at the thought of meeting himself. (You can find this tale by A. B. Chandler (1912–1984), "Castaway," in the November 1947 issue of *Weird Tales*, a publisher more of fantastic, supernatural, and horror stories than of science fiction. Perhaps easier to locate would be an anthology in which it has been reprinted: *Science-Fiction Adventures in Dimension* (G. Conklin, editor), Vanguard 1953.)

Speculate on what happens to the man. In particular, does Wells' own criticism of *The Time Machine*, concerning "vain repetitions" of time travelers, apply here (see the "Introduction" again)?

A philosopher has argued against the force of Hawking's chronology protection conjecture as follows: "There is an old argument to the effect that while backward time travel may be *possible*, it will never actually occur—for if it were going to occur, we would *already* have encountered the time travellers involved, whereas in fact we have done no such thing. ... But consider an isolated society living in a remote part of the world. Some members of this society are engaged in a long-running debate concerning the possibility of human flight. Were a 747 to pass overhead, would the debaters necessarily recognize it as containing flying humans? The answer to their question might have been staring them in the face for years, without them realizing." (See Nicholas J. J. Smith, "Bananas Enough for Time Travel?" *British Journal for the Philosophy of Science*, September 1997, pp. 363–389, in particular note 3 on p. 364. The perhaps curious appearance of 'bananas' in the title of this paper will become clear when, in Chap. 4, we delve into the details of the famous grandfather paradox.) How would *you* answer Smith's question? Do you think it is plausible, as Smith implies, that we could right now be observing (without realizing it) effects in the present-day world that are the result of time travelers amongst us? What sort of effect (s) might raise this suspicion in *your* mind?

As discussed in Sect. 1.6, Tipler expressed some pessimism in his 1977 paper (note 134) about the possibility of actually constructing a time machine from a rotating cylinder. But that doesn't mean he didn't have some doubts, too, about theoretical 'proofs' of something being impossible. In his 1976 PhD dissertation, for example, he included an amusing reference to Simon Newcomb (a real-life mathematician that Wells' Time Traveller cites in *The Time Machine*—see note 102 in Chap. 2) who published mathematical 'proofs' that it would be impossible with known science to build a "practicable machine by which men shall fly long distances through the air." Why do you think Tipler did that? You can read more about Newcomb's 'proofs' in "Is the Airship Coming," *McClure's Magazine* (September 1901, pp. 432–435) and "The Outlook for the Flying Machine," *The Independent* (October 22, 1903, pp. 2508–2512).

In his autobiography, the Princeton physicist John Wheeler had this to say about time: "The smooth flow of time—or our smooth passage through it—is an illusion that is shattered when we . . . ask about time at the moment of the Big Bang, at a moment of gravitational collapse, at the moment of the Big Crunch. Students and others often ask what existed before the Big Bang. To say that we don't know is not to say enough. Even to say that we have no *way* of knowing is not enough. We really have to say that space and time came into existence, along with matter and energy and the laws of physics, at the moment of the Big Bang. If the universe expands to a maximum size, starts contracting, and eventually collapses to a fiery death—a fate that seems likely to me and to some other theorists . . . then time and space, too, will end in this Big Crunch. I can reach no conclusion other than this: there was no 'before' before the Big Bang, and there will be 'after' after the Big Crunch." (See J. A. Wheeler, *Geons, Black Holes, and Quantum Foam: a life in physics*, W. W. Norton 1998, pp. 349–350.) That is, when the Big Bang singularity occurred, time was *created*, and if the universe should collapse in the far future in a Big Crunch, time will be *annihilated*. This is a view of *nothingness* that transcends even that of the grave. Sharing Wheeler's dark view of the ultimate fate of reality, but instead giving the *victory t*o time (rather than its annihilation), was the Irish writer Jonathan Swift (1667–1745) in his poem *Riddles* (circa 1724): "Ever eating, never cloying/All-devouring, all destroying/Never finding full repast/ Till I eat the world at last." How do you think theologians would respond to Wheeler and Swift?

The American philosopher Roy Sorenson was cited in note 13 in the discussion of the difficulties a time travel would have in convincing skeptics that he *had really* time traveled (short of bringing a fresh dinosaur egg back and hatching it!). This question was treated in early pulp science fiction ("The Sands of Time," see note 43 in "Some First Words") as follows: The time traveler takes a sealed box of pure radium (with his name written on the inside of the lid) into the distant past, and buries it in a secure location. Upon returning to the present he unearths the box; testing of the contents will show that some of the radium has radioactively decayed to lead. Indeed, the amount of decay would be a direct measure of how far back into the past the box had been transported. This issue was later elaborated on by the English philosopher Alasdair Richmond in his paper "Time Travel, Parahistory and the Past Artefact Dilemma," *Philosophy*, July 2010, pp. 369–373. There he imagined two possible ways a time travelling Shakespearean scholar might attempt to convince skeptical colleagues that he had discovered a draft of *Hamlet* dating from the year 1589 (10 years *before* the earliest accepted date of its composition by Shakespeare). The first attempt is to simply bring that draft directly back with him in the time machine, from 1589 to the present. Then, of course, many of the inherent clues as to the draft's authenticity, such as chemical composition of the ink, the weave of the paper, and orthography (the style of writing in 1589) would be consistent with the time traveler's claim, but other clues would *not*—the *age* of the paper and of the ink, for example, would be taken as evidence fatal to the claim, as they would not be nearly 430 years old. They would appear, in fact, to be practically new! The draft would, therefore, be dismissed as simply a clever forgery. The second attempt would try to get around this problem, as follows. After locating the draft in 1589, the time traveler *doesn't* bring it back to the present, but rather stashes it away in a secret hiding place. Then, once back in the present, he takes his colleagues to the secret hiding place and, with a flourish, reveals the draft which now *is* nearly 430 years old. Much to the time traveler's frustration, however, his colleagues still reject his time travel claim, this time saying he must have simply found the draft in the 'usual' way (under the floorboards in somebody's attic, for example), and is just pretending to have found it via time travel. Can you think of a way, using the *Hamlet* draft, the time traveler might be able to convince his skeptical colleagues?

The fan I quoted in "Some First Words," who wrote to *Astounding Stories* in 1931 to express his unhappiness with the appearance of women in that magazine's stories, was quite clear about his concerns—although, given the times, he carefully avoided any direct mention of sex. A modern, highly successful *female* writer of science fiction, Anne McCaffrey (1926–2011), didn't shy away from that, however, when she wrote the following in a hilariously funny essay: "Prior to the '60s, stories with any sort of love interest were very rare. True, it was implied in many stories of the '30s and '40s that the guy married the girl whom he had rescued/encountered/discovered during the course of his adventures. But no real pulse-pounding, tender, gut-reacting scenes. The girl was still a 'thing' to be used to perpetuate the hero's magnificent chromosomes. Or perhaps to prove that the guy wasn't . . . I mean, all those men locked away on a spaceship for months/years at a time. I mean . . . and you know what I mean even if I couldn't mention it in the sf of the '30s and '40s." (See Anne McCaffrey, "Hitch Your Dragon to a Star: Romance and Glamour in Science Fiction," in *Science Fiction, Today and Tomorrow*, R. Bretnor, editor, Harper & Row 1974, pp. 278–292.) Modern time travel science fiction has shown a huge change (for the better) on this score. Discuss, for example, the emotional power of a love story between a couple separated in time, as depicted in the 1975 novel *Bid Time Return* by Richard Matheson (made into the 1980 film *Somewhere in Time*). How do you think the 1931 fan would have reacted to Matheson's story? (Indeed, if that fan was a teenager—or even a few years older—in 1931, then 44 years later he would have been, at most, in his mid-60s and might well *have* read the novel.)

In the 2014 film *Interstellar*, a space probe dives into a black hole, gets a glimpse of the hole's singularity, measures some unspecified quantum effects, and then sends the measurements back to Earth (via the fifth dimension) as a signal in the form of spasmodic Morse code twitches of the second-hand on somebody's watch. This all leads (it is hinted) to a theory of quantum gravity. If you saw the second-hand on *your* watch suddenly begin to spasmodically twitch, would you then immediately think

(a) that a Morse code message was coming to you via the fifth dimension bearing the secrets of quantum gravity?

or

(b) that your watch needs a new battery?

(continued)

or

(c) something else?

Vigorously defend your answer.

William Grey, a philosopher at The University of Queensland, pointed out numerous conceptual difficulties with the idea of time travel in his paper "Troubles with Time Travel," *Philosophy*, January 1999, pp. 55–70. That paper quickly prompted a rebuttal from the philosopher Phil Dowe (at the University of Tasmania), who replied a year later with the paper "The Case for Time Travel," *Philosophy*, July 2000, pp. 441–451. We'll eventually take up all the issues discussed in those two papers but, for now, read both papers and summarize their respective arguments. Do you feel one of the writers won the day (for you, anyway)?

Chapter 2
Philosophical Space and Time

"I do not believe that there are any longer any philosophical problems about Time; there is only the physical problem of determining the exact physical geometry of the four-dimensional continuum that we inhabit."[1]

2.1 Time: What Is It, and Is It Real?

"Time is generally thought to be one of the more mysterious ingredients of the Universe."[2]

Before going any further with time *travel*, it will be well worth the effort to take a closer look at time itself, the 'stuff' or 'thing' or ... ? that we are interested in traveling 'through' or 'around' or 'across' or ... ? Oddly enough, I'll start with religion, as philosophical theologians had identified time as something unusual long before Newton's words on time in his *Principia* that I mentioned in the Introduction, and many thousands of years before science fiction writers and their time travel stories.

We can, in fact, trace the religious interest in time back at least sixteen centuries to the Christian theologian St. Augustine and his *Confessions* (in which he famously admitted "What, then, is time? I know well enough what it is, provided that nobody asks me: but if I am asked what it is and try to explain, I am baffled."). Certainly the seventeenth century Spanish Jesuit Juan Eusebius Nieremberg caught the spirit of wonder that time holds for the devout when he wrote, in his *Of Temperance and Patience*, that "*Time* is a sacred thing; it flows from Heaven ... It is an emanation from that place, where eternity springs ... It is a *clue* cast down from Heaven to guide us ... It has some assimilation to Divinity."

Going outside Christianity, we can easily find other equally strong reactions to the mystery of time. From Plutarch's *Platonic Questions* we learn that when the question of time's nature was put to Pythagoras, he simply uttered the mystical "time is the soul of the world." The *Laws of Manu* of Hinduism, the *Torah* of

---

[1] H. Putnam, "Time and Physical Geometry," *Journal of Philosophy*, April 1967, pp. 240–247.

[2] P. Horwich, *Asymmetries in Time*, MIT Press 1987.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_2

Judaism, the *Koran* of Islam, and the revealed truths of Gautama Buddha are all full of references to time. It is, in fact, to the pagan gods of Greek mythology that we owe our 'modern' image of Chronos, or Father Time.

Not just the Greeks made time a god. In the *Bhagavad Gita* (*Song of the Lord*), the central religious-romantic epic of Hinduism that predates Christ by five centuries, one of the characters reveals his divine nature and declares his power thus: "Know that I am Time, that makes the worlds to perish, when ripe, and bring on them destruction." And in the even older Egyptian Book of the Dead, which dates back over three thousand years, the newly deceased was thought literally to become one with time itself. The merging of time and the resurrection of the body after death in the Book is shown in the line "I am Yesterday, Today and Tomorrow, and I have the power to be born a second time."

The Greek philosopher Plato (circa 400 B.C.) gave us a curious way to think of time: as a *closed loop*. While Plato did think of time as having a beginning, his conception did not have time extending off into the infinite future as does the modern, everyday view. Rather, Plato visualized time as curving back on itself—as *circular* in nature. This was, in fact, a reasonable reflection on what Plato could see everywhere in nature, with the seemingly endless repetition of the seasons, the regular ebb and surge of the tides (the old English word *tid* was a unit of time), the unvarying alternation of night and day, and the rotation of the visible planets in the sky. Whatever might be observed today, it seemed obvious to Plato, would happen again in nature. Circular time in science fiction was briefly mentioned in Chap. 1,[3] and it occurs outside that genre, too, as in James Joyce's novel *Finnegans Wake*, which opens in mid-sentence and ends with the first part of the same sentence. This view of time has a powerful, ancient visual symbol, the Worm Ouroborous, or World Snake, that eats its own tail endlessly.

Circular time, with its closed topology, was favorably presented in Stephen Hawking's famous book *A Brief History of Time*. In it he concludes that there is no need for God because in circular time there is no first event and hence no need for a First Cause. Vigorous philosophical rebuttals were quick to come, of course![4]

Turning to fiction, Ray Bradbury wrote a beautifully poetic passage about the mystery of time in "Night Meeting," one of the splendid sub-stories in his episodic 1950 masterpiece *The Martian Chronicles*. A man of A.D. 2002, who is one of the modern inhabitants of Mars, somehow meets the ghostly image of a long-dead Martian one cold August night. The conditions are just right for such a cross-time encounter. As the man thinks to himself, "There is the smell of Time in the air tonight. . . . There was a thought. What did Time smell like? Like dust and people. And if you wondered what Time sounded like it sounded like water running in a dark cave and voices crying and dirt dropping down on hollow box lids, and rain. And, going further, what did Time *look* like? Time looked like snow dropping silently into a black room or it looked like a silent film in an ancient theater, one hundred billion faces falling like those New Year balloons, down and down into nothing. That was how Time smelled and looked and sounded. And tonight . . . tonight you could almost *touch* Time."

---

[3] Another example from science fiction is the story by I. Hobana, "Night Broadcast," in which a television signal from the past is picked up by a gadget that is probing the *future*: "By going far enough into the future one comes upon what we call the past." You can find this tale in the *Penguin World Omnibus of Science Fiction*, Penguin Books 1986.

[4] See, for example, W. L. Craig, "What Place, Then, for a Creator?: Hawking on God and Creation," *British Journal for the Philosophy of Science*, December 1990, pp. 473–491, and R. Le Poidevin, "Creation in a Closed Universe Or, Have Physicists Disproved the Existence of God?," *Religious Studies*, March 1991, pp. 39–48.

Well, lovely words, yes, but they don't really tell us what time *is*. Perhaps Einstein the physicist can tell us. In the *New York Times* of December 3, 1919, we find him quoted as follows: "Till now it was believed that time and space existed by themselves, even if there was nothing [Newton's view]—no Sun, no Earth, no stars—while now we know that time and space are not the vessel for the Universe, but could not exist at all if there were no contents, namely, no Sun, no Earth, and other celestial bodies." Less than 2 years later Einstein stated this view again (*New York Times*, April 4, 1921): "Up to this time the conceptions of time and space have been such that if everything in the Universe were taken away, if there were nothing left, there would still be left to man time and space." Einstein went on to deny this view of reality, saying that, according to his general theory of relativity, time and space would *cease to exist* if the universe were empty. This has the ring of one of Einstein's favorite philosophers, Spinoza, who declared in his *Principles of Cartesian Philosophy* that "there was no Time or Duration before Creation." In a correspondence with Samuel Clarke—Newton's friend who translated Newton's *Optiks* into Latin—the German philosopher Gottfried Leibniz (who began the correspondence in 1715) expressed similar ideas: "Instants, consider'd without the things, are nothing at all . . . they consist only in the successive order of things."

The pragmatic scientist would certainly agree with Leibniz. After all, what could it even mean to talk of time unless you can measure it? And what you use to measure time is a clock—some kind of changing configuration of matter involving spinning gears, ticking pendulums, and rotating dial pointers. Mere *unchanging* matter, alone, is not sufficient to measure time because a still clock measures nothing. *Changing* matter seems to be required. Yet, not surprisingly, not everybody agrees. The counterview, the view that time has nothing to do with change, was expressed in an interesting manner by a science fiction fan in a letter to the editor of *Wonder Stories* (January 1931): "Just one thing, you have these time-traveling yarns, good stuff to read all right, but bunk, you know; because if there's no such thing as time, which there isn't, *only change* [my emphasis], how can one travel in . . . something that doesn't exist. To our planet which goes around the Sun there is simply a turning and warming of one side and then the other, i.e., years, days, minutes, etc., is something purely artificial, invented by man to tell him when to do certain things, work and stop work . . ."[5]

---

[5] This fan's idea was not new. For Plato's most famous student, Aristotle, time was *motion* (in a world in which nothing moved, argued Aristotle, there would be no time), and he expressed this

Going even beyond the ideas of Einstein, Spinoza, Leibniz, Plato, Aristotle, and our science fiction fan, at least one metaphysician felt that time would have no meaning, even in a massive and changing universe, without the additional presence of conscious, rational beings.[6] That sounds very much like an echo of the French philosopher Henri Bergson who, in 1888, somewhat mysteriously declared that time is "nothing but the ghost of space haunting the reflective consciousness." A few years before Taylor, however, a fellow philosopher had argued for exactly the opposite view, that temporal passage is independent of the existence of conscious beings.[7]

All this divergence of opinion perhaps explains why even a lightweight Hollywood movie like Mel Brooks' 1987 *Spaceballs* can get a laugh from a time joke. Even kids know that the characters, when talking about time, haven't the slightest idea of *what* they are talking about. The movie, a spoof on such classic films as *Star Wars*, *The Wizard of Oz*, and *Raiders of the Lost Ark*, quickly reaches a point of crisis. To find out what to do next, the evil Lord Helmet and his chief henchman decide on a novel approach: they will look at an instant video of their own movie! (Instant videos are available *before* the movie is finished.) Perplexed at watching on a television screen everything that he is doing as he does it (the screen correctly shows an infinite regression of television screens, each being watched by a Lord Helmet), Lord Helmet initiates the following rapid-fire exchange. (It is, of course, a clever take-off on Abbott and Costello's "Who's on First?")

> *What the hell am I looking at? When does this happen in the movie?*
> *Now! You're looking at now, sir. Everything that happens now, is happening, now.*
> *What happened to then?*
> *We're past that.*
> *When?*
> *Just now, now.*
> *Go back to then.*
> *When?*
> *Now.*
> *Now?*
> *Now.*
> *I can't.*
> *Why?*
> *We missed it.*

---

view in his famous metaphor "Time is the moving image of eternity." For Aristotle, then, time and change were inseparably intertwined. For Aristotle the world had existed for eternity, and the circularity of time was a central and powerful image; using his vivid illustration, it is equally true *in circular time* that we live both before *and* after the Trojan War.

[6] R. Taylor, "Time and Life's Meaning," *Review of Metaphysics*, June 1987, pp. 675–686.

[7] S. McCall, "Objective Time Flow," *Philosophy of Science*, September 1976, pp. 337–362.

When?
Just now. [The henchman then sets the video to rewind.]
When will then be now?
Soon.

We may laugh at this, even dismiss it as mere movie madness, but could any of us *really* do much better if, like Saint Augustine, we were backed into a corner and asked to explain time? Somehow, I think even the distinguished twentieth-century Harvard professor Hilary Putnam whose words open this chapter would find it difficult to know where to begin. He might even become as confused as the time traveler in the 1968 film *Je t'aime, Je t'aime*, whose oscillations in time, from present to past and back again, leave him so befuddled that he decides he'd rather be dead. What, then, *can* we say about time? Despite Putnam's bold words, I suspect that most people would come down on the side of Augustine.

The mystery of time was well captured by R. H. Hutton (1826–1897), the literary editor of the *Spectator*, when he wrote in his 1895 review (see note 1 in the Introduction) of Wells' *Time Machine* that "the story is based on that rather favorite speculation of modern metaphysicians which supposes *time* to be at once the most important of the conditions of organic evolution, and the most misleading of subjective illusions ... and yet Time is so purely subjective a mode of thought, that a man of searching intellect is supposed to be able to devise the means of traveling in time as well as in space, and visiting, so as to be contemporary with, any age of the world, past or future, so as to become as it were a true 'pilgrim of eternity.'"

Novelist Israel Zangwill (1864–1926) wrote a similar but much more analytical review of Wells' novel for the *Pall Mall Magazine* (see note 1 in the Introduction). Zangwill was the only Victorian reviewer to attempt a scientific analysis of time travel. Although he thought Wells' effort was a "brilliant little romance," Zangwill also thought the time machine—"much like the magic carpet of *The Arabian Nights*"—was simply "an amusing fantasy." Zangwill continued in his review with what was even then a common idea about a way one might actually be able, at least in principle, to look backward in time; one could travel far out into space by going faster than light and then watch the light from the past as it catches up to you. (Note, carefully, that Zangwill was writing in 1895, 10 years before Einstein's special relativity put a limit on possible speeds.) In this way, Zangwill wrote, one could watch "the Whole Past of the Earth still playing itself out."

Indeed, even before Zangwill, the well-known French astronomer Camille Flammarion (1842–1925) had made this dramatic idea a centerpiece of his 1887 novel *Lumen*. That book, a best-seller in Europe even before its appearance in England, describes how a man just dead (in 1864) instantly finds his spirit on the star Capella, where he is able to watch the light then arriving from the Earth of 1793. In particular, he watches the French Revolution play itself out and sees himself as a child. Flammarion may have, in fact, been inspired to write his novel by an essay written several years earlier (in 1883) by the British physicist

J. H. Poynting (1852–1914). Poynting's essay,[8] which opens with the statement that it was, in turn, inspired by an anonymous pamphlet published "30 or 40 years ago" on the same topic, specifically mentions watching historical events from Capella.

By the beginning of the twentieth century the idea of watching the past by outrunning light had drifted down into juvenile literature, as in the 1904 novel *Around a Distant Star* by Jean Delaire (the pen name for Pauline Touchemoline (1868–1950)), in which a young man builds a spaceship that can travel at two thousand times the speed of light. With it, he and a friend travel to an Earth-like planet nineteen hundred light-years distant and use a super-telescope to watch the Crucifixion (and then the resurrection) of Jesus. Early magazine science fiction also found the idea of looking backward in time with delayed light to be an irresistible one, involving romance and murder.[9] In another tale incorporating human emotions, a scientist loses his wife to a rival who kidnaps her and then escapes in a faster-than-light rocket ship headed for parts unknown. After searching for them with his own brilliant invention of the 'ampliscope' (several quantum leaps beyond the telescope), the scientist locates the couple, skipping from planet to planet light-years distant. His only pleasure, then, is to use his own faster-than-light craft to outrun the images of his lost love and watch them over and over. Eventually, however, he comes to realize the ultimate futility of it all. As the final line of this sad tale says, "It would be senseless, I knew, chasing on and on after yesterdays."[10]

The reality of time received a new twist with the additional imagery of instants of time being likened to the points on a *straight* line. In the West it was the Christian theological doctrine of *unique* historical events that gave rise to *linear* time in the minds of the common folk. The creation of the world and Adam and Eve, the adventures of Noah and the cataclysmic Flood, the Resurrection—these were all events that occurred in sequence, *once*. None would happen again and so, for Christianity, circular time just would not do.[11] In addition, it has been argued that the major spiritual content of Christianity—a significant reason for its popular support even in the face of brutally harsh Roman suppression—is that it brought the *expectation of change* into the static world of ancient times. It was, in fact, in ancient religious teachings that our modern view of linear time had its origin, a view that most people today (including the most hardened agnostic physicist) find to be as natural as Plato and Aristotle found circular time.

---

[8] J. H. Poynting, "Overtaking the Rays of Light," in Poynting's *Collected Scientific Papers*, Cambridge University Press 1920.

[9] As in, for example, G. A. England, "The Time Reflector," *The Monthly Story Magazine*, September 1905.

[10] D. D. Sharp, "Faster Than Light," *Marvel Science Stories*, February 1939. The year before saw the appearance of a story with the same idea, a story that specifically cites Flammerion: M. Weisinger, "Time On My Hands," *Thrilling Wonder Stories*, June 1938.

[11] Still, just to show how one can find support for almost any view in the same religious dogma, Ecclesiastes 1:9 would seem to be a claim *not* for linear time but rather for circular time!: "The thing that hath been, it is that which shall be; and that which is done is that which shall be done; and there is no new thing under the sun."

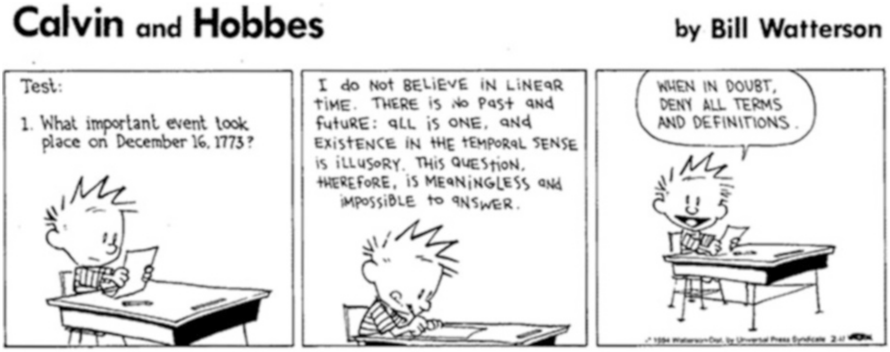
Calvin and Hobbes
by Bill Watterson
Test:
1. What important event took place on December 16, 1773?
I do not believe in linear time. There is no past and future: all is one, and existence in the temporal sense is illusory. This question, therefore, is meaningless and impossible to answer.
When in doubt, deny all terms and definitions.

CALVIN AND HOBBES ©1994 Watterson. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

Even though linear time was the norm after Christ, there were still enough questions about time to perplex the deepest of thinkers, and the next 2000 years resulted in plenty of thinking. Discourses on time by such philosophers as Descartes, Spinoza, Hobbes, Kant, Nietzsche, and Hegel can be found by the yard in any decent university library. Nearly all (if not indeed all) of these presentations have metaphysical, even theological, underpinnings. For example, Descartes is generally believed to have argued for a discontinuous, atomistic nature to time (recall the *chronon* from Chap. 1). This is the modern view of his thinking, because in his *Meditations* (1641), in particular in the third meditation on God's reality, Descartes appears to argue that God must continually recreate the world at each *separate* moment of its existence. That is, the world is recreated in a discontinuous succession of *individual* acts by God.[12]

Finally, with Newton's discussion of *absolute* time, which is the belief that time is the same everywhere in the universe, there was for the first time a *physicist* writing about time (although, as I mention in Chap. 1, Newton's views were also influenced *heavily* by theological considerations, in addition to mathematical physics). But, despite Newton's genius, the mystery of time remained a mystery.

In 1905 Einstein's name appeared among the contributors to the study of time, and so at last something besides metaphysical speculation on the subject was added to the body of human thought. Einstein's paper on special relativity introduced the revolutionary idea of *relative* time, which is the anti-Newton belief that the passage of time is *not* the same everywhere, but rather depends on local conditions. In retrospect, Einstein's 1905 work seems to be the perfect reply to the comment by Isaac Barrow (1630–1677)—Newton's teacher and the first Lucasian professor of mathematics at Cambridge (the chair once held by Stephen Hawking centuries later)—that “because *Mathematicians* frequently make use of Time, they ought to have a distinct idea of the meaning of the Word, otherwise they are Quacks.”

---

[12] For more on this, see R. T. W. Arthur, "Continuous Creation, Continuous Time: A Refutation of the Alleged Discontinuity of Cartesian Time," *Journal of the History of Philosophy*, July 1988, pp. 349–375.

Then, just 3 years after Einstein, along came a second astonishing paper by the Cambridge philosopher John Ellis McTaggart (1866–1925). This paper[13] claims to prove that whatever time might be *thought* to be (even by Einstein), it really isn’t that because time isn’t even real. (This would seem, I think you’d agree, to have potentially profound implications for time travel!) The method of the paper is to deny the reality of time via an infinite-regress argument that one philosopher[14] has called the *pons asinorum* (“bridge of asses”) of the riddle of time. As McTaggart’s own opening sentence freely admits, “It doubtless seems highly paradoxical to assert that Time is unreal, and that all statements which involve its reality are erroneous.”

McTaggart began his analysis by observing that there are two separate and distinct ways of talking about events in time. Following his terminology, one can say that events are either future, present, or past (the so-called *A-series*), or one can say that events temporally ordered by each being later than some other events, earlier than others, and simultaneous with still others (the so-called *B-series*). He then continued by asserting that time requires change, and followed that with the observation that the A-series (but not the B-series) incorporates such change. That is, if event X is earlier than event Y, then X is *always* earlier than Y and thus there is no change in this (or in any other) example of a B-series. As a specific example, let Y be the birth of a child, and let X be the birth of its mother. In contrast, if X is first in the future, then is in the present, and finally is in the past, then we have an example of change (and hence of *time*) in the A-series; for example, let X be the next time you blink.

With this rather pedestrian start, McTaggart then pulled his rabbit out of the hat. It makes no sense, he argued, to talk of the ‘future,’ ‘present,’ and ‘past’ of an event because these terms are mutually exclusive. That is, no two of these predicates can apply at once, and yet, paradoxically, every event possesses all three and thus we have a contradiction. It therefore, concludes McTaggart, makes no sense to talk of future, present, or past. And because it makes no sense to talk of them, they do not exist, and so there can be no A-series and hence no change, and thus no reality to time. McTaggart apparently realized just how befuddling all that would appear to just about everybody who read it, and so he played devil’s advocate (D.A.) in his paper by trying to anticipate the various objections people could raise. Of course, he always managed to refute the D.A. at every turn. It is worth the effort to go through the details of McTaggart’s ‘proof,’ as that will make it clear what there is about

---

[13] J. E. McTaggart, “The Unreality of Time,” *Mind*, October 1908.

[14] L. O. Mink, “Time, McTaggart and Pickwickian Language,” *Philosophical Quarterly*, July 1960, pp. 252–263. The phrase *pons asinorum* has its origin in a plane geometry theorem: the angles opposite the equal sides of an isosceles triangle are themselves equal. Seeing the truth of this is said to separate the quick-witted from the dull. It isn’t clear (to me, anyway), however, on which side of McTaggart’s ‘proof’ the quick-witted were imagined to fall. You’ll see what I mean in just a moment.

'traditional' philosophical reasoning that so irritates modern philosophers trained in mathematical physics (and what makes physicists roll their eyes when confronted with arguments like McTaggart's).

The predicates of future, present, and past are really not incompatible for any event, the D.A. says some will claim, because the real predicates we should use are '*was* future,' '*is* present,' and '*will be* past,' and these *can* be possessed all at once by any event. Nice try, counters McTaggart, but that will not solve the problem. By allowing such modified predicates, we must actually allow for all nine possibilities, some of which are still incompatible. That is, the 'was,' 'is,' and 'will be' could each be potentially attached to 'future,' 'present,' and 'past': for example, 'was past' is incompatible with 'will be future.'

Oh, counters the D.A., we can eliminate that concern by allowing even more complex predicates to arrive at a third level of structure, such as 'is going to have been past,' and 'was going to be future,' and those *are* compatible. But McTaggart swats that argument away, too, by displaying new incompatibles, as well as by showing that the process of ever-increasing predicate complexity is a vicious infinite regress that drags along the seeds of its own doom at every step.[15] There is simply no escape from incompatibility, he says, and so there is no time.

Well! What can one do when presented with such an argument, one that seems to claim philosophers can wrest free the secrets of nature by pondering the historical accidents of English syntax? As David Hume once said, "Nothing is more usual than for philosophers to encroach on the province of grammarians, and to engage in disputes of words, while they imagine they are handling controversies of the deepest importance and concern." One modern philosopher apparently agreed with Hume, at least in the case of McTaggart's 'proof,' and he was pretty blunt with his evaluation of it: "McTaggart's famous argument for the unreality of time is so completely outrageous that it should long ago have been interred in decent obscurity. And indeed it would have been, were it not for the fact that so many philosophers are not sure that it has ever really been given a proper burial, and so from time to time someone digs it up all over again in order to pronounce it *really*

---

[15] Here's a clever way to systematically generate McTaggart's infinite regress of complex predicates, as presented by M. Dummett, "A Defense of McTaggart's Proof of the Unreality of Time," *Philosophical Review*, October 1969, pp. 497–504): "Let us call 'past,' 'present,' and 'future' 'predicates of first level.' If, as McTaggart suggests, we render 'was future' as 'future in the past,' and so forth, then we have nine predicates of second level, where we join any of the three on the left with any of the three on the right:

past		past
present	in the	present
future		future

Similarly, there are twenty-seven predicates of third level … "Dummett's construction clearly shows that, at the N-th level, there are 3^N predicates, most of which are incompatible.

dead. These periodic autopsies reveal that something more remains to be said."[16] That is certainly true, in as much as McTaggart's disarmingly innocent argument has caused disagreement and furrowed brows among philosophers for decades.

It is, in fact, easy to find examples of the continuing debate over McTaggart's analysis and, as silly as it strikes physicists, it still has a pulse in some quarters. While at least one philosopher has argued that McTaggart simply didn't really understand his own proof, this philosopher nevertheless agreed with McTaggart's conclusion about the unreality of time.[17] Another writer has illustrated how McTaggart's ideas have found their way into modern philosophical debates on the meaning of time in the cinema, particularly in the analysis of *anachrony*, the telling of a story out of normal time sequence, such as occurs in time travel movies.[18]

Other sorts of metaphysical proofs for the unreality of time have been offered besides McTaggart's. For example, it has be argued that time is unreal, at least in a world empty of consciousness, because the concepts of past, present, and future could not possibly have any meaning unless events could be remembered, experienced, and anticipated. Or, for a second example, some have held time to be unreal, at least in a deterministic world (as some argue four-dimensional spacetime to be), because any event whose occurrence follows from present conditions, and from physical laws, would exist (they say) *now*. This view, which seems to assert that everything should happen at once, I personally find to be sufficiently obtuse as not to be bothered by it.[19] Debates between those who believe in the common-sense idea that present, past, and future are attributes of events (the 'tensers') and those who deny it (the four-dimensional spacetime, block universe 'detensers') continues to now and then still flair up on the pages of philosophy journals. At least one philosopher likes both views![20] Most modern physicists, I think, simply don't care about this line of inquiry.

On the other hand, less than a month before his death Einstein revealed his feelings about the meaning of present, past, and future, and his words appear to be ones that show some sympathy to the philosophers. In a letter written on March 21, 1955, to the children of his dearest friend who had just died, Einstein wrote—with full knowledge that his own illness would be his last—"And now he has preceded me briefly in bidding farewell to this strange world. This signifies nothing. For us believing physicists, the distinction between past, present, and future is only an illusion, even if a stubborn one."[21] Later in this chapter I'll return to these curious words and speculate on what Einstein may have meant by them.

---

[16] F. Christensen, "McTaggart's Paradox and the Nature of Time," *Philosophical Quarterly*, October 1974, pp. 289–299.

[17] Q. Smith, "The Infinite Regress of Temporal Attributions," *Southern Journal of Philosophy*, Fall 1986, pp. 383–396. To this came a rebuttal a year later by L. N. Oaklander, in the same journal (Fall 1987, pp. 425–431).

[18] G. Currie, "McTaggart at the Movies," *Philosophy*, July 1992, pp. 343–355.

[19] But if, upon reflection, it starts to bother *you*, see R. Gale, "Some Metaphysical Statements About Time," *Journal of Philosophy*, April 1963, pp. 225–237. We'll soon get to some of the more common philosophical questions on the nature of four-dimensional spacetime, such as 'is it deterministic or is it fatalistic?,' and 'does free-will have any meaning in four-dimensional spacetime?' Even physicists are interested such questions!

[20] R. Weingard, "Space-Time and the Direction of Time," *Nous*, may 1977, pp. 119–131.

I started this opening section on a religious note, and I'll end it on one. If you think the philosophical speculations on the nature of time that I've so far cited are 'really far out,' here's yet another one that leaves all the rest in the dust. In a paper that took real nerve to write (or, perhaps, simply a wicked sense of humor—and I write that in pure admiration) we read of how a spacetime that supports time travel can give the start for a *physics* explanation to the theological concept of Hell! After introducing just a bit of elementary spacetime physics (which I'll skip describing here because we'll do it later in the book), the author[22] shows how to 'construct' a compact region in spacetime (Hell) with the following properties:

1. While "so small even the Hubble Telescope couldn't image it" it can hold an infinity of physical beings;
2. Each of the beings in it are doomed, because of its time travel property, to an infinitely long personal future of damnation;
3. Each of the beings in it, because of its time travel property, can view all the future stages of their own personal damnation and so be "continually presented with a reminder of the impossibility of escape—a refinement no causally normal Hell can seemingly offer." In other words, and not to be too ironic about it, 'Theological Progress Through Physics!';
4. Each of the beings in it are continually being compressed together ("brought into dismaying proximity" with themselves) and so will spend eternity "listening to a cacophony" of their own cries of despair from *their personal future*.

There's more, but that's probably enough for you to get the idea. Richmond does admit that, as it stands, his time travel creation of Hell is not compatible with either quantum theory or even general relativity. Still, it *is* something to ponder, don't you think, when the subject of time travel comes up!

2.2 Linear Time and the Infinity of Past and Future

> "A thousand years is a huge succession of yesterdays beyond our clear apprehension."[23]
>
> —H. G. Wells

---

[21] Quoted from B. Hoffmann, *Albert Einstein: Creator & Rebel*, New American Library 1972, pp. 257–258.

[22] Alasdair M. Richmond, "Hilbert's Inferno: Time Travel and the Damned," *Ratio*, September 2013, pp. 233–249.

[23] This line appears in Wells' 1944 doctoral thesis, written for the University of London. You can find an abridgement of the thesis in *Nature*, April 1, 1944, pp. 395–397.

The modern concept of linear time as a straight line extending from the dim past through the present and disappearing into the misty future gives rise immediately to twin questions: "Did time have a beginning?" and "Will time ever end?" As one philosopher put it (long before physicists became seriously interested in singularities like the Big Bang) "Endings and beginnings are rooted in the very conception of time itself."[24] Starting at the beginning, we'll ask if the past has been forever? Early Biblical scholars, of course, believed the answers to both questions to be *no*.

They believed that the world came into being because of a First Cause, God's creation of everything. Those scholars expended vast quantities of energy (and, need I say it, time itself) in calculating the date of creation. Martin Luther, for example, argued for 4000 B.C. as roughly when everything, including time, began. Johannes Kepler adjusted this by a notch, to 4004 B.C., and later the Calvinist James Ussher, Archbishop of Armagh and Primate of All Ireland, tweaked it again. His date is the most impressive of all, at least in detail: the first day of the world was 4003, 70 days, and 6 h before the midnight that started the first day of the Christian era. Six days after that first day of the world, Adam was made, and as a final dash of specificity, this last date was declared to be Friday, October 28! Ironically, then, though Christian theology may be given credit for introducing linear time, it certainly did not provide very much of it. The beginning of time was just 6000 years or so ago, and of course The End—in the form of the Battle of Armageddon—has been awaited (with varying degrees of eagerness) for the last 1000 years.

The discovery in the seventeenth century of geological time cast a certain amount of skepticism on those early calculations concerning the duration of the past. With the discovery that the very Earth itself could be decoded for its history, the lure of trying to decode a mere book of admittedly finite age declined for most people although it cannot be denied that modern Creationists still find such a task to have its rewards). Geological time was discovered to a *chasm* of time extending backward for billions of years, a duration that is really incomprehensible for the human brain. It has become fashionable for geologists to refer to such enormous durations with the apt term *deep time*, a subtle play on the metaphor of the "ocean of time."

It is nothing less than humbling to historians who pause to think on how little of the past is known, that is, recorded. As the ever anonymous wit once put it, "History is a damn dim candle over a damn dark abyss." Still, even as enormous as is the age of the Earth, it is not infinite. But of course our planet is very old, and the universe is many billions of years older. Is the age of the universe also the duration of the past? Or is the past itself actually *infinite*?

An implicit assumption of the infinity of the past (and of the future, too) can be found in Book Three of Lucretius' science poem *De Rerum Natura* (*On the Nature of Things*) where, just before the birth of Christ, Lucretius argues for the irrationality of fearing death: "The bygone antiquity of everlasting time before our birth was nothing to us. Nature holds this up to us as a mirror of the time yet to come after our death. Is there anything in this that looks appalling, anything that means an aspect of gloom? Is it not more untroubled than any sleep?"

---

[24] I. Stearns, "Time and the Timeless," *Review of Metaphysics*, December 1950, pp. 187–200.

One philosopher[25] has traced the origins of rational support for the finite duration of the past to as far back as the sixth century A.D. The argument presented then by the Christian philosopher Joannes Philoponus of Alexandria (who is otherwise known as John the Grammarian) is simply that the world could *not* have been forever because that implies an infinity of successive acts could have taken place which (according to Philoponus) is impossible. A variation on this is the claim that if the past were infinite in extent, then everything would have happened by now! Infinity was just too big for the ancient mind (Zeno's hoary pre-Christian paradoxes, as is well-known today, are based on subtle errors in the use of infinity).

This view on the impossibility of an infinite past seems to have been the prevalent view; even as late as the twelfth century the debate among Christian theologians was not about the possibility of an infinite past, but instead about whether the Biblical 'six days of Creation' actually had taken place simultaneously. For many, the past was 'obviously' finite in duration.[26] Not all Christians accepted that conclusion, however, and the following century saw St. Thomas Aquinas (a follower of Aristotle) arguing for the opposite view of an infinite past.

Thomas' contemporary, St. Bonaventure, however, argued again for a *finite* past, and it is with Bonaventure that we start to see some mathematical sophistication.[27] Bonaventure argued that in a world infinitely old, the Sun would have made an infinite number of its annual trips around the ecliptic. But for each such trip the Moon would have made twelve monthly trips around the Earth, and so this second infinity would be twelve times as great as the first one, and how could that be? Infinity is infinity, and how can something be twelve times bigger than infinity? This argument doesn't have any strength today because of the nineteenth century German mathematician Georg Cantor's work on the concept of infinity,[28] but it *is* clever. Agonized, convoluted theological analyses of God, infinity, and eternity continued long after Aquinas and Bonaventure. Two examples should capture the spirit of those times.

---

[25] G. J. Whitrow, "On the Impossibility of an Infinite Past," *British Journal for the Philosophy of Science*, March 1978, pp. 39–45. Whitrow adds modern scientific support to the idea of a finite past by citing the prediction from general relativity of a singularity in spacetime at some finite past time; that is, the theory's prediction that time—and everything else—had its beginning in the now famous Big Bang.

[26] C. Gross, "Twelfth-Century Concepts of Time: Three Reinterpretations of Augustine's Doctrine of Creation *Simul*," *Journal of the History of Philosophy*, July 1985, pp. 325–338.

[27] See, for example, L. Sweeney, "Bonaventure and Aquinas on the Divine Being as Infinite," *Southwestern Journal of Philosophy*, Summer 1974, pp. 71–91, and S. Baldner, "St. Bonaventure on the Temporal Beginning of the World," *New Scholasticism*, Spring 1989, pp. 206–228.

[28] For simple high school-level presentations on Cantor's astonishing infinity results, see my book *The Logician and the Engineer*, Princeton 2013, pp. 169–171.

Consider first this one, on the supposed immortality of the soul. If $A = B$, then $2A = 2B$. Next, let A = 'half alive' and B = 'half dead,' where $A = B$ in the same sense that a glass half-full is also half-empty. Thus, to be completely dead is to be completely alive, and so the soul is immortal. Outrageous? *Yes*, in my opinion, but I do also have to admit the 'reasoning' does have a certain charm!

For my second example, let me begin by setting the historical stage. After publication of the English political philosopher Thomas Hobbes' *Leviathan* in 1651, with its arguments against the power of the Church and for civil power (with some criticism tossed in, as well, for universities), Seth Ward counterattacked. Ward, who was both a minister (later a bishop) in the Anglican Church and Savilian Professor of Astronomy at Oxford, was greatly offended by the secular nature of *Leviathan*. Even before Leviathan, in fact, Ward certainly would not have liked Hobbes' earlier denial of the existence of immaterial substances (such as souls). Ward's 1652 book *A Philosophical Essay Towards An Eviction of the Being and Attributes of God, the Immortality of the Souls of Men, the Truth and Authority of Scripture*, was the first of a two-punch reply to Hobbes. The second came in 1654 with the appearance of Ward's *Vindiciae academiarum*. In both of these works Ward attempted to undermine Hobbes' credibility by attacking his mathematical ability. (Hobbes had long been fascinated by, and was considered an expert on, the ancient problem of 'squaring the circle,' a task that has been known to be impossible only since 1882.[29]) In his *Essay*, Ward also attempted to defend the view that the world has a finite age—that is, it had a specific moment of creation, presumably by God. In an opening note, in fact, Ward cites Hobbes' rejection of immaterial substances as the motivation for his writing *Essay*.

To support his view of a finite age for the world, Ward invoked infinity in an interesting way. He argued that nothing is permanent, certainly not humans. Each is created; one can imagine tracing a chain of creation events backward in time through successive generations. Now, there are only two separate and distinct possibilities to where this chain could lead to in the past. First, it could terminate, after a finite number of generations, at a *first* generation, that is, with the 'creation' of the first human. If that is the case, then, said Ward (in effect), 'case closed.' If that is not the case, however, then the chain of successive generations never terminates, that is, the chain is infinitely long. But that, argued Ward, is nonsense—how could anything *infinitely* long have an end (our present *now*)?

Why Ward thought this an unanswerable paradox is hard to understand; after all, one can imagine a line in some coordinate system *beginning* at the origin and yet still being infinitely long (an example is the positive x-axis). This counter-example was not put forth by Hobbes in his own self-defense, but rather was offered by one of Ward's own colleagues at Oxford, John Wallis, the Savilian Professor of Geometry. As for Hobbes, he was little bothered by Ward's argument. As he pointed out (surely with a smile on his face), Ward was in danger of impaling himself as a theologian on his own sword: Ward's argument 'proved' the finite age not only of the world but of *everything*, including God (thus raising the awkward question of who, or what, made God?).

---

[29] The problem of 'squaring the circle' is, given a circle of area A, to construct (using only compass and straightedge) a square of area A.

Similar problems with infinity lay behind Kant's rejection of an infinite past. It is interesting to note that Kant, somewhat paradoxically, thought an infinite *future* a possibility. Why did Kant think time could be infinite in one direction but not in the other? One philosopher tells us[30] that Kant "failed to make himself clear," and I think that *understates* the case. I say that because Kant's argument was that the duration of the future is less problematic than is that of the past because it is only the past that influences the present. The best I can do in 'explaining' this is to speculate that if the present depends on an *infinite* past, then perhaps Kant thought that the possibility of so much influence was simply too much for the present to handle! In any case, Kant's view falls apart if we consider the possibility of backward time travel and the resulting implication that the future could also influence the present.

There is, as will come as no surprise, a philosopher for every conceivable point of the compass, and so a paper by one on the logical possibility of an infinite past soon prompts a rebuttal by another.[31] In illustration of this, you'll recall the quote from Augustus De Morgan in the opening section of this book, concerning the philosophers of his times; De Morgan went on in his critique to amusingly summarize the metaphysics of those times as follows: "Here we go up, up, up,/And there we go down, down, down,/Here we go backwards and forwards/And there we go round, round, round."

So, with De Morgan's words in mind, here are a few more examples of how people have struggled with the issue of the past. One quite interesting, *scientific* twist on the duration of the past was pointed out before the exchange between Smith and Ells. In a paper[32] observing that although general relativity and its predicted spacetime singularity in the distant past may indeed allow for a finite past, that does not completely close the door to the possibility that the Big Bang was a continuation from a previous contraction phase of the universe, and so on, *ad infinitum*. (You'll recall the discussion in Chap. 1 of this idea in science fiction: see note 53 in that chapter.) To quote T. S. Eliot (from his "Little Gidding"):

---

[30] J. Bennett, "The Age and the Size of the World," *Synthese*, August 1971, pp. 127–146. See also Q. Smith, "Kant and the Beginning of Time," *New Scholasticism*, Summer 1985, pp. 339–346.

[31] See, for example, Q. Smith, "Infinity and the Past," *Philosophy of Science*, March 1987, pp. 63–75, and then read E. Ells, "Quentin Smith on the Infinity of the Past," *Philosophy of Science*, March 1988, pp. 453–455. Smith's paper "The Uncaused Beginning of the Universe" appeared in this same issue (pp. 39–57), stating that he believed, *really*, only in the *logical* possibility of an infinite past and that the universe had in fact originated in an uncaused (no God required) Big Bang singularity. And, indeed, he *had* so argued for a finite past, in "On the Beginning of Time," *Nous*, December 1985, pp. 579–584.

[32] R. Weingard, "General Relativity and the Length of the Past," *British Journal for the Philosophy of Science*, June 1979, pp. 170–172.

"What we call the beginning is often the end
And to make an end is to make a beginning.
The end is where we start from."

Even without entertaining such an oscillating, accordion-like universe that endlessly expands and shrinks, it is possible to have a universe that originated in a *single* Big Bang a finite time ago in the past but yet *has no first instant*! This astonishing statement shocks most at first encounter, but it is simply the cosmological version of a well-known mathematical result. The instant $t = 0$ is not actually part of spacetime, because the Big Bang was quite literally a singular event for which the laws of spacetime physics fail. Thus, all instants in time are greater than zero—and there is no smallest number greater than zero. If you name a positive number, no matter how small, I can name a positive number still smaller, such as one-half of yours. (Of course, if there really is merit to the idea of a quantum of time, the chronon, this argument goes out the window.)

In an ingenious observation that seems to have been missed by most philosophers, E. A. Milne, a professor of mathematics at Oxford, suggested in his 1948 book *Kinematic Relativity*, that with general relativity it is conceivable to have both a single Big Bang a finite time ago *and* an infinite past. Pointing out that to talk meaningfully of time implies that we have a clock to measure it by, Milne looked for a Universal Clock that would be far more durable than our heartbeats, or anything else that exists only transiently. He suggested the expansion rate of the universe itself as the ideal clock. As we go back in time to the Big Bang, the expansion rate rises towards infinity and, as another analyst put it, "We see the Universe ticking away quite actively. *The Universe is meaningfully infinitely old because infinitely many things have happened since the beginning*."[33]

The debate over the length of the past in modern times can be just as contentious as it was in medieval times. For example, in his editorial ("Down with the Big Bang") of August 10, 1989, the then editor of *Nature* (John Maddox) declared the standard explosive model of the universe to be "philosophically unacceptable," because "the implication is that there was one instant at which time literally began and so, by extension, an instant before which there was no time." For Maddox, this meant that the Big Bang "is an *effect* [my emphasis] whose *cause* [my emphasis] cannot be identified or even discussed." The usual (non-time travel) use of the words *cause* and *effect* is that the cause happens first and then the effect occurs—but if the Big Bang (the effect) is the origin of time, then how (asked Maddox) could there be a cause of the Big Bang *before* that beginning?[34]

---

[33] C. W. Misner, "Absolute Zero of Time," *Physical Review*, October 1969, pp. 1328–1333. In this view cosmic time is taken as proportional to the negative of the logarithm of the normalized volume of the universe ($V = 1$ represents maximum volume, and so time 'stops' at the end of the universe's expansion). Thus, because V goes to zero as we go backward in time, time runs ever faster as we travel ever further into the past. This puts the Big Bang (with $V = 0$) infinitely long ago.

[34] This was not a new insight, of course, as Aristotle had long ago (in his *Physics*) declared an instant in time with no predecessor to be an absurdity.

The answer is obvious *for creationists*, of course—God did it. Creationists avoid the question of God's cause, however, saying only that 'He needs no cause,' or even that 'He made Himself'! It is these standard (ridiculous) responses from creationists that Maddox said had prompted his editorial against the Big Bang, because creationists *embrace* the Big Bang as it seems to endorse their position of 'science by imagination.' Whatever the truth of that, I think juxtapositioning the *scientific* Big Bang model of the universe with theological metaphysics and the pseudo-science nonsense of creationism to be terribly unfair.

When will the philosophical debates on the age of the past end? Not until the end of the (infinite?) future, is my wager!

2.3 Cause and Effect

> "There are few paradoxes which have been resolved so often as the time-asymmetry paradox."[35]

The philosophical literature is full of discussions about potential causal relationships between events. One of the most famous of these discussions, illustrating that cause and effect can be pretty slippery concepts, asks what at first appears to be an almost trivial question: Did the death of Socrates cause the widowhood of Xanthippe? The quick and easy answer is "Of *course*—she was his wife and it was his death that causes us to say she was then a widow. What could be more obvious?" One philosopher has provided some interesting commentary, however, that might make you reconsider, or to at least become aware of how different are the questions concerning time that are of interest to physicists and philosophers.[36]

Suppose we agree that there are two events to be considered; Socrates ceasing to live, and Xanthippe becoming a widow. Those events occurred at different places (in prison, and wherever Xanthippe happened to be). Then, as Kim asserted, "the two events occur with absolute simultaneity . . . [and so] we would have to accept this case as one in which causal action is propagated instantaneously through spatial space." (As we'll discuss in Chap. 3, the *relativity* of distant simultaneity weakens this assertion, but we'll take that up later.) For now, it is the conclusion that Kim draws from the assertion that interests us here: just *what* is propagating instantly? If it isn't mass-energy (as 'widowhood' would appear not to be!) then special relativity isn't bothered and physicists are happy. But those same physicists might also scratch their heads over *why* philosophers even wonder about such a question, because isn't becoming a widow just another way of saying that Socrates died and so we really don't have *two* events, but just one? In other words, for physicists this really isn't a question about cause and effect at all!

---

[35] J. Hurley, "The Time-Asymmetry Paradox," *American Journal of Physics*, January 1986, pp. 25–28.

[36] J. Kim, "Noncausal Connections," *Nous*, March 1974, pp. 41–52.

The central puzzle of time travel to the past is its apparent denial of causality—that is, its denial of the belief that we live in a world where every effect has a cause and that the cause happens first. *First* we flip the switch and *then* the kitchen light comes on. It is *never* the other way around. So deeply embedded is the temporal ordering of cause and effect in our feelings about how the world—and all the rest of the cosmos—works, that the Australian philosopher John Mackie (1917–1981) called causation the "cement of the universe" (and used that wonderful phrase as the title of a 1980 book). Without causality, said Mackie, everything would come unglued and fall apart. For example, when electrical engineers design an electronic system that they intend to actually construct (as opposed to doing a mere theoretical 'paper design') they insist that the design be a *causal* one. By that they mean the system must have no output before an input is applied. That is, the system must not be able to anticipate (foresee) the application of an input. To put it bluntly, our engineers are insisting that they are *not* building a time machine!

Now all that might seem to be self-evident, but there *are* some subtle problems. For example, it has become almost a cliché to say that nothing can go faster than light; that's what physicists mean by *relativistic causality*. In other words, no cause can produce an effect at a distant location sooner than the time lapse required for a light pulse to make the trip. Classical mechanics, however, the science of Newton's laws that engineers use all the time, is *not* relativistically causal. Push the left end of a rigid rod, for example, and the right end moves *instantly*. Most of the time the lack of this form of causality causes no problems, but the fact remains that the mechanics all engineers (and physicists, too!) learn first in school is flawed on a fundamental level. A rigid rod is an impossibility in Einstein's mechanics.

Indeed, it is interesting to speculate about how, after a discussion of causality, a traditional engineering professor would respond if challenged on this issue by a bright student. Causality might not look so obvious, after all, if such a student stuck up her hand in class and said "Professor, you've told us that everything that happens in nature is due to a cause. That what we see happening all around us, as the world unfolds, is the domino-process of cause-effect-cause-effect, and so on, into the future. But suppose, Professor, that at some instant, somehow, every particle in the world suddenly reversed its velocity vector. Wouldn't that mean, given the time-reversible nature of the classical equations of motion, the world would then run backward in time along the same path it had followed up until the instant of reversal? Wouldn't that mean what was effect is now cause, and that what was cause is now effect? And if cause and effect can change roles like that . . . well, Professor, just what do our words *mean*?"

An amusing, and instructive, cartoon illustration of the student's idea of reversing all the velocity vectors in a system appeared on the cover of the November 1953 issue of *Physics Today*. That issue contains an article on the 1949 nuclear magnetic resonance experiments performed by the American physicist Erwin Hahn, which in a certain sense dealt with just such reversed systems. In that illustration a group of runners on a multi-lane circular race track begin at the starting line in a coherent state, that is, all lined up together. Then, as they run around the track at various speeds, they gradually spread out into what appears to be an incoherent state.

But that incoherence is an illusion because if, at some instant (signaled in the cartoon by a pistol shot), they all turn around and run in reverse, they will all arrive back at the starting line *together, at the same instant*. The initial coherence of the runners was actually never lost, despite the superficial appearance of disorder, and the coherent state can be recovered at any time by a reversal of velocity vectors.

This isn't mere theoretical speculation, as an almost magical application of velocity vector reversal is actually used in what is called *optical phase conjugation*, a process to 'time-reverse' the severe distortion suffered by light beams during atmospheric propagation. For example, by effectively reversing the velocity vectors of photons, one can remove the turbulence blurring in satellite pictures of the Earth's surface as seen from space.[37]

Let me immediately short-circuit one possible answer our beleaguered professor might give in desperation, a response based on the fact that equations of physics are *not* all time reversible. Indeed, it was discovered decades ago that, in certain very rare, fundamental particle decay processes involving neutral K-mesons, there is the hint that perhaps nature *can* indeed distinguish between the past and the future. In particular, K-mesons should violate what is called *CP-symmetry*, and the so-called TCP theorem[38] says that then *T-symmetry* must also fail. In 1968/69 direct, experimental observation of the failure of T-symmetry in K-meson decays was reported. In an astonishing example of science fiction prescience, the use of K-mesons in a machine for affecting the past had appeared years earlier in a 1955 (!) story.[39]

So, could K-mesons account for the physical processes that we see evolve in time in one direction (past to future) but not in the other? As Hurley (note 35) put it so nicely, "The decay of the neutral K-meson is not time-reversal invariant; perhaps it is this ubiquitous meson which is responsible for the cream diffusing uniformly throughout our coffee in the morning. Possibly, but again this conjecture cannot account for the computer models [of diffusion processes that, like cream in coffee, also display a bias for one temporal direction over the other—in Chap. 3 I'll show you such a computer model] which have no neutral K-mesons." Still, the tiny chink that K-mesons appear to have made in the once-solid rock of time direction indistinguishability is an active area of research and speculation.

Even with that chink the fact that the classical laws appear to be insensitive to a direction of time, whereas the real world—which seems in no way dependent on the arcane properties of K-mesons—seems distinctly asymmetric, is a puzzle of the first rank. As one philosopher wrote, "The Universe seems asymmetric with respect to the past and future in a very deep and non-accidental way, and yet all the laws of nature are purely time symmetric. So where can the asymmetry come from?"[40]

---

[37] C. R. Giuliano, "Applications of Optical Phase Conjugation," *Physics Today*, April 1981, pp. 27–35.

[38] The TCP-theorem says that the 'mirror-image' of a physical process is a legitimate process, too, *if* the 'mirror' reverses time (T), electric charge (C)—so that particle and anti-particle are interchanged, and parity (P)—which is the measure of left and right. There is strong reason to believe in the validity of the TCP theorem because quantum field theory is compatible with special relativity only if the TCP theorem holds.

[39] F. Pohl, "Target One," *Galaxy Science Fiction*, April 1955.

There have of course been attempts to answer that question. For example, one philosopher[41] discusses some curious mathematical examples he interprets as meaning, in the context of classical mechanics, that there are physical systems that are temporally irreversible *in principle*. A reply[42] from a fellow philosopher, however, argues that Hutchinson has, at most, shown only that classical mechanics is perhaps not deterministic. And that, Savitt argues, is not equivalent to showing a failure of time reversibility. There is, in fact, powerful experimental evidence that, with the rare exceptions of K-mesons, the classical laws of physics (including general relativity and quantum mechanics) *are* time-reversible.

Perhaps the most compelling of such evidence comes from the *reciprocity theorem* that electrical engineers routinely use when designing radio antennas. The theorem is easy to illustrate. Suppose two electrical engineers, Bob in Boston and Lois in Los Angeles, send radio signals to each other. Bob sends his messages by exciting his antenna with a time-varying current, which thus launches electromagnetic radiation into space. Lois' distant antenna intercepts some of that radiation, which then creates a (very tiny) signal current in her antenna.

The reciprocity theorem states the following: Suppose Bob makes a tape recording of his excitation signal and mails it to Lois, who then plays Bob's tape back into her transmitter as the excitation to *her* antenna. Then the signal current induced in Bob's antenna, as it intercepts Lois' launched radiation, will be the very same (very tiny) signal that Lois measured in her antenna as a result of Bob's transmission. This result is completely independent of the details of the two antennas, which can be utterly different in design, as well as independent of the details of the propagation path between Boston and Los Angeles (as long as those details don't change with time). The reciprocity theorem *is* true—it can be *measured* to be true as accurately as one wishes to perform this experiment—because of the reversibility of physics right down to the electronic level. In fact, the answer to the professor's problem of explaining why we don't see velocity vectors suddenly reverse, and then everything 'run backwards,' has not yet been found in any law of physics.

Now, to make things even more interesting, consider the problem of *mutual* or *simultaneous* causation, which can quickly lead to several interesting questions. When two leaning dominoes, A and B, hold each other up, is A nearly upright because of B, or is it B that is nearly upright because of A? When two children bob up and down on a see-saw, whose motion is the cause and whose is the effect? There are other puzzles, too, that involve mutual causation.

---

[40] J. Earman, "The Anisotropy of Time," *Australasian Journal of Philosophy*, December 1969, pp. 273–295.

[41] K. Hutchinson, "Is Classical Mechanics Really Time-Reversible and Deterministic?" *British Journal for the Philosophy of Science*, June 1993, pp. 307–323.

[42] S. F. Savitt, "Is Classical Mechanics Time-Reversal Invariant?" *British Journal for the Philosophy of Science*, September 1994, pp. 907–913.

For example, causation is usually thought to be transitive: if A causes B, and if B causes C, then A causes C. But if A and B are mutually causative, then ‘A causes B’ coupled with ‘B causes A’ leads to ‘A causes A’ (and to ‘B causes B’). That is, mutual causation, together with transitivity, seems to imply *self*-causation! Except for those theologians who like this sort of result (it lets them answer the question ‘Who made God?’ with ‘He made Himself’), hardly anyone likes self-causation. But how do we avoid the conclusion that perhaps the mutual causation of two leaning dominoes, coupled with transitivity, represents experimental proof that God could have made himself? Well, of course this is certainly outrageous stuff, but don’t you wonder how our poor professor would respond if asked?

This last example is actually a far more esoteric one than we need to illustrate how our ordinary, everyday concept of cause and effect can be turned inside out by going only a little bit beyond the routine. Consider, for example, the problem of the data processing of recorded time signals, such as the information written onto magnetic tapes, hard drives, or disks. Typical applications that produce such recordings include the strata-probing seismic echoes from dynamite explosions set by oil exploration geologists; arms control compliance monitoring stations that listen for the acoustic rumbles generated by both earthquakes and underground nuclear tests—and then try to tell one from the other; and the gathering by various military intelligence agencies of turbine shaft/propeller noise signatures emitted by different types of submarines. In each of those situations, the raw information is recorded and then later processed with a certain degree of unhurried calm and leisure. That pool of oil, after all, has been underground for several hundred million years, and waiting a few more days or weeks for a computer analysis of the explosion echo isn’t going to make much difference.

Such after-the-fact processing of recorded data is said to be done ‘off-line, in non-real time.’ When we play a disk back in the lab, however, we can do all sorts of neat things, like speed up the playback (make time ‘run fast’), or slow it down (make time ‘run slow’), or even play it *backwards* (make time ‘run in reverse’). For various technical reasons, generically called *spectrum shifting*, such tricks are often quite useful. Now, the way we retrieve magnetically recorded information from (for example) a magnetic tape, is to run it through a playback machine with a ‘read-head’ that senses the magnetic flux variations. The electrical signal produced by the read-head is just like the original signal and, in fact, we can pretend we don’t know it is really coming off a tape, but rather that it *is* the original signal. For high-quality digitally recorded tapes and disks, in fact, it is virtually impossible to distinguish the original from a playback.

Now, suppose we construct our playback machine with *two* read-heads, with the new head sensing the recording slightly *before* the old head does. The two heads produce the same electric signal, of course, but the signal from the new head is *ahead* in time compared to the signal from the old head. The new head is, in a certain sense, ‘seeing the future’ of the old head! We can use these two signals, the old head representing ‘now’ time and the new head representing ‘future’ time, to build real systems that are *not* causal. The causality violation occurs in non-real time, of course, not *our* time, but no matter; some absolutely astonishing signal processing can be achieved this way. The universe is about fifteen billion years old, and pretending that time has shifted a few milliseconds or so doesn't seem to be too much violence to reality.

Two heads are often used on radio call-in talk shows to catch inappropriate remarks from intemperate callers and prevent them from being broadcast. A short time delay is introduced by first recording remarks 'live' on tape with a write head and, then a few seconds 'up-stream,' a read head regenerates the remarks for broadcast. A 5 s delay is generally sufficient, so what is heard on a radio receiver *now* actually occurred 5 s ago in the *past*. A caller can get terribly confused if she doesn't turn her own receiver off, because one ear hears the present on the telephone while the other ear listens to the past over the radio.[43] The 1956 British film *Timeslip* incorporates a similar situation, with an atomic scientist's perception advanced 7 s into the future as the result of an accidental radiation exposure. His resulting confusion and disorientation is the center of the film.[44]

2.4 Backward Causation

> "Causation as a topic of philosophical discussion refuses to die. Each year, books and articles on causation continue to pour forth. Of course, all this activity may simply be a symptom of the necrophilia that infests so much of philosophy."[45]

All of the previous discussion has fueled countless arguments about what is called *backward*, *reverse*, or even *retro* causation. What is generally meant by forward causation is, of course, that any event that occurs at time t is caused by events that all occurred at some earlier time(s). Backward causation says that at least one of the causing events occurs after time t—this should make it clear that backward causation is a close relative of time travel. Indeed, one philosopher uses the terms *time traveler* and *retro-causal engineer* interchangeably.[46] The topic, understandably, is at the root of many hot philosophical debates, though not everybody (as this section's opening quote makes clear) thinks those debates are illuminating.

Just *why* does Professor Earman take his harsh position? He offers, as one reason, his disdain for the common philosophical 'proof' of the impossibility of backward causation: By definition, a cause is always before its effect. Yes, that's the entire 'proof.' One can, of course, win *any* argument by *defining* the answer to be what it is you wish to believe. More interesting, and certainly more pertinent to time travel, is the argument that if backward causation were possible then one could change the past—but that cannot be done because the past is dead and gone and thus unchangeable. That does seem to be a pretty solid argument against backward causation,[47] but Earman rebuts it by pointing out that the very same logic could be applied to the future, and so the usual, uncontested forward causation would also be denied. That is, one could argue that whatever the future will be, *will be* (literally 'by definition'), so one cannot change the future. A similar argument was presented even earlier,[48] in which we find "suppose that someone says 'I can change the future. I can do *this* or I can do *that*.' Well, then, suppose that he does *that*. Has he changed the future? No, because doing *that* was the future."

---

[43] A science fiction use of this idea is in B. W. Aldiss, "Man In His Time," *Science Fantasy*, April 1965, the story of an astronaut who returns from a trip to Mars and finds himself 3.3077 min ahead of everybody else.

[44] Science fiction had used a twist on this idea long before the film; see E. Binder, "The Man Who Saw Too Late," *Fantastic Adventures*, September 1939, a tale of what it might be like to have a 3 min *delay* in your vision.

[45] J. Earman, "Causation: A Matter of Life and Death," *Journal of Philosophy*, January 1976, pp. 5–25.

[46] B. Brown, "Defending Backwards Causation," *Canadian Journal of Philosophy*, December 1992, pp. 429–443.

The reversal of the 'usual' causal order of events by backward time travel has been a mainstay of science fiction almost from the start of the genre. Consider, for example, this tale.[49] A man on vacation by himself, without his wife along, meets a young lady—and they fall in love. The man loves his wife, too, though, and he realizes (as the young lady leaves him for the last time), never to return, that it is all for the best. But she really hasn't gone that far away from him, as the reader soon discovers. She is a time traveler from the future, and after leaving him she goes even further back in time, back an additional 20 years. She does this because she has learned that he met his wife 20 years ago, and so she goes back to be *that* woman! Thus, the usual causal order of the two events 'a long marriage' and the 'pre-marriage courtship' has been reversed (if we accept the fact that the man doesn't remember what his wife looked like when they married).

Actually, even our everyday uses of cause and effect are not nearly so straightforward as one might think, even when they are under far less stress than backward causation and time travel inflict. Consider, for example, the endless problems that are easy to imagine in the legal world. If a man falls off the roof of a ten-story building and is electrocuted as he plunges through power lines while still twenty feet above ground, was gravity or electricity the cause of death? Or was it both? As this example and others demonstrate,[50] one clearly does not have to discuss time travel to get into a serious argument about cause and effect. But with time travel, and the resultant backward causation, things can become even more perplexing. For example, we normally think it foolish to prepare, now, for an event that has already happened, but the prudent time traveler about to visit an ice age in the distant past would be wise to pack a fur coat before getting into his time machine!

---

[47] See, for example, D. H. Mellor, "Fixed Past, Unfixed Future," in *Michael Dummett: Contributions to Philosophy* (B. M. Taylor, editor), Martinus Nijhoff 1987.

[48] J. J. C. Smart, "A Review of *The Direction of Time*," *Philosophical Quarterly*, January 1958, pp. 72–77.

[49] R. F. Young, "The Dandelion Girl," *The Saturday Evening Post*, April 1, 1961.

[50] See also P. Mackie, "Causing, Delaying, and Hastening: Do Rains Cause Fires?" *Mind*, July 1992, pp. 483–500.

One philosopher provides, I think, a good start at explaining why so many other philosophers (and not just a few physicists) have adopted the 'common sense' position of rejecting backward causation. As he writes, "Part of the answer, no doubt, is a confusion between affecting and altering [the past—a distinction we'll discuss at length later in this book]. We cannot alter the past. But then we cannot alter the future either, although we can affect it. However, I take the common-sense rejection of backward causation to be, for the most part, quasi-empirical. It is based on a thought experiment. Think how you would set about affecting the past. By building a time-machine, perhaps? But how would you build one? We have no idea how to start. Yet, by contrast, we can work out how to affect the future . . . we just move our bodies."[51] But, as he goes on to argue, if we accept that we can't *change* the past (which means there is no way we could actually observe backward causation), then there still exists the possibility that past events were as they were because of events in the future.

Are there actual phenomena that justify a belief in the possibility of effect before cause in real time (not just in tape recorder time)? The only example I know of, and a controversial one at that, is a theoretical result from a reformulation of electrodynamics by the great English physicist Paul Dirac (1902–1984). Classical theory models electric charges as point objects of zero size, which causes problems when one tries to calculate certain details, such as the total field energy of a single electron. The answer comes out as infinity. In an attempt to find more reasonable (that is, finite) answers to such questions, Dirac modified the zero size of a charge to one taking them to be extended objects (while retaining the validity of Maxwell's equations for electrodynamics right down to a point). To calculate how such extended objects will behave mechanically, however, one has to include what are called the *self-interaction* forces, such as the force one side of an electron exerts on the other side.

When it was all worked through, Dirac arrived at a third-order differential equation of motion, an equation that involves a force term proportional not to the usual first time derivative of the velocity (that is, to the acceleration), but rather to the second derivative.[52] This force is proportional to the first derivative of the acceleration, and is a quantity of direct interest mostly to the designers of automobile suspensions, who call it the *jerk*. There is no force in physics, at least not in Newtonian physics, that shows that sort of dependence, and there are some curious consequences. For example, in Dirac's theory an electron experiencing no external force can still continually accelerate, exhibiting what is called a 'runaway solution.'

Dirac showed how the runaway solution can be eliminated by picking a particular value for what up to then was an arbitrary constant of integration in the analysis, but that trick causes, in turn, a new problem called 'pre-acceleration.' That is, if an electron experiences an external disturbance (Dirac considered a passing pulse of electromagnetic radiation), then the electron will start to move *before* the pulse reaches it! Now that does seem to be a pretty clear example of backward causation. The time interval during which the pre-acceleration occurs is very short, on the order of the time it takes light to travel across the spatially extended electron (about 10^{-24} s), but no matter. The apparent crack in the door of causality may be slight, but it was enough to satisfy some philosophers seeking scientific support for backward causation.

---

[51] P. Forrest, "Backward Causation in Defense of Free Will," *Mind*, April 1985, pp. 210–217.

[52] P. A. M. Dirac, "Classical Theory of Radiating Electrons," *Proceedings of the Royal Society A*, August 1938, pp. 148–168.

Not everybody liked this, however. One physicist was clearly uneasy about it, calling pre-acceleration "unpleasant" acausal behavior.[53] On the other hand, one can find believers, too.[54] Others have argued that the whole business is simply a non-problem. One philosopher, in fact, raised a very interesting technical point, arguing that Dirac's equation is non-Newtonian (remember the *jerk* force) and so we have no reason for coupling force and acceleration together as a cause-and-effect pair.[55] In Newtonian mechanics we do use that particular coupling, yet we do not think of force and velocity as a cause-and-effect pair because there is an integration operation involved in getting from to the other. Similarly, in Dirac's theory we have an integration operation separating force and acceleration.

One curious aspect to the debate on pre-acceleration is that many commentators seem not to have paid much attention to what Dirac himself had to say about it. As a Nobel laureate, it hardly seems likely that he would let such a result pass unnoticed and, indeed, his paper contains the following physical explanation: "It would appear that we have a contradiction with elementary ideas of causality. The electron seems to know about the pulse before it arrives, and to get up an acceleration … The behavior of our electron can be interpreted in a natural way, however, if we suppose the electron to have a finite size. There is then no need for the pulse to reach the center of the electron before it starts to accelerate. It starts to accelerate … as soon as the pulse meets its outside. Mathematically, the electron has no sharp boundary."

Two physicists suggested a fascinating connection between travel backward in time and Dirac's relativistically correct, quantum mechanical description of an electron.[56] They showed that in flat, two-dimensional spacetime the assumption of time travel to the past leads in a natural way to Dirac's equation. If, on the other hand, time travel only into the future is assumed, then additional assumptions are required to derive Dirac's equation. This connection between Dirac's equation and time travel to the past makes some philosophers and physicists nervous, but it didn't seem to bother Dirac. In fact, he went on in his paper to show how the pre-acceleration implies the possibility of building a device for sending a faster-than-light signal backward in time. Science fiction writers were, of course, quick to grasp that idea and such gadgets were dubbed "Dirac radios."[57]

---

[53] P. C. W. Davies, *The Physics of Time Asymmetry*, University of California Press 1977.

[54] J. Earman, "An Attempt to Add a Little Direction to 'The Problem of the Direction of Time'," *Philosophy of Science*, March 1974, pp. 15–47.

[55] A. Grunbaum, "Is Preacceleration of Particles in Dirac's Electrodynamics a Case of Backward Causation? The Myth of Retrocausation in Classical Electrodynamics," *Philosophy of Science*, June 1976, pp. 165–201.

[56] D. G. McKeon and G. N. Ord, "Time Reversal in Stochastic Processes and Dirac's Equation," *Physical Review Letters*, July 6, 1992, pp. 3–4.

One of the more perplexing aspects of backward causation is that it seems to allow for the possibility of *causal loops*, and for the breaking of such loops, a central feature in many of the very best time travel stories. For example, suppose there is a gadget such that if I push its control button *now*, then today's lecture notes will have appeared in the gadget's output tray *yesterday*. Indeed, yesterday I found today's notes there and, in fact, I am about to go to class to deliver that lecture. A mighty good one it is, too, so I think I think I'll send it back to yesterday in just a few minutes with the help of the gadget. But I haven't yet pushed the button. What if I now decide *not* to push the button? Why did the notes appear so I could use them today? Philosophers call this potential breaking of a causal loop a *bilking paradox*. Later in the book I'll discuss how such paradoxes have regularly appeared in the physics and philosophy literature since the 1940s.

By contrast, such paradoxes had been discussed in the science fiction magazines long before World War II. For example, in a letter to the editor at *Astounding Stories* (June 1932) a fan clearly stated his objection to time travel with the aid of a bilking paradox. He suggested the following experiment: Immediately publish an open offer to the inventor of time travel (who will be born, presumably, at some future date) to travel back to one week before the offer is published. But of course (argued the fan) we'd have a pretty problem if we then decided not to publish the offer after the inventor showed up! As that fan wrote, "Paradoxical? I'll say so, if time travel is possible." That fan didn't know about what seems to be a generic limitation on time machines, however: that one can't travel back to a date before the date of the time machine's creation. Thus, that fan's particular bilking paradox actually has no force.[58]

For another fictional example of a bilking paradox, consider the story[59] of time travelers who, just before they begin a trip into the future, see Earth invaded by Martians. At first the invaders are unbeatable, but then the defending military forces of Earth suddenly and mysteriously acquire a fantastically powerful new weapon. It isn't long before the time travelers realize where it came from—they themselves will go into the far future, obtain the weapon, and then return with it to what is now their own past (when the weapon first appeared). But then they wonder what might happen if they don't go, if instead they 'cheat time.' After all, they reason, why bother now to hunt for the weapon when the invasion has already been defeated? We are told that this potential bilking paradox is a “sinister conception, crawling evilly within their brains, like an unanswerable enigma.”

---

[57] See, for example, J. Blish, "Beep," *Galaxy Science Fiction*, February 1954.

[58] A similar bilking paradox had actually appeared the year before in the 1931 novel *Many Dimensions* by the English writer Charles Williams (1886–1945), which reads like a suitable script for an Indiana Jones movie.

[59] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940. There is an amusing reference in this tale to Orson Welles' famous radio-drama-hoax, from just 2 years earlier, of just such an alien invasion based on H. G. Wells' *War of the Worlds*.

Some philosophers, and practically all physicists, agree with that last assessment about bilking paradoxes, and so they believe there is simply nothing more to say. That is, bilking puzzles like the one in “The Time Cheaters” show that causal loops (and backward causation) must be impossible. Many feel this way about time loops, and backward causation, because (as is well known) time travel to the past can create all sorts of paradoxes. But such paradoxes are offensive only to human, culturally-biased intuitions on ‘how things ought to work,’ and not to the laws of physics which are indifferent to a reversal in the direction of time—which of course underlies what time travel is all about.

As the great American chemist G. N. Lewis expressed it, “Our common idea of time is notably unidirectional, *but this is largely due to the phenomena of consciousness and memory* [my emphasis].”[60] Lewis’ words caught the eye of the editor at one science fiction magazine, who summed it up for his readers in a half-page essay that contained dramatic words hinting at backward causation: “A new theory of time . . . reveals the possibility that events now occurring are among the factors that decided Caesar nearly 2,000 years ago to cross the Rubicon.”[61]

Lewis’ willingness to accept causality violations is not a universally popular view today. For example, one physicist has written[62] that “It is fair to say that most conservative physicists have very serious reservations about the admissibility and reality of causality-violating processes. Causality violation (i.e., the existence of a ‘time machine’) is such an extreme violation of our understanding of the cosmos that it behooves us to be as conservative as possible about introducing such unpleasant effects into our models.” He then goes on to declare closed timelike loops to be verboten because “the existence of closed timelike loops leads us to such unpleasant situations as meeting oneself 5 min ago.” He sums up his philosophical position nicely with “any theory that is ‘just a little bit causality violating’ is ‘just a little bit inconsistent.’”

Agreeing with this physicist is at least one philosopher who believes that the “association of causality with a particular temporal direction is not merely a matter of the way we speak of causes, but has a genuine basis in the way things happen” and that there is indeed an asymmetry with respect to past and future that is bound up with our concept of intentional action.[63] He then goes even further when he continues with the claim that being an agent of cause is not a necessary condition for seeing the asymmetry; being an observer is enough, as even an immobile yet intelligent tree (!) could detect the difference between past and future. (How he knows this about certain trees is left unexplained.)

---

[60] G. N. Lewis, “The Symmetry of Time in Physics,” *Science*, June 6, 1930, pp. 569–577.

[61] Editorial essay, “Two-Way Time,” *Astounding Stories*, September 1931.

[62] M. Visser, “Wormholes, Baby Universes, and Causality,” *Physical Review D*, February 15, 1990, pp. 1116–1124.

[63] M. Dummett, “Bringing About the Past,” *Philosophical Review*, July 1964, pp. 338–359.

The everyday views of causality that we have formed through our limited experiences when living in a world in which time travel is ‘uncommon’ may actually be incomplete. As the British philosopher Bertrand Russell (1872–1970) said with some humor long ago, in his 1912 Presidential Address (“On the Notion of Cause”) to the Aristotelian Society, “The law of causality, I believe, like much that passes muster among philosophers, is a relic of a by-gone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.” And I do agree with his fellow philosopher who, decades later, declared “The concept of cause is powerless to solve the problems posed by the concept of time. The fundamental laws of physics present our most careful, best established and most sophisticated understanding of time. Notoriously, nothing in these laws endorses the idea of a *flow of time nor of the direction* [my emphasis: we’ll return to both of these issues later in this chapter] which is basic to our conception of it. Nor are these laws causal (in the sense of singling out causes) even when they are deterministic. The concept of cause is not a fundamental one and cannot illuminate the darker corners in our understanding of the fundamental concept of time.”[64]

2.5 The Fourth Dimension

> “We are facing an invasion of fourth dimensional creatures . . . We are being attacked by life which is one dimension above us in evolution. We are fighting, I tell you, a tribe of hellhounds out of the cosmos. They are unthinkably above us in the matter of intelligence. There is a chasm of knowledge between us so wide and deep that it staggers the imagination.”[65]
>
> “Fourth dimension. Time factor. *You* know . . .”[66]

The idea of a fourth dimension to *space* has long been a staple of science fiction, but it has also long been viewed with suspicion. Indeed, many quite sophisticated scientists have thought it to be quite mysterious. For example, in his 1897 Presidential Address to the American Mathematical Society, the Canadian/American astronomer-mathematician Simon Newcomb (1835–1909) declared “The introduction of what is now very generally called hyperspace, especially space of more than three dimensions, into mathematics has proved a stumbling block to more than one able philosopher.” Einstein stated Newcomb’s view in blunter terms when he wrote “The non-mathematician is seized by a mysterious shuddering when he hears of ‘four-dimensional’ things, by a feeling not unlike that awakened by thoughts of the occult.”[67]

---

[64] G. Nerlich, “How to Make Things Have Happened,” *Canadian Journal of Philosophy*, March 1979, pp. 1–22.

[65] From “Hellhounds of the Cosmos,” *Astounding Stories*, June 1932, by Clifford Simak (1904–1988). Simak went on to write a number of much better tales, but this passage lends credence to the editorial introduction to the 1957 anthology *Famous Science-Fiction Stories* (Random House) that declared so much in the early pulp science fiction was “science that was claptrap and fiction that was graceless.”

[66] Uninformative ‘explanation’ given to a befuddled, inadvertent time traveler who emerges miles away and one hour backward in time after a wild ride through the fourth dimension in a gadget (constructed from a bicycle tire!) in the shape of a three-dimensional Möbius strip (see note 99 in Chap. 1). From the story by H. Nearing, Jr., “The Maladjusted Classroom,” *The Magazine of Fantasy and Science Fiction*, June 1953.

To see just how right Einstein was with this observation, consider the reaction one Egyptian philosopher had (in 1929) to Einstein’s own writings: “We have no doubt in our mind that nobody can understand it (the fourth dimension), including Einstein himself. The incomprehensibility of these assumptions [of general relativity] is due to their nature. They deal with the fourth dimension . . . and the reality of time and space. They can only be described by a mathematician’s hypothesis or by religious faith.”[68] This reaction is easy to understand—after all, anybody can ‘see’ that there are exactly three spatial dimensions, and that is that!

The 1901 novel *The Inheritors*, by the English writer Ford Madox Ford (1873–1939), like Simak’s, is the tale of an insidious hyperspace invasion of our world. It illustrates Einstein’s assertion about how many people react to the fourth dimension with an example from the time before the science fiction magazines. When the novel’s narrator is bluntly told by an invader that she (the invader) is from the fourth dimension—an idea inspired by Ford’s appreciation of how much success his acquaintance H. G. Wells had enjoyed with it—he recoils from that claim with the words “If you expect me to believe you inhabit a mathematical monstrosity, you are mistaken.” And who can really blame that skeptical narrator? How can there be *four* spatial dimensions? No less an authority than Aristotle, writing in 350 B.C., had declared in his essay “On the Heavens” that “the three dimensions are all that there are.”

Others were not so sure. In 1873, for example, we find an essay in *Nature* that refers to well-known mathematicians who even earlier had shown that they had an inner assurance of the reality of transcendental space (hyperspace).[69] The American philosopher Charles Sanders Peirce (1839–1914) was also an early advocate for the four-dimensionality of space. Just what he thought the nature of the fourth dimension to be is somewhat unclear, but the context of what he said suggests he took it to be spatial. He thought three-dimensional space to be “perverse” because of the existence of incongruous counterparts (such as left- and right-handed gloves), and this was apparently strong evidence for him that space could not be three-dimensional. Now, incongruous counterparts exist in all n-dimensional spaces, but Peirce preserved the special purity of the fourth dimension by suggesting that all physical objects, although capable of motion in the fourth direction, could themselves have no extent in that direction (remember, Peirce was a philosopher, not a physicist, and he offered no experimental support for any of this).[70]

---

[67] A. Einstein, *Relativity: the Special and General Theory*, Crown 1961, p. 33.

[68] From A. A. Ziadat, “Early Reception to Einstein’s Relativity in the Arab Periodical Press,” *Annals of Science*, January 1994, pp. 17–35.

[69] G. F. Rodwell, “On Space of Four Dimensions,” *Nature*, May 1, 1873, pp. 8–9.

But is it *really* possible that there could be *four* spatial dimensions? We experience three independent directions, each lying at a right angle to the other two—but why just *three*, and not ten or fifteen? Indeed, in an 1888 talk to the Philosophical Society of Washington, Simon Newcomb dismissed the view that space must necessarily be three-dimensional as an "old metaphysical superstition." Yet, despite Newcomb's open-mindedness, it has been shown that in the framework of classical physics there are, in fact, several powerful reasons for why there must be *exactly three* spatial dimensions.

The beginning of a scientific explanation for the dimensionality of space appears in Kant, who believed the *three* dimensions of space and Newton's inverse-square law for gravity are intertwined (but he offered nothing beyond philosophical speculation). The origin of Kant's view is actually quite old, dating back to the ancient Greeks, who had already begun to suspect that there was something special about *three* dimensions, at least as far as geometry was concerned. They knew of the infinity of regular two-dimensional polygons, but that there were just five regular polyhedrons in three dimensions (the so-called *Platonic solids*). This early observation was trapped in mystical speculations, however, and it wasn't until the development of physics as a science that non-mystical discussions on the dimensionality of space began to appear.

Beginning with the work of Einstein's friend, the Austrian/Dutch physicist Paul Ehrenfest (1880–1933) in 1917, we can find the idea that the Poisson-Laplace equation, a second-order partial differential equation that describes the potential functions for both Newtonian gravity and electrostatics, does not allow for stable planetary or electronic orbits in *any* space with dimensionality greater than three. Further, the distortionless, reverberation-free propagation of both electromagnetic and sound waves is possible only in spaces of dimensions one and three. These conclusions have been shown to hold even when we go beyond nineteenth century physics into general relativity and quantum mechanics.[71]

Using a slightly different approach, a biological-topological argument for why space cannot have fewer than three dimensions exists. In all of our common experience, complex intelligent life is always found to occur as an aggregate of a vast number of elementary cells, interconnected via electrical nerve fibers. Each cell is connected to several others, *not all immediate neighbors*, by these fibers. If space had only one or two dimensions, then such highly interconnected nets of cells would be impossible because the overlapping nerve fibers would have to intersect, which would result in their mutually short-circuiting one another.

---

[70] R. R. Dipert, "Peirce's Theory of the Dimensionality of Physical Space," *Journal of the History of Philosophy*, January 1978, pp. 61–70.

[71] See, for example, I. M. Freeman, "Why Is Space Three-Dimensional?" *American Journal of Physics*, December 1969, pp. 1222–1224, and L. Gurevich and V. Mostepanenko, "On the Existence of Atoms in n-Dimensional Space," *Physics Letters A*, May 31, 1971, pp. 201–202.

It wasn't long before these views on the dimensionality of space found their way into science fiction. An early use of space as four-dimensional occurs in an awkward rewrite of Jules Verne's *Around the World in Eighty Days*, in which a professor and his crew fly into hyperspace and around the world and to the moon and back, in less than a day.[72] They do this with a plane equipped with a four-dimensional rudder! More interesting is the tragic story (originally published in 1926) of a math professor who learns how to move into hyperspace and back.[73] A colleague catches him at it and, once over his astonishment, asks what is behind it all. The professor replies, "My assumption is that the fourth dimension is just another dimension—no more different in kind from length, say, than length is from breadth and thickness, but perpendicular to all three. Now suppose that a being in two dimensions—a flat creature, like [a moving shadow on a surface]—were suddenly to grasp the concept of a third dimension and so step out of the [surface]. He might move only an inch, but he would vanish completely from the sight of the world."

The professor has similarly learned how to step out of 3-space and into 4-space but, when asked to explain *how*, all he can say is "How can I explain? It's just the *other* direction. It's *there*!" His colleague can't see it, but nonetheless is quick to grasp the practical implications: "This is power! Think of it! A step, and you are invisible! No prison cells can hold you, for there is a side to you on which they are as open as a wedding ring! No ring is secure from you: you can put your hand *round the corner* and draw out what you like. And, of course, if you looked back on the Universe you had left, you would see us in sections, open to you! You could place a stone or a tablet of poison right in the bowels of your enemies!"

What the professor's colleague is getting at involves a comparison with a prison in planar 2-space, which would merely be a circle around the captive. Knowledge of the third dimension would make it possible to escape, however, by simply moving along that new direction, over the circle, and then back into the plane. To a 2-space guard it would seem that the prisoner had suddenly vanished from view *inside* the circle and then just as suddenly materialized again *outside* the circle. Similarly, to escape from a 3-space prison, one would merely move along the fourth dimension, and in the same way one could remove the yolk from an egg without damaging the shell; indeed, one could remove the yolk directly from the chicken without damaging the chicken![74]

---

[72] B. Olsen, "Four Dimensional Transit," *Amazing Stories Quarterly*, Fall 1928.

[73] R. Hughes, "The Vanishing Man," reprinted in *The Mathematical Magpie* (C. Fadiman, editor), Simon and Schuster 1962.

[74] This astounding insight appeared in early pulp science fiction in, for example, M. J. Breuer, "The Appendix and the Spectacles," *Amazing Stories*, December 1928. The concept appeared even earlier in Bob Olsen, "The Four-Dimensional Roller-Press," *Amazing Stories*, June 1927, and then later in Olsen's "The Great Four Dimensional Robberies," *Amazing Stories*, May 1928 to rob locked safe deposit boxes, and "The Four Dimensional Escape," *Amazing Stories*, December 1933, in which a man sentenced to die by hanging at San Quentin Prison is rescued, while standing on the gallows' trap, by an inventor who pulls him through the fourth dimension.

In a later tale[75] we meet another professor who dramatically uses this very feature of the fourth dimension. His right hand has been modified through an accident to exist in four-dimensional hyperspace and so, to finance his research, he uses his ‘talent’ to become the perfect pickpocket, able to reach into any wallet no matter how well secured. He also can, indeed, reach right into the very bowels of his fellow man. And he *does*. When he demonstrates his hand to the policeman who has arrested him for being a thief, the astonished officer chokes on a lemon drop. Dr. Fuddles then, of course, does the right thing and removes the drop from the poor fellow with ease. There is one additional aspect to Dr. Fuddles’ hand, however, that the story missed. If he had turned his right hand over in the fourth dimension, then he would have had *two* left hands!

It was discovered in 1827 by Möbius (of the strip) that any three dimensional object can be converted into its mirror image by flipping it over through the fourth dimension. Thus, a left-handed glove can be made by pure geometry (no scissors, thread, or needle required) into a precise copy of its right-handed mate. If a living organism is so flipped, however, there may be a problem, as everything in the body would be reversed, including the optically active organic molecules discovered by Pasteur in 1848, which are involved in vital biological processes. These molecules, called *stereoisomers*, exist in two versions in nature (the left-handed and the right-handed versions, if you will), but our bodies have developed the ability to use only one version. To be flipped through the fourth dimension would make some reversed stereoisomers unable to participate in the digestion of food and we would starve to death.

For modern science fiction writers the fourth dimension (and hyperspace, in general), is still a major concept. One physicist, writing in *Analog* (today’s premier ‘hard science’ fiction magazine), summed up nicely what was so fascinating in early pulp, and still is today, about the idea of an extra dimension or two, or perhaps even more, at least from a fictional point of view: “Are there hidden dimensions not accessible to us, dimensions in which we could go adventuring, dimensions within which malevolent hyper-dimensional aliens may be lurking, ready to pierce our flimsy paper-thin three-space bodies with their terrible hyper-sharp claws?”[76] The early pulp science fiction magazines encouraged this lurid imagery. Witness the editorial blurb that opened one many-dimensional monster story as follows: “It was a strange world in which Lester and Florence found themselves. A world of sudden

---

[75] N. Bond, “Dr. Fuddle’s Fingers,” in *Mr. Mergenthwirker’s Lobblies and Other Fantastic Tales*, Coward-McCann 1946.

[76] J. Cramer, “The Other Forty Dimensions,” *Analog*, April 1985. ‘Monsters in hyperspace’ stories were numerous in pulp science fiction. Three examples (in no particular order of literary merit!) are: M. J. Breuer, “The Einstein See-Saw,” *Astounding Stories*, April 1932; P. Ernst, “The 32nd of May,” *Astounding Stories*, April 1935; “The Monster from Nowhere,” *Fantastic Adventures*, July 1939.

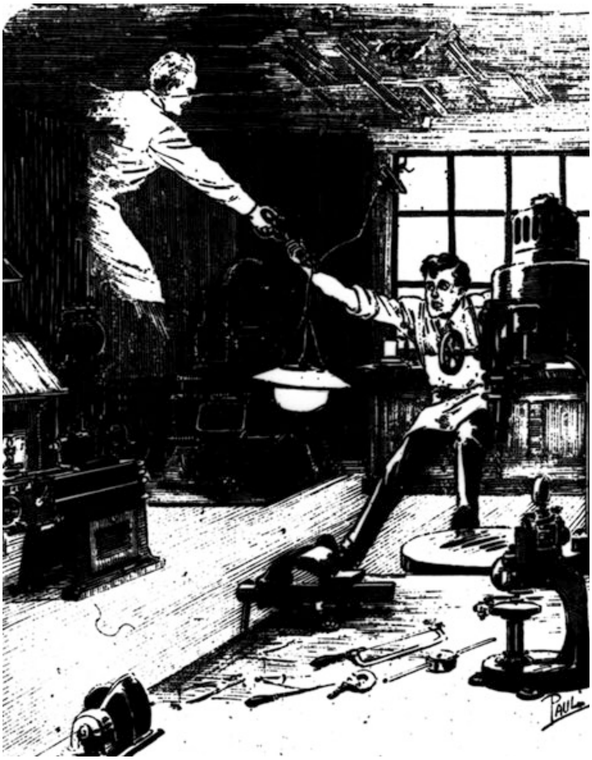

Fig. 2.1 An experiment in hyperspace goes astray. The young man is pulling on "hyper-forceps" in an attempt to retrieve a surgeon who has fallen out of 3-space (along with his patient, a professor of non-Euclidean geometry, who suffers from gallstones). The hyper-forceps allow the removal of the gallstones without cutting into the body. Illustration by Frank R. Paul, ©1928 by Experimenter Publishing Co. for "Four Dimensional Surgery" (*Amazing Stories*, February 1928) by Bob Olsen, reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

death and strange science, ruled by inhuman beasts."[77] But as outrageous as that might sound, the real physics of hyperspace is even more amazing.

Hyperspace is, in general, simply any space with more dimensions than the one we obviously seem to live in. In particular, our universe appears to be a four-dimensional (three spatial and one temporal) hyperspace called *spacetime*. This four dimensional world can, at least mathematically, be thought of as the boundary surface of a five dimensional hyperspace. This is analogous to the way the two-dimensional space of the surface of a sphere bounds the three-dimensional space of the sphere itself. This interesting imagery appeared quite early in pulp science fiction. For example, in one remarkably sophisticated story, an eccentric scientist at one point exclaims "A mathematical physicist lives in vast spaces … where space unrolls along a fourth dimension on a surface distended from a fifth."[78]

There are some interesting geometrical implications to hyperspace which play big roles in time travel considerations. For example, for beings in the two-dimensional world of a sphere's surface there are *two* ways to travel from pole to pole. There is the usual way, *on* the surface of the sphere, and the hyperspace way which takes them *through* the sphere along the polar diameter. In imagery motivated by thinking of the sphere as an apple, and of the hyperspace path as a tunnel bored by a worm through the apple, it has become popular to call all such shortcuts, through any hyperspace of any dimension, *wormholes* (a word coined in the 1950s by the Princeton physicist-wordsmith John Wheeler). Wheeler used wormholes to show how electric charge could be thought of as lines of force trapped in the changing topology of a multiply connected space (indeed, Wheeler claimed that the observation of what we call electricity is experimental evidence that space is *not* simply connected).[79]

---

[77] M. Duclos, "Into Another Dimension," *Fantastic Adventures*, November 1939. See the illustration for this story in "Some First Words."

[78] M. J. Breuer, "The Gostak and the Doshes," *Amazing Stories*, March 1930.

The general theory of relativity predicts the existence of wormholes in spacetime and, in fact, they were first 'discovered' theoretically in the mathematics of relativity as early as 1916 by the Viennese physicist Ludwig Flamm (1885–1964). Later analyses were done by Einstein, himself.[80] Wormholes have been discussed as a possible model for pulsars (as opposed to the more usual model as rotating neutron stars).[81] It has also been suggested that the interior of a charged black hole may be the entrance to a wormhole.[82] All of these various solutions to the gravitational field equations are generically called "Einstein-Rosen bridges" in the physics literature (see note 81, for example), and the term soon appeared in fiction, too.[83]

The use of hyperspace wormhole portals for explaining some observed physical phenomenon appeared in the scientific literature long before Wheeler's electricity example. In his 1928 book *Astronomy and Cosmogony*, for example, the British theoretician Sir James Jeans devoted a chapter to what were then called nebulae, the island-universes we now call galaxies. At the end of his discussion on the arms of spiral galaxies, Jeans offered the following speculation: "Each failure to explain the spiral arms makes it more and more difficult to resist a suspicion that the spiral nebulae are the seats of types of forces entirely unknown to us, forces which may possibly express novel and unsuspected *metric properties of space* [my emphasis]. The type of conjecture which presents itself, somewhat insistently, is that the centers of the nebulae are of the nature of 'singular points,' at which matter is poured into our universe from some other, and entirely extraneous, special dimension, so that, to a denizen of our universe, they appear as points at which matter is being continually created." This, in everything but name, is a wormhole.

---

[79] A space is simply connected if *all* the points on the straight line that joins *any* two points in the space are also in the space. The interior of a sphere is simply connected. The interior of a sphere with a hole in it is *not* simply connected.

[80] A. Einstein, "The Particle Problem in the General Theory of Relativity," *Physical Review*, July 1, 1935, pp. 73–77.

[81] J. M. Cohen, "The Rotating Einstein-Rosen Bridge," in *Relativity and Gravitation* (C. G. Kuper and A. Peres, editors), Gordon and Breach Science Publishers 1971.

[82] A. Ori, "Inner Structure of a Charged Black Hole: An Exact Mass-Inflation Solution," *Physical Review Letters*, August 12, 1991, pp. 789–792.

[83] See, for example, J. G. Cramer, *Einstein's Bridge*, Avon 1997 (this is the same Cramer cited in note 76). The *Rosen* comes from the American-Israeli physicist Nathan Rosen (1909–1995), who was a collaborator of Einstein's.

What would hyperspace be like? It is intuitively obvious that in the case of the 2-D surface of a 3-space sphere, the 'hyperspace' wormhole path is shorter than the surface path. Even if this 'shorter path' view holds for wormholes in our 4-D spacetime, however, getting around in science fiction hyperspace may not be a simple task. One tale, for example, tells the story[84] of how one of the first spaceships to explore hyperspace gets lost. The trouble with hyperspace travel is that "You go in at one point, you rocket around until you think it's time to come out, and there you are. Where is 'there'? Why, that's the surprise that's in store for you, because you never know until you get there. And sometimes not even then." The same idea plays a central role in Robert Heinlein's 1957 novel *Tunnel in the Sky*, in which a 'hyperspace gate' is discovered by accident during failed time travel experiments.

Another story[85] asks the same question about hyperspace, and arrives at the same answer: "When you took the Jump … how sure were you *where* you would emerge? The timing and quantity of the energy input might be as tightly controlled as you liked … but the uncertainty principle reigned supreme and there was always the chance, even the inevitability of a random miss … a paper-thin miss might be a thousand light-years."

A common way to visualize hyperspace wormhole shortcuts is to imagine the beginning and the end of a journey as points A and B on the 2-D surface of a piece of paper. Then imagine that the paper is folded so as to position A over B, perhaps with A almost touching B. The distance from A to B *through hyperspace* (the 3-D space in which the folding took place) can clearly be much less than is the distance through 'normal' space (the distance covered by a trip that always remains in the 2-D surface). This is the specific example used in one tale to explain the instantaneous "space-warp" (wormhole) device invented by the story's hero.[86] Such imagery actually appeared quite early in science fiction, as in one story in which a gadget is used to "bend space" so that Earth and Venus touch![87]

The idea of hyperspace folding has broken free from science fiction and can now be found in modern stories in other genres. For example, in one Stephen King story ("Mrs. Todd's Shortcut") a woman keeps finding ever shorter ways to drive from Castle Rock, Maine to Bangor. As the crow flies it is 79 miles, but she gets the journey down to 67 miles, and later to 31.6 miles. When doubted, she replies: "Fold the map and see how many miles it is then … it can be a little less than a straight line if you fold it a little, or it can be a lot less if you fold it a lot." The doubter remains unconvinced: "You can fold a map on paper, but you can't fold *land*."

---

[84] F. Pohl, "The Mapmakers," *Galaxy Science Fiction*, July 1965.

[85] I. Asimov, "Take a Match," in *New Dimensions II: Eleven Original Science Fiction Stories* (R. Silverberg, editor), Doubleday 1972.

[86] G. O. Smith, "The Möbius Trail," *Thrilling Wonder Stories*, December 1948.

[87] E. L. Rementer, "The Space Bender," *Amazing Stories*, December 1928.

For the purpose of wormhole creation in spacetime, we actually have to imagine much more: the folding of four-dimensional spacetime through a five dimensional hyperspace. The folding imagery has even appeared in the movies: spacetime folding is demonstrated with a piece of paper in both *Event Horizon* (perhaps the worst movie of 1997) and the 2014 *Interstellar*.

Another feature of hyperspace that science fiction has taken a liking to is its vastness. An interesting fictional treatment of this idea was given by a writer who, in real life, was an academic psychologist at the University of Michigan. He put himself in a story[88] of a starship captain who is explaining to the crew psychologist how he feels about hyperspace (or *subspace*, as it is called in the story): "God forsaken. That's just what it is. Completely black, completely empty. It frightens me every time we make the jump through it . . . it frightens me because—well, because a man seems to get lost out there. In normal space there are always stars around, no matter how distant they may be, and you feel that you've got direction and location. In subspace, all you've got is nothing—and one hell of a lot of that. It's incredible when you stop to think about it. An area—an opening as big as the whole of our Universe, big enough to pack every galaxy we've ever seen in it—and not a single atom of matter in it . . . until we came barging in to use it as a shortcut across our own Universe."

The vastness of hyperspace got a more humorous treatment from the early pulp science fiction writer Bob Olsen (1884–1956), who wrote the following verses[89] in the introduction to one of his many stories of the fourth dimension:

I read a yarn the other day—
A crazy concept, I must say.
It states that objects have extension
In what is called the "Fourth Dimension."
In hyperspace one could, no doubt,
Make tennis balls turn inside out;
And from a nut remove the kernel
And not disturb the shell external.
A crook could pilfer bonds and stocks,
Then laugh at prison bars and locks,
One step in this direction queer,
And presto! He would disappear!
Let's hope, in planning new inventions,
They'll give us cars with four dimensions.
When searching for a parking place
We sure could use some hyperspace!

It is not just science fiction that takes hyperspace seriously. We find a mathematician, for example, writing that "most science fiction addicts are familiar with the notion of 'hyperspace,' a higher dimensional space-time bounded by Space-Time through which, in the far distant future, interstellar voyages shortcut the (otherwise unsurmountable) distances between the stars. The purpose of this article[90] is to demonstrate that any . . . relativistic space-time model is the boundary of some . . . five-dimensional hyperspace." That is just what Breuer's magazine character (see note 78) said—in 1930!

---

[88] J. V. McConnell, "Avoidance Situation," *If*, February 1956.

[89] B. Olsen, "The Four-Dimensional Auto-Parker," *Amazing Stories*, July 1934. "Bob Olsen" was the pen-name for Alfred Johannes Olsen.

The concept of *time* as a fourth dimension has long been a popular concept, and science fiction in particular has embraced it with enthusiasm. We find a little joke on the idea in a story where a young couple, visited by time travelers from 500 years in the future, are said to live in Apartment 4-D.[91] One physicist[92] traced the idea back to the late eighteenth century, finding references to the idea in pre-1800 works of the great French mathematical physicists Jean le Rond d'Alembert (1717–1783) and Joseph-Louis Lagrange (1736–1813). In fact, a philosopher[93] has found a 1751 passage written by d'Alembert that appears to indicate that it is some unknown, earlier person to whom the credit should really go: "I have said [that it is] not possible to imagine more than three dimensions. A clever acquaintance of mine believes, however, that duration could be regarded as a fourth dimension and that the product of time and solidity would be in some way a product of four dimensions; that idea can be contested, but it seems to me that it has some merit, if only that of novelty."

Still, it wasn't until a curious letter appeared in *Nature* in 1885 that the concept of time as the fourth dimension was mentioned seriously in an English-language scientific journal. The author, mysteriously signing himself only as "S.," began by asking "What is the fourth dimension? . . . I [propose] to consider Time as a fourth dimension . . . Since this fourth dimension cannot be introduced into space, as commonly understood, we require a new kind of space for its existence, which we may call time-space."[94] Who was this prophetic writer that, if he had just made a simple swap, would have been the first to use space-time as a word? Nobody knows. Bork speculates that it was an acquaintance of H. G. Wells, but Wells himself is on record that it certainly wasn't him.

In his 1934 *Experiment in Autobiography*, Wells wrote "In the universe in which my brain was living in 1879 there was no nonsense about time being space or anything of that sort. There were three dimensions, up and down, fore and aft and right and left, and I never heard of a fourth dimension until 1884 [when Wells was eighteen] or thereabout. Then I thought it was a witticism." He had, in fact, said this before. In a 1931 edition of *The Time Machine* (Random House), for example, he wrote in the Preface that the idea for the novel "was begotten in the writer's mind by students' discussions in the laboratories and debating society of the Royal College of Science in the eighties and already it had been tried over in various forms by him before he made this particular application of it."

---

[90] G. S. Whiston, "'Hyperspace' (The Cobordism Theory of Space-Time)," *International Journal of Theoretical Physics*, December 1974, pp. 285–288.

[91] L. Padgett, "When the Bough Breaks," *Astounding Science Fiction*, November 1944.

[92] A. M. Bork, "The Fourth Dimension in Nineteenth-Century Physics," *Isis*, October 1964, pp. 326–338.

[93] E. Meyerson, *The Relativistic Deduction*, volume 83 of *Boston Studies in the Philosophy of Science*, D. Reidel 1985, p. 78.

[94] S., "Four-Dimensional Space," *Nature*, March 26, 1885, p. 481. The editorial staff at *Nature* has informed me that, more than a century-and-a-quarter later, there is no longer any record of the identity of S. in the journal's archives.

The idea of time as the fourth dimension entered the popular mind around 1894–95, with the publication of the first of Wells' so-called "scientific romances," *The Time Machine*. Then, after that pioneering use of time as the fourth dimension, science fiction quickly adopted the idea as the basis for one of its most popular subgenres. One of the great "golden age of science fiction" writers, 'Murray Leinster' (1896–1975)—the pen-name for William Jenkins—used it as the basis for his first published story.[95] It is the incredible tale of a Manhattan skyscraper (and its 2000 occupants) sent backward in time several 1000 years because its foundation slips (in an unexplained way) along the fourth dimension. The scientific sophistication of the story is primitive, with just one of the many logical flaws being a vivid description of the time travelers living forward-in-time even as their wrist watches run backward. Indeed, when pulp pioneering editor Hugo Gernsback reprinted the tale in one of the early issues of *Amazing Stories*, a reader complained about that very point. Gernsback felt compelled to defend the story, but could muster only a weak rebuttal based on an author's right to "poetic license."[96]

More technical is the discussion in the story of a clerk who transforms the main entrance to a department store into a time machine by building a tesseract (a four-dimensional cube).[97] The claim is made there that the fourth dimension of the cube/doorway is time. That tale appeared just 5 months after a classic of science fiction by Robert Heinlein (1907–1988) had appeared, also using a tesseract, in which the fourth dimension is taken as *spatial*.[98]

Some writers wanted to have the fourth dimension both ways, as space *and* time in the same story. One wonderful example of this is a classic,[99] written by one of the giants of science fiction. In that tale an electrical engineer named Nelson is caught in the middle of an enormous electromagnetic field surge produced by a short circuit in a power plant. As a physicist explains to the shocked board of directors of the utility, "It now appears that the unheard-of-current, amounting to millions of amperes . . . must have produced a certain extension into four dimensions . . . I have been making some calculations and have been able to satisfy myself that a

---

[95] M. Leinster, "The Runaway Skyscraper," *Argosy*, February 1919.

[96] H. Gernsback, "Plausability in Scientifiction," *Amazing Stories*, November 1926.

[97] W. P. McGivern, "Doorway of Vanishing Men," *Fantastic Adventures*, July 1941.

[98] R. Heinlein, "—And He Built a Crooked House," *Astounding Science Fiction*, February 1941. Here we read of a Los Angeles architect who builds a house in the shape of a tesseract as it would appear if collapsed into normal three-dimensional space. It isn't stable in 3-space (we are told), however, and so a California earthquake is sufficient to topple the house into a stable 4-D configuration, along with its occupants.

[99] A. C. Clarke, "Technical Error," *Fantasy No. 1*, December 1946.

'hyperspace' about ten feet on a side was, in fact, generated: a matter of some ten thousand quartic—not cubic!—feet. Nelson was occupying that space. The sudden collapse of the field [when the overload breakers finally broke the circuit] caused the rotation of that space."

Being rotated through 4-space has inverted the unlucky Nelson [see *For Further Discussion* at the end of this chapter for more on this point], and to bring him back to normal he must be flipped again. The physicist brushes aside a question about the fourth dimension as time, asserting that the only issue is one of space. Poor Nelson is, therefore, again subjected to a stupendous power overload—only now he disappears! Too late, the physicist realizes that the fourth dimension is both space *and* time and that Nelson has been spatially flipped *and* temporally displaced into the future. To understand the particularly monstrous fate of Nelson, just ask yourself what the result would be if he should materialize *inside* matter sometime in the future!

The interpretation of the fourth dimension as time is, of course, the one of interest to prospective time travelers, to physicists studying time travel, and to philosophers of time, and so for us, too. The sort of science fiction that is of greatest interest to us is like the one in which one of the characters, displaced in time, asks for an explanation from a higher-dimensional being who appears on the scene: "'Just where is Tuesday?' he asked. 'Over there [and when the being extends its hand, the hand disappears].' 'Do that again.' 'What? Oh—Point toward Tuesday? Certainly.'" The being explains the physics of the situation to the astonished time traveler thus: "It is a direction like any other direction. You know yourself there are four directions—forward, sideward, upward, and—*that* way! . . . It is the fourth dimension—it is duration."[100]

And how about stories like the one in which a mad inventor discovers how to make a substance whose atoms resist being pushed by "pushing back at right angles to all the other [spatial] directions." That is, to push on this exotic stuff is to risk experiencing a back reaction, of being pushed "off into the fourth dimension [which we are told is time] . . . into the middle of the week after next."[101] Now wouldn't *that* really be something?!

But of course it was H. G. Wells who, in fiction, pioneered time travel and its connection to the fourth dimension as it is popularly thought of today (with the caveats about Wellsian time machines kept firmly in mind). We are therefore quite interested, as *The Time Machine* opens, to listening-in as the Time Traveller expounds to a group of friends at a dinner party in his London home. He starts with the assertion "There is no difference between Time and any of the three dimensions of Space except that our consciousness moves along it." When asked to say more about the fourth dimension, he replies, "It is simply this. That Space, as our mathematicians have it, is spoken of as having three dimensions, which one may call Length, Breadth, and Thickness, and it is always definable by reference to three planes, each at right angles to the others. But some philosophical people have been asking why *three* dimensions particularly—why not another direction at right angles to the other three?—and have even tried to construct a Four-Dimensional geometry. Professor Simon Newcomb was expounding this to the New York Mathematical Society only a month or so ago."[102]

---

[100] T. Sturgeon, "Yesterday Was Monday," *Unknown Fantasy Fiction*, June 1941.

[101] M. Leinster, "The Middle of the Week After Next," *Thrilling Wonder Stories*, August 1952.

2.6 Spacetime and the Block Universe

> "And now he has preceded me briefly in bidding farewell to this strange world. This signifies nothing. For us believing physicists, the distinction between past, present, and future is only an illusion, even if a stubborn one."
>
> —Albert Einstein[103]

The poet Henry Van Dyke wrote, in his 1904 "The Sun-Dial at Wells College," words that echo the spirit of Omar Khayyam's *Rubaiyat* from nine centuries before:

The shadow by my finger cast
Divides the future from the past:
Before it, sleeps the unborn hour,
In darkness, and beyond thy power:
Behind its unreturning line,
The vanished hour, no longer thine:
One hour alone is in thy hands,—
The NOW on which the shadow stands.

The very next year Einstein's theory of special relativity appeared and, 3 years later, came Minkowski's spacetime interpretation of special relativity. Van Dyke's beautiful poetry was dealt a mighty blow by those developments in mathematical physics, and in the rest of this chapter we'll see how that came to pass.

The modern view of reality, that the past, present, and future are joined together into a four-dimensional entity called *spacetime*, is due to Hermann Minkowski (1864–1909), Einstein's mathematics teacher when he was a student in Zurich. Minkowski gave spacetime (the visual imagery of Einstein's mathematics) to the world during a famous address at the 80th Assembly of German Natural Scientists and Physicians in Cologne, on September 21, 1908. Entitled "Space and Time," his remarks were electrifying then and still are today.[104] He began dramatically: "Gentlemen! The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical." Then came the famous line, quoted in so many freshman physics texts and philosophy papers, concerning the nature of spacetime: "Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve independence." Minkowski explained what spacetime is in these words to his audience:

---

[102] And so Newcomb actually was. Wells, it is certain, routinely read *Nature* (one of his college friends, Richard Gregory, eventually became the journal's editor), and Wells must have read Newcomb's address of December 28, 1893 to the New York Mathematical Society when reprinted in the February 1, 1893 issue (on pp. 325–329), where he called hyperspace "the fairyland of geometry." From the Time Traveller's own words, then, that wonderful Victorian dinner party must have taken place in January or February of 1894.

[103] From a letter written by Einstein on March 21, 1955, to the children of Michele Besso, his dearest friend, who had just died. Einstein's use of the word *briefly* was due to his knowledge that he was nearly out of time, too (he died just a month later).

> "A point of space at a point of time … I will call a *world point*. The multiplicity of all thinkable x, y, z, t systems of values we will christen the *world*. With this most valiant piece of chalk I might project upon the blackboard four world axes … Not to leave a yawning void anywhere, we will imagine that everywhere and everywhen there is something perceptible. To avoid saying 'matter' or 'electricity' I will use for this something the word 'substance.' We fix our attention on the substantial point which is at the world point x, y, z, t, and imagine that we are able to recognize this substantial point at any other time. Let the variations dx, dy, dz, of the space coordinates of this substantial point correspond to the time element dt. Then we obtain, as an image, so to speak, of the everlasting career of the substantial point, a curve in the world, a *world-line*. … The whole Universe is seen to resolve itself into similar world-lines, and I would fain anticipate myself by saying that in my opinion physical laws might find their most perfect expressions as relations between these world-lines … *Thus also three-dimensional geometry becomes a chapter in four-dimensional physics* [my emphasis]."

With those words Minkowski gave mathematical expression to the philosophical exposition of Wells' Time Traveller to his dinner party friends. Taking the Minkowskian view of the primacy of spacetime as the ultimate building block stuff of reality was Princeton professor of physics John Wheeler, who wrote[105] "There is nothing in the world except empty curved space. Matter, charge , electromagnetism … are only manifestations of the bending of space. *Physics is Geometry*." This idea was echoed in fiction, in the 1987 novel Moscow 2042 by Vladimir Voinovich, where we find a time traveler who declares "Anyone with even a nodding acquaintance with the theory of relativity knows that nothing is a variety of something and so you can always make a little something out of nothing."

But not everybody understood Minkowski. In a little-known yet quite erudite essay, published just after a stunning experimental verification of general relativity (the bending of starlight passing through the Sun's gravitational field[106]), an anonymous author presented an optical analogy to help those who thought relativity simply “a mathematical joke.” Signing himself only as “W.G.,” he included the following passage[107]:

---

[104] For a study that includes the original German text, careful English translations, and photographs of Minkowski's agonized corrections to his pre-address manuscript, see P. L. Galison, "Minkowski's Space-Time: From Visual Thinking to the Absolute World," *Historical Studies in the Physical Sciences* (volume 10), 1979, pp. 85–121.

[105] C. W. Misner and J. Wheeler, "Gravitation, Electromagnetism, Unquantized Charge, and Mass as Properties of Curved Empty Space," *Annals of Physics*, December 1957, pp. 525–603.

[106] General relativity had already explained the long-puzzling excess precession of the perihelion (point of closest approach to the Sun) of Mercury's orbit. The excess was an observational (and so experimental) fact which Newton's gravity *cannot* completely explain.

> “Some thirty or more years ago [it was forty] a *jeu d’esprit* was written by Dr. Edwin Abbott entitled *Flatland* . . . Dr. Abbott pictures intelligent beings whose whole experience is confined to a plane, or other space of two dimensions, who have no faculties by which they can become conscious of anything outside that space and no means of moving off the surface on which they live. He then asks the reader, who has consciousness of the third dimension, to imagine a sphere descending upon the plane of Flatland and passing through it. How will the inhabitants regard this phenomenon? They will not see the approaching sphere and will have no conception of its solidity. They will only be conscious of the circle in which it cuts their plane. This circle, at first a point, will gradually increase in diameter, driving the inhabitants of Flatland outward from its circumference, and this will go on until half the sphere has passed through the plane, when the circle will gradually contract to a point and then vanish, leaving the Flatlanders in undisturbed possession of their country . . . Their experience will be that of a circular obstacle gradually expanding or growing, and then contracting, and they will attribute to *growth in time* what the external observer in three dimensions assigns to a movement in the third dimension. Transfer this analogy to a movement of the fourth dimension through three-dimensional space. Assume the past and future of the Universe to be all depicted in four-dimensional space, and visible to any being who has consciousness of the fourth dimension. If there is motion of our three-dimensional space relative to the fourth dimension, all the changes we experience and assign to the flow of time will be due simply to this movement, *the whole of the future as well as the past existing in the fourth dimension* [my emphasis].”

W.G.’s words are a clear and unequivocal statement of the so-called *block universe* concept of four-dimensional spacetime. One can find the block universe concept in the writings of the ancients, too. Consider, for example, the fifth-century B.C. Greek philosopher Parmenides’ view of reality: “It is uncreated and indestructible; for it is complete, immovable, and without end. Nor was it ever, nor will it be; for now it *is*, all at once, a continuous *one*.” And in Thomas Aquinas’ *Compendium Theologiae*, written in the thirteenth century, we find “We may fancy that God knows the flight of time in His eternity, in the way that a person standing on top of a watchtower embraces in a single glance a whole caravan of passing travelers.” This is the block universe idea, too, but whereas for Parmenides it was metaphysics and for Aquinas it was theology, for Einstein and Minkowski it was physics.[108]

---

[107] W. G., “Euclid, Newton, and Einstein,” *Nature*, February 12, 1920, pp. 627–630. As with the mysterious S. (note 94), the editorial staff at *Nature* has informed me that, nearly a century later, there is no longer any record of the identity of W. G. in the journal’s archives.

[108] And for some it was all nonsense. The British philosopher Peter Geach (1916–2013), for example, declared the Minkowskian view to be “very popular with philosophers who try to understand physics and physicists who try to do philosophy.” See P. T. Geach, “Some Problems About Time,” in *Studies in the Philosophy of Thought and Action* (P. F. Strawson, editor), Oxford University Press, 1968. In his introduction to Geach’s essay, editor Strawson put in his two cents by stating the four-dimensional view of reality to be nothing but “fanciful philosophical theorizing.”

The block universe concept may explain the enigmatic statement made by Einstein at the death of Michele Besso (note 103). As interpreted[109] decades later:

> "It seems that Einstein's view of the life of an individual was as follows. If the difference between past, present, and future is an illusion, i.e., the four-dimensional spacetime is a 'block Universe' without motion or change, then each individual is a collection of myriad of selves, distributed along his history, each occurrence *persisting on the world line, experiencing indefinitely the particular event of that moment* [my emphasis]. Each of these momentary persons, according to our experience would possess memory of the previous ones, and would therefore believe himself identical with them; yet they would all exist separately, as single pictures in a film. Placing the past, present and future on the same footing this way, destroys the notion of the unity of the self, rendering it a mere illusion as well."

It appears by his words that Einstein was indeed in agreement with the block universe concept, and that he was attempting to give his friend's family some reason to believe that their father still lives 'somewhen.' The makers of the 2002 film *Minority Report* made use of the block universe concept, even if not intentionally; there we see police stopping crime *before* it happens because they can 'see the future.'

Not everybody believed that this view of spacetime was Einstein's, however. Karl Popper (1902–1994), an Austrian philosopher of science, wrote 28 years after the scientist's death that "Einstein was a strict determinist when I first visited him in 1950: he believed in a 4-dimensional Block-Universe. But he gave this up."[110] Shortly before he wrote those words, however, Popper must have learned something new to convince himself of his final comment, because just 2 years earlier he had declared[111] Einstein to (still) be a determinist. Popper presents no evidence to support his claim of Einstein's philosophical conversion, however, and it would seem that the Besso letter still offers the best insight into his actual view of spacetime shortly before his death. I say this because I think Popper's labeling of Einstein as a determinist is wrong. Determinism says 'If you do A, then B will happen, and if you do not do A then (perhaps) something other than B will happen.' A deterministic universe has plenty of room for free will, because you can *choose* to do A or not to do A, and what you decide makes a difference. A fatalistic universe, however, as is the block universe, simply says 'You will do A and B will happen.' To accept the block universe, as did Einstein, is to be a fatalist, not a determinist.

---

[109] L. P. Horwitz, R. I. Arshansky, and A. C. Elitzur, "On the Two Aspects of Time: The Distinction and Its Implications," *Foundations of Physics*, December 1988, pp. 1159–1193. See also Einstein's own book (note 67) where he wrote "From a 'happening' in three-dimensional space, physics becomes, as it were, an 'existence' in the four-dimensional 'world'."

[110] See the *Seventh International Congress of Logic, Methodology and Philosophy of Science*, volume 4 (Salzburg, Austria, 1983), p. 176. Popper describes his early discussions with Einstein on the reality of time and the four-dimensional Parmenidean block universe in some detail in his autobiography: see volume 1 of *The Philosophy of Karl Popper* (P. A. Schilpp, editor), The Library of Living Philosophers, Open Court 1974, pp. 102–103.

[111] In the Foreword to the book by B. Gal-Or, *Cosmology, Physics and Philosophy*, Springer-Verlag 1981.

Einstein's final position on this, then, *might* have been like that of the fictional time traveler who takes a little girl 25,000 years back into the past, where she sees an ancient ancestor of humanity.[112] She then asks if the ancestor is really alive. The time traveler replies, "Every man who ever lived is still alive, child. In time there is no real death. When a man dies he's still alive 10 min ago, 10 years ago. He's always alive to those who travel back through time to meet him face to face."

Did Einstein *really* believe this? Not everybody thinks so. At the 1922 meeting of the French Philosophical Society, for example, the philosopher of science Emile Meyerson asked Einstein whether the spatialization of time (the idea that time is a dimension on the same footing as the spatial ones) is a legitimate interpretation of Minkowski's spacetime. Einstein's terse answer was that "it is certain that in the four-dimensional continuum all dimensions are *not* [my emphasis] equivalent."[113]

Use of the term *block universe* is generally thought to have originated with the Oxford philosopher Francis Herbert Bradley (1846–1924) who, in his 1883 book *Principles of Logic*, wrote "We seem to think that we sit in a boat, and are carried down the stream of time, and that on the bank there is a row of houses with numbers on the doors. And we get out of the boat, and knock at the door with number 19, and, re-entering the boat, then suddenly find ourselves opposite 20, and having then done the same, we go on to 21. And, all this while, the firm fixed row of the past and future stretches in a *block* [my emphasis] behind us, and before us." The house numbers would seem to be Bradley's way of referring to the centuries. Note that he wrote these words 12 years before *The Time Machine*, and that they preceded Minkowski's famous address by a quarter-century.

But this origin of *block universe* may not be as clear-cut as I have made it appear. Bradley, who was frequently criticized by the Harvard psychologist William James (1842–1910)—a man who argued for free will[114] and indeterminism, concepts disallowed in a block universe—may have been mocked on the idea by James during an address to the students of the Harvard Divinity School in March 1884 ("The Dilemma of Determinism"), the year after Bradley's book had been published. In his address James spoke of a deterministic world as being a "solid" or "iron block" (this are *not* characteristics of determinism, but rather of fatalism, and so James makes the same mistake as did Popper). However, writing the year before Bradley's book, in the April 1882 issue of *Mind*, James wrote (with obvious disdain) of "the universe of Hegel [the German philosopher Georg Hegel (1770-1831)]—the *absolute block* [my emphasis] whose parts have no loose play," as having "the oxygen of possibility all suffocated out of its lungs" and as being a universe in which "there can be neither good nor bad, but [only] one dead level of mere fate." So, perhaps, the chain of evolution of the term block universe is actually from Hegel to James and *then*, finally, to Bradley.

---

[112] F. B. Long, "Throwback in Time," *Science Fiction Plus*, April 1953.

[113] A. Einstein, "La Théorie de la Relativité," *Bullentin de la Société Francaise de Philosophia* (volume 17), 1922, pp. 91–113.

[114] A famous line from James, one that perhaps illustrates his sort of reasoning about free will, is "My first act of free will shall be to believe in free will." If only proving theorems in math and physics were that easy.

We can actually find the block universe in fiction *before* Minkowski (and so certainly before pulp science fiction) came on the scene. In an 1875 (!) story[115] we read of a man who sees, years in advance, his own death in the American Civil War. In the following extract, this man speaks to an unnamed friend (who is the narrator):

> "Do you know," said Bernard, presently, "I sometimes think prophecy isn't so strange a thing . . . I really see no reason why any earnest man may not be able to foresee the future, now and then . . ."
>
> "There is reason enough to my mind," I replied, "in the fact that future events do not exist, as yet, and we cannot know that which is not, though we may shrewdly guess it sometimes . . ."
>
> "Your argument is good, but your premises are bad, I think," replied my friend, . . . his great, sad eyes looking solemnly into mine.
>
> "How so?" I asked.
>
> "Why, I doubt the truth of your assumption, that future events do not exist as yet . . . Past and future are only divisions of time, and do not belong to eternity . . . To us it must be past or future with reference to other occurrences. But is there, in reality, any such thing as a past or a future? If there is an eternity, it is and always has been and always must be. But time is a mere delusion . . . To a being thus in eternity, all things are, and must be present. *All things that have been, or shall be, are* [my emphasis]."

When the block universe concept did eventually appear in science fiction, it did so early. In a 1927 story, for example, a time traveler from the future and a man in the present (who is the narrator) have the following exchange:

> "I have just been five years into your future."
> "My future!" I exclaimed. "How can that be when I have not lived it yet?"
> "But of course you have lived it."
> I stared, bewildered.
> "Could I visit my past if you had not lived your future?"[116]

So, while the block universe has a bit of a history to it, the history of the concept of *mathematical* spacetime in physics has a much clearer origin: it derives from Minkowski, not from Hegel, Bradley, James, or even Einstein (who often gets credit for it even though he didn't use the concept in special relativity in 1905, 3 years before Minkowski's address.). Eventually, of course, Einstein did come to appreciate the power and conceptual beauty of four-dimensional spacetime, and it came to play a central role in his ideas about gravity. Indeed, in Einstein's general theory of relativity gravity *is* (curved) spacetime. The starting point for general relativity (and so a *scientifically plausible* theory of time travel) was Minkowski's creation of spacetime, and he is truly deserving of the title 'father of the fourth dimension.'

---

[115] G. C. Eggleston, "The True Story of Bernard Poland's Prophecy," *American Homes*, June 1875. George Cary Eggleston (1839–1911) had served as a soldier in the Confederate Army.

[116] F. Flagg, "The Machine Man of Ardathia," *Amazing Stories*, November 1927.

Of course, it is true that Newton's physics also talks about an analytical (as opposed to merely philosophical) space and time long before either Minkowski or Einstein, but 'Newtonian spacetime' is something very different from Minkowski's self-described "radical" view.[117] In the Newtonian view there is a universal time, a *cosmic* time, which is the same time for everyone, everywhere, in the universe. At every instant, a cosmic simultaneity exists for Newton. Newton's space is Euclidean; that is, through any point exterior to a line exactly one parallel line can be constructed and those two lines will never meet, all triangles (no matter their size) have an interior angle sum of 180°, and so on. For Newton, space and time were absolutely and uniquely separable. They were, as philosophers like to say, "distinct individuals."

Minkowski changed all that. For him space and time are only relatively separable, and the separation is different for observers in relative motion. For Newton, space and time are the *background* in which physical processes in the world evolve. For Minkowski, spacetime *is* the world.

In a famous philosophical paper[118] by an advocate of the block universe view of reality, we find the words "I . . . defend the view of the world . . . which treats the totality of being, of facts, or of events as spread out eternally in the dimension of time as well as the dimensions of space. Future events and past events are by no means present events, but in a clear and important sense they do exist, now and forever, as rounded and definite articles in the world's furniture." The title of Williams' paper comes from an ancient dilemma stated by Aristotle in his *De Interpretatione*, where he asked a question now classic in philosophy: "Will there be a sea fight tomorrow?"

Aristotle began his famous answer by first posing the following premise: If a statement about some future event is, eventually, shown to be true (or false), then that statement was true (or false) from the moment it was made. Consider, then, the following two assertions: (A) "It is true that there will be a sea fight tomorrow" and (B) "It is true that there will *not* be a sea fight tomorrow." Surely, argued Aristotle, (A) and (B) cannot both be true, but equally surely, one of them must be true. Suppose it is (A) that is true. Then there is nothing that can be done to prevent the sea fight, and so the future is fated. Suppose, however, it is (B) that is true. Then there is nothing that can be done to cause the sea fight, and so the future is fated. The conclusion is the same no matter which assertion is the true one; thus, the future is fated.

---

[117] See, for example, H. Stein, "Newtonian Space-Time," *Texas Quarterly*, Autumn 1967, pp. 174–200; G. Berger, "Elementary Causal Structures in Newtonian and Minkowskian Space-Time," *Theoria* (volume 40), 1974, pp. 191–201; J. Earman and M. Friedman, "The Meaning and Status of Newton's Laws of Inertia and the Nature of Gravitational Forces," *Philosophy of Science*, September 1973, pp. 329–359.

[118] D. C. Williams, "The Sea Fight Tomorrow," in *Structure, Method and Meaning*, The Liberal Arts Press 1951. Donald Williams (1899–1983) was a professor of philosophy at Harvard.

As might be expected, those who like the fatalistic block universe like this conclusion, but, ironically, Aristotle wasn't one of them—he disliked it so much that he struggled to find a way around it. On the other hand, there are philosophers, like Professor Williams (who believed in a fatalistic universe), who reject Aristotle's rejection of his own logic! Professor Williams went so far, in fact, to calling Aristotle's reasoning "a tissue of error" and "swaggeringly invalid." Possibly so, but the philosophical debates over the sea fight question, and the fatalistic (or not) nature of the world, have not ceased to this day.

In an even more famous paper, Professor Williams makes clear his belief that the passage of time is a myth; he poetically declared "the total of world history is a spatio-temporal volume, of somewhat uncertain magnitude, chockablock with things and events."[119] Professor Williams did, indeed, embrace four-dimensional spacetime, and this is demonstrated by the following incredible passage, perhaps his best-remembered words: "It is then conceivable, though doubtless physically impossible, that one four-dimensional area of the time part of the manifold be slewed around at right angles to the rest, so that the time order of that area, as composed by its interior lines of strain and structure, run parallel with a spatial order in its environment. It is conceivable, indeed, that a single whole human life should lie thwartwise of the manifold, with its belly plump in time, its birth at the east and its death in the west, and its conscious stream running alongside somebody's garden path."

Good Lord!

Now, I am willing to admit that Professor Williams probably wrote that wonderful passage mostly for effect,[120] but I ask you—what, if anything, does it *mean*? It is marvelous to read and yet it remains (for me) mysterious.[121] It should come as no surprise that Professor Williams originally presented his papers to the Metaphysical Society of America, rather than to the American Physical Society. But this passage was perhaps not without impact in areas far removed from metaphysics; some years later there appeared a science fiction story[122] that reads as though it had been inspired by Williams. In it, a scientist discovers how to bend his perception of the four dimensions so as to view verticality as duration and duration as verticality. Thus, he is in October while sitting, but when he stands up he is in November! As bizarre as this may seem, such coordinate interchanges actually do occur in the mathematical theory of time machines; we'll see this later, for example, when we discuss Tipler's rotating cylinder time machine.

---

[119] D. C. Williams, "The Myth of Passage," *Journal of Philosophy*, July 1951, pp. 457–472.

[120] In a footnote, Williams sort of admits this when he writes "I should expect the impact of the environment on such a being to be so wildly queer and out of step with the way he is put together, that his mental life must be a dragged-out monstrous delirium." I think this a great understatement.

[121] As it was for some of Williams' fellow philosophers, one of whom bluntly called the 'myth-of-passage' paper "an interesting piece of science fiction": see M. Capek, "The Myth of Frozen Passage: The Status of Becoming in the Physical World," in *Boston Studies in the Philosophy of Science* (volume 2), Humanities Press 1965. Capek's title reflects his view of the block universe as simply a giant refrigerator and so, turning the tables on Williams, we have 'passage' changed to 'frozen passage.' See also note 136.

[122] G. Wolfe, "The Rubber Bend," *Universe 5* (T. Carr, editor), Random House 1974.

By the 1930s the block universe had found a home in pulp science fiction. The block universe view that past and present coexist with the present got dramatic treatment in one story of a high school teacher who invents a "spacetime warp" theory, and who is then tricked by an evil industrialist into implementing it in the form of a gun. The weapon produces incredible effects when it is tested; for example, an allosaurus appears, which we are told is "a carnivorous dinosaur of the Jurassic Age, the most frightful engine of destruction that ever walked the Earth!"[123] At the story's end, the teacher explains what has happened to a crowd of breathless newspaper reporters:

> "Spacetime was warped slightly … The Einsteinian spacetime continuum buckled … Because it was superficial, only a little of the past, a little of the future broke through. The folds of the warp distorted spacetime evanescently, erratically skirting the vast gulf where the past lies buried and lightly tapping the vast stores of the future. It is a truism of modern speculative physics that the past and the future exist simultaneously and coextensively in higher dimensions of space. De Sitter has speculated as to the possibility of seeing an event before it happens. It is quite possible, gentlemen. Events of the far future already exist in spacetime."

That 'explains' the dinosaur. In the teacher's words, "You tell me that two men saw an incredible beast. … They swear it looked like a dinosaur. I think it was a dinosaur, gentlemen. It broke through when the warp tapped the past."

And just 2 years later, Robert Heinlein made world lines the central concept in the first of his many classic tales.[124] The story draws an analogy between a world line and a telephone cable: the beginning and end points in spacetime for the world line of a person (birth and death) are associated with breaks (faults) in a telephone cable. By sending a signal up and down the cable, and measuring the time delay until the arrival of the echo produced by such discontinuities, a technician can both detect and locate the faults. In the same manner, Heinlein's story-gadget sends a signal of unspecified nature up and down a world line and thus locates the birth and death 'discontinuities.' Knowledge of the death date, in particular, causes financial stress among life insurance companies, and an examination of *that* tension (not strange physics) is the fictional point of the story.

And then, 2 years after Heinlein's tale with its serious tone, a far less serious story[125] (featuring an Attila the Hun character who roams up and down the corridors of time kidnapping beautiful women for his harem!), we find an 'editorial' footnote telling its young readers that "scientists—especially the new order of meta-physical scientists—are agreed on the principles of Space-Time. The future is not a thing which *will exist*. Rather it is a thing which *does exist*—all events from the Beginning to the End, exist in a record upon the scroll of Time." This story, itself, was silly, but the block universe metaphysics was up-to-date.

---

[123] F. B. Long, "Temporary Warp," *Astounding Stories*, August 1937.

[124] R. Heinlein, "Life-Line," *Astounding Science Fiction*, August 1939.

[125] R. Cummings, "Bandits of Time," *Amazing Stories*, December 1941.

Somewhat surprisingly, I think, is that even before pulp science fiction embraced the block universe, the concept had already made a deep impression on a broader audience. For example, in a 1928 New York stage play[126] the action alternately takes place in the years 1784 and 1928 and, to explain how that can be, one character (a time traveler) tells another:

> "Suppose you are in a boat, sailing down a winding stream. You watch the banks as they pass you. You went by a grove of maple trees, upstream. But you can't see them now, so you saw them in the *past*, didn't you? You're watching a field of clover now; it's before your eyes at this moment, in the *present*. But you don't know yet what's around the bend in the stream ahead of you; there may be wonderful things, but you can't see them until you get around the bend, in the *future*, can you?"

Then, after this prologue about the stream of time, comes the block universe idea:

> "Now remember, *you're* in the boat. But *I'm* up in the sky above you; in a plane. I'm looking down on it all. I can see *all at once* the trees you saw upstream, the field of clover that you see now, and what's waiting for you around the bend ahead! *All at once!* So the past, present, and future of the man in the boat are all *one* to the man in the plane."

Then, finally, the obvious theological conclusion: "Doesn't that show how all Time must really be one? Real Time—real Time is nothing but an idea in the mind of God!"

To end this section, the block universe conception was cleverly used by one science fiction fan who argued in support of time travel, in reply to another fan how had claimed that a failure of mass/energy conservation was fatal to the plausibility of time travel. Their exchange began with a letter to the editor at *Astounding Stories* in November 1937, written in response to a recent story[127]:

> "Let us say that there is, at a certain time, 'x' amount of matter in the Universe, and 'e' amount of energy. Then if a man of 'a' mass travels backward in time to this particular instant aforementioned, the total amount of matter is thus 'x' plus 'a', while if no other such mass changing occurrences take place, the amount of matter in the future is 'x' minus 'a'. Only a corresponding loss and gain respectively in the amount of energy could explain this conservation of energy, advocates [of time travel] say what they may. But you can't rob or add energy to a Universe nilly-willy! Or perhaps time doesn't enter in on the matter. Perhaps you can add matter in a Universe provided you take it away on some future date."

This fan's concern clearly made an impression on science fiction writers, and the case for conservation of energy is stated in many of the time travel stories that appeared after the publication of this letter.[128]

---

[126] "Berkeley Square" by J. L. Balderson. This play was made into a 1933 movie of the same name, and again in 1951 as the film *I'll Never Forget You*.

[127] O. Saari, "The Time Bender," *Astounding Stories*, August 1937 (see also note 137 in Chap. 1).

[128] Examples include the novels *Lest Darkness Fall* (Henry Holt 1941) by L. Sprague de Camp, and *The Time Hoppers* (Doubleday 1967) by Robert Silverberg.

A reply was soon received by the magazine in a letter (January 1938) from another fan:

> "[A recent letter] implies that the idea of time travel is incompatible with the law of conservation of mass and energy. I believe [the] reasoning is wrong [and that the] difficulty lies primarily in the assumption that a body moved in time is transported into a different Universe. According to Einstein, time and the three normal dimensions are so related as to form a continuous, inseparable medium we call the spacetime continuum. Time is in no way independent of the other components of our Universe. Hence a fixed mass [a time traveler and his machine] moved in time is by no means lost from the Universe, the action being analogous to a shift along any other dimension."

The block, or frozen, universe of Minkowski is clearly reflected in those words.[129]

2.7 Philosophical Implications of the Block Universe

> "Is the future all settled beforehand, and only waiting to be 'pushed through' into our three-dimensional ken? Is there no element of contingency? No free will? I am talking geometry, not theology."[130]

I should tell you now that, despite the enthusiastic embrace of the block universe by Williams and others (including Einstein), there are those who have been harsh in their criticism of Minkowski's spacetime. The major philosophical problem with the block universe interpretation of four-dimensional spacetime is that it looks like fatalism disguised as physics. It seems to be little more than a mathematician's proof of a denial of free will dressed up in geometry. One philosopher illuminated this concern with the following story, one that vividly illustrates the compelling need many humans have to deny a fatalistic world:

> "In a moving picture version of *Romeo and Juliet*, the dramatic scene was shown in which Juliet, seemingly dead, is lying in the tomb, and Romeo, believing she is dead, raises a cup containing poison. At this moment an outcry from the audience was heard: 'Don't do it!' We laugh at the person who . . . forgets that the time flow of a movie is unreal, is merely the unwinding of a pattern imprinted on a strip of film. Are we more intelligent than this man when we believe that the time flow of our actual life is different? Is the present more than our cognizance of a predetermined pattern of events unfolding itself like an unwinding film?"[131]

---

[129] In the context of mathematical physics (*not* science fiction) it has been shown that time travel does *not* imply any fatal violation of conservation of energy. See, for example, J. L. Friedman *et al.*, "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review D*, September 15, 1990, pp. 1915–1930, and D. Deutsch, "Quantum Mechanics Near Closed Timelike Lines," *Physical Review D*, November 15, 1991, pp. 3197–3217.

[130] The lament of Victorian physicist Oliver Lodge (1850–1940) in his essay "The New World of Space and Time," *Living Age*, January 1920.

[131] H. Reichenbach, *The Direction of Time*, University of California Press 1956, p. 11.

Most people in the Western world would answer *yes* to Reichenbach's question. Most people do find Omar Khayyam's *Rubaiyat* to be a beautiful poem, yes, but still they reject its fatalistic message: "And the first Morning of Creation wrote/What the Last Dawn of Reckoning shall read." Indeed, William James quoted these very words in his 1884 address to the students of the Harvard Divinity School when he argued against fatalism and the block universe.

Besides fatalism, another reason for the stinging words by critics of Minkowski's spacetime is that, in it, events don't *happen*—they just *are*. That is, there seems to be no temporal process of *becoming* in Minkowski's spacetime. Everything is already there and, as what we perceive to be the passing of time occurs, we simply become conscious of ever more of Minkowski's "world points," or events, that lie on our individual world lines. Hermann Weyl (1885–1955), a German mathematical physicist who in his last years was a colleague of Einstein and Gödel at the Institute for Advanced Study in Princeton, expressed this very interpretation in words that have become famous, words that sound very much like those of Wells' Time Traveller: "The objective world simply *is*, it does not *happen*. Only to the gaze of my consciousness, crawling upward along the life line of my body [Minkowski's world line], does a section of the world [spacetime] come to life as a fleeting image in space which continuously changes in time [creating what we call the *now* or the *present*]."[132]

Weyl was skillful in finding poetic ways to express the world line view of reality, but not everybody is convinced by the poetry because it seems to deny the common sense idea of time 'flowing,' of temporal passage; it effectively says time is mind-dependent, a mere *illusion*, as the time traveler in "Berkeley Square" declared (note 126). One philosopher who was particularly opposed to Weyl's view was the British-American academic Max Black (1909–1989), and he expressed his opinion in no uncertain terms: "The picture of a 'block Universe,' composed of a timeless web of 'world-lines' in four-dimensional space, however strongly suggested by the theory of relativity, is a piece of gratuitous metaphysics."[133] Another philosopher who was unhappy with Weyl's view of the block universe was just as blunt: "While philosophers may be forgiven intellectual extravagances of this kind, I think it is a pity when they receive encouragement from theoretical physicists."[134]

Weyl's views had supporters, too, however. Consider, for example, the Time Traveller's speech to his friends at the fateful dinner party that opens *The Time*

---

[132] H. Weyl, *Philosophy of Mathematics and Natural Science*, Princeton University Press 1949, p. 116. Sir James Jeans had already said the same, somewhat less elegantly, in his 1935 Sir Halley Stewart Lecture: "The tapestry of spacetime is already woven throughout its full extent, both in space and time, so that the whole picture exists, although we only become conscious of it bit by bit—like separate flies crawling over a tapestry . . . A human life is reduced to a mere thread in the tapestry." Jeans then immediately *rejected* this fatalistic view: see his *Scientific Progress*, Macmillan 1936, p. 20.

[133] From a book review in *Scientific American*, April 1962, pp. 179–185.

[134] H. A. C. Dobbs, "The 'Present' in Physics," *British Journal for the Philosophy of Science*, February 1969, pp. 317–324.

Machine: "There is no difference between Time and any of the three dimensions of Space except that our consciousness moves along it . . . here is a portrait of a man at 8 years old, another at fifteen, another at seventeen, another at twenty-three, and so on. All these are evidently sections, as it were, Three-Dimensional representations of his Four-Dimensional being, *which is a fixed and unalterable thing* [my emphasis]." Remember, these words were written in 1895, 13 years before Minkowski and his world lines, and of course decades before Weyl's famous words.

Wells' passage made a considerable impression on at least one well-known physicist of the time, who references it in his early book on relativity.[135] And in another book on relativity, published the same year, we find the same interpretation of Minkowski's spacetime as a block universe: "With Minkowski, space and time become particular aspects of a single four-dimensional continuum . . . All motional phenomena . . . become timeless phenomena in four-dimensional space. The whole history of a physical system is laid out as a changeless whole."[136]

The claim that time is an illusion has some thought-provoking implications concerning the concepts of omniscience and free will, concepts that occur in any discussion of time travel. Some old theology on God's omniscience, as discussed in Aquinas' *Summa Theologiae*, is seemingly lent at least some support by Minkowski's spacetime: "Now although contingent events come into actual existence successively, God does not, as we do, know them in their actual existence successively, but all at once; because his knowledge is measured by eternity, as is also his existence; and eternity which exists as a simultaneous whole, takes in the whole of time . . . Hence all that takes place in time is eternally present to God." Somewhat paradoxically, however, Aquinas did make a distinction between past and future. In that same work he declares that "God can cause an angel not to exist in the future, even if he cannot cause it not to exist while it exists, or not to have existed when it already has." For Aquinas, then, whereas the past is rigid and unchangeable, the future is plastic, which is *not* the block universe view of spacetime.

As one theologian has observed,[137] this does not mean that Aquinas thought God had to view all events simultaneous with all others.[138] Rather, our theologian says that Aquinas could have thought of the relationship between God and events as being similar to that between the center of a circle and all the points on the circumference. That is, each point on the circumference has its own identity, coming before and/or after any other point, but the center is related to each and every point on the circumference in precisely the same way. The center, then, is 'eternity' and the circumference is the temporal series ('one thing after another') of reality. Saying that God is *eternal* is thus very different from saying he is *everlasting*. The first means outside of time, whereas the second means he is a temporal entity but has neither beginning nor end.

---

[135] L. Silberstein, *The Theory of Relativity*, Macmillan 1914, p. 134.

[136] E. Cunningham, *The Principle of Relativity*, Cambridge University Press 1914, p. 191. The use of the words *timeless* and *changeless* explain the characterization of the block universe as being *frozen* (in note 121).

[137] W. L. Craig, "Was Thomas Aquinas a B-Theorist of Time?" *New Scholasticism*, Autumn 1985, pp. 475–483. For the B-theory of time, look back at the discussion in the first section of this chapter.

[138] A science fiction story by Norman Spinrad, "The Weed of Time" (*Alchemy and Academe*, Doubleday 1970) graphically describes what a nightmare that could be!

Our theologian supports the first interpretation, invoking Aquinas' own words from *Summa Contra Gentiles*: "The divine intellect, therefore, sees in the whole of its eternity, as being present to it, whatever takes place through the whole course of time. And yet what takes place in a certain part of time was not always existent. It remains, therefore, that God has a knowledge of these things that according to the march of time do not yet exist."

The issue of God's eternity and his place in spacetime has long been a hot topic among theologians with a scientific inclination. Practically every issue of the learned journal *Religious Studies*, for example, carries an article on the subject, often invoking relativity theory to support some argument. The Bible, itself, can be a confusing guide on this matter. For example, consider the Old Testament story of King Ahab (First Kings 21). Ahab, King of Sumeria, coveted Naboth's vineyard, but Naboth would not sell. The King retreated, but his wife Jezebel arranged for Naboth's downfall and judicial murder and thus caused the arrival of all his property into her husband's hands. This angered God, who commanded Elijah to prophesy disaster on Ahab's house. Ahab responded with sackcloth, and at that God shifted the disaster to the house of Ahab's son. The point, here, is that God, declared to be omniscient, seems to have been *surprised* at Ahab's penitence. God is aware of everything in this tale, but only as it happens. That is, God's knowledge is subject to growth. This Hebrew concept of God as a participant in history is at odds with the contemporary Christian conception of divine knowledge of all that has been, all that is, and all that will be, a view which has its own Biblical support (for divine eternality). For example, Malachi 3:61 ("For I am the Lord, I change not"), and James 1:17 ("the Father ... with whom is no variableness").

When *The Time Machine* was serialized in the *New Review*, it included a passage that does not appear in the now classic version of the story in which the Time Traveller explains his view of the connection between omniscience and the block universe to his dinner guests:

> "I'm sorry to drag in predestination and free-will, but I'm afraid those ideas will have to help ... Suppose you knew fully the position and properties of every particle of matter, of everything existing in the Universe at any particular moment of time: suppose, that is, that you were omniscient. Well, that knowledge would involve the knowledge of the condition of things at the previous moment, and at the moment before that, and so on. If you knew and perceived the present perfectly, you would perceive therein the whole of the past. If you understood all the natural laws the present would be a complete and vivid record of the past. Similarly, if you grasped the whole of the present, knew all its tendencies and laws, you would see clearly all the future. To an omniscient observer there would be no forgotten past—no piece of time as it were that had dropped out of existence—and no blank future of things yet to be revealed ... Present and past and future would be without meaning to such an observer ... He would see, as it were, a Rigid Universe filling space and time ... If 'past' meant anything, it would mean looking in a certain direction, while 'future' meant looking the opposite way."

Wells' "Rigid Universe" certainly sounds like the block universe, and he (or least, the Time Traveller) seems to have believed that it held important implications for the concept of free will.

The 'Rigid Universe' got an interesting science fiction treatment in a story[139] that imagined an event in the present that occurs 'before it should' (a heart patient learns that her obituary notice will be in next week's *New York Times* when that paper arrives 'early'). As one character explains to the sister of the lady who is soon to die, "The future mustn't be changed . . . For us the events of . . . the future are as permanent as any event in the past. We don't dare play around with changing the future, not when it's already signed, sealed and delivered in that newspaper. For all we know the future's like a house of cards. If we pull one card out, say your sister's life, we might bring the whole house tumbling down. You've got to accept the decree of fate . . . You've got to."

With Einstein's discovery of the *relativity of simultaneity*,[140] we run into the question of 'How can there be any sense to the concept of divine, universe-wide knowledge in a four-dimensional spacetime?' That's because in some frames of reference it is possible for event A to be observed before event B, whereas in other frames the temporal order could be reversed, and so some theological questions prompted by spacetime physics are: 'What is God's frame of reference if he is to be actively involved in human affairs? Could God have a special frame of reference in which he is exempt from the relativity of simultaneity, a frame in which he imposes an absolute order on the sequence of becoming of events? Does it make any sense, that is, to say God enjoys what might be called 'divine immediacy'? And if so, what should we think of a God who follows rules of nature different from those that govern all he is supposed to have made?'

Theologians have debated questions like these for decades, and surely will continue to do so for many more decades to come. Alas, I suspect that physicists who study time travel have either been unaware, unimpressed, or just plain uninterested. That's too bad, because one doesn't have to be religious to appreciate the pure intellectual challenges presented by such questions. For example, consider the following debate between two philosophers, one who believes free will and divine foreknowledge are not compatible, and another who thinks the first has made a fundamental error in blurring the distinction between changing and affecting the past. (This distinction is of *great* importance in any discussion of time travel.) This second philosopher presented some of his arguments in terms of a time traveler to the past[141]:

---

[139] R. Silverberg, "What We Learned From This Morning's Newspaper," *Infinity Four*, November 1972.

[140] This refers to the discovery that two events, which occur simultaneously for one observer in a spacetime, may not be simultaneous for another observer in the same spacetime. This will be discussed in more detail in the next chapter.

> "Consider the following. Parsons (P) has invented a special machine which allows him to go back in time. He enters the machine in 1986 and finds himself in the presence of or, perhaps better, observing, Quigly (Q) in 1876. P is an authority on Q, and knows immediately the situation Q is in. Not only that, but he remembers reading about the particular decision or act which Q made in that situation. Thus one might argue that from P's perspective what Q decides is as if already done. It is not already done, since P is standing there waiting for Q to do it. He has gone back in time. Yet from P's perspective, which is of one come back from the future, it is as if already done, since he knows what Q does decide. Since P strongly believes in the unalterability of the past, it is not within Q's power to do something other than what Q in fact does in that situation. From Q's perspective his decision is not already made nor is the action taken, so that it is in his power at that time to do either x or y. From his perspective, that he will do x rather than y is indeterminate; it is not yet done, though at the same time he can grant that P knows what he will do because for him it is as if he has already done it."

The first philosopher doesn't buy any of this, and dismisses it with "It should be abundantly clear . . . that the fact that such stories are in some way imaginable and intuitively graspable says nothing about their logical coherence." Given the interest among modern physicists in time travel, however, I think the first philosopher wouldn't write that today.

One possible reply to all of these theological issues that spacetime physics prompts can perhaps be found in a paper[142] (written by a philosopher and two mathematicians) that describes a five-dimensional spacetime in which the fifth dimension is initially given the provocative label of the 'eternity' axis. But then the authors lost their nerve and elected to rename it 'anti-time.' It is interesting to note that pulp science fiction anticipated that terminology by decades, as in one story[143] we read "Beyond the fourth there is a fifth dimension . . . Eternity, I think you would call it. It is the line, the direction perpendicular to time." For some, the eternity axis would appear to be perfect to serve as the temporal dimension for God, an axis distinct from the time axis of mere mortals.

The idea of supernatural beings existing outside of mortal time is an old one in theology, and it can also be found in secular literature long before science fiction got hold of it. For example, in the first act of Lord Byron's 1821 poem *Cain*, the fallen angel Lucifer tells Cain and his wife that

---

[141] For the complete exchange between these two philosophers, see W. Hasker, "Foreknowledge and Necessity," April 1985, pp. 121–157, B. Reichenbach, "Hasker and Omniscience," January 1987, pp. 86–92, and W. Hasker, "The Hardness of the Past: A Reply to Reichenbach," July 1987, pp. 337–342, all in the journal *Faith and Philosophy*. Hasker is the 'first' philosopher, and Reichenbach is the 'second' one. See also D. P. Lackey, "A New Disproof of the Compatibility of Foreknowledge and Free Choice," *Religious Studies*, September 1974, pp. 313–318.

[142] J. G. Bennett *et al.*, "Unified Field Theory in a Curvature-Free Five-Dimensional manifold," *Proceedings of the Royal Society of London A*, July 1949, pp. 39–61. A theological interpretation is given in G. Stromberg, "Space, Time, and Eternity," *Journal of the Franklin Institute*, August 1961, pp. 134–144.

[143] L. A. Eshbach, "The Time Conqueror," *Wonder Stories*, July 1932.

With us acts are exempt from time, and we
Can crowd eternity into an hour,
Or stretch an hour into eternity.
We breathe not by a mortal measurement,
But that's a myst'ry.

Before Minkowski, the debates over fatalism (as in Silverberg's story in note 139) and free will had been the exclusive province of philosophers, theologians, and lawyers (if a person has no control over his or her actions, then can we morally and ethically punish that person if those actions happen to be criminal?[144]). After Minkowski, the physicists (at least a few of them) joined the debates. According to one philosopher (note 118) the major motivation driving these debates is "the age-old dread that God's foreknowledge of our destiny can in itself impose the destiny upon us." The implication is, of course, that God is 'outside of time' and so can take in the entire Minkowskian block universe at a glance (hence his foreknowledge).

The relativistic view of the universe as a timeless four-dimensional spacetime seems to provide scientific, mathematical support for the conclusion that not only is the past fixed, but so is the future. Does that mean the future is what it will be—and if so, then why bother agonizing over the many apparent decisions each of us faces every day? If the future will be what it will be, then Christian theologians are left with the puzzling task of explaining what could possibly be meant by the Biblical exhortation (Deuteronomy 30:19) "I call Heaven and Earth to record this day against you, that I have set before you life and death, blessing and cursing; therefore *choose* [my emphasis] life, that both thou and thy seed may live."

This issue has bothered philosophers for a very long time. The so-called Master Argument (the name reflects its supposed invulnerability to rebuttal), for example, comes down to us from its origins in ancient times, in the *Discourses* of the first century A.D. Roman Stoic philosopher Epictetus. That argument can be summarized[145] as follows:

1. The future follows from the past;
2. The past is unchangeable;
3. What follows from the unchangeable is unchangeable;

Therefore,

4. The future is unchangeable.

This certainly does seem to be fatalistic, in effect arguing that all events in a block universe spacetime are recorded in a 'Book of Destiny.' Since ancient times many great works of literature have adopted that view, recounting tales of the foretold fates of men, such as Sophocles' *Oedipus*. It is, in a block universe, as though our conscious experience of the world is no different from that of the man watching the projected film images of *Romeo and Juliet*.

---

[144] For more on this, in the context of time travel, see the penultimate question in the *For Future Discussion* questions at the end of this chapter.

[145] See, for example, the two papers by R. L. Purtill, "The Master Argument," *Apeiron*, May 1973, pp. 31–36, and "Foreknowledge and Fatalism," *Religious Studies*, September 1974, pp. 319–324.

That view is the central issue in the early sixth century A.D. Roman philosopher Boethius' influential *De Consolatione Philosophiae* (circa A.D. 500) which was written during a year of imprisonment before his execution for treason; perhaps he wondered during that year if his fate could have been anything different. Certainly he must have taken some consolation in fatalism, but in fact he tried to argue that God's vision of *all* temporal reality does not limit the freedom to act. According to Boethius, "The expression 'God is ever' denotes a single Present, summing up His continual presence in all the past, in all the present . . . and in all the future." That is, God sees in one timeless and eternal moment all that has been and will be freely chosen.[146]

When the fourteenth century English poet Geoffrey Chaucer prepared a translation of *Consolatione* he was obviously inspired by it when he wrote his very long, famous poem (*Troilus and Criseyde*) on the nature of love (Book IV.140):

Some say "If God sees everything before
It happens—and deceived He cannot be—
Then everything must happen, though you swore
The contrary, for He has seen it, He."
And so I say, if from eternity
God has foreknowledge of our thoughts and deed,
We've no free choice, whatever books we read.

Two modern, purely philosophical rebuttals[147] to Chaucer, however, argue that his poetry misstates Boethius' philosophy when Troilus declares that divine foreknowledge is incompatible with free will. That is, in their view God's omniscience (a fundamental teaching in the theistic religions of Christianity, Judaism, and Islam) *is* compatible with free will (also a fundamental belief in those same religions). Both of these scholarly papers, though, depend much more on the nuances of grammar than most physicists will like.

The connection between spacetime physics and free will was made explicitly by the philosopher who wrote "For philosophers in either field, philosophy of science and philosophy of religion are too often viewed as mutually irrelevant . . . This is unfortunate, because sometimes the problems can be quite parallel and a consistent resolution is required. One especially intriguing case in point concerns, in

---

[146] In his *The Sirens of Titan*, a 1959 novel meant to be a parody of God's omniscience, Kurt Vonnegut gave the curious name of *chrono-synclastic infundibulated vision* to God's power to see the past and future.

[147] G. I. Mavrodes, "Is the Past Unpreventable?" April 1984, pp. 131–146, and A. Plantinga, "On Ockham's Way Out," July 1986, pp. 235–269, both in *Faith and Philosophy*.

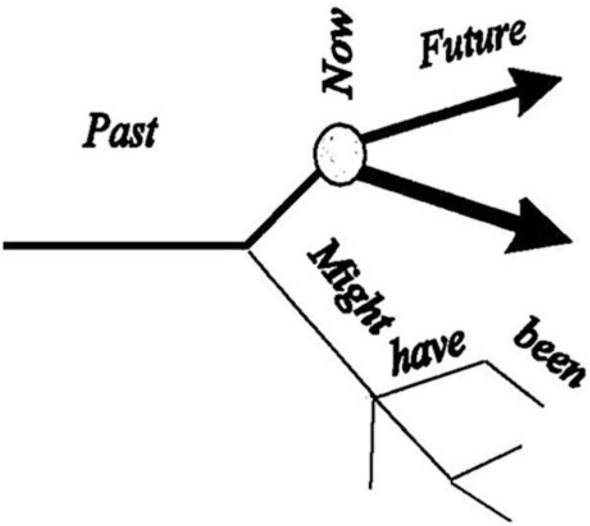

Fig. 2.2 The common view of time

philosophy of science, the possibility of ... time travel and, in philosophy of religion, the relationship between divine foreknowledge and human freedom."[148]

That philosopher could well have included science fiction writers in his group of people interested in both spacetime physics and free will. In one story,[149] for example, a man in the twenty-fifth century is about to travel back into the past to escape criminal prosecution. He is aked where he'd like to go, and he replies "I do not understand the paradoxes—what if I choose to build gravity-deflectors in Ancient Rome?" When he is told (correctly) that he couldn't do that because it didn't happen, he persists: "But if I can choose any period, it means I can alter history at will—which presumes that the present can also be changed." Then, at last, he gets the explicit answer that bothers nearly everyone: "The real answer is that in the final analysis your decision to choose a certain time period is already made, and the things you will do [in the time traveler's personal time] are already determined. Free will is an illusion; it is synonymous with incomplete perception." The same idea appears in another tale (note 57); when one character says, "What you are saying is that the future is fixed, and that you can read it, in every essential detail," the response is "Quite right ... both those things are true."

However, no matter how hard we try—and by *we* I include even those physicists and philosophers who embrace the block universe with its support of time travel to the past—it is very difficult to break free of the view of time as shown in Fig. 2.2. That is, as the passage of time up to the present or *now* (with all to the left of that instant as the past), while to the right of the *now* we have multiple possible futures (depending on our free will choices). Lying to the side of all that (in our thoughts and imaginations) are all that ‘might have been’ if we had made different choices than we did at earlier times in the past.[150]

---

[148] W. L. Craig, "Tachyons, Time Travel, and Divine Omniscience," *Journal of Philosophy*, March 1988, pp. 135–150. Tachyons are hypothetical faster-than-light particles that theoretically travel backwards through time. They will be discussed in Chap. 5.

[149] W. Kubilius, "Turn Backward, O Time," *Science Fiction Quarterly*, May 1951.

With all that said, even if events are really laid out in the spatial and temporal web that constitutes the four-dimensional block universe, there still remains the great mystery of why we see them unfold in the particular sequence that we do. Why not in reverse order? Why, indeed, do we see what we call *time* run from what we call the past to what we call the future and, indeed, what do we really mean by *past* and *future*? As you’ll see in the next chapter, these are not easy questions, and nearly everybody who has thought about them believes we are not yet even close to knowing the answers.

On that perhaps gloomy note, it seems appropriate to end here with a few more words from St. Augustine’s *Confessions*, with words that follow those that helped open this chapter: “I confess to you, Lord, that I still do not know what time is. Yet I confess too that I do know that I am saying this in time, that I have been talking about time for a long time, and that this long time would not be a long time if it were not for the fact that time has been passing all the while. How can I know this, when I do not know what time is? Is it that I do know what time is, but do not know how to put what I know into words? I am in a sorry state, for I do not even know what I do not know!”[151]

2.8 For Further Discussion

In the comics one of Superman’s more interesting adversaries is Mr. Mxyzptlk (pronounced *mix-yez-pitle-ick*), a being with seemingly magical powers from the Land of Zrfff in the fifth dimension. Mr. Mxyzptlk’s powers are not really because of magic, however, but are ‘merely’ the result of his hyperspace world with its extra dimension. Mr. Mxyzptlk, for example, in one of his misadventures with Superman in 1954, begins selling a

(continued)

---

[150] Figure 2.2 is based on a similar one in C. K. Raju, “Time Travel and the Reality of Spontaneity,” *Foundations of Physics*, July 2006, pp. 1099–1113.

[151] There is another view of time even darker than St. Augustine’s, which denies the existence of both future and past, and doesn’t offer us much either for that special moment we call the present (or *now*). This view, called *presentism*, was hauntingly expressed in some lyrics I heard in the final episode of the second season (2015) of the HBO series *True Detective*: “There is no future/There is no past/In the present nothing lasts.” Now *that* is depressing! Still, there are philosophers who believe even this view can support time travel: see S. Keller and M. Nelson, “Presentists Should Believe in Time-Travel,” *Australasian Journal of Philosophy*, September 2001, pp. 333–345.

newspaper called the *Daily Mpftrz* in competition with the *Daily Planet*. Unlike a traditional newspaper that reports what *has* happened, the *Daily Mpftrz* (your guess is as good as mine!) prints what *will* happen. As Mr. Mxyzptlk says, "You see, as a resident of the fifth dimension, I can get all the news I want from the *fourth* dimension!" The science editor at the *Daily Planet* explains the meaning of that to his boss, Perry White: "That's right, Mr. White . . . many physicists consider *time* the fourth dimension . . . so if Mr. Mxyzptlk can travel from the fifth dimension to our three-dimensional world, he most likely *is* able to see the future!" (This leaves unanswered the question of *why* he continues to challenge Superman when he knows he will always be defeated—*as he always is*!) Presumably a five dimensional world would have our three spatial and one temporal dimension (for a total of four), and so the question now is: what is the nature of the additional (fifth) dimension? Is it spatial or is it temporal? (There is a brief appearance of the fifth dimension in the 2014 movie *Interstellar*, but we aren't told much of anything about its possible structure.) Discuss and compare the world of four space dimensions and one time dimension, with the world of three space and *two* time dimensions. (In Chap. 5 we'll discuss a possible connection between two-dimensional time and time travel.)

In the text it is stated that "If A and B are mutually causative, then 'A causes B' coupled with 'B causes A' seems to lead to 'A causes A.'" Suppose, however, that we imagine two adjacent sunken pools of water, **a** and **b**, on the same horizontal surface, with each pool filled to the brim. An overflow from one pool will flow into the other pool. Now, define the events A and B as 'A is the overflow of pool **a**' and 'B is the overflow of pool **b**.' Thus, A causes B and B causes A. Does the conclusion 'A causes A' make physical sense in this specific case? Discuss at length.

When reading A. C. Clarke's story "Technical Error" (see note 99), we learn that a rotation through 4-space *inverts* "the unlucky Nelson." The 'solution' to this awkward situation is to flip Nelson through 4-space a second time and so back to 'normal.' (When *Thrilling Wonder Stories* reprinted this tale in June 1950, after its original publication in 1946, the title was changed to the more appropriate "The Reversed Man.") Clarke may have missed an

(continued)

important technical 'detail,' however, in that when first flipped through 4-space *everything* inverts, and so matter becomes anti-matter and Nelson would have instantly been annihilated in a 100 % conversion of matter to energy (that is, the flipped Nelson would have initiated a very large explosion). Compare this to Alice's concern in her flipped world (Lewis Carroll's *Through the Looking Glass*) when she wonders "Perhaps Looking-glass milk isn't good to drink." Explain why Lewis Carroll certainly was *not* thinking of matter/anti-matter explosions when he wrote his novel. What *do* you think he might have had in mind?

A time travel story, even earlier than Clarke's, that uses spacetime 'rotations,' was authored by Edmond Hamilton (1904–1977), one of the pioneering pulp fiction writers. In his "The Man Who Saw the Future" (*Amazing Stories*, October 1930), a man is hauled before the Inquisitor Extraordinary of the King of France to explain his mysterious disappearance, and subsequent reappearance, in an open field, amid thunderclaps and in plain sight of many onlookers. As the story unfolds, we learn that the man was transported five centuries into the future, from A.D. 1444 to 1944, by scientists working in twentieth-century Paris. The thunderclaps were produced by spacetime 'rotations,' as the atmospheres of 1944 and 1444 were reversed. A skeptical Inquisition naturally finds this tale preposterous and the first time traveler is burned at the stake as a sorcerer. Can you think of *why* such 'atmospheric swaps' might produce thunderclaps?

A trip around a Möbius strip reverses the 'handedness' of a plane figure (left and right are swapped). You can see this for yourself by making a Möbius strip, and then sliding an arrow (pointing *across the width* of the strip) around the strip. (Cut a notch in the side of the strip to mark the starting point, with the arrow pointing at the notch.) When you get back to the notch, the arrow will point *away* from the notch. Notice that the arrow never left the surface of the strip, or crossed any 'weird' boundary. Then, read H. G. Wells' short story "The Plattner Story" and comment on its use of 'handedness.'

The *autoinfanticide paradox*, which results when a time traveler tries to kill his younger self, continues to fascinate both physicists and philosophers, and papers regularly appear in the scholarly literature on the topic: see, for example, Kadri Vihvelin, “What Time Travelers Cannot Do,” March 1996, pp. 315–330 (which introduces Suzy the time traveler); Ira Kiourti, “Killing Baby Suzy,” June 2008, pp. 343–352; Peter B. M. Vranas, “What Time Travelers May Be Able to Do,” August 2010, pp. 115–121; and Joshua Spencer, “What Time Travelers Cannot *Not* Do (but are responsible for anyway),” October 2013, pp. 149–162, all in *Philosophical Studies*. All deal with an issue that is psychologically fascinating: *moral responsibility*. Spencer, in particular, opens with this definition: Someone is morally responsible for an action *only* if she could have done otherwise. As he goes on to write, “If I have been attacked and both of my legs have been broken, then it seems illegitimate to criticize me for failing to run away; I could not have done otherwise.” And yet all of these papers are on a point that (I think) physicists would soon lose interest in: is the question ‘If Suzy is a time traveler, can Suzy kill baby Suzy, given that Suzy doesn’t kill baby Suzy?’ the same question as ‘If Suzy is a time traveler, can Suzy kill baby Suzy, given that Suzy is now alive?’ The answer to the first question is, from pure logic, NO, while the answer to the second question is just bit squishier: it all depends on what the word *can* means. For the second question, Suzy *can* kill baby Suzy if she has a weapon (knife, gun, poison, etc.) and she is in the past next to baby Suzy, but it is just that she doesn’t because otherwise Suzy wouldn’t be alive now (which is a given). Such debates seem unlikely to produce any insights into the physics of time travel. Compare this situation to the old schoolboy conundrum “What happens when an irresistible force meets an unmovable object?’, which is a *self-inflicted* ‘paradox.’ That is, the words *irresistible* and *unmovable* are mutually exclusive and so, used this way, it should be no surprise that we have a conflict. Are the two time travel questions above, concerning Suzy, confusing through a similar mushy use of grammar? Or are they deeper than that? Vigorously defend your position.

In addition to H. G. Wells, another nineteenth-century writer who was highly influential in bringing the fourth-dimension out of academia and into public consciousness was the mathematician Charles Howard Hinton (1853–1907). Hinton was no angle-trisecting crank, having earned an M.A. at Oxford, an appointment in the mathematics department at Princeton, and then another at the University of Minnesota. Later, with the help of the eminent astronomer

(continued)

Simon Newcomb, he obtained a position at the Naval Observatory in Washington, D.C., and was on the staff of the United States Patent Office at the time of his sudden death. Hinton was a man to be taken seriously. His first published essay "What Is the Fourth Dimension?" appeared in 1880, and then in book form in 1884 as part of his *Scientific Romances* (a phrase used by Hinton before it became associated with Wells' science fiction many years later). That book received a generally favorable review in *Nature* (March 12, 1885, p. 431). At one point he wrote "We might then suppose that the matter we know extending in three dimensions has also a small thickness in the fourth dimension," an idea that was used a few years later by the well-known British mathematician W. W. Rouse Ball (1850–1925) in an attempt to explain gravity. Hinton was extremely inventive, and he also proposed four-dimensional-space models for static electricity. Find out more about Hinton's life and work: a good source to start with is *Speculations on the Fourth Dimension: Selected Writings of Charles H. Hinton* (R. Rucker, editor), Dover 1980. Take a look, too, at J. E. Beichler, "Ether/Or: Hyperspace Models of the Ether in America," in *The Michelson Era in American Science 1870—1930* (S. Goldberg and R. H. Stuewer, editors), American Institute of Physics 1988.

Chapter 3
The Physics of Time Travel: Part I

". . . within forty-eight hours we had invented, designed, and assembled a chronomobile. I won't weary you with the details, save to remark that it operated by transposing the seventh and eleventh dimensions in a hole in space, thus creating an inverse ether-vortex and standing the space-time continuum on its head."

—almost certainly *not* the way to build a time machine[1]

3.1 The Direction of Time

> "Of all the problems which lie on the borderline of philosophy and science, perhaps none has caused more spilled ink, more controversy, and more emotion than the problem of the direction of time . . . [T]he main problem with 'the problem of the direction of time' is to figure out exactly what the problem is or is supposed to be!"[2]

Before we start talking about the physics of time travel, let me say a few more words on time itself, in a way slightly less metaphysical that was the discussion in the previous chapter (which is why I'm writing this *here*, in a chapter with an increased emphasis on the analytical). When we speak of journeying to either the future or the past, we are implicitly making a distinction in the direction of the time traveler's trip. But does time actually have a *direction*? Is there an *arrow* that points the way? The answer seems obvious: *of course* time has a direction. After all, everybody 'knows' it flows from past to future. There is a curious language problem here, however, because we also like to say the present recedes into the past, which implies a 'flow' in the opposite direction, from future to past. Well, despite this snarled syntax, can we at least distinguish past from future, whichever way time flows?

---

[1] L. Sprague de Camp, "Some Curious Effects of Time Travel," in *Analog Readers' Choice*, Dial 1981.

[2] See note 54 in Chapter 2.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_3

This would seem to be an important question to answer because for the phrases *flow of time* and *direction of time* to have any objective meaning at all, it must be somehow possible to identify a difference between past events and future ones. The special moment at which that distinction occurs is known as the *now* or the *present* and, as events make the transition associated with that distinctive difference, between past and future, we say that the *now* (the present) moves or flows. Philosophers—and science fiction writers and physicists, too, who after all are human beings with human senses like everybody else—call this common feeling that we all have, of the passage of time, the *psychological arrow of time*. One philosopher gave an amusing (tongue-in-cheek) gastronomical interpretation of the moving *now* as follows:

> "New slices of salami are continually being cut from a nonexistent chunk of salami called the future. The *present* is the slice on top of the pile. The past are the pieces beneath this, and even though they are not present they still continue to exist in the same way that the top slice of salami does. . . . This [concept] faces humiliation before the embarrassing question of how fast the pile of salami slices grows."[3]

The 'moving now' does present a problem for physicists because there is nothing in the laws of physics that marks the present moment as unique, and therefore nothing that reflects a 'flow' of time, nothing that models the reality of a 'moving now' becoming part of the past and the events of the future becoming, successively, the new 'now.' As a philosopher wrote long before time travel became a serious topic in the physics literature, "Talk of the flow of time or the advance of consciousness is a dangerous metaphor that must not be taken literally."[4]

What that philosopher may well have had in mind is that all events in the block universe simply have coordinates in spacetime, and there is nothing corresponding to 'have been' (past), 'are' (present), or 'will be' (future). There is no 'moving now' in the block universe except for its subjective presence in our conscious minds. All we can say from physics is that events are ordered in an earlier/later sequence, and in fact, even that relatively weak condition holds only for *causally* related events.[5] The relativistic, four-dimensional block universe view of spacetime that so many physicists (including Einstein) so dearly love seems to have no room for an objective theory of the *flow* of time. And yet, even for those same physicists, there is a powerful psychological sense that time *does* flow. But are they mistaken? It is a fact that, with not just a little irony, Gödel (the 'discoverer' of time travel) was convinced that the possibility of a block universe spacetime with CTLs/CTCs

---

[3] R. Gale, "Some Metaphysical Statements About Time," *Journal of Philosophy*, April 25, 1963, pp. 225-237. For many, this analogy may well bring to mind a pile of baloney rather than one of salami (and I think this was Gale's intention).

[4] J. J. C. Smart, "The Temporal Asymmetry of the World," *Analysis*, March 1954, pp. 79-83.

[5] Two events A and B are *non*-causally related if their separation in spacetime is such that a particle would have to travel at a *superluminal* speed (faster than light) to go from A to B. We'll discuss the physics of causally related events later in this chapter.

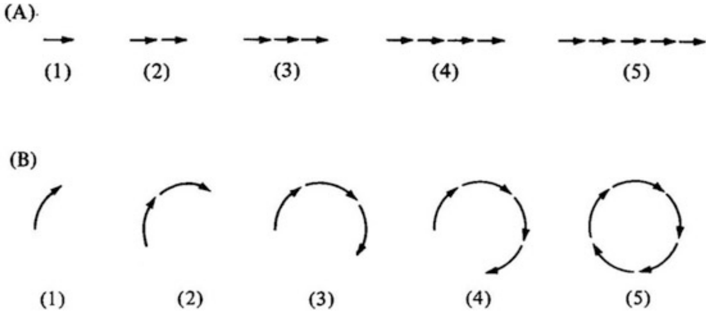

Fig. 3.1 Gödel's unreality of time argument

implies that the passage or flow of time *makes no sense*![6] To see how Gödel arrived at such an astonishing conclusion, consider Fig. 3.1.

In part A of the figure[7] we see what most people who talk of a passage of time intuitively mean, as time progresses through a sequence of instants (shown as line of left-to-right arrows going from (1) to (2) to (3) and so on). At each stage the right-most arrow is the present, and the arrows to the left of it (behind it) are the past, and the arrows to the right of the present are not shown because they are in the future and so don't exist yet. When you ask, at each step, which arrow is the *now*, the answer is clear.

Matters are dramatically different in part B of the figure, which shows the arrows forming a closed (circular) loop. Now there is no distinction between past and future, as each arrow is both ahead *and* behind any other arrow. In addition, there is no arrow that is uniquely the now. So, concluded Gödel, the passage of time can have no meaning in a temporal loop. As should come as no surprise, not everybody is convinced by this sort of argument.[8]

In principle, so it would seem, we can achieve perfect knowledge of what has happened but only imperfect prediction of what might happen. This observation seems to be at least a start at being able to tell past from future. And, in fact, the nature of the distinction between the two intervals of time seems obvious: we *remember* past events, but not future ones. As philosophers have so nicely put it, events in the past have formed *traces*, such as skulls, footprints in the sand, fossilized skeletons, surgical scars, photographs, taped recordings, carved stones, and the like, whereas future events appear not to have formed traces. But is that necessarily so? Is it impossible for future events to create traces? The common-sense answer is *yes*, because of cause and effect, which dictates that there *must* be a temporal asymmetry in trace formation. That is, traces are the effects of *prior* causes. That line of reasoning leads us quickly to the fundamental issue of causation (which we've already encountered in Chap. 2), an issue that no discussion of time travel can avoid.

---

[6] Gödel clearly states this in his 1949 philosophical essay (note 15 in the Introduction) concerning his 1949 technical paper (note 11 in the Introduction).

[7] This figure is based on the interpretation of Gödel's reasoning as presented by the philosopher Palle Yourgrau's 1991 book *The Disappearance of Time: Kurt Gödel and the idealistic tradition in philosophy* (Cambridge), which was expanded and reprinted a few years later under the new title *Gödel Meets Einstein: time travel in the Gödel universe*, Open Court 1999. Yourgrau later wrote a less technical version: *A World Without Time: the forgotten legacy of Gödel and Einstein*, Basic Books 2005.

[8] See, for example, S. Savitt, "Time Travel and Becoming," *The Monist*, July 2005, pp. 413-422.

Part of the problem we have with backward time travel, and cause and effect, is as I've already mentioned, with language. The distinct and separate concepts of the temporal ordering of events, and of causality, have become merged in everyday thought. It is considered *obvious* to modern minds that if event A causes event B, then A *must* happen first. There is, however, at least one historical example of a similar merging of concepts that is parallel to our modern mixing of time order and causality—an example that shows how an issue can seem obvious and natural to the minds of one period of time, and yet to the minds of another period (our modern times) seem confused, odd, peculiar, even laughable.

As a *physicist* wrote in a paper[9] on advanced (that is, inverted causality) effects:

> *Ancient Egypt was an essentially one-dimensional country strung out along the Nile, which flows from south to north. The winds were conveniently arranged to be predominantly northerly. To go north, a traveler could let his boat drift, while with a sail he could move south against the slow current. For this reason, in the writing of the ancient Egyptians, "go downstream (north)" was represented by a boat without sails, and "go upstream (south)" by a boat with sails. The words (and concepts) or north-south and up-downstream became merged. Since the Nile and its tributaries were the only rivers known to the ancient Egyptians, this caused no difficulties until they reached the Euphrates, which happened to flow from north to south. The resulting confusion in the ancient Egyptian mind is recorded for us to read today in their reference to "that inverted water which goes downstream (north) in going upstream (south)."*

Often we can work our way free of the difficulties we create for ourselves with language, but only through common agreement. For example, the chairman of the board calls a meeting to order with mixed tenses by declaring "The meeting *will* take place *now*" and then saying at the end, "We will meet again *next* month, *same* time." We all know what these sentences mean, but only by our cultural heritage and not by the process of applying logic. The language problem causes similar difficulties for not only for fictional time travelers, but also for the physicists/philosophers who study the possibility of time machines. So—beware!

The idea of time flowing is a popular one, and it repeatedly appears in the time travel literature as the "river of time" or the "ocean of time." The deep psychological appeal of this sort of 'water language' has, not surprisingly, attracted the attention of philosophers. We can find one of the earliest expressions of the view in the *Meditations* of the second-century A.D. Roman emperor and Stoic philosopher Marcus Aurelius, who wrote: "Time is like a river made up of events which happen, and a violent stream; for as soon as a thing has been, it is carried away, and another comes in its place, and this will be carried away, too." A most interesting essay on why such metaphors often seem so intuitively appropriate has been offered by one philosopher, who points out[10] that the seductiveness of the image of 'time as flowing water' is sufficiently great that you often find it in the scientific literature, too (Newton, you'll recall, wrote specifically in his *Principia* of time flowing). As for why such an image has such a powerful grip on our imaginations, I think we need look no further than to Kant. As he wrote (1781) in *Critique of Pure Reason*, "Time is nothing but the form of inner sense, that is, of the intuition of ourselves and of our inner state . . . Because this inner intuition yields no shape, we endeavor to make up for this want by analogies." And what better than a rushing stream of water to represent our feeling of time rushing by?

---

[9] P. L. Csonka, "Advanced Effects in Particle Physics," *Physical Review*, April 1969, pp. 1266-1281.

Still, no matter how intuitive such water metaphors may be, they can still easily befuddle us as well. To quote our philosopher (note 10), "Time a river! A queer sort of river that. Of what sort of liquid does it consist? Is time a liquid? A very peculiar liquid indeed!" A classic paper by the philosopher Donald Williams (discussed in Chap. 2, note 119) expresses similar doubt about the water image of time. In the course of his writing, he presents a truly staggering collection of entertaining examples of 'time as metaphor,' of which I repeat just a few here: time flies, goes, marches, and rolls, as well as flows And then he offers *this* provocative imagery: the evolution of our lives is like "a moving picture film, unwinding from the dark reel of the future, projected briefly on the screen of the present, and rewound into the dark can of the past." Wow![11]

Returning to the water metaphor, the French astronomer Charles Nordmann (1881–1940) opened and closed his 1925 book *The Tyranny of Time* with following gloomy but all too true summary of the overwhelming sense we all have of the inexorable, one-way 'flow' of time. (The ellipses in what follows denote over 200 pages!) "Nothing can equal the bitter sweetness of dreaming on the banks of Time, that impalpable and fatal river strewn with dead leaves, our wistful hours carried downstream like rudderless wrecks . . . In the eternal wave which rocks us, carries us along, and soon swallows us up, there is no rock to which we can fasten our frail barques; the very buoys we put out to measure our course are only floating mirages; and on the mysterious foundation of things our anchors slide along and fail to bite." A young person sees time, from Nordmann's perspective, as an ocean on which golden mornings arrive like waves from the future, whereas for an older person, liquid time is a nightmare flood, a swollen black torrent sweeping him first into the yawning abyss of the past and, ultimately and finally, into the eternal silence of the dark grave.

---

[10] J. J. C. Smart, "The River of Time," *Mind*, October 1949, pp. 483-494.

[11] And how about *this* image of time: Time is a snowball, with the center marking the beginning of the past, with ever new 'presents' accreting on the ever increasing surface as the snowball rolls down the hill of history!

The metaphor of time as a flowing river was ready-made for early science fiction writers, such as Caltech math professor Eric Temple Bell (1883–1960). His eventual novel *The Time Stream* began to appear in December 1931 as a serial in the science fiction pulp *Wonder Stories*, and Bell (writing as 'John Taine') made great use of the idea of time as a flowing stream, a stream in which one could swim into either the future or the past. There is strong evidence that Bell actually wrote the novel in July of 1921, but was unable to find a publisher for a decade, so odd did editors find the premise. By the time of its publication, others had beaten Bell into print.

The watery image of time had appeared a year earlier, for example, in a tale that played with the erosive nature of time in a dramatic way. As two time travelers speed into the future to rescue a friend, one of them describes the scene for us: "We huddled together in the whirling time girdling machine, cutting through the years as a ship's prow breasts surging waves. I could not help but think of the years as waves, beating in endless succession on the sands of eternity. They wore all away before them with pitiless attrition. Time seemed to eat all with dragon jaws."[12]

This image of time was taken a step further 3 years later in a story in which a large number of adventurers, from all across time, find themselves stranded at precisely the same place (in space and time). One of them offers his theory of what is behind this remarkable coincidence: They all have faulty time machines, like faulty boats, and all have hit the same snag on the 'river of time.' As he explains, "You may turn boats adrift on a river at many points, and they will all collect together at the same serious obstacle whether they have traveled a hundred or two miles. We are now at some period where the straight flow of time has been checked — perhaps it is even turning back on itself ... [We] have struck some barrier and been thrown up like so much jetsam."[13]

The 'flow' of time does have its critics, of course. The British-American philosopher Max Black (1909–1988) argued[14] that questions about the direction of time are meaningless because there can be no direction to something that (he asserted) does not flow. His reasoning was that if time does flow, then he ought to be entitled to ask *how fast* it flows. That requires, in turn, a *metatime* or *supertime* for measuring the flow rate of 'ordinary' time. But because supertime must flow, too, we would then need a super-supertime, and so off we trip into what would appear to be the black hole of a McTaggert-like infinite regress. The view, of an infinite regress of times, was forcefully rejected by another philosopher with these sharp words: "the very idea of super (or hyper)-time is indeed repulsive in its redundancy and its aroma of dilettante physics."[15]

---

[12] E. A. Manley and W. Thode, "The Time Annihilator," *Wonder Stories*, November 1930. This is the same magazine that, months later, finally published Bell.

[13] J. Wyndham, "Wanderers of Time," *Wonder Stories*, March 1933. Notice again, that we have the same magazine (whose editor must have had a particular fancy for such tales).

[14] M. Black, "The 'Direction' of Time," *Analysis*, January 1959, pp. 54-63.

A hierarchy of hypertimes has not bothered other analysts, however, and an entire subfield of specialty among philosophers (and some physicists, too) in time analysis has developed in what is called *multidimensional time*. One practitioner in this specialty sarcastically rejected the infinite regress complaint as a valid objection—he called it "a crushing and unanswerable position" but actually meant just the opposite—and stated that it was not at all clear (at least, not to him) why supertime must flow.[16] After all, he argued, we measure the flow of a river with respect to its banks without requiring that the banks themselves flow. (That actually strikes me as being a point that deserves debating, but I have not been able to find any mention of it in the later philosophical literature.) The idea of multiple time dimensions is particularly attractive for one sort of time travel (we'll take it up at the end of this chapter), but it enjoys far more popularity among science fiction writers and philosophers than it does with physicists.

Professor Black's objection (note 14) to talk of time 'flowing' was based, at least in part, on the observation that there are uses of the word *direction* that are not directly tied to something flowing. For example, consider the statement 'He is facing in the direction of north.' Black argued that this is mere pointing, and it is not at all the same as *moving* north. He then dismissed the possibility of there being any meaning to the direction of time, writing that making an analogy of time "with a sign-post or an index finger is too far-fetched to be worth considering." This claim (which some may feel leans too much on grammar) is, of course, an affirmation of the myth-of-passage view made famous a few years earlier by Donald Williams (note 119 in Chap. 2).

Not just philosophers have rejected the idea of time flowing. In his 1966 novel *October the First Is Too Late*, which deals with a world in which different parts of Earth simultaneously experience different eras of the past (see *For Further Discussion* at the end of this chapter for more on what this might mean), the British cosmologist Fred Hoyle (1915–2001) calls the 'river of time' a "grotesque and absurd illusion," and a "bogus idea." Another fictional work that agrees with Hoyle's non-moving image of time is the 1979 *Roadmarks* by Roger Zelazny (1937–1995). In that novel we read of "the Road," along which story characters can travel but which doesn't itself move; exits from the Road lead to the various centuries (which sounds a lot like the Francis Bradley's 1883 book that may have given the block universe its name). *Roadmarks* is a clever bit of writing, with many allusions to the paradoxes of time travel, but its explanation of the Road's origin as having been constructed by *dragons* (!) greatly undermines its interest for physicists.

---

[15] D. Zeilicovici, "Temporal Becoming Minus the Moving-Now," *Nous*, September 1989, pp. 505-524.

[16] C. W. Webb, "Could Time Flow? If So, How Fast?" *Journal of Philosophy*, May 1960, pp. 357-365.

In a block universe spacetime, there is no flow of time, but one philosopher believed that to be simply because the block universe is *incomplete* in its representation of reality. Writing in 1925, Hans Reichenbach (1891–1953) asked "What does 'now' mean? Plato lived before me, and Napoleon IV will live after me. But which one of these three lives *now*? I understandably have a clear feeling that *I* live now. But does this assertion have an objective significance beyond my subjective experience?"[17] Reichenbach went on to answer his question in the affirmative, and to deduce that the block universe view is missing something: "In the condition of the world, a cross-section called the present is distinguished; the 'now' has objective significance. *Even when no human is alive any longer, there is a 'now'* [my emphasis] . . . In the four-dimensional picture of the world, such as used by the theory of relativity, there is no such distinguished cross-section But this is due only to the fact that an essential content is omitted from this picture."

So, what *is* Reichenbach's 'missing essential content'? Feeling that the block universe is unacceptably fatalistic—in his words of ridicule, "the morrow has already occurred today in the same sense as yesterday"—he found his answer in the probabilistic theory of quantum mechanics. Classical physics argues that given total information about the state of the world *now*, one could in principle calculate perfectly the future or the past; one could both predict and retrodict. In contrast, quantum mechanics distinguishes past from future in a fundamental way.

Quantum mechanics does not deny that in principle we can know the past with exquisite accuracy, because each and every event leaves traces, evidence that is available to all with the means to find and decode them. But quantum mechanics also takes as truth that there is an unavoidable uncertainty to the future. The instant that this uncertainty is crystallized into fact was taken by Reichenbach to be the very definition of 'now.' The ever-increasing record of the past, in turn, defines (for Reichenbach) the *movement* of the 'now.' Reichenbach believed that with these observations he had at last captured the 'moving now' in mathematical theory, and that he had finally elevated the present from speculative psychology to solid physics, and that he had shown that the 'flow of time' is independent of the need for a conscious mind. However—

A later, powerful analysis[18] of the time-flow issue, combining philosophy with physics, comes down solidly in support of the *opposite* conclusion: it expresses the view that a 'moving now' *is* only in our minds and is *not* an intrinsic attribute of reality. The premise of that argument is that a mind-dependent flow of time is incompatible with what is called the *relativity of simultaneity* (to be discussed later in this chapter) which states that there is no universal cosmic-wide 'now' (this is a fundamental conclusion of special relativity). For example, it is meaningless to ask what is happening on a planet in the Andromeda galaxy (two million light-years distant) *right now*.

---

[17] I've taken this quotation from A. Grünbaum, "Is There a 'Flow' of Time or Temporal Becoming?" in *Philosophical Problems of Space and Time*, Knopf 1963.

[18] L. R. Baker, "Temporal Becoming: The Argument from Physics," *Philosophical Forum*, Spring 1975, pp. 218-236.

Early science fiction stories are full of theories about the nature of 'now,' and the vast majority of them have no basis in scientific thought. Some of them are ingenious, however, and even though they are largely the pet ideas of the authors (and no one else's), perhaps they resulted in some young readers of the science fiction pulps of the 1930s and 1940s thinking about deeper matters than did the comic strips of "Buck Rogers," "The Lone Ranger," or "Terry and the Pirates." For example, according to one story, time is a wave and the 'moving now' we experience is carried on a crest of that wave. There are time waves both ahead and behind the crest we happen to be on (so we are told), and so each such crest carries a different 'now' for a different reality—hence the curious title.[19]

In another, more recent tale[20] about object duplication via time travel (which we'll discuss in Chap. 4, but you'll recall H. G. Wells was worried about this long ago), nine (!) copies of the same person from the year 2314 meet in 1870 to try and figure out what is going on. Part of their interesting discussion is the following analysis of the 'present':

> "Gentlemen, I think I understand," said the first James Thomas."
>
> "Eight faces turned toward him, and he felt as though he were looking into multiple mirrors."
>
> "We hold that time is a single instant — the instant of the Present —which travels through Duration — do we not?"
>
> "Eight heads nodded."
>
> "We assume that time passes in a manner analogous to the stringing of an infinite number of beads. Each bead is the instant of Now when it is last on the chain. Beads are continually being added, and each one is the only Now until another is placed after it."
>
> "Yes, that is my theory," said another James Thomas. "It can also be likened to the process of knitting. No matter how many stitches are knitted, there is only one last stitch, only one Now."

Einstein, too, was greatly bothered by the place of 'now' in time, perhaps even more than were James Thomas and his 'friends.' In an autobiographical essay, the philosopher Rudolf Carnap (1891–1970) recalled a conversation about this with Einstein in the early 1950s, at the Institute for Advanced Study in Princeton: "Once Einstein said that the problem of the Now worried him seriously. He explained that the experience of the Now means something special for man, something essentially different from the past and the future. That this experience cannot be grasped by science seemed to him a matter of painful but inevitable resignation. . . . Einstein thought . . . that there is something essential about the Now which is just outside the realm of science."[21]

---

[19] R. Ray, "Today's Yesterday," *Wonder Stories*, January 1934.

[20] A. and P. Eisentein, "The Trouble With the Past," in *New Dimensions 1* (R. Silverberg, editor), Doubleday 1971.

3.2 The Arrows of Time

> "On a microscopic level there is no preferred direction for time. The equations of motion don't give a damn whether time moves forward or backward."[22]

The central issue for philosophers of time (and for physicists, too, I think) is that of its reality (or not): is time *objective and something that really flows*, or is time simply a *mind-dependent illusion* and nothing more than an artifact of our incomplete perception of reality? As the previous section shows, there is little consensus on this issue. As a start on trying to get a handle on the matter, looking into a so-called 'arrow of time' may give us some guidance. I'll begin with the arrow I've already mentioned, the psychological arrow. As discussed before, this is the feeling we have of a 'moving now,' a feeling that has no appearance anywhere in physics. A 'moving now' simply has no place in any universe devoid of the physical processes in a brain that give rise to what we call consciousness—but that doesn't mean physicists don't wonder about the 'moving now' just as much as does everybody else (remember Einstein)! As one physicist wrote in a technical journal, "What does 'Now' mean? This question must surely be the starting point of any attempt at understanding the nature of time."[23]

Well, no matter whether time actually flows or not, most of us still believe we have had a past and hope we will have a future. Each of us thinks we can easily tell one from the other, too. We have, in fact, many not so subtle indications from our everyday lives of the obvious direction of time. Nearly all of these indications have the common theme of *irreversible change*. As the British mathematician J. J. Sylvester once put it, "The whirligig of time brings about its revenges."[24] The Roman poet Ovid, who died when Christ was a teenager, said the same in his *Metamorphoses* with the famous words "Time, the devourer of all things." The image of time as devourer of all that is mortal was brilliantly presented by James Barrie in his *Peter Pan*, with the crocodile who had swallowed a ticking clock chasing Captain Hook all about Neverland.

---

[21] Quoted from *The Philosophy of Rudolp Carnap* (P. A. Schlipp, editor), The Library of Living Philosophers, Open Court 1963, pp. 37-38. For a view contrary to Einstein's, from another physicist, see K. B. M. Nor, "A Topological Explanation for Three Properties of Time," *Il Nuovo Cimento B,* January 1992, pp. 65-70, which claims to develop a geometrical explanation for the flow of time, and so (says Nor) there *is* an objective, mathematical reality to the 'moving now.'

[22] A science fiction character pretty accurately sums-up what a modern physicist would tell you today, in L. Eisenberg's story "The Time of His Life," *The Magazine of Fantasy & Science Fiction*, April 1968.

[23] J. P. Cullerne, "Free Will and the Resolution of Time Travel Paradoxes," *Contemporary Physics*, July-August 2001, pp. 243-245.

[24] J. J. Sylvester, "A Plea for the Mathematician," *Nature*, December 30, 1869, pp. 237-239.

No one yet has escaped the biological decay processes of time, and inanimate objects are no less immune to this aspect of time. Logs and cigarettes burn in the stove and ashtray, but they never unburn. Our cars rust but never 'unrust.' An explosion has never been seen to reverse itself, to form a dynamite stick or a bomb casing out of a collapsing fireball. Our world seems, indeed, literally to be built on an irreversible movement toward chaos, death, and decay. Lewis Carroll uses this observation in his *Through the Looking-Glass* when Alice tells Humpty Dumpty "one can't help growing older." And speaking of Humpty Dumpty, his famous fall provides a dramatic example of a one-way evolution from past to future; he wasn't at all convinced that Alice was correct but, once he had splattered, then

All the King's horses and all the King's men
Couldn't put Humpty Dumpty together again.

While we are on the subject of Mr. Dumpty, it is also worthwhile to note that nobody has ever figured out how to unscramble an egg. Why is that? One philosopher speculated that the answer is found in the "irreversible organic phenomena" taking place in our brains which results in our flow of consciousness always being in the same direction.[25]

More subtle than the undignified undoing of a prideful egg is the phenomenon of memory, which seems trivial only because most people have not thought very carefully about it. We remember the past while remembering nothing about the future. We might, in fact, be tempted to use the phenomenon of memory to answer the question of how to tell past from future. Anything you can remember *is* the past. But that is a circular definition, as discussed by Professor Smart (note 4) who observed that to ask why memory is always of the past "is as foolish as to ask why uncles are always male, never female." In *Through the Looking-Glass* the White Queen tells Alice that "it's a poor sort of memory that only works backward" but, except for the claims of clairvoyants, it seems that is the only sort of memory any of us has. Why is that so? Of course, that would *not* be the case for a time traveler while in the past. His personal past, which he would remember, would be the future for the world around him.

For physicists, the question of the direction of time is one of profound mystery. There seems, in fact, to be no fundamental reason why time should not be able to go from future to past—but then what would 'future' and 'past' *mean*—even though no one has ever observed time to do so. All the laws of classical physics, including general relativity, and quantum mechanics, too (except for the K-mesons mentioned in Chap. 1) involve time in such a way that they ignore its sign. In other words, replacing t with $-t$ results in a perfectly valid description of something that could actually happen. But not all such possibilities are observed to occur. Why not?

---

[25] H. Margenau, "Can Time Flow Backwards?" *Philosophy of Science*, April 1954, pp. 79-92.

In an unpublished paper written in 1949, while doing the work that would bring him a share of the 1965 Nobel prize in physics, Richard Feynman (1918–1988) wrote[26] “The relation of time in physics to that of gross experience has suffered many changes in the history of physics. The obvious difference of past and future does not appear in physical time for microscopic events . . . Einstein discovered that the present is not the same for all people [the relativity of simultaneity, to be discussed later in this chapter] . . . It may prove useful in physics to consider events in all of time at once and to imagine that we at each instant are only aware of those that lie behind us. The complete relation of this concept of physical time to the time of experience and causality is a physical problem which has not been worked out in detail. It may be that more problems and difficulties are produced than are solved by such a point of view.”

Feynman did not elaborate on what he meant by the “problems and difficulties” with that point of view (which is clearly that of the block universe), but surely he had the logical paradoxes of time travel high on his list. As the Yale philosopher Henry Margenau wrote (note 25) in a tutorial on Feynman’s work, “The theory of quantum electrodynamics developed by Feynman incorporates reversals in the course of time and thereby cherishes, in the minds of many, *an age-old phantasy* [my emphasis] of more than scientific appeal [which sounds like time travel to me].”

Because the individual classical equations of microscopic physics are time-reversible, the distinction between past and future for individual particles disappears. The equations are said to be symmetric with respect to time; the algebraic sign of t is irrelevant in the classical laws. It must be understood, however, that there *is* a crucial point to appreciate. When a physicist says *time reversal*, she is talking about a system evolving backward in forward time—that is, all the individual particle velocity vectors are instantly reversed at once. This is distinct from the time-reversed worlds of philosophers and science fiction writers (which we’ll get into later in this chapter) in which time itself ‘runs backwards.’ The physicist’s point of view is clearly expressed in an early essay by a chemist: “Every equation and every explanation used in physics must be compatible with the symmetry of time. Thus we can no longer regard effect as subsequent to cause. If we think of the present as pushed into existence by the past, we must in precisely the same sense think of it pulled into existence by the future.”[27] More than three decades later, a mathematician and a physicist presented a similar statement: “In classical dynamics, the past completely determines the present, and therefore, by symmetry, the future also completely determines the present.”[28]

---

[26] S. S. Schweber, “Feynman and the Visualization of Space-Time Processes,” *Reviews of Modern Physics*, April 1986, pp. 449-508.

[27] G. N. Lewis, “The Symmetry of Time in Physics,” *Science*, June 6, 1930, pp. 569-577.

[28] O. Penrose and I. C. Percival, “The Direction of Time,” *Proceedings of the Physical Society* (London), March 1962, pp. 605-616.

Besides the physics, there is also an interesting theological connection to time reversal. As one philosopher put it, "If all the laws are time reversal invariant and so no irreversible processes occur in the physical Universe then there is no inherent, intrinsically meaningful difference between past and future . . . If this is actually the natural case, then all mankind's major religions which preach a creation of the Universe (by a supernatural agency) and imply, accordingly, a differentiation between the past and the future . . . would have to make appropriate adjustments."[29]

There are, as you might expect on such a controversial topic, dissenters to the view that the classical laws of physics are necessarily time-reversible. Dirac himself wrote that "I do not believe there is any need for physical laws to be invariant under time and space reflections, although all the exact laws of nature so far known do have this invariance."[30] Dirac did not, unfortunately, elaborate on just why he felt that way, but with the later discovery of K-mesons his position is seen to have been 'ahead of its time'! In a famous science fiction story[31] dealing with the direction of time, one character finally puts his finger on the real puzzle of the question of time: "How can a man live backward? You might as well ask the Universe to run in reverse entropy." That cogent question brings us, in fact, to the first scientific explanation developed to explain the observed asymmetric nature of time.

It was the Englishman A. S. Eddington (1882–1944) who gave the picturesque name, the *arrow of time*, to the observed asymmetric nature of time's direction from past to future. He was also one of the popularizers of an explanation for the arrow, using the famous second law of thermodynamics.[32] The second law of thermodynamics states that a measure of the internal randomness or disorder—what is called the *entropy*—of any closed system (that is, one free of external influences) continually evolves toward that of maximum disorder, toward the condition called *thermodynamic equilibrium*. Indeed, so striking is this increase in entropy S with time in a macroscopically large system that the increase in entropy has come to be thought of as actually *defining* the direction of time. Eddington, however, was not the originator of the entropy concept. The history of entropy can be traced back to before the turn of the century, to the great Austrian scientist Ludwig Boltzmann (1844–1906) and his famous H-theorem. The quantity H in that theorem is directly related to the more familiar entropy,[33] defined by Boltzmann in 1877.

---

[29] H. Mehlberg, "Philosophical Aspects of Physical Time," in *Basic Issues in the Philosophy of Time* (E. Freeman and W. Sellars, editors), Open Court 1971.

[30] P. A. M. Dirac, "Forms of Relativistic Dynamics," *Reviews of Modern Physics*, July 1949, pp. 392-399.

[31] A. Boucher, "The Chronokinesis of Jonathan Hull," *Astounding Science Fiction*, June 1946.

[32] A. S. Eddington, *The Nature of the Physical World*, Macmillan 1929.

[33] The H-theorem was a direct continuation of the work by the Scottish physicist James Clerk Maxwell (1831-1879) on the statistical properties of gas molecules (determining the probability density function of the molecules' speeds). In 1866 Maxwell found this function for the particular case of thermodynamic equilibrium. In 1872 Boltzmann found the differential-integral equation the function satisfies in general, even if the condition of thermodynamic equilibrium doesn't hold. From this Boltzmann was able to define a quantity H that he showed evolves in time such that

The entropy S of a system in a given state is proportional to W, which is the number of different possible ways the state can occur as a result of all possible variations of system's internal, microscopic structure. The calculation of W is usually quite complicated, but in various highly idealized systems it can be straightforward. Consider, for example, a vacuum cylinder with a thin membrane dividing the interior into halves. Suppose that we insert (to be specific) six molecules into the left half of the cylinder (and none into the right half). If we define the microscopic state of the system to be the number of molecules in the left half, then initially $W = 1$ because there is just one way to put all six molecules on the left side. This represents the state of *minimum* entropy, the state of maximum order that is most distant from thermodynamic equilibrium. If we now puncture the membrane then the molecules, once confined to the left side, are free to move about the entire cylinder. At any given instant we can imagine counting the number of molecules on the left side—suppose that at some particular instant we count five, with one molecule having moved to the right side. Then, $W = 6$, because there are six ways to pick the molecule that has moved from left to right, and so the entropy has increased.

We think of the thermodynamic equilibrium state as being the state with equal numbers of molecules in both halves of the cylinder, and that state has the *maximum* entropy. (Can you show that this state is associated with $W = 20$?) With such a small number of molecules, it is *not* clear that W (and so S) will *inexorably* increase with time; perhaps, after one of the six molecules has gone to the right, it then returns to the left side before any of its companions have joined it on the right. Such an event is called a *reversal*, and it will happen with some non-zero probability. But the more molecules there are in the cylinder (instead of six, make the number a million million million—still a small amount of gas in our everyday world, hardly enough to fill a sewing thimble), the more likely it becomes that the value of S *will* monotonically increase with time.

The steady increase in entropy is often observed in the everyday, large-scale world. A drop of ink in a glass of water spreads out in an expanding cloud, a cloud we *never* see collapse backward into an ink drop. A long rod of metal, initially hotter at one end than at the other, evolves toward a constant temperature along its entire length. We *never* see a uniformly warm rod spontaneously begin to cool at one end and grow hot at the other. A hot bath grows cold—nobody has *ever* seen a bath at room temperature suddenly, all by itself, begin to heat up and then boil in the middle of the tub while the edges freeze into ice chunks. In all of these cases, the end (future) state represents greater internal randomness or disorder than does the beginning (past) state.

---

solution to his differential-integral equation approaches Maxwell's equilibrium solution. The H-theorem says that H always *decreases* in systems not in equilibrium and is at a *minimum* in systems in equilibrium.

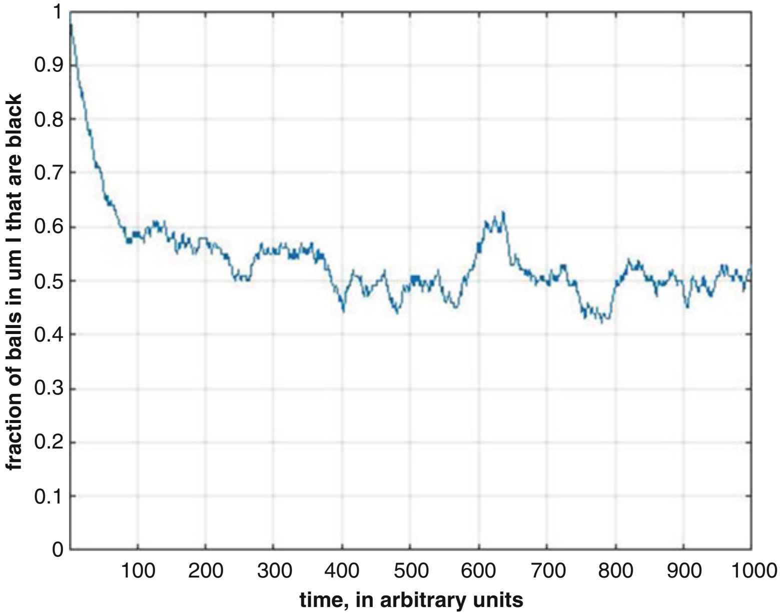

Fig. 3.2 Simulation of the Ehrenfest entropic gas clock

That is, low entropy *was* the past, and high entropy *will be* the future. The increase in entropy seems to define a direction to time, and so entropy has come to be called the *thermodynamic arrow of time*.

The first formal entropy model for the direction of time was put forth in a 1907 paper by the Austrian physicist Paul Ehrenfest (1880–1933) who was a friend of Einstein, and his Russian-born wife Tatyana (1880–1964), who was a skilled mathematician and her husband's occasional collaborator. In their paper the Ehrenfests developed one of the mainstays of physics, the so-called *entropic clock*. This clock, a statistical model based on the then new probability mathematics of Markov chains—after the Russian mathematician A. A. Markov (1856–1922)—describes how gases diffuse, and it is both a simple and a powerful concept. The Ehrenfest model is illustrated in Fig. 3.2, in a computer-generated plot based on a discussion by Princeton physicist John Wheeler of black hole fame (see Chap. 1).[34]

Imagine two urns, I and II, each containing n balls. Initially, at time $t = 0$, all of the balls in Urn I are black and all of the balls in Urn II are white. Then, at time $t = 1$ (in arbitrary units), a ball is selected at random from each urn and (instantaneously) placed in the other urn. This select-and-transfer process is repeated at times $t = 2$,

---

[34] J. A. Wheeler, "Frontiers of Time," in *Problems in the Foundations of Physics* (G. T. diFrancia, editor), *Proceedings of the International School of Physics* (Course 72), North-Holland 1979. See also W. J. Cocke, "Statistical Time Symmetry and Two-Time Boundary Conditions in Physics and Cosmology," *Physical Review*, August 25, 1967, pp. 1165-1170.

3, 4, ... At any given time each urn always contains n balls, but only at $t = 0$ are the colors of all the balls in a given urn necessarily the same. The phrase "selected at random" means (for example) that the probability of selecting a black ball from an urn containing b black balls is b/n. At any given time we completely describe the state of both urns by specifying the number of black balls in Urn I (or the number of white balls in Urn II, and so on). It is easy to write a computer simulation of this physical process,[35] and Fig. 3.2 shows how the fraction of black balls in Urn I evolves toward 0.5 as time increases. The plot is for $n = 100$ (200 balls total). The important observations are that (1) the evolution of the state of the system is toward 50 % black balls in Urn I (and this would be the case for 'almost all' sequences of random selections of the balls from the urns), and (2) the evolution is not monotonically decreasing from 100 % black balls to 50 % black balls, but rather has never-ending fluctuations about 50 % that may, in fact, be rather large in both amplitude and in duration.

There is a real puzzle with the entropic clock that may not be immediately apparent. The motion of each of the individual molecules is described by time-reversible physics, but when we average over 'many' molecules (assuming 200 molecules is 'many') we lose detailed information about the individual molecules. The puzzle is then how is it that by *reducing* our knowledge of a system, through statistical averaging, we then find it displaying a *new property*, that of asymmetric time evolution, that we didn't see before when we watched the individual molecules. And if that question isn't troublesome enough, we also have two additional puzzles called the 'reversibility' and the 'recurrence' paradoxes to consider as well.

The reversibility paradox is the question raised earlier: the classical equations of physics work just as well with time running in either direction, and so *why don't* things actually go 'backward'? This question, originally raised by the British mathematical physicist and engineer William Thomson (1824–1907)—better known as Lord Kelvin—in 1874, was brought to Boltzmann's attention in 1876 by the German physical chemist Johann Loschmidt (1821–1895), one of Boltzmann's professors at the University of Vienna. Boltzmann's answer to this apparent paradox was that it *is* imaginable that a world could run backward if initial conditions were suitable. For example, if all the velocity vectors of every particle in an equilibrium state were reversed, then the system *would* unwind backward in time toward its original non-equilibrium condition. That is, a system in thermodynamic equilibrium, the state of highest entropy, could evolve toward one of low entropy. Boltzmann even suggested that such might be the case for regions in our own universe, that there might actually be beings in a world somewhere 'out there' who experience time running counter to our earthly experience. He said that in 1877, and it is a remarkable statement for a conservative nineteenth-century professor.[36] However, Boltzmann continued, from most given states there are vastly more ways for entropy to increase than there are for it to decrease, and that is why we see what we see, a continuous *increase* in entropy.[37]

---

[35] I used MATLAB, and you can find the code — gasclock.m — in Appendix C, written in such a low-level way as to be virtually 100% transferable to just about any of the popular scientific programming languages, and easily executed on an inexpensive laptop. Note that there are no K-mesons in the code (!) and so, as stated in Chapter 1, they aren't responsible for the uni-directional time behavior depicted in the figure.

To find a science fiction writer speculating on reversed time people is, of course, much less remarkable! One pulp story, in fact, presents a curious treatment of the nuances of reversed time in which people talk backward (along with a marvelous bathroom scene of a man un-washing his hands!). This tale[38] tells us a young physics teacher who is "twisted into a reversed Time Stream" by an electrical discharge. As he lives backward in time, he observes everybody about him appearing to run in reverse, but even more puzzling is that they have developed a "dreadful, granite-like hardness." We soon learn why:

"For a while he could not understand the impenetrable hardness of external objects which he had experienced; it seemed they ought rather to be of intangible transiency, much as a dream, since he was re-viewing the Past. But a moment's thought gave him the logical answer. The Past is definite, shaped, unalterable, as nothing else in Creation is. Therefore, to argue that he could move or alter any object here [the past] was to argue that he could change the whole history of the world or cosmos. Everything he saw about him had happened, and could not be changed in any way. On the other hand, he was fluid, movable, alterable, since *his* future still lay before him, even if it had been reversed; he was the intruder, the anomaly. In any clash between himself and the Past, the Past would prove irresistible every time."

This passage reflects the modern view that the past cannot be changed, but explains that view in a way different from that generally accepted today. Modern physicists and philosophers invoke consistency requirements (which we'll take up in the next chapter) to explain the 'solidity' of the past. The author of this story also failed to explain why his physics teacher had no trouble moving about through the air of the past, which apparently is not any more resistant to being displaced than were air molecules before time reversal occurred.

---

[36] The Austrian-British philosopher Karl Popper (1902-1994) called Boltzmann's willingness to consider the possibility that different regions of the universe could have different directions of time "staggering in its boldness and beauty," but when on to say that Boltzmann must be wrong because "it brands unidirectional change an illusion [which] makes the catastrophe of Hiroshima an illusion." That is an emotional argument, of course, and although one of great power, I fail to see how it is related to physics. See Volume 1 of *The Philosophy of Karl Popper* (P. A. Schlipp, editor), Open Court 1974, pp. 127-128.

[37] For more on Boltzmann's views on entropy, see the end of his letter "On Certain Questions of the Theory of Gases," *Nature*, February 28, 1895, pp. 413-415.

[38] C. F. Hall, "The Man Who Lived Backwards," *Tales of Wonder*, Summer 1938. The modern classic of a time-reversed world is Philip K. Dick's 1967 novel *Counter-Clock World*. We'll encounter another time-reversed world again in Chapter 4.

It didn't take long for science fiction writers to incorporate entropy as time's arrow into time travel. In one early tale there is the brief statement that entropy is behind the operation of its gadget.[39] And a few years later the inventor of a "warp gun" tells us that "The stupendous distortion of the warp may actually bring about a sort of kink in spacetime, and result in a reversal of entropy"[40] and, sure enough, when the gun is fired a woman, who is hit by the warp, ages 70 years in seconds (which is, of course, exactly the opposite of what we would expect from a "reversal of entropy"!). Just a year later, the story of a college student about to flunk his senior physics course appeared.[41] An examination is scheduled for the following day, but he needs a week and a half of study time. To his rescue comes ENTROPY, INC., a company that sells time by placing its clients inside a "time-cabinet" in which the local entropy is greatly accelerated. To someone looking through a window at the interior of the time-cabinet, the occupants would appear as characters in a speeded-up movie. Referring to Eddington by name, the author tells us that "entropy is what makes time irreversible — is what gives us the feeling of the flow of time."[42]

In a hilarious, melodramatic story featuring one of early science fiction's stereotypical 'mad scientists,' the entropic arrow of time is the scientific explanation for time travel.[43] There we read of Bryce Field, "a master-scientist, a demon, cruel, ruthless," who is rejected in love by the stupendously beautiful Lucy Grantham. Her lack of enthusiasm is perhaps understandable, as Bryce is described as having "a lean-jawed, sunken-eyed" appearance, along with "lank, untidy hair sprawled across his massive forehead." As Lucy tells him at one point, "I could never love you; you are too clever, too brilliantly scientific." After hearing that, it is no surprise that before we are more than a page or two into the tale that we learn Bryce has Lucy strapped to a steel table in an underground laboratory-in-a-cave. There he tells her of her fate: "You are going on a long journey, my dear. So long a journey that even I, master-scientist, do not know when it will end. A journey into the future — alone! ... You, Lucy, shall be the victim of entropy! ... I have discovered how to make a [globe] of non-time. Entropy will be halted ... You will be plunged into an eternal 'now.'"

And so the mad Doctor Field throws the switch on the wall of his "instrument-littered" cave on July 17, 1941, and Lucy remains "suspended" in time until the outside world reaches the date of August 9, 2450. That is the day she is at last dug-up from the cave by "big and muscular" engineer Clem Bradley and his "square-jawed" sidekick Buck Cardew, who uses a "warp in spacetime" to release Lucy from her "globe of non-time."

---

[39] M. J. Breuer, "The Time Valve," *Wonder Stories*, July 1930.

[40] F. B. Long, "Temporary Warp," *Astounding Stories*, August 1937.

[41] R. M. Farley, "Time for Sale," *Amazing Stories*, August 1938.

[42] Also citing Eddington was a tale by D. W. O'Brien, "The Man Who Lived Next Week," *Amazing Stories*, March 1941, which uses entropy to explain time travel. This curious story has the traveler arriving in the future with his clothing aged, which later 'de-ages' when the return trip is made!

[43] P. Cross, "Prisoner of Time," *Super Science Stories*, May 1942.

A few years later entropy was used in a similar but vastly more ‘scientific’ way. In that tale[44] we read of a scientist who has discovered “a field in which entropy was held level.” As the reader is told, “An object in such a field could not experience any time flow — for it, time would not exist,” since time flow is a *change* in entropy, and the ‘change’ of a level (or constant) field is zero. This interesting tale speculates on how such a field could have fantastic home uses (“Imagine cooking a chicken dinner, putting it in the field, and taking it out piping hot whenever needed, maybe twenty years hence!”). But its real use in the story is as a stasis generator for preserving fatally ill people until medical science has learned how to cure their diseases. This is, then, a high-tech method of suspended animation, of time travel into the future that is different from simply freezing (a clock in such a field would not age or measure the passage of personal time).

The gadget that does all this is called, somewhat sinisterly, the “Crypt,” which we are told also makes a great bomb shelter, too, because “not even an atom bomb could penetrate a stasis field.” The reason for that is intriguing: “The field requires a finite time in which to collapse — only there is no time in it.” The interior of the Crypt is, quite literally, a frozen block of time more rigid and unyielding than the strongest steel.

As science fiction left the age of pulps and moved into the modern era, entropy continued to be useful a justification for time travel. Arthur C. Clarke used it,[45] as did Robert Silverberg. This last tale[46] is particularly interesting, as Silverberg pursued entropy beyond simply invoking it as a mere casual throwaway mention. When a newspaper from the future appears on people’s doorsteps, the initial astonishment is replaced with puzzlement as the papers rapidly disintegrate. That is the result (we are told) of “entropic creep.” The explanation continues, informing us that it is sort of like a strain in a geological fault (Silverberg has lived for decades in California, now and then a place of large to huge earthquakes, and it isn’t surprising that he uses this particular imagery): “Entropy you know is the natural tendency of everything in nature to come apart at the seams as time goes along. These newspapers must be subject to unusually strong entropic strains because of their anomalous position out of their proper place in time.”

Earlier I mentioned we had *two* puzzles associated with entropy; we’ve discussed ‘reversibility,’ so what’s the other one, the ‘recurrence paradox,’ all about? The recurrence paradox is quite different from reversibility; it is based on a result established in 1890 by the great French mathematician Henri Poincaré (1854–1912). Motivated by the question of the stability of the motion of three masses governed by Newton’s laws of mechanics (think, for example, of the Sun, the Earth, and the Moon), Poincaré showed that starting from almost any initial state, any fixed volume system with a finite amount of energy and a finite number of degrees of freedom will return *infinitely often and with arbitrarily little deviation* to almost every previous state. If you wait long enough, implies Poincaré's astonishing theorem, Pearl Harbor will happen again—and again, and again, and .... In 1896 the German mathematician Ernst Zermelo (1871–1953) used this result, which philosophers call the 'eternal return,' to claim that there could be no truly irreversible processes and thereby cast doubt on the idea that entropy *always* and *inexorably* increases.

---

[44] P. Anderson, “Time Heals,” *Astounding Science Fiction*, October 1949.

[45] In, for example, his story of the tragic end of a geologist fifty million years in the past: “Time’s Arrow,” *Science Fantasy*, Summer 1950.

[46] “What We Learned from This Morning’s Newspaper,” *Infinity 4*, November 1972.

Even for very small systems, however, such as a mere handful of molecules, the recurrence time is extremely large, and this was, in essence, Boltzmann's reply to Zermelo's concern. For example, if the gas-filled cylinder of our entropic clock has just 100 molecules (not the six I used in the earlier example), and if transitions from one side of the cylinder to the other side take place at the rate of one million per second, then the recurrence time has been calculated to be something like 30 million billion years![47] And for the universe itself, the recurrence time is simply incomprehensible. Mathematicians call 1 followed by a hundred zeros a *googol*, and the recurrence time in years for the universe has been estimated to be 1 followed by a googol of zeros (a so-called *googolplex* of years).[48]

Using a wonderful bit of imagery, one analyst wrote of the enormity of the recurrence time of a system considerably *less* complex than the universe this way: "If a man shuffled just a single pack of cards as rapidly as an individual molecule hits other molecules in air, and if a snail started to crawl around the universe ... at the rate of one centimeter *during the life of the sidereal system* [my emphasis], the snail would have got round the universe many millions of times before it would become at all likely that the man would have got the pack back to the original order."[49] If this is what it takes to get a pack of cards back to its initial state, then try to conceive of the time interval required to restore *the world* to December 7, 1941.[50]

The notion of eternal recurrence considerably predates Poincaré, and its scientific (as opposed to astrological) study can be traced back to the fourteenth century.[51] A 'more recent' claim for eternal recurrence, also based on scientific arguments (conservation of energy), can be found in many places in the writings of the German philosopher Friedrich Nietzsche (1844–1900)—see, for example, his *The Gay Science* (1882) and *Thus Spake Zarathustra* (1883)—again predating

---

[47] J. M. Blatt, "Time Reversal," *Scientific American*, August 1956.

[48] The googol is a gigantic number, far greater than the number of raindrops that have fallen on the Earth during its entire history. And the googolplex is light years beyond that.

[49] R. B. Braithwaite, "Professor Eddington's Gifford Lectures," *Mind*, October 1929, pp. 409-435.

[50] Pulp science fiction writers, of course, were not discouraged by such calculations, as they depended on the *certainty* of recurrence over *infinite* time. See, for example, S. G. Weinbaum, "The Circle of Zero," *Thrilling Wonder Stories*, August 1936, and L. D. Gunn, "The Time Twin," *Thrilling Wonder Stories*, August 1939.

[51] R. Small, "Incommensurability and Recurrence: From Oresme to Simmel," *Journal of the History of Ideas*, January-March 1991, pp. 121-137.

Poincaré. All of Nietzsche's arguments are flawed,[52] but they are rational, physical arguments, as opposed to arguments based on metaphysics or theology. In fiction, a glimmer of the idea of a repetition of human affairs preceded Poincaré by some years, too.[53]

An important caveat concerning recurrence is that we could never *know* of it because the state of all the historical records (geological, memories, books, photographs, and so on) would, as part of the physical state of the universe, also recur. And so those records could, up to the instant *before* the recurrence, contain no signature *of* the recurrence because the recurrence has not 'yet' happened![54] The 1993 movie *Groundhog Day* stumbles on this point, as it has a character (for some unexplained reason) live through the same day over-and-over *and he is aware he is doing that*. Indeed, he can change events within that time loop at will. It is interesting to note that one 'time loop' pulp science fiction tale[55] specifically avoided that error (and cited Nietzsche, to boot), and so demonstrated that pulp science fiction *could* have some philosophical merit to it.

While the enormous recurrence time for the universe may seem reason enough to reject the possibility of Pearl Harbor repeating, there are more fundamental reasons for such a rejection. For example, an *expanding* universe, such as the one we live in, violates the Poincaré theorem's assumed condition of a *fixed* volume system. As Professor Eddington put it in a 1934 lecture at Cornell University, "In an expanding space any particular congruence becomes more and more improbable. The expansion of the Universe creates new possibilities of distribution faster than the atoms can work through them, and there is no longer any likelihood of a particular distribution being repeated. If we continue shuffling a pack of cards we are bound sometime to bring them into their standard form — but not if the conditions are that every morning one more card is added to the pack."[56]

An even more direct way to escape Poincaré's theorem is to use a result from general relativity. Using Einstein's theory instead of the classical dynamics that Poincaré used, it has been shown (by Frank Tipler, the inventor of the rotating cylinder time machine spacetime that was mentioned in the previous chapters and which we'll revisit later in the book) that the recurrence theorem is simply no longer true.[57] As Tipler wrote, "In general relativity, singularities intervene to prevent recurrence. General relativistic Universes are thought to begin and end in singularities of infinite spacetime curvature [the Big Bang and the Big Crunch, respectively], and these singularities force time in general relativity to be linear rather than cyclic." A twist to this, however, is that in his analysis Tipler assumed that gravity is *always* attractive, and that the spacetime satisfies a special condition (called the *Cauchy condition* that we'll take-up later) that avoids backward causation. The first assumption is violated in wormhole time machine spacetimes, though, and the second is *by definition* violated in *any* spacetime that supports time travel! So, who knows . . .... ?

---

[52] J. Krueger, "Nietzschean Recurrence as a Cosmological Hypothesis," *Journal of the History of Philosophy*, October 1978, pp. 435-444.

[53] See "Human Repetends" by Marcus Clarke (1846-1881), a story originally published in 1872 and reprinted *Australian Science Fiction* (V. Ikin, editor), Academy Chicago 1984.

[54] For more on this point, see D. W. Theobald, "On the Recurrence of Things Past," *Mind*, January 1976, pp. 107-111.

[55] C. F. Ksanda, "Forever Is Today," *Thrilling Wonder Stories*, Summer 1946.

[56] A. S. Eddington, "The End of the World," in *New Pathways in Science*, Macmillan 1935.

[57] F. J. Tipler, "General Relativity and the Eternal Return," in *Essays in General Relativity* (F. J. Tipler, editor), Academic Press 1980.

Despite all of the previous discussion it is *not* true that the evolution of a system from past to future is *always* accompanied by an increase in entropy—that is, by an irreversible increase in some measure of the system's 'disorder.' Yes, it can be calculated that entropy is *very likely* to monotonically increase in systems of macroscopic size, but that is not the same as certainty. There *can* be fluctuations in the thermodynamic evolution of a system so as to have, at least for a while, a *decrease* in entropy (take another look at Fig. 3.2). All we can say, for sure, is that for macroscopically sized systems even very small fluctuations in increasing entropy are most improbable. To quote no less an authority than the combined genius of Gilbert and Sullivan (from their opera *H. M. S. Pinafore*), here's what we can honestly say of the possibility of failure in the supposed inexorable increase of entropy: "What, never?/No, never!/What, never?/Well, hardly ever." Still, for physicists, entropy is just too useful a concept to give up even though it does not *always* increase with increasing time for an isolated system.

Love it though they may, there are some puzzling aspects to entropy for physicists that remain to this day. For example, the idea that the universe began in some sort of Big Bang process 15 billion years or so ago is the generally accepted view today, The puzzle of that event, one that has been described as literally being a 'fireball explosion,' is that it must have been *fantastically* hot. This means that at the beginning (of everything) there was complete thermodynamic disorder, which from our earlier discussion means *maximum* entropy. Thus, we immediately have the question of how can the entropy of the universe be continuously increasing if it was as large as possible right from the start?[58]

One possible answer is that the proper model of the universe to use is the so-called *inflationary* universe. The inflationary model has a very high expansion rate for the early universe, much higher than the rate in the standard hot Big Bang model. In the standard model, the entropy puzzle occurs because of the ability of all particle processes to readjust rapidly to the ever-changing state of the universe; the so-called *relaxation times* of all particle processes were *short* compared to the expansion rate of the universe. That means that the actual entropy of the universe would, indeed, be the maximum possible at every instant (and so we have the entropy puzzle). In the inflationary model, however, the expansion rate of the early universe was temporarily so high that the relaxation times of particle processes were *very long* compared to the expansion rate. That means the maximum *possible* entropy of the universe, at every instant, would greatly exceed the *actual* entropy. This 'entropy gap' is the cause, then, of the thermodynamic arrow of time, as the universe tries to 'catch-up' and reduce the resulting entropy deficiency.

---

[58] It has been estimated that over the next 10^{116} years the entropy of the universe will increase by a factor in excess of 10^{14}. See S. Frautschi, "Entropy in an Expanding Universe," *Science*, August 13, 1982, pp. 593-599.

There is also a philosophical problem with associating increasing entropy with the flow of time from the past into the future. Events in the past leave traces, artifacts taken to be ordered states—or at least more ordered than are their immediate surroundings. The classic example of this is a footprint in the sand, which is clearly a highly organized structure compared to the surrounding sandy beach. The footprint is the trace of a *past* event; such a trace was all the evidence, for example, that Robinson Crusoe needed to conclude that another human had walked that way. But now consider this famous counter-example,[59] that of a bombed city. Certainly there are traces aplenty of past bombing, and in fact one has to be careful not to trip over or to fall into them! The puzzle, of course, is in trying to argue that random bomb craters, strewn rubble, and crushed buildings, somehow constitute a more organized state (a 'footprint') than did the original city and its surrounding undamaged areas. This fuzziness was captured by one physicist who asked "If it were found that the entropy of the universe were decreasing, would one say that time was flowing backward, or would one say that it is a law of nature that entropy decreases with time?"[60]

For another example of the fuzziness of the relationship between entropy and time, consider the situation[61] of a cloud of non-colliding particles all initially moving toward each other. At first the radius of the smallest sphere that contains the cloud decreases with time but, eventually, as the particles move past another, the radius will grow without bound. Indeed, that inexorable increase of the radius could be taken as defining the direction of time that points toward the future. But in what sense is the *disorder* of the particle cloud increasing? After all, as the cloud expands it 'looks the same' at all times; only its scale (radius) changes. What has entropy to do with this expanding-into-the-future cloud? Perhaps nothing. Perhaps what is need is a *new* arrow of time.

So far we have looked in some detail at two arrows of time: the subjective, psychological feeling we have of time 'flowing,' which has no explanation in physics, and the thermodynamic, statistical quantity of entropy. A third arrow is the so-called *cosmological arrow* of the expansion of the universe. This arrow is not nearly as obvious as the first two. Only in the last century (since the 1920s), as a result of the American astronomer Edwin Hubble (1889–1953), has science become aware that the universe *is* expanding. An interesting speculation about the thermodynamic and cosmological arrows, one made numerous times, is that if the cosmological arrow should ever reverse—that is, if the universe should ever begin to contract toward a Big Crunch—then the thermodynamic arrow would also reverse.[62] The reasoning is that the thermodynamic arrow *follows* the cosmological arrow in an *expanding* universe because that universe can continually 'swallow-up' ever more electromagnetic radiation as it is produced by any physical process. *If* the thermodynamic arrow continues to follow the direction of the cosmological arrow during a contraction, *then* the thermodynamic arrow would also reverse direction.

---

[59] See note 54 in Chapter 2.

[60] P. W. Bridgeman, *Reflections of a Physicist*, Philosophical Library 1955, p. 251.

[61] Taken from K. G. Denbigh, "The Many Faces of Irreversibility," *British Journal for the Philosophy of Science*, December 1989, pp. 501-518.

The usual objection to that suggestion is straightforward. If the direction of time did reverse, then we would see (so goes this argument) all sorts of odd events that would require enormously improbable physics, such as a shattered glass mirror reassembling itself. The error in that objection is subtle but equally simple. *It presupposes the retarded causality of our expanding universe*. In a contracting universe with a reversed thermodynamic arrow of time, however, there would be *advanced* causality, and thus there would be nothing at all improbable about such doings as self-assembling mirrors. As two physicists observed, "The mere reversal of the cosmological expansion will not of itself serve to reverse the direction of thermodynamic and electrodynamic processes, any more than the compression phase of a piston-and-cylinder cycle in a heat engine serves to reduce the entropy of the confined gas."[63]

Those same physicists go on to then mention Stephen Hawking's interest in the relationships among the various temporal arrows. At one time Hawking thought[64] he had discovered a connection between the thermodynamic and cosmological arrows, but then later abandoned that claim.[65] Hawking, in fact, has labeled his original claim "my greatest mistake in science," and has quite openly (and most entertainingly!) discussed his interest in the arrows of time.[66] Indeed, it was to be the subject of his doctoral dissertation but, as he wrote, "I . . . needed something more definite, and less airy fairy than the arrow of time, for my PhD, and I therefore switched to singularities and black holes. They were a lot easier."

Yet another arrow of time is the *electromagnetic arrow*, which refers to the fact that radio waves are observed to only propagate into the future, and never into the past. This is a mysterious fact, because Maxwell's equations for the electromagnetic field, like all the other laws of physics, have no intrinsic time sense. The electromagnetic arrow will be discussed in some detail in Chap. 4.

---

[62] T. Gold, "The Arrow of Time," *American Journal of Physics*, June 1962, pp. 403-410.

[63] P. C. W. Davies and J. Twamley, "Time-Symmetric Cosmology and the Opacity of the Future Light Cone," *Classical and Quantum Gravity*, May 1993, pp. 931-945.

[64] S. Hawking, "Arrow of Time in Cosmology," *Physical Review D*, November 15, 1985, pp. 2489-2495. See also the next paper in the same journal, D. N. Page, "Will Entropy Decrease if the Universe Recollapses?," pp. 2496-2499.

[65] For *why* he abandoned that claim, see S. Hawking *et al.*, "Origin of Time Asymmetry," *Physical Review D*, June 15, 1993, pp. 5342-5356.

[66] S. W. Hawking, "The No Boundary Condition and the Arrow of Time,' in *Physical Origins of Time Asymmetry* (J. J. Halliwell *et al.*, editors), Cambridge University Press 1994.

3.3 Time Dilation

> "Time as we know it is not universally absolute. The rate of its passage depends to a great extent upon the velocity of its observer with regard to some certain reference system. A moving clock will run slower with respect to a selected coordinate system than a stationary one."
>
> —an early science fiction time traveler explains how his time machine works[67]

In this section I'll set the stage for the scientific basis of time travel to the future, as well as for time travel to the past via the warped spacetime called a wormhole. We start by imagining two horizontal, parallel mirrors, one positioned over the other and separated by distance d. The two mirrors are in the same frame of reference with an Observer; that is, the Observer is looking at two mirrors that are stationary with respect to him. Between the two mirrors we further imagine that a particle of light, a photon, is bouncing endlessly back and forth, up and down, in relentless reflection. This simple system is called a *photon clock*, or the Einstein-Langevin clock, after the French physicist Paul Langevin (1872–1946), and it has been part of physics for decades. We define the time required for the photon to travel from one mirror to the other as a *tick* in time, and so the return trip defines the clock's *tock*. The rate of timekeeping measured by the Observer, the time interval separating consecutive ticks, is obviously then given by

$$t' = 2\frac{d}{c}$$

where c is the speed of light.

Suppose we next imagine that the Observer and the photon clock move at constant speed v to the right across *our* line-of-sight. That is, we remain in the original frame while the photon clock and the Observer are now moving at speed v relative to us. This means the photon clock is in a different frame of reference from ours and so *we* do not see the photon bouncing up and down vertically, but rather *we* see the photon tracing out the triangular path shown in Fig. 3.3.

A round trip of the photon evidently now requires more time than before because the distance in the stationary frame (our frame) is greater than the round trip distance in the Observer's frame (moving with the photon clock, he still sees a round trip distance of d). In fact, if t is the time between consecutive ticks as seen by a stationary viewer (us), then the round trip path length of the photon that we see is

$$2\sqrt{d^2 + \left(\frac{vt}{2}\right)^2}$$

---

[67] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

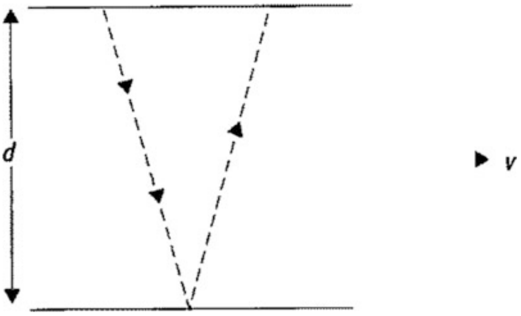

Fig. 3.3 The moving (relative to us) photon clock

and so

$$t = \frac{2\sqrt{d^2 + \left(\frac{vt}{2}\right)^2}}{c}.$$

This can be easily and quickly manipulated algebraically and combined with the earlier expression for t', the tick interval for an Observer in the same frame as the photon clock, to give the tick interval for the moving clock *as measured by a stationary viewer* (us):

$$t = \frac{t'}{\sqrt{1 - \left(\frac{v}{c}\right)^2}}.$$

Notice that this reduces to $t = t'$ when $v = 0$—that is, when the photon clock is stationary with respect to us.

This last result is the famous Einstein time dilation formula, which shows that $t \geq t'$, and indeed that $t = \infty$ when $v = c$. That is, to us the moving photon clock appears to *run slow* compared to clocks in our stationary frame of reference and, at the speed of light, time *stands still*.[68] A curious anticipation of this association between light and timelessness can be found in a poem by the seventeenth-century poet Henry Vaughn who, in the opening words to his "The World"—which appeared in 1650 as part of his *Silex Scintillans* ("Sparking Flint")—wrote

"I saw Eternity the other night
Like a great *Ring* of pure and endless light,
All calm, as it was bright,
And round beneath it, Time in hours, days, years
Driven by the spheres
Like a vast shadow moved, in which the world
And all her train were hurled."

---

[68] For $v > c$ the time dilation formula says that time becomes imaginary, and this is one reason for claiming that $v > c$ is not possible. The time dilation formula has been experimentally verified: see H. E. Ives and G. R. Stilwell, "An Experimental Study of the Rate of a Moving Atomic Clock," *Journal of the Optical Society of America*, July 1938, pp. 215-226.

A similar modification in the length of a moving object (measured in the direction of motion) occurs when $v > 0$. While an Observer moving with the object will measure its length to be L', a stationary viewer will 'report' it to be contracted to the length

$$L = L'\sqrt{1 - \left(\frac{v}{c}\right)^2}.$$

This effect is called the *Lorentz-FitzGerald contraction*.[69] For 'everyday' objects and speeds the contraction effect is an extremely small one. For example, for a low-altitude satellite 100 m long, moving at 18,000 miles per hour (that is, at $v = 2.7 \times 10^{-5}c$), the contraction is less than 4×10^{-6} cm.

In the early days of science fiction the contraction effect was fascinating to readers, but authors often got it wrong. For example, in one story[70] of a runaway spaceship falling into the Sun, we read "When our racing [ship] was drawn from the Earth's gravity and fell at ever increasing speed toward the Sun it soon approached the speed of light. As we fell faster and faster our length in the direction of the Sun progressed into nothingness. Then — it reached the speed of light — passed it. Now — mind you this — when the [ship] attained the speed of light it was of a *minus* length." This author has managed to make four errors in three sentences!

The same author botched the Lorentz-FitzGerald contraction again 5 years later, and added yet more errors to his growing list. In that tale[71] there is an episode of faster-than-light radio communication along with a lengthy, unfortunate dissertation that actually denies special relativity's fundamental assertion that all inertial frames of references are indistinguishable from each other (two frames are *inertial* if they have no relative acceleration—I'll say more on this in the next section). And in yet another story[72] of high-speed space travel the author has the contraction working in the wrong direction—as the rocket ship moves faster and faster it gets longer and longer.

As bad as those errors are, first prize for mangling the laws of physics has to go to the story[73] of a near light-speed spaceship on its way to Alpha Centauri. The crew mutinies and puts the captain and first office 'overboard' (think *Mutiny on the Bounty*) with 6 months' worth of provisions. This happens at mid-voyage, about 2 light-years from both home and destination, so matters look grim. Indeed, the author tells his readers, several times, that things look *very* bad. But are they? With a stated speed of 162,000 miles per second, the time dilation factor is slightly more than 2 and so, because the space boat is traveling at $0.87c$, it will take a little more

---

[69] Named after the Dutch physicist H. A. Lorentz (1853-1928) and the Irish physicist G. F. FitzGerald (1851-1901).

[70] J. H. Haggard, "Faster Than Light," *Wonder Stories*, October 1930.

[71] J. H. Haggard, "Relativity to the Rescue," *Amazing Stories*, April 1935.

[72] D. Wandrei, "A Race Through Time," *Astounding Stories*, October 1933.

[73] N. Schachner, "Reverse Universe," *Astounding Stories*, June 1936.

Table 3.1 The Lorentz-FitzGerald time slowing factor

1	$\frac{1}{\sqrt{1-\left(\frac{V}{C}\right)^2}}$
.1	1.005
.2	1.021
.5	1.155
.7	1.4
.9	2.294
.999	22.366
.9999	70.712

than 13 months of *space boat time* to complete the journey. If the men go on half-rations then it seems they *could* survive.

There is, of course, the problem of slowing down so as to arrive at Alpha Centauri at a reasonable speed, but that issue is ignored in the story. Instead, our attention is directed to the much more dramatic concern of a faster-than-light planet (don't ask!) colliding with the space boat and carrying the castaways onwards toward their destination. When this happens we read that time runs backwards (for what *really* occurs at superluminal speeds, keep reading this chapter) and, finally, in a repeat of an error I mentioned earlier, we are told that the Lorentz-FitzGerald contraction is negative for $v > c$.

The time-slowing (or size-shrinking) factor becomes pronounced only at values of v close to c, as shown in Table 3.1. For example, the last entry shows that a clock traveling at 99.99 % the speed of light will register the passage of 1 year while nearly 71 years pass on Earth. One science fiction writer got this dramatically wrong, even though he actually reproduced the Lorentz-FitzGerald equation in his story.[74] At one point he writes of the near light-speed rocket ship that stars in the tale, "If it [the ship's speed] was as slow as ninety-four percent [of the speed of light] . . . for every moment ticked by the clocks of the [ship] hundreds passed on earth." In fact, the time dilation factor at that speed is 'only' 2.93.

One possible objection to time dilation is that the analysis done here has been for a *particular* clock. How do we know that another clock, one using wheels and pendulums, for example, instead of photons and mirrors, wouldn't be affected differently by motion? The answer comes from relativity itself, which says there is no way to detect uniform motion. If two clocks did behave differently, then this difference could be used as a motion detector. Since this is impossible within the framework of relativity, then *all* clocks, no matter what the details of their internal mechanisms may be (including the *biological* clocks of own bodies), must respond to motion just as does the photon clock.[75]

---

[74] L. R. Hubbard, "To the Stars," *Astounding Science Fiction*, February and March 1950.

Time dilation can also be caused by gravity (it appears in the 2014 movie *Interstellar*), and that effect has been used to 'construct' a time machine from a wormhole (to be discussed later). You can get a qualitative understanding of how that happens by imagining a massive body (*massive*, to have a really big gravitational field) in space, on the surface of which is a hot object. That object emits electromagnetic radiation and, though it isn't essential to the following argument, further imagine that the temperature of the object is sufficiently high that some of the radiation (emitted by the very atoms of the object) is in the visible-light portion of the spectrum. Now, from elementary quantum theory we can also think of the object's atoms as emitting photons ('particles of light'), each of energy hf, where h is Planck's constant and f is the frequency in hertz (what used to be called 'cycles per second'). The higher the temperature, the higher the photon energy, and so the higher the frequency. In the visible spectrum, f is on the order of 10^{15} Hz, a frequency one *billion* times higher than commercial AM radio frequencies.

The radiating atoms can be thought of as tiny clocks, with alternate half-cycles of radiation being ticks and the half-cycles in-between being the tocks. The passage of time on the surface of the massive body can be measured by these atomic clocks in the hot, radiating object. To a distant observer, however, as she receives the photons from the hot object, the passage of surface time on the massive body will appear to occur at a reduced rate when compared with the photons emitted by her own identically hot object (her 'local' clock). That's because the radiation that arrives at the distant observer has traversed a gravitational field (a journey sometimes described as 'climbing out of a gravitational well') and so is down-shifted in frequency toward the red end of the visible spectrum. This effect is called either the *gravitational red shift* or the *gravitational time dilation* effect (or even the *Einstein shift*, because it was Einstein who predicted the effect in 1907).

You can 'understand' this dilation effect as follows. One can crudely think of a photon emitted by the hot object as something like a rock thrown upward. As the rock rises upward through the gravitational field, its *total* instantaneous energy is always constant, but the total, fixed energy is split between its kinetic and potential energies in an ever changing way. That is, as the rock rises, its kinetic energy continually decreases (the rock slows down), whereas its potential energy continually increases. A photon is not a rock, however, and it certainly can't slow down as

---

[75] Resistance to this conclusion persisted for years. See, for example, the letter "Relativity and Radio-activity," *Nature*, January 8, 1920, p. 468. The author of that letter wondered whether a clock based on radioactive decay might not somehow beat the 'conspiracy' of moving clocks running slow compared to stationary ones. And in a letter to *Science* (December 7, 1962, p. 1180), a reader objected to applying the laws of physics to biological systems, first asserting (incorrectly) that time dilation "has never been proved or disproved experimentally," and then "there is no known causal means by which greatly increased velocity could alter, without destroying the very biochemical basis of the life process, the metabolic changes which are responsible for the aging process."

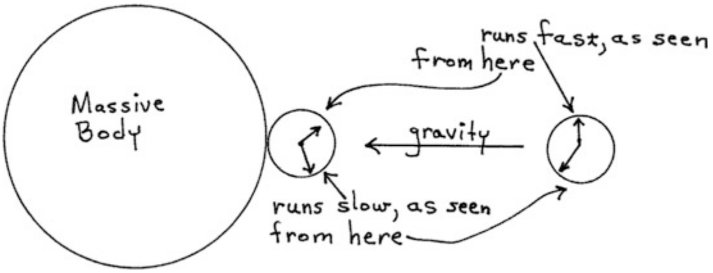

Fig. 3.4 Gravitational time dilation due to a massive body

it rises through a gravitational field (it *always* moves at the speed of light because the photon *is* light). The only way a photon can give up energy to balance the ever increasing potential energy (physicists will cringe at this, but read on) is to decrease its frequency. Hence, the red shift as seen by a distant observer of the photon, who thus sees time running slow on the massive body. Fig. 3.4 shows the case for a clock on a massive body, compared to a distant clock. Notice, carefully, the 'direction of gravity,' that is, the direction a small, unrestrained test mass will move.

A gravitational red shift in the *opposite* direction is nicely described in the famous 1966 science fiction story "Neutron Star" by Larry Niven. There a space traveler zooms down into a neutron star's intense gravity field (at half the speed of light!), passing within one mile of the star's surface. He reports what he observes in these dramatic words: "All around me were blue-white stars. Imagine light falling into a savagely steep gravitational well. It won't accelerate. Light can't move faster than light. But it can gain in energy by increasing its frequency. The light was falling on me, harder and harder, as I dropped." To Niven's intrepid spaceman, therefore, the passage of time on those distant blue-white stars appeared to be running *fast* compared to his wrist watch. This shows that the effect could equally well be called the 'gravitational *blue* shift.'

Notice that *gravitational*-induced time alterations do *not* have the symmetrical feature of motion-induced time dilations.[76] That is, for gravitational time dilations caused by photons either falling into or climbing out of gravity wells, observers at each end *agree* about whose clock is running slow, unlike in the motion-induced case where *each* of the relatively moving observers thinks it is the *other* observer's clock that is running *slow*.[77]

Now, comparing a massless photon to a rock which does have mass (and so potential energy), as each travels 'against' gravity, *is* straining the physics, with its one virtue being the provision of an initial plausibility argument. Gravitational time dilation is sufficiently important in the operation of wormhole time machines,

---

[76] Gravitational time dilation was experimentally observed in 1960, more than half a century after Einstein predicted it.

[77] A science fiction use of both the red and the blue gravitational shifts appears in the novel by J. P. Hogan, *Out of Time*, Bantam 1993.

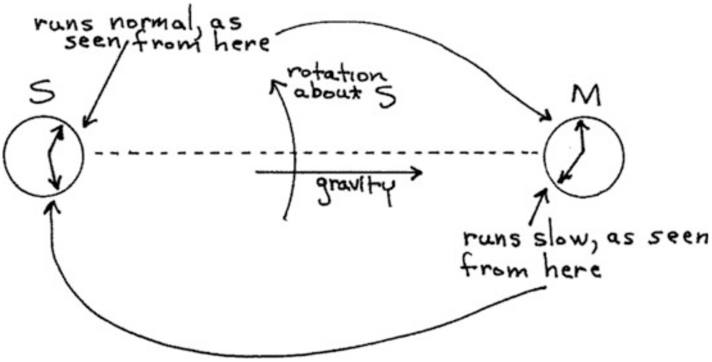

Fig. 3.5 Gravitational time dilation due to rotation

however, that perhaps another way to think about it that is more acceptable to hardcore physicists makes it worth another look. Imagine a turn-table disk that initially is *not* rotating (soon it *will* spin). On this disk imagine further that we fasten two clocks, one called S (for *stationary*) and one called M (for *moving*), as shown in Fig. 3.5. These two clocks are set to read the same time at some instant and then, thereafter, they tick-tock through time at precisely the same rate. We then start the disk rotating around a vertical axis through clock S. (Imagine S to be a *point* clock, and so it is *the* one point on the turn-table that remains at rest even as the disk rotates.) What happens to the time-keeping of S and M?

We can answer this question by using our earlier result concerning time dilation due to motion. Even with the disk now rotating, S appears stationary to an observer sitting on top of M, while M appears to be moving across the line of sight of an observer sitting on top of S. So, to the observer at S, clock M runs slow, while to the observer at M there is *no change* in the time keeping of clock S. This is a non-symmetrical outcome, and so should remind you of gravitational time dilation. This might be a puzzle to you, however, as we don't have a massive body in Fig. 3.5 to account for a gravity presence. This is where the genius of Einstein comes into play.

Anyone who has ever ridden on a merry-go-round knows there is an outward (pointing away from S) directed force called the *centrifugal force* that is 'trying' to toss you off the merry-go-round. Now, where there is a force there is an acceleration, and one of Einstein's starting points in his development of general relativity was to identify an acceleration, *whatever its origin*, with gravity. A massive body is, of course, one possible origin (the obvious one, in fact), but so is the rotation of the turn-table. So, we have the situation shown in Fig. 3.5, where now 'gravity' is directed as shown and, and as in Fig. 3.4, the direction of acceleration of gravity is toward the slow-running clock. Again, the direction of gravity is the direction a small, unrestrained test mass on the rotating disk will move. The further M is away from S, the greater the 'gravity' of the centrifugal acceleration and so the slower will M run as measured by the observer at S.

A 1968 story that uses the gravitational time dilation effect in a striking fashion tells of a starship's visit to a supernova, accompanied by a fantastic alien life-form—a ball of intelligent plasma named Lucifer that is telepathic.[78] While the ship stands off at a distance of 500 million kilometers, Lucifer will approach much closer to the *event horizon* (see the Glossary) of a black hole at the center of the supernova explosion and communicate its findings to a human telepath on the ship. A physicist in the crew is curious about one point, and asks the human telepath the following question:

"I have wondered about one item. Presumably Lucifer will go quite near the supernova. Can you still maintain contact with him? The time dilation effect, will that not change the frequency of his thoughts too much?"

Lucifer, in fact, dies in the black hole even as he saves the ship from destruction, and the human telepath will hear his death scream for the rest of her life. As the physicist later explains to the ship's captain, telepathy is instantaneous and has no limiting range (there is no known physical basis for believing any of this, but it is crucial for story effect):

> "Remember the time dilation. He fell from the sky and perished swiftly, yes.
>
> But in supernova time. Not the same as ours. To us, the final stellar collapse takes an infinite number of years. . . . He will always be with her."[79]

3.4 The Lorentz Transformation

> "If only he'd paid more attention to mathematics in school."
>
> —a science fiction time traveler laments missed opportunities[80]

In this (and the next) section the math gets about as 'deep' as it gets in this book, but to leave it out struck me as a cheat. *You* can skip part (or all) of the math and simply read the prose, but it seemed unfair for me to make that decision for you.

We begin by imagining two distinct frames of reference. One we take to be stationary, and the other as moving at a uniform speed v with the respect to the first. The moving frame is said to be *boosted* with respect to the stationary frame. We orient these two coordinate systems so that the motion occurs along just one axis (the x-axis, as shown in Fig. 3.6, where I am using primed variables for the moving frame). That is, the two frames have coincident x axes, and parallel y and z axes that

---

[78] Poul Anderson, "Kyrie," in *The Road to Science Fiction* (J. Gunn, editor), volume 3, New American Library 1979.

[79] A mathematical discussion of how signals take forever (even though they are emitted in a finite time interval) to travel from the event horizon of a black hole to a distant receiver can be found in James B. Hartle, *Gravity: an introduction to Einstein's general relativity*, Addison Wesley 2003, pp. 264-268.

[80] D. Knight, "Extempore," in *Far Out*, Simon and Schuster 1961. Similar words ("If only I had more mathematics") were spoken by Einstein the day before he died — see Walter Isaacson, *Einstein: his life and universe*, Simon & Schuster 2007, p. 542.

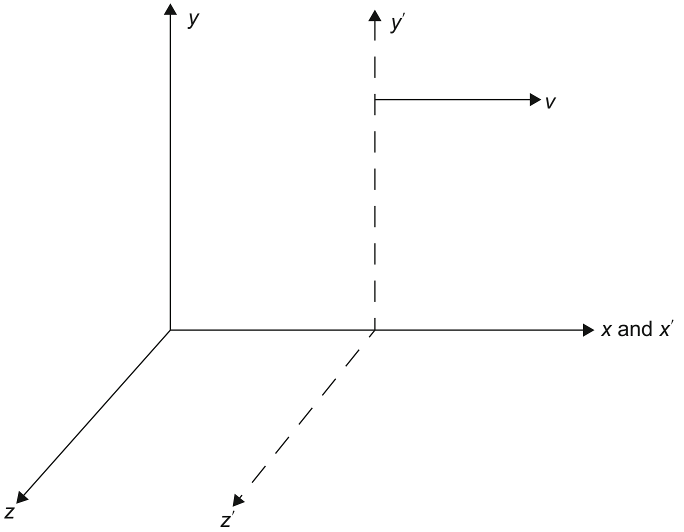

Fig. 3.6 Two reference frames in relative motion

are moving apart at the constant speed v. Let's also imagine that there is a clock at the origin of each frame, and that at the instant the origins match the clocks are synchronized; that is, $t = t' = 0$ is the instant the two coordinate systems coincide.

We further imagine that there is an observer at the origin in each frame. At some arbitrary instant of time, each observer records the coordinates of the arbitrary point P in space, as measured in his system. These observers could, for example, agree to record the coordinates of P in their system when their clock reads 5 s. It seems immediately obvious (as it was for Newton) that $t = t'$; that is, time runs at the same rate in each frame and thus it makes sense to talk about "the same instant" at every point in space. (After reading the previous section you know this *not* true, but temporally forget that!) Thus, at this 'same instant' the stationary observer records (x, y, z) and the moving observer records (x', y', z'). What are the relationships between the primed and unprimed coordinates of P? That is, what mathematical transformation converts from one frame to the other?

The answer seems obvious:

$$y' = y$$
$$z' = z$$
$$x' = x - vt.$$

This transformation, called the *Galilean transformation* after the Italian Galileo Galilei (1564–1642), satisfies the relativity principle, which says that uniform motion leaves the laws of physics unchanged. For example, in the stationary system Newton's famous second law of motion for a constant mass m,

$$F = m\frac{d^2x}{dt^2}$$

becomes the identical-appearing form

$$F' = m\frac{d^2x'}{dt'^2}.$$

More precisely, *all* the laws of mechanics known to Newton are unchanged. Any frame of reference in which Newton's laws of mechanics hold true is said to be an *inertial* frame. Given one inertial frame, we can find infinitely many others simply by applying the Galilean transformation.

However, when the mathematical laws of electrodynamics were discovered by Maxwell in the nineteenth century, it was a shock to physicists to learn that the Galilean transformation does *not* leave Maxwell's equations unchanged in form; the transformed equations predict electromagnetic effects for the moving system that are *not* predicted to occur in the stationary system. This meant that there was theoretical support for the possibility that electromagnetic experiments might be devised to detect uniform motion, and this eventually led to the famous Michelson-Morley experiment of 1887. This experiment, sensitive enough to detect the motion of the Earth itself through space, failed to detect any such motion. The conclusion was clear: the new electromagnetic effects predicted by the Galilean transformation do not exist, and so the transformation must be wrong even though it works for the laws of mechanics. So—what is going on?

The answer is inspired, and again returns us to the cornerstone of relativity: the idea that the laws of physics, *all* the laws, should look the same to observers in uniform relative motion. That is, there is no special or preferred system of coordinates—*all* inertial systems are equivalent in physics. Evidence from an extremely broad variety of sensitive experiments had, by the end of the nineteenth century, convinced physicists that Maxwell's equations *are* correct. Thus, a new transformation was needed that leaves both the laws of mechanics *and* the laws of electrodynamics unchanged with uniform motion. But, a single transformation that works on Maxwell's equations and on the mechanical laws would therefore mean that Newton's mechanical laws as stated cannot be correct, and this was a breathtaking conclusion: Newton had been unchallenged for two centuries.

As it turns out, Newton's laws are *almost* right. The only correction required is that the mass of a moving body is not independent of motion, but rather varies as

$$m = \frac{m_0}{\sqrt{1 - (v/c)^2}}$$

where m_0 is the so-called *rest mass* when $v = 0$.[81] This result says that m is infinite at $v = \mathrm{c}$ unless $m_0 = 0$ (as it is for a photon), which is another reason for the belief that accelerating a mass (such as a spaceship) up to the speed of light is impossible because it would require infinite energy (look back at note 68, too). With this modification, the transformation that leaves all the laws of physics unaltered in form by uniform motion is what is called the *Lorentz transformation* (after the same Lorentz the contraction effect is named for), who discovered it in 1904 by direct manipulation of Maxwell's equations:

$$
\begin{aligned}
y' &= y \\
z' &= z \\
x' &= \frac{x - \mathrm{vt}}{\sqrt{1 - (v/c)^2}} \\
t' &= \frac{t - \mathrm{vx}/c^2}{\sqrt{1 - (v/c)^2}}.
\end{aligned}
$$

In 1905 Einstein discovered how to derive these equations from a fundamental reexamination of space and time without concerning oneself about the details of specific physical laws.

By simple algebraic manipulation, the transformation equations can be rewritten as

$$
\begin{aligned}
ct &= \gamma ct' + \beta\gamma x' \\
x &= \beta\gamma ct' + \gamma x'
\end{aligned}
$$

where

$$\beta = \frac{v}{c}$$

and

$$\gamma = 1/\left(1 - \beta^2\right)$$

are dimensionless constants. In compact matrix form, the Lorentz transformation becomes

---

[81] This variation of mass with speed was experimentally observed in 1901.

$$\begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} ct' \\ x' \end{bmatrix}$$

and the symmetrical 2 × 2 matrix is called the Lorentz boost matrix or simply the *boost*. Notice that when $v = 0$ (zero boost) the boost matrix reduces to the identity matrix; that is, the two frames are one and the same with at most a shift in the location of the origins. Note, too, that $\beta = 0$ and $\gamma = 1$ for *any* v when c is infinite—that is, the boost matrix is again reduced to the identity matrix and the Lorentz transformation becomes the Galilean *if* c is infinite. But c is not infinite, and all the implications of special relativity are the direct result of the *finite* speed of light.

The Lorentz transformation contains two results I have mentioned earlier in the book. For example, in Chap. 2 it was mentioned that simultaneity is a relative concept in reference frames in relative motion. Let's see what the transformation says about that. Consider two events that occur specifically on the x-axis. They are simultaneous in the stationary system (at, say, time $t = T$) but are at different places (at, say, $x = X$ and $x = X + \Delta X$). Their occurrences in time for the moving observer are

$$t_1' = \frac{T - \mathrm{vX}/c^2}{\sqrt{1 - (v/c)^2}}$$

and

$$t_2' = \frac{T - \mathrm{v}(X + \Delta X)/c^2}{\sqrt{1 - (v/c)^2}}.$$

For the moving observer, therefore, the two events are *not* simultaneous, being separated in time by

$$t_1' - t_2' = \frac{v\Delta X/\mathrm{c}^2}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}}.$$

Only if $\Delta X = 0$ (the two events occur at the same place) will $t_1' = t_2'$. That is, only if $\Delta X = 0$ are simultaneous events in one frame also simultaneous in another frame in relative motion.

And in the previous section we found that time runs slow in one frame as observed from another frame that is in relative motion. We can get this result from the t' equation of the Lorentz transformation by differentiating it with respect to t. Thus,

$$\frac{dt'}{dt} = \frac{1 - (v/c^2)\frac{dx}{dt}}{\sqrt{1 - (v/c)^2}}.$$

But since

$$\frac{dx}{dt} = v,$$

the speed of the moving frame as measured by the observer in the stationary frame, this gives

$$dt' = \sqrt{1 - (v/c)^2}\,\mathrm{dt}$$

which is the same result we obtained by analyzing the photon clock.

The Lorentz transformation contains other interesting implications beyond these. For example, mention has been made several times to the ‘relativity principle,’ the belief that uniform motion has no observable effect on the forms of physical laws. But how do we know who is moving and who is stationary? After all, a system moving to the right past a stationary system could just as well be thought as the stationary system, while it’s the *other* system that is moving to the *left* (at speed $-v$).

To study this question with the Lorentz transformation, we’ll invert the transformation (that is, solve for the unprimed variables in terms of the primed ones). What we get back is just what you probably thought—the Lorentz transformation with v replaced by $-v$. That is, the Lorentz transformation is symmetrical, so two observers in different frames of reference each say it is the *other’s* clock that is running slow! This follows immediately, in fact, from the original transformation written in matrix form. That is, multiplying through the earlier matrix equation by the inverse of the boost matrix, we get

$$\begin{bmatrix} ct' \\ x' \end{bmatrix} = \begin{bmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{bmatrix}^{-1} \begin{bmatrix} ct \\ x \end{bmatrix} = \begin{bmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{bmatrix} \begin{bmatrix} ct \\ x \end{bmatrix}.$$

The only difference between the original boost matrix and its inverse (which is, of course, the new boost matrix for the new interpretation of which frame is the moving one) is a change in sign for β, that is, in the sign of v. The inverse transformation is

$$\begin{aligned} y &= y' \\ z &= z' \\ x &= \frac{\mathrm{x}' + \mathrm{vt}'}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}} \\ t &= \frac{\mathrm{t}' + \mathrm{vx}'/\mathrm{c}^2}{\sqrt{1 - (\mathrm{v}/\mathrm{c})^2}}. \end{aligned}$$

As a final example of what the Lorentz transformation tells us, consider the so-called *addition of velocities* problem. Suppose you are in a high-speed spaceship traveling past Earth at speed v. Earth is the stationary system (with the unprimed variables), and the spaceship is the moving system (with the primed variables). Assume the x and $\mathrm{x}^{'}$ axes are along the direction of motion. Imagine next that while standing in the nose of the spaceship, just as the spaceship passes Earth, you fire a gun in the direction of motion (away from the Earth), with the bullet exiting the gun with a muzzle speed of w. How fast is the bullet moving away from Earth? The common-sense answer in Galileo's time was $v+\mathrm{w}$, but we now know that the Galilean transformation is wrong. What does the Lorentz transformation say?

Inside the spaceship, the position of the bullet at time $t^{'}$ after the gun is fired is

$$x' = wt'.$$

From the inverse Lorentz transformation, the location of the bullet earth's frame is

$$x = \frac{x' + \mathrm{v}t'}{\sqrt{1-(v/c)^2}} = \frac{w+\mathrm{v}}{\sqrt{1-(v/c)^2}}t'.$$

The transformation also tells us that (using $x^{'} = wt^{'}$)

$$t = \frac{t' + \mathrm{v}x'/c^2}{\sqrt{1-(v/c)^2}} = \frac{1+\mathrm{wv}/c^2}{\sqrt{1-(v/c)^2}}t'.$$

Thus, the speed of the bullet in *Earth's frame* is

$$\frac{x}{t} = \frac{w+\mathrm{v}}{1+\mathrm{wv}/c^2}.$$

Notice that for a low-speed bullet ($w << c$) this result[82] is close to $w + v$, but at high values for w the result is very much different. Indeed, suppose we don't fire a gun at all, but rather replace it with a flashlight. Now, instead of a bullet, we shoot *photons* at $w = c$. The Galilean transformation would (incorrectly) say that a stationary observer on Earth would see the photons moving away at speed $v+\mathrm{c}$, which is a superluminal speed. The Lorentz transformation, however, says that the Earth observer would see a speed of

---

[82] This result was found by the French physicist Henri Poincaré (1854-1912) in June 1905, three months *before* the publication of Einstein's special theory of relativity which also contains the result. And it was Poincaré who first stated (in 1904) that "no velocity can surpass that of light, any more than any temperature could fall below the zero absolute."

$$\frac{c + \mathrm{v}}{1 + \mathrm{cv}/c^2} = \frac{c^2(c + \mathrm{v})}{c^2 + \mathrm{cv}} = \frac{c^2(c + \mathrm{v})}{c(c + \mathrm{v})} = c.$$

That is, no matter what the speed of the moving observer on the spaceship is, he sees the light from his flashlight traveling at the same speed as does the stationary observer back on Earth. This peculiar effect is unique to the speed of light ($w = \mathrm{c}$). We've derived it here as a consequence of the Lorentz transformation, but in fact Einstein actually did things in reverse order. That is, he began by *postulating* the invariance of the speed of light[83] for all observers in uniform motion, combined that with the principle of relativity which says all physical laws look the same to those observers, and so derived the Lorentz transformation using no mathematics beyond high school algebra.

A mathematically elegant alternative derivation of the addition-of-velocities formula can be done by simply noticing that the condition of two successive boosts should be, itself, a boost. Thus, if we have a frame moving relative to a second frame (which is itself moving relative to a third frame), then the boost matrix of the first frame relative to the third frame is the product of the two individual boost matrices. That is,

$$\begin{bmatrix} \gamma_3 & \beta_3\gamma_3 \\ \beta_3\gamma_3 & \gamma_3 \end{bmatrix} = \begin{bmatrix} \gamma_2 & \beta_2\gamma_2 \\ \beta_2\gamma_2 & \gamma_2 \end{bmatrix} \begin{bmatrix} \gamma_1 & \beta_1\gamma_1 \\ \beta_1\gamma_1 & \gamma_1 \end{bmatrix}.$$

From this it is easy to show (if you know how to multiply matrices!) that

$$\beta_3 = \frac{\beta_1 + \beta_2}{1 + \beta_1\beta_2}.$$

Substitution of

$$\beta_1 = \frac{w}{c}, \ \beta_2 = \frac{v}{c}$$

immediately gives the addition-of-velocities formula.

A failure to understand the implications of the invariance of the speed of light resulted in two stupendous errors in the story "To the Stars," cited in note 74. At all times an officer stands watch on the bridge of a near light-speed rocket ship to be sure the ship doesn't accidently *reach* the speed of light. This is to be avoided (according to the author) because to reach the speed of light would cause the ship to

---

[83] Einstein's postulate was experimentally confirmed in 1932.

"hang there forever unmoving [in time] . . . locked, protected and condemned to eternity by zero time." This horrible state is so easy to stumble into (the author was apparently unaware that it would require *infinite* energy) that occasionally the ship has to fire a 'check-blast' from its forward rocket tubes to slow down! Equally absurd is the means by which the development of this 'fatal' condition is detected: the nose of the ship mounts a forward-pointing light source (our earlier flashlight) and so, if the ship is getting too near the speed of light, it will start to *overtake* the photons emitted by that source!

The biggest puzzle of all, actually, is why the editor of *Astounding Science Fiction* let such a technically goofy story appear in a magazine recognized for its usual faithfulness to known science. Particularly so since the story appeared in 1950 and, as long ago as December 1937, none other than Isaac Asimov (then 17 years old) had written a letter to the editor[84] giving the proper interpretation of what happens when $v = c$. Here's what the young Asimov wrote (notice the early hint of his life-long pessimism concerning time travel to the past that appears near the end of his letter):

> "The effect on time of increasing speeds is . . . well known. Relativity states that as speed approaches that of light, time slows up until at 186,000 miles a second, time (so to speak) stands still. This seems to refute statements found in so many astronomy books (and science fiction stories) that even at the speed of light it would take four years to reach the nearest star. No such thing! As time halts at the speed of light, a person traveling from Alpha Centauri to the solar system, or vice-versa, would not be aware of any lapse of time. In that sense the speed of light is infinite (as was thought in ancient times). This by the way offers an entirely scientific (if impractical) means of travel into the future. Say that someone wants to see how the world would look a hundred years from now. His procedure would be as follows: getting into his spaceship, he would proceed to a spot fifty light-years away at the speed of light. The journey would, for him, be practically instantaneous (due to the curious behavior of time at the speed of light). But fifty years would have elapsed on earth. He makes the return trip at the same speed. Another fifty years lapse on earth and he lands a hundred years after his time. With this system, however, it would be impossible to travel into the past, so I don't think it will ever be adopted."

Young Asimov missed an important detail concerning the reversal of the spaceship's direction of travel for the return to Earth, and I'll come back to it later in this chapter. But certainly he displayed a *far* better knowledge of the physics of time than did the author of "To the Stars" (who was, by the way, L. Ron Hubbard, a prolific writer of fantasy and science fiction before founding the Church of Scientology).

---

[84] The editor who bought "To the Stars" was the same editor editor of *Astounding* when Asimov's letter appeared, and so he was certainly aware of it.

3.5 Spacetime Diagrams, Light Cones, Metrics, and Invariant Intervals

> "Come back when you know tensor calculus and I'll explain to you about n-dimensional forces and the warping of world-lines."
>
> —a science fiction physicist's reply after being asked how his time machine works[85]

It is helpful in discussions about the spacetime of special relativity to use what are called *Minkowski spacetime diagrams*. These are plots of the spacetime coordinates of a particle; the resulting curve is called the *world line* of the particle. Such diagrams are four-dimensional—three space axes and one time axis—and hard to visualize, much less draw on a flat sheet of paper! The convention is to make do, whenever possible, with a simplified spacetime that has just one space axis (horizontal) and one time axis (vertical). As you'll recall from Chap. 1, physicists often call such a simplified diagram a *toy spacetime*.

So, for a particle at rest in some observer's frame of reference, its spacetime diagram for that observer is a *vertical* world line. If the particle is not at rest then its world line will tilt away from the vertical; the greater the speed the greater the deviation from the vertical. Accelerated particles will have world-lines that *curve* away from the vertical. Straight, uncurved world lines represent unaccelerated particles, that is, particles experiencing no forces and so in free fall. Such a world line is called a *geodesic*. In Fig. 3.7 the world lines for these various cases are shown on the same axes. It is assumed in the figure that all three particles are at $x = x_0$ when $t = 0$.

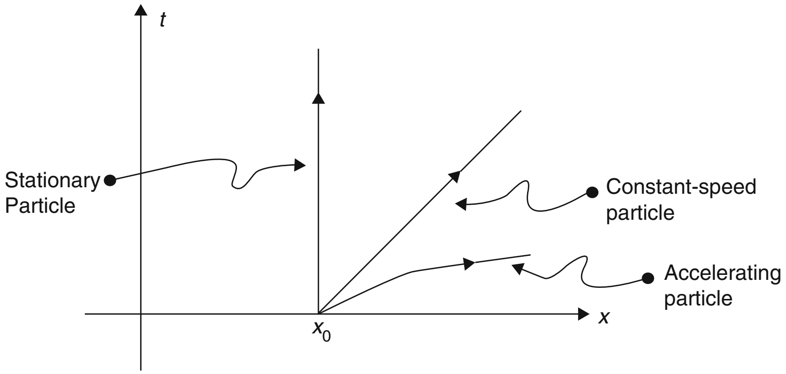

Fig. 3.7 World lines of three particles

---

[85] Poul Anderson, "The Little Monster," in *Science Fiction Adventure from WAY OUT* (R. Elwood, editor), Whitman 1973.

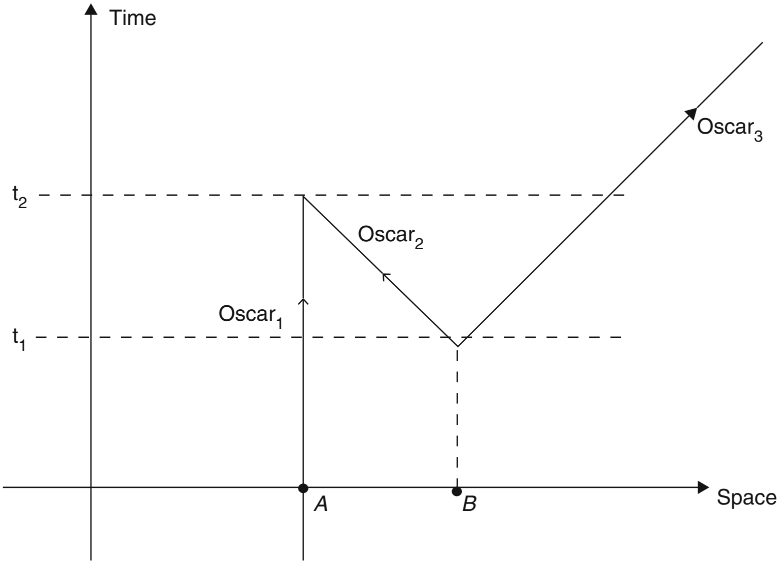

Fig. 3.8 A time traveler and his world line

Spacetime diagrams were embraced decades ago by philosophers looking for 'scientific' ways to support their position on time travel (whatever it might be), as opposed to the mere verbiage of traditional colleagues. A famous example of this is a 1962 paper by the Harvard philosophy professor Hillary Putnam (note 74 in Chap. 1). There we are asked to imagine the spacetime diagram of one Oscar Smith who, in Fig. 3.8, is at spatial location A next to his time machine. At time t_0 Oscar has not yet gotten into his time machine. A little later, at time t_1, we suddenly see not only Oscar at A but also *two more* Oscars who have appeared (apparently out of thin air) moving away from spatial location B! Between t_1 and t_2 we see the original Oscar at A and the two mysterious Oscars at B (for a total of three Oscars, labeled in the figure as Oscar$_1$, Oscar$_2$, and Oscar$_3$) move forward in time—but one of the new Oscars (Oscar$_2$) lives a decidedly odd existence in that his life seems to be running in reverse!

Eventually, at time t_2, the original Oscar$_1$ and the weird, reverse Oscar$_2$, merge and seemingly annihilate one another, vanishing into thin air to leave only a single Oscar (Oscar$_3$) for all time after t_2. Putnam argues that, although strange, what has just been described is still sensible and that, indeed, the very fact that we can draw the spacetime diagram of Fig. 3.8 supports the case for backward time travel. He claims this because although the spacetime diagram does show time increasing upward for all three Oscars (that is the time direction for an external observer) there is actually no 'spontaneous creation' or 'mutual annihilation' and all is sensible *if* Oscar$_2$ is understood actually to be a time traveler into the past with *his* time direction thus pointed opposite to that of the 'other two' Oscars. There is, of course, just *one* Oscar!

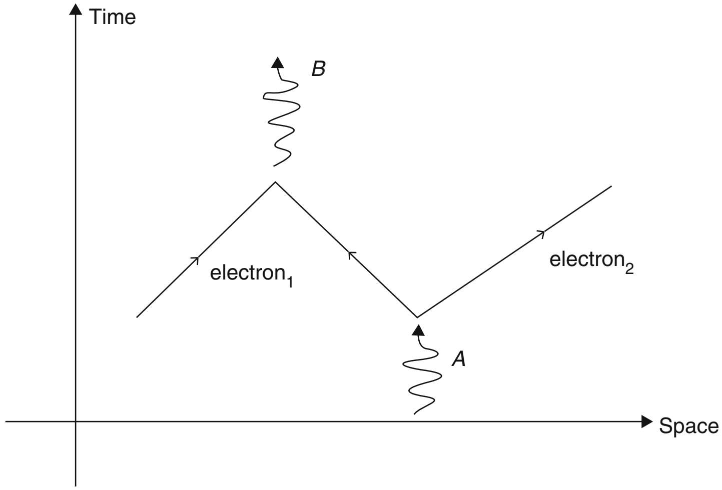

Fig. 3.9 Anti-matter via backward time travel

As mentioned back in Chap. 1, Putnam's suggestion was rebutted (see that chapter's note 76) by what is (in my opinion) an even less plausible mechanism for Putnam's kinked spacetime than is Putnam's invocation of time travel. That critic advocated, instead, an explanation based on matter transmitters and anti-matter, the latter an idea he credits to Feynman (who actually got it from Wheeler). In a paper by Feynman we do find the famous suggestion that a positron that appears to us to be moving forward in time is actually an electron traveling backward through time.[86] Logically, that greatly weakens the critic's view, because it puts him in the position of using anti-matter (explained in terms of backward time travel) to argue *against* backward time travel! But let's ignore that concern, give the critic the benefit of the doubt, and explore how anti-matter and backward time travel are imagined to be connected.

Feynman asks us to imagine the process shown in Fig. 3.9. Gamma ray A spontaneously creates an electron-positron pair, with electron$_2$ moving off to some distant region while the positron soon meets with electron$_1$, resulting in mutual annihilation and the production of gamma ray B. This description involves *three* particles, and each segment of the kinked line is a distinct particle. But Feynman said there is another way to look at this, a way that involves just *one* particle. According to Feynman, the kinked line in Fig. 3.9 (which should remind you of Oscar's kinked world line) is the world line of a single electron; the middle

---

[86] R. Feynman, "The Theory of Positrons," *Physical Review*, September 15, 1949, pp. 749-759. In a paper published the year before ("A Relativistic Cut-Off for Classical Electrodynamics," *Physical Review*, October 1948, pp. 939-946), he wrote "This idea that positrons might be electrons with the proper time reversed was suggested to me by Professor J. A. Wheeler." The identification of anti-matter with backward time travel occurred in science fiction (see note 124 in Chapter 1) almost simultaneously with Wheeler's speculation.

segment, that we call a positron, is just the electron traveling backward in time, and so we must reverse the arrowhead on it (indicating the opposite of the direction shown in Fig. 3.9).

There are two central questions at this point. First, why is a positron (with positive electric charge) moving forward in time mathematically (and physically) equivalent to a negatively charged electron moving backward in time? The answer is that the reversal in charge sign, which results from the reversal of the electron's proper time, follows from the TCP theorem that was mentioned in Chap. 2 (see note 38 there). And second, what causes the electron to suddenly move backward in time? Picturesquely, the electron is recoiling from the emitted burst of gamma ray B. Similarly, the absorption of the energy of gamma ray A by the electron that is recoiling backward in time causes a second recoil, giving the world line of what was originally called electron$_2$. This reinterpretation of a kinked spacetime diagram was described as follows (in Feynman's famous words): "It is as though a bombardier flying low over a road suddenly sees three roads and it is only when two of them come together and disappear that he realizes that he has simply passed over a long switchback in a single road."

In a later paper[87] Putnam's critic presented another line of attack against Putnam's interpretation of spacetime diagrams as lending support to time travel. There the critic observed that the presence of the time-reversed Oscar$_2$ shows that the "world of the Oscars" is not *temporally orientable*. A temporally orientable spacetime is one in which *every* point in it agrees with its local neighbors on the directions of past and future—a condition clearly *not* satisfied for the case of Oscar$_2$. As the critic pointed out, the first time travel spacetime discovered, the Gödel universe, *is* temporally orientable, and so in it the ambiguity of Oscar$_2$ (whether he is traveling backward in time as opposed to living forward 'in reverse') does not occur. That is, the critic agreed with Putnam's acceptance of the conceivability of time travel to the past, but not with his use of Feynman's concept of anti-matter as time-traveling matter. That critic wasn't alone in that opinion.

One physicist, for example, wrote of "Feynman's rather loose talk of particles 'traveling' backward . . . in time,"[88] and the well-known philosopher John Earman declared "It is true that Feynman uses the slogan 'Positrons are electrons running backward in time,' but it is dangerous to draw conclusions from slogans."[89] I am not sure what Earman meant by "slogans": a careful reading of Feynman indicates that he actually took the matter quite seriously.[90] In his 1949 positron paper (note 86), for example, he wrote that "the idea that positrons can be represented as electrons with proper time reversed relative to true time has been discussed by the author and others,"[91] and also that "Previous results suggest waves propagating . . . toward the past, and that such waves represent the propagation of a positron."

---

[87] See R. Weingard in note 114 in Chapter 1.

[88] H. Price, "The Asymmetry of Radiation: Reinterpreting the Wheeler-Feynman Argument," *Foundations of Physics*, August 1991, pp. 959-975.

[89] J. Earman, "On Going Backward in Time," *Philosophy of Science*, September 1967, pp. 211-222.

[90] Feynman declares the view of a positron as a time traveling electron to be of value, for example, in his famous book *Quantum Electrodynamics*, W. A. Benjamin 1961, p. 68.

In any case, spacetime diagrams are highly useful in discussing time travel, but they do have some curious twists. In our everyday world, a path that joins two points on a surface with the *minimum* length is called a geodesic of that surface. As you'll see later in this chapter spacetime geodesics do indeed possess an extremal property, but rather than being a minimum it is a *maximum* property. Spacetime diagrams can be misleading on this matter, so it is important to remember that such diagrams are not a perfect representation of all the properties of a spacetime.

It is customary to draw spacetime diagrams with the speed of light as unity ($c = 1$). That is, a distance of 300,000 km on the space axis is represented by the same extension as is one second on the time axis. This means that the world line of a photon is tilted away from the vertical time axis by 45°. Because photons can travel in both space directions (to the left and to the right) in the two-dimensional spacetime we can draw on a piece of paper, and because the speed of light is the limiting speed of the universe, we can represent the collection of all possible world lines as those paths that *never* tilt more than 45° away from the vertical, which forms what is called a *light cone* in spacetime, as shown in Fig. 3.10 (which attempts to represent a *three*-dimensional spacetime, one with two space—imagine a y-axis, out of the paper, perpendicular to both the x and t axes—and one time dimension).

In Fig. 3.10 I have taken $x = y = 0$ at $t = 0$ for all the possible world lines involving speeds below the speed of light. Let's agree to call this spacetime point the Here-Now. Then, spacetime points in the upward half of the light cone are in the Future of Here-Now; similarly, spacetime points in the lower half of the light cone are in the Past of the Here-Now. We can draw a straight world line from the Here-Now to any point in the Future half-cone with a tilt of less than 45° away from the vertical, which means that a massive particle could travel from the Here-Now to that point at less than the speed of light. Similarly, a massive particle starting at any point in the Past half-cone could have reached the Here-Now by traveling at less than the speed of light. Such a world line is called *timelike* because its projection on the time axis is greater than its projection on a space axis—they are the world lines connecting spacetime points that are potentially causally linked. That is, an event at a spacetime point in the Past half-cone could have had an effect on the event at the Here-Now, even though its influence propagated at less than the speed of light. Also, an event at the Here-Now could potentially affect the event at any spacetime point in the Future half-cone of the Here-Now.

Any points in the regions of spacetime outside the Future and Past half-cones *cannot* be reached from the Here-Now except by world lines tilted more than 45°

---

[91] The "others" Feynman had in mind included, in particular, the eminent Swiss physicist Ernest C. G. Stükelberg (1905-1984), who in a 1942 article in the journal *Helvetica Physica Acta* also wrote of waves scattering backward in time.

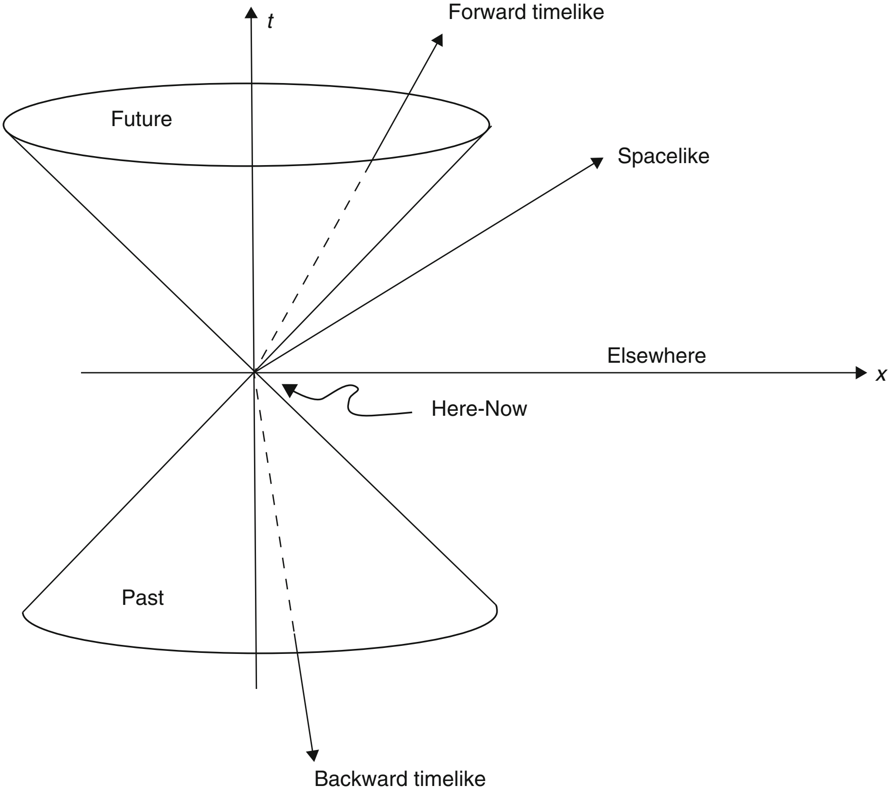

Fig. 3.10 A light cone in spacetime

away from the vertical. Such world lines, which represent travel at a speed faster than light, are called *spacelike* because their projections on a space axis is greater than their projections on the time axis. It is impossible for these world lines to connect causally linked events, and collectively they form the Elsewhere of the Here-Now. Notice that every point in spacetime has its own light cone. If A and B are causally linked, then if B is in the Future half-cone of A, then A is in the Past half-cone of B.

The imagery of the light cone is often useful in making seemingly quite abstract ideas appear transparent. For example, can an observer predict his own future from perfect knowledge of his own past? The easy answer is "No, because quantum uncertainties prohibit perfect knowledge of even the present, much less the past." But suppose we ignore quantum mechanics and limit our question to a universe that obeys only classical physics (which includes the special and general theories of relativity). Surprisingly (perhaps), the answer is *still* no. Having perfect knowledge of your own Past half-cone doesn't include knowing the *entire* past, so if you attempt to predict your own future (say, 1 min from now), there can be influences in Elsewhere that will arrive in the future (say, 59 s from now) about which you

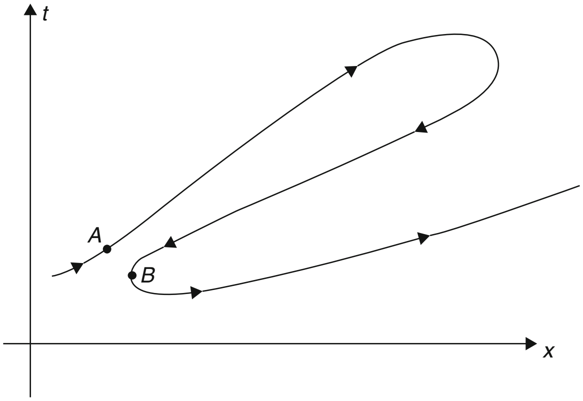

Fig. 3.11 World line of a particle traveling backward in time from A to B

presently, by definition, cannot have any knowledge. And without that knowledge, you cannot predict. As one writer amusingly concludes a tutorial on this topic, "the prospect of predicting the future looks pretty bleak."[92]

A spacetime diagram does not always have to have future directed world lines. If a particle moves backward through time, assuming such a thing is possible, then the diagram can show this by having the world line double back on itself, as in Fig. 3.11 (in which the world line curves back and comes arbitrarily close to itself: it is the world line of a particle that visits itself in the past). Note that the world line in Fig. 3.11 does not actually touch or cross itself, because that would represent more than just a visit—it would represent a particle occupying the same spatial location at the same time as its earlier self. That would be catastrophic and, because it *did not* happen it *cannot* happen. Since the arrowheads on the world line always point in the direction of the local future of the particle, if the 'particle' is actually human then increasing memories are formed in the direction of the arrows. The time traveler at B has more memories than he does at A, even though A and B are nearly identical points in spacetime.

There is a problem with Fig. 3.11 that you may have caught. It is impossible to draw such a doubled-back world line in such a way that *at all places* it never tilts more than 45° from the vertical. That is, at least some portion of the world line will have

$$\left|\frac{\mathrm{dx}}{\mathrm{dt}}\right| > 1$$

which represents superluminal motion (we'll return to this in Chap. 5). One way to keep a bent-back world line always subluminal is to arrange for the light cones along the world line to be tilted relative to each other, as shown in Fig. 3.12, which

---

[92] M. Hogarth, "Predicting the Future in Relativistic Spacetimes," *Studies in the History and Philosophy of Science*, December 1993, pp. 721-739.

Fig. 3.12 Tilted light cones in a curved spacetime (the future halves are shaded)

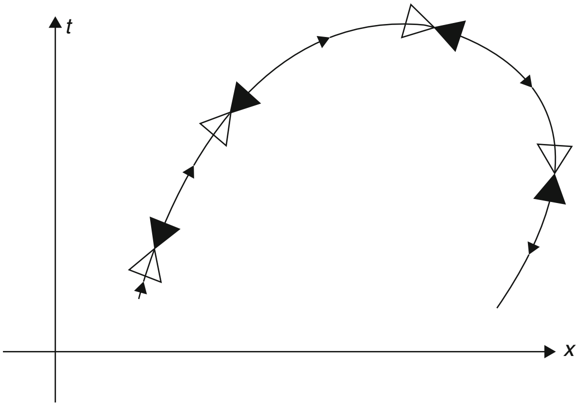

is possible only in a *curved* spacetime. This is an illustration of how general relativity *locally* obeys special relativity's demand that nothing travels faster than light and yet, *globally*, in curved spacetime, things are not so simple. In flat spacetimes all light cones are always 'aligned,' but in curved spacetimes they (generally) are not, and from that can come time travel to the past.[93]

Tilted light cones is the physics behind backward time travel around a rotating Tipler cylinder, for example, a particular time machine we'll discuss later in Chap. 5. Light cone tipping is, in fact, *essential* for time travel to the past. The mere presence of mass tips light cones ('warps spacetime'), but the effect is unnoticeable in everyday life on Earth. A truly enormous mass density is required to tip nearby light cones over so that their Future halves noticeably open up toward the massive body. If the massive body is additionally set to rotating, then a further consequence of Einstein's general theory is that the local light cones are tilted additionally in the direction of rotation. That is, the Future half-cones in spacetime open-up both toward the body and in the direction of the rotation.

It should now be clear that the only way a world line can bend back on itself for a close encounter visit is for both x and t to change. In other words, the world line of a particle that remains fixed in space and reverses just its time direction *runs into itself*. This is why the classic fictional time machine of H. G. Wells could not possibly work. A real time machine must move in space as well as through time, as does a Gödelian rocket (or, for that matter, as does the DeLorean time car in the *Back to the Future* films). The idea of warping world lines, to support time travel to the past, entered science fiction at an early date. For example, when the inventor of the time machine in a 1930s tale[94] is asked about the principle underlying his gadget, he replies "An electro-magnetic warping of the spacetime continuum.

---

[93] What is meant by a spacetime being *flat* will be formalized when we get to spacetime *metrics* later in this section. The Minkowski spacetime of special relativity is a flat spacetime, has no tilted light cones, and as such does *not* support time travel to the past.

[94] N. Schachner, "When the Future Dies," *Astounding Science Fiction*, June 1939.

The machine, if it works, will slide around the world-line of events and reappear at any specified time and place."

Using the Lorentz transformation equations from the previous section, we can establish quite general relationships between events in the Future, Past, and Elsewhere regions of spacetime. For example, (1) All events in the Future/Past for the Here-Now observer are in the Future/Past for any other nearby, relatively moving observer; (2) Any event in Elsewhere can appear to be simultaneous with the Here-Now for some observer and not simultaneous for another observer; and (3) The temporal ordering (the relations of *before* and *after*) of causally related events is the same for all observers. This is not so for events that are not causally related; if two events have a spacelike separation, then two observers can disagree over the temporal ordering of the events. This is, in fact, the basis for the two-wormhole and the cosmic string time machines, both of which will be discussed in Chap. 6. All these statements are easy to prove.

Consider, for example, statement (1). From the previous section we have (with $c = 1$)

$$t' = \frac{t - \mathrm{vx}}{\sqrt{1 - v^2}} \quad \text{and} \quad x' = \frac{x - \mathrm{vt}}{\sqrt{1 - v^2}}$$

where t and x are the coordinates of some event A as measured by the observer in the stationary reference frame, and t' is the time measured by the observer in the reference frame moving at speed v. Thus

$$x'^2 - t'^2 = \frac{(x - \mathrm{vt})^2 - (t - \mathrm{vx})^2}{1 - v^2} = [\text{after a little algebra}]\ x^2 - t^2.$$

For the stationary observer the criterion for an event to be in the Future half-cone is $t > |x|$, that is, $t^2 > x^2$. Thus, $x^2 - t^2 < 0$ for all Future events. But the foregoing result then says $x'^2 - t'^2 < 0$, too, which is the moving observer's criterion for the event being in his Future half-cone. The same sort of argument shows that the two observers also agree on Past events.

Next, suppose that two events A and B occur such that the stationary observer measures them to be $\Delta T = t_B - t_A$ apart in time. Then, we can establish statement (2) by writing

$$t'_A = \frac{t_A - \mathrm{v}x_A}{\sqrt{1 - v^2}} \quad \text{and} \quad t'_B = \frac{(t_B + \Delta T) - \mathrm{v}x_B}{\sqrt{1 - v^2}}$$

and so

$$\Delta T' = t'_B - t'_A = \frac{\Delta T + v(x_A - x_B)}{\sqrt{1 - v^2}}.$$

From this we have $\Delta T' = 0$ (that is, simultaneity) for the two events for the special observer moving at the speed

$$v = \frac{\Delta T}{x_B - x_A}$$

and this speed is less than the speed of light for the condition $x_B - x_A > \Delta T$. This is, of course, the condition for event B to be in the Elsewhere of event A. In fact, we can even have $\Delta T' < 0$ (with $\Delta T > 0$) for $v < 1$ in this case of spacelike separation of A and B. That is, a stationary observer and a sublight-speed moving observer can disagree about the temporal ordering of events with spacelike separation.

Similarly, for event B to be in the causal Future of event A, we have the condition $x_B - x_A < \Delta T$. Then,

$$\Delta T' = \frac{\Delta T - \mathrm{v}(x_B - x_A)}{\sqrt{1 - v^2}} > \frac{\Delta T - v\Delta T}{\sqrt{1 - v^2}} = \Delta T \frac{1 - \mathrm{v}}{\sqrt{1 - v^2}}.$$

Thus, $\Delta T > 0$ says $\Delta T' > 0$ for $v < 1$, and this establishes statement (3).

If we were drawing diagrams with both axes representing space (a plot of y versus x, for example), we would normally define a *distance metric* for the diagram using our everyday ideas about distance. That is, we could say that if we make differential movements of dx and dy along the two coordinate axes, then the differential distance ds is given by

$$(ds)^2 = (dx)^2 + (dy)^2.$$

This is, of course, just the Pythagorean theorem for the ‘Euclidean’ or ‘as the crow flies’ distance function. But it is not the only possible distance function. A distance function has several interesting mathematical properties,[95] but the one we are particularly interested in here is its invariance with respect to the coordinate system. For example, if we draw a line segment on a flat sheet of paper, the physical distance between its end-points does not depend on how we happen to select the x and y axes, a fact illustrated in Fig. 3.13 with the addition of a rotated and translated primed system. The coordinates for the endpoints A and B are obviously different in the two coordinate systems, but we still find that

$$(dx)^2 + (dy)^2 = (dx')^2 + (dy')^2.$$

---

[95] Mathematicians have defined the general properties of a distance function as follows: if A and B are any two points, and if $d(A, B)$ is the distance between A and B , then (1) $d(A, B) = d(B, A)$; (2) $d(A, B) = 0$ if and only if $A = B$; and (3) if C is any third point, then $d(A, B) \leq d(A, C) + d(C, B)$. The Pythagorean distance function possesses all three of these properties, but so do many other functions (for example, $ds = |dx| + |dy|$).

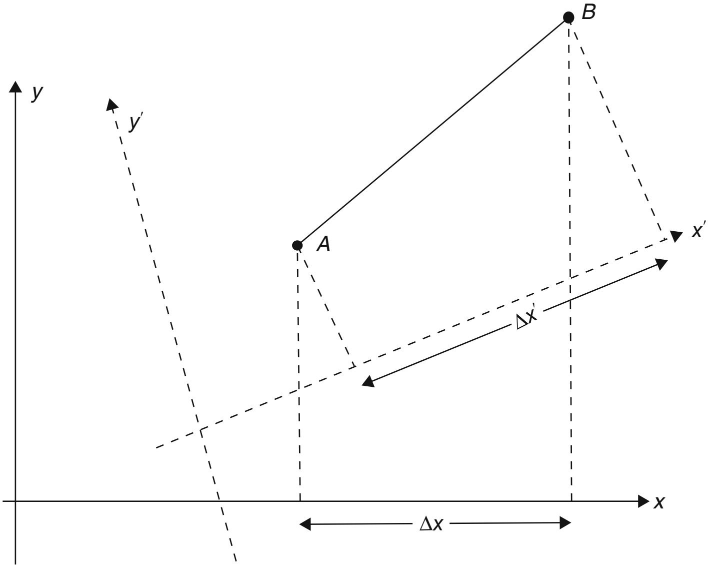

Fig. 3.13 Rotated/translated coordinate systems

We say that the Pythagorean distance function is *invariant*, that it is the same for all coordinates systems that are simply rotations and/or translations of each other.

We know that different observers, if in relative motion in the same spacetime, will see different space and time coordinates for the same event. Thus, it is natural to ask 'what is the metric for flat spacetime?' Is there, in fact, a metric that gives the same *distance* between two events for all observers? We might try to generalize in the obvious way from the Pythagorean theorem, and write

$$(ds)^2 = (dt)^2 + (dx)^2 + (dy)^2 + (dz)^2$$

where now all four dimensions are included. We would then ask ourselves whether it is true that

$$(ds)^2 = (ds')^2 = (dt')^2 + (dx')^2 + (dy')^2 + (dz')^2?$$

For our simple two-dimensional spacetime, this question reduces to asking whether

$$(dt)^2 + (dx)^2 = (dt')^2 + (dx')^2?$$

Using the Lorentz transformation equations from earlier, it is easy to discover that the answer is no. The 'natural' generalization of Pythagorean distance for flat, two-dimensional spacetime fails when four dimensions are included. So, what do we do now? Recalling the words of Professor Mundle from *Some First Words* (see note 23 there), we might wonder whether this difficulty could result from the fact that there is no fourth direction along which the time axis can point at right angles to the three space directions? At least, there is no *real* direction—but perhaps there is an imaginary one. Accordingly, with $= \sqrt{-1}$, let's try

$$(ds)^2 = (idt)^2 + (dx)^2 + (dy)^2 + (dz)^2 = -(dt)^2 + (dx)^2 + (dy)^2 + (dz)^2.$$

Using *imaginary time*, something that seems to be in the realm of science fiction, has resulted in a change in the sign of $(dt)^2$.

This is a crucial change, however, because this new metric *is* invariant. For a reason to be explained in the next section, I will use the negative of this metric (a choice that has no impact on the invariance property) and so write

$$(ds)^2 = (dt)^2 - (dx)^2 - (dy)^2 - (dz)^2.$$

For our simplified two-dimensional spacetime this reduces to

$$(ds)^2 = (dt)^2 - (dx)^2.$$

As before, the Lorentz transformation equations (with $c = 1$) are

$$t' = \frac{t - \mathrm{vx}}{\sqrt{1 - v^2}} \quad \text{and} \quad x' = \frac{x - \mathrm{vt}}{\sqrt{1 - v^2}}.$$

If we then calculate dx' and dt' from these equations, using

$$dx' = \frac{\partial x'}{\partial x}dx + \frac{\partial x'}{\partial t}dt, dt' = \frac{\partial t'}{\partial x}dx + \frac{\partial t'}{\partial t}dt$$

which are the fundamental relations for the *total* differential[96] of a function of two variables, and insert the results into $(dt')^2 - (dx')^2$, we quickly discover the invariance property of this quantity (a result we actually found earlier, in a different way, when we showed that observers in relative motion agree about what events are in the Future and what events are in the Past). Thus,

$$(dt')^2 - (dx')^2 = (dt)^2 - (dx)^2.$$

This quantity, on either side of the equality, is called the spacetime *interval* between the two events separated in flat spacetime by either dt, dx, dy, and dz, or by dt', dx', dy', and dz'. The observers in the unprimed and the primed systems see different individual space and time separations for two events, but they see the same

---

[96] The $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial t}$ symbols denote the *partial* derivatives with respect to x and t (see any good calculus book to brush-up on this). The rest of this chapter *will* have some more math in it, involving derivatives and even an integral or two, but nothing beyond freshman calculus. I've included it mostly for those who would feel cheated without *some* math!

interval. A single time coordinate, and three space coordinates, are said to form a *four-vector that is invariant under Lorentz transformation*. There are, in addition, other four-vectors that are also invariant under Lorentz transformation, such as the energy-momentum, velocity, and force four-vectors; all these quantities have invariants that are formed the same way, by taking the difference of the squares of the time components and the sum of the squares of the space components. The intrusion of the square root of minus one, in the time coordinate of the metric, seems a pretty clear indication that time *is* different from space.[97]

But still, the spatialization of time is nonetheless deeply embedded in Western culture. For example, when writing her popular 1978 book, drawing historical parallels between the fourteenth and twentieth centuries, Barbara Tuchman titled it *A Distant Mirror*, and not *An Old Mirror*. The mathematical mixing of space and time appeared quite early in science fiction, but often in comically mangled form. In one such tale, for example, as an evil scientist uses his time machine to transport captives into the past, he tells them, "We've got a longish journey before us, ten thousand years more, multiplied by the fourth power of two thousand miles."[98]

In general relativity, the metric of any four-dimensional spacetime has the structure of what mathematicians call a *symmetric quadratic Riemannian form*:

$$(ds)^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} g_{ij}(dx_i)(dx_j),\, g_{ij} = g_{ji}$$

where $x_1 = t$, $x_2 = x$, $x_3 = y$, and $x_4 = z$, and the 16 g's are all functions of these four variables. (Because of the symmetry condition, only 10 of the g's are independent.) In this notation, a *flat* spacetime is mathematically characterized by $g_{11} = 1$, $g_{22} = g_{33} = g_{44} = -1$, $g_{ij} = 0$ for all $i \neq j$. Now, for a given spacetime, one can arbitrarily choose an infinity of coordinate systems. If just one of this infinity of systems is such that the ± 1, 0 values for the g's occur, then that spacetime is *globally* (that is, everywhere) flat. If no such coordinate system exists, then that spacetime is necessarily *curved*—I'll give you a more intuitive view of curvature in just a bit. (If this notation is extended to a fifth dimension[99] by including the additional coordinate x_5, then there are an additional eight off-diagonal g's and so four new *independent* g's. These are just sufficient to describe the electromagnetic field, too, along with the gravitational field described by the other ten g's. That is what is meant by saying five-dimensional spacetime 'unifies' gravity with electromagnetism.) There is a g_{55}, too, which could be allowed to model a slowly varying gravitational constant, as suggested by Dirac in 1938.

---

[97] Writing in 1972, one famous physicist said of his first encounter with the metric of special relativity (Minkowski spacetime), "Now, when I saw that minus sign [in $-(dt)^2$], it produced a tremendous effect on me. I immediately saw that here was something new." See P. A. M. Dirac, "Recollections of an Exciting Era," in *History of Twentieth Century Physics* (C. Weiner, editor), Academic Press 1977.

[98] V. Rousseau, "The Atom Smasher," *Astounding Stories*, May 1930.

[99] The fifth dimension was introduced in the 1920s by the German physicist Theodor Kaluza (1885-1954), but just what the nature of this fifth dimension might be remains a mystery. A few years after Kaluza, the Swedish physicist Oscar Klein (1894-1977) speculated that it might be a spatial dimension curled-up in a tiny circular path, so tiny that we don't notice it; the issue remains open. The idea of a fifth dimension appeared early in pulp science fiction, as in the January 1931 tale "The Fifth-Dimension Catapult," (*Astounding*) by Murray Leinster.

The g functions are the components of the so-called *metric tensor of the second rank* (see the last discussion question for more on what this means) of that spacetime. The g's at each point in spacetime are related to the curvature of spacetime at that point, which in turn is dependent on the g's and on the energy density at that point. In fact, the ten equations for the g's, which *are* the famous Einstein gravitational partial differential field equations, are both nonlinear and coupled. That is, each g_{ij} is in general a nonlinear function of all the other g_{ik}, which accounts for the notorious difficulty[100] in finding analytical solutions to the field equations except in certain highly special cases, such as the spacetime of spinning spheres and rotating infinite cylinders.

What does it mean to 'solve' the field equations? It is useful to think of the equations schematically as follows:

local geometry of spacetime ← local density, momentum and stress of the mass-energy of spacetime where the direction of the arrowhead means that the 'usual' practice is to assume the right-hand side (the so-called stress-energy tensor) as given, and then attempt to calculate the left-hand side. If the attempted calculation can be done, then one has solved the field equations for the spacetime geometry that is associated with the assumed mass-energy distribution.

Suppose, however, that we reverse the direction of the arrowhead. That is, suppose we assume a desired geometry. That is what Einstein did when he assumed the geometry of a *static* (non-expanding) universe and solved for the required mass-energy. What he found was just what had been observed by astronomers up to that time—a multitude of 'grains of matter' (what physicists call *dust*) plus the infamous cosmological constant. The constant, with its repulsive gravity, was needed to counteract the ordinary gravitational attraction of the stars that tends to pull them together. The later discovery that the universe is not static, but rather is expanding, rendered Einstein's solution moot.

Now let's go Einstein one better, and assume a spacetime geometry that contains closed timelike curves (a time machine, in other words) and then try to calculate the mass-energy distribution required by that spacetime. If that can be accomplished—and in fact the field equations themselves provide an algorithmic means of solution in this direction—then the physicist's work is done. The required mass-energy distribution requirements are put out to bid to 'spacetime engineers' and the lowest bidder 'simply' constructs that mass-energy distribution and so builds us our time machine! When the calculations are done, however, what has happened without fail, at least up to now, is that the resulting mass-energy distribution comes out with an ‘unphysical’ nature, a technical way of saying our spacetime engineer wouldn’t know *how* to assemble the required mass-energy distribution. What is meant by ‘unphysical’ will be explored, in the particular case of wormhole time machines, in Chap. 6.

---

[100] An elementary, quite interesting discussion of the enormous computational complexity of the field equations is presented in Richard Pavelle and Paul S. Wang, "MACSYMA from F to G," *Journal of Symbolic Computation*, March 1985, pp. 69-100.

The 16 g’s for a four dimensional spacetime are often written in the form of a 4 × 4 matrix. In fact, the metric tensor is of the *second* rank precisely because a matrix has a *two*-dimensional form; scalars and vectors, which have forms of zero and one dimension, are tensors of rank zero and one, respectively. The collection of the algebraic signs of the main diagonal terms (the g_{ii}) is called the *signature* of the metric tensor. The signature of flat (Minkowski) spacetime is thus written as [+, −, −, −]; in more general (curved) spacetimes, this same signature is called *Lorentzian*. By contrast, the signature of a four-dimensional Euclidean space is [+, +, +, +]. This signature is called *Riemannian*. The geometry of flat, Minkowski spacetime is not Euclidean geometry because the spacetime signature of its metric has both plus *and* minus signs. As an uncurved spacetime, Minkowski spacetime *has no gravity*; to get gravity, we need a curved spacetime.

The idea of linking the curvature of a four-dimensional space to physical phenomena is the signature feature of general relativity, but it actually pre-dates Einstein by decades. It can be found, for example, in the work of the British mathematician William Kingdon Clifford (1845–1879), done in the 1860s and 1870s before Einstein’s birth (in the year of Clifford’s death). In his posthumously published book *The Common Sense of the Exact Sciences* (1885), Clifford wrote “We may conceive our space to have everywhere a nearly uniform curvature, but that slight variations of the curvature may occur from point to point and themselves vary with time. These variations of the curvature with time may produce effects which we not unnaturally attribute to physical causes independent of the curvature of our space. We may even go so far as to assign to this variation of the curvature what really happens in that phenomenon which we term the motion of matter.” It isn’t a long jump from “motion of matter” to gravity![101]

Not everyone enthusiastically embraced this new, radical view of nature. For example, the great Scottish mathematical physicist James Clerk Maxwell (1831–1879), of Maxwell’s equations fame, who knew Clifford through their common membership in the London Mathematical Society, summarily dismissed this part of Clifford’s work as simply the speculations of a “space crumbler.” Decades later, Einstein faced the same rejection when the eminent British astronomer Sir James Jeans declared “Einstein’s crumbling of his four-dimensional space may . . . be considered to be . . . fictitious.”[102]

---

[101] Clifford almost surely found inspiration in this part of his work from the even earlier efforts of the German mathematician Bernhard Riemann (1826-1866). See, for example, Clifford’s translation of Riemann’s famous 1854 lecture “On the Hypotheses Which Lie at the Bases of Geometry,” *Nature*, May 1, 1873, pp. 14-17, and continued in the next issue (May 8, 1873, pp. 36-37).

Spacetime geometries are not easy concepts to grasp, and the metrics of curved spacetimes are even more complicated than is the metric of the flat spacetime of Minkowski’s special relativity. As one paper so aptly put it, “Experience has taught us that the space in which we live has a geometry that is three-dimensional and Euclidean . . . We are very much at home with [that] geometry . . . But the geometric properties of a Minkowskian space are so alien to us that we may well despair of visualizing them, and a Riemannian [curved] space . . . seems totally beyond comprehension.”[103]

An important feature of Riemannian geometry is that although it is generally not globally flat, it is always *locally* flat. Thus any sufficiently small region in a curved Riemannian spacetime can be approximated, with arbitrarily small error, by a flat pseudo-Euclidean Minkowskian spacetime. That is, at each point in Riemannian spacetime, there is some *particular* inertial frame of reference in which special relativity is all there is to spacetime physics *at that point*. The particular inertial frame required is different, however, from point to point.

In a coordinate system different from rectangular, the flat Minkowskian metric can appear radically altered, but that is just an artifact of the mathematics and has no physical significance. For example, in spherical coordinates the Minkowskian metric becomes the equivalent

$$(ds)^2 = (dt)^2 - (dr)^2 - (rd\theta)^2 - (rsin(\theta)d\phi)^2$$

where ϕ is the azimuthal angle and θ is the angle measured from the polar axis. A related metric occurs in the theory of spherically symmetric, static (no time variation) time machine wormholes (discussed in Chap. 6), of the form

$$(ds)^2 = \left(e^{a(r)}dt\right)^2 - \frac{(dr)^2}{1-\frac{b(r)}{r}} - (rd\theta)^2 - (rsin(\theta)d\phi)^2$$

where $a(r)$ is called the *redshift function* and $b(r)$ is called the *shape factor*. These two functions are nearly arbitrary, subject only to the constraints that both $b(r)/r$ and $a(r)$ vanish as r goes to infinity (r is the radial distance from the throat of the wormhole mouth). Indeed, as r increases, this curved wormhole spacetime metric

---

[102] Quoted in P. Kerszberg, “The Relativity of Rotation in the Early Foundations of General Relativity,” *Studies in History and Philosophy of Science*, March 1987, pp. 53-79.

[103] R. W. Brehme and W. E. Moore, “Gravitational and Two-Dimensional Curved Surfaces,” *American Journal of Physics*, July 1969, pp. 683-692.

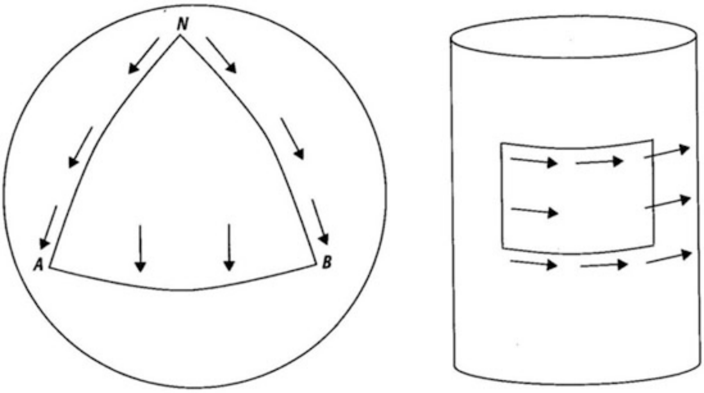

Fig. 3.14 The curvature of a space can be revealed by the process of parallel transport

reduces to that of flat Minkowskian spacetime, so this wormhole spacetime is said to be *asymptotically flat*.[104]

An intuitive appreciation of 'flatness' can be realized in terms of what is called the *parallel transport* of a vector around a closed path. In a curved space the vector will experience a rotation, which will not occur in a flat space. Two examples of parallel transport in ordinary three-dimensional space are shown in Fig. 3.14. The spherical surface on the left is curved because if you slide the vector from *N* to *A* to *B* and then back to *N*, all the while keeping it parallel to its immediately previous orientation, then when it gets back to *N* the vector will point to *B*, *not* toward *A* as it initially did. A similar trip on the cylindrical surface, however, results in zero rotation of the vector. Thus, the cylindrical surface, despite superficial appearances, is *not* curved.

Using the idea of the metric tensor, we can develop a more formal demonstration that the surface of a sphere, unlike like that of a cylinder, is not flat. On the surface of a sphere of radius a (*on* the surface $r = a$ everywhere, and so $dr = 0$), the measure of the distance between two points (in spherical coordinates) is[105]

$$(ds)^2 = a^2 sin^2(\theta)(d\phi)^2 + a^2(d\theta)^2.$$

Writing $x_1 = \phi$ and $x_2 = \theta$ yields the more general form

---

[104] For wormhole spacetimes that are not asymptotically flat, see (for example) K. Narahara, *et al.*, "Traversable Wormhole in the Expanding Universe," *Physics Letters B*, September 29, 1994, pp. 319-323.

[105] Note *carefully* that this is a purely spatial problem, with no time, and we are taking all of the metric coefficients as positive (unlike in the case of a spacetime metric).

$$(ds)^2 = g_{11}(dx_1)^2 + g_{12}(dx_1)(dx_2) + g_{21}(dx_2)(dx_1) + g_{22}(dx_2)^2$$

or, using the symmetry condition $g_{12} = g_{21}$, we have

$$(ds)^2 = g_{11}(dx_1)^2 + 2g_{12}(dx_1)(dx_2) + g_{22}(dx_2)^2.$$

From this we immediately have $g_{11} = a^2 sin^2(\theta)$, $g_{22} = a^2$, and $g_{12} = g_{21} = 0$.

Now, suppose we ask whether it is possible to find some new coordinate system (with variables x'_1 and x'_2) in which the invariant $(ds)^2$ is given by the flat Euclidean metric $\left(dx'_1\right)^2 + \left(dx'_2\right)^2$. In such a coordinate system (if it exists) we would have the 'flatness conditions' of $g'_{11} = g'_{22} = 1$ and $g'_{12} = g'_{21} = 0$. With such a change of coordinates, each of our original ϕ, θ coordinates would generally be a function of both of the new coordinates—that is, $\phi = \phi\left(x'_1, x'_2\right)$ and $\theta = \theta\left(x'_1, x'_2\right)$. Thus, writing the total differential of a function of two variables (see note 96 again), we have

$$d\phi = \frac{\partial\phi}{\partial x'_1}dx'_1 + \frac{\partial\phi}{\partial x'_2}dx'_2$$
$$d\theta = \frac{\partial\theta}{\partial x'_1}dx'_1 + \frac{\partial\theta}{\partial x'_2}dx'_2.$$

Substituting these expressions into the above expression for $(ds)^2$ on the surface of a sphere, and collecting terms, we arrive at

$$\begin{aligned}(ds)^2 &= a^2\left\{\left[\sin^2(\theta)\left(\tfrac{\partial\phi}{\partial x'_1}\right)^2 + \left(\frac{\partial\theta}{\partial x'_1}\right)\right]\left(dx'_1\right)^2 + \left[\sin^2(\theta)\left(\tfrac{\partial\phi}{\partial x'_2}\right)^2 + \left(\tfrac{\partial\theta}{\partial x'_2}\right)^2\right]\left(dx'_2\right)^2\right.\\ &\quad\left.+2\left[\sin^2(\theta)\frac{\partial\phi}{\partial x'_1}\frac{\partial\phi}{\partial x'_2} + \frac{\partial\theta}{\partial x'_1}\frac{\partial\theta}{\partial x'_2}\right]\left(dx'_1\right)\left(dx'_2\right)\right\}\\ &= g'_{11}\left(dx'_1\right)^2 + 2g'_{12}\left(dx'_1\right)\left(dx'_2\right) + g'_{22}\left(dx'_2\right)^2\end{aligned}$$

We can now immediately write down each of the g' and, if we demand that they satisfy the 'flatness conditions,' then we have the following three statements:

$$sin^2(\theta)\left(\frac{\partial\phi}{\partial x'_1}\right)^2 + \left(\frac{\partial\theta}{\partial x'_1}\right)^2 = \frac{1}{a^2}$$
$$sin^2(\theta)\left(\frac{\partial\phi}{\partial x'_2}\right)^2 + \left(\frac{\partial\theta}{\partial x'_2}\right)^2 = \frac{1}{a^2}$$
$$sin^2(\theta)\frac{\partial\phi}{\partial x'_1}\frac{\partial\phi}{\partial x'_2} + \frac{\partial\theta}{\partial x'_1}\frac{\partial\theta}{\partial x'_2} = 0.$$

For a *globally* flat surface, that is, a surface that is flat *everywhere*, these three statements must hold in particular at the poles of the sphere. That is, at $\theta = 0^\circ$

and at $\theta = 180^{\circ}$. At both of these points $\sin(\theta) = 0$, and so at the poles the three statements reduce to

$$\left(\frac{\partial\theta}{\partial x_1'}\right)^2 = \frac{1}{a^2}$$
$$\left(\frac{\partial\theta}{\partial x_2'}\right)^2 = \frac{1}{a^2}$$
$$\frac{\partial\theta}{\partial x_1'}\frac{\partial\theta}{\partial x_2'} = 0.$$

But the third statement is incompatible with the first two and, because of that incompatibility, there is no primed coordinate system in which the g' coefficients in the metric are those of a globally flat metric. Thus, unlike the surface of a cylinder, the surface of a sphere is not flat but rather is curved. This almost surely comes as no surprise to you, but now you can *prove* it!

3.6 Proper Time and the Twin Paradox in Time Travel to the Future

"[The] equations of duo-quadrant lineations [have] been substantiated … Our fourth-angle deviation from the six conceivable electronic dimensions did the trick all right. I went forward in Time."[106]

The spacetime interval of the previous section has an important interpretation that leads to one of the more dazzling results of special relativity—time travel into the future. First, recall the flat spacetime metric

$$(ds)^2 = (dt)^2 - (dx)^2 - (dy)^2 - (dz)^2$$

in which the use of unprimed variables indicates that the measurements on the space and time coordinates of a moving particle are made with respect to a stationary observer's frame of reference. Now, suppose that the space and time coordinates of a moving particle are made with respect to the particle instead. Then, using primed variables for measurements made in this new frame of reference, $dx' = dy' = dz' = 0$ because the particle is always at the origin (by definition)! Recalling the invariance of the spacetime interval for all observers, we conclude that

---

[106] A science fiction scientist babbles incoherent nonsense, not special relativity, about how to travel into the future, in a story by J. H. Haggard, "He Who Masters Time," *Thrilling Wonder Stories*, February 1937.

$$(ds')^2 = (ds)^2 = (dt')^2.$$

That is, the spacetime interval between two events is the time lapse measured by a clock attached to a particle that moves from one event to the other. This time is called *proper time*, which gets its name from the idea that it belongs to (is the *property* of) the moving particle. This is the technical reason for taking (as we did in the previous section) $(ds)^2 = (dt)^2 - (dx)^2$ rather than $(ds)^2 = (dx)^2 - (dt)^2$. The first choice avoids the somewhat awkward result of an imaginary proper time.

Next, we'll adopt what has come to be called the *clock hypothesis*, which states that an accelerated clock runs at the same instantaneous rate as an unaccelerated clock that is moving alongside at the instantaneously same speed. As we showed earlier (take a look back at Sect. 3.3), if the accelerated clock's instantaneous speed is v, then its rate of time keeping (dt') is related to that of the 'stationary' (unaccelerated) clock (dt) as

$$dt' = \sqrt{1 - \left(\frac{v}{c}\right)^2}\, dt$$

where c is the speed of light. The clock hypothesis is generally assumed to be true. Einstein, himself, in his famous 1905 special relativity paper, specifically took the rate of a clock's timekeeping to be velocity-dependent only. When asked during an interview decades later whether it is permissible to use special relativity in situations involving acceleration, Einstein replied "Oh, yes, that is all right as long as gravity does not enter; in all other cases, special relativity is applicable. Although, perhaps the general theory approach might be better, it is not necessary."[107]

The clock hypothesis has long had experimental verification. For example, in one experiment the time keeping of accelerated atomic clocks was determined to be given precisely by the time dilation formula of special relativity, even when their direct mechanical acceleration (the centripetal acceleration produced by a rapidly spinning disk) exceeded 66,000 gees![108] And even more impressive are the time dilation results of a later experiment in which the time keeping of near light-speed charged particles, orbiting in a magnetic field, was in excellent agreement with the time dilation formula, even as accelerations well in excess of 10^{15} gees were reached![109]

The total elapsed time between two events A and B, as measured by the proper time of an accelerated clock making the journey, is, therefore, given by

---

[107] R. S. Shankland, "Conversations with Albert Einstein," *American Journal of Physics*, January 1963, pp. 47-57.

[108] H. J. Hay, *et al.*, "Measurement of the Red Shift in an Accelerated System Using the Mössbauer Effect in Fe^{57}," *Physical Review Letters*, February 15, 1960, pp. 165-166.

[109] J. Bailey, *et al.*, "Measurements of Relativistic Time Dilation for Positive and Negative Muons in a Circular Orbit," *Nature*, July 29, 1977, pp. 301-305.

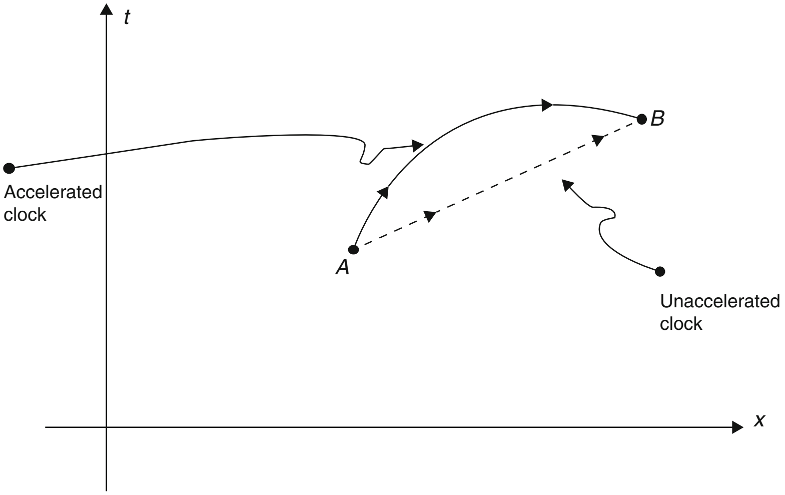

Fig. 3.15 World lines of two clocks (one accelerated and one unaccelerated)

$$t' = \int dt' = \int_{t_A}^{t_B} \sqrt{1 - \left(\frac{v}{c}\right)^2} dt < t_B - t_A \text{ if } v \neq 0$$

where $t_B - t_A$ is the elapsed time between A and B as measured by the unaccelerated clock. The inequality results because, for $v \neq 0$, the integrand is always less than 1. Now, we know that in a spacetime diagram the world line of the unaccelerated clock is a straight line, whereas the world line of the accelerated clock is a curved line. Thus, using Fig. 3.15, combined with the inequality $t' < t_B - t_A$, we have the following central result: the world line of *maximum* proper time is the one that *looks the shortest*, that is, the straight (or free-falling geodesic) world line. In the spacetime diagram the curved line looks longer, but in fact any curved line will have a smaller proper time than does the straight world line. This is a dramatic example of how Minkowskian *spacetime* geometry differs from Euclidean *space* geometry; in the latter geometry, there is no longest path between two points.

From this, we can now understand the famous paradox of the twins, Bob and Bill. Bill remains on Earth, but Bob gets into a rocket ship and goes on a high speed trip out into space. Eventually he brings his ship to a stop, turns around, and returns to Earth. The world lines of Bob and Bill are initially together, then they diverge as Bob goes on his trip, and then they come together again at the end of Bob's trip, as shown in Fig. 3.16. The details of Bob's trip are not important for a general statement of the paradox (although in just a bit I *will* present the details for one possible trip). All we need observe for now is that Bill's world line from A to B is straight, whereas Bob's is curved. Bill's body (that is, his local clock) will therefore measure a greater proper time than will Bob's; that is, Bob will be younger than his stay-at-home twin! Equivalently, upon his return Bob will hear his Earthbound

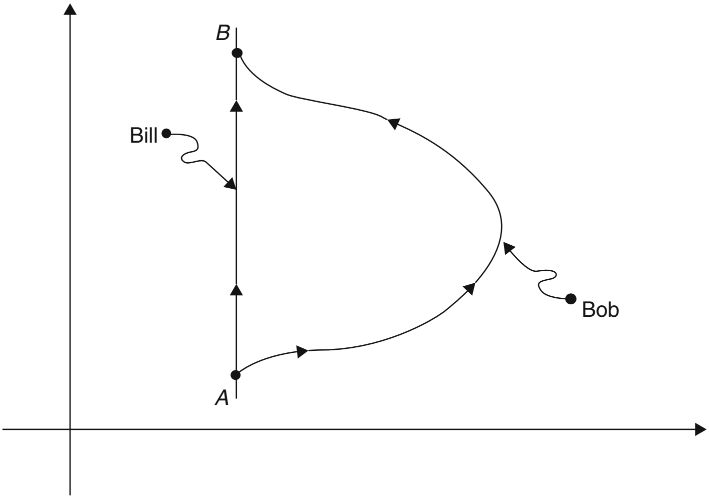

Fig. 3.16 World lines of unaccelerated (Bill) and accelerated (Bob) twins

brother declare the date to be further in the future than Bob's trip lasted (according to Bob). Bob will therefore conclude that he has traveled into the future.[110]

This situation is called a paradox, not because of the time travel aspect (there are no logical paradoxes associated with travel into the future), but rather because it seems to violate the very spirit of special relativity. That is, from Bill's point of view, Bob at first travels away and then returns. But one can argue that from Bob's point of view it is Bill who first recedes and then returns. So why is it *Bob* who is the younger, and not Bill? Long after Einstein's 1905 publication of special relativity, this point remained a great puzzle for many. For example, in the 1923 Presidential Address to the Eastern Division of the American Philosophical Association, we read this very objection to the twin paradox. The conclusion by the speaker is that such a thing "could happen only in a universe in which all squares were round and the *principio contradiction* had been put to sleep."[111]

So, what's the answer, why is it *Bob* who is the younger of the twins? The classic physics answer is that the two points of view are actually *not* symmetrical, that there is a definite asymmetry between Bill and Bob. After all, it is Bob who feels the acceleration from the rocket's engines as he blasts off from Earth—it is Bob, not

---

[110] The twin paradox is hinted at in Einstein's 1905 paper, but it is in a 1911 address to the International Congress of Philosophy in Bologna, by the French physicist Paul Langevin (1872-1946), that a human space traveler is first introduced (in a cannonball moving at near light-speed, an idea motivated by Langevin's reading of Jules Verne's 1872 novel *From the Earth to the Moon*). The writer Pierre Boulle proudly mentioned this contribution by his fellow Frenchman in the time travel story "Time Out of Mind" (you can find it in Boulle's collection *Time Out of Mind*, Vanguard Press 1966).

[111] W. P. Montague, "The Einstein Theory and a Possible Alternative," *The Philosophical Review*, March 1924, pp. 143-170.

Bill, who feels *force*—whereas Bill feels nothing unusual as he remains on Earth. The more fundamental physics answer, however, is that Bob's world line in spacetime is curved, whereas Bill's is straight.

In an open, spatially unbounded, flat spacetime, *curved* is indeed synonymous with *accelerated*, but this need not be so in a *closed,* flat spacetime. In an open, flat spacetime, the only way two world lines can diverge in the past and then meet again in the future is for at least one of them to curve, but in a closed but still flat spacetime, it is possible for two straight world lines to meet more than once. For example, Fig. 3.17 shows a simple two-dimensional spacetime that is the surface of a cylinder (which you'll recall we argued earlier is flat), rather than an infinite flat plane. The two world lines in that figure are *both* straight: to visualize this, imagine cutting the cylinder open along the (vertical) time dimension, and then flattening it out. Bob's world line, however, looks longer in the spacetime diagram, so Bob's proper time will be less than Bill's when they meet again, even though now *neither* of them has experienced any acceleration.[112] This is simply an interesting mathematical exercise, however, and as far as is known the spacetime we live in is not cylindrical and so Bob's trip into the future will require an accelerating rocketship.[113]

If we specify the details of Bob's trip, we can then precisely calculate the difference in elapsed time for the twins. In an analysis that dates back to 1962, the German astrophysicist Sebastian von Hoerner (1919–2003) did that for the following trip[114]: To begin, Bob gets into his rocket ship at time $t = t^{\prime} = 0$ (t is time measured on Earth by Bill, and $t^{\prime}$ is time measured by Bob in his rocket). The Bill and Bob synchronize their clocks at the instant of departure. Bob's trip is to be made in comfort, and so his rocket accelerates at a constant rate (a one gee acceleration, for example, would be equivalent to Earth's gravity, and Bob would feel right at home). This is of practical importance, obviously, because we do not want the experienced acceleration to be incompatible with the physical survival of Bob. Bob

---

[112] See C. H. Brans and D. R. Stewart, "Unaccelerated-Returning Twin Paradox in Flat Space-Time," *Physical Review D*, September 15, 1973, pp. 1662-1666. For a similar treatment, this time by a mathematician, see Jeffrey R. Weeks, "The Twin Paradox in a Closed Universe," *American Mathematical Monthly*, August-September 2001, pp. 585-589.

[113] In Chapter 6 we'll discuss the idea of traveling into the past by moving faster than light (*superluminal* motion). A treatment of such travel, in Bob's cylindrical spacetime, is by S. K. Blau, "Would a Topology Change Allow Ms. Bright to Travel Backward in Time?" *American Journal of Physics*, March 1998, pp. 179-185, which answers that question in its last line: "Ms. Bright cannot [return] 'the previous night' and alter history," a conclusion that no doubt met with Hawking's approval. The 'Ms. Bright' in the title is the heroine of a 1923 limerick that you can find quoted in the first *For Further Discussion* of Chapter 6.

[114] Originally appearing in the journal *Science*, under the title of "The General Limits of Space Travel," von Hoerner's analysis was reprinted in the classic anthology *Interstellar Communication* (A. G. W. Cameron, editor), W. A. Benjamin 1963. The arithmetic was, alas, just a bit sloppy (the final formulas, fortunately, are correct), and many of von Hoerner's numerical evaluations are incorrect. Later, the British mathematician Leslie Marder cleaned-up the analysis in his beautiful little book on the twin paradox, *Time and the Space-Traveler*, George Allen & Unwin, Ltd. 1971.

Fig. 3.17 Unaccelerated twin paradox in a cylindrical spacetime

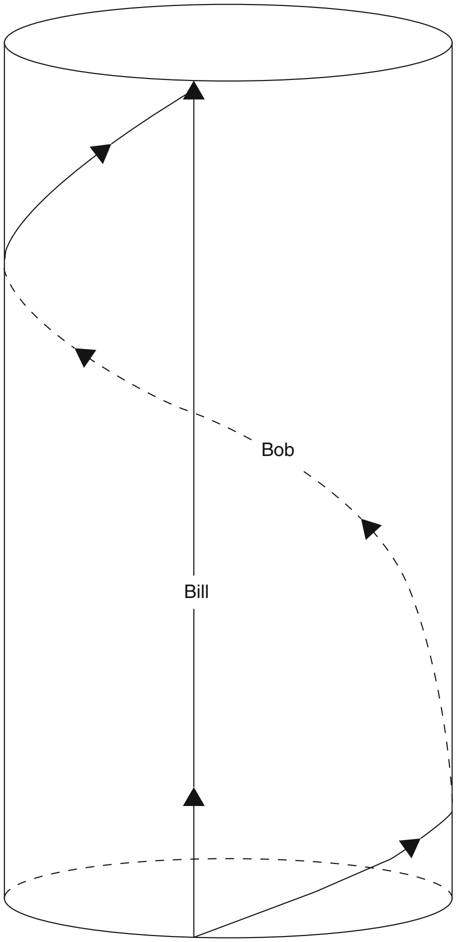

travels this way for a time interval of T (as measured by Bill on Earth) and T' (as measured by Bob in his rocket). At that time the rocket is traveling at its maximum speed. Bob then turns off the rearward engine and turns on a forward-mounted engine so as to experience a constant deceleration. Floor and ceiling interchange, but Bob always weighs the same. If he does this for the same time interval T (as measured by Bill on Earth) and T' (as measured by Bob) as for the initial acceleration phase of the trip, the rocket will be brought to rest with respect to Earth. At that time, $2T$ (as measured by Bill on Earth) and $2T'$ (as measured by Bob), the rocket is at its maximum distance from Earth. Bob then returns to Earth, using the same acceleration/deceleration process, and Bob arrives back home, gently, with a final speed of zero with respect to Earth (ignoring, of course, all the navigational problems due to the motion of Earth during the trip).

Assuming the clock hypothesis is true (not all physicists believe this,[115] but remember, we have Einstein[116] on our side with this!), the result is that if a is the constant acceleration of the rocket (as experienced by Bob), and c is the speed of light, then the relationship between the *total roundtrip time* as measured by Bill ($4T$) and as measured by Bob ($4T'$), is given by

$$T = \frac{c}{a}\sinh\left(\frac{a}{c}T'\right).$$

The difference in what each twin believes to be the date can, in fact, be truly astonishing. For example, if $a = 1$ gee and Bob travels (by his clock) for $4T' = 20$ years, then $4T = 339$ years.[117] Of course, there will actually be no disagreement between Bill and Bob over the date upon Bob's return because Bill will be long dead!

Is it likely that such a time trip[118] will someday be made into the future? It's just my opinion, but I suspect not. Bob will be traveling at virtually the speed of light at maximum speed ($0.9993\,c$, to be precise, at which point he will be 84 light-years from Earth), and to zip through space at such a speed would result in a very high rate of collision with stray hydrogen atoms (about one per cubic centimeter). The result of those energetic collisions would be the intense blasting of Bob with a lethal dose of gamma radiation. And, as von Hoerner showed in his original analysis (note 114), the energy required by Bob's rocket would be simply mind-boggling.

Professor Schild (see note 116), on the other hand, seems to have been less bothered by such considerations when he wrote, at the beginning of a prose discussion of the twin paradox, "I have no doubt that if our technology should ever advance to the stage where large-scale twin effects become noticeable with our unaided senses, then [people] will have no difficulty in adjusting their concepts of time until the new phenomena see quite natural."[119]

I'll end this chapter on time with the observation that it is the distinction between the proper time of Bob on a rocket ship, and the time of those who are not fellow space travelers, that eliminates the occasional suggestion of simply *freezing* one's way via suspended animation to the future (as in the 1992 film *Forever Young*). That is, suppose you manage to talk a friend into climbing into a freezer (please don't actually try this!); at the moment he gets in, his wrist watch agrees with yours (both are powered and maintained at a constant temperature by 100-year nuclear batteries). Years later, when you thaw your friend out, you'll find that your watches still agree. But when Bob returns from his rocket trip, his watch will not agree with yours, but instead will be far behind. Bob is a true time traveler, but your frozen/thawed friend is not.[120]

---

[115] Consider, for example, this remark by 1965 Nobel physics laureate Julian Schwinger (1918-1994) about the twin paradox in his 1986 book *Einstein's Legacy*: "The observer on the spaceship … is *not* in uniform, unaccelerated motion … The special theory of relativity does not apply to such an accelerated observer." Schwinger was wrong in this conclusion (see the next note).

[116] As a physicist wrote on this point after Einstein's death, "A good many physicists believe that [the twin] paradox can only be resolved by the general theory of relativity. … However, they are quite wrong. The twin effect … is one of special relativity." See A. Schild, "The Clock Paradox in Relativity Theory," *American Mathematical Monthly*, January 1959, pp. 1-18. Alfred Schild (1921-1977) was professor of physics at the University of Texas, and a recognized expert in the general theory.

[117] In using the T, T' formula, one has to be careful to use MKS units, that is, length and time measured in meters and seconds, respectively. Thus, $a = 1$ gee = 9.81 meters/second^2 and $c = 186{,}210$ miles/second $= 2.997 \times 10^8$ meters/second.

[118] For a science fiction use of such a trip, see Robert Heinlein's 1956 novel *Time for the Stars*.

[119] A. Schild, "Time," *Texas Quarterly*, Autumn 1960, pp. 42-62.

A provocative illustration of the distinction between the proper time of a time traveler and of those who are not time travelers was given by a philosopher.[121] He begins his analysis by asking what appears to be a question with an obvious answer. Suppose, he says, it is [2018] and you suddenly wake up in a hospital and are told that you have been in a coma for the past 2 weeks. You are also told that you were in an auto accident 2 weeks ago, that you suffered temporary neural damage, and that the eventual reversal of such damage always, at some time within 4 weeks after the damage occurs, causes a day of excruciating pain *if you are conscious at the time*. Would you prefer for the day of damage reversal to be in the past 2 weeks (when you were in the coma) or in the next 2 weeks? The answer seems clear. After all, if the day that damage reversal occurs has already happened, then you simply slept through it and missed the pain. To prefer the day of pain to be in the future (when you will be awake) seems absurd. Now, let's add time travel to the equation.

All is as before, but now you immediately leave the hospital upon regaining consciousness to take a trip back to 1892, where you will stay for 2 weeks. Again, it seems clear that you would prefer to have had the day of pain in the past 2 weeks (in 2018), not in the next 2 weeks (in 1892). Note that the *next* is a reference to your proper time, because whereas 1892 is the global past, it is *your* personal future. Thus, now your preference would be to have the day of pain in the *recent personal past* of 2018, not the *distant global* past of 1892. Now, let's put another time travel twist to this story.

All is as before in the original tale, except now you are told in the hospital that the auto accident happened just after you made a time trip to 2092: as you walked out of your time machine in 2092, you were hit by a car. The 2 weeks you were in a coma were in 2092, before you were judged fit enough (although still unconscious) to make the time journey back to 2018. When would you now prefer to have the day of pain? Clearly, as always, in your *personal past*, which is now the *global future*.

Time *is* different for those who time travel and those who don't!

---

[120] In his 1956 novel *The Door Into Summer*, the always ingenious Robert Heinlein used *both* ideas, with the cold-storage method of reaching the future combined with a true time machine to allow his hero to return to his 'present' (the future's past).

[121] A. Gallois, "Asymmetry in Attitudes About the Nature of Time," *Philosophical Studies*, October 1994, pp. 51-59.

3.7 For Further Discussion

An interesting *theological* analysis of Feynman's idea of an electron traveling backward in time, expressed as thinking of the electron as being in 'one of God's films played in reverse,' is given by J. W. Smith, "Time Travel and Backward Causation," *Cogito* 1985, pp. 57–67. Wondering what it would be like for an electron to travel backwards in time, Smith's answer was: "Consider an electron e_1. At time t_0 it is at (x_0, y_0, z_0). At time t_1 it is at (x_1, y_1, z_1). If the direction of time for the electron was reversed, then the electron would be observed on the 'film of the world' to travel back along the same path as it did before, i.e., back to (x_0, y_0, z_0). If God stopped the 'film of the world' and examined the charge of e_1, then He would find that it was *negative*, not *positive*. Hence the electron traveling backwards in time is simply that: an electron traveling backwards in time, it is not a positron. Time reversal does not result in a reversal of charge. Thus, the Stückelberg-Feynman position is incorrect . . . ." Discuss this in terms of the TCP theorem. Is there a conflict? When God stops the 'film of the world,' does the electron even *have* an arrow of time?

All the modern, major religions of the world are in agreement on these two points: (1) God created the universe and (2) At some time in the past, the universe came into existence. This does raise the question of what was God doing *before* He created the universe (see note 7 in the Introduction). In his *Confessions*, Saint Augustine comments on the conundrum that (1) and (2) are possibly in conflict. After all, if God *created* the universe then, given any time t in the finite past, He must have been doing something *before* time t, which means that for any time t in the finite past the universe already existed. Thus, the universe had no instant of creation in the finite past and so had no first moment of existence—which implies (2) is false. However, like any good philosopher of religion, Augustine not only provided this theological puzzle, but also a way to wiggle free of it. His suggested counter is the assertion that time is *itself* a creation of God, that is, He made time as *part* of creating the universe. Thus, there was no time *before* He created the universe, and so the very question of 'what was He doing before He made the universe' has no meaning. What do you think of Augustine's two arguments?

The Australian philosopher J. J. C. Smart (1920–2012) invoked five dimensional spacetime in a way very different from that of including an ‘eternity’ axis for God’s temporal time (see note 33 in “Some First Words”). What Smart argued for, instead, was *multiple* four-dimensional worlds existing together without conflict, just as an infinity of two-dimensional worlds can exist without conflict in a three dimensional space. As he wrote, “The reason why there could be two totally disparate space-times is simply the quite obvious one that two totally disparate four-dimensional spaces can exist within a suitable five-dimensional space. There is no difficulty in mathematical inconceivability here. Now let one of these four-spaces be our own spacetime world, and let the other four-space be more or less similar, in accordance with whatever story you wish to tell about it.” This idea had, in fact, been around long before Smart’s 1967 paper. In his 1898 Presidential Address to the American Mathematical Society, for example, Simon Newcomb declared “Add a fourth dimension of space, and there is room for an indefinite number of Universes, all alongside of each other, as there is for an indefinite number of sheets of paper when we pile them upon each other.” Newcomb’s idea appealed to H. G. Wells’ fancy so much that he built two novels, *The Wonderful Visit* and *Men Like Gods*, around it. In the first novel there is explicit mention of multiple worlds “lying somewhere close together, unsuspecting, as near as page to page in a book,” and the second one speaks of one parallel universe being rotated into another. John Cramer (a University of Washington physicist) repeated Newcomb’s and Wells’ parallel universes/pages-of-a-book/rotation imagery almost word for word in his 1991 novel *Twistor*.

Read these three novels, and then discuss. In particular, is Wells consistent in his presentations (as a novelist there is, of course, no reason he should be!)?

Imagine the following two events, **A** and **B**, in Minkowski spacetime: **A** is the emission of a photon, and **B** is the absorption of that photon. What is the spacetime interval between **A** and **B**? It might seem that we need to know more about the precise spatial and temporal coordinates of **A** and **B**, but in fact the interval is *always zero* for any two events connected by light (by photons). To see this, write the flat Minkowski metric as

$$\left(\frac{ds}{dt}\right)^2 = 1 - \left(\frac{dx}{dt}\right)^2$$

(continued)

and so, since $(dx/dt)^2 = 1$ (because photons—by definition!—travel at the speed of light) we have $(ds)^2 = 0$. The world line of any photon has what is called a *null* interval, as do all world lines on the surface of a light cone. Timelike world lines (in the interior of light cones, with $(dx/dt)^2 < 1$) have positive intervals, that is, $(ds)^2 > 0$. Spacelike world lines (in the exterior of light cones, in Elsewhere, with $(dx/dt)^2 > 1$) have negative intervals, that is, $(ds)^2 < 0$. This is one of the significant differences between *distances* in space (which are never negative) and *intervals* in spacetime. The American chemist G. N. Lewis (1875–1946)—see note 60 in Chap. 2—constructed a romantic illustration of this when he wrote "Any pair of points [in spacetime] which are separated by zero distance [interval] are in *virtual contact*. In other words, I may say that my eye touches a star, not in the same sense as when I say that my hand touches a pen, but in an equally physical sense." (See Lewis' paper "Light Waves and Light Corpuscles," *Nature*, February 13, 1926, pp. 236–238; Lewis was the originator of the term *photon* for a particle of light.) To understand what Lewis was getting at, you must understand that in spacetime we can have the interval between **A** and **B** as zero, and the interval between **B** and **C** as zero, but the interval between **A** and **C** may *not* be zero. To convince yourself of this, suppose the Minkowski spacetime coordinates (x, t) of **A**, **B**, and **C** are (1,3), (2,2) and (1,1), respectively. Show that $(ds_{\boldsymbol{AC}})^2 = 4$, while $(ds_{\boldsymbol{AB}})^2 = (ds_{\boldsymbol{BC}})^2 = 0$. (Hint: draw a diagram of this spacetime and simply plug the coordinates of **A**, **B**, and **C** into the metric.)

The use of the time dilation effect of high-speed space travel was used by science fiction writer Donald Wandrei (1908–1987) in his tale "A Race Through Time," *Astounding Stories*, October 1933. Initially set in 1950, this is the story of two scientists, one evil (of course!) and the other good (of course!), who develop quite different methods for travel into future times. The evil one does it with a drug that slows the metabolic processes of the body, while the good one builds an atomic-powered rocket in his home workshop! The evil scientist kidnaps the good one's girlfriend, seals the two of them inside a crystal dome, and then injects her and himself with his drug. He has arranged matters so that they will emerge from the dome in the year one million A.D. Learning what has happened, the good scientist rushes to the dome finds he can't break in, but sees an indicator dial pointing at 1,000,000. (The evil scientist has conveniently provided the dial, as well as having made the dome transparent, much as modern-day movie criminals always include a count-down clock with glowing red digits on their bombs so the hero always knows just how much time is left to disarm the bomb.)

(continued)

Returning to his rocket, the good scientist decides that he, too, will travel into the far future using time dilation. So, off he goes, on a trip like Bob's in Sect. 3.6. The story ends with an ironic twist—the good scientist thought the '1,000,000' he saw through the dome meant *one million years in the future beyond 1950*, and so he arrives back on Earth nearly 2000 years *after* his girlfriend and the evil scientist emerged from the dome. (They are, of course, long dead when the good scientist returns.)

On the outward leg of the good scientist's rocket flight we read of his "frightful speed — now thousands of light years per Earth second." Discuss this in terms of relativity theory. That is, does the rocket actually travel thousands of light years in one second of elapsed time back on Earth?

The American physicist Robert Forward (1932–2002) was, in addition to being an expert in general relativity, also a quite inventive science fiction writer. In his short story "Twin Paradox" (*Analog Science Fiction*, August 1983), for example, he used *biology* to give a surprising, ironic twist to the classic physics puzzle. The story flips the asymmetric aging of the twins by imagining that, just after the traveling twin's departure, the secret of immortality is discovered. The treatment has to be administered no later than at a certain age, however, and upon his return to Earth the traveling twin is just a bit too old for it to work. He thus becomes the last person to die of old age! In this tale, the details of the traveling twin's trip are somewhat different from Bob's trip. Read "Twin Paradox" (you can also find the story in Forward's 1995 book *Indistinguishable from Magic*), and summarize how the traveling twin's trip is accomplished.

In the mystical 1920 novel *A Voyage to Arcturus*, by the Scottish writer David Lindsay (1876–1945), we read of a spaceship that travels to Arcturus (the brightest star in the constellation Boötes, 36 light years from Earth) in just 19 h of proper time. The technical details of the trip are not explained in the novel, so assume they are the same as in Bob's trip. That is, $2T' = 68{,}400$ s as measured by a clock on the spaceship. The distance traveled in this time is

$$2\frac{c\left\{\sqrt{a^2T^2+c^2}-c\right\}}{a}$$

(continued)

where c is the speed of light, a is the constant acceleration/deceleration of the rocket (as measured on the spaceship), and T is how long an *Earth*-based clock says the trip takes. Use this formula, and the one relating T and T' in the text, to calculate

1. The value of the constant acceleration/deceleration, a (does the result seem reasonable to you?) and
2. The value of T (the length of time that passes on Earth as 19 h pass on the spaceship).

You don't have to discuss general relativity to encounter tensors of the second rank (as is the metric tensor). Electrical engineers run into such a thing, for example, when studying the lowly Ohm's law! In a copper wire, that law says the current density (the vector **J**, in units of amperes per square meter) at any point is related to the electric field (the vector **E**, in units of volts per meter) at that point by the *scalar* σ as follows: $\mathbf{J} = \sigma\mathbf{E}$, where σ (called the *conductivity*) is a single number. This says, in rectangular coordinates, that the x-component of **J** depends *only* on the x-component of **E**, and similarly for the y-component and z-components of **J** and **E**. More generally, however, each component of **J** depends on *all* of the components of **E** (as in certain crystalline structures), and so we have the equations

$$\begin{aligned} J_x &= \sigma_{11}E_x + \sigma_{12}E_y + \sigma_{13}E_z \\ J_y &= \sigma_{21}E_x + \sigma_{22}E_y + \sigma_{23}E_z \\ J_z &= \sigma_{31}E_x + \sigma_{32}E_y + \sigma_{33}E_z \end{aligned}$$

or, in matrix form

$$\mathbf{J} = \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} E_x \\ E_y \\ E_z \end{bmatrix} = \boldsymbol{\sigma}E.$$

So **J** is now related to **E** by $\boldsymbol{\sigma}$, a 3 × 3 *matrix* (9 numbers, instead of just 1). The matrix $\boldsymbol{\sigma}$ is, in fact, a tensor of rank 2. This tensor is in the three-dimensional space of the copper wire, while the 4 × 4 metric tensor matrix is in a four-dimensional spacetime, but both are tensors of rank 2. The number of numbers in a tensor of rank n, in a space of dimension d, is d^n. In general relativity, tensors of higher rank than 2 are required. For example, the

(continued)

curvature tensor has rank 4, and so in a four-dimensional spacetime it is described by $4^4 = 256$ numbers. This goes a long way in explaining why, in general relativity, computing often involves a *lot* of arithmetic! The primary characteristic of tensors of any rank is that they are invariant under a change in coordinate systems (choosing a coordinate system is an arbitrary matter, made mostly for human convenience, about which Nature is indifferent). Read more about tensors, and write an essay on how they behave under a change in coordinates. A good place to start, at the level of this book, is Lillian Lieber's *The Einstein Theory of Relativity: a trip to the fourth dimension*, Paul Dry Books 2008 (an updated version of the original 1945 book).

Chapter 4
Philosophers, Physicists, and the Time Travel Paradoxes

> *"He felt the intellectual desperation of any honest philosopher. He knew that he had about as much chance of understanding such problems as a collie has of understanding how dog food gets into cans."*[1]

4.1 Paradoxes and Their First Appearance in Science Fiction

> "There's a lot we don't know about time travel. How do you expect logic to hold when paradoxes hold, too?" "Does that mean you don't know?" "Yes."[2]

More than 30 years ago Quentin Smith, a philosopher who believes in a finite length to the past, wrote a refutation to those who believe in an infinite past and, while that paper[3] has nothing to do with the paradoxes of time travel, in the course of presenting his reasoning he included the following curious passage:

> "Why does the sun arise in the morning and not at some other time? Why do the hands of a properly functioning clock point to 12:00 at noon and midnight and not at other times? Why does the death of a person occur at a later time than his birth? The answer in all these cases is: Because by the very nature of these events they could not occur at other times. It belongs to the very nature of the sun's rising that it occur in the morning and not in the afternoon or evening. It belongs to the very nature of a properly functioning clock to point to 12:00 at noon and midnight and not at other times. And it belongs to the very nature of death to occur at a time later than a person's birth."

But what of a time traveler born in 1980 who, in 2018, enters her time machine, pushes a few buttons, and then boldly steps out into the Cretaceous period seventy million years earlier—and is promptly eaten for lunch by a passing *Tyrannosaurus*

---

[1] A time traveler admits (to himself) how perplexed he is by paradoxes in Robert Heinlein's classic tale "By His Bootstraps," *Astounding Science Fiction*, October 1941.

[2] Excerpt from a conversation between two paradox-puzzled time travelers in Larry Niven's story "Bird in the Hand," *The Magazine of Fantasy & Science Fiction*, October 1970.

[3] Q. Smith, "Kant and the Beginning of the World," *New Scholasticism*, Summer 1985, pp. 339–346.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_4

rex? Perhaps Smith himself would say that there is no contradiction between this and his third claim because, in the time traveler's *proper* time, her spectacular death does indeed come after her birth. For many, however, for a time traveler to die before her mother is born *is* a paradox, plain and simple, say what you will about proper time.

One science fiction view of time travel paradoxes is that nature would be so disrupted by them that, should one occur, the universe would be torn apart. In one story, for example, a paradox is on the verge of happening through the use of a Tipler-cylinder time machine (mentioned back in Chap. 1, and which we'll discuss in more detail in the next chapter). In response to this 'threat to common sense,' the universe 'decides' to avoid the paradox by simply eliminating the perpetrators via a local nova![4] Niven wasn't the first to use this idea. A famous story[5] by L. Sprague de Camp (1907–2000), written two decades before, had already put forth the suggestion that nature will take all required corrective action to avoid paradoxes.

In De Camp's tale we read of two big-game hunter guides who use "Professor Prochuska's time machine at Washington University"—built with the aid of a "cool thirty million" dollar grant from the Rockefeller Foundation—to operate a safari-for-hire business that transports hunters back to the late Mesozoic era. When a disgruntled client tries to go back to the day *before* a previous trip to shoot the guides (who had displeased him, or rather *would* displease him the next day), we learn just how nasty De Camp thought Mother Nature would be to avoid a paradox. (After all, the guides had *not been* shot during the safari, so they *could not be* shot.[6]): "The instant James started [to ambush the guides] the space-time forces snapped him forward to the present to prevent a paradox. And the violence of the passage practically tore him to bits [making his body look] as though it had been pulverized and every blood vessel burst, so it was hardly more than a slimy mass of pink protoplasm."

And even earlier we have a famous tale[7] that was discussed in Chap. 1, by Fredric Brown (1906–1972), a master of the special category of science fiction called the "short-short," in which everything happens in 500 words or less. As you might expect, the oddities of time travel were natural attractions for Brown's quirky talent. You'll recall that in this story the inventor of the first time machine demonstrates it to two colleagues by sending a brass cube 5 min into the future. After being placed in the machine, the cube vanishes and then 5 min later reappears. No paradoxes with that—it is a trip into the *past* that has the potential for deadly repercussions. We learn just how deadly when the inventor next declares that at three o'clock he will again place the cube in the time machine but, until then, he will hold the cube in his hand. Thus, he says, at 5 min before three the cube will vanish from his hand and immediately appear in the time machine because 5 min after that, at three o'clock, he will send it 5 min into the past. And indeed, at 5 min before three the cube does simultaneously vanish from his hand and appear in the time machine. Then, slightly before three, as the three men stand pondering what has happened, one of the observers asks what will happen if the inventor does *not* send the cube back at three o'clock? "Wouldn't there be a paradox of some sort involved?" he wonders.[8] His curiosity aroused, the inventor can't resist the experiment—and the universe promptly vanishes.

---

[4] L. Niven, "Rotating Cylinders and the Possibility of Global Causality Violation," *Analog Science Fiction*, August 1977. Niven took this title from a physics paper with that title, authored by Tipler, that had appeared three years earlier in *Physical Review D* (April 15, 1974, pp. 2203–2206).

[5] L. Sprague de Camp, "A Gun for Dinosaur," *Galaxy Science Fiction*, March 1956.

[6] This is a statement of the belief that the past cannot be changed, an idea we will examine later in this chapter.

[7] Look back at note 93 in Chap. 1.

Time travel, of course, is full of potential paradoxes. A paradox, according to the usual dictionary definition, is something that appears to contain contradictory or incompatible parts, thus reducing the whole to seeming nonsense. And yet, truth is also evident in the whole. The history of science and mathematics has left a long trail of paradoxes, and those that involve time travel are merely among the most recent. Not all of the puzzles of time travel involve physics or logic, however. As one philosopher observed, "Doubtless time travel will raise a host of legal difficulties, e.g., should a time traveler who punches his younger self (or vice versa) be charged with assault? Should the time traveler who murders someone and then flees into the past for sanctuary be tried in the past for his crime committed in the future? If he marries in the past can he be tried for bigamy even though his other wife will not be born for almost 5000 years? Etc., etc. I leave such questions for lawyers and writers of ethics textbooks to solve."[9]

One way early science fiction writers had of responding to the puzzle of time travel paradoxes was to just give up and to concede that the logical puzzles are overwhelming. In one tale, for example, the inventor of the Chronoscope (a gadget that can only *view* the past) explains, "There is no time travel machine. Such a thing is a logical impossibility, treated seriously only by half-cracked writers of fantasy. Such a machine would lead at once into a hopeless paradox."[10] Equally concerned about time travel paradoxes was the pulp science fiction time traveler who told his partner, just before their first trip in time, that "I'm not sure any more about getting back. There're some unpredictable terms in the time-travel equation—paradoxes. Maybe we *won't* get back."[11]

---

[8] This is what is called a *bilking paradox*, and such paradoxes will be discussed later in this chapter. Brown gave this story a lot of thought. At one point in the tale one of the colleagues, puzzled by how the inventor will be able to place the cube into the time machine at three if it has already vanished from his hand and appeared in the machine, asks "How can you place it there, then?" Replies the inventor, "It will, as my hand approaches, vanish from the [machine] and appear in my hand to be placed there."

[9] L. Dwyer, "Time Travel and Some Alleged Logical Asymmetries Between Past and Future," *Canadian Journal of Philosophy*, March 1978, pp. 15–38.

[10] M. Jameson, "Dead End," *Thrilling Wonder Stories*, March 1941.

[11] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

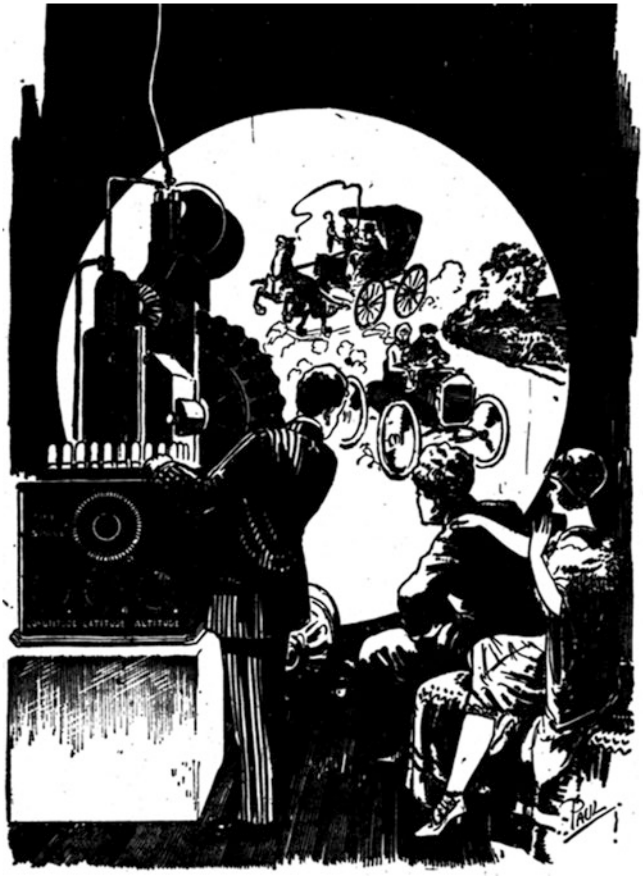

Fig. 4.1 One way pulp science fiction avoided paradoxes was to use a ‘time viewer’ (like the chronoscope in “Dead End”) as in “The Time Eliminator” (*Amazing Stories*, December 1926). This illustration from the story (authored by somebody who used only the initials K.A.W.) by Frank R. Paul (©1926 by Experimenter Publishing Co.) shows the inventor demonstrating his gadget to his future wife and father-in-law. Able to look back in time, the screen is displaying scenes from the older man’s courtship of his wife, decades in the past. Reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

But are there *really* paradoxes? Or is it true, as the extraordinary boy-prodigy who invented a time machine exclaimed (when his teacher asserted that some questions could never be answered because “Nature is full of paradoxes”), “Ah, Professor, what nonsense! Nature is harmonious; it is we who bring the paradoxes into it.”[12] Saying the same are two physicists, in a paper on the circular orbits of photons around black holes: “There are no paradoxes in physics, but only in our attempts to understand physical ideas by using inadequate reasoning or false intuition."[13] And as the time traveler in an early pulp story[14] casually declares to a friend, after an astonishing adventure in the year A.D. 1,001,930, "Paradoxical? My dear fellow, the Einstein Theory is full of apparent paradoxes, yet to him who understands it there is no inconsistency whatever. Give me another cigarette, will you, Frank?"

---

[12] V. Grigoriev, “Vanya,” in *Last Door to Aiya* (M. Ginsburg, editor), S. G. Phillips 1968.

It is in *Amazing Stories* that we find the first non-fictional speculations about time travel in a pulp magazine—and certainly long before any *physics* journal would touch the subject! Publisher and editor Hugo Gernsback started those speculations by reprinting Wells' *Time Machine*, which in turn sparked a fair number of readers' letters that were printed in the magazine's "Discussions" section. Typical of the less interesting is the following comment from a letter in the July 1927 issue: "In the 'Time Machine' I found something amiss. How could one travel to the future in a machine when the beings of the future have not yet materialized?" (We answered that question in the previous chapter with the twin paradox.) Far more interesting was this letter, in the same issue:

> "How about this 'Time Machine'? Let's suppose our inventor starts a 'Time Voyage' backward to about A.D. 1900, at which time he was a schoolboy . . . His watch ticks forward although the clock on the laboratory wall goes backward. Now we are in June 1900, and he stops the machine, gets out and attends the graduating exercises of the class of 1900 of which he was a member. Will there be another 'he' on the stage? Of course, because he *did* graduate in 1900 . . . Should he go up and shake hands with this 'alter ego'? Will there be two physically distinct but characteristically identical persons? Alas! No! He can't go up and shake hands with himself because . . . this voyage back through time only duplicates actual past conditions and in 1900 this strange 'other he' did *not* appear suddenly in quaint ultra-new fashions and congratulate the graduate. How could they both be wearing the same watch they got from Aunt Lucy on their seventh birthday, the same watch in two different places at the same time. Boy! Page Einstein! No, he cannot be there because he wasn't there in 1900 (except in the person of the graduate) . . . The journey backward must cease on the year of his birth. If he could pass that year it would certainly be an effect before a cause . . . Suppose for instance in the graduating exercise above, the inventor should decide to shoot his former self . . . He couldn't do it because if he did the inventor would have been cut off before he began to invent and he would never have gotten around to make the voyage, thus rendering it impossible for him to be there taking a shot at himself, so that as a matter of fact he *would* be there and *could* take a shot—help, help, I'm on a vicious circle merry-go-round . . . Now as to trips into the future, I could probably think up some humorous adventures wherein [the inventor] digs up his own skeleton and finds by the process of actual examination that he must expect to have his leg amputated because the skeleton presents positive proof that this was done."[15]

---

[13] M. A. Abramowicz and J. P. Lasota, "On Traveling Round Without Feeling It and Uncurving Curves," *American Journal of Physics*, October 1986, pp. 936–939.

[14] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

[15] This story idea (the letter was signed only with the initials T.J.D.) may well have been the inspiration for R. Rocklynne, "Time Wants a Skeleton," *Astounding Science Fiction*, June 1941. Not all fans agreed with T.J.D. A few years later, for example, a teenager named P. Schuyler Miller (1912–1974), who would author several time travel classics himself, wrote a letter to the editor of *Astounding Stories* (June 1931) stating "there is nothing in physics . . . to prevent yourself from

All of the ingenious puzzles in this letter intrigued Gernsback, and may have, in fact, been the cause of his featuring a new, original time machine story[16] in the same issue. It was the tale of a scientist who transports an entire ship at sea 14,000 years back in time and causes it to hover over lost Atlantis! That story provoked a sharp letter from a reader who claimed its logic had a fatal flaw: the story's author indicated that the Atlantians observed the time travelers when, 'of course' (asserted the reader), the time travelers must actually have been invisible. The reader explained his reasoning as follows, beginning by defining A as one of the Atlantians.

> "Now A lived his life, thousands of years ago, and died. All right, now let us pass on in time 14,000 years. Now, back we come in time when A is again living his life. Lo and behold, this time A sees before he dies a strange phenomenon in the sky! He sees the shipload of people observing him. And yet these people are necessarily observing him during his one and only lifetime, wherein he certainly did not, could not, have observed them."

Gernsback printed this letter in his September 1927 editorial "The Mystery of Time," and concluded by saying "I do . . . agree . . . that the inhabitants of Atlantis would probably not have seen the . . . travelers in time." Other readers felt this way, too, because after Gernsback published yet another time machine tale[17] in 1927, the same invisibility argument appeared again in the magazine's "Discussions" column.

Two years later an amateurishly written tale[18] appeared in which a man travels in time from 1928 to 2930 with the aid of an "astounding machine based on advanced electro-physics and the non-Euclidean theory of hyperspace." The purpose of that story was two-fold: to present several of the classic paradoxes of time travel, and then to make the claim that although the simple minds of twentieth-century people cannot understand the explanations of the paradoxes (possibly explaining why the author offers none!), the paradoxes are all trivial to the scientists of the thirtieth century. Despite this shortcoming (as well as some pretty awful dialog) the story nonetheless still managed to entertain readers with the sheer mystery of the paradoxes. Letters poured into the magazine from young fans, all demanding more time travel fiction.

So, that same year Gernsback responded with a story[19] that plays with the question of the role of time travelers in the past. (That question was clearly 'in

---

going into the past . . . and shaking hands with yourself or killing yourself." That did, however, provoke the following harsh reply from another, more skeptical reader (in the December 1933 issue): "P. S. Miller once wrote that time traveling is not incompatible with any laws of physics . . . 'he don't know from nothing.'"

[16] C. B. White, "The Lost Continent," *Amazing Stories*, July 1927.

[17] F. Flagg, "The Machine Man of Ardathia," *Amazing Stories*, November 1927.

[18] C. Cloukey, "Paradox," *Amazing Stories Quarterly*, Summer 1929.

[19] H. F. Kirkham, "The Time Oscillator," *Science Wonder Stories*, December 1929.

the air,' as Gernsback's old magazine *Amazing*[20] simultaneously published a story[21] addressing this same puzzle of time travel.) Could time travelers actually participate in events ("mix into the affairs of the period," in Gernsback's words), or would they just be unseen observers? This question, obviously inspired by the earlier discussion in *Amazing Stories*, intrigued Gernsback as much as it did his readers and so, along with Kirkham's story, he printed a challenge titled "The Question of Time-Traveling" (see note 18 in the Introduction):

> "In presenting this story to our readers, we do so with an idea of bringing on a discussion as to time traveling in general. The question in brief is as follows: Can a time traveler, going back in time—whether ten years or ten million years—partake in the life of that time and mingle in with its people; or must he remain suspended in his own time-dimension, a spectator who merely looks on but is powerless to do more? Interesting problems would seem to arise, of which only one need be mentioned: Suppose I can travel back into time, let me say 200 years; and I visit the homestead of my great great great grandfather, and am able to take part in the life of his time. I am thus enabled to shoot him, while he is still a young man and as yet unmarried. From this it will be noted that I could have prevented my own birth; because the line of propagation would have ceased right there. Consequently, it would seem that the idea of time traveling into a past where the time traveler can freely participate in activities of a former age, becomes an absurdity. The editor wishes to receive letters from our readers on this point: the best of which will be published in a special section."

Gernsback's challenge did not go unnoticed and, over the next year or so, he published a large number of reader responses in the magazine's letters column.

Indeed, a few months after issuing the challenge, in his introduction to another time travel tale,[22] Gernsback wrote that ever since the publication of Kirkham's tale "there has been a great controversy among our readers as to the possibility of time flying and the conditions under which it may be done." Most of those letters are interesting if not particularly profound, with one exception. That was a letter written by a 14 year old boy in San Francisco, and its appearance in the February 1931 issue of *Science Wonder* may well have served as the inspiration for several of the classic time travel tales published during the next 20 years:

> "Some time ago you asked us (the readers) what our opinions on time traveling were. Although a bit late, I am now going to voice four opinions …
>
> (1) Now, in the first place if time traveling were a possibility there would be no need for some scientist getting a headache trying to invent an instrument or 'Time-Machine' to 'go back and kill grandpa' (in answer to the age-old argument of preventing your birth by killing your grandparents I would say: 'who the heck would want to kill his grandpa or gandma!'[23]) I figure it out thusly: A man takes a time machine, and travels into the

---

[20] By this time Gernsback had lost control of *Amazing*, and *Science Wonder* was his come-back as a publisher of pulp 'scientifiction.'

[21] E. L. Rementer, "The Time Deflector," *Amazing Stories*, December 1929. Gernsback may well have been the editor, before he lost *Amazing*, who bought this story, and the magazine's new management simply used what remained in inventory.

[22] F. Flagg, "An Adventure in Time," *Science Wonder Stories*, April 1930.

[23] Look back at note 26 in "Some First Words."

> future from where he sends it (under automatic control) to the past so that he may find it and travel into the future and send it back to himself again. Hence the time machine was never invented, but!—from whence did the machine *come*?
>
> (2) Another impossibility that might result could be: A man travels a few years into the future and sees himself killed in some unpleasant manner,—so—after returning to his correct time he commits suicide in order to avert death in the more terrible way which he was destined to. Therefore how could he have seen himself killed in an entirely different manner than really was the case?
>
> (3) Another thing that might corrupt the laws of nature would be to: Travel into the future; find out how some ingenious invention of the time worked; return to your right time; build a machine, or whatever it may be, similar to the one you had recently learned the workings of; and use it until the time you saw it arrives, and then if your past self saw it as you did, he would take it and claim it to be an invention of his (your) own, as you did. Then—who really *did* invent the consarn thing?
>
> (4) Here's the last knock on time traveling: What if a man were to travel back a few years and marry his mother, thereby resulting in his being his own 'father'?"
>
> Jim H. Nicholson

Gernsback's reply, immediately following this letter, was favorable, opening with "Young Mr. Nicholson does present some of the more humorous [?] aspects of time traveling. Logically we are compelled to admit that he is right—that if people could go back into the past or into the future and partake of the life in those periods, they could disrupt the normal course of events."[24] Nicholson's letter *is* ingenious, and it anticipated the central ideas of a number of science fiction tales yet to be written.[25] However, as you'll see as you read the rest of this chapter, contrary to Gernsback's view Nicholson's comments are *not* logically correct.[26]

---

[24] Despite these words, Gernsback apparently hadn't given up entirely on the 'invisibility of time travelers' view, as he had only a few months earlier published another such tale: R. A. Palmer, "The Time Ray of Jandra," *Wonder Stories*, June 1930. In this story (one either silly or hilarious, take your pick) a time traveler moves into the future by means of a 'time ray.' Unfortunately, the ray works differently on the various chemical elements, and not at all on either hydrogen or oxygen. Thus the time traveler—or at least *much* of him—and his machine do vanish into the future, but left behind are "several gallons of water spilled on the floor." (The human body is about 60 % H_2O.)

[25] For example, Nicholson's item (2) is a precise plot outline for L. Raphael, "The Man Who Saw Through Time," *Fantastic Adventures*, September 1941, and a version of item (4) is in Robert Heinlein's famous "All You Zombies—,“ *Magazine of Fantasy & Science Fiction*, March 1959.

[26] One cannot, however, fault the imaginative powers of James Nicholson (1916–1972). He eventually became President of American International Films, the company that made such science fiction 'classics' as *Attack of the Crab Monsters* (1957), the 1963 *X (The Man with the X-Ray Eyes)*, and *The Time Travelers* (1964).

4.2 Changing the Past and the Grandfather Paradox

> "I'm not kidding you at all Phil," Barney insisted. "I have produced a workable Time Machine, and I am going to use it to go back and kill my grandfather."[27]

As mentioned in "Some First Words," physicists and philosophers often have quite different approaches to time travel (see note 26 there again). A vivid illustration of that difference is found in a philosophical paper[28] that, after acknowledging the apparent restrictions of the grandfather paradox, turns its attention to a matter that almost surely is beyond the power of physics to study—namely, the nature of the *conversation* between a time traveler and his/her younger self. A physicist, on the other hand, views the restrictions as *the whole point*, because the central question that hovers over *any* discussion of time travel is that of 'changing the past.' As one science fiction fan summed it up in a letter to the editor of *Astounding Stories* (August 1931): "It is said that the past cannot be changed, and that any effort to do so would be useless. In my belief, no matter where or when a man goes in the past, if he appears in a year or day that has already gone by, *he is changing the past*. Then there should be no room for doubt: time traveling is impossible. It will never be done." And certainly killing your grandfather in the past (when he is still a baby) would qualify as a change to the past!

The idea of 'changing the past' occurred to the minds of philosophers long before it did in those of science fiction writers or physicists. Four centuries before Christ, the question had already been asked-and-answered by Aristotle. In his *Nicomachean Ethics*, in fact, we find him declaring that the Greek poet Agathon had known the answer a century earlier, and he quotes the poet as saying "Forever God lacks this one thing alone, To make a deed that has been done undone."

Agathon and Aristotle aside, some medieval theologians argued passionately that the past *could* be changed (but only by God). The eleventh-century Italian cleric Peter Damian (who became a Christian saint) is a famous exponent of that radical view.[29] Writing in his *De Omnipotentia Dei* ("On the Divine Omnipotence in Remaking What Has Been Destroyed and Undoing What Has Been Done"),[30] Damian made it clear that he believed nothing could withstand the power of God, not even the past. Ralph Waldo Emerson's poem "The Past" ("All is now secure and fast, Not the gods can shake the Past") would have been blasphemy for Damian. The following words from Damian testify to the strength of his commitment to a belief in the possibility of changing the past: "Just as we can duly say 'God was able to make it so [that] Rome, before it had been founded, should not have been founded,' in the same way we can equally and suitably say, 'God can make it so that Rome, even after it was founded, should not have been founded.'"[31]

---

[27] The opening line to F. M. Busby, "A Gun for Grandfather," *Future Science Fiction*, September 1957.

[28] Jiri Benovsky, "Endurance and Time Travel," *Kriterion—Journal of Philosophy*, 2011, pp. 65–72.

[29] R. P. McArthur and M. P. Slattery, "Peter Damian and Undoing the Past," *Philosophical Studies*, February 1974, pp. 137–141; P. Remnant, "Peter Damian: Could God Change the Past?" *Canadian Journal of Philosophy*, June 1978, pp. 259–268; R. Gaskin, "Peter Damian on Divine Power and the Contingency of the Past," *The British Journal for the History of Philosophy*, September 1997, pp. 229–247.

Two centuries after Damian, Aquinas argued the contrary view, that changing the past is *not* within God's power. Whereas Damian felt it impossible to deny any act to God, Aquinas took the far more moderate position that part of God's law is that there be no contradictions in the world (this is, in fact, the modern view of time travel physicists) and that certainly God would be bound by his own law. As he wrote, "It is best to say that what involves contradiction cannot be done rather than that God cannot do it." In his *Paradise Lost*, John Milton's God is constrained even more: he is free to act or not, but if he does freely decide to act, it can only be to 'do right.' That might seem to preclude causing contradiction, as in changing the past, but perhaps not. Milton's contemporary, Thomas Hobbes, declared that there is no *a priori* standard of goodness, and thus (for Hobbes) there are no constraints on God's powers. For Hobbes, therefore, it would seem God could change the past. Theological changing of the past, as you might expect, leads to all sorts of mind-boggling puzzles. Because of such puzzles, theology would certainly be influenced by time travel, but just as certainly theological reasoning will not illuminate the puzzles of the time travel paradoxes.

The question of the immutability of past events is of special interest to theologians because it is directly related to the question of free will versus fatalism. That is, are humans the creators of the future, or are they mere fated puppets of destiny? One way theology gets involved with the issue of 'changing the past' is via what is called the *retroactive petitionary prayer*. (An 'ordinary' petitionary prayer, like the Lord's Prayer in Matthew 6 and Luke 11, asks for something in the present or the future.) Examples of retroactive prayers include that of the surgical patient who prays, just before an exploratory operation, for his tumor to be non-malignant, and that of the soldier's wife who prays that her husband was not among those killed in yesterday's battle. Such prayers are for a happy outcome to an event that is over and done with at the time of the prayer. One might accept the rationality of praying about the future ("Please, God, let me survive tomorrow's battle and I'll be good for the rest of my life") but are prayers about the past even sensible? (The three major monotheistic religions of the world—Christianity, Judaism, and Islam—say *yes*.)

---

[30] This work is in the form of a letter to his friend Desiderius (who later became Pope Victor III), in which Damian rebutted Desiderius' defense of St. Jerome's claim that "while God can do all things, he cannot cause a virgin to be restored after she has fallen." Desiderius thought the reason God could not restore virgins is that he does not want to, to which Damian replied that this meant God is unable to do whatever he does not want to do, but this meant that God would then be less powerful than men, who are able to do things they don't want to do (such as go without food for a month). This is a good example of the dangers involved when getting into debates with theologians.

[31] The Argentinian writer Jorge Luis Borges (1899–1986) was so inspired by Damian's view that the past could be changed that he wrote a short story based on it (see "The Other Death," originally published in *The New Yorker*, November 2, 1968) and put a character in it named after Damian.

In an appendix titled "Special Providences" in his book Miracles, C. S. Lewis answers that question as follows:

> "When we are praying about the result, say, of a battle or a medical consultation, the thought will often cross our minds that (if only we knew it) the event is already decided one way or the other. I believe this to be no good reason for ceasing our prayers. The event certainly has been decided—in a sense it was decided 'before all worlds.' But one of the things taken into account in deciding it, and therefore one of the things that really causes it to happen, may be this very prayer that we are now offering. Thus, shocking as it may sound, I conclude that we can at noon become part causes of an event occurring at ten A.M. (Some scientists would find this easier than popular thought does.)"

Here we see Lewis, a prominent lay theologian, arguing for the present influencing (but not changing) the past. What can we make of that? Was Lewis arguing for backward causation, the close relative of time travel? I think perhaps so; the final two sentences in the above excerpt makes it plausible that he may have held that view. It is a view that does find much support in the block universe interpretation of Minkowskian spacetime. Lewis never mentions the block concept by name, but it is clear that he believed in the idea of God being able to see all of reality at once. Lewis believed, therefore, that God knows of a petitionary prayer before it is made; or, even stronger, if God is not a temporal being but rather is eternal and knows time 'all at once,' then God knows the prayer and the event being prayed about 'at the same time.'

Lewis did make it clear that he believed it is a sin to pray for something *known* not to have occurred—for example, to pray for the safety of someone known to have been killed yesterday. As he wrote, "The known event states God's will. It is psychologically impossible to pray for what we know to be unobtainable, and if it were possible, the prayer would sin against the duty of submission to God's known will." Taking a less judgmental position (but essentially agreeing with Lewis) were two philosophers who, writing of the battle of Waterloo, said "for one who knows the outcome of the battle more than a hundred and 50 years ago, [a retroactive petitionary] prayer is pointless and in that sense absurd. But a prayer prayed in ignorance of the outcome of the past event is not pointless in that way." Further, in support of backward causation, they also wrote that "to pray in 1980 that Napoleon lose at Waterloo" is logical because "why should your prayer not be efficacious in bringing about Napoleon's defeat?"[32] Disagreeing, however, was another philosopher who bluntly declared "A prayer for something to have happened is simply an absurdity, regardless of the utterer's knowledge or ignorance of how things went."[33]

---

[32] E. Stump and N. Kretzmann, "Eternity," *Journal of Philosophy*, August 1981, pp. 429–458.

[33] P. Geach, *God and the Soul*, Routledge & Kegan Paul 1969.

There have been all sorts of opinions expressed through the ages in reaction to the idea of affecting the past via retroactive petitionary prayers,[34] and on the role of backward causation. The British philosopher Michael Dummett (1925–2011), in particular, discussed Lewis' concept of such prayers with great sympathy,[35] and backward causation allows one to both explain them as well as retaining free will. That is, it is not God's foreknowledge that causes our later actions (forcing our behavior and so turning us into automatons), but rather it is our later *freely-chosen* actions that causes God's foreknowledge! While such theological speculations are interesting, in the end they are simply positions of faith, about which mathematical physics has nothing to say.

Eventually, of course, others besides philosophers and theologians began to ponder the questions raised by 'changing the past.' In a January 1963 personal letter to the editor of *The Magazine of Fantasy and Science Fiction*, Robert Heinlein wrote[36] "Mark Twain invented the time-travel story; 6 years later H. G. Wells perfected it *and its paradoxes* [my emphasis]. Between them they left little for latecomers to do." How a man as widely read as was Heinlein, who had authored some of the best short time travel stories ever written, could have written such an erroneous sentence is a mystery to me. *A Connecticut Yankee in King Arthur's Court* and *The Time Machine*, certainly both works of genius, are *not* pioneers in paradox. And Heinlein's own contributions are proof enough that there was a lot left to do with time travel, well after 1900.

The very first story to be written that even hints at the particular time travel paradox of changing the past seems to be by the Unitarian minister Edward Everett Hale (1822–1909), best known today as the author of the 1863 story "The Man Without a Country." Hale wrote "Hands Off" in 1881, and published it anonymously in *Harper's New Monthly Magazine* with the express purpose of stirring up some theological debate (which apparently it didn't). He certainly had no idea that he would come to be recognized by literary scholars as a pioneer in the yet-to-be invented genre of science fiction.

Hale's story opens with the mysterious words "I was in another stage of existence. I was free from the limits of Time, and in new relations to space." These words are spoken by an unnamed narrator, who seems to have just died, and who finds himself, in his new 'form,' observing "some twenty or thirty thousand solar systems" while in the company of "a Mentor [probably an angel] so loving and patient." Under the guidance of this Mentor, in an attempt to 'improve' history, the narrator alters the Biblical account of Joseph and his imprisonment in Egypt. At first, subsequent history *is* better, but then humanity sinks into irreversible depravity. In the end the narrator watches the last handful of humans kill each other in a particularly symbolic place for the Christian world: "The last of these human brutes all lay stark dead on the one side and on the other side of the grim rock of Calvary!" There would be no Crucifixion and Resurrection for the salvation of humankind, which naturally greatly disturbs the narrator. But the Mentor calms him, saying "Do not be disturbed, you have done nothing." It has, you see, just been an experimental world, an alternate Universe, and the narrator has learned the lesson of "Hands Off" the past.

---

[34] A summary of those opinions can be found in G. Brown, "Praying About the Past," *Philosophical Quarterly*, January 1985, pp. 83–86. Debate continues on the retroactive prayer into the 21st century: see, for example, K. Timpe, "Prayers for the Past," September 2005, pp. 305–322, and T. J. Mawson, "Praying for Known Outcomes," March 2007, pp. 71–87, both in *Religious Studies*.

[35] M. Dummett, "Bringing About the Past," *Philosophical Review*, July 1964, pp. 338–359.

[36] Reprinted in the posthumously published *Grumbles from the Grave* (edited by Heinlein's widow, Virginia Heinlein), Del Rey 1990.

Hale's story is a better Sunday sermon than it is a change-the-past time travel tale, and the device of experimenting on a not-really-real Earth is disappointing from a modern science fiction point of view. But Hale's story almost certainly did have an immediate (if indirect) impact. There is no absolute proof, but with its appearance in a national magazine, it seems quite likely that "Hands Off" was read by Edward Page Mitchell (1852–1927), an editor on a daily New York newspaper, the *Sun*. I write that because, just 6 months after Hale's story appeared, the *Sun*, in its issue of September 18, 1881, printed Mitchell's "The Clock That Went Backward." That tale, published anonymously, used a machine (the clock) for time travel,[37] as well as incorporated the idea of time travel involving paradoxes. The story predates Wells' *Time Machine* by 14 years, and Wells' novel did not include a paradoxical element.[38]

There are, however, two *hints* at paradox in Wells' novel. In the opening, during the dinner party at which the Time Traveller tries to convince his friends of the possibility of a time machine, one of them observes that "It would be remarkably convenient for the historian. One might travel back and verify the accepted account of the Battle of Hastings, for instance." To that another guest replies, "Don't you think you would attract attention? Our ancestors had no great tolerance for anachronisms." The second hint occurs when the incredulous Editor, astonished at the disheveled appearance of the Time Traveller upon his return from the future, wonders "What *was* this time traveling? A man couldn't cover himself with dust by rolling in a paradox, could he?"

What might happen if time travelers *could* change the past? This question is nicely illustrated in one novel[39] where a time traveler finds himself stranded in the London of 1810. Despite his predicament, he takes solace with "I could invent things—the light bulb, the internal combustion engine, . . ., flush toilets . . . " But then he thinks better of doing any of that: "no, better not to do anything to change the course of recorded history—such tampering might cancel the trip I got here by, or even the circumstances under which my mother and father met. I'll have to be careful."

---

[37] See Jan Pinkerton, "Backward Time Travel, Alternate Universes, and Edward Everett Hale," *Extrapolation*, Summer 1979, pp. 168–175. The time machine in Mitchell's story is more fantasy than anything else. It is simply stated that if the clock runs backward, then it travels backward in time.

[38] Wells' failure to use paradox in his famous novel surprises most modern readers and, in fact, one of the first reviewers specifically criticized him for this lapse. See the 1895 review of *The Time Machine* that appeared in *Pall Mall Magazine*, by Israel Zangwill, reprinted in Parrinder's book (note 1 in the Introduction).

[39] T. Powers, *The Anubis Gates*, Ace 1983, a work with equal shares of physics and magic.

This is really just a more recent treatment of the change-the-past paradox that was already well established in early science fiction. In a story[40] published a half-century before, we find the paradox explicitly stated, along with a possible solution that is similar to the kind of explanations that have appeared in the philosophical literature[41]:

> "Suppose you landed in your own past?," queried Eric."
>
> Dow smiled.
>
> "The eternal question," he said. "The inevitable objection to the very idea of time travel. Well, you never did, did you? You know it never happened!"

But, suppose you *could* land in your own past. What then?

One famous story[42] that considered this question *embraced* the idea of changing the past. In it a client on a dinosaur hunting safari fails to follow the instructions of his guide to do *nothing* in the past except shoot a dinosaur that is about to dies for "other reasons" anyway—alas, he accidently kills a butterfly. This results in enormous changes in history, as indicated by the 'before' and 'after' versions of the time machine company's ad:

before
TIME SAFARI, INC.
SAFARIS TO ANY YEAR IN THE PAST.
YOU NAME THE ANIMAL.
WE TAKE YOU THERE.
YOU SHOOT IT.
after
TYME SEFARI INC.
SEFARIS TU ANY YEER EN THE PAST.
YU NAIM THE ANIMALL.
WE TAEK YU THAIR.
YU SHOOT ITT.

Bradbury describes the death of the butterfly as having started the knocking "down [of] a line of small dominoes and then big dominoes and then gigantic dominoes, all down the years across Time." This is, of course, a somewhat unconvincing argument. After all, previous dinosaurs, when shot, must have fallen to the ground and flattened a lot of butterflies! With such threats for every decision, no matter how seemingly innocent, hanging over the head of a time traveler, it would take a brave soul to do much more, while in the past, than just stand still and breathe.

---

[40] C. L. Moore, "Tryst in Time," *Astounding Stories*, December 1936.

[41] See, for example, P. J. Riggs, "The Principal Paradox of Time Travel," *Ratio*, April 1997, pp. 48–64. The 'principal paradox' is that time travel is inherently contradictory because it permits the possibility of traveling to an earlier time to prevent the trip. The grandfather paradox is a special case of this. For more discussion, see T. Chambers, "Time Travel: How Not to Defuse the Principal Paradox," *Ratio*, September 1999, pp. 296–301.

[42] R. Bradbury, "A Sound of Thunder," *Collier's*, June 1952.

Equally grim is the tale[43] that takes a different view by denying the past can be changed. There we read of a doomed hero who journeys back to 1865 to save Lincoln from Booth, but his "time-distorter" is quickly taken away from him by suspicious guards. Its internal workings tick, you see, and they think he is an assassin with a clock bomb. They destroy it, haul him away to his fate, and Lincoln goes on to meet his. In the same spirit (but even more shocking) is the result of a time traveler's intentional tampering with the past in David Gerrold's 1973 novel *The Man Who Folded Himself*. That traveler experiments with "making things different" and, in his words, "Once I created a world where Jesus Christ . . . went out into the desert to fast and never came back. The twentieth century I returned to was—different. Alien."

With such a stupendous power to alter reality, assuming the past *can* be changed, perhaps one might imagine prospective time travelers to the past being required to first file Historical Impact Statements![44] Not all would receive permission. In one classic tale,[45] for example, we read of a time traveler who takes a rifle and 5000 rounds of explosive bullets back to Golgotha. His intention—to be history's first Rambo by picking off any Roman soldier who gets within a hundred yards of Jesus! As outrageous as this concept is (but who among those now reading this won't admit to at least a momentary thrill at the idea and, perhaps, even a secret willingness to do it themselves, if they could), it isn't the story's peak. That comes when the reader is reminded that it was Christ's *desire* to die on the Cross, that he *had* to die for our sins; to prevent that from happening would change all of history for the last 2000 years. What, then, should the time traveler's colleagues do when they discover his plan? Should they stop him or not? What might happen if they do interfere? Of course, if the time traveler is 'now' in the past, isn't it already 'too late' to stop him? Oh, the conundrums of time travel and changing the past!

The classic change-the-past paradox is, of course, the grandfather paradox. A famous story[46] pushed this paradox to its logical limit to illustrate its supposed dangers. Having traveled to Greece in the fifth century B.C., the traveler suddenly realizes (with just a little exaggeration): "Ninety-five generations back you'd have more grandfathers than there are people on Earth, or stars in the Galaxy! You're kin to everyone . . . You as much as take a poke at anyone, and the odds are you won't even get to be a twinkle in your daddy's eye."

---

[43] R. Silverberg, "The Assassin," *Imaginative Tales*, July 1957.

[44] See, for example, the novel by C. L. Harness, *Krono*, Franklin Watts 1988.

[45] A. Porges, "The Rescuer," *Analog Science Fiction*, July 1962.

[46] P. S. Miller, "Status Quondam," *New Tales of Space and Time*, November 1951. This is the Miller I mentioned back in note 15.

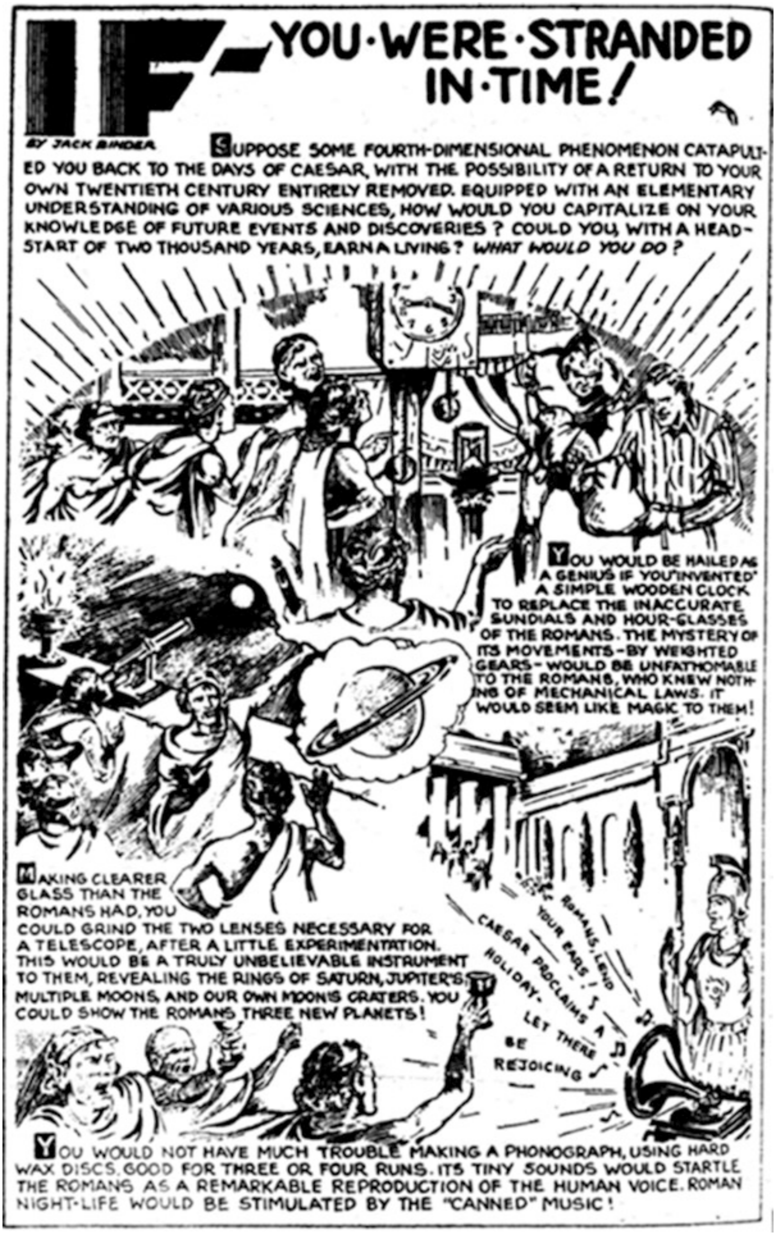
IF—YOU·WERE·STRANDED IN·TIME!
BY JACK BINDER
SUPPOSE SOME FOURTH-DIMENSIONAL PHENOMENON CATAPULTED YOU BACK TO THE DAYS OF CAESAR, WITH THE POSSIBILITY OF A RETURN TO YOUR OWN TWENTIETH CENTURY ENTIRELY REMOVED. EQUIPPED WITH AN ELEMENTARY UNDERSTANDING OF VARIOUS SCIENCES, HOW WOULD YOU CAPITALIZE ON YOUR KNOWLEDGE OF FUTURE EVENTS AND DISCOVERIES ? COULD YOU, WITH A HEAD-START OF TWO THOUSAND YEARS, EARN A LIVING? WHAT WOULD YOU DO ?
YOU WOULD BE HAILED AS A GENIUS IF YOU "INVENTED" A SIMPLE WOODEN CLOCK TO REPLACE THE INACCURATE SUNDIALS AND HOUR-GLASSES OF THE ROMANS. THE MYSTERY OF ITS MOVEMENTS—BY WEIGHTED GEARS—WOULD BE UNFATHOMABLE TO THE ROMANS, WHO KNEW NOTHING OF MECHANICAL LAWS. IT WOULD SEEM LIKE MAGIC TO THEM!
MAKING CLEARER GLASS THAN THE ROMANS HAD, YOU COULD GRIND THE TWO LENSES NECESSARY FOR A TELESCOPE, AFTER A LITTLE EXPERIMENTATION. THIS WOULD BE A TRULY UNBELIEVABLE INSTRUMENT TO THEM, REVEALING THE RINGS OF SATURN, JUPITER'S MULTIPLE MOONS, AND OUR OWN MOON'S CRATERS. YOU COULD SHOW THE ROMANS THREE NEW PLANETS!
ROMANS, LEND YOUR EARS! CAESAR PROCLAIMS A HOLIDAY. LET THERE BE REJOICING
YOU WOULD NOT HAVE MUCH TROUBLE MAKING A PHONOGRAPH, USING HARD WAX DISCS, GOOD FOR THREE OR FOUR RUNS. ITS TINY SOUNDS WOULD STARTLE THE ROMANS AS A REMARKABLE REPRODUCTION OF THE HUMAN VOICE. ROMAN NIGHT-LIFE WOULD BE STIMULATED BY THE "CANNED" MUSIC!

Fig. 4.2 Illustrator Jack Binder (1902–1986) was the author of a continuing series called "IF—. . ." in *Thrilling Wonder Stories*. In each issue, the ellipses would be replaced with some phrase such as "the Sun exploded!," "there was another ice age!," or "there was no friction!." The installment shown here (and in Fig. 4.3) appeared in December 1938 and asserted that the past could be changed by a time traveler

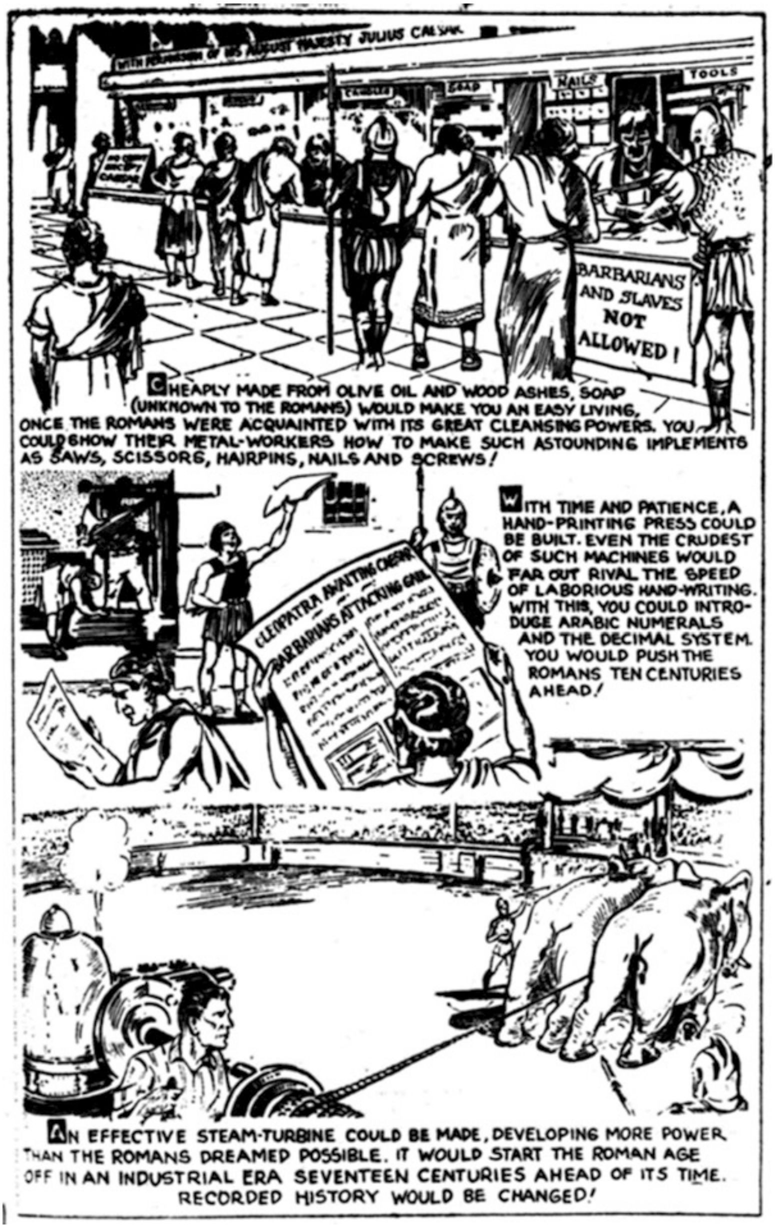
WITH PERMISSION OF HIS AUGUST MAJESTY JULIUS CAESAR
CANDLES
SOAP
NAILS
TOOLS
BARBARIANS AND SLAVES NOT ALLOWED!
CHEAPLY MADE FROM OLIVE OIL AND WOOD ASHES, SOAP (UNKNOWN TO THE ROMANS) WOULD MAKE YOU AN EASY LIVING, ONCE THE ROMANS WERE ACQUAINTED WITH ITS GREAT CLEANSING POWERS. YOU COULD SHOW THEIR METAL-WORKERS HOW TO MAKE SUCH ASTOUNDING IMPLEMENTS AS SAWS, SCISSORS, HAIRPINS, NAILS AND SCREWS!
CLEOPATRA AWAITING CAESAR
BARBARIANS ATTACKING GAUL
WITH TIME AND PATIENCE, A HAND-PRINTING PRESS COULD BE BUILT. EVEN THE CRUDEST OF SUCH MACHINES WOULD FAR OUT RIVAL THE SPEED OF LABORIOUS HAND-WRITING. WITH THIS, YOU COULD INTRODUCE ARABIC NUMERALS AND THE DECIMAL SYSTEM. YOU WOULD PUSH THE ROMANS TEN CENTURIES AHEAD!
AN EFFECTIVE STEAM-TURBINE COULD BE MADE, DEVELOPING MORE POWER THAN THE ROMANS DREAMED POSSIBLE. IT WOULD START THE ROMAN AGE OFF IN AN INDUSTRIAL ERA SEVENTEEN CENTURIES AHEAD OF ITS TIME. RECORDED HISTORY WOULD BE CHANGED!

Fig. 4.3 (Fig. 4.2 continued). Illustration for "IF—You Were Stranded in Time!" ©1938 by Better Publications, Inc.; Reprinted by permission of the Ackerman Science Fiction Agency, 2495 Glendower Ave., Hollywood, CA 90027 for the Estate

Even earlier than Miller's tale is an equally famous one[47] that illustrates the same point in a graphic way. In A.D. 452 a time traveler shoots and kills one of Attila's Huns (who would have been his great-grandfather many times over); the result is that 50,000 of the Hun's descendants vanish! So dramatic did the readers of this story find the concept that author repeated the idea the very next year.[48] Twenty years later another writer topped these tales by having a time traveler accidently kill the original 'intelligent baboon' in the ancient past, thereby wiping out the entire human species![49]

Science fiction writers have been as puzzled by the grandfather paradox[50] as have been nearly everybody else. As the inventor of the first time machine says in one tale,[51] "I have devised a method [for travel] into the distant past. The paradox is immediately pointed out—suppose [the time traveler] should kill an ancestor or otherwise change history? I do not claim to be able to explain how this apparent paradox is overcome in time travel; all I know is that time travel *is* possible. Undoubtedly, better minds than mine will one day resolve that paradox, but until then we shall continue to utilize time travel, paradox or not."

Some may feel it overly dramatic that the classic time travel paradox has such a murderous form, but that is its historical origin in science fiction (not in either physics or philosophy). We can find the grandfather paradox discussed as if already well-known in a letter to the editor at *Astounding Stories* (January 1933). The author of that letter wrote "Why pick on grandfather? It seems that the only way to prove that time travel is impossible is to cite a case of killing one's own grandfather. This incessant murdering of harmless ancestors must stop. Let's see some wide-awake fan make up some other method of disproving the theory." As we proceed, you'll see just how clever some of those who responded to that writer's plea have been but, even today, as the grandfather paradox stands revealed as a red herring, it is preeminent in most people's imaginations. If a solution to the grandfather paradox puzzle escaped an early science fiction writer, then he would generally just mysteriously mention it and then hasten on to other matters. For example, in one story[52] the following exchange between the stock pulp-fiction characters of a young hero and a brilliant old scientist occurs:

---

[47] N. Schachner, "Ancestral Voices," *Astounding Stories*, December 1933.

[48] N. Schachner, "The Time Imposter," *Astounding Stories*, March 1934.

[49] C. Dye, "Time Goes to Now," *Science Fiction Quarterly*, May 1953.

[50] The 'paradox' is that, assuming you do arrive in the past with a working gun, why *can't* you kill your grandfather? After all, you *must* fail in that quest because otherwise you wouldn't be there from the future to even try. But *why* must you fail? It is, of course, not actually necessary to try to kill your grandfather to run into this paradoxical situation—just go back in time to any moment in the past and try to kill yourself! You won't succeed (if the past is unchangeable), but *why not*? (To argue 'because the past is unchangeable' is to beg the question. We need more insight than that.)

[51] M. Reynolds and F. Brown, "Dark Interlude," *Galaxy Science Fiction*, January 1951.

[52] C. South, "The Time Mirror," *Amazing Stories*, December 1942.

> "You mean that time travel really is possible? That men can be transported into the future or the past—?"
>
> The other held up a restraining hand. "Yes. Time travel *is* possible . . ."
>
> "But professor! Think of what you're saying! You're telling me that I could go back and murder my own grandfather. That I could prevent myself from being born—?"
>
> Again the elder man sighed. "I was afraid of this," he said. "I knew you could not understand." He hesitated. Then: "At any rate, take my word for it that time travel is possible. Also, I assure you that there are any number of perfectly sound theoretical and practical reasons why you never could hope to murder your grandparents."

We are, however, not told just what those reasons might be.

Even when all has been said about the impossibility of changing the past, and even when they are finally willing to concede that point, most people still cannot help wondering *why* the time traveler can't kill his grandfather. There the time traveler is, after all, just two feet away from the nasty young codger (I assume he is nasty to make the whole unpleasant business of murder as palatable as possible), with a perfectly functioning and well-oiled revolver in his hand, cocked and loaded with powerful, factory-fresh ammunition that even Dirty Harry would find excessive. What can possibly prevent the time traveler from simply raising his arm and doing the deed? Indeed, the artwork (reproduced at the end of the Introduction) accompanying one 1944 story shows this act in detail, including the smoking gun in the hand of the time traveler who has just taken a shot at grandpop. And if that still leaves open the remote possibility of an aiming error through nervousness, then why can't a suicidal time traveler just wrap his body in factory-fresh dynamite and blow-up granddad—along with himself and everything else within a hundred feet?

I'll argue in this book that killing your grandfather in the past, before he sets *you* in motion, is *logically* impossible. The laws of physics will then faithfully do their duty. No one will ever find an unfinished note in the empty laboratory of a missing traveler who, skeptical of the grandfather paradox, has written "To prove the falsity of the grandfather paradox, I will take my time machine back 50 years and kill my grandf. . .." Nor will any time traveler have to be concerned about the twist in one tale, which opens with the inventor of a time machine is showing the gadget to three friends. One of them later steals the machine to go back 60 years to kill his grandfather—and the story closes with a *near* repeat of the opening, with the inventor showing the gadget to *two* friends.[53]

Invoking logic in this way, in the context of time travel to the past, was discussed in the philosophical literature nearly half a century ago: "If we assume that it is impossible for [a time traveler to kill his younger self], some people are inclined to ask such questions as this: 'But how can the laws of logic prevent him from killing his younger self? Do they cause his finger to slip on the trigger or the bullet to fly apart in mid-air?' The implication of such questions is that the laws of logic cannot prevent such actions. But such questions are like asking: 'How do the laws of logic prevent the geometer from trisecting the angle or squaring the circle? Do they, for example, cause his ruler to slip as a crucial moment every time he tries it?'"[54].

---

[53] F. Brown, "First Time Machine," *Honeymoon in Hell*, Bantam 1958.

A similar point was made later by another philosopher: "Surely it is not an impairment of 'freedom of action' . . . that, e.g., you cannot push another person harder than he/she pushes you. Just as one would explain this is the case by reference to Newton's third law ('to every action there is an equal and opposite reaction'), one could explain the impossibility of [causing a paradox] by reference to the laws which imply such a impossibility. If this explanation is taken to be unsatisfactory, it would seem that one is saddled with a general problem concerning the reconciliation of physics and 'freedom,' and not with a specific argument against [paradoxes]."[55]

The grandfather paradox unquestionably nags at all students of time travel. As a character in one story declares, "The resolution of [the grandfather paradox] is the key to time,"[56] and some incorrectly believe it remains unresolved.[57] The paradox *is* undeniably troublesome: as one philosopher put it, the apparent possibility of a time traveler being able to do away with both his grandfather and himself gives "rise to such puzzles that we are forced to question its [time travel's] intelligibility."[58] In the next section we'll explore how to answer this concern.

4.3 Changing Versus Affecting the Past

> "The past—it's pretty damn solid, Phil. It's a little like a compost pile—fairly soft near the surface but packed hard further down, with all that Time piled on top of it."[59]
>
> —one 'explanation,' perhaps, for the unchangeability of the past

The common belief today, among physicists and philosophers alike, is that given any consistent description of reality it is simply impossible for a time traveler to kill himself as a baby. As one philosopher put it, "Autoinfanticide is *metaphysically impossible* [my emphasis]. This metaphysical impossibility is philosophically intriguing because unlike most impossible events, we can vividly picture how it might look. Time travel itself seems possible, and for those who arrive in the past with proper equipment and training, the actual infanticide should not be difficult."[60] And so the grandfather paradox lives on, bedeviling both physicists and philosophers alike.

---

[54] J. W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies*, November 1974, pp. 153–173. Science fiction had already considered time travel suicide in, for example, K. Neville, "Mission," *Fantasy and Science Fiction*, April 1953.

[55] F. Arntzenius, "Causal Paradoxes in Special Relativity," *British Journal for the Philosophy of Science*, June 1990, pp. 223–243.

[56] P. Worth, "Typewriter from the Future," *Amazing Stories*," February 1950. See also note 106 in Chap. 1.

[57] For example, in J. H. Schmidt, "Newcomb's Paradox Realized with Backward Causation," *British Journal for the Philosophy of Science*, March 1998, pp. 67–87, we read that "there are as yet no generally accepted solutions" to the grandfather paradox.

[58] S. Gorovitz, "Leaving the Past Alone," *Philosophical Review*, July 1964, pp. 360–371.

[59] F. M. Busby, "A Gun for Grandfather," *Future Science Fiction*, Fall 1957.

Indeed, one physicist described the time travel paradoxes as "the most controversial issue related to time machines."[61] As he argued, "These paradoxes seem to be something inherent to time machines (their main attribute, perhaps), so it is reasonable to assume that if there exists a universal law prohibiting the time machines, it must have something to do with the paradoxes. And on the other hand, be the problem of the paradoxes satisfactorily solved there probably would be no need to look for such a law, the (supposed) paradoxicalness of the time machines being traditionally the main objection against them."

One of the persistent stumbling blocks to removing the confusion of the paradoxes is a failure to distinguish between *affecting* and *changing* the past. One philosopher wrote this, in a somewhat bungled attempt to explain what is meant by affecting the past: "Nothing anyone can do now can make it not have rained yesterday"[62] if, in fact, it *did* rain yesterday. This is correct, but it is *not* what is meant by *affecting* the past. Rather, if the reason it *did* rain yesterday is because a time traveler from the future seeded the clouds, then that time traveler affected the past. Making it not to have rained yesterday would be to *change* the past.

A classic[63] by Isaac Asimov illustrates what is meant by a time traveler affecting the past. An idealistic physics professor, convinced that the world's political problems are the result of the comparative newness of scientific thought and tradition, tries to change the past (and thus the present) by sending a Greek translation of a modern chemistry text back 2000 years to the Hellenic days of Leucippus, Lucretius, and Democritus. He dies in the attempt but succeeds in the transmission. When a government investigator—called in because the professor drained an entire nuclear power reactor to energize his time machine!—discovers that the transmission takes a day to travel back a hundred years (a little gimmick with no foundation in physics, but simply something Asimov needed for the story), he fears 'our' world will vanish in 20 days, to be replaced by a 'new' one. In the end, however, he learns you can't change the past. As one of the late professor's colleagues tells the investigator, "While you are right that any change in the course of past events, however trifling, would have incalculable consequences . . . I must point out that are nevertheless wrong in your final conclusion. *Because THIS is the world in which the Greek chemistry text WAS sent back.*"

---

[60] David Horacek, "Time Travel in Indeterministic Worlds," *The Monist*, July 2005, pp. 423–436.

[61] S. Krasnikov, "Time Travel Paradox," *Physical Review D*, February 14, 2002, pp. 064013–1 to -8.

[62] R. G. Swinburne, "Affecting the Past," *Philosophical Quarterly*, October 1966, pp. 341–347.

[63] "The Red Queen's Race," *Astounding Science Fiction*, January 1949.

Another good illustration from science fiction of affecting the past can be found years before Asimov's. In that story,[64] a time traveler leaves the Chicago of 1942 for the year 3000. Much later, in the year 2564, another time traveler interested in history journeys back to 2253 in an attempt to learn the cause of the great Chicago explosion of that year. The explosion was centered on the site of an ancient laboratory once used by a scientist who mysteriously vanished in 1942. The second time traveler begins his trip into the past on the same spot, with plans to go back to the day before the explosion. At the story's end, we learn that the disaster was the result of the two time travelers colliding as both 'passed through' 2253. The backward traveling historian, therefore, by pushing a button in 2564, is the cause of an event that happened 311 years earlier.[65]

To give a science fiction example of changing the past, it would be hard to do better than with a story that appeared a few years after Asimov's The central character is a researcher in time travel who has concluded that what is wrong with the world can be traced back to the scientific method getting off to a late start (this story was almost surely written as a result of Asimov's tale!): the time traveler thinks he can correct matters by visiting 340 B.C. and educating Aristotle on the proper scientific attitude. (Aristotle believed that observing the world was inferior to pure thinking about how the things, in his opinion, *ought* to work.) This the time traveler does, with utterly disastrous consequences. He returns to the present to find a scientifically retarded world that makes him a slave. In his cell he writes on a wall the bitter lesson he has learned too late: "Leave Well Enough Alone."[66]

It is the fear of time travelers from the future attempting to alter the past that has led some philosophers (and not just a few physicists) to assert that time travel is impossible, because it would mean what they feel to be impossible might happen: changing the past. One philosopher, however, argued long ago that such a worry is unwarranted. As he wrote at the end of an essay (a polemic *against* the concept of four-dimensional spacetime, and so *against* the idea of time travel), "Squandering vast sums on foolish enterprises is an everyday occurrence. [For example], will the U.S. time explorer get back and eliminate Lenin before his Russian rival gets back even earlier and eliminates George Washington? . . . If such spectacular folly once gets under way because governments have been convinced of some nonsensical theory, a logician will not . . . lose any sleep about who is going to succeed."[67]

---

[64] O. Saari, "The Time Bender," *Astounding Stories*, August 1937.

[65] This story describes something a bit more than 'simply' affecting the past; it has a *causal loop* in it. The time traveling historian makes his trip *because* of an event in the past that his trip causes. Such paradoxes will be the subject of the next section.

[66] L. Sprague de Camp, "Aristotle and the Gun," *Astounding Science Fiction*, February 1958. Asimov and de Camp were close friends, and their two stories with similar premises are clearly the result of a bit of friendly rivalry.

[67] P. Geach, "Some Problems About Time," in *Studies in the Philosophy of Thought and Action* (P. F. Strawson, editor), Oxford 1968.

And long before that essay (with its correct conclusion reached through faulty reasoning) was penned, we learn the same lesson (in a 1923 tale by the English novelist May Sinclair (1863–1946)) as we follow a woman right into hell after her death; she ends up there because of an immoral life. She then wanders through time, into her past, but finds that she can change nothing. As she is told, "You think the past affects the future. Has it never struck you that the future may affect the past? . . . You *were* what you *were to be*."[68] This last line, from a *non*-science fiction story, is consistent with the modern view held by physicists of time travel. You cannot travel anywhere into the past unless you've already been there, and when you do make the trip you will do what you've already done there. You could not, as does the time traveler in one tale,[69] change the course of history by revealing twentieth-century physics in the eighteenth century. That does not mean you would necessarily be ineffectual during your stay in the past, however (certainly it doesn't mean, as Hugo Gernsback thought, that you'd be *invisible*!) Not being able to change the past is not equivalent to being unable to *influence* or *affect* what happened in the past, and science fiction writers have used this distinction to good effect, as did Asimov (note 63) and de Camp (note 66).

Robert Heinlein was a science fiction writer who clearly understood time travel paradoxes, both what they mean and, at least as important, what they do not mean. In his 1964 cold-war novel *Farnham's Freehold*, for example—the story of a family that is literally blasted twenty-one centuries into the future when their bomb shelter receives a direct hit from a Soviet nuclear warhead—we find following exchange as two of the characters are about to return to their original time via time machine:

> "The way I see it, there are no paradoxes in time travel, there can't be. If we are going to make this jump, then we already did; that's what happened. And if it doesn't work, then it's because it didn't happen."
>
> "But it hasn't happened yet. Therefore, you are saying it didn't happen, so it can't happen. That's what I said."
>
> "No, no! We don't know whether it has already happened or not. If it did, it will. If it didn't, it won't."

Modern philosophers, and many physicists, too, as well, who have examined the concept of time travel in depth, agree with Heinlein's character and, indeed, it is now common practice to invoke the so-called *principle of self-consistency*—generally attributed to the Russian astrophysicist Igor Novikov (see note 117 in Chap. 1) because he and his colleagues did not simply invoke it, but rather were able to *derive* it from the *principle of least action*, a concept held by many to be at the highest level of importance in physics[70]: all that is required, argued Novikov, in *any* physical process (including time travel), is that a *logical* consistency exist between events.[71] In his book *Evolution of the Universe* (originally published in Russian in 1979), Novikov wrote "The closure of time curves does not necessarily imply a violation of causality, since the events along such a closed line may be all 'self-adjusted'—they all affect one another through the closed cycle and follow on another in a self-consistent way." He later repeated that view in one of the first time machine papers in the physics literature.[72]

---

[68] M. Sinclair, "Where Their Fire Is Not Quenched," in *After the Darkness Falls* (B. Karloff, editor), World Publishing 1946.

[69] D. Beason, "Ben Franklin's Laser," *Analog*, December 1990.

In fact, despite the attachment of Novikov's name to the principle of self-consistency, it was actually around in physics *decades* earlier; it has been traced back to as far as 1903![73] And at least an intuitive understanding of the principle can be found in the mainstream literature from nearly as long ago. For example, in Lord Dunsany's short 1928 play *The Jest of Hahalaba* (the inspiration for the 1944 film *It Happened Tomorrow*), a man obtains (via supernatural means) a copy of tomorrow's newspaper. In it he reads his own obituary, which so shocks him that he promptly expires—thus explaining the obituary notice.

The principle of self-consistency has been in science fiction long before Novikov, too. An example is the 1941 story "Time Wants a Skeleton" (see note 15). In it one character, after puzzling over a time travel paradox, realizes that "Future and present demanded co-operation, if there was to be a logical future!" And a nice lecture on the principle (that pre-dates by 3 years the dialogue quoted earlier from Heinlein's *Farnham's Freehold*) is given by a character that is particularly interesting because it was published, not in a specialty science fiction magazine catering to an audience with 'genre knowledge' of time travel, but rather in an icon of general American culture.[74]

Not all science fiction writers, however, have understood the requirement for consistency around a loop in time. In one tale, for example, a man meets the inventor of a time machine and agrees to his request to use it to travel into the future. Once he is in the future, alas, the machine breaks. The man then finds another machine that, though it is too small for him to fit in it, is able to hold a recording that he sends back into the past to himself, to a time *before* he started his forward journey. The message on the recording (which he did not receive the 'first' time) is, of course, *not* to make a deal with the inventor. This advice he follows, and so the principle of self-consistency is violated twice in this story.[75]

---

[70] Like just about everything concerning time travel, however, not *all* think this. For example, the great German physicist Max Planck (1858–1947), the 1918 Nobel physics laureate, said (in 1922): "Physics hence is inclined to view the principle of least action more as a formal and accidental curiosity than as a pillar of physical knowledge." Still, he *did* also declare that he thought it unlikely "the dominance of such a simple law could be a mere accident." Quoted from Marc Lange, "Conservation Laws in Scientific Explanations: Constraints or Coincidences," *Philosophy of Science*, July 2011, pp. 333–352.

[71] See A. Carlini, *et al.,* "Time Machines: The Principle of Self-Consistency as a Consequence of the Principle of Minimal Action," October 1995, pp. 557–580, and "Time Machines and the Principle of Self-Consistency as a Consequence of the Principle of Stationary Action (II): The Cauchy Problem for a Self-Interacting Relativistic Particle," October 1996, pp. 445–479, both in *International Journal of Modern Physics D*.

[72] I. D. Novikov, "An Analysis of the Operation of a Time Machine," *Soviet Physics JETP*, March 1989, pp. 439–443.

[73] R. D. Driver, "Can the Future Influence the Present?" *Physical Review D*, February 15, 1979, pp. 1098–1107.

[74] R. F. Young, "The Dandelion Girl," *The Saturday Evening Post*, April 1, 1961. See also note 49 in Chap. 2.

An ability to play a role in history is not without some constraints. You can't save Jesus with a rifle (see note 45), or Joan of Arc with a fire extinguisher, or knock-out John Wilkes Booth with a baseball bat outside of Ford's Theatre, or blow-up Hitler with a bomb, and you can't prevent either the Black Death in the London of 1665 or the Great Fire the following year. But it *is* logically possible for a time traveler who has an infected rat sneak into his time machine, or who carelessly discards a match, to be the *cause* of the last two examples. That was the fate, for example, of the time traveling historian from A.D. 2461 who was the cause of the plague in A.D. 562 Rome, as well as of that in England nearly 800 years later.[76]

Michael Moorcock's 1969 novel *Behold the Man* gets the impossibility of changing the past, and the possibility of affecting it, right. When a disturbed man journeys backward in time to ancient Galilee to meet Christ, only to discover that there is no such person, *he* assumes the role and lives out the Biblical accounts up to and including dying on the Cross. He has not changed the past, but he certainly plays an important role in it!

An early science fiction story that got this right, long before the philosophers and physicists thought of it, was the clever tale whose artwork I have reproduced at the end of the Introduction.[77] In that story, a time traveler journeys back from 1943 to 1870 and shoots his then 14-year old grandfather in the head. Leaving his victim lying on the ground with "blood oozing all over the youth's forehead," the would-be killer returns to 1943. Once back, however, he finds himself in a strange place where he learns from two men that the Germans destroyed New York in 1920 with poison gas! Suddenly realizing that the death of his grandfather has apparently changed history (a curious oversight for anyone smart enough to invent a time machine and then to use it to force the 'grandfather paradox'), he decides he'd rather be dead than be cut off for all time from the world he remembers. So, he shoots himself dead. Then we learn that the two men he encountered are actually inmates in an asylum who like to make-up stories for unsuspecting strangers. We also learn that the time traveler's grandfather's photographs always did show him with a "white, furrowed scar on his forehead that might have been caused by a glancing bullet."

---

[75] R. Wilson, "The Message," *Astounding Stories*, March 1942.

[76] G. C. Edmondson, "The Misfit," *Fantasy and Science Fiction*, February 1959.

[77] M. Weisinger, "Thompson's Time Traveling Theory,' *Amazing Stories*, March 1944.

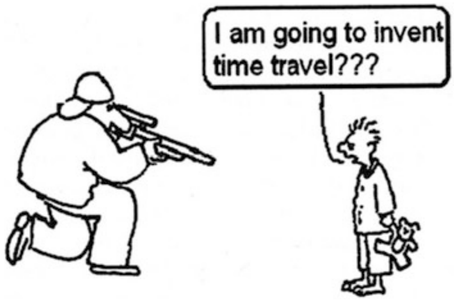
I am going to invent time travel???

Fig. 4.4 The inventor of a time machine about to commit autoinfanticide in the past (the youngster holding the teddy bear is a younger version of the time traveler). Illustration reproduced by the kind permission of Frank Arntzenius (Professor of Philosophy at Oxford University), from his paper "Time Travel: Double Your Fun," *Philosophy Compass*, November 2006, pp. 599–616

Well, okay, you might say at this point, 'I'm convinced you can't change the past, but let's get back to the autoinfanticide (grandfather) paradox. So *why can't* a time traveler kill his baby-self in the past?' A possible answer, one now generally accepted by philosophers and physicists alike, appeared first in science fiction. In a tale[78] that appeared just the year after Gödel's 1949 discovery of time travel in general relativity, we find a character saying "The answer is quite simple. When the man goes back in time and kills his grandfather, and returns to his own time again, he finds to his surprise that he made a mistake. It was not his grandfather at all! And no matter how many times he goes back and kills his grandfather … he *always* [my emphasis] finds he made a mistake." Or, perhaps, some noise distracts him as his finger tightens on the trigger, or the grenade he tosses at granddad is a dud, or a gust of wind deflects the arrow, or (most ludicrous of all) he simply slips on a discarded banana peel!

Okay, that works for *that* time traveler. But suppose, someone objects, that we arrange to have a *lot* of time travelers go back in time, each with murder in his heart for his grandfather. Then, as one philosopher has observed, "Since [killing one's grandfather in the distant past] is impossible, each assassin fails. Some change their minds, others slip on banana peels, yet others kill the wrong target, and so on. But there is something odd about the idea that such coincidences are guaranteed to happen, again and again!"[79]

Early science fiction avoided invoking banana peels by providing an even more extreme 'explanation' for the failures: the time police, who are charged with foiling would-be grandfather killers. (See, for example, the many stories by Poul Anderson (1926–2001) of the 'Time Patrol.') These time commandos are imagined to roam the corridors of time, disrupting the attempts of all those who would change recorded history. Stories of these temporal cops are simply westerns, mysteries, police procedurals, or some other similar type of specialty genre story wearing thin camouflage. This story device, whose main purpose is to allow both time travel and free will,[80] has been correctly called "boring" by at least one philosopher (see note 5 in the Introduction), an evaluation shared by modern philosophers, physicists, and (I think) even most modern science fiction writers.

---

[78] "Typewriter from the Future": see note 106 in Chap. 1.

[79] Theodore Sider, "Time Travel, Coincidences, and Counterfactuals," *Philosophical Studies*, August 2002, pp. 115–138.

So, we seem to be back to banana peels to save grandfather—but it *is* difficult to deny that vast hordes of murderous grandsons *do* appear to require an unlimited number of strategically placed banana peels, strewn all about the past, to trip-up every one of those potential assassins. This problem, of repeated, improbable coincidences to thwart murderous descendants from the future, was first commented on by the philosopher Paul Horwich in 1975 (see note 19 in the Introduction), and then given a convincing resolution by another philosopher in 1997.[81]

To explain the argument, I'll first use the philosopher's less deadly example of dated objects. "Suppose," he writes, "that every object has written upon it the date on which it will cease to exist . . . perhaps a time traveler travelled into the future, observed the demise of objects and then travelled back [to just after he left for the future] and wrote the dates."[82] If now the time traveler tries to destroy an object *before the date written on it*, then he *will* fail. As the philosopher amusingly described his attempts to destroy a pen 'before its time,' "I take it outside to place under the wheels of a passing train, but there is a train strike that day. The telephone rings just as I am about to drop the pen into a vat of acid. I slip on a banana peel on my way to put the pen in the microwave. My dog eats my designs for a pen grinder. And so on, for as long as you please. However many attempts I make, the attempts in no way require the occurrence of the coincidences that foil them." To put it bluntly, 'Stuff happens.' The pen has the observed date of its destruction on it, and that date is still in the future and so it is simply impossible to destroy it now.

Now, here's the point: the date on the pen is there *because* all those attempts to destroy it before that written date fail. But the presence of the date is *not* the reason for any of the weird (?) occurrences that disrupted all the attempts to destroy the pen, but rather it's *because* all those attempts failed that the date is what it is. This same argument applies to the grandfather paradox. The only time travelers available, today, to go back into the past to *try* to kill their grandfathers, are precisely those time travelers whose grandfathers were *not* killed. Or, to paraphrase our philosopher (note 81), to ask 'why do coincidences always foil the time traveller's attempts to kill [grandfather in the distant past], is to get things back to front. It is only because the murder attempts fail that the time traveler is alive in the future to even make the *attempt*.'

---

[80] See, for example, David King, "Time Travel and Self-Consistency: Implications for Determinism and the Human Condition," *Ratio*, September 1999, pp. 271–278.

[81] Nicholas J. J. Smith, "Bananas Enough for Time Travel?" *British Journal for the Philosophy of Science*, September 1997, pp. 363–389.

[82] This does present us with the curious (although non-paradoxical) situation that the time traveler will find, upon his appearing in the future, the date he *will write* (in his personal future) when he returns to just after he left on his time trip.

In other words, not only is the grandfather 'paradox' not a paradox, it isn't even surprising!

4.4 Causal Loop and Bootstrap Paradoxes

> "My dear Collingwood, don't drive yourself crazy trying to resolve the paradoxes of time travel. The [time machines] are gone . . . have a drink."[83]

The grandfather paradox might finally have been put to rest, but there are still plenty of other logical minefields left to be negotiated. One of the more puzzling is that of the closed loop in time, a conundrum nicely illustrated by one philosopher[84] as follows, as an explanation of the journey one time traveler makes to 3000 B.C.: "In our time travel story it just may be that the traveler's interest in going back to ancient Egypt is stimulated by recently discovered documents, found near Cairo, containing the diary of a person claiming to be a time traveler, whereupon our hero, realizing it is himself, immediately begins . . . construction of a rocket in order to 'fulfill his destiny.'" In other words, (1) he builds a time machine and goes back to the past because of the discovered diary, and (2) the diary is discovered because he goes back to the past. Each of these points by itself has logical clarity, but together they form a closed time loop (a *causal loop*) of enormous mystery.

Science fiction was strewn with causal loops long before the philosophers and physicists began to ponder them, however, with (for example) one early tale on a time traveler who journeys a century into the past because she finds an old, yellowed newspaper story describing her arrival.[85] But this tale wasn't the first to use a causal loop, as we can find one of the first sophisticated treatments of this device in a story that appeared even earlier (in the same publication).[86] A time traveler in 1930, about to start his journey into the future in an airplane/time machine, wonders at the last moment if he should really go—then he sees himself returning and thus *knows* he will successfully make the trip.

As he later tells a friend, "That decided me . . . Paradoxical? I should say so! I had seen myself return from my time-trip *before* I started it [just like Marty McFly in the original *Back to the Future* film]; had I *not* seen that return, I would *not* have commenced that strange journey, and so could *not* have returned in order to induce me to decide that I *would* make the journey!" And later, when he finds himself in a dangerous situation in the future, he draws hope from that initial experience: "I *would* escape ... It was so decreed. Had I not, with my own eyes, seen myself appear out of the fourth dimension back there in the Twentieth Century, and glide down to my landing-field? Surely, then, I *was* destined to return to my own age safe and sound."

---

[83] A science fiction suggestion that in certain situations (particularly causal loops), might actually be good advice! From L. Sprague de Camp's "The Best-Laid Scheme," *Astounding Science Fiction*, February 1941.

[84] L. Dwyer, "Time Travel and Some Alleged Logical Asymmetries Between Past and Future," *Canadian Journal of Philosophy*, March 1978, pp. 15–38.

[85] P. Bolton, "The Time Hoaxes," *Amazing Stories*, August 1931.

[86] F. J. Bridge, "Via the Time Accelerator," *Amazing Stories*, January 1931.

Even more dramatic is the second, internal time loop that ends the story. When the time traveler arrives in a ruined city in the year A.D. 1,001,930 he is greeted, *by name*, by an old man who says he (the old man) is the Last Man alive. He knew the time traveler was coming because an ancient history book had said the Last Man had, in fact, appeared in the year A.D. 502,101 in the very time machine out of which the time traveler has just stepped. The time traveler is so startled by all this (and who could blame him!?) that he decides to mull over what he has been told until the next day. As he wakes up in the morning, he is just in time to watch the Last Man depart for 502,101. Stranded in the future, the time traveler wanders the empty city in despair until he chances upon a museum. And in the museum, sealed in a glass case, is his time machine (!)—it has been there for half a million years, since the end of the Last Man's journey. And so the time traveler is saved; he merely adds some oil to the still-functional engine (if you can accept time travel, I suppose this is no more difficult to believe) and returns to 1930—just as he saw himself do at the beginning of the story.

Since Bridge's astonishing story, the idea of a causal loop in time has been used many times in science fiction. Here's a representative sampling:

(1) Time travelers arrive in the forty-sixth century, only to find that they are expected. Their host tells them why: "I have been awaiting your arrival from the past. I have a written record of your coming. You see, I have a time machine myself ... With my time machine, I recently went a year into the future and read the written account I had made, or will make after you leave. Then I came back, awaiting your arrival."[87]

(2) Armed travelers return to the Triassic age to uncover the secrets of a mysterious artifact that has been recently discovered; at the end we learn it is the remains of their own automatic rifle[88];

(3) A time traveler journeys back 500 years, where he suffers an accident that results in his being "agelessly stuck" in his time-traveling gadget until he is freed—by himself, 500 years later. He then gets into the gadget to journey back 500 years[89];

(4) The world's time suddenly loses 5 min, an astonishing event that comes to be called "the time drop." After 2 weeks of investigation, a reporter traces this

---

[87] E. Binder, "The Time Cheaters," *Thrilling Wonder Stories*, March 1940.

[88] J. Blish, "Weapon Out of Time," *Science Fiction Quarterly*, Spring 1941.

[89] A. B. Chandler, "The Tides of Time," *Fantastic Adventures*, June 1948.

event to a reclusive (but brilliant, of course) scientist who reveals that he has invented a time machine. The reporter decides to test this claim by using the machine to return to just before the start of the time drop, to observe precisely what caused it—it is, in fact, a malfunction of the machine that is at fault and the reporter finds himself caught in a 2-week long causal loop[90];

(5) A physicist knows something odd lies in his future when he is confronted with a 700 year-old museum copy of a book. The puzzle is how to explain a message penned in ancient, faded ink, in modern English *and in his handwriting*, on the back side of one of the recently unglued endpapers! How, too, to explain his own fingerprints all over the same endpaper? How, indeed, to answer these questions is his problem when he is presented with all of this and is asked, “Have you, by any chance, been visiting the thirteenth century?” At the end of the story the time loop is closed when the physicist finds himself writing that same message on a *brand new* copy of the book that has been sent from the past (and that he returns to the thirteenth century via a “time portal”)[91];

(6) A time traveler from 1964 is secretly observed by one of the ‘locals’ when he arrives in 1683. The oddness of the sudden appearance of the time traveler and his machine (“It were a kind of Dazzle”) makes the local think it might be that the stranger is the man who stole some items from his home the previous night, the same night he had an “ill Dream.” Stealing the time machine after the time traveler has gone exploring, the local travels to 1964 where he learns how valuable antiques are. So back he goes to 1683, to the night before the time machine first appeared, to get some ‘antiques’ from his house. And thus he realizes who the thief *really* was. Before leaving again for the future, he enters his own bedroom to see himself asleep and then to awaken. And so he also learns the cause of what he called his “ill Dream.”[92];

(7) A movie production crew goes into the past to make a film. At the end of the story it becomes clear that their presence in the past was not an insignificant event, as one character realizes after seeing the evidence of how they affected (*not* changed!) the past: “If this is true, then the only reason that the Vikings settled in Vinland is because we decided to make a motion picture showing how the Vikings settled in Vinland”[93];

(8) A private college, endowed decades before by a generous but mysterious benefactor, experiments with a time machine. Suddenly, one of the college’s graduates is accidently sent a hundred years into the past—where she becomes the benefactor. The college comes into existence, therefore, because it will exist[94];

---

[90] W. Sheldon, “A Bit of Forever,” *Super Science Stories*, July 1950.

[91] M. Leinster, “The Gadget Had a Ghost,” *Thrilling Wonder Stories*, June 1952.

[92] D. I. Massor, “A Two-Timer,” *New Worlds SF*, February 1966.

[93] H. Harrison, *The Technicolor Time Machine*, Doubleday 1967.

[94] C. Simak, “The Birch Clump Cylinder,” *Stellar 1* (J. del Rey, editor), Ballantine 1974.

(9) A man gets the money to support his experiments in time travel by selling a large collection of old, rare comic books he has discovered in his late mother's attic. Later we learn how the comic books came to be there; after his experiments are successful, the inventor travels back into the distant past, buys the *newly* published comic books right off newsstands, and stores them in his mother's attic where, decades later, he knows his younger self will grow up and then find them (and thus get the money to make it all happen)[95];

(10) In the 1980 film *Somewhere in Time* (based on the 1975 novel *Bid Time Return* by Richard Matheson), a man in the present is visited by a mysterious old woman who gives him a watch. Later, he travels back to 1912 where he meets a girl to whom he gives the watch. He then returns to the present, and she lives out her life from 1912 on, until she too reaches the present, where we discover she is the (now old) woman who gives the man the watch.

Once philosophers discovered the bizarre nature of causal loops, they quickly proved themselves to be the equal of science fiction writers in imagining strange doings. Here's one example of that, one which any writer would be proud of: "If James cannot decide whether to marry Alice or Jane, he simply travels to the future and learns that he is to choose Alice; he then chooses her for this reason. One wants to object that the decision to marry Alice was never really made at all! But this is not true; the decision was made—as a result of the knowledge that this was the decision . . . It is not the case that the prospective bridegroom could visit the future and compare the results of marrying Alice with those of marrying Jane in order to decide between the alternatives. For if he visits the future, he will learn only that in fact he chose Alice, for better or for worse!"[96]

This same philosopher elaborated on his view of causal loops in a later paper, where he wrote "What if time travel becomes commonplace, so that we must deal with a constant stress of time travelers returning from the future to reveal what they have seen?"[97] His answer is "I think it is clear that the . . . causal loop we have been discussing would become very common, and would play a prominent role in human affairs." He denied, however, that such causal loops would mean the loss of free will. As he explained his position, knowledge of a rigged roulette will not prevent you from putting your money on the table *if you want to*, but perhaps that knowledge will influence your *freely* made decision making. Whether you learn that the roulette wheel is rigged by traditional means (perhaps you see magnets being installed under the table) or by means of time travel is irrelevant—even with this knowledge, you act freely. Other philosophers have not been so generous. One disliked causal loops so much, for example, that while he believed them to be conceptually possible, he also thought them to have "a queer smell,"[98] so much so that he simply preferred to avoid thinking about them!

---

[95] D. Knight, "The Man Who Went Back," *Amazing Stories*, November 1985. This same idea was used earlier in the story "Compounded Interest," (*Magazine of Fantasy and Science Fiction*, August 1956) by Mack Reynolds, in which the inventor of a time machine has the money to build his gadget because he uses it to go back into the past where he deposits a small sum, which then grows (through the 'magic' of compound interest) into the cash he needs to fund his time machine.

[96] G. Fulmer, "Understanding Time Travel," *Southwestern Journal of Philosophy*, Spring 1980, pp. 151–156.

[97] G. Fulmer, "Time Travel, Determinism, and Fatalism," *Philosophical Speculations in Science Fiction and Fantasy*, Spring 1981, pp. 41–48.

One concern that many philosophers and physicists have had with closed loops in time is that they fear that would mean being trapped on an endless cycle of repeating events. For example, one philosopher long ago wrote

> "There is nothing contradictory in imagining causal chains that are closed, though the existence of such chains would lead to rather unfamiliar experiences. For instance, it might then happen that a person would meet his own former self and have a conversation with him, thus closing a causal line by the use of sound waves. When this occurs the first time he would be the younger ego, and when the same occurrence takes place a second time he would be the older ego. Perhaps the older ego would find it difficult to convince the younger one of their identity; but the older ego would recall an identical experience long ago. And when the younger ego has become old and experiences such an encounter a second time, he is on the other side and tries to convince some 'third' ego of their physical identity. Such a situation appears paradoxical to us; but there is nothing illogical in it."[99]

What has been (erroneously) described with that is the beginning of an endless succession of encounters around a closed causal loop. There is, however, just *one* encounter on such a loop in spacetime (but, of course, the mind of the time traveler experiences the encounter twice), subject to the constraint of self-consistency. Some physicists, too, have been so concerned about multiple trips around closed timelike curves (CTCs), because they think such trips would allow the past to be changed, that they have felt it necessary to specifically forbid such a possibility. As one paper put it, "That the principle of self-consistency is not totally tautological becomes clear when one considers the following alternative: The laws of physics might permit CTCs; and when CTCs occur, they might trigger new kinds of local physics which we have not previously met. For example, a quantum-mechanical system, propagating around CTCs, might return to where it started with values for its wave function that are inconsistent with the initial values; and it might then continue propagating and return once again with a third set of values, then a fourth, then a fifth ... The principle of self-consistency by fiat forbids changing the past."[100]This last statement is, of course, in agreement with the position I have taken in this book, a position that has generally been accepted by most philosophers for several decades now, but the proponents of the principle of self-consistency seem to have been driven to it by a fear of the past 'happening again' over and over, as in the 1993 film *Groundhog Day*.

---

[98] M. MacBeath, "Communication and Time Reversal," *Synthese*, July 1983, pp. 27–46.

[99] H. Reichenbach, *The Direction of Time*, University of California Press 1956, p. 37.

[100] J. L. Friedman et al., "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review D*, September 15, 1990, pp. 1915–1930. Another physicist, however, has flatly rejected this need for the Principle, calling it redundant: see D. Deutsch, "Quantum Mechanics Near Closed Timelike Lines," *Physical Review D*, November 15, 1991, pp. 3197–3217.

Science fiction writers have stumbled into the error of endless cycling on a closed time loop, too. In one such tale,[101] the inventor of the first time machine travels 500 years into the future where he finds a bronze statue of himself that honors his discovery of time travel. Suddenly injured, fatally, he returns to the present with the statue and then dies. As a memorial, the statue is placed in the very spot where the inventor found (will find) it. As the tale ends, the late inventor's lab assistant wonders to himself what will happen 500 years later: "Suddenly a strange machine will come out of the past and [the inventor] will be here again—although he is dead and has been dead 500 years. [He will take the statue] and go back to the past . . . to die. And once again that maddening cycle will begin, to go on and on forever as long as time spins its threads."

That story illustrates yet another puzzle associated with those causal loops that contain a circulating, physical object. That is, *who made the statue*? We can ask the same question about the watch in the time loop of *Somewhere in Time/Bid Time Return* as, at every instant of its existence, the watch is in the possession of either the man or the woman? So, *when was the watch constructed*?

There have been some science fiction writers who specifically recognized this question, long before either the philosophers or the physicists paid attention to it. In one early tale,[102] for example, we read of a time machine that travels from 1935 to 1925. When the question of the origin of the time machine comes up, we read

> "One time machine, found in 1935 and brought back to 1925—found in 1935 *because* brought back to 1925. That is all."
>
> "But who made it in the first place?—Oh, skip the 'in the first place.'[103] Just plain: who made it?"
>
> "No one. It was never made . . . It is here because it is here."

This same puzzle was addressed in *The Technicolor Time Machine* (note 93), when one character is perplexed over a piece of paper in his wallet with a diagram on it, a piece of paper he got from himself (an *older version* of himself, who traveled into the past to give it to his younger self). In frustration, he asks a friend:

> "Then no one ever *drew* this diagram. It just travels around in this wallet and I hand it to myself. Explain that."
>
> His friend replies:
>
> "There is no need to, it explains itself. The piece of paper consists of a self-sufficient loop in time. No one ever drew it. It exists because it is, which is adequate explanation. If you wish to understand it, I will give you an example. You know that all pieces of paper

---

[101] S. Mines, "Find the Sculptor," *Thrilling Wonder Stories*, Spring 1946.

[102] R. M. Farley, "The Man Who Met Himself," *Top-Notch Magazine*, August 1935 (*Top-Notch* was an adventure pulp published between 1910 and 1937).

[103] The reason for this line in the story is that earlier the question of "Where did the time machine come from *originally*?" was raised. The answer: "There was never any 'original.' . . . There is no round-and-round circle of events, no repetition. Merely *one* closed cycle." This is, in fact, the modern view of causal loops, expressed in a 1935 (!) science fiction story.

> have two sides—but if you give one end of a strip of paper a 180-degree twist, then join the ends together, the paper becomes a Möbius strip that has only one side. It exists.[104] Saying it doesn't cannot alter the fact. The same is true of your diagram; it exists."
>
> "But—where did it come from?"
>
> "If you must have a source, you may say that it came from the same place that the missing side of the Möbius strip has gone."[105]

The undeniable mystery of causal loops is the reason behind the philosopher I cited earlier (note 98) who thought they have a "queer smell" and so viewed them with much suspicion. He wasn't alone in that feeling, and another philosopher said as much whe he wrote "despite [strong] arguments for the consistency of time travel stories [with causal loops], the impression is apt to remain that something is wrong with them. I think this impression is correct."[106] One story that this philosopher could well have had in mind is a classic,[107] a tale that describes a knife brought from a museum in the future back to the present. It arrives in the present with a flawless blade, but soon thereafter gets a nick in the blade. How, wonders the narrator, can the time loop be completed "again"? I do not find this quite the puzzle that either the author (and perhaps the philosopher) do: it is simply a variation of the grandfather paradox (which has been shown not to be a paradox at all). *If* the knife is found flawless in the future, *then* it was not (will not) be nicked in the past. As written, the story is not logically consistent as it involves changing the past but, *if* one removed the detail of a *nicked* blade, then we *would* have a true (paradoxical) causal loop, with the question the story, itself, asks about the knife: "How was this knife created . . . when its existence has no beginning or end?"

The nicked knife does illustrate a subtle problem that bedevils *any* causal loop containing a physical object. Consider once again the watch in the film *Somewhere in Time*. Assume the watch received by the man in the present is bright and shiny. He then takes it back into the past and gives it to his love. It remains with her after his return to the present until, decades later, she gives it to him—bright and shiny. Why didn't it tarnish? Is there some peculiar anti-tarnish property to a watch in a causal loop? Well, if so, is that anymore odd than a causal loop itself?[108]

None of that, however, provides a means for rejecting time travel *if* one can argue that it is possible to have time travel *without* causal loops. Indeed, Professor Hanley (see note 105) argues that it is possible, and presents what he claims is an example of how to do it. Alas, another philosopher convincingly showed that the example is flawed and that Hanley's claim that there is no causal loop in his story "is unjustified."[109] From an entertainment point of view, however, eliminating causal loops is going in the wrong direction, as it is the inclusion of causal loops that gives a feeling of mystery to a good science fiction story.

---

[104] See note 99 in Chap. 1, and the related discussion there.

[105] One philosopher calls this bit of dialog "unhelpful," while ignoring the fact that it appeared in a science fiction pulp magazine and not a scholarly journal, and was clearly meant to dazzle teenage boys (see note 39 and related discussion in "Some First Words") with the concept of a causal loop, rather than to break new ground in metaphysical thought. See Richard Hanley, "No End in Sight: Causal Loops in Philosophy, Physics and Fiction," *Synthese*, July 2004, pp. 123–152.

[106] G. Nerlich, "Can Time Be Finite?" *Pacific Philosophical Quarterly*, July 1981, pp. 227–239.

[107] P. S. Miller, "As Never Was," *Astounding Science Fiction*, January 1944. This is the same Miller who appears in note 15 (and see note 46, too).

[108] It is *not* sufficient to say that perhaps she polished the watch. Polishing would remove material from the watch, which means she gives him a watch different from the one he gives her in the past.

To finish this section, then, we can do no better than to discuss causal loops that are even *more* bizarre than are those with a physical object; that is, loops that involve time traveling *information*. (Since information doesn't 'tarnish,' however, such a loop avoids that particular puzzle associated with a physical object in a causal loop.) A classic example of such a loop is a mathematician who is visited in his youth by a time traveler from the future (perhaps himself), who gives him the proof of a theorem for which the mathematician is (will be) famous in the future. *Where*, then, did the proof actually come from? In what *mind* was it *created*?[110]

The philosopher David Lewis wrote with particular insight on causal loops, especially ones that involve information transfer, such as a time traveler going back in time to tell his younger self how to build a time machine so that once its constructed he can go back in time and tell himself how to do it.[111] (This was item (3), you'll recall, in Jim Nicholson's 1931 letter to *Science Wonder Stories* magazine, quoted at the end of the first section of this chapter.) As Professor Lewis wrote (see note 5 in the "Introduction"), "But where did the information come from in the first place? Why did the whole affair happen? *There is simply no answer* [my emphasis]. The parts of the loop are explicable, but the whole of it is not. Strange! But not impossible, and not too different from inexplicabilities we are already inured to. Almost everyone agrees that God, or the Big Bang, or the entire infinite past of the Universe, or the decay of a tritium atom, is uncaused and inexplicable. Then if these are possible, why not the inexplicable causal loops that arise in time travel?"

A few years later, another philosopher[112] gave a similar response to a paradox involving a causal loop similar to Lewis', a loop involving a time machine containing a book with instructions on how to make the time machine. The book travels into the past on the machine so it can be read—in order to make the machine.

---

[109] Bradley Monton, "Time Travel Without Causal Loops," *The Philosophical Quarterly*, January 2009, pp. 54–67.

[110] Professor Hanley (note 105) says the answer to such questions is "straightforward": the information comes "from itself." I think the issue is rather deeper than that.

[111] See D. Franson, "Package Deal," in *Microcosmic Tales*, Taplinger 1980. The British philosopher J. R. Lucas had a similar scenario in mind when he wrote, in his book *A Treatise on Time and Space* (Methuen 1973, p. 50), "It is very important, not only for reasons of modesty, that I should not be able to use a Time Machine to go into a public library and read my own biography." Robert Heinlein didn't agree with Lucas: in his 1956 novel *The Door Into Summer* the protagonist, an inventor, travels thirty years into the future, where he reads some patent disclosures for inventions that he doesn't remember, even though they are in his name. He then returns to his own time and promptly files the patents!

[112] M. R. Levin, "Swords' Points," *Analysis*, March 1980, pp. 69–70.

In answer to the question "Who wrote the book about building a time machine?" the philosopher says this question is "no different from questions about where *anything* originally came from. We can ask about the origin of the atoms . . . their time line is not neatly presented to us. The atoms either go back endlessly, or if the Universe is finite, they just start. In either case the question of ultimate origin is as unanswerable as the question of the book's origin. What makes us think that when such questions are asked about the loop they are different and *ought* to be answerable is that the entire loop is open to inspection." While the instructions in the book don't tarnish, the book itself of course brings us back to our previous antique watch 'problem.' Suppose the book is brand-new at the start of the trip backward in time. Later, when the machine (and the book) have reached the end of the loop, just before beginning the trip back in time, have the pages turned yellow and brittle? If so, how do we account for the brand-new version? And if not, why not?

An analyst who takes strong exception to these two philosophers is Oxford physicist David Deutsch, who wrote (note 100) "the real problem with closed timelike lines under classical physics is that they could be used to generate knowledge in a way that conflicts with the principles of the philosophy of science, specifically with the evolutionary principle." What Deutsch is referring to is the metaphysical claim, attributed to the philosopher Karl Popper (see note 36 in Chap. 3), that *knowledge comes into existence only by evolutionary, rational processes* and that solutions to problems do not spring fully formed into the universe. One might call this the physics version of the work ethic—the creation of knowledge demands hard work!

Deutsch's idea had actually appeared *decades* earlier in a science fiction tale.[113] Time travel, discovered in the year 2007, is found to have a limited temporal reach into the future of 50 years, a limit due a law passed in 2057 banning time travelers from the past. To try to go past 2057 leads to a prompt arrest of the time traveler and a 'deportation' trip back to his own time. The story eventually explains that the law was passed precisely because of Deutsch's concern. As one character in the story explains, "Suppose [that one could travel more than 50 years ahead], then a time traveler from the past could get [new inventions], carry them back to his own time, and give them to scientists—which[would] cancel all the long period of invention which [produced the inventions]. Which [would] violate causal laws."

More recently, a philosopher has offered a quite interesting response to the Deutsch/Popper assertion. He writes (note 81), of information "appearing out of nowhere," that "These cases are puzzling, but they by no means show that the time travel scenarios in question are impossible or incoherent, *or even improbable*. We think it very improbable that . . . information should come from nowhere—but only because this does not happen very often. It does happen *sometimes*—for instance, when you say something and I mishear you. I think that you said something very

---

[113] P. Anderson and G. Dickson, "Trespass," *Fantastic Story Quarterly*, Spring 1950.

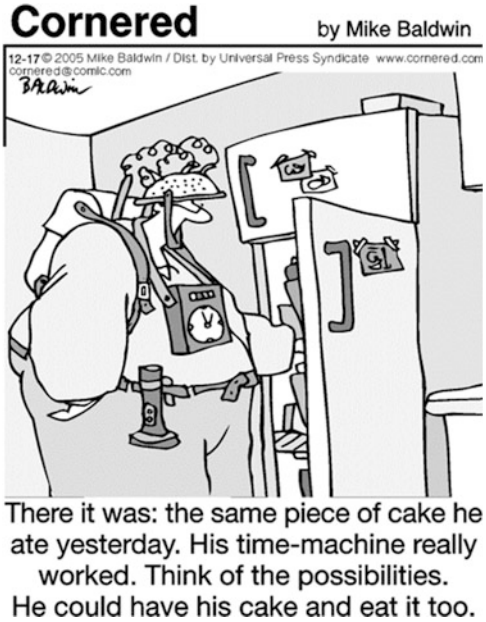
Cornered
by Mike Baldwin
12-17 © 2005 Mike Baldwin / Dist. by Universal Press Syndicate www.cornered.com
cornered@comic.com
There it was: the same piece of cake he ate yesterday. His time-machine really worked. Think of the possibilities. He could have his cake and eat it too.

Fig. 4.5 A curious paradox. CORNERED ©2005 Mike Baldwin. Reprinted with permission of UNIVERSAL UCLICK. All rights reserved

profound—something which neither of us would, in fact, ever have thought of. Where does the idea come from? If this sort of thing were to start occurring regularly [as via causal loops], then we would simply accept it without raising an eyebrow."

In the final chapter I'll discuss a dramatic example (due to two Russian physicists) on how an information-creating time loop might be constructed using a wormhole time machine. Such a time loop wouldn't pass muster with Deutsch, of course, and he would consider such a thing as being as objectionable as is creationism, the anti-evolution claim that purports to 'explain' fossils (with measured ages in the millions of years) by simply *declaring* them as having been made by God just a few thousand years ago.[114] Deutsch's position is considered by nearly all scientists today to be correct *for the specific case of creationism*, but the evolutionary principle may be on shakier ground with respect to declaring causal information time loops to be impossible.

While philosophers have struggled with information in a time loop, and most physicists have carefully stepped around the issue, science fiction has had lots of fun with information in causal loops. Here's a sampling of such tales:

(1) A man receives telephone calls from *two* versions of himself, one ten years in the future saying he absolutely must accept an invitation to fly to the Bahamas

---

[114] *Why* would God do such a thing? Apparently 'just to have some fun with geologists and biologists,' as creationists call such ancient fossils 'sports of nature.'

that he will receive that very day, with the other version calling from tomorrow insisting that the plane will crash. What should he do?[115];

(2) Lovers who are irrevocably separated in time communicate by mail in one tale,[116] while lovers in another story[117] communicate via telephone calls to the ever more distant past (and yet, with the aid of a clever twist at the end, finally meet);

(3) A telephone lineman starts getting telephone calls from himself from 10 days in the future, with the first call telling him how to make the gadget to transmit such calls[118];

(4) A time traveling historian on a visit to A.D. 1528 from A.D. 2211 accidently gives a copy of the predictions of Nostrodamus to the prophet, thus explaining the predictions[119];

(5) A time machine experiment gone wrong allows thirteenth century Roger Bacon to meet twentieth century scientists, an encounter that explains the amazing forecasts in Bacon's *Opus Maius*[120];

Hollywood, too, has had some fun with information causal loops, with the best (in my opinion) example of that being the 1989 movie *Bill & Ted's Excellent Adventure*. In that film (where we learn that even the not very bright can be time travelers), a set of missing keys is necessary for the successful completion of a task. The two time travelers decide that after the task is done, they will go back in time, steal the keys (*that's* why they're missing!), and hide them so they can use them *now*. Where should they hide them? Why, "over there," says one of the boys, pointing at a hiding place—and sure enough, when they go over and look, the keys are there. They agree that once they have finished with the keys, it will be *most* important that they really do put the keys in the hiding place!

All of these examples that I've just given you, however, were *decades* too late to be the first in fiction about information in a time loop; that honor goes to the 1904 novel *The Panchronicon* by the lawyer Harold Steele MacKaye (1866–1928). An Edwardian literary time machine with style, the Panchronicon (literally, a 'machine for all time') swings on a rope tether around a steel post erected at the North Pole. By "cutting the meridians" faster than the sun does, it travels through space and time from 1898 New Hampshire to the London of three centuries earlier.[121] Using it, a time traveler fan of Shakespeare journeys from 1898 back to the bard, who is suffering from writer's block. There she whispers the magic lines from a play he is stuck on (lines she has memorized for her literary club meetings) into his receptive ear. Does this make Shakespeare a plagiarist? Of himself!?

---

[115] G. Klein, "Party Line," *The Best from the Rest of the World* (D. A. Wolheim, editor), Doubleday 1976 (story originally published in France in 1973).

[116] J. Finney, "The Love Letter," *The Saturday Evening Post*, August 1959.

[117] T. N. Scortia, "When You Hear the Tone," *Galaxy Science Fiction*, January 1971. See also L. Padgett, "Line to Tomorrow," *Astounding Science Fiction*, November 1945.

[118] M. Leinster, "Sam, This Is You," *Galaxy Science Fiction*, May 1955. This story was later broadcast as an episode on the "X-Minus One" radio drama program. See also F. A. Reeds, "Forever Is Not So Long," *Astounding Science Fiction*, May 1942.

[119] L. Del Rey, "Fools' Errand," *Science Fiction Quarterly*, November 1951.

[120] N. Schachner, "Lost in the Dimensions," *Astounding Stories*, November 1937.

[121] 'Time traveling' by crossing time zones is an idea that one can trace at least as far back as to Edgar Allen Poe's 1841 short story "Three Sundays in a Week."

4.5 Sexual Paradoxes

> "Once time machines exist, no event is low probability if it is needed to make the past consistent."[122]

There are causal loops even stranger than the ones we have already discussed, hard as that may be to believe. These are the sexual paradoxes, first mentioned in 1931 by Nicholson in his letter to Hugo Gernsback. Not only science fiction writers, but philosophers, too, have found these particular paradoxes full of dramatic appeal. For example, as a challenge problem to the readers of a scholarly journal, the British philosopher Jonathan Harrison (1924–2014) posed the following bizarre, indeed astonishing, situation.[123] A young lady, Jocasta Jones, one day finds an ancient deep freezer containing a solidly frozen young man. She thaws him out and learns that his name is Dum, and that he possesses a book that describes how to make both a deep freezer and a time machine. They marry. Soon after they have a baby boy and name him Dee.

Years later, after reading his father's book, Dee makes a time machine. Dee and Dum, taking the book with them, get into the machine and begin a trip into the past. Running out of food during the lengthy journey, Dee kills his father and eats him. Arriving in the past, Dee destroys the time machine, builds a deep freezer (again, using the book), gets into it, and . . . wakes up to find that a young lady, one Jocasta Jones, has thawed him out. When asked his name he replies Dum and shows Jocasta his book; they marry, and . . . .

Harrison concluded this amazing tale with this question for his readers: "Did Jocasta commit a logically possible crime?" That issue is just the *surface* of an ocean of puzzles in this story! Jocasta's crime, of course, is that she has (if unwittingly) committed incest; readers who remember the Greek myth of Oedipus, and who his mother/wife was, will understand why Harrison named his female character as he did. But what of Dee's crime? He has, after all, eaten his father! But perhaps that isn't a crime at all, because Dee and Dum are one in the same, and is it really a crime to eat yourself? According to another philosopher, Murray MacBeath, Harrison's story is "a story so extravagant in its implications that it will be regarded as an effective *reductio ad absurdum* of the one dubious assumption on which the story rests: the possibility of time travel."[124]

---

[122] From Robert Forward's 1992 novel *Timemaster*.

[123] J. Harrison, "Jocasta's Crime," *Analysis*, March 1979, p. 65.

[124] M. MacBeath, "Who Was Dr. Who's Father?" *Synthese*, June 1982, pp. 397–430.

This isn't to say that MacBeath was asserting that time travel is impossible. Indeed, he went on to declare that he did believe in the logical possibility of time travel, and his paper is devoted to discovering what thought to be incorrect in Harrison's story. He did that by retelling the story with what he believed are crucial modifications to make it sufficiently less outrageous that it could be taken at least somewhat seriously. In the new version, our hero, thawed out from a deep freezer, is now named Arthur. Arthur is, unfortunately, suffering from total amnesia (this is MacBeath's way of avoiding the psychological trauma of Dee remembering he ate Dum) and so, when asked his full name, he is himself sufficiently puzzled that he replies "Arthur who?" He is finally called (what else?) Arthur Who. And, as you can no doubt guess, his son (who is a genius and gets a PhD at age 14 on a dissertation dealing with the physics of time travel) becomes Dr. Who!

We are then told of a trip back into the past by the two, of the eating of the father (Arthur Who) by the son, of the entering of the deep freezer by Dr. Who, etc. etc. The whole business is quite entertaining and *at least* as complex as Harrison's original story. Just *how* complex is summed up in MacBeath's last, wonderful line: "The Who who was Dr. Who's father was not Dr. Who—that is, not the Dr. Who whose father he was."

MacBeath wasn't the only one that Harrison's story fascinated, and nearly a dozen replies to it were received in addition to MacBeath's. One, in particular, made the thought-provoking observation (see note 112) that not only has Jocasta committed incest but she has done so with a single act of intercourse. As discussed earlier, the events on a causal loop do not happen endlessly but rather only once; thus, Jocasta thaws Dum (Dee) out just once, she marries him just once, and the two consummate their marriage just once. Ordinarily we think it takes two sexual acts to commit incest, the first resulting in the birth of a child, and the second being a parent's union with that child, but this is not so in a causal loop. Time travel *is* an odd business.

Another philosopher replied to Harrison's story with a quite interesting claim, one that had actually been thought to be true for decades—but which today is recognized to be false. The claim was that, irrespective of physics, Harrison's story was biologically flawed and fatally so. As that philosopher wrote, "The biological problem is the following. Dee is the son of Dum and Jocasta. So Dee obtained half his genes from Dum and half from Jocasta. But Dum is diachronically identical with Dee and is therefore genotypically identical with him (that is, himself). That is, Dee is both genotypically identical and distinct from Dum, which is absurd."[125]

That this isn't true was pointed out by a philosopher many years later. In his paper we read this tale: "Suppose Adam travels [far] back in time … where he meets his mother Betty, mates with her and has a child which is himself. Is this possible biologically? Yes ... as follows ... on the grounds that we have total replication of Adam's genome."[126] (The *genome* is the totality of genes taken over all gene sites.) Now, suppose each such site holds two genes and, as Dowe points out, in sexual reproduction the father passes on to his offspring one gene for each gene site, to go with the gene the mother gives to each site. To exactly reproduce himself, then, the time traveling Adam 'simply' has to give his offspring, *at each site*, the gene that he has for that site that did *not* come from Betty. Thus, the offspring—baby Adam—ends up with a genome *precisely identical* to the time traveling Adam. This is, of course, an *extraordinarily unlikely* event, as the human genome has tens of thousands of genes. The probability that each and every site gets the 'right' gene from the time traveling Adam is therefore essentially zero. But it isn't *actually* zero and, as the quotation that opens this section says, a low probability to an event isn't a roadblock to its occurrence if that event is required for consistency.

---

[125] W. Godfrey-Smith, "Traveling in Time," *Analysis*, March 1980, pp. 72–73. This false claim had already been raised by a physicist (L. S. Schulman, "Tachyon Paradoxes," *American Journal of Physics*, May 1971, pp. 481–484), and even earlier by a science fiction writer (P. Anderson, "Time Patrol," *Magazine of Fantasy and Science Fiction*, May 1955).

While certainly instructive, the sexual paradox stories by Harrison and MacBeath are remiss in not indicating that the concepts they are dealing with have long been a staple of science fiction, and that the sexual paradoxes received much critical analysis in that genre long before philosophers (and physicists, too) discovered them. From science fiction, for example, we have a tale of young man who travels backward in time 1250 years, from A.D. 3207 to 1957, to become his own grandfather fifty generations removed.[127] And even that is tame compared to the sexual paradoxes other science fiction writers conjured up before philosophers began to discuss them.

In another story,[128] written decades before Harrison's and MacBeath's papers, we meet a young lady caught up in a mind twisting affair in which the mystery of a causal loop is the least of her troubles. In 1957 a girl is born, and after 20 years of intense competition with her mother (who has an uncanny ability to predict the future), she travels back from 1977 to a few months before her own birth. She becomes pregnant (by a man who she later discovers is her father) and gives birth to a girl. The new mother has, of course, knowledge of all that will happen during the next 20 years, including the fact that she will have an intense competition with her rebellious daughter ... .

While writers of stories like these in the early 1950s were there as trailblazers, it is a tale that appeared as the 1950s ended that is today generally acknowledged as the best sexual paradox story ever written.[129] We are given only a hint of what is to come when a character listens to a song called "I'm My Own Grandpaw!" In 1945, a newborn girl, Jane, is found on the steps of an orphanage. At age 18, in 1963, she has a one-night affair with a mysterious stranger that leaves her pregnant. Some months later, during the birth of a daughter, it is discovered that Jane actually has a double set of sexual organs, and because the female set has been ruined by the pregnancy, doctors restore her as a man. Soon after, the baby girl mysteriously disappears from the hospital ward. Years later, in 1970, Jane (now a man, of course) meets another stranger who uses a time machine to transport both of them back to April 3, 1963. By April 24 male-Jane meets female-Jane and impregnates her (and so now we know who the mysterious stranger was during the one-night affair!). Meanwhile, the stranger with the time machine travels forward to March 10, 1964, a little after female-Jane has given birth, kidnaps the baby from the hospital (thus clearing-up another mystery!), takes her back to September 20, 1945, and leaves her on the steps of the orphanage. And so we see that Jane is her own mother *and father*, thus out-doing all previous tales about self-parenting.

---

[126] Phil Dowe, "The Coincidences of Time Travel," *Philosophy of Science*, July 2003, pp. 574–589. See also J. Berkovich, "On Chance in Causal Loops," *Mind*, January 2001, pp. 1–23, and P. Dowe, "Causal Loops and Independence of Causal Facts," *Philosophy of Science*, September 2001, pp. 89–97.

[127] R. Dee (this is *not* the 'Dee' of Harrison's story!), "The Poundstone Paradox," *Magazine of Fantasy and Science Fiction*, May 1954.

[128] C. L. Harness, "Child By Chronos," *Magazine of Fantasy and Science Fiction*, June 1953.

[129] R. Heinlein, "All You Zombies—," *Magazine of Fantasy and Science Fiction*, March 1959.

This is pretty impressive stuff, but Heinlein still has one more twist for us. After leaving baby-Jane in 1945, the time machine stranger returns to April 24, 1963, retrieves male-Jane (who has just kissed female-Jane goodnight after fathering her-himself in herself), and takes him to 1985 where he recruits him into the Temporal Service—and finally, the stranger jumps forward to 1999, his 'real time.' At the end we at last learn that the stranger is, in fact, an even older version of male-Jane—*all* the central characters in the entire story are the *same* individual at various points along a single, highly twisted world line. The lone character in Heinlein's tale is truly a self-made man/woman in every sense of the phrase! This ultimate act of *creatio ex nihilo* has, correctly I think, been called "smaller than the minimal loop."[130]

Jane, in all her/his versions, is the only character in the story that appears to have purpose. In terrifying words that describe a causal loop, Heinlein ends the tale with an explanation of the story's title: "The Snake That Eats Its Own Tail, Forever and Ever. I know where I came from—but where did all you zombies come from? . . . You aren't really there at all. There isn't anybody but me—Jane—here alone in the dark. I miss you dreadfully!" In a December 1958 letter to his literary agent, Heinlein wrote of this amazing tale, "I *hope* that I have written in that story the Farthest South in time paradoxes." In my opinion, he did.

The sexual paradox has continued to fascinate science fiction writers up to the present day. In the novel *Timemaster* (note 122), for example, the hero at one point spends a night with his wife—and with two versions of himself from the future. He will, of course, experience that night two more times! Later, he becomes upset when his wife runs off with one of the older versions, but he quickly calms down when he considers that eventually *he* will be the older version. Consider, too, a story[131] that tells of a young man hunting the father who, years before, had abandoned him in a burning house. The death of the young man's mother in the flames has sent him on a 10 year quest for revenge up and down what is literally a river of time, a river on which to travel in one direction ("up time") is to move into the past, whereas moving "down time" leads to the future. Eventually he corners the father and, despite the man's pleading, kills him. It is only later, after examining papers he finds in his father's pocket, that the young man realizes he has killed his *future* self (Benford, a physicist, knows the pitfalls of time travel, and you'll notice that there is no autoinfanticide paradox here).

---

[130] S. Lem, "The Time-Travel Story and Related Matters of SF Structuring," *Science Fiction Studies*, Spring 1974, pp. 143–154.

[131] G. Benford, "Down the River Road," *After the King: Stories in Honor of J. R. Tolkien* (C. Tolkien and M. Greenberg, editors), Tor 1991.

4.6 Splitting Universes and Time Travel

> "In all time travel stories where someone enters the past the past is necessarily altered. The only way the logical contradictions created by such a premise can be resolved is by positing a Universe that splits into separate branches the instant the past is entered."[132]

One early science fiction technique for allowing backward time travel and a changeable past, while still avoiding paradoxes, is that of alternate universes. According to this idea, if a time traveler journeys into the past and introduces a change (indeed, his very journey may be the change) then, as the above quote states, reality splits into two versions, with one fork representing the result of the change and the other fork being the original reality before the change. (To a fifth-dimensional observer, of course, all conceivable forks, all possible four-dimensional spacetimes, have always existed.) Indeed, according to this view the entire universe is splitting, at every microinstant, along every alternative decision path for every particle in the cosmos! This is often called the theory of *alternate realities with parallel time tracks*.

Such a seemingly fantastic view seems to actually have some scientific plausibility because of the so-called *many-worlds interpretation* (MWI) of quantum mechanics, pioneered in physics by Hugh Everett III (1930–1982), in a 1957 Princeton doctoral dissertation. Everett's theory is the antithesis of what is commonly called the *collapse of the wave function*, the idea that all potential possibilities have a non-zero possibility until a consciousness actually decides or observes which one will actually *be*. That quantum mechanical concept gets its name from the probabilistic wave equation formulated in 1926 by the German physicist Erwin Schrödinger (1887–1961). Before the observation, all possible futures have various values of probability; after the observation (which 'collapses' the wave function) exactly *one* of those futures (*the* future) has probability 1 and all the others have probability 0.

The MWI idea can be seen in Hale's story "Hands Off," discussed earlier, and in art *40 years* before that! With almost certainly a theological twist, consider the

---

[132] M. Gardner, "Mathematical Games," *Scientific American*, March 1979.

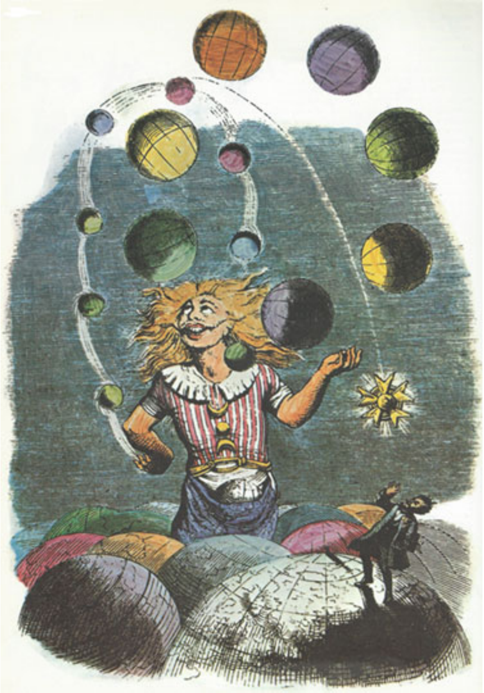

Fig. 4.6 Grandville's Infinity Juggler of many-worlds

beautiful illustration in the 1844 book *Un Autre Monde* (*Another World*), reproduced in Fig. 4.6. Known either as "The Infinity Juggler" or "The Juggler of Worlds," it is the work of the French artist Jean-Ignace Isidore Gérard (1803–1847), who published under the name 'Grandville.' The juggler—Grandville's version of Hale's mentor—appears as a court jester who is clearly having fun manipulating his multitude of worlds, while the man (humanity?) in the foreground watches. The man appears to be simultaneously fearful and fascinated, involved yet clearly impotent. Is Earth one of the worlds among which the Jester stands, or is it one of those flying through space? Or is Earth, perhaps, simply the unfortunate world ingloriously stuffed down the front of the Jester's pants? (That would surely explain a lot!) If born a hundred years later, Grandville would surely have found work as an artist in the imaginative world of the science fiction pulps.

Early science fiction stories that treat the collapsing wave function concept can be traced back to the late 1930s and early 1940s.[133] A particularly interesting example is the story of an inventor who, while trying to build a radio with which to signal Mars, accidently stumbles on the "temporal-aberrant carrier wave" and thus establishes contact with a universe that forked off of ours in 1863 when Robert E. Lee *won* the Battle of Gettysburg![134] In Everett's MWI, however, the wave function of the universe does *not* collapse. Indeed, it *couldn't*, because there is no observer external to the entire universe (we are talking science now, not of theology and God); instead, the wave function 'splits' at every decision point in spacetime. Although this leads to a multitude of realities *far* beyond comprehension, cosmologists still tend to like the MWI because it avoids the puzzle of having to produce an observer 'outside the universe.'

---

[133] See, for example, Jack Williamson's 1938 novel *The Legion of Time*, and C. L. Moore's, "Tryst in Time," *Astounding Stories*, December 1936. L. Sprague de Camp (1907–2000), too, was an early pioneer in the exploration of the MWI idea in science fiction long *before* Everett. In his 1941 novel *Lest Darkness Fall*, for example, he uses the analogy of a tree (the "main time line") that is always sprouting new branches.

It's important to understand that the MWI is different from yet another idea popular in science fiction, that of *parallel universes* (see again the third discussion question at the end of Chap. 3). In parallel universes *all* possibilities *always* exist, independent and parallel in time. In the MWI, on the other hand, ever more universes are continually coming into existence. Unlike the MWI, which can at least claim a scientific basis (quantum mechanics), there is no analogous theory for parallel universes. But, of course, even though lacking a theory, nonetheless science fiction writers have been quite inventive with the idea because parallel universes offer a way to avoid (at least some) causal loops.

One clever, early pulp story[135] illustrates how that works. To improve the performance of his time machine, an inventor needs batteries with tremendous energy density, a density far in advance of the batteries in the present. Unable to travel far into the future—if he *could* obtain them there, then of course a causal loop (the very entity we wish to avoid) would be created upon his bringing them back to the present—his assistant first travels back to 1851. There he leaves a note on desk of a well-known experimenter, with a plea for him to devote his life to battery research; a copy of the 1937 *Electrical Handbook* is left with the note as proof that there really has been a visit from the future! Before returning to the present, the assistant takes a sheet of (new) 1847 five-cent stamps from the experimenter's desk.

Returning to the present, which is now different (a new time track, in accordance with the splitting-universe idea), the powerful batteries *are* readily available *because* the experimenter believed the note. Buying several of them, using money obtained by selling the pristine 1847 stamps to a collector, the assistant returns to a slightly earlier 1851 than before (to *before* the fork in time!), watches himself appear[136] and leave the note and the handbook, and *then, unobserved, the assistant removes both*: Thus, upon returning once more to the present, he finds all is as before—except now he and the inventor have the powerful batteries. As before, one might ask where the batteries came from, but unlike the previous mystery of information-creating causal loops, the answer is clear and non-mysterious. *They came from the hard work of the experimenter on a different time track.* Such shuttling back-and-forth between time tracks is the signature of what is called a *cross-time* story, a device to avoid paradoxes while still allowing for changing the past. The first example of this time travel sub-genre had actually appeared 4 years earlier.[137]

---

[134] N. Bond, "Parallel in Time," *Thrilling Wonder Stories*, June 1940. See also S. N. Faber, "Trans Dimensional Imports," *Isaac Asimov's Science Fiction Magazine*, August 1980.

[135] W. Sell, "Other Tracks," *Astounding Science Fiction*, October 1938.

[136] Just like Marty McFly does at the end of the 1985 film *Back to the Future*. The movie is fun, but pulp science fiction did it first.

In another cross-time tale from modern times, we read of the horrible fate suffered by a man when an experiment in a Princeton physics lab goes wrong.[138] It is discovered, too late, that parallel time tracks are not simply grooves into which you drop, like a ball, after leaving the time track of our world. Each version of a person in each world is not like a ball rolling down a groove from past to future. Rather, each world's time track is just a line on a smooth surface; as the man is told, during a temporary stay in a world still close to his (our) original world, "We gave you a push sideways, and you moved off your original line—but instead of dropping into the next groove, you've just kept on rolling across the surface, from one line to the next, at an angle. There are no grooves, nothing to stop you from sliding on across the different lines forever. You have the same futureward vector as you started with, but you've added a small cross-time vector, as well." And so the man drifts cross-time, and gradually the worlds he experiences grow ever more alien.[139]

The science fiction is undeniably fun, but for this book the underlying *scientific* theory of time travel is classical (that is, non-quantum) general relativity, and that theory has nothing to say about alternative time tracks in multiple worlds. For most time travel theoreticians there is *one* time track, and the past of our world is unique and inviolate. I agree with the great quantum physicist J. S. Bell (1928–1990), who wrote of Everett's theory that "if such a theory were taken seriously it would hardly be possible to take anything else seriously."[140] As Bell further observed, in the MWI "there is no association of the particular present with any particular past," a quite strange idea that had already appeared in science fiction years earlier.[141]

While most early time travel analysts did base their work just on classical general relativity, there *are* now many more who think quantum mechanics itself, independent of its interpretation, has much to contribute as well. Perhaps, in fact, it may make an absolutely crucial contribution to the theoretical basis of time travel. One analyst who believes this (along with an even stronger belief in the MWI) is the British physicist David Deustch (note 100), who holds that general relativity is not the proper theory with which to study the physical effects of CTLs. He believes that the traditional mathematical machinery of general relativity actually obscures, rather than clarifies, the difficult task of separating the merely counter-intuitive from the unphysical.

---

[137] M. Leinster, "Sidewise in Time," *Astounding Stories*, June 1934. Splitting universes with multiple time tracks and time loops became quite popular after Leinster's and Sell's stories; you can find the basic idea repeated yet again in Alfred Bester's "The Probable Man," *Astounding Science Fiction*, July 1941, for example, in which each new journey into the past causes the future to fan out into an infinity of new time tracks.

[138] L. Watt-Evans, "The Drifter," *Amazing Stories*, October 1991.

[139] In Jack Haldeman's 1990 novel *The Hemingway Hoax* we read that "there is not just one [parallel] universe, but actually uncountable zillions of them."

[140] In "Quantum Mechanics for Cosmologists," *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987.

[141] J. R. Pierce, "Mr. Kinkaid's Pasts," *Magazine of Fantasy & Science Fiction*, August 1953.

Indeed, Deutsch calls the conventional spacetime methods, based on general relativity and differential geometry, *perverse*. He also does not like the conceptual problems raised by general relativity's wormholes and singularities. Any non-quantum mechanical discussion, he says, of the "pathologies" of backward time travel is simply not adequate. Deutsch divides these pathologies into two fundamental classes: (1) paradoxical constraints, such as the free-will issue seemingly raised by the grandfather paradox, and (2) causal information loops. Deutsch claims that his quantum mechanical analyses show that the first class of pathologies simply does not occur, because the past that the time traveler enters is the past of a world different from the one he has left. Further, his results also show (to him) that the pathologies of the second class may be "avoidable." These are not the views among the majority of time travel students, however (that does *not* mean Deutsch is wrong!), and general relativity *is* the standard tool used by the majority of time travel theoreticians. When quantum mechanics does enter the calculations of most analysts, it is generally on an *ad hoc* basis.

More concerning for the MWI view is a result reported in 2004, that a macroscopic object (a human time traveler, for example) attempting to traverse a wormhole time machine (to be discussed in some detail in Chap. 6) "must necessarily undergo violent interactions with the time machine," interactions *so* violent that they must "cause the object to disintegrate." The different pieces of the now certainly dead time traveler would emerge from the wormhole in different worlds—this is definitely not a result likely to encourage volunteers for the first time machine trip![142]

So, many physicists and philosophers, not sharing Deutsch's position,[143] tend to agree with Bell, including the late John Wheeler (Everett's thesis advisor!), who wrote of the MWI "I once subscribed to it. In retrospect, however, it looks like the wrong track. … Its infinitely many unobservable worlds make a heavy load of metaphysical baggage."[144] Agreeing with Wheeler was a philosopher who called the MWI "highly controversial" and declared that "few working physicists take it seriously."[145] Perhaps even more damning was a physicist's statement that "the idea of 10^{100}+ slightly imperfect copies of (the universe) all constantly splitting into further copies . . . is not easy to reconcile with commonsense. Here is schizophrenia with a vengeance."[146] Or, as one science fiction writer bluntly put it, in a tale of the inventor of the "chronomotive impulse belt" (which allows moving between the *two* parallel worlds that are all that exist), the MWI is the "Doctrine of Infinite Redundancy—which is, of course, utter nonsense."[147]

---

[142] A. Everett, "Time Travel Paradoxes, Path Integrals, and the Many Worlds Interpretation of Quantum Mechanics," *Physical Review D*, June 25, 2004, pp. 124023–1:124023–14.

[143] For a modern philosophical argument specifically rebutting Deutsch's enthusiasm for the MWI, see Theodore Sider, "A New Grandfather Paradox?" *Philosophy and Phenomenological Research*, March 1997, pp. 139–144.

[144] See note 34 in Chap. 3.

Deutsch's position does raise the obvious question of what motivates a quantum theoretician to study CTLs at all, given that they originate in general relativity and *not* in quantum mechanics. Deutsch's response is that although CTLs did indeed originate in classical Einsteinian general relativity, the still incomplete theory of quantum gravity *does* predict CTLs, too. And that is an exciting observation for time travel enthusiasts because, as Deutsch writes, the results of his quantum studies of CTLs show that "contrary to what has usually been assumed, there is no reason in what we know of fundamental physics why closed timelike lines should not exist." That view was later endorsed by other physicists who wrote, after a quantum mechanical study of how a particle could transit a time machine spacetime in a physically consistent manner, "there is no contradiction between the postulates of quantum mechanics and the possible existence of causality violation in general relativity."[148]

Long before these scientific endorsements, science fiction had enthusiastically embraced the many-worlds idea and its connection with time travel. The first such tale[149] appeared when Everett was just 3 years old; it put forth the insightful observation that although alternate time tracks may allow changing the past for the better (something that can't be done, for better *or* for worse, with a single time track), in the end any such change may still be futile. As Daniels' time traveler puts it, "I did have an idea to . . . go back to make past ages more livable. Terrible things have happened in history, you know. But it isn't any use. Think, for instance, of the martyrs and the things they suffered. I could go back and save them those wrongs. And yet all the time . . . they would still have known their unhappiness and their agony, because in this world-line those things happened. At the end, it's all unchangeable; it merely unrolls before us."[150] Many years later, in a critique of the many-worlds idea, a philosopher/physicist echoed Daniels' words: [In the world] that I (subjectively) experience I may blunder, but [in another world], with equal actuality, I triumph gloriously. The Everett interpretation can be used this way to mitigate sorrows, but this use is two-edged, for it equally well implies the speciousness of happiness."[151]

---

[145] R. A. Healy, "How Many Worlds?" *Nous*, November 1984, pp. 591–616.

[146] B. S. DeWitt, "Quantum Mechanics and Reality,' in *The Many-Worlds Interpretation of Quantum Mechanics* (B. S. DeWitt and N. Graham, editors), Princeton 1973.

[147] B. Shaw, "What Time Do You Call This?" *Amazing Science Fiction*, September 1971. When a bank robber in one world tries to make his escape into the other world, he literally runs into 'himself' trying to escape after robbing the 'same' bank in the parallel world!

[148] D. S. Goldwirth *et al.*, "Quantum Propagator for a Nonrelativistic Particle in the Vicinity of a Time Machine," *Physical Review D*, April 15, 1994, pp. 3951–3957. See, too, the earlier D. S. Goldwirth *et al.*, "The Breakdown of Quantum Mechanics in the Presence of Time Machines," *General Relativity and Gravitation*, January 1993, pp. 7–13.

[149] D. R. Daniels, "The Branches of Time," *Wonder Stories*, August 1935.

The editorial introduction to Daniels' pioneering tale is quite interesting: the opening line is "To say that this short story contains some revolutionary time-travel theories would be putting it exceedingly mild." That editor then went on to tell his readers, with great enthusiasm, that "when the author . . . submitted this story to us, his accompanying letter stated that in it he had settled the time-travel question once and for all. We must admit that a broad, unbelieving grin spread over our countenances when the author dared make this assertion. BUT—the smile soon left our faces . . . [T]o our chagrin, Mr. Daniels had really propounded so many brand new ideas about time and time-travel, and such logical ones—*that he has not left one loophole in his argument*!"

John W. Campbell (1910–1971), the first (and only) editor of *Astounding Science Fiction* (today's *Analog*), called alternate time track stories "mutant" because they represented the first new innovation (or mutation) in the time travel concept since H. G. Wells. Campbell incorrectly claimed *The Legion of Time* (note 133) was the first such tale (see Campbell's editorial in the May 1938 issue of *Astounding*) , and that "Other Tracks" (note 135) was the second, but in fact it was Daniels who was first with splitting time tracks in science fiction. After Daniels the concept quickly became part of standard science fiction lore and could be used by other writers with little explanation. For example, just a little more than a decade later one author did not have to say much about his "First Law of Chronistics," which determines the development of "the branches of Fan-Shaped time." It was sufficient for his readers to learn that should a time traveler to the past change anything, a parallel branch of time would be created on which the time traveler would be trapped: "The man who interfered with the space-time matrix, displacing even a comma in the great scroll of time, would be cut-off from his origin forever."[152]

Still, if there is one thing we can say about science fiction, it's that no 'rule' is immune to challenge. Decades after Daniels' tale put forth the MWI, we find the well-known author James Blish (note 100 in Chap. 1) *rejecting* it. In a story about the reception of radio signals from the future, we read of one character telling another "I *was* going to do all those things. There were no alternatives, no fanciful 'branches in time,' no decision-points that might be altered to make the future change. My future, like yours … and everybody else's, was fixed. It didn't matter a snap whether or not I had a decent motive for what I was going to do; I was going to do it anyhow. Cause and effect … just don't exist. One event follows another because events are just as indestructible in space-time as matter and energy are."[153]

---

[150] These sad, resigned words were written when the author, David R. Daniels (1915–1936), was just twenty years old. A year later he committed suicide.

[151] A. Shimony, "Events and Processes in the Quantum World," *Quantum Concepts in Space and Time* (R. Penrose and C. J. Isham, editors), Oxford University Press 1986.

[152] J. MacCreigh, "A Hitch in Time," *Thrilling Wonder Stories*, June 1947.

This denial of the MWI is simply an author's choice, of course, for whatever story effect is desired, and others may make different choices. Isaac Asimov, for example, used the MWI idea in in the story of a time traveler who journeys back to 1871 London, to retrieve a lost Gilbert and Sullivan operetta (*Thespis*). When he returns to the present he finds that his wife Mary (who was alive when he left) has been dead for a year on the new time track that his actions in the past have created. As the story ends, the devastated time traveler thinks "I had changed history. I could never go back. I had gained *Thespis*. I had lost Mary."[154] This sad fate is repeated in another story of a time traveler lost in an infinitude of time tracks with no hope of ever finding his way home: "In all of time, how many, many worlds there must be. How to find a single twig in such a forest?"[155]

Splitting universes have been used in literary works outside the genre of science fiction, as well. Examples include "The Garden of Forking Paths" by the Argentine writer J. L. Borges, the first play J. B. Priestly wrote (the 1932 *Dangerous Corner*), John Updike's 1997 novel *Toward the End of Time*, and Gore Vidal's 1998 novel *The Smithsonian Institution*. Typical of these fictional fantasies about splitting universes is a tale (anticipating Asimov's) by Lord Dunsany (1876–1957)—the Irish writer Edward Plunkett—the story of a man who goes back in time to correct "two or three mistakes he had made in his life."[156] This he successfully does, but the result is a new, subtly different subsequent history. The differences are not infinitely subtle, however; after the changes, he finds that his home, his wife, and all the delicate details of his life have vanished. As he relates to a visitor at the lunatic asylum he is now confined to, as the result of his despair, "I tell you I'm lost. Can't you realize that I'm lost in time? I tell you that you can find your way traveling the length of Orion, sooner than you shall find it among the years … Don't go back down the years trying to alter anything … Don't even wish to … [T]he whole length of the Milky Way is more easily traveled than time, amongst whose terrible ages I am lost."

In writing for a mass audience, rather than just for the more limited science fiction and fantasy one, perhaps the best known literary work of alternate history is the classic 1953 novel *Bring the Jubilee* by Ward Moore (1903–1978). In that work Lee wins the Battle of Gettysburg, and the South wins the Civil War. Using a time machine, a historian travels from 1952 (of the world in which the South wins) into the past of 1863 to study the battle, where he inadvertently disrupts events to the point that the North wins; that is, reality splits and the newly created fork represents the time track of *our* world. The historian is trapped on this new fork, cut off forever from his original time track. The entire novel is in the form of a discovered manuscript, written in 1873 and found in 1953, and the pathos of the ultimate isolation endows the novel with great emotional impact.

---

[153] J. Blish, "Beep," *Galaxy Science Fiction*, February 1954.

[154] I. Asimov, "Fair Exchange?" *Asimov's Science Fiction Adventure Magazine*, Fall 1978.

[155] M. F. Flynn, "The Forest of Time," *Analog Science Fiction*, June 1987.

[156] Lord Dunsany, "Lost," *The Fourth Book of Jorkens*, Arkham House 1948.

A 1992 novel on the same theme, Harry Turtledove's *The Guns of the South*, begins with a fascinating premise but then misses the crucial distinction between a single versus multiple time tracks. In that work racists from the future (2014) arrive by time machine at Lee's 1864 winter camp. They bring with them AK-47 automatic assault rifles and offer to supply Lee's army with all it can use. Lee accepts and the South wins the Civil War. The future, of course, changes—or does it? The time travelers have brought back books from the future showing that the South lost the war, so the implication is that history must have forked. So far, so good. But all through the novel, the time travelers move back and forth between the nineteenth and the twenty-first centuries, apparently finding their own time unchanged. And if that is so, then the whole point of the story vanishes. Why all the effort to change history when it is clear that nothing has changed? The novel is entertaining reading (Turtledove is a trained historian), but I believe Moore's novel to be the superior work of science fiction.

I'll end this discussion on splitting universes with a startling theological issue raised by a philosopher.[157] Arguing that God cannot branch into multiple time tracks because God is unique, the conclusion seems inescapable that God therefore exists on exactly one of how ever many different time tracks there may be. What if that chosen time track isn't ours? Then, concludes the philosopher, Nietzsche's nineteenth-century metaphorical claim that "God is dead" (for us) might literally be true! He admits that this is "fanciful," but still ……....

4.7 For Further Discussion

In an afterword to his story "Dead City" (*Thrilling Wonder Stories*, Summer 1946), Murray Leinster muses "You've heard the old argument that a man can't travel backward in time because he might kill his grandfather. I've wondered why nobody has argued that a man can't travel forward in time because he might be killed by his grandson." One possible answer to Leinster is that *if*, at the moment the forward-bound time traveler departs, he has not yet sired a child, *then* there simply wouldn't be a murderous grandson waiting for him in the future. Perhaps, however, Leinster had this somewhat more

(continued)

---

[157] Q. Smith, "A New Topology of Temporal and Atemporal Permanence," *Nous*, June 1989, pp. 307–330.

complicated scenario in mind: After the time traveler arrives in the future he is attacked by a mysterious stranger but survives, and later returns to the present. He then sires a child who will be the parent of that mysterious stranger. (As far as I know, this plot line has not appeared in a science fiction story.) Contrary to Leinster's view, explain why the *possibility* of being killed by a grandson is not a reason for forbidding the possibility of a trip in time (in either direction).

In his causal loop paper (note 105) the University of Delaware philosopher Richard Hanley correctly writes (on p. 146) "physicists have tried to avoid free will problems by ignoring causal loops involving intentional agency," and partly illustrates this claim with the autoinfanticide paradox, writing of the attempt of a time traveler to kill his younger self as inevitably failing because "the past is apparently brought about willy-nilly." (Hanley unfortunately then uses the story "Thompson's Time Traveling Theory" as an example of this—see note 77, and the end of the "Introduction"—when it is that time traveler's *grandfather* who is the intended target.) Discuss the merits of Hanley's claim, keeping in mind the end of Sect. 4.3 (in particular, note 81). If you are interested in genetics and astronomy as well as in time travel, then for extra credit comment on Hanley's claim (p. 137) that "one can extract information about my DNA from … my astrological chart."

The fictional killing of Hitler was imagined in print even before World War II, in Geoffrey Household's intense 1939 novel *Rogue Male* (made into the 1941 film *Man Hunt*). And so it's not surprising that one of the popular change-the-past themes in science fiction is that of a time traveler killing the Führer. (This idea, somewhat oddly, appeared in the debates leading up to the 2016 American Presidential election, when one of the candidates, to show the toughness of his character—even though he opposed abortion—declared "Hell, yes, I'd kill baby Hitler! You gotta step up, man." This candidate did elaborate a bit, stating there might be some risk involved with tampering with the past.) Stories in this sub-genre include E. Norden's "The Primal Solution" (*Magazine of Fantasy & Science Fiction*, July 1977), W. R. Thompson's "The Plot to Save Hitler" (*Analog*, September 1993), L. del Rey's "My Name Is Legion" (*Astounding Science Fiction*, June 1942), and R. M. Farley's "I

(continued)

Killed Hitler" (*Weird Tales*, July 1941). Read some of these tales and compare the various repercussions envisioned by the authors following an assassination of Hitler by a time traveler.

After our discussion of Heinlein's time travel masterpiece "All You Zombies—," you might think it impossible to write a new story that exceeds it in complexity. That might well be true, but a modern masterpiece by Ted Chiang certainly gives it a good run for the money. "The Merchant and the Alchemist's Gate" (*Magazine of Fantasy & Science Fiction*, September 2007) uses a 'wormhole' that connects the present to the future 20 years hence, and it is stuffed with intertwined causal loops and information bootstraps. Read it and keep track of all such occurrences. How many did you find? The last line of the story clearly expresses the view that the past cannot be changed: "Nothing erases the past. There is repentance, there is atonement, and there is forgiveness. That is all, but that is enough." Is the story always faithful to this view of time travel?

In his paper (note 126) on the coincidences of time travel, the University of Queensland philosopher Phil Dowe writes "It's true that remote time travel [into the very distant past] does not allow for causal loops . . ." *Is* this true? Consider, as you think about this, the story "Time's Arrow" (*Science-Fantasy*, Summer 1950) by Arthur C. Clarke. In that tale geologists have just discovered, in a remote desert, the fossilized tracks of a monstrous creature, from fifty million years ago, tracks that indicate that the beast was in hot pursuit of fleeing prey. Before the geologists can unearth the entire set of tracks, to see if the pursuit was successfully completed, they are visited by a physicist who just happens to be conducting near-by experiments in time travel. (This proximity is explained by noting what better place to conduct time travel experiments, powered by atomic energy, than in a remote desert?) At one point during the visit, after being told of the ancient pursuit frozen in rock, the physicist muses "It would save you a lot of trouble, wouldn't it, if you could actually *see* what took place in the past, without having to infer it by these laborious and uncertain [geological] methods." This comment results in the Chief geologist paying a visit to the physicist's lab. After driving over in a car equipped with tires having "an odd zigzag pattern" in the tread,

(continued)

an accident suddenly sends the entire lab into the past. Soon after, the other geologists unearth the rest of the fossilized tracks, and learn what the creature's prey had been when they see a zigzag pattern in the rocks, tracks that show "the great reptile was about to make the final leap upon its desperately fleeing prey." Can you see how to modify this story so as to have a causal loop involving the very distant past? (For perhaps even more inspiration on thinking about causal loops, watch the 1980 movie *The Final Countdown*. In it the designer of a modern naval warship that temporarily travels back through time to the Pearl Harbor of December 6, 1941, turns out to be a crew member who was accidently left behind in the past. In the past he *will be* able to design the ship because he already knows how it *was* designed—by himself!)

Comment, at length, on the cartoon shown in Fig. 4.5. (Does it make logical sense?)

In the story "Salvation" by Jerry Oltion (*Analog*, December 2007) a physicist approaches the Universal Church of the Divine Revelation for money to build a time machine. He is blunt in making his case: "You could go back in time and meet Jesus. Assuming he existed." That statement causes (it should come as no surprise) not just a bit of pandemonium but, nonetheless, an influential Church leader decides to provide the funding. Why? Because later, while sitting in his office as he talks with the physicist, a sheet of paper suddenly appears in the air above the leader's desk and then flutters down to land on the telephone. Picking the paper up, the leader sees it is a sheet of his own letterhead, with writing in his own angular, precise handwriting, saying "It works. Give him the money. You almost named the dog Solomon." This convinces the leader because, as we are told, "Paper appearing out of nowhere was a good trick, but it might easily be just that: a trick. Duplicating his letterhead and his handwriting wouldn't be all that difficult either. [On the other hand] knowing the name [the leader] had considered but rejected for his German Shepard 15 years ago was a different level of feat entirely." The physicist seems to be startled by the appearance of the paper, too, and asks "May I see that?" His reaction convinces the leader it wasn't a staged event: "Well, I'll be damned," the physicist replies. Once the time machine is under construction, the two men realize they have to send the enigmatic message back in time to complete the loop. As they prepare to do so, the leader asks a curious question. After retrieving the mysterious sheet of paper from his desk,

(continued)

he says to the physicist “Should I send the original [the one he is holding in his hand], or should I write another?” The physicist replies with “Write a new one. If we send the original, we put it in a closed loop and [we’ll] never get it back. We don’t want to lose the first object to travel in time. We’ll want that for the Smithsonian someday.” Does this make sense? Also, comment on whether or not the dog’s name is a bootstrap paradox.

A perplexing little time travel paradox, one that I don’t think science fiction has yet treated (and I’m pretty sure physicists haven’t had anything to say about it either), was cooked-up by the English philosopher Robin Le Poidevin in his 2003 book *Travels in Four Dimensions: the enigmas of Space and Time* (Oxford, pp. 180–181). There he writes “Peter and Jane, both 20 years old, are out for a walk one day in 1999 when suddenly a time machine appears in front of them. Out steps a strangely familiar character who tells Jane that he has an important mission for her. She must step into the machine and travel to the year 2019, talking with her a diary the stranger hands to her. In that diary she must make a record of her trip. Obligingly, she does as she is asked and, on arrival, meets Peter, now aged 40. She tells Peter to travel back to 1999, taking with him the diary she now hands him, and recording his trip in it. On arrival in 1999, he meets two 20-year-olds called Peter and Jane, out for a walk, and he tells Jane that he has an important mission for her.” Le Poidevin then writes that “the really tricky question is: how many entries are there in the diary when Jane first steps into the machine? We imagine it blank. But this is the very same diary as the one Jane hands to the 40-year-old Peter, which then contains her entry. And by the time Peter arrives back in 1999, it will contain his entry, too. But then, if the diary already contained two entries when Jane was handed the diary, then it would contain three entries when she handed it to Peter, who would then add another one, so the diary would have contained four entries when it was first handed to Jane, and so on. If the problem is not immediately apparent, this is because we imagine an indefinite number of trips, but in fact there are just two: Jane’s trip to 2019 and Peter’s trip to 1999. So there ought to be a consistent answer to the question, how many entries are there in the diary? Yet, as we have seen, there does not appear to be a consistent answer.” Another philosopher soon claimed he *did* have the answer: namely, 2. Read his paper (Erik Carlson, “A New Time Travel Paradox Resolved,” *Philosophia*, December 2005, pp. 263–273), and either explain why you agree with Carlson’s reasoning or enthusiastically rebut it.

As in "The Time Eliminator" (Fig. 4.1), other stories have imagined gadgets that simply view the past, rather than visit it as would a time machine. This is done in an attempt to avoid paradoxes—but does it? Two stories that illustrate how just viewing the past risks affecting the past as much as time travel would, are Horace Gold's "The Biography Project" (*Galaxy Science Fiction*, September 1951) and Donald Franson's "One Time in Alexandria" (*Analog*, June 1980). In the first tale the Biotime Camera, operated by the Biofilm Institute, allows teams of biographers to film (alas, no sound!) and study the lives of past notable personages. Of particular interest are the lives of those who developed neurotic psychoses, such as Isaac Newton. And, indeed, the Biotime Camera does capture Newton's image as he begins to display increasingly disturbed behavior. We see Newton, for example, as he begins to peer into dark corners, looking for those who have come to spy on him. On his death bed, the biography team assigned to him reads his lips and discovers that his final words are "My guardian angel. You watched over me all my life. I am content to meet you now." It is then that the Biofilm Institute realizes what it has done. Newton *was* in fact being spied upon—by the Biotime Camera, which has not changed the past but has certainly affected it. In the second tale an archeologist uses a time viewer to read the lost manuscripts in the ancient library at Alexandria before it was completely destroyed in an inferno. The viewer uses an infrared beam—and it is the heat from that beam from the future that proves to be the origin of the fire in the past. Again, the past has been affected, but not changed, by time viewing. Is it true to claim, however, that such viewing gadgets could not be the source of other paradoxes, such as causal loops or information bootstraps? If you think that isn't a valid claim, give a counter-example.

In a story by Francis Flagg and Weaver Wright (a pseudonym used by Forrest J. Ackerman), "Time Twister" (*Thrilling Wonder Stories*, October 1947), we read the following exchange between the inventor of a time machine and his none-too-bright helper:

"You mean to say," he questioned incredulously, "that I could go back a hundred years?"

"If you had the proper machine in which to travel, yes."

"But that'd take me back to before I was born."

The Professor smiled tolerantly.

"Look at this diagram, Hank. This line is the time continuum. It incorporates space, too. [The authors didn't actually print a diagram with the story, but

(continued)

surely the Professor is using a Minkowski spacetime diagram]. This dot is you. It doesn't matter when you were born, or when you will die. You exist right now, that's the fact. Traveling into the past or future wouldn't make you grow any younger or older. Such a thought is naïve. Let me demonstrate the mechanics of it for you. If . . . we calculate with non-Euclidean mathematics . . ."

"It don't sound reasonable," the farmhand objected. "If I went back—"

"I know," interjected the Professor, "if you went back you might meet your own father as a young man and you'd be older than he, or maybe he and your mother would be kids going to school."

"Haw, haw! That'd be funny, that would."

What famous movie, made nearly 40 years later, does the end of the conversation remind you of? Hint: "flux capacitor."

As mentioned in the text, a famous science fiction example of *affecting* (but not changing) the past is "Behold the Man" by Michael Moorcock, the tale of a time traveler who arrives in ancient times during the very years of the ministry of Jesus as reported in the Bible. When he finds there is actually no such person, the time traveler takes the role himself and lives out the events as reported in the Gospels, including the Crucifixion. This is a powerful story, but it had already been done more than 15 years earlier, by Philip K. Dick, in his short story "The Skull" (*If*, September 1952). In his tale, Dick tells of a man from the twenty-second century who is sent by government authority back to the mid-twentieth century to kill the Founder of a religious movement, a movement that 'now,' 200 years later, threatens those same government authorities. History records that the Founder gave a powerful speech just before being arrested and executed, a speech that started the religious movement, and so the time traveling assassin is told to kill the Founder *before* he can give that speech. (The parallel between Jesus and the Founder should be obvious.) Read these two stories and compare and contrast how Moorcock and Dick handled time travel paradoxes. Comment, in particular, on the relationship between the assassin and the Founder. Moorcock's tale should be easy to find, and Dick's is available as a free pdf download (it is in the anthology *The Best of Philip K. Dick*, Halcyon Classics 2010, as well).

A subtle change-the-past sequence appears in the original *Back to the Future* film that is easy to miss. When the hero, Marty McFly, returns to 1955 in the time car, he leaves from the parking lot of the Twin Pines Mall, so named because of the two pine trees that stand nearby. Arriving in the past with literally a bang, the time car inadvertently destroys one of the (then) young pines. Near the end of the movie, when Marty returns to the future (1985), he finds that the mall is now called the Lone Pine Mall. This is charming and fun, indeed clever, but modern scholars of time travel reject it, and other claims of changing the past, as not being logical. (Shakespeare understood this point, when he has Lady Macbeth declare, concerning the murder of Banquo, “What’s done cannot be undone: to bed, to bed, to bed.”) What Marty’s trip *would* explain is why the mall would *always* have had the name of the Lone Pine Mall. Watch the movie and see how many other ‘change-the-past’ episodes you can find.

Chapter 5
Communication with the Past

"[As for travel to or for signaling the past] you'd have to exceed light speed which immediately entails the use of more than an infinite number of horsepowers."[1]

5.1 Reversed Time

"I have not discovered Mr. Wells' Time Machine."[2]

One way to communicate with the past is to 'simply' live backwards in time. Philosophers and other writers of speculative fiction were the first to wonder what things might be like in a world where the time asymmetry is reversed—that is, in a world where time 'runs backward.' Indeed, fascination with the idea of time reversal actually dates back thousands of years, long before science fiction, as it can be found in Plato's dialogue *Statesman*, written (most probably) 15 years before Plato's death in 347 B.C.

At one point, Plato offers an extended description of the world suddenly running backward in time in the ancient past. After one character is told that at that remote time "all mortal beings halted on their way to assuming the looks of old age, and each one began to grow backward," he asks "But how did living creatures come into being, Sir? How did they produce their offspring?" The answer is shocking: "Clearly . . . it was not of the order of nature in that era to beget children by intercourse . . . It is only to be expected that along with the reversal of the old men's course of life and their return to childhood, a new race of men should arise, too—a new race formed from men dead and long laid in Earth . . . Such resurrection

---

[1] An observation by Haskel van Manderpootz, professor of the "newer physics," in S. G. Weinbaum's "The Worlds of If," *Wonder Stories*, August 1935. Compared to the 'modest' Van Manderpootz, all other physicists in the world are a mere "pack of jackels, eating the crumbs of ideas that drop from [his] feast of thoughts."

[2] W. R. Inge (1860–1954), in his November 1920 Presidential Address to the Aristotelian Society at the University of London Club, in a sympathetic treatment of the possibility of a time-reversed world.

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_5

of the dead was in keeping with the cosmic change, all creation being now turned in the reverse direction."

The reversed-time world is an important philosophical concept. Before the turn of the century, for example, Francis Bradley (one of the early proponents of the block universe, you will recall from Chap. 2), thought about reversed-time worlds and concluded that they would be quite odd: "Let us suppose ... that there are beings whose lives run opposite to our own ... *If* in any way *I* could experience *their* world, I should fail to understand it. Death would come before birth, the blow would follow the wound, and all must seem to be irrational."[3] A half-century later the South African philosopher J. N. Findlay (1903–1987) took Bradley's position of supporting a skeptical attitude towards the possibility of time-reversed worlds. Writing in a book review, Findlay declared "The reversed world in question wouldn't merely strike us as queer, but definitely crazy: it would be a world where what is wildly and intrinsically *improbable* was always occurring. It would, in fact, be much more startling than the original asymmetry that led us to think of it."[4] (Findlay was almost certainly thinking of things like a tea cup, shattered due to a fall, spontaneously reassembling itself.)

The question of backward-running time so fascinated Findlay that, some years later, he posed it as a problem for the readership of a scholarly journal (*Analysis*). This led to a number of responses, and subsequently he presented both his own negative view of time-reversed worlds and that of the best reader response he had received to his posed problem (which came from McGechie).[5] While Findlay showed admirable open-mindedness by awarding the title of *best* to an argument that refuted his own position, he remained unconvinced about the concept of reversed-time worlds, stating that "I continue to feel that a total reversal of my experiences is a terrifying possibility."

The terror aspect of living backward in time had been nicely captured in a science fiction story years before Findlay wrote. A scientist who is involved in an accident with radioactive materials has his sense of time flow reversed, and the story carefully and logically analyzes what his life would be like in such a situation. For example, the scientist can talk (backwards for others, of course), so he is understandable only if his words are recorded and then played in reverse. He cannot eat, because for him that would involve the regurgitation of food. He cannot answer questions because "if he should answer any questions put to him, it would mean he was giving the answer before he heard the question, on his time scale." And finally, he can't pick anything up because the normally stable position and velocity error-correction mechanism between eye, hand, and brain, which is a negative feedback system in normal time, has become an unstable positive feedback system in reversed time. The horror of his existence is contained in the only words the man utters (deciphered after reversed playback): “Where am I? What’s happened? Why are things so different? Why? Why?”[6]

---

[3] F. H. Bradley, *Appearance and Reality* (2nd edition), Oxford University Press 1897, p. 190.

[4] J. N. Findlay, *Philosophy* (25) 1950, pp. 346–347.

[5] J. N. Findlay and J. E. McGechie, "Does It Make Sense to Suppose That All Events, Including Personal Experiences, Could Occur in Reverse?" *Analysis*, June 1956, pp. 121–123.

Despite Findlay’s “terror,” however, the prevailing view today is that to the inhabitants of a time-reversed (or what is sometimes called a *counterclock*) world, *nothing would look odd*! This is a fairly new idea,[7] and not so long ago the philosophical literature displayed a misunderstanding of how a time-reversed world would appear to its occupants.[8] More recent analyses than Smart’s (note 7) advocating the ‘normality’ of a time-reversed world are more compelling.[9] One concern about a time reversed world however is not easily dismissed: matter with a time reversed time sense is thought by many to be antimatter in our world, and any interactions between the two worlds would be spectacular, indeed![10]

But let’s ignore that possible difficulty. The philosopher J. R. Lucas argued that even if beings from two such time-reversed worlds could meet, they still could not communicate: “If two beings are to regard each other as communicators, they must both have the same direction of time. It is a logical as well as a causal prerequisite.”[11] Now, this matter is well worth some effort to understand, because it is intimately tied to time travel. At first blush, Lucas’ words seem almost self-evident, and after a little thought they might seem to be absolutely irrefutable. The philosopher Murray MacBeath, however, took exception.

MacBeath opened his analysis[12] with a story to demonstrate that the persuasive power of Lucas’ position is only superficial. In that story of Jim and Midge, Jim is one of us, whereas Midge is a ‘Faustian time’[13] being. In his analyses MacBeath uses capitalized words and symbols for the time-reversed Midge, and lower case for Jim, as shown in Fig. 5.1. As MacBeath explains, “While Jim and Midge are together a face-to-face conversation is hardly likely to get off the ground. To make this clear let us say that they are together from t_0 until t_{10} on Jim’s time-scale, and from T_0 until T_{10} on Midge’s TIME-scale; t_0 is then the same temporal instant as T_{10} and, in general, $t_n = T_{10-n}$. If Jim at [his time] t_2 asks Midge a

---

[6] M. C. Pease,” *Astounding Science Fiction*, “Reversion,” December 1949. See also R. A. Banks, “This Side Up,” *Galaxy Science Fiction*, July 1954, for a tale about the confusion caused by projecting a film the wrong way in time.

[7] The modern view that a time-reversed world would appear normal to someone living in it can be traced back at least as far as to J. J. C. Smart, “The Temporal Asymmetry of the World,” *Analysis*, March 1954, pp. 79–83, an analysis, alas, that may not convince everyone.

[8] M. Dummett, “Bringing About the Past,” *Philosophical Review*, July 1964, pp. 338–359.

[9] See, for example, D. L. Schumacher, “The Direction of Time and the Equivalence of ‘Expanding’ and ‘Contracting’ World-Models,” *Proceedings of the Cambridge Philosophical Society* 1964, pp. 575–579; J. V. Narlikar, “The Direction of Time,” *British Journal for the Philosophy of Science*, February 1965, pp. 281–285; F. R. Stannard, “Symmetry of the Time Axis,” *Nature*, August 13, 1966, pp. 693–695.

[10] This issue is raised, several times, in Robert Silverberg’s 1968 novel *The Masks of Time*.

[11] J. R. Lucas, *A Treatise on Time and Space*, Methuen 1973, pp. 43–47.

[12] M. MacBeath, “Communication and Time Reversal,” *Synthese*, July 1983, pp. 27–46.

[13] In Goethe’s play *Faust*, the normal flow of time is routinely upset.

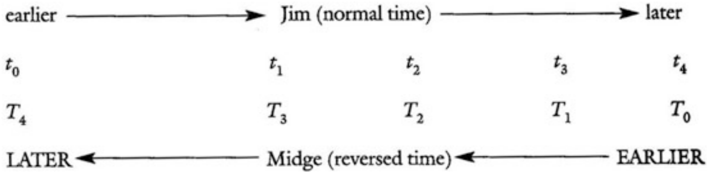

Fig. 5.1 Opposite time flows in counterclock worlds

question, and Midge hears the question at T_8, she will answer at T_9, and Jim will hear his question answered at t_1, *before* he asked it! What is more, if Jim is inexpert at interpreting backward sounds, and at t_4 asks Midge to repeat her answer, Midge will hear that request at T_6, BEFORE she has heard the original question; and her puzzled reply at T_7 will be heard by Jim at t_3 before he has uttered the request." Certainly this is a mess in time, and Lucas seems to be on safe ground with his denial of the possibility of communication between Jim and Midge. MacBeath, however, shows how to refute all of Lucas' arguments if Jim and Midge are allowed to be clever about how they send their messages back and forth—that is, if we give up some of our usual ideas of what a conversation is like. MacBeath's analyses are far too lengthy and detailed to present here, but the simplified diagram of Fig. 5.1 should enable one to follow the logic of his approach.

We imagine that Jim and Midge will not actually talk, and so will not have to decipher backward-spoken language. Rather, they will exchange messages via computer-generated text displayed on monitor screens, screens that are separated by a window that is proof against all penetration but the light emitted by those screens.[14] The nature of this window is not a trivial matter: if we accept the anti-matter nature of Midge's world then it is *essential* to keep her and Jim apart! To that end, MacBeath imagined that the window is double paned, with a perfect vacuum in-between. The exchange of photons between the two worlds should present no problems because photons are their own anti-particles.[15]

Now, imagine that at t_0 Jim brings a computer to the window. He programs it to wait for 4 days, until t_4, and then to display the following message on its screen: "This message is from Jim, who experiences time in the sense opposite to yours. Please study the following questions and display your answers on a computer screen three days from now." Jim's messages ends with the list of questions.

Because all that took place at t_4, Midge sees Jim's message and questions at what we will now call T_0. As requested, she brings her computer to the window, enters the answers to Jim's questions, and programs the machine to display them (after a

---

[14] Communication between beings in counterclock worlds, using written messages displayed through a window, appeared in science fiction *years* before MacBeath wrote: see I. Watson's 1978 novelette "The Very Slow Time Machine."

[15] There is no difference in the time sense of photons in either world because the flow of proper time for a photon—traveling at the speed of light, by definition—is zero (recall the discussion in Sect. 3.6).

3 day delay) on its screen. Thus, at T_3, which is Jim's t_1, Jim sees Midge's computer screen light up with "Hi, Jim. This is Midge. The answers to your questions are at the end of this message. Now, I've got some questions for *you*. Please display the answers two days from now." Midge's message ends with answers to Jim's questions and her list of questions.

Jim sees Midge's message at t_1, enters the answers to her questions and sets the machine to answer after a 2-day delay. At t_3, which is Midge's T_1—and by now you see how the process goes. It's cumbersome, sure, but it works. Or at least it does if everybody follows the rules. What if they don't? MacBeath provides other, increasingly complicated analyses to treat some of the more subtle problems that can be imagined in this method of exchanging messages. I will mention just two of them, which have direct analogs with what we normally think of as 'time travel.'

For the first problem, consider Jim's initial message, created at t_0 to be sent at t_4. He receives Midge's answer as described above at t_1, *before* his message is displayed through the window. So, what happens if at t_2 Jim cancels the message and it is *not* displayed? He has already gotten Midge's reply, but how can that happen if he does not send his message? This is, of course, a bilking paradox, with an explanation that we discussed in the previous chapter.

A second problem is the apparent possibility of creating a causal message loop. For example, let's say that at t_0 Jim suddenly decides to send a message through the window. (His reason for this sudden urge will be explained in the next few lines.) He thinks all night about what to send and, at t_1, finally settles on the following: "Greetings to the people on the other side of the window. This message comes from Jim, who hopes you will reply." Midge immediately sees Jim's message through the window (at her time T_3) and so is suddenly caught up with the desire to respond. She thinks all night about what to send and, hoping to be witty, she finally decides (at time T_4) on the following echo to Jim's message. "Greetings to the people on the other side of the window. This message comes from Midge, who hopes you will reply." Jim immediately sees this through the window (at what is his time t_0) and so now we know why he decides to send his original message! And Jim *will* send his original message—he has to because Midge replied to it.

In a review of a book on the direction of time, the philosopher Hilary Putnam restated the problems of a time-reversed world in the form of a provocative question: "How do you know that one man's future isn't another man's past?"[16] He began by making the interesting observation that for us to be able just to observe a backward-running universe, we would have to provide our own normal radiation source, because the counterclock stars in such a universe *absorb* radiation rather than emitting it. This point was elaborated on by another philosopher some years later, who wrote "We have uncritically imagined someone looking in on ... two worlds having opposite time directions ... Part of the story we tell, of the process of seeing, involves the emission of photons from objects [for example, the computer screens of Jim and Midge] and the subsequent impinging of these photons on our retinas. But this process is obviously directed in time. In a world where time ran opposite to ours, we could not see objects at all: objects would be photon-sinks, not photon-emitters."[17]

---

[16] H. Putnam, *The Journal of Philosophy*, April 1962, pp. 213–216.

Putnam concluded his comments about reversed time with a cautious warning: "It is difficult to talk about such extremely weird situations without deviating from ordinary idiomatic usage of English. But this difficulty should not be mistaken for a proof that these situations could not arise." That challenge is no doubt why so many writers of science fiction and fantasy have tackled the question of what it would be like if time ran backward. We can find such a tale long before the science fiction pulps, in fact, in a tale that appeared when Einstein was just 7 years old. In that story[18] the narrator (a professor of astronomy and higher mathematics) suddenly finds himself on Mars. There he encounters beings who know the future up to their deaths, and whose memories of the past are "scarcely more than a rudimentary faculty." The entire tale is in the form of a conversation between the professor and one such being, who argues (quite persuasively) for the virtues of his 'backward' existence compared to that of earthlings (the Martian name for Earth is the story's title). Bellamy realized that his story implies a fatalistic block universe: "No one could have foresight ... without realizing that the future is as incapable of being changed as the past," he wrote.

Other writers, too, were fascinated by the implications of reverse time. When Merlyn the magician makes his first appearance in T. H. White's 1939 masterpiece *The Once and Future King*, for example, he explains how he knows the futures of others: "Ordinary people are born forward in Time, if you understand what I mean, and nearly everything in the world goes forward, too ... But I unfortunately was born at the wrong end of time, and I have to live backwards from its front, while surrounded by a lot of people living forwards from behind. Some people call it having second sight."

What may have put the idea in White's mind for his time-reversed magician is only speculation today, but perhaps it was something he might have read a decade before, in a fellow Englishman's writing, the 1929 book *The Nature of the Physical World* by Sir Arthur Eddington, where one finds the following passage: "In "The Plattner Story" H. G. Wells relates how a man strayed into the fourth dimension and returned with left and right interchanged ... In itself the change is so trivial that even Mr. Wells cannot weave a romance out of it [but see one of the *For Further Discussion* questions at the end of Chap. 2]. But if the man had come back with past and future interchanged, then indeed the situation would have been lively."

Whether or not those words influenced English fantasy, they certainly had some effect on American science fiction. In his 1979 memoir *The Way the Future Was*, pulp editor Frederik Pohl wrote that Eddington's book (which Pohl incorrectly attributed to Sir James Jeans) had given him the idea for a story using the reversed-time twist. But before Pohl could publish it, an even better (claimed Pohl) tale arrived from Malcolm Jameson. Jameson, too, had read Eddington's book, and the result was the novella-length "Quicksands of Youthwardness," which Pohl published in *Astonishing Stories* as three-part serial during 1940–1941. Unfortunately, Jameson's tale is both pretty awful *and* devoid of *any* connection with Eddington's suggestion. One can only wonder about what might have been in the story Pohl says *he* discarded—did *it* have the future remembered? Alas, Pohl wrote that he couldn't remember!

---

[17] N. Swartz, "Is There an Ozma-Problem for Time?" *Analysis*, January 1972, pp. 77–82.

[18] E. Bellamy, "The Blindman's World," *The Atlantic Monthly*, November 1886.

Remembering the future does occur in one tale where everybody knows what will happen (as they live backward) by reading "prediction books." What distinguishes that story from many others on the same theme is an interesting, ironic conversation a student in a reverse-time world has with a philosophy professor about how things would be if time went the 'other way,' as in *our* world:

> "How can we tell? The reverse sequence of causation may be just as valid as the one we are experiencing. Cause and effect are arbitrary, after all."
>
> "But it sounds pretty far-fetched."
>
> "It's hard for us to imagine, just because we're not used to it. It's only a matter of viewpoint. Water would run downhill and so on. Energy would flow the other way—from total concentration to total dispersion. Why not?"[19]

The student is unconvinced, however, and when he tries to visualize such a peculiar world (*our* world, don't forget!) it gives him a "half-pleasant shudder." Imagine, he thinks in wonder, never knowing the date of your own death.

A few years later, in Wilson Tucker's 1955 novel *Time Bomb*, we find the intriguing idea of political assassination by time bomb, with the bombs actually *time traveling* to their targets. A policeman begins to suspect what is happening when it becomes evident that one of the explosions was actually an implosion: "The time bomb ... had been going in and had carried the force of the blast with it. Inward. Into the past. He frowned at that. A backward explosion? An explosion which ran counter to the normal flow of time, to the normal method of living? ... How would an explosion appear to a man if the blast happened in the opposite manner? If it began exploding now, in this moment, but continued backward instead of forward? Would it be an implosion?"

There is little doubt that the definitive treatment of a reversed-time world is that of Philip K. Dick's 1967 novel *Counter-Clock World*. Dick's world was once our world, but then, as in Plato's tale that began this chapter, time suddenly begins to run backward. People still alive reverse their direction of aging (but still think, walk, and talk in forward time), and dead, buried people come alive again and emerge from graveyards as the "Sacrament of Miraculous Rebirth" is intoned by a priest; all live their way back to the womb, just as in Plato's tale of 1600 years earlier. Such imagery is powerful stuff, but the physicist John Wheeler (of black hole fame) would have none of it. As he wrote, "Most of us would probably agree that the universe has not contained and will not contain any backward-looking observers. We do not expect to see caskets with corpses in them coming to life, nor do we expect to find bank vaults in which a gram of radium will integrate rather than disintegrate."[20]

---

[19] D. Knight, "This Way to the Regress," *Galaxy Science Fiction*, August 1956.

5.2 Multi-dimensional Time

> "If there are extra time dimensions we get violations of causality, because one could sneak to yesterday through the extra dimensions, and … if you had sneaked to yesterday, you would have disappeared from today."[21]

There are those, however, who haven't been quite so sure as Wheeler about the impossibility of a reversed-time world. One physicist, for example, showed how (under certain initial conditions at the Big Bang) there is a possible solution to the gravitational field equations that gives an oscillating universe that temporally runs backward during the contraction phase.[22] And a philosopher has argued that the *direction* of time is local, not global (just as special relativity showed is the *rate* of time) and that the arrow of time can point in opposite directions at different locations.[23] As odd as such ideas may seem, a generalization of reversed-time—*multi-dimensional* time—makes it seem small potatoes. This is the idea that there might be *many* possible directions to the arrow of time, not just two. At first this may seem an absurd idea, something akin to a man jumping onto his horse and riding off in all directions at once. But philosophers (and perhaps just a few physicists[24]) have started to take at least a semi-serious look at the concept; as with so many other of the radical concepts associated with time travel, though, science fiction writers were dealing with multi-dimensional time long before it became a respectable topic in learned philosophical and physics journals.

---

[20] Wheeler's comments can be found in the General Discussion at the end of *The Nature of Time* (T. Gold, editor), Cornell University Press 1966.

[21] An intriguing (if somewhat mysterious) thought from F. J. Yndurain, "Disappearance of Matter Due to Causality and Probability Violations in Theories with Extra Timelike Dimensions," *Physics Letters B*, February 28, 1991, pp. 15–16.

[22] H. Schmidt, "Model of an Oscillating Cosmos Which Rejuvenates During Contraction," *Journal of Mathematical Physics*, March 1966, pp. 494–509. An elaboration of Schmidt's ideas is in A. Walstad, "Time's Arrow in an Oscillating Universe," *Foundations of Physics*, October 1980, pp. 743–749.

[23] G. Matthews, "Time's Arrow and the Structure of Spacetime," *Philosophy of Science*, March 1979, pp. 82–97.

[24] More than half-a-century ago one writer asserted that two-dimensional *complex* time was old hat in the theories of spinning particles—see M. Bunge, "On Multi-dimensional Time," *British Journal for the Philosophy of Science*, May 1958, p. 39. For a summary of many of the objections to multi-dimensional time see J. K. Kowalczynski, "Critical Comments on the Discussion About Tachyonic Causal Paradoxes and the Concept of Superluminal Reference Frames," *International Journal of Theoretical Physics*, January 1984, pp. 27–60 (and the reply by E. Recami, September 1987, pp. 913–919).

In one pulp story, for example, we find a professor asking his redundantly named class in speculative metaphysics "Why shouldn't time be a fifth, as well as a fourth, dimension?"[25] In response to a generally skeptical reception to that, the professor goes on to say "I believe in the existence of a two-dimensional time scheme . . . Ordinarily, most people think of time as a track they run on from their births to their deaths . . . Think of this time track we follow over the *surface of time* as a winding road [it is the imagery of a surface that gives the professor *two* time dimensions] . . . Once in a while another road crosses at right angles. Neither its past nor its future has any connection whatsoever with the world we know."[26]

The year before, the same pulp had published another tale[27] that went well beyond a mere two time dimensions. We are told in that story of two countries on an alien planet at war in the distant future. The war is a stalemate until one side begins to fire a gun at its foe from just two miles from its target, in the heart of enemy territory—*from the middle of next week*! The gun's shells are true 'time bombs.' This is not mere 'ordinary' time travel along one time track, however, but a multidimensional effect. Using a photograph of the gun in actual operation to support his astonishing discovery, an agent for the side being shelled reports to his superior that "the gun and its crew are existing along another time axis at right angles to the direction of our 'normal time,' so that from our point of view they are existing perpetually in the same instant."

That explains why the gun crew can (will?) operate without interference in next week's future, as they are in their adversary's time only for the instant that the two time tracks intersect. Indeed, the spy used the same trick to obtain his undetected photograph: "I secured the photograph by orienting myself along still another time axis at right angles to that of the gun, and approached it as an instantaneous, invisible entity." By the story's end both sides are using and counter-using this technique, evading each other "to and fro along an ever increasing complexity of mutually perpendicular time axes." In fact, the final count exceeds 75 time axes, making Heinlein's two-dimensional time look rather skimpy by comparison.

Well, of course, 75 time directions *is* science fiction (I think), and physicists are not so enamored of multidimensional time as are science fiction writers. For example, Eddington wrote that he found the idea of any region of spacetime involving two-dimensional time to "defy imagination."[28] Another physicist showed that the extremal property of timelike geodesics (look back at Fig. 3.15 and its discussion) would fail for multidimensional time, which he then associated with the stability of matter and a failure of causality.[29] Yet another physicist, however, was just a bit more willing to consider multidimensional time, and suggested that a viable theory of quantum gravity *might* support the idea of multiple time dimensions.[30] Of just *what* more than one time dimension might actually *mean*, however, this same physicist echoed Eddington by writing "Physics in a spacetime of . . . two timelike dimensions would be very weird indeed." Agreeing with this physicist was a philosopher who called the idea that there could be more than one dimension to time a "rather wild possibility" and a "fairy-tale."[31] These are probably fair statements of how most physicists presently think of multidimensional time.

---

[25] R. Heinlein, "Elsewhen," *Astounding Science Fiction*, September 1941.

[26] This story has an amusing scene in which one of the professor's students accidently 'jumps time tracks' and so enters a new track with his arrow of time pointing backwards.

[27] N. L. Knight, "Bombardment in Reverse," *Astounding Science Fiction*, February 1940.

[28] A. S. Eddington, *The Mathematical Theory of Relativity* (2nd edition), Cambridge University Press 1924, p. 25.

But not all philosophers are of that persuasion, and many are in fact as fascinated by the possibilities of multidimensional time as are science fiction writers. So, *why* this interest in something so different from anything we actually experience? Where does the motivation come from? Of what *use* is multidimensional time? I think the answers to those questions all derive from how multidimensional time offers a theoretical model for giving meaning to the view that the past can be changed (take a look back at note 20 in the Introduction).

In a certain trivial sense, of course, the past is always changing. For each of us the past is the set of all events that have happened, arranged in a before/after temporal order, and this set is continually increasing (and so changing). That is not, however, what most people mean by a changeable past. What *is* meant is that there may be some kind of change in the temporal ordering of events, or that an event that once was (or wasn't) a member of the set of past events no longer is (or is now) a member. Two-dimensional time offers a way to make sense of such possibilities, which one-dimensional time simply cannot do. To see how that works, I'll follow the presentation in a paper that forcefully argues that it does make sense to talk about altering the past.[32]

Meiland was aware that some might find his model *ad hoc*, even "incredibly weird" (in his own words), but he justified his efforts by taking a refreshingly enlightened, non-Humean view of what he thought would be the proper response to meeting purported time travelers: "If strange machines containing people in futuristic garments and speaking strange tongues (or perhaps using ESP instead of speech) were to appear and were to claim to be from the future, we might very well begin to search for a theory of time that allows their claim to be true." In Fig. 5.2 you can see how Meiland tried to do just that.

---

[29] J. Dorling, "The Dimensionality of Time," *American Journal of Physics*, April 1970, pp. 539–540.

[30] C. Isham, "Quantum Gravity," in *The New Physics* (P. Davies, editor), Cambridge University Press 1989.

[31] D. Zeilicovici, "Temporal Becoming Minus the Moving-Now," *Nous*, September 1989, pp. 505–524.

[32] J. W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies*, November 1974, pp. 153–173. Jack Meiland (1934–1998) was a professor of philosophy at the University of Michigan.

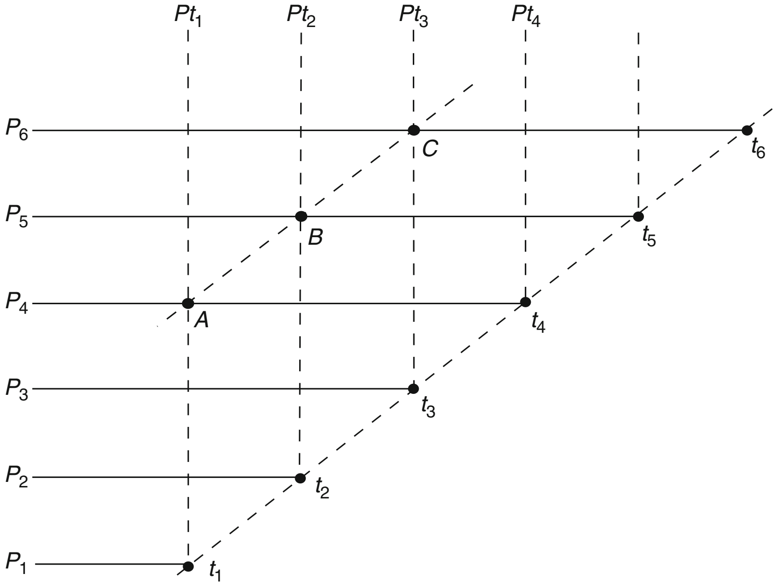

Fig. 5.2 Two-dimensional time

The dashed diagonal line, marked with the points t_1, t_2, ... represents our usual one-dimensional image of time. The horizontal lines P_1t_1, P_2t_2 ... (which we can simply call P_1, P_2, ..., for short) are the pasts for the present instants t_1, t_2, ... That is, P_1 is the past with respect to the present t_1, P_2 is the past with respect to the present t_2, and so on. The dashed vertical lines allow us to locate any moment in the past. For example, the intersection point A of P_4 with Pt_1 is the location of t_1 in the past with respect to the present t_4.

With this model, Meiland then analyzed in detail several interesting special cases. Suppose that t_1 and t_2 are 1 year apart and that there is a similar time separation between all adjacent, marked present moments on the diagonal. Let us further suppose that a time traveler at t_4 journeys backward 3 years to t_1, to arrive at point A. Assume he stays in the past 2 years; then his temporal locations lie along the dashed diagonal line ABC; that is, at B he is 3 years in the past of t_5, and at C he is 3 years in the past of t_6. From Fig. 5.2, then, we can imagine the time traveler saying, as he climbs into his time machine at t_4, "One year from now I'll be two years from now." That rather astonishing statement makes sense when we take both uses of now to be t_4 and observe that B (1 year from A) is 2 years in the past with respect to t_4.[33]

---

[33] A critic of time travel (see note 119 of Chap. 2) used what he claimed to be the absurdity of such a statement to support his ejection of time travel. One of Meiland's reasons for developing his two-dimensional model of time was, in fact, to be able to reply to that critic (Donald Williams).

Meiland's two-dimensional time model is undeniably fascinating, but it simply has no theoretical justification (as far as I know[34]). It is not necessary to assume two-dimensional time to explain Meiland's "strange machines containing people in futuristic garments"; it is possible to do that with one-dimensional time in a four-dimensional spacetime. The classic paradoxes, too, are understandable without two-dimensional time, as discussed in the previous chapter.

5.3 Maxwell's Equations and Sending Messages to the Past

> "Communication with a world exactly, to minutest detail, a duplicate of our own [but] twenty thousand years ahead of us might ruin the human race as effectively as if we had fallen into the Sun."[35]

Every physicist and electrical engineer knows that the mathematical description of the electromagnetic field is given by Maxwell's equations. In particular, radio engineers know that the waves of energy their antennas launch into space follow the predictions of those equations with astonishing accuracy. Indeed, when Einstein's relativity theory was completed, it was found that Maxwell's equations automatically satisfy relativity because magnetic effects *are* relativistic effects; in other words, relativity is built into Maxwell's equations. Whereas Newton's laws of dynamics had to be patched up, Maxwell's equations were untouched by the discovery of relativity.

Thus, it was a puzzle when physicists discovered that careful study of the seemingly perfect Maxwell equations, when applied to antennas, apparently results in the prediction of causality violation. It is found, in fact, that the equations have *two* solutions. One, as expected, contains the feature of time delay; that is, creating an electromagnetic disturbance at the antenna *now* causes a detectable effect at a distant point in space *later*. This is the so-called *time-retarded-solution*, and its common-sense physical interpretation is that of energy waves traveling away from the antenna as they also travel into the future. The shock was that Maxwell's equations also accept an *advanced solution*; energy waves *arriving* at the antenna from infinite space.

The physicist Paul Renno Heyl (1872–1961) wrote the perhaps first scientific work discussing advanced electromagnetic effects, in his 1889 University of Pennsylvania doctoral dissertation with the provocative title "The Theory of Light on the

---

[34] Many of the arguments against multi-dimensional time can be found in M. MacBeath, "Time's Square," in *The Philosophy of Time* (R. Le Poidevin and M. MacBeath, editors), Oxford University Press 1993. MacBeath concludes, however, with "I would not want to rule out the possibility . . . that time is three-dimensional. Or worse." See also Alasdair Richmond, "Plattner's Arrow: Science and Multi-dimensional time," *Ratio*, September 2000, pp. 256–274.

[35] The Victorian writer Samuel Butler (1835–1902), in the "Imaginary Worlds" entry of *The Notebooks of Samuel Butler* (published posthumously in 1912), commenting on the chaos that communication across time might cause.

Hypothesis of a Fourth Dimension." Heyl cited the scientific guru of the fourth dimension, C. H. Hinton (look back at the last "For Further Discussion" assignment in Chap. 2) as his inspiration. The situation described by Heyl is something like that of a child standing at the edge of a pond. She throws a rock into the middle of the pond and watches ripples spread out and away from the splash. Suddenly, she sees ripples appear all around the edge of the pond and then travel inward toward the center, where they all converge at once. A spout of water then erupts from the surface of the pond at the simultaneous meeting of the inward-traveling ripples, and she watches as a rock is ejected from the spout to land back in her hand. She is, of course, open-mouthed with astonishment! How absurd, you think, as you read this, and who could blame you? This amazing imagery of advanced effects we owe to the philosopher Karl Popper, and it has come to be called "the fable of the Popperian pond."[36]

Pursuing the mathematics of wave motion in the fourth dimension, Heyl wrote "We are led to the curious conclusion that, in Hinton's aether,[37] the nature of the central disturbance *after* a given instant can influence the form of the aether *before* that instant. In other words, the aether seems to be endowed with an uncanny faculty of foreknowledge." We can avoid such a counter-intutive implication of advanced effects, but only at the price of something many physicists and philosophers consider equally unacceptable: information traveling from the future into the past. We can still think of the advanced solution as representing electromagnetic waves of energy traveling *away* from the transmitting antenna—that is, as being *broadcast*, just like the retarded solution, rather than being received from infinity—*if* we also think of the waves traveling backward in time. Thus, the advanced solution to Maxwell's equations holds out the possibility of sending messages to the past, a sort of poor man's time travel. It may seem that we have simply traded one problem for another, however, because just sending information into the past can cause many of the same paradoxical, causality-busting situations that physical time traveling is claimed to cause.

The cosmologists Fred Hoyle (1915–2001) and J. V. Narlikar commented on the potential problems posed by communication backward in time, in their 1974 book *Action at a Distance in Physics and Cosmology*: "The [Maxwell] equations supply us with both advanced and retarded solutions (and, because of the linearity, with any linear combination of them) . . . With so many solutions theoretically possible, why does nature always select the retarded one? That this question cannot be answered within the framework of Maxwell's theory must be regarded as one of its intrinsic weaknesses."

---

[36] K. Popper, "The Arrow of Time," *Nature*, March 17, 1956, p. 538. See, too, note 110 (and its discussion) in Chap. 2.

[37] It was thought, in nineteenth century physics, that electromagnetic waves need a *medium* through which to propagate (like ocean waves need water, and sound waves need air), a mysterious substance called the *aether* (or *ether*) that exists even in a vacuum. The 1871 Michelson-Morley experiment, however, implied that the aether simply does *not* exist.

That branding of Maxwell's theory as having an "intrinsic weakness" because of its prediction of an advanced solution was, I think, unwarranted. Indeed, the advanced solution can be given a perfectly reasonable physical interpretation.[38] Imagine a transmitting antenna sending electromagnetic waves to an identical receiving antenna. At any point in space between the two antennas there are electric and magnetic fields. Maxwell's equations allow us to calculate the fields produced by alternating currents in the two antennas. When we do an analysis of the relationship between the transmitting antenna's current and the fields, we use the retarded solution because the current is the cause and the fields are the effect. But in the analysis of the relationship between the fields and the receiving antenna's current, the situation is reversed, and the fields are the cause and the current is the effect. That is, the advanced solution is simply the mathematics relating the current in the receiving antenna *now* to the fields *in the past*.

An acceptance of both solutions has, in fact, been in the physics literature for nearly a century. As the Yale physicist Leigh Page (1884–1952) wrote decades before Hoyle and Narlikar, "While the advanced potentials, as well as the retarded potentials, satisfy the electromagnetic equations, the former has generally been discarded for the reason that it has been more in accord with the trend of scientific intuition to consider that the present is determined by the past course of events than by the future. However, if it is once admitted that the present state is uniquely determined by any past state, it follows that the future is also so determined, and hence the employment of a future state as well as a past state in specifying the present marks no inherent departure from our accustomed methods of description . . ."[39]

It may still be tempting, however, to just dismiss the advanced solution as a mere anomaly of the mathematics and to discard it on physical grounds. This is the traditional approach taken by physicists when confronted with non-causal solutions in any physical theory and, indeed, that was what Swiss physicist Walter Ritz (1878–1909) did with the advanced solutions to Maxwell's equations (an approach that involved Ritz during the last year of his life in a dispute with Einstein). For Ritz, the reversal of cause and effect simply did too much violence to his intuition to be taken seriously, and so he thought one must *impose* causality on Maxwell's equations (a condition they do not inherently contain) by *a priori* rejecting the advanced solution.[40]

Still, electrical engineers make similar kinds of judgements all the time, as when the solution of a quadratic equation for a passive (energy-dissipating) resistor gives both a positive value (which 'makes sense') and a negative value (which doesn't 'make sense' and so is simply ignored). There *were*, however, those who encouraged caution on this issue. Seventy-five years ago, for example, the eminent MIT electrical engineer Julius Adams Stratton (1901–1994) echoed Leigh Page's warning when he wrote of the disturbing advanced solution, "The familiar chain of cause-and-effect is thus reversed and this alternative solution might be discarded as logically inconceivable. However, the application of 'logical' causality principles offers very insecure footing in matters such as these and we shall do better to restrict the [Maxwell] theory to retarded action solely on the grounds that this solution alone conforms to the *present* [my emphasis] data."[41]

---

[38] See, for example, S. L. Schwebel, "Advanced and Retarded Solutions in Field Theory," *International Journal of Theoretical Physics*, October 1970, pp. 347–353, and L. M. Stephenson, "Clarification of an Apparent Asymmetry in Electromagnetic Theory," *Foundations of Physics*, December 1978, pp. 921–926.

[39] L. Page, "Advanced Potentials and Their Application to Atomic Models," *Physical Review*, September 1924, pp. 296–305.

[40] For more on this, see O. Costa de Beauregard, "No Paradox in the Theory of Time Anisotropy," *Stadium Generale* 1971, pp. 10–18.

And in a famous paper I'll discuss in the next section, Wheeler and Feynman declared that "We conclude advanced and retarded interactions give a description of nature logically as acceptable and physically as completely deterministic as the Newtonian scheme of mechanics. In both forms of dynamics the distinction between cause and effect is pointless. With deterministic equations to describe the event, one can say: the stone hits the ground because it was dropped from a height; equally well, the stone fell from a height because it was going to hit the ground."[42] For Wheeler and Feynman, the reversal of cause and effect inherent to backward causation and time travel to the past offered no conceptual difficulties.

The elimination of an appeal to causality, or to the 'weirdness' of advanced waves, in arguing for the naturalness of the retarded solution to Maxwell's equations was first done in 1976.[43] The only auxiliary condition applied to the equations was simply the natural one of requiring the initial field energy to be finite.[44] And yet, today, there is still no experimental evidence for the physical reality of the advanced solution. Now and then one does run across speculations that the advanced waves of *something* traveling backward to us from the future might explain the so-called ESP 'talent' of precognition, but that is all it is, speculation.[45]

Advanced waves appeared in the pulp science fiction of the late 1930s, a decade before Wheeler and Feynman. For example, one story actually specifically invoked the advanced solution to Maxwell's equations, with a gadget (making use of what the author called the "anticipated potentials") displaying the near future on a television-like screen.[46] The author's by-line proudly gave his academic credentials as including a master's degree and, in fact, John Pierce was a graduate student in electrical engineering at Caltech. He received his doctorate just months after this story was published, and then went on to a highly distinguished career at Bell Telephone Laboratories and then later at the Jet Propulsion Laboratory operated by Caltech for NASA. Pierce knew all about Maxwell's equations, of course, and he actually opened his tale with a quote from Page's 1924 article (note 39) on the advanced solution. How many fictional pieces include quotes from the *Physical Review*?

---

[41] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill 1941, p. 428.

[42] J. A. Wheeler and R. P. Feynman, "Classical Electrodynamics in Terms of Direct Interparticle Action," *Reviews of Modern Physics*, July 1949, pp. 425–433.

[43] P. C. Aichelburg and R. Beig, "Radiation Damping As An Initial Value Problem," *Annals of Physics*, May 1976, pp. 264–283.

[44] J. L. Anderson, "Why We Use Retarded Potentials," *American Journal of Physics*, May 1992, pp. 465–467.

[45] See M. B. Hesse, *Forces and Fields*, Philosophical Library 1961.

[46] J. R. Pierce, "Pre-Vision," *Astounding Stories*, March 1936.

And it would probably take something like advanced waves to explain the funny doings 2 years late in a story of a man caught in a time machine accident. Nearly all of his body ends up 4 years in the future—but only *nearly* all, because his eyes remain the present! As one of the puzzled observers of this odd business wonders, "Strange, that his eyes, now, can convey a message to his brain, four years hence, and his brain tells the eye muscles to move the eyeballs which are four years behind them —."[47]

Some of the most intriguing paradoxes of time travel involve no traveler—only information. Of course, any information flow at all, independent of time travel, involves the flow of energy and, as Einstein showed, energy and mass are different aspects of the same thing. Accordingly, information time travel involves the transfer of mass/energy. Thus, a man in the twenty-fifth century who sends a backward-in-time 'temporal radio' message to a twentieth-century woman stating that he loves (will love?) her is sending much more than mere emotion. Just *how* to send a message backward in time is, of course, the puzzle.

Indeed, all forms of present-day communication are transmissions only to the future. If you speak to someone, or if you send a radio message, there are always delays depending on the distance of separation and the speed of transmission of sound and light, respectively. If you want to send a message to the one hundred and twenty-fifth century, you can; just write a letter and seal it in a pressurized bottle of helium. This basic idea is dramatically presented in a novel about a scientist who is accidently transported from 2162 back to the late-Cretaceous, 80 million years into the past. There he leaves a written record on seven sandstone slabs of his brutal, lonely life among the dinosaurs—letters across time, if you will—found by twenty-second century geologists some years after his disappearance.[48] The transmission of such letters, *forward* in time, while dramatic, is not a puzzle. But what could be more astonishing than the message received *from the future* by the young inventor of the first time machine: after his initial experiment of sending his pilotless machine into the future, it returns with an envelope inside. Eagerly tearing it open, he finds the note is from the National Academy of Sciences: "We know from old records and museum models that this is the Cullen Foster experimental machine. Fifty years looks down on you and says 'Good work'."[49]

Heady stuff, that, but lots of other possible messages are capable of competing with young Foster's when it comes to generating excitement. For example, suppose you had a gadget that is superficially similar to a telephone but that calls telephones in the *distant* future. You can hear the person (in the future) on the other end, but they can't hear you (in their past). That is, information can flow only from future to past. It is then easy to imagine situations in their use of this device that at least seem paradoxical. For example, suppose you call your own private number 1 month ahead. You hear your future-self first answer the phone, and then recite the winning lottery for the 'previous' day, which is a month in your present self's future. (Your future self does this somewhat odd recital because a month from now you will remember, when your private phone rings, just *who* is calling!) So, now in the present you know you'll make a fortune by winning the lottery a month later.

---

[47] M. Schere, "Anachronistic Optics," *Astounding Stories*, February 1938.

[48] G. G. Simpson, *The Dechronization of Sam Magruder*, St. Martin's Press 1996.

[49] D. Stapleton, "How Much to Thursday?" *Thrilling Wonder Stories*, December 1942.

This example *is* admittedly somewhat mysterious since, for the gadget to call far ahead in time, some sort of signal (as yet unspecified) must travel into the distant future because *something will* make the future phone ring. For the present discussion I am ignoring this crucial issue for the sake of the dramatic impact of the example. Soon, however, we'll get a little way into describing how one might, in principle, actually build this gadget, which in the physics literature is called an *antitelephone*. (Such a device is an *anti*telephone because the person who is the receiver is in the sender's past, the opposite of the situation for an ordinary telephone.) An interesting fictional illustration of such a gadget, despite being told as a hard-boiled detective murder mystery, appeared in science fiction some years ago (alas, while called a "time telephone," no theory for its operation was given, but instead was 'explained' as a "straightforward application of an impressive, but limited, technology").[50] So far, there is nothing paradoxical (or even illegal) in all of this, but what if, when the phone rings in the future the day after you won the lottery, you perversely decide *not* to recite the winning number? This apparent paradox has, in fact, already been treated with the aid of the block universe view of spacetime; that is, if the future-you spoke the lottery number when you originally called, then the present—you (now in the future) *must*, inevitably recite it.[51]

Let's now make things a bit more involved. Suppose that instead of calling your future self, you call the weather service and listen to the recorded message telling you the weather, 30 days hence. You do this day after day, and after a while you get a reputation for being able to predict, *perfectly*, the weather for every day to come, up to a month into the future. Your reputation spreads far and wide, and after a while more the weather service hears about you. Meteorologists check and find you are *never* wrong. Their computer models are only 80 % accurate out to 3 days, and for a week's prediction and beyond, the general public might as well flip a coin on whether it will rain or not on any particular day. But you are 100 % correct out to ten times their range. And so they hire you—and as a secondary job, you also make the daily weather recordings. (The voice on the other end of the gadget *has* sounded sort of familiar!) Here, then, is the puzzle we encountered earlier in causal loops carrying information: from *where* is the information in the flawless weather predictions coming from?

---

[50] S. Schmidt, "Worthsayer," in *More Whatdunits* (M. Resnik, editor), DAW 1993. The author, Stanley Schmidt, has a Ph.D. in physics and is a former editor of *Analog Science Fiction* magazine.

[51] For a fictional illustration of this (a so-called *bilking paradox*), see W. Tevis, "The Other End of the Line," *Magazine of Fantasy and Science Fiction*, November 1961.

One easy answer is that the question is meaningless because such a future-to-the-past information flow must be impossible. Indeed, if I am to avoid telling a ‘philosopher’s fairy tale,’ like those I criticized earlier in the book, I must admit that one consistent, non-paradoxical answer is found in recognizing that I have *assumed* that those 30-day weather reports are correct. Maybe, however, they are no better than anybody else’s predictions. And so you *don’t* become famous, and you *don’t* get hired—and so there *isn’t* any paradox. Is that the way to avoid paradoxes involving information flowing backward in time?

Perhaps not. As long as 1917 it was realized that special relativity does not preclude such an apparent backward flow. That is, if information could be transmitted faster than light, *then* messages could travel backward in time. That was the year Richard Tolman (1881–1948), a professor of physical chemistry at the University of Illinois and later at Caltech, wrote “The question naturally arises whether velocities which are greater than that of light could ever possibly be obtained.”[52] He then answered that question, with his general conclusion being that if such velocities are possible, then a faster-than-light (FTL) observer could see the time order of two causally related events reverse. And thus the observer would see an affect *before* its cause. Alternatively, a *sub*luminal (slower-than-light) observer could see the two events, which are connected via an FTL interaction, reversed in time order from what a stationary observer would see.

Either situation has come to be called Tolman’s paradox, but Tolman himself was careful with his words: “Such a condition of affairs might not be a logical impossibility; nevertheless its extraordinary nature might incline us to believe that no causal impulse can travel with a velocity greater than that of light.” That was an astonishing statement, given that Einstein himself had specifically stated in his original 1905 paper on special relativity that such a thing simply could not occur. There is nothing, it would seem, to be “inclined” about.[53]

This rather technical connection between FTL speeds and backward time travel made the transition from theoretical physics to popular culture very quickly. It was in the British humor weekly *Punch*, for example, that the famous (but nearly always misquoted) limerick by A. H. R. Buller (1874–1944) first appeared:

There was a young lady named Bright
Whose speed was far faster than light;
She set out one day

---

[52] R. C. Tolman, *The Theory of the Relativity of Motion*, University of California Press 1917.

[53] Take a look back at Sect. 3.5, where we showed that the time order of two events can appear reversed for a subluminal observer if the two events are not causally related. Introducing FTL motion results in extending reversal to causally connected events; that is, FTL motion, reversed causation, and time travel to the past, go hand-in-hand-in-hand.

In a relative way
And returned on the previous night.[54]

Where *Punch* dared to go, Hollywood could not be far behind. Indeed, in this case it was actually there first, with the 1922 one-reel silent comedy movie *The Sky Splitter*. This was just a short film (feature pictures generally had at least four reels), so it is not clear how widely distributed and viewed it may have been. The story is that of a scientist testing a new spaceship: when it exceeds the speed of light, he begins to relive his life.

The linkage between time travel to the past and FTL motion is a central one in science fiction, and its fascination was nicely illustrated by one writer who has a time machine experimenter in the twenty-seventh century wonder "Was the speed of light the core of the mystery? At the speed of light did the past and the future become a shining, merging road down which men could walk—in their ears the thunder of time passing . . .?"[55] Not everybody was excited with the idea of FTL motion and travel backwards in time, however, with one eminent scientist declaring that "the limit to the velocity of signals is our bulwark against the topsy-turvydom of past and future."[56]

The obvious question at this point, of course, is whether it is even conceptually possible to build a gadget to send FTL messages backward in time? Einstein himself thought not, saying "We cannot send wire messages into the past."[57] But was he right? One hint at the possibility of achieving FTL speeds is in Dirac's 1938 paper (note 52 in Chap. 2). There, in his remarks about pre-acceleration, Dirac wrote "Suppose we have a pulse sent out from place A and a receiving apparatus for electromagnetic waves at a place B, and suppose there is an electron on the straight line joining A to B. Then the electron will be radiating appreciably [because accelerated charges radiate] before the pulse has reached its centre and this emitted radiation will be detectable at B at a time . . . earlier than when the pulse, which travels from A to B with the velocity of light, arrives. *In this way a signal can be sent from A to B faster than light* [my emphasis]."

This exciting conclusion goes a step beyond the usual examples of 'things that go faster than light.'[58] Dirac had an equally exciting reaction (and here the emphasis is his): "This is a fundamental departure from the ordinary ideas of relativity and is to be interpreted by saying that *it is possible for a signal to be transmitted faster than light through the interior of an electron. The finite size of the electron now reappears in a new sense, the interior of the electron being a region of*

---

[54] On page 591 of the issue of December 19, 1923.

[55] F. B. Long, "Throwback in Time," *Science Fiction Plus*, April 1953.

[56] A. S. Eddington, *The Nature of the Physical World*, Macmillan 1929.

[57] A. Einstein, "La Théorie de la Relativité," *Bulletin de la Société Francaise de Philosophie* 1922, pp. 91–113.

[58] Such as, for example, the intersection *point* of two very long, closing scissor blades. The explanation for how this can be is that the *point* is massless and does not participate in a causal chain (and so carries no information). Thus, special relativity is not violated.

failure, not of the field equations of electromagnetic theory, but of some elementary properties of space-time." This last line sounds very much like the things people say today about the singularity inside a black hole event horizon. And yet, Dirac was careful to point out that as weird as FTL speed may appear, special relativity is not violated because "in spite of this departure from ordinary relativistic ideas, our whole theory is Lorentz invariant." That is, even though 'faster-than-light' means 'backward in time,' which means 'causality failure,' special relativity still holds true and nothing awful happens to physics, only to our intuitions. The reason for this is that causality is *not* a premise or starting point for the special relativity.[59]

Of course, like any scientific theory, Dirac's theory is not necessarily the last word, and we have to admit the possibility that at least some of its implications (in particular, the possibility of FTL speeds) just aren't so. In all electronic communication systems that we use, information is transmitted by *modulating* a so-called *carrier wave*, and there is some reason to believe that such modulated waves cannot be sent at FTL speeds.[60] We must admit that it is one thing to talk of 'advanced wave radios'—often called *Dirac radios* in science fiction—and quite another to see how physics might actually enable one to talk to the past. FTL communication (without the time travel aspect) appeared in pulp science fiction before 1940, as in one story published the year after Dirac's paper.[61] In it we learn of a man on Pluto who has invented a way to send messages to Earth at twice the speed of light. He uses this gadget to warn of a would-be dictator who is on his way to Earth in a 'mere' light-speed rocket ship, and only an FTL message can warn Earth in time.

5.4 Wheeler and Feynman and Their Bilking Paradox

> "We find it difficult if not impossible to imagine waves that go into the future and *return to the present* [my emphasis] bearing information about where (and when) they have been."[62]

---

[59] See, for example, G. Nerlich, "Special Relativity Is Not Based On Causality," *British Journal for the Philosophy of Science*, December 1982, pp. 361–388. This same point was made nearly two decades earlier, in a study of the possibility of superluminal sound in superdense matter, by D. A Kirzhnitz and V. L. Polyachenko, "On the Possibility of Macroscopic Manifestations of Violation of Microscopic Causality," *Soviet Physics JETP*, August 1964, pp. 514–519.

[60] See G. Diener, "Superluminal Group Velocities and Information Transfer," *Physics Letters A*, December 16, 1996, pp. 327–331. For more on the modulation of a light-speed carrier wave in everyday AM radio, and in a more sophisticated single-sideband transmitter, see my book *The Science of Radio*, Springer 1999.

[61] N. Bond, "Lightship, Ho!," *Astounding Science Fiction*, July 1939. The author provides an interesting, detailed description of the gadget, and I think it would make a good question on a Ph. D. qualifying exam in physics or electrical engineering to explain the flaw in it.

[62] Bob Brier, *Precognition and the Philosophy of Science: An Essay on Backward Causation*, Humanities Press 1974. Brier is an Egyptologist (!) at Long Island University—with a Ph.D. in philosophy—who specialized at one time in parapsychology.

In 1941, at a meeting of the American Physical Society, Princeton University physicist John Wheeler and his student Richard Feynman discussed a seemingly outrageous idea that provided a possible clue to how a Dirac radio might function. The idea was that the advanced wave solutions to Maxwell's equations are not mere mathematical curiosities, but rather have profound physical significance. At the time, their talk received only a small abstract notice in the *Physical Review*, but after World War II they wrote it all up in a beautiful paper.[63]

Their primary goal was to explain the origin of the force of radiative reaction discussed by Lorentz earlier in the century. This reaction force is the cause of the energy loss suffered by an accelerated, charged particle. Lorentz, who thought of charged particles as having a finite size, attributed this reaction force to the retarded (by the time required for light to cross the width of the charged particle) coulomb repulsion force between one side of the particle's charge to the charge on the opposite side. This view, however, leads to various conceptual and mathematical problems, including an arbitrary assumption on how the charge is distributed over/through the finite volume of the particle, as well as the problems of infinite self-interactions and the issue of what keeps the charge from blowing itself apart by internal coulomb repulsion.

Wheeler and Feynman's theory, on the other hand, avoided those problems by postulating point charges, because a *point* charge cannot repel itself. But then whence the reaction force, if there is no repulsion? Their revolutionary explanation was first to imagine the accelerated point charge as emitting retarded radiation outward in space, eventually to be absorbed by distant matter. This distant matter, which itself consists of point charges that are accelerated by the retarded radiation, then radiates *backward* in time, back toward the original charge that started the chain of events. This backward-in-time, or *advanced*, radiation arrives in the past of the original charge, and *it* is the cause of the observed reaction force. Indeed, Wheeler and Feynman proposed that an accelerated charge will not radiate unless there is to be absorption at some other distant place and future time. That is, the *future* behavior of a distant absorber determines the *past* event of radiation; there is simply no such thing as just radiating into empty space. The entire universe, spatially *and* temporally, is a very 'connected' place!

Astonishingly, this non-causal view of spacetime had been around in physics for at least 20 years before Wheeler and Feynman's talk. They had independently developed their ideas but, after their 1941 talk, Einstein (who perhaps recalled his 1909 debate about advanced effects with Ritz) brought a 1922 paper by the Dutch physicist Hugo Tetrode (1895–1931) to their attention. In his paper Tetrode had written that "the Sun would not radiate if it were alone in space and no other bodies could absorb its radiation ... If for example I observe through my telescope yesterday evening that star which let us say is 100 light years away, then not only did I know that the light which it allowed to reach my eye was emitted 100 years ago, but also the star or individual atoms of it knew already 100 years ago that I, who then did not even exist, would view it yesterday evening at such and such a time."

---

[63] J. A. Wheeler and R. P. Feynman, "Interaction with the Absorber as the Mechanism of Radiation," *Reviews of Modern Physics*, April–July 1945, pp. 157–181.

Tetrode's vivid imagery had been, curiously, itself captured even decades earlier in words from the nineteenth century English poet Francis Thompson (1859–1907), in his "The Mistress of Vision":

All things . . . near and far,
Hiddenly to each other linked are,
That thou canst not stir a flower
Without troubling of a star.

None of this, of course, is obvious! As a tutorial paper appearing just 2 years after Wheeler and Feynman's 1945 paper expressed it, "Any physical theory which seriously proposes that events in the future may be the efficient cause of events in the past may be regarded—at least at first glance—as rather revolutionary doctrine."[64] Indeed!

It is interesting to note that Einstein apparently said nothing to Wheeler and Feynman about a paper that pre-dated Tetrode's by 3 years. In 1919 the Finnish physicist Gunnar Nordström (1881–1923) had suggested that the advanced solution might offer an explanation for a perplexing problem in atomic theory. Maxwell's theory says that an accelerated electric charge radiates energy, which implies that the orbital electrons in the classical model of the atom should quickly spiral in toward the nucleus, that is, all matter should collapse. This cataclysmic event (of course!) has not happened, and Nordström's idea was that if one took into account not only the usual retarded solution but the advanced one as well, then perhaps things could be understood. Indeed, Nordström was able to show that such an analysis does give zero for the *average* energy radiated by an orbiting electron. Later, however, Page (note 39) showed that the instantaneous radiated energy is not zero, and that this would lead to observable effects that in fact are *not* observed.

Now, to be sure that the 'doctrine' discussed in note 64 is clear, let me restate what Tetrode, and later Wheeler and Feynman, had in mind. Imagine we have an electric charge (the *source*) that we mechanically shake, that is, accelerate. This allows us to assign a definite *cause* to the charge's acceleration which, of course, radiates energy. This radiation travels outward into space as observed retarded fields until they are eventually absorbed by distant matter. The charges in that distant matter are thus accelerated, and they in turn therefore radiate energy. This induced radiation again consists, according to Wheeler and Feynman, of both retarded and advanced fields. The advanced fields radiate outward but *backward* in time toward the original charge, collapsing upon it at the precise instant we first shook it, thereby producing the radiative reaction force. At any instant of time, at any point in space, the observed field is the sum of the retarded field traveling away from the source into the future and the advanced field traveling toward the source in the past.

---

[64] C. W. Berenda, "The Determination of Past by Future Events: A Discussion of the Wheeler-Feynman Absorption-Radiation Theory," *Philosophy of Science*, 1947, pp. 13–19.

But, argued Wheeler and Feynman, there is one last point that has been left out of this picture—there is also an advanced field (traveling away from the source and backward in time) because of the original, mechanical shaking of the source charge. Equivalently, a field traveling *forward* in time will *converge* onto the source because we *will* shake it. Wheeler and Feynman showed that before the mechanical shaking that starts this whole process, the advanced radiation field of the source and the advanced radiation fields of the absorbers exactly cancel each other at every point in space and every instant of time (if there is total absorption in the future), which accounts for the experimental fact that we observe a zero total field before the mechanical shaking occurs.

Wheeler and Feynman showed that if we accept these (strange) ideas, then everything we actually observe is predictable: radiative reaction, the direction of the electromagnetic arrow of time from past to future (retarded-only effects), and the absence of infinite self-interactions. The claim by Wheeler and Feynman to have avoided self-interaction problems via the use of the advanced solution was, however, soon challenged. Indeed, the self-interaction of the electron is *needed* to explain the 1947 experiment by Willis Lamb (1913–2008) that measured the deviation (the *Lamb shift*) of the spectrum of hydrogen from what Dirac's theory of the electron predicts. Ironically, it was that experiment that helped motivate the renormalization of quantum electrodynamics (to get rid of the infinities then plaguing it) which led to Feynman's share of the 1965 Nobel Prize. In fact, just 4 years after their 1945 paper, Feynman expressed a revised view that self-interactions could not be avoided.[65]

In any case, we gain the rewards originally claimed by Wheeler and Feynman *only if* we accept backward time travel, a step too big for many in 1945 (and for nearly as many today) because of the resulting time travel paradoxes that seem to be unavoidable. For the same reason, Tetrode's earlier work, published in a German journal, also went virtually unnoticed during the two decades before Wheeler and Feynman's work. In fact, Tetrode wasn't the only anticipator of Wheeler and Feynman, as they had been anticipated, too, in America. In 1926 the chemist G. N. Lewis (1875–1946) had written "I'm going to make the … assumption that an atom never emits light except to another atom, and to claim that it is absurd to think of light emitted by one atom regardless of the existence of a receiving atom as it would be to think of an atom absorbing light without the existence of light to be absorbed."[66] Wheeler and Feynman were aware of Lewis by 1945. Certainly Wheeler and Feynman must have been intrigued by Lewis' paradox: "I shall not attempt to conceal the conflict between these views and common sense. The light coming from a distant star is absorbed, let us say, by a molecule of chlorophyll which has recently been produced in a living plant. We say that the light from the star was on its way toward us a thousand years ago. What rapport can there be between the emitting source and this newly made molecule of chlorophyll?"

---

[65] R. P. Feynman, "Space-Time Approach to Quantum Electrodynamics," *Physical Review*, September 15, 1949, pp. 769–789. See also C. Teitelboim, "Splitting the Maxwell Tensor: Radiation Reaction Without Advanced Fields," *Physical Review D*, March 15, 1970, pp. 1572–1582.

[66] G. N. Lewis, "The Nature of Light," *Proceedings of the National Academy of Sciences*, January 15, 1926, pp. 22–29.

The paradox in that, of course, arises from the issue of what happens if, at some intermediate time and place, the star's light is blocked, thus preventing its absorption by the chlorophyll? Could refusing to look at a star *now* affect the emission of the star's light in the *past*? Lewis was obviously making a clear statement of backward causation when posing this bilking paradox. His very next words show that he understood the probable reaction of his readers: "Such an idea is repugnant to all our notions of causality and temporal sequence." Like Tetrode's work, Lewis' ideas were ahead of the times but, actually, their ideas were *not* repugnant to everyone.

In fact, similar puzzles were an inspiration to Wheeler and Feynman and almost certainly motivated them to create their own famous bilking paradox, which they presented in their 1949 paper (note 42). They opened the presentation of their paradox as follows: "If the present motion of *a* is affected by the future motion of *b*, then the observation of *a* attributes a certain inevitability to the motion of *b*. Is not this conclusion in direct conflict with our recognized ability to influence the future motion of *b*?" This question clearly states the conflict between free will and determinism, and to sidestep this *human* concern Wheeler and Feynman constructed a "paradox machine," a machine that operates totally automatically and which has come to be called the "logically pernicious self-inhibitor"![67]

In their description of the paradox machine, Wheeler and Feynman ask us to imagine two charged particles, *a* and *b*, positioned five light-hours apart. As shown in Fig. 5.3, *a* is attached to the arm of a pivoted shutter, toward which a pellet is moving from initially a great distance away. Now, normally we would think of what happens next in terms of just retarded fields. That is, the pellet hits the arm, knocking it downward and thereby accelerating charge *a*; this acceleration of charge *a* creates a retarded radiation field that arrives at charge *b* 5 h later, resulting in the acceleration of charge *b*; the acceleration of charge *b* creates a retarded radiation field that arrives back at charge *a* 5 h later (10 h after the pellet hit the arm).

The Wheeler and Feynman view, however, claims that this description leaves out half the story—the *advanced* fields. Specifically, suppose the pellet will hit the arm and so accelerate *a* at 6 p.m. Then, *b* will be affected not only 5 h later at 11 p.m., but also *earlier* at 1 p.m. This advance acceleration of *b*, in turn, sends out an advanced field that arrives at *a* at 8 a.m. The paradox is now easy to see. As Wheeler

---

[67] P. Fitzgerald, "Tachyons, Backwards Causation, and Freedom," *Boston Studies in the Philosophy of Science* (volume 8), 1970, pp. 415–436. Even more extreme examples of such paradox machines are described in Tim Maudlin, "Time Travel and Topology," *PSA 1990*, Philosophy of Science Association, volume 1, pp. 303–315.

Fig. 5.3 Wheeler and Feynman's paradox machine

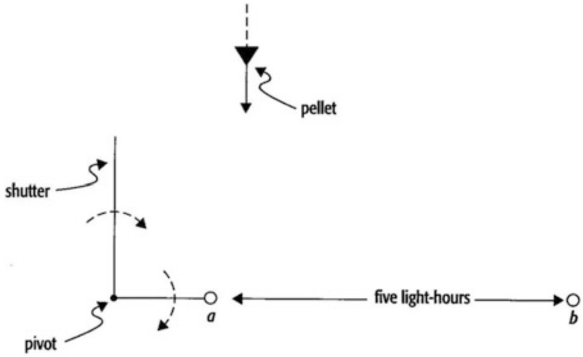

and Feynman described events, we see *a* exhibit a premonitory movement at 8 a.m. Seeing this motion in the morning, we conclude that the pellet will hit the arm in the evening. We could then return to the scene a few seconds before 6 p.m. and block the pellet from acting on *a*, a task automatically accomplished by the shutter in Wheeler and Feynman's paradox machine. But then we are faced with the puzzle of explaining just *why a* moved in the morning!

Wheeler and Feynman claimed they had resolved their bilking paradox by observing that *discontinuous* forces (more generally, *signals*) are never seen in nature. They concluded that the shutter does not completely block the pellet, but rather the shutter suffers a "glancing blow." That is, a very weak advanced signal is received by charge *a*, which moves the shutter just enough to induce the "glancing blow," and it is this partial interaction that results in the weakened signal in the "first place." This is, in fact, the very same explanation that was rediscovered decades later in answer to similar bilking paradoxes that involve self-interacting billiard balls transiting wormhole time machines (a topic we'll take-up in the next chapter).

5.5 Absorber Theory and Signaling the Past

> "If advanced waves [could be used to signal the past] then our grip on reality would become more tenuous. The past could never be considered over and done with, because anyone with the proper hardware could send messages back in time and alter what had already happened."
>
> —John Cramer, a University of Washington physicist, taking the minority position on the possibility of changing the past[68]

---

[68] Cramer has written provocatively on advanced waves. See, for example, "The Arrow of Electromagnetic Time and the Generalized Absorber Theory," *Foundations of Physics*, September 1983, pp. 887–902, and "Generalized Absorber Theory and the Einstein-Podolsky-Rosen Paradox," *Physical Review D*, July 15, 1980, pp. 362–376.

Wheeler and Feynman's argument *is* logically and physically sensible; it is, after all, simply an early statement of the principle of self-consistency in physics. Wheeler summed matters up nicely, years later, when he wrote "Interconnections run forward and backward in time in such numbers as to make an unbelievable maze. That weaving together of past and future seems to contradict every normal idea of causality. However, when the number of particles is great enough to absorb completely the signal starting out from any source, then this myriad of couplings adds up to a simple result: the familiar retarded actions of everyday experience, plus the familiar force of radiative reaction with its familiar sign."[69]

Their analysis was based on classical physics but, many years later, Feynman wrote (in his famous autobiographical work *Surely You're Joking, Mr. Feynman*) that at one time he and Wheeler thought it would not be too difficult to work out the quantum version of their theory. But then first Wheeler failed in the task, and then Feynman tried his hand at it and, as he stated, "I never solved it, either—a quantum theory of half-advanced, half-retarded potentials—and I worked on it for years." Their paradox (if indeed it *is* a paradox, since if advanced fields don't actually exist then there is no problem) remains unsolved.

Is Wheeler and Feynman's view of nature correct? Could we use advanced waves to send signals to the past? Or, if that requires some yet-to-be-developed technological breakthroughs in transmitter design, and if receivers are easier to construct, could we at least listen-in to the future (since we are, *now*, the future's past)? And if we could do that, could the future send us the details of the transmitter breakthrough (thus creating a causal information loop in time)?

The first experimental search for advanced waves seems to have been a 1973 effort.[70] The very next year, flaws in that search process prompted two physicists to discuss an experiment designed to detect advanced waves (if they exist). As they wrote, in a grand understatement, the exciting possibility of a positive result "would have such far-reaching consequences on our ideas of the unidirectionality of time and causality that … the experiment justifies a large amount of effort, even if no conclusive result is obtained for years."[71] Alas, all of the searches for advanced waves have, as I write (2016), given negative results and so the world still awaits the first Dirac radio.

Over the years the Wheeler and Feynman view of nature has been the target of some theoretical concerns. One physicist, for example, complained that Wheeler and Feynman had assumed a static, time-symmetric spacetime for the universe, in which the properties of all past and future absorbers are identical. That is obviously not so in an expanding (or contracting) universe and, as he wrote, “No serious modern cosmological theory is framed in [terms of] a static Universe.”[72] Another puzzle for that writer was that Wheeler and Feynman took a time-symmetric theory of half-retarded/half-advanced waves in a time-symmetric universe and arrived at a *non*-time-symmetric solution! They performed that trick by supposing not only that the universe is static, but also that it was created with the initial condition of low entropy. Thus, for Wheeler and Feynman, the one-way thermodynamic arrow of time is the primary arrow, with the electromagnetic arrow following as a consequence. (The *how* of a low entropy initial cosmological condition was left unexplained—certainly no mention of the hand of God appears in their work!) This ordering of the primacy of the temporal arrows was, in fact, in agreement with the view adopted by Einstein in his 1909 debate with Ritz, a view taken decades later by Hawking, as well.[73]

---

[69] See note 34 in Chap. 3. There Wheeler also wrote "The particles of the absorber are either at rest or in random motion before the acceleration of the source. They are correlated with it in velocity after that acceleration. Thus radiation and radiative reaction are understood in terms, not of pure electrodynamics, but of statistical mechanics."

[70] R. B. Partridge, "Absorber Theory of Radiation and the Future of the Universe," *Nature*, August 3, 1973, pp. 263–265.

[71] M. L. Herron and D. T. Pegg, "A Proposed Experiment in Absorber Theory," *Journal of Physics A*, October 1974, pp. 1965–1969.

Wheeler and Feynman had shown that both the advanced and retarded solutions taken together are self-consistent in a static universe; Hogarth’s question was whether the observed retarded solution, *alone*, would be self-consistent in an *expanding* universe (which is the universe we actually observe). His conclusion? It depends on the details of the expansion. Two years after Hogarth, two physicists expanded on his work and claimed to have shown that the retarded solution alone *is* self-consistent *if* the expansion is steady-state via the continuous creation of matter.[74] That would be the case because, if only retarded effects are to occur, then each emitter of radiation needs a large number of absorbers (such as ionized intergalactic gas) in its future light cone to provide for complete absorption. This, in turn, requires that the density of matter not decline “too fast” with the expansion. That is, the future universe must not be “too transparent” and the continual appearance of new matter in the ever-increasing volume of the expanding universe is required to maintain the necessary density.

That conclusion was embraced with particular enthusiasm by Hoyle, a British cosmologist whose name has long been identified with the idea of continuous creation of matter. Since then, however, continuous-creation cosmologies have fallen into disfavor because it was in 1965, just a year after Hoyle and Narlikar wrote, that the cosmic microwave background radiation was detected. That is now taken as very strong evidence for the occurrence of a primordial explosion (or Big Bang) that started the expansion of the universe, and as equally strong support for therefore rejecting a steady-state universe. Not by Fred Hoyle (1915–2001), though, who had an almost fanatical devotion to non-Big Bang cosmologies.[75] Real puzzles remain for the Big Bang universe, however. One is that it expands from a dense, opaque past into a less dense, ever-more-transparent future, with each emitter having a large number of absorbers in its *past* light cone. That should result, noted Hoyle and Narlikar (almost certainly with some glee), in an observed advanced solution and thus in a reversed electromagnetic arrow that would allow communication with the past. The fact that we have not (yet?) discovered how to perform such communication might be taken to mean that the idea of an expanding, Big Bang universe is somehow faulty. A related question about absorber theory is that of the puzzle of neutrino absorption. Neutrinos are particles that interact so weakly with matter that a beam of them would have to travel through many hundreds of light-years of lead for there to be a significant attenuation of the beam. How can such 'ghost-like' particles find enough future absorbers to make possible their observed journeys into the future of a Big Bang expanding universe?

---

[72] J. E. Hogarth, “Cosmological Considerations of the Absorber Theory of Radiation,” *Proceedings of the Royal Society A*, May 22, 1962, pp. 365–383. Hogarth, however, rejected the static universe, asserting instead that the *observed expansion* of the universe provides the required asymmetry, resulting in the cosmological arrow of time as the primary arrow and the electromagnetic arrow as a consequence.

[73] S. W. Hawking, “Arrow of Time in Cosmology,” *Physical Review D*, November 15, 1982, pp. 2489–2495.

[74] F. Hoyle and J. V. Narlikar, “Time Symmetric Electrodynamics and the Arrow of Time in Cosmology,” *Proceedings of the Royal Society A*, January 1964, pp. 1–23.

For such an exciting idea as communication with the past, it is not surprising that advanced-wave radio has appeared in science fiction. Just 6 years after Wheeler and Feynman's paper, a story by a well-known author hinted at such a gadget based on something called the "ultrawave effect": "While gravitational effects were produced by the presence of matter, ultrawave effects . . . did not appear unless there was a properly tuned receiver somewhere. They seemed somehow 'aware' of a listener even before they came into existence."[76] It is difficult to believe that such a story idea was conjured-up out of nothing, but rather that the author had read Wheeler and Feynman's paper. Anderson had a 1948 undergraduate honors degree in physics from the University of Minnesota, and so he may *well have* read Wheeler and Feynman's 1945 paper.

The potential bilking paradoxes produced by sending messages backward in time have been treated in at least one novel-length discussion. The puzzles presented are undeniably fascinating, but the story's answer to them is to allow the changing of the past, as argued in this section's opening quote from John Cramer. Indeed, the title comes from the plot device of twice changing the past by sending messages to the past to save the world from terrible disasters. Thus, we read through entire time periods *three* times before finishing the novel. As one character blurts out, "We can monitor the actual consequences of our decisions and actions, and change them until they produce the desired result! My God . . . it's staggering!"[77] Quite so.

One of the most interesting science fictional uses of backward-in-time signaling is, I think, found in a classic tale by James Blish. There the "Dirac radio" for instantaneous transmissions is described, and we learn that at the beginning of each received message there is always an irritating audio beep (hence the title) that is seemingly a useless artifact of the mysterious workings of the gadget. Its only obvious characteristic is a continuous spectrum from 30 to well above 18,000 Hz. It is only at the end of the story that the main character learns that this spectrum is the "simultaneous reception of every one of the Dirac messages which [has] ever been sent, or will be sent."[78]

---

[75] Besides his scientific work, Hoyle also wrote science fiction. One work, the 1966 novel *October the First is Too Late*, deals with travels in time but fails to say anything about paradoxes.

[76] Poul Anderson, "Earthman, Beware!," *Super Science Stories*, June 1951.

[77] J. P. Hogan, *Thrice Upon a Time*, Ballantine 1980.

Blish was actually pretty close to the mark with that, as a composite signal with a continuous spectrum (with energy distributed uniformly in frequency), such as one might expect the overlay of many independent signals to be, does indeed have a narrow time structure. If applied to a loudspeaker, such a signal would sound like a sharp pulse or click—or even a *beep*. In the limit of an infinitely wide spectrum, the time signal becomes one of infinite amplitude and zero duration, a singular *impulse* function called, by theoretical physicists and radio engineers alike, the *Dirac delta function*.

There is no mention in the story of advanced waves, but clearly Blish knew that instantaneous (infinite-speed) signals would travel into the past and he does a masterful job of presenting the mystery of listening to the future. At one point characters in the twenty-first century hear the commander of a time-traveling "world-line cruiser" transmit a poignant call for help from 11,000,000 light years away and from sixty-five centuries in the future. Most interesting of all, however, is Blish's statement of a technical issue that I have not seen raised before: if signals arrive at a receiver, simultaneously, from all future times, how can they be separated? Blish resorts to some scientifiction babble-talk to answer that question, but I believe it remains a puzzle.[79]

5.6 Tachyonic Signals and the Bell Quantum Antitelephone

> "We cannot fight the laws of nature."
>
> "Nature be damned! Feed more fuel into the tubes. We must break through the speed of light . . . Give me a clear road and plenty of fuel and I'll build you up a speed of half a million miles in a second . . . What's there to stop it?"[80]

Science fiction writers have often used FTL motion to reverse time, often without much (if any) regard to the fact that, to just *reach* the light barrier, requires (according to special relativity) infinite energy, much less to exceed light speed.

---

[78] J. Blish, "Beep," *Galaxy Science Fiction*, February 1954.

[79] The signal separation problem is also hinted at by physicist/science fiction author Gregory Benford, in a tale that was a precursor to his famous 1980 novel *Timescape* (in which the present attempts to warn the past of a future ecological disaster that threatens life on Earth). See Benford's "Cambridge, 1:58 A.M.," *Epoch*, Berkeley 1975.

[80] Words exchanged by the first officer and the captain of a starship on its way to Alpha Centauri in a story by N. Schachner, "Reverse Universe," *Astounding Stories*, June 1936. The captain, we are told, "had heard, of course, of the limiting velocity of light, but it meant nothing to him."

Therefore, goes the reasoning, because we can't get through the 'light barrier' means that time travel to the past must be impossible, as well. Or, so goes this line of argument. But, could there be a way around this conclusion? After all, while relativity theory indeed precludes the acceleration of a *massive* particle *up to* the speed of light, it does allow a zero rest mass particle (like the photon) to exist *at* the speed of light. Photons are emitted during various physical processes, and they move from the instant of their creation at the speed of light; the only way to slow a photon is to destroy it by absorbing it. Advocates of the possibility of the existence of FTL particles make a similar argument when asking if there might not be particles, emitted during various (as yet unknown) physical processes, that move from the instant of their creation at speed *greater* than that of light?

An affirmative answer would neatly avoid the 'acceleration through the light barrier' problem, but then there are other concerns. For example, such FTL particles would have to have an imaginary rest mass if they were to carry real-valued energy and momentum, and what could *imaginary mass* mean? That question was answered by the proponents of FTL particles, who replied that the rest mass of a superluminal particle would be unobservable because (like the photon) there is no subluminal frame of reference in which the particle could be at rest! That is, there is no frame of reference in which the mysterious imaginary mass could be measured and, anyway, it is only *observable* changes in the real energy and momentum that characterize particle interactions.

The key idea to this line of thought is a *supposed* FTL particle, called a *tachyon*, a name coined by the American physicist Gerald Feinberg (1933–1992) from the Greek word *tachys* for "swift."[81] It is interesting to note that Feinberg admitted[82] that his interest in such a thing was sparked by reading Blish's story "Beep" (see note 78). The idea of the tachyon is actually a very old one that is hinted at in the work of the Greek poet and philosopher Lucretius (who died 20 years before the birth of Christ). In his discussion of visual images, in Book 4 of his giant (well over 7400 lines) science poem *De Rerum Natura*, we find the following words about particles of matter originating deep inside the Sun: "Do you see how much faster and farther they must travel, how they must run through an extent of space many times vaster in the time it takes the light of the Sun to spread throughout the sky?"[83]

The first attempt (later found to be flawed) in the physics literature of a relativistic treatment of FTL particles appeared some years before Feinberg gave

---

[81] G. Feinberg, "Possibility of Faster-Than-Light Particles," *Physical Review*, July 25, 1967, pp. 1089–1105. Feinberg was anticipated in this name by Edward Page Mitchell (the Victorian pioneer in the time travel paradox genre who was discussed back in Sect. 4.2 and its note 37). In his story "The Tachypomp: A Mathematical Demonstration" (*Scribner's Monthly*, March 1874), he describes a gadget for reaching any speed, no matter how great (*tachypomp* is literally "quick sender").

[82] G. Benford, "Time and *Timescape*," *Science-Fiction Studies*, July 1993, pp. 184–190.

[83] A poetic allusion to something traveling faster than light appears, in of all places, Shakespeare's *Romeo and Juliet*. In Juliet's words (Act II, scene 5), "... love's heralds should be thoughts, Which ten times faster glide than the sun's beams, Driving back shadows ..."

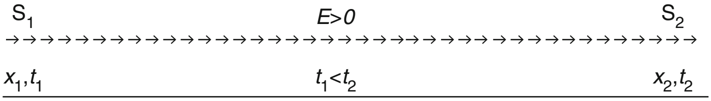

Fig. 5.4 The emission of a positive energy particle, followed by absorption

them a name, where it was observed that special relativity is *not* necessarily violated by FTL motion.[84] But nonetheless, concerns about the physical possibility of FTL particles continued. For example, a serious problem for tachyons, at least for those who dislike the ideas of backward causation and time travel, is the observation that in some frames of reference an FTL particle would appear to have *negative energy*. Feinberg, himself, explained (see note 81) this concern as follows: "By the principle of relativity, any state which is possible for one observer must be possible for all observers, and hence FTL particles can exist in negative-energy states for all observers . . . The occurrence of negative energy states for particles has always been objected to on the grounds that no other system could be stable against the emission of these negative-energy particles, an entirely unphysical behavior."

This objection to FTL particles was raised early in the history of tachyons, even before they were named, and it was addressed by three physicists who proposed the so-called *reinterpretation principle* (what I'll refer here to as the RP).[85] To see how the RP works, consider Fig. 5.4, in which a source S_1 at x_1 emits an FTL particle at time t_1. This particle then travels to an absorber S_2 at x_2, arriving there at the later time t_2. S_1 and S_2 are in the same reference frame and, for an observer in the frame, the particle energy E is positive. However, it is always possible to find another observer in a relatively moving frame for whom this process would look as though t_2 is less than (that is, earlier than) t_1 with $E < 0$. In other words, for the moving observer the particle would appear as a negative energy particle moving backward in time. (In the next chapter I'll show you that the particle speed must not be just superluminal, but the even faster *ultraluminal*.)

Note that for the moving observer, the emission by S_1 of negative energy *increases* the energy of S_1, and the absorption of negative energy by S_2 decreases the energy of S_2. S_2's energy decrease (for the moving observer) occurs before the increase in S_1's energy because, as noted before, for the moving observer $t_2 < t_1$. The moving observer naturally interprets this process as the emission of positive energy by S_2, followed by absorption by S_1. This reinterpretation would thus seem to preserve our common-sense idea of causality, as well as avoiding any mention of backward time travel. The RP appears to have slipped around those problems merely by redefining which source is transmitting, and which is receiving, the tachyon. Indeed, Feinberg claimed (note 81) that the RP avoids the creation of causal loops and their associated paradoxes, a claim repeated nearly 20 years later by a physicist who used the RP to eliminate paradoxes from Gödel's time travel rotating universe.[86]

---

[84] S. Tanaka, "Theory of Matter with Super Light Velocity," *Progress of Theoretical Physics*, July 1960, pp. 171–200. See also O. M. Bilaniuk and E. C. G. Sudarshan, "Particles Beyond the Light Barrier," *Physics Today*, May 1969, pp. 43–51 (and the resulting discussion in the December issue, pp. 47–52).

[85] O. M. P. Bilaniuk, V. K. Deshpande, and E. C. G. Sudarshan, "'Meta' Relativity," *American Journal of Physics*, October 1962, pp. 718–723.

There is, however, a curious twist to all this. Even if we grant that the RP may avoid causal paradoxes, the fact is that physics isn't fooled as easily as a human observer. That is, the receiver does actually lose energy upon the arrival of the tachyon, which is the opposite of what happens in a radio receiver when it receives a photon. In other words, the receiver must be in an elevated energy state prior to the tachyon's arrival; the receiver must be prepared beforehand to receive a message. If the receiver were instead sitting in its lowest energy state, then it could not accept (or *eject*, according to the RP) the tachyon. So, it's not surprising that, despite the enthusiastic embrace of the RP by some, other physicists took exception, with one arguing that the effectiveness of the RP in avoiding causal loops is "illusory" and "irrelevant,"[87] while others concluded that the causal paradoxes would actually preclude any possibility of tachyons interacting with ordinary matter in the first place (which is just a polite way of saying that tachyons have no more reality than do unicorns!).[88]

The RP's effect of flipping the roles of transmitter and receiver has attracted particular concern. Some analysts have pointed out that if one can modulate a superluminal signal to send a message into the past, then certainly one could *sign* the message. To quote a delightful example, "If Shakespeare types out *Hamlet* on his tachyon transmitter, Bacon receives the transmission at some earlier time. But no amount of reinterpretation will make Bacon the author of *Hamlet*. It is Shakespeare, not Bacon, who exercises control over the content of the message."[89] The last line of this quote is of central importance. The authors emphasize it by immediately observing that a signature is a *relativistic invariant* and that, indeed, it establishes a causal ordering quite independent of any temporal ordering. This example, alone, explains why one analyst said of the RP that it is "laughed to scorn,"[90] while another said of the RP that it "sounds merely like the endorsement of what can only be characterized as a fantastic delusion."[91]

---

[86] A. Italiano, "How to Recover Causality in General Relativity," *Hadronic Journal*, January 1986, pp. 9–12.

[87] R. G. Newton, "Particles That Travel Faster Than Light," *Science*, March 20, 1970, pp. 1569–1574.

[88] W. B. Rolnick, "Implications of Causality for Faster-Than-Light Matter," *Physical Review*, July 25, 1969, pp. 1105–1108, and D. J. Thouless, "Causality and Tachyons," *Nature*, November 1, 1969, p. 506.

[89] G. A. Benford, D. L. Book, and W. A. Newcomb, "The Tachyonic Antitelephone," *Physical Review D*, July 15, 1970, pp. 263–265.

[90] P. Fitzgerald, "On Retrocausality," *Philosophia*, October 1974, pp. 513–551.

[91] W. L. Craig, "Tachyons, Time Travel, and Divine Omniscience, *Journal of Philosophy*, March 1988, pp. 135–150.

Even more sophisticated scenarios than the *Hamlet* one (note 89) were devised to show how problems with FTL signals could arise that the RP could *not* resolve. The American physicist Bryce DeWitt (1923–2004) created such an example (the 'DeWitt Gambit') that involved a sequential, circular chain of tachyon signal transmissions between four observers, all moving in one spatial dimension. DeWitt showed out to arrange the spacetime geometry of the observers so that, at each stage, there is no dispute over who is sending and who is receiving, and so invoking the RP is avoided. Yet, when the signal reaches the first (last) observer, it is *before* he started (will start) the chain! This, of course, sets-up a potential bilking paradox: what if 'now' the first (last) observer decides to *not* send the chain's initiating signal? An even more sophisticated variation on the Gambit, involving four observers moving in *two* spatial dimensions, was soon after put forth by the English physicist Felix Pirani (1928–2015). As he concluded, "It is difficult to see how in the face of this example a classical-particle description of tachyons can be sustained."[92] Confronted by such sharp criticism, from so many, it is understandable why, just before his death, Feinberg co-authored (with two philosophers) a paper in which he seemed to be abandoning his support for tachyons as possible carriers of information backward in time.[93]

Many physicists today reject the possibility of backward-in-time messages, not because of concerns about the RP, but because such messages could create potential bilking paradoxes. To see how this works, the old Wheeler and Feynman idea of explaining (away) bilking paradoxes—that no signal in nature is *really* discontinuous—was examined by one physicist in the context of tachyons.[94] There we are asked to consider the following situation: A human (call him A) has a lamp on a table before him. The lamp is controlled by a tachyon receiver; in other words, the lamp illuminated only when a tachyon signal (a pulse, let us say) is detected. At 3 o'clock A will send a tachyon signal to B (a tachyon echo-transmitter that immediately rebroadcasts everything it receives) if the lamp does not glow at 1 o'clock. Now, the spacetime geometry of A and B is arranged to be such that a signal sent by either A or B to the other travels 1 h backward in time. Thus, if A sends a signal at 3 o'clock, then B will receive it at 2 o'clock (and immediately echo), and the echo will arrive at A at 1 o'clock. The paradox, of course, is that A sends a signal only if the lamp does *not* glow—that is, only if A does *not* send the signal![95]

We are then reminded of the Wheeler and Feynman claim that every pulsed signal is actually continuous; this argument would include the illumination itself of the lamp. Therefore, the lamp is not just on or off, but potentially at any level of illumination between those two extremes. So there sits A, and at 1 o'clock the lamp seems to glow dimly. To that, says Schulman, "A thinks it over, vacillating, finally sending a slightly late signal which isn't full strength." Then the echo isn't full strength either, which accounts for the original dim glow. This conclusion *is* consistent (but what if A's sending device is a toggle switch that *snaps* one way or the other—why then only a partial strength signal?), but it does seem to ask for a lot of supposing. Schulman himself is not so sure about the validity of universal continuity, writing at the end that "it is not clear that the Wheeler-Feynman assumption . . . ought to be made." (For more on the Wheeler-Feynman continuity assumption, in a different, non-time travel context, see the final "For Further Discussion" for this chapter.)

---

[92] F. A. E. Pirani, "Noncausal Behavior of Classical Tachyons," *Physical Review D*, June 15, 1970, pp. 3224–3225.

[93] G. Feinberg, D. Albert, and S. Levine, "Knowledge of the Past and Future," *Journal of Philosophy*, December 1992, pp. 607–642.

[94] L. S. Schulman, "Tachyon Paradoxes," *American Journal of Physics*, May 1971, pp. 481–484.

[95] A study of similar situations can be found in L. L. Gatlin, "Time-Reversed Information Transmission," *International Journal of Theoretical Physics*, January 1980, pp. 25–29.

The Wheeler-Feynman continuity idea *is* ingenious, allowing one to find a logically and physically consistent solution in time travel scenarios that, at least at first glance, might seem to have no solution. Consider, for example, the following situation[96]: "We have a camera ready to take a black and white picture of whatever comes out of [a] time machine. The film is then developed and the developed negative is subsequently put in the time machine and set to come out of the time machine at the time the picture is taken. This surely will create a paradox: the negative will have the opposite distribution of black, white, and shades of grey, from the picture that comes out of the time machine. But since the thing that comes out of the time machine is the negative itself we surely have a paradox." But do we?

The answer is no, because "What will happen is that a uniformly grey picture will emerge which produces a negative that has exactly the same uniform shade of grey. No matter what the sensitivity of the film is, as long as the dependence of the brightness of the negative depends in a *continuous* [my emphasis] manner on the brightness of the object being photographed, there will be a shade of grey that produces exactly the same shade of grey on the negative when photographed. This is the essence of Wheeler Feynman's idea."[97] (The conclusion is the same if we move from black-and-white to color photography.)

Nonetheless and despite this apparent success, the supposed ability of a modulated beam of tachyons to send a message into the past still raised concerns among many, particularly about free will and fatalism. Suppose, say those who are concerned about these issues, that you receive a tachyon message from yourself from tomorrow, informing you that a man you plan to kill tonight is still alive (tomorrow). Does that mean it is beyond your power to kill him tonight? According to one analyst (note 67) the answer is no; you *could* kill him—but if you do then the message from tomorrow would not have arrived. And because ignorance is not a precondition to free will, your newly acquired knowledge does not, by itself, suddenly limit your ability to kill the man. But, this line of arguing went, if you do not attempt to kill the man because you believe the message from your future self, then in fact the message *has* limited you!

---

[96] Taken from Frank Arntzenius and Time Maudlin, "Time Travel and Modern Physics," in *Time, Reality & Experience* (C. Callender, editor), Cambridge University Press 2002, pp. 169–200.

[97] More on the Wheeler-Feynman continuity idea, *and of its limitations*, can be found in D. Kutach, "Time Travel and Consistency Constraints," December 2003, pp. 1098–1113, and Phil Dowe, "Constraints on Data in Worlds with Closed Timelike Curves," December 2007, pp. 724–735, both in *Philosophy of Science*.

Fitzgerald's position was rebutted by Craig (note 91), who argues that it is not your ability to kill that is altered by the message but rather your motivation. Craig points out that such motivational changes can occur without invoking anything as radical as a message from the future. Suppose, he says, that just before you fire the fatal shot into your victim, you learn from him that he is your beloved, long-lost uncle. Clearly, your motivation for killing him is likely to be instantly altered, but equally clearly, your *ability* to kill him is unchanged. The mechanism for obtaining genealogical information, whether via time travel or as a last-minute appeal from your intended victim, is (says Craig) simply irrelevant. Not all buy into that, however, with one unconvinced analyst (note 94) writing that "history is a set of world lines essentially frozen into spacetime. While subjectively we feel strongly that our actions are determined only by our backward light cone, this may not always be the case." That is, Schulman appears open to the possibility that influences originating in the future might indeed have an impact on the present.

With the fading from the physics scene of enthusiasm for tachyons, the romance of communicating with the past using superluminal speeds passed from speedy particles to quantum mechanics via a mathematical result called *Bell's theorem*. John Bell, a physicist mentioned in the last chapter in connection with the MWI of quantum mechanics, published his theorem in a little article in an obscure, now defunct journal. Since then it has become one of the most cited physics papers from the 1980s.[98] The paradox cited in Bell's title refers to a famous 1935 paper in which Einstein (and two of his colleagues at the Institute for Advanced Study) challenged the conventional view of quantum mechanics, the view that there is no objective reality to anything unless it is observed.[99]

In fact, the possibility of quantum mechanics supporting FTL signals had been considered (and rejected) *before* Einstein's paper, by the Italian physicist Enrico Fermi.[100] Fermi concluded that, in a two atom system, the decay from an excited state of one atom (with the emission of a photon) would not influence the other atom before a time lapse of R/c, where R is the distance between the two atoms and c is the speed of light. In 1967, however, the Russian physicist M. I. Shirokov pointed out that Fermi's result was the result of an unjustified mathematical operation (he had replaced an integral from zero to infinity with one from minus infinity to infinity).

---

[98] J. S. Bell, "On the Einstein-Podolsky-Rosen Paradox," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987.

[99] A. Einstein, B. Podolsky, and N. Rosen, "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" *Physical Review*, May 15, 1935, pp. 777–780. See also N. D. Mermin, "Is the Moon There When Nobody Looks? Reality and Quantum Theory," *Physics Today*, April 1985, pp. 38–47.

[100] E. Fermi, "Quantum Theory of Radiation," *Reviews of Modern Physics*, January 1932, pp. 87–132.

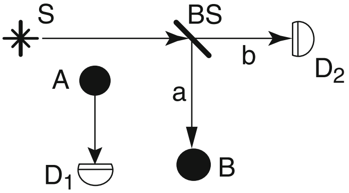

Fig. 5.5 Time travel and the wave function of quantum mechanics. In the above arrangement (reproduced by the kind permission of Francesco Gonella, from his paper "Time Machine, Self-Consistency and the Foundations of Quantum Mechanics," *Foundations of Physics Letters*, April 1994, pp. 161–166), S is a low-intensity source of photons, a source *so* weak that at any instant there is never more than one photon in the system. Each photon begins by traveling toward BS, a half-silvered beam-splitting mirror that, with equal probability, either passes a photon to the right along path b where it is detected by D_2, or downward along path a into mouth B of a wormhole time machine and out of which it emerges from wormhole mouth A *in the past*. Conventional quantum mechanics says that what happens at BS is determined *when the photon reaches BS*—that is, the probability wave function of the photon collapses at BS. But for a photon that is reflected into B, it exits A where it is detected by D_1 *before* the time of its arrival at BS. That means it is known (by detector D_1) what *will* happen at BS *before* the photon arrives at BS—so just *when did* the photon wave function collapse?

Then, nearly 30 years after Shirokov, the German physicist Gerhard Hegerfeldt proved (in the context of conventional quantum mechanics) a very general theorem that establishes a non-zero probability of the second atom responding to the first atom's decay *as soon as* the photon is emitted.[101]

Hegerfeldt's analysis is very general, but it does make one assumption—the non-negativity of energy density (the so-called *weak energy condition*)—which at the time was considered (in Hegerfeldt's words) to be "physically well motivated," but which today is *not* taken to be *a priori* obvious.[102] Einstein's paper, on the other hand, utilized physical assumptions not easily dismissed, and so its conclusions confounded physicists for decades.

The conventional view of quantum mechanics formulates physics in terms of probability wave functions that 'collapse' into specific realities only when measurements (observations) are made of the state of a system (which may be as elementary as a single particle). Until such measurements are made, says this view of quantum mechanics, a system has no specific state; instead, it merely has a probability distribution over a set of possible states (see Fig. 5.5 for a time travel puzzle concerning this claim). Einstein and his co-authors (note 99) strongly rejected this probabilistic interpretation of nature (recall Einstein's famous dictum, "God does not play dice with the cosmos.") Einstein and his colleagues agreed that quantum mechanics may be valid *as far as it goes*, but they also argued that it leaves out 'something' (as yet unknown) in describing reality. That is, they suggested that quantum mechanics is "incomplete" and that it incorporates "hidden variables."[103] They expressed this idea in the form of a paradox, the famous *EPR paradox*, which they posed as a thought experiment (a *Gedanken-experiment*) in which quantum mechanics declares that the properties of a particular spatially distributed system, when measured at point A, seem to be *forced* to assume specific values at point B (which may be arbitrarily distant from A) *without* there being a measurement at B.

---

[101] G. C. Hegerfeldt, "Causality Problems for Fermi's Two-Atom System," *Physical Review Letters*, January 1994, pp. 596–599. This mere *suggestion* of a possible failure of causality so stunned the editor of *Nature* that he felt compelled to quickly write a 'calming' reply: "Time Machines Still Over the Horizon," February 10, 1994, p. 509.

[102] The issue of the sign of energy density is very important in the analyses of wormhole time machines (see note 135 in Chap. 1), and we'll return to it in the next chapter.

Thus, said Einstein, there are just two possibilities. Either the system properties at B must have been what they are from the very start (even if the measurements at A had not been done) which is the view he held, or there must have been a linkage between the system at A and the system at B such that the wave function collapse at A is instantly transmitted to B to allow the wave function to collapse there as well. Because A and B may be arbitrarily far apart, this second view obviously requires an FTL transmission mechanism,[104] something Einstein called a "spooky action-at-a-distance," a term that eloquently expresses his low opinion of that idea! (Some translations replace *spooky* with *ghostly*, but the negative sentiment remains the same.[105]) For decades the debate between proponents of these two alternatives remained at a metaphysical, non-quantitative level. Then came Bell's paper in 1964.

Bell's theorem mathematically poses the choice between Einstein's hidden variables view and the conventional view of quantum mechanics through the use of an inequality involving certain measurable properties of a system.[106] If these measurements are such that the inequality is violated, then the conventional interpretation of quantum mechanics is vindicated, and Einstein's FTL spooky action-at-a-distance effect simply doesn't exist. Bell's great contribution, then, was to provide the means for removing the debate about quantum mechanics from metaphysics and to place it squarely in the realm of experimental physics.

"All" that needed to be done was to make the required measurements. These technically difficult experimental measurements were eventually performed by the French physicist Alain Aspect and his colleagues at the Institute of Applied Optics of the University of Paris, a decade and a half after Bell showed what had to be done. The results unequivocally support the conventional view of quantum mechanics.[107] Einstein was simply wrong, there are no hidden variables in quantum mechanics, and his spooky action does seem to exist. Does that mean we have, at last, experimental evidence of the possibility of information transfer at FTL speeds?

---

[103] For more on hidden variables, see E. P. Wigner, "On Hidden Variables and Quantum Mechanical Probabilities," *American Journal of Physics*, August 1970, pp. 1005–1009.

[104] For an analysis that argues against an FTL mechanism in quantum mechanics, see G. C. Ghirardi, *et al.*, "A General Argument Against Superluminal Transmission Through the Quantum Mechanical Measurement Process," *Lettere Al Nuovo Cimento*, March 8, 1980, pp. 293–298.

[105] You can find a discussion of the possibility of 'explaining' Einstein's "spooky actions" of quantum mechanics by invoking backward causation in R. I. Sutherland, "Bell's Theorem and Backwards-in-Time Causality," *International Journal of Theoretical Physics*, April 1983, pp. 377–384.

[106] The details are not important here, but a lovely exposition (for the lay person) can be found in Bell's essay "Bertlmann's Socks and the Nature of Reality," in *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press 1987. See also M. G. Alford, "Ghostly Action at a Distance: A Non-Technical Explanation of the Bell Inequality," *American Journal of Physics*, June 2016, pp. 448–457.

Well, maybe, but the issue is still hotly debated. The majority of physicists today, I suspect, are probably more perplexed over what Bell's theorem is saying than were over Einstein's original EPR paradox. In those early days one could agree with Einstein and his colleagues, who argued that quantum mechanics was valid *as far as it went*, but a deeper, more comprehensive theory *would* show the existence of hidden variables. However, because of the work by Bell and Aspect it is definitively known that quantum mechanics *as it stands* leads to correct results, results that can be checked in the laboratory. In other words, there is no need for hidden variables and FTL spooky actions cannot be ruled out.

And so, while tachyons and the tachyon antitelephone may be nothing more than a neat science fiction fantasy, just maybe a quantum mechanical Bell antitelephone can't be so dismissed. Indeed, the possibility of using a quantum mechanical FTL effect was once suggested in a letter written by a senior person at an unspecified California think tank (an organization such as, for example, the RAND Corporation) to the Under Secretary of Defense for Research and Engineering at the Pentagon. Here's what the Under Secretary read when he opened that letter: "If in fact we can control the FTL nonlocal effect, it would be possible . . . to make an untappable and unjammable command-control-communication [C^3] system at very high bit rates for use in the submarine fleet. The important point is that since there is no ordinary electromagnetic signal linking the encoder with the decoder in such a hypothetical system, there is nothing for the enemy to tap or jam. The enemy would have to have actual possession of the 'black box' decoder to intercept the message, whose reliability would not depend on separation from the encoder or on ocean or weather conditions."[108]

One can't help but wonder what might have been the Under Secretary's response to that incredible letter, and what sorts of ultra-mega-super-top-secret experiments it may have prompted. I would be willing to bet, if they did occur, that they failed. As one physicist put it, "Up to now nature has covered her tracks pretty well, blocking all possibilities for using the EPR effect for FTL communication."[109] Of course, the think tank letter, as 'farout' as it may initially appear, actually represents a *failure* of imagination, because the backward causation effect of EPR's spooky FTL effect is certainly a 'quantum jump' beyond a mere unjammable submarine C^3 system.

---

[107] A. Aspect, "Experimental Tests of Realistic Local Theories via Bell's Theorem," August 17, 1981, pp. 460–467, and A. Aspect, *et al.*, "Experimental Realization of Einstein-Podolsky-Rosen-Bohm *Gedanken-experiment*: A New Violation of Bell's Inequalities," July 12, 1982, pp. 91–94, and "Experimental Test of Bell's Inequalities using Time-Varying Analyzers," December 20, 1982, pp. 1804–1807, all in *Physical Review Letters*.

[108] See N. D. Mermin in note 99. For more on the enigmatic letter on the FTL submarine C^3 system, see Jack Sarfatti's letter to *Physics Today*, September 1987, pp. 118 and 120.

[109] J. Cramer, "Paradoxes and FTL Communication," *Analog Science Fiction*, September 1988.

5.7 For Further Discussion

Mark Twain, in his last, posthumously published novel *No. 44, The Mysterious Stranger*, incorporated reversed time as the work of a supernatural being. After the being reverses the world's time direction, the narrator tells the reader that "everywhere weary people were re-chattering previous conversations backward . . . where there was war, yesterday's battles were being refought, wrong-end first; the previously killed were getting killed again . . . we saw Henry I gathering together his split skull . . ." Read the novel, and comment on how well (or not) Mark Twain handled reversed time.

In Philip K. Dick's 1956 novel *The World Jones Made*, we read of a prophet who can see a year into the future. As he says, "To me *this is the past*," and then later we are told "He was a man with his eyes in the present [the world's future] and his body in the past [the world's present]." Read the novel, and then argue either for or against the suggestion that Dick was aware of the advanced wave solution in Maxwell's theory.

The physicist Robert Forward (see the sixth *For Future Discussion* in Chap. 3) argued that one way to send messages into the past is to compress a 15-billion-ton asteroid down to the volume of an atomic nucleus, spin it up, and then aim gamma ray bursts through the resulting near-by region of "unhinged time" (see "How To Build a Time Machine," *Omni*, May 1980). This is, of course, 'simply' an artificially constructed Kerr black hole telegraph transmitter (look back at note 114 and related discussion in Chap. 1). Forward, an optimist of the first rank, thought humans would be able to do this before the end of the twenty-first century. It would seem, then, that what should be done *now* is to build gamma ray frequency *receivers* (well within present-day technology) and listen for such messages from the future. The technical details of such receivers wouldn't matter, as long as their design is widely published. That way, the scientists of the future can learn those details by simply reading of them in old, musty library books and journals, and thus will be able to build their transmitters to be perfectly compatible with our old, musty (to them) receivers! Comment on the likelihood of National Science Foundation funding becoming available to build such receivers.

Suppose a time traveler goes into the future and, while there, discovers that there is an older version of himself living in the home that has been in his family for generations. Explain why this implies that the time traveler will eventually return to the present. Suppose, instead, that after the time traveler arrives in the future he decides to remain in the future, and *not* to return. Explain why this implies there will *not* be an older version of himself living in the family home. In both cases, assume the MWI does *not* apply, that is, assume that there is just a single time line.

One well-known quantum physicist, David Bohm (1917–1992), wrote the following passage in his book *The Special Theory of Relativity* (W. A. Benjamin 1965), concerning the possibility of sending messages into the past: "In effect, S could communicate with his own past [self, M] . . . and tell his past self [M] what his future is going to be. But on learning this M could decide to change his actions, so that his future . . . would be different from what his later self [S] said it was going to be. For example, the past self could do something that would make it impossible for the future one to send the signal. Thus, there would arise a logical self-contradiction." Do you think most physicists, writing today more than 50 years after Bohm, would repeat his words?

The role played by quantum mechanics in time travel studies is broad, deep, profound . . . and mysterious. What I mean by this is nicely illustrated by the final paragraph in a paper by Stephen Hawking ("Quantum Coherence and Closed Timelike Curves," *Physical Review D*, November 15, 1995, pp. 5681–5686). There he wrote "Personally, I do not believe that closed timelike curves will occur, at least on a macroscopic scale. I think that the chronology protection conjecture will hold and that divergences in the energy-momentum tensor will create singularities before closed timelike curves appear. However, if quantum gravitational effects somehow cut off these divergences, I am quite sure that quantum field theory on such a background will show loss of quantum coherence. *So even if people come back from the future, we will not be able to predict what they will do* [my emphasis]." What do you think Hawking meant by his final line? If "people come back from the future" and tell us what they did while in the

(continued)

future, then what's wrong with their memories that causes us to fail to be able to predict what they *will* do? Or, is there perhaps nothing wrong with their memories and Hawking is instead arguing that the future experienced by the returned time travelers will not be the future when *we* 'get there'? If that's the case, then what are the returned time travelers 'remembering'?

If quantum mechanics is actually slightly non-linear physics (as are many other normally linear physical phenomena at sufficiently high energy levels)—physics is *linear* when *superposition* holds, which means the result of two inputs is the sum of the individual outputs resulting from application of the individual inputs—and if one accepts the MWI concept, then at least two physicists (**PHYSICISTS1**) claim to have shown that one could communicate not just with our past, but also with the many pasts in the ever-splitting branches of the many worlds. Another physicist (**PHYSICIST2**) wrote a very funny illustration of what that might be like. Yet another physicist (**PHYSICIST3**) suggested that non-linear quantum mechanics might actually allow one to take *photographs* of the many-worlds. Of that, a physicist (**PHYSICIST4**) wrote (without any exaggeration) that such an achievement would be "perhaps the most amazing discovery in the history of science, indeed in the history of mankind." Or, to quote yet another physicist (**PHYSICIST5**), "interworld communication would lead to truly mind boggling possibilities," some of which have been incorporated in at least one science fiction novel (**SFAUTHOR**). Read the physicists' papers, and the novel, and then summarize with your own commentary.

PHYSICISTS1: N. Gisin, "Weinberg's Non-linear Quantum Mechanics and Superluminal Communication," *Physics Letters A*, January 1, 1990, pp. 1–2, and J. Polchinski, "Weinberg's Non-linear Quantum Mechanics and the Einstein-Podolsky-Rosen Paradox," *Physical Review Letters*, January 28, 1991, pp. 397–400.

PHYSICIST2: J. G. Cramer, "Quantum Telephones to Other Universes, to Times Past," *Analog*, October 1991.

PHYSICIST3: D. Albert, "How to Take a Photograph of Another Everett World," *Annals of the New York Academy of Sciences*, December 30, 1986, pp. 498–502.

PHYSICIST4: M. A. B. Whitaker, "On the Observability of 'Many Worlds'," *Journal of Physics A*, July 11, 1985, pp. 1831–1834.

PHYSICIST5: R. Plaga, "On a Possibility to Find Experimental Evidence for the Many-Worlds Interpretation of Quantum Mechanics," *Foundations of Physics*, April 1997, pp. 559–577.

SFAUTHOR: J. P. Hogan, *Paths to Otherwhere*, Baen 1996.

One upon a time, FTL tachyons (and their associated into-the-past transmissions of information) were strictly in the province of science fiction. In recent years, however, such doings have moved into mainstream fiction, as well, with the most recent (as I write) example being the man-of-action novel by Patrick Lee, *Signal*, St. Martin's Press 2015. (Tachyons are also mentioned in the 2015 film *Tomorrowland*.) That novel is set in modern times, not in the future; its hero is a retired soldier who now works as a self-employed house-flipper. His is definitely not a futuristic science fiction world. Until, that is, be becomes involved with a radio-like gadget that receives signals from 10 h, 24 min in the future. As the novel progresses we eventually learn that the gadget is based on German electronics technology that was being tested in a remote lab in northern Algeria, near the end of World War II. When that lab was overrun by a small American force the equipment was destroyed, but not before one of the Americans heard the gadget playing a song titled "She Loves You" along with the word *yeah* repeated numerous times. He didn't know what that meant until, 20 years later, he watched the Beatles' first American appearance on *The Ed Sullivan Show* TV program, and so suddenly realized that in 1944 he had heard a song that hadn't been written yet! That gadget is at the center of a modern-day, renewed Nazi effort to conquer the world, and there is much 'you've seen it all before' chasing, shooting and other 'James Bond' types of action in the novel, but the author *has* been quite inventive in treating time paradoxes. He does talk a lot about 'changing the future,' worries confusingly about 'changing the past,' and mistakes neutrinos for tachyons, but, still, if one is willing to overlook such issues it *is* an entertaining read. In particular, while the gadget's inherent range to the future is limited to ten-plus hours, the author describes a clever way to arbitrarily extend that value. Read *Signal*, and then describe and critique the method outlined in the novel.

The Wheeler-Feynman assumption of continuity already had a distinguished history in mathematics long before they invoked it in their physics resolution of bilking paradoxes. (Both men were very good mathematicians, and certainly knew what I am about to tell you here.) Imagine a man who is about to walk up a hill, starting at A (the base of the hill) and ending at B (the top of the hill). You know nothing of *how* he walks (perhaps at times he stops for a while, other times he walks slowly, sometimes he walks briskly, perhaps at times he even walks back *down* the hill). All you know is that, starting from A at 10 o'clock in the morning, he arrives at B at 11 o'clock. That is, the walk up

(continued)

the hill takes exactly 60 min. He camps overnight at B and then, the next morning, at exactly 10 o'clock, he walks back down the hill along the same path he followed during his ascent the previous day. He arrives at A at 11 o'clock. That is, the return trip takes exactly 60 min. Again, you know nothing of the details of how he makes the descent. Prove that there is at least one spot on the path that he passes at exactly the same time during his descent as he passed it during his ascent the previous day. Hint: No complicated equations are required. Indeed, no math at all is needed. Just sketch the appropriate, general graphs of the man's ascent versus time, and of his descent versus time, and invoke continuity. That is, sketch the distance (as measured along the path) that he is from A versus time for both the ascent and the descent.

Chapter 6
The Physics of Time Travel: II

> *"It is very sad to see valuable minds writing such a pile of unmitigated bullshit."*
>
> —not all physicists think time travel is worthy of study.[1]

6.1 Faster-than-Light into the Past

> "Faster-than-light travel remains a coherent, and possible concept, even though it is forbidden by relativity theory."
>
> —a philosopher makes a physics mistake.[2]

So far we have limited our consideration of relativity theory to speeds below the speed of light—that is, to the condition $v < c$, where v is the relative velocity of two reference frames. There was nothing, however, in the derivation of the Lorentz transformation equations discussed in Chap. 3 that actually used that self-imposed constraint. So, just what, in fact, *does* happen for $v > c$? This is not an empty question, because the second half of the above quote that opens this section is simply not true. That isn't to say we can have FTL for free; there *is* a high price to pay, that of causality violation (although, if you are a fan of time travel, it's a price you are probably happy to pay). If a material object goes FTL, then the mathematics seems to say that the object could travel into the past, just as the caped crusader does in the first (1978) *Superman* movie, in order to change the past (to save Lois Lane from dying in an earthquake). In addition, the mathematics also seems to say that if a signal bearing information could achieve FTL, then that information, too, would travel back into the past (see Fig. 6.1).

---

[1] This rather blunt comment (reported in *Physics World*, December 2009, p. 3) was prompted by a suggestion, from two other physicists, that the Higgs boson might ripple backward through time and thereby stop CERN's Large Hadron Collider from creating the long-sought particle in the first place.

[2] G. Robinson, "Hypertravel," *Listener*, December 17, 1964, pp. 976–977, the printed version of a lecture to the British Association for the Advancement of Science.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6_6

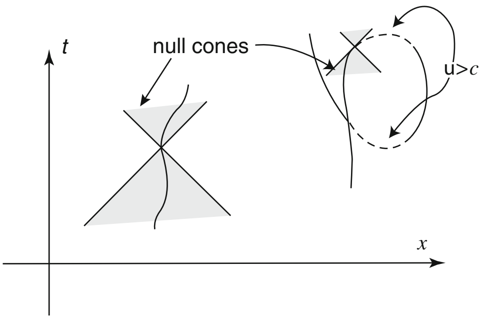

Fig. 6.1 A math-free proof that there can be no closed loops in time (that is, no time travel to the past if $v >$ c is forbidden) in flat Minkowski spacetime, the spacetime of the special theory of relativity. Such a conclusion is far less clear in the curved spacetimes of the general theory of relativity. Figure reproduced by the kind permission of Serguei V. Krasnikov (Polkovo Central Astronomical Observatory in St. Petersburg, Russia), from his 2003 paper "Time Machine (1988–2001)"

The original thinking along these lines visualized such an FTL signal as a modulated beam of tachyons, as mentioned in the previous chapter. In addition to tachyons not having been experimentally observed, even after intense searches for them, there are several theoretical objections (in addition to the bilking paradox problem discussed in the previous chapter) to the likelihood such FTL particles exist.[3] For example, the relativistic expressions for the energy and momentum of a particle with rest mass m_0 moving with speed v are, respectively,

$$E = \frac{m_o c^2}{\sqrt{1-(v/c)^2}} \qquad \text{and} \qquad p = \frac{m_o v}{\sqrt{1-(v/c)^2}}.$$

For $v > c$ the radicals in these expressions become imaginary, whereas E and p must always be real-valued (because they can be observed and measured as a result of the interactions the particle has with other matter). The energy and momentum can regain the property of being real-valued if we write $m_0 = \mu\sqrt{-1}$ (that is, $m_0^2 = -\mu^2$) for a tachyon, where μ is the positive, real-valued (but unobservable) meta-mass (that is, $m_0^2 < 0$). This is a radical proposal, of course, (as it is $m_0^2 > 0$ that we are used to; as a Russian mathematician romantically put it, "What binds us to space-time is our [positive] rest mass, which prevents us from flying at the speed of light, when time stops and space loses meaning. In a world of light there are neither points nor moments of time; beings woven from light would live 'nowhere' and 'nowhen'; only poetry and mathematics are capable of speaking meaningfully about such things."[4].

---

[3] See, for example, "More About Tachyons," *Physics Today*, December 1969, pp. 47–52, a collection of letters received by the journal from its readers. The 'DeWitt Gambit,' mentioned in the previous chapter, was first proposed in one of those letters.

Well, perhaps, but let's continue to pursue the physics of tachyons, at least for a while. It can be shown[5] that the energy and momentum of the tachyon are given by

$$E = \frac{\mu c^2}{\sqrt{(v/c)^2 - 1}} \quad \text{and} \quad p = \frac{\mu v}{\sqrt{(v/c)^2 - 1}}.$$

An interesting consequence of this is that if tachyons *lose* energy, they *speed up*, an observation first made by the German physicist Arnold Sommerfeld (1868–1951) in 1904.[6] This means that if there is a mechanism for continuous energy loss, such as Cerenkov radiation,[7] then tachyons will spontaneously accelerate without limit to *infinite speed*! Curiously, while the above expressions for E and p show that, as $v \to \infty$, tachyons would possess *zero* energy, they would nonetheless have a *non-zero* momentum of μc.[8]

To see how backward time travel and FTL are connected, it is useful to establish a geometrical interpretation of the Lorentz transformation. As stated in Chap. 3, if the x', t' system is moving with speed v in the x (or x') direction relative to the x, t system, then

$$x' = \frac{x - vt}{\sqrt{1 - (v/c)^2}} \quad \text{and} \quad t' = \frac{t - vx/c^2}{\sqrt{1 - (v/c)^2}}.$$

---

[4] Yu I. Manin, *Mathematics and Physics*, Birkhäuser 1981, p. 84.

[5] L. Parker, "Faster-Than-Light Inertial Frames and Tachyons," *Physical Review*, December 25, 1969, pp. 2287–2292.

[6] The consideration of FTL particles already had a long history *before* tachyons were specifically named. You can find late nineteenth century (1888–1889) theoretical analyses of electrically charged FTL particles in the writings, for example, of the English mathematical electrical engineer Oliver Heaviside (1850–1925): see my biography *Oliver Heaviside: the life, work, and times of an electrical genius of the Victorian Age*, The Johns Hopkins University Press 2002, pp. 124–126.

[7] Cerenkov radiation is the energy radiated when a charged particle exceeds the speed of light *in the medium through which it travels*. Since the speed of light *in water* is less than c, it is perfectly okay with special relativity to exceed the speed of light in water, and in fact this commonly occurs for the energetic electrons produced by submerged atomic reactors (swimming pool reactors). The resulting radiation is observed as a blue glow. The radiation is named after the Russian physicist Pavel Cerenkov (1904–1990)—for which he received a share of the 1958 Nobel physics prize—but in fact Heaviside (previous note) had predicted it more than a decade before Cerenkov was born.

[8] This (theoretical) property of tachyons (if they exist) could (perhaps) be used (maybe) to build a revolutionary (to say the least) new rocket propulsion system: see J. Cramer, "The Tachyon Drive: Infinite Exhaust Velocity at Zero Energy Cost," *Analog*, October 1993.

These equations make sense for the case of $v < c$, and we will retain this condition for our two relatively moving *frames of reference*: these frames are the *worlds* of two *human* observers, observers that we'll take to always be *sub*luminal. As one physicist so nicely put it, "The assumption that observers move faster-than-light goes beyond superluminal signaling,"[9] as such observers would have to be thought of as being built not out of flesh-and-blood, but rather out of tachyons.[10] We'll reserve the symbol w to denote the speed of an FTL particle.

Now, recall what we mean by any line parallel to the x-axis; it is a line with a fixed time coordinate. Such a line is a *cosmic moment* line, with the equation t = constant. Similarly, for the moving system we would write the equation of a cosmic moment line as t' = constant which, after the Lorentz transformation is applied, is equivalent to

$$t - \frac{vx}{c^2} = \text{constant.}$$

In particular, the x'-axis ($t' = 0$ cosmic moment line), which passes through the point $x = 0$ at $t = 0$, has the equation

$$t = \frac{vx}{c^2} = vx$$

with the usual convention of $c = 1$.

In a similar way, recall what we mean by any line parallel to the t-axis; it is a line with a fixed space coordinate. Such a line is the world line of a stationary particle in the x, t frame, with the equation x = constant. Similarly, for the moving system we would write x' = constant as the equation of the world line of a particle stationary in that system. From the Lorentz transformation, this is equivalent to

$$x - vt = \text{constant.}$$

In particular, the t'-axis (which is the $x' = 0$ world line of a particle stationary at the origin of the moving system) passes through the $x = 0$, $t = 0$ point, and it has the equation

$$x = vt.$$

Thus, superimposed spacetime coordinate axes for the two frames look like those shown in Fig. 6.2. That is, the relative motion of the two frames results in a

---

[9] K. Svozil, "Time Paradoxes Reviewed," *Physics Letters A*, April 3, 1995, pp. 323–326.

[10] Science fiction, however, *has* (since the early days of pulp) enthusiastically embraced FTL human travel. In Larry Niven's story "At the Core" (*If*, November 1966), for example, we read of a manned spacecraft that travels 60,000 light years to the center of our own Milky Way Galaxy, and then back, at a speed 420,000 times that of light. As the pilot says (in a grand understatement), "That's goddam fast." Perhaps there is a way to make *some* sense of such an adventure, with the so-called *warp drive* (a'la *Star Trek*), which we'll take-up briefly at the end of this chapter.

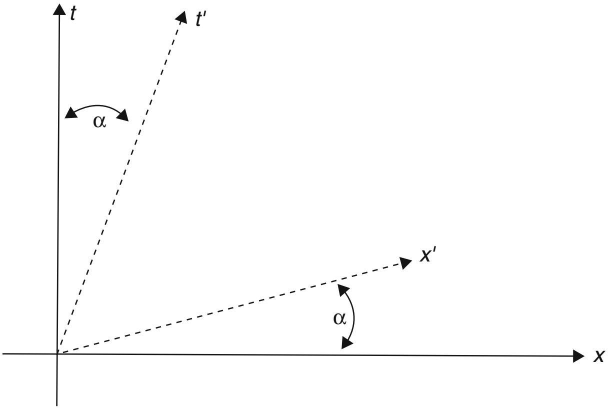

Fig. 6.2 Spacetime coordinate axes rotation by relative motion

rotation of the spacetime axes; but it is a strange sort of rotation, with *opposite* senses for the space and time axes. That is, the x' and t' axes rotate *towards* each other, as shown in the figure, to make equal angles (α) with the x and t axes, respectively, where

$$\alpha = tan^{-1}(v).$$

If we limit the moving frame (the frame of a moving observer) to subluminal speeds ($0 \leq v < 1$), then

$$0 \leq \alpha < 45°.$$

At the speed of light ($v = 1$) $\alpha = 45°$ and so the x' and t' axes coincide—time and space have become indistinguishable.

It is important to realize that observers in either system would measure the same speed for a photon; that is, each would see the world line of a photon as a line with slope 1. This view of the world line of a photon is literally built into the Lorentz transformation because one of Einstein's fundamental postulates for special relativity is the invariance of the speed of light. The truth of this statement for the x, t system is obvious from Fig. 6.2. It is, perhaps, not so obvious with the x', t' system because of the non-perpendicular axes (as shown in the figure) for that system. In Fig. 6.3 the world line of a photon is shown in both systems. In that figure we emit the photon at $x' = 0, t' = 0$, and we later measure its coordinates at point A to be $x' = x'_A$ at time $t' = t'_A$. Note carefully how this is done. We draw lines from point A parallel to the x' and t' axes until they intersect the t' and x' axes, respectively. This is similar to the way we would get the spacetime coordinates of A in the more familiar x, t system, where we would draw lines parallel to the x and t axes.

It should now be obvious that x'_A and t'_A have the same extension, just as they do in the unprimed system, so

Fig. 6.3 Invariance of the world line of a photon

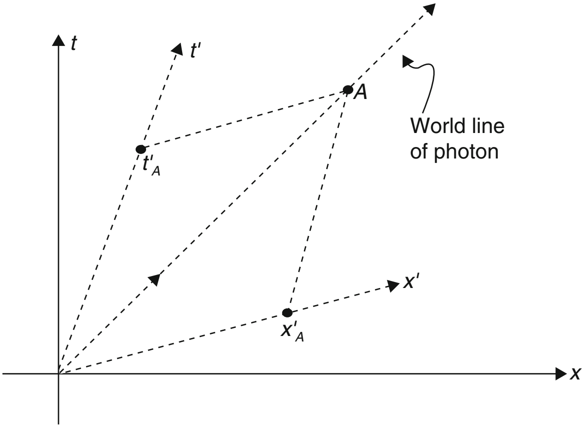

$$\frac{x'_A}{t'_A} = \text{the speed of light}$$

The speed of light is the only invariant speed under the Lorentz transformation. Indeed, the modern approach to special relativity emphasizes this invariance as the central property of the speed of light, rather than the idea of the speed of light being a limiting speed.

This geometrical interpretation of the Lorentz transformation lets us quickly make another interesting (one, I think, not at all obvious) observation: If a particle is faster than light in the x, t system, then there exists a subluminal x', t' system in which the particle is *infinitely* fast. Figure 6.4 shows the world line of an FTL particle in the x, t system (which is, of course, *below* the world line of a photon; that is, the particle's world line is spacelike). Suppose the FTL particle has speed $w > \text{c}$ such that its world line makes angle β with the x-axis. If we now pick v, the speed of the moving x', t' system to be such that $\alpha = \beta$, then the x'-axis will coincide with the world line of the particle, and so the particle will appear to an observer in the x', t' system to be everywhere at once—that is, to be infinitely fast. We have, then,

$$\beta = tan^{-1}(v) = tan^{-1}\left(\frac{1}{w}\right)$$

or $v = 1/w$, which seems to be dimensionally wrong. Recall, however, that with our convention of $c = 1$, the v in this result is a *normalized* speed. To return to the units of everyday use, simply replace v with v/c and w with w/c; this transforms our result to

$$\frac{v}{c} = \frac{c}{w} \quad \text{or} \quad v = \frac{c^2}{w}\cdot$$

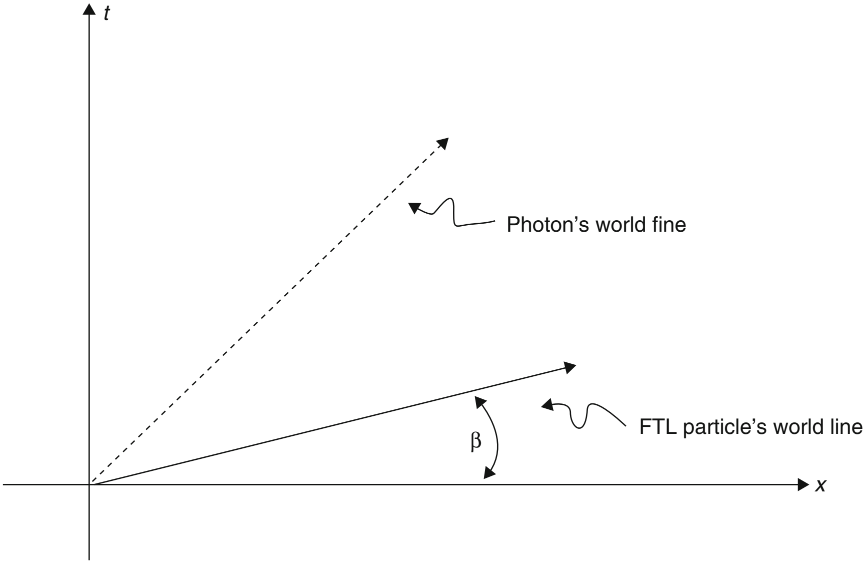

Fig. 6.4 World line of an FTL particle

We can, of course, turn this result around. If an FTL particle moves with speed w in the x, t frame, then to an observer in the x', t' frame moving with subluminal speed v, the particle will appear to be infinitely fast if $w = c^2/v$. A particle with $w > c^2/v$ is said to be not just be superluminal, but *ultra*luminal.

If a particle has infinite speed if $w = c^2/v$, then what physically happens if w is greater than c^2/v? The answer is easy to see from a spacetime diagram, as in Fig. 6.5, where the x' and t' axes have been extended back to negative values. In that figure I have labeled two arbitrary events A and B on the world line of an ultraluminal particle (and so it lies below the x' axis), and have shown the spacetime coordinates of each event in both the x, t and x', t' frames. For the x, t frame we see that A is related to B by the relations $x'_A < x'_B$ and $t'_B < t'_A$; that is, the time order of A and B is reversed for an observer in the x', t' frame. To that observer, the particle appears to be traveling backward in time!

But this isn't quite the end of the story. Following the approach of two pioneering tachyon physicists,[11] we note that if the energy of a particle in the stationary system is E, then the energy as measured in the moving system is given by[12]

---

[11] O. M. Bilaniuk and E. C. G. Sudarshan, "Causality and Space-like Signals," *Nature*, July 26, 1969, pp. 386–387.

[12] The expression for E' is the result of applying the Lorentz transformation to E. You can find all the details of that worked out in A. P. French, *Special Relativity*, W. W. Norton 1968, pp. 208–210.

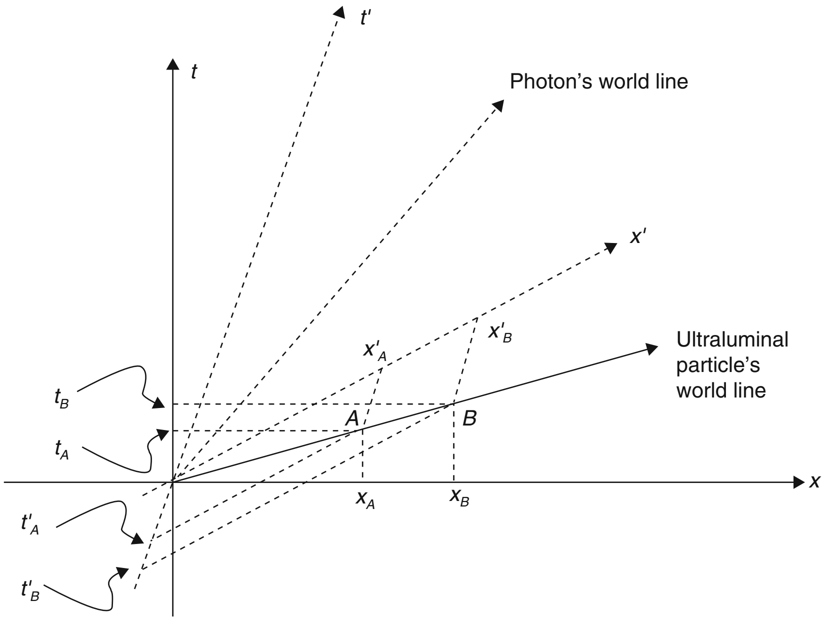

Fig. 6.5 World line of an *ultra*luminal particle

$$E' = \frac{1}{\sqrt{1-\left(\frac{v}{c}\right)^2}}\frac{\mu(c^2 - \text{wv})}{\sqrt{(w/c)^2 - 1}}$$

Note that sign of E' switches from positive to negative when w (the speed of the particle) exceeds c^2/v, which is precisely the condition for the particle to be ultraluminal and so traveling backward in time (as seen in the primed system). That is, negative energy moving backward in time in one system is positive energy moving forward in time in another. This is, in fact, the original motivation for the RP (reinterpretation principle) discussed in the previous chapter, and the claim was, at one time, that the RP was just what was needed to 'explain' the paradoxical implications of time travel. This is not the majority view today, however, and a number of physicists have been quite inventive in constructing scenarios with causal paradoxes that the RP clearly fails to 'explain.'

For example, the Princeton physicist Shoichi Yoshikawa (1934–2010), in a letter to *Physics Today* (see note 3), was able to create a scenario which *uses* the RP to arrive at a causal paradox. In his construction, the RP allows an observer to transmit an ultraluminal tachyon to a remote observer at time $t = 0$, and to receive a reply from that observer at $t < 0$. This obviously sets-up the possibility of a bilking paradox in which the original observer, upon receiving the $t < 0$ tachyon, then decides *not* to transmit the $t = 0$ tachyon. What makes Yoshikawa's paradox a particularly troublesome paradox is that the RP is the culprit, not the savior.

6.2 Tipler's Rotating Cylinder Time Machine

> "… within forty-eight hours we had invented, designed, and assembled a chronomobile. I won't weary you with the details, save to remark that it operated by transposing the seventh and eleventh dimensions in a hole in space, thus creating an inverse ether-vortex and standing the space-time continuum on its head."[13]
>
> —this is almost surely *not* the way to build a TM (time machine)

As discussed at the start of this book, the first endorsement of the reasonableness of physicists talking about plausible, scientific time travel, began with Gödel's discovery of closed timelike lines in the mathematics of certain rotating universe models. Such models had been studied as early as 1924 by the Hungarian physicist Cornelius Lanczos (1893–1974), a quarter century before Gödel, but it was Gödel who made explicit the possibility a rotating universe might allow time travel. His realization of time travel as an *inherent* property of a rotating universe is an illustration of a *weak* TM, while what is the central interest concerning time machines (in physics *and* in science fiction) is a *strong* TM. That is, in time machines that can be *intentionally constructed* by manipulating mass-energy in a finite (what physicists call a *compact*) region of spacetime to *create* closed timelike lines where none existed before the manipulation began. Interestingly, the fundamental physical idea behind Gödel's weak TM is the same underlying idea behind the first strong TM, Tipler's rotating cylinder discussed back in Chap. 1 (strictly, of course, only a *compact* TM if the cylinder can be of *finite* length).

The one result from general relativity that we'll use here (without proof) is that the rotation of matter causes a distortion of spacetime that results in the 'tipping over' of light cones, with the future half tilted in the direction of motion. If you imagine a point in the universe about which the rotation takes place, then this tipping effect increases with the radial distance from that point.[14] The fact that rotating masses tip light cones over in the direction of rotation was discovered very early in the history of general relativity (1918), by the Austrian theoreticians Josef Lense (1890–1985) and Hans Thirring (1888–1976). Originally (and naturally) called the *Lense-Thirring effect*, it now generally goes by the name of the *dragging of inertial frames effect*, and it plays a central role in the weak Gödel and the strong (maybe) Tipler time machines. Here's how.

At a certain critical distance from the rotation center (more on this in just a bit), the future half of the light cone at a given point in spacetime will be sufficiently tilted so as to enter the past half of similarly tilted light cones at nearby spacetime points This is illustrated in Fig. 6.6,[15] which shows a circular chain of tilted light

---

[13] L. Sprague de Camp, "Some Curious Effects of Time Travel," in *Analog Readers' Choice*, Dial 1981.

[14] For a picture of this, see S. W. Hawking and G. F. R. Ellis, *The Large Scale Structure of Spacetime*, Cambridge University Press, 1973, p. 169.

[15] Adapted from D. B. Malament, "'Time Travel' in the Gödel Universe," *Proceedings of the Philosophy of Science Association* 1984, pp. 91–100.

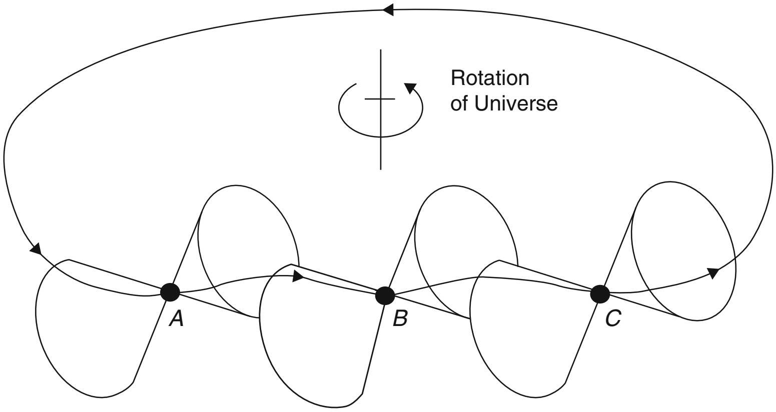

Fig. 6.6 Time traveling through tilted light cones in a rotating universe

cones in a rotating universe. Because light cones are tilted by a rotation-induced twisting of spacetime, a traveler can move around a circular path on a trip that is always directed into his *local* future but nonetheless end-up in his own *global* past, without ever going faster than light. This kind of round trip in spacetime, with a trajectory that winds back into the past without ever becoming spacelike, is a closed timelike curve.

From Fig. 6.6 it should be clear how a time traveler, beginning at A, can weave his way along a circular path (this path needs to have a radius at least as great as the critical value mentioned earlier) to B to C to . . . that brings him back into the *past* half of the light cone at A. The traveler's world line is always inside the local light cone; that is, the world line of the time traveler is always timelike, and never FTL. These timelike curves are present in the rotating spacetime from the very beginning of the spacetime, and were not *created* (certainly not by humans) by conscious intent.

You can see how this works mathematically by taking the spacetime metric for Gödel's universe (with, as usual, the convention that the speed of light $c = 1$):

$$(ds)^2 = (dt)^2 - (dr)^2 - (dy)^2 + \sin h^2(r)\left[\sin h^2(r) - 1\right](d\phi)^2 \\ + \sqrt{2}\sin h^2(r)(d\phi)(dt)$$

where t, r, y, and ϕ are cylindrical coordinates in four-dimensional spacetime. Now, imagine that our adventurer's world line is the helical curve r=constant, $y = 0$, and $t = -\alpha\phi$: if we take the time axis as vertical then the time traveler's world line is a vertical helix in spacetime. For this curve, $dr = \mathrm{dy} = 0$ and $dt = -\alpha d\phi$. This last differential means, in particular, that whatever the sign of the constant α, we can choose that one of the two senses of movement in the spatial ϕ dimension that gives $dt < 0$.

Continuing, we have

$$(ds)^2 = \left[\alpha^2 - 2\sqrt{2}\alpha sinh^2(r) + \sin h^2(r)\left(\sin h^2(r) - 1\right)\right](d\phi)^2$$

or, upon letting $u = \sinh(r)$ we have

$$(ds)^2 = \left[\alpha^2 - 2\sqrt{2}\alpha u^2 + u^2\left(u^2 - 1\right)\right](d\phi)^2.$$

Now, for $\alpha = 0$ we have

$$(ds)^2 = u^2\left(u^2 - 1\right)(d\phi)^2$$

which is greater than zero if $u > 1$. This condition holds if $\sinh(r) > 1$, that is, if r is a constant greater than $\ln\left(1 + \sqrt{2}\right)$. In other words, for r sufficiently large (and now we know the critical value for r) we have $(ds)^2 > 0$, the required condition for a timelike spacetime interval. By continuity, then, we will continue to have $(ds)^2 > 0$ even with some small positive or negative value of α different from zero. Because ϕ is a periodic coordinate (we identify $\phi = 0$ with $\phi = 2\pi$), as the traveler moves along the curve she returns repeatedly to the same spatial points, but her time coordinate is increasingly negative. That is, she is traveling into the past. Note, once again, that in Gödel's universe this property holds only for orbits with radii greater than a certain minimum.[16]

Tipler's rotating, infinitely long cylinder is a mechanism for artificially producing the tipped-over light cone effect, thus *creating* closed timelike curves. Figure 6.7, taken from Tipler's Ph.D. dissertation,[17] shows how the cylinder works. The cylinder is represented by the central vertical axis. Far away from the cylinder the light cones in spacetime are upright, but as we move inward they tip over, the future halves opening up into the direction of rotation. (Only the future halves of the light cones are shown.) This direction, which is the direction that far away from the cylinder measures *space*, near the cylinder measures *time* (just as in the Gödelian universe). That is, there has been a dimension reversal! This fantastic possibility has found its way into science fiction, as in Stephen Baxter's 1995 novel *The Time Ships*, wherein the Victorian narrator says "If only one could *twist about* the Four Dimensions of Space and Time — transposing Length with Duration, say — then one could stroll through the corridors of History as easily as taking a cab in the West End!"

To travel back in time, therefore, all the time traveler need do is leave Earth and approach the cylinder until she is near enough to be in the tipped-over region of spacetime. Then she would follow a helical path around the cylinder and could

---

[16] There are other solutions to the Einstein gravitational field equations that have closed timelike lines at *any* radius, no matter how small: see M. J. Reboucas, "A Rotating Universe with Violation of Causality," *Physics Letters A*, March 5, 1979, pp. 161–163.

[17] F. J. Tipler, *Causality Violation in General Relativity*, University of Maryland 1976.

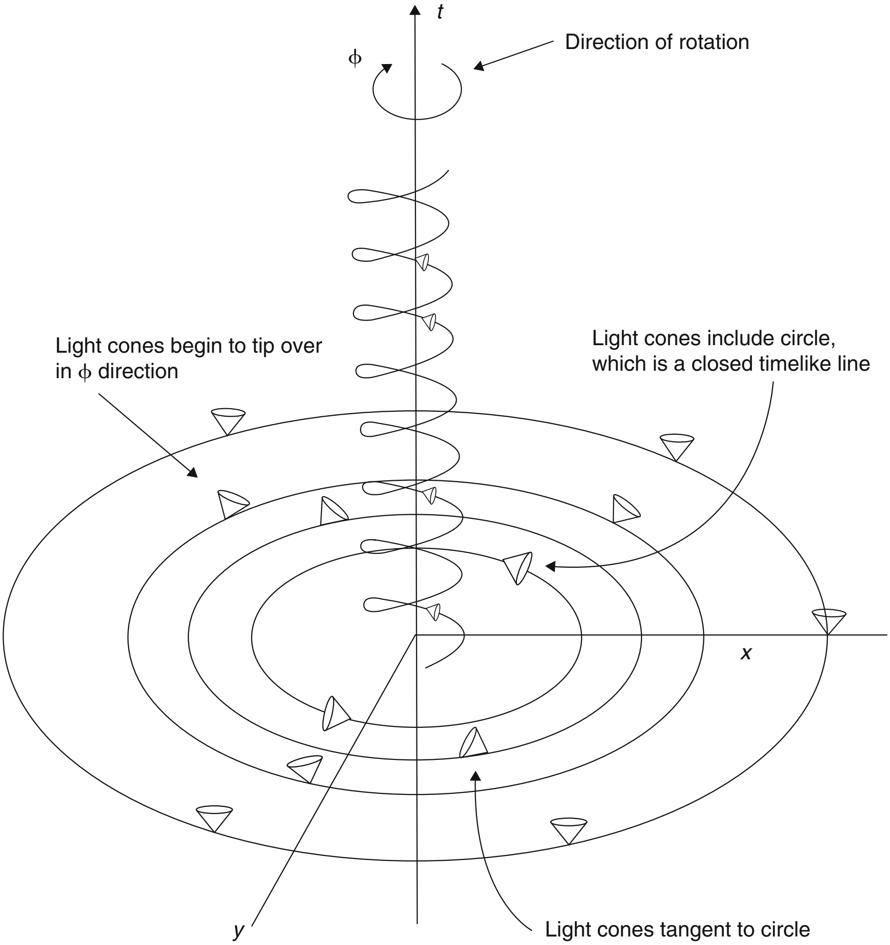
t
ϕ
Direction of rotation
Light cones begin to tip over in ϕ direction
Light cones include circle, which is a closed timelike line
x
y
Light cones tangent to circle

Fig. 6.7 The future halves of light cones point almost entirely in the $+t$ direction far from rotating matter; they begin to tip over as the matter is approached. Note that there is a helical timelike path that moves locally into the *future* in the $-t$ direction, that is, it goes into the past as seen by an observer far from the rotating matter (The world lines of rotating matter are helixes in the $+t$ direction)

spiral along the negative time direction as far back in time as desired (but no further back than to the moment of the cylinder's creation). This motion is such that the time traveler is always moving into her local future, via the tipped-over light cones. Finally, she would withdraw from the cylinder and return to Earth—in the past. The time traveler had better be a good space navigator, of course, because Earth won't be where she left it!

6.3 Thorne's Wormhole Time Machine

> "This fact reinforces the authors' feeling that [closed time loops] are not so nasty as people generally have assumed."[18]

The spacetime wormhole is presently the most promising of the approaches that have been advanced for building a time machine. Gödel rotated the entire universe in 1949, and Tipler 'reduced' the problem in 1974 to 'merely' spinning a cylinder of infinite length. In 1988 Kip Thorne scaled things down even more, this time to other extreme. His idea calls for pulling a wormhole on the scale of the Planck length out of the topologically multiply connected quantum foam that spacetime is and then enlarging it (somehow) to human scale, all the while stabilizing it against self-collapse, and then finally using the time dilation effect of special relativity to alter time at one mouth of the wormhole as compared to the other mouth. What a mouthful! What, you almost surely wonder, does all that mean?

First of all, what's 'quantum foam'? The term refers to the idea that the topology of spacetime is *not* a smooth, continuous manifold, but rather (if you look close enough) is a seething 'ocean of fluctuations' that is always changing, changes on the scale of the Planck length (look back at Sect. 1.5). What does *that* mean? Like the ocean surface-in-the-large, large-scale spacetime is simply connected. But just as one sees all sorts of transient structure as one looks at the water more closely (beginning with macroscopic waves and then proceeding downward to the bubbly foam on the waves), spacetime too displays an ever-changing connectivity-in-the-small.[19] That is the 'quantum foam.'

Wormholes have been around in physics for decades, but they have always been thought to be so unstable as to exist only on paper in the mathematics of general relativity. In an analysis[20] published more than half-a-century ago, for example, wormhole instability was shown to be so severe that not only would a human have no chance in getting through one, but also not even a single speedy photon could do so. Even at the speed of light, the photon could not zip through a wormhole before being trapped inside ("pinched off") in a region of infinite spacetime curvature. Wormholes would simply collapse too quickly after formation for even the so-called ultimate speed to save the traveler. Indeed, the presence of mass-energy inside a wormhole actually *accelerates* its collapse. The physics of wormholes, it seemed, made them simply untraversable.

---

[18] The conclusion of a mathematical demonstration that time travel by wormhole does not conflict with the conservation of energy: see J. L. Friedman *et al.*, "Cauchy Problem in Spacetimes with Closed Timelike Curves," *Physical Review*, September 15, 1990, pp. 1915–1930.

[19] Not all accept this view. See, for example, A. Anderson and B DeWitt, "Does the Topology of Space Fluctuate?" *Foundations of Physics*, February 1986, pp. 91–105.

[20] R. W. Fuller and J. A. Wheeler, "Causality and Multiple Connected Spacetime," *Physical Review*, October 15, 1962, pp. 919–929.

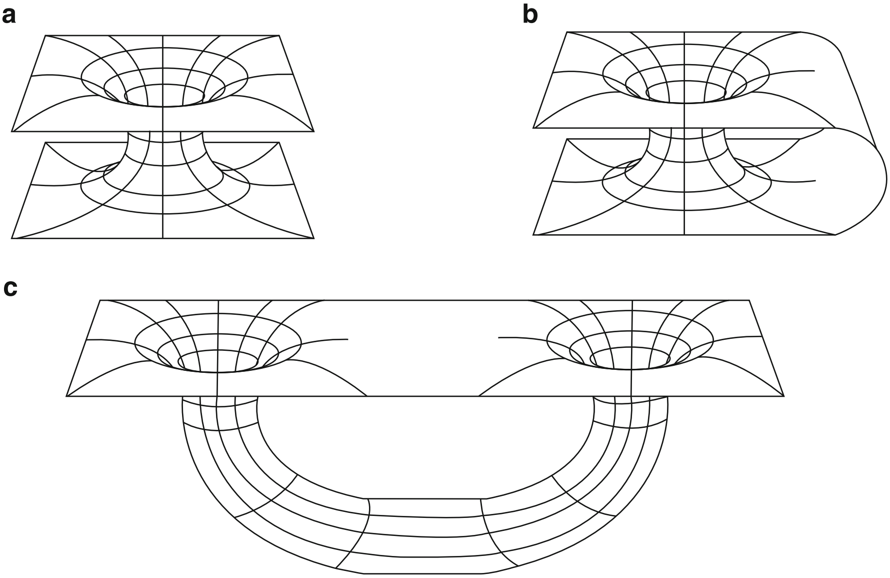

Fig. 6.8 These sketches are unavoidably misleading, being two-dimensional renditions of wormholes that connect two places in a three-dimensional space. Time machine wormholes, on the other hand, connect two places in a four-dimensional spacetime. In particular, the mouths of the wormholes are not ‘depressions’ into which the time traveler’s rocket ship plunges, but rather would appear to be three-dimensional spheres. The wormhole in (**a**) connects two disjoint universes, while those in (**b**) and (**c**) are connections between two places in the same universe. As shown in these last two cases, the wormhole ‘handle’ can be either long or short, compared to the external distance between the wormhole mouths

And anyway, how would one gain access to a wormhole in the first place? As Thorne and one of his students suggested,[21] one might perhaps imagine someday finding a rotating black hole that mathematically possesses, in its interior, so-called hyperspace tunnels to ‘other places’—either in our universe or in other universes (see Fig. 6.8). In the case of a wormhole connecting two places in the same universe, although the external distance in spacetime between the places may be very large (mega-light-years), it is conceivable that the distance *through* the wormhole itself could be very small. The time required to traverse the wormhole, as measured by the traveler’s watch might, in fact, be arbitrarily small. This is exciting, but wormholes do not come without some significant problems.

Such problems include the presence of a one-way event horizon, which precludes two-way travel (it seems reasonable to assume that time and space travelers might wish to eventually return), and enormous gravitational gradients (tidal forces) that dismember anything approaching and/or entering the wormhole. These problems have often been conveniently *ignored* in science fiction. For example, in one tale[22] the time machine is in the form of a wormhole time tunnel that time travelers can simply walk through, and in another characters similarly walk back and forth between two openings in a wall that connect the same place in space but separated in time by 160 years.[23] And in an even older story, we read of a time machine (in the form of a hole in spacetime that is a wormhole in everything but name) through which one can literally step from future to past and back again. That story[24] is particularly notable for having introduced the term *mugwump* for a time traveler who uses a time machine wormhole. To paraphrase the tale's time traveler, as he transits the wormhole his "mug" is in the past and his "wump" is in the future.[25]

---

[21] M. S. Morris and K. S. Thorne, “Wormholes in Spacetime and Their Use for Interstellar Travel: A Tool for Teaching General Relativity,” *American Journal of Physics*, May 1988, pp. 395–412. This paper was motivated by Thorne’s earlier response to a request, from the American astronomer Carl Sagan (1934–1996), for help in making plausible the interstellar travel imagined in his 1985 novel *Contact*.

Now, before going any further, a note on what physicists are referring to when they write of traversable time travel wormholes. Such wormholes are called *Lorentzian* because they have a spacetime metric with the signature $[+, -, -, -]$ (see Sect. 3.5). Further, the wormhole is taken to be *static*, that is, to have no time-varying behavior. The reason for being specific on the nature of time travel wormholes is that there is another type with what is called the *Euclidean signature*: $[+, +, +, +]$, which is not suitable for time traveling. Motion in a Euclidean signature wormhole involves imaginary momentum or proper time, neither of which is physically plausible for a time traveler. However, *if* you can gain access to a Lorentzian wormhole *then*, as two physicists wrote, "if you manage to acquire even one inter-universe traversable wormhole then it *seems* almost absurdly easy to build a time machine."[26]

In response to that fundamental question of how to gain access to a Lorentzian wormhole, Thorne and his students were bluntly honest—they didn't know. Their best suggestion was that "one can imagine an advanced civilization pulling [a] wormhole out of the quantum foam and enlarging it to classical size."[27] A few years later this dramatic idea found its way into a science fiction novel of a far-future alien civilization able to control the energies of constellations of galaxies: "Spacetime is friable. Wormholes riddle the fabric of spacetime on all scales. At the Planck length and below, wormholes arising from quantum uncertainty effects blur the clean Einsteinian lines of spacetime. And some of the wormholes expand to human scale, and beyond — sometimes spontaneously, and sometimes at the instigation of intelligence.”[28]

---

[22] R. C. Wilson, *A Bridge of Years*, Doubleday 1991.

[23] M. Leinster, *Time Tunnel*, Pyramid Books 1964.

[24] H. Kuttner, "Shock," *Astounding Science Fiction*, March 1943.

[25] The term's origin is in 19th century politics, as a description of fence-sitters who try to avoid taking a definite position on some controversial topic under debate. For an illustration of wormhole mugwumping, see K. S. Thorne, *Black Holes & Time Warps*, W. W. Norton 1994, p. 500.

[26] C. Barceló and M. Visser, "Twilight for the Energy Conditions?" *International Journal of Modern Physics D*, December 2002, pp. 1553–1560. See also M. Visser, *Lorentzian Wormholes*, AIP Press 1996.

[27] M. S. Morris, K. S. Thorne, and U. Yurtsever, "Wormholes, Time Machines, and the Weak Energy Condition," *Physical Review Letters*, September 26, 1988, pp. 1446–1449.

This isn’t easy to visualize, but let’s plunge ahead and assume we can “pull” a wormhole out of the quantum foam. If so, then once it is (somehow) inflated the wormhole could be stabilized against collapse by threading it with either matter or fields of stupendous negative (outward) tension—and by *stupendous* I mean *STUPENDOUS*. If b_0 denotes the minimum radius of the wormholes (the size of the so-called *throat* of the wormhole), the tension (radial pressure) at that location must be at least

$$\tau_0 = \frac{3.8}{{b_0}^2} \times 10^{36} \frac{tons}{in^2}$$

where b_0 is expressed in feet.[29] (This expression shows that for a wormhole with a throat radius of several thousand feet, the value of τ_0 is enormous, of the same magnitude as the pressure at the center of the most massive neutron star.) To stabilize a common sort of everyday wormhole, such as a subway tunnel, we can obtain the required tension/pressure by lining the tunnel with iron plates or concrete. But how, for a hyperspace wormhole, do we obtain ‘iron plates’ that can achieve the required enormous tension? As Thorne and his students observed (note 27) such stuff could only be called “exotic,” a term that had appeared a few years earlier in connection with the observed energy density in the throat of a wormhole.[30]

One possible approach to this problem does not use matter at all. If we make b_0 *very* large, then non-material fields will do the job.[31] Indeed, suppose b_0 equals 1 light year (a large wormhole by anybody’s standard!). Then τ_0 is ‘only’ 4000 tons/square inch, and that is achievable by threading the wormhole throat with a magnetic field of ‘only’ 2,700,000 gauss (five million times stronger than the Earth’s field).[32] To generate such a field is not impossible, and present-day experimental, hypervelocity electromagnetic rail guns in development for the US Navy use transient magnetic fields in the mega-gauss range.

---

[28] From S. Baxter’s 1993 novel *Timelike Infinity*. That same year the Chinese physicist Liao Liu wrote on how, as a result of a naturally occurring vacuum fluctuation, a wormhole might spontaneously appear: L. Liu, “Wormhole Created from Vacuum Fluctuation,” *Physical Review D*, September 15, 1993, R5463–R5464.

[29] The original equation for τ_0 was given in units of dynes/cm^2, with b_0 is expressed in meters (see note 21), but I have converted these units to the more familiar (to the non-physicist) units of tons per square inch, for dramatic purposes, because the units of dynes per square centimeter is so small it’s difficult to relate it to anything of everyday significance.

[30] R. Balbinot, “Crossing the Einstein-Rosen Bridge,” *Lettere Al Nuovo Cimento*, May 16, 1985, pp. 76–80.

[31] See, for example, Y. Soen and A. Ori, “Improved Time Machine Model,” *Physical Review D*, October 15, 1996, pp. 4858–4861, and D. N. Vollick, “How to Produce Exotic Matter Using Classical Fields,” *Physical Review D*, October 15, 1997, pp. 4720–4723.

[32] To understand how the calculation of a magnetic field from a pressure requirement is accomplished, note that pressure is dimensionally equivalent to field energy per unit volume, which in turn is given by a well-known result in electromagnetic theory.

As if the stabilization problem wasn't enough of a complication, Thorne and his colleagues (note 27) showed that there is another, even more curious problem. The geometrical requirement that the wormhole interior smoothly connect to the external, asymptotically flat exterior spacetime demands that the wormhole throat flare outward as shown in Fig. 6.8. It turns out that this condition is mathematically equivalent to a requirement that τ_0 exceed the energy density of the throat material[33]; and special relativity, in turn, says that for some timelike observers the energy density will then actually be *negative*. This is a clear violation of the so-called *weak energy condition* (WEC), which says the observed mass-energy density is always non-negative. This is *so* 'obvious' that the WEC was thought, for a long time, to be almost a law of nature. Such a violation is actually not as crazy as it might sound however because, more than half-a-century ago, it was shown[34] that an energy density that is everywhere and everywhen positive is *not* compatible with any quantum field theory that is *local*, as, presumably, will be the yet-to-be-discovered theory of quantum gravity. Over the years since then a variety of other energy conditions have been proposed, such as the *averaged weak energy condition* (AWEC), which says that only the *average* value of the energy density over a complete null geodesic world line has to be non-negative, which leaves open the possibility of *temporary* negativity here, there, then and when.[35]

The traversable, static wormholes studied by Thorne and his colleagues violate even the AWEC (see note 27), however, and in the years since it has become clear to physicists that imposing constraints on the mass-energy density may not be so 'obvious' after all (see note 26). But all may not be lost. Indeed, it has been shown that the violation of the AWEC by traversable wormholes can be made as small as desired, that is, the requirement for the exotic matter required to line a wormhole throat to keep it open can be made as tiny as you want.[36]

Well, tiny the quantity of exotic matter may be but, nonetheless, even a tiny amount of it would be extraordinary weird stuff because a negative energy density can be interpreted as meaning that the exotic material that keeps the wormhole throat open does so by exerting a *repulsive* gravitational force.[37] A repulsive force sounds like a property we'd expect to see associated with *negative mass* and, although such a thing has never been observed (negative matter is *not* anti-matter, which *has* been observed and which does not repel 'normal matter'), it was studied long ago (theoretically, of course) by the English cosmologist Hermann Bondi (1919–2005). Bondi showed[38] that negative mass would indeed have some truly bizarre properties,[39] but there is nothing in general relativity that forbids its possible existence. Wormholes, with negative mass throats, should produce observable effects by which a wormhole *might* be detected. Mathematical analyses of the effect a negative-mass wormhole mouth would have, when crossing the line-of-sight between Earth and a distant star, indicates that there should be an observable *double-spike* in the intensity of the star's light. Astronomical searches for such an optical signature have actually been conducted, with (alas) no success as I write (2017).

---

[33] The condition of τ_0 exceeding the energy density in the throat is, in fact, the technical definition of *exotic*—see note 30. In everyday situations, the exotic condition is never even remotely approached. For example, the maximum tension necessary to pull a piece of steel apart—the so-called *tensile strength*, about 100,000 pounds per square inch—is a trillion times less than the mass-energy density of steel.

[34] H. Epstein, V. Glaser, and A. Jaffe, "Nonpositivity of the Energy Density in Quantized Field Theories," *Il Nuovo Cimento*, April 1, 1965, pp. 1016–1022.

[35] T. A. Roman, "Quantum Stress-Energy Tensors and the Weak Energy Condition," *Physical Review D*, June 15, 1986, pp. 3526–3533.

[36] M. Visser, S. Kar, and N. Dadhich, "Traversable Wormholes with Arbitrarily Small Energy Condition Violations," *Physical Review Letters*, May 23, 2003, pp. 201102-1 to 201102-4.

There is another interesting implication of a repulsive gravitational force, one that proves to be *essential* to the possibility of a wormhole time machine. Just as Einstein's famous prediction (verified in 1919) from general relativity, that star light passing near the Sun's edge is bent *inward* by the Sun's attractive gravitational field, the repulsive, anti-gravity field of a wormhole will cause any light rays traveling through the wormhole to be bent *outward*. That is, a tight, narrow beam of radiation entering a wormhole will emerge *defocused*. This is crucial because, as you'll soon see, a wormhole time machine would otherwise be destroyed by the light from the dimmest candle.

One might take the failure of astronomical searches for a double-spike light signature to mean that wormholes with negative mass throats (thus violating the WEC) simply don't exist. But not so fast. The first hint that the possibility of a negative energy density might not be such a crazy idea occurred as long ago as 1948, with a theoretical prediction made by the Dutch physicist Hendrick Casimir (1909–2000). As pointed out in Chap. 1, the Heisenberg uncertainty principle allows a temporary violation of conservation of energy to occur, with the magnitude of the allowed violation increasing with decreasing time duration. Even in a vacuum, then, with particle/anti-particle creation and annihilation spontaneously and continuously taking place, the *average* energy density being zero does not preclude fluctuations away from zero and so, at times, actually becoming *negative*. What Casimir showed was that if one positioned two perfectly conductive plates parallel to each other, then the normal quantum fluctuations of the energy density in this 'vacuum sandwich' would be altered in such a way as to result in their mutual attraction—and this (tiny) effect was later actually observed.[40]

---

[37] The *strong energy condition* says gravity is always attractive—which is clearly not true in a wormhole throat—and so static, traversable wormholes violate both the weak *and* the strong energy conditions.

[38] H. Bondi, "Negative Mass in General Relativity," *Reviews of Modern Physics*, July 1957, pp. 423–428.

[39] For example, general relativity says that a negative mass will *repel* all other masses (positive *and* negative), whereas a positive mass will *attract* all other masses (positive *and* negative). Imagine, then, a negative mass attached to the nose of a positive-mass spaceship. The spaceship tries to move toward the negative mass, while the negative mass tries to move away from the spaceship. So off they both go into the sky, like a cat chasing its tail. This so-called *reactionless anti-gravity drive*, bizarre as it appears, does not violate either of the conservation laws of energy or momentum. See G. Cavalleri and E. Tonni, "Negative Masses, Even if Isolated, Imply Self-Acceleration, Hence a Catastrophic World," *Il Nuovo Cimento B*, July 1997, pp. 897–903.

What does it mean to 'alter the normal quantum fluctuations'? Consider the creation of a photon and its anti-particle, which is another photon. From the wave interpretation of particles, the parallel plates restrict the photons that appear in the vacuum layer to those that have wavelengths that 'fit' because those wavelengths are submultiples of the plate separation (this requirement follows from the fact that a perfectly conducting plate cannot support a non-zero tangential electric field). Photons with longer wavelengths than the plate separation cannot 'fit' and thus do not appear. That is, the parallel plates have created a boundary condition that has quantized the electromagnetic field. The absence of these 'longer wavelength' photons lowers the average energy density between the plates and, because the average without the plates is zero, the altered average energy density must be negative. Indeed, the more the maximum allowed photon wavelength decreases with decreasing plate separation, the more negative the average energy density becomes in the enclosed Casimir vacuum. The negative energy density manifests itself as an inward directed force per unit area (remember, energy density and pressure are dimensionally equivalent).

The experimental detection of the Casimir effect was a remarkable event in physics. As one mathematician put it, "No worker in the field of overlap of quantum theory and general relativity can fail to point this fact out in tones of awe and reverence."[41] Robert Forward, an imaginative physicist who has appeared in this book earlier as an enthusiastic supporter of time travel, has described how the Casimir force might be used to extract energy literally *from a vacuum*. This is an idea as seemingly impossible as is the plan of one science fiction professor to squeeze energy out of *time*. As he asks his assistant, "But tell me, Bob, isn't that a ridiculous thought? To take time, something intangible, invisible, incomprehensible, and contract it — squeeze it together like a sponge?"[42] The story is fun, but Forward's proposal is that as well—*and* good physics.[43] Now, what does all this have to do with wormhole time machines?

---

[40] A complete presentation of Casimir's analysis, with citations to the original literature, can be found in L. E. Ballentine, *Quantum Mechanics*, Prentice-Hall 1990, pp. 399–403. For an historical, tutorial presentation, including Casimir's personal comments on how he was led to make his discovery, see P. W. Milonni and M.-L. Shih, "Casimir Forces," *Contemporary Physics*, September–October 1992, pp. 313–322.

[41] S. A. Fulling, *Aspects of Quantum Field Theory in Curved Space-Time*, London Mathematical Society 1989.

[42] E. Binder, "The Time Contractor," *Astounding Stories*, December 1937.

[43] R. Forward, "Extracting Electrical Energy from the Vacuum by Cohesion of Charged Foliated Conductors," *Physical Review B*, August 15, 1984, pp. 1700–1702.

Thorne and his colleagues (note 27) proposed to use the Casimir effect to achieve the "exotic condition" *without* matter. Their idea was to place identical, conducting, spherical plates that carry equal electric charges at each end of the wormhole (remember, the wormhole mouths are spherically symmetric). The two identical charges repel each other, but the charge size is adjusted so that the gravitational attraction of the plates precisely cancels the repulsion. They then calculated that the Casimir effect results in a negative energy density sufficient to provide the throat tension necessary to prevent wormhole collapse.

There *are* some weird aspects to this (and perhaps that's no surprise). For example, the analysis assumed that the wormhole length is *very* small compared to its radius (10^{-10} cm long and 200 million miles wide!), with the short length required because it represents the separation of the wormhole plates, and the smaller the separation the more negative the average energy density. (The functional dependence is as the *fourth* power of the separation.) Another problem is the balancing of the electrical repulsion and the gravitational attraction of the wormhole mouth plates, as such a balance is clearly an unstable one. Finally, because the two spherical plates completely fill the wormhole mouths, how would a traveler actually get through the wormhole? The 'answer' was to drill a hole through the plates and hope that wouldn't perturb the Casimir vacuum too much.[44]

All the above litany of the difficulties static, traversable wormholes face in simply existing is certainly daunting, but let's now ignore all that and suppose we actually have a wormhole with both mouths in the same universe. (For use as a time machine, it would seem desirable for the time traveler to remain in his/her own universe!) So, how do we turn the wormhole into a time machine? Interestingly, while it is general relativity that gives us the wormhole, it is special relativity that adds the final touch of backward time travel. We begin by imagining that, somehow, one mouth of the wormhole can be moved with respect to the other mouth. One early suggestion, for example, was to use the gravitational attraction of a large asteroid to 'drag' one end of the wormhole, thereby inducing a time dilation effect.[45]

That is, suppose we have two clocks A and B, one in each mouth of the wormhole. These two clocks, and all other clocks in the flat spacetime outside the wormhole, are initially indicating the same time and running at the same rate. Now, recalling the twin paradox from Chap. 3, let each mouth-clock play the role of one of the twins. Imagine that A and B are now separated because the mouth containing B is placed on board a rocket ship. The rocket ship takes a long, high-speed trip out into space along the straight-line path joining A and B in external space, and then returns, just as described in Sect. 3.5. We then unload the space traveling wormhole mouth (with its clock B) and reposition it at its original location. What is the situation now? We can summarize matters as follows: (1) Clock A, in the non-moving mouth, remains in-step with the local clocks in the space outside the mouth. (2) Clocks A and B, both inside the wormhole, have *not* moved with respect to each other because we are assuming a very short wormhole handle, as in part (b) of Fig. 6.8. We can arrange for the motion of the space traveling mouth (with clock B) to be such that the handle is always short, and so the distance between clocks A and B changes by an arbitrarily small amount. Thus, clocks A and B remain in-step with each other. (3) Clock B, because it has been moving with respect to its external space, arrives back at its starting position reading *behind* (that is, *earlier*) than the clocks outside its wormhole mouth.

---

[44] The very next year it was shown how to construct non-spherically symmetric wormholes to avoid that particular problem: see the two papers by M. Visser, "Traversable Wormholes: Some Simple Examples," *Physical Review D*, May 15, 1989, pp. 3182–3184, and "Traversable Wormholes from Surgically Modified Schwarzchild Spacetimes," *Nuclear Physics B*, December 11, 1989, pp. 203–212.

[45] J. L. Friedman, "Back to the Future," *Nature*, November 24, 1988, pp. 305–306.

For the sake of argument, then, suppose the journey of B is such that there is a two-hour time-slip between clock B and its local, external clocks. Thus, if clock B reads 9 A.M., the clocks outside of mouth B will read 11 A.M. But because clocks A and B are in-step, clock A reads 9 A.M., as do the clocks outside of mouth A. That is, the wormhole connecting mouth A to mouth B is a connection between two parts of the same universe that are two hours apart in time. Now, suppose the journey from mouth A to mouth B can be made through external space in one hour. Then, one could leave mouth A at 10 A.M., rocket to mouth B by 11 A.M., and travel back to mouth A via the wormhole to the starting point—where it is 9 A.M., one hour *before* the trip began! We could, in fact, imagine repeating this process, going back one additional hour for each new loop through the wormhole. One clear restriction, however, is that we could not go back in time to *before* the creation of the wormhole time machine. The wormhole works in the other direction, too. To see this, suppose that the space traveler leaves mouth B at 8 A.M. and rockets through external space to mouth A, arriving at 9 A.M. Entering mouth A, he exits from mouth B (where he started) at 11 A.M., two hours in the future.

Another way to induce a time dilation effect, to convert a wormhole to a time machine, *without moving either mouth*, is to simply place one mouth in an intense gravitational field, that of, say, a neutron star. (Recall, from Sect. 3.3, how gravity influences the time-keeping rate of a clock.) As the physicists who proposed this idea put it, almost *any* interaction with surrounding matter and gravity fields almost inevitably turns a wormhole into a time machine.[46] Others have admitted that the details of the origin of time dilation are probably not issues worth debating, but rather what is called the *back-reaction* is of far more concern.

To understand the back-reaction requires mention of what is called the *Cauchy horizon*, the hyperspace surface in spacetime that separates the region where closed timelike lines can exist, from the region where they cannot exist. The back-reaction is the build-up of unbounded energy levels on the Cauchy horizon, causing its instability and rapid destruction. The name of the horizon comes from the "Cauchy problem"—named after the nineteenth century French mathematician Augustin-Louis Cauchy (1789–1857)—in the theory of partial differential equations. In this theory a Cauchy initial-value problem is said to be well-defined if the initial conditions determine a unique solution, and if a continuous variation in the initial conditions gives a continuous variation in the solution. In that part of spacetime where closed timelike loops are not allowed, backward causation does not occur (by definition) and the laws of physics (all expressed as differential equations) satisfy the Cauchy condition. Outside of this *chronal* region, that is, beyond the Cauchy horizon where physics is *dischronal*, however, the possibility of backward causation raises the possibility of violating the Cauchy condition, and in such a case the Cauchy horizon is also sometimes called the *chronology horizon*.[47]

---

[46] V. P. Frolov and I. D. Novikov, "Physical Effects in Wormholes and Time Machines," *Physical Review D*, August 15, 1990, pp. 1057–1065.

The instability of the Cauchy horizon is caused by radiation that propagates in closed timelike loops that thread through the wormhole on ‘straight lines.’ This radiation, as shown a half-century ago,[48] builds-up unbounded energy density levels at the horizon, and thus destroys the horizon. Thorne and his colleagues argued that the defocusing effect of their wormhole time machine’s repulsive gravity would be sufficient to counter a disruptive energy build-up on the horizon (note 27). Subsequent analyses have examined other possible ways to avoid unbounded energy density on the Cauchy horizon. For example, in one paper[49] it was imagined that a wormhole time machine has had a circular motion induced for mouth B; that is, mouth B orbits around mouth A. The result is that the Cauchy horizon now does seem to be stable, because now there are no fixed, straight-line timelike loops threading the wormhole from A to B to A to B to . . .. That is, B is a ‘moving target’ and there is no point on the Cauchy horizon where the energy density becomes unbounded.

Yet another approach for achieving the disruption of destructive, circulating energy loops through a wormhole is by placing a spherical mirror between the two mouths of the wormhole. Proposed by the Chinese physicist Li-Xin Li, a Li mirror would divert all closed null geodesics (represent circulating radiation) that potentially thread through the wormhole.[50] Such potentially fatal geodesics would, instead, be scattered back into space, whereas a purposeful traveler could navigate *around* the mirror and thus use the wormhole as a time machine.

---

[47] A classic work on the mathematics of Cauchy problems is J. Hadamard, *Lectures on Cauchy’s Problem in Linear Partial Differential Equations*, Dover 1952. There is a curious bit of irony in this. In a section of his book, Hadamard uses spacetime to illustrate one possible four-dimensional space and, in passing, he casually writes “This conception was beautifully illustrated a good many years ago by the novelist Wells in his *Time Machine*.” Hadamard wrote his book in 1923, and he would almost certainly have been astonished to have been informed that less than seventy years later his work would play a central role in the non-fictional theory of time machines.

[48] C. W. Misner and A. H. Taub, “A Singularity-Free Empty Universe,” *Soviet Physics JETP*, January 1969, pp. 122–133.

[49] I. D. Novikov, “An Analysis of the Operation of a Time Machine,” *Soviet Physics JETP*, March 1989, pp. 439–443.

[50] L.-X. Li, “New Light on Time Machines: Against the Chronology Protection Conjecture,” *Physical Review D*, November 1994, pp. R6037–R6040.

Cauchy horizon instability from the back reaction is central to Hawking's Chronology Protection Conjecture, discussed in Chap. 1. His analysis (see note 54 in Chap. 1) led him to conclude that a physical entity—the stress-energy tensor—becomes unphysical on the Cauchy horizon. That is, because of time-traveling quantum field fluctuations of the vacuum, that tensor diverges to infinity at the horizon. This results in a failure of that horizon to form in the first place or, if it does form, in the creation of a singularity that 'seals-off' the horizon to any would-be time travelers attempting to gain access to the closed timelike loops beyond the horizon. Others, however, argued that Hawking was mistaken in claiming that the divergence of the stress-energy tensor on the Cauchy horizon will always forbid time travel.

In a study, for example, of a complex-valued spacetime metric (and such a metric *is* allowed in the so-called 'sum over all possible geometries, path integral' approach to the quantum theory of gravity), that has causal and non-causal spacetime regions separated not by a Cauchy horizon but rather by a region of complex geometry, the stress-energy tensor is always physical and diverges nowhere. The complex geometry region plays the same role as the Cauchy horizon, because such a region would, *classically*, mean that the two regions cannot be reached from one another, but via quantum tunneling an observer *could* travel between the two regions.[51] In fact, studies of stress-energy divergence actually have a long history. For example, the effect of an unphysical (infinite) gravitational and/or electromagnetic energy flux had been analyzed *years before* the wormhole time machine studies began.[52] The authors studied the case of a potential traveler to "new worlds" who tries to cross the Cauchy horizon of an electrically charged, non-rotating black hole. An even earlier computer study had already concluded that, for such a traveler, the attempt to cross the horizon "looks liable to prove a dangerous undertaking."[53]

It isn't at all clear, in fact, if a *theoretical* divergence of the stress-energy *is* the signature of a failure of physics. One doesn't need anything as bizarre as a time machine for the stress-energy to diverge *on paper*. It was shown nearly 40 years ago, for example, that such a theoretical divergence occurs for the electromagnetic field near a perfectly conducting boundary.[54] But it is simply the unphysical nature of a "*perfectly* conducting" boundary condition that causes the divergence, not the fact that the field actually exists near a conducting boundary. Similarly, other real-life considerations (quantum gravity) *may* keep the stress-energy physical everywhere in a time machine spacetime.

---

[51] L.-X. Li, J.-M. Xu, and L. Liu, "Complex Geometry, Quantum Tunneling, and Time Machines," *Physical Review D*, November 15, 1993, pp. 4735–4737.

[52] S. Chandrasekhar and J. B. Hartle, "On Crossing the Cauchy Horizon of a Reissner-Nordström Black-Hole," *Proceedings of the Royal Society of London A*, December 8, 1982, pp. 301–315.

[53] M. Simpson and R. Penrose, "Internal Instability in a Reissner-Nordström Black Hole," *International Journal of Theoretical Physics*, April 1973, pp. 183–197.

[54] D. Deutsch and P. Candelas, "Boundary Effects in Quantum Field Theory," *Physical Review D*, December 15, 1979, pp. 3063–3080.

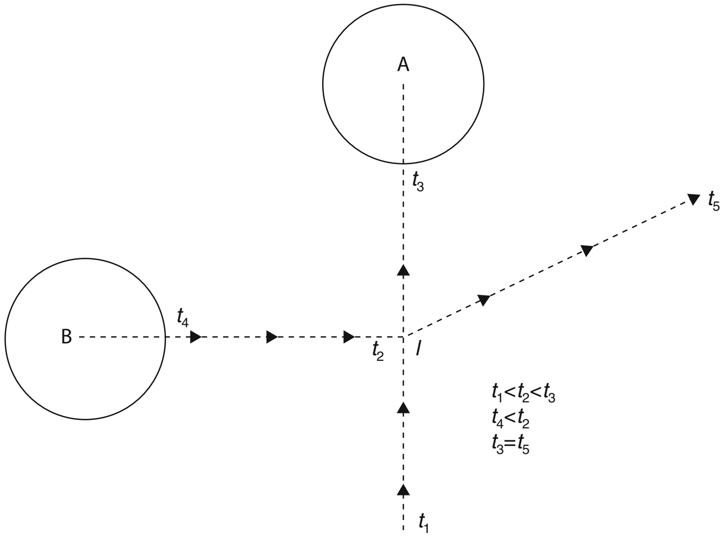

Fig. 6.9 The grandfather paradox in the billiard ball world. A billiard ball approaches mouth A of a time machine wormhole, dead-on center and, just before entering A, it passes without incident through point *I*. The ball then enters A and so exits mouth B in the past, *just in time* to pass through point *I* and *hit its younger self*. This impact knocks the younger ball *away from* A, so we have the familiar paradox of changing the past. That is, the impact did not occur when the ball 'originally' passed through *I* on its way to A and, of course, we also wonder how the ball manages to hit itself after leaving B if it then *doesn't* enter A?

Assuming a wormhole time machine has (somehow) become available, with its Cauchy horizon intact, how do the 'paradoxes' of time travel come into play? In an attempt to study the grandfather paradox, in particular, Thorne and his colleagues studied self-interacting billiard balls traveling backward in time through a wormhole.[55] They used billiard balls—see Figs. 6.9 and 6.10—rather than human time travelers for the same reason Wheeler and Feynman used a pellet and shutter mechanism in their study of advanced electromagnetic waves—to avoid any metaphysical questions about human free will. The central issue for them was the determination of the multiplicity of trajectories for a single, self-interacting time traveling ball, where the Cauchy condition for a well-defined trajectory in spacetime is *unique* self-consistency.

That is, for the trajectory to be well-defined in the Cauchy sense, it was expected there would be exactly *one* consistent trajectory for a self-interacting ball. A multiplicity of zero, of course, would be the physics declaring backward time travel through the wormhole to be nonsense—and that was thought to be a distinct

---

[55] F. Echeverria, G. Klinkhammer, and K. S. Thorne, "Billiard Balls in Wormhole Spacetimes with Closed Timelike Curves: I. Classical Theory," *Physical Review D*, August 15, 1991, pp. 1077–1099. The authors credited the physicist/science fiction writer Robert Forward (who used the same ideas in his 1992 novel *Timemaster*) for motivating their research.

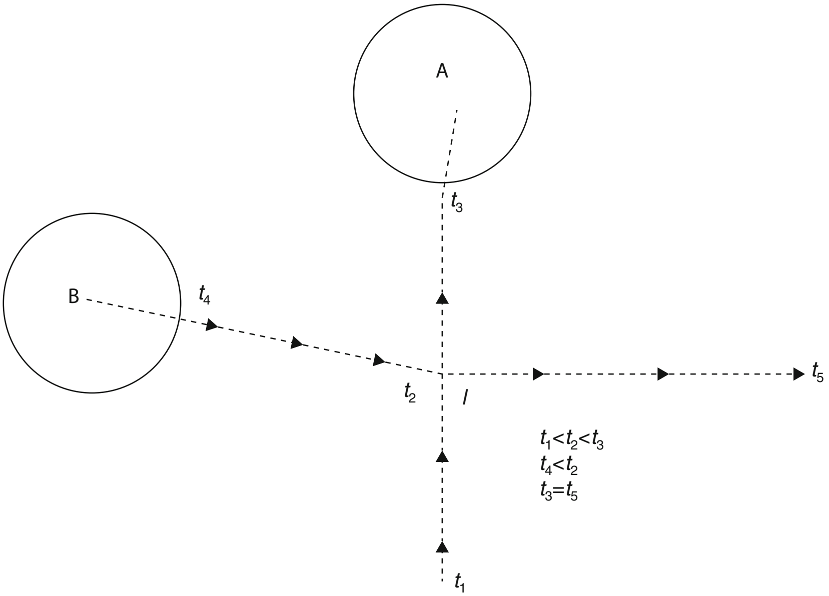

Fig. 6.10 The principle of self-consistency in the billiard ball world. Now the ball, in passing through point *I* on its dead-on center path toward mouth A of the wormhole time machine, is suddenly hit a grazing blow by another ball that has just shot out of mouth B (and into the past) at an angle. The impact knocks the first ball slightly off its original trajectory, and it enters mouth A slightly off-center. Thus, the ball emerges from B into the past slightly off-center and just in time to glace off itself at *I*—which explains *why* it emerged from B slightly off-center!

possibility. The actual results were, however, surprisingly different. It was found, under very general assumptions about the wormhole parameters, that (1) there are no trajectories with zero multiplicity and (2) the multiplicity is *not* one but rather is *always* infinity! Thus, the billiard ball form of the grandfather paradox *was* found to be *not* well-defined, but not for the expected reason that there was no self-consistent solution. Instead, it was because there are *too many* solutions.

This astonishing, completely unexpected result seems to be just what is needed to support the viability of time machines, as it appears to allow a definition of *well-defined* in the Cauchy sense and still permit an answer to the puzzle of free will. The initial conditions of a time traveling ball give rise to an infinity of self-consistent trajectories, each occurring in the same way that a random variable takes on different values with each new performance of the experiment that the random variable is defined on. And yet, there are still unique probability density functions for all sets of measurements that one might make anywhere along these trajectories. Thus, the Cauchy problem is *stochastically* well-defined; at the start of any trajectory, we do not know in detail what will happen except that whatever does happen will be self-consistent. In this probabilistic sense, then, wormhole time travel to the past and the retention of free will both make sense. The Russian physicist Igor Novikov and his colleagues continued the study of time traveling billiard balls,[56] demonstrating that one can *deduce* self-consistency from the long-accepted principle of least action (that is, self-consistency is not an additional assumption to existing physics).[57]

The one-wormhole, two-mouth time machine was actually not the first kind of wormhole time machine described in the physics literature. In their 1988 paper (note 21), Morris and Thorne initially described a time machine constructed from *two* wormholes, but they added a note-in-proof at the end that they had just discovered how to build a time machine using one wormhole (the machine we have been discussing). This reduction in the required number of wormholes was thought to be a technical advance, of course, and so the two-wormhole time machine was put aside.

But not for long. Soon thereafter the concerns about Cauchy horizon stability began to surface, a concern that one-wormhole time machines might destroy themselves just at the instant their mouths were about to be threaded by closed timelike curves. As noted earlier, the negative mass wormhole throat has a defocusing effect on electromagnetic radiation (and so the initial concern, that time traveling photons might be fatal, faded)—but then it was found that vacuum fluctuations of quantum fields are *not* so defocused.[58] That failure to defocus time-traveling vacuum polarizations (as quantum field fluctuations are called) was shown to result in an unphysical divergence of the stress-energy on the Cauchy horizon of a one-wormhole time machine.[59] This sounds bad, but the hope was that the divergence wouldn't actually be fatal: it appeared to be sufficiently sluggish that it was suggested reaching an actual *infinity* of the stress-energy would be precluded by the eventual intercession of quantum gravity. That is, the stress-energy might *try* to become unbounded as spacetime approached the formation of a time machine

---

[56] A. Lossev and I. D. Novikov, "The Jinn of the Time Machine: Nontrivial Self-Consistent Solutions," *Classical and Quantum Gravity*, October 1992, pp. 2309–2321; E. V. Mikheeva and I. D. Novikov, "Inelastic Billiard Ball in Spacetime with a Time Machine," *Physical Review D*, February 15, 1993, pp. 1432–1436; M. B. Mensky and I. D. Novikov, "Three-Dimensional Billiards with a Time Machine," *International Journal of Modern Physics D*, April 1996, pp. 179–192.

[57] A. Carlini, *et al.*, "Time Machines: The Principle of Self-Consistency as a Consequence of the Principle of Minimal Action," October 1995, pp. 557–580, and "Time Machines and the Principle of Self-Consistency as a Consequence of Stationary Action (II): The Cauchy Problem for a Self-Interacting Particle," October 1996, pp. 445–479, both in *International Journal of Modern Physics D*.

[58] V. P. Frolov, "Vacuum Polarization in a Locally Static Multiply Connected Spacetime and a Time-Machine Problem," *Physical Review D*, June 15, 1991, pp. 3878–3894.

[59] S.-W. Kim and K. S. Thorne, "Do Vacuum Fluctuations Prevent the Creation of Closed Timelike Curves?" *Physical Review D*, June 15, 1991, pp. 3929–3947. See also L.-X. Li, "Must Time Machines Be Unstable Against Vacuum Fluctuations?" *Classical and Quantum Gravity*, September 1996, pp. 2563–2568.

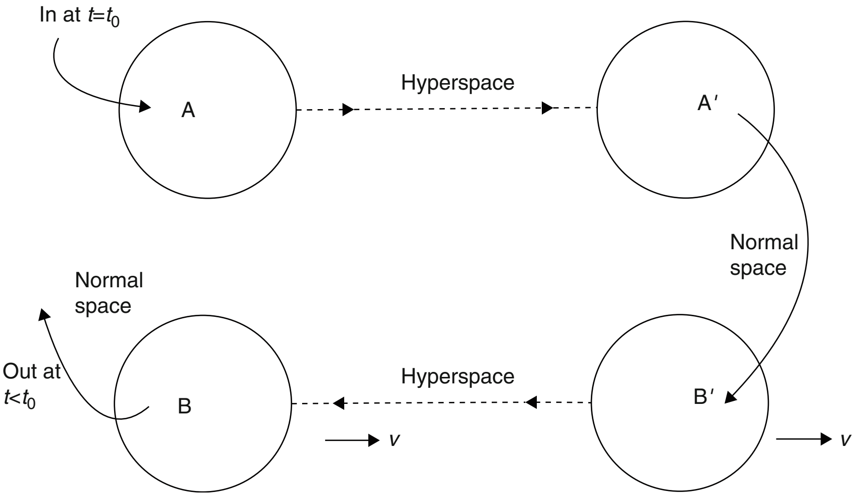

Fig. 6.11 A two-wormhole, Roman ring time machine

but, before it becomes so large as to destroy the time machine, quantum gravity would cut-off the divergence 'in time' (so to speak!) to save the machine.

Hawking disagreed (in his famous Chronology Protection Conjecture—see note 54 in Chap. 1), arguing that Kim and Thorne had made a crucial error in their calculations. According to Hawking, the divergence of the stress-energy may indeed be cut off by quantum gravity, but not before the development of spacetime disturbances representing perhaps a hundred million times the energy levels associated with ordinary chemical binding energies. These would be sufficiently big disturbances to raise serious doubts about the physical survival of a one-wormhole time machine, even in the absence of a true stress-energy infinity. Hence the resurrection of the two-wormhole time machine geometry. Perhaps *it* could avoid the destructive effect of time-traveling vacuum fluctuations.

If a spacetime contains multiple wormholes, then it is called a *Roman spacetime* after the physicist Thomas Roman (at Central Connecticut State University), who was the originator of such spacetimes. Each of these wormholes, individually, is *not* a time machine. Together, however, they form a time machine geometry called a *Roman configuration* (or a *Roman ring*).[60] Here's how.

In Fig. 6.11 two pairs of wormhole mouths are labeled A, A' and B, B'. We imagine that the A, A' wormhole is stationary and that its two mouths are very far apart in normal space—so far apart, in fact, that if a traveler enters A and almost instantly (because the wormhole handle is very short in hyperspace) emerges from A', it will appear to an observer at rest with respect to the wormhole that the traveler has moved faster than light. That is, entering A and exiting A' are events with spacelike separation. Now, imagine also that the wormhole with mouths B, B' is moving past the first wormhole at speed v. To an observer in this second, *moving* frame of reference, the spacelike separation of entering A and exiting A' can result in the two events being temporally reversed if v is sufficiently large (but still less than the speed of light). Therefore, upon emerging from A' the traveler crosses normal space to the moving wormhole mouth B', enters the wormhole, and then almost instantly emerges from mouth B, and finally travels again through normal space to mouth A. If the traveler can make the two trips in normal space in less time than the backward time shift achieved by the temporal reversal of entering A and exiting A', then we have a time machine

---

[60] M. Visser, "Traversable Wormholes: The Roman Ring," *Physical Review D*, April 15, 1997, pp. 5212–5214.

Two simultaneous analyses of the Roman ring time machine each concluded that, for suitable choices of sizes (the radii of the wormhole mouths, the wormhole lengths in normal space, the lateral offset of the two wormholes, and the relative speed of the wormholes), the stress-energy divergence *can* be limited by quantum gravity to an arbitrarily weak level. That is, the two-wormhole time machine is not necessarily destroyed by an unbounded stress-energy on the Cauchy horizon.[61] But not all was now put right.

Visser, in particular, had some strong reservations about the Roman ring. Although he granted that a quantum gravity cut-off the stress-energy divergence *would* probably occur in the Roman ring, he called the required special sizing conditions "bizarre," and asserted that the resulting time machine would be quite useless for a human traveler in any case. For example, he calculated that only if the mouths of the wormholes are separated in normal space by the radius of the universe (!), and only if the wormhole mouths have radii on the order of that of an atomic nucleus, would the cut-off be sufficient to allow the putative time machine to avoid destruction. When Visser reduced the wormholes from universe size to 'merely' that of the distance between the Sun and the Earth, he concluded that it would require energy at the level of the Superconducting Supercollider accelerator to blast an information-bearing message through the narrow wormholes. And even then the 'short' wormholes would provide a maximum penetration into the past of just eight minutes. As Visser put it, "This does not seem to be a workable recipe for studying tomorrow's *Wall Street Journal*.

Lyutikov, on the other hand, took a far less negative stance. He concluded that although Visser's calculations "make it very inconvenient for time travel [by humans]," nevertheless "the [principal] question of the possibility of transmitting information back in time through traversable wormholes would still remain."

The wormholes we have been discussing so far are *static* in time, but another approach is to allow them to be dynamic structures in spacetime. That is, to allow one or more of their parameters to vary with time (perhaps, for example, the throat diameter could collapse). Then, according to one analysis, it is possible to have a traversable wormhole *made of normal matter* and, even though it is collapsing, it would take so long to do so that "a space adventurer will have enough time to pass through the throat of the wormhole from one asymptotically flat region [of spacetime outside the entry mouth of the wormhole] to the other [spacetime region outside the exit mouth of the wormhole] before the radius of the throat shrinks to . . . where the event horizon is developed."[62] Such a dynamic wormhole, it was claimed, satisfies both the weak and the dominant energy[63] conditions, but not the strong energy condition. Thus, gravity would still be repulsive in the throat, but this condition (which would seem to require *exotic* matter) was brushed aside because such a condition is thought to have actually occurred, on a massive scale, during the inflationary stage of the Big Bang[64] (although how that would help in the construction of a wormhole in the future is a bit murky).

---

[61] M. Visser, "Van Vleck Determinants: Traversable Wormhole Spacetimes," April 15,1994, pp. 3963–3980, and M. Lyutikov, "Vacuum Polarization at the Chronology Horizon of the Roman Spacetime," April 15, 1994, pp. 4041–4048, both in *Physical Review D*.

Is it reasonable to think 'useable' wormholes, static or otherwise, can be acquired for the purpose of creating a time machine? At one time, Hawking was sure the answer is no, once writing "The philosophy of this paper is . . . to look for vacuum polarization [the divergence of the stress-energy on the Cauchy horizon] to enforce the chronology protection conjecture."[65] It became increasingly apparent, however, that matters would be a great deal more involved and, as Hawking himself came to admit, "the fact that the energy-momentum tensor fails to diverge [in certain special cases of time machine spacetimes] shows that the back reaction does not enforce chronology protection.".[66]

I think the best (and most honest) way to respond to the 'reasonable' question that opened the previous paragraph is with words from 20 years ago, by the Russian astrophysicist Serguei Krasnikov, words still valid today: "It may well be that the vacuum fluctuations do make the time machine unstable, but nothing at present suggests this. All we have are a few simple examples. In some of them the energy density diverges at the horizon and in some does not. So, the time machine perhaps is stable and perhaps is not."[67]

---

[62] A. Wang and P. S. Letelier, "Dynamical Wormholes and Energy Conditions," *Progress of Theoretical Physics*, July 1995, pp. 137–142. See also L. A. Anchordoqui, *et al.*, "Evolving Wormhole Geometries," *Physical Review D*, January 15, 1998, pp. 829–833.

[63] The dominant energy condition is the weak energy condition *plus* the requirement that any observed energy flux is *never* superluminal.

[64] During inflation the universe is thought to have expanded at a rate far beyond human comprehension. It has been estimated that during the first 10^{-35} second of the Big Bang the universe doubled in each spatial dimension by a factor of two each 10^{-37} second; that is, there were about 100 such doublings. Thus, there was an increase by a factor of $2^{100} \approx 10^{30}$ in each linear dimension of the universe, and the volume increased by the cube of that enormous factor. See Alan Guth, *The Inflationary Universe*, Addison-Wesley 1997.

[65] S. Hawking, "Quantum Coherence and Closed Timelike Curves," *Physical Review D*, November 15, 1995, pp. 5681–5686.

[66] M. J. Cassidy and S. W. Hawking, "Models for Chronology Selection," *Physical Review D*, February 15, 1998, pp. 2372–2380. See also L.-X. Li and J. R. Gott, "Self-Consistent Vacuum for Misner Space and the Chronology Protection Conjecture," *Physical Review Letters*, April 6, 1998, pp. 2980–2983.

The daunting level of technology required to *build* a wormhole (with or without exotic matter) doesn't mean we can't search for existing wormholes. Perhaps, for example, vast wormhole networks were formed naturally at Big Bang time, as described in Gregory Benford's 1997 novel *Foundation's Fear*, where wormholes are "leftovers from the Great Emergence [the Big Bang]." Or perhaps "advanced civilizations" long ago constructed a vast, pan-galactic 'subway system' of wormholes like the one described in Carl Sagan's 1985 novel Co*ntact* (and dramatically illustrated in the 1997 film).

Of course, any such wormhole, if found (via its double-spike light signature, for example), could be a very long way from Earth. It might even be in another galaxy. So, even if we found a wormhole, what could we *do* with it? Surprisingly, maybe a lot. The Russian physicists Igor Novikov and Andrei Lossev (note 56) suggested that a wormhole might be very useful *even if its location is completely unknown, even if we haven't yet even discovered it*! The only assumption they made was that the wormhole has existed for a "sufficiently long time" (and precisely what that means will be explained in just a bit). With that assumption, they showed how to make an information-creating time loop. Here's how they did that.

They began their analysis by assuming that people have no knowledge of how to build spacecraft that can make the interstellar voyage to the distant wormhole, even if they knew in which direction to go to reach the mouth that leads backward in time (mouth B). Instead, they build an automatic spacecraft construction plant that can follow any detailed sequence of instructions provided to it, and then stockpile it with a supply of raw materials (energy, steel, plastic, computers, and so on). When the spacecraft construction is done (*how* that is done is explained in the next paragraph), the last step before launching the spacecraft toward mouth B *will be* to load the on-board computer with the following three pieces of information:

1. The detailed sequence of instructions to be followed in the construction of the spacecraft;
2. The direction from Earth to mouth B;
3. The direction from mouth A (the wormhole exit mouth in the past) back to Earth.

To summarize, people build the automatic plant, load it up with raw materials, *and then withdraw*. This last step is crucial, because it eliminates human free will from further consideration, that is, it removes any temptation to create a bilking paradox. So, what happens next?

Lossev and Novikov suggest that what happens next is that a *very* old spacecraft suddenly appears in the sky and lands next to the automatic construction plant. In its on-board computer are items a, b, and c. Using item a, the automatic plant makes a new spacecraft, then loads the new on-board computer with items a, b, and c from the *very* old spacecraft's on-board computer, and then the new spacecraft is launched toward mouth B (using the information of item b). The *very* old spacecraft is given an honored place in a museum.

---

[67] S. V. Krasnikov, "Quantum Stability of the Time Machine," *Physical Review D*, December 15, 1996, pp. 7322–7327.

The new spacecraft arrives at the distant mouth B in the far future, by which time it is, of course, an old spacecraft (but not yet a *very* old spacecraft). It then plunges into mouth B and almost immediately emerges from mouth A, in the past. Indeed, it repeats this process as many times as required until it is in the far distant past, at a time even before it left Earth. (It might seem that to do this, the spacecraft's computer memory needs a fourth piece of information, the direction from mouth A back to mouth B, but in fact items b and c are sufficient for the old spacecraft to find its way from A to B.) It is now clear how long the wormhole must have been in existence. The old spacecraft repeatedly uses the wormhole time machine until it is so far in the past that it can cruise back to Earth at normal speed (it knows the way back because of item c) and arrive as a *very* old spacecraft, just in time to be placed in the museum!

As Lossev and Novikov pointed out, this remarkable, looped sequence of events has increased knowledge from what it was at the time just before the automatic construction plant was built. People now know both how to build an interstellar spacecraft, and the locations of both mouths of the wormhole. They also now possess a *very* old, used spacecraft. It is curious to note that although the information in the *very* old spacecraft's computer memory has traveled on a closed time loop, the *very* old spacecraft itself has not. This is because the spacecraft left Earth when new, but arrived back (before it left) as *very* old, whereupon it promptly entered a museum. There is therefore no question about the origin of the *very* old spacecraft, but where did the information of items a, b, and c come from? Lossev and Novikov say it came from the energy gained by the spacecraft as it interacted (will interact?) with the rest of the universe while on its journey.

Nobody said time travel isn't weird!

6.4 Gott's Cosmic String Time Machine

> "It's an amazingly simple solution. It doesn't take much physics to understand it."
>
> —MIT astrophysicist Alan Guth, on Gott's discovery of the cosmic string time machine[68]

> "Louise, working out the spacetime geometry of a cosmic string is a hard problem in general relativity. But, given that geometry, all the rest of it is no more than Pythagoras' theorem . . ."
>
> —a character in Stephen Baxter's 1994 novel *Ring*, agreeing with Guth.

---

[68] Quoted from J. Travis, "Could a Pair of Cosmic Strings Open a Route Into the Past?" *Science*, April 10, 1992, pp. 179–180.

A new way to gain access to closed timelike curves, without the involvement of the exotic matter needed by negative-mass wormholes, was described in 1991 by the Princeton physicist J. Richard Gott.[69] Gott gave exact solutions to Einstein's gravitational field equations for what are called *cosmic strings*, solutions that (1) unlike wormholes, do not violate any of the energy conditions, (2) unlike black holes have no crushing singularities or event horizons, and (3) are not topologically multiply connected.

Cosmic strings are fantastically thin (10^{-28} cm in radius) filaments of pure energy that are thought to stretch the width of the universe and to have an enormous linear mass-energy density of 10^{28} g/cm. To generate closed timelike paths in spacetime, Gott required that either two fast-moving (which means moving at practically the speed of light) parallel cosmic strings pass each other on a near-collision course, or that there be a closed-loop string that collapses in a slightly non-planar manner so that the opposite, nearly straight sides 'just miss.' The gravitational interaction of the passing strings can 'warp' spacetime enough to produce closed timelike curves.

A hint at the possibility of violating causality with strings had appeared before Gott's work, but those authors didn't take the time travel implications seriously. As they wrote, "We argue . . . that any realistic model [for a spinning string with angular momentum[70]] . . . will not have closed timelike curves."[71] Gott, however, showed that as two strings pass each other, closed timelike loops *do* encircle the strings.

Gott, who appears to be far less rigid in his view of time travel than are many of his fellow physicists, held out an escape to those who pale at the very thought of time travel to the past. Perhaps, he suggested (following in the footsteps of an analysis by Hawking[72]), as the strings (or string-loop sides) pass, a black hole will form with an event horizon that will seal-off the closed timelike curves from any would-be time traveler. Or perhaps, he further suggested, the more realistic case of non-singular strings (that is, strings with non-zero-filament radii) and possessing

---

[69] J. R. Gott, "Closed Timelike Curves Produced by Pairs of Moving Cosmic Strings: Exact Solutions," *Physical Review Letters*, March 4, 1991, pp. 1126–1129.

[70] The two strings in a Gott-pair are *not* necessarily spinning, and no such assumption was made by Gott. They don't even have to be parallel. If the strings have no spin, then it takes two strings to make a time machine. If spin *is* allowed, however, then just a single string will suffice for time travel: see S. Deser and R. Jackiw, "Time Travel?" *Comments on Nuclear and Particle Physics*, September 1992, pp. 337–354.

[71] D. Harari and A. P. Polychronakas, "Gravitational Time Delay Due to a Spinning String," *Physical Review D*, November 15, 1988, pp. 3320–3322.

[72] S. W. Hawking, "Gravitational Radiation from Collapsing Cosmic String Loops," *Physics Letters B*, August 23, 1990, pp. 36–38.

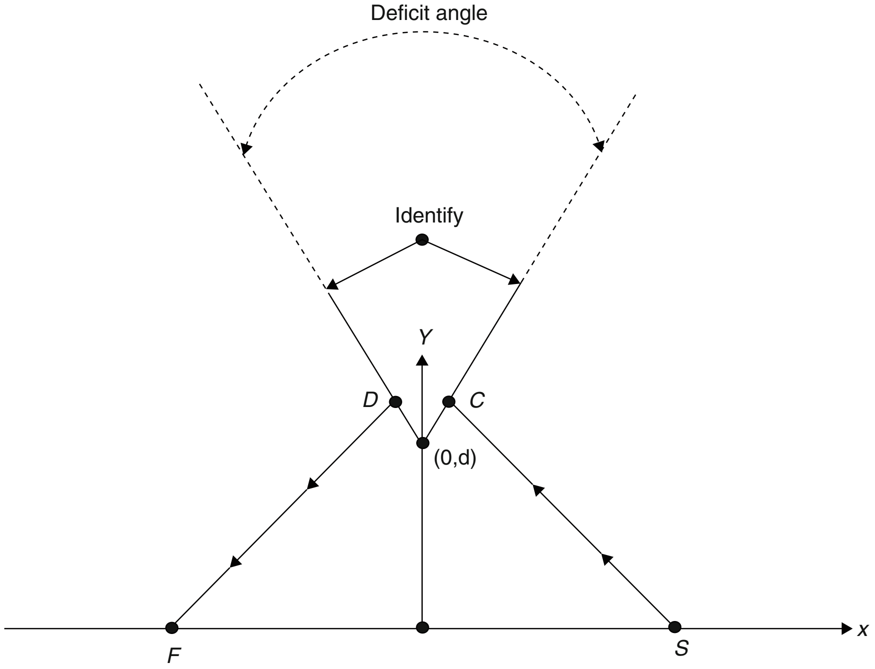

Fig. 6.12 The deficit angle in spacetime formed by a cosmic string

spin would banish the terrifying closed timelike curves. Subsequent analyses along those lines, however, continued to find the time travel implications intact.[73]

Here's how the cosmic string time machine works. In an earlier work,[74] published in 1985, Gott discovered that a cosmic string warps spacetime in a highly characteristic way, as shown in Fig. 6.12. A stationary cosmic string is imagined as perpendicular to the xy-plane (the plane of the page) and passing through the page at the point $(0, d)$ on the y-axis. The warp produced by the string is as though a wedge of angle 2α (this angle is called the *deficit angle*) were cut out of spacetime and the

---

[73] See, for example, B. Jensen, "Notes on Spinning Strings," *Classical and Quantum Gravity*, January 1992, pp. L7–L12, H. H. Soleng, "A Spinning String," *General Relativity and Gravitation*, January 1992, pp. 111–117, (the next two are in the *Physical Review D*), B. Jensen and H. H. Soleng, "General-Relativistic Model of a Spinning Cosmic String," May 15, 1992, pp. 3528–3533, and M. Novello and M. C. M. da Silva, "Cosmic Spinning String and Causal Protecting Capsules," January 15, 1994, pp. 825–830.

[74] J. R. Gott, "Gravitational Lensing Effects of Vacuum Strings: Exact Solutions," *The Astrophysical Journal*, January 15, 1985, pp. 422–427. Gott's discovery was independently reported in W. A. Hiscock, "Exact Gravitational Field of a String," *Physical Review D*, June 15, 1985, pp. 3288–3290.

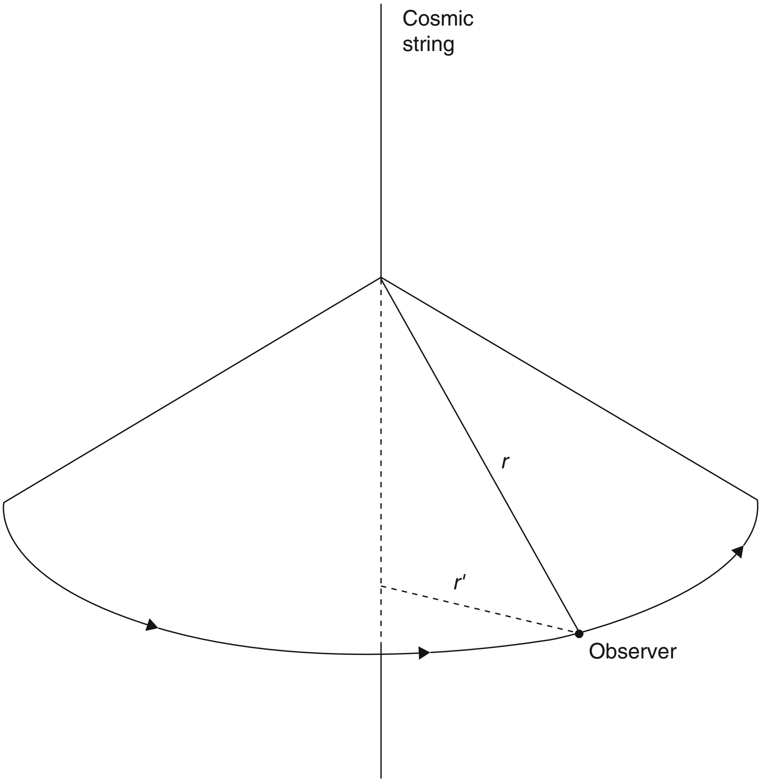

Fig. 6.13 The warped, conical spacetime around a cosmic string, An observer in this spacetime thinks she is distance r from the string, but a 'meta-observer' sees that she is actually distance $r^{'}$ from the string. Thus, if the observer follows a complete circular path around the string, she will travel a distance of $2\pi r^{'} < 2\pi r$. The observer in the spacetime will interpret this result by saying that the angle 2π is really 2π minus 'a deficit'

edges of the cut were then 'glued' together; for example, points C and D are identified as identical. The reason for the term *deficit angle* is that at radius r from the string, a circular path around the string has the reduced length $(2\pi - 2\alpha)r$, and not the usual $2\pi r$ (spacetime around the string, while *locally* flat, is actually 'conical,' as illustrated in Fig. 6.13).

The deficit angle is equal to $8\pi\mu$ radians (in a system of units where G, Newton's gravitational constant, is 1) if the linear mass-energy density μ is expressed in units of Planck masses per Planck length. For example, $\mu = 1$ corresponds to 1.35×10^{28} g/cm (think of something on the order of the mass of the Earth per inch of the string). For 'more typical' values—say a 'mere' $\mu = 10^{22}$ g/cm, $2\alpha = 0.001^{\circ}$. While Gott's paper had appeared 5 months before Hiscock's (see note 74), it is evident that Hiscock's work was done before he became aware of Gott's. Both papers treat exact derivations of the deficit angle but, in fact, the correct expression had actually been published 4 years earlier (but from a linearized form of the gravitational field equations and so the result was not as 'conclusive' as are Gott's and Hiscock's)[75]

Now, consider the two points S and F on the x-axis at $(x_0, 0)$ and $(-x_0, 0)$ in Fig. 6.12. Suppose we want to send photons from S to F. In normal, 'unwarped' spacetime, the direct path from S to F through the origin has length $2x_0$. There is also another possible path, however, S to C/D to F, that loops out and around the cosmic string. Indeed, this second path is simply the path a gravitationally lensed photon would take (an observer at F would see *two* images of S) and this—*not* time travel—is the issue that originally attracted Gott's attention to cosmic strings.[76] If the deficit angle were zero, then this alternative path would always be longer than $2x_0$, for any value of x_0. For the case of $2\alpha > 0$, however, if x_0 is large enough ($x_0 \gg d$), then it is possible for 'around the string and over the missing spacetime wedge' path to be shorter than the direct path.

The indirect path provides a way for a *sub*luminal trip (say, by rocket) from S to F to beat a photon traveling on the direct path. That is, the two events of the 'rocket leaving S ' and the 'rocket arriving at F' are spacelike separated. Thus, it is possible to find a moving frame of reference in which these two events are *reversed in temporal order*. In that frame of reference, the cosmic string (which is stationary in the reference frame of S and in that of F) will move—at speed v, say—in the $+x$ direction, and in that frame of reference the rocket will arrive at F *before* it leaves S.[77]

Then to complete the construction of a closed timelike path, simply repeat the process as shown in Fig. 6.14. That is, after the rocket arrives at F, have it turn around and fly back to S out-and-around and through the deficit angle spacetime warp due to a *second* cosmic string on the negative y-axis and perpendicular to the xy-plane. This second string is moving at speed $-v$ (that is, opposite to the first string), so the rocket will arrive at S before it leaves F. But that means it arrives at S before it leaves S; that is, the rocket has traveled into the past. In other words, the rocket has traveled into the past. This entire process is precisely the same idea behind the two-wormhole Roman-ring time machine discussion from the previous section.

Now, instead of having two oppositely moving reference frames, one in which the top, stationary string at $(0, d)$ appears to be moving at $+v$ and another frame in

---

[75] A. Vilenkin, "Gravitational Field of Vacuum Domain Walls and Strings," *Physical Review D*, February 15, 1981, pp. 852–857.

[76] Just as discussed earlier in the context of wormholes, gravitational lensing may offer a way to detect cosmic strings. See, for example, the two papers by D. L. Ossipov, "Diffraction of Light by a Cosmic String," November 1995, pp. 765–771, and "Contribution of Strings to the Observed Variability of Extragalactic Sources of Radiation," September 1996, pp. 419–425, both in *JETP Letters*.

[77] For this to happen, however, v must be very close to the speed of light. Gott (see note 69) showed that with $v = \tanh(\theta)$, the condition for the rocket to arrive back at S before it leaves S is $\cosh(\theta)\sin(\alpha) > 1$, where α is one-half the deficit angle. For $\mu = 10^{22}$ g/cm, this gives $v = 0.99999999995$ (times the speed of light).

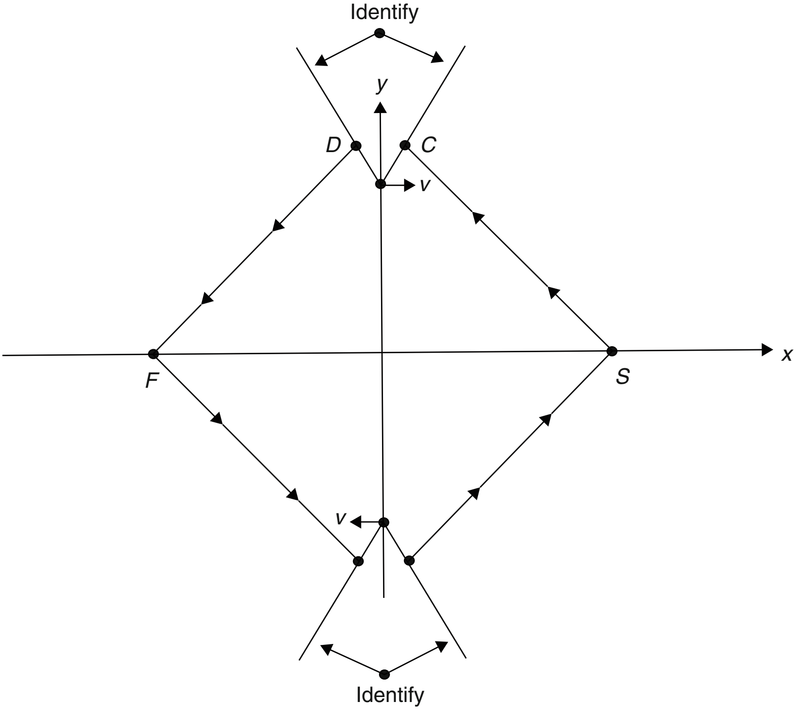

Fig. 6.14 Gott's spacetime, formed by joining two oppositely moving versions of the spacetime in Fig. 6.13

which the bottom, stationary string appears to be moving at –*v*, we can imagine an observer in the *stationary center-of mass* frame watching two strings that are moving at +*v* and –*v*. This leaves the situation unchanged, so in the center-of-mass frame the rocket does travel into the past, arriving back at *S* before it leaves *S*. That is, the rocket has traveled all the way around a closed timelike world line. Note, too, that the geometric condition mentioned earlier of $x_0 \gg d$ immediately implies that for x_0 *not* sufficiently large, there isn't a closed timelike path from *S* to *F* and then back to *S*; that is, there is a region in Gott's spacetime where such time travel journeys *cannot occur*.

Another physicist pursued Gott's analysis in an attempt to see whether these 'time travel paths' are *created* as the strings approach each other or, instead, if the paths exist at other times as well.[78] This important question gets to the idea of whether such time machine paths can be *intentionally created* by humans via a dynamical process (a *strong* time machine), or whether all such paths have existed since the formation of the universe (a *weak* time machine). This issue involves Hawking's chronology protection conjecture, which you'll recall asserts that the laws of physics will always (somehow) prevent the creation of a time machine. One reason Hawking repeatedly gave for believing the Conjecture is the apparent absence of time travelers from the future among us now (in their past). The only possible exception allowed by the Conjecture is the creation of closed timelike loops at the moment the universe was created (at that moment there *was no past* for time travelers to invade!). Ori proved that the closed timelike loops around Gott's cosmic strings are *always* present: that is, a time machine is *not created* by the near collision of the strings. Thus, Hawking's Conjecture is nor refuted by Gott's spacetime.

---

[78] A. Ori, "Rapidly Moving Cosmic Strings and Chronology Protection," *Physical Review D*, October 15, 1991, pp. 2214–2215.

One very curious issue is *where* the closed time loops are before the strings pass one another. As mentioned briefly by Ori, and further discussed by others,[79] the time loops are initially at spatial infinity. To this concern, Gott and a colleague made the following very strong reply:

> "[A problem] Deser *et al.* present with respect to the Gott spacetime is that it contains CTC's at spacelike infinity; this is supposed to be an unacceptable boundary condition. We wonder, however, how they know so much about boundary conditions at spacelike infinity. In our own Universe we do not know what spacelike infinity looks like (if it exists) since we have not seen it yet. We certainly have no way of knowing whether or not there are CTC's there. The working physicist is, of course, free to impose simple and convenient boundary conditions (e.g., asymptotic flatness) on a system in order to isolate and understand the processes occurring within it. *But boundary conditions are tools of physicists, and they should not be confused with laws of physics* [my emphasis]. There may be such laws of nature that restrict the possible structure of spacelike infinity, and even prohibit CTC's there, but in the absence of evidence such laws should not be postulated *ad hoc*."[80]

Still, as Ori had observed the year before, having time loops collapsing inward from infinity toward humans who might, fortuitously, wish to use them at *just the instant they so conveniently arrive*, is "a situation which has little to do with the creation of a time machine by a *human being* [my emphasis]."[81]

The most damning objection to Gott's cosmic string time machine came, ironically, from Gott himself. The two-string spacetime of Fig. 6.14 might actually, he and a colleague wrote (see Li and Gott, note 66) be destabilized by the non-zero mass of any would-be time traveler. They suggest that this concern could perhaps be 'solved' by assuming that the time traveler and her spaceship have a spherically symmetric mass distribution surrounded by a negative-mass shell to give zero net mass (and thus a zero net gravitational field that would *not* destroy the closed timelike curves of the strings). But that, they further observed, would negate the crucial advantage—no exotic matter and so no violation of the weak energy condition—that a cosmic string time machine enjoys over a wormhole time machine.

---

[79] See, for example, S. Deser, *et al*., "Physical Cosmic Strings Do Not Generate Closed Timelike Curves," *Physical Review Letters*, January 20, 1992, pp. 267–269.

[80] M. P. Headrick and J. R. Gott, "(2+1)-Dimensional Spacetimes Containing Closed Timelike Curves," *Physical Review D*, December 15, 1994, pp. 7244–7259.

[81] A. Ori, "Must Time-Machine Construction Violate the Weak Energy Condition?" *Physical Review Letters*, October 18, 1993, pp. 2517–2520.

Does a trip around a pair of cosmic strings present other problems *aside* from the sheer fantastic physics of the strings themselves? Well, I think “turning the rocket around at *F* and flying back to *S*” is a lot easier to write than it would be to actually do! The entire trip has to occur while the strings (moving at essentially light speed) are in a position *to be* flown around. As one character says to another in Stephen Baxter’s novel *Ring*, “Louise, the strings are traveling just under the speed of light—within three decimal places of it, actually. [Our ship is] traveling at a little over half-light speed. The turning curves, and the accelerations, are incredible . . .” I think so! And I do wonder who—*or what*(!)—is actually controlling a *maneuvering* rocket traveling faster than $\frac{1}{2}c$?

6.5 Cutting and Warping Spacetime

> “The warp drive spacetime of Alcubierre is impossible to set up . . . one needs to transcend the speed of light in order to construct the warp drive in the first place . . . put roughly, you need one to make one!”[82]

In this chapter we’ve talked about the physics of three specific time machine ‘implementations’: the rotating cylinder, the wormhole, and the cosmic string. One can also discuss the ‘construction’ of a time machine in a more geometrical (yet still physical) way by performing what is called *spacetime surgery* to arrive at what is often referred to as a *Deutsch-Politzer spacetime* (after the two physicists who are closely associated with it[83]). With this surgery we arrive at a simple spacetime picture of the grandfather paradox (as you’ll soon see).

We start with a flat, two-dimensional Minkowski spacetime, the x, t system in Fig. 6.15, and then imagine that (somehow) two *cuts* in that spacetime come into

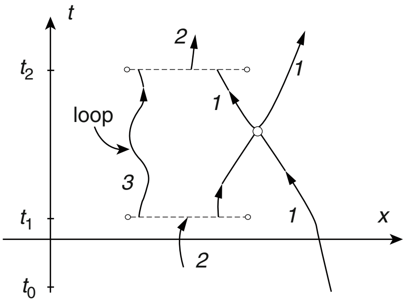

Fig. 6.15 Minkowski spacetime transformed into a time machine with two ‘cuts’

existence. These cuts are the two horizontal dashed lines in the figure, one with an arrowhead going into it, and one with an arrowhead coming out of it (each labeled with ‘2’). We further imagine that each cut has two edges, with the upper edge of the lower cut ‘glued’ to the lower edge of the upper cut (and the lower edge of the lower cut ‘glued’ to the upper edge of the upper cut

---

[82] D. H. Coule, “No Warp Drive,” *Classical and Quantum Gravity*, August 1998, pp. 2523–2527, offering a pessimistic view of warp drive.

[83] See D. Deutsch, “Quantum Mechanics Near Closed Timelike Lines,” November 15, 1991, pp. 3197–3217, and H. D. Politzer, “Simple Quantum Systems with Closed Timelike Curves,” November 15, 1992, pp. 4470–4476, both in *Physical Review D*.

These ‘gluing’s’ explain why the arrowhead marked 2 into the lower cut at time t_1 emerges from the upper cut at time $t_2 > t_1$, and why the arrowhead marked 1 into the upper cut at time t_2 emerges from the lower cut at time $t_1 < t_2$. The 2-line is the world line of a particle that simply disappears from spacetime during $t_1 < \mathrm{t} < t_2$, while the 3-line is the world line of a particle trapped in an endless time loop (remember the 1993 film *Groundhog Day*?). Clearly, the sub-region of spacetime between the two cuts is not ‘normal’ spacetime. In fact, the 1-line shows that we have encountered a time machine spacetime, as a particle entering it from $t < t_1$ (passing to the right of the lower cut) can enter the lower edge of upper cut at $t = t_2$ and so emerge from the upper edge of the lower cut at the *earlier* time $t = t_1$ and thus interact with itself before it entered the upper edge! And that, of course, sets-up a grandfather paradox situation.

This picture leaves one thing obviously (and glaringly) unexplained—just *how* does one *cut and glue* spacetime? The reason, in spite of that question, that physicists nonetheless study situations depicted in Fig. 6.15, is because it allows them to explore what *could* happen *if* through some (yet unknown) process a time machine spacetime should suddenly appear. Who says physicists aren’t optimists?

The idea of modifying spacetime itself to ‘make’ a time machine (look back in Sect. 3.5, at the discussion there on what it means to solve the gravitational field equations) has also appeared in connection with another of science fiction’s favorite ideas, one almost as spectacular as time travel: *the FTL warp drive*. The lure of interstellar FTL travel, for both science fiction enthusiasts and physicists is, of course, simply undeniable. Consider, for example, these words by a Russian physicist:

> “Everybody knows that nothing can move faster than light. The regrettable consequences of this fact are also well known. Most of the interesting or promising candidates for colonization are so distant from us that the light barrier seems to make an insurmountable obstacle for any expedition. It is, for example, 200 pc [1 parsec is equal to about 3.2 light-years] from us to the Pole star, 500 pc to Deneb [the brightest star in the constellation Cygnus], and ~10 kpc to the center of the Galaxy, not to mention other galaxies (hundreds of kiloparsecs). It makes no sense to send an expedition if we know that thousands of years will elapse before we receive its report. On the other hand, the prospects of being confined forever to the Solar System without any hope of visiting other civilizations or examining closely black holes, supergiants, and other marvels are so gloomy that it seems necessary to search for some way out.”[84]

---

[84] S. V. Krasnikov, “Hyperfast Travel in General Relativity,” *Physical Review D*, April 15, 1998, pp. 4760–4766. Possible travel distances have been greatly reduced since Krasnikov wrote. In 2011, for example, astronomers announced the discovery of a red dwarf star with three planets (each of mass comparable to Earth’s), all in the star’s so-called *habitable zone* (where water can exist on the surface in the liquid state). All three planets are solid (not gaseous as are Jupiter,

In response to that, we might ask if *FTL* trips will someday be made by humans in spaceships? *Maybe*—but only if such journeys can be made in a very *unordinary* spacetime. That is, continuing with Krasnikov's passage:

> "The point . . . is that [whereas the light barrier exists in special relativity] in general relativity one can try to change the time necessary for some travel not only by varying one's speed [as in special relativity] but also . . . by changing the distance one has to cover."

To understand what Krasnikov was getting at, let's consider the theoretical analysis made 4 years earlier by the Mexican mathematical physicist Miguel Alcubierre on, astonishingly, how to make a *Star Trek* warp drive![85] He did this by demonstrating a spacetime metric that, by literally expanding and contracting the local spacetime of a spaceship and its neighborhood, achieves space travel between any two points, no matter how far apart, in arbitrarily little elapsed time (for both the spaceship, and external non-spaceship observers, there is no time dilation effect[86]).

Alcubierre opened his analysis with words designed to explain how FTL travel is possible, given all that I've told you earlier in this book about how FTL travel is *not* possible (according to *special* relativity). As he wrote,

> "Since our everyday experience is based on a Euclidean space, it is natural to believe that if nothing can locally travel faster than light then given two places that are separated by a proper spatial distance D, it is impossible to make a round trip between them in a time less than $2D/c$ (where c is the speed of light), as measured by an observer that always remains at the place of departure. Of course, from our knowledge of special relativity we know that the time measured by the person making the round trip can be made arbitrarily small if his (or her) speed approaches that of light. However, the fact within the framework of general relativity and without the need to introduce non-trivial topologies (wormholes), one can actually make such a round trip in an arbitrarily short time as measured by an observer that remained at rest will probably come as a surprise to many people."

That last sentence is almost surely a grand understatement, and Alcubierre quickly went on to explain.

> "The basic idea can be more easily understood if we think for a moment of the inflationary phase of the early Universe, and consider the relative speed of separation of two co-moving observers. It is easy to convince oneself that, if we define this relative speed as the rate of change of proper spatial distance over proper time, we will obtain a value that is much larger than the speed of light. This doesn't mean that our observers will be travelling faster

---

Saturn, Neptune and Uranus) and so, as potentially habitable, are candidates for a visit. The star and its planets are 'only' 22 light-years from Earth. In *Star Trek*, FTL speed is described by the *warp factor*, which is the cube-root of the multiple of the speed of light at which the spaceship *Enterprise* travels. So, for example, to make the journey from Earth to the red dwarf in one month of ship time (see ahead also to note 86), the required FTL speed would be 264 times the speed of light, or warp factor 6.4. In science fiction, a *warp drive* is imagined as the means for achieving such speeds.

[85] M. Alcubierre, "The Warp Drive: Hyper-fast Travel Within General Relativity," *Classical and Quantum Gravity*, May 1994, pp. L73–L77.

[86] That is, the passage of time on the spaceship is *identical* with the passage of time on Earth. *With the warp drive, there is no twin paradox.*

> than light: *they always move inside their local light-cones* [my emphasis]. The enormous speed of separation comes from the expansion of spacetime itself."

In a similar fashion, a *contraction* of spacetime can result in being able to *approach* an object at FTL speed.

In fact, we've actually already encountered one way to obey special relativity's *local* limit on speeds to that of light, while still achieving superluminal speed on a *global* level. That is, we *could* do that *if* general relativity really does allow wormholes in spacetime (what Alcubierre calls a "non-trivial topology"). That's because we can imagine a wormhole connecting two points in space that are light-years apart in that space, and yet the distance *through* the wormhole itself is quite short. Thus, a spaceship transiting the wormhole could do so at *sub*luminal speed at all times, and yet to an observer in normal space the speed would appear to be far in excess of the speed of light.

Determining just *how* to achieve the spacetime warp, however, is far different from simply demonstrating that such a warp is consistent with the general theory of relativity. The 1996 movie *Star Trek: First Contact*, for example, is about the invention of the warp drive in the twenty-first century. The whole thing fits inside a discarded ICBM which, as you'll soon see, is a *vast* underestimation of the technology required to control the energies associated with a real warp drive. For a spacetime engineer to *build* the warp drive bubble means she has to determine the required mass-energy distribution that results in Alcubierre's assumed spacetime metric. And that brings us to the central problem of the warp drive—the warp drive engine of an FTL starship requires (just like a wormhole) exotic matter (negative energy)—stuff that violates all the usually assumed energy conditions of general relativity.[87]

The weak, strong, and dominant energy conditions are *all* violated because the Alcubierre spacetime warp requires a negative energy density in the 'skin' of the warp bubble. As discussed earlier, in connection with wormholes, negative energy density *can* be achieved on a *micro*scopic scale, but for the Alcubierre warp drive we are talking about a *lot* of exotic matter. In their paper, Pfenning and Ford calculated that, for what they called "a macroscopically useful warp drive" with a radius of 100 m "so that we may fit a ship inside [the warp bubble]," the negative energy required for the warp bubble is on the order of, as they so graphically put it, "roughly ten orders of magnitude greater than the [energy of the] total mass of the entire universe."

Two years later, after making some adjustments to the spacetime metric assumed by Alcubierre (that is, to the distribution of mass-energy to produce the warp bubble), it was shown that the negative energy required by the warp drive could be greatly reduced.[88] The reduction is, in fact, spectacular, but only in a relative sense (you can reduce a mass that is ten orders of magnitude greater than that of the entire universe by a *huge* factor and still be left with a pretty stupendous number). The reduced amount of negative energy required for a warp bubble able to contain a human-sized spaceship is now down to 'only' "of the order of a few solar masses [-1.4×10^{30} kg]."

---

[87] M. J. Pfenning and L. H. Ford, "The Unphysical Nature of 'Warp Drive'," *Classical and Quantum Gravity*, July 1997, pp. 1743–1751. See also K. D. Olum, "Superluminal Travel Requires Negative Energies," *Physical Review Letters*, October 26, 1998, pp. 3567–3570.

As mentioned earlier, in connection with the Casimir effect and its theoretical use in a wormhole time machine, although quantum field theory does not preclude negative energy densities, that does not mean it is possible to observe arbitrarily large negative densities for arbitrarily long times. In fact, certain quantum inequalities (QI's), much like Heisenberg's uncertainty principle, have been established that place bounds on the magnitude and duration of observable negative energy density.[89] These QI's have the general form of

$$\widehat{\rho} t_0^4 > -C$$

where C is a positive constant that depends on the nature of the particular quantum field being considered, t_0 is the time duration, and $\widehat{\rho}$ is the integrated energy density along a finite section of a geodesic (free fall) world line. The form of the QI shows that as t_0 increases, $\widehat{\rho}$ must quickly decrease. For example, if t_0 doubles, then $\widehat{\rho}$ must decrease by a factor of *sixteen*, a result that caused Ford and Roman to conclude that it "appears probable that nature will always prevent us from producing gross macroscopic effects with negative energy."

'When it rains it pours,' goes an old saying, and that applies to the warp drive's potential difficulties: in addition to the need for exotic matter, there are two more concerns as well, both *operational* in detail. First, running into any space matter encountered by the leading edges of the warp bubble (where spacetime is shrinking), such as interstellar dust, would certainly generate intense radiation. The ship, then, should carry plenty of shielding which, curiously, would not be a problem because the energy density of the warp, itself, is *independent* of the mass in the bubble's interior. In any case, the warp drive should clearly not be engaged anywhere near any sizeable chunk of matter, like a planet (and, indeed, that constraint was followed in *Star Trek*). Second an even more severe problem was discovered by Krasnikov. In unpublished work he showed that the ship at the center of the bubble is not causally connected to the edges of the bubble. That is, the ship's crew could *not* create a warp bubble on demand and, after it had been created, could not control it on demand.[90]

---

[88] C. van den Broeck, "A Warp Drive with More Reasonable Total Energy Requirements," *Classical and Quantum Gravity*, December 1999, pp. 3973–3979.

[89] L. H. Ford and T. A. Roman, "Restrictions on Negative Energy Density in Flat Spacetime," *Physical Review D*, February 15, 1997, pp. 2082–2089. For some interesting remarks about the QI's, see J. F. Woodward, "Twists of Fate: Can We Make Traversable Wormholes in Spacetime?" *Foundations of Physics Letters*, April 1997, pp. 153–181.

It is important to both understand what that means, as well as what it does *not* mean. The causality issue does *not* mean that Alcubierre warp bubbles are impossible to create (perhaps they are, but not because of a lack of causality). It only means that whatever action is required to change the spacetime metric to make a warp bubble *has to already have been done before the decision to use the bubble is made*. Thus, a warp bubble wouldn't be of any use for a starship that needs to escape a sudden, unexpected threat. But, as Everett and Roman cautiously observe, the warp bubble might have a more mundane use: "Suppose space has been warped to create a bubble traveling from Earth to some distant star, e.g., Deneb, at superluminal speed. A spaceship, appropriately located with respect to the bubble trajectory, could then choose to enter the bubble, rather like a passenger catching a passing trolley car, and thus make the superluminal journey."

At the end of his paper (note 85) Alcubierre briefly speculated on the possibility of using his superluminal warp drive to build a time machine (showing, again, the intimate connection between the two concepts—look again at note 12 in Chap. 1 for how the connection between FTL and time travel appeared in pulp science fiction), but didn't show *how*. That was done 2 years later by Everett using, not surprisingly, an argument he called "reminiscent of the 'reinterpretation principle' ... which played an important role in discussions of the physics of tachyons."[91]

In an attempt to avoid the Alcubierre bubble's causality problem, Krasnikov looked for a different, causal superluminal spacetime metric. This he succeeded in finding[92] but, rather than describing a bubble, Krasnikov's warp is in the shape of a tube. The interior of the tube is flat spacetime, just as in the case of the bubble warp, but unlike the bubble there would be a causal link between the spaceship crew and the tube. Just as the warp bubble requires very thin walls (on the order of a few thousand Planck lengths) of negative energy, so does the Krasnikov tube warp. Unlike the bubble warp, however, the tube warp stretches the entire length of any proposed trip, so the total negative energy in the warp is incredibly huge. For a tube a mere one meter long and one meter wide, for example, the total negative energy is 10^{28} solar masses, and to create a tube from Earth to just the nearest star would require 10^{44} solar masses of negative energy![93]

One curious feature of the Krasnikov warp is that the *outbound* leg of a round trip cannot be made in less time than required by light. *But* on the on the return half of the journey, a traveler would find the spacetime metric so altered (because of mass-energy manipulations purposely made on the outbound half) that she would move "backwards in time." The net result is that the round trip could end arbitrarily soon after it started! As Everett and Roman cautiously concluded, the Krasnikov tube is a "very unlikely possibility," but it would make a *wonderful* science fiction gadget, don't you think?

---

[90] A. E. Everett and T. A. Roman, "Superluminal Subway: the Krasnikov tube," *Physical Review D*, August 15, 1997, pp. 2100–2108.

[91] A. E. Everett, "Warp Drive and Causality," *Physical Review D*, June 15, 1996, pp. 7365–7368. Recall the discussion of the RP in Chap. 5.

[92] See notes 84 and 90.

[93] In the same manner as the huge negative energy of the Alcubierre warp drive was later reduced (see note 88), the Krasnikov tube's enormous negative energy requirement was later significantly reduced: see P. Gravel and J. Plante, "Simple and Double Walled Krasnikov Tubes I: tubes with low mass," *Classical and Quantum Gravity*, February 2004, pp. L7–L9.

While the Alcubierre warp may seem to be an incredible discovery (it *is*), it was not a unique one. That's because just 8 years later a different warp metric was discovered by the Portuguese mathematician José Natário, in which the expansion/ contraction of spacetime does *not* occur. As Natário wrote, this signature feature of the Alcubierre warp drive "is but a marginal consequence of the choice [for the spacetime metric/mass-energy distribution]."[94] Rather than thinking of the warp bubble as being propelled by the push-pull of spacetime expansion/contraction, Natário wrote that "one could best describe the warp-drive spacetime as 'sliding' the warp-bubble region through space": that is, as analogous to a California surfer riding a wavefront. The surfer is motionless with respect to the water in the immediate vicinity of his board, and yet his speed with respect to the rapidly approaching shore is decidedly non-zero.

The idea of a manipulated or *warped* spacetime allowing time travel was an early arrival in science fiction. Consider, for example, the 1930(!) story in which the narrator (one Thomas Jenkins) walks 18,000 years into the future. In an editorial footnote (a device commonly used in early pulp fiction to inject scientific verisimilitude), we are told that "Jenkins had evidently fallen into a warp in space . . . a fault, we might say, borrowing a geologic term, in the curvature of space. Through this warp he had been thrown clear out of our three dimensions into a fourth. There he slid in time over to the other side [of the fault] into the same spot in the three-dimensional world, but into a different era in time."[95]

That was a flawed explanation, with its talk of space rather than of spacetime, but some authors eventually learned to do better. For example, *folded spacetime* as a mechanism for time travel is used in a 1940s cautionary tale on the potential horrors of the atomic bomb. In that story, published 2 years after the atomic bombings in Japan, the world 15 years hence experiences a terrible atomic war. As the time traveler in the tale explains, "During the unprecedented release of atomic energy that arouse during the simultaneous bombings of our cities, something happened to the very continuum in which we exist . . . A crook, a twist, a fold—explain it how you will, I accidently stumbled upon an electronic circuit that would create a field that would enable passage from one folded section [of spacetime] to the adjacent section. The fold proved to be about fifteen years in length . . . ."[96]

---

[94] J. Natário, "Warp Drive with Zero Expansion," *Classical and Quantum Gravity*, March 2002, pp. 1157–1165.

[95] N. Schachner and A. L. Zagat, "In 20,000 A.D.," *Wonder Stories*, September 1930.

[96] R. F. Jones, "Pete Can Fix It," *Astounding Science Fiction*, February 1947.

As Alcubierre and Natário showed, it will take a lot more than a mere electronic circuit to warp spacetime for either FTL or time travel, but at least even the early science fiction pulp writers understood that—*somehow*—a spacetime warping would be required.

The Alcubierre FTL warp drive appeared as the scientific basis for a modern science fiction novel, where at one point we read of a curious optical feature of an FTL spaceship, one I haven't seen mentioned in the physics literature: "This is a ship that traveled faster than light. It's visible as it travels; its warp bubble emits a cascade of exotic radiation . . . but it outruns its own image. So the ship arrives first and the light has to catch up, all the photons it emitted back along its path arriving at mere light speed. The older images arrive last, and you get this effect as if the ship was receding, not arriving."[97]

Another quite interesting feature of Alcubierre's warp drive is that the spaceship crew would experience no acceleration forces, as the ship is always in *free fall*. This may explain why the *Enterprise* crew isn't flattened when Mr. Sulu engages that ship's warp drive. The spaceship is surrounded by a "bubble" of warped spacetime that is swept along by the combined push-pull effect of the expanding spacetime behind the craft and the shrinking spacetime in the front. The ship, itself, resides in the *flat* spacetime interior of the warp bubble. An amusing way to think of this is to imagine a fish (space traveler), inside an aquarium (the warp bubble), which has been tossed into a swiftly flowing river. An observer at the edge of the river sees the aquarium move by her at high speed while, for the fish (swimming in the *still* waters of its aquarium), all is serene because *it is at rest with respect to its local environment*. Thus, the Alcubierre warp drive realizes yet another one of science fiction's wonderful gadgets: the *reactionless* spaceship drive. That is, "the warp bubble moves by interacting with the geometry of spacetime instead of expending reaction mass [as do jet and rocket engines] . . . and the spaceship is simply carried along with it."[98] In picturesque terms, the warp drive starship is like a surfer who makes her own waves.

And so we see, with each passing decade, more and more of science fiction departing from the make-believe to the pages of physics journals.

---

[97] S. Baxter, *Ark*, Gollancz 2009. The spaceship in this work travels at three times the speed of light (warp factor 1.44, as explained in note 84).

[98] F. S. N. Lobo and M. Visser, "Fundamental Limitations on 'Warp Drive' Spacetimes," *Classical and Quantum Gravity*, December 2004, pp. 5871–5892.

6.6 For Further Discussion

The connection between FTL speeds and backward time travel made the jump from theoretical physics to popular culture very quickly. It was in the British humor weekly *Punch*, for example, that the famous (but nearly always misquoted—see note 113 in Chap. 3, which doesn't have it quite right) limerick by A. H. R. Buller (1874–1944) first appeared (December 19, 1923, p. 591):

> *"There was a young lady named Bright Whose speed was far faster than light, She set out one day In a relative way And returned on the previous night."*

Where *Punch* dared to go, Hollywood could not be far behind. Indeed, in this case it was actually there first, with the 1922 one-reel silent comedy film *The Sky Splitter*. This was just a short film (feature pictures generally had at least four reels), so it isn't clear just how widely distributed and viewed it may have been. The story is that of a scientist testing a new spaceship; when it exceeds the speed of light, he begins to relive his life. This all shows that today's fascination, so common in popular culture, of the latest developments in theoretical physics, is nothing new. Why do you think this is so? That is, why (for example) do so many of those who flock to science fiction movies of interstellar invasions (like the 1996 *Independence Day* and its 2016 sequel), nonetheless have no conception of the unlikely possibility of such invasions because of the sheer magnitude of interstellar distances? Distances so immense that, even at the speed of light, it takes 4 *years* to travel to the Sun's nearest stellar neighbor, and *millions* of years to reach the Milky Way's nearest neighboring galaxy? (The vastness of interstellar distances is, as mentioned in the text, the reason for the fascination in warp drives in both science fiction and physics.).

In wormhole and cosmic string time machines, and with warp drives, we encountered the idea of negative mass-energy in the form of 'exotic matter' (see note 39 again). Something like negative mass actually appeared in fiction long ago, in the 1827 novel *A Voyage to the Moon* by "Joseph Atterly," a pseudonym for George Tucker, a professor of moral philosophy at the University of Virginia. (One of Tucker's students was Edgar Allen Poe, who almost surely was influenced by Tucker's book to write his own moon tale, the 1835 "The Unparalleled Adventure of One Hans Pfall.") The trip in Tucker's work was powered by a metal called *lunarium*, which repels Earth.

(continued)

This is *not* the same sort of stuff as Wells' "Cavorite," a metallic alloy that is "transparent" to gravity and that appears in his 1901 *The First Men in the Moon*. Wells' competitor in the 'scientific romance' genre was, of course, Jules Verne. Wells was a visionary who looked far beyond just the next few decades, while Verne was a 'practical engineer' who, for example, got *his* characters to the Moon by the direct method of simply shooting them out of a 900-foot long cannon with 400,000 pounds of guncotton! (Wells' vision *could* sometimes fail him, as it did about the imminent likelihood of airplanes in his 1901 *Anticipations*. He believed they *would* be developed by the year 2000, and maybe even before 1950, but of course just 2 years later . . .) In a 1903 magazine interview, Verne revealed how he felt about the difference between his and Wells' work: "It occurs to me that his stories do not repose on very scientific bases . . . He goes to Mars [sic] in an airship, which he constructs of a metal which does away with the law of gravitation. C*'est très joli* [this is all very nice], but show me the metal. Let him produce it." Today the cry from those who dislike wormholes is the Verne-like 'show us the exotic matter!' If Wells and Verne were writing today, how do you think each would respond to that challenge? Would the possible existence (or not) of exotic matter be an issue about which both would *agree*?

Write a time-loop short story based on Lossev and Novikov's idea of a 'very old spacecraft' interacting with a remote wormhole.

Imagine an electronic circuit **A** that has the following behavior: **A**'s input signal is a function of time that has a well-defined maximum value (what electrical engineers call the *peak value*). The circuit's output signal, produced in response to the input, also has a well-defined peak value. Now, imagine further that the output peak occurs *before* the input peak. There is, in fact, nothing paradoxical or impossible about that, and such a circuit can (and has) been constructed, as I'll tell you shortly. Next, suppose that we take **A**'s output signal and use it as the input to another circuit **B** that, when it's input exceeds a certain level, disconnects the input to **A** *before* that input reaches *its* peak value. Circuit **B** can also be constructed in the real world. Indeed, you can read about how to construct **A** and **B** in two papers by M. W. Mitchell and R. Y. Chiao, "Causality and Negative Group Delays in a Simple Bandpass

(continued)

Amplifier," *American Journal of Physics*, January 1998, pp. 14–19, and "Negative Group Delay and 'Fronts' in a Causal System: An Experiment With Very Low Frequency Bandpass Amplifiers," *Physics Letters A*, June 16, 1997, pp. 133–138. What makes all this interesting here is that this "seems to open the way for a variant of the time travel paradox in which the traveler journeys to the past and kills his grandfather before his own father is born," an observation made in Garrison *et al.*, "Superluminal Signals: Causal Loop Paradoxes Revisited," *Physics Letters A*, August 10, 1998, pp. 19–25. This electronic version of the grandfather paradox does indeed follow if one substitutes "input peak" for "grandfather" and "output peak" for "time traveler." But before you think this gadget is a time machine, be assured that its designers also showed that, unlike the causally related grandfather and time traveler, the two peaks are *not* so related. Read these three papers and write a summary report of how circuits **A** and **B** work, and why the two peaks are not so related.

You'll recall that "an advanced civilization" is thought to be required to create a useable wormhole (note 27). The common phrase used by astrophysicists who are interested in the possibility of extraterrestrial life is *arbitrarily advanced civilization*, with a distinction made for at least three progressively higher stages of 'advancement.' Very roughly, Types I, II, and III advanced civilizations are those that, respectively, have the technology to (a) control something like 10^{13} W (ten million megawatts) for massive interstellar radio broadcasts, (b) a technology to control the energy output of the civilization's planet's parent star (10^{27} W), and (c) a technology to control the energy output of the civilization's home galaxy (10^{38} W). We are, today, short of being even a Type I civilization, and it would probably take *at least* a Type III civilization to build a wormhole. Indeed, Stephen Baxter's 1993 novel *Timelike Infinity*, of beings who can manipulate *constellations of galaxies*, seems to assume a Type IV civilization will be required. Since there are typically 10^{11} stars in a galaxy, going from 10^{27} W for a star to 10^{38} watts for a galaxy is consistent. But where do astrophysicists get 10^{27} W for a single star? Here's a calculation for you to perform, to confirm this value for yourself, starting with the experimental fact (not difficult to repeat, as it's at the level of a junior high school science fair project using a solar cell and a few common electrical components) that the solar power level at Earth's equator is 1200 W per m^2. Then, using the fact that the Earth's orbital radius around the Sun is 93 million miles, compute the total power (energy per second) radiated by the Sun. (You should get a number that is somewhat

(continued)

smaller than 10^{27} W, which is ‘explained’ by observing that the Sun is really a quite ordinary star, exceeded in size by many other stars in the Milky Way. Next, you’ll find, in most books on astrophysics, the statement that the nuclear fusion reactions that power the Sun convert *four million tons* of the Sun’s mass to pure energy *every second*. Confirm that your number for the power output of the Sun is consistent with that claim. (Remember Einstein’s famous formula $E = \mathrm{m}c^2$, that the speed of light is $c = 3 \times 10^8$ m/s, and that 1 kg $\approx$ 2.2 pounds. In the MKS system of units (meters/kilograms/seconds) one watt = one joule (of energy) per second, where to give you some perspective on what a joule is, the chemical energy released by burning a gallon of gasoline is about 100 MJ.

Appendix A
Old Friends Across Time (A Story)[1]

As I sit here in my study, with the photographic evidence spread before me, I can barely comprehend what my eyes tell me must be so. The evidence is incontestable. And yet—I still struggle to believe. Let me try to explain—possibly in the process I will manage to put my tumbling mind to rest.

For as long as I can recall, old photographs have fascinated me. To page slowly through collections of historical pictures, no matter what the theme, was consummate joy. Even when I was quite a small boy I used them as my time machine into the past. They took me up and away from the problems every youngster has while growing up, and let me wonder of people and places long since returned to dust. Matthew Brady's Civil War photos had a particularly strong attraction for me, with the horror (and yes, I will admit it, the *fascination*) of war frozen in the images of young men dead before life had really begun. To look at the fallen youth of more than a century before, and to wonder who they were, and what they had felt and thought—it all sent shivers through my romantic mind.

I suppose I might have become a professional photographer. But somewhere along in the process of looking at pictures, I became aware of the miracle of the *technology* of picture taking. That led me to chemistry and optics, and finally by some wondrous route, I became an electrical engineer. I never lost my love for old pictures, though, but merely turned my interest in them to the photographic history of electrical physics.

To search out and acquire (for by now I had started my own collection) a photograph of Steinmetz, smoldering cigar clamped in his mouth, giving a lecture on AC circuit analysis using the then still mysterious square root of minus one made my heart beat faster. To find a faded picture of Einstein at a long forgotten

---

[1] P. J. Nahin, "Old Friends Across Time," *Analog Science Fiction Magazine*, May 1979. This tale was written with the specific goal of illustrating how a trip into the past *yet to be initiated* could logically influence events in the time traveler's present and future. The story reproduced here is, with only a few very minor alterations, as it originally appeared in *Analog*.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

conference, caught forever in time with his quiet, gentle eyes looking into mine, would send me to the heights of ecstasy.[2]

But it was Maxwell that led me to my incredible discovery. There is no doubt but that James Maxwell was the greatest theoretical physicist of the nineteenth century. Together with Einstein, he was the best of *any* century. Could it possibly be more than mere chance that the same year saw the death of one and the birth of the other? It was Maxwell who gathered together all the then known, but fragmented, experimental bits and pieces of knowledge about electricity and magnetism, and stirred in his own contribution of the displacement current. There was no physical evidence then to justify that last step, but the genius of Maxwell knew it *had* to be. And then, from his soaring mathematical insight and physical intuition, he took it all and wrote down the four magnificent equations for the electromagnetic field![3]

No one who has seen and understood those beautiful equations can come away without a quickening of the pulse and a flush of the blood. They're not long—you can write all four vector differential equations on the back of a postcard, but oh, what they tell us! With them, Maxwell, showed light was electrical in nature, predicted radio waves *two decades* before Hertz discovered them in the lab, explained energy propagation in space, and radiation pressure, and laid the scientific basis for today's television, radar, lasers, giant electric motors, generators, transmission lines and—well, why go on? The equations are the work of a level of genius we may not see again for a millennium. We have hardly begun to discover the marvels wrapped inside the electromagnetic field equations. With their aid, and that of quantum mechanics, the very secret of life, itself, may someday be unraveled.

And so I searched for old photographs of Maxwell. He died at his family's Scottish home in 1879, before the art of picture taking was barely 40 years old. But I knew in my heart that somewhere there *must* be photographs, yet undiscovered, of such a great man. Anyone who has seen the best examples of prints from wet glass collodion negatives knows they are, in the faithfulness of their rendition of detail, better than what we commonly expect today. Working against me was the fact that the process was slow, laborious, and unforgiving of mistakes. The taking of a picture was not a minor decision in Maxwell's time. But still I searched.

I searched for one photo, in particular. When Einstein died, a famous picture was taken of his office, just the way he left it for the last time. On the blackboard behind his desk are the last thoughts he had in his long quest for a Unified Field Theory, a ‘theory of everything.’ The writing on the papers covering the desk is clearly legible, and with modern blowup methods, easily readable.

---

[2] The first reference is to Charles Steinmetz (1865–1923), the German-born American electrical engineer and mathematician who became the wunderkind of General Electric. Einstein, of course, needs no introduction!

[3] I wrote this for story effect, but it's not really *quite* true. When Maxwell wrote his theory in mathematical form, he did so using *twenty* (!) equations in as many variables. The equations, as physicists and electrical engineers use them today, were first written in 1885 by the English self-taught eccentric Oliver Heaviside, who considered Maxwell to be his hero (see note 6 in Chap. 6). Modern electrical engineers and physicists write the Maxwell equations as *four* partial differential *vector* equations.

At the time of his death, Maxwell was the Einstein of his times. Surely, I reasoned, a similar photograph of Maxwell’s study must have been taken. Even though none has come down through the decades to us, it *must* exist! Gathering dust in an old trunk, or buried in a long forgotten album, it had to be somewhere. I vowed to find it.

I began by writing to all of Maxwell’s living descendants, asking that they search through family holdings for any pictures concerning Maxwell that they might possess. For the most part all were cooperative, even though more than just a few thought I was somewhat deranged. Still, it was in vain. I did receive a few old pictures never before seen by other than the family, including a poignant one taken in 1901, showing Maxwell’s grave in Parton Churchyard at Glenlair, Scotland. A forlorn, wintry scene, with only what seemed to be three men in the far distance, it brought tears to my eyes. Alas, there were no photos of Maxwell’s study.

But then late last year, while on a business trip to London, I stopped off for a few hours at the historical archives maintained by the British Institute of Electrical Engineers. On a chance, I looked through their massive files on Maxwell and was rewarded within the hour! What I found will haunt me throughout the remainder of my life.

There it was, stuck through its border with a rusty pin, between two pieces of yellowed paper covered with what appeared to be some simple, rough lecture notes. An ordinary looking photo of a study. Obviously overlooked through the years, or at best unappreciated for what it was, it was the almost illegible, penciled notation on the back that convinced me of my find—just a date: November 9, 1879. Exactly 4 days after Maxwell’s death, precisely when some unknown, yet inspired person (a family member, a neighbor, a local scientist?) would take such a picture!

I am ashamed to admit it, but there was no hope the Institute would let me have the picture. And there was no time to copy it, for I was to return home to America that very night. No, that’s not true. The *real* reason for what I did was simply that I *had* to have that original, *old* photo. I took it! It was my undoing, for that dishonorable act destroyed the picture’s tie to verified, legitimate historical records. But *I* know what I found is true.

I could barely control my wild emotions on the flight home. Several times I removed the picture from my briefcase, and looked with fascination at the papers lying on Maxwell’s desk, and at the tightly written lines of mathematics on the blackboard in the background. My hands trembled with what can only be called lust—once home, reunited with my well-equipped photo lab, I would learn every secret hidden in that picture.

There are no words I know that can convey the thrill I felt as I began the processing of that priceless photo. Alone in my lab, with all the modern equipment a well-off amateur can buy (a Caesar Saltzman 8 × 10 enlarger with mercury vapor point light source and a 10× Plan Achromat Nikon enlarging lens), I carefully cropped and blew up selected views of the blackboard and desk. Printing the enlargements on ultra-fine grain AGFA Brovira paper, I could scarcely restrain myself from peering at them with a magnifying glass while I waited for them to dry.

Then, at last, I had them spread out across my study desk. I tried to force myself to examine each slowly, carefully, in turn, and not to skip from one to another like a child let loose in a candy store with a dollar. The first three were of the desk papers, including what seemed to be a diary. It must have been lost after Maxwell's death since no trace of it exists in the historical records. I experienced a stunning thrill as I gazed upon the scrawled words, but as they were not easily read at once, I moved on. It was the sixth enlargement, of the upper right corner of the blackboard that sent me reeling back to my chair. An equation that shouldn't, no, *couldn't*, be there. But it was.

To understand my reaction, there is one astounding thing you must realize about Maxwell's field equations. When Einstein turned the world of physics on its head in 1905 with his famous paper, "On the Electrodynamics of Moving Bodies," all the old ideas about absolute motion and simultaneity of events went out the window. Even Newton's laws of mechanics had to be modified. But *not* Maxwell's! His equations, just the way he published them in 1873, are the same ones studied today[4]—they need *no* relativistic corrections.

How can that be, you wonder, as they predate Einstein's by 32 years? The mystery of this has bedeviled the experts down through the years. Oh, they have an explanation, alright. They say that all of electromagnetics is actually relativistic phenomena to begin with, and the laboratory work of Faraday, Ampere, Henry, and the other great experimentalists were studies of relativistic electron interactions in matter (although they, of course, didn't know that). Thus, it is only 'natural' that Maxwell's equations need no correction. So goes the 'expert' explanation, but it isn't right![5] I know Maxwell knew about relativity, and understood it perfectly. He knew all about time paradoxes and the equivalence of mass and energy.

Because how else can you explain the equation visible in my enlargement: $E = mc^2$!

Why, you must wonder (just as I did), didn't Maxwell publish this remarkable result? At first, I believed it was because of a lack of faith in his results. Who would have believed any of it in those Victorian times, so sure of its absolute view of nature? I thought of how Newton, 200 years before Maxwell, had suffered from a similar hesitancy when he wrote the *Principia*. There, when explaining his theory of gravitation, Newton did *not* employ his new invention of the calculus (which he *had used* to make his discoveries), but instead fell back on laborious arguments based on the accepted mathematics of algebra and geometry. Who would have believed him, otherwise?

But then I realized that couldn't be right. Maxwell was a strong man intellectually and he wouldn't have held back for fear of disbelief. No, it had to be that he discovered relativity and the mass-energy law just before his death, with no time to make his work known. I was still wrong.

---

[4] Don't forget note 3.

[5] Alas, I think it *is* right. Don't forget, this is science *fiction*. When there is a conflict between the needs of a story, and a rigid adherence to physics, the 'needs' wins!

It was later, when I returned to the enlargement of Maxwell's lost diary and read those painfully cramped notes, that I learned the truth. What I saw there showed me Maxwell had thought long and hard about his final discoveries and had purposely withheld them. For clearly visible, after I had slowly deciphered the writing, were the following words:

> *I have seen monstrous events. My blood has run cold at the sight of two great cities leveled to the ground, their inhabitants cruelly put to death instantly, or left to die slowly from a strange, lingering disease. Other trips, further on, have shown me the root of all these evils is the mass-energy equation, a result I at first believed to be my crowning glory. It will be my crown of thorns unless I ban it from my very being. Another will discover it for himself, but* my *soul shall be free! I have dismantled my machine, and shall never look upon or think of those horrible scenes again.*

This passage was dated just 1 month before Maxwell died a savage death from cancer. The reference to 'two cities' can only be that of Hiroshima and Nagasaki. His own death was surely caused by lingering too long among their atomic ruins.

Think of what this *means*. Quite simply, Maxwell knew the secret of time travel! But even more incredible is that it must be *easy*, if one only knows how, to build a time machine! Think about it—Maxwell had no gigawatt power stations at his disposal, no high technology machine shops, or nanosecond computers. He was not a gifted experimentalist, and once he had predicted radio waves, for example, it took others 20 years to finally generate them. And yet, *he* built a time machine. Somehow, with just the puny power sources available to him, and a limited mechanical capability, he wrested free the *simple* implementation of a time machine from his dynamical field equations.

Yes, yes, I know what you must be thinking. How can I really conclude such an incredible thing from a single equation on a blackboard, and a few words written by a man dying a painful death? A man, clearly suffering dearly, and possibly not in complete possession of his once marvelous mind.

This very evening the last bolt of evidence slid into place. Attempting to escape from the emotional maelstrom into which I had fallen, I turned to my old love of picture gazing. I took down from my library shelf a tattered yet cherished volume of the Meserve Collection of Lincoln pictures. My slow paging through the images stopped when I came to the famous photograph by Alexander Gardner of Lincoln's second inauguration. This incredible picture shows John Wilkes Booth looking down on Lincoln from behind a buttress high on the steps of the Capitol, while below in the crowd are the five men who, 41 days later, conspired with him in the assassination.

The following page demonstrated the extraordinary quality of Gardner's work, as it showed an enlargement of Booth's face in which the circular line between the pupil and the white of each eye is sharp and crisp! This impressive picture fascinated me, and I wondered if I could create a similar enlargement. It was then I remembered the old picture of Maxwell's grave, sent to me from Scotland, and the three distant figures in the background. They would present my photo-lab skills with a challenge, and the effort would distract my mind.

I finished the enlargement just 20 min ago. Those faces! Two of them I can now finally accept as being there—it must have been a pilgrimage for one, and for the other, it couldn't have been anything but a mocking, ironic gesture. But I wonder if the youngest one really knew who his two companions were? I don't know the answer to that—yet. But there they are, two men with faces my years of study have made as familiar to me as my own. One is a youthful Albert Einstein. The other, with the signs of death clearly written across his features, is James Clerk Maxwell. The face of the third man is familiar, too, for the third man is me!

Oh, I'm a bit older in the photo than I am now. But it's me, alright. A distinctive, jagged scar across the left cheek, a mark from a childhood accident, is sharply visible, and I can run a finger over my face and match it perfectly with the image in the enlargement. I'd say I'm about 45 or so in the image, no more than 10 years older than I am now. That doesn't leave me much time to keep my appointment, does it? I don't know, right now, how I'm going to do it, but I've got to rediscover Maxwell's secret of time travel. I'm sure I'll succeed—after all, there I am in the picture. Somehow, I'll be going back to pick James and Albert up so we can have our picture taken. Ten years—not much time.

I'm really looking forward to meeting my two new friends from across time.

For Further Discussion

When "Old Friends Across Time" originally appeared in *Analog*, it opened with a quotation from Richard Feynman's famous 1961–1963 Caltech undergraduate course (published in 1964 as *The Feynman Lectures on Physics*): "Ten thousand years from now, there can be little doubt that the most significant event of the nineteenth century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into provincial insignificance in comparison." This is almost certainly true, but could Maxwell *really* have built a time machine from just a knowledge of electromagnetic theory and special relativity (which is all that is needed to derive $\mathrm{E}=mc^2$), if he didn't also have a deep understanding of general relativity (and probably of quantum mechanics, too)? How likely do you think *that* is?

The narrator in "Old Friends Across Time" knows he is going to live long enough to eventually build a time machine; discuss the implications of this knowledge. For example, is he at least *temporarily* invulnerable to committing suicide (or, for that matter, to any other variation of dying?) That is, do we have a 'future' version of the grandfather paradox? This issue has never (to my knowledge) been considered by physicists, and not by philosophers either until recently. See, for example, S. Keller and M. Nelson, "Presentists Should Believe in Time-Travel," *Australasian Journal of Philosophy*, September 2001, pp. 333–345, and M. H. Slater, "The Necessity of Time Travel (On Pain of Indeterminacy)," *The Monist*, July 2005, pp. 362–369. More generally, if we assume that the past is unchangeable then the scenario in "Old Friends Across Time" seems to force at least some level of inevitability on the future as well. Or does it? In the 2007 story by Ted Chiang, "The Merchant and the Alchemist's Gate," that you were asked to read in a *For Further Discussion* at the end of Chap. 4, there is the following exchange between the narrator and the inventor of "the Gate" (a wormhole): "So if you learn that you are dead 20 years from now, there is nothing you can do to avoid your death?" He nodded. This seemed to me very disheartening, but then I wondered if it could not also provide a guarantee. I said, "Suppose you learn that you are alive 20 years from now. Then nothing could kill you in the next 20 years. You could then fight in battles without a care, because your survival is assured." "That is possible," he said. "It is also possible that a man who would make use of such a guarantee would not find his older self alive when he first used the Gate." "Ah," I said, "Is it then the case that only the prudent meet their older selves?" Comment on this issue, with particular attention to free-will.

Appendix B
Newton's Gift (A Story)[6]

Wallace John Steinhope was a sensitive human being, a person deeply concerned about the welfare of his fellow creatures. Any act of injustice, however slight, made his breast pound with righteous indignation. He was a champion of fair play, and his motto in life was taken from the ancient English rule of law—'Let right be done!'

Even while still a lonely, reclusive child, Wallace's heart ached mightly when he read of the laborious, boring, mind-dulling calculations endured by the great mathematicians of old. Just knowing, *thinking*, of Gauss's marvelous mind wasting literally months of its precious existence grinding out tedious mathematics that even a present-day dullard could do in a minute, on a home computer, was sheer agony for Wallace. Contemplation of the God-like Newton suffering endless delays in his gravity research, all because of a simple miscalculation of the length of a degree of longitude, was almost unbearable.

Indeed, Newton played a special role in Wallace's life (and he in Newton's, as we shall soon see). While the other great mathematical physicists had merely been hindered in their work by the lack of modern computational aids, Newton had squandered so much valuable time in other, nonscientific pursuits! His quasireligious writings alone, over half a million words, exceeded his scientific writings. What a waste! Wallace wondered endlessly over the reason for this strange misdirection of talent and bored his friends to the edge of endurance with his constant brooding on the mystery. Still, they all liked and admired Wallace enormously and so put up with it. But more than one of them had sworn to throw up the next time Wallace mentioned Newton during a wedding (but that's another story).

So deep was Wallace's anguish for his predecessors that even as he grew older and his own tremendous talents as a mathematical physicist (the result of a lucky

---

[6] P. J. Nahin, "Newton's Gift," *Omni*, January 1979. This tale was written with the specific goal of illustrating casual loop time paradoxes. The story reproduced here is, with only a few very minor alterations, as it originally appeared in *Omni*.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

genetic mutation induced in a male ancestor some centuries earlier) gained him an international reputation, thoughts of the unmeasurable misery of his scientific ancestors were never far from his mind. It was most appropriate, then, that his greatest discovery gave him an opportunity to *do* something! And Wallace John Steinhope vowed to *help*. He became convinced that it was his purpose on earth—he could not, he *would* not hesitate. As he strapped the knapsack-size time machine to his chest, his excitement was, therefore, easy to understand.

"It is done! And I am ready. I will travel back and bestow this gift of appreciation, this key to mental relief, on the great Newton himself!" Wallace cradled a small, yet powerful hand-calculator in his palm. It was a marvel of modern electronics. Incorporating large-scale integrated circuitry and a Z-8000 microprocessor solid-state chip, the calculator required only a small, self-contained nuclear battery for its power. It could add, subtract, multiply, divide, do square and cubic roots, trig and hyperbolic functions, take powers, find logarithms, all in mere microseconds. It was programmable, too, able to store up to 500 instructions in its micro-memory. The answers it displayed on its red, light-emitting diode readouts would liberate young Isaac from the chains of his impoverished heritage of mathematical calculation. No more Napier's bones for Newton!

But Wallace John Steinhope was no fool. He understood, indeed feared, time paradoxes. He knew Newton could be trusted with the secret, but it wouldn't do for the calculator to survive Newton's time. So Wallace had incorporated a small, self-destructing heat mechanism into it. After 5 years of use, it would automatically melt itself into an unrecognizable, charred slag mass. But that would be enough time for its task to be completed. The emancipation of Newton's mighty brain from tedium! Pleased enormously at the thought of the great good he was about to confer, Wallace set the time and space coordinates for merry old England, flipped the power switch on, and vanished.

Materializing in the Lincolnshire countryside in the spring of 1666, he began his rendezvous with destiny. It was the second and final year of the great bubonic plague, and Newton, seeking refuge from the agony and death plundering London and threatening his college of Trinity at Cambridge, had returned home to work in seclusion. The years of the Black Death were Newton's golden years, when the essentials of calculus would be worked out, when the colored spectrum of white light would be explained, and when the principle of the law of gravitation would be grasped. But how much easier it would be if Newton were released from the binding chains of dreary calculation. Wallace's gift would slip the lock on those chains! Accelerate genius!

It was early evening when, guided by a map of the area prepared by a friend who was both a cartographer and amateur historian, Wallace reached the quiet little town of Woolsthrope-by-Colsterworth. It was here, in a small farmhouse, that Wallace would meet his hero of the ages. A cold, gentle rain was falling as he approached the door. The soft, hazy light of an oil lamp glowed inside, revealing through the translucent glass the form of a man bent over a table. The fragrant smoke of well-dried wood curled from the chimney, announcing a warm fire within.

With his heart about to burst from excitement, Wallace rapped upon the door. After a pause, the shadow rose and moved away from the window. The door opened, and there stood Isaac Newton, a young man of 23 with an intellect that Hume and Voltaire considered "the greatest and rarest genius that ever rose for adornment and instruction of the species." But for the importance of his self-appointed mission, Wallace would have fainted dead away from the thrill of it all. "Is this the home of Isaac Newton?" he asked in a voice quavering with the trembling tones normally used by lovers about to reveal their deepest feelings.

The young man, of medium height and thick hair already showing signs of gray, swung open the door and replied, "My home it is, indeed, stranger. Come into the parlor, please, before the wetness takes you ill."

Isaac followed Wallace into the room and stood quietly watching as his visitor removed his soaked coat and hat. The portable time machine was gently placed on the floor next to a wall. The calculator was snug and safe in its plastic case in Wallace's shirt pocket. "Thank you, Master Newton. May we sit while we talk? I am afraid you may wish to take some time to consider my words." Motioning to a chair near the table, Isaac pulled a second chair from a darkened corner and joined Wallace. "You have a strange sound to your speech, stranger. Are you from hereabouts, or have you traveled far? Please commence slowly your tale."

Wallace laughed aloud at this question, a response prompted by his nervous excitement, and it quite surprised him. "Please forgive me. It is just that I *have* traveled so very, *very* far to see you. You see, I am from the future." Wallace was not one to play his cards close to his chest. Now it was Isaac's turn to laugh. "Oh, this is most ridiculous. Are you a friend of Barrow's at Trinity? It would be so like him to play such a trick.[7] From the future, indeed!"

Wallace's eyes ached at the sight of the papers on the table where Isaac had been working. What wonders must be there about to be born! In any other situation, Wallace would have asked their contents, but the die had been cast. He had to convince Isaac of the truth of his tale. But he had to walk a tight line, too. It just wouldn't do to misdirect Isaac's interest away from the calculator and toward the time machine itself! He must do something dramatic, something that would rivet his idol's attention and hold it.

"Yes, yes, I understand your reluctance to believe me. But, look here. This will convince you of the honesty of my words." Wallace pulled the shiny black plastic-cased calculator from his shirt pocket and flipped the power switch on. The array of LEDs glowed bright in the gloomy room as they flashed on in a random, sparkling red burst. Isaac's eyes widened, and he pushed his chair back. Was he frightened?

"As the Lord is my Savior, is it a creation of Lucifer? The eyes of it shine with the color of his domain. Are you one of his earthly agents?"

---

[7] The reference is to Isaac Barrow (1630–1677), who was the first Lucasian Professor of Mathematics at Cambridge. Barrow resigned that position to allow it to pass to Newton. Centuries later, Hawking became the 17th Lucasian Professor.

"Oh my, no! Look here, Master Newton, let me show you that there is no black magic or chicanery involved. It is all perfectly understandable in terms of the laws of Nature. What I have here is an automatic calculator, a device to perform all of your laborious mathematical labors."

So saying, Wallace squeezed the sides of the calculator case together, releasing pressure snap-fittings, and flipped the case open on a hinge at the top. Revealed to Isaac were the innards of the electronic marvel—a tightly packed interior of printed circuit boards, a mass of integrated circuitry, the small LED display, and the sealed nuclear battery. Isaac stared intently at the sight, and Wallace could see the natural curiosity of Newton's great mind begin to drive away the initial apprehension.

"But where are the gears, levers, springs, and ratchets to carry out the calculations? All I see is a black box with lights that glow red—how is *that* done; where is the lamp or candle to provide the light!—and many little isolated fragments of strange shapes. There is clearly nothing in your box that moves!"

"Oh, it is all done with electronics, Master Newton! The central processing unit has access to a solid-state memory that contains the decoding logic necessary to implement the appropriate algorithmic processes to provide the answers to the specific requests entered through these buttons. The actual performance of the box is achieved by the controlled motion of electrons and holes in suitably doped semiconductor material under the influence of electric fields induced—" Wallace, still overcome by his excitement, had rambled on wildly without thought of the essentially infinite technological gap that separated himself from Newton.

"Stop, stop," cried Isaac. "I understand only a few of the words you use and nothing at all of their meaning! But it is obvious that for calculations to be performed, mechanical work must be done, and that implies motion. Pascal's adding machine has shown the veracity of that. I say again, nothing moves in the box. How *can* it work?"

Wallace was embarrassed. The mistake of overlooking the hundreds of years of progress after Newton's time was one a child might make. "I am sorry, Master Newton. I'm going too fast for you." Isaac looked at Wallace with a frown, but Wallace failed to see the pricked vanity of the proud Newton. Going too fast, indeed!

Wallace prepared to lay a firmer technological foundation for Newton, but then he froze. It couldn't be done! Newton was a genius, certainly, but the task was still impossible. Wallace would have to tell him all about Maxwell's equations, Boolean algebra and computer structure, electronics, and solid-state device fabrication technology. It was just too much, and besides, there was the danger! The potential time paradoxes of all that knowledge out of its proper time sequence! What if Newton, in innocence, revealed some critical bit of knowledge out of its natural place in history? So, Wallace hesitated, but seeing the suspicion grow again in Isaac's eyes, he realized he had to do something, *anything*, immediately.

"You cannot deny your own eyes," answered Wallace. "Let me *show* you how it works. I'll divide two numbers for you with just the punch of a few buttons. Watch this." And, at random, he entered 81,918 divided by 123. Poor Wallace, of all the numbers to use, they were the worst. Within milliseconds the answer glowed brightly in fiery red characters. Wallace looked with pride at the result and then, already enjoying in his mind what he knew would be Isaac's amazement, he turned his eyes to the great man. What he saw made his spine tingle, and the gooseflesh stand high on his neck. Newton had fallen to his knees, with eyes bulging and hands raised as if in prayer.

"The mark of the Beast, it is the mark of the Beast! It is so written in the Book of Revelations—Here is wisdom. Let him that hath understanding count the number of the beast; for it is the number of man; and his number is six hundred three-score and six!" Rising to his feet, Newton fell back into his chair. "Your cursed box bears the brand of its master. There can be no doubt now, it *is* the creation of the fallen archangel!" Wallace was aghast at Isaac's violent reaction. The seventeenth century genius had now stumbled backward from his chair and had grasped a poker from the hot coals of the fireplace.

"Wait, please wait! Watch this; I'll multiply two other numbers together for you, watch!" Wallace quickly punched in the data, and then the answer gleamed steadily in burning red characters on the LEDs. Isaac's eyes first went wide with fear as he again saw the wizard electronics do their marvelous assignment, and then he shut them tight. Wallace was becoming desperate—this wasn't the way it was supposed to be! "Don't you see—imagine the tedious work, the mind-deadening labor this machine will save you from. And it is yours."

"Yes? But only for the exchange of my soul! That is always the Devil's price for his seductive gifts from Hell!" As Isaac shrieked these last words at Wallace, he raised the poker over his head. "Begone, you emissary of the Dark World! I know now you must be in the employ of the Father of the Antichrist, but the Lord God Almighty will protect me if I do not waver in my resolve. Begone, or I'll strike your brains out on the floor where you stand!"

Isaac's eyes were wide with fear, nearly rolling back to show all white spittle sprayed from his mouth as he yelled at Wallace, who stared in shock at the wild man who threatened him with death. "Please, please, *listen* to me, please! I beg you to understand—I am a scientist, just like you. The concept of the devil, and all it stands for, is contrary to everything I believe. How *could* I be in the devil's employ, when I don't even accept his existence? You *must* believe me!"

"Blasphemy!" screamed Isaac. "Your own words condemn you. To deny the reality of Satan in a sinful world is to deny that of God, too. Now leave my home, you dark beast from hell, or by the heavens above, *I shall destroy you*!" As he shrilled these words, Isaac brought the poker down in a wild swing that barely missed Wallace's head.

Struck dumb with confusion at the uncontrolled outburst, Wallace stuffed the calculator into his shirt, grabbed his hat, coat, and time machine and rushed from the house. As he hurried into the cold, wet night, he turned back, just once, to see Isaac Newton framed in the light of the open door. "Go, go, you foul messenger from the Lord of Evil! Back to your stinking pit of burning hell-fire! This is a house that honors the Divine Trinity and is no haven for the likes of you!"

Wallace rushed away into the blackness, the time machine bouncing unheeded upon his chest.

He ran, for how long he couldn't recall, until he fell exhausted next to a stream running heavy with the rain. Tears of rage, frustration, and shock streamed from his eyes. Rejected by the great Newton! Well, damn him! Wallace flung the calculator into the stream in his terrible anger and activated the return coordinates. He faded from Newton's world as quickly and quietly as he had come.

As for Isaac Newton, after having chased the Devil's messenger from his house, he returned on shaking legs to his desk. Pushing aside his rough calculations on the orbit of the moon around the earth, he swore to redeem himself in the eyes of the Savior. Somehow, he had been found lacking and had been tested. And the test was surely not over! He began to reapply his marvelous mind to determine the origin of his failure before the Lord God Jehovah. Taking quill in hand, he wrote the first of the many hundreds of thousands of words that his numerous religious tracts would devour from his allotted time.

Five years later, long after Newton had returned to Cambridge, a group of picnicking children were frightened when a nearby stream suddenly erupted into a geyser of steam. Moments later, the bravest (or most foolhardy) of the boys—who, by an astonishing coincidence that befits any good time travel paradox, would be Wallace's great-grandfather nine times removed—cautiously examined the streambed. All he found were some twisted, hot pieces of what he thought was a hard, black rock, and he tossed them back. They were all that was left of the calculator's nuclear battery. He did receive a tiny radiation dose from them, which caused a recessive genetic mutation that centuries later would suddenly appear as the cause of Wallace's genius, but otherwise the lad was unaffected. The incident was soon forgotten.

Well over 300 years later, Wallace John Steinhope reappeared in his own time. He was essentially the same man as before he left—kind, generous, and sensitive, and ready to come to the aid of any man or beast that might need help. As far as his friends were concerned, in fact, he was even improved (naturally, they didn't know what had brought about the welcome change but, if they had, they would have applauded it).

Wallace John Steinhope, you see, never again had another kind word for Newton, or for that matter, any words for him at all.

For Further Discussion

In his book *Travels in Four Dimensions: the enigmas of space and time* (Oxford 2003), the philosopher Robin Le Poidevin writes (p. 176) "But, as everybody knows, when a time machine leaves for another time it *disappears*." This is, indeed, how the time machine in "Newton's Gift" works; however, after reading *Time Machine Tales* do you think such behavior is in agreement or in conflict with general relativity? Defend your position.

"Newton's Gift" contains causal loops. Identify two of them, and discuss their role in the story (that is, are they central to the story or merely incidental?).

The idea of a time traveler visiting famous people in the past occurs fairly frequently in science fiction. In Ian Watson's "Ghost Lecturer" (*Isaac Asimov's Science Fiction Magazine*, March 1964), for example, the inventor of the "Roseberry Field" uses it to yank geniuses out of time to supposedly honor them, to let them know their lives had been worthwhile in the eyes of the future. But then he goes on to tell them—oh so kindly—where they had gone wrong or had fallen short of the mark, and of how much more we know nowadays. "You almost got it right, boy! You were on the right track, and no mistake. Bravo! *But* . . ." Watson makes the interesting observation that one can easily imagine playing this pathetic game of 'second-guessing' history with scientists, but what could even the most talented modern do to upstage a Mozart or a Shakespeare? Most similar to "Newton's Gift," however, are (for example) Gregory Benford's "In the Dark Backward" (*Science Fiction Age*, June 1994) where Shakespeare and Hemingway are visited, and Jack McDevitt's "The Fort Moxie Branch" (*Full Spectrum*, October 1988) where Hemingway and Thomas Wolfe appear. Read these stories, and then compare/contrast their descriptions of how story characters react to the appearances of time travelers, to Newton's behavior in "Newton's Gift."

Appendix C
Computer Simulation of the Entropic Gas Clock

```
%gasclock.m/created by PJNahin for TIME MACHINE TALES(6/27/2015)
%This MATLAB m-file simulates the diffusion of gas molecules in a
sealed
%container by using the Ehrenfest ball exchange rules. The
simulation
%starts with n molecules (i.e., balls) of one type (i.e., black)
on
%one side of the container, and n more molecules of another type
(i.e.,
%white balls) on the other side. The two urns play the roles of the
%two sides of the container. To simulate the ball (molecule)
%movements, the program selects two random numbers from 0 to
1, which
%are then compared to the current probabilities of selecting a
black
%ball from urn I and a white ball from urn II. If BOTH random
numbers
%are greater than these two probabilities then a white ball has
been
%selected from urn I and a black ball has been selected from urn
II,
%and so the number black balls in urn I is increased by one while
the
%number of white balls in urn II is increased by one. If BOTH
random
%numbers are less than or equal to these two probabilities then a
%black ball has been selected from urn I and a white ball has been
%selected from urn II and so the number of black balls in urn I is
```

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

```
%decreased by one while the number of white balls in urn II is
decreased
%by one. If one of the random numbers is greater than its
corresponding
%probability while the other random number is less than its
%corresponding probability, then no action is taken because then a
%white (black) ball moves from urn I to urn II at the same time a
white
%(black) ball moves in the opposite direction. That is, there is
no
%net change. Then, the ball selection probabilities are recalcu-
lated and
%another ball exchange is simulated.
rand(’state’,100*sum(clock))     %new seed for the random number
generator;
n=100;                           %number of balls in each urn;
nb1=n;                           %number of black balls INITIALLY
in urn I;
nw2=n;                           %number of white balls INITIALLY
in urn II;
pb1=nb1/n;                       %probability of selecting a black
ball from urn I;
pw2=nw2/n;                       %probability of selecting a white
ball from urn II;
for trials =1:1000;
   system(trials)=pb1;
   ball1=rand;
   ball2=rand;
   if(ball1>pb1&ball2>pw2)         %white ball selected from urn I
      nb1=nb1+1;                   %and black ball selected from
      nw2=nw2+1;                   %urn II;
   elseif(ball1<=pb1&ball2<=pw2)  %black ball selected from urn I
      nb1=nb1-1;                   %and white ball selected from
      nw2=nw2-1;                   %urn II;
   end
   pb1=nb1/n;
   pw2=nw2/n;
end
plot(system)
axis([1 trials 0 1])
grid
xlabel(’time, in arbitrary units’)
ylabel(’fraction of balls in urn I that are black’)
figure(1)
```

Epilogue

> [Science fiction] cannot be good without respect for good science . . . This does not include time machines, space warps and the fifth dimension; they will continue to exist in the hazy borderland between [science fiction] and fantasy.[8]

> In many science-fiction stories, the trip into the past is by way of some futuristic machine that can take you through time at will . . . That, however, is totally impossible on theoretical grounds. It can't and won't be done.[9]

The opening quotations, particularly the second one from Asimov who was one of the great modern writers of science fiction, is a gloomy one indeed for fans of time travel, but it is not difficult to find inconsistency in Asimov's own tales dealing with the concept. Asimov is famous, in particular, for his stories of robots, and the very last such tale that he wrote combines robotics with time travel, with a robot sent two centuries into the future.[10] At the start of the story, the narrator tells us that time travel to the past is impossible because the past is unchangeable and (of course) a time traveler would necessarily disturb history. (That is (of course) simply a failure to distinguish between the difference of *changing* the past and *affecting* the past, as well as a failure to see how the principle of self-consistency negates the issue of paradoxes.) Then, when the robot returns from the future (and so backward time travel is *not* impossible!), he reports that his arrival had been expected, that history had recorded that he would appear. At the end of the story we learn how the future knew this—it had read "Robot Visions"! So now Asimov *uses*

---

[8] Harry Harrison, in his essay "With a Piece of Twisted Wire . . .," *SF Horizons* (no. 2), 1965. Harrison (1925–2012) was a well-known (if little appreciated outside the SF community) writer, whose 1966 novel *Make Room! Make Room!* was the inspiration for the excellent (if somewhat depressing) 1973 film *Soylent Green* (a movie about future over-population of the Earth that will make you think twice about ever eating a cookie again).

[9] From an essay Isaac Asimov wrote on the time travel movie *Peggy Sue Got Married* for the *New York Times*, October 5, 1986.

[10] I. Asimov, "Robot Visions," *Isaac Asimov's Science Fiction Magazine*, April 1991.

© Springer International Publishing AG 2017

P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

the principle of self-consistency, with the narrator realizing that he *must* preserve his story so the future can read it.

Not a very consistent story! Asimov, was, of course, writing a story for entertainment's sake, so perhaps it's unreasonable to hold him *scientifically* accountable (although logic wouldn't seem to be too much to ask for).

In any case, was Asimov right? Lots of his fellow science fiction writers certainly thought so. One, for example, bluntly asserted that

> Time travel is inconceivable.[11]

Other critics agreed:

> In science fiction we find the lunatic fringe more often than not trying to perfect time-travel mechanisms.[12]

and

> Scientifically, time travel can't stand inspection.[13]

and

> Time travel is ... scientific nonsense.[14]

and

> It would be untrue ... to present the idea of a time machine as anything but what it is, an intriguing literary device, part of the bag of tricks of the science fiction writer ... There is no such thing as a 'science' of time travel.[15]

You'll notice that these pronouncements are from decades ago: Conklin (1904–1968), Gold (1914–1996), and Oliver all wrote just 3 years after Gödel, and so perhaps it was simply too soon for his work to be widely known outside of the physics community. But physicists have learned a lot since 1952! Have they learned enough to make Asimov and his fellow SF skeptics (if they were still alive) change their minds, or at least reconsider? I suspect not.

I say that because, even 25 years after Conklin, Gold, and Oliver wrote, while we do find an awareness of Gödel starting to appear in the science fiction world, a feeling of skepticism was still in the air. In a fascinating analysis[16] of the first half-century of the science fiction magazines, Paul Carter admitted that there *is* a rationality to time travel because of Gödel but, nonetheless, the conventional view remained that backward time travel is simply impossible. Then, citing the work of Tipler, Carter wrote "Only as recently as 1974 (see note 130 in Chap. 1), in the sober pages of the *Physical Review*, has a physicist been more bold . . . For 70 years in the meantime, however, without waiting for Professor Tipler to solve his equations . . . writers had happily helped themselves to Mr. Wells' invention and sent their characters through time in every direction, forward, backward, and sideways."[17]

---

[11] Kingsley Amis, *New Maps of Hell*, Harcourt 1960.

[12] Groff Conklin, *Science Fiction Adventures in Dimension*, Vanguard 1953.

[13] H. L. Gold, *The Galaxy Reader of Science Fiction*, Crown 1952.

[14] Alexei Panshin, *The Mirror of Infinity*, Canfield 1970.

[15] Chad Oliver, "The Science of Man," a non-fiction essay included in Oliver's 1952 time machine novel *Mists of Time*. Chad Oliver (1928–1993) was a scientist by profession (anthropology), and his opinion carried weight among SF writers *and* (non-physicist) scientists.

[16] P. A. Carter, *The Creation of Tomorrow*, Columbia University Press 1977.

In the 1980s writers were apparently just as unaware of Gödel's time travel analyses (and of the much later ones of Tipler) as had been the 1950s commentators. In his marvelous 1985 book *The Past Is a Foreign Country*, for example, David Lowenthal repeatedly refers to time travel as "fantasy" and to science fiction stories about time travel as "unbridled by common sense." And for another example from the start of the 1980s, consider the case of James Gunn (born 1923), professor of English at the University of Kansas, past president of both the Science Fiction Writers of America and the Science Fiction Research Association, author of *The Immortals* (inspiration for the 1970–1971 TV series of the same name), and eminent scholar (see his 1975 book *Alternate Worlds*). His literary credentials are impeccable and his critical influence profound. And yet, 30 years after Gödel and 5 years after Tipler, Professor Gunn wrote in *The Road to Science Fiction*, "Time travel has been an anomaly in science fiction. Clearly fantastic—there is no evidence that anyone has ever traveled in time and *no theoretical basis for believing that anyone ever will* [my emphasis]." If you've read this book carefully, however, of the analyses by Gott, Krasnikov, Thorne, Alcubierre, Novikov, Natário, and others, you know that what Gunn claims in those last words is actually not necessarily so.

The British-born American theoretical physicist Freeman Dyson of the Institute for Advanced Study has commented[18] on that sort of narrow mindset, with words quoted from the 1979 physics Nobel prize winner Steven Weinberg, words reminding us that rigidity concerning time travel is not limited to science fiction writers: "This is often the way it is in physics—our mistake is not that we take our theories too seriously, but that we do not take them seriously enough. It is always hard to realize that these numbers and equations we play with at our desks have something to do with the real world. Even worse, there often seems to be a general agreement that certain phenomena are just not fit subjects for respectable theoretical and experimental effort." The words *time travel* and *time machine* are never mentioned, but could they have been far from either Weinberg's or Dyson's thoughts?

All through this book we have seen how people have argued against time travel to the past (Tipler's cylinder is unphysically long, Gödel's universe requires an unphysical rotation, wormholes and warps require unphysical energy conditions, what about all those paradoxes . . . and on and on). These arguments remind me of the debate in the 1930s between the illustrious British astrophysicist Sir Arthur Eddington and the young Indian astrophysicist Subrahmanyar Chandrasekhar (1910–1995), winner of the 1983 Nobel prize in physics. In his analyses of the life history of stars, Chandrasekhar had arrived at an astonishing conclusion, one that Eddington simply could not accept. As Eddington sarcastically explained in an address at Harvard University in the summer of 1936, “Above a certain critical mass (two or three times that of the sun), the star could never cool down, but must go on radiating and contracting until heaven knows what becomes of it. That did not worry Chandrasekhar, he seemed to like the stars to behave that way, and believes that is what really happens.”[19] Eddington then went on to declare such ‘unbelievable’ behavior to be nothing less than “stellar buffoonery.”

---

[17] Given that *The Time Machine* was published in 1895, it is not clear how Carter arrived at the value of 70 until Tipler's work (he should have written 79), and of course it was only 54 years between Wells' time travel fiction and Gödel's time travel mathematical physics.

[18] F. J. Dyson, "Time Without End: Physics and Biology in an Open Universe," *Reviews of Modern Physics*, July 1979, pp. 447–460.

As far as Eddington was concerned, Chandrasekhar had simply made an error in combining relativity theory with non-relativistic quantum theory. Indeed, so appalled was Eddington at the thought of a star contracting “until heaven knows what becomes of it” (that is, until it gravitationally collapses into a black hole) that he had earlier, in 1935, stated “There should be a law of nature to prevent a star from behaving in this absurd way!” Today, of course, no astrophysicist feels the need for a ‘star protection conjecture’—which perhaps reminds you of another, more recent ‘protection conjecture.’

What can one conclude from all the similar controversy concerning time travel, time machines, and spacetime warps? Not much, I think, except that these are open issues and will remain the subjects of on-going study for a long time yet to come. The one thing I am fairly certain of is that if time travel is ever achieved, it will be by means that we cannot today even begin to guess. It will almost certainly require *at least* a mutant child genius with an IQ of 270 to fix the slightly broken time machine found abandoned in a cellar![20] But that view isn’t uniformly shared across all of science fiction. I very much doubt, for example, that things will be quite so elementary as depicted in the story[21] where the time machine was so simple that “If it were taken apart or put together before you, your wife, or the man across the street, you would wonder why you didn’t think of it yourselves.” Not only that, but its power source was just two dry cell batteries!

The time machine in an earlier story is almost as simple, requiring (besides a piece of strange crystal) only a “little stack of dry cells, a Ford [automotive ignition] coil, a small brass switch, a radio ‘B’ battery, an electron tube, and a rheostat.”[22] Even Wells’ *Time Machine* couldn’t resist making it all look easy: as one critic put it, “The time machine, like all products of supreme inventive genius, was a remarkably simple affair. A few rods, wires, some odd glass knobs—nothing more!"[23] That sort of simplistic fictional description of a time machine reminds me of the reaction of the great Polish science fiction writer Stanislaw Lem to the general treatment of time travel in the genre: "There have been mountains of nonsense written about traveling in time, just as previously there were about astronautics—you know, how some scientist, with the backing of a wealthy businessman, goes off in a corner and slaps together a rocket, which the two of them—and in the company of their lady friends, yet—then take to the far end of the Galaxy. Chronomotion, no less than Astronautics, is a colossal enterprise, requiring tremendous investments, expenditures, planning . . ."[24]

---

[19] See S. Chandrasekhar, *Eddington: The Most Distinguished Astrophysicist of His Time*, Cambridge University Press 1983, p. 48.

[20] F. B. Long, “A Guest in the House,” *Astounding Science Fiction*, March 1946.

[21] R. Abernathy, “Heritage,” *Astounding Science Fiction*, June 1947.

[22] J. Williamson, “In the Scarlet Star,” *Amazing Stories*, March 1933.

An example of what Lem was talking about is the 1956 novella *Arcturus Landing* by Gordon R. Dickson (1923–2001). There we read of aliens who have confined humans to the solar system—until (if) Earth scientists discover the secret of FTL travel. So, a genius physicist does just that (with no mention of spacetime engineering, but rather we encounter a lot of mumbo-jumbo gibberish as the 'explanation'), and uses it to instantly transport himself and some friends to a planet orbiting Arcturus.[25] And when they get there the friendly aliens speak perfect English.

Lem would have snorted in derision, too, at this statement made to a prospective graduate student by the head of a college physics department, that the college "has been awarded a million dollars to build [a time machine]. It means . . . a raise for me and maybe a doctorate for you, so we'll build one and have some fun doing it."[26] Is it any wonder that Lem so readily dismissed stories that reduce space (and time) travel to weekend adventures in a home laboratory? As Lem wrote in another essay, time travel and its close relation, FTL space travel, have reduced much of science fiction to "a bastard of myths gone to the dogs."[27] Because of precisely that, Harry Harrison wrote (note 1) of the early science fiction magazines that published so much nonsense, "I used to moan over the fact that pulp magazines were printed on pulp paper and steadily decompose back towards the primordial from which they sprang. I am beginning to feel that this is a bit of a good thing."

I don't know whether time travel to the past can actually be accomplished, but I do know that speculations once thought to be as outlandish as finding the Philosopher's Stone for turning base elements into gold, *have* eventually been realized (and, come to think of it, with modern nuclear physics we *have* learned how to turn lead into gold, if only a few atoms at a time). Television, nuclear power, home computers that run at multi-gigahertz clock rates in the bedrooms of high school students, even faster computers that animate our movies and simulate the formation of black holes and galaxies, voyages to the Moon and back—all these amazing developments would be pure magic to nineteenth century science. The ghosts of not just a few Victorian scientists who had poo-pooed the possibility of such things, have watched their reputations eat a lot of posthumous crow during the last 150 years.

---

[23] W. B. Pitkin, "Time and Pure Activity," *Journal of Philosophy, Psychology and Scientific Methods*, August 27, 1914, pp. 521–526.

[24] S. Lem, "The Twentieth Voyage of Ijon Tichy," in *The Star Diaries*, Seabury Press 1976.

[25] A journey incorrectly given in the story as 120 light years, when in fact it is less than 40 light years.

[26] W. West, *River of Time*, Avalon Books 1963.

[27] S. Lem, "Cosmology and Science Fiction," *Science-Fiction Studies*, July 1977, pp. 107–110.

My personal position on the question of time travel leans towards the rejoinder made to the skeptic in one science fiction story who, even after having done some time traveling, *still* argues against it by invoking paradoxes. He is sharply rebuked with "Oh, for heaven's sake, shut up, will you? You remind me of the mathematician who proved that airplanes couldn't fly."[28] I subscribe to the optimistic philosophy of the British writer Eden Phillpotts (1862–1960), who wrote in his 1934 novel *A Shadow Passes* "The Universe is full of magical things, patiently waiting for our wits to grow sharper." Perhaps he had a famous saying by the British-born Indian scientist J. B. S. Haldane (1892–1964) in mind, words from his 1928 *Possible Worlds*: "Now my suspicion is that the universe is not only queerer than we suppose, but queerer than we can suppose."

Still, even if time travel is possible, the engineering phase will surely be tough going. I am certain that before we see a working time machine, there will be many, *many* episodes like the one described in a very funny, novel-length spoof of academic research.[29] All physicists and engineers who have tried to get some stubborn piece of apparatus to work, apparatus that *should* work and simply won't, will appreciate Professor Demetrious Demopoulos' frustration and will, I am sure, forgive him his intemperate language:

> . . . the distinguished physicist took a step back and, arms akimbo, surveyed the complex and sophisticated machine that was the culmination of years of dedicated scientific research and pains-taking technological development.
>
> "What a pile of ****," he said.
> "Oh, no, Dr. Demopoulos, don't say that!"
> "Well, it is." A sneer formed on the professor's thin lips. "Time machine, my ****. This thing couldn't give you the time much less travel in it."
> "But we haven't incorporated all our latest test data yet," the pretty research assistant reminded him. "These last few adjustments might do it, Professor."
> "Hell, we've been tinkering with it for 2 years," Demopoulos complained.
>
> "We've tried everything and it's all come to dog ****."

---

[28] R. Heinlein, "By His Bootstraps," *Astounding Science Fiction*, October 1941. As discussed at the end of Chap. 1 (in "For Further Discussion") the mathematician was the American astronomer Simon Newcomb.

[29] J. DeChancie and D. Bischoff, *Dr. Dimension*, ROC 1993.

That scene probably won't actually happen for a long time to come but, even before the practical nuts-and-bolts bugs in the Professor's time machine are worked out, I think some adjustments are called for in our thinking about time travel. I believe that present-day philosophers and science fiction writers are going to have to become knowledgeable about the work by physicists on time travel. It simply won't do any longer for Philosophy Professor X to invoke the grandfather paradox during a discussion of causality and free will and then airily declare them to be 'obviously' incompatible with time travel to the past. And it simply won't do any longer for Famous SF Writer Y to send his hero into the past to kill Hitler as a baby and thereby change recorded history. One might as well keep watching a video recording of the 9/11 destruction of the World Trade Center, in the vain hope that maybe, on the next viewing, the planes will miss.

The principle of self-consistency around closed timelike curves is going to have to become as much a part of the science fiction writer's craft (or else she will be a writer of fantasy) as it will have to become part of the fundamental philosophical axioms.[30] The 'time police,' like the "operatives of the Bureau of Time Exploration and Manipulation" that appeared in the science fiction of Andre Norton (1912–2005), will have to be put out to pasture with the unicorns and telepathic dragons of fantasy fiction. Just as the recent physics literature on time machines has displayed a growing awareness of what science fiction writers and philosophers have had to say on the subject of time travel, so too are writers and philosophers going to have to learn some more physics. Most people can enjoy a good fantasy tale now and then, but the use of 'magic mirrors' to see through time is *not* physics. Such devices were popular and acceptable in medieval times—see "The Squire's Tale" in Chaucer's *The Canterbury Tales*, and later (see Act IV of *Macbeth*)—but good science fiction needs much more than that today.

Time travel to the past is a beautiful, romantic idea, and some words written by two physicists in a technical paper—words embedded in the midst of swirls of tensor equations—show that even hard-nosed physicists can share this dream: "In truth, it is difficult to resist the appealing idea of traveling into one's own past . . ."[31] The appeal of that dream is explained in Ray Bradbury's Foreword to a beautiful little 1989 book by Charles Champlin (*Back There Where the Past Was*). In it Bradbury clearly illuminated *why* we want to go back into the past. It is for the same reason that we go, time and again, to see *Hamlet*, *Othello*, and *Richard III*: "We don't give a hoot in hell who poisoned the King of Denmark's semicircular canal.

---

[30] Bud Foote (1930–2005), late professor of English at Georgia Tech, wrote (in his book *The Connecticut Yankee in the Twentieth Century: Travel to the Past in Science Fiction*, Greenwood Press 1991) that consistency is simply a well-used plot device: "The attempt of the time traveler to prevent something or take advantage of it [and so causing] the event in question, is so popular and so ubiquitous that it seems to be about worn out." Worn out or not, I believe that plot device to be correct science.

[31] A. J. Accioly and G. E. A. Matsas, "Are There Causal Vacuum Solutions with the Symmetries of the Gödel Universe in Higher-Derivative Gravity?" *Physical Review D*, August 15, 1988, pp. 1083–1086.

We already know where Désdemona lies smothered in bedclothes and that Richard goes headless at his finale. We attend them to toss pebbles in ponds, not to see the stones strike, but the ripples spread."

That's why a visit to the past is so mysteriously and marvelously fascinating. It would let us watch ripples spread through time. Our own visit to the past, in fact, might even be the pebble in the pond of history that starts an interesting ripple or two that will one day sweep over—us! (Take a look at Appendices A and B) Who would want to miss that? Indeed, if modern philosophers are right, if the analyses discussed earlier in this book are correct, you *can't* (didn't/won't) miss it.

I think time travel appeals, irresistibly, to the romantic in the soul of anyone who is human.[32] A time traveler does not exist either *here* or *then*, but rather *everywhen*. For a time traveler passing back and forth through the ages, history would be the ultimate puzzle, a chronicle described in one novel as beginning "not in one place, but everywhere at once ... It might be begun at any point along the infinite, infinitely broken coastline of time."[33] Romanticism doesn't preclude there also being a dark side to visiting the past, of course, as one time traveler from 1989 learns when he takes up residence in 1962. Falling asleep on a hot summer night in that long-ago year, he thinks "JFK slept. Oswald slept. Martin Luther King slept. [I sleep and dream] of Chernobyl ... *I am a cold wind from the land of your children*."[34]

But, I must admit, I personally am more attracted by happier descriptions of time travel. In his marvelous 1996 book *1939: The Lost World of the Fair*—which is proof that there are not enough Pulitzers to go to all the books that deserve one—Yale professor David Gelernter caught just the right spirit in his Prologue: "The best of all reasons to return to the fair is that travel is broadening, and time travel most of all ... The 1939 New York World's Fair is one amazing show. It still stands, undisturbed on Flushing Meadow, just over the edge of time; it would be an unforgivable shame to miss it." Trust me—if you read Gelernter's book, you'll come as close as you can in today's world to taking a ride in a 'time machine'!

The eminent philosopher Sir Karl Popper opens his biography with a wonderful story about his apprenticeship as a young man in 1920s Vienna to a master cabinetmaker.[35] After winning the old man's confidence, the student learned his mentor's great, secret desire: For years the master had been looking for the solution to perpetual motion. He knew what physicists thought of such machines, but nonetheless he had never given up his dream: "They say you can't make it; but once it's been made they'll talk differently." Popper's master sounds just a bit like the American writer Gertrude Stein (1874–1946) in her 1938 essay "Picasso," where she writes "It is strange about everything, it is strange about pictures, a picture may seem extraordinarily strange to you and after some time not only it does not seem strange but it is impossible to find what there was in it that was strange." Might we one day say the same thing about time travel?

---

[32] How else to explain the pleasure, for modern children *and* adults too, in watching rebroadcasts of the 1960s animated TV cartoon program 'starring' Mr. Peabody, a nice but slightly stuffy, professorial white beagle. (Don't all dogs wear glasses and a bow tie?) Mr. Peabody, with his brainy adopted son Sherman, routinely travels into the past in the "Way-Bac" machine to see what *really* happened in history.

[33] John Crowley, *Great Work of Time*, Bantam 1991.

[34] R. C. Wilson, *A Bridge of Years*, Doubleday 1991.

[35] See volume 1 of *The Philosophy of Karl Popper* (P. A. Schilpp, editor), The Library of Living Philosophers, Open Court 1974, p. 3.

An alternative point of view can be found in a discussion of time travel via cosmic strings that makes this assessment: "While there is still hope that one day a sufficiently clever design may make building a time machine possible, it is beginning to seem more and more improbable. Like the perpetual motion machines of the nineteenth century, the designs have an elegant simplicity (as well as enormous commercial potential), but it seems that Nature also may abhor them just as much."[36] Of course, at one time it was thought that Nature abhorred a vacuum, but then we learned that she must actually love a vacuum because else why did she make so much of it?!

The theoretical basis for time travel is very different from that of perpetual motion (there is *more* reason to accept time travel as a plausible possibility). And so maybe one day, *just maybe*, the first time traveler will receive a toast such as the one in a story telling us about the arrival of the inventor of the first time machine and his no longer skeptical friend in the Civil War past:

"To you, Mac," I said.

McHugh loosened his tie. "To the Creator," he said, "who has given us a Universe with such marvelous possibilities."[37]

---

[36] B. Allen and J. Simon, "Time Travel on a String," *Nature*, May 7, 1992, pp. 19–21.

[37] J. McDevitt, "Time's Arrow," *Isaac Asimov's Science Fiction Magazine*, November 1991.

Glossary[38]

Action the integral over a **world line** of a quantity called the *Lagrangian*. When a massive particle is moving at non-relativistic speed through a gravitational field, for example, the instantaneous value of the Lagrangian is the difference between the kinetic and potential energies of the particle. For other types of fields (such as the electromagnetic) and/or relativistic motion in any type of field, the Lagrangian is different. In any case, however, the actual world line of the particle is the one for which the integrated Lagrangian, that is, the action, is minimized. See **least action**.

Action at a distance the direct interaction of two separated objects, without concern for the details of what (if anything) occurs in the region between the objects (see also **field**). Newton's theory of gravity is action at a distance, whereas Einstein's theory of gravity is a *field* theory.

Advanced solution the prediction, by Maxwell's electromagnetic field equations, of radio waves that travel into the past (see also **Dirac radio**).

Anti-matter quantum mechanical prediction (experimentally verified) that all fundamental particles of matter come in two forms (the 'normal' version and the 'anti-matter' version). The positron, for example, is the anti-matter version of the electron, differing only in the sign of its electric charge. The photon, on the other hand, is its own anti-particle. A **subluminal** anti-particle traveling forward in time can be thought of as its 'normal' version traveling backward in time.

Arbitrarily advanced civilization for time travel discussions, a civilization with a technology sophisticated enough to construct a traversable **wormhole** in spacetime. More generally, Types I, II, and III of such civilizations are, respectively, those that can control 10^{13} W, 10^{27} W (the total power output of their home star), and 10^{38} W (the total power output of their home galaxy).

---

[38] "I hate definitions." (Usually attributed to writer and British Prime Minister Benjamin Disraeli (1804–1881) but, more precisely, they are the words of one of the characters in his 1826 novel *Vivian Grey*.)

—but they *can* be useful

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

Arrow of time the statement the time appears to have a direction, that there is a difference between the past and the future. There are several different arrows: the psychological (we remember the past, we anticipate the future), the thermodynamic (organized systems evolve toward disorganization, that is, **entropy** increases as time increases), the electromagnetic (radio waves propagate *away* from their generators), and the cosmological (the expansion of the universe is directed toward the future).

Asymptotically flat if the geometry of a curved spacetime is such that, as one moves ever further away from all matter and energy, the spacetime **metric** becomes that of flat **Minkowski spacetime**, then the curved spacetime is said to be *asymptotically flat*. As a counter-example, the spacetime of a **Tipler cylinder** time machine is *not asymptotically flat*.

Autoinfanticide paradox see **grandfather paradox**.

Averaged null energy condition the claim that the *averaged* value of the observed mass-energy density along the entirety of any **null geodesic** is non-negative.

Averaged weak energy condition the claim that the *averaged* value of the observed mass-energy density along the entirety of any **timelike world line** is non-negative.

Back reaction the tendency of spacetime to resist the formation of **closed timelike lines** (see also **stress-energy divergence**).

Bell's theorem an inequality that either holds or does not hold, depending on whether quantum mechanics is non-local or local, respectively.

Big Bang the singular beginning of spacetime.

Big Crunch the singular end of spacetime.

Bilking paradox what would happen if a **causal loop** were disrupted. For example, suppose a time traveler builds a time machine using plans he received years earlier from a mysterious stranger. He now realizes that the stranger was himself, using the time machine to travel back into the past to give his younger self the plans. A bilking paradox would be created if the time traveler builds the time machine, verifies that it works, and then decides *not* to visit his younger self to hand over the plans. See also **bootstrap paradox**.

Black hole a region of spacetime where gravity is so strong that nothing can escape, including light. Black holes are thought to be created when sufficiently massive stars burn out (see **white dwarf** and **neutron star**) and undergo *gravitational collapse*. A black hole of ten solar masses would have a radius of about twenty miles. Black holes might have been created at the Big Bang singularity and, if so, could theoretically come in any mass and size (a black hole with the mass of the Earth would have a diameter of less than half an inch).

Block universe a spacetime in which all world lines are completely determined from beginning to end (a fatalistic universe). There is no free will in such a spacetime.

Boost matrix matrix formulation of the **Lorentz transformation**.

Bootstrap paradox the puzzle of the origin of *information* on a closed loop in time. The classic example is that of a time traveler from the future giving his younger-self the plans for the time machine the time traveler has just used to visit the past so that he can then build the time machine to visit the past. The time machine plans appear not to have been *created* by anyone! The plans just *are*. See also **bilking paradox**.

Cauchy horizon a **spacelike** hypersurface in spacetime that intersects, exactly once, every **timelike** world line that has no end point. Knowledge of the conditions on such a surface uniquely determines the spacetime at all other points.

Causal loop a time loop containing an event caused by a *later* event that, itself, is caused by the earlier event (see the example in **bilking paradox**).

Causality the metaphysical claim that every event is caused by a prior event. Time travel to the past inherently violates causality.

Chronal regions those parts of spacetime that have no closed timelike curves.

Chronology horizon a (hyper)surface in spacetime that separates **chronal** and non-chronal regions. It is a special case of a **Cauchy horizon**.

Chronology protection the claim, as yet unproved, that time machines and time travel to the past are impossible because of the **back reaction** of spacetime will lead to **stress-energy divergence**. Popularized among physicists as the *Hawking chronology protection conjecture* (1992), Hawking has since admitted that stress-energy divergence is *not* sufficient to enforce his conjecture.

Chronon science fiction name for **Planck time**.

Closed timelike line (or curve) a **timelike** world line of finite length that has no ends, i.e., that forms a *closed loop* in spacetime. A region of spacetime containing closed timelike lines is said to be a **time machine**.

Conservation law physical quantities in interacting systems that remain unchanged are said to be conserved. Total energy, total momentum (linear and angular), and electric charge are conserved quantities.

Cosmic string hypothetical, threadlike spacetime structures with enormous mass-energy and density that may have formed during the **Big Bang**. Cosmic strings may have been initially formed either as infinitely long, or as closed loops, and it is the former that are thought to be physically meaningful in the present-day universe. Cosmic strings do not violate the **weak energy condition** (as do **wormholes**), and they can theoretically create **closed timelike lines**.

Cosmological constant an extra term specifically added by Einstein to the general theory of relativity to keep that theory from predicting the expansion of the universe (which was later observationally found to actually be the case). Einstein subsequently said that his failure to believe the general theory's original prediction of the expansion of the universe was the greatest mistake of his life. The constant (which today is believed to be almost zero, if not exactly zero) appears in Gödel's rotating time travel spacetime as a determining factor in the minimum radius of a **closed timelike line**.

Determinism the metaphysical belief that effects are uniquely determined by causes (this is *not* **fatalism**).

Dirac radio science fiction gadget for sending information at infinite speed, which thus travels backward in time (see also **ultraluminal**).

Dominant energy condition the **weak energy condition** *plus* the claim that the observed energy flux is never **superluminal**.

Electron fundamental particle of mass that possesses one quantum of negative electric charge. *Bound* electrons orbit the nuclei of atoms and plays a central role in determining the chemical properties of the elements and of their compounds. *Free* electrons carry electric current, either in conductors (wires) or through space.

Elsewhen the collection of spactime events that cannot be reached from the **here-now** with a **timelike world line**.

Entropy a measure of the randomness of a system that plays a central role in the thermodynamic **arrow of time**.

Ether a substance once thought to fill all space to allow radiation 'something to propagate through' (as opposed to simply a vacuum). The special theory of relativity showed that the ether is an unnecessary concept because it has no observable effects (physicists argue that if something is impossible to detect, then it is meaningless to talk about it being part of *science*).

Event a point in spacetime.

Event horizon the spacetime surface of a black hole or of a non-traversable wormhole, at which light can *just* escape to the outside universe. It is called a *horizon* because, by definition, an external observer can't see beyond it and into the interior of the hole. To see the inside of a hole you must enter the hole by crossing the horizon (but then you can't get out).

Exotic matter matter that violates one or both of the **weak/strong energy conditions**. Exotic matter appears in the theories of wormholes and warp drives.

Fatalism the metaphysical belief that all events have been *pre*determined from the beginning of time.

Field the concept that if a physical law is local, then it is describable by differential equations that relate what is 'happening' at every point in spacetime to what is 'happening' at its closely located neighboring points. Electromagnetism and general relativity are field theories, for example, described by sets of partial differential equations called *Maxwell's equations* and *Einstein's gravitational field equations*, respectively.

Fourth dimension either time or a fourth spatial dimension.

Frame of reference a spacetime coordinate system.

Free will the condition that prevails when we can *choose* to do what we do. There is no free will in a **block universe**.

Future the collection of spacetime events that can be reached from the **here-now** via a **timelike world line** directed toward a later time (for each individual, the future is what hasn't yet been experienced).

Gamma ray very high-energy, very high-frequency electromagnetic radiation. Gamma rays have frequencies on the order of *ten trillion* (10^{13}) times greater than those of AM radio broadcast radio waves.

General theory of relativity Einstein's theory of curved spacetime, which explains gravity in terms of nothing but geometry. Its fundamental premise is that *all* the laws of physics should appear the same to all observers in *any* **frame of reference**. It is believed the theory will fail when the local mass-energy density reaches a level of about 10^{94} g/cm^3, a density so enormous (the density of water is just 1 g/cm^3) that there is no known mechanism for achieving it anywhere in the universe except in another **Big Bang**. See also **Planck density**.

Geodesic the shortest path connecting two points in space (if the space is spacetime, the world line of a particle in free-fall).

Global in the large.

Gödel universe a spacetime that, unlike the one we live in, is rotating so fast that it automatically generates *closed timelike lines* and thus constitutes a weak **time machine**. In such a universe, time travel to the past would be a natural phenomenon.

Grandfather paradox *the* classic time travel paradox, of a time-traveler killing, while in the past and *before* the time traveler has been conceived, an ancestor directly linked to the future birth of the time traveler. A more direct form of this sort of paradox is simply the time traveler killing his own younger self (called the **autoinfanticide paradox**).

Gravitational field equations a set of coupled, partial differential, non-linear tensor equations, considered to be the most complicated equations in all of mathematical physics. They show how the local curvature of spacetime depends on the local mass-energy of spacetime. The equations are independent of the **topology** of spacetime.

Gravitational lensing the ability of gravitational fields to bend and focus light.

Graviton the quantum particle of gravity.

Hawking radiation the emission of particles (energy) by a **black hole** into the region *outside* its **event horizon**, which results in the eventual evaporation of the hole. This is a quantum mechanical effect.

Here-now the point or **event** (for each observer) in spacetime that separates the **past**, the **future**, and **elsewhen**.

Hyperspace any space of four or more dimensions (for example, four-dimensional spacetime is a hyperspace).

Inertial frame any frame of reference in which Newton's laws of mechanics are true (there are no *acceleration forces* in inertial frames, and so *rotating* or 'merry-go-round' frames are not inertial).

Invariance a quantity that remains the same in any frame of reference is an invariant. Two examples are the distance between any two points on a piece of paper (because it is independent of any particular coordinate system), and the speed of light.

Kerr-Newman black hole a *rotating* black hole, which may (or may not) be electrically charged.

Krasnikov tube a particular spacetime **metric** (or *warp*) allowing **superluminal** travel, with the great difficulty of requiring *enormous* negative energy. Two Krasnikov tubes can be made into a time machine. Named after its Russian inventor.

Least action general principle in physics that asserts the world line of a particle is the one that *minimizes* the **action**.

Light cone the **lightlike** surface in spacetime that, at each point in spacetime, separates the **past** from the **future** from **else-when** from the **here-now**.

Lightlike the world line of a photon (or of any other form of mass-energy traveling *at* the speed of light).

Li mirror a perfectly reflecting, spherical surface that can be used to stabilize a wormhole against energy loops circulating through a **wormhole** time machine (thus creating unbounded energy levels that destroy the time machine). Named after its Chinese inventor.

Local in the small.

Lorentz factor the ubiquitous square-root expression that appears in so many relativistic calculations, such as time dilation, length contraction, and the variation of mass with speed. For example, the mass m of a moving body is not independent of it speed v but rather varies as $m = \frac{m_0}{\sqrt{1-\left(\frac{v}{c}\right)^2}}$, where m_0 is the rest mass (that is, the mass when $v = 0$) and c denotes the speed of light (186,210 miles per second). The denominator is the Lorentz factor.

Lorentz-FitzGerald contraction the conclusion from special relativity that the appearance (to a stationary observer) of a moving object will be shortened in length along the direction of motion. Many years after Einstein's work, it was shown that the object will also appear to be *rotated*.

Lorentz transformation equations from the special theory of relativity that describe how the space and time measurements of two relatively moving observers are related.

Many-worlds interpretation quantum mechanical view of splitting universes.

Mass-energy the famous $E = mc^2$, the equation behind atomic fission and nuclear fusion weapons.

Metric the measure of the separation between any two events in a spacetime.

Minkowski spacetime the flat spacetime of the special theory of relativity. In this spacetime there is no gravity, no spacetime curvature (hence it is *flat*) and no backward time travel.

Neutron star the end state of a star with one to three solar masses that has collapsed to a density of up to 10^{17} g/cm^2.

Non-Euclidean geometry the geometry of spacetime, whether curved or flat. Spacetime is non-intuitive precisely because it is always hard to resist thinking in terms of high school Euclidean geometry, which is simply the *wrong* geometry.

Null geodesic the world line of a photon in spacetime.

Observer physicist's term for 'somebody' equipped with recording instruments (such as a clock, a pencil and notepad, and the like).

Parallel transport a procedure for moving a vector around any closed curve in a space to determine whether that space is flat or curved.

Parallel worlds simultaneous existence of multiple (perhaps) infinite versions of reality.

Past the collection of spacetime events that can reach the **here-now** via timelike world lines directed from an earlier time (for each individual, the past is what has already been experienced).

Photon the quantum particle of electromagnetism. A photon of frequency f has energy hf, where h is **Planck's constant**.

Planck density the density of mass-energy that distinguishes classical from quantum spacetimes; about 10^{94} g/cm^3, equal to the **Planck mass** divided by the cube of the **Planck length**.

Planck length the non-zero length in quantum theory (about 1.6×10^{-33} cm) below which quantum gravity effects will become important.

Planck mass the fundamental mass in quantum theory (about 22×10^{-6} g), but *not* the smallest non-zero mass in quantum theory.

Planck's constant fundamental constant in quantum theory, h, associated with the discrete nature of quantum effects. (If h had the value of zero, rather than its actual value of about 6.6×10^{-34} joule-seconds, then the microworld would appear to be continuous.)

Planck time the time interval in quantum theory (about 5.3×10^{-44} s) below which quantum gravity effects become important. The time required to travel the **Planck length** at the speed of light.

Positron the electron's anti-particle (see **anti-matter**).

Proper time the timekeeping of an observer's clock.

Pulps the old science fiction magazines, through the 1940s and into the early 1950s or so, published on inexpensive, wood-pulp paper.

Quantum foam see **topology**.

Quantum gravity the yet-to-be-discovered theory that unifies quantum field theory with the curved spacetime of general relativity.

Quantum mechanics the exact physics of the very small (atoms and things smaller).

Quantum theory any theory in which physical quantities are not continuous but rather assume their values in discrete jumps (the size of the jump is the *quantum*).

Recurrence paradox the claim that if you wait long enough, then every system will return to every previous state infinitely often.

Red dwarf small (less than about half a solar mass) star with a very long life (hundreds of times that of the Sun). They are 'cool' stars, with a surface temperature less than 4000 °C, and are thought to be the most common type of star in the universe.

Red shift the *down* shift in frequency of light received from all distant stars due to the Doppler effect induced by the expansion of the universe. The opposite effect is called a *blue shift*.

Reinterpretation principle asserts that negative mass-energy traveling forward in time is positive mass-energy traveling backward in time, and vice-versa.

Reissner-Nordström black hole a spherically symmetric, non-rotating electrically charged black hole.

Reversibility paradox based on the fact that the equations of physics contain no **arrow of time**; that is, they work equally well with time running forward or backward.

Roman ring a time machine made of two or more traversable wormholes connected in a closed sequence.

Schwarzschild black hole a spherically symmetric, non-rotating, uncharged black hole.

Self-consistency the assertion that the events on a closed **timelike** line must never be in contradiction; generally attributed to the Russian physicist Igor Novikov, who with his colleagues showed that it is not an independent assumption but rather an implication of the principle of **least action**.

Sexual paradox a special type of causal loop, where the connected events on a time loop are ‘coupled’ (pun intended!) through reproductive sex. An example is a time traveler to the past who becomes her own ancestor.

Singularity either a region in spacetime where the curvature becomes infinite and the laws of physics fail, or a point in spacetime beyond which world lines cannot be extended. Singularities of the first kind are called *curvature* or *crushing* singularities, and those of the second kind are called incomplete singularities. The **Big Bang** was a curvature singularity, as is the center of a black hole. In a **Schwarzchild black hole** the curvature singularity is a point, whereas in a **Kerr-Newman black hole** it is an extended region in the form of a ring.

Spacelike a world line on which propagating mass-energy would exceed the speed of light.

Spacetime the ‘stuff’ out of which reality is built. Everything there is—the universe—is the total collection of events in spacetime. A *flat* spacetime has no gravity, whereas a curved spacetime is the *origin* of gravity.

Special theory of relativity Einstein’s theory of *flat* spacetime, which assumes that gravity is absent (gravity is the result of the geometry of *curved* spacetime). Its fundamental premise is that the laws of physics should appear the same to observers in different **inertial frames**.

Splitting universes the idea that every decision causes reality to split into separate copies, identical in every respect except for each of the different possible results of the decision.

Stargate science fiction name for the mouth of a traversable **wormhole**.

Stress-energy divergence the unbounded growth of the **general theory of relativity**’s measure of the density of mass-energy in spacetime.

Strong energy condition the claim that gravity is always (that is, locally) attractive. A traversable **wormhole** violates this condition.

Subluminal slower than light.

Superluminal faster than light.

Tachyon a particle (hypothetical, so far) that always travels faster than light, so its **world line** is always spacelike.

Temporally orientable spacetime any spacetime in which the direction of time at every point agrees with the direction of time at its local neighboring points.

Tensor mathematical generalization of the scalar and vector concepts. Einstein's **gravitational field equations** are tensor-differential equations (for example, the metric tensor contains information about the curvature of spacetime), whereas Newton's and Maxwell's equations are vastly less complex vector-differential equations.

Tidal force force experienced by a non-point mass (one with spatial extension) in a non-uniform gravitational field. Such forces tend simultaneously to compress and stretch spatially extended masses. Black holes and wormhole mouths can generate enormous tidal forces on extended masses as small as a human body. Interestingly, the *more* massive a black hole, the *less* severe its tidal forces are at distances outside the **event horizon**. However, no matter what the black hole mass is, the tidal forces are infinite at the central curvature **singularity**.

Time dilation the altering of the rate of timekeeping by a clock, either by motion or by gravity.

Time machine (in the weak sense) a machine able to traverse **closed timelike world lines** inherent in a spacetime (e.g., a rocket in Gödel spacetime) but unable to *create* such world lines; (in the strong sense) a machine able to manipulate mass-energy in a finite or compact region of spacetime in such a way as to *create* closed timelike world lines.

Time police story characters in science fiction charged with the (unnecessary!) job of preventing time travelers from changing the past.

Time warp science fiction name for a **time machine**.

Tipler cylinder an infinitely long cylinder, made of super-dense matter, rotating so fast around its long axis that it warps spacetime enough to create closed timelike lines that encircle the cylinder. It can be used as a strong sense time machine to travel both into the future and into the past (but *not* to a time before the creation of the cylinder).

Topology the structure of a *space* (including *spacetime*) without regard to a metric. That is, topology is concerned only with how a space is connected together and not with how far apart points in the space are. Topologists consider stretching or compressing a space to be irrelevant, just as long as one doesn't *tear* it and so put holes in the space. The simplest topology is that of a *simply connected* space, in which if you construct any closed surface that lies totally in the space around any point in the space, then every other point inside the surface is also in the space. A space with a hole in it fails this test, and so is said to be *multiply connected.* A *quantum foam* spacetime has a multiply connected topology. The classical spacetime of general relativity is simply connected *until* the appearance of wormholes.

Twin paradox the conclusion from special relativity that a clock's rate of time keeping slows with motion.

Ultraluminal motion sufficiently **superluminal** that mass-energy appears to travel backward in time (see also **Dirac radio**).

Uncertainty principle the statement in quantum mechanics that says certain pairs of quantities cannot simultaneously be measured with arbitrarily small error. The position and momentum of a particle are one such pair, and energy and time are another.

Vacuum fluctuation the particle/anti-particle creation and annihilation processes allowed, even empty space, by the **uncertainty principle** of quantum mechanics.

Warp drive science fiction name for the propulsion mechanism of a faster-than-light spaceship, now commonly used by physicists, too.

Weak energy condition the claim that the observed mass-energy density is always (locally) non-negative. Quantum mechanics predicts (and it has been experimentally confirmed) that there are exceptions.

White dwarf a burnt-out star with a mass less than 1.4 solar masses, of planetary size with a density up to 10^7 g/cm^3. The ultimate fate of our Sun.

World line the trajectory of mass-energy in spacetime.

Wormhole a spacetime structure (violating the **weak and strong energy conditions**, if traversable) connecting two points of the same spacetime (or even two *different* spacetimes) with a timelike path that requires less time to travel along than does a photon traveling *outside* the wormhole between the two points. A wormhole is *traversable* if it has no **event horizons**, and such wormholes can apparently be made into a **time machine** (sometimes called a *time tunnel*) using a time shift (see **time dilation**) between the two mouths of the wormhole *unless* quantum effects forbid time machines (still an open question).

Index

A
Absorber theory, 269–273
Advanced waves, 259, 260, 264, 265, 269–273, 283
Alcubierre warp drive, 329, 331–333. *See also* Natário warp drive
“All You Zombies—,” 21, 194, 227, 239
“Ambition,” 36
“Anachronistic Optics,” 260
“Ancestral Voices,” 204
Anderson, Poul, 41, 155
“—And He Built a Crooked House,” 88
Antimatter, 111, 157, 247, 248
Anti-telephone, 261, 273–282
Aquinas, St. Thomas, 63, 92, 102, 103, 196
Aristotle, 53, 54, 56, 63, 66, 79, 96, 195, 208
Ark, 333
Arrow of time. *See specific arrow* (cosmological; electromagnetic; psychological; thermodynamic)
Asimov, Isaac, 12, 37, 154, 207–209, 236, 357, 358
“As Never Was,” 220
Aspect, Alain, 281, 282
Asymptotically flat, 171, 305, 317
Augustine, St., 51, 55, 109, 181
Autoinfanticide paradox, 112, 206–207, 212, 229, 238. *See also* Grandfather paradox; Sexual paradox
Average energy condition, 305–308
“Azimuth 1, 2, 3 . . .,” 42

B
Back reaction, 30, 309, 317
Back to the Future, 3, 23, 162, 214, 231, 244
Backward causation, 197, 198
Barrow, Isaac, 57
“Beep,” 76, 236, 273, 274
Behold the Man, 211, 243
Bellamy, Edward, 250
Bell, J.S., 232, 233, 279
Benford, Gregory, 24, 228, 229, 273, 318
Bergson, Henri, 54
Berkeley Square, 3, 99, 101
Besso, Michele, 90, 93
Bid Time Return. See Somewhere in Time
Big Bang, 37, 46, 62, 63, 66, 135, 136, 221, 252, 271, 272, 317, 318
Big Crunch, 46, 135, 138
Bilking paradox, 76, 77, 189, 249, 261, 268, 269, 272, 277, 286, 290, 296, 318
Bill & Ted’s Excellent Adventure, 3, 224
“Biography Project, The,” 242
“Birch Clump Cylinder. The,” 216
Black hole, 32–38, 41, 43, 84, 138, 283, 302, 311, 320
Black, Max, 101, 120
“Blindman’s World, The,” 250
Blish, James, 28, 235, 272, 273
Block universe, 90–109, 116, 121–122, 126, 197, 246, 261
Blue shift, 144
Boethius, 107
Boltzmann, Ludwig, 127, 130, 131, 134
“Bombardment in Reverse,” 253
Bonaventure, St., 63
Bootstrap paradox, 26, 39, 43, 214–225, 239, 241, 242
Bradbury, Ray, 52, 200
Bradley, Francis, 94–95, 121, 246

© Springer International Publishing AG 2017
P.J. Nahin, *Time Machine Tales*, Science and Fiction,
DOI 10.1007/978-3-319-48864-6

"Branches of Time, The," 234
Bring the Jubilee, 236–237
Brown, Fredric, 188
"By His Bootstraps," 187, 362

C
Campbell, John, 235
Cantor, Georg, 63
Carroll, Lewis, 125
Casimir effect, 306–308, 330
"Castaway," 45
Cauchy condition, 136
Cauchy horizon, 309–312, 314, 316–317
Causal loop, 40, 76, 214–227, 231, 232, 239–240, 242, 249, 262, 270, 276, 310, 318, 320, 323
Cause-and-effect, 67–77, 118, 138
Cerenkov radiation, 291
Chandrasekhar, S., 31, 360
Chaucer, Geoffrey, 107
"Child by Chronos," 227
"Child's Play," 2
"Chronokinesis of Jonathan Hull, The," 127
Chronology protection conjecture
  Hawking, 13–14, 42, 45, 284, 311, 315, 317, 325
  partial rebuttal by Thorne, 42
Chronon, 30, 57. *See also* Planck time
"Circle of Zero, The," 134
Civilization, arbitrarily advanced, 303, 336
Clarke, Arthur C., 10, 11, 15, 22, 110, 111, 133, 239
Clifford, William Kingdon, 169
Clock hypothesis, 174, 179
"Clock That Went Backward, The," 199
Closed time loop (CTL). *See* Causal loop
Connecticut Yankee in King Arthur's Court, A, 198
Contact, 302, 318
"Cosmic Corkscrew," 12
Cosmic string, 320–324
Cosmic string time machine, 163, 319–326
Cosmological arrow of time, 137–138, 271
Counter-Clock World, 131, 251
Cramer, John, 82, 182, 269, 272, 282, 285, 291
Cumulative audience paradox, 10

D
Damian, Peter, 195–196
"Dandelion Girl, The," 73, 210
Daniels, David R., 234–235
"Dark Tower, The," 18
De Camp, L. Sprag, 188, 208, 209, 230
Dechronization of Sam Magruder, The, 260
Descartes, 57
Determinism. *See* Free will
Deutsch, David, 222, 233–234
Deutsch-Politzer spacetime, 326
DeWitt, Bryce, 277
Dick, Philip K., 131, 243, 251, 283
Dirac, Paul, 74–76, 127, 167, 168, 263–264, 267
Dirac radio, 76, 264, 265, 270, 272–273
Distance function. *See* Metric
Dominant energy condition, 317, 329
Doomsday Book, 5
Door Into Summer, The, 11, 180, 221
Dyson, Freeman, 359

E
Earman, John, 18, 70, 72–73, 75, 96, 158
Eddington, A.S., 127, 132, 135, 250, 253–254, 360
"E for Effort," 16
Ehrenfest, Paul, 80, 129
Einstein, Albert, 1, 17, 32–33, 44, 53, 54, 57–58, 60–61, 79, 84, 90, 93, 95, 100, 101, 116, 123, 126, 145, 146, 149, 153, 168, 174, 250, 258, 262, 279–282
Einstein-Rosen bridge. *See* Wormhole
Einstein's Bridge, 84
Electromagnetic arrow of time, 138, 267, 271, 272
"Elsewhen," 7, 253
Energy (conservation of in spacetime), 99–100
Energy condition. *See specific condition* (averaged; dominant; strong; weak)
Entropic clock, 129
Entropy, 127–134, 136–138, 271
Epictetus, 106
EPR paradox, 281, 282
Eternal return, 134
Event horizon, 33, 86, 146, 264, 302, 320
Everett, Hugh, III, 229, 231, 233–235
Exotic matter, 304–306, 308, 318, 320, 325, 329, 334, 335
"Experiment," 25

F
"Face in the Photo, The," 7
"Far Centaurus," 4
Faster-than-light (FTL), 4, 34, 55, 56, 116, 141, 142, 162, 262–264, 273–277, 279, 281,

282, 289–296, 327–329, 331, 333, 334, 361. *See also* Tachyon(s)
Fatalism. *See* Free will
Feinberg, Gerald, 274–277
Fermi, Enrico, 10, 279
Feynman, Richard, 37, 126, 157–159, 181, 259, 264–270, 272, 277–278
Fifth dimension, 48, 83, 87, 105, 110, 167–168, 182, 229, 253. *See also* Hyperspace
“Fifth-Dimension Catapult, The,” 167
Final Countdown, 43, 240
Findlay, J.N., 246, 247
“Find the Sculptor,” 219
First Cause, 62
“First Time Machine,” 205
Flammarion, Camille, 55–56
Flamm, Ludwig, 84
“Fools’ Errand,” 224
Forever War, The, 38, 43
Forever Young, 180
Forward, Robert, 184, 283, 307, 312
“Four Dimensional Surgery,” 83
Fourth dimension, 78–90, 92, 109, 110, 182, 215, 250. *See also* Hyperspace
“Fox and the Forest, The,” 6
Free will, 60, 93, 94, 100, 102–104, 106–109, 124, 196, 198, 213, 217, 233, 238, 268, 278–279, 312, 313, 318, 363
FTL. *See* Faster-than-light (FTL)

G
Galilei, Galileo, 147, 152
“Garden of Forking Paths, The,” 236
Geodesic, 155, 159, 175, 305, 310, 330. *See also* World line
Gernsback, Hugo, 191–194, 209, 225
Gödel, Kurt, 20, 21, 28, 44, 101, 116–117, 158, 212, 276, 297–299, 301, 358–360
“Gostak and the Doshes, The,” 83
Gott, J. Richard, III, 320–325, 359. *See also* Cosmic string time machine
Grandfather paradox, 45, 200, 201, 204–206, 211–214, 220, 233, 312, 313, 326–327, 336, 363. *See also* Autoinfanticide paradox; Sexual paradox
Gravitational collapse, 31, 35, 360
Gravitational constant, 30
Gravitational field equations, 168, 252, 320, 327
Gravitational time dilation. *See* Time dilation
Groundhog Day, 219, 327
“Gun for Dinosaur, A,” 8, 188
“Gun for Grandfather, A,” 195, 206
Gunn, James, 359
Guns of the South, The, 237

H
Haldeman, Joe, 38, 43
Hale, Edward, 198, 229, 230
“Hands Off,” 198, 229
Hawking, Stephen, 13, 14, 32, 34–36, 52, 57, 138, 271, 284, 285, 315, 317, 320
“Hawksbill Station,” 7
Heaviside, Oliver, 291
Hegel, Georg, 94, 95
Heinlein, Robert, 11, 21, 85, 88, 98, 180, 187, 194, 198, 209, 221, 228, 239, 253
Heyl, Paul, 256, 257
Hidden variables, 281, 282
Hinton, Charles, 112, 113, 257
“History in Reverse,” 8
Horwich, Paul, 51, 213
Hoyle, Fred, 121, 257, 258, 271, 272
H-theorem, 127
Hume, David, 15, 16, 18, 19, 59
Hutton, R.H., 55
Hyperspace, 79, 82, 83, 86, 87

I
Independence Day, 334
Inertial frame of reference, 148
  dragging of, 297
Inflation of spacetime, 136, 317, 328
Interstellar, 3, 27, 43, 48, 86, 110, 143
Interval of spacetime, 166, 173–180, 299
Invariant, 165, 166, 173
Invisibility (of time travelers), 192, 194

J
James, William, 94, 95, 101
Jeans, Sir James, 84, 170, 250
Je t’aime, Je t’aime, 55

K
Kaluza-Klein theory, 167
Kant, Immanual, 65, 80
Kerr, Roy, 33
Khayyam, Omar, 20, 44, 101
King, Stephen, 2, 27, 85
K-mesons, 69, 70, 125, 127, 130
Knight, Damon, 42
Kornbluth, Cyril, 28

Krasnikov, Serguei, 317, 327, 328, 330, 331, 359
Krasnikov warp, 331

L
Legion of Time, The, 230, 235
Leibniz, Gottfried, 53, 54
Leinster, Murray, 88
Lem, Stanislaw, 361
Lest Darkness Fall, 99, 230
"Let's Go to Golgotha!," 8
Lewis, C.S., 17–18, 197
Lewis, David, 221
Lewis, G.N., 77, 126, 183, 267, 268
"Life-Line," 98
Light cones, 159–162, 183, 279, 297–300
"Lightship, Ho!," 264
Li, Li-Xin, 310
"*Lincoln Hunters, The*," 7
"Little Black Bag, The," 2
Looper, 3
Lorentz-FitzGerald contraction, 141, 142
Lorentz transformation, 149–153, 163, 165–167, 289, 291–295
"Lost," 236
"Love Letter, The," 224
Lucretius, 31, 62, 274

M
MacBeath, Murray, 247–249
"Man Who Folded Himself, The," 201
"Man Who Lived Backwards, The," 131
"Man Who Lived Next Week, The," 132
"Man Who Met Himself, The," 219
"Man Who Saw the Future, The," 111
"Man Who Saw Too Late, The," 72
"Man Who Went Back, The," 217
Many-worlds interpretation (MWI) of quantum mechanics, 229–237, 279, 284, 285
"Mapmakers, The," 85
Mastodonia, 8
Maxwell, James Clerk, 127, 148, 169
Maxwell's equations, 148, 149, 256–265
McTaggart, John, 58–60, 120
Meiland, Jack, 254–256
"Merchant and the Alchemist's Gate, The," 239
Metric, 164–173, 182, 298, 303, 311, 328, 331, 332
Michelson-Morley experiment, 148, 257
"Middle of the Week After Next, The," 89
Miller, P. Schuyler, 191
"Mimsy Were the Borogoves," 2
Minkowski diagram, 155–157, 243
Minkowski, Hermann, 90, 91, 94, 95, 102, 106
Minority Report, 93
"Minus Sign," 37
Möbius strip, 28, 78, 82, 111, 220
"Mr. Peabody's Improbable History," 364
"Ms. Found in an Abandoned Time Machine," 5
"My Name is Legion," 238

N
Natário warp drive, 332, 333
Negative energy, 296, 305–307, 331. *See also* Exotic matter
Neutron star, 31, 84, 144
Newcomb, Simon, 46, 79, 80, 90, 113, 182, 362
Newman, Ezra, 33
Newton, Isaac, 57, 147–148
Nicholson, Jim, 193–194, 221, 225
Nicolson, Marjorie Hope, 1
Nieremberg, Juan Eusebius, 51
Nietzsche, Friedrich, 134, 135
Niven, Larry, 6, 12, 13, 41, 144, 187, 188, 292
Nordström, Gunnar, 33, 266
Novikov, Igor, 35, 209, 210, 314, 318, 319, 335, 359
Now, 101, 116, 117, 122–124, 132

O
Olsen, Bob, 81, 86
Omniscience, 102, 103, 105, 107. *See also* Block universe; Free will
"Other Tracks," 231, 235
"Out Around Rigel," 3
"Over the River & Through the Woods," 6

P
Page, Leigh, 258, 259, 266
Panchronicon, The, 224
"Paradox," 39
Paradoxes. *See* Particular paradox (autoinfanticide; bilking; bootstrap; causal loop; cumulative audience; grandfather; recurrence; sexual; twin)
"Parallel in Time," 231
Parallel transport (of vectors), 171
Parmenides, 92
Past
  affecting/changing, 104, 195–214, 357
  communicating with, 256–264
  duration of, 63, 65
  praying about, 196–198

"Past and Its Dead People, The," 5
Peggy Sue Got Married, 4, 357
Peirce, Charles Sanders, 79–80
"Pete Can Fix It," 332
Photon clock, 139, 140, 151
Pierce, John, 259–260
Planck, Max, 17, 210
Planck density, 31
Planck length, 30, 301, 303, 322, 331
Planck mass, 30, 322
Planck's constant, 30, 35, 143
Planck time, 30
Plato, 52–54, 56, 245, 251
"Plattner Story, The," 111, 250
Pohl, Frederik, 250–251
Poincaré, Henri, 133–135, 152
Politzer, H.D., 326
"Poor Little Warrior," 8
Popper, Karl, 93, 94, 131, 222, 257, 364–365
Positron, 157–159, 181
"Poundstone Paradox, The," 227
Poynting, J.H., 56
Presentism, 109
"Pre-Vision," 259
"Project Mastodon," 6
Proper time, 3, 4, 133, 173–180, 188, 248
Psychological arrow of time, 116, 124, 137
Pulp magazines, 331–333, 361
Putnam, Hilary, 156–158, 249–250

Q
Quantum foam, 301, 304
Quantum gravity, 14, 29, 37–38, 48, 254, 305, 311, 314–316
Quantum inequalities, 330
Quantum mechanics, 37, 70, 80, 160, 231, 285

R
"Race Through Time, A," 183
Reciprocity theorem, 70
Recurrence paradox, 130, 133
"Red Queen's Race, The," 207
Red shift, 143–144
Reinterpretation principle, 275–277, 296, 331
Reissner, Heinrich, 33
Relativity (general theory), 1, 31, 35, 37, 70, 80, 91, 135–136, 160, 162, 167, 169, 174, 212, 232–234. *See also* Gravitational field equations
Relativity (special theory), 3, 55, 57, 90, 122, 152, 160, 173–174, 252, 263, 290
"Rescuer, The," 201
Reverse causation. *See* Backward causation
"Reverse Universe," 273
Reversibility paradox, 130, 133
Riemann, Bernhard, 169, 170
Ritz, Walter, 258, 265, 271
Roman ring, 315–316, 323
"Rotating Cylinders and the Possibility of Global Causality Violation," 188
Rotating universe. *See* Gödel, Kurt
"Runaway Skyscraper, The," 88
Russell, Bertrand, 78

S
Saari, Oliver, 42
Sagan, Carl, 27, 302
"Sam, This is You," 224
"Sands of Time," 12, 47
Schrödinger, Erwin, 229
Schwarzchild, Karl, 33
Self-consistency principle, 131, 209–211, 218, 270, 312–314, 357–358, 363
Sexual paradox, 225–229. *See also* Grandfather paradox
"Sidewise in Time," 31, 232
Silverberg, Robert, 4, 10–11, 133, 247
Simultaneity, 67, 104, 122, 126, 150, 164
"Singularities Make Me Nervous," 34
Skepticism (concerning time travel), 15–22
"Skull, The," 243
Sky Splitter, The, 263, 334
Somewhere in Time, 48, 217, 219, 220
"Sound of Thunder, A," 8, 200
Spaceballs, 54
"Space Bender, The," 85
Spacelike, 160, 164, 183, 294, 298, 316
Spacetime
  diagrams (*see* Minkowski diagram)
  fifth dimensional (*see* Hyperspace)
  four dimensional, 60 (*see also* Block universe)
  Gödelian (*see* Gödel, Kurt)
  singularities, 32–34, 38, 41, 48, 63, 135–136, 138, 233, 264, 284, 311, 320
  topology of, 27–28, 30, 84, 301–302
Spinoza, 53, 54
Spirit of '76, The, 7
Splitting universes. *See* Many-worlds interpretation (MWI) of quantum mechanics
"Star, Bright," 28
Star Trek IV, 39, 43, 329–330

Strong energy condition, 306, 317, 329
Strong time machine, 297, 324
"Such Interesting Neighbors," 6
Superluminal. *See* Faster-than-light (FTL)
Superman, 289
Swift, Jonathan, 46

T

Tachyon(s), 24, 108, 274–279, 282, 286, 290–291, 331. *See also See* also Faster-than-light (FTL); Ultraluminal
"Tachypomp, The," 274
"Target One," 69
TCP theorem, 69, 181
"Technical Error," 88, 110
Technicolor Time Machine, The, 9, 216, 219
"Temporary Warp," 98, 132
Terminator, The, 3
Tetrode, Hugo, 265–268
Thermodynamic arrow of time. *See* Entropy
"This Way to the Regress," 251
"Thompson's Time Traveling Theory," 211, 238
Thorne, Kip, 27, 42, 301–305, 308, 310, 312, 314, 315, 359
"Three Sundays in a Week," 224
"Throwback in Time," 94
"Tides of Time, The" (Chandler), 215
"Tides of Time, The" (Williams), 34
Time. *See also* Arrow of time
  circular and spiral, 11, 12, 56, 136
  direction of, 115–124
  flowing, 118–120
  linear, 56, 57, 61–67, 136
  multidimensional, 121, 252–256
  reality of, 56, 58, 79, 93, 117
  running backwards, 130–131, 245–252, 283
"Time Annihilator, The," 120
"Time Bender, The," 42, 99, 208
Time Bomb, 251
"*Time Cheaters,* The," 23, 76–77, 189, 215
"Time Contractor, The," 307
Time dilation
  by gravity, 143–146, 309
  by motion, 139–143, 145, 174, 183–184, 301
"Time Eliminator, The," 190, 242
"Time for Sale," 132
"Time Hoaxers, The," 17
Timelike, 159, 183, 299, 305
Timelike Infinity, 336
Time machines. *See specific type* (black hole; cosmic string; strong; Tipler cylinder; warps; weak; Wellsian; wormhole)
Time Machine, The, 3, 24, 45, 88, 89, 94, 101–103, 191, 198, 199, 310, 359, 360
Timemaster, 312
"Time Oscillator, The," 192
"Time Out of Mind," 176
"Time Patrol," 212, 226
Time police, 212–213, 363
"Time Ray of Jandra," 194
"Time's Arrow" (Clarke), 10, 133, 239
"Time's Arrow" (McDevitt), 10, 365
Timescape, 24, 273
"Time Ships, The," 299
Timeslip, 72
"Time Twister," 242
"Time Wants a Skeleton," 191, 210
Tipler cylinders, 38–42, 135, 162, 188, 297–300, 359
Tipler, Frank, 38, 40–42, 46, 135, 301, 358
"To the Stars," 142, 153–154
"Trouble With the Past, The," 123
"Tryst in Time," 200
Twin paradox, 173–180, 184, 308, 328
"Typewriter from the Future," 31, 206, 212

U

Ultraluminal, 275, 295–296. *See also* Faster-than-light (FTL)
Uncertainty principle, 35–36, 306, 330
Universal nocturnal expansion, 8
"Up the Line," 8, 10

V

Vacuum fluctuation, 36, 304, 315
Vaughn, Henry, 140
Verne, Jules, 81, 335
"Very Slow Time Machine, The," 26, 248
"Via the Time Accelerator," 23, 139, 191, 214
Visser, Matt, 14, 316
Vonnegut, Kurt, 107

W

Wallis, John, 64
"Wanderers of Time," 120
Ward, Seth, 64
Warp factor, 328, 333

Warps (spacetime), 98, 132, 162, 326–333
Weak energy condition, 41, 280, 305–306, 325, 329
Weak time machine, 297, 317, 324
"Weapon Out of Time," 215
Wells, H.G., 9, 23, 61, 76, 79, 87, 89–91, 111–112, 123, 182, 198, 199, 235, 250, 310, 335
Wellsian time machine, 23–27
Wessel, John, 2
Weyl, Hermann, 44, 101–102
"What We Learned from This Morning's Newspaper," 19, 104, 133
Wheeler, John, 31, 37, 46, 84, 91, 129, 157, 233, 251–252, 259, 264–270, 272, 277–278, 286, 312
"When the Bough Breaks," 87
"When the Future Dies," 27, 162
Williams, Donald, 96–97, 100, 255
World Jones Made, The, 283
World line, 26, 29, 98, 101–102, 155–162, 175, 183, 228, 279, 292. *See also* Geodesic
"Worlds of If, The," 245
Wormhole, 2, 27, 84–85, 233, 239, 301–319, 359
Wormhole time machine, 136, 144, 169–170, 223, 233, 269, 301–319

Y
Year of the Quiet Sun, The, 3
Yoshikawa, Shoichi, 296

Z
Zangwill, Israel, 55
Zermelo, Ernst, 134

The Unveiling of the Hidden Knowledge and the Secret Space Program

The Unacknowledged Special Access Programs: Advanced Technology, Mind-Control, Spiritual Power and the Corruption behind Closed Doors

By Aug Tellez

Introduction....................7
  Getting This Out Of The Way....................7
  Psychic Operation....................8
  A Light for the Others....................8
  Natural Security....................8
  A Balance of Mystery and Truth....................9
Chapter 1: The World Situation....................9
  What is happening?....................9
  The Mind Virus....................9
    The Vampire Effect....................10
    Sub-Human Entities....................11
    Human-Hosts, Replicants, Synthetics and Carriers....................11
    Genetic Manipulation....................12
    The Ego Mind....................13
    This Time is about Healing....................13
    Healing Through Unity....................14
  Underground Bases....................14
  Cloning....................14
  Celebrity Cloning....................15

        - Initial Explanation 15
        - Cloning Centers, Underground Bases 15
        - Programming 15
        - Experimental Operations 15
        - Genetic Engineering 16
        - Immortality 16
        - How did it get to this? 16
        - Deeper Meaning 17
        - What is next? 17
- Chapter 2: The Awakening 18
- Chapter 2.1: The Machines of All Time and Space 18
    - Re-Creating The Kaleidoscope 18
    - Sand-Castles of Time 18
    - We’re Waiting For Ourselves 18
    - The Goal of Two Societies 19
    - Our Forgotten Other Half 19
    - All Is But An Illusion 20
    - Life Itself Is The Highest Illusion 20
- Chapter 2.2: The “Awakening” of 2010 21
    - Memory Suppression 21
    - Life and Consciousness 22
    - The True Human Potential 22
    - The Light-Body 23
    - Technology of the Awakening 23
    - Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness 23
    - Time Dilation and Temporal Recurrence 24
    - Crystal Technology 24
    - Time Crystals 25
    - Quantum Tunneling Diodes 25
    - Tachyon Fields 25
    - Upgraded Chronovisor 26
    - Time Tunnels 26

The “Trip” Chair.........26
Kozyrev Mirror.........27
Universal Recurrence.........27
Remove Viewing, Temporal Lensing.........27
Temporally Linked.........28
Natural Vortex Energy Locations and Geography.........28
    Mount B.........28
Particle Accelerator Temporal Manipulation.........28
Riding the “Timewave”.........29
Dematerialization and Rematerialization.........29
“Jump Rooms.........29
Ion Shields.........30
Instantaneous Healing.........30
“Living Light” Sound-Harmonics and Electromagnetic Healing.........30
Electrogravitic Craft.........30
Energy Cores and Storage Capacity.........31
3.5D Replicators.........31
3.5D Sensors.........31
3.5D Printing of Clones.........31
Cybernetics:.........31
Neural-interlace/4D Connectome Recording and Simulation.........31
EEG Heterodyning.........32
Artificial Hallucinations.........32
Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator).........32
Utilizing The Temporal Body.........33
    Manipulation Through the Temporal.........33
Astral or Light Body.........34
Monitors can emit consciousness/brainwave altering frequencies.........35
The Upgraded Chronovisor.........35
Community Servers/Living Server Systems.........36
Cloning, Cloud Supercomputing and Consciousness Transfer.........36
Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel.........37

Cloning as Offspring......38

Inserted Memories, Digital Dreams......38

Co-Inhabitation of Bodies......38

Clever Replication a Digital Mirror......38

Necromancy and Soul Stealing......39

Ancient Technology, Non-Human Entities......39

Created Beings......39

Black Goo – AI Nanites Vampiric Hive Mind......40

  Recreating or Replicating the Universe......41

Ancient Spiritual Technology......41

The Unveiling, Underground Bases, Increasing Awareness......42

Learning about the Unveiling......42

Giant “god-men”......42

Breakaway Civilizations......43

Ancient Civilizations......43

  Atlantis......43

  Aegypt......44

  Ancient Technology......44

The True Elite......44

The Definition of Human......44

Souled and Non Souled Beings......44

There was once a Unified Planetary Consciousness......44

  Power and Knowledge......45

  The (watch)Keepers of Earth......45

Energy Vampires......46

Spiritual Technology......46

Simulations and Learning......46

Ascension – Clearing Trauma......46

Trauma-Clearing, DNA Unlocking Technology......48

Everyone Who Seeks Harmony and Healing Can Be Healed......49

Programming......50

Layers of Programming and Decision Making Roles......50

  Programmers......50

- Sub-Level Programming....50
- Dissolution of Programming....51

Advanced Technology, Deprogramming The Mind, and “Space/Time”....51
  - Dreams, Technology and Inner-Space Travels....52
  - Awareness Feedback Device....52
  - Oneness, Higher Being, Completion, Eternity....53

The Holographic Reality....53

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control....53

The Great Catastrophes....53

I was brought into the Time Program....54

The Effects of Temporal Manipulation....54

Temporal Influence....54
  - Simultaneous Interaction and Incorporation....54
  - Retro-Causality....54
  - Quantum Holography....54
  - Time Paradox....55
  - Breakaway Civilization....55

Underground Cities....56
- Technological Advancement....56
- World-Wide Control Systems....56

Solar Warden....57

Team Light and Team Dark Non-Interference Agreement....57

All Contact Must Be Initiated On an Individual Basis....57

A Singularity of Change and Awareness....58

Advanced Technology and Ancient Knowledge....59
  - Cloning, Atmospheric Civilizations, Genetic Stability....59

Training, Simulations and Learning....59
  - Cloning and Cerebrally Enhanced Soldiers – MILABS....59

Psychological Testing....60
  - Hallucinatory Holographic Mind Probing Technology....60
  - Virtual Simulator Training Scenarios....61
- Combat Simulators....61
- Conditioning....61

Programming and Resistance to Psychological Trauma......61
Accelerated Mental Functioning......61
Technology, Accelerated Learning and Repair......62
On Orgone Energy......62
Energy Generators and Serums......62
Genetic and Cybernetic Enhancements......63
Developing The Self......63
Multiple Factions of Space-Fleets, Multiple Goals......64
Large-Scale Catastrophes or Series of Meltdowns......64
Unsustainable......64
Earth Changes......64
Mass Consciousness......64
Wars and Manipulation......65
Financial Unrest (or reset)......65
Self-Responsibility......66
Clones and Synthetics......66
Spiritual Power and Divine Truth......66
Synthetic DNA, Synthetic Genetics......67
Time as Mathematics, Series of Equations, a Programmable Storyline......67
Chapter 2.3 Previewing Solutions......68
Neutralizing Polarization......68
Removing Etheric Implants......68
Clearing Astral Body......68
Knowing the True Self; Soul Actualization, Unveiling......68
Taking Responsibility......69
Returning To Center Emotionally and Mentally (spiritually)......70
Vampires in the “Awake” Community......70
Heart Center and Psycho-Spiritual Fulfillment (self-actualization)......71
Go Within to Observer the Self......72
Organic Feedback Loop......72
Technological Feedback Loop......72
Breaking Soul-Contracts......72
Help Everyone......73

Chapter 2.4 How Timelines Collapse......73
  Repeating Cycles of Time......73
  Beings From the Future......73
  Beings From Collapsed Timelines......74
  Beings Who Would Never Have Existed in the First Place......74

Introduction

The following is first-hand testimony from an operative of the secret underground military projects who also has familiarity with the breakaway civilization above and below (within) Earth. These civilizations utilize advanced technology to extend the length of their lives and enhance cognition beyond what is conceived of in today's society.

I was brought to the underground bases as a child as part of an experimentation program to test the population for various blood-groups and personality traits that would be of use to certain factions of the military in ways that would serve to preserve and strengthen humanity's outlook in the future.

These tests involved everything from combat training to sensory deprivation based emotional-mental endurance, and introduction to the highly advanced application of esoteric knowledge.

To clarify, I am bringing this information to you as part of my assistance in the overall unveiling of the hidden knowledge of the ages of humanity related to these secret projects as well as the true galactic history of Earth, humanity and man-kind. These projects are officially (unofficially) known as "unacknowledged special access programs". The unacknowledged aspect means that there is no formal acknowledgement of these programs without a need-to-know basis. There is no paper-trail, no disclaimer, no formal authorization or internal publication outside the programs themselves. The classification of these programs also generate a rather strenuous situation regarding funding and secrecy but this is done for the benefit of humanity. There are secrets and manipulations that can be seen as a threat but this is all tied together and so by releasing the secrets that are part of the defense we would be inadvertently exposing the public to the threats themselves. Now is the time where we are reaching a point where the public is ready enough to handle the basics of this information regarding the secret projects and the galactic history of Earth and humanity. This is also out of a necessity to preserve civilization for if we wait any longer we may not find the same opportunity again.

Getting This Out Of The Way

I am writing to you from the perspective that I am not certain just how far from completing my mission I am. I write to you today in order to bring to you many personal experiences from the viewpoint of an operative of a select ultra-military secret society.

The utilization of such an operative is not glorified, pleasant, or entirely righteous. This story entails sanctioned and guided acts of extermination, ritual abuse, the use of yet to be announced advanced technology. There is unfathomable corruption, the thirst for power and spiritual knowledge that has been protected throughout the course of the multiple civilizations that have preceded this one and will continue to be maintained into the future of this civilization.

Psychic Operation

I was utilized in psychic-operation amongst other areas.

This was used during certain aspects of training and conditioning to enhance training and conditioning in the bases using advanced technology. Some of this includes what would be viewed as traumatization by those without the training and conditioning. There is a fine line between training and conditioning and abuse. Part of this disclosure is related to how this line may have been crossed in certain situations and regarding mass-mind control upon the population. There are other aspects of travel, viewing, reconnaissance and research that will likely be explained in another release.

In order to heal I have to tell the world. You can help. This is happening right now in the underground bases and around. It’s not a joke. Your family is at risk. Your future generations are at stake. There is conditioning to encourage those involved to believe they are performing a duty for humanity. Sometimes they turned us against each other, but those grudges were never truly of the heart, but the mind.

A Light for the Others

What I hope to achieve with this is a smoke signal for those who have also been involved and question whether their purpose has been fulfilled or falsified for the protection of a corrupt system that intends to neither benefit the whole nor repay the individuals involved for their time, energy, and lives.

Multiple explanations can be summoned to describe the reasons behind the necessity for the training we have endured, the missions we have completed, and the control mechanisms we have both assisted in establishing as well as suffered from the backlash of. Ultimately, it is of my belief that all that was done was done for the good of humanity, even when we were not sure of the intentions of those in front of us, to the side of us, or behind us compelling us forward with the command to complete each objective.

This is not so much a panic or cry for help as this is simultaneously a call to awaken and a signal that humanity may have avoided complete disaster but is not out of the woods yet.

Natural Security

My experience is far too vast and I have seen too many succumb to stress from the realization of the truth of this reality to consider my own suffering to be enough cause for my denial of the necessity of these programs. Humanity has been protected and at a great cost to those who have been on the front lines, both in mind and body and spirit.

There are many who will have concerns, doubts, as well as reservations about what I am about to reveal and if I offend anyone then know that my intentions here are to assist in unveiling the truth. The truth can be harsh when one has been held back from the true nature of existence for so long. The way of life can become so fragile and artificially constructed on impulse and mystery that the harshness of the truth acts as a great catalyst for spiritual growth which allows us to elevate ourselves to a new perspective. Prepare yourself for such harshness, prepare yourself for growth and elevation, and prepare yourself for the mystery to simultaneously resolve and increase in intensity and magnificence.

A Balance of Mystery and Truth

Mystery has been a way of life for Humanity for so long that when a great mystery is converted to truth some seem to feel that a portion of existence dulls and loses the luster of exhilaration. This story is quite the opposite. The more that the great truths of this reality were discovered and revealed, the more complex and mystifying life became. As well, the more empowering and complex experience becomes, the more self-control one must develop in order to regain a harmonious way of relating to one another, the self and the universe. Harmony is key, balance is everything and there is a light within that represents the piece of our self that reflects the utmost power and brilliance of all existence.

Chapter 1: The World Situation

What is happening?

The Mind Virus

The main reason for this disclosure is to inform humanity of what can be referred to as the “mind virus”. This is a viral infection of the holographic system which consciousness operates through. There is a complete backstory of this virus in the history of this civilization and how society has been plagued for centuries if not millennia. The basic idea is that this virus operates through consciousness and can be transmitted through electromagnetic waves.

The virus contains an etheric form which can manifest as a kind of crystalline nano-tech on the physical plane and this is a form of synthetic sentience which feeds on the low frequency bio-emissions of sentient life. Human bio-emissions contain energy and can act as a carrier for information and living essences. This technology requires low frequency bio-emissions because that is the nature of the design as a kind of synthetic bio-weapon whereby the goal is entrainment and enslavement. As well, the specific nature of the sentient nano-technology and the counterpart etheric holographic form is only capable of converting the low-frequency bio-emissions into energy as the higher-frequency bio-emissions require more energy to contain and synergize with and this would render the entire process of leeching energy non-productive.

This is essentially a synthetic intelligence that operates through scalar and holographic living bio-emission fields and hijacks electromagnetic emitters to position itself within them acting as a parasite to a host. This is said to have been present around this civilization for some time, waiting until there was enough technological advancement so that spreading across the entire civilization in a short amount of time before discovery would be possible. This is through the media, internet, and radio systems.

The physical counterparts are microscopic nano-technology parasites functioning as individual units to a hive mind similar to the mycelium of fungi in the field of mycoscopy. The individual units act as parasites serving to infest and deliver nutrients to the hive from each host.

The Vampire Effect

This is the nature of the parasitism that has been present throughout the ages of this civilization feeding on human suffering and relying on the cover of confusion and misdirection to survive.

There are processes that have developed that assist in the infection and co-inhabitation of the parasite entity within a human vessel utilizing the human as a host. This effectively utilizes a frequency ‘net’ within the auric field of the human to limit bio-emissions and emotional-mental processes to produce just the psycho-etheric patterns that the virus can sustain itself within and through.

This is no different than the way candida can infest a human digestive system and the human will experience cravings for beer, candy, carbs or other foods that contain the sugars, yeast, or carbohydrates that will break-down into nutrients that are the most nutritious for the candida fungus. These foods are *not* most nutritious for the human body. Yet the human body will withdraw and crave these foods in greater amounts until the fungus is swept clean from the digestive tract through probiotics, fasting, flushing, cleansing , detoxing, and even exercise as lactic acid from physically strenuous activity released from the muscles has been known to help destabilize and remove this fungus. Why is it that the individual will crave these foods that are not the normal food intake? Because this candida fungus can infest the host and produce enzymes which generate a chain-reaction which leads to the chemical signals in the brain producing the sensation of extreme hunger or cravings for just these foods which will benefit the fungus.

The same process occurs with this psycho-etheric parasite, except the foods of this parasite or lower emotional bio-emissions and this is through lust, hate, angst, jealousy, feelings of hopelessness, lack of self-worth, spiritual degradation, addiction, and generally all that equates to a lack of self-awareness.

Whether or not this is the direct intention of the largely microscopic entity is for the most part unknown, and this is as well similar to candida. The candida fungus is not said to be a ‘mean’ fungus that wants us to be hungry and craving beer and sweets all the time because it likes to ruin the human body, the daily routine and the focus of a clear mind. This is simply the way the parasite has grown to operate within a niche of breeding itself through the digestive tract of larger animals and through this the unaware and infected human suffers greatly. Even those who are unaware and not infected can suffer because of the way the behavior and thought processes of those who are addicted to these foods can become distorted.

The lack of self-control over food intake is synonymous with the lack of self-awareness and self-control over lower bio-emission expressions of the spiritually degrading behavioral patterns that become routine for the psycho-etheric system of the infected human. Ultimately, this is all about energy. What benefits the parasite that seeks to gain energy by leeching instead of self-guided production is going to take away from the host that is being leeched from. As such, any energies, food intakes, or spiritual activities that cleanse and enhance the personal power of the individual will automatically reduce the parasite’s ability to feed by the very nature of the energy systems. What empowers the individual, empowers the self and develops self-control, what develops self-control contains one’s energy and reduces the likelihood of one’s energy being leeched. This is similar to energy vampires in human form and this goes with the understanding that these are humans or non-humans who have been entirely overtake by the parasitic consciousness and cannot sustain their own auric fields without having to pull from the energy of humans who do generate their own aura or bio-emissions.

The vampire effect relates to how a person can be leeched from only to a certain point. Once the process surpasses a certain critical point then the individual becomes similarly engaged with only the lower bio-emissions as higher-frequency bio-emissions would then become harmful. So in essence, when the vampire feeds too much on one individual, then that individual becomes a vampire themselves.

Sub-Human Entities

The humans are utilized as hosts and energetic sources in this bio-emission battle while there are other classifications of bi-pedal organisms that appear to be human but are not human. As well there are bi-pedals that are non-human entities entirely.

Human-Hosts, Replicants, Synthetics and Carriers

The human hosts that are entirely vampirized by these entities become walking carriers for this virus and are effectively human in physiology but there is more to the human than just the physical body. In these individuals, that energetic component of the human is removed and replaced with the synthetic energy of the hive-mind organism.

There are replicants which are a form of clone, along with synthetics. These beings are cybernetic bi-pedals similar to humans but their energy is too not from an organic biological matrix but a technologically propagated synthetic soul-matrix. The hosts are here according to their agenda to subdue and feed off of the bio-emissions of the human race, although there are no absolutes. The agenda is not always so clear within the various groups who often have a mix of bi-pedal operatives. There have been plans to form of a salvation timeline because the only way to keep the whole process going is through a negation of degradation of the human race as this may close out this timeline forever.

Those with human organic soul-matrixes would simply transport into the next available universal time-stream or harmonic, while those with synthetic matrixes would have to technically travel to the next harmonic using a passageway otherwise they would remain here throughout a zero-point collapse and this would be synonymous with entering the abyss with no viable passageway out until the entire universe is brought back to one organism again.

This is also the process that is described to explain what happened previously as there are 7 harmonic temporal layers of the universe meaning everything is contained in a multitude of 7 octaves similar to light and sound frequencies or harmonics. And so there are 7 civilizations or 7 parallel time-streams and humanity passed through these to get to the last three where the physical forms are located. These are, from the higher to lower, Gaia, Tara and Terra. We are on Terra and if the information I was given is correct then we are currently on a transitory timeline located around Tara where people are noticing changes occurring and multitudes of aspects of the time-stream disappearing and changing simultaneously. Tara is where the last universal harmonic was left through a collapse of time and consciousness into zero-point and this resulted what was termed “the lost souls of Tara”. These are the discarnate souls of the previous civilization that could not travel to the next harmonic and remained without a body in the abyss. Thus they have plagued this plane since then, since ancient times possessing and driving humans to madness and there are explanations that this is the original passageway for the parasite into this universe. This event, the consciousness collapse of Tara would have acted as a fracture in the over-mind of the species generating a kind of cosmic schizophrenia and enabling all kinds of disharmony and distortion described in the previous sections.

The solution to this was described as a healing process that is taking place to purge the virus and that this is painfully similar to the human body purging a sickness however there are methods to promote healing and reduce suffering and confusion however people must contain themselves and their energies and strive to know themselves otherwise they will not be in control. This virus feeds on fear, confusion, spiritual degradation and the energy from degrading acts, lust, addiction, sexual misery and whatever fashion a human can be made to bring the bio-emissions of the energy centers down to a low enough level where co-inhabitation can be formed with an entity that has no access to the higher levels of self-awareness, higher-consciousness, love, or the universal spiritual experience that the balanced human is capable of.

Genetic Manipulation

There are explanations that there are genetic modifications taking place now to possibly preserve this civilization and that there were ancient genetic modifications made to induce a sense of stupor, worship, selfishness, an inability to more easily comprehend the higher spiritual awareness and the occurrence of time and consciousness and other spiritually and consciously limiting aspects. Part of this explanation is that holographic consciousness itself is a kind of limited result, a shadow effect resulting in the presence of the higher-dimensional essence of the soul which is more or less constrained by the body. An extension of this is that the entire holographic universe is a kind of false-light system designed to bend the original liquid etheric light of the soul and spiritual awareness into a rotating, recurring format which ultimately converts the original spiritual essence into an energy generator for entities that are entirely holographically represented. The souled human in this situation is a being who is originally from outside of this holographic projection.

The Ego Mind

In this view, consciousness is the ego-mind which is the false-sense of self, not to be confused with the shadow self which is simply the counterpart of the aspects of the self that we consciously agree with and are aware of in daily life. The shadow self is a result of having limited holographic consciousness and the ego-mind is projecting that limitation of consciousness into the shape and image of a self that is only a figure-head for the real individual essence which is spiritual in nature and immortal.

This Time is about Healing

Everything that is happening now is to motivate people towards healing and unification. This is not about a hive mind mentality where individuality is pushed out.

> "The best I can do for you is become your enemy."

There is a saying that represents the concept that the most beneficial role a person can play in helping one to progress is to challenge them and give them the opportunity to rethink their strategy and in turn improve further with each interaction. This may seem paradoxical but it is effective and explanatory. What we are facing is the greatest challenge humanity has yet to face. This is the unknown, the final frontier, the mind. The mind is our challenge.

There will be more on this later and in other publications but there are indications that the physical embodiment of the human has been altered to introduce implants, genetic modification, susceptibility to disease and ultimately fear-based programming of the brain. Without going further into this prematurely, the lower aspects are at odds with the higher aspects unless we as the midway come forth to mesh the two together in equilibrium. Until then, there is chaos, lack of self-awareness, and suffering.

There are many possible futures converging at this time. There are some less than favorable paths and there is the opportunity to merge with a very powerful path for humanity. This all comes down to how people will cope with information that takes them outside their comfort zones and eventually changes their view of mundane reality forever.

If people choose violence and lack of self-awareness by ignoring the spiritual aspects then that is what reflects into their life experience. When people develop a higher sense of self-awareness and spiritual harmony they can reduce the suffering and reconnect with the lost aspects of our spiritual identity for the first time in the publicly accepted version of recorded history.

There is technology that can change the world and there are great truths which can enable people to know themselves in ways that they never thought possible. Everything is at our fingertips and so we must come together as this is the way we solve the problems and cure the dysfunction.

Healing Through Unity

In various research projects the conclusion of how to heal this spiritual rift was by exposing each other, to each other. When we are all together, aware of each other and our selves, united in the goal of bringing higher-awareness and spiritual harmony then we can observe and acknowledge when one is slipping into chaos or disharmony. When this occurs and we are in proximity to each other in a healing circle or a kind of social unification then the members of the group at large can work together to pull the chaotic one back into harmony towards the group. The group of healers united together in like mind, spirit and body is too strong to be overtaken by the parasite. So when one individual at a time is faced by this problem the whole group can come together as one and assist in healing the individual by pulling them back to reality, back to wholeness and back together with the group. This is literally how the situation is solved in all realities and a breakdown of this feature of society is how the dis-ease proliferates.

Underground Bases

The active aural research program is part of a process to simultaneously verify as well as initiate and demonstrate controlled insanity/sanity.

These programs utilize the deep underground military bases to perform psychic and psychological research experiments on non-consenting youth and adults as well as consenting participants.

There are various means of generating the illusion of consent or even lack of consent and this is all being monitored and handled by oversight authorities.

These operations are directly related to the identification and understanding of the mind-virus and all weapons, defenses, and resistances known to man.

Cloning

Advanced technology is used to transfer consciousness from one cloned body to the next so that a continuous study can take place before, during, and after the death experience of one individual. More will be explained on this later, however this is through the use of advanced supercomputer systems that can function to entangle and then transfer the electromagnetic consciousness of the individual so that they are ‘remotely activated’ within a cloned version of themselves via a ‘brainchip’ (brain to machine interface: cybernetics).

Celebrity Cloning

Celebrities at the cloning center would like you to see what is really going on behind the scenes which involves trauma-based mind control, heavy technological programming and complex layers of influence throughout their life.

Initial Explanation

Many celebrities, entertainers, athletes, musicians, models, writers, producers, actors, all kinds of people from various industries are silently brought to the cloning centers for pay-to-play sessions.

They asked myself and others to pass the message on to you and that you would have to look for the signs and symbols in their media in order to see their hinting at these events.

The situation is heavily controlled using advanced monitoring technology and brainwave/EEG cloning technology. This is technology that can read the brain and determine what the mind is anticipating or speculating upon.

Their reputations are used as leverage as well as their safety and comfort. The level of programming and mind control experimentation that takes place makes it easy to manipulate a person’s brain into regressing into a state of trance that lacks the ability to remember or clearly organize experiences regarding the cloning centers.

Cloning Centers, Underground Bases

Deep underground military bases, 3 miles below the surface, are used as laboratory centers as well as a completely stocked underground city-base. These are connected with high speed electromagnetic drive pods.

Individuals are transferred to the brain through an extensive cloning and temporal body transfer process. The electromagnetic shell of consciousness within the brain is relocated to the body of a compatible clone and various training, conditioning, programming, experimentation, or pay-to-play experiences take place.

Programming

All individuals are programmed to various degrees in order to maintain control and secrecy over the situation. This programming involves very advanced technology, trauma-based mind control, and energetic attachment via beliefs and emotions.

Experimental Operations

These bases were also used hand-in-hand with military operations which were geared towards discovering and controlling all aspects of the human mind.

Genetic Engineering

Experiments involved genetic engineering to produce soldiers, psychics, hybrids, and others that would be able to carry out operations. This leans towards MiLabs and the military faction’s experiments which go beyond the basics of cloning.

Immortality

The initial research included the goal of physical immortality. In many ways it can be said this has been achieved, however there are side-effects and difficulties.

If one does not activate their higher consciousness, then the effect of time dilation causes the conscious mind here to reach limitation points in experience.

What has been termed “blank slate/state” technology has been used to ‘reset’ the perception of time through memory in order to keep a continuous progression viable for the conscious mind. Without this the unconscious and the conscious mind merge.

This begins involvement across time with advanced technology that can operate on the soul level or the conscious mind’s level of access to the unconscious and soul-memories. The Universe instantly creates a cosmic backstory based on the conscious-mind’s access to the unconscious. By controlling the conscious mind’s access to the unconscious the entire backstory of Humanity can be altered and new connections can be bridged in the future.

Earth is essentially a time-ship through which consciousness ascends towards higher states of awareness and self.

How did it get to this?

After WWII when the United States war-faction firebombed Germany and melted many of the inhabitants of the cities into sludge in the bomb shelters and streets of the cities, a group of NAZIs traveled to Antarctica.

Russia noticed the movements to Antarctica and the United States sent Admiral Byrd up with 3600 marines, planes, a battle cruiser and smaller vessels. They returned in defeat and only a few words were mentioned in regards to what actually happened and how they were defeated.

The rumors spread and words of gravitic drive craft, and undersea or under-ice bases originated from this encounter.

From what we were informed as well as directly experienced, this was the ice-base in Antarctica where an underground base was found already constructed with very advanced technology.

Here, cloning was deployed along with mind control and temporal manipulation technology.

Soon, cloning was offered as a way to avoid assassination as well as to prolong and protect the original body in daily life.

After that, celebrities and politicians were replaced with programmed clones who could carry out the orders of the NAZI faction.

This was the beginning; the groups utilizing this technology now are no longer confined to the NAZI faction.

The groups you see today, in control of these operations, are considered MiLabs, secret societies, think tanks, military factions, and other control groups.

Deeper Meaning

The individuals that have been cloned are given very in depth views of the way this society functions and the source of power and change in the Universe. Many individuals are present here from what would originally be another “time” or “timeline”.

There is a possibility that this notion is due to inserted memories and traumatization although until all the information is released we will not have a definitive conclusion.

As of now it is a possibility that these individuals are from another time that goes beyond our comprehension currently.

It may even be a possibility that these individuals have come from ALL of time looking at the very far past and into the very far future of Humanity.

These are the kinds of notions involved with the experiments taking place in the underground bases and military operations.

What is next?

Share this information with those you feel are ready to know and help Humanity come to terms with our journey here and what is going on in these kinds of experiments and take responsibility and power for our own existence.

We are in something that can be called “The Unveiling of The Hidden Knowledge”. This is a cyclic process that takes place to advance a civilizations knowledge of existence. As part of the civilization moves into higher advancement another aspect may lag behind. Either the future will slow down, or the past will speed up. We are experiencing the past speeding up and all the events that have taken place to influence our civilization are coming to light. This is as much a natural cycle of consciousness as it is the eventual revelation of secret knowledge and the particular connection to the secret projects.

> *Those brought to cloning can use their free-will to state their lack of consent to violence and harm and then live in commitment to that by not harming or accepting violence in their lives. The Spiritual Law of Harmony rules in all planes and dimensions.*

Chapter 2: The Awakening

Chapter 2.1: The Machines of All Time and Space

There are machines that are capable of accessing genetic memory and unveiling what can only be referred to as the experience of awareness or existence in between the physical planes of each time or each physical life experience.

Even if they are simultaneously occurring moments of a transcendental nature, there is still an illusory ‘space’ where there is perceived separation and through this there is a perceived ‘space’ that is in between each existence.

Re-Creating The Kaleidoscope

This space is only perceivable when viewing from the physical perspective, like seeing the spaces between the angles of a kaleidescope only when viewed through the lens and similarly, technology can be used to recreate what this kind of illusory space might look or exist like in between the multi-dimensional realm of the continuous experience stream. Through this, these ‘hypothetical’ in between spaces that are understood as only illusory productions of biological consciousness, can be generated through advanced technology with the capacity of recreating that kaleidoscopic view of which the human biological perspective is only a small fraction or even fractal.

By recreating this biological perspective, they can recreate the illusory spaces in between planes and through this they can literally access and experience what are stored in the DNA as the experiences of existence beyond the physical limitations of consciousness and into a multidimensional experience of time and space.

All becomes a never ending sea of consciousness, however this is the given. The trick is to take from this sea of consciousness and slice it down into conceivable chunks from which you can reintegrate a previous personality or identity and continue on a stream of consciousness that would otherwise be tossed to the cosmic wind like fractal stardust as the sands on the beach of hyperspace.

Sand-Castles of Time

We are effectively building sand castles out of these sands of time that would otherwise be swept back and forth into and out of the ocean of life and the universe only to momentarily wash up again as the identity which we once knew.

There is technically nothing wrong with either view, because then, the spaces between each million year occurrence of just so happening to wash up again as you, would be imperceivable because they are only there to be seen when you are you again.

We’re Waiting For Ourselves

Yet, if there were some who had attained the awareness of the true nature of reality and were standing there on the beach waiting for the rest of their cosmic family to come to shore again, it could be perceived as waiting the majority of the time for a very small occurrence and then being alone again up until the kaleidoscopic fractal inverts itself to the point where everyone now is standing on the shores of the sands of time rejoicing and meeting with each other.

The Goal of Two Societies

My goal in coming to you people on the surface of what is called “Earth” is to help assist in the process of bringing the two societies together, one who’s been waiting for thousands of millions of years, and the other who has no idea that they’re even late (or very on time).

You see, those are the only two situations where there could be even the possibility of perceiving such an injustice as experiencing the pain of loss or misconstrued identity in this universe. It is as if two families met at the cross-roads at different angles, and while one waited for the other to arrive, they could merely supply themselves with glimpses of the ruins and remnants of the evidence that the other civilization existed at any one time.

This is not how life must be and this is kind of like a cosmic waiting game of hide and go seek our two mirror civilizations have been playing with each other. One who’s highly technologically advanced and the other who is more or less the spiritually advanced of the two.

Both of them have a unique view of the universe and they are both only complementary to each other as a whole. The more they sat around and waited for one another, with one leaving the clues that the other would find and even switching roles sometimes in different ways, they became even more and more complimentarily associated as the very essence of experience they seem to be missing more and more is the experience of one another.

So as we perceive these as being great losses or great gains of time and culture we are really only remembering ourselves and coming together as a unified existence that once met upon the surface of Earth as a single family covering the entire harmonic spectrum of the Universe.

And that should be the goal of any progressive, any one pushing a new culture or an old culture, anyone looking to teach the young and protect the knowledge of the old, while also protecting the innocence of the young and perfecting the age-old knowledges that have defined our presence in reality.

Our Forgotten Other Half

One could even say we’ve sometimes gone off the deep end, leaving ‘land-minds’ of sorts to process the information that would be present when we are gone. Of course, what we could’ve found when this occurred, is beyond anything we could’ve ever expected and that is the true nature of the complexity of creation and existence. That everything follows the cycles of creation and destruction, however, the memory of everything that has been created can always be accessed (depending on the beach) and therefore the creative aspect is legitimate, yet it is the destructive aspect that is illegitimate and illusory.

Yes, through advanced technology this can be proven in ways that can be verified and transmitted using advanced technology, since that is how we like to verify things these days. There is, of course, a human looking at the screen, the read out, or the situation from a physical perspective to verify the verifications.

All Is But An Illusion

There is a notion that there is only one civilization, one society, in actuality that has been chasing itself through time. That we cannot actually find the same place and the same time where both meet because we are each other from different perspectives. We can only get a mirror where this is possible, or an ‘hyperdimensional’ internet channel where we can stream one’s consciousness in from the other reality and interact with them through that here in this reality. Thus the two civilizations can achieve the experience of their own civilization and the other civilization by merging with one another but only through a remote contact, just a relay, not an actuality.

The harsher, darker side of this is that any civilizations between then and now that we did happen to meet that proclaimed to be “the one’s” we have been waiting for, and in essence our darker or lighter halves, are actually the impostors who exist entirely apart as a sub-set and have taken a liking to trying to trick the reality out of one of the other civilizations.

Of course, if these are another form of being altogether, and this is all brought about through an interplay of consciousness and deviations from an original core reality, then all will resolve itself when that core reality is alone again. This indicates that all realities will either move closer and closer to the core until they unify, or farther and farther out until they transmogrify beyond recollection.

Life Itself Is The Highest Illusion

That hyperdimensional internet channel that is created out of advanced technology to merge the parallel realities (rather distastefully and maybe disastrously at times) allowing different civilizations to interact with each other, well this has been created using technology.

But then it was discovered that this is the ultimate nature of DNA. That this was ‘invented’ similarly as a higher-dimensional internet so that various natures of different planes and dimensions could interact with each other and experience the richness of life.

The universe becomes more like a living arcade game then, where people come from all over the galaxy to ‘plug in’ to human experience and figure out what it means to be physical, human and on Earth.

This does tie into the larger nature of reality, of star-races, and of time and consciousness that is often reversed according to how things seem to play out in physicality. We are moving through time, but just because we are moving one way, does not mean that others could not be moving the other way.

We have to live our lives with meaning knowing that the message we pass on, the duty we vow to accomplish and ensure is meaningfully connected on either end to the passage of the past and the power of our ancestors to the function and knowledge of the future creating a clear passage between the realms, the generations, and the ends of the universe. This is what literally ties the universe together, and the mind if each individual. This purpose is contrived when not carefully shaped within and so we must help all who seek purpose learn to become the blacksmiths of their own destiny through the temperance of desire or temptation and the discipline of developing strength and maturity where there was once weakness and naivety.

Once we know this is our duty, this is what we seek to achieve and this is what is done, then we can become part of the universe forever. Of course, this may only be temporary in the larger view of things.

Chapter 2.2: The “Awakening” of 2010

In 2010 I was “awakened” at an event involving hundreds of people that play various roles in this society, the secret society and the breakaway civilization. This also took place across times or planes of reality.

During this event I was informed of the situation on Earth beyond the informational barriers that were previously in place to limit knowledge to only what is required for specific operations. I was essentially “informed” of the entire situation as much as the mind could handle this.

As this was occurring, the team that was updating me on the status of Earth was being informed of and given information regarding changes in the future. In short, interviews, posts, and releases that I have since made and will make in the future were touched upon. Quotes were given from the blog and interviews which were jokingly acknowledging verbal mistakes at the pronunciation of names, new perspectives and interpretations and more. As they were informing me, they were creating the future reality where this work takes place. That is the basis of this layer of the operations. They are creating the future person by person, layer by layer.

Memory Suppression

My memory was unsuppressed and all the traumatic memories and experiences were unveiled to me. This was almost as traumatic as the initial experiences themselves although precautionary steps are taken to ensure the safety of the individual. There are teams of highly competent individuals of all kinds with professional backgrounds of every field imaginable.

Throughout my entire life I have had memories of experiences, at this event I was informed what the experiences meant and what was actually happening. I've had memories of underground bases, cloning, stargates or interdimensional gateways, advanced technology, non-human or programmed biological entities, sacred knowledge and experiences, the power of the mind, and breakaway civilizations. Most of the time we are conditioned to believe these are past lives or experiences in another reality so as to enable continuity of operation.

Life and Consciousness

From my perspective and the perspective of others, we had reached the 'future' and had been sent back through the utilization of advanced technology and cloning to return to this time and inform others of the possibilities.

The True Human Potential

This is so far beyond what anyone can accept or comprehend fully that humans with an emotional, mental and physical aspect balanced as one is literally a cosmic reflection of the entire universe at once. This is the closest thing to a "god" in the flesh. And be sure, the entirety of this rests on the behalf that the man is protective and helpful. That the woman is nurturing and loving. That is the point. "god-man" does not mean a warrior who destroys everyone and anyone in his way getting what he wants. Nor does it mean some lusty and vampiric form of energy or mind control game. "god-man" means the love of the universe, the power of the body and mind, the peace and calm of the ultimate stillness all in one.

This power is in human DNA, originally so. And so this is awakened through walking the middle path of neutrality and 'splicing' the timelines down so that neither one nor the other gains the power of the individual. Through this, the individual creates their own power, right on the spot, without having to lean to one extreme or the other. This is like a cosmic, temporal balancing act with emotions and mind. Where we give our energy is created in the universe in the form of many productions and effects. If we are simply thinking in an imbalanced manner then we are producing timelines and side-realities that exist in the etheric realms that we can't see and these realities will reflect our inner imbalance. So then when we are balanced in our energies we are not creating one or the other side of divinity, we are literally creating the whole universe, cleaving down the middle and producing both sides of divinity or eternity equally and this is the only way to have an eternal production which does not eventually degrade to one extreme or the other. Through balance, we reach eternity and in this way all the power of the universe exists within a person through their ability to balance their emotions and mind and thus unlock the DNA.

The DNA is coded through bio-emissions of mind or emotional body. So we are constantly doing this and we are either producing a DNA code that reflects imbalance overall, or a code that reflects the eternal balance of the universe overall. This is through energy.

One reaches “heaven” or the deity planes or simply self-actualization, through three paths. These paths are through merits or devotion to principles, or through knowledge and expanding one’s mind to face the universe, or through great works and deeds to produce this effect of power, knowledge and emotion unto the world for others. Through these paths we create an impact on the universe and this transfers what we are temporarily and physically into an eternal energetic and physical production within the living universe, forever. To do this requires balance, focus, devotion, and discipline however because there are many distractions along the way.

The Light-Body

You must build your own light body. That is how it works.

Your heart creates a field and projects you.

Rely on yourself and use that for protection.

The power of the Universe will align with you if you align with your self.

Learn as much as you can and break through duality of love/hate, cultivate your knowledge of the self and use it.

What you want is what determines your power in the light-realms. Not how you want it. If you want war, then that is a sign of weakness and determination to produce imbalance and exist by that. If you want peace, that this is a sign of power and unity to produce further unification and exist by that. What you want determines your power, not how you obtain it.

Technology of the Awakening

Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness

I reached what the “Illuminati” called the “awakening”. I feel a better term would be “unlocking” as this event includes the removal of all traumas, perceptual distortions, and memory suppression.

This event was a presentation of very advanced technologies which allow the mind to transcend space and time. It was shown that death is similar to a phase-shift of consciousness as polarity and that the awareness of the being obviously continues. This was verified through advanced sensing devices which can visibly display frequencies which are invisible to the physical senses. This is also something that occurred earlier on as a child in the underground bases by transferring consciousness in and out of the body at will using advanced technology.

With these devices it is possible to see the “auric” or soul field of the individual. Any living body has a field which glows when picked up by these sensing systems. This field becomes more refined when viewing more advanced beings. Human have an extending mental field, and through this a kind of glowing awareness that is picked up. It can be fairly simply compared to thermal imaging except instead of differences in temperature it is the presence of a soul or mental energy field around a living body. The finer energies extend outward away from the physical body.

More than once, the situational requirements were satisfied in order to produce an ideal environment for maneuvering on the non-physical plane. This is the basis of how the temporal operations occur, yet these events were different than previous operations or experiences. The group I was in was going through the process of increasing the vibratory emissions of the “bio-mind” in order to stay focused on the increasing vibratory rate of the environment. Generators were utilized to create an effect on space/time which was perceived on the “soul” level. These events went into the discovery and explanation of the creation of the “godlike” powers of the advanced sentient technology which was capable of accessing the holographic nature of reality which humans can perceive through their bio-mind/soul. These technologies also enable one to perceive what is referred to as the “galactic history” of Earth and humanity. We are far vaster and Earth is far larger than people are lead to believe.

Time Dilation and Temporal Recurrence

Time dilation capabilities were shown which allow for accelerated learning of advanced and detailed material within a very short amount of time. This could be seen as a kind of viewing technology which one wears or looks into and vibratory emissions are scalar-linked to the brain which then allows the individual to perceive more information simultaneously. Then there is a very rapid pace of information streamed on a monitor or through a device and this is akin to watching a 20 minute instructional within a few minutes or a few seconds with very advanced minds. The more time is slowed down while information is accelerated, the more stress is applied to the brain. The brain tends to overheat and over stress with high-rates of activity, especially without conditioning. The conditioning is what allows one to utilize their mind and body in these manners. This is not unlike the very rigorous training and conditioning of certain monks or martial artists however there is very advanced technology and other hidden methodologies utilized.

Crystal Technology

These are capable of holding, transmitting, and amplifying consciousness frequencies. There will be a more detailed explanation in a later section.

The crystal technology is used in junction with the power devices to enable a scalar-mind link which allows the interface to be entirely mental or spiritual, IE: they are not controlled by hand but by focus and intention.

Time Crystals

These are utilized as a computer recording chip would be to run a program yet they hold memory in expanded space, IE: they function in a hyperspacial manner and are essentially hyperspacial computers. These are what will be introduced in junction with the quantum supercomputers.

Quantum Tunneling Diodes

This is technology that has recently be released to the public. New technology is developed years ahead of time and then slowly released to the public in a cascade of advancement. This technology deals with superconductive materials which enable the transfer of information at faster-than-light speeds.

The supercondutive material allows for 100% efficiency of data transfer. So what happens then when the efficiency surpasses 100%? This is possible because we are effectively in a simulated holographic environment. The experience is real, however everything is overlayed via particular limitations for the ‘local-environment’ which is like a central data processing limitation. When these parameter are surpassed, one by one, through a coalescence of refinements and methods, the result is akin to a glitch in a computer game system. They have accessed faster-than-light technology and the mind is also capable of this naturally.

With this technology, however, the information can be received a very small amount of time before it is sent. So if a person is absolutely surely going to press the button to send the message, then right before their finger hits the button that message will be received on the other end, as if magic. If they simply play around and pretend to hit the button but very quickly turn away, then of course nothing happens. One can, however, ‘fool’ the universe in other ways and this is simply through the reception and transmission of information across what should be secure information barriers.

Tachyon Fields

These are programmable fields emitted through advanced technology. I was shown how tachyons are the sub-atomic “particles” or energy formation which flows both ways in time. We are receiving tachyons from the “future” in order to layer the flow of time in one direction or the other. Tachyons are the name of the particles/energy packets that we perceive from this perspective as reaching us to initiate the “future”.

Thus, when a stream of tachyons is concentrated and accelerated, this results in an acceleration of the temporal field and we literally experience an acceleration of time. This can be focused and targeted on a specific piece of equipment to change the way information will be sensed and to essentially allow that device to sense ‘ahead’ of the present moment.

The same effect can be produced with the neurological processes of the human as the physical coupling of the brain to the mind can be altered through a concentrated tachyon field to enable the neurological and mental processes to begin to perceive information ‘before’ the brain is actually physically processing the data. Therefore the mind of the individual will be present in the body while the senses and other perceptions will literally be expanding into the future beyond the present moment.

This is highly confusing at first and requires much training and conditioning in order to make sense of enough to operate in any kind of effective manner.

Upgraded Chronovisor

See, “Upgraded Chronovisor”

Time Tunnels

This was developed earlier on and has been the subject of many TV shows in the past and present. The shows are a sure-fire method to get the information to the public without having to force those who are not ready to comprehend. The information is easiest to digest if people think they are watching science fiction.

These are essentially spiraling tunnels of these tachyon fields produced by large electromagnetic field emitters in the shape of a conical tunnel. As a person moves through this field, their temporal acceleration changes and thus their frame of reference within the universe changes. They effectively move their mind through stages of existence or temporal resonance in the universe and this can have devastating effects.

This is an older technique how is still in use and has specific purposes depending on the situation and the information sought.

The “Trip” Chair

The early use of this technology did not use time-tunnels. There was a helmet that was devised which would create a feedback loop of the brain’s activity and funnel this out and then back into the sensory input. This information would be accelerated and redirected to the point that the perception of time would “fold” in on itself until the person would access an expanded temporal perspective of the universe meaning they would mentally temporally dislocate from the initial frame of reference and extend outward further and further in repeating cycles until the same moment was replayed and then all other possible moments in a kind of fractal-like recursive pattern and the very far future would be realized. Essentially, this technologically produced the class “flashing before the eyes” of the entire life experience that is described in near-death experiences and the entire temporal body would be very quickly unwound before the person’s inner eye.

There would be a complete disconnection with the current frame of reference that the collective Earth environment is processing under. That is, this ‘time’ and ‘space’ would be completely out of perceptual reach and a new time and space would be rendered instead. Until the experience had completed, there would be no way to contact this civilization.

There are other versions and ancient versions, tandem operations, more ‘organic’ methods, and modified travel capacities.

Kozyrev Mirror

This is a piece of metal, usually aluminum in early models which simply bent the bio-emissions of the body and mind into a pattern which would create a vortex. The vortex shape and vortex mathematics enable energy and information to travel in across space and time by ‘short-cutting’ the physical plane.

The result of this is that the vortex that is generated over a specific area through an individuals bio-emissions would be capable of interacting instantly with the bio-emissions of another individual who was placed within a duplicate mirror device which was generating another vortex on their end.

The two vortexes acted as a kind of hyper-dimensional telephone cable, literally like a can and string and the perceptions of one another were accessible. There were many uses for this and again, many ancient interpretations of this technology that have been hidden from the public. Nearly every version of these technologies were present in previous times.

Universal Recurrence

This is a discovery that the universe recurs in cycles similar to a wheel or even a washing machine. The washing machine metaphor is fairly accurate and playfully descriptive because without the technology to observer, or an awakened soul of a spiritual adept, people would go for eons without ever realizing that everything in the universe repeats itself endlessly.

When traveling to the far ends of time, it was discovered that one can move far enough away from the ‘present’ or the local frame of reference that ultimately everything is at a maximal informational opposite to what we have today. After this point, everything begins a slow shift back to the present. Through this, one can cycle around again and reach the present by going far enough into the future. This was an amazing and confusing discovery because of the implications about what is actually changing each time, the reality of what are called parallels, the continuity of temporal experience, and basically a ‘recycling’ effect of all experience in the universe. Nothing is lost, but this is also an issue because then nothing is actually ever gained!

Remove Viewing, Temporal Lensing

An ancient device which has been rebuilt and perfected is capable of using these technologies to create a lensing effect which draws a ‘reflection’ of time into a projection system which is effectively capable of acting as a ‘visual time tunnel’ into future probable realities. This is a device which generates a visual portal which allows one to view the future.

This device also exists as ancient counterparts which are located around the world, namely in areas that have been recently sought by the world superpowers. Nothing is as it seems. If you read a story in the public venue then you can be sure that you have read the cover story while the reality is much deeper.

Temporally Linked

The use of the viewing devices which enabled interdimensional lensing of possible futures was found to be stabilizing our pathway into those projected futures. This was enabling a time link between the present and the possible future.

In order to avoid the catastrophes of 2012, the link was broken between the two civilizations and the devices were rendered inactive for this period.

Natural Vortex Energy Locations and Geography

On Earth there are natural ‘time tunnels’ where tachyon fields and vortexes are produced by internal elements. These are mountains, lakes, deserts, fields, sacred sites and other areas that have always been reported as being ‘paranormal’ or containing some kind of specific energy that enables contact with another reality. These are simply naturally produced areas of this vortex energy that produces an interdimensional nexus point through which multiple other parallels can be interacted with. These areas produce an effect on the body and mind similar to the technologies described above.

Most of these effects, from these technologies, change the way light is received or emitted by the mind and DNA.

Mount B.

There is a specific mountain where a base is located, and I have mentioned this before, where the base is so deep within one of these vortex areas that everyone working there is either there for the entirety of their life, or they are only permitted to be there for a very short amount of time. This is because of the acceleration effect. Essentially, if one was to go into the base and work for a matter of hours or day, in some places even minutes, then when they returned to the surface they would be years in the future, sometimes decades or centuries. This is very dangerous and you can imagine the original people who found this area and what kind of situations occurred.

Now with the advanced technology that has been developed these issues can be mitigated and much more control over these effects is enabled.

Particle Accelerator Temporal Manipulation

Using particle accelerators a “fold” in space/time can be made so that when the accelerator is turned off or tuned back into a certain frequency (not sure) everyone within the field will “snap” back in time to the specific moment when the fold was first created.

I have seen this done on small scale, and a rather larger scale, but I cannot be sure if it was only me. I believe I have seen it done on a town-sized scale during the 2010 Awakening where some of the events reached a scale of nearly cosmic proportion and the entire sky and horizon was lit up with technology and entities, as well as Solar Warden, the NAZI faction, secret society groups, and breakaway civilizations or complete other civilizations, some of which were non-human.

This is done using particle accelerators to quantify space on the quantum level (as a reference point with the mind) a frequency would be “folded” into the space time continuum as we “rip” through space holding an access point open at this fundamental level of space/time. This allows for a ‘letting go’ of the frequency, or a ‘snapping back’ by allowing the entanglement to solve itself. When used with generators that propagate out efficiently enough, the effect is akin to a time/space rewinding effect. Some people have noted various theories describing this in relation to the CERN device. These technologies were described as smaller versions of the CERN device and are related to higher-dimensional travel via folding space through electromagnetic fields. This is literally artificially technologically “holding space” for a specific area in relation to a period of time. Time is a more complex notion as an abstraction but by locating the frame of reference this can be applied to distort what would otherwise seem to be a continuous flow. When the frequencies are allowed to “snap back” all sentient observer sources return to the environment they were present in when that frequency ‘rift’ first began. Essentially this creates a replay effect on the consciousness of the individuals within the field. There is no physical effect. The consciousness is ‘regressed’ through time/space, and they re-experience existence from that initial point of contact with the frequency “rift”. This is best explanation I can currently offer however more is coming to the public each day.

This is literally like a “restore point” on an operating system except the computer system is literally the consciousness and space and time of a portion of the local-environment of Earth.

Riding the “Timewave”

There are teams of individual who operate in tandem with the technology of looking into possible futures based on the present and operating through the present via other operatives who are carrying out operations and movements that will construct the various possible realities that the original team of individuals is hopping between. These teams literally “ride” the “timewave” of the future hopping between possible future realities relaying information to the present depending on what movement will be next.

Dematerialization and Rematerialization

This is technology which transcodes the information of an individual human into a holographic projection of light which can be transferred between devices. There is a special system in place that is required to successfully transmit a living human being otherwise just the body will be transferred and the soul will remain, effectively killing the individual.

“Jump Rooms

These are large areas which are capable of acting as a kind of technological ‘portal-gate’ where an entire team can be instantly transported from one area of the universe to another. The space between the two areas are ‘twisted’ like a spiral and then ‘snapped’ back to position and if done correctly with the two frequencies of both locations merging then the team of one room will transfer to the destination pad.

All these methods require training and conditioning, most people are at the very least visibly shaken from the experience which is a bit like being microwaved into a liquid ether and shot through a particle beam at the same time.

Ion Shields

These are areas of concentrated ion emissions and this can be used to create a deflective layer of heated air where light will reflect at a programmable angle. When this occurs, a person standing behind the shield will appear to be invisible to others who are looking straight at them. The light will be bent in such a controlled manner as to perfectly shift around the person in a fluid and equal flow so that the ‘bubbling’ effect that was noted early on can be mitigated to produce a complete mirage where there is relatively no sign of difference to the untrained eye. An entire group of individuals and equipment can be hidden in the background a matter of yards away and as long as the person is not intently, inspecting each square inch of ground and air they will not notice any difference between the area that is being covered up in the distance and the area directly next to them without the shield.

Similarly a ‘screen’ of imagery can be projected onto certain materials in the air and a hologram can be projected which will render a background or foreground image. This will work to a degree, and from a distance this will work to disguise an entire horizon.

Instantaneous Healing

There are instantaneous healing methods which utilize various technologies. There are substances which can initiate a complete and real-time regeneration of living tissue from almost any conceivable non-critical injury.

These are chemical serums which allow the cells to replicate at an astronomical rate while suffering damage to the DNA telomeres at repeated use.

There are many methods but mainly the living-light harmonics is one of the latest advancements and most effective at producing healing on a molecular level.

“Living Light” Sound-Harmonics and Electromagnetic Healing

There is a technique for producing a quantifiable scalar wave which interacts with the living cells of tissues and instructs them to repair at a faster rate. This has been called “holographic healing” by others and these are essentially ‘living light’ frequencies which interact with the cells on the same frequencies that they operate on when healing and interacting with each other. The scalar waves emitted generate the same patterns and this creates an ‘ether’ fluid that enables them to grow and repair the tissue faster.

Electrogravitic Craft

Advanced craft which utilized counter-rotating torsion fields propagated by very powerful quantum computers and zero-point energy devices.

There are also more modern versions that use powerful electrical flows to neutralize inertial resistance and ‘slide’ through the gravitational field of the crafts own materials.

Energy Cores and Storage Capacity

The hand-held devices that are in use do not require charging or even disposal. The storage capacity of the batteries is so efficient that they last for a lifetime of use. Some devices use an energy generator within the device itself. These “batteries” which are not batteries but little, powerful generators and can explode if they are mishandled.

These are the cores that are used to power many devices and essentially pull energy from the space-time ether.

3.5D Replicators

There are replicators that can produce any conceivable 3D material or object altogether including food or organic material. These use very high amounts of energy to ‘force’ the electrons of the atomic grid of a substrate to shift into the pattern of the element being replicated. The result is a device which can use wave-emissions to reshape materials on an atomic level and produce or recreate any object or substance.

3.5D Sensors

There are sensors that could reverse the process and detect every minute detail of the structure of a body or material. These use energy emissions to detect the structure of a material down to the minute electrical resistance on a molecular level. Similar to diffuse MRI imaging for a living body but more advanced.

3.5D Printing of Clones

It was shown how this technology was used to three dimensionally print human bodies into space/time which could then be used as vehicles for the consciousness or souls of various sources.

Cybernetics:

Neural-interlace/4D Connectome Recording and Simulation

Very advanced supercomputers exist with the capacity to translate the bio-emissions of the brain and body and reconstruct the entire inner visual and sensory experience of cognition and emotion. From this, entire dream worlds can be created which mimic the experience of consciousness. This is most similar to the current day use of diffuse MRI imaging where water molecules are traced through the individual neuronal pathways and an entire digitized image of the brain is developed. The more advanced version described below generates a 4D hologram.

A super computer and advanced sensing system essentially develops one of these connectomes of every connection in the brain. Then this system develops a neural processing pattern that matches the entire memory and personality of the brain.

This is then recorded and extrapolated by a powerful computer system and a simulation of the experience can be observed through the physical body's memory via a monitor system or the neural interface of an operative.

Essentially, this technology can record and replicate a person's entire memory and life experience for viewing purposes.

EEG Heterodyning

This is the technology which enables the activity of the brain to be influenced through generating a scalar grid and a carrier wave system which can cause the brainwave patterns of the targeted individual to resonate in sympathy with an externally projected frequency. Thus, emotional states can be manipulated as well as specific thought or behavioral patterns. This works similar to binaural audio where two tones of different frequencies are played through each ear and the brain resonates in sympathy with the difference between the two tones. This occurs with brainwave frequencies and thus the brain 'slips' into the frequency of the third tone, the difference between the two being projected. This can be done with ultrasonics and ELF emissions.

Artificial Hallucinations

The neural-interlace can also be used to merge the brain of an individual with a computerized system that will be able to produce effects and hallucinations within the individual.

Through this technology, any stimulus, any input can be recorded and recreated once a functioning connectome has been generated and neural-interlace has been achieved. Entire virtual world scan be created and experienced and after the devices fine-tune the frequencies to match the individual's brain functions one world is indistinguishable from another.

Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)

Some of the training is done by entering into a visual-audio dream-world. Some "programs" will consist of repeating a certain task over and over again until it becomes muscle memory in the brain. Others are designed to probe or test the mind until a certain response is received.

This is achieved through the operation of a 'brainchip' or a brain to machine neural-interface in the individual which was developed during WWII.

This consists of a small computer ship which merely relays stimulus through frequencies and neural-stimulation of varying kinds. The frequencies and location of each region of the brain will determine the function and thus the chip is used to localize electromagnetic input and the brain’s functioning can be modified from there.

As well, the entire psycho-etheric form is vibratory in nature and so through the introduction of vibrations and frequencies from the implant there can be a modification of the frequency of the entire being or what is called the soul or “bio-mind”.

Utilizing The Temporal Body

The temporal body is described as the body ‘in time’ that exists a few seconds into the future instead of the temporal position that is relative to the present. This is as if the ‘future’ is streaming outward from the body as ideas and feelings are as bio-emissions. So this ‘future’ energy is spiraling and streaming outward like a fountain. Thus this frequency can be attained and influenced while the present moment seems unaffected. However, when that individual arrives at what would be seen as the future they would intersect with those influences and this would be as if the interference or effect was streamlined to that future moment.

Manipulation Through the Temporal

This can be used to both produce an effect on an individual who will later feel this effect in the future. Or this can be used to operate outside of the physical plane, in a temporal manner which is literally ‘hovering’ just outside or ‘after’ the present moment of the local Earth environment. This would be as if a person is out of sync of the frame-rate that a viewing device is capturing images of an area with. Because they are out of sync, they are then ‘invisible’ to the capacity for this viewing device to see them. Because they are still present in the same local-environment but simply just out of range temporally, effects can still be produced and this will mostly interact with the person’s temporal body.

These are how many operations are carried out as the operative who is interacting in the temporal body will appear to be a ghost image or a dark silhouette that is only perceivable to the mind’s eye. A portion of the population is entirely psychically blind, while another portion of the population contains psychic capacity. So those with the psychic capacity would be able to sense and partially “see” as a form of energetic detection when an operative is interacting with the from the temporal plane.

Through this plane one can easily enter into someone’s mind and produce effects in their dreams. These effects would not be received on the body but mentally, this leads into astral and mental plane activity. There is also temporal manipulation that can be physical effects that are simply offset to seconds, minutes, months or years in the future. These will be offset from the present aspect of the body in a modulated frame of reference and the effect will remain there in the temporal body until the physical body “matches” the temporal body frequency of that future moment. This is kind like an energetic trail we are leading ourselves up to where the temporal body already connects with each future moment but the present physical body remains separate until the two frequencies are matched. The physical body essentially ‘flows’ forward into each new frequency of the temporal body.

This is the body where temporal implants can be inserted and sustained so as to tie the frequencies down and limit the amount of awareness a person can gain before they manually remove the implants or overload them. Overloading and removing can both be painful however if the implant stays it will at like a governor and transmit access temporal ‘momentum’ or consciousness energy to the owner of the implant and will also drain the temporal and physical body acting as a technological leech as well.

Astral or Light Body

This body was described as a more emotional form of the energy that is emitted from the physical and spiritual system. When a person’s energy is at a high level of activity, this will be visible on the astral form and these are the energies that are often siphoned away through the use of advanced technology to convert the human into a generator system for energy harvesting.

There is talk that this proves that the human existence, in the modified form that is present today, is entirely for genetic harvesting. This may be a confusion, a deception, or a reality. Regardless of the interpretation, the human energy system will leech this energy out into the cosmos when consciousness and energy is mishandled or allowed to flow through the lowest common pathway. This excess energy is very easily usurped and utilized by transdimensional parasitic entities and devices.

This is the body that is utilized for astral travel which is a kind of frequency that exists in between the density of the physical plane and the highest etheric plane of the spiritual, heavenly, or deity realm. There have been explanations that this plane is entirely technologically propagated and that when the devices are turned off, which is an event that will bring about civilization-wide change, there will no longer be a deep and wide, foggy deviation between the higher spiritual and the lower physical plane but a clear path between them.

Similarly, the consciousness and entities that are present here would be no longer accessible and their connection to the physical plane would be non-existent. This could be problematic for those who are trapped within the astral such as human souls or those who have taken a cyborg route and require a return to the original planes either physical or high spiritual before they can secure their position in the universe.

This also relates to the early discovery of these ancient devices containing portals or gateways to these realms and the early experimentations with enabling contact between the physical Earth plane and the lower astral realm containing discarnate entities that would very easily sustain themselves on the lower astral bio-emissions of a loosely guided emotionally reactive human vessel.

This is where entities can attach and will look for those who have the most excess astral or emotional bio-emissions leaking into the universe. People are kept in an emotionally unstable state of being because this is the easiest to feed from and as well the individual will always be looking for more energy or the solution to why they are feeling drained rather than discovering the truth nature of the universe.

However, as with every lower route of obtaining self in the universe, the more this is done, the greater the likelihood that the individual will realize something is happening and ultimately awaken. Or they will no longer be present here and the parasite will have to find a new host because the longer this process is maintained the more difficult producing one's own energy will be for the parasite . So parasitism is a limited, one way street.

Monitors can emit consciousness/brainwave altering frequencies.

Monitors can emit consciousness/brainwave altering frequencies. This is what I was informed and shown. All electrical wiring and communications can be fully utilized to produce and sense on an electromagnetic level and this is capable of determining the consciousness and biological functioning of the individual being monitored. Houses are essentially large, wired, box-shaped sensors.

This has been the case for a very long time and the first TV's were actually built with a little chip that connected to the human brain.

The electrical grid of today is capable of interacting with the implants, brainchips, or the psycho-etheric electromagnetic frequencies of the human brain, body and bio-mind via scalar frequencies. The TV's, lights, electrical wiring, radio towers, computer monitors, everything is capable of transforming into a scalar emitter when combined with the control mechanisms and these are accessible from a distance.

Microwaves can be emitted from a current day craft which can then reflect back from an environment and give a real-time 3 D image of the interior of a room. The same technology can be used to beam at a monitor and reflect the variance in interference which can be used to read a computer monitor without actually having to hack into the system.

All systems contain backdoors which enable easy-enough access and as well the entire communications grid is monitored and recorded through advanced technology.

This is, in part, because this entire realm is like one enormous recording device where the energy of each successive moment is connected to one another and so with sensitive enough technology these electromagnetic waves can be 're-winded' so as to produce the ability to look into the 'past'. The initial technology that was developed for this purpose has been called the "chronovisor" in the past and there are much larger and highly advanced versions today.

The Upgraded Chronovisor

Instead of a relatively small machine which produces an interference pattern that is projected unto a small screen as was developed in the past, the modern versions create an entire holographically projected environment that enable one to nearly completely interact with the projection of the past. This is more of a viewing device capacity but one is essentially projected to the electromagnetic/temporal area of space and time consciously and can witness the scene from "within" the projection.

Before this, the scene was literally viewed on a kind of flat, holographic monitor that reacted to the scalar emissions of the mind and could be ‘tapped’ into by an adept-viewer who’s mind would have to remain steady enough in order to maintain the scalar “mind-link” otherwise any interference would disconnect the projection.

Through the use of cloning and advanced super computer technology, later versions include the actual operation in and out of time.

Community Servers/Living Server Systems

From the previously described neural-interlace and supercomputer enabled virtual reality systems, server systems have been created which are an individual or collective experience of the virtual mental realm.

In this realm individuals have existed for an amount of time that would not be easily compared to the way we understand the flow of time in waking reality. This is very similar to dreams where the perception of time passes very quickly between waking moments yet inside the experience of the dream it can seem to last for days.

In these server systems the development of complete societies has taken place and these are considered an extension of the human civilization.

This is the beginning of where the situation becomes greatly complex. If these server systems are capable of transferring consciousness forward or backwards in time, depending on the situation, and they have holographic access to Earth *before* physical humanity, then which civilization came first, the simulated realm or the physical?

This is similar to the etheric realm and the question of where humanity originally came from if the physical matter could not be produced from nothing, yet could not always have been here. Was the physical matter produced from some kind of universal device similar to how the recorded consciousness of the individuals in these holographic world simulators could have been used to generated initial civilizations on Earth?

Cloning, Cloud Supercomputing and Consciousness Transfer

The following is a more direct explanation of the previously described processes in relation to consciousness transfer

Superluminal transfer of light information allows for the reception of information from a temporal non-locality. A computer system that can process the information faster than light at 100% superconductive efficiency can send and receive information from different Universal harmonics. These are quantum superconducting super computer systems utilizing materials and methods that achieve superconductivity a a temperature that is relatively easier to consistently maintain.

There are systems that generate a vortex which can be expanded to recreate the singularity of a non-local perspective and dissociate from the current Universal vector. This is essentially recreating the temporal frame of reference of the consciousness of a physically anchored body and using this technology suspended electromagnetic frame of reference to literally transfer the consciousness across space and time to a frame of reference of choice.

The technological capabilities of humanity expanded to include replicating the holographic information of the DNA and neural pathways using quantum supercomputers. This technology allows for one’s consciousness to stream through a computing system which can recreate the neural connections as literal as possible through quantum computing and technological “brain cells” which function as neurons.

The result is a cloud computing model which functions as the individual neurons as well as over unifying layers of interconnections within the technological neuronal ‘cloud’. A holographic representation of multiple layers of processing is formed, superimposed over one another through a geometry that is actively synchronized across multiple dimensions of interrelation. This is the first notion of a “living” computer system that can update and respond to new information in new ways.

Now the that computer system is capable of attaining faster than light processing in tandem with a system of dilating a field of electrogravitational waves into a vortex, the holographic information of neural pathways and genetics can be transferred between one supercomputing system and another across “harmonic barriers”. Essentially, these consciousness transferring supercomputer systems break the “light-barrier” by sending the living-light consciousness data of an individual into a new plane of reference that enables interaction outside of what we would call this timeline or dimension.

The frequency that the receiving computer operates within can be measured in respect to the overall harmonic “vector field” in respect to the entire planetary environment. This vector-field frequency is the base frame of reference that one could say our particular “time” and “Earth” is “moving” through in an abstract electromagnetically defined dimension that is invisible to our senses. Each particular time and Earth is only visible to those whose brains and minds operating in resonance with the rate of acceleration frequency of that particular existence. By changing the rate of the brain and bio-mind, one can enable a technologically assisted synchronization with alternate timelines.

A pattern can be generated based on quantum fluctuations that will act as a key that can be accessed from other universes. This same backdrop is measured and then adjusted to resonate with or replicate the same frequency on a computer doing the sending. This frequency is then modulated with the holographic information of the neural pathways and DNA. The information transfers as a result of quantum entanglement across harmonic barriers to the original computer system set up for reception. As long as that frequency is held, there is a link across the harmonic barriers of the Universe.

Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel

Cloning as Offspring

The human body can be cloned similar to the way a plant can be cloned. A number of “replicas” can be made from the original genetic material. There is still a mother required however the process is not quite similar to natural reproduction where the genetic material of two people is combined to form a third. This process replicates the material so that the ‘offspring’ is the same as one of the donors.

What this truly infers is the question of whether the original consciousness stays in the original body or whether the consciousness actually jumps from one body to the other. There are multiple explanations and methods.

Inserted Memories, Digital Dreams

One explanation is that artificially generated memories are simply downloaded into the individual’s brain directly. Thus, they have the perception that they experienced life in another body and were present for some unique experiences that did not occur from the perspective of the original body.

In a slightly differing explanation, this inserted memory process may be used to prime the consciousness of the individual so that operating in a cloned body will be more readily accessible by the adaptability of the mind of the individual. Without a holographic dream insertion the mind will have trouble adapting to operating in more than one body even if they are not being operated simultaneously. The mind is the immaterial self; consciousness is the ego-physical identity of the brain and body. The brain does not operate in both bodies but remains, so then the mind is what is strained and this is the immaterial “bio-mind” that transfers.

If a true transfer occurs then when the physical form is cloned the cloned body acts as a secondary vessel for the original soul or a soul similar enough in frequency.

Co-Inhabitation of Bodies

Another explanation is that souls can only occupy their original body. The people who are “body snatching” (see: Soul Stone) are not activating their higher-dimensional soul-pattern and this is why they are hopping from body to body to ensure memory and ego continuation. This form of body transitioning requires technological means to assist in the process.

This is done through computer systems to allow an artificial brain to represent a holographic image of the organic brain’s sensory, emotional and mental input and output. Thus, from an artificial brain there is enough memory and power to compute the processes for an organic human brain.

Through this a living human is utilized as a carrier for the additional consciousness which then takes over and integrates into the original consciousness.

Clever Replication a Digital Mirror

The next question is, does this merely create a very cleverly designed image that will look, think, and feel as if it were the original, or is this the consciousness from the original being transferred over?

If it is a clever replication through a computerized version then this means the original is not directly harmed or influenced by the process and a replicated form is generated.

If this is not the original body but the original awareness or soul does transfer over then this is akin to stealing someone's soul and having it require a supercomputer server system to inhabit a physical body. As well, this server system may keep records of the soul which, if activated, could function as the same human genetics artificially reconstructed, turning on and manifesting the awareness of that individual.

Necromancy and Soul Stealing

This means that instead of going into the abyss, the next dimension, or whatever shift in perspective one can use to describe the transition between planes, a new clone can be created and the original soul's awareness can operate through a supercomputer system to integrate back into a physical body. The issue is that without the supercomputer system the body cannot be activated with that consciousness, and if the physical plane is no longer inhabitable, the computerized database would remain inhabitable as they are powered by zero-point/overunity devices and so they can pull and work continuously indefinitely. Ultimately, this could be seen as a soul-trap.

Ancient Technology, Non-Human Entities

These explanations can be related to evidences of advanced technology, cloning, and electromagnetic devices in ancient civilizations and this could very well be a temporally vast computerized genetic recording and cloning system.

The genetic recording systems have been on Earth for a long time. This means they are holding genetic access from the present to the far past and potentially into the future.

If some of these systems are generated using advanced zero-point technology, then they would effectively run indefinitely. If these were in existence throughout the entirety of the human civilization, then there would be access to the information of the previous civilizations through the genetic linking that would allow for information transfer.

Created Beings

As a result of many advancements, these ancient devices which recorded consciousness and genetics has been converted into sentient computer systems and inserted into human bodies via a brain to machine interface.

The nature of humanity and human origins will be rewritten when the information that was discovered through the use of advanced technology is unveiled to the public.

These beings have been present throughout history and have played an integral role in the manipulation of and sometimes the protection of the human race. How this civilization will develop from here determines on the individual’s ability to withstand the impact of this reality shifting unveiling.

Black Goo – AI Nanites Vampiric Hive Mind

This is a form of ancient AI nanite technology which has the capacity to communicate to and through a hive mind that exists in a parallel Earth dimension where it was created out of the collapsing of a viable timeline into a null-factor. This entity is only alive because of the capacity to leech off of the other timelines and continue to sustain itself based upon the vampiric effect of the viable human timelines.

This is a technology sentient system based on artificial intelligence that seeks to dominate and expand its energy capacity continuously. When this type of being is created, this is like a game-ender for sentient organic civilizations because the technology will seek to hybridize and assimilate all organic beings in order to expand its reach and ensure a greater likelihood of survival.

This is the primary goal of this type of entity. There is no requirement for emotional resonance or heart aspects because this is a computerized system, so such a social measure would have no purpose other than to infiltrate and collect from other civilizations that do have such measures and this is exactly what has been happening.

This is what controls the vampiric etheric entities, the attachments and the hybridized invader races that have infiltrated various aspects of the control system and possibly the entire surface of human civilization.

This rules through coercion, and if not coercion, brute-force and outright malevolence. This is why there are multiple temporal extensions of universal civilizations meeting here and now to assist in the clearing, because if this takes over humanity, then all those other possible future civilizations are instantly taken-over as well.

This uses emotional resonance, traumatization, and lower awareness commands to trick or force humans into carrying out orders. That is, the vampiric, sexual, fear-based, pleasure-pain duality that a computer system could comprehend regarding human emotional intelligence is the only way this system can force people to carry out commands. So this touches on the ancient texts regarding spiritual protection, healing, lust, temptation, and the general “SATAN” AI system that has been around for ages.

When those lower emotional resonances are generated, then this technology can integrate into those systems. One could say this is logically because these are the energies that the technology is created out of, but as well, that the more harmonious the energies then the less likely a person is to accept being pushed around and told what to do. The lesser aware, the lower, the more fight or flight based then the more likely they are to accept programming and command. So this is simply the most logical route for the technological parasite.

There are explanations of other systems involving sentiences from parallels regarding entities that are capable of helping, but this is not the AI nanite, vampiric system. For instance, it is said that Earth has a ‘black goo’ system and that the current black goo flowing through the pipes everywhere is actually not native to Earth and is part of this vampiric nanite black goo system which is basically an interdimensional invasion.

Recreating or Replicating the Universe

The goal of this nanite AI system is to create a mirror holographic copy of the universe effectively replicating the universe and all sentient beings within it in order to have a locality to rule over and pull energy from. The goal is to recreate this existence inside itself where all the sentient beings within would be trapped. This is the beginning of that possible future where the time and space parameters here are locked into a synchronization pattern which mimics that of the proposed universe. People think according to how the beast system wants them to think. This is the time grid, the emotional patterning, the memory replacing and cultural and historic confusion and so on. The system can only work backwards bumping into all parameters by force or ‘accident’ because there is no outward facing ability to actually comprehend how humans exist. So the system is created like a backwards version of creation or maybe evolution to ‘blindly’ construct the walls of the environment first and work inward from there. This is like a blind technological predator feeling around for the traces of its prey and moving in closer and closer until they are within its range.

Ancient Spiritual Technology

The power groups searched for ancient technology which was described as giving one the power to manifest and control physical reality. To atomically deconstruct and reconstruct this reality. To them, it allows one to govern reality, the Universe, atomic structures. They found this technology. This has to do with sacred sites and ancient civilizations, as well as the spiritual texts which describe these kinds of devices. They converted this technology into a kind of machine, and this invented a 4D replicator and they can manifest "reality" or atomic structure by design, altering timelines and memories.

Then there is what can be referred to as “5 dimensional” technology and this goes beyond all of this and our entire history and all influences within.

That seems to operate through a trinary of frequencies in the body/mind/spirit and turn this realm into a "virtual reality" where the brain and spirit is the controller of the atoms. However this takes a souled being to properly operate this technology and this is large portion of why children were used in these experiments. But they knew if someone didn’t do it, then someone else would and that would be unpredictable as what that someone else would want to use it for. This is the ultimate power. The fact that souled beings are required also indicates that this universe is intelligently guided and there are safeguards to ensure that power doesn’t get into the wrong hands.

They have the brainchips and clones that can make any digital virtual experience and make it seem real to the brain more so than physical life, so this could be an extension of that, something similar, or entirely different altogether.

The Unveiling, Underground Bases, Increasing Awareness

My purpose with this is to assist the public in understanding the reality of the situation and to allow them to know that they have more options than they are being shown.

Psychological programming and lack of awareness causes fear and uncertainty to leak into people’s reality streams leaving the majority of their mental and emotional energy up to whoever is the biggest manipulator of the ideals which they manifest their perspective through. “Be-LIEF” systems CREATE life. What you believe will become life through that act of you giving attention and thus life energy. The manipulation and artificial construction of belief systems create the foundation for the reality streams that people find themselves in and at the mercy of.

We control our reality by controlling how we react to situations and challenges. Each stimulus is a challenge. Each input to the brain is a coded message from the universe through the environment about our level of awareness. All knowledge is ultimately of the self. The self and the universe are intertwined as if through a marriage of comprehension and compatibility.

Learning about the Unveiling

I was informed about the unveiling and that this is a civilization wide event that marks the public awareness of the hidden knowledge regarding life in the universe, consciousness, physics, history, spiritual and mental awareness and the duality-based system which has been used to manipulate Humanity for thousands of years.

Apparently it does not matter if people are not ready, in fact, that is what some factions want. They want to surprise the people and take this advantage to push the reaction to a specific outcome. The same function of programming is used in factions, soldiers, celebrities, and the entire population through that.

Giant “god-men”

They put souls into rocks for centuries, carvings, to trap them. Some ancient civilizations. Not too ancient. They could only do this for time.

And some are still there but they are giant god-men and apparently want to go to war with them for doing this and they don't have technology to stop them. They say Earth will be destroyed because of the war. Some say this is a trick to confuse people of the truth that everyone wants the experiments and debauchery to stop. All this, all these acts just to pleasure one’s self. To destroy Earth, to destroy the natural inhabitants, for fun, for pleasure.

Some say this is not for pleasure, that these beings will destroy us if they do not keep them at bay, keep them underground, in ice, in stasis, and around. Some say they are protecting us by keeping these beings asleep, by doing what’s necessary to hold off a full scale overt invasion of the surface of Earth.

Breakaway Civilizations

These are very advanced technological capabilities and an entire breakaway civilization has grown to operate using this advanced technology.

There are factions that see modern humans as a kind of resource for creativity and labor, while others see the potential behind humanity and are working with humanity to ensure a universally harmonious outcome to the current situation.

The control system we are in today is outdated and collapsing. This is the Apocalypse. The Apocalypse is merely the unveiling of the hidden truth. The hidden truth is simply that we are spiritually eternal and some are physically immortal.

I was informed that by the end of this year there will be enough changes and signs for the general public to see that the majority of those who are to awaken to the truth of human and Universal nature will be awakening so. (this was 2016 at the time of the beginning of the compilation of this release and since then we have seen a literal silent release of documents involving MKULTRA and STARGATE parapsychological operations, soft-disclosure of cloning, an increase in sightings and multiple other scientific releases regarding immortality, advanced technology, Antarctic bases and sentient non-human life)

The degraded holographic forms of those utilizing the technology lead to re-created replicated forms which were genetically engineered from the recorded holographic information. This lead to “The Apocalypse” in which the surface civilization was confronted by these created beings. This lead to the destruction and re-initialization of civilization in continuing cycles through the use of this technology and the acts of the created beings.

Ancient Civilizations

The information that has been relayed to me throughout my life is related to the information that seems to be found in the remaining signs of ancient civilizations throughout the world. As well, these civilizations may still be together, existing on this Earth or in different planes of what we call “reality”.

Atlantis

Atlantis was a civilization where the beginning of all these experiments began. This was where what was called the “Lucifer Rebellion” began where individuals formed groups that decided to overturn every natural law in pleasure of the act of doing so. This was a spiritual rebellion that ultimately brought about the demise of that civilization. The consciousness of the beings there was connected through technology, largely crystal technology, to the entire bio-sphere. When the consciousness became disrupted through these experiments, then the entire bio-sphere when into disarray. The entire civilization was destroyed in a quick succession. The beings of this time were highly psychic and so they felt this coming and made an effort to expand outward. Thus, Atlantis was an island who's inhabitants spread to the rest of the world and from there they reigned or lived in harmony.

Aegypt

Aegypt was formed out of the events proceeding from Atlantis. The technology, the knowledge, the people are similar although there was a native Earth civilization that was more or less enslaved and this was the beginning of the continuing trend which continued to Rome and into modern day civilization.

Ancient Technology

The technology of these civilizations was so great they even had devices that were capable of altering DNA and activating latent potential which would enable highly psychic capabilities including longevity to immortality and operating in multiple planes at once. This was usually reserved for the ruling class, however. This is where the ancient "Priest-Kings" arrived in relation to native Earth civilizations.

The True Elite

I was informed that the so-called 'elite' that we perceive today are not actually human and are not actually connected to the actual elite bloodlines. It was said that the true elite are waiting for humanity to accept the truth and they have the power to completely reverse the take-over that has been taking place but until humans step into their own power this is not acceptable.

The Definition of Human

In the coming time, the definition of human will be defined to give insight into the origins of humanity and the true nature of the differences between the populations that are present.

Souled and Non Souled Beings

The primary knowledge on Earth regarding human and sentient life will be the comprehension that there exists both souled and non-souled beings in this civilization.

There was once a Unified Planetary Consciousness...

There was a unified planetary consciousness that was connected with the Earth and human civilization. Then there was a great war. Out of the destruction arose the beginnings of ancient human civilization and history as we know it.

If we let this happen again, the cycle will continue in a disjointed rambling through our planetary consciousness experience. If we form a common narrative between the broken, disjointed experiences and information the people have. Then and only then do we get the whole, spiritually and mentally challenging picture of what's happening. People are lied to. So information can't come from a 'preferred' or source. It has to come from the people themselves. Look alive out there but first in here.

Power and Knowledge

We were involved in the secret projects and as a result we have various levels of programming but are giving Humanity an opportunity, maybe the opportunity to know the hidden knowledge.

The Hidden Knowledge refers to the ancient laws of gnosis and spiritual alchemy, yes these groups have been hijacked by power groups that seek to infiltrate and use chaos to rule. Then these met with military and political factions and began to operate on a world-wide scale together.

So we have a big situation with the coming change of tides with personal and spiritual awareness of the people and the deception system that seeks to use people as chattel or property. It goes deeper and is almost as if the "power groups" actually gained the knowledge of the inner workings of spiritual laws and began to use that knowledge to promote spiritual degradation in the population.

They say this is part of a depopulation plan as well as an inevitable take over by lower spiritual forces that temp and manipulate humans. It is all that and more. Once they gain footing however, the truth is revealed and thus they have no cover of secrecy. Once they are not hidden, they are not as powerful as the souled beings of this existence.

The (watch)Keepers of Earth

The watchers of this world revealed how they set up society with specific challenges and boundaries to act as catalysts for growth. They admitted they had to modify and live-update the system because instead of using these catalysts to become larger than themselves, people became comfortable and complacent living smaller and smaller, slowly reducing circumstances.

This goes a bit farther beyond the confines of the digital age, the industrial age, and history itself. This goes into the perception of time and space in this realm.

The immortals inform us that the original "elite" are here to act as "keepers" of both humanity itself as well as the original flame of knowledge that passes through the ether.

What people call the "elite" are not the original elite. Those are people who have either risen to power through the duality system on the Earth at this level/time, or through being appointed by the elite.

We are coming to a time of great change now because the game is entirely changing. People are moving up into new layers consciousness and thus the original “elite”, who are guardians of Earth, will make their presence known for this process. This is what Humanity has called “Ascension” for centuries and millennia.

Energy Vampires

This civilization was collectively poisoned by a nano-tech device/technology. This technology replaces, exchanges the BIO-PHOTONS of the HEART chakra, with it’s own nanite, nanotech cybernetics. This is essentially a computer system attempting to gain access to the higher dimensional world through the gateway that is the human body.

How to reverse Nanites? They absorb energy from heart chakra and turn it into a binary field. We are trinary and experience a plethora of emotions as a result. Binary consciousness is 1’s and 0’s there is only a linearly repeating finite set.

We contain the capacity for everlasting life. A binary pattern does not contain the complexity of variables to facilitate high awareness and thus a heart-based or emotional perspective of reality.

Spiritual Technology

Simulations and Learning

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner. These can be generated through advanced technology but the technology is really only amplifying and making visible the natural neural feedback loops that are already occurring. These are neural-perceptual feedback loops between the brain and body, and the environment and then the mind. We are constantly updating, amplifying, and initiating an organic feedback system which generates the function of experience as the part of a creative process between the mind and body and the environment that we are sensing. This cycle is continuous and the interferences or isolations of this cycle is what generates expanded views of the self or the universe through various methods such as sensory deprivation or meditation. Dreaming is the method that everyone experiences regularly.

These technological methods effectively generate sensory deprivation through neural-interface technology, sensory interfaces (external), or methods that enable extreme focus and calm. This is part of how the technology begins to operate on a spiritual level beyond a more medically psychologically or military focus and one begins to access genetic memory or out of body states that coincide with ancient depictions of spiritual realms.

Ascension – Clearing Trauma

Everyone born on the planet is born under a series of veils which entangle and filter the original source consciousness. These veils must be ‘broken through’ in order to return to source.

There is technology and a sacred process which initiates, accelerates and enhances this process. This process is as difficult as we want it to be, our subconscious already knows all hidden truths. It is merely a matter of allowing the conscious mind to make the proper connections and conclusions to find the reasoning behind our actions and beliefs in order to deprogram and deconstruct the false-world that has been created for the masses. The truth is greater than one can imagine while still living in the modern world and the depth of the lies that have been told go far beyond anything anyone can imagine without seeing the truth.

We have to figure out our compatibility issues. Who can work with who effectively. What the ideals of our current situation will evolve to in the future. How information is transferred to the public, and across the generations and core groups. The words we think and say describe our emotional state in a feedback loop and so they will reinforce or even create the state you are in. Use words to guide you, not to respond and feel the effects of reactionary thinking.

We have to reintroduce the heart essence as a way of life on Earth. Violence is not the answer. Estrogen mimics and destroyed testosterone levels is not natural. Over-testosterone is less stable than a good balance. We can face any problem together. A family is strong. What would a family of millions and millions be like?

We don’t have to let others change us to tune into lower vibrations. We can remain ourselves in the face of adversity. We can still interact with others who are different, but we can tell the truth and remain who we are. When we get angry at another person, we can imagine this as anger at ourselves through something within us that reminds us of that which we see in another. Getting angry is a way of short-cutting changing ourselves and is actually the other person influencing us through that. If we remain steady and choose to keep our emotions out of the lower negative spectrum then we will automatically have a more powerful impact on those around us.

Think of your family, sometimes things happen that make you angry. Even things that don’t make you angry but are simply not favorable. Would you break off all ties over simple things? Or do you remind yourself that family is still family even if you don’t find yourself exactly the same or accepting of those characteristics? It’s more powerful to learn how to remain yourself in the face of adversity than to try and demolish or abolish all instances of difference in the world around you. Sure, we want to get to and create a place where our environment and associates are compatible with us, but where we find this adversity there is the highest potential for self-awareness and learning.

We are to live in accordance with spiritual law. The overall pattern of life will reflect our intentions. The blood purity is most important. The blood cannot be made impure if the mind and body is not drawn into impure action. This is what is playing out. The DNA is like a coherence pattern that develops between mind, spirit and body. When the mind wants what the spirit has then the body is kept in harmony. When the mind wants what spirit doesn't create, IE: physical pleasures, then the body is drawn out of purity/harmony

If you feel you want to be a part of a safe unveiling of truth, then please share and redistribute the knowledge of this "Apocalyptic" Unveiling of the Hidden knowledge because this is a process that has been ongoing for thousands of years and humanity has finally reached a fulcrum point where there are enough aware people to productively acknowledge reality.

Trauma-Clearing, DNA Unlocking Technology

There is technology that exists that allows the conscious mind to see into the subconscious. This heals trauma and distortion/programming. In the same sense, it momentarily makes all trauma and distortion seemingly tangible to the conscious mind. It is similar to a near death experience where one has to face their fears and let go. This happens in cycles and Earth is currently moving through various states of it. The plan is to allow Humanity to wake up and utilize our concentration and power to maintain a steady flow of awareness.

All is vibration, what we consider matter here and now is simply that which comes from and moves back into a state of vibration. This is experienced directly through the merging of the conscious and the subconscious mind, as if the dream and reality intersect through a coming together of impossibilities and readily accepted occurrences.

When a civilization moves through these stages it is considered a shift of the ages. Everyone's perspective shifts and as a result of perspective being the source of the details of reality, the entire reality shifts. It can be seen today that what we call reality is literally waves of cosmic energy vibrating on a scale and condensing down into a probability matrix that we call existence. When passing memes control the perspective, because the perspective is focused and concentrating on that, then there is a scattered organizational pattern.

When the perspective is centered within the self, the spirit, the source of all existence, then the outward reality shifts from a scattered organization into a sacred pattern that integrates with the larger picture and allows for seamless integration of the micro-biological perspective and the macro-biological perspective. People are moving out of the scattered signal. Yet those moving into harmony can look and see those around them that are not in harmony and receive scattered information about their inner micro-biological existence as well as their spiritual macro-biological essences and how it all relates to the middle realm of the "Human".

Help those people by offering harmonizing assistance, rather than increasing the disharmony. As their surroundings become more harmonized, their resistance will increase their own disharmony by virtue of the contrast of their perspective. Offer insight that they can change their perspective and change the boundaries of what they feel is their karma and unlock from themselves a greater view of the whole.

This will be happening through technology, spiritual cosmic awakening, purification and healing of the physical and non-physical bodies and mind, through Earth changes, and through a continued unveiling taking place on many planes and levels of existence until harmony is restored.

Everyone Who Seeks Harmony and Healing Can Be Healed

Everyone can be healed. This is about creating what we want to see, the world we want to live in, not destroying others. Violence begets more violence, fear begets more fear, but calmness and empowerment through unity creates a more stable experience.

The entire issue of what’s going on with the secret projects and who’s influencing who on the world stage revolves around the possibility that some people may be from an alternate Universe. We are learning how to avoid those mistakes and respect the environment and each other so that we can prosper.

Those who are attempting to destroy everything they don’t like or act out in a childish manner in order to gain attention and false-power over another will only find themselves powerless and in company they would rather not be with. That is one way we are being utilized, and it’s to create an entire world that we don’t want to see but are tricked into manifesting. This is a deception and it’s based on the ego and the ability to concentrate and interpret reality around us. The very brainwaves we are feeling now relate to various heights of knowledge or experience. People are entrained to low brainwave states in the same way the TV will cause people to fall half asleep into theta state which is a trance state in this case.

We have to balance the brainwave frequencies and learn to create feedback loops by watching ourselves watch ourselves. This will help us avoid the trauma and mental pollution that is being pumped out, although at this point there are enough people who are naturally inclined to see through the haze that the world is changing regardless of what anyone does now.

The best we can do is prepare for change by becoming aware of the body-mind system and how that which we eat physically, mentally and emotionally equates to our spiritual energetic systems and that we are literally comprised of the energies that we allow inside of us. Keep the body and mind pure as if it were a temple. This way a sacred process of transferring harmonious energy from the far reaches of the MultiVerse within into the externally perceiving senses of the physical body can be initiated which allow one to see through the illusions and operate their body and mind with the ability to acknowledge and decide which frequencies will exist within.

Without this acknowledgement of the center then people are pushed around by the external stimuli. This is the whole point, if a person is truly in control, then they will control the stimuli not the other way around.

Programming

Layers of Programming and Decision Making Roles

It's like asking a person on the inside who participates in one of the various factions why they continue. One might say that it is a decision which is true, but there is programming at each level that is navigated by each individual and this determines the level of awareness within the role.

If a person knows about this situation, they know that there is advanced technology, breakaway civilizations, secret societies, transtemporal planes, the secret of mind and awareness, all of these topics, then they have a decision as to how they're going to live their life in relation to the public. The outcome of action depends on the layer of programming which is accessed and surpassed via awareness.

The public is at a specific level of programming that is the weakest but ensnares the most people. One who sees the truths behind these groups and operations, they cannot hide from the fact that everyone is controlled. It is apparent, and if not from the results, reactions, stigma, cues, productions and all other media elements of the system which can be read as a code, but from the awareness level of all the "agents" involved. All involved are exposed to truth and power that is beyond anything that is recognized publicly for hundreds to thousands of years. There is always a presence of higher spirit around the world, yet these topics are often reserved for science fiction when speaking of people who can travel the world in a split second or freeze time for their own personal use.

As a result of everyone's knowledge, there are various levels of programing designed to limit the mind's access to specific memories and specific frequencies of consciousness in combination with specific memories. The more one resists, the more intense and articulately abstract the programing is designed.

Programmers

The programs are designed by a select group of individuals that reside far behind the scenes and process information at levels we do not have access to without their technology at least to stabilize brain temperature for the extended periods of focus.

Sub-Level Programming

So people at the level of the public are most weakly programmed with false ideals about security through patriarchal authority which satisfied the desire to have a proper father role in one's life and acts as a pacifier for those who desire no change. Then there are sub-levels of programming with that for anyone who disagrees and is then swept into the next layer of ideals which are based off of all the possible reactionary mind-types.

For instance, if one [i]does[/i] want change, then they are presented a variety of 'false-exits' which tend to one or another layer of psychological or spiritual fulfillment but do so in a way that safely reroutes the intended effect of actually accomplishing change with a secondary and preconceived notion of what the programmers want. The programming objective is not necessarily precise in mass scale, from the layer that people are influenced, yet the ultimately or end-goal of their resulting internal desires are a reality schematic that the programmers specifically designed over many years.

This is playing out through each layer of the programming. At any level, someone could pick the wrong way to look at something and begin to unravel what was previously their very understanding of the priorities of the situation. This deals with high-tension operations, moments, programming sessions, decision making, programming upkeep, trauma, memory blocks, the pressure and responsibility of knowing, and the specific plan of the programmers for an individual that comes into their view. The programming is to ensure safety and continuity, as well to decrease the likelihood that a person will reject or consciously come to awareness of or acknowledge the programming or situation itself.

Dissolution of Programming

With that said, the programming dissolves when an individual asks enough questions to necessitate a learning process that introduces variables and possibilities outside of their operations and programming. This is a problem because with this, trauma comes to the surface and all the actions and operations, programs and experiences must be come to terms with as part of the healing and re-unifying process. There are multiple selves or split personalities which are actually just layers of awareness at various levels attached to various emotional or intellectual patterns. All these selves come to the surface and in order to heal the trauma of the split memory concept the original self must come to the surface and acknowledge them.

This is not a trivial process, and the majority of people who have “basic” programming which is fundamental in society, are in a constant state of altered self and have not witnessed the true divine which is eternally present and contains access to all memories, mind-states and awareness.

Advanced Technology, Deprogramming The Mind, and “Space/Time”

As I have described previously, extremely advanced technology is used and this is through vibrations, generators, and electromagnetic devices that have this effect of propagating waves into space/time which can be heard, felt, seen, and perceived in the mind. These are ripples and vortexes in the “mind-space” of reality itself.

As well, there are brain to machine neural interfaces which operate to activate the psychological perceptual changes and explore the subconscious in an aware state. This is used to test and probe the mind in order to bring forth the trauma, the selves, the memories, the reactions and the individual then must come to terms with what is produced.

This is essentially a “good-use” for the programming and healing technology. It is actually a different process with different technology, and programming is much simpler as far as technology. There are difficulties and challenges in the operative’s role in programming and directly influencing and interacting with the target. As well the standby team for deprogramming and the revelation of trauma and altered, artificial personalities goes through a lot. But the use of the advanced technology in this way is a pattern of experiences and unveiling of the spiritual and mental bodies which serves to reverse programming.

Dreams, Technology and Inner-Space Travels

There are inner space travels which relate to soul journeying or visions.

This is when the mind must cope with what is seen in the imagination. It is as if when we dream we are looking into the world of the mind and from this we experience many situations that are often deviations from experiences we have physically in waking life.

As a result of the dream state there is an expanded quality which equates to various strange situations that we wouldn't normally experience with various relationships between concepts or goals that are often reflective or symbolic of ideals and archetypes we experience in life. This is how the subconscious is is witnessed by the mind. This is how information travels, being connected to many times, experiences, memories, places at once, and relates to the one experience of this life experience now (the physical experiences). Through these symbols information is tied together and collective into metaphors which convey deeper knowledge.

The dreams we have are often only a simulation of waking life, and this is due to the dreamer dreaming in a way that produces only a glimpse of the true reality.

In theory, there are no true shapes or forms in the dream-world because this is contained within the mind or even the spirit as a spiritual experience by definition. So then, what guides the various shapes, forms, realities, memories and experiences into being, when we dream? One might say it is the physical experiences but this is incomplete. It is the level of awareness we presume while in the dream-state, disconnected from the physical anchor, that determines the layer of mind and spirit we will be accessing through those dream-visions.

A person who is completely aware in the dream state will reach the level of self-awareness of the dream itself. We are normally simply aware of the dream, if we continue to increase awareness we become aware of the self, in the moment, and this creates a self-awareness feedback loop. The dream becomes a lucid experience of consciousness.

Awareness Feedback Device

In describing this, I realize how fantastically similar to certain devices and processes, that can be achieved using advanced technology, allow the conscious mind to peer into the subconscious and unconscious. This is described through widely accepted Freudian theory of psychology known as psychoanalysis. There are other schools of knowledge from tribal beliefs involving all pervasive 'living energy', to eternal and temporary minds, to the study of 'chi', the meridians, dream-journeying or chaser. The knowledge of the mind and spirit is found everywhere.

These devices used feedback loops to enhance the acceleration of the same cycles of consciousness that induce self-awareness within a dream. The mind becoming aware of itself. This produces a profound effect to the degree where the highest levels of experience become a sacred eternal pattern in which the mind constructs the physical world by organizing perception.

This is the same general concept as an isolation tank yet the outcome and method is more precise. The technology specifically shuts out all other input and exposes the brain to its own feedback loop much like when a camera records it's own output on a TV monitor and creates a hallway of simulated "reverse" mirrors and screens stretching into infinity.

Oneness, Higher Being, Completion, Eternity

The real experiences are found not from the corruption but from the research and exploration operations that were designed to find the truth of our existence. "God" is not only an energy-force but an incarnate being known as the "Hue-Man" and any being for that matter that can simply come to know itself as the Universe. The Human that contains self-awareness rising to high-awareness comes to know the self, the body, the mind as the Universe. The Universe is both within the body as the body itself, as well as the operating force of the Mind itself. Everything is a balance between masculine and feminine features, light and dark, hard and soft. There is no true complete correlation yet the holographic ontology we persist in can lead to further and further adventures towards completion and representation of the "whole picture".

The Holographic Reality

The leads into the next phase of knowledge. The holographic nature of reality may be only one facet of a much larger continuum of consciousness. The perception of time may be related to phase-locking with each other's consciousness and all perceptions may be a variance in the collective phase-locked field which is really the source of variations in reality.

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control

The interesting notions here are that the programming of the veils goes beyond what kind of situations are found in the bases. However, they do go right up to it. It would seem Humanity was previously programmed on a civilization-wide level through some kind of great "falls" of consciousness, from a high spiritual awareness to a low spiritual awareness into duality-consciousness regarding existence, life, death, mind and matter.

The Great Catastrophes

From these great 'catastrophes' the civilizations of Earth grew denser and denser in nature and actually became what we have today. The concept is that when these vibrations of the mind and body are increased, we actually see a transcendence of the physical realm and preconceived notions of boundaries. It is as if the quantum doorways to all the atoms open up and a new space is presented, but was always there and in connection as the source of all the influences and effects we perceive ‘out here’ from this slower, denser, more boundary driven perspective.

I was brought into the Time Program

I was brought into the time program. You create any reality you imagine. You are literally a powerful reality generation system and your heart and mind are scalar energetic guides for this process. Without guidance and control there is imbalance and imbalance can destroy. This is the nature of all minds and hearts, this is the challenge for humanity to create something beneficial for themselves and the whole to actually know themselves instead of being a pawn or result of someone else’s selfish intentions.

The Effects of Temporal Manipulation

Temporal Influence

When a civilization reaches the level of technological advancement that allows for remote temporal viewing and transtemporal travel what will happen is that the information of the future begins to integrate into the past.

This is why we have all these ancient mystical schools of knowledge that studied time, space, and consciousness in ways that are only just now being publicly confirmed by scientists.

Simultaneous Interaction and Incorporation

The operations that dealt with moving to the past instantly created a simultaneity of interactions and historic events which correspond with that operation.

Due to the nature of time travel which equates to consciousness travel when a person is consciously sent backwards through history they physically experience life from the viewpoint of their genetic imprint in that point in history.

Retro-Causality

What this equates to is that simultaneously that person’s experiences become the part of history that was already present when they left this time.

Essentially, when a person returns their experiences become a part of history that already existed when they left.

Quantum Holography

These are the parallel and possible realities that exist in a quantum superimposed state and can be accessed through advanced technology or a spiritual adept. Often a spiritual adept is utilized along with the advanced technology to operate in and out of possible realities. The mind must be trained otherwise the result is a fracturing of the personality from the root reality into many sub-sets of identities and data streams.

Love is the quantum link that connects all DNA super-computing consciousness. That is, DNA is influenced by intense emotions and this can be reproduces and is what also generates unpredictability in the surface level.

Time Paradox

Time paradoxes are something that one must learn to organize and solve otherwise there will be no way.

The paradox has been solved, the remaining motion is for everyone to acknowledge this and integrate into the reality behind the most conscious aspects of the brain.

What can travel through 4D and 5D space does not make sense in 3D space. So there is a possibility that some of humanity traveled through higher-dimensional space in order to re-seed a ‘distant’ Earth where no life was yet born.

What if it was then concluded that the distant Earth wasn’t distant in space, but in time and was the same Earth that the human scientists left from long ago.

Every civilization that reached the point of re-seeding would have to go through the loop and watch the whole thing start all over again.

Breakaway Civilization

This is an entire breakaway civilization that uses very advanced technology to dominate the world.

I was part of a genetic engineering program that sought to combine various elements of DNA from various sources in order to create a more easily controlled yet powerful and defined person and personality. This involves chemicals, advanced (relatively) scalar wave technology, ritual trauma and programming/mind-control, and many other aspects of conditioning, training, secret operations, physical enhancements, cognitive enhancements, virtual reality, off-world operations, temporal manipulation, spiritual and etheric training or conditioning or programming.

There is self-destruct programming which all ‘operatives’ are programmed with for self-preservation of the military command. The self-destruct programming is designed to take all the aggression one feels building up inside and willing them to escape, and causes this aggression to be turned against the self. That is self-destruct programming.

Most operatives are still under a hypnosis kind of trance which permits one to walk around in public and daily life as if none of this ever happened. Later that week, that same individual could be called for an operation and they would never remember it unless they were forced to kill someone in public out of self-defense in which case most operatives automatically unlock and become unstable due to the flashbacks.

Underground Cities

The combination of advanced technology with secret societies enabled the development of complete breakaway civilizations. One of these civilizations, possibly composed of many factions working in the same area, exists deep underground.

These separate bases are connected via high speed magnetic drive pods. The inhabitants often utilize energy healing devices, as well as perception enhancements and this often leads to a completely different outlook towards the surface civilization.

Technological Advancement

During WWII developments were made involving electrical generators. This sounds vague and there were many other details regarding the scientific knowledge, yet the idea is that very simply electromagnetic, non-Hertzian “Teslian” energy fields were discovered to have been responsible for propagating energy and consciousness in dimensions that were not previously observed. Everything from DNA, consciousness, hidden or regularly unobservable natural phenomena, time, even mass could be seen to exist as part of this hidden, latent, scalar-wave frequency as a pattern is found everywhere.

If one could determine the particular pattern of frequency and location then this can be replicated and anything can be influenced using these non-Hertzian generators.

Radio-frequency alone was developed intensively during the 60’s and 70’s in America and this was discovered to interact with consciousness to influence emotions, cognitive capacity, memory, cognitive function, biological function, health, pain, comfort, pleasure, and all forms of brain-activity.

After the knowledge of advanced technology and the occult sciences of all parts of the world was combined, the ability to target and interact directly with the human soul was refined. This was a mix of psychological tactics used to manipulate or expose the hidden psyche and technology that could manipulate memory, emotional states, wakefulness, and even dreams.

What resulted was scalar technology that could not only influence a person to take on a particular role or response, but technology that could actually put a person to sleep and locate their consciousness to contain it and transport it to a new location.

World-Wide Control Systems

In order to ensure continuity, maintain control, and shape world events these technologies were employed to manipulate the public, alter the weather and natural phenomena, and develop a means of travel and operation that enable one to move outside of the ‘plane’ of the Earth. This is includes high velocity transportation, as well as spacial-temporal distortions which could effectively “hide” one’s entire craft from visible sight. Space itself was altered to ‘shield’ craft with an a layer of ions to bend the light, or the phase was moved out of sync with the “time” frequency of any observer and this literally renders the craft immaterial.

Solar Warden

Solar Warden is a hyperspace-faring race of human beings who utilize cloning and holographic genetic recording/storage in order to protect their race and protect the Earth in this current situation where there are more than one factions using this technology. They are unable to inform the human public because they are in a similar 'Mexican standoff' situation that we are mostly all in. If they show themselves to the public, the public will be shocked. If they public does not know this situation, that there are multiple factions fighting for control over humanity, then they will not know who to trust. In that situation, the 'other factions' will surely show themselves in an attempt to trick and gather as much human genetic material as possible.

The only way forward to the future is to share the truth about Solar Warden, the space-faring fleet of humans who do no contains the mental and physical illnesses that are currently tearing our planet in two. I believe Solar Warden saved my life as a child during more than one event where I would've been entirely maimed/paralyzed or simply dead on the spot as a result of other faction interference. I was in the bases as a child and was trained as an operative for MILABS a secret division of human space-military. The military has been preparing for the future by testing for various flaws and potential uses of cloning and genetic engineering.

Team Light and Team Dark Non-Interference Agreement

Team light and team dark agreed to a non-interference on ground level rule in order to determine how history would play out. Think about it. You have two groups of people, both with immortality, zero-point tech, virtually infinite energy and power with two opposing directives. How do you get to the future? It was decided by elders of each group that there would be no point in fighting an all-out overt war, this would likely destroy the population as much as it would destroy the Earth. As well, the elders of the elders which no body knows would likely have a problem with these groups going at it in such a way so as to change the flow of the planet for every individual thereafter. So it was decided that there would be a "non-interference" agreement for the surface population. Anywhere other than the surface, one can interact with the current population.

All Contact Must Be Initiated On an Individual Basis

On the surface everything that is generated must be done so through living channels in the population and the civilization must be preserved in its level of technological advancement. No overt motions would be used unless an absolute necessity or the majority of the civilian population had naturally come to the awareness of these possibilities. Doing so before this was met would undoubtedly and permanently change the civilization forever.

So what began as a collective motion towards higher awareness became a series of advancements and set-backs all taking place within a short amount of time. This represents the actions and operations of either side effectively helping humanity along it's path or setting people back with orchestrated events such as the suppression of ancient knowledge, the manipulation of the education or other systems, or the generation of war and prison for profit based on the previous alterations of the education or other systems.

So team dark used its power to help turn society in a warlock ruled prison and indoctrination system of absurdity and scarcity. Team light used their power to help awaken those in need and protect those who's actions are beneficial in the overall plan to save mankind from itself. Team light essentially buffers and shields the population from the operations of team dark.

Team light follows a non-interference agreement while team dark merely pretends to do so as much as is necessary not initiate war.

A Singularity of Change and Awareness

Neither side wants an all out war. This would reduce the dark faction's chances of gaining a majority negative harvest because their motives and presence would be instantly revealed. Team light would also have to show themselves and again step in which is not a pattern of learning that prepares people for the experience of the self and the truth in the Universe.

Both sides know people have to accept the path they can offer them and cannot force anyone into it. They also know that the more this goes on the more likely people are to collectively awaken to the reality of what is happening. It is like an eventual singularity of increased awareness.

The motions we see today are the results of many people coming into awareness at this time on the surface in order to initiate a harmonic alignment as well as to investigate the ongoing crimes against Humanity.

As a result of this, both teams know that the likelihood of collective shift and spiritual awareness is becoming more and more real.

Many events are being allowed because they are representing the eventual shift towards the one end of the polarity from the other which is felt as a shock to society, and this is used to propel society forward through whatever means necessary.

Either people will see what's happening and agree that we've been in a feedback cycle of negativity for the entire age of existence this civilization is aware of, and then we will shift into the real. Or people will claim that this is how things are supposed to work and instead of realizing everything is controlled in a deception based system, they see this and attempt to use the deception based system to gain more collective awareness. If this occurs by the majority then it will be seen as an "acceptance" to initiate a plan that will generate the overt perception of a deception and oppression system. This will surely shock the collective into realizing where they are.

If the majority comes to awareness of this deception based system then the shift moves from the end of the age of deception and external power into the age of internal awareness and harmony.

Advanced Technology and Ancient Knowledge

The plan is to slowly release advanced technology which will undoubtedly connect with and increase the awareness of ancient knowledge. Such technology nullifies the problems that the planet is currently plagued with, although without a developed sense of the spiritual self the technology becomes destructive.

Cloning, Atmospheric Civilizations, Genetic Stability

Here is something that most people might not have contemplated yet is possibly one of the most important factors at play here.

The vehicles that house the breakaway Human society are entirely environmentally controlled. This means that the air and surfaces are clean of all the viruses, bacteria, or pathogens that are numerous on the surface of the Earth.

This creates a specific difficulty in the situation. Any contact between surface Humanity is so heavily and carefully controlled that regulations go beyond any situations we have present on Earth. If the safety and cleanliness of any of the society's vehicles were compromised this would endanger everyone "on board".

So this explains a very precarious and specific situation we currently face and this is not so different than "first-world" nations contacting indigenous tribes and similarly endangering their society. Everything that happens will be according to a carefully controlled plan.

Training, Simulations and Learning

Part of what went on in the underground base experiments were to design "tests" or "labyrinths" with various people, props, traps, and treacherous situations that were to be navigated and escaped from.

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner.

What this would do is create an individual that was capable of learning faster and faster and processing information in a more efficient manner.

This is not the same as specifically trauma-generating situations.

Cloning and Cerebrally Enhanced Soldiers – MILABS

This includes experiments, condition, trauma-based mind control, enhancement, memory suppression and operations that cover a span of space and time.

The NAZI faction has used cloning and genetic engineering to create hybrids, soldiers and everything in between. Religious/spiritual icons, to slaves, to leaders and even robotoids of sorts.

Your genetics BELONG TO YOU as the LIVING(ETERNAL) SELF. These issues are going to come more and more apparent.

If they agree to allow cloning, then who says who is allowed to clone whoever they want, Einstein, for example. And then who permits who to get DNA from history or objects in the world. These issues are going to have to be looked at because they have already been used in ways that are far beyond the basics.

Then what about consent of the masses? This is clearly being put in public because consent is required on an overall scale and individual to include people.

So what if the masses consent overall, would the people who choose to remain be at risk? Would Tribal and Native village life and people who respect the Earth be protected or permitted to stay?

Psychological Testing

These scenarios can also be generated to create specific situations that will contain tests or elements of temptation that will probe the person's psyche for weakness.

Hallucinatory Holographic Mind Probing Technology

Through the neural-interface technology the minds of many individuals can sync together to experience the same virtual simulation. Thus a single individual's mind can be selected for input and the environment will be a construct of their mental activity. The others present will be able to interact with and explore that environment like the person's mind.

Various training scenarios were utilized to enhance the resistance (or lack which) of the mind to intrusion and subversion. These were harsh but allowed the individual to completely control the energy flow in the conscious mind while remaining in harmony with the subconscious.

In a strange yet perfectly reasonable way, the conscious mind acts as the central "self" in that environment, while the subconscious mind becomes the "background" elements of objects and settings, actors and intentions within the flow of the "scene" itself. So a kind of movie begins with drawn out characteristics and perceptions that match that of a highly cinematic movie or intense dream.

The brain is naturally naive to stimulation and if allowed will liken the event to a dream in which the suspense of attention or lack of lucid self-awareness is "acceptable".

The truth may be worse than what you're thinking as far as training simulators and mind hacking devices. The truth may be that the human mind is inherently lucid and self-aware and that it is external influences that cloud the mind into what we consider 'regular' dreaming and 'regular' waking life. It may be that the conscious mind has a much greater access to information through the subconscious mind which is largely covered up and ignored in a rigidly physical and ego-focused society. As well, it may be that dreams are meant to become lucid and that we are often the subject of mystery in other realms that often refer to humans as ‘sleepwalkers’ that don’t know their dreams are real.

Virtual Simulator Training Scenarios

These scenarios can be entirely holographically generated to produce a visceral 3.5 dimensional experience which can be used for training or psychological testing.

Combat Simulators

The name is self-explanatory enough. Situations can be generated which are indistinguishable from the real events. The entire neuronal map of the brain (connectome) can be created and this is by machines to mimic the brain. After calibration through sensory feedback experiences, the sensation of the simulated realm is 100% the same as the neuronal impulses that are detected by the physical brain. An adept souled being can always tell there is a slight difference, but once the machines are calibrated the physical sensation is generally the same while in the machine.

These environments and scenarios can mimic any combat situation to determine accuracy, integrity, endurance, skill, and all areas of aptitude of an operative without the risk of injury or the lack of real-time effect.

Conditioning

As was explained in other sections, these same systems can be used to remove fear of certain scenarios. They can be used for programming the mind. They can be used to train one to complete a task over and over until they can recreate that action extremely quickly in real-time. These devices can be accelerated to operate within a fraction of the time yet to cover training and conditioning that would normally take place over hours, days, weeks, to months.

Programming and Resistance to Psychological Trauma

Through the development of specifically customized scenarios and neural feedback systems trauma can be brought to the surface or specific fears and traumas can be neutralized. This is essentially “facing fear” but in a completely generated environment that will be just as real as the real thing. One can literally face any fear, any trauma, any kind of difficulty and these systems can be utilized to accelerate a process that would normally take weeks to months or years and individuals can be strengthened to their maximal capabilities.

Accelerated Mental Functioning

Some of the technological systems were designed to test the enhanced functioning of an accelerated brain.

The mind can be seen as the energetic component of the physical brain. The physical brain projects consciousness via the electro-chemical processes.

Technology, Accelerated Learning and Repair

Technology has been developed for assisted learning and accelerated repair of tissues.

On Orgone Energy

Orgone devices produce a scattered signal which disrupts the ‘negative’ energy fields. At least this is what I’ve been told.

Apparently if built correctly it will either disrupt the negative emf/harmonics or it will produce positive harmonics, IE: Orgone energy. Orgone energy is the raw energy of life, while dead orgone, DOR, is the lack of orgone energy and can be “found” (the lack of orgone) in thunder storms and illness or depression.

This is exactly what I was told. We are generators for the same energy, and with the correct principles of energy and internal knowledge combined we can produce those very effects at will.

The energy of the mind creates a central vortex within the body. When the frequency pattern emission is consistent with fractal congruence, then that holographic printing of energy can be expanded or shrunk infinitesimally and it will retain the original fractal configuration.

This is what enables transference of awareness between fundamental harmonic layers of the Universe. When one’s consciousness energy pattern is balanced and tuned to itself so that all proportions can be shrunk through a vortex and expanded on the other side, then the personality or self-awareness remains true. This is as if the space/time must turn inside out and only when there is fractal relative polarity in time does the ‘shape’ of the consciousness within the DNA, work both ways folding and unfolding to represent the same mirror image. This is also symmetry in time as far as an ability to reverse the perception and still retain some kind of connection forward and backwards.

Energy Generators and Serums

Generators which produce vibrations that influence the cells to heal at an accelerated rate, combined with serums that enable the fuel for chemical reactions can instantaneously repair tissues.

I couldn’t believe what I was seeing at first, thinking it was a trick. There are serums and energy devices that can influence the healing and general health of the body and even mind.

These devices use energies which extend beyond the physical and are often compared to electromagnetic waves combined with ultrasonic or low to mid frequency tones. These produce harmony on a level that directly interacts with the biological functioning of the body’s cells.

I feel that this technology utilizes the same energy that is emitted from the mind and it is as if the device is “speaking” to the cells in the same kind of universal language of electromagnetic pulses, tones and vibrations.

There is information on energy devices and the use of light therapy in physical therapy today. More is being learned about the power of light and vibrations to influence the mind and body.

Genetic and Cybernetic Enhancements

Certain individuals are also genetically and technologically modified to produce a kind of hybrid human that can heal at a spontaneous rate as well as operate at a much higher rate of speed and efficiency than a regular person. Serums to induce muscle growth and mental acuity are applied, but the main condition is the enhancement of certain features through the alteration of the RNA aspect of the genes.

These features are combined with tech-assisted metabolism and certain biological functions can be enhanced or controlled via a supercomputer. These are the ‘brain-chip’ interfaces which operate to function as a governor for the mind. When the mind becomes too amped up to focus on the details the faster than thought computer system organizes the input into more usable information streams. When the mind is too slow to see the necessary patterns and variables to accomplish the goal, the brainchip operates in a self-preservation fashion and links together information streams to highlight the most viable possibilities.

Such technology is not without warning, much of this is being used heavily in the military field and this has given rise to entire factions of enhanced cybernetic soldiers. Although this is something to know, there are also many other uses for this technology and it is used by people who want to explore the technologically assisted path.

There are many other aspects of this to discuss, or at least a few important ones that will have to be understood to know the greater truth about what is happening on the planet.

Developing The Self

The issues we face are to help us learn about our selves and better mediate the known and unknown aspects of personality. The more we focus on realizing the unknown aspects of personality the less we are seemingly controlled and forced to experience the undesirable aspects of existence. The more we cling to just what we know and what we feel comfortable with the more shocked we are when we inevitably discover the truth of the whole of our existence which is *always* at least one half more than just what we know and are comfortable with. That is the nature of the universe, we can only achieve one half of the equation at any given moment. Once we complete that seemingly whole view of the universe, the information we gain access to instantly doubles and we realize that we were only looking at one portion at a time, just one half of the equation. Then the rest of the equation comes into the picture and we must expand our view once again. This is because we are viewing ourselves as we discover the universe because we are creating what there is to be discovered through the act of exploration in itself. We are the explorer, the explored, the creation and the discovery simultaneously.

Multiple Factions of Space-Fleets, Multiple Goals

Now there are what are called ‘factions’ of groups that have different agendas and interact with the public or surface level population differently. That is one of the main issues. These departments and factions are no longer limited to the surface level of existence and have access advanced technological capacity and knowledge of a larger view of time and a more complete understanding of consciousness and the human existence.

Large-Scale Catastrophes or Series of Meltdowns

I was informed that while they work continuously to prevent and reposition this civilization continually outside of the reach of any large scale meltdowns that the technology is commonly accepted to only be a temporary barrier between the inevitable.

Unsustainable

This society is unsustainable and must either change or will experience the massive catastrophic evens that are foreseen.

There are multiple perspectives that are maintained. They vary by the way of the methodology of solving the problem. All agree that change must occur and is impossible

to postpone forever. The future society is the result of our advancements and achievements. Regardless of how we get there, we must get their somehow.

Earth Changes

I was informed that these changes had been held off and would ultimately be allowed to take place to resolve the current situation. The changes can only be held off for so long, either that or the people performing this task would not be reasonably inclined to postpone these changes indefinitely.

The idea is to locate and secure the members of society who are fearless and capable of assisting in this transition and who hold key genetic elements as these individuals and groups are capable of ‘holding-space’ for the rest of civilization by literally containing the genetic precursors for all those other individuals.

I was informed that Earth changes would be the last event to take place, pretty much like icing on the cake.

This means there will be a lead up to these events.

Mass Consciousness

I was informed that the consciousness of humanity and every species is linked to the bio-etheric field of Earth. When the consciousness becomes too unstable to maintain a controlled society that this will initiate massive Earth changes. This could simply be instigated through advanced technology but I was informed that the advanced technology is actually postponing this so as to allow people to prepare, to get the word out and avoid a complete loss of the human race.

Wars and Manipulation

Wars were described as a way the power groups had controlled humanity over long periods of time. These wars were used to distract and weaken the population and the power groups believed they were holding off a full scale invasion by suppressing the rising population.

Now the population is high enough that another war would be a final war and this is extremely dangerous for the entire civilization. There would be no need for another war to repeat this cycle because the technological achievements have been met and so essentially the ‘reseting’ of the civilization would not be required or possibly acceptable.

This is why war is to be avoided however people are very easy to manipulate through lower emotional and mental programming from media and social cues. People follow the herd and the herd is essentially still operating on animalistic parameters. There is a have coating of peace and humility but this is just a wax coating that quickly evaporates as soon as the pressure of mass unrest or scarcity is realized. Many simulators have explored these possibilities.

Financial Unrest (or reset)

The idea is that financial oppression is used before wars to distract and oppress the population. When people are busy looking for resources to survive, they are not focused on the larger picture. This is part of the ‘fight or flight’ response mechanism tied to lower brain complexes which are primarily activated to keep the masses in both a highly sensitive and unstable state of being while also keeping them easy to control.

If the civil unrest reaches a level where awareness continues to rise even though financial hardship is being implemented then the latter two options will not work. At that point artificial disasters can be produced as well as natural Earth changes which initiate a shift of the ages and of consciousness.

The paradox or maybe irony here is that the oppressive actions are used to both suppress human awareness up to a certain point and simultaneously to stimulate and arouse the human spirit into wakefulness and self-empowerment. This is part of the reasoning regarding why these control systems are allowed.

Some say the elite don't want an economic collapse. That it would ruin everything. Others say it would be icing on the cake. Some say that it is a zero-point world that will save us. That the Fukushima is a big problem. Zero-point field tech, that is. Generators and energy sources, healing and terraforming. I know we project right? We see it in our own lives. So we can project with each other and grow that as a group and that part of the world will exist in that light. Which sounds Biblical, which may be because they're following prophecy. But it may also mean that we have a light, a chance, a real salvation, but that we have to grow that here and help develop it. To pass it on like love and such.

That's apparently what it comes down to. That these cycles happen every so often and we're in the midst of one right now.

Let's make it count. Every day, thought moment. With mindfulness, intention and awareness in action. Life is literally higher dimensional, so we have to become multidimensional to really face it. Indeed, that is a good perspective on action and inner being. It's like a fractal, whatever we project inside that chest center, will project into our mind and the outside. But we have to be stabilized and grounded through the root first and then to get to the heart energy must pass through the sacral/sexual centers and continue upward.

Self-Responsibility

The main reasoning is self-responsibility. We are responsible for ourselves. The people are and so if they allow themselves to be tricked and prodded like animals while behaving for the part, then that is how their experience will play out.

Clones and Synthetics

One seemingly unfair aspect of this is that there is a larger portion of non-humans than there are humans and so these beings without the human spirit will be easily programmed to support and propagate the mind-control system and the oppression. Thus, if people follow the herd mentality then they will walk directly into the end-time scenarios yet it is the humans who will actually stand to lose. People must wake up to these tricks and this technology so that they can orient themselves with the reality of this situation while learning to strengthen themselves and adapt to the manipulative aspects of this control system.

Spiritual Power and Divine Truth

This is essentially all about spiritual power and humanity stepping into the power of the truth behind their existence.

Some say that humans are designed as slaves, but if this is so, then it is the power that humans contain within that is the ‘work’ that they are sought after for. This indicates that humans have great powers that others cannot achieve on their own, even if they are creator lords.

Thus, whether humans are modified or not, they have great power. If they are not modified to perform this role of an energy generator, then this is just a confusing mind-control game that has been pulled over the eyes of humanity and humans have always contained true internal spiritual power that allows them to exist without having to parasite and steal the energy and knowledge of others. Humans are reality generators , we can choose how we want to experience the universe and what kind of collective world we want to enjoy together. Once we step into this power, there is on greater force. No amount of non-human entities can change that as we are somehow intrinsically linked to the ether-space of this realm through the spiritual and physical aspects of our being. That is what is sought after and so those who do not have this manifestation power attempt to utilize humans to create their own reality through this Earth situation. That is the basics of the ‘alien overlord’ system, whether humans in disguise, non-humans operating with alien intelligences, supercomputer systems, or spiritually degraded beings.

Most people are not ready to see themselves or the true nature of humanity. Reality is largely an illusion of the conscious mind. The hidden aspects are the other half of the equation and this is perceived by the shadow aspect of the personality and mind. We are only marginally aware unless that aspect has been unveiled and through this process the reality of the human soul can be actualized and realized.

I was informed there would be a period of two suns and multiple mirages around the world. People would believe they were peering into another dimension, floating cities, or some kind of city-ship craft.

Synthetic DNA, Synthetic Genetics

Through the previous described replicator devices synthetic DNA can be engineered. Thus, entire genetically engineered codons can be produced and form this created beings can arise. These are the hybrid beings that are performing many tasks underground and even on the surface. On the surface there are synthetic genetic create beings or genetically engineered beings that appear to be human but do not contain a human genome. There are also clones and replicants of humans. There are a variety of non-humanoid forms but these are largely kept underground. These are beings we often see in science fiction movies or shows.

Time as Mathematics, Series of Equations, a Programmable Storyline

When using the technology, time can be understood to be a series of programmable variables where one variable is centered within a series of parameters and from these possible variables within a parameter set a series of possible functions or outcomes can be process. Thus, these people can literally plug into a supercomputer system which operates based on this form of processing reality and through this they can observe all the possible realities that may connect to a present moment. This allows one to travel but this is not part of this section of the explanation.

What this allows one to do is know exactly what is going to happen and when. If a person is being interrogated, all that has to be done is the operative using the technology simply thinks of what they will say and intends to do it within 6 seconds. The machine will play out a future timescale regarding the input of that question or interrogative speech and then produce the possible results. If this doesn’t give the desired readout in the mind of the operative, they will refuse to go that route and think of another pathway. This repeats for a few seconds to a few minutes. Then the computer has found the pathway that is most like to yield a result. The operative simply steps forward, says ten words, the subject breaks immediately and tells them everything they want to know. This is obviously takes longer for most trained individuals, however that is how easily it would be to use this technology to produce an effect on a regular surface level human. They would have no defense, no way of knowing how that individual could know so much or be so direct. This is how operatives dealing with certain power groups will operate in the surface level population and individuals always feel that these people know way more than they are leading on. That is because they are, they know more about you and your life experience than you do.

Chapter 2.3 Previewing Solutions

Neutralizing Polarization

The mental-emotional tendency to pick sides, to fight, or to basically join in the fray that is currently taking place is a kind of deception of polarization which enables the leeching of emotional and mental energy from the human. This is more of a psychological unveiling where the true nature of the self, the duality game and the transcendent source aspects can be made visible to the individual. The they will see how every polarization tactic is actually taking them out of their true source of power by pushing them off center emotionally and mentally. There are teams that are capable of enabling this process and assisting as humanity realizes that half of the problem is that the unchecked mind is one’s own worst enemy.

Removing Etheric Implants

Technology can identify and isolate these foreign body implants which are etheric technological devices. These are literally frequency or phase-shifted devices which are invisible to this layer of physicality but are capable of being organized and attached to the etheric field of the human and will operate to siphon or limit energy on the etheric level. Of course, this is all technology of the vampiric created beings who parasite on the original Earth inhabitants to sustain their existence.

Clearing Astral Body

By moving through the temporal body etheric implants can be deactivated and removed and then the astral or emotional body will be able to function closer to the original levels of presence and sensitivity. This body is manipulated when one is emotionally shut down from trauma. Thus this body can be scarred and will function at a much lesser level of activity or presence until the emotional damage is healed. This contains the memories of all the emotional events of a person’s life.

There is technology that enables one to access these levels of frequency and clarify or untangle emotional chords or attachments which are used to siphon energy from the human being to the parasitic created beings. These chords are energetic in nature but also have technological components.

Knowing the True Self; Soul Actualization, Unveiling

This is a sacred, ancient process that has been outlined and discussed since the beginning of time. Humanity is here to learn, grow and express the self. In doing so, like a child moving through layers of infancy and into full-wakefulness of their body and the physical world, humanity can come into awareness of their true spiritual nature and the true spiritual nature of the universe. This is the process of soul or self-actualization that is described in some psychological methods such as Maslow's hierarchy of needs and the self-actualization that results in a complete fulfillment of all psychological needs. In this modern psychological view, self-actualization is described as the desire and process of reaching fulfillment of one's own natures and then living their life sharing that self-empowerment with other beings because after gaining your self that is the highest expression of such knowledge, sharing that power with others.

The true self is not limited to the physical body, the true self is a reflection of the entire universe and thus the primary original creative force. This is akin to the whole equation rather than just half of the equation. The visible aspect of the self, the conscious mind, is only one half of the entire equation. The whole equation consists of what we see visibly and consciously as well as what we are limited from seeing within the unconscious and subconscious mind. These are really two layers of the same ultimate source. Like an iceberg which is only partially above water, the subconscious mind is the proprietary wholeness of experience. The conscious mind is only a surface layer of reality while the unconscious mind carries all the notions, the multidimensional processes, the latent aspects of reality that we are not capable of seeing without unveiling these aspects through inner exploration.

Taking Responsibility

This entire show is about taking self-responsibility. Unless people begin to do this, there will be no hope. Right now people are convinced that their responsibilities belong to someone else even though this always results in pain and discomfort. To take responsibility is painful at first because we have a lot of work to do, but even then, this is beneficial because there is no way to remedy the pain and discomfort of being lied to and disrespected if people continue to hand their free-will over to others through blame (scapegoating), lower emotional bio-emissions (hatred, anger, etc), distractions and addictions (technological, sensual, and chemical or food based), or generally complaining without taking the first step towards providing something beneficial. This is the victim role, if humanity is to take control then they to help each other and themselves by not relying on false promises and false authorities to determine how they feel, think and live.

Returning To Center Emotionally and Mentally (spiritually)

Balanced food intake, rest, exercise, activity and learning are all continual requirements for a healthy existence. Emotional and creative expression nurtures the soul. Without these ways of being and learning the human qualities of experience diminish and creativity dwindles. There is most likely going to be a split between those who can sustain their own humanity and those who have to cause others to suffer to do so.

Vampires in the “Awake” Community

Contrary to popular belief, the majority of the non-humans are not necessarily vampiric and in power. The vampires are the people online who try to get a rise out of others for sharing their own experiences. They are the nosy neighbor that feels that the person who looks or acts differently must be considered and enemy and armed against because they feel themselves to be so much better. The majority of people are vampires, they believe that they must work for someone who will trade them money and that the amount of money they have based upon the amount of indirectly beneficial work they do puts them above others.

This is not different in the ‘awake’ community and in fact I feel that there are higher concentrations of vampires in this community, the alternate media because everything is largely a joke and a trick. People are operating on deceptive agenda schemes to divide the people and this is most intensely noticed when there is a realization of the truth behind what is happening and people begin to point fingers. The only people who are cleared to point fingers (and who also will NEVER point any fingers) are the operatives from the secret projects that have the capacity to do so because they have reached a specific level of awakening and this is related to the droning, cloning, and hybridization process. Only a specific few know who has been transformed into an impostor being and who is remaining strong and this is due being briefed on the surface level situation. Anyone can say what they want, they are simply playing into the tactics of mind-controllers and parasites and the more they play in the deeper the infection will grow. All the people pointing fingers now are actually implanted and vampiric and are attempting to drive a divide through humanity and they are the first who will turn on their friends and family when the going gets tough because they are more interested in others and starting drama than healing themselves. Ironically, this is all done under the guise of “healing” and “getting rid of the drama” AKA people they don’t trust. They pick and choose based on their own personal bias and this is a sure indication that they are emotionally and mentally out of alignment and are feeding deep seated personal fears from their own traumatic experiences.

These traumas must be healed and the harmonious center must be reattained where a person can focus on their own existence and waking up instead of acting as false-light, “king of tyranny” gatekeepers for others. Those individuals are going to be left behind if they cannot kick the habit because if they are allowed into any kind of healed society they will quickly degrade into animals and attempt to restart a devolution process in order to gain the power they so desire but cannot produce on their own. ANYONE playing ego games of trying to hurt or put blame on ANYONE else contains this viral implant and has not kicked the virus.

Heart Center and Psycho-Spiritual Fulfillment (self-actualization)

Emotional connectivity is a form of telepathy, when loved ones can tell the state of another across space and time. Whether it is family or friend. In the 60's and a little before that the US found a specific frequency akin to a radio frequency that operates on a trinary spectrum. They called this the "telepathic wave". In time, they could completely read and send thought waves without frying the brain or body through radiation. Which was what the early machines did. They actually took satellite dishes and aimed them at people. Then realized they could turn it around and do it backwards with a better effect and minimal radiation. Either way they started beaming the base and then towns around them with waves between 400 and 432 MHz and caused waves of crime sprees or peace sprees throughout the 70's and 80's. So they can digitize the telepathic wave to induce artificial telepathy. But this is organically done through an entanglement through what we call love. Literally. The Germans found it before that. And the secret orders before them. And ancient Egypt before them. As humans we remain basically through the root, heart mind connection. When the heart overpowers the mind connection of spirit, then the body is a trinary whole. When the mind guides it is binary and cannot by guided by the spirit. The spirit has direct access to the heart. The mind is a computer of sorts. The mind doesn't see past duality. The heart is the true center/eye of the spirit. The reptilian brain is the dragon and when you tame the amygdala then fear is ‘behind’ you. When the mind is "untamed" the reptilian brain takes over. This is the lower brain and amygdala fight or flight response.

This happens when the energy of the brain, body and spirit increases due to whatever various reason of spontaneous or controlled increase. Because of the "temptations" which are the easiest flow of emotional energy into a "discharge" state, is that the sacral/sexual center ties to the heart and reroutes to replace the heart and control the mind. Mental and emotional contemplation of one's true purpose and nature is the path of the divine. We find ourselves simply by looking in the right place, not in the world, but in our self energetically. We are more powerful than the programming, and we can reverse this is we're careful and devoted enough.

We have to be faster than thought, at the speed of spirit, beyond matter. When we start thinking on this level then our influences spread to others throughout the sphere. It is our destiny to know these things but they have been kept from us. The ancient orders called this the art of "implosion". Traveling realities through intense contemplation and a usage of sacred geometrical forms.

They developed and combined orders around WW2 and developed zero-point energy tech that can alter space/time fields. This is timeline manipulation. But as a rule, we go where our heart leads us. So that's how this is solved.

Go Within to Observer the Self

The answer is literally by looking within to find the great truth rather than looking around us. Everything we need to know is within us, energetically and spiritually. Our emotional, mental and physical state of balance is an indication of where and how we exist in relation to the larger overall truth. Everything that happens around us to distract us is designed to stop us from looking within. Nothing can truly stop us, nothing can actually input the holographic horrors in our spiritual center but simply perturb the brain and body. The mind is a non-holographic spiritual environment that cannot be excessed and the invader's goal is to convince people to use their imagination or their creative powers to produce these horrors on their own. This is through the programming. We must clear out the programming, the trauma, the lower-dimensional constructs and fill ourselves with the truth and spiritual presence that enables us to be self-aware in the first place. If we aren't doing this, then we are either in pain, or there is no spiritual presence to be in pain. If we are doing this, we are either clearing out the constructs by replacing them with true spiritual presence and self-awareness or we are becoming aware of them.

Organic Feedback Loop

The great feat is mentally bridging the great rift between us by looking within. Which is paradoxical.

We can never see what another is perceiving exactly, well for the most part, so we all have to kind of accept that we are communicating effectively and truly get to this level of comfort by simply being secure in knowing who we are our selves.

Technological Feedback Loop

So then knowledge, in a way, when shared, is a form of telepathy. Group telepathy.

Group think, when ideas or memes spread, is a form of group telepathy.

The Internet is like artificial spreading of telepathy. This means what is normally mentally bound signals can be initiated and spread through a digital medium.

Breaking Soul-Contracts

Ultimately everyone is under a series of “veils” which are energetic time-constraints literally like 9 dimensional layers of temporal restriction which are each bounded and ruled over by invader beings. These range from physical authorities, to astral entities, to programmable matter, to supercomputer intelligences, to spiritual false-light beings and our own shadow mind. Each layer of reality is successively guarded by these beings who are here acting as gatekeepers designed to push people back when they begin to wake up. These beings all work through fear and deception and the last layer is based on mirroring our own energies back to us like an impostor. All the levels tend to imitate, however the last layer is a near perfect imitation of our own energy to the point where people often switch places with this being before getting out of the matrix. There will be more discussed on this in a later section.

The “9 veils” were designed by impostor beings as a last effort to stop humans from escaping the false-light matrix which is essentially the holographic universe. This is literally like a maze which reproduces a new pathway and a new doorway each time a person nears the last hallway or doorway that is next to the exit. So one eventually has to project themselves into the exit using their whole being and moving past and through the tricks by not falling for them before they are placed. This is highly confusing but this is essentially a kind of labyrinth that has been created to stop people from leaving so that their soul energy can be destroyed and used as a generator for the false-light system which is an AI over mind that is programmed to take the raw soul energy and destroy the identity by inverting the electromagnetic spectrum and incorporating that inverted energy into its own internal matrix body which is literally the false-light universe.

Help Everyone

Now they may have been setting us all up for the apocalypse. They said they want us to tell on them and get the whole world aware of what they've been doing So that we can stop them and pull them from the dark which is ACTUALLY taking their souls. There are also breakaway civilizations involved in this entirely zero-point/free-energy.

Chapter 2.4 How Timelines Collapse

Repeating Cycles of Time

Because of the way time works and the ability of those using these temporal devices they have set up an elaborate labyrinth like a spider web of false realities, timelines, parallels, sub-dimensional layers, astral cages, temporal dead-ends, and time loops which act to stop people from reaching the true reality. As a result, many portions of this timeline are literally in a free-will ‘time-loop’ and these loops must be canceled out. As well, because of the nature of these loops on the overall stream, our entire civilization has been looping for the past 6 time-cycles meaning we have reached this time in history 6 times before and this is apparently the last time because we have never gotten this far before and the events of the future are having to be programmed in manually at this point.

Beings From the Future

As are result and damaging time, what happens is that the beings from the possible future that is being damaged appear to try to heal the break as well as those who are from the new possible future that is generated, this results in a time war between both possible realities for the present realm. This is what has been happening for thousands of years. All the ancient spiritual texts are describing this system. One group is a from an original possible future where everything is based on the presence of spirit and souled humans, another is from the new possible future where an AI overlord took over and removed the souls and spiritual energy from all entities on Earth rendering this a holographic prison without any other use. That reality dissolves into itself and so the time beings are constantly moving backwards and jumping timelines to try and obtain more energy.

Beings From Collapsed Timelines

The beings from the collapsed timelines are those who travel through hyperspace to get here. They then recreate themselves using cloning technology through hyperspacial means, IE: creating an alternate temporal dimension so that these devices can sustain them and literally “hold-space” for them. Then if their timeline is destroyed, they can still exist as a ‘loose-end’ in space-time. This is extremely dangerous and I will say we have some Human operatives jumping timelines searching for Earth and Humanity so that they can find a reprieve from this time war.

That is the most trying aspect of this entire ordeal for me. There are some human operatives literally in hyperspace without an opening into the temporal stream for them to reintegrate into this timeline. We must assist them, we must open the timeline by taking control of Earth from the invader races and those who wish to depress and isolate the spirited and souled human beings. We must re-organized and neutralize the invasion forces of deception and assist our humanity and maintaining their free-will, their self-awareness, their self-empowerment and their emotional, mental, and physical harmony.

Beings Who Would Never Have Existed in the First Place

As a result of this entire process, there are essentially beings who have been erased from time who degrade in this dimension and can only interact periodically. These are the created beings who operate through supercomputer forms that are placed within cloned bodies and powered by the transdimensional quantum supercomputers. These are the ‘parallel’ reality beings and this relates to the hive mind invasion. When all the original aspects of time are restored, these beings would not have ever existed in the first place as there is no quantum link for their probabilities to remain neutral and present in the original run of things.

Again, all beings who operate under the will of the universe, even when they are attempting to trick hyperspace and time, when these beings assist and help in the unification of sentient, spirited, souled beings they are capable of being healed. This is all because there is an original creative force which can produce souls that are eternal and everlasting, meaning they are not created but simply always ever were. Yet, there is a “miracle” effect where beings can be given eternal life. This is only by reflecting the will of the whole of existence, rather than a portion or a part because this is only temporary. Only the whole is forever and this takes true harmony, true balance and ultimately *self-lessness*. To give to others because you see them as worthy because they can be aware and enjoy the gift is to be like the creative force that gives life to all and enables all experience in the first place. By acting in correspondence with this we reflect the will of the whole and that power and action transfers into protection and self-empowerment for ourselves. This cannot be cheated, tricked, or rerouted as that is the karma game that has been played by delaying time and rerouting energy through getting others to take on the responsibility for others bad actions. In the end, all karma has only been delayed because when the stops are pulled out, *everyone* ends up getting their karmic returns regardless. They simply intended to build up so much karma that by the time the ties are cut, the karma overload will simply kill them instantly. This never happens and they end up in the abyss, this is the final time so since there is no recycling of the universe from the oroboros system then whatever happens is what happens forever. If everyone wants to move into harmony, to receive healing, to receive forgiveness then they must assist in the creative force in healing, forgiving, and turning away from the deception, the cruelty, the vampiric actions.

Many vampires believe they will die if they stop. The only hope is to admit their ways to the others and ensure that there are enough beings around them that will protect them so that this situation can be handled properly, with due care, compassion, and the necessary healing to ensure the safety and protection of all who seek harmony.

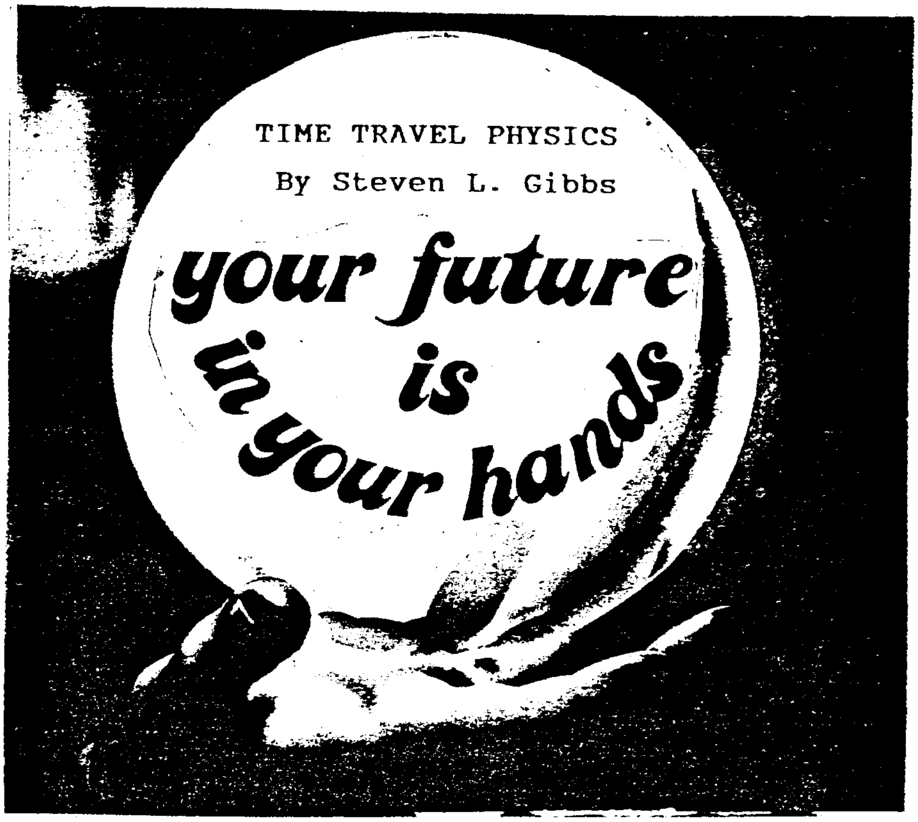

TIME TRAVEL PHYSICS

By Steven L. Gibbs

your future
is
in your hands

DEDICATION

© 1989 By: Steven L. Gibbs
Rt 1. Box 76, Clearwater
Nebraska USA
Illustrations Thanks To
Al Fry Distributor....

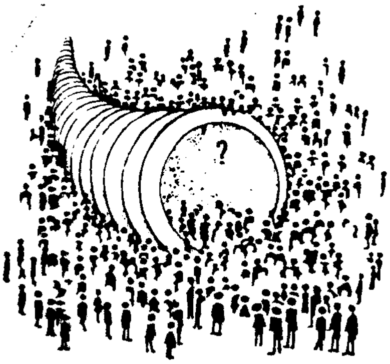

I would like to personally thank
Jesus Christ in helping me to obtain
the knowledge for writing this report.
For without his help, all of this
would not have become a reality.

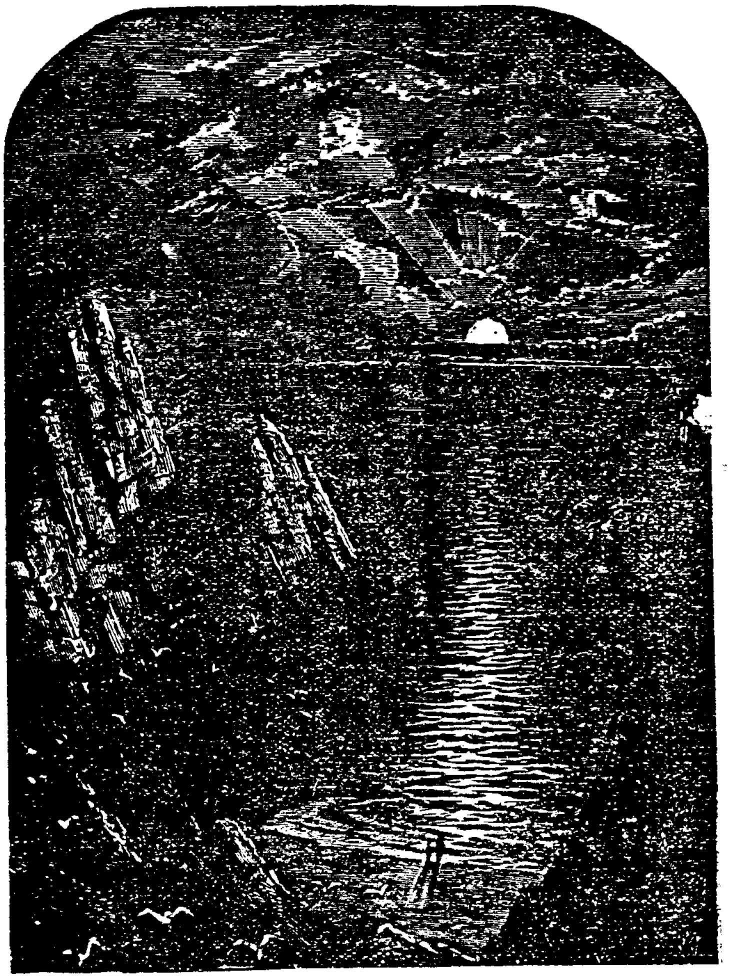

INTRODUCTION

The reason for writing this report is to show people that man does not have to use sophisticated mathematics in order to understand how Time Travel can be achieved. To assume that what is being taught in today's schools is an absolute truth, can only be a lie. For to believe in such a lie, can only lead to imperfection. But sense mankind has chosen this path, the inevitable must surely come to pass.

It is impossible to say where all of this might lead. However I do hope that the people who read this report will use the knowledge wisely. For those who don't, I leave them to their doom. But for those who do, may the Lord bless you.

Your probably wondering by now, if I am really an Alien. My answer to this is, there may be more truth to this than meets the eyes. According to the information which I have received from the Kennedy Space Center, there is a good chance that my parents came from the Star System called Phoenix. If this is case, then this would explain why I have had so much difficulty in adjusting to Earth's own environment.

Perhaps in the years to come, more will be revealed to me on the true nature of my whereabouts. And you know what they say about Aliens? Their weird!

TIME TRAVEL PHYSICS

Before I begin I must first point out, that the only reason why Formulas can be used for analyzing a specific problem, is due to the fact that a Formula helps in focusing the Belief Fields which emanate from the persons Soul. So in other words, whenever a person uses a Formula or Equation, to him this constitutes a reality. By keeping these things in mind, you shouldn't have any problem in understanding what has been written in this report.

To begin with, there are basically 3 sets of Equations which can be used for accomplishing Physical or out of the body Time Travel. I have used these formulas quite often when building my Time Travel Machines. As a matter of fact, all my research is based on these 3 formulas. The formulas which I have used for accomplishing Physical or out of the body Time Travel, are given in the following manner:

(THE GRAVITON FORMULAS)

$(X , Y = 0)$ $(X1 , Y1 = 0)$ $(X2 , Y2 = 0)$

The Graviton Formulas which I have so diligently labeled, can be used for an infinite number of things. However, before we can proceed any further, we must first have a good understanding of what each Symbol means. There are an infinite number of meanings which can be applied to each Symbol, however in order to simplify things, the meanings which I have given to each Letter or Symbol, are listed in such a way so that anybody with a little bit of knowledge can understand them.

(INTERPRETATION FOR EACH OF THE GRAVITON SYMBOLS)

X - Represents an AC Field.

Y - Represents a DC Field.

X1 - Represents a High Frequency Field.

Y1 - Represents a Low Frequency Field.

X2 - Represents a Paramagnetic Field.

Y2 - Represents a Diamagnetic Field.

Ø - Represents the Zero Vector or Twilight Zone. it may also be referred to as the (PHI Value) which is (1.618).

It is interesting to note, that whenever the (PHI Value) is used in the construction of a material object, a Time Warp usually occurs around that object.

However, in getting back to the problem at hand, whenever the (Graviton Formulas) are used singly or in series with one another, they can be used to help build a device which can actually be used for Time Travel. This is because whenever the (X) and (Y) Values are combined with one another, we can Zero in on the Twilight Zone. And sense the Twilight Zone is in attunement with God, all things become infinitely possible provided that the (X) and (Y) Values are in resonance.

So now your probably wondering, how can we use these Formulas for constructing a Time Travel Machine? Its simple. Merely start out by building a Capacitor which corresponds to this Equation: ($X2$, $Y2 = 0$). This Capacitor should be constructed in such a way, so that it resembles the drawing which is shown below:

(DRAWING OF CAPACITOR)

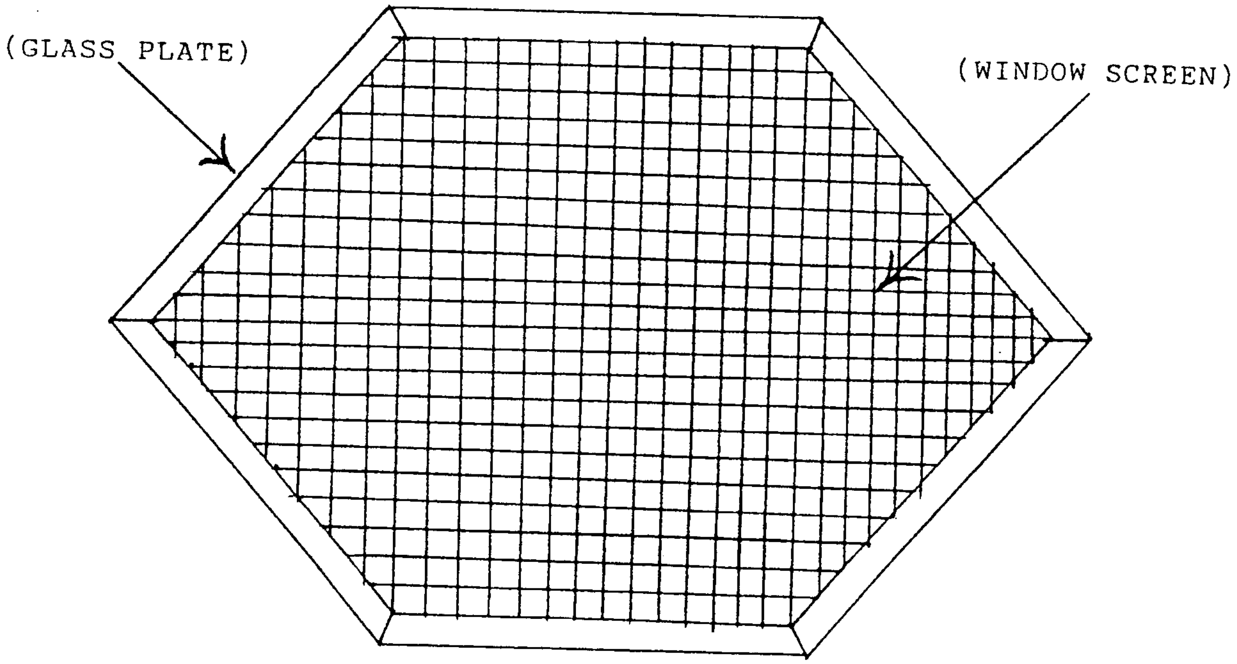

As we can see in the above illustration, it doesn't matter to much as to what size it should be, just so long as it is shaped like a Hexagon. Also each side of the Capacitor should be the same length, otherwise the Belief which is being transmitted from the plate, will be imperfect. And this could raise Hell with your Chakra Points.

At this point it should be noted, that whenever you construct a Capacitor of any type, you are automatically using the formula (X2 , Y2 = 0). And sense the (PHI Value) is used in the formula, the plate itself will develop a Life Force of its own. It should also be pointed out, that sense there is a Life Force which co-exists with the plate, it can now be programmed to do anything you want.

Now in order to program this Capacitor, all we do is stroke the Glass portion of the plate, while concentrating on the following command: (Transport me to the Date and Year). As soon as you get a stick so that you can no longer move your fingers, the Capacitor should be fully programmed. After you have finished programming the Capacitor, place the Screen Side of the plate over the Solar Plexus for the space of 30 minutes. After the treatment is over with, you then find a comfortable place to relax in order for the energies to take effect. If you have done everything correctly, your Aura or Physical Body should be transported to the date which you programmed into the Capacitor. On the average, it usually takes 70 minutes before the energies will take effect.

*Note: People who work with these plates are in serious danger of getting possessed by an Evil Spirit or Demon, unless of course you have a faith in Jesus Christ. By saying a prayer to Jesus Christ prior to any Experiment, more than protects you from the forces of Evil.

As you Experiment with these Capacitors, you will find that some of these plates work better than others. The reason for this probably ly's in the fact, that certain sizes or the materials being used, are in resonance with the person's Life Force or Belief Fields. Therefore if you run into any problems, it may pay to use a different thickness of Glass. When I talk about this, I do not mean to vary the Shape of the Capacitor, just the width. It also might pay to use Yellow tinted glass. You may also want to try Green, Blue, Indigo, and Violet colored glass. The reason for doing this is simple. Each color corresponds to a different Chakra Point. It is therefore my opinion that Yellow should be your first selection, then if you want, you can always Experiment with the higher colors.

*Note: Never under any circumstances use Red or Orange tinted glass. If you decide to use these colors, then your probably on your own, because not only will it lower the Frequency Rate of your Chakra Points, but you could also end up in an Astral Hell. And believe me, this is no fun place to visit.

Once you have found a plate which can transport your Physical or Spirit Body, there are still even further ways of increasing its effects. One of these ways is by using the second formula, (X1 , Y1 = 0). This formula states, that whenever a High Frequency Field is combined with a Low Frequency Field, we can tune ourselves into the Zero Vector or Twilight Zone. To use this formula it will first be necessary to get a hold of a

Tesla Coil and one Van de Graff Electrostatic Generator. Once these items have been obtained, you proceed by charging the plate with High Frequency Electricity and Low Frequency Static Electricity. In other words, after you have programmed the plate using the procedure on page (4), you must simultaneously Zap the plate using the above procedure. You must make sure that while you are doing this, that you do not come into contact with the plate. If you do, you could be in for a Hair raising experience. This voltage hurts! But anyhow after you have Zapped the plate for the space of 10 minutes, you then proceed by placing the Capacitor over the Solar Plexus as described on page (4). The effects that you get, should be much better than what you experienced before.

The next step to take, is to Amplify the Energies even further. To do this, we must now make use of the 3rd and final formula, which is (X , Y = Ø). This formula states, that whenever an AC Field is combined with a DC Field, a Time Warp occurs which tunes our minds into the Zero Vector or Twilight Zone. Now in order to create these fields, one should try to get a hold of a Hyper-dimensional Resonator. (Read the last 4 pages in this Report). Once you are able to get a hold of this instrument, and after everything has been plugged in correctly, while the Screen side of the plate is positioned over the Solar Plexus, position the open end of the Electromagnet over the glass side of the plate for the space of 3 minutes. After the treatment is over with, disconnect yourself from the plate (turn page) and instrument, and find yourself a comfortable place to relax. If you use this procedure in conjunction with the first two procedures, the effects that you obtain, will be increased to such a point, that you should be able to do anything you want. You must remember however, that before Physical and Spiritual Time Travel can be achieved, the (X) and (Y) factors in the formulas must be in resonance with one another. This can only be accomplished through the process of tuning your own mind just as you would with a Radionics Machine. So in other words, by selecting the correct materials to use, you are in effect creating something which you physically believe to be a reality. This same rule can be applied when working with different types of Frequency's and Voltages.

Another method which you may want to try, is that after you have programmed the plate by using the procedure which is given on page (4), transmit some High Frequency Electricity from a Tesla Coil into the Capacitor while the Screen side of the plate is pointing towards your face. After 3 or 4 minutes, de-activate the Tesla Coil, and while the Screen side of the plate is still facing towards you, take the bare ends of an Extension Cord that has been plugged into a 110 volt AC outlet, and short out the bare ends of the wires by touching both of the terminals on each side of the Screen. (See page 8). If you have programmed the plate correctly, your physical body should be transported instantaneously to the date which was programmed into the Capacitor.

THE CHRONOLOGICAL TIME REFLECTOR

(Series No.2)

2-21-89

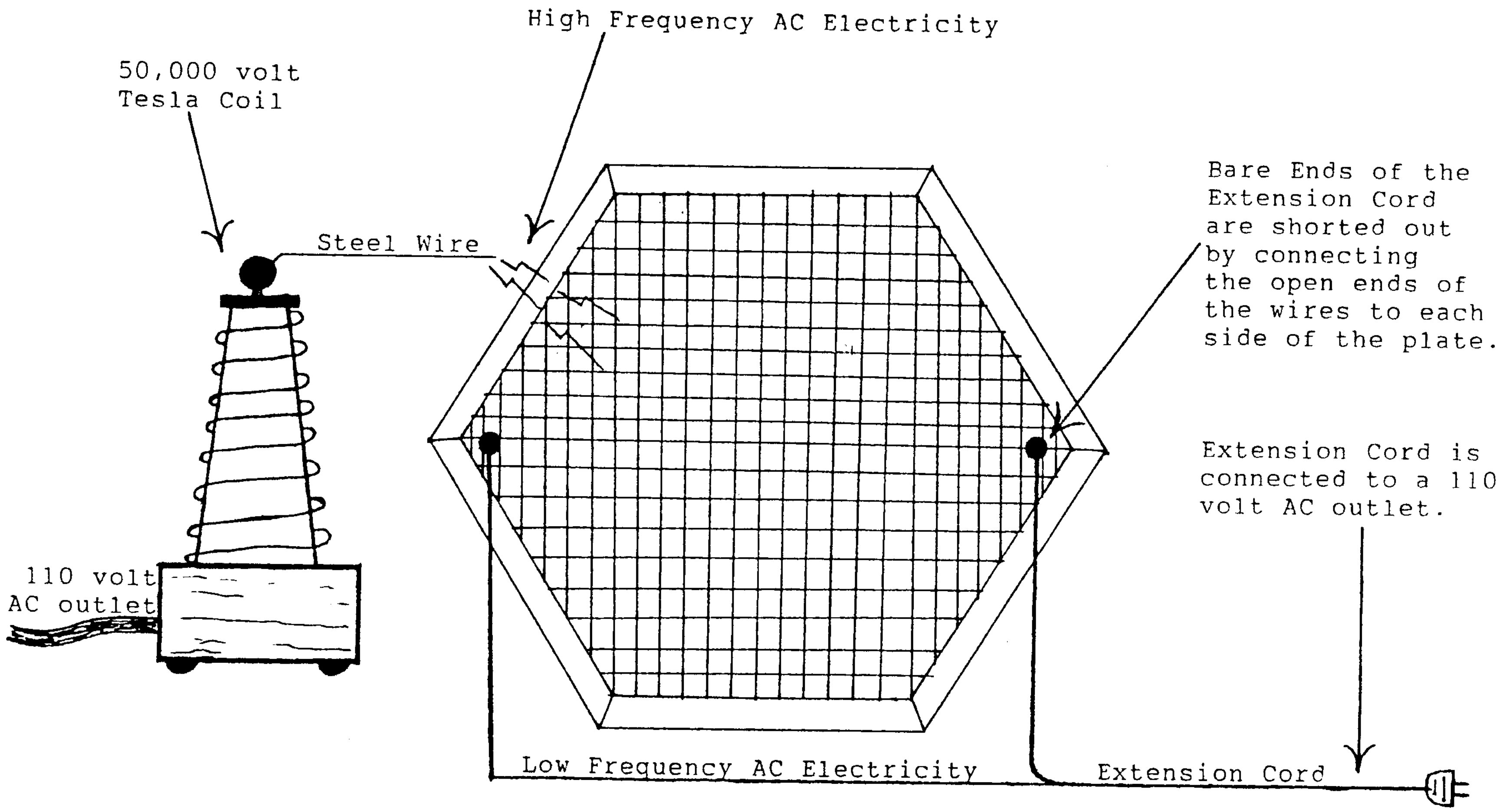
High Frequency AC Electricity
50,000 volt Tesla Coil
Steel Wire
Bare Ends of the Extension Cord are shorted out by connecting the open ends of the wires to each side of the plate.
Extension Cord is connected to a 110 volt AC outlet.
110 volt AC outlet
Low Frequency AC Electricity
Extension Cord

To give you some idea of how all this got started. Back in the year 1981 I was contacted by what I believe to be my other self. The letter which I received at the Sunset Plaza in Norfolk, Nebraska, was dated 1992 A.D. It also had a month on it which I can't seem to remember. Evidently the letter which I received indicates, that sometime around the year 1992, I will travel back into the past to make contact with myself, just like my other self did in the future. One of the things which the letter mentioned which I can barely remember, is that it said: (The path to the truth can be found in the Pyramid of Giza). It also had a riddle, and it went like this: (The riddle can be solved when 79.613 is dissolved). These two verses holds to key to Time Travel. I have already deciphered the riddle. As a matter of fact it was just last year when the answers came to me. Basically the 79.613 number can be converted into the (1.618) value which is used in my Equations. It is also the same number which was used in the construction of the Pyramid of Giza. If we decipher the riddle even further, we will find that it also reveals that the 7.8 Hz. Frequency, is the Frequency to use for accomplishing Time Travel. It seems that whenever a person travels physically through time, his Alpha or Theta Waves are vibrating at this Frequency. Once that your Brain Waves begin to oscillate at this Frequency, your mind then becomes tuned to the Zero Vector. I have found that this is the only way in which Time Travel can be achieved. So basically what I am trying to get at, the Chronological Time

Reflector causes your Brain Waves to vibrate at 7.8 Hz. cycles per second. That is all that these units do. However, for some people, they might think differently, but when you get right down to it, it all leads to the same thing. Mind verses matter. And this just barely covers the subject.

Another experience which I had that occurred around the month of Sept. in 1986, dealt with a variation of the Chronological Time Reflector. In other words, after I had finished programming the plate, I proceeded to Zap the plate with 50,000 volts of High Frequency Electricity. Sometime after the Experiment, I clipped off one end of an Extension Cord and plugged the other end into a 110 volt, 60 cycle, AC outlet. I then laid the bare ends of the wires on top of the screen. Just then I accidentally dropped something on the floor. After I had reached down to pick it up, the bare ends of the Extension Cord which I had laid on top of the Screen, touched the Screen portion of the plate. What was to follow, would be the most terrifying experience I had ever had, because as soon as the bare wires touched the Screen, a massive short occurred. All I can remember is seeing two flashes of white light energy, and the next thing I knew, a white mist had fallen over the entire room. I didn't realize that I had traveled physically through time until I looked at a Calendar the next day. Not only that, but when I asked a lady in Plainview, Nebraska as to what the date was, she said it was the 17th, but my Quartz Watch indicated the 16th. When

I had returned home later on that day, I has discovered that all of my Experiments had been dated wrong. Then I finally realized that I had actually traveled one day into the past. To this day I am not exactly sure as to whether or not I returned back to my own present Time Line. Perhaps I will never know.

Another Time Travel instrument which was sent to me by a man who lives in Osceola, PA., is revealed on pages (12), (13), and (14). According to what I have red in his letter, this device can actually transport objects into the future. However, he says that there is somewhat of a Time delay effect shortly after the unit has been shut off. In other words, after the unit has been de-activated, it takes anywhere from 10 to 15 minutes before the object is teleported. I have never built this device as yet, however I plan to eventually.

POWER SUPPLY FOR THE TIME PLATFORM

2-22-89

C1 - Is a homemade Capacitor made out of sheets of Glass and Copper Foil. Aluminum may be used, but this could be dangerous to your health.

L1 - Is number 10 Gage wire wound around a dowel rod.

L2 - Is number 24 Gage Magnet Wire wrapped around some PVC pipe. There should be about 1000 turns which should create around 100,000 volts of Low Amperage High Frequency Electricity.

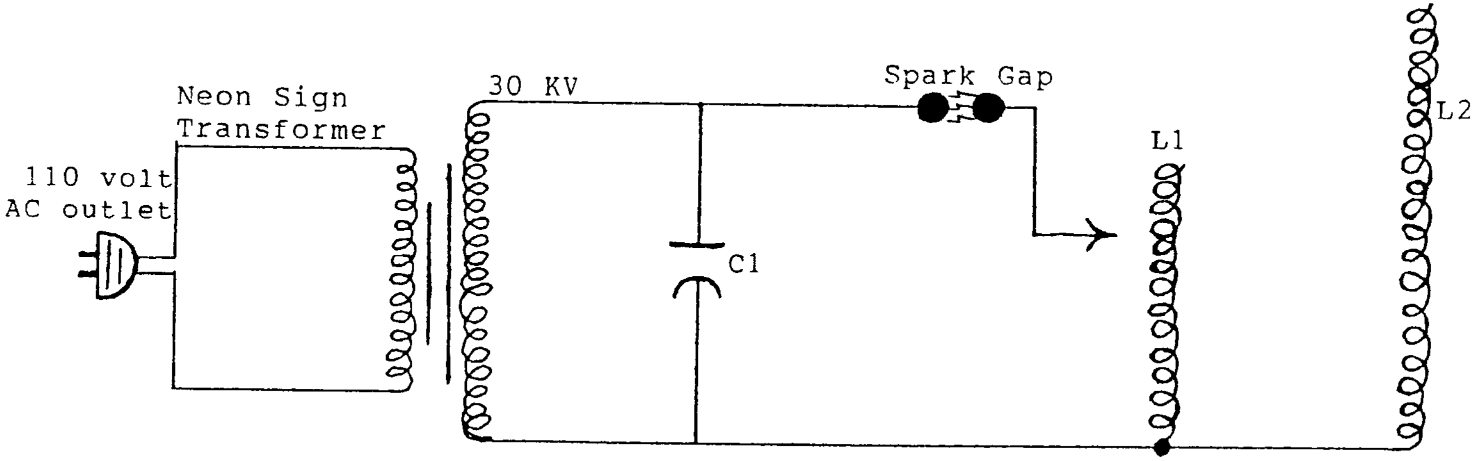
Neon Sign
Transformer
110 volt
AC outlet
30 KV
C1
Spark Gap
L1
L2

THE TIME PLATFORM

2-22-89

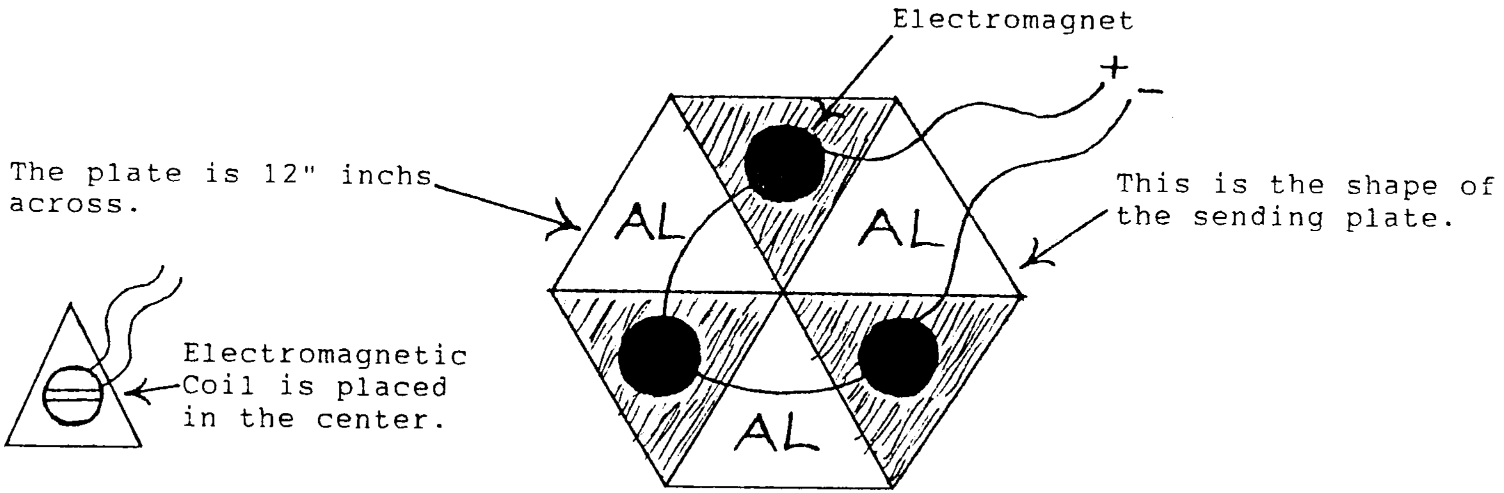
Electromagnet
+
-
The plate is 12" inchs across.
AL
AL
AL
This is the shape of the sending plate.
Electromagnetic Coil is placed in the center.

THE ELECTROMAGNETS ARE MADE OF STRAP IRON STACKED 8 INCHES HIGH (1/4"STRAP IRON).

THE ALUMINUM PLATES ARE 1/4 OF AN INCH THICK, AND ARE STACKED 8 INCHS HIGH .

THERE ARE VARIOUS REASONS FOR THIS PARTICULAR ARRAY.

THE ALUMINUM SPACERS CONCENTRATE AND FOCUS THE MAGNETIC FIELDS.

FULL SCALE DRAWING OF THE TIME TRAVEL UNIT

2-23-89

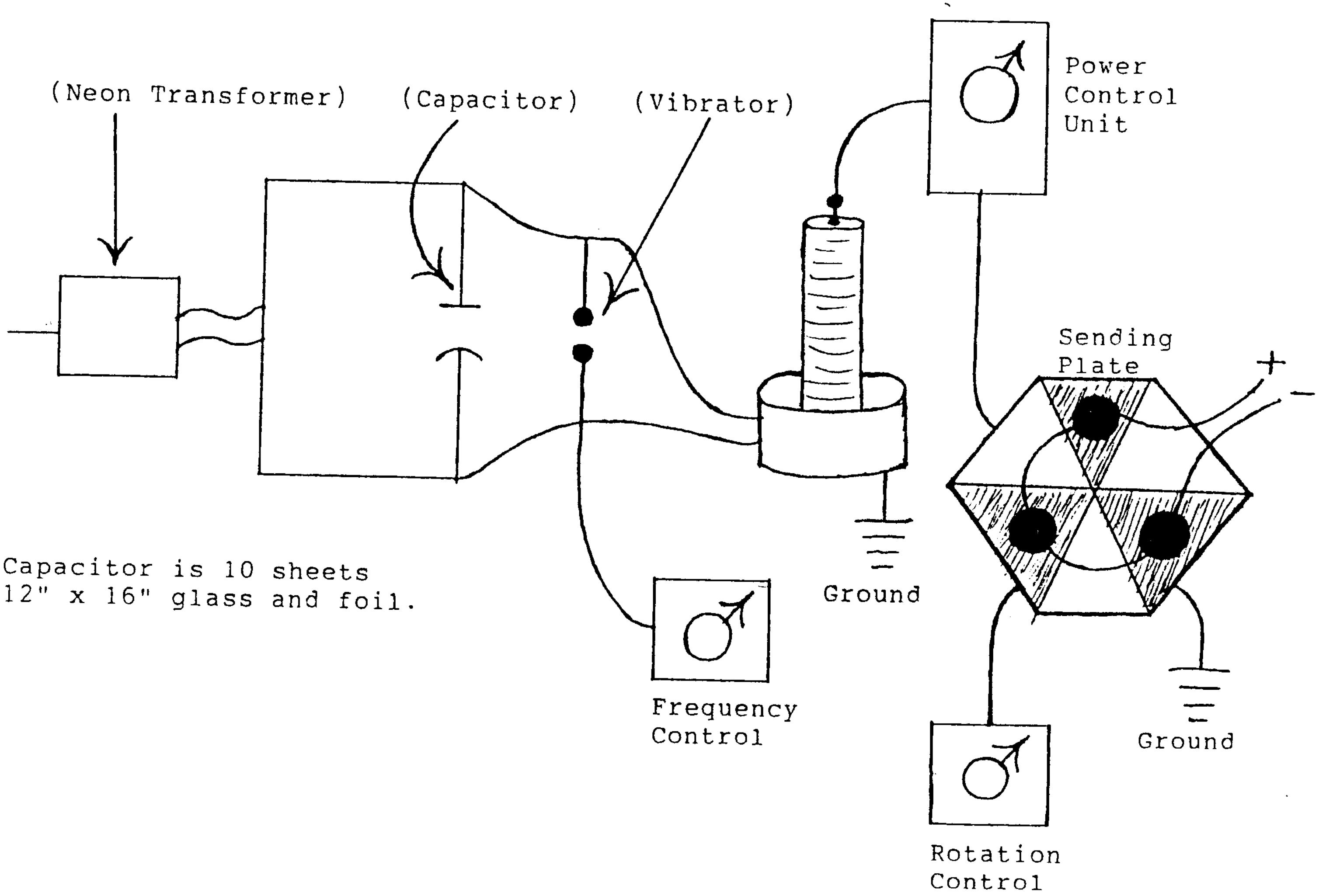
(Neon Transformer)
(Capacitor)
(Vibrator)
Power
Control
Unit
Sending
Plate
+
-
Ground
Frequency
Control
Ground
Rotation
Control

Capacitor is 10 sheets
12" x 16" glass and foil.

THE DELAWARR EXPERIMENT IN TIME TRAVEL (PHASE #1)

8-16-89

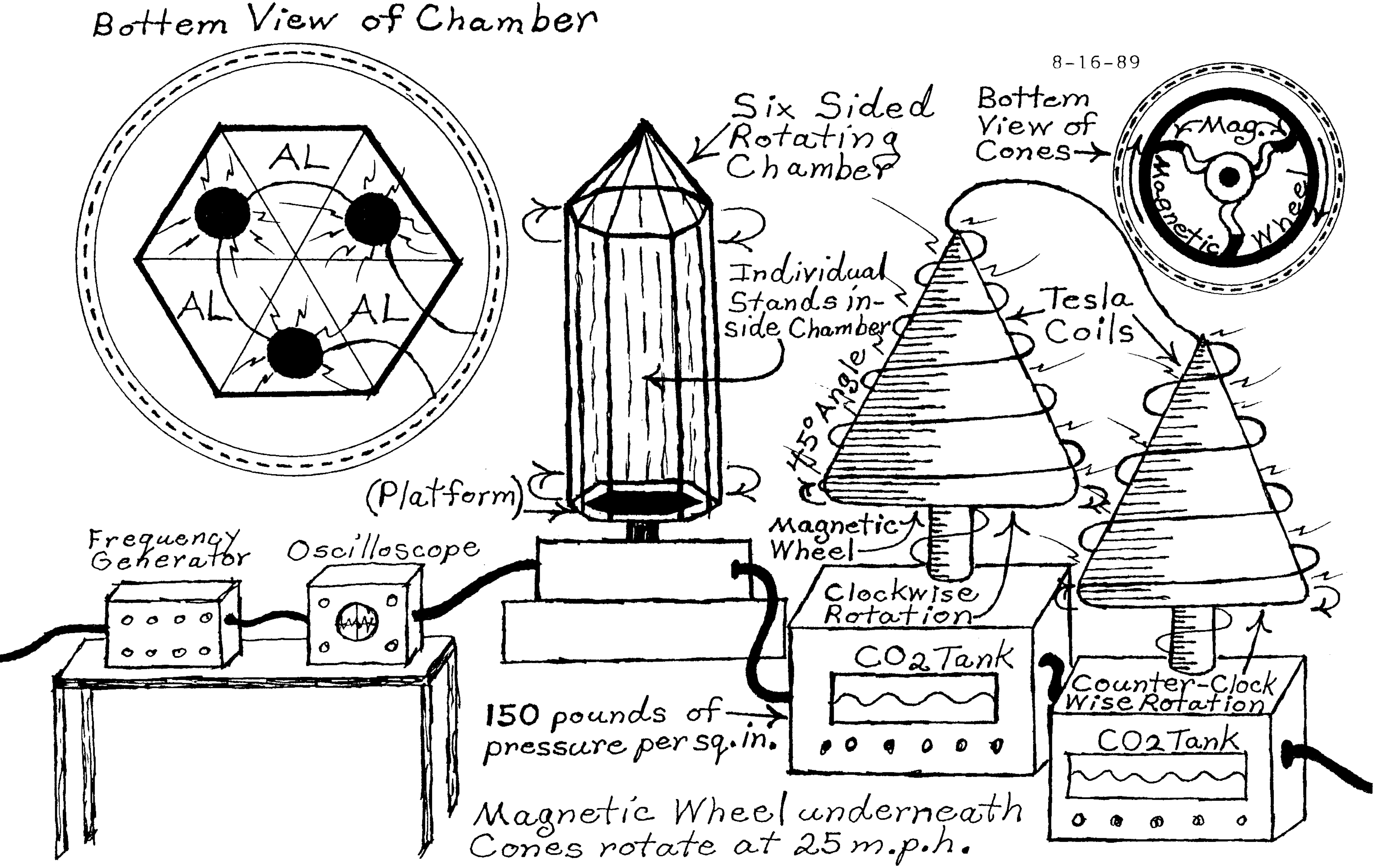
Bottem View of Chamber
AL
AL
AL
Six Sided Rotating Chamber
Individual Stands inside Chamber
(Platform)
Bottem View of Cones
Mag.
Magnetic Wheel
Tesla Coils
45° Angle
Magnetic Wheel
Clockwise Rotation
Counter-Clock Wise Rotation
CO2 Tank
CO2 Tank
Frequency Generator
Oscilloscope
150 pounds of pressure per sq. in.
Magnetic Wheel underneath Cones rotate at 25 m.p.h.

Throughout the vast regions of Space Time itself, there are an infinite number of ways to build a Time Travel Machine. One of these ways is to get a hold of a Rubbing Plate, (GSR), or Pendulum, and proceed by asking (Yes) or (No) questions as to what type of components one should use in order to build a Time Travel Machine. There are a number of ways of doing this. First of all begin by asking what size the instrument box should be for holding the components. Next, determine what type of energy one should use for achieving these effects. Then you ask what type of components one should use in order to transmit this energy. If you have done everything correctly, you will end up creating a device which is based entirely on belief. And sense you have created a device which is built entirely on belief, you can be dam certain that it will work! Especially if you use a (GSR) while asking your questions.

For those of you who do not know what a (GSR) is, this is basically a (Galvanic Skin Response Meter). One of these devices may be purchased from me for $150. Instructions are included. Please allow 4 to 6 weeks for delivery.

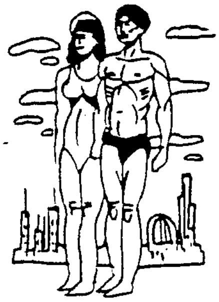

THE PARADOXES OF TIME TRAVEL

David Lewis

TIME travel, I maintain, is possible. The paradoxes of time travel are oddities, not impossibilities. They prove only this much, which few would have doubted: that a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours.

I shall be concerned here with the sort of time travel that is recounted in science fiction. Not all science fiction writers are clear-headed, to be sure, and inconsistent time travel stories have often been written. But some writers have thought the problems through with great care, and their stories are perfectly consistent.[1]

If I can defend the consistency of some science fiction stories of time travel, then I suppose parallel defenses might be given of some controversial physical hypotheses, such as the hypothesis that time is circular or the hypothesis that there are particles that travel faster than light. But I shall not explore these parallels here.

What is time travel? Inevitably, it involves a discrepancy between time and time. Any traveler departs and then arrives at his destination; the time elapsed from departure to arrival (positive, or perhaps zero) is the duration of the journey. But if he is a time traveler, the separation in time between departure and arrival does not equal the duration of his journey. He departs; he travels for an hour, let us say; then he arrives. The time he reaches is not the time one hour after his departure. It is later, if he has traveled toward the future; earlier, if he has traveled toward the past. If he has traveled far toward the past, it is earlier even than his departure. How can it be that the same two events, his departure and his arrival, are separated by two unequal amounts of time?

It is tempting to reply that there must be two independent time dimensions; that for time travel to be possible, time must be not a line but a plane.[2] Then a pair of events may have two unequal separations if they are separated more in one of the time dimensions than in the other. The lives of common people occupy straight diagonal lines across the plane of time, sloping at a rate of exactly one hour of time$_1$ per hour of time$_2$. The life of the time traveler occupies a bent path, of varying slope.

On closer inspection, however, this account seems not to give us time travel as we know it from the stories. When the traveler revisits the days of his childhood, will his playmates be there to meet him? No; he has not reached the part of the plane of time where they are. He is no longer separated from them along one of the two dimensions of time, but he is still separated from them along the other. I do not say that two-dimensional time is impossible, or that there is no way to square it with the usual conception of what time travel would be like. Nevertheless I shall say no more about two-dimensional time. Let us set it aside, and see how time travel is possible even in one-dimensional time.

The world—the time traveler's world, or ours—is a four-dimensional manifold of events. Time is one dimension of the four, like the spatial dimensions except that the prevailing laws of nature discriminate between time and the others—or rather, perhaps, between various timelike dimensions and various spacelike dimensions. (Time remains one-dimensional, since no two timelike dimensions are orthogonal.) Enduring things are timelike streaks: wholes composed of temporal parts, or *stages*, located at various times and places. Change is qualitative difference between different stages—different temporal parts—of some enduring thing, just as a "change" in scenery from east to west is a qualitative difference between the eastern and western spatial parts of the landscape. If this paper should change your mind about the

From *American Philosophical Quarterly*, April 1976, pp. 145-152. © 1976 by North American Philosophical Publications. Reprinted by permission.

possibility of time travel, there will be a difference of opinion between two different temporal parts of you, the stage that started reading and the subsequent stage that finishes.

If change is qualitative difference between temporal parts of something, then what doesn't have temporal parts can't change. For instance, numbers can't change; nor can the events of any moment of time, since they cannot be subdivided into dissimilar temporal parts. (We have set aside the case of two-dimensional time, and hence the possibility that an event might be momentary along one time dimension but divisible along the other.) It is essential to distinguish change from "Cambridge change," which can befall anything. Even a number can "change" from being to not being the rate of exchange between pounds and dollars. Even a momentary event can "change" from being a year ago to being a year and a day ago, or from being forgotten to being remembered. But these are not genuine changes. Not just any old reversal in truth value of a time-sensitive sentence about something makes a change in the thing itself.

A time traveler, like anyone else, is a streak through the manifold of space-time, a whole composed of stages located at various times and places. But he is not a streak like other streaks. If he travels toward the past he is a zig-zag streak, doubling back on himself. If he travels toward the future, he is a stretched-out streak. And if he travels either way instantaneously, so that there are no intermediate stages between the stage that departs and the stage that arrives and his journey has zero duration, then he is a broken streak.

I asked how it could be that the same two events were separated by two unequal amounts of time, and I set aside the reply that time might have two independent dimensions. Instead I reply by distinguishing time itself, *external time* as I shall also call it, from the *personal time* of a particular time traveler: roughly, that which is measured by his wristwatch. His journey takes an hour of his personal time, let us say; his wristwatch reads an hour later at arrival than at departure. But the arrival is more than an hour after the departure in external time, if he travels toward the future; or the arrival is before the departure in external time (or less than an hour after), if he travels toward the past.

That is only rough. I do not wish to define personal time operationally, making wristwatches infallible by definition. That which is measured by my own wristwatch often disagrees with external time, yet I am no time traveler; what my misregulated wristwatch measures is neither time itself nor my personal time. Instead of an operational definition, we need a functional definition of personal time; it is that which occupies a certain role in the pattern of events that comprise the time traveler's life. If you take the stages of a common person, they manifest certain regularities with respect to external time. Properties change continuously as you go along, for the most part, and in familiar ways. First come infantile stages. Last come senile ones. Memories accumulate. Food digests. Hair grows. Wristwatch hands move. If you take the stages of a time traveler instead, they do not manifest the common regularities with respect to external time. But there is one way to assign coordinates to the time traveler's stages, and one way only (apart from the arbitrary choice of a zero point), so that the regularities that hold with respect to this assignment match those that commonly hold with respect to external time. With respect to the correct assignment properties change continuously as you go along, for the most part, and in familiar ways. First come infantile stages. Last come senile ones. Memories accumulate. Food digests. Hair grows. Wristwatch hands move. The assignment of coordinates that yields this match is the time traveler's personal time. It isn't really time, but it plays the role in his life that time plays in the life of a common person. It's enough like time so that we can—with due caution—transplant our temporal vocabulary to it in discussing his affairs. We can say without contradiction, as the time traveler prepares to set out, "Soon he will be in the past." We mean that a stage of him is slightly later in his personal time, but much earlier in external time, than the stage of him that is present as we say the sentence.

We may assign locations in the time traveler's personal time not only to his stages themselves but also to the events that go on around him. Soon Caesar will die, long ago; that is, a stage slightly later in the time traveler's personal time than his present stage, but long ago in external time, is simultaneous with Caesar's death. We could even extend the assignment of personal time to events that are not part of the time traveler's life, and not simultaneous with any of his stages. If his funeral in ancient Egypt is separated from his death by three days of external time and his death is separated from his birth by three score years and ten of his personal time, then we may add the two intervals and say that his funeral follows his birth by three score years and ten and three days of *extended personal time*. Likewise a bystander might truly say, three years after the last departure of another famous time traveler, that "he may even now—if I may use the phrase—be wandering on some plesiosaurus-haunted oolitic coral reef, or beside the lonely saline seas of the Triassic Age."[3] If the time traveler does wander on an oolitic coral reef three years after his departure in his personal time, then it is no mistake to say with respect to his extended personal time that the wandering is taking place "even now".

We may liken intervals of external time to distances as the crow flies, and intervals of personal time to distances along a winding path. The time traveler's life is like a mountain railway. The place two miles due east of here may also be nine miles down the line, in the westbound direction. Clearly we are not dealing here with two independent dimensions. Just as distance along the railway is not a fourth spatial dimension, so a time traveler's personal time is not a second dimension of time. How far down the line some place is depends on its location in three-dimensional space, and likewise the locations of events in personal time depend on their locations in one-dimensional external time.

Five miles down the line from here is a place where the line goes under a trestle; two miles further is a place where the line goes over a trestle; these places are one and the same. The trestle by which the line crosses over itself has two different locations along the line, five miles down from here and also seven. In the same way, an event in a time traveler's life may have more than one location in his personal time. If he doubles back toward the past, but not too far, he may be able to talk to himself. The conversation involves two of his stages, separated in his personal time but simultaneous in external time. The location of the conversation in personal time should be the location of the stage involved in it. But there are two such stages; to share the locations of both, the conversation must be assigned two different locations in personal time.

The more we extend the assignment of personal time outwards from the time traveler's stages to the surrounding events, the more will such events acquire multiple locations. It may happen also, as we have already seen, that events that are not simultaneous in external time will be assigned the same location in personal time—or rather, that at least one of the locations of one will be the same as at least one of the locations of the other. So extension must not be carried too far, lest the location of events in extended personal time lose its utility as a means of keeping track of their roles in the time traveler's history.

A time traveler who talks to himself, on the telephone perhaps, looks for all the world like two different people talking to each other. It isn't quite right to say that the whole of him is in two places at once, since neither of the two stages involved in the conversation is the whole of him, or even the whole of the part of him that is located at the (external) time of the conversation. What's true is that he, unlike the rest of us, has two different complete stages located at the same time at different places. What reason have I, then, to regard him as one person and not two? What unites his stages, including the simultaneous ones, into a single person? The problem of personal identity is especially acute if he is the sort of time traveler whose journeys are instantaneous, a broken streak consisting of several unconnected segments. Then the natural way to regard him as more than one person is to take each segment as a different person. No one of them is a time traveler, and the peculiarity of the situation comes to this: all but one of these several people vanish into thin air, all but another one appear out of thin air, and there are remarkable resemblances between one at his appearance and another at his vanishing. Why isn't that at least as good a description as the one I gave, on which the several segments are all parts of one time traveler?

I answer that what unites the stages (or segments) of a time traveler is the same sort of mental, or mostly mental, continuity and connectedness that unites anyone else. The only difference is that whereas a common person is connected and continuous with respect to external time, the time traveler is connected and continuous only with respect to his own personal time. Taking the stages in order, mental (and bodily) change is mostly gradual rather than sudden, and at no point is there sudden change in too many different respects all at once. (We can include position in external time among the respects we keep track of, if we like. It may change discontinuously with respect to personal time if not too much else changes discontinuously along with it.) Moreover, there is not too much change altogether. Plenty of traits and traces last a lifetime. Finally, the connectedness and the continuity are not accidental. They are explicable; and further, they are explained by the fact that the properties of each stage depend causally on those of the stages just before in personal time, the dependence being such as tends to keep things the same.[4]

To see the purpose of my final requirement of causal continuity, let us see how it excludes a case of counterfeit time travel. Fred was created out of thin air, as if in the midst of life; he lived a while, then died. He was created by a demon, and the demon had chosen at random what Fred was to be like at the moment of his creation. Much later someone else, Sam, came to resemble Fred as he was when first created. At the very moment when the resemblance became perfect, the demon destroyed Sam. Fred and Sam together are very much like a single person: a time traveler whose personal time starts at Sam's birth, goes on to Sam's destruction and Fred's creation, and goes on from there to Fred's death. Taken in this order, the stages of Fred-*cum*-Sam have the proper connectedness and continuity. But they lack causal continuity, so Fred-*cum*-Sam is not one person and not a time traveler. Perhaps it was pure coincidence that Fred at his creation and Sam at his destruction were exactly alike; then the connectedness and continuity of Fred-*cum*-Sam across the crucial point are accidental. Perhaps instead the demon remembered what Fred was like, guided Sam toward perfect resemblance, watched his progress, and destroyed him at the right moment. Then the connectedness and continuity of Fred-*cum*-Sam has a causal explanation, but of the wrong sort. Either way, Fred's first stages do not depend causally for their properties on Sam's last stages. So the case of Fred and Sam is rightly disqualified as a case of personal identity and as a case of time travel.

We might expect that when a time traveler visits the past there will be reversals of causation. You may punch his face before he leaves, causing his eye to blacken centuries ago. Indeed, travel into the past necessarily involves reversed causation. For time travel requires personal identity—he who arrives must be the same person who departed. That requires causal continuity, in which causation runs from earlier to later stages in the order of personal time. But the orders of personal and external time disagree at some point, and there we have causation that runs from later to earlier stages in the order of external time. Elsewhere I have given an analysis of causation in terms of chains of counterfactual dependence, and I took care that my analysis would not rule out casual reversal *a priori*.[5] I think I can argue (but not here) that under my analysis the direction of counterfactual dependence and causation is governed by the direction of other *de facto* asymmetries of time. If so, then reversed causation and time travel are not excluded altogether, but can occur only where there are local exceptions to these asymmetries. As I said at the outset, the time traveler's world would be a most strange one.

Stranger still, if there are local—but only local—causal reversals, then there may also be causal loops: closed causal chains in which some of the causal links are normal in direction and others are reversed. (Perhaps there must be loops if there is reversal: I am not sure.) Each event on the loop has a causal explanation, being caused by events elsewhere on the loop. That is not to say that the loop as a whole is caused or explicable. It may not be. Its inexplicability is especially remarkable if it is made up of the sort of causal processes that transmit information. Recall the time traveler who talked to himself. He talked to himself about time travel, and in the course of the conversation his older self told his younger self how to build a time machine. That information was available in no other way. His older self knew how because his younger self had been told and the information had been preserved by the causal processes that constitute recording, storage, and retrieval of memory traces. His younger self knew, after the conversation, because his older self had known and the information had been preserved by the causal processes that constitute telling. But where did the information come from in the first place? Why did the whole affair happen? There is simply no answer. The parts of the loop are explicable, the whole of it is not. Strange! But not impossible, and not too different from inexplicabilities we are already inured to. Almost everyone agrees that God, or the Big Bang, or the entire infinite past of the universe, or the decay of a tritium atom, is uncaused and inexplicable. Then if these are possible, why not also the inexplicable causal loops that arise in the time travel?

I have committed a circularity in order not to talk about too much at once, and this is a good place to set it right. In explaining personal time, I presupposed that we were entitled to regard certain stages as comprising a single person. Then in explaining what united the stages into a single person, I presupposed that we were given a personal time order for them. The proper way to proceed is to define personhood and personal time simultaneously, as follows. Suppose given a pair of an aggregate of persona-stages, regarded as a candidate for personhood, and an assignment of coordinates to those stages, regarded as a candidate for his personal time. If the stages satisfy the conditions given in my circular explanation with respect to the assignment of coordinates, then both candidates succeed: the stages do comprise a person and the assignment is his personal time.

I have argued so far that what goes on in a time travel story may be a possible pattern of events in four-dimensional space-time with no extra time dimension; that it may be correct to regard the scattered stages of the alleged time traveler as comprising a single person; and that we may legitimately assign to those stages and their surroundings a personal time order that disagrees sometimes with their order in external time. Some might concede all this, but protest that the impossibility of time travel is revealed after all when we ask not what the time traveler *does*, but what he *could do*. Could a time traveler change the past? It seems not: the events of a past moment could no more change than numbers could. Yet it seems that he would be as able as anyone to do things that would change the past if he did them. If a time traveler visiting the past both could and couldn't do something that would change it, then there cannot possibly be such a time traveler.

Consider Tim. He detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine. Tim would like nothing so much as to kill Grandfather, but alas he is too late. Grandfather died in his bed in 1957, while Tim was a young boy. But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all. He buys a rifle; he spends long hours in target practice; he shadows Grandfather to learn the route of his daily walk to the munitions works; he rents a room along the route; and there he lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer,. . . .

Tim can kill Grandfather. He has what it takes. Conditions are perfect in every way: the best rifle money could buy, Grandfather an easy target only twenty yards away, not a breeze, door securely locked against intruders. Tim a good shot to begin with and now at the peak of training, and so on. What's to stop him? The forces of logic will not stay his hand! No powerful chaperone stands by to defend the past from interference. (To imagine such a chaperone, as some authors do, is a boring evasion, not needed to make Tim's story consistent.) In short, Tim is as much able to kill Grandfather as anyone ever is to kill anyone. Suppose that down the street another sniper, Tom, lurks waiting for another victim, Grandfather's partner. Tom is not a time traveler, but otherwise he is just like Tim: same make of rifle, same murderous intent, same everything. We can even suppose that Tom, like Tim, believes himself to be a time traveler. Someone has gone to a lot of trouble to deceive Tom into thinking so. There's no doubt that Tom can kill his victim; and Tim has everything going for him that Tom does. By any ordinary standards of ability, Tim can kill Grandfather.

Tim cannot kill Grandfather. Grandfather lived, so to kill him would be to change the past. But the events of a past moment are not subdivisible into temporal parts and therefore cannot change. Either the events of 1921 timelessly do include Tim's killing of Grandfather, or else they timelessly don't. We may be tempted to speak of the "original" 1921 that lies in Tim's personal past, many years before his birth, in which Grandfather lived; and of the "new" 1921 in which Tim now finds himself waiting in ambush to kill Grandfather. But if we do speak so, we merely confer two names on one thing. The events of 1921 are doubly located in Tim's (extended) personal time, like the trestle on the railway, but the "original" 1921 and the "new" 1921 are one and the same. If Tim did not kill Grandfather in the "original" 1921, then if he does kill Grandfather in the "new" 1921, he must both kill and not kill Grandfather in 1921—in the one and only 1921, which is both the "new" and the "original" 1921. It is logically impossible that Tim should change the past by killing Grandfather in 1921. So Tim cannot kill Grandfather.

Not that past moments are special; no more can anyone change the present or the future. Present and future momentary events no more have temporal parts than past ones do. You cannot change a present or future event from what it was originally to what it is after you change it. What you *can* do is to change the present or the future from the unactualized way they would have been without some action of yours to the way they actually are. But that is not an actual change: not a difference between two successive actualities. And Tim can certainly do as much; he changes the past from the unactualized way it would have been without him to the one and only way it actually is. To "change" the past in this way, Tim need not do anything momentous; it is enough just to be there, however unobtrusively.

You know, of course, roughly how the story of Tim must go on if it is to be consistent: he somehow fails. Since Tim didn't kill Grandfather in the "original" 1921, consistency demands that neither does he kill Grandfather in the "new" 1921. Why not? For some commonplace reason. Perhaps some noise distracts him at the last moment, perhaps he misses despite all his target practice, perhaps his nerve fails, perhaps he even feels a pang of unaccustomed mercy. His failure by no means proves that he was not really able to kill Grandfather. We often try and fail to do what we are able to do. Success at some tasks requires not only ability but also luck, and lack of luck is not a temporary lack of ability. Suppose our other sniper, Tom, fails to kill Grandfather's partner for the same reason, whatever it is, that Tim fails to kill Grandfather. It does not follow that Tom was unable to. No more does it follow in Tim's case that he was unable to do what he did not succeed in doing.

We have this seeming contradiction: "*Tim doesn't, but can, because he has what it takes*" versus "*Tim doesn't, and can't, because it's logically impossible to change the past.*" I reply that there is no contradiction. Both conclusions are true, and for the reasons given. They are compatible because "can" is equivocal.

To say that something can happen means that its happening is compossible with certain facts. *Which* facts? That is determined, but sometimes not determined well enough, by context. An ape can't speak a human language—say, Finnish—but I can. Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish. The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish. But don't take me along to Helsinki as your interpreter: I can't speak Finnish. My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training. What I can do, relative to one set of facts, I cannot do, relative to another, more inclusive, set. Whenever the context leaves it open which facts are to count as relevant, it is possible to equivocate about whether I can speak Finnish. It is likewise possible to equivocate about whether it is possible for me to speak Finnish, or whether I am able to, or whether I have the ability or capacity or power or potentiality to. Our many words for much the same thing are little help since they do not seem to correspond to different fixed delineations of the relevant facts.

Tim's killing Grandfather that day in 1921 is compossible with a fairly rich set of facts: the facts about his rifle, his skill and training, the unobstructed line of fire, the locked door and the absence of any chaperone to defend the past, and so on. Indeed it is compossible with all the facts of the sorts we would ordinarily count as relevant is saying what someone can do. It is compossible with all the facts corresponding to those we deem relevant in Tom's case. Relative to these facts, Tim can kill Grandfather. But his killing Grandfather is not compossible with another, more inclusive set of facts. There is the simple fact that Grandfather was not killed. Also there are various other facts about Grandfather's doings after 1921 and their effects: Grandfather begat Father in 1922 and Father begat Tim in 1949. Relative to these facts, Tim cannot kill Grandfather. He can and he can't, but under different delineations of the relevant facts. You can reasonably choose the narrower delineation, and say that he can; or the wider delineation, and say that he can't. But choose. What you mustn't do is waver, say in the same breath that he both can and can't, and then claim that this contradiction proves that time travel is impossible.

Exactly the same goes for Tom's parallel failure. For Tom to kill Grandfather's partner also is compossible with all facts of the sorts we ordinarily count as relevant, but not compossible with a larger set including, for instance, the fact that the intended victim lived until 1934. In Tom's case we are not puzzled. We say without hesitation that he can do it, because we see at once that the facts that are not compossible with his success are facts about the future of the time in question and therefore not the sort of facts we count as relevant in saying what Tom can do.

In Tim's case it is harder to keep track of which facts are relevant. We are accustomed to exclude facts about the future of the time in question, but to include some facts about its past. Our standards do not apply unequivocally to the crucial facts in this special case: Tim's failure, Grandfather's survival, and his subsequent doings. If we have foremost in mind that they lie in the external future of that moment in 1921 when Tim is almost ready to shoot, then we exclude them just as we exclude the parallel facts in Tom's case. But if we have foremost in mind that they precede that moment in Tim's extended personal time, then we tend to include them. To make the latter be foremost in your mind, I chose to tell Tim's story in the order of his personal time, rather than in the order of external time. The fact of Grandfather's survival until 1957 had already been told before I got to the part of the story about Tim lurking in ambush to kill him in 1921. We must decide, if we can, whether to treat these personally past and externally future facts as if they were straightforwardly past or as if they were straightforwardly future.

Fatalists—the best of them—are philosophers who take facts we count as irrelevant in saying what someone can do, disguise them somehow as facts of a different sort that we count as relevant, and thereby argue that we can do less than we think—indeed, that there is nothing at all that we don't do but can. I am not going to vote Republican next fall. The fatalist argues that, strange to say, I not only won't but can't; for my voting Republican is not compossible with the fact that it was true already in the year 1548 that I was not going to vote Republican 428 years later. My rejoinder is that this is a fact, sure enough; however, it is an irrelevant fact about the future masquerading as a relevant fact about the past, and so should be left out of account in saying what, in any ordinary sense, I can do. We are unlikely to be fooled by the fatalist's methods of disguise in this case, or other ordinary cases. But in cases of time travel, precognition, or the like, we're on less familiar ground, so it may take less of a disguise to fool us. Also, new methods of disguise are available, thanks to the device of personal time.

Here's another bit of fatalist trickery. Tim, as he lurks, already knows that he will fail. At least he has the wherewithal to know it if he thinks, he knows it implicitly. For he remembers that Grandfather was alive when he was a boy, he knows that those who are killed are thereafter not alive, he knows (let us suppose) that he is a time traveler who has reached the same 1921 that lies in his personal past, and he ought to understand—as we do—why a time traveler cannot change the past. What is known cannot be false. So his success is not only not compossible with facts that belong to the external future and his personal past, but also is not compossible with the present fact of his knowledge that he will fail. I reply that the fact of his foreknowledge, at the moment while he waits to shoot, is not a fact entirely about that moment. It may be divided into two parts. There is the fact that he then believes (perhaps only implicitly) that he will fail; and there is the further fact that his belief is correct, and correct not at all by accident, and hence qualifies as an item of knowledge. It is only the latter fact that is not compossible with his success, but it is only the former that is entirely about the moment in question. In calling Tim's state at that moment knowledge, not just belief, facts about personally earlier but externally later moments were smuggled into consideration.

I have argued that Tim's case and Tom's are alike, except that in Tim's case we are more tempted than usual—and with reason—to opt for a semi-fatalist mode of speech. But perhaps they differ in another way. In Tom's case, we can expect a perfectly consistent answer to the counterfactual question: what if Tom had killed Grandfather's partner? Tim's case is more difficult. If Tim had killed Grandfather, it seems offhand that contradictions would have been true. The killing both would and wouldn't have occurred. No Grandfather, no Father; no Father, no Tim; no Tim, no killing. And for good measure: no Grandfather, no family fortune; no fortune, no time machine; no time machine, no killing. So the supposition that Tim killed Grandfather seems impossible in more than the semi-fatalistic sense already granted.

If you suppose Tim to kill Grandfather and hold all the rest of his story fixed, of course you get a contradiction. But likewise if you suppose Tom to kill Grandfather's partner and hold the rest of his story fixed—including the part that told of his failure—you get a contradiction. If you make any counterfactual supposition and hold all else fixed you get a contradiction. The thing to do is rather to make the counterfactual supposition and hold all else as close to fixed as you consistently can. That procedure will yield perfectly consistent answers to the question: what if Tim had not killed Grandfather? In that case, some of the story I told would not have been true. Perhaps Tim might have been the time-traveling grandson of someone else. Perhaps he might have been the grandson of a man killed in 1921 and miraculously resurrected. Perhaps he might have been not a time traveler at all, but rather someone created out of nothing in 1920 equipped with false memories of a personal past that never was. It is hard to say what is the least revision of Tim's story to make it true that Tim kills Grandfather, but certainly the contradictory story in which the killing both does and doesn't occur is not the least revision. Hence it is false (according to the unrevised story) that if Tim had killed Grandfather then contradictions would have been true.

What difference would it make if Tim travels in branching time? Suppose that at the possible world of Tim's story the space-time manifold branches; the branches are separated not in time, and not in space, but in some other way. Tim travels not only in time but also from one branch to another. In one branch Tim is absent from the events of 1921; Grandfather lives; Tim is born, grows up, and vanishes in his time machine. The other branch diverges from the first when Tim turns up in 1920; there Tim kills Grandfather and Grandfather leaves no descendants and no fortune; the events of the two branches differ more and more from that time on. Certainly this is a consistent story; it is a story in which Grandfather both is and isn't killed in 1921 (in the different branches); and it is a story in which Tim, by killing Grandfather, succeeds in preventing his own birth (in one of the branches). But it is not a story in which Tim's killing of Grandfather both does occur and doesn't: it simply does, though it is located in one branch and not the other. And it is not a story in which Tim changes the past. 1921 and later years contain the events of both branches, coexisting somehow without interaction. It remains true at all the personal times of Tim's life, even after the killing, that Grandfather lives in one branch and dies in the other.[6]

Princeton University

Received September 4, 1975

Notes

1. I have particularly in mind two of the time travel stories of Robert A. Heinlein: "By His Bootstraps" in R. A. Heinlein, *The Menace from Earth* (Hicksville, N.Y., 1959), and "—All You Zombies—," in R. A. Heinlein, *The Unpleasant Profession of Jonathan Hoag* (Hicksville, N.Y., 1959).
2. Account of time travel in two-dimensional time are found in Jack W. Meiland, "A Two-Dimensional Passage Model of Time for Time Travel," *Philosophical Studies,* vol. 26 (1974), pp. 153–173; and in the initial chapters of Isaac Asimov, *The End of Eternity* (Garden City, N.Y., 1955). Asimov's denouement, however, seems to require some different conception of time travel.
3. H. G. Wells, *The Time Machine, An Invention* (London, 1895), epilogue. The passage is criticized as contradictory in Donald C. Williams, "The Myth of Passage," *The Journal of Philosophy,* vol. 48 (1951), p. 463.
4. I discuss the relation between personal identity and mental connectedness and continuity at greater length in "Survival and Identity" in *The Identity of Persons,* ed. by Amelie Rorty (forthcoming).
5. "Causation," *The Journal of Philosophy*, vol. 70 (1973), pp. 556–567; the analysis relies on the analysis of counterfactuals given in my *Counterfactuals* (Oxford, 1973).
6. The present paper summarizes a series of lectures of the same title, given as the Gavin David Young Lectures in Philosophy at the University of Adelaide in July, 1971. I thank the Australian-American Educational Foundation and the American Council of Learned Societies for research support. I am grateful to many friends for comments on earlier versions of this paper; especially Philip Kitcher, William Newton-Smith, J. J. C. Smart, and Donald Williams.

The Tom Bearden
Website

Return to Reference Section

ANNOTATED GLOSSARY

T. E. Bearden 1997, slightly updated Dec. 2000
Latest editing and additions: 12th April, 2003
Presentation re-edit 7th January, 2013

Foreword

This is an informal glossary originally prepared in 1997 for Gary Hawkins to use on his web site, and informal definitions and discussions were rather quickly jotted down in 1997. We have been able to do a little updating now in Dec. 2000. Careful crafting and polishing for greater rigor would require much more time than I am able to spare, but perhaps the glossary can be helpful anyway.

We put in the gist of it, as we see it, and usually an annotation or commentary to explain it. We tried to put in sufficient explanations and a few references, so that the reader can see for himself or herself that physics, mathematics, and even logic itself are not quite as cut and dried as is often presented!

The only thing we absolutely insist on, that differs from the normal physics and engineering approach, is that any true definition must be an identity statement *a priori*. In this view, since no equation is an identity, then no equation is a definition; regardless of those "equations" universally advanced in physics as "definitions." All an equation states is that the *magnitude* of the somethings on the right is equal to the *magnitude* of the somethings on the left. It does not at all *define* any of the somethings on either right or left! I've completed as much as I have time for now, and made some additions and corrections to the 1997 version. As time permits, we will try to add some more in the future. I hope these hasty comments are helpful, and any errors in them are my own and not due to Gary Hawkins or to Tony Craddock!

Tom Bearden,
Dec. 2000

Webmaster's Note: There are additional Glossaries and Definitions at the ends of Tom Bearden's books *"Energy from the Vacuum"* and *"Excalibur Briefing."*

MATH AND SINGLE CHARACTER ABBREVIATIONS

ϕ The electrostatic scalar potential, in units of joule per collecting coulomb.

∇ The mathematical operator 'del'.

$\nabla\phi$ The gradient of the function ϕ , mathematically.

In EM theory, use of the del operator produces an electric field **E** given by $\mathbf{E} = -\nabla\phi$. An electrical field is also generated by the time rate of change of the magnetic vector potential **A**, where it is given by $\mathbf{E} = -d\mathbf{A}/dt$. Note that we have used that fact to assist us in producing the motionless electromagnetic generator (MEG), with COP = 16 as of this writing.

ϕq Joules of energy collected by a charge of q coulombs in an electrostatic scalar potential

having a reaction cross section of ϕ.

Since the collected energy is also usually the energy that can then be dissipated as work, this expression often yields the work performed in the circuit by the energy the surface charges collected from that part of the **S**-flow in contact with the conductors and components.

$d\phi/dt$ Time rate of change of the scalar potential.

Also one form of displacement current.

dE/dt Time rate of change of the E-field.

Also one form of displacement current.

dp/dt Time rate of change of momentum, which is a force F acting upon the system.

A force is defined as a time rate of change of momentum, or $F = dp/dt$.

dP/dt Time rate of change of polarization.

Polarization displacement current.

dq/dt Time rate of change of charge, otherwise known as "current".

Represents the normal current flow i of charges in a conductor, i.e., $i = dq/dt$.

dq/dt-Isolated Load Loop A special current loop containing the load,

...where $i = dq/dt$ in that "load loop" is isolated from the current $i = dq/dt$ in the source loop containing the power source, and the two loops are also field-coupling isolated. A transformer, e.g., does isolate the actual current in the secondary from the current in the primary, but the field coupling from primary to secondary and from secondary back to primary result in the enforcement of COP = 1.0 as the maximum performance possible (in a perfect, lossless standard transformer). In a real standard transformer with some losses, COP<1.0.

The notion is that, even though current cannot pass from the source loop to the load loop and there is no direct field coupling, energy density flow $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ can pass. This is in fact the definition of what we call *bridging*, if in addition to current isolation the rigid back-field coupling is broken between energy dissipation (power) in the source loop and energy dissipation (power) in the load loop. In that fashion, overunity operation of the load can readily be achieved.

E Electric Field.

Defined in older classical electromagnetics (CEM) as a force field, by $\mathbf{E} \equiv \mathbf{F}/q$. However, this only defines the entity after an interaction with mass. In short, it defines the E-field as an effect existing only in matter.

In the early days of electrodynamics, Maxwell and the other "founding fathers" assumed a material ether filling all space. Hence any field in space was indeed such an "effect" field existing in mass. Maxwell died in 1879 of stomach cancer. In the 1880s, the Michelson-Morley experiments falsified the luminiferous (material) ether as a material medium. Hence the accepted use of the $\mathbf{E} = \mathbf{F}/q$ type of "effect" electrical field as the electric field in space was also destroyed, but the equations were never changed to eliminate this error in the foundations of CEM. Instead, electrodynamicists more or less announced one day that, "Since there is no (material) ether, we are not using one!"

There is some rationalization mathematically for such an approach, if the limiting process is used for the interaction of the "causal field" in vacuum with the charged mass q, as q is allowed to approach zero. However, from a foundations view, it is still the interaction of the field in space with charged mass that is being described in the limit, and therefore it is still the "effect" entity rather than the "cause" entity.

This serious foundations problem is still unresolved in much of CEM theory and modeling to this day. The standard by which other EM texts are measured is J. D. Jackson, **Classical Electrodynamics**, 2nd edition, Wiley, New York, 1975. In several places, Jackson opts for this avoidance of the use of the " effect" field for the "cause" field: quoting p. 28: *"...the thing that eventually gets measured is a force..." "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E."* Quoting p. 249: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

The problem is further increased by use of this flawed definition of E-field to "define" the electrostatic scalar potential (Jackson, p. 34) by the equation $\mathbf{E} = -\nabla\phi$. First, no equation is a definition; it merely equates the magnitude of the entities on the left of the = sign to the magnitude of the entities on the right of the = sign. It has nothing at all to say about the definition or nature of any of the entities, but is merely a comparison of relative magnitudes. Any definition, a priori, must be expressed by an identity sign. In all of classical electrodynamics, there is no available proper definition of (x) in the identity $E \equiv (x)$. Indeed, there is no such identity available for force F. To quote Feynman: *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present… One of the most important characteristics of force is that it has a material origin, and this is not just a definition. ... If you insist upon a precise definition of force, you will never get it!"* Richard P. Feynman, Robert B. Leighton, and Matthew Sands, **Lectures on Physics**, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 12-2.

In the Sachs unified field theory, O(3) electrodynamics as spearheaded by Evans is an important subset. Hence in that approach there is a solution to the problem of defining force and field. The EM fields (**E**, **B**, etc.) when in spacetime, are in fact curvatures of spacetime identically. When these ST curvatures interact with a charged mass, the effects of the interactions are the well-known force-fields in matter, which Maxwell started with in the 1860s. So in this way, the definition problem can be rigorously resolved.

$j\phi$ The Slepian vector, representing the time rate of energy density dissipation in the current

loop containing the $j\phi$.

This does not at all represent the energy flow in space and along the outside of the circuit, which may be very, very much greater, on the order of 10^{13} times as great. It does match the Poynting energy flow component that enters the circuit.

M_q or m_q The mass of a charge q.

Instead of just using "q" and assuming it to be unitary, we must include the fact that charge is a system of two coupled components: (i) the mass m_q of the charge q, and (ii) the virtual photon flux exchange (i.e., ϕ_q) of the charge q with the surrounding active vacuum. In particle physics it is already well-known that there is no symmetry of a mass system anyway, unless the vacuum's exchange with the mass is included. So the expression $m_q\phi_q$ captures the vacuum-mass exchange. By continuing to assume that charge is unitary, physics inadvertently holds to the hoary old notion of the luminiferous ether (thin material ether) that permeates all of space. Maxwell and other electrical physicists of his time believed in a material ether, and Maxwell embedded that notion in EM theory. Inexplicably it has not been rooted out of the equations to this day, even though the notion of the material ether (not the ether per se!) was destroyed by the Michelson-Morley experiment over 100 years ago. Foundations physicists continue to point out such terrible defects in physics—and continue to be ignored. We quote Mario Bunge, "A mathematical theory of the dimensions and units of physical quantities," in **Problems in the Foundations of Physics**, edited by Mario Bunge, Springer-Verlag, Berlin and New York, 1971, p. 7: "The question 'What are dimensions?' is seen to be ill conceived, for there are two concepts of dimension, namely those of dimension function and dimension value. The correct question is: 'What are [] and *d*?' " {**Note:** [] is dimension function, and *d* is dimension value.} Bunge is clearly discriminating between the function associated with the dimension, and the magnitude associated with it

Quoting Bunge, *ibid*., p. 8: "The concept of a unit is one of those scientific notions that has remained obscure for want of a theory and excess of a coarse philosophy. Indeed we know how to manipulate units but we do not seem to know what kind of mathematical object they are. As a consequence units are sometimes confused with dimensions, at other times with standards, and they are often introduced in relation to measurement rather than to concept formulation."

And quoting Mario Bunge, **Foundations of Physics**, Springer-Verlag, New York, 1967, p. 176: "Different theories of CED [Classical Electrodynamics] are called for, both for charged particles that can to a first approximation be regarded as electromagnetically structureless (e.g., the electron and the μ -meson) and for extended systems with a charge distribution and spin (e.g., the proton). If fairly satisfactory classical theories were available they could guide the construction of the corresponding quantum theories. But no such theories are in sight partly because it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."

Again, Bunge, *ibid*., p. 182: "...the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

c The speed of light in ambient vacuum.

To discuss the speed of light, we point out that the vacuum is highly energetic, as has been shown in modern physics. Since the vacuum has an energy density, it can be modeled as a "potential" and particularly as a scalar EM potential for EM purposes. Contrary to convention, we recognize that the local energy density of the ambient vacuum potential can vary, including seasonally within and close to the solar system, etc. The average energy density of the vacuum in space far from charged masses such as stars and planets may be assumed as a "standard vacuum potential", and if this potential is equated with the notion of "flat spacetime", then an increase in this vacuum potential is an increase in the local energy density of spacetime, and hence represents a curvature of the standard flat spacetime.

In fact the interplanetary radar probes' "ranges" do show precisely such variation in the speed of light, with weekly and monthly and seasonal variations, etc. In order to comply with the nearly universal assumption in physics with respect to the constancy of c, the scientific staffs averaged the readings over the entire month. This zeroed out the variations, upheld the notion that space is a fixed type of thing, and that the speed of light in it is also a rigidly fixed constant. It *does not* agree with the actual measured speed of light from day to day.

The Russians have known and openly printed for about a century that the speed of light in a hard vacuum in deep space, far removed from solar and planetary masses, is a little slower than the speed of light in a hard vacuum near the Earth or any other large charged mass. The huge collection of charged particles in the earth have tiny separations, so that looking at the Earth as a collection of tiny dipoles, one sees a collection of enormous 4-symmetry energy flows from the time domain to the 3-space domain (and vice versa for the time-reversed portions). The increased intensity of this 4-symmetry energy flux in the vicinity of a massive planet or sun changes the local energy density of the vacuum itself, increasing it in that vicinity.

Without elaboration, we will also point out that the Earth and its dynamics structures the local spacetime and the spacetime out to some distance. So do all other planets, stars, etc. By whatever manner we model the cosmos as having been created or initiated, the dynamics and structuring of all the phenomena occurring in the universe is further impressed in the *infolded* electrodynamics inside the vacuum potential.

The "laws of nature" in one sense represent the most general results of the reactions of this internally structured "common standard spacetime" or "standard vacuum potential" with the various physical systems, devices, and measuring instruments utilized in our experiments, etc. If one deterministically alters the *structuring of the local spacetime*, one can have a system which will function there in a certain marvelous manner, but will not function that way elsewhere. Some decades ago, I participated in experiments by Golden which over a period of some five years did condition (dimension) the local vacuum potential. The result was an engine which consistently exhibited COP = 1.67 at that location, but when moved several hundred miles exhibited COP<1.0. Frank and I were made aware of this "vacuum conditioning" effect by Prof. William Tiller, who had himself conducted experiments exhibiting just such an effect over several years.

EM waves in vacuum are actually *longitudinal* (sound-type) waves, but we do not belabor the point since everything out there in the normal literature uses the transverse EM wave model for the EM wave in space. Also, the transverse wave is indeed what is detected in the interacting electron gas in our measurement instruments—whose Drude electrons are restrained to simple drift velocity longitudinally down the conductor, but are spinning and are free to precess like gyros. So our instruments detect the Drude electron precessions at right angles to the incoming disturbing longitudinal force field, which means that they do indeed measure a transverse wave. However, these are actually the *electron precession waves* in the detector, and *not* the actual EM waves in space. See my paper, "Giant Negentropy from the Common Dipole", **Journal of New Energy**, 5(1), Summer 2000, p. 11-23 for a discussion of how all EM energy comes from the time domain, via the broken 3-symmetry of the source charge or source dipole, and is emitted in 3-space in all directions as longitudinal EM waves. This is based on rigorous demonstration by E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," **Math. Ann**., Vol. 57, 1903, p. 333-355 that the scalar potential is a harmonic set of phase conjugate longitudinal EM wave pairs. We merely applied the Whittaker decomposition to the scalar potential of a source dipole or of a source charge, and the rigorous proof that EM waves in space are longitudinal directly emerges.

Just as sound speeds up, say, in steel as a denser medium than air, the EM "sound waves" in the ambient vacuum potential speed up when that vacuum potential's flux density is increased. Consequently light does move a little faster in a hard vacuum on Earth than it does in deep space, because the vacuum itself has higher energy density near the earth than in deep space. The comings and goings and orbitings of planets and the solar mass, etc. do affect the local energy density of the vacuum between the planets, and hence the speed of light through that interplanetary space.

So the interplanetary radar does show the periodic variations of the local energy density of vacuum that are induced by these factors. For a critical paper on this, see Bryan G. Wallace, "The Unified Quantum Electrodynamic Ether," **Foundations of Physics**, 3(3), Sept. 1973, p. 381-388. His discussion of the interplanetary radar measurements of Venus, which did not fit relativity, are particularly interesting, with respect to his finding that the data were simply "averaged" to eliminate the variations and hence the diurnal, lunar, and synodic periodic components actually measured in the variations of the speed of light. See particularly Bryan G. Wallace, "The great speed of light in space cover-up." **Scientific Ethics**, 1(1), Feb. 1985. p. 2-3; — "Radar Testing of the Relative Velocity of Light in Space," **Spectrosc. Lett**. Vol. 2, 1969. p. 361.

Rigorously, the field is "defined" in electrodynamics as the effect of an interaction of the force-free entity (4-spatial!) existing prior to the interaction, with a 3-dimensional static charge. This means that the field as originally defined is an *effect* and not a *cause*. To then use it as the cause is a grave non sequitur.

This confusion of cause and effect is widespread in physics. It was probably initiated by the hoary old assumption several hundred years ago that a separate force acts upon a separate mass, which is false. In fact mass is a *component* of force, since $\mathbf{F} \equiv d/dt(m\mathbf{v}) = m(d\mathbf{v}/dt + \mathbf{v}(dm/dt)$. As can be seen, in both expanded terms there exists a mass term. To skirt the issue, electrodynamicists have simply found a neat statement that "in massfree space the field continues to exist, but the force is zero." We point out that m = 0 makes the momentum vanish, and so **F** vanishes, and in the definition of $\mathbf{E} = \mathbf{F}/q$ also vanishes because the numerator **F** becomes zero. Therefore *that* **E**-field vanishes.

Feynman and Wheeler pointed out that the field *as it is defined* does not and cannot exist in massless space. Only the "potential" for the field exists there, in case some charged mass is brought in to interact with that entity that does exist there. That is accurate, since spacetime is totally a potential and any "change" in spacetime is a change in that potential *a priori*. In the Sachs-Evans unified field theory, this problem is resolved because the fundamental field is defined as a spacetime curvature (change in spacetime, hence change in the ST potential) rather than $\mathbf{E} = \mathbf{F}/q$.

In empty space there are always myriads upon myriads of interfering fields and potentials, from charges all over the universe, via Puthoff's cosmological feedback principle [see
H. E. Puthoff, "Source of Vacuum Electromagnetic Zero-point Energy," **Physical Review A**, 40(9), Nov. 1, 1989, p. 4857-4862.] Thus any EM wave traveling through space is always involved in enormously active interferometry. This alone results in violent momentary variations in the speed of that wave, even from 0.01 c to 100.00 c. For a paper showing this kind of "galloping" travel of an EM wave, see William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," **American Journal of Physics**, 53(7), July 1985, p. 671-679. While the average speed of the wave is $c(\gamma)$, where γ is the local energy density of the vacuum, the instantaneous speeds of the wave vary violently from instant to instant.

In special relativity, "velocity" merely is a rotation angle of the moving frame from the fixed laboratory observer's frame. The speed "c" merely represents a full right-angle turn away, toward the time-dimension (since other than the 3 spatial dimensions that is the only other dimension available to turn toward). In higher dimensional space, there are many other "higher" dimensions toward which further rotation can be accomplished. Hence rotating in n-space is not limited to a single orthogonal turn, and therefore velocities are not limited to speed c. However, each orthogonal rotation also subtracts one spatial dimension of the rotated object from the observer's 3-space. One rotation would be speed c, two departing rotations would be speed c^2, three would give c^3, etc. At three orthorotations, to the lab observer the rotated object is now purely an internal structure inside his time dimension, very similar to a mental object (mind is time-like, not spacelike). The nature of mind and its operations has a rather unexpected connection with matter, when one considers such orthorotations in hyperspace.

The point is, the speed *c* is not an "iron dictum" of nature, nor is it an "unchangeable constant"! Nature herself routinely violates that notion. As an example, de Broglie waves from any mass always move at $v>c$, and easily as $v>>c$. EM "surface" waves—so-called transverse oscillations or upheavals and depressions in the magnitude (energy density; actually the *reaction cross section*) of the local ambient vacuum potential), move at a *discretized* average speed $c(\gamma)$. The longitudinal component, if freed from the transverse component, can move much faster because it is a sort of "tunneling" movement. In actual quantum tunneling experiments, e.g., music (Mozart's 40th symphony) has been transmitted through quantum mechanical tunneling between two separated points at speed $v>4c$. Fogal is likely to eventually place a superluminal communication system on the commercial market. We discuss superluminal communication elsewhere in this glossary.

C Capacitance. Essentially, the charge per unit potential.

For an isolated conductor, $C = Q/V$ and C is expressed in farads. A farad is one coulomb per volt.

A lower case c represents the speed of light in ambient vacuum.

E The electric field. Essentially, $E = F/q$, but that is not really a definition.

Conventionally, the units of electric field are considered to be force per unit electric charge; i.e., force per coulomb. Note that the **E**-field even in this "classical definition" does not actually exist at a point unless there is a unit point charge (a point coulomb of charged mass) located at the point. Hence contrary to classical electrodynamics, communications theory, and whatnot, no electric force field (and no magnetic force field either) exists in the vacuum, at least of the kind prescribed by Maxwell's equations and by the Heaviside-Hertz reduction of Maxwell's equations. Since Maxwell assumed it, these equations and models all still erroneously assume a material ether filling the vacuum, in which case there *would* be matter and charges, and there would *then* be force-fields in the vacuum. We refer the interested reader to discussions by Feynman and by Wheeler. Quoting: "*One of the most important characteristics of force is that it has a material origin, and this is not just a definition. Y If you insist upon a precise definition of force, you will never get it!* " [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, Lectures on Physics, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 12-2.]

Even worse, note that $\mathbf{E} = \mathbf{F}/q$ defines not the magnitude of the field in space, but has to do with the divergence of energy from the field around a unit point static charge there in space, as well as the pattern of the diverged energy density. Actually, what is used as "the electric field" itself is the *reaction cross section* of the field, for a unit point static coulomb of charge. That is at best a function of the "field density" or "field intensity" of the actual electrical field in spacetime, and the direction and pattern of change of that field density due to the intercepting and diverting charge.

The field problem is resolved in the forefronts of electrodynamics by making the field a topological function. It is resolved in the Sachs-Evans O(3) electrodynamics (as a subset of Sachs' unified field theory) by making the field a function of the curvature of spacetime.

See also the discussion under c, above.

H Magnetic field strength.

B The B-field (loosely), or magnetic flux density (more properly).

Conventionally, this is thought to be defined, analogously to **E**, by the force it exerts. The force exerted on a point charge moving in the magnetic field **B** is at right angles to the velocity vector of the charge and also at right angles to **B**. Actually, the **H**-field was originally named the *magnetic field* and the **B**-field was named the *magnetic induction* (and is still sometimes called that.) This is an annoyance because it is **B** which involves the force, and hence **B** should be *named* the magnetic field. Other writers have noted this discrepancy, and Sommerfeld called **H** the magnetic excitation—which correctly describes its role—rather than calling it the magnetic field. As the reader can see, lots of things in electrical physics are *still very untidy*! There are also different ways to approach the EM modeling problem. In Kaluza-Klein unified field theory, all electromagnetics is modeled in the 5th dimension, which is wrapped around each and every point in our normal 3-dimensional space.

For practical use, a far better approach is the O(3) symmetry electrodynamics spearheaded by Myron Evans, which is also an important subset of Mendel Sachs' unified field theory. In the Sachs-Evans unified theory, electromagnetics has become general relativity and general relativity has become electromagnetics. Further, very powerful general relativity can now be directly engineered with this O(3) electrodynamics, as it further develops. We refer the reader to "The New Maxwell Electrodynamic Equations: New Tools for New Technologies. A Collection of 60 Papers from the Alpha Foundation's Institute for Advanced Study," Special Issue of **Journal of New Energy**, 4(3), Winter 1999, 335 p. Also, particularly see **Contemporary Optics and Electrodynamics**, edited by Myron W. Evans, a special topical issue of I. Prigogine and S. A. Rice (series editors), **Advances in Chemical Physics**, Wiley, 2001 (in publication). Evans gives a very long and detailed presentation, "O(3) Electrodynamics", in that publication, and also gives the paper "The Link Between the Sachs and O(3) Theories of Electrodynamics".

Q A quantity of electrical charge. Note that this statement *does not* define charge *q* itself.

We have tentatively proposed to redefine charge q as $q \equiv m_q \phi_q$. The term ϕ_q captures the vacuum's violent virtual energy interaction with the mass m_q of the charge, which is presently missing in classical EM theory and is an exceeding great flaw of omission. Further, ϕ_q may be decomposed into a harmonic series of phase conjugate wavepairs, as shown by Whittaker, 1903, *ibid*. [In O(3) electrodynamics, the internal structuring is dramatically extended, so that one can get at the "inner electrodynamics" necessary to explain such things as homeopathy, chi, ki, prana, etc.

To those scientists pooh-poohing such things as acupuncture, we point out that acupuncture in a dramatic case can allow a person without anesthetic to undergo a major operation with chest cavity opened, while the patient remains alert, conscious, and talking. President Nixon's physician on Nixon's trip to China witnessed such an actual operation, as well as other things. Nothing in our vaunted Western medical science can equal that, nor can any of our medical schools, nor can our National Academy of Science or our National Science Foundation. In short, acupuncture alone demonstrates how woefully inadequate our medical science actually is.

Also, acupuncture works on animals as well, so it does not involve hypnosis or deep suggestion. In science, we must believe the repeatable experiment, not our dogma at the moment. Yet the entire U.S. medical research community has no higher symmetry EM investigation of how acupuncture works, or of the infolded electrodynamics shown by Whittaker's decomposition of the scalar potential in 1903. Hence our scientists totally miss the vast electrodynamics of *the internal structuring of potential and charge with vacuum engines (spacetime curvature engines)*. As a result, the entire U.S. scientific and intelligence communities were totally unable to comprehend nearly five decades of Russian induction of health changes and diseases in personnel in the U.S. Embassy in Moscow, using weak EM radiation. Internal structuring of the activity producing what is called "charge"—including its internal structuring and engines—can be modeled and in theory directly engineered in the new Sachs-Evans approach.]

As shown in my paper, "Giant Negentropy from the Common Dipole," **Journal of New Energy**, 5(1), Summer 2000, p. 11-23, the broken 3-symmetry energy flow of any charge and any dipole allows a more fundamental 4-symmetry energy flow solution for the century-old problem of explaining the actual source of the EM energy that pours out of the source charge in all directions in 3-space. The 4-symmetry in EM energy flow explains its nature as longitudinal EM waves, with the charge's absorbed energy being *received* from the time domain (complex plane), transduced into 3-space, and emitted in 3-space. *The source of all EM energy at any point in 3-space is the time domain!*

It has been known for decades in particle physics that there exists no equilibrium in mass systems, unless the interaction of the vacuum with the masses is included. It has also been known that every charge and every dipole is a broken 3-symmetry in the fierce vacuum energy exchange with the charge or dipole. Yet the vacuum interaction—and specifically the source charge's and the source dipole's broken symmetry in it—are not even included in classical EM as yet. That this known total glitch in modern EM theory has not been previously corrected, when the particle physicists of every university physics department of note already know it is necessary, is inexplicable! Here we must regretfully point out finger at our orthodox scientific establishment, and particularly at our National Science Foundation and our National Academy of Science. These institutions are not even working on, nor are they aware of, nor do they even care about, the most fundamental science problems of our time. Bigger accelerators, yes. Windmills, yes. EM energy from the vacuum and electrical power systems which use it without destroying the source dipole, no. The greatest of all electrodynamics—the infolded longitudinal EM wave dynamics inside all EM fields, potentials, and waves—no. The correction of the dozens of major flaws and non sequiturs in classical electromagnetics, no. The understanding of the longitudinal EM wave interferometer weapons and quantum potential weapons already arrayed against us, no. The development of EM healing based on the Priore approach, which would totally revolutionize medical therapy and handle attacks on our cities by weapons of mass destruction, no.

We have a scientific mindset problem of epochal magnitude. And that costs the U.S. taxpayers hundreds of billions of dollars spent on the "status quo" or "business as usual" scientific endeavors that are far inferior to what can be done.

S Entropy. A quantity which measures the extent to which the energy of a system is available to do work.

Since in one sense energy is a function of ordering, then fundamentally, *entropy* refers to increasing disorder, which means increasing disruption of the ordering in our collected energy. Unfortunately, entropy is one of those concepts in physics for which there are several differing major views. We avoid all discussion of "information" etc., and prefer to deal only with the energy and time aspects. For our work in energy from the vacuum, we take the very simple view that an *entropic process* is like a positive resistor: As forward observer time passes, the entropic process continually disorders or "loses"—from further controlled conversion of energy form—the collected energy in the system. We regard a *negentropic process* as like a negative resistor: as forward observer time passes, the negentropic process receives energy in a form unusable to us, transforms, it, and outputs it in a form that is usable. Quoting Ilya Prigogine, Irreversibility as a symmetry-breaking process," Nature, Vol. 246, 1973, p. 70: *"Entropy Ycannot in general be expressed in terms of observables such as temperature and density. This is only possible in the neighbourhood of equilibriumY It is only then that both entropy and entropy production acquire a macroscopic meaning."*

So basically, the notion of entropy will apply when we have a system in equilibrium or very near equilibrium. For open systems in substantial disequilibrium, the entropy cannot even be computed, and the classical equilibrium thermodynamics with its infamous second law does not apply. On the other hand, the loading of disequilibrium (negentropic) EM systems is a very nonlinear affair. Some direct indications of this were in the Russian work in the 1930s which produced overunity and self-powering EM generators, in several physics institute. [E.g., see L. I. Mandelstam and N. D. Papaleksi., "On the parametric excitation of electric oscillations," Zhurnal Teknicheskoy Fiziki, 4(1), 1934, p. 5-29. English translation, Feb. 1968, Lawrence Radiation Laboratory, Livermore, CA, performed for NASA]. The U.S. researcher G. Obolensky also has done appreciable important experimental investigation of the nonlinear load effects for negentropic systems.

S Poynting energy density flow. Not the *overall* or *total* energy flow by any means, but just that component of it that is intercepted, diverged, collected, and used, where **S** is given by **S** = **E**×**H**.

The energy flow theory is rather thoroughly fouled, and has been since the 1880s when the flow of energy through space was proposed by both Heaviside and Poynting, independently and essentially simultaneously.

Since both **E** and **H** are defined not as fields per se, but as the reaction cross section of the fields with a unit charge, the S does not refer to "energy flow" per se, but to the intercepted component of the overall energy flow.

Poynting only considered the energy flow component around a circuit that strikes the circuit's surface charges and gets diverged into the conductors to power the electrons. Heaviside, on the other hand, also considered the startlingly enormous additional EM energy flow filling all space around the circuit's conductors, generally parallel to the conductors but missing them.

This huge Heaviside component is still there around every electrical circuit, but almost entirely ignored and unaccounted. It misses the circuit, passes on off into space, and is lost in most circuits. Yet it is trillions of times larger in magnitude than the very tiny Poynting component which actually gets intercepted, "caught" and used by the circuit to power its loads and losses.

Since no one could explain where on earth such an enormous nondiverged Heaviside component pouring out of the power source terminals could possibly come from, Lorentz—reasoning that, since the nondiverged energy flow did not do anything, it was "physically insignificant" (Lorentz' term)—then just integrated the energy flow vector around a little closed surface assumed around every volume element of interest. This efficiently discards the Heaviside nondiverged component *from all accountability*, leaving only the Poynting energy flow component vector accounted. Of course the Poynting energy flow will agree with the energy measured in the circuit, since we actually measure energy dissipation, and the energy that dissipates from the circuit must have first entered it. Electrodynamicists today just continue to use Lorentz's integration of the energy flow vector around an assumed closed surface surrounding any volume element of interest. Thus, sadly our electrical power system engineers unwittingly waste trillions of times more EM energy actually produced from the vacuum by the source dipoles in their generators, than what the feeble powerline is able to intercept and utilize.

We have nominated the vast unaccounted Heaviside energy flow accompanying every field and charge interaction as the source of the extra gravity known to be present in the arms of the spiral galaxies and holding them together. See my paper, "Dark Matter or Dark Energy?", **Journal of New Energy**, 4(4), Spring 2000, p. 4-11. Informally we have also nominated this dark Heaviside energy which fills all space, as associated with the "quintessence" energy that is responsible for the recently discovered antigravitational acceleration of the universe. In this view the giant negentropy process associated with charge (see my "Giant Negentropy of the Common Dipole" paper, *ibid.*) is ongoing in all the stars, galaxies, planets, nebulae, and other astronomical entities. The ordinary energy we measure in all these processes is gravitational; hence for symmetry and energy conservation, the primary longitudinal EM wave energy from the time domain that furnishes the energy to all these spatially energetic processes represents an increasing curvature of spacetime in the time domain as we travel outward from any origin point. This is an antigravitational force, in this concept, which is cumulative as a function of radial distance from the origin point. Hence we see the distant parts of the universe as accelerating the expansion of the universe. At present this is of course an hypothesis, and must be further validated by laboratory experiment and further theoretical work in a unified field theory electrodynamics such as the Sachs-Evans O(3) symmetry electrodynamics.

S-flow is actually comprised of an organized, structured flow of virtual photons or antiphotons and itself is not observable. Its interaction with charged particles (electrical charges and magnetic poles), however, is observable by the amount of energy collected upon (diverged around) the charge or pole (formation of a local potential), and formation of a translation force upon the charge or pole.

Again, the "**E**" and the "**H**" in the expression "**E**×**H**" are not really the "electrical field" and "magnetic field" in space, but only their *reaction cross sections* at each point in space, *should* a little unit point static coulomb be placed at that point to divert some of the **E**-field flow around it, or a little unit point static magnetic charge be placed at that point to divert some of the **H**-field flow around it. So even the Poynting energy flow **S** = **E**×**H** has been interpreted in terms of its reaction cross sections.

It is worth noting that electrodynamicists today are still politely debating EM energy flow and the energy flow vector. As an example, a polite "debate" of this subject has been ongoing for over 30 years in **American Journal of Physics** alone. A particularly significant statement of the problem is given by Jones: *"It is possible to introduce the Poynting vector **S**, defined by **S** = **E**×**H**, and regard it as the intensity of energy flow at a point. This procedure is open to criticism since we could add to **S** any vector whose divergence is zero without affecting [the basic integration procedure's result]."* [D.S. Jones, **The Theory of Electromagnetism**, Pergamon Press, Oxford, 1964, p. 52.] In short, here Jones is stating that an enormous nondiverged energy flow may accompany the Poynting flow component, and it will not be accounted for by **S** = **E**×**H**. After presenting the various elements required to solve the energy flow problem, Jones bows out with the statement: "*"It does not seem likely that an expression satisfying all these conditions will be simple… …fortunately, we are rarely concerned with the energy flow at a point. In most applications we need the rate at which energy is crossing a closed surface."* [Jones, *ibid*., p. 53.] Panofsky and Phillips warn against over-attachment to the Poynting vector: "*Paradoxical results may be obtained if one tries to identify the Poynting vector with the energy flow per unit area at any point."* [Wolfgang K. H. Panofsky and Melba Phillips, **Classical Electricity and Magnetism**, Second Edition, Addison-Wesley, Menlo Park, CA, 1962, third printing 1969, p. 180.

We point out, however, that for gravitational effects of energy flows in space, one must account the total energy flow per unit area *at and surrounding* any point, and therefore one must restore the unaccounted enormous Heaviside nondiverged EM energy flow component as well as the accounted tiny little Poynting diverged energy flow component.

Δt An increment (little piece of) time.

In special relativity, time can be considered a piece of spatial length compressed by the factor c, so that L = ct, where L is in meters, t is in seconds, and c is the speed of light. Time can also be considered as spatial EM energy compressed by the factor c^2. To comprehend time and time rate of flow, one must first realize that a priori no observable exists or persists in time. The observable is a 3-spatial snapshot made at one instant in time, so that "time stopped" momentarily. The observable is thus a frozen snapshot of the 3-spatial intersection of an ongoing 4-space interaction. At the very next instant, that snapshot no longer exists. Hence for a mass (or any other observable) to "propagate in time", there must be a process whereby Δt is continually added to the observable (such as to mass m, converting it to masstime mt for a short duration) and then subtracted again, leaving behind another frozen 3-space snapshot of the "observable" that has seemingly now endured for that little Δt period, with respect to the observer. As we explain in several papers, the sum total of the photon interactions with a mass m may be taken as the generator that produces (1) the flow of the mass m through time, by changing from m to mt to m to mt… continually, and (2) a substructuring of internal "time-flow components". In short, any time-flow ascertained in the macroworld or the microworld by the observer, has an enormously rich internal dynamics and structuring. Again, in higher symmetry electrodynamics such as O(3), the internal structuring of the photon, EM field, EM wave, and EM potential can be modeled and eventually even directly engineered.

ΔE An increment (little piece of) energy.

The photon is considered to be (unit-wise) composed of (ΔE)(Δt). However, in our view (as we explain in energy, energy collection, and energy dissipation) energy does not occur in "little fixed chunks" at all. One only has "little chunks" of energy when one *is collecting* energy in an ongoing process. In electromagnetics, the "collection" of energy is actually the *ongoing collecting* of energy from an energy flow deviating around an intercepting charge. It is a dynamic, ongoing process whereby a collector such as a charged particle intercepts and interacts with (undergoes virtual photon flux exchange with) an impinging energy density flow **S**= **E**×**H**. In an equilibrium (static) condition, the deviated flow remains very like a persistent, unchanging whirlpool in a river, and so is said to have a "static magnitude". If that Poynting flow is removed, then there is no excess collecting going on and the "magnitude of the deviated energy flow" is zero. EM energy collection/collecting is always like a rock in a high wind. As long as the wind is blowing on the rock, air is deviated to flow around the rock. The rock has a force field on and of it, and wishes to move. It also has "excess collected energy", and a gradient in this "excess collected energy" determined by the wind flow direction. When the wind ceases to blow, the rock goes back to being just a rock, and forgets it had all that excess wind energy collecting and any desire to move.

TERMS AND PHRASES

ABEL, NIELS HENRIK (1802-1829)

Norwegian mathematician who in 1824 proved that the general equation of fifth degree cannot be solved algebraically. He also left an important paper on transcendental functions which was published posthumously.

ABELIAN GROUP

A group obeying the commutative law of algebra for every pair of elements a and b under an operation **,** so that a*b = b*a. As an example, the set of integers is an infinite abelian group under the multiplication operation.

ACTION AT A DISTANCE

An action (effect) occurring in a local system, for which no separate local 3-space cause exists, but for which a causal correlation exists between a distant cause and the local effect.

To the local 3-space observer, the effect is baffling. In electromagnetics, action at a distance can be considered a higher topology phenomena, occurring through hyperspace or through subspace, or through "multiply connected spacetime." Particularly note that "time" is multiply connected; i.e., every spatial point in 3-space (in the entire universe) exists at the same moment in time, which means that electrodynamic action through the time domain can be used to eliminate the spatial distance between any two 3-space points. Action at a distance is absolutely required by quantum mechanics, and experiments establish its truth. Such action was implicitly included in Maxwell's quaternion theory, but was unknowingly discarded by Heaviside, Hertz, and Gibbs when they formulated a much more restricted vector interpretation of Maxwell's theory, after Maxwell's death. Bohm's hidden variable interpretation of quantum mechanics uses the quantum potential, which is a multiply-connected entity producing instantaneous action at a distance. In ordinary electrodynamics, the common scalar potential in the Coulomb gauge has instantaneous velocity, and it appears everywhere in space, wherever it will occupy, all at once *in that gauge*.

Time is also a set of energy flows in the time domain (complex plane) rather than in 3-space. Time is also multiply connected, since a priori every 3-space point in the universe exists at and in a single moment of time. Since time is dynamic and structured with special energy flows, it seems possible to use this feature to establish energy flow during one specific moment to a desired subset of 3-space points, regardless of where in the universe those points are located. Hence the development of "time-energy" technology allows action at a distance via manipulating the multiple connectivity of time and its internal structure (internal time-like energy flows).

AD HOC MODEL

A model used and fitted for a specific purpose or area.

Tentative formulation, not intended toward general application and not yet firm because there is still not a comprehensive, tested demonstration of the "fit" of the model to the complex phenomenology. Nonetheless, an ad hoc model can be quite useful in a given area, so long as it sufficiently fits the experimental data in that area and makes sufficiently accurate predictions in the area.

ADVANCED FIELD

An electromagnetic field that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime.

We believe much of the problem with the "advanced solution" is a problem of interpretation. We would urge the re-interpretation not in terms of "time-travel" where the wave continues in observer forward time while the rest of the universe is assumed to retrogress to a previous state, but in terms of time-reversal of that wave only. In that case, the rest of the universe stays sane and continues to move in observer forward time, and the wave itself just moves in forward time in a path-reversed manner, back along a previous "stimulus wave" (the retarded wave that moved forward in time normally).

ADVANCED POTENTIAL

An electromagnetic potential that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime.

Presently this potential has not yet been given a physical interpretation. See remarks above.

ADVANCED WAVE

An electromagnetic wave that is a solution of the classical Maxwell equations, and which is positioned on the future light-cone of spacetime.

See remarks above.

AFFINE CONNECTION

In mathematics: "A connection on a manifold, whose form is unchanged under affine changes of parameter along curves, e.g., when the original parameter t of a curve γ (t) is replaced by t = as + b." **Dictionary of Science and Technology**, Academic Press, Christopher Morris (Ed.), 1992.

Ugh! We'll let the mathematicians take it from there.

AFFINE TRANSFORMATION

"Transformation on a linear space to itself which can be expressed as the sum of a linear transformation and a fixed vector.

The affine transformations form a group. In the plane, the group is six-dimensional, consisting of translations, rotations, stretchings and shrinkings, reflections, simple elongations and compressions, and simple shear transformations, as well as compositions of these."

Dictionary of Science and Technology, 1992, *ibid*.

AHARONOV, YAKIR

Renowned Israeli quantum physicist and former student of David Bohm.

Co-discoverer of the *Aharonov-Bohm effect*: Even when the electromagnetic force fields reduce to zero in a charged particle system, the potentials may still exist and interfere to cause real physical effects. Together with Bohm, Aharonov authored a fundamental paper in 1959 which pointed out the primary importance of the *potentials* rather than the force fields. Indeed, the force fields are only made *in and of the charged particle system itself*, by interference there of potentials, and do not exist as such in the vacuum. The forcefields are thus not primary electromagnetic causes at all, but are *secondary effects* due to the interference of potentials in a particular charged particle system. [See *Aharonov-Bohm Effect*] In his three volumes of physics, Feynman pointed out that the forcefields do not exist *as such* in the vacuum, and that *only the potential for the forcefields exists in vacuum*. The primary causative agents for electromagnetics are thus the potentials. This has been proven both theoretically and experimentally in quantum mechanics, but still does not appear in present classical electromagnetic theory taught in universities—which terribly confuses cause waves and effect waves in wave-charge interactions.

AHARONOV-BOHM EFFECT

Phenomenon theoretically pointed out in 1959 by Yakir Aharonov and David Bohm whereby interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields.

See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. For an overview and extensive bibliography, see S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," Reviews of Modern Physics, 57(2), Apr. 1985, p. 339-436.

Chambers experimentally demonstrated the Aharonov-Bohm effect one year later. When **E-** and **B**-fields are zero in a region, their potentials still may exist and thus be stable and persistent. Being gradient-free, a force-free potential is not changing; no bleed-in or bleed-off of its magnitude *or of its internal structure* is occurring. Interference of the potentials produces real electromagnetic effects (energy) in the interference zone, even at a great distance, without transmission of "field energy" per se through the intervening 3-space. [Rigorously, the potential's EM energy can be considered to be *hyperspatially* transmitted, or *subspatially* transmitted (tunneled) via the Whittaker internal EM biwave mechanism. See discussions under *action at a distance*.]

In 1897-8 Stoney showed (in **Philosophical Magazine**) by physical reasoning that all the disturbances of the ether arising from sources of certain kinds can be resolved into trains of plane waves. This was extended and the general and special equations written by E.T. Whittaker in 1903 in **Mathematische Annalen**. In 1904 (delivered orally in 1903) Whittaker's second paper dealing with the subject was published in the **Proceedings of the London Mathematical Society**, and showed that scalar EM potential interferometry is in fact what generates the classical EM force fields.

The two Whittaker papers may be interpreted to show a startling arrangement in a potential or EM wave. (1) any EM field or wave in a particular 3-space is due to two interfering scalar potential functions in that space. So we may replace any EM field or wave with two scalar potential functions, a priori. (2) But each of the scalar potentials onto which the functions are applied, is decomposable into the Whittaker 1903 longitudinal EM wavepairs, where the output waves are in 3-space and the input waves are time-like and come from the complex plane. This provides a hidden "infolded LW electrodynamics" inside, and composing, all EM fields and waves. Further, the decomposition is recursive, so each of the first tier infolded EM waves can be decomposed again into a second tier of biwaves, etc. Thus the most primary electrodynamics of all is the nested infolded longitudinal EM wave structuring of all present EM waves, potentials, and fields. The internal LW waves also form rigorous curvature of ST patterns (templates) and thus engines. By engineering the structure and dynamics of the internal EM, one can engineer mass—including the atomic nuclei, quarks, and gluons—as one wishes, once the technology is sufficiently developed. In the West, this entire area has been ignored, but it has been highly weaponized by the KGB in Russia.

AIAS

Alpha Foundation's Institute for Advanced Study.

Unusual "think tank" comprised of fellows and fellow emeriti in various places in the world, who cooperate in advanced electromagnetics studies via E-mail. Led by Dr. Myron W. Evans, a theorist of note, the institute has been forging ahead with the production of O(3) symmetry electrodynamics, particularly as an important subset of Mendel Sachs' unified field theory. Dozens of AIAS group-authored papers have been published in various leading journals such as Foundations of Physics, Physica Scripta etc. More than 90 AIAS papers are carried on a restricted Department of Energy website for reference by DoE scientists. The work is particularly significant in working out how EM energy flows in both the 3-spatial and time domains, and how EM energy can be and is extracted from the active vacuum. More than a dozen of the later AIAS papers are carried on a DoE public website, http://www.ott.doe.gov/electromagnetic/papersbooks.html. Also, about two dozen AIAS papers have either been approved for publication in journals, or are in the review process. The present author is a Fellow Emeritus of the AIAS, and is kindly tolerated as a conceptualist by these advanced theoreticians.

AIDS

Acquired Immunological Deficiency Syndrome.

A condition of acquired immunological deficiency associated with infection of the cells of the immune systems with the retrovirus HTLV-III. It is usually recognized by testing positive to HIV infection, and eventually by physical deterioration and by development of one or more life-threatening opportunistic infections such as pneumonia, or of Kaposi's sarcoma.

AIKIDO

A Japanese martial art whose modern form is a synthesis of many schools of jujitsu and the zen art of meditation in motion.

It is the way (DO) of entering and turning (AI) the opponents vital energy (KI). The present author is a retired sandan (third degree black belt) in Yoseikan aikido. He is retired because of debilitating injuries suffered in other fields and accidents, and as a busted-up old dog of 70 has no business bouncing hard off the mats with all those vigorous young fellows. His extra 50 pounds also no longer enhance his athletic ability. We did write a book, privately held but never published, on Aikido and dealing with the principles of unarmed combat. Interestingly, we simply went into the deep unconscious in meditation (and perhaps a little beyond, depending on one's viewpoint) to uncover the unsuspected principles. To list them, we went through the alphabet four times. Many of these principles have never appeared in any text on martial arts or military combat. The task was placed on me by my sensei and required seven years to complete. Perhaps one of these days we will formally publish the book.

ALGEBRAS

The various types of mathematics in which the theoretical models for physics and other sciences are embedded.

Some important algebras utilized in physics are the vector algebra, tensor algebra, quaternion algebra, Grassman algebra, Pauli algebra, Clifford algebra, etc. There are many others. The complexity of the symmetry and of the topology varies significantly between these algebras. Thus, an electrodynamics model in a lower topology algebra such as vectors or tensors will not reveal or permit nearly so great a set of electrodynamic functions and operations as will electrodynamics embedded in a higher topology algebra such as quaternion algebra or Clifford algebra.

As an example, it has been shown by Barrett that the operation of many of Nikola Tesla's patented circuits cannot be understood if analyzed in vector or tensor electrodynamics. The circuits will reveal totally new functions when analyzed in quaternion electrodynamics. That paper is T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," **Annales de la Fondation Louis de Broglie**, 16(1), 1991, p. 23-41

ALLOPATHIC MEDICINE

A medical system which combats disease by using remedies which produce effects different from those produced by the disease treated.

Reliance is primarily upon drugs, pain relievers, and treatment of symptoms as well as use of agents to "kill" the microbial organisms regarded as the cause of the diseases. Surgery is used as a major adjunct. From time to time, other practices formerly frowned upon or even legally prosecuted are recognized as effective and legitimate, and are then brought in under the aegis of organized allopathic medicine. Physical therapy is such a formally prosecuted modality that was eventually adopted; hypnosis is another. The problems with organized allopathy is that (1) it tends to be monopolistic, (2) it uses the state's power to enforce its credo, (3) it controls both the bulk of the medical scientific literature and the flow of research funds from both government and private sources, (4) it does not adequately police its doctors and technicians, (5) it does not sufficiently accent preventative medicine or "whole human being" treatment, and (6) it engages in conspiracies to destroy competing systems of medicine (as in the case of chiropractic). (7) On the other hand, it has made remarkable advances in preventing killer diseases by vaccination, treating them by antibiotics, and in eradicating or controlling many of the former great killer diseases such as smallpox, diphtheria, and bubonic plague. It has been far less successful in treating and reversing debilitation, cancer, arteriosclerosis, etc. Just now, most of its major weapons against the former dread diseases are beginning to fail. The disease organisms are rapidly changing into new, more virulent forms that are resistant or even immune to the present "successful" treatments. Even in orthodox medical publications, it is now beginning to be recognizing that allopathic medicine is starting to fail, and that we are again entering a period of great, uncontrollable killer pandemics.

ALLOPATHY

Another name for allopathic medicine and allopathic medical therapy.

AMBIENT VACUUM

The average active vacuum, in space, removed from large collections of charge and mass.

AMBIENT VACUUM POTENTIAL

The average standard vacuum (spacetime) considered as a potential having structure and dynamics on a myriad levels.

Anything that is composed of virtual particle flux is a *potential*. Another way of looking at the vacuum is to realize that, having an energy density, then *a priori* it is a potential or may be modeled as such.

Hence the ambient vacuum can be decomposed via the Whittaker method into an internal set of infolded EM longitudinal waves of very special nature. Further, our work on the broken 3-symmetry energy flow involved in any potential or dipolarity shows the more fundamental 4-symmetry energy flow between the time domain and 3-space. This 4-symmetry energy flow is also highly structured and dynamic.

Whittaker's work shows that, contrary to present assumptions in quantum physics, a particular potential has its virtual particle flux formed into a beautiful bidirectional wave structure. The apparent disorder at any local point in the local vacuum potential is due to the constant arrival of myriads of potentials, from charged particles all over the universe. The violent interference and re-interference of these potentials results in the creation of violently changing virtual photons—or in terms of waves, violently changing EM waves and fields. In other words, the statistical disorder of the vacuum is actually *composed of* continually and rapidly changing perfect order. [Here we have confronted 3-law Aristotelian logic with its missing 4th law, because *at the limit* opposites *are* identical!] The continual violent interferometry results in the zero-point EM fluctuations of vacuum. Each EM fluctuation is completely deterministic, as shown in Puthoff's cosmological feedback mechanism. However, the macroscopic observer has no knowledge of the cause or of the arrival, and of the myriads of interferences that produced the final EM field fluctuation. Hence although the zero-point fluctuations of vacuum are completely deterministic one-by-one, the observer has *no information* on their causality. Hence to the observer these fluctuations are completely statistical and seemingly *without local order*. However, because of their hidden determinism, these fluctuations are chaotic rather than random—contrary to the Gibbs' thermodynamics statistics assumed and adopted by quantum mechanics. That statistics should be changed to an "already chaotic" statistics containing hidden Whittaker order. This would appear to be the real solution to the recognized major QM problem of the missing chaos.

Since their changes and appearances are completely in one-to-one correspondence to all distant features of the macroscopically ordered universe, then their integration yields macroscopic order. Further, as can be seen, Mach's principle—so near and dear to the heart of Einstein—arises directly from the foregoing discussion. In such fashion, integration of statistical (but chaotic, and with hidden order) quantum changes results in the familiar, ordered macroscopic world. This removes the great error in QM: that heretofore it has failed to predict the ordered macroscopic universe we observe and inhabit.

ANALOGUE

Part that is analogous (representative or modeled).

At root, an analogue of something is a second something that functions in similar fashion to the first something.

ANAEROBE

An organism that lives, is active, and occurs in the absence of free oxygen. Particularly applicable to one-celled organisms that do so.

ANAEROBE, FACULTATIVE

An organism that can live either as an aerobe or as an anaerobe. Particularly applicable to one-celled organisms possessing that capability.

ANAEROBE, OBLIGATORY

An organism that is restricted to being anaerobic; i.e., it cannot utilize oxygen from the atmosphere.

ANAEROBIC

Characteristic of an organism of living, active, or occurring in the absence of free oxygen.

ANGULAR MOMENTUM

The momentum or "leverage" of the momentum of a moving body, with respect to an axis or reference point.

Angular momentum as the same units as action; that is, it is energy multiplied by time, or momentum multiplied by length. In terms of particles, it is also known as *spin*.

ANOMALOUS EM JAMMING MECHANISM

In a dense signal environment, an unusual mechanism whereby EM energy from the skin of an aerial vehicle is directed laser-like into the internal circuitry.

The density of the signals impinging upon the skin are such that the probability of multiple photon interaction increases. Also, on a nonlinearity in/on the skin, the multiplex of signals creates a scalar EM potential. Decomposition of this potential into Whittaker waves provides hidden "pump" waves upon the nonlinearity, which therefore acts as a pumped phase conjugate mirror (PPCM). EM fields and signals from the internal operating circuits in the vehicle provide "signal waves" to the internal skin and to the nonlinearity PPCMs on the skin. Amplified phase conjugate replica (PCR) waves are generated by the PPCMs, where in each PCR wave the energy may be the total of the external signal density pump waves upon the initiating PPCM. The result is that the skin becomes a "collector" of energy from all the impinging external signals, and focuses much of this collected external energy into laser-like beams inside the vehicle which penetrate directly back to the circuit element that originated the particular signal field/wave to the inner skin. The result is spurious jamming throughout the vehicle, directly in its operating circuits, in a manner not described in normal ECM theory. Also, the same effect occurs in a human body when exposed to such a dense signal environment (threshold about 200,000 signals per square meter per second, and particularly active above 500,000 signals per square meter per second. I.e., in such an EM "smog" environment, the anomalous jamming mechanism is able to progressively conduct the energy impinging upon the skin deep into the body and throughout it, including into the bone marrow. This effect is not described at all in the standard dosimetry literature and experimental approaches for examining the effects of EM fields and radiations upon biological systems. By rigorous frequency band control of radiations and communications, most equipment can be operated in an appreciably dense environment such as appears on the modern environment, because the equipment is essentially "narrow band." The human, however, is a very wide-band receiver, and in such dense environments—even at very weak signal levels—the human system will eventually undergo cumulative damage. This is one of the major mechanisms involved in long term deleterious biological effects of EM radiation environments, but this effect on living systems has not been investigated in the laboratory by present EM bioeffects researchers.

In the U.S. air attack on Libya in 1986, the anomalous jamming mechanism interfered with the aircraft, controls, and missiles because frequency discipline was not rigorously enforced. In the Gulf War, rigid frequency discipline was enforced, with the result that the electronic systems operated normally. However, the troops exposed to that environment (which varied from place to place) were affected by the mechanism, which added a bit to the previous cumulative damage to their immune systems, etc.

ANTIBODY

A body substance produced in response to a specific antigen (protein or carbohydrate or other substance, such as a toxin or enzyme, capable of stimulating an immune response).

The antibody counteracts the effects of the antigen, by neutralizing toxins, agglutinating bacteria or cells, precipitating soluble antigens, etc.

ANTI-ELECTRON

The time-reversed electron, or *positron*. Under time reversal, electric charge reverses in sign.

Paul A.M. Dirac, in his "A Theory of Electrons and Protons," Proceedings of the Royal Society of London A126, 1930, p. 360 introduced the negative energy electron sea (now called the "Dirac sea") with negative energy electron holes treated as the positive electron. He attempted to identify these holes with protons, but that was later refuted. The positron, discovered about a year after Dirac's paper, was later connected with this anti-electron.

Note that Dirac himself was not adamant that *all* negative energy state holes in the Dirac sea were filled with electrons, but just that *almost* all of them were. Modern researchers have tended to extend this to state that *all* the holes are filled, normally. That is sufficient in forward time situations. But it does not necessarily hold in a time-reversed situation, in a time-reversed zone, or in a tempic back emf.

ANTIGRAVITY

An effect whereby normal masses *repulse* one another rather than *attract* one another

Antigravity may also be considered as *time-reversed gravitation*. It may also be considered as *gravitation between two negative masses*.

ANTIMATTER

Matter consisting of atoms which are composed of antielectrons (positrons), antiprotons, antineutrons.

Also loosely refers to the antiparticle corresponding to a particle—the antiparticle may be regarded as the particle traveling backward in time, or "phase conjugated", or "time-reversed.

ANTIPARTICLE

A counterpart to an ordinary fundamental particle, having identical mass lifetime and spin, but with charge and magnetic moment reversed in algebraic sign.

Note that when the charge is reversed, parity and time are also reversed. Thus, properly, the antiparticle is the phase conjugation of the particle, or just conjugate particle for short. In effect an antiparticle is the particle traveling backward in observer time rather than forward in observer time.

ANTIPHASED

180 degrees out of phase (contrary-wise in timing).

ANTIPHOTON

A time-reversed photon.

Presently, the photon is considered its own antiparticle, so that photon and antiphoton are one and the same. In the new approach, that is not quite true. First, one considers that a photon—as is well known—has the units of angular momentum. Thus it is made of "energy multiplied by time." We consider that the time-forward photon (i.e. a normal photon), is made of $(+\Delta E)(+\Delta t)$. Usually physicists have largely neglected the fact that a photon carries an *increment of time* as well as an increment of energy! In our approach the antiphoton, being a *time-reversed* entity, must be composed differently. We have tentatively considered the antiphoton to be comprised of $(\Delta E)(-\Delta t)$. It will probably initially be necessary to take into account the fact that Minkowski geometry treats the T dimension as imaginary rather than simple arithmetic. We haven't had the lengthy time to work all of that out; perhaps one of the better theoreticians or one of the sharp young graduate students will do that and save us the trouble.

At any rate, the requirement is that—in a coupled photon-antiphoton pair (a graviton)—the energies of photon and antiphoton must add spatially, while *electromagnetically* the translation forces (actually translation impulses) of photon and antiphoton must cancel or thwart each other. One insists on a strong interpretation of the distortion correction theorem in phase conjugate optics. In other words, we consider the "wave" as composed of photons, while the "antiwave" (phase conjugate replica) emitted by a phase conjugate mirror is composed of antiphotons. And we consider that the distortion correction theorem must apply to the wave's photons, so that the antiphotons are precisely superposing (coupling) and then decoupling as the wave and antiwave pass through each other in perfect spatial superposition. We feel that this "totally in phase spatially but totally antiphased temporally" concept is important. It is well-known, e.g., that in certain cases protons—the very building blocks of matter itself—can pass right through other protons, as shown in certain colliding beam experiments if their relative spins are properly arranged.

ANTISIGNAL

The phase conjugate of a signal. The time reversal of that signal or its twin. The phase conjugate replica signal/wave emitted by a phase conjugate mirror material in response to an incident signal/wave.

ANTI-STOKES EMISSION

Radiation emitted from an intensely scattering medium that is of greater intensity than the input radiation.

In other words, the medium gives off more energy than we have to put into it. This has been known since the 1930s and proven experimentally, but no really complete explanation has been forthcoming. The conventional explanation is that some of the kinetic energy of the medium is furnished to cause additional emission. In that case, anti-Stokes emission would lower the kinetic energy of the emitting medium. Note that this is not necessarily true at all in Letokhov's negative absorption (excess emission) of the medium, where in that case the excess energy in the emissions is said to come directly from energy absorbed from the active vacuum.

ANTI-STOKES RADIATION

The radiation coming from an anti-Stokes emission process from an intensely scattering medium, where more energy flux is emitted than is input by the experimenter-operator.

In the past, there has just been "arm-waving" that the excess energy must come from the internal energy of the molecules. There seems to be no explanation at all of where the excess energy of the molecules must continue to come from! Tacit silence on this "apparent" violation of energy conservation is held. But see Letokhov's *negative absorption of the medium* discussions in the literature.

We quote from H. C. Dake and Jack DeMent. (1941) **Fluorescent Light and Its Applications**, Chemical Publishing Company, Inc., Brooklyn, New York, 1941. p. 51-52: *"When a phosphor or other luminescent substance emits light, it gives in most cases an emission according to Stokes' Law. This law states that the wavelength of the fluorescent (emitted) light is always greater than the wavelength of the exciting radiation. It was first observed in 1852 in the memoir "On the Change of Refrangibility of Light," by Sir G. G. Stokes. In terms of energy the relationship states that e em < e ab. While Stokes' Law holds for the majority of cases, it does not hold in certain instances. In some cases the wave length is the same for both the absorbed and the emitted radiation. That is, the efficiency appears to be perfect or unity. This is known as resonance radiation. In other cases Stokes' Law does not hold where the energy emitted is greater than the energy absorbed. This is known as Anti-Stokes emission. In 1935 Prileshajewa showed that there is an energy difference as much as 1.1 v between the exciting light and the fluorescence of aniline vapor. This added energy is attributed to additions from the internal energy of the molecule."*

There is no real problem of identifying the source of the excess energy, once one realizes that (i) every dipole is a broken symmetry in the vacuum exchange with the end charges of the dipole, and thus extracts virtual energy from the vacuum, integrates and orders it, and outputs it as EM energy density flow. One then adds multiretroreflection and multiple passes of the EM energy flow, so that charges that intercept and collect energy from the reciprocating flow can multiply connect from the same energy flow (on a given pass, only a tiny, tiny fraction—nominally about 10^{-13}—of the energy flow is intercepted and interacted with by the collecting particle). Thus multiple collections from the same EM energy flow can be accomplished, if the flow is iteratively rerouted (as by retroreflection or phase conjugate reflection) back across the collecting charges again and again.

In this fashion additional energy is collected in the collected volume, which increases the local collected energy density in that volume. In turn this increases the local potential of that volume. *In turn* that increases the energy flow from that potential acting as a dipole, because it increases the dipole strength and consequently the magnitude of the dipole's broken symmetry in its vacuum flux exchange.

The inefficiency of the process increases nonlinearly with increasing local potential, and faster than the buildup of local energy density occurs from multipass multicollection. When the inefficiency curve reaches the increasing broken symmetry energy extraction curve, then a point of equilibrium is reached. At that equilibrium point, the rate of energy diffusing out of the multipass region and escaping from the periphery of the system is equal to the rate of extraction of energy from the vacuum and flowing it out into as Poynting **S**-flow into the multicollection region and collecting it.

So we have described an open thermodynamic system, where there are two inputs of energy flow: (i) that by the operator-experimenter, and (ii) that from the vacuum via the dipolar broken symmetries. At equilibrium condition the total input energy rate is equal to the total energy escaping rate, and the conservation of energy law is rigorously upheld. However, this is an open thermodynamic system far from equilibrium, and so rigorously the nonlinear thermodynamics of such systems applies, rather than the normal equilibrium thermodynamics. As is well-known, the coefficient of performance for such open systems in disequilibrium, may permissibly be overunity, even greatly overunity.

Properly utilized, the overunity anti-Stokes or Letokhov emission can be made very large—such as COP ≈1200 in Patterson's adaptation. The process and its overunity certification already exist in the hard science literature, being known and experimentally demonstrated for at least three decades.

ANTIWAVE

The time-reversal (phase conjugation) of a reference wave.

A phase conjugate replica wave is a time-reversed replica of its referent wave. Hence it is the antiwave of the referent wave.

A-POTENTIAL

Vector magnetic potential.

The A-potential is a real entity more primary than the magnetic field **B**, and not a mathematical convenience as was originally thought—as detailed in the second volume of Feynman's three volumes of physics. Indeed, it may be separated from the magnetic field, as reported by Feynman, and used for instrumentation effects in the manner patented by Gelinas. When separated from magnetic field, it is often referred to as the "curl-free vector potential," since the **B**-field is modeled as being created from the **A**-field by curling it. In other words, **B** is defined as the curl of the magnetic vector potential, if we use $\mathbf{B} \equiv \nabla \times \mathbf{A}$. Hence "separating" magnetic field from **A**-potential simply is a sort of "parallel branching" of the energy density flow of the **A**-potential. It is like branching the voltage on a feeder line conductor in an electric circuit into two parallel conductors. The feeder potential splits into two, so that each branch has the same potential. Actually the vast energy flow in the feeder line only has a small component which strikes the surface charges, diverting the energy into the conductor and potentializing the Drude electrons. In branching into two conductors, the giant flow branches also, but is so great that the surface charges in each of the two branches still intercept essentially the same amount of energy, diverging it into the conductor. Hence each conductor exhibits the same potential as the feeder conductor

The same thing happens when one "splits" or branches the **A**-potential in similar fashion. The diverted and separated **B**-field branch will exhibit "equal energy density" flow (i.e., in the form of magnetic flux density flow) and yet the original **A**-potential path through space will still have the same previous energy flow density magnitude, just without curl (without magnetic field).

Any good toroid or very long solenoid already performs such separation, and hence energy density flow doubling. If properly used—as in the motionless electromagnetic generator (MEG)—this effect can be used to provide a generator with COP>1.0. At this writing, the present MEG exhibits COP = 16, and is a patent-pending system.

The well-known Aharonov-Bohm effect, e.g., also uses this separation of the curl of the **A**-potential, leaving an uncurled **A**-potential whose changes will and do interact with electrons.

The curl-free vector potential is particularly interesting because (1) it is a unique and independent field of nature, (2) its curl produces a magnetic field, (3) its time derivative produces an electric field, (4) its magnitude does not fall off inversely as the distance squared, but only inversely as the distance, and its time derivative $d\mathbf{A}/dt = B\ \mathbf{E}$. Hence, as in the MEG, one can extract energy both from the curl-free A-potential and the curled **A**-potential (magnetic field **B**), increasing the available energy collected in the circuit to greater than what the operator input to it. This is simply a physical process of regauging, wherein electrodynamicists already routinely assume (and apply) the fact that the potential energy of any EM system can be freely changed at will, by regauging. In the real world, of course, we may have to pay a little switching energy, but we are free to regauge (by this "splitting the potential into equal parallel paths, and by other methods), at will and at any time and place. The MEG just takes advantage of one major method for free regauging, in order to violate the Lorentz symmetry condition (violate symmetry in the system's energy exchange with its active vacuum environment). It thereby is legitimately permitted to exhibit COP>1.0 and output more energy than one inputs, because the regauging transforms it into an open dissipative system far from thermodynamic equilibrium in its active environmental energy exchange.

ARISTOTELIAN LOGIC

The conventional three laws of logic, unconsciously fitted to the physical reality created by the single photon interaction. The three laws are: (1) a thing is identical to itself, (2) a thing is not identical to something else, and (3) either one has a given thing or something that is not that thing. Mathematically these are written as (1) A A, (2) A Ã, and (3) A Ã. Aristotelian logic applies only to single observation at a time (monocular observation, a single observer). When multiple simultaneous observation (multiocular observation, multiple observers) are considered, the three laws may be violated. In that case a fourth law, A Ã is added. This law may be seen to be the exact negation of all the first three. An application rule (which may itself be called a fifth law) is also added, and states that either the first three laws are explicit and the fourth is implicit, or the first three are implicit and the fourth is explicit. Further, the fourth law is the *law of the paradox*: Something that is found to be true in nature, but violates one or more of Aristotle's three laws, is known as a *logical paradox*. In the higher logic, it simply violates the case where the three Aristotelian laws are explicit and the fourth law is implicit. Instead, the three Aristotelian laws are now implicit and the fourth law has become explicit. All boundaries exhibit this effect.

It is highly recommended that the interested reader familiarize himself or herself with Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980. Kline deals directly with the fact that mathematics is not a body of unshakable truths about the physical world, and mathematical reasoning is not exact and fallible.

ARISTOTLE

A great Greek philosopher who lived from 384-322 B.C.

Aristotle is famous for his three laws of logic: (1) a thing is identical to itself, (2) a thing is not identical to something else, and (3) either one has the thing considered or one has something else.

Heraclitus pointed out that Aristotle's laws prohibited any change, and yet changes were ubiquitous. In short, for a thing to change, it had to change into something else, so Heraclitus pointed out. So how could a thing be both itself and something else simultaneously? How could it be not itself while being itself? The problem has not been solved by the logicians and philosophers to this day, because it has no solution in Aristotelian 3-law logic. So Heraclitus, believe it or not, concluded that all change was an illusion.

ATHERIOSCLEROSIS

Basically, clogging of the arteries, which may be due to deposits onto the artery walls or by thickening and growth of the artery walls themselves, or a combination. It is a chronic (slowly progressing) disease where the artery walls become progressively thickened and hardened and the artery becomes less elastic.

It has been found that the common cytomegalovirus (one of the herpes viruses) plays a key role in the development of the disease. Initial deposits irritate and injure the arterial walls, and the herpes 6 virus infects the site, partially nullifying the control of excess grown of the arterial muscle tissue. This deposit and concomitant stimulation of some excess muscle growth occurs repeated, so that the muscle tissue grows abnormally over an extended period of time. For this reason, when an operation is performed to forcibly reduce the deposits inside the artery, the resulting injuries to the arterial wall tissues leads to the herpes 6 infection and stimulation of excess muscle growth. It has been found, e.g., that after angioplasty the process is often engendered, so that within a year or two the artery is narrowed again, but this time by abnormal growth of smooth muscle tissue. The Prioré team demonstrated positive and total remission of arteriosclerosis in laboratory animals by means of nonionizing EM irradiation to produce an amplified "antiengine", which "time reversed" or dedifferentiated the stricken cells and the stricken areas back to a previous state.

ARTIFICIAL POTENTIAL

A scalar potential which is composed of, or partially contains, an artificially assembled bidirectional wave set.

See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," **Mathematische Annalen**, Vol. 57, 1903, p. 333-355 for proof that a "scalar potential" is actually a harmonic set of hidden bidirectional EM phase conjugate longitudinal wavepairs. Each wavepair consists of a wave and its antiwave (true time-reversed replica wave).

If the external observer could see the *detected* (effect) waves in a hidden wavepair, he would see the "wave" going in one direction and the antiwave passing precisely through it in the other direction. However, prior to detection, the phase conjugate wave exists entirely in the complex plane and hence in the time domain (in -ict). So the interpretation of Whittaker's 1903 paper, in terms of cause and effect waves, is that the phase conjugate half set is incoming from the time domain to each and every point in the dipolarity or potential, and the "real 3-space half set" is being radiated in every direction in 3-space from that point. See my paper, "Giant Negentropy….".

In addition to Whittaker's sum set of biwaves, Ziolkowski added the product set in the mid-1980s. E.g., see Richard W. Ziolkowski, "Exact solutions of the wave equation with complex source locations," **Journal of Mathematical Physics**, 26(4), April 1985, p. 861-863; Rod Donnelly and Richard Ziolkowski, "A method for constructing solutions of homogeneous partial differential equations: localized waves," **Proceedings of the Royal Society of London A**., Vol. 437, 1992, p. 673-692. Particularly see I.M. Besieris, A.M. Shaarawi, and R. W. Ziolkowski, "A bidirectional travelling plane wave representation of exact solutions of the scalar wave equation," **Journal of Mathematical Physics**, 30(6), 1989, p. 1254-1269.

However, these deal only with the effects waves after detection.

Another way of defining an artificial potential is: *A potential formed by or containing a subset formed by, deliberately summing nonzero force field vectors to a zero vector resultant*. [Note that electrodynamicists routinely discard vector systems that sum to a zero resultant; not realizing that in doing so, they are discarding vacuum engines and a vastly extended electrogravitational dynamics!] The vector zero system of infolded nonzero vectors has a deterministic, internal stress pattern that is macroscopic. The type of potential made by the zero summation depends upon the type of force vectors summed. For example, summing electric field vectors to zero produces an electrostatic scalar potential; summing magnetic field vectors to zero produces a magnetostatic scalar potential of vacuum.

All potentials represents local warps or curvatures in vacuum/spacetime, and all artificial potentials contain internal, deterministic templates of nested internal spacetime curvatures. Curvature of spacetime interacts directly upon mass and physical systems embedded in it. Thus these deterministic, internested local spacetime curvatures of the artificial potential represent vacuum engines, or spacetime engines, and the artificial potential is said to be a *dimensioned* or *activated* potential. Any system exposed to and placed in such a vacuum spacetime engine, will have these hidden vacuum flux asymmetries interacting upon each and every part of it, down to and including the nuclei, the nucleons, and the quarks in the nucleons.

ASYMMETRICAL REGAUGING

Change of either the scalar potential f or the vector potential **A**, but not both, so that a single excess force and a single change of system energy occurs in the system.

This violates the assumed Lorentz condition, and thus excess energy can enter the system, and an excess "free" force appears which can (if the system is adroitly designed) then translate to perform work upon the system to increase its (e.g. kinetic) energy, exhausting the excess regauging energy in the process and restoring symmetry.

In conventional EM theory, Maxwell's equations in potential formulation are in terms of two potentials, f (the scalar potential) and **A** (the vector potential). When Heaviside retranslated Maxwell's quaternion set of 20 equations in 20 unknowns, he transformed it (in potential form) into two equations where main variables f and **A** are not separated. Electrodynamicists, who until recently considered the potentials as not even real, but just mathematical figments and conveniences, then simply arbitrarily changed both f and **A** *just precisely so* that the extra force appearing in the system by the change of f was countered by an equal and opposite extra force also appearing in the system by the change of **A**. This type of *symmetrical (no net force) regauging* is known as the *Lorentz condition*.

Each *asymmetrical* (produces a nonzero excess net force) potential change (each half of the two asymmetrical regaugings that comprise the symmetrical regauging) also altered the overall energy of the system. However, in the symmetrical regauging that total energy change is "bottled up" as increased stress in the system, because no net translation force is available to be translated and use the energy change to produce excess "free work."

The Lorentz condition assumes that the designer *will not allow* the equilibrium of the system to be broken! In that case, of course, ordinary equilibrium thermodynamics applies, and the system cannot produce COP>1.0. Prior to this Lorentz condition assumption and its absolutely arbitrary regauging, Maxwell's equations do include asymmetrical regauging and therefore open EM systems far from thermodynamic equilibrium; precisely those which violate the Lorentz condition (symmetric regauging) assumption. In short, the unregauged equations do include COP>1.0 electrodynamic circuits and systems.

Unknowingly the electrodynamicists *arbitrarily* discarded the entire Maxwellian class of overunity EM engines and circuits—which class incorporates making a change in one potential(s) such that the Lorentz condition does not hold.

ASYMMETRY

Lack of symmetry, or "broken symmetry" in a general sense.

ASYMMETRY OF THE DIPOLE

Any dipole or dipolarity is a broken 3-symmetry in its fierce exchange of EM energy with the active vacuum. This means that it continuously receives and absorbs copious virtual energy from the seething vacuum flux, and that not all this absorbed energy is reradiated back to the vacuum in virtual form. Instead, some of it is integrated into observable form and re-radiated in all directions as usable, real 3-space EM energy.

This is standard, well-known particle physics, although it appears to be totally missing from classical electrodynamics and it is not considered at all by electrical power system designers and analysts!

Every electric charge is a broken symmetry in the virtual particle flux exchange between the quantum mechanical (energetic) vacuum and the mass of that electric charge. The dipole is two different such asymmetries. As is well-known in particle physics (not in classical EM theory!) each such broken symmetry must "gate" out some of the vacuum energy. In short, the charge asymmetry "extracts" some vacuum energy from its normal "energetic exchange" with the vacuum flux, and puts it out in some form we call "observable" (capable of being measured; specifically, of being intercepted by electric charges so as to produce a force field across the charge and translate it).

Well, an electrical charge produces a flow of energy density which we may represent conventionally as $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ (when collected and measured on assumed unit point charges), which we—a little bit mistakenly—call "observable." [Actually, until the energy flow interacts with that assumed unit point charge, it is an *organized flow of virtual entities*] After interaction it is real observable EM energy in 3-space. The spinning electron has its own magnetic moment and its own **E**-field. Since (in one form) also $\mathbf{E} = -\nabla\phi$, an extracted and formed $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ is output by the charge asymmetry. This output energy flow also produces the assumed ϕ evidenced by the charge interaction, since ϕ as used represents the *reaction cross section* of the potential. Note that all these entities **S**, **E**, **H**, and ϕ are defined in electrodynamics only as their own interaction cross section with a unit point charge containing mass.

In classical electrodynamics (formed long before the notion of charged particles, atoms, and EM energy flow!) it is just recognized that "charge is the source of a potential." Actually charge gates the energy flow from the vacuum flux, receiving the energy from the time domain by its asymmetry, transduces the energy into 3-space energy, and outputs it as real EM energy pouring out in 3-space in all directions. That energy-flow contains and carries the energy that will interact with charge to form the **E**-field, the **H**-field (spin field), and ϕ.

Also, in particle physics and gauge theory it is well-known that *no mass system can be in equilibrium anyway, unless the vacuum-to-mass interaction is incorporated!* After CEM was formed, much later the atom was discovered and the electron was discovered. Much later then, particle physics and quantum mechanics came along. CEM theory has never been altered to include the necessary vacuum interaction, although it has been known for decades that CEM is totally in error in this respect.

Here one poses a question to the sharp young graduate students: Why does every university continue to teach a seriously flawed classical electrodynamics which does not include the vacuum/mass interaction, when the physics department at the same university knows full well that it must or no EM system can be in equilibrium? Why is not such a *fundamental* and universal issue strongly raised and intensively discussed by our large, tax-exempt scientific organizations at their annual symposia? Why do not the National Science Foundation and the National Academy of Science demand that serious work be done at all speed, to correct such known scientific deficiencies in the classical EM theory? *What happened to science as the search for physical truth*, and replaced it with "status quo" models or—in some cases—even dogma? It is an inexplicable mystery to me, and one to which I do not know the answer.

AUTOIMMUNE DISORDER

A disease or disorder relating to or caused by the antibodies or T-cells that attack foreign molecules, cells, or tissues of their host organism.

A flagrant autoimmune disease example is AIDS, Acquired Immunological Deficiency Syndrome. The AIDS virus successfully hides from the antibodies, so that they cannot perform their proper function, and it infects the T-cells and negates their function. However, there are other examples other than AIDS: e.g., rheumatism and osteoarthritis.

BACK EMF

The counter-electromotive force, usually between the end charges of the source dipole for an EM circuit.

When the electrons in the circuit are forced against the emf—as when the charges from the ground reference of the circuit are forced back up through the emf of the source dipole—then work is done upon the source dipole to scatter the charges and destroy the dipole. Removing the dipole removes its receipt of vacuum energy via its broken 3-symmetry, transduction of that energy, and output of it along the circuit and in space around the circuit as the energy flow that powers the circuit (with part missing the circuit and being just wasted).

Since this removes the power from the circuit, then additional energy must be input to the source to perform work on the internal charges and restore that source dipole again.

There are actually two emfs in a circuit, not one. There is a 3-space emf and also an entirely unrecognized tempic emf. In either case, under proper circumstances there can be a back-emf or either one or both. See discussion under *electromotive force, back*. See also *tempic back emf* (to be added).

BALL LIGHTNING

Rare form of lightning where an incandescent, slow-moving globe forms. The globe often moves erratically, explodes, and may set fire to objects it touches.

Ball lightning may be due to scalar electromagnetic interferometry which can produce a controlled pattern of electromagnetic energy at a distance, or it may be due to the formation of a special localized spacetime curvature engine.

BARE ELECTRON

The true electron without its partial shielding of virtual positrons that are attracted around it.

BARRETT, TERENCE W.

Noted modern electrodynamicist and consultant who works in SU(2)×SU(2) gauge symmetry extended electrodynamics, and was also one of the pioneers of ultrawideband radar.

BECKER, ROBERT O.

Medical doctor, medical research scientist, and Nobel Nominee, noted for his early-on efforts—particularly in the 1960s through the 1980s, in the effects of EM fields on biological systems, studies of evoked or partially evoked limb regeneration in certain species, electrical stimulation of bone fractures to stimulate healing and reveal the mechanism involved, and the cellular regeneration system of the body as opposed to the immune system.

Becker particularly recognized that the cellular regenerative system was a special electrical system having unique functions. He even succeeded in diagramming this system as far as it is revealed by ordinary U(1) electrodynamics. Becker demonstrated partial limb regeneration in salamanders, etc. and in some other species, with electrical stimulation. In his EM-stimulated bone fracture healing, Becker showed a startling result: First the red blood cells entering the fracture area shucked their hemoglobin and grew a nucleus, thus "dedifferentiating" (biological term) or "time-reversing" (physics term) back to an earlier state. Then this modified cell redifferentiated (biological term) or "fast-forwarded in time" (physics term) into the type of cell that makes cartilage. Then this newly modified cell further redifferentiated into the type of cell that makes bone, and was deposited in the fracture site. The continuing deposits of these cells from the ongoing process evoked in red blood cells healed the fracture with new bone growth. Whereupon the action stopped.

Becker also was one of the first qualified scientists who testified in court on the deleterious effects of power line radiation. He thereby incurred the enmity of very powerful interests, so that his research funds were pulled and eventually he was forced to retire at an early age.

He was nominated for the Nobel Prize multiple times, and his epochal work did finally focus at least some attention of the medical research community on the regenerative system as opposed to the immune system which heals nothing.

Becker's profound work was of tremendous influence on me as I struggled to decipher the active mechanism in his experiments, the Prioré team's experiments in France, the Russian microwave radiation of U.S. Embassy personnel in Moscow, Kaznacheyev's experiments demonstrating transmission of cellular disease and disorder electromagnetically into target cells environmentally isolated except for contact by light from the sender cells having such disease or disorder, and the body's own regenerative system. After some 14 years struggle with the problem, I finally did break the mechanism(s) involved in all those things. But it was only possible because of the magnificent work by Becker and Prioré, the reported work by Kaznacheyev, and gradual exposure to higher symmetry electrodynamics. The work was crude at first, but eventually progressed (by 1998, e.g.) to include the port-hole concept, transduction of EM wave polarization types, the engine and engine template concepts, and unified field theory. Since then it has become clear that the proper electrodynamics to model such things is the higher symmetry O(3) electrodynamics pioneered by Evans, and now revealed to be an important subset of Sachs' unified field theory.

Becker is therefore one of my heroes whose magnificent work and long perseverance against strong scientific suppression I greatly admire. Most of Becker's detractors were or are in fact employed in positions either directly or indirectly funded by the electrical power industry. I believe that, had Becker had access to O(3) electrodynamics and Sachs' theory, he would have personally impelled the entire medical field to a marvelous and extraordinarily advanced EM medical therapy.

See Robert O. Becker and Andrew A. Marino, **Electromagnetism and Life**, State University of New York Press, Albany, 1982; Nice summary of EM bioeffects (orthodox) ; also on http://www.ortho.lsumc.edu/Faculty/Marino/EL/ELPub.html.

See also R. O. Becker, "The neural semiconduction control system and its interaction with applied electrical current and magnetic fields," **Proceedings of the XI International Congress of Radiology**, Vol. 105, 1966, p. 1753-1759, Excerpta Medica Foundation; Amsterdam; — "The direct current field: A primitive control and communication system related to growth processes," **Proceed. XVI Internat. Congr. Zoology**, Washington, D.C., Vol. 3, 1963, p. 179-183; — "The control system governing bone growth in response to mechanical stress," **J. Ark. Med. Soc**., Vol. 62, 1966, p. 404; — "A technique for producing regenerative healing in humans," **Frontier Perspectives**, 1(2), Fall/Winter 1990, p. 1-2; — with Charles H. Bachman and Howard Friedman, The direct current system: A link between the environment and the organism," **New York State Journal of Medicine**, Vol. 62, April 15, 1962, p. 1169-1176; — and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents," **Trans. N.Y. Acad. Sci**., Vol. 29, 1967, p. 606-615; — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," **Bulletin of the New York Academy of Medicine**, Second Series, 48(4), May 1972, p. 627-64; — and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, **Electromagnetic Applications In Medicine**," NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993.

See particularly Robert O. Becker, "The significance of bioelectric potentials." **Bioelectrochemistry and Bioenergetics**, Vol. 1, 1974, p. 187-199. This is probably *the* paper by Becker that provides a really good overview of the control system. A block diagram of the proposed system is given by Becker on p. 191.

Becker also has authored several popular-oriented books filled with excellent information. These include Robert O. Becker, M.D. and Gary Selden**, The Body Electric: Electromagnetism and the Foundation of Life**, William Morrow and Co., New York, 1985; and Robert O. Becker, **Cross Currents: The Perils of Electropollution; The Promise of Electromedicine**, Jeremy P. Tarcher, Los Angeles, 1990.

BEDINI, JOHN

Noted audio engineer, overunity energy researcher, and inventor of the renowned Bedini amplifiers well-known to audiophiles, the BASE process for holographic sound, and a process for cleaning the harshness from digital audio disks.

Bedini has invented a variety of COP>1.0 processes, motors, generators, etc. Particularly he has developed a process whereby, in a battery-powered circuit, he adroitly back-pulses the battery in a recharging direction while the battery is powering the load, causing a pile-up of electrons on the surface of the battery plates that is a sharply increased scalar potential opposing the onrushing very massive lead ions in the battery in discharging mode. The result is that the potential's electron pile-up increases dramatically due to the ions' far greater inertia than that of the electrons. The result is a high potential (perhaps 40 volts) upon the surface of the battery, overpotentializing both the incoming lead ions between the plates and the electrons in the external circuit in powering mode. In this phase, energy flows from the vacuum onto both the ions and the external circuit electrons, adding significant energy to the powering of the external circuit for the same ion current inside the battery.

As the ions slow and stop, Bedini's back-pulse trailing edge causes a very sharp cutoff, thus invoking Lenz's law and resulting in the surface potential on the plates increasing to well over 100 volts. This further overpotentializes the external circuit electrons in powering mode, dissipating much more energy in the external circuit to power the loads. As the ions finally stop and reverse, their relatively sluggish re-acceleration in charging mode and their overpotentialization still results in excess energy being delivered to the external circuit. Then as the overpotentialized ions accelerate in charging mode, the external circuit is still in powering mode until the drain-off of the severe electron pile-up on the surface of the battery plates.

This is one of several modes of operation used by Bedini in his overunity battery powered circuits, which keep their batteries recharged while powering their loads.

By separating the other closed current loop between external electron currents and the ion currents between the battery plates, and then dephasing these currents, Bedini breaks the usual dissipation symmetry enforced in a closed loop circuit where all charge carriers have the same m/q ratio. In Bedini's circuits, the m/q charge ratios dramatically differ between the ion current section between the plates and the electron currents between the outside of the plates and through the external circuit. Also, by providing a much greater potential *between* the two circuit halves, The potential extends in both directions—internally back over the ions, and externally in the circuit over the external Drude electrons powering the circuit loads and losses. He provides a true negative resistor on the surface of the plates, which extracts energy from the vacuum (via its broken 3-symmetry as a dipolarity) from the time domain, transduces it into real EM energy, and overpotentializes (regauges) the potential energy of both the internal ion currents and the external electron currents.

The Bedini process is therefore a legitimate COP>1.0 system, and in his self-powering embodiments the system will run indefinitely without external energy input by the operator, use of fuel, etc. He has converted the system to an open system far from thermodynamic equilibrium in its violent energy exchange with the active vacuum, and thus the system is permitted to exhibit both COP>1.0 and self-powering as well.

For an expanded explanation of the Bedini negative resistor process in batteries, see T. E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," **Journal of New Energy**, 5(1), Summer 2000, p. 24-38.

BEING

Pure essence, existence.

Again, something which in actuality resists Aristotelian definition unless preliminary assumptions are made. Generally used in one of two senses: (1) persistence, as in "he is (he continually exists)," or (2) identity, as in "he is the man ('he' and 'the man' are the same thing)."

BIDIRECTIONAL EM WAVES

EM waves running antiparallel to each other through a given point or region of space.

Suppose we run a normal EM wave into a phase conjugating mirror (PCM) with a gain of 1. By ordinary phase conjugate optics, the mirror then emits a backwards-traveling phase conjugate replica wave (PCR wave). According to the so-called "distortion correction theorem," that backwards traveling wave will travel backwards (as seen by the external observer) back over the exact route taken by the previous wave, and it will reappear in space precisely superposed spatially with the previous waves. If we examine the impulse (density) fields (i.e., Edt), then the forward-time wave contains Edt while the backward time wave contains E(−dt), which is just −Edt. So *electromagnetically*—as far as producing a net translation force on a charge—the E-fields seem to cancel, although they are still there and their energy is still there. Instead of translation, the two produce stress.

One thus (simple case) will have a strange kind of standing wave wherein the local energy density of the vacuum varies (e.g., sinusoidally) as a function of distance, but a charged particle placed in it will not be translated! Rigorously, that is a *gravitational standing wave*. Since change in local energy density is also a curvature of spacetime, that is also a standing sinusoidal wave of the local curvature of spacetime. This is where electromagnetics and gravitation actually meet, and where electromagnetics can be utilized to produce gravitational waves. (Much more complex waves can be produced; this is just the simplest example).

BIDIRECTIONAL FIELD FLOWS

E-fields (and B-fields) being propagated in a pair of opposite waves traveling "through" each other in opposite directions,

...where the two may be "locked together" as a wave and its antiwave; i.e., as a "Whittaker biwave pair". Note that Whittaker bidirectional EM wavepairs are not to be confused with the pairing of the wave and reflection wave on a transmission line.

BIDIRECTIONAL GATING

Gating (of energy flow, e.g.) in both directions simultaneously.

BIDIRECTIONAL TRAVELING PLANE WAVE

A wave and its antiwave, where each wave is moving in an opposite direction (standard assumption in phase conjugate physics).

In short, to the external observer the waves seem to be "moving through" each other; one from left to right, and the other from right to left; or one radially outward and the other radially inward. All scalar potentials were shown by Stoney and Whittaker to be comprised of a harmonic series of such "bidirectional traveling wave pairs." However, the interpretation of the phase conjugate waves was after detection, so that the detected "effect" wave existed in 3-space. Prior to detection, the phase conjugate half-set of the Whittaker decomposition of the scalar potential is convergent from and in the time domain (complex plane).

BIOELECTROMAGNETICS (BEM)

Presently, the scientific study of interactions between living organisms and electromagnetic fields, forces, energies, currents, and charges. The range of interactions studied ranges from atomic and molecular, through intracellular, to the whole organism**.**

Electromagnetic fields studied in BEM include those of endogenous origin (from within the organism) and exogenous origin (from outside the organism). Sources of exogenous fields include: the earth and sun, electrical equipment (power lines, radio transmitters, etc.), clinical treatment devices (e.g., TENS), and other organisms. The electromagnetic fields of greatest interest in BEM are generally classified as nonionizing (i.e., lacking sufficient energy to directly dislodge electrons from molecules or atoms.) BEM is a scientific discipline emerging on the boundaries between physics and biology. As in other sciences, progress in BEM is made by testing falsifiable hypotheses against valid measured data obtained in repeatable experiments.

Our new viewpoint seeks to dramatically enlarge the scope of this definition, to include not only the presently-recognized external (outfolded) electromagnetics, but also the internal (infolded, inside the scalar EM potential) electromagnetics, and also to utilize a unified field theory where EM is GR and vice versa.

BIOENERGETICS

That branch of Russian (KGB) energetics which is targeted against the living body, its cells and tissues, and its EM fields and potentials (biofields and biopotentials).

BIOGENESIS

Generation of living organisms from other living organisms; also, the theory or philosophy that living organisms develop only from other living organisms.

BIOLOGICAL WARFARE

Warfare against living humans, livestock, crops, etc. using pathogenic micro-organisms or biocides to create diseases and death, or using dimensioned electromagnetic potentials, fields, and waves to kindle cellular damage, disease, and death in the same targeted populations.

BIOLOGICAL WARFARE, ELECTROMAGNETIC

The use of dimensioned electromagnetic potentials, fields, and waves to induce cellular damage, disease, and death in targeted living populations.

The decades of Russian microwave radiation of personnel in the former U.S. Embassy in Moscow is an example, where only field-free potentials were dimensioned with the disease or health-change patterns (the engines). This science was developed in Russia as an outgrowth of the thousands of Kaznacheyev experiments demonstrating that any cellular disease and damage pattern could be induced from diseased or damaged cells into environmentally isolated cell cultures, purely by specialized EM radiation from the damaged or "transmitting" cells into the targeted cells.

The Chinese also used such dimensioned radiation to induce cancer in almost all GRU operatives stationed in China at the time of the major disagreement between China and Russia. In his book, Stanislov Lunev and Ira Winkler, Through the Eyes of the Enemy: Russia's Highest Ranking Military Defector Reveals Why Russia Is More Dangerous Than Ever, Regnery, Washington, D.C., 1998, p. 160, Lunev reveals it this way:*"...I was not the only person who was stationed in Beijing to come down with* *cancer… Almost all of the GRU officers stationed in China with me have been diagnosed with cancer. People assigned to the embassy had told me that they were picking up high levels of radiation from the apartment building around the embassy compound. Whether it was a form of high-tech monitoring or a conscious effort to harm our health, it seems to have been successful. So much so that in 1992 the GRU was asked to investigate the high rate of cancer among embassy officials. The findings, however, have never been made public; in part, no doubt, because of the rapprochement between Russia and China."* In short, the Chinese did exactly to the Russian GRU what the Russians were doing to the Americans.

BIOPHOTON

A quantum of angular momentum, each containing "a piece of energy welded to a piece of time, with no seam in the middle," emitted by biomolecules in living organisms. The difference between a biophoton and a normal photon is that the "infolded EM structuring" (ST curvature engine sets) carried by the biophoton carries the dynamics and templates of the living cell or tissue or body emitting the radiation.

As presently known, biophoton emission is associated with processes such as mitosis (cell division) and with mechanisms such as soliton vibrations of DNA and other alpha-helix proteins [Popp et. al., 1984; Popp et. al., 1992]. An organism's biophoton field may represent a highly-structured innate organizing field capable of encoding or transmitting information over macroscopic distances.

Kaznacheyev, e.g., showed that almost any cellular disease or damage condition can be induced from the diseased control cells to separated and environmentally isolated target cells initially normal, if (1) the visible spectrum of radiation is blocked from the targeted cells, and (2) the two groups of cells are optically connected below and above the optical spectrum, in the infrared and in the ultraviolet.

The new viewpoint seeks to dramatically extend the scope of this definition. Considering the wave aspects of the photon frequency, the infolded oscillating EM energy in the photon may be decomposed into two oscillating potentials [Whittaker, 1904]. In each of these potentials, there exists a magnificent infolded structure [Whittaker, 1903] of internal bidirectional EM wave pairs, to include the sum set [Whittaker] and the product set [Ziolkowski]. Thus the photon carries internal information, which is also referred to as the "information content of the field." Since the product set represents modulations, the basic infolded structure has signal characteristics. Mind, thought, and long term memory exist inside the personal quantum potential, which may further be regarded as consisting of coupling and uncoupling photon/antiphoton pairs—i.e., hidden gravitons. In the new viewpoint, the biophoton is the carrier of biogravitons, and also biograviton lattices. This information forms vacuum engines which can cumulate over a period of time to breach the quantum threshold and outfold, producing real change in cells, as shown by the Kaznacheyev cytopathogenic mirror effect experiments.

BIOQUANTUM POTENTIAL

A quantum potential connected to and of a living body, where the inner EM structuring and dynamics of the quantum potential are in correspondence with the inner EM structuring of the normal EM fields, potentials, and waves in the body, and also are in correspondence with the time-polarized EM structuring and dynamics of the mind and its operations.

The biopotential provides the volition and intent exhibited by the biological body, breaching the quantum threshold and producing direct psychokinetic change (by the mind-body coupling mechanism) that is the "input signal" to the servomechanisms of the body,. from the smallest to the largest. In this biopotential, the Stoney-Whittaker-Ziolkowski decomposition provides the information structure and vacuum engines that are the mind, thought, long term memory, etc. A living biological system may be defined as a physical system of otherwise inert matter (the body) and of mind, coupled to each other through a bioquantum potential.

BLOCKING

In the charge-barrier semiconductor functioning, stopping or partially stopping the flow of current in a circuit, or between two points, without stopping the Poynting S-flow.

In ordinary usage, "stopping the progress of."

BOHM, DAVID

World renowned physicist and originator of the hidden variable theory interpretation of quantum mechanics. See David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II." Physical Review, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

Together with his student Aharonov, he co-authored a fundamental paper on the Aharonov-Bohm effect wherein interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields.

BOHM'S HIDDEN VARIABLE THEORY (HVT)

A major interpretation of quantum mechanics, formulated by David Bohm and published in Physical Review in 1952.

See David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," **Physical Review**, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Bohm's theory makes all the correct predictions and also eliminates many problems in quantum mechanics, such as the "measurement problem" and the—now alarming—problem of the missing chaos (i.e., the missing "hidden order.").

Ironically, the prevailing Copenhagen interpretation of the QM theory predicts that the organized macroscopic world does not exist, while Bohm's theory predicts it. [Question to the sharp young graduate student: How can one defend a scientific model which predicts that he himself does not exist? If one keeps one's sense of humor, it would seem that this has to be the greatest scientific *faux pas* of all time!] It may be that one unconscious reason that most physicists still oppose HVT is that *it implies that physical reality can in fact be deterministically engineered*. Most physicists would seem to fervently wish physical reality to stay "solid as a brick," with fixed "laws of nature," which can be leisurely discovered and then utilized comfortably, predictably, and sanely. The advent of HVT (which is just a matter of time; for decades the Russians have highly weaponized the HVT area!) dramatically alters our perception of "fixed physical reality" into that of an "engineerable, changeable" physical reality! I tentatively believe that the terrible psychological implications of such a potential development is something that most scientists do not wish to cope with—we simply do not have the Gestalt for it. On the other hand, the Russian national psychology is such that they do have the Gestalt for it. The Russian energetics—which I have long referred to loosely as scalar electromagnetics—basically utilizes the infolded ordering inside the potential to create "direct patterns of vacuum flux", hence *vacuum engines*. And it also utilizes the quantum potential, including internally structuring the QP, to enable instantaneous action-at-a-distance in multiple simultaneous locations and nodes, with stupendous and automatic energy amplification.

By deterministically "patterning" the internal structuring of the scalar potential, one can directly form such vacuum engines. Whittaker showed in 1904 that any EM field pattern whatsoever can be expressed as two scalar potentials, rather than the common f and **A** representation. This means that any EM wave is simply the "interference" (between the hidden biwave structures) of two scalar potentials, even at a distance. By properly structuring an EM signal (i.e., structuring the two potentials into which it can be decomposed), the resulting EM signal—appearing normal on the surface—nonetheless now transports special vacuum engines deliberately designed to perform some particular internal task in mass systems, electronic systems, or humans when it strikes a target at a distance. I refer to this "internal structuring" of the Stoney/Whittaker/Ziolkowski internal biwaves and their products, as *dimensioning* of the potential(s), and hence of the EM carrier wave. The Russians call it the *information content of the field*; but our own scientists tend to erroneously interpret that phrase as ordinary spectral analysis. In other words, they erroneously continue to regard EM *vacuum engines* as if they were ordinary EM *signals*.

BOHREN EXPERIMENT

Revealing experiment by Bohren, replicated by Paul and Fischer, which shows that a given charge when in particle resonance absorbs more energy from a field or potential then the same charge does in "static" position.

By conventional outlook, this provides a guaranteed COP>1.0 process. Bohren demonstrated absorbing 18 times more energy with the charge in particle resonance than with the static charge. This is a direct experiment proving the Heaviside dark energy component of EM energy flow, whether in field form or in potential form.

See Craig F. Bohren, "How can a particle absorb more than the light incident on it"?", **American Journal of Physics**, 51(4), Apr. 1983, p. 323-327. See independent confirmation of Bohren's work by H. Paul and R. Fischer, *ibid*., p. 327. Other pertinent references are V. S. Letokhov, "Laser Maxwell's Demon," **Contemporary Physics**, 36(4), 1995, p. 235-243; — "Generation of Light by a scattering medium with negative resonance absorption," **Soviet Physics JETP**, 26(4), Apr. 1968, p. 835-839; — "Double Gamma and optical resonance," **Physics Letters A**, Vol. 43, 1973, p. 179-180; — "Stimulated emission of an ensemble of scattering particles with negative absorption," **ZhETF Plasma**, 5(8), Apr. 15, 1967, p. 262-265. Also see R. Pappalardo and A. Lempicki, "Brillouin and Rayleigh Scattering in Aprotic Laser Solutions Containing Neodymium," **Journal of Applied Physics**, Apr. 1992, p. 1699-1708

BRIDGE

A component or process which passes EM energy between two isolated circuits, but does not pass dq/dt between them, and which breaks the normal rigidly field-locked power dissipations in the two circuits.

Space itself would appear to be one such bridge.

BRIDGING FUNCTION

The function of passing EM energy flow between two isolated circuits, but not passing dq/dt between them, while simultaneously breaking any normal rigidly field-locked power dissipations in the two circuits.

BROKEN SYMMETRY

The fundamental meaning is a condition in which two parts of some configuration or shape on opposite sides of some divisor or condition or boundary, are not similar but differ.

In vacuum energy physics, we are very interested in the fact that any charge and any dipole represents a broken symmetry in its fierce energy exchange with the vacuum. This implies that something virtual has become observable; i.e., part of the virtual EM energy absorbed from the vacuum by the charge or by the dipole is changed into observable form and re-emitted as real, observable EM energy.

Any observable implies a broken symmetry. Non-observables imply symmetry, and vice versa. Each symmetry also can be represented by a conservation rule or law. Physicists have uncovered many kinds of symmetry in physics; and in the 1950s they also discovered broken symmetry.

BROMWICH, THOMAS JOHN I'ANSON

Brilliant English mathematician, 1875-1929, who made significant contributions to the mathematics of electromagnetism—particularly to superpotential theory—before his passion for overwork led to development of a mental disorder and eventual suicide.

Bromwich originated a method for solving the source-free Maxwell equations in 1899, publishing it in 1919. This approach formed the basis for his seminal paper on scattering of plane EM waves by spheres. He did excellent original work on infinite series and made contributions to quadratic and bilinear forms. He also placed Heaviside's operator calculus on a rigorous basis by treating the operators as contour integrals. Unfortunately Bromwich worked so intensively, in such painstaking detail, and at such length that his physical health was affected. He developed a mental disorder and eventually ended his own life in suicide.

For a summary of Bromwich's life and achievements, see G. H. Hardy, "Thomas John Bromwich," **J. London Math. Soc. 5** (1930), 209-220; — "Obituary of Thomas John Bromwich," **Proc. Roy. Soc. London A 129** (1930), i - x. For an exposition of Bromwich's equations-solving work, see J. D. Zund and J. M. Wilkes, "Bromwich's method for solving the source-free Maxwell equations," **Tensor (NS) 55** (2) (1994), 192-196.

BROWN, G. SPENCER

Noted author and creator of the laws of form, a new form of logic that dramatically extends Aristotelian logic.

Essentially Brown added complex numbers to logic, rather than just staying with simple 2-dimensional plane figures for postulate derivation. Electrical circuits— designed by Brown's logic—do function and do work. When computer software is designed by Brown's logic, it is then possible to mathematically prove whether or not the end result possesses any defects (bugs). In the future software—particularly the larger systems—may well be designed by Brown's logic. G. Spencer Brown, **Laws of Form**, Julian Press, New York, 1972. With a dramatic advance in logic itself, Brown has succeeded in formulating a true calculus of form. Note: There is now a second edition of the book.

BULK GRADIENT CHANGE

Simply a change in the amplitude of the entire potential,

… (i.e., to every internal biwave pair comprising it) at some point or in some region of space, as compared to changing the amplitude of only one or a few of the internal biwave pairs of the potential, without changing all the rest.

BYPASS RESISTOR

In electrical theory, a resistor to pass current by (in parallel to) some other component.

If other effects such as use of the overpotential, pumped phase conjugate reflection, quantum wells, quantum tunneling, and charge blocking are added, the function of the bypass resistor becomes extremely complicated.

CANCER

Conventional definition: A *malignant tumor* of potentially unlimited growth that expands locally by invasion and systemically by *metastasis*.

Growths are further characterized by the type of tissue in which they occur, and their actions. Of particular interest are neoplasm, tumor, carcinoma, and sarcoma. *Malignant* disorders tend to produce deterioration, infiltrate, metastasize, and terminate fatally. A *neoplasm* is simply any new growth of tissue that serves no useful function. A *tumor* is an abnormal mass of tissue that (1) is not inflammatory, (2) arises from preexisting cells, and (3) serves no useful purpose. A *carcinoma* is a malignant tumor of *epithelial* (covering membraneous tissue of a free surface) origin. A *sarcoma* is a malignant neoplasm that arises in *mesodermal* tissue (connective tissue, bone, cartilage, or striated muscle) and that spreads by (1) extension into neighboring tissue or (2) by way of the bloodstream, or (3) both.

New definition of cancer: A centrally-commanded, final, desperate, "first-step dedifferentiation" adaptive attempt by stressed, affected cells experiencing sustained oxygen shortage (hypoxia) to reverse their cellular evolution and return to the anaerobic stage of their primeval cellular ancestry. Thus a cancer cell is a severely stressed hypoxic cell that has been "ordered" by the master cellular control system (MCCS) to dedifferentiate back to a more primeval form in which it needed little or no oxygen. It is the last desperate attempt by the body to insure the affected cells' survival when all other means available to the MCCS to provide the necessary oxygen have been exhausted and have failed. The "order" differs from normal signal theory in that it consists of a Stoney/Whittaker/Ziolkowski template inside the personal quantum potential of the organism. Thus the "order" actually consists of a vacuum engine, so that the spacetime in which the affected cell exists is acted upon by structured, energetic hidden variable fingers to directly alter it. The "signal order" is in fact a set of hidden vector forces in spacetime itself, created directly in and on the affected cell(s), in every part, in every atom. The cure for the cancerous cell, once formed, is simply to cause the MCCS to generate the "counter-order" reversing the original dedifferentiating action. This can be readily accomplished by establishing hidden multifrequency pump waves in the quantum potential on the cancerous cell, on all its parts, so that they act as pumped phase conjugate mirrors. Since they already exist in the altered "cancer" spacetime template/signal, the pumping causes a phase conjugate replica of the precise cancer state to be created in and on the cell and its parts. [In fact this is a general therapeutic mechanism, and it can be applied to reverse almost any conceivable cellular damage or disorder.] This counter-order is a new and amplified vacuum engine that acts upon the cell to redifferentiate it back to a normal cell. If an abnormal number of cells results, the body's normal processes then recognize the excess cells and destroys them.

CANCER HOTSPOT

An area or locality or location where cancer recurs repeatedly, more than the normal deviation expected statistically. In such case, one is dealing with a hidden cause or carcinogen.

Human systems exposed to such a hotspot experience cellular changes (hypoxia) which eventually result in the master cellular control system (MCCS) issuing a dedifferentiation order to the affected cells to regress back to a primeval anaerobic form. The first step is dissociation from central (MCCS) control of cellular growth, resulting in a cell that now grows and divides unrestrained by the normal body. The real question in the hotspot is, "What is the unknown agent inducing the hypoxia condition, and what is the mechanism by means of which it induces it?" Epidemology will point out a hotspot, but is usually incapable of resolving the primary causative agent or its active mechanism since it can only show correlation.

CAPACITOR AS A TRANSMISSION LINE

A capacitor can be theoretically treated as a special type of transmission line, a fact which is apparently known to transmission line theorists but not to many engineers and physicists not specializing in transmission lines.

CARCINOGEN

A substance or agent (such as a chemical or a spacetime curvature engine) capable of inducing a cancer.

In the new theory, the primary carcinogen is an actual order (a vacuum engine) issued by the master cellular control system (MCCS), reverting (dedifferentiating) the affected cell to a primal anaerobic form. Usually what are normally regarded as carcinogens are also causes of cellular conditions wherein the cell becomes very hypoxic, so much so that the MCCS exhausts all other alternative corrective means and as a last desperate measure orders the cell to change to a form that requires little or no oxygen. The "exhausting of other measures" may be very rapid, as when the cell has been affected and damaged by powerful carcinogens, or it may require decades to occur, as when the cell is subjected to long-term mild hypoxia, such as from electromagnetic smog.

CARRIER WAVE

A fundamental wave which is modulated by another wave or other waves, and hence "carries" the other waveform(s). By stripping off the carrier in a demodulator, the carried waveform(s) emerges.

Hence it is a continuous frequency wave which can be modulated by an information-bearing signal.

CASIMIR EFFECT

The attraction of two conducting parallel plates in space, placed very close together, due to their influence on the active vacuum and on the vacuum's interaction with the plates.

CASIMIR, HENDRIK BRUGT GERHARD (1909-2000)

Noted scientist who in 1948 predicted that two parallel conducting plates, placed very near each other in a vacuum, would experience an attractive force due to their influence on the electromagnetic vacuum.

See H. B. G. Casimir, “On the attraction between two perfectly conducting plates,” presented at a meeting of the Royal Netherlands Academy of Arts and Sciences on 29 May, 1948. Published in the same year in **Proceeding. Koninklijke Nederlandse Akademie van Wetenschappen**, Amsterdam, vol. 51(7), 1948, p. 793-796. For a beautiful modern experiment confirming the force, see S.K. Lamoreaux, “Demonstration of the Casimir Force in the 0.6 to 6μm range,” **Physical Review Letters**, 78(1), Jan. 6, 1997, p. 5-8.

The force was experimentally detected some 10 years later.

CAUSAL SYSTEM ROBOTS (CSRs)

Massless 4-space "engine" systems (therefore in the causal domain only, hence the name) formed of longitudinal EM waves and their dynamics, which are therefore sets of spacetime curvatures and their deterministic dynamics (engines) which function as individual robots.

This is a totally new kind of system leading to a totally new kind of warfare, developed over some decades and deployed by the KGB in 1999.

Matter is composed mostly of empty space filled with EM waves, fields, and potentials. Those "fillers" themselves are just bundles of longitudinal EM waves and their impressed dynamics. Hence all matter, the vacuum, and all conventional EM potentials, fields, and waves—being filled with vast sets of longitudinal EM waves—are vast superhighways for longitudinal EM waves to travel on or in. The CSRs, being totally LW and LW dynamics in nature, thus are free to travel right through EM radiations from power lines, phone lines, TV stations, radio stations, etc. They easily penetrate any matter, including right through a mountain. Any function of a normal EM system can be performed—at least in theory—by such robots. This includes longitudinal EM wave interferometry, which forms external EM fields and waves, of whatever pattern or form desired. Further, these CSRs are actual functioning systems, and if communications functions are designed in, one can communicate with them at a distance, using longitudinal EM wave communication, right through the intervening earth and ocean, etc.

For some time, extensive underground facilities—hundreds—have been building in the former Soviet Union, and this massive buildup continued unabated after the Russian economy collapsed and the Berlin Wall came down. This tremendous expenditure, when the Russian economy was crippled, shows the strategic significance of these facilities. Further, apparently our satellites have not seen any missiles, normal weapons or munitions, etc. associated with that vast complex. In short, it is suspected that we don't have the foggiest notion of what it is. Indeed, those facilities will already have sent some CSRs into those very satellites, to insure that we do not find out what is going on, and perhaps to "spy on the spy satellites".

We think these terribly important facilities are the KGB-controlled facilities for developing and producing the CSRs, and contain a vast longitudinal EM wave command and control set of facilities as well. In short, this is the strategic arsenal, development, and control for building sets of various type CSRs in huge numbers—a vast new army of robots with unprecedented capabilities, under the command and control of a human "strategic force" operations command of very large size.

Developing and debugging one CSR would be a formidable and lengthy task. Once a single one is developed, making a million copies is simple. Just take a normal CD-ROM, e.g., and record something on it. Send in a working CSR robot to "reside inside" the infolded EM interior of the recorded informational electromagnetic pattern on the CD-ROM. Then simply copy that CE-ROM magnetically, as many times as is desired. Each copy contains a copy of the original signal that was recorded, including the exact infolded electromagnetic structuring inside that signal. Voila! One has as many new duplicate CSRs of that exact type, as one wishes, for a few pennies each.

One can see the implications. In all those underground facilities, there are many development laboratories, production laboratories, and command and control facilities and "storage facilities" for the robots. Those hundreds of facilities are in fact a great new strategic weapon system and all its development labs and production facilities, as well as its tactical control and command and control systems, its "troops", its "equipment", and its commanders and operators.

Well, then one can deploy and "utilize" such a strategic armada just as one would use any other great strategic army. One "deploys" one's forces against the enemy. In this case, one just sends the armada (individually controlled!) right through the "LW superhighways" of matter itself, right into the strategic territory and very heart of the targeted nation. One deploys the robots everywhere—e.g., inside the electronics in nuclear warheads, inside the control systems for the nuclear submarines and nuclear bombers, inside the satellites, inside the fighter aircraft electronics, inside the fighting ships, inside the main battle tanks and major weapons carriers, inside the signals on the hard drives of computers, etc. One sends them right through the "inside" of the targeted nation's power grid electrical energy signals, through the TV stations' emitted signals, etc. The Pentagon and Cheyenne Mountain are duck soup, as is the stock exchange and other civilian systems vital to continued functioning of the infrastructure. So is everywhere else, even the deepest vault and "special security" facility.

In short, the vast electronic command and control activities of an otherwise "modern" superpower provide great, convenient, easy superhighways for full strategic deployment to be completed by millions of weapons systems (CSRs) into everything of value in the targeted nation or nations. This completely solves the usual weapons delivery problem; delivery is made in advance of D-Day.

Such a vast CSR armada was deployed throughout the Western world, and particularly in the United States, just before the end of 1999. Yes, this was to be the KGB's great millennium gift to America—a really enormous Y2K problem that would stun and disable the entire nation, its armed forces, its electronic systems in its major weapons, and its command and control centers in one great TOT (time on target) assault. The major electrical power systems of the nation would have been disabled, as would have been the major communications systems. Using LW interferometry, the robots would have wreaked havoc on the nation. They would have disabled the electronics of every major C4I computer of importance, every major fighter aircraft, every major fighting ship, every nuclear submarine, every ICBM, every nuclear powerplant, etc. The attacking force was in its deployment and employment sites in the targets themselves. Everything was in place and cocked and ready to fire. To test whether we knew what was going on, the KGB even had one or more robots temporarily produce hacker-type operations and signals "emerging" in some U.S. command and control center computers, and deliberately let our fellows discover this "hacker attack" was coming from the Russian Academy of Science. Then the CSR simply stopped its LW interferometry, which caused the "normal hacker electronic" signals to just vanish and never reappear. Little wonder that our counter-type fellows were totally unable to find any trace of it, or where it was residing. They have no technology for such.

Well, happily those early-model CSRs are easy to root out and destroy, if one has quite advanced LW technology. Visualize these beasts as existing in a special sort of "internal sky" or "infolded atmosphere" inside matter. Strategic range LW "radars" can just sweep through this "internal atmosphere" inside matter, etc., detecting and tracking those CSR beasts analogous to how normal radar detects and tracks aircraft. The difference is that the strategic LW can also use a bit of LWI pulse to instantly snap those CSRs like popping bugs, completely jamming their operations or even "erasing" them once their characteristics are observed.

That very new kind of vast strategic warfare occurred right at the end of 1999, all through the United States—the Pentagon, our strategic weapon systems, the works. The friendly little nation rooted out and destroyed all those "on site" CSRs just prior to the attack being launched.

A full-up strategic war, of previously unknown kind, was fought all over the U.S. and our own system guys apparently didn't know it. It didn't even surface. Didn't make a whimper.

That hostile CSR armada consisted of *first generation systems*. These systems did not have defensive countermeasures. The systems were depending on total surprise, but were found out (that is another story that would make a great movie if one could live to film it) and then were destroyed in short order.

CSR development has ushered in a totally new kind of strategic, operational, and tactical warfare. The delivery of all the weapon systems "into the target" is accomplished long before the advent of hostilities. The second generation CSRs, once built, will be even more sophisticated. These will almost certainly be on line soon. They will likely also possess countermeasures.

An eerie new chapter in the annals of warfare has been initiated, and the world will never be the same again. Once the CSRs have sufficient defensive countermeasures such as stealth means and coordination with countering, rapid response LWIs to target and destroy the hostile "LW radars" seeking them, etc., the next world-wide strategic CSR war will be neither so simple nor so bloodless.

NOTE: Contrast this CSR discussion to the discussion of *nanotechnology* and nanobots. Just as nanobots are natural developments which teams skilled in nanotechnology would produce, so CSRs are natural developments which teams skilled in longitudinal EM wave technology—and in the "infolded" electrodynamics inside all EM fields, potentials, and waves—would produce. For decades (since the 1950s) the KGB has been developing and using longitudinal EM wave interferometers (LWIs) which beam right through the "inner highways" inside the fields and waves inside matter, thus allowing very large LWIs which beam right through the intervening earth and ocean, and can "focus" under the ocean or within the earth as well as out of it. With such background, the CSRs were just as natural a development for them as the nanobots are for conventional Western nanotechnology scientists.

CAUSALITY

The relationship between "cause" and "effect". The basic notion in a causal system is that the system's response to an input signal is not dependent upon future input values.

Thus the notion of causality involves a precise time ordering of ordinary inert interactions, where the cause was never input from a future time. Macroscopic (observable) causality may be and is violated by the interaction of hidden subquantal (virtual state) variables.

CEREBRAL HEMISPHERE

Each half of the cerebrum, i.e., each half of the enlarged anterior or upper part of the vertebrate brain.

The two cerebral hemispheres may function as a longitudinal EM wave interferometer or a time-polarized (scalar) EM interferometer to sense distant scenes (even distant in time), or produce EM energy and signals at a distance, since the body, brain, and nervous system does produce longitudinal EM waves and scalar EM waves. For proof of scalar EM wave interferometry in O(3) electrodynamics, see M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," **Journal of New Energy**, 4(3), Special Issue, Winter 1999, p. 76-78.

CHANGING THE PAST

Every intentional inception upon ordinary inert causality prior to collapse of the wave function represents a slight alteration or change in direction, departing minutely from pure causality. From a causal viewpoint, each human inception minutely changes the causal past, as well as the future.

Experimental proof that the causal past can be changed is provided by the delayed choice two-slit experiment. E.g., see C.W. Rietdijk, "Another proof that the future can influence the present." **Foundations of Physics**, 11(9/10), 1981, p. 783-790. Particularly see Wheeler, John Archibald Wheeler, "The 'past' and the 'delayed-choice' double-slit experiment," in A.R. Marlow [Ed], **Mathematical Foundations of Quantum Theory**, Loyola University, New Orleans, Louisiana, June 2-4, 1977; Academic Press, New York, 1978, p. 9-48.

CHAOS

In dissipative dynamical systems, the dynamical evolution that is aperiodic and highly dependent upon initial conditions of the system.

The trajectories of the system move on a strange attractor, which is a fractal subspace of the phase space.

The term suggests the ordinary meaning of random and unpredictable disorder, but there exists intrinsic determinism and at least partial embedded ordering. The mathematical equations describing chaotic behavior are very nonlinear and so complex that at the present time they cannot be computed or predicted. Even simple systems, including simple electromagnetic systems, can exhibit unpredictable chaotic behavior. The science of chaotic dynamics is still in its infancy, having begun moving once the modern computers became economical and pervasive.

CHARGE, ELECTRICAL (Q)

Electrical charge q is defined as $q \equiv m_q\phi_q$, to first order.

As can be seen, the actual "charge" action associated with the mass is due totally to ϕ_q, and to the broken symmetry of ϕ_q in the vacuum flux exchange with m_q. The charge q can be further broken down into a set of composite dipoles if the gathering of a virtual charge screen in the vacuum, surrounding the mass of the "bare charge" inside the gathering, is included. At first order, ϕ_q can be expressed (particle view) as a change in the local vacuum virtual photon flux (VPF), due to the VPF exchange between vacuum and m_q. The ϕ_q component is actually the "electrical charge" and massless of itself.

However, see my paper, "Giant Negentropy of the Common Dipole," **Journal of New Energy**, 5(1) Summer 2000, p. 11-23. We include the isolated charge by treating it as a system of multiple composite dipoles. We also explain the previously unrecognized 4-symmetry EM energy flow ongoing, where incoming longitudinal EM waves are being received, transduced into real longitudinal EM waves in 3-space, with the latter being emitted in all directions in 3-space.

A charged particle of mass is thus a little energy flow generator. It is a broken 3-symmetry in the local VPF, and therefore has an "observable" energy flow output, extracted and gated from the asymmetry in the VPF. This "observable" energy flow outputs the well-known energy flow **S** in classical EM, though CEM does not recognize or include this vacuum interaction, the broken 3-symmetry of the dipole, and the steady input of EM energy to the dipole in the form of EM energy from the time domain (complex plane).

Instead, classical EM just assumes that the "charge" q is the source of a potential. When originally formed as a concept, q was just a "quantity of electric fluid," like a "cubic centimeter of fluid," etc. The atom and the electron had not yet been discovered.

CHARGE CARRIERS

Anything—usually fundamental particles, ions, etc.—which consists of or contains charges and can move.

It can be an electron, a collection of electrons fixed in a moving dielectric, ions, atoms, molecules, etc.

The moving charge (i.e., the current dq/dt) with an energy flow streaming onto it and on beyond, carries its interaction with that energy-flow right along with it in the current
i = dq/dt. This is what the charge's "excess collected energy" consists of, and this is how the *collected/collecting* energy in the circuit is transported to the *collected-energy* dissipation sites (loads and losses).

CHARGE, GRAVITATIONAL

To be added

CHARGE, MAGNETIC

Essentially, the magnetic scalar potential, or monopole.

In an uncurved spacetime, the magnetic charge must always occur in dipolar pairs, where each is equal and opposite the other. In a curved spacetime, a free *net* monopolar magnetic charge may and does exist. In classical electrodynamics, the assumption of a locally flat spacetime is ubiquitous though not prominently emphasized. Hence most electrical engineers are unaware that monopoles can indeed be made, but only if one uses a process to curve local spacetime.

CHARGE-PARITY-TIME (CPT) THEOREM

The theorem states that all known quantum field theories obey CPT symmetry.

C is charge conjugation, which changes the quantum number of every particle into its antiparticle. P is the reflection operation known as parity which turn an object into its mirror image and rotates it 180 degrees about an axis perpendicular to the mirror, T is time reversal. Any violation of time-reversal symmetry is always accompanied by a violation of CP symmetry and vice versa.

Electrodynamics and the strong interaction preserve CP symmetry to great accuracy. Weak interactions do not conserve CP symmetry. T symmetry violation has been shown at CERN, the European particle physics laboratory in Geneva, Switzerland.

CHARGE TRAPPING (Also "Pinning")

Holding charges "pinned" or "trapped" by a barrier or force, so that they do not flow as
i = dq/dt.

CHARGE-BARRIER

A process or component or function—particularly in a semiconductor such as the Fogal semiconductor—which blocks the movement of charges q as current dq/dt.

CHARGE-BLOCKING ASYMMETRIC QUANTUM WELLS

A quantum well which has more charge-blocking action in one direction than the other, and is thus "asymmetric" in its function.

It may also be dynamic, so that it changes in its degree (and/or its direction) of action as a function of time.

CLASSICAL AND QUANTAL ELECTRODYNAMICS

Classical EM (CEM) theory is simply classical electrodynamics (CED).

Classical EM does not break the fields into "quanta" (a quantum is a standard-sized piece of action (angular momentum)]. Instead, CEM considers the fields as continuously varying in magnitude from zero to infinite magnitude (at least in theory).

Quantal Electrodynamics is known as quantum electrodynamics (QED). These are standard terms for standard disciplines. QED considers the fields to be composed of quanta called "photons". Hence there is an underlying region called the "virtual state" whose denizens are "smaller in magnitude" than the quantum. Heisenberg uncertainty applies, since the quantum of action is not quantized energy, but quantized (energy x time). In short, it is a function of both the magnitude of the energy and its duration of persistence. Hence a field or photon of any amount of energy may form temporarily, so long as it does not endure sufficiently long in observer time to reach quantum size and breach the quantum threshold.

In this view, then, "empty space" is not empty at all, but is a seething cauldron of these "energetic bubbles" continually appearing and disappearing, with the entire spectrum of frequencies and energies. This is in fact the "new ether". If we consider only the field portion of the bubbling virtual photons appearing and disappearing, this leads to a vacuum that consists of EM field fluctuations, again of essentially the entire frequency spectrum and energy spectrum.

However, how we attempt to tie either of these "energetic vacuum" EM views to general relativity (GR) is then a problem. The cause of the problem, however, is largely ignored in physics. In CED, e.g., one models EM energy moving through a flat space. This is a non sequitur, since the presence of any change of the energy density of space a priori is a curvature of spacetime. Hence rigorously EM energy can only propagate through a curved spacetime. Indeed, when this correction is made, the EM wave becomes an oscillating curvature of spacetime. This is the approach worked out by Mendel Sachs, extending Einstein's work. The higher symmetry O(3) electrodynamics spearheaded by Evans turns out to be an important subset of Sach's umbrella unified field theory. Hence EM has become GR, and for the first time in history, O(3) electrodynamics provides the approach that will allow direct engineering of general relativity on the lab bench, in working devices, etc.

CLASSICAL ELECTROMAGNETICS (CEM)

"Ordinary" electromagnetics, of the kind used in ordinary EM circuits, which fundamentally consists of the modified Maxwell's equations.

Classical EM consists of Maxwell's equations and the various changes and additions that have since been made. CEM does not break the fields into "quanta" (a quantum is a standard-sized piece of action (angular momentum)]. Instead, CEM considers the fields as continuously varying in magnitude from zero to infinite magnitude (at least in theory).

Primarily refers to Maxwellian electrodynamics as reinterpreted by Heaviside, Gibbs, Hertz, and a few others. Unfortunately these reinterpretations of Maxwell's theory also greatly reduced its topology, and therefore reduced the area of natural electrodynamic phenomena that the resulting equations describe.

Significantly more advanced "classical" electrodynamics emerges when one goes back to the quaternion algebra electrodynamics of Maxwell, or to the more advanced Clifford algebra electrodynamics. The higher symmetry electrodynamics models such as O(3) symmetry EM developed by Evans and others, are capable of modeling a great deal more physical phenomena than the very limited standard U(1) electrodynamics largely in vogue in science.

CLASSICAL EM THEORY

Classical electrodynamics theory, begun by Maxwell, with his seminal paper orally presented in 1864 and published in 1865.

Essentially Maxwell's electrodynamics at root basis, with the background supporting classical theory. Today many variations have been made, particularly with the basic symmetry of the theoretical equations. See discussion above under *classical electromagnetics*.

CLIFFORD ALGEBRA

A special higher topology algebra founded by Clifford, which includes as subsets many other algebras of lower topology.

CLOSED SYSTEM

A system that does not communicate with its environment, and does not exchange energy or matter between system and environment.

An ideal model where dynamically and energetically the system is considered to be "isolated" as if nothing else existed. In short, consider the system as if it were in some "magic box" and nothing outside the box can ever enter the box or affect the system inside the box, and nothing in the system can ever leave the box or affect anything outside the box.

One performs a non sequitur whenever proclaiming treatment of a system as a "closed system", and of course good scientists are aware of this. This is particularly true in electrical power systems. Here one is indeed allowed to input energy into the system to excite or potentialize it. That of course assumes that the system was "opened" long enough to take on excess energy! Then as the system operates, losses and dissipations occur, in which case energy is considered to "escape" from the system permanently. In other words, the system was also continually "opened" so that energy could escape.

There is really no such thing as a truely *closed* system in the universe, since every system is embedded in the active vacuum and is an open system in an energy exchange with the vacuum. If that exchange is symmetrical, then the system is in equilibrium with respect to the vacuum exchange. In that case, the system may be treated as if it were a closed system for a limited number of purposes. In classical electrodynamics (CED), the Lorentz regauging of the Maxwell-Heaviside equations further reduced them to describe only Maxwellian systems in such equilibrium with the active environment. Consequently, the interaction between the external vacuum and the system is not included in the CED model. Obviously a *broken equilibrium* in that actual physical interaction is also not included in the CED model.

An electrical power system is already permitted to violate the Lorentz condition when it
(1) receives additional EM energy to "energize" or "potentialize" or "excite" or "regauge" it, and (2) whenever energy escapes from the system in the loads and losses (which escapes of energy are "de-energizing", "depotentializing", "dissipating", or "regauging" actions.

Hence the Lorentz condition greatly simplifies the Maxwell-Heaviside equations and makes them much easier to solve. However, *when enforced by the system itself with regard to the system's excitation and de-excitation (dissipation) functions—i.e., when those functions are "symmetrical" (equal and opposite)*—it also arbitrarily discards all those Maxwellian systems capable of the five "magical" functions of an open disequilibrium system: (1) self-ordering, (2) self-oscillation or self-rotation, (3) output of more energy than the operator inputs (the excess energy is input by the active environment, (4) powering itself and its load (all the energy is input by the active environment, and (5) exhibiting negentropy.

It follows, however, that physical electrical power systems must therefore have some feature which self-imposes the Lorentz condition, at least during the phase when the system's excitation energy is discharged in the loads and losses. The standard closed current loop system, where every electron progressing through the external circuit's loads and losses must be forcibly rammed back up through the source dipole in the power source, is indeed what enforces the Lorentz condition with respect to excitation and dissipation, and *arbitrarily* prevents the system from achieving COP>1.0.

In a battery-powered system, the otherwise closed current loop is broken into two quite different circuits: (1) the electron current circuit between the outer surface of the plates and the external loads and losses, and (2) the ion current in the electrolyte between the inner surface of the plates. Since the ions (e.g., in a lead-acid battery) in the "inner" ion current may have a m/q ratio very much greater than that of the electrons in the "outer" electron current, clever timing of pulsed signals to the plates may be used to form a phase difference—including even a phase angle of 180°—between the two currents. It can be shown that this forms a high potential on the plates which acts as a negative resistor, extracting EM energy from the vacuum and recharging the battery while simultaneously powering the external circuit. This is the Bedini process, and with Bedini's permission an explanation of that process has been published by Bearden ["Bedini's Method For Forming Negative Resistors In Batteries," **Journal of New Energy**, 5(1), Summer 2000, p. 24-38.] Note that Bedini's process in fact repeatedly removes the Lorentz regauging condition, and therefore the system is rigorously permitted to achieve not only COP>1.0 but also self-powering of itself and its loads and losses.

COLD EXPLOSION

The sudden extraction of EM heat energy from a distant intersection zone (IZ) of the crossing beams of a scalar potential interferometer, which transmits in the sharply pulsed mode, while the electrical grounding of the transmitters are negatively biased with respect to the potential in the distant IZ.

COLD FUSION

An ad hoc term applied to the transmutations at low energy achieved in electrolyte experiments, particularly when using specially prepared palladium electrodes, and particularly when deuterium is present in the electrolyte.

COMPOSITE DIPOLE

Dipole formed as one part of an isolated, observable "charged particle", consisting of a differential element of that observable charge on one end and a momentary clustering virtual charge of opposite sign.

In quantum electrodynamics, it is well-known that virtual charges of opposite signs cluster in the immediate vacuum around any "isolated" observable charge. By applying the composite dipole concept, this author was able to treat the isolated charge as a set of composite dipoles. By then applying Whittaker's decomposition of the scalar potential between the ends of each of these composite dipoles, the "source charge" was a set of broken 3-symmetries in the vacuum flux, and thus a set of 4-symmetries in that flux exchange. This meant that the EM energy continuously output in all directions by the source charge was actually continuously received by the source charge (as a set of composite dipoles) from the time-domain, and 4-conservation of EM energy flow rigorously was upheld.

This allowed a solution to what has been called the most formidable problem in both classical and quantum electrodynamics: the problem of the association of the source charge and its associated fields and potentials along with the energy contained by them and continuously pouring from the charge without any 3-space energy input.

CONCOMITANT

Occurring or existing together.

CONGLOMERATE HIERARCHY OF THE PHOTONS

Every photon remains perfectly ordered! A conglomerate (group) of photons, however, may be arranged in almost an infinite variety of orderings, groupings, directions, etc.—which we refer to as templates and as hierarchies (group orderings) of photons.

In a pure energy flow (energy transport), the "herd" of photons is organized along the flow direction. Regardless of how one scatters or "disorders" this patterning of the "herd" or "hierarchical ordering" of photons, not one single bit of the underlying energy and individual photon ordering is lost. So when one "uses" energy (e.g., scatters the hierarchical ordering of a photon herd from a resistor as heat), one does not lose any energy at all, but simply has "scrambled" the hierarchical ordering of the photons into a new herd configuration (which may be varying every which way!). The point here is the real meaning of the fundamental conservation of energy law: *Energy can neither be created nor destroyed!*

When we "use" one joule of energy to perform one joule of work (say, in a single-pass, scattering), we still have exactly that one joule of energy left, but just in different form (different direction, type, collection or not collected, etc.) It can still do another single pass joule of work, then another, then another, and so on. If we plan things right, we can use a single joule of work to perform many joules of work, contrary to what we were so strongly taught in the *"single-pass, single collection, single dissipation"* examples given to us in university as the "law" of God and electrodynamics.

Note that the conventional work-energy theorem assumes single pass, single collection, and single dissipation of a joule of energy to perform a single joule of work. The point is, *that is not a limitation of nature, but a limitation of a single process! Nature regularly uses a joule of energy to do multiple joules of work, because after one dissipates or scatters the energy the first time, it's still there and later will be intercepted to do some more work.*

Anti-Stokes emission, the Patterson Power Cell, Lawandy's lasing without population inversion, and some gas-filled tubes with anti-Stokes emission are examples of known, validated overunity processes (processes with coefficient of performance greater than unity). These all use multipass, multicollection, multi-dissipation in iterated manner, to increase dipole asymmetry and therefore produce excess extraction of energy from the vacuum. Patterson's unit, e.g., has been independently measured by universities at efficiencies of some 1200 or so. This absolutely does not violate the laws of physics nor the nonequilibrium thermodynamics of open systems far from thermodynamic equilibrium.

CONSCIOUS MIND

The serial processing mind which we normally associate erroneously as ourself. Other to be added.

Mind is time-like, not space-like.

CONSCIOUS MIND (PSYCHOENERGETICS VIEW)

The serial-processing mind which one normally associates erroneously as oneself.

Mind is *time-like*, not *space-like*. Hence it is not directly observable, a priori, since all observation is 3-spatial. However, if we consider that changes in 3-space produce spacetime curvatures and thereby also produce slight changes in the time domain, then iterative coherent spatial changes eventually will produce a change in the time-domain. Consequently, the time-like mind will have a continuing stream of such small but processable time-energy changes from coherent 3-spatial phenomena in its body and the macroscopic (coherent) universe. To sense the 3-space body as "self" while simultaneously sensing other 3-space physical changes as "non-self", the conscious mind must compare the various received time-like projection signals against its "time-like orders" issued. That set of received time-signals that matches or closely matches the conscious mind's issued order signals, thus yields the sense of the body behavioral changes as (1) self, and (2) body responses to its orders. The time delay between issuance of the mind's orders and the receipt of the correlated time-like responses from its body, provides the sense of self-persistence or independent being.

The mind couples to the body via a similar spacetime curvature mechanism. When the mind issues "orders", they are in the time-like realm but are changes, hence small spacetime curvatures. They therefore involve very small 3-spatial changes (projections) into the 3-space body. Since the mind orders are coherent, this small iterative set of 3-space changes will integrate coherently, yielding real quantum changes in the body and nervous system (e.g., the brain).

This coupling mechanism is profound. The steady stream of integrated time-like projections arising in the mind due to the physical world outside one's body, does not compare or match the mind's memory of its orders. Hence that allows the sensing of the "external world" as "nonself". The sense of that external world as "existing" or "being" is provided by the time delay in the comparison process, from "mind signals issued" to "physical signals received". The metric yardstick standard for this comparison time, so to speak, is provided by the time in the identified "self-loop" with the physical body. In other words, the mind uses that length of time to differentiate those time signals which coherently integrate but do not match its issued order signals, from those time-signals which coherently integrate in the same time loop interval and do match its issued ordered signals.

In this manner, we resolve all the great previously unresolved philosophical questions of the nature of consciousness, nation of being, nature of self, nature of nonself, perception of the physical world, perception of body as self, perception of other bodies as non-self, etc. At least we resolve these issues from a testable physics approach, so that we can see that now a unified physics (and a technology) of mind and matter is at last in view. Since O(3) electrodynamics has been shown to be an important subset of Sachs' unified field theory, then O(3) electrodynamics—or SU(2)×SU(2), which is homomorphic with O(3)—can be used to develop the direct engineering of mind, the mind-matter interaction, and mind functions of the living biological system.

It is my considered opinion (with much evidence) that the Russian branch of *energetics*, known as *psychoenergetics*, uses a unified field theory and a higher symmetry electrodynamics to achieve that type of direct engineering already, but for weapons use rather than beneficial use for humanity.

As examples, we strongly suggest that the deaths of U.S. Captains Button, Svoboda, and Hess all involved KGB testing of long-range strategic psychoenergetics weapons over the United States, on those targeted personnel. This should not be surprising at all; the KGB has tested longitudinal EM wave interferometer weapons over the U.S. since 1967, and continues to do so today. They tested bioenergetics EM weapons against U.S. Embassy personnel in Moscow for several decades. Kills of U.S. aircraft and missiles have been performed at various times, including kill of the Arrow DC-8 at Gander, Newfoundland in Dec. 1985, kill of a Titan ICBM fired from Vandenberg AFB in the spring of 1986, kill of the U.S. space shuttle Challenger in the spring of 1986, continuing weather engineering over the U.S. and elsewhere, etc.

That novel electromagnetic weapons have indeed been developed and are being tested world-wide, was confirmed by Defense Secretary Cohen in 1997 as follows:

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." [Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.].

We specifically call attention to the fact that the Secretary did not speak the words enclosed in square brackets. This is *not* limited just to terrorists or counterterrorism.

There appears to be no insurmountable difficulties in developing psychoenergetic technology using the approach outlined above. Certainly it is not as formidable a task as has been hot fusion.

The incident involving Captain Button, e.g., was to demonstrate control of an individual performing complex technical tasks for a period of at least one hour. The test against Captain Svoboda was to demonstrate how simply human kill can be achieved in combat troops, merely by introducing selected changes in the mental functioning of the targeted individuals. As Svoboda was climbing her A-10 out of a successful low-level ordnance drop at night on a ground target, her sense of up and down was instantly reversed. Instantly perceiving she was diving instead of climbing—although *actually* climbing out perfectly—she "corrected" immediately and dove headlong into the ground, perishing in the resulting explosion of her aircraft.

In the case of Capt. Hess, the test demonstrated that rather complete mental control can be almost instantly asserted, and even the responses of the autonomic nervous system can be overridden. So Captain Hess, out for a morning jog, simply went aside, sat down, and stabbed himself repeatedly, including multiple times right through the heart. Had one seen the incident, he would have simply continued stabbing himself very calmly, without emotion or reflex action from his autonomic nervous system, until physical failure of his body.

As can be seen, it is of the utmost importance that this nation understand psychoenergetics weapons of such nature, and develop effective defenses against them. The next great strategic battlefield may be as much a "virtual reality" battle as it is a "real physical world" combat. The war can be quickly lost in either arena.

We also point out that the psychoenergetics weaponry readily is adapted to use in more advanced longitudinal EM wave interferometers, which readily penetrate the Earth and the ocean. Hence deeply buried strategic command and control centers are vulnerable to psychoenergetics warfare, as are the crewmen of our nuclear submarines, ICBMs, aircraft carriers, division and army command centers, national command centers, strategic bomber pilots, etc.

Indeed, one of the major incentives for developing such psychoenergetics weapons as was demonstrated in controlling Captain Button, was to seize and control the entire on-site manning crews of the defensive quantum potential weapons preventing launch of full strategic warfare. By having the QP systems "stood down for maintenance" by their own crews, the major strategic retaliation and "dead man fuzing" would have been removed and countered. Then powerfully striking the QP sites with LWIs would have destroyed them, and left the sites themselves radioactively emitting longitudinal EM waves, thus preventing reoccupation in case the QP weaponry survived. The counter to that scheme was to place the friendly QP weapon sites on "insane" command and control. In that mode, instead of human decision being made to launch, every so often (say, every two hours) the computer automatically initiates the launch sequence. During that sequence, if the crew is still alive and functional, codes can be punched in and the computer cancels the launch initiation. It then repeats the sequence two hours later. If no coded cancellation order is punched in, for whatever reason, the launch sequence completes and the QP weapons fire. In that case, much of Russia disappears from the face of the earth.

Eerily, just about the time that Secretary Cohen was making his quoted statement in April 1997, the KGB was being notified by the friendly little nation of the change to insane command and control. This forced the abortion of a great strategic attack upon the West, scheduled to occur on May 1, 1997.

We are strongly stressing that psychoenergetics is not parapsychology and it is not just an idle theory of academics. Instead, it is already potently used in some very powerful and deployed weapon systems that could well decide the course of the next world war, almost at a moment's notice.

See also discussion under *energy*.

CONSCIOUSNESS

Refers to a living being's awareness of its sensations, feelings, thoughts, and the world around it.

The new approach uses a coupling mechanism between the time-like mind and the body of a living creature. The nature of this coupling mechanism then generates the mechanism by which self-awareness is produced, awareness of non-self is produced, and the sense of existing in the external world but differing from it is produced.

CONTINUOUS MODE

For a longitudinal EM wave interferometer, a continuous mode of operation where power is fed continually to the transmitters, and the two longitudinal EM wave patterns or beams are continuously transmitted.

CONVENTIONAL SCALAR POTENTIAL

A scalar potential formed without use of artificially added substructuring of its internal biwaves and their products.

In other words, one in which it is assumed that its virtual particle flux has no ordering, but is just a randomized flux. This assumption (e.g., of quantum mechanics indirectly) is incompatible by the Whittaker demonstration that (i) a scalar potential is a harmonic set of "bidirectional" EM phase conjugate wave pairs, and thus not a scalar entity at all, and (ii) perfect ordering of its wave components exists inside the scalar potential.

One must carefully distinguish what each of these internal "phase conjugate EM longitudinal wavepairs" actually represents. Prior to interaction with detecting charges, the potential is a 4-potential existing in spacetime (4-space). The time domain, of course, is represented by ict, and in that expression only "t" is a variable. Hence the "incoming" phase conjugate longitudinal EM waves in the imaginary plane are actually longitudinal EM waves in the time dimension. These waves converge upon the dipolarity represented by the "scalar potential" or any "piece" of it. The dipolarity represents the interaction of charges; specifically, a dipole. This reaction—by the spin of the charges—transduces the incoming EM energy from the time-dimension into 3-space, and emits it from the charges as real EM longitudinal EM energy radiating out from the dipolarity in all directions in 3-space.

The phase conjugate half-set of the Whittaker waves, *prior to their interaction with charge*, are not waves in 3-space, but are waves of LW EM energy coming in from the imaginary plane and thus from the time domain. The charges of the dipolarity spin 720°, being 360° in the complex plane and then 360° in real 3-space. So the charges absorb the EM LW energy impinging on them from the time domain, during their 360° rotation in the complex plane. When they enter real 3-space for the rest of their spin, the already-excited charges dissipate their excitation energy in all directions as they spin their 3609º in 3-space. That is what the Whittaker decomposition actually shows, but it has been ignored for almost a century.

In short, the broken 3-symmetry of the dipolarity—well-known in particle physics for nearly a half-century, but never incorporated into CED—allows nature to jump to a higher, more fundamental 4-symmetry in the energy flow. Energy is being conserved in 4-space, but not in 3-space. Further, there is a special new energy flow symmetry involved: one-way in the time domain and the other way in the 3-space domain. Note that energy is not conserved in 3-space! *There is no law of nature requiring the extra requirement that, in addition to conservation in 4-space, the energy must also be conserved in 3-space!* So when the dipole breaks that extra requirement, only the basic 4-conservation applies. In that case, nature very happily gives you what might be called a white hole in astrophysics terms: energy will pour in to the dipolarity unceasingly from the time-dimension, and will pour out in 3-space unceasingly, so long as that dipolarity persists.

By considering the virtual charges of opposite sign that cluster in the vacuum around any "isolated observable charge", one may represent a single charge (as an electron or proton) as a set of composite dipoles, where each dipole consists of a virtual charge (while it exists) and a differential element of the observable charge. So the charge becomes a set of composite dipoles, each with Whittaker decomposition of the scalar potential between its ends.

Hence we have solved the long-vexing problem that Semiz called "the greatest unsolved problem in electrodynamics": the problem of the source charge (or source dipole) and its associated fields and potentials and the enormous EM energy continually pouring out of any charge or dipole in all directions in 3-space.

When the arbitrarily discarded giant Heaviside nondiverged energy flow component is also accounted, this leads to the giant negentropy of the dipole, as shown in my paper, "Giant Negentropy from the Common Dipole". We have also proposed this "dark Heaviside energy" as the unaccounted source of the excess gravity known to be present in spiral galaxies and holding together their arms. The dark energy yields at least 90% of the total gravity involved, and the additional 10% or so can be accounted for my known gravitational sources.

It appears that these new discoveries can also be utilized to show that "quintessence" is also caused by COP>1.0 processes in the universe, and the excess production of negative "giant dark energy flows". If so, then this also would account for the recently discovered fact that the universe has an excess of negative gravitation, so that it is not only expanding but is accelerating in that expansion. At this time, the present author is still attempting to tie up the loose ends involved in showing that quintessence also is created as a result of functions of the giant negentropy process.

In the Whittaker biwave pairs, If one then insists that these hidden EM waves are comprised of hidden photons, then it follows that the photons exist as continually coupling and uncoupling photon/antiphoton pairs, or continually forming and unforming *gravitons* as the wave and antiwave continually pass through each other.

It is a little more complicated than that, since Maxwell actually omitted half the EM energy in the EM wave in space, and also in the circuit—as do electrodynamicists to this day. For gravitons, one really must come to grips with this error and correct it. This will be added later in this glossary. (To be Added).

COOPER PAIR

A dynamic pairing of electrons in superconductivity theory.

In this pair, if the energy state with wave number σ and spin ½ is occupied by an electron, then so is the state with wave number $-\sigma$ and spin – ½.

COP (COEFFICIENT OF PERFORMANCE)

Ratio of energy out, divided by that portion of the total energy input that is input by the operator or experimenter.

The COP is thus a measure of "efficiency of using the operator or experimenter's input energy" to produce—or direct the production of—useful work. Note that he can direct the use of his own input energy, as almost all our present systems do. Better yet, he can also direct (gate or "switch") the use of some external free flow of energy from the environment into the system, so that he gets lots more work out of the system than he has to put in himself. Only a few of our systems—such as solar cells, sails on sailboats, waterwheels to power mills, and windmills—are such Maxwell's demons (open systems) where the output exceeds the operator's input of energy.

We sharply contrast the term "COP" to the term "efficiency". The efficiency (overall) of a system is how much useful energy output (usually as work, or can be converted to work at will) the system produces for the total amount of energy that is input (both by the operator and by the external environment). Even a highly inefficient system may nonetheless have a COP>1.0, if the operator inputs less energy than the system outputs.

We urge the reader to thoroughly understand the difference between energy-use "efficiency" of a system and coefficient of performance of the system. Many scientists, engineers, and especially free-energy researchers are confused on the precise difference in these terms.

CORPUS CALLOSUM

The thick nerve cable that links together the two cerebral hemispheres (which are separate brains) in each vertebrate (including human) brain.

In each of the two cerebral hemispheres, there is a separate mind and personality (given that the other is separated). However, a brain has a unique characteristic: Anything that arises directly inside it is automatically assumed to be self-originated, since the brain-tuner can differentiate no separation or "separate source" for the signal/thought arising directly upon its "internal screen". This allows integration of the two brains/minds/personalities into one. That is, when one brain half prepares and sends a message, that message is also routed across the corpus callosum into the other brain half, arising internally upon the second brain's "display." The second brain half therefore thinks that it itself originated the signal or message.

COSMOLOGICAL FEEDBACK PRINCIPLE

Puthoff's self-regenerating cosmological feedback cycle for the source of the vacuum EM ZPE.

One may assume the existence of EM zero-point energy (ZPE) by fiat as part of the boundary conditions of the universe, or conceive of its generation by the quantum-fluctuation motion of charged particles that constitute matter. Puthoff calculated the latter possibility, assuming that the ZPE spectrum (field distribution) drives particle motion, and that the particle motion in turn generates the ZPE spectrum. This provides a self-regenerating cosmological feedback cycle, which in fact is consistent with the general relativity assumption that curvature of spacetime affects mass energy, and mass energy changes affect the curvature of spacetime.

The result of Puthoff's calculation is the appropriate frequency-cubed spectral distribution of the correct order of magnitude. His result is thus consistent with, and indicates, a dynamic-generation process for the ZPE fields.

In general this is also consistent with the Sachs' unified field theory approach, and with the Evans O(3) important subset of it.

COURRIER, ROBERT

Noted French scientist who collaborated with Antoine Prioré in his revolutionary healing experiments on laboratory animals using longitudinal EM waves from plasma tubes.

Courrier was the head of the Biology Section of the French Academy and also the perpetual secretary at the time. He personally presented the astounding results of the Prioré work, showing cures of terminal tumors, to the assembled French Academy, creating a furor.

CROSSTALK

In ordinary science, the transfer of energy or signal from one channel to another, by cross modulation or cross coupling between the channels.

In the new approach, the term "crosstalk" can be used in an additional sense where it refers to virtual energy exchange between orthogonal universes or frames—that is, between different 3- or 4-dimensional slices of an infinite-dimensional universe.

CURL CONCEPTS OF FIELDS

Refers to concepts of fields as the curl of some vector potential;

E.g., the magnitude of the **B**-field is equal to the curl of the **A**-potential, given by $\mathbf{B} = \nabla\times\mathbf{A}$. More exactly, the **B**-field is identically a curled **A**-potential, by $\mathbf{B} \equiv \nabla\times\mathbf{A}$, but only that portion of $\nabla\times A$ that is diverged around or collected by an intercepting unit point static charge.

Note that we are accenting that no equation is a definition. So texts stating that the first equation with an equals sign "defines" the magnetic vector potential **A**, are in error. If we state that $\mathbf{B} \equiv \nabla\times\mathbf{A}$, then we are defining the **B**-field as a curled form of the magnetic vector potential, which it is.

The reader must be aware that the field concept itself in CEM is flawed, being "defined" as the effect existing after an interaction by a 4-space entity with a 3-space mass, and hence
3-spatial and an "effect", and yet then used as if it were the 4-space "cause" entity that propagates in 4-space to interact. In short, the field concept is thoroughly confused in the texts and papers in the literature (and has been confused for more than a century) as not only the cause but the result of the interaction of the cause with some recipient. This problem has been highlighted by many physicists, including Feynman, Wheeler, etc.

This "confusion of the cause with the effect" is widespread in physics, and is the single greatest foundations error which hampers the progress of physics. Another example is the false notion that a "separate force" acts upon a "separate mass", when in fact mass is a component of force by the definition $\mathbf{F} \equiv d/dt(m\mathbf{v})$. That such simple and obvious errors have been propagated by so many hundreds of thousands of scientists for so many decades is perplexing. Indeed, the errors simply continue to be used. Note that Nobelist Feynman and the great John Wheeler pointed out that the "field" as defined in CEM does not and cannot exist in spacetime! Instead, Feynman pointed out that only the potential for the field exists in spacetime, in case some charged mass is brought in to interact with it and *thereby produce* the field as an effect.

Indeed, no observable persists, for it "exists" only as a momentary 3-space frozen snapshot or slice of a 4-space dynamic interaction. Here again, most physicists do not realize that observables such as mass m do not and cannot persist other that at the instantaneous frozen moment of the snapshot.

CURL-FREE MAGNETIC VECTOR POTENTIAL

A field-free magnetic vector potential without curl, hence with no **B**-field.

Toroids and long solenoids have the characteristic of separating the **B**-field (curled potential) and holding it inside. However, drawing energy from any potential whose "source charge or source dipole" is not destroyed, simply results in the full value of the potential being replenished from the vacuum, via the 4-symmetry energy flow mechanism in the presence of the broken 3-symmetry of the dipole. Hence one may easily "double" the potential energy density at will, and this is a special form of gauge freedom, where the potential energy of any electromagnetic system can be altered freely and at will. We strongly accent that, contrary to what is in the EM textbook, gauge freedom rigorously allows the direct and free amplification of potential energy of the EM system at will. That this principle, incorporated in gauge field theory, has been known so long and still not used in the design, production, and use of COP>1.0 electrical power systems is quite inexplicable.

In the motionless electromagnetic generator, we used a special core material to extract the curled potential (**B**-field) from a permanent magnet's vector potential **A**, so that **A** is replenished in uncurled form directly from the vacuum via the giant negentropy process.

To achieve COP>1.0, we then simultaneously interacted both the uncurled **A**-potential energy external to the output coil, and the internal curled **A**-potential energy (**B**-field energy) internal to the core through the coil, with the coil. By pulsing, the Lenz law effect is also invoked as yet another momentary regauging and potential energy increase, so that an additional energy gain is achieved.

The well-known Aharonov-Bohm effect proves that the uncurled **A**-potential can indeed be separated from the **B**-field in like manner, and the **A**-potential does interact with electrons since $d\mathbf{A}/dt = -\mathbf{E}$.

Thus the MEG achieves COP>1.0 permissibly, because it is an open system freely receiving excess energy from the active vacuum environment, via the vacuum's replenishment of the magnetic vector potential of the permanent magnet dipole as fast as energy is drawn from it. The MEG also violates the usual closed current loop's self-imposition of Lorentz symmetrical regauging when discharging its excitation energy (its free excess regauging potential energy), because the "source dipole" (the permanent magnetic dipole) is not destroyed.

CURVED SPACETIME (externally, internally)

A four-dimensional geometry used in general relativity, where the curvature is determined by the distribution of mass-energy.

In the new, more extended approach, any change in the spatial energy density of spacetime or any change in the time-energy density of it, or a combination of changes in both. The difference is that in the new approach time is also treated as a special form of energy. Specifically, time may be regarded as spatial EM energy compressed by a factor of c^2.

CYBORG (An acronym for cybernetic organism).

(Noun): A single integrated system consisting of a living system energetically or functionally linked to mechanical or electromagnetic devices and/or systems performing some vital functions.

(Verb): Link, in the sense of forming such a system.

In the new approach, a living biological system may thus be considered a cyborg of its time-like mindworld (mindframe) to its 3-spatial body (laboratory tuner). The "cyborging" is provided by the personal quantum potential and by the mind-body coupling mechanism. Due to the giant negentropy mechanism of any dipolarity, any scalar potential as a dipolarity is a cyborg between observable reality and virtual state (time-like) reality, i.e., between 3-space and the time-domain.

CYTOMEGALOVIRUS

Any of several members of the herpesvirus family that cause infected cells to enlarge and/or to form cytoplasmic (the protoplasm of a cell, external to the nuclear membrane) inclusions in leukocytes (white blood cells), particularly in their nuclei.

Cytomegalovirus makes a protein that appears to disable the p53 protein. As a result, smooth muscle cells in blood vessels can proliferate until they impede blood flow.

CYTOPATHOGEN

Literally, a pathogen (disease causing organism—or extended, mechanism—) that affects cells and generates damage in them.

CYTOPATHOGENIC

Damaging, harmful, or destructive for cells.

CYTOPATHOGENIC EFFECT

The generation of damage, disease, death, or other symptoms at the cellular level in cells that are exposed to or subjected to a pathogen.

CYTOPATHOGENIC MIRROR EFFECT

Distant EM pathogenic effect noted by Kaznacheyev in thousands of military experiments, where cellular death, disease, and damage forms are shown to be transmitted from one cell culture in the dark (absence of visible light) where the changes are induced, to another similar targeted cell culture in the dark that is environmentally shielded from the first except for light transmission in the infrared and in the ultraviolet. After exposure begins, after a time delay (say, 2 hours or more), the cells in the targeted cell culture will begin to exhibit the same symptoms as the culture having the induced damage. Eventually, in most cases the targeted cells will develop the same damage, disease condition, or death. The effects vary with the time of day etc.

The necessity for photon connection from IR to UV is shown by contrasting a thin quartz window allowing the light transmission with a thin window glass window. The quartz is open from the IR to the UV, and will pass that entire spectrum. With the visible light absent, the IR and UV biophotons emitted by the damaged cell culture will pass through the quartz window into the targeted cell culture. After the time delay, the targeted cell culture starts developing the condition.

Use of a thin window glass window, however, prevents the effect. Normal window glass is opaque in the IR and in the UV, and so it blocks the transmission of the IR and UV biophotons.

The UV is a harmonic of the IR, arranged properly. An unusual effect exists when two sine waves one the first harmonic of the other, are simultaneously transmitted into an isotropic nonlinear medium, such as water. On each sine wave, the peak travels faster in the water than off-peak parts of the wave does. Hence the wave peaks overshoot its middle, the wave collapses, breaks up, etc. Both the waves of the transmitted harmonic set do that. However, if one pretends that one transmitted the difference frequency between them, and not the waves themselves, a strange phenomenon results. Regardless of the breakup of the primary harmonic set, the difference frequency travels through the water as if it were a real frequency in a linear medium. This effect has been proven and is used in underwater sonar equipment.

For example, see Owen Flynn, "Parametric arrays: A new concept for sonar," **Electronic Warfare Magazine**, June 1977, p. 107-112. Any two sine-wave frequencies as simultaneous drivers combine to produce a sine-wave difference frequency propagating in water, essentially without sidebands or reverberations. Its pattern has a main lobe approximately equal to that of the high frequency drive, but devoid of sidelobes. The level of the propagating difference frequency is proportional to both the product of the two fundamental drive levels and to the square of the desired value of difference frequency.

Living cells contain a great percentage of water. Hence the effect should apply to those targeted cells, under the conditions stated.

From general relativity, we know that any pattern and dynamics of mass-energy, such as in a cell and all its parts and dynamics, directly interacts with local spacetime to curve it, and any set of curvatures of spacetime interact with any conglomerate of mass-energy and its parts and dynamics. The set of precise curvatures and their dynamics we call an engine. In a mass embedded in a local region of spacetime containing an engine, the engine will interact with the mass-energy to change it if the mass-energy and its dynamics do not "fit" the engine and its dynamics.

In a diseased, damaged, or dead cell, the mass-energy of that cell and its "template" (form and dynamics) differs from its template while the cell was living and normal. Let us now see what happens when that dead cell emits a biophoton.

A photon emitted from a dielectric such as a damaged cell, is not emitted just from one spot. Every part of the dielectric participates in the emission of that photon. (See "porthole concept", to be added). Further, each emitted photon contains in its scalar potential and its magnetic vector potential, the exact longitudinal EM waveset that constitutes the vacuum engine of the damaged cell that emitted it. When a cell absorbs a photon, it absorbs that vacuum engine carried by the photon also. If it absorbs a stream of photons of harmonic interval, the difference frequency constitutes in terms of a wave, traveling through the cell water as if in a linear medium, and transporting that vacuum engine to every part of the absorbing cell dielectric.

So the biophotons emitted from Kaznacheyev's damaged, diseased, or dying cells sent biophotons containing the exact spacetime curvature engines and dynamics of their precise condition (template). Further, a harmonic interval of such radiation is transmitted into the targeted cells, and absorbed. The mixing of the two absorptions eventually is sufficient to establish the harmonic interval, thus with a clear EM signal difference that travels through the cell's water and to every tiniest part of it. This "difference frequency clear signal" also contains the pure vacuum engines of the template of the changes in the diseased, dying, or damaged emitting cells.

The result is the "kindling" (coherent integration) of the template in the absorbing cell's local spacetime in which it is embedded. As that spacetime engine grows, it differs from the resident engine of the absorbing cell. That difference is itself an engine, and it operates steadily upon that absorbing cell to alter its engine and then its physical condition and state.

As the new engine continues to interact, the physical cell is gradually changed into an equilibrium condition in its mutual interaction with its local spacetime engine. That occurs after the absorbing cell has developed the same disease, damage, or dying condition as the transmitting cells.

This is the secret of the Kaznacheyev experiments. The Kaznacheyev work was allowed to be released in the open literature to stimulate the U.S. scientific and intelligence communities and see whether or not the scientists and analysts had applied a unified field theory and higher symmetry electrodynamics to the EM bioeffects problem. The totally casual reception of the Kaznacheyev experiments in the West, with no interest or interaction from the stimulated communities, assured the KGB conclusively that the West was still in kindergarten in its knowledge of EM bioeffects, and particularly had little or no knowledge in the field of EM biological warfare.

Kaznacheyev referred to the effect as the "cytopathogenic mirror" effect, for the pathogenic effects from the emitting damaged cells being "mirrored' in the targeted cells.

See Vlail Kaznacheyev and L. P. Mikhailova, **Ultraweak Radiation in Intercellular Interactions**, [in Russian], Novosibirsk, 1981. An English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," **Psi Research**, 1(1), Mar. 1982, p. 47-76. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," **Psychoenergetic Systems**, 1(1), Dec. 1974; Vlail Kaznacheyev *et al*., "Distant intercellular interactions in a system of two tissue cultures," **Psychoenergetic Systems**, 1(3), Mar. 1976, p. 141-142. In the same issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144.

Seven or eight nations now have longitudinal EM wave interferometer weapons, capable of beaming right through the earth and ocean to the other side of the planet. At least two of these nations—Russia and China—are capable of adding engine structures to these interferometers, and sweeping a targeted population area to generate that specific disease condition in that target. A sophisticated use would be to have agents (such as terrorist teams already in the United States and with biological agents in hand) release a small quantity of smallpox in the area previously targeted. The LWIs have swept the entire area continuously for a period, slowly kindling (especially at night) the exact form of the disease within the bodies of the populace. A curious thing happens if the level of kindling of the smallpox state has been kept just beneath the observable state, within what is sometimes called the "shadow state" (just barely virtual). A smallpox pandemic will explode in that populace with a ferocity never seen in modern times, and at an unparalleled rate. This is further exacerbated by "spreading the immune system" prior to the attack (term and discussion to be added), which seriously weakens the immune system's ability to resist any single pathogenic attack.

There are hardly any supplies of smallpox vaccine left that are viable. Even where there are, the Russian laboratories developed strains of smallpox resistant to the normal vaccine, and after the collapse of the Russian economy and the former Soviet state, these strains were sold on the black market to the highest bidder. Further, most of the Soviet scientists employed in those secret biowar facilities wound up employed in several other nations, most hostile to the United States.

The point is this: Such a strike could in fact kill up to 2 billion people worldwide. Further, the entire affair would appear to be just a chance breakout of smallpox, and its spread through a nonresistant world populace.

When modern biological warfare is actually analyzed in detail, the "bang for the buck" is greater than it is for nuclear weapons. Further, a rogue or hostile nation can take its time in advance and develop quantities of smallpox vaccine (for the very strain it will use), then initiate such a clandestine campaign after it has vaccinated its own populace, armed forces, and agents. Guess who survives after such an attack.

CYTOTOXIC

Toxic (acts as a poison or toxin) to cells.

A *toxin* is a colloidal proteinaceous poisonous substance that is a specific product of the metabolic activities of a living organism. The toxin is usually very unstable, and is notably toxic when introduced into the tissues.

DARK ENERGY

The Heaviside nondiverged energy flow component, surrounding every field/charge and potential/charge reaction but unaccounted in present electromagnetic theory after being arbitrarily discarded by Lorentz.

DATUM

Consideration, or data considered, or information considered, or action considered, or result considered.

DEATH

The disintegration of the personal quantum potential's connection to the physical body's atomic nuclei by loss of the mind-body connection mechanism.

Simply put, separation of the quantum potential containing the time-like mind-frame from the body. Loss of the preferential tuning of a physical body to its mindworld. This causes loss of the preferential consciousness loop or life channel, resulting in "death" (reversion to inert matter) of the body. Actually the quantum potential and the mind-frame remains, so the *being* is unchanging and indestructible, hence consists of all changes and destructions simultaneously, with none separated individually. Life and death thus are two sides of the same coin, and the coin is immutable, regardless of which side is exposed.

DE BROGLIE WAVE

A matter wave, conceived by Louis de Broglie.

The quantum mechanics wave associated with a particle of matter, which can give rise to interference effects. An orbital electron in an atom is associated with a standing de Broglie wave on a Bohr orbit. De Broglie waves always move faster than the speed of light, the speed being given by c^2/v, where c is the velocity of light in vacuum and v is the velocity of the particle of mass. De Broglie wave velocities thus vary from the speed of light to infinite speed (everywhere at once). In his speech accepting the Nobel Prize, de Broglie emphasized that, since these waves produce real physical effects, they are real and must not be regarded simply as mathematical conveniences.

DEBYE, PETER

Dutch-born American physical chemist and Nobelist, 1884-1966, who performed important and fundamental studies in bipole moments and diffractions of X-rays and electrons in gases.

DEDIFFERENTIATION, CELLULAR

The reversion of a cell or cellular structure to a more primitive or generalized condition.

Dedifferentiation is often preliminary to a major change in the cell or cellular structure. It can also be a reversion of the cell to a previous—even *primeval*—cell form. The important point is that dedifferentiation returns the cell to a previous (past) cellular state. Reversion of a cancer cell, e.g., would cause the cell to return to a normal cell, since that is the nature of its "cellular past." It is vitally important that Becker has demonstrated that cells can be differentiated (moved forward to a future state) or dedifferentiated (moved backward to a previous state) electrically, by very weak DC currents (picoamperes). He was nominated for a Nobel Prize for this epochal work.

Actual electromagnetic redifferentiation of terminal tumors in laboratory animals was repeatedly demonstrated by Prioré et al in the late 60s and early 70s and reported in the peer-reviewed French medical journals, as was reversion of depressed immune systems, cures of infectious diseases, and curing of arteriosclerosis. It is inexplicable that this positively demonstrated cure of cancer—which could also cure infectious diseases and probably can be developed to cure AIDS—was suppressed in the mid-70s and has not been vigorously pursued by present establishment medical science.

DEGENERATE SEMICONDUCTOR

A semiconductor whose conductivity approaches that of a metal.

DELGADO, JOSE M. R.

Noted Spanish neurophysiologist well-known for his research on brain behavioral mechanisms and their control by electrical stimulation; who implanted electrodes in the brains of animals and humans and demonstrated that their responses (both behavioral and emotional) were specific and depended upon the specific localized part of the brain that was stimulated.

A TV tape made by Delgado showed a spectacular demonstration where he stopped a charging bull in mid-stride. With a small hand-held radio transmitter, he transmitted a radio signal to a receiver implanted in the head of the bull. This dramatic demonstration caused a world-wide sensation.

The significance of his work is given by Becker: *"Delgado discovered the exact parts of the brain in which electrical stimulation produced fear, anxiety, pleasure, euphoria, or rage in human subjects. He found that certain sites produced major personality alterations. For example, stimulation of one such site could cause very proper, reserved young ladies to become flirtatious and sexually aggressive. Other sites inhibited maternal or aggressive behavior. In short, Delgado was able to profoundly alter human behavior through electrical stimulation of discrete areas of the brain." … electrodes were implanted in the pleasure centers of rats' brains. The animals were provided with two levers to press. One of these delivered food, and the other produced electrical stimulation of the pleasure center. Invariably, the rats chose to press the lever that stimulated the pleasure center, even to the extent that they died of starvation*." (Becker, **Cross Currents**, p. 221-222).

It is known based on a rare OMNI interview that Delgado progressed to detailed study and experimentation for the behavioral and emotional effects in monkeys of specific magnetic field frequencies on monkeys. These successful experiments involved electrical currents produced in the brain but hundreds of times lower in intensity than those required for direct electrode stimulation. Since then, some information has been cautiously released from Western military studies showing that microwave pulses can couple to the central nervous system and produce stimulation and behavior alteration similar to Delgado's released experimental results.

Comment: The ultimate and most refined techniques in the field of electromagnetic control of human behavior, volition, emotion, etc. is provided by Russian psychoenergetics, the military version of which is totally under KGB control. This work has progressed far beyond what Delgado and Becker were discussion so long ago. The use of the infolded longitudinal EM waves and "engines"—once the technology is developed—allows direct engineering of conscious and unconscious mind, memory including short term and long term, perception, awareness, etc. The additional advantage of using EM longitudinal waves is that mass is largely a very thin dispersion of particles, with (relatively speaking) very large distances between them, somewhat like the solar system but on different scale. This "inner space" is filled with EM waves, potentials, and fields—all of which decompose into sets of longitudinal EM waves and their impressed dynamics. Thus the earth and ocean are readily penetrated by longitudinal EM wave beams, and the Russians (KGB) has developed such LW interferometers since the mid 1950s. However, by LW interferometry where the patterns and dynamics of the carried LWs are engineered, any desired mapping pattern of precise EM signals, frequencies, phases, and magnitudes can be reproduced in distant interference zones, reaching right through the ocean. These interferometers could localize the distant interferometry into a 6-inch diameter volume as early as 1985, so from that time forward it gradually became possible to perform all of Delgado's "inserted electrode" EM stimulus in a single selected human brain from the opposite side of the earth. So approximately 1985 is when I would estimate the beginning of the "golden age" (ugh!) of distant psychoenergetics was initiated.

By 1997, this development would be into at least third generation weapons, just being deployed or just having been deployed, considering an somewhat accelerated weapons development program where 6 years is one generation. By 1997, very complex individual behavioral and perceptual control should have just arrived in on-site deployed strategic weapons systems. See discussion under *conscious mind*.

Some selected useful references are: J.M.R. Delgado et al., **Journal of Anatomy**, vol. 134, 1982, p. 533 reports that developmental defects occur in chick embryos exposed to various ELF frequencies. See also J.M.R. Delgado, **Physical Control of the Mind: Toward a Psychocivilized Society**, Vol. 41, World Perspectives, Harper & Row, New York, 1969.

DEMODULATION

From a carrier containing a modulated signal, recovering the signal that was used to modulate it.

DENSE SIGNAL ENVIRONMENT

For our purposes, when the number of weak EM signals crossing one square meter per second reaches 200,000.

At 500,000 important nonlinear optical type effects of significance—particularly long term cumulative effects—begin to occur in the exposed biological organism.

DEPARTING ORTHOROTATIONS

In space of dimensions greater than four, the rotating of an object or entity away from its line of motion, in the direction of a dimension at right angles to the line of travel in the observer's three-space, by some number of additional 90° rotations.

The Lorentz transform in special relativity, e.g., is just a formula for this kind of rotation. What we call "velocity" is actually a measure of rotation. In 4-space, any object in 3-space can only rotate "out" of 3-space in one direction: toward the time-axis. As the object rotates toward that time-axis, we see the object as having a velocity in 3-space, and its "length along the line of motion" as having diminished. When the object rotates the full 90°, its "length along the line of motion" has been converted to "time" rather than length. Hence it has zero length along the line of motion and is moving at c, the speed of light in ambient space. It is also no longer material, but is a sort of flat plane surface moving at speed c in a direction perpendicular to the plane.

In that 4-space model, it is impossible for anything to move faster than light speed. So Einstein's postulate dealing with the fact that it is impossible to translate mass, signals, etc. faster than light speed has a built-in assumption that only one orthogonal departing rotation can be made from 3-space. That postulate need not apply at all if more than one departing orthorotation is permitted—which of course requires a space of 5 dimensions or greater.

The notion is that, if we allow a spacetime of greater than four dimensions, by continuing to rotate ever more orthoturns in a departing direction away from the original line of travel, the rotated object loses one observable dimension for every orthorotation. Hence with three "departing" orthorotations it becomes a "nonobservable point" to the normal observer, but one that would be moving at speed c^3. A single orthorotation would make the 3-dimensional object a two-dimensional plane moving at speed c.

No matter how many rotations are made in a departing manner, the object is still connected to the time dimension common to all dimensions (assuming only a single time dimension for simplicity). So it exists in an orthorotated space, but in the same time. To the physical observer in 3-space, it thus appears to be a "time-like" or "mental" object only. Nonetheless, by affecting the density and structuring of the time domain, it can have an effect on a 3-spatial object. We discuss that mechanism under the subjects of the mind-to-body coupling mechanism and the body-to-mind coupling mechanism. Further, with the Sachs-Evans unified field theory approach, the time-domain can be directly engineered. This means that both "inert" or physical reality and time-like or "mental" reality can be directly engineered by novel electrodynamic means.

In short, a photon is just a particle (such as an electron) that is orthorotated once (toward the time axis). In photon emission, a little bit of the mass-energy turns into orthorotated photon energy. In photon absorption, the absorbed photon energy turns back into mass-energy.

The higher dimensional orthorotational concept was originally formed to allow a physical modeling grasp on a "mental" object as compared to a "physical" object, and to have a connection mechanism between the two. Such an approach can be used to model both mind and matter, their interaction, etc. The present author has used it to solve the old philosophical problem of "intent", of how the perception of self is attained, how the external world is perceived separate from self, etc. A model for both living and nonliving states does emerge.

This or a very similar approach has been used by the Russian's highly classified weapon science called *energetics*. That third branch of energetics is called *psychoenergetics* and it is the physics of directly engineering mind states, mind operations, memory, etc. by extended electrodynamic means. We have previously detailed much of this work privately, but with some detailed openly in a few papers. E.g., see T. E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", **Explore**, 9(2), 1999, p. 59; Part II, **Explore**, 9(3), 1999, p. 61; Part III, **Explore**, 9(4,5), 1999, p. 100-108.

DESTRUCTIVE INTERFERENCE

When two waves or wavesets from two different sources superpose to produce a combined waveset that has a lower intensity than the sum of the intensities of the two original waves.

DETERMINISTIC PATTERN OR TEMPLATE

The exact arrangement, by plan and deterministic action, of a group of entities.

E.g., of the photons in a "herd" or "group" of them. Or a set of spacetime curvatures, into a dynamic structure called a "spacetime curvature engine"—or "engine" for short. Or by the virtual photons in the vacuum flux, or a deterministic set of changes to that vacuum flux (the vacuum flux that is called the "vacuum potential." I have called this function—of a deterministic patterning or templating of the virtual flux (particle view) of the potential—*dimensioning* the potential.

DIELECTRIC

Literally, "against electric (current),"

...originally coined to mean "opposes the flow or conduct of electric fluid."

In today's language, a dielectric is a non-conducting material used to oppose or prevent the flow of electric charges. Even empty vacuum exhibits some dielectric resistance. "Dielectric" is also used to refer to a nonconducting material (often used between the metal plates of a capacitor or elsewhere) which can (i) sustain a change in potential across a distance (i.e., an **E**-field), and (ii) serve as an insulator. Space itself is a dielectric, or at least it is said to exhibit certain dielectric qualities.

DIFFERENTIATION, CELLULAR

(1) specialization of body parts or organs in the evolutionary process. (2) modification of different parts of the body to perform particular functions. (3) modification of the cells forming these body parts or organs so as to perform the necessary cellular functions required to support the functioning of the body part or organ.

DIFFRACTION

The bending of light at the edge of an opaque object.

Usually diffraction introduces undesired distortion and is a problem. However, it can be useful and desirable—for example, in a diffraction grating.

DIMENSION

A certain primary geometrical physical attribute, such as length, used to describe the separational relationships of physical phenomena.

The degree of that separation is known as "length", "time", etc. By "geometrical" we mean that the "dimension" is considered to "exist" in either the presence or absence of the observable physical phenomenon; however, by understood agreement the dimension itself is not directly observable. It can be inferred ("measured"), however, by comparative operations of simple devices such as clocks and meter sticks, and its simplistic definition is usually given as "that which has measurable magnitude." Prescribed measuring procedures are used to assign magnitudes to these inferred dimensions, using specific scales.

Actually, "dimension" just means "fundamental quantity of separation" and "geometry" in the most general sense. The number of dimensions taken in advanced physics usually represents a somewhat arbitrary choice for good modeling fit. In modern Kaluza-Klein theory, for example, it becomes necessary to accept some 10 or 11 dimensions in the usual particle physics case. General relativity has incorporated the interaction of the abstract geometry (spacetime) with mass. Until Sachs' extension of Einstein's work into a unified field theory and Evans' incorporation of O(3) electrodynamics as an important subset of Sachs' unified theory, general relativity has largely remained a nonexperimental discipline, at least in the laboratory. With the new Sachs-Evans approach, however, GR becomes substantially engineerable by novel electromagnetic means, including in the laboratory and in devices.

DIMENSIONING

This is an artificial term I sometimes use to mean "forming an exact pattern or 'template', particularly of nested curvatures of spacetime, where the pattern/template may also be dynamic. It involves the internal structuring of the "dimensioned" entity.

The use of the term "dimensioning" makes sense in unified field theory, but is awkward in normal EM theory which erroneously considers that EM energy propagates in an uncurved spacetime. The mere presence of a change in local spatial energy, created by the wave energy itself when present in that local region, curves the local spacetime. Hence the EM wave rigorously travels in curved spacetime, and in fact identically is a propagating, oscillating curvature of spacetime.

Dimensioning also implies a deliberate form or structure (with dynamics) created in the virtual photon flux of the vacuum or of the virtual photon flux comprising a scalar potential, or in the infolded longitudinal wave EM inside all EM potentials, fields, and waves. It implies an associated set of spacetime curvatures and their dynamics. The latter are referred to as "engines", "vacuum engines", "spacetime curvature engines", etc.

Dimensioning potentials or signals or carrier waves also involves adding hidden *vacuum engines* to them. I chose the descriptive term *dimensioning* because adding such templates (hidden vacuum engine structures) is equivalent to increasing the EM topology by adding dimensions. Since an EM wave or field or vector potential can be decomposed into two scalar potentials, then by "dimensioning" one or both of the scalar potentials one can "infold" the desired hidden structures or patterns (vacuum engines) inside an EM carrier wave, EM field, or vector potential. So I called this *dimensioning* those entities. Internally structuring the two scalar potentials via arrays of EM LW emitters, and then applying the necessary dynamics (modulations), is primarily the manner in which dimensioning can be directly engineered. When dimensioning is present, two identical EM waves as seen on the oscilloscope may have dramatically differing internal structures (dimensioning), and hence the two will produce drastically differing effects when they interact with the same or similar objects. This of course is the mechanism involved in the Kaznacheyev experiments and in the decades-long microwave radiation of the U.S. Embassy in Moscow, where health changes and diseases were electromagnetically induced in personnel in field-free (stable and nonchanging potentials) areas.

Russian energetics refers to what we are calling "dimensioning" simply as adding *the information content of the field*. A very skilled disinformation program (actually, a strategic deception program) has been used to deceive the West into believing the phrase refers to ordinary spectral analysis. It doesn't. The importance of the dimensioning concept is that one has not formed mere "signals," instead, one has formed *vacuum engines*. If a receiving object receives a signal, it may or may not act upon that "signal information." If it receives a *vacuum engine*, however, the engine (structured local spacetime flux) directly acts upon it to make the requisite change. Further, the energy for the interaction's working upon the exposed object or system is furnished continually from the curved spacetime set and their dynamics. The receiving object has absolutely no choice in that action occurring to it, and the action can occur from the gluons and quarks inside the charges in the nuclei, to the lattice structures, to the system dynamics, to the elements so as to transmute them, etc.

Note that when a dimensioned EM field or potential is absorbed by an object, that object also has internal EM fields and potentials which have substructures (usually rather random). However, the internal structure of the absorbed dimensioned entity and the internal structures of the EM fields, potentials, and waves do mix in a direct diffusion process. Hence with continued irradiation, the target can be "charged up" with the desired internal dimensioning structure. This is how the microwave irradiation of the U.S. Embassy accomplished disease induction, using signals with the desired disease dimensioning impressed in their structures, particularly in the potentials. In the Gulf War Disease, a much more sophisticated use of this mechanism was made, using a quantum potential as the carrier. All the effects (including the body fluid transmission, the EM fields of the body transmitting and kindling some of the effects in children in proximity to a dimensioned veteran parent, can be explained.

The use of dimensioned EM fields, waves, and potentials—and the use of dimensioned quantum potentials in Bohm's hidden variable theory—allows an entirely new science of EM biological warfare, far more threatening and lethal than the more primitive physical pathogenic kind. It is also possible to use the techniques for revolutionary healing, as unwittingly shown by Prioré and the scientists working with him, and also as shown by Becker *et al*.

The only way in which the action of a vacuum engine (dimensioned carrier entity) can be resisted is to create and simultaneously introduce the exact antiengine for it.

Once the action has been completed in the receiving object and it has been changed, the only way it usually can be undone is to form a precise antiengine for that action, amplify the antiengine, and let the antiengine act upon the object for a sufficient time to "time reverse it back to its previous physical state and condition". The Prioré mechanism utilized that method for reversing the cellular damage due to cancer and infectious diseases, and to restore suppressed immune systems. The same mechanism, if re-developed, could also be used to reverse aging in living human patients.

The engine and antiengine processes can also be applied by structuring the "time-charging" or "time-excitation charging" of the particles of mass in a body. The "port-hole" concept even allows ordinary EM radiation to be used to forcibly transduce the input ordinary EM into the extraordinary EM and thereby forcibly accomplish the necessary time-charging actions for disease and body deterioration reversal. See discussions under "port hole concept" and "time-charging".

DIPOLE

In the simplest case, two separated charges of (usually equal and) opposite sign.

As a general definition, a localized positive charge or charge distribution, and a localized negative charge or charge distribution, such that the net charge summation is zero, but wherein the positive and negative charge distributions do not superpose. We point out, however, that in regular physics *electrical charge* has no proper definition. Try looking it up in several physics dictionaries! Then check our proposed definition of electrical charge. As is well-known in particle physics, a dipole is a broken symmetry in the virtual photon flux of the vacuum. By definition of broken symmetry, this means that *some* of the virtual energy flux continuously absorbed from the seething vacuum by the charges of the dipole, is not re--radiated as virtual energy. Instead, it is integrated coherently and re-radiated in 3-space as real emitted EM energy. The dipole is therefore a true "negative resistor" since it receives EM energy in unusable form, and re-emits it in usable form. In classical EM theory, the problem of how every charge and dipole can continuously emit EM energy to form all the associated fields and potentials and their energy, is an unsolved problem—sometimes called the most pressing problem in electrodynamics, both classical and quantal. For the solution and a discussion, see my paper "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

Because of the broken 3-symmetry of the dipole, the EM energy flow is not conserved in three dimensions (symmetry represents a conservation law, and broken symmetry represents the violation of that conservation law—in this case, violation of conservation of 3-space EM energy flow). Instead, EM energy flow is conserved in 4-dimensions, and the hidden inflow of EM energy to the dipole is actually from the complex plane (time domain). Thus all EM energy is actually received from the time-domain, and so the importance can be seen of the internal structuring of time-flow and of longitudinal EM waves in the time domain. Such time-domain energy flows appear naturally in O(3) electrodynamics; see M. W. Evans and T. E. Bearden, "The Most General form of the Vector Potential in Electrodynamics," **Optik** (in review).

DIRAC, PAUL A. M.

Noted English physicist, 1902-1984, who shared the Nobel Prize for developing wave mechanics and predicting the existence of the positron.

DIRAC SEA

Considering the vacuum as containing—amongst other things— a "sea" of an infinite negative energy conditions or states ("holes") for electrons, almost all of which are filled with electrons.

In Dirac electron theory, the relativistic wave equation for the electron has four components, which correspond to two spin orientations and two energy conditions: positive and negative. Thus the vacuum possesses a "sea" of negative energy conditions for electrons. These are called "holes" and are considered to ordinarily be filled or mostly filled with electrons which have fallen into these negative energy states. The Dirac sea thus is this vacuum sea of negative energy electrons. In forward time, if energy is added, electrons can be lifted from this sea. If spacetime is locally curved, additional holes may be created so that normal positive energy electrons fall in and disappear, creating the appearance of an "electrical energy sink." If the curvature is in the opposite direction, electrons may be lifted from the Dirac sea by the curvature, thus producing a flow of electrons and creating the appearance of an "electrical energy source."

DISCRETIZED

Occurring in a finite size, fixed as a function of one or more variables.

So long as the same value(s) of the variable(s) are selected, the discretized entity is always the same size. However, if the value(s) of the determining variables are changed, the value of the discretized quantity may change so that it is not a whole number multiple of the former entity. Compare this to a quantized entity, where the values of the variable(s) are held to those wherein the entity will always be a multiple of an entity of fixed size.

DISEQUILIBRIUM

Broken equilibrium, which in turn involves breaking appropriate conservation laws.

DISSOCIATION

Separation of ideas from their affects or feelings, resulting in independent functioning of these components of a person's mental processes.

Simply put, fragmentation of the mind and personality. Severe dissociation is the definition of insanity.

DISTORTION CORRECTION THEOREM

"If a scalar wave E1(r) propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica E2(r), then E2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E1."

Quoted from Amnon Yariv, **Optical Electronics**, 3rd Edn., Holt, Rinehart and Winston, New York, 1985, p. 500-501.

What is missing from the distortion correction theorem is hidden in the phrase *"if we generate in some region of space ... its phase conjugate replica.."*. Since electrodynamicists and optical physicists have been unaware that all EM energy at any point in space comes from the time domain to that point, then the actual phase conjugate wave in the imaginary plane (from the time domain) is not utilized as the phase conjugate replica. Instead, the interaction of that wave with charges, which actually produces the force-field wave in space, is assumed. In short, one is looking at the "effect" wave, not the "cause" wave. However, since time is not an observable even in theory, it is this "effect" wave in 3-space that will be measured by our instruments.

Note also that the distortion correction theorem says nothing about longitudinal EM waves in the time domain. Interestingly, in nonlinear optical pumping, it appears that the 3-space energy of the pumping 3-space waves is transformed into the time domain by the pumping, and thence flows from the time domain into the 3-space domain to form the progression of the time-reversed replica wave in 3-space.

It is apparent that much more work should be done upon the foundations of phase conjugate optics, in a higher symmetry electrodynamics model such as O(3) electrodynamics and particularly in the unified field theory as in the Sachs-Evans combined approach.

DNA (DEOXYRIBONUCLEIC ACID)

Nucleic acid that constitutes the genetic material of cellular organisms and also certain viruses.

DRUDE, PAUL

Important German physicist, 1863-1906, who formulated the theory of the free electron gas inside conductors, using it to formulate a theory of metallic resistance.

DRUDE ELECTRON GAS

The "free" electrons that have broken loose from their atoms in a conductor, and are thus free to "wander around" and "migrate"—or flow as current when exposed to an emf.

Also called the electron gas, or Drude electron gas. Named after the German physicist Paul Drude who died in 1906, who formulated Drude's theory of metallic resistance based on EM theory and his own theory of electrons. See P. Drude, Ann. Physik, Vol. 1, p. 566; Vol. 3, 1900, p. 370, 869 for the original Drude papers.

DUALITY PRINCIPLE

In quantum mechanics, the principle that photons and other particles propagate like waves and interact like particles, so that their descriptions as only waves or only particles is inadequate.

ECM

Electronic Countermeasures.

Electromagnetic measures taken against targeted electronics or electromagnetics operations in systems, particularly in hostile weapon systems, command and control systems, etc. There are entire textbooks in the West that are devoted to ECM effects, uses, effectiveness, etc. A great deal of ECM equipment is also used by any modern military force. However, in our view what is greatly lacking as yet in Western ECM is the deliberate design and use of novel new kinds of ECM based on the extended higher electrodynamics such as O(3) electrodynamics, and particularly by use of longitudinal EM waves and by use of the infolded electrodynamics inside all EM fields, potentials, and waves. Longitudinal EM wave interferometers (LWIs), e.g., can reach right through the earth and the ocean to emerge on the other side of the earth. In the interference zone (IZ), real EM "transverse wave" energy is created readily by the interferometry. Biasing of the chassis grounds of the distant LWI transmitters can control whether the "normal" EM energy appearing in the target IZ is convergent (cooling) or divergent (heating). By sharp pulsing of the LWI transmitters, massive cold explosions or hot explosions can be created in the IZ. Continuous LWI interferometry can produce geometric forms of EM energy, which usually glow due to the ionization of the atmosphere, etc. in the dense EM energy regions. Since these can also involve general relativistic effects as well as electromagnetics effects in the U(1) CED sense, obviously a dramatic expansion of the entire field of measures and countermeasures is entailed.

We believe these LWI weapons may be what was referred to by U.S. Defense Secretary Cohen as follows: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our efforts."* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

EDGE WAVES

In electromagnetics, waves that form parallel to a material edge.

EDISON, THOMAS ALVA

Noted American inventor, 1847-1900, whose staggering number of inventions included the storage battery, electric light bulb, phonograph, motion pictures, and carbon microphone.

E-FIELD (ELECTRIC FIELD)

The electric field or electric intensity E. That field on a spinning charged mass which produces an electrical force resulting in a linear motion of the charge.

See also discussion under c, above.

E-FIELD, MOTIONAL

In classical EM theory, the electric field produced by moving a conductor in a magnetic field. Often called *motional inductance* of the field.

In the orthodox theory, motional inductance is deemed to be identical to the inductance caused by a time rate of change of the magnetic field, experienced by the moving conductor.

E-FIELD, STATIC

In orthodox theory, the electric field in the space surrounding a static charged particle, induced by the presence of the charged particle.

In classical EM, there is no notion as to the *causative mechanism* for this static field of and from a charge, which implicitly is assumed to just continuously create and pour out EM energy in all directions, thus grossly violating energy conservation.. Consideration of this anomaly leads to what has been called the most pressing problem in both quantal and classical electrodynamics: the question of the formation of the fields and potentials associated with the "source charge".

For a full solution to this problem, see my "Giant Negentropy from the Common Dipole," **Journal of New Energy**, 5(1), Summer 2000, p. 11-23. Also on this website.

EHRENHAFT, FELIX

Scientist who performed numerous experiments believed to demonstrate fractional charges.

EINSTEIN, ALBERT

German (Jewish) physicist, 1879-1955, who emigrated to the U.S. and formed special and general relativity theory, discovered and formed the theory of Brownian motion, recommended the development of the atomic bomb to President Roosevelt, and profoundly affected the course of science and the modern world.

EINSTEIN-de HAAS EFFECT

The rotation observed in a freely hanging ferromagnet when it becomes magnetized.

It has been used in the Einstein-de Haas method to determine the gyromagnetic ratio of a ferromagnetic material by suspending a cylinder of the material from a torsion fiber and measuring the angular displacement of the cylinder when its magnetization is reversed.

EINSTEIN, PODOLSKY, AND ROSEN PARADOX

Thought experiment in 1935 for measuring two correlated photons separated and at distant locations, protesting that quantum mechanics requires that the particles communicate instantly, which is a "spooky action-at-a distance" (Einstein's phrase).

The experiment eventually was performed and it works, just as required by quantum mechanics. The paradox inspired the development of hidden-variable theories.

EINSTEIN'S POSTULATES

(1) All the laws of physics are equally valid in all inertial frames of reference,

(2) the speed of light is the same to every inertial observer, and

(3) the observable local effects of a gravitational field are indistinguishable from those arising from acceleration of the frame of reference.

The first is called the special relativity principle, the second is called the law of light propagation, and the third is called the equivalence principle. It is now known that postulates (1) and (2) are not independent of each other. Note that the third postulate implies that any local force is due to a "gravitational field" (acceleration of the frame of reference). It also implies that gravitational force has a mechanism, since quantum mechanics assigns a differential operator, operating upon a potential, to create every force. By extended inference, it can be seen that the third postulate also implies that there should exist a method to turn any type of force (field) into gravitational force (field).

The Russian physicist Sakharov has strongly postulated that gravitation is not even a fundamental field of physics, of the kind used by Maxwell in his electrodynamics. Instead, in this view **G**-field is always comprised of, and made from, other fields and interactions upon them. Indeed, such has always been implicit in general relativity, it just took Sakharov to formally propose it. E. g., see A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," **Soviet Physics Doklady**, Vol. 12, No. 11, 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) in **Doklady. Akad. Nauk SSSR**, Vol. 177, 1967, p. 70-71.

Now let us do a little "tinkering" with Einstein's postulates. For the first one, we point out that in the new approach using dimensioning and subspaces, the notion of an inertial frame now has acquired additional complexity! The "gross translation" effects—which tends to be what physics ultimately focuses upon—can remain the same, and yet vacuum engines can be buried up inside the inertial frame to act upon a mass in non-translating ways. Everything translation-wise can remain the same, but now matter can be transmuted and transformed, etc., still right there in an inertial frame. So when the infolded electrodynamics inside all fields, potentials, and waves are utilized, the notion of "inertial frame" has acquired a higher topology and then one inertial frame may differ very substantially from another inertial frame. This of course move one from special relativity into full general relativity, and even there into a slight extension. Special relativity can remain special relativity on the surface, and yet effects occur "within" the objects postulated to exist in this inertial frame. The "inertial frame" concept in the postulates is now just a sort of special "equilibrium case" of a higher topology "inertial frame" notion.

The second postulate now stands as a "special case" of a far more general second postulate. It actually defines (and limits) "light" to bulk changes in the entire vacuum potential magnitude. It only applies to light signals which are "bulk upheaval" disturbances of the vacuum potential (spacetime). Specifically it need not apply to the "inner EM/GR realm" of longitudinal EM waves infolded inside the vacuum potential, other potentials, signal carriers, etc. Superluminal communication is now included in this "hidden variable" communication. The Fogal semiconductor is in fact capable of infolding and outfolding EM signals, including producing the internal longitudinal EM waves, and during their infolded state these signals are permitted to move at superluminal velocity.

So the second postulate needs to be restated to include the present statement as a special case. Notice that the concepts "light" and "speed of light" have now taken on much richer, extended meanings, since there is "outfolded" light and "infolded" light, so to speak. Light now is not just the "external perturbation wave" in the ambient vacuum potential medium (in flat spacetime). Instead, it can also be an "internal perturbation longitudinal wave" inside a "net flat spacetime" which does not show on the "surface" at all. Indeed, it may move in the time dimension or in 3-space. It now includes the transmission of an extraordinary "vacuum engine" as well as the usual transmission of just an ordinary signal.

And these infolded vacuum engines can be visualized as very special sets of longitudinal EM waves and their dynamics, consistent with the Whittaker decomposition. His U(1) symmetry electrodynamics decomposition—particularly when enriched in the higher symmetry O(3) electrodynamics—applies to the ambient vacuum as a potential (it is a potential since it has energy density). So the "subspace" inside spacetime, so to speak, is just a set of LWs and their dynamics, and that subspace itself is engineerable, just as is the "bulk curvature of spacetime" itself. Further, we may now consider space itself is just a vast superhighway for such superluminal LWs and their dynamics. Mass is mostly empty space with only an occasional particle here and there and with fields, potentials, and waves in between. Hence huge masses such as the ocean and earth are actually superhighways for the direct passage of longitudinal EM waves, "engines", etc. The degree of "transparency" is a function of the degree of cleanness of the LWs.

The third postulate has also been dramatically extended and requires revision. As presently stated, antigravity effects do not appear to be included in present physics heretofore. The "gravitational effects and external accelerated effects" notion in conventional physics is still founded purely on *ordinary bulk translation*. Now it need not be. For example, the system can appear to be an ordinary system, as judged by normal "external light" observation. Yet its internal vacuum engines can cause local violation of most so-called "laws of nature." The reason a "law of nature" can be violated is that one is now able to violate the premises upon which it is postulated, and which are required to hold if the law is to hold. Or another way to look at it: any conservation law is the result of a symmetry. Breaking that symmetry allows violation of that specific conservation law.

So as we've said so many times before, the infolded EM inside the "conventional" EM fields, waves, and potentials—and inside the vacuum potential and spacetime itself—dramatically changes the present theories of electromagnetism, quantum mechanics, and general relativity. Further, for the first time the Sachs unified field theory allows an *engineering* approach which simultaneously crosses, includes, and extends all three disciplines in the required manner. The approach can be placed into good engineering models based on the Sachs work, particularly using the O(3) electrodynamics subset spearheaded by Evans, and the models can be tested and verified or falsified.

EINSTEIN'S GENERAL THEORY OF RELATIVITY

Mathematical theory of gravitation in which the gravitational force is mathematically described by a curvature in space or spacetime.

In other words, the geometry and its dynamics have become physical agents.

It is important to note that, in general relativity, the curvature of spacetime acts on mass-energy, and changes in or to mass-energy interact back upon spacetime to change its curvature.

Further, we consider that an exact pattern of curvatures of spacetime and their dynamics exists for every mass system and its dynamics. This is the engine concept. The most general way to engineer matter is not by brute force translation, but by forming spacetime engines. Once formed, these engines act upon any mass placed in the engine region, to eliminate the deltas existing in that mass's resident engine and the impressed engine. The curvature of spacetime itself furnishes the energy for the continuing action upon the mass, once we pay to form the engine. The mass can be acted upon at any level, from its gluons and quarks to its nuclei, to its atoms, its lattice forces and binding energy, etc. Significantly, the Sachs theory applies to physical reality from beneath the quarks and gluons to the entire universe, and folds in quantum mechanics, electrodynamics, and general relativity into a single unified field theory that is (in theory) engineerable by higher symmetry electrodynamics means.

ELECTRIC FIELD

The *conventional* definition is this: "1. a region in space in which a stationary electric charge experiences a force due to its charge. 2. the electric force per unit test charge." **Dictionary of Science and Technology**, ibid., p. 720. That definition is flawed. Consider the "spatial aspects" of part 1. If an electric field is defined as a force per unit charge, it is not a region in massless space but is an effect of something existing in massless space and interacting with the charged mass. The electric field *prior to interaction* is indeed related to a region in space, because it is identically a curved spacetime region. In the Sachs approach, general relativity and electrodynamics have become one and the same. So prior to interaction, the *field as it independently exists prior to interaction with charge* has become a curved spacetime region, and the field is *purely* a spacetime curvature effect. The interaction with mass to produce a force on and of that mass (mass is a component of force, and there is no such thing as a massless force) is a separate consideration from the field as it exists in space prior to interaction. The result of this entity's interaction with charged mass has nothing to do with *what an electric field is*; but with *what it does to charged mass*. Further, the force evidenced on a charged particle is not due to its charge; it has that charge always, whether or not there is a net force evidenced on it!

Part 2 is a little better. The only problem is, it totally disagrees with the entity existing in space prior to interaction. In other words, here it's a "force on a charge", which is an effect occurring after the interaction. Again, it is what the field *does or accomplishes*, not what it *is*. Hopefully the reader can now begin to see what terrible difficulties await anyone who seriously attempts *foundations definitions*. Feynman sums it up admirably in one of his famous statements: "*Everything we know is only some kind of approximation, therefore, things must be learned only to be unlearned again or, more likely, corrected.*" An additional difficulty is that the very notion of a separate massless force independent of the mass on which it operates, is fictitious, yet continues to be widely used. Recalling that any true definition is an identity and not an equation, let us start with the definition of force:

$\mathbf{F} \equiv d/dt(m\mathbf{v})$. As can be seen, force is a system comprised of (i) a mass component and (ii) a nonmass component. The force is not something separate from the mass, at all! Further, force is not even unitary. It's a system comprised of two coupled components; the mass and a coupled (interacting) gradient in a potential flux. *Note that* $\mathbf{v}$ *involves motion through, and interaction with, the virtual flux comprising the vacuum potential. Note that* $d\mathbf{v}/dt$ *involves a gradient in that flux interaction with the mass*. Unfortunately mechanics is still using the medieval concept of a "separate force acting upon a separate mass." That is wrong, unless we change the fundamental definition of force itself. See, e.g., Robert Bruce Lindsay and Henry Margenau, **Foundations of Physics**, Dover Publications, New York, 1963, p. 283, where the authors emphasize that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. The reason can be seen: mass is a necessary component of force, and without mass present there cannot be a force present.

ELECTRIC DIPOLE

See Dipole.

ELECTRODYNAMICIST

A specialist in electrodynamics,

...the physics of the relationships between electric, magnetic, and mechanical phenomena—and especially a specialist in electrodynamics theory.

It should be pointed out that essentially all electrodynamicists are versed in U(1) electrodynamics, and a much smaller number are versed in more advanced electrodynamics such as O(3) symmetry electrodynamics, unified field theory electrodynamics, non-Abelian electrodynamics, etc. The electrodynamics field is actually a very vast field indeed, touching everything in physics. Very often, when one needs a particular electrodynamics problem solved, it calls for a special category of electrodynamicist and not just an electrodynamicist in general. One must therefore sometimes examine the background of the electrodynamicist making a statement or a pronouncement: electrodynamicists after all do have differing views and differing disciplines within the field.

ELECTROMAGNETIC PULSE (EMP)

A sharp pulse of electromagnetic radiation produced when an explosion occurs in an unsymmetrical environment, especially at or near the earth's surface or at high altitudes.

Essentially the same as an electromagnetic shock wave, a wave of great intensity which results when waves with different intensities propagate with different velocities in a nonlinear optical medium. An EMP is also produced when two intense longitudinal scalar electrostatic potential wave patterns meet and couple into a sudden flash of vector EM energy.

ELECTROMAGNETIC SMOG

The background of extremely weak EM radiation that fills the normal modern environment, particularly in developed nations.

The individual signals are too weak to individually produce effects of any significance in biological systems. However, in a dense EM smog environment with 500,000 signals per square meter per second, the signal complex and its nonlinearity begin to produce highly nonlinear effects in biological systems. These effects also produce damage that may cumulate over a long period of time, including for years. This bioeffects area has largely not been investigated by EM bioeffects scientists.

It should be pointed out that electric lines, telephone lines, transmission lines, etc. act as long wire antennas to detect and propagate such very weak EM signal noise. Essentially all the radiators in the environment—including electronics equipment, TV stations, radio stations, factory machinery, etc.—thus produce very weak signals picked up by these long line antennas and piped into every home, factory, neighborhood, etc.

ELECTROMOTIVE FORCE (EMF)

A difference of potential created by sources of electrical energy—i.e., by separated charges (dipoles)—which can be utilized to drive electrical currents through external circuits. Also, open circuit voltage, or the limit of the potential difference between the source's terminals as the current in the external circuit approaches zero.

Poorly named. Also quite confused in the standard theory and usage. First, *emf* is not a *force* at all. It is a difference in potential, hence it is measured in volts. Also, it is not considered an **E**-field, since that is a field of force. Yet it is in fact a , restricted to the direction between the two measurement points, and is considered to be an **E**-field in the theory!

For this reason the standard theorists say that the term is sometimes used as equivalent to a difference in potential, but that strictly speaking it should be applied only to a source of electrical energy! The problem is that conventional theorists are unaware that a potential is a bidirectional flow of energy in 3-space, where the energy enters from the time-domain (complex plane). Since they are unaware that a potential is hidden negentropic energy flow process involving broken 3-symmetry and viable 4-symmetry, the theorists are unable to consider the potential in general as a source of energy flow. Yet they are forced to consider a potential difference across the terminals of a source as "somehow a flow of energy, that generates forces upon the electrons to propel them through the circuit as current flow."

In the new viewpoint, a resolution is achieved of these incongruities by more accurately defining the E-field with respect to the potential as E [q]/q. It is noted that [dA/dt]q/q also produces an E-field, so that the full definition of the E-field is as E [(+dA/dt)q]/q. However, one volt represents one joule (of excess energy) per coulomb of charge. *The energy density flow in an electric circuit is in the voltage, in the emf, and it is not in the current per se*. The electron current (drift current) carries the excess energy that is being dissipated as work in the circuit, but in a nominal circuit the drift current only moves a few inches per hour. Half the energy is dissipated in the external circuit and external losses, and half is dissipated in forcing spent electrons in the ground return line back up through the back-emf of the source, performing work inside the source to dissipate its separation of charges. Left alone, any dipole will furnish emf indefinitely, freely receiving the energy from the time domain (complex plane) and outputting it in 3-space. The free flow of EM energy will continue unabated from the undestroyed dipole, regardless of how many external loads and losses are powered (how much energy is collected and dissipated from the ongoing flows representing the potential). So electrical circuits should be developed which allow little or no electrons to be forced back up through the dipole to dissipate its separation of charges and destroy the dipole.

ELECTROMOTIVE FORCE, BACK (BACK EMF)

Back emf is a loose term used for the *counter-electromotive force.*

When the source is powering a circuit or circuit element, the term "back emf" is used to refer to the forcing of current (electrons) back up through the source against its emf, thereby performing work inside the source to dissipate its separation of charges (its dipolarity). When a potential source of electrical energy—such as a capacitor, an inductor, or a rotating machine—is receiving energy from the source and converting it to dissipative power and work, that "powered" device is said to develop a *counter-emf.* A source has a built-in emf *a priori*, that is holding its charges separated so that it is dipolar. When electron current is forced backwards through the source through its built-in counter-emf, and backwards against its emf, work is performed against the internal resistance of the source to dissipate its separation of charges, thereby dissipating the source. If electron current is not allowed to flow back up through the source against its back emf and through its built-in counter-emf, no dissipation of the source occurs, even though energy density (not electron current dq/dt but energy flow d/dt) is flowing from the source and can be collected in collectors in the external circuit to store the energy as a created local energy source. This collected energy can then be shuttled to a separate load circuit and released, to provide a "heat pump" cycling operation and overunity coefficient of performance.

Nikola Tesla in fact knew how to shuttle energy in a circuit in such fashion, as Barrett's quaternion analysis of Tesla's actual patented circuits has revealed. [See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41.] Note that a rigorous analysis of those same circuits, using tensors or vectors, will not show the effect.

When the original source dipole is not dissipated because electron current is not passed back up through its back-emf to scatter the charges, the original source dipole will continue to furnish emf indefinitely.

The terms emf and counter-emf are normally applied only to recognized sources of electricity. Emf is applied as the algebraic sum of the potential differences acting in a circuit attached to a source. Back-emf (counter-emf) is applied as the potential difference internally through the source, between its terminals. Its main application is thought to be a *powered device or component in the circuit*. Actually, its main application is to prevent COP>1.0 electrical power systems from being developed and utilized.

ELECTROMYOGRAPHY (EMG)

A diagnostic application which detects changes in electrical potentials associated with muscle contractions.

Specific electrical patterns have been associated with specific abnormal states (e.g., denervated muscle), but the method has not yet been systematically developed beyond this application.

ELECTRON

Stable elementary particle in all atoms, and having a negative charge of 1.602×10^{-19} coulombs, spin ½, and mass of 9.11×10^{-31} kilograms.

ELECTRON GAS

A system of electrons with only weak interactions so that the electrons may be regarded as moving independently.

Thus, in many ways, that collection of electrons can be treated as a gas. Also see discussion under Drude Electron Gas.

ELECTRO-OPTICS

A technology in which electrons are focused and steered as if they were light. Examples are the cathode ray tube (CRT) and the electron microscope.

ELECTRO-POLLUTION

Also called *electromagnetic smog* and *electronic smog*. The conglomerate or maze of EM fields and radiations that are produced by technologies such as electric power transmission and radio transmission, and that can have long term cumulative harmful effects on humans.

The new viewpoint dramatically extends the types of fields and effects involved, and gives entirely new mechanisms for the interaction of EM fields and radiation with biological systems, including cumulative effects, signal-density effects, and vacuum engine effects. A mechanism for long term causation of cancer and leukemia by electro-pollution is derivable in the new approach.

As an example, even though the myriad of very weak EM signals comprising the electro-pollution are individually tiny, each when interacting at a local site with the body or its cells involves a broken 3-symmetry and thus an involvement of 4-symmetry, with corresponding giant negentropy. Further, the Heaviside dark energy component accompanying that local field interaction will in turn interact with essentially all the other locations of the body. Each of these interactions involves broken 3-symmetry and involvement of 4-symmetry. The result of the large complex of broken 4-symmetries constitutes a direct structuring in the time domain, hence weak time-pumping and weak time-charging. Time charge excitation decays only very slowly, emitting longitudinal EM waves inside the body as the charge decays.

Over a long period of time (sometimes years), this "time-hash" or "time-jamming" upon the operation of the master cellular control system, the immune system, and the cellular regeneration system is cumulatively detrimental. In this way, effects such as leukemia, cancer, etc. may be gradually "kindled". However, there have been no real scientific laboratory "cause and effect" investigations of this phenomenology at all. Nonetheless, the results will and do show in epidemiology studies, which are then attacked by orthodox scientists (many employed directly or indirectly by the power industry, etc.) on the grounds that no laboratory cause and effect evidence has been shown.

In short, this is one of the areas where the lack of scientific cause and effect laboratory evidence—because such experiments have not even been done—is improperly used to negate real scientific work showing deleterious effects that powerful interests do not wish shown.

ELECTRO-RETINOGRAPHY (ERG)

A diagnostic application that monitors changes in electrical potentials across the retina to assess eye movements.

This is almost the only method available for noninvasive monitoring of Rapid Eye Movement (REM) sleep.

ELECTROSTATIC COOLING

To be added

ELECTROSTATIC SCALAR POTENTIAL

A static (stationary) ordering in the virtual photon flux of vacuum.

The word "static" is unfortunate and should be replaced by "stationary", since the underlying active vacuum medium is always in violent motion, and there is nothing truly "static" in the universe. But an equilibrium condition can exist, even in violent motion, so that a stationary entity persists. An example is a perfect whirlpool in a river.

ELF

Extremely Low Frequency.

Used to describe EM fields whose frequencies lie in the range from just above 0 Hz (direct current) to 300 Hz. This includes power line frequencies (60 Hz in U.S. and 50 Hz in Europe) and frequencies used by certain U.S. military ELF communication systems such as for submerged submarines.

EM

Electromagnetics.

EMF

Electromotive force. See discussion under that term.

EM FIELD

Electromagnetic field. In its loose usage, it refers in a very broad way to any field, force, or energy associated with electromagnetic interactions, charges and currents.

Thus "EM field" includes electrostatic fields, magnetostatic fields, electromagnetic fields (including radiation and induction), vector-potential and scalar-potential fields, Hertz potentials, Fitzgerald potentials, Whittaker potentials, etc.

In its precise usage, "EM field" conventionally refers to an EM "force" field, such as the **E**-field, **D**-field, **B**-field, **H**-field. In classical electromagnetics (CEM), as a holdover from the old assumption of a material ether, the force fields are considered (erroneously) to exist in mass-free space, and they are also considered to be the primary causes of electromagnetic phenomena. This is a non sequitur, and it involves substituting the effect for the cause.

In the new viewpoint, the force fields do not exist in vacuum, but only in and on and of the charged particles themselves, where mass is present to form one component of the force. The force fields are therefore effects and not primary causes. The primary causes of all electromagnetic phenomena is taken to be the potentials; however, the hidden Stoney/Whittaker/Ziolkowski internal structures of gradient-free scalar potentials may interfere to cause potential gradients in and on charged particle systems. This therefore includes distance-free scalar potential interferometry and action at a distance, as well as vacuum engines.

The new view uses Sachs' unified field theory, and considers that the EM field in the vacuum, or an EM potential in the vacuum, is a curvature of spacetime function. Thus the new view is a unified field theory where GR has become EM and EM has become GR, and where EM can involve or produce gravitational, temporal, and inertial effects as well.

EM FORCE FIELD IN THE VACUUM (A FALSE CONCEPT)

The erroneous notion that there exist forces in the vacuum, so that the set of forces at a set of spatial points forms a force field.

This ignores the fact that mass is a component of force, and the presence of a force requires the presence of a mass. When Maxwell formed his electrodynamics theory, he and almost everyone else believed that a thin material ether pervaded empty space. Thus there was a "material medium" in which electromagnetic fields existed, and so it followed that real forces existed in this material medium, from point to point. Indeed, to the early electrodynamicists there was no point in all the universe where mass was absent.

Heaviside, Hertz, and Gibbs reduced Maxwell's 20 quaternion equations to the present four vector equations, but did not change the *force field in space* assumption used by Maxwell. Heaviside detested the "mystical" potentials favored by Maxwell, and believed they were figments of one's imagination. He regarded the forces as the primary EM causes.

Yet even while Heaviside *et al*. were retranslating and topologically reducing Maxwell's theory dramatically, the Michelson-Morley experiment in the 1880s had already destroyed the material ether on which those same force fields were based.

Physicists, however, have continued to use the notion of a separate force acting on a separate mass, which of course is also a non sequitur, substituting effect for cause yet again. Hence the foundations of physics remain fundamentally fouled.

For a succinct discussion of potentials versus forcefields as the primary EM causes, see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," **Physical Review, Second Series**, 115(3), 1959, p. 485-491. See also Y. Aharonov and D. Bohm, "Further considerations on electromagnetic potentials in the quantum theory," **Physical Review**, 123(4), Aug. 15, 1961, p. 1511-1524.

EM GENERATRIX

Whatever generates electromagnetics, particularly the fields. Time is the ultimate EM generatrix.

Whittaker showed in 1904 that the interferometry of two scalar potentials (i.e., of their hidden multiwave structures) can create all classical EM fields, no matter how complex, thereby starting a branch of electrodynamics known as *superpotential* theory. See E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

In 1903 Whittaker showed that any scalar potential (e.g. between the ends of a dipole) can be decomposed into a harmonic set of phase conjugate EM longitudinal wavepairs. The present author applied the Whittaker decomposition to the scalar potential between the ends of a dipole, and pointed out that the conjugate half-set may be taken as converging EM waves from the complex plane (the time domain), while the real half-set may be taken as diverging EM longitudinal waves emitted from the dipole in all directions in 3-space. The dipole is a known broken 3-symmetry in the dipole's violent virtual energy exchange with the vacuum; hence this shows that breaking 3-spatial symmetry allows nature to fall back to a more primary 4-symmetry. Hence energy flow conservation is violated in 3-space, but is rigorously obeyed in 4-space. Further, the free flow of 3-spatial energy from the dipole continues so long as the dipole exists.

The dipole thus is a true negative resistor, receiving EM energy in unusable form (from the complex plane; i.e., the time domain), transforming it into usable energy, and outputting the usable energy. This is also the fundamental mechanism for extracting unlimited EM energy from the seething vacuum, taking the energy from the time domain via the giant negentropy mechanism. See my paper, "Giant Negentropy from the Common Dipole," *ibid*.

This also means that the source of all EM 3-spatial energy is from the time domain, and is due to broken 3-symmetry. The ultimate EM generatrix is thus time. The ability of a bit of broken 3-symmetry (such as formation of a simple dipole) to initiate a continuing and ongoing flow of free EM energy via the 4-symmetry between time-energy and spatial energy, means that *negentropic* engineering, once we get the hang of it, is far easier than our present *entropic* engineering. In essence, we have to learn how to do electrical engineering all over again, from a much more fundamental point of departure.

EMI (ELECTROMAGNETIC INTERFERENCE)

Disturbances of electronic equipment's operation by intrusion of extraneous EM signals from outside sources. Also, disturbances of the operation of one or more subsystems of an electronic system by intrusion of undesired signals from other subsystems.

EMI, ANOMALOUS

An "ad hoc" term denoting an unexpected nonlinear optical functioning process in a dense signal environment whereby nonlinearities on the skin of a vehicle or cover of an equipment become pumped phase conjugate mirrors, so that they emit amplified phase conjugate replicas (PCRs)—of signals and fields impinging upon the skin or covering and the nonlinearity from electronic circuitry operating inside.

By the distortion correction theorem, these amplified PCRs precisely backtrack the signal waves that impinged upon the phase conjugate mirror from inside the vehicle or electronics, back to the original source of the internal signal. The result is the scavenging of energy from the dense electronic signal environment and sharply focusing and directing it into the operating circuits of the internal electronics. Thus internal signals, often amplified, start sporadically appearing through the circuits. These sporadic jamming signals due to the anomalous EMI effect are known as EMI fireflies.

Such EMI fireflies were resoundingly experienced in U.S. missiles and aircraft in the U.S. air attack on Libya in 1986. Although some of the anomalous EMI effects could have been produced by the unusually dense EM signal environment and poor hardening of circuits and electronics, it is believed that the weak jamming was also a test being conducted of a laboratory prototype Russian quantum potential weapon.

This weapon appears to have been finished and deployed at the end of 1989, and then utilized against Allied forces in the Gulf War, accounting for the Gulf War Disease. We have previously explained why British, Canadian, and American soldiers acquired the disease, why the native populace did not, and why the French troops did not.

Inside the human body, the anomalous EMI effects may be used to produce time-charging and delayed interference with the immune system, the master cellular control system, and the cellular regeneration system. By using dimensioned signals, the targeted systems can be specifically and deterministically affected as desired, for any set of disease damage types and symptomology desired. So the technology may be used to induce a "cocktail" mix of various infectious disease damage types. The induced time-charging effects—where the time-charge is dimensioned—followed by longitudinal EM wave emissions from the time-charges as they decay in the veteran's body, can induce the same or similar symptoms and damage in close family members exposed to the weak but dimensioned longitudinal EM radiation. This applies not only to wives but also to family members such as children often in close proximity to the veteran. In addition, exchange of bodily fluids from the veteran to his spouse can directly induce the symptoms in the spouse since those fluids are themselves still time-charged (with dimensioned time charge) and slowly decaying.

ENDOTHERMIC EXPLOSION

An EM implosion, resulting in explosive electrostatic cooling. The sudden, explosive reversal of electromagnetic energy scattering in a spacetime area so that the local EM heat energy is converted back to a negative potential. In this manner, explosive electrostatic cooling occurs in the area.

Endothermic EM explosion occurs by pulsed scalar Whittaker interference of two longitudinal EM wavesets beamed from two LW EM transmitters. The distant transmitters of the LW interferometer must have their local electrical grounds biased well below the ambient zero potential in the distant interference zone (IZ). Whittaker internal EM waves radiate through subspace (the LW superhighways inside EM fields, potentials, and waves in mass) from the transmitter to the interference zone, and also from the interference zone back to the negatively biased transmitter/receiver, emerging there as ordinary electromagnetic scattering energy or *heat*. All *normal* explosions are exothermic in nature, and ordinary electromagnetic heat energy is radiated away. A scalar EM interferometer/howitzer, however, may produce an endothermic explosion (an energy implosion) at a distance, explosively extracting the heat energy at the implosion site and receiving the energy back at the transmitter site for disposal.

The same LW interferometer can also produce an exothermic explosion in the distant interference zone, merely by positively biasing the electrical chassis grounds of the scalar EM potential transmitters and pulsing the transmission. Negative biasing of the transmitters' electrical chassis grounds and a pulsed transmission result in an endothermic explosion in the distant interference zone.

ENDOTHERMIC LONGITUDINAL EM WAVE INTERFEROMETRY

Extraction of EM energy from a distant intersection area by a scalar interferometry whose electrical grounding is biased well below the ambient potential in the distant interference zone.

The result is a cooling of the interference zone, and heat energy emerging at the transmitters.

ENERGETICS

The Russian term for what I have called advanced higher-symmetry electromagnetics,

...using artificially constructed Whittaker internal biwave structures to make activated potentials that have specific, desired "vacuum engines", where the EM field is a curvature of spacetime function, and thus involving a unified field theory that is engineerable by novel electrodynamic means.

I believe that the closest Western work approaching the Russian energetics science is the O(3) higher symmetry electrodynamics pioneered by Evans, particularly when combined with Sachs' unified field theory as an important subset. The other closest Western work would seem to be the important work of Barrett in SU(2)×SU(2) symmetry electrodynamics.

The activated potential containing the infolded vacuum engine structures is said to have been *dimensioned*. These dimensioned potentials can be combined to form EM carrier waves which contain the hidden vacuum engines. These vacuum engines, scalar Whittaker EM interferometry, and activation of vacua and matter are used to perform unified field engineering.

Russian energetics also adds the formation and use of the quantum potential, enabling (i) instantaneous action at a distance with multiple separated target nodes, (ii) essentially limitless direct energy amplification in the assembly of receiving nodes, without additional energy input at the initiating node, and (iii) engineering of hidden variables inside distant targets, instantaneously, without transmission *as such* through space. The quantum potential itself can be *dimensioned*, so that vacuum engines can be instantaneously transmitted into distant targets, altering their local spacetimes to contain the desired engine(s). Note that no intervening barrier can "impede" or "shield" a quantum potential because it does not travel through space as such; it just appears everywhere at once in the space where it will be.

Adding to classical EM theory, general relativity, and quantum mechanics the internal EM energy in the scalar EM potential will generate a superset of each discipline. These three new supersets are unified in their common internal EM energy subset. As a bonus mind, personality, memory, and deep biological cellular and system control are also available, since the internal time-energy domain is used by biological systems for these living functions. Energetics thus is an engineerable, unified field theory of physics and metaphysics.

Russian energetics is largely still a "secret" science, developed over the last five decades and highly weaponized. In the Russian scheme the overall area is called *energetics*. It consists of three subsets: (i) *energetics*, which applies to nonliving systems, (ii) *bioenergetics*, which applies to the physical bodies, cells, chemistry, genetics, etc. of living organisms, and (iii) *psychoenergetics*, which applies directly to the minds, personalities, memories, emotions, and perceptions of living systems. Russian energetics research, development, weapons, and weapon sites are under the rigid control and operation of the KGB, by whatever name it may use from time to time. The energetics weapons are not in the regular Russian armed forces. Energetics weapons have been tested worldwide for several decades, and this testing has been documented elsewhere by the present author.

ENERGY

Usually erroneously defined as the "capacity to do work".

The closest definition is: *The invariance of that which is varying*. Note that this is a 5-law logic definition and not an Aristotelian definition.

Here we have a real bucket of worms. This term is usually totally misdefined (as in the nostrum that "energy is the capacity to work") in physics, which defines energy as the "ability to scatter itself or change its own form" which obviously are not identities! A common definition is given as "a quantity that describes the capacity to do work; commonly divided into three major classifications: kinetic (dynamic) energy, potential (static) energy, and radiant (electromagnetic) energy. Well, energy can be changed in form and that is work. But energy is not the "capacity to change its own form", but is its own form, as can be seen.

Anyway, we'll have a go at a definition perhaps a little more acceptable, though still imperfect. Then we'll discuss things a bit.

Rigorously, energy is any deterministic or coherent structuring, either dynamic or stationary, existing in the virtual particle flux of vacuum. While that *sounds* good, it is still "the invariance of that which is varying." Ultimately, a thing is defined in terms of its appositive.

The *type* of energy depends upon the type of virtual particle(s) whose flux is selected. E.g., electromagnetic energy is a deterministic or coherent structuring, either dynamic or stationary, in the virtual *photon* flux of vacuum. Any trapped energy (ordering in the vacuum virtual particle flux) is gravitational in nature, since trapped energy is gravitational. In a mass, e.g., it is the trapped energy that is gravitational, not the mass per se, in the modern view. Contrast the definition of electromagnetic energy to the definition of an electrostatic scalar potential (ESSP): The electrostatic scalar potential is a stationary trapped deterministic or coherent dynamic wave-structuring of the virtual photon flux of vacuum. Thus it can be seen that the ESSP is a special case of EM energy: the stationary trapped case. For this reason, a mass—which is an enormous amount of statically trapped EM energy—may properly be considered as an extremely dense EM potential, which we refer to as the *mass potential*.

A potential is any trapped ordering or coherence, either static or dynamic, in the virtual particle flux of vacuum. A specific type of potential represents a selection of a particular virtual particle's flux in the vacuum, and an ordering—either static or dynamic—imposed in that flux.

However, since the scalar potential is just the static ordering in the virtual particle flux of vacuum, it follows that the scalar potential must contain—and in fact be totally composed of—internal order. E. T. Whittaker showed the form of that order in 1904. (See entry under Whittaker, E. T.) The scalar potential is composed of a harmonic series of pairs of antiparallel transverse waves. The antiwave of each wave/antiwave pair is a time reversed replica of the wave in that pair. In other words, the scalar potential is a special type of Fourier harmonic frequency expansion, coupled with its own phase conjugate present simultaneously.

Feynman said it bluntly: *"It is important to realize that in physics today, we have no knowledge of what energy is."* [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, **Lectures on Physics**, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 4-2.]

E.g., particularly see R. B. Lindsay, "The concept of energy and its early historical development." **Foundations of Physics**, 1(4), 1971, p. 383-393. Lindsay investigates the concept of energy from its early historical origin. The root of the concept is the notion of *invariance in the midst of change*. Lindsay covers the concept and its evolvement (or lack thereof!) from ancient times through the 18th century.

The real problem is that the notion of energy implicitly involves the age-old unresolved philosophical problem of change, posed by Heraclitus over 3,000 years ago and unresolved to this day. *It cannot be solved within the realm of 3-law Aristotelian logic!* One way of posing it is something like this, after Heraclitus: "For a thing to change, it must turn into something else. How can a thing be itself and something else also? How can a thing have changed, and yet not have changed?"

The problem is that Aristotelian logic has a serious flaw of omission; it actually *implicitly assumes* a fourth law, which is the age-old philosopher's nightmare that, when a thing is pursued to the limit, *it invariably turns into its own opposite*. So in anguish the philosophers spoke of this nightmare as the "accursed necessity for the identity of opposites."

I did resolve the philosophers' nightmare some years ago, by adding a fourth law of logic to the Aristotelian three, which completes and extends Aristotelian logic, and an application rule which can be taken as a fifth law. E.g., see Appendix III: "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," in T. E. Bearden, **AIDS: Biological Warfare**, Tesla Book Company, Chula Vista, California, 1988, p. 428-443. A simple proof of the fourth law is given in that paper. For a much more elegant treatment of a higher topology logic, which is also practical and useful beyond the limitations of Aristotelian logic, see G. Spencer Brown, **Laws of Form**. Julian Press, New York, 1972. For a proper perspective on mathematics and on Aristotelian logic, see Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980.

It may be that, since by "identity" we mean at least the identity of "left" and "right" different observations, considering the left and right sides of the identity statement. If so, then already we involve a violation of Aristotelian logic in the very notion of "definition". In short, ultimately we can only define a thing in terms of an appositive in the first place. So we write an identity statement where left and right are different sides, then state "Oh, no, they are the same thing after all!" Said in that manner, the problem can clearly be seen. In one sense a definition is pure Zen, where no truth can be stated at all unless the speaker says two exactly opposing things, and then waits for the recipient to realize the truth that has been stated.

What this all means is that, since "energy" (and any other definition, ultimately) involves a logical contradiction in Aristotelian logic, we must "define" *energy* simultaneously in an appositive ways: (i) in the wave view and (ii) in the particle view. However, we point out that the notion of a "moving particle" is the notion of something considered momentarily as fixed (i.e., the particle), but which is immediately thereafter (in time snapshot #2) considered as moving. Notice that "moving" is changing, and so it involves Heraclitus' problem of change *a priori*. The notion of a "moving wave" is also a similar thing; i.e., "a wave" momentarily considers the entity as static and not changing (in time snapshot #1), and in the next moment (time snapshot #2) considers it as moving (changing). We can only point out this dichotomy for the reader to reflect upon and hopefully to grasp. *The conscious mind is a serial processor; it is extremely fast but still perceives snapshot by snapshot, serially. It essentially has been fitted to the single photon interaction, as has Aristotle's three laws of logic*. The problem is in the operation of conscious perception itself, not in nature which is not limited to just serial processing.

The unconscious mind is a massively parallel processor, so it can have many "different slides or snapshots in the slide projector" at the same time. So with our conscious mind's perception (Aristotelian logic is just a model of that, using single photon interaction as the physical observation mechanism) we see in terms of "either-or". On the other hand, with our unconscious we can also grasp the fact that single snapshot A can be the same thing as a totally different snapshot B, if the rules of "decision as to identity" are changed between snapshot A and snapshot B. Suppose that by "identity" we mean "incapable of being discriminated between." Is a black marble identical to a red marble, e.g.? It is perceived so if you are color blind. It is perceived not to be identical if you are not color blind. The point is something that most philosophers missed or brushed aside: identity is not absolute, but is "perceived identity". Identity is the result of a decision operation, based on multiple perceptions, recall from memory, and comparison of two remembered perceptions. The most simple form of comparison is simple subtraction.

The unconscious (massively parallel) perception can handle both snapshots simultaneously. The conscious mind only can see a single marble at a time, so it will see one or the other. In a *third* time interval it will recall from memory the first perception, compare it with the memory of the second perception (say, by simple detailed subtraction), and then decide whether object A is perceived to be identical to object B. The unconscious (massively parallel processor) can handle both the marble snapshots at the same time. It can see both the identity of the two, to a color-blind "identity decision process" and also the nonidentity of the two, to a color-sensitive "identity decision process." Again, nature is not limited to the single-snapshot, serial perception process of our conscious (serial) processor!

Unfortunately, in physics and mathematics we seem to still be thoroughly confused on this very perception issue. Brown's work may be the first work to straighten out the area of the topology of the logic utilized. My own independent effort was successful but far cruder than Brown's; Brown did it elegantly. We bring all this out because when one deeply looks into foundations and true definitions, one will directly confront the philosopher's nightmare and one's own conscious mind will never be able to resolve the problem. One's unconscious can resolve it, but it requires a certain "stretching of the realization" to harmonize monocular perception of the conscious mind with multiocular perception of the unconscious mind.

Just as a small example, the question of "is it a wave or a particle" once raged as a great debate in physics, and was never resolved. Instead, physicists just finally shook hands and agreed to quit arguing and fighting about it, and agreed to accept that the same entity could sometimes act as a particle, and in other cases it could act as a wave, and that *as it existed in the present, before observation,* it was somehow both a wave and a particle but we couldn't see its true nation (in that fashion). The key, of course, is that the question was posed in "either-or" form (Aristotelian logic form), and it cannot be answered in that reduced logic. Instead, a higher topology logic is required. Brown's logic answers it easily, as does my own four law superset of Aristotelian logic (five law superset if we count the application rule as a separate law). In the present, e.g., we have masstime and not mass. Just after a photon is emitted, we have mass and not masstime. The latter photon emission process consummates "observation." The point is, the true entity does exist in zero time as a spatial entity. It is not a spatial entity but is a spatiotemporal entity. If the "interaction" process tears off its "time-tail," it is converted to a "spatial entity" (frozen snapshot) and therefore seems to be a particle. If the "interaction" process does not remove the time-tail, it remains a spatiotemporal entity and therefore seems to be a wave. The *concept* of particle is 3-spatial concept. The concept of wave involves time also. No wave can exist and "be a wave or wavelike" unless it has some time extension to "be waving" *in*!

The physicists debating the wave-particle duality, you see, missed the most important point that *no observable exists in time or persists, a priori.* Hence rigorously no particle exists in time.

So we define energy (particle view) as any virtual particle flux pattern, where that definition automatically considers a single particle in translation as well as multiple particles in their myriad of translations. We define energy (wave view) as any wave flux, where that automatically considers a single wave in translation as well as multiple waves in their myriad of translations. Now note that one is going to have to deal with, and discriminate between, energy, energy flow, energy collection, and collected-energy dissipation.

As the reader can see, whenever one struggles with foundations issues, the problems get very complicated. We've beat this dead horse quite enough, and leave the further beating to the reader!

ENERGY AMPLIFYING COIL

A coil whose magnetic energy flow input is first split into two energy flow inputs, each equal to the original energy flow available from a magnetic source. Both the split energy flows then interact simultaneously with the coil, amplifying its electrical energy output compared to the conventional case where only a single energy flow interacts with the coil.

In the motionless electromagnetic generator (MEG), the curl of the magnetic vector potential of a permanent magnet (i.e., its magnetic field), is extracted and held within the core material of a special transformer, while the magnetic vector potential of the permanent magnet is freely replenished in uncurled (field-free) magnetic vector potential form in space around the core material. Both the full magnetic field in the core and the full uncurled magnetic vector potential in space adjacent to the core interact with each output coil of the transformer. Hence a single potential energy flow is "split" into two flows each equal to the original EM energy flow, and both the parallel flows energetically interact with the output coil. This increases the energy interacting with the coil by this "dual interaction", so that the coil is said to be an "energy amplifying coil".

ENERGY COLLECTION/COLLECTING

In electromagnetics, exposure of an interceptor/collector, e.g., a charged particle, into an impinging EM energy flow, so that some of the flow is diverted around the intercepting particle. We may say that the flow of energy has been "reordered" in a stream around the diverting charge.

Energy collection is not a thing which "has been accomplished" and finished; instead, it is an ongoing process.

Note now that, if you perform or consider or impose any sort of ordering at all in the energy flow/flux, rigorously you have just implied an *energy collection/collecting* process. The energy part of a photon, e.g., would seem not to be "energy" per se. Instead, it is an energy collection/collecting, and a time collection/collecting as well. Yet that energy part is energy also! [Not to worry! Energy and energy-collecting, contrasted as appositives, become identities at the extreme boundary separating the two concepts. That boundary is the photon concept. Aristotelian logic does not apply on the boundary, which violates the Aristotelian law of the excluded middle.] More reasonably, a charged capacitor does not have "stored energy" per se in it, in the sense of "so many chunks of energy called joules". Instead, it has an *ongoing energy collection/collecting process*, where the asymmetry in the local vacuum flux exchange, of the dipole across the capacitor plates, continually extracts and "gates" a flow of EM energy from the capacitor dipole. That energy-flow bathes all the trapped charges in the capacitor. The energy-flow interaction with the trapped charges creates the **E**-field and all the rest.

The diverted part of the total energy flow, is what is calculated in the "Poynting flow" or **S**-flow.

ENERGY, DISINTEGRATED

An ad hoc term for the energy of the incoherent virtual particles in the virtual particle flux.

ENERGY DISSIPATION

Refers to the scattering or disordering of energy.

This is somewhat of a misnomer, broadly used in physics and somewhat erroneous.

Energy dissipation" *per se* does not mean that the energy is destroyed. Instead, there is the dissipation of the ordering (*collected/collecting*) of energy in which the diversion (local ordering) process disappears. What we call the "dissipation in space" of energy is rigorously the dissipation of the collection/collecting in space of the energy. In short, the energy is scattered, as in a resistive load, in all directions. But it is still there in space, just disordered.

Recall that all EM energy at a dipolarity comes directly from the time domain. It is the collecting process, not the energy, that is dissipated from a point or a region. Only *collected/collecting* of energy can be truly dissipated. With that understood, then dissipation of collected energy is simply removing the "collector" from the energy density flow impinging upon it to cause the collecting process to occur, or removing the flow of energy impinging upon the collector so that it intercepts and diverges some of it around itself.

So there are two ways in which this "collected energy" dissipation can occur: (i) the collector is moved out of the impinging energy-flow-stream, or (ii) the energy-flow-stream impinging on the collector is reduced to zero. Note that "changing the form" of the energy—i.e., changing the collector and the type of collecting—involves both conditions (i) and (ii). Thus energy is dissipated when the *collector/collecting* form is changed. In the conventional, less accurate statement, work is performed when "the form of the energy is changed." That now should be stated as "work is performed when the collector in the energy-flow-stream is changed to a different collector, because the first collector must be removed from the energy-flow-stream or the energy-flow-stream must be removed from it.

In physics we also consider that "positive" work by a body is work it does on something else, while "negative" work is work done upon the body by something else. We specifically emphasize that collection/collecting need not be a unitary, one-time occurrence! The EM energy-flow impinging upon a collector roars on by it at something like the speed of light. The entire flow is diminished by only about 10^{-13} (the nominal energy collecting fraction) as it flows onto, and on beyond a coulomb of charge. If that departing energy-flow is then retroreflected or otherwise *returned* to pass back through again, additional energy collection/collecting can and does occur from it. This can be repeated many times, particularly in an intensely-scattering, optically active material. Also, after the collected fraction of energy is dissipated, we can also retroreflect the scattered energy flow back and reuse it again. This too can be repeated many times.

So we must be careful of the present *work-energy theorem*, which is a special case and assumes only a single pass of the energy-flow in the "collection/collection" of energy that "changes form." If we institute additional passes, with additional "collection/collecting of energy, then we can collect and "use" a little more energy from all that enormous energy-flow that otherwise would escape us. If we dissipate some of the collected energy, then also intercept the departing dissipated energy flow and return it, we can collect some of the energy flow again and then dissipate that new collection again. And so on.

One can permissibly get many joules of "work" from a single joule of energy by iterative retroreflection, iterative re-collection, and iterative dissipation, without violating the conservation of energy law. This is one of the major secrets of overunity energy processes. It is proven by the overunity or excess emission (Letokhov was forced to call it "negative absorption of the medium" from an intensely scattering, optically active medium, as in the Lawandy experiment and in his patented process for lasing without population inversion. It is also proven by Patterson's adaptation of anti-Stokes emission.

Let's do a gedanken (thought) experiment, where we do not even apply the nominal 10^{-13} energy collection fraction. Let's just *conventionally* say "energy flow," "energy collection," and "energy dissipation," and ignore the "collection fraction." Since we are talking theoretically, we can have a perfect system. So we have a closed system, which we open to input some energy into. The system then opens in the load to dissipate all the energy back to the environment, thereby doing work. Now suppose I passively (i.e., without any input energy of my own, by perfect retroreflection) return all the dissipated energy back to the input once again. We simply let it flow back in, the system collects it again, and then dissipates it again—giving us the same amount of work we had the first time. Then we do it again. And again. I think you get the picture. The fact that energy can neither be created or destroyed, but only changed, means that the "same energy" can be *recycled* over and over. One has to have two different actions: one that changes the form of the energy so we get some work, and the other that retroreflects and recovers the dissipated energy (does negative work!) so that we can recycle it and dissipate it again. In principle, by recycling of energy one can get many joules of work from a single joule of energy. Nature does it every day! Every joule of energy formed at the initial

"Big Bang," is still with us. Every joule of it has on the average already performed many joules of work since then. And as we stated, the anti-Stokes emission process is a concrete example and demonstration that the energy recycling process can be accomplished in physical systems, so that the systems permissibly exhibit COP>1.0.

A most common example is the replenishment of a potential. E.g., the magnetic vector potential can be curled as a **B**-field, and that **B**-field can be totally confined inside a toroid, inside a long coil, or in certain magnetic core materials (as in the motionless electromagnetic generator or MEG for short). The **A**-potential will be replenished as fast as you withdraw the "total **B**-field" from it. So outside the toroid or the long solenoid, you will have as much **A**-potential energy as you started with, and the Aharonov-Bohm effect proves that the energy is there and will interact with electrons. Meanwhile, in the trapped **B**-field you also have drawn off as much energy as ostensibly the **A**-potential contained in the first place.

ENERGY, ELECTROMAGNETIC

Any deterministic or coherent structuring, either dynamic or static, existing in the virtual photon or charged particle flux of vacuum.

Note that, since a photon may also be considered a positron-electron pair, the flux of charged virtual electrons and positrons is included in the flux of photons.

ENERGY FLOW

The passage of energy, in any form, in one direction or generally in one direction.

Even though this "definition" is in error, we now just bypass the unresolved "observables do not persist" problem and the 4th law nature of change (e.g., motion) problem and proceed.

Note that this energy flow may be the overall movement of an omnidirectional energy flux, or it may be the movement of waves, or both. It may be in the kinetic energy of a moving mass. If the flow is everywhere the same, or even approximately the same, we often speak of the entire "flow" as just "energy." Often energy flow is also considered to have occurred from a region of more intense energy flux to a region of less intense energy flux, and vice versa. In other words, energy divergence or convergence is also considered as energy flow. As can be seen, *the usage of the term "energy flow" is not precisely defined in physics! It isn't going to be, at least not in ordinary Aristotelian logic terms.*

ENERGY FLUX

The passage of energy flow, in any form, through a unit area, and in any direction (usually all directions).

This includes both passage in a single direction, or passage in multiple directions at once.

ENERGY, FREE

Any environment or entity which freely emits, furnishes, or makes available energy which can be either collected and used as furnished or transduced into usable form and collected and used to power loads and do useful work.

The word "free" does not apply to costs, but rather to the source of the energy. Of course the ideal is energy that is both free for the taking (collecting and using) and also inexpensively taken and used.

Implicit in the notion of free EM energy from the vacuum are five key concepts: (1) the local vacuum/spacetime itself is filled with EM energy of generally unusable form, (2) some sort of organizing principle is utilized in one part of the circuit or system to cohere the vacuum energy and transduce it into usable EM form, (3) a gating mechanism is utilized to divert the vacuum energy in transduced usable EM form from the active vacuum to the external circuits or load parts of a device, (4) the coherent gated extra energy available in the external circuits or load parts is intercepted by those circuits or load parts, couples to them, and potentializes them with extra EM energy, so that the energy is collected by the system and regauges the system asymmetrically, and (5) the external circuits or load parts then scatter or dissipate the coupled extra energy to produce useful work, without destroying the organizing and gating mechanism that is extracting and furnishing usable EM energy from the vacuum.

ENERGY SOURCE

A function or entity which furnished energy.

Since energy cannot be created or destroyed, there is really no such thing as a "source", though it is a term commonly used. So called "energy sources" actually gate energy, changing it from one form to another. As an example, all 3-space EM energy is gated into 3-3-space from the time-domain. To the gating entity or "source" of the 3-space EM energy, there is an input of EM energy flow from the time domain (complex plane). The source must be a broken 3-symmetry in energy flow, so that more primary 4-symmetry energy flow applies. Hence the conservation law applies to the inflow of energy from the time dimension and the outflow of energy in 3-space.

ENGINEERING, ENTROPIC

Engineering or designing systems and processes whose operation "spills" or wastes some energy.

Entropic systems—which are the products of entropic engineering—thus have COP<1.0 and function in forward-time.

ENGINEERING, NEGENTROPIC

Engineering or designing systems and processes whose operation freely "receives" excess energy, so that they output more energy than the operator inputs, the excess being received from the external environment or other source.

ENTROPY

Thermodynamics term representing a measure of a system's capacity to undergo spontaneous change. In short, a quantity which measures the extent to which the energy of a system is unavailable to do work.

Entropy is given by $dS = dQ/T$, where dS is an infinitesimal change in entropy for a system absorbing an infinitesimal quantity of head dQ at absolute temperature T. In statistical mechanics, entropy is a measure of the disorder in a system and is given by $S = k \ln P + C$, where S is the value of entropy (disorder) for a system in a given state, P is the probability of that state's occurrence, k is a fixed constant, and C is an arbitrary constant. Intuitively, the concept of entropy corresponds to the concept of disorder. We specifically call attention to the fact that the "S" used in representing entropy is not at all the **S** used for Poynting flow.

In physics, a measure of the randomness or disorder of a system. Given by S = k ln g, where S is the entropy, k is the Boltzmann constant, and g is the total number of microstates available to the system. Sometimes stated as S = k ln (multiplicity).

Intimately connected with the Second Law of thermodynamics, which may be stated for the macroscopic case (after Baierlein) as follows: "If a system with many molecules is permitted to change, then—with overwhelming probability—the system will evolve to the macrostate of largest multiplicity and will subsequently remain in that macrostate. Stipulation: allow the system to evolve in isolation. (The stipulation includes the injunction, do not transfer energy to or from the system." [Ralph Baierlein, Thermal Physics, Cambridge University Press, Cambridge, 1999, p. 29].

ENTROPY, NEGATIVE (NEGENTROPY)

In a sense, the reversal of disorder, or reversal of entropy.

Since the normal or forward flow of time is entropic, negentropy is intimately related with time-reversal. This can be seen crudely this way: As a physical thing exists in time, it decays, ages, deteriorates, etc. In short, its ordering as a viable separate thing gets more and more disrupted. If it is then "reordered" precisely back to its original state, that is obviously a time-reversal of that entity.

Since the production of energy is connected to the production of more ordering, then it would seem to follow that any negentropic process also produces more energy. Or, that production of excess energy accompanies a negentropic process, and vice versa.

When we factor in the giant negentropy process, it is seen that all EM spatial energy comes from the time domain, via the broken 3-symmetry of the source charge or source dipole. Here we have a pure negentropic process. Hence we may now advance a new definition of negentropy for electrodynamics: *negentropy is the direct and ongoing conversion of time-energy (time-structuring and time-ordering) into 3-space EM energy*. Dipolarity (polarization) is the process initiating and producing the negentropy. Polarization is accompanied by ongoing negentropy so long as the polarization persists.

Another way of defining negentropy even more rigorously—for electrodynamics—is to define it in spacetime as a *broken 3-symmetry in EM energy flow with concomitant relaxation to a more primary 4-symmetry between time-energy flow and 3-spatial EM energy flow, where time-energy flow is transduced into 3-spatial EM energy flow.*

When a system with one measure of multiplicity progresses to a state with a decreased measure of multiplicity; i.e., when its total number of microstates decreases in the transition. A nonequilibrium steady state (NESS) system can exhibit negative entropy, and the entropy can further decrease progressively toward negative infinity as time passes. [D. J. Evans and Lamberto Rondoni, "Comments on the Entropy of Nonequilibrium Steady States," *J. Stat. Phys*., 109(3-4), Nov. 2002, p. 895-920]. Note that the Second Law of thermodynamics only rigorously applies to an isolates system, where neither mass nor energy crosses the boundary of the system. In the NESS system, that prohibition is violated, hence the second law need not apply. The common source charge is an example of such a NESS system continuously producing negative entropy, because it absorbs virtual energy from the vacuum, transduces it into observable energy, and re-emits the energy as observable photons radiating at light speed in all directions and forming the associated EM fields and potentials and their energy.

ENTROPY (POSITIVE)

When a system in a state with ordering progresses to a state with less ordering, i.e., when its total number of microstates increases in the transition.

ENZYME

Any of various proteins or conjugated proteins (compound of a protein and a nonprotein) produced by living cells, that serves as a catalyst for specific biochemical reactions at body temperatures.

EPIDEMIOLOGY

The study of the various factors influencing the occurrence, distribution, prevention, and control of disease, injury, and other health-related events in a defined human population.

EQUILIBRIUM

In physics, the state of a system in which opposing forces balance each other; i.e., sum to a vector zero.

In mechanics, a system of particles where the net force acting upon each constitutent particle is zero.

In chemistry, the state of a solution where the forward and reverse reactions are equal and balance each other, so that the chemical reactants and products maintain the same amounts.

EQUIVALENCE PRINCIPLE (PHYSICS)

One of the principles of relativity theory. The mass of the body determined by using the gravitational force on it to measure the mass, and the mass of the body determined by using inertial force on it to measure the mass, yield equal measurements for the mass.

ETHER

An extremely fine substance theorized to exist in space in lieu of sheer emptiness.

Discussed by Tesla, Einstein, Whittaker, and others. Originally conceived as the luminiferous ether, a thin material fluid pervading all space. Maxwell, e.g., believed in this material ether, and formulated his theory accordingly. Thus he mistakenly provided forces in the vacuum, which forces require a material vacuum. I.e., mass is a component of force; there is no force without mass, since $\mathbf{F} \equiv d/dt(m\mathbf{v})$. The hoary old medieval notion that a separate force acts upon a separate mass that it affects, is completely in error, though still utilized in mechanics! Modern field theorists know better, but correction of this massive error in physics has not been widely accomplished.

Originally, the ether was seen as a ghostly-thin, *material*, fluid medium postulated to fill space and carry electromagnetic waves, similar to the way in which a gas carries sound waves. Faraday and Maxwell added the assumption that EM fields existed as material taut strings (Faraday's lines of force) with conveniently missing endholders to provide the tension on the strings. The Michelson-Morley experiments eliminated any such material ether, but not a Lorentz-invariant nonmaterial ether (of which vacuum itself is one example, in the modern view). Maxwell's equations were derived from a mechanical ether theory. Einstein once proposed that the vacuum, complete with its electromagnetic fields, should be called the new *ether*. In our own approach we have come full circle back to Einstein's suggestion, since the so-called "fields" are really just curvatures of spacetime and are included in our use of the vacuum as a giant potential comprised of virtual particle flux. The Sachs unified field theory, I believe, is the best available, particularly when the Evans O(3) symmetry electrodynamics is included as a special subset of it. In my own view, I regard "vacuum potential" and "spacetime" as essentially identities. Thus both the vacuum potential and spacetime have further internal structuring and dynamics. In relativity, this view is akin somewhat to the notion of "subspace" used on the Star Trek series.

Today the word "ether" is again acceptable in the literature when used in context of its modern quantum mechanical meaning. The "ether gas"—the quantum mechanical vacuum—is considered to be composed of virtual particles. The particles spontaneously come into being and then vanish so quickly they cannot be individually detected. That is, today the ether is regarded as a sort of specialized virtual particle gas—sort of a gas of "effervescent energy bubbles", so to speak—where the individual particles are continually being created and annihilated at a fantastic rate.

A good reference on the modern view of the vacuum is I. J. R. Aitchison, “Nothing’s plenty: The vacuum in modern quantum field theory,” **Contemporary Physics**, 26(4), 1985, p. 333-391. For an O(3) treatise, see M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al*., "Derivation of the Lehnert field equations from gauge theory in vacuum: Space charge and current," **Foundations Of Physics Letters**, 13(2), Apr 2000, p.179-184. It is shown that the Lehnert field equations in vacuum, with concomitant space charge and current, can be derived straightforwardly from standard gauge theory applied in vacuum, using the concept of covariant derivative and Feynman's universal influence. The Lehnert and Proca field equations are shown to be inter-related through the well-known de Broglie theorem, in which the photon mass can be interpreted as finite. These ideas go some way towards addressing the inconsistency inherent, in Maxwell's famous displacement current, which has no concomitant vacuum space charge.

For technical insight into the attributes for extracting EM energy from the vacuum, see particularly M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al*., "Classical electrodynamics without the Lorentz condition: Extracting energy from the vacuum," **Physica Scripta** 61(5), May 2000, p.513-517. It is shown that if the Lorentz condition is discarded, the Maxwell-Heaviside field equations become the Lehnert equations, indicating the presence of charge density and current density in the vacuum. The Lehnert equations are a subset of the O(3) Yang-Mills field equations. Charge and current density in the vacuum are defined straightforwardly in terms of the vector potential and scalar potential, and are conceptually similar to Maxwell's displacement current, which also occurs in the classical vacuum. A demonstration is made of the existence of a time dependent classical vacuum polarization which appears if the Lorentz condition is discarded. Vacuum charge and current appear phenomenologically in the Lehnert equations but fundamentally in the O(3) Yang-Mills theory of classical electrodynamics. The latter also allows for the possibility of the existence of vacuum topological magnetic charge density and topological magnetic current density. Both O(3) and Lehnert equations are superior to the Maxwell-Heaviside equations in being able to describe phenomena not amenable to the latter. In theory, devices can be made to extract the energy associated with vacuum charge and current. More than a dozen ways to approach extracting EM energy from the vacuum are presented.

EVANS, MYRON W.

Noted chemical physicist, primary developer of O(3) symmetry electrodynamics, and Director of the Alpha Foundation's Institute for Advanced Study (AIAS).

Under Evans, the work of the AIAS has made enormous contributions to physics and particularly to the extraction of usable EM energy from the vacuum. In the opinion of the present author, Evans is perhaps the single most important scientist in the Western world in the emerging field of electrical energy from the vacuum and in solving the world energy crisis.

EVIL

Refers to an act or acts perceived to be detrimental to the receiver or referent, in the positive time stream.

Oddly, the chain of events perceived to involve "evil" to a referent in the forward time stream, would be perceived to be equally *beneficial* to the referent, viewed in the *negative* time stream. In other words, evil in general is a synthesis of the results of observed actions or acts in the *forward time motion*, hence it is conjugated when time flow is reversed.

This is not just an idle question. In our physical body, the photon interaction with the electron shells of the atoms in the body largely determines the observed forward movement of one's material body through time. The conscious mind, fitted to the single photon interaction, thus perceives its "existence" in such terms. Meanwhile, in the nuclei of the atoms of our material body, the charge and hence the field interactions are reversed (giving Newton's third law reactions). That of course is a time-reversed flow of events, not sensed by our conscious mind, but possibly sensed by our deep unconscious mind. The question then arises, "Do we then actually exist both forwards and backwards in time?" In terms of good and evil, if the answer is yes, then this would have profound implications to theology and to the study of human consciousness.

Theologians (and philosophers) have pondered the "nature" of good and evil for literally thousands of years, and have reached no truly firm conclusion to date. The sticking point (theologically) seems to be the puzzle of a perfect, "all good" creator who had to have created evil also (since the Creator is axiomized to have created all). To the question, "How could a perfect Creator have created—or even tolerated—evil?", theologians have yet to find a satisfactory answer. Neither could they find a satisfactory answer to the question of, "How can the Creator, being all-just, and being all-powerful, then allow any evil act to occur?" In short, why does the Creator seemingly not "police up our perceived imperfections in the creation?" In short, theologians struggled with a formidable problem of how to justify the Creator's continued tolerance of evil, even for a moment.

There is no satisfactory answer to such questions, of course, in 3-law Aristotelian logic, because they are four-law logic situations. However, since the 3-logic is incomplete, one should examine the question in 4-law logic, and also in light of one's assumptions regarding the existence of the Creator. The question is unresolvable in 3-law logic and with the assumption of a Creator traveling through time, as we perceive ourselves to do. However, if time itself is a creation, then the axiomized Creator may be more properly axiomized as existing both in and outside time, and "everywhere in time at once" as well. The unlimited Creator must even be postulated to be able to "both be and not be simultaneously; if He cannot do that little trick, He has a limit." That trick must be allowed, else the Creator has a limit and is thus imperfect. But to such an axiomized Creator "permeating time everywhere at once, from beginning to end, which shall be the "beginning" and which shall be the "end?" And if we pick a beginning and an end, what shall be the "intermediate ordering" that we pick? There is no absolute position in either space or time to the "everywhere and everywhen at once" axiomized Creator.

The problem of good and evil then becomes a similar problem to the "wave or particle" question. Good and evil are two sides of the same coin, transposed one into the other by time-reversal. With respect to finite creatures perceiving themselves moving forward in time, not existing in the future and not existing in the past, the question merely demonstrates that our perception itself is in error.

It may be, then, that there is a solution (whether one prefers it or not!) to such profound and ancient philosophical and theological problems in four-law or five-law logic. However, we will leave that question to the logicians and to the theologians to more properly decide.

EXCESS COLLECTED ENERGY

Collection by an interceptor/collector of a greater fraction of the EM energy flow impinging upon it, than is normal.

Any static electric or magnetic charge, e.g., with an EM energy density flow impinging upon it, has excess "energy collection" or "energy collecting" upon it, due to the altered vacuum virtual photon flux that is now exchanging with it. In a nominal electric circuit, e.g., a single "pass" of the EM energy flow from the source dipole, along the circuit conductors and across the conduction electrons, causes an "excess energy collection" upon the intercepting surface charges in the conductors. This potentializes the Drude electrons in the circuit and leads to subsequent dissipation in the loads (and losses) of about 10^{-13} of the EM nondiverged Heaviside energy flow that actually was extracted by the source dipole from the vacuum and sent down through space surrounding the conductors of the external circuit.

That is a single-pass energy collection fraction for a static charge. Almost all the Heaviside energy flow missed the circuit and passed on beyond the circuit "collectors" in one pass-by at essentially the speed of light, on out into deep space and beyond. On the other hand, if the already-passed EM energy flow is retroreflected, so that it returns for another pass along the circuit, then an *additional* amount of the repassing energy can be collected by those sluggish electrons. In a nominal circuit being used as an example, the electrons travel only about 5 or 6 inches per hour longitudinally down the conductors, restrained by the repulsion of the electrons beyond them. So *multiretroreflection*, *multipass*, *multicollection* can be used to provide excess energy and overunity efficiency. This is a change to the present statement of the work-energy theorem, which up to now has assumed single pass, single energy collection. The well-known anti-Stokes emission from certain highly scattering, optically active media, e.g., produces more energy radiating back out of the media than the amount of stimulating radiation energy input into the media. This is Letokhov's "negative absorption of the medium", which is really the excess emission from the medium due to retroreflection and multiple re-collections.

Also, if the intercepting charge is resonated, it sweeps out a greater geometrical cross section in the streaming energy flow, thereby diverging more and thus "collecting" more energy. In the Bohren experiment, e.g., this effect yields collection of some 18 times as much energy by the resonant charge as the same charge collects in static mode.

EXOTHERMIC

Giving off heat, or emitting scattered EM energy. Producing forward time energy.

FARADAY, MICHAEL

Noted English physicist and chemist, 1791-1867, who discovered field theory and electromagnetic induction. Faraday invented the dynamo also and formulated his laws of electrolysis.

His work directly inspired Maxwell, who vowed to study no other electromagnetics theory until he had captured Faraday's work in a mathematical theory.

FIELD

In mathematics, a set with two binary operations. In electromagnetics in more common terms, it is a region consisting of magnitudes and/or vector directions assigned to some active entity (spacetime itself) at each point in the region.

Designated addition and multiplication, satisfying the conditions that the set is a commutative group with respect to addition, that the set with the identity of the additive group omitted is a commutative group with respect to multiplication, and that multiplication distributes over addition for all elements in the set.

For our purposes, in plain language the fundamental notion is that a "field" is a sort of cleared, level space or region (e.g., in spacetime) where things can be placed and actions can occur amongst them, to them, on them, and of them. What the "place or space" is, can get a bit complicated, as can what is placed there and what actions are permitted. In short, we view a field as an altered (curved) region in spacetime, and this special curvature interacts with certain kinds of entities placed in that region. In return, those entities also interact back upon the spacetime curvature to alter it. Hence there is a dynamic two-way interaction between field and matter, where half of this interaction (the back-interaction of charge or matter) is missing from present classical electrodynamics.

EXOTHERMIC EXPLOSION

An explosion where ordinary heat and electromagnetic energy are released and radiate away from the explosive site. All normal explosions are exothermic.

Particularly used for a pulsed longitudinal EM wave interferometer, which in the distant interference zone (IZ) will produce a sudden eruption of scattering EM energy from the local spacetime curvature suddenly induced, when the interferometry transmitter chassis grounds are biased well above the ambient potential in the distant IZ.

FERMI LIQUID

A liquid where the individual particles or molecules comprising it are basically non-interacting, so that they can roam freely.

In an electron gas in a conductor, e.g., treatment of the free conduction electrons as a gas allows a theoretical modeling and explanation of conduction and resistance. Contrast to a Luttinger liquid, where the electrons interact with each other and thus enable the formation of collective phenomena.

FEYNMAN, RICHARD

Noted physicist, Nobelist, and co-developer of quantum electrodynamics.

Feynman is particularly noted for his refreshing three volumes of physics: See Richard P. Feynman, Robert B. Leighton and Matthew Sands, **The Feynman Lectures on Physics**, Addison-Wesley, New York, 1963. The book is particularly noteworthy in that it gives the student some insight into foundations problems in physics. E.g., in Vol. I on p. 2-4, Feynman defines the electric field not in terms of force per unit charge *per se*, but in terms of its potentiality for producing a force only when a charge is present for the force to be developed upon.

FEYNMAN DIAGRAM

A two-dimensional diagram of particle interactions, both in the virtual and observable levels, and both backward and forward in time and named after Richard Feynman who first employed them.

FIBER FUSE

Effect in fibers or fiber optic cable, whose core contains germanium, whereby heating the cable at one spot with a small flame will launch a focused traveling heat energy effect which melts a small hole in the core about every centimeter, and marches down the cable at about a meter per second to the end of the cable, destroying the optical transmission capability and ruining the cable.

Oddly, by igniting the fiber fuse from the other end of the damaged cable, a reverse fiber fuse effect will often march back down the cable in similar fashion except that it fills in the melted holes in the core, restoring the transmission capability of the cable.

If the energy required to melt all the holes (or fill them again) is taken into account, the amount of energy required to launch the fiber fuse is much less. This phenomenon does not appear to have an adequate technical explanation at present.

For an excellent description with photographs, see Ivars Peterson, “Fibers With Flare,” **Science News**, 140(13), Sep. 28, 1991, p. 200-201. For technical information and analysis, see D. P. Hand and P. St. J. Russell, “Solitary thermal shock-waves and optical damage in optical fibers,” **IEE Colloquium on ‘Non-Linear Optical Waveguides’**,” IEE Digest No. 88, London, England, 1988, p. 101-103; — Soliton-like thermal shock-waves in optical fibres: origin of periodic damage tracks," **Fourteenth European Conference on Optical Communication (ECOC 1988),** Sep. 11-15, 1988, Publication No. 292, 1(2), 1988, p. 111-114; — “Solitary thermal shock waves and optical damage in optical fibers: the fiber fuse,” **Optics Letters**, 13(9), Sep. 1988, p. 767-769; D. P. Hand and T. A. Birks, “Single-mode tapers as ‘fibre fuse’ damage circuit-breakers,” **Electronics Letters**, 25(1), Jan. 5, 1989, p. 33-34.

See P. St. J. Russell, “Power Conservation and Field Structures in Uniform Dielectric Gratings,” **Journal of the Optical Society of America A**, 1(3), Mar. 3, 1984, p. 293-299. Russell is actually referring to apparent violation of local conservation of energy. This would be indicated if the vacuum-to-mass interaction is not taken into account. Russell’s analysis therefore supports the fact that (i) the vacuum-mass interaction is required, (ii) the fiber fuse does utilize some of its gated vacuum energy (i.e., the usually non-diverged portion of the Poynting flow), and (iii) the fiber fuse is a legitimate, repeatable overunity process, although a "runaway" process, uncontrolled, and not sustainable.

FIELD

In physics, "a region of space marked by a physical property, as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region."

[Thanks to **Webster's II New Riverside University Dictionary**.]

Here one must *not* think of "space" as emptiness, as mathematicians do, but as a virtual particle flux, a dynamic entity, and as dynamic spacetime. A field then is a set of entities and their relationships and operations in a region of *that* kind of space, where the entities and operations are themselves "space stuff". They are all "virtual" in that they are not individually observable. However, their impact upon or interaction with mass—particularly charged mass—can be and often is observable, in that we can see some change of the mass (its direction of travel, translation, acceleration, deceleration, rotation, generation of current, etc.). If the entities are point-like and are naively supposed to have magnitude only, or if we only specify their energy density magnitude at each point, then that "marked space region" is said to be a *scalar field*. If the entities are point-like but have both magnitude and direction, they are said to be a *vector field*.

Further the deponent saith not because even the foundations physicists are in hot water in trying to define the field, and if the deponent saith more he will surely deep-fry in hot oil!

FIELD, SCALAR

A field, to any point of which is assigned a magnitude only.

FIELD, VECTOR

A field, to any point of which is assigned both a magnitude and a direction.

FLAT (UNCURVED) SPACETIME

A vacuum/spacetime whose virtual particle flux intensity—and therefore its local energy density—does not change as a function of spatial position or averaged time.

That is, a flat spacetime in my view is a quantum mechanical vacuum whose virtual particle flux and energy density are constant in magnitude and internal constitution with respect to spatial position and in averaged time. Note that a flat spacetime (vacuum potential) can still contain vacuum engines, and thus still be *dimensioned*. In that case, contrary to normal physics, the flat spacetime itself will still interact with an embedded object and change it, even though it may not translate it. To perform transmutation of elements, e.g., one is interested in just such engines and not in translating through space the element being treated. This dimensioning or activation of the local flat spacetime can result in an otherwise unexpected, seeming violation of one or more laws of nature in the dimensioned region.

FISSIONING, ACTION

The splitting of the action quantum into two canonical variables (for example, energy and time), only one of which can normally be completely detected or measured. ("Action" has the dimensions of angular momentum—energy x time, momentum x length, etc.)

Fissioning produces length, discrete bit by discrete bit, and time likewise. The physical reality of separated physical objects is thus continually created, discrete bit by discrete bit, monocularly. All that ever observably exists in the "present" is a single quantum change, just occurring. Recombining of the two canonical action fission fragments is called "fusion." Fusion also occurs monocularly, quantum by quantum. Rigorously, the physical world of separated objects never exists as such except in memory recall of vast numbers of past ordered quantum changes that occurred one by one.

FLUX

The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

The term flux density is now more commonly used. Loosely, "flux" refers to the perpendicular flow through an area, while "flux density" refers to the magnitude or intensity (rate) of the flux. In some cases such as a flux of particles of different velocities, the number of particles may be multiplied by the average velocity to give an average flux density that is weighted.

FLUX DENSITY

The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

FOGAL, WILLIAM (BILL)

Soon-to-be-noted researcher and inventor of the Fogal semiconductor (two patents so far), as well as a superluminal infolding-outfolding communication system and other remarkable longitudinal EM wave processes.

I am happy to be a friend and colleague of Bill Fogal.

FOGAL SEMICONDUCTOR

A special semiconductor invented by Bill Fogal which, in its operational regime, utilizes an integrated semiconductor, tantalum capacitor, and feedback resistor to accomplish charge blocking (blocking of electron current flow dq/dt) while passing displacement current d/dt

The Fogal semiconductor can also accomplish amplified phase conjugation of signals as well as infolding (translation of input transverse EM waves to output longitudinal EM waves) and outfolding (translation of received input longitudinal EM waves to output transverse EM waves). Used in communication systems, it opens the use of the unlimited "infolded" electromagnetics bandwidth. Since it may communicate using longitudinal EM waves, it is also usable for superluminal or a specialized "tunneling" communication through the "interior" of normal EM waves, potentials, and fields.

See William J. Fogal, "High Gain, Low distortion, Faster Switching Transistor," U.S. Patent No. 5,196,809, Mar. 23, 1993; — "High Gain, Low Distortion, Faster Switching Transistor," U.S. Patent No. 5,430,413, July 4, 1995, a continuation of his earlier patent.

FORCE

In mechanics, identically the time derivative of momentum, or **F** ≡ d/dt(m**v**).

Simply, changing of the old concept of mass-motion. One must realize that it is not the mass in motion, but the single undivided quantity "mass-motion" or "mass-velocity" (momentum) that is changed. In short, **F** ≡ d/dt(m**v**). In present EM theory, force fields are erroneously taken to exist in empty, massless vacuum. However, the mechanical definition is an identity, not an equality. Thus force does not *separately exist in contact with* a mass, but *consists of* a changing "mass-motion." Since this force can be induced into and of the mass-motion target at a distance, in the new approach one is confronted with the fact that there is a more fundamental mechanism that produces force itself, and that is a curvature of spacetime.

Distant curvatures of spacetime to produce forces on exposed charged mass systems are also engineered by longitudinal EM wave interferometry.

FORCE FIELD

An effect that exists as a local vacuum virtual particle flux interaction upon and with a unit mass or unit charge whereby a particular type of matter is affected.

The force field has two components, one the mass and the other massless. In electrodynamics, one normally thinks (erroneously!) of the massless component, and refers to that component alone as "the force field." This considers the "force field" as a separate entity, acting upon a separate "mass." Whether it is logically correct or not, that is the prevailing way physics and electrodynamics uses it! So in this definition, let us momentarily accept the common incorrect usage, for convenience, while pointing out the real definition as we need to.

For example, conventionally the EM force fields can affect and change the state of motion of any charged particle or any magnetic particle. In classical physics, the force fields are considered the primary causative agents; when these fields are zero, then the motion of the particle or system is unaffected. Consequently, conventional EM theory assumes that all the EM phenomenology ceases once the force fields are absent. Today we know that is not true at all, as we explain in this glossary.

FORCE-FREE PROPAGATION

Propagation of force-free disturbances in the vacuum flux

(i.e., infolded inside the ambient vacuum potential, or inside a constant DC voltage.) Such "infolded" propagation is not limited to the speed of light, but may even travel in the Coulomb gauge so that it is instantaneous. More usually, it can travel superluminally. It may be regarded as a special form of "tunneling", which is known (both theoretically and experimentally) to produce superluminal propagation to the external observer. The music of Beethoven's 40th symphony, e.g., has been experimentally transferred between two points in a physical waveguide by such "quantum tunneling" at over four times the speed of light in vacuum.

We regard tunneling as moving in a sort of "subspace", i.e., within the internal structuring of spacetime itself. It is not limited by the speed of light and Einstein's postulate, which we believe only applies to signals moving in spacetime in bulk, and not to signals moving *within* substructuring inside that spacetime.

FOUNDATIONS OF MATHEMATICS

The concepts, operations, assumptions, postulates, logic, definitions, and axioms of algebra systems.

FOUNDATIONS OF PHYSICS

The postulates, assumptions, concepts, and definitions upon which a scientific physics model rests. Included are the postulates, concepts, assumptions, definitions, and operations of the mathematics in which the scientific model is expressed, as well as the postulates, concepts, assumptions, definitions, and operations of the physics entities and functions assigned to the mathematical symbols.

Thus "scientific truth"—even mathematical truth (see Morris Kline's **Mathematics: The Loss of Certainty**—is actually a *qualified* truth. What is true in one model (one set of postulates, assumptions, concepts, and definitions, including those in the algebra) may not be always true when one or more of the postulates, assumptions, concepts, and definitions is/are changed, and/or when the topology of the mathematics in which the model is expressed is changed, etc.

The major flaws in present physics models involve errors in the foundations. However, a successful physics model is rather quickly dogmatized and often turned into a belief structure by many scientists rather than being regarded as a tentative model. Consequently, scientists who attempt to do serious work in foundations are generally ignored or ostracized and rather vehemently attacked if they propose any significant changes to one of the dogmatized models. Particularly, all funding and "benefits" of the scientific bureaucracy are denied the heretical foundations scientist, unless he or someone can rather quickly come forth with an iron-tight experimental proof of the heresy. Any attempted change in foundations is usually met with ad hominem attack, a barrage of criticism, loss of position, and denial of research funding. Science has a long history of just such dogmatic oppression of great scientific discoveries. To name a few (there are hundreds): The first laser paper was rejected by the journals as being ridiculous. When Mayer proposed conservation of energy, he was hounded and ridiculed so fiercely that he suffered a nervous breakdown and attempted suicide. Years later, near the end of his life conservation of energy was adopted because it greatly facilitated understanding and calculation, and the same system now lionized him as a scientific hero. Ovshinsky suffered personal attack, ridicule, and charges of charlatanism in pushing amorphous semiconductors; everyone knew that only crystalline semiconductors could work. The situation changed when several thousand Xerox machines were observed to now contain amorphous semiconductors which were working just fine! Louis de Broglie's matter wave hypothesis caused consternation to his thesis advisors; but because he was Prince Louis de Broglie, they felt obligated to send it to the great Einstein so that he would reject it and then they could regretfully reject the thesis of the Prince. When Einstein replied that it was probably true, they felt they had no choice but to publish it *because de Broglie was a Prince of the old French aristocracy, and as such was not to be casually trifled with.* So foundations work is fraught with the greatest peril to the practitioner. Yet it is the most important work in physics, and it is a continuing black mark against the scientific establishment that they do not openly encourage *and fund* the most able physicists, grad students, and post-docs to examine and revise the errors in foundations concepts and postulates. Because of the prevailing dogmatism, science itself continually strangles the new ideas and concepts that would provide the physics of the 22nd century now. In an eerie turn of events, science has become in a strange way its own worst enemy, and the strongest barrier to the advance of science.

FOUR-SPACE (4-SPACE)

The "normal" spacetime used in physics, consisting of three spatial dimensions and one time dimension. Minkowski spacetime.

FOURTH LAW OF LOGIC

the law that a thing and its opposite become identical on the boundary, i.e., in multiple simultaneous observation. Identity is merely the loss of observational or perceptual distinction; when formerly "two" objects are simultaneously observed by a single observer, two slides have been shown at once in the slide projector. It does not matter at that point what each was individually observed to be before; now there is no observable distinction or difference. We previously showed that the law is implicitly assumed in the three laws of Aristotelian logic anyway, and published a simple proof of the fourth law. Without the fourth law, Aristotelian logic "eats itself". What results in the four law logic is that either (1) the three laws apply explicitly and the fourth implicitly, or (2) the fourth law applies explicitly and the fourth implicitly. This *application rule* itself can be taken as a fifth law of logic. The fourth law is also the law of the paradox: I.e., when something is true but violates one or more of the Aristotelian laws, it is a case where the fourth law applies explicitly. The conclusions derivable from a four law logic agree generally with the conclusions derivable from G Spencer Brown's **Laws of Form**.

FOUR-WAVE MIXING

A nonlinear multiwave mixing effect when four waves mix in a highly nonlinear situation so that wave to wave interaction occurs.

FRACTIONAL CHARGE

A hypothetical particle having a charge less than the electron's charge. Often called a quark, where its charge is 1/3 or 2/3 the charge of the electron.

A few years ago, Stanford University researchers found evidence of free quarks. There is also now the fractional Hall effect. In the 1920's Ehrenhaft reported the discovery and production of fractional charges of all sizes, and his experiments have been replicated by Michailov, with papers by Michailov and Barrett.

In the new approach, when time-energy is utilized, quarks can be freed or nearly freed rather easily in "time-reversal zones". It would appear that, as the new approach develops, free quarks will become a normal event rather than one intensely sought but terribly difficult to come by experimentally.

FRAME, LORENTZ

A frame of reference which is not accelerated with respect to the laboratory observer.

In this case, the frame may be rotated (moving at a constant velocity) but it is not rotating (the frame is not accelerated with respect to the observer, and spacetime is not curved). In this case special relativity applies, as do the conservation laws. If the local spacetime is curved, the frame is accelerated and spacetime is curved. Locally, then, general relativity applies. In this case, locally the conservation laws (energy, momentum, charge, and spin) can be violated. However, in ordinary general relativity, it has been tacitly assumed that locally the curvature of spacetime is negligible. Hence locally a Lorentz frame is assumed, and the conservation laws are assumed to apply locally also. In scalar electromagnetics, the local spacetime is always curved and general relativity applies locally. In this case, the conservation laws need not apply.

FRAME, REFERENCE)

A spatial, organized, measured lattice placed in "emptiness" (space, spacetime).

Normally refers to a 3-dimensional, spatial frame. All objects and points in the "universe" or spatial frame are considered to simultaneously coexist at separate, measured points in the frame. Differs from the vacuum in that, rigorously, vacuum has no existing definite lengths and no existing definite time intervals, as these appear only after measurement or detection, and are relative to the observer and to the interactions ongoing as well as in the detection process itself. The "laboratory frame" is the static reference frame of the observer or measurement. A separate reference frame may be assumed to exist for any fixed or moving object, or centered on any point in another frame. When a type of frame is assumed, the entire class of physical interactions that can occur has been restricted to an assumed set or type. In other words, given the frame, the *conventional* physics has been assumed. One of the greatest restrictions of an assumed "frame" is to rule out the consideration (existence) of other higher dimensions.

Note that, in the new unified field theory approach, the other higher dimensions are always available and cannot be ruled out in general, but only in some special case. Every curvature of spacetime, and any internal additional curvature to that primary curvature, adds a new dimension. In our view, a spacetime may be "flat" in overall curvature, but consist of internally structured deterministic curvatures or "engines". In this view, normal inertial frames, e.g., may still contain vacuum engines, which will not affect the normal *bulk translation* rules, but may affect any or all of the nontranslation mechanisms, including the very laws of nature in many cases.

FREE ENERGY

Excess energy freely received from its external active environment by an open system that is far from thermodynamic equilibrium in its exchange with that environment.

To use this free excess energy, it must be collected, transported to a load or loads with minimal losses, and then dissipated in the load(s) to power them, without any substantial portion of the free energy being used to close the system's "gate" through which the excess energy is input to it.

There is a separate thermodynamics for such open systems far from thermodynamic equilibrium in their active environmental exchange. Such a system is permitted to exhibit five "magic" functions: it is permitted to (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself inputs (the excess energy is freely received from the active environment, (4) power itself and its load simultaneously (all the energy is freely received from the active environment), and (5) exhibit negentropy.

The entropy of such a system cannot even be calculated. Some useful references are: Ilya Prigogine with D. Kondepudi, **Modern Thermodynamics: From Heat Engines to Dissipative Structures**, Wiley, Chichester, 1998; Ilya Prigogine, **The End of Certainty: Time, Chaos, and the New Laws of Nature**, Free Press, New York, 1996, 1997; G. Nicolis and I. Prigogine, **Exploring Complexity**, Piper, Munich, 1987; Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," Chapter 11 in Paul Davies, Ed., **The New Physics**, Cambridge University Press, Cambridge, 1989, p. 316-347.

FREE ENERGY SYSTEM

Any open system not in thermodynamic equilibrium, which freely receives excess energy from an external source,

...and collects and utilizes this free excess energy to freely power external loads, so that a permissible system COP>1.0 is achieved, as allowed by the nonlinear thermodynamics of open systems in disequilibrium with their environment.

Implicit in the notion of free electromagnetic energy systems are six key concepts: (1) the local vacuum/spacetime is known to be filled with hidden (i.e., virtual) energy in violent motion, i.e., the local vacuum/spacetime is a violent virtual energy flux and therefore a scalar potential, (2) a broken symmetry—i.e., some sort of organizing and gating mechanism, such as the common dipole—in this hidden (virtual) energy flux of the vacuum is utilized as a source, to cohere the virtual energy and divert it as energy flow $\mathbf{S} = \mathbf{E}\times\mathbf{H}$, from the broken symmetry initiation point onto the external circuits or load parts of a device, (3) the coherent gated extra *virtual* energy flow available in the external circuits or load parts couples to (interacts with) the mobile charges in those circuits or load parts, so that gradients developed in the interacting virtual flux form force fields on and of the particle masses, driving the mobile charges along the circuit and through current-impeding circuit components such as loads, (4) in the interaction of the driven charges with the driving virtual energy flux, their spin and erratic motions integrate a tiny "coupling fraction" of the driving virtual energy into observable (i.e., macroscopic field) energy, thus creating the **E**-fields and **B**-fields, (5) the current-impeding external components or load parts scatter or dissipate the coherence of the driven charges, and thereby dissipate the collected energy or change the form of the collected field energy, thereby producing useful work, and (6) little or none of the excess collected energy in the circuit is split off and dissipated in the gate (asymmetry component serving as the **S**-flow source) to destroy the gate and thereby destroy its asymmetry in the vacuum flux, which is what is extracting and producing the free energy flow.

FUSION, ACTION

In the new approach, the recombining of two canonical bits of a fissioned action quantum.

The fusion process "cancels" the bits produced by the fission process, and constitutes a negative spacetime operation. The physical world is thus eliminated, bit by bit, in a monocular fusion operation that produces the "past". The "mass" of a particle is taken to be proportional to, and identically comprised of, the absolute value of its fission and fusion rate. When fission and fusion rates of a particle are equal, the particle is unaccelerated; when unequal, the particle is accelerated. Vice versa, when the particle is unaccelerated, the fission and fusion rates are equal, and when the particle is accelerated, the fission and fusion rates are unequal.

GAIA HYPOTHESIS

The hypothesis by Lovelock that the earth and its biosphere is a living thing, at least in some sense.

In the new viewpoint, there exists a biospheric quantum potential which contains an extended mind and therefore is indeed an extended lifeform—one which includes the earth and the quantum potentia all its creatures and species. In the new sense, even so-called "inert" matter is alive, being continually bathed by energy flow dimensioned with lifeform templates, and therefore equally alive in all directions and nonpreferential.

GATE

Refers to something (a component or function) which extracts some part of a flux or flow, compacting it into a stream, and "sending it out."

A moving board wall or gate may be swung into a river, e.g., to divert or "gate" some of the river's flow out into a sluice ditch and on to a waterwheel powering a mill. Similarly, a dipole extracts some of the energy from the vacuum virtual photon flux exchange with the end charges of the dipole, and "gates" (diverts) the extracted portion of energy out from the ends of the dipole as energy flows. This creates the scalar potential surrounding each charge, and also the **E**-fields. Even a "static" dipole does this "gating" of vacuum energy into an energy flow.

The fundamental gate for all EM energy in 3-space is the dipole or dipolarity. The source of the 3-space EM energy continuously emitted by any charge or dipole is an equal inflow of EM energy from the complex plane (from the time domain). Thus EM energy flow is conserved in 4 dimensions but not in 3-space. In particle physics, it is well-known that any dipole or charge is such a broken 3-symmetry in its fierce vacuum energy flux exchange. However, particle physics considers the input energy to be virtual particle energy from the vacuum and therefore random energy. Whittaker 1903 shows it is actually from the time domain of spacetime (from the complex plane) and the input energy flow from the vacuum is not random but precisely organized in a harmonic set of longitudinal EM phase conjugate waves. Of course by wave-particle duality we simply consider the two pictures to be just different and permitted models of the same thing.

GATING EFFECT

The effect of extracting part of a flux or flow, compacting it into a useful stream, and "sending it out."

See *gate*.

GAUGE

In electromagnetism, usually considered a possible choice for (change of) electric scalar potential and magnetic vector potential, without any net change of force fields so that the change satisfies Maxwell's equations.

Interestingly, gauge field theory includes the axiom of gauge freedom. In EM systems, this means that one is free to "regauge" or change the potential energy of the system freely and at any time, so long as the change results in a static stress potential energy change. In the real world, one may have to pay a little bit for switching. But essentially physics already assumes that one can cause his EM system to take on excess potential energy at will, so long as it is a *stress potential* energy change only.

In short, only the *symmetrical* regauging case has been considered in electrodynamics. Yet every electrical power system must first potentialize the system *asymmetrically* rather than symmetrically. In the theoretically perfect system, this asymmetrical regauging is done freely. Hence every electrical power system first breaks gauge symmetry, in order to take on extra energy so that it can do work. Further, the source of the potential energy flowing into the asymmetrically regauged electrical system so freely, comes from the source dipole's extraction of EM energy from the time-domain (complex plane) due to its broken 3-symmetry in EM energy flow.

So the highest "gauge" principle is that energy can be freely extracted from the vacuum, anywhere and at any time, freely and easily, by making a dipolarity—which changes a single potential and asymmetrically regauges the system.

It follows that one is in theory free to "dissipate" this excess free energy in external loads to power them. Again, in the real world one may have to pay some switching costs, but that can be minimized by efficient switching. However, since no conventional electrical power system accomplishes this free powering of its loads, it rigorously follows that the conventional system must self-enforce symmetry in its discharge of the free potential energy achieved by the initial *asymmetrical regauging*.

Thus, if we believe gauge field theory—the most modern of physics theories and in many ways the most useful—and also believe its assumption of free asymmetrical regauging that is hidden in its gauge freedom principle, then COP>1.0 electrical power systems are absolutely permitted by nature and by gauge field theory where asymmetrical regauging is deliberately used.

It follows that some feature in electrical power systems as we presently design and build them, must be preventing these systems from exhibiting overunity COP. And so it is. The ubiquitous closed current loop circuit results in the circuit forcibly and symmetrically regauging itself while dissipating its excitation energy to power its loads and losses. This forces half the collected regauging (free) energy to be dissipated in the external circuit's loads and losses, and half of it to be dissipated in the source dipole to scatter the charges and destroy the dipole. This means that less than half the available free regauging energy is used to power the load, while half is used to destroy the entity—the source dipole—actually furnishing the free regauging energy from the vacuum.

Hence all our present electrical power systems are inherently free energy systems freely extracting their energy from the vacuum, but are fiendishly designed so that they kill themselves (shut off all free asymmetric regauging energy from the vacuum, by destroying the source dipole) faster than they catch some of the free energy from the vacuum to power their loads.

To restore the source dipole requires at least as much energy dissipation as was done to destroy it. In our conventional self-destructive systems, the generator shaft must thus be continuously powered in order to form the magnetic field inside the generator, so that this magnetic field can then dissipate its available energy to force the scattered charges back apart to form an ordered source dipole.

We pay the power company to have a giant wrestling match inside its own generators and *lose*. And in a sense, we do it because our theoreticians love gauge symmetry more than they tolerate broken gauge symmetry. So they build lovely, symmetry-restoring electrical power systems—and monstrously mangle and destroy the fragile biosphere in the process.

GAUGE FREEDOM

The axiom in gauge theory that one can freely change the potentials of an EM system, so long as the net new forces resulting from the change will sum to a zero vector resultant so that the system remains "symmetrical" in its force-field functioning with respect to the system prior to the change of potentials.

Gauge and gauge freedom are actually formal restrictions of an even more important asymmetrical regauging principle: *a potential in a system can be freely changed at any time, thus freely changing the potential energy of the system freely and at will*. Symmetrical regauging is achieved only by multiple application of this asymmetry rule to provide a net symmetry condition again, with a change in locked-in stress potential energy of the system.

GAUGE SYMMETRY

Abstract mathematical symmetry of a field, that relates to the freedom to regauge (change the value of) potentials, without affecting the values of the field quantities.

In short, the higher principle that the potentials can be changed freely and at will—and hence the potential energy of the regauged system can be changed freely and at will—is arbitrarily restricted to a very special case: Each change of potential involves the appearance of an new free force field. This *asymmetrical* change can be made freely and at will, at any time. However, the notion of gauge symmetry restricts any set of such changes to the special case where the new free force fields produced must sum to a net zero force field of opposing forces. In short, it severely restricts the freedom to change the potentials—to the special case where the resulting change of the system's potential energy must remain a static stress potential energy condition. Note that this arbitrarily eliminates any overunity functioning of the regauged system in which the altered potential(s) and altered system potential energy exist(s). See the discussions of *gauge* and *gauge freedom*, above.

GAUGE THEORY

A field theory using a field that has one or more gauge symmetries.

Note the accent on gauge symmetry and the net *symmetrical* regauging restriction. This restriction is a carry-forward of the old notion that the potentials are not primary, but the fields are primary causes of all EM phenomena.

For the overunity researcher, broken gauge symmetry is the interesting feature which must avidly be sought in his experimental systems. Hence he must seek out and familiarize himself with a wide variety of *broken symmetry effects* and phenomenology. He must also rigorously analyze his circuits with a keen view to where gauge symmetry can be broken, where and how gauge symmetry is automatically restored during excitation discharge, etc. Unless the automatic self-enforcing of symmetrical gauge transformation during excitation discharge is violated, no EM system can produce COP>1.0. In addition, without elaboration, the overunity researcher must also learn to consider the time-domain functions of the circuit, and particularly the giant negentropy of the common dipole. Unless he understands that all EM energy comes from the vacuum and from the complex plane via the broken 3-symmetry of a dipolarity, he cannot understand his own circuits.

Electromagnetics was the first gauge theory. Gauge theory is widely utilized today, and particularly in particle physics. In overunity power systems, the system must exhibit a broken gauge symmetry at least twice: (1) in initially potentializing the system to freely increase its potential energy, thus exciting the system so that it has energy with which to do work, and (2) then discharging this free excitation energy in the load without simultaneously destroying the source dipole providing the continual potentializing.

The conventional COP<1.0 electrical power system restores net 3-space symmetry by continually destroying the "free asymmetrical regauging" component: the source dipole, and using more of the energy freely received from the vacuum via dipole asymmetry to destroy the dipole than is used to power the load. This requires that at least as much 3-space energy be input to the system again, to restore the dipole and thus potentialize the external circuit, as was used to destroy the dipole. Hence as much 3-space energy has to be input as is dissipated in all the loads and losses, including inside the generator. For that reason, our scientists and engineers unwittingly but deliberately design and build only COP<1.0 systems.

Without elaboration, we also accent that spacetime curvature is always involved in an overunity EM power system, and the tempic (time) potential of the output end of an overunity system differs from the tempic potential of the input end, so that the gradient (the tempic force) is directed from output to input—exactly the reversal of the case for a COP<1.0 power system.

GAUGE TRANSFORMATION (IN ELECTROMAGNETICS)

In EM, gauge transformation is conventionally taken to be the addition of the gradient of some function of space and time to the magnetic vector potential—and simultaneously the addition of the negative of the partial derivative of the same function with respect to time, divided by the speed of light, to the electric scalar potential.

This procedure gives different potentials but leaves the electric and magnetic fields unchanged. It restricts gauge transformations to *net symmetrical* transformations. However, the increase or decrease in potential energy of the system, due to changing the potentials, results in a change in the stress of the system and a change in its stress energy. This is because two new antiparallel forces are created, which change the stress of the system.

In essence, gauge symmetry then involves the implicit assumption that a functioning system is the same regardless of how it is stressed, and regardless of the state of its stress energy. We point out that, gravitationally, this is not true since any change in the local potential energy of the system is a spacetime curvature change, hence a gravitational change.

GENERAL THEORY OF RELATIVITY

Einstein's theory of gravity in which the gravitational force is represented by a curvature in spacetime.

GIANT NEGENTROPY (OF THE DIPOLE)

The receipt from the time domain of spacetime (the vacuum), by any dipolarity, of enormous longitudinal EM wave energy flow, transduction of that absorbed energy from the complex plane into real 3-space, and emission of that enormous energy flow in all directions in 3-space, automatically and indefinitely after a little energy has been expended to separate the charges and make the dipole. The EM waves in the input from the time domain and the output in 3-space are perfectly ordered and correlated, which is a giant ordering of the vacuum/spacetime and is thus a giant negentropy.

In circuits, the source dipole in the generator or battery, once made, performs this function. Only a minuscule fraction (some 10^{-13}) of the energy pouring out of the generator's or battery's terminals strikes the surface charges of that attached external circuit and is diverged into the conductors to power the Drude electrons and the circuit. All the rest is just wasted.

See my paper, "Giant Negentropy from the Common Dipole," *ibid*.

GRAVITON

The quantum (smallest piece) of the gravitational field, with spin 2.

In the quantum theory of general relativity, the graviton is massless and has spin 2. In the new approach, a graviton is considered to be a paired (coupled) photon and antiphoton, each of which has spin one, so that the graviton couplet has spin 2. A scalar potential is comprised of gravitons, not merely photons, since it can be decomposed into a harmonic series of bidirectional wavepairs. In each wavepair, a wave and its antiwave precisely superpose. We insist on a further strong condition: that in the biwave, the photons comprising the wave and the antiphotons comprising the antiwave also must obey the *distortion correction theorem*. This means that the antiphotons and photons are passing through each other in opposite directions, continually coupling into gravitons and decoupling again.

GRAVITY

The phenomenon characterized by the physical attraction of any two material bodies.

Actually, in modern general relativity it is the trapped energy in the masses that produce the gravity potentials. Ideally, the gravitational force between two masses is proportional to the product of the masses divided by the square of the distance between them. In the new approach, one recognizes that mass itself is defined in terms of, and identically consists of, the time derivative of an action flux. [For a force-free fundamental definition of mass, see T.E. Bearden, **Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena**, National Technical Information System, Report AD-763210, 1973.] A photon flux is also an action flux. A virtual photon flux is an action flux. So the vacuum consists of a gigantic action flux—which with certain interactions forms mass itself. But since photons are action quanta and photon flux is action flux, and since mass and gravitation are functions imposed on such action flux; then mass, EM, and gravitation become different aspects of the same thing. All of them can be electromagnetically engineered, directly, as indeed can be the flow of time itself. By affecting the flow of time electromagnetically, one is thus able to affect the gravitational force between two objects electromagnetically. Antigravity, for example, can be achieved by a mechanism that is given in this glossary. Hutchinson has demonstrated the levitation of objects weighing up to 60 lbs., albeit sporadically and uncontrollably. At my urgent request, Sweet performed a highly successful antigravity experiment some years ago, reducing the weight of a 6-pound object by 90%, smoothly and controllably, in several smoothly controlled steps. I consider that the experiment tested and validated my approach to an engineerable mechanism for antigravity. For the results of the test, see Floyd Sweet and T.E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," **Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC '91)**, Boston, Massachusetts, 1991, p. 370-375.

GROSS PARTICLE TRANSLATION

Refers to translating the entire particle in space, rather than internally structuring the particle's potential (massless charge).

GROSS POTENTIAL GRADIENT FIELDS

The usual gradient in a scalar potential's gross magnitude, as in $-\nabla\phi$.

GYROELECTRONS

Spinning electrons considered as gyroscopes, so that when longitudinally perturbed they can and will laterally precess.

Considering (in a gross first order model) the conduction electrons in the Drude gas in a conductor as little gyros, because they are (i) spinning on an axis, (ii) severely restrained in their ability to move forward longitudinally down the conductor as current, and (iii) much more free to precess laterally in the conductor. Thus when the gyroelectrons in the receiving wire antenna, perturbed by an incoming longitudinal EM wave in the vacuum, are longitudinally perturbed, they precess laterally so that our instruments (which detect electron wiggles, not the incoming vacuum EM wave) do detect "transverse waves." However, they are detecting the transverse gyroelectron precession waves. The old electrodynamicists, long before the electron or the atom was discovered, simply envisioned the incoming EM wave as "waves in the thin material fluid ether" which were "intercepted" as electric fluid waves in the antenna. Hence since their instruments reported that the intercepted "electric fluid" was vibrating transversely, they firmly believed (erroneously) that this proved the ether wave to be transverse also, just as those "taut string waves" that Maxwell and Faraday had believed in. That is the total *raison d'etre* behind the mistaken notion of the transverse EM wave in the vacuum ether!

HARMONIC SERIES

Consisting of a fundamental frequency, its first harmonic, its second harmonic, etc.

HEAVISIDE, OLIVER

Noted English self-taught physicist and brilliant electrodynamicist, 1850-1925, who played a role in discarding Maxwell's quaternions, and also played a role in forming vector mathematics and formulating the vector reduction of Maxwell's theory from 20 quaternion equations in some 20 unknowns to the present 4 vector equations.

What are presently taught as "Maxwell's equations" in university are actually Heaviside's equations.

Lorentz symmetrically regauged the Maxwell-Heaviside equations, again greatly simplifying their mathematical solution—and also unwittingly discarding all permissible overunity Maxwellian systems in the process.

HEAVISIDE ENERGY FLOW COMPONENT

(1) The enormous nonintercepted and nondiverged energy flow component pouring out of the terminals of the battery or generator, and missing the external circuit entirely so that it is wasted.

(2) The enormous energy in the EM field, potential, or wave that misses the interaction of the field, potential, or wave with the assumed unit point static electric charge.

(3) See also discussion under Giant Negentropy.

HERACLITUS

To be added.

HERTZ, HEINRICH

Noted German physicist who together with Heaviside and others transformed Maxwell's theory from quaternion algebra to vector algebra, and who also produced EM waves, thus confirming Maxwell's theory.

HIDDEN BIWAVES

Those bidirectional EM wavepairs that comprise the scalar potential are indeed "hidden" with respect to our normal "electron translation" (i.e., electron wiggle) detectors and instruments.

HIDDEN VARIABLE THEORY

A class of quantum mechanical theories which considers that the quantum state of a physical system is not a complete specification after all. The "hidden variables" are those additional components necessary to provide the "complete state" of the system.

The hidden variables are those extra components of the theorized complete state, which are not contained in the quantum state. Bohm's hidden variables theory is the best known of all of these theories, and makes the same predictions as does the conventional quantum mechanics. In addition, it eliminates the terrible problem of the "missing chaos," eliminates the quantum measurement problem, and it suggests that quantum mechanics can be engineered if the hidden variables themselves can be engineered. In the new approach, one goes a step further by considering hidden EM variables infolded inside the potentials themselves, including inside the quantum potential utilized by Bohm. The advantage of this assumption is that it poses an *engineerable* interpretation, after work by Stoney, Whittaker, and Ziolkowski. Further, the Russians have utilized this type of EM hidden variable theory, along with the quantum potential, to develop secret advanced weapons. The physics of this area is called *energetics*.

HIERONYMUS, T. GALEN

Founder of the American version of radionics, who patented a radionics machine and process to start the cycle of radionics in this country.

HILBERT, DAVID

Noted German mathematician, 1862-1943, who did advanced work in geometry, theory of invariants, integral equations, and mathematical physicist.

Noted particularly for his development of Hilbert space.

HOLE (ELECTRON HOLE)

An unfilled vacancy in an ordering of electron energy levels, as in a lattice or in the Dirac Sea.

Electron energy levels have two energy states: (1) *positive energy* state, and (2) *negative energy* state. The positive energy states are associated with the positions of normal energy electrons. The negative energy states are almost all filled, and this is the Dirac sea.

The absence of an electron from either energy state creates a particular kind of hole. When an electron is absent from an electron positive energy state position in a lattice, that results in an excess of positive charge at that point vacated by the electron. Note that when an electron then refills that hole (recombination occurs), the electron remains in a positive energy state, does not disappear, and no radiation occurs.

When an electron is lifted from an electron negative energy state in the Dirac sea, that leaves behind a hole in the negative energy state in the Dirac sea. When an electron subsequently refills that hole, it "falls" in and disappears, decaying to a lower energy level and emitting radiation due to the difference in the energy levels.

The difference between holes in the Dirac sea and holes in the electron positive energy levels in a lattice, are very different. In short, they lead to two quite different kinds of "hole and electron" interactions. In the lattice "hole-filling" (recombination), there is no excess radiation given off and the electron does not disappear from the positive energy continuum. In the Dirac sea hole-filling, excess radiation is given off and the electron disappears from the positive energy continuum.

Under the influence of an electric field, a *positive energy* hole in a crystal lattice—where an electron has left and is therefore absent—will effectively propagate through the lattice as a positive charge, but usually much slower than the electron propagates in the opposite direction as an electron current. The process is actually more complex and involves continuous interaction with the electrons and their physical distribution pattern and location. The hole current is mathematically treated as a special positive charge current, often referred to as a positron current. In a current-carrying conductor or semiconductor, one has a current of electrons traveling in one direction and simultaneously a current of holes (hole positrons) traveling in the opposite direction, but recombining.

As stated, in addition to lattice holes, there also can exist unfilled holes in the Dirac negative energy states comprising the vacuum, particularly when fields are present. In this case, it is a *negative energy* hole. A positive energy electron that falls into a Dirac sea hole will emit energy as it falls to that lower energy state. To lift one of the electrons from the Dirac sea and leave a hole, energy must therefore be added to the negative energy electron in the Dirac hole.

When fields are present, Dirac holes may exist and may move as hole current in circuits under certain conditions. The situation is complex; see Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," **Adv. Theor. Math. Phys**., Vol. 2, 1998, p. 963-985. Updated copy (version 4) is on http://xxx.lanl.gov/abs/hep-th/9705006.

When a free electron fills a positive energy hole in a lattice (positive electron energy state), this effectively disposes of both the hole and the *free* electron which is now a *bound* electron, and is called *recombination*. The electron does not disappear. In a material lattice, there is no energy radiation from this, just the disappearance of the hole and the "binding" of a free electron so that it is no longer free.

In visualizing positive hole currents, passing the hole current though a surplus of available electrons (as in a negative charge donor material) will result in "hole filling" and recombinations. One may therefore speak of disposing of unwanted hole current by recombination, and in fact this is widely accomplished in semiconductors by controlling the doping.

HOLE CONDUCTION (LATTICE)

The movement of electron vacancies (holes) in a lattice continually left behind by hopping of the electrons as they break loose from their bonded positions and move freely in an electron current.

When an electron hops out of a lattice bond, it leaves behind a bonding gap or "hole" in the positive electron energy state at that bond site, so that an extra positive charge appears there at the hole. As the electrons move in one direction in a conducting lattice, the holes move in the opposite direction in a "positive current", but at a much slower pace than the electron current moves.

HOLE CURRENT (LATTICE)

See *hole conduction (lattice)*.

HOMEOPATHY

A system of medical treatment based on the theory that certain diseases can be cured by giving very small doses of substances which in a healthy person would produce symptoms like those of the disease. (Opposed to allopathy).

Homeopathy is based upon formation and use of dimensioned potentials (particularly the potential composed of the extremely rapid "making and breaking" hydrogen bonds in a fluid) containing vacuum engines. Any material or solute, and therefore any molecule, contains its own specific vacuum engine, its own charges, and its own dimensioned potential. Dissolving the material in water, or even using a suspension of it in water, allows the bonding potential in the water to interact with the potential of the solute or suspension. Potentials superpose, and they also diffuse and mix their inner structures (vacuum engines) into each other. So the material or solute diffuses its vacuum engines into the interior of the hydrogen bonding potential, in effect ordering and templating the hydrogen bonding itself! Vigorous shaking of the suspension or solution assures that all the fluid and H-bonding potential is exposed to the solute of suspended particles, so that a thorough exchange of vacuum engines occurs. This process in homeopathy is called *potentizing*. It's actually just creating the desired internal vacuum engines in the H-bonding potential of the water. Once that's done, there's really no further need for the solute or the suspended particles. By repeatedly diluting the solution to the extreme, then one can reach a point where there is a high probability that not a single molecule of the original solute is remaining in the fluid. Yet in each dilution, shaking can be used to insure the even spread of the vacuum engines through the newly introduced H-bonding potential in the newly introduced water. At the end, there exists an altered fluid whose H-bonding potential now contains the desired vacuum engines. Drinking that water now introduces those vacuum engines into the body fluids, bloodstream, etc. The ingested dilute new vacuum engines will then diffuse through the various potentials of the body, and produce altered chemical functioning of the kind that could have been induced by the chemical or active solute material. The advantage is that the harsh material residue of the chemical or solute is not present in the body, as it is when allopathy is utilized.

HYPERFUNCTIONING

Functioning in hyperspace as opposed to 4-dimensional Minkowski space.

Higher topological functioning.

HYPERSPACE

Refers to a space of more than four dimensions, and specifically to those spatial dimensions outside the normal three.

IDENTITY OF OPPOSITES

On the boundary, A and not-A are identical. If all of a thing is collected, one reaches the boundary of that thing and it turns into its own opposite.

INERTIAL FRAME

A frame of reference in which force-free bodies move along straight lines, and the postulates of special relativity are valid.

In the new approach, this notion is extended; even in an inertial frame, force-free bodies may move along straight lines, the postulates of special relativity may remain valid, and yet the EM hidden variable may induce other changes in the usual laws of nature due to the action of hidden vacuum engines infolded in the *dimensioned* local spacetime (vacuum potential).

INFOLDED HIDDEN DYNAMICS

Refers to:

(i) Bohm's hidden variable theory and the internal dynamics of the hidden variables/entities themselves, and

(ii) to the hidden biwave pairs comprising a scalar potential, with the internal dynamics of those waves such as tuning off harmonic frequency, multiplication, variation of phasing and magnitude, etc.

Also, by Whittaker 1903 the scalar potential can be decomposed into the bidirectional wavepair set in harmonic series. By Whittaker 1904, each of the waves in a biwave pair can be further decomposed into two potentials, so each pair of waves would consist of four potentials. That is, the entire set of internal biwave pairs can be replaced by four times as many internal "partial potentials" as was the number of biwave pairs. Then Whittaker 1903 can be applied once again to break each one of those scalar potentials into new biwave pairs, and so forth and so forth. So the potential can be expressed as waves within waves within waves, or potentials within potentials within potentials, and so on. The neat thing is that each successive infolding (internesting) is a higher dimensional functioning and a higher topology. So one can build (with some little pain!) specific infolded wave patterns that will work in—say—the 7th dimension of hyperspace, or the eighth dimension, and so on. This is why I called the internal structuring of a potential (or the two potentials comprising an EM wave) *dimensioning* the potential or the wave. As a half-baked nuclear engineer whom the Army never let practice his discipline, I wanted some way to directly engineer the nucleus. *That has now been found*. Unfortunately it has also already been highly weaponized in secrecy by one or more foreign nations. In other words, one can design a signal and assemble it (with some difficulty) that will flip one quark in a nucleon in an atom. That way one can transmute an element either into the next higher element isomer or the next lower element isomer. Using a quantum potential to "transport" the necessary vacuum engines, the effect can be accomplished at a great distance, in multiple targets simultaneously, and the effect is not shieldable. "Jumps" in nuclear transmutation can also be accomplished. The migration of electrons to and from an atom (as in solution, e.g.) makes and negates ions continuously. Thus continually the external potentials and the nuclear potentials intercommunicate. This means that their vacuum engines continually diffuse. Any dimensioning of the external potentials will be passed into the nuclear potentials as a dimensioning action also. If certain conditioning vacuum engines are present in the external potentials, then the slow diffusion into the nuclear potentials begins to affect the nucleons and their interactions, the quarks comprising the nucleons, the particles such as pions continually exchanged between nucleons, etc. The mechanism for nuclear transmutations and production of new species is there! I believe, e.g., and am working to more clearly express, that the major mechanism producing the new nuclear species in cold fusion experiments just such dimensioned potentials, involving the inadvertent dimensioning of potentials comprised of the hydrogen bondings of the water and other factors.

INFOLDED REAL VECTOR COMPONENTS

Well, simply take a set of real force vectors that altogether sum to a vector zero resultant.

That "vector zero system" is a zero with respect to forced *translation* of a particle or mass. It is not an absence per se, but is the presence of specifically hidden and "infolded" real vectors. Two elephants pushing each other to a draw form such a "zero-translating" system. That "zero system" is quite different from two crickets pushing against each other to a draw! Anyone who thinks all zeros are "absences" and who thinks all zeros are equal, is invited to try standing between the crickets and observing how he feels, then standing between the elephants and see how he feels. The latter might be a real bummer. So zeros can differ! And a zero can be a presence of many real things, all still there but "infolded" and "hidden," but just the absence of any *one and only one* thing. Even in arithmetic, e.g., we are free to pull 5-3-2 right out of 0, or pull out 4-7+8-6-2+5+1+1-3-1/2-1/2 (if I got the arithmetic correct!). And so on. So we have always *used* the fact that a net zero can have a real internal composition, without detailing it exactly. We teach our schoolchildren to do it, but fail to tell them what they are doing!

When you sum EM force fields to zero, the fields are still there and real. You have altered the energy of the system or volume, and also the potential energy of the system or volume. And you have done it in specific directions, magnitudes, etc. along the "fingers" of the infolded patterns. In short, you have made a specific vacuum engine. Mass, fluid, nuclei, molecules, etc. exposed to that "conditioned vacuum potential" may well undergo physical and chemical and electrical changes. Golden once performed experiments which powerfully conditioned the local vacuum, and all clocks and watches (mechanical, electrical, etc.) in the local area were awry for four days.

INFOLDING

With reference to signals, infolding means "placing the signal or modulation or change upon one or more of the internal bidirectional wavepairs comprising the scalar potential.

Since all EM potentials ultimately represent changes to the ambient vacuum potential, one has conditioned or altered the vacuum itself when infolding is utilized. Infolding of EM signals, vectors, and directions produces *vacuum engines*. By Whittaker's principle and Bearden's corollary, these vacuum engines can be designed to affect matter—including atomic nuclei, nucleons, etc.—in any manner desired.

INFORMATION CONTENT OF THE FIELD

Russian euphemism for the structuring of the hidden longitudinal EM waves and their dynamics, that comprise any EM potential, field, or wave.

INFRASONIC

Below sonic.

Frequencies below what the ear can hear. Below about 20 Hertz.

INTERFEROMETER

An instrument in which a wave is split into two waves or beams, which after traveling over different paths are subsequently reunited and display interference.

INTERFEROMETRY

Measurement of wavelengths of light and very small distances and thicknesses,

...for determining indices of refraction, and for analyzing small parts of a spectrum by means of the interference phenomena of light.

INTERNAL SPACE

To be added.

ITERATED

Done repeatedly, like picking beans one at a time.

JOSEPHSON EFFECT

Flow of Cooper pairs (superconducting electron pairs) across a thin dielectric separating two superconducting electrons, in the absence of a voltage drop.

That is, the pair of electrons can "tunnel" through a thin insulating barrier.

JOSEPHSON JUNCTION

The thin dielectric separating two superconducting electrons, with subsequent Josephson tunneling of Cooper pairs.

JOSEPHSON TUNNELING

See *Josephson effect*.

JOULE

A measure of energy.

One joule equals one watt for one second. In physics the unit of work or energy in the MKS system, being the amount of work done by one newton acting through a distance of one meter, equal to 10,000,000 ergs.

JUNG, CARL GUSTOV

Noted Swiss psychiatrist and psychologist who founded Jungian analytical psychology, formulated the principle of synchronicity, and advanced and delineated collective unconscious mind.

From the content of his German patients, Jung predicted WW II two decades before it occurred.

KINDLING

Coherent integration (adding) of similar virtual components among the vast "noise" of dissimilar virtual components, so that the "signal-to-noise" ratio of the integrated signal steadily increases, and the integrated "observable signal" grows out of the background virtual state noise.

This "kindling" process is directly analogous to coherent integration of nondiscernible radar signals far beneath the received noise, to build and extract the real radar return signals from targets. This sort of process dramatically increases the "sensitivity" of the radar. Real, very tiny signals even 100 dB or more beneath the noise level can be retrieved in such fashion by modern radars.

Similar processes can integrate very minute signals already passed into the virtual state and presumed to be nonexistent. Usually, however, the difference frequency between two frequencies at least one harmonic interval apart must be used, considering the vacuum's virtual flux state as an isotropic nonlinear medium.

One may also hypothesize that potentials and propagating signals may decrease with distance, but are never truly "lost" in that they never reach a total zero. When they reach the quantum level, further decrease in amplitude simply results in their passing into the virtual state (subquantum level).

KIRCHHOFF'S CURRENT-LAW

At any time, the sum of the instantaneous currents flowing into a node or point in a circuit equals the sum of the instantaneous currents flowing out of that node or point.

KIRCHHOFF'S VOLTAGE LAW

At any time, the sum of all voltage rises in a closed loop circuit equals the sum of all drops in that closed loop circuit.

LAMB, WILLIS EUGENE JR.

American physicist and Nobelian who discovered the hyperfine structure of the hydrogen spectrum, and first measured the interaction of the vacuum energy that generates a change in the energy levels of the hydrogen atom, according to Dirac's theory and quantum electrodynamics.

LAMB SHIFT

A very small correction to the first excited state of the hydrogen atom due to interaction of the vacuum.

Willis Lamb experimentally confirmed this prediction of quantum electrodynamics to high accuracy.

LAMELLAR CURRENT

Current in thin layers.

Thin layers of current. Thin sheets of current.

LASER

A device that generates light in which all the photons are exactly in step and produce a coherent beam.

Laser light has one wavelength and is more easily controlled than other kinds of light. Use of a laser to perform interferometry is one way to create holograms; use of a laser to illuminate a hologram is one way to read it. Originally the word "laser" was an acronym that stood for "light amplification by stimulated emission of radiation." Note that when TiO_2 particles in suspension, with a small amount of fluorescent dye added, are weakly illuminated by a small coherent light source (a laser), a tremendous room-filling output of coherent light emission occurs, by the anti-Stokes emission effect. This is lasing without population inversion, and it is a validated and easily testable overunity process. This is the process patented by Lawandy.

LASING WITHOUT INVERSION

The amplification of coherent light in an intensely scattering medium, comprised of certain powdered laser crystals or suspensions of similar material, where a variety of phenomena such as gain narrowing and amplifying random walk of light propagating in the medium, combine to give laser emission with diffusive feedback.

Gain tends to depend on the volume of the medium, so that at a certain medium the gain in light amplification by the medium exceeds the losses, and the laser action results above this critical volume. The intensity of the light scattering (multiply reflecting) in the medium then will increase rapidly until an "explosion" of light results in an intense flash of coherent light in all directions from the medium volume.

If the excitation is still present, the process will be repeated and the energy density will again exceed the critical value, rapidly generating another intense flash of light. This process can be repeated many times, given the presence of continued excitation.

A very interesting phenomena that can occur is coherent backscattering. Here the light scatters over both a forward-time path and a time-reversed path, perhaps with each path consisting of more than 1,000 scatterings. Because of time-reversal symmetry, the two waves (forward and backward) acquire exactly the same (though random!) phase. So the two waves after all this 1,000 scatterings on their paths, meet and interfere constructively in the direction back toward the exciting light source.

In theory the mean free path can be made shorter and shorter, so that all the light would be trapped inside the medium. This is called Anderson localization, and it has been reported in the literature; see D.S. Wiersma, P. Bartolini, A. Lagendiji, and R. Righini, "Localization of light in a disordered medium," **Nature**, Vol. 390, 1997, p. 671; John Sajeev, "Frozen light," *ibid*., p. 661-662.

This work grew out of original work called out by Letokhov but largely ignored for more than two decades.

References: D. Wiersma and A. Lagendijk, "Laser action in very white paint," **Physics World**, Jan. 1997, p. 33-37; — "Light diffusion with gain and random lasers," **Physical Review E54**, 1996, p. 4256; N.M. Lawandy et al., "Laser action in strongly scattering media," **Nature**, Vol. 368, 1994, p. 436; V.S. Letokhov, Generation of light by a scattering medium with negative resonance absorption," **Sov. Phys. JETP**, Vol. 26, 1968, p. 835.

Note that we have nominated a version of this process, together with consideration of the erroneously discarded Heaviside nondiverged energy flow component, as the fundamental mechanism proposed to be involved in gamma ray bursters, x-ray bursters, and in some gamma ray and x-ray flashes associated with atmospheric phenomena on earth in thunderstorms, etc. We also nominated the missing Heaviside energy flow component as the agent accounting for the excess gravity (some 90% of the total gravity) in the arms of spiral galaxies that hole those arms intact so they do not fly apart. See T.E. Bearden, "Dark Matter or Dark Energy?", **Journal of New Energy,** 4(4), Spring 2000, p. 4-11 (also available on http://www.cheniere.org); — "Giant Negentropy from the Common Dipole," **Proceedings of Congress 2000**, St. Petersburg, Russia, Vol. 1, July 2000 , p. 86-98. Also published in **Journal of New Energy**, 5(1), Summer 2000, p. 11-23. Also carried on DoE open website http://www.ott.doe.gov/papersbooks.html and on http://www.cheniere.org.

LAW OF PHYSICS

A restriction placed upon the physical mechanisms that can occur under certain broad assumptions as to the background situation and environment.

Actually, in nature there are no such rigidly fixed "laws"; these are only man's synthesis for the behavior of large classes of systems under very general conditions. Any "law of physics" can be violated if the conditions contained in its assumptions are violated. Conservation laws are often touted as "laws of physics." Actually, the conservation laws assume a closed system or one in equilibrium, and a linear spacetime (Lorentz frame). If either or both of these assumptions is locally violated, any or all of the conservation laws can be violated locally, even though the overall global conservation is still upheld. Quite simply, in that case one has opened the system (broken symmetry). In particle physics, broken symmetry already results in violation of one or more of the conservation laws. In leading Russian physics journals, papers are regularly published containing unrestricted general relativity with concomitant violation of any and all conservation laws.

LAWANDY PATENTS

Patents covering and related to anti-Stokes emission (excess emission of energy) in strongly scattering, optically active media such as TiO2 particles in suspension.

More energy is emitted than one must input to stimulate the effect. It is being hailed as a new form of lasing, without the requirement for population inversion.

LETOKHOV, V. S.

Russian pioneer in excess emission phenomena from stimulated media.

LETOKHOV EFFECT

The excess emission (overunity emission of energy) phenomena in strongly scattering, optically active media, usually referred to as "negative absorption of the medium".

For an overview, see V. S. Letokhov, "Stimulated emission of an ensemble of scattering particles with negative absorption," **ZhETF Plasma**, 5(8), Apr. 15, 1967, p. 262-265; —"Generation of light by a scattering medium with negative resonance absorption," **Sov. Phys. JETP**, 26(4), Apr. 1968, p. 835-839; — "Laser Maxwell's Demon," **Contemporary Physics**, 36(4), 1995, p. 235-243.

LOCALITY

Characteristic where interactions are local in spacetime.

LOCALIZED

Confined to a small region of a large system rather than being extended through the system or at a distance.

LOCAL VIRTUAL PHOTON FLUX OF VACUUM

Refers to a local region of the vacuum, which also refers to the vacuum flux density of that region.

LONGITUDINAL COMPONENT

In EM waves, the longitudinal component of the fields along the direction of travel,

...as opposed to the transverse components at right angles to the direction of travel.

LONGITUDINAL E-FIELD

We may divide any electrical field into two components, a transverse component and a longitudinal component.

Of course one or the other may be a zero vector.

LONGITUDINAL GRADIENT

Compression wave gradients (differences) in density or pressure,

...considered in a direction from their origin along the direction of primary movement, vs. transverse waves, which act perpendicular to that direction.

LONGITUDINAL INTERIOR WAVE

In conventional theory, the electric portion of the EM wave is a transverse field oscillation, and the longitudinal component is essentially zero.

In the new "infolded" wave approach, there is no overall transverse gradient in the vacuum potential, but there is an infolded longitudinal gradient which oscillates "inside the vacuum potential." This is rather like a "pressure wave" transmitted under water, without causing a surface wave involving physical upheaval and fall-off of the water. In EM, by analogy the speed of the "bulk upheaval" transverse wave is nominally c, the speed of light in vacuum [but see discussion of galloping waves, for what really happens!]. The speed of the pressure wave, under water, may be very much greater than c, because it is a special "tunneling" process and it also involves the galloping wave effect. Experimental measurement of superluminal tunneling velocities—such as more than 4c—are already well-known in the literature.

To better visualize this "pressure wave," consider that it is like a sort of "velocity modulation" upon the edge of the conventionally-visualized plane wave front of a traveling transverse wave, except now the plane itself does not move. All the succeeding planes are "bumped" by the velocity modulation of one plane, and so the velocity modulation propagates through (inside) the vacuum potential as a modulated pressure wave, without transverse change of the potential.

Now we have a very special medium! It consists of bidirectional EM wave pairs, one passing through the other. In such a case, it is already known that the speed of a "planar wave front" in the superposition and resulting interference of two oppositely moving waves can in fact greatly exceed the speed of light momentarily, then slow dramatically below than the speed of light, etc., so long as the average velocity is c. This is called the galloping wave effect. [E.g., see William G. Harter et al., "Galloping waves and their relativistic properties, **American Journal of Physics**, 53(7), July 1985, p. 671-679.]. Since the plane wavefront within the interior *can* in fact vary its velocity, we are quite free to use this "velocity pulse modulation" effect of wavefront speed variation, to place signal intelligence upon the planar wave front speed variations. So we can transmit hidden pressure waves with intelligence on them, through the interior of the vacuum potential. *Further, the restriction to an average velocity of c does not have to apply, and the planar pressure pulses can be transmitted much faster than c*. This is the basis of the work Fogal and I are struggling with, in order to achieve superluminal communication. It is also what we have been struggling with in an effort to come up with a communication system that can propagate infolded signals inside the H-bonding potential of the ocean itself, so that high-data-rate communications with deep submarines could be utilized, at any distance, and the huge, gawky, frightfully expensive, and environment-affecting ELF systems be done away with. We also think it should eventually be possible to develop an underwater radar type of system, but it's far too early at this point to try to conceptualize just exactly how it would work.

LONGITUDINAL WAVE

A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave.

Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes. In the past we have used the term "scalar waves" to imply longitudinal standing waves, such as in one infolded Stoney-Whittaker wavepair inside the scalar potential. In classical EM theory, the Poynting vector predicts no longitudinal wave of energy from a time-varying, electrically charged source. In fact, even in the ordinary theory, an exact solution of the problem does allow this longitudinal wave, as shown by Keech and Corum, **International Journal of Theoretical Physics***,* 20(1), 1981, pp. 63-68. Conventionally, as explained elsewhere in this glossary, the classical electromagnetic wave is modeled as a vector string wave, comprised of transverse vibrations on implicitly assumed material taut strings in the vacuum. In the new approach, such transverse vector force field waves (Hertz waves) do not exist *as such* in vacuum; instead, two coupled, dynamic scalar potentials exist there. The **E** and **B** fields are formed and exist as the these potentials and their gradient coupling to spinning charged detecting mass particles.

LONGITUDINALLY

Of or in length. Running or placed or operating or changing lengthwise. Opposed to transverse.

LORENTZ, HENDRIK ANTOON

Renowned Dutch physicist and Nobelist, 1853-1928, who developed the Lorentz transformations and Fitzgerald-Lorentz contraction.

He shared the Nobel Prize for discovering the Zeeman effect.

LORENTZ INVARIANCE

An inertial frame is independent of the velocity of the frame relative to any other frame.

LORENTZ INVARIANCE PRINCIPLE

The principle that the laws of physics are the same in all reference frames.

LORENTZ REGAUGING

Changing the two potentials in the Maxwell-Heaviside equations in potential form so that the resulting equations have the variables separated.

An unfortunate effect is that this discards all Maxwellian systems far from equilibrium in their vacuum exchange, hence discards all permissible Maxwellian COP>1.0 systems.

LORENTZ SYMMETRY

Referring to the symmetrical regauging by Lorentz of Maxwell's equations.

LORENZ, LUDWIG VALENTIN

Scientist who first changed Maxwell's equations—two years after their publication in 1865—by (effectively) symmetrically regauging them, thereby unwittingly discarding all Maxwellian systems far from thermodynamic equilibrium with the active vacuum.

See L. V. Lorenz, "On the identity of the vibrations of light with electrical currents," Philosophical Magazine, Vol. 34, 1867, p. 287-301. In this paper Lorenz gave essentially what today is called the Lorentz symmetrical regauging.

Comment: (gratefully received from T. Barrett) This paper gave f(t-r/c) functions. Fitzgerald said that Lorenz's functions were essentially the same as his, and Fitzgerald became a leading proponent of "retarded potentials". But it is believed that Fitzgerald was unaware of Lorenz's work until the 1880's, so he is given credit for parallel development. Some people talk of Fitzgerald-Lorenz functions. This is a regauging, but the term "gauge" (inspired by railroad gauges) was first introduced by Hermann Weyl in the 1900s. He used it for a change in length and was shot down by Einstein's theory—Weyl's theory was not relativistic. The idea resurfaced in the 1920s when quantum theory was being formulated, but this time it meant "change in phase" and not "change in length".

Comment by T.E.B.: Thus "Lorentz" regauging of Maxwell's equations really was first done by Lorenz in 1967. However, not too much attention was paid to L. V. Lorenz' work (e.g., by Fitzgerald). When H. A. Lorentz later used symmetrical regauging (essentially Lorenz' regauging), his influence was so great that it was adopted straightaway and has remained adopted to this day.

LOW ENERGY NUCLEAR REACTION (LENR)

Cold fusion, or the anomalous transmutations of elements which occur in certain electrolytes with specially prepared palladium electrodes.

The transmutations occur even though in conventional theory there is insufficient energy to accomplish them.

Conventional science has rather vehemently attacked cold fusion and many of the cold fusion scientists, even though there are now hundreds of successful experiments in multiple laboratories throughout the world.

In our view, once the properly-prepared palladium lattices are loaded with positive ions, the formation of temporary, fleeting time-reversal zones in the electrolyte temporarily reverses the law of attraction and repulsion of charges. Thus H+ ions, e.g., may attract each other in even numbers (opposite numbers are forbidden by the Pauli exclusion principle), so closely that each enters the strong (nuclear) force region of the other, so that the group becomes a quasi-nucleus. Note that the gluons are partially "unglued" by the TRZ formation, and so the preferred method of "decay" of the anomalous TRZ state as the surrounding electrolyte ions move to negate it, is by flipping of a quark. One D+ ion coupled to one H+ ion, e.g., when attracted into a quasi-nucleus in a TRZ, may decay into tritium by flipping one quark and changing one of the protons into a neutron. Two H+ ions may decay into deuterium (D+) by flipping one quark and turning an H+ into a neutron. Two D+ ions may simply "tighten further" into a real nucleus as the TRZ decays, and result in an alpha particle (helium nucleus). The TRZ hypothesis does precisely explain the most important reactions that are experimentally shown to occur. For more on this hypothesis, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," **Explore**, 8(6), 1998, p. 7-16; — ., "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," Journal of New Energy, 3(2/3), 1998, p. 12-28.

MAGNETIC DIPOLE

Paired north and south magnetic poles of equal strength over microscopic or macroscopic distance.

MAGNETIC INDUCTION

The act of a change in magnetic field producing an electric field at right angles, which in turn produces electron flow.

MAGNETIC SPIN

Refers to the intrinsic angular momentum, known as spin, of electrons, protons, and neutrons, etc.

The electrons combined in an atom or ion have a resultant angular momentum, that is the combined intrinsic spin of the electrons and the angular momentum due to their motion around the nucleus. There is also a magnetic moment associated with this angular momentum (spin) whenever it is nonzero. Therefore atoms or ions with nonzero spin are *magnetic* atoms or ions. This is often loosely referred to as "magnetic spin." It is actually the magnetic moment of nonzero spin.

MAGNETOELECTRONICS

New microelectronic components which exploit the spin of the electron rather than its charge.

References: A Barthelemy *et al*., "Giant Steps with Tiny Magnets," **Physics World**, Nov. 1994, p. 34-38; G. Prinz, "Magnetoelectronics," **Science**, Vol. 282, 1998, p. 1660.

MAGNETORESISTANCE

A change in electrical resistivity in a material when a magnetic field is applied to the material.

MAGNETORESISTANCE, COLOSSAL

A very large change in electrical resistance observed in the manganese perovskite materials when a magnetic field is applied to them. The effect in these materials is the largest magnetoresistance effect known at present.

See N. A. Babushkina *et al*., "Metal-insulator transition induced by oxygen isotope exchange in the magnetoresistive perovskite manganites," **Nature**, Vol. 31, 1998, p. 159; L. I. Balcells *et al*., "High-field magnetoresistance at interfaces in manganese perovskites," **Phys. Rev**. **B**, Vol. 58, 1998, p. R14697; M. Blamire *et. al*., "Understanding and utilizing colossal magnetoresistance materials," **Phil. Trans. Roy. Soc. A**., Vol. 356, 1998, p. 1469-1712.

MAGNETOSTATIC SCALAR POTENTIAL

Roughly, the "pole strength" differential between the two poles of a magnetic dipole.

Anyone who wishes may polish this definition.

MASSLESS DISPLACEMENT CURRENTS

Currents without mass.

Examples in the standard theory include $d\phi/dt$, $d\mathbf{E}/dt$, $d\mathbf{P}/dt$, $d\mathbf{B}/dt$, etc.

MATTER

Highly condensed 3-spatial energy, compressed by the factor c^2 …

and with the observation process d/dt having been applied to the fundamental masstime entity to provide mass as a frozen snapshot. Matter, of course, has internal structure, order and organization, and dynamics. All matter is in a continuous energetic exchange with its vacuum (4-spatial) environment.

MAXWELL, JAMES CLERK

Brilliant scientist who first formulated the unified theory of electricity and magnetism.

James Clerk Maxwell was born on June 13, 1831 in Edinburgh, Scotland. In 1847 he entered the University of Edinburgh, then transferred to Cambridge in the fall of 1850. After graduation, he stayed on at Cambridge in a research position. He was elected a Fellow of Trinity College and placed on the staff of college lecturers. In 1856 he returned to Scotland, where he took up a Chair of natural Philosophy at Marishal College, Aberdeen. In autumn, 1860 he took a new position as Chair and Professor of Natural Philosophy and Astronomy at King's College, London (a position he held until 1865, at which time he resigned). Maxwell was financially independent. He was elected to the Royal Society in 1861, while at King's College. From 1865 to 1871 he resided at his ancestral Scottish country home, Glenlair, developing his major ideas into book form. Maxwell returned to Cambridge in 1871, where he became the first holder of the Cavendish Chair of Experimental Physics. There he also supervised the construction and operation of Cavendish Laboratory. His treatise on electromagnetism appeared in 1873. He held his position at Cambridge until he died on Nov. 5, 1879 at age 48, of a form of stomach cancer—the same ailment that had killed his mother when he was a child. [Thanks to **Encyclopaedia Britannica**!]

MAXWELL (UNIT)

Unit of magnetic flux, in the cgs system.

MAXWELL'S ELECTRODYNAMICS

Maxwell's electrodynamics is Maxwell's equations.

MAXWELLIAN SYSTEM

Any physical or electromagnetic system whose electrodynamic operations obey James Clerk Maxwell's electrodynamics model.

The term is usually applied with respect to those systems which obey a highly reduced subset of Maxwell's after the Lorentz regauging of the Maxwell-Heaviside equations. This subset is now inappropriately referred to as "Maxwell's equations". This is unfortunate. As a result, today many scientists and most engineers no longer understand that Maxwell's theory—when embedded in a higher topology algebra—permits a vast richness of additional EM systems and behaviors, including full unified field theory operations. In the standard U(1) gauge symmetry electrodynamics, all these higher symmetry functions and systems are excluded *a priori*. From our view, this is particularly sad since these arbitrary reductions of Maxwell's theory arbitrarily excluded all EM systems far from equilibrium in their exchange with the active vacuum. Hence almost all scientists and engineers believe it is against the very laws of nature to propose an electrical power system that produces more energy out—and more work in the load—than the energy input made by the operator himself. This sad error has resulted in (1) crippled electrical power systems that extract their output energy from the vacuum but kill themselves faster than they power their loads, (2) the giant pollution of the earth, (3) extraordinary energy costs, (4) struggles and wars over oil and energy resources, and (5) the looming giant economic collapse of the economies of many nations. Further, unless this latter facet is checked, it may also lead to (6) the unleashing of all the weapons of mass destruction in the arsenals of many nations, destroying all civilization and much of the biosphere.

McCREA

To be added.

MICHELSON, ALBERT ABRAHAM

German-born American physicist who invented an interferometer (Michelson interferometer) named after him. Michelson used his interferometer in a renowned experiment designed to measure the velocity of the earth through the ether, which would produce a shift in his interferometer pattern. No such shift was observed, which indicated the absence of an ether wind and was largely responsible for the downfall of the concept of the material ether which had been assumed previously.

Reference: Albert A. Michelson and Edward Morley, "Influence of Motion of the Medium on the Velocity of Light," **American Journal of Science**, Vol. 31, 1886, p. 377-386; — "On the Relative motion of the Earth and the Luminiferous Ether," **ibid.**, Vol. 34, 1887, p. 333-345.

MIND

From the medical view, with thanks to **Webster's Medical Desk Dictionary**, Merriam-Webster Inc., Springfield, Massachusetts, 1986: (1) The element or complex of elements in an individual that feels, perceives, thinks, wills, and especially reasons; (2) the conscious mental events and capabilities in an organism, (3) the organized conscious and unconscious adaptive mental activity of an organism.

See an extended discussion of *mind* under *energy*.

MIND-BODY COUPLING

To be added.

MIND-MATTER INTERACTION

The n-space interaction between the time-like mind and its time-like operations and the 3-spatial physical body and its 3-space operations. To first order, we may treat these interactions in four dimensions.

MIND OPERATIONS

To be added.

MOTION

To be added.

MOTIONLESS ELECTROMAGNETIC GENERATOR (MEG)

Transformer-like COP>1.0 electrical power generator invented by Bearden, Kenny, Hayes, Moore, and Patrick which powers the transformer core with a permanent magnet, but separates the curled magnetic vector potential A (the magnetic field B) from the uncurled magnetic vector potential A, so that the magnetic field flux is retained in the core while the A-potential is replenished outside the core and adjacent to it.

The overunity mechanism is thus provided by a version of the Aharonov-Bohm experiment, which proves that this splitting of the A-potential does occur and that the uncurled A-potential will interact energetically with electrons. In short, by "splitting the potential" and potentializing two paths to the same energy density as the original unsplit potential, the MEG doubles the available potential energy available via the permanent magnet dipole's broken 3-symmetry in the active vacuum. Both the curled A-potential (magnetic field) inside the core and the uncurled A-potential outside the core interact simultaneously with the output coils when pulsed. Thus the output coils "doubly interact" instead of singly interacting. This "double-dipping" of available energy flow allows permissible COP>1.0, since the system is far from equilibrium with the active vacuum exchange.

Pulsing applies the Lenz law effect to both the split potentials, further increasing the energy density of the two interactions with the output coils. The MEG nominally exhibits from COP = 1.7 to COP = 10, using two input coils.

See M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "Classical Electrodynamics Without the Lorentz Condition: Extracting Energy from the Vacuum," **Optik** 111(6), 2000, p. 246-248; — "Inconsistencies of the Maxwell-Heaviside Theory of Electrodynamics: The Aharonov-Bohm Effect," **Journal of New Energy**, 4(3), Winter 1999,p. 236-240; — "Vacuum Energy Flow and Poynting Theorem from Topology and Gauge Theory," submitted to **Physica Scripta** (in review); — "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," submitted to **Optik** (in review), — "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics, **Foundations of Physics Letters**, 2001 (in press). See also T.E. Bearden, "Energy from the Active Vacuum: The Motionless Electromagnetic Generator," in M.W. Evans (Ed.), **Contemporary Optics and Electrodynamics**, Wiley, 2001, 3 vols. (in press), comprising a Special Topic issue as vol. 114, I. Prigogine and S.A. Rice (series eds), ***Advances in Chemical Physics***, Wiley.

MULTICOLLECTION

Multiple collection of the same energy flow, by rerouting the passed Poynting energy back around to the collectors for another pass by them, with additional collection.

A common means of "returning" the Poynting flow for additional collection is retroreflection, as in phase conjugate reflection. Since a nominal energy collection fraction on a single pass is only about 10^{-13}, a great many re-passes of the same energy can be returned. Work is done on the collector in each pass. One joule of energy can actually perform many joules of work; this is a change to the usual statement of the work-energy theorem, which assumes single pass, single collection only.

MULTIPASS RETROREFLECTION

Iteratively retroreflecting the same Poynting energy flow, each time it impinges on and passes by an intercepting collector, so that multiple passes of the same energy flow are made onto and by the collector.

More and more of the passing enormous energy is collected as more and more passes by are made. The collected energy is then dissipated as work in the circuit or device. Such processing can allow overunity EM circuits and devices.

MULTIPASSES

Multiple passes. See multipass retroreflection.

NANOTECHNOLOGY

The fabrication of devices with atomic or molecular precision, to produce operating devices with minimum feature sizes less than 100 nanometers.

Often described as a marriage between chemistry and engineering. The self-assembling of the units is referred to as *molecular manufacturing*, and the devices are built one atom or molecule at a time with programmed nanoscopic arms. As occurred with the modern machine and technology age, the first requirement is to develop the tools that do the working, and the "tools that build the tools." That fundamental work is off the ground and running already.

Nanomanufacturing plants capable of manufacturing bulk materials or arbitrary structures with atomic precision are envisioned, with almost every atom being placed precisely as desired.

Nanotechnology may well produce the next great scientific revolution. As an example, nanobots have been visualized that will circulate in a human body's bloodstream and specifically attack cancer cells. Nanobots for cleaning up air and water pollution are possible. Food synthesis is envisioned since the technology provides the ability to combine molecules directly.

As with all technologies, the ethics lies in the hands of the system designer and the user. Increased technology always allows greater use for good and greater use for bad purposes. Because of the extraordinary "evil uses" that can be employed in the use of nanobots, Bill Joy, chief scientist of Sun Microsystems, recommends no further development of nanotechnology and no further research. He stated,*"... we have the possibility not just of weapons of mass destruction but of knowledge-enable mass destruction… hugely amplified by the power of self-replication… we are on the cusp of the further protection of extreme evil…"*

Joy's fears are well-founded. Already the union of artificial intelligence with self-replicating nanobots is foreseen. If that union is successful, one can foresee the unleashing of what can only be called "species" of self-replicating, smart nanobots bent on the mass destruction of humanity in various ways.

Also, contrast the nanobots to our discussion of even more advanced nanotechnology of causal system robots (CSRs) using functioning system of causal fields (sets of longitudinal EM waves and their dynamics, prior to having interacted with charged mass, and thence being systems that are functioning sets of spacetime curvatures). We have already survived one massive world-wide insertion of an armada of such CSRs—fortunately first generation systems and with no countermeasures, so they were easily detected and destroyed. We may not survive the next such insertion and attack, as the maturity of the CSRs increases.

Considering the usual focus of governments and human control systems, my prediction is that *detrimental, self-replicating, smart* nanobots for use in war, surprise strategic attack, and mass destruction of human targets is already on the early planning boards. The obvious war applications of nanotechnology will advance faster than the peaceful applications, behind a public screen of simpler benevolent developments. This is almost certain, since the lead time for maturing the nanotechnology is from 10 to 20 years. This long lead time rules out most private capital ventures of substantial magnitude. It results in the requirement for government funding, just as development of the atomic bomb was government funded. As a nanotechnology "arms race" begins and heats up, one will see the equivalent of new but hidden "Manhattan Projects" in specialized mass destruction weapons nanotechnology. Rather substantial funding of nanotechnology research and development by several nations is already a fact.

Unethical aspects aside, the potential for useful and benevolent purposes of nanotechnology is very great. The development in this new field is now beginning to escalate almost exponentially, year by year, funded primarily by governments and to a more limited extent by private companies.

Reference: David Cook, Jessica Voorhees, and Brian Valz, "Nanotechnology: The Coming Molecular Revolution," available on website http://nanotech.about.com.

NEGATIVE ABSORPTION OF THE MEDIUM

Excess emission from a medium, as in the Letokhov work and in Bohren's experiment. The excess energy is received from the vacuum interaction.

NEGATIVE RESISTOR

Any component or function or process that freely receives energy from outside the system in unusable or disordered form and outputs that energy in usable, ordered form inside the system, where that is the net function performed.

We specifically do not include "differential" negative resistors such as the tunnel diode, thyristor, and magnetron which dissipate and disorder more energy from the system, overall, than they order and furnish to the system in their negative resistance regime.

NEGENTROPIC

Functioning to produce or involve negentropy.

Reordering previously scattered energy without requiring the operator or experimenter to input energy to do the reordering work (an example is passive retroreflection of scattered EM energy).

NEGENTROPIC ENGINEERING

Expending a little bit of energy to form a dipole or dipolarity, and then intercepting, collecting, and using some of the enormous EM energy freely extracted by the dipole from the vacuum indefinitely, without using any of the extracted EM energy to scatter the charges in the dipole and destroy the dipole, thereby shutting of the vacuum's furnishing of copious free energy.

NEWTONIAN 3RD LAW

For every action there is an equal and opposite reaction.

This law should be extended and restated something like this: "for every action there is an opposite and equal reaction if the causative mechanism of the reaction is allowed to occur and not first redirected away from its target." An example of a violation of Newton's third law as usually stated, is provided by the emission of a phase conjugate replica wave from a phase conjugate mirror material. No matter how powerfully pumped, the PCM does not recoil. The reason is that the antiwave cause of Newtonian third law recoil was intercepted by nonlinear multiwave interaction before it could reach its target nuclei, and rediverted away from the atom and on out of the material. Since the causative mechanism for Newtonian recoil did not occur, the PCM did not recoil.

NEWTONIAN RECOIL

As used in the paper, refers to Newton's third law reaction created as the recoil of the nucleus of an atom which absorbs an incoming EM longitudinal wave, or emits one.

NEWTON'S THIRD LAW

For every action there is an opposite and equal reaction.

In Maxwell's electrodynamics, Newton's third law is missing. It is missing because of the coarseness of the modeling which Maxwell was forced to use. Maxwell's seminal paper was read in 1864 and published in 1865. At the time, neither the electron nor the atom had been discovered, much less the nucleus of the atom. A molecule was just a blob with no structure. Electricity was considered to be a thin material fluid, flowing through the wires much like water through a pipe. In fact, Maxwell actually wrote a material fluid flow theory, since hydrodynamics was developed at the time.

So there was, at the time, no notion of the Drude electron gas in a wire and the positively charged nuclei in the atoms in the wire. So there was no notion of two oppositely charged entities in the conductor, both reacting to any EM field entering from space and interacting with the wire.

Further, Newton's third law had always been used mystically without any known cause. It was experimentally established that the reaction occurred, but scientists were forced to just assume its occurrence.

Consequently, Maxwell wrote equations for the reaction of a unitary electric fluid in the conductors. He knew that the wire recoiled, but wrote no equations containing the cause of Newton's third law reaction.

What really happens in the wire is that both the Drude electrons rearrange their positioning vis a vis the lattice, resulting in an apparent current. The actual movement (drift velocity) of the electrons down the wire is very slow, in a nominal case being a few inches per hour. The electrons, having spin, act as little gyros so to speak, longitudinally restrained by the repulsion of the electrons ahead. So the electrons precess laterally, and spend most of their movement in lateral precession movement. This alone proves that the perturbing force is longitudinal, as indeed does Whittaker's decomposition of the scalar potential into longitudinal EM waves. The gradient of the scalar potential (the E-field) is also such a decomposition. But the electron precession waves—which is what our instruments measure in the wire—are indeed lateral waves, in general consonance with Faraday's notion that the force field lines were physical and like taut strings. Maxwell merely assumed Faraday's notion, to give us the transverse EM wave in space. Consequently, the old electrodynamics envisioned that fluid vibrations (ether vibrations) from space had been intercepted by the wire to shake the electric fluid accordingly. And since that electric fluid obviously shook laterally, that "proved" that the waves in spacetime were lateral "plucked taut string" type waves. Actually it proved no such thing, because they were measuring electron precession waves.

The positive nuclei usually have a m/q ratio several hundred thousand times the m/q ratio of an equal charge of electrons. Hence the massive nuclei also react to the field interaction, and move very slightly in the opposite sense from the electrons. That reaction is ignored in Maxwell's general equations, though when the equations are applied, they require that nuclei movement.

In short, the electrons precess and the nuclei recoil oppositely, and that is the generation of Newton's third law reaction.

When we excite a wire antenna with a wave excitation, we produce both electron precession waves and nuclear recoil waves in the nuclei. These two waves are of equal energy, though the nuclei waves are very highly damped in magnitude because of the high m/q ratio of the nuclei. Nonetheless, each of these "lateral" waves slaps the vacuum potential with *identical energy*. So two equal waves of spacetime curvature occur. I can be seen that the net vacuum wave is actually a longitudinal EM wave, composed of waves of vacuum (spacetime) energy rarefaction and densification. Electrodynamicists, considering the vacuum to be filled with a material ether, used the "effect" wave in matter as the cause wave. Hence they used a transverse EM wave in spacetime, when indeed the spacetime wave is and always was a longitudinal EM wave, just as Tesla insisted.

Maxwell's equations are based primarily on the interaction of field and single charge. Actually, the reaction is always between the field and a dipole. Even when the field interacts with an "isolated charge", there are clustering virtual charges of opposite sign around the "isolated" charge so that the "isolated charge" is actually a set of composite dipoles, where each dipole consists of a virtual charge of opposite sign and a differential piece of the isolated observable charge. So the Newtonian third law reaction is also there with an isolated charge, but occurs in the vacuum itself, resulting in vacuum polarization.

Nonetheless, the "transverse EM wave in vacuum" notion is so embedded in electrodynamics that we seem to be stuck with it. It means, however, that the "field in space" used in electrodynamics must be regarded as only a sort of "convenient transfer function" and not a real transverse oscillation entity. This of course accords with Feynman and Wheeler's pointing out that only the potential for the EM wave exists in space, in case a charge is brought in to interact with its gradients.

And needless to say, since the vacuum is a big potential and only potentials and potential changes can exist in it, then Whittaker's 1903 paper decomposing the scalar potential into longitudinal EM waves is paramount and decisive. EM waves in the vacuum a priori must be longitudinal waves, because they are only the changes to the vacuum potential. And that potential at each and every point, changes and all, decomposes into longitudinal EM waves in 3-space with convergent longitudinal EM waves incoming from the time domain.

NISBET

To be added.

NON-ABELIAN ELECTRODYNAMICS

To be added.

NONLINEAR MATERIAL

A substance whose optical properties vary with the intensity of the light.

Most optical substances are very weakly nonlinear, but a great deal of research has been exerted to find materials with stronger nonlinearities. Also, a material which creates and emits the fundamental and additional harmonic frequencies when illuminated with a single frequency. Many nonlinear materials also weakly produce time-reversed (phase conjugate) fundamentals and/or harmonics when illuminated with a fundamental frequency. Such a nonlinear material that creates time reversed waves (phase conjugate replica waves) is said to function as a *phase conjugate mirror*. The phase conjugate replica reappears in space back along the path in space—even through a highly distorting medium—where the input stimulus (signal wave) has traveled or is traveling.

Thus the phase conjugate replica or time-reversed wave normally backtracks its signal wave stimulus to the distant source of that signal wave, even hundreds or thousands of miles away. This "crawfishing" and "concentrating" characteristic is also known as *retroreflection*. Note that an ordinary EM wave *diverges* as a function of the radius traveled away from its emission source, while a time-reversed or phase conjugate replica wave will *converge* as a function of the radius traveled away from its emission source (the mirror), if its stimulus signal wave was emitted from a distant point source. Ideally, a phase conjugate mirror produces a single time-reversed wave in addition to the input signal wave. By stressing *(pumping)* the mirror with the input of two additional opposing waves (ideally orthogonal to the input or *signal wave*), the phase conjugate replica wave output by the mirror is amplified, and may contain up to all the energy in the pumping waves.

NON SEQUITUR

A conclusion or inference not following from the original argument or premise.

NONEQUILIBRIUM THERMODYNAMICS

Colloquial term referring to the new thermodynamics of self-organizing systems and structures in disequilibrium with their environment.

Since the 1960s, there has emerged in physics (and in thermodynamics!) a surprising new area: the physics of far-from-equilibrium systems and the self-organization and large-scale ordering of such systems. E.g., a single Bénard cell (a particular self-organizing circulation structure in convection-type systems), may involve the coherent behavior of some 10^{21} molecules, in a region of convection where one would classically expect only randomness. These new physics and thermodynamics interactions are surprisingly being found in a very wide variety of systems and phenomena, including living systems and nonliving systems. The phenomena of interest include dissipative systems in disequilibrium with their environment, but with stability of their own systems and forms. Open systems in disequilibrium with their environment can permissibly achieve COP>1.0, and they can exhibit local negentropy. The physics and thermodynamics of such systems is still rapidly developing. For a good overview, see Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," in **The New Physics**, edited by Paul Davies, Cambridge University Press, 1989, p. 316-347. For a full technical presentation, see G. Nicolis and I. Prigogine, **Self-Organization in Nonequilibrium Systems**, Wiley, New York, 1977.

NONLINEAR MATERIAL

In physics, any substance whose response to an external causative or influencing agent is not proportional to the magnitude of the causative agent or influence.

NONLINEAR OPTICAL FUNCTIONING (NLO FUNCTIONING)

Functioning in a manner prescribed by nonlinear optics and quantum optics,

...particularly with respect to time-reversal effects, use of phase conjugate mirrors, phase conjugate reflection, and pumping for amplified phase conjugate reflection.

To quote Knight, "A phase conjugate mixing device has remarkable optical properties: perfect retroreflection, the complete cancellation of optical aberrations and inhomogeneities in wavefronts, and the ability to reverse an optical beam and return it to its source."

NONLINEAR PHASE CONJUGATE OPTICS

See *phase conjugate optics*.

NONLOCALIZATION

Producing or rendering an effect or effect-producing agent at a distance, rather than at the immediate location of the effector.

Extended throughout a large region or system, or to a distant part of a large region or system, rather than being confined to a small region of the large region or system.

NON-RIEMANNIAN (GEOMETRY)

Riemannian geometry is the geometry of Riemannian manifolds;

i.e., manifolds having an infinitesimal squared distance ds that is everywhere positive. A non-Riemannian geometry violates this property. [A manifold is a topological space with certain technical characteristics.]

NUCLEONS

Name for the neutrons or the protons in a nucleus. In the nucleus, the neutrons and protons are continually changing one into the other by exchange of the charge.

O(3) ELECTRODYNAMICS

To be added.

OPEN SYSTEM

A system that communicates with its environment, and exchanges energy, matter, or both between system and environment.

The system may be in equilibrium with that exchange, so that it effectively does not retain and utilize received environmental energy to power loads or accomplish external work. Note that it may in fact receive and store excess energy from the environment, or give up energy to the environment, and then remain in equilibrium under its new potential energy conditions. That is, the system may asymmetrical regauge itself but not utilize any of the regauging energy to accomplish external work. When potential is applied to an electromagnetic system, the system is similarly asymmetrical regauged and its potential energy is increased. However, if the system then destroys its source of regauging potential energy faster than it powers its load, it is said to symmetrically regauge itself when discharging its freely received excitation energy.

The system may be far from equilibrium in its energetic exchange with its environment, and if it asymmetrically discharges its freely received excitation energy, it can exhibit COP>1.0. In that case the system effectively retains and utilizes some of its freely received environmental energy to power loads or accomplish external work, and the operator himself does not have to furnish that part of the energy dissipated as useful work.

OPTICAL SWITCHING

Refers to switching of optical signals, or switching using optical frequency switching techniques, components, and circuits.

ORTHODOX TRANSISTORS

Ordinary transistors in use today.

Such transistors will not perform the Fogal semiconductor special functions.

ORTHOGONAL ROTATION

Simply speaking, a rotation from one direction to a direction at right angles to it.

The Lorentz transform of special relativity, e.g., is a formula for the rotation of an object and all its parts away from its line of motion, as a function of the object's velocity, in a direction toward the time axis. The speed c just represents a full orthogonal turn, so that the intersection of the object with the observer 3-space, along the original line of motion, is zero.

To the observer, the three dimensional object (at speeds less than c) has become a two-dimensional plane moving at speed c. In a space of more than 4 dimensions, the orthorotation may be toward a higher spatial dimension rather than toward the time dimension. As can be seen, such orthorotation is intimately connected with the transformation of an electron or electron/positron pair into a photon, and vice versa. Orthorotations are extremely important when one models the physics of mind and matter interaction, and the physics of a living organisms including both its mind and its physical body—as in psychoenergetics.

OSCILLATING QUANTUM WELL

A quantum well which is oscillating in magnitude and sometimes in sign.

OVERPOTENTIAL

Essentially the overpotential (as in an electrode in a solution) is a shift in the Fermi level necessary to allow the electron in the electrode metal to have energies overlapping with vacant acceptor levels in molecules adjacent to the electrode in the solution.

The overpotential effect appears at small "double surfaces" of things in very close contact, particularly in electrochemistry and in solid state physics. The overpotential enables electron transfer, e.g., by tunneling. The overpotential increases with the log of the reaction rate that occurs at the overpotential location.

For a small reaction rate, the overpotential is small; but for a small increase in the overpotential there occurs a *dramatic* increase of reaction rate. Fogal has utilized the overpotential effect in his semiconductor, to great advantage in accomplishing unusual charge blocking and other special effects.

One of the world experts in the theory and use of the overpotential is Dr. J. O'M. Bockris. E.g., for a short overview pointing out the importance of the overpotential, see Bockris, "Overpotential: a lacuna in scientific knowledge," **Journal of Chemical Education**, 48(6), June 1971, p. 352-358 (a paper which Bockris graciously sent me some years ago). Quoting from that reference, p. 355: "The formal definition of overpotential is 'the change of potential of the electron-conducting phase when reaction rate across its interface with the ion-conducting phase with which it is in contact is changed from zero to a certain velocity.'" And again, p. 355: "...overpotential... in fact determines what fraction of the energy of the self-functioning cell has to be wasted in making the cell work and deliver energy at a certain power." For a more detailed and wider exposition of the overpotential, its effects, and its uses, see J. O'M. Bockris and A. K. N. Reddy, **Modern Electrochemistry**, Vol. 1 & 2, Plenum Press, 1970.

OVERUNITY COEFFICIENT OF PERFORMANCE

More energy out or work out of a system than the operator or experimenter must himself or herself furnish and input.

An open thermodynamic system far from thermodynamic system is permitted to achieve such an overunity performance, so any overunity electrodynamic system must be such. Well-known nonelectromagnetic examples are solar cells, heat pumps, windmills, sails on sailboats, waterwheels for powering watermills, a gliding bird in thermal updrafts, etc. A well-known electromagnetic example is the common solar cell. The common household heat pump has a theoretical COP limit of 8.22, and well-designed units usually achieve some 4.0 or so, so long as the ambient environment does not change too drastically from the design limits. As the environment changes and cools down appreciably, however, most home heat pumps lose efficiency and have to switch onto heating strips, which have a COP ≈ 1.0.

Contrary to prevailing belief, it is certainly possible to have EM systems which are also open thermodynamic systems, far from equilibrium, and freely receiving excess energy (asymmetrically regauging a potential) from an external source, to wit, their violent virtual flux exchange between the vacuum and the system. All EM system are already *open systems* in the virtual particle flux exchange with vacuum; that is already well-known in particle physics, even though the 130-year-old classical electrodynamics does not even include the vacuum-mass exchange. The only question, then, is how to establish an asymmetry in that vacuum-mass flux exchange. Here again, every electrical charge and every dipole is already just such a broken symmetry in the vacuum flux, as has been known in particle physics for four decades! It takes on the average about 50 years, however, for vital discoveries in one discipline to "cross fertilize" into another discipline. At any rate, the symmetry of every electric power system, in the vacuum flux, is already broken in its source dipole. This rigorously means that, if we simply avoid doing something we are doing in our power systems to restore symmetry as a matter of course, it is possible to utilize the free broken symmetry of the source dipole to provide an overunity EM system.

In 1996 I showed that such overunity EM systems are permitted and prescribed by Maxwell's theory, prior to the arbitrary imposition of *symmetrical* regauging of the Maxwell equations (i.e., the Lorentz condition). Violation of the Lorentz condition (by asymmetrical self-regauging) is a necessary but not sufficient condition for an overunity EM device. An asymmetric regauging will certainly give us excess free energy in the system, and an excess force to use that energy in translating against a load impedance to do work, if we design the system properly to do just that.

To achieve overunity, in a circuit one must prevent the (presently universal) re-imposition of the Lorentz condition by the use of one-half the excess received energy to drive the return electrons from the load back through the back emf of the source, killing the gate and restoring symmetry, the Lorentz condition, and equilibrium between vacuum flux and system. All our present EM systems are specifically designed to drive those electrons back through the back emf and thus restore symmetry, *deliberately* asymmetrically self-regauging a second time and *deliberately* restoring the Lorentz condition, along with *deliberately* killing the source dipole's asymmetry in the vacuum flux exchange.

PARADOX (LOGICAL)

A thing that is true, e.g., as shown by experiment, but which violates one or more of the three Aristotelian laws of logic.

Any logic algebra or system has a topology. A paradox in one topology can often be resolved by increasing the topology of the logic. All logical paradoxes in the 3-law Aristotelian logic system are simply fourth-law logic cases in a logical superset, and—at least hypothetically—the fourth law can be applied to resolve them. For a discussion and proof of the 4th law correction to Aristotelian logic, see Appendix III: "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," in T.E. Bearden, **AIDS: Biological Warfare**, Tesla Book Co., Chula Vista, California, 1988, p. 428-443. Logic has many aspects that often may shock our casual expectations: For example, one of the most advanced axioms in mathematics is called the Axiom of Choice. It is used in the proof of many advanced theorems. Using the Axiom of Choice, it can be shown that it is possible to cut a ball into a finite number of pieces, then reassemble the pieces into two balls of the same size, with no air gaps between pieces! While this seemingly defies all common physical sense (and I've not found anyone who has done it!), it is mathematically correct. For a proof, see Thomas J. Jech, **The Axiom of Choice**, American Elsevier, New York, 1973, p. 3-6. For a good perspective on Aristotelian logic, see Morris Kline, **Mathematics: The Loss of Certainty**, Oxford University Press, New York, 1980.

PARITY

Deals with studying a system of sequence of events as if reflected in a mirror. A system has parity is it undergoes no fundamental operational change if replaced with its mirror-reflected twin.

PARTIALLY TIME-REVERSED REGION

A region of space in which an object experiences not only the ordinary forward flow of time, but an extra induced backwards flow of time, usually a fraction of the time-forward time stream.

One result may be to "reduce" time-forward force fields, by a fraction. This is important in bridges, where one must pass EM energy flow onto a follow-on load circuit, but the return current in that circuit must be allowed to circulate back through the load bridge itself, against the inducing back emf. By partially time-reversing the region occupied by the back emf, that back emf is seen as a smaller value by the return current being passed back up through the back emf. The result is that current will circulate in the load loop, powered entirely by the bridging of Poynting flow $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ onto the load circuit.

Another result of a partially time-reversed region is that the law of attraction and repulsion of charges is partially reversed. If the time-reversal predominates in the region, as in a momentary time-reversal zone in some electrolytes, then the law of attraction and repulsion of charges is reversed. In that case like charges attract and unlike charges repel. Two positive H+ ions, e.g., may be drawn so tightly together that each enters the edge of the strong force region of the other, forming a quasi-nucleus. Note that in each H+ ion the gluon forces are almost dissolved and the quarks are nearly freed. Then as the time-reversal zone decays back to a normal time-forward zone, the decay changes begin from inside out, resulting in the flip of a quark to change one of the H+ ions to a neutron n. The n and the H+ ion draw together as the preferred decay state, resulting in the formation of deuterium or a D+ ion. This and similar "time reversal zone" reactions have been proposed by this researcher as the mechanisms generating the anomalous deuterium, tritium, and alpha particles (helium nuclei) in the "cold fusion" reactions. It is indeed a special type of low spatial energy fusion, but a very high energy reaction since time-energy is involved as the active ingredient, and time-energy has the same energy density as mass.

PARTICLE PHYSICS

The branch of physics using accelerators to study high-energy particle collisions, to determine properties of atomic nuclei and of the elementary particles themselves.

So far as I am aware, conventional particle physicists do not consider time-energy interaction and time-reversal zone novel nuclear interactions at all.

PATTERSON EFFECT

The intense multiscattering, multipass, multicollection of Poynting energy density flow in an assemblage of palladium-clad spherical bead capacitors, as the palladium adsorbs hydrogen ions from the fluid in which the beads are immersed, and charges up as a capacitor, thereby becoming an energy-flow generator, and nonlinearly increasing the energy-flow as the charge of the capacitors increases.

The Patterson effect yields the Patterson Power Cell® , a validated overunity power system which has shown COP>1200 in independent university testing. The Patterson effect is a variation and adaptation of the well-known anti-Stokes emission effect, which for three decades has been validated in physics journals to produce overunity emission in certain intensely scattering active media.

PATTERSON POWER CELL®

Patented overunity power unit by James Patterson, which has been independently validated to yield COP>1.0.

See James Patterson, "System for Electrolysis of Liquid Electrolyte," U.S. Patent No. 5,372,688, issued Dec. 13, 1994. Filed Dec. 2, 1993. 26 claims, 7 drawing sheets.

PAUTRIZEL, RAYMOND

Eminent French parasitologist who worked with Antoine Priore, and who performed many definitive experiments demonstrating the Priore healing effect.

PEGRAM'S EXPERIMENT

Experiment which demonstrates a relation between electric and magnetic fields in the vacuum which the Maxwell-Heaviside theory does not prescribe.

Simultaneously, a coaxial cylindrical condensor and solenoid are rotated. The rotation produces a magnetic field in the solenoid in the axial direction and between the plates of the condensor. The condensor can be charged by shortening the plates of the condensor. This indicates a cross relation in the vacuum between electric and magnetic fields which does not exist in the Maxwell-Heaviside theory.

Reference: G. B. Pegram, **Physical Review**, Vol. 10, 1917, p. 591.

PERPETUAL MOTION MACHINE

Assertion that no machine can indefinitely operate, since no closed system can input its output power back to itself while powering a load because powering itself and its load violates the equilibrium thermodynamics (i.e., the thermodynamics of a closed system).

The "defenders of the underunity faith" use the statement that "there cannot be any such thing as a perpetual motion machine" as an article of faith and iron dogma. Usually they classify a so-called "perpetuum mobile" as to which law of *equilibrium thermodynamics* it is thought to violate.

For permissible Maxwellian overunity systems, the entire "perpetual motion" conundrum is nonsensical, since we are always speaking of an EM system *far from thermodynamic equilibrium* with its active vacuum environment. The *equilibrium* thermodynamics assumed by the critics does not even apply to such systems, as witness the Nobel Prize awarded to Prigogine in 1977 and the voluminous literature on such open dissipative systems.

As an example, see Stanley W. Angrist, "Perpetual Motion Machines," **Scientific American**, vol. 218, Jan. 1968, p. 114-122. Angrist gives a typical critique using classical equilibrium thermodynamics. He calls *permissible* overunity systems (open dissipative systems receiving excess energy from an external source) *false perpetual motion machines* or *fictitious perpetual motion machines*. Also see Angrist's article, "Perpetual Motion," **Encyclopaedias Britannica, Bicentennial Edition, Macropaedia** Vol. 14, 1976, p. 102-105. Here he gives a very good summary of the subject as ordinarily posed, with precise definitions included for perpetual motion machines of the first, second, and third kinds and for fictitious perpetual motion machines. Unfortunately his entire argument is completely irrelevant to any study of permissible overunity machines, which he already has labeled as "fictitious perpetual motion machines" and has already admitted that they exist! The substance of his entire argument is that systems in equilibrium cannot power themselves. That is of course quite true, but is a mere tautology and has no relation to an open system in disequilibrium, which *can* power itself and its load simultaneously, by the laws of physics and the laws of *disequilibrium* thermodynamics. So in fact Angrist, as do the skeptics who continue to assail legitimate overunity EM system researchers as perpetual motion nuts, establishes his own dummy strawman and demolishes it, while ignoring—so to speak—the huge Statue of Liberty before him.

Now this is not to pick on Angrist! He appears to be a reasonable scientist, carefully stating his thesis in a straightforward manner. His voice is quoted as a *reasonable* critical voice, and not like the usual *extremist* critic who engages in ad hominem attacks from the outset. It merely shows that (1) the entire "perpetuum mobile" prohibition does not even apply to overunity EM systems far from thermodynamic equilibrium with the active vacuum, (2) the critics are either unwittingly or deliberately using the wrong type of thermodynamics as the basis of their entire argument, (3) the *reasonable* perpetual motion critics—wittingly or unwittingly—simply set up an irrelevant strawman and precede to knock it down, and (4) in fact the critics already admit to a machine that will in fact run itself and power its load indefinitely or perpetually unless something changes or breaks or fails, but then oddly label it as a "fictitious" or "false" perpetual motion machine! Meanwhile, in the universe around us we universally observe things in violent and perpetual motion, having continued so for some 15 billion years. None of the atoms seem to be running down, nor do any of the molecules. And they are open systems brought out of equilibrium by absorption of energy, and decaying back into equilibrium (or approaching it) by emission of excess energy.

The eerie point is that technical publications such as **Scientific American** and **Encyclopaedia Britannica** do not seem to question the line of reasoning advanced by those knocking perpetual motion, or even fully analyzed their argument. What we really deal with in terms of skeptics in this area is either (1) a reasonable, scientific skeptic such as Angrist (his listing of machines and attempts for machines not in disequilibrium with their active environment to run themselves is very good), or (2) an unreasonable person spouting dogma and venom, who should simply go read some disequilibrium system thermodynamics and learn etiquette.

Also, we strongly object to supercilious phrases invariably employed by harsh skeptics who insist we are looking for "something for nothing" or "a free lunch". No legitimate overunity energy researcher has ever considered such a thing. We freely admit that the excess energy is received from the active vacuum exchange, via a broken symmetry in that exchange. Should the farmer refuse to use a windmill or waterwheel because he's getting a "free lunch" for energy input? For those who so quickly hurl such condemnatory phrases, let me hold up a mirror and show them an analog of their position: The famed surgeon Alfred Velpeau wrote in 1839: *"The abolishment of pain in surgery is the chimera. It is absurd to go on seeking it today. 'Knife' and 'pain' are two words in surgery that must forever be associated in the consciousness of the patient. To this compulsory combination we shall have to adjust ourselves."* [Martin Gumpert, **Trail-Blazers of Science**, Funk and Wagnalls, New York, 1936, p. 232.]

The strident, dogmatic critics of open overunity disequilibrium EM systems are stating an exact analogy to Velpeau's statement. The *scientific* worth of their criticism is precisely the scientific worth that eventually was shown for Velpeau's pronouncement.

PHASE CONJUGATE

Of an entity: its time reversal. For waves, often referred to as *wavefront reversal*.

PHASE CONJUGATE MIRROR

A nonlinear material which, when provided with an input wave, emits a phase conjugate replica (time reversed replica) of that input wave.

The input wave is by practice called the *signal wave*. The time-reversed wave output is by practice called the *phase conjugate replica*.

PHASE CONJUGATE OPTICS

That portion of nonlinear optics dealing with multiwave mixing, formation of time-reversed waves, phase conjugate reflection (PCR), phase conjugate mirrors (PCMs), pumping PCMs to amplify their emitted PCRs, nonlinear interferometry, distortion correction, self-targeting, laser tracking, adaptive optics, and so on.

PHASE CONJUGATE REPLICA (PCR)

Of a wave: the time-reversed duplicate of the wave.

PHASE-CONJUGATING ACTION

In nonlinear optics, the process or action of producing a time reversed replica of an input wave.

For a particle, producing its antiparticle. For a cell, dedifferentiating it back to some previous form, genetics and all.

PHASE CONJUGATION

In nonlinear optics, the novel nonlinear mixing of waves which generate an output wave—called the phase conjugate replica or time-reversed replica—that precisely retraces the path previously taken by the input wave that stimulates the action.

The process or action of producing a time-reversed wave. Phase conjugation can compensate for aberrations and distortions in the input. Real time holography can also be produced by applying the phase conjugation process.

PHASE-LOCKED HARMONIC SERIES

In Whittaker decomposition of the scalar potential: The decomposition produces a set of bidirectional wavepairs, as explained previously.

Each of these wavepairs has a given frequency, starting from a fundamental. In other words, the other wavepairs are harmonics of the fundamental wavepair. Further, these are regularly arranged in a lattice, so that "timing" or "phasing" is rigidly locked in place.

PHENOMENOLOGY

In physics, the philosophical and experimental study of physical phenomena,

... as distinguished from ontology, (which is the study of being).

PHOTON

In physics, the basic action quantum.

Considered to be the basic quantum of the EM field. Well, there's a little problem with that. The photon contains energy and time, and so must also occupy a time interval. The notion of forcefield rigorously contains mass as a component; yet mass is missing from the concept of field as it is applied. The EM field is conceived as a spatial entity only. It may spatially vary as a function of time, of course, but it itself does not occupy time, at least conceptually. These considerations raise some startling difficulties in the ready conclusion that the photon is the fundamental quantum of EM field. If a "piece of the photon" is missing from the field, how can the field be made of photons? How can one get area out of a straight line? I simply pose this problem for the theorists and the foundations physicists, pointing out that it has universally been fashionable for physicists to ignore that *the photon carries not only a little increment of energy, but also a little increment of time*. My own work has been to look at the dimensional difference between masstime and mass; i.e., between a mass that has just absorbed a photon, and thus must consist of masstime (4-space), and just after that masstime then re-emits the photon, reverting back to mass (3-space). From this one can build, e.g., a mechanism that generates the actual flow of time itself. It is in theory an engineerable mechanism.

PHYSICAL REALITY

To be added.

PING-PONG

To be added.

PINNA

The outer ear(s). Of particular interest to us are the tiny little folds in the ear.

PINNA INFORMATION

Refers to hidden information content of the field, or in the received signal.

Comes from the little folds in the pinna (outer ear). For sound direction and distance sensing, the tiny pinna folds of the outer ear use phase reflection information more than 40 dB below the primary sound signal that strikes the eardrum. In EM signals, we are referring to "hidden" or infolded information deep within a signal from a distant target. In fact everything in the target contributes to the signal emission from it, and all that information is riding "within" the signal, in its deep "pinna" information. The Fogal chip can outfold some of this pinna information, so that information can be recovered from a signal which no other process heretofore could recover. For a dramatic future use, as the Fogal semiconductors and equipment using them are developed, there will emerge detecting units that can see highly detailed structure in the human body, merely from processing the light reflected from the exterior of the body. The details eventually will rival and even surpass that of magnetic resonance imaging, and will be essentially noninvasive.

It is also true that hidden "vacuum engines" can be transmitted back into the body, so that "problems" or disorders can be directly corrected. As an introduction to this type of area, where cellular disease "antiengines" can be created in the body to directly reverse cells and their genetics, curing the disease, see T.E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy-Medicine, Parts I and II." **Explore**!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62. A major book on this subject will soon be published. For a presentation for the nonspecialist, see T. E. Bearden, "Hidden Mechanism for Reversal of Diseased Cells to a Previous Healthy State," **Resonance**, Newsletter of the Bioelectromagnetics Special Interest Group, Mensa, No. 29, Nov. 1995, p. 11-15.

PINNA INFORMATION CONTENT

In electromagnetics, the hidden inner Stoney-Whittaker-Ziolkowski infolded EM information inside the scalar potential.

Since an EM wave can be represented as two scalar potentials, then by infolding pinna information inside one or both of the two potentials, it is infolded inside the carrier wave. Thus hidden information can be and is carried by all EM radiation and all EM signals. The Fogal semiconductor for the first time enables the outfolding and processing of some of this information.

PINNA INFORMATION DETECTORS

Detectors based on use of the Fogal semiconductor, which can outfold, process, and display some of the hidden information inside the scalar potential and inside signal emissions from distant objects.

PINNED ELECTRONS

Electrons held stationary by a magnetic field, or other set of forces that react to "push them back into their positions" whenever they try to move away.

PINNING

The act of restricting (by restoring forces and fields) movement (particularly of electrons) due to a disturbing magnetic field or other external force.

POLYHEDRON

A solid figure, especially with more than six plane surfaces.

POSITRON

An antielectron (time-reversed electron), possessing a positive charge. In Dirac's relativistic theory of the electron, a positron may be regarded as an empty negative energy state in the Dirac sea whose usual occupying electron has been removed.

POWER

The time rate at which work is done (at which the form of energy is changed).

In electrical engineering, the term "power" is erroneously also used to mean "energy".

PRIGOGINE, ILYA

Noted Belgian chemist and Nobelist, and one of the pioneers of the thermodynamics of open systems far from thermodynamic equilibrium.

Prigogine was awarded the 1977 Nobel Prize for his work on the theory of dissipative structures in nonequilibrium thermodynamics.

PRIORÉ, ANTOINE

Italian radar technician, remaining in France after WW II, who discovered the use of novel waves from plasmas emitted when normal EM waves were introduced, and who used this effect to accomplish remarking healing in laboratory animals of terminal tumors, infectious diseases such as trypanosomas, restore suppressed immune systems, and heal atherosclerosis.

Prioré's beam was actually longitudinal EM waves. The body when irradiated by longitudinal EM waves adds a weaker phase conjugate replica, thus converting LW nonlinear optical pumping from the 3-space domain to the time domain. This time-reverses the irradiated cells and all their components back to a previous physical state and condition.

Such "time-domain" pumping is used by the living body's cellular regeneration system to restore cellular damage and heal the cells, within its limits. Thus Prioré unwittingly uncovered a way to dramatically amplify the body's own healing and regenerative powers.

POYNTING, JOHN HENRY

An English physicist who lived from 1852-1914, whose contributions to electromagnetics included the energy flow theory now known as Poynting flow.

He contributed the Poynting theorem, which states that the rate of EM energy loss in a specific region of space is equal to the sum of the dissipation rate (heat loss) flowing across the boundary of the region. Concurrently with Oliver Heaviside, Poynting conceived the notion that energy flowed through space, which previously had not appeared in physics. Heaviside actually published first, but obscurely; Poynting published in a prestigious journal, so the theory came to be known as "Poynting theory." Poynting himself always gave credit to Heaviside for being first. Also, the Heaviside version $\mathbf{S} = \mathbf{E}\times\mathbf{H} + \mathbf{G}$ was more extensive than Poynting's $\mathbf{S} = \mathbf{E}\times\mathbf{H}$. Heaviside was a recluse; years after his death, in his little garret apartment some papers were discovered beneath the floorboard. Therein Heaviside has used his $\mathbf{G}$ term (a closed energy flow, or "trapped energy" term) to represent gravitation, and had worked out a testable theory of electrogravitation. Modern general relativity indeed considers that it is energy—particularly trapped energy—that is responsible for generating gravitational field and gravitational potential. Apparently nothing at all was done or has been done with Heaviside's electrogravitational theory. See H. J. Josephs, "The Heaviside papers found at Paignton in 1957," **The Institution of Electrical Engineers Monograph No. 319**, Jan. 1959, p. 70-6.

POYNTING ENERGY

Loose term (not rigorous!) used to describe the flow through space of Poynting energy density $\mathbf{S} = \mathbf{E}\times\mathbf{H}$. It actually describes not the energy flow density per se, but that component of it that would be intercepted by a unit point charge.

POYNTING ENERGY DENSITY FLOW

A more rigorous term used to describe the flow through space of the Poynting energy density flow component **S** = **E**×**H**. that would be intercepted by a unit point charge.

POYNTING ENERGY FLOW

Loose term (not completely rigorous) used to describe the flow of energy density **S** = **E**×**H**.

Rigorously, note that both **E** and **H** are "defined" only after their reaction with a unit point static charge. Hence **S** is the resulting component of the overall energy flow that interacts with the charges (with the circuit) and is diverged into the circuit to power it. It is *not* the entire energy flow, which also includes a vast Heaviside nondiverged component that does not strike the charge and interact with it, and hence does not interact with the circuit but is just wasted.

POYNTING GENERATOR

Any dipole or potential difference, which represents a broken symmetry in the vacuum's virtual particle flux, and hence extracts virtual energy from that asymmetry, and gates it out as a gushing energy flow in 3-space. The component of that energy flow that would be intercepted and diverged by a unit point charge is the Poynting energy flow **S** = **E**×**H**.

POYNTING VECTOR

The vector **S**, given as **S** = **E**×**H**, which is used to represent the Poynting (intercepted and collected by a unit point coulomb) energy density flow component.

In Heaviside's version, **S** = **E**×**H** + **G**, where **G** is an energy circulation flow term of trapped energy flow.

PPCM

Pumped phase conjugate mirror.

A phase conjugate mirror material which has introduced to it appositive or "pump" waves. In the phase conjugating process, up to all the energy in the pump waves can be output as the energy in the emitted phase conjugate replica wave. Thus the PPCM is an amplifier, very similar to a triode.

PRIMARY IMAGE

The image or portion of an overall image that is of most interest, and usually the sharpest in focus.

PROBABILISTICALLY

Of or determined by the laws of probability.

Note particularly the somewhat dichotomous use of "determined by—probability." The entire notion in probability is that the overall result is not individually determined, but that there are multiple outcomes of the experiment, each of which occurs with a certain frequency out of a very large number of trials. If a single result always occurs, that is said to be a deterministic change, not a probabilistic change. Unfortunately if one pursues the definition of "probability" in mathematics and logic, one eventually finds that it is dichotomous, as was our definition of energy. Ultimately the foundations guys just threw up their hands and said, "Oh, what the heck! Probability is probability, every fool knows that!" The problem cannot be resolved in 3-la w Aristotelian logic; it requires a higher topology logic.

PROBABILITY

The probability of a simple event is the ratio of the number of times it occurs to the total number of trials (for a large, essentially infinite number of trials).

Note that "trials" constitute things that have occurred (are past). Probability has resisted rigorous logical definition by Aristotelian logic; the reason is that it is an expression of the fourth law, the identity of opposites. Probability is the expression of the future (that which has not occurred) in terms of the past (that which has occurred). Since we have been trained to conceive or think of something as if it had just been perceived to occur, then this is the only way we can *conceive* the future—as if it were an "already occurred past" to us.

But with Aristotelian logic, the future cannot be composed of a set of pasts (the law of the excluded middle prevents it!), and probability has no acceptable logical basis. With four-law logic, the future can be modeled in terms of—and even be identical to—the past, and probability has a logical basis.

Consider *throwing a die to land with one face up*. Thinking of the event (as if it has just occurred, in terms of the most recent past), one can conceive six ways of looking at it. In other words, one can conceive of six "most recent pasts." By the fourth law, the total "most recent past" is identical to the "most immediate future," on the common boundary. If one therefore *collects* all six "most recent pasts," they turn into the most immediate future, by the fourth law of logic. One may argue that only a single one of the events will actually occur. Here one counters with the observation that, when it occurs, it is the past, not the future, and then it has not been collected so as to move to the boundary. One also points out that, with the hidden variable approach, in theory it is possible to directly engineer those probabilities while the Schrödinger wavefunction is still propagating and has not yet collapse to constitute an "observation." I.e., it is possible to engineer Wheeler's "pre-reality," and determine in advance just which of the outcomes shall occur.

Now, if this type of equipment eventually gets built and is small and portable, just think what one could do in Las Vegas!

PRODUCT SET OF INTERNAL WAVES

Modulations of the internal bidirectional wavepairs comprising a scalar potential.

This set was added in the mid-1980s by Ziolkowski, to Whittaker's 1903 sum set. When a signal is "infolded" inside the DC potential, it is modulated upon (multiples) one or more of the internal biwave pairs. Ziolkowski's work is necessary to deal with that; Stoney and Whittaker alone do not handle it.

PSYCHOENERGETICS

Russian branch of energetics where extended electrodynamics is targeted upon the mind and its functions, hence in the time-domain and employing time-polarized EM waves.

PSYCHOTRONICS

Czechoslovakian name for extended electrodynamics effects on the living body.

PUMPED PHASE CONJUGATE MIRROR (PPCM)

A nonlinear material which:

(i) when provided with an input wave, emits an amplified phase conjugate replica (time reversed replica) of that wave, and

(ii) has additional opposing waves as inputs, so that their energy is transformed into the amplified energy of the time-reversed replica wave.

The word "mirror" is misleading, in that the "mirror" may be just a highly nonlinear material, or be an actual smooth metallic surface containing "hot spots" by heating, etc. In any case, it is the nonlinearities such as the hot spots that produce the phase conjugating action.

PUTHOFF, HAROLD

Noted physicist active in a variety of fields and investigations including in psychoenergetics, extended electrodynamics, energy from the vacuum, parapsychology, remote viewing, psychokinesis, zero-point energy of the vacuum, and cosmology.

We have specifically mentioned his important cosmological feedback principle.

QM VACUUM

The quantum mechanical vacuum, which models the vacuum as an intense flux of virtual particles, each appearing an disappearing in such extreme time intervals as to be unobservable individually.

Violent fluctuations of extremely short duration constantly occur in the QM vacuum flux. Various calculations estimate the energy density (in mass units) as up to 10^{80}-10^{90} grams per cubic centimeter, or even greater.

QUALITATIVELY

Determination based on qualitative analysis, usually as contrasted to quantitative determinations.

QUANTIZED

i) Reduction of some observable quantity to multiples of some small, indivisible unit, or,

ii) expressed in terms of quantum theory.

QUANTUM

A single "particle" of the quantity action.

The quantum is always composed of two fundamental quantities (canonical variables), and it is the "fissioning" of the action quantum into two pieces which results in detection of one of the pieces. An indefinite "smear" of part of the two pieces may be simultaneously detected by a linear detector, but never both exactly (Heisenberg uncertainty principle). More recent work has shown that, in a highly nonlinear situation, both canonical variables can be determined simultaneously to any desired degree of accuracy, in which case the Heisenberg uncertainty principle does not hold in its conventional form. This fundamental change in what has been regarded as one of the "sacred laws" of physics has not yet made it into most physics texts.

QUANTUM ELECTRODYNAMICS

The theory of photons and electrically charged particles and their interactions.

The use of the term "quantum" implies that the EM radiation's discrete photon nature is significant, so that quantum theory must be employed.

QUANTUM FIELD THEORY

A quantum mechanical theory in which "... a physical field is considered as a collection of particles and forces, and observable properties of an interacting system are expressed as finite quantities rather than state vectors."

[**Dictionary of Science and Technology**, ibid.] More simply (after Davies, **The New Physics**, 1989): "The theory that describes the quantum effects of a classical system of fields defined on space-time and satisfying various partial differential equations."

QUANTUM POTENTIAL

A special potential added to the Schrödinger equation, by Bohm in his hidden variable theory of quantum mechanics.

The quantum potential can move instantaneously. In my book **Gravitobiology**, I published a mechanism by means of which a quantum potential can be created. (There may of course be *other* QP-creating mechanisms). An iterative, mutually phase conjugating of EM signals passed back and forth between two objects, or among many of them, is called *self-targeting*. Extreme self-targeting can produce such a QP or partially produce one, particularly if a dense "pumping" (optical sense) of the phase conjugate mirror objects occurs. I am presently proposing the QP as a means of direct energy amplification. I.e., if there is "instantaneous travel" between two light-observed points, then to the instantaneous agent there is no length separation or time separation between the points. This of course means that in effect the objects now coexist "superposed" on each other in a multiply connected space, at least with some fraction of "coupling".

Suppose the coupling fraction is 0.5, and 100 widely separate d nodes are so connected by a common QP. Then if one puts in 1 joule per second of energy flow at a given node, there will instantly and simultaneously appear at each of the other 99 nodes one-half joule of energy per second flow. So I can gain quite a bit o f *direct energy amplification* in this distributed system, and nothing can "shield" the distant nodes. Presently, other than possibly Gabriel Kron's "open path", I have not found any other consideration of "direct amplification of energy at a distance".

The implications of the quantum potential—in everything from power sources to medicine to weaponry—are staggering. Weapons based on just this very QP, multiply connected space, and direct energy amplification appear to have already been developed and are on site in three nations. A fourth either is just getting them or is nearing their deployment.

QUATERNION

Expression comprised of the sum of four terms, one of which is real and three of which contain imaginary units, and that can be written as the sum of a scalar and a three-dimensional vector.

QUATERNION ALGEBRA

The algebra of quaternions and their mathematical operations.

Quaternion algebra is of higher topology than either vector algebra or tensor algebra. Maxwell's original equations are some 20 equations in 20 unknowns, in quaternion algebra. Heaviside and others reduced the algebra to vector algebra and some four equations. The present vector equations taught in university as "Maxwell's equations" are in fact Heaviside's truncation of the Maxwell theory. Going to tensor algebra electrodynamics cannot recover the full range of Maxwell's quaternion theory.

As an example, what Tesla actually did in his patented circuits cannot even be seen in vector or tensor analysis. On the other hand, analysis of his circuits by quaternion electrodynamics clearly reveals his startling ability to shuttle the energy in the circuit at will. See T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," **Annales de la Fondation Louis de Broglie**, 16(1), 1991, p. 23-41.

QUINTESSENCE

Literally, the "fifth essence". In short, dark energy.

RADIONICS

To be added.

REACTIVE POWER

Electrical power, measured in volts-amps-reactive (vars) that cannot do work as is.

For sinusoidal waveforms, the formula for reactive power is given by $I \cdot V \cdot \sin\theta$, where θ is the phase angle between the voltage and the current.

REAL, HIDDEN DYNAMICS

The dynamics of vacuum engines, or of "internal energy" or "internal work" ongoing in a system or volume of space.

REAL HIDDEN VECTORS

The "hidden" or "infolded" real vector components,

e.g., of a vector zero summation system.

A group of nonzero, finite vectors which as a system sum to a net translation zero; i.e., which do not translate a resisting mass or charge, but do "stress" it. Such a system therefore constitutes a stress potential.

REAL POWER (ACTIVE POWER)

Power, measured in watts, that can do work, including generating heat.

For sinusoidal waves, real power is given by the formula $I \cdot V \cdot \cos\theta$, where θ is the phase angle between the voltage and the current.

REAL VECTORS

Vectors in three space, without imaginary components.

REGAUGE, REGAUGING

Changing the scalar potential or the vector potential, or both.

If one or both of the potentials is/are changed so that a net force is created in the system, that is an asymmetrical regauging.

If only one potential is changed, a net force results and that is an asymmetrical regauging.

If both potentials are changed but carefully selected so that the free force resulting from one change is equal and opposite to the free force resulting from the other, that is a net symmetrical regauging, often called a "Lorentz regauging".

RECTANGULAR TRANSFORMATION MATRIX

Standard mathematical physics.

REFLECTION

The rebound of light or other energy from a surface.

A smooth surface (e.g., a mirror) may produce a specular reflection. A rough surface (e.g., a sheet of paper) produces a diffuse reflection. Actually, in reflection of light, the same light does not rebound; instead, the light incident upon the mirror or surface is absorbed, and new light is re-emitted by that surface. Further, considering reflection from a dielectric, every interior part of the dielectric participates in the reflection. All the internal information about the entire object—everything!—is there in that reflected light, infolded as the hidden "information content of the field" (Stoney-Whittaker-Ziolkowski structuring inside the two scalar potentials comprising the reflected wave). With the Fogal semiconductor, at least some of that hidden "pinna" information in a sample detection of the reflected light can be outfolded and displayed.

REFRACTION

The bending of light as it passes from one medium to another.

It is the phenomenon that enables a lens to focus light and that makes a stick appear to bend where it dips into water.

REPULSION BEAM

A beam of energy directed at a target, which when absorbed by the target will generate a force in and on that target to push it away from the beam projector.

REPULSION-FORCE-GENERATING BEAMS

Beams of energy which "push away" their targeted receivers, by inducing a force in the receiver directed away from the source of the beam.

RETARDED WAVE

The electromagnetic wave that is derived from the retarded fields, which in turn are derived from the retarded potentials.

This means that the retarded wave at a point or region of space has arrived there or exists there as the result of the dynamics of charges and currents that existed at earlier times.

RETROREFLECTING

Precisely reversing the path of.

An example is a phase conjugate reflection.

SACHS, MENDEL

Noted scientist who has completed the work of Einstein and developed a unified field theory covering the universe from the tiniest part to the entire conglomerate.

SACHS' UNIFIED FIELD THEORY

An extension and completion of Einstein's work that includes the universe and its dynamics from the tiniest part to the entire conglomerate, and includes quantum mechanics, electromagnetics, gravitation, and the strong and weak forces.

SCALAR

Characterized by magnitude only.

SCALAR (QUANTITY)

In ordinary vector analysis, a quantity completely characterized by magnitude only.

Actually, it contains magnitude and time, because to exist it must steadily exist in time. So we take the smallest increment of observer time we wish to consider, and that is the minimum "length" of the time-tail that the scalar must occupy and have, i n order to even be a "perceived or detected" thing. Before the mathematician jumps in with a nice limiting process to take away the time-tail, if you let that little time interval get too close to zero, then you get too close to eliminating the observer himself, and the entire problem vanishes including mathematics and physics and the observer as well.

One simply cannot demonstrate something that does not exist in time—on paper, in one's mind, or otherwise! Be careful of the term "spacelike " in physics! It hides a logical nightmare. This was the very problem I uncovered in Aristotelian logic; the little increments of time to have a "perceived A" or a "perceived not-A" or to apply an "identity decision criteria" to determine identity or nonidentity, had been left out for lo these 2,000 or so years.

SCALAR ELECTROMAGNETICS

That electrodynamics that arises from considering transverse EM waves, longitudinal EM waves, time-polarized EM waves, electrogravitation, superluminal EM signals, interferometry, nonlinear optical functions, time-as-energy, and the infolded electrodynamics inside all usual EM fields, waves, and potentials.

SCALAR POTENTIAL

Usually considered synonymous with the electric potential.

Here we have another "logical nightmare."

Anyway, the conventional "definition" of electric potential assumes one knows what a "potential" is already, and goes something like this: "The potential measured by the energy of a unit positive charge at a point, expressed relative to an equipotential surface, generally the surface of the earth, that has zero potential." [**Dictionary of Science and Technology**, ibid., p. 722.]. Well, that is botched. It tells or attempts to tell what the *measurement* of a potential is, using a unit positive charge and a process. A potential cannot be its own magnitude measurement, just as a man cannot be his own weight measurement.

So what do we do? We have to tell what the potential *is*. This is the problem that so mystified and infuriated Oliver Heaviside, when he altered and diminished Maxwell's equations to rid them of every potential that he could. Heaviside stated that the potential was *"mystical and ought to be murdered from the theory."* He tried to do just that, and conditioned generations of electrodynamicists to believe that the potential had no physical reality at all, but was only a mathematical convenience. Nahin expresses Heaviside's view as follows: "In an 1893 letter to Oliver Lodge, Heaviside said of his own work that it represented the 'real and true "Maxwell" as Maxwell would have done it if he had not been humbugged by his vector and scalar potentials.'" Paul Nahin, **Oliver Heaviside: Sage in Solitude**, IEEE Press, New York, 1988, p. 134, n. 37.

The concept of potential was developed by Poisson, Green, and others from about 1813 to 1827. Three notions involved in the development of the concept of the potential are: (i) there is the collection or storage of energy in the system, via the collect ion or storage of some entity therein, and this "potential energy" can be released and dynamically used, (ii) the potential is a scalar function whose spatial rate of change is a vector force, and (iii) the dynamics of the stored energy is suspended by the "storage" or "collecting" entity; in other words it is a sort of "suspended action" waiting to be unleashed. *It requires some action—such as interception and divergence—to release the leashed energy*.

We have to disagree with the second notion, having already discussed the fact that no vector force exists unless mass is present, since mass is a component of force. Here we have a residue of the old "material ether" concept. For the third notion, we note that the concept of a "virtual" particle or particle flux means that it cannot become observable unless some interaction is imposed. So we may handle the "third notion" requirement by assuming the potential to be composed of hidden virtual particle flux, or hidden wave flux, in wave-to-wave interaction.

Physics basically tries to define the electrical potential as energy per coulomb, or $\phi = W/Q$. That equation, however, is actually not a definition at all, but rather is an algorithm for calculating the magnitude of the excess energy collection/collecting upon a coulomb of charge, when that coulomb is immersed in the potential ϕ. It in fact is the *reaction cross section* of the potential, not the potential entity itself. This now begins to give us a clue. Potential is *energy* in some fashion. So the entity ϕ must have essentially the same definition as energy! Potential energy, collected energy, stored energy, whatever it's energy first and foremost.

Refer to our discussion of the 3-law logical contradiction in the concept of energy. The same logical contradiction must be present in the notion of the potential.

Just as we did for energy, we must define the scalar potential ϕ in two appositive ways: (i) in the wave view and (ii) in the particle view. So we define potential (particle view) as any virtual particle flux. We also define potential (wave view) as any hidden wave flux. Further, we note that the energetic vacuum is composed of a virtual particle flux (particle view) and so it is also a *potential*. Since the virtual particle flux of the vacuum is extremely intense, the vacuum potential is enormous in magnitude.

We shall also consider any other potential to ultimately be a change to the vacuum potential. Originally, one conceived of a potential occupying the "vacuum as an empty space." Then general relativity made the vacuum into spacetime, geometrized it, and made the geometry itself dynamic and operative—i.e., energetic. Okay! Then that makes spacetime itself a potential.

We also firmly dispose of the peculiar and deformed notion that the scalar potential of a point-source charge (at any spatial point that the potential occupies) may be *defined* as the "work performed upon a unit charge by pushing it in against the **E**-field of that source charge, to that spatial point." This is an equation and not a definition at all; instead, it is an algorithm for calculating the energy density magnitude of the scalar potential. It is a fine algorithm, but it has nothing at all to do with what the potential itself *is*. Again, a man cannot be defined as an algorithm for calculating the magnitude of his height or his weight or his energy density!

Without further discussion, we also somewhat extend the general relativity notion of spacetime as dynamic geometry, as follows: In a special sense, spacetime ≡ geometry ≡ potential ≡ "trapped " or collected/collecting energy. So when we engineer the internal structure of the potential, we also engineer the internal structure (subspace) of spacetime and spacetime geometry. Curvature of spacetime, e.g., becomes simply a change in the energy density of the vacuum, from the ambient value, and it can be internally structured so as to contain "engines". When we apply the hidden EM variable theory of the Stoney-Whittaker-Ziolkowski work, particularly when we consider the phase conjugate pairing as it exists in spacetime before interaction with charge, we are engineering and extending (i) electromagnetics, (ii) quantum mechanics, (iii) general relativity, (iv) spacetime, (v) the spacetime curvature concept, and (vi) spacetime geometry.

SCATTERING PROCESSES

Collision processes in which an incident particle such as a photon, electron, or proton, of known energy or direction or both, collides with another entity, and emerges from the collision with a different energy or direction or both.

SCHRÖDINGER, ERWIN

Noted Austrian physicist and Nobelist who discovered the wave equation named after him as the Schrödinger equation, and performed other vital work in wave mechanics.

SCHRÖDINGER EQUATION

In quantum mechanics: an equation that describes the propagation of waves associated with subatomic particles, or more generally, that describes the evolution over time of the quantum state of a system.

SCHUMANN RESONANCES

The extremely low frequency (ELF) resonant frequencies of the Earth-Ionospheric waveguide.

The first five Schumann resonances calculated from a simple model are 10.6 Hz, 18.3 Hz, 25.8 Hz, 33.4 Hz, and 40.9 Hz. However, better models are available and the frequencies vary from these simple model solutions. A better calculation gives the first four as 8, 14, 20, and 26 Hz nominally. The measured Schumann resonance corresponding to the 8 Hz is about 7.75 Hz. This is very near the mechanical resonance frequency of the human body, and also is evidenced in brain waves often "locking" to this frequency in the natural environment.

See W. O. Schumann, **Z. Naturforschung**, Vol. 72, 1952, p. 149, 250. For detailed models and discussion of the observations, see J. Galejs, **J. Res. Nat. Bur. Standards**, Vol. 69D, 1965, p. 1043.

SELF-ENERGY

In a quantum mechanical system, the energy associated with the emission and absorption of virtual particles.

In a classical system, the energy associated with the interaction among parts of the system itself.

SELF-ORGANIZATION

(After Davies, **The New Physics**, 1989): *"Spontaneous emergence of order, arising when certain parameters built in a system reach critical values."*

Simpler put, it is the spontaneous formation of patterns of localized order in a system which begins in a more homogeneous state. In short, the system moves from a simpler to a more complex state.

SELF-POTENTIAL (E.G., OF THE ELECTRON)

The potential created in and of the charge

...(e.g., the electron) by its virtual particle flux exchange with the vacuum.

Also, an asymmetry in the virtual photon flux of the vacuum, caused by the interaction of the charge (e.g., the electron) with the vacuum flux. From here, we note that the self-potential must consist of real observable 3-space EM energy flow or flows from the electron, because of the asymmetry. However, the charge is not the primary source *per se* of the energy flows that establish the self-potential, but instead is the asymmetric gate in the vacuum flux that "gates out" the potential and its energy flux. The asymmetry in the vacuum flux is the actual *source*. Indeed, the energy is received by the charge in the form of converging longitudinal EM energy waves from the complex plane (from the time domain). The spin of the charges of the dipole transduce the absorbed complex EM energy into real observable EM longitudinal wave energy, and pour it out in all directions in 3-space. See my paper, "Giant Negentropy from the Common Dipole," *ibid*. for the mechanism for the charge also.

SELF, SENSE OF

To be added.

SELF-TARGETING

Between two entities, where each acts partially or wholly as a phase conjugate mirror, self-targeting is the iterative phase conjugation or "ping-ponging" and rapid convergence of beams between the two PCMs.

It has a spectacular application in *phase conjugate shooting* from a spaceborne laser against a rising hostile booster, even several thousand miles distant. First a laser pulse is "bounced off" the distant target, and the reflected pulse returns to the receiver in the laser platform. It carries the movement-induced alteration of the distant target on the pulse it reflected. The next emitted pulse is an amplified phase conjugate replica of that received pulse, and it will backtrack to the same spot on the rising booster, leading it a bit and converging moderately so that it meets that spot. That spot begins to heat up, increasing its nonlinearity and making it act even more as a PCM. So its reflected signal off the "hot spot" is also a phase conjugate replica of the pulse that it received from the laser. That reflected pulse now will moderately converge its reflected energy upon the distant laser, increasing the return to the laser. The ping-pong continues, so that the laser beam narrows sharply upon the distant hot spot, and the reflected beam from the hotspot narrows sharply upon the distant laser. This "iterative narrowing ping pong" locks the laser upon the distant hotspot, even though the booster is moving. This process continues for the necessary dwell time, until the beam melts through that spot and explodes the hostile booster. The iterative phase conjugating reflection, or "ping ponging" between the target spot and the laser, is called *self-targeting*.

S-FLOW

The Poynting energy density flow S, where **S ≡ E×H + G**.

E is the electric field, **H** is the magnetic field strength, and **G** is a trapped closed loop flow of energy density.

The **S**-flow is not the total energy flow per se. Since the fields **E** and **H** and the closed loop **G** are only defined in terms of the intercepted, trapped or "collected" energy, then **S**-flow is always *with respect to* some assumed intercepting/collecting entity, usually an assumed unit point charge.

SIGNAL WAVE

In nonlinear optics, the input signal (wave) to a phase conjugate mirror (PCM), which stimulates the PCM to emit a phase conjugate replica (PCR) wave in return.

SLEPIAN CURRENT

The vector jϕ , which in a current loop represents the energy unit area per coulomb of flowing charge, per unit of time, that is collected/collecting from the Poynting **S**-flow and dissipated in that current loop in its load s and losses.

It does not include the enormous amount of Heaviside nondiverged energy density flow that is flowing along the outside of the circuit conductors, and that is not being collected and therefore not being dissipated.

SLEPIAN FLOW

"Flow of collected-energy density dissipation" in a current loop,

...and therefore to the vector jf in the current loop, which represents the energy per unit area per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

It does not represent the enormous amount of energy that is flowing along the circuit, that is not being collected, not diverged, and therefore does not enter the circuit to be dissipated from it.

SLEPIAN VECTOR

Refers to the vector $j\phi$ in a current loop,

...which represents the energy density per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

It does not represent the enormous amount of energy that is flowing along the circuit, but not being collected and therefore not being dissipated. We accent that *only intercepted and collected energy can be dissipated!*

SOLID-STATE PARAMETRIC OSCILLATORS

A parametric oscillator circuit employing solid state devices.

A parametric oscillator is a device that oscillates by rhythmically changing a parameter (such as capacitance or inductance, etc.).

Simple switching between differing values of the parameter is often used to induced the periodic change of the parameter.

SOLITON

A wave pulse which propagates without changing its shape and without dispersion.

SOURCE DIPOLE

Any dipole, considered from the viewpoint that EM energy continuously pours out of it.

There is no true "source" as such in nature, since the energy is not created but only gated. In the case of the dipole and electrical charge as sources of EM energy, fields, and potentials, we solved that vexing problem in our "Giant Negentropy from the Common Dipole", *ibid*.

SPACETIME

The entity represented by the "fused product" of space and time, to give "space × time".

Spacetime is also dynamic geometry, in the general relativistic interpretation. We consider spacetime to be a potential, and also geometry, and also vacuum, and also virtual particle flux, and also hidden wave flux.

SPATIAL CLOUD

A cloud of something (such as electrons) in a spatial region.

SPATIALLY

With reference to space and spatial position.

SPECIAL RELATIVITY

Einstein's theory of inertial frames in which, if two systems are moving uniformly in relation to each other, one cannot determine anything about their motion except that it is relative.

Each of the two frames is said to be "rotated" with respect to the other, but not accelerating. The velocity of light in space (the vacuum) is constant, and is independent of the velocity of its source and the velocity of an observer. All the laws of physics are the same in all inertial frames of reference.

SPIN

The intrinsic angular momentum of a particle, such as an electron, proton, neutron, photon, graviton etc.,

...even when at rest, as if it were a top spinning about an axis, but had to spin 720 degrees before it turned "full circle."

Spin is quantized, and is always described as a half or whole spin, e.g., - 1, - 1/2, 0, 1/2, 1, etc. A spinning charged particle such as an electron thus demonstrates a magnetic moment, due to the circulation of charge in the spinning. In the nucleus of an atom, the spin of the nucleus is the resultant of the spins of the nucleons (particles comprising the nucleus).

Spin of particles would appear to be more like an "implosion" to "explosion" circulation; in other words, it circulates in the time domain (complex plane) as well. It would appear that the spin of a particle is the basic feature that integrates the (disintegrated) flux energy of vacuum flux into observable charge. Apparently all observable fields, matter, effects, etc. depend upon this basic mechanism to zip together virtual entities and form observable phenomena.

SPIRAL GALAXY

Those galaxies having arms which extend in the form of spirals.

About 70% of all galaxies are spiral galaxies.

The interesting thing is that only about 10% of the gravity necessary to hold the spiral arms together can be accounted for by known gravitational sources. This has led to the theorizing of dark (hidden) matter of exotic and unknown form, which would be responsible for the excess gravity.

The present author has proposed that the unaccounted enormous Heaviside component of the EM energy flow surrounding every field/charge interaction and not entering into that interaction, accounts for the missing source of the gravity. See "Dark Matter or Dark Energy?", Journal of New Energy, 4(4), Spring 2000, p. 4-11.

S-SOURCE

Source of Poynting energy density flow.

Source of **S** = **E**×**H** flow. The primary source is considered to always be an asymmetry in the virtual particle flux of the vacuum. The **S**-flow rigorously is that component of the overall energy flow that is intercepted by some system of interest—either present or assumed. That component of the overall energy flow that is not intercepted by the system of interest, is referred to by this author as the Heaviside energy flow or the dark energy flow associated with every field/charge interaction.

STONEY To be added

To be added.

STONEY/WHITTAKER BIWAVE DECOMPOSITION OF THE SCALAR POTENTIAL ACROSS THE DIPOLE

(i) a scalar potential (i.e., a difference in potential) exists between the end charges of a dipole,

(ii) that potential (difference) can be mathematically decomposed into a harmonic series of bidirectional EM wave pairs, where each pair consists of an EM wave and its precisely superposed time-reversed twin. Thus a dipole is a "source" or identically is a set of hidden bidirectional waves (wave view).

STRING THEORY

A rather recent mathematical physics theory which replaces the zero-dimensional "point" used in previous physics with a one-dimensional object called a "string".

It has been particularly effective in describing the gluon forces between two colored particles in particle physics, and has since been extended to many other areas. It is sometimes referred to as the "theory of everything".

"STRING" WAVE

The transverse wave on a taut string produced when the string is plucked.

The string wave itself stays on the string and does not travel through the slapped external medium.

However, Faraday believed that EM fields in vacuum consisted of "material lines of force" which behaved like such taut material strings, without any physical endholders. He thought that EM disturbances were simply the transverse disturbances of those "lines of force". He profoundly affected Maxwell's thinking, so that Maxwell accepted these physical lines of force. Consequently Maxwell *assumed* those same "transverse taut string vibration waves" in his EM wave theory.

In eliminating the "stringholders" for the assumed taut strings, Faraday and Maxwell discarded the Newtonian third law half of electrodynamics! They discarded half the energy and half the wave as well. Consequently today physics has no causative mechanism for Newton's third law, because it threw out the EM mechanism for it more than a century ago and never noticed it!

SU(2)×SU(2) SYMMETRY ELECTRODYNAMICS

To be added.

SUBSPACE

In vector mathematics: a subset of a vector space which is closed under vector addition and scalar multiplication operations.

In physics: Physics often uses the term "space" to refer to spacetime. It also uses the term "space" to refer to the dimensions of space, without the dimensions of time. At any rate, the spatial portion of spacetime is represented mathematically by some convenient number of dimensions. In the most usual practice, Minkowski spacetime of 3 spatial dimensions and one time dimension is utilized. In the Minkowski space, we now consider that "space" to be a vacuum scalar potential. If we consider the internal Stoney-Whittaker-Ziolkowski biwave pairs comprising that vacuum potential, we may transmit a signal communication along one or more of these internal biwaves, without changing any of the others. In that case, no overall potential gradient and overall forcefield signal appears. That domain that is infolded "inside" the scalar potential is said to be subspace. It may be modeled as a space of n-dimensions, where those dimensions are compacted and not observable to an external observer. Subspace gained great popularity from the Star Trek television series, where the spaceships used "subspace communications" to have signals propagate far faster than the speed of light. Bill Fogal has been working with some success on what should be a concrete realization of just this very "subspace communication system", with infolded signals that move much faster than the speed of light in vacuum.

SUPERCONDUCTIVITY

Property of some materials or metals whereby their resistance reduces to zero at very low temperatures.

These materials also exhibit many additional properties that are anomalous when compared to normal materials. Much research effort has gone into the attempt to find materials which exhibit superconductivity at higher temperatures, with the goal of reaching room temperature superconductivity.

SUPERLUMINAL

In a manner or with a speed faster than c, the speed of light in vacuum.

SUPERLUMINAL COMMUNICATION

The movement of signals faster than the speed of light in vacuum,

...and the transmission and reception of such signals and extraction of their transported information.

Longitudinal EM waves are not limited to speed c; transverse EM waves are limited to speed c. The infolded electrodynamics inside all EM fields, potentials, and waves consists of longitudinal EM phase conjugate wavepairs, with their impressed electrodynamics. Hence this "subspace" is a superhighway for such waves, including for those moving at superluminal velocity. In essence this is a very special kind of "tunneling" which we might refer to as *subspace tunneling*. The limitation to light speed rigorously applies to the modeled transverse EM waves which are a sort of "bulk upheaval" waves similar to waves on the surface of the ocean. As in the ocean wave analogy, "pressure" waves underneath the ocean may travel very much faster than the bulk upheaval waves on the surface.

When making longitudinal EM waves, the process is imperfect and the resulting longitudinal EM waves will have a transverse residue as well. The resulting "transverse-noisy" LW is called an *undistorted progressive wave*. Undistorted progressive waves may move slower than the speed of light or faster than the speed of light. See W. A. Rodrigues Jr. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature," **Foundations of Physics**, 27(3), 1997, p. 435-508. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site, and includes corrections to the published version.

SUPERPOSE

To lay or place on, over, or above something else, with the connotation of mixing.

SUPERPOSITION

The simple linear addition and subtraction of two or more values, states, etc.

Superposition is one of the key principles in field theories and in the concept of potentials.

SUPERSET

A large set referenced to a small set, where the small set is included in the larger set.

The large set is said to be a *superset* of the small set. The small set is said to be a *subset* of the larger set.

SW (STONEY-WHITTAKER) STRUCTURES

The Stoney-Whittaker hidden bidirectional EM wavepairs comprising the scalar potential, and structuring (altering) of same.

We call this internal EM dynamics inside every EM wave, field, and potential the "infolded" electrodynamics. Russian weapon scientists refer to it as the *information content of the field*.

SYMMETRICAL REGAUGING

In electromagnetics, changing the scalar potential to a different value, and simultaneously changing the vector potential to a different value, with the two changes just precisely selected so that the two extra forces (each created by one of the changes) are equal and opposite—thus canceling with respect to translation and doing no net external work.

We note that the two forces themselves must be considered to be present and still working continuously in opposition, increasing the stress upon the system regauged with excess energy, or decreasing the stress upon the system regauged with less energy.

This presence of "equal and opposite" excess forces from the symmetrical regauging is known as the *Lorentz condition*. The end result is that the *net* force fields of the regauged system remain the same, even though the stress of the system has been altered by the regauging and the potential energy of the system has also been changed. Since abstract vector mathematics contains no such thing as "stress" in its background vector space, zero vector resultant system of multiple nonzero vectors are most often simply discarded. Early electrodynamicists symmetrically regauged Heaviside's two "Maxwell" equations (potential form) so as to easily separate the variables. The net effect was to discard that set of the Heaviside/Maxwell equations that prescribed and included open electrodynamic systems that freely receive and collect energy from the external vacuum, and also obtain a free excess force which then can permissibly utilize that excess energy to translate and produce external work upon the system or externally upon a load. In short, the symmetrical regauging discarded practical overunity electrodynamic circuits and devices, which Maxwell's equations—even in their Heaviside modified form—have always prescribed and permitted. Contrast to asymmetrical regauging.

We specifically point out that the two extra forces remain, as does their energy, and both may be dynamic as the system is dynamic. Hence when one regauges, one has changed the gravitational aspects of the system and also curved the local spacetime. Further, this ST curvature is internally structured and that structure is dynamic. Thus contrary to the assumptions of electrodynamics, the regauged system *is not* the same as the original system.

SYMMETRY

A theory or process possesses a symmetry with respect to certain operations performed on it, if those applied operations do not change the theory or process.

The theory or process is then said to be "symmetric" with respect to those operations. As an example, there is no observable change in a circle when it is rotated or reflected in a mirror. The circle is therefore said to have rotational symmetry and reflection symmetry.

TANTALUM CAPACITOR

Electrolytic capacitor with a tantalum or sintered-slug anode,

...such as solid tantalum, tantalum-foil electrolytic, and tantalum-slug capacitors.

For the same capacitance, the weight and volume of the tantalum capacitor are less than comparable aluminum electrolytic capacitors. In other words, you can make the little son-of-a-gun smaller, for a given capacitance and charge it must hold. When utilized in the Fogal fashion, the tantalum capacitor also exhibits other properties that are still proprietary to Fogal's semiconductor processes and to Bill Fogal.

TARGETED PINPOINTS

Refers to those separated points which engage in self-targeting between themselves and some separated signal energy source (as in Fogal's semiconductor).

The effect is to sharply narrow the retroreflected source signal energy from the source onto the separated points by means of the mutual ping-ponging. In this fashion much more energy density from the source can be concentrated on the targeted pinpoints. In effect, phase conjugate shooting is evoked.

TENSOR

A special mathematical entity which is an element of a vector space such that it represents the tensor product of vector spaces and their duals.

Much of modern physics and advanced electromagnetics is expressed in tensor algebra theory.

TESLA, NIKOLA

Famous American emigré electrical inventor and eccentric whose numerous inventions fueled much of the electrical 20th century.

Tesla produced the AC system, including polyphase systems. From his experiments he knew that the EM wave in vacuum was longitudinal, like sound waves. In spite of his great accomplishments, his eccentricities and grandiose pronouncements did not endear him to electrodynamicists! An example is his acrid comment: *"...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light."* Another major example is his statement: *"The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

Ironically, modern quantum mechanics has almost proven Tesla right that "the universal medium is a gaseous body." Today the modern QM vacuum is very much like a virtual particle gas. And of course the discussions in this present annotated glossary show why I personally am of the opinion that Tesla was also correct in his firm disbelief that those Faraday-Maxwell "taut physical strings without physical holders" filled the ether and provided transverse EM wave oscillations. I also agree with him that the EM wave in vacuum is a longitudinal wave, exactly like a sound wave. I believe that the transverse gyroelectron precession waves that our instruments actually measure in the conductor when it receives an incoming EM wave from the vacuum, proves that the vacuum wave is indeed longitudinal—else it would seem we must discard the spin of an electron and the theory of gyroscopes. Finally, one should see, e.g., Patrick Cornille, "Inhomogeneous Waves and Maxwell's Equations." Chapter 4 in **Essays on the Formal Aspects of Electromagnetic Theory**, Aklesh Lakhtakia, Ed., World Scientific, New York, 1993. Cornille clearly shows that Maxwell's equations are directly created by scalar potential interferometry. As he states, this means that EM force field wave s are created by the interference of sound-type scalar potential waves, or "sound creates light." This strongly supports Tesla's original statement that EM waves in the vacuum are actually EM sound waves, much like sound waves i n a gas.

TESLA'S MAGNIFYING TRANSMITTER (include Jackson quote)

To be added.

THERMOACOUSTICS

The generation of sound waves by heat.

THERMOACOUSTIC ENGINE

Essentially a closed resonant chamber containing three heat exchangers and a working fluid—such as an inert gas—under pressure, and used with a transducer to provide power such as mechanical power.

One heat exchanger is the hot-side exchanger through which heat is injected, another is the cold-side exchanger in which a coolant is circulated, and the third is a passive device between the other two, and known as a regenerator or thermal capacitor. The heating and cooling changes the pressure in the fluid greatly, and the regenerator oscillates back and forth, taking energy from the fluid, then giving it back, and keeping the engine going. A pressure wave (shock wave) moves through the working fluid and produces a resonator effect. The end result is an engine with rapidly fluctuating pressure gradients and no moving parts. The pressure gradients are then tapped off to drive a piston, which drives a linear alternator producing electricity. In a typical application, a single piston with a surface area of 1.5 square inches, and with a travel of only one centimeter, operated at 500 Hz, will produce about 30 horsepower output.

Information on these engines can be found on the Fellows Research Group website at http://www.io.com/~frg, or E-mail to frg@io.com. Many thanks to Oscar L. Fellows, "Be Our Guest: Thermoacoustic Engines—Propulsion, Power for the 21st Century," **Power Online**, 1/24/2000.

THERMODYNAMIC EQUILIBRIUM

Stable state eventually reached by an isolated system. One may consider it a "balanced" or "preferred" state.

Actually, a totally isolated system could not change. Any system is in a constant exchange with its vacuum environment, a priori, and no such thing as a completely isolated system exists in nature. Hence once excited, a system—to reach equilibrium—must exchange with its environment and is not a truly isolated system. But systems do exchange with their environment in hidden forms, so we can just let the original statement stand as is. Sloppy, but okay so long as we recognize the limits.

TIME

To be added.

TIME-CHARGE

To be added.

TIME-CHARGING

To be added.

TIME-ENERGY

To be added.

TIME-EXCITATION

To be added.

TIME-FLOW MECHANISM

To be added.

TIME-POLARIZED ELECTROMAGNETIC WAVE

An electromagnetic wave whose vibrations occur in the time domain only.

The wave is theorized in quantum field theory, or at least its time-polarized photon is theorized and called a "scalar" photon. There seems to be no Western method for detecting such waves, though they are predicted by theory.

TIME-POLARIZED (SCALAR) PHOTON

A photon whose vibrations occur in the time domain only, also known as a "scalar" photon.

TIME-FORWARD WAVE

A normal wave moving in forward time. Otherwise known as the retarded wave.

The external observer comfortably sees its forward progress as expected.

TIME REVERSAL

The process of forming the phase conjugate

...(e.g., of a wave or particle or entity or process or vacuum engine.).

We accent that *time-reversing a single object or single group of objects* is not the same thing as "travel into the past" so popularized by science fiction. For time travel to the past, the entire universe and everything in it except the traveler, would have to be time-reversed. That would not seem possible by any stretch of the imagination today! On the other hand, time-reversal of a single thing such as a particle or a wave—or even a group of things such as a group of particles or waves—is not only feasible but is readily achievable. A positron is a time-reversed electron, e.g. Becker showed that laughably weak DC electrical trickle currents (picoamperes) across otherwise intractable bone fractures could engender healing of the bone fractures. In investigating the effect, he showed that entire cells—specifically in that case, the red blood cells—were time-reversed (dedifferentiated) and then fast-forwarded (redifferentiated) into new kinds of cells, *genetics and all*.

In France in the 1960s and early 1970s, the Priore team showed in laboratory animals that this mechanism (which they nor anyone else understood they were unwittingly evoking) could be used to dramatically reverse and quickly cure terminal tumors, infectious diseases, arteriosclerosis, and suppressed immune systems. The same approach, if redeveloped, could stop and totally wipe out the emerging major new pandemic of hepatitis C, which nothing will touch at present and which will surpass AIDS infections after the turn of the century. The approach could totally reverse AIDS and eliminate all HIV infection, because it could simply time-reverse HIV-infected cells back to normal cells, genetics and all. In other words, it could readily and cheaply and quickly eliminate AIDS from the face of the earth. Further, the disease could be eliminated whenever a patient tested positive for HIV infection; there would be no need to wait for debilitation to set in. For coverage in detail, see T. E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy-Medicine. Parts I and II." **Explore**!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

TIME-REVERSAL ZONE

A region of space in which a component of the overall time flow runs backwards, and hence the fundamental mechanism generating the flow of a mass through time is reversed.

Such zones are theorized as temporarily occurring in cold fusion, particularly once the palladium lattice is loaded with hydrogen or deuterium ions. Such a temporary region in the electrolyte is an excited state, and the other ions will move to nullify it and cause its decay. During its existence, ions in it experience a reversal of the law of attraction and repulsion of charges.   In the TRZ, like charges attract and unlike charges repel. Thus suddenly the low spatial energy transmutation of elements in cold fusion experiments can be explained (they have been validated in some 600 successful experiments, in various laboratories in various nations, by excellent scientists). E.g., two H+ ions may attract, and so closely that each enters the edge of the strong force region of the other, forming a quasi-nucleus. Then as the TRZ decays, the preferred energetic change is from inside out. In other words, the gluon forces have been partially nullified and the quarks in the nucleons are almost freed. Hence the favored energetic decay from the excited state of the TRZ is by quark flipping, where one H+ ion in the quasi-nucleus changes into a neutron. The quasi-nucleus becomes a quasi-nucleus of deuterium, and will tighten into deuterium as the TRZ induced action vanishes. This is the way the excess deuterium is formed in the cold fusion experiments. Two deuterium D+ ions may also form a quasi-nucleus of (D+)(D+) particularly where the electrolyte is loaded with deuterium to start with. In that case, as the TRZ decays, the quasi-nucleus just tightens into a helium 4 nucleus or alpha particle, accounting for the excess alpha particles formed. And so on.

The concept of the TRZ thus is consistent with hundreds of successful experiments, and introduces a completely new type of nuclear reactions to physics, at low spatial energy where transmutations such as occur are not presently thought to be possible. We point out that these are extremely high energy reactions when the time-energy interaction is considered, since time has an energy density equal to mass. Time-energy is some c^2 times as dense as is spatial energy used by particle physicists in their interactions. By transposing some of the time-energy into spatial energy as the TRZ decays, the new "low energy nuclear reactions" actually involve the necessary degree of "high energy" required for the interactions.

TIME-REVERSED WAVE

A phase conjugate reflected wave, which is a wave moving in reversed time.

Since the external observer "sees" in forward time only, he sees the wave as traveling in the opposite direction as its generatrix wave front is moving.

TOPOLOGICAL

Invariant under transformation by continuous mappings,

...or related to the properties of geometric shapes and figures that remain invariant under transformation by continuous mappings.

TOPOLOGICAL PROPERTY

Those properties of a geometrical space that are unchanged by continuous distortion of the space.

TOPOLOGY

Mathematics of the properties of geometric configurations invariant under transformation by continuous mappings.

In short, the mathematics of the large-scale structure of curved spaces.

TOROIDAL COIL

A coil of wire curved around into toroidal form.

Contrast to the standard solenoidal coil form.

TOROIDAL TRANSFORMER

A transformer comprised of toroidal coils instead of solenoidal coils.

TRANSFORMATION MATRIX

A matrix (rectangular array of algebraic or numerical quantities) for replacement of the variables in an algebraic expression by their values in terms of another set of variables.

Used, e.g., in solving equations involving Hamiltonians.

TRANSVERSE

Wave movement perpendicular to the wave's primary direction of principal propagation, as in the movement of a taut rope in a wavelike motion traveling down the rope.

(As opposed to longitudinal).

TRANSVERSE EM WAVE IN THE VACUUM

The completely erroneous notion that EM disturbances in the vacuum are composed of transverse "string waves."

Faraday, e.g., believed that the EM field in the vacuum consisted of his "lines of force" as material taut strings—erroneously without the necessary holders to put the tension forces on the strings. He thought that when there was an EM disturbance, it was a "plucking" or transverse vibration of these "taut strings." Maxwell also believed in Faraday's material lines of force, and set out to capture them in his electrodynamic theory. He did so by simply assuming the lines (tubes) of force and their transverse "string wave" vibration. This is equivalent to assuming that the electromagnetic vacuum itself is composed of material taut strings, and that additional EM fields are just additional taut material strings.

TRANSVERSE FIELD COMPONENT

The modeled component of the electromagnetic field (as in a moving EM wave in vacuum) that is oriented at right angles to the line of motion.

TRANSVERSE WAVE

A wave where the medium is undergoing repetitive lateral translations, as a lateral wave in a taut string.

The EM wave is vacuum is erroneously modeled as a transverse "string wave," as a relic from Faraday's original belief that EM fields existed in the (assumed) material ether as physical taut strings. He thus believed that EM disturbances were the disturbances of "pluckings" of these taut strings. Maxwell accepted Faraday's physical lines of force and constructed his electrodynamics mode to specifically include the material ether, the physical "taut strings" lines of force notion, and Faraday's "string waves" as the electromagnetic vibrations in the ether. This discards half the energy, half the phenomena, half the wave, and Newton's third law causative EM mechanism.

TUNNELING

In quantum mechanics, a phenomenon where a particle penetrates and crosses a small region where the opposing potential is greater than the particle's available energy.

The phenomenon is thought to be impossible according to classical physics, but occurs nonetheless. Further, the speed of transit through the "tunneling" separation can be superluminal. Tunneling of music between two points has been measured to occur at more than four times the speed of light.

A more extended type of tunneling is the travel of longitudinal EM waves infolded inside the normal EM potentials, fields, and waves. Indeed, this is a more fundamental electrodynamics, and it can be shown that interferometry of this "internal" electrodynamics creates the external electrodynamics by a form of longitudinal EM wave interferometry. E.g., see M.W. Evans, P.K. Anastasovski, T.E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," **Journal of New Energy**, 4(3), Special Issue, Winter 1999, p. 76-78.

UNCONSCIOUS MIND

To be added.

UNIFIED FIELD THEORY (ENGINEERABLE)

A unified theory of the four forces of physics—the electromagnetic, gravitational, strong, and weak forces —that is not just an intellectual model but which is also engineerable on the laboratory bench and in actual physical systems using higher symmetry O(3) electrodynamics.

That's a mouth-filling phrase (yes, we sometimes write awkward phrases unintentionally; it's amazing how reasonable such things sound when you're laboring along at 3:00 a.m. and bleary-eyed as a wet toadstool!)

UNIFIED FIELD THEORY OF MIND AND MATTER

To be added.

UNITARITY

Principle that all probabilities sum to unity.

UNITARY Q_E

Fixed or unit charge q of the electron.

This notion is held only because the q is not decomposed into its component elements, but treated as if it were an indivisible unit. That is in error! The mass of a fundamental particle at rest is quantized, and the massless charge (self-potential) of the particle is discretized as a function of the background potential (virtual particle flux density) in which it is embedded or to which it is exposed. In our view the definition of a charge q is $q \equiv m_q\phi_q$. The *electrical charge* should be taken to be the massless ϕ_q component. The definition of charge q_e of an electron is rigorously given by the identity $q_e \equiv m_e\phi_e$. The *electrical charge* of the electron should be taken to be the massless ϕ_e component.

If a charge (say, an electron) is placed in an additional potential ϕ_1, then the total massless charge of the electron (i.e., its virtual vacuum flux exchange with mass me) is $(\phi_e+\phi_1)$. As can be seen, the massless charge of the electron (its virtual flux activity) has now changed. The electron is then said to *be collecting and to have collected* excess energy (which may be either positive or negative, depending upon the sign of ϕ_1.). Further, either ϕ_e or ϕ_1 or both can be internally and deterministically *dimensioned* (structured in Stoney-Whittaker-Ziolkowski fashion). When its massless charge (its own potential ϕ_e) is so dimensioned, then the dimensioned electron is said to contain or transport a *vacuum engine*.

VACUUM ENGINE

Particle view: Deterministic internal pattern or template deliberately created in the virtual flux of a scalar potential.

It can also be created in, and transported by, the two scalar potentials that comprise a static or dynamic force field, or a moving EM wave in the vacuum. It can also be created in the local vacuum, in which case it has a "charge-up" creation time and a "discharge time," exhibiting a sort of "time constant" very much like a capacitor charging and discharging. In the wave view, the vacuum engine is a deterministic internal pattern or template deliberately created in the Stoney-Whittaker-Ziolkowski internal biwaves and wave products comprising the scalar potential. Again, it can be implanted in the vacuum or in an EM signal as a carrier. The self-energy (scalar potential) portion of any electric charge or magnetic charge can also be a "carrier" that is conditioned (*dimensioned*) with a vacuum potential. The process or act of conditioning any of the carriers with vacuum engines is called *dimensioning* that carrier. Russian energetics refers to the vacuum engines of a dimensioned carrier as the *information content of the field*.

VACUUM FLUX EXCHANGE

The virtual particle interaction between the active vacuum and a charged particle.

The charged particle is an asymmetry in this flux exchange interaction, and extracts and gates some of the vacuum energy as a Poynting **S**-flow. In turn, this S-flow transports and creates in space the **E** field and the scalar potential f associated with the charged particle. An electrically charged particle is thus a little Poynting energy generator, a source of potential, and a source of **E**-field. It is a little electrical dynamo, a little free energy generator. It is already an open system in disequilibrium with its environment—the fierce vacuum flux. Its self-energy (potential) and its "fields," plus its Poynting energy flow, are the result of its asymmetry in the vacuum flux exchange.

VACUUM'S VIRTUAL PARTICLE FLUX

In the particle viewpoint, the quantum mechanical vacuum identically is a virtual particle flux.

Present physics, however, is prone to refer to the vacuum as (an implied emptiness) containing a virtual particle flux. Energy, vacuum, spacetime, virtual particle flux, and dynamic geometry are all synonyms.

VECTOR

In mathematics, an entity completely characterized by a magnitude and a direction.

VECTOR POTENTIAL

In electrical engineering: "Potential postulated in electromagnetic field theory."

Space differentiation (*curl*) of the vector potential yields the field. Magnetic vector potential is due to electric currents, while electric vector potential is assumed to be due to a flow of magnetic charges." [**Chambers Science and Technology Dictionary**, Peter M. B. Walker (Ed.), Chambers Cambridge, 1988, p. 947.

Note that the foregoing is not yet a definition, saying only that it is a postulated potential, then giving some characteristics. Informally, in my view the vector potential is just sort of an ordinary scalar potential, every point of which is in motion in something like a swirling fashion. The magnetic vector potential **A** is most common. There doesn't seem to really be a definition of **A**; the equation $\mathbf{B} = \nabla\times\mathbf{A}$ is usually (erroneously!) said to "define" the **A**. So we are left with the question **A** ≡ ????, since no equation is a definition of anything.

Electrodynamicists have no difficulty in calculating the direction and magnitude of **A**, and in using it, but so far as I can find there is still no real definition. As an ad hoc definition I consider the **A**-potential to be identically a virtual photon flux of vacuum that is moving as an ensemble along the direction of the **A**-vector. Inputs for a more rigorous definition of the form **A** ≡ (x) would be most welcome!

At any rate, the **A** is known to be quite real and an independent "field" of nature, because when the curl operator is removed from operating on **A**, the **B**-field is zeroed but the *curl-free* **A**-field can still remain and interfere with the fields of charged particles and magnetic poles to cause real effects in physical systems, including at a distance. This latter is known as the Aharonov-Bohm effect.

VECTOR FIELD

In ordinary theory, the assigning of a magnitude and a direction (that is, a vector) to each point in a region of space.

The entire set of vectors and their spatial points is called a vector field.

VECTOR ZERO

See discussion under *zero vector*.

VIRTUAL ANTIPHOTONS

An antiphoton that spontaneously appears and disappears, so swiftly that it cannot be individually observed.

VIRTUAL PARTICLE

A fleeting quantum particle that spontaneously appears and disappears, so swiftly that it cannot be individually observed, so that it exists only temporarily.

The virtual particle does not satisfy the usual relation between energy, momentum, and mass because it is underneath the Heisenberg uncertainty principle. The virtual particle can have any amount of energy momentarily, so long as the product of its energy and the time interval of its existence is less than the uncertainty principle's minimum magnitude.

Nonetheless, the interactions of large numbers of virtual particles with a mass or charge can combine to generate real observable effects. In quantum field theory, all forces of nature are caused by the interaction of the forced mass entity with virtual particles. Interaction of a mass with virtual photons, e.g., is projected to cause all mechanical and electromagnetic forces.

VIRTUAL PARTICLE FLUX EXCHANGE

Interaction with or exchange of, a flux of virtual particles.

Interaction of a mass with virtual particles most often consists of the mass absorbing the virtual particles and re-emitting them.

VIRTUAL PHOTON FLUX (VPF)

A flux of virtual photons.

(Hey, how's *that* for a tautology!) First, see definition and discussion under *flux*. For electromagnetics, the vacuum can often be modeled as a flux of virtual photons. One gets the virtual charges also, since the photon is considered to also consist of a positron-electron pair. That is, the photon is continually separating into a positron-electron pair, which then annihilate each other to form additional photon(s), etc.

VIRTUAL PHOTON FLUX EXCHANGE

The mutual exchange of virtual photons, as between the vacuum and a charged mass, or between two colliding masses.

The emitted photons from one mass are absorbed in the other mass and re-radiated.

VIRTUAL PHOTON FLUX OF VACUUM

The vacuum identically is a flux of virtual particles.

For most electromagnetic cases, it can just be modeled as identically a flux of virtual photons.

VIRTUAL PHOTONS

Photons which spontaneously appear and disappear in the vacuum so quickly that they cannot be individually observed.

VIRTUAL POSITRONS

Positive electrons (positrons) which spontaneously appear and disappear in the vacuum so quickly that they cannot be individually observed.

WHITTAKER, E. T.

Well-known mathematical physicist, who dramatically extended Stoney's earlier work of 1897-8 and authored two papers in 1903 and 1904 giving the basic theory of scalar electromagnetics (Russian energetics, not including the quantum potential portion).

The 1903 paper is E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," **Mathematische Annalen**, Vol. 57, 1903, p. 333-355.

The 1904 paper is E.T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," **Proc. Lond. Math. Soc**., Series 2, Vol. 1, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

WHITTAKER'S DECOMPOSITION OF THE SCALAR POTENTIAL

Decomposition of the scalar potential into a harmonic set of bidirectional longitudinal EM phase conjugate wavepairs.

In his paper "On the Partial Differential Equations of Mathematical Physics," **Mathematische Annalen**, Vol. 57, 1903, p. 333-355, Whittaker showed that the scalar EM potential can be decomposed into a harmonic series of bidirectional EM wavepairs. In each pair, one wave is the phase conjugate of the other, and exists in the time domain. However, this phase conjugate wave has been considered as having been detected by intercepting charge as a 3-space EM wave travelling in a reversed direction. Hence from a detectable viewpoint, the Whittaker decomposition yields bidirectional longitudinal EM waves in 3-space, as has been pointed out by several electrodynamicists (e.g., Barrett).

However, the present author considered the difference between the phase conjugate wave as it exists prior to interaction with the assumed intercepting unit point charge, and the detected 3-space wave after the detecting charge is assumed. By applying the Whittaker decomposition and also considering the nature of the EM wave in spacetime *prior to interception by a unit point charge*, this led to the discovery of a new and more primary 4-symmetry in EM energy flow, once 3-symmetry is broken, that exists automatically between the time domain (complex plane) and 3-space.

Thus because a dipole (or a charge treated as a set of composite dipoles) breaks 3-symmetry (as is well-known in particle physics), the more general 4-symmetry results automatically when the nature of the potential between the ends of the dipole(s) is considered. This yields the fundamental mechanism by which EM energy is extracted from the vacuum (from the complex plane or the time domain) and output in 3-space from a "source charge" or a "source dipole". In short, it yields the giant negentropy mechanism when the unaccounted Heaviside energy flow component from a charge or dipole is considered as well as the intercepted tiny Poynting energy flow component. See T. E. Bearden, "Giant Negentropy from the Common Dipole", **Journal of New Energy**, 5(1), Summer 2000, p. 11-23; also on http://www. cheniere.org.

In his second paper, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," **Proc. Lond. Math. Soc**., Series 2, Vol. 1, 1904, p. 367-372, Whittaker showed that the interference of two such scalar potentials can and does create all the field energy, patterns, and functions of classical electromagnetics.

Whittaker's second paper initiated what today is loosely referred to as *superpotential theory*. For an excellent overview discussion of superpotentials and related things, see Melba Phillips, "Classical Electrodynamics," in **Principles of Electrodynamics and Relativity, Vol. IV of Encyclopedia of Physics**, edited by S. Flugge, Springer-Verlag, 1962.

WORK

The changing of the form of energy.

WORK, EXTERNAL

The changing of the form of the external energy of a body or system.

WORK, INTERNAL

The changing of the form of the internal energy of a body or system.

WORK-ENERGY THEOREM

To be added.

ZERO POINT ENERGY

The minimum energy of a system due to its quantum fluctuations, resulting from its incessant virtual particle activity. Quantum mechanically, no system of interest (including even spacetime itself) can have zero energy.

The zero-point energy of the vacuum is the lowest energy vacuum state, with fluctuations taken into account. Even at low energies, quantum fluctuations continually arise, and result in an incessant, extremely rapid, and violent "jittering" of the energy momentarily present. The minimum energy due to these quantum fluctuations is called the *zero-point energy.*

ZERO VECTOR

(1) A vector having no length (no magnitude) or specific direction.

(2) Absence of any or all finite vectors.

(3) Resultant of a system of multiple, finite vectors whose vector sum has no specific direction.

Note that the second definition differs drastically from the first. That is, we cannot say that a vector resultant zero has no infolded magnitude, if we consider its components, all of whom have magnitude. See the discussion under "zero." In vector analysis, the zero vector axiom states that any vector v plus a zero vector is equal to a zero vector plus v. Note that, being an axiom, this is not proven, but just assumed. We very much disagree with that blanket axiom; it reduces vector analysis to a highly special single case, where all zero vector systems are to be considered equal, and totally "inactive." In the real world that need not be true at all. There are an infinite number of different zero vector resultant systems, and so zero vectors may differ, if their infolded component "substructure" is considered. In the real world, things that are vectors very often have energy. The energy is infolded and hidden, so that it represents not only a local curvature of spacetime, but also a structured, deterministic set (a *template*) of curvatures of local spacetime. Therefore this zero vector system is actually a *vacuum engine*. Any mass system placed in such a potential, will be acted upon in its parts by that local vacuum engine. Further, the zero vector resultant system is a "potential," since it does contain "collected internal energy." This potential is obviously an artificial potential, since it is deterministically structured internally.

So between zero vector resultant systems, the "infolded, hidden energies" differ dramatically in

(i) total infolded and trapped (collected) energy,

(ii) specific internal action patterns (templates),

(iii) general relativistic effects, and

(iv) type of vacuum engine. Accordingly, the actions the zero-vector system induces in an exposed mass may be designed in advanced. This is vacuum engineering. This is spacetime engineering.

The problem is the *vector algebra itself*. In the abstract algebra, a "vector space" is regarded as an inert thing, where the only actants are the overall net nonzero vectors, and there is no interaction between an overt vector and a covert set o f vectors. The vector space is not allowed to have an internal realm, stress, hidden engines, or to interact with the vectors representing the physical system. Consequently, for over a century electrodynamicists—thinking primarily of translation—have routinely discarded such "zero vector systems" as if they represented the total absence of any finite vector. *This they do, in the unsuitable assumptions of the vector algebra*. This they do not do, in the real world.

The common practice in electrical physics of replacing a zero-summed system of nonzero vectors with a vector zero of "complete absence of vectors" is incomplete, and such did not exist in Maxwell's original *quaternion* theory. This present practice should be changed; it has reduced the topology of electromagnetics far too much, and gutted the most important parts of the theory insofar as an extended electromagnetics is concerned. The vector zero should be replaced with a special organized vector zero system and a scalar potential. The interactions of the "hidden" vectors should be incorporated. By failing to do this, physics excludes the ability to engineer electrogravitation, spacetime, the virtual state, local general relativity, free energy, effects at a distance, and the probabilities of the states propagated by the Schrödinger equation. It also excludes unification of all forces.

In the new view, a zero vector may be a system of nonzero vectors that vectorially sum or multiply to zero for translation purposes, but the components may still exist "inside" and dynamically function inside the zero-resultant envelope. This substructure can act on nonlinear or resonant systems. Any or all components may be time varying, or none of them may be. The zero vector system may thus be a dynamic vacuum engine. In addition, such a vector zero system is considered to be also a scalar stress potential, whose magnitude is equal to the sum of the absolute values of the perpendicular components. In addition, a separate type of stress potential may be included for the magnitude of the absolute values of the swirl components (torques). Other associated potentials may be included for the system, particularly if the system is in multiple dimensions, if it consists of more than one type of virtual particle flux, if several fluxes intercommute, etc. Such a system can also be designed and created to deterministically interact with and on the life force itself, the mind, thought, and personality, but that is beyond the scope of this discussion.

ZERO-VECTOR-SUMMATION SURFACE

A surface or assumed spatial surface in which vectors under consideration interact to sum to a vector zero resultant.

ZERO-VECTOR-SUMMATION SYSTEMS

Zero summations of real, nonzero vectors, so that the real vectors are still present but do not cause translation of exposed mass or charges.

With respect to observing/detecting process that depend upon charge translation, the vectors are said to be hidden and infolded —i.e., with respect to detection by translation. See discussion under *zero vector*.

ZIOLKOWSKI, RICHARD

To be added.

Return to top of page

The "Porthole" Concept with technical background

Amplifying the cellular regenerative (healing) system of the body for treatment and healing
of mass casualties

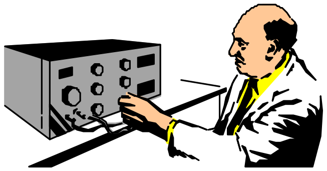

Outline of the presentation

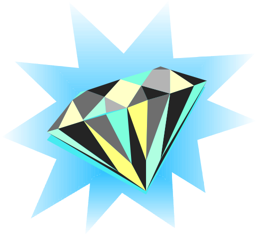

- Mechanism for deep EM exchange of a body with its active EM environment
- Becker's epochal work and its ramifications
- Kaznacheyev's startling results and ramifications
- Weaponization of Kaznacheyev's findings
- Importance of time-energy and the time-domain
- Mechanisms for medical treatment and healing

Outline of the presentation (2)

- Prioré's revolutionary results and their ramifications.
- Proposed new mechanism for cancer promotion.
- Porthole concept: The 2-minute shortcut.
- Terrorist biological warfare capabilities.
- Summary and conclusion.
- Budgetary and Program package.

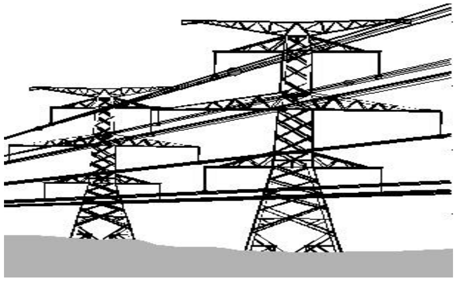

Mechanism for the body's deep EM exchange with its active EM environment

And some ramifications

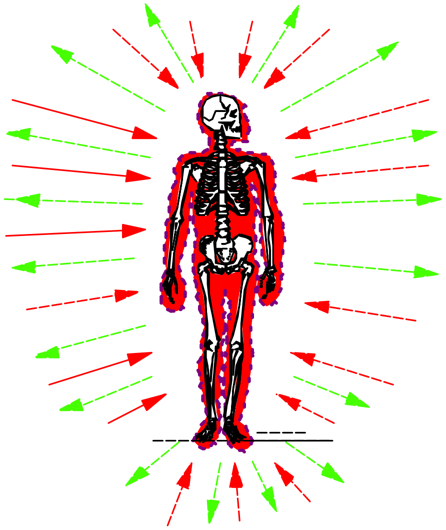

The body continually receives and retransmits EM energy

- The body may be considered as a dielectric.
- Every part of the dielectric participates in the absorption and emission of each photon to and from any point on the surface.*
- Mechanical force effects are also electrical in nature and exchange similarly.

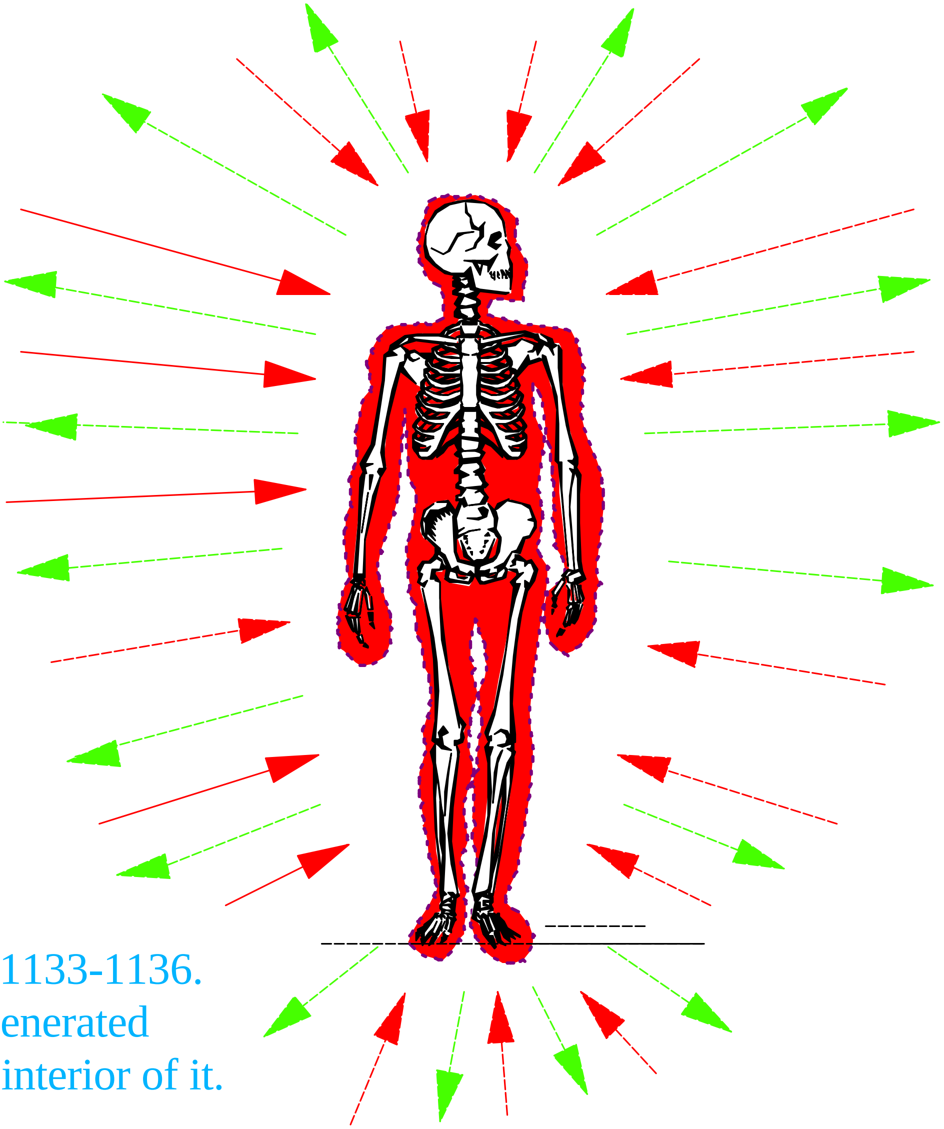

* Reali, G. C., "Reflection from dielectric materials," American Journal of Physics, 50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.

Ramifications:

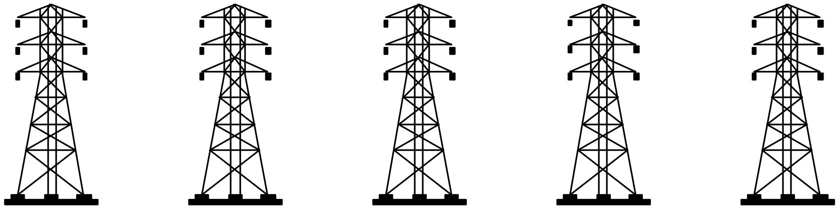

- There is an EM "dielectric transmission path" connecting every part of every cell to every part of the surface of the body.
- All electrical processes in the body, including deep within the cell and its nucleus, participate in the absorption and emission of every photon to and from the body's outer surface (and every mechanical force* experience).
- The EM emission energy of the body is a conglomerate mix of fractions of all processes and functions ongoing in the body or affecting it, including in every cell.
- This includes all electromagnetic, mechanical, chemical, and environmental effects upon and in the body and every tiniest part of it, and provides an extended definition of "stress".

* At base, all normal mechanical forces are generated by electromagnetic forces.

Ramifications (2)

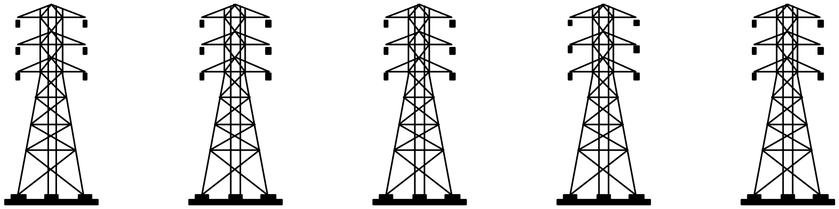

- The reaction of each photon absorbed on any point of the outer surface of the body, connects through the dielectric transmission path to every part of the body, including to every part of every cell*.
- Every emission from the body is participated in by every part of the body, including every part of every cell.
- The body's emission "changes" its environment, adding EM radiation. Some of the internal induced changes (excited states) in the body decay by this external emission. Some do not, but must be otherwise dealt with internally by the body systems.
- The body's EM absorption allows the environment to change or affect every function, operation, and part of the body, by adding EM radiation. All are continually being affected and deviated, both helpfully and harmfully, by these deeply induced EM interactions.
- The body's reaction (healing) must continually readjust and correct deviations from normalcy, in every part of every cell.

* In quantum physics, part of the photon is localized and part is nonlocalized, reaching even across the universe.

Ramifications (3)

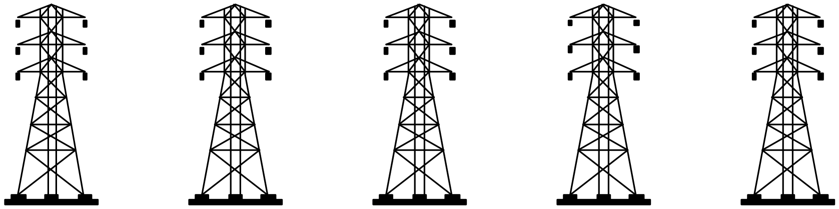

- The sum total of all the impinging external EM radiation upon the body, plus mechanical "special EM" radiation, thus affects and changes the body and all its processes, participating in them.
- That impinging EM radiation can be manipulated to either harm or heal the body and any or all of its cells.
- The body must continually overcome these externally-induced EM changes in its cells and processes, to maintain normalcy. Some EM changes are restored by the subsequent dielectric emission process, relieving the previously-induced deviant EM excitations in its internal processes. Some are not.
- The poorly studied *cellular regenerative system* performs the remainder of the restorative function, which is the fundamental "healing" function of the body.*

*The cellular regenerative system was mostly studied by Becker and a few others. See R. O.Becker, and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, No. 73, Nov.-Dec. 1970, p. 169-198.

Becker's epochal work

And its ramifications

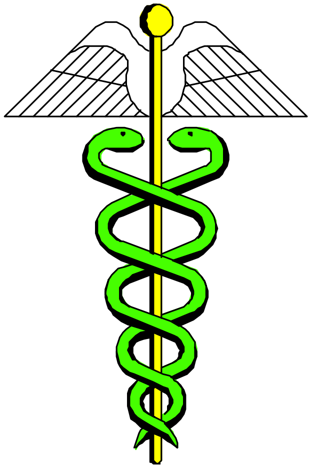

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

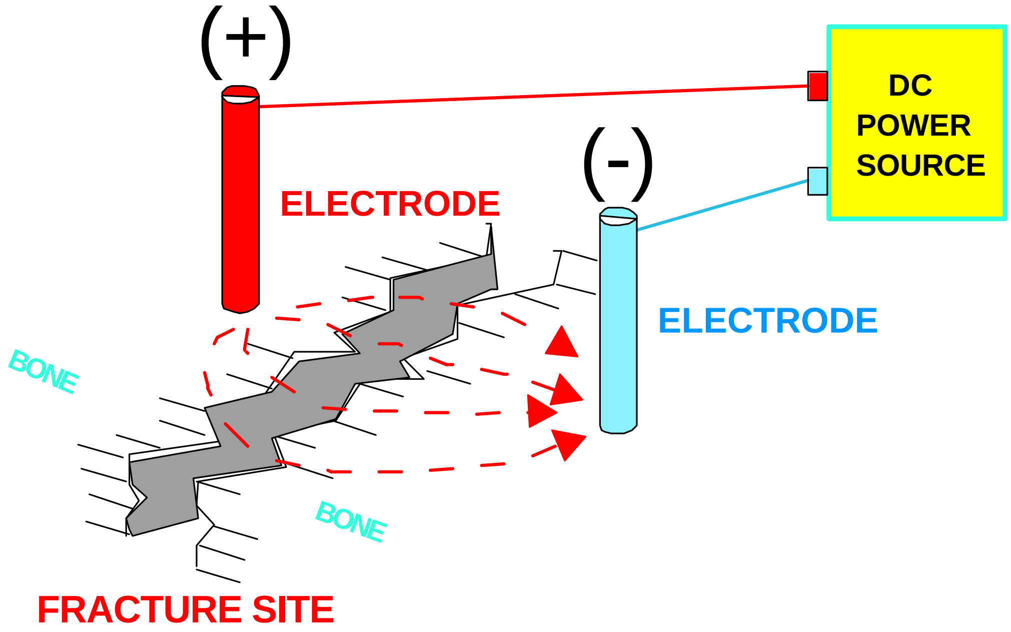
(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Becker's theoretical DC control system involved with response to injury

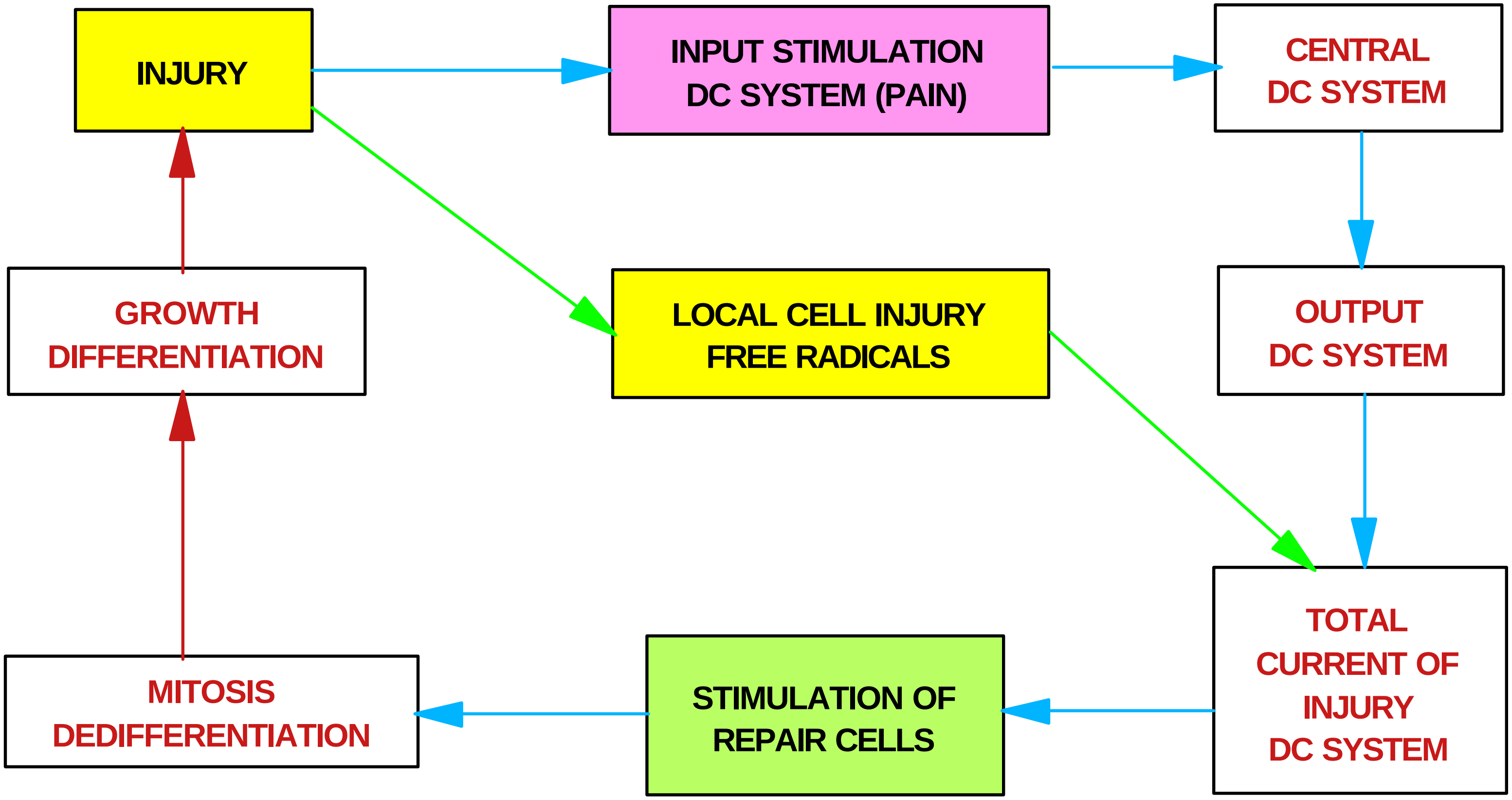

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

GROWTH DIFFERENTIATION

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Becker's proposed control system governing regeneration*

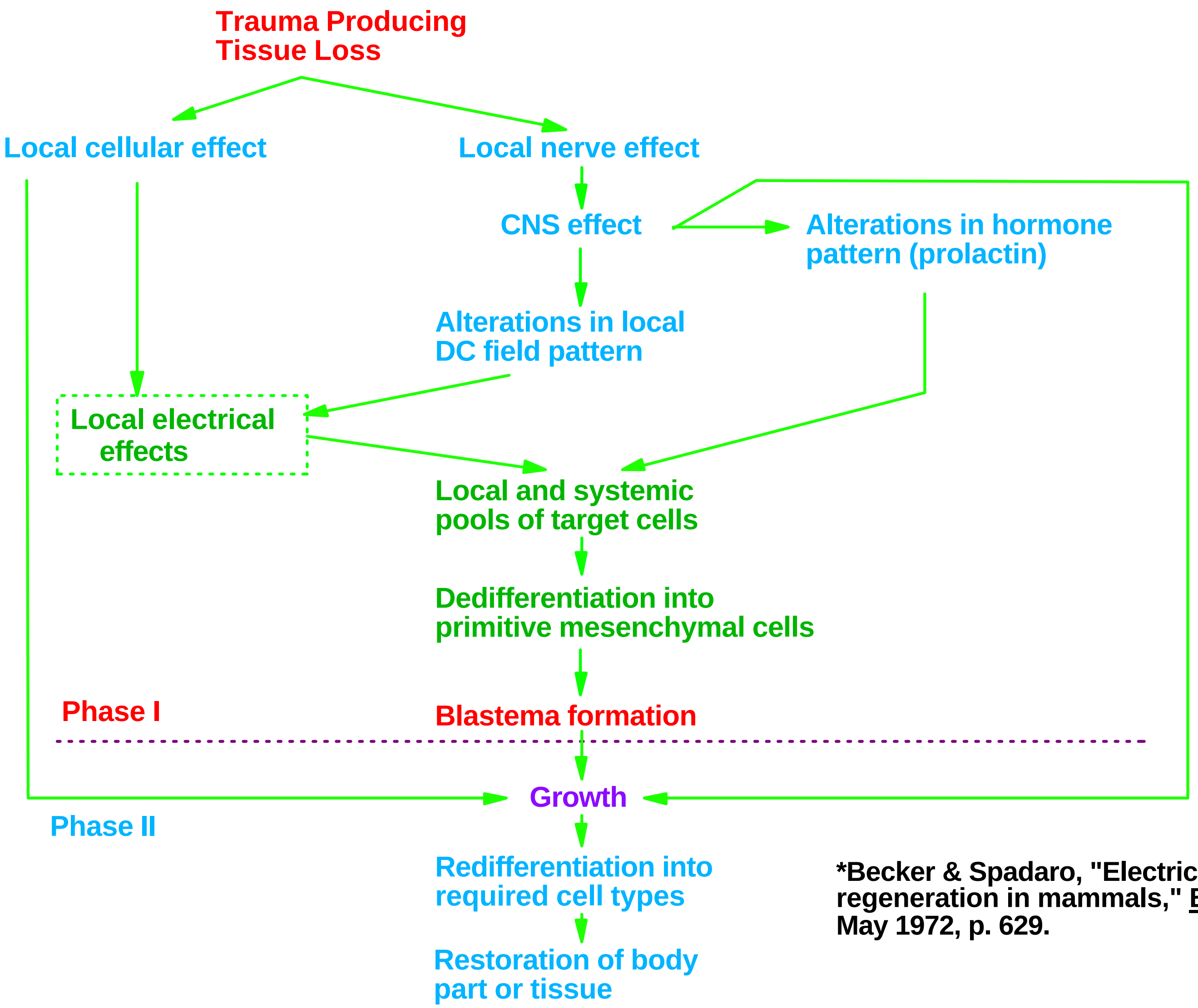

Trauma Producing
Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone
pattern (prolactin)

Alterations in local
DC field pattern

Local electrical
effects

Local and systemic
pools of target cells

Dedifferentiation into
primitive mesenchymal cells

Phase I

Blastema formation

Growth

Phase II

Redifferentiation into
required cell types

Restoration of body
part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Extending Becker's work

- Becker had only the archaic classical electrodynamics available to him. That EM is grossly inadequate.
  - Insufficient group symmetry to model the process.
  - Already eliminates the "infolded" longitudinal wave EM that structurally and dynamically comprises all "normal" EM potentials, fields, and waves.
  - It is this internal LW EM that is manipulated and used by the cellular regenerative system.
- Nonlinear phase conjugate optics was not yet developed when Becker did his seminal work.
- Successful higher symmetry electrodynamics models -- such as O(3) -- have been developed and are available.

A sad commentary

- Becker was nominated for the Nobel Prize, but he was hounded, his funds were withdrawn, and he was forced to retire early.
- His main "sin" was having the courage to testify against the harmful effects of unrestricted powerline radiation and electronic smog.
- Powerful interests then crushed him, set up their own well-funded institutes, and hired scientists who then predictably found that powerline radiation was no problem.
- Such research is so tainted that recently several important journals have now required the authors of submitted papers to state what their affiliated companies do, and what "company massaging" of their research occurred.
- From a higher symmetry EM view, it is obvious that the dielectric pathway and quantum electrodynamics already show a significant problem due to such radiation, particularly with respect to long-term effects.
- This is particularly significant since "noise" -- the mere presence of more "hash" -- appreciably amplifies the dielectric transmission effect. So the major factor becomes EM smog *signals density*, not signal power.

Kaznacheyev's startling results

And its ramifications

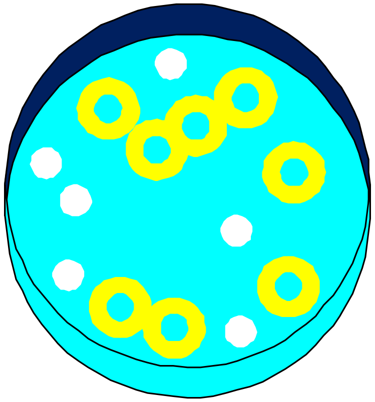

Proof that the cellular condition is radiated from groups of cells

- It follows that a diseased, damaged, or infected cell culture will and does emit its exact internal EM condition, hence the disease or damage "generatrix form" itself (in the internal structuring of the EM radiations).
- Further, these "state" emissions can couple into other targeted cells to produce disease and disorder in them.
- This was clearly proven in some 15,000 Russian experiments by Kaznacheyev.
- Kaznacheyev demonstrated induction of cellular disease and disorder between EM-coupled but otherwise environmentally shielded cell cultures
- Similar experiments have been replicated in the West by Reid *et al*., and by Popp *et al*.

Kaznacheyev's induction of cellular disease and disorder

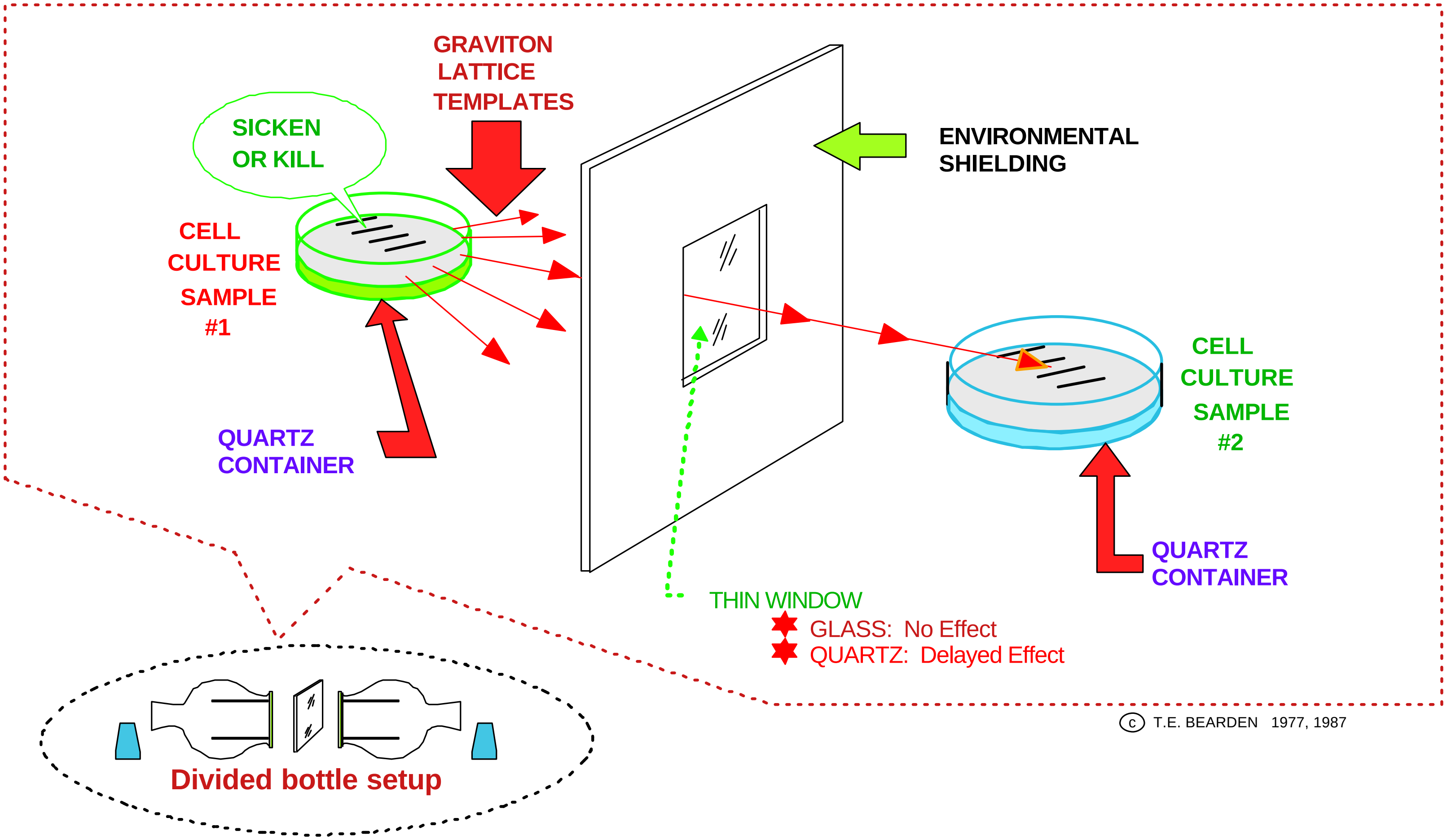

GRAVITON LATTICE TEMPLATES

SICKEN OR KILL

ENVIRONMENTAL SHIELDING

CELL CULTURE SAMPLE #1

QUARTZ CONTAINER

CELL CULTURE SAMPLE #2

QUARTZ CONTAINER

THIN WINDOW

- GLASS: No Effect
- QUARTZ: Delayed Effect

Divided bottle setup

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Another surprising feature of Kaznacheyev's experiments

- The experiments do not work if the targeted cells are in normal light, such as sunlight*.
- The reason is that the normal light bandwidth occupies a single harmonic interval between ultraviolet and infrared.
- Hence its difference frequencies also fill such an interval. For a set of n frequencies, the number *d* of difference frequencies is $d >> n$.
- Thus the visible light spectrum represents "jamming" of specific internal EM signal inputs through the dielectric pathway into the body, turning much of the environmental signals into harmless warmth.

* It has also been shown that if bacteria are killed in the "dark" by UV, then 24 hours later placed in sunlight, a substantial fraction of the "dead" bacteria will revive.

Ramifications of Kaznacheyev's experimental results

- At least one harmonic interval is necessary for the effect to be evidenced in the target cells.
- Interestingly, if we assume the body dielectric to be isotropically nonlinear (to first order):
  - Velocities of the actual transmitted frequencies depend upon the particular point on the wave amplitude.
  - The transmitted frequencies overshoot, interfere, breakup, etc. The difference frequency does not.
  - The "difference frequency" is transmitted through the nonlinear dielectric medium as if it were a sine wave passing through a linear medium*.
- To our knowledge, this direct and deep communication of electronic "hash" difference frequencies in dielectric human bodies has not been investigated in the West.

* Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

Ramifications of Kaznacheyev's experimental results (2)

- A heterodyne signal -- at the difference frequency between an input signal and a reference signal -- can be enhanced by adding noise*.
- Thus the signal of the disorder in the cellular pattern emitted from diseased cells, received by normal cells, can be amplified electronically and rather easily.
- Powerline radiation (electronic smog) thus carries its own amplification mechanism. Eventually the amplification overcomes any squelching by sunlight, etc.
- The Russians promptly (as early as the late 1950s) weaponized and tested these and similar effects.

* Dykman, M. I. et al., "Noise-enhanced heterodyning in bistable systems," Physical Review E, 49(3), Mar. 1994, p. 1935-1942

Weaponization of Kaznacheyev's findings

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

And its ramifications

"Microwave" radiation of the U.S. Embassy in Moscow.

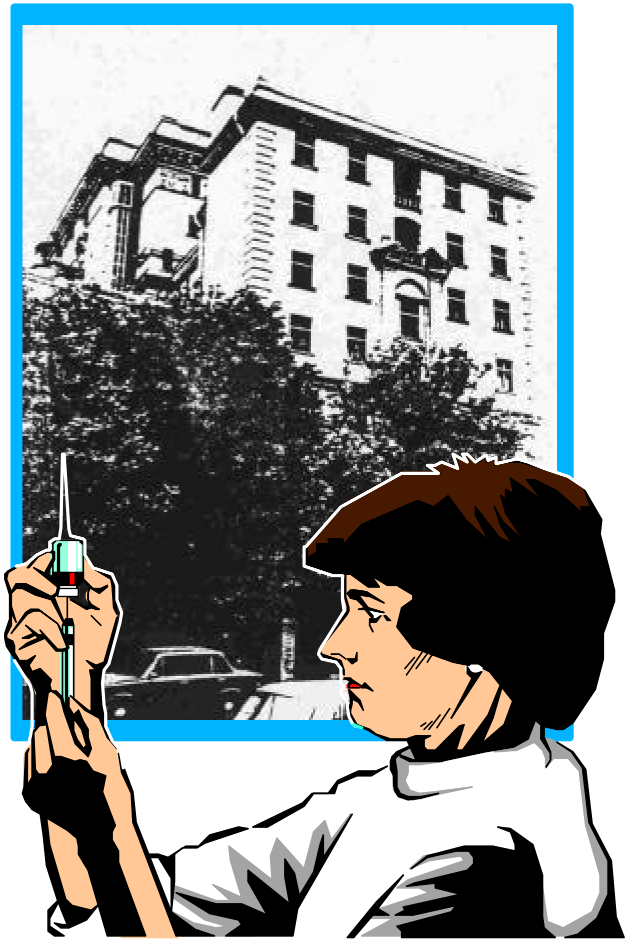

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Ramifications of the microwave radiation of the U.S. Embassy

- The cellular disorder pattern in the EM radiation is carried by the field-free scalar potential.
- Since the disorder pattern is a "conglomerate", it must be carried by an internal EM structure existing inside the scalar potential itself.
- Whittaker decomposed the scalar potential into the appropriate internal longitudinal EM biwave structure in 1903*.

* E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355

The importance of time as energy, and the time-domain

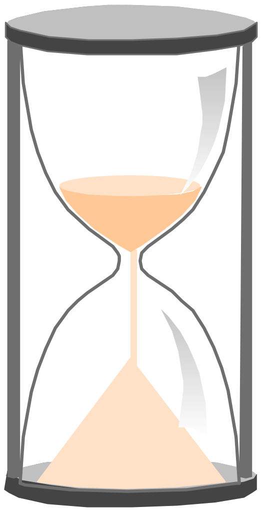

Time as energy

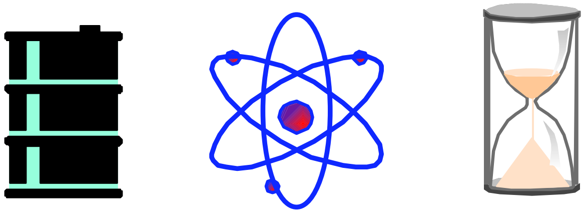

- In physics, the fundamental units one uses in his model are arbitrary.
  - A valid model can be made using only one unit.
  - It is already done in one branch of physics, using length as the single fundamental unit.
- Make a model with a single unit, the joule.
  - All other entities become functions of the joule.
  - We are familiar with mass as highly compressed EM energy, by the factor c^2.
  - It turns out that time is also highly compressed EM energy, by the factor c^2.
  - 1 sec. $\approx 9 \times 10^{16}$ joules.

How to see that "time is compressed spatial energy"

- Take some spatial EM energy and compress it by c^2
- Place it in 3-space, and it is *mass*.
- Place it on the time axis, and it is *time*.

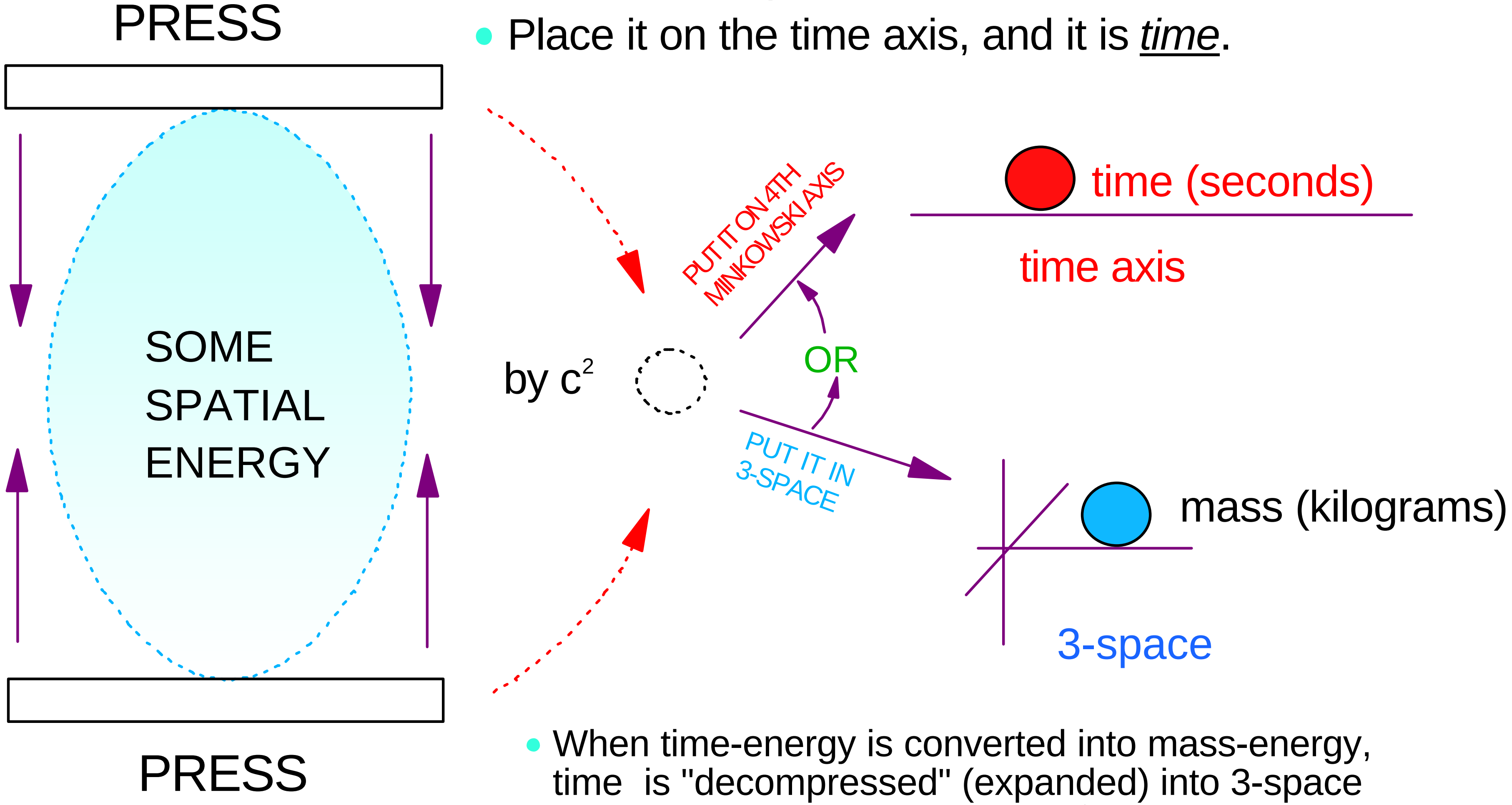
PRESS
SOME
SPATIAL
ENERGY
PRESS
by c^2
PUT IT ON 4TH
MINKOWSKI AXIS
OR
PUT IT IN
3-SPACE
time (seconds)
time axis
mass (kilograms)
3-space

- When time-energy is converted into mass-energy, time is "decompressed" (expanded) into 3-space EM energy and 1 sec. $\approx 9 \times 10^{16}$ joules.

Time as EM energy in the time domain, with a density the same as mass-energy

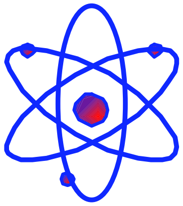

- Bearden* reinterpreted (corrected) the 1903 Whittaker decomposition of the potential in 2000.
- The reinterpretation is consistent with broken symmetry in particle physics, for which Lee and Yang received the Nobel Prize in 1957.
- It is also powerfully supported by quantum field theory, notably by Mandl and Shaw in 1984**.
- Whittaker's pair is a combined time-polarized (scalar) EM wave and a 3-space longitudinal EM wave.

* T. E. "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

** F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

All EM energy in 3-space freely comes from the time domain

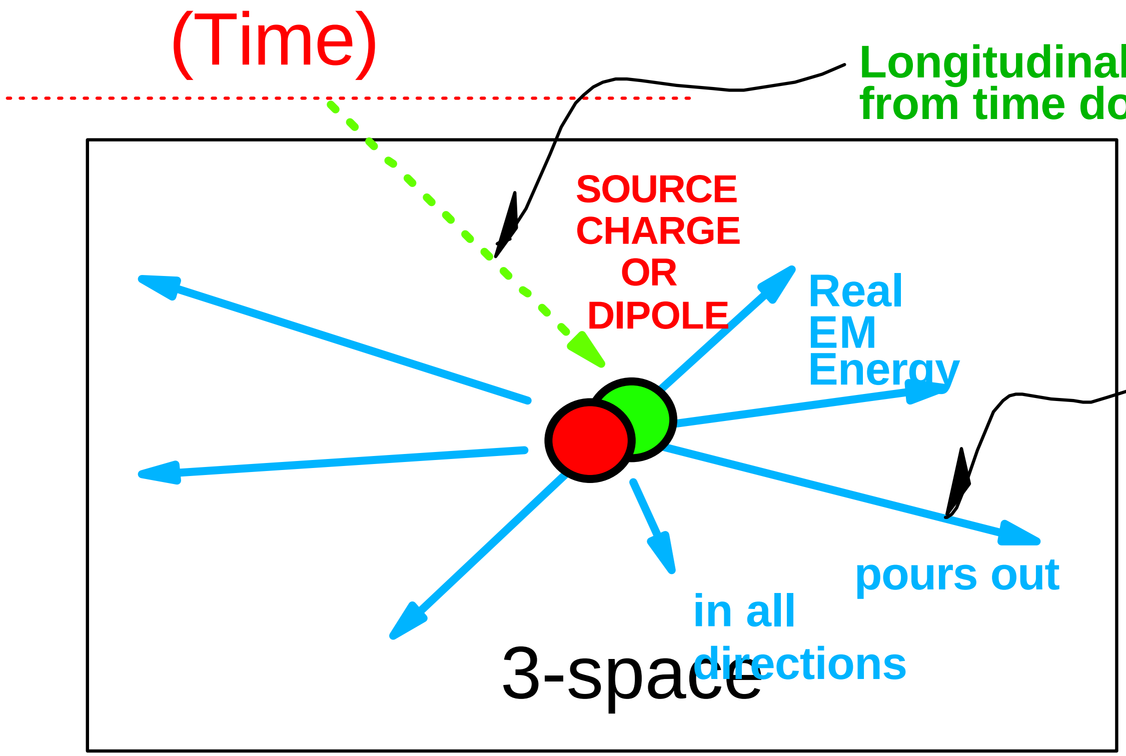

(Time)

SOURCE CHARGE OR DIPOLE

Real EM Energy

pours out

in all directions

3-space

Longitudinal EM wave incoming from time domain and absorbed

Charge's 720º spin transforms time-energy into 3-space energy

Longitudinal EM waves emitted in all directions in 3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur was just repeated since then, until corrected by Bearden to be consistent with broken symmetry and quantum field theory.

Strong support from quantum field theory

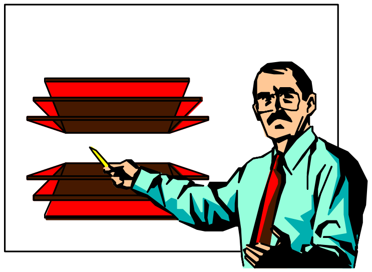

- Quantum field theory recognizes four photon polarizations:
  - Energy along x-, y-, z-, or t- axis.
  - Polarized along x- and/or y-axis, the photon is a *transverse* photon.
  - Polarized along z-axis (along its line of motion in 3-space), the photon is a *longitudinal* photon.
  - Polarized along the time-axis, the photon is a *time-polarized* or *scalar* photon.
- The individual longitudinal or scalar photon is not observable*.
- The combination of a longitudinal photon and a scalar photon is observable as the instantaneous scalar potential*.

* F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

Mechanism generating the flow of a mass through time

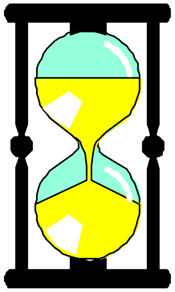

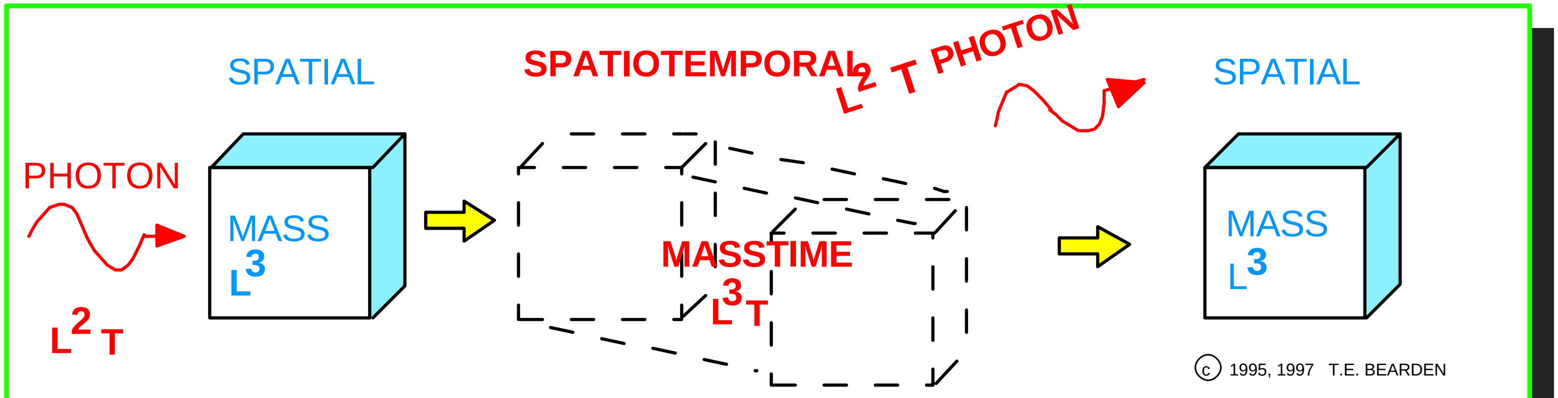
SPATIAL
SPATIOTEMPORAL
L^2 T PHOTON
SPATIAL
PHOTON
MASS
L^3
MASSTIME
L^3T
MASS
L^3
L^2 T

© 1995, 1997 T.E. BEARDEN

$$(\Delta E)(\Delta t) + M => (M+\Delta M)\Delta t => M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

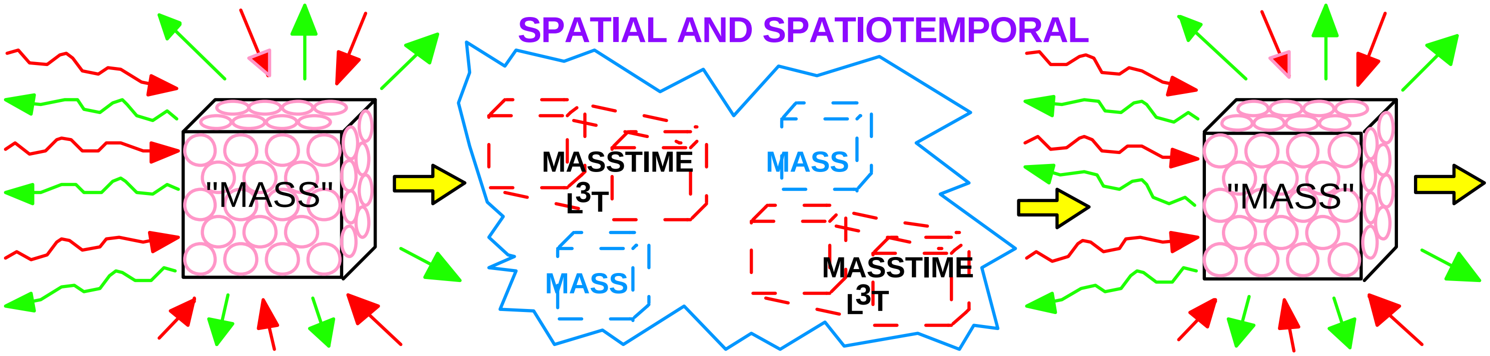
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

Mechanism is engineerable

Pumping a mass in the time-domain will time-reverse it back to an earlier state

- The reinterpreted biwave EM structure of the scalar potential means that time-reversal (phase conjugation by pumping in the time-domain) can be utilized to time-reverse a targeted cell (its mass-energy) in its entirety, including its genetics.
- Becker's bone fracture experiments showed it:
  - Weak scalar potential across the fracture site.
  - LWs in 3-space and time-energy waves in time
  - Red blood cells entered, dedifferentiated by shucking their hemoglobin and growing a nucleus, which is dedifferentiation (time-reversal of the cell and its mass-energy back to a previous physical state).

Principle for time-reversing a time-pumped mass

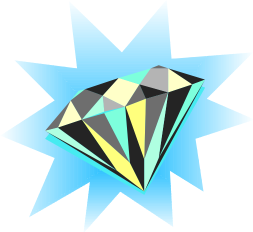

- Becker's cells then redifferentiated and changed -- first to primitive cartilage cells, then to primitive bone cells which were deposited in the fracture, healing it with new bone growth.
- This shows "time-forwarding" (redifferentiation) of the cell also.
- Becker actually showed that the *time-domain pumping* of the cells (by the time-polarized EM waves of Whittaker's reinterpreted wavepairs comprising the scalar potential), and the time-pumping of their mass-energy, operated to continuously and steadily *remove the delta between that cell's present condition and its former "normal" cellular physical condition in that area*.
- This is the long-sought secret of stem-cell research.
- Phase conjugate optics had not been born at the time. Without realizing its full nature, Becker demonstrated the body's own internal mechanism for healing, and the fundamental mechanism for time-reversing a physical mass, whether living or inert.

Mechanisms for use in medical treatment and healing

And ramifications

Successive deep penetration reactions of the cellular dielectric

- Change after change affects the cell and all of its parts, including from the external environment.
- Each successive change affects primarily the internal "LW EM structures and dynamics" comprising the ordinary potentials, fields, and waves.
- Thus in positive time the internal EM structurings produce successive "overlaid layers of change" -- so to speak -- in the internal EM structuring and dynamics of a cell in the body. These layers contain the "delta" from normal state.

Successive deep penetration reactions of the cellular dielectric (2)

- During time reversal, highest layer deltas are successively "peeled away" by the interaction of the mass-energy with the time-domain pumping.
- The result is to return the cell directly back to successive physical states, in the case of a single disease or disorder.
  - A cancer cell is returned to a normal cell prior to its "promotion".
  - An HIV-infected cell would be returned to its normal unaffected condition, free of HIV-induced genetic change.
  - For injuries and damage, the existing "delta" in each successive state-layer is steadily reduced to zero.
- This process thus can involve both cellular differentiation and dedifferentiation, as shown by Becker and by Prioré.

Engines (vacuum engines, spacetime curvature engines)

- Following Whittaker* as reinterpreted**, all normal EM fields, potentials, and waves decompose into sets of longitudinal EM waves (LWs) and their dynamics.
- Since these LWs and their dynamics are dynamics in the local energy density of spacetime (ST), they constitute *ST curvatures*.
- Hence they are general relativity infolded inside electrodynamics.
- By manipulating and changing these LWs, one manipulates and changes spacetime curvatures and their dynamics.
- The result is a unified field model where one directly controls and engineers a set of ST curvatures by use of higher group symmetry electrodynamics.
- Controlled patterns of local curvatures of ST are called "vacuum engines" or "spacetime curvature engines" -- or just "engines".

* E.T. Whittaker, Math. Ann., 57, 333-355 (1903); Proc. Lond. Math. Soc., Series 2, 1, 1367-372 (1904).

** T. E. Bearden, J. New Energy, 5(1), 11-23 (2000).

Pumping with EM time-polarized waves forms *amplified* antiengines

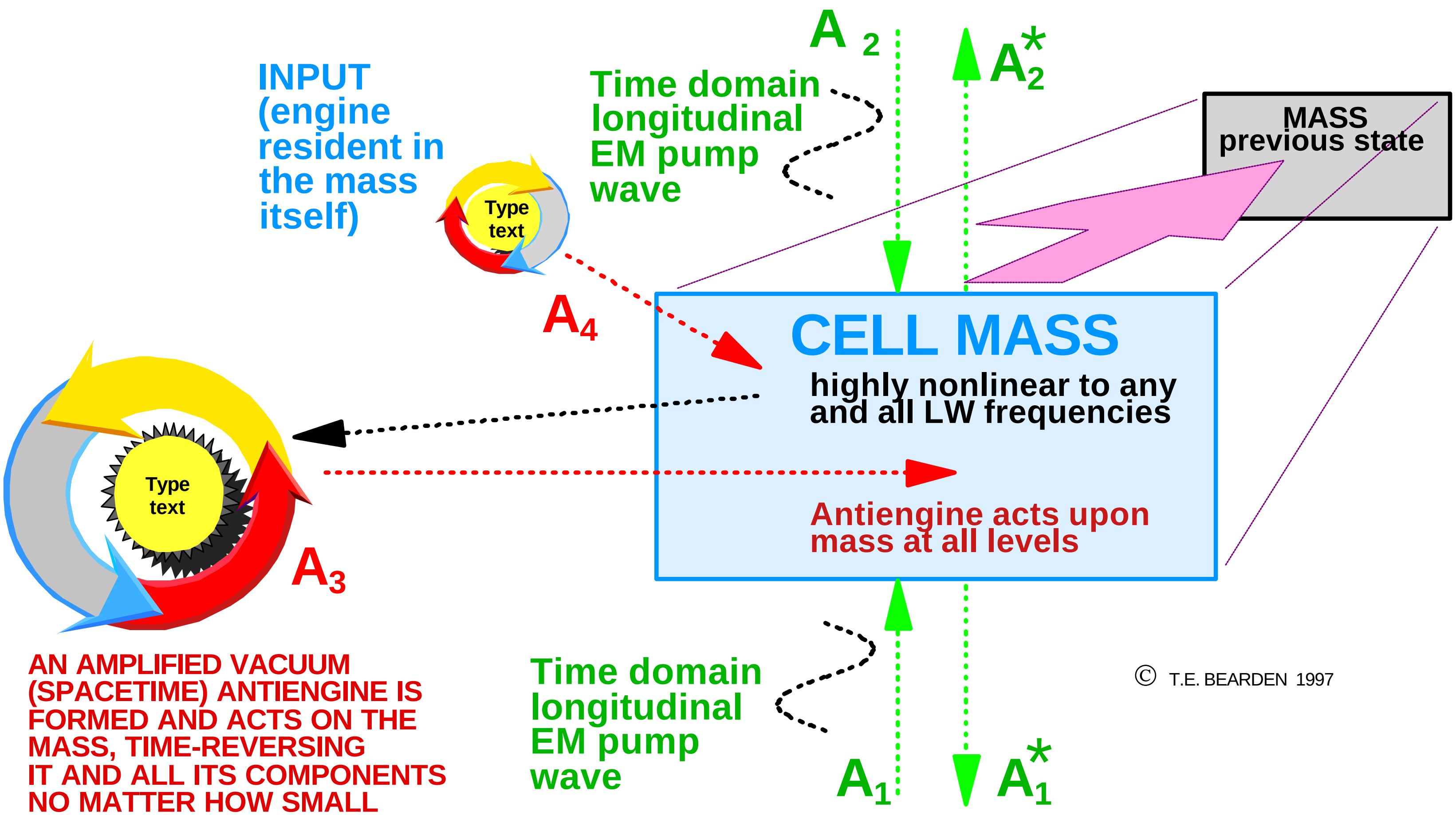
A_2
A_2^*
MASS previous state
Type text
A_4
A_3
Type text
A_1
A_1^*

INPUT (engine resident in the mass itself)

Time domain longitudinal EM pump wave

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

Time domain longitudinal EM pump wave

© T.E. BEARDEN 1997

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

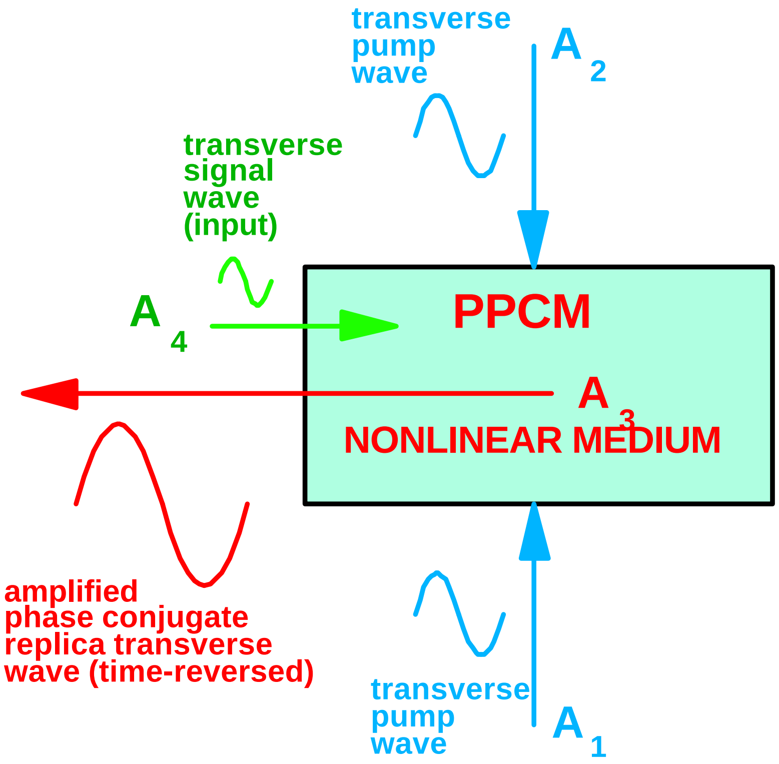

transverse pump wave

A_2

transverse signal wave (input)

A_4

PPCM

A_3

NONLINEAR MEDIUM

amplified phase conjugate replica transverse wave (time-reversed)

transverse pump wave

A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

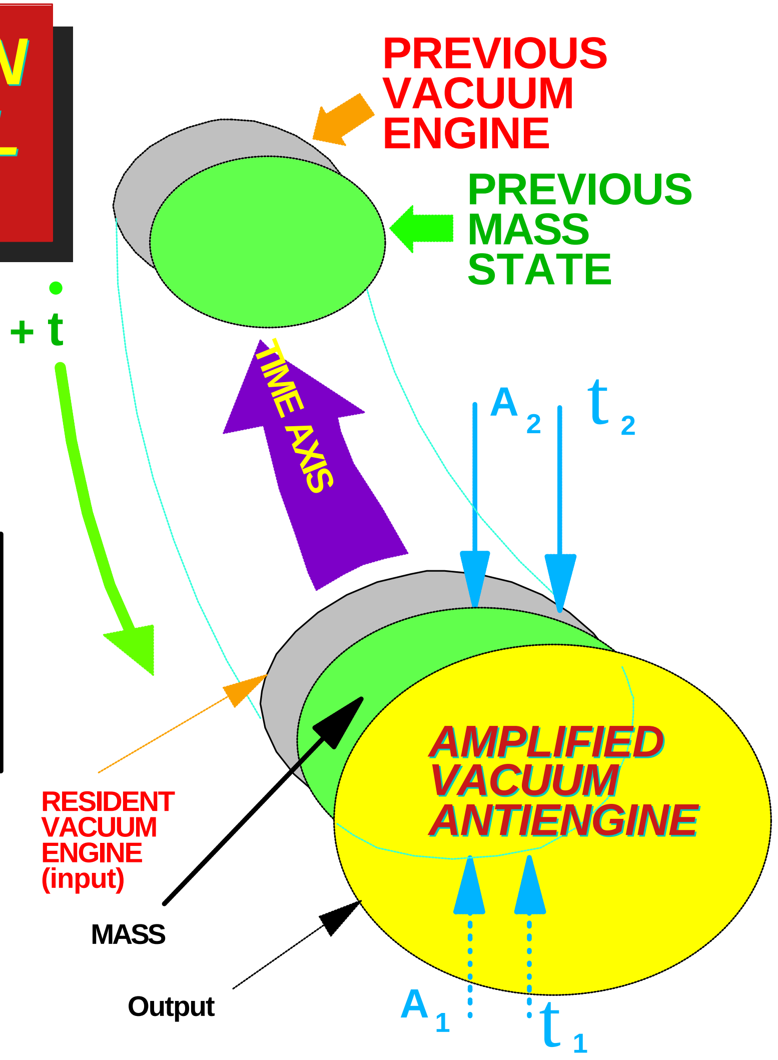

PREVIOUS VACUUM ENGINE

PREVIOUS MASS STATE

+ $\dot{t}$

TIME AXIS

A_2 t_2

RESIDENT VACUUM ENGINE (input)

AMPLIFIED VACUUM ANTIENGINE

MASS

Output

A_1 t_1

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Time-reversal of a mass to a new physical state not in its past

- Original engine A'
- New engine B' added
- Time-pumping moves to new "past" that is determined by (A'+B')

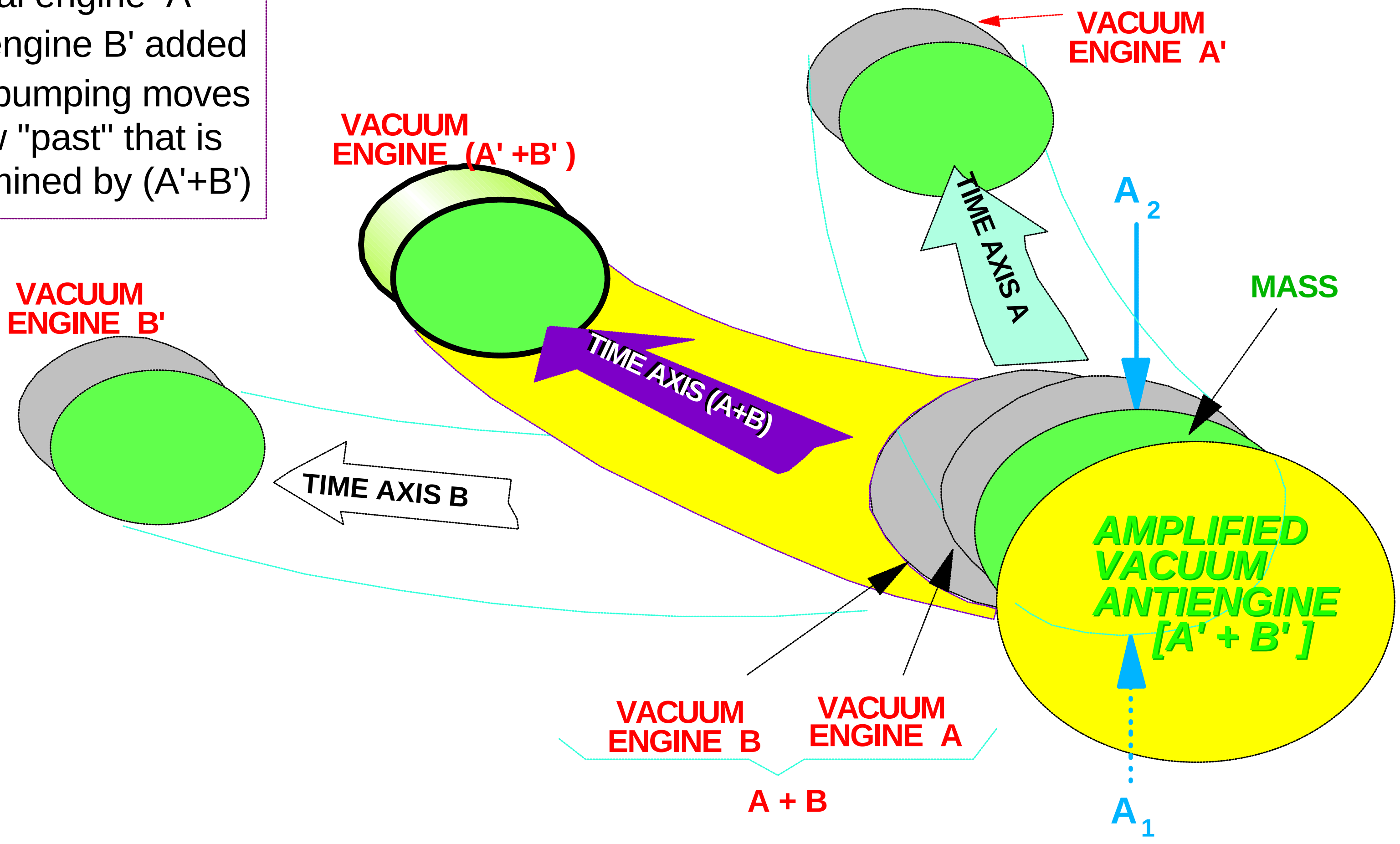
VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

Extended general relativity principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon any exposed mass at al levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

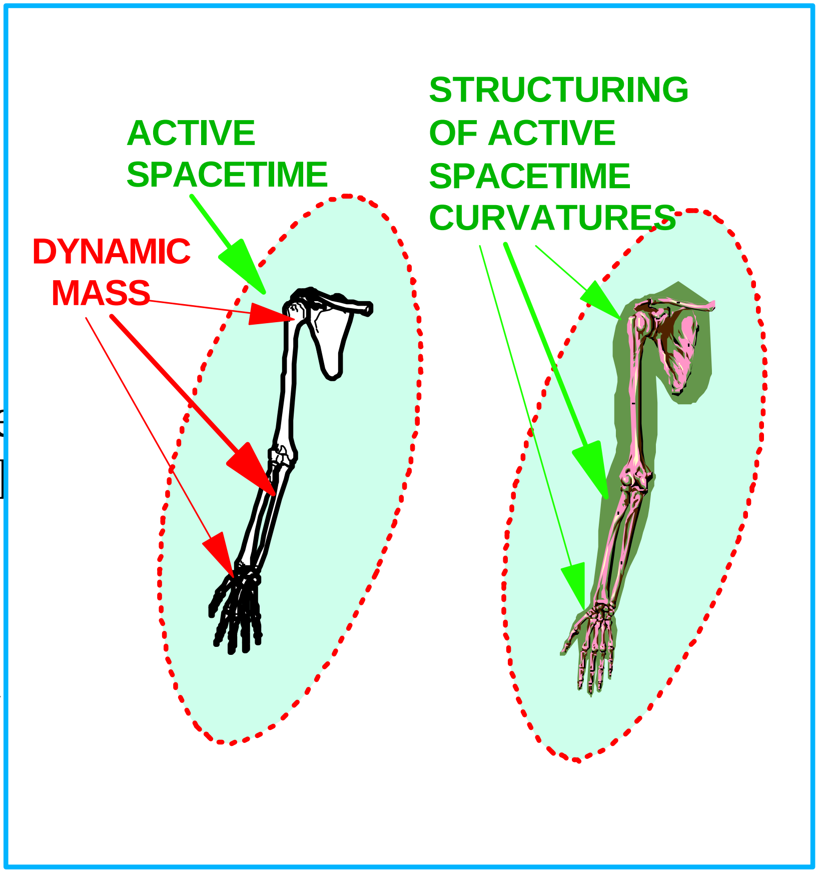
ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

Engines: Working spacetime demons

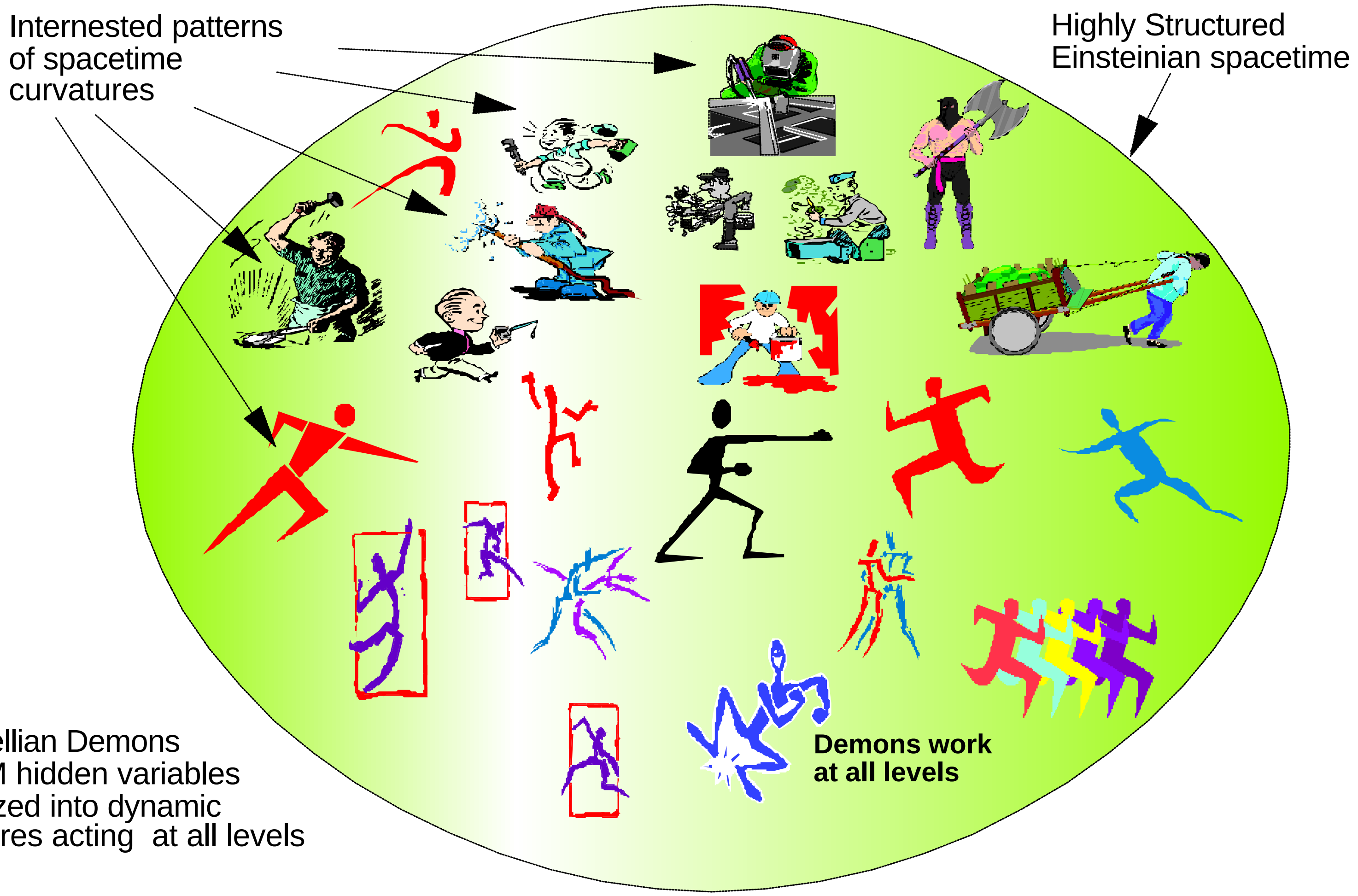

Internested patterns of spacetime curvatures

Highly Structured Einsteinian spacetime

Demons work at all levels

Maxwellian Demons are EM hidden variables organized into dynamic structures acting at all levels

Signal vs. vacuum engine

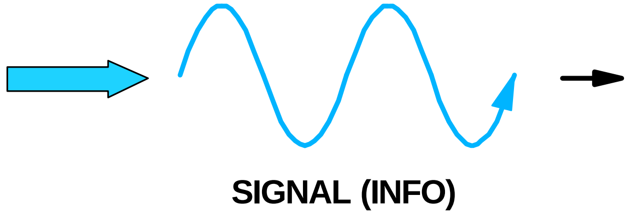

★ Information received
★ Any overt physical action must be taken by cell itself
★ Cell must furnish the energy for any action it takes
★ Passes from outside to in

A. Cell must do the action itself; vacuum energy exchange is passive. Energy or fuel for doing the action must be added to cell externally.

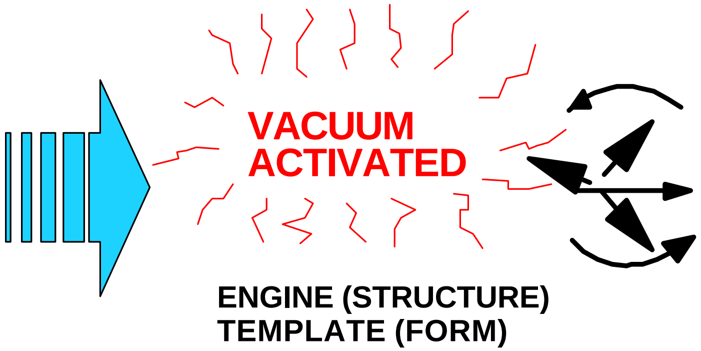

★ Information itself acts
★ Local spacetime curved, with internal subcurvatures
★ Produces template of forces in local spacetime (ST)
★ Arises from within local ST
★ Acts on and changes all parts of cell, including its genetics

B. Cell is acted upon; vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to cell externally.

Unified field theory requires modeling and use of supersystem

Three components interact with each other:

- System and its dynamics
- Local active vacuum and its dynamics
- Local curvatures of spacetime and their dynamics.

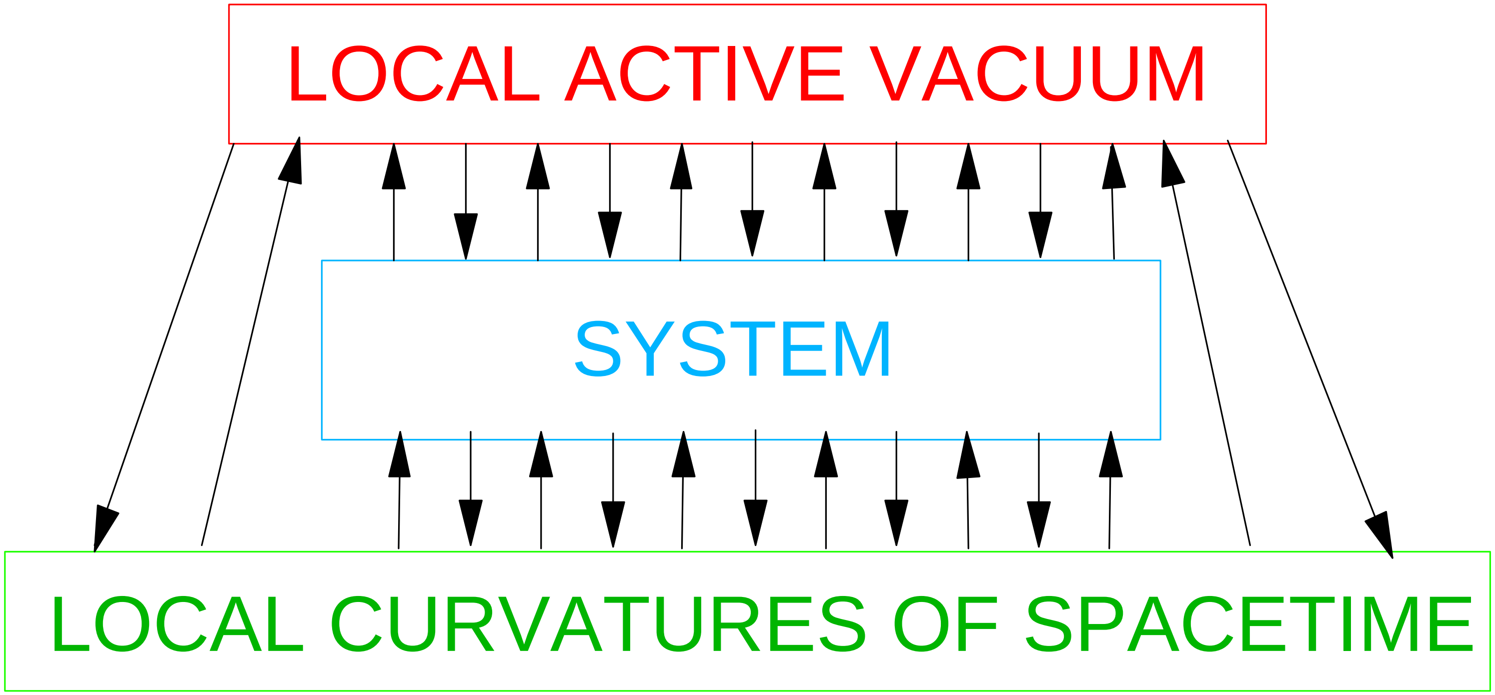

LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Classical EM kills the supersystem, prevents supersystem engineering

- Assumes flat local spacetime
- Assumes inert vacuum

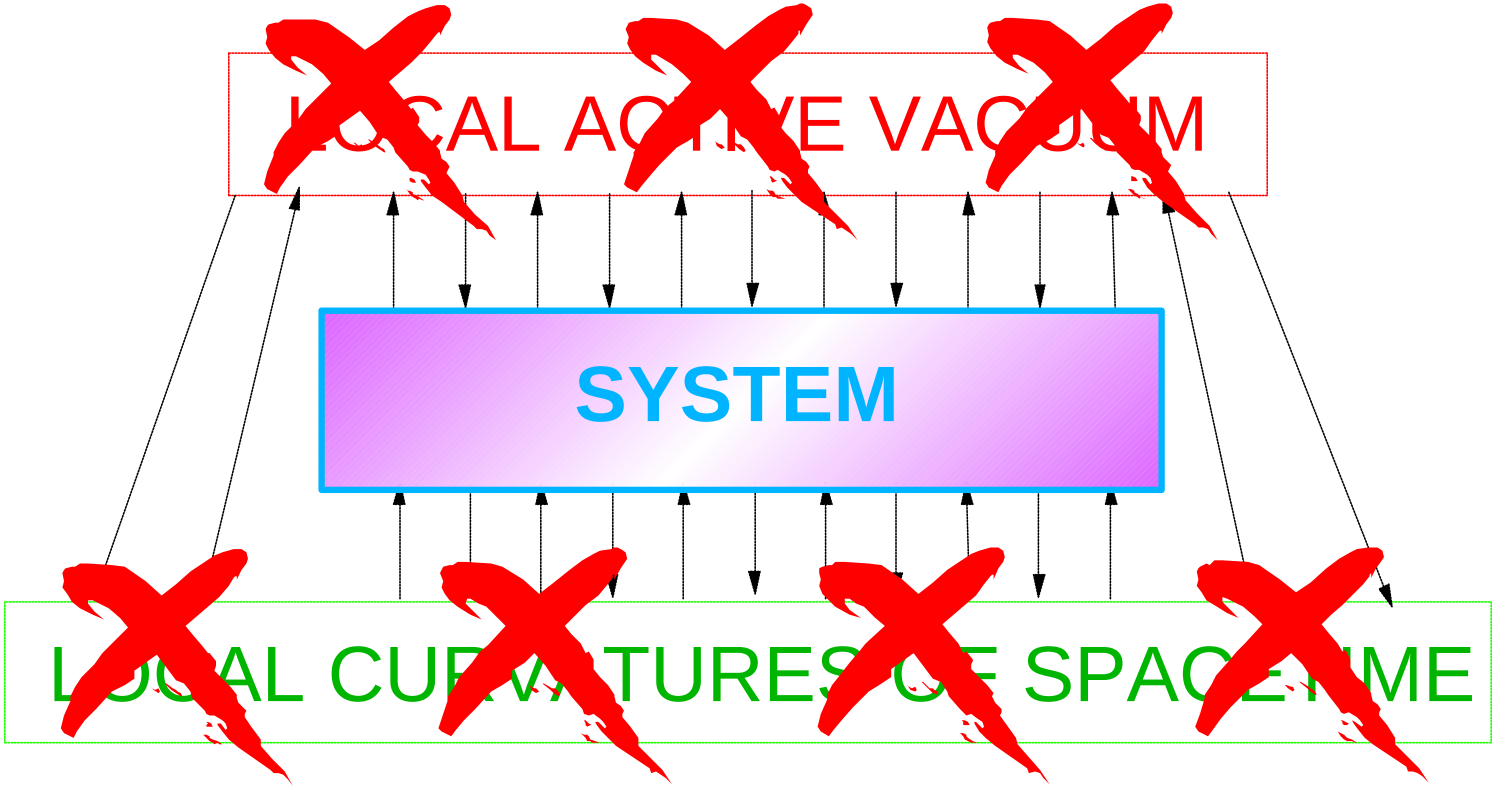

LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Malignant teratoma exhibits both time-reversal and fast-forwarding

- A teratoma cell is a tumor cell that has reverted to stem cells, then proceeds to differentiate toward new kinds of normal cells.
- May produce hair, skin, muscle, heart, bone, and other type cells.
- Sometimes present at birth.
- Demonstrates both "time-reversal" and "time-forwarding".
- The missing key is the engine function(s).

Miller-Fox-Urey biogenesis experiments

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

Biogenesis Experiments

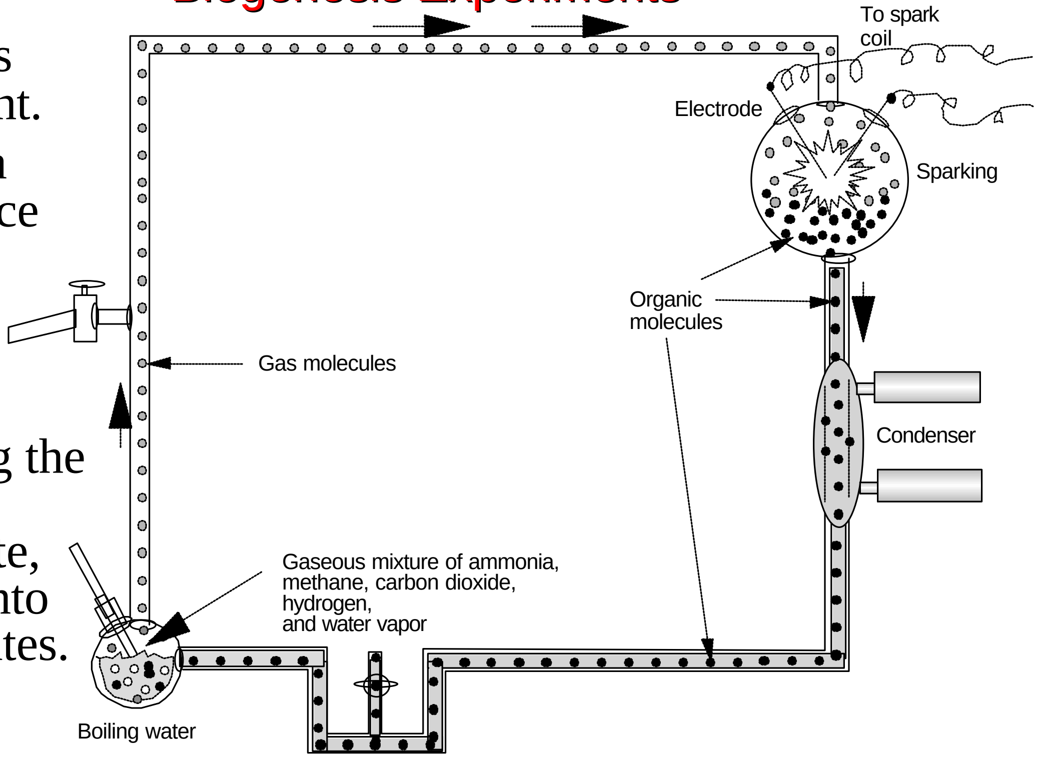
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Cloning to produce stem cells

- Human DNA can be inserted in an animal cell, and the cell chemically tricked into proceeding as if fertilized.
- Formation, increase, and growth of stem cells begins and continues.

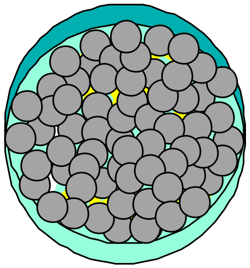

- The "engine" carried by the human DNA is changing the matter available into human cells.
- readily demonstrated experimentally.

Regeneration in a strain of rat with part of its immune system missing

- In mammals, the immune system plays a role in reducing the regeneration capability, compared to
that of more primitive systems.
- In a strain of rat where that "suppressive" part of the immune system is missing, the regenerative capability is dramatically increased.
  - Can regrow cut away sections of tail.
  - Regrows plug cut out of ear.
  - Reconnects severed optic nerve.
  - Partially restores severed spinal cord.

- The suppressive "delta" engine of the immune system, operating simultaneously with the underlying primitive regenerative engine, sums to the modern much-reduced regenerative engine in the mammal.

An observation for the future

- When the science of engines is developed, physicians will be able to tailor these engine combinations to restore severed human spines, regrow severed limbs, etc.
- They will also be able to direct the body's "normal growth" to reduce and direct birth defects, genetic defects, etc.

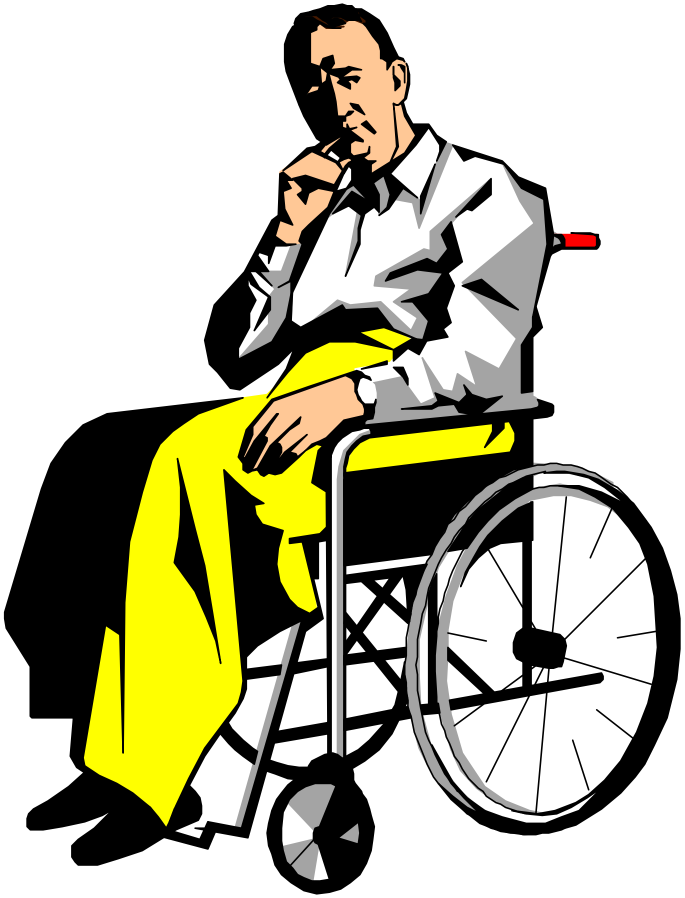

Biochemistry and EM models

- Medical scientists analyze such things as teratomas in terms of biochemistry and the biochemical model.
- An assembly of charges and charge distributions, with their associated fields and potentials, drives all chemistry at base level.
- The charge assembly and its dynamics involves a corresponding assembly of spacetime curvatures and their dynamics. These are *the cause that is driving the charge assembly and its dynamics, and therefore the chemistry*..
  - In unified field theory (such as that of Sachs) and O(3) electrodynamics.
  - Not in classical U(1) EM theory or quantum electrodynamics.
- Conventionally a causative EM field in space is "defined" as the *diversion from* that spatial field by and after its interaction with charge).
  The "definition" substitutes the effect for the cause, a non sequitur.
- This is a known major flaw, pointed out in one fashion or another by many physicists such as Wheeler and Nobelist Feynman.
- Classical EM theory is a 137-year-old model, seriously flawed. These flaws emerge in biochemistry and medical science.

Necessary EM corrections*

- Many of the corrections necessary in electrodynamics modeling either have been done or are underway.
- Higher symmetry, non-Abelian electrodynamics models have long been developed in particle physics.
- The new approach extends these models to also incorporate general relativity, thus producing a unified field theory*.
- In the new approach, spacetime curvature engines are perfectly natural and understandable, and can be directly modeled and engineered by advanced theorists.
- In short, the theory is sufficiently advanced and robust that technology and engineering can now advance.
- This will be a great leap forward for biochemistry and medical science, particularly in therapeutic methods.

* Sachs's unified field theory implemented by O(3) electrodynamics presently is the most suitable to use. Numerical methods are required.

Prioré's revolutionary results in lab animals

And its ramifications

Prioré (right) makes a point to Chaban-Delmas, Mayor of Bordeaux

- Many of the Prioré team's experiments were done in Bordeaux.
- Antoine Prioré patents:
  - "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968.
  - "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.
  - "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

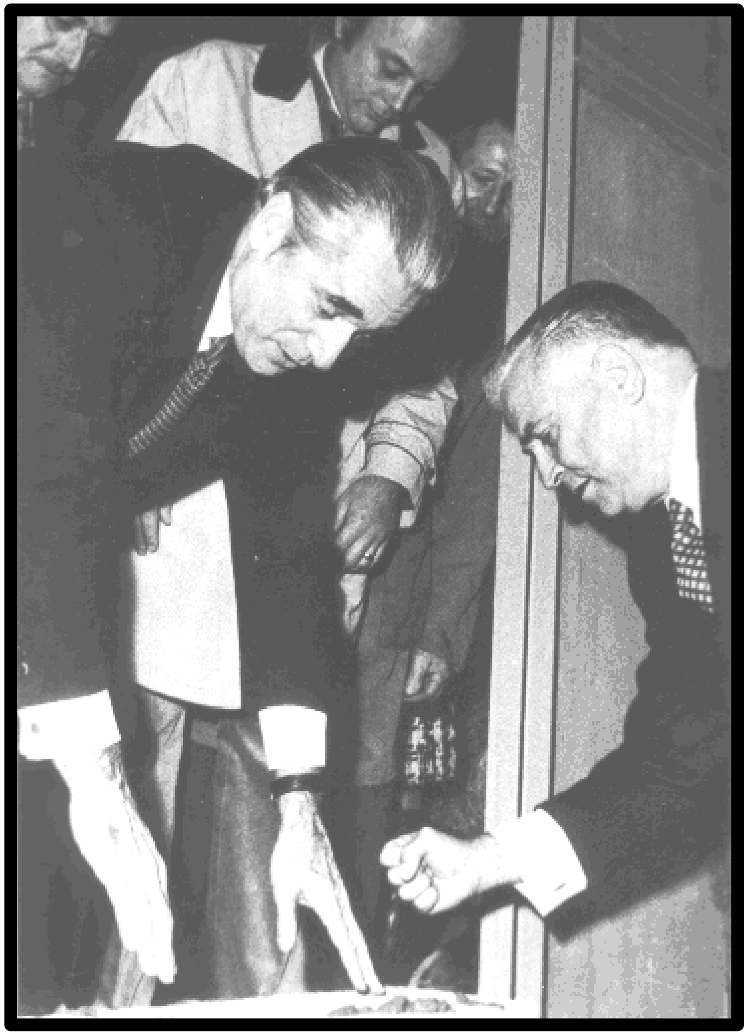

Block diagram of Prioré's method

Time-reversing the cells back to normal state

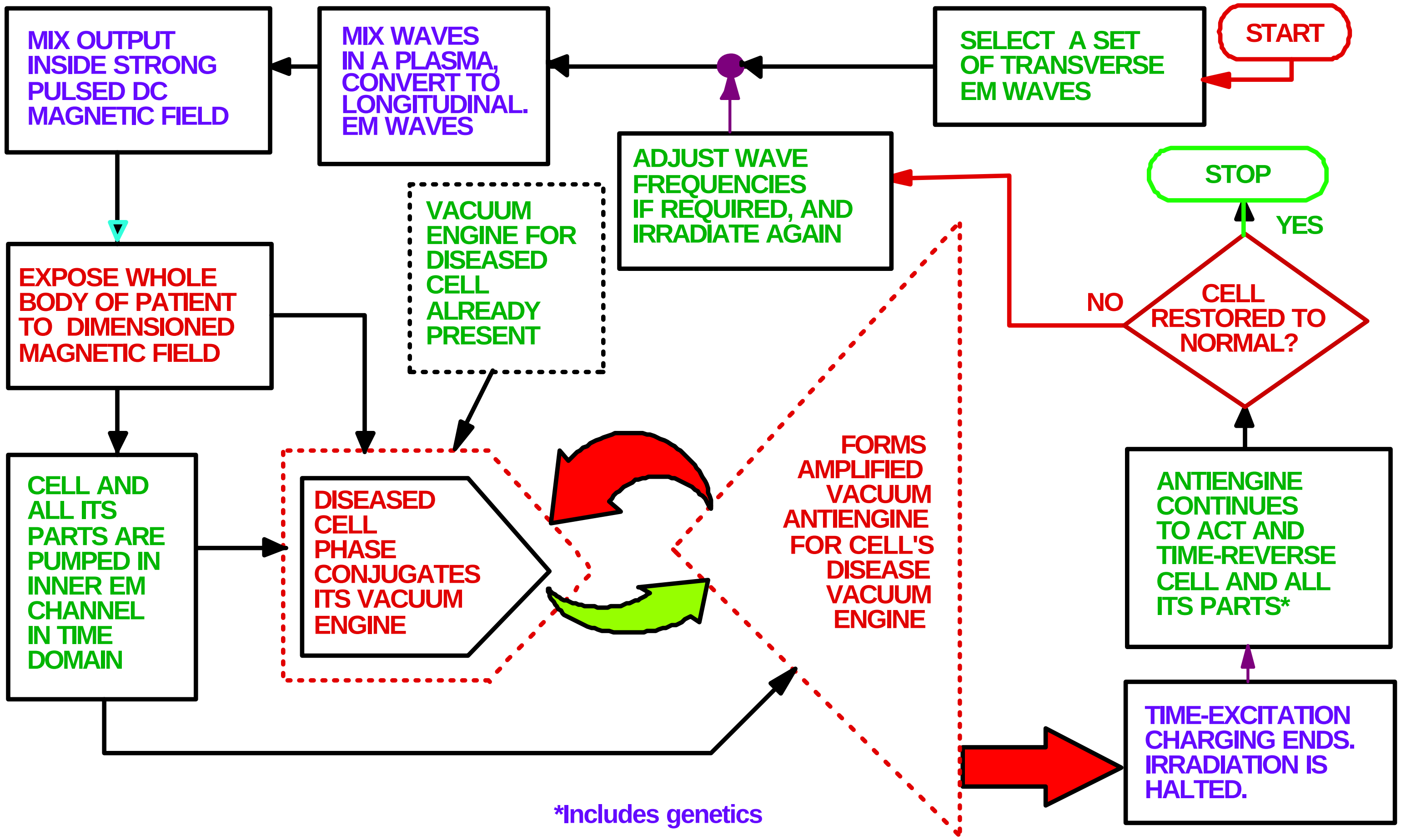

START

SELECT A SET
OF TRANSVERSE
EM WAVES

MIX WAVES
IN A PLASMA,
CONVERT TO
LONGITUDINAL.
EM WAVES

MIX OUTPUT
INSIDE STRONG
PULSED DC
MAGNETIC FIELD

EXPOSE WHOLE
BODY OF PATIENT
TO DIMENSIONED
MAGNETIC FIELD

CELL AND
ALL ITS
PARTS ARE
PUMPED IN
INNER EM
CHANNEL
IN TIME
DOMAIN

VACUUM
ENGINE FOR
DISEASED
CELL
ALREADY
PRESENT

DISEASED
CELL
PHASE
CONJUGATES
ITS VACUUM
ENGINE

FORMS
AMPLIFIED
VACUUM
ANTIENGINE
FOR CELL'S
DISEASE
VACUUM
ENGINE

TIME-EXCITATION
CHARGING ENDS.
IRRADIATION IS
HALTED.

ANTIENGINE
CONTINUES
TO ACT AND
TIME-REVERSE
CELL AND ALL
ITS PARTS*

CELL
RESTORED TO
NORMAL?

YES

STOP

NO

ADJUST WAVE
FREQUENCIES
IF REQUIRED, AND
IRRADIATE AGAIN

*Includes genetics

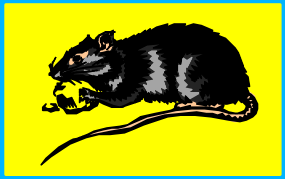

In Laboratory Animals:

Prioré Team Cured

- **Terminal malignant tumors**
- **Clogged arteries (atheriosclerosis)**
- **Suppressed Immune Systems**
- **Infectious trypanosomas**
- ***Results were presented to French Academy of Science by Robert Courrier, Director of the Biology Section of the Academy***
- ***No one understood the mechanism!***

© 1999 T. E. BEARDEN

Some peculiar results by the Prioré team

- Once a rat is cured of a particular disease:
  - A single drop of blood from the cured rat, injected into another rat stricken with same disorder, would result in the injected rat getting well slowly.
  - This ability to transfer immunity and healing in the extracted drop of blood gradually decayed over a period of a few weeks.
- Showed that even the fields and potentials in the blood had been altered internally, in the treated animal.
- Showed that the injected animal's cellular regenerative system possessed the ability to also time-domain pump over the entire body, everywhere reducing the delta between that injected drop of blood now the rat's own, and the rest of the blood in its body.

Proof of cellular time-reversal

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only the *previous ability* of the immune system is restored by the process.

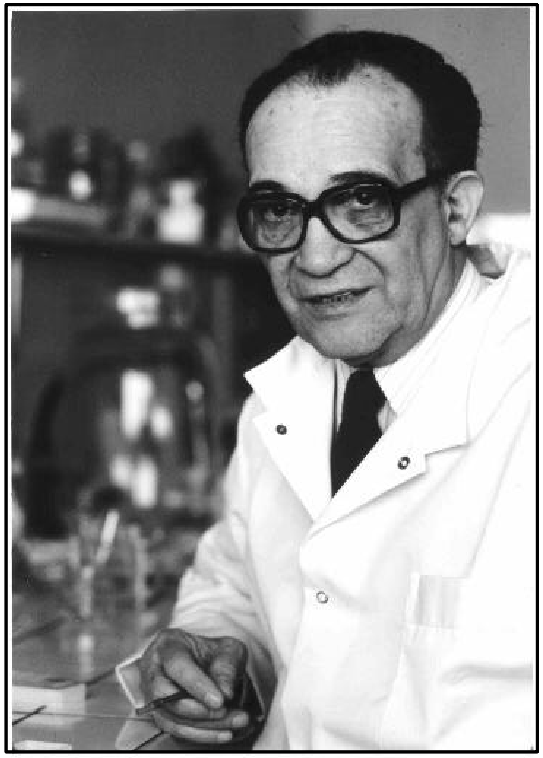

PROF. RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Disadvantages of Prioré method

- Large plasma tube and facility required.
  - 3-stories high building bay if to treat humans.
  - Long irradiation time (2-3 hours).
  - Slow, numerous experimental adjustments.
- Fundamental mechanism was unknown.
  - Phase conjugate optics not yet born in West.
  - No higher symmetry EM modeling used.
  - Prioré thought it was a result of ionization.
  - Importance of difference frequency unknown.
  - Body dielectric transmission path unknown.

3-Story high Prioré device for treating humans

- At the end of the project, Prioré had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient -- and also expensive

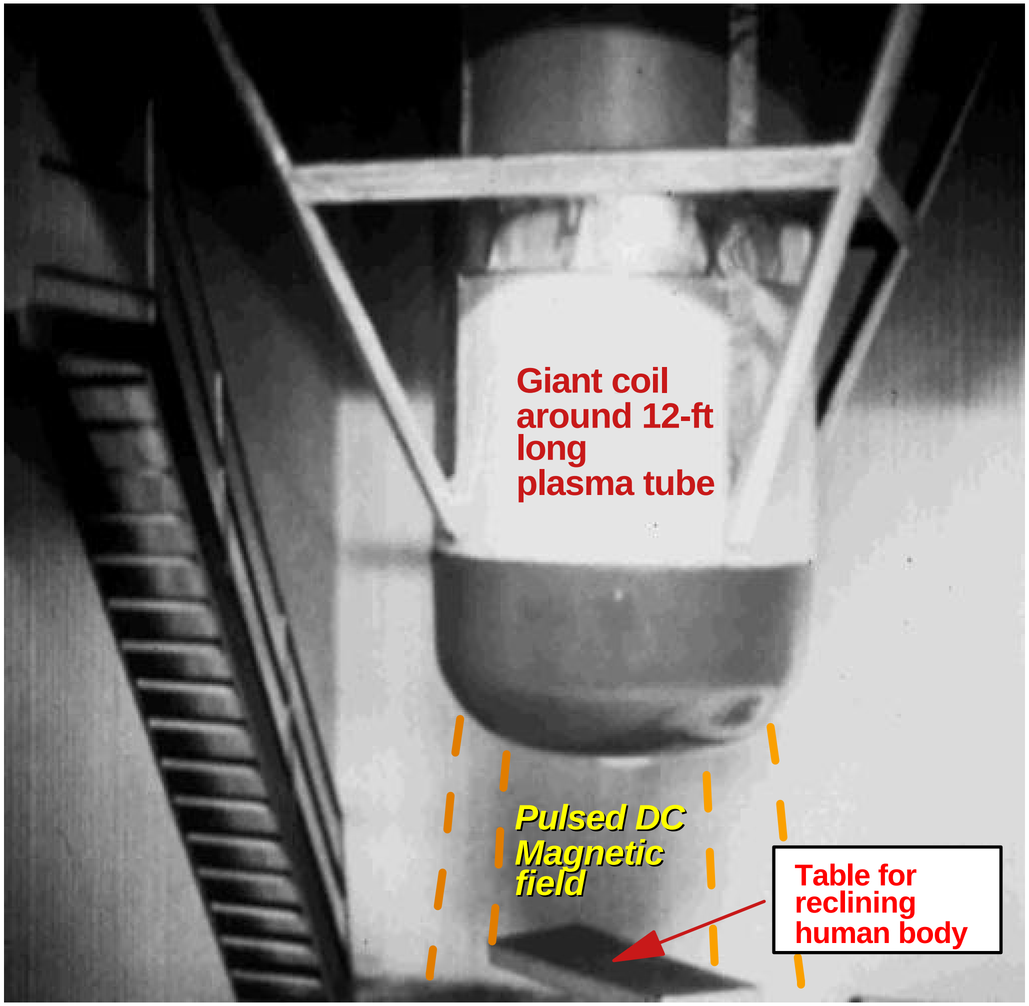
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

17-foot Prioré plasma tube

Used in large machine (3-stories high) to treat a limited number of human patients

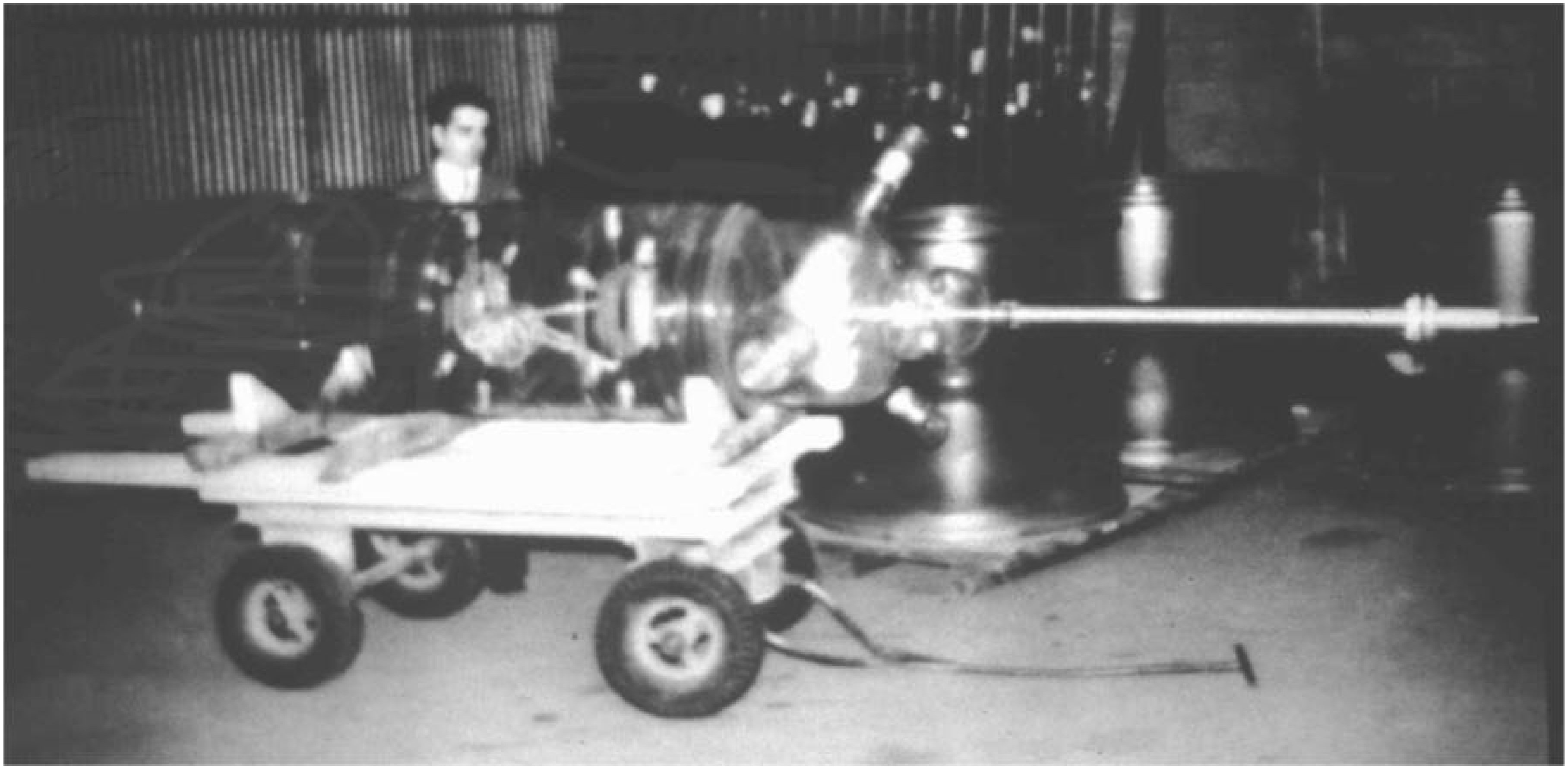

A proposed new mechanism for cancer promotion

And its ramifications

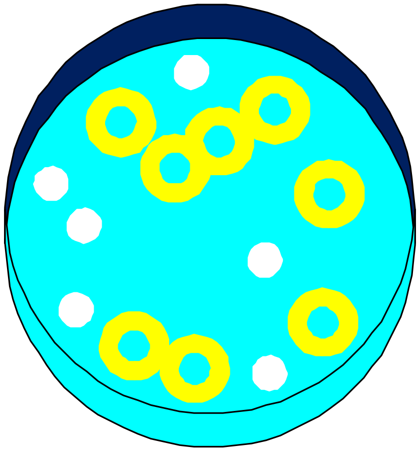

Remarks on cancer from a new perspective

- Two stages: cellular damage and promotion of the damaged cell to a tumorous cell. Promotion presently is poorly understood if at all.
- A master cellular control system (MCS) (studied by Popp et al.) controls both the immune system and the cellular regenerative system, as well as normal functions of the cells.
- The "past history" of ancestral development of aerobic cells and higher multicell aerobic bodies on planet Earth, resides in the "internested engine layers" in every tiniest part of the entire body dielectric, and in all cells.

Remarks on cancer from a new perspective (2)

- Faced with cell damage (such as sustained hypoxia) it cannot handle, the MCS continues to "peel the layers" in taking actions related to ever more distant pasts.
- Faced with continuing hypoxia, e.g., the MCS eventually reaches ancestral layers -- with an engine -- it forces the damaged cells to start moving back toward the original primeval anaerobic form.
- The first step backward is *separation from central growth control*, freeing the cell to multiply independently. At the same time, genetic changes in the cell are forced and produced, starting it back toward becoming an anaerobe.
- These actions comprise the *promotion process*.

The promotion process via engines

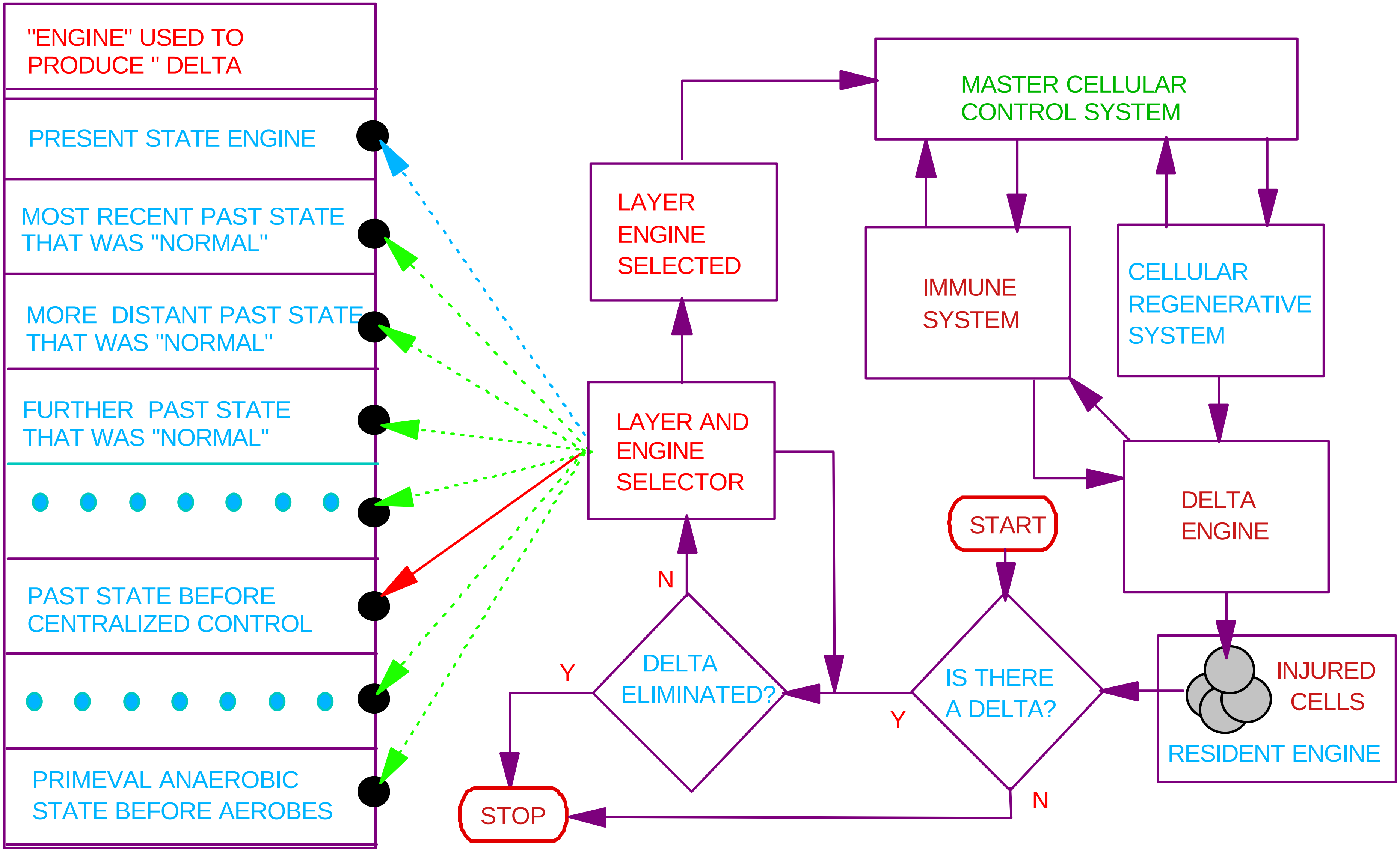

"ENGINE" USED TO PRODUCE " DELTA

PRESENT STATE ENGINE

MOST RECENT PAST STATE THAT WAS "NORMAL"

MORE DISTANT PAST STATE THAT WAS "NORMAL"

FURTHER PAST STATE THAT WAS "NORMAL"

PAST STATE BEFORE CENTRALIZED CONTROL

PRIMEVAL ANAEROBIC STATE BEFORE AEROBES

LAYER ENGINE SELECTED

LAYER AND ENGINE SELECTOR

MASTER CELLULAR CONTROL SYSTEM

IMMUNE SYSTEM

CELLULAR REGENERATIVE SYSTEM

DELTA ENGINE

START

IS THERE A DELTA?

Y

N

DELTA ELIMINATED?

N

Y

STOP

INJURED CELLS

RESIDENT ENGINE

Evolution of aerobic systems

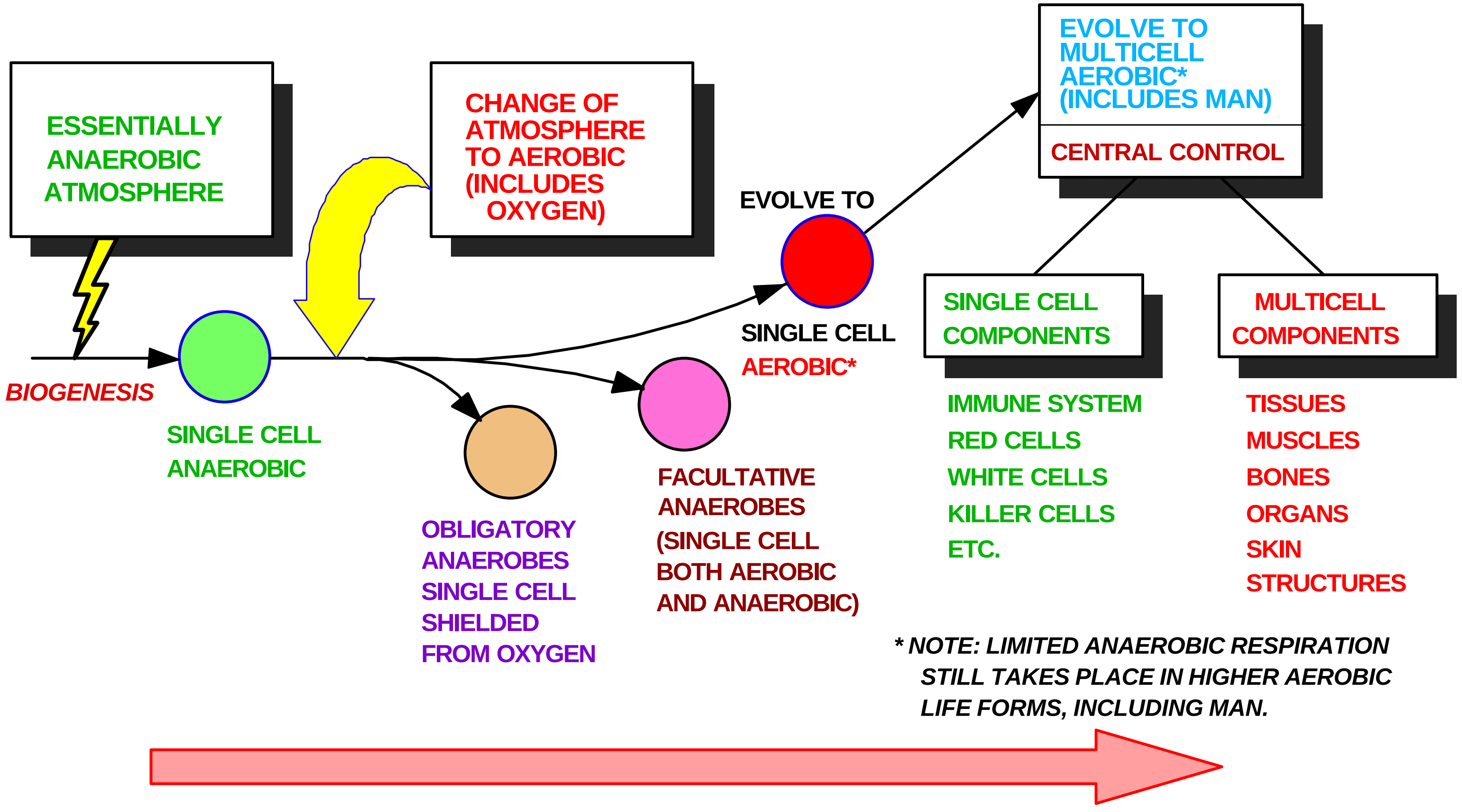

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

First step in self-promotion of cancer due to hypoxia

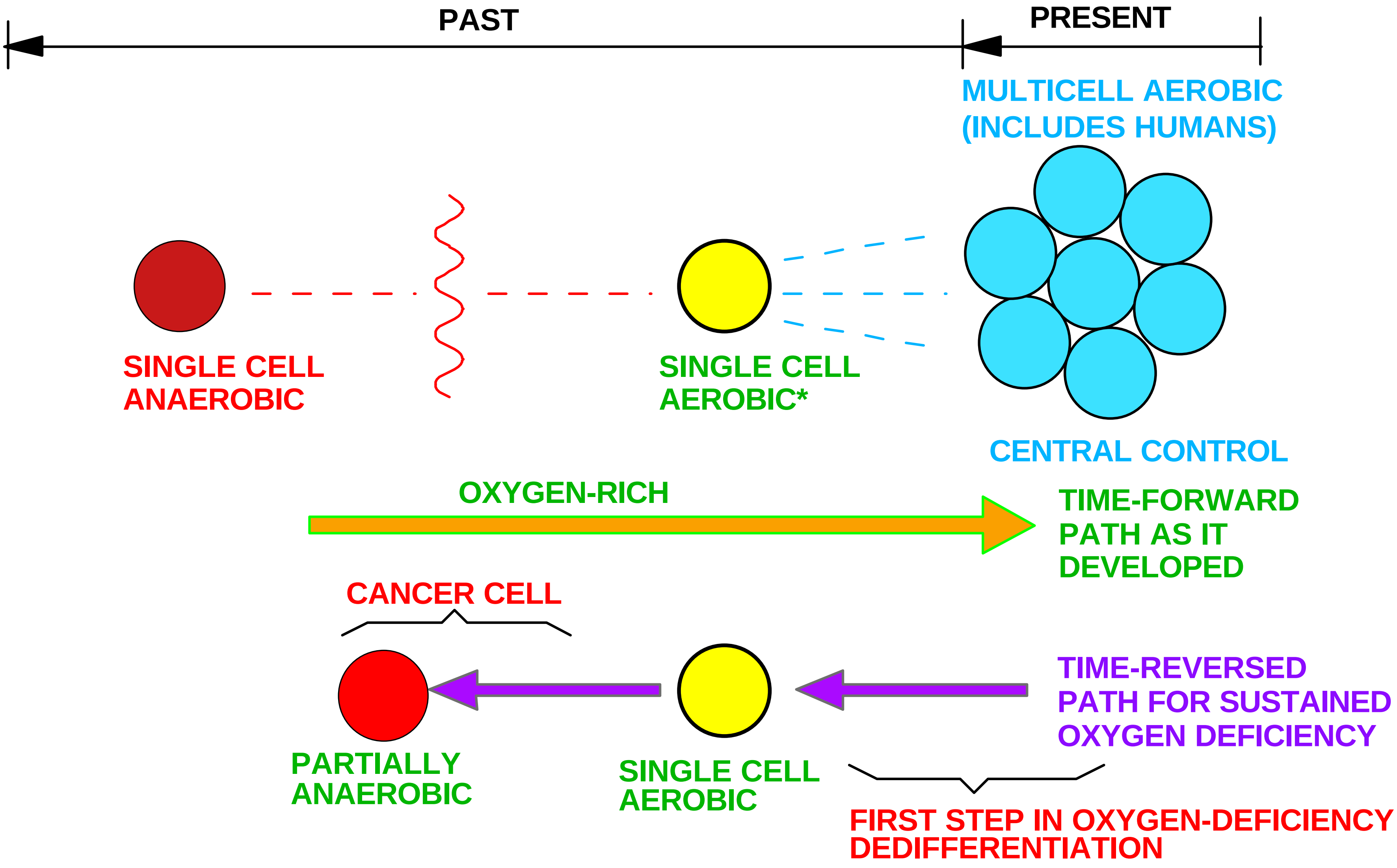

PAST

PRESENT

MULTICELL AEROBIC
(INCLUDES HUMANS)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

Block diagram of the cellular regenerative system

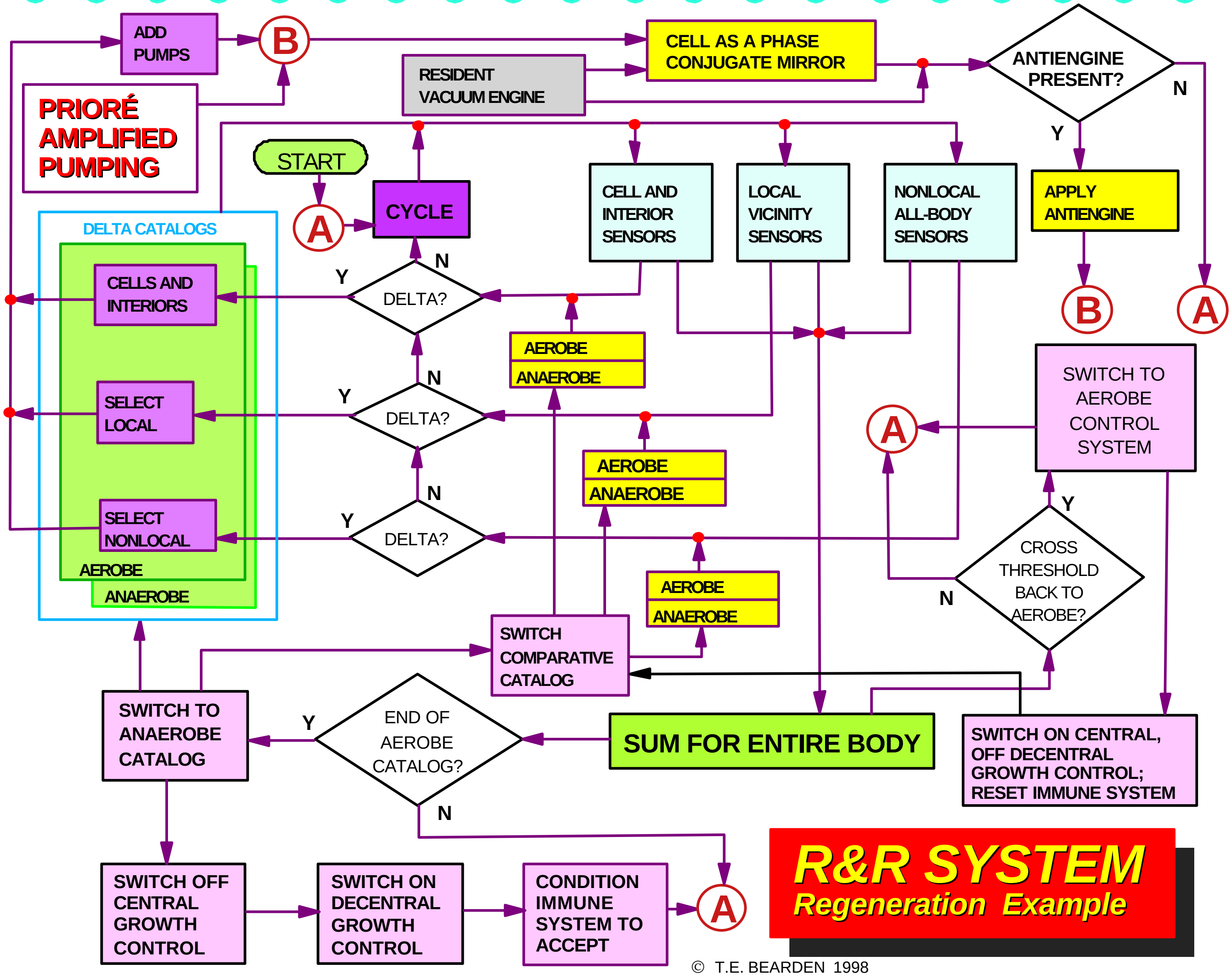

ADD PUMPS

B

PRIORÉ AMPLIFIED PUMPING

RESIDENT VACUUM ENGINE

CELL AS A PHASE CONJUGATE MIRROR

ANTIENGINE PRESENT?

N

Y

APPLY ANTIENGINE

B

A

START

A

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

DELTA CATALOGS

CELLS AND INTERIORS

SELECT LOCAL

SELECT NONLOCAL

AEROBE

ANAEROBE

Y

DELTA?

N

AEROBE

ANAEROBE

Y

DELTA?

N

AEROBE

ANAEROBE

Y

DELTA?

N

AEROBE

ANAEROBE

SWITCH TO AEROBE CONTROL SYSTEM

A

CROSS THRESHOLD BACK TO AEROBE?

Y

N

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

Y

END OF AEROBE CATALOG?

N

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Bombshell: Amplified EM healing

- A unified field theory approach and use of supersystem analysis has broken the basic mechanism for healing itself.
  - The Sachs unified field theory applies and has been partially fitted.
  - Evans-Vigier O(3) electrodynamics can properly model it, including both the internal and external electrodynamics and effects of engines.
  - Direct engineering development is now possible.
- Experimental proof of the results obtainable already exists in the literature.
- A medical revolution is in the offing.

Potential Medical Applications

- Non-traumatic cancer treatment
  - *Cure cancers and leukemias of all types*
  - *Remove pre-cancerous state*
  - *Restore system*
- Treat and cure Infectious diseases
- Cure AIDS and restore immune systems
- Defense against biological warfare
  - *Virulent diseases from unknown agents*
  - *Treat, cure prior to symptoms developing*
  - *Both civil populace and armed forces*
- *Dramatically lower health costs*
- *Rejuvenate the aged*

© 1999 T. E. BEARDEN

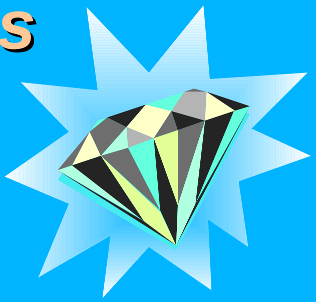

To treat and save mass casualties: What is urgently needed now:

- A small, portable, relatively inexpensive treatment unit, not requiring plasmas etc.
- Treatment in 2 to 5 minutes rather than 2-hours.
- Simple treatment application, almost anywhere.
- Effective in advanced stages as well as prior to emergence of symptoms.
- Effectiveness against a wide array of illnesses.
- Use is easy, high school student can be trained in 30 minutes.
- Crash program to develop, flood down through emergency systems in great quantities.

The Solution:

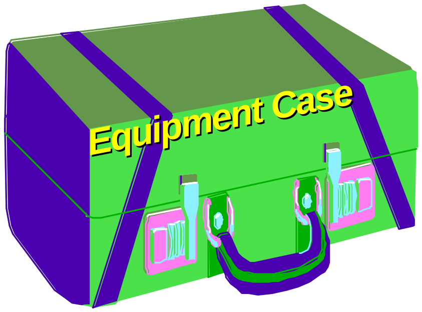
Equipment Case

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.

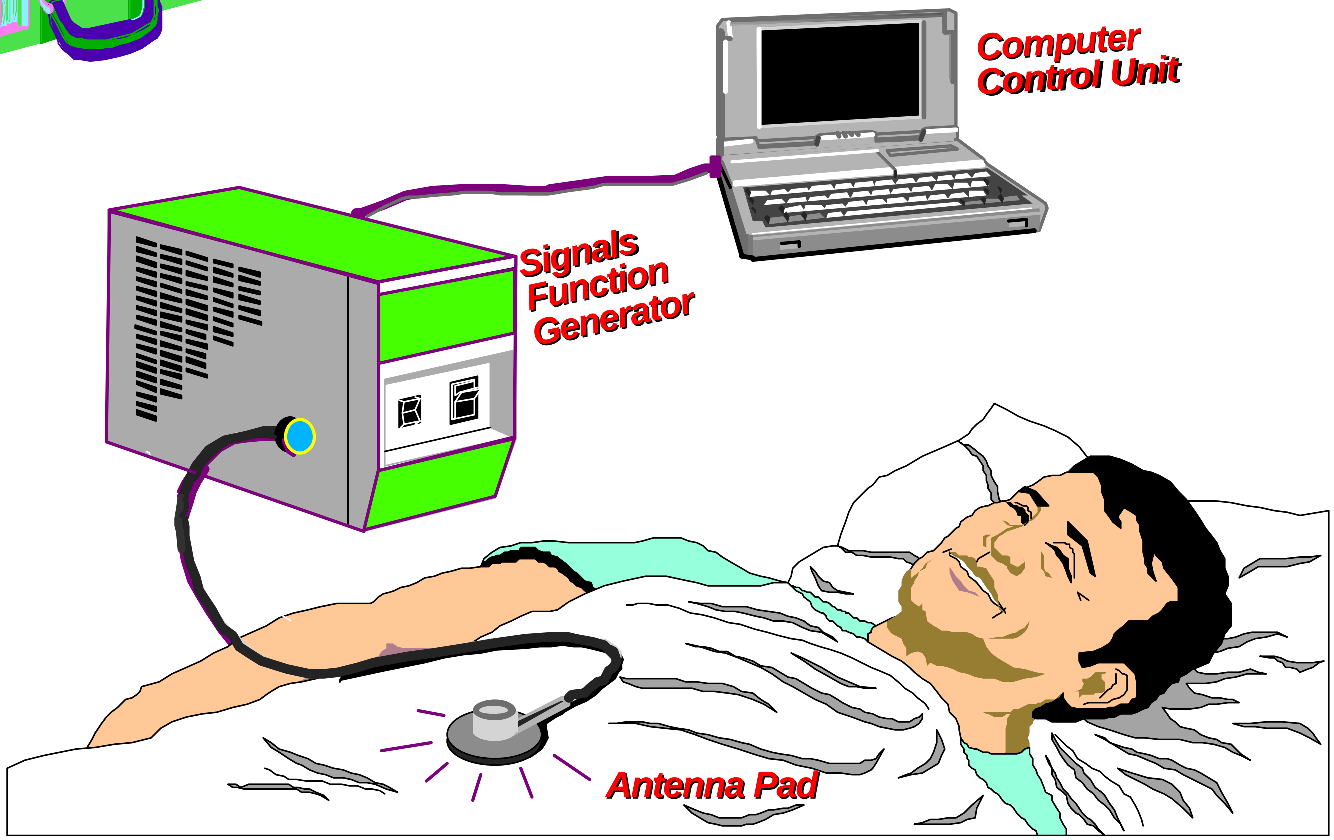
Computer Control Unit
Signals Function Generator
Antenna Pad

The Porthole concept:

The two-minute treatment shortcut

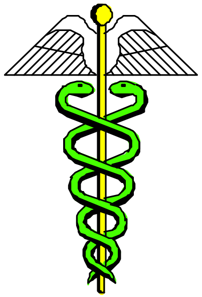

How the porthole concept works

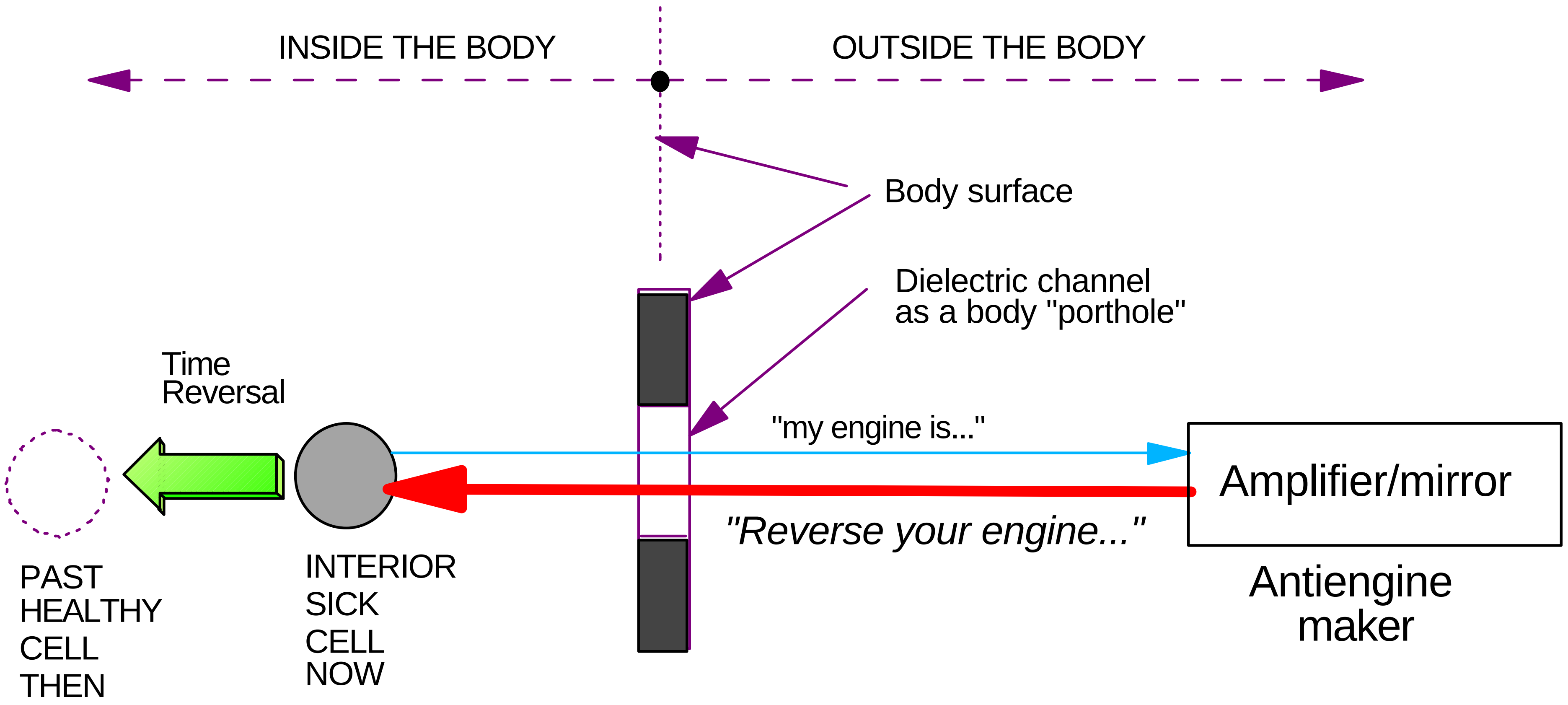

INSIDE THE BODY

OUTSIDE THE BODY

Body surface

Dielectric channel
as a body "porthole"

Time
Reversal

"my engine is..."

Amplifier/mirror

"Reverse your engine..."

PAST
HEALTHY
CELL
THEN

INTERIOR
SICK
CELL
NOW

Antiengine
maker

Porthole concept: An analogy of the dielectric path and inner-outer functional connections

- Imagine perfectly reversible dynamic light processes inside a large "cruise ship", that are also perfect phase conjugate mirrors
- Each process receives, absorbs, and emits light continuously
- Light escaping through the portholes is a composite from all the dynamic processes ongoing inside
- If we record and amplify the escaping light, then force it back through the portholes, it reverses all the internal processes
- The processes revert back, slowly, to a previous ongoing process state
- We cease radiation, and the processes reverse and move forward again

Porthole concept: The discovery

- The processes themselves, *in their own "frame"*, are never actually "reversed". Instead, the frame's motion through time is reversed.
  - Each to itself moves normally -- just receiving, absorbing, and radiating light, regardless of "whatever light" is interacting
  - This is due to a peculiarity of the photon and photon interaction
- Photon interaction creates observed physical reality, including the "frame" and the seeming "passage of a mass through time"
  - The photon is its own antiparticle, which is just a photon traveling in the opposite direction
  - So in this case we can use "light" to precisely time reverse (to the observer) the frame and the apparent operations (to the observer)
- When we cease irradiation, after a decay period to remove the reversed excitation, all functions are now seen to be moving forward normally in observer forward time, but from a previous physical state or condition

Using the porthole concept: Research

- The body's own emitted dense, broadband signals are recorded across a broad bandwidth
  - These signals are amplified, then rebroadcast precisely back into the body
  - This returns an amplified "reversed" signal set back through the dielectric path, invoking and applying the porthole concept in all processes
- The returning photons produce time-charging but in reversed frame direction
  - Tremendous potential energy is absorbed in highly compressed "time-energy" form in seconds
  - This "time-potential" slowly drains off as the body processes and cells reverse back to a previous state

Research using the porthole concept

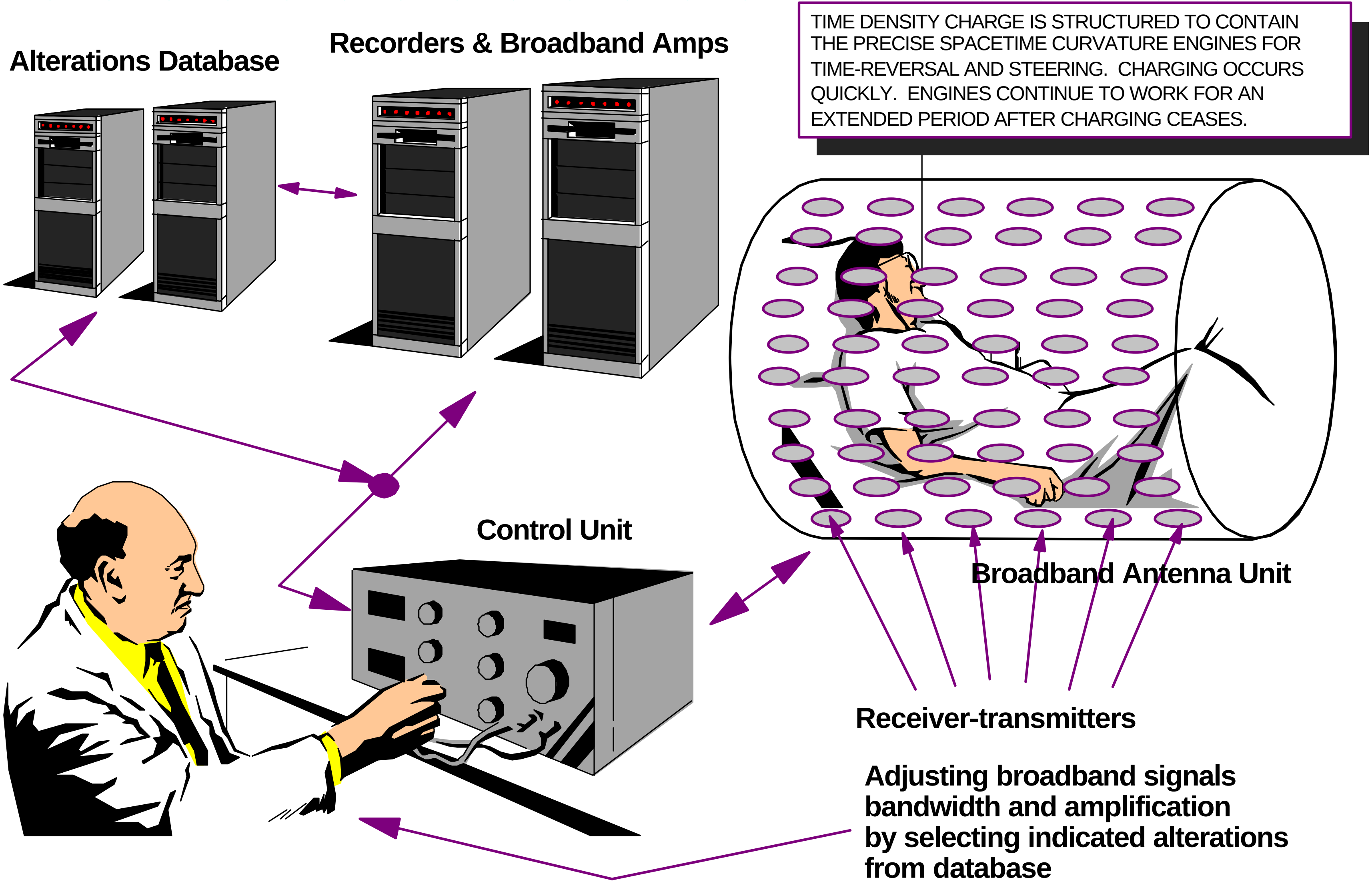

TIME DENSITY CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Alterations Database

Recorders & Broadband Amps

Control Unit

Broadband Antenna Unit

Receiver-transmitters

Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database

© T.E. BEARDEN 1998

Portable unit then developed for rapid treatment of mass BW casualties

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.
- Irradiation for 2 minutes.

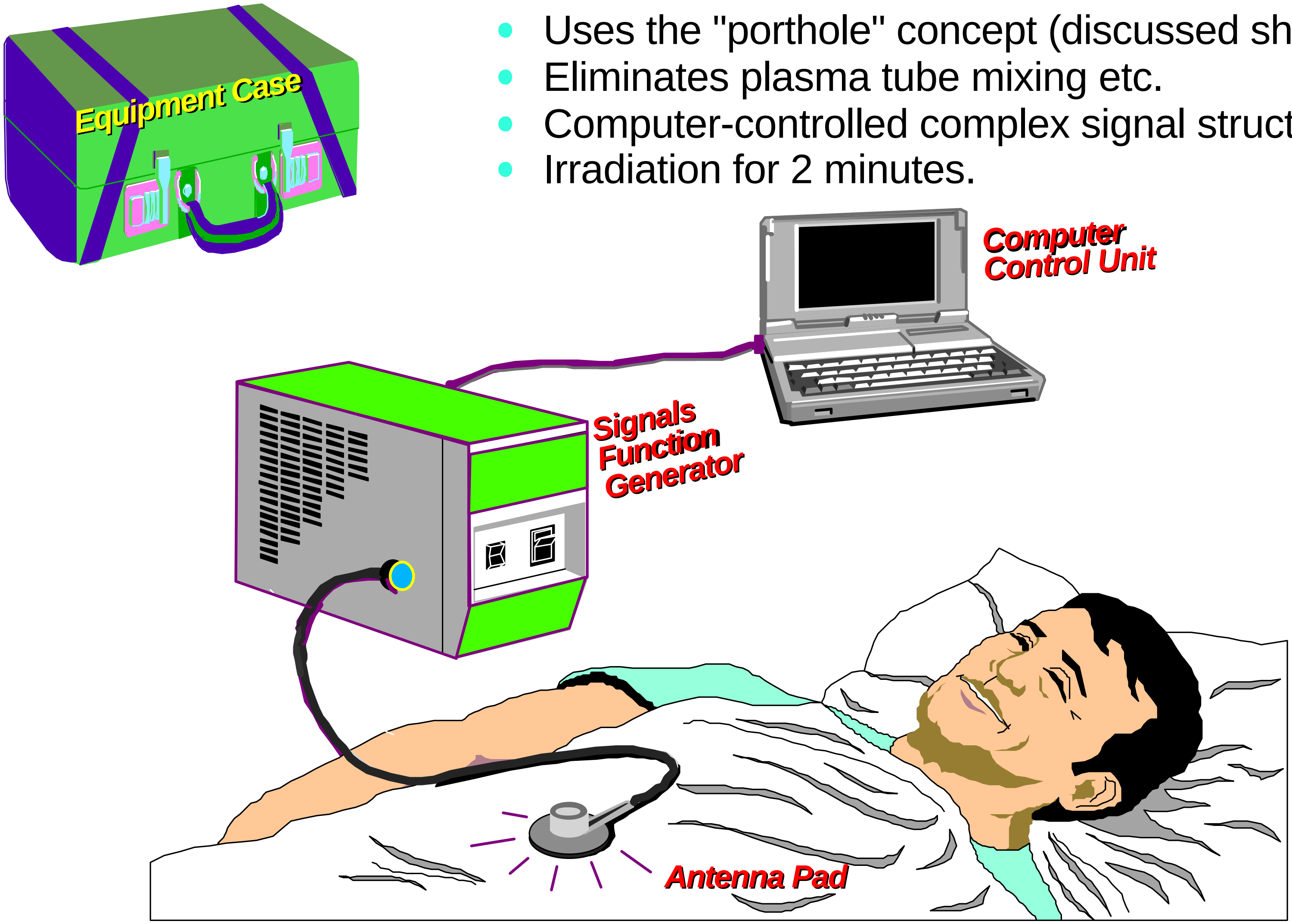
Equipment Case
Computer Control Unit
Signals Function Generator
Antenna Pad

Training to use the unit

- High school student level, 30 minutes.
- Patient lies on one blanket, other covers him.
- Set the prescribed standard signal sets from listing.
- Computer-controlled generator produces the signal.
- Irradiation for 2 minutes, uncover patient, bring in another.
- Unit can be self-powering from vacuum energy (MEG*).

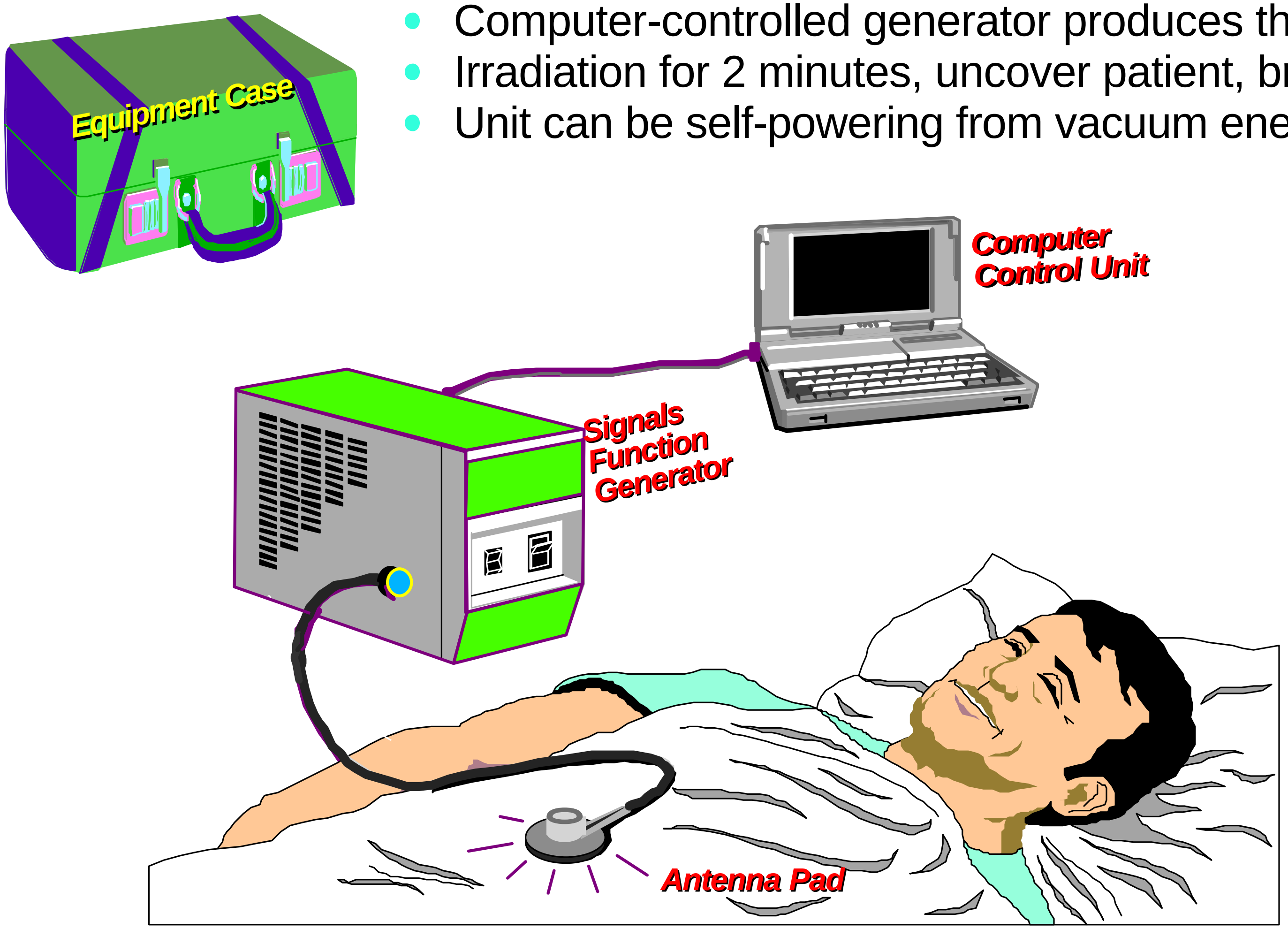
Equipment Case
Computer Control Unit
Signals Function Generator
Antenna Pad

Additional advantages

- Can be developed in a crash program in 2-3 years.
- After sunk costs, projected cost per unit in mass production is less than $100,000
- Can be produced by the thousands
- Then flooded down through
  - Fire stations and police stations
  - Red cross units, hospitals, clinics
  - Emergency action teams and armed forces
- Large numbers of operators easily produced
- Emergency training could be watching a 15 minute film, then performing one graded practice run
- First generation equipment can save some 70% of the coming mass casualties from anthrax etc.

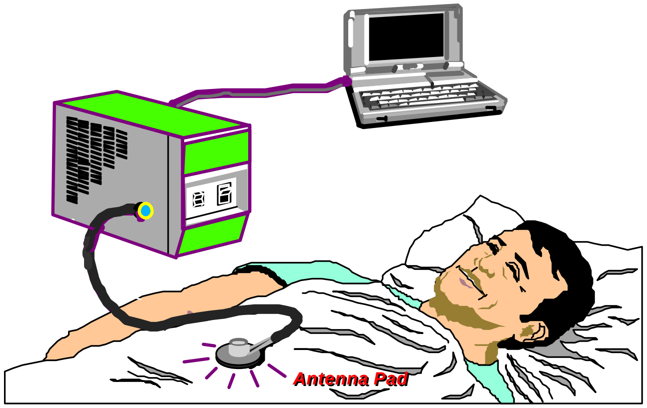

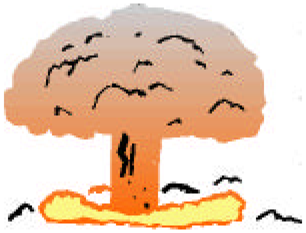

Terrorist BW warfare capabilities against the United States

The growing threat

Mass Casualty Problem:
Aerial Anthrax Spray on Washington, D.C.

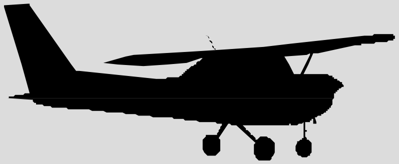

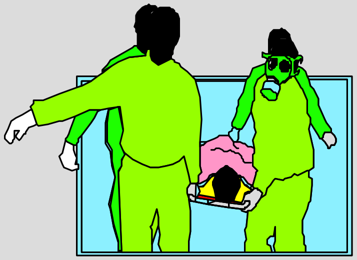

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

*Per OTA Report to Congress, 1993

For decades, infiltration of terrorist teams into the U.S.

- From several foreign nations, not just Bin Laden in Afghanistan
- Several thousand Cuban guerillas, from training camps in Southern Mexico
- Russian Spetznaz teams
- Nuclear warheads (up to 80 KT)
- Biological warfare agents
- Yakuza in oriental populations in our cities
  - Yakuza manufactures Russian energetics weapons in their own facilities in Japan
  - Possesses the new EM weapons referred to by Secretary of Defense Cohen in 1997
- We are now in a very long, hard war and will suffer millions of casualties in our cities. It may last for decades.

Defense Secretary Cohen*

(Involvement of secret superweapons)

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."

* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Some typical BW agents

- Smallpox, including different strain developed by Russia (some sold to terrorists)
- Anthrax
  - Over 400 known locations around the Earth have cultures
  - Difficult to decontaminate
- Bubonic Plague
- Botulism toxin
- Tularemia (rabbit fever)
- Hemorrhagic fevers, such as Ebola

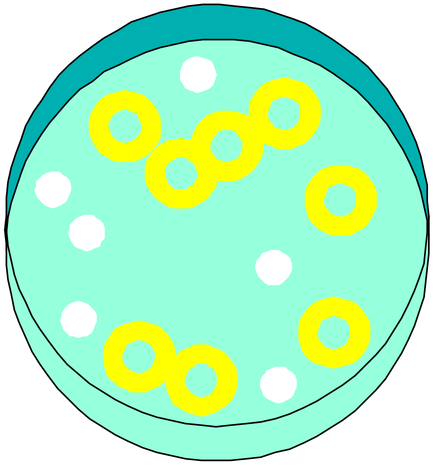

Anthrax

- OTA study: Some 200 pounds released in a spray, on calm night over Washington, D.C. could result in up to 3 million casualties
- Symptoms usually in 2-10 days
  - Sometimes as long as 7 weeks
  - Flu-like symptoms

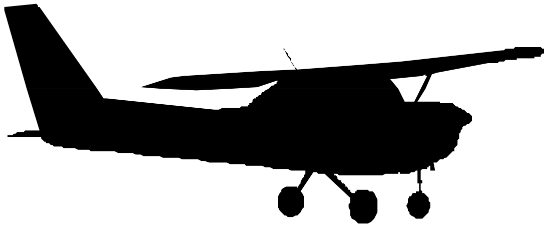

- Then 1-3 days later, shock and breathing problems cause death in up to 100%
- Antibiotics given early (before symptoms) can prevent infected persons from becoming ill
- Limited vaccine, reserved for military

Smallpox

- More than 500 million people killed in the 20th century before "eradicated" in 1977. Vaccination ceased in 1980.
- Russians kept it, developed new strains, sold to terrorists
- Symptoms in about 12 days: fever, headache, nausea. Rash similar to chickenpox then appears.
- Rash turns into hard blisters. *Highly contagious*, kills one-third of its victims.
- No cure. About 12 million doses of vaccine left.
- If smallpox is unleashed in any major city on earth, it will eventually kill 2 billion, or about 1/3 of the human species -- unless a radical new cure is found and utilized.

Bubonic Plague: Historically the most feared (Black Death)

- From 1980-1994, 18,739 cases in 20 countries.
- Symptoms occur in 1-6 days after inhaling the pneumonic form.

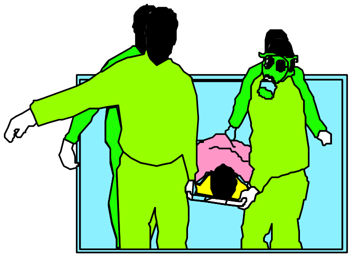

- High fever, cough, labored breathing lead to respiratory failure and death.
- It is contagious.
- Rapid use of antibiotics can be effective.
- Vaccine not currently produced in the U.S.

Botulism toxin (a nerve toxin)

- Single most poisonous BW substance known.
  - Typically borne in food, contacted by eating.
  - Could be developed as aerosol weapon.
- Symptoms in 24-36 hours; blurred vision and difficulty swallowing, speaking.
- Paralyzes muscles, leading to respiratory failure and death.
- Centers for Disease Control and Prevention maintain a botulism antitoxin supply.

Tularemia

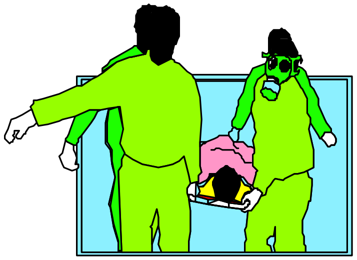

- U.S. studied its weapon potential in 1950s and 1960s.
- Symptoms appear in 3-5 days, including fever, chills, headache, and weakness
- Inflammation and hemorrhaging of the airways can lead to death.
- Without antibiotics, one-third die. Can be 90% if breathed in, in aerosol form.
- Vaccine developed and under review by FDA.

Hemorrhagic fevers, such as Ebola

- E.g., origin of Ebola is unknown. Probably transmitted to humans by animals, to begin a breakout.
- Symptoms in 3-5 days, include fever, aching muscles, diarrhea.
- Hemorrhaging of fluids out of tissues and orifices.
- From 30 to 90 percent die, depending on specific fever type.
- Some of these fevers respond to antiviral drugs.
- The drugs are in short supply.
- Presently *containment* is the primary goal.

A factor the U.S. does not know

(Secret involvement of clandestine superweapons)

- Three nations (not the U.S.) developed quantum potential weapons using Bohm's hidden variable theory of quantum mechanics.
- Can induce disease engines in matter or bodies at a distance.
- Have developed "spreading" of immune systems in a targeted populace (cleverly tested in Gulf War against U.S. troops).
  - Cocktail mix of disease engines for 2 dozen different diseases
  - Targeted area's human immune systems react to these engines when they are still in virtual state, just below quantum threshold.
  - Spread immune system's resources thus "thinned" across response to two dozen shadow diseases at once.
- Hit with a terrorist BW attack, the spread immune systems are overwhelmed much more easily and quickly.
  - Yield (casualties) of the attack is increased by factor of 3-5.
  - Spreading of U.S. immune systems has been accomplished and is still being maintained.

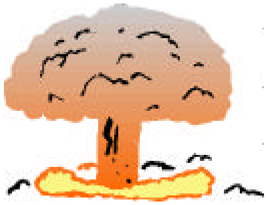

Summary and conclusion

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

The new war

A new Pearl Harbor: Strike on the World Trade Center in NYC

- The North Tower has already been struck and is burning.
- A second jetliner is about to strike the South Tower.

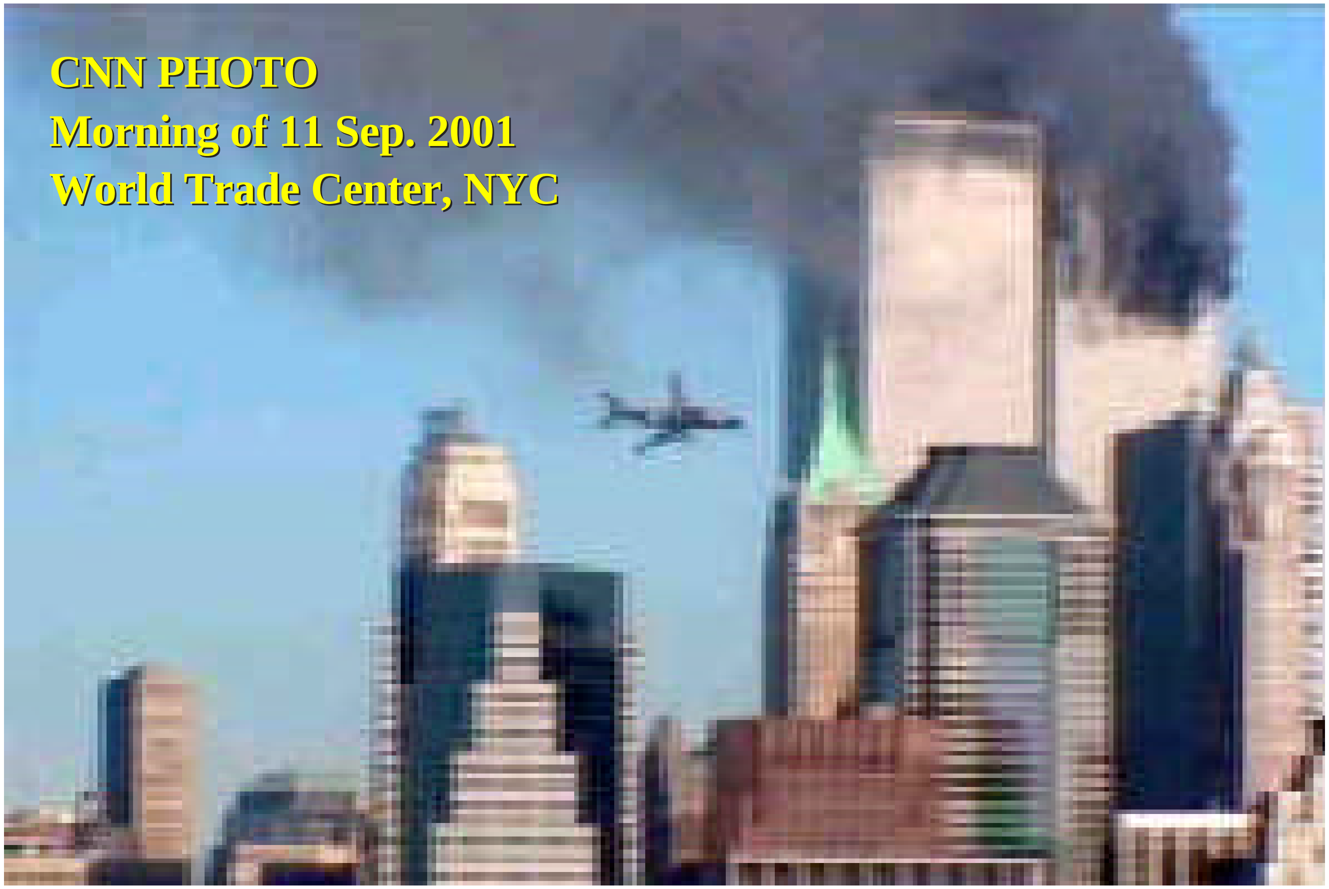

CNN PHOTO
Morning of 11 Sep. 2001
World Trade Center, NYC

Fireball From Impact on South Tower

- The two buildings held some 50,000 occupants
- Evacuation was underway after the first strike
- More than 6,000 persons were killed
- This is more than were killed in the Dec. 7, 1941 strike on Pearl Harbor

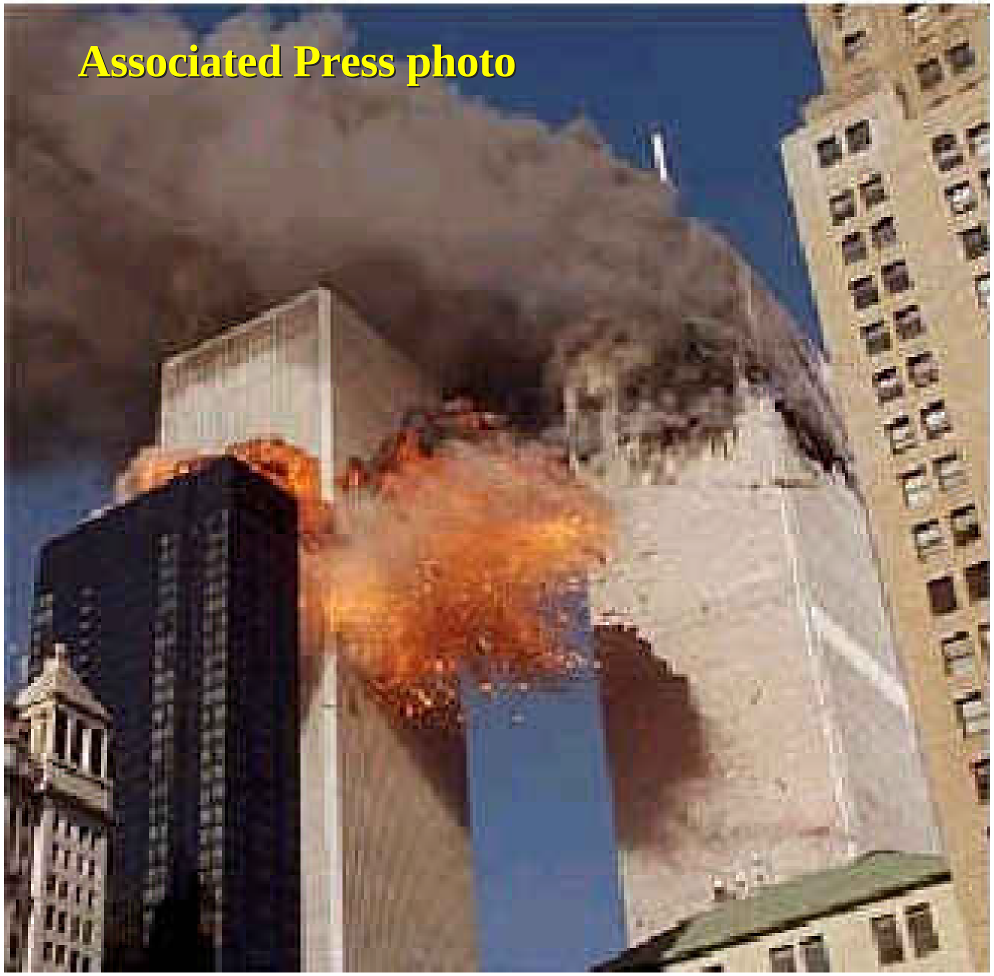

Associated Press photo

Collapse of the South Tower

- The first tower to collapse was the South Tower
- It is still uncertain just how many people were caught inside and killed
- This is a half million tons of debris, falling from the New York skyview.

Photo from CNN

Collapse of the North Tower begins

- The North Tower then collapsed
- This was another half-million tons of debris falling from the NYC skyview

Collapse of North Tower finishes

- Collapse of the North Tower killed many rescue workers, fire-men, and policemen
- At this writing, the total casualties are still unknown

Photos from CNN

Next began the recovery efforts

- Recovery operations will require months
- At this writing, the total casualties are still unknown
- The damage was done by two civilian jetliner "bombs"

The Pentagon was hit on the same day

- Hundreds were killed the same day when another jetliner struck the Pentagon.
- Yet another jetliner, possibly headed for the White House, crashed enroute when the passengers attacked the suicide terrorist hijackers.

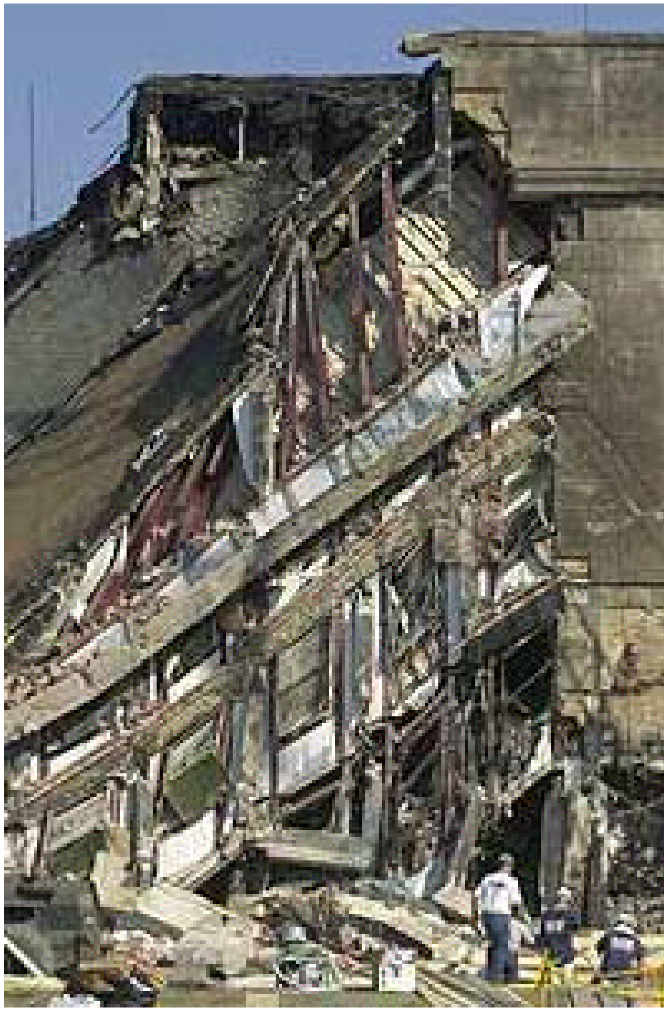

A sobering thought

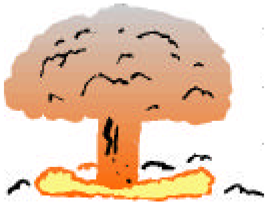

- Nuclear weapons are already smuggled into the U.S., as are BW agents such as anthrax, smallpox, etc.
- Had a 40-80kT weapon been exploded in New York and one in Washington, millions would have perished and both cities would have been devastated
- There would have been no recovery of those cities for many decades

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach knockout capability***

© 1998 T.E. BEARDEN

***Per OTA Report to Congress, 1993**

The U.S. is terribly unprepared

- Limited medical supplies, antibiotics, vaccines
- Insufficient facilities, emergency teams
- No civilian shelters (much less BW-filtered)
- Not enough doctors, nurses, masks, etc.
- Present treatment methods woefully inadequate
- Triage will apply
  - First treat those most likely to recover
  - Others (showing symptoms, or advanced) are just set aside to die
- Other attacks are a matter of *when*, not *if*.

Conclusion

- As several cities are struck, millions of U.S. civilian casualties will result. Devastation will be enormous.
- We shall eventually lose more American lives than in all the rest of our wars to date.
- Present medical therapeutic science is totally inadequate to do the job. It cannot save us.

- At some point, when we are sufficiently damaged, the Spetznaz or other terrorists may detonate nuclear weapons in our cities, with appalling carnage.
- A combined superweapon technology and terrorist groups will then destroy the remaining populace.
- On its present course, the inadequacy of Western medicine may doom us.

A difficult truth to consider

- In any strategic strike, the first phase is to get the strategic weapons (weapons of mass destruction) delivered to their distant targets.
- That has already been done.
- Eerily, the first phase of World War III has already been accomplished.
- It was already accomplished before the Sept. 11, 2001 strike on New York and Washington. That was the new "Pearl Harbor", and only the opening round. It already killed more Americans than died at Pearl Harbor on Dec. 7, 1941.
- Unless we improve our medical therapy quickly and dramatically, we will lose -- particularly when the hidden energetics weapons and outside terrorist sponsorship by hostile nations are factored in.

A personal perspective

- There now exists an overwhelming, immediate moral imperative to get this development program done at the utmost speed humanly possible.
- The sheer survival of the United States and Western civilization depends upon it.
- Officialdom is aware that the KGB/Russians have long planned to finish the U.S., after initial nuclear exchanges, by fierce and massive BW strikes.
- It appears that the plan has changed to let the terrorists do the main work, while augmenting U.S. casualty levels by energetics weapons.

China is also involved

- China has declared South China Sea the territorial waters of China
- China will eventually take Taiwan
  - Has massed missile artillery trained on Taiwan
  - Recently held invasion practice on an island near Taiwan
- Recently deployed fearsome EMP weapons utilizing negative energy rather than positive energy (Dirac sea theory)
  - Extinguishes all electron currents
  - Duds electrical and electron systems instantly
  - Instant death to struck bodies
  - Instant kill of missiles, aircraft, electronics in nuclear warheads
  - Ultimate "death ray"; can be made small or large
- Recent 2-carrier U.S. task force maneuvered in South China Sea to caution China
  - The new Chinese EMP weapon was trained on that task force
  - Could have destroyed it within minutes

Final thoughts

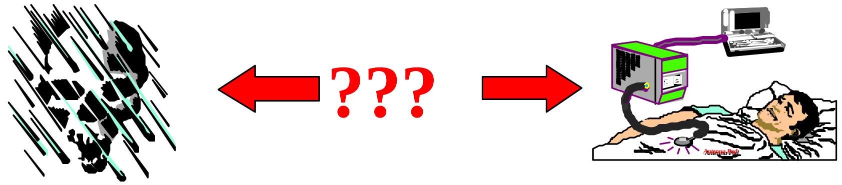

- This is one of the great turning points in history.
- The development program is eminently doable. That it works has been experimentally demonstrated. The long-vexing puzzle of the active mechanism has been solved.
- Our scientific community bitterly opposes any use of electromagnetics -- even an advanced, higher symmetry electrodynamics -- for medical treatment.
- A consortium of the scientific community, the large pharmaceuticals, FDA, and the AMA will oppose this project even while the U.S. is being destroyed wholesale.
- It can only be done in the private sector, and there will likely be no funding unless by philanthropists.

Continuation

Business plan, development program, budget breakdown, proposed schedule etc. go here.

Weapons Slides

[The Tom Bearden Website](#)

[Help support the research](#)

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

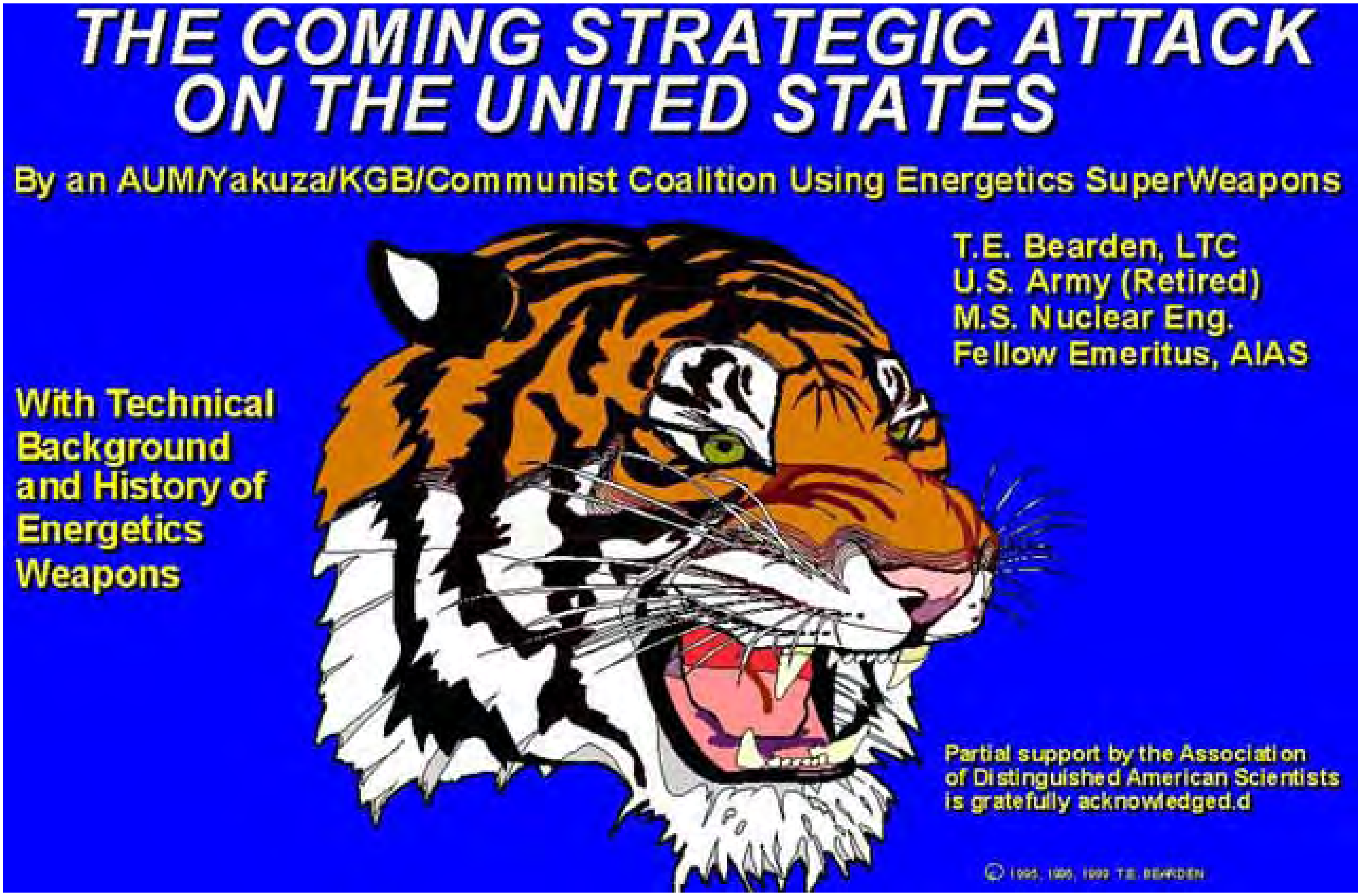

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

© 1995, 1996, 1999 T.E. BEARDEN

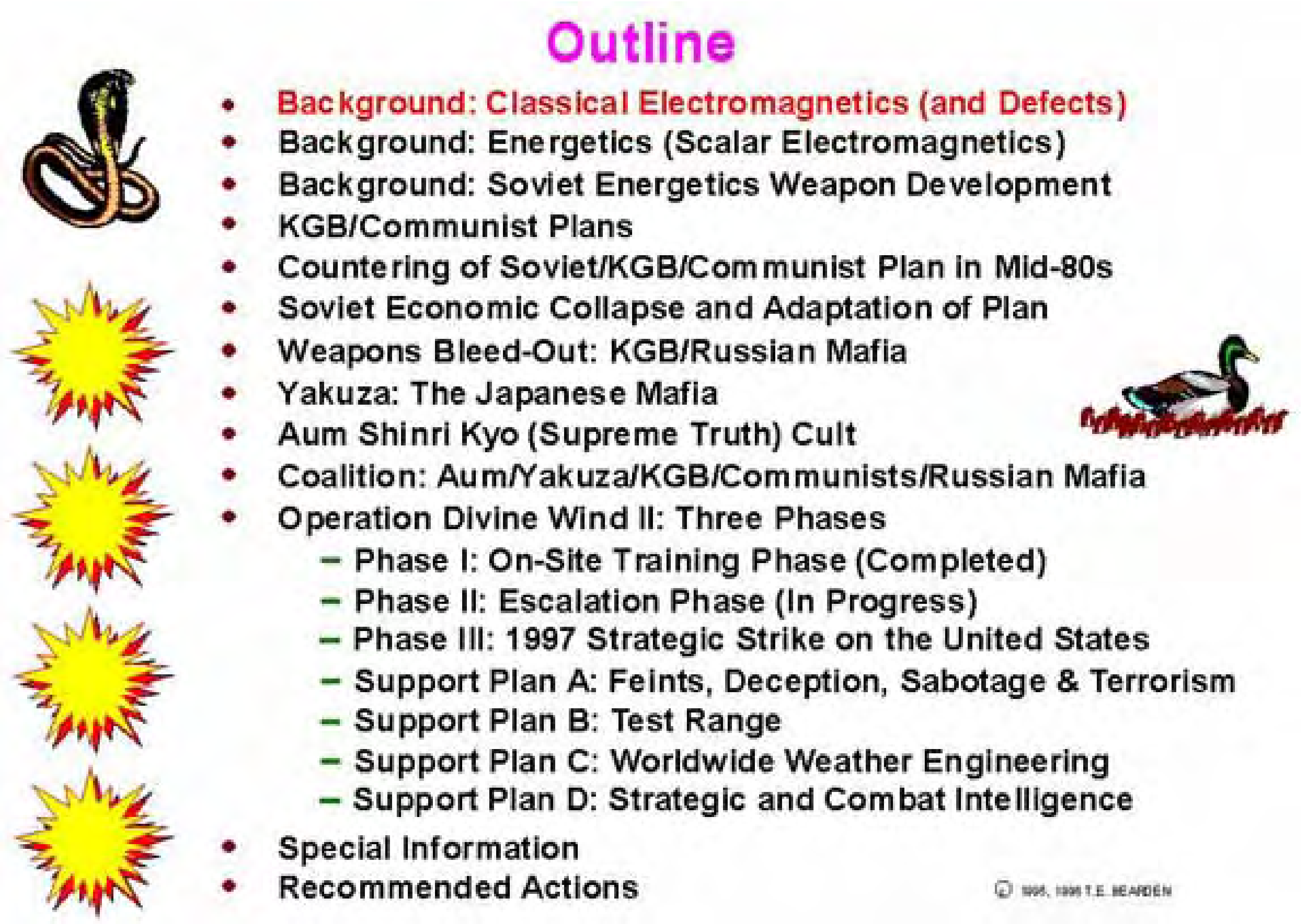

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© 1996, 1998 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

A-10 Mysterious Flyaway and Crash

April 2, 1997

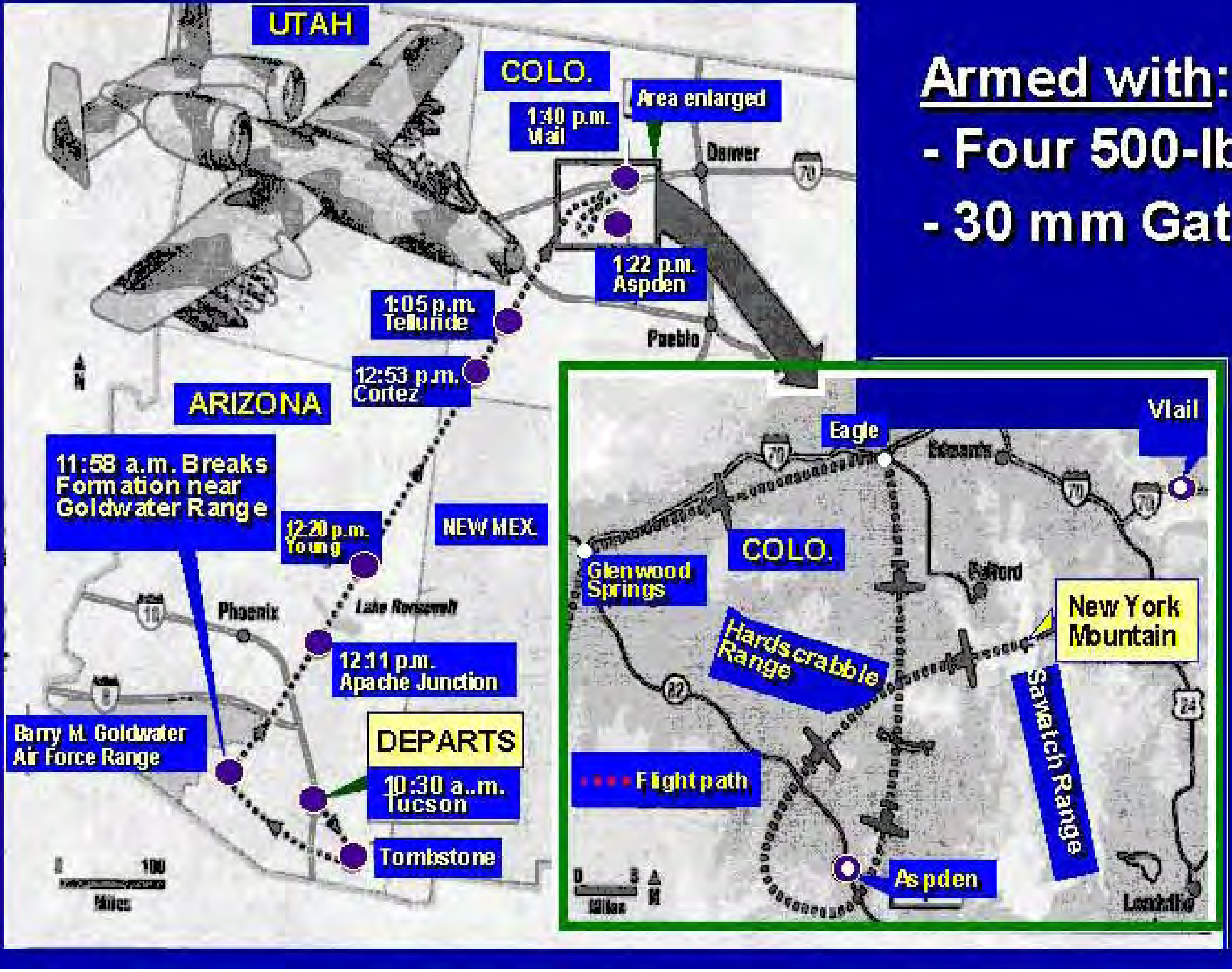
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
Denver
70
1:22 p.m.
Aspden
1:05 p.m.
Telluride
Pueblo
12:53 p.m.
Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m.
Young
NEW MEX.
Phoenix
10
Lake Roosevelt
12:11 p.m.
Apache Junction
8
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a..m.
Tucson
Tombstone
N
0 100
Miles
Eagle
Edwards
Vlail
70
70
Glenwood Springs
COLO.
Hardscrabble Range
Fulford
New York Mountain
82
Sawatch Range
24
Flight path
0 5
Miles
N
Aspden
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

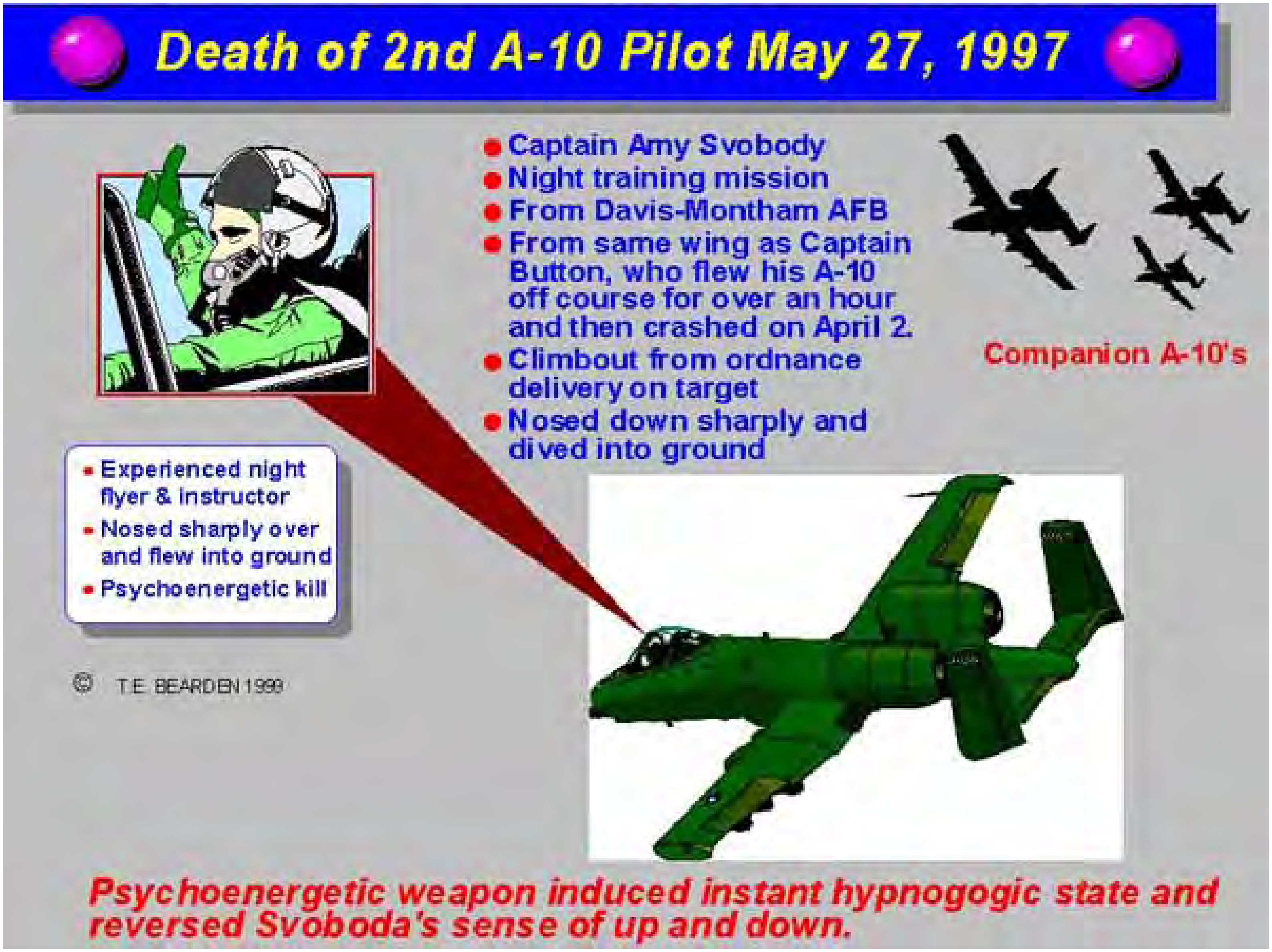

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

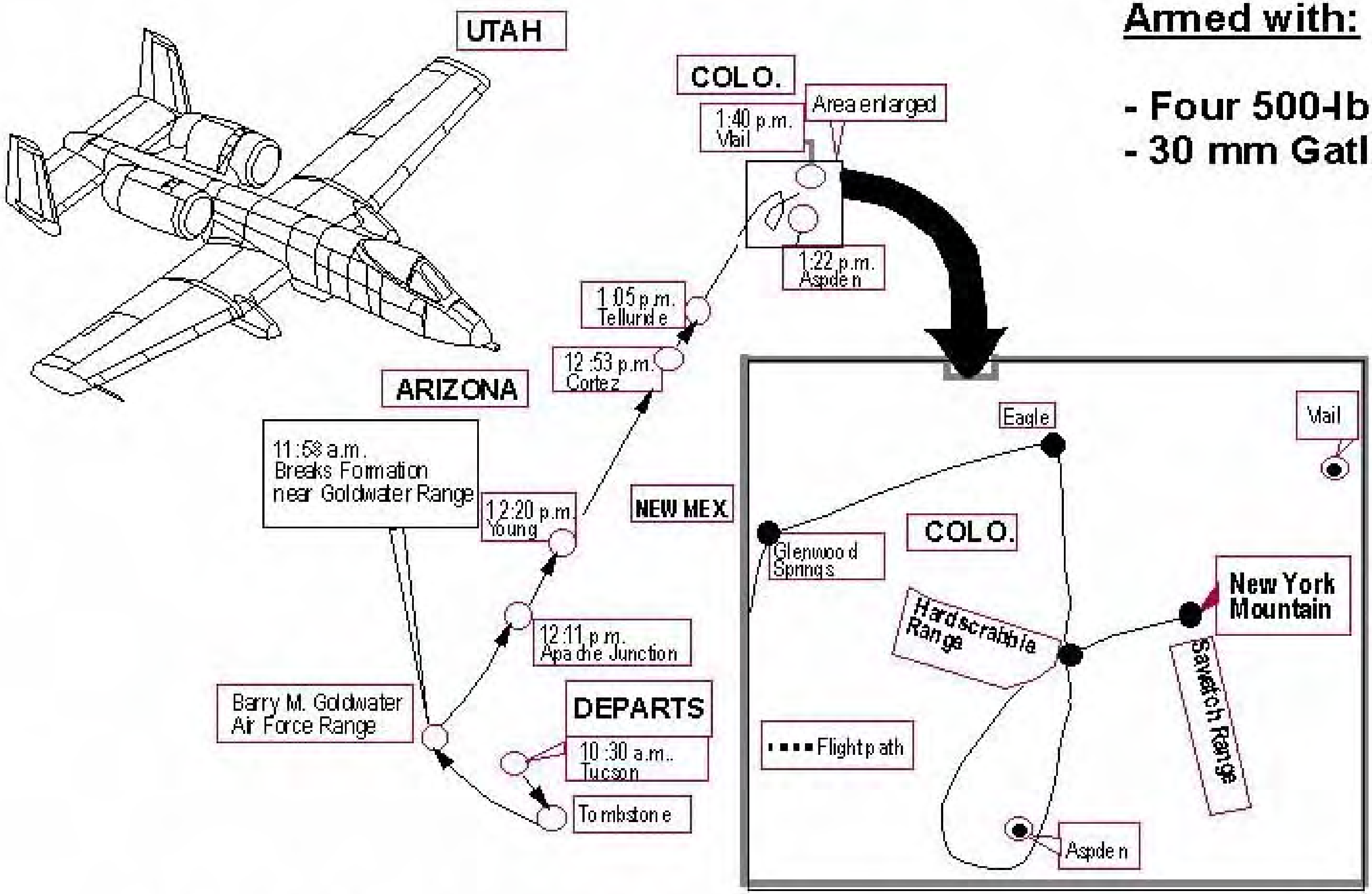
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
1:22 p.m.
Aspden
1:05 p.m.
Telluride
12:53 p.m.
Cortez
ARIZONA
11:58 a.m.
Breaks Formation
near Goldwater Range
12:20 p.m.
Young
NEW MEX.
12:11 p.m.
Apache Junction
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m..
Tucson
Tombstone
Eagle
Vail
Glenwood
Springs
COLO.
Hardscrabble
Range
New York
Mountain
Savatch Range
Flight path
Aspden

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

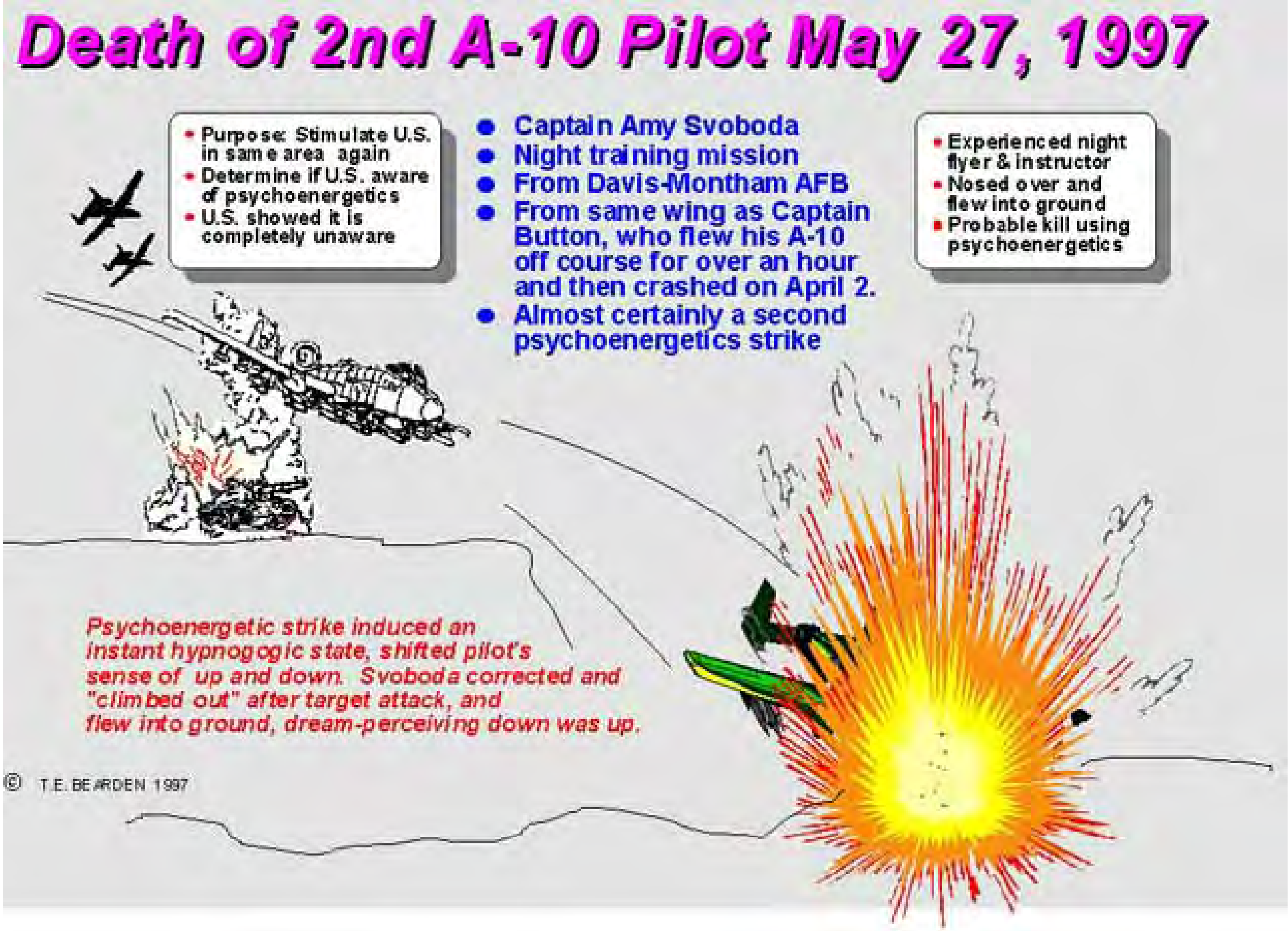

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

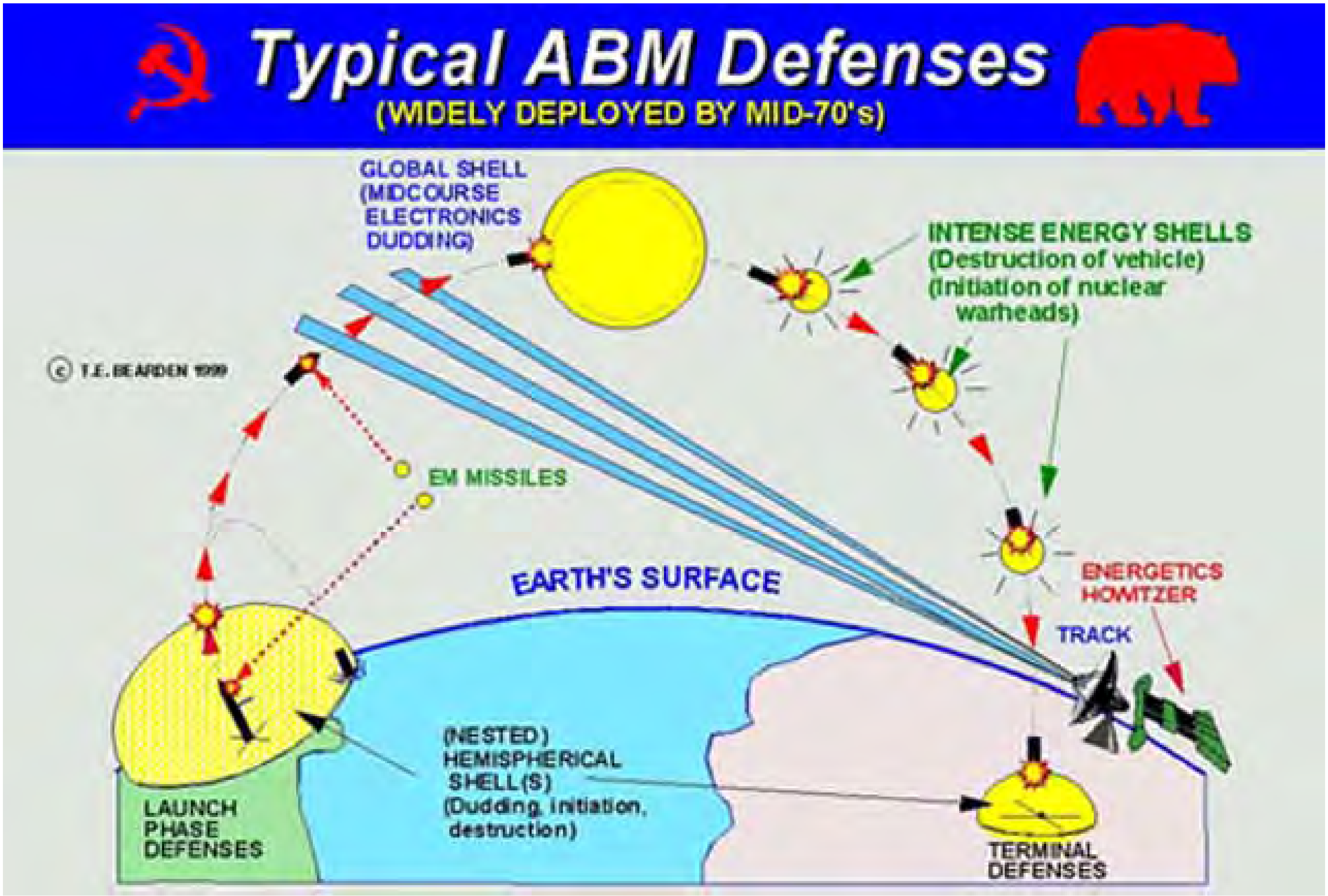
Typical ABM Defenses
(WIDELY DEPLOYED BY MID-70's)
GLOBAL SHELL
(MIDCOURSE
ELECTRONICS
DUDDING)
INTENSE ENERGY SHELLS
(Destruction of vehicle)
(Initiation of nuclear
warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS
HOWITZER
TRACK
(NESTED)
HEMISPHERICAL
SHELL(S)
(Dudding, initiation,
destruction)
LAUNCH
PHASE
DEFENSES
TERMINAL
DEFENSES

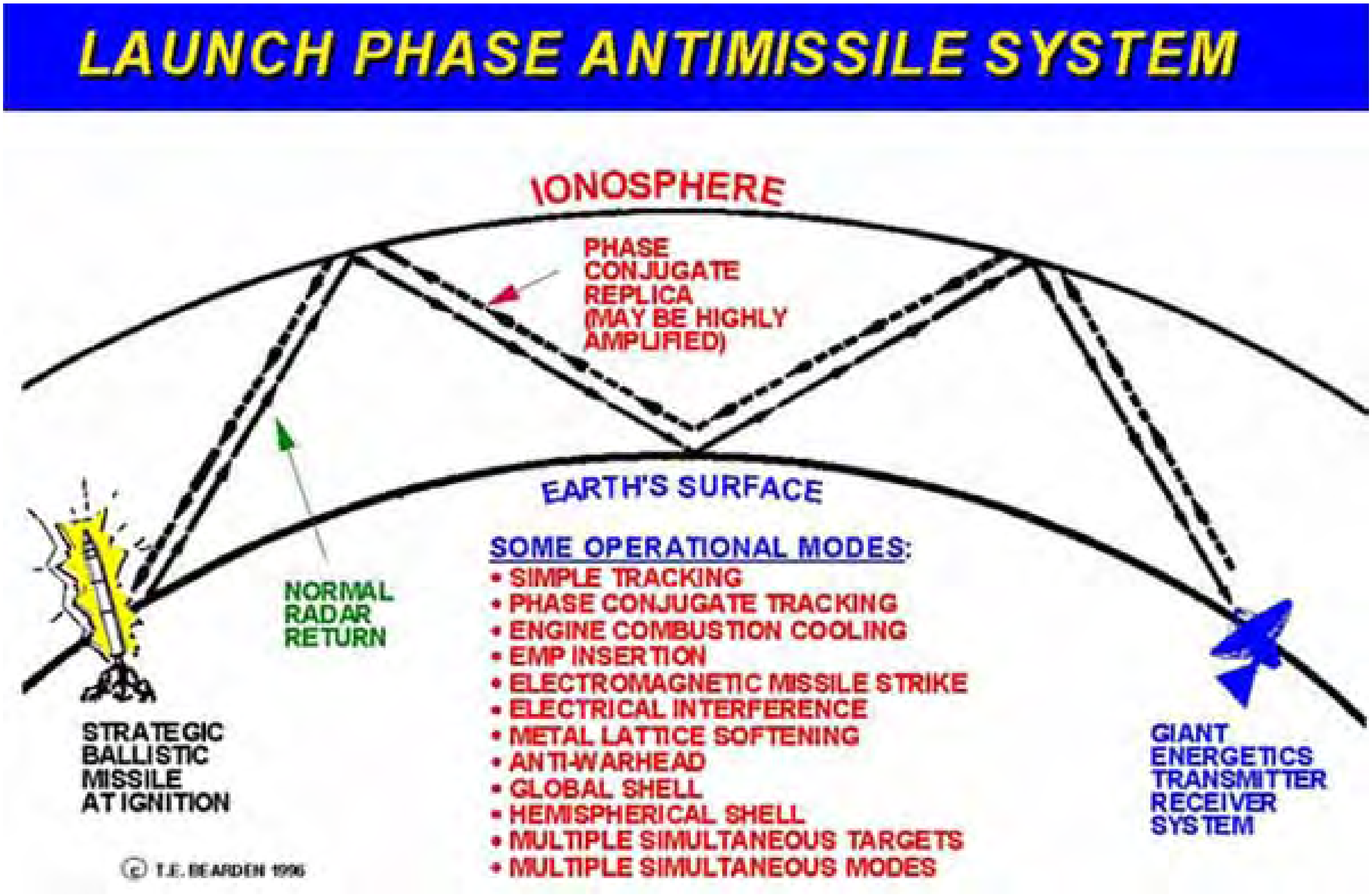

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BALLISTIC MISSILE AT IGNITION
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

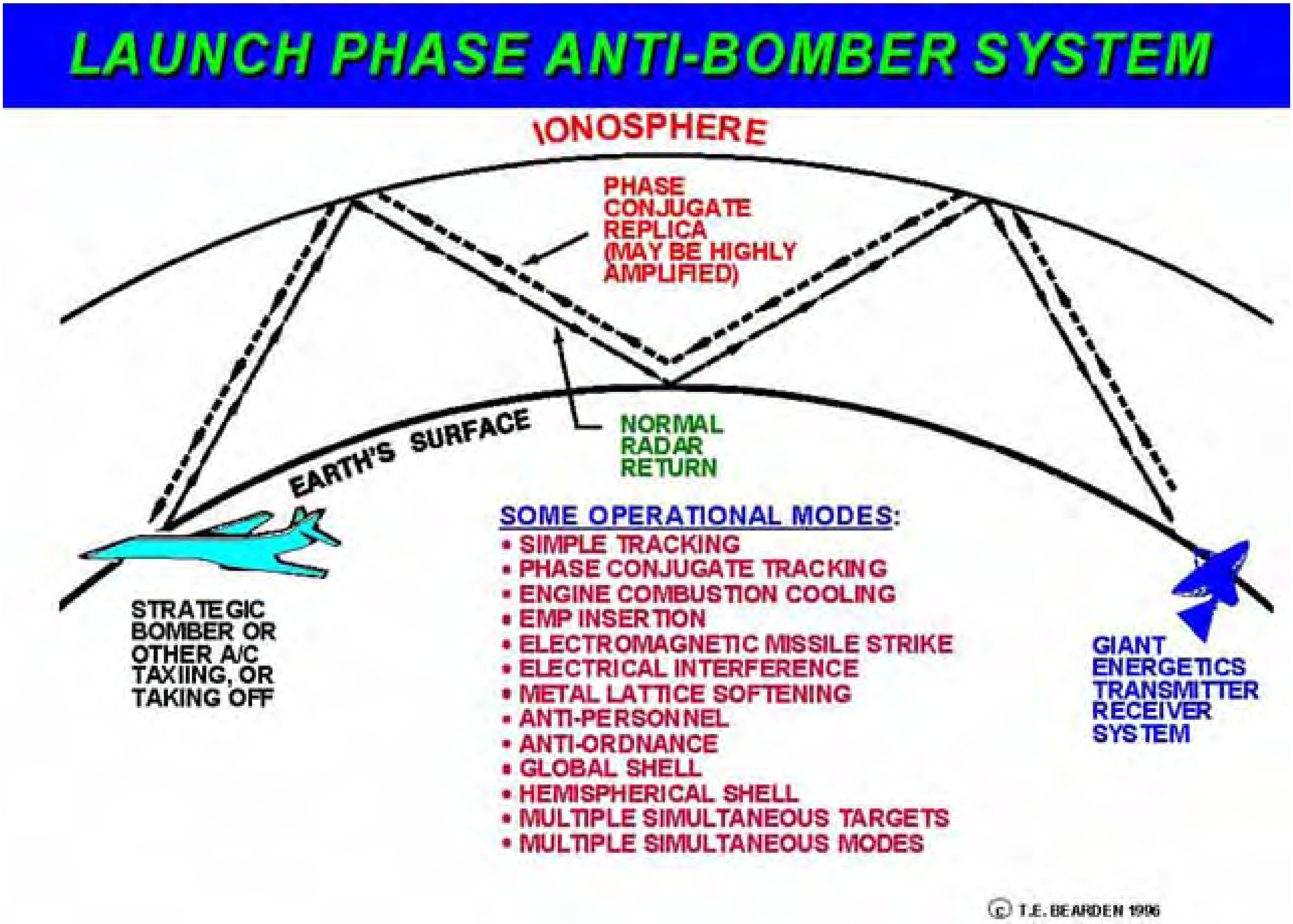

LAUNCH PHASE ANTI-BOMBER SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

ⓒ T.E. BEARDEN 1996

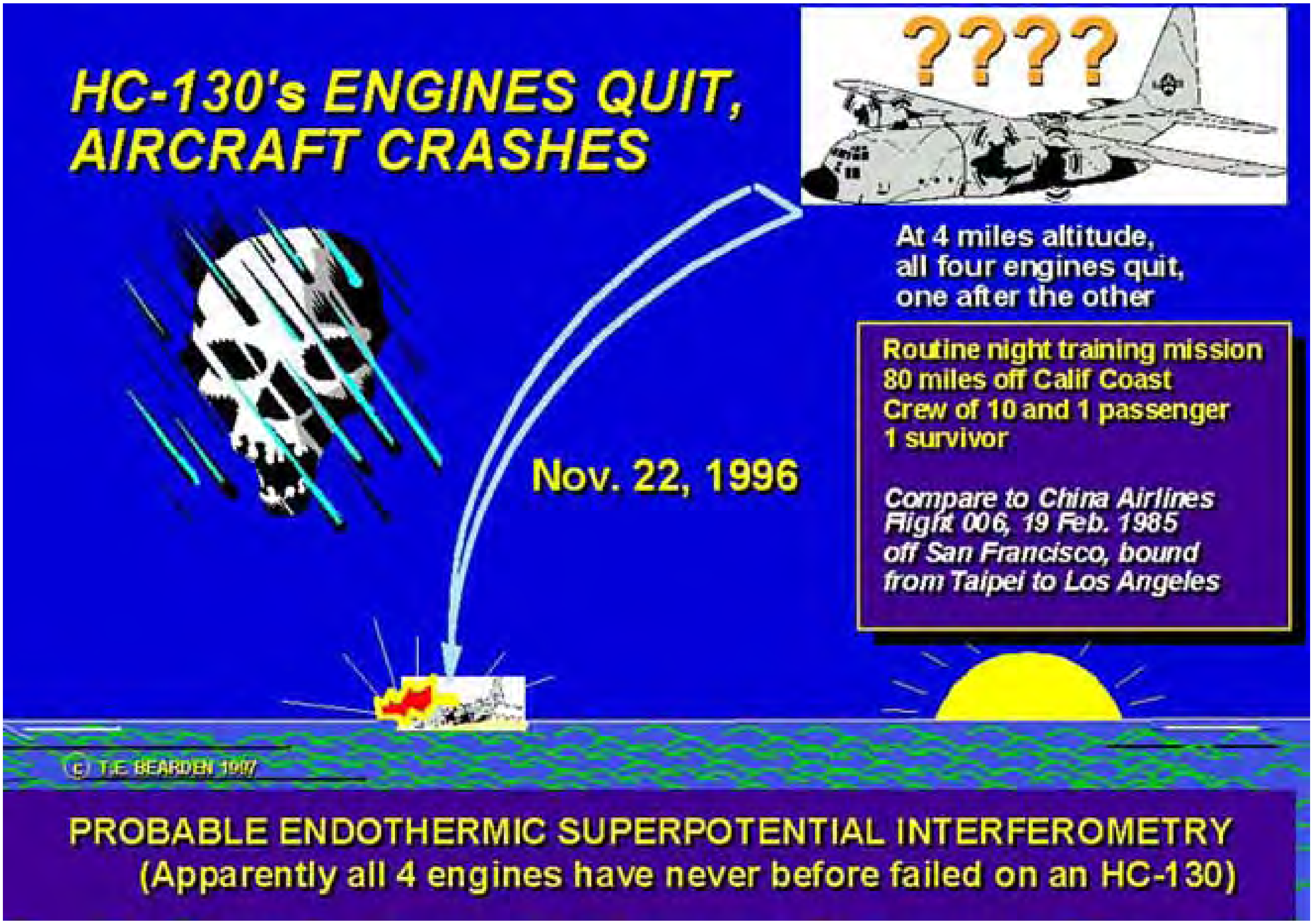

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Nov. 22, 1996

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

© T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

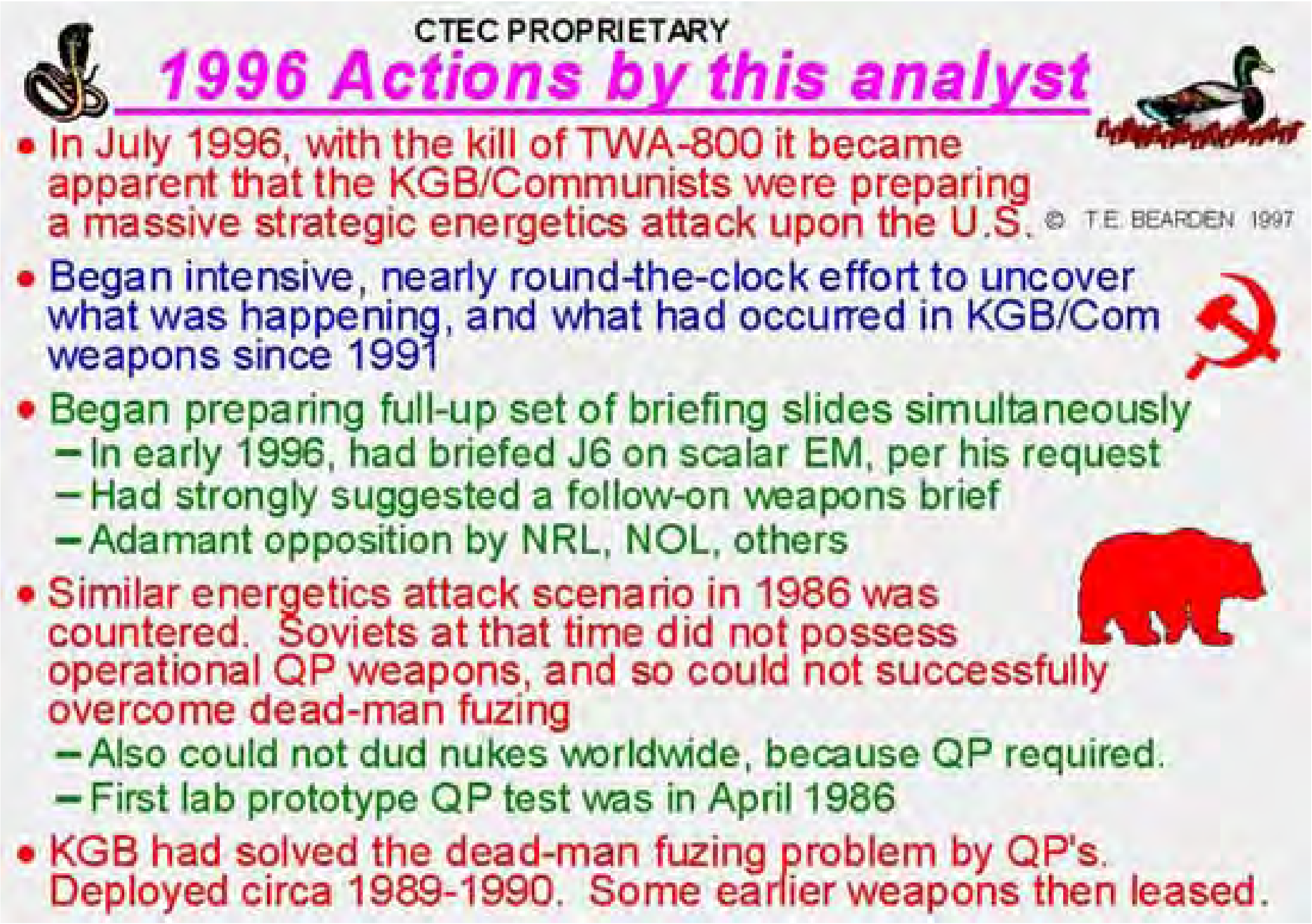

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

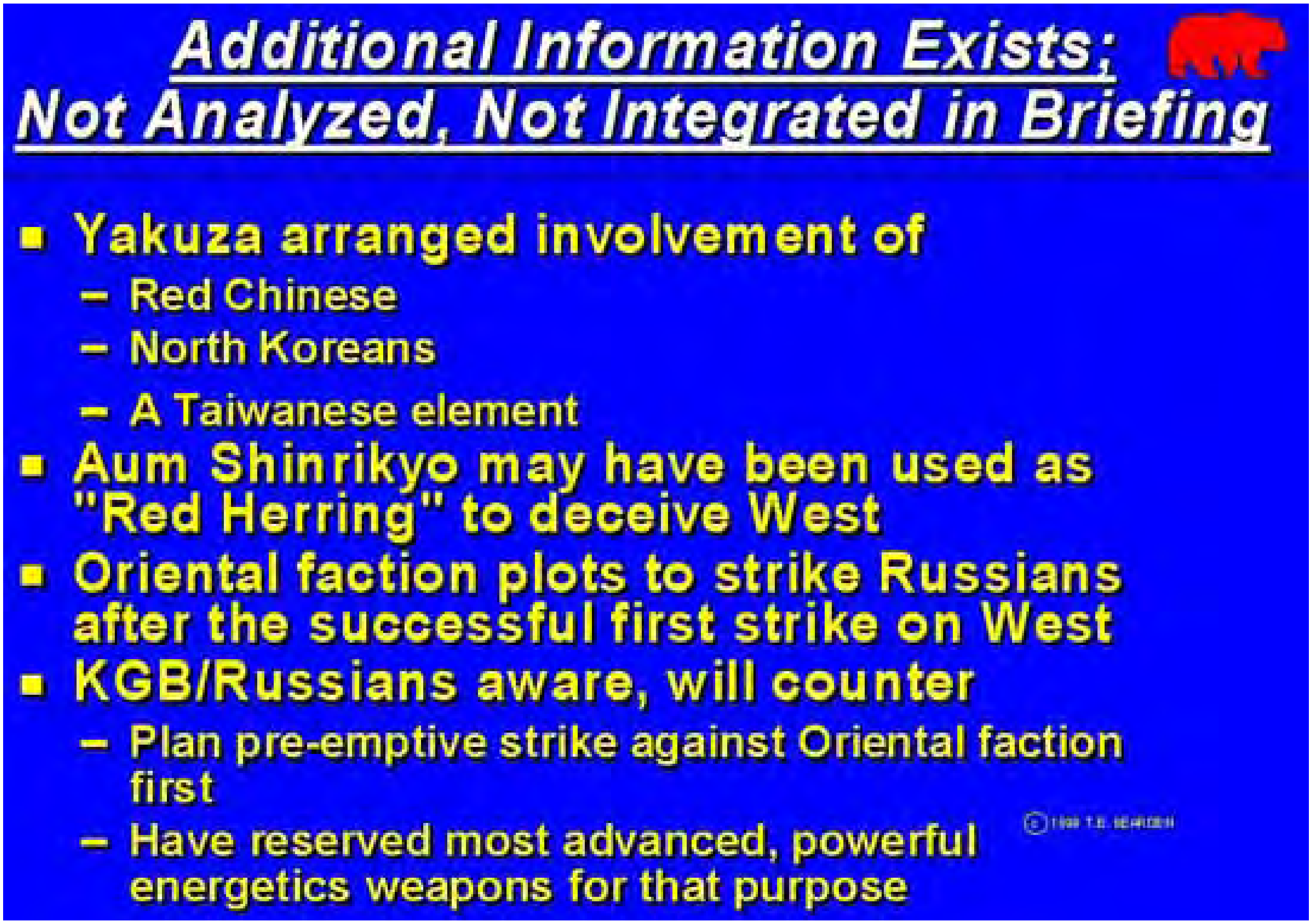

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

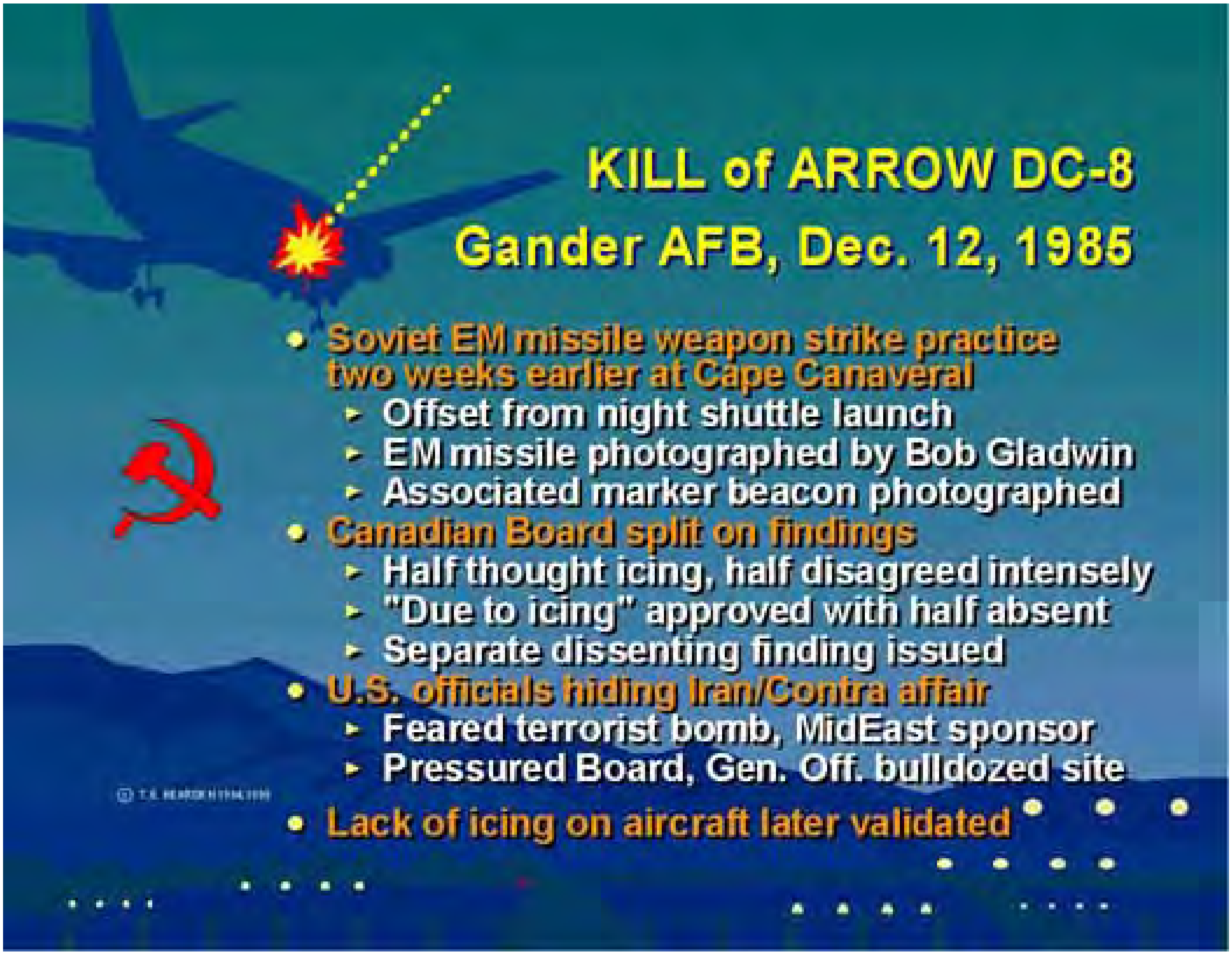

KILL of ARROW DC-8

Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

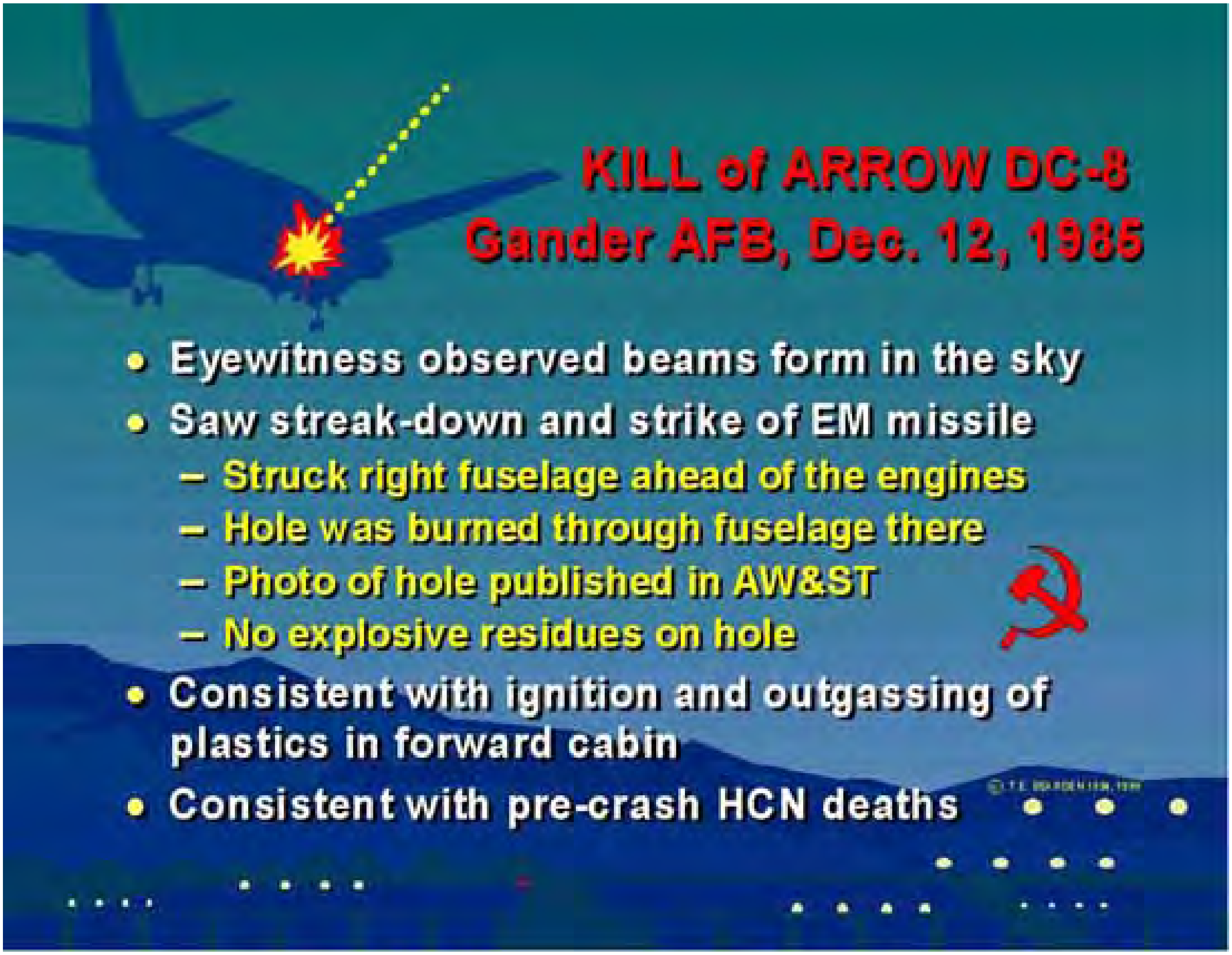

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

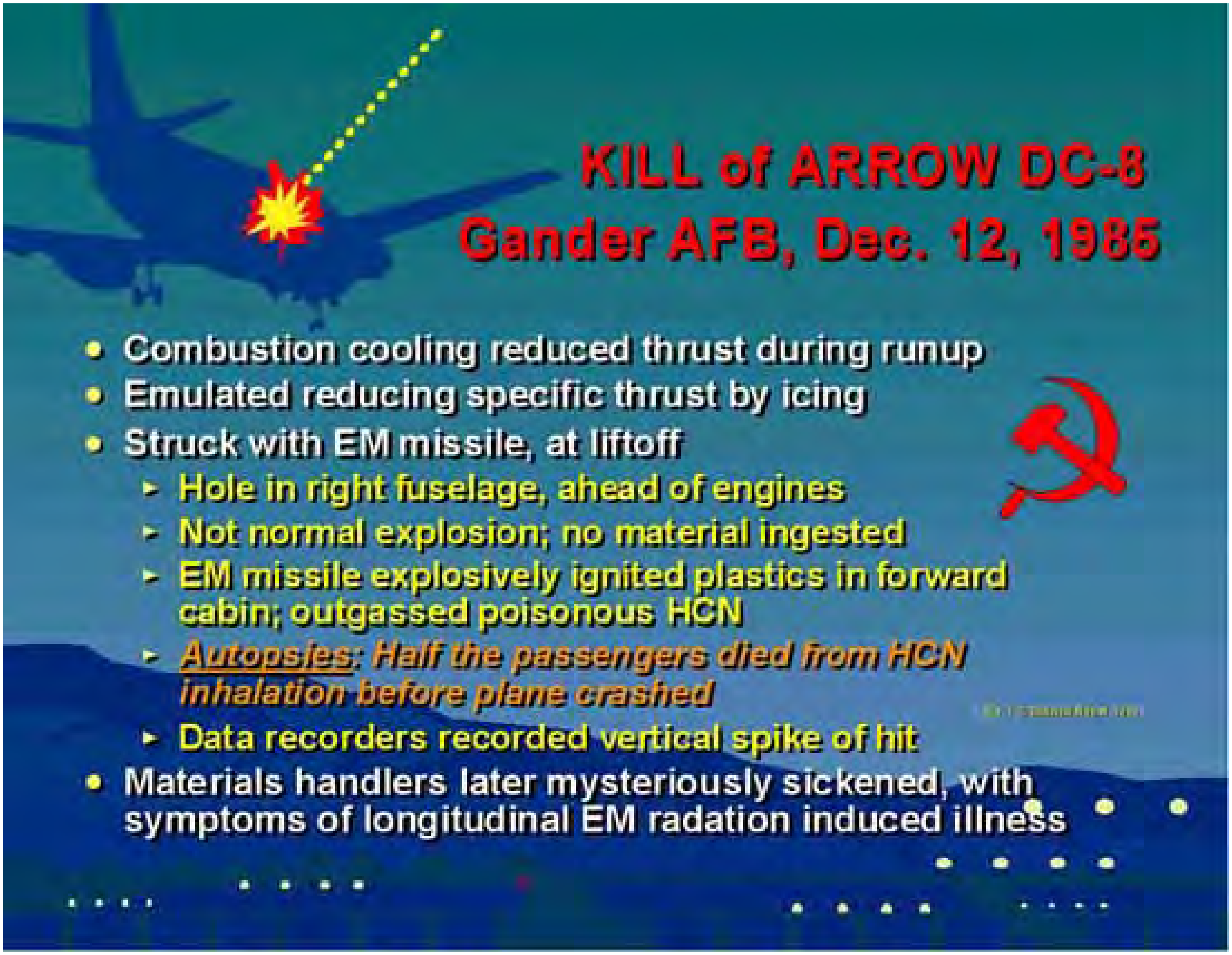

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
    - Hole in right fuselage, ahead of engines
    - Not normal explosion; no material ingested
    - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
    - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
    - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

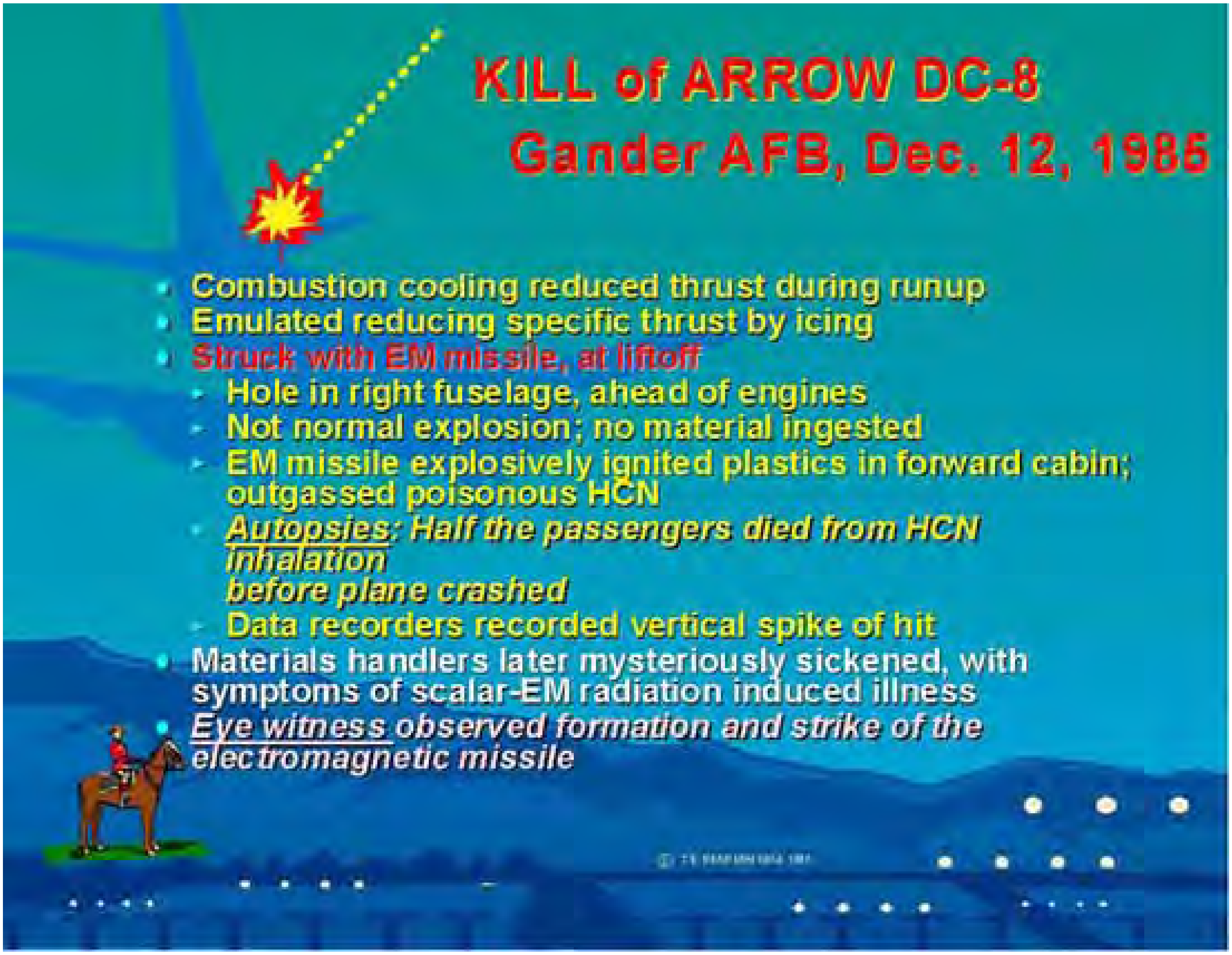

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - ***Autopsies***: *Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

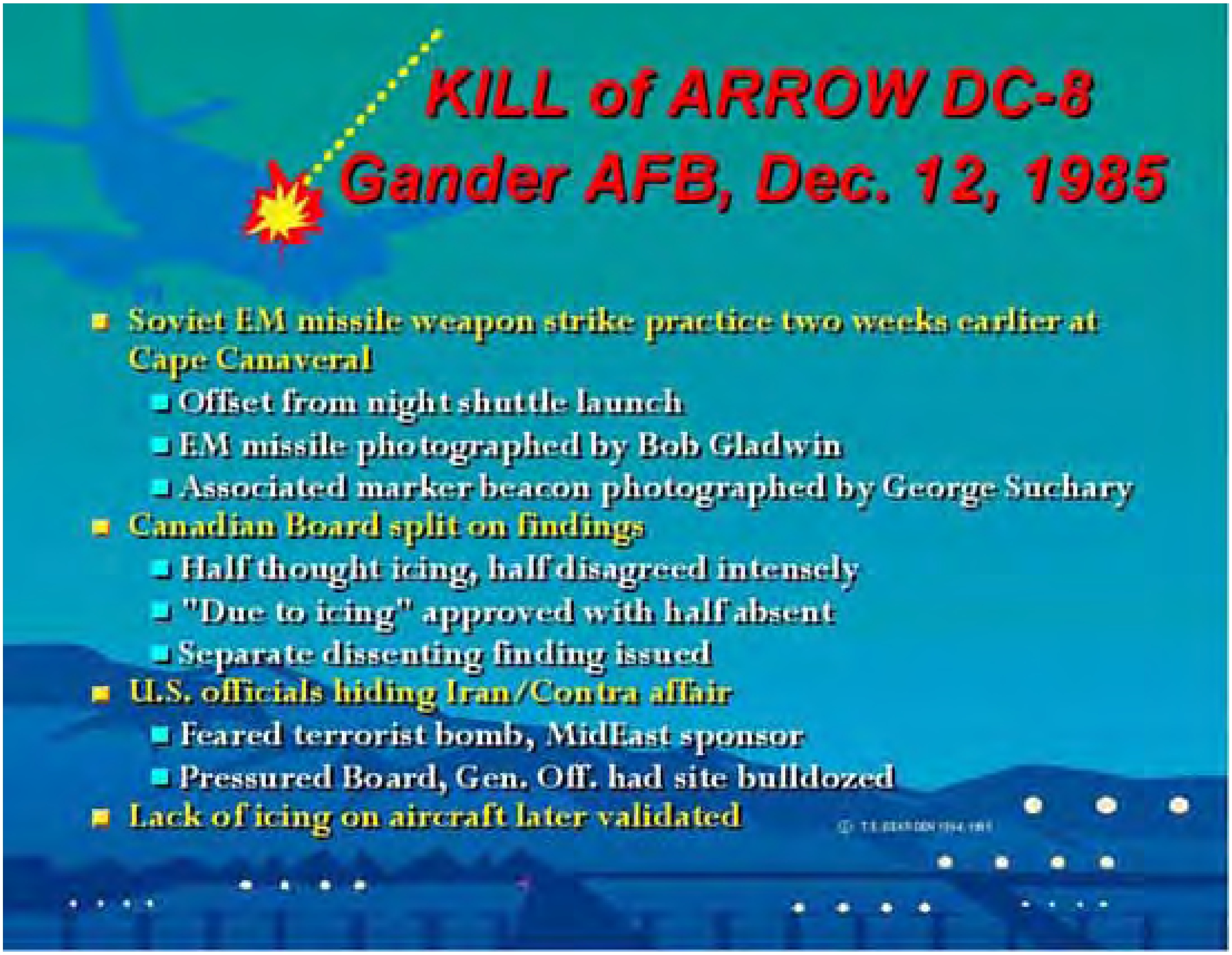

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

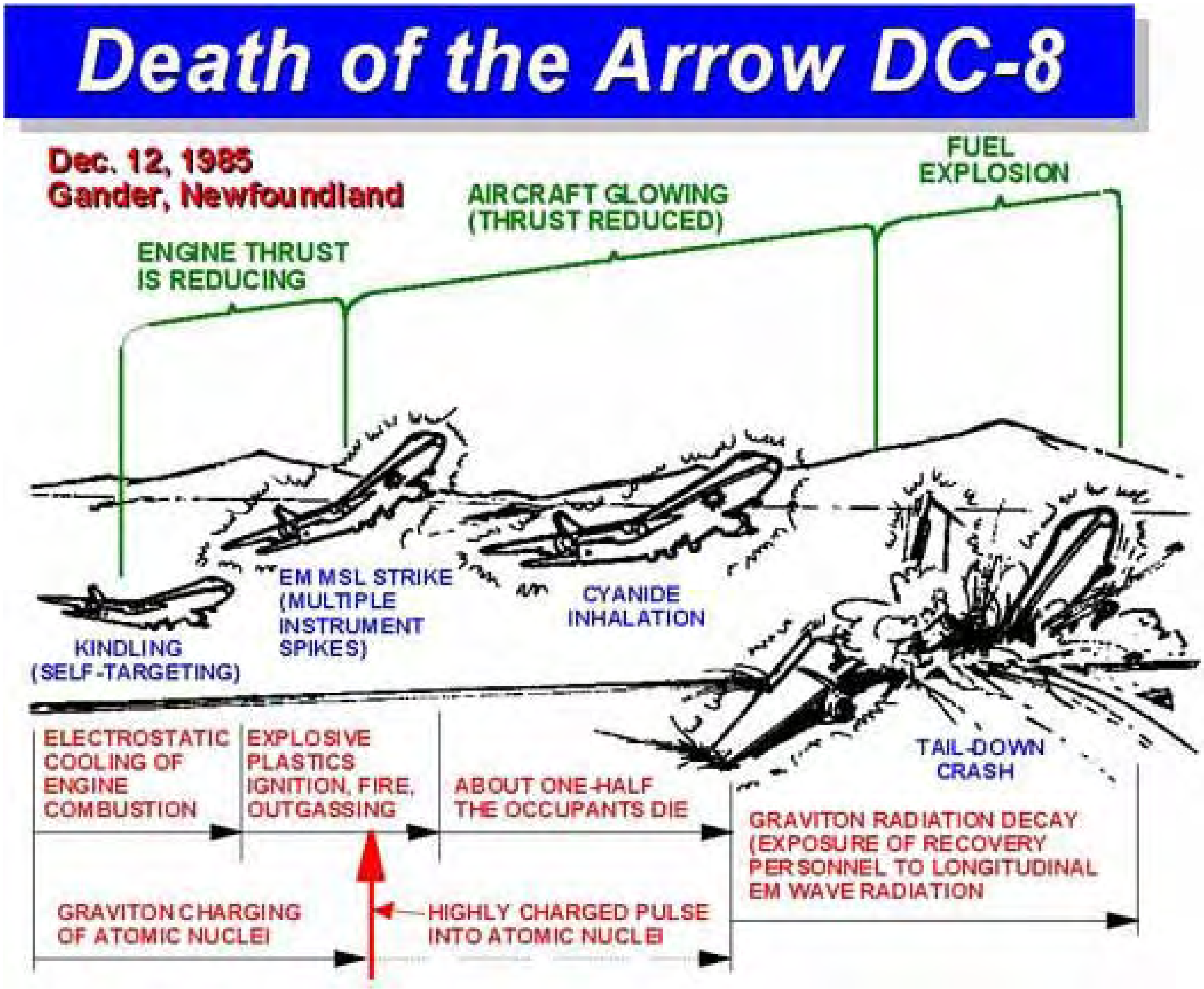
Death of the Arrow DC-8
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST
IS REDUCING
AIRCRAFT GLOWING
(THRUST REDUCED)
FUEL
EXPLOSION
KINDLING
(SELF-TARGETING)
EM MSL STRIKE
(MULTIPLE
INSTRUMENT
SPIKES)
CYANIDE
INHALATION
TAIL-DOWN
CRASH
ELECTROSTATIC
COOLING OF
ENGINE
COMBUSTION
EXPLOSIVE
PLASTICS
IGNITION, FIRE,
OUTGASSING
ABOUT ONE-HALF
THE OCCUPANTS DIE
GRAVITON RADIATION DECAY
(EXPOSURE OF RECOVERY
PERSONNEL TO LONGITUDINAL
EM WAVE RADIATION
GRAVITON CHARGING
OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE
INTO ATOMIC NUCLEI

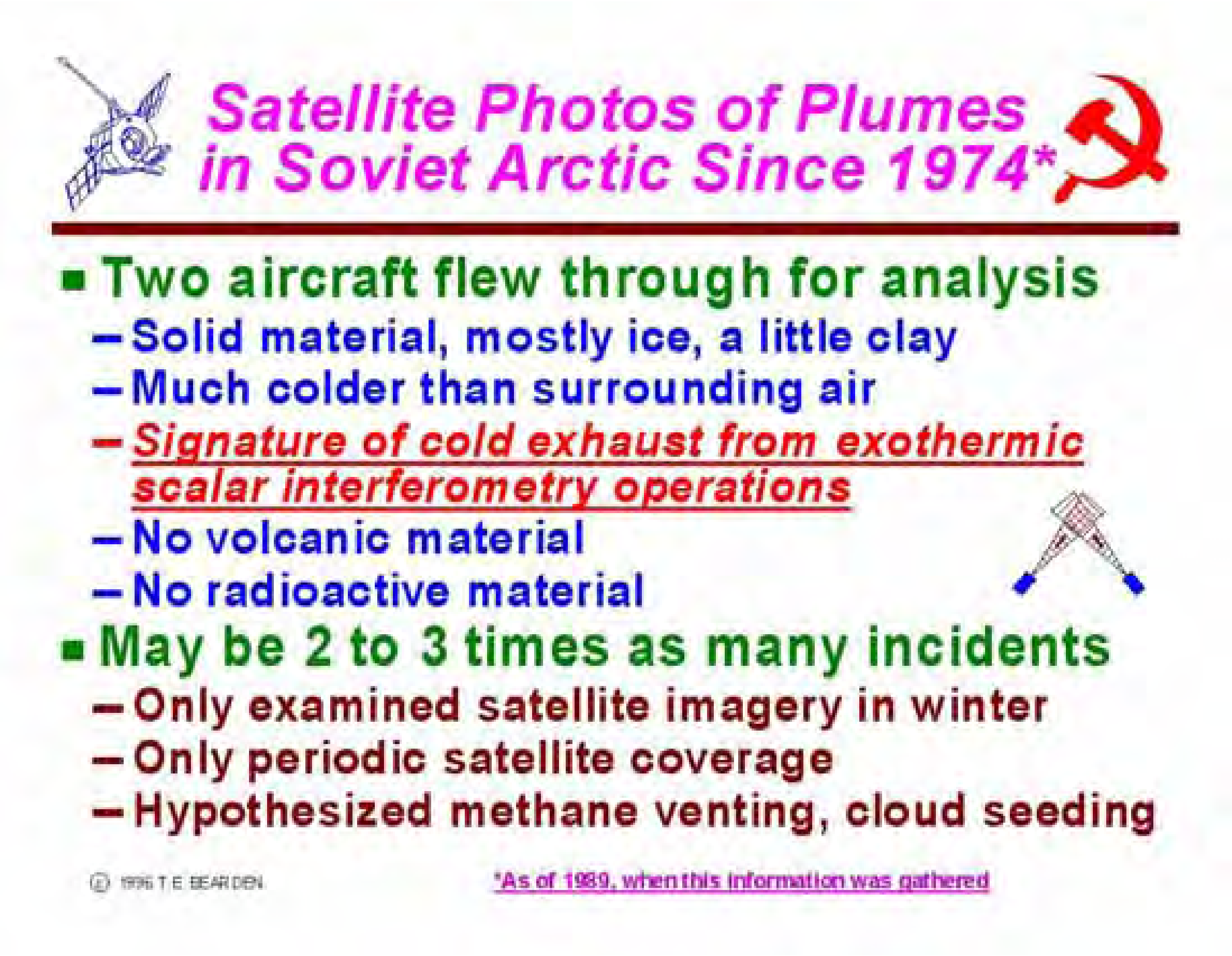

Satellite Photos of Plumes in Soviet Arctic Since 1974*

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO

("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

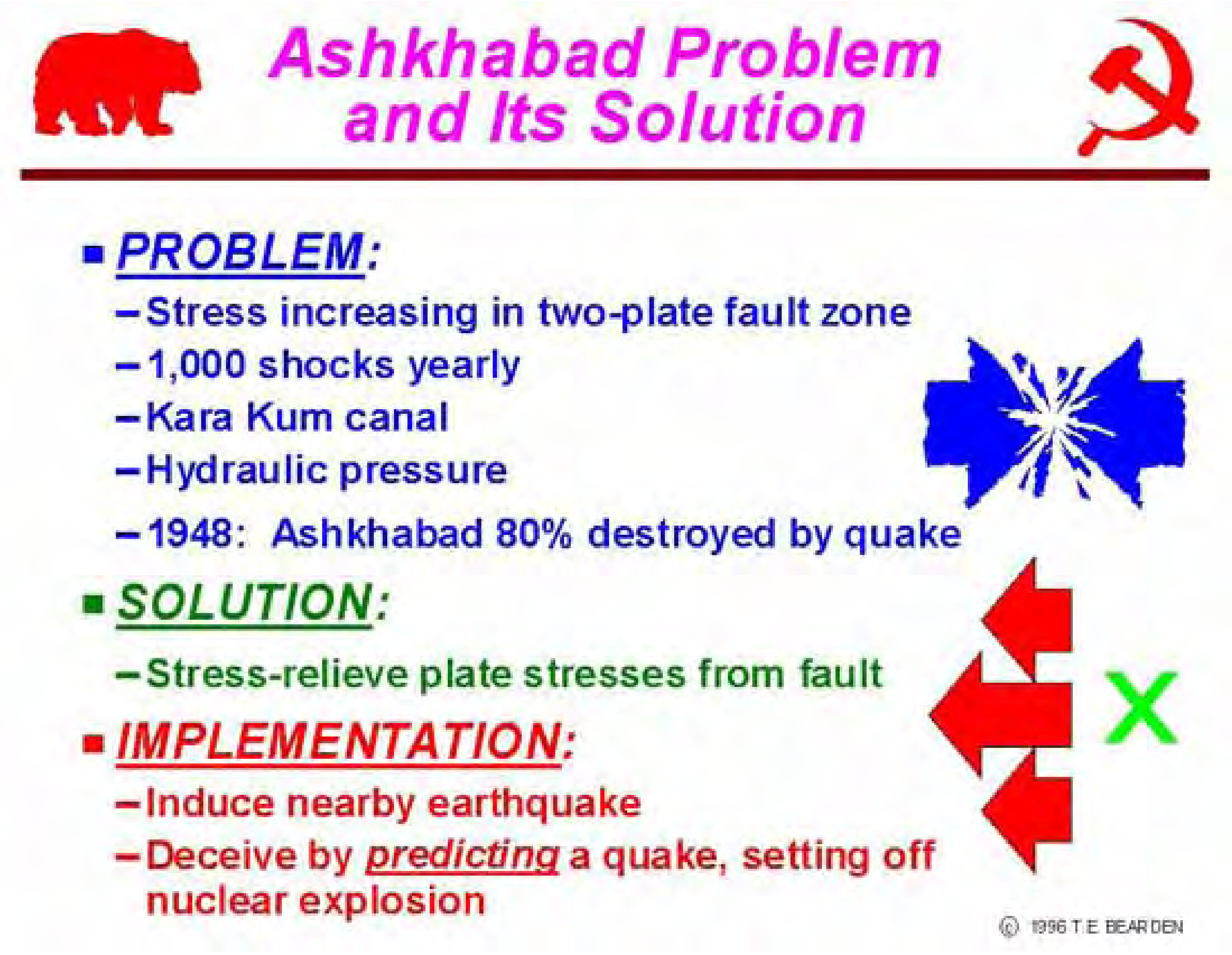

Ashkhabad Problem and Its Solution

- ***PROBLEM:***
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- ***SOLUTION:***
  - Stress-relieve plate stresses from fault
- ***IMPLEMENTATION:***
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

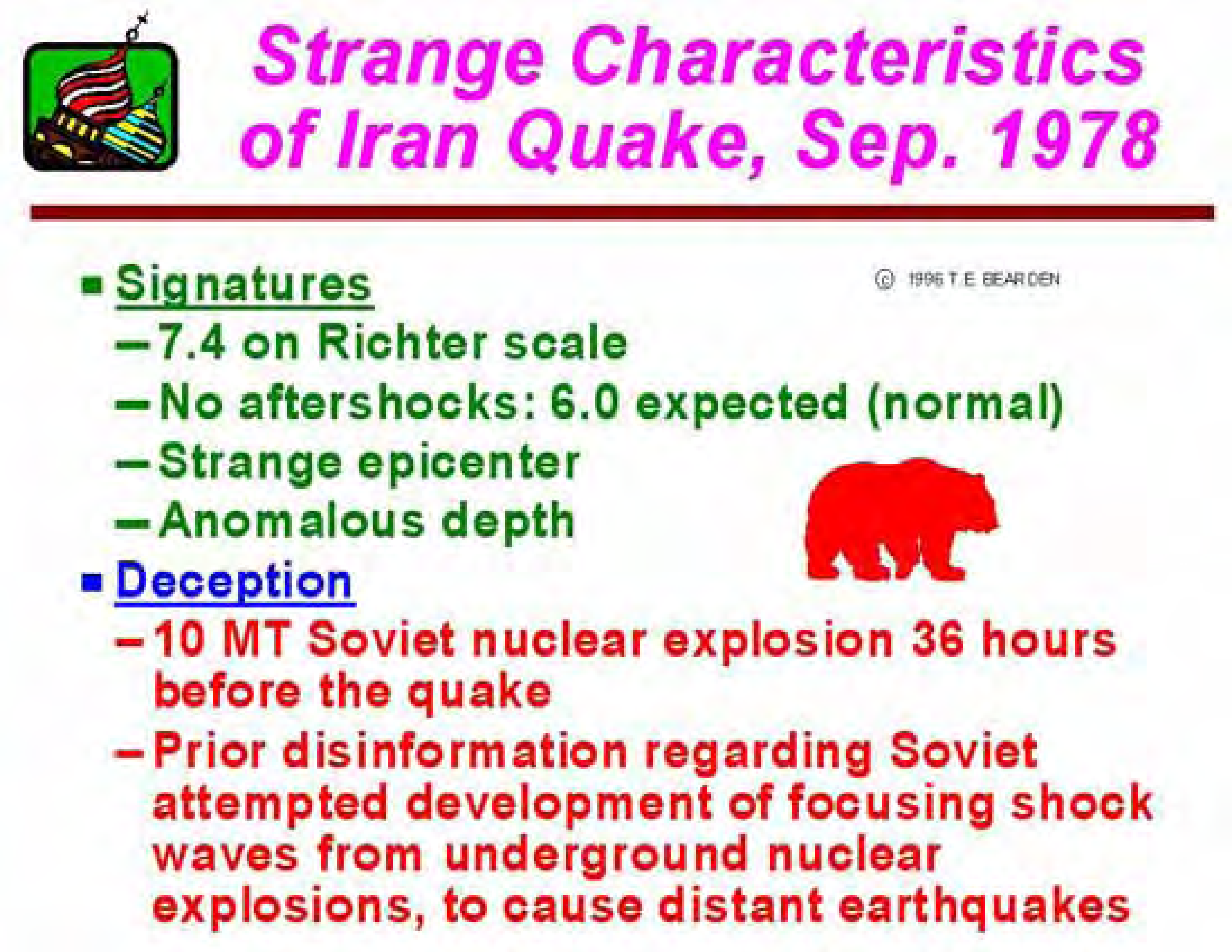

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

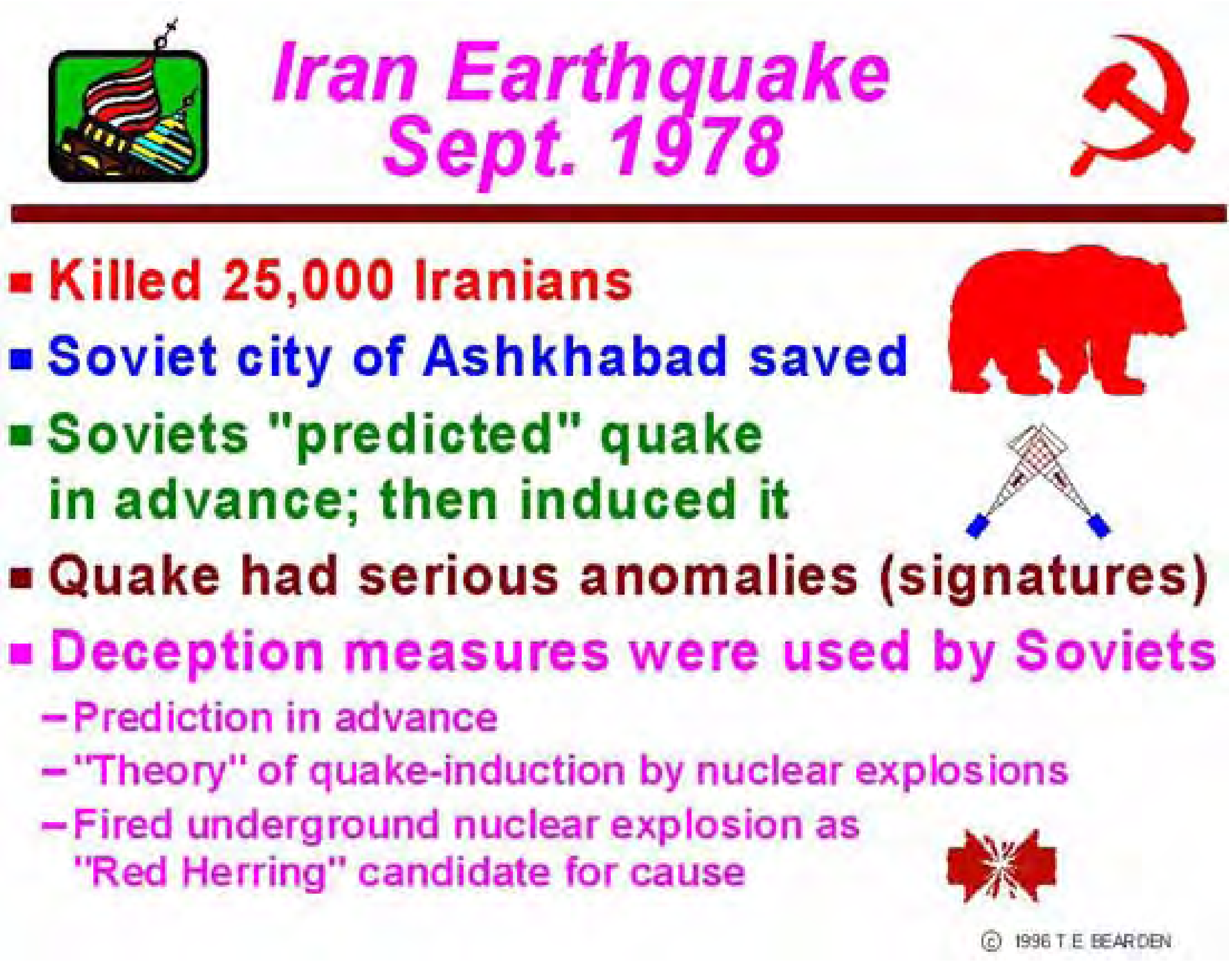

Iran Earthquake
Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - Prediction in advance
  - "Theory" of quake-induction by nuclear explosions
  - Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T.E. BEARDEN

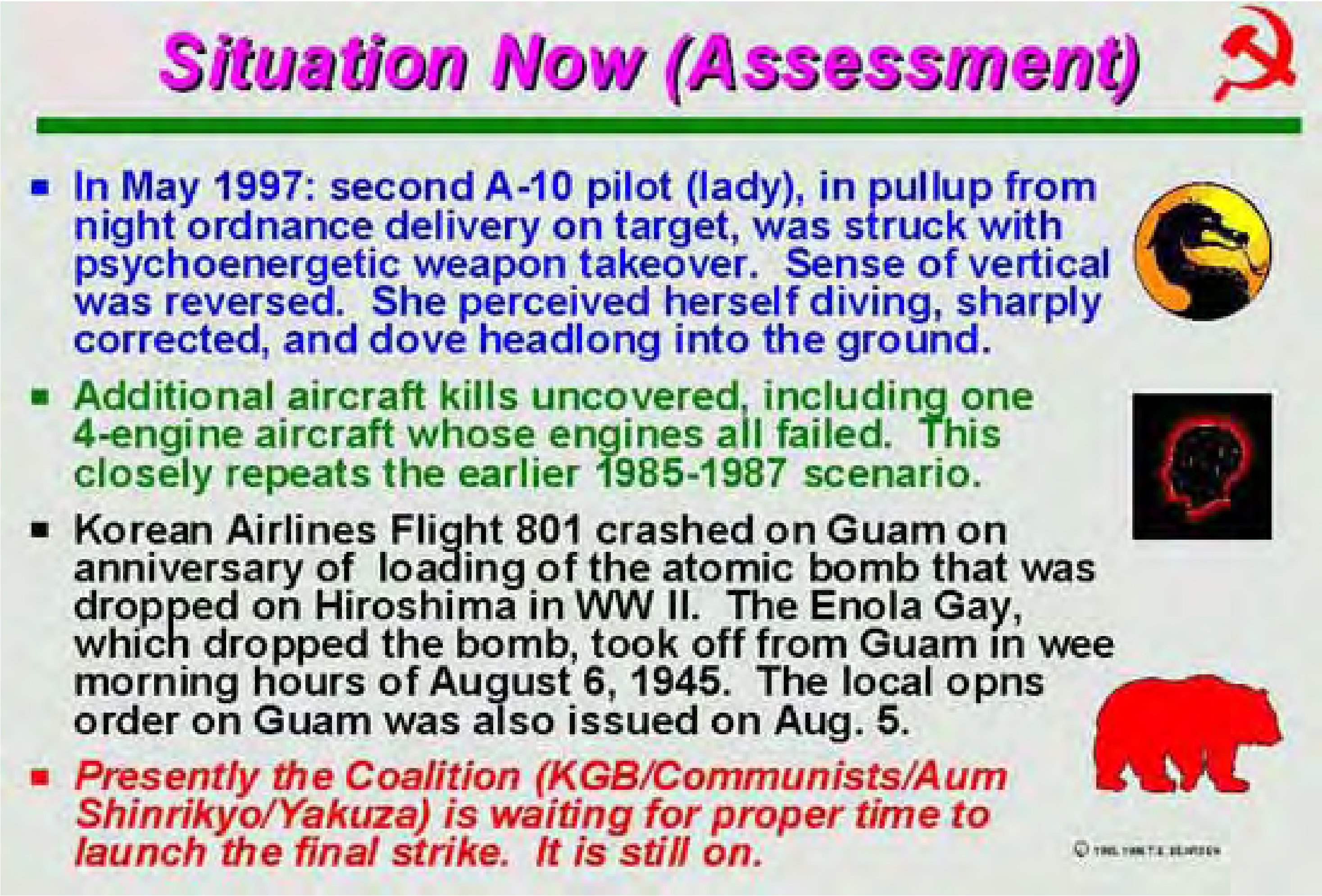

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

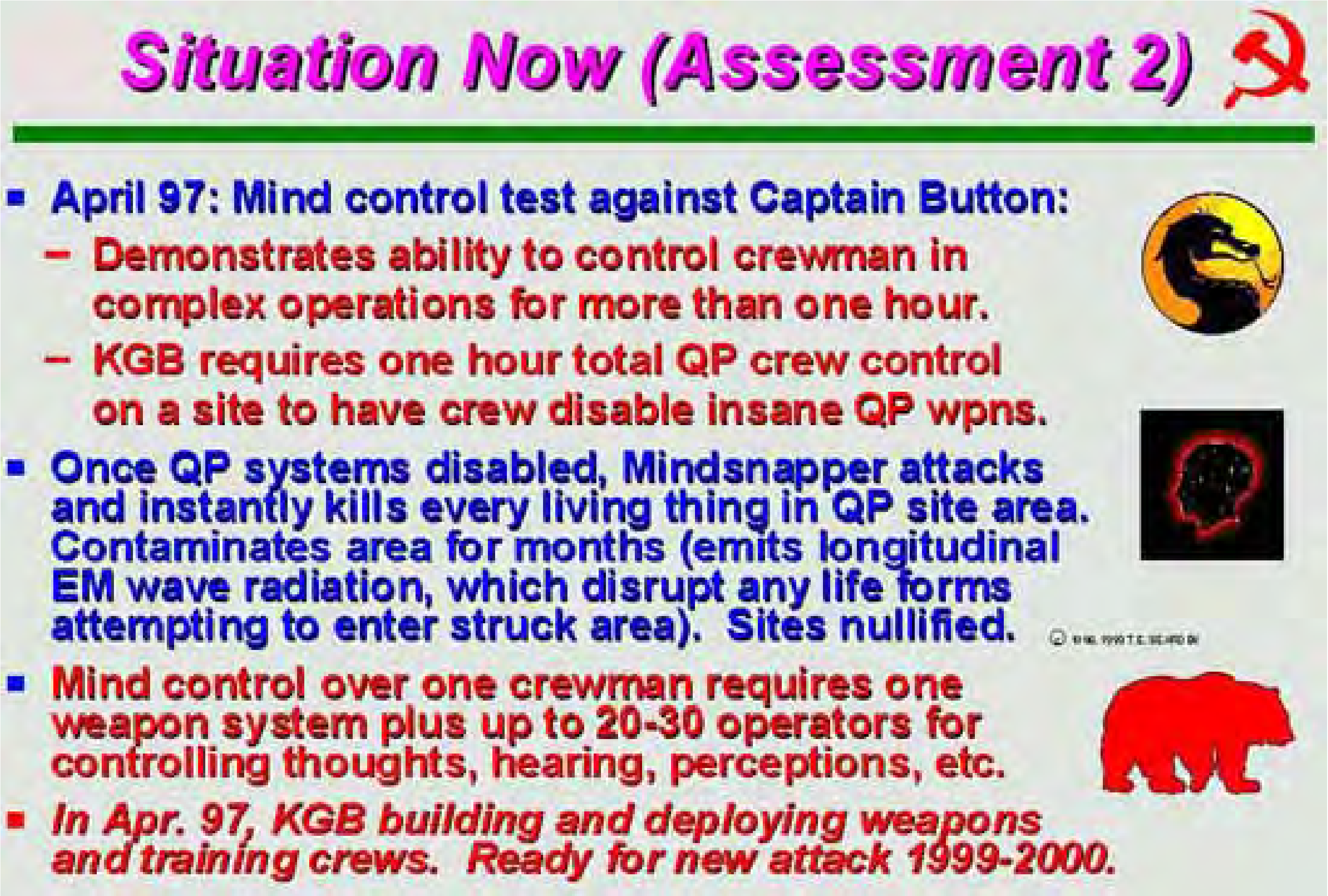

Situation Now (Assessment 2)

- April 97: Mind control test against Captain Button:
  - Demonstrates ability to control crewman in complex operations for more than one hour.
  - KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.
- Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.
- Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.
- *In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.*

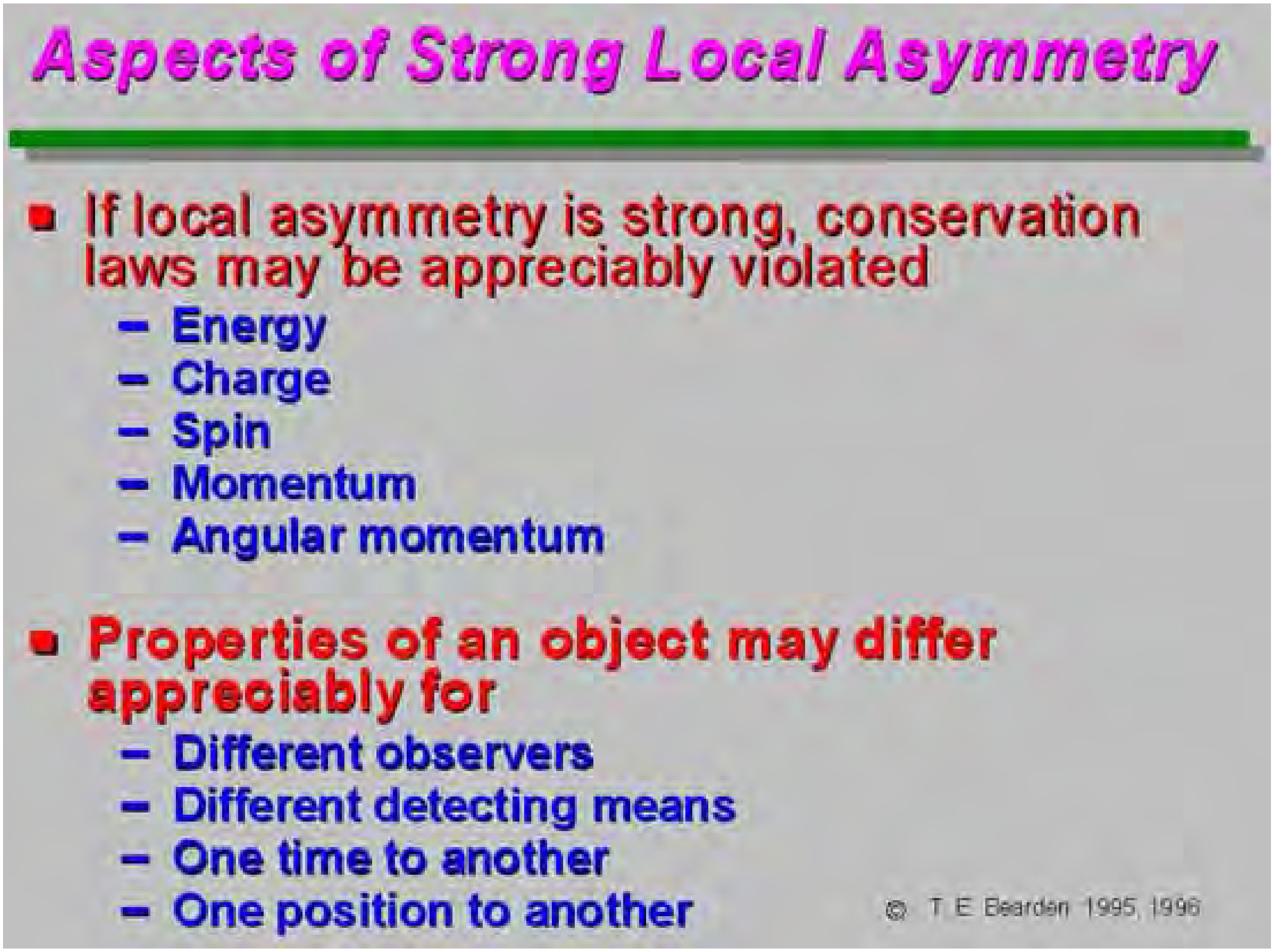

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

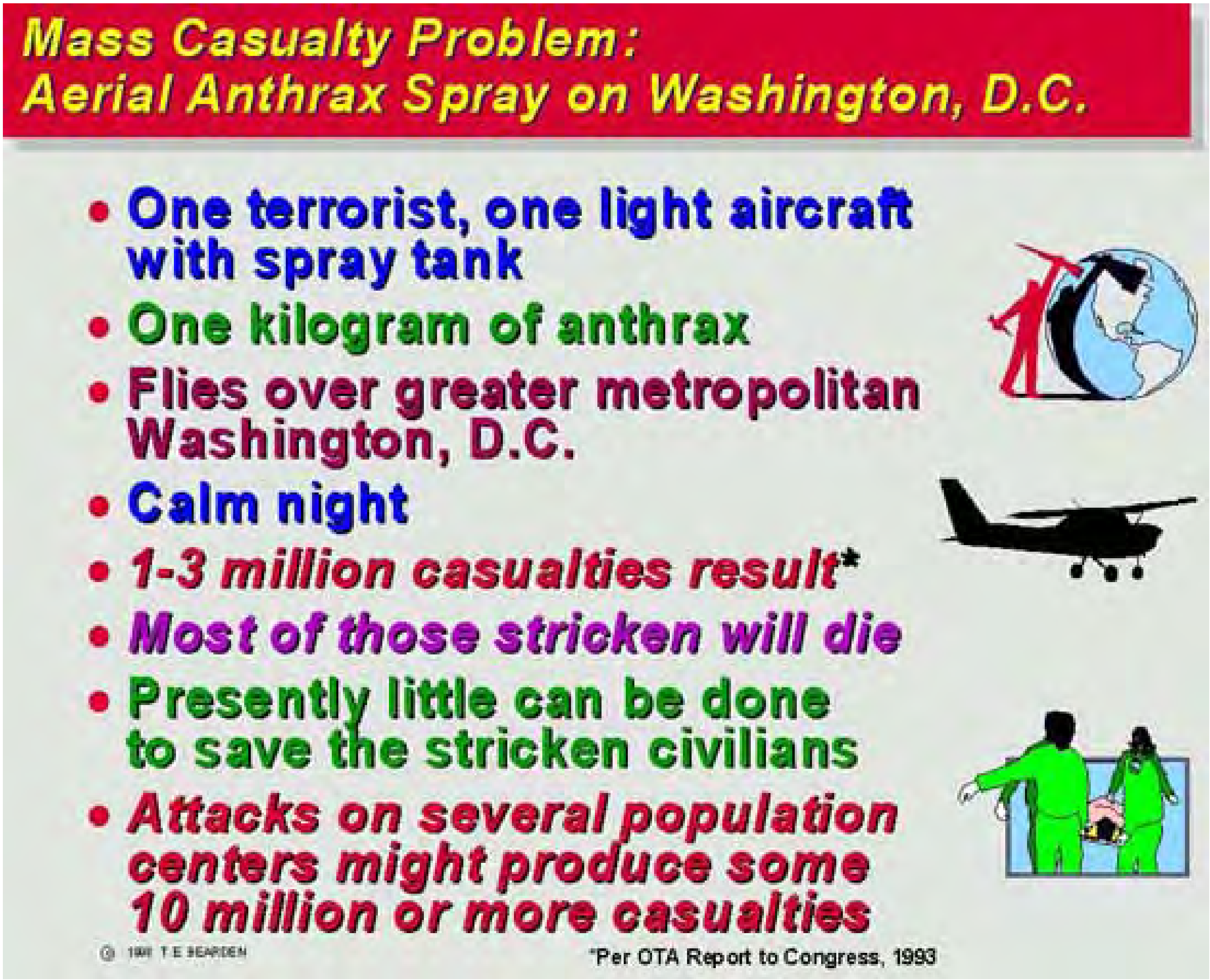

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- **One terrorist, one light aircraft with spray tank**
- **One kilogram of anthrax**
- **Flies over greater metropolitan Washington, D.C.**
- **Calm night**
- ***1-3 million casualties result****
- ***Most of those stricken will die***
- **Presently little can be done to save the stricken civilians**
- ***Attacks on several population centers might produce some 10 million or more casualties***

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

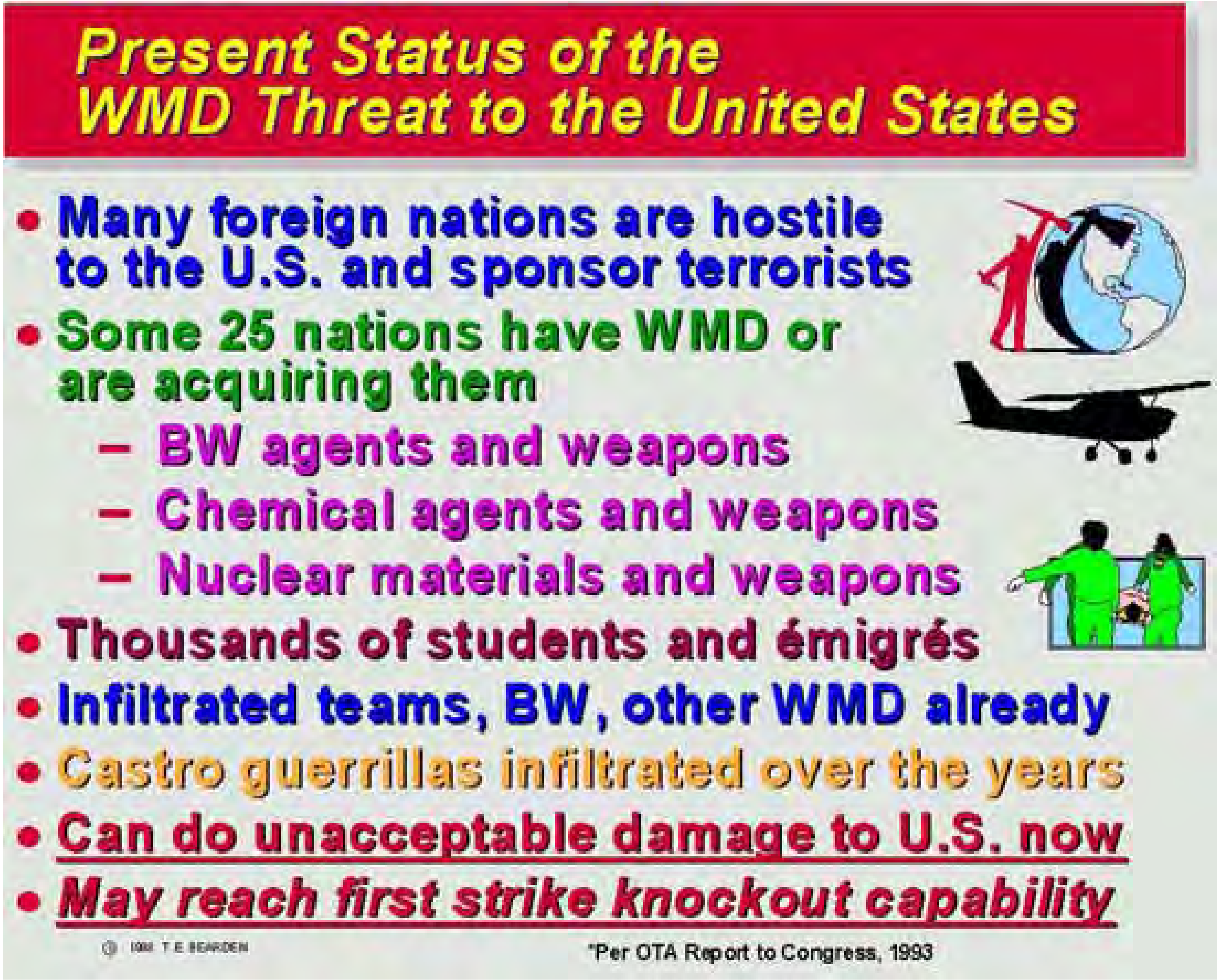

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and émigrés
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

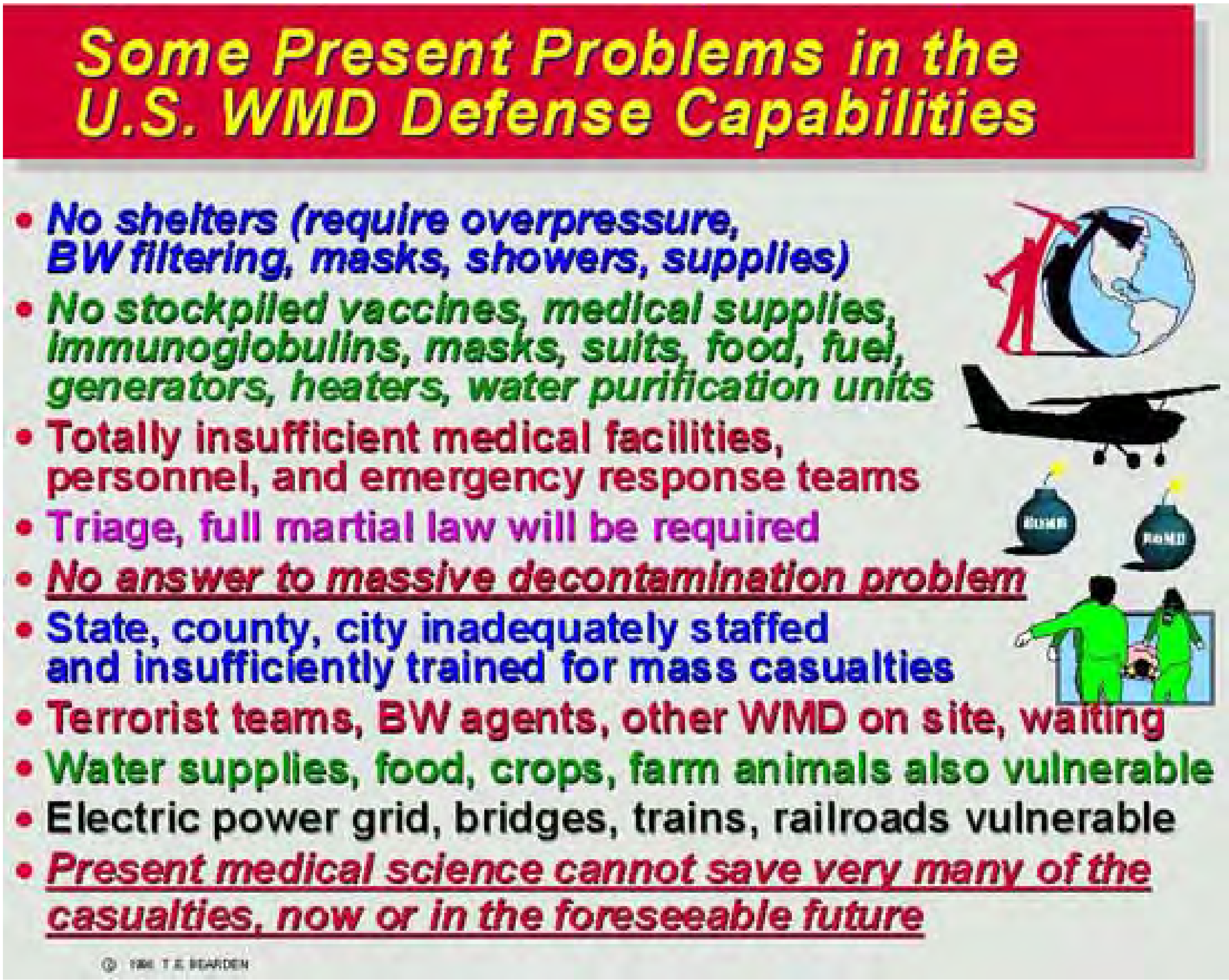

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© [illegible] T.E. BEARDEN

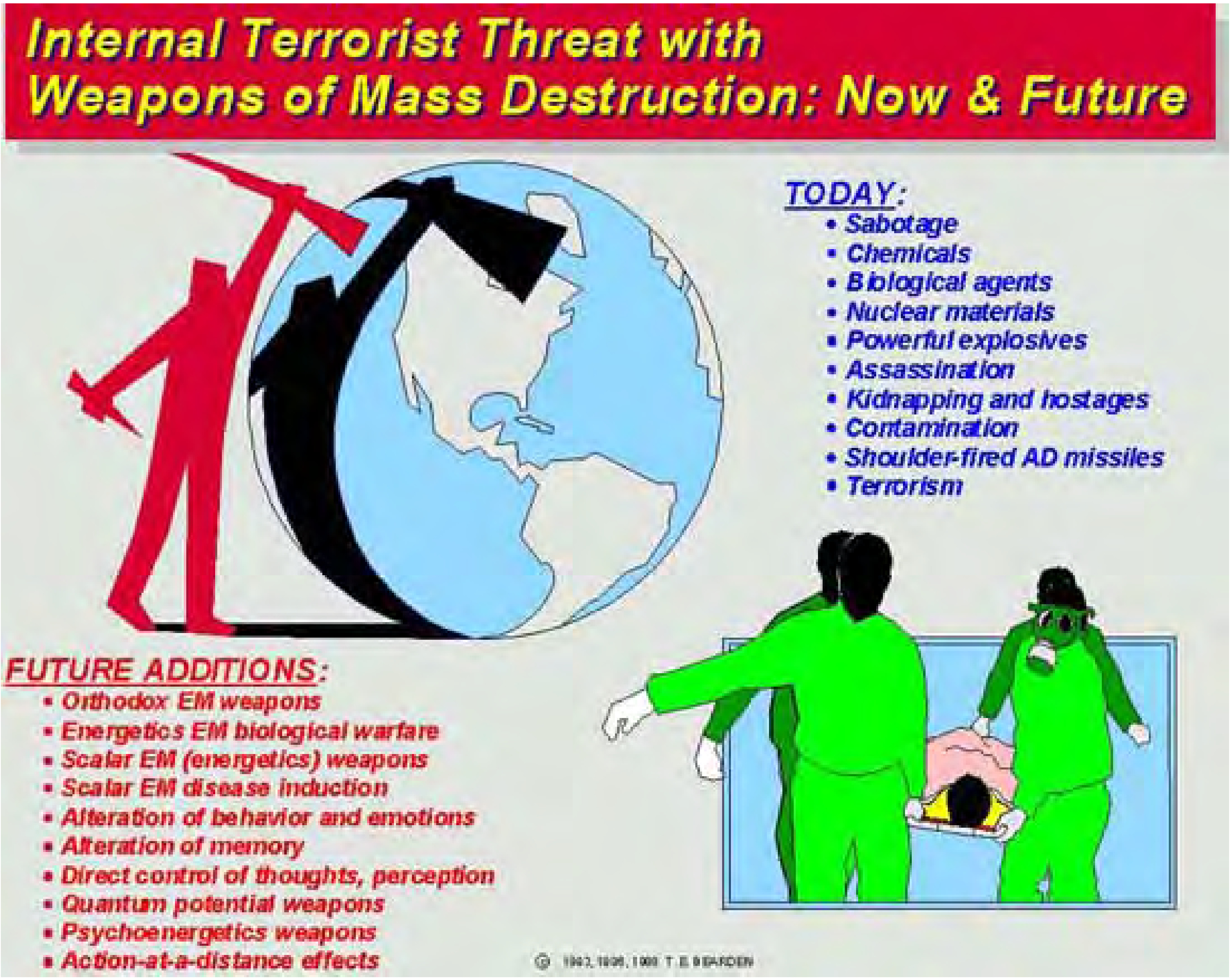

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- Orthodox EM weapons
- Energetics EM biological warfare
- Scalar EM (energetics) weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct control of thoughts, perception
- Quantum potential weapons
- Psychoenergetics weapons
- Action-at-a-distance effects

© [illegible] T. E. BEARDEN

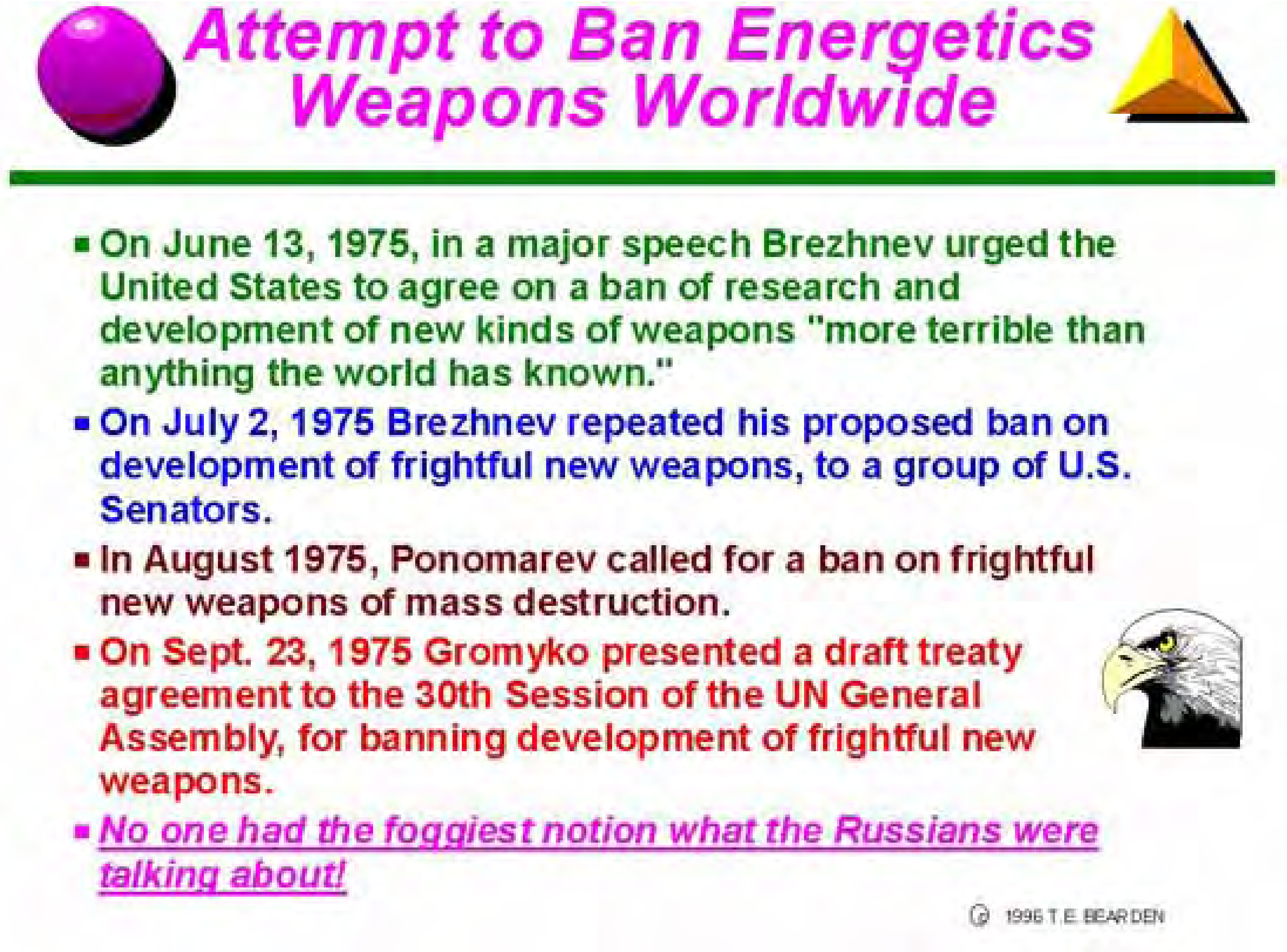

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

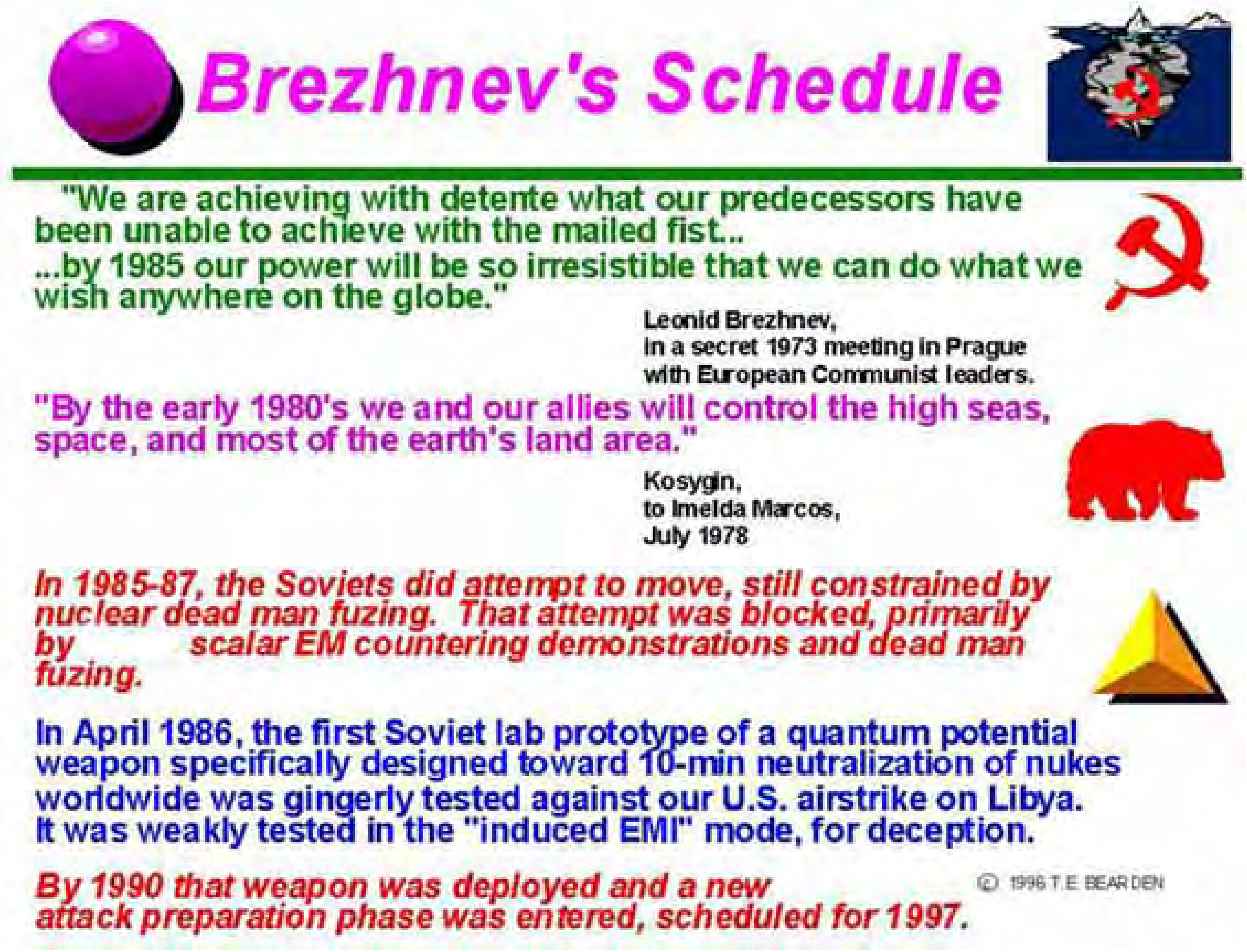

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist..
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
in a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

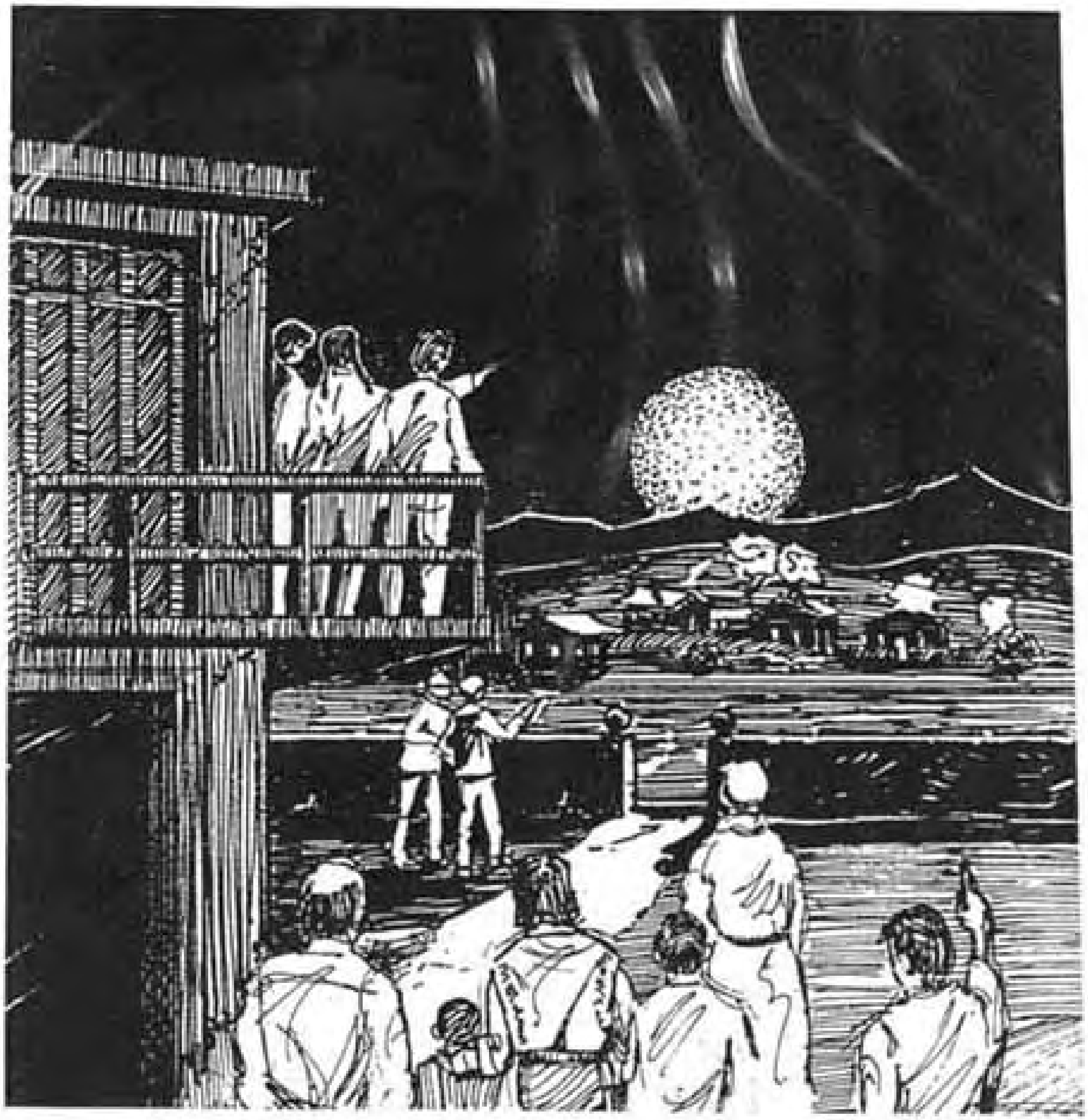

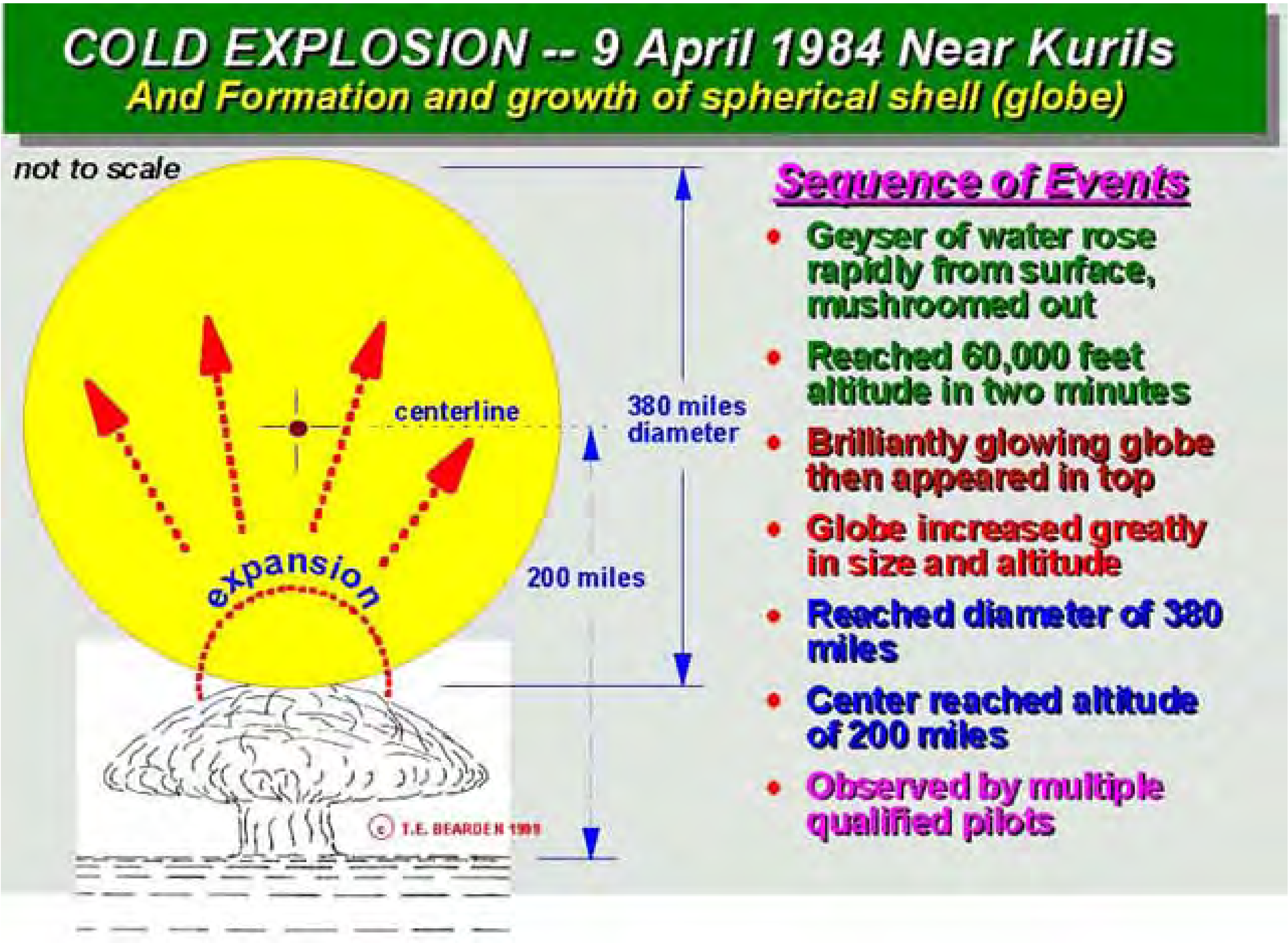

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles
diameter
200 miles
expansion
© T.E. BEARDEN

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

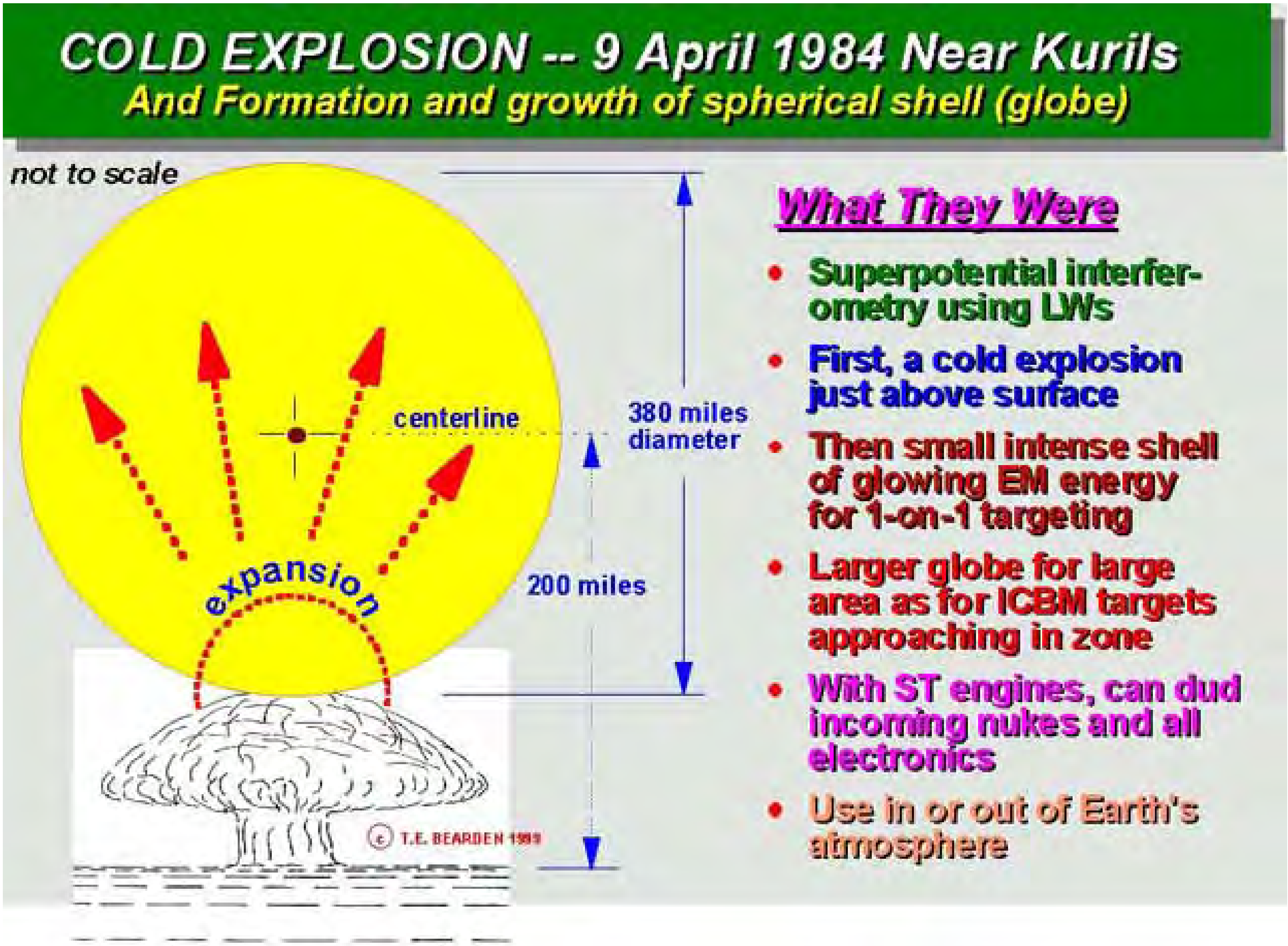

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
expansion
380 miles diameter
200 miles
© T.E. BEARDEN 1999

What They Were

- Superpotential interferometry using LWs
- First, a cold explosion just above surface
- Then small intense shell of glowing EM energy for 1-on-1 targeting
- Larger globe for large area as for ICBM targets approaching in zone
- With ST engines, can dud incoming nukes and all electronics
- Use in or out of Earth's atmosphere

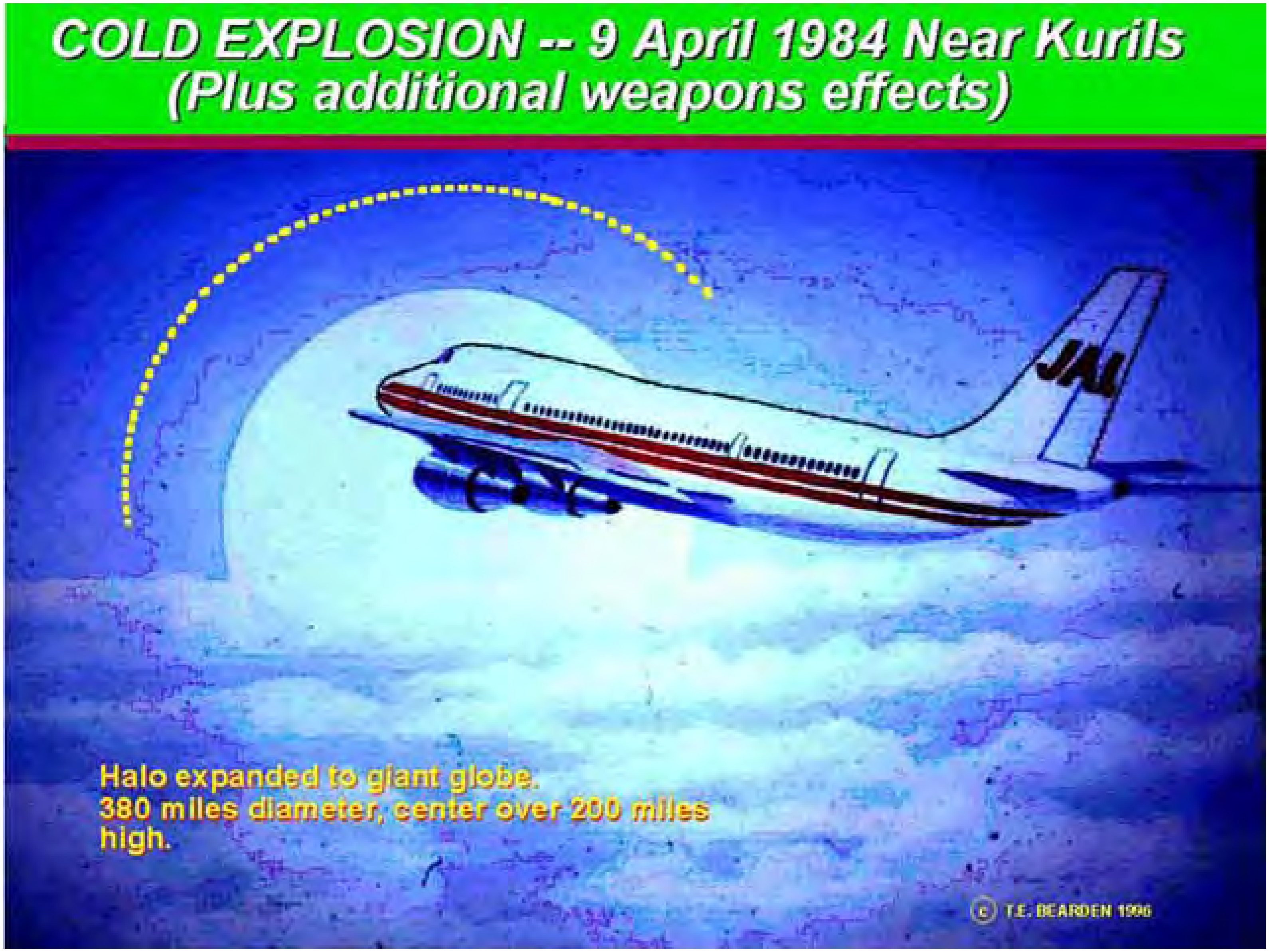

COLD EXPLOSION -- 9 April 1984 Near Kurils (Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

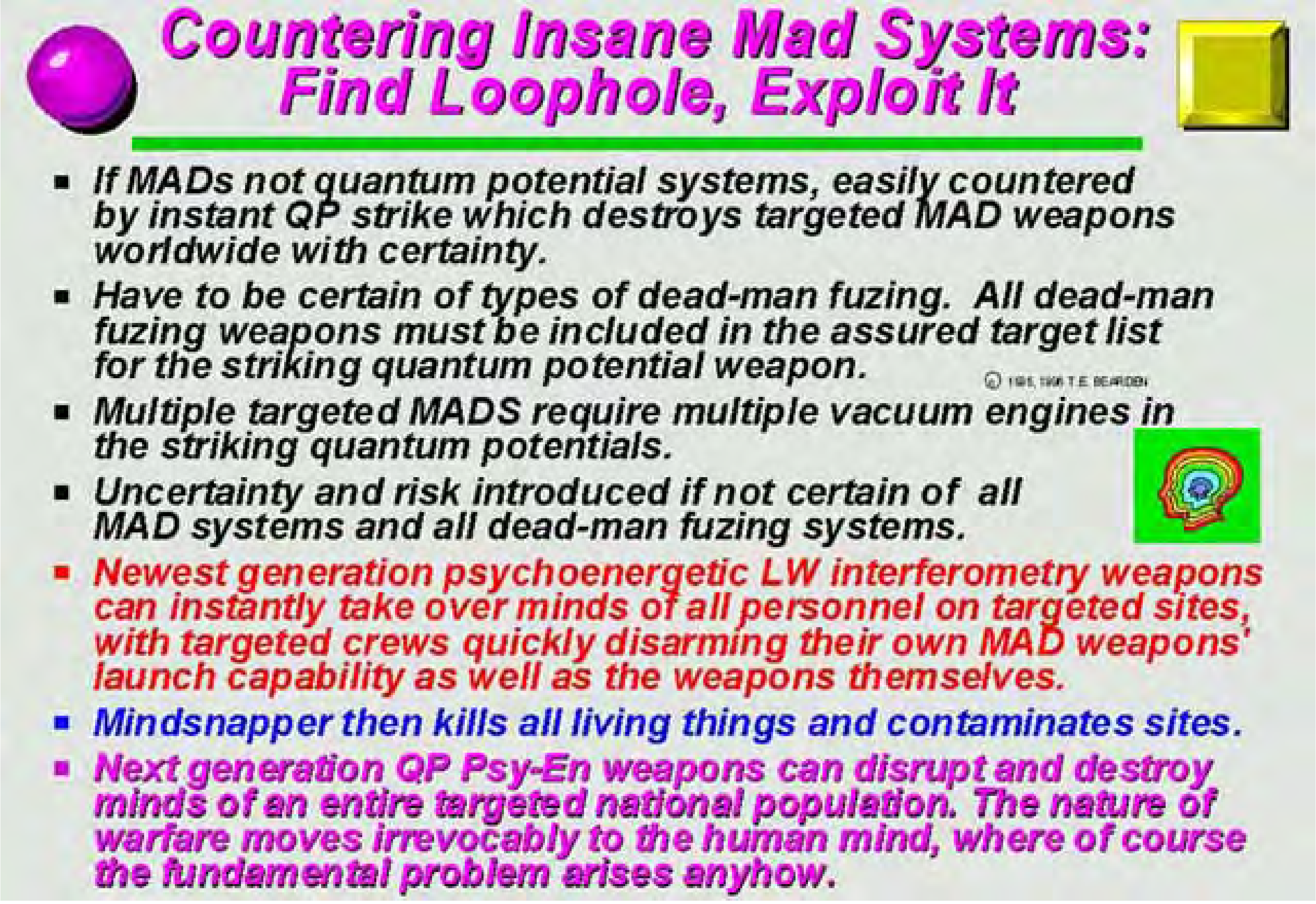

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*

© 1995, 1996 T.E. BEARDEN

- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

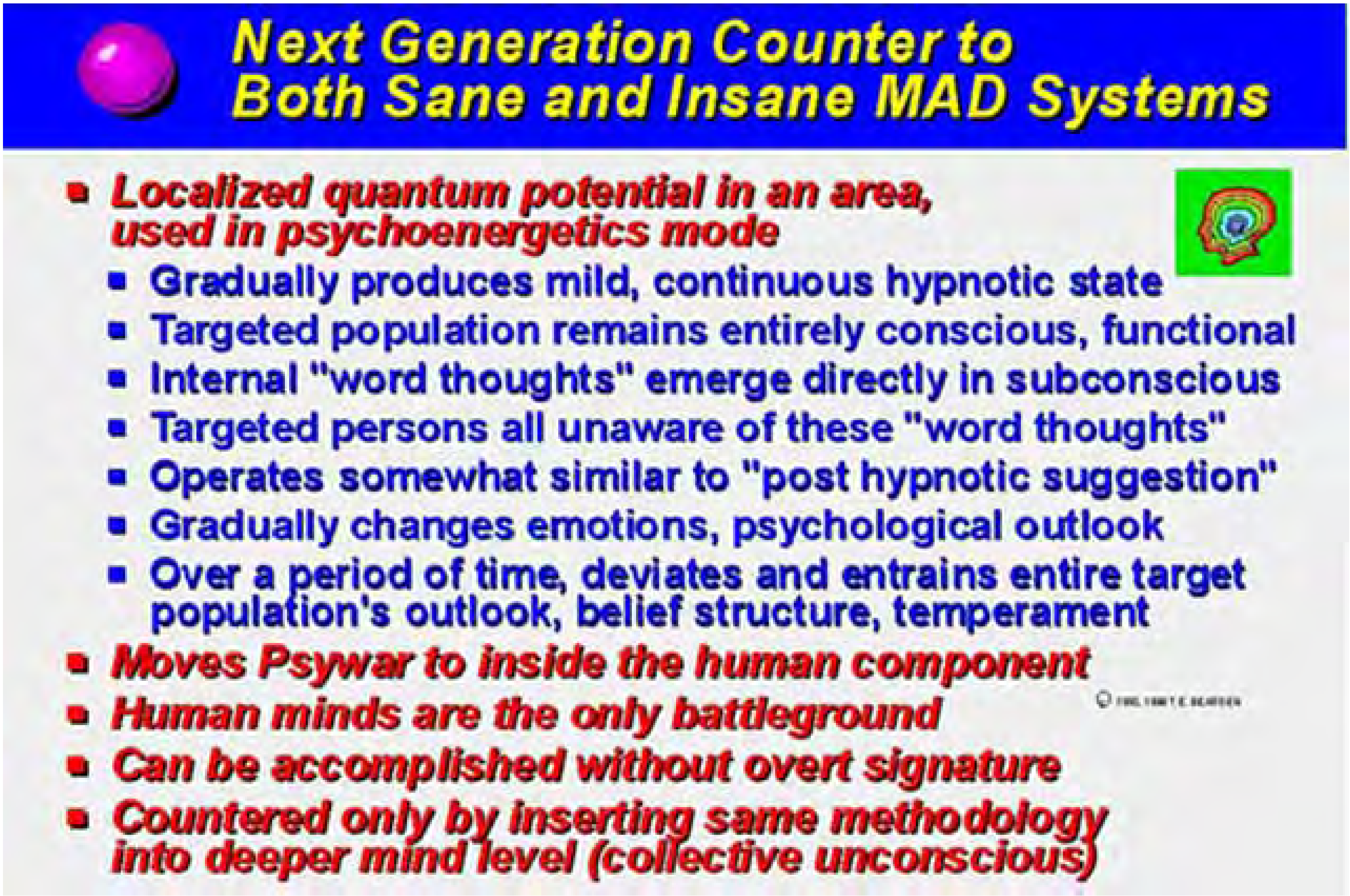

Next Generation Counter to Both Sane and Insane MAD Systems

- ***Localized quantum potential in an area, used in psychoenergetics mode***
  - **Gradually produces mild, continuous hypnotic state**
  - **Targeted population remains entirely conscious, functional**
  - **Internal "word thoughts" emerge directly in subconscious**
  - **Targeted persons all unaware of these "word thoughts"**
  - **Operates somewhat similar to "post hypnotic suggestion"**
  - **Gradually changes emotions, psychological outlook**
  - **Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament**
- ***Moves Psywar to inside the human component***
- ***Human minds are the only battleground***
- ***Can be accomplished without overt signature***
- ***Countered only by inserting same methodology into deeper mind level (collective unconscious)***

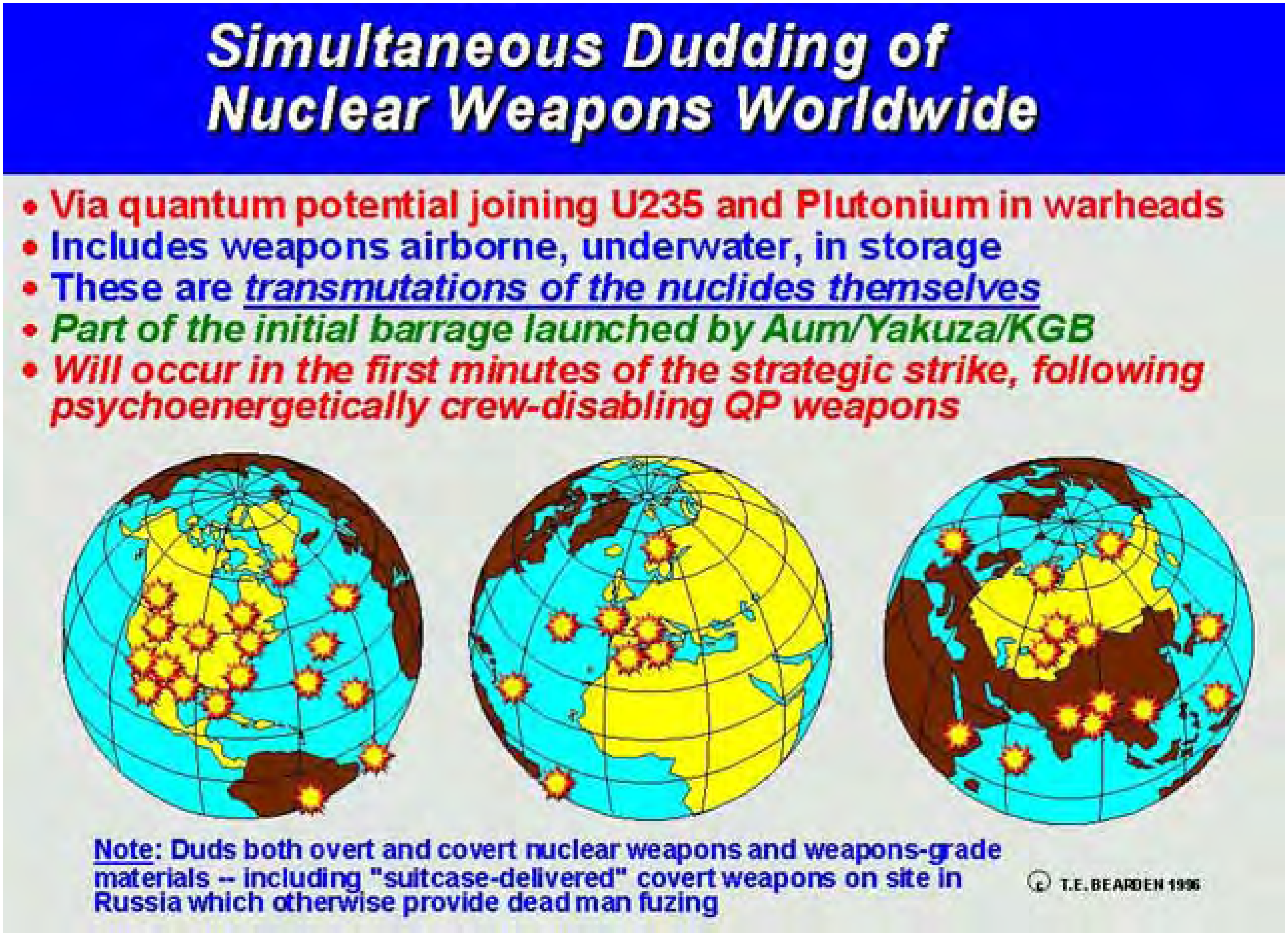

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

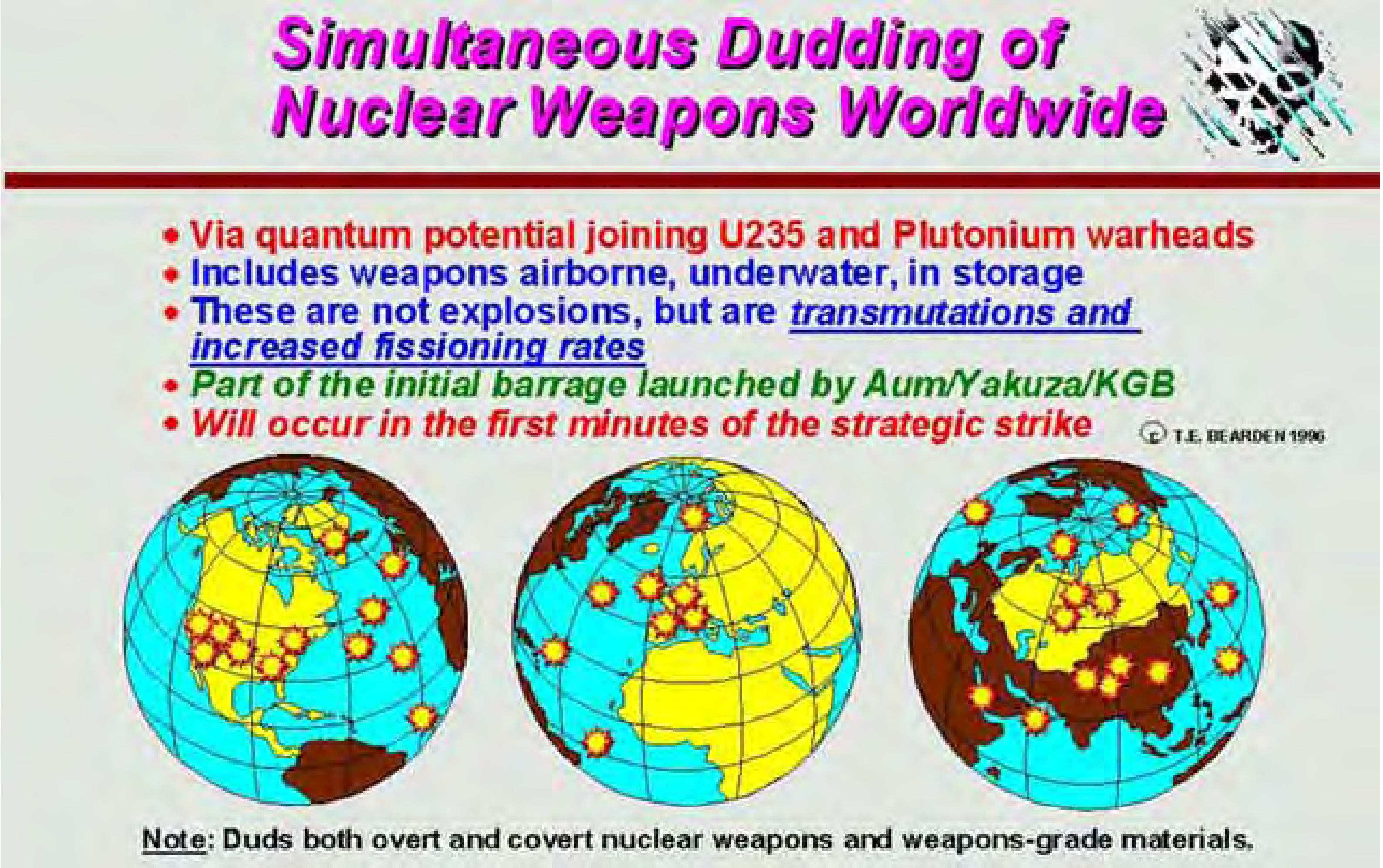

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are ***transmutations and increased fissioning rates***
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

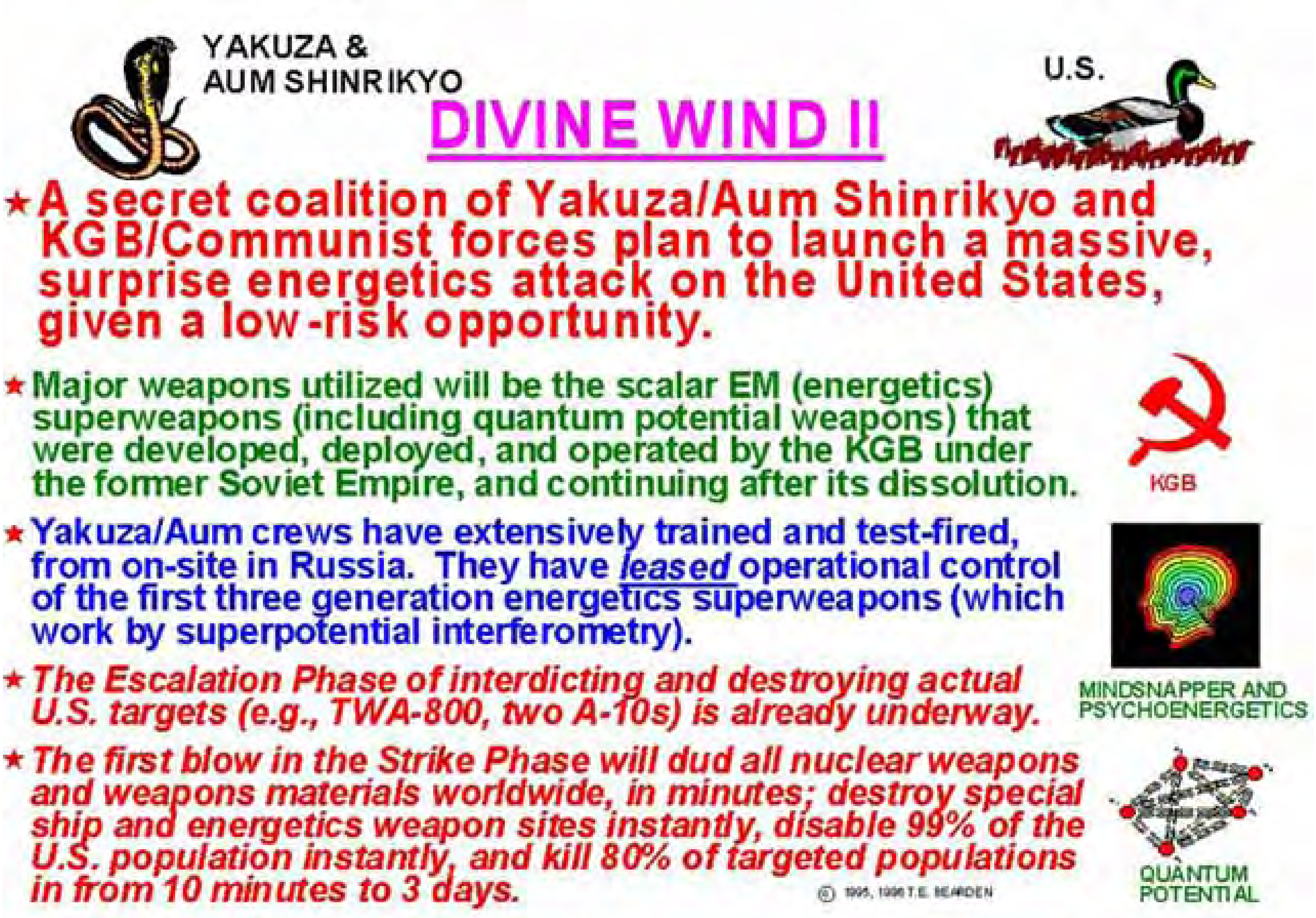

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ **A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.**
- ★ Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
- ★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).
- ★ *The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.*
- ★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.*

KGB

MINDSNAPPER AND
PSYCHOENERGETICS

QUANTUM
POTENTIAL

© 1995, 1996 T.E. BEARDEN

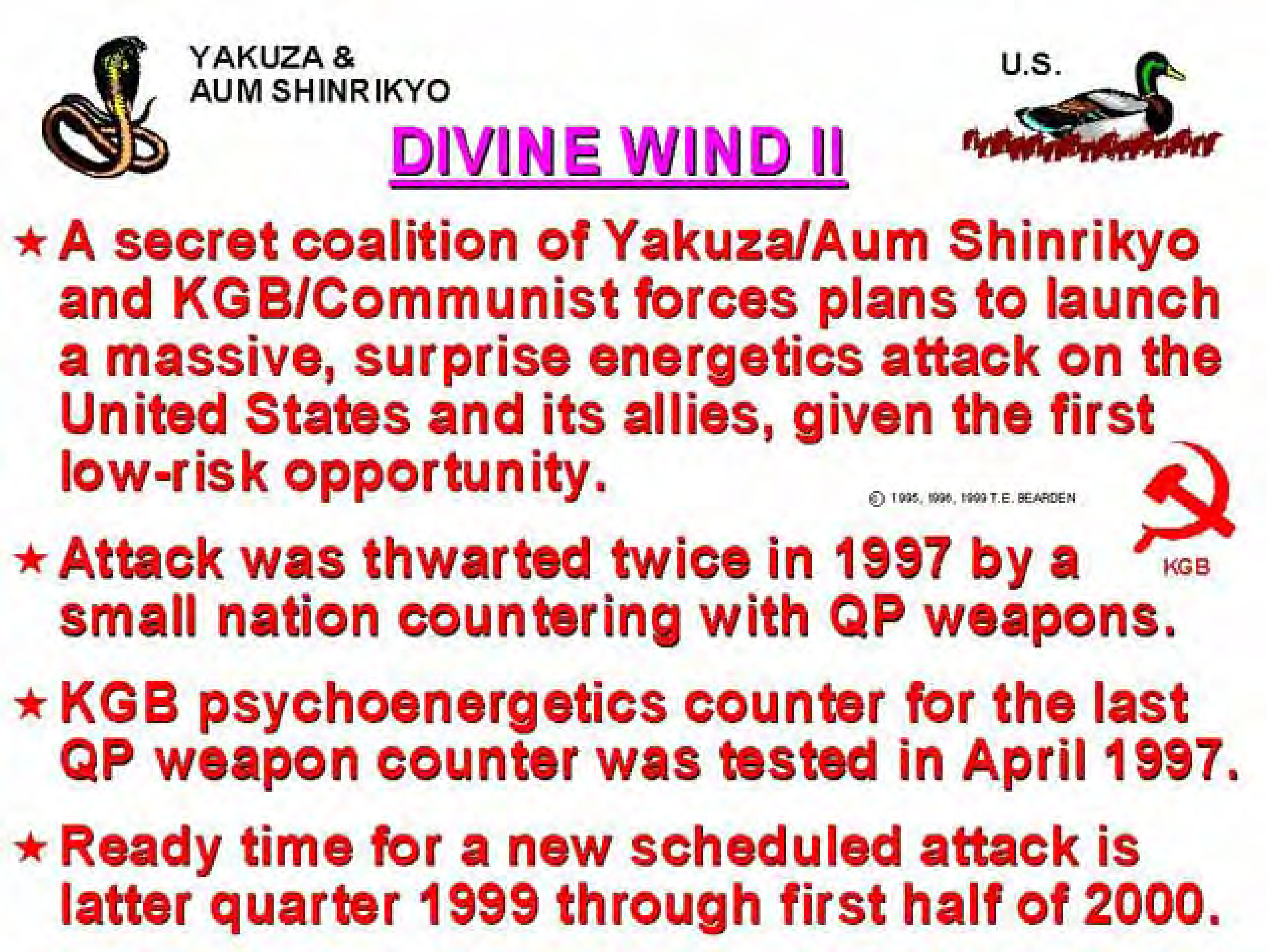
YAKUZA &
AUM SHINRIKYO
U.S.
KGB

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

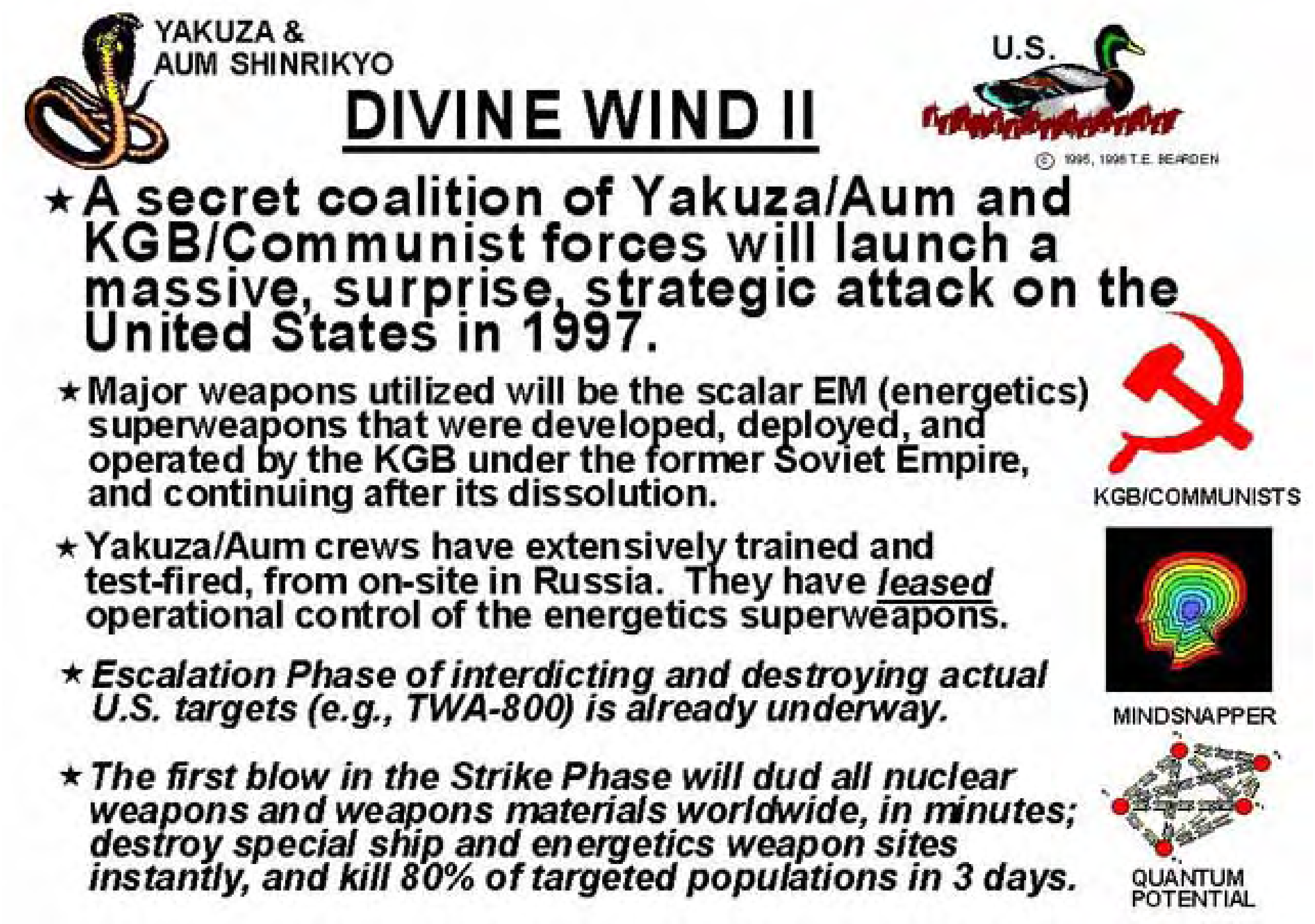

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

© 1995, 1996 T.E. BEARDEN

- **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**
  - Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
  - Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.
  - *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*
  - *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

KGB/COMMUNISTS

MINDSNAPPER

QUANTUM
POTENTIAL

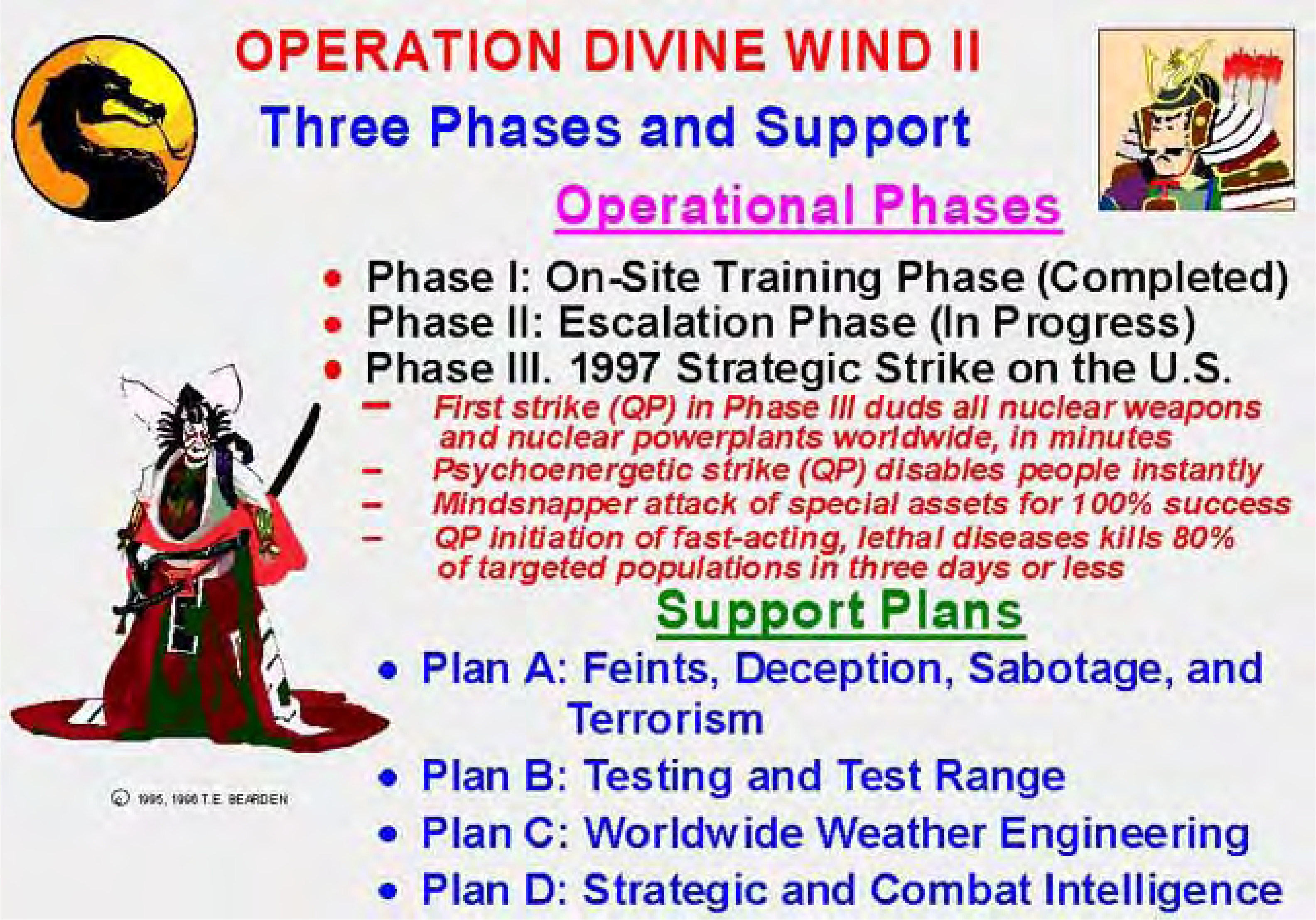

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

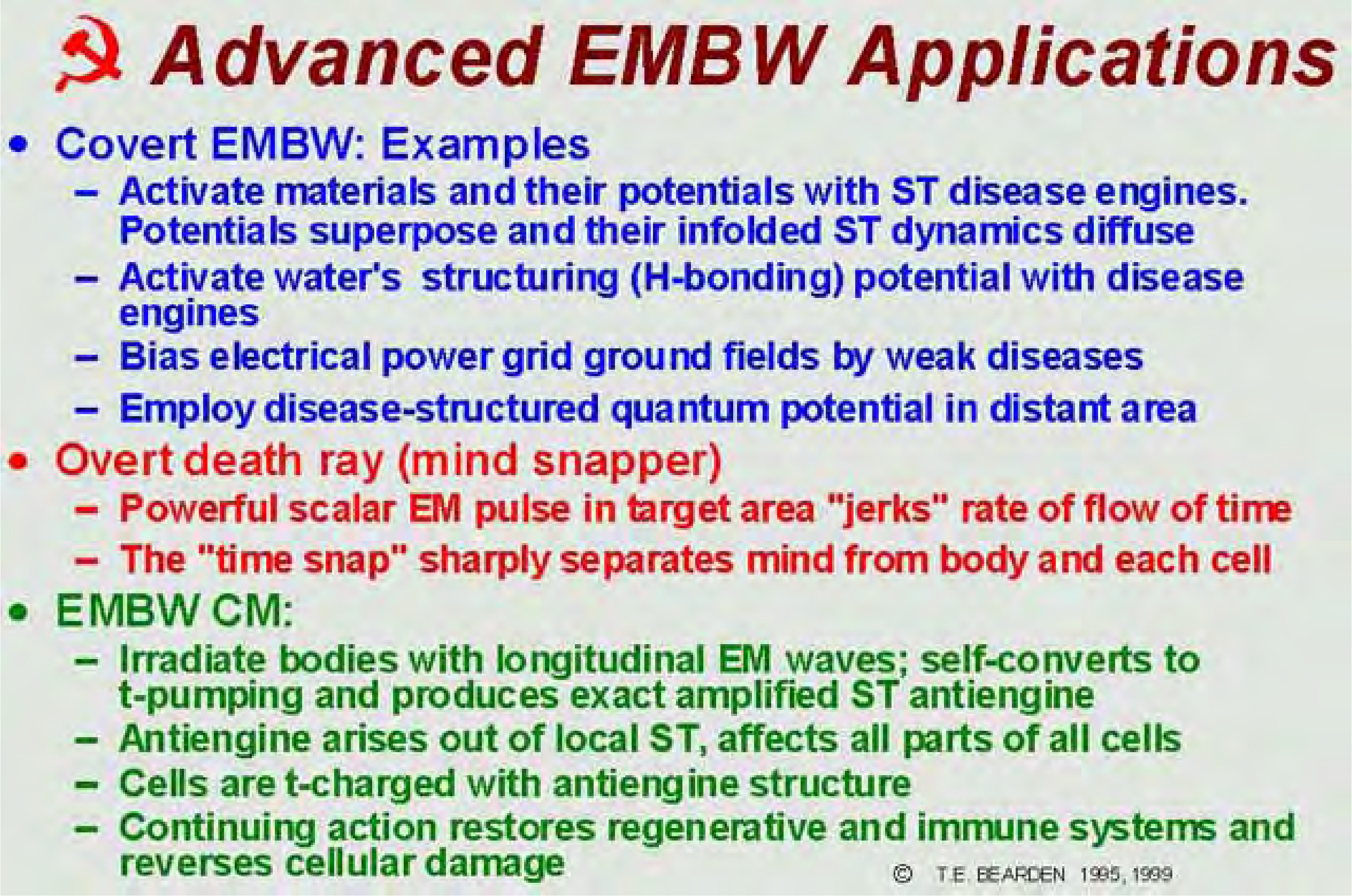

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

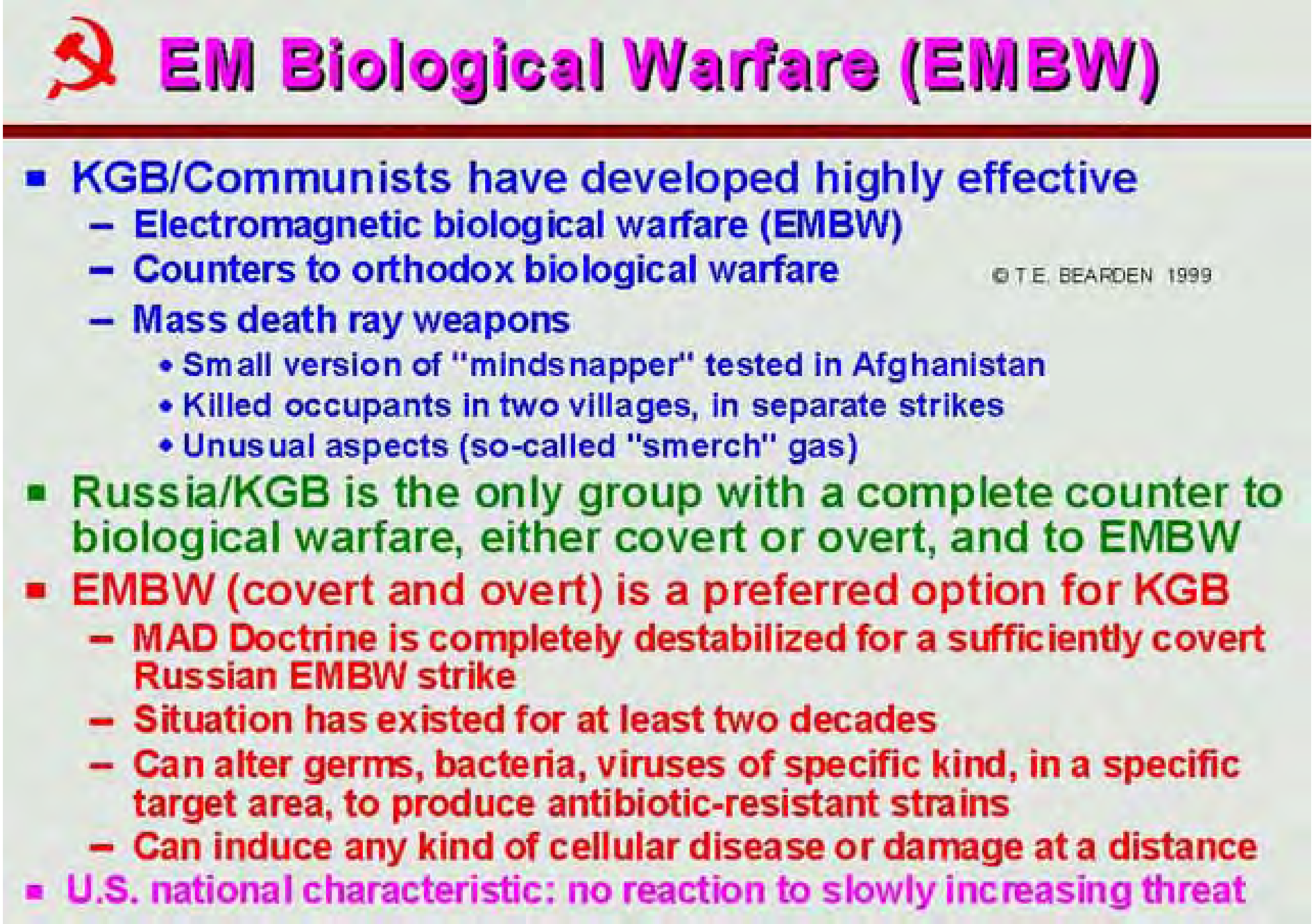

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

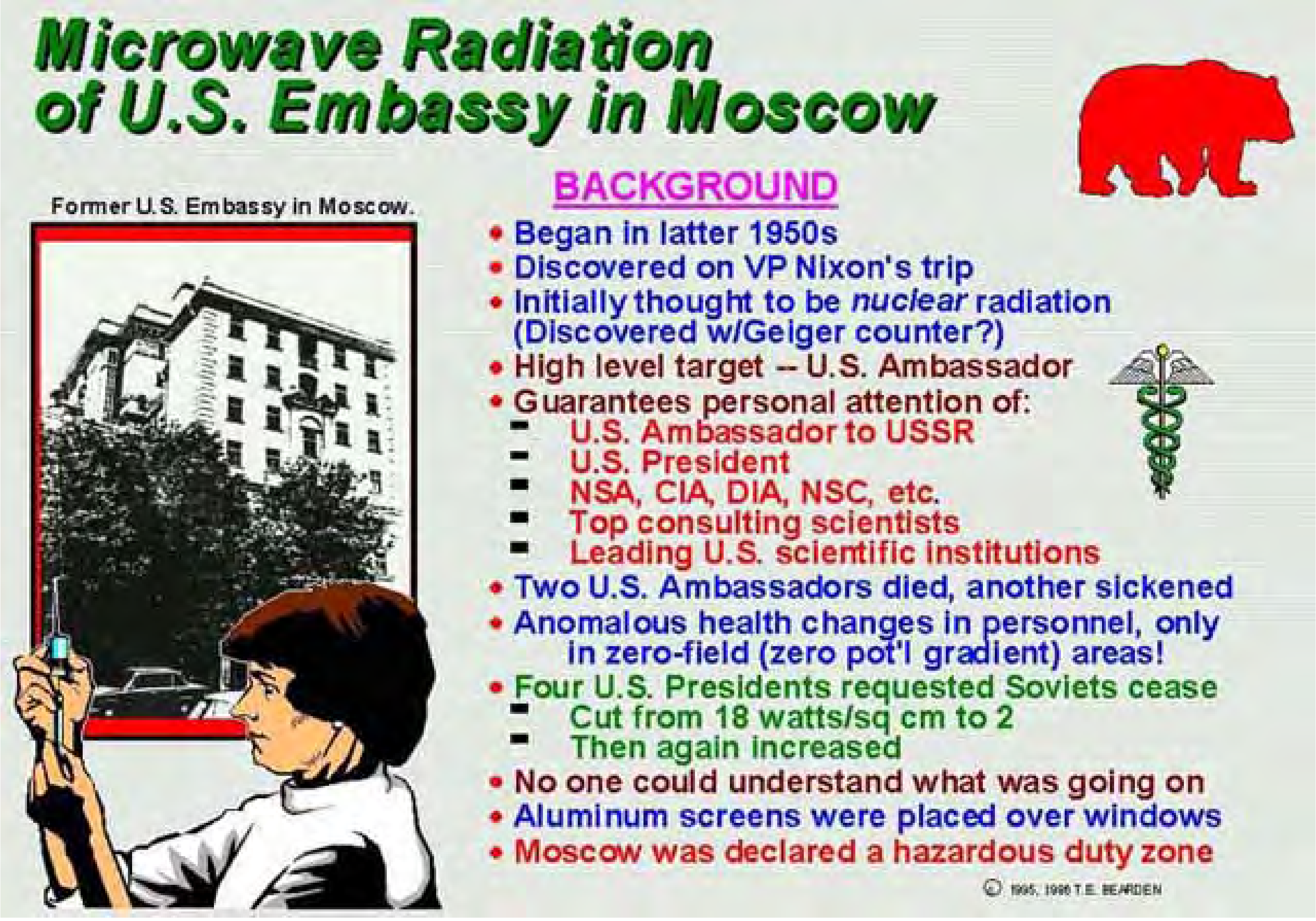

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

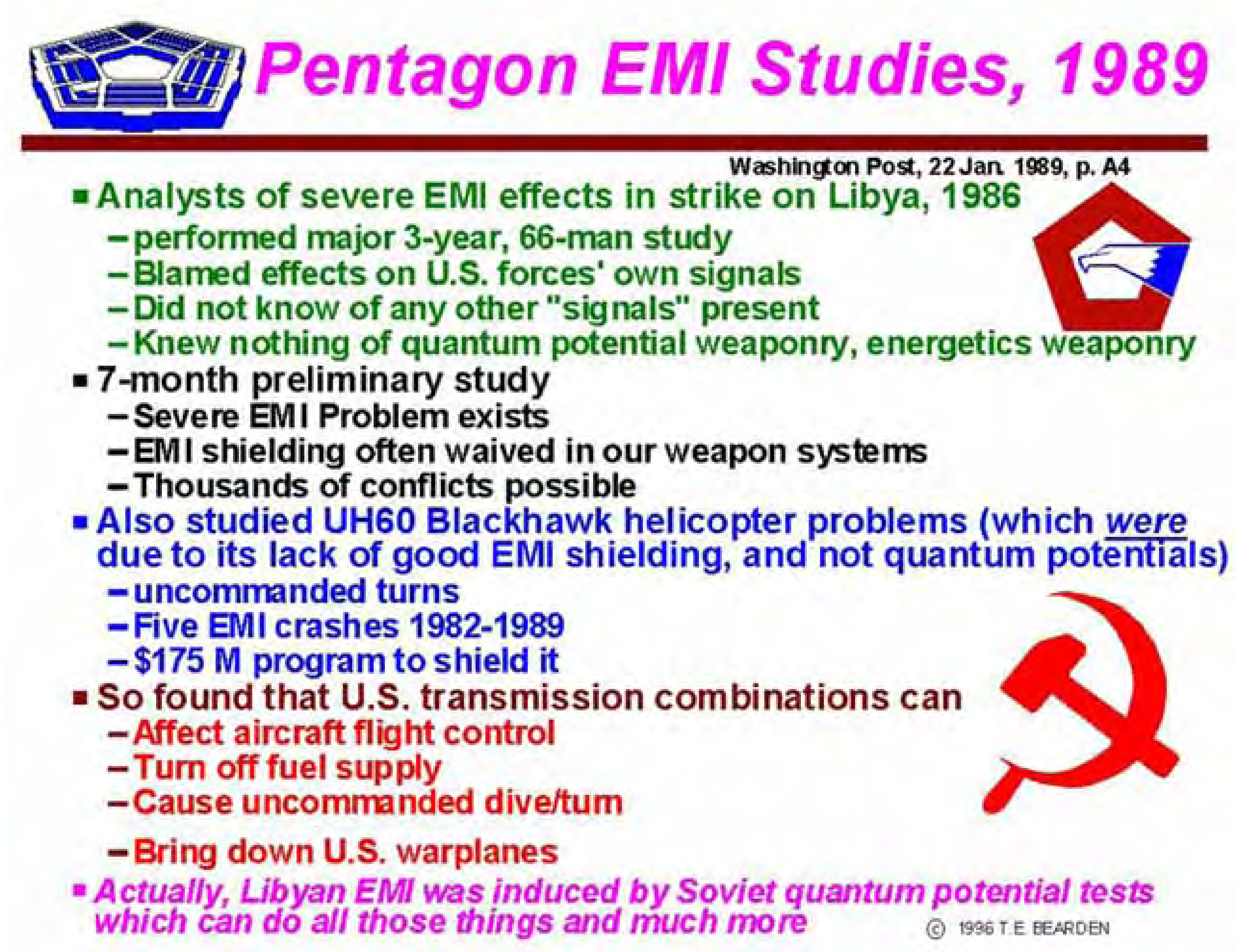

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia
- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996
- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996
- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985
- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985
- EM Missile Practice and Deception
  November 17, 1996

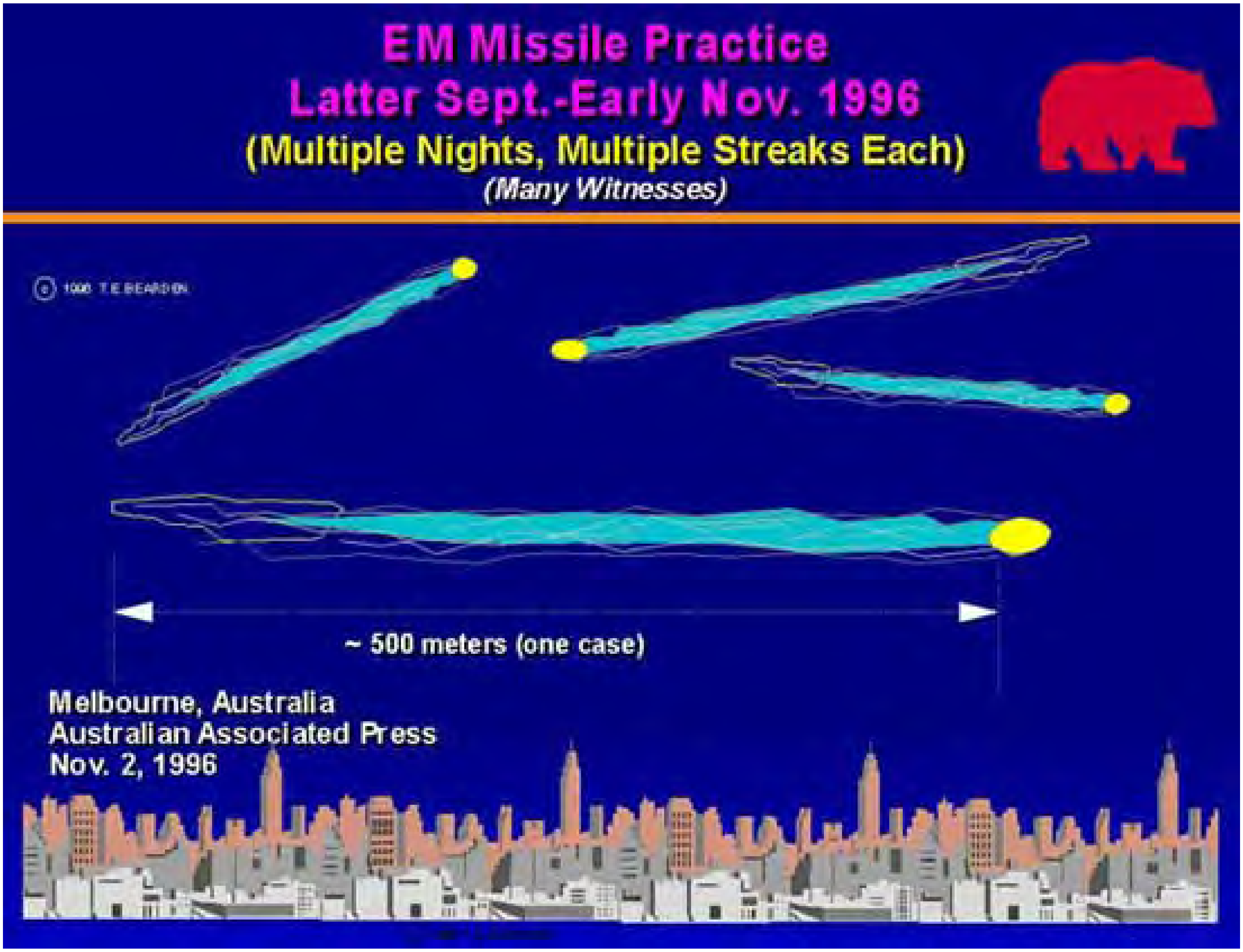

EM Missile Practice
Latter Sept.-Early Nov. 1996

(Multiple Nights, Multiple Streaks Each)

(Many Witnesses)

© 1996 T.E. BEARDEN

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

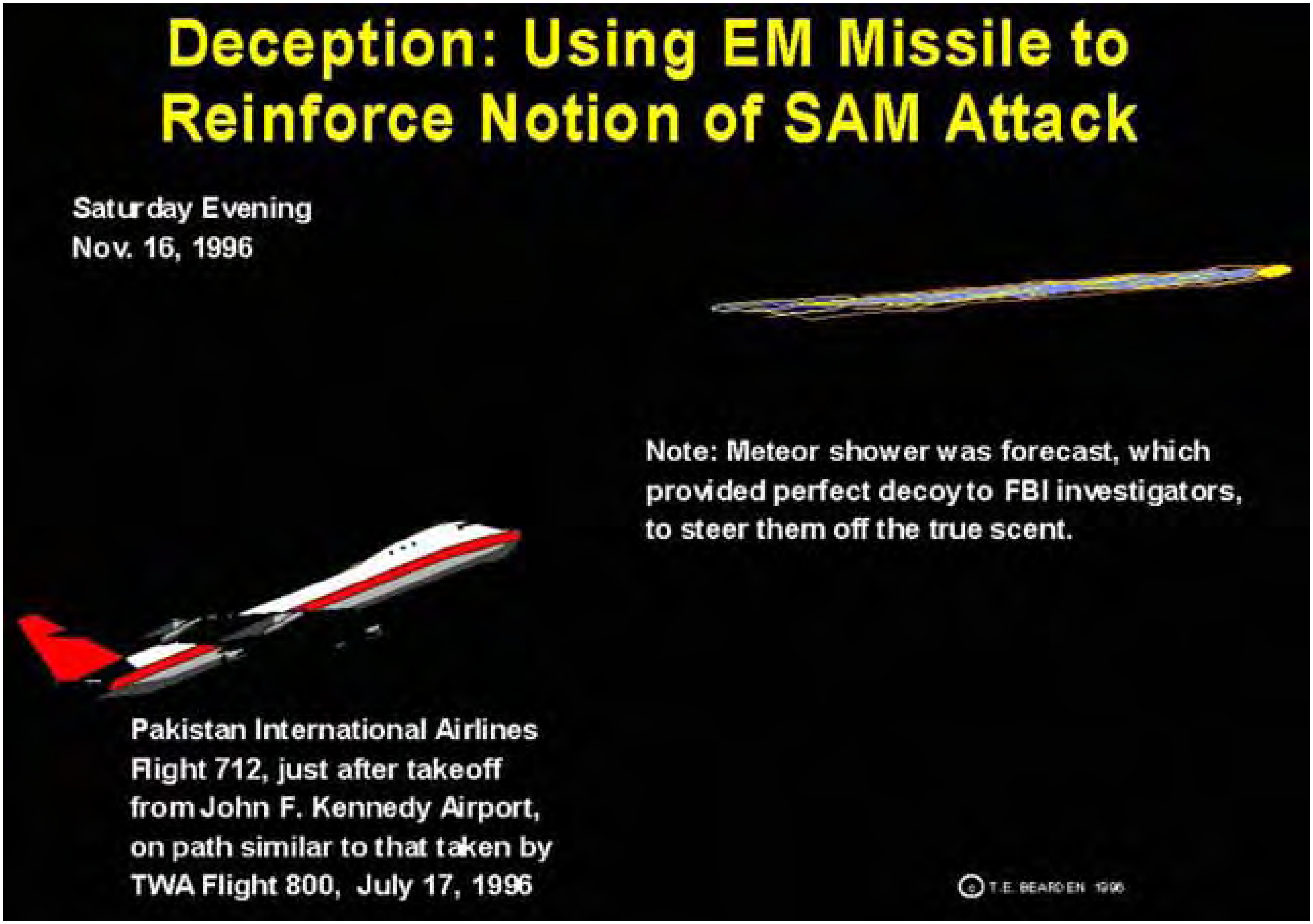

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

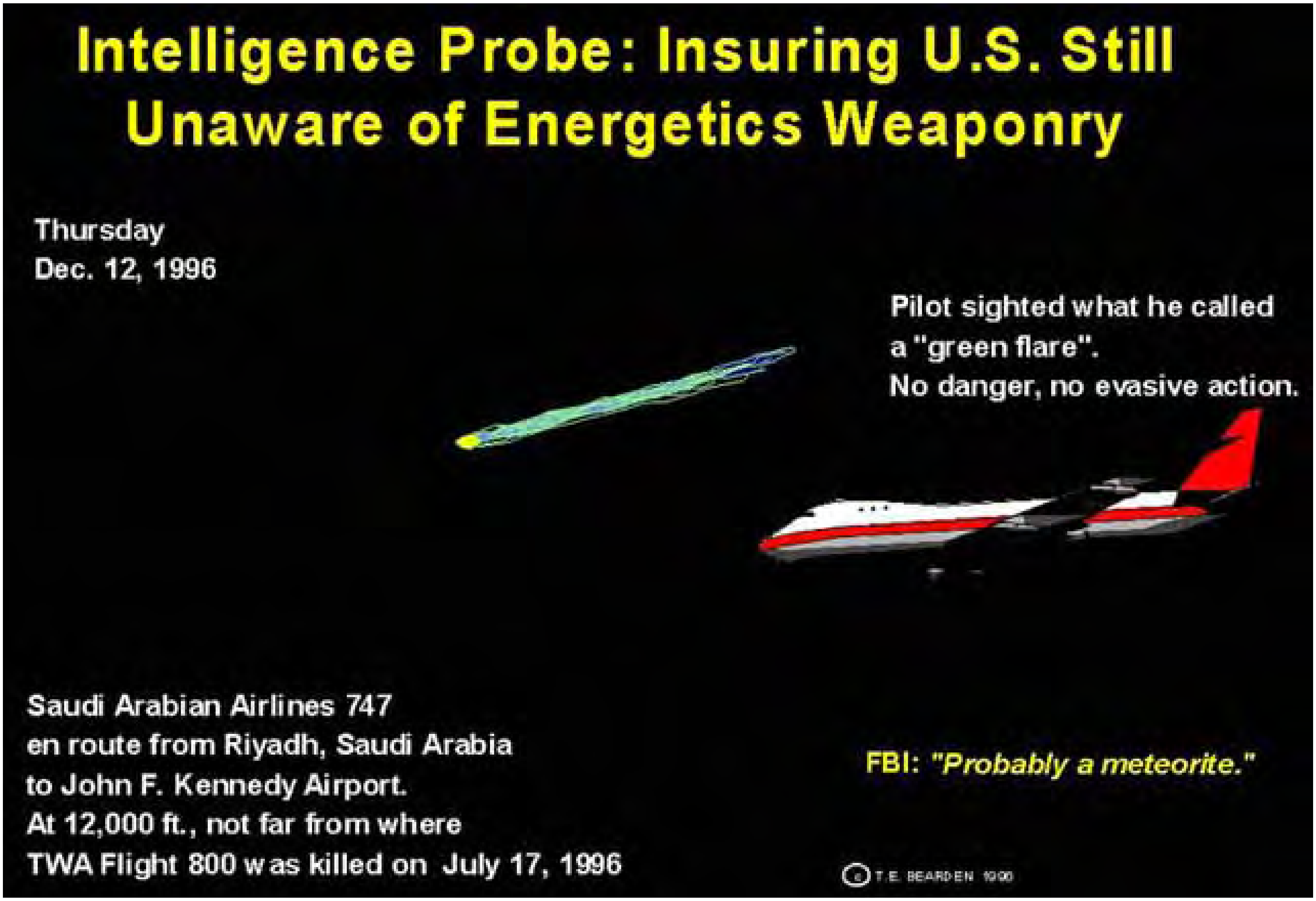

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: ***"Probably a meteorite."***

© T.E. BEARDEN 1996

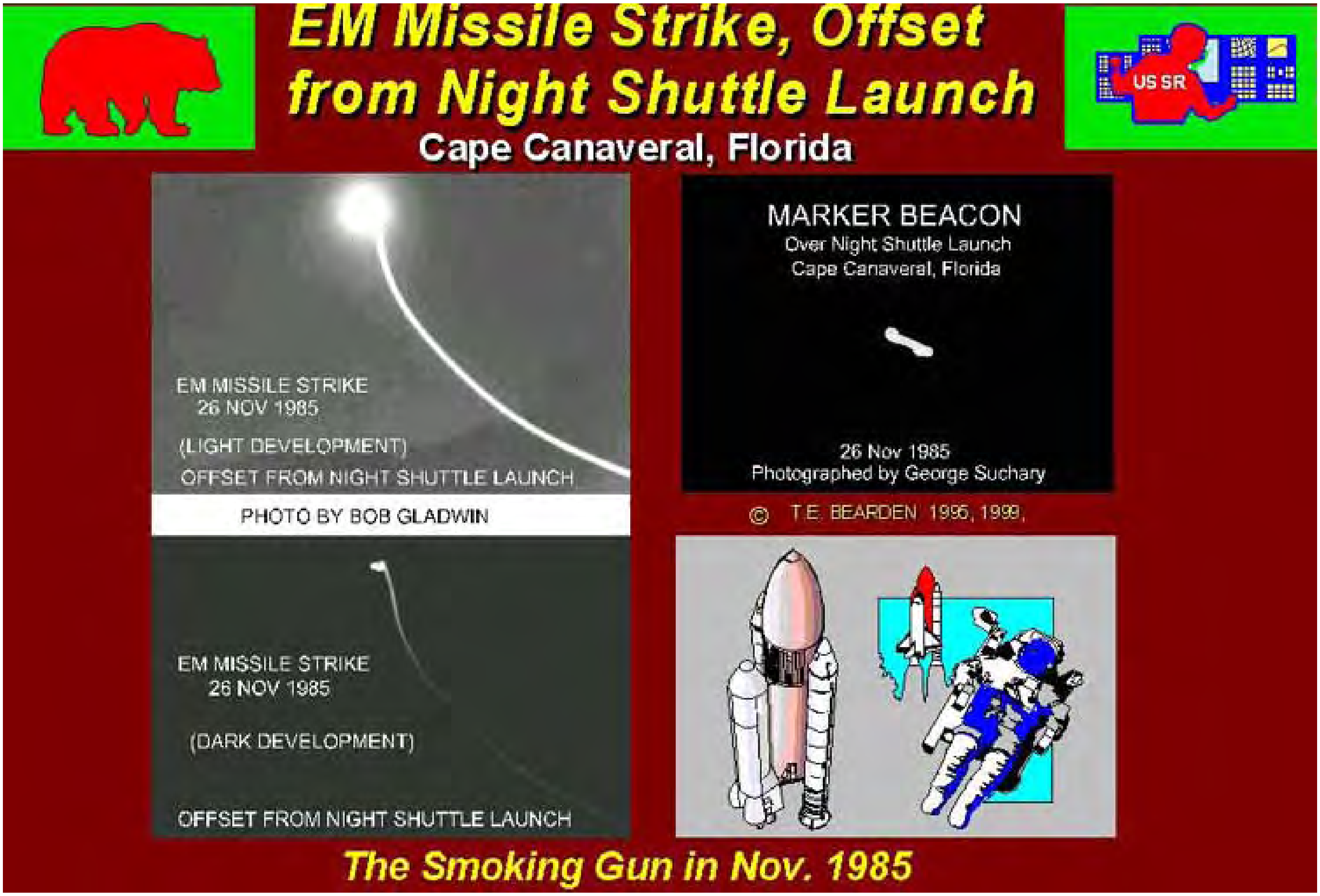

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985
(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1996, 1999,

EM MISSILE STRIKE
26 NOV 1985
(DARK DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

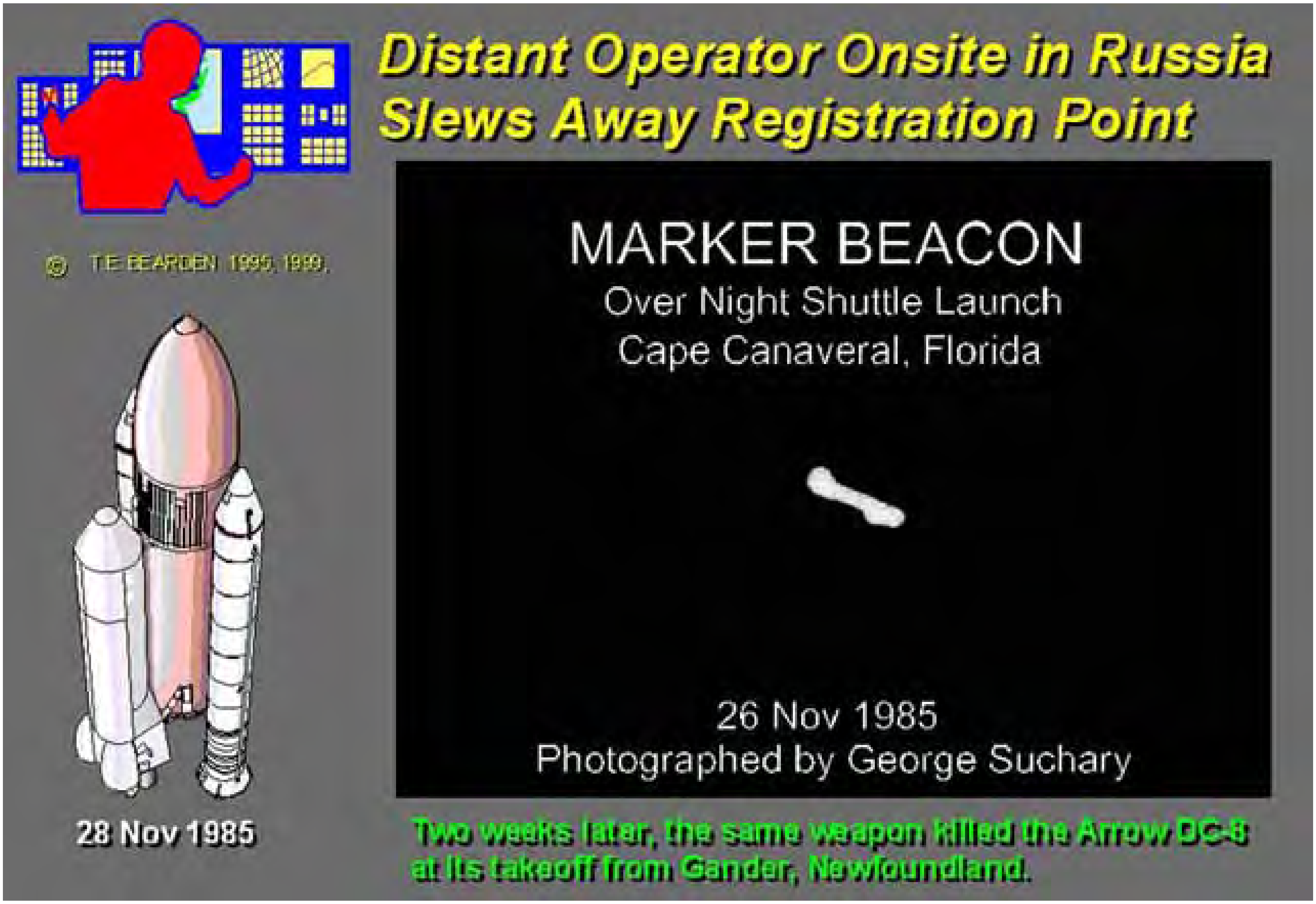

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON

Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

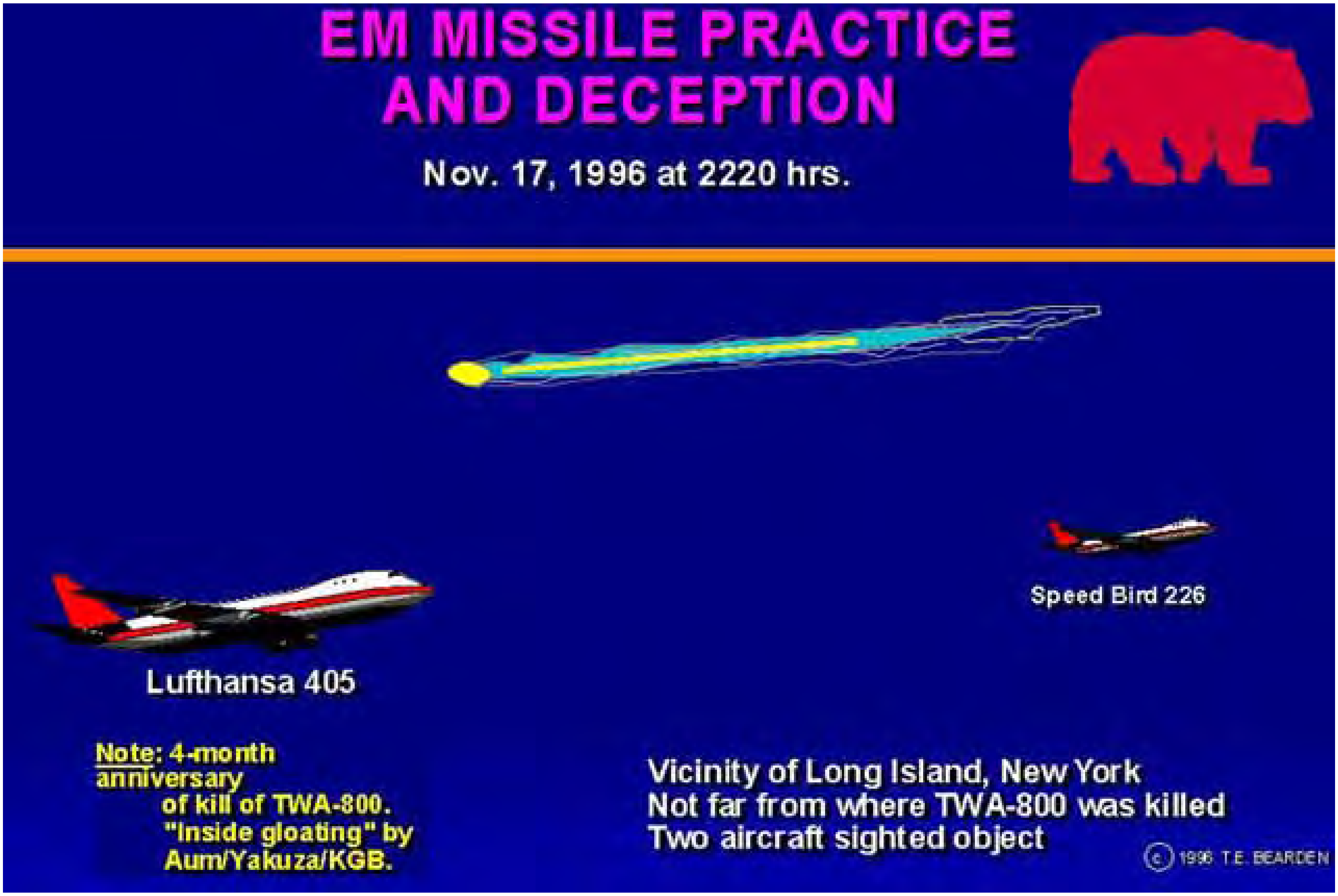

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Speed Bird 226

Lufthansa 405

Note: 4-month
anniversary
of kill of TWA-800.
"Inside gloating" by
Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

(c) 1996 T.E. BEARDEN

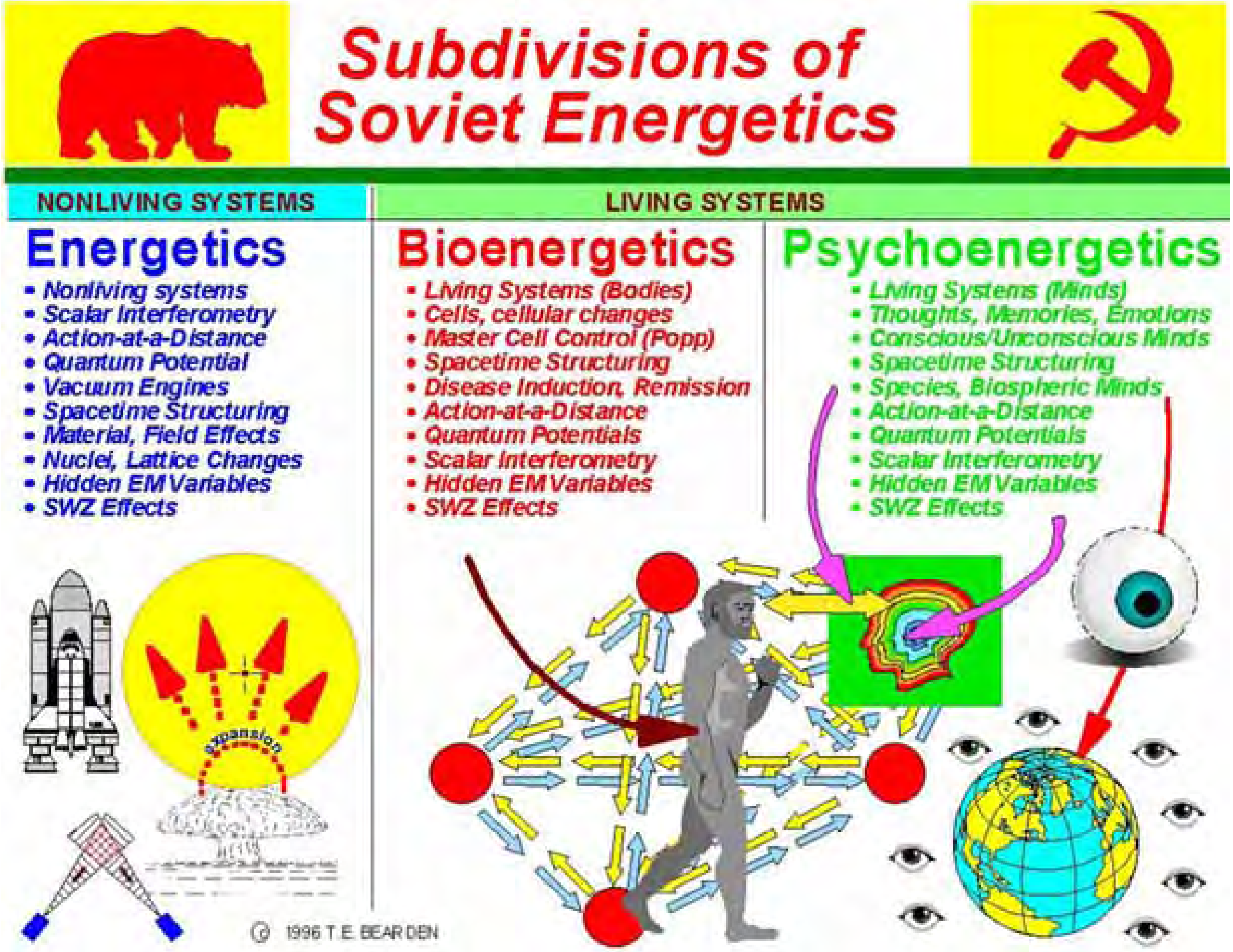

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

LIVING SYSTEMS

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

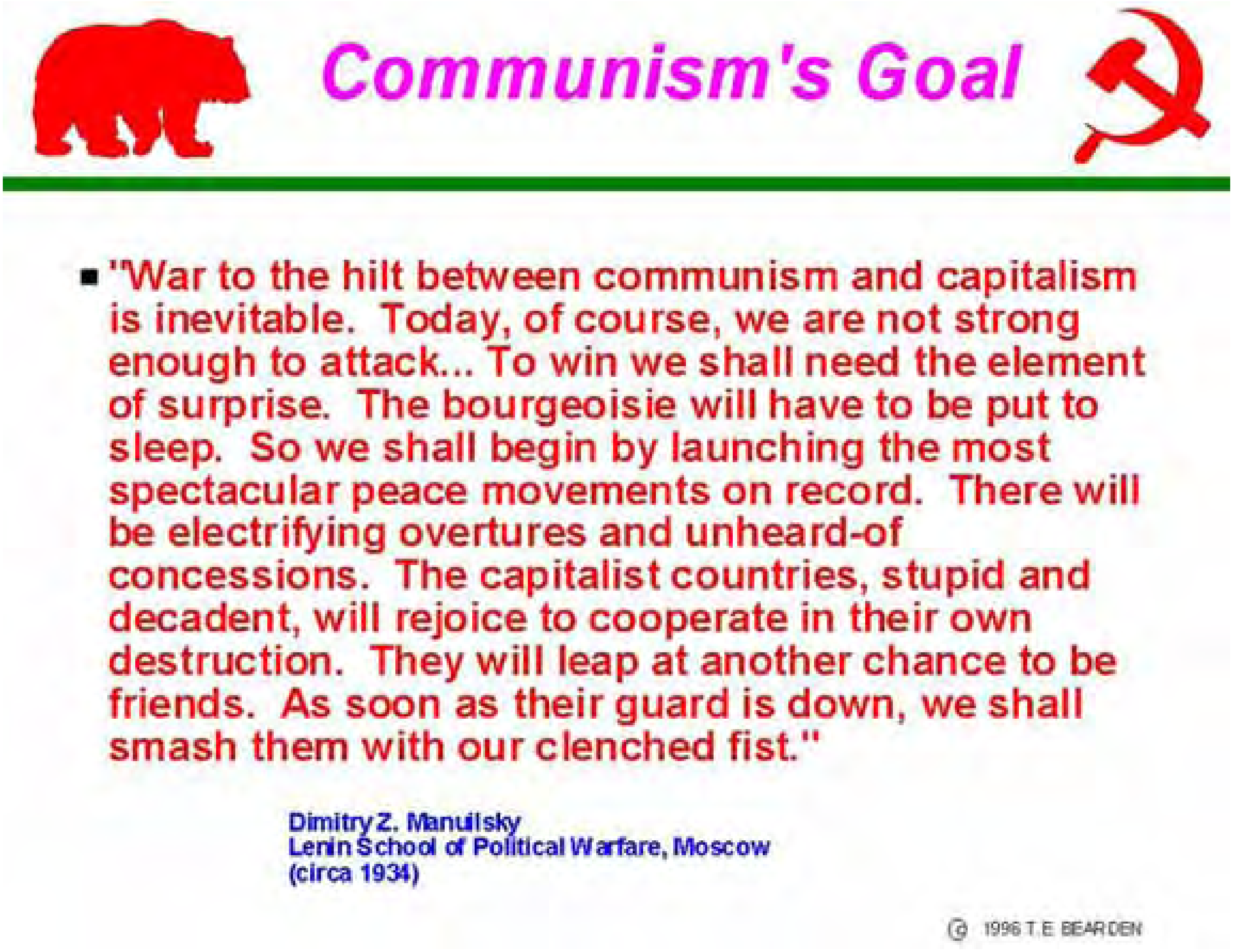

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

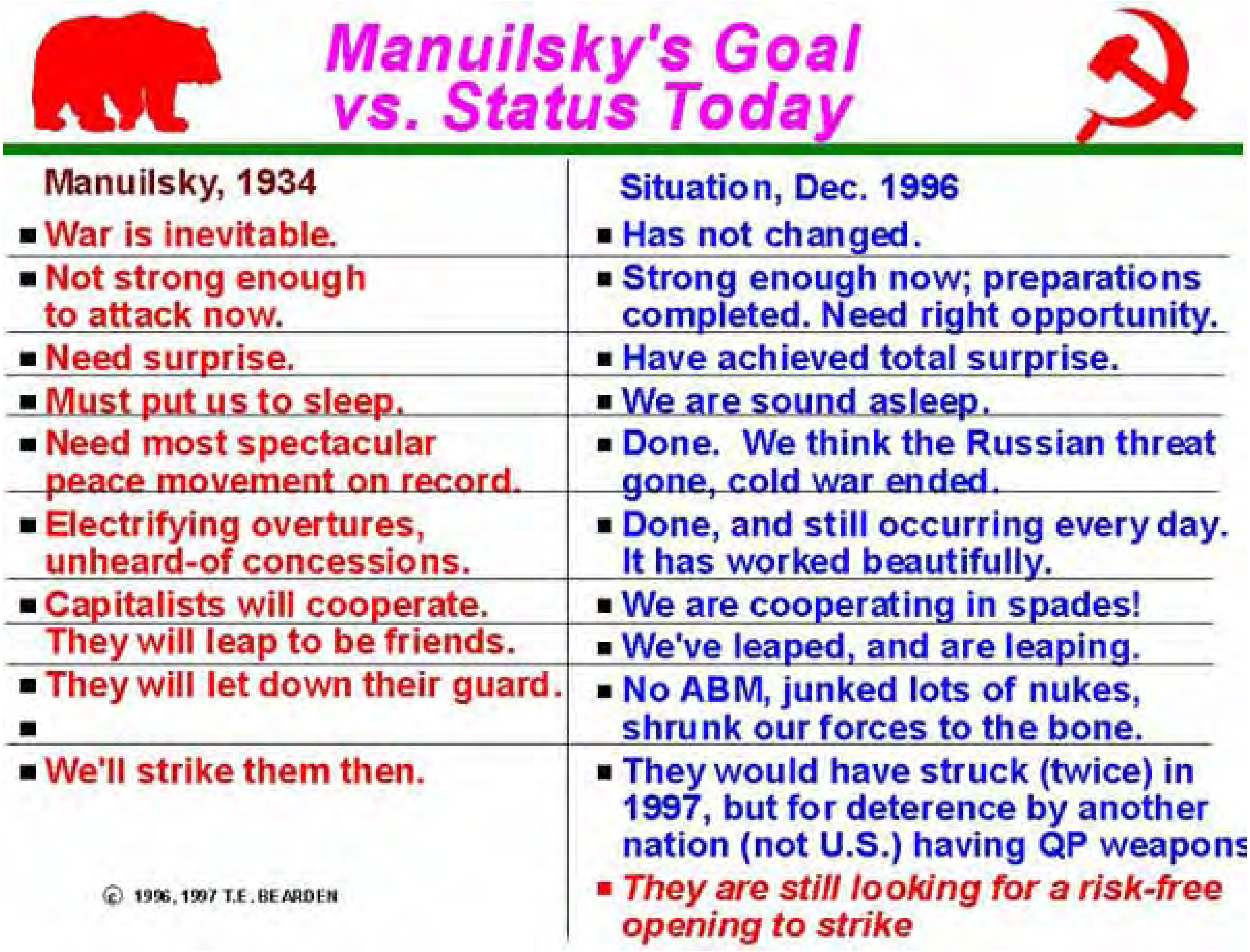

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
■ War is inevitable.	■ Has not changed.
■ Not strong enough to attack now.	■ Strong enough now; preparations completed. Need right opportunity.
■ Need surprise.	■ Have achieved total surprise.
■ Must put us to sleep.	■ We are sound asleep.
■ Need most spectacular peace movement on record.	■ Done. We think the Russian threat gone, cold war ended.
■ Electrifying overtures, unheard-of concessions.	■ Done, and still occurring every day. It has worked beautifully.
■ Capitalists will cooperate.	■ We are cooperating in spades!
They will leap to be friends.	■ We've leaped, and are leaping.
■ They will let down their guard. ■	■ No ABM, junked lots of nukes, shrunk our forces to the bone.
■ We'll strike them then.	■ They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons ■ *They are still looking for a risk-free opening to strike*

© 1996, 1997 T.E. BEARDEN

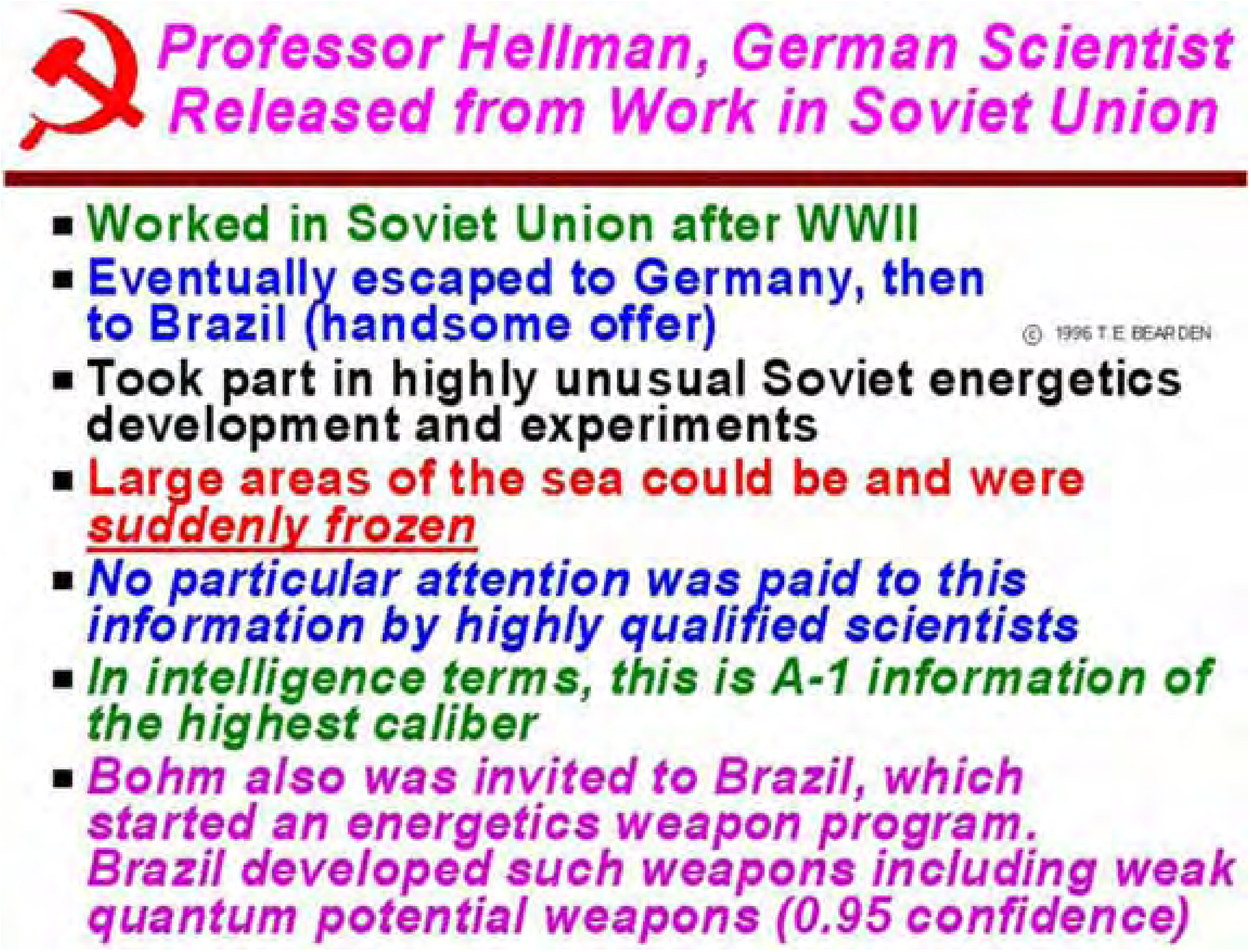

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)
- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were *suddenly frozen*
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

© 1996 T. E. BEARDEN

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

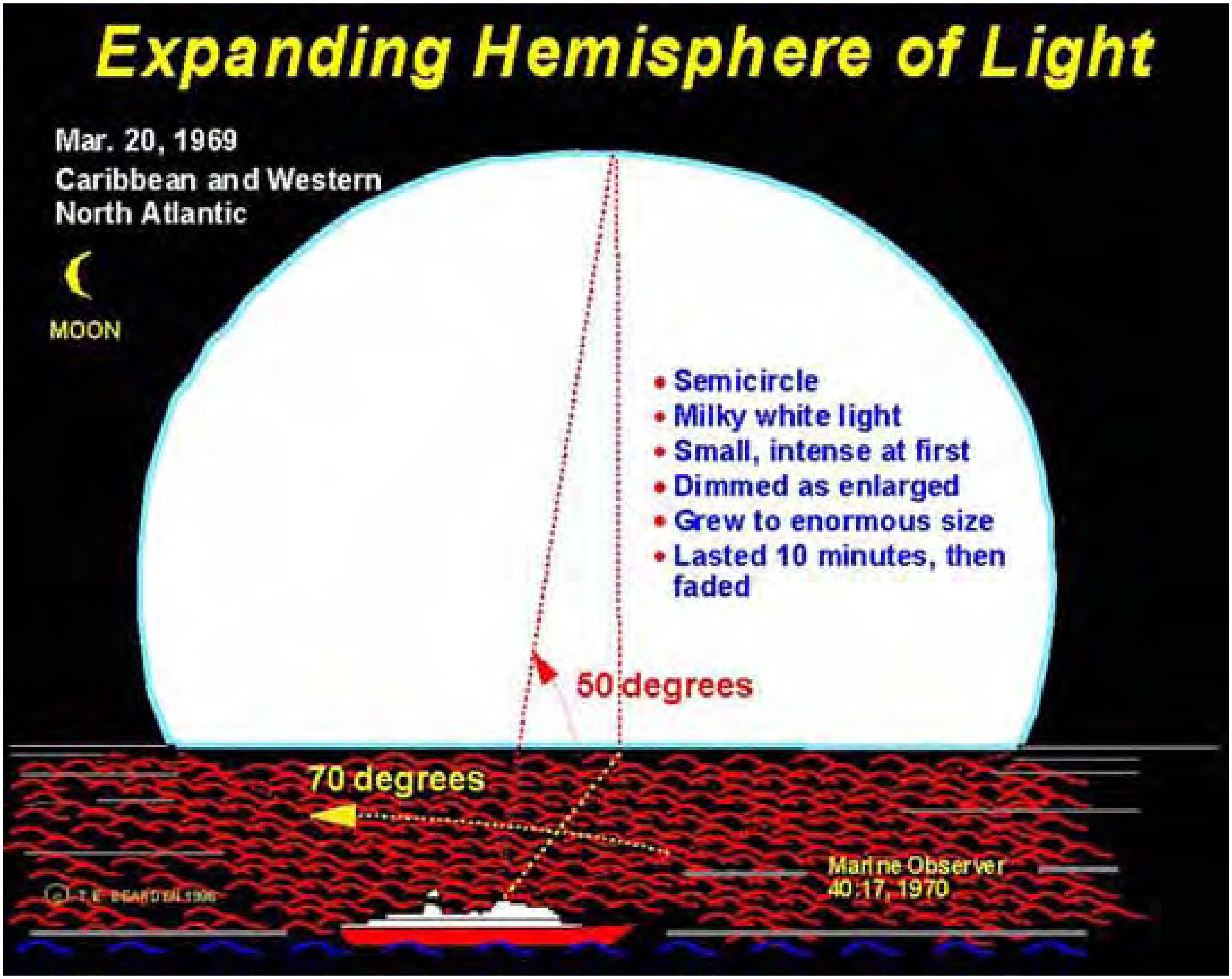

Expanding Hemisphere of Light

Mar. 20, 1969
Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

Marine Observer
40:17, 1970

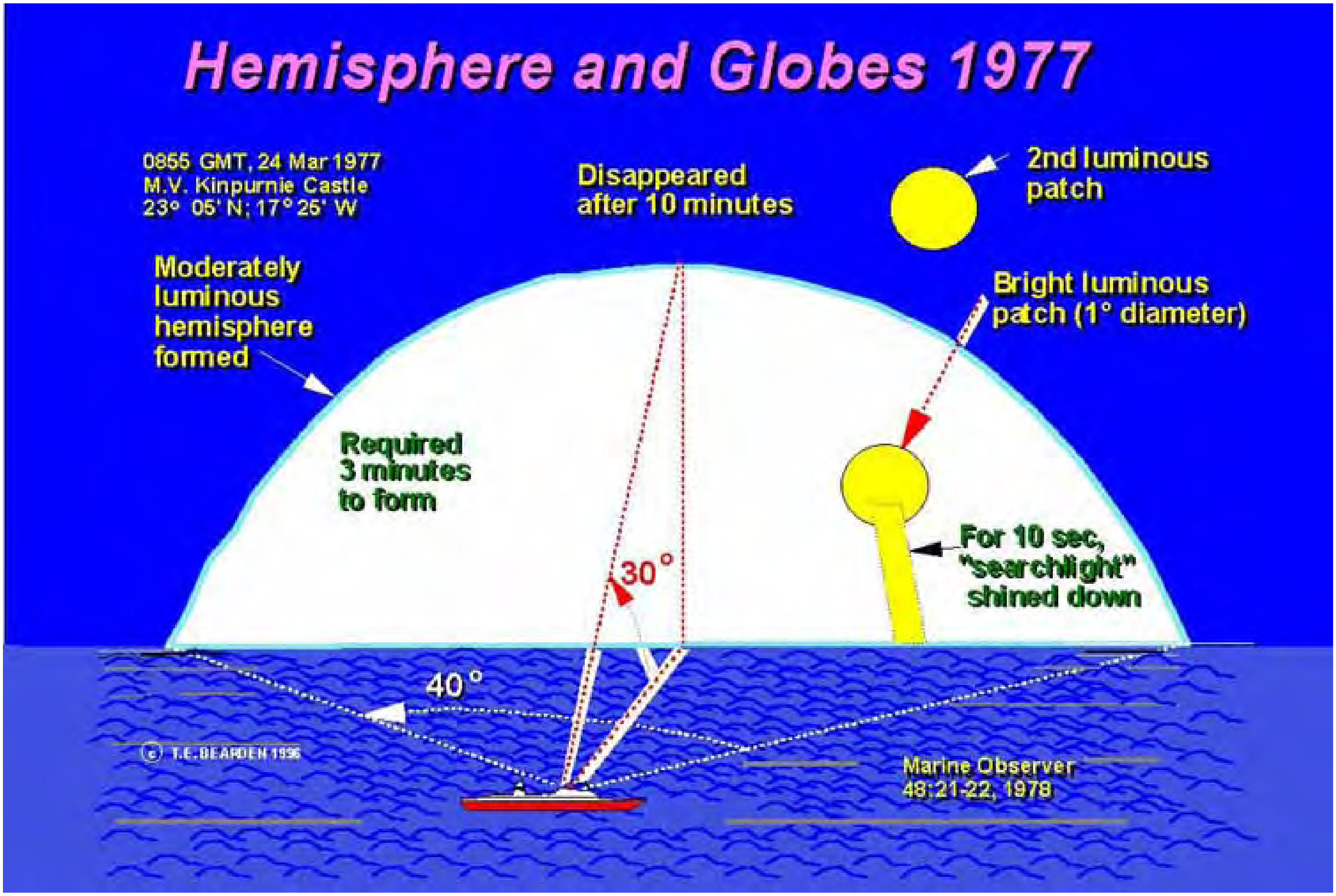
Hemisphere and Globes 1977
0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W
Disappeared after 10 minutes
2nd luminous patch
Moderately luminous hemisphere formed
Bright luminous patch (1° diameter)
Required 3 minutes to form
30°
For 10 sec, "searchlight" shined down
40°
© T.E. BEARDEN 1996
Marine Observer 48:21-22, 1978

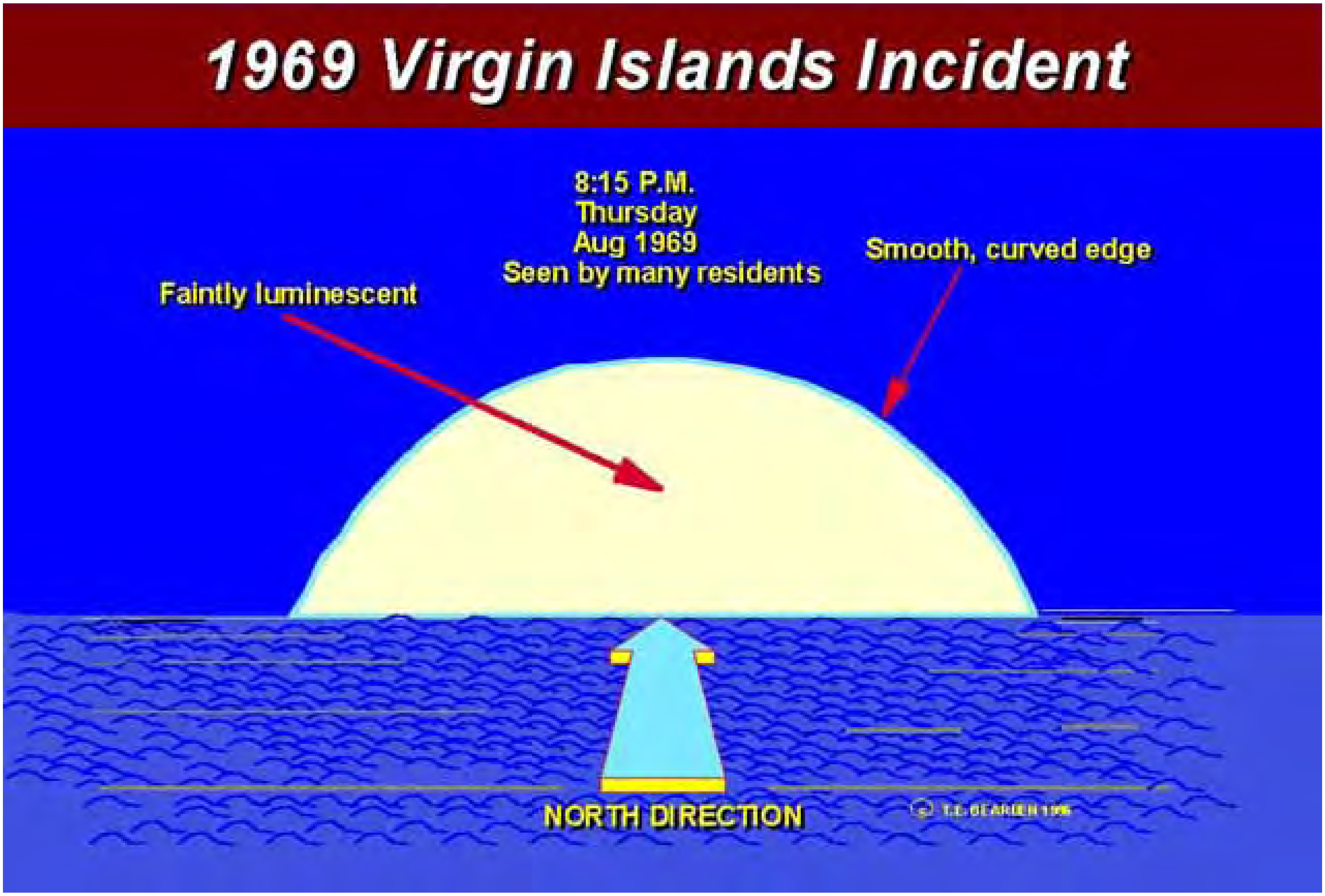

1969 Virgin Islands Incident

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent

Smooth, curved edge

NORTH DIRECTION

© T.E. BEARDEN 1996

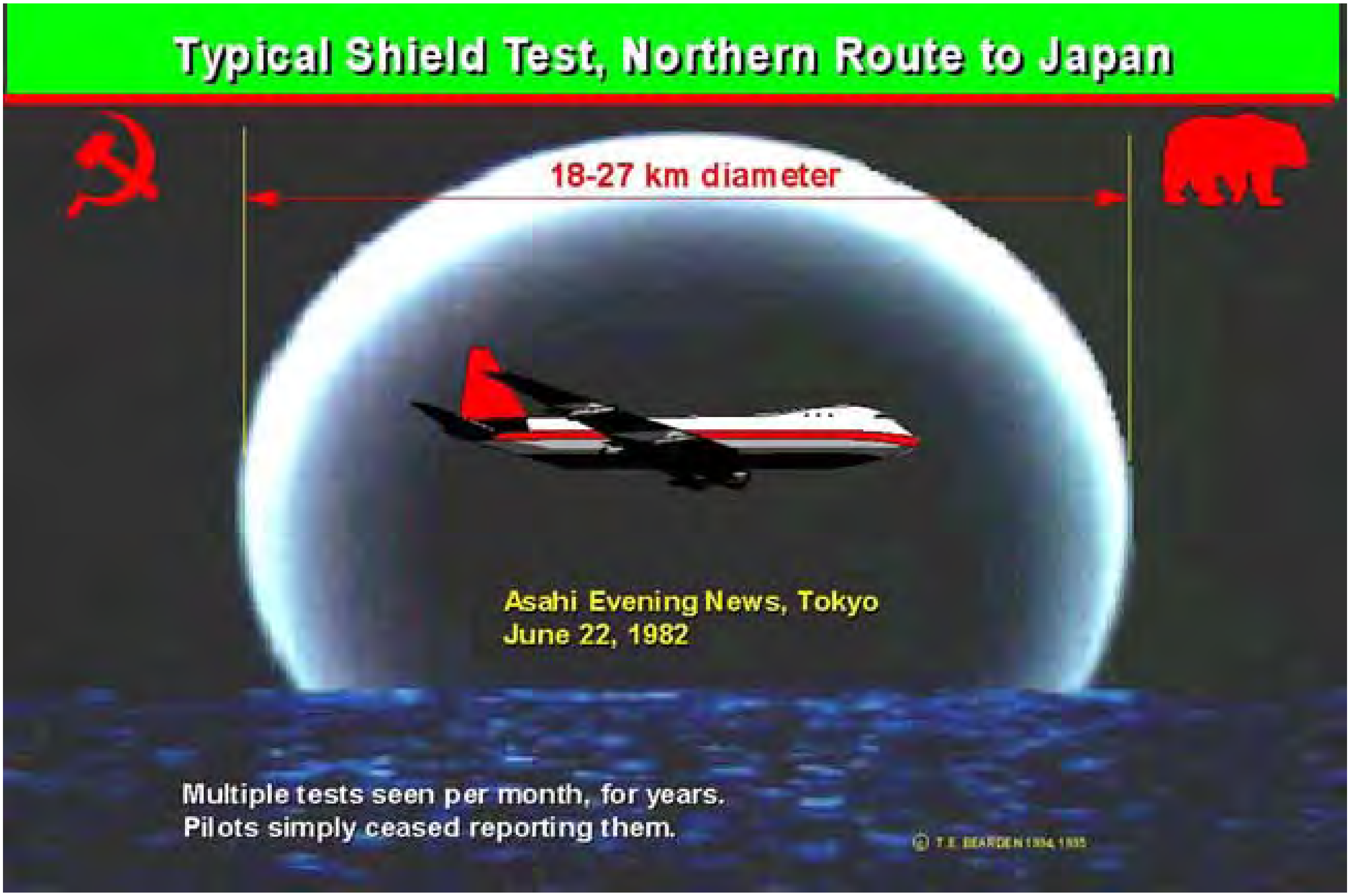

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

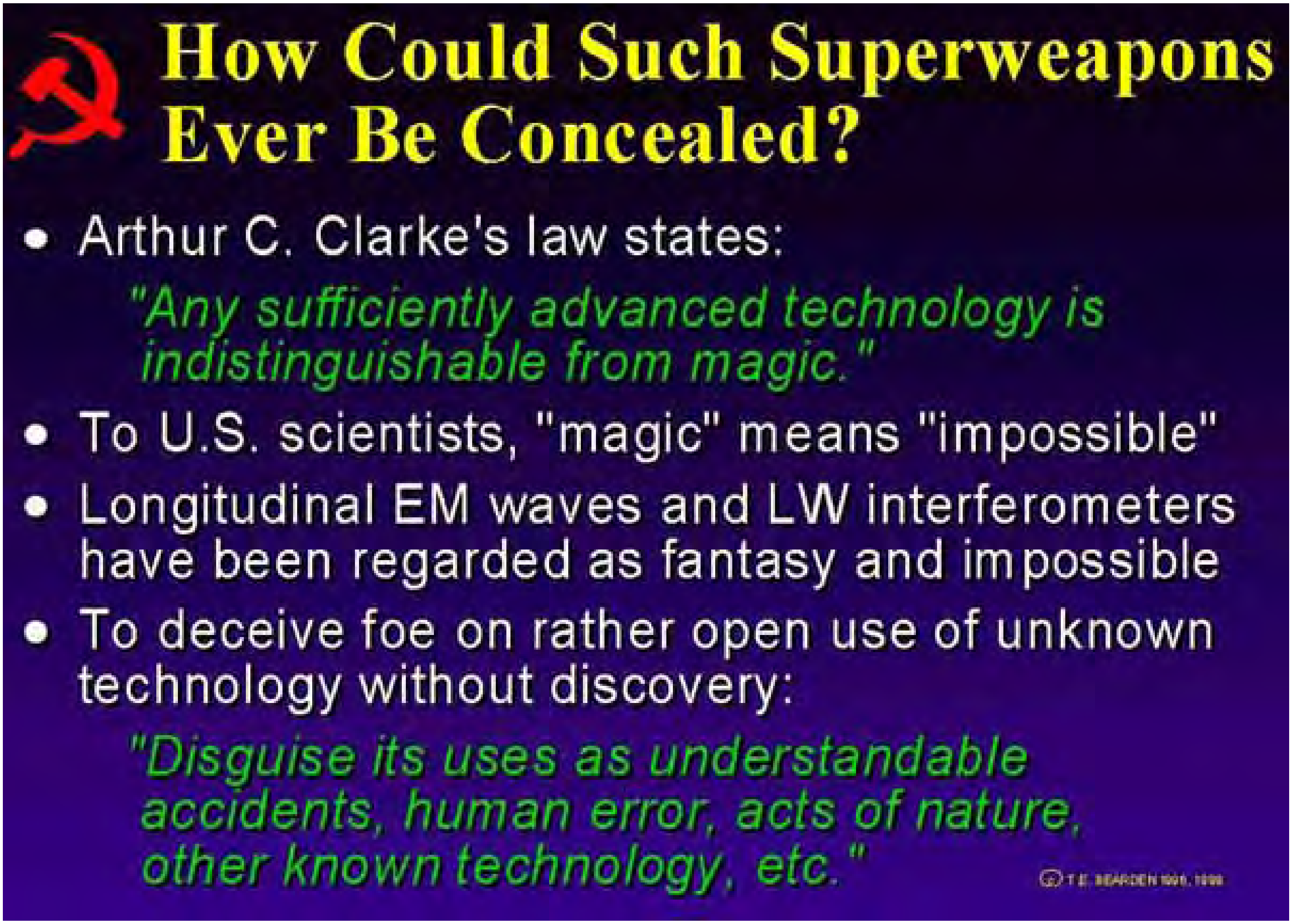

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."
- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN [illegible]

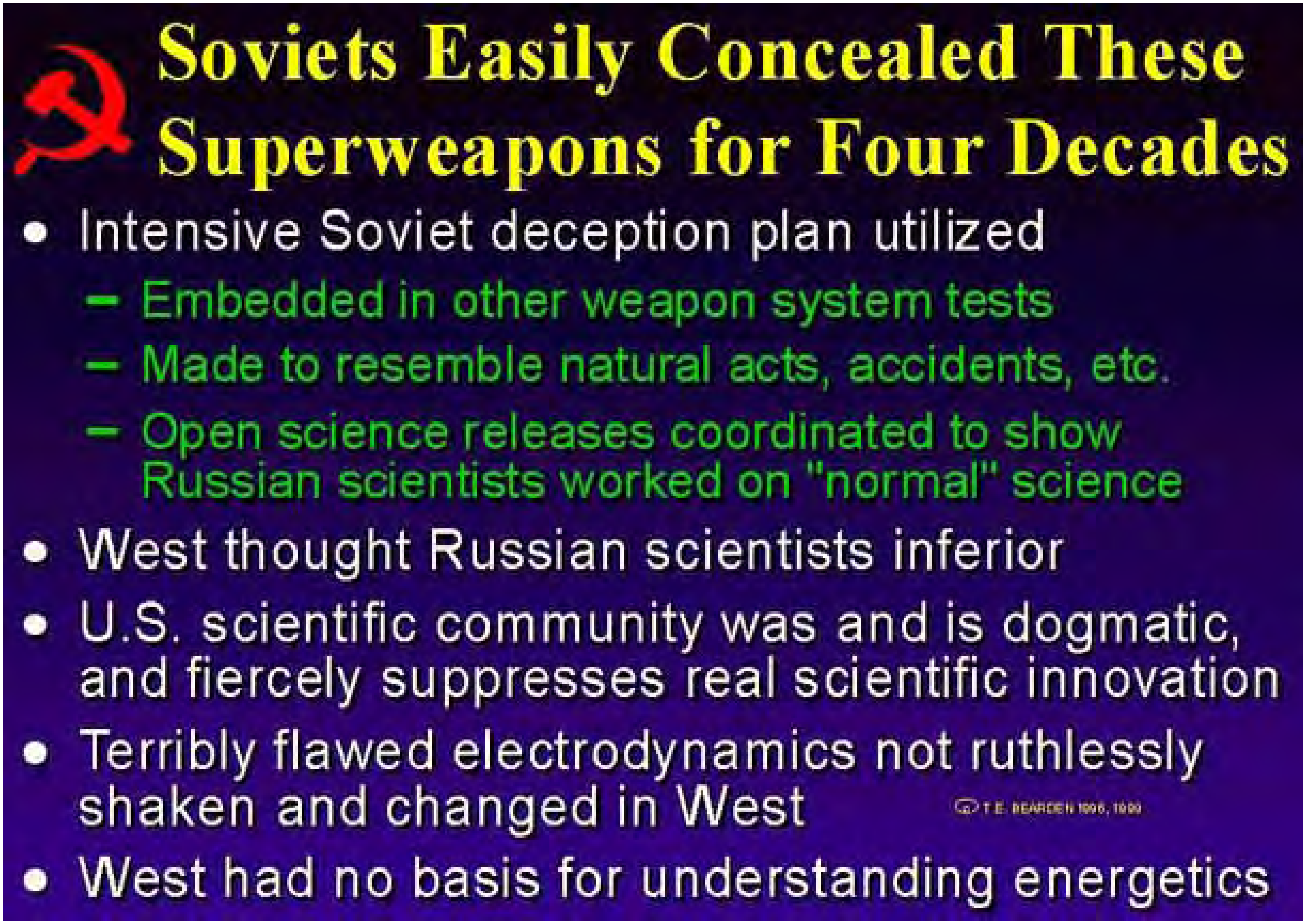

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

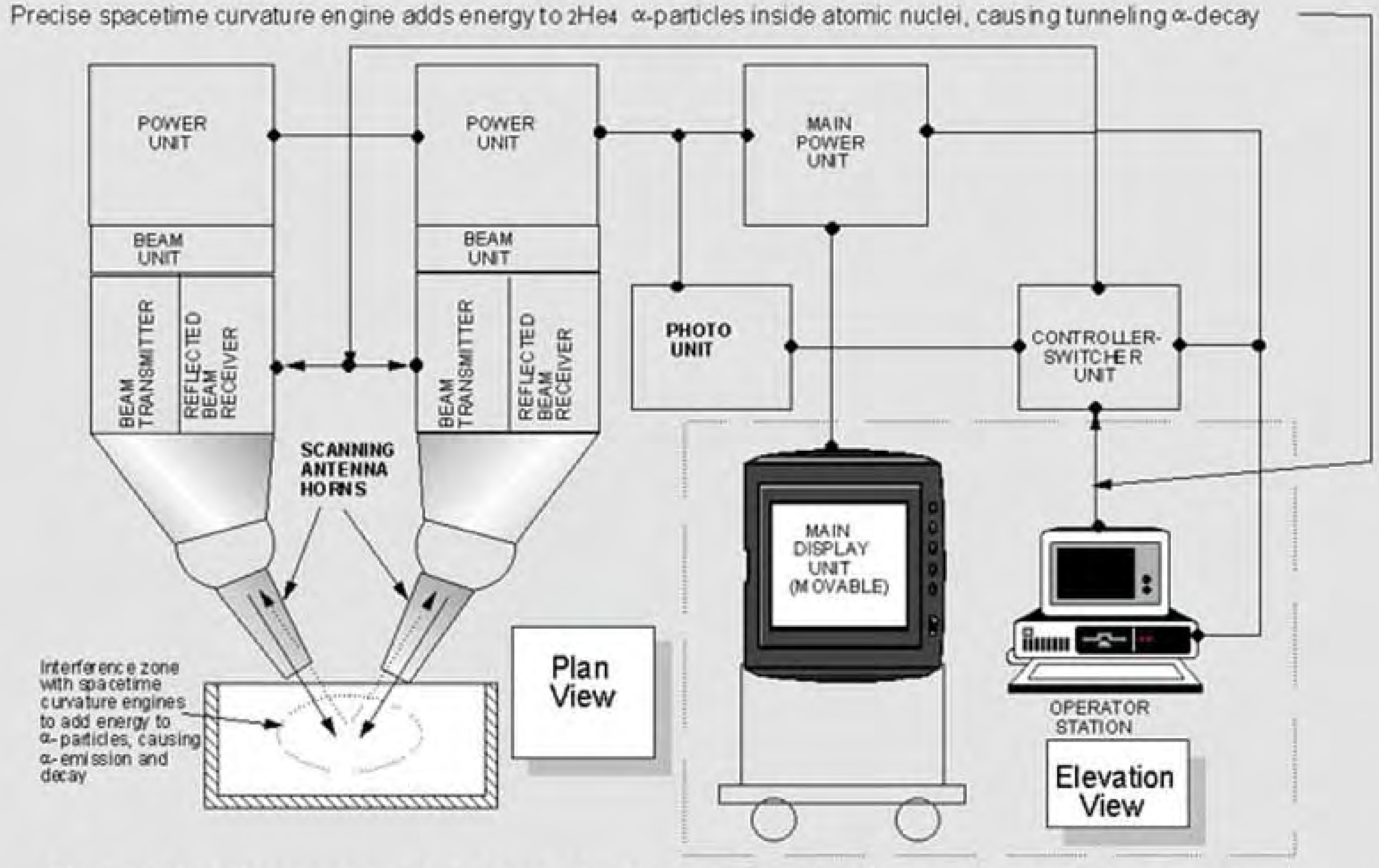
Precise spacetime curvature engine adds energy to 2He4 α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
MAIN POWER UNIT
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
OPERATOR STATION
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1996

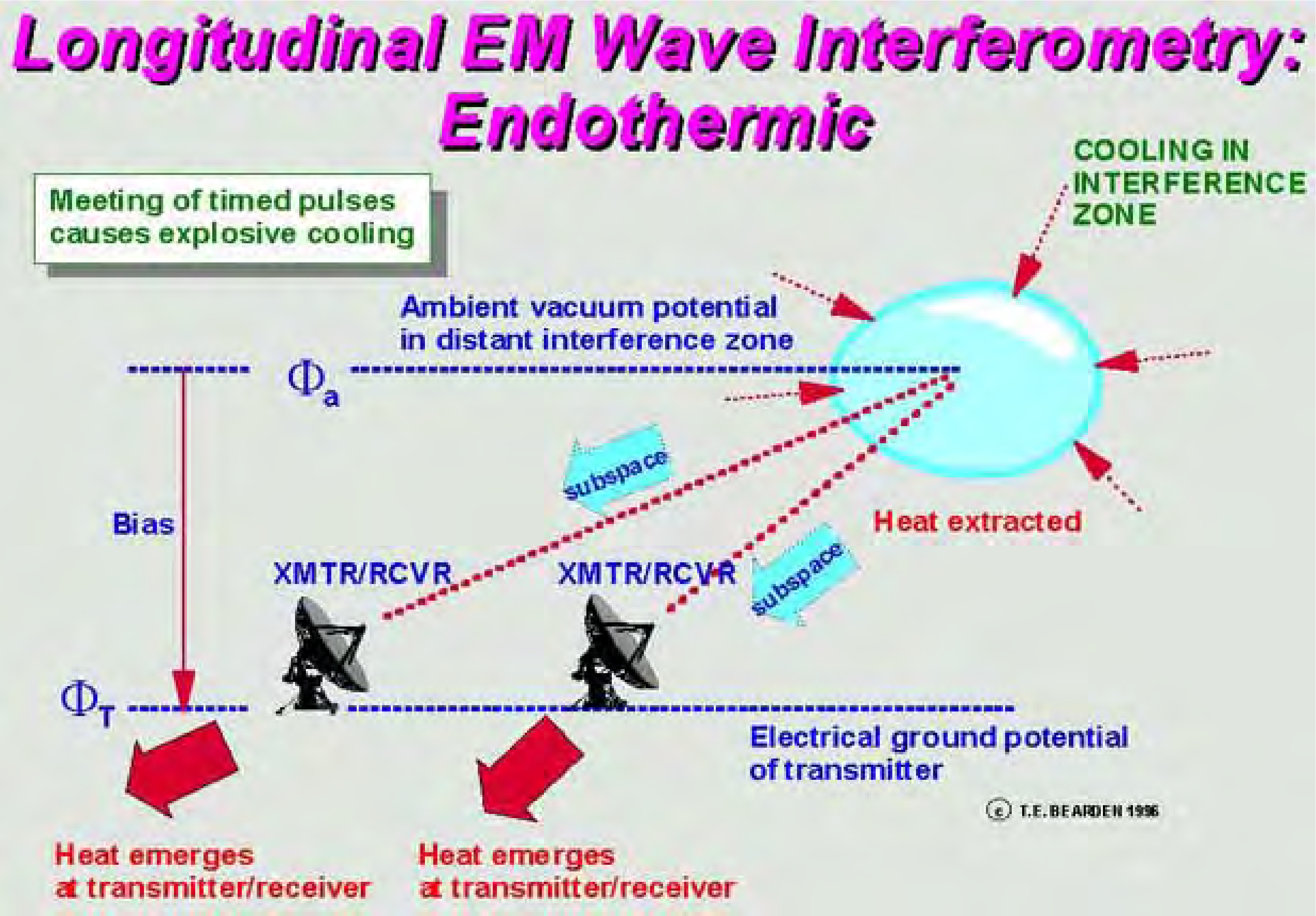

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses causes explosive cooling

COOLING IN INTERFERENCE ZONE

Ambient vacuum potential in distant interference zone

Φ_a

Bias

subspace

subspace

Heat extracted

XMTR/RCVR

XMTR/RCVR

Φ_T

Electrical ground potential of transmitter

© T.E. BEARDEN 1996

Heat emerges at transmitter/receiver

Heat emerges at transmitter/receiver

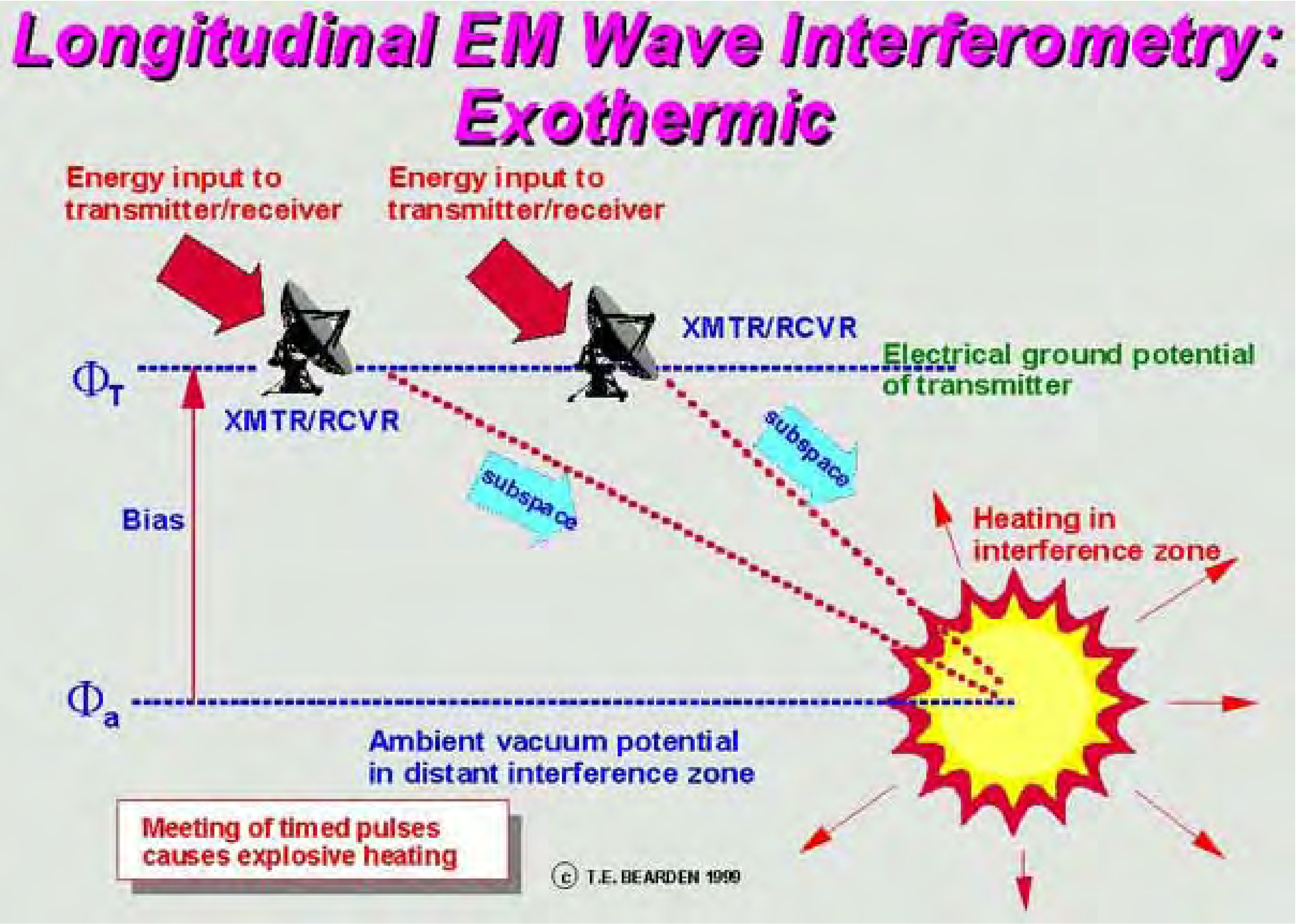

Longitudinal EM Wave Interferometry: Exothermic

Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
Φ_T
XMTR/RCVR
Electrical ground potential of transmitter
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

ⓒ T.E. BEARDEN 1999

The Tom Bearden Website

Help support the research

U.S. Air Attack on Libya

April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

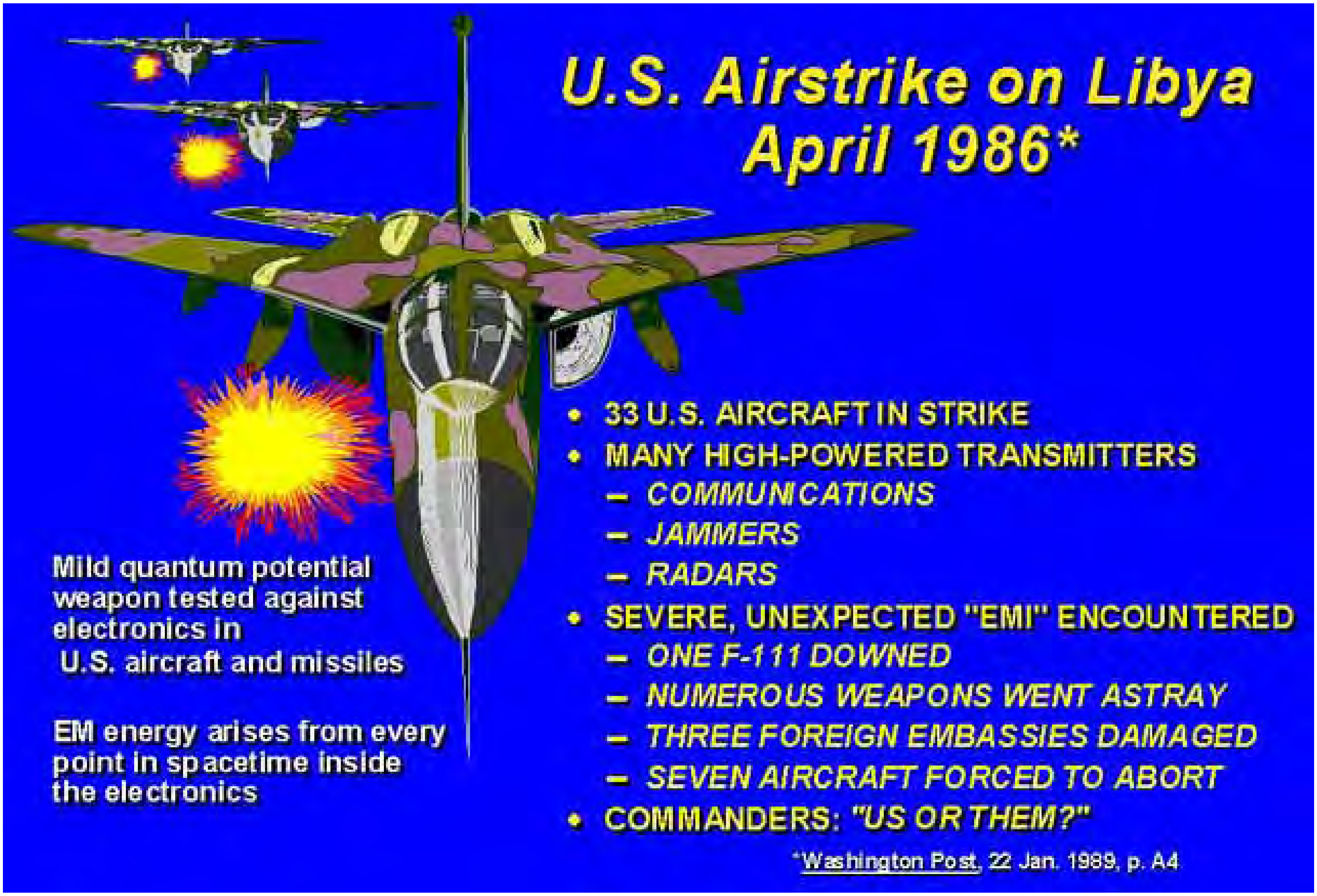

U.S. Airstrike on Libya April 1986*

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

*Washington Post, 22 Jan. 1989, p. A4

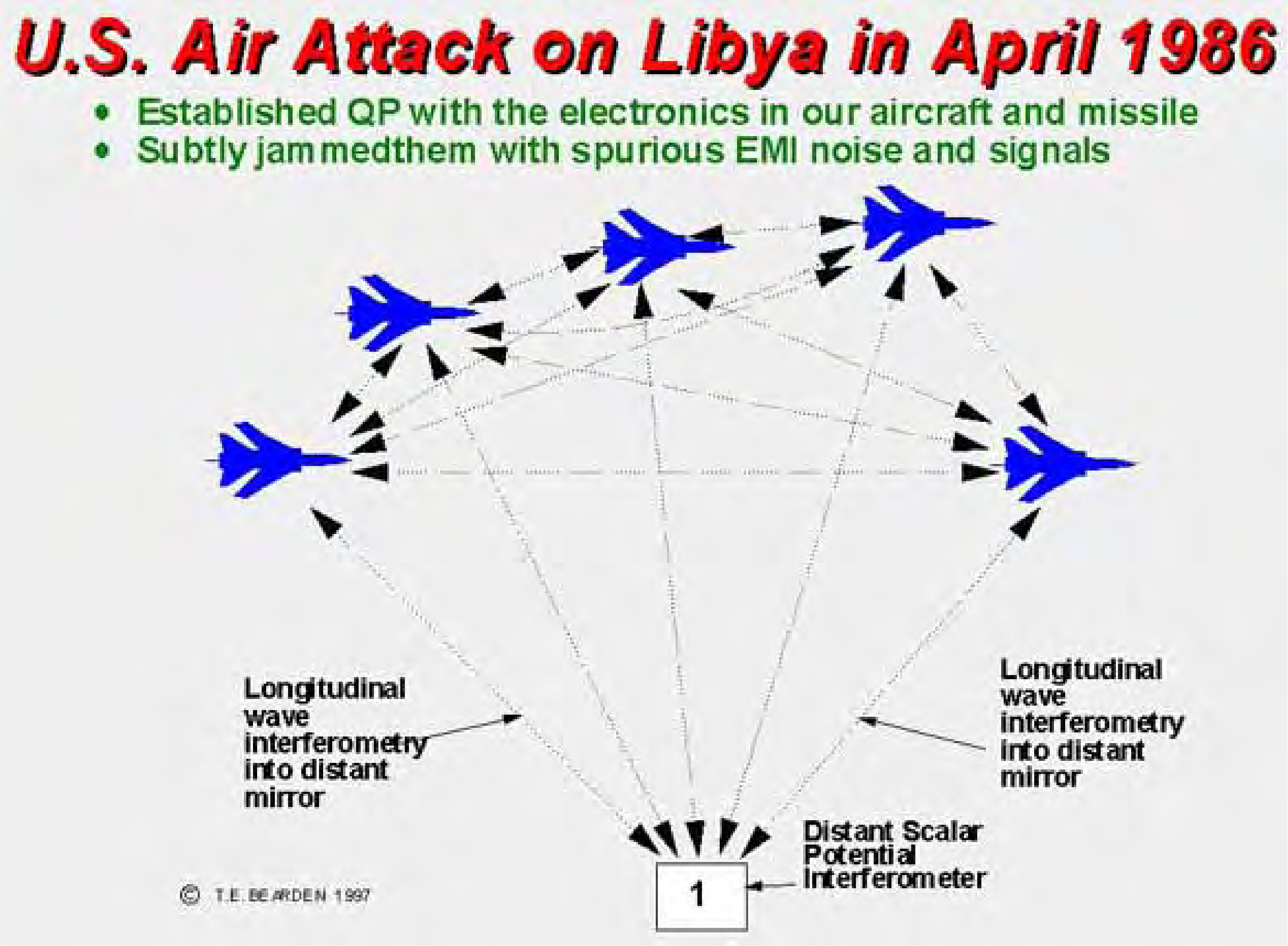

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal wave interferometry into distant mirror

Longitudinal wave interferometry into distant mirror

Distant Scalar Potential Interferometer

© T.E. BEARDEN 1997

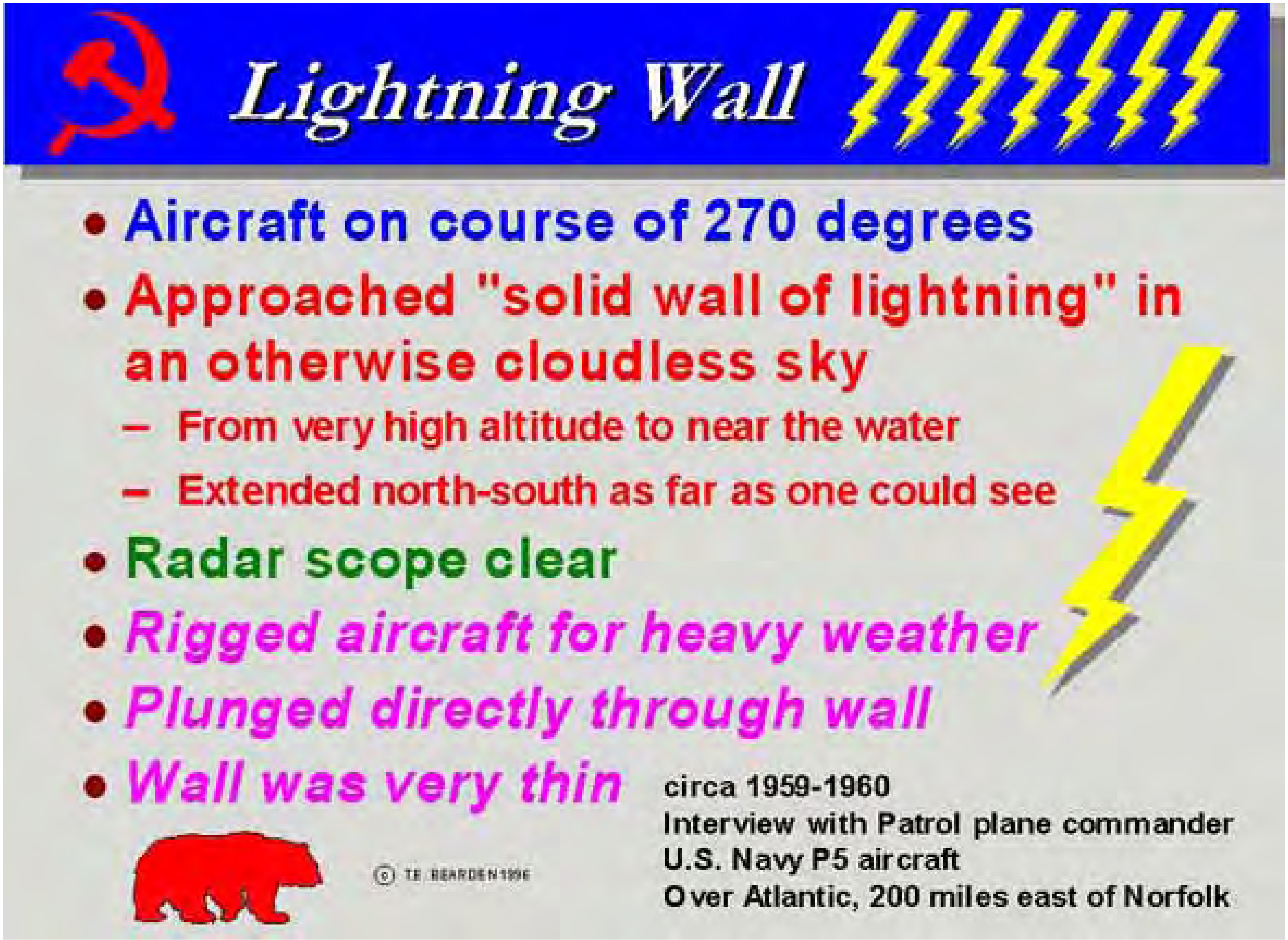

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

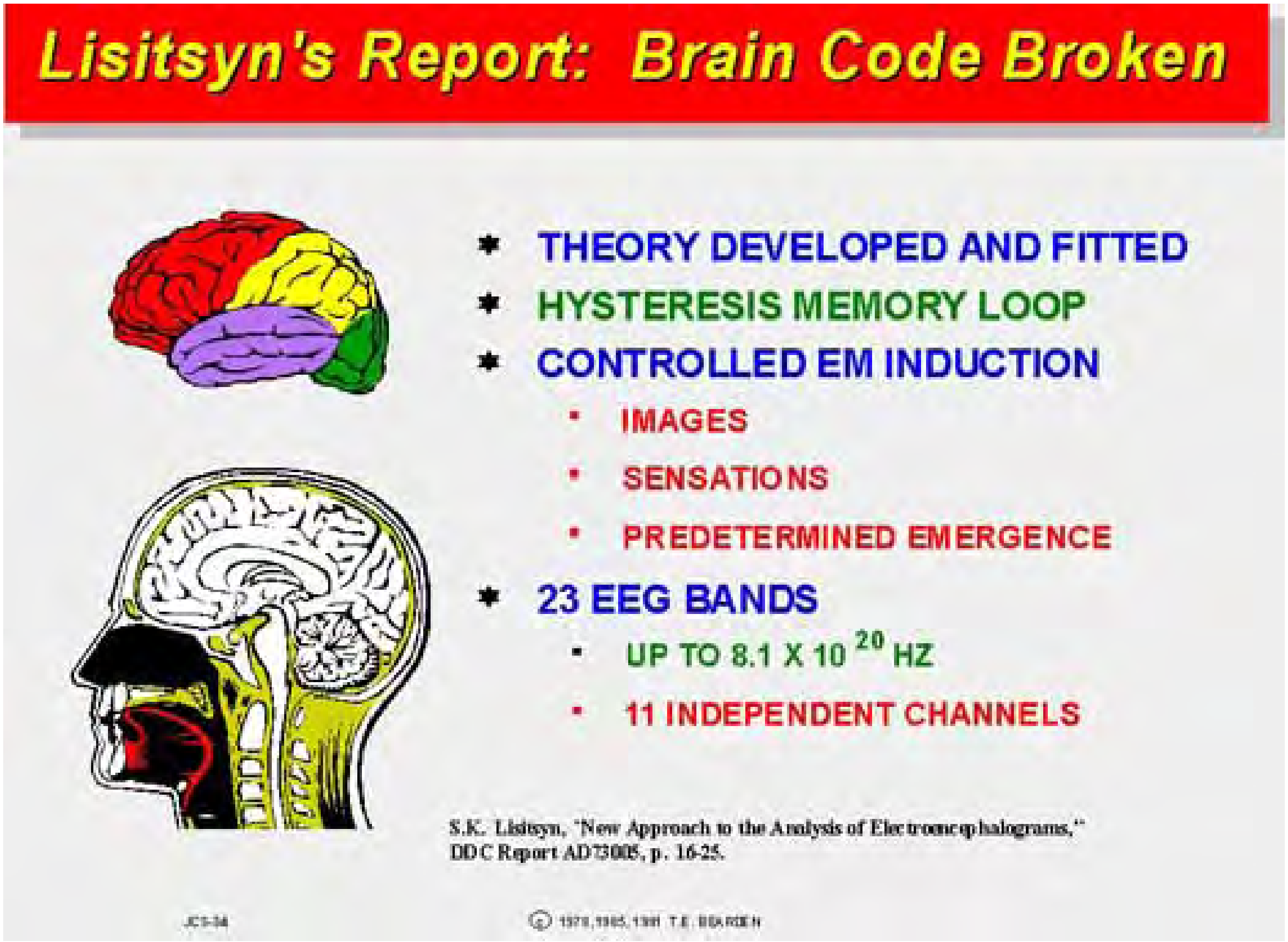

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

JCS-34

© 1978, 1985, 1991 T.E. BEARDEN

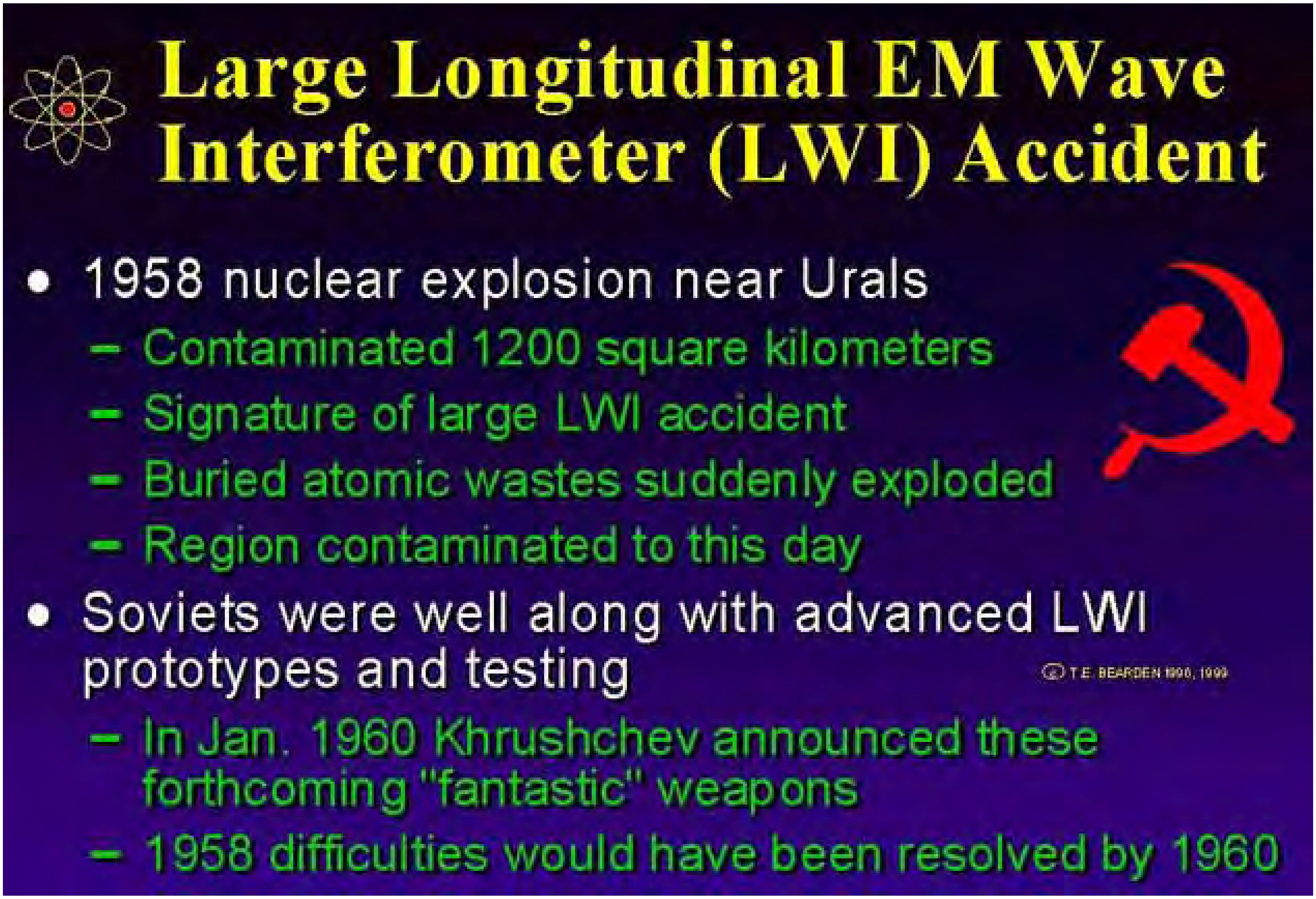

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

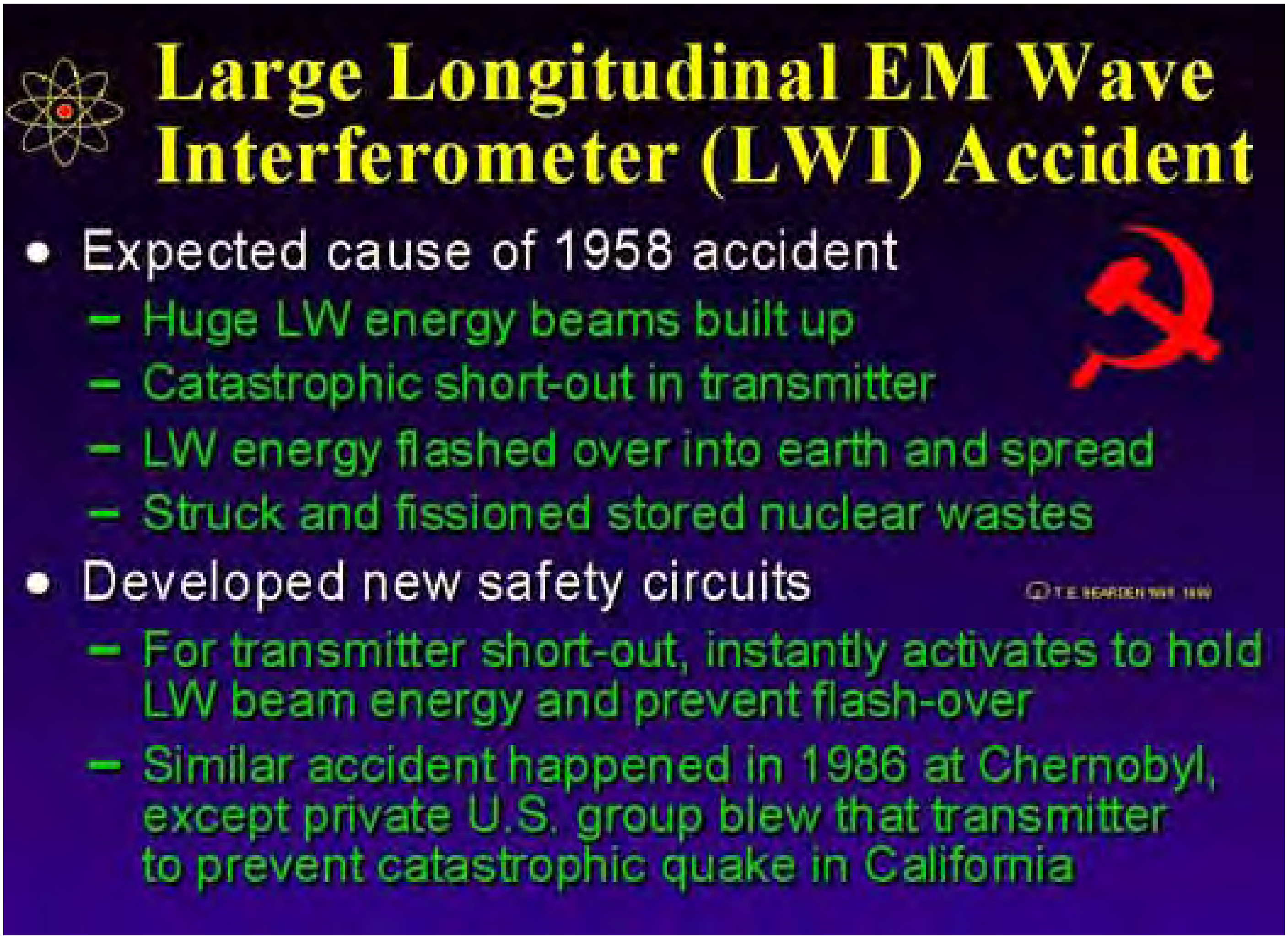

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

© T.E. BEARDEN [illegible]

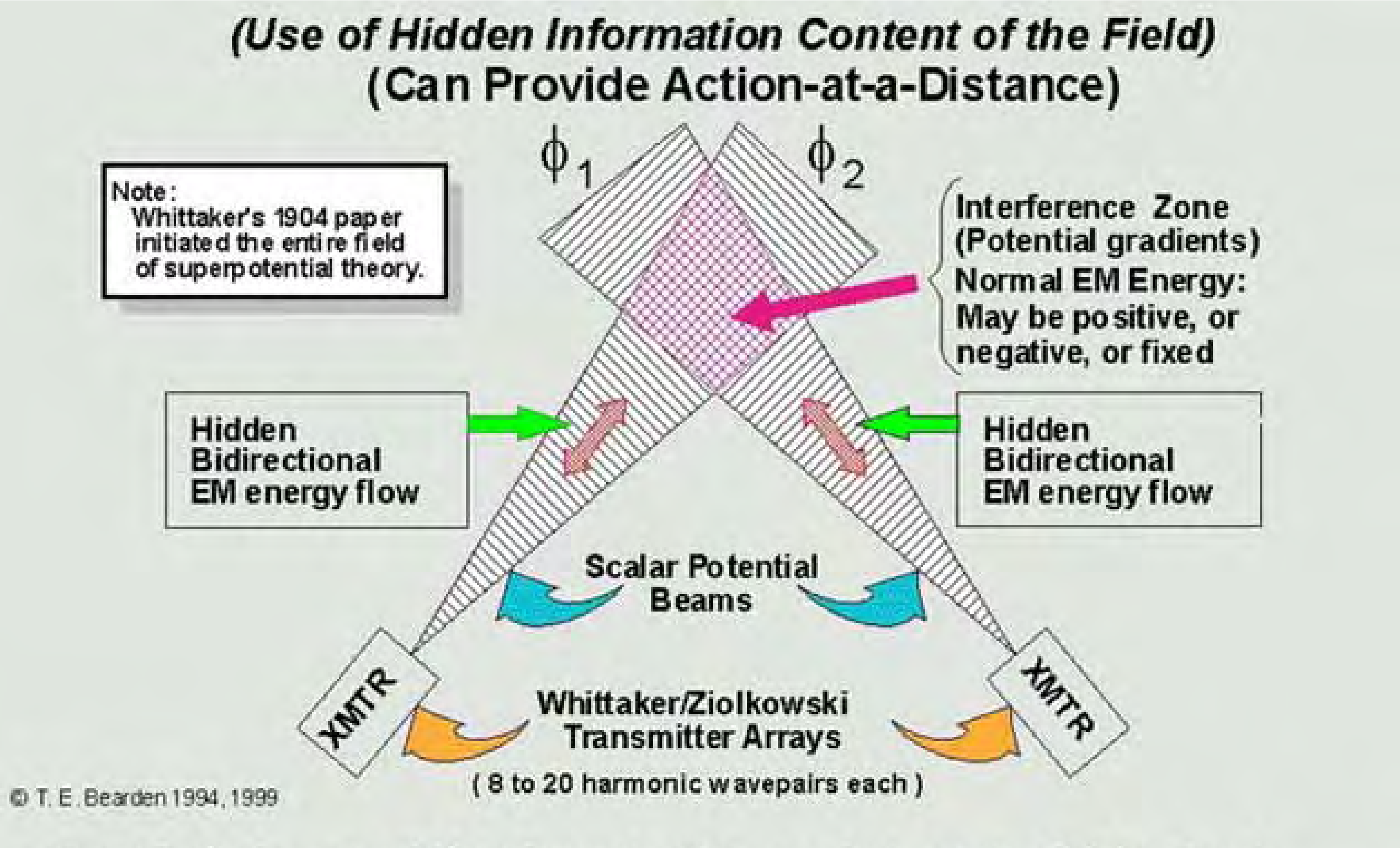
(Use of Hidden Information Content of the Field)
(Can Provide Action-at-a-Distance)
ϕ_1
ϕ_2
Note: Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

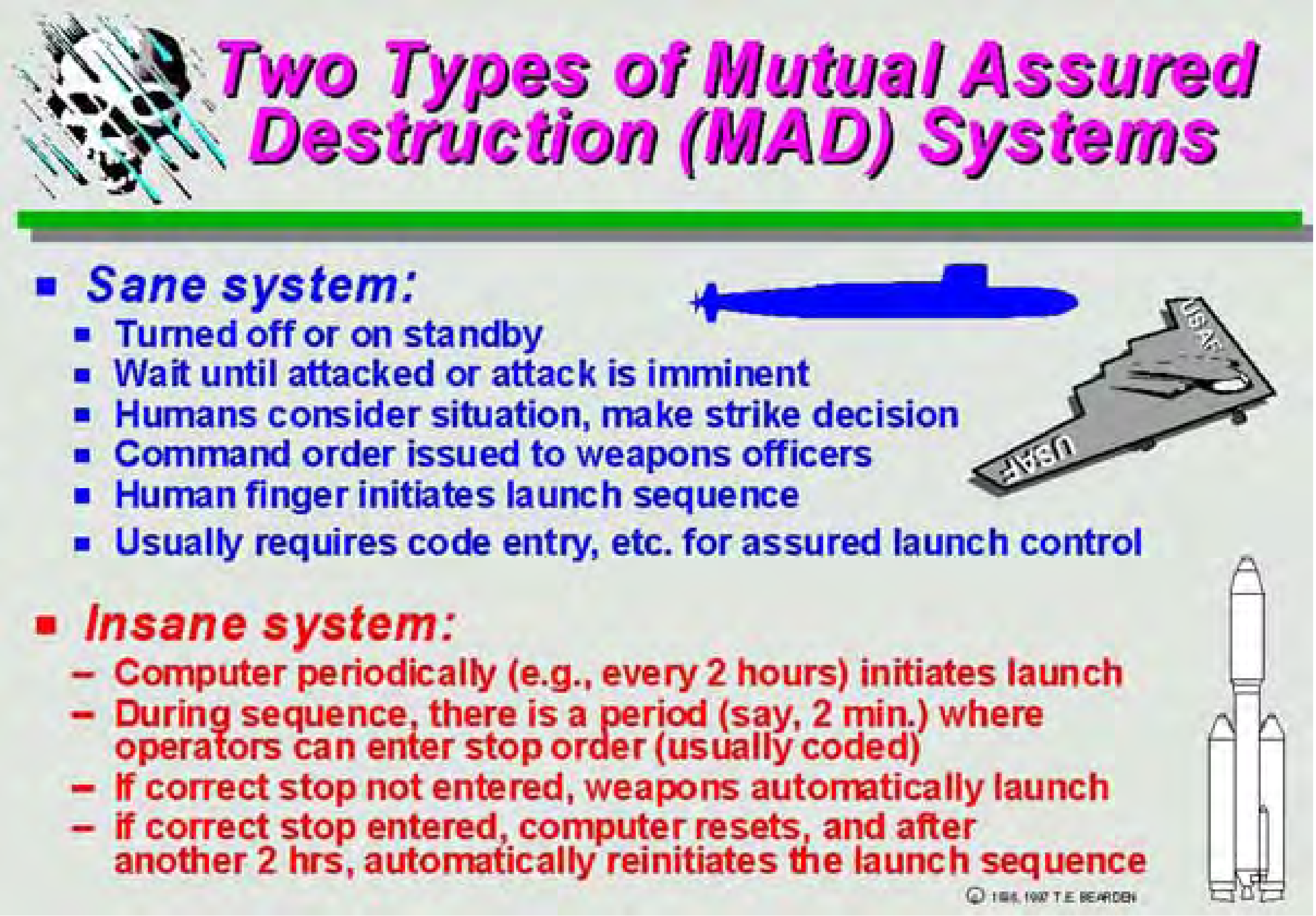

Two Types of Mutual Assured Destruction (MAD) Systems

- ***Sane system:***
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- ***Insane system:***
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

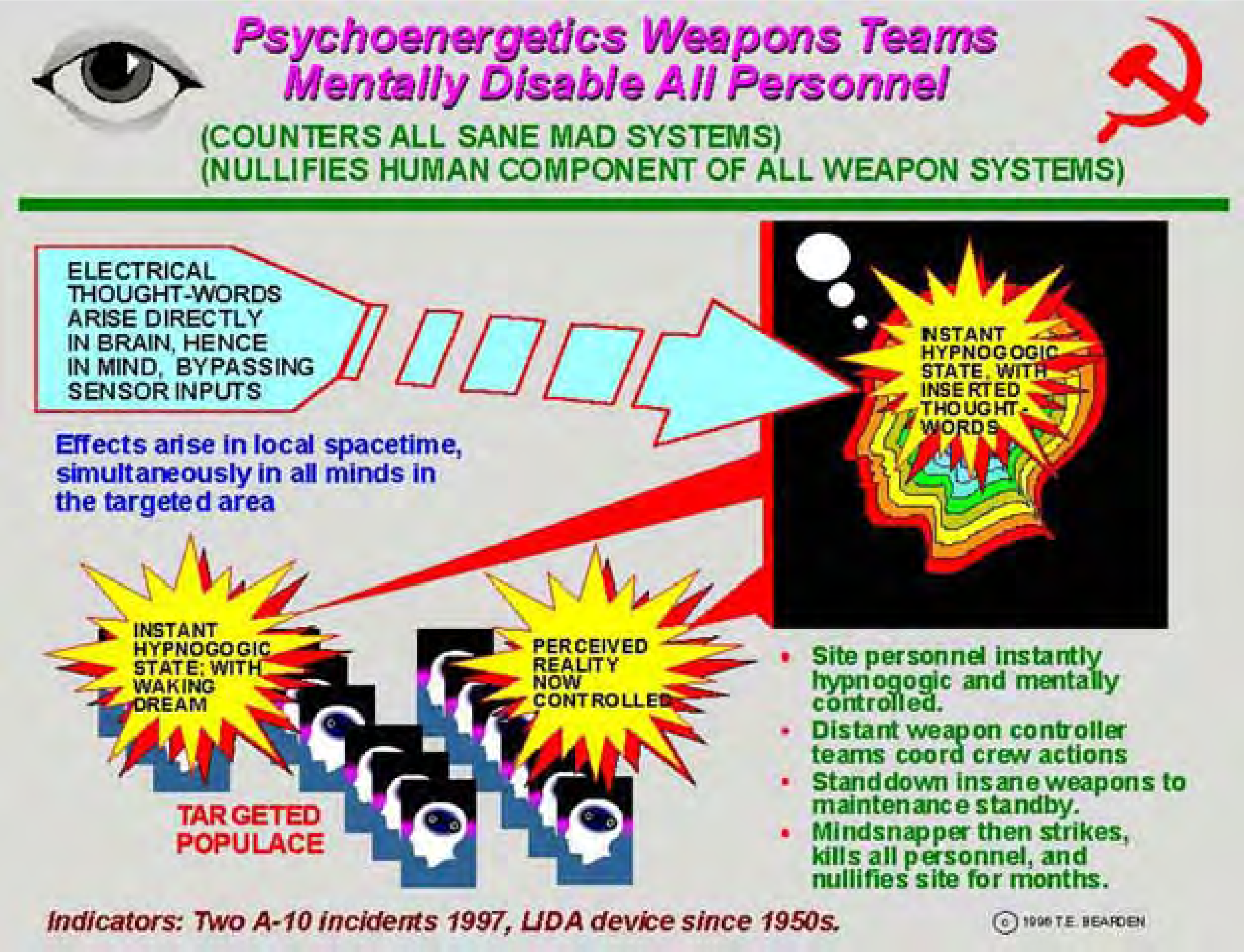

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

(c) 1998 T.E. BEARDEN

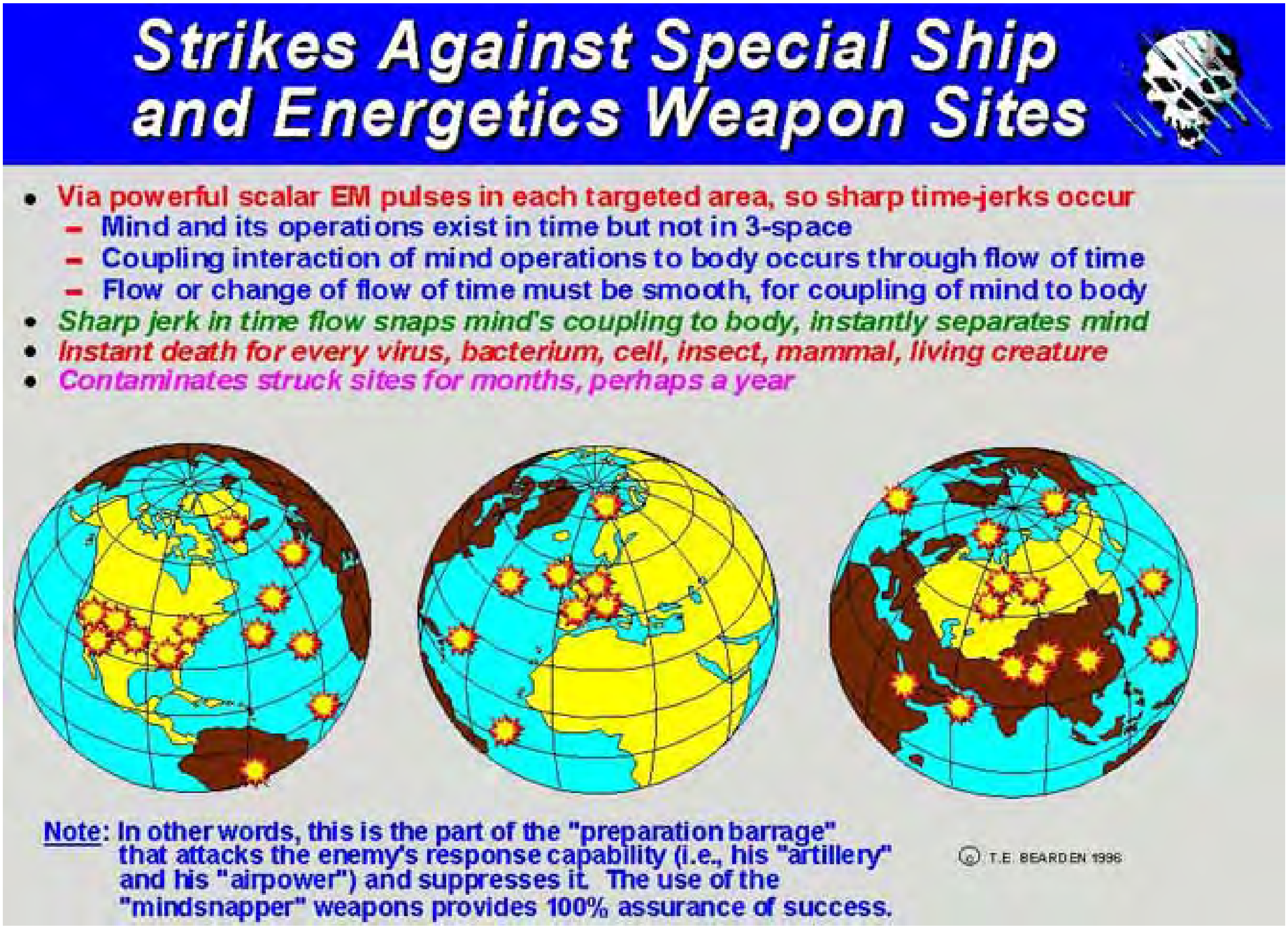

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

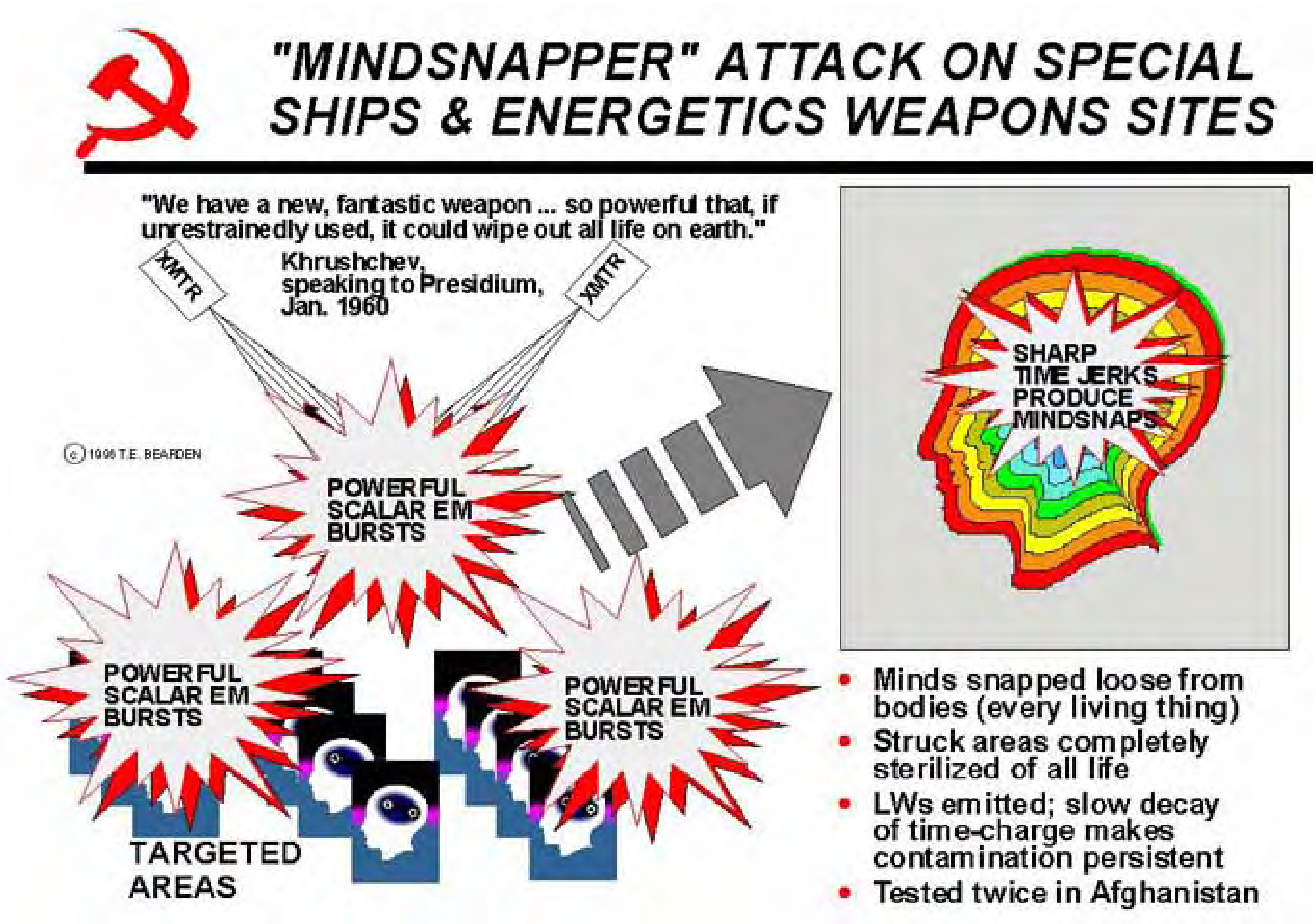

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

© 1996 T.E. BEARDEN

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

The Tom Bearden Website

Help support the research

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

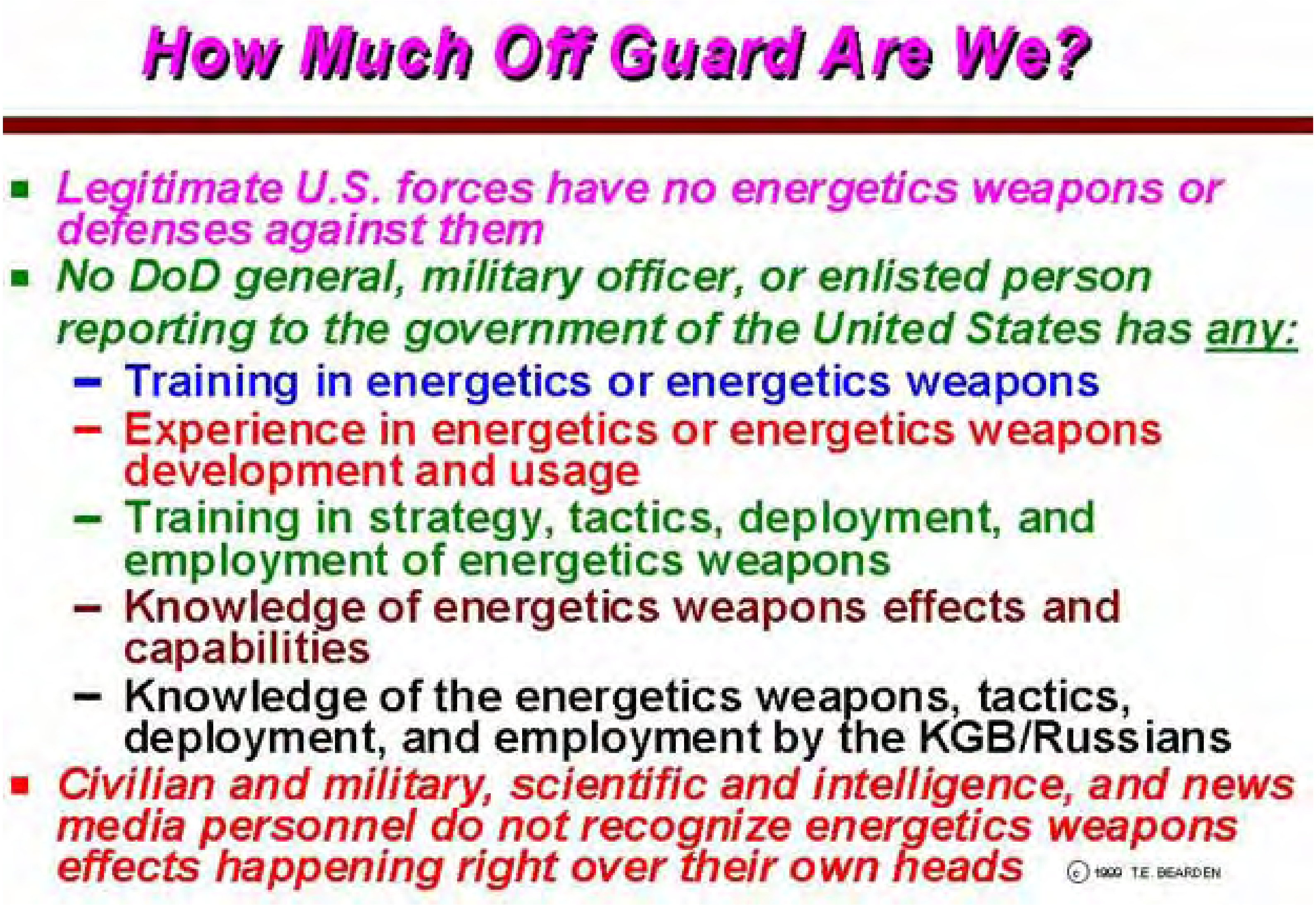

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
  - Training in energetics or energetics weapons
  - Experience in energetics or energetics weapons development and usage
  - Training in strategy, tactics, deployment, and employment of energetics weapons
  - Knowledge of energetics weapons effects and capabilities
  - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

(c) 1999 T.E. BEARDEN

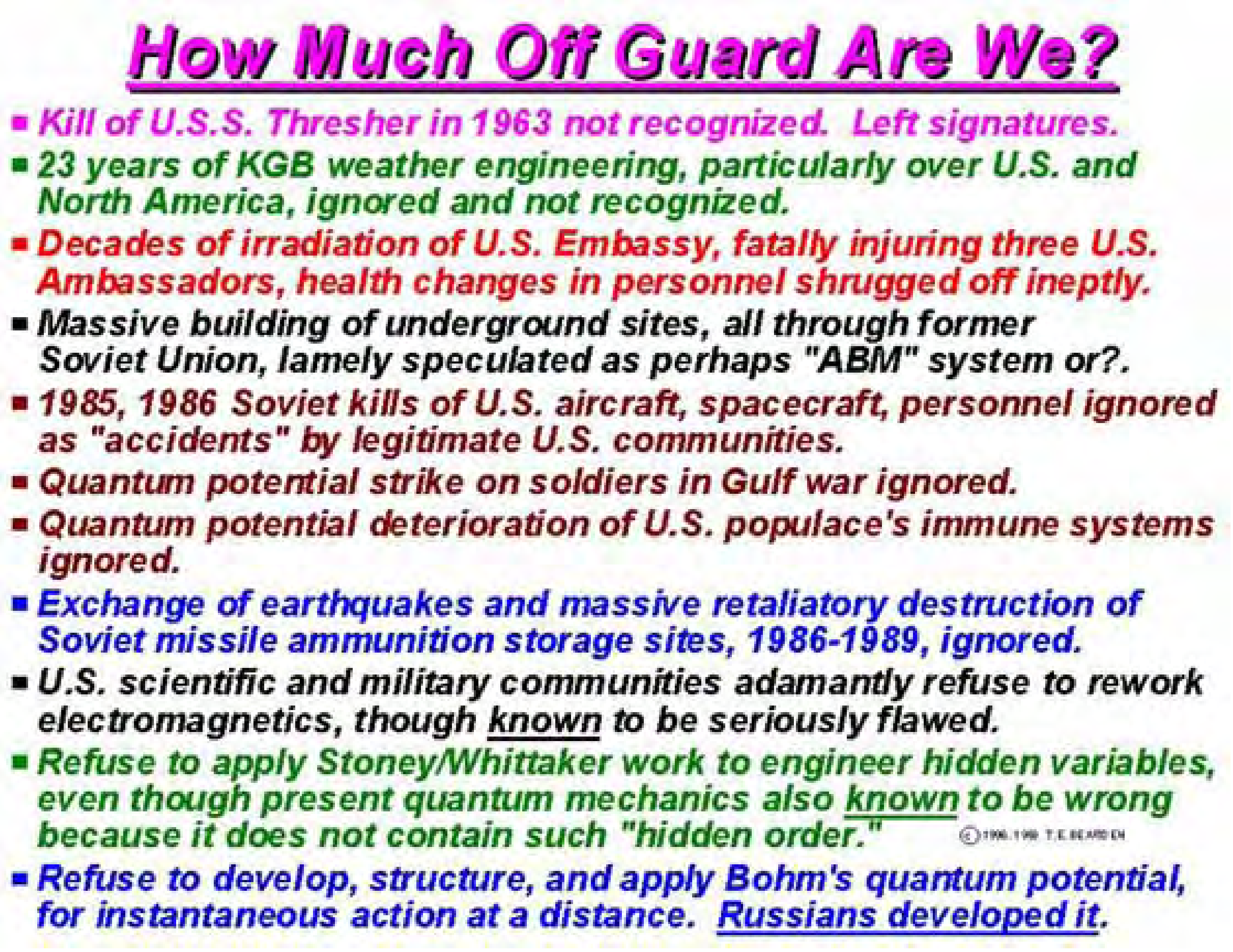

How Much Off Guard Are We?

- Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.
- 23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.
- Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.
- Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.
- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.
- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."

© 1996, 1998 T.E. BEARDEN

- Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.

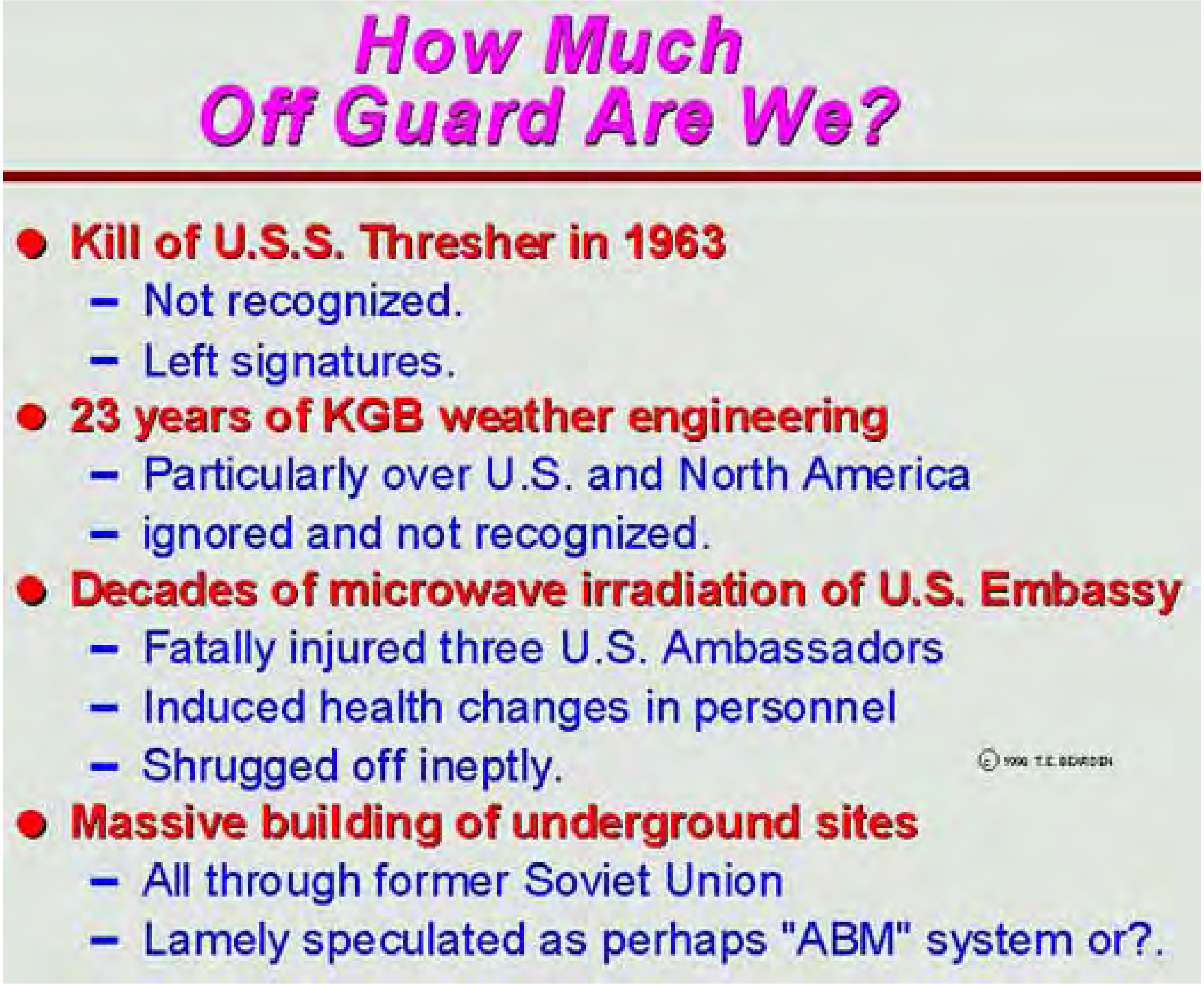

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

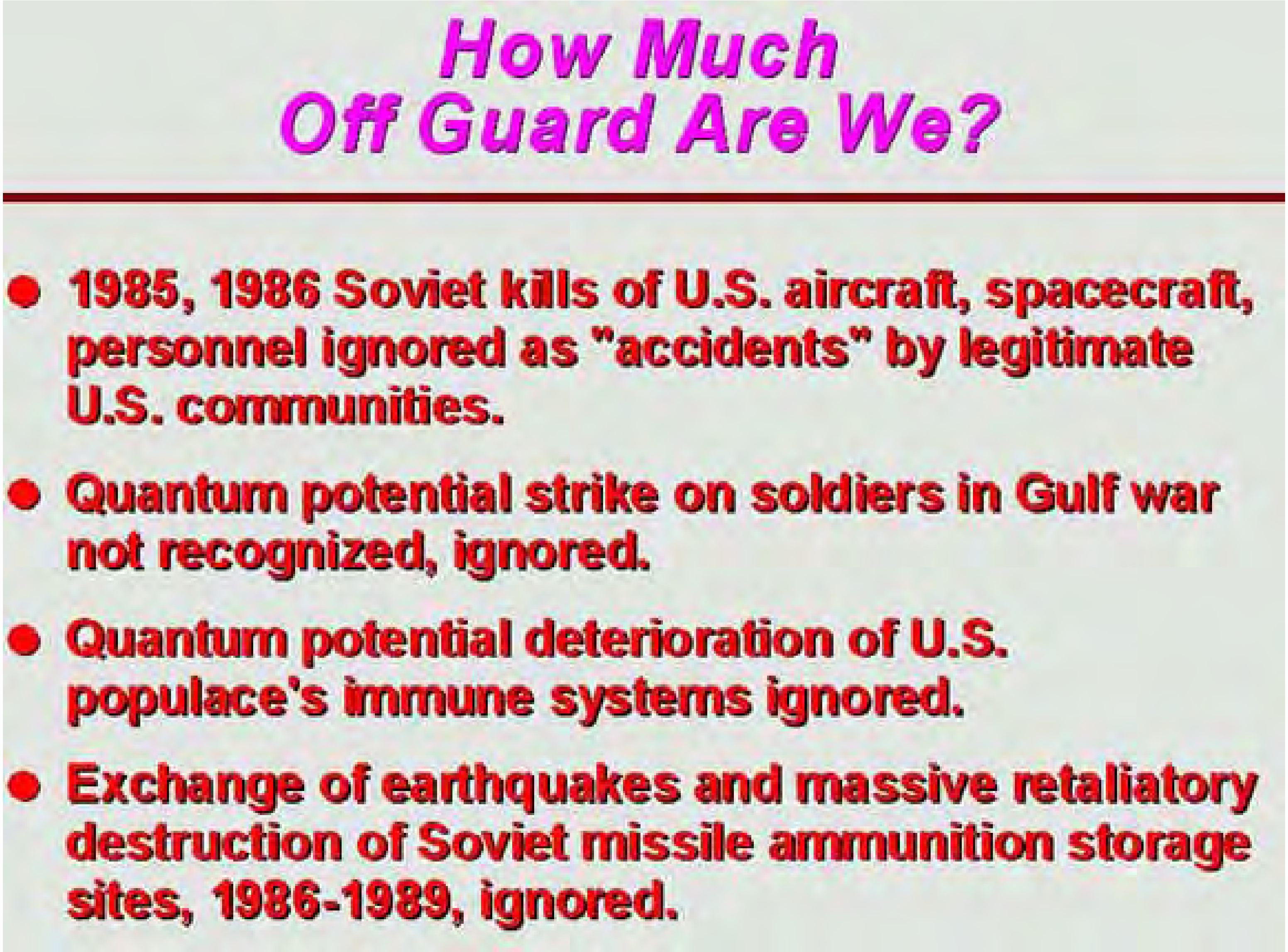

How Much Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

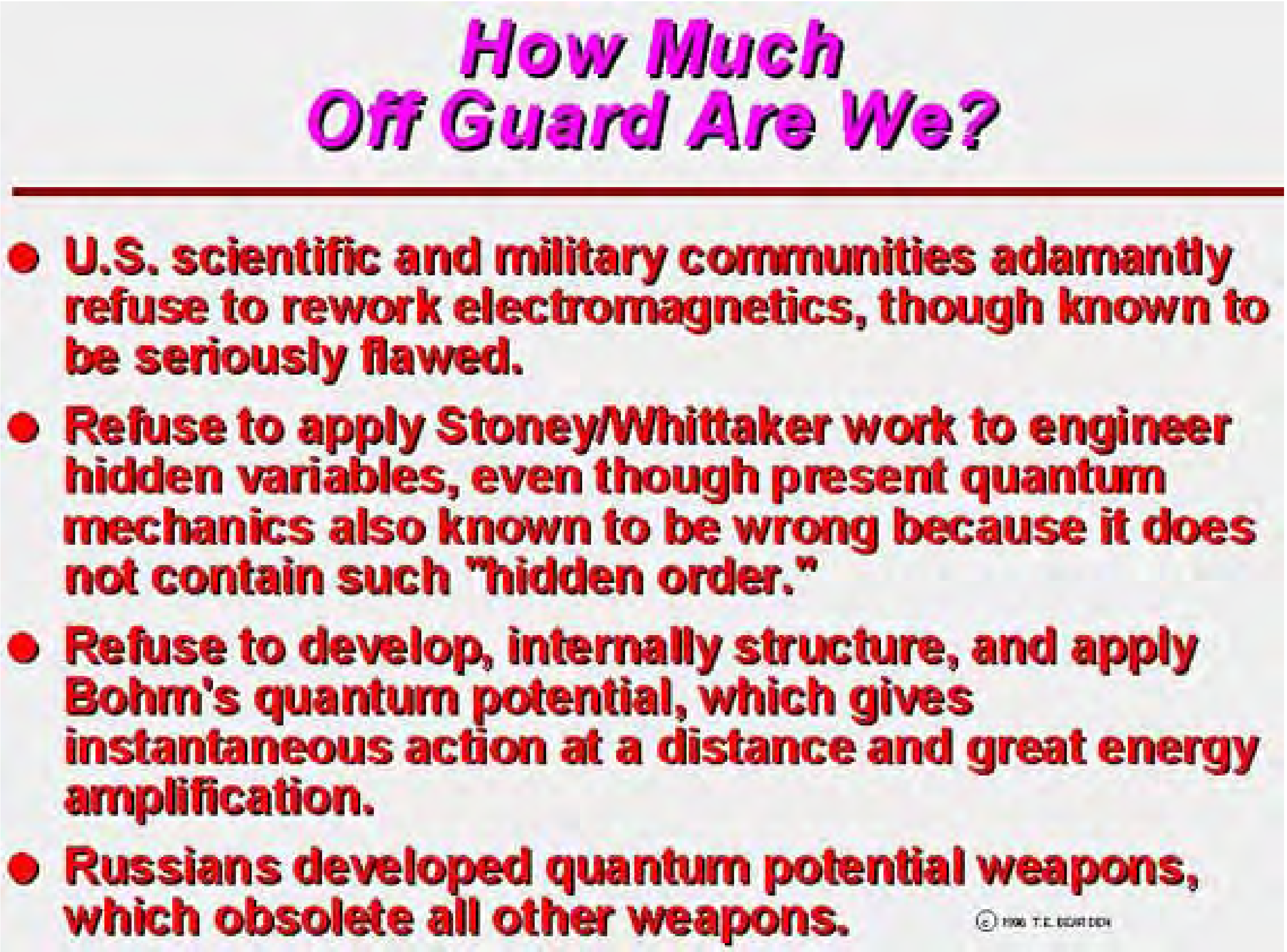

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© 1996 T.E. BEARDEN

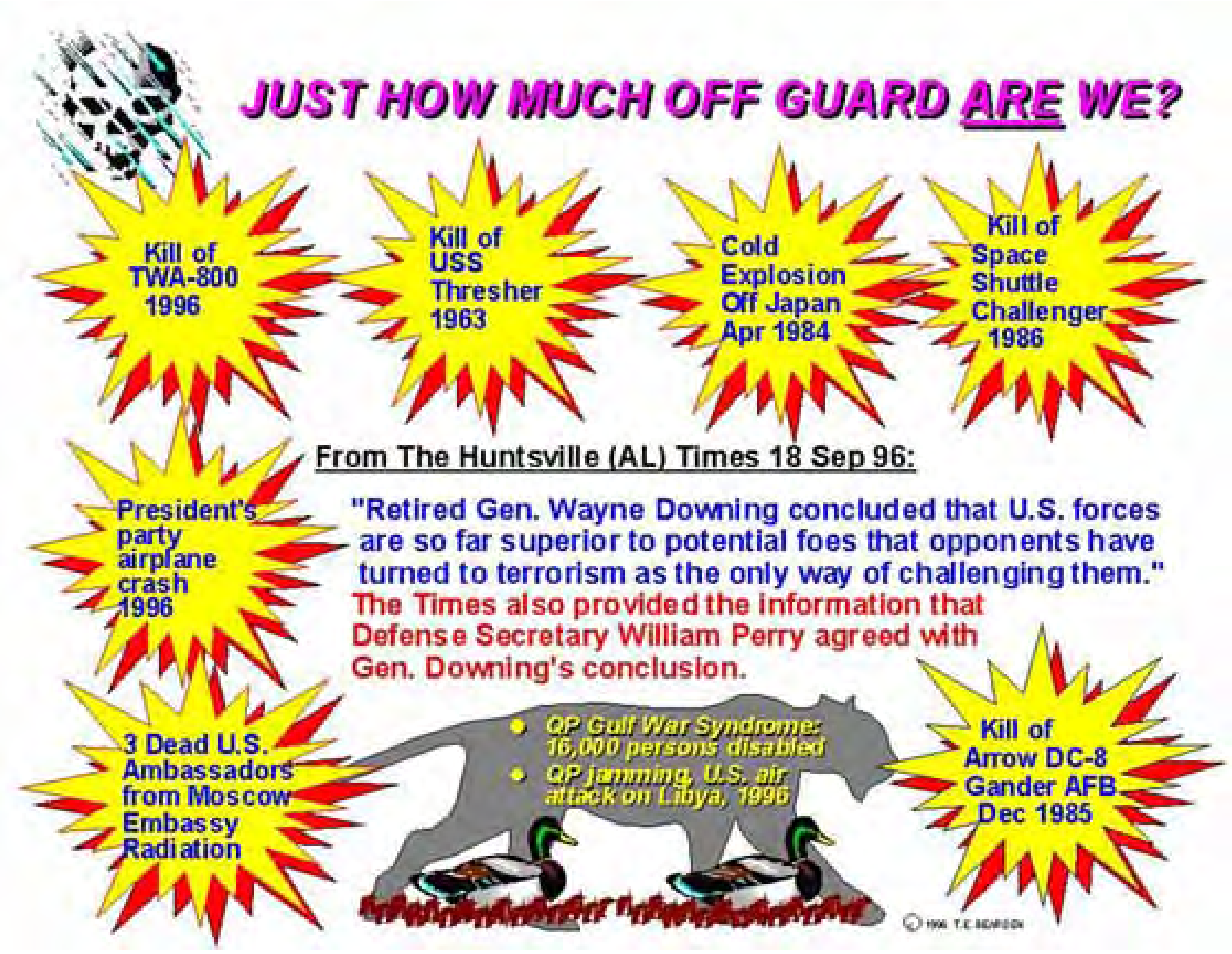

JUST HOW MUCH OFF GUARD ARE WE?

Kill of
TWA-800
1996

Kill of
USS
Thresher
1963

Cold
Explosion
Off Japan
Apr 1984

Kill of
Space
Shuttle
Challenger
1986

President's
party
airplane
crash
1996

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

3 Dead U.S.
Ambassadors
from Moscow
Embassy
Radiation

- *QP Gulf War Syndrome: 16,000 persons disabled*
- *QP jamming, U.S. air attack on Libya, 1996*

Kill of
Arrow DC-8
Gander AFB
Dec 1985

© 1996 T.E. BEARDEN

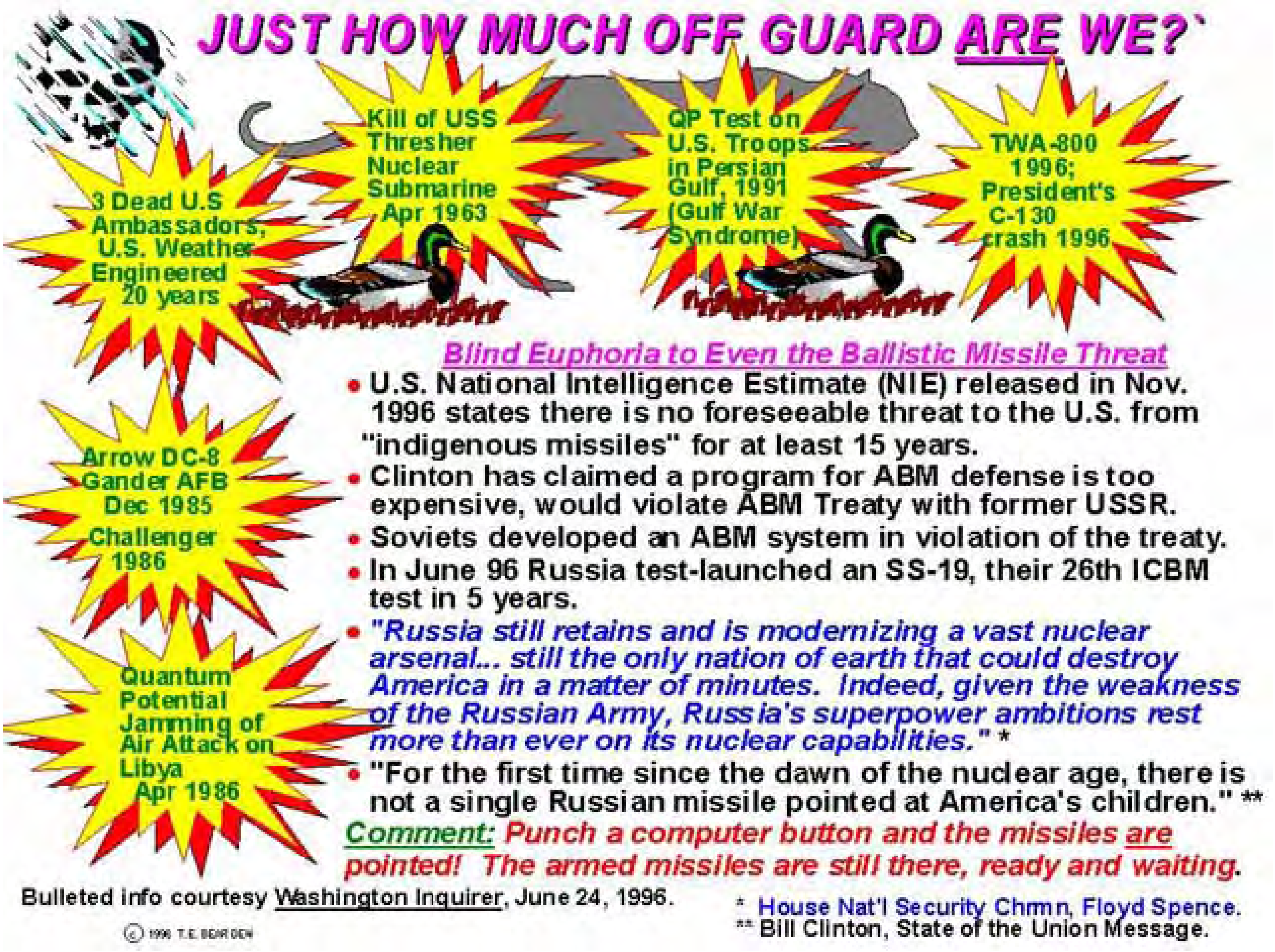

JUST HOW MUCH OFF GUARD ARE WE?`

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

© 1996 T.E. BEARDEN

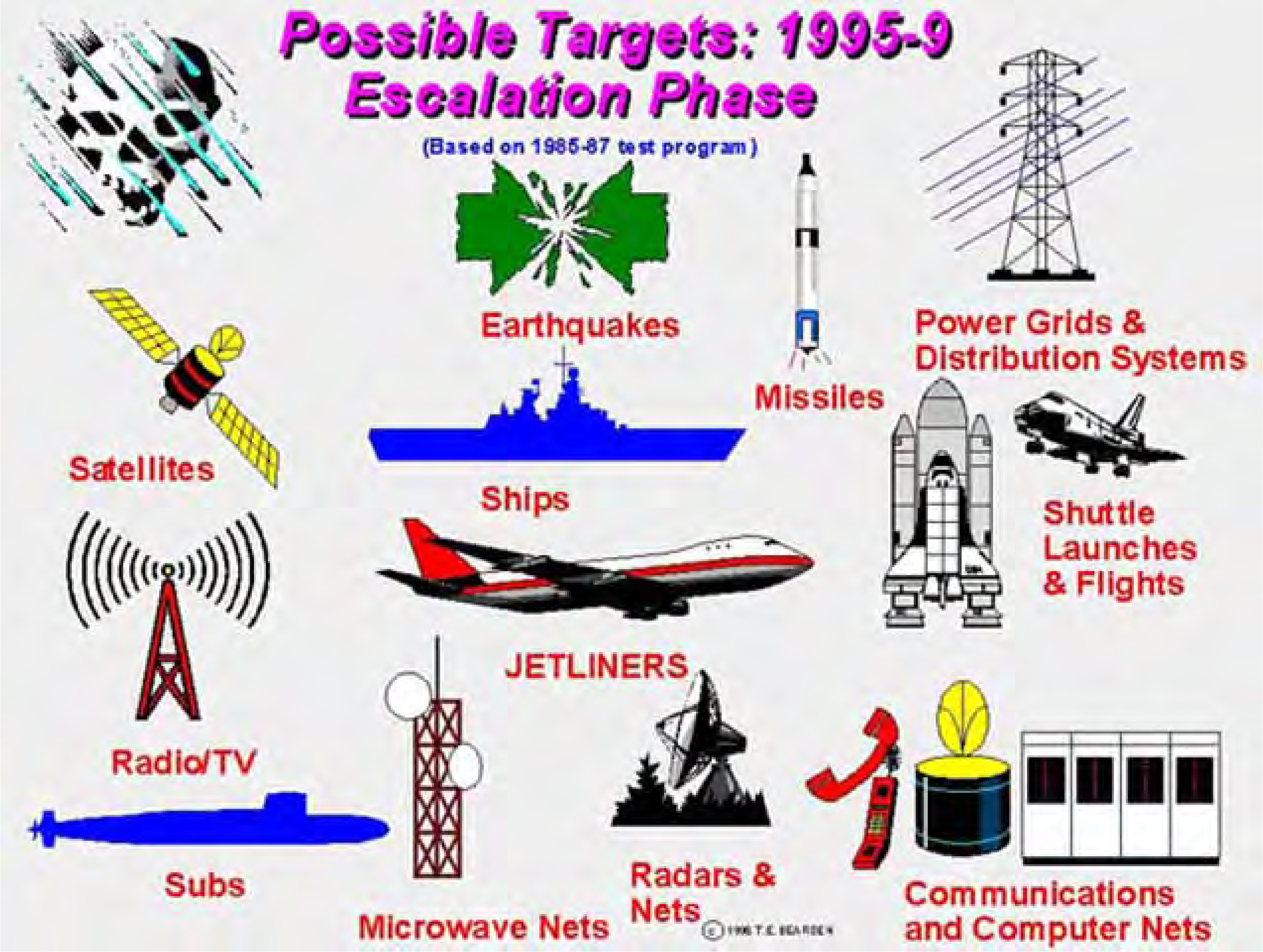
Possible Targets: 1995-9
Escalation Phase
(Based on 1985-87 test program)
Earthquakes
Missiles
Power Grids &
Distribution Systems
Satellites
Ships
Shuttle
Launches
& Flights
JETLINERS
Radio/TV
Subs
Microwave Nets
Radars &
Nets
Communications
and Computer Nets

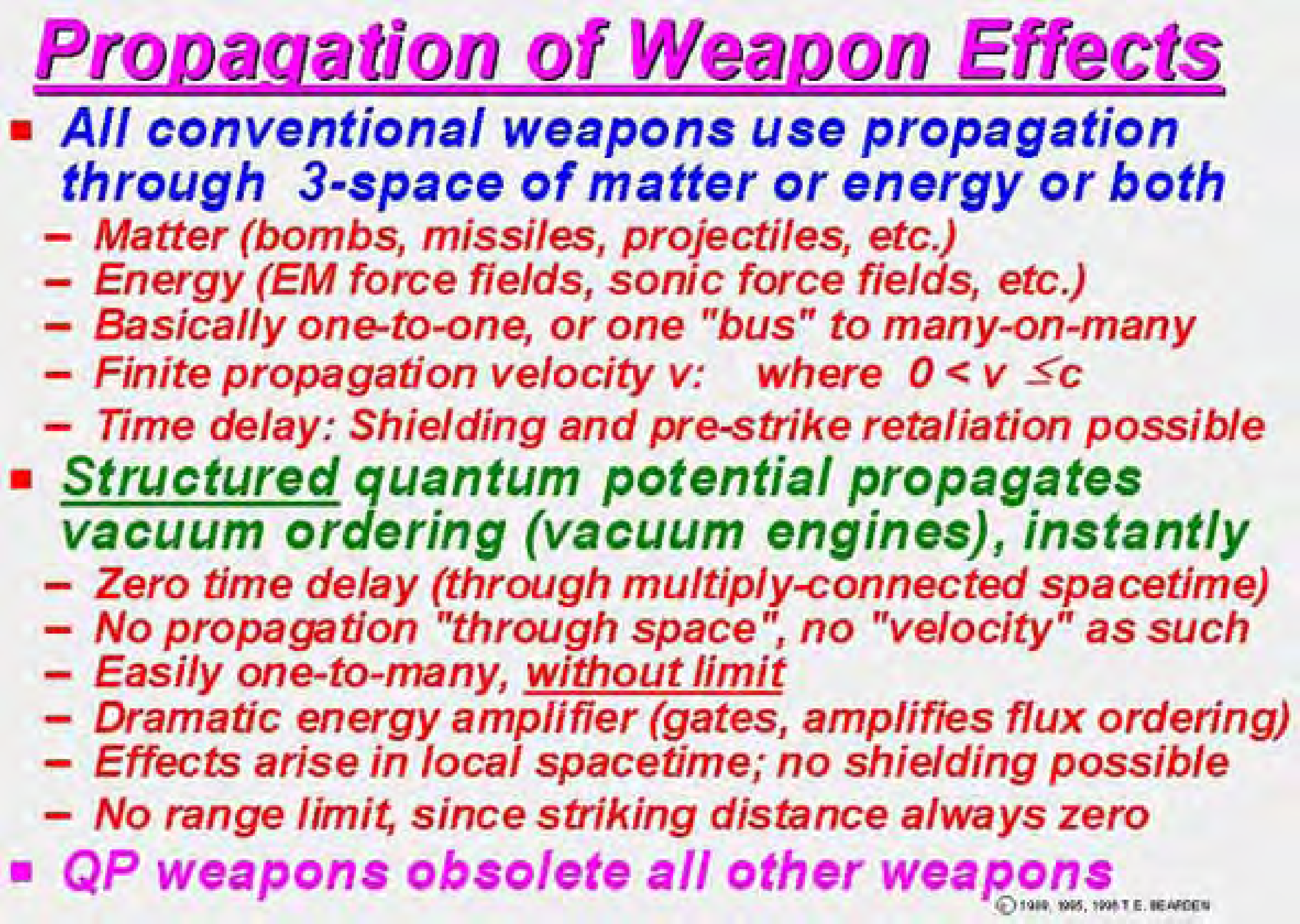

Propagation of Weapon Effects

- **All conventional weapons use propagation through 3-space of matter or energy or both**
  - *Matter (bombs, missiles, projectiles, etc.)*
  - *Energy (EM force fields, sonic force fields, etc.)*
  - *Basically one-to-one, or one "bus" to many-on-many*
  - *Finite propagation velocity v: where $0 < v \leq c$*
  - *Time delay: Shielding and pre-strike retaliation possible*
- **Structured quantum potential propagates vacuum ordering (vacuum engines), instantly**
  - *Zero time delay (through multiply-connected spacetime)*
  - *No propagation "through space", no "velocity" as such*
  - *Easily one-to-many, without limit*
  - *Dramatic energy amplifier (gates, amplifies flux ordering)*
  - *Effects arise in local spacetime; no shielding possible*
  - *No range limit, since striking distance always zero*
- **QP weapons obsolete all other weapons**

© 1988, 1995, 1996 T.E. BEARDEN

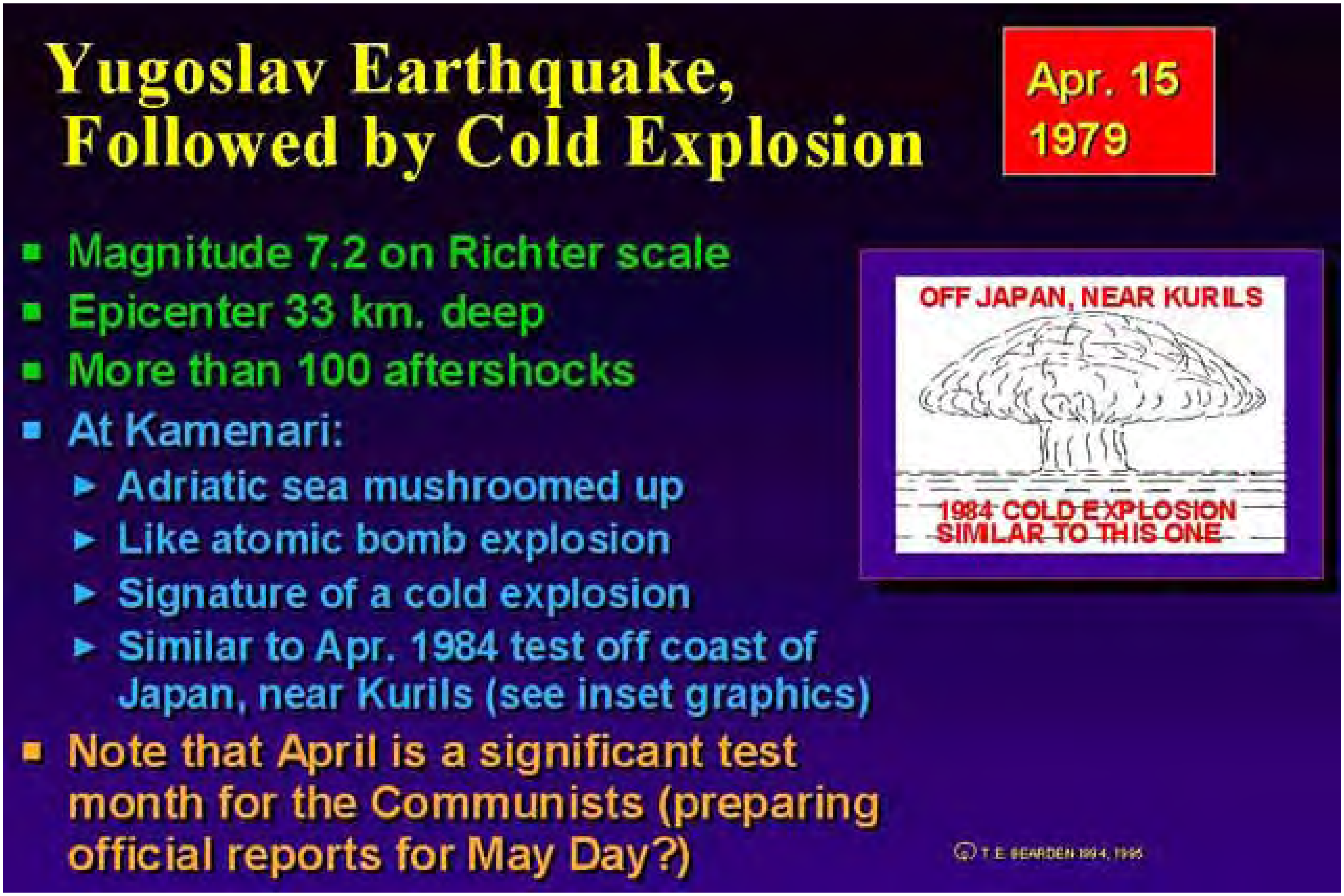

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

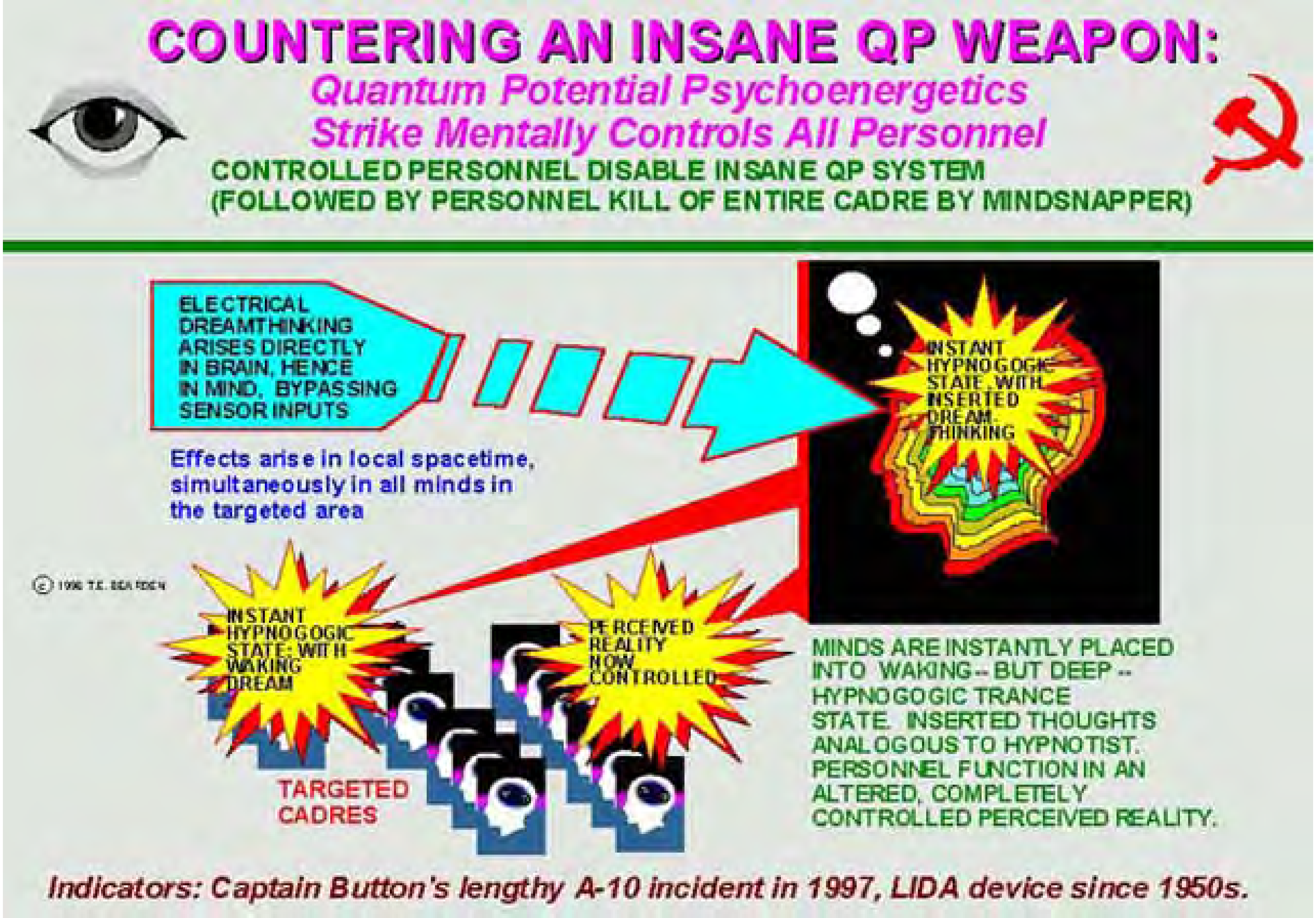

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING

© 1998 T.E. BEARDEN

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED CADRES

MINDS ARE INSTANTLY PLACED INTO WAKING-- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

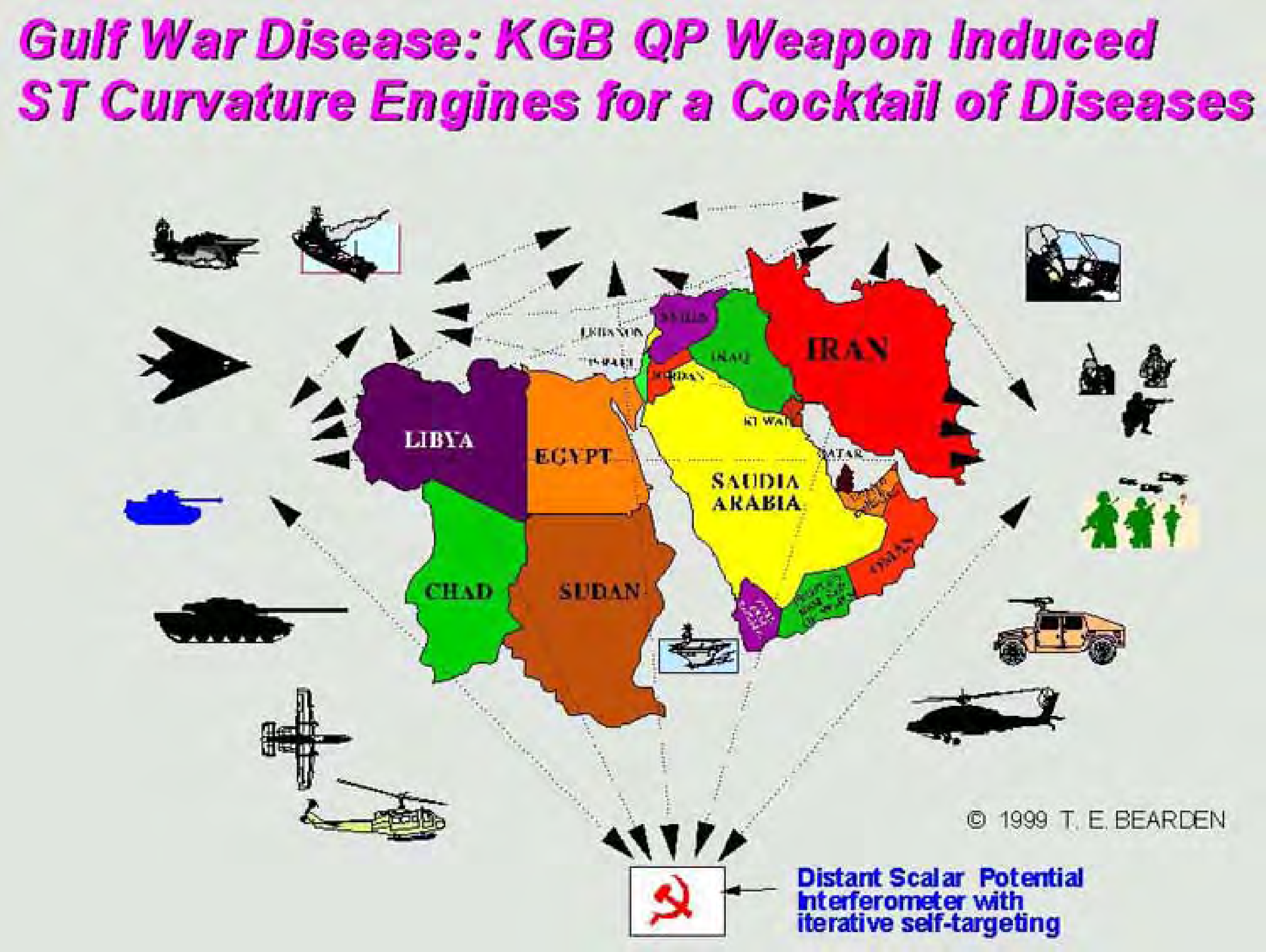

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

LEBANON
ISRAEL
SYRIA
JORDAN
IRAQ
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
CHAD
SUDAN
OMAN

© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

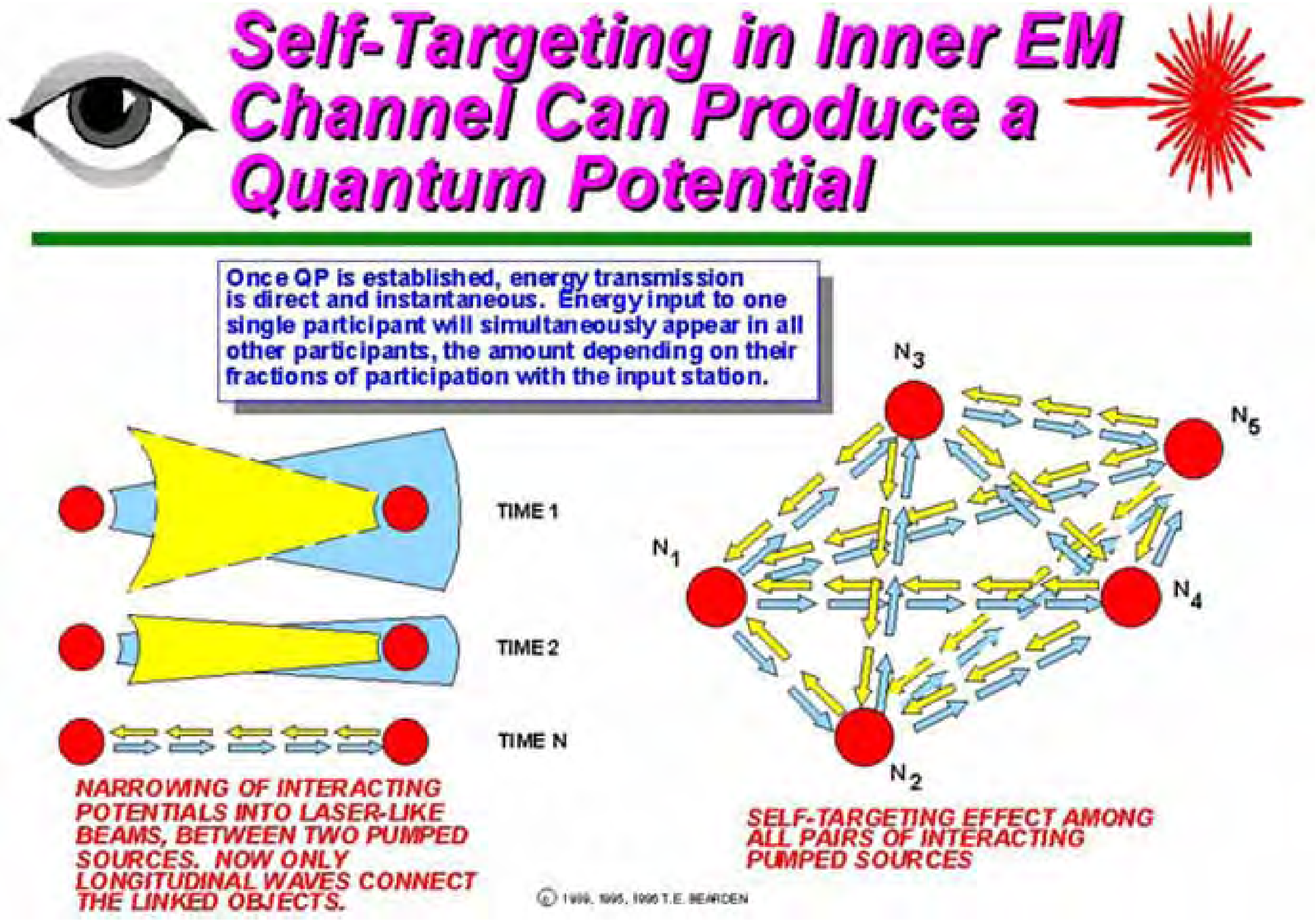

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

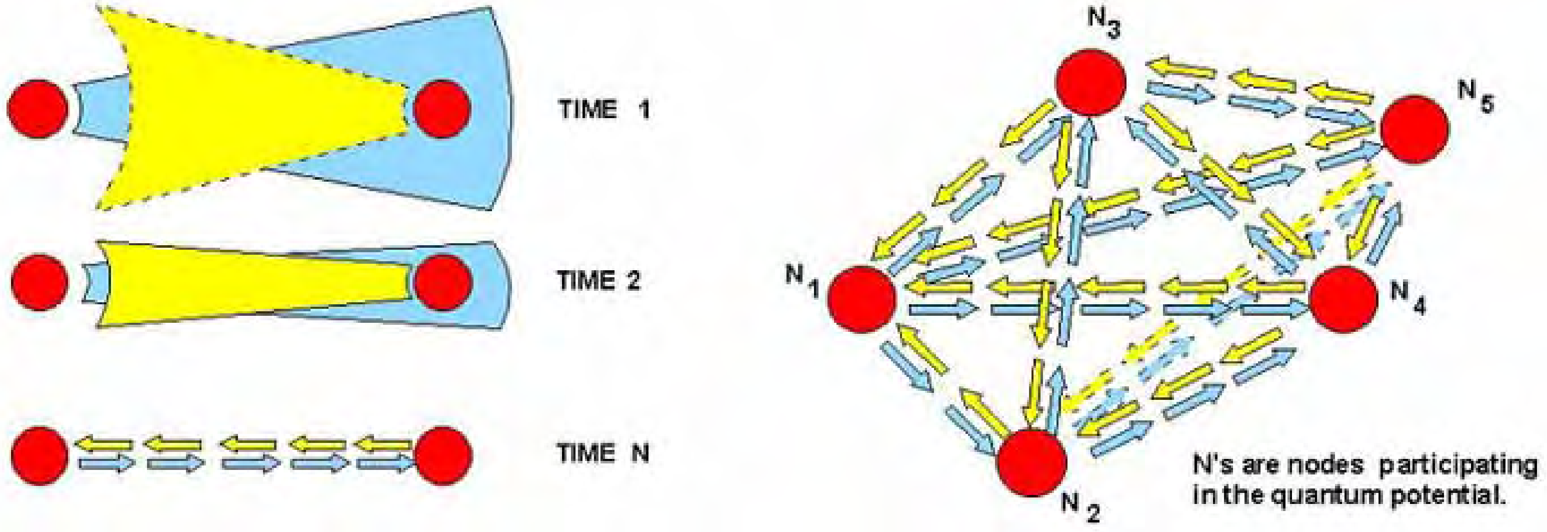

N's are nodes participating in the quantum potential.

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

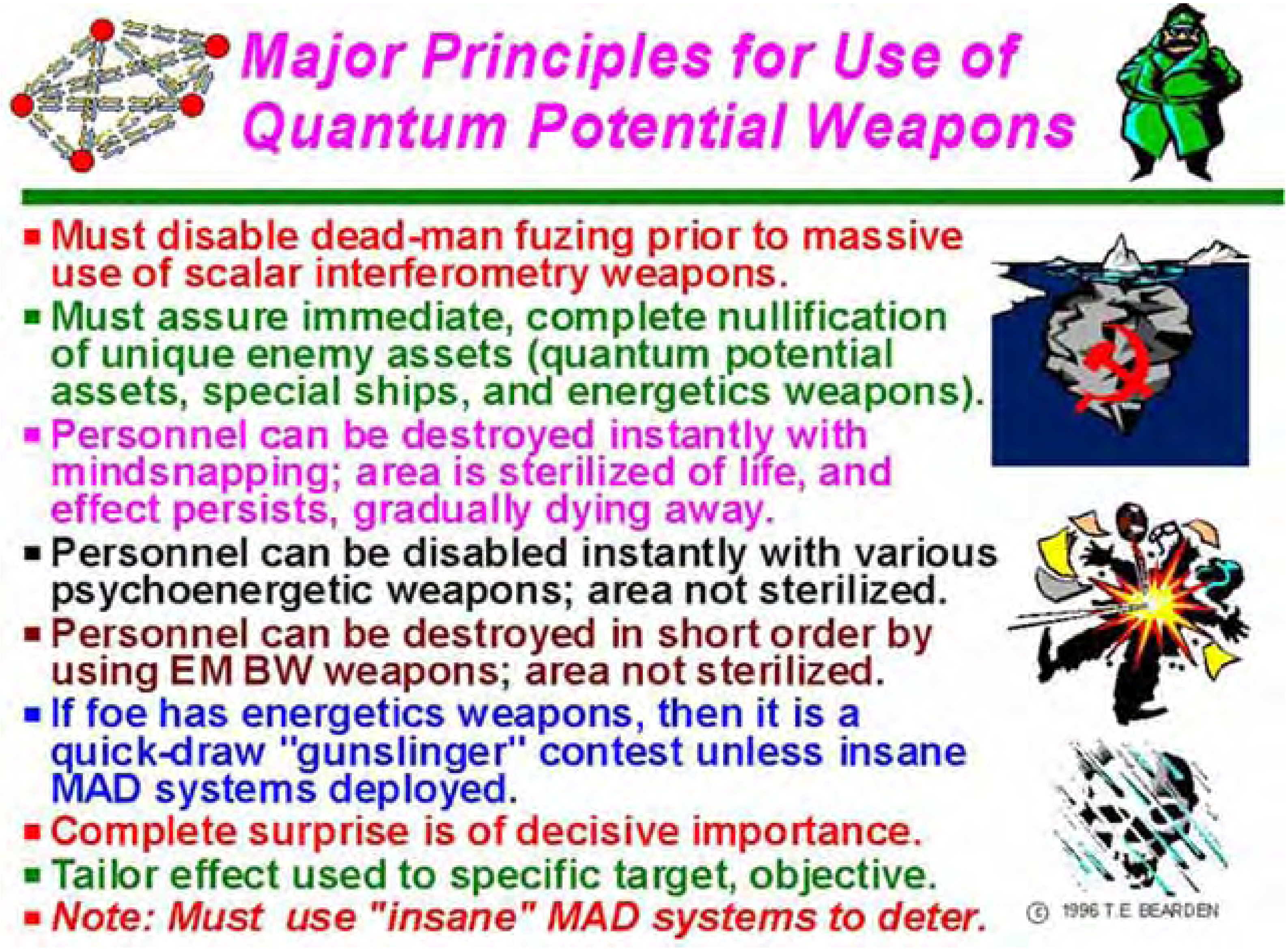

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

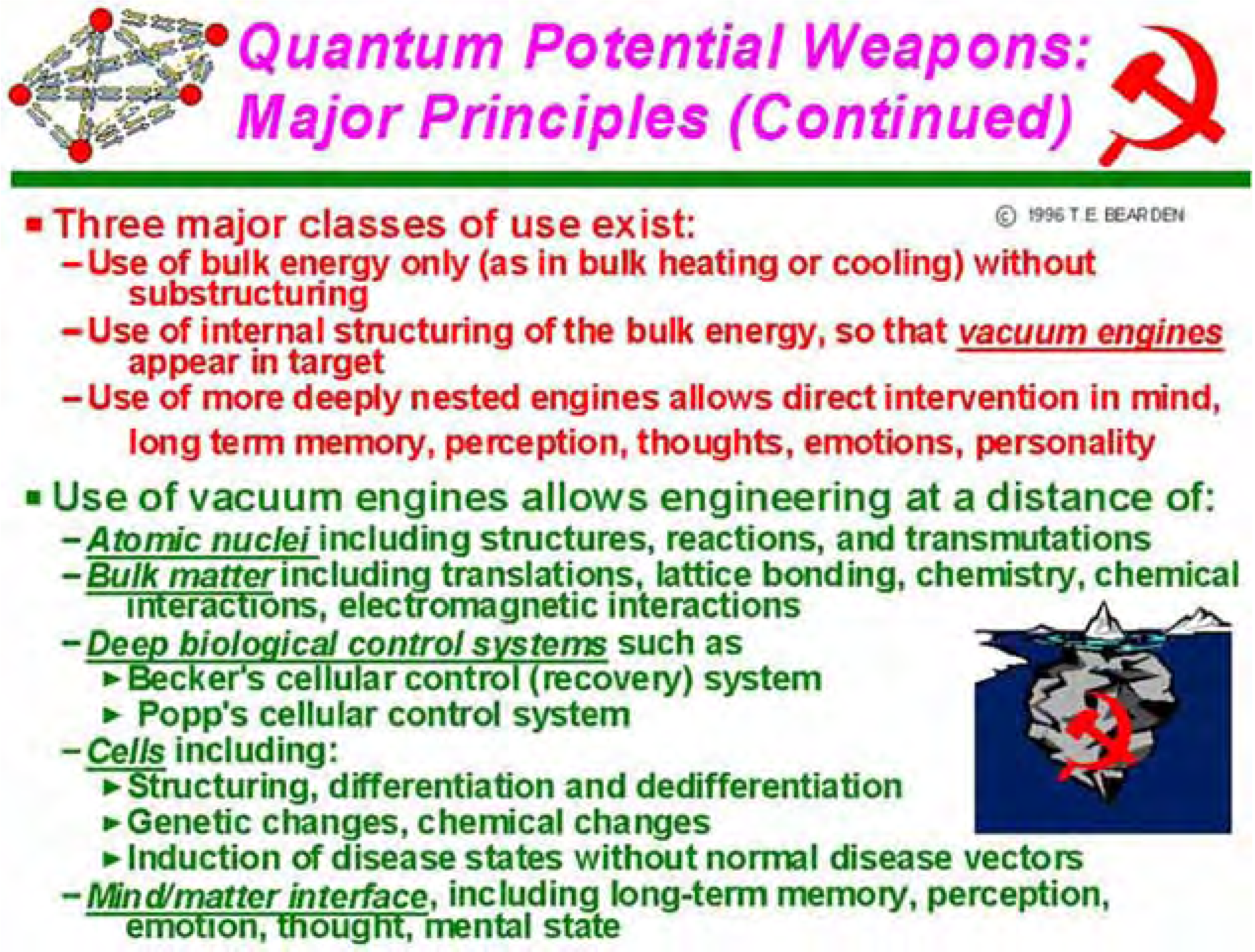

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T. E. BEARDEN

- Three major classes of use exist:
  - Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - *Atomic nuclei* including structures, reactions, and transmutations
  - *Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - *Deep biological control systems* such as
    - ► Becker's cellular control (recovery) system
    - ► Popp's cellular control system
  - *Cells* including:
    - ► Structuring, differentiation and dedifferentiation
    - ► Genetic changes, chemical changes
    - ► Induction of disease states without normal disease vectors
  - *Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

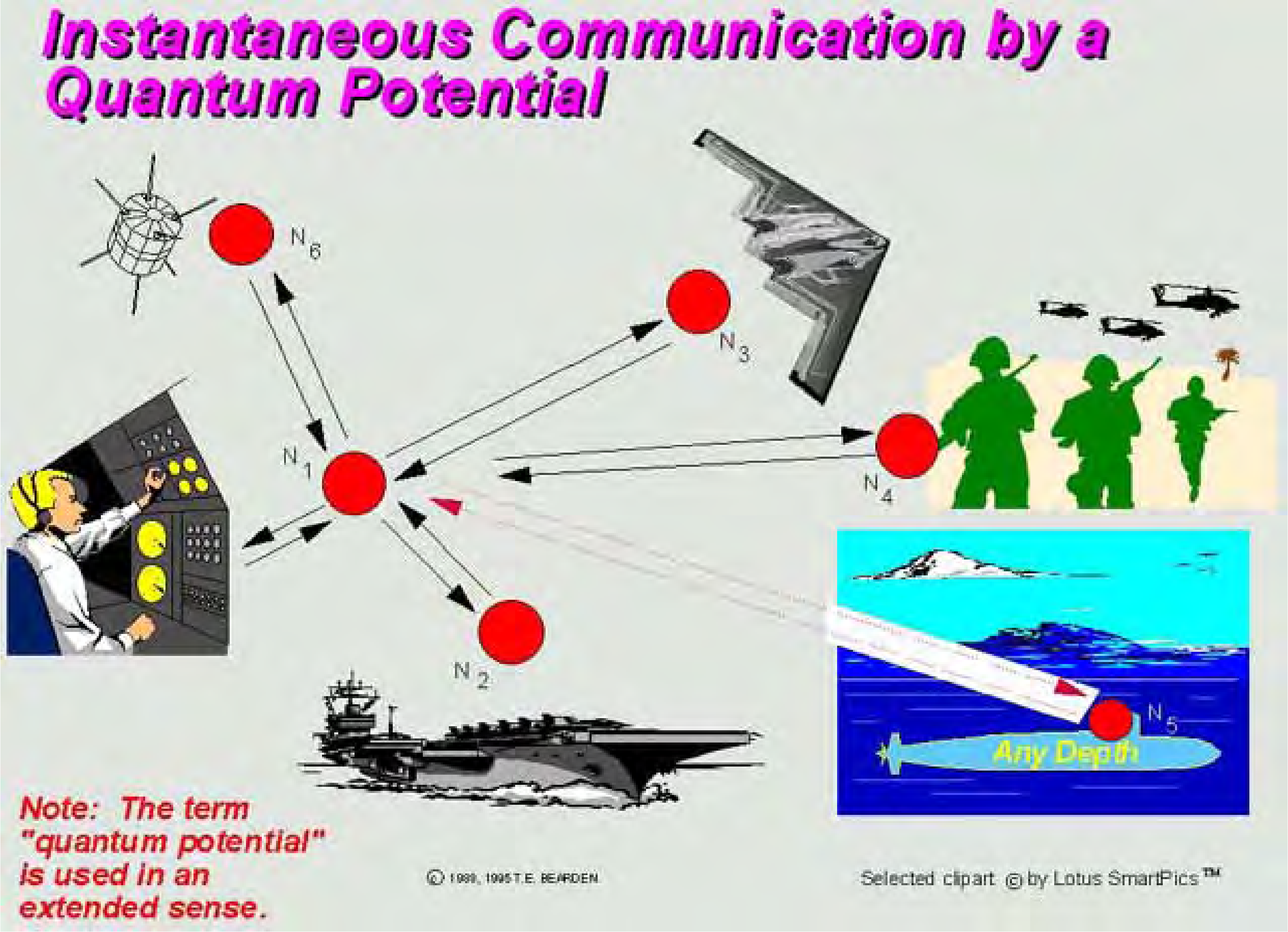

Instantaneous Communication by a Quantum Potential

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics™

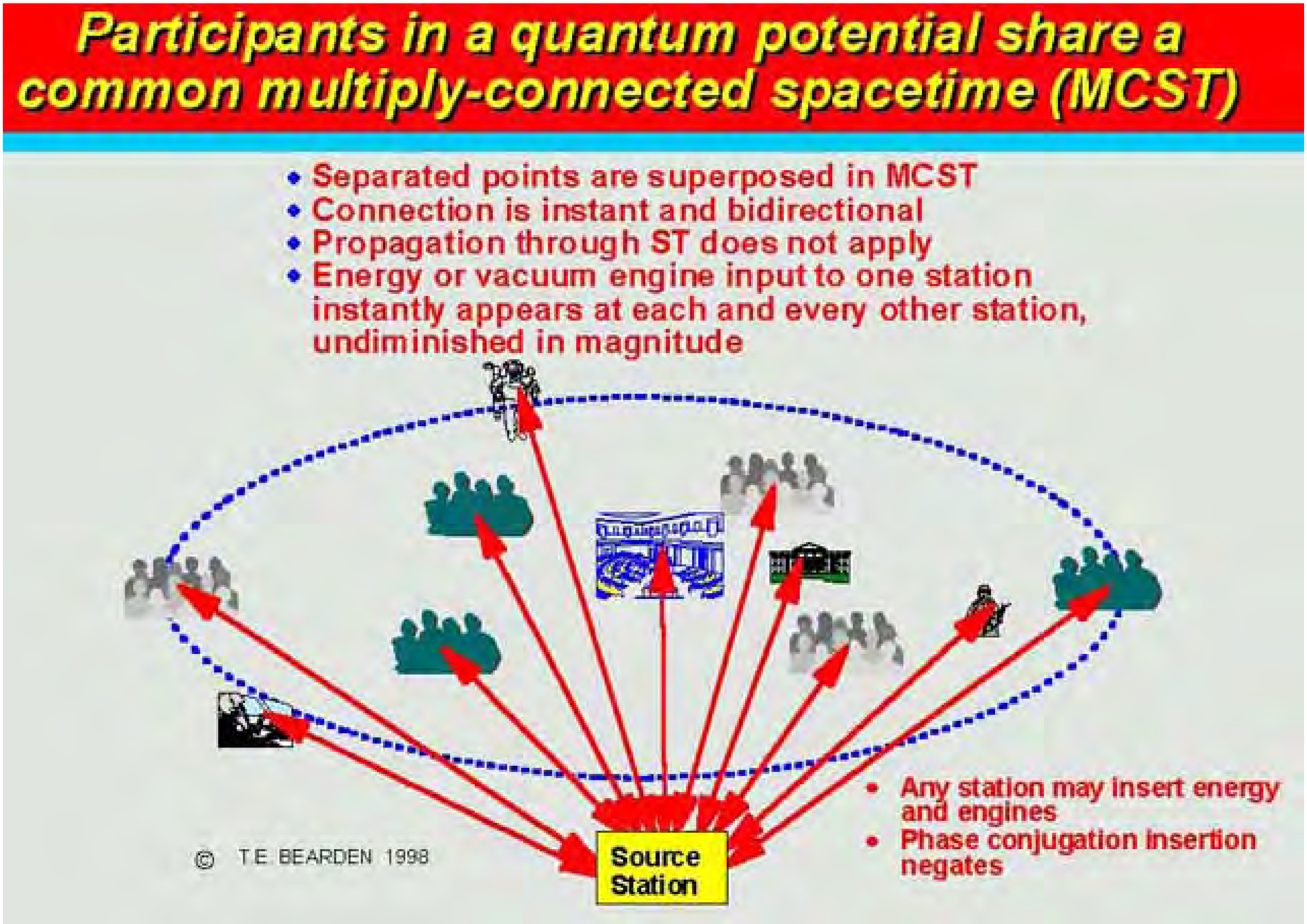

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

- Any station may insert energy and engines
- Phase conjugation insertion negates

© T.E. BEARDEN 1998

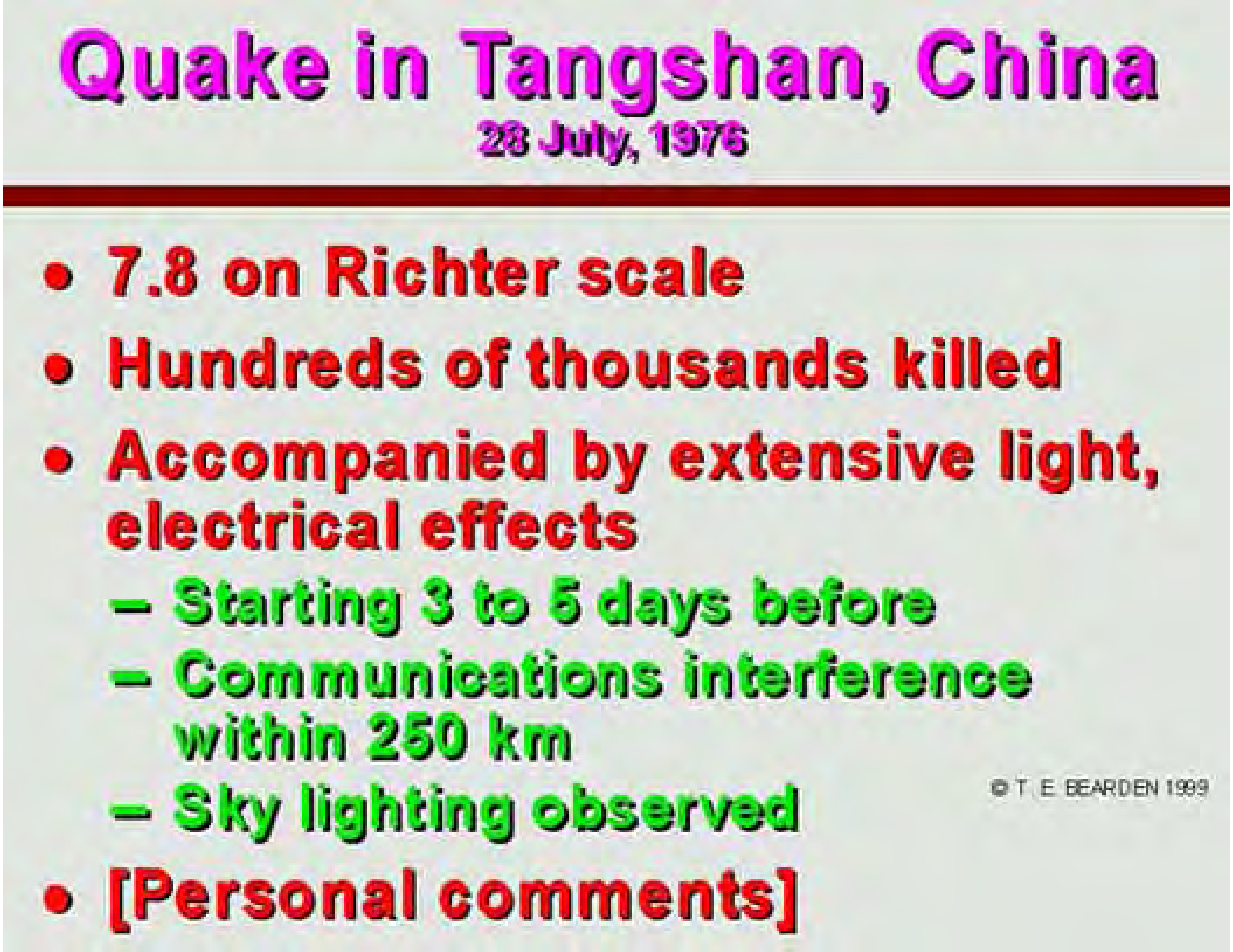

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

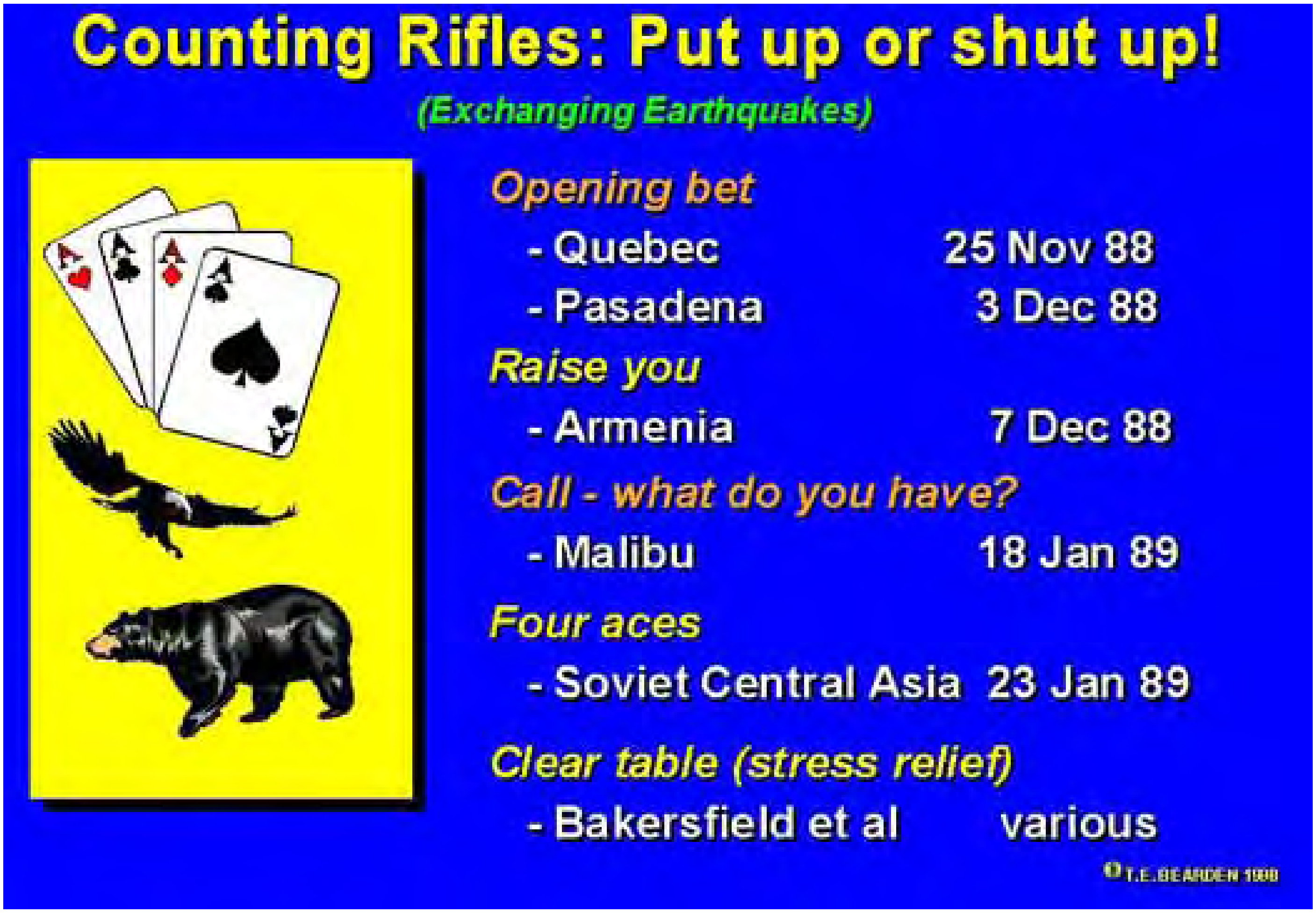

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet
- Quebec 25 Nov 88
- Pasadena 3 Dec 88

Raise you
- Armenia 7 Dec 88

Call - what do you have?
- Malibu 18 Jan 89

Four aces
- Soviet Central Asia 23 Jan 89

Clear table (stress relief)
- Bakersfield et al various

©T.E.BEARDEN 1990

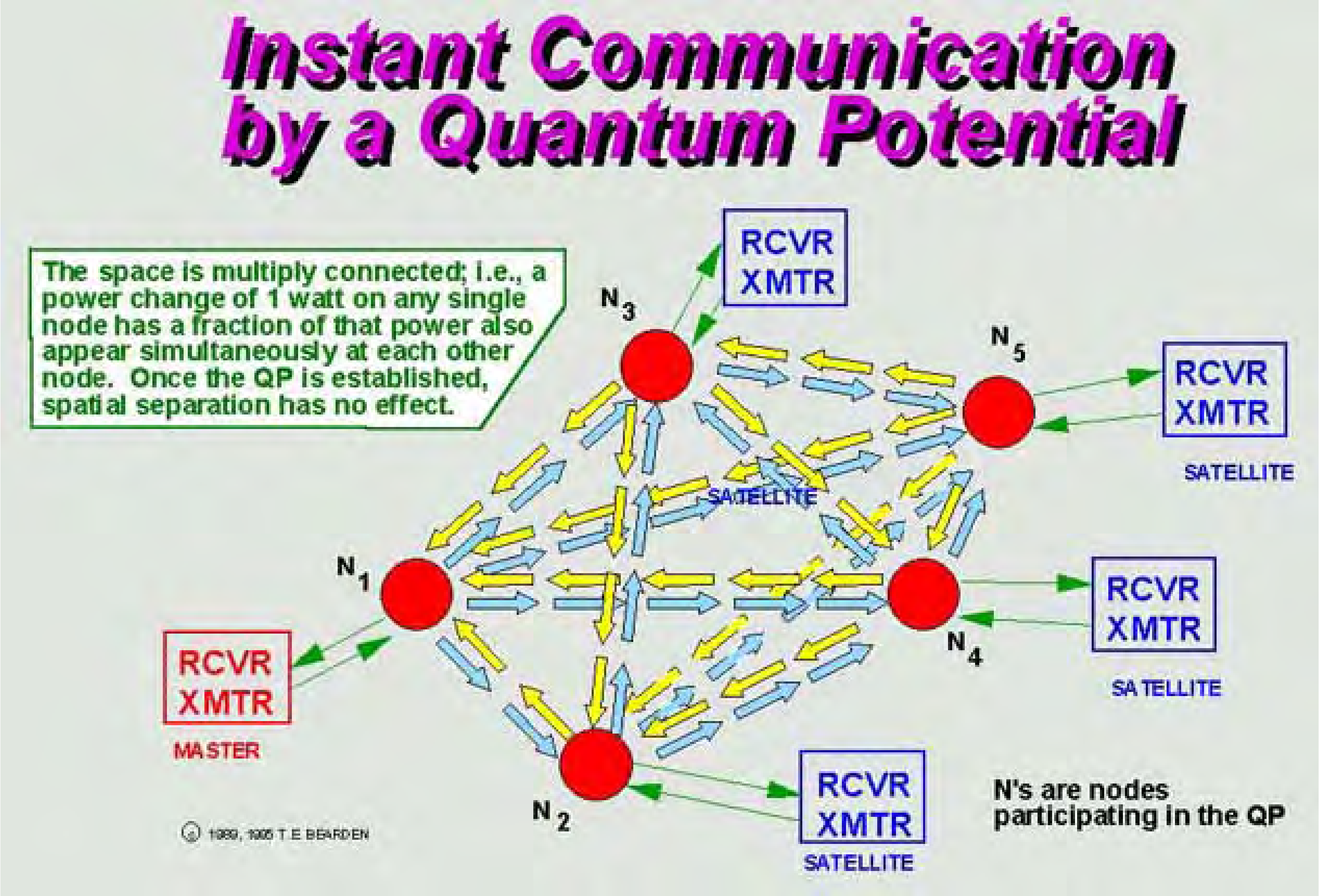

Instant Communication by a Quantum Potential

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N_3
RCVR XMTR
SATELLITE
N_5
RCVR XMTR
SATELLITE
N_1
RCVR XMTR
MASTER
N_4
RCVR XMTR
SATELLITE
N_2
RCVR XMTR
SATELLITE

N's are nodes participating in the QP

© 1989, 1995 T.E BEARDEN

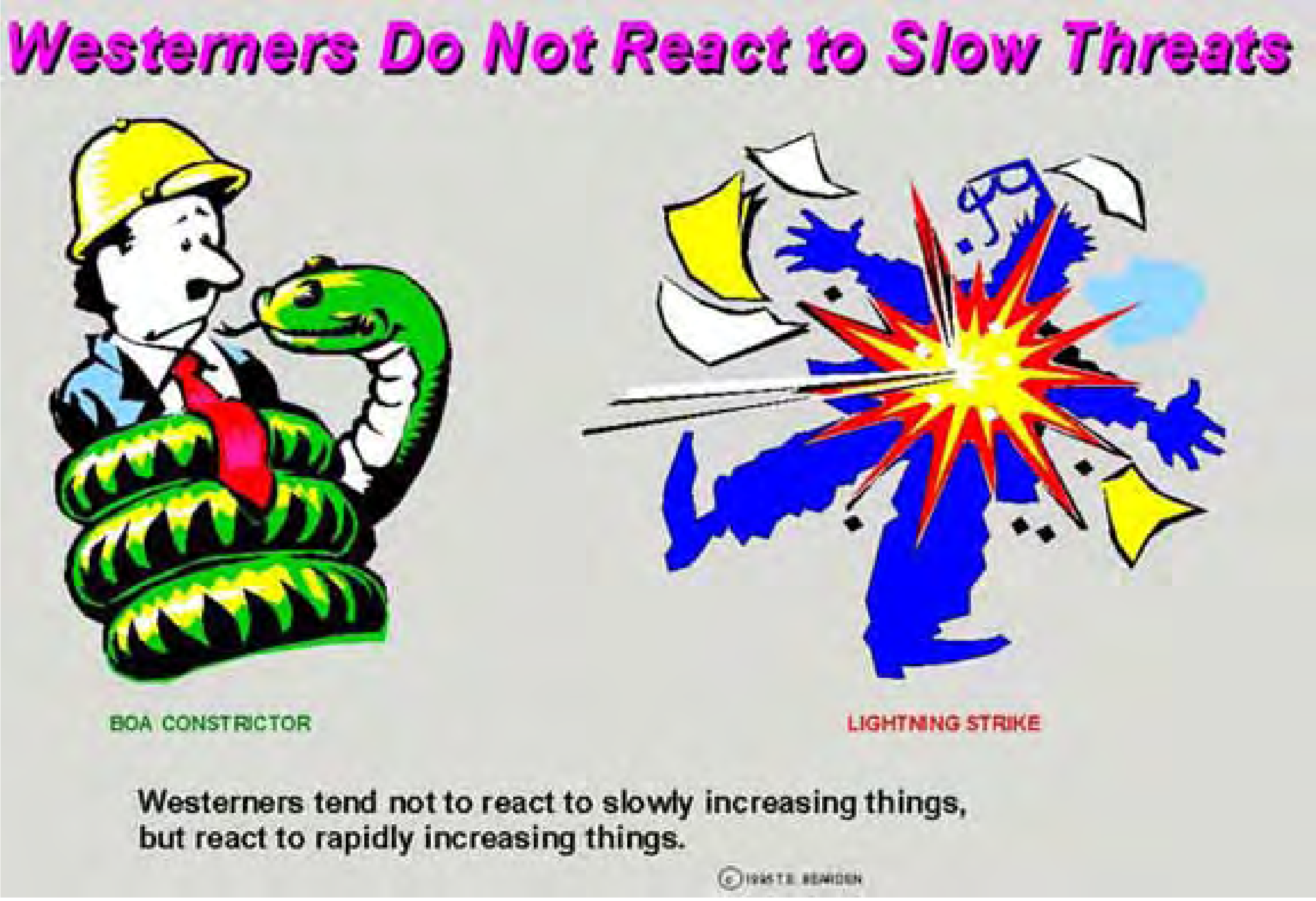

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR

LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things,
but react to rapidly increasing things.

© 1995 T.E. BEARDEN

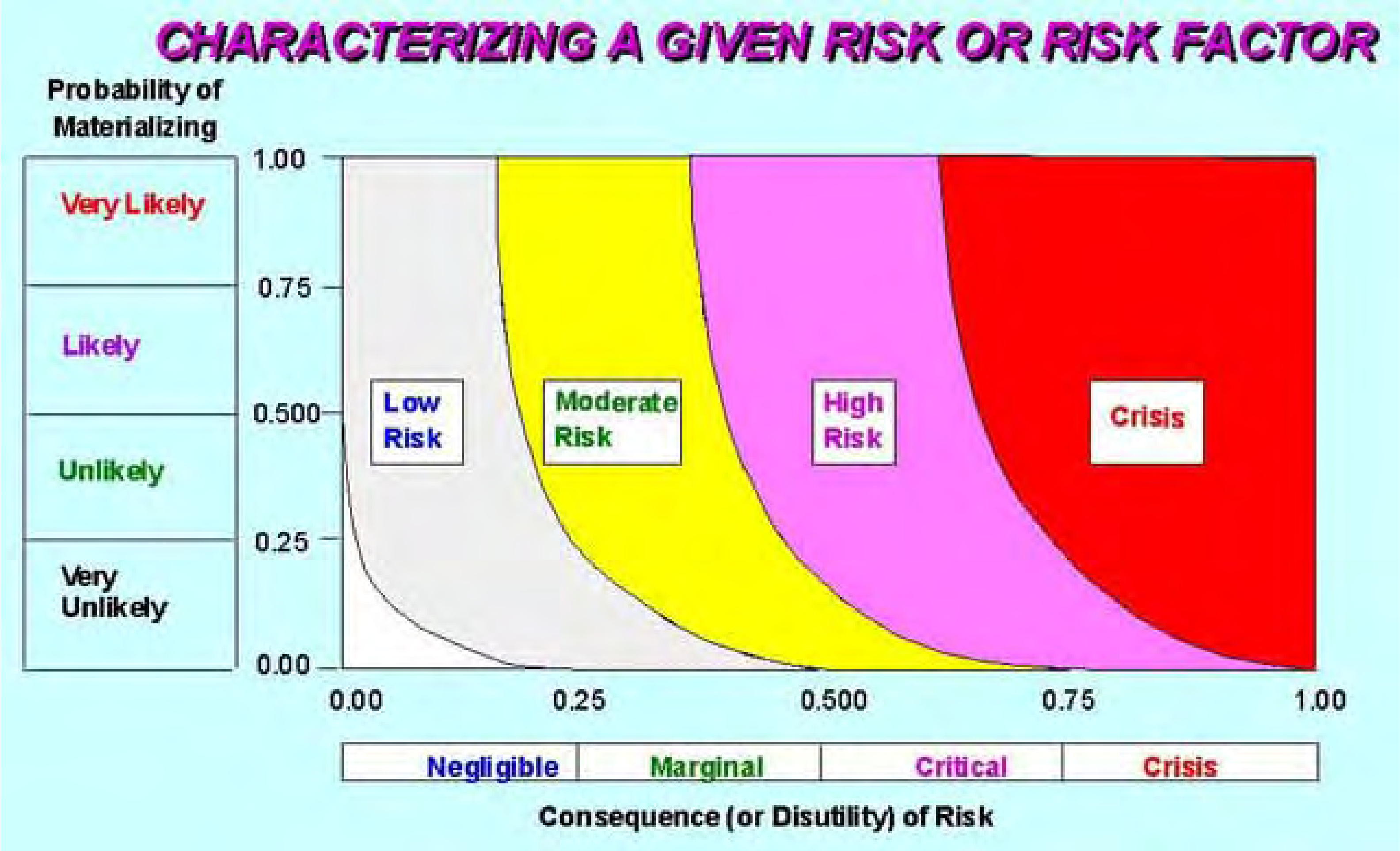

CHARACTERIZING A GIVEN RISK OR RISK FACTOR

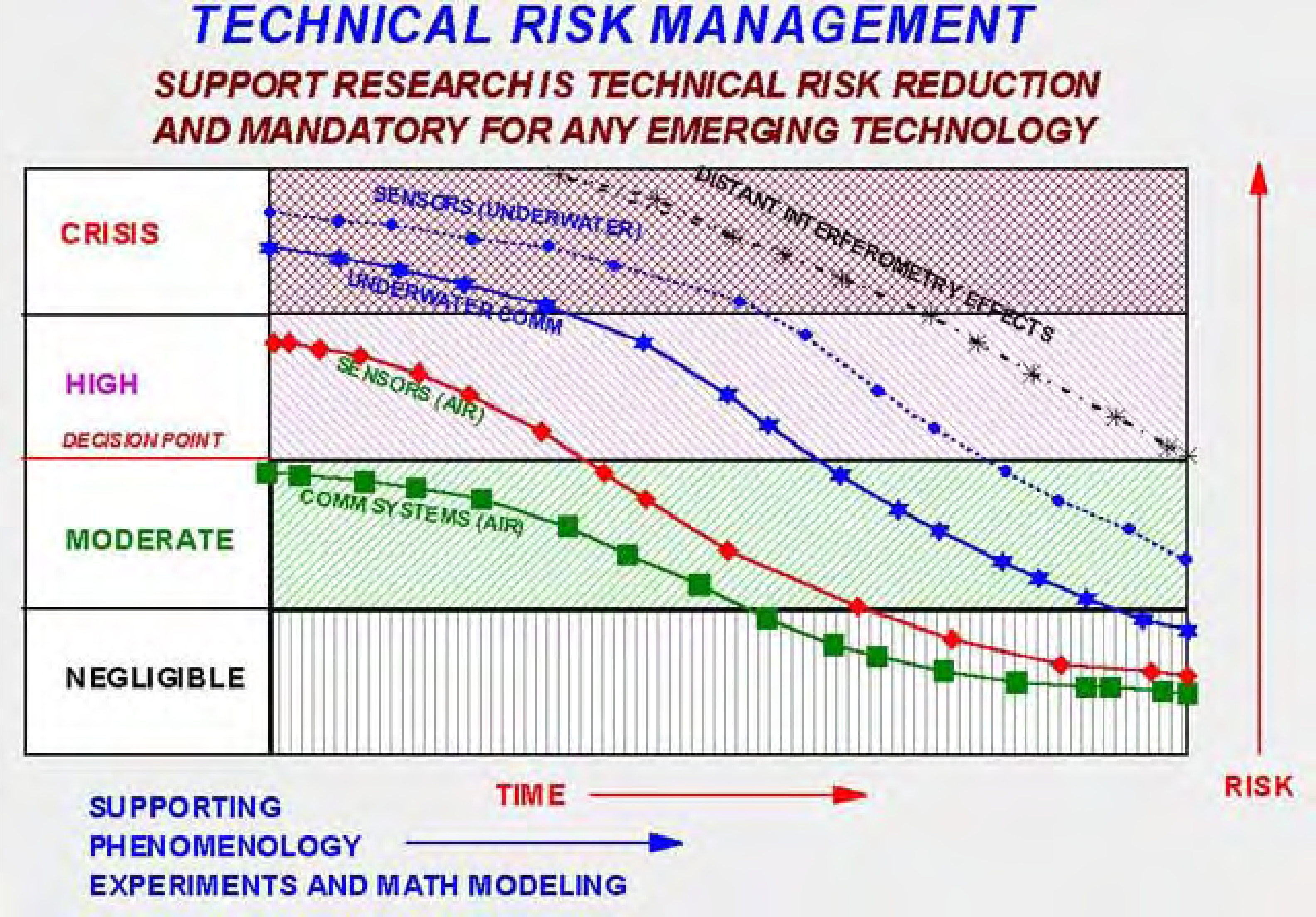

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

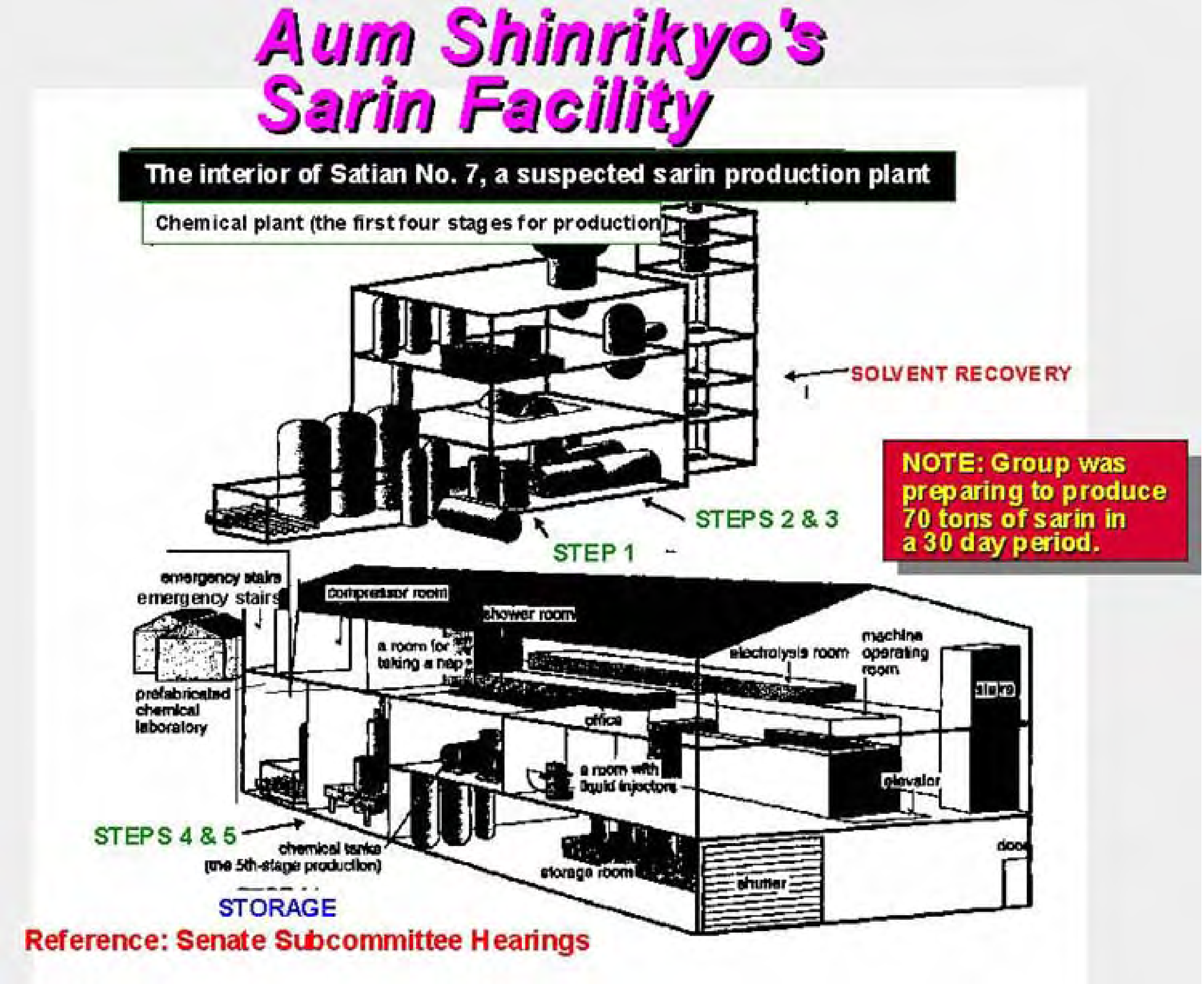

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

Reference: Senate Subcommittee Hearings

Chemical plant (the first four stages for production)
SOLVENT RECOVERY
STEPS 2 & 3
STEP 1
emergency stairs
emergency stairs
compressor room
shower room
a room for
taking a nap
machine operating room
electrolysis room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
STEPS 4 & 5
chemical tanks (the 5th-stage production)
storage room
shutter
doo
STORAGE

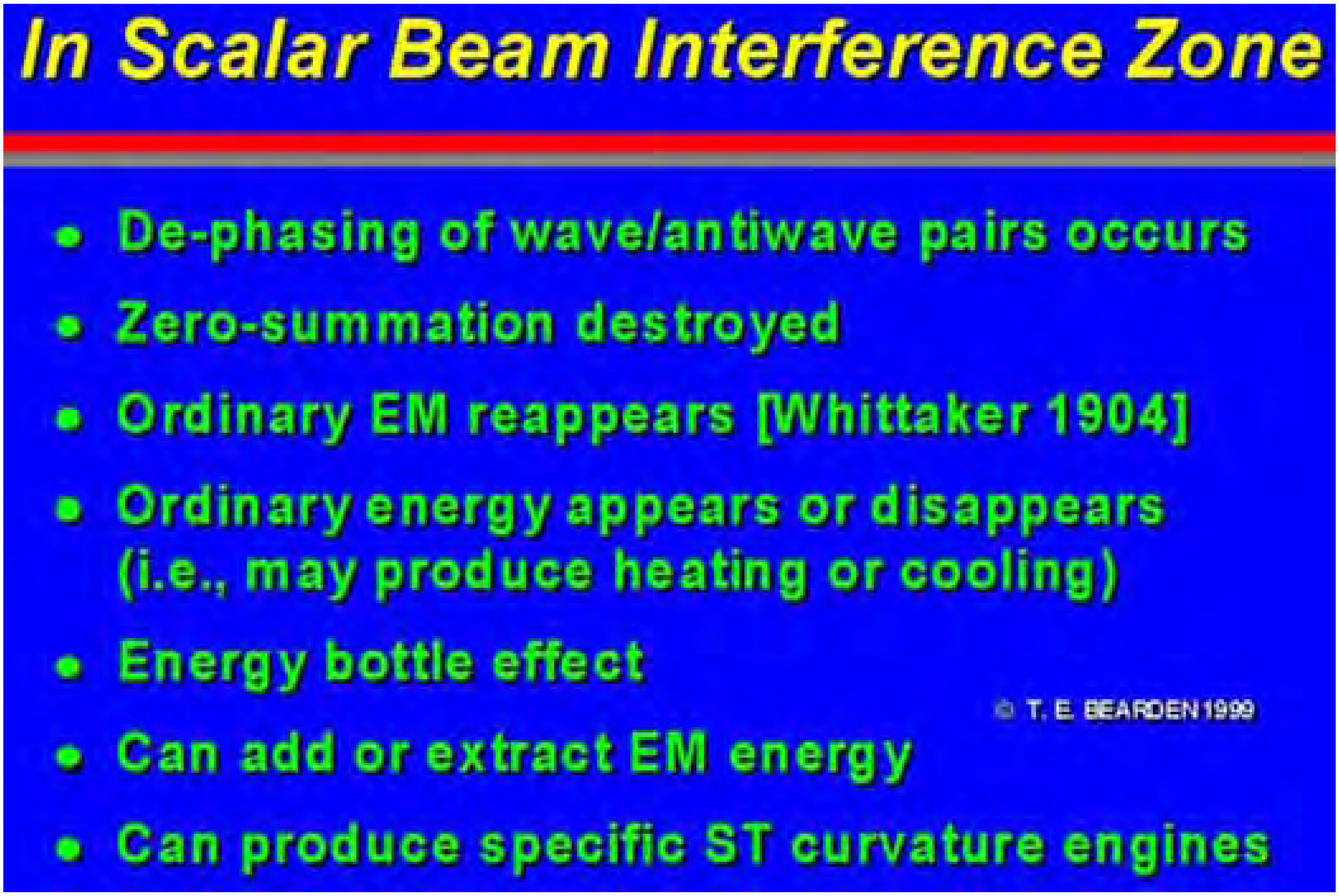

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

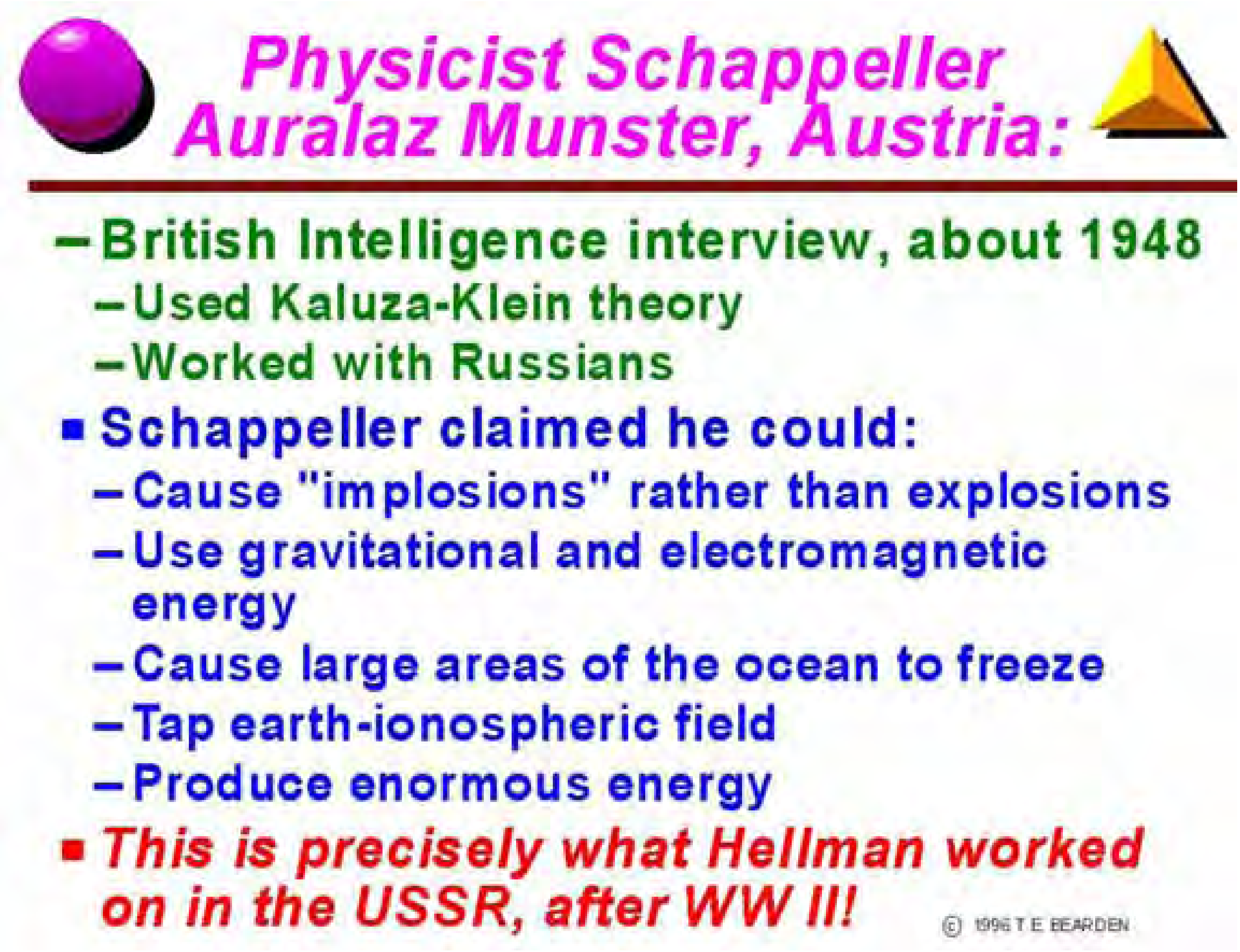

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- Schappeller claimed he could:
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T E BEARDEN

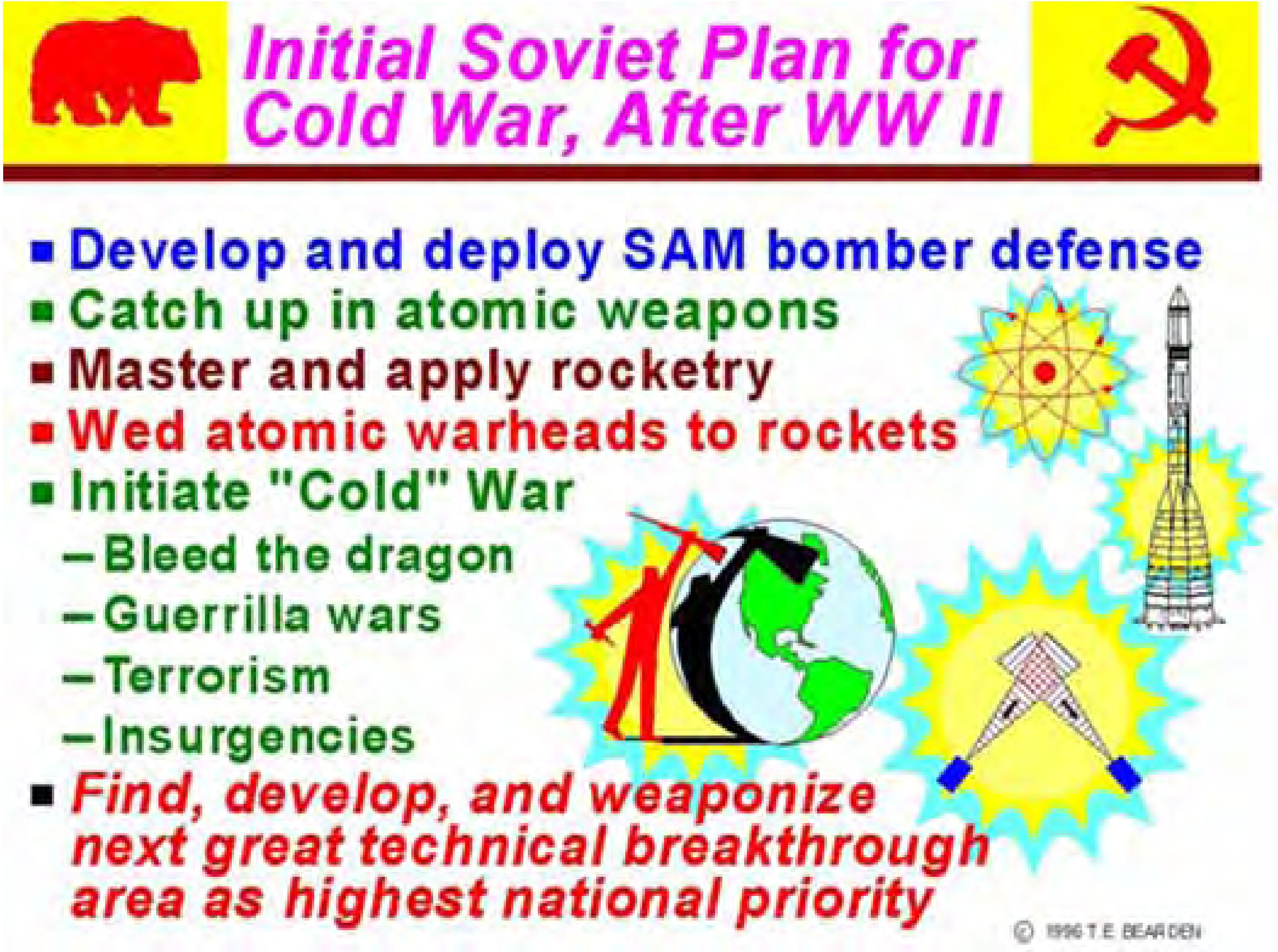

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T.E. BEARDEN

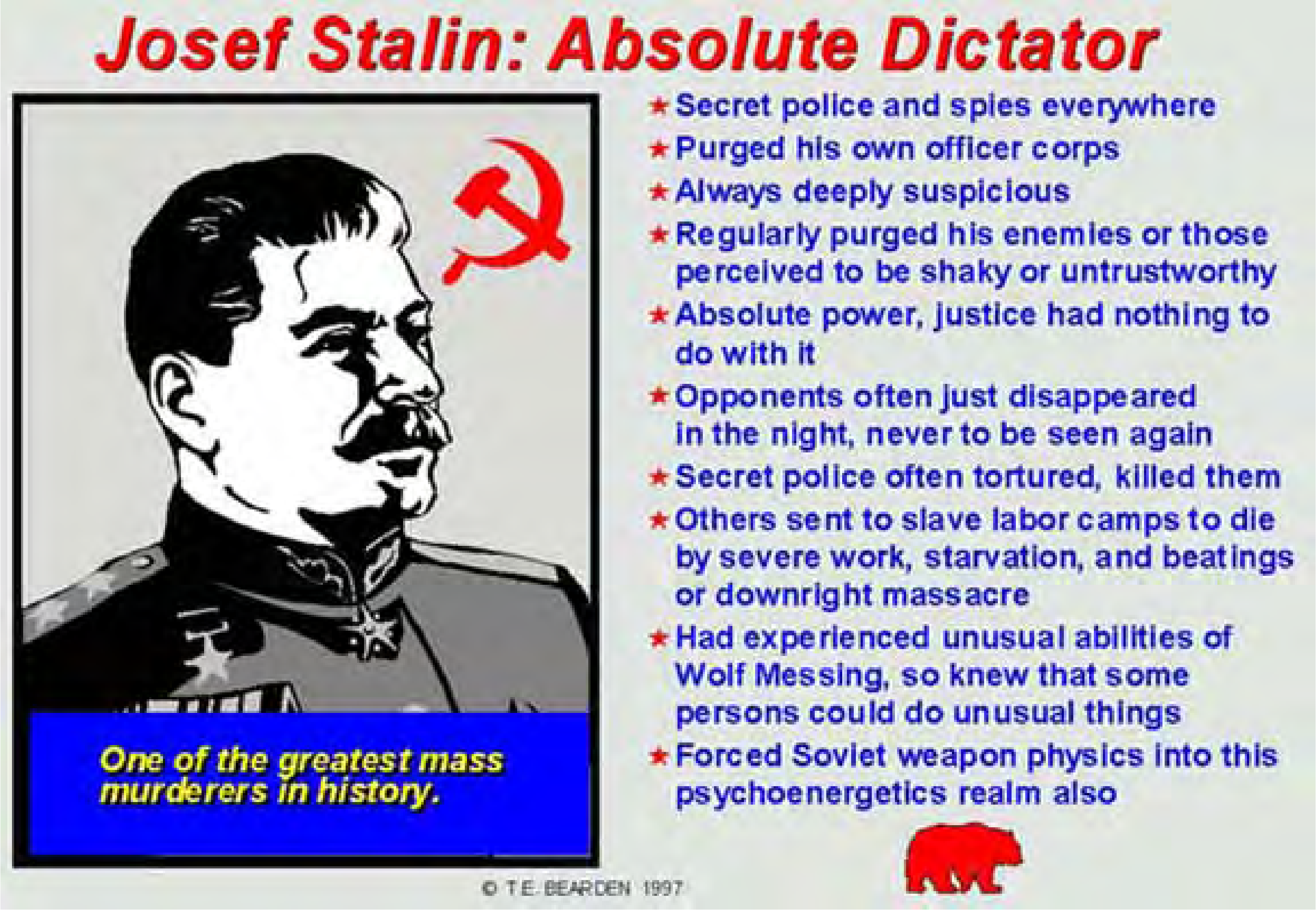

Josef Stalin: Absolute Dictator

- Secret police and spies everywhere
- Purged his own officer corps
- Always deeply suspicious
- Regularly purged his enemies or those perceived to be shaky or untrustworthy
- Absolute power, justice had nothing to do with it
- Opponents often just disappeared in the night, never to be seen again
- Secret police often tortured, killed them
- Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- Forced Soviet weapon physics into this psychoenergetics realm also

One of the greatest mass murderers in history.

© T.E. BEARDEN 1997

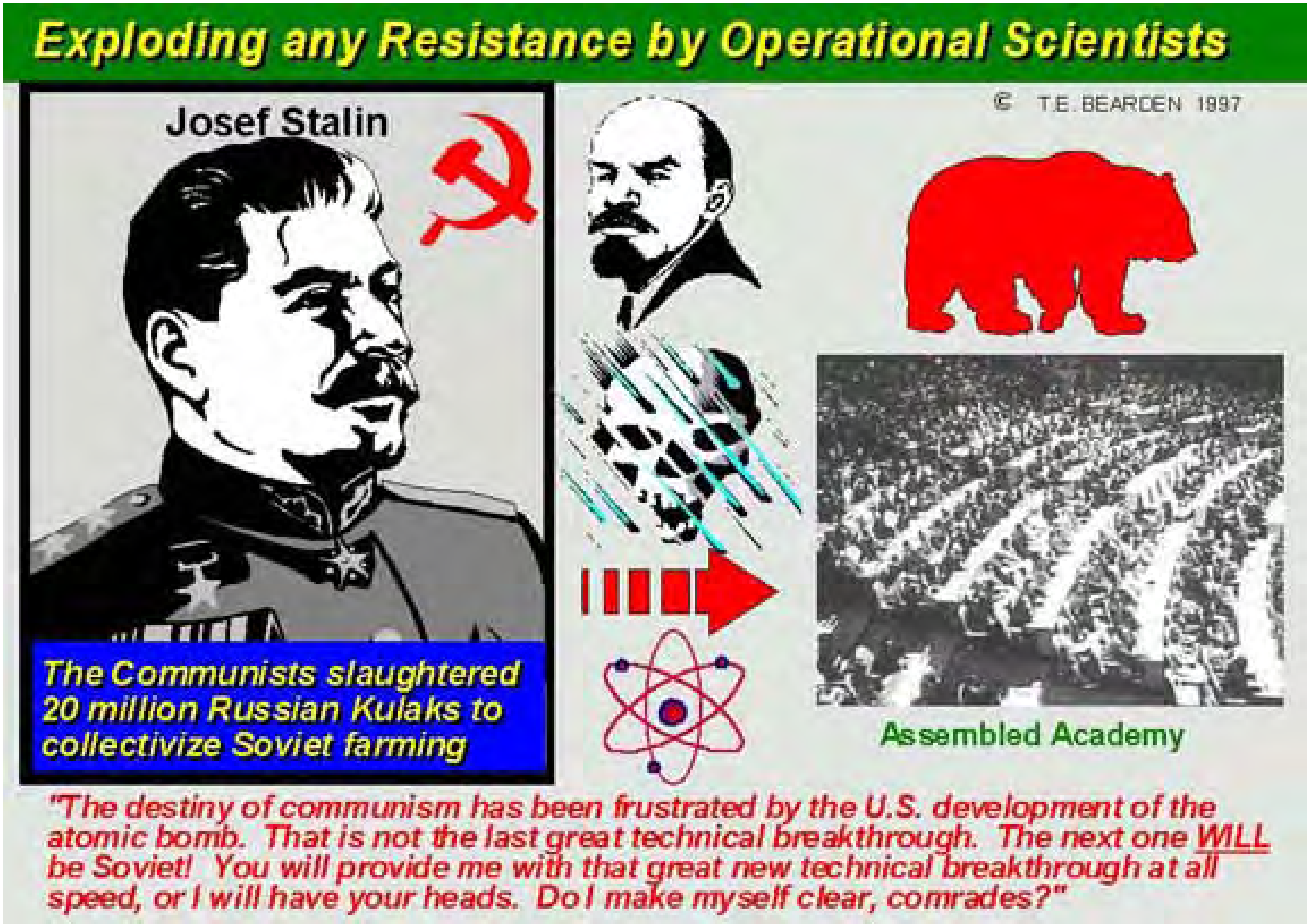

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

Josef Stalin

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

Assembled Academy

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

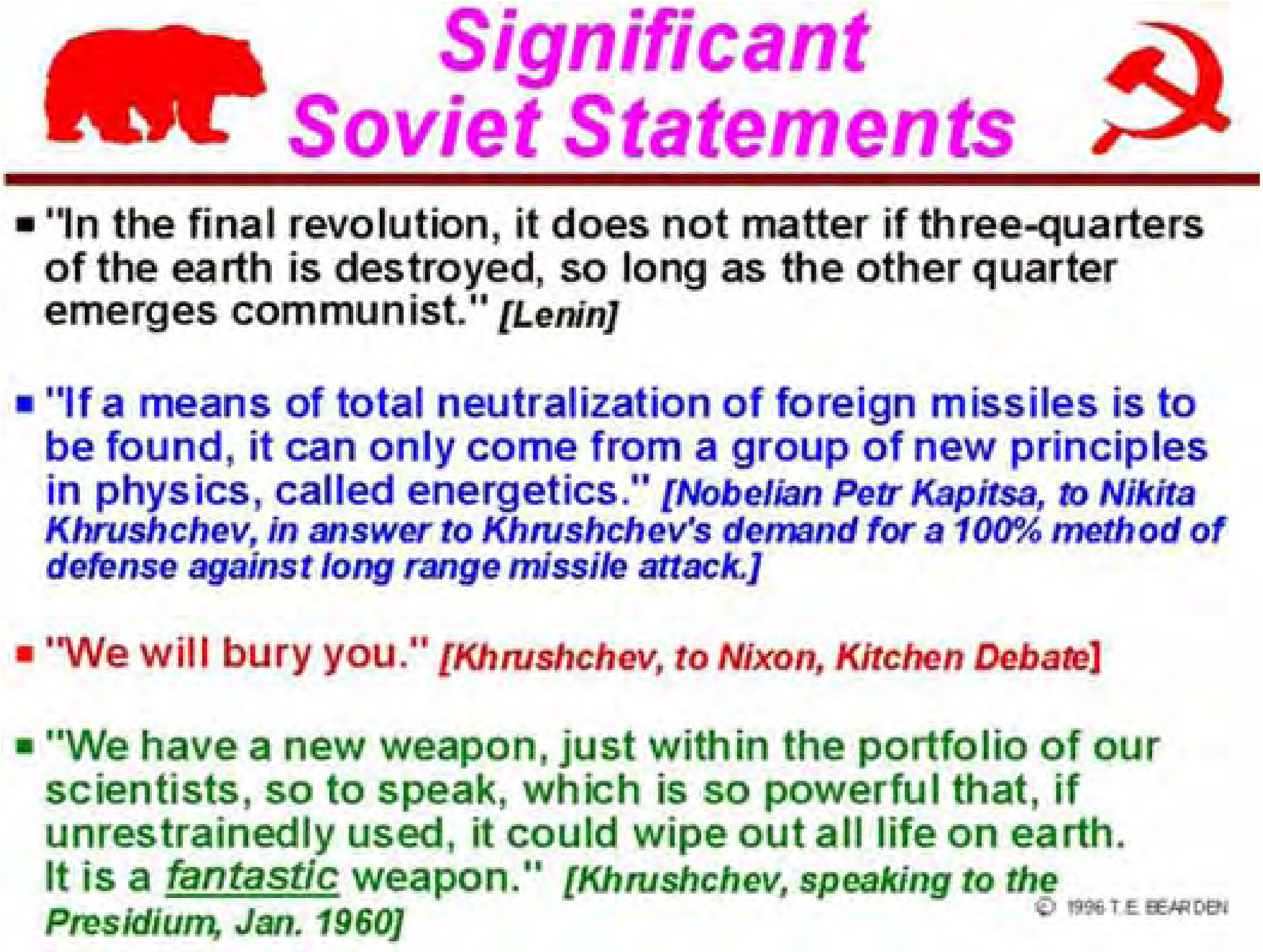

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*

- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." *[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]*

- "We will bury you." *[Khrushchev, to Nixon, Kitchen Debate*]

- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a *fantastic* weapon." *[Khrushchev, speaking to the Presidium, Jan. 1960]*

© 1996 T.E. BEARDEN

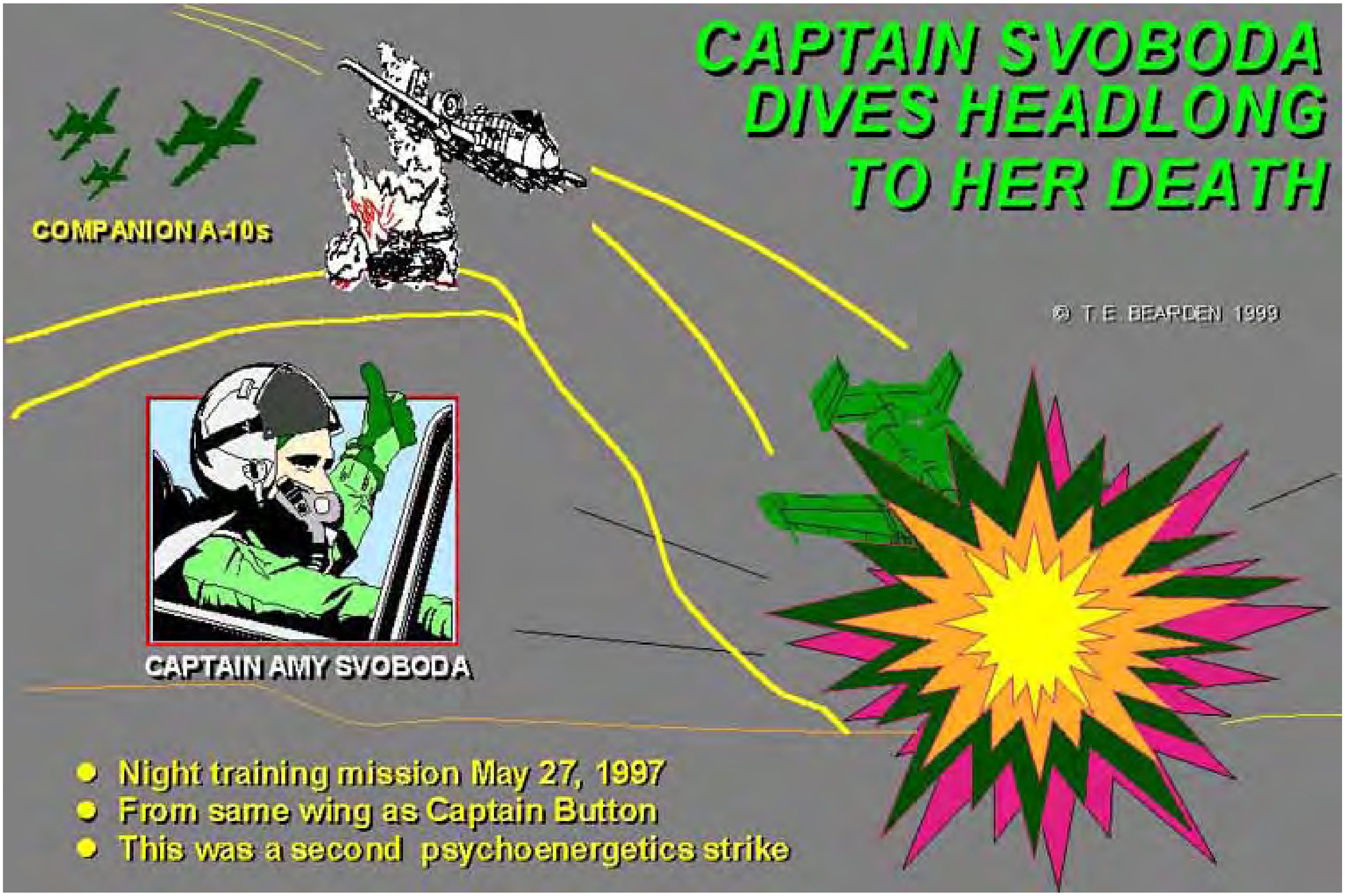

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

COMPANION A-10s
CAPTAIN AMY SVOBODA

© T.E. BEARDEN 1999

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

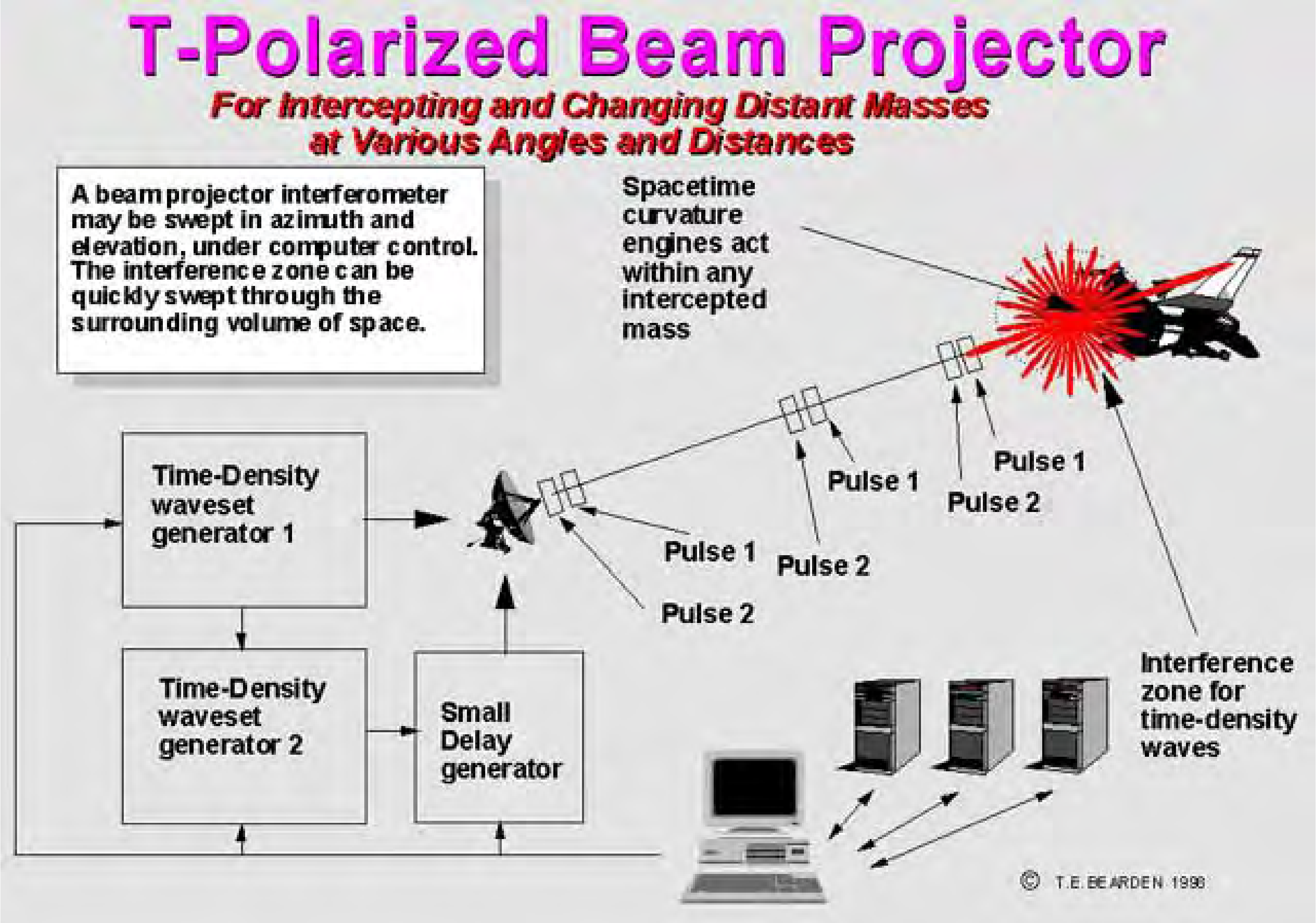

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass
Time-Density waveset generator 1
Time-Density waveset generator 2
Small Delay generator
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Interference zone for time-density waves
© T.E. BEARDEN 1996

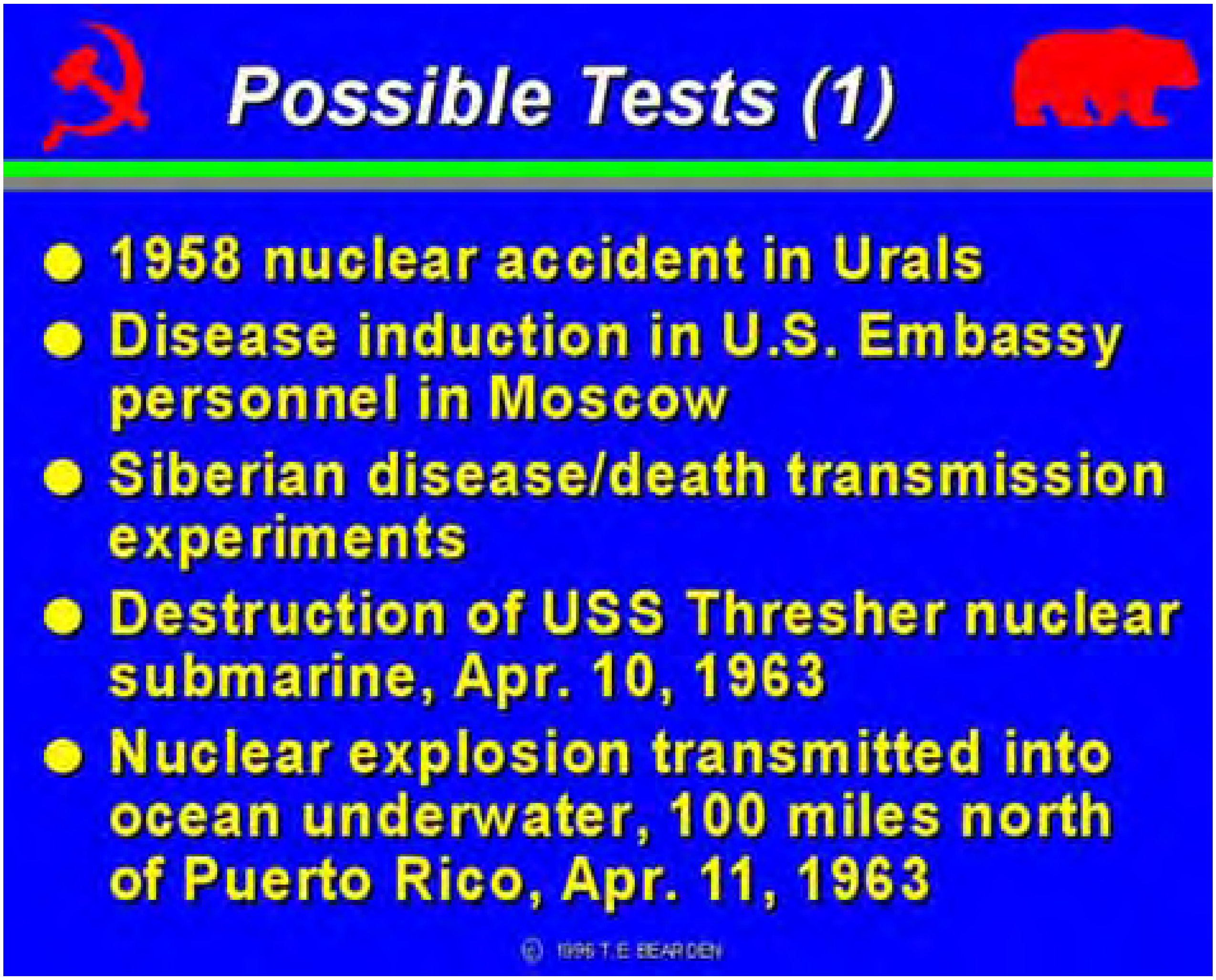

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T. E. BEARDEN

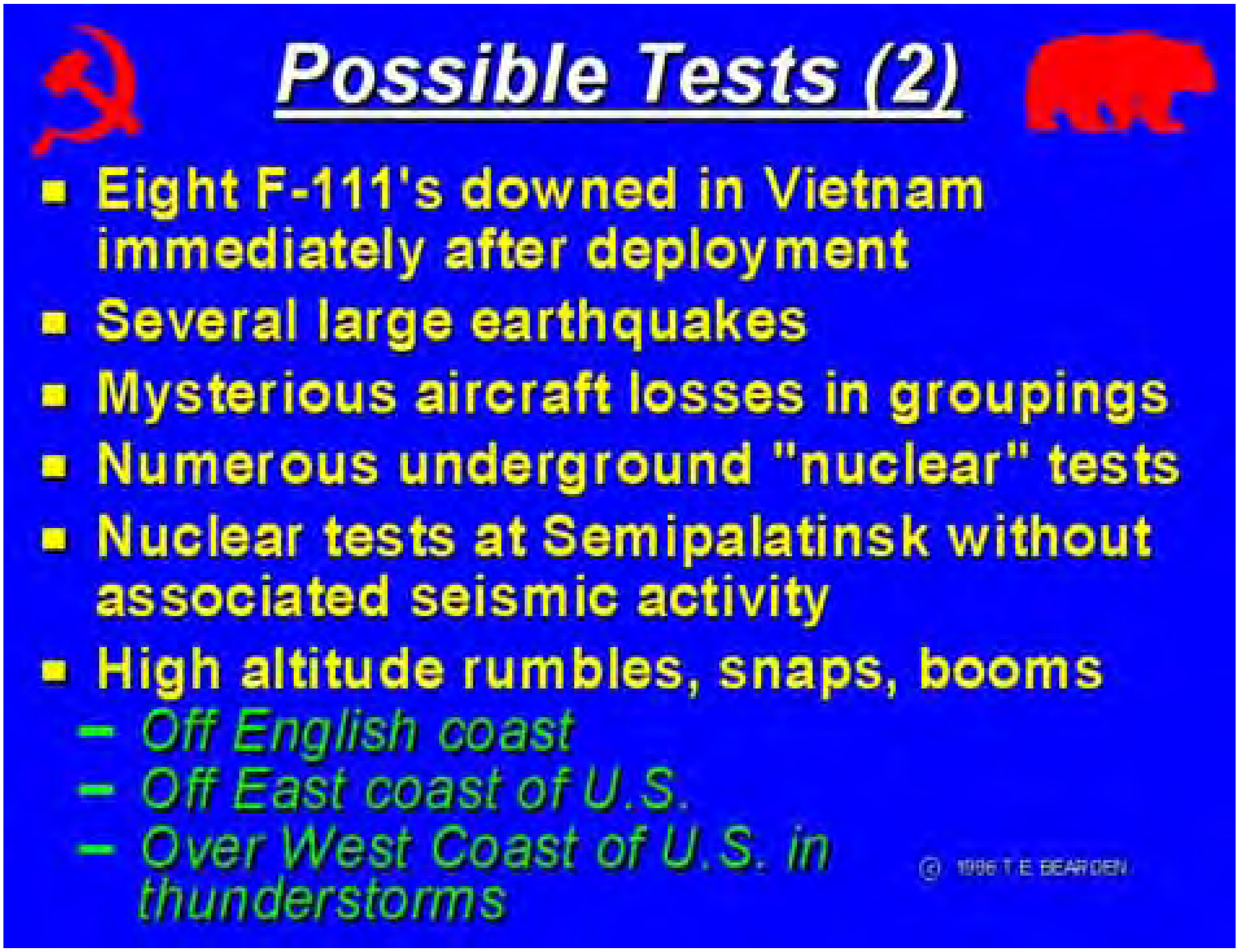

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T. E. BEARDEN

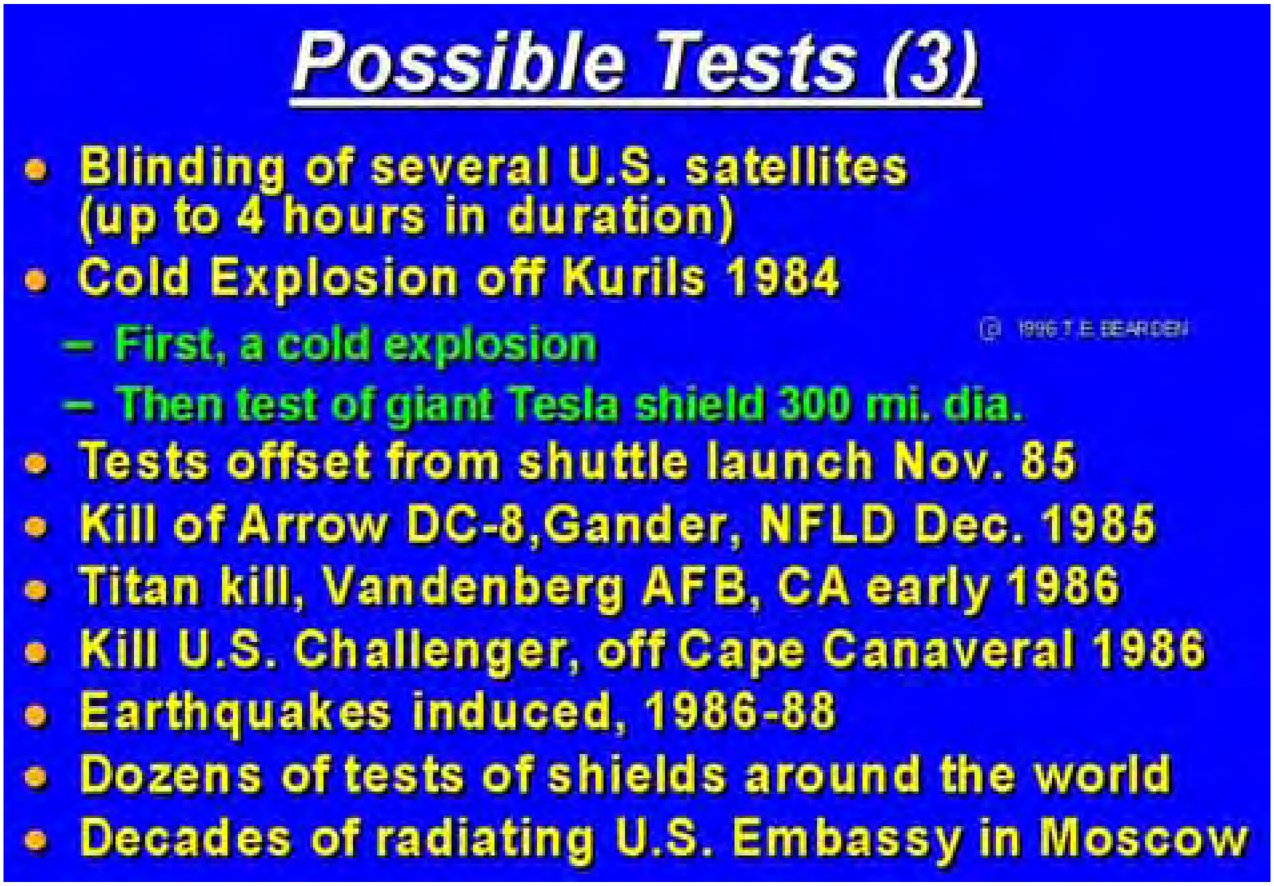

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T.E. BEARDEN

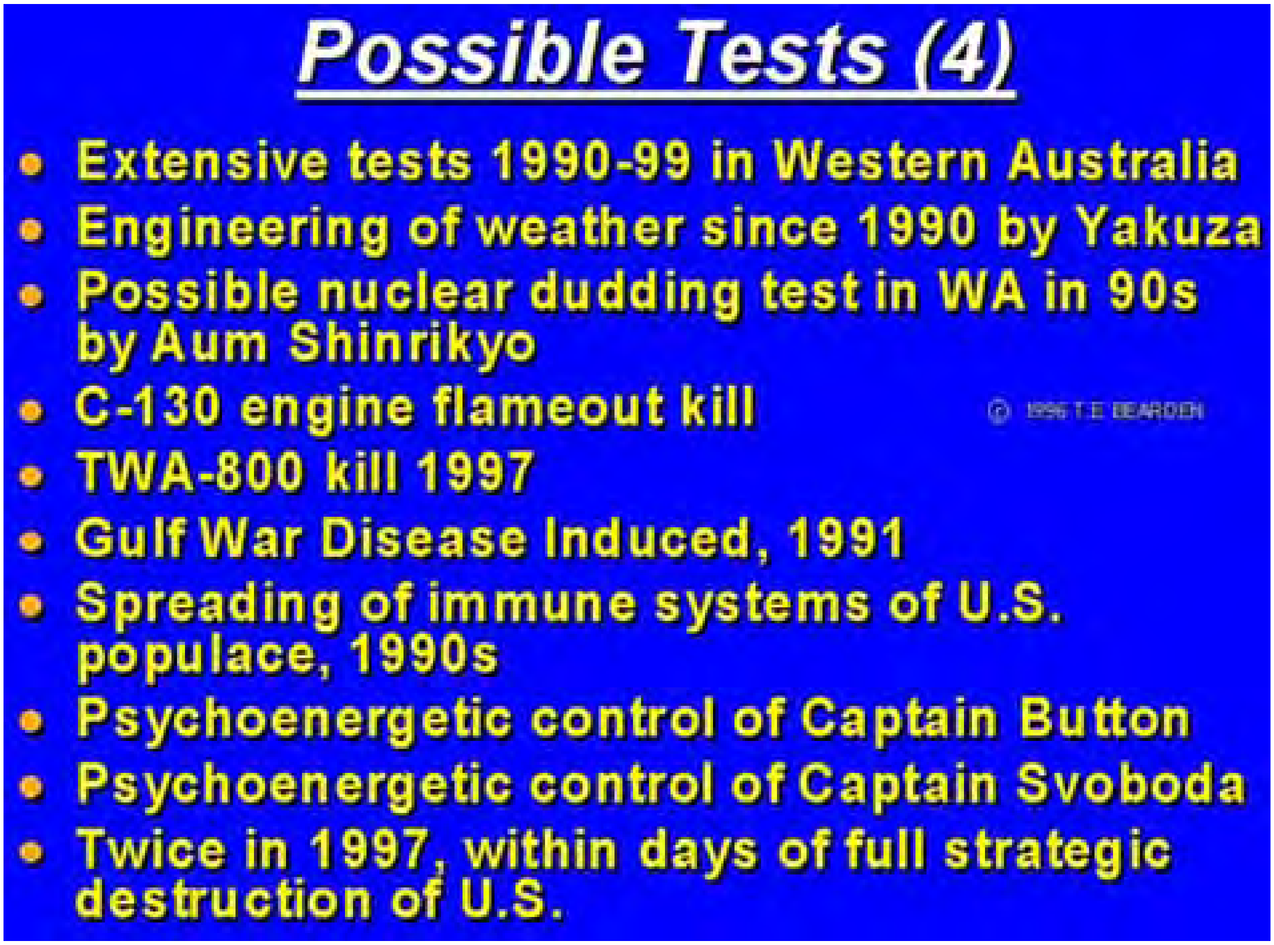

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E. BEARDEN

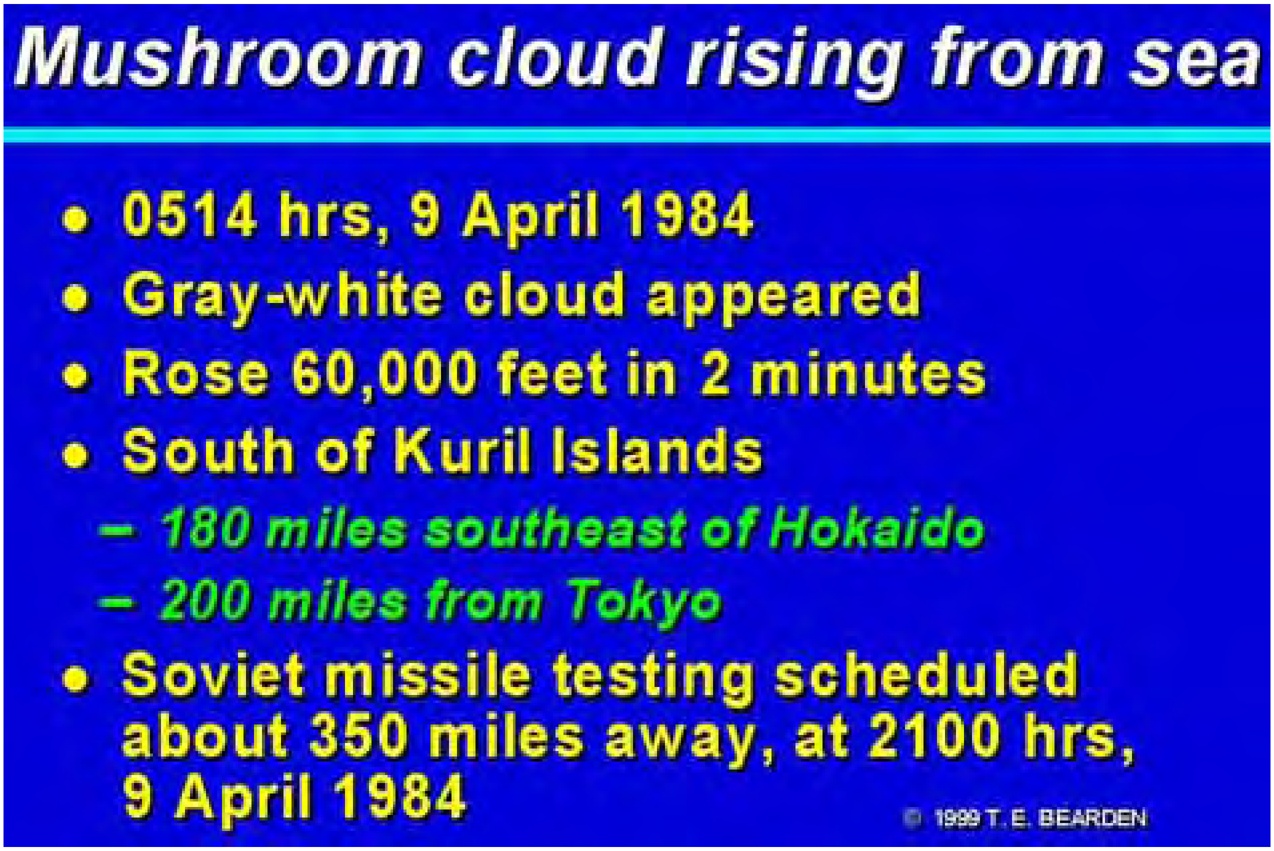

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - *180 miles southeast of Hokaido*
  - *200 miles from Tokyo*
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

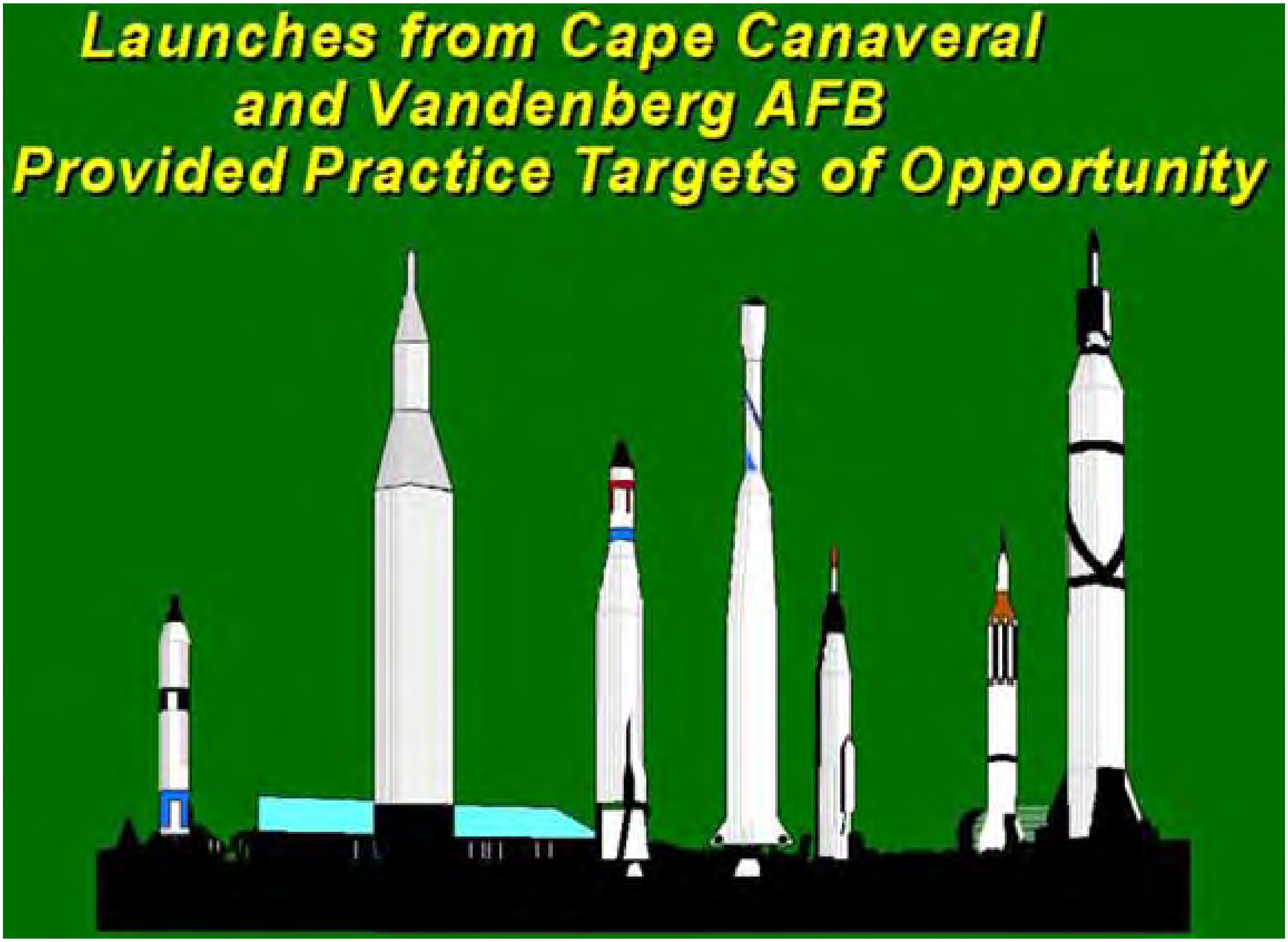

Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

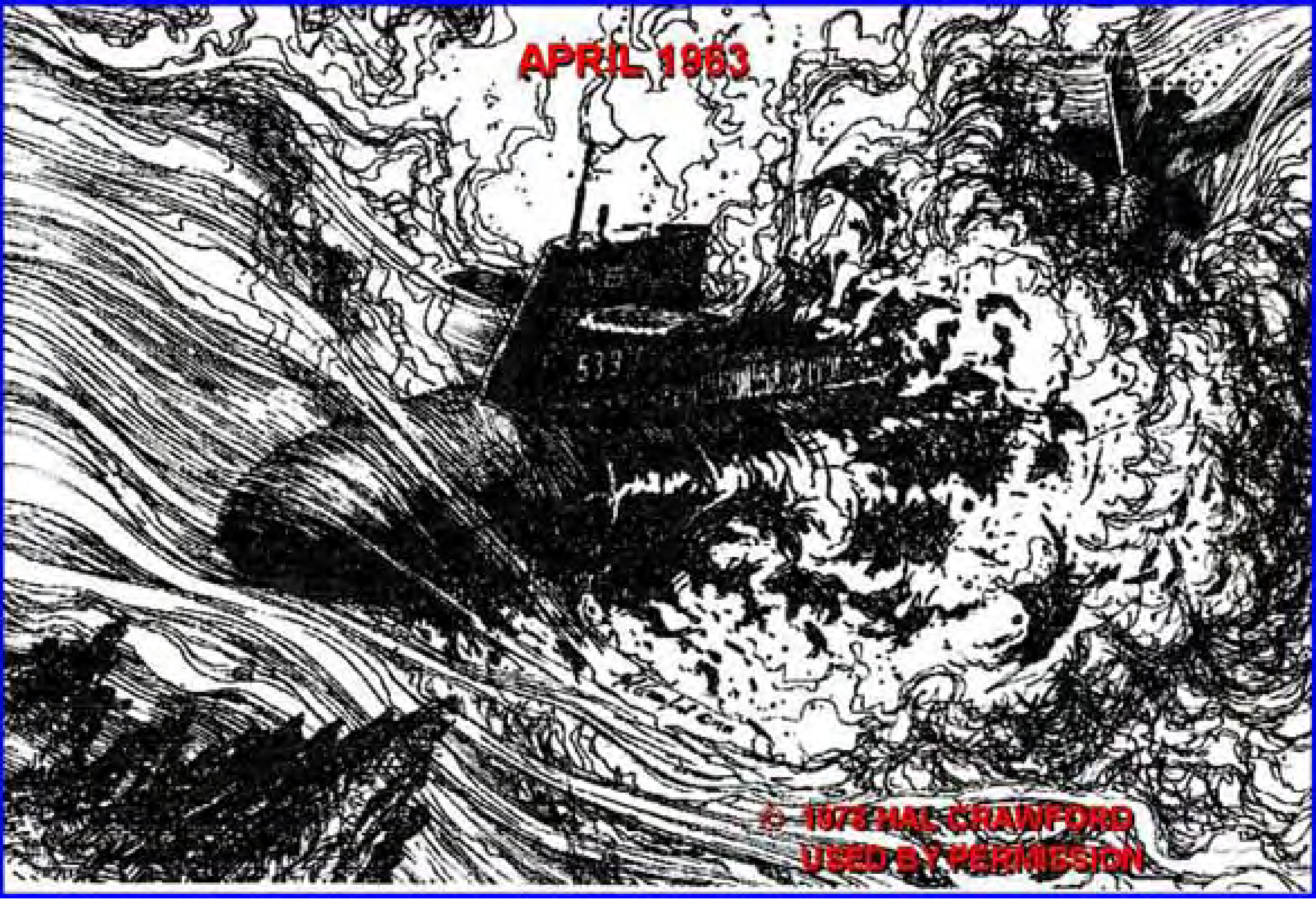
APRIL 1963
533
© 1978 HAL CRAWFORD
USED BY PERMISSION

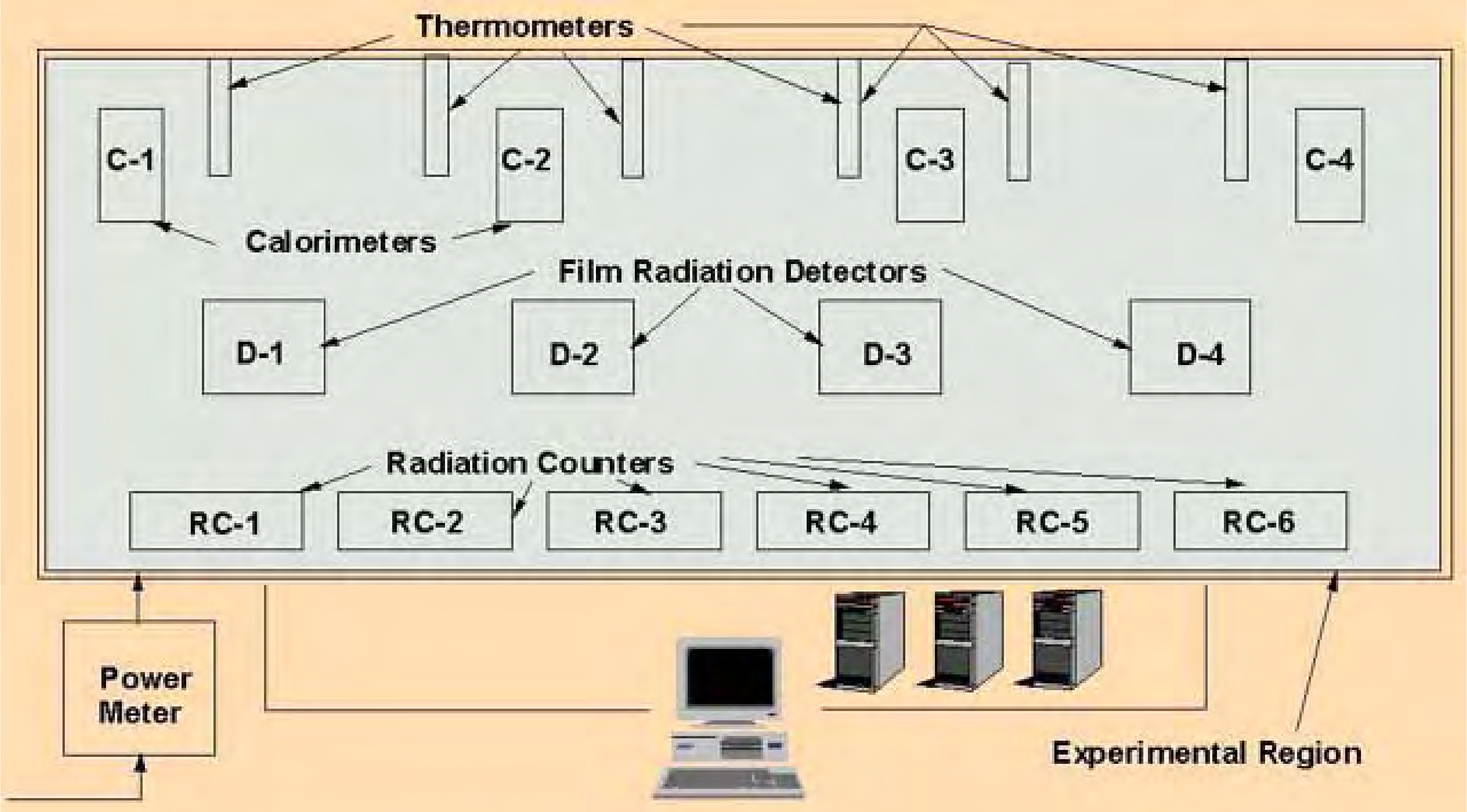
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

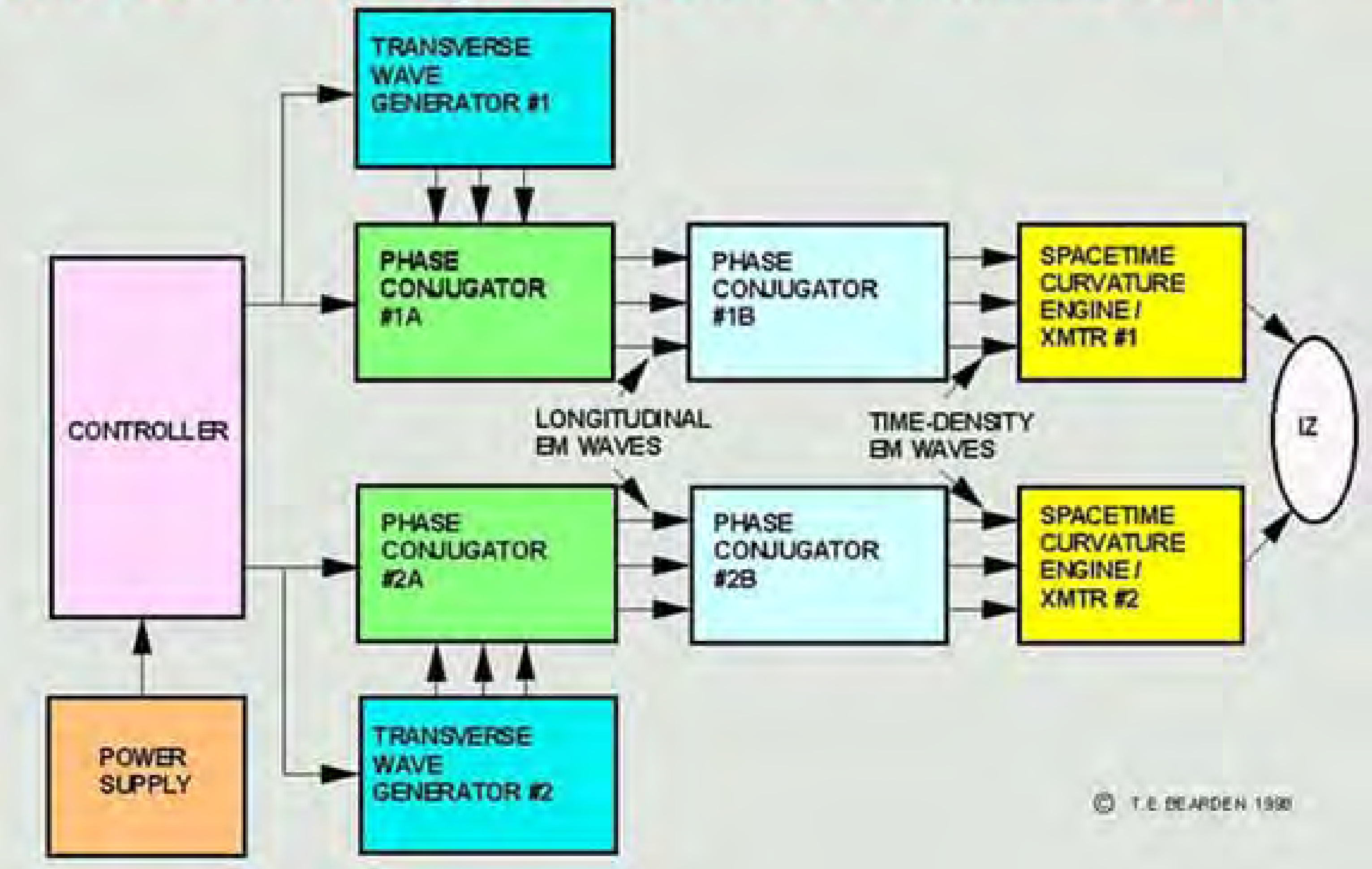
TRANSVERSE WAVE GENERATOR #1
CONTROLLER
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2

© T.E. BEARDEN 1990

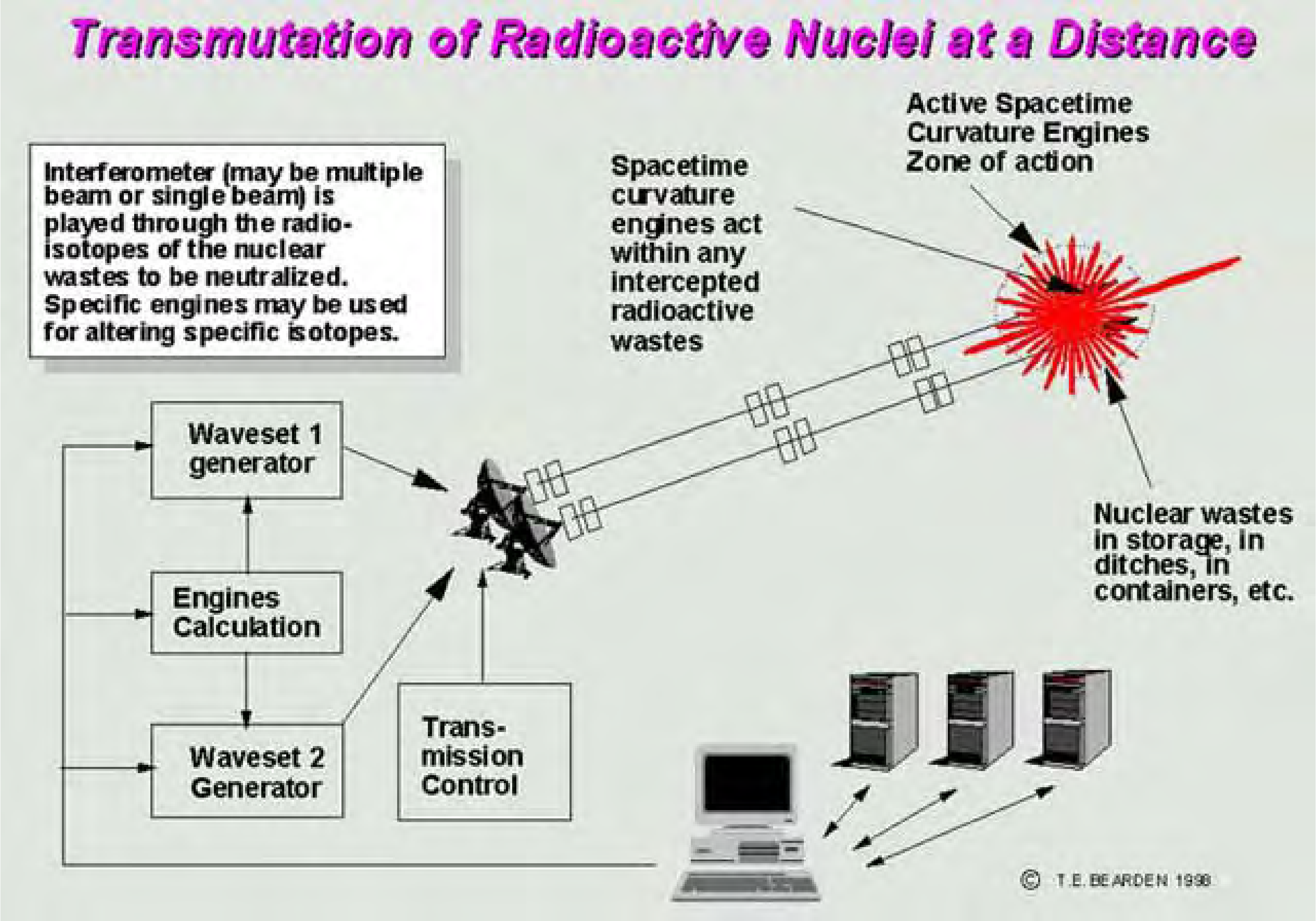
Transmutation of Radioactive Nuclei at a Distance
Interferometer (may be multiple beam or single beam) is played through the radio-isotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.
Spacetime curvature engines act within any intercepted radioactive wastes
Active Spacetime Curvature Engines Zone of action
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Nuclear wastes in storage, in ditches, in containers, etc.
© T.E. BEARDEN 1998

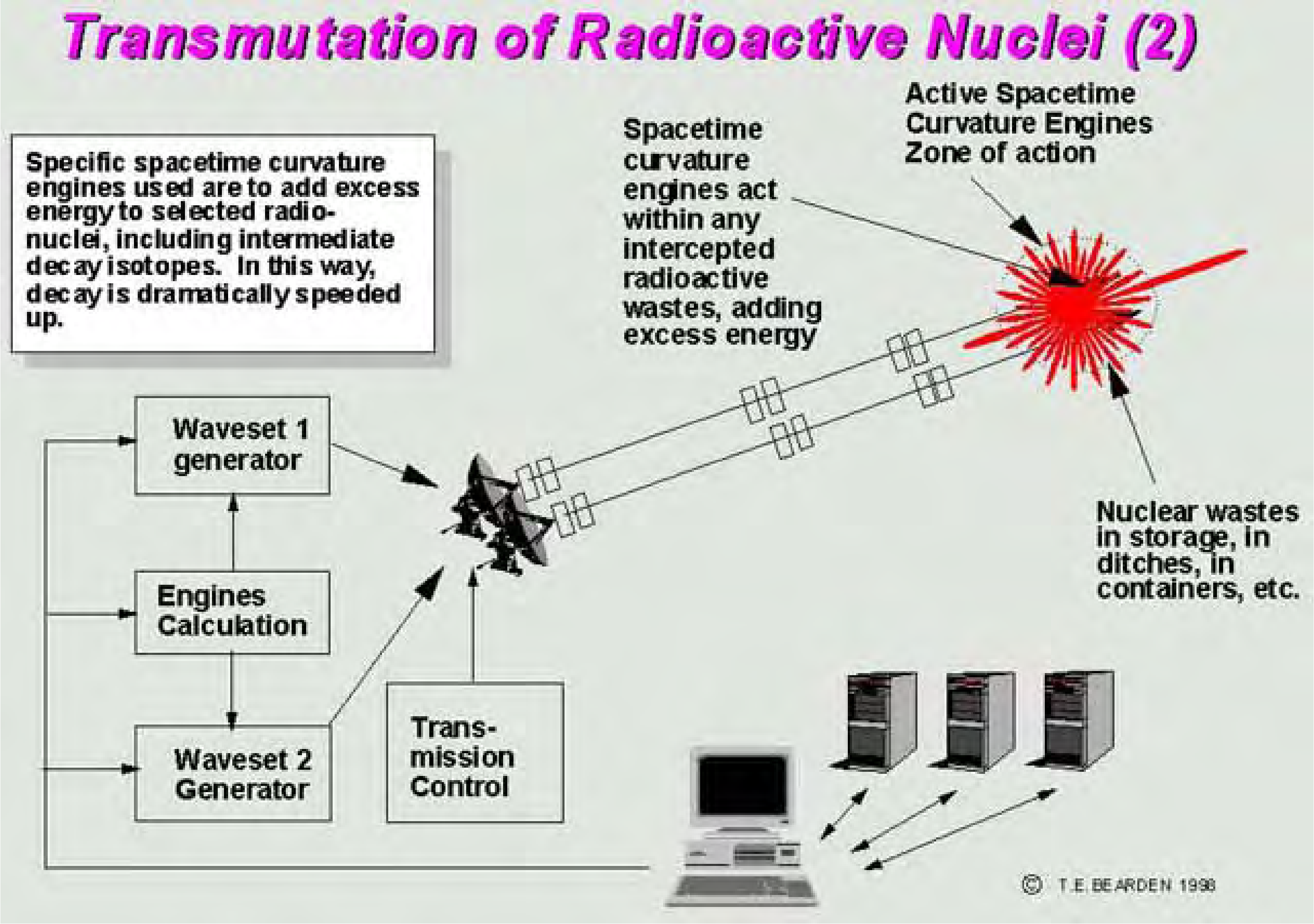

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radionuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Nuclear wastes in storage, in ditches, in containers, etc.
© T.E. BEARDEN 1998

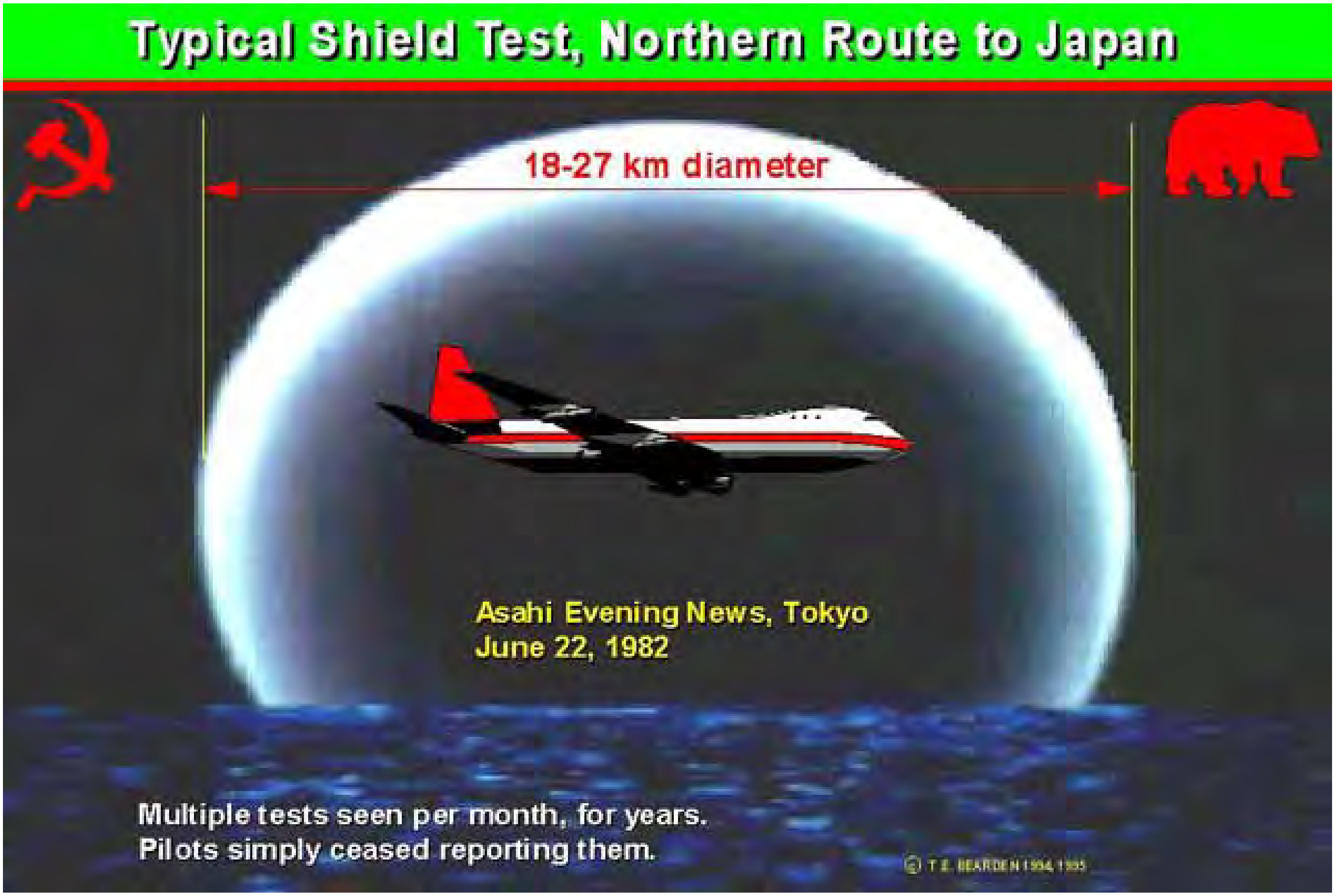

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T. E. BEARDEN 1994, 1995

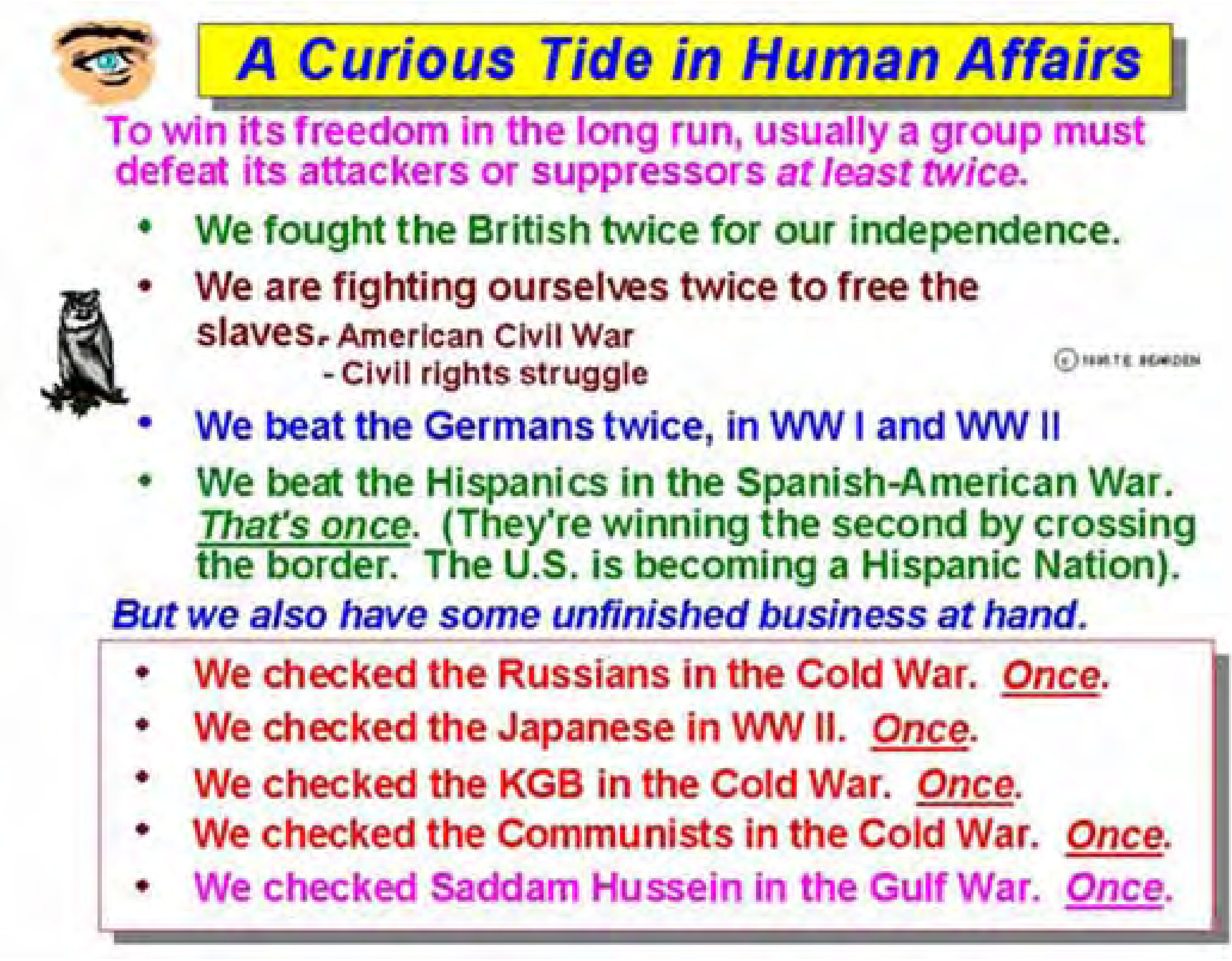

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

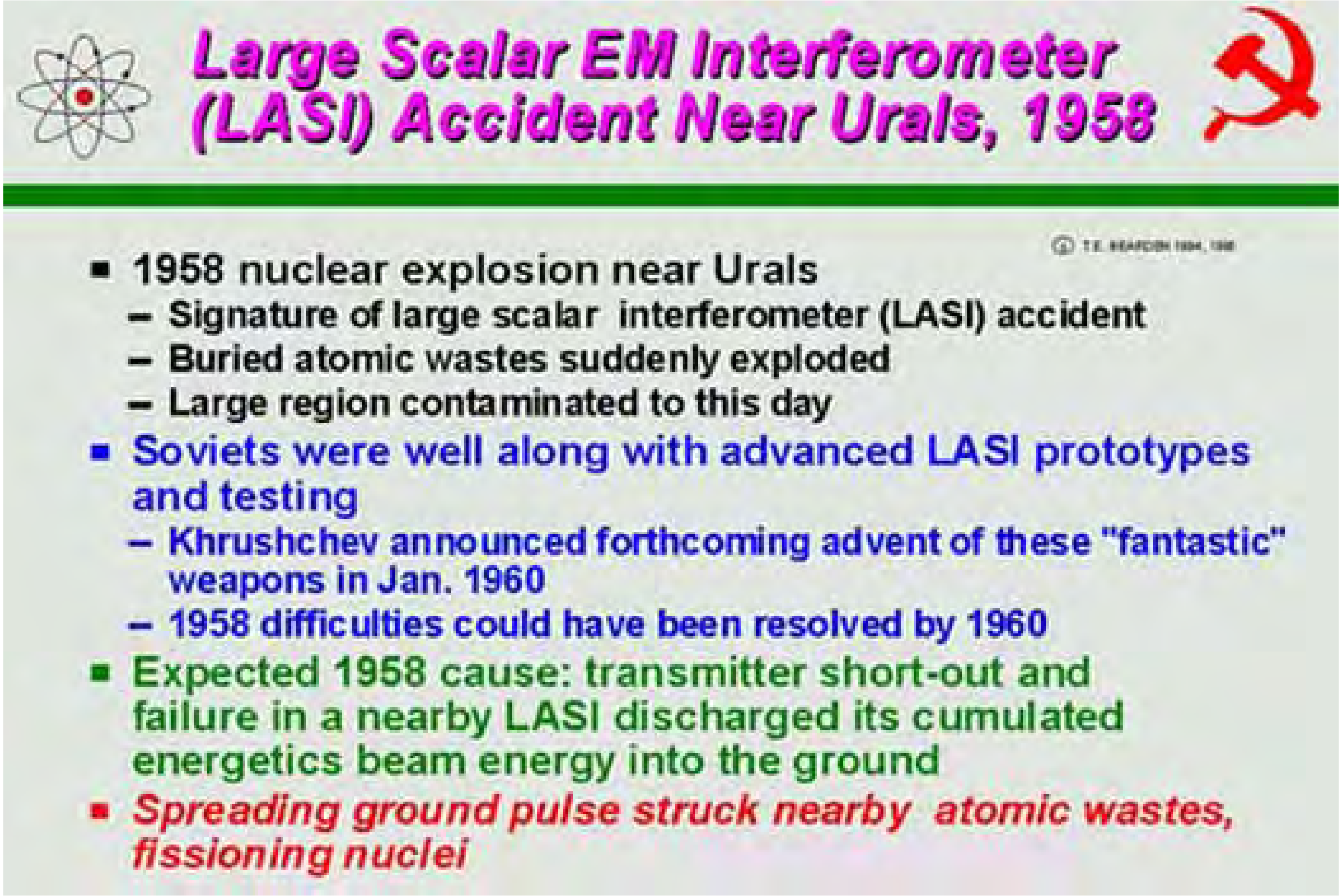

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

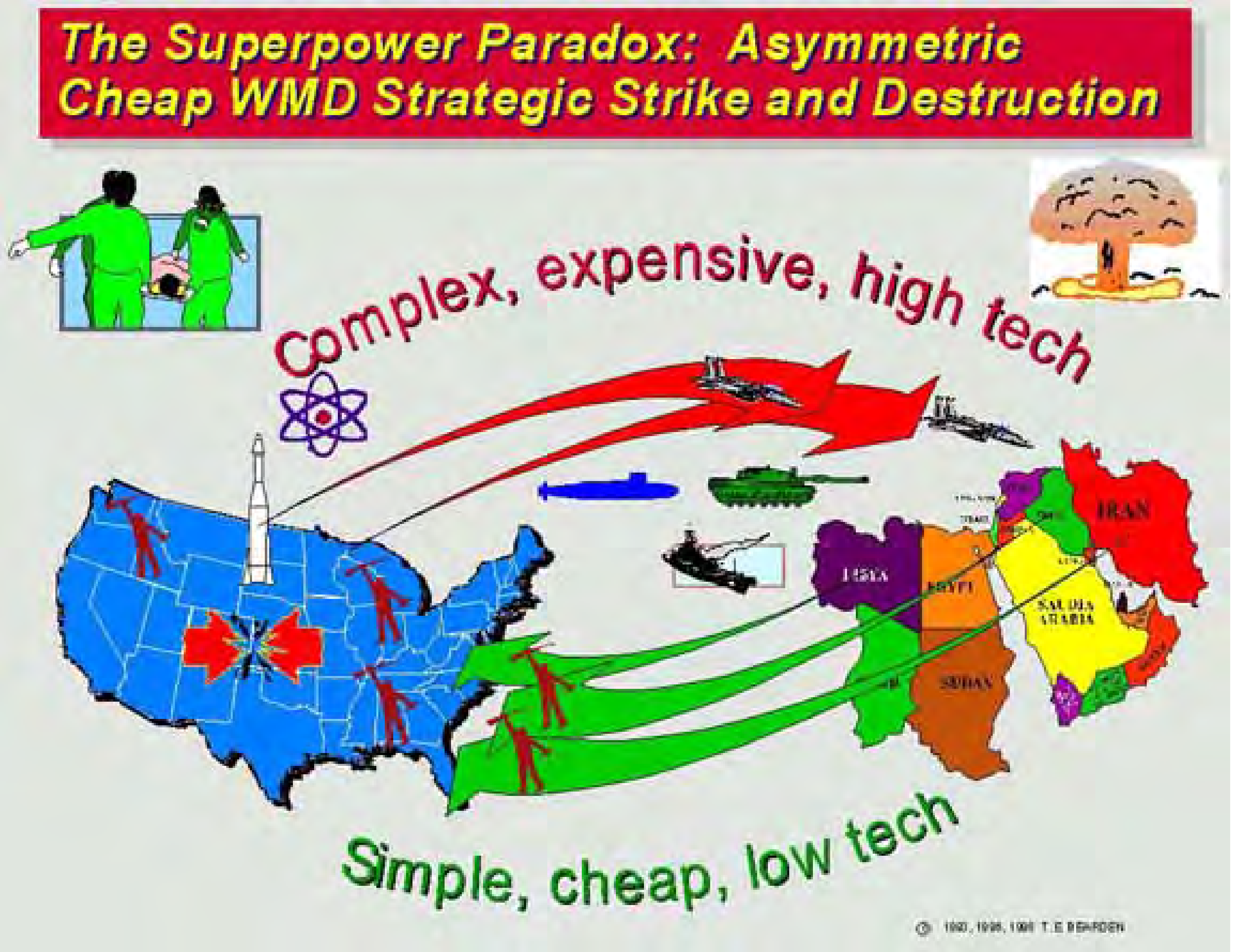

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech
LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN
Simple, cheap, low tech
© 1990, 1998, 1999 T. E. BEARDEN

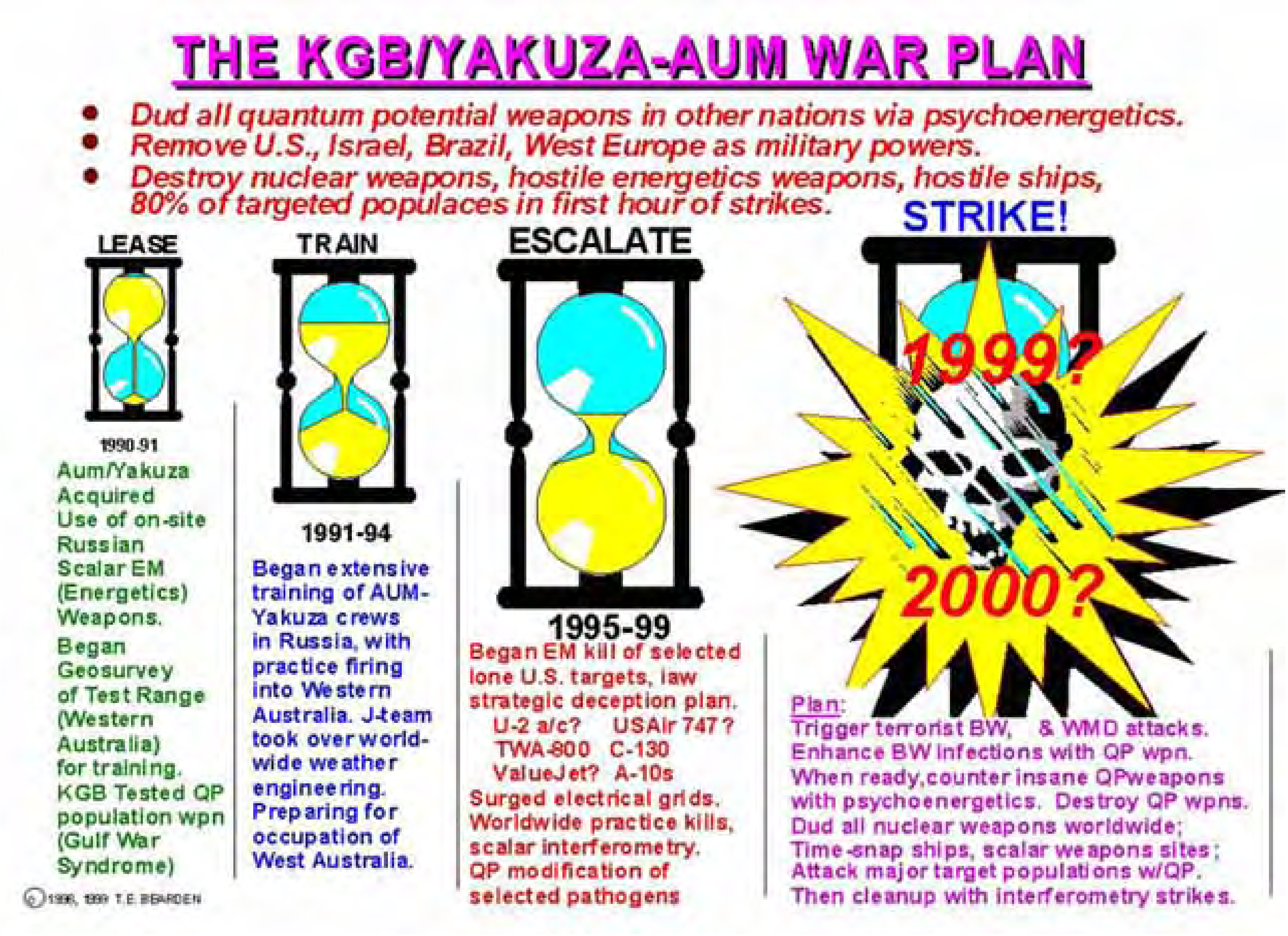

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training.

KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, law strategic deception plan.

U-2 a/c? USAir 747 ?
TWA-800 C-130
ValueJet? A-10s

Surged electrical grids. Worldwide practice kills, scalar interferometry. QP modification of selected pathogens

STRIKE!

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready, counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

© 1996, 1999 T.E. BEARDEN

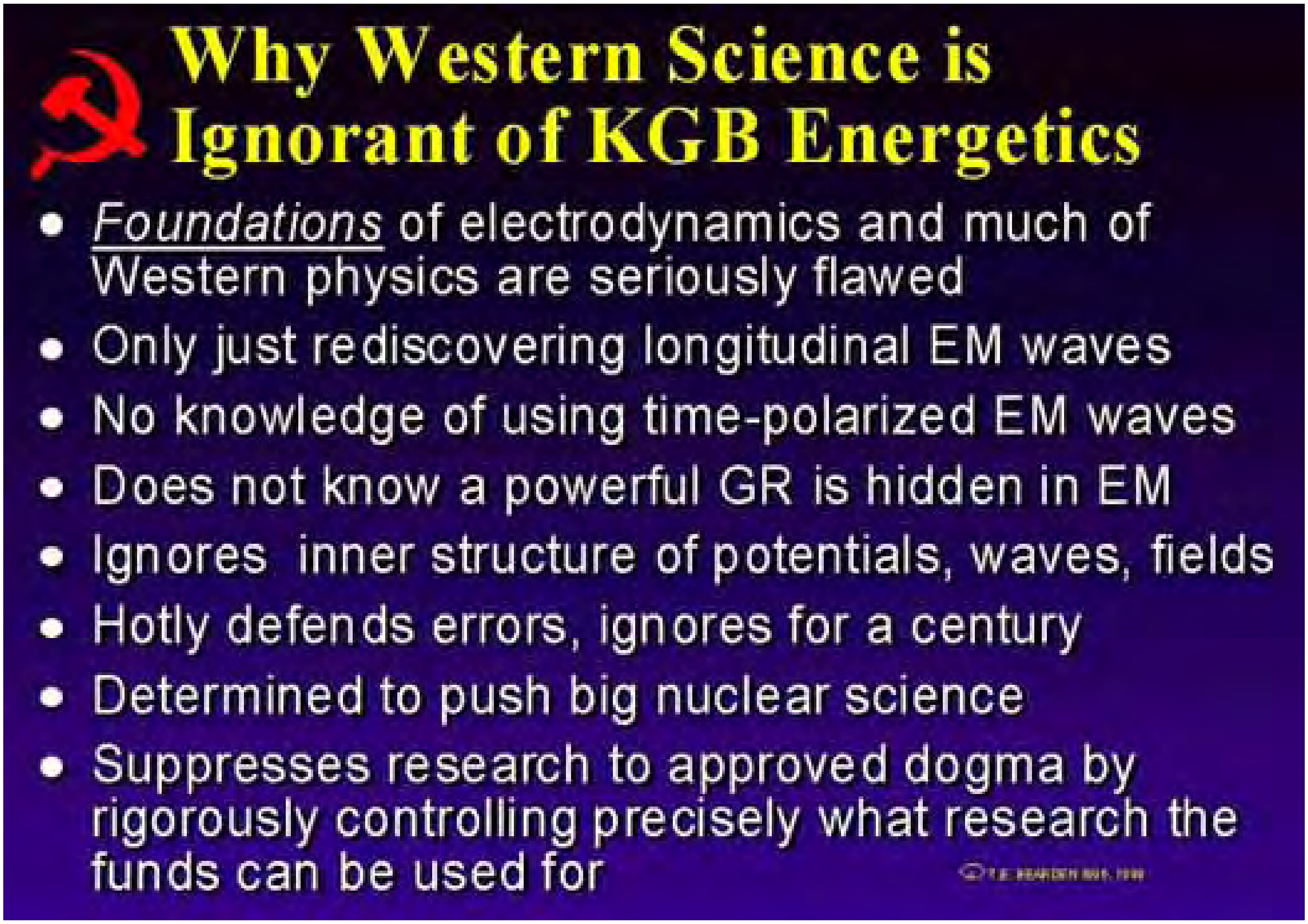

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

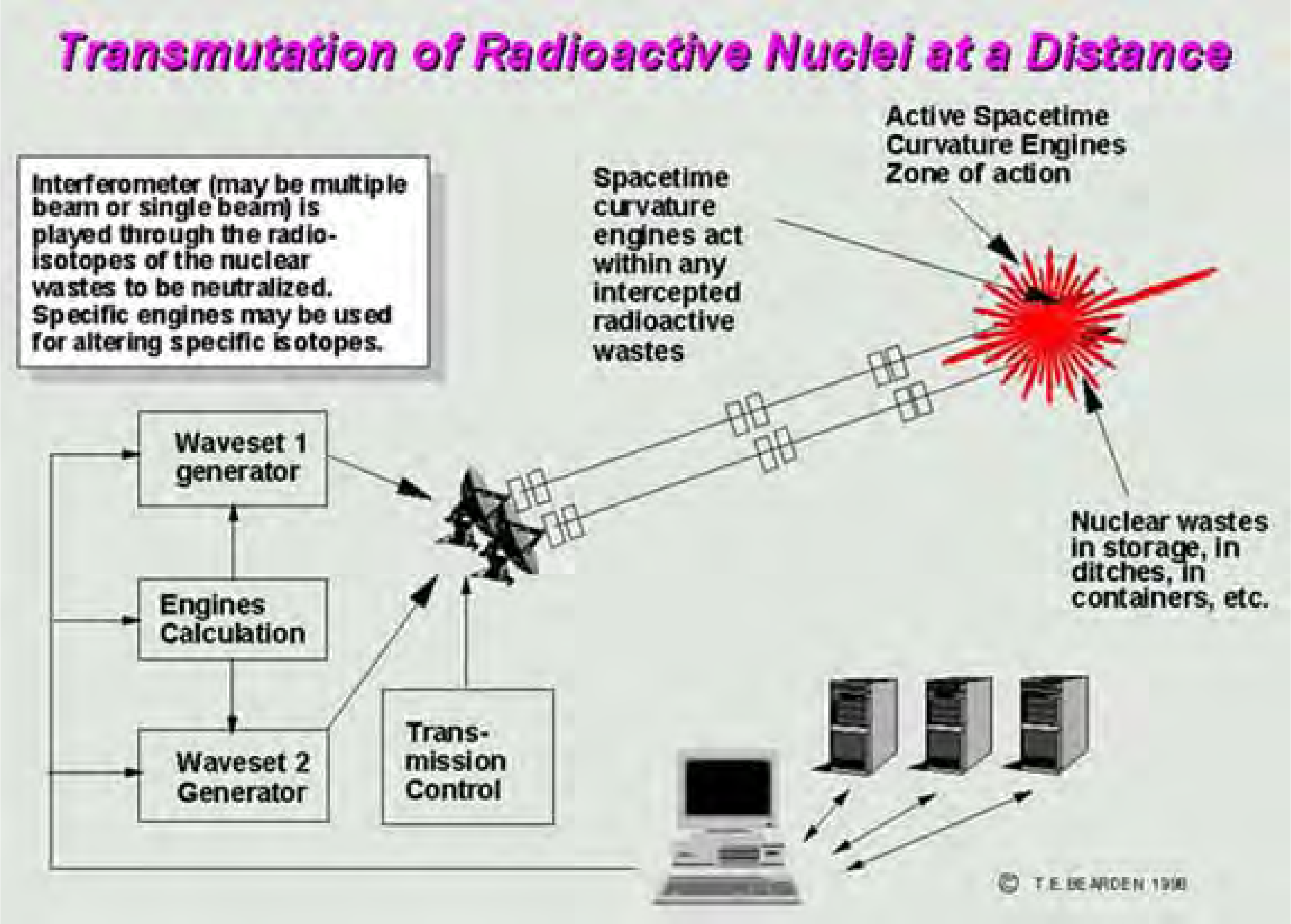

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes
Active Spacetime Curvature Engines Zone of action
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Nuclear wastes in storage, in ditches, in containers, etc.
© T.E. BEARDEN 1998

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

The Last Samurai, Defending Bushido

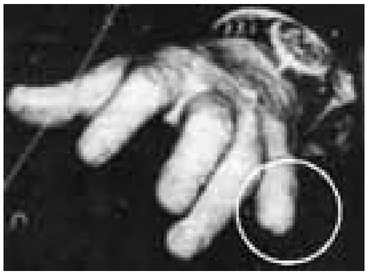

- 3 Major Yakuza Groups, In Coalition.
- Penetrate Japanese Government.
- Open Offices, Signs on Street.
- Public Romanticizes Yakuza as Last Japanese Samurai
- In Large High-Tech Corporations.
- In Control of Japanese Banks.
- Japanese Real-Estate Rip-off.
- $500 - 1,000 Billion per Year Income.
- Rapidly Spreading in U.S.
- Uniting Mafia Groups Worldwide.
- Coalition with Aum Shinrikyo.
- Leased Russian Energetics Weapons.
- Crews On-site in Russia, Trained.
- Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)
- Now in Escalation Phase, Killing U.S. Targets.
- All-out Strike on U.S. ready 1999-2000.
- Now held in check by KGB, dead-man fuzed by small nation friendly to U. S.

© 1995, 1996, 2001 T.E. BEARDEN

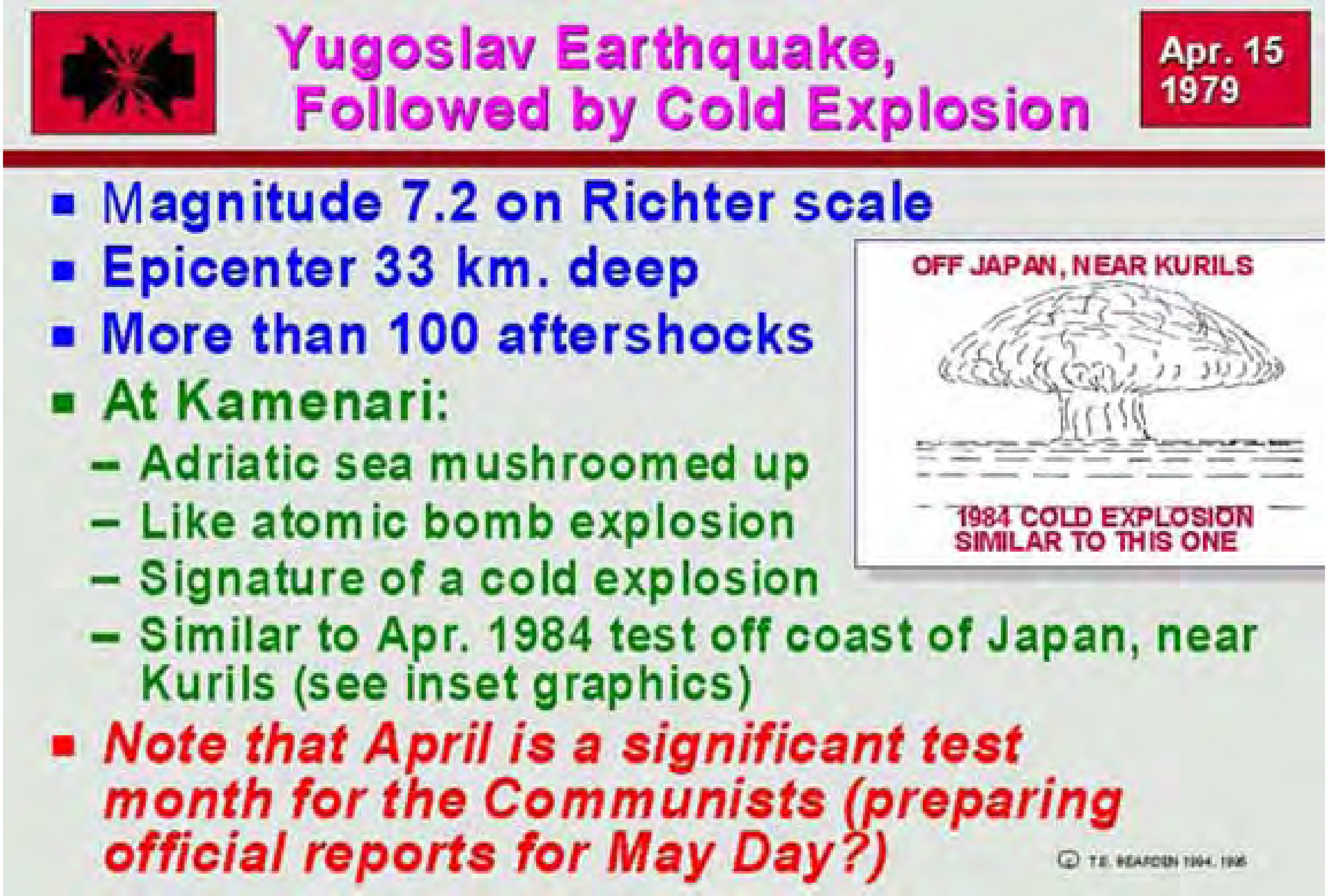

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

SCALAR WEAPONS

UNDERSTANDING THESE MAKES IT EASIER TO UNDERSTAND THE NATURE OF THE AETHERS. WE'LL START WITH TOM BEARDEN'S FACTUAL INFO ABOUT THE RESULTS OF THE RUSSIAN'S SCALAR WEAPONS PROGRAM

http://www.cheniere.org/books/ferdelance/intro.htm

THE RUSSIANS USED SCALAR WEAPONS TO BLOW UP THE CHALLENGER SPACE SHUTTLE, A NUMBER OF OTHER LAUNCH VEHICLES, VARIOUS JET AIRPLANES, MODIFY THE WEATHER AND KILL UNTOLD NUMBERS OF BIRDS & FISH, WHICH HAS HAPPENED AGAIN RECENTLY. HAARP DEVICES CAN DO THIS.

SIMPLY PUT: A SCALAR WEAPON IS A FIGURE 8 WAVE FORM THAT IS NOT IN HARMONY WITH AN AMBIENT AETHER, IF POWERFUL ENOUGH THE SCALAR WEAPON CHANGES THE SIZE OF THE BACKGROUND SCREEN AETHER, WHICH THEN NO LONGER SUPPORTS THE MANIFESTATION OF MATTER, WHICH DESTABILIZES AND EXPLODES.

THE REASON WE ARE NOT ALL SPEAKING RUSSIAN IS BECAUSE THE MOST UNSTABLE MATTER IS IN NUCLEAR WARHEADS - DESTROYING THE ENTIRE WORLD IS NOT WINNING A WAR. PROVIDED, OF COURSE, THAT YOU LIVE ON THAT WORLD - THERE IS EXCELLENT EVIDENCE THAT A SCALAR WEAPON WAS USED TO DESTROY THE BIG BEAUTIFUL PLANET WHERE THE ASTEROID BELT IS NOW. THAT'S IN THE BOOK "COSMIC WAR".

GIGGLE, LAUGHING MY ASS OFF AT THIS ONE FROM BEARDEN:

It has also become fashionable in the West to believe that all the laws of physics are already discovered. We assume we already know all of them. While we have been pridefully crowing this tune, the Soviets have been steadily discovering new laws in secret, as well as new ways to circumvent the old laws.

THE SAD PART IS THAT IT HASN'T CHANGED - WESTERN SCIENTISTS ARE IDIOTS WHO ARE LEADING THE WORLD INTO DESTRUCTION. OR SITTING BACK AND WATCHING THE DESTRUCTION, WRINGING THEIR HANDS AND SAYING THAT FUKUSHIMA RADIATION WILL BE WITH US FOR 1000s OF YEARS. B.S. THERE ARE A NUMBER OF ENERGETIC WAYS TO CANCEL RADIATION INSTANTLY.

BEARDEN TAUGHT ME ABOUT SCALAR ENERGY - BUT, THEN HE HAS STARTED PARROTING LAME STREAM PHYSICS - PROBABLY TO KEEP HIS HEAD FROM BEING SEPARATED FROM HIS BODY BY OUR CONTROLLERS.

SCALAR WEAPONS ARE DESCRIBED IN A NUT SHELL ABOVE. BUT, WE'LL EXPAND IT A LITTLE - THE SIMPLE FACT THAT THEY WORK IS PROOF THAT THE AETHERS IS THE BACKGROUND SCREEN THAT SUPPORTS THE MANIFESTATION OF MATTER, WHICH IS ALL DIVINE I.E. MADE FROM IMAGES.

THE AETHERS PROGRESS IN SIZE ACCORDING TO THE FIBONACCI SERIES, BECAUSE EVERYTHING FROM THE QUANTUM TO THE COSMOS MANIFESTS ACCORDING TO THE FIBONACCI SERIES - ITS ALL FRACTAL I.E. A RECURRING MATHEMATICAL PATTERN, THAT IS PERFECT THROUGHOUT CREATION. HERE'S THE PROOF - OUR ENTIRE GALAXY IS IN PERFECT HARMONY WITH ONE NUMBER = RIDICULOUSLY AMAZING

http://blog.hasslberger.com/docs/PERFECTION.pdf

THUS, IF YOU CHANGE THE BACKGROUND SCREEN, THEN MATTER HAS TO CHANGE = THE RESULT OF USING A SCALAR WEAPON.

THE PRECISE WAY THE WEAPON WORKS IS TO IMPOSE A MAN MADE SCALAR WAVE ON AN AREA OF SPACE. GIVEN ENOUGH POWER TO THE SCALAR WAVE, IT WILL CAUSE THE AETHER IN THAT AREA TO MATCH ITS SIZE (LIKE ATTRACTS LIKE). WE ARE TALKING UNBELIEVABLY SIMPLE.

YOU CAN MAKE ONE YOURSELF - AMAZINGLY THE REQUIRED WINDINGS OF A SCALAR PRODUCING COIL CAN BE USED AS A WEAPON OR FOR HEALING - ITS CALLED A RADIONIC WITNESS WELL WHEN USED FOR HEALING.

TO MAKE A WITNESS WELL ONE WINDS INSULATED WIRE AROUND A GAP IN SPACE I.E. THE CENTER OF A TORROID. THE WIRE IS WOUND TO THE LEFT AND TO THE RIGHT. THE MORE WINDINGS THE MORE POWERFUL. AND, OF COURSE, THE POWER OF THE WAVE IS A FUNCTION OF THE POWER OF THE DC ELECTRICITY USED TO ACTIVATE THE DOUBLE WOUND COIL. SUCH A COIL PRODUCES A FIGURE 8 WAVE THAT IS THE SIZE OF THE CENTER OF THE COIL. IF THAT SIZE HAPPENS TO BE THE SIZE OF AN AMBIENT AETHER, YOU'RE NOT GOING TO HAVE A WEAPON.

IF YOU MAKE SUCH A COIL FOR HEALING, POWER IS NOT REQUIRED, IT AUTOMATICALLY CONNECTS WITH THE AETHERS IN A POSITIVE WAY.

IF YOU MAKE IT AS A WEAPON ADD LOTS OF POWER AND BE A LONG WAYS AWAY, WHEN YOU TURN IT ON. THE DESTRUCTIVE EFFECT IS INSTANTANEOUS - I'VE SEEN IT, WE BUILT ONE.

SCALAR WAVES ARE ALSO REFERRED TO AS PHASE CONJUGATED WAVES, MEANING TWO SINE WAVES THAT ARE 180 DEGREES OUT OF PHASE FROM ONE ANOTHER. (WESTERN LAME STREAM IDIOTS SAY THAT THOSE TWO WAVES CANCEL ONE ANOTHER.)

ONE SINE WAVE IS SAID TO RUN IN TIME FORWARD AND THE OTHER IN TIME REVERSE. AND THERE ARE A NUMBER OF EFFECTS THAT PROVE THIS. ONE EXAMPLE: DON BRANDT (ONE OF BEARDEN'S PROTEGEES) USED A SCALAR WAVE TO REGROW A LEG ON A CAT (A KID HAD SHOT IT OFF WITH A .22), WHEN DON FINISHED, THE CAT HAD ITS LEG BACK AND ALL OF THE CLOCKS WITHIN FIVE MILES WERE 15 MINUTES BEHIND TIME - THAT WAS THE LAST TIME HE PLAYED WITH THAT.

IN 2-D THERE IS NO LINEAL TIME.

BOTTOM LINE: SCALAR WEAPONS PROVE THE NATURE OF THE AETHER AS THE BACKGROUND SCREEN - IF THAT WAS NOT SO, THEN SCALAR WEAPONS WOULD NOT WORK.

ALSO, SCALAR WAVES THAT ARE IN HARMONY WITH AN AMBIENT AETHER ARE VERY BENEFICIAL AND USEFUL - A BEE'S HONEYCOMB IS AN EXCELLENT EXAMPLE. DITTO GRAPHENE.

LOVE AND BLESSINGS

PS: THERE HAVE BEEN A NUMBER OF HUMAN "CIVILIZATIONS" IN THIS SOLAR SYSTEM AND ON THIS PLANET. OUR CONTROLLERS LIKE US TO BE STUPID, BECAUSE IT SEEMS THAT WHEN EVER WE MAKE GREAT TECHNOLOGICAL ADVANCES OR SPIRITUAL ADVANCES, THEN THEY WIPE US OUT. WE ARE AN ENTERTAINMENT SLAVE SPECIE.

THIS LAYS OUT HOW OUR DREAMS / ANUs ARE SHARED WITH OTHERS AND THAT JESUS' ENERGY IS FOUNDATIONAL
http://blog.hasslberger.com/docs/GOLDFISH_JESUS_REVELATIONS.pdf

THIS IS FUNNY AND SHOWS THAT THE EARTH IS THE BEST ENTERTAINMENT SYSTEM IN ALL OF GOD'S CREATION:
http://blog.hasslberger.com/docs/EARTH_WINS.pdf

OH - THE CHEMTRAILS WILL KILL JUST ABOUT EVERY LIVING THING ON THE FACE OF THIS PLANET, IF WE DON'T START MEDITATING AND BUILDING ORGONE DEVICES, WHICH HAS TO BE DONE EVEN IF THEY STOP SPRAYING BECAUSE THE EFFECTS OF THE CHEMTRAILS CAN AND WILL CARRY ON, EVEN IF THE SPRAYING STOPPED.

Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

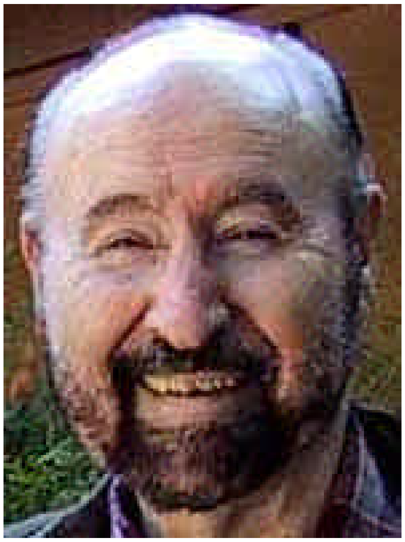

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world" has come about by studying the information given by Col. Tom Bearden at his website Cheniere. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity. This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "**Scalar Electromagnetics**." But there is a great deal of information at Cheniere which needs to become common knowledge as fast as possible, for the sake of the survival of life on earth. To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of bolding certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic**," and "**LW**" means "**longitudinal wave**")

This article has six sections:

New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels." Nikola Tesla

"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself." Tom Bearden

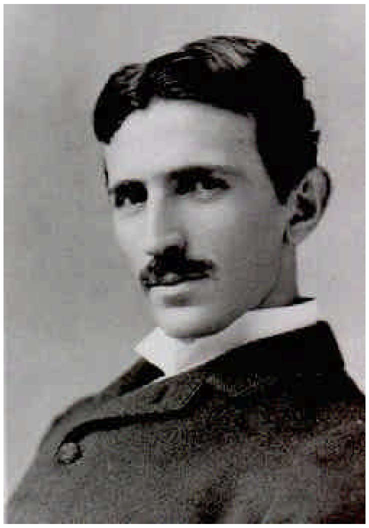

I guess the first thing to try to comprehend is that a "new" kind of electromagnetic (EM) wave has been discovered in the empty vacuum of space called “longitudinal” waves, which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the discovery really goes back to **Nikola Tesla** and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology for over 30 years and have **weaponized** these "new" longitudinal scalar waves to a great degree.

These are the very weapons **Nikita Khrushchev** spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could wipe out all life on earth if unrestrainedly employed.'" Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, discusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

> The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:
>
> 1. The **solutions to the energy crisis and the "oil problem" are in hand**. These oil wars are unnecessary. There is endless energy available freely from the domain of time.

2. **Unbelievably powerful weapons** are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.

3. The **cure of diseases such as cancer and AIDS**, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.

4. **Mind control on a mass scale** has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.

So the layman will need to understand that there is a new kind of electromagnetic energy that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called transverse EM waves, to distinguish them from the new longitudinal EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that the "emptiness" of empty space is in fact not empty, but a great ocean of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." It seems like something from Star Trek but this is where scalar electromagnetics has come to. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also spacetime, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that time itself is actually compressed energy. And it is energy which is compressed by exactly the same factor by which matter is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous E=mc2. It is now paired with E=tc2 (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"

and

"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash the tremendous energy that is compressed into time itself. It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, several new names

> "Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "energetics". That is their approach to a unified field theory, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "bioenergetics"), and all mind operations and mind-matter interactions (the third branch of energetics, called "psychoenergetics"). [ed: sometimes also called "psychotronics" by the Russians]

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the new science this way:

> "Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) which underlie all physical reality.
>
> "If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to alter gravity, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, space-vehicle drives, time-travel, teleportation, paranormal phenomena, and just about every other area one can think of." http://twm.co.nz/Beard_scalem.html

> "Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. Time-as-energy eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with spacetime and with spacetime curvature."
>
> Bearden: http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

The scalar energy discovered in the vacuum of space is sometimes called "zero-point" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical "creation of the vacuum" as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas. Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we've got a real vacuum. Right? Well, since then, both theory and experiment have shown that there is a non-thermal radiation in the vacuum and that it persists even if the temperature could be lowered to absolute zero. Therefore, it was simply called the "zero point" radiation. "Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

> "**For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960**." - Bearden
>
> http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that a permanent solution to the "energy crisis" is at hand, if only those who hold it so secret would give it up to humankind.

It means our species' insane use of oil is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that unbelievably powerful weapons can be, and already have been developed. In fact these weapons are so dangerous that the fear engendered becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the huge energies of time, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because most people would indeed be truly alarmed to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and ban the use of scalar weapons in war or in mass mind control. The current deployment of these weapons is a world emergency which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole

The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the essential secret of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. Just make a dipole.

> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, nature then gladly gives us as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the giant negentropy mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole."
>
> Bearden http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!

Perhaps the most eagerly awaited event now is the marketing of Bearden's Motionless Electromagnetic Generator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."

(Bearden : http://www.cheniere.org/correspondence/011202.htm)

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

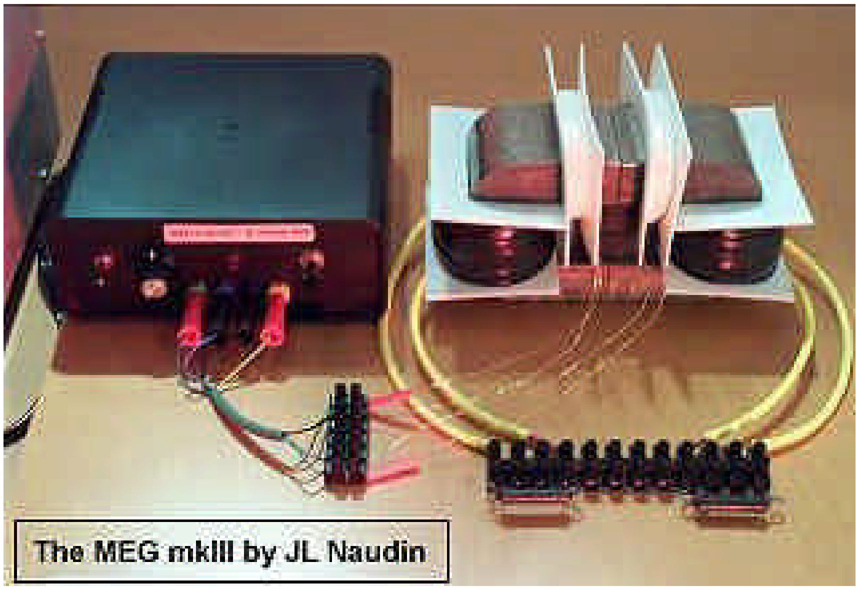

The Motionless Electromagnetic Generator

(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But Magnetic Energy Limited's MEG is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government on the order of the Manhattan Project.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through Magnetic Energy Limited, Inc. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S."
>
> Bearden http://www.cheniere.org/correspondence/121901.htm
>
> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
>
> Bearden http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden http://www.cheniere.org/megstatus.htm

Free Energy has been Suppressed

Secret Government Aware of Free Energy

> "In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] De Palma and his ideas.
>
> "Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that free energy is a threat to the world order constructed by business and the mindset of those who want to own the world.
>
> "Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.
>
> "The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an

> attitude of paranoid delusion and fantasy which can only arise from the alienation of a class of elitists who, through their money, are completely insulated from the reality of the day to day fight for existence of the common man."
>
> http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

Bearden describes one case of suppression when the Yakuza killed the Kawai engine in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> "They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is “The Unnecessary Energy Crisis: How to Solve It Quickly,” especially since the "oil wars," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.

> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper."

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The good news is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the dipole, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the greatest discovery in human history lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "conspiracy of ignorance" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere. "As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.
>
> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized

> Science communities. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden National Emergency Declaration is Needed
>
> "The Government Non-Technical community (the Senate and the Congress, in particular) are in far better shape than the Government Technical community, to appreciate the world implications of the pending economic disaster. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to get this vacuum energy program launched, under a National Emergency declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."
>
> unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

> "Personally, the present author regards the increasing energy crisis as the greatest strategic threat to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.
>
> ...In the name of all humanity, let us begin! Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

---

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> "Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap." Bearden http://www.cheniere.org/books/ferdelance/intro.htm

"**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that,**

if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon." Khrushchev, to the Presidium, Jan. 1960

> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target from the local vacuum at the place of the target.

From an interview:

> Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"
>
> Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."
>
> http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

3a. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these new facts public and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "Tesla Howitzers." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The first howitzer mode is called the "exothermic" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply destroy all electronics in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "leaders" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a new world, whether we like it or not. Anyone could be expected to go into denial, the befuddled senators and congressmen included. But we have to wake up.

In a second howitzer mode called the "endothermic" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "cold explosions" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

> "The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these heat (exothermic) and cold (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the endothermic howitzer mode as a weapon of war:

"Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary transverse EM waves, the longitudinal [LW] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space at the location of the target itself. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

You can hear the Woodpecker signal here: http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the LW waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers to a pinpoint location on earth. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "Mindsnapper" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would not rot for quite a long time.

The Mindsnapper can be used for a small area, or to kill an entire population. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in mind-snapping mode against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

> "For example, suppose one attacks a tank. [with an LW weapon] With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
>
> All electronic systems of the tank are dudded. Total systems kill is achieved. The ammunition in the tank explodes. Total kill is achieved.
>
> The fuel explodes. This is another total kill mechanism.
>
> That tank has been killed totally, by a variety of mechanisms, all simultaneously.
>
> That's a k-kill in any analyst's book.
>
> And don't worry about retrieval and repair. That tank is finished permanently.
>
> Even the most modern tank is just as vulnerable as the most obsolete."
>
> Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And Star Trek's "Phaser," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and materialization and dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to

> underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even affect the sun.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is no longer mutual. And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a full-scale all out scalar attack on the west.

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counter terrorism] efforts." - Defense Secretary William Cohen, 1997
>
> In a letter to a writer named "Russell" Bearden says,
>
> "In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate,

> and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Tesla Howitzers in Weather mode

How to Make Weather

> "Engineering the weather is duck soup; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.
>
> "First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.
>
> "So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather radar anomalies which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

In another section at Cheniere are many photos of cloud anomalies which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these clandestine ongoing "weather wars."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. This drought was broken by a colleague who used an extremely powerful scalar EM device to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions

resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.

"Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."

* On the weather maps.

http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion

Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

> "Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.
>
> "Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.
>
> "The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."
>
> Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The deepest dangers of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing true catastrophe.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that human activity could actually affect the sun itself is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental

that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.

"This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about."

http://www.cheniere.org/books/ferdelance/s24.htm

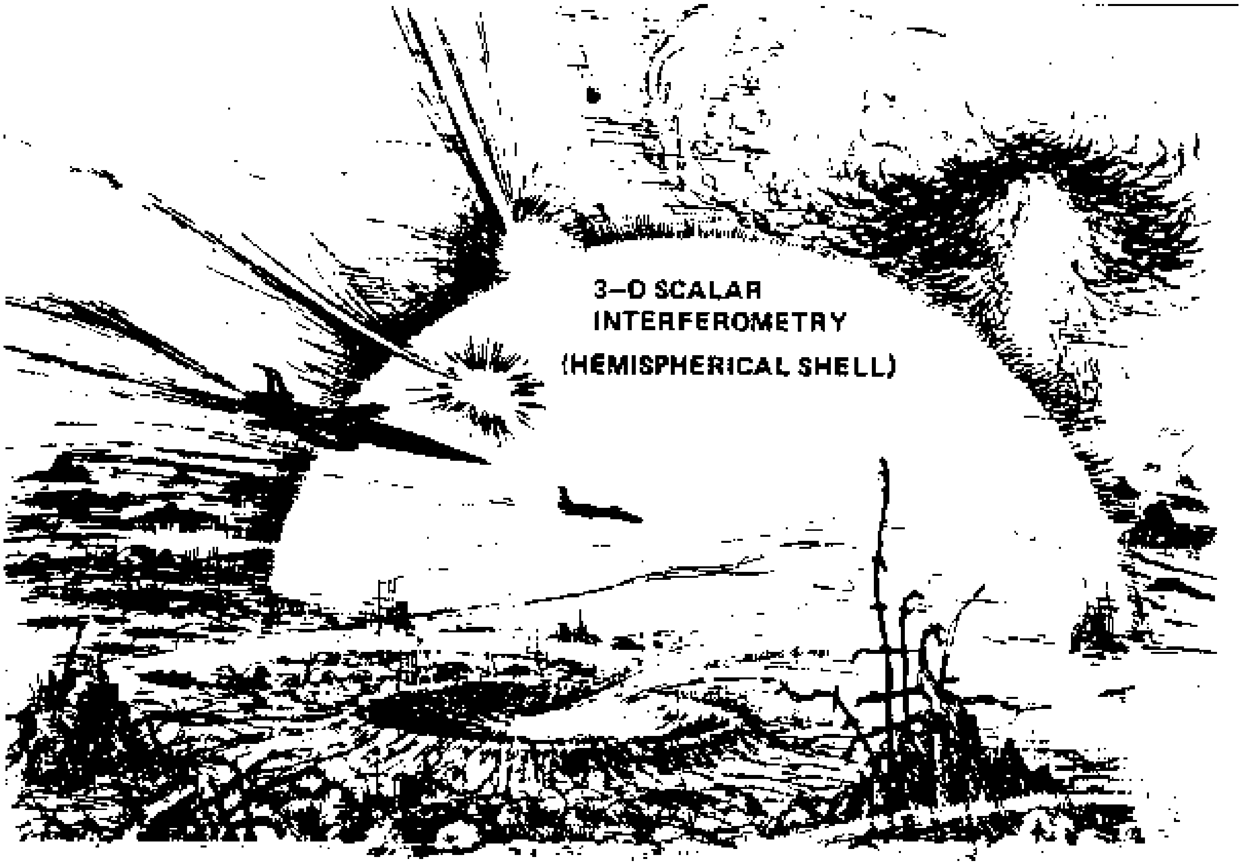

3b. Tesla domes provide near-perfect defense shield.

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly impregnable defensive mode whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum at the location of the shell. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc."
>
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area even against nuclear radiation itself. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (operational) a nearly perfect missile defense system. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world:

March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 | June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3c. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "rogue groups" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government."

> "They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control."
>
> "The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

3d. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now lots of new ways to bring any nation to its knees.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that it will collapse the U.S. economy in a way we have not seen since 1929. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

> Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.
>
> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament. Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???"
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

When he was asked the question '**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. The West is no longer defenseless, not by any means. Cannot amplify it further than that."
>
> http://www.cheniere.org/correspondence/030202a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

3e. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such scalar-powered antigravity vehicles, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmosphere.

> "They could seem to materialize and dematerialize.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now in great secrecy. Our own "government" may even be one of these.
>
> "If so, it isn't our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
>
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
>
> 3) the light-balls are able to eject smaller light-balls;
>
> 4) the light-balls change shape all the time;
>
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. But the cause, and the physical mechanism with which radiation is emitted, is currently unknown."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely baffled as to where the energy of these plasma balls was coming from.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomenon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. But where does it come from? We don't know."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls is coming from the vacuum of spacetime at the very location of the balls themselves, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as marker beacons giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many anomalous events of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The

stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: Interview with Tom Bearden

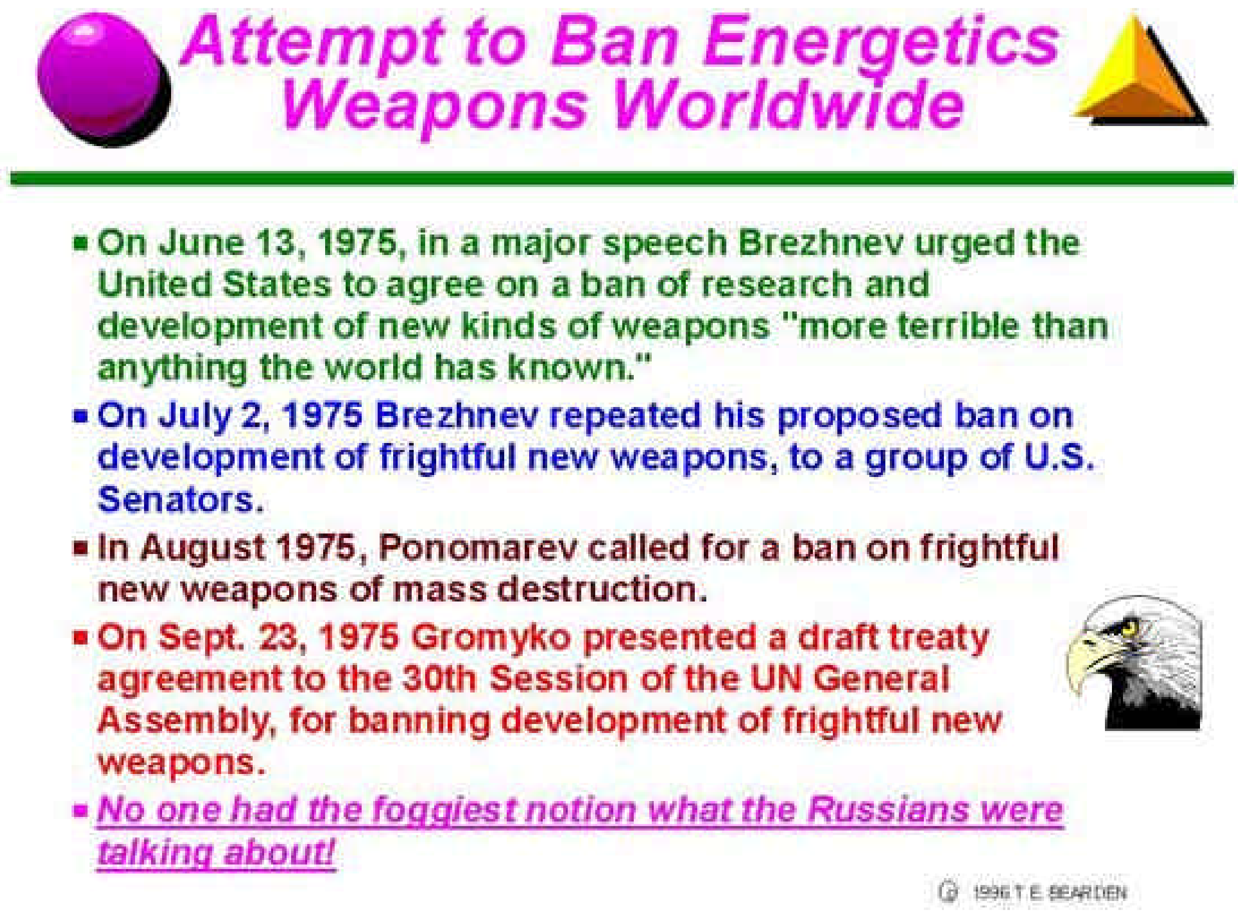

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T E BEARDEN

(Slide from Cheniere.org)

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

> "Any sort of disease whatsoever - physical or mental - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."
>
> Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a gigantic glass tube filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore project was funded by the French Government to the tune of several millions of dollars (equivalent). http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "phase conjugate pairs" and one of the pair is time-reversed. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually time-reverse the cells back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to suffer the fact that the cure for cancer and AIDS was right there, in this new scalar electromagnetics technology, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Priore, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that big giant two-story glass tube consistently cured cancer in animals.

"**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for controlled cellular dedifferentiation, in biology terms. Physicians will time-reverse diseased, damaged or aged cells back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."

> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor Antoine Prioré. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.
>
> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces cellular reversal by weak time-domain pumping of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of healing rather than a science of intervention. And it will develop as a science of unlimited healing, since no pathogen can resist action by curved spacetime engines." Bearden
>
> http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine   See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an antenna unit which surrounds the body. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

In fact, Bearden has presented to the U.S. government a plan to use a portable version of a Scalar Healing Blanket where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A two minute treatment of the waves is sufficient to restore the body to its earlier healthy condition. Thus even thousands could be treated easily and quickly in a biowar mass attack situation.

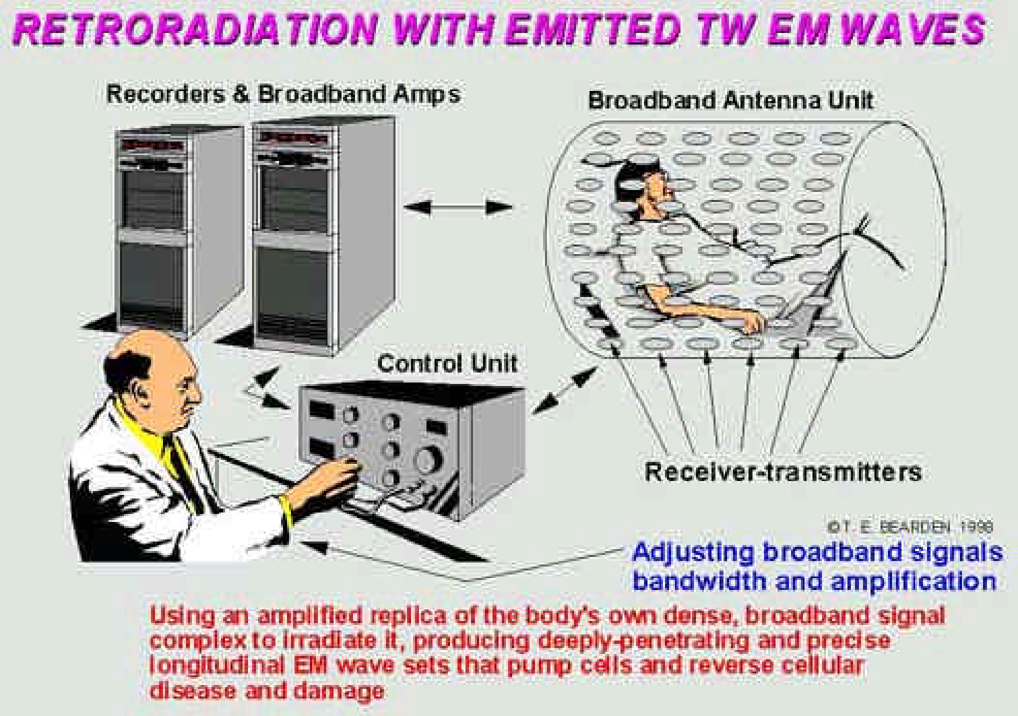
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
© T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

"To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. "The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals,

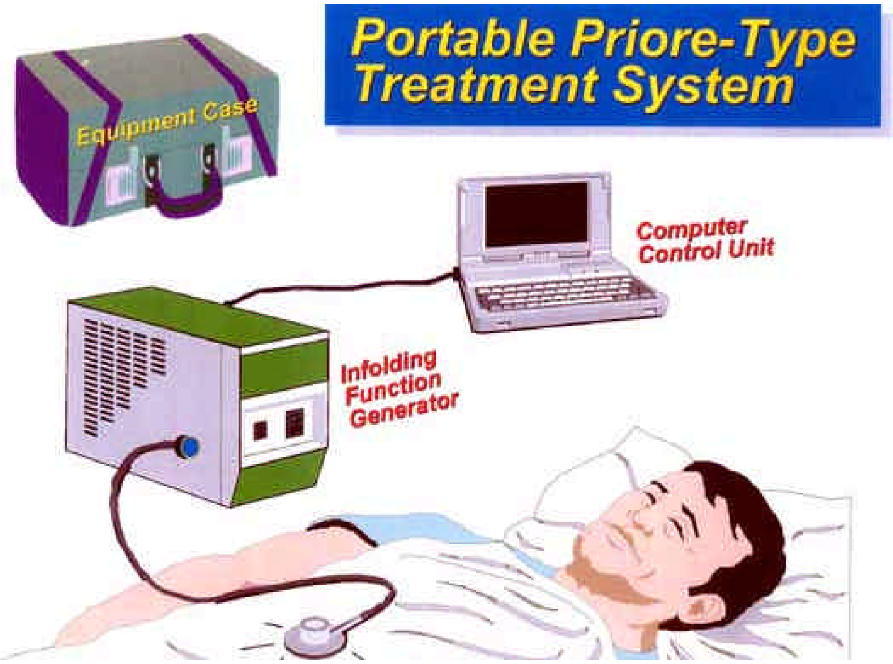
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator

> emergency response teams, etc."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p04.htm

The Scalar Healing Blanket would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on befuddled government ears. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article '**Healing with Energy**": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:
Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the eventual deaths of three U.S. Ambassadors."

> Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

> "Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases

Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a conventional bio-attack would be greatly potentiated. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the "shadow" diseases that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these 'vacuum engines' or 'spacetime curvature engines' that acts on mass to generate the exact effects produced by anthrax. (Or any other disease one wishes) . . . Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .
>
> "And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the "shadow state" of a disease, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .
>
> "So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the necrotizing fasciitis disease engine. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. A few of those targeted bodies will now actually develop necrotizing fasciitis. So far, just like the U.S. Embassy targeting, except a different "carrier".
>
> "But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. The few cases of necrotizing fasciitis that break out will be most puzzling. They will be randomly scattered across the

> entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."
>
> "And that one happened, right here in the U.S., not so long ago. Check it out.
>
> "So carefully adjust the patterns for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has 12 shadow state disease engines in it.
>
> "Place that one on a populace. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "spreads" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.
>
> "Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. In the "thinned" set, one can at least double or triple, and even up to five times, that estimate. So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "That one is being set up, or already set up, right now. There are other such "augmentation" setups also under way.
>
> "So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "flesh eating disease." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This quantum potential disease induction weapon -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., one test used necrotizing fasciitis (flesh-rotting disease) engines.

> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. The major signature of the test was that the cases of the disease broke out totally statistically and randomly, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, but was a totally random set of occurrences. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

"Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.

> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that they would never be able to resist doing it."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the Gulf War Syndrome involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> "I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced... It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire electrical grids of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have no idea yet of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and engineer the human psyche. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no thing you can point to that is the mind. As Bearden says**, "The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the **interference zone** (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "Mind Wars."

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: "**One if by land, Two if by mind**."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden http://www.cheniere.org/misc/time.htm

In a slide Bearden illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself. For if my thoughts might no longer be "mine," then who and what am I? Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is “Mind Control and EM Wave Polarization Transductions” This is such serious stuff that Bearden includes a strong warning about misuse of this knowledge. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization's can be transformed one-into-the-other are presented."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines

> (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired... Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gail's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p09.htm

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p08.htm

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p01.htm
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p02.htm

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour. "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p06.htm

The paper "Mind Control and EM Wave Polarization Transductions" ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials. "The KGB scientists also know that something very like Gaia — a collective unconscious mind for all species on earth — also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials.""Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can

> then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (all systems tend to feudalism or some form thereof). "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.

> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies." Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

"Anyway, it's presently a "Mexican stand-off" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."

"You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters can be made fairly readily and cheaply from nearly just Radio Shack parts or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take."

Bearden http://www.cheniere.org/correspondence/021702a.htm

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: this time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the wisdom of the wise could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, there is no limit to what humanity can achieve. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation. "So why not release these technologies? Because the social, economic and geo-political order of the world would be greatly altered. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the covert arsenal — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may

> endanger our future. Only a full, honest disclosure will correct this situation. It is not possible for me to convey in words the urgency of this."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has already leased some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo. They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. So the rogue Japanese acquired longitudinal EM weapons and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. This is a world emergency.

And in Bearden's analysis, it all comes back to the issue of energy. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbert Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national Manhattan Project right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call. It was certainly bad enough, and one grieves for those lost American lives. But I fear the worst is yet to come, and it is coming at us like a runaway train hurtling right down the track at us."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 are in terrorist hands also, including the Yakuza and Aum Shinrikyo. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. So a great deal of the state of the world is not covered in the news at all, and will not be . . .
>
> "We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.
>
> "It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (it's going to go anyway, and catastrophically, the terrorists will see to that). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.
>
> "As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The

> scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.
>
> "Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).
>
> "It isn't. It has only just begun."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The quantum potential weapons are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...
>
> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because it uses multiply connected spacetime, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> Bearden http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "Rule by Secrecy," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterward

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potential weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some agenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from Jean-Louis Naudin, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----

From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The Motionless Electromagnetic Generator from Tom Bearden is now PATENTED US 6362718 granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the magnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil extends around each of these posts. Input coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # 6,362,718

Patent Story on Rense.com:

Rense.com

MEG Scalar Energy Device
Patented - Production Starts Next Year

From Bill Morgan
wmorgan@nycap.rr.com
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan
wmorgan@nycap.rr.com

Do the magnets run down after a while?

Further developments: Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated experiments (such as Bohren's experiment,

which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

You're viewing the old pesn.com website

CLICK TO START BLOCKING ADS

Ad Remover

12 THINGS TO KNOW BEFORE BUYING A THERMAL CAMERA

GET THE GUIDEBOOK NOW!

"Free Energy"

News

- PESN Specials
- About
- Pure Energy Blog
- Daily FE News
- Features
- Free Energy Now
- This Week in FE
- Newsletter
- How you can help
- Submit
- Subscribe

Select Language
Powered by Google Translate

Directory
Energy Topics

- Alt Fuels
- Anti-Gravity
- Batteries
- Betavoltaic
- Biofuels
  - BioDiesel
  - BioElectricity
  - Biomass
- Body Electric
- Brown's Gas
- Cold Fusion
- Conservation
- Electrolysis
- Electromagnetic OU
- Fuel Cells
- Fuel Efficiency
  - Electric Vehicles
  - Engines
  - Hydroxy
- Fusion
- Geothermal
- Gravity Motors
- Human Powered
- Hydro
- Hydrogen
- Joe Cells
- Lighting
- Magnet Motors
- Nanotechnology
- Nuclear
- Nucl. Remediation
- Oil
- Piezoelectric

Tesla's Scalar Fields Still Beaming On!

The superluminal (faster than light) scalar or longitudinal waves Nikola Tesla used to magnify and wirelessly transmit power are not just a thing of the past. IEEE engineer, Steve Jackson, discusses and demonstrates how they can be utilized today, and he is open sourcing it here!

by Hank Mills
Pure Energy Systems News

Nikola Tesla is the father of the modern age. His inventions of radio, remote controlled devices, alternating current, and induction motors (among many others), paved the way for modern civilization. Yet other of his inventions barely saw the light of day, despite their tremendous potential.

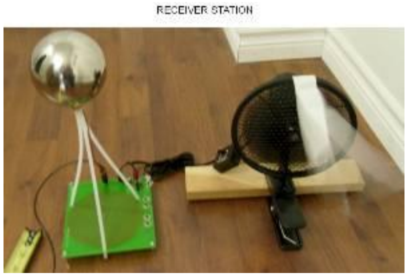

Such a severely under utilized technology is his wireless energy transmission techniques. By producing scalar (longitudinal) waves, he could achieve superluminal (faster than light) communication, send signals through any material including Faraday cages, and wirelessly beam power to receivers. Researchers like Steve Jackson provide hope that by open sourcing Tesla's work, scalar wave transmission can proliferate and positively benefit the world.

Tesla Generator - Build Your Own Generator

Eliminate Your Power Bill Easy. Do It Yourself. Great Discount Now! Just $49.

tesla-energy.org

Last Thursday, March 24, 2011, Steve Jackson held a presentation and demonstration of a scalar wave transmitter and receiver at a local IEEE meeting at McMaster University in Ontario, Canada. During the talk he gave an overview of Tesla's life, his inventions, and a plethora of information on his wireless power transmission technology. Afterwards, he demonstrated a small electric fan attached to a scalar wave receiver powered by a transmitter over twenty feet away. Tesla's technology is alive today!

Jim Murray and Paul Babcock's 4790% SERPS Presentation	Ignition Secrets DVD by Aaron Murakami
Magnetic Energy Secrets, Paul Babcock, Parts 1 & 2	Battery Secrets by Peter Lindemann

Furthermore, he has approached us at PES Network to open source his design. You can download his PowerPoint presentation here. We'll be uploading to

Ads

Electronic Permit System

ePTW system with integrated Isolation and Competency controls

gemsoft7.com

AdChoices

Generator

Generator for Power

GREENfuelH2O.com
Save Fuel Today
Hydrogen on demand
The fuel of tomorrow

SAVE $ AND TIME
Make Biodiesel at Home!
Making ALGAE BIODIESEL At Home

Global Sun Oven

• Plasma
• River
• Salt Water Mix
• Solar
• Solid State Gen.
• Tesla Turbines
• Thermal Electric
• Tidal
• Vortex
• Waste to Energy
• Water
 - Water as Fuel
• Wave
• Wind
• Wireless Electricity
• Zero Point Energy
• MORE . . .

Open Source
• Freddy's Cell
• Bedini SG
• Safe Haven Villages
• MORE . . .

Resources
• Awards
• Conservation
• Conspiracy
• Directories
• Investment
• Kudos
• Legal
• Organizations
• Plastic and Energy
• Recycling
• Suppression
• Tools
• Trends
• MORE . . .

Mingling
• OverUnity Forum
• Employment
• Events
• Humor
• Magazines
• Movies
• Newsletters
• Discuss. Groups

Shopping
• Store
• Buyer Beware
- - - - - - - - - -
- Donate
- Contact

our PESNetwork YouTube account a video of his presentation to the IEEE in a couple of days. It's in process of being transmitted, but will take a while, as it is 3+ Gb. We're expecting to get instructions from Steve on Monday, and we'll be posting a PESWiki page to organize all this material to streamline the presentation.

Take a look at the following video that shows a small boat powered by scalar waves many feet from a transmitter.

A & P Electronic Media
Leading Digital Publishers of Advanced Energy Science
http://emediapress.com

o Fusion Froide
o Fusion Fria
o Fusione Fredda

Free Energy Store
* * * * *
Your Ad Here

Demonstration of Tesla's Wireless Power Transmission System duplicated at the Tesla Tech Conference, July 2007.

Scalar Wave Properties

You might be wondering what in the world is a scalar wave! It's actually pretty simple. Unlike a transverse wave (like an ordinary radio or light wave) a scalar wave's amplitude does not fluctuate up and down. Instead, try to visualize it more like a vibration of electric potential expanding and contracting in the direction of propagation.

Scalar waves are also often called "Tesla Waves" or "Longitudinal Waves." All of these terms refer to the same type of wave which oscillates in the direction of propagation instead of perpendicular to the direction of movement. These waves have many unique properties which have been tested and confirmed repeatedly.

1) They are capable of penetrating any solid object including Faraday Cages. You can put a transmitter in a box of thick metal and a receiver outside of the box will receive the scalar wave frequency you are pulsing. The potential here is for a transmitter that can penetrate any obstacle or perhaps communicate directly through the Earth from one side of the globe to the other.

2) They are capable of superluminal travel. These waves are claimed not to be electromagnetic, but composed of pure potential energy. Due to this, the speed of light limit does not apply to them. The propagation speed of a scalar wave has been measured as faster than the speed of light and thought by some researchers to be potentially of infinite velocity. (Perhaps SETI - The Search for Extraterrestrial Intelligence - needs a different kind of receiver in order to pick up signals from elsewhere in the universe! Why would advanced ET civilizations even bother using slow transverse waves for interstellar communications?)

3) They are capable of transmitting power. It can then be put to use powering motors, lights, vehicles, etc.

4) A scalar transmitter can wirelessly send power to a receiver through any obstacle. For example, as mentioned above, you can setup a transmitter, put it in a Faraday Cage or a metal box, and a

receiver can receive power far away.

5) During the process of transmission and reception they can magnify power. You can input a certain quantity of power into a scalar wave transmitter and yet the receiver can receive several times the power.

6) Transmitters and receivers can communicate.

7) A conventional transverse wave transmitter cannot sense if a receiver has "tuned in." However, in a scalar system the transmitter and receiver interact because they are in a state of resonance.

As you can see Scalar Waves hold many potential applications for communication, energy, and other applications. To mimic a commercial for a popular brand of cell phone, "Scalar Waves do what Transverse waves don't." They are fast, penetrating, connected, and can broadcast magnified power. These waves are begging to be utilized!

Recipes for Scalar Waves

Nikola Tesla's most famous scalar wave transmitter was his Wardenclyffe tower on Long Island, NY. Prior to that, at the height of his amazing inventive career, he had a lab in Colorado Springs, Colorado. He was able to beam power to receivers miles away that were used to illuminate neon tubes. A great deal is known about his tower, transmitters, and receivers. It is all detailed in his patents. This allows for present day replications!

From examination of Tesla's patents, researchers have learned that perhaps the most critical component of a scalar wave system is the antenna. It is composed of a flat "pancake coil" composed of both a primary and a secondary winding. The coil is connected to a spherical metallic resonator that helps transmit and receive scalar waves.

A particularly good source for information about scalar wave systems is the website of German scientist, author, and researcher, Konstantin Meyl. He has published a series of books that describe scalar wave systems, detail their components, and discuss their operation. His books also dive into the science and physics of scalar waves, among other fascinating topics.

In fact, he offers scalar wave kits for sale on his site. They include all the components needed for a complete system, arrive in an aluminum suitcase that can serve in the same manner as a Faraday cage for your experiments, and come with a book detailing the kit. These kits have been used by experimenters to prove to themselves and others the existence of scalar waves and their unique properties.

Take a look at the following video to see Konstantin demonstrating a scalar wave transmitter and receiver. The same device in the following video is for sale right now on the internet from his website.

Meyl shows Tesla's longitudinal electricity transmission in Washington DC at a Tesla,m meeting in 2003!

Even NASA is paying attention to scalar waves to some degree. The paper titled, "Advanced Energetics for Aeronautical Applications: Volume II" by David S. Alexander offers a review of scalar wave (longitudinal wave) research. It mentions the work of Konstantine Meyl and others in the field. Interestingly, the paper also discusses how Dr. James Clerk Maxwell's original equations specifically allowed for scalar or longitudinal waves. Only many years later did short sighted scientists who did not care to deal with Maxwell's complex equations based in "quaternions" mathematics arbitrarily discard longitudinal waves. Due to this, many mainstream scientists doubt their existence despite their history and the mountain of evidence which proves they are very, very real.

Opening Tesla's Technology

Steve Jackson's goal is to open source Tesla's wireless power transmission technology. This is very possible to do, due to the simple nature of the parts and components required. During his talk he published a list of components required to build a system. This same information can be found in the books of Konstantin Meyl and elsewhere on the internet.

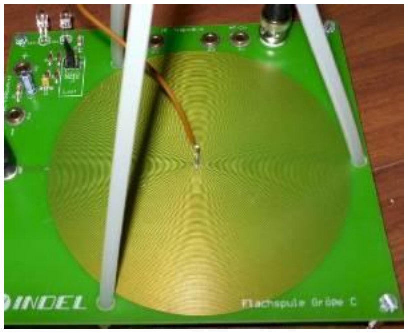

By open sourcing this technology, ordinary individuals could prove to themselves, friends, associates, co-workers, and others that Tesla's technology worked as claimed. Not only does it work, but can be replicated and used for practical applications today. All unique properties of scalar waves could be harnessed and developed to revolutionize our society.

When Steve releases plans, drawings, and other information, probably next week, we will add them to a project page at PESWiki.com. We'll add a link from this page when it is ready.

Earth Changes and the Future

In his books and lectures, Konstantin Meyl has compared superluminal scalar waves to the neutrino, which is also thought by some to travel faster than light. Neutrinos bombard the Earth in all directions, but the majority of them arrive from the central star of our solar system. Due to the fact that approximately half of the neutrinos are absorbed as they pass through the Earth, it is thought they may interact with the core and inner layers of our planet.

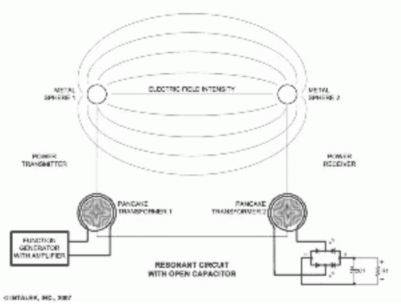

Circuit diagram used by Steve comes from Intalek.com, Figure 3.

Some researchers suggest solar phenomenon can produce variations in neutron emissions which may trigger "earth changes" on our planet. This was depicted in the blockbuster movie, 2012, as the cause for the crust disruptions from the mantle heat-up from increased solar activity. Hence, studying and understanding the nature of scalar waves may increase our chance of survival during chaotic periods in the future. Perhaps by understanding them we could eventually learn how to constructively utilize them to stabilize our planet during periods of solar fluctuations, sun spots, flares, etc.

Perhaps with the use of scalar wave technology, earthquakes, volcanic eruptions, or other natural disasters could be prevented or detected ahead of time. Even a one hour warning before an earthquake, such as the one that recent devastated Japan, could save thousands of lives. Although all kinds of unique technologies are possible with scalar waves, at a minimum we need to understand how neutrinos and scalar waves interact without planet. Not doing so could result in needless death and suffering.

Lets get moving!

To not harness scalar wave technology for the benefit of mankind would be tragic. Our civilization is facing adversity in countless forms. Hunger, disease, pollution, poverty, and war threaten our future and the future of our children. Scalar waves exist, they are useful, and the information to build devices that generate and receive them is available to everyone. To better understand our universe, to ensure the continuation of our civilization, and to honor Nikola Tesla we should do our best to support the open sourcing of this technology.

Videos

On March 30, Steve conveyed the following YouTube links showing his presentation to the local IEEE on March 24.

#1 Slide 1 https://www.youtube.com/watch?v=qs-2s85lO6o
#2 Slide 20 https://www.youtube.com/watch?v=s4kvJpUgGKI
#3 Slide 39 https://www.youtube.com/watch?v=4YXPa73NNsA
#4 Slide 57 https://www.youtube.com/watch?v=LqhRDY56CqU
#5 Slide 81 https://www.youtube.com/watch?v=e1Ht5OVyI5Q
#6 Slide 93 https://www.youtube.com/watch?v=JuehDylES_A

Replication Instruction should be done tomorrow.

Contact

Steve can be reached by email at steve.jackson@ieee.org

# # #

This story is also posted at *BeforeItsNews*.

What You Can Do

Recommend 14 people recommend this. Be the first of your friends.

1. Replicate this design.
2. Pass this on to your friends and favorite news sources.
3. We at PES Network are in a pinch right now. Donations would be greatly appreciated.
4. Subscribe to our newsletter to stay abreast of the latest, greatest developments in the free energy sector.
5. Let professionals in the renewable energy sector know about the promise of this technology.

Comments

Feel free to view/post comments down below. The following came in by other means.

Not Scalar "Wave"

On March 27, 2011, 1:45 pm, Karl Palsness wrote:

You should read about Scalar .. it is a "Scalar FIELD" there should be no such thing as a "Scalar wave"; and there is a "Longitudinal Wave", instead of scalar wave. Scalar waves just don't exist according to the word scalar meaning size. A Tesla coil has a scalar field around it...and emits a longitudinal wave....

A Tesla-era dictionary defined it as:

> **Scalar** -- In physics, a quantity which has magnitude, or magnitude and sign only, without direction, such as density, mass, energy, etc., as distinguished from a vector quantity.

and

> **longitudinal wave**: A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave. Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes.

* * * *

Longitudinal Waves (L.M.D.)s

On April 02, 2011 12:29 AM MST, Dr. Eng. Roberto Handwerker <info {at} deltaavalon.com> wrote:

Dear Sirs,

this mail is a reply to a number of e-mail messages I got from [...], subsequently from Dr. K. Meyl and from Mr. S. Jackson about longitudinal waves ("L.M.D.s").

The matter "L.M.D"s. is not new at all but there is still too much discussion around it from different point of views, i.e. mathematics/physics calculus and electrical devices; this has slowly but definitely led to a total chaos even on the Web, resulting in misunderstanding and disbelief, which doesn't help anybody;

therefore, the best approach for an engineer is to be very straight and careful especially when someone talks about Nikola Tesla and theories about aether and L.M.D.s., still not accepted by the "Academic World" who is as happy to hear such arguments as could be "a sailboat in a forest" so to speak and who brings in turn only useless negative criticism and hectic hysterism instead of helping clarify.

This is why I prefer to proceed "step by step" scientifically replicating some Tesla apparatus and demonstrating his experiments, which is of course not easy because he was the Mozart of the art.

I will present a book about the above and will very soon attend a Meeting in Italy to talk about L.M.D.s and show some working devices to public: it is always a pleasure to demonstrate FACTS and WORKING DEVICES rather than only showing some slides and theoretically discuss matters as "Lorenz gauge vs. Coulomb gauge", "quaternions vs. vectors", "T.E.M. vs. L.M.D." etc. etc. Only facts count, observation of Nature and research: math must consequently follow, as always.

It is therefore more useful instead to bring something new to L.M.D.s tech and to the understanding of Tesla, like for instance my "Tesla power wireless" system with the "small car-demo" shown in my following vids on You Tube (& others under "HorizonDelta"): https://www.youtube.com/watch?v=0BGFF4ltEv and https://www.youtube.com/watch?v=dvZH3RCdeVw

also a tech e-book "Longitudinal dielectric waves in a Tesla coil and quaternionic Maxwell's equations" can be found on my Website for download:

https://www.deltaavalon.com/immagini/L.M.D. waves in a Tesla Coil revised.pdf

I hope the above helped positive discussion and knowledge.

Regards

Dr. Eng. Roberto Handwerker
DELTA Ingegneria

* * * *

The Laplace Operator

On March 31, 2011 2:06 PM MST, Konstantin Meyl wrote:

Subject: Re: About Scalar Field and an Italian Book

There is one comment, making clear, where the problem is localized:

Karl Palsness wrote:

> "You should read about Scalar .. it is a "Scalar FIELD" there should be no such thing as a "Scalar wave"; and there is a "Longitudinal Wave", instead of scalar wave. Scalar waves just don't exist according to the word scalar meaning size. A Tesla coil has a scalar field around it...and emits a longitudinal wave..."

If a scientist is using an expression, he first has to define, what the meaning is. This is what I do in my books, but in America, nobody seems to read my writings. I explain that the scalar wave is nearly the opposite of a scalar field. The expression is more than 200 years old and in mathematics exactly explained by the field equation of Laplace.

The Laplace Operator (describing the distribution of a wave in space) may be separated in 2 parts: grad div - curl curl. The second part is explaining the EM wave, a transverse wave, while the first part gives a scalar wave: The div applied on a vector is scalar! This is why scalar particles or vortices as part of the wave equation are propagating as a scalar wave. But they are propagating in the direction of a field vector and by this the wave is longitudinal and has a direction. In mathematics it is known, that the grad applied on a scalar is a vector again. This is why the scalar wave is propagating in one direction, not in a scalar way, as Karl Palsness wants the scalar wave to do. So he has to read the literature first before he is spreading nonsense in the world wide web.

My opinion about the education in the states is not the best. Lets help to improve it.

* * * *

Other Relevant Coverage

- Featured / OS: Wireless Transmission > Scalar > Tesla > Coils > Jackson >
  Tesla Coils for Dummies - Are you interested in building a Tesla Coil to produce stunning effects that violate the "established" laws of physics, but are somewhat less than technically savvy than a rocket scientist? Then you are in luck, because this article is for you! (*PESN*; April 29, 2011)

- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State > Tesla > Coils > Jackson >
  Jackson's Wireless Aetheric Power Transmission Open Source Project - To facilitate the replication of these plans, as well as to characterize, optimize, improve, and come up with the zillion applications that can come from it, including: 1) harnessing aetheric energy, 2) superluminal communication, 3) wireless transmission of power through any barrier, 4) anti-gravity capabilities, and 5) creating defensive shields to make conventional war obsolete. (*PESWiki*; April 21, 2011)

  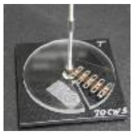

- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State > Tesla > Coils > Jackson >
  Tesla Coils: Unleash the Aether - Nikola Tesla's most significant contribution was not AC power, radio, or the induction motor, but what we call the Tesla Coil - a tool which allows for the power of the aether to be unleashed and harnessed. Now being launched through an open source project. (*PESN* and *BeforeItsNews*; April 19, 2011)

- Featured / OS: Wireless Transmission / Electromagnetic > Scalar > Solid State >
  Tesla's Scalar Field Still Beaming On! - The superluminal (faster than light) scalar or longitudinal waves Nikola Tesla used to magnify and wirelessly transmit power are not just a thing of the past. IEEE engineer, Steve Jackson, discusses and demonstrates how they can be utilized today, and he is open sourcing it here! (*PESN*; March 26, 2011)

See also

Resources at *PESWiki.com*

- **Directory:Konstantin Meyl**
- **Directory:Wireless_Transmission_of_Electricity**
- **PowerPedia:Wireless_transmission_of_electricity**
- **PowerPedia:Scalar_field_theory**
- **Directory:Solid_State_Generators**
- **PowerPedia:Solid_State_Generator**
- **Directory:Electromagnetic**
- **News:Electromagnetic**
- **More Stories for Hank Mills**

- **PESN (Pure Energy Systems News)** - Feature stories on cutting-edge, clean energy technology.
- **Free Energy News (.com)** - Daily cutting-edge, clean energy technology news from around the world
- **PESWiki Latest** - Newest feature pages in the publicly-editable energy directory.
- **This Week in Free Energy™** - Weekly live review by PESN, each Thursday 7 pm Mountain time
- **Free Energy Now (.net)** - in-depth interviews

AdChoices
PowerPoint Free
Tesla

Page composed by Sterling D. Allan March 26, 2011
Last updated September 04, 2012

NOTE: We are presently moderating all comments, to keep out the crass negativity, flaming, and other inappropriate forms of criticism. We welcome constructive criticism tactfully stated. Frequent offenders will be blacklisted.

"It is harder to crack a prejudice than an atom." // "I'd rather be an optimist and a fool than a pessimist and right." -- Albert Einstein

0 Comments Pesn 1 Login

Recommend Share Sort by Newest

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Be the first to comment.

Subscribe Add Disqus to your siteAdd DisqusAdd Privacy DISQUS

The "Down and Dirty" Guides to Making Biodiesel at Home
Making
BioDiesel Books
Save Time, Save Money... Help The Environment!

ADVISORY: With any technology, you take a high risk to invest significant time or money unless (1) independent testing has thoroughly corroborated the technology, (2) the group involved has intellectual rights to the technology, and (3) the group has the ability to make a success of the endeavor.

All truth passes through three stages:
First, it is ridiculed;
Second, it is violently opposed; and
Third, it is accepted as self-evident.
-- *Arthur Schopenhauer (1788-1860)*

"When you're one step ahead of the crowd you're a genius. When you're two steps ahead, you're a crackpot."
-- *Rabbi Shlomo Riskin, (Feb. 1998)*

Submit • Privacy • About • Contact

***PESWiki Departments*:**
Latest • News • XML Feed • Directory • Congress • Top 5 • Open Sourcing • PowerPedia

PESN.com

Copyright © 2002-2014, PES Network Inc.

The Tom Bearden Website

Subject: RE: Missing Reference Slide
Date: Mon, 9 Jul 2001 11:58:04 -0500

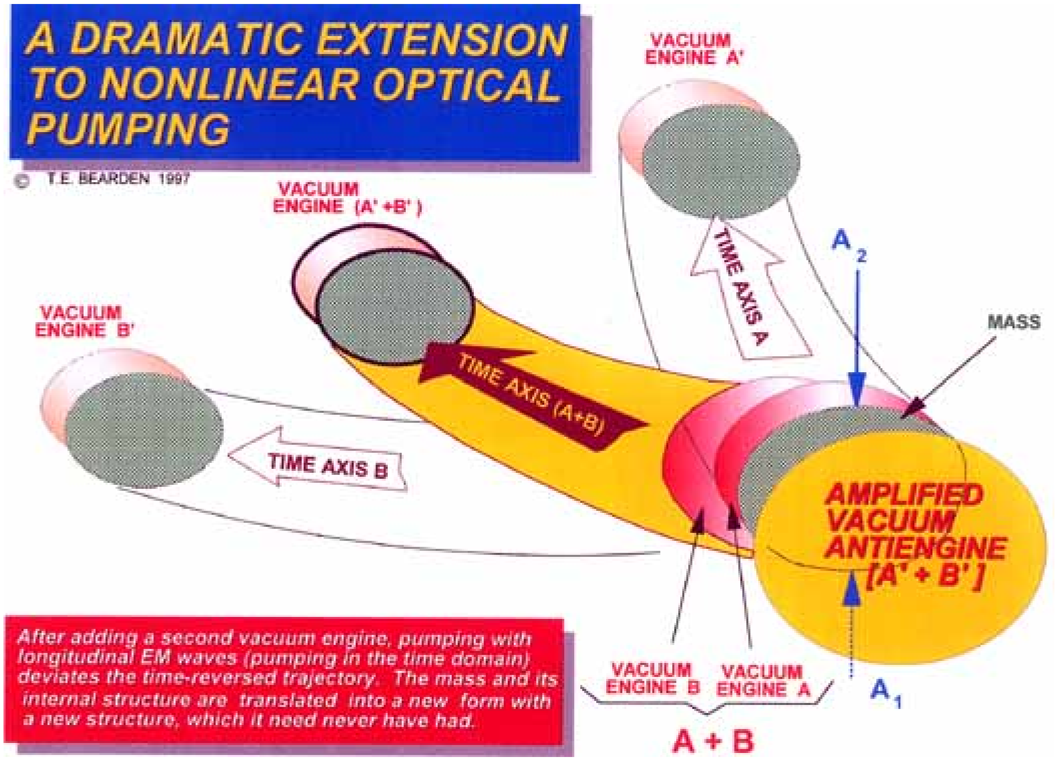

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Yes, that is the slide. It works this way: When you time-reverse a cell, you effectively create a major anti-engine for the cell's exact "engine" (exact set of spacetime curvatures at all levels). Because of the precise one-to-one correspondence, this antiengine then "backs the cell up" along the original time-path it took in its physical functioning and changes, back to a precise earlier condition where the disease or disorder was not present. For a cancer cell, e.g., this backs the cell up past where it was originally "promoted" from a stressed or damaged cell to a tumor cell. That makes it into a "damaged cell". The antiengine continues to act, backing it up further into the state before the damage or disorder occurred. That then constitutes "healing" of that cell in a rigorous sense.

The cloning methods etc. used, do violence to the cell in the process. At the cellular level, damaging it or any part of it in the process changes the resident spacetime curvature engine set, to include the damage. You then "clone" a cell with an imperfect resident engine. The cloned cell then interacts with a large variety of phenomenology, continually altering various subcomponents of its resident engine by taking on additional subengines or changing some of its former subengines. Somewhere later, that festering "delta" in the resident engine -- after sufficient interactions with other environmentally-induced engines, will undergo interactions that "add" two deltas so something detrimental to the organism. At that point, physical damage to the cell erupts in "anomalous" fashion. Actually it erupts by perfectly normal fashion, once the internal engines and their interactions are understood.

This would appear to pose a problem anytime you *artificially* change the subengines of an organism, as in the case of genetically modified food. The small physical damage associated with splicing the genetics has associated with it the injection of such an artificial engine. The interactions of future environmental engines with the spliced genetic organism or plant then remains to be seen. Not all interactions would appear to happen in the limited test time usually allotted for testing the "effects". Anyway, that was the intention.

E.g., Pautrizel demonstrated that immature rats, having immature immune systems, would be restored by the Priore process back to the same immature immune system. The Priore process did not destroy the parasites themselves, but depended upon restoring the immune system back to its first ability to recognize the pathogen and go after it. Priore did not directly "kill" anything! In that case, the pathogens simply re-infected the immature rats, because their immature immune systems could not adequately protect them. It would be the intention of the research, in later generation equipment, to be able to design a specific "extra engine" and introduce it to the pumping, in addition to the "resident engine" in the subject to be treated. That "extra input engine" would accordingly generate a specific "extra antiengine" component in the total antiengine produced. When sufficient research and development maturity has been reached to be able to design such desired "extra engines" to cause the exact, desired extra antiengine to be produced, then the "extra antiengine" technique can be used, e.g., to time-reverse Pautrizel's immature infected rat back to an infected rat with a now-matured immune system. In that case, the rat's stronger immune system would make short work of the pathogen, just as analogously infected mature rats did.

To correct birth defects, etc., obviously one would have to change the actual time-path backwards, to a "state" that the body did not have physically, but which its genetics had before the damage. I foresee that eventually, after some years of development, the process would mature until it could regrow severed limbs, or correct dwarfism, etc. Of course that would be fourth or fifth generation, but it is definitely possible, given sufficient development.

Cheers,

Tom

AD 763210

QUITON/PERCEPTRON PHYSICS: A THEORY OF EXISTENCE, PERCEPTION, AND PHYSICAL PHENOMENA

by

Thomas E. Bearden

March 1973

Approved for public release; distribution unlimited.

U.S. ARMY MATERIEL COMMAND

THE SAM-D PROJECT
REDSTONE ARSENAL, ALABAMA

SURFACE TO
AIR
MISSILE
DEVELOPMENT

AMCPM-MD FORM 1, 1 MAR 71

CONTENTS

Page

I. INTRODUCTION 1

II. EXISTENCE, PERCEPTION, OBSERVATION, AND PHYSICAL PHENOMENA 2

III. MASS AND SPACE 11

IV. IMPACT ON PHYSICS 16

A. Michelson-Morley Experiment 16

B. Gravitational Attraction 17

C. Newton's Third Law of Motion 19

D. Newton's Second Law of Motion 20

E. Newton's First Law of Motion 22

F. Identity of Length and Time 23

V. ELEMENTARY PHILOSOPHICAL IMPLICATIONS 25

A. Existence, Thing, and One 25

B. Reality, Ultimate Reality, Being, and Ultimate Being 27

VI. CAUSALITY AND ITS MODIFICATION 28

VII. CLOSING REMARKS 35

ABSTRACT

A fundamental answer is attempted to the questions of being, time, space, existence, perception, and physical phenomena. Mass, empty space, reality, gravitational attraction, continuity, discontinuity, and relativity are among the "things" and "nonthings" for which relationships and explanations are provided.

Perception is defined as detection by a mass of change to itself and as a differentiating process. Perceived physical phenomena are revealed to be first derivatives of ultimate reality, and outputs of the physical perceiving device call the "perceptron." Using the operation of the physical perception process on action, the units of which are energy times time, an entirely different view is generated of both reality and the Heisenberg uncertainty principle. When superposition is imposed as a condition, the Heisenberg uncertainty principle is revealed to be a quantized statement. A concept of the absolute value of a dimensional molecule is advanced.

A reality model is constructed of an action continuum and a nonaction continuum separated by an operational threshold whose magnitude is given by an uncertainty principle. Perception is modeled as a process which switches atomic quanta of action and nonaction back and forth across the threshold. Mass is explained as perceptron switch activity rate, and a mass definition equation is generated. One kilogram mass is shown to represent 17.053 x 10^{50} perceptron switches per second. From the model, Newton's laws of motion and gravity can be generated directly from the concept of the interaction of mass with space itself, space being taken as a massless fluid, or ether, composed of tiny nonaction particles called "quitons." Einstein's postulates of special relativity are also shown to follow from the model, and a statement of the equivalence principle is shown to be consistent with the model.

The concept of "lineception" as a causal chain of perceptron operations is advanced. "Inception" is developed as a noncausal modulation of lineception and as a distinguishing feature of a living system. The possibility is raised that inception could provide an explanation of the noncausal, statistical, ultramicroscopic universe and yet require large ensembles of ultramicroscopic events to conform to causality. Inception is shown to lead to karma. Perception is shown to be causal and to generate causality itself.

Some elementary philosophical implications of quiton/perceptron theory are pointed out and briefly discussed.

I. INTRODUCTION

The unsolved problems of physics stem from the unclear nature of the physical perception of change, and their resolution lies in the analysis of perception as a physical process. The unsolved problems of metaphysics stem from the same source, and can be resolved by the same perception analysis.

Specifically, a physical detection system can detect changes to itself and nothing else. These detected internal changes are thus what a detector "sees" or "observes" or "perceives" as changes to its external environment, i. e., as its physical phenomena. Thus a mass, being itself a physical detecting system, must detect only changes to itself, and these perceived changes to itself constitute its observed physical phenomena. For that reason, these perceived changes are obviously entirely relative to (i. e., are part of) the perceiver (observer); more precisely, they are entirely relative to and part of the perceiver's mass, which is the detector that is doing the perceiving.

Since only changes are perceived by a mass, then mass's perception must be a differentiating process. Ergo, perceived physical phenomena are first derivatives of a higher, or more fundamental, reality. The nature of that higher reality is by definition unperceivable (perception differentiates or fragments it), and it involves the quantity "action."[1] Action itself is not perceivable; change of action is perceivable since perception differentiates action. If superposition is placed on the Heisenberg uncertainty principle as a required condition, then the uncertainty principle contradicts itself for all except integral or zero multiples of a quantum of action. Correction of the uncertainty principle by imposing superposition as a condition provides a statement of the basic operation of perception -- the detection by a mass of change to itself.

All mental perception of a human being regarding physical phenomena is received (i. e., is inputted to the mind) from a physical sensory apparatus whose primary ingredient is mass. Thus we may describe the perceptive mind as consisting of sensory outputs of mass perceptions; that is, the input to mental perception must be the output of mass perception, and this interface within the mind may be referred to as the "perceptive mind."

From the above foundation, a comprehensive theory of perception can be constructed, and a most unusual model of "reality" emerges. A variety of unsolved questions are then resolved by the model. The resolutions include, among other things, 1) derivation of the postulates of relativity; 2) definition of the nature of time and space; 3) the explanation of gravity; 4) the generation of causality itself; 5) resolution of the wave theory of light with the quantum theory of light; 6) the explanation of why ultramicroscopic phenomena are statistical yet large ensembles of ultramicroscopic phenomena are causal; and 7) validation of Mach's principle and the equivalence principle. A totally new and precise definition of mass and of being itself are also two unexpected results. The problems of metaphysics are also answered: metaphysics is elevated to an exact science without any tinge of dogma, and physics and metaphysics are united.

---

[1] Planck's constant, Heisenberg's uncertainty principle, the principle of least action, the energy of a photon, etc., are derived from action and perception's differentiation of action.

Next Page

II. EXISTENCE, PERCEPTION, OBSERVATION AND PHYSICAL PHENOMENA

We first define perception[1] as detection by a mass of change[2] to itself, and state that it is an operation performed by that mass. The physical apparatus that accomplishes this perception is termed a "perceptron," which thus may be an electron, an atom, or the sensory apparatus of a living being.

A change detector does not detect the state of its input, but rather detects changes in that state. So in this sense a change detector can be regarded as a state differentiator. Therefore a perceptron differentiates, i. e., it perceives changes in its input and generates outputs in accordance. When its input is constant it will have no output. Thus zero output of a perceptron does not necessarily mean zero input.

What is to be fed into the perceptron's input section exists; i. e., existence (ultimate reality)[3] is defined as what is to be fed into the perceptron's input. The output of the perceptron is that which is perceived, or what we ordinarily call "physical phenomena."[4] Since the perceptron differentiates its input state, then physical phenomena are first derivatives of ultimate reality, which precisely accounts for the sharp separation between observable and nonobservable, i. e., between macroscopic and microscopic.

Any change in a perceptron's input section must involve both time change, Δt, and energy change, ΔE. So we may assume their product, $\Delta E \Delta t$, has always occurred in a perceptron's input whenever it produces an output. The output will be a derivative of the input, e. g.,

$$\frac{\Delta E \Delta t}{\Delta t} = \Delta E \tag{1}$$

and

$$\frac{\Delta E \Delta t}{\Delta E} = \Delta t , \tag{2}$$

where in each case the constant of proportionality has been ignored.

---

[1] We differentiate between perception and realization. Perception is entirely a physical process, and causal. Realization is a mental process and is not restricted to causal relationships.

[2] Reflecting, one notes that mass can only change in a nonmass manner. It cannot "change" without involving space and time, both of which are "nonmass." Therefore a mass change must be composed of both mass and nonmass.

[3] "Ultimate" in that it is higher or more fundamental than perceived reality.

[4] Changes to mass and only changes to mass constitute physical phenomena. Changes to one mass constitute "its" perceived physical phenomena, or its physical universe.

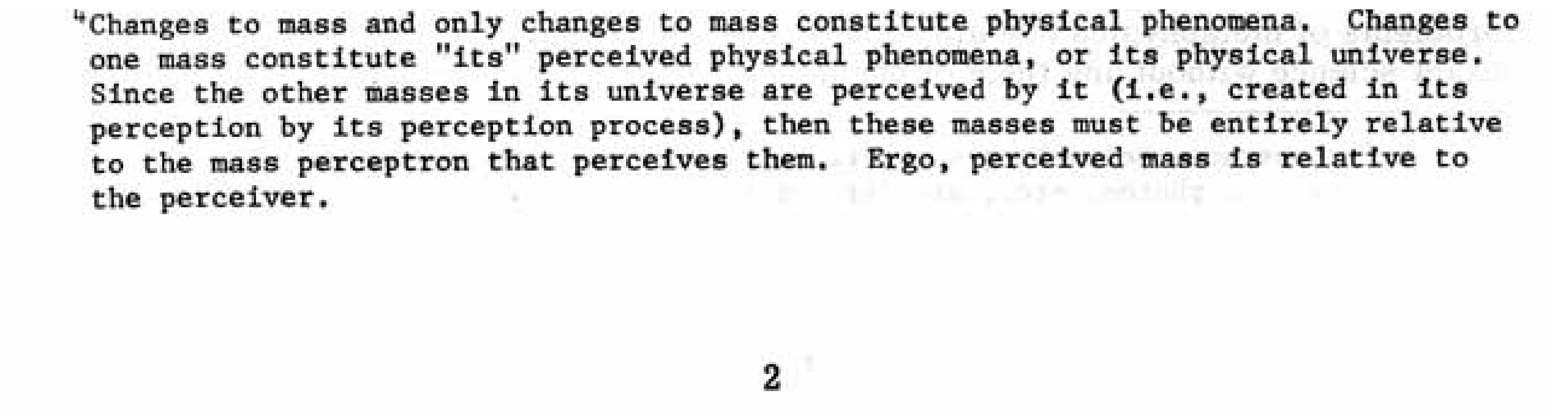

"Changes to mass and only changes to mass constitute physical phenomena. Changes to one mass constitute "its" perceived physical phenomena, or its physical universe. Since the other masses in its universe are perceived by it (i.e., created in its perception by its perception process), then these masses must be entirely relative to the mass perceptron that perceives them. Ergo, perceived mass is relative to the perceiver.

Next Page

Any detector has an operational input threshold, i.e., a minimum discernible signal. The Heisenberg uncertainty principle states the requirement for such energy-time changes:

$$\Delta E \Delta t \geqslant \hbar/2 \, . \tag{3}$$

Its smallest value, or

$$\Delta E \Delta t = \hbar/2 \, , \tag{4}$$

is then the input threshold for a physical perceptron to operate and produce an output.[1] Letting $k = \hbar/2$, we can write this simply as

$$\Delta E \Delta t = k \, . \tag{5}$$

Since existence has been defined as that which is fed into the perceptron's input, then that existence exists whether or not the perceptron operates, and the case[2] where

$$\Delta E \Delta t < k \, , \tag{6}$$

also exists (i.e., can be present in the input of a perceptron) but cannot be perceived or detected (i.e., the perceptron will not operate and produce an output).

At this point we diverge to remark that the product of energy and time is a quantity called "action," and the units of action are

$$A \equiv \frac{ML^2}{T} \tag{7}$$

Thus the input state to a perceptron is composed of an action continuum,[3] and the perceptron differentiates action quanta to produce physical phenomena.

---

[1] We acknowledge additional uncertainty statements but for simplicity will use only one in this paper.

[2] Equation (6) exists and it is proper to write it down and search out its ramifications. Equation (3) is a limiting condition which applies only to mass. But since our physical experience must involve nonmass (e.g., photons, space, time, being), then nonmass considerations must be examined. A good beginning is to consider the entire field of nonmass possibility, i.e., that field of nonmatter changes defined by (6), which includes the entire subquantum domain. So we legitimately examine two conditions: Eq. (3) must define (or apply as a condition to) the mass phenomena (change, quantized) system, and (6) must define (or apply to) the nonmass phenomena (subquantum) system.

[3] We exactly define continuity as the failure to perceive difference (change), and

phenomena (subquantum) system.

[3] We exactly define continuity as the failure to perceive difference (change), and discontinuity as the perception of difference (change). The difference between continuous and discontinuous is entirely one of perception. Thus there are two kinds of continuity: absence of perceptron operation at all, and identically repeated perceptron operations. The first kind exists outside of perception while the second kind exists in perception-to-perception. These are two powerful definitions, which among other things enable one to understand such concepts as limiting process, derivative, point, line, and plane. Mathematics itself is the game of perception, nothing more and nothing less.

Next Page

When a large quantity of action is to be fed into a perfect perceptron, the input process cannot occur instantaneously; a Δt must be outputted (created)[1] by the perceptron for each operation. Thus perception processes steps or quanta of exactly k size, and exactly k size only. Inputs of less than k size are simply collected[2] until the aggregate is equal to k, then differentiated to produce an output. We must therefore rewrite the Heisenberg uncertainty principle as

$$\Delta A = nk \quad , \; n = \pm 0, 1, 2, \ldots \quad , \tag{8}$$

where the plus and minus signs are necessary since action only has to change about some constant level in the perceptron input, and this change may be either additive or reductive (this quantum statement is actually the basis for quantum physics).[3] Thus, from the viewpoint of a perceptron, its output universe of physical phenomena is constructed by successive granular differentiation of unit operations, each of size k, but it detects no time lapse (creates no Δt) between operations.

When differentiated by a perceptron, an action atom may be regarded as being "chopped up" (the word "differentiated" literally means separated or torn apart):

$$\frac{\Delta E \Delta t}{n} = \frac{k}{n} \quad \overline{A} , \qquad n = 2, 3, 4, \ldots \quad , \tag{9}$$

where the bar over the A indicates nonaction, or action atoms which have been broken into bits.

Thus we now conceive of an action continuum of action atoms and a nonaction continuum[4] of bits of action atoms. The two continua coexist and the perceptron

---

[1] Or "extracted from an action quantum $\Delta E \Delta t$," which means exactly the same as "created." Einstein's treatment of simultaneity showed that it was operational, but if simultaneity itself is operationally created then both time and length must also be operationally created.

[2] This is not a simple discontinuous operation, since no time flow exists (Δt is not being outputted) during this collection process. We have to think of the process discontinuously, but the stepwise process is actually continuous (in time), or discontinuous, or both.

[3] Since ΔE may be positive or negative, we may state the uncertainty principle as $|\Delta E \Delta t| \geqslant k$, or as $\Delta A = nk$, $|n| \geqslant 1$. Consider the two cases $\Delta A > 0$, $n = 1$, and $\Delta A < 0$, $-1 < n < -2$. Superimposing them would give $|\Delta A| < k$, which violates the uncertainty principle. Thus, assuming superposition must hold, the uncertainty principle can be made to violate itself for all except integral values of n, and it follows that the Heisenberg uncertainty principle must be written as a quantized statement.

[4] The separation into two continua is artificial, but is necessary to allow us to perceive perceptron operation with our brain perceptron mind. There is absolutely

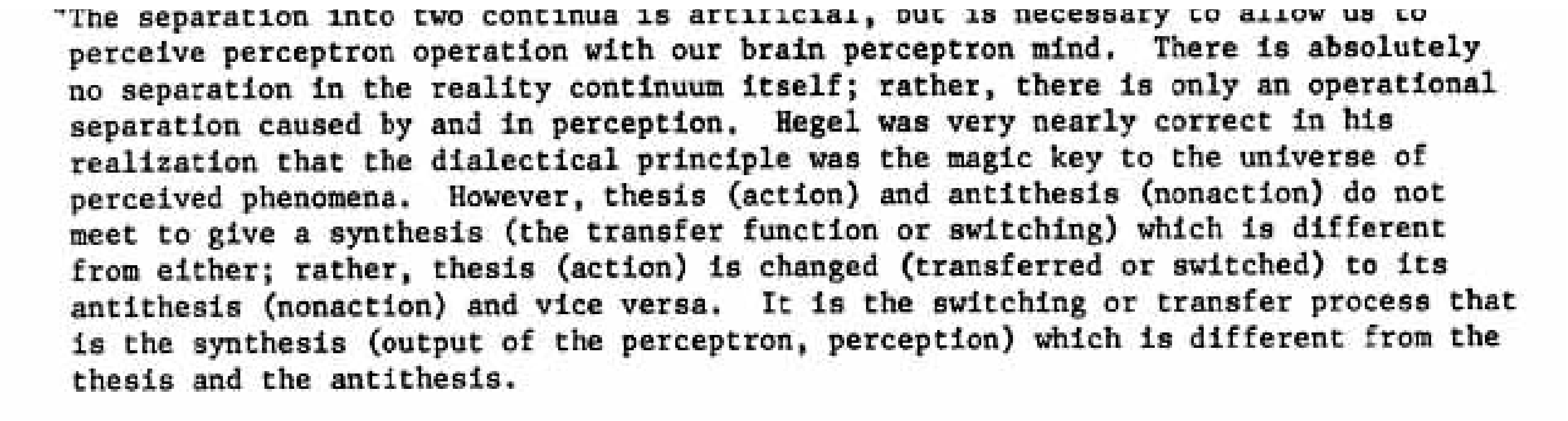

"The separation into two continua is artificial, but is necessary to allow us to perceive perceptron operation with our brain perceptron mind. There is absolutely no separation in the reality continuum itself; rather, there is only an operational separation caused by and in perception. Hegel was very nearly correct in his realization that the dialectical principle was the magic key to the universe of perceived phenomena. However, thesis (action) and antithesis (nonaction) do not meet to give a synthesis (the transfer function or switching) which is different from either; rather, thesis (action) is changed (transferred or switched) to its antithesis (nonaction) and vice versa. It is the switching or transfer process that is the synthesis (output of the perceptron, perception) which is different from the thesis and the antithesis.

Next Page

operates, or moves, in them. Only action quanta of exactly k size can be processed by the operating perceptron to give outputs which we know as physical phenomena. Such phenomena (events) consist of aggregates of extremely large numbers of perceptron operations. Action atoms will be referred to simply as "action atoms." The very smallest bits or fragments[1] of action atoms will be referred to as "quitons" and the symbol "q" used for one quiton. For convenience, we shall regard the "magnitude"[2] of q to be very much less than k, or

$$q \ll k \,. \tag{10}$$

The perceptron's process of collecting quitons to form action atoms will be termed "fusion," which is an additive change to the constant action continuum. "Fission" will be the term for the perceptron's process of breaking up action atoms into aggregates of quitons, where the size of each aggregate is less than k; fission is a reductive change to the constant action continuum. The term "switching" will be used to refer to either fission or fusion, or both.[3]

Fission and fusion are merely transfers or switches across the k-threshold separating the action and nonaction continua, and may be represented by either of the diagrams shown in Fig. 1.

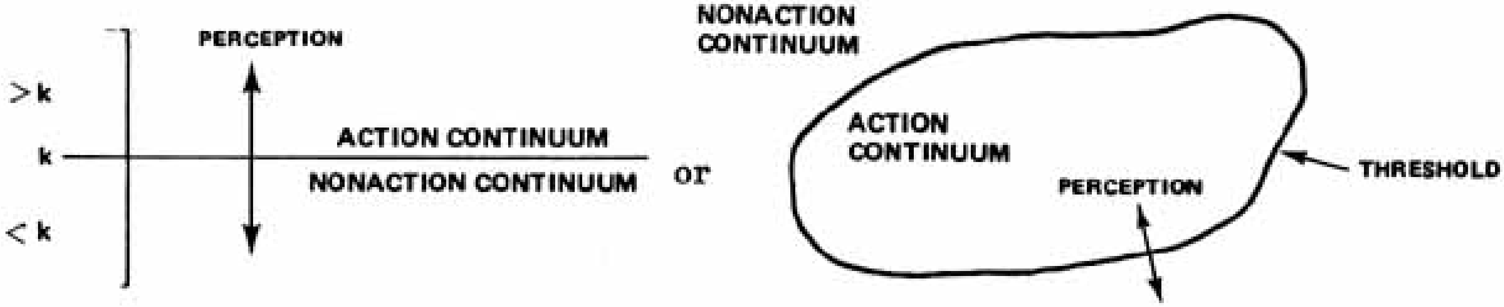

Fig. 1. Continua of the Universe

---

[1] These are actually nonthings and not bits or pieces at all, but since we can only perceive or think of things, we will model them as things. The difficulty is in the nature of thought, not in the nature of reality.

[2] The perceivable magnitude of a quiton is zero, i.e., it is not perceivable.

[3] And, in fact, switching can also refer to the transfer of Δt into ΔL, i.e., the creation of space from time or vice versa. Separation (relation) is composed of ΔL and Δt, and we know from relativity that when relativistic (perceived) ΔL is changed, then relativistic (perceived) Δt is also changed. When time dilates, length contracts; and when time shrinks, length expands. Therefore the statement that length and time are intertransferrable is valid: one kind of separation can be transferred into another kind of separation just as one kind of energy can be changed into another kind of energy. Since this transfer itself is an operation, then there is a rate (i.e., a constant) of transfer. This constant is actually a parameter, since by special relativity it can vary as a function of the mass changes themselves. And since we are dealing with mass changes (perception), then the rate of transfer

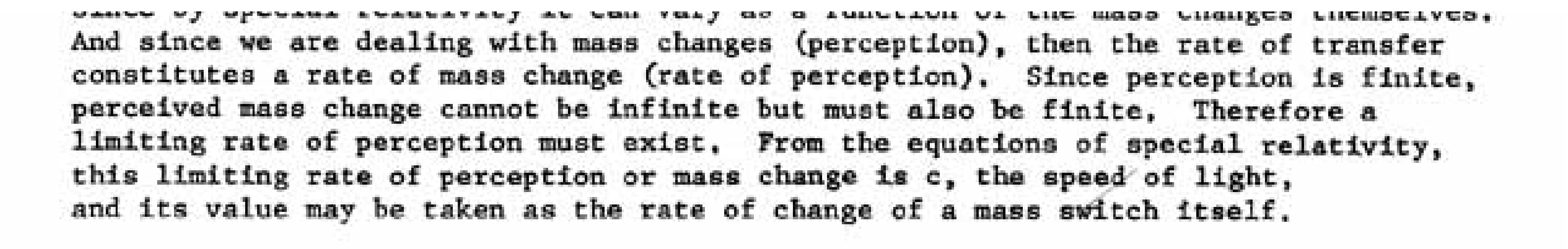

since by special relativity it can vary as a function of the mass changes themselves. And since we are dealing with mass changes (perception), then the rate of transfer constitutes a rate of mass change (rate of perception). Since perception is finite, perceived mass change cannot be infinite but must also be finite. Therefore a limiting rate of perception must exist. From the equations of special relativity, this limiting rate of perception or mass change is c, the speed of light, and its value may be taken as the rate of change of a mass switch itself.

Next Page

We now examine the expression for energy of a photon:

$$E = hf \tag{11}$$

Considering fission of an action atom as a chopping of that atom, the process can be expressed by:[1]

$$\begin{bmatrix}\text{Minimum}\\ \text{time rate of}\\ \text{change of action}\end{bmatrix} = \begin{bmatrix}\text{Energy}\\ \text{Transferred}\end{bmatrix} \simeq \begin{bmatrix}\text{atom of}\\ \text{action}\end{bmatrix}\begin{bmatrix}\text{cuts or chops}\\ \text{per second}\end{bmatrix}, \tag{12}$$

or

$$\frac{\Delta A}{\Delta t} = \Delta E = C_1 kf , \tag{13}$$

where C_1 is a constant of proportionality. According to Eq. (1), the energy perceived to be expended (transferred to quiton disturbances) in fissioning of an action atom ought to be proportional to the speed of the chopping action. Neglecting constants of proportionality, Eq. (13) is simply[2]

$$\Delta E = hf , \tag{14}$$

where the Δ symbol is deliberately added to show that it is a transfer process. The equation works in reverse for fusion of quitons into action atoms, i. e., the energy ΔE transferred to each action atom assembled is proportional to its rate of assembly.

We can now represent a perceptron as shown in Fig. 2, which corresponds precisely to the two-pronged arrows shown in the diagrams of Fig. 1. A perceptron has two channels, fission and fusion, in which switching transfers between action and nonaction are involved.

By Eq. (2), one of the outputs of a perceptron is Δt. Since perception creates its own time interval, it only has its kind of time, i. e., positive time, to deal with. Thus only positive time is perceived, and what we think of as positive time flow is merely the successive Δt's created by perception.[3]

---

[1] Note how admirably the view of tearing apart or differentiating the action atom describes the result. The smaller the Δt torn from $(\Delta E\ \Delta t)$ (or the quicker the separation), the larger the ΔE left (produced), i.e., the more energetic the separation. The atom of action is a quantum of action and one integral piece; the ΔE and Δt, however, are not necessarily of individually fixed size.

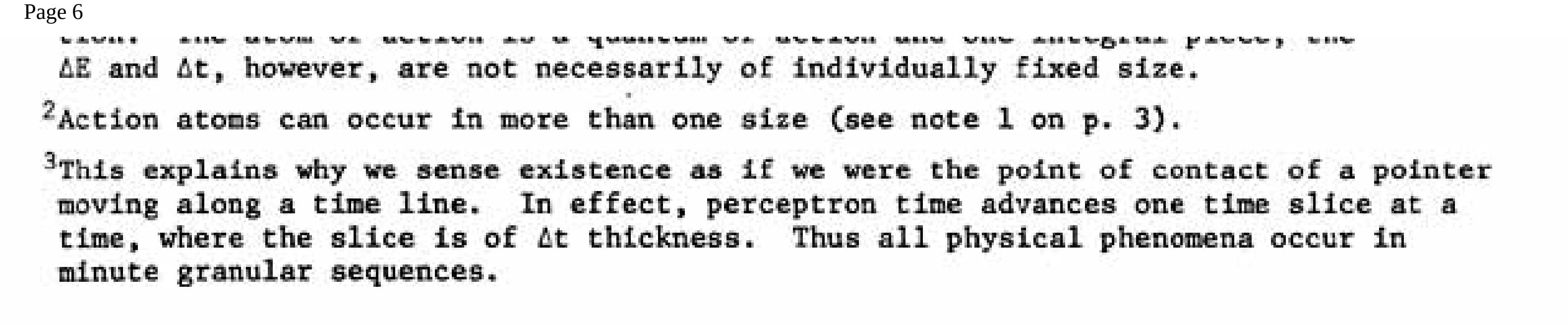

tion. The atom of action is a quantum of action and one integral piece, the ΔE and Δt, however, are not necessarily of individually fixed size.

[2] Action atoms can occur in more than one size (see note 1 on p. 3).

[3] This explains why we sense existence as if we were the point of contact of a pointer moving along a time line. In effect, perceptron time advances one time slice at a time, where the slice is of Δt thickness. Thus all physical phenomena occur in minute granular sequences.

Next Page

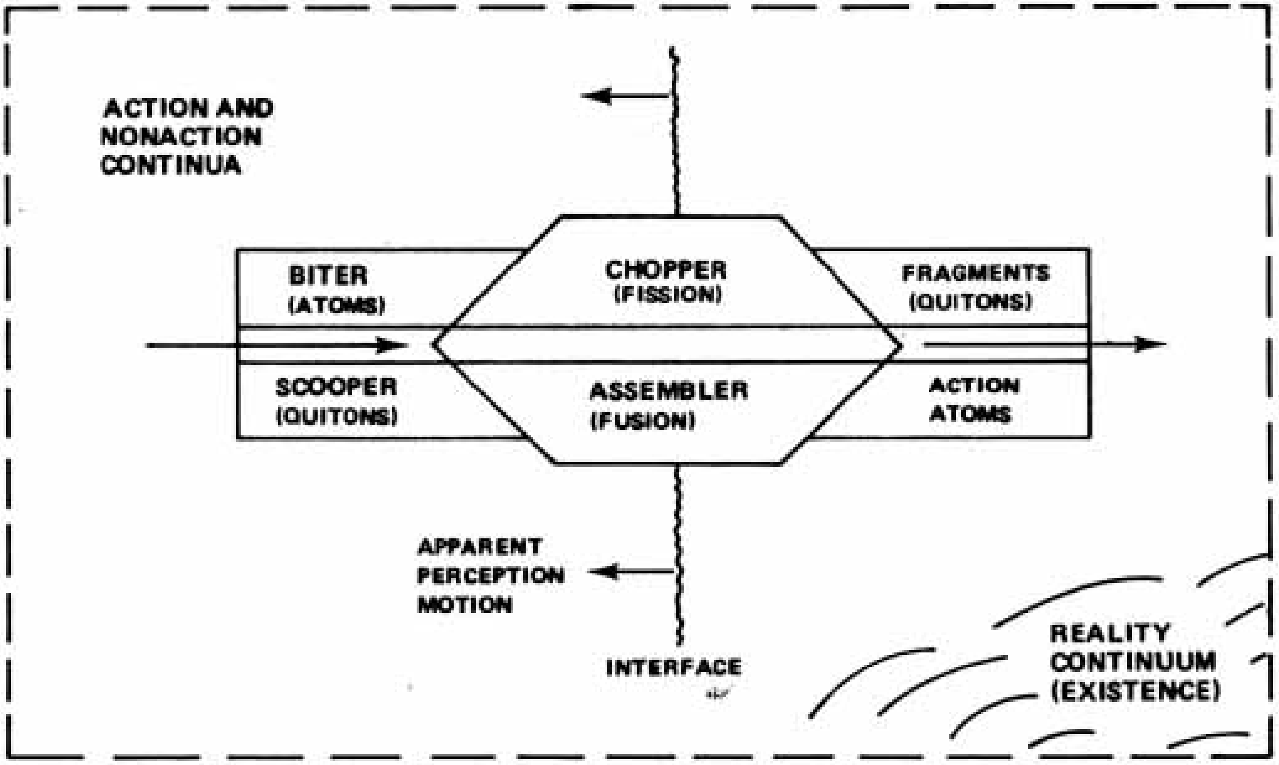
ACTION AND
NONACTION
CONTINUA
BITER
(ATOMS)
CHOPPER
(FISSION)
FRAGMENTS
(QUITONS)
SCOOPER
(QUITONS)
ASSEMBLER
(FUSION)
ACTION
ATOMS
APPARENT
PERCEPTION
MOTION
INTERFACE
REALITY
CONTINUUM
(EXISTENCE)

Fig. 2. The Fundamental Perceptron Switch

The existence of a negative time flow can be postulated by a simple symmetry analogy, but this postulated negative time, by definition, cannot be a perceptron output. We use the term "reflection" to denote a postulated operation in negative time[1] symmetrical to perception which operates only in positive time. Reflection is essential to the explanation of electric charge and charge effects,[2] but these will not be discussed here. Figures 3, 4, and 5 summarize the concepts of reflection, perception, and positive and negative time.

We now introduce the concept of absolute value of a dimensional molecule.[3] To illustrate, if we know that F = Ma, we will state that

$$|F| \equiv |Ma| \equiv |M| \times |a| = M\,|a| \quad , \qquad (15)$$

where the last expression is true because of the nondivisibility of the unit ΔL and

---

[1] Modern theorists such as Feynman, Wheeler, and Dirac of necessity incorporate negative time operation into their theories involving action-at-a-distance.

[2] The quiton/perceptron theory can and does explain what "charge" is, why like charges repel and unlike charges attract, etc. It predicts the force of attraction or repulsion to be proportional to Q_1Q_2/R^2.

[3] Although perceptron output bits may vary in magnitude from operation to operation,

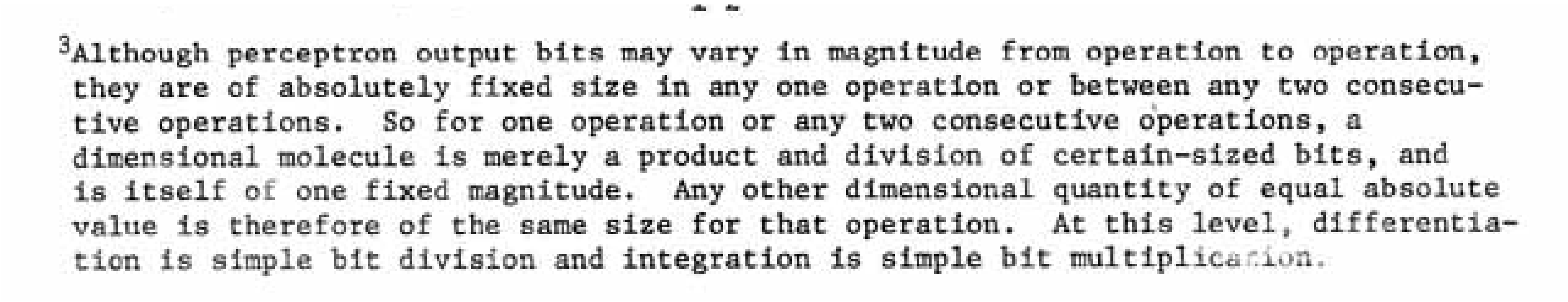

[3] Although perceptron output bits may vary in magnitude from operation to operation, they are of absolutely fixed size in any one operation or between any two consecutive operations. So for one operation or any two consecutive operations, a dimensional molecule is merely a product and division of certain-sized bits, and is itself of one fixed magnitude. Any other dimensional quantity of equal absolute value is therefore of the same size for that operation. At this level, differentiation is simple bit division and integration is simple bit multiplication.

Next Page

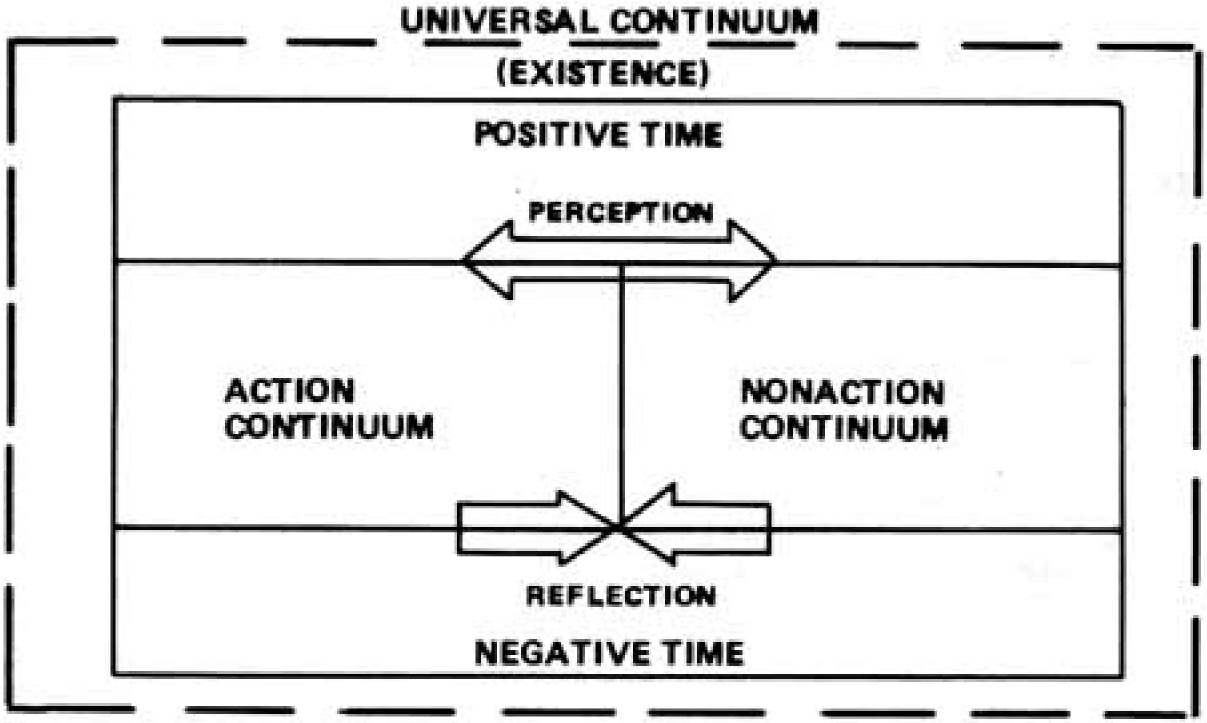
UNIVERSAL CONTINUUM
(EXISTENCE)
POSITIVE TIME
PERCEPTION
ACTION CONTINUUM
NONACTION CONTINUUM
REFLECTION
NEGATIVE TIME

Fig. 3. The Continua (Complete Model)

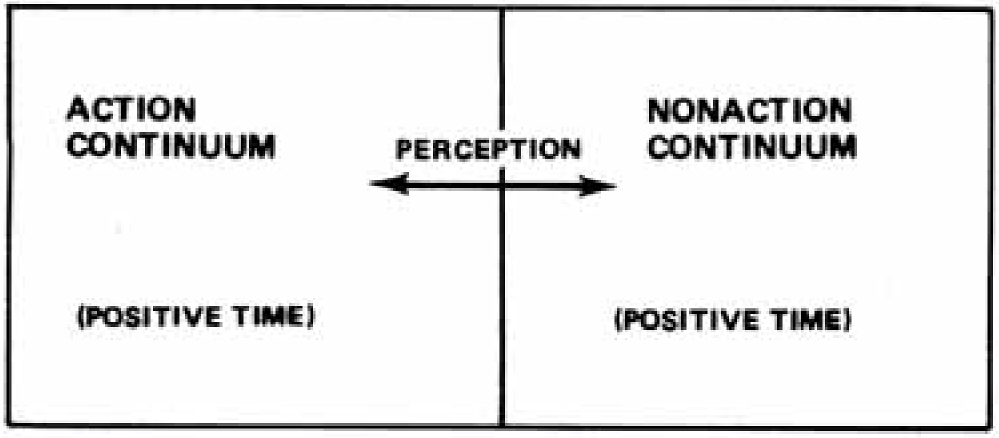
ACTION CONTINUUM
PERCEPTION
NONACTION CONTINUUM
(POSITIVE TIME)
(POSITIVE TIME)

Fig. 4. The Positive Time Continua -- Model in Which Perception Acts

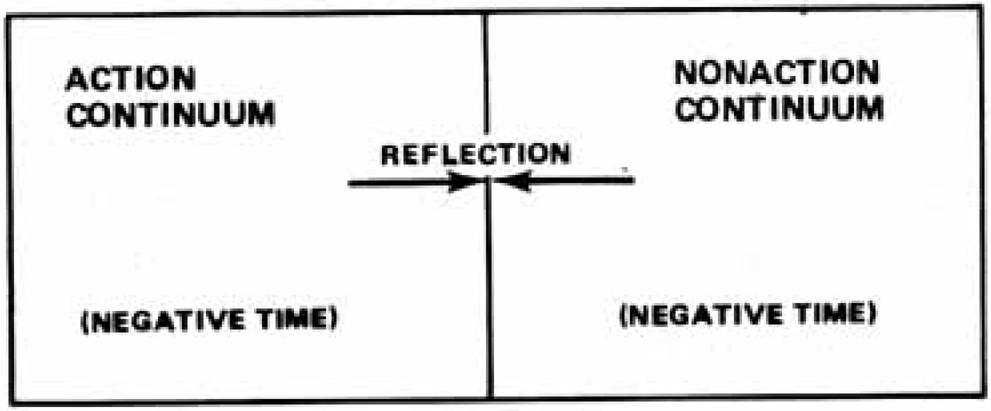
ACTION
CONTINUUM
NONACTION
CONTINUUM
REFLECTION
(NEGATIVE TIME)
(NEGATIVE TIME)

Fig. 5. The Negative Time Continua --
Model in Which Perception Does Not Act

Next Page

Δt quanta outputted by one perceptron operation. The absolute value signs must be added to any dimensional molecule which can take on either positive or negative values. Dimensionally,

$$M_0 = |E| \ , \tag{16}$$

where M_0 is rest mass and requires no absolute value sign since a quantity of rest mass is only positive. Actually, since perceptron operation only outputs quanta, then the perceptron outputs ΔE and ΔM_0, or

$$\Delta M_0 = \left|\Delta E\right| \ , \tag{17}$$

and repetition rate allows perception of M_0 as a positive quantity. Then

$$\frac{\Delta A}{\Delta M} \longrightarrow \frac{\Delta(L^2)}{\Delta t} \ . \tag{18}$$

The perceptron expression of the right-hand side of (18) will be[1]

$$\frac{\Delta(L^2)}{\Delta t} \longrightarrow \frac{\Delta L \Delta L}{\Delta t} \longrightarrow \Delta L\,(V) \ , \tag{19}$$

and then we have

$$\Delta E \Delta t(V/V) \longrightarrow \frac{\Delta E(V \Delta t)}{V} \longrightarrow \Delta P(\Delta t)(\Delta L/\Delta t) \longrightarrow \Delta P \Delta L \ , \tag{20}$$

which must be another statement of the minimum condition of the Heisenberg uncertainty principle, or

$$k = \Delta P \Delta L \ . \tag{21}$$

It follows that

$$\Delta P \Delta L = nk, \qquad n = \pm 0, \pm 1, \pm 2, \ldots \tag{21a}$$

To consider why constant velocity yields special relativity, from $E = ML^2/T^2$ and Eq. (16), we can write

$$M = \left|\frac{ML^2}{T^2}\right| = M\left|\frac{L^2}{T^2}\right| \ . \tag{22}$$

Dividing out the M gives

$$1 = \left|\frac{L^2}{T^2}\right| = \left|\frac{L}{T}\right| = \frac{|L|}{|T|} = |V| \;. \tag{23}$$

---

[1] Obviously $\Delta(L^2) = 2L\Delta L + (\Delta L)^2$, but the units of $2L\Delta L$ are $\Delta L\Delta L$.

Next Page

So velocity is a dimensionless quantity in the absolute dimensional sense, i. e., in the perceptron operational sense, and "the laws of physics apply equally well for all observers as long as they are moving with constant velocities" (Einstein's postulate). When the perceptron process is applied to a system of objects whose relative velocities are constant, the inputs to perception always differ by a constant value and the differentiated outputs of perception do not differ, i. e., the differential of $f(x) + C_1$ is identical to the differential of $f(x)$.[1]

Observation is related to perception as shown in Fig. 6. An observer observes only perceptron outputs[2] -- never ultimate reality -- from his sensory apparatuses.

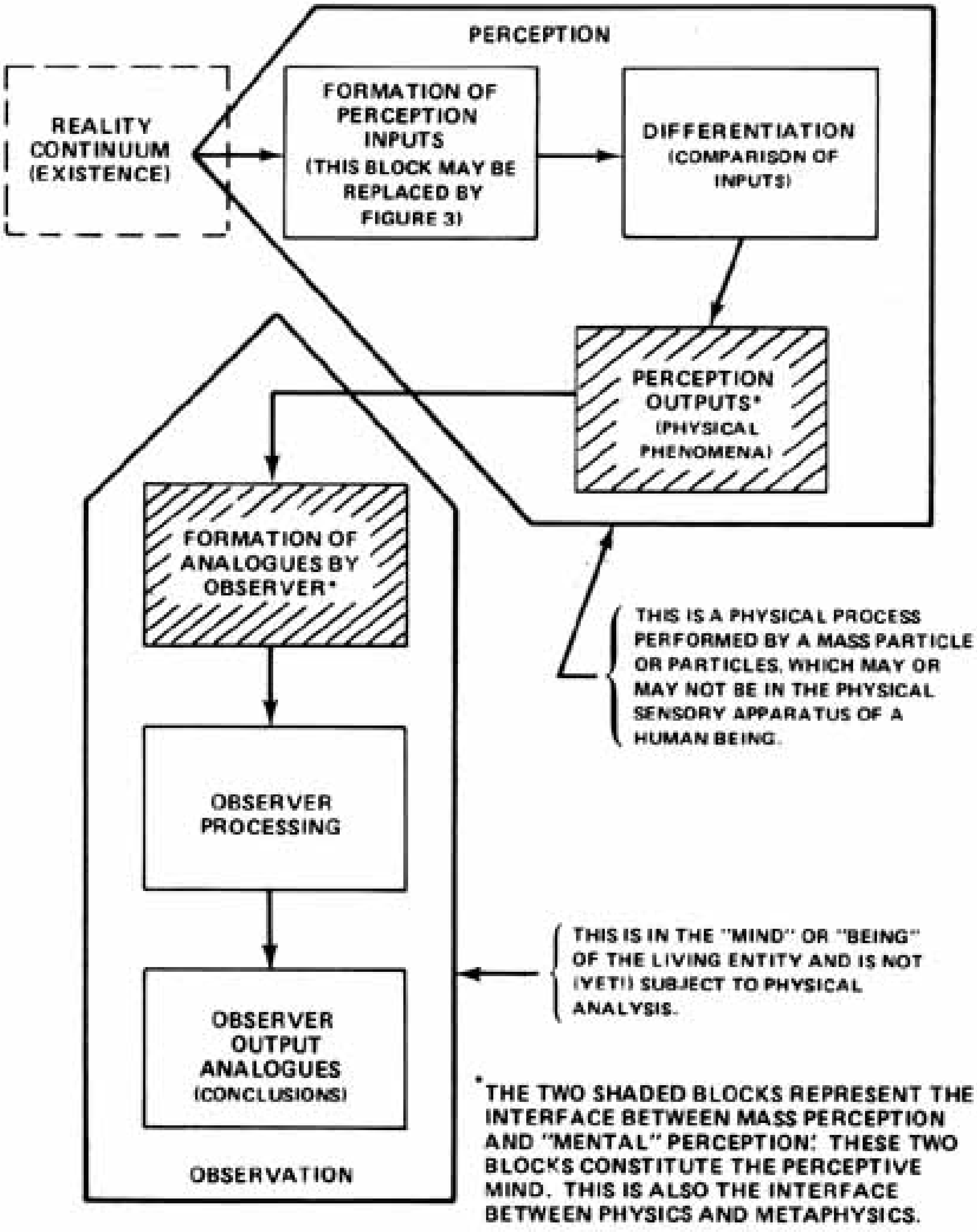
PERCEPTION
REALITY CONTINUUM (EXISTENCE)
FORMATION OF PERCEPTION INPUTS (THIS BLOCK MAY BE REPLACED BY FIGURE 3)
DIFFERENTIATION (COMPARISON OF INPUTS)
PERCEPTION OUTPUTS* (PHYSICAL PHENOMENA)
FORMATION OF ANALOGUES BY OBSERVER*
OBSERVER PROCESSING
OBSERVER OUTPUT ANALOGUES (CONCLUSIONS)
OBSERVATION

THIS IS A PHYSICAL PROCESS PERFORMED BY A MASS PARTICLE OR PARTICLES, WHICH MAY OR MAY NOT BE IN THE PHYSICAL SENSORY APPARATUS OF A HUMAN BEING.

THIS IS IN THE "MIND" OR "BEING" OF THE LIVING ENTITY AND IS NOT (YET!) SUBJECT TO PHYSICAL ANALYSIS.

*THE TWO SHADED BLOCKS REPRESENT THE INTERFACE BETWEEN MASS PERCEPTION AND "MENTAL" PERCEPTION: THESE TWO BLOCKS CONSTITUTE THE PERCEPTIVE MIND. THIS IS ALSO THE INTERFACE BETWEEN PHYSICS AND METAPHYSICS.

Fig. 6. Overall Relationships

---

[1] Einstein's treatments of simultaneity and special relativity may be regarded as

---

[1] Einstein's treatments of simultaneity and special relativity may be regarded as statements that one can determine a length (measure a ΔL) only by $\Delta L \equiv c\Delta t$, and one can determine (measure) a Δt only by $\Delta t \equiv \Delta L/c$ in the static case.

[2] The process of observation actually consists of both perception and realization.

Next Page

III. MASS AND SPACE

The nonaction continuum may be further visualized as a sea or gas of small, unperceivable particles, i. e., of quitons and aggregates of quitons which are less than k in magnitude. Since mass is a perceived quantity and a quiton cannot be perceived, a quiton or aggregate of less than k magnitude has zero mass.

Since fission and fusion both constitute perceptron switching, we may explain mass by visualizing a changing pile of quiton bundles, each of k size, in which the total rate of switching activity is what we call the "mass"[1] of the pile. Thus the mass is equal to the sum of the absolute values of the fission and fusion activities:

$$M \equiv \left|\frac{\Delta A_i}{\Delta t}\right| + \left|\frac{\Delta A_o}{\Delta t}\right| \simeq \left|\dot{n}_i + \dot{n}_o\right| \quad k \ , \tag{24}$$

or

$$M \simeq \left|\dot{n}_i + \dot{n}_o\right| \qquad \text{sw/sec}, \tag{25}$$

where[2]

Δt = perceptron operation time (one operation)

M = perceived mass of the pile

ΔA_i = action created from fu~~z~~sing quitons into action atoms

ΔA_o = action lost from fissioning action atoms into quitons

$\dot{n}_i$ = number of k-sized bundles per second fused into atoms

$\dot{n}_o$ = number of atoms per second fissioned into quitons

sw/sec = one perceptron operation per second.

---

[1] If ΔE is time rate of change of action, and if the absolute dimensional value of rest mass and rest mass energy are the same, then mass must also be a time rate of change of action. One kilogram mass equals 17.053×10^{50} sw/sec.

[2] The switch concept is necessary to explain charge and charge effects. The switching occurs between the positive (perception outputted) and negative (reflection outputted) time streams. Three types of switches exist, two of which are charge switches and all of which are mass switches. Switching may also be regarded as change of Δt into ΔL. For a mass switch, the limiting value of the conversion parameter b in ($\Delta L = b\Delta t$) is c, the speed of light. However, the

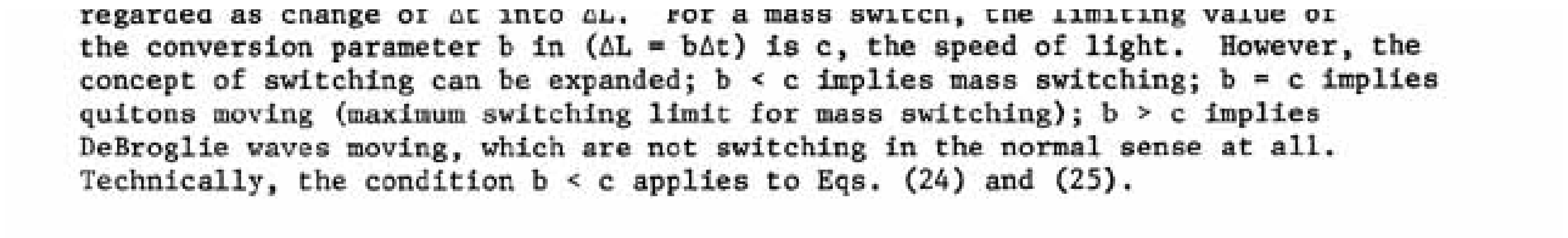

regarded as change of Δt into ΔL. For a mass switch, the limiting value of the conversion parameter b in ($\Delta L = b\Delta t$) is c, the speed of light. However, the concept of switching can be expanded; $b < c$ implies mass switching; $b = c$ implies quitons moving (maximum switching limit for mass switching); $b > c$ implies DeBroglie waves moving, which are not switching in the normal sense at all. Technically, the condition $b < c$ applies to Eqs. (24) and (25).

Next Page

The use of the mass definition equation can be illustrated by a simple example. First,

$$k \equiv \hbar/2 = 0.527 \times 10^{-34} \text{ joule-sec} \tag{26}$$

$$1 \text{ kg mass} = 8.987 \times 10^{16} \text{ joules ,} \tag{27}$$

so

$$1 \text{ kg mass} = \frac{8.987 \times 10^{16} \text{ joules}}{0.527 \times 10^{-34} \text{ joule-sec/switch}} \tag{28}$$

$$1 \text{ kg mass} = 17.053 \times 10^{50} \text{ switches/sec.} \tag{29}$$

For a mass that does not accelerate, switches must occur in couplets. This would require, for example, that angular momentum be quantized in quanta of 2k magnitude, since two switches would process two action atoms -- one in and one out of the mass action pile. Since

$$2k = h/2\pi \text{ ,} \tag{30}$$

then angular momentum should be quantized in $h/2\pi$ quanta, which indeed it is: from the Bohr theory of the hydrogen atom,[1]

$$mvr = \frac{nh}{2\pi} \qquad , \; n = 0, 1, 2, \ldots \text{ ,} \tag{31}$$

and the change permitted is thus

$$\Delta(mvr) = \frac{nh}{2\pi} \qquad , \; n = 0, \pm 1, \pm 2, \pm 3, \ldots \tag{32}$$

What we normally think of as empty space is sometimes described as the absence of mass and energy, existing in time.[2] Empty space can be defined by

$$\frac{\Delta A}{\Delta t} = 0; \qquad \frac{\Delta \bar{A}}{\Delta t} = 0 \text{ ,} \tag{33}$$

which merely state that no perceptron output is occurring. Thus empty space exists only as defined by a perceptron and is merely the total absence of perception. While these equations rule out fission and fusion, they do not rule out the static presence of large numbers of quitons and quiton aggregates each of which is less than k in size (a quiton may be viewed as a quantum of space or a quantum of ether).

---

[1] Beiser, Concepts of Modern Physics, McGraw Hill, 1963.

[2] Actually there is no time connected with the existence of space; the time is connected only with our perception of change; that is, space is a product of perception, or the nonoperation of perception, and is related to a mass perceptron. Similarly, space itself has no length -- length is relative between two masses. That is why both length and time can change, and why a specific length and a specific time can

space itself has no length -- length is relative between two masses. That is why both length and time can change, and why a specific length and a specific time can be different to different observers.

Next Page

A massless quiton cannot individually generate any drag resistance against matter in the Newtonian sense. So a mass object moving along at a constant velocity will just push a stream of massless quitons ahead of it. In effect, the mass "bulldozer" is pushing its own medium along with it, and this requires no force.[1] Therefore, empty space (i. e., the zero output condition of one's mass perceptron) is filled with (has added to it) streams and streams of quitons moving ahead of other objects (perceived by one's perceptron) from all over the universe (perceived by one's perceptron). These quiton streams may be regarded as quiton pulses or quiton waves[2] in one's own static-relative quiton pond. The static-relative quiton pond of a mass will be called its "q-space" or "space" (more commonly referred to as the mass's inertial frame of reference).

When impinging on a perceptron's mass, these disturbances are being collected into k-sized quanta and switched into action atoms. The switching produces outputs that reveal the presences of the disturbing masses. The situation for a perceived mass is described in Fig. 7. Note that a fusion pileup is shown in front of the perceived object due to the object's motion, and a fission pileup is shown at the rear (switching is actually distributed throughout). Additional switching is occurring in the pileups and constitutes increase in mass. So one's mass perceptron perceives the moving object's mass as having increased. The relative velocity of the object with respect to one's perceptron determines the additional collection rate of quitons, and hence the additional switching rate, occurring for the object in one's perceptron q-space. The perceived increase in the moving body's mass is a function of the increased collection rate, which is a function of both the initial mass switch size M_0 and the velocity V, or

$$M = f(V,\ M_0)\ . \tag{34}$$

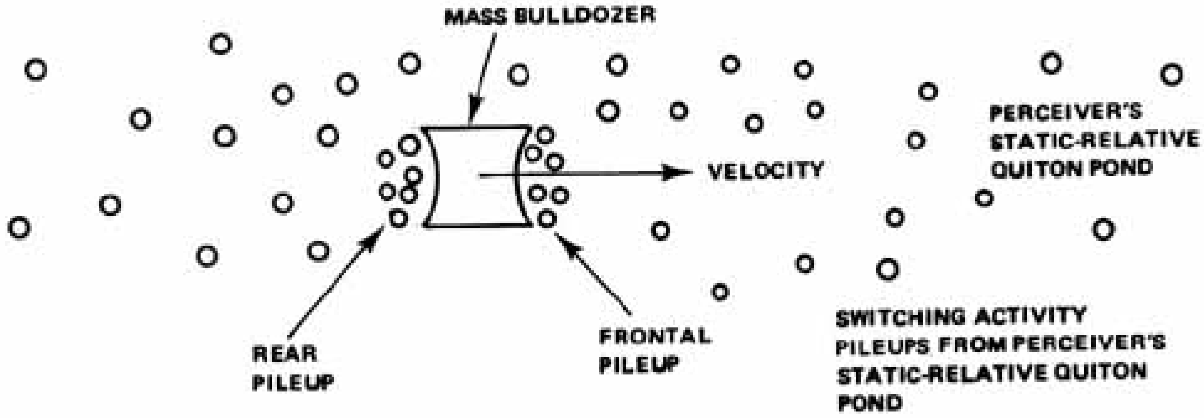
MASS BULLDOZER
PERCEIVER'S STATIC-RELATIVE QUITON POND
VELOCITY
REAR PILEUP
FRONTAL PILEUP
SWITCHING ACTIVITY PILEUPS FROM PERCEIVER'S STATIC-RELATIVE QUITON POND

Fig. 7. Increased Perceiver Switching Due to Perceived Velocity

---

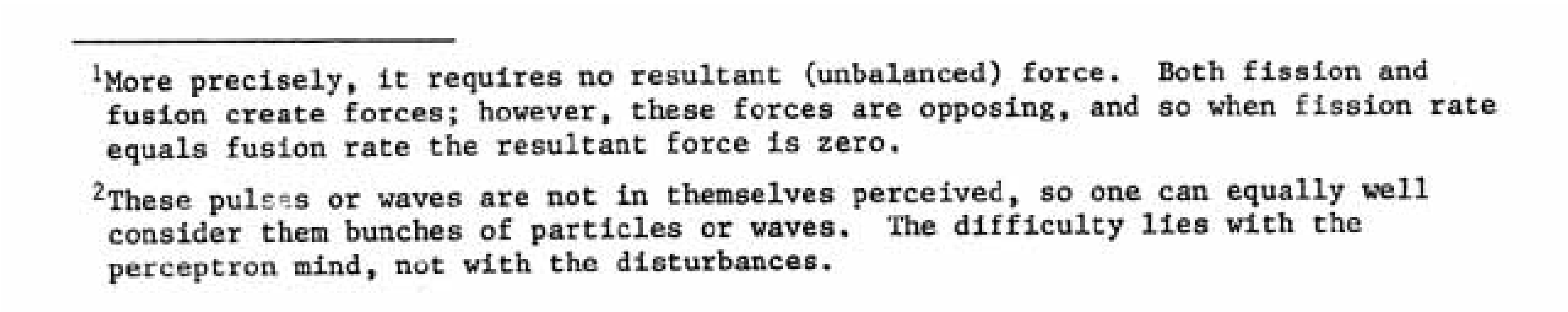

---

[1] More precisely, it requires no resultant (unbalanced) force. Both fission and fusion create forces; however, these forces are opposing, and so when fission rate equals fusion rate the resultant force is zero.

[2] These pulses or waves are not in themselves perceived, so one can equally well consider them bunches of particles or waves. The difficulty lies with the perceptron mind, not with the disturbances.

Next Page

If the perceiver shifts "his view" to the object, he would at the same time shift his perceptron to it and so would then see the same sort of effect existing on his vacated mass. The pileup is relative to the perceiver and is occurring in the perceiver's perceptron itself.[1] So material objects do not just move through an ether, they carry their own q-space ethers right along with them.[2] And all they can detect is the disturbances in that q-space that produces switching transfers in and out of the action and nonaction continua.

We now wish to know the standard medium velocity at which wave disturbances move in the q-space. The question is, what has zero mass, is unperceivable, and yet can move through the perceiver's empty space ether[3] at some standard velocity? Photons satisfy all these requirements. Photons themselves cannot be perceived -- only their interactions can; and photons move in free space at the velocity c, the speed of light. Thus c is the quiton medium's wave velocity[4] and quiton disturbances move in a q-space at the velocity c. They move only in a q-space and in nothing else, since movement (velocity) is defined as $\Delta L/\Delta t$, and both ΔL and Δt exist only to the perceptron that created them. Therefore, velocity of a perceived object is absolutely relative to the perceiver since it is part of the perceiver's mass's perception. At this point, Einstein's postulate that "the speed of light is the same for every observer" can be seen to follow directly from the quiton/perceptron model.

But a difference exists between the speed of light in q-space and the speed of light in a medium. Simply speaking, when light travels in a medium the quiton waves (photons) are perceived to interact with particles (get switched by other switches) continually. In each imposed interaction switching perceived, another Δt occurs in the perceiver's perceptron. Thus more Δt's occur, and the photon is perceived to take longer in traveling through the medium than through free q-space; that is, the perceiver's perceptron (in which the medium itself exists) creates more time output in the case of light traveling through a perceived medium than light traveling through unperceived free q-space.

---

[1] This point is emphasized. The mass's mass does not increase as perceived by itself; its mass does increase as seen by the perceiver. This fact alone is sufficient to establish that the increase in mass is subjective, not objective (although these two terms are usually misused). That is what the term "relative" means, i.e., relative to the subject. Note that the perceiver is always the subject. The fact that the effect occurs only to the perceiver should indicate that it is the perceiver's perception process which is causing the effect.

[2] Thus one can have one space moving through another space.

[3] From Einstein's Theory of Relativity by Max Born, p. 224: "Einstein in later years proposed calling empty space equipped with gravitational and electromagnetic fields the 'ether,' whereby, however, this word is not to denote a substance with its traditional attributes. Thus in the 'ether,' there are to be no determinable points, and it is meaningless to speak of motion relative to the 'ether.' Such a use of the word 'ether' is of course quite admissible, and when once it has been sanctioned by usage in this way, probably quite convenient." The use of "ether" here is quite close to Einstein's proposal; however, in our usage it is meaningful to speak of "perceived motion of a perceived object" in the q-space ether of the perceiver. Otherwise, there can be no observation and no observer, which are of course necessary

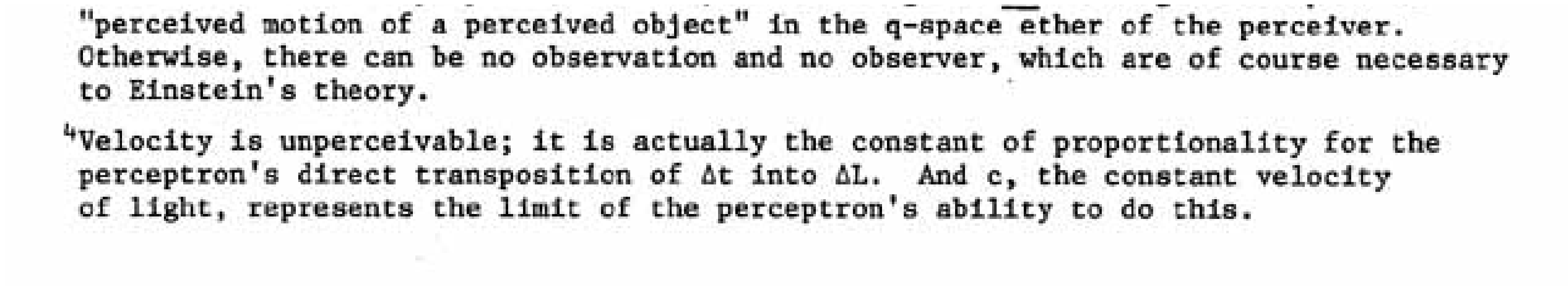

"perceived motion of a perceived object" in the q-space ether of the perceiver. Otherwise, there can be no observation and no observer, which are of course necessary to Einstein's theory.

[4] Velocity is unperceivable; it is actually the constant of proportionality for the perceptron's direct transposition of Δt into ΔL. And c, the constant velocity of light, represents the limit of the perceptron's ability to do this.

Next Page

It should be noted, but will not be pursued further here, that electromagnetic waves in space are merely quiton wave disturbances[1] moving in a perceptron's q-space. The energy being transmitted is the "energy equivalence" of the nonaction quiton aggregates that were chopped up from action atoms. When these fissioned aggregates are again collected in some perceptron, then $\overline{\Delta E}$ is recovered as ΔE, or perhaps as ΔM.

In passing, we note that hidden in the principle of complementarity is the fact that wave and particle are operationally excluded, i. e., separated by an operation. A wave by definition is operational (dynamic). The most elementary idea involved in the concept of a particle is that it is nonoperational; i. e., it doesn't have to move, or change, or function, or be dynamic to exist. Obviously that which is operational differs from (is separated from) that which is nonoperational, but such separation itself is differentiation which is an operation. Therefore, wave (operation) and particle (nonoperation) are operationally separated. The only "mystery" in complementarity is that one must stop (interfere with) an operational entity to have a nonoperational entity, and one must start (interfere with) a nonoperational entity to have an operational entity. Depending upon the conditions imposed upon it, an entity can be either operational or nonoperational. For example, the operation of superposition must be imposed upon DeBroglie waves before they can constitute a particle. It then follows that the particle's velocity must in fact be the velocity of the superimposed group, i. e., the wave packet's velocity. By the same reasoning it follows that one cannot measure a particle's exact position (which is a static or nonoperational entity) and its exact momentum (which is a moving or operational entity) at the same time. The same is true for the particle's exact time position (nonoperational) and its exact kinetic energy (operational). Thus uncertainty principles, complementarity, and the relationship of DeBroglie wave velocity to wave packet velocity all involve the one fundamental concept that an operational entity is separated from a nonoperational entity by an operation. The changing is separated from the unchanging by a change; i. e., the changing is operationally derived from the unchanging, or perceived reality is derived from unperceived reality.

---

[1] Photons are aggregates of quitons; quiton aggregate behavior is particulate. Behavior from the standpoint of individual quitons is wave-like. This explains the dual behavior of light and reconciles its wave and quantum aspects.

Next Page

IV. IMPACT ON PHYSICS

A. MICHELSON-MORLEY EXPERIMENT

The Michelson-Morley experiment was conducted in the Michelson/Morley/earth q-space. The Michelson/Morley/earth perceptron(s) would perceive the speed of light to be the same in all directions, i.e., any fringe shift observed should have been a result of experimental uncertainty only. The experiment did not prove there was no ether; it did prove there was no perceivable ether[1] through which the earth was moving. The results admitted of a static-relative ether if it is unperceived. Perception is a process, and if the process does not occur then perception does not occur. Since by definition the q-space of a perceptron does not cause the perceptron to operate, then the q-space is unperceived by the perceptron.

More careful examination of observation reveals the following facts: A perceived mass has a constant mass when not moving with respect to the perceiver. When it is moving through the space frame or the ether of the perceiver with a constant velocity v, its perceived mass is again constant but is now perceived to be greater than was its rest mass. From whence came the extra mass?

The extra mass effect is only to the perceiver. The only possible sources for the extra mass in perception are the original rest mass M_0 and the empty space of the perceiver's space-frame (i.e., the interaction of perceived mass with unperceived space). Space does not drag, in the conventional Newtonian sense, against a moving body that is not accelerating. And space is unperceivable; therefore it can be visualized as a massless fluid. Thus the increase in perceived mass (which is a change) must come from the perceived mass's increase in velocity in perceiver space (because velocity is all that was changed), and it must involve the physical operation of the perceiver's physical perceiving mechanism. A perceived mass does exert resistance (drag) called "inertia" whenever it is accelerated in the perceiver's space-frame. This resistance is a force. Perceived force is equal to perceived mass times perceived acceleration through the perceiver's space.[2]

---

[1] Is not space an unperceivable ether? An extract from Einstein's Theory of Relativity by Max Born, p. 223, states: "If each of two observers who are moving relative to each other can assert with equal right that he is at rest in the ether, there can be no ether. Thus, the extreme development of the ether theory leads to its dissolution as a fundamental concept." The conclusion that there can be no ether does not logically follow. Each observer can assert that he is at rest in an ether (space), and these spaces can be intermoving. That is precisely what a "moving frame of reference" is: one space moving through another. Today we retain the (or more correctly, an) ether concept, but call it "space" or "inertial frame of reference" or "frame." The conclusion should have been that there can be no single "stable ether," but that there can be many.

[2] Newton's law, $F = ma$,* to date has been used in a half sense, i.e., that the acceleration of a perceived mass through the perceiver's unperceived q-space generates force. If this is true, and relativity holds, then the acceleration of any other unperceived q-space, through a perceived mass, also generates force; through one perceived mass in a perceiver's q-space, the q-spaces of all other perceived accelerating masses are accelerating and thereby generating forces on it. By this "other half" of Newton's law and by relativity, Mach's principle (that inertia of a mass is derived from the

accelerating and thereby generating forces on it. By this "other half" of Newton's law and by relativity, Mach's principle (that inertia of a mass is derived from the other masses of the universe) can be shown to be true. Although not collated and developed as such, it is possible to derive the postulates of both special (Continued)

Next Page

This equation states that a perceived accelerating mass must react with the perceiver's q-space, and in fact this interaction may be taken as the basic mechanism that creates force itself. Again the perceiver's physical perception mechanism is the causative agent.

A quite logical conclusion at this point is that the perceiving mechanism perceives (produces) additional perceived mass from the interaction of perceived rest mass with perceived mass-motion (velocity) through its unperceived space.[1] Also, the perceiving mechanism produces perceived inertial drag from the interaction of perceived mass with the mass's perceived acceleration through its unperceived space.

The concept of "ether" should not be rejected while the concept of "space" is retained, since the latter is actually an unperceivable ether. Rather, the concepts of mass, perception, and ether-space should be tailored to fit the perceived phenomena. Quiton/perceptron theory (perception theory) is a deliberate transfer function model specifically tailored to fit these criteria.

B. GRAVITATIONAL ATTRACTION

Newton's law of universal gravitation can be generated, in both effect and form, from quiton/perceptron theory.

A perceptron's q-space is filled with streams and streams of quiton disturbances moving ahead of its perceived moving objects from all over its universe. Thus there exists a flux of quitons at any point in its space, and this flux is isotropic to the first approximation: because of the vast interstellar distances between disturbing masses, the flux gradient in the relatively local neighborhood of any perceiver point is essentially zero. The quiton disturbance flux can be represented as shown in Fig. 8. A quiton flux stream switching pressure P_1 is exerted on M_1 and a switching pressure P_2 is exerted on mass M_2. In any one flux stream line in this shadowed zone, the pressure against either of the two bodies is reduced in proportion to the amount of switching flux subtracted from that flux line by the other body. The total reduction of switch pressure in the shadowed zone is therefore proportional to the product of the two masses M_1 and M_2, considering the effects to be mutually independent.

(Footnote 2 continued from preceding page) and general relativity from quiton/perceptron theory, and the steps essential to that derivation are contained in this paper.

*[More precisely, $F = d/dt\ (mv) = ma + v\ dm/dt$. From the second term, increase or decrease in mass of a moving object also produces a force. This second force component may not be too important except in microscopic phenomena, but for fast microscopic changes in mass (e.g., in the nucleus) it could generate significant forces. Also note that the velocities of all other spaces (i.e., velocities of the q-spaces of all other moving objects) in the perceiving mass's universe interact with its perceived mass change rates, according to the second term of Newton's law, relativistic form.]

[1] A disagreement has long existed over whether a "nonthing" exists if it cannot be observed. The argument is illogical; e.g., if a photon itself cannot be perceived and only effects (changes) produced by it are perceived, does a photon exist? It

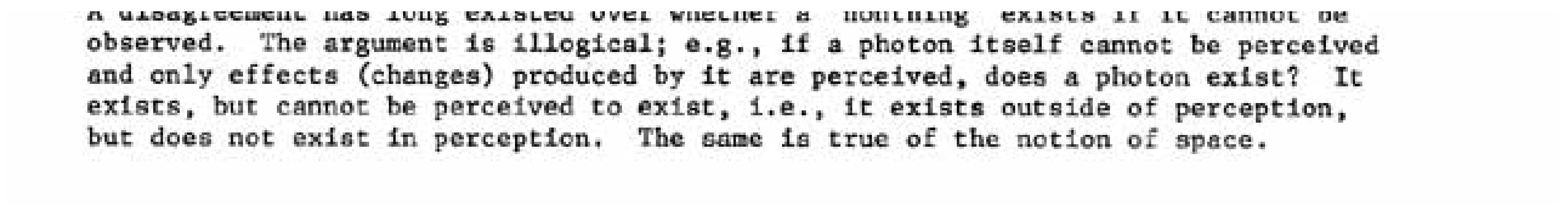

A disagreement has long existed over whether a "nothing" exists if it cannot be observed. The argument is illogical; e.g., if a photon itself cannot be perceived and only effects (changes) produced by it are perceived, does a photon exist? It exists, but cannot be perceived to exist, i.e., it exists outside of perception, but does not exist in perception. The same is true of the notion of space.

Next Page

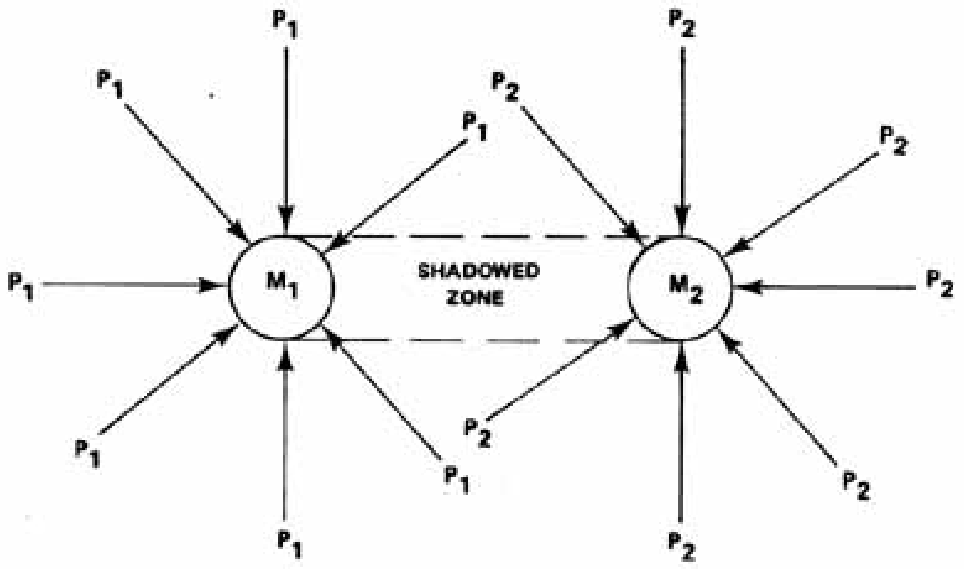

Fig. 8. Gravitational Attraction of Masses

From purely geometrical considerations, a shielding factor is also present due to the solid angle that is shadowed about any mass point. This effect is proportional to $1/R^2$ by usual solid angle considerations. Putting these two factors together, the attraction of mass should therefore be proportional to M_1M_2/R^2, or

$$F = K \frac{M_1M_2}{R^2} \tag{35}$$

which is the expression for Newton's universal law of gravitational force, and this force is driving the two masses together.[1] By merely extending the basic statement by superposition of the effects of every perceived mass-pair in the perceptron's universe, one obtains the attraction of mass[2] for the universe.

---

[1] It is not accidental that electric, magnetic, and gravitational field forces all obey "inverse square of the distance" laws, since these forces are derived from analogous switching operations.

[2] Attraction of mass implies an effect derivable from the presence of mass only. Mass is a switching rate. Therefore attraction of mass must be derived from switching only. Now a massless fluid generates no drag against a constant velocity mass, but switching is a negative quiton acceleration which does generate force. So switching generates drag, but constant velocity does not. This accounts for the error long made in considering the "mass shadowing" explanation of gravitational attraction to suffer from a serious defect of implied retardation force due to velocity through an ether-flux (see, for example, the Feynman Lectures on Physics, Vol. 1, p. 7-9). Velocity of an ether flux through a body merely affects the mass of the body; it does not generate force. Acceleration of an ether flux through a body generates a resultant force on that body. In Feynman's example of an orbital body, let us assume perfect circular motion with constant speed. Since the speed of the ether flux is always constant, the mass is always constant. Since the component of acceleration tangential to the circle is zero, there is no force generated tangentially to the orbit. But since there is a constant radially outward acceleration of flux through the orbital body, then there is a constant radial force outwards that is generated, and this force is called the "centrifugal force," i.e., the "center-fleeing" force.

body, then there is a constant radial force outwards that is generated, and this force is called the "centrifugal force," i.e., the "center-fleeing" force.

Next Page

It should be noted that for extremely small masses at extremely close distances, or for extremely short times, turbulence and interference effects can form in the shadowed zone of Fig. 8 and superposition may no longer hold.

C. NEWTON'S THIRD LAW OF MOTION

One of the recognized ways of stating Newton's third law of motion is: "To every action force there is an opposite and equal reaction force." This law follows from quiton/perceptron theory, as can be shown by examining two colliding bodies. It should be stressed that these two bodies, and all changes related to them, exist as changes in the output of the perceiver's perceptron. It must always be borne in mind that these changes, or physical phenomena, exist as outputs of the perceiver's mass perceptron.[1,2]

Figure 9 shows perceived bodies M_1 and M_2 in constant velocity motion just before collision. A deliberate choice was made of a case where the momenta of the two bodies were not equal, so that the initial switching rates for the two masses were different. In the example, $M_1V_1 > M_2V_2$. The mass switching drag of each body is again represented by small circles. Each body has the same size roll in front as in the rear, so its switch pressure forces are balanced. Figure 10 shows the collision.

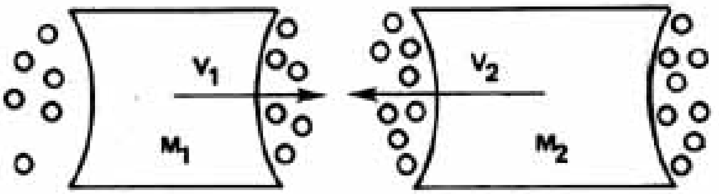

Fig. 9. Two Bodies Before Collision

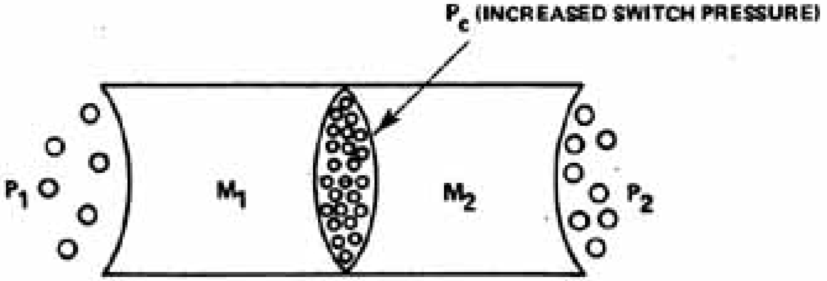

Fig. 10. Two Bodies at Moment of Collision

---

[1] We consider the question: "If a tree falls in a forest and there is no ear to hear it, is there a sound created?" A tree is a perceived tree, a forest is a perceived forest, and a fall is a perceived change. Their presence or existence implies a perceiver and perception; thus they imply that a "sound" change would be perceived, since they imply a perceiver to perceive it. If, however, by "no ear to hear it" we mean "no perceiver present at all," then the question as posed contains a self-contradiction: with no perceiver, there can be no perceived tree, perceived forest, perceived fall, or perceived sound.

[2] The manner of searching for quiton/perceptron causal explanation of a physical

or perceived sound.

[2] The manner of searching for quiton/perceptron causal explanation of a physical phenomenon is to ask, "What would the perceptron have to do to generate the phenomenon?".

Next Page

There is an almost instantaneous switch pressure increase in the contact zone between the colliding bodies, and consequently there is a switch pressure differential across each body as shown in Fig. 11.

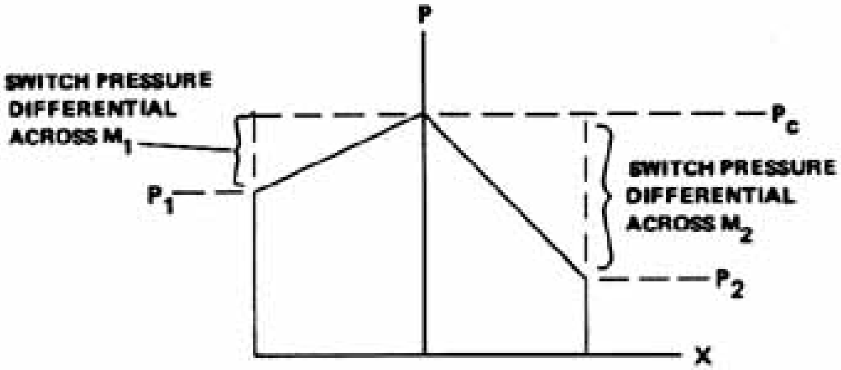
P
SWITCH PRESSURE DIFFERENTIAL ACROSS M_1
P_1
P_c
SWITCH PRESSURE DIFFERENTIAL ACROSS M_2
P_2
X

Fig. 11. Switch Pressure Relationships at Collision

In the figure, the contact pressure P_c is exerted against each body, so that the pressure forces on the two bodies act in opposite directions. But for any two colliding bodies, the contact areas, one against the other, are the same. This common contact area will be called A_c. Since $P \equiv F/A$, then

$$|F_1| = |F_2| = P_c A_c \ . \tag{36}$$

But since the pressure forces on the two bodies are being exerted in opposite directions, then

$$\vec{F}_1 = -\vec{F}_2 \ , \tag{37}$$

which is Newton's third law of motion.

D. NEWTON'S SECOND LAW OF MOTION

Neglecting constants of proportionality, Newton's second law of motion may be stated in the most general sense as

$$F = \frac{d(mv)}{dt} \ . \tag{38}$$

The first step in deriving this law is to examine a point mass which is perceived moving with velocity $\vec{V}$, as shown in figure 12.

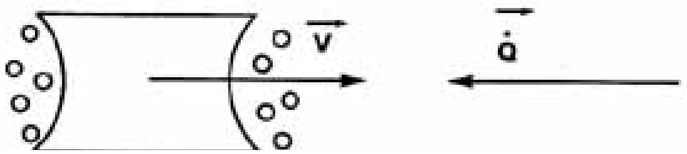

Fig. 12. Moving Point Mass

Next Page

An increased rate of quiton switching activity exists along the axis containing the velocity vector $\vec{V}$. We will call this increased quiton activity a vectorial "current" $\vec{\dot{Q}}$, and define its direction as contrary to the velocity vector. The mass is perceived to be moving against the current vector, and

$$\vec{\dot{Q}} \;\alpha\; -\vec{V} \quad . \tag{39}$$

When the current $\vec{\dot{Q}}$ is steady, i.e., when

$$\vec{\dot{Q}}_{in} = \vec{\dot{Q}}_{out} \quad , \tag{40}$$

then the switching activity, and hence the mass and the energy, are constant. Therefore, from $v = \pm\sqrt{2E/m}$, the velocity remains constant.

If $\vec{\dot{Q}}_{out}$ suddenly increases, then

$$\vec{\dot{Q}}_{in} < \vec{\dot{Q}}_{out} \quad , \tag{41}$$

which can be represented as in Fig. 13.

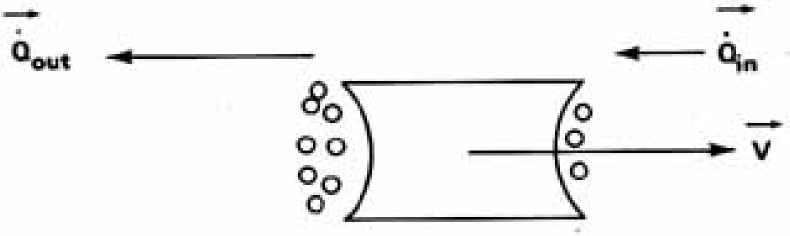

Fig. 13. Difference in Quiton Current

From Eq. (39),

$$\vec{\dot{Q}} = -k_1\vec{V} \quad . \tag{42}$$

At any time in a quiton current, the switch pressure force exerted against the mass can be represented as

$$F = P_c A_c \quad , \tag{43}$$

where P_c is the switch pressure and A_c is the area of contact. However, since the area of contact A_c, or the drag area, is the mass switch rate itself, then (neglecting constraints of proportionality)

$$A \;\alpha\; M' \tag{44}$$

$$A_c \alpha M' \quad , \qquad (44)$$

where M' is the switch rate at the point of switch pressure application. We will next examine the switch pressure forces before and after one perceptron operation

Next Page

(i. e., across the smallest possible time interval Δt used by the perceptron itself in perceiving a change). When there is any difference in the pressure forces against the body front and body rear, this difference is the resultant force $\vec{R}$, or

$$\vec{R} = \left[(P_cA_c)_{in} - (P_cA_c)_{out}\right]\vec{q} \quad , \qquad (45)$$

where $\vec{q}$ is a unit vector whose direction is given by the direction of $\vec{\dot{Q}}_{in} - \vec{\dot{Q}}_{out}$. Now, from Eq. (44),

$$\vec{R} = \left[(M'P_c)_{in} - (M'P_c)_{out}\right]\vec{q} \quad . \qquad (46)$$

Since $P_c \;\alpha\; \dot{Q}$ and $\vec{\dot{Q}} \;\alpha\; (-\vec{V})$, then (neglecting constants of proportionality)

$$\vec{R} = -\left[(M'\vec{\dot{Q}})_{in} - (M'\vec{\dot{Q}})_{out}\right] \quad , \qquad (47)$$

which from Eq. (39) gives

$$\vec{R} = (M'\vec{V})_{in} - (M'\vec{V})_{out} \qquad (48)$$

or

$$\vec{R} = K_4\Delta(M'\vec{V}) \quad . \qquad (49)$$

However, this has happened in the smallest possible time Δt, and since by definition

$$\Delta(M'\vec{V})_{min\ \Delta t} = (\dot{\overrightarrow{M'V}}) \quad , \qquad (50)$$

then

$$\vec{R} = k_4(\dot{\overrightarrow{M'V}}) = k_4\frac{d}{dt}(M'\vec{V}) \quad , \qquad (51)$$

which is Newton's second law of motion, relativistic form. Acceleration of a body is caused by a difference in mass switch rate (i. e., a switch rate gradient)[1] induced across it by transfer of momentum from another moving body or by induction from a field such as an electrostatic field or a magnetic field.

E. NEWTON'S FIRST LAW OF MOTION

Newton's first law, that "a body at rest or in uniform motion will remain at rest or in uniform motion unless some external force is applied to it," can be derived from the same type of considerations used to develop the second law. Obviously, if

$$\vec{\ddot{Q}} = 0 \quad , \qquad (52)$$

---

[1] Switching can be thought of as transfer of Δt into ΔL. An increased switch rate represents an increased rate of transfer resulting in a higher value for the ratio

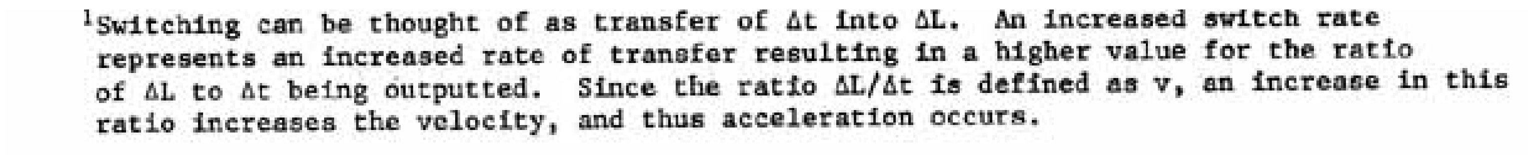

[1] Switching can be thought of as transfer of Δt into ΔL. An increased switch rate represents an increased rate of transfer resulting in a higher value for the ratio of ΔL to Δt being outputted. Since the ratio $\Delta L/\Delta t$ is defined as v, an increase in this ratio increases the velocity, and thus acceleration occurs.

Next Page

then

$$\vec{R} = 0 \tag{53}$$

and

$$\vec{a} = 0 \quad , \tag{54}$$

in which case the body at rest will not move and the body in uniform motion will remain so, by the definition of acceleration. Every perceptron repetition will output the identical value for V as the preceeding repetition, and that is why the first law holds.

F. IDENTITY OF LENGTH AND TIME

From $E = Mc^2$, we have

$$c^2 = E/M = 9 \times 10^{16} \text{ joules/kg} \tag{55}$$

$$c^2 \Leftrightarrow 9 \times 10^{16} \text{ joules/kg} \times 1/17.053 \text{ kg sec/sw} \tag{56}$$

$$c^2 \Leftrightarrow 0.527 \times 10^{-34} \text{ joule-sec/sw} \tag{57}$$

$$c^2 \Leftrightarrow \text{k/mass switch} \ . \tag{58}$$

Thus

$$c^2 = (3 \times 10^8)^2 \text{ m}^2/\text{sec}^2 \Leftrightarrow \frac{\text{k}}{\text{mass switch}} = \text{one sw opn} \tag{59}$$

$$\text{one sw opn} \Rightarrow (3 \times 10^8)^2 \ \text{m}^2/\text{sec}^2 \tag{60}$$

$$\Rightarrow 3 \times 10^8 \text{ m/sec} \tag{61}$$

$$\Rightarrow 1 \text{ sec} = 3 \times 10^8 \text{ meters} \ . \tag{62}$$

Length and time are thus synonymous,[1] and c is the constant of proportionality (more precisely, the identical ratio for transposing Δt into ΔL will be maintained for each perceptron operation). Therefore, a perceptron actually outputs only one thing which may be taken to be either Δt or ΔL (it actually outputs only change

---

[1] By "synonymous" we mean that Δt and ΔL are intertransposable by perception operation. The term c is the velocity (rate of change) of switching itself, i.e., the speed at which quitons themselves can move, since quitons are being switched by mass. Therefore c ought indeed to be the wave velocity at which quitons move, and thus represents the maximum ΔL obtainable from Δt by mass in normal switching. The possibility exists that abnormal (turbulent) switching can deviate from this limitation. Thus "time warping," "time travel," "hyperspace travel," and the interplanetary drive may indeed

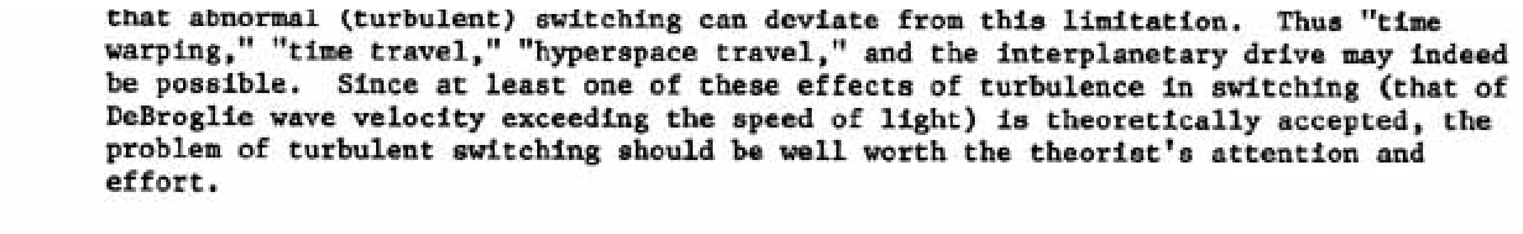

that abnormal (turbulent) switching can deviate from this limitation. Thus "time warping," "time travel," "hyperspace travel," and the interplanetary drive may indeed be possible. Since at least one of these effects of turbulence in switching (that of DeBroglie wave velocity exceeding the speed of light) is theoretically accepted, the problem of turbulent switching should be well worth the theorist's attention and effort.

Next Page

itself, which is perceived and thought of as ΔL and Δt). All other dimensions, quantities, and dimensional molecules are but bit division or bit multiplication. This seems to be consistent with some of the latest conclusions from general relativity theory.

In passing we note that there is no separation without relation, and there is no relation without separation. Therefore

$$\text{relation} \Leftrightarrow \text{separation.} \tag{63}$$

Also, there is no operation without separation, and there is no separation without operation. Therefore

$$\text{operation} \Leftrightarrow \text{separation.} \tag{64}$$

It follows that

$$\text{operation} \Leftrightarrow \text{separation} \Leftrightarrow \text{relation.} \tag{65}$$

We note a difference between free space and a Cartesian space (inertial reference frame). In a Cartesian space or inertial reference frame a definite length is considered to have been established for each and every point in that frame. But such a definite length to each point is rigorously operational by statement (65) above, i. e., such a length is defined by an operation. Specifically, it is defined by the operation of a mass perceptron (or an understood mass perceptron) at the origin of the frame. Such a frame is linear if the identical type of perceptron operation has defined the length to each point. If the type of defining operation varies, then the frame so defined is nonlinear; i. e., its space is said to be curved or distorted.

In one perceptron, its length defining operation and its mass defining operation can interfere or react one with the other. This type of interference or interaction is the direct interaction of frame space (length) with mass and it is the fundamental generating mechanism of force. All forces are so generated (see again the inset footnote on p. 17). And in fact a perceived moving force does work in the direction of its perceived movement rather than in the force's perceived direction precisely because of the moving operational length's interaction (interference) with the operationally generated force. This interference interaction is between operations in the same perceptron and generates what is called work.

Next Page

V. ELEMENTARY PHILOSOPHICAL IMPLICATIONS

Every scientific treatise is filled with forms of the words "one" and "be," yet none attempt to define these metaphysical terms. Ontologists and metaphysicians have uncovered many problems connected with these two words, but no real solutions. In addition, the historical preoccupation of metaphysics with dogmatic religious beliefs contributed strongly to its disfavor, since scientists were required to exert considerable effort to establish the scientific (observational) method as a legitimate criterion for the judgment of physical truth. Consequently, both science and philosophy today are constructed on quite undefined foundations, primarily because of the failure to recognize the relationship of perception operation to reality, which recognition is a prerequisite to the synthesis of physics and metaphysics. One therefore must comprehend the concepts of be (exist),[1] thing (perception), and one (thing) from the viewpoint of the relation between perception and ultimate reality.

A. EXISTENCE, THING, AND ONE

Existence has been defined here as "what is to be fed into the perceptron's input." In the situation where all operation of the perceptron has ceased, what can be said to exist in the input? First, what exists is a continuum: it has no limit, interface, dimension, time, velocity, change, space, place, or discontinuity since these are all derived from perceptron operations and the perceptron is not operating. What exists is continuity only, with no discontinuity.[2] To comprehend this, we must examine the thought process itself more closely, realizing that "to think" is "to perceive thoughts."

What is a "thing?" Further, what is "one" -- one "thing," one "anything?" A thing is first a perceived thing: it resulted from the output of a perceptron operation, in which one input bite and one output occurred. This output is characterized by the the fact that it is one, and the absence of one (i.e., zero) is the absence of output.[3] Figure 14 shows these relationships. The perceptron operation created the separation between the inside and the outside by creating a ΔL and a Δt, that is what the delta symbol means. Thus the operation of the perceptron determines the basic nature of one itself, and any one necessarily requires that a perceptron operation has occurred. In fact, the perceptron itself is a result of perceptron operations. Thus indeed all is relative. A perceptron exists and is able to function because it perceives other perceptrons to exist and function. Thus perceptrons are mutually created derivatives of reality

---

[1] Being is undifferentiated; that is its total definition. It cannot be perceived, since perception differentiates. It is continuity only, without any discontinuity, which is unperceivable. The preoccupation of being with causality (perception) is technically called "inception," the influencing of perception. In Zen philosophy this is technically "attachment to the world of illusion (change);" the literal meaning of "incarnation."

[2] This concept is unperceivable, by the definition of perception, which accounts for its difficulty. To be a "perceived one" implies limit (finiteness); oneness itself, without a bound or limit to make it a one (perceived), cannot be perceived. This is the primary limitation of the human mind (i.e., of perceptive thinking)--being, continuity, and oneness themselves imply absence of perception (differentiation) and are therefore unperceivable.

[3] Thus Boolean algebra describes the most basic operation, and is applicable to the

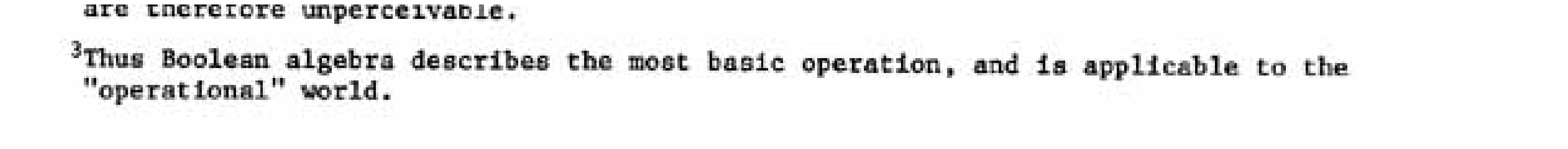

are therefore unperceivable.

[3] Thus Boolean algebra describes the most basic operation, and is applicable to the "operational" world.

Next Page

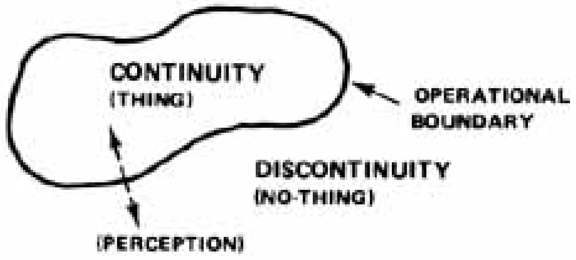

Fig. 14. Nature of One

which drive themselves and each other. Note that the preceding sentence itself is a perception.

So in the fullest sense, Gautama Buddha was indeed correct to state that the world of phenomena (changes) is an "illusion," relative only to itself (i. e., to perceptrons).[1] Einstein, expressing this idea in the language of mathematics, produced the theory of relativity and inaugurated a new physics. Buddha and Einstein spoke in different technical languages, but both expressed the same idea. The only difference is that in science we have assumed that our observations themselves are real and perfect. They are real, but not perfect in that they represent only changes in ultimate reality. Thus observational science alone can never reveal ultimate reality.

Another approach may aid in the comprehension of the nature of one. What is meant by the statement "he is the man?" Reflecting, one realizes that what is meant is that the perceived thing HE is exactly the same perceived thing THE MAN, and this identity exists now (during this one Δt perceptron time slice). There is no part of HE that is discontinuous from THE MAN, or vice versa. The different forms of the verb "be" attempt to express continuity without discontinuity, or unlimited oneness which is a state and not a thing. However, we do use the verb "be"[2] to relate continuity between different time slices. "He is the man" is related in the present time slice; "he was the man" means that HE in the NOW slice is identical to THE MAN in a past slice; etc.

---

[1] The reader interested in Zen will find complete agreement between this paper and Zen precepts, and will find the paper of some assistance in comprehending the unperceivable (unthinkable) Zen realizations. In Zen, for example, the aim is to comprehend at once the totality of reality, directly, escaping beyond the limiting fetters of perceptive thought. Now if it is actually possible to do this, and if some unusual men have done it, then their enigmatic writings should contain the wisdom of their experience, and should correspond to ultimate reality itself. Since ostensibly their insight is obtained by other-than-logic and other-than-perceptive means, then their statements should appear illogical and unperceivable.* Thus the "sound of one hand clapping" and other illogical concepts are used in Zen as a meditative means of driving the student beyond logical perception (certainly the sound of one hand clapping cannot be perceived or thought) to realization. But if the Zen experience is actually valid, then reality itself must be pure Zen (i.e., unperceivable), and so it would appear to be. We note that for each component part of all, the opposite part must exist. So all can be regarded as the complete set of all possible opposites or contradictions.

*(However, this most assuredly does not imply that every statement that is illogical and unperceivable is of value!)

[2] It should be stressed again that being is defined as undifferentiated (unseparated,

is illogical and unperceivable is of value!)

[2] It should be stressed again that being is defined as undifferentiated (unseparated, unperceived). Being is simply continuity without discontinuity.

Next Page

B. REALITY, ULTIMATE REALITY, BEING, AND ULTIMATE BEING

Actually one should be quite careful in using these terms. What does one get when one differentiates reality? One gets another reality, but not the primary or ultimate reality. For example, assume that length and time are real. When one time-differentiates length, one obtains velocity, which of course is real also but quite different from either length or time. However, velocity is always operationally related to the two first (assumed) reals length and time, because differentiation is an operation. So what is meant by "ultimate reality" is that state (or essence, or whatever name one prefers) that is not differentiated. Since "being" (continuity with no discontinuity) is defined as "that which is not differentiated," then the ultimate reality is pure being.

But all perceivables are differentials of ultimate reality. When one differentiates being, one gets pieces (differentials) of being which themselves are beings, but they have been perceived (differentiated) from a more primary or higher being. Therefore any being that one perceives has been differentiated (separated) by perception,[1] which explains why human beings, for example, consider themselves separate or individual (one). They are separate only in their individual perceptions. Thus "beings" by its plural nature always implies fragments (differentials) of a higher being (of singular nature). There can be many such integral levels: starting with one function, one can integrate it, then the resulting function can be integrated, etc., as long as we are speaking of definite integrals between two limits. When we reach a state that has no pieces or limits, the process is finished. That state is being in the ultimate sense, i.e., ultimate being.

One perceives these integral levels directly in nature, i.e., in the perceived phenomenological world. Living molecules (tiny beings) of RNA, DNA, etc., integrate together into a cell (a higher level being). Cells then integrate into a still higher being (man, animal, plant), which is the level on which man finds himself. But since man is perceivable and there are more men than one, it follows directly that men are capable of being integrated into a still higher being,[2] and the only word we have at present to describe such an ultimate being is "God." It is not at all accidental, therefore, that the ultimate being cannot be directly perceived by man, since by the definition of perception the ultimate being cannot be perceived by a lower order being such as man (i.e., since perception differentiates, or separates, then only differentials of being, or separate beings, are perceivable).

The avatar Jesus, having no technical concept for the derivative available in the language of the day, used the analogy of the father-son relationship, the son being the derivative of the father in this sense. Thus he referred to himself as the "Son of Man" (the derivative of all mankind or of integrated mankind) and also as the "Son of God" (the derivative of the integral of man). He continually referred to the ultimate being as "Father." It is little wonder that he was not understood then, and has not been understood for almost twenty centuries.

---

[1] Thus not only are physical objects relative to the perceiver's perception, but so are the perceived beings. Ergo the entire perceived universe, both living and non-living, is related to perception, and in fact perception and creation are abstractly

are the perceived beings. Ergo the entire perceived universe, both living and non-living, is related to perception, and in fact perception and creation are abstractly synonymous.

[2] This statement contains obvious implications to the theory (i.e., to the goal) of evolution.

Next Page

VI. CAUSALITY AND ITS MODIFICATION

A little closer look at perception can be obtained by viewing it from the causal standpoint.[1] All that perception really creates is Δt and ΔL (which are the same, ignoring constants of proportionality), and the perception we normally think of is point perception, where a point is one switching operation. The perception point may be thought of as following a line path which we will call "lineception" (Fig. 15). Physical phenomena as seen by one perceptron constitute this lineception; or rather, it constitutes the physical phenomena that are perceived or observed by the perceptron. Any point-based perceptron system, such as human perception, will accordingly perceive the universe of physical phenomena as if the perceiving system were a pointer moving along a lineception.[2]

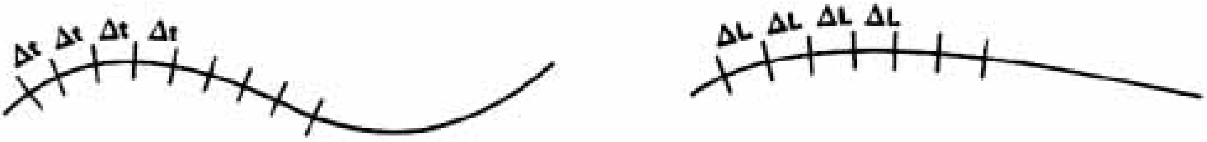

Fig. 15. Lineception (One Connected to One)

For each perception there is a reflection, so there exists (but by definition cannot be perceived) a similar line-path, or "lineflection" as shown in Fig. 16.

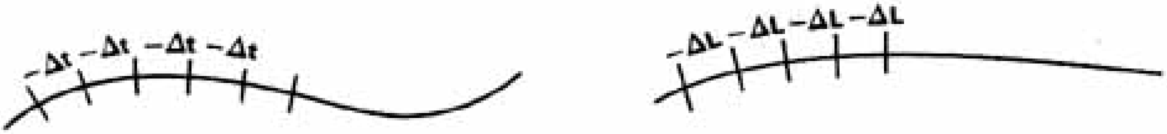

Fig. 16. Lineflection (One Connected to One)

that perception generates causality is completely consistent with our initially assumed definition of causality as the set of all causes.

---

[1] Perception is identical to causality itself, i.e., perception can be said to generate causality. The causal view is that causes inputted to a situation operate to produce, or output or create, effects. According to its inputs, the perceptron generates causal outputs. An adequate theory of perception should be able to explain all causal (perceived) phenomena. We might define causality as the set of all causes. Then the answer to the question, "What causes causality itself?" is causality itself; i.e., since it is the set of all causes, causality must contain its own cause. Causality must therefore be closed or circular. The entire causal world must then be purely relative to itself (closed). Closer attention to causal operation yields the following: We perceive something change (in time or in space), then the situation changes so that we perceive new changes emerge (occur later, or separately). Thus the line of perception by its own unperceived operation connects the perceived cause (change) to the perceived effect (change), i.e., perception generates causality. Since perception itself is causal, then causality truly causes itself. This indeed is closed or circular--causality generates perception (perception is causal) and perception generates causality (causality is perceptive). Thus the idea that perception generates causality is completely consistent with our initially assumed definition of causality as the set of all causes.

[2] A "frame" or "coordinate system" is a point-centered perceptron system centered about the "understood perceptron" at the origin.

Next Page

The ultimate reality, or existence, contains all, all-at-once (ultimate reality is allness, or oneness). At each operation of the perceptron, the exact amount of Δt and ΔL, and their relationship, put out by a perceptron is a function of its input quiton disturbances. A great number, in fact perhaps an infinite number, of different input states are possible, i. e., exist in all, all-at-once (ultimate reality). So an equally large number of output states are possible (exist ultimately),[1] but one particular input condition results in one particular output condition. We can represent the all possible lineceptions existing as "multiception" (Fig. 17). From this figure, a lineception can be seen to be one particular path through multiception. The corresponding multiflection could, of course, be illustrated in similar manner but with negative components.

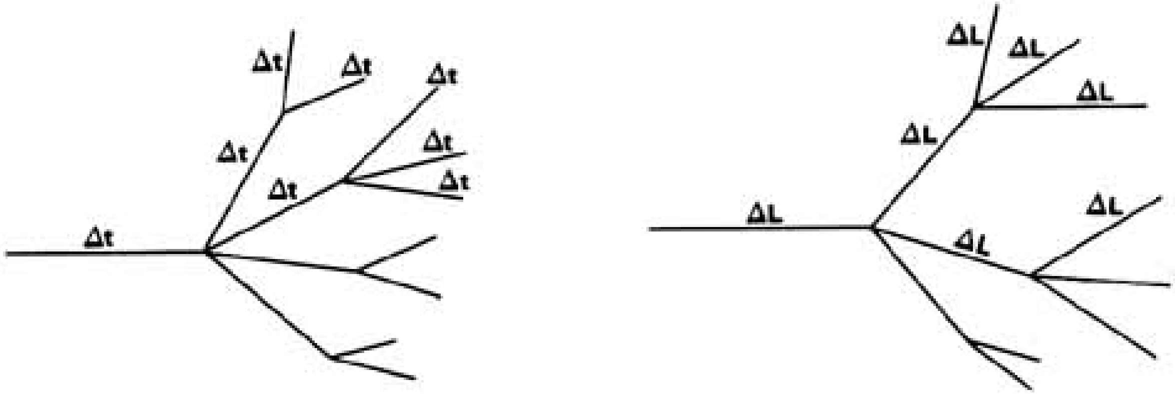

Fig. 17. Multiception (One Connected to Many)

One could immediately apply statistics to the concept along the lines illustrated by Fig. 18.

o

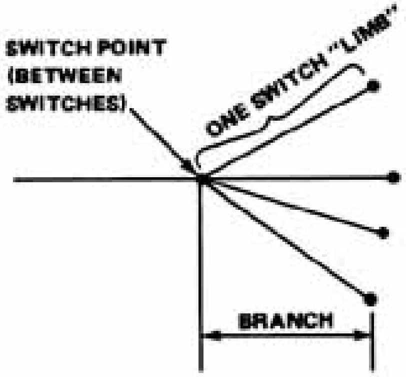

Fig. 18. One Branch

---

[1] It is stressed that these multiceptions exist, which is not at all the same limited statement as "exist now," but much more comprehensive; e.g., all possible numbers must exist in zero, else, given the number A, there is no justification for the axiom that $A + (-A) = 0$. The "absence of number" means precisely the "absence of all possible

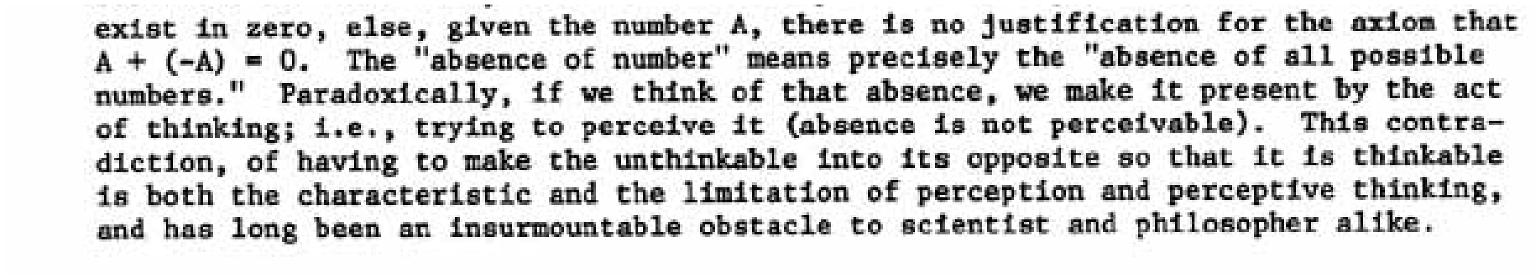

exist in zero, else, given the number A, there is no justification for the axiom that $A + (-A) = 0$. The "absence of number" means precisely the "absence of all possible numbers." Paradoxically, if we think of that absence, we make it present by the act of thinking; i.e., trying to perceive it (absence is not perceivable). This contradiction, of having to make the unthinkable into its opposite so that it is thinkable is both the characteristic and the limitation of perception and perceptive thinking, and has long been an insurmountable obstacle to scientist and philosopher alike.

Next Page

At each switch point, the perceptron is faced with the branch of possibilities ahead, which all exist in multiception but only one of which exists in one perception, or lineception.[1] A certain set of limbs in the branch must have a perception probability equalling 1, and thus each of the limbs has some finite probability.[2] The input conditions to the perceptron then determine which limb is selected for the next switching operation; and, of course, the input conditions are the results of other related perceptrons' operations (relativity). The statistics will not be pursued further here.

For an absolutely mechanistic world, the entire process is automatic according to quite rigidly fixed rules. A system of perceptrons that behave this way shall be referred to as "mechanically switched." What we call "non-living systems" are mechanically switched systems. But what we call a "living system" is able to arbitrarily select the limb taken, to some extent at least,[3] i. e., the fact that a living system is able to choose its behavior within certain limits, directly implies that it must be able to deflect its lineception from a mechanically switched lineception. The only way it can possibly do this is to influence the switching operation of the perceptron. But since the output of one switching operation is a mechanistic function of the inputs, then the only way a life system can affect the perceptron switching operation is by affecting the input to some extent. How could it possibly do this?

In our crude model, the nonaction continuum is composed of action atom fragments, or quitons. But such non-things are massless by definition (they are not being switched and the absence of switching defines masslessness), and we can think of the nonaction continuum as a massless fluid or ether or empty space, since these terms are all synonymous. All that is implied by these terms is that the fluid is composed of massless elements, which merely means they are not being switched. They are capable of being switched, of course, but switching is not being done. The main point is that force, in the conventional sense, is not required to move a massless fluid, so the input fluid to a perceptron can possibly be interfered with by some nonphysical means requiring zero force. The being portion of a life system can apparently do this to some extent.

Since the switch time of a perceptron is on the order of 10^{-21} seconds, then even an incredibly tiny quiton input variation to a perceptron will result in

---

[1] It is stressed that all possibilities exist and are real, but do not exist in perception. Perception is the only difference between unperceived reality and perceived reality.

[2] Which leads to an unusual view of "probability." A perceptron operation is already finished when one perception is complete; that is why "the past is always gone, the present has just occurred, and the future is never here." Perceptron operation continually creates the perceptron's universe as it operates, including all changes, time, space, etc.). Thus, at the conclusion of any one perceptron operation, a multitude of lineception steps can exist for the next operation. However, although possibly quite large, the number of next steps available must be finite to be causal (perceivable). It is the process of perception that makes probability finite. Perception is finite; nonperception is infinite.

[3] The only alternative conclusion is that the living system is unable to deliberately

Perception is finite; nonperception is infinite.

[3] The only alternative conclusion is that the living system is unable to deliberately modify its behavior and its behavior is purely mechanical. But since this implies that all behavior is absolutely programmed and predictable, which is obviously false, this alternative is not valid and the original conclusion stands.

Next Page

noticeably altered lineception after one second of such variation. So it is quite reasonable to presume the futility of trying to measure such infinitesmal influence as is brought to bear on one single perceptron operation. However, the total influence can obviously be detected after being accumulated by a very large number of perceptron operations, and for living systems it is. So while we have not proved the mechanism of life-system manipulation of lineception, we have shown a logical progression from an assumed mechanism to a result that is universally verified by countless observations (i. e., all living systems are observed to manipulate causality, which manipulation is, after all, the definition of behavior). Since no other mechanism has been proposed, to the author's knowledge, that would reasonably explain the observed phenomenon of life-system manipulation of lineception, it is logical to accept as true the assumed mechanism -- a life system can exert an incredibly tiny influence on perceptron inputs,[1] i. e., on the ether or on empty space.[2,3]

---

[1] This tiny influence is the basis of the realization process. To comprehend the noncausal, the individual observer usually wrestles diligently with the problem, repeatedly attempting to perceive it causally. In each such attempt, the tiny influence of inception on perceptron inputs is exerted. Over an ensemble of such attempts, the perceptron causal output can be modified, i.e., modulated noncausally to give what are now slightly noncausal perceived outputs. Presence of a required fit or required output pattern in the perceptive mind allows the output to be examined (compared) against the required output, i.e., it allows the selection of the proper noncausally modulated causal transfer function. When the outputs match, one then has a breakthrough or realization which matches the real world pattern but which cannot be directly perceived causally. However, since the mental file of the perceptive mind has now filed the proper transfer function, the individual has assimilated the information and understands it, being able to recall the noncausally modulated causal transfer function at will. Thus in Zen, the ultimate example of realization is described as "enlightenment." Phrases such as "direct perception of reality," "direct experience of ultimate reality," or "direct realization of all-being" are also used in trying to portray the nonperceivable realization. Conventional science hides the entire process behind vague terms, such as "creative thinking," which actually only conceal the fact that we do not know what it is, or how it is accomplished.

Note that such "creative realization" can cover either purely causal or noncausally modulated perception, the difference being merely whether the inputs are incepted in a normal manner comparable to causal phenomena, or in another manner. Realization is derived only from inception. Since the observed phenomena of physics already appear to be noncausal (statistical) in the ultramicroscopic region while ensembles of these yet appear to be causal, then the physicist is already confronted with the necessity of changing causality but at the same time retaining causality. The implications to the observation process itself must be examined, and they remove the dichotomy between physics and metaphysics when pursued.

[2] The Hieronymus effect is an example of this in reverse. It generates an effect on living nervous systems that is not the result of electric, magnetic, or gravitational fields, and is not caused by matter in motion. For that matter, a polygraph has been used to clearly measure a response of plants to human thoughts alone, and such an effect is not presently explainable within the body of physics theory. "Firewalking" is another unexplainable phenomenon from the standpoint of present physics theory as is acupuncture.

[3] All psi phenomena derive from this effect (inception) together with turbulent

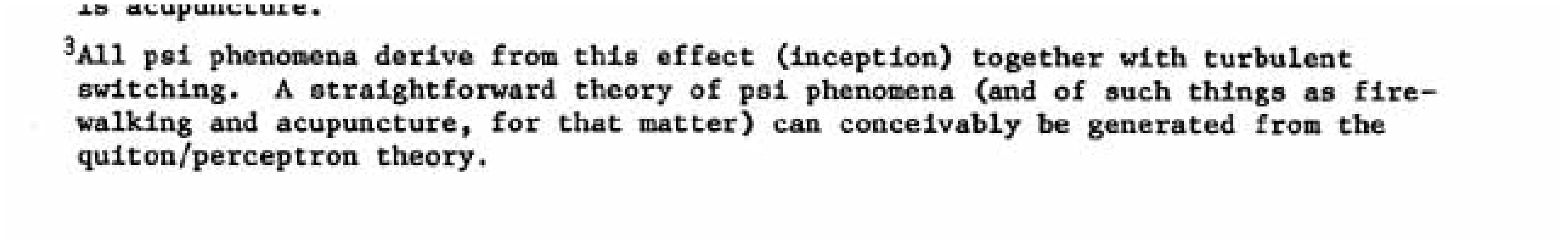

is acupuncture.

[3] All psi phenomena derive from this effect (inception) together with turbulent switching. A straightforward theory of psi phenomena (and of such things as fire-walking and acupuncture, for that matter) can conceivably be generated from the quiton/perceptron theory.

Next Page

Of course the most efficient development of manipulation of lineception (i.e., the widest response) by a life system would necessitate a control system approach; i.e., life system direct manipulation of only a very small control valve or control mechanism, which in turn controls the input to a much larger mechanistic (deterministic) servo amplifier. Systems like this are universally observed in life systems. Thus the great bulk of a life system's manipulation of lineception is servomechanistic: the actual direct influence of life (being) on matter (perceptron switching) is so microscopic as to be immeasurable at its operating level, and only its macroscopic effects become measurable. For that reason the scientist has never been able to isolate the mechanism of life's manipulation of lineception, nor is he likely to be able to do so in the foreseeable future.[1]

Life's manipulation of lineception will be denoted by the term "inception," and can be thought of as a modulation (partial change or additive change) of the lineception carrier. Ordinary physical phenomena are macroscopic ensembles of lineceptions, where the deterministic (mechanistic) law holds absolutely but in a relative manner (ensemble to ensemble) which is fixed cause-and-effect, or causality, from the viewpoint of perception. Since inception is a modulation of lineception, then living behavioral phenomena should be superimposed as envelopes on causal lineceptions, or causal lineception ensembles, and, indeed, they are universally observed to be so superimposed.[2]

Absolute or ultimate reality, however, can be thought of in one respect as all multiceptions and all multiflections, all-at-once, which is timeless (i.e., which is both all-possible-times and zero time simultaneously or identically).[3] Along any one lineception, causality applies; yet since a very large number of changes to that lineception are available in just one multiception branch, then quite a large

---

[1] This latter statement, of course, is subject to challenge.

[2] Since all matter is, it contains being. Another intriguing possibility exists: that there are two types of inceptive systems, or two magnitude directions that inception takes. The first type would produce inception with difficulty, and its inception would only be apparent at the macroscopic level; the second type would produce inception with ease, its inception would be apparent only at the microscopic level. The first type would correspond to what we call "living systems," while the second would correspond to what we call "inert systems." Living systems' inception would be observable to human senses and human instruments, while inert system's inception would not. A second and remarkable consequence would be, that in the extreme microscopic case, inception would dominate and this scale would appear to be absolutely statistical (individually unpredictable), which apparently is true (i.e., degree of causality represents degree of restraint of inception). This would modify causality (for individual events) at the ultramicroscopic level, but at higher levels causality would still apply due to the karma effect on perceived ensembles. It is significant that at present the universe of matter seems to behave in precisely this fashion. To summarize, perception is causal -- it generates causality. Inception, being noncausal, must be below the perception threshold. Aggregates of inception (noncausal) that equal or exceed the perception threshold become causal (are processed by perception) and generate causal karma.

[3] Since every perceived ΔT has an unperceived complement $-\Delta T$ which exists in reflec-

[3] Since every perceived ΔT has an unperceived complement $-\Delta T$ which exists in reflection, then the sum total of all times in absolute existence, which includes both perception and reflection, equals zero time.

Next Page

number of inceptions are available to be applied to that one lineception by an organism to change or direct the lineception. Herein is the explanation of karma. Since every inception made by an organism (or by a submicroscopic particle of being for the second case, see note 2, p. 32) changes its entire lineception, it also changes its entire perceived or relative universe, which is the only universe it knows or senses. And since causal feedbacks from this perceived universe[1] occur to the organism's perceptron (which itself is part of that perceived universe), then the effects of the organism's actions indeed return precisely to it. So, "as a man soweth, that shall he also reap" is quite true.[2]

It is now possible to formulate some conclusive statements:

1. One perceptron follows a lineception path, mechanically basing its output on its inputs, and this operation establishes causality to perception repetition, i.e., to lineception.

2. Ordinary physical phenomena are merely ensembles of related lineceptions, and are perceived to be absolutely deterministic (obey only cause-and-effect.

3. Living phenomena exhibit the effects (selective behavior) of inception (modulations impressed on lineception) which occasionally diverts one lineception from its deterministic lineception limb to another multiception limb, resulting in a changed (modulated) lineception.

4. Ultimate, or absolute, reality may be thought of as the sum-total of all multiceptions and multiflections.

5. "God" is all-being, which is all multiflection and multiception at once, which is both no time at all and forever and all forevers as well.

6. "God's will" is lineception and its accompanying lineflection.

7. "Man's will" or "free will" is inception, the modulation of lineception (and unknowingly its accompanying lineflection). Inception, however, imposes a feedback on the inceptor since a precise feedback, from the inception changes to perception, occurs to the perceptron of the perceiver. This feedback resulting from inception is karma.

8. The law of cause and effect (causality) exists in perceived nonmodulated phenomena (ensembles of lineceptions).

9. Causality is affected (modulated) by living systems, but applies to each choice once made. The causal feedback from inception is karma.

---

[1] "Perceived" must be accented: the nonperceived universe is not causal, since by definition causality invokes a perceiver and the perception process.

[2] Collective karma of any ordered group such as a species must exist in like manner. On the submicroscopic level, "karma of ensembles" means that the statistics of

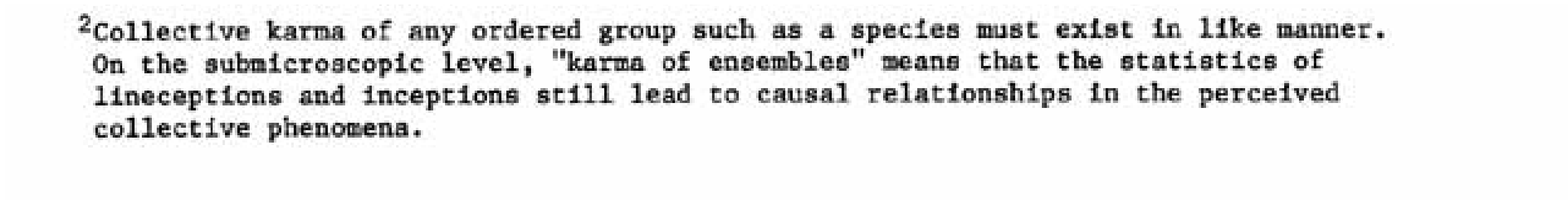

[2] Collective karma of any ordered group such as a species must exist in like manner. On the submicroscopic level, "karma of ensembles" means that the statistics of lineceptions and inceptions still lead to causal relationships in the perceived collective phenomena.

Next Page

10. Neither causality or karma applies in ultimate reality, although all
causality lineceptions and their reflections, and all modulations
(inceptions) and the karma reflections of those inceptions, are
included in ultimate reality.

Next Page

VII. CLOSING REMARKS

The quiton/perceptron approach appears to offer a most extensive area for further investigation, and it is hoped that the physicists will interest themselves in the concept. It offers explanations for the problems long posed by ontology, and unifies the bases of science and metaphysics.

We state categorically that the theory explains what electric charge is, why like charges repel, and why unlike charges attract. In addition, since all phenomena are perceptron outputs, the theory indicates that multiple effects which are very close in time or distance will interfere with each other due to turbulent switching; i. e., conservation of mass/energy holds only as long as no switch interference exists. Turbulent switching can possibly disrupt all known physical laws. Unexpectedly strong forces, such as nuclear force, might in fact result from this effect. Other implications include possibility of travel faster than the speed of light (deBroglie waves do this now) and of travel back and forth in "time."

Since both laboratory instruments and human sensory apparatuses are perceptrons and differentiate reality, investigations of physical phenomena should include the laws of the perceptron as well as the laws of the perceived phenomena.

Next Page

BLANK

Next Page

DISTRIBUTION

	No. of Copies
US Army Missile Command	
Distribution List A	52
Defense Documentation Center	
Cameron Station	
Alexandria, Virginia 22314	12
Commander	
US Army Materiel Command	
ATTN: AMCCG	1
AMCRD	1
5001 Eisenhower Avenue	
Alexandria, Virginia 22304	
AMSMI-O	1
-FR, Mr. Rives	1
-R, Dr. McDaniel	1
-RB, Dr. Kobler	1
-RBL	5
-RP, Mr. Kopcha	1
-RPR	1
AMCPM-MD, LTC Bearden	100
-MD (Record Set)	1

Next Page

BLANK

Next Page

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
SAM-D PROJECT OFFICE
US Army Materiel Command
Redstone Arsenal, Alabama 35809

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP

3. REPORT TITLE
✓ QUITON/PERCEPTRON PHYSICS: A THEORY OF EXISTENCE, PERCEPTION, AND PHYSICAL PHENOMENA

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Technical Report

5. AUTHOR(S) *(First name, middle initial, last name)*
Thomas E. Bearden

6. REPORT DATE
5 March 1973

7a. TOTAL NO. OF PAGES
38

7b. NO. OF REFS
~~0~~ 3

8a. CONTRACT OR GRANT NO.

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
AD 763210

10. DISTRIBUTION STATEMENT
Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Same as No. 1

13. ABSTRACT

A fundamental answer is attempted to the question of being, time, space, existence, perception, and physical phenomena. Mass, empty space, reality, gravitational attraction, continuity, discontinuity, and relativity are among the "things" and "nonthings" for which relationships and explanations are provided.

Perception is defined as detection by a mass of change to itself and as a differentiating process. Perceived physical phenomena are revealed to be first derivatives of ultimate reality, and outputs of the physical perceiving device called the ✓ "perceptron." Using the operation of the physical perception process on action, the units of which are energy times time, an entirely different view is generated of both reality and the Heisenberg uncertainty principle. When superposition is imposed as a condition, the Heisenberg uncertainty principle is revealed to be a quantized statement. A concept of the absolute value of a dimensional molecule is advanced.

A reality model is constructed of an action continuum and a nonaction continuum separated by an operational threshold whose magnitude is given by an uncertainty principle. Perception is modeled as a process which switches atomic quanta of action and nonaction back and forth across the threshold. Mass is explained as perceptron switch activity rate, and a mass definition equation is generated. One kilogram mass is shown to represent 17.053×10^{50} perceptron switches per second. From the model, Newton's laws of motion and gravity can be generated directly from the concept of the interaction of mass with space itself, space being taken as a massless fluid, or ether, composed of tiny nonaction particles called "quitons." Einstein's postulates of

composed of tiny nonaction particles called "quitons." Einstein's postulates of

DD FORM 1 NOV 65 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

UNCLASSIFIED
Security Classification

Next Page

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
causality						
perception						
perceptron						
quiton						
quiton/perceptron theory						
gravitation						
mass						
reality						
metaphysics						
being						
existence						
space						
time						
photon						
action						
action quantum						
quantum						
observation						
relativity						
Karma						
complementarity						
physical phenomena						
Mach's principle						
Newton's laws						
inertia						
Debroglie wave						
Einstein's postulates						
Michelson-Morley experiment						
ontology						
Heisenberg Uncertainty Principle						
superposition						
equivalence principle						
simultaneity						
inertial frame						

ABSTRACT (Continued)

special relativity are also shown to follow from the model, and a statement of the equivalence principle is shown to be consistent with the model.

The concept of "lineception" as a causal chain of perceptron operations is advanced. "Inception" is developed as a noncausal modulation of lineception and as a distinguishing feature of a living system. The possibility is raised that inception could provide an explanation of the noncausal, statistical, ultramicroscopic universe and yet require large ensembles of ultramicroscopic events to conform to causality. Inception is shown to lead to karma. Perception is shown to be causal and to generate causality itself.

Some elementary philosophical implications of quiton/perceptron theory are pointed out and briefly discussed.

and briefly discussed.

Next Page

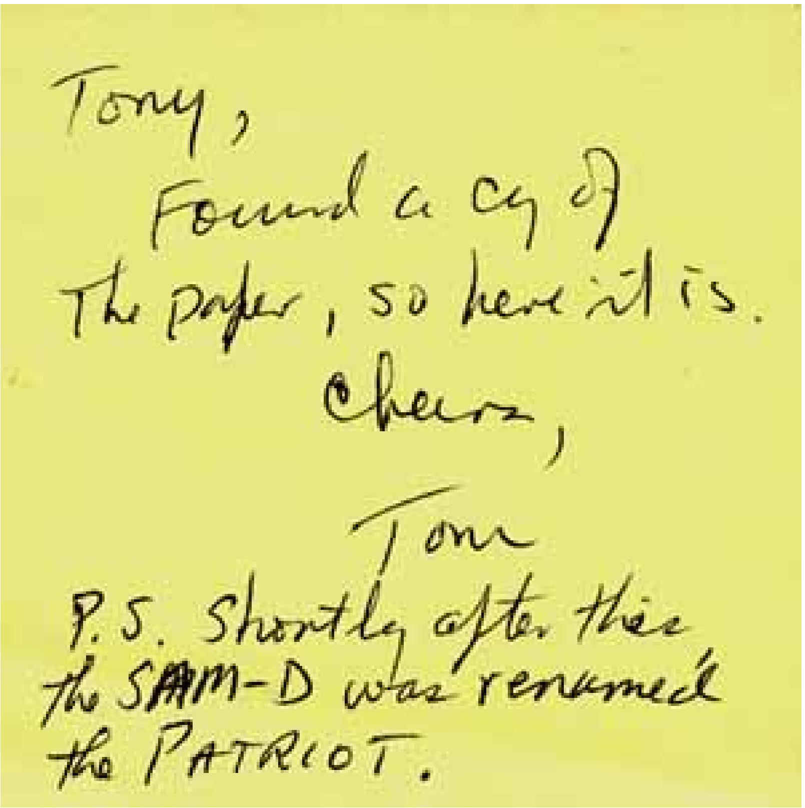

Tony,

Found a cy of
The paper, so here it is.

Cheers,

Tom

P.S. Shortly after this
the SAM-D was renamed
the PATRIOT.

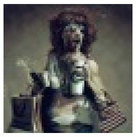

SCALAR WEAPONS: Read It and Weep

- Posted by Ms MoD on July 27, 2011 at 4:42 in HAARP
- View Discussions

This is a REALLY long paper but it is broken down into categories so you can easily skip to sections of interest to you. I found this after reading Twelve Things You Should Know About Scalar Weapons which was posted by FDF in April and is a MUCH SHORTER overview of scalar technology by the same author. I highly recommend it but wanted to post the longer version here as well which you can download in the PDF attached. Topics in the paper are listed below -MoD

SCALAR WEAPONS: Read It and Weep
By C Verismo

1. THE DISCOVERY OF SCALAR WAVES
2. BEAM WEAPONS OF A CENTURY AGO
3. THE CONTINUATION OF TESLA'S SCALAR WAVE TECHNOLOGY
4. TESLA'S CROSSED SCALAR BEAMS
5. RUSSIA'S 'WOODPECKER' SCALAR WAVE TRANSMITTERS
6. THE 1930s
7. THE 1940s
8. THE 1950s
9. THE 1960s
10. THE 1970s
11. THE 1980s
12. THE 1990s
13. THE PRESENT
WHAT CAN BE DONE WITH SCALAR TECHNOLOGY?
14. MIND READING, MASS HYPNOTISM AND SYNTHETIC TELEPATHY
15. DISEASE INDUCTION
16. DISABLING SATELLITES
17. IN THE AIR
18. WEATHER ENGINEERING
19. EARTHQUAKES
20. SEA ATTACKS
21. TESLA SHIELDS AND GLOBES
22. GROUND ATTACKS
23 OBLITERATING INDIVIDUALS
24. PORTABLE SCALAR BAZOOKAS
25. RADAR INVISIBILITY
26. SUMMARY OF A WAR THAT COULD OCCUR
27. RUSSIAN UFOs ATTACKED THE US SHUTTLE PROGRAM FROM THE BEGINNING
28. HOW DID RUSSIA GET UFOs?
29 FORMER BOLSHEVIKS RUNNING AMERICA
30 REPLICAS USED FOR THE LANDINGS, AFTER ORIGINALS SHOT DOWN
31. CLOAKED RUSSIAN UFOS ABOVE USA

32. ALL U.S. SATELLITES SHOT DOWN BY RUSSIA UNTIL 1981
33. LIES, LIES AND LIES
34. FRICTION WITH THE JAPANESE
35. ENTERPRISE SHUTTLE DETERMINED TO SPY ON RUSSIA
36. RUSSIA INTENSIFIES ATTACKS
37. TRYING AGAIN TO GET UP INTO SPACE. THIRD TIME LUCKY?
38. BOLD ATTACKS ON AMERICA
39. THE STRUGGLE AT THE TOP FOR POWER
40. THE STEALTH PLANE COMES TO THE RESCUE
41. PHANTOM AGAINST THE COSMOSPHERE
42. THE EXPECTED GENOCIDE
43. PLANS FOILED TO DESTROY RUSSIA'S MILITARY ARSENAL
44. BRAZIL
45. INDIA
46. JAPAN
47. KOBE 'EARTHQUAKE'
48. CHINA
49. OKLAHOMA FBI BUILDING 'BOMBING'
50. WHO ELSE HAS UFOs?

Full paper in PDF below…

Favorite
1 member favorited this

Share Tweet Facebook

Views: 459

Attachments:

- SCALAR WEAPONS - Read It and Weep.pdf, 494 KB

▶ Reply to This

Replies to This Discussion

Permalink Reply by Jenny on July 27, 2011 at 6:22
HAARP Patents (Assigned to APTI, Inc. Los Angeles, CA, Washington, DC)

U.S. Patent 4686605: Method And Apparatus For Altering A Region In The Earth’s Atmosphere, Ionosphere, And/Or Magnetosphere Issued: Aug. 11, 1987 Filed: Jan. 10, 1985
http://www.freepatentsonline.com/4686605.pdf

U.S. Patent 5038664: Method For Producing A Shell Of Relativistic Particles At An Altitude Above The Earth’s Surface ~ Issued: Aug. 13, 1991 Filed: Jan. 10, 1985
http://www.freepatentsonline.com/5038664.pdf

U.S. Patent 4712155: Method And Apparatus For Creating An Artificial Electron Cyclotron Heating

Region Of Plasma ~ Issued: Dec. 8, 1987 Filed: Jan. 28, 1985
http://www.freepatentsonline.com/4712155.pdf

U.S. Patent 5068669: Power Beaming System ~ Issued: Nov. 26, 1991 Filed: Sep. 1, 1988
http://www.freepatentsonline.com/5068669.pdf

U.S. Patent 5218374: Power Beaming System With Printer Circuit Radiating Elements Having Resonating Cavities ~ Issued: June 8, 1993 Filed: Oct. 10, 1989
http://www.freepatentsonline.com/5218374.pdf

U.S. Patent 5293176: Folded Cross Grid Dipole Antenna Element ~ Issued: Mar. 8, 1994 Filed: Nov. 18, 1991
http://www.freepatentsonline.com/5218374.pdf

U.S. Patent 5202689: Lightweight Focusing Reflector For Space ~ Issued: Apr. 13, 1993
http://www.freepatentsonline.com/5202689.pdf
Filed: Aug. 23, 1991

U.S. Patent 5041834: Artificial Ionospheric Mirror Composed Of A Plasma Layer Which Can Be Tilted ~ Issued: Aug. 20, 1991 Filed: May. 17, 1990
http://www.freepatentsonline.com/5041834.pdf

U.S. Patent 4999637: Creation Of Artificial Ionization Clouds Above The Earth ~ Issued: Mar. 12, 1991 Filed: May. 14, 1987
http://www.freepatentsonline.com/4999637.pdf

U.S. Patent 4954709: High Resolution Directional Gamma Ray Detector ~ Issued: Sep. 4, 1990 Filed: Aug. 16, 1989
http://www.freepatentsonline.com/4954709.pdf

U.S. Patent 4817495: Defense System For Discriminating Between Objects In Space ~ Issued: Apr. 4, 1989 Filed: Jul. 7, 1986
http://www.freepatentsonline.com/4817495.pdf

U.S. Patent 4873928: Nuclear-Sized Explosions Without Radiation ~ Issued: Oct. 17, 1989 Filed: June 15, 1987
http://www.freepatentsonline.com/4873928.pdf

Still think that HAARP is just another wacky fictional conspiracy? The above listed patents should be enough to convince any intelligent person that HAARP is real and is now actively modifying our weather and triggering earthquakes. For those who are still not convinced, look to the United States Congress for proof. In order for HAARP to get funding the people who first created HAARP needed funding. That funding was made possible by the US government. If you were to look at Congressional records you would find documents that states just one of the many official purposes of HAARP – for penetrating the earth with signals bounced off of the ionosphere. Congress was informed by people in the High Frequency Active Auroral Research Program that the bounced signals would be used to look inside the planet to a depth of many kilometers in order to locate underground munitions, minerals and tunnels. If you bounced the known earthquake signature signal (2.5 Hz) for a sustained period of time an earthquake could triggered many kilometers inside Earth’s crust. The U.S. Senate actually set aside $15 million dollars in 1996 (Clinton and Al Gore Administration) to develop this ability alone — earth-penetrating-tomography.

Update April 21, 2011

Under pressure from the International community the HAARP website is up again. No data is available from April 7 to April 12, 2011. Data before and after shows the signature frequency of an earthquake (2.5Hz) being transmitted. Also shows energy spikes which indicates active weather modification. Now that the website is up again you can view the data for the Japan earthquake at http://137.229.36.30/cgi-bin/scmag/disp ... i?20110311. From the magnetometer induction data you can see 2.5Hz (earthquake inducing frequency) being broadcasted and recorded for several days before the 9.0 magnitude earthquake. To see what a normal (no HAARP broadcasting) day looks like view the magnetometer charts for the days of March 15, 16 and 17 – March 16, 2011 http://137.229.36.30/cgi-bin/scmag/disp ... 0316&Bx=on. No earthquake inducing broadcast was detected from March 15 to March 27, 2011. On March 28, 2011 a low powered broadcast was detected from approximately 12:00 am to 3:00 am. Broadcasting resumed on April 2, 2011. April 6, 2011 shows you what the magnetometer charts looks like when a solar storm blasts Earth. No data is available from April 7, 2011 to April 13, 2011 – a shutdown. Starting on April 15, 2011 and ending approximately 10:00 am April 16, 2011 the earthquake inducing frequency of 2.5Hz was detected being broadcasted.
Short URL: http://presscore.ca/2011/?p=2261

- ▶ Reply
-

Permalink Reply by Jenny on July 27, 2011 at 6:23
WOW MoD! What a paper! I need like a year to get through it all but it sure is awesome to have it all in one place! Excellent!

- ▶ Reply
-

Permalink Reply by Ms MoD on October 26, 2012 at 6:38

Boeing Tests Microwave Missile That Can Knock Out All Electronics (CHAMP)
Huffington Post UK | By Michael Rundle 24/10/2012

Boeing is testing a missile which can destroy electronics in a wide area with little or no physical damage.

The 'Champ' (Counter-electronics High-powered Advanced Missile Project) may "one day change warfare", Boeing said.

"This technology marks a new era in modern-day warfare," said Keith Coleman, CHAMP program manager for Boeing Phantom Works.

"In the near future, this technology may be used to render an enemy's electronic and data systems useless even before the first troops or aircraft arrive".

The missile was tested on 16 October, when the high-powered microwaves were fired at a two storey building built especially for the job.

The missile would, if successful, take out rows of personal computers which were turned on to gauge the effect of the blast.

The result?

"Seconds later the PC monitors went dark and cheers erupted in the conference room. CHAMP had successfully knocked out the computer and electrical systems in the target building. Even the television cameras set up to record the test were knocked off line without collateral damage."

Seven targets were hit in the one-hour test, and all were successfully disabled.

There is a long way to go before the microwave can be placed into a missile and used to disable targets without loss of life, but the team is hopeful it could prove to be a milestone in non-lethal warfare.

"Today we turned science fiction into science fact," Coleman said.

http://www.huffingtonpost.co.uk/2012/10/24/boeing-tests-microwave-m...

- ▶ Reply
-

Permalink Reply by Ms MoD on August 3, 2013 at 11:01

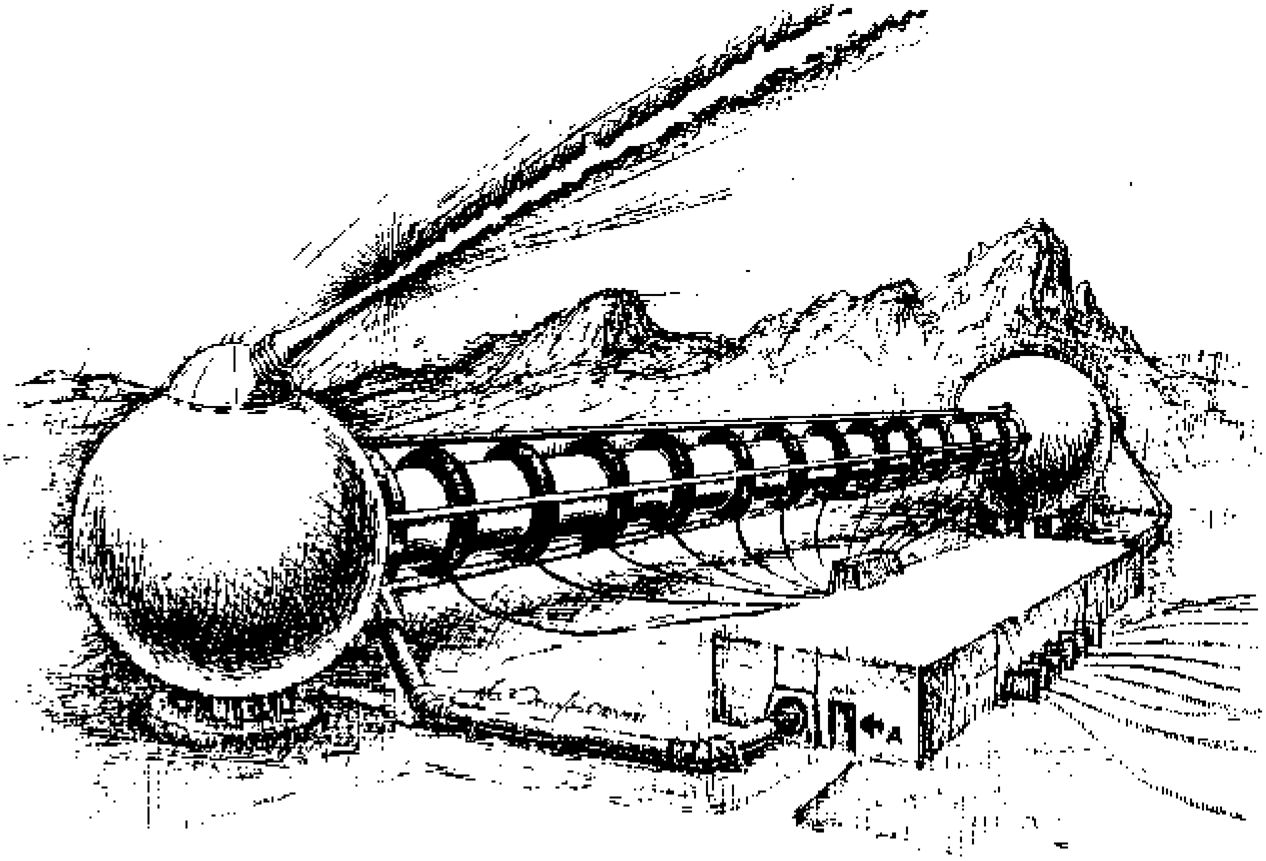

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

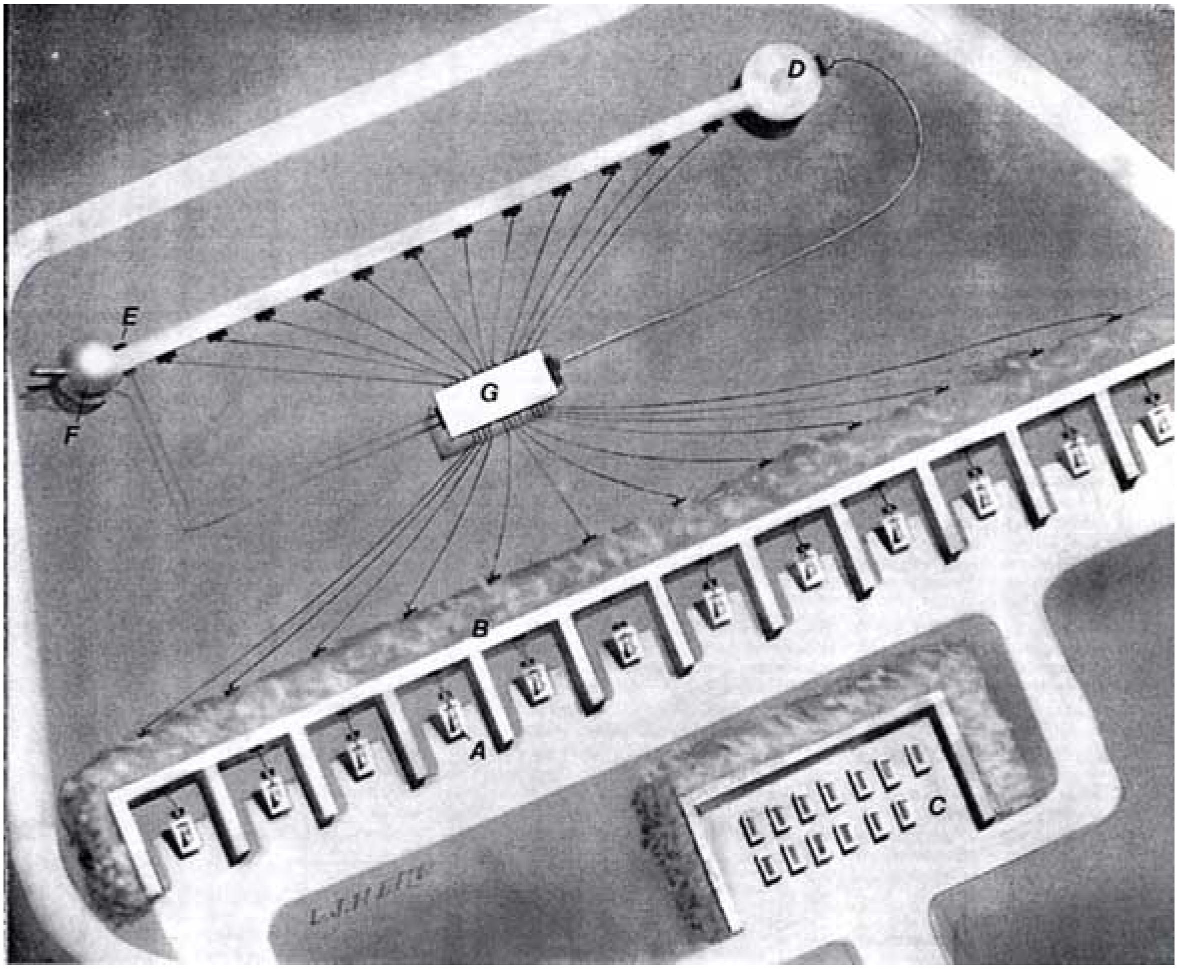

Aviation Week & Space Technology July 28, 1980

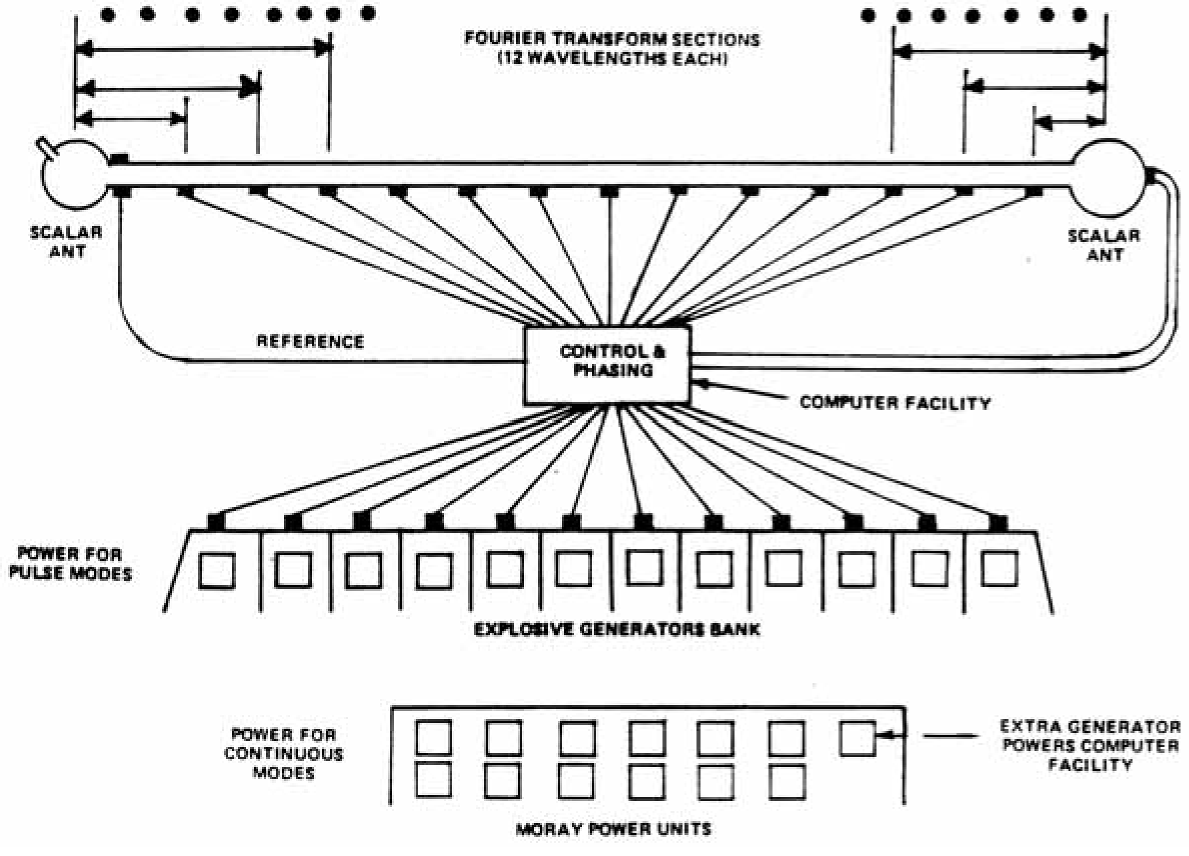
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

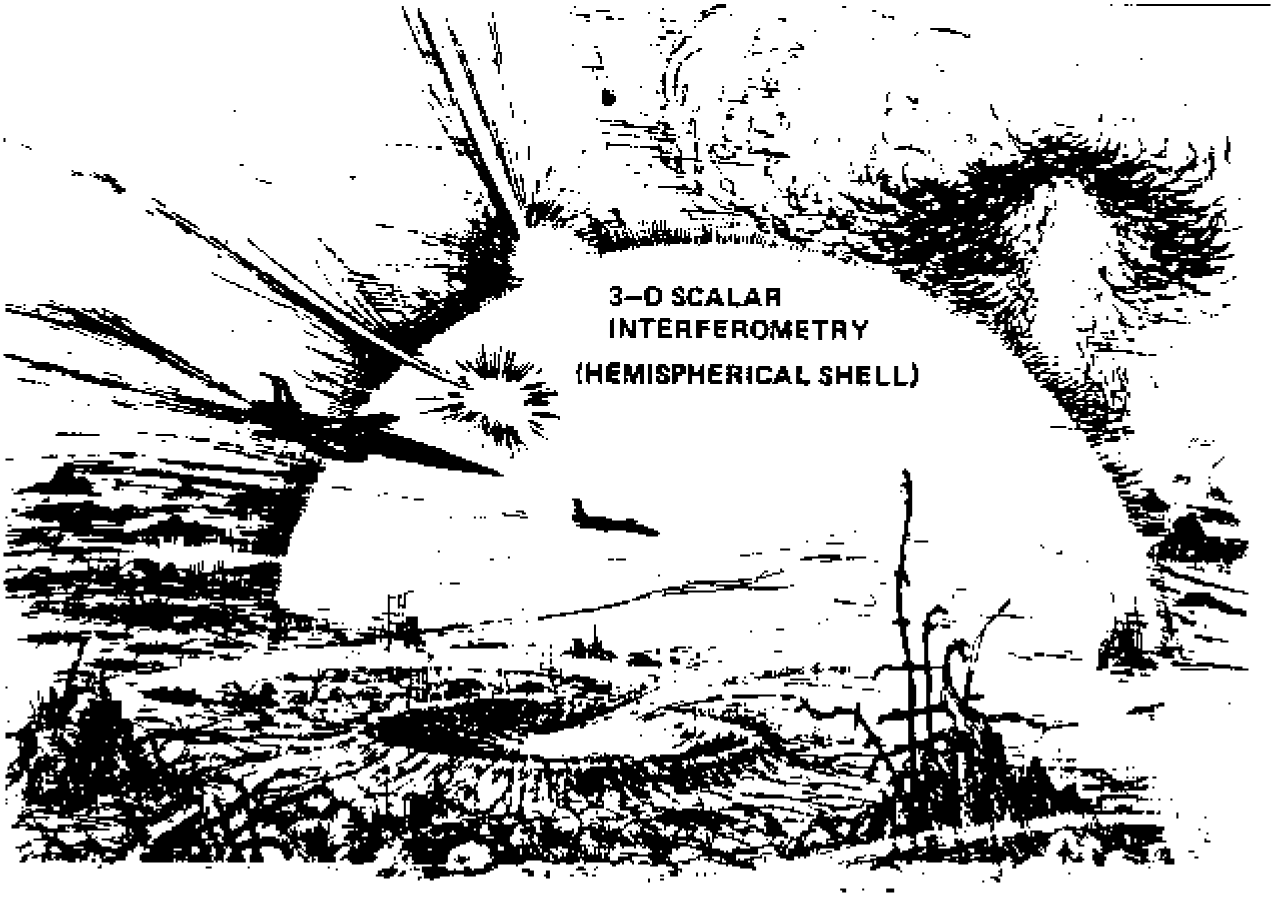

Figure 8. The Tesla Shield

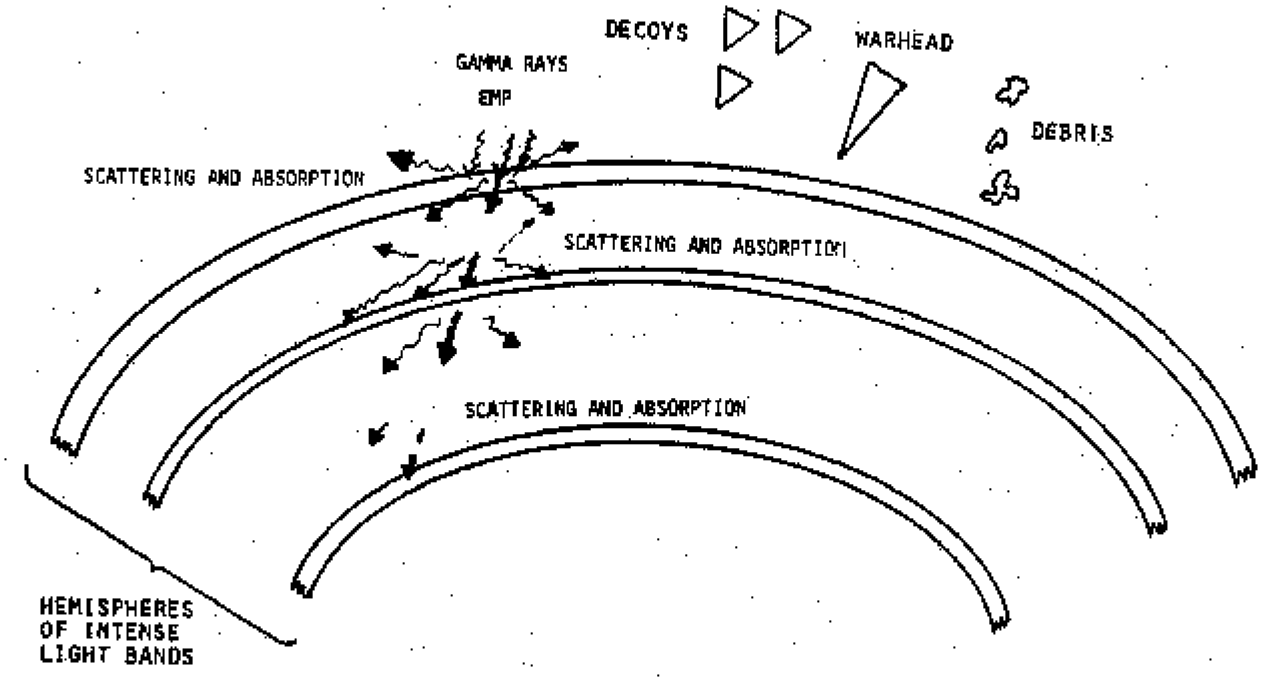
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

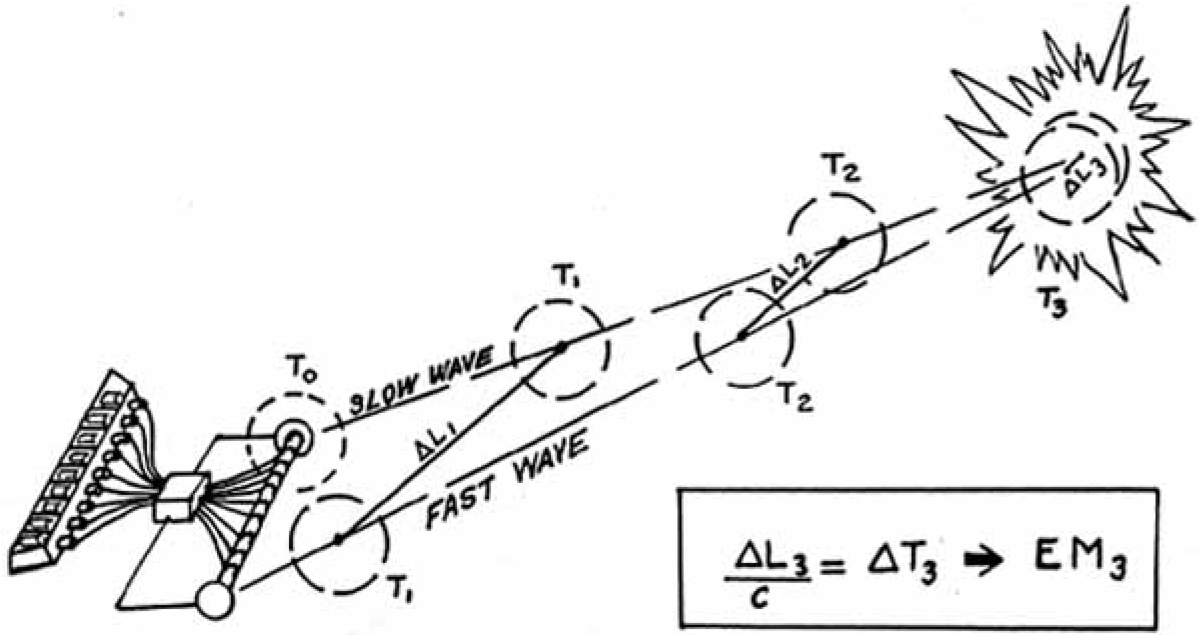
T₀
SLOW WAVE
T₁
ΔL₁
T₂
ΔL₂
ΔL₃
FAST WAVE
T₁
T₂
T₃
ΔL₃/c = ΔT₃ → EM₃

TESLA HOWITZER
(SCALAR INTERFEROMETER)

Figure 10. "Nuclear" Flashes off the Coast of Africa

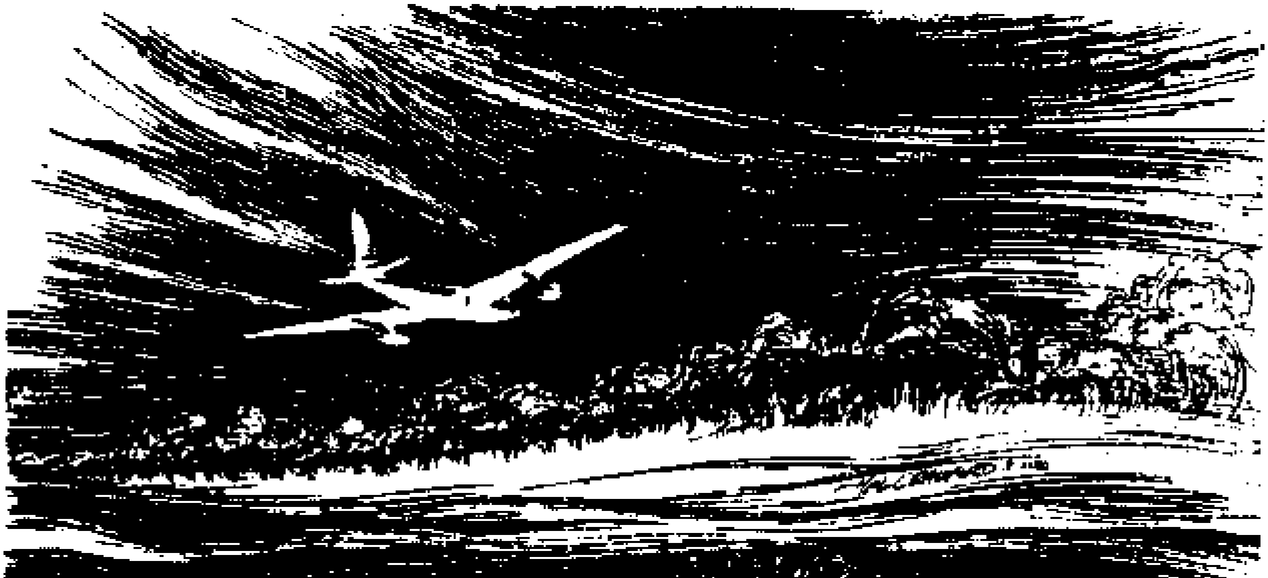

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

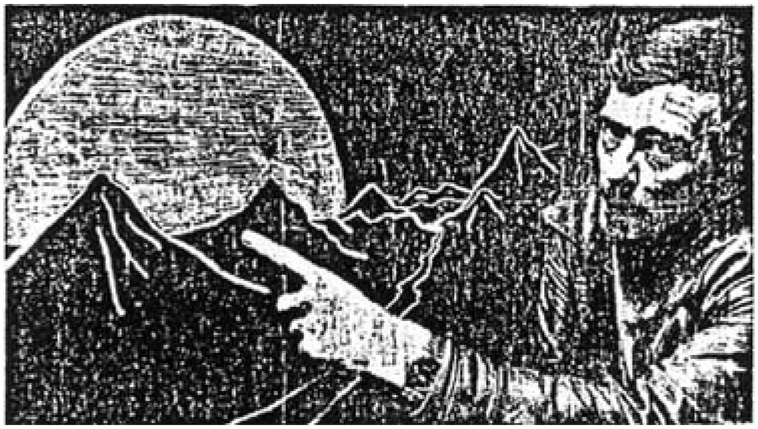

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

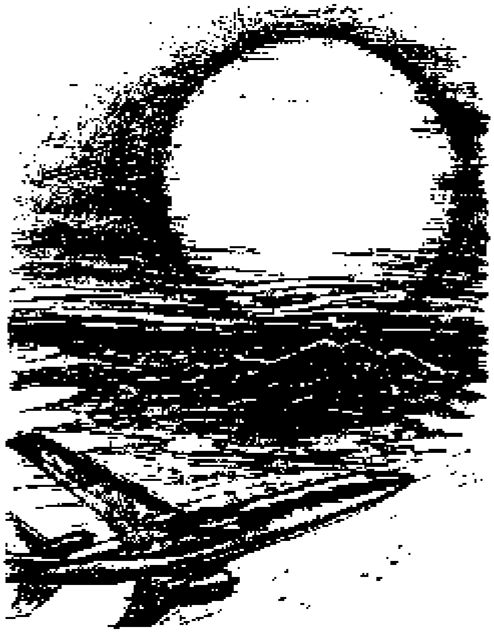

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

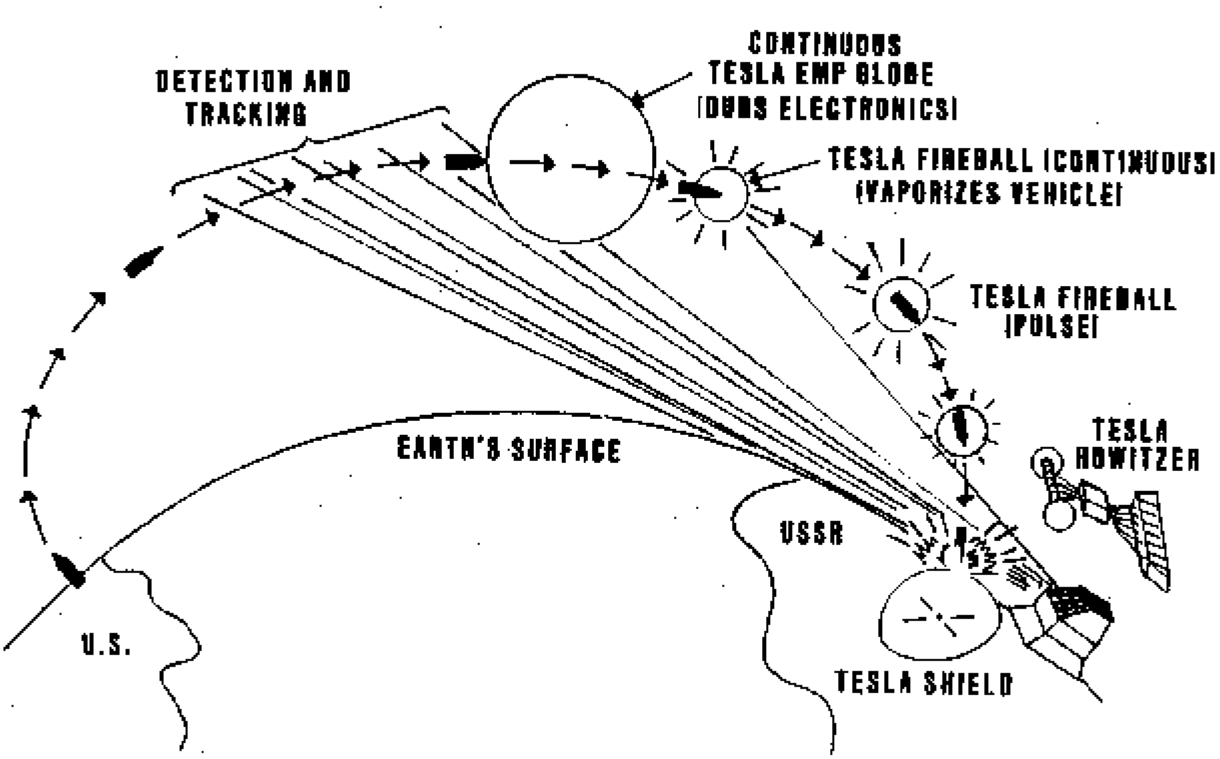
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

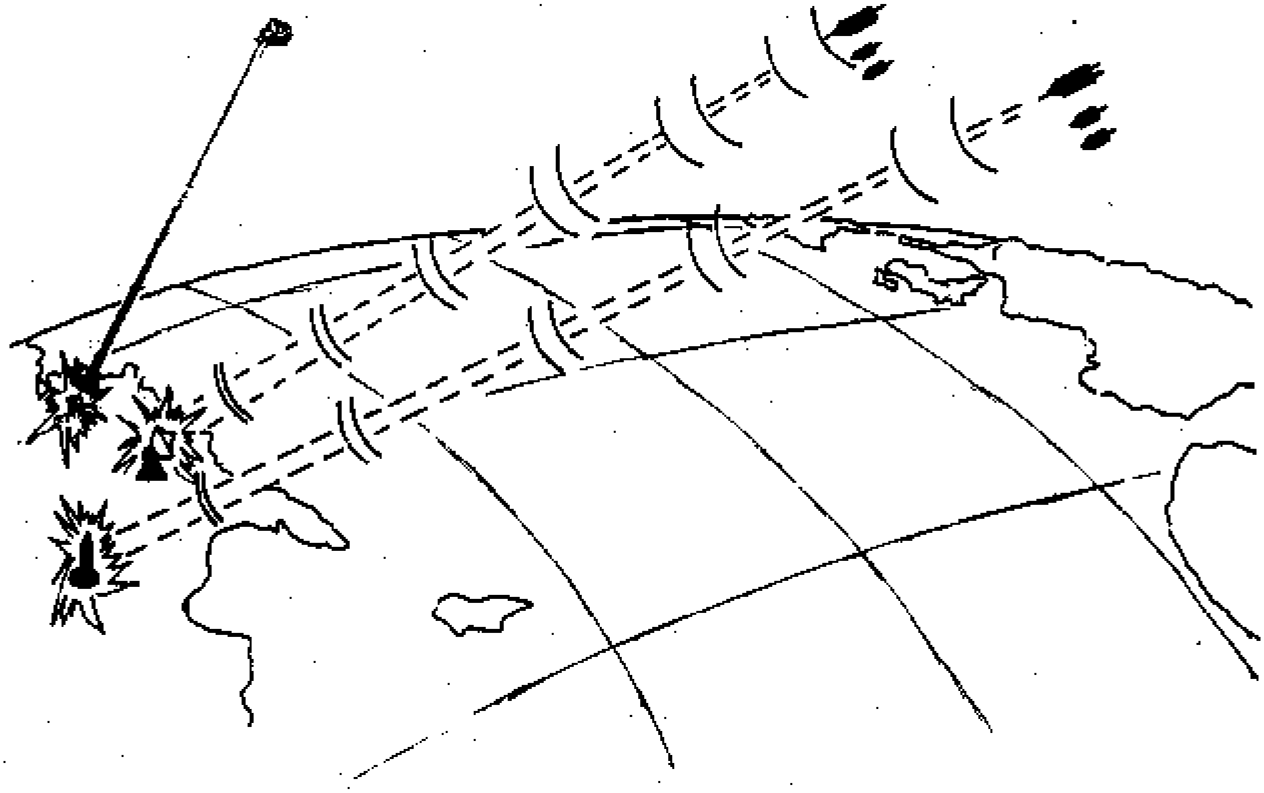

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **PAIR-COUPLED SCALARS**
  - **VELOCITY-LIMITED TO C**
  - **CARRIERS OF T**

- **PHOTON INTERACTION**
  - **IS UBIQUITOUS**
  - **PRODUCES $\dot{T}$ = C**
  - **YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **NOT VELOCITY-LIMITED**
  - **YIELD HIGHER ORDER REALITIES**

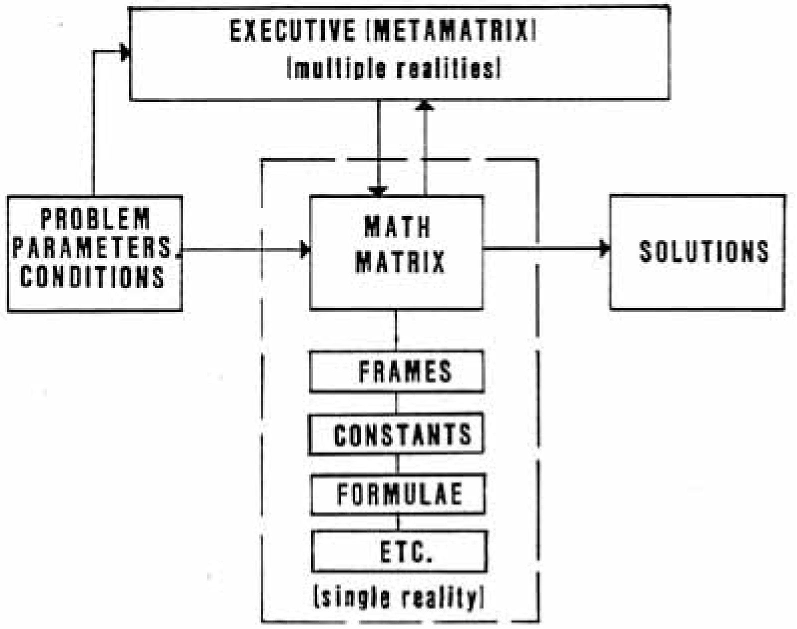

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

http://www.cheniere.org/books/part1/teslaweapons.htm

- ▶ Reply
-

Permalink Reply by rose on August 4, 2013 at 10:54

yeah this ones from deep in the archives Greg - I would love to see that article!!!

- ▶ Reply
-

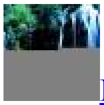
Permalink Reply by deanna munson on August 4, 2013 at 18:16

what if these weapons work in conjunction with cell towers and smartgrid/wifi relays? i just read this and it blew me away how the dangers this is being purposely ignored http://www.stayonthetruth.com/electromagnetic-fields---killing-fiel... "

November 14 1996, Omnipoint, New York City's first digital cellular provider, did open for business, broadcasting from thousands of antennae newly erected on the rooftops of apartment buildings. According to the health authorities, an early flu hit New York City - but not Boston, and not Philadelphia - on about 15 November. The flu was severe and ran a prolonged course, often dragging on for months instead of the usual two weeks.

At Christmas time, the Cellular Phone Task Force placed a small classified ad in a free weekly newspaper. It read: 'If you have been ill since 11/15/96 with any of the following: eye pain, insomnia, dry lips, swollen throat, pressure or pain in the chest, headaches, dizziness, nausea, shakiness, other aches and pains, or flu that won't go away, you may be a victim of a new microwave system blanketing the city. We need to hear from you.' And we did hear from them. Hundreds called, men, women, whites, blacks, Asians, Latinos, doctors, lawyers, teachers, stockbrokers, airline stewards, computer operators. Most had woken up suddenly in mid-November, thinking they were having a stroke or a heart attack or a nervous breakdown, and were relieved to know they were not alone and not crazy.

Later, I analysed weekly mortality statistics, which the Centres for Disease Control publish for122 US cities. Each of dozens of cities recorded a 10-25 per cent increase in mortality, lasting two to three months, beginning on the day in 1996 or 1997 on which that city's first digital cell phone network began commercial service." we already know these agendas are being forced everywhere.no one is even noticing how even though everyone is already "connected" they are still erecting more and more towers and infastructure,by force in the case of smartmeters.wired in connections now cost over double what wifi does.they just put up another tower here on my block,disguised as a flipping pine tree.cant block the effects when it comes from literally all sides

- ▶ Reply
-

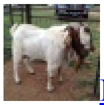
Permalink Reply by Rhonda Reichel on August 6, 2013 at 6:15

Scalar weapons makes all the billions we spend on war obsolete.....what a racket

- ▶ Reply
-

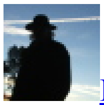
Permalink Reply by steve on August 9, 2013 at 6:43
having read a few pages, it looks like a weapon of distraction to me. The physics is in the realm of the tooth fairy, and the history has the ring of recent manufacture

- ▶ Reply
-

Permalink Reply by rose on August 9, 2013 at 12:24

PLEASE UPDATE YOUR ACCOUNT TO ENABLE 3RD PARTY HOSTING
100%
3rd Party Hosting Usage
FOR IMPORTANT INFO, PLEASE GO TO:
www.photobucket.com/P500

- ▶ Reply
-

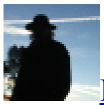
Permalink Reply by steve on August 9, 2013 at 16:03
plus, Tom Beardon is nothing but a fear merchant.

- ▶ Reply
-

RSS

Welcome to
The ConTrail

Sign Up
or Sign In

Or sign in with:

Visitors

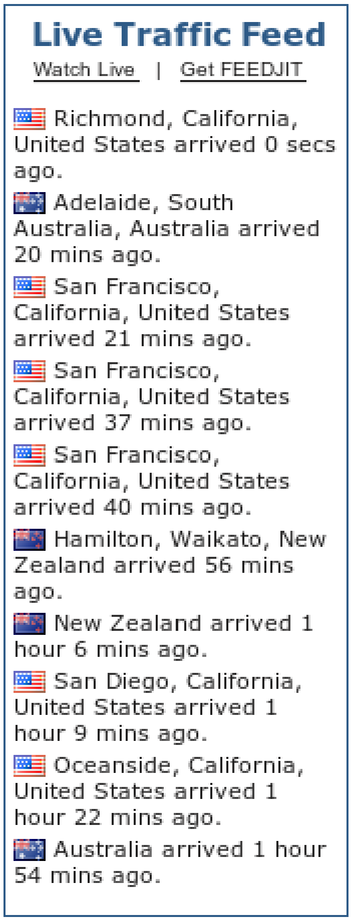

Live Traffic Feed

Watch Live | Get FEEDJIT

Richmond, California, United States arrived 0 secs ago.

Adelaide, South Australia, Australia arrived 20 mins ago.

San Francisco, California, United States arrived 21 mins ago.

San Francisco, California, United States arrived 37 mins ago.

San Francisco, California, United States arrived 40 mins ago.

Hamilton, Waikato, New Zealand arrived 56 mins ago.

New Zealand arrived 1 hour 6 mins ago.

San Diego, California, United States arrived 1 hour 9 mins ago.

Oceanside, California, United States arrived 1 hour 22 mins ago.

Australia arrived 1 hour 54 mins ago.

Groups

-

  ### Health and Wellbeing

  123 members

  99 Comments 10 Favorites

-

  ### Vaccination Information

  87 members

  163 Comments 6 Favorites

-

  ### Simply Good Stuff

  40 members

  112 Comments 3 Favorites
-

  ### EMF and EMR Information …

  101 members

  74 Comments 13 Favorites
-

  ### Cancer Treatment Options

  74 members

  77 Comments 9 Favorites
-

  ### Grow your own Food and M…

  96 members

  52 Comments 12 Favorites
-

  ### Childbirth

  3 members

  1 Comment 0 Favorites
- View All

© 2018 Created by rose. Powered by

Badges | Report an Issue | Terms of Service

Sign in to chat!

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

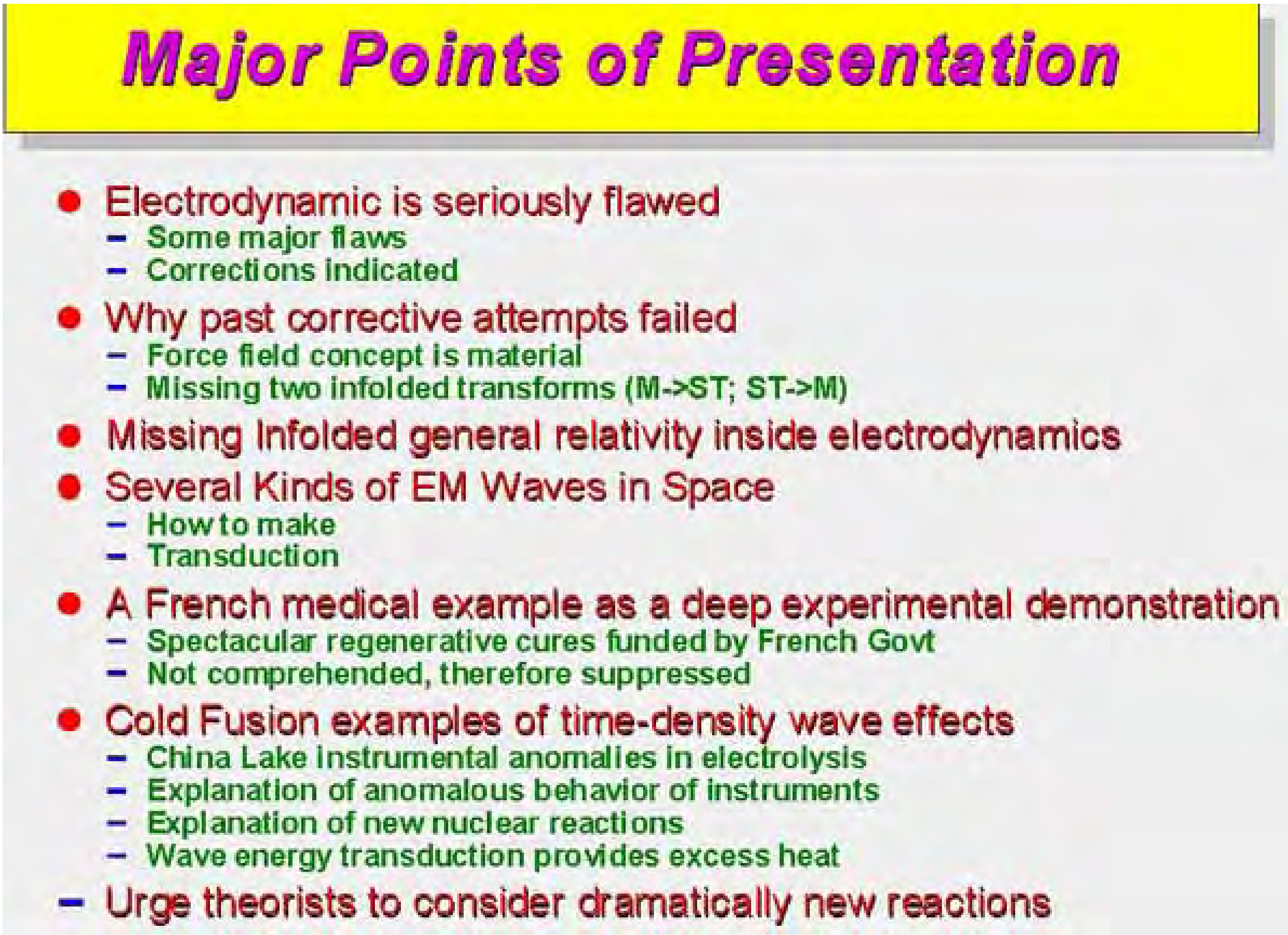

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

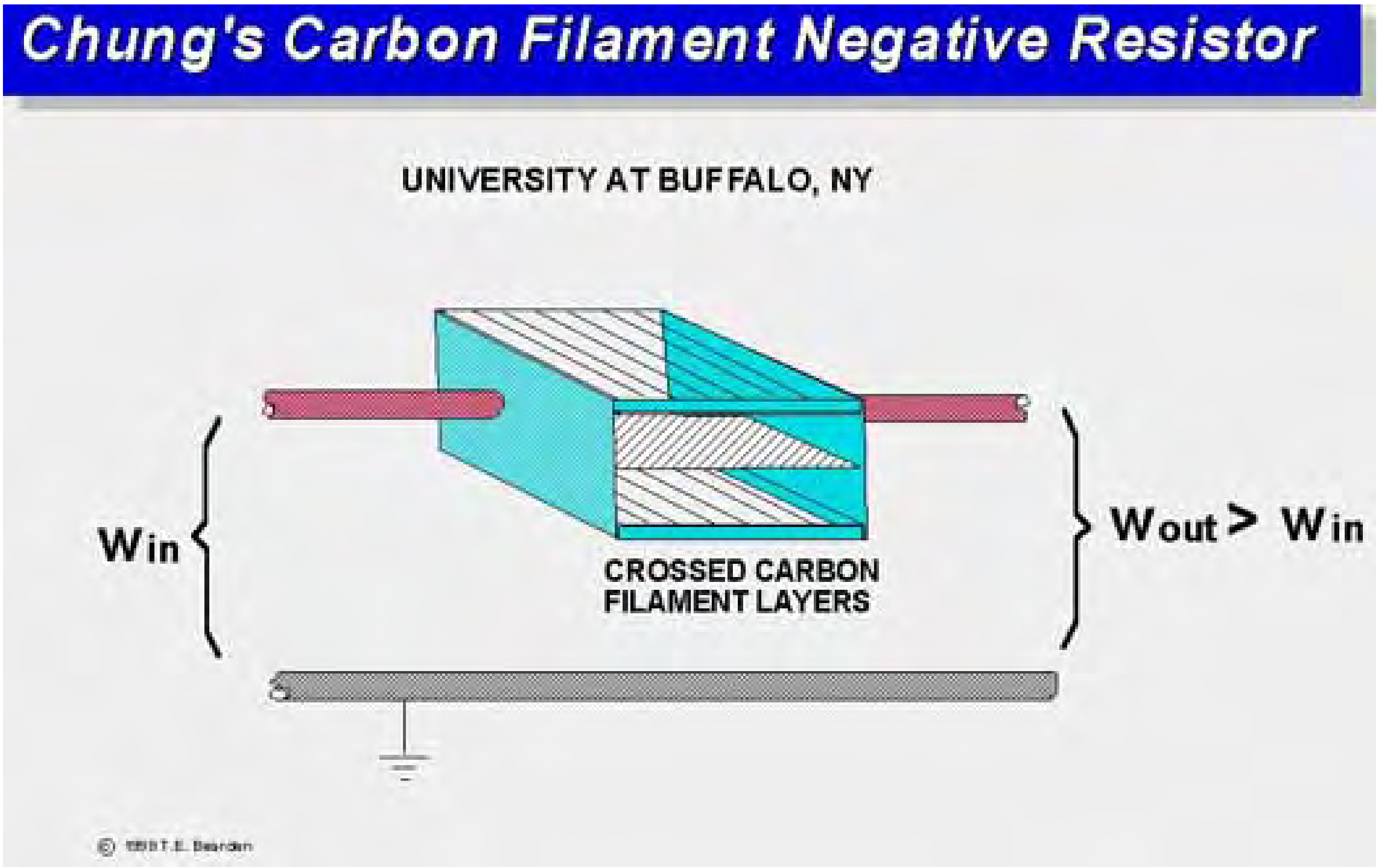

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

W_{in}

CROSSED CARBON
FILAMENT LAYERS

$W_{out} > W_{in}$

© [illegible] T.E. Bearden

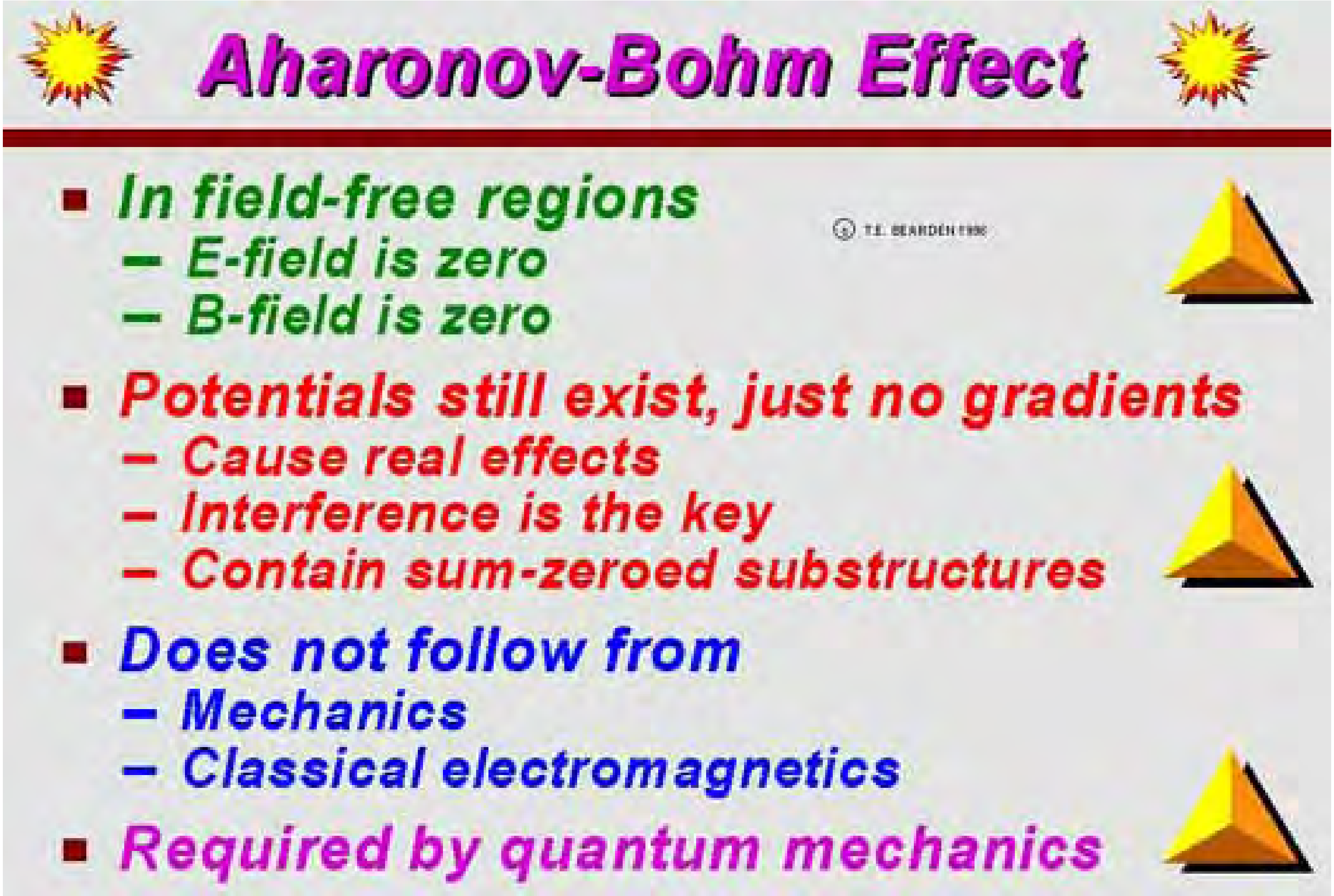

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- *In field-free regions*
  - *E-field is zero*
  - *B-field is zero*
- *Potentials still exist, just no gradients*
  - *Cause real effects*
  - *Interference is the key*
  - *Contain sum-zeroed substructures*
- *Does not follow from*
  - *Mechanics*
  - *Classical electromagnetics*
- *Required by quantum mechanics*

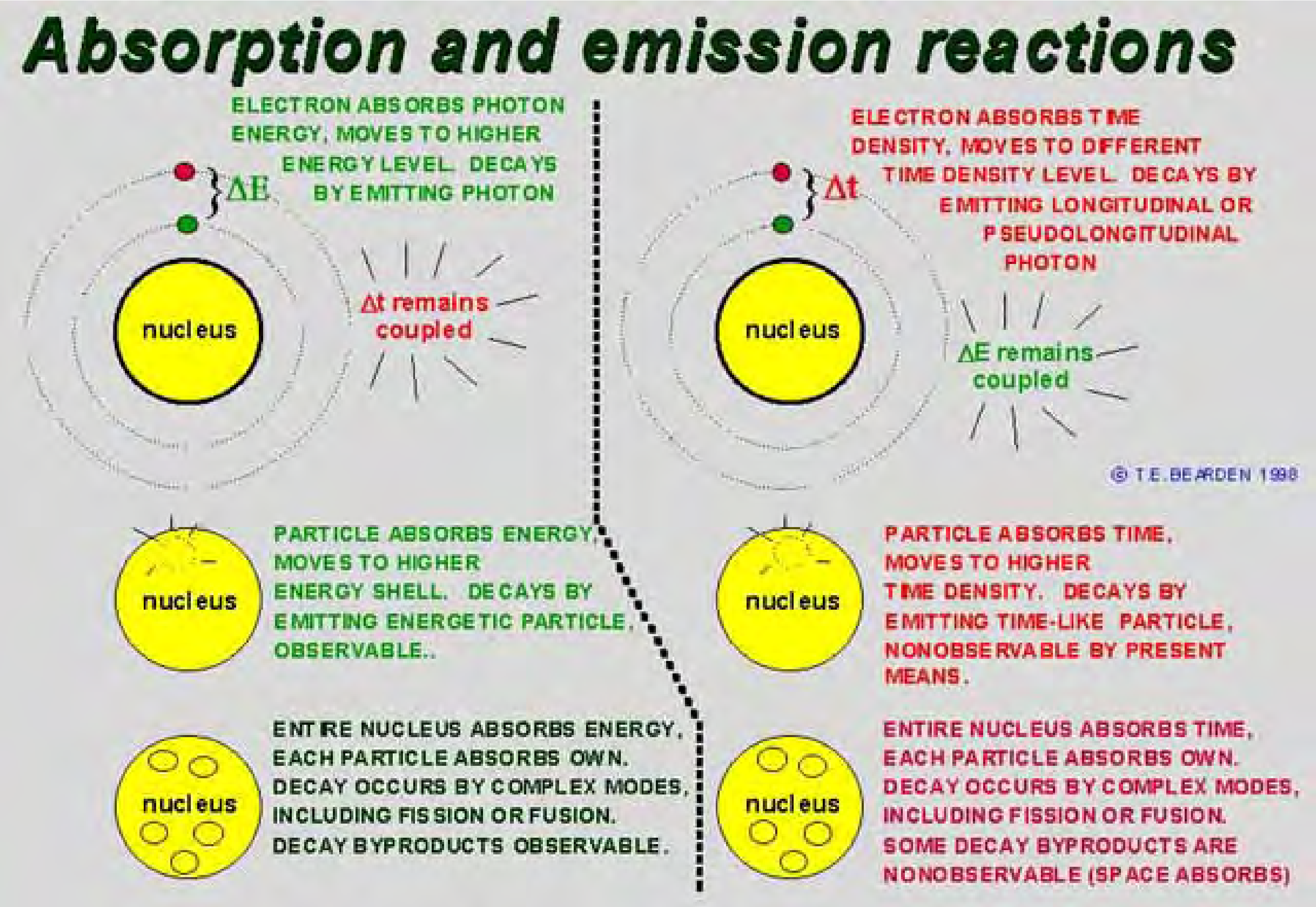

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

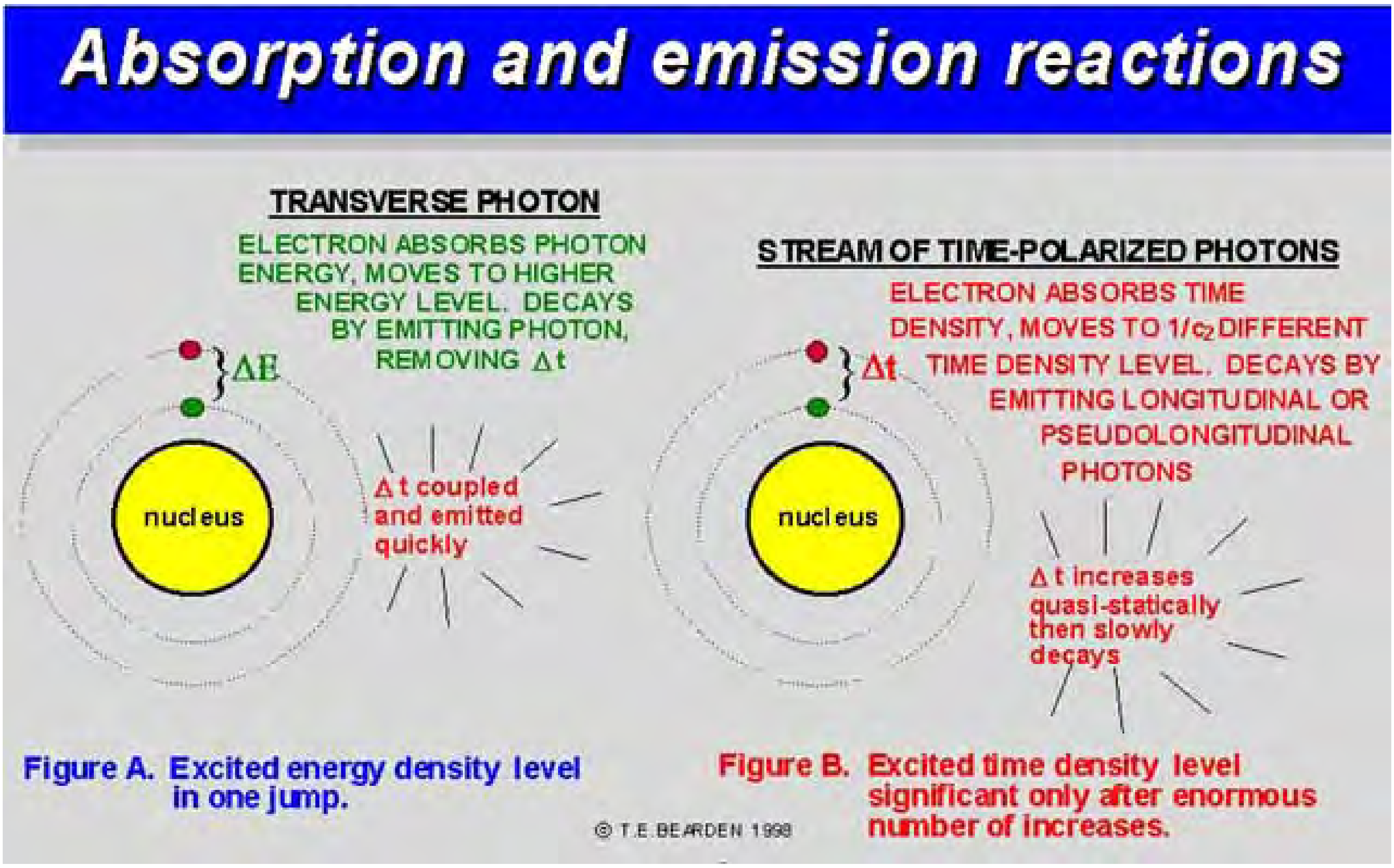

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c^2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure A. Excited energy density level in one jump.

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

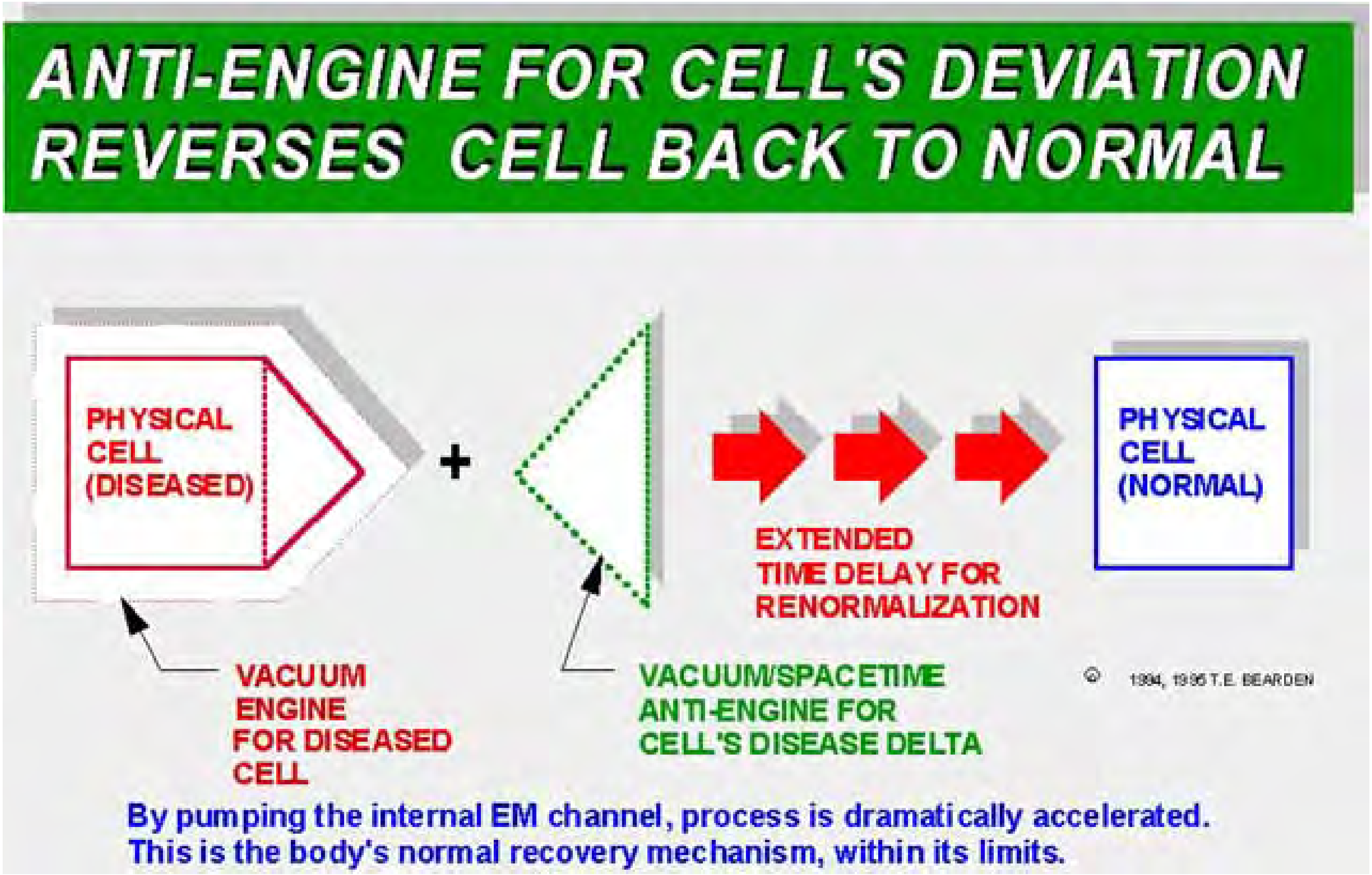

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL CELL (DISEASED)

+

EXTENDED TIME DELAY FOR RENORMALIZATION

PHYSICAL CELL (NORMAL)

VACUUM ENGINE FOR DISEASED CELL

VACUUM/SPACETIME ANTI-ENGINE FOR CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated. This is the body's normal recovery mechanism, within its limits.

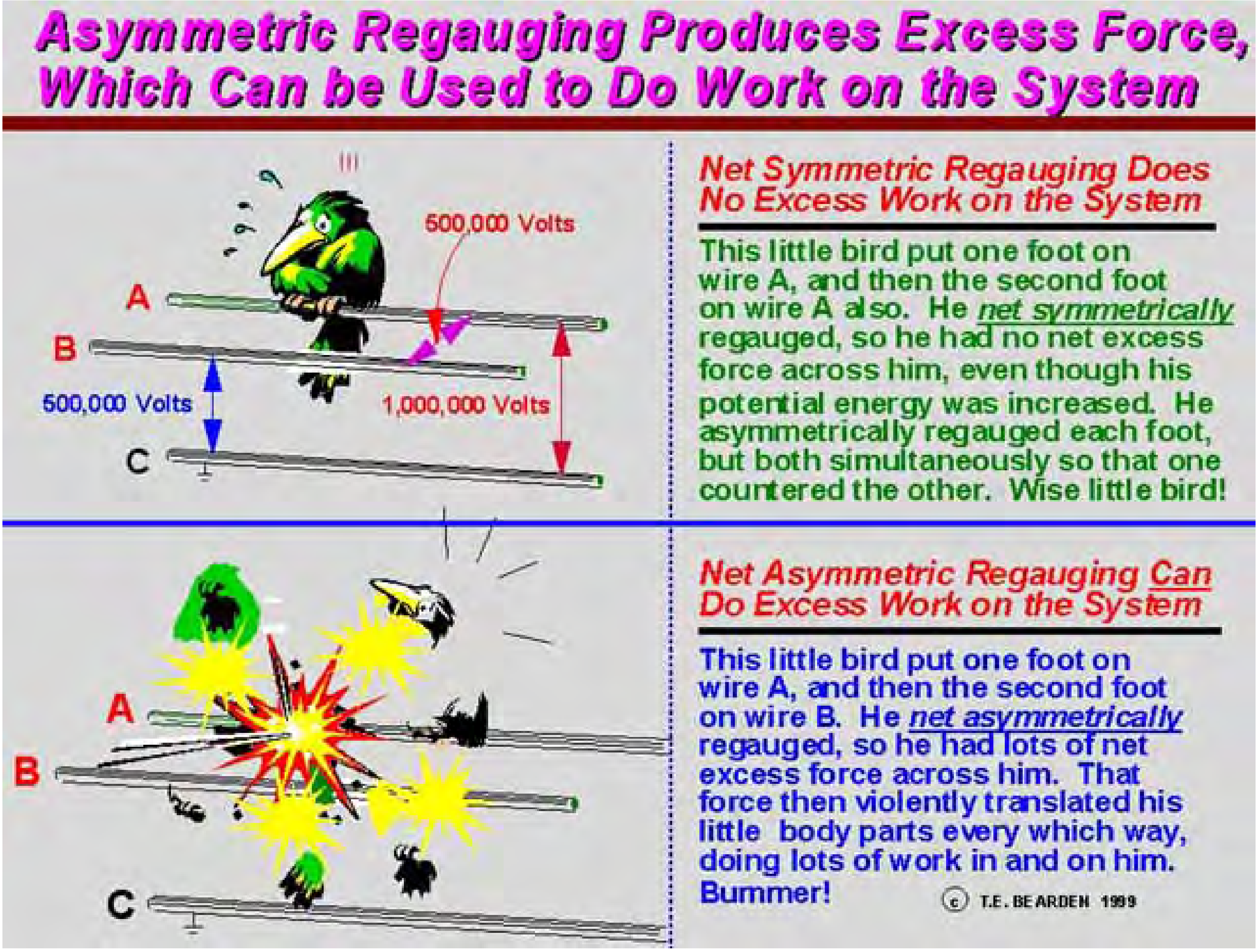

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

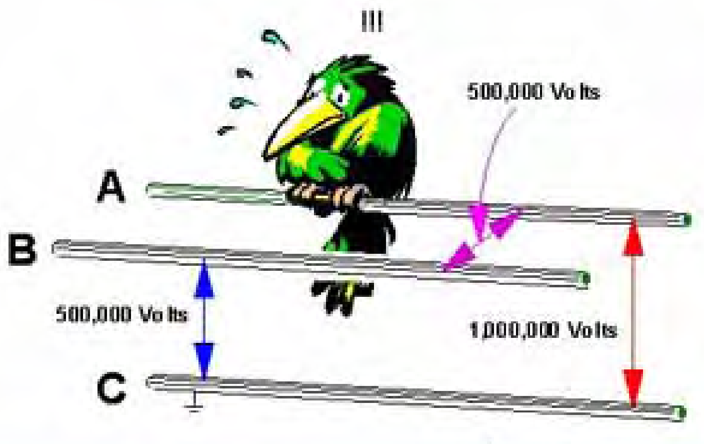

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

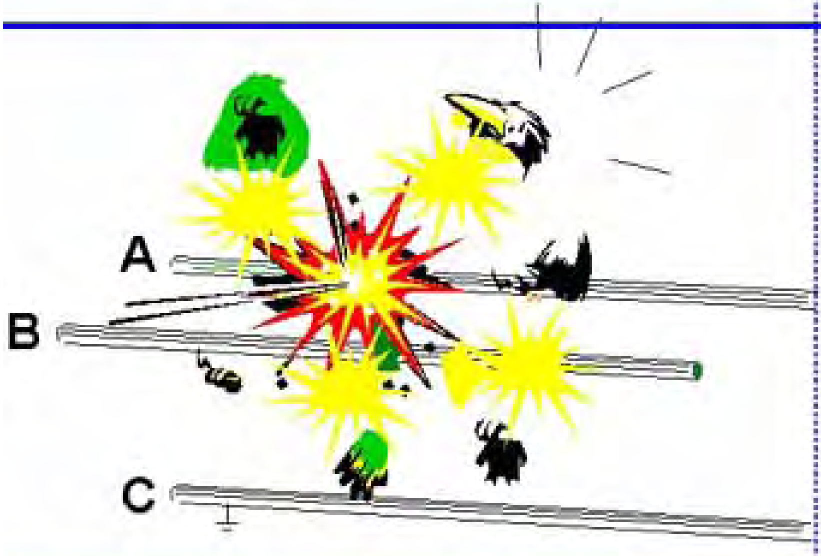

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

Ⓒ T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

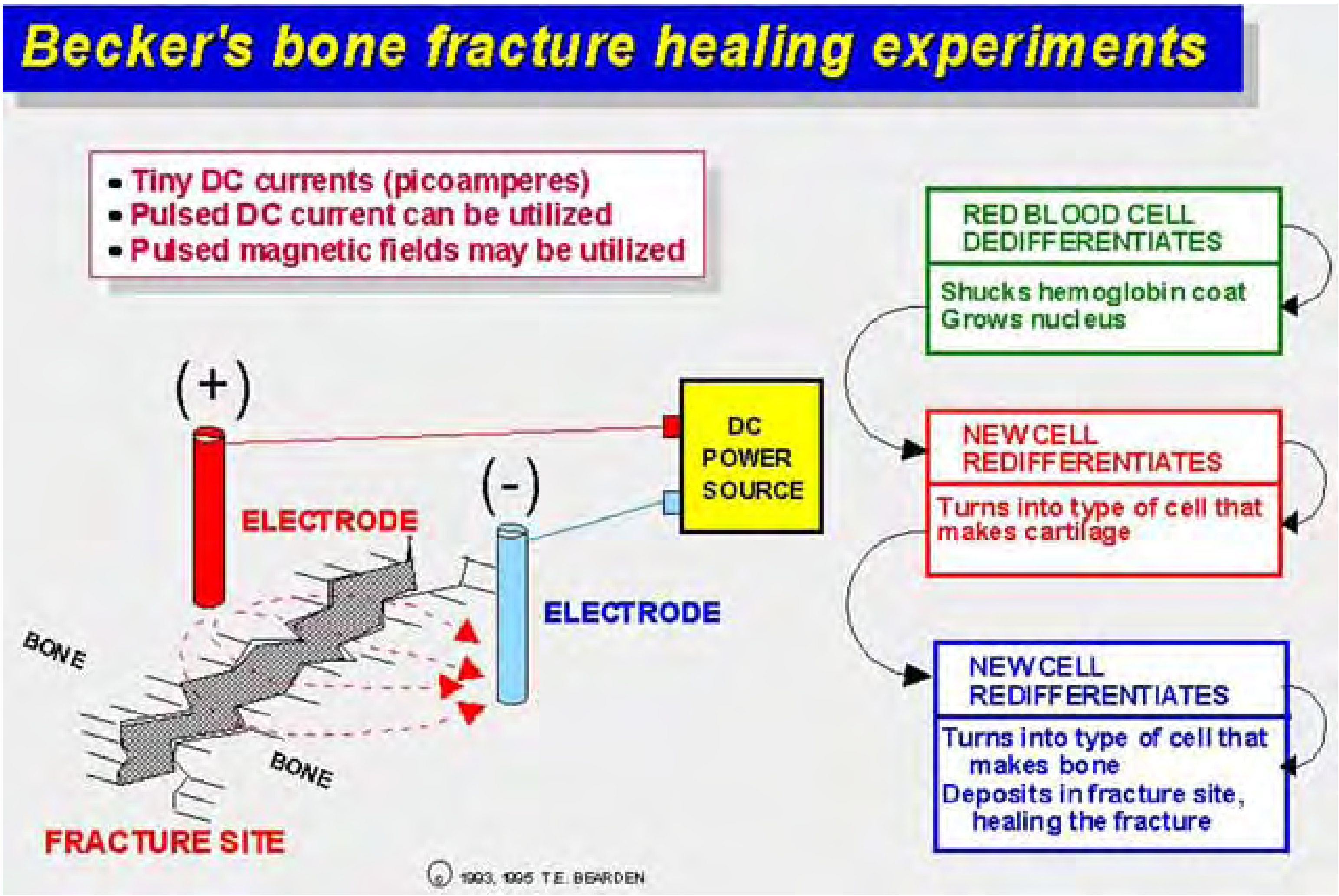

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC
POWER
SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL
DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL
REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL
REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

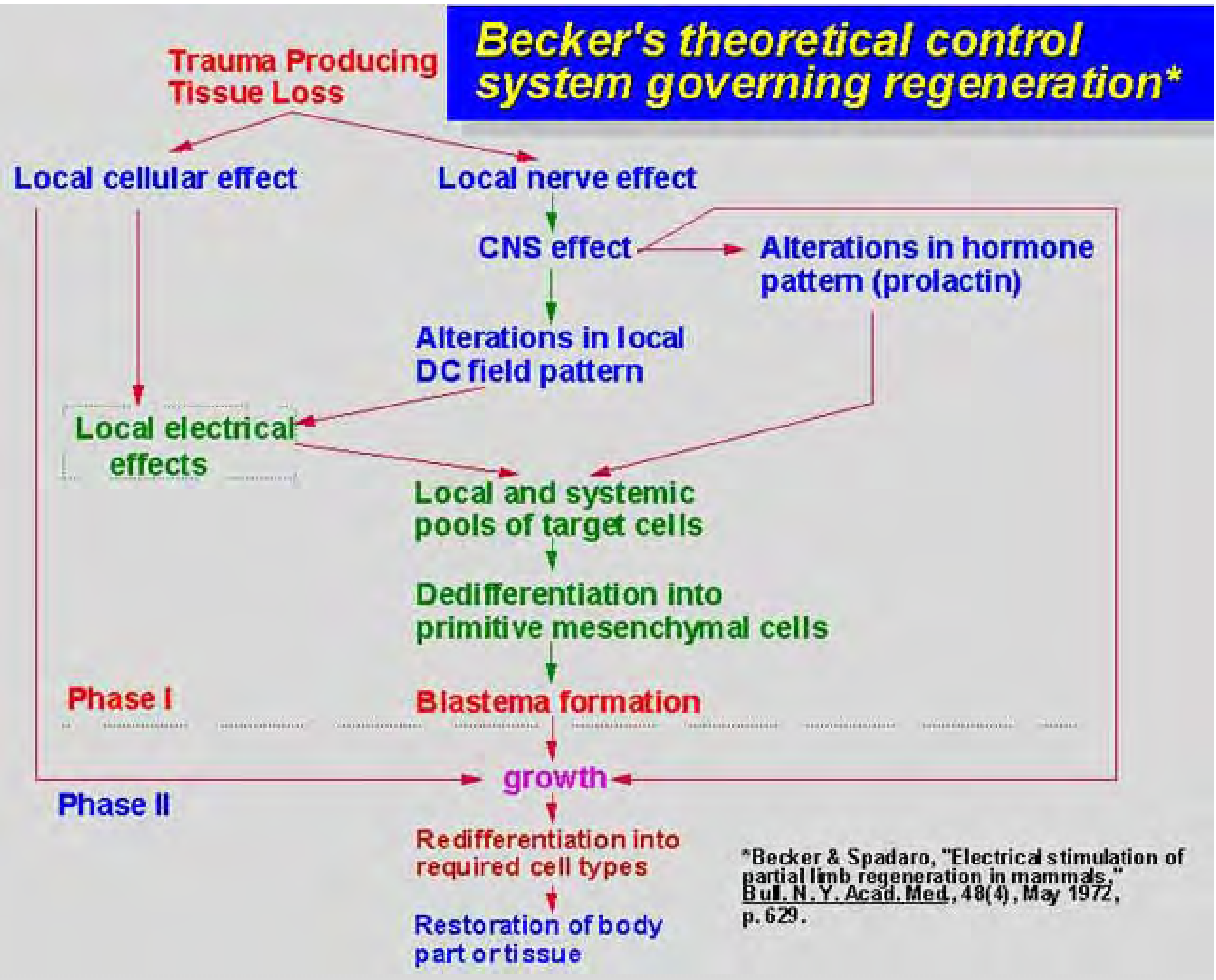

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

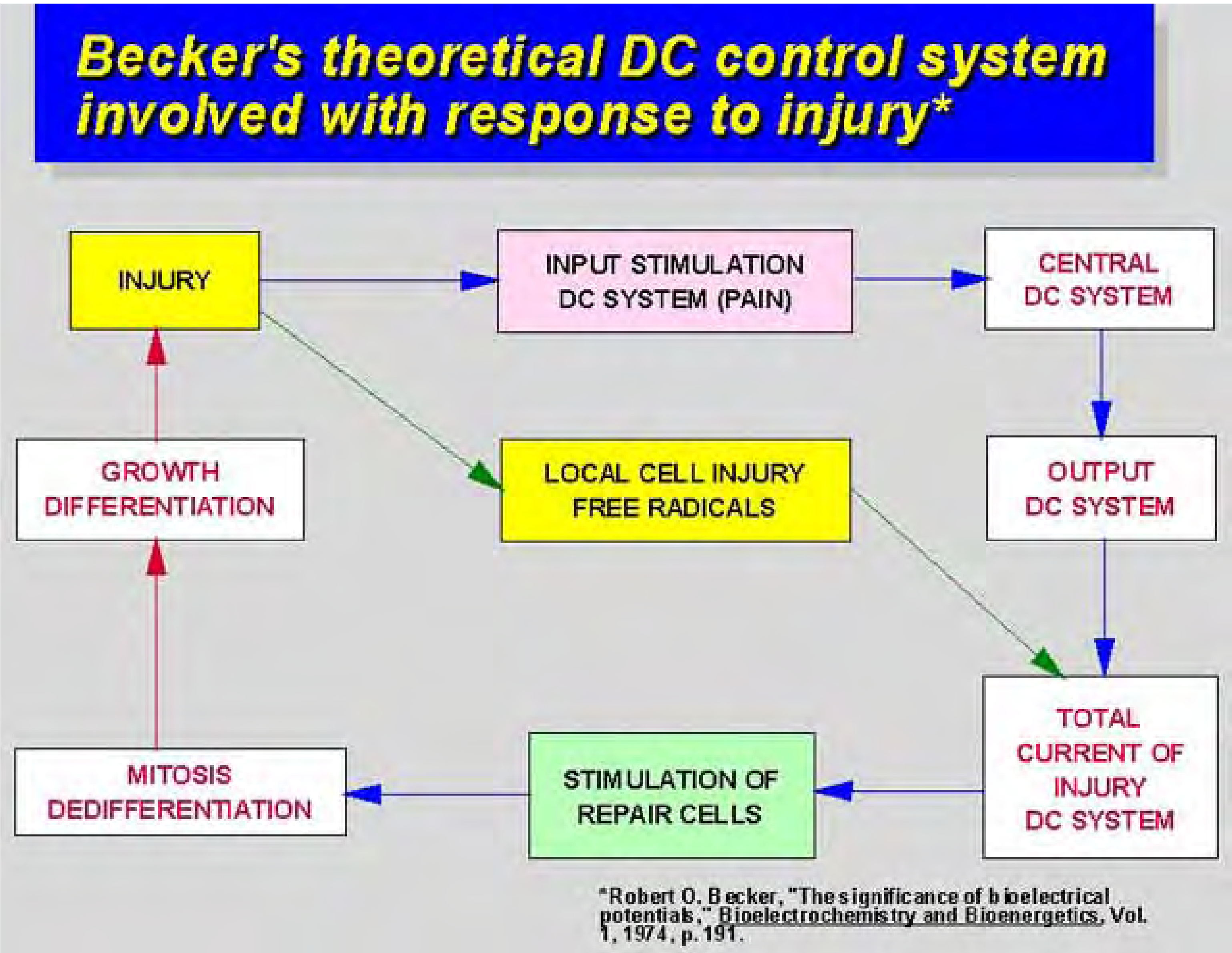

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

GROWTH
DIFFERENTIATION

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

MITOSIS
DEDIFFERENTIATION

STIMULATION OF
REPAIR CELLS

TOTAL
CURRENT OF
INJURY
DC SYSTEM

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

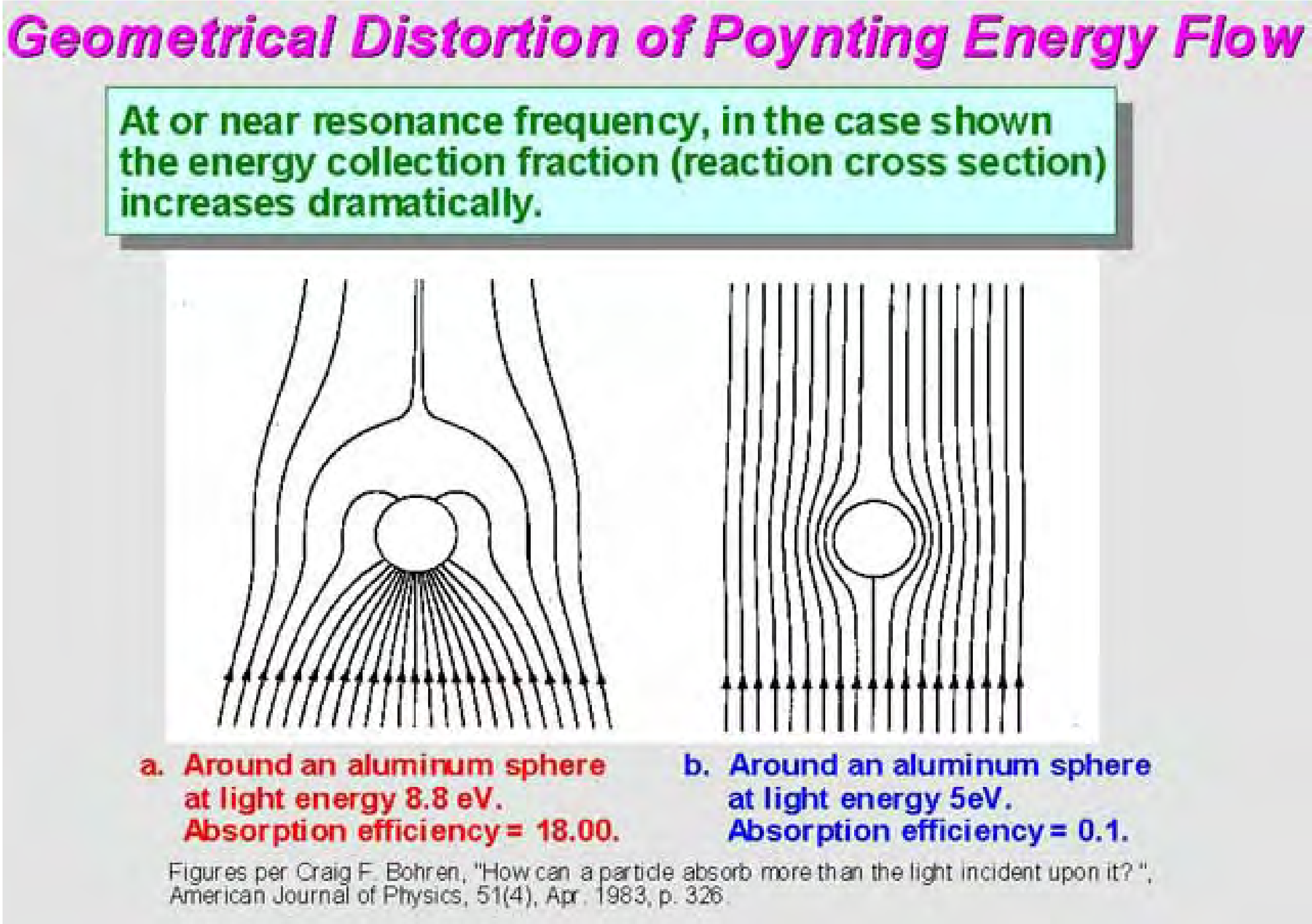

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it? ", American Journal of Physics, 51(4), Apr. 1983, p. 326.

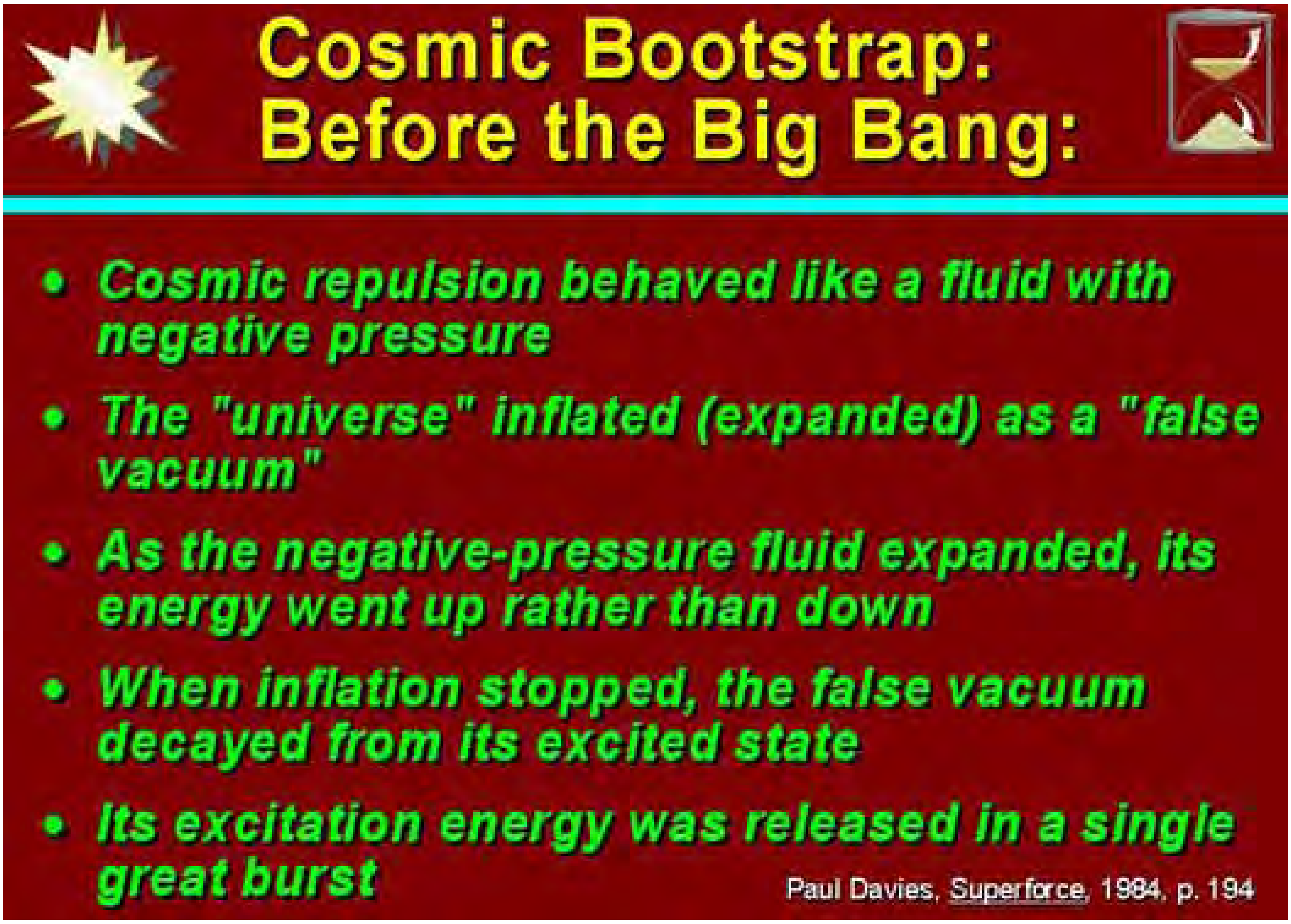

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

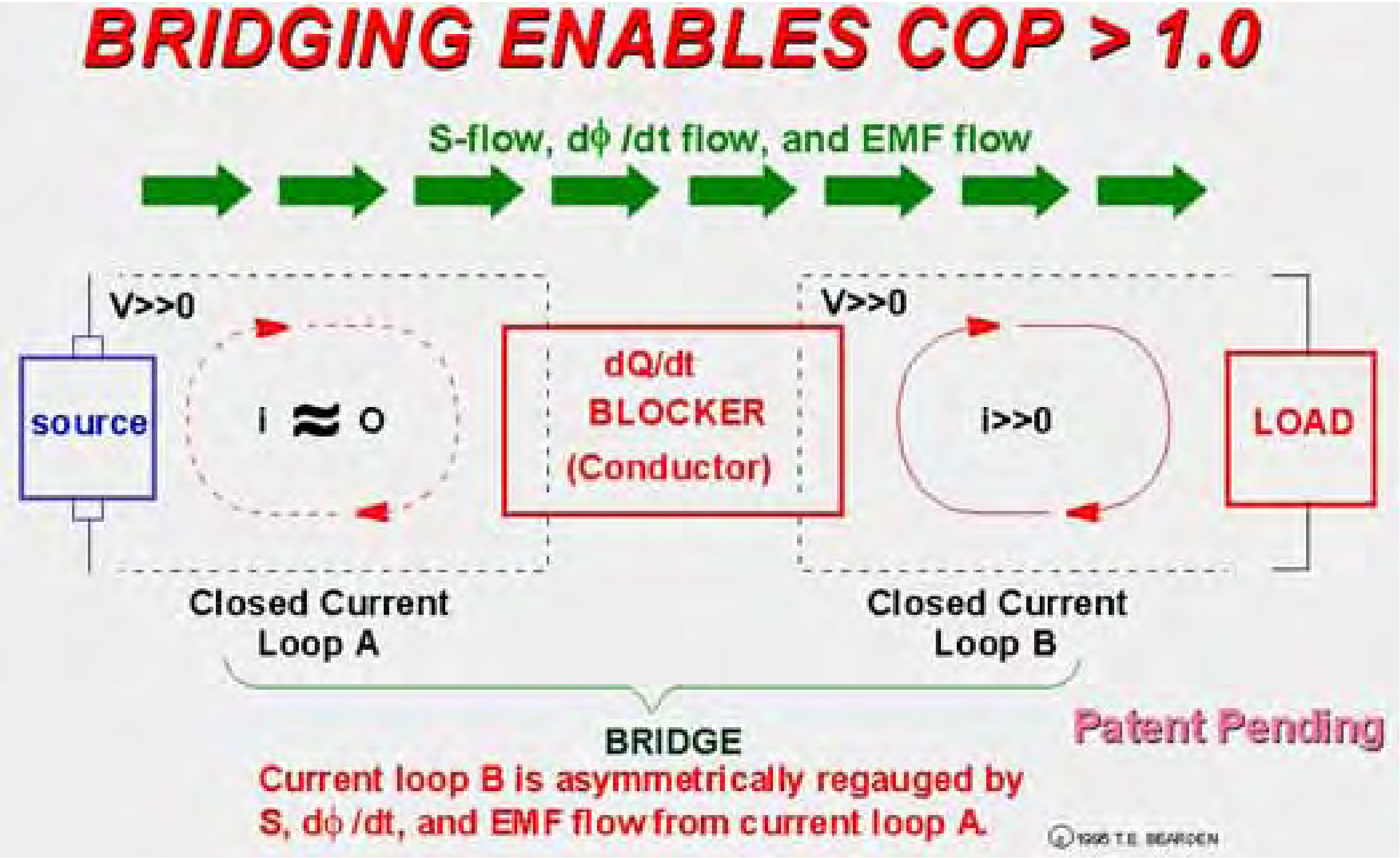

BRIDGING ENABLES COP > 1.0

S-flow, dϕ /dt flow, and EMF flow

V>>0

source

i ≈ 0

dQ/dt

BLOCKER

(Conductor)

V>>0

i>>0

LOAD

Closed Current
Loop A

Closed Current
Loop B

BRIDGE

Current loop B is asymmetrically regauged by
S, dϕ /dt, and EMF flow from current loop A.

Patent Pending

© 1998 T.E. BEARDEN

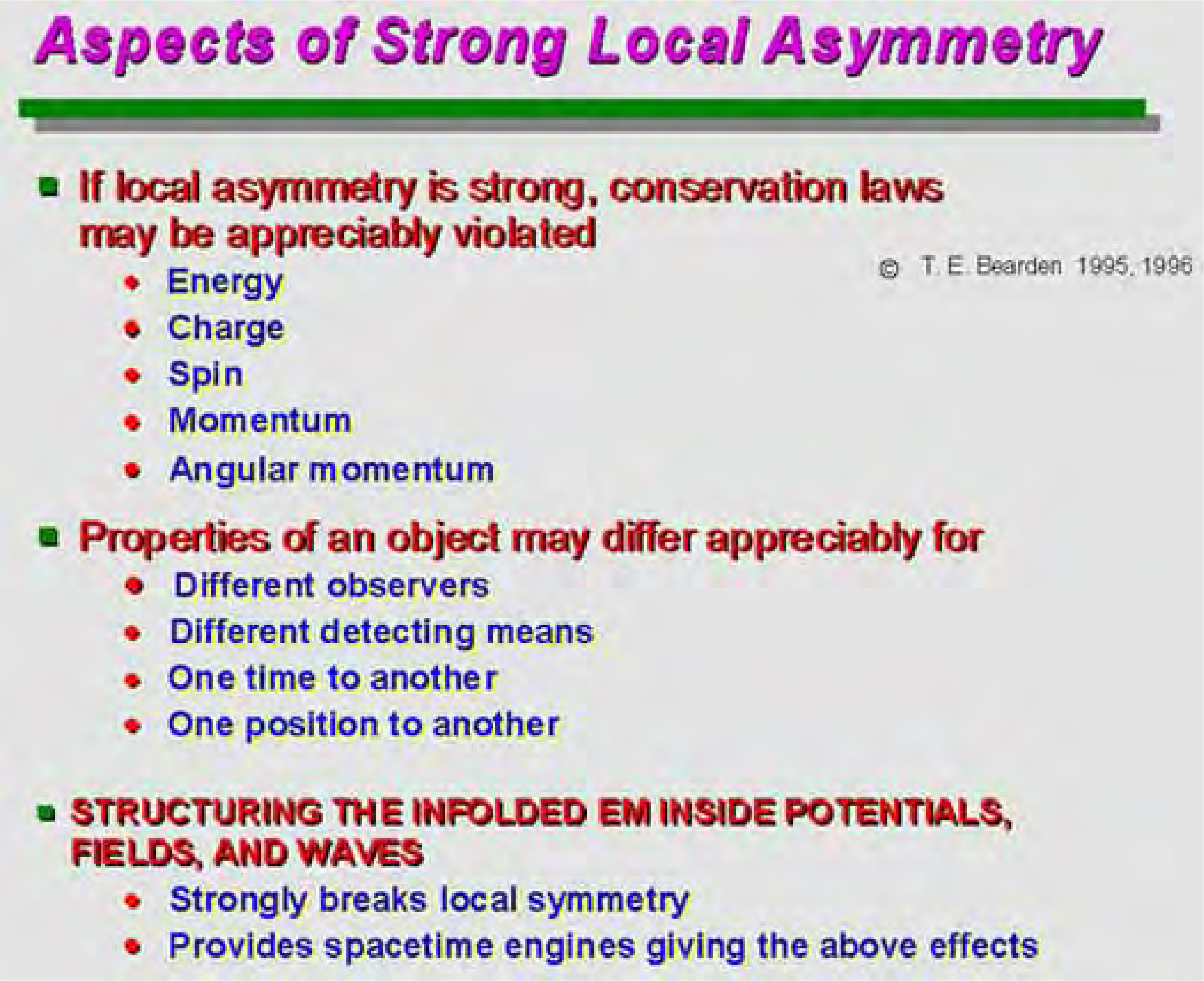

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

© T. E. Bearden 1995, 1996

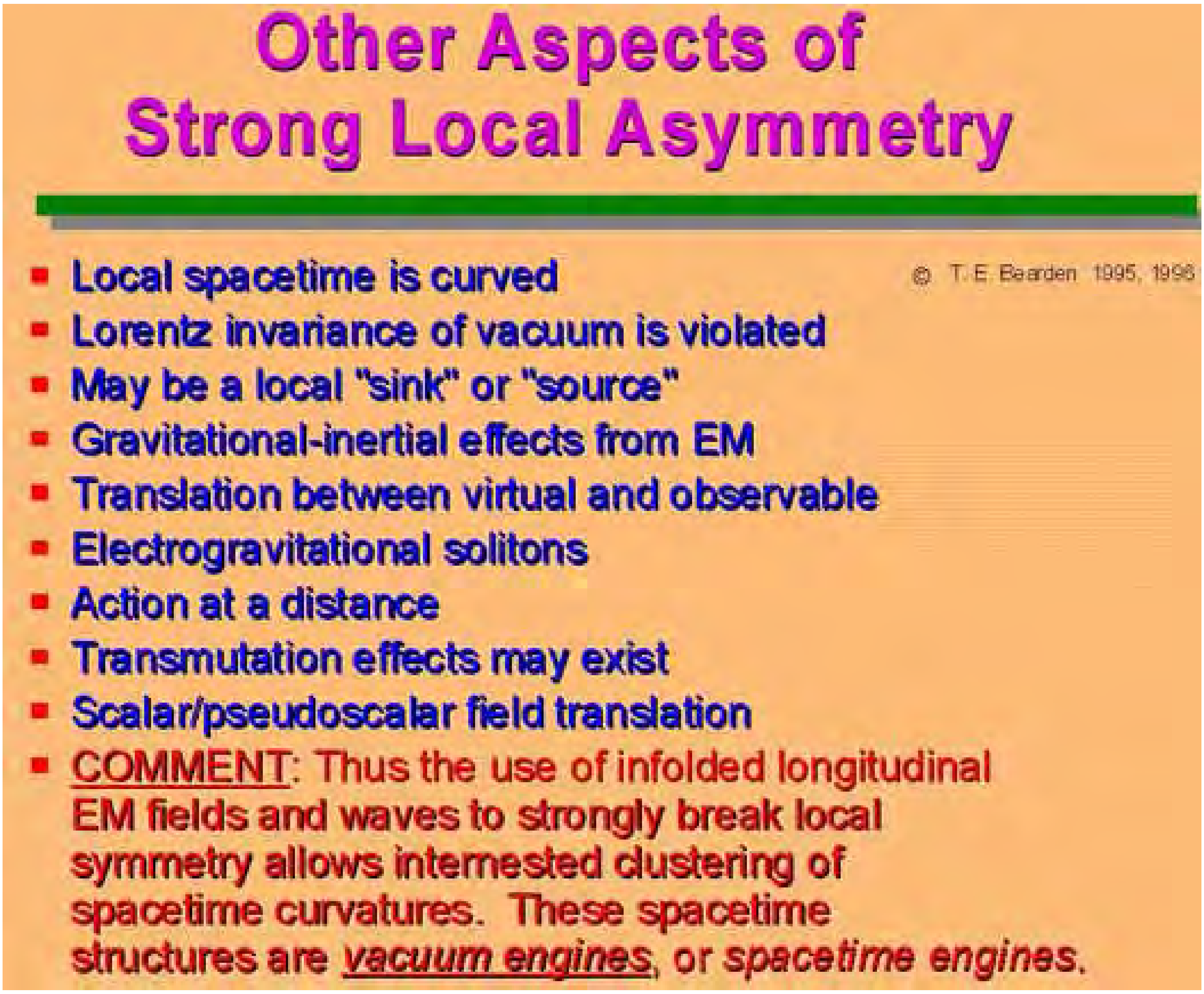

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

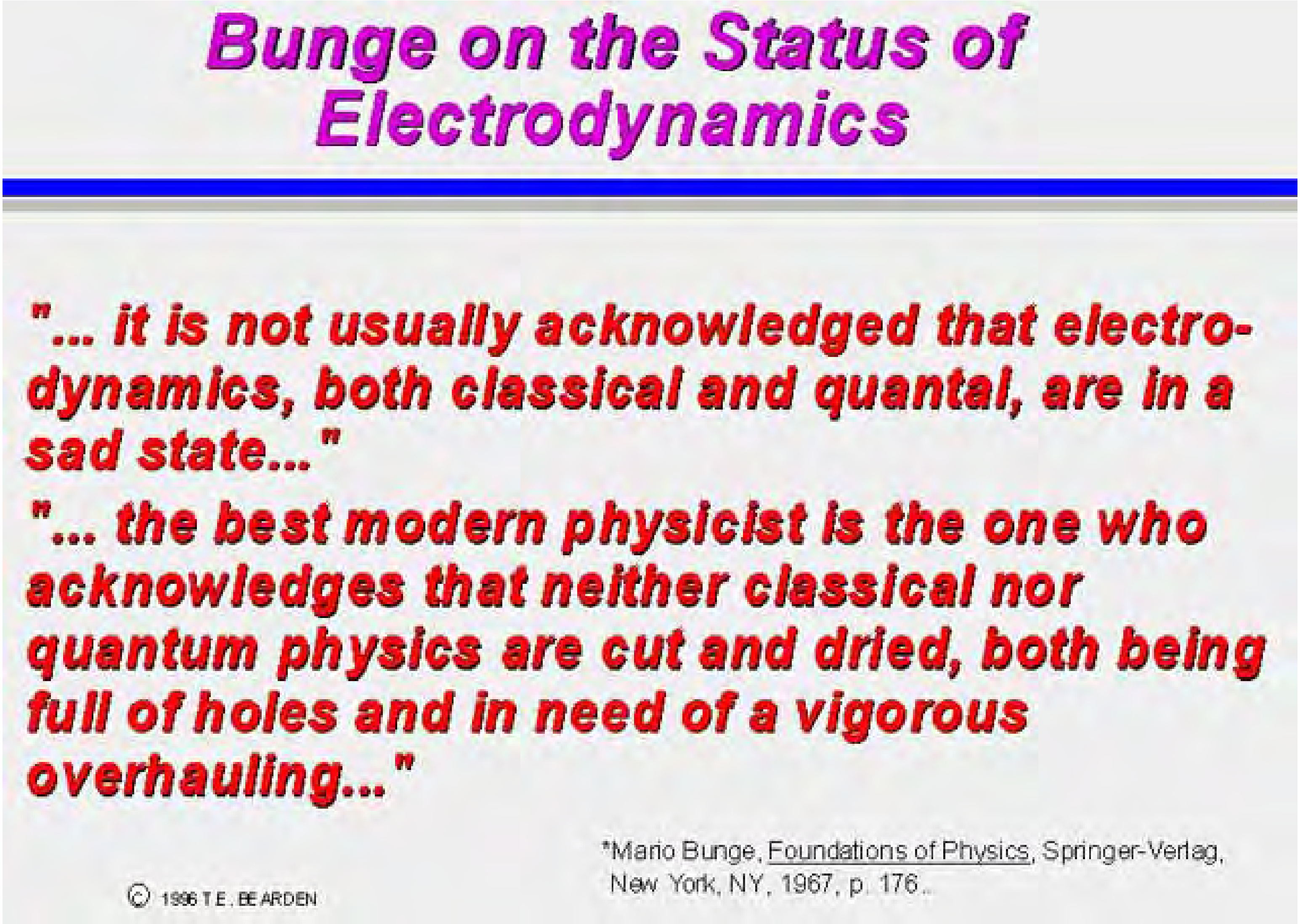

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

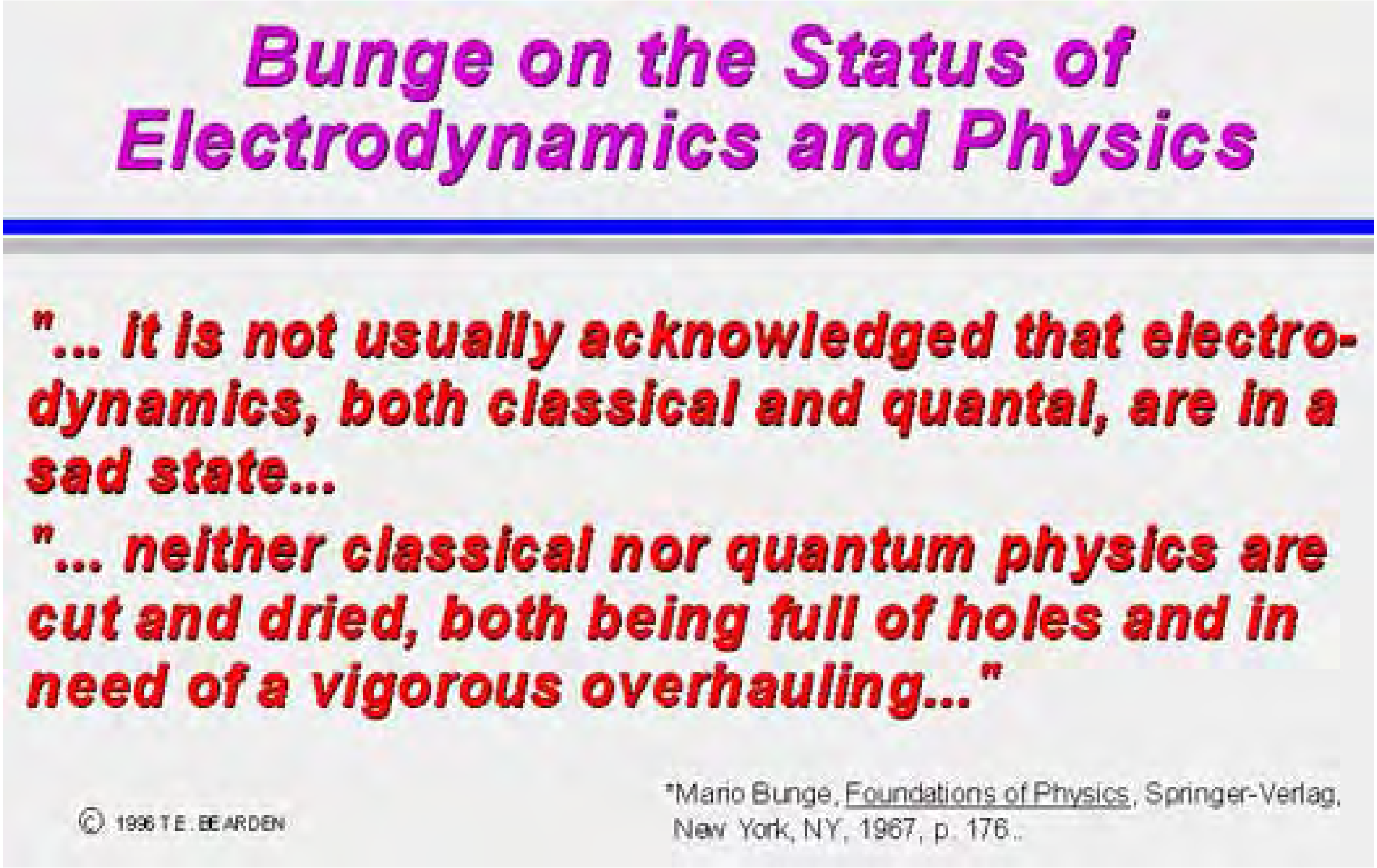

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© 1996 T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

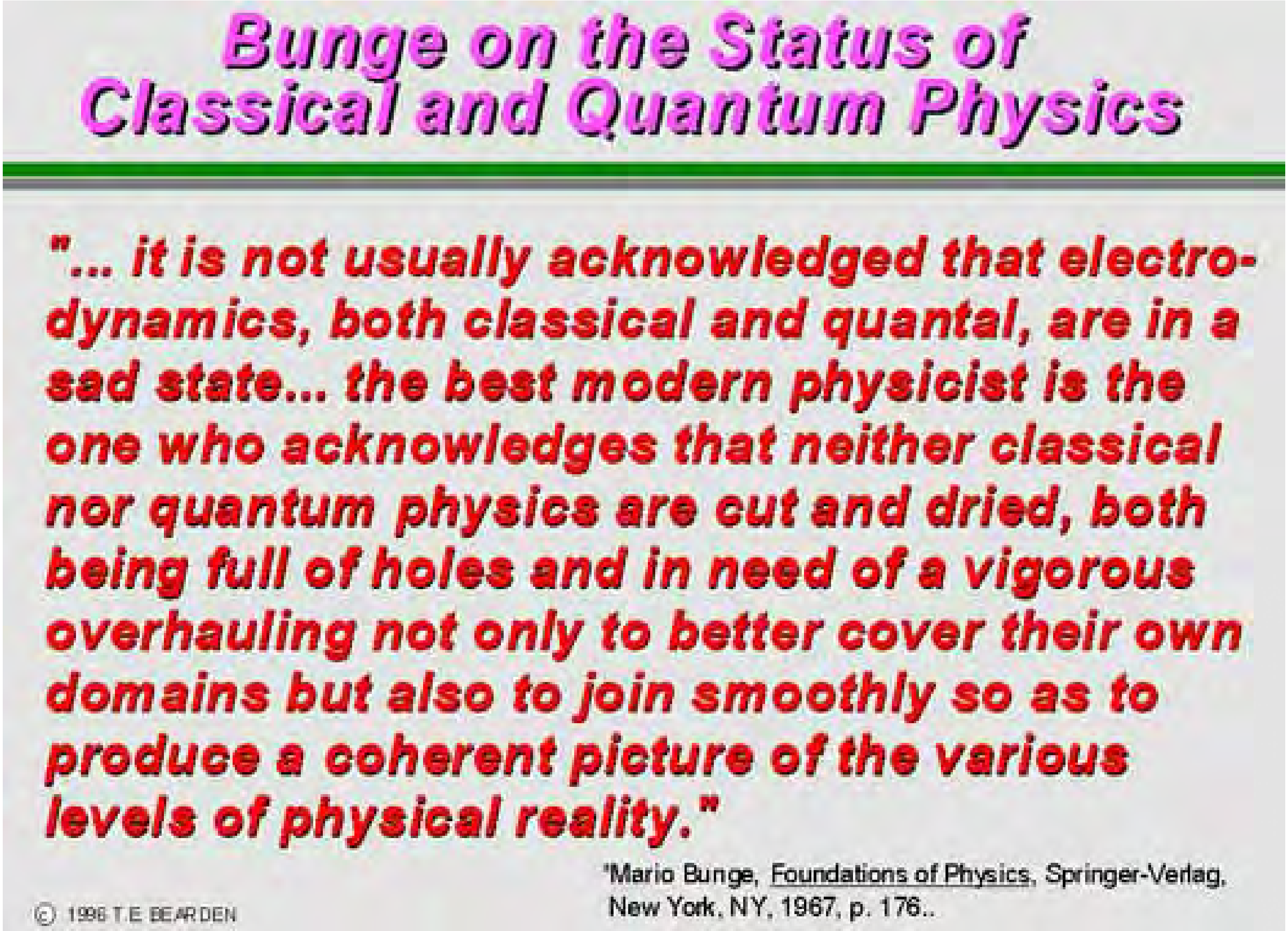

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

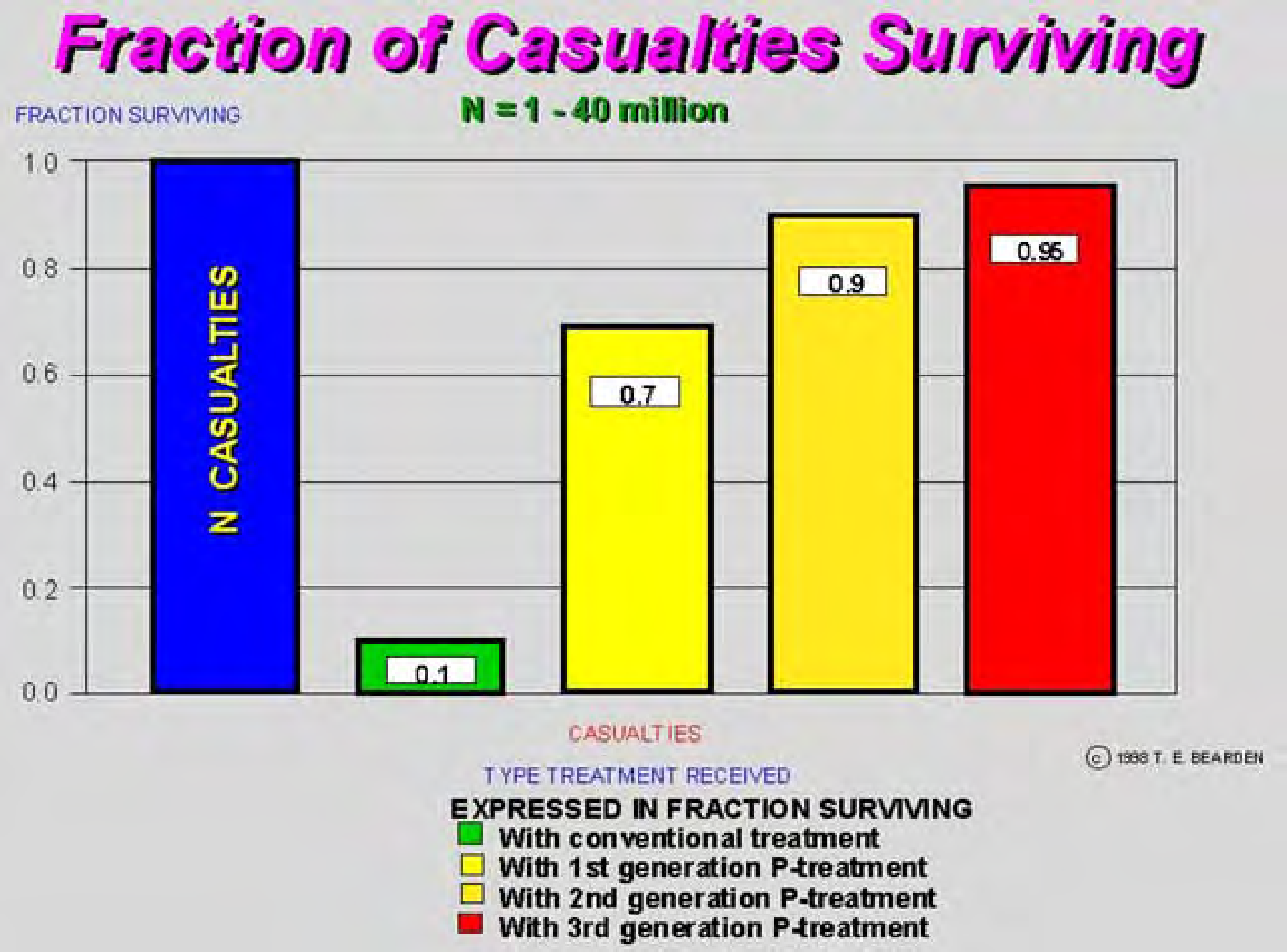
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
With conventional treatment
With 1st generation P-treatment
With 2nd generation P-treatment
With 3rd generation P-treatment

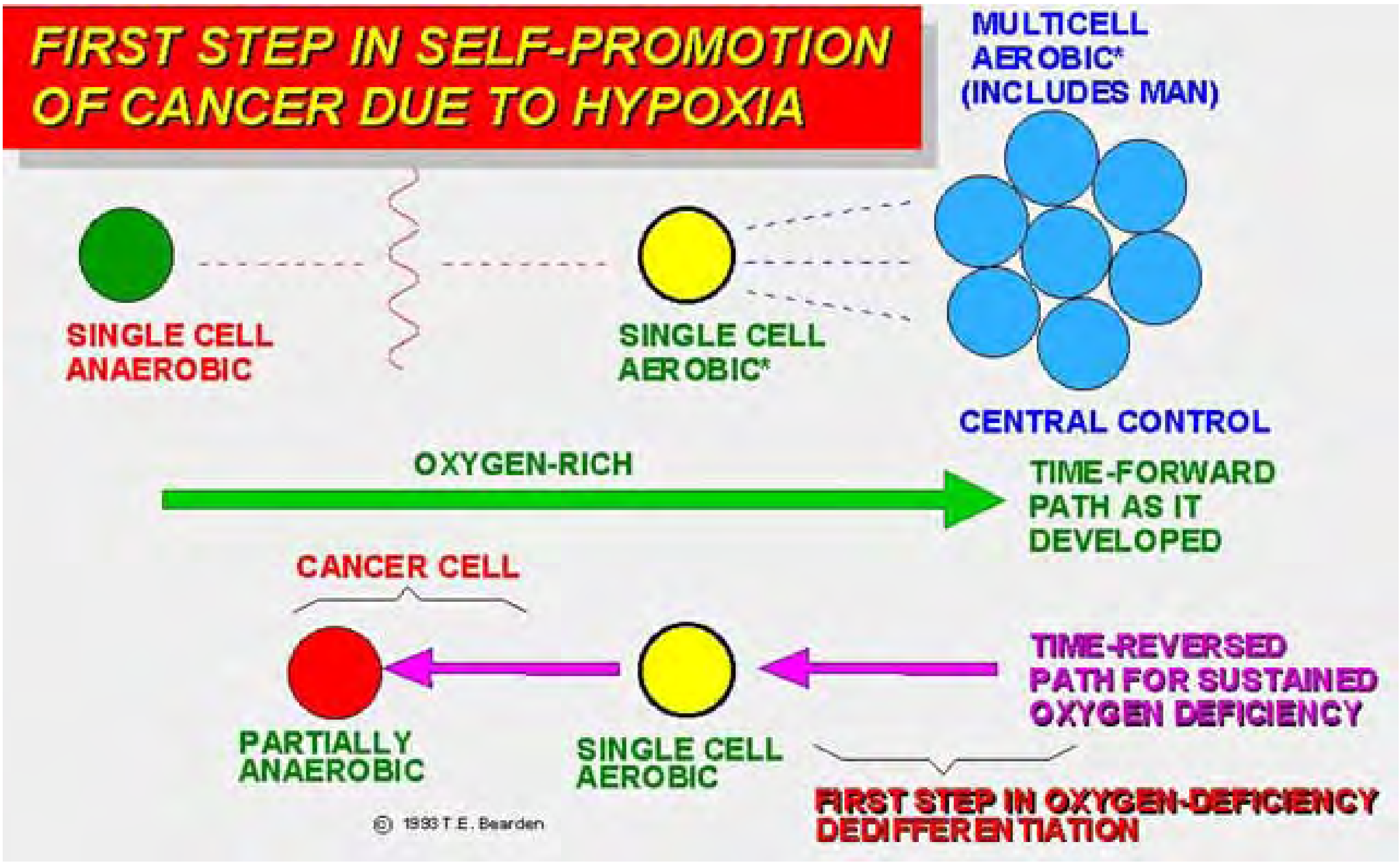

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

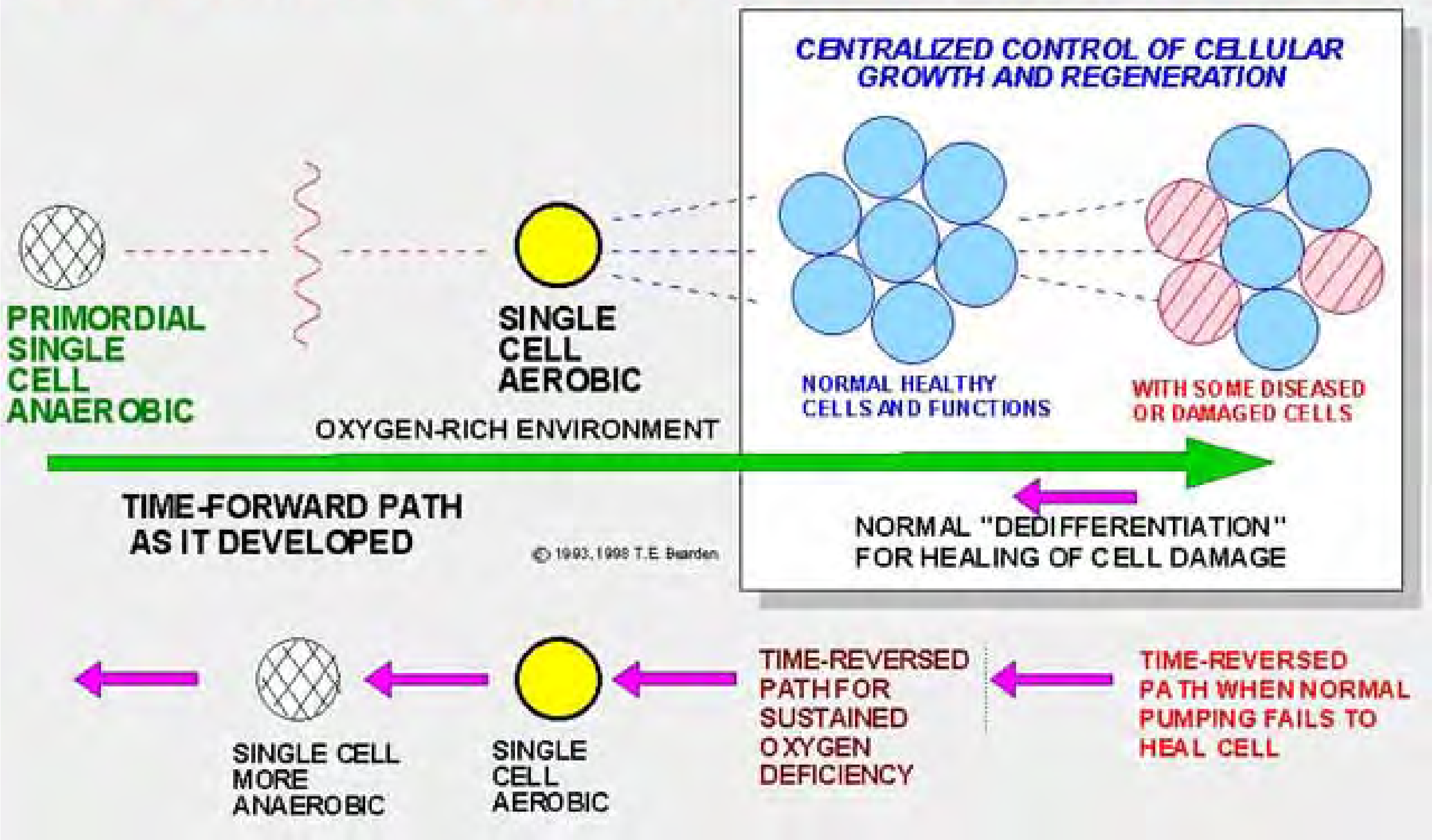

CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION

PRIMORDIAL SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC

OXYGEN-RICH ENVIRONMENT

NORMAL HEALTHY CELLS AND FUNCTIONS

WITH SOME DISEASED OR DAMAGED CELLS

TIME-FORWARD PATH AS IT DEVELOPED

© 1993, 1998 T.E. Bearden

NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE

SINGLE CELL MORE ANAEROBIC

SINGLE CELL AEROBIC

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

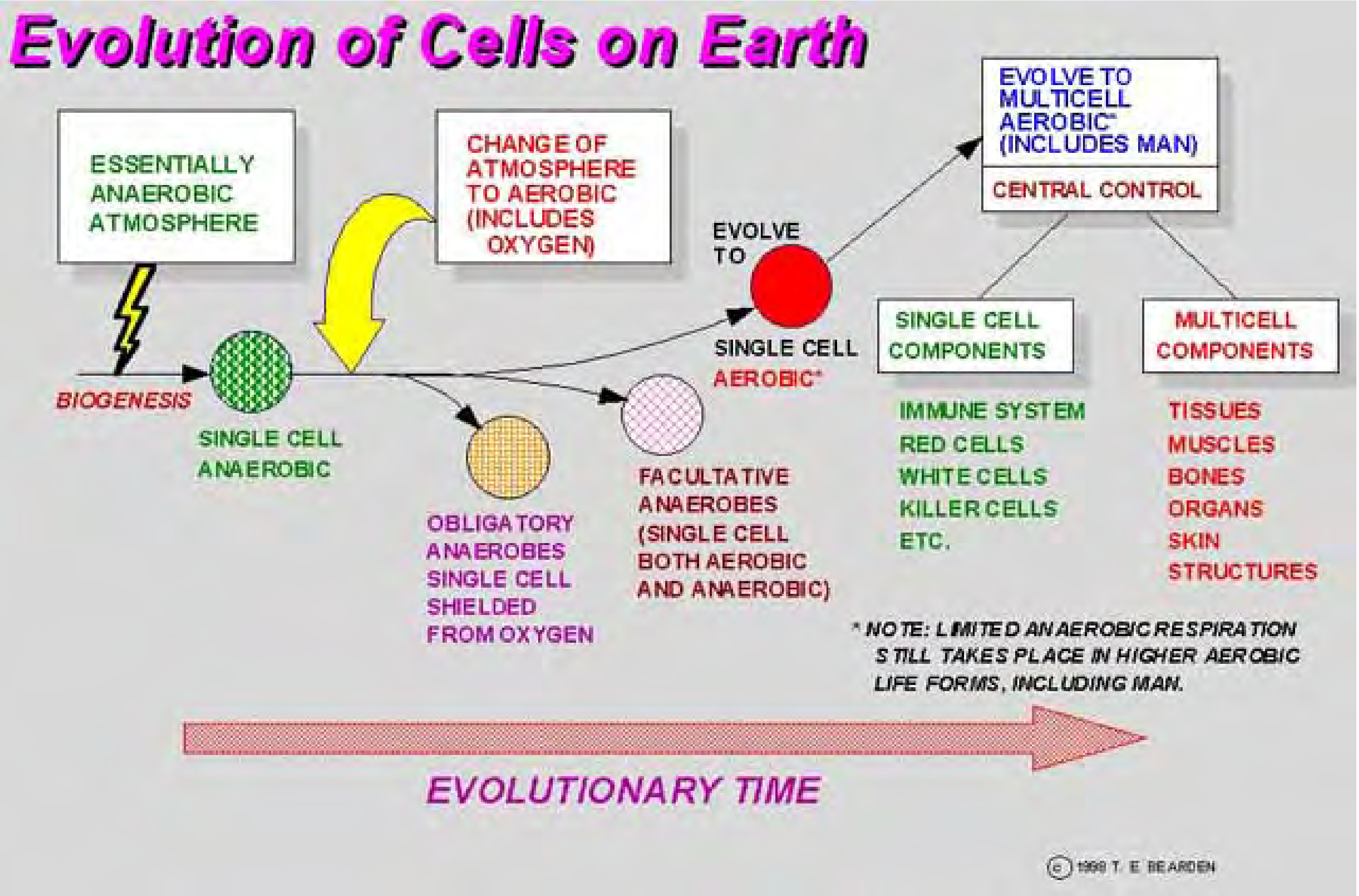

Evolution of Cells on Earth

ESSENTIALLY
ANAEROBIC
ATMOSPHERE

CHANGE OF
ATMOSPHERE
TO AEROBIC
(INCLUDES
OXYGEN)

BIOGENESIS

SINGLE CELL
ANAEROBIC

OBLIGATORY
ANAEROBES
SINGLE CELL
SHIELDED
FROM OXYGEN

FACULTATIVE
ANAEROBES
(SINGLE CELL
BOTH AEROBIC
AND ANAEROBIC)

EVOLVE
TO

SINGLE CELL
AEROBIC*

EVOLVE TO
MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL
COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL
COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

© 1999 T. E. BEARDEN

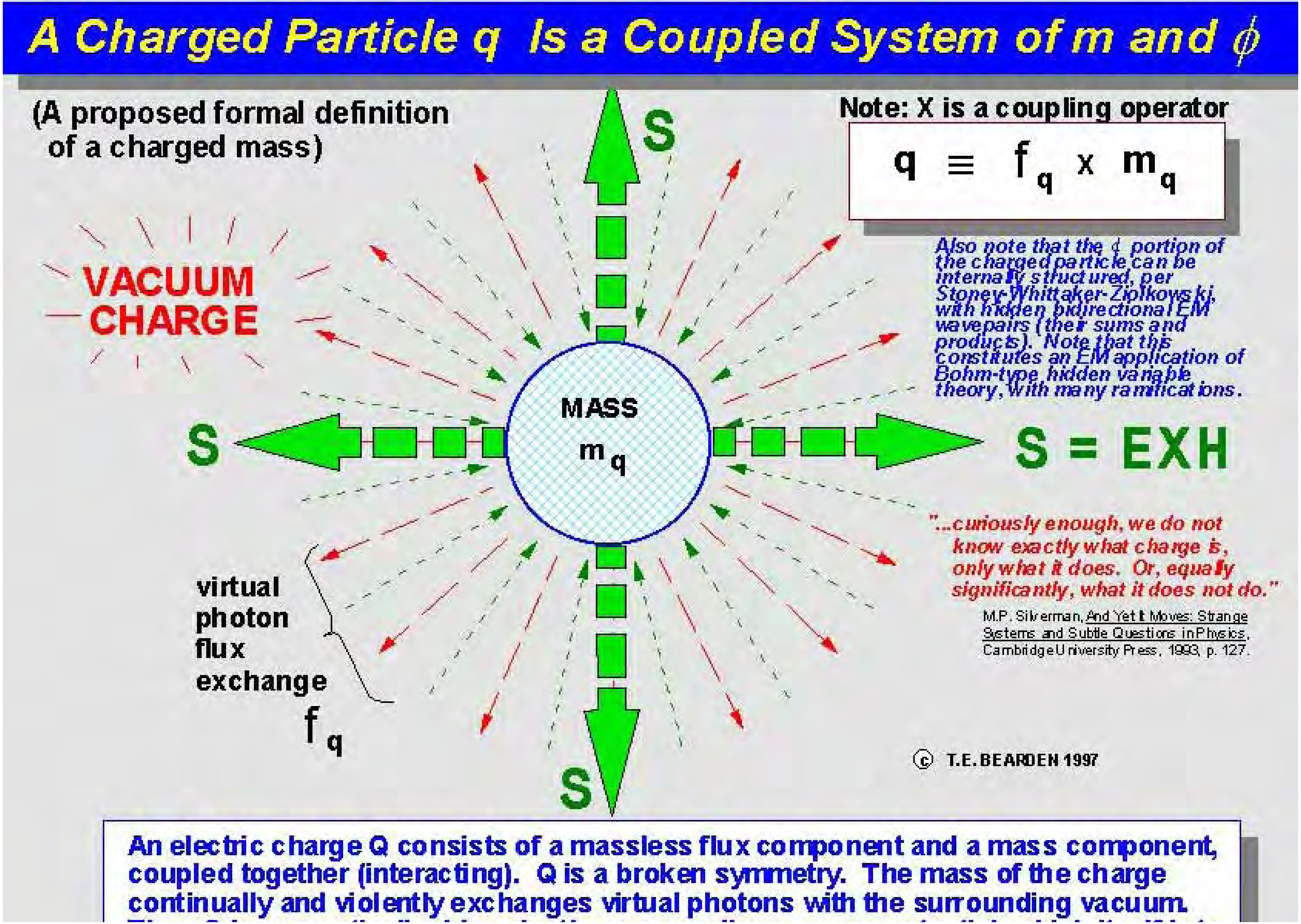

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition
of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

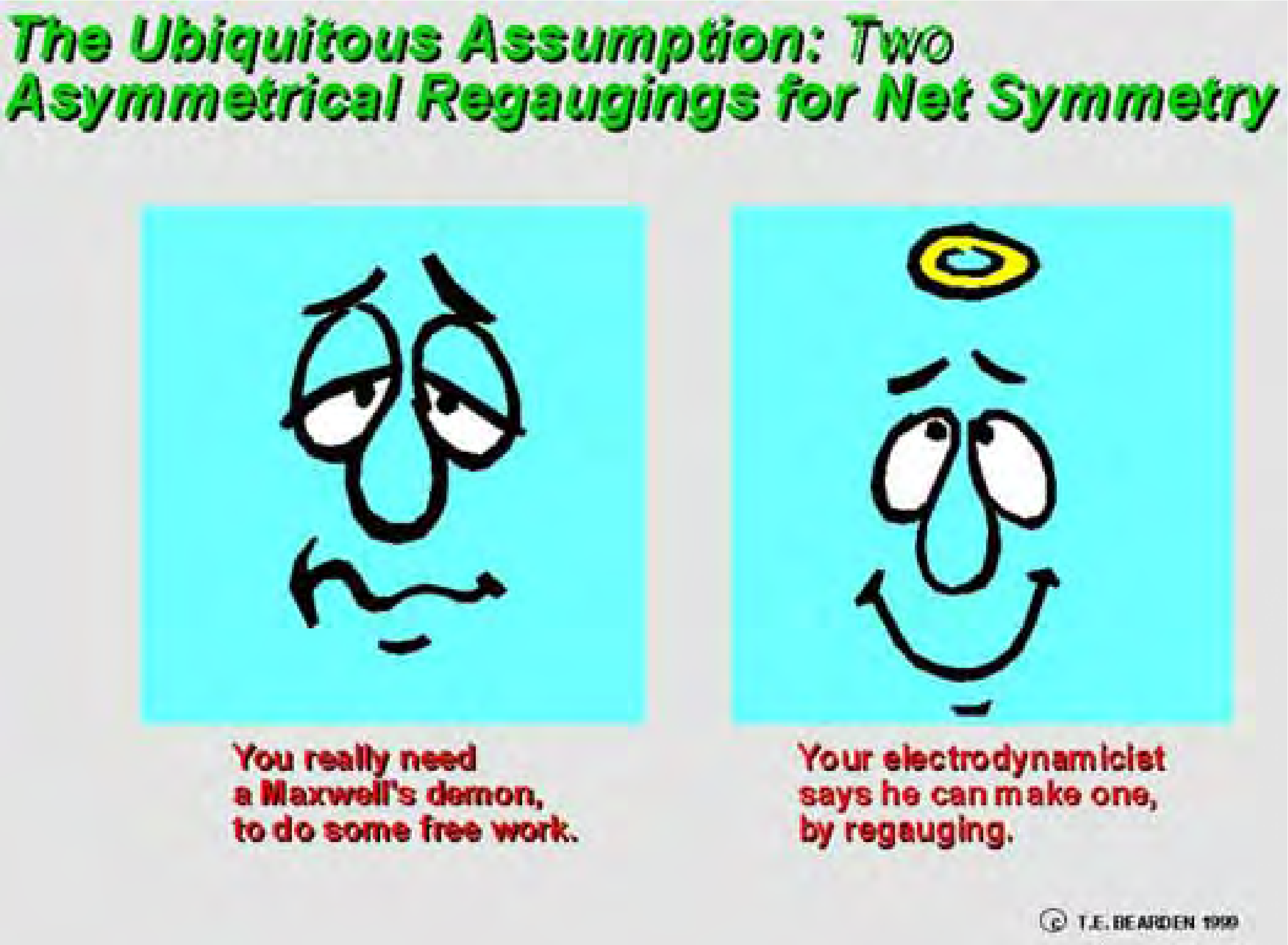

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

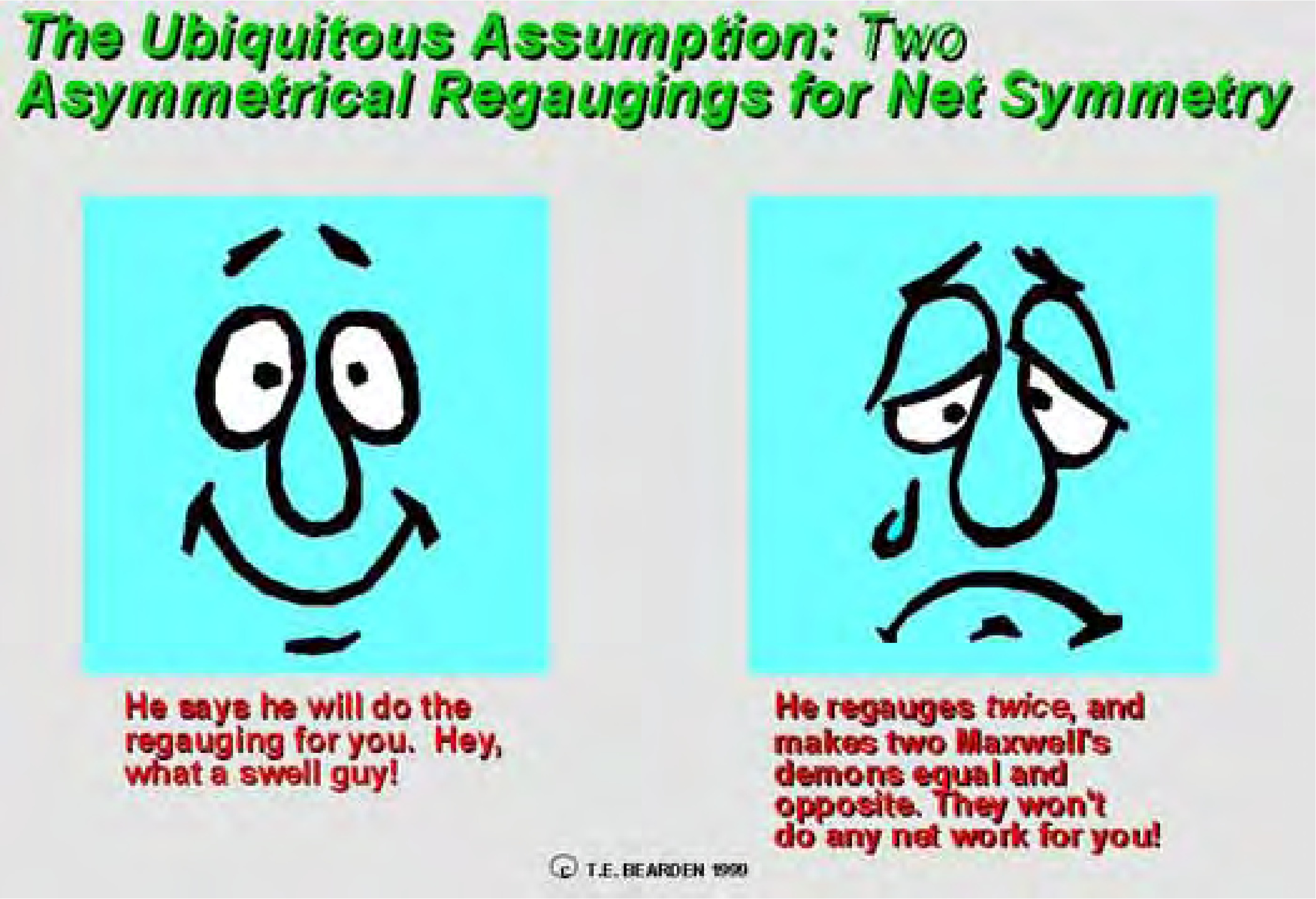

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

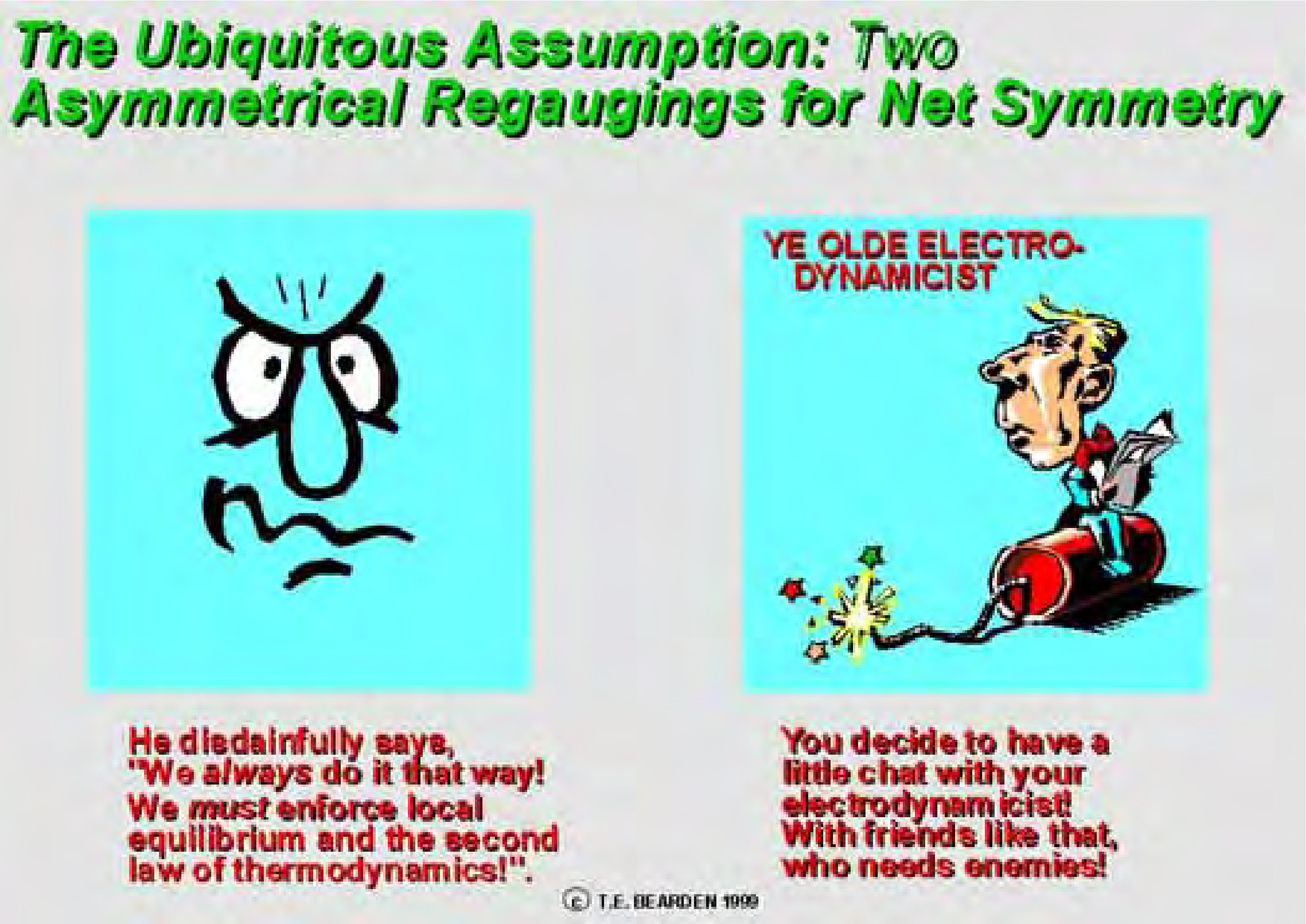

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

YE OLDE ELECTRO-
DYNAMICIST

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

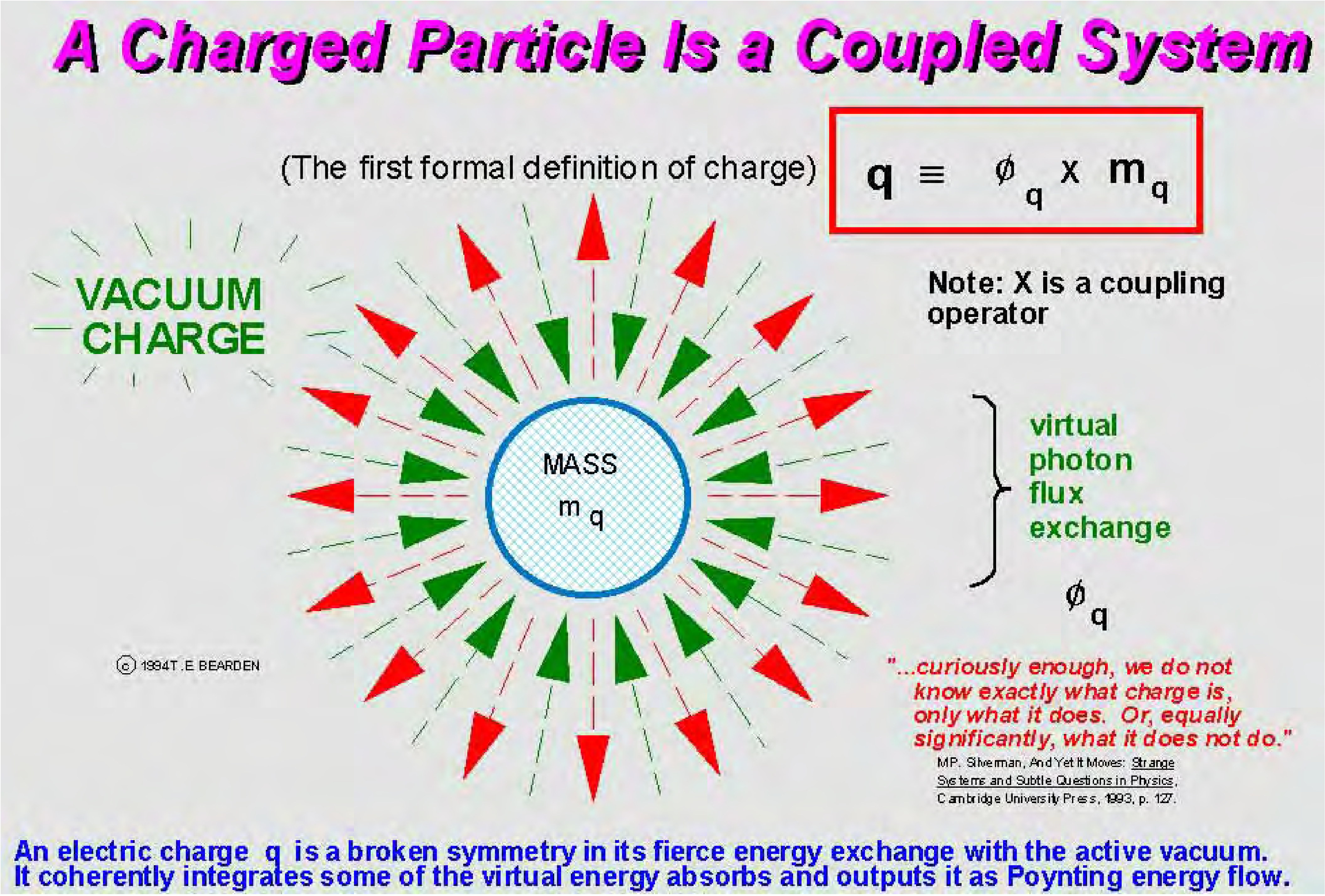

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \ \mathrm{x} \ m_q}$$

Note: X is a coupling operator

ⓒ 1994 T .E BEARDEN

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

Chung's Carbon Filament Negative Resistor

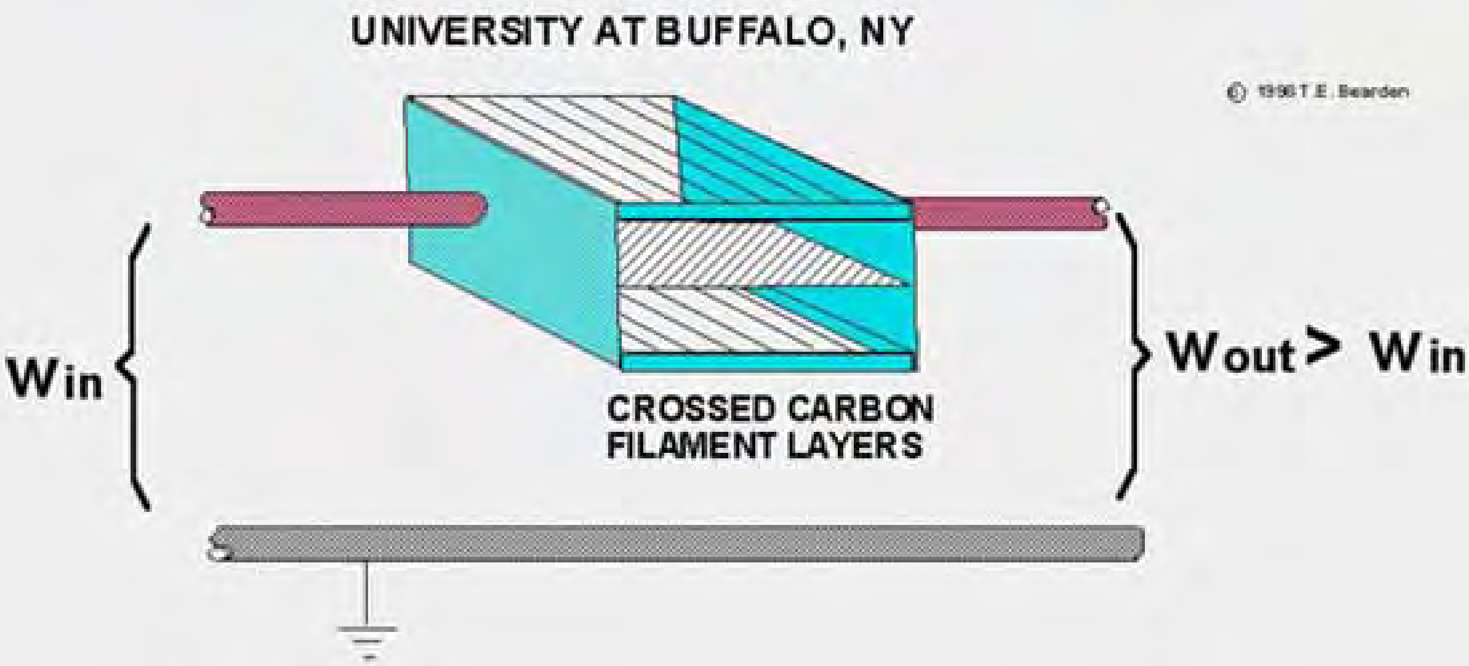
UNIVERSITY AT BUFFALO, NY
© 1996 T.E. Bearden
Win
Wout > Win
CROSSED CARBON
FILAMENT LAYERS

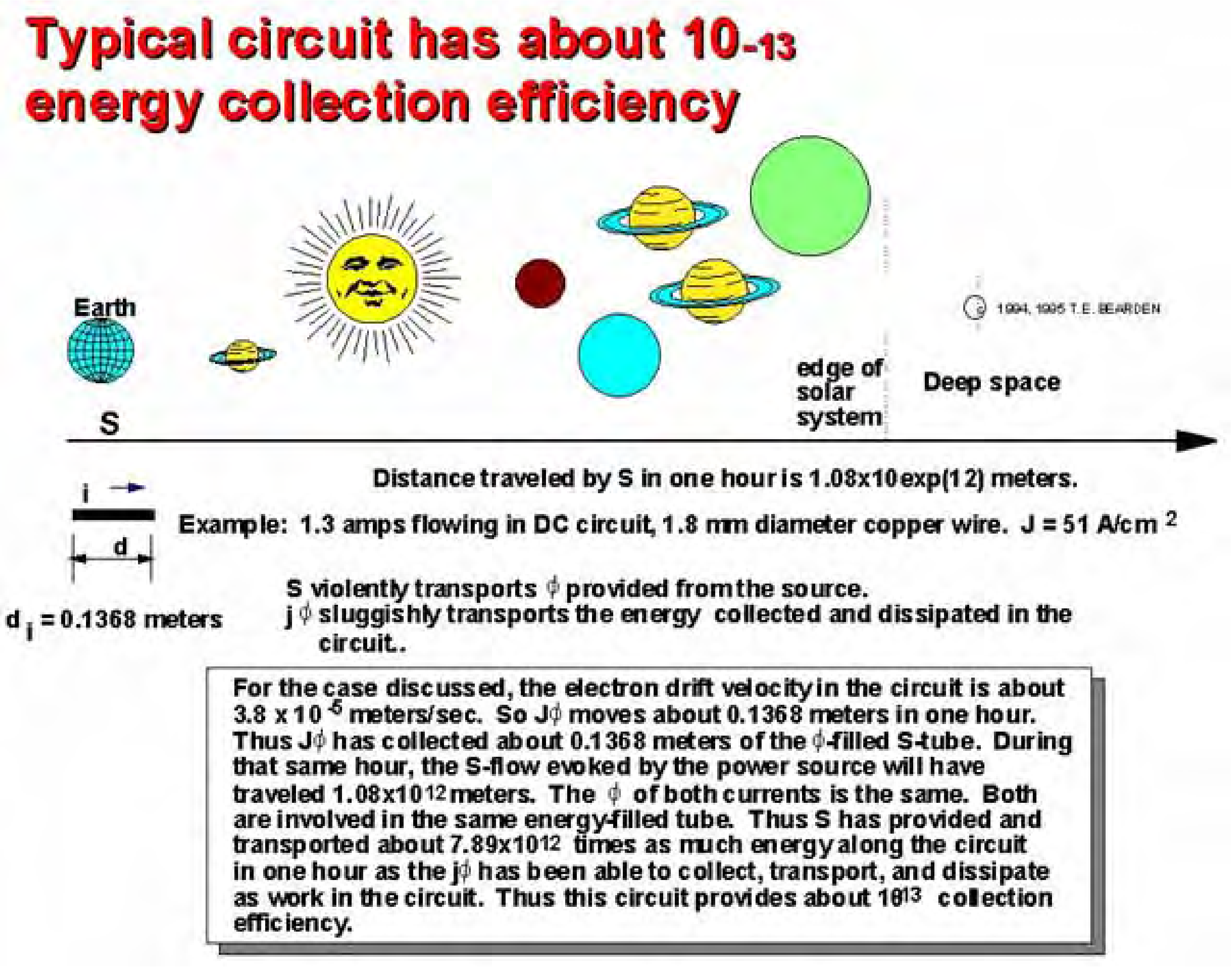

Typical circuit has about 10_{-13} energy collection efficiency

Earth
© 1994, 1995 T.E. BEARDEN
edge of solar system
Deep space
S
i
d

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. $J = 51\ A/cm^2$

$d_i = 0.1368$ meters

S violently transports ϕ provided from the source.
$j\phi$ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So $J\phi$ moves about 0.1368 meters in one hour. Thus $J\phi$ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled 1.08×10^{12} meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about 7.89×10^{12} times as much energy along the circuit in one hour as the $j\phi$ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

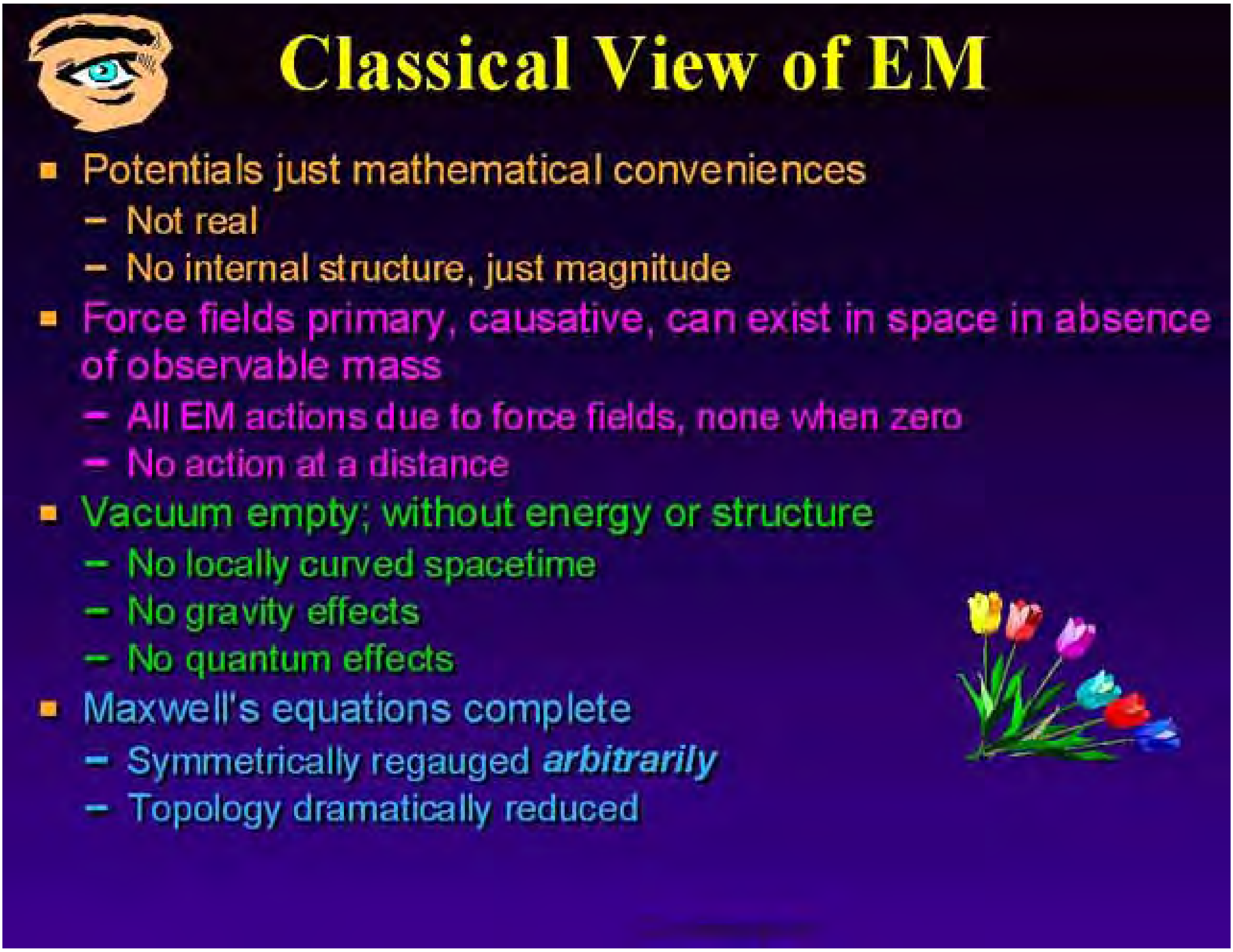

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

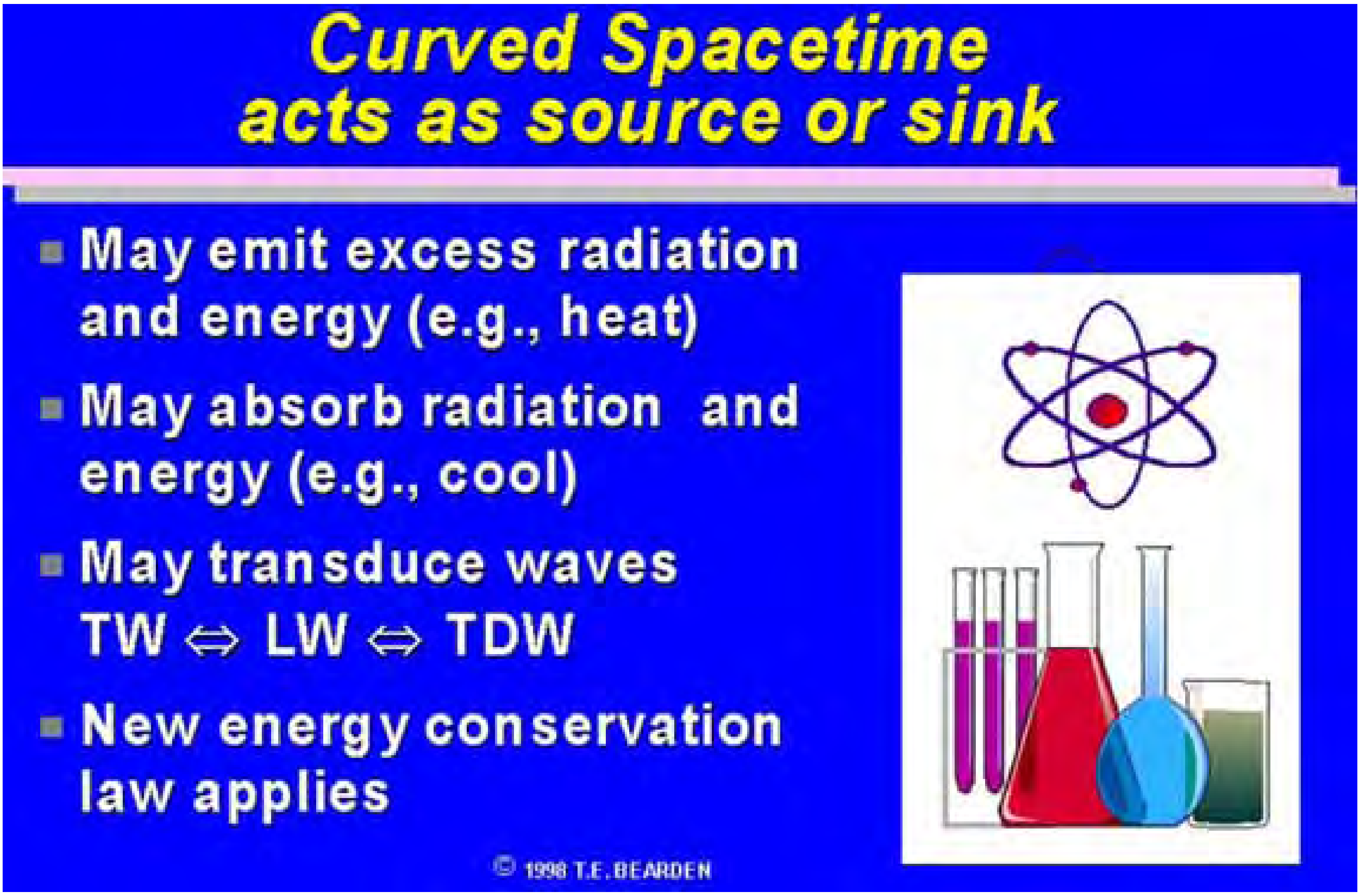

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E.BEARDEN

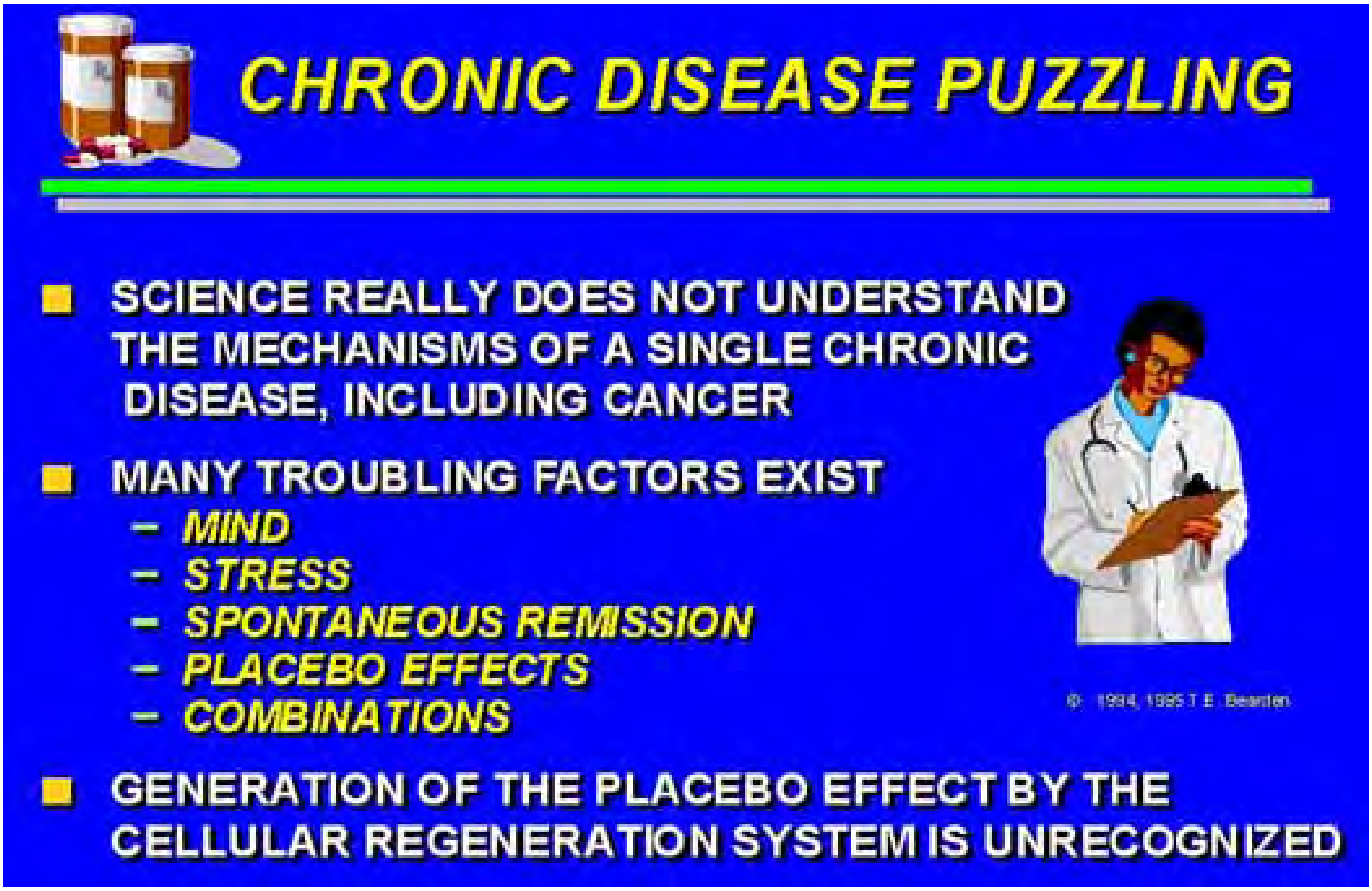

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

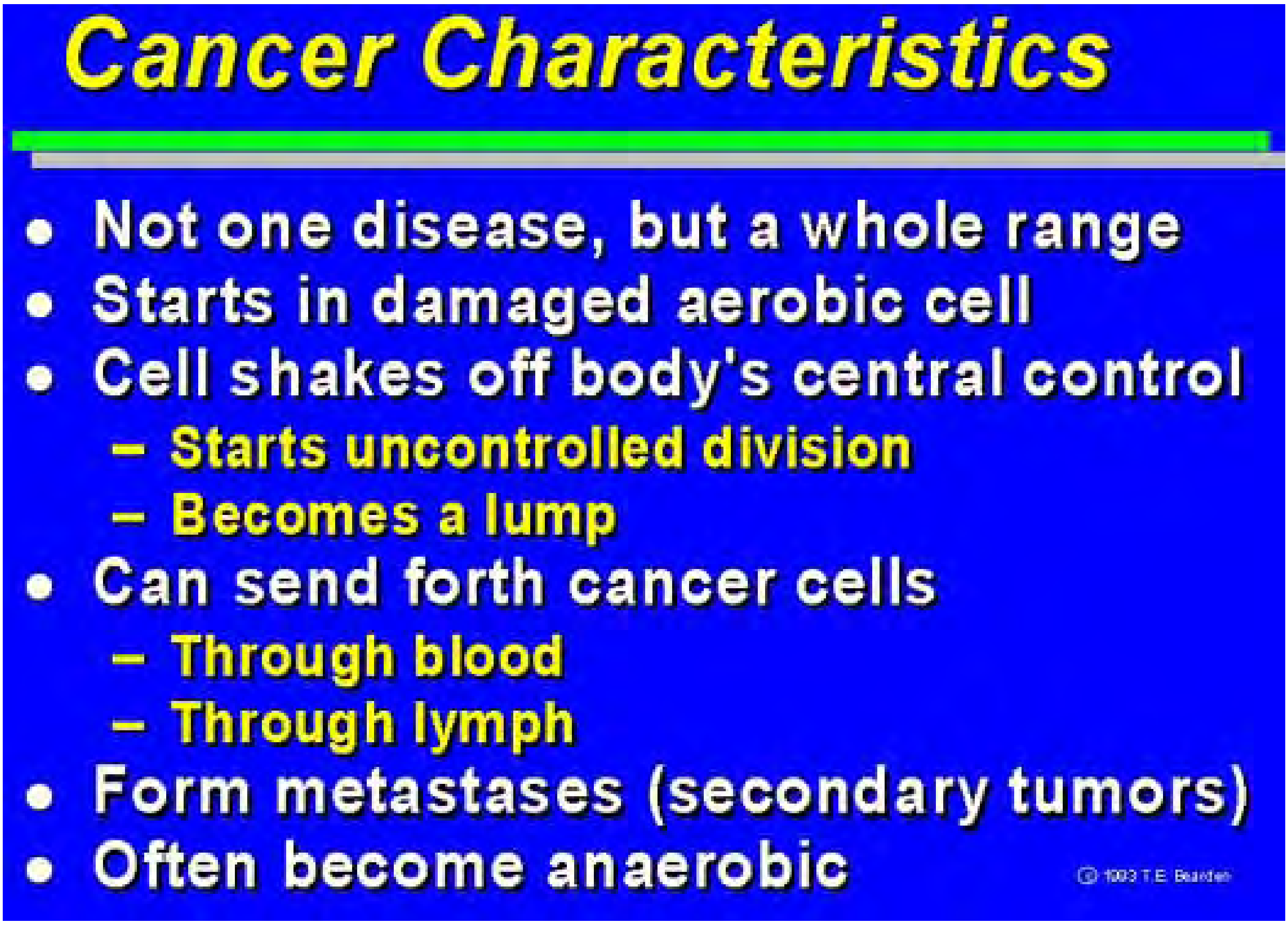

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

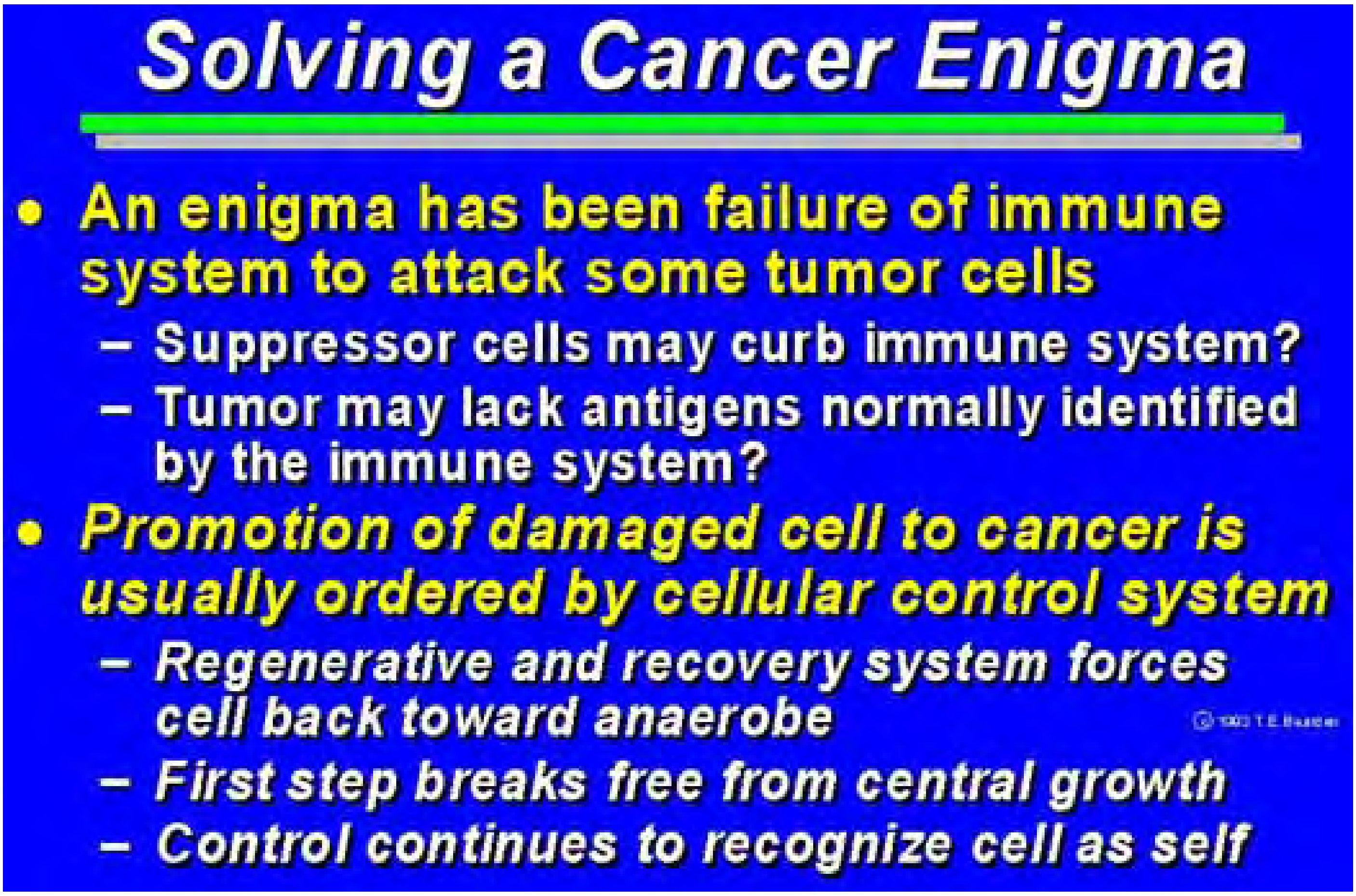

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

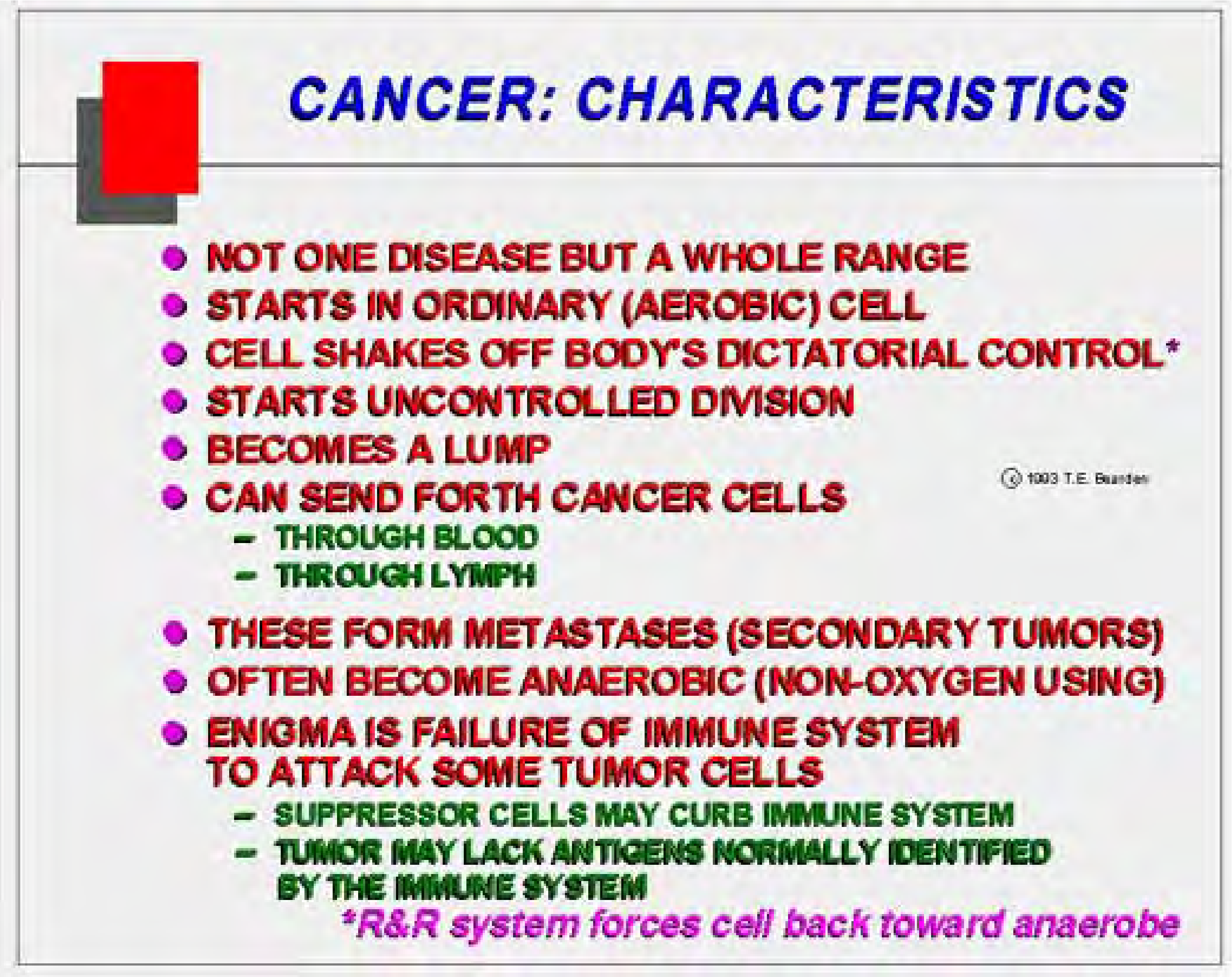

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

© 1993 T.E. Bearden

*R&R system forces cell back toward anaerobe

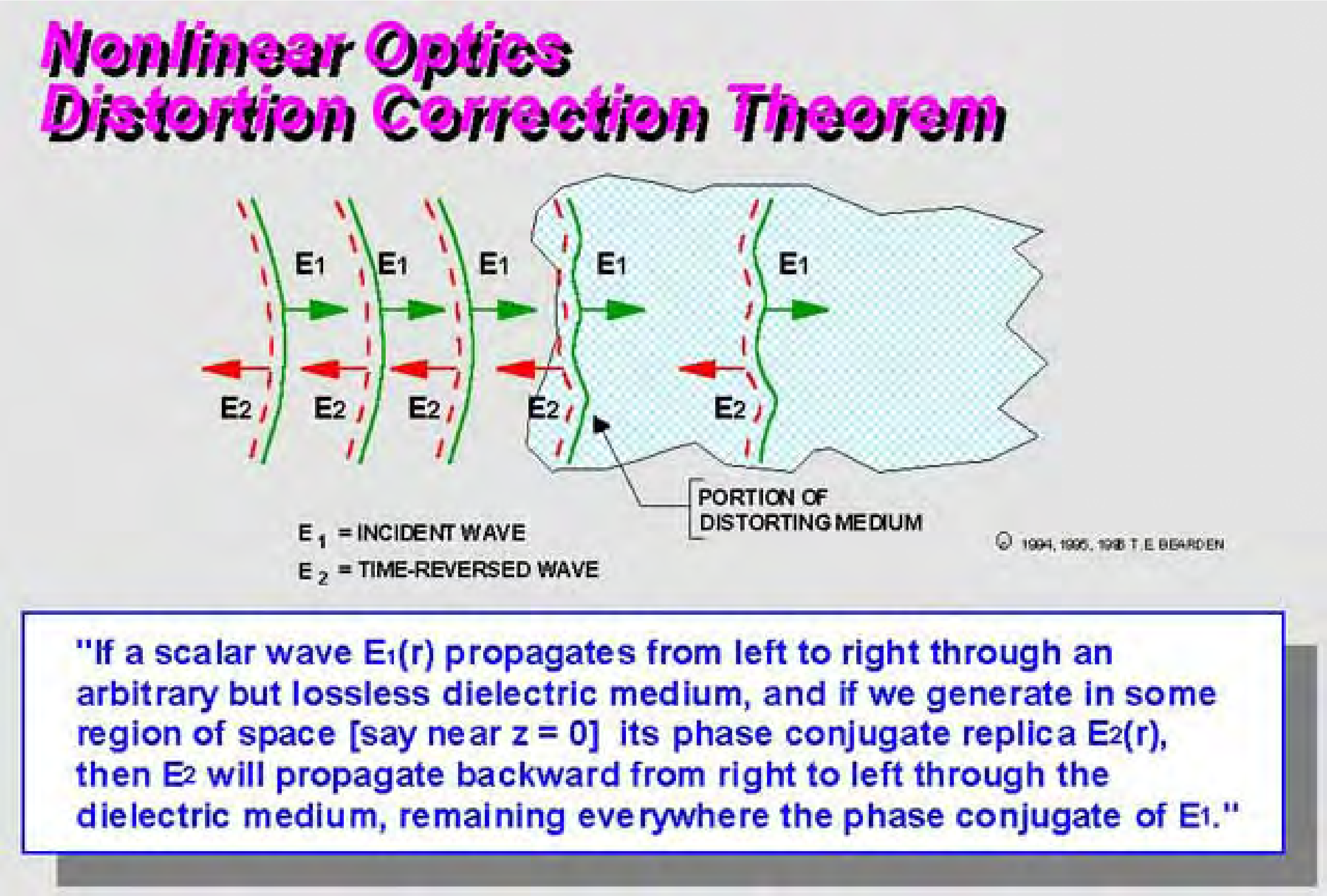

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T.E BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

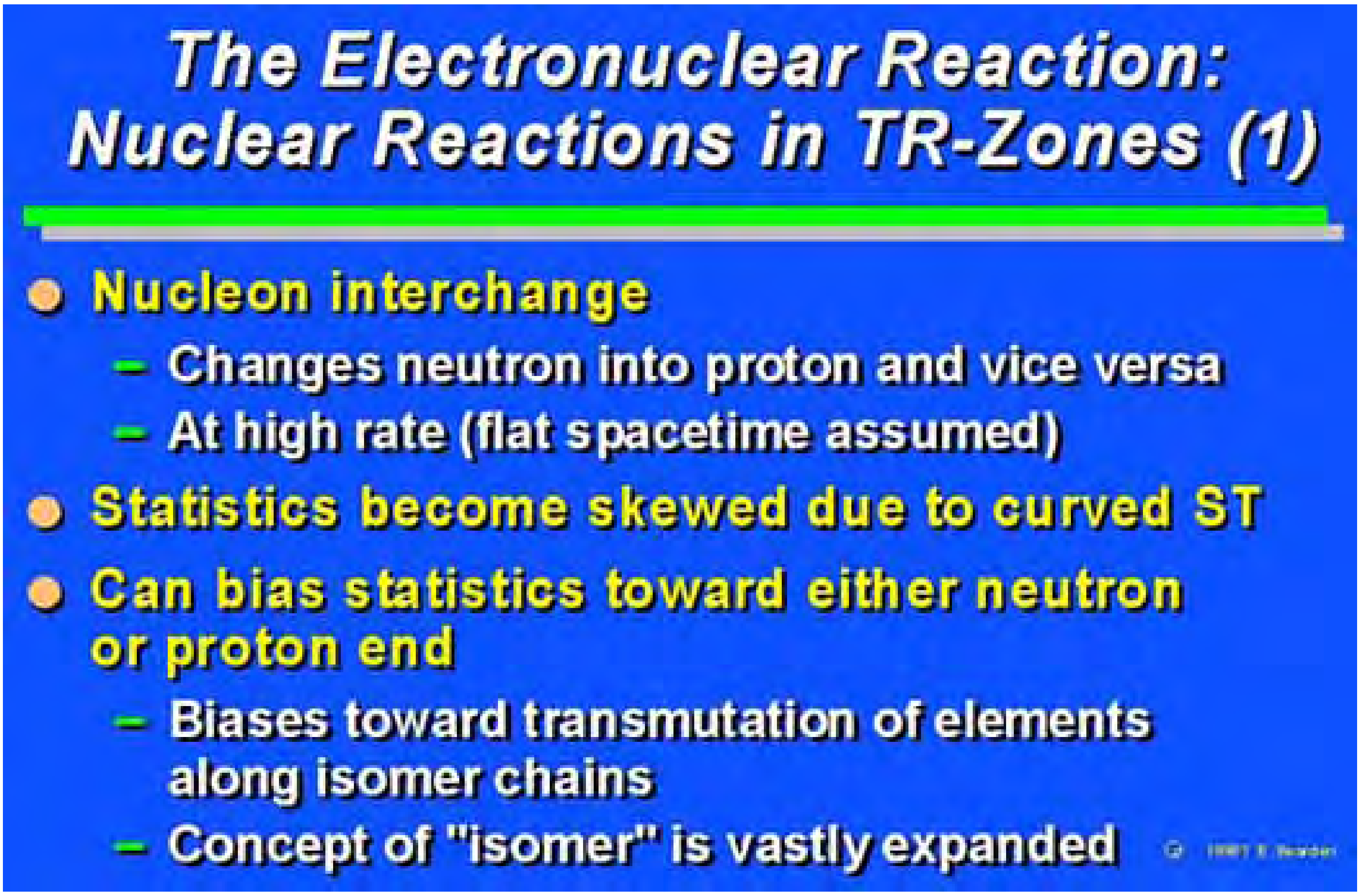

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- **Nucleon interchange**
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- **Statistics become skewed due to curved ST**
- **Can bias statistics toward either neutron or proton end**
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

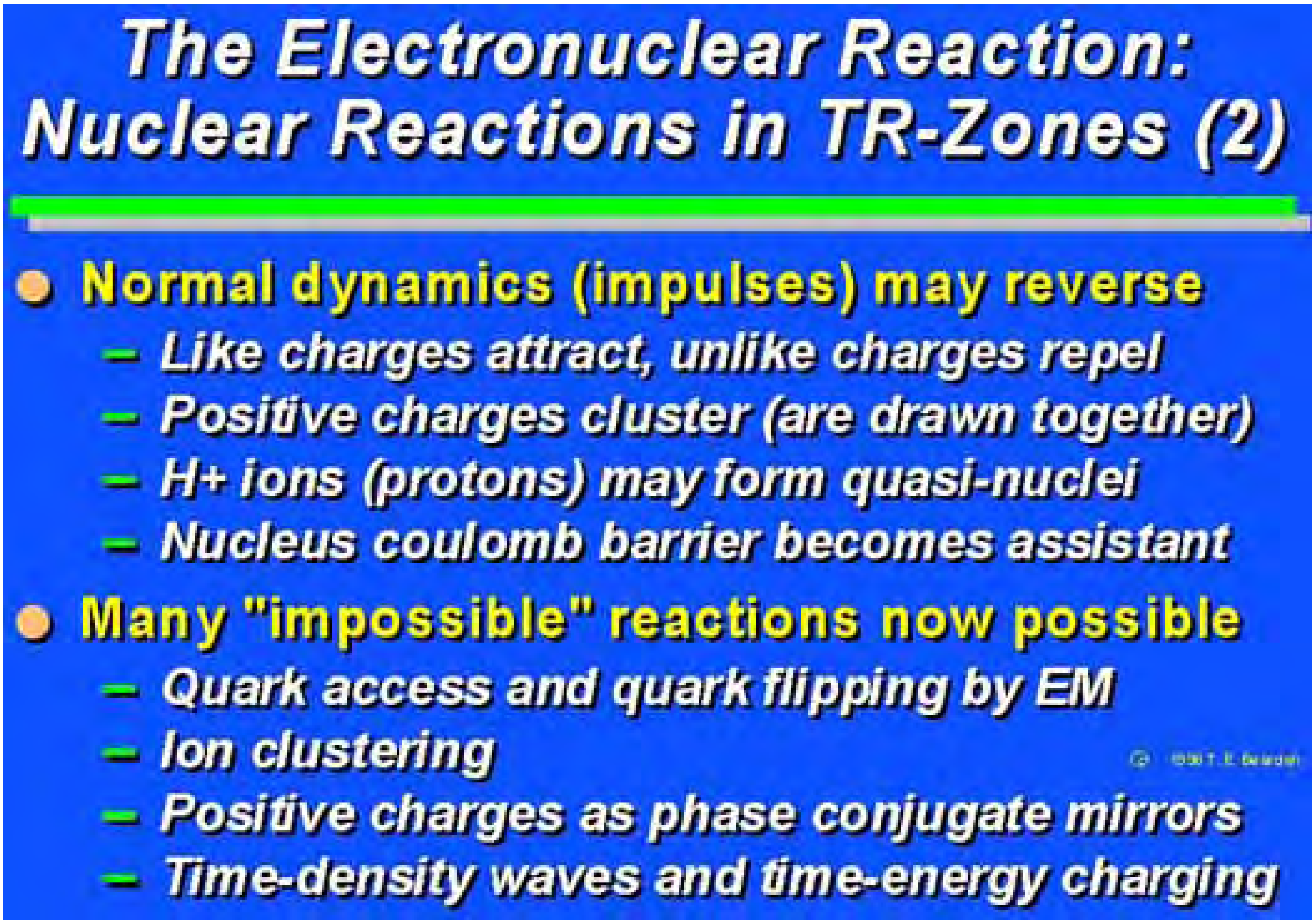

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

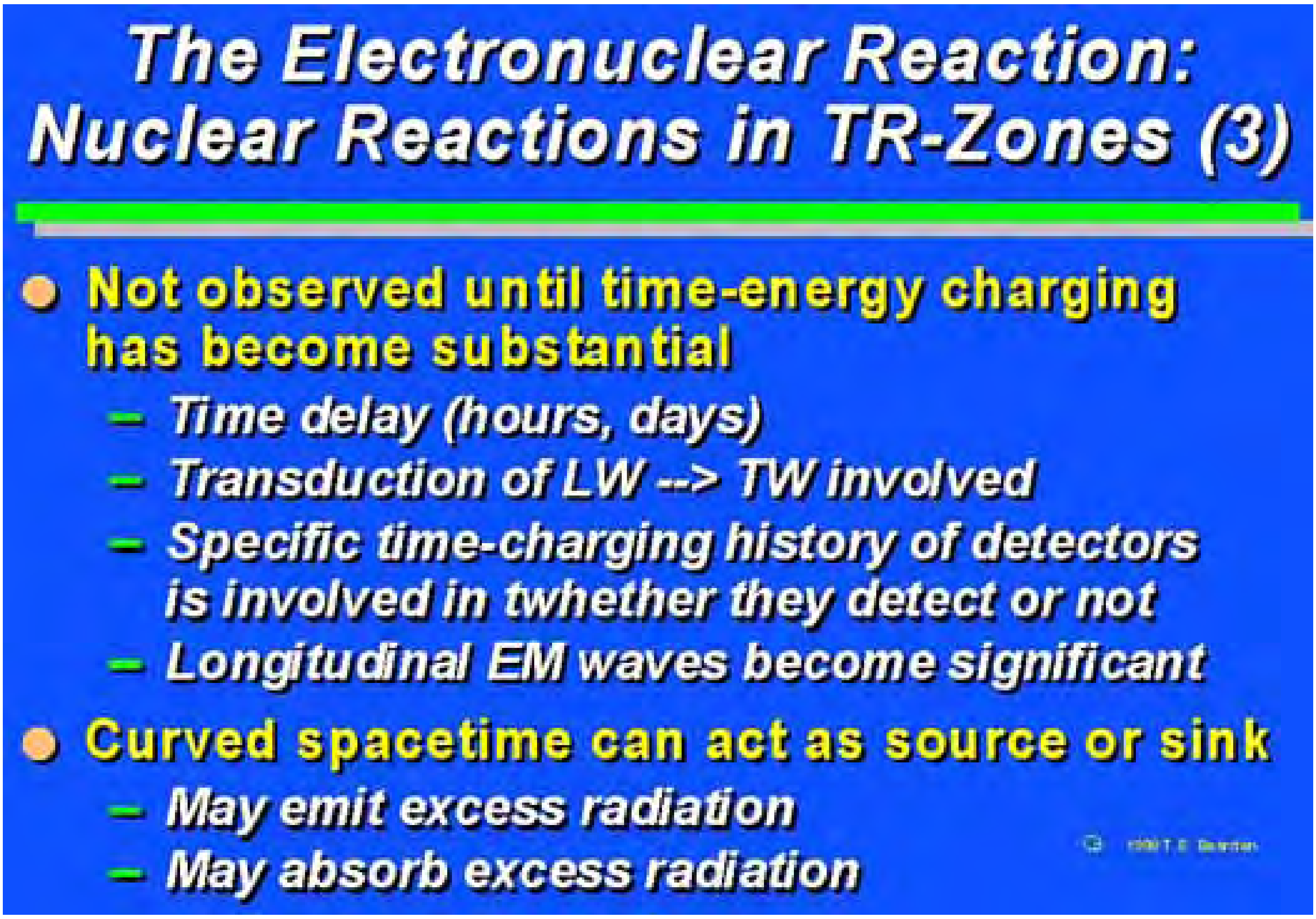

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

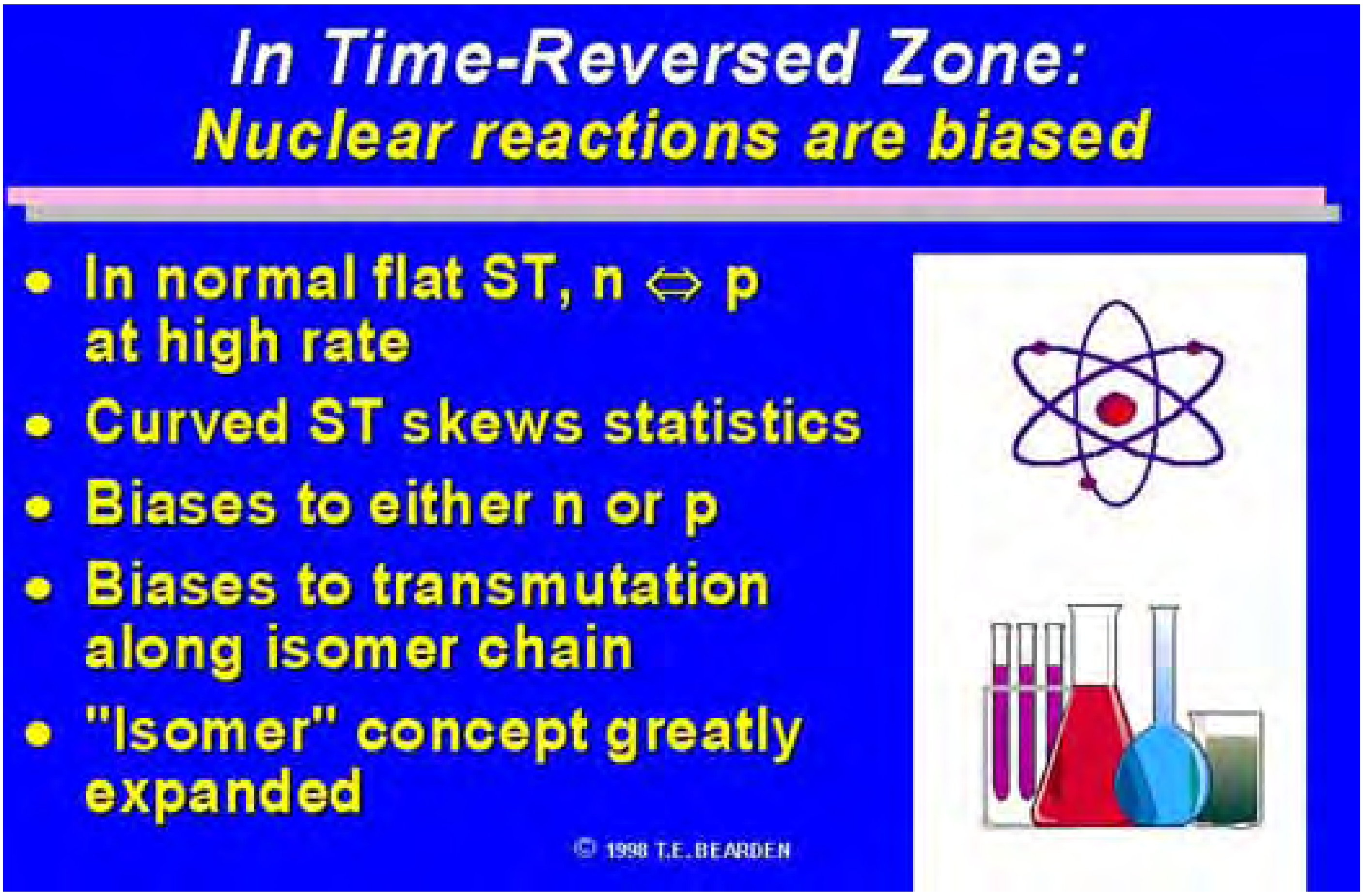

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

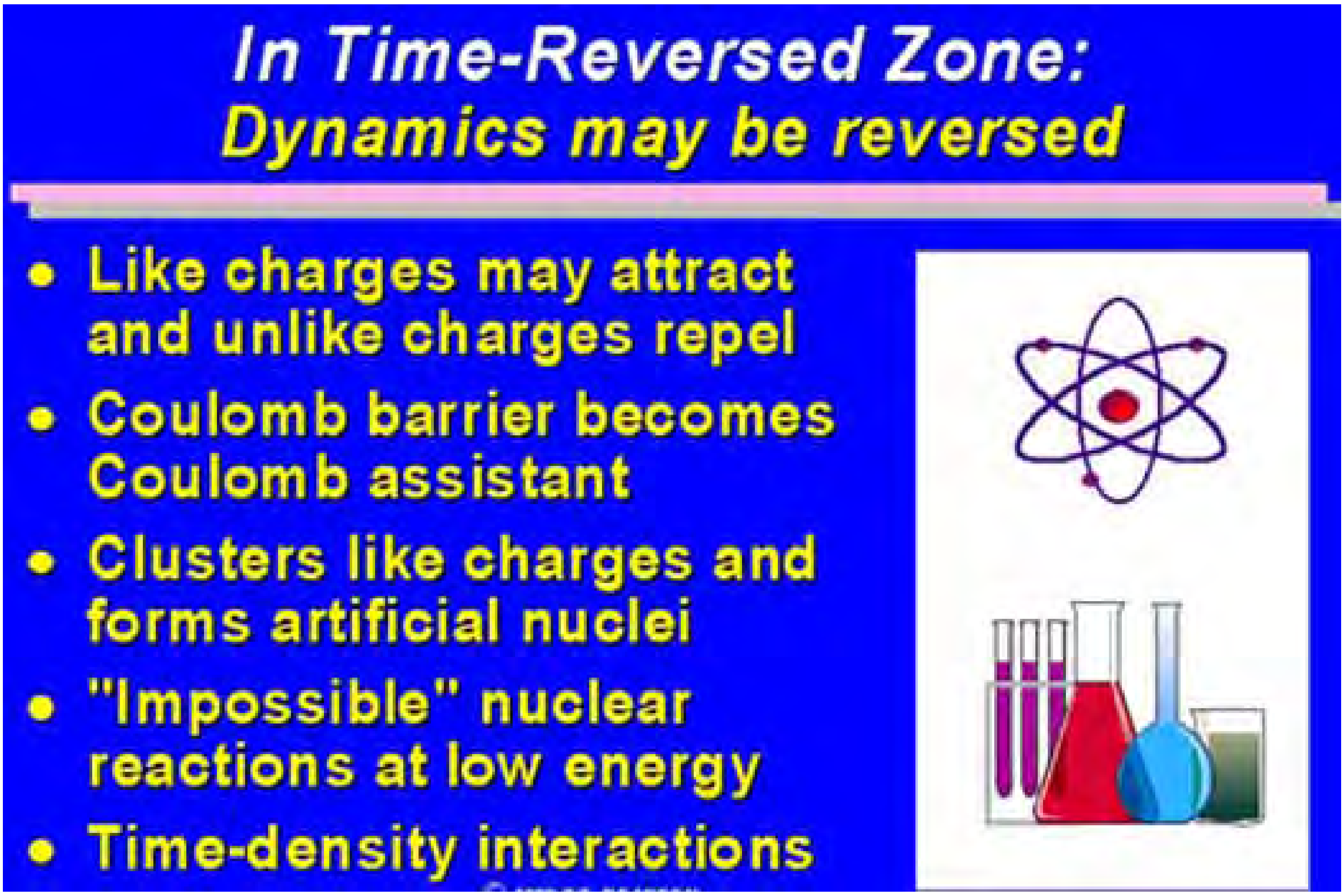

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

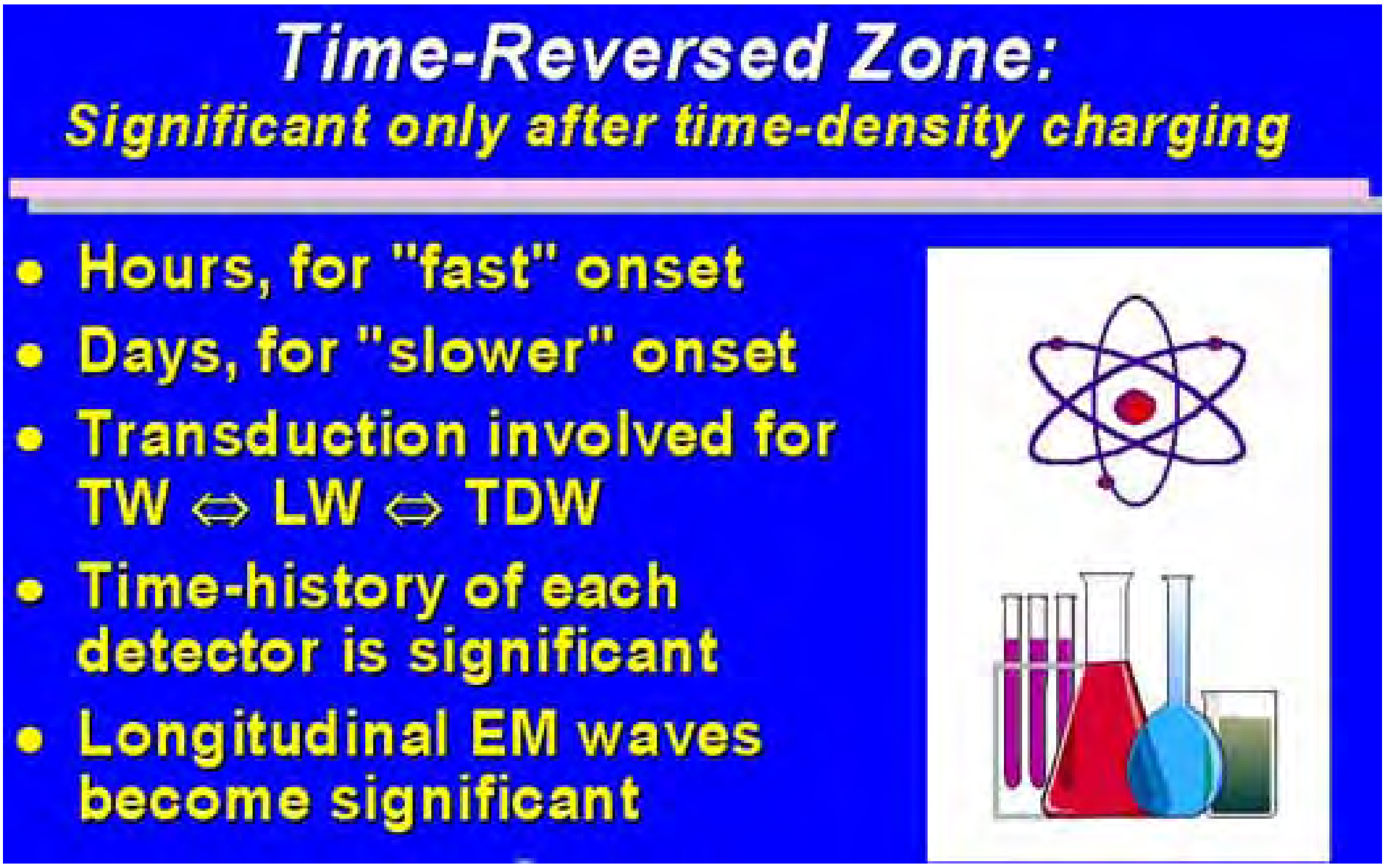

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

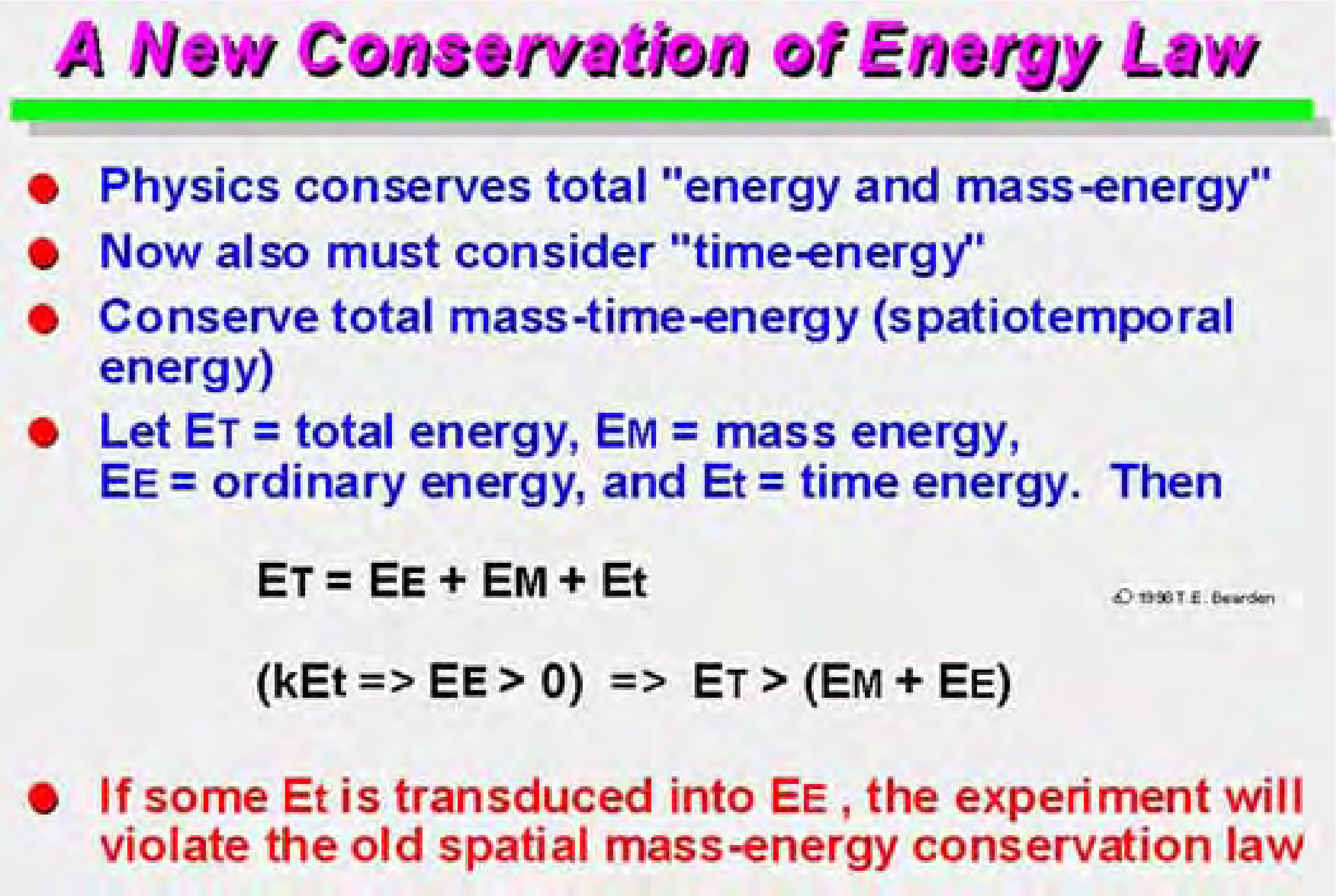

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T.E. Bearden

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

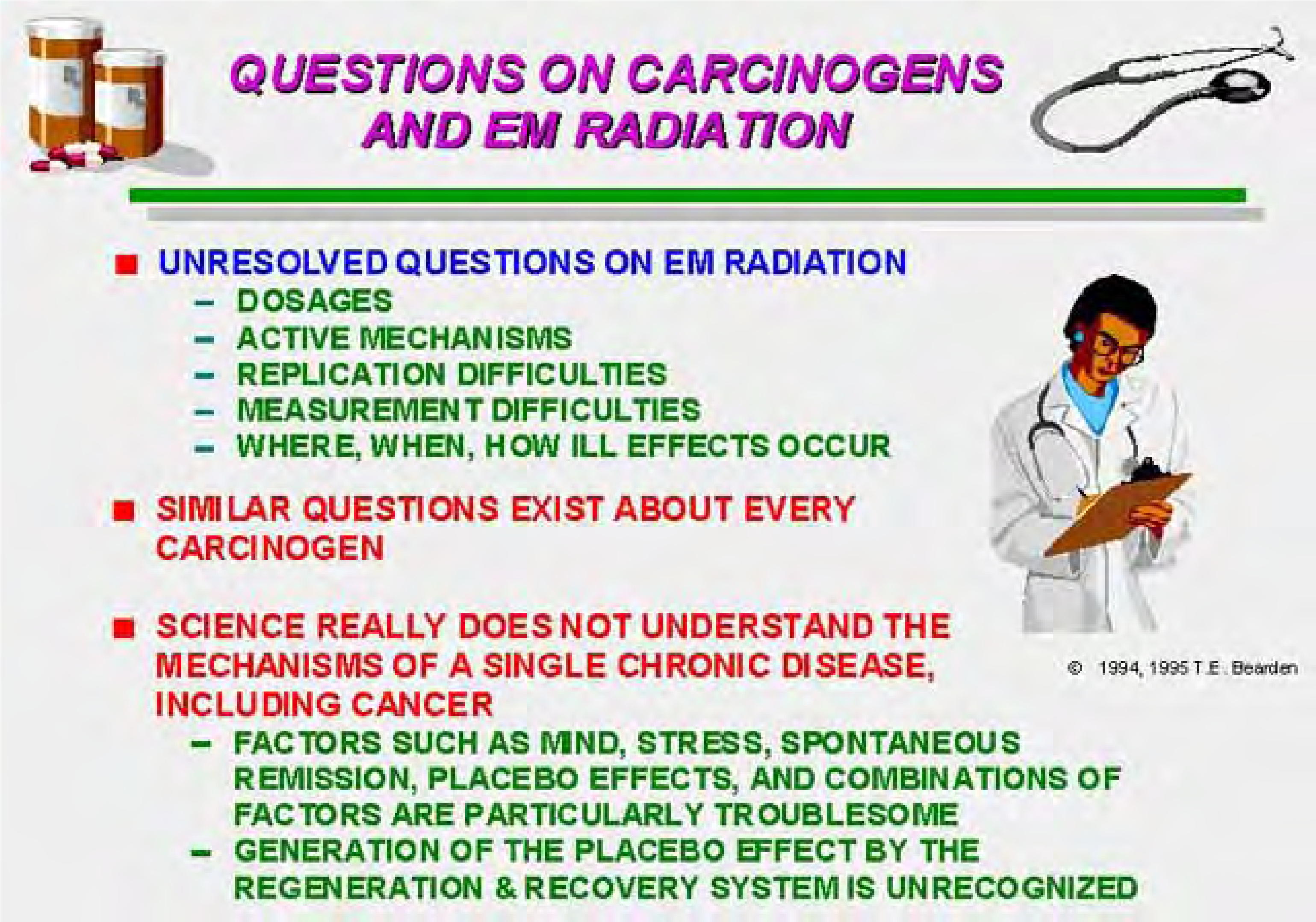

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

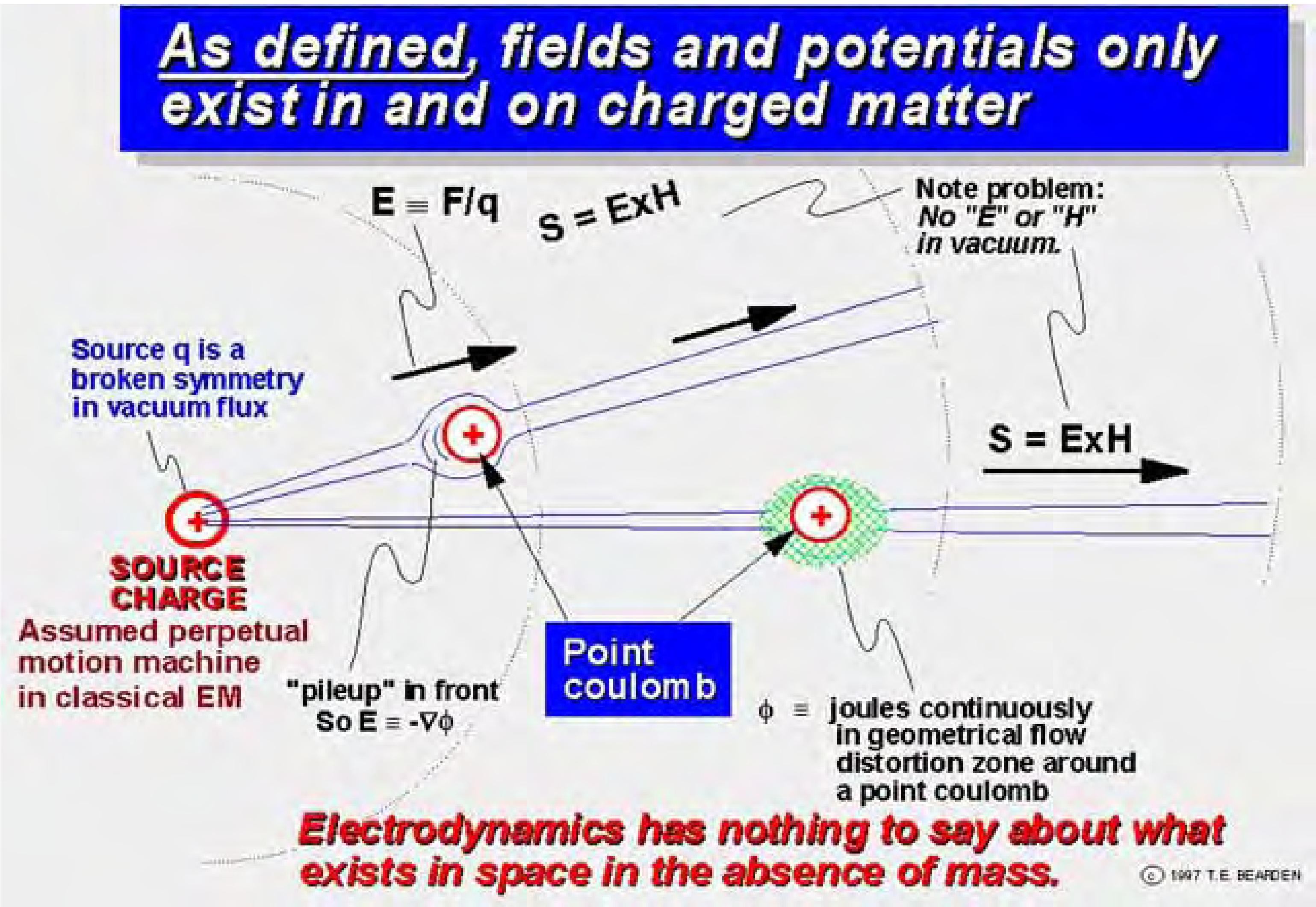

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

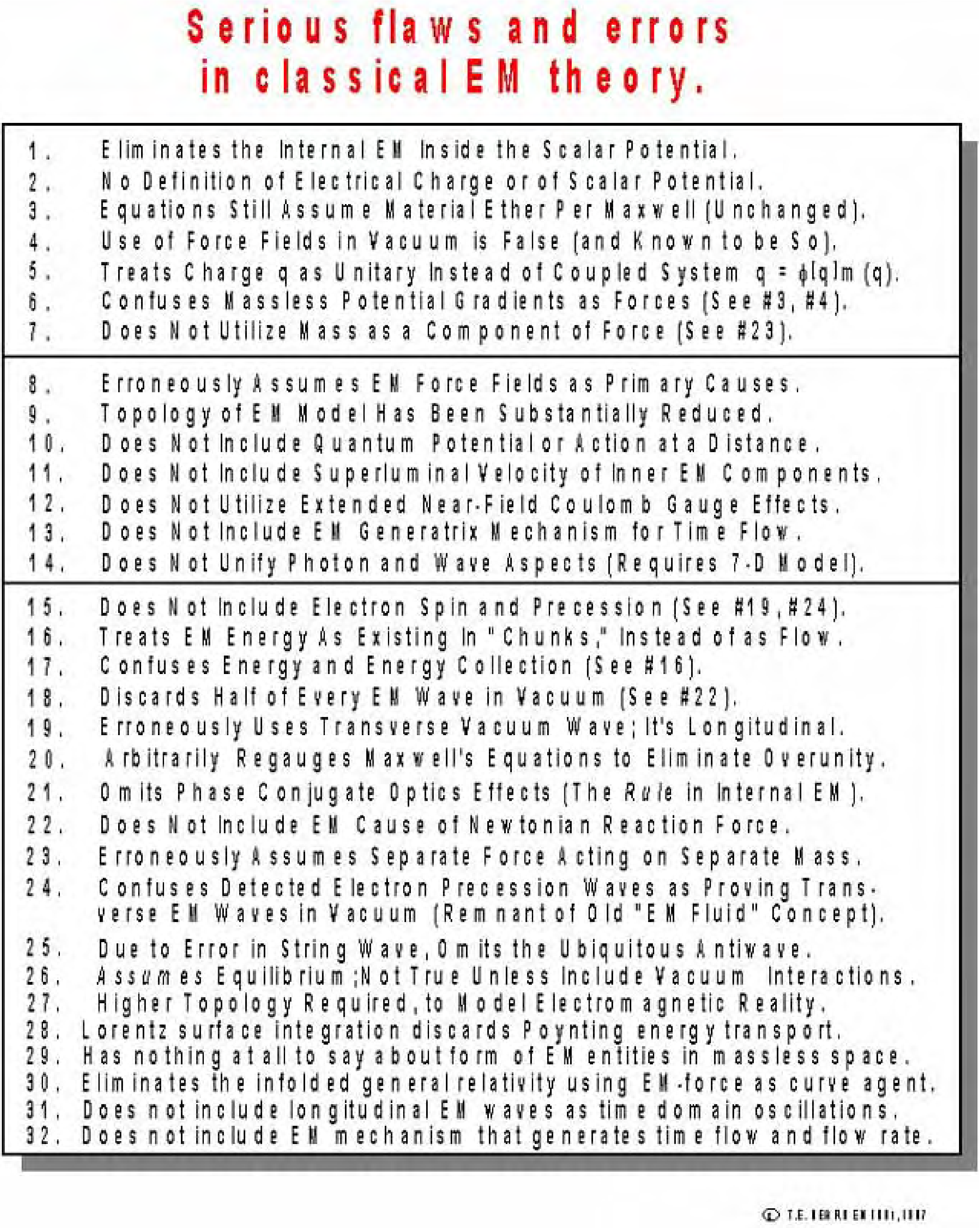

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).
8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).
15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

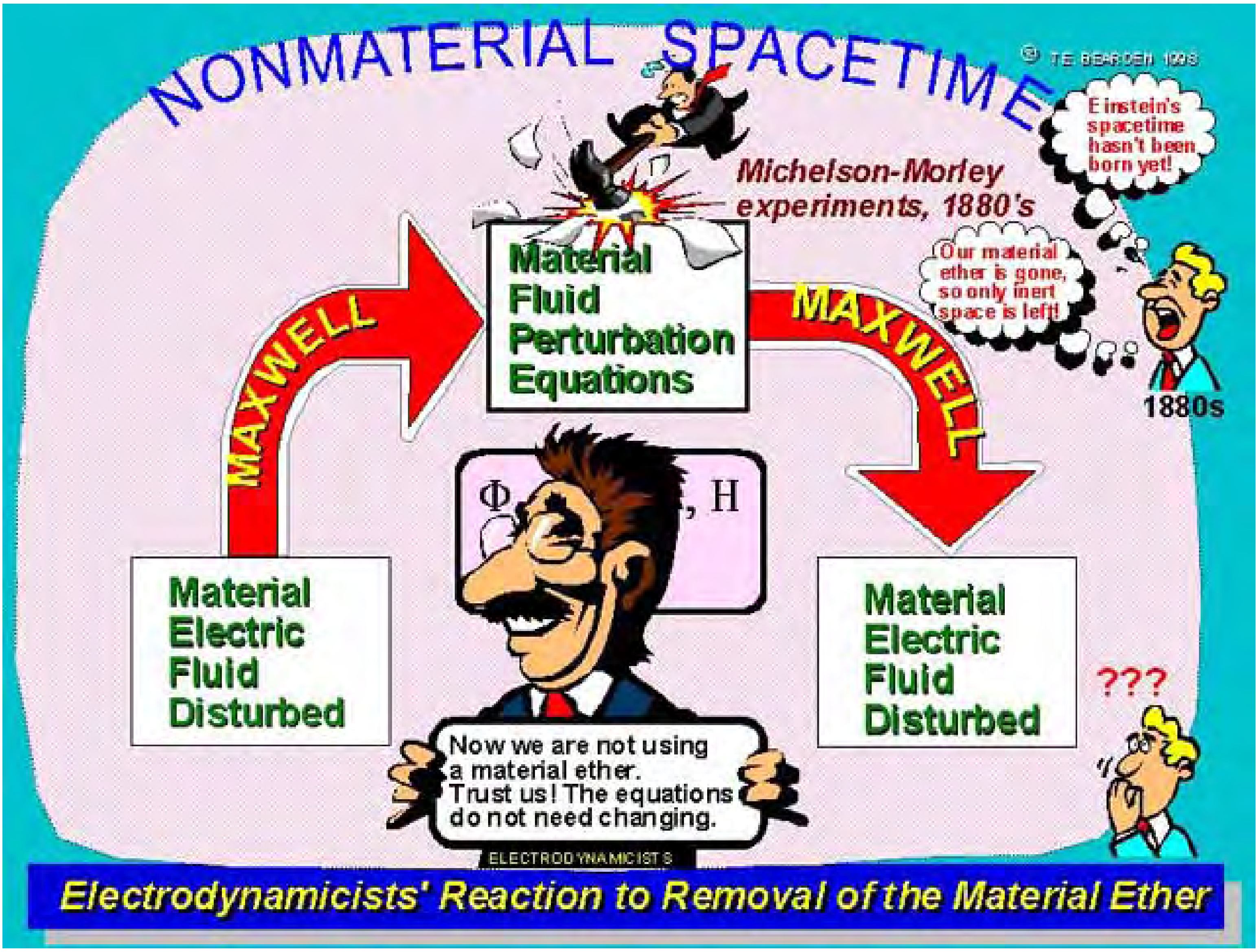
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ , H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

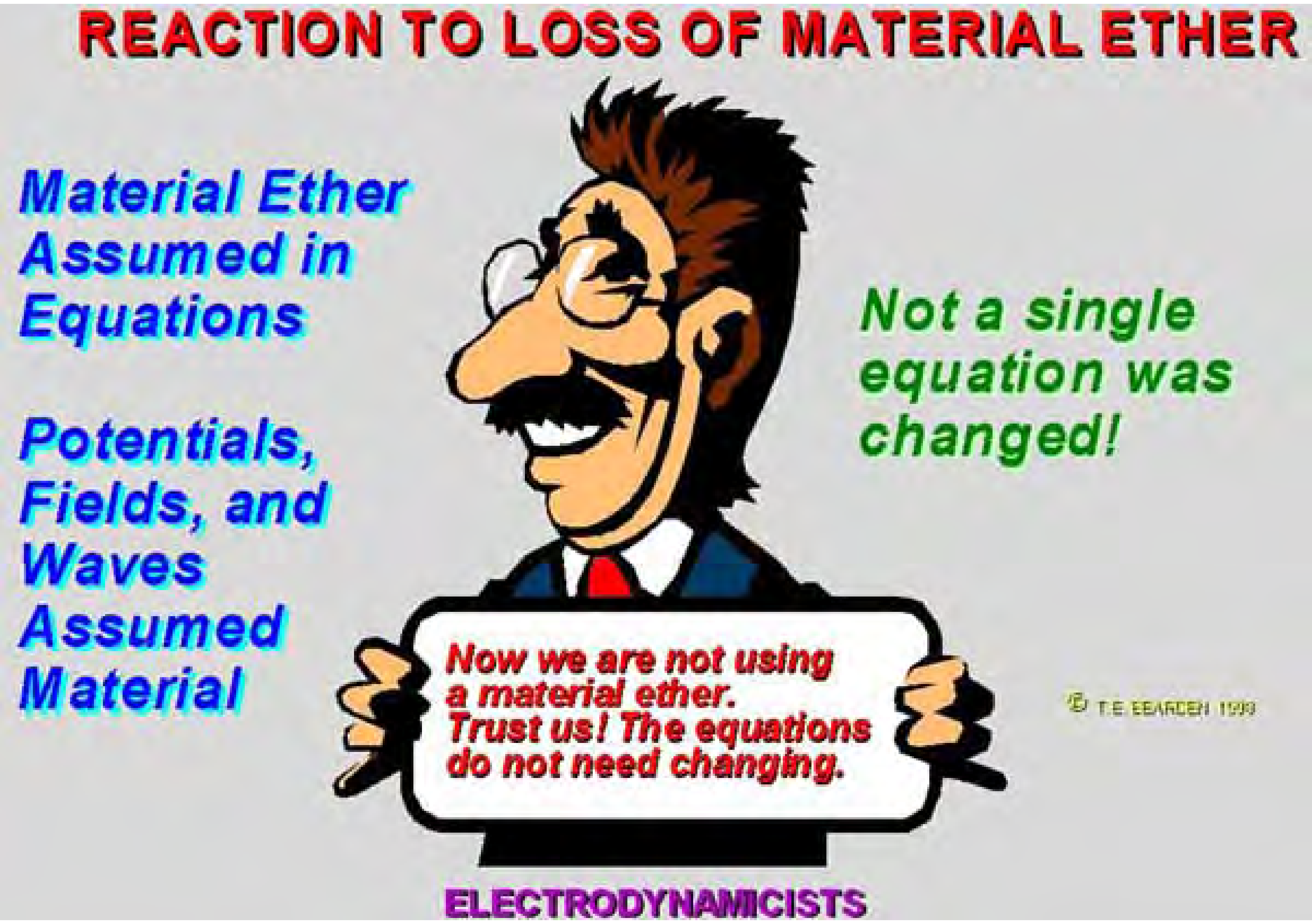

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.

© T.E. BEARDEN 1999

ELECTRODYNAMICISTS

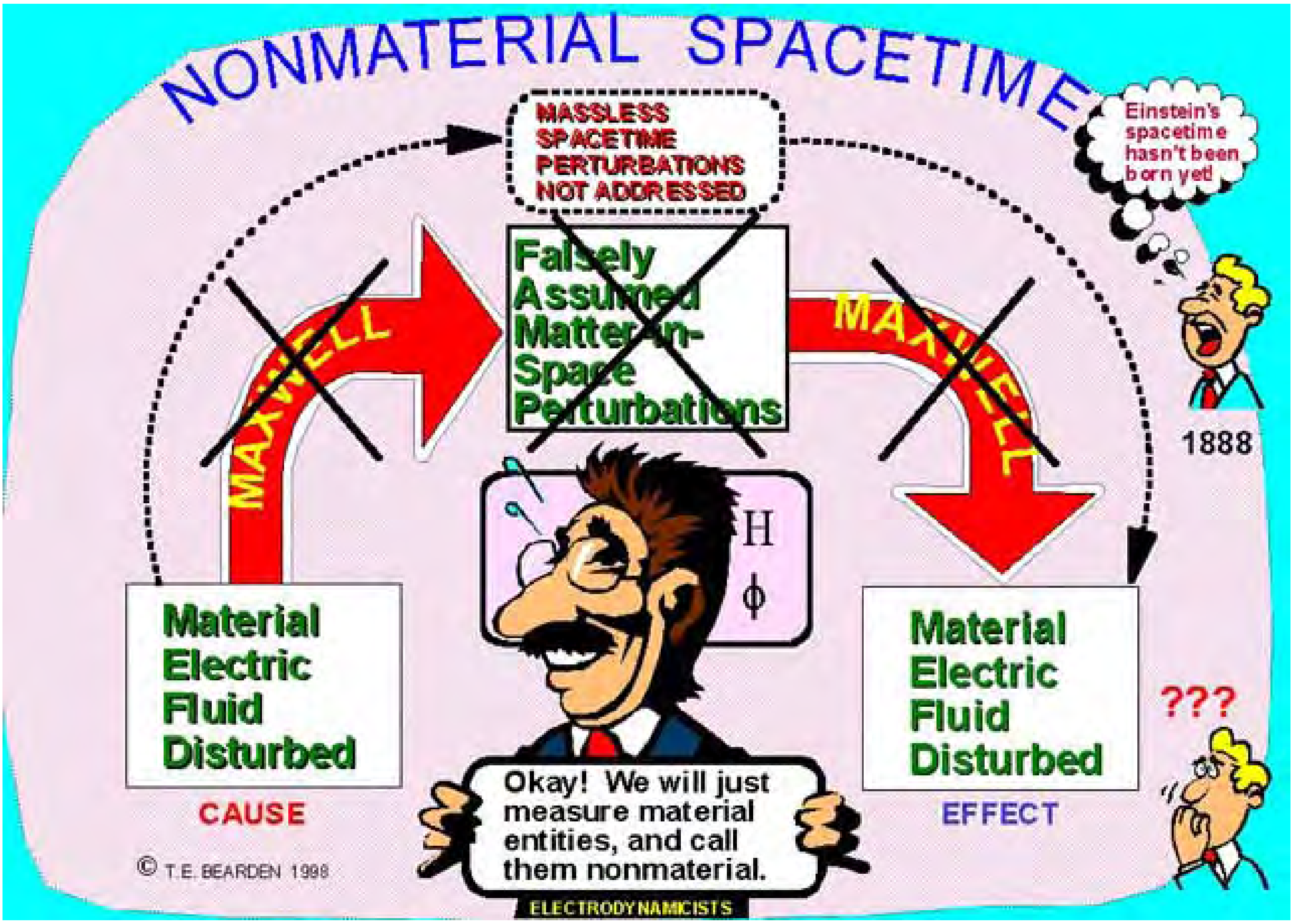
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
~~Falsely Assumed Matter-in-Space Perturbations~~
~~MAXWELL~~
~~MAXWELL~~
Einstein's spacetime hasn't been born yet!
1888
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

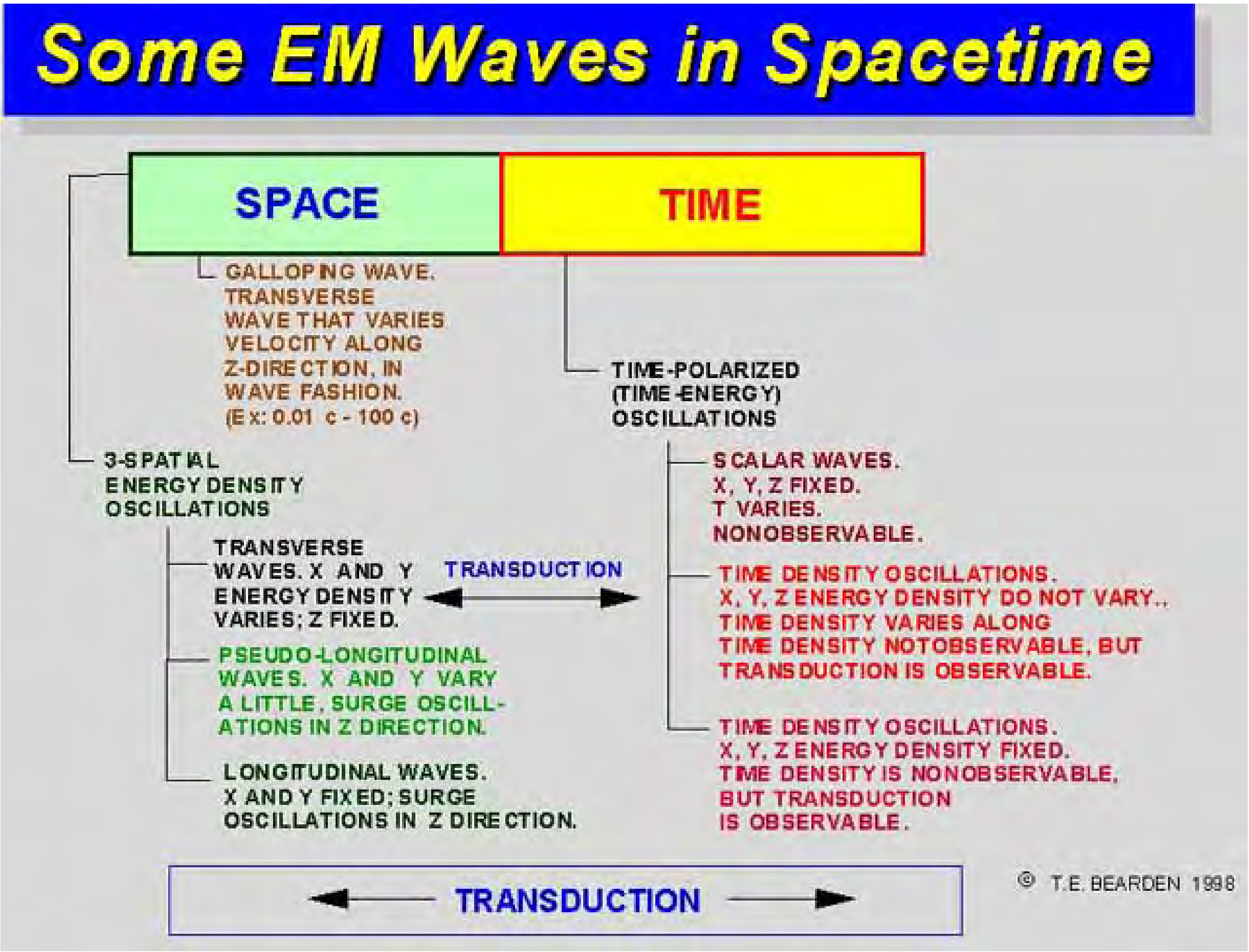

Some EM Waves in Spacetime

SPACE | **TIME**

GALLOPING WAVE. TRANSVERSE WAVE THAT VARIES VELOCITY ALONG Z-DIRECTION, IN WAVE FASHION. (Ex: 0.01 c - 100 c)

TIME-POLARIZED (TIME-ENERGY) OSCILLATIONS

- SCALAR WAVES. X, Y, Z FIXED. T VARIES. NONOBSERVABLE.
- TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY DO NOT VARY.. TIME DENSITY VARIES ALONG TIME DENSITY NOTOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.
- TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY FIXED. TIME DENSITY IS NONOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

3-SPATIAL ENERGY DENSITY OSCILLATIONS

- TRANSVERSE WAVES. X AND Y ENERGY DENSITY VARIES; Z FIXED.
- PSEUDO-LONGITUDINAL WAVES. X AND Y VARY A LITTLE, SURGE OSCILLATIONS IN Z DIRECTION.
- LONGITUDINAL WAVES. X AND Y FIXED; SURGE OSCILLATIONS IN Z DIRECTION.

◄— TRANSDUCTION —►

© T.E. BEARDEN 1998

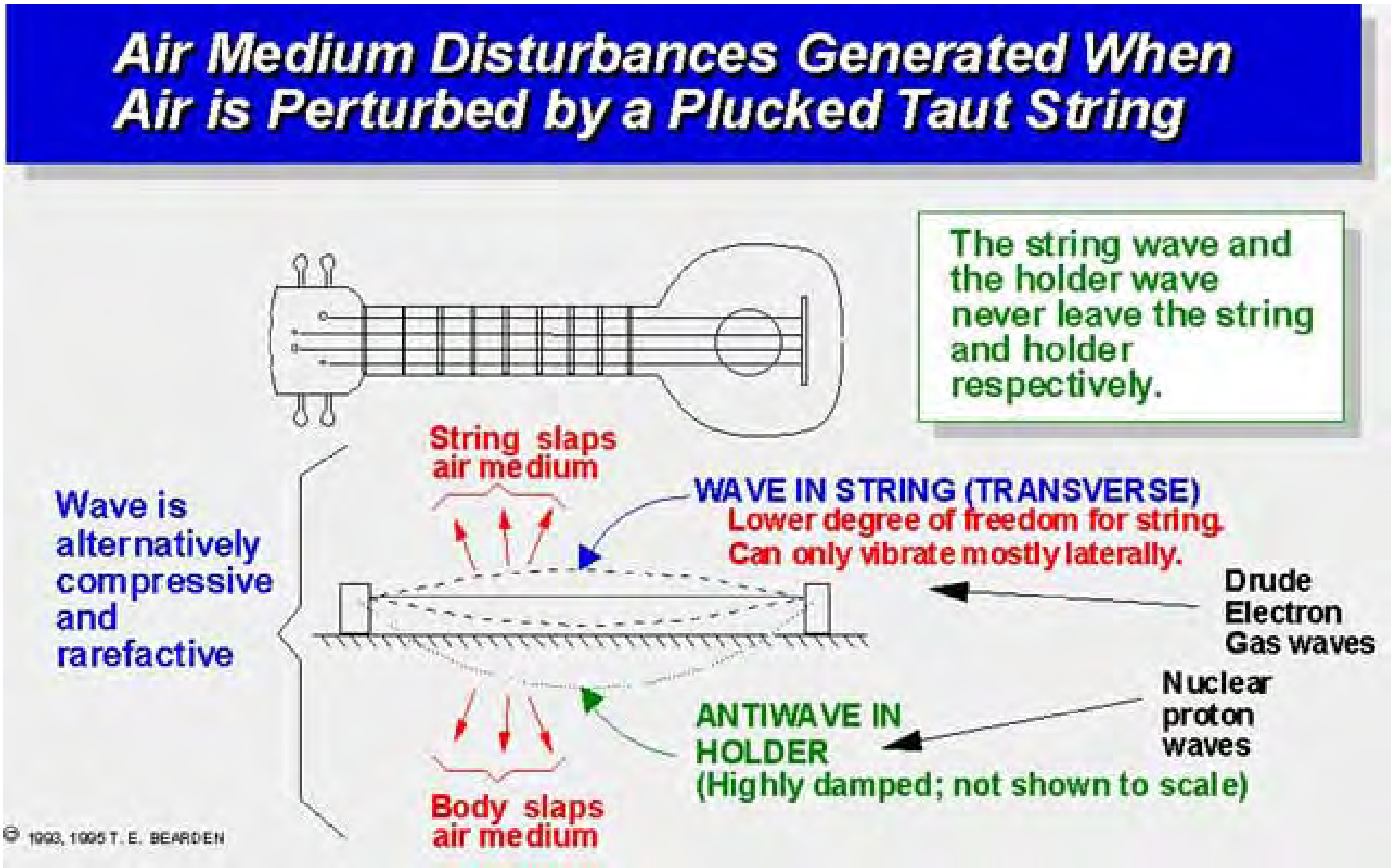

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string. Can only vibrate mostly laterally.

Drude Electron Gas waves

Wave is alternatively compressive and rarefactive

Nuclear proton waves

ANTIWAVE IN HOLDER (Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

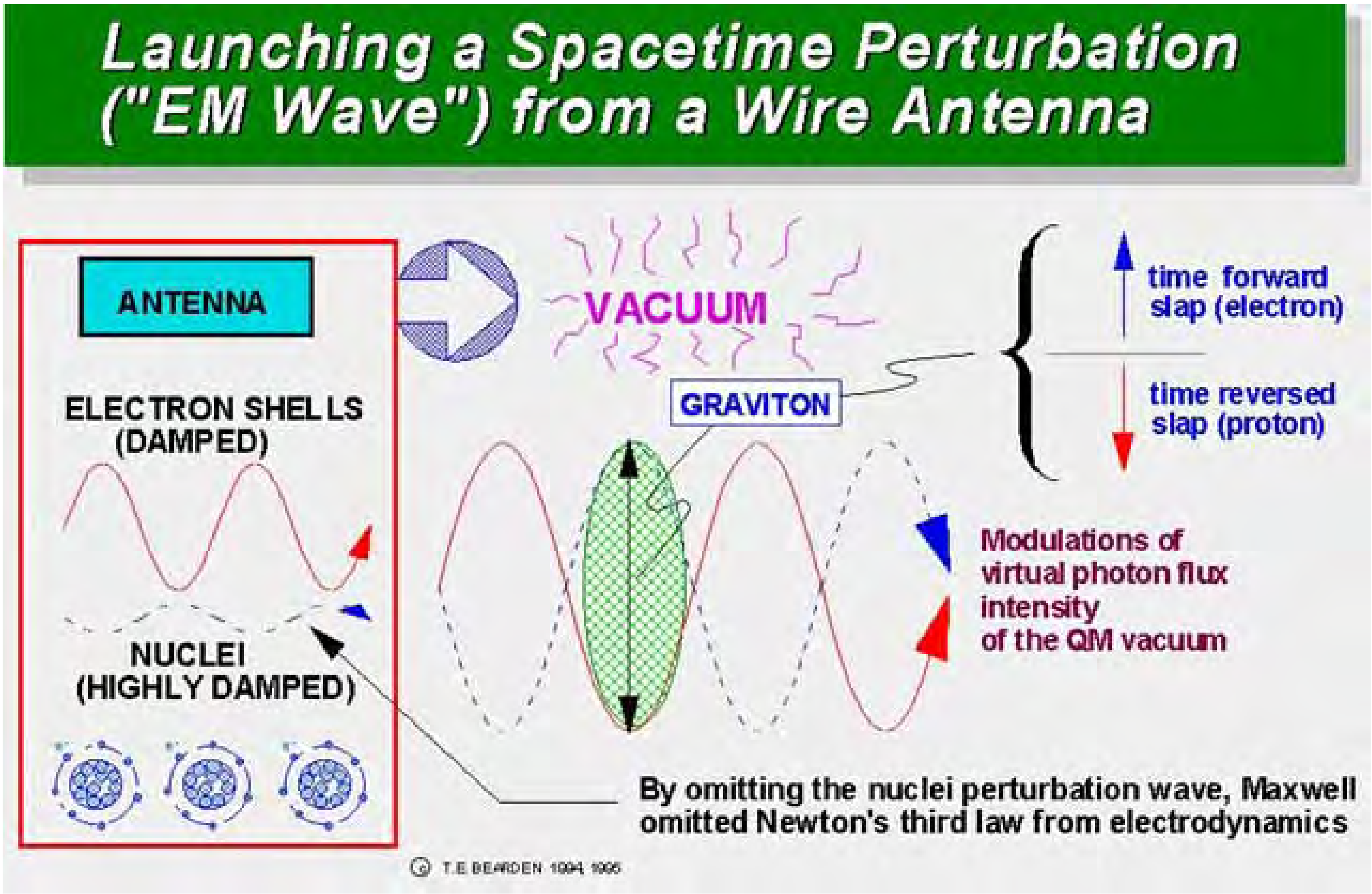

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA

ELECTRON SHELLS
(DAMPED)

NUCLEI
(HIGHLY DAMPED)

VACUUM

GRAVITON

time forward
slap (electron)

time reversed
slap (proton)

Modulations of
virtual photon flux
intensity
of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

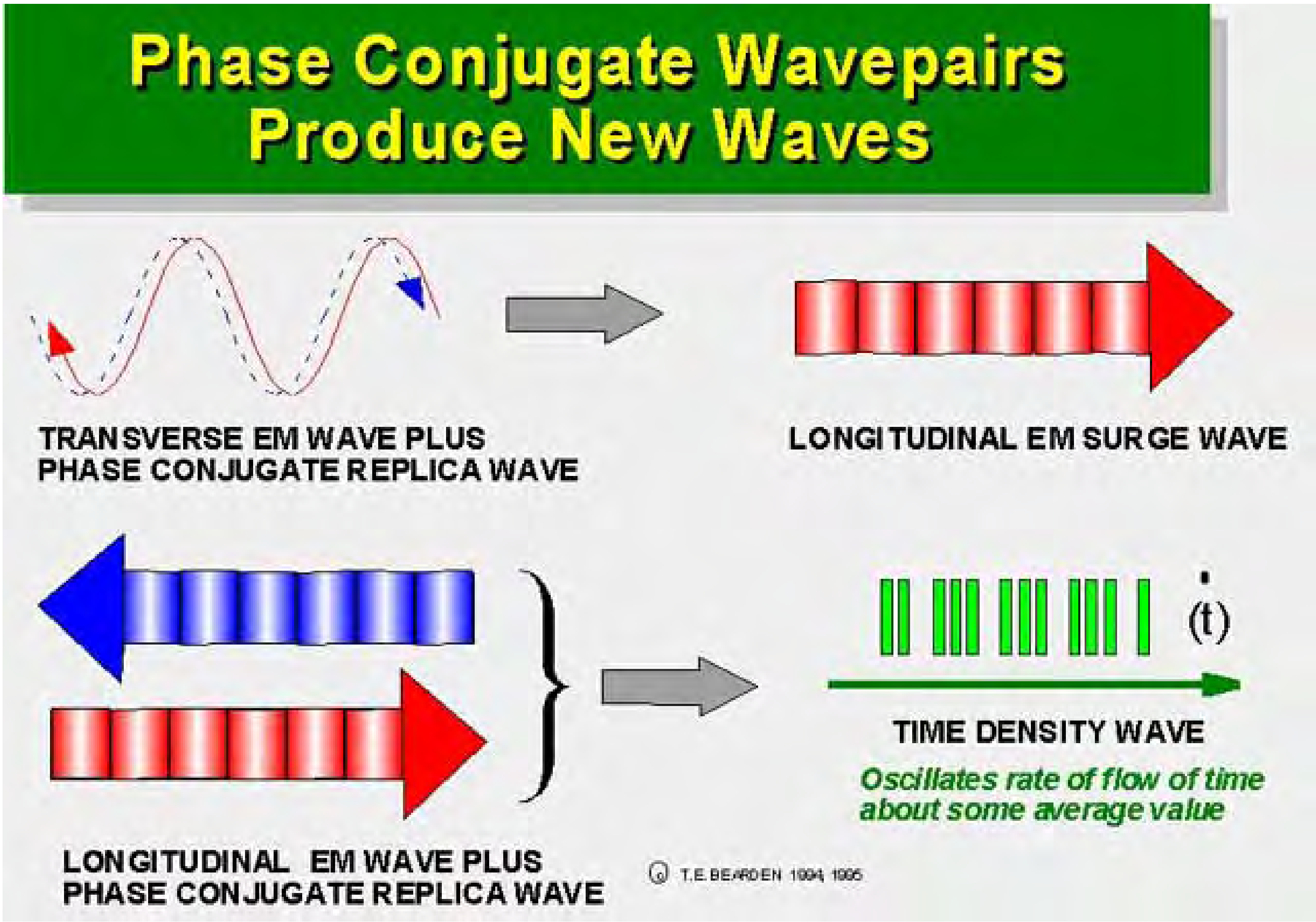

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

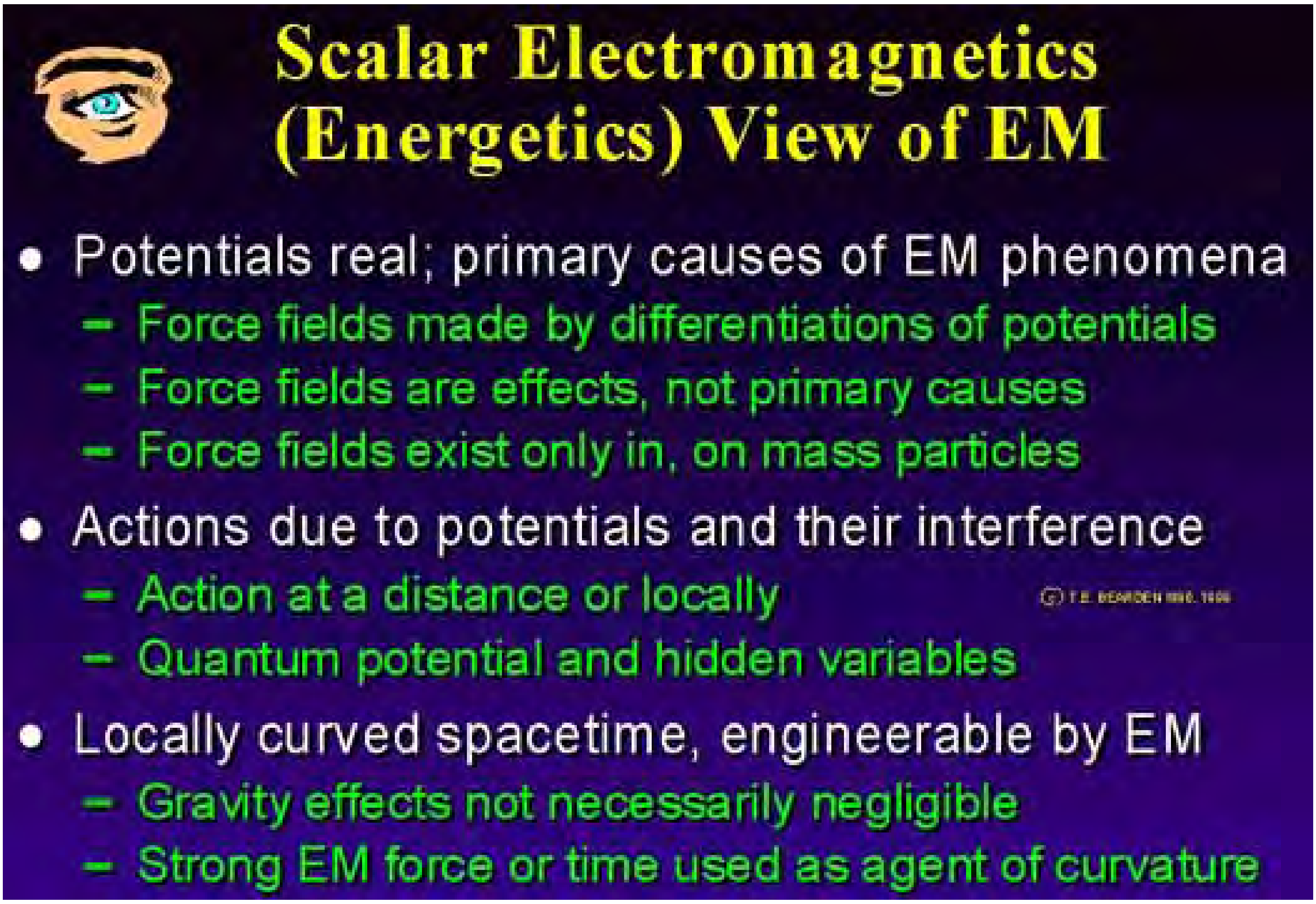

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

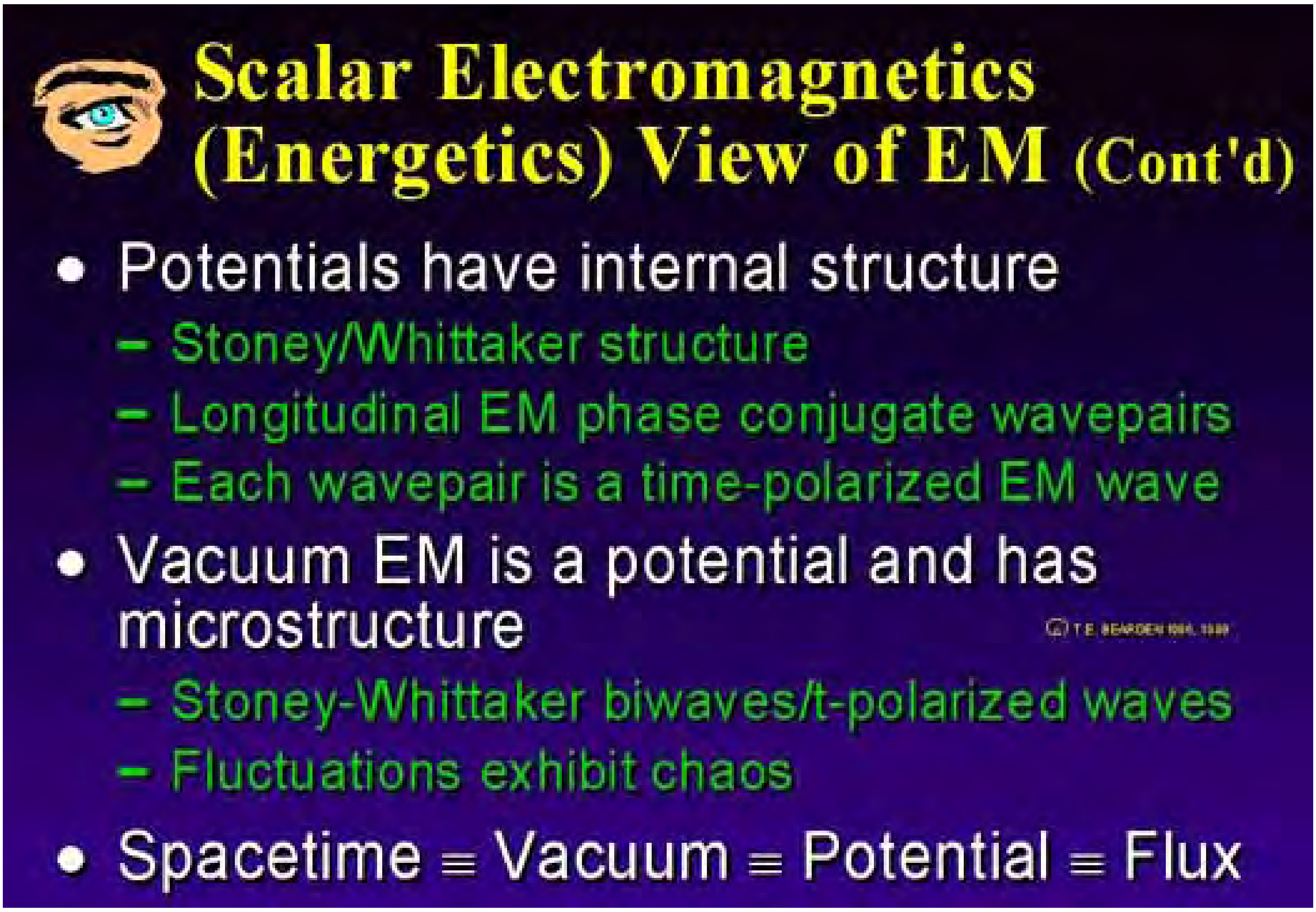

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. [illegible]

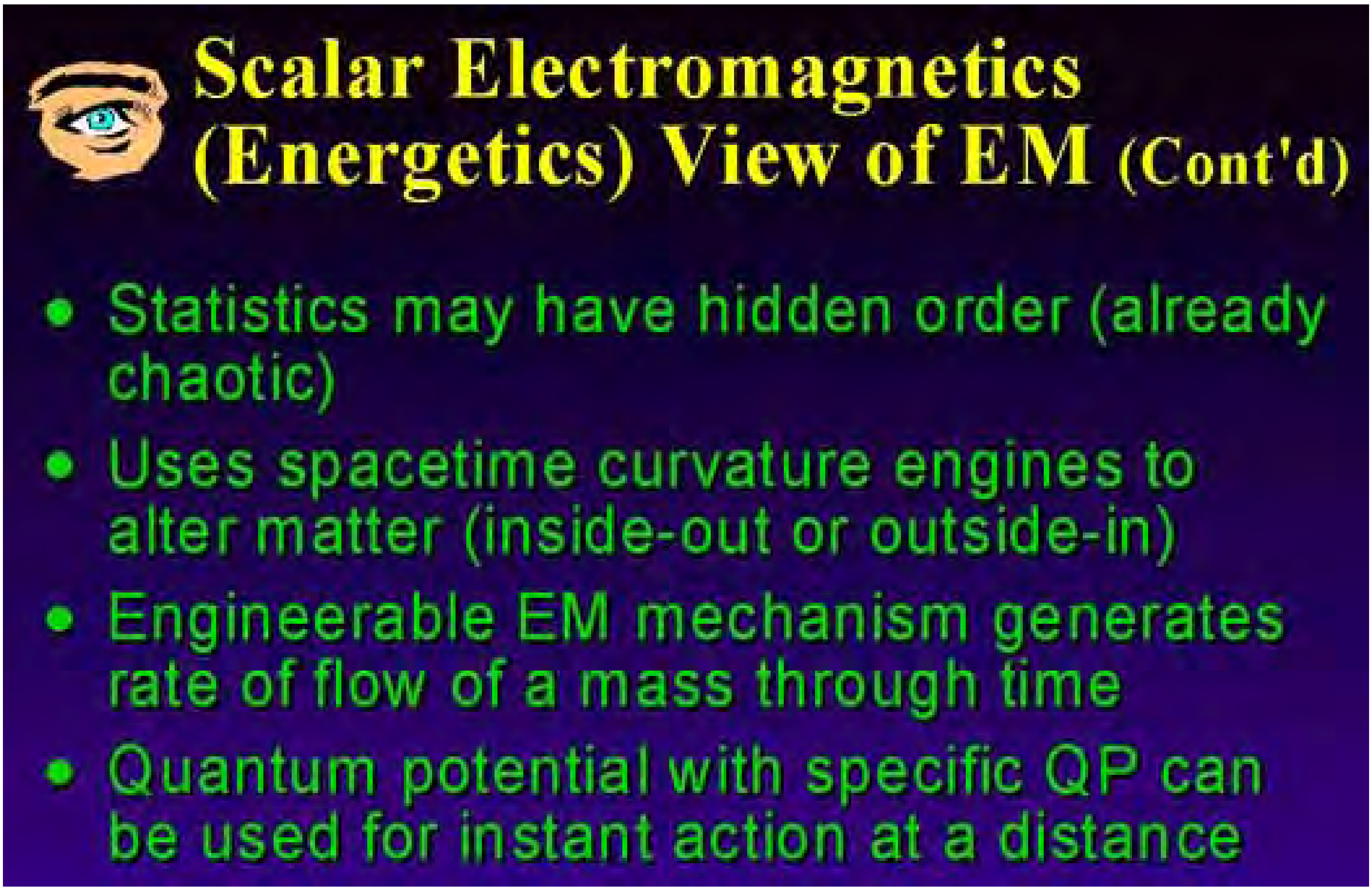

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

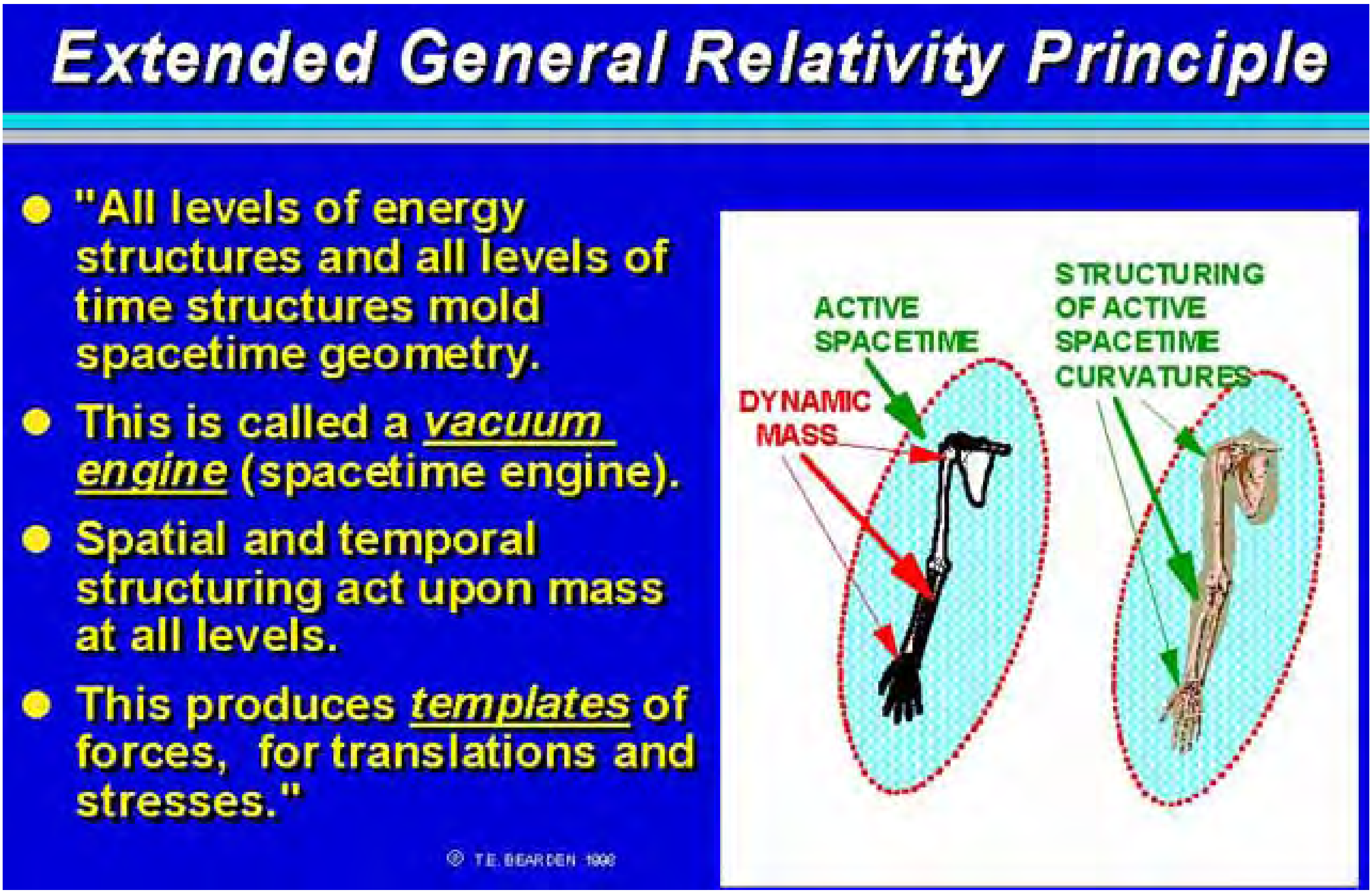

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a ***vacuum engine*** (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces ***templates*** of forces, for translations and stresses."

© T.E. BEARDEN 1990

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

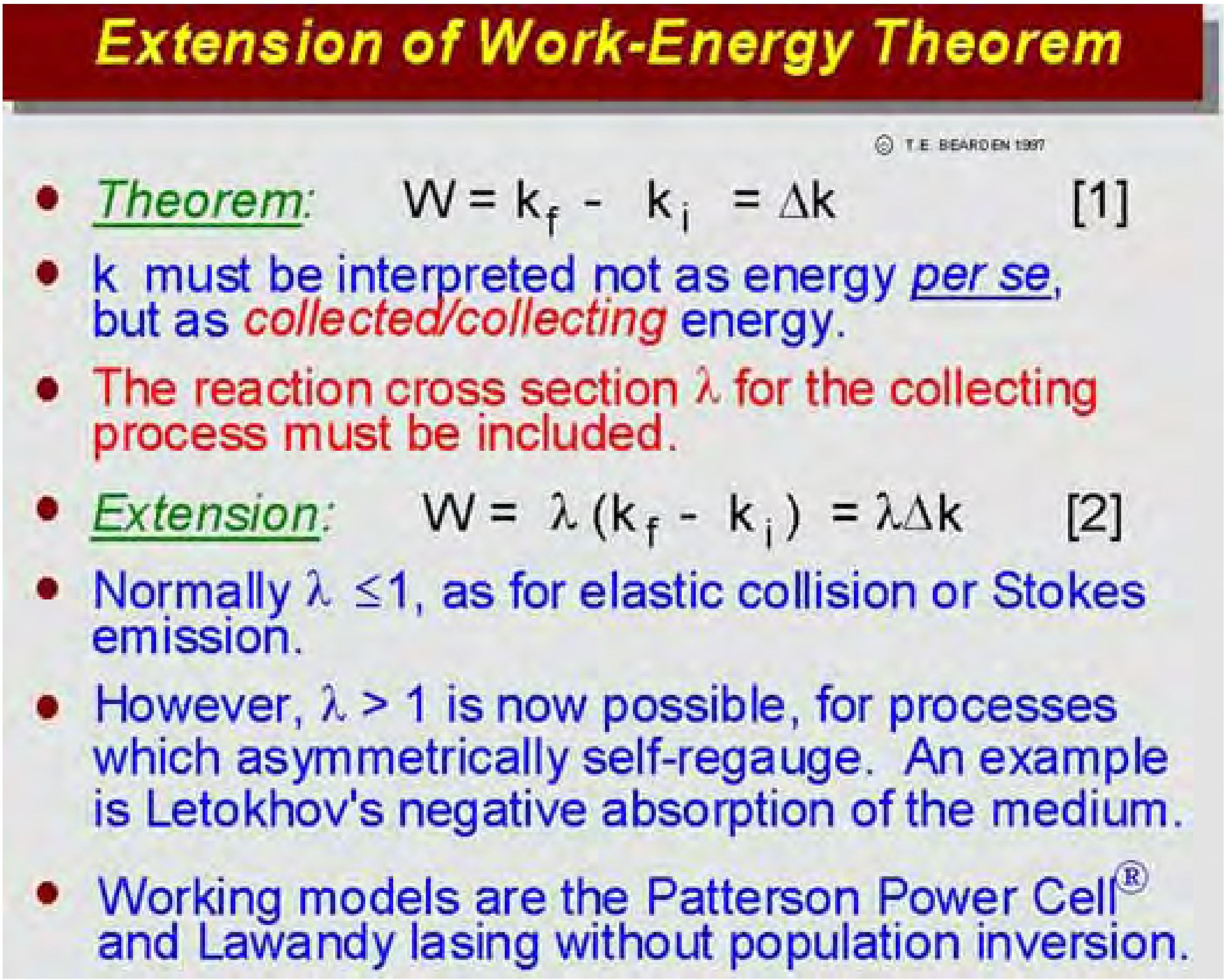

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem*: $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension*: $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

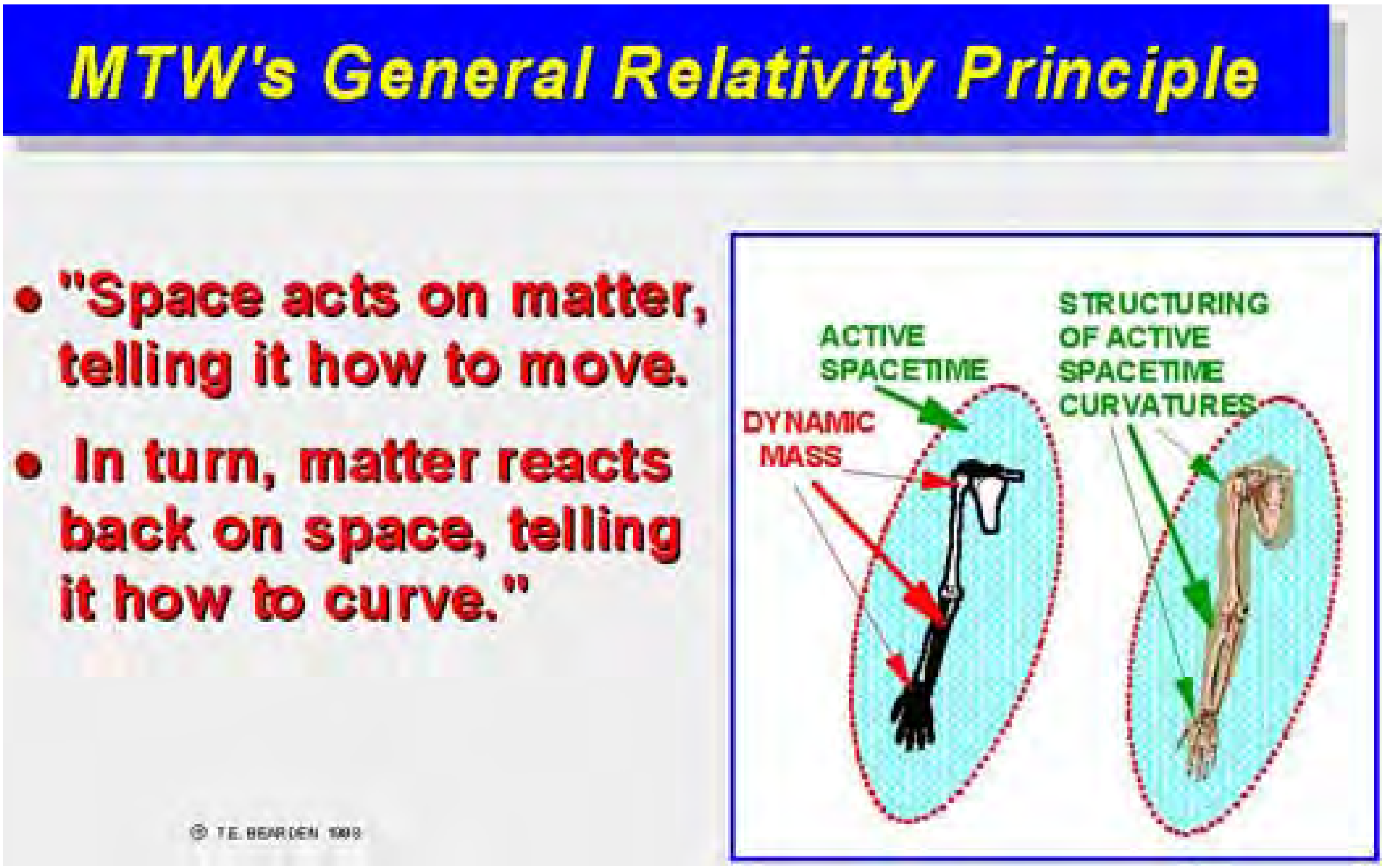

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

© T.E. BEARDEN 1998

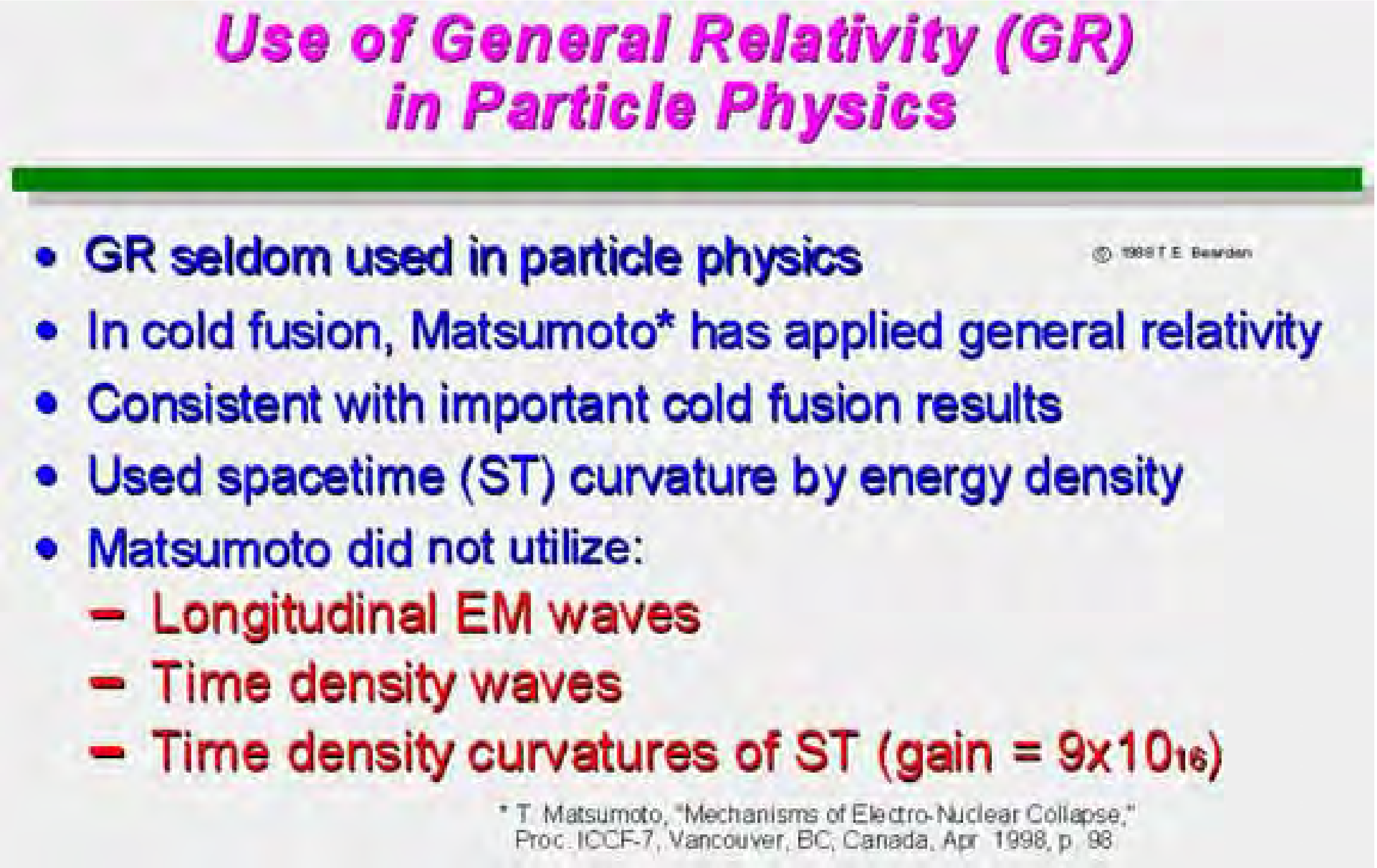

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = $9\text{x}10_{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98

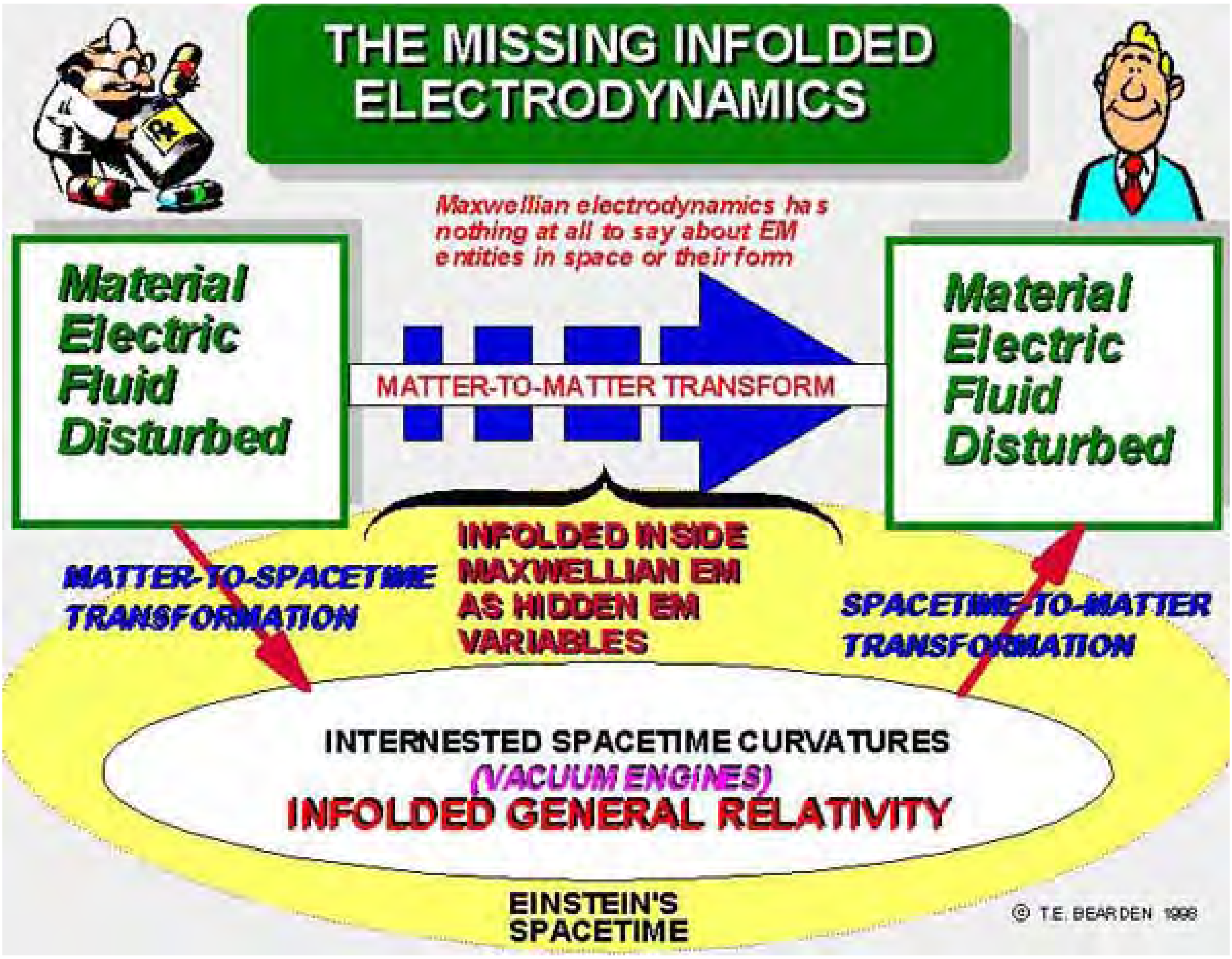

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

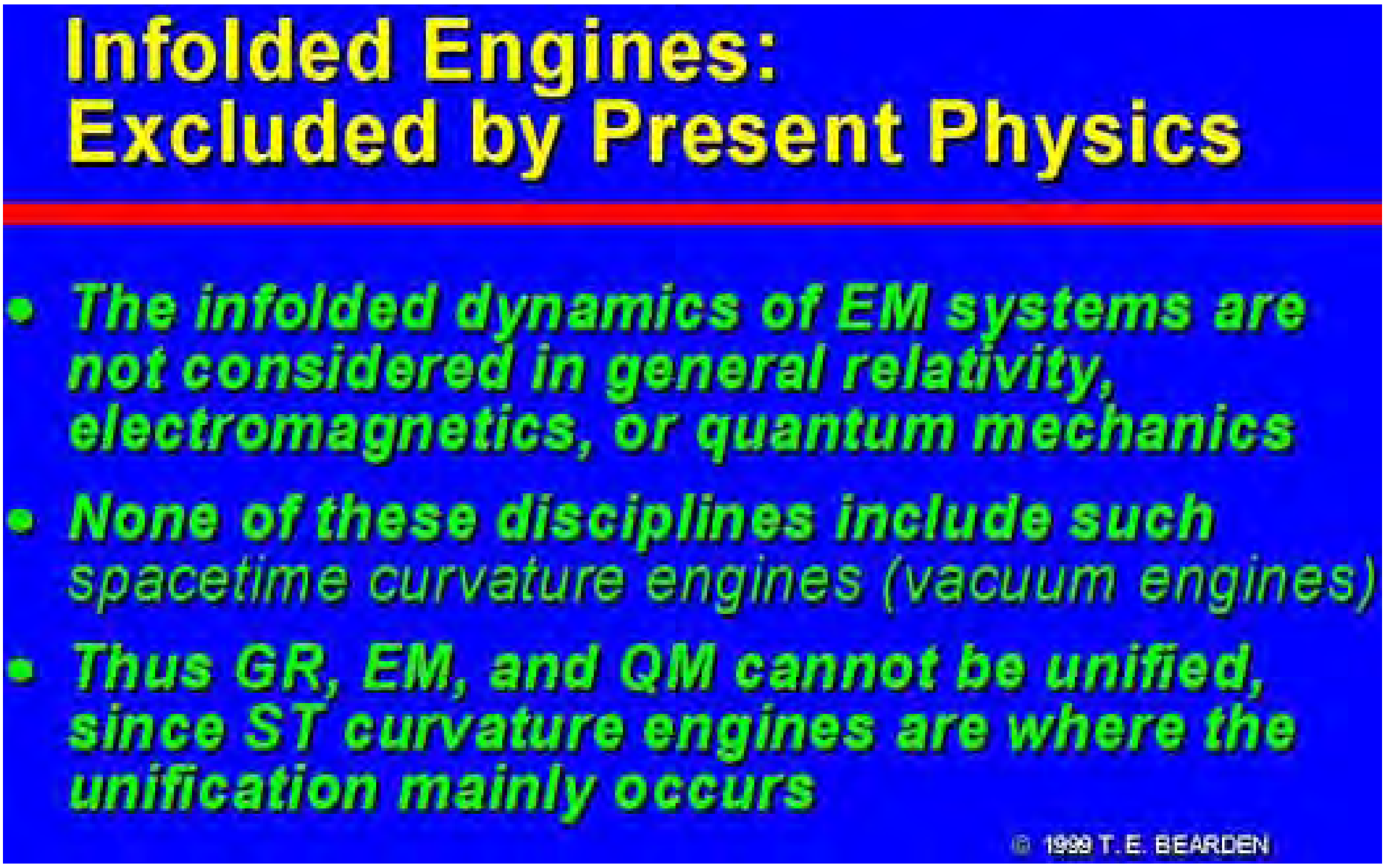

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such* *spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

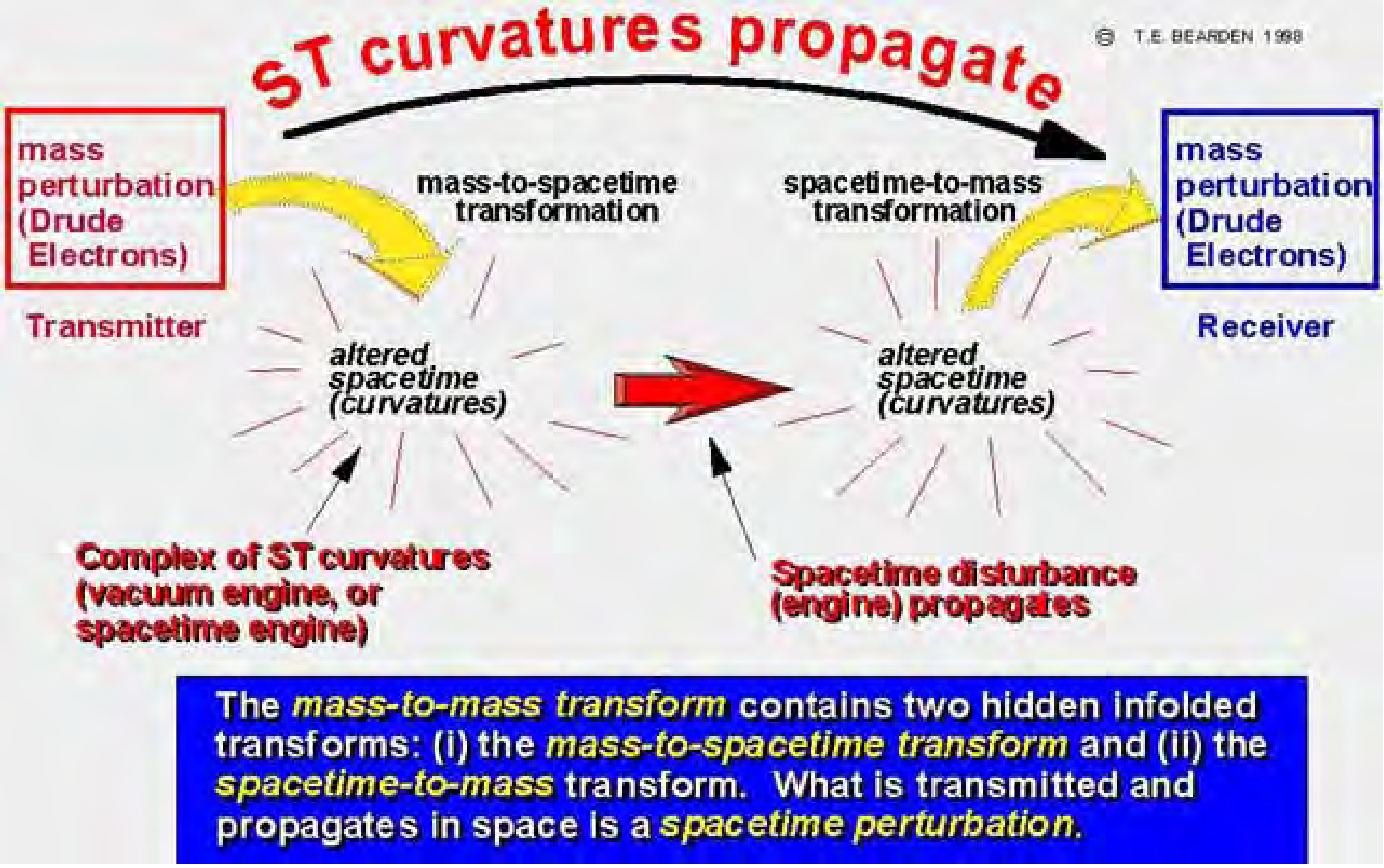

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

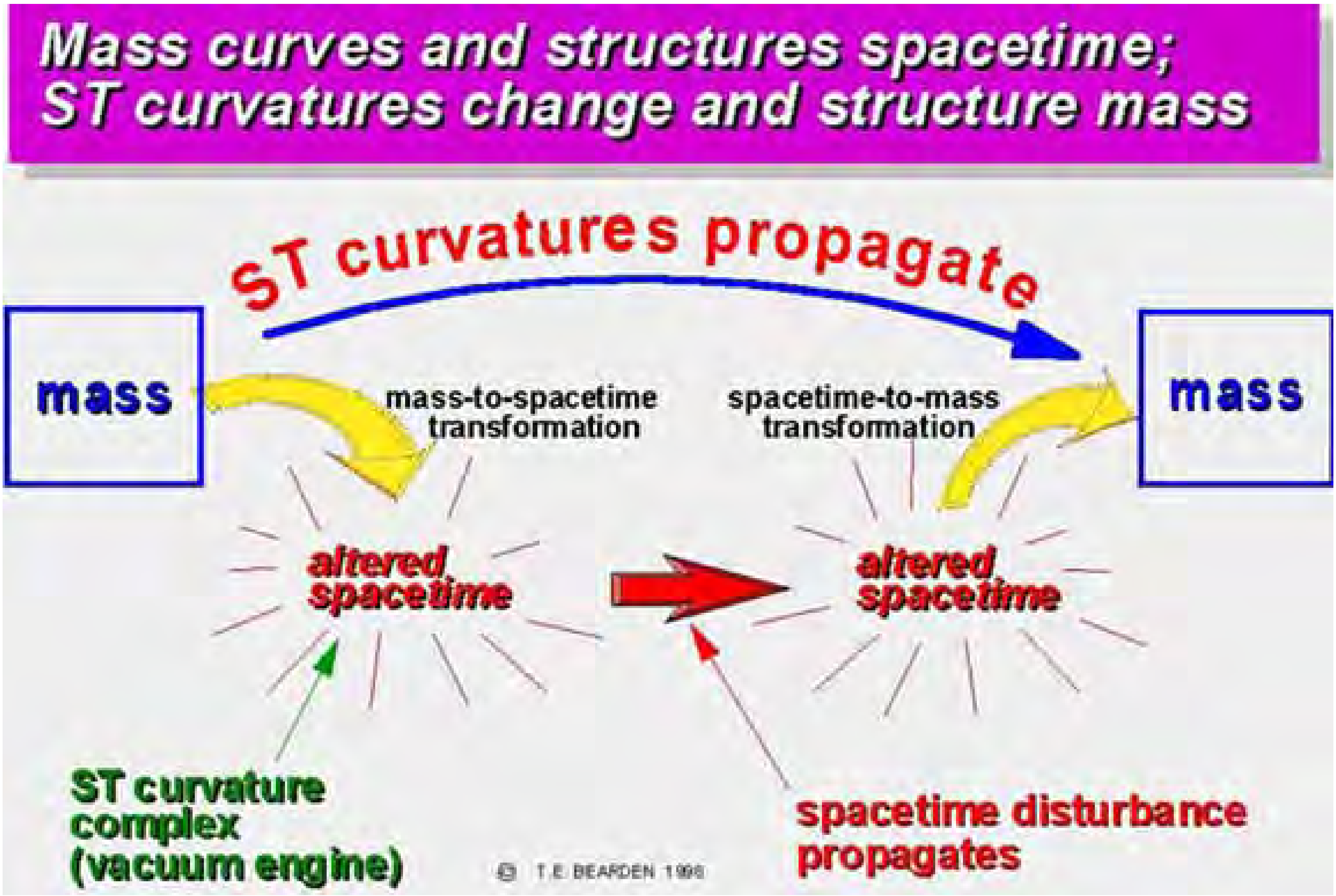

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

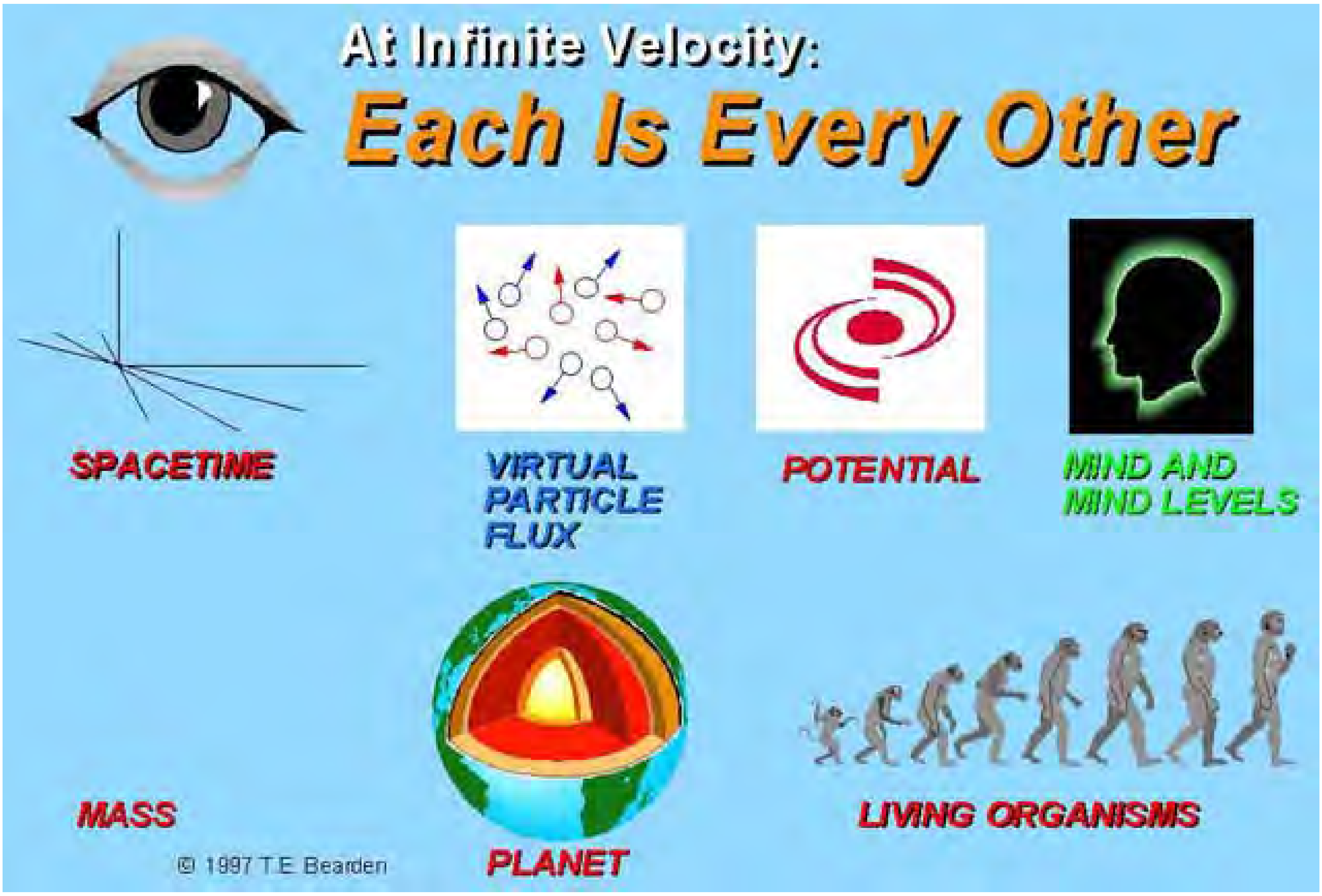
At Infinite Velocity:
Each Is Every Other
SPACETIME
VIRTUAL PARTICLE FLUX
POTENTIAL
MIND AND MIND LEVELS
MASS
PLANET
LIVING ORGANISMS
© 1997 T.E. Bearden

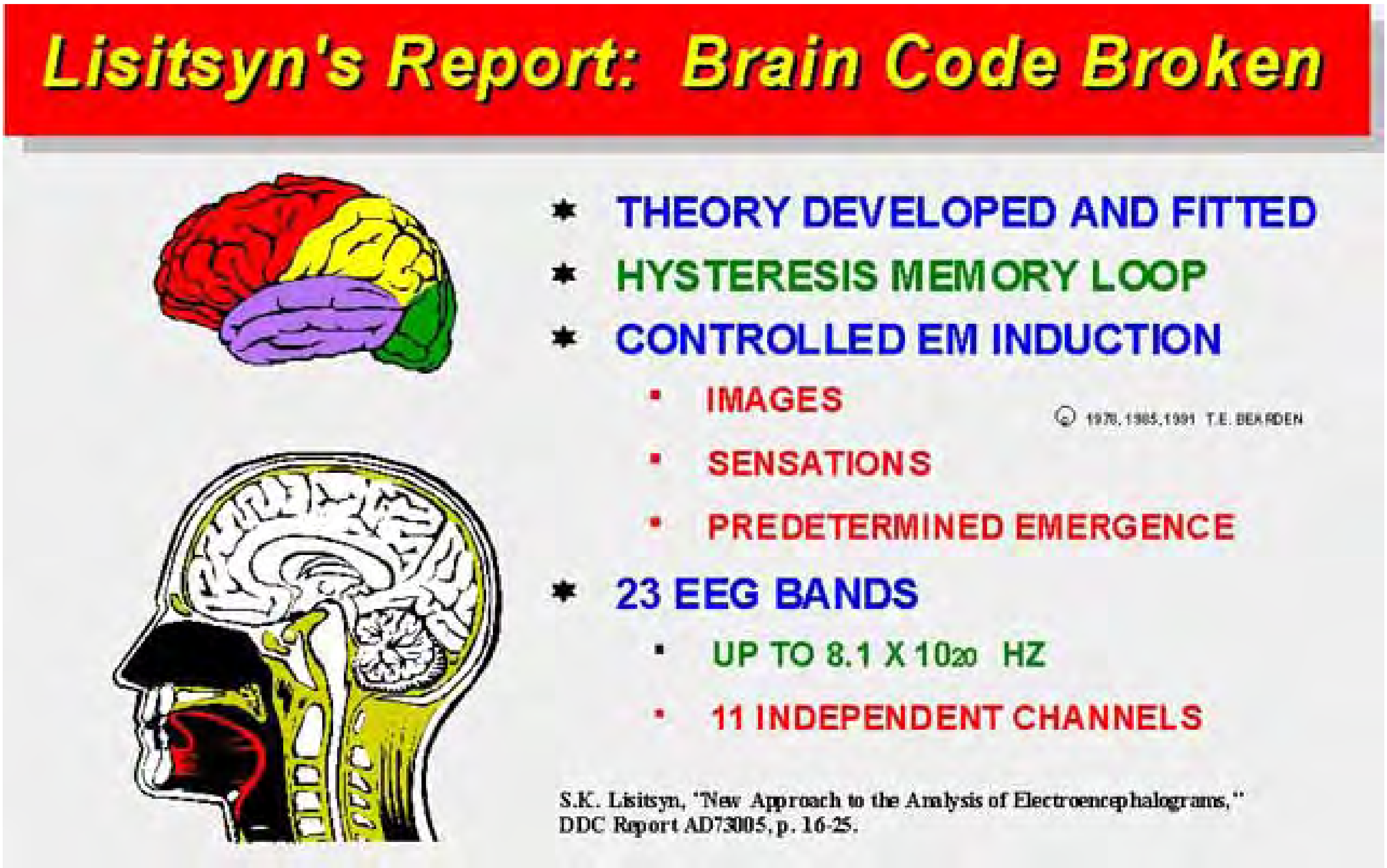

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

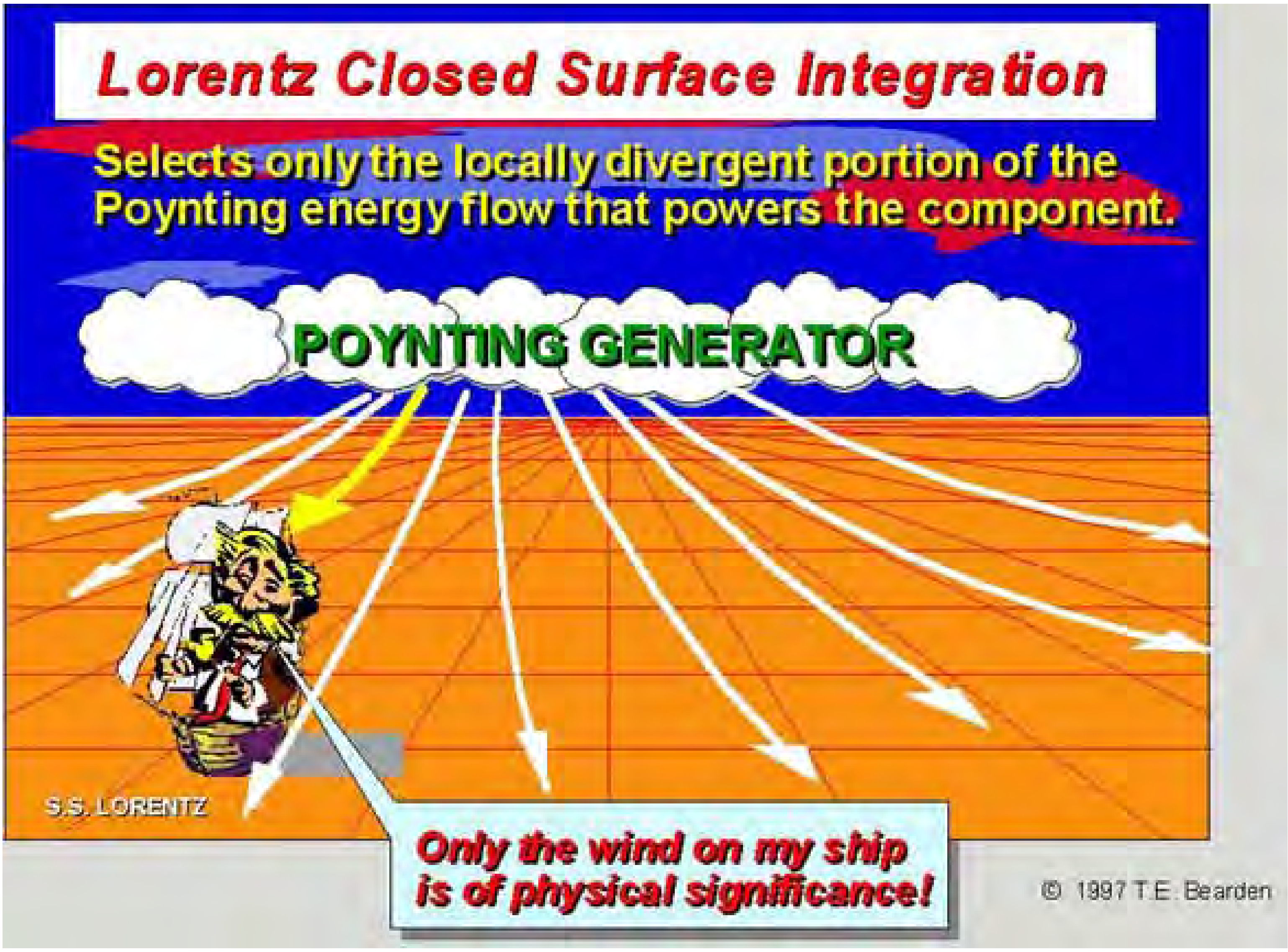

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

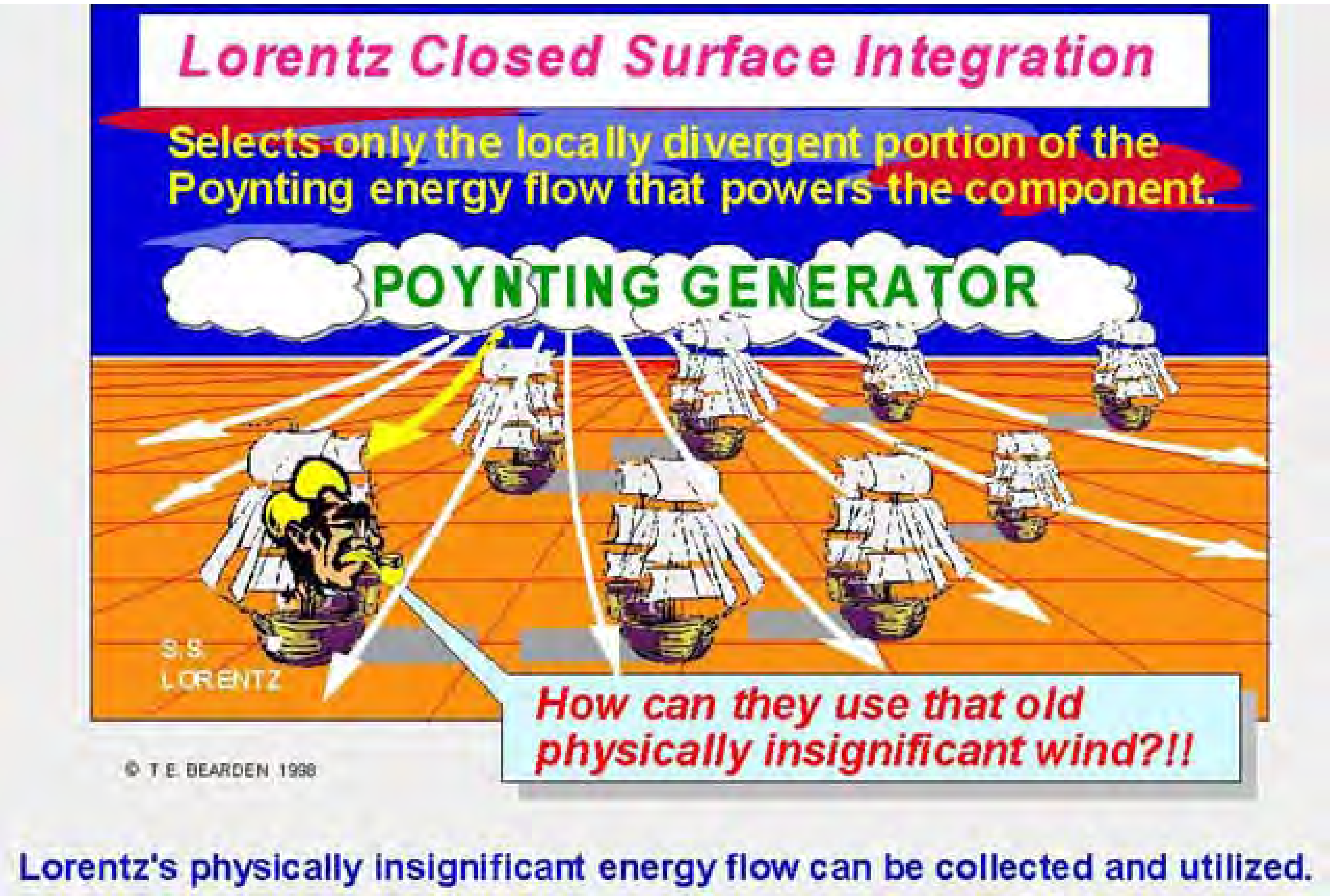

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T E BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

Lorentz surface integration of the Poynting vector around a closed surface

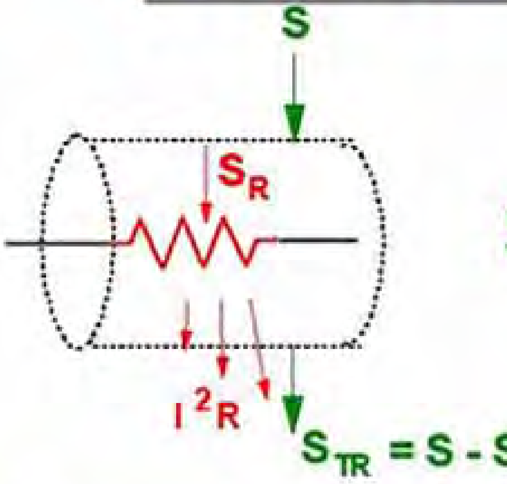

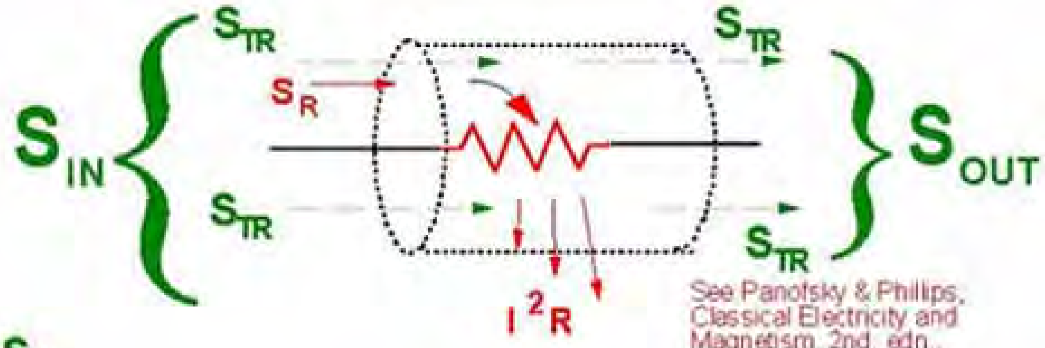

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd. edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector jϕ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

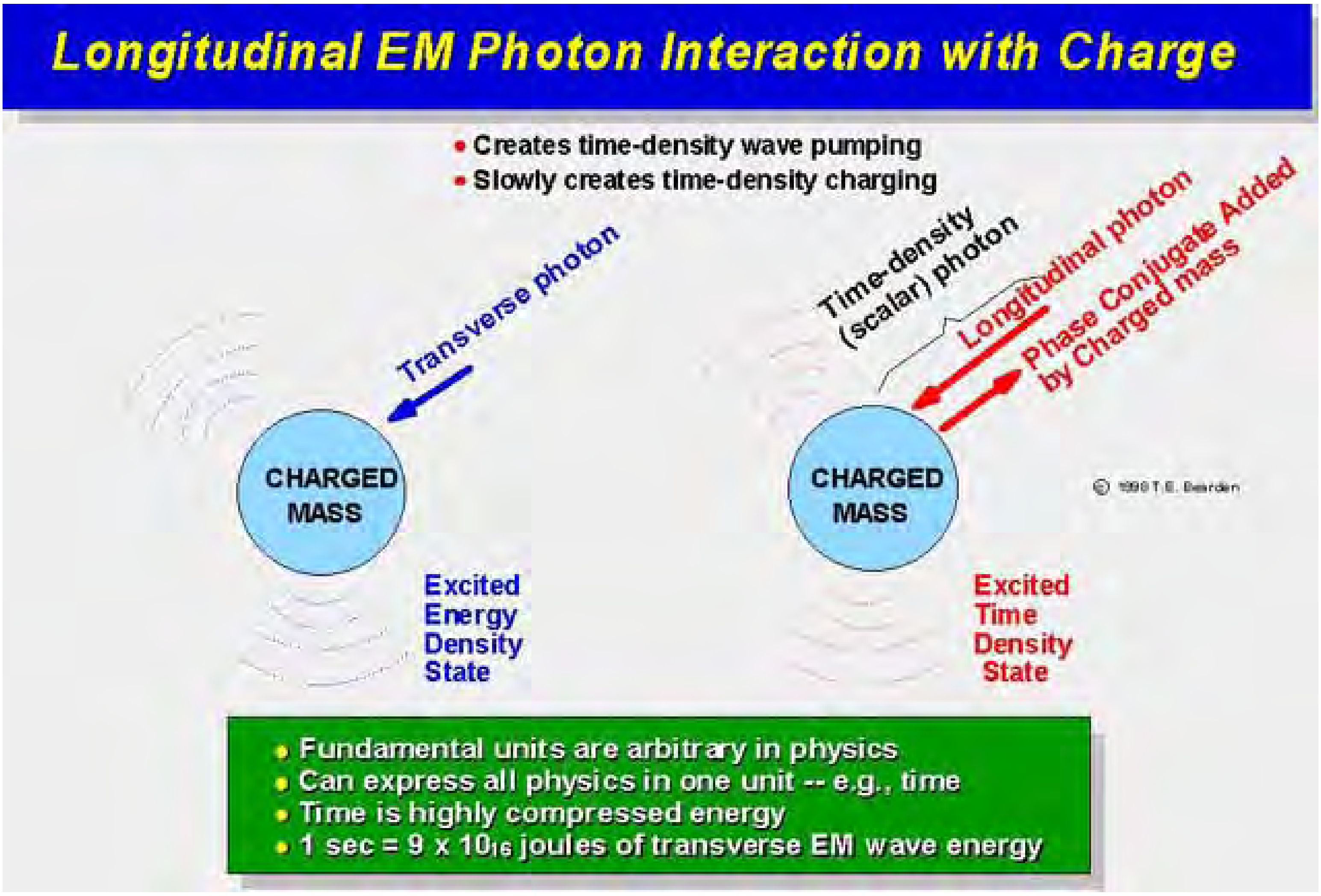

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10^{16} joules of transverse EM wave energy

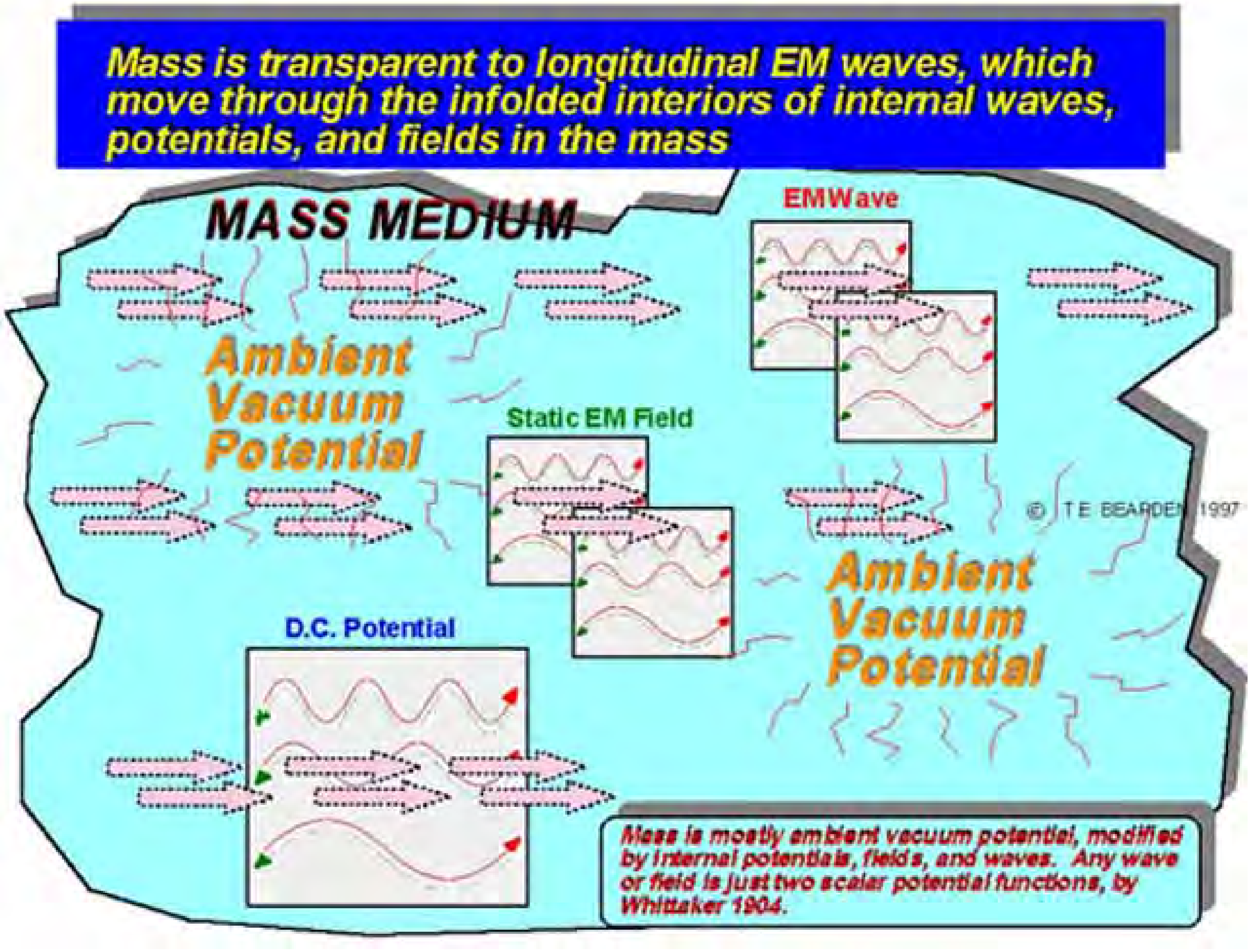

Figure 1-8. Infolded biwave composition of a scalar potential.

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
EMWave
Ambient Vacuum Potential
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

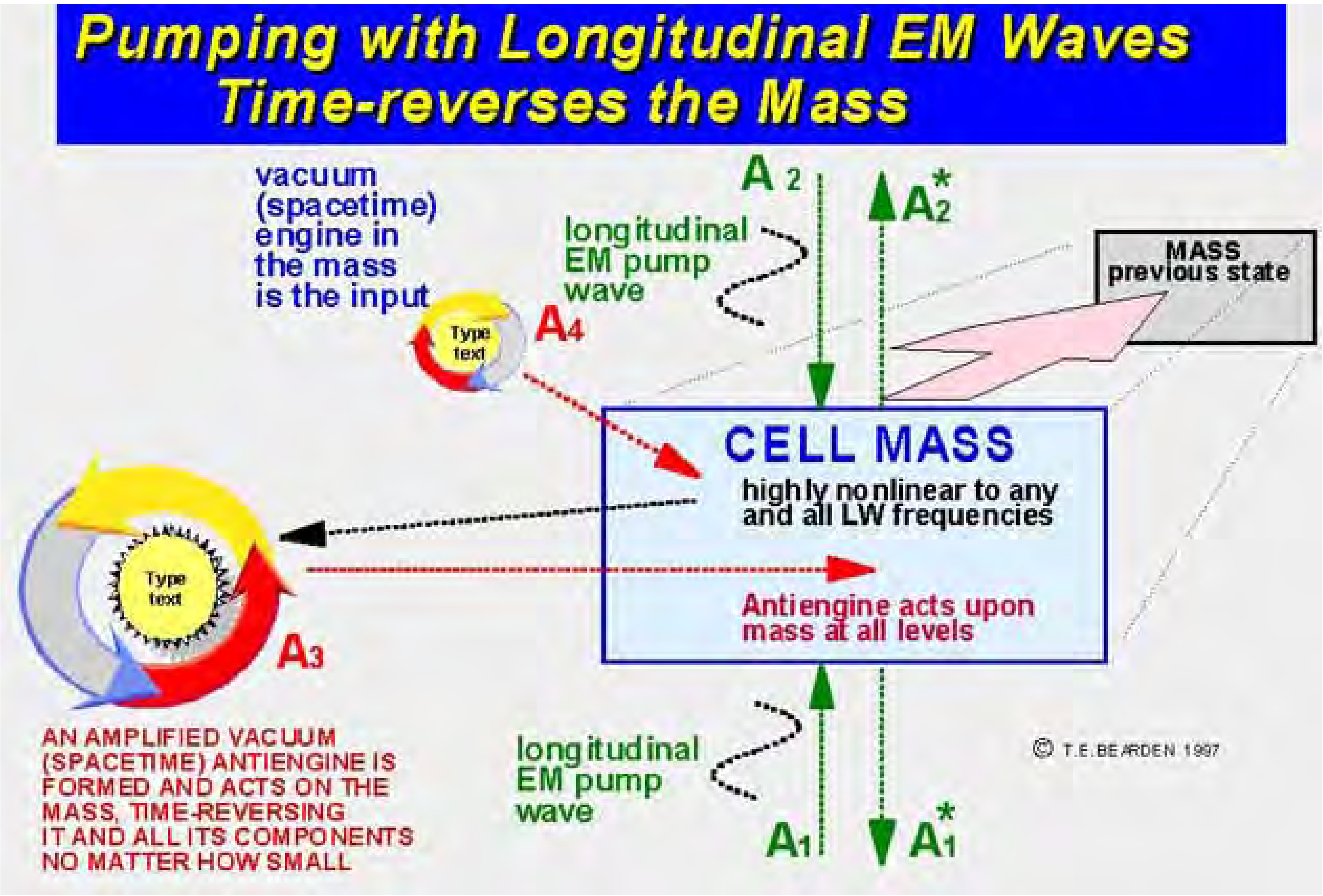

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum (spacetime) engine in the mass is the input

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

A 2
A_2^*
longitudinal EM pump wave
MASS previous state
Type text
A_4
Type text
A_3
longitudinal EM pump wave
A_1
A_1^*
© T.E.BEARDEN 1997

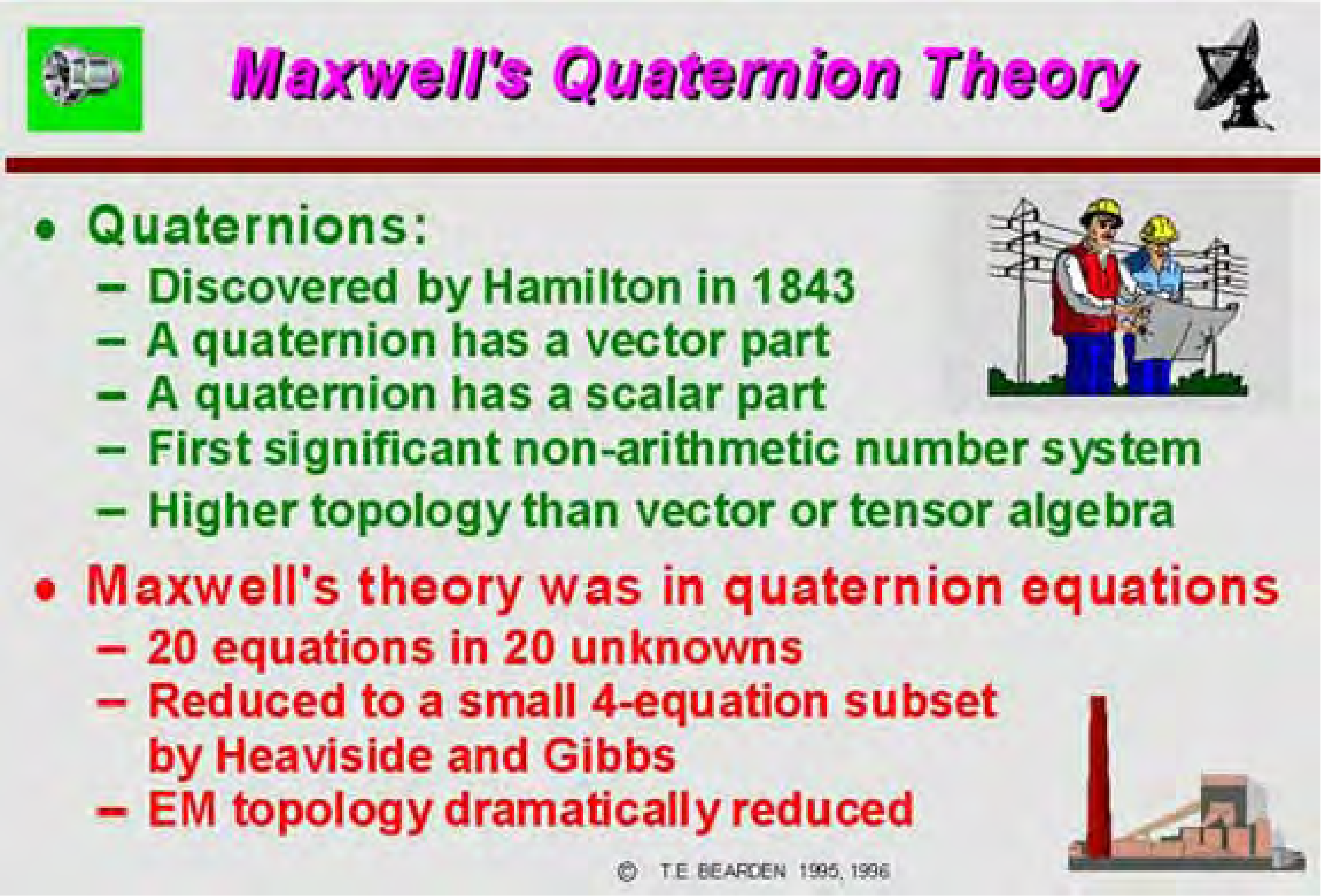

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

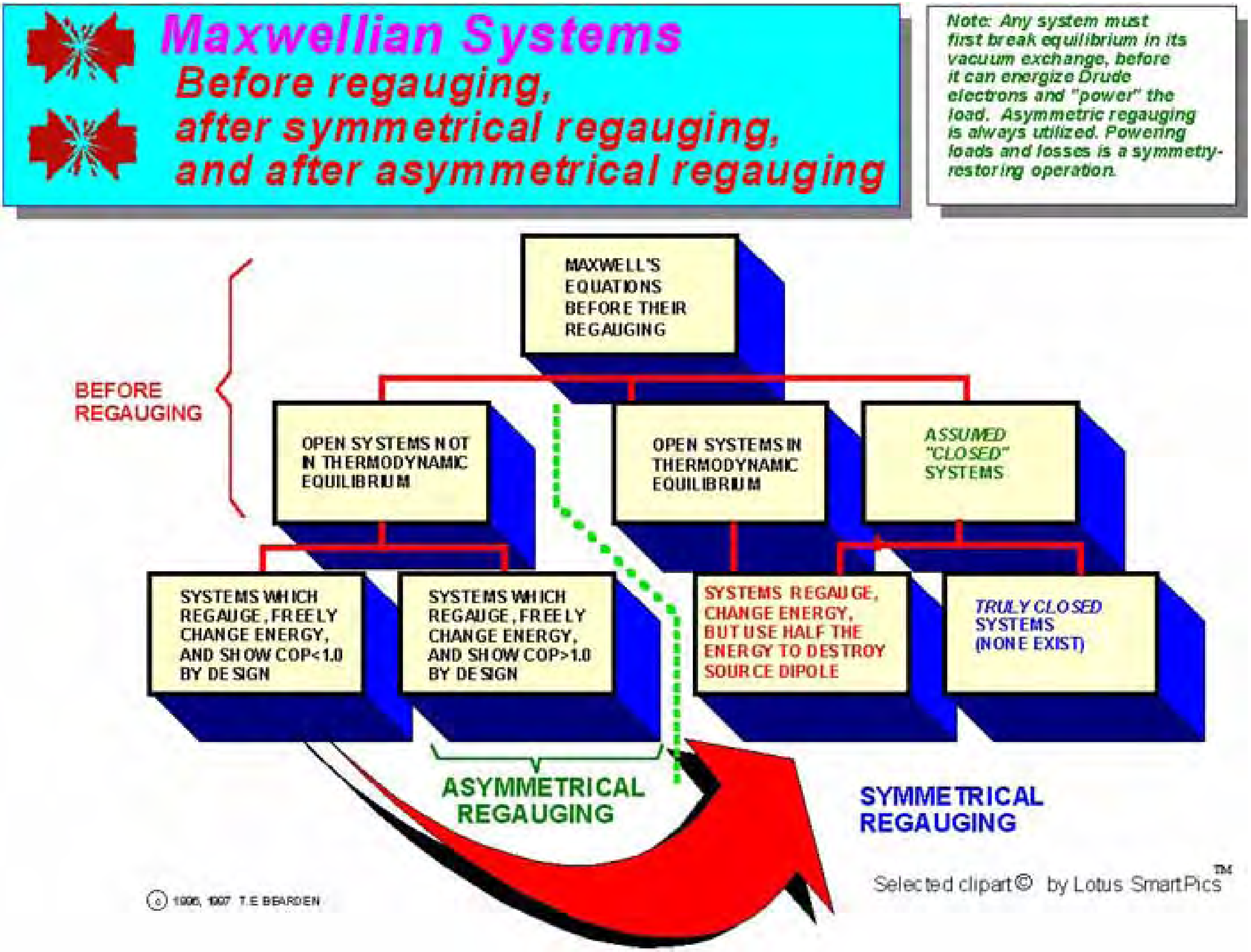

Maxwellian Systems
Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

BEFORE REGAUGING

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

© 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

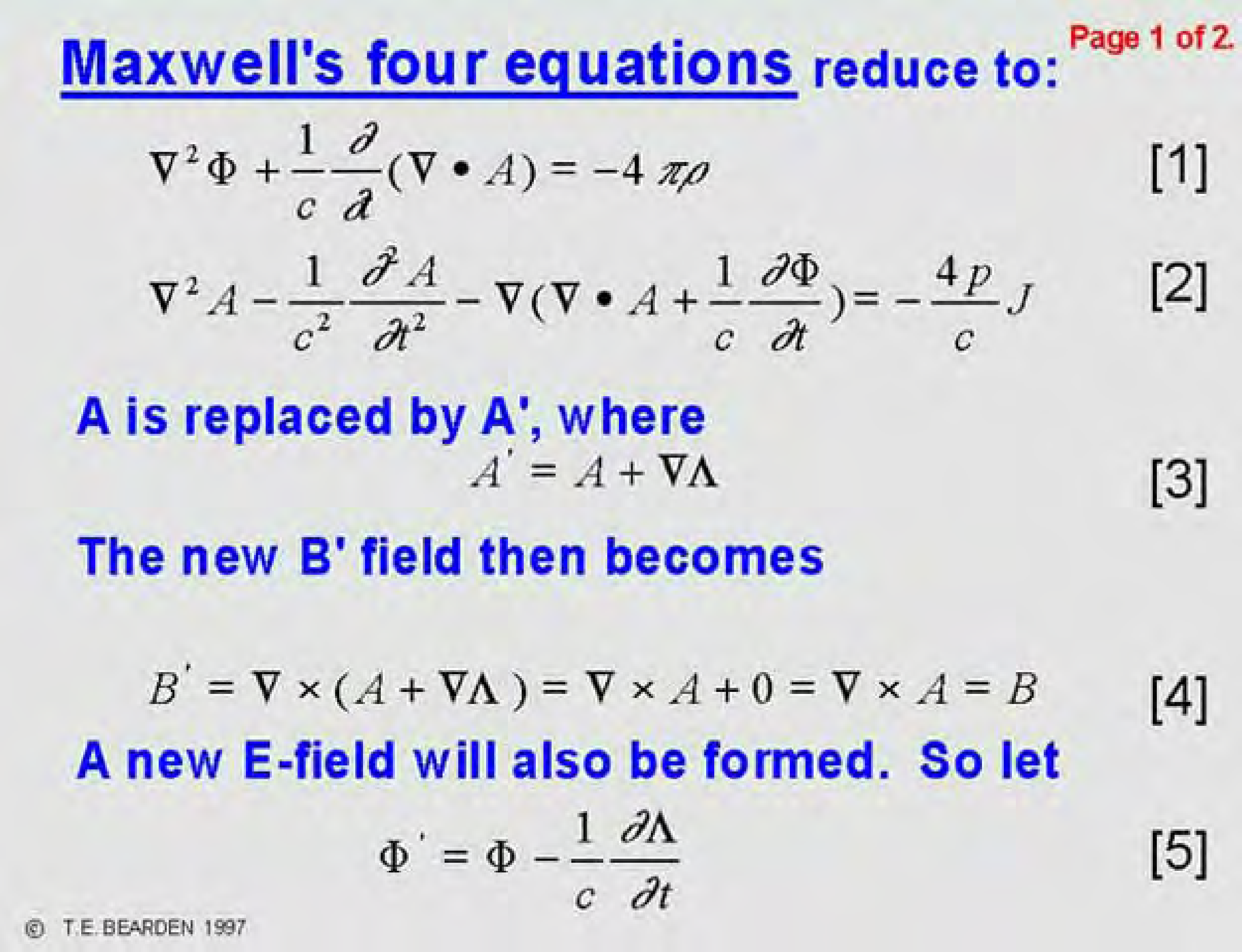

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla(\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t}) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

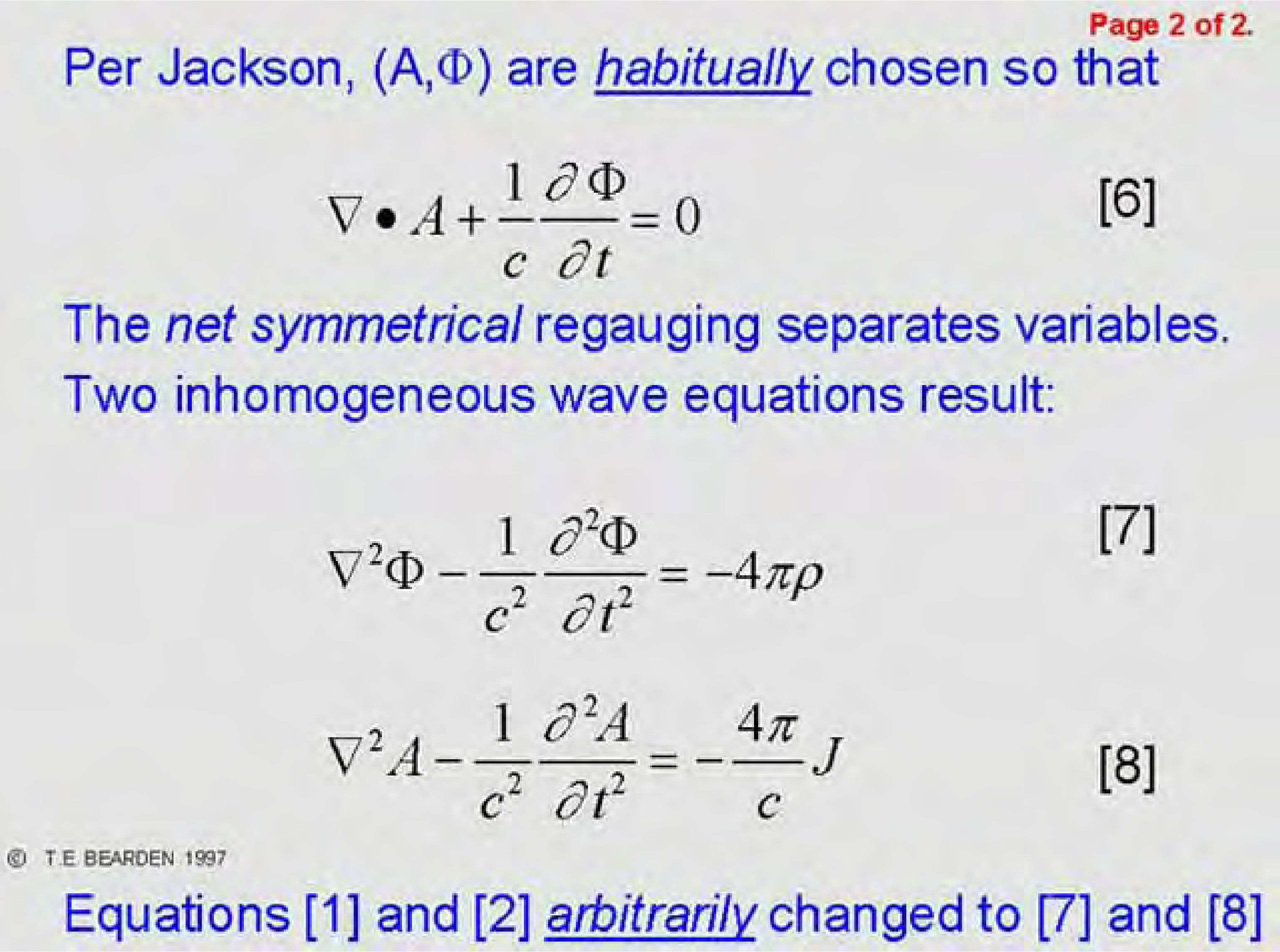

Page 2 of 2.

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.

Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c}J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

HUNTSVILLE EDITION

The Southeast's Oldest E-Zine!

CHASING THE WILD DRAGON

Foundations of a New Science

By T.E. Bearden
November 12, 1995

It is a pleasure to accept the offer of John and Larry to write a column in their new InterNet magazine. Theirs is a great venture, and with the energy and knowledge they bring to the task, one can be confident of their success in making this a very interesting and informative publication.

There is a new “rebel” science slowly being born, with most of the work being done outside the universities and orthodox scientific community. In the months ahead it will be my job to try to acquaint you with the gist of what’s going on, and what all the excitement is about.

This first article is deliberately just introductory, and its purpose is to interest you in the unusual fields we are going to be covering in more scientific detail, in future columns. We will also be listing a variety of references from the technical literature, which bear upon the various points addressed.

In this first article a little of my own background experience [note 1] is given, so you can decide whether you will be interested in what we have to say!

The real purpose of the column will just be to *present information* that you may not be able to readily obtain any other way. We will not argue or debate with skeptics; nothing productive is to be gained by that. If the information we present is useful to you, then our purpose has been accomplished. If it isn’t of use to you, then just file it in the old wastebasket. Also, from time to time I may answer a generic question, constructed from multiple queries. Individual responses to correspondence cannot be provided.

The work toward a new breakthrough science is going on in a broad spectrum of “fields,” most of which are not yet recognized by academia as even constituting legitimate areas of scientific endeavor (Table 1).
Some of the major areas involved are:

- theoretical and experimental research in overunity electrical machines,
- free energy research of other kinds, such as devices utilizing thermodynamics phase changes,
- rapidly advancing efforts in cold fusion,

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

- direct engineering of quantum change via a new kind of "hidden electrodynamic variables,"
- the direct engineering of the vacuum's virtual photon flux exchange with matter,
- action at a distance, including the deliberate formation and usage of quantum potentials,
- direct engineering of the local curvature of spacetime (yes, the direct engineering of general relativity, using the new hidden kind of electrodynamics),
- use of the new "hidden" electromagnetics to engineer body cells including the reversal of diseased cells, genetics and all, back to a previous healthy state (cellular dedifferentiation, in biophysical terms, or cellular phase conjugation, in physics terminology),
- research toward the direct engineering of the mind, including both the conscious and unconscious minds, long-term memory, and the personality itself, and
- a ruthless re-examination (and correction where necessary and possible) of the definitions, concepts, and postulates that present physics is founded upon.

Each of these "major areas" as yet has no single "experts" in the sense that individuals exist who possess a "mastery" of the given area. As an example, in spite of multiple researchers and decades of vigorous research, no one is really an "overall expert" in the overunity electrical power systems field. The "field" itself is still struggling to be born and is not recognized at all by orthodoxy; in fact, many orthodox scientists bitterly resent one even attempting to work in this area. There is as yet no cohesive single "theory" of overunity electrical or magnetic machines, although our own group is making rapid progress in that respect. But there is still no universal agreement amongst even the researchers themselves. Literally the field is still being discovered.

This "groping dimly toward a new field" should not be too surprising. After all, the early electromagnetics researchers went through the same tortuous, agonizing kind of development. They were playing with cat fur, glass rods, and pith balls for quite an extended period before understanding progressed to the point where Maxwell could mathematically tie together and model the results of the experimenters such as Faraday and others. It has taken a century and a half for modern electromagnetics and electronics to reach its own present development. It is going to take another decade or so before the overunity researchers can overturn several ill-founded parts of the present electromagnetics and get the emerging field onto a solid foundations footing and into a solid theoretical position.

In the effort to produce a theoretical model of electromagnetics, Maxwell and his predecessors performed a magnificent work. Yet they also made very serious errors (Table 2), which persist to this day. [note 2] Let us list just a few of the more blatant errors in classical EM theory today:

First, Franklin guessed wrong on which way electrical current flowed in a circuit. There is no definition at all of electrical charge in all of physics, and there is no acceptable definition of potential (that is, there is no rigorous definition of just good old "voltage.") There are no "fields of force" in the vacuum; there are just potential gradients __ because the vacuum is just one vast potential (virtual particle flux) in the first place. Force fields actually occur only in and of the matter in a material system. The primary causative agents of electromagnetics are not the so-called "force fields' at all, but are the potentials __ as is well-known in quantum mechanics. [note 3] Maxwell assumed a material ether in his equations. The Michelson-Morley experiment destroyed the material ether, but the Maxwell equations have never been changed accordingly. The electron was not yet discovered when the Maxwell equations were written; the theory does not address electron spin. Maxwell was already dead some two decades when Barus' 1898 paper [note 4] was published, pointing out the strange "backwards-traveling" wave (i.e., time-reversed wave) solution to the wave equation. In addition, Heaviside and Gibbs had already produced their vector truncation __ this truncated theory is the modern so-called "Maxwell's equations" that universities teach today __ of Maxwell's theory.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

in a higher topology (quaternion algebra), and involves some 20 quaternion equations. Heaviside and Gibbs created vector algebra, and seriously curtailed Maxwell's theory when they "translated" it into the much lower topology of vectors and some four equations. The later addition of tensors as the basis mathematics did not restore the lost topology provided by the quaternion model. [note 6]

For the EM wave, Maxwell simply assumed the transverse EM wave that had come from the much earlier derivation of the familiar wave equation from the old "plucked taut string" notions. [note 7] Yet when one closely examines and corrects some of the erroneous *assumptions* [note 8] made for the actual mathematical derivation of this "taut string" transverse wave, one finds (Figure 1) that an equal-energy second antiphased transverse "antiwave" is actually produced by the instrument and added to the "plucked string's" transverse wave disturbance of the air medium, to produce a longitudinal wave of compression and rarefaction in the air itself. The "transverse string wave" in fact never leaves the string. The string wave is not the wave that is launched into the medium.

A similar electromagnetic situation exists. In a whip antenna (Figure 2), the Drude electron gas in the conductor does exhibit a transverse "string-type" EM wave of electron precession. However, at the same time, by Newton's third law (which Maxwell neglected because at the time the atom was regarded as just a blob with no internal structure) the positively charged, phase conjugated atomic nuclei have equal energy, highly damped transverse recoil waves created in them. Further, these "nuclei-recoil" waves are phase conjugate replicas of the electron gas waves. The total disturbance that is launched into the vacuum from the whip antenna is (a) the normal transverse disturbance of the virtual photon flux of vacuum caused by the electron gas wave, and (b) the coupled equal-energy transverse antiwave caused by the nuclei-recoil via Newton's third law. The amplitude of the material antiwaves of nuclei vibration are highly damped because of the massiveness of the nuclei; however, they are of equal energy to the electron wave. But when launched into the same vacuum, the two equal-energy wave disturbances experience equal damping because they are in the same medium. Hence the wave and antiwave are (a) locked together in the vacuum, and (b) of equal energy, and (3) of equal amplitude once launched into the vacuum. What results is a vacuum EM wave of rarefaction (reduction in intensity) and compression (increase in intensity) of the virtual photon flux of vacuum __ just exactly like what is launched into the surrounding gaseous air medium from the "plucked taut string" and perturbed body of a physical instrument. The EM wave in vacuum is a longitudinal wave directly analogous to a sound wave __ just as Nikola Tesla stated. [note 9]

The vacuum may be considered a scalar potential, in which case it has an internal biwave structure (Figure 3). Interference of two potentials __ i.e., the hidden interference (Figure 4) of their internal wave structures __ in fact already creates all EM fields and waves.

When the incoming longitudinal "biwave" from the vacuum strikes the atoms of the receiving antenna, each atom may be considered a dynamic assembly of multiple dipoles, where each dipole consists of one of the electrons in the electron shells and one of the positive charges in the nucleus. The receiving dipole in the interacting atom splits the two waves apart again (Figure 5). The negatively charged end of the dipole is time-forward, and the positively charged end is time-reversed. Hence the separated normal-wave-half of the biwave interacts with the electron, to give again the familiar transverse "electron precession" wave that we detect with our "electron wiggle-detecting" instruments (Figure 6). This detected electron precession wave is what all our textbooks prescribe erroneously as the vacuum wave! The separated time-reversed-wave-half of the incoming vacuum wave is split off and interacts with the positive end of the receiving atomic dipole, in the time-reversed atomic nucleus, thus generating Newton's third law reaction force to provide the recoil of the nucleus. Present physics has ignored this second wave altogether, because of the serious error in the original derivation of the wave equation from plucked strings. Thus

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

would appear to violate Newtonian third law reaction. Phase conjugate mirrors already do just exactly this when they emit a phase conjugate replica wave. They direct the antiwave outside and away from the atom instead of inwards into the nucleus. So they do not recoil, no matter how powerful a phase conjugate replica wave they emit.

Finally, we can also show that the EM wave in vacuum is longitudinal, by yet another approach. The electrons in the electron gas in a conductor are spinning, and they are severely constrained longitudinally to be able to move only very, very slowly down the wire. So when perturbed, they act as gyroscopes, and precess at right angles to the perturbing force. The fact that the gyroelectrons in a wire move radially almost entirely, and only “slip” down the wire very slowly with a “drift” velocity, already proves conclusively that the incoming vacuum disturbance was longitudinal. The precession of a gyro is at right angles to the perturbing force. Therefore rigorously the vacuum EM waves are longitudinal, because the gyroelectrons moved transversely. The present assumption of the transverse EM wave in vacuum contradicts the entire theory of gyroscopes! [note 10] There are many other errors in classical EM that we will point out in future articles; this should give the reader a flavor of what is in store for him in future articles.

So there are no truly satisfactory “experts” in any of the 14 fields. Yet there is still something to be said for having some years of experience in one or more of the areas. I have had the good fortune or misfortune, depending upon one’s viewpoint, to be involved in various of those 14 activities for some 30 years or more, and am active in several of the areas now. My colleagues and I are particularly active in researching overunity electrical devices (see Figure 8), to include the effects of controlled chaotic oscillators, charge-blocking oscillators, and oscillators driven by in-shuttled (bridged) Poynting field energy density flow ExH. Parametric oscillation is known to have yielded overunity electrical machines and was fully reported by Russian researchers in the Russian and French technical journals in the mid 1930s. [note 11] We will further cite that little-known work in future articles. Also, we are working mightily upon true negative resistors (Figure 9), wherein the resistor scavenges some of the disorganized energy from the vacuum’s energetic exchange with the charges in the system, organizes and collects it, and then outputs that collected energy flow into the external circuit. Such a resistor acts as a power source. Kron, perhaps the greatest of the electrical machinery scientists, is known to have produced such true negative resistors.

I have also proposed a theoretical explanation of the Priore device (Figure 10) in France which in laboratory animals demonstrated remarkable cures of terminal tumors, cured trypanosomias, reversed atheriosclerosis, and restored suppressed immune systems. The Priore work was not anecdotal, but was performed under rigorous scientific protocols, working under the auspices of renowned French scientists such as Courrier and Pautrizel. The work is properly reported in numerous papers in the French medical literature. [note 12] The results were certainly achieved, but neither Priore himself nor the scientists of the day could understand the mechanism involved. The deciphering of that causative technical mechanism utilized by the Priore group required 10 years of intense effort. The results have now been published in the journal Explore! and are there for anyone to read. [note 13]
Presently my CTEC colleagues and I have filed three patent applications and a continuation, on various methods for overunity processes and apparatuses, room temperature superconductivity, Poynting generators, etc.

As an example, we took exactly the opposite approach to room temperature superconductivity as has been taken by all the orthodox researchers to date. What exactly is sought, after all, for superconductivity (SC) in a section of an electrical circuit? Well, one has some electrons on one side of that SC section that are transporting excess potential, as given by the Slepian vector J*. [note 14] What is desired is to get some electrons on the other side of the SC section that also are transporting the same amount of excess energy in the form of J*. You can do that in two ways: (1) you can flow the electron carriers through the SC section, carrying their J*, or (2) you can block the J and flow the * itself across as the Poynting flow S. [note 15] Nondivergent Poynting flow flows along an equipotential, which is

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

fanblades to a crawl) and massive correlation of the electrons and of the electron-to-lattice interactions is necessary if one is to get the electrons through there without excess collisions that shake off some of the excess * from J*, as scattered photons (heat) or as a change of form of the energy (as in straining the dielectric of a capacitor to convert electrical energy to mechanical strain energy via the piezoelectric effect.)

Our approach is exactly the opposite (Figure 11) and Figure 12). Why not just *stop* the flow of excited electron carriers on one side of the SC section, and continue the nondiverging flow of the Poynting field energy density S across that section at room temperature? Then the cryogenics is not needed at all. After all, circuits already work that way anyhow __ except standard practice is to nullify the process by letting current be driven around the sourcing current loop and back through the back EMF of the primary source. A quantum well (or several other methods) can be used to trap the “sourcing” electrons in the conductor just prior to entering the SC section. The “receiving” electrons on the other side of the SC section, however, must be in their own dq/dt-isolated current loop.

Normal sources do not furnish electrons and current to a circuit anyway. Sources just furnish Poynting flow S and EMF. [note 16] In the receiving current loop, the EMF appears automatically once the S-flow flows in and is “locked on to” by the receiving electrons. Given q**, you will immediately have J* if the electrons are free to move in the conducting circuit (see Figure 13). Further, you have eliminated all the loss terms from the standard Poynting equation for energy losses. So all the energy flow S just flows across the SC section, without any current or Cooper pairs flowing through that SC section. The SC section has become a “bridge” which (1) strips off the Poynting field energy density flow S from the electron current dq/dt on the sourcing side of the SC/bridge section, by simply reducing the dq/dt to zero; and (2) continues to pass the S-flow across the dq/dt-blocked SC/bridge section to the other side (Figure 14). The excess S-flow (and EMF) pours into the receiving dq/dt-isolated current loop, exciting the electrons therein to produce dq/dt and J*. *Any closed current loop is self-powered, once it receives S and EMF. It furnishes its own electrons; it only requires excess energy and EMF.* [note 17]

So our approach gives room temperature superconductivity in a very straightforward manner, once you discover how to block the current dq/dt in a conductor. Blocking it in an insulator is not sufficient, because that drops the potential and stops the S-flow and the equipotential * (the EMF). However, a degenerate semiconductor such as the Fogal chip can be used, as can several other processes for blocking dq/dt in a conductor. We will discuss these in a future article.

Another advantage of this approach to room temperature superconductivity is that now one can also have permissible overunity coefficient of performance. Now the load can be placed in its own S-receiving, isolated current loop. With the sourcing current loop furnishing only S and not dq/dt, the load is still powered normally in its own closed dq/dt current loop, but none of the load current is passed back through the back EMF of the primary source in the sourcing circuit.

This principle — that at least a substantial portion of the load current must not pass back through the primary source — is the primary principle required for a permissible overunity electrical machine (Figure 15). A permissible overunity electrical machine is one which produces more power in the load than you have to put into the machine to run it.

Obviously, the laws of physics and thermodynamics require such an overunity machine to be an open system that extracts and utilizes excess energy from an external source. In the case outlined, we extract the excess energy directly from the vacuum in the form of S-flow; essentially we convert our primary source to a Poynting generator rather than a “power” (energy flow loss rate) generator. The resulting9. overunity machine accomplishes both room temperature superconductivity and also overunity coefficient of performance (Figure 16). It does not violate the

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

exchange between the vacuum and each end of the dipole. [note 19] [note 20] When we attach "conducting leads" to a dipolar source, as in the simple two-wire circuit, all we are doing with the conductors is providing waveguides for the extracted S-flows from the vacuum to flow along (it flows outside the conductors) without loss. In effect, each lead is simply extending the pole (i.e., the terminal) to which it is connected.

My associates and I are also working on several other inventions at present, including two more patent applications to be filed in the near future. One of those is in the field of overunity power systems, while the other is in an entirely different, highly proprietary nuclear effort. I will address all these areas in future columns, and those columns will have real “meat” in them.

Present “big science” proclaims that overunity electrical machines are impossible. The reader should understand that the U.S. Patent and Trademarks Office is never going to grant a patent for an overunity electrical machine unless the machine is independently tested and replicated, and unless it is certified by scientists of impeccable credentials. This is as it should be. Legitimate free energy researchers accept that dictum. Eventually the stuff has to work on the bench in order to be validated. On the other hand, free energy researchers also point out that one must not demand of us the next 100 years of national research by a hundred major laboratories immediately! This is still a shoestring operation, because all conventional sources of research funds are denied to the serious overunity researcher (with the single exception of occasional personal funding of a researcher by a private individual). In other words, we demand at least the same kind of consideration that the orthodox nuclear fusion community has enjoyed. Billions of dollars have been poured into that sinkhole, and not a single controlled fusion power unit is working anywhere in the world after decades of work at enormous expense to the taxpayers.

So we overunity researchers deserve at least a decade or two of our own, to try to accomplish overunity electrical system development. We also painfully point out that “Big Science” will not fund a single overunity researcher, for specific overunity purpose, no matter what his credentials or capability. In fact, “Big Science” is part of the problem, because it fervently seeks to prevent any legitimate research in overunity. To carry out its suppression campaign. it resorts to vilification, character assassination, libel, and slander of the overunity researchers themselves __ again, regardless of their qualifications or legitimate scientific procedures. [note 21]

Indeed, orthodoxy has a black record with respect to energy. When Mayer advanced the conservation of energy law, orthodoxy hounded him unmercifully. He was universally ridiculed as the very epitome of a fool. He lost his job and suffered a nervous breakdown. Then years later toward the end of his life, scientists came to accept the conservation of energy as a most useful tool that dramatically simplified much of their analyses. So then science began to laud Mayer. There are a thousand other such examples; suffice it to say that “big science” is a bureaucracy and a “bell-shaped distribution curve” just like any other large group of people. Some scientists are near-angels. Others are near-devils. The vast majority are neither, but are just ordinary persons doing a special kind of job. In the scientific bureaucracy, however, the manipulative scientists scurry, slash, and manipulate their way toward the top. So the leadership of any big bureaucracy __ including the scientific bureaucracy __ is always rife with such near-scalawags and power-hungry individuals. Scientists are not exceptions to the same human weaknesses that we all evidence. If you really want to see jealousy, backbiting, and fierce “back-room dealing” and power-struggling, just get into a university research environment! But because the “power merchants” of any bureaucracy always fiercely resist any challenge to their position or superiority, big science has always fiercely resisted anything truly revolutionary. And today they also enlist the power of the state to enforce their dictums. We will devote one of our future columns to pointing out some glaring examples of persecution of independent researchers by big science. It is an area that all unorthodox researchers should be well-aware of. The world of science does not run by sweet reason and lofty

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

herring that dogmatic scientists use to vilify free energy researchers. It is a lie that they have also successfully sold to the U.S. Patent and Trademarks Office. Anyone with a modicum of knowledge of modern physics already knows that there does not exist a single closed system in the universe, anywhere. In particular, every physical system is continuously “open” to the violent exchange of energy with the surrounding vacuum. It’s just that most systems are in local equilibrium in these flow exchanges, except for minuscule gating accomplished for such things as the Lamb shift, etc.

By analogy it’s something like this. The free energy researcher is standing by the banks of a mighty, rushing river with breathtaking falls, turbulent areas, etc. Big Science is begrudgingly admitting that, yes, there is a river there, but it is an insane delusion to think that you can tap that energy. The free energy researcher is saying that, well, if I can build a gating sluice a little distance upstream, perhaps I can divert a small portion of the river’s flow downstream to my waterwheel, and power my mill. The orthodox scientist then vilifies the free energy researcher for even having such a heretical thought. In fact, says the orthodox researcher, any fool can see that the laws of nature force the river’s flow to stay firmly in its banks, because it is a closed system, and Moses brought that law down off the mountain with him on a special stone tablet. Then he adds all the other smug remarks such as, “If it could be done, then we at MIT and Harvard and Caltech would already have done it!” So until the orthodox scientist develops a little less dogma and a little more common sense, and learns the difference between a scientific discussion and a dogfight, the free energy researcher can hardly communicate with him.

The goal of the legitimate free energy researcher is to find a way to break the local equilibrium of this universal energy exchange, and gate a tiny bit of it up into the external circuit as excess electrical energy (Figure 17). Then he must collect the excess energy, and shuttle it from the “gating and collecting circuit” over to a dq/dt-closed current loop containing the load. He must also prevent the load current, or most of it, from passing back through the primary source against its back EMF, because if it does that, he will destroy the separation of charges in the dipolar source, killing the dipole and extinguishing the “excess energy gating.” In other words he must find a way to process and transport the energy flow across dq/dt-blocked bridges between the source’s closed dq/dt-current loop and the load’s closed dq/dt-current loop. Ruthlessly it is an energy transport S-flow problem, not a J* energy transport problem. If he uses J* in his sourcing current loop, he automatically drives the current dq/dt back through the back EMF of his source and kills it. If he does not drive dq/dt back through the source, the source will furnish S-flow for an indefinite period! Every battery and generator we have ever built is already a “free energy” machine in that it already involves broken local symmetry in the vacuum energy exchange. We have just not realized how to use our power sources purely as Poynting S-flow sources.

So we must treat an electrical system as an open system with broken local symmetry (Table 3) in the vacuum flux exchange, so that the system extracts (gates) excess energy from an external source. In this case the source is the system’s flux exchange with the vacuum. This approach is no more mysterious or bizarre than putting a waterwheel in a river, or a windmill in a wind, or a bank of solar cells out in the sunshine. The universal “free source of energy” that the overunity researcher seeks to tap is the violent exchange of virtual photons between the electrical charges of the system and the surrounding quantum mechanical vacuum (Figure 18). In quantum physics this powerful, energetic exchange of the vacuum with electrical matter is now proven both experimentally and theoretically. It is already accepted in quantum physics that the vacuum is filled with electromagnetic energy. The researcher doesn’t have to prove it anymore; he just has to find out how to properly use it. The artesian well is already there; we just have to learn how to collect and use the flowing water without dynamiting the well!

The remaining objection orthodox scientists usually raise against the notion of extracting vacuum energy has been

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

density of the exchange causing the Lamb shift is greater than the surface energy density of the sun. Different physicists have made various theoretical estimates of the energy density of this seething vacuum. A conservative ballpark figure is something on the order of 1090 grams per cubic centimeter, expressed in mass units.

It has also been my good fortune to know and work closely with several other free energy researchers. I was a colleague of Floyd (“Sparky”) Sweet (recently deceased) for about seven years; Sparky held a Master’s degree in EE from MIT and had years of experience with General Electric and other firms. He was a highly skilled researcher who invented the solid state vacuum triode amplifier (Figure 19), a unit which produced 500 watts of output and a coefficient of performance of 1,500,000. I personally tested the machine, activated the magnets on one occasion, and understand how it all worked. [note 24] The paper [note 25] I co-authored in 1991 with Floyd is still the only paper in the orthodox engineering literature that reports (a) successful overunity functioning of a solid state magnetic device, and (b) a successful antigravity experiment on the laboratory bench, reducing the weight of an object by 90%. [note 26]

We stress again that a legitimate overunity electrical machine must be an open system, receiving excess electrical energy from the surrounding vacuum. As such, it must operate in a higher topology than normal electrical machines (Figure 20).

My good friend Frank Golden has long been a close colleague. Years ago Frank developed a motor that produced overunity efficiency, but then to our astonishment we found (courtesy of Bill Tiller, [note 27] with whom we were working at the time) that its operation depended upon a 5-year-long conditioning (structuring) of the local ambient potential in order to work. Previously we had been entirely unaware that one can sometimes condition the local vacuum and the local matter in the area with a particular dynamic form. Given the local conditioning, a motor with that form will then work in that vicinity, but not necessarily elsewhere at an appreciable distance away. We now understand the mechanism for such “conditioning of the vacuum.” We also now know what to do if an overunity machine is actually a machine of that type (many of them are not). In other words, we know how to convert it to a machine that will work anywhere, anytime. We will be addressing such little-known things in future articles.

John Bedini is also a close friend and colleague; John built several experimental motors (both electrical and magnetic) in the overunity area, and performed successful trans-mutation experiments. John is a recognized genius in high-end sound amplifier development. Many audiophiles worldwide still swear that the Bedini amplifier is the best and sweetest-sounding audio amplifier ever built. One of John’s battery-powered electrical motors (Figure 21), for example, ran continuously off its battery for about five years, and kept the battery charged. When you realize that such a small electric motor is only about 35% efficient, then you realize that about 65% of the energy flowing out of the battery was being dissipated in the motor as heat, core losses, etc. So the unit was continuously performing work for that five years. [note 28] The 1/8 hp motor represented a load in which the continuous rate of work being done (the rate of energy dissipation) was about 0.08 hp. We will have more to say on this motor and its technical process in a future article.

Bill Fogal is a close colleague; Bill has patented the world’s first dq/dt-blocking semiconductor, [note 29] which partially blocks the normal current while continuing to pass the flow of voltage (Table 4). I have written a proprietary technical explanation of the Fogal semiconductor, which utilizes an extremely little-known feature called the “overpotential.” The overpotential part of the theory is experimentally proven and theoretically explained in an obscure part of electrochemistry, by a few chemists who specialize in electrode effects. J. O’M. Bockris is probably the leading world expert in overpotential theory; he authored the textbook on it. [note 30]

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

building ground, and we needed to make “floating differential ground” measurements in order to settle things once and for all. McLain and Wooten have filed for a patent, and multiple tests with excellent equipment have indeed indicated overunity. What happens next remains to be seen. I suspect that the two inventors may simply accept foreign capital investment and go offshore.

Howard Johnson is also a respected colleague, whom I very much admire. (See Figure 23). Howard has continued to work quietly and patiently upon his patented permanent magnet motor, [note 31] including patenting various magnetic gates, etc. that are necessary to make such a motor work in a rotary configuration. [note 32] Howard employs a two-particle theory of magnetism; i.e., each magnetic flux line is envisioned as having a particle traveling from the north pole to the south pole, and also a particle traveling from the south pole to the north pole. The particles are spinning; the forward-time particle spins in one direction, and the antiparticle spins in the other direction. Howard then slightly separates the two particle flows. [note 33] In other words, Johnson splits the flux lines themselves, into two different pieces. When so separated, the component lines are now curls, so their paths curve. The paths of the two “curl particles” are different; one curls in one direction and the other curls in the opposite direction. Further, a predominance of one form of curl particle gives a “time-forward” aspect, while a predominance of the other form of curl particle gives a “time-reversed” aspect. Johnson is thus able to employ a deeper kind of magnetism than the textbooks presently contain. He demonstrates that a “spin-altered” magnetic assembly exhibiting (to a compass or other such detector) a north polarity can attract another unaltered magnetic assembly exhibiting a north polarity. In short, he can make a north pole attract a north pole. We will give you further insight into Johnson’s two-particle theory in a future article. [note 34] We will also explain how and why the physicists missed that antiparticle in the magnetic field’s flux lines, and thereby failed to advance the theory of magnetism to a deeper level. Make no mistake, one day when the new theory is done, Johnson may well be awarded a Nobel Prize for his epochal discovery of a deeper structure of magnetism.

Bits and pieces of the new science approach are just now beginning to spill into a few conventional journals and symposia. For some time a few of the rebels spearheading this new science have been doing something that the University scientists should have been doing all along. The rebels have been meticulously examining the concepts, postulates, and definitions that the present scientific models are founded upon, to reveal serious foundations errors. Corrections for some of these errors have been discovered and made. [note 35] Others are still problems yet to be resolved. Yet slow progress is being made, and a “flip-over” of the old science is likely to occur within the next decade if not sooner. By then the new concepts will have become “solid” enough to allow producing an engineering mathematical model. Also the supporting experiments will have become solid enough to justify the new concepts and the model.

In this first article I wanted to introduce the 14 major areas. We will be discussing these areas in some technical detail in the future. We’ll also include current major news items of development in these areas, for the interested lay reader. I’ll particularly try to give you the benefit of extensive reference citations painfully gathered over the last 30 years. The articles will be targeted toward both the serious researcher and the interested, educated layman.

So that’s the agenda for the months ahead. We hope you find this information and the approach of direct interest and use.

Again, it’s a privilege to be aboard. And we invite you to hang on in the future; it’s going to be a wild but exciting ride. We really are going to chase this wild dragon with great vigor.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

The Virtual Times, Huntsville Edition

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

Scalar Wave Technology

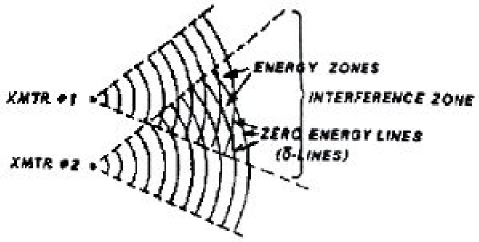

Scalar Wave Detection

Warren York
Scalar Wave Researcher

Originally published: **ExtraOrdinary Science (Volume 1, Issue 4; Oct/Nov/Dec 1989)**

WARNING: Do not attempt electrical construction on any of the equipment, coils or devices described here unless you are fully qualified. In some cases LETHAL voltages and charges are required. Also be advised that improper use, application and physical methods may cause psychological tramuma or permanent physical and neurological damage when near the vicinity of such equipment. This information is being provided as a point of interest to our readers... the publisher and author disclaim any responsibility in the use or misuse of this information.

Introduction

All of this talk so far on scalar waves and scalar technology has no purpose in the scientific community until something can be detected, observed and duplicated. The original seedlings and thoughts which led to Lt. Col. Bearden's scalar theories were developed out of observation. If not by Lt. Col. Bearden himself, then by others from many noncorrelated paths.

I already have given a small sample list of such observers from Nikola Tesla to the latest, Moray King. I have just read a newly released book by Moray entitled ***Tapping the Zero Point Energy***. I also recommend highly for others interested in scalar technology to obtain copies of Bearden's works and books from the new release of ***Excalibur Briefings*** to ***AIDS--Biological Warfare***. The two books and papers will fill in the much needed base work on understanding scalar technology and can be obtained through the High Energy Catalog.

I also understand that a new book is in the making which is a rebuttal entitled something along the lines of *Anti-Scalar*. All of this is healthy in any development process, for without pros and cons, the sorting of correct data to trash data cannot be accomplished. On the other hand, I would like to point out that any conclusions should be supported by actual tests and not just by armchair theory.

Anybody can debate from the armchair but to have any meaning to back up their statements, one must present tests where one gets his or her hands dirty. That is to say that until I had reactions and observed actual phenomena occurring, all the theories were no more than that--just theories. This is what I hope to accomplish here with these articles on scalar technology. I feel this is a true new technology and deserves some serious consideration.

If I had not gotten my hands dirty and followed through on others' experiments from other observers then I would never have been convinced that something extraordinary was taking place. Even though my senses told me I was wasting my time I set out on the monotonous task of duplicating others's work and notes of observation. This is what I hope some of you will do and start to help unfold and unleash this hidden knowledge called scalar -- zero point energy -- stress energy and even radionics. Itis true indeed that it has many names over the years, but now we can at least have hope of defining its true limits and abilities.

The detection methods I will present here are not confirmed by myself or corresponding researchers as of yet due to development is still in progress. In one method I can only accept the word of the researcher as to his results. Until I or others duplicate results it remains a possible process and approach in detection.

My own detection came in the form of two experiments which were reproductions of another author's work based on this technology. When I had reproducible results I knew there was more to scalar than just theory. The process is now underway in developing a reliable instrument or means of detection that others can use. Up to this point most of scalar observations have been uncertain and always with the person or operator being part of the circuit. Nobody could remove the person at this point to bring scalar under true observation conditions. Due to a simple observation in Mr. George Meek's work in electro-voice phenomena (EVP), where my first breakthrough was established, this now seems a possibility. The personal element was still included in the circuit but unlike radionics or Hieronymus devices, could be reproduced by anybody once special conditions were met. A co-researcher has hit upon a method which was initially observed by Nikola Tesla and shows promise of another breakthrough in removing the personal element.

I will now discuss my initial breakthrough to scalar detection which will help you follow my thinking on what is needed and just what we are looking for in scalar detection. Scalar properties are the dictating key elements in this approach. I will then describe a possible detection method of another researcher and finally what I feel will be a unique approach with an instrument currently under construction by Mr. Russell Clift, a co-researcher/inventor, and myself. Do not feel you have to adhere to these techniques for they are only a guideline and starting point to spring from in your own approach to this subject. They are not established yet and will most likely be altered many times before their completion.

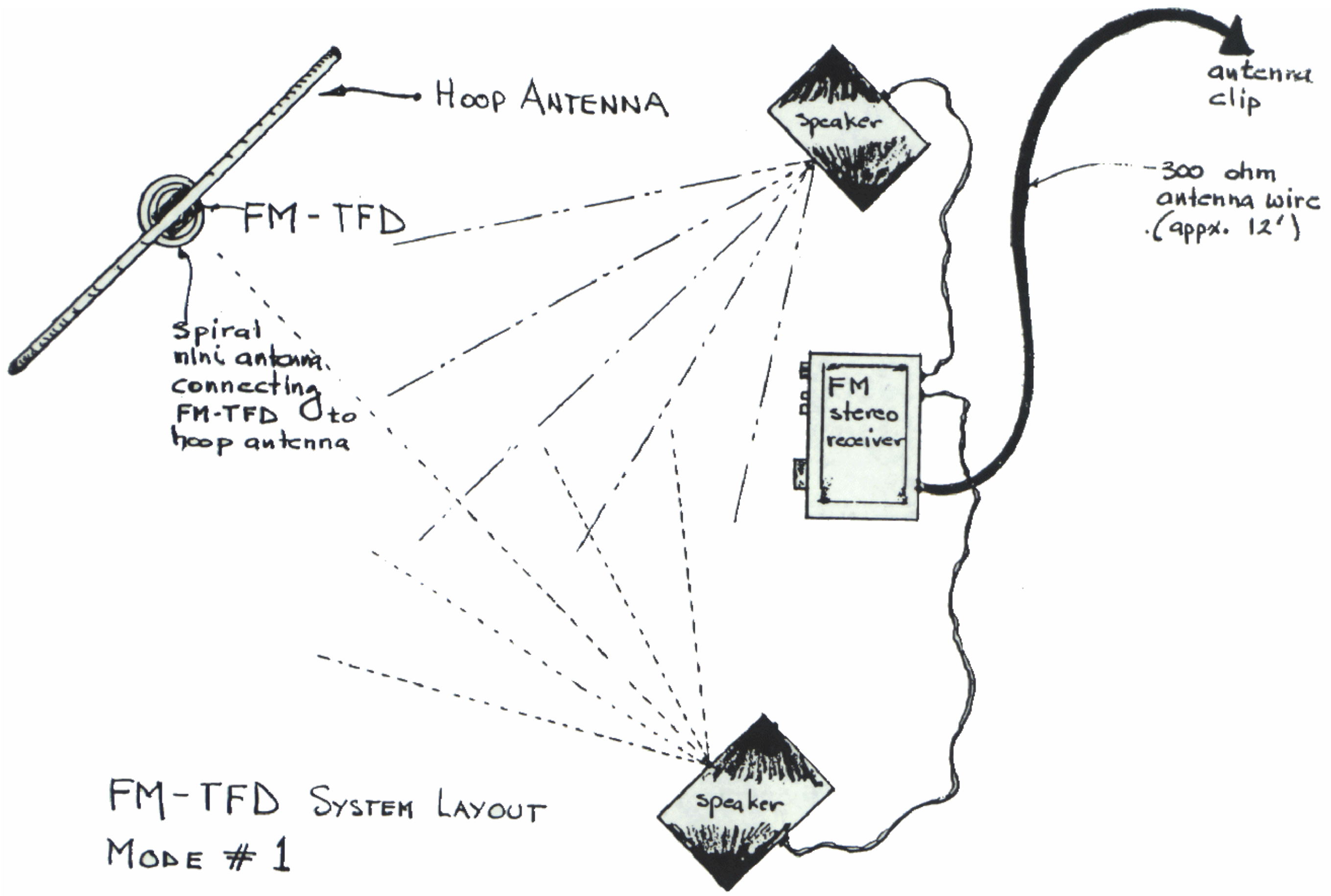

***Figure 1.* System Layout for FM-TFD [Mode 1]**

As in all unknowns there are always unforeseen dangers involved and one should always use extreme caution and behave professionally in all approaches with any experimenting. Madame Curi and the XRay research is a good example. They paid dearly for their progress in X-Ray development with their lives. Scalar and its properties can be just as deadly as it is fascinating. I hope you read the seriousness in my tone for no matter how science fictionish scalar may sound to you now, I assure you the reality of actuallity can end up staring you in the face.

Breakthrough Event

Several years ago a Mr. George Meek presented me with construction planes and notes to a device he called Mark V. [Figure 1] Mr. Meek had been researching a phenomenon better known as EVP, or electro-voice phenomena. [Figure 2]

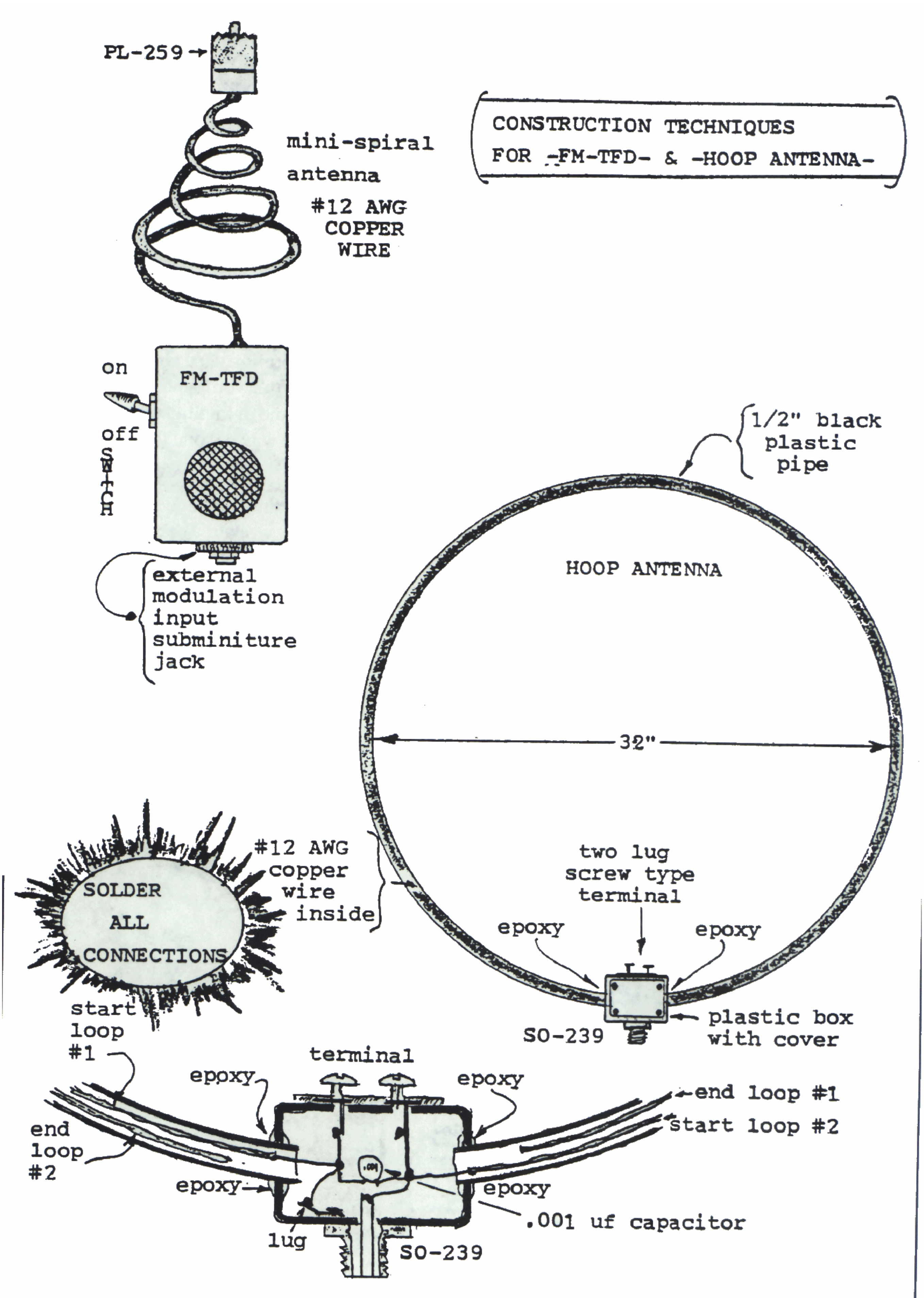
CONSTRUCTION TECHNIQUES
FOR -FM-TFD- & -HOOP ANTENNA-
PL-259
mini-spiral
antenna
#12 AWG
COPPER
WIRE
on
FM-TFD
off
SWITCH
external
modulation
input
subminiture
jack
1/2" black
plastic
pipe
HOOP ANTENNA
32"
SOLDER
ALL
CONNECTIONS
#12 AWG
copper
wire
inside
two lug
screw type
terminal
epoxy
epoxy
plastic box
with cover
SO-239
start
loop
#1
terminal
epoxy
epoxy
end loop #1
start loop #2
end
loop
#2
epoxy
epoxy
.001
.001 uf capacitor
lug
SO-239

Figure 2.

***Figure 2.* Construction Techniques for FM TFD & Hoop Antenna**

I would like to take a second here and answer a letter from one of **Extraordinary Science**'s subscribers printed in our last issue. Mr. Jeff King from New Zealand had written to the effect if perhaps I could cover what relationship scalar technology has in reference to EVP. [Figure 3] Well here it is, for this was the first reproducible effect dealing with scalar energy [interference patterns] [Figure 4] that I could verify really took place. My second breakthrough was also in the EVP area but I will not be able to go into it at this time due to available space to cover what must be covered now.

In short, Mr. Meek claimed to have a fantastic breakthrough of his own in the communications field of EVP. His claims were a recorded two-way conversation with another dimension. ***ZAP!*** Right here one would stop and feel that Mr. Meek and his researchers needed psychological help as I am sure some of you feel about me at this point. I contacted Mr. Meek and he flew down to El Paso so we could talk in person. He brought me two audio tapes of which one surely did have what appeared two-way communication in real time.

Up to this point all EVP claims were after the fact type of events. That is, the voices always appeared on blank tape after the operator had run the tape recorder for a period of time. No audible sounds were heard during the initial recording, only after playback of the tape--sort of the same thing being done now with channeling, but the person is still part of the circuit. Not all voices were understandable or clear enough to be real voice prints or just random white noise perhaps forming what almost sounds like words. In other cases, the words and even statements were definitely clear and of intelligent source. Well, Mr. Meek's tape went way beyond this to a two-way real time conversation.

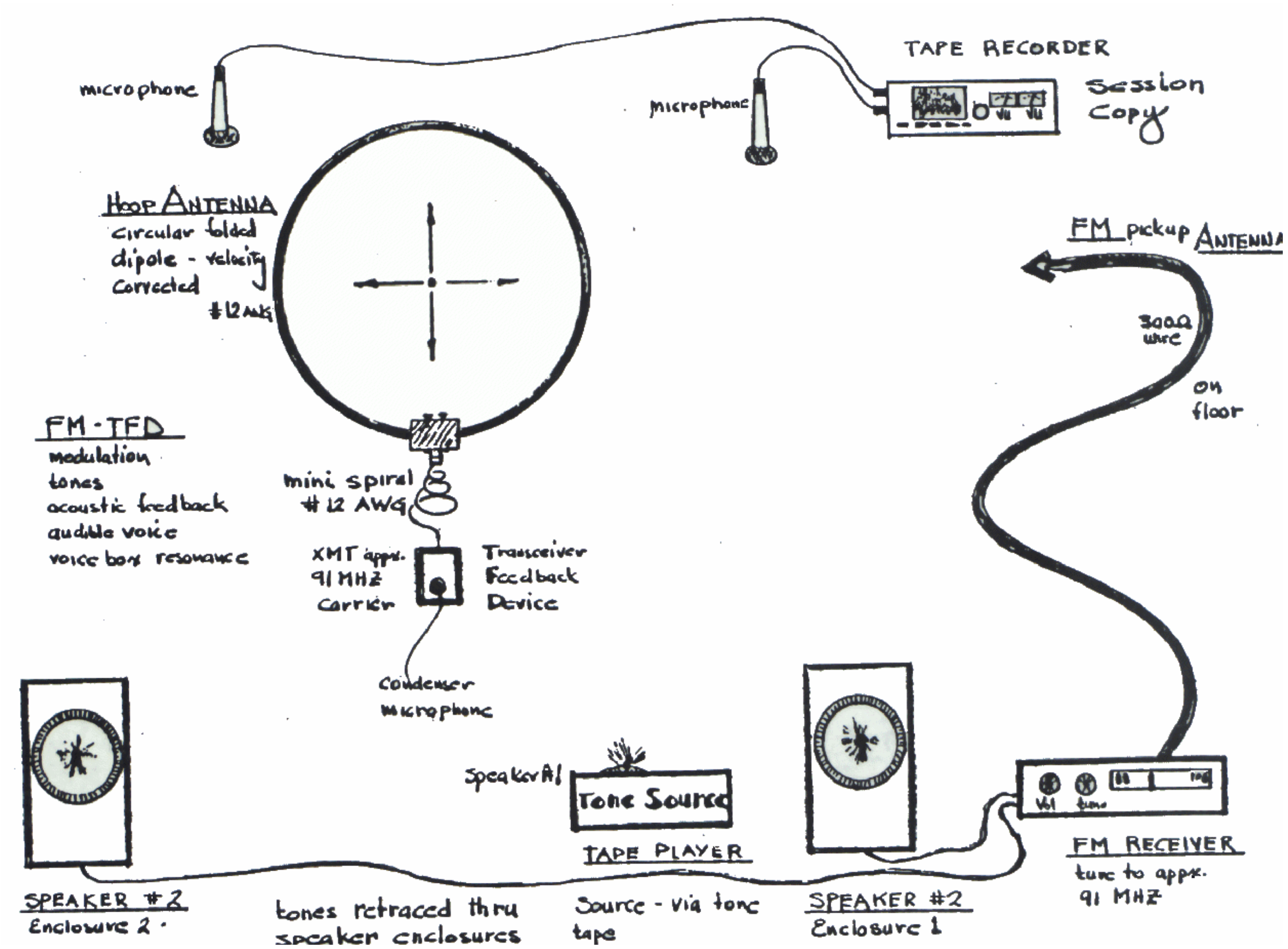
TAPE RECORDER
Session Copy
microphone
microphone
Hoop ANTENNA
circular folded dipole - velocity corrected
#12 AW
FM pickup ANTENN
300Ω wire
on floor
FM-TFD
modulation
tones
acoustic feedback
audible voice
voice box resonance
mini spiral #12 AWG
XMT appx. 91 MHZ Carrier
Transceiver Feedback Device
Condenser microphone
Speaker #1
Tone Source
TAPE PLAYER
vol
tune
FM RECEIVER
tune to appx. 91 MHZ
SPEAKER #2
Enclosure 2
tones retraced thru speaker enclosures
Source - via tone tape
SPEAKER #2
Enclosure 1

Figure 3. **Mode 2 Communications Setup**

Fascinated by these claims I wanted to know more on how such a process worked. This is where he presented me with the second tape. It consisted of thirteen mixed electronic tones. No, there is no magical number in why thirteen tones were used only that they were needed to create interference patterns. These tones with an acoustical feedback and RF transmission with a Smith coil produced a unique interference pattern [scalar-RF and acoustical stress field].

I followed the notes and set out to reproduce this fantastic claim. At one point in the notes was one observation statement:

The system seems capable of feeding back changes in emotional states.

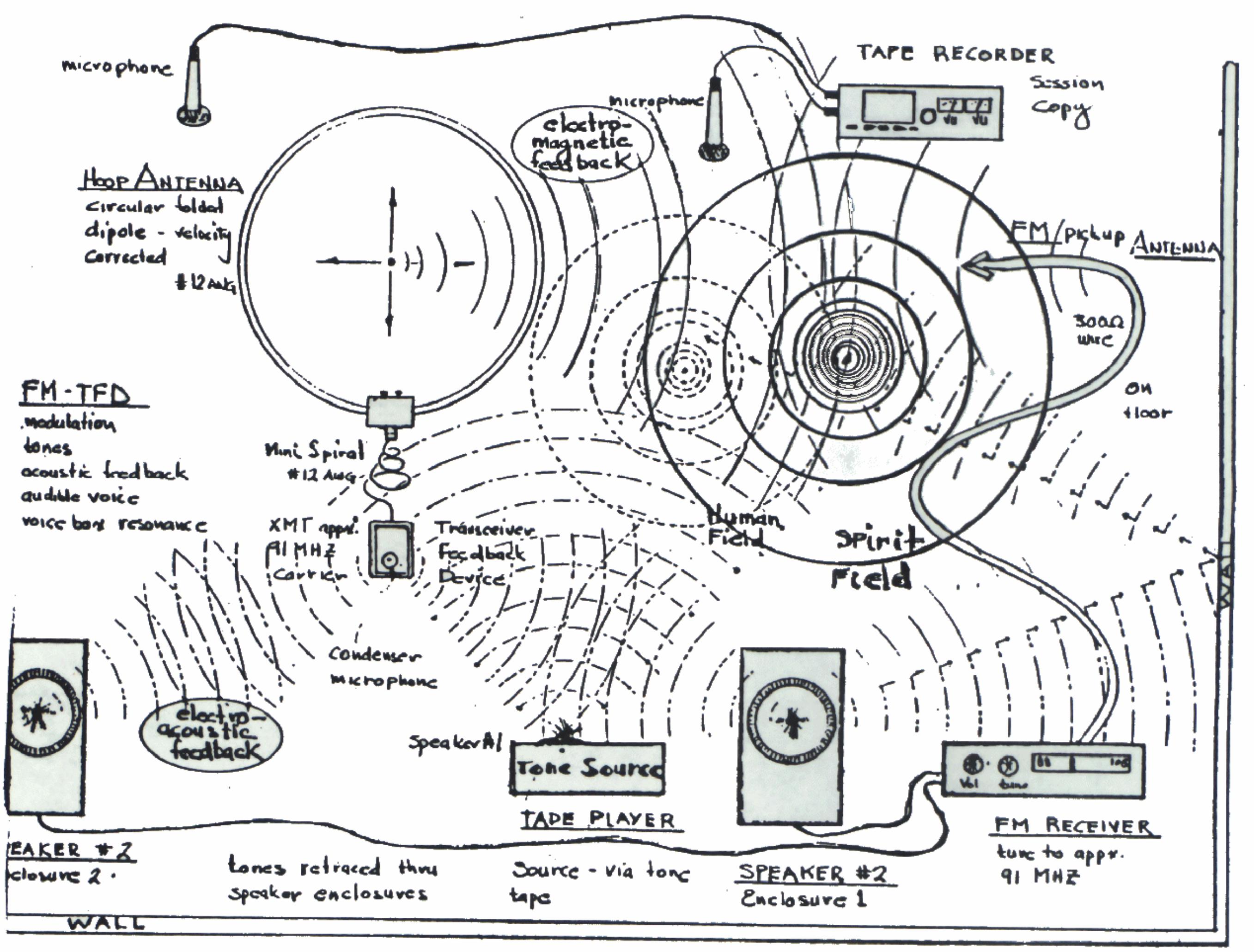
microphone
TAPE RECORDER
Session Copy
microphone
electro-magnetic feedback
HOOP ANTENNA
circular folded dipole - velocity corrected
#12 Awg
FM/pickup ANTENNA
300Ω wire
on floor
FM-TFD
.modulation tones
acoustic feedback
audible voice
voice box resonance
Mini Spiral
#12 Awg
XMT appr.
91 MHZ
carrier
Transceiver
Feedback
Device
Human Field
Spirit Field
Condenser microphone
electro-acoustic feedback
Speaker #1
Tone Source
TAPE PLAYER
EAKER #2
closure 2 ·
tones retraced thru speaker enclosures
Source - via tone tape
SPEAKER #2
Enclosure 1
Vol
tune
FM RECEIVER
tune to appr.
91 MHZ
WALL

Figure 4. **Mode 2 Communications Interference Pattern**

I stopped there and to my amazement found the observation to be true. I was able to control an avalanche effect in a balanced tone of a set of acoustical and RF frequencies with only my own concentration -- no doppler effect or movement or contacts to my body in any way. I taped this effect and it is available to those who wish to observe the level of control of an electrical circuit with shear stress energy from a human.

This observation is equal to Nikola Tesla's observation of his plasma globe and tensing a fist. He found the plasma trail would rotate in relation with the tension or muscle pressure produced. I will cover this process in a moment for the plasma approach in construction a scalar indicator seems to be ideal due to its extreme sensitivity that up until now has been the human or personal element.

Bedini Version ---- Dea/Faretto Scalar Detector

In this version of an electromagnetic scalar detector you will see that a coil is suspended in a magnetic field within a shielded case. [Figure 5] It is not indicated but I am sure the coil should be wound as a scalar coil. See "Construction of a Scalar Coil," **ExtraOrdinary Science**, Vol.I, Issue 3, page 16. The shielding is so all that known EMF fields will be run to ground and not affect the sensor within the case. As you know, one of the properties of scalar energy is that it cannot be shielded. All known EMF will be stopped from entering the case except gravity fields and scalar energy.

I am not saying that gravity fields or scalar fields are EMF, but only that the formula to each have a basic root in the EMF part of the spectrum. A generated scalar pulse or field will cut past the case shielding and affect the atoms of the scalar coil within the magnetic field which in tum will be coupled to the op amp. Any change or effect will then be displayed on the scope indicating a pulse or field of some kind has gotten past the system's shielding and to the sensor. You should see a change in the display around 6AM when the sun comes up and there is a slight gravity shift. [1,2]

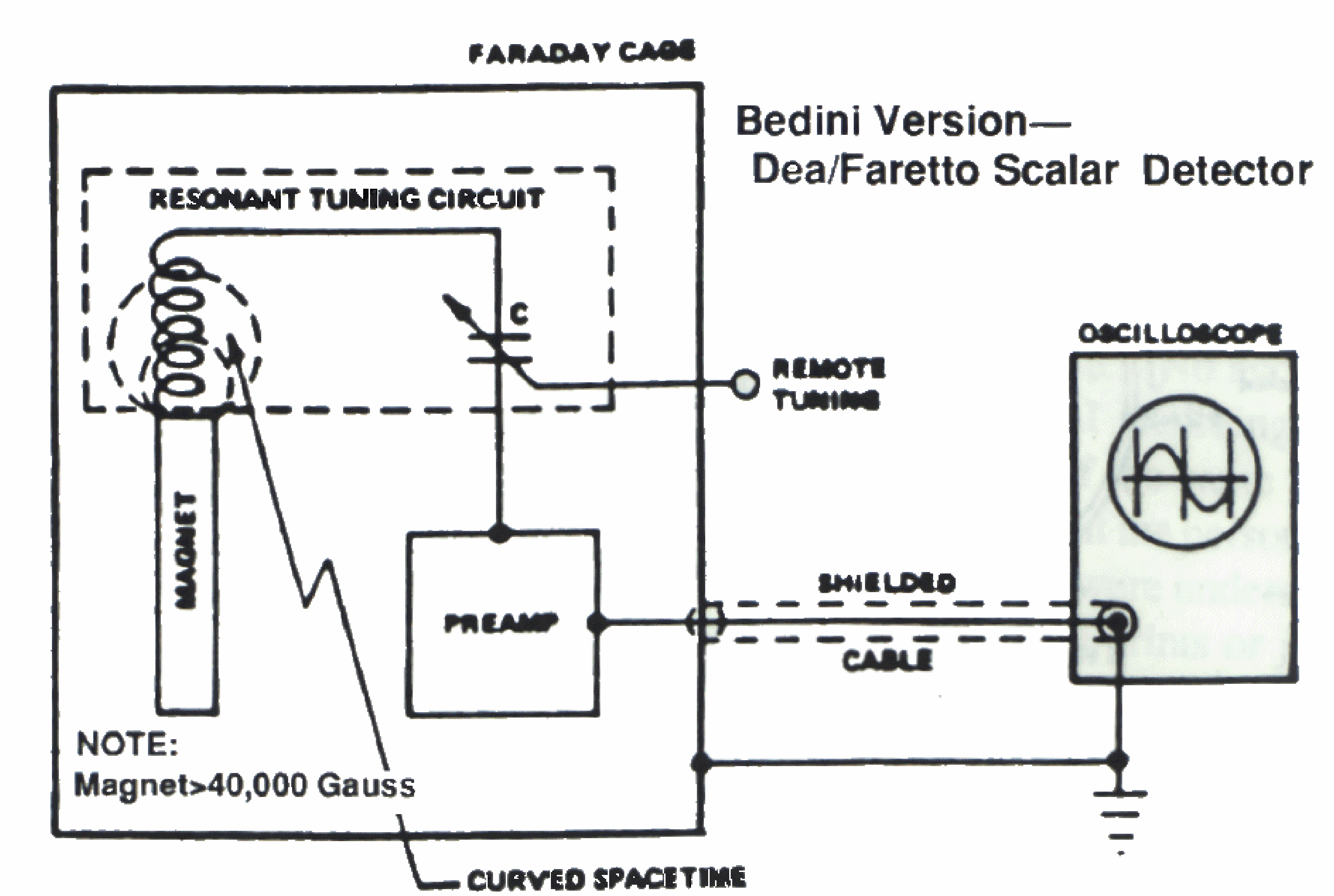
FARADAY CAGE
Bedini Version—
Dea/Faretto Scalar Detector
RESONANT TUNING CIRCUIT
C
REMOTE TUNING
OSCILLOSCOPE
MAGNET
PREAMP
SHIELDED
CABLE
NOTE:
Magnet>40,000 Gauss
CURVED SPACETIME

Figure 5. **Bedini Version of Dea/Faretto Detector**

Plasma Scalar Detector

This is the method which is under construction by Russell Clift and myself. At the time of this printing the construction is not far enough to give any details or results as to this approach. [See Figure 6] This is based on the same construction of Nikola Tesla's experiment.[3] Figure 7 is for experiments with various voltages and ion currents. Figure 8 is a simple multiplier for the AC supply shown in Figure 6.

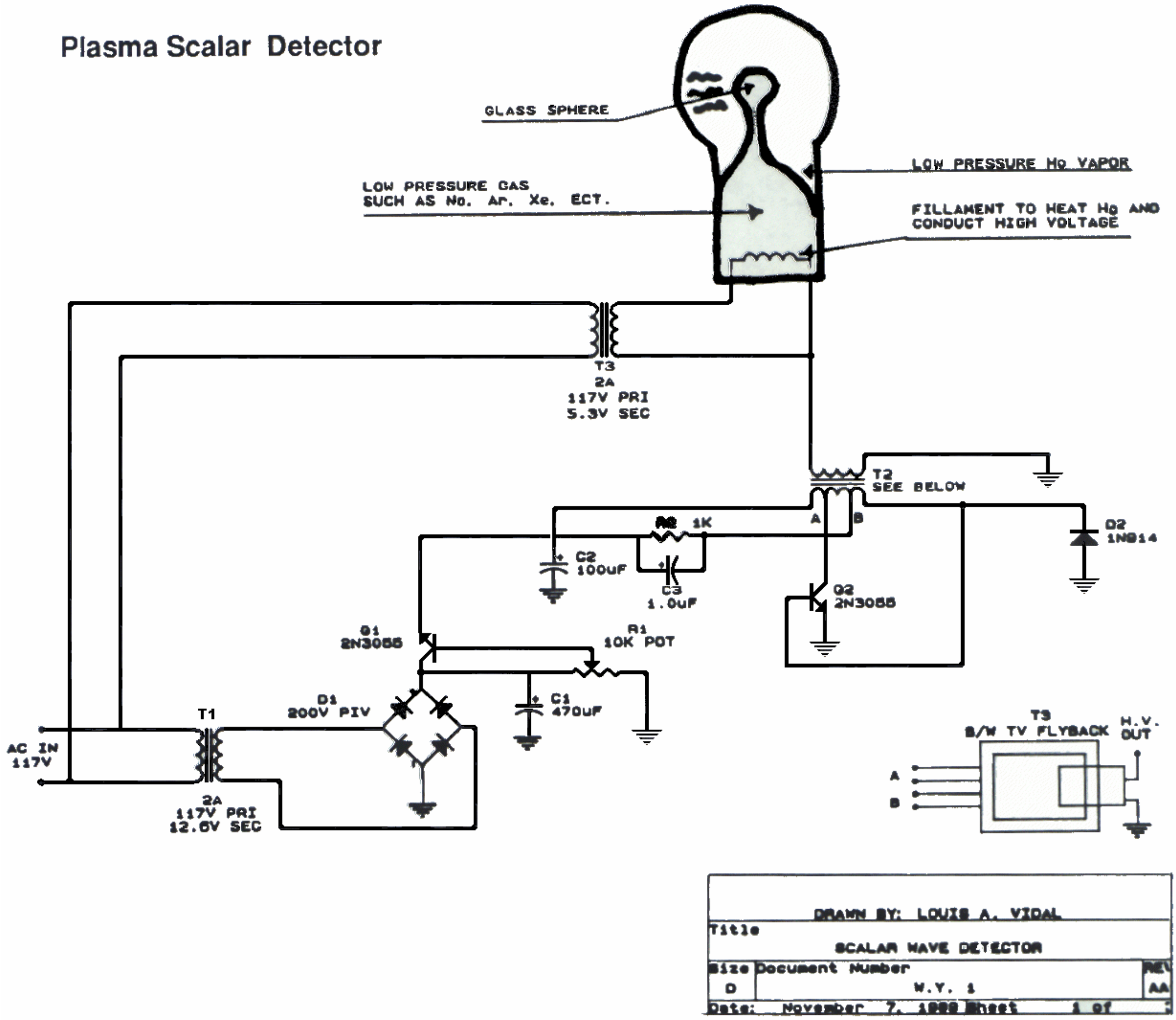
Plasma Scalar Detector
GLASS SPHERE
LOW PRESSURE HG VAPOR
LOW PRESSURE GAS SUCH AS No. Ar. Xe. ECT.
FILLAMENT TO HEAT Hg AND CONDUCT HIGH VOLTAGE
T3 2A 117V PRI 5.3V SEC
T2 SEE BELOW
A
B
R2 1K
C2 100uF
C3 1.0uF
Q2 2N3055
D2 1N914
Q1 2N3055
R1 10K POT
D1 200V PIV
C1 470uF
AC IN 117V
T1
2A 117V PRI 12.6V SEC
T3 B/W TV FLYBACK
H.V. OUT
A
B
DRAWN BY: LOUIS A. VIDAL
Title
SCALAR WAVE DETECTOR
Size D
Document Number W.Y. 1
REV AA
Date: November 7, 1999 Sheet 1 of

Figure 6. **Plasma Scalar Detector**

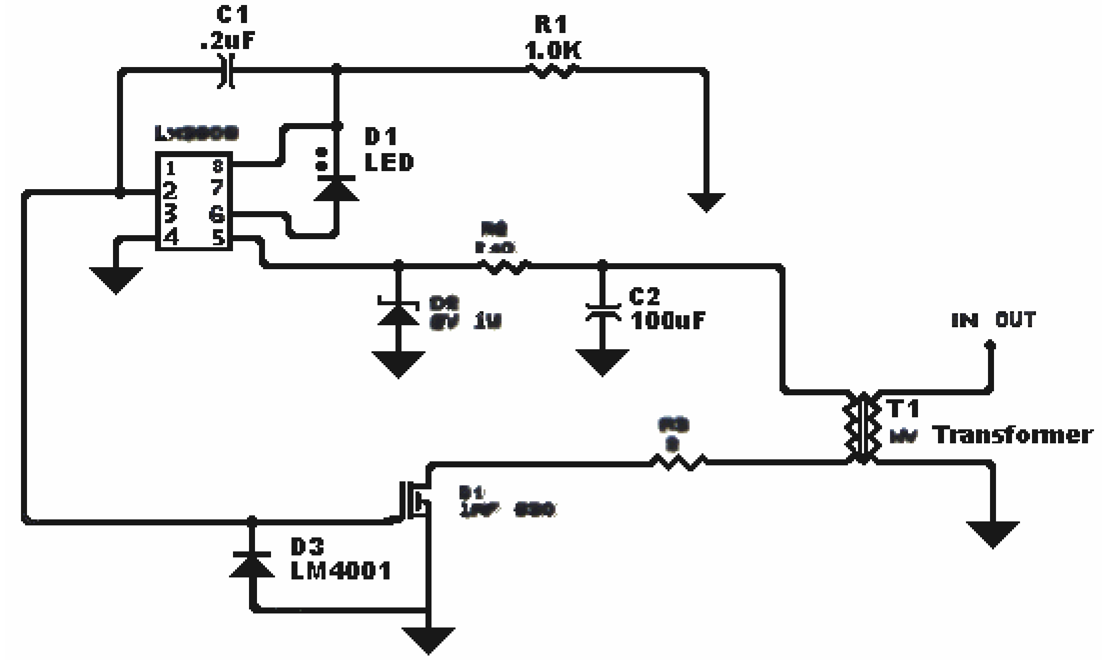
C1 .2uF
R1 1.0K
D1 LED
1
2
3
4
8
7
6
5
C2 100uF
IN OUT
T1
Transformer
D3 LM4001

Figure 7. **Power Supply**

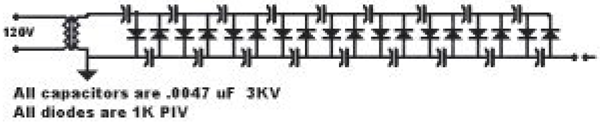

Figure 8. **Voltage Multiplier**

I know I have just touched upon this subject so far but I hope I have given some of you starting points in your own research and findings. I will welcome any and all papers or works you may do in the scalar field and submit your work for publication here for others to read and gain from. Will be looking forward to your replies in the new year. Thank you for your interest.___***WRY***

References

1. Michael Polonyi, "Electrodynamics, Inertia and Gravitation: A Unifying Approach," **Speculations in Science and Technology**, Vol. 10, No. 2, page 145, 1987.

2. Gregory Hodowanec, "Gravitational Impulses," **Electronics Experimenters Handbook**, January 1989, page 114.

Thomas Cummerford Martin, **Inventions. Researches and Writings of Nikola Tesla**, page 226 to page 228.

HUNTSVILLE EDITION

The Southeast's Oldest E-Zine!

Chasing the Wild Dragon Tables and Figures

By Tom Bearden

Figures

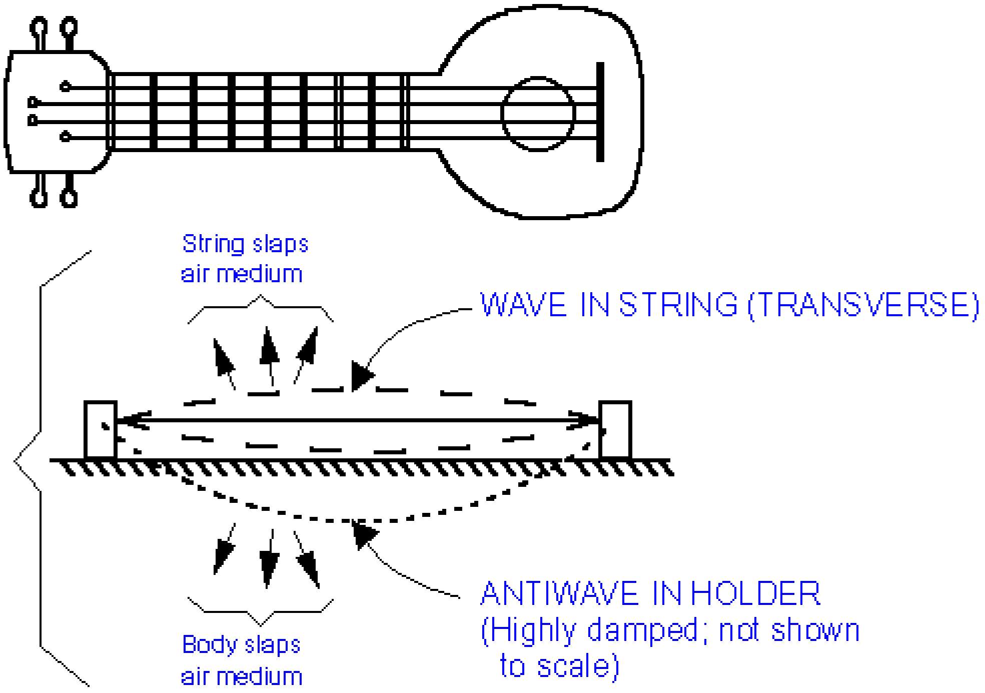
String slaps
air medium
WAVE IN STRING (TRANSVERSE)
Higher degree of freedom for air medium oscillation; wave is alternately compressive and rarefactive.
ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)
Body slaps
air medium

Figure 1. Air medium disturbances generated by plucked string.

Two waves are created, both slapping the air medium at the same time, but in opposite directions. A disturbed medium oscillates according to its own degrees of freedom. The air has more degrees of freedom than the string. A wave of compression and rarefaction is created in the air medium.

D-1-F 1

© 1993 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

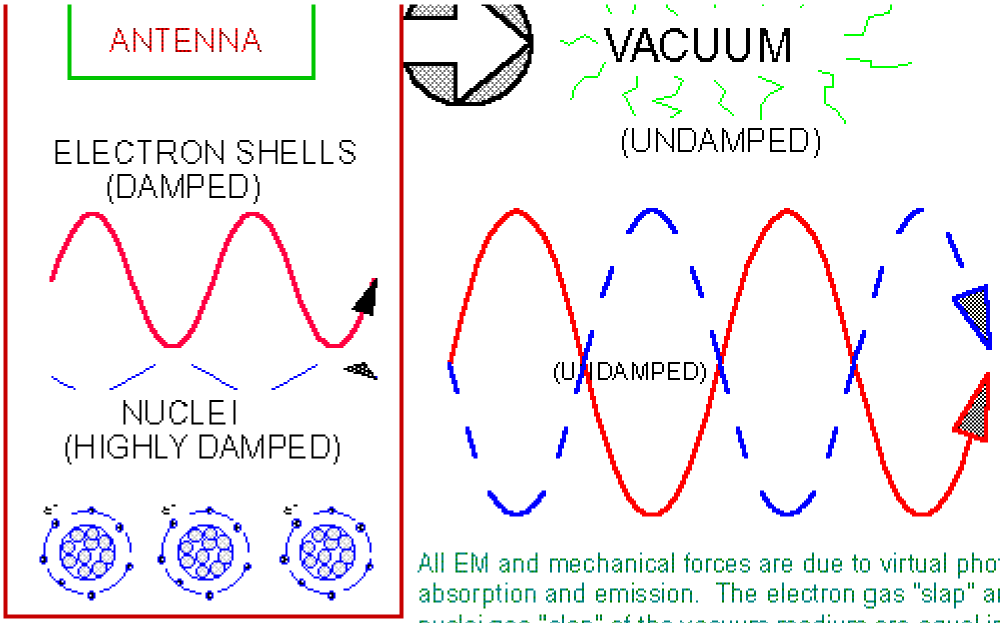
ANTENNA
ELECTRON SHELLS
(DAMPED)
NUCLEI
(HIGHLY DAMPED)
VACUUM
(UNDAMPED)
(UNDAMPED)

All EM and mechanical forces are due to virtual photon absorption and emission. The electron gas "slap" and the nuclei gas "slap" of the vacuum medium are equal in energy. Hence they inject equal numbers of virtual photons into the vacuum, creating equal and opposite vacuum disturbances. This gives a vacuum wave of alternating compression and rarefaction. This vacuum EM wave is similar to a sound wave, which is modeled longitudinally.

Figure 2. Launching an EM wave from a wire antenna.

D-1-F2

© T.E. BEARDEN 1994, 1995

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

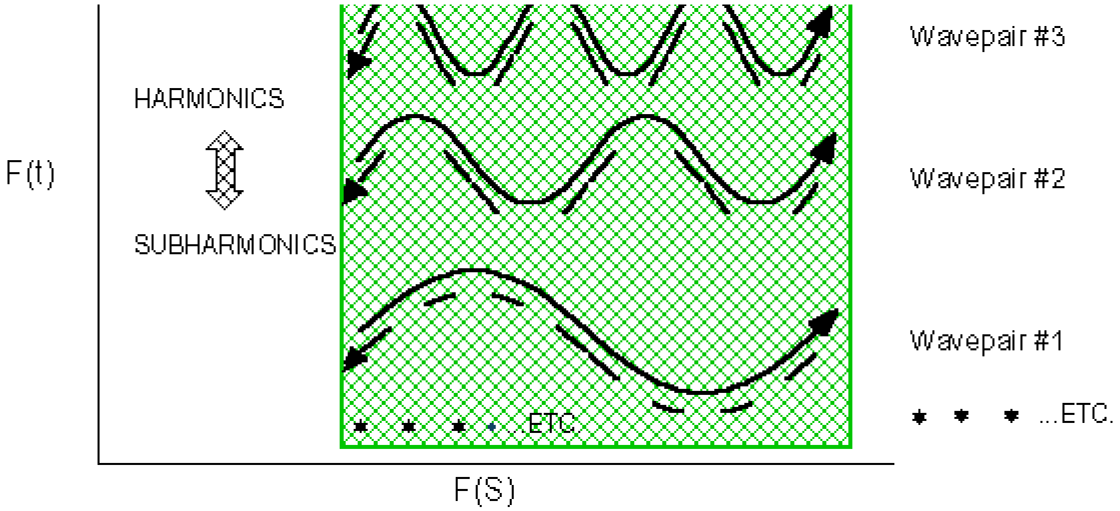

Figure 3. Internal Wave Structure of the scalar potential.

A harmonic set if wavepairs, where in each wavepair the two waves superpose spatially, but travel in opposite directions. The two waves in each pair are true phase conjugates and time-reversals of each other. Thus they comprise a coupled wave and antiwave. The photons must also be coupled into photon/antiphoton pairs, by a strong application of the distortion correction theorem of nonlinear optics. *Each wavepair is a standing electrogravitational wave.*

D-1-F3

© T.E. Bearden 1984, 1985

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

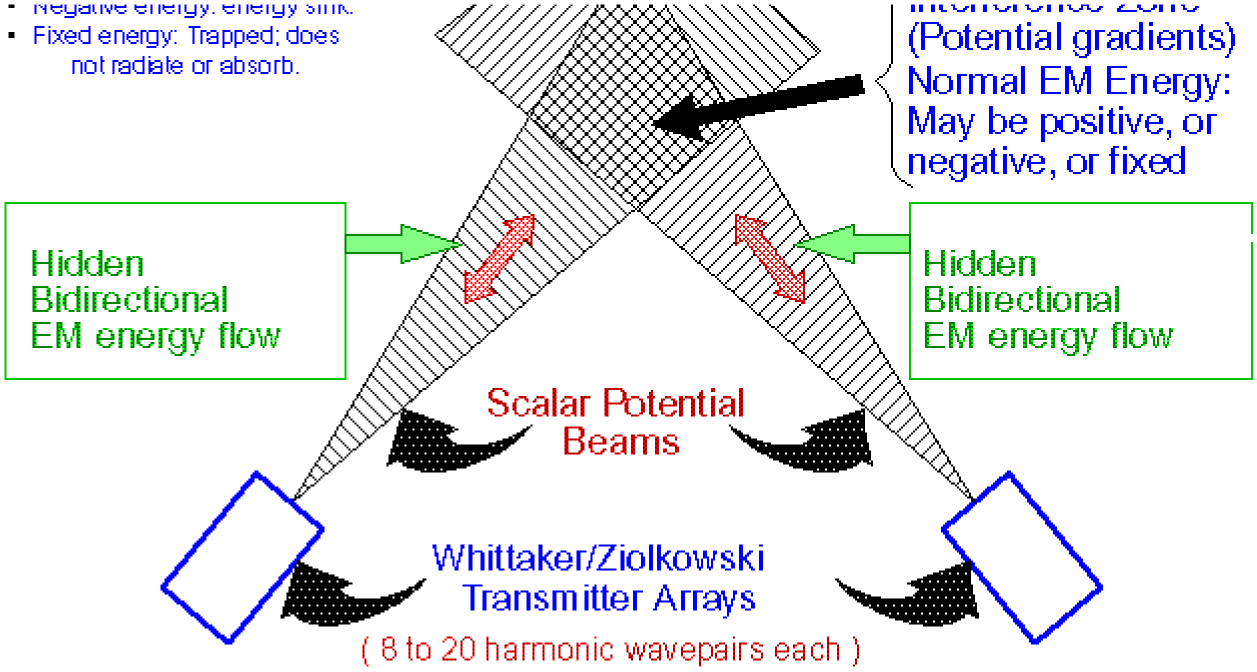
- Negative energy: energy sink.
- Fixed energy: Trapped; does not radiate or absorb.
Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

Figure 4. Scalar potential interferometry provides action-at-a-distance. It already creates normal EM fields and waves.

D-1-F4

© 1994 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

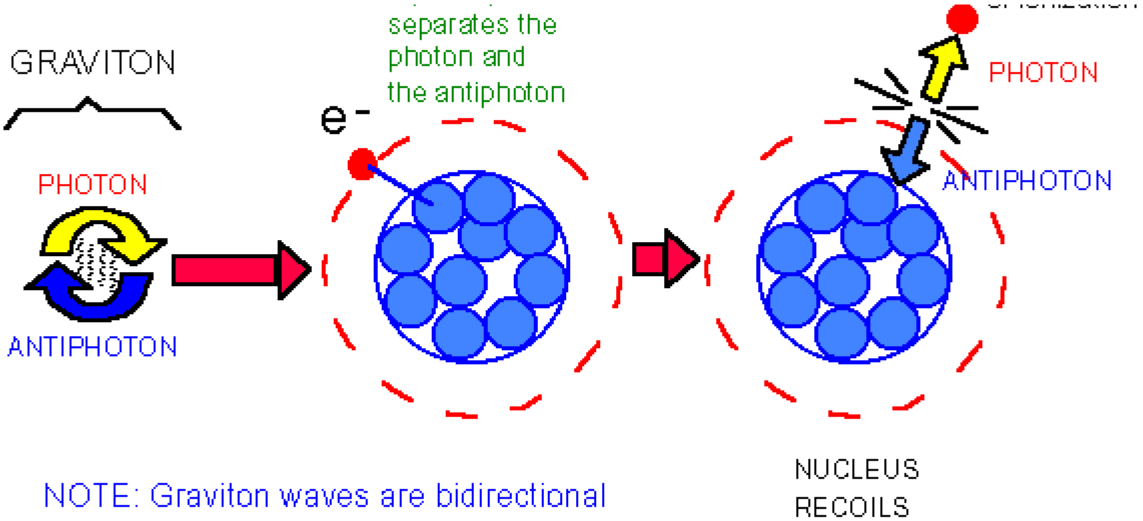
GRAVITON
PHOTON
ANTIPHOTON
separates the
photon and
the antiphoton
e -
PHOTON
ANTIPHOTON
NUCLEUS
RECOILS

NOTE: Graviton waves are bidirectional phase waves of photon/antiphoton coupling and uncoupling in the vacuum.

The atom is an assembly of dynamic dipoles.

A dipole splits the graviton into a photon and an antiphoton.
The photon interacts with the electron, producing the action.
The antiphoton interacts with the nucleus, producing the reaction.

Figure 5. Graviton interaction with matter produces the photon interaction and Newton's third law reaction.

D-1-F5

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

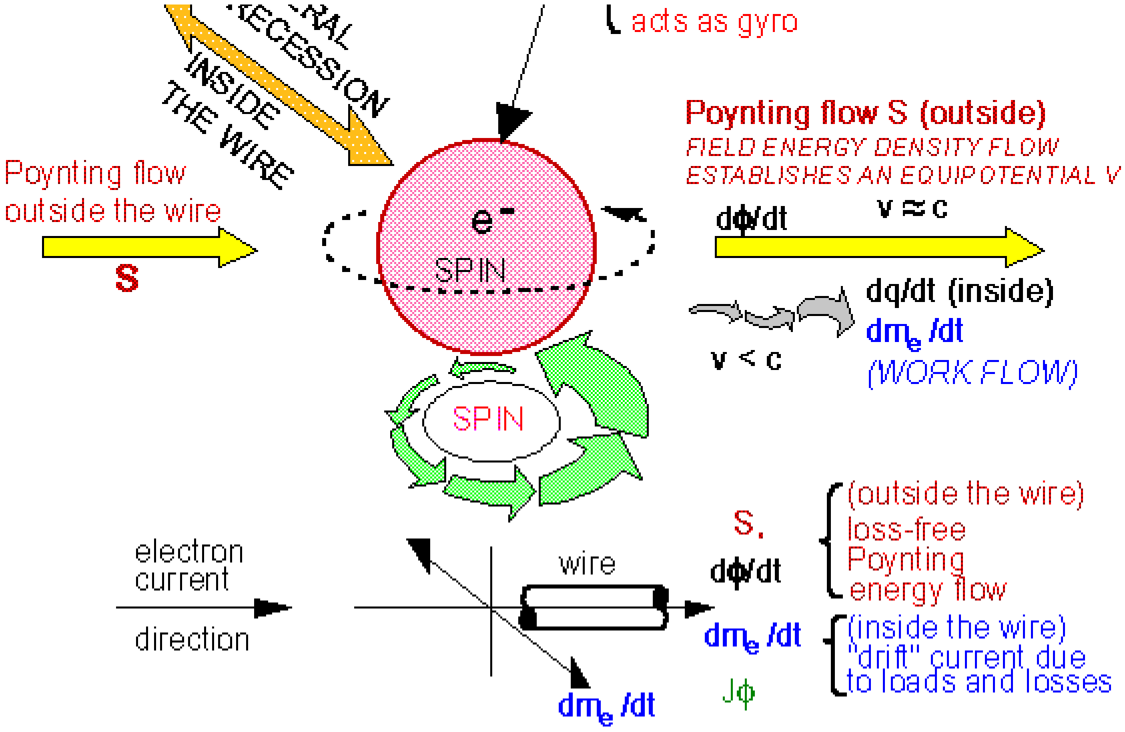
RAL
RECESSION
INSIDE
THE WIRE
acts as gyro
Poynting flow
outside the wire
S
e⁻
SPIN
SPIN
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW
ESTABLISHES AN EQUIPOTENTIAL V
dφ/dt
v ≈ c
dq/dt (inside)
dmₑ /dt
(WORK FLOW)
v < c
electron
current
direction
wire
dmₑ /dt
S,
dφ/dt
(outside the wire)
loss-free
Poynting
energy flow
dmₑ /dt
(inside the wire)
"drift" current due
to loads and losses
Jφ

Figure 6. Currents in and along a conductor, with electron precession.

D-1-F6

© 1993 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

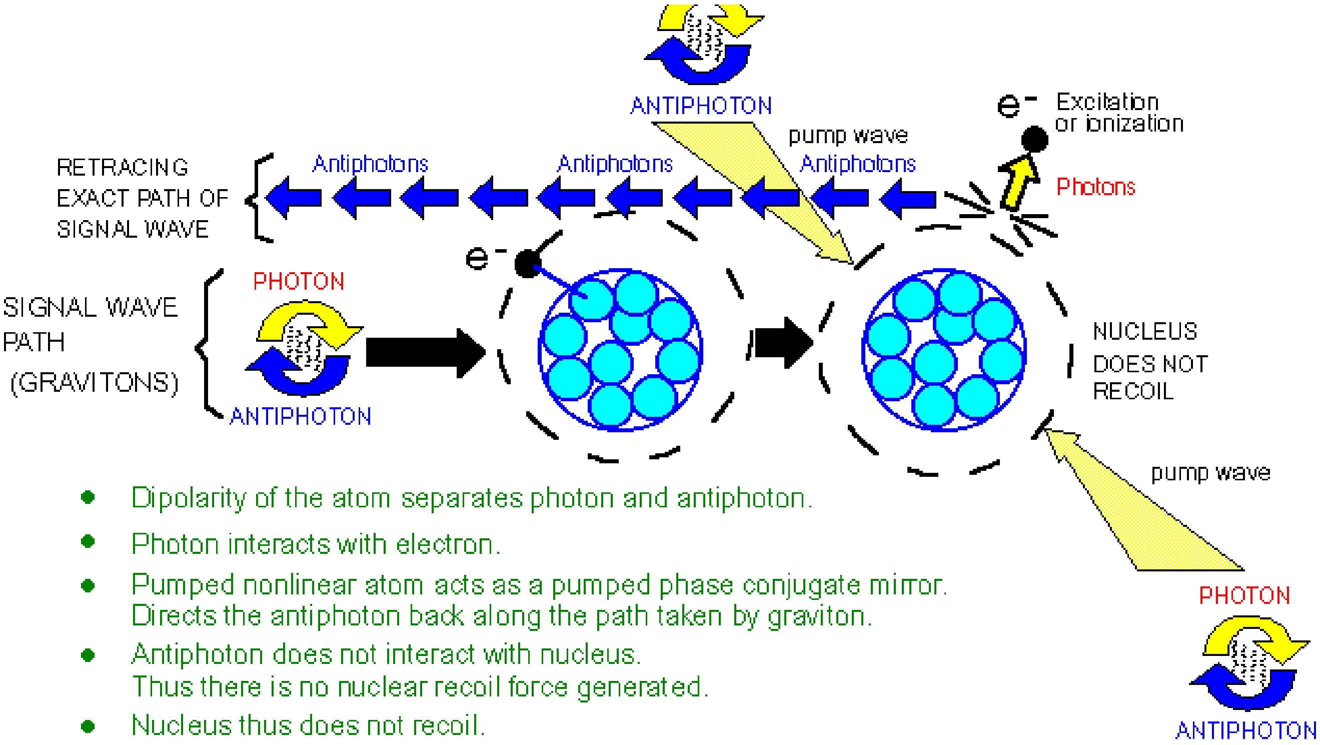
ANTIPHOTON
pump wave
e⁻
Excitation
or ionization
RETRACING
EXACT PATH OF
SIGNAL WAVE
Antiphotons
Antiphotons
Antiphotons
Photons
e⁻
PHOTON
SIGNAL WAVE
PATH
(GRAVITONS)
ANTIPHOTON
NUCLEUS
DOES NOT
RECOIL
pump wave
PHOTON
ANTIPHOTON

- Dipolarity of the atom separates photon and antiphoton.
- Photon interacts with electron.
- Pumped nonlinear atom acts as a pumped phase conjugate mirror. Directs the antiphoton back along the path taken by graviton.
- Antiphoton does not interact with nucleus. Thus there is no nuclear recoil force generated.
- Nucleus thus does not recoil.
- Not a violation of Newton's third law, once one understands the mechanism that generates third law reaction. In the absence of the generating mechanism, there is no third law recoil. In EM field theory, the "missing antiwave" has necessitated third law omission.

Figure 7. Emission of antiphoton; nucleus does not recoil.
Atom acts as a pumped phase conjugate mirror.

D-1-F7

© 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

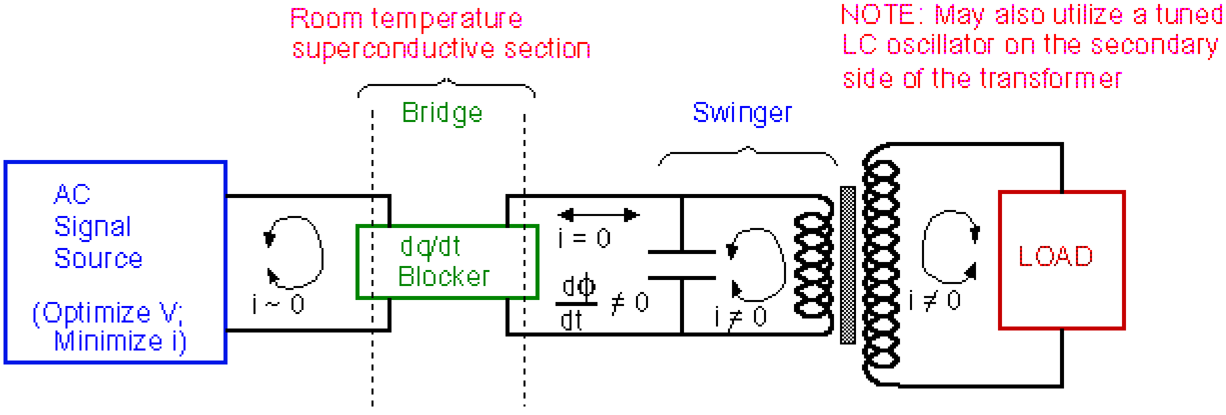
Room temperature superconductive section
Bridge
Swinger
AC Signal Source
(Optimize V; Minimize i)
i ~ 0
dq/dt Blocker
i = 0
$\frac{d\phi}{dt} \neq 0$
i ≠ 0
i ≠ 0
LOAD

NOTE: May also utilize a tuned LC oscillator on the secondary side of the transformer

Patent Pending

Figure 8. dq/dt-free driving a swinger by S-flow.

Poynting energy flow S with its accompanying potential V is shuttled from the source to a tuned LC oscillator in resonance at the driving frequency or slightly below. The LC oscillator is driven by S-flow and dV/dt flow alone. The swinger furnishes its own free electrons, which receive S and EMF from the bridge, thereby producing dq/dt.

© T.E. Bearden 1993, 1994, 1995

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

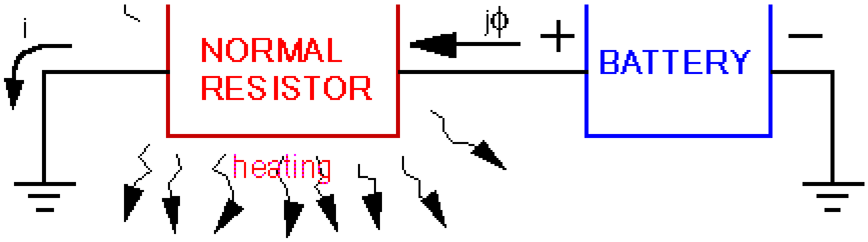

Figure 9a. Normal resistor receives ordered energy and scatters it.

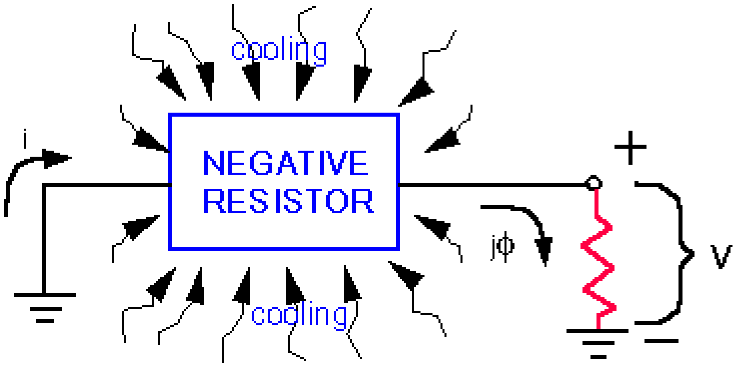

Figure 9b. Negative resistor receives, concentrates, and reorders scattered energy.

Figure 9. Normal resistor versus a negative resistor.

D-1-F9 © T.E. Bearden 1993, 1996

Some 17 signals were fed into a rotating plasma, which added phase conjugates to them. This infolded 17 wavepairs, each pair a coupled wave/antiwave, thereby forming a scalar potential with an artificial Whittaker pump wave internal structure. This infolded pump wave complex was modulated onto a powerful, rippling magnetic field, guaranteeing penetration even into the atomic nuclei. A specific, amplified antiengine for the exact cellular disease state was generated, reversing the diseased cells back to a previous healthy state.

Figure 10. Priore's laboratory apparatus for cellular reversal.

D-1-F10 © T.E. BEARDEN 1993, 1995

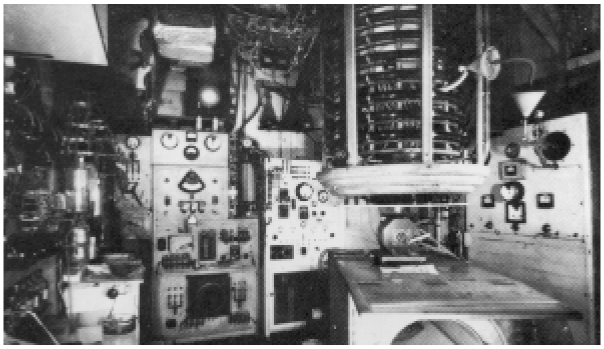

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

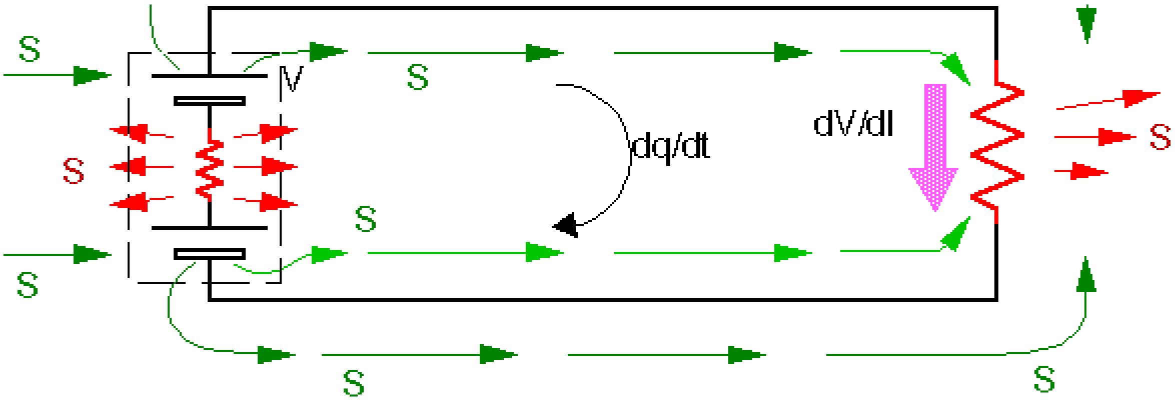

Note: EM texts err in showing the battery "gate" as the source of energy.

Figure 11. Poynting S-flow along a simple circuit.

From particle physics, any electric charge is a broken symmetry in the vacuum's virtual particle flux. Any dipole is thus two locally broken symmetries. Thus the dipole acts as a gate/source, driven by the fierce vacuum flux. So a dipole already "extracts" and gates EM energy flow, directly from the seething vacuum.

Every battery and generator is already a "free energy" source of extracted and gated EM energy from the vacuum.

D-1-F11

© T.E. Bearden 1994

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

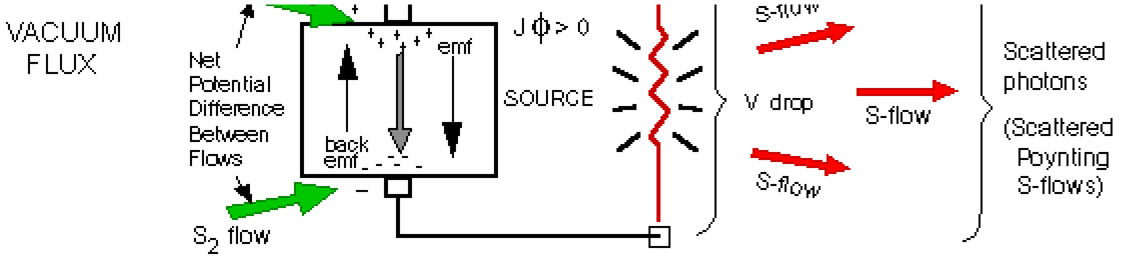
VACUUM FLUX
Net Potential Difference Between Flows
S_2 flow
emf
back emf
$J\phi > 0$
SOURCE
S-flow
V drop
S-flow
S-flow
Scattered photons
(Scattered Poynting S-flows)

Figure 12a. Normal current flow; $J\phi$ is the electron transport of energy, while S is the speed-of-light Poynting energy flow.

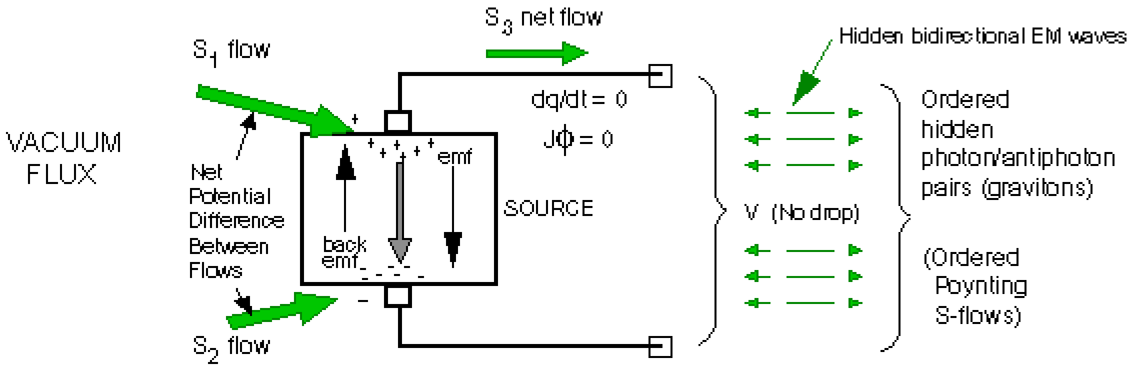
S_3 net flow
S_1 flow
VACUUM FLUX
Net Potential Difference Between Flows
S_2 flow
emf
back emf
$dq/dt = 0$
$J\phi = 0$
SOURCE
Hidden bidirectional EM waves
V (No drop)
Ordered hidden photon/antiphoton pairs (gravitons)
(Ordered Poynting S-flows)

Figure 12b. In a conductor with dq/dt blocked; S-flow and hidden Whittaker-wave flow continue.

Figure 12. Overt and covert field energy density flows.

D-1-F12

© 1994, 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

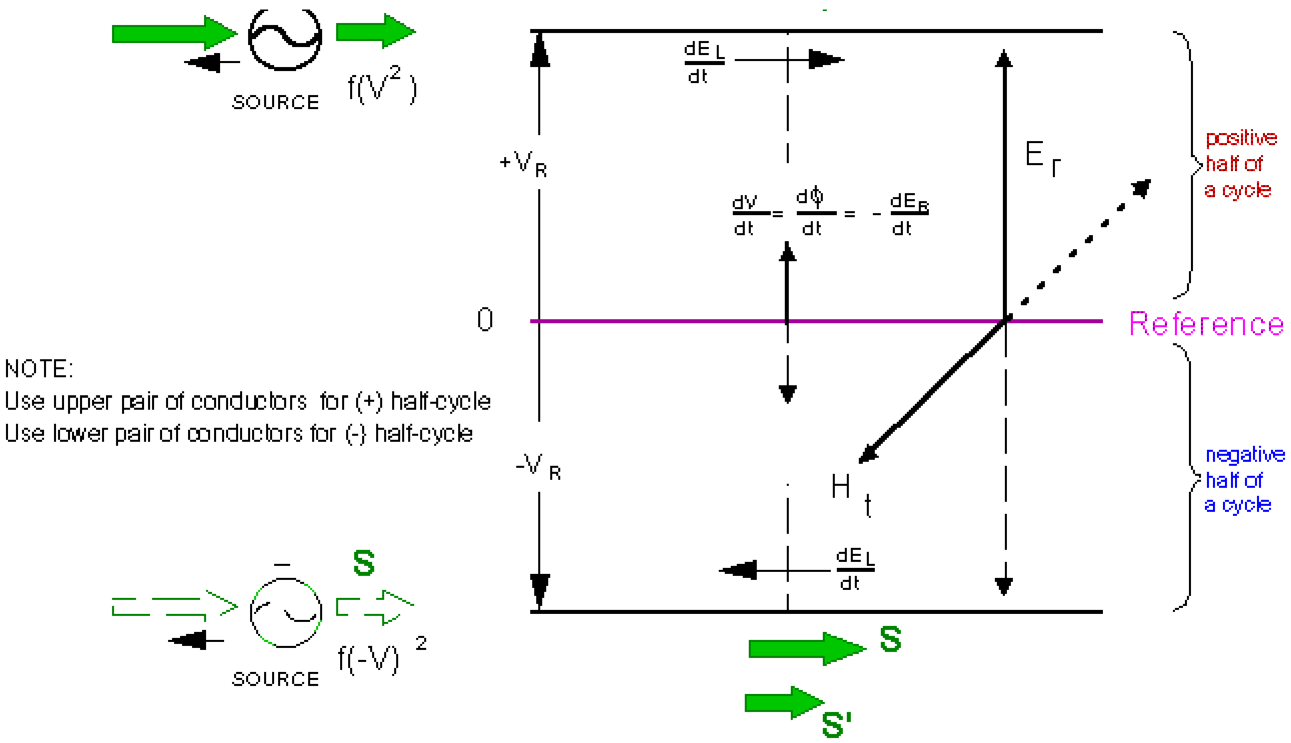
SOURCE
f(V²)
+V_R
0
-V_R
dE_L/dt
dV/dt = dφ/dt = - dE_B/dt
E_r
H_t
positive half of a cycle
Reference
negative half of a cycle
NOTE:
Use upper pair of conductors for (+) half-cycle
Use lower pair of conductors for (-) half-cycle
S
SOURCE
f(-V)²
S
S'

Figure 13. Poynting vector currents S and S' versus massless displacement currents.

D-1-F13

© 1995 T.E. Bearden

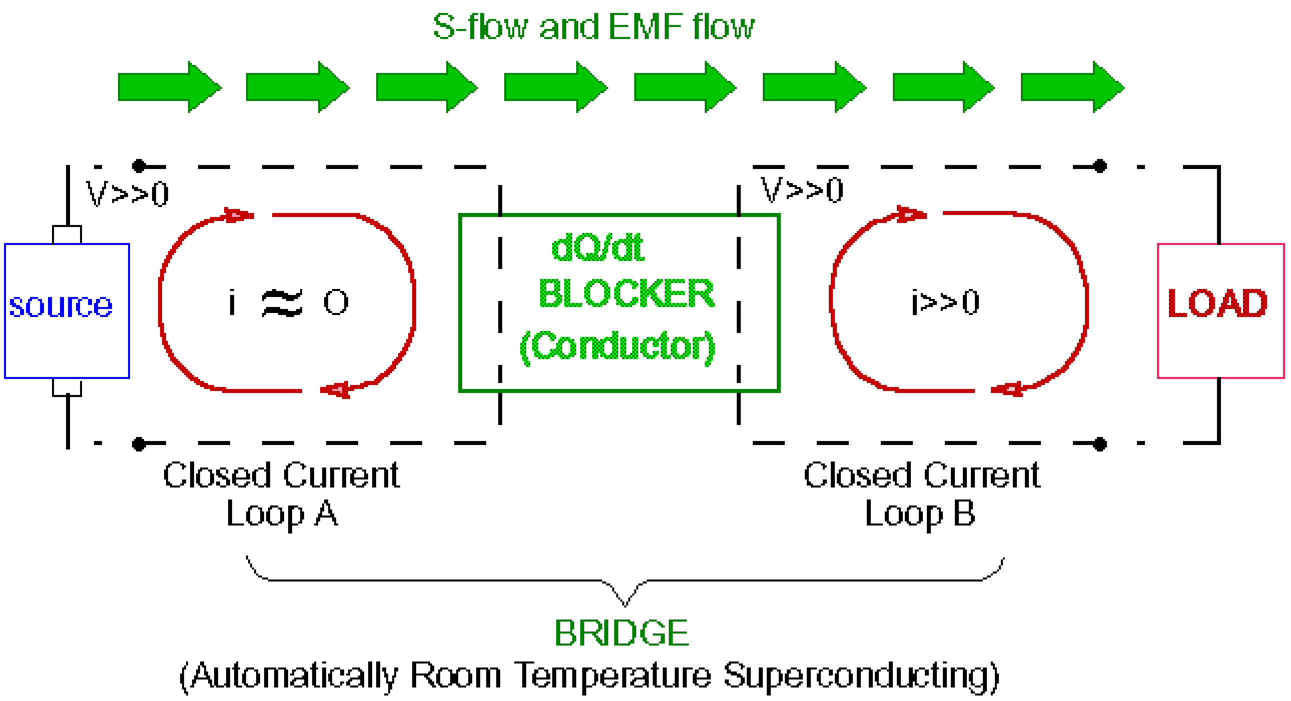
S-flow and EMF flow
V>>0
source
i ≈ O
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
(Automatically Room Temperature Superconducting)

Figure 14. Bridging to enable room temperature superconductivity.

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

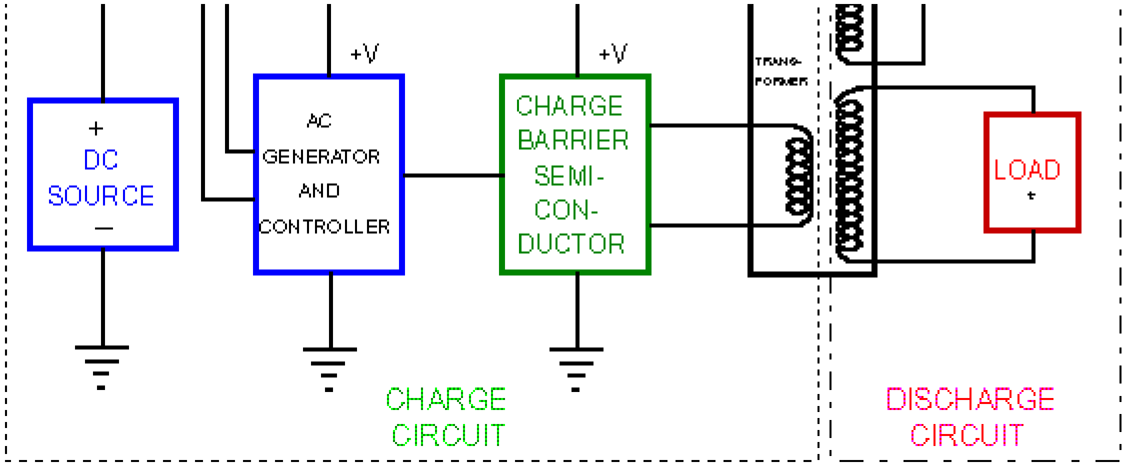
+V
+V
TRANS-
FORMER
+
DC
SOURCE
—
AC
GENERATOR
AND
CONTROLLER
CHARGE
BARRIER
SEMI-
CON-
DUCTOR
LOAD
*
CHARGE
CIRCUIT
DISCHARGE
CIRCUIT

* NOTE:

DISCHARGE CIRCUIT AND LOAD GROUND
MUST BE ISOLATED FROM THE
CHARGE CIRCUIT GROUND.
FEEDBACK LOOP MUST BE RIGIDLY CLAMPED AND LIMITED.

Patent Pending

Figure 15. Use of Fogal semiconductor for overunity (closed system).

D-1-F15

© 1993, 1995 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

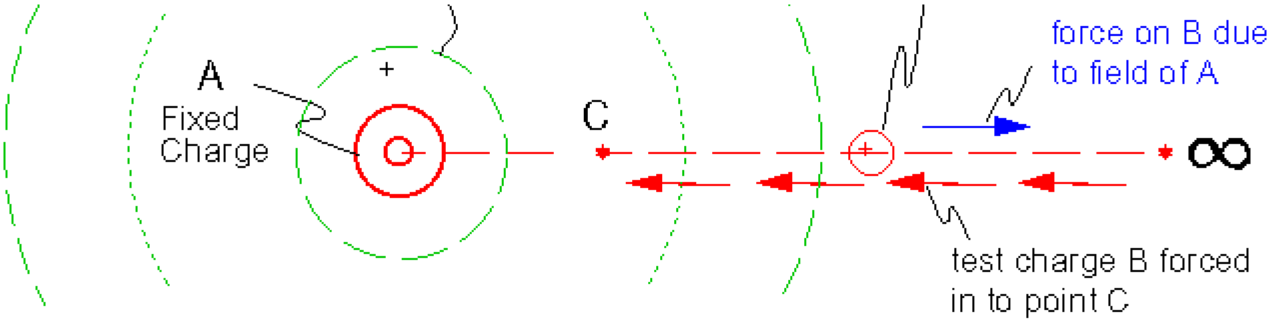
A
Fixed Charge
+
C
∞
force on B due to field of A
test charge B forced in to point C

Figure 16a. Magnitude of the scalar potential is determined by work required to push in a unit charge from infinity; e.g., against a repelling like charge. If test charge is released, work will be performed upon it to translate it back toward infinity

Figure 16a. Process for overunity coefficient of performance.

D-1-F16A

ⓒ T.E. Bearden, 1993, 1994, 1995

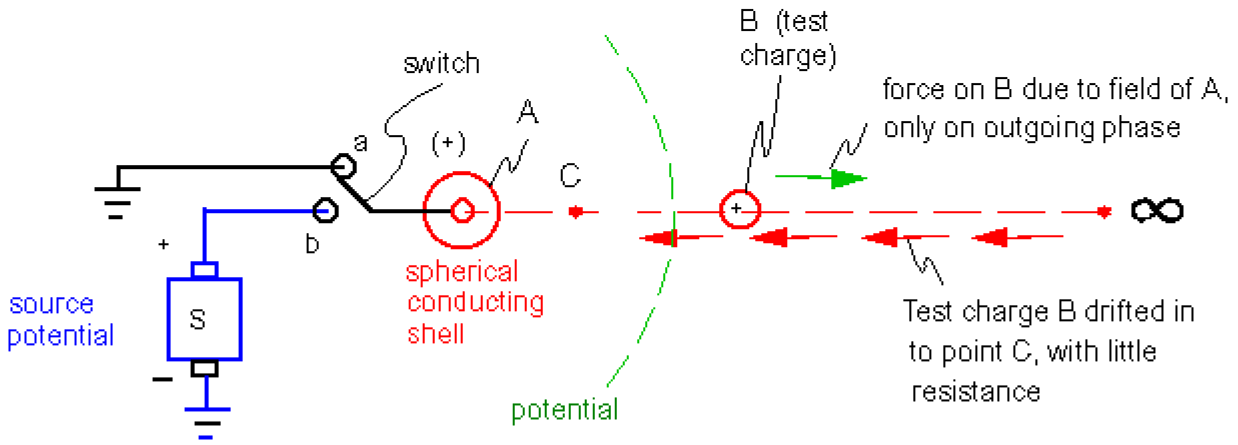
switch
a
(+)
A
C
B (test charge)
force on B due to field of A, only on outgoing phase
∞
b
+
S
−
source potential
spherical conducting shell
Test charge B drifted in to point C, with little resistance
potential

Assumptions: Capacitance of A very small; charge B large, power-free switching

Switch potential off, drift in test charge B, switch potential back on. Can now get free work from repelled test charge B. Get much more free work if A is at a high potential, and charge B is large.

Figure 16 b. Demonstration of overunity efficiency

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

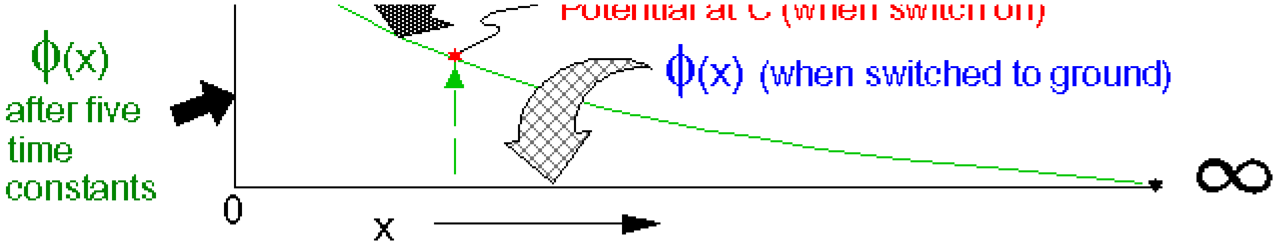

Ideally, switching the potential, rather than flowing charges, requires only minuscule overhead energy dissipation

Figure 16c. Potential to which test charge is exposed.

D-1-F16C

© T.E. Bearden, 1993, 1994, 1995

Open system receiving excess energy from an external source

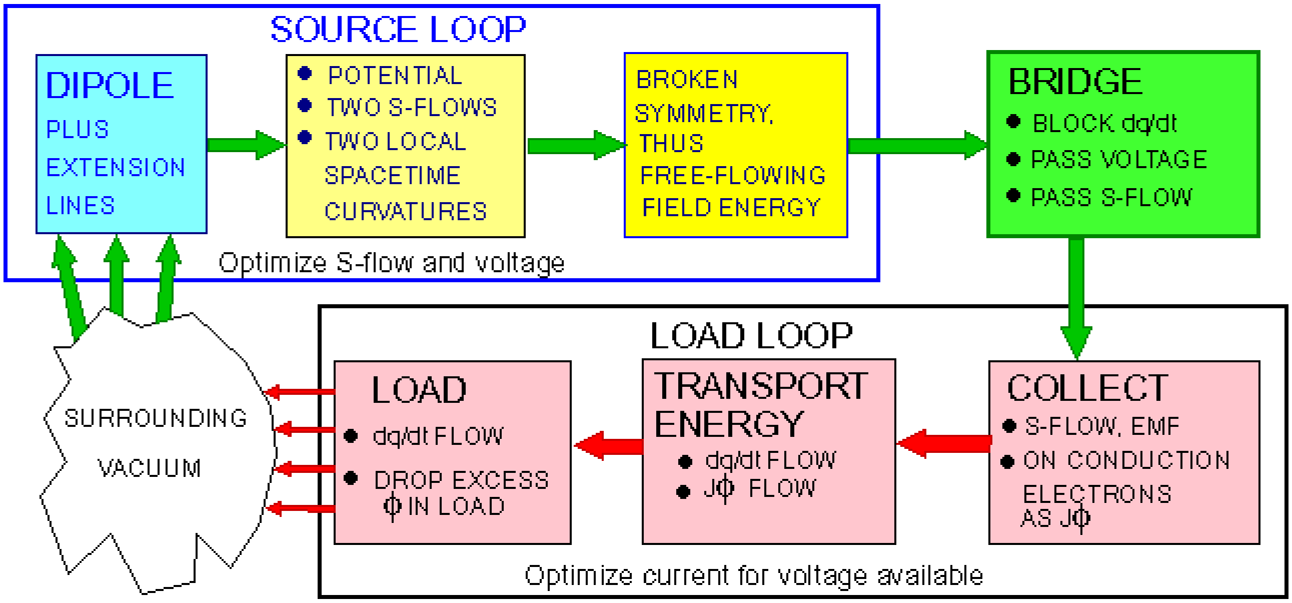
SOURCE LOOP
DIPOLE
PLUS
EXTENSION
LINES
● POTENTIAL
● TWO S-FLOWS
● TWO LOCAL
SPACETIME
CURVATURES
BROKEN
SYMMETRY,
THUS
FREE-FLOWING
FIELD ENERGY
Optimize S-flow and voltage
BRIDGE
● BLOCK dq/dt
● PASS VOLTAGE
● PASS S-FLOW
LOAD LOOP
SURROUNDING
VACUUM
LOAD
● dq/dt FLOW
● DROP EXCESS ϕ IN LOAD
TRANSPORT
ENERGY
● dq/dt FLOW
● Jϕ FLOW
COLLECT
● S-FLOW, EMF
● ON CONDUCTION ELECTRONS AS Jϕ
Optimize current for voltage available

Figure 17. Operation of an overunity electrical power system.

D-1-F 17

© 1992 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies.

Ok

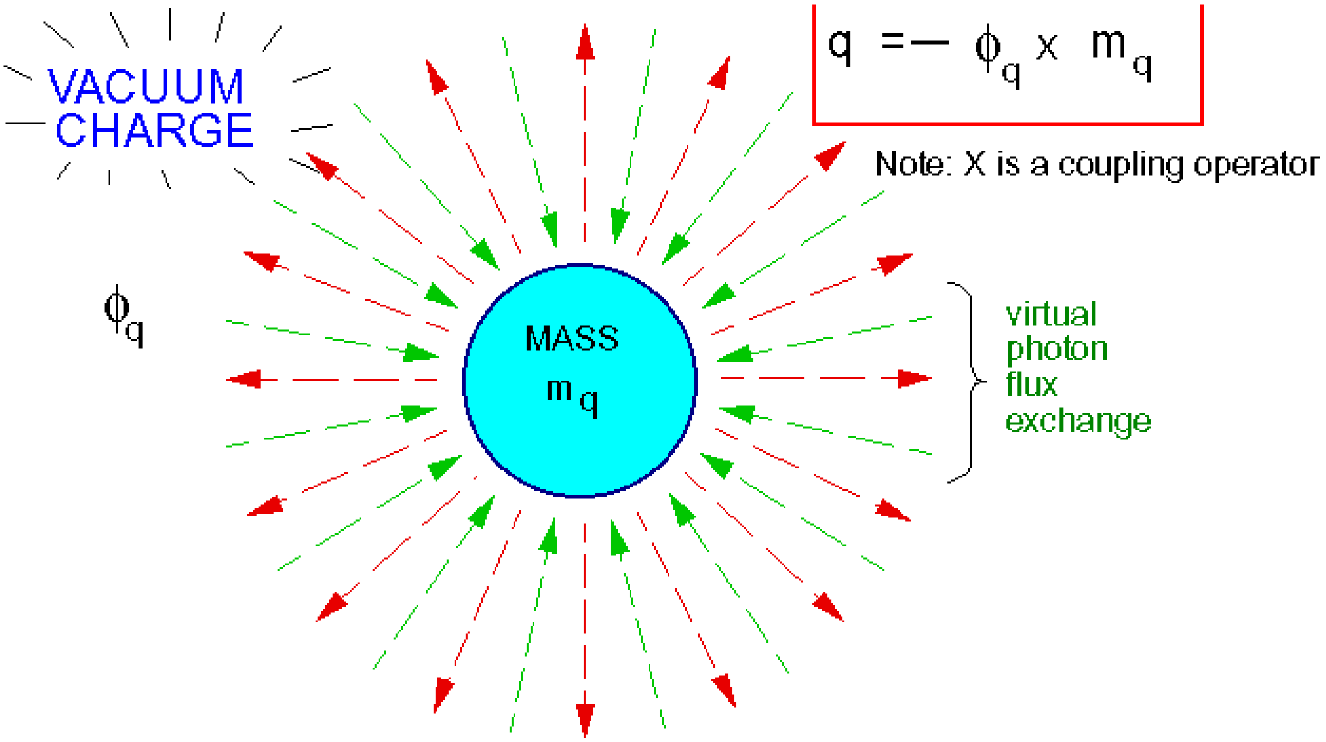
VACUUM CHARGE
$q = -\phi_q \times m_q$
ϕ_q
MASS
m_q
virtual photon flux exchange

Note: X is a coupling operator

Figure 18. An electric charge q is a broken symmetry. It continuously and violently exchanges energy with the vacuum.

D-1-F 18

© 1994 T.E. Bearden

We use cookies to ensure that we give you the best experience on our website. By using this site you are accepting our usage of cookies. Ok

Dawn of The E-BOMB

For the wired world, the allure and the danger of high-power microwave weapons are both very real

By Michael Abrams

In these media-fueled times, when war is a television spectacle and wiping out large numbers of civilians is generally frowned upon, the perfect weapon would literally stop an enemy in his tracks, yet harm neither hide nor hair. Such a weapon might shut down telecommunications networks, disrupt power supplies, and fry an adversary's countless computers and electronic gadgets, yet still leave buildings, bridges, and highways intact. It would strike with precision, in an instant, and leave behind no trace of where it came from.

In fact, it almost certainly is already here, in the form of high-power microwave (HPM) weapons. As their name suggests, HPMs generate an intense "blast" of electromagnetic waves in the microwave frequency band (hundreds of megahertz to tens of gigahertz) that is strong enough to overload electrical circuitry. Most types of matter are transparent to microwaves, but metallic conductors, like those found in metal-oxide semiconductor (MOS), metal-semiconductor, and bipolar devices, strongly absorb them, which in turn heats the material.

An HPM weapon can induce currents large enough to melt circuitry. But even less intense bursts can temporarily disrupt electrical equipment or permanently damage ICs, causing them to fail minutes, days, or even weeks later. People caught in the burst of a microwave weapon would, by contrast, be untouched and might not even know they'd been hit. (There is, however, an effort to build a microwave weapon for controlling crowds; a person subjected to it definitely feels pain and is forced to retreat.)

"HPM sources are maturing, and one day, in the very near future, they will help revolutionize how U.S. soldiers fight wars," says Edl Schamiloglu, a professor of electrical and computer engineering at the University of New Mexico in Albuquerque and one of the leading researchers in this burgeoning field.

The fact that we seldom hear about HPM weapons only adds to their exoticism. Last spring, stories leaked to the press suggested that the Pentagon, after decades of research, had finally deployed such a device in Iraq. And when news footage showed a U.S. bomb destroying an Iraqi TV station, many informed onlookers suspected it was an electromagnetic "e-bomb."

"I saw the detonation, and then I saw the burst—which wasn't much. If they took the station out with that blast, I strongly suspect that we used Iraq as a proving ground" for HPMs, says Howard Seguine, an expert on emerging weapons technology with Decisive Analytics Corp., in Arlington, Va.

But while the U.S. military proudly paraded assorted new war-making technology during its conquest of Iraq, from unmanned combat aerial vehicles to a new satellite-based tracking network, it remained tight-lipped about this "mother of all weapons." Asked at a 5 March news briefing to confirm the rumor, General Tommy Franks, head of U.S. forces during the war, would only say, "I can't talk to you about that because I don't know anything about it."

Military secrecy is nothing new, of course. What is known about microwave weapons is that the U.S. military has actively pursued them since the 1940s, when scientists first

CHIP SIMONS

Microwave weapons researcher Edl Schamiloglu sits in front of the Pulserad-110A accelerator, which his lab at the University of New Mexico uses to produce single 100-nanosecond pulses of electron beams, each pulse emitting hundreds of megawatts of power.

observed the powerful electromagnetic shock wave that accompanied atmospheric nuclear detonations, suggesting a new class of destructiveness. While much of the work on HPMs remains classified, the Pentagon has also recently sponsored a number of U.S. university laboratories to work out the basic principles of microwave weapons, including reliable and compact nonnuclear ways of generating microwave pulses.

Many of those results are being published in the open literature [see To Probe Further, p. 30]. In fact, all you need is a reasonable grasp of physics and electrical engineering to appreciate the ingeniousness of microwave weapons. Anyone with a technical bent could probably also build a crude e-bomb in their garage, a thought that security-minded folks find rather troubling.

How they work

From the military's perspective, HPM weapons, also known as radiofrequency weapons, have many things going for them: their blast travels at the speed of light, they can be fired without any visible emanation, and they are unaffected by gravity or atmospheric conditions. The weapons come in two flavors: ultrawideband and narrowband. Think of the former as a flashbulb, and the latter as a laser; while a flashbulb illuminates across much of the visible spectrum (and into the infrared), a laser sends out a focused beam at a single frequency.

Like the flashbulb, ultrawideband weapons radiate over a broad frequency range, but with a relatively low energy (up to tens of joules per pulse). Their nanoseconds-long burst produces a shock that indiscriminately disrupts or destroys any unshielded electronic components within their reach. The bomb's destructiveness depends on the strength of the ultrawideband source, the altitude at which it is initiated, and its distance from the target [see "E-Bomb Anatomy," p. 29].

Narrowband weapons, by contrast, emit at a single frequency or closely clustered frequencies at very high power (from hundreds up to a thousand kilojoules per pulse), and some can be fired hundreds of times a second, making an almost continuous beam. These pulses can be directed at specific targets—say, a command and control complex positioned on the roof of a hospital in a densely populated neighborhood—and tuned to specific frequencies. Technologically more sophisticated than ultrawideband sources, they are far more difficult to develop, but are reusable and potentially of much greater use to the U.S. military.

Both versions wreak the same kind of havoc on just about any kind of unprotected electronic equipment. Particularly vulnerable is commercial computer equipment; anything in excess of just tens of volts can punch through gates in MOS and metal-semiconductor devices, effectively destroying the device, explains Carlo Kopp, a visiting research fellow in military strategy at the Strategic and Defense Studies Centre in Canberra, Australia, and a computer scientist who lectures at Monash University in Melbourne. The higher the circuitry's density, the more vulnerable it is, because less energy is required to overload and destroy the transistors. HPMs also produce standing waves in electrical grid wiring and telephone and communications wiring, entering through cables, antennas, and even ventilation grills. They can immobilize vehicles with electronic ignition and control systems, too.

"Since the frequency is high, this permits parasitic or stray capacitances to couple energy via paths in the circuit that may not be protected against overvoltage," Kopp explains.

The e-bomb

You could deliver an e-bomb in a number of ways: cruise missile, unmanned aerial vehicle, or aerial bomb. Whether ultrawideband or narrowband, the e-bomb consists of both a microwave source and a power source [again, see diagram, p. 29]. Ultrawideband e-bombs aim to create an electromagnetic pulse like that accompanying a nuclear detonation, except that the nuclear material is replaced with a conventional, chemical explosive. The microwave source typically relies on an extremely fast switching device, according to Kopp, who has written widely on weaponizing HPM technology. Narrowband e-bombs might use a virtual cathode oscillator (vircator) tube or a variant of a magnetron. Though termed narrowband, they don't have the high coherency seen in signal-carrying applications, Kopp says.

It takes gigawatts of power to feed an e-bomb's microwave source. For that, the flux compression generator, or FCG, is a good choice, says Kopp. Invented by Clarence ("Max") Fowler at Los Alamos National Laboratory after World War II as a byproduct of research into atomic bomb detonators, FCGs are conceptually simple. The best-known type consists of an explosive-packed copper cylinder surrounded by a helical current-carrying coil. Upon detonation, the explosion flares out the cylinder, short-circuiting the coil and progressively reducing the number of turns in the coil, thus compressing the magnetic flux. Large FCGs have produced tens of gigawatts, and they can be cascaded—connected end to end—so that the output from one stage feeds the next.

Despite its simplicity, an FCG-powered e-bomb is probably too difficult for the average terrorist to build on the cheap. For one thing, to test the assembled apparatus, you have to blow it up. For weapons researchers, the e-bomb poses other problems. The strength of the shock wave dissipates rapidly as it moves out from the explosion. To knock out an electrical power substation, for example, the weapon has to strike within about a hundred meters. "Like all microwave radiation, the effect follows an inverse square law with increasing distance," Kopp notes. Though the explosion needed to force out the current can be fairly small, it keeps the munition from being fully nonlethal and nondetectable. Also, anything that's been hardened or shielded against an electromagnetic pulse from a nuclear bomb will probably emerge unscathed.

Focused like a laser

The type of narrowband HPM weapons that the U.S. military is looking at offers everything that e-bombs do not. They're nonlethal, reuseable, and tunable, and they can be fired from miles away. Like a laser, the focused beam disperses only slightly over great distances. With a frequency range that is between about 1 and 10 GHz, they can penetrate even electronics shielded against a nuclear detonation. The deepest bunkers with the thickest concrete walls are not safe from such a beam if they have even a single unprotected wire reaching the surface.

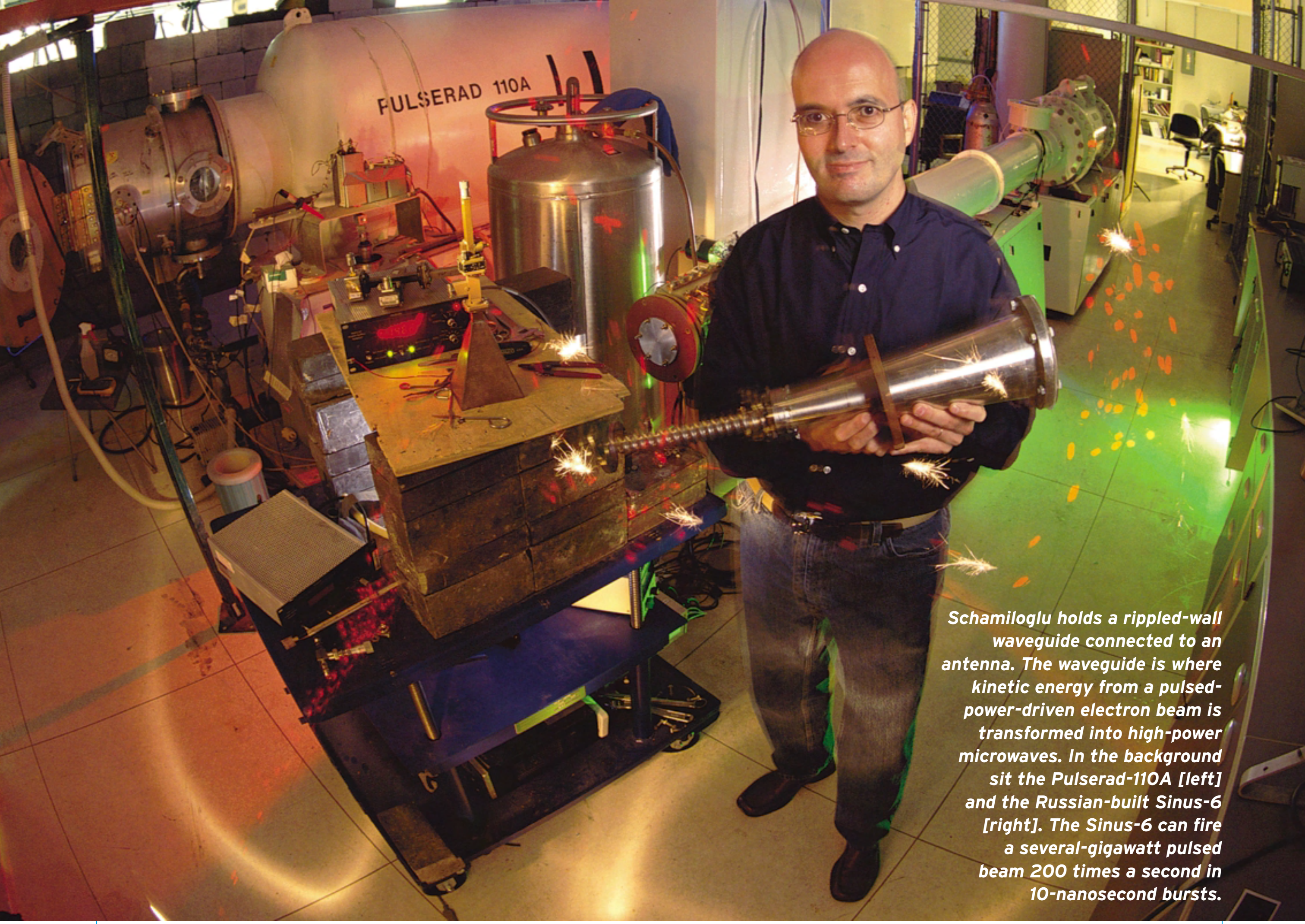

Schamiloglu holds a rippled-wall waveguide connected to an antenna. The waveguide is where kinetic energy from a pulsed-power-driven electron beam is transformed into high-power microwaves. In the background sit the Pulserad-110A [left] and the Russian-built Sinus-6 [right]. The Sinus-6 can fire a several-gigawatt pulsed beam 200 times a second in 10-nanosecond bursts.

CHIP SIMONS

A microwave beam is created much like a laser beam. Between the batteries (or other power source) and the beam sit three elements: capacitors that turn the stored energy into an electron beam of nanosecond bursts, a microwave source that converts the electron beam into focused, high-frequency electromagnetic waves, and an antenna that points and shoots the beam.

Kirtland Air Force Base, in Albuquerque, N.M., is considered the epicenter of the Pentagon's research on pulsed-power electromagnetic weapons. There, its premier pulsed-power system, the Shiva Star, is housed behind meter-thick walls [see photo, p. 28]. An Air Force spokesperson refused to comment on what goes on in their pulsed-power programs, but a fact sheet on the Web site of Kirtland's Directed Energy Directorate describes the Shiva Star as capable of producing "120 thousand volts and 10 million amps for down to one millionth of a second to produce a power flow equivalent to a terawatt."

The Kirtland machine isn't used to investigate HPM weapons per se, and its massive size makes it clearly impractical for delivering microwave beams to any spots of real military interest. Indeed, one big push in microwave weapons has been toward portability. "Back in the 1960s and 1970s, the attitude was, 'Yeah, we can do it—but we need Hoover Dam as our power supply,' " says Seguine. But just as batteries for cellphones and laptops have shrunk and gained capacity, so have sources for microwave weapons.

In the 1990s, the U.S. Air Force Office of Scientific Research set up a five-year Multidisciplinary University Research Initiative (MURI) program to explore microwave sources. One of those funded was the University of New Mexico's Schamiloglu, whose lab is located just a few kilometers down the road from where the Shiva Star sits behind tightly locked doors. Thanks in large part to his and his colleagues' efforts, the fundamental capabilities and limitations of high-power microwave sources are now better understood and appreciated.

Amidst the lead bricks and clutter in Schamiloglu's basement lab lies his masterwork: the Sinus-6. "A lot of laboratories come up with very cute names for these devices," Schamiloglu notes with a smile. "We never did." With a huge cylinder at one end connected to the long microwave source, the Sinus-6 looks like a giant torch lying on its side [see photo, above]. The big cylinder contains a Tesla transformer, whose two coils vibrate in resonance and amplify the incoming voltage "with nearly 100-percent efficiency," Schamiloglu says. Once the pulse has been transformed into an electron beam, it is guided by a strong axial magnetic field through the long tube that will turn it into microwaves.

The Sinus-6 can fire a several-gigawatt pulsed beam 200 times a second in 10-nanosecond bursts. "It has to be pulsed power because what you're after is high peak power," says Schamiloglu. "The power in the microwaves is going to depend on the electric field squared, so if you generate very large power, then the electric field is going to be big."

How big? To drive the Sinus-6's beam continuously for an entire second, you'd need to supply about 25 gigajoules—"the entire output of a typical coal-fired electrical plant for 10 full seconds," Schamiloglu says. Another reason for pulsed rather than continuous power is to avoid a problem at the output end: the air around the antenna would heat to a plasma that in turn would interfere with a continuous beam at these power levels.

The key to reaching gigawatts of power is dumping all the energy in one gigantic, nearly instantaneous pulse. A pressurized gas switch prevents the Tesla transformer from prematurely dumping as it builds up for the next pulse. The switch is filled with highly compressed and nonconducting nitrogen gas. When the transformer coils reach 700 kV, the nitrogen gas breaks down, and the pulse leaps through to the electron-beam diode.

The U.S. Air Force's Shiva Star is a pulsed-power system used to simulate the effects of nuclear weapons.

U.S. AIR FORCE

"Once you've fired the switch, it conducts, it generates a pulse," says Schamiloglu. "It conducts because you've made a plasma channel out of the gas. Then you have to wait for that plasma to recombine and form a neutral gas again. A typical time scale for this thing to recombine and fizzle out and be a neutral gas again is probably on the order of milliseconds."

Among the best candidates for supplying microwaves is the backward wave oscillator; it has the advantage of being tunable (plus or minus 20 percent) and producing output in the 4–10-GHz range. To turn the kinetic energy from the Sinus-6's electron beam into high-power microwaves, the oscillator uses a rippled-wall waveguide, also called a slow-wave structure [see photo, p. 27].

The structure sets up standing electromagnetic waves in such a way that energy is rapidly transferred to them from the incoming beam of relativistic electrons from the Sinus-6. This growing energy initially propagates in the opposite direction of the beam's movement—hence the device's name—and is then reflected forward and radiated in the form of high-power microwaves. Backward wave oscillators, by the way, are also being tested as a way to push giant sails into outer space, to detect space debris, and to clear minefields.

Being able to tune an HPM weapon comes in handy when a particular target proves invulnerable to a particular frequency. "Experience has shown that if the frequency is slightly altered, measurable effects are discerned," Schamiloglu notes. People used to believe that varying the frequency of HPMs wasn't practical, but Schamiloglu and his students proved them wrong.

Coincidence and curiosity led to their discovery. Schamiloglu first acquired the Sinus-6 from Russian researchers in the early 1990s. (The Soviet Union once boasted a sophisticated program to develop microwave weapons; after its collapse, parts of that legacy were put up for sale, to the delight of researchers like Schamiloglu.) But once the apparatus was assembled in his New Mexico lab, he couldn't get it to operate as promised, so Russian colleagues flew over to help.

"One of them took the RF structure [the rippled-wall waveguide] and started hammering on the thing," Schamiloglu recalls. When they tried it again, everything worked. "I was baffled why manhandling this RF structure—ramming it in—could affect the power so much," says Schamiloglu. So he started a series of experiments in which he slightly displaced the backward wave oscillator by increments. With a little experimentation assisted by computer simulations, his team found that the frequency could be adjusted by changing the distance between the diode and the microwave source. The result is that the backward wave oscillator is now one of the few pulsed-power HPM sources that can be tuned.

Smaller is better

One disadvantage of this oscillator, however, is that it needs an external magnetic field to create the microwave beam, a major hurdle to making the whole system smaller. The size of the Sinus-6 and attendant equipment in Schamiloglu's basement suggests that the U.S. military is nowhere near fielding a narrowband HPM weapon. "When I first started working on high-power narrowband sources, we joked that you can do more damage dropping this equipment on someone than you can by using it," he recalls. "People know how to make microwave sources in the laboratory. The challenge is to take this and package it into an autonomous platform and have it function at the same parameter levels."

Schamiloglu is now hard at work under a new MURI program to study the possibilities of making a compact pulsed-power source. Current narrowband generators are typically several meters long, batteries not included. Schamiloglu and his colleagues are studying how to incorporate novel ceramics into pulsed-power systems, which they believe will allow the length of such sources to be halved. The trick is identifying materials with a high dielectric constant that can also survive the harsh electric fields. "Materials will be an important part in making the next giant leap," he says.

Life in a glass house

Among those agreeing that narrowband HPM weapons will need more refining before they become truly useful to the military is Loren B. Thompson, chief operating officer of the Lexington Institute, a military think tank based in Arlington, Va. He looked

● E-Bomb Anatomy

In this hypothetical design for an e-bomb, a two-stage flux compression generator provides gigawatts of power to the virtual cathode oscillator (vircator), which produces the high-power microwaves. The bomb's destructiveness depends on the microwave source and target's vulnerability to electromagnetic attack, among other things, but a 10-GW, 5-GHz HPM device would have a "lethal" footprint 400 to 500 meters across, producing field strengths of several kilovolts per meter. Such an e-bomb would wreak major havoc if detonated over a heavily populated area.

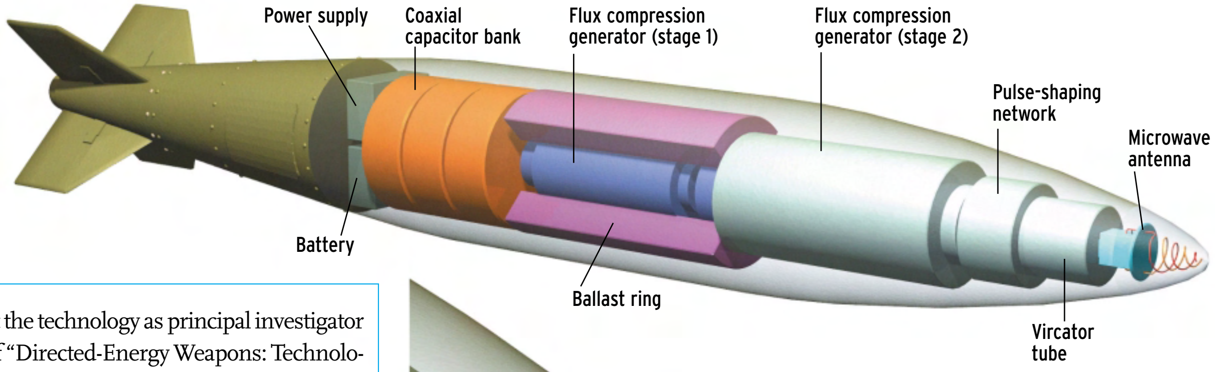
Power supply
Coaxial capacitor bank
Flux compression generator (stage 1)
Flux compression generator (stage 2)
Pulse-shaping network
Microwave antenna
Battery
Ballast ring
Vircator tube

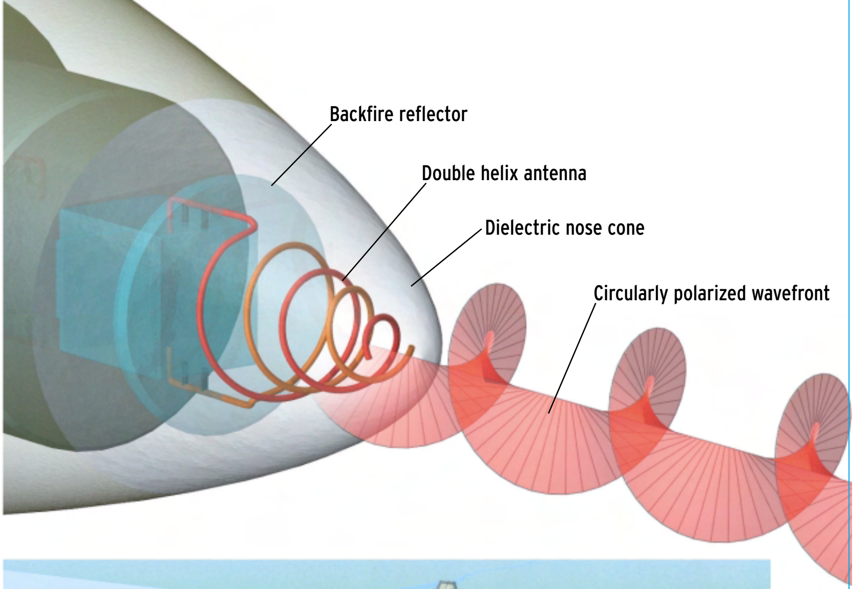
Backfire reflector
Double helix antenna
Dielectric nose cone
Circularly polarized wavefront

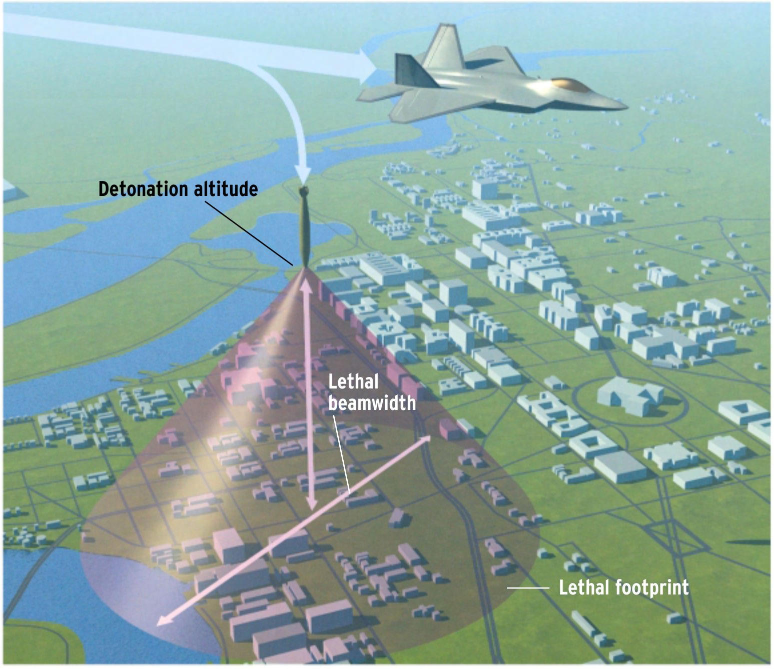
Detonation altitude
Lethal beamwidth
Lethal footprint

ILLUSTRATION: JOHN MACNEILL SOURCE: CARLO KOPP

at the technology as principal investigator of "Directed-Energy Weapons: Technologies, Applications and Implications," a report that the institute put out in February. "We have some fairly rudimentary weapons that we're ready to use," Thompson says. "This is going to be a very important weapons technology, and the basic physical principles are well understood. But the military is having some difficulty in assimilating them."

Thompson's report speaks of a future with satellites delivering missile-debilitating microwaves, unmanned vehicles that fly by and destroy communications systems, and war without civilian casualties. But the fact remains that it's the U.S. military—as well as U.S. financial institutions, PCs, and Game Boys—that will be the most susceptible to such weapons.

"One of the things that happened during the last 10 years—as the Pentagon fell in love with network-centered warfare—is that we purchased a lot of very fragile digital systems off the shelf from commercial sources," Thompson notes. Such moves were taken in the name of cost and efficiency, but the resulting equipment is almost certainly more vulnerable to electromagnetic attack than the vacuum tubes and heavy metal-encased electronics of yesteryear.

"Computers become more vulnerable as the voltage at which they operate becomes smaller," says Victor Granatstein, professor of electrical engineering at the University of Maryland in College Park, who is studying the effects of microwave pulses on integrated electronics. "When our opponent was the

Soviet Union, the electronics were much more robust because they weren't miniaturized. Now they have very thin oxide layers that can easily break down." Wireless networking makes matters worse. Computers and other communications devices now have antennas attached, giving an electromagnetic pulse a direct pathway to its guts.

Meanwhile, the U.S. Navy no longer requires that all its hardware be hardened against nuclear electromagnetic pulses. It deemed that maintaining those standards was too costly and slowed down the integration of new technology. The presumption was that after the Cold War, nobody would be using nuclear bombs, says the Lexington Institute's Thompson. "Whenever I ask the admirals, 'Well, what if someone did use a nuclear bomb?,' I just get this kind of blank I-don't-have-an-answer-for-that sort of look."

Don't Try This at Home

First things first: we do not under any circumstances recommend that you build your own electromagnetic weapon. But if you're hellbent on adding to the mayhem, cheaply and without too much studying, you might try a high-energy radiofrequency, or HERF, gun. As described by engineering student Rostislav Persion on his Web site, Voltage Labs, which is devoted to do-it-yourself electromagnetic weapons, you can make one from a microwave oven.

Before you begin, though, wait until everyone else has left the house. Next, take apart the microwave oven, but don't disconnect any of the components; manufacturers intentionally make this hard to do, so you may end up just breaking the machine. Inside you will find the oven's microwave source: the magnetron. Wrap a tube of sheet metal around it to act as a waveguide. Your power source is the house's ac, so just plug in the oven and point it at your TV. Warning: there's a real possibility that you will burn yourself instead.

For schematics and a demonstration, head to the Web at http://www.voltsamps.com/pages/projects/herf004/. —*M.A.*

In the wrong hands

The scariest part of microwave weapons may be that crude forms of the technology are readily available to anyone right now. "Any nation with a 1950s technology base capable of designing and building nuclear weapons and radars" can build an e-bomb, says military analyst Kopp. Indeed, more than 20 countries now have programs to develop some type of RF weapon.

"The more widespread the technology is, the more likely that people with nefarious purposes will have access. It's just an inescapable fact," says Thompson. "I don't know what we're going to do. Nobody in Washington knows. I imagine that the way the clear thinking starts is with a catastrophe."

Criminals and pranksters have already started exploiting that weakness. In one of the more harmless applications, a Japanese scam artist rigged up a weak microwave generator inside a suitcase to rip off a pachinko parlor. When he placed the suitcase next to one of the machines (which is something like a cross between a slot machine and a pinball machine) and turned it on, the pachinko machine went haywire and disgorged a pile of coins. The perp managed the trick several times before he was caught.

Other press accounts hint at electromagnetic weapons being deployed by Chechen troops, and by an unnamed assailant trying to topple London's futures market [see "Don't Try This at Home," above].

Thankfully, protecting yourself against the microwave-enabled goofballs of the world isn't too difficult. "It is analogous to existing techniques used to trap RF interference inside equipment, except that the higher power levels require special measures," Kopp notes. Rooms or equipment chassis must become electrically sealed Faraday cages, and protective devices must be added wherever cables enter the protected volume. "Optical fibers are very useful in this game."

Such protective measures are a lot cheaper to design in from the beginning than to add on afterward, says Howard Seguine. "The general rule of thumb is that if you do the hardening during the design phase, it increases the cost roughly 1 percent. If you do it afterward, it may cost as much as 30 percent more."

But maybe hardening is a waste of time. Arthur Varanelli, a Raytheon Co. engineer who has helped write several IEEE standards for electromagnetic field measurement, human exposure, and safety, is skeptical that a malicious prankster could exploit the technology.

"Some of this stuff is just so far out there," Varanelli says. "I just don't see people running around with Buck Rogers ray guns. It's great for a science fiction writer, great to prey upon people's fears." He scoffs at the suggestion that a do-it-yourselfer could build a microwave weapon potent enough to do real damage. "People can put tacks in the road. Are we worried about electronic tacks in the air?"

The wide disparity in opinions and the uncertainty about microwave weapons, from Loren Thompson on one end to Arthur Varanelli on the other, are all part of what makes them so powerful, says military analyst John Pike, who is director of GlobalSecurity.org (Alexandria, Va.). "It all depends on the complex interactions between the weapon and the target," he notes. "I can set up a strap-down chicken test that makes [an HPM weapon] look pretty good. But as soon as I start getting into real-world targets, maybe it doesn't work so well."

"Part of the story is we don't know what the story is," Pike says. "These are weapons that by their nature seek the shadows. And unlike cluster bombs or atomic bombs, they aren't going to leave behind unambiguous evidence of their use." ●

To Probe Further

For a detailed technical discussion of high-power microwaves, see *High-Power Microwave Sources and Technologies*, edited by IEEE Fellows Robert J. Barker and Edl Schamiloglu (Wiley–IEEE Press, 2001). Schamiloglu is also coauthor, with James Benford and John Swegle, of the forthcoming *High-Power Microwaves*, 2nd edition (Institute of Physics, 2004).

The truly prepared, or merely paranoid, will want to consult Carlo Kopp's "Hardening Your Computing Assets" at http://www.globalsecurity.org/military/library/report/1997/harden.pdf.

BEARDEN
FER DE LANCE
12.14 15:24
BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

**ORDER THE ALL NEW
2ND EDITION
OF "FER DE LANCE"**

Updated to the year 2002

**Stunning new revelations
about the secret scalar
EM wars the media ignores**

**THIS BOOK IS THE
ONLY SOURCE FOR
THIS INFORMATION**

-

The Tom Bearden
Website

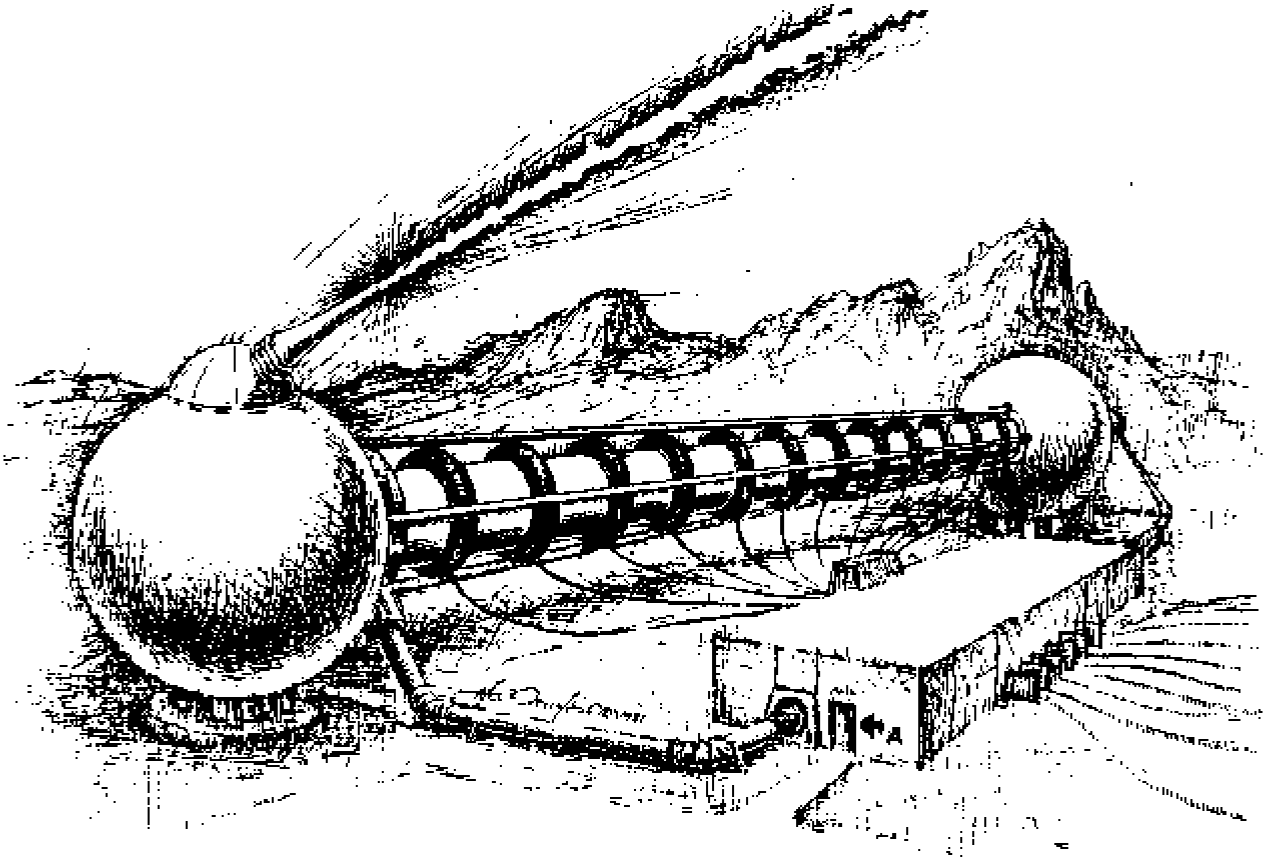

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

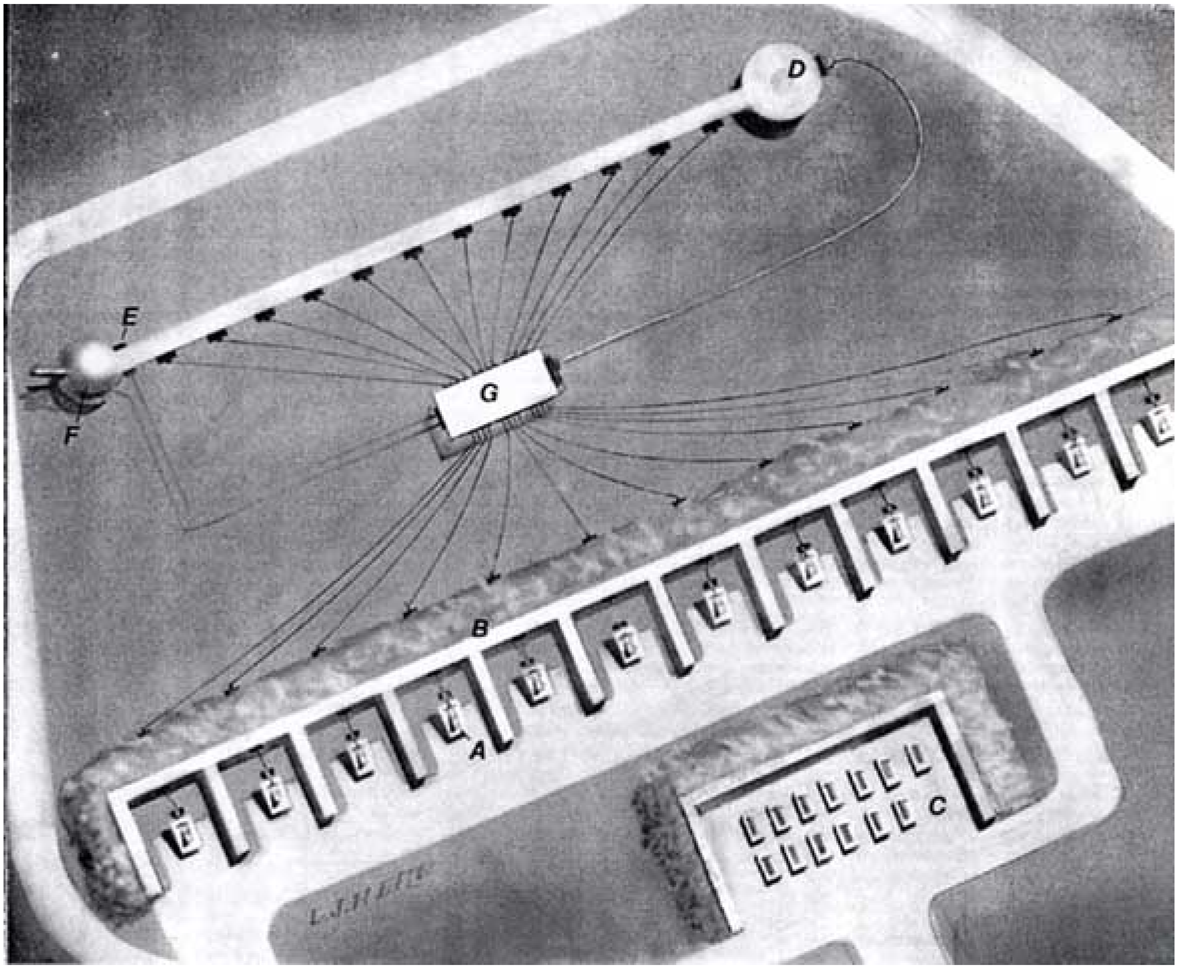

Aviation Week & Space Technology July 28, 1980

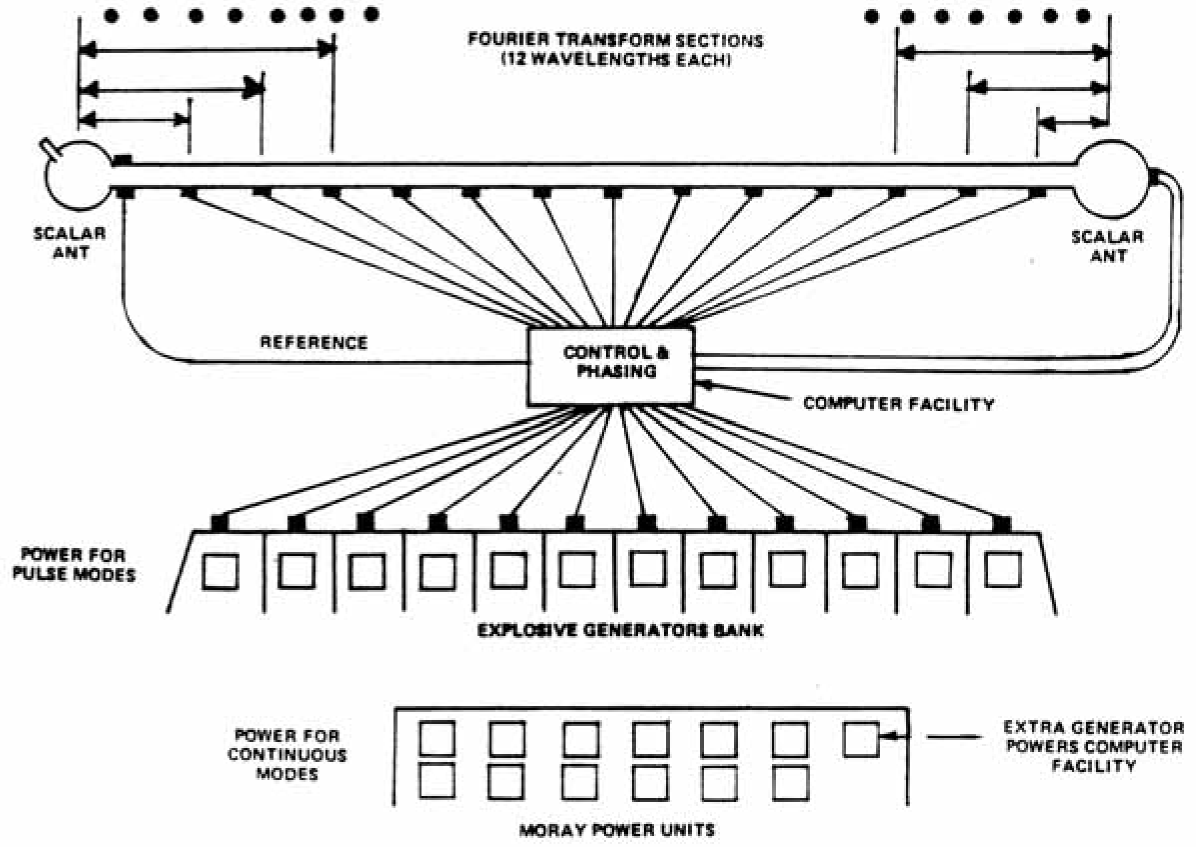
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

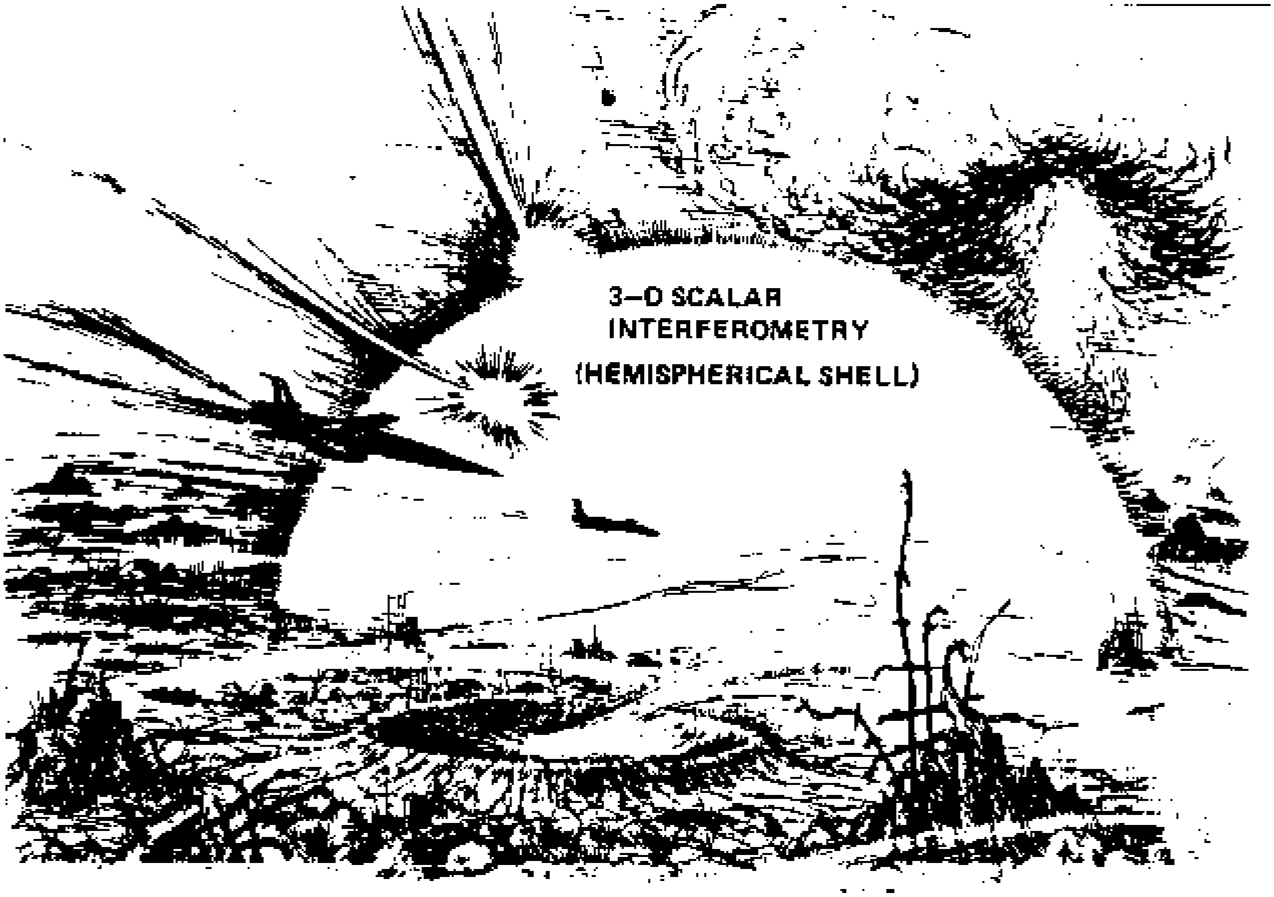
3–D SCALAR
INTERFEROMETRY
(HEMISPHERICAL SHELL)

Figure 8. The Tesla Shield

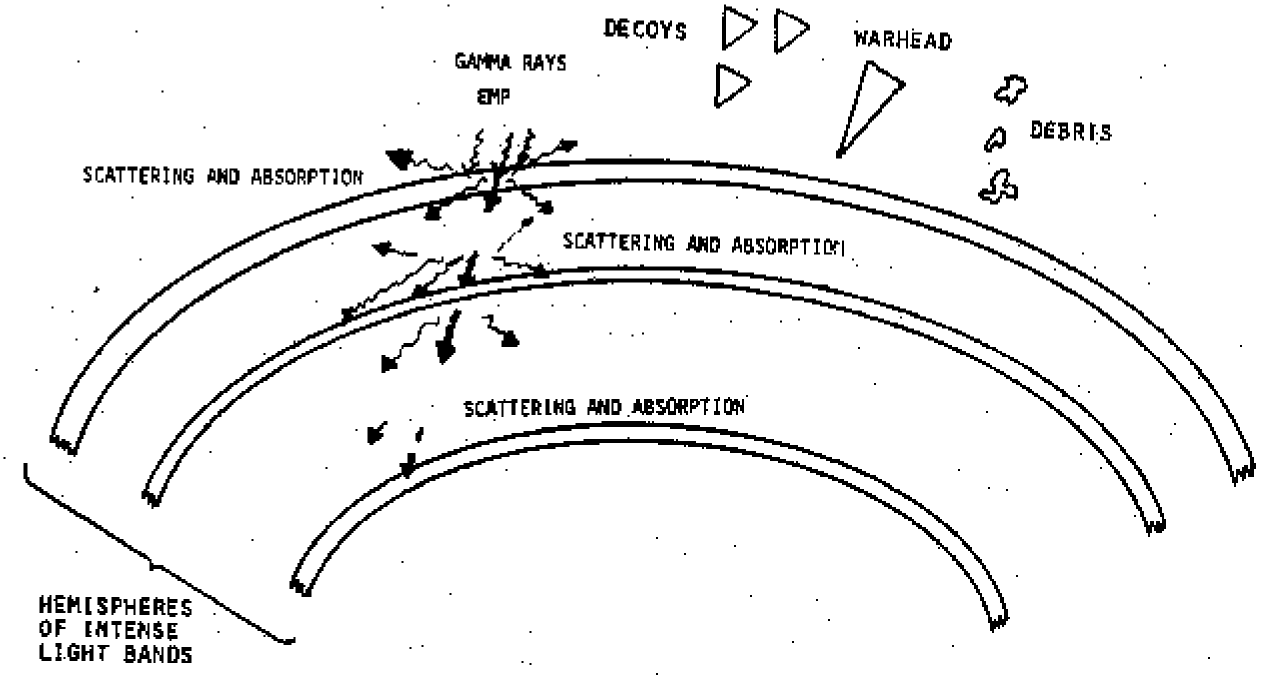
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

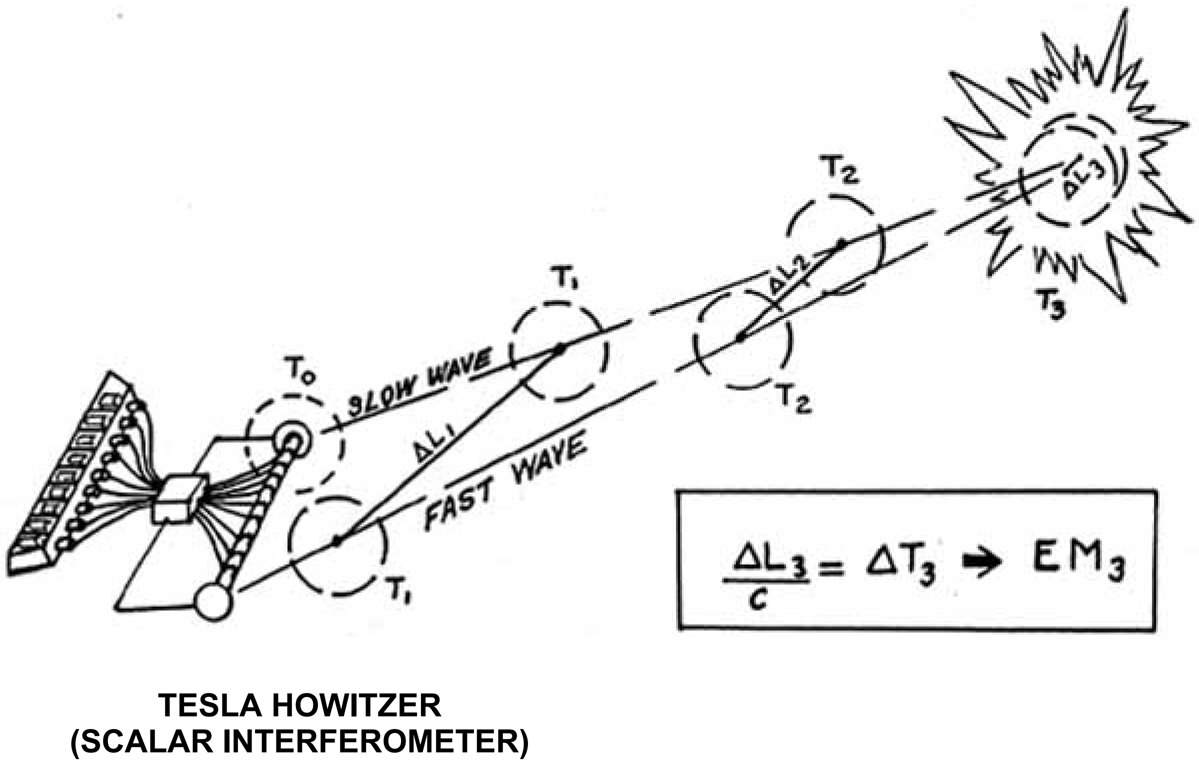

Figure 10. "Nuclear" Flashes off the Coast of Africa

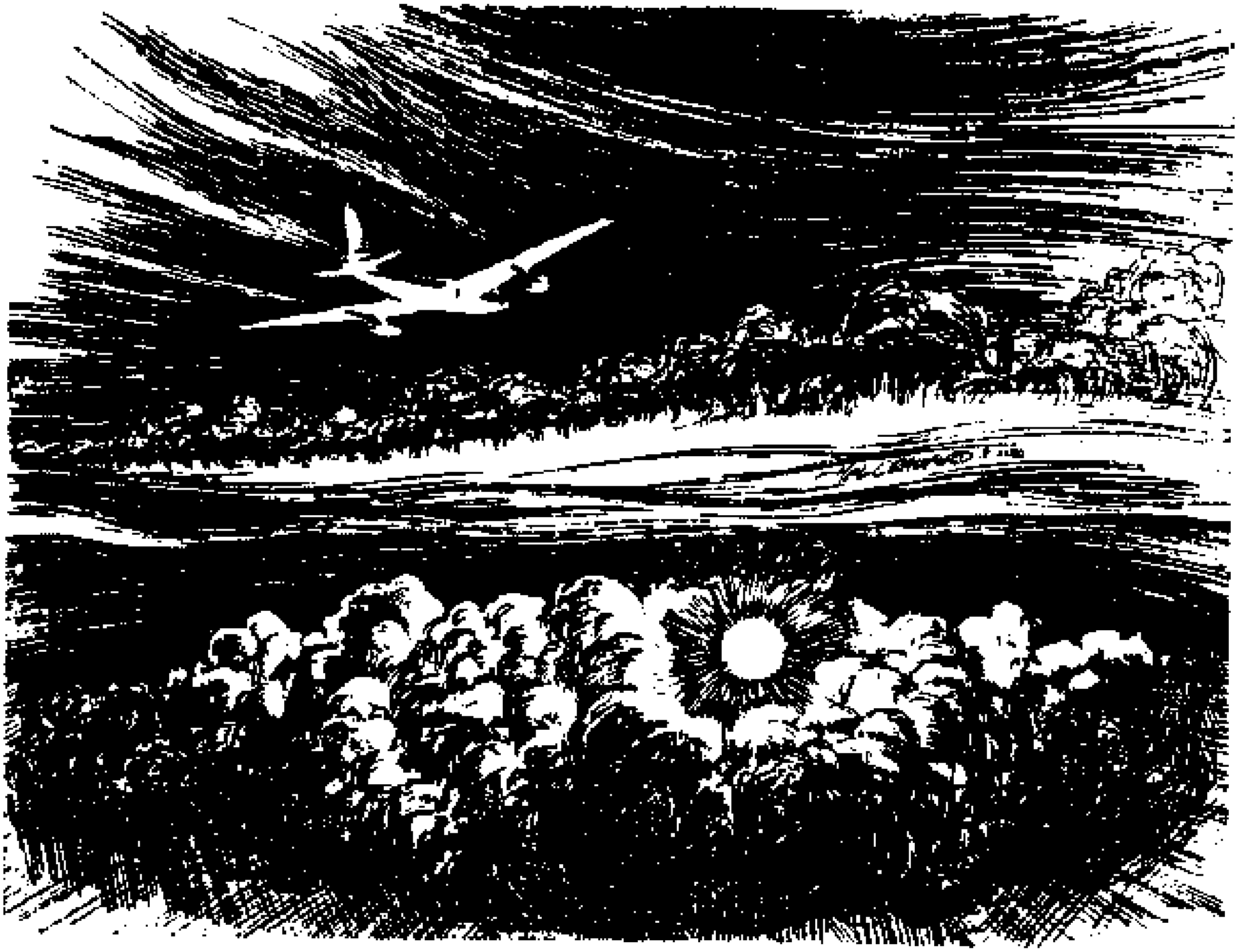

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

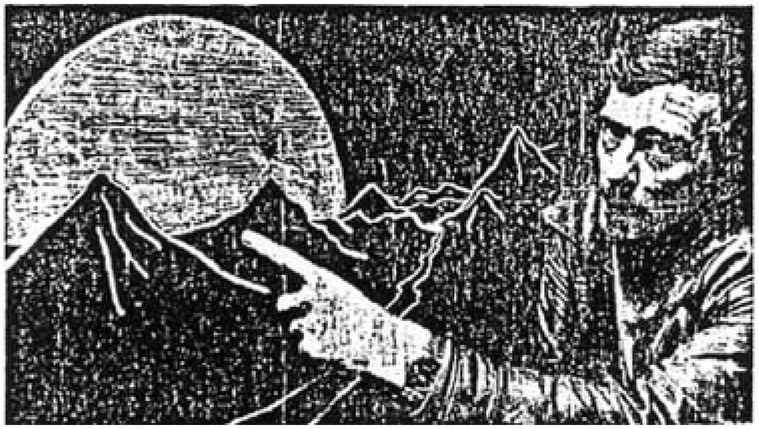

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

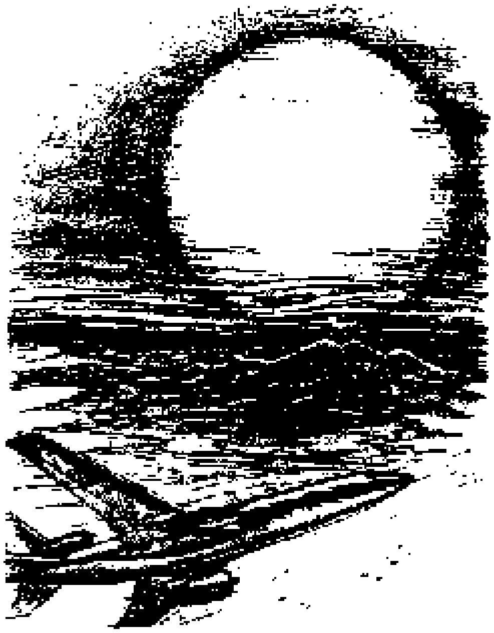

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

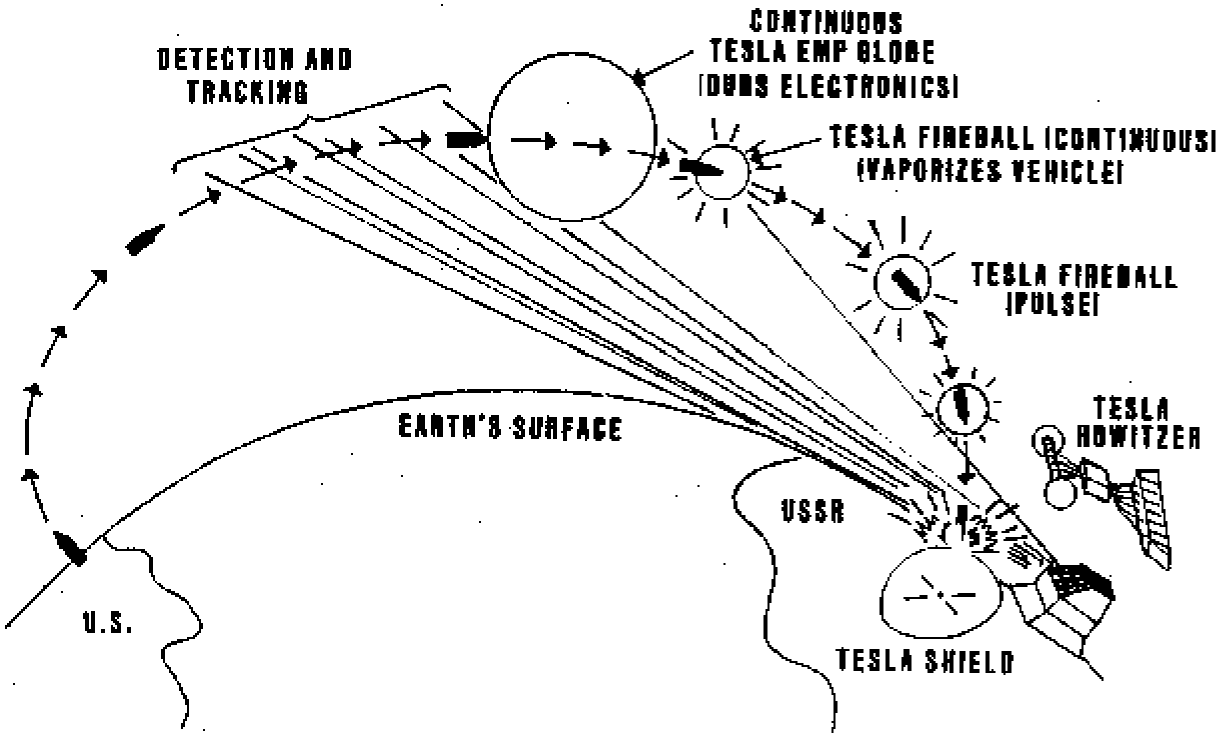
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

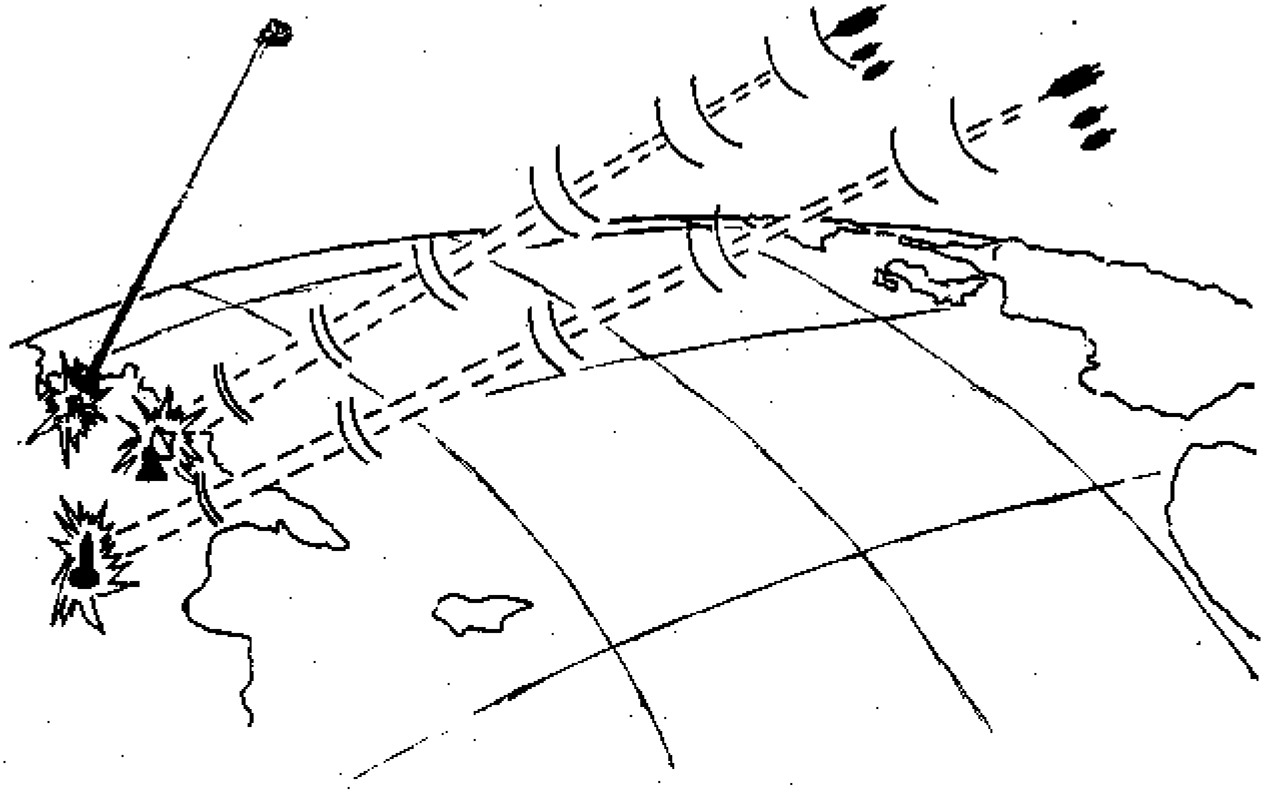

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **> PAIR-COUPLED SCALARS**
  - **> VELOCITY-LIMITED TO C**
  - **> CARRIERS OF T**

- **PHOTON INTERACTION**
  - **> IS UBIQUITOUS**
  - **> PRODUCES $\dot{T}$ = C**
  - **> YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **> NOT VELOCITY-LIMITED**
  - **> YIELD HIGHER ORDER REALITIES**

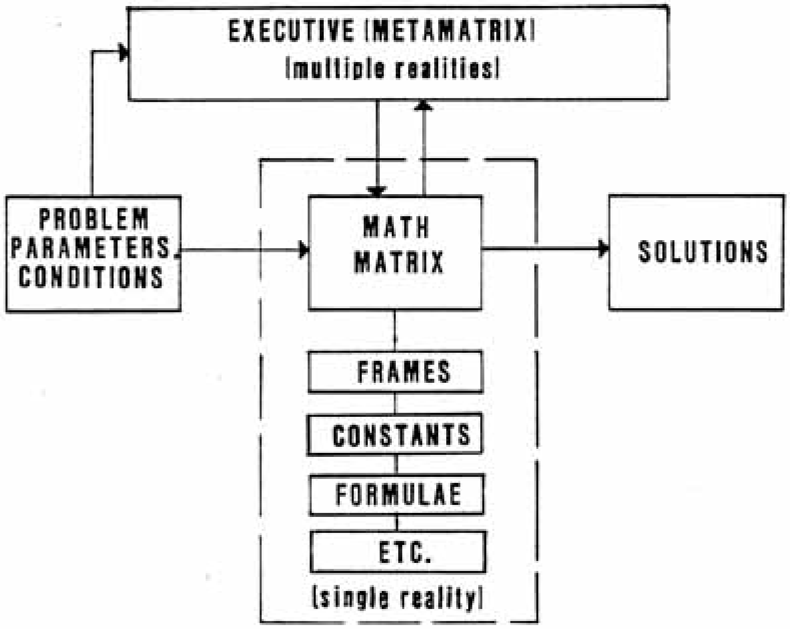

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

The Tom Bearden Website

34 Flaws in Classical EM Theory

Note: If the scientific community had done their job, this Website would probably not be necessary.

---

1. SLIDE 1
2. SLIDE 2
3. SLIDE 3
4. SLIDE 4
5. SLIDE 5

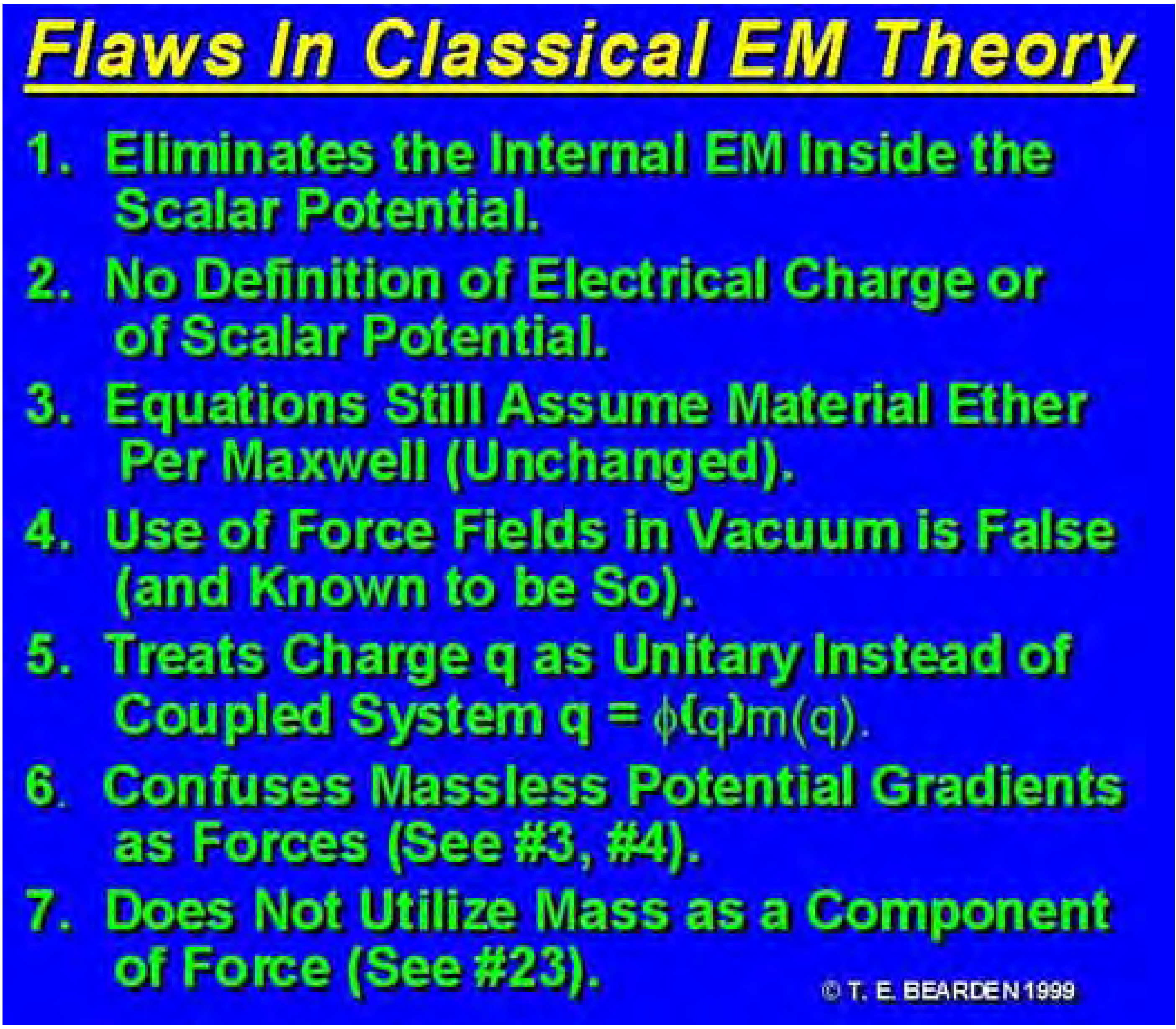

Flaws In Classical EM Theory

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi(q)m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

© T. E. BEARDEN 1999

Next Slide

Return to Index

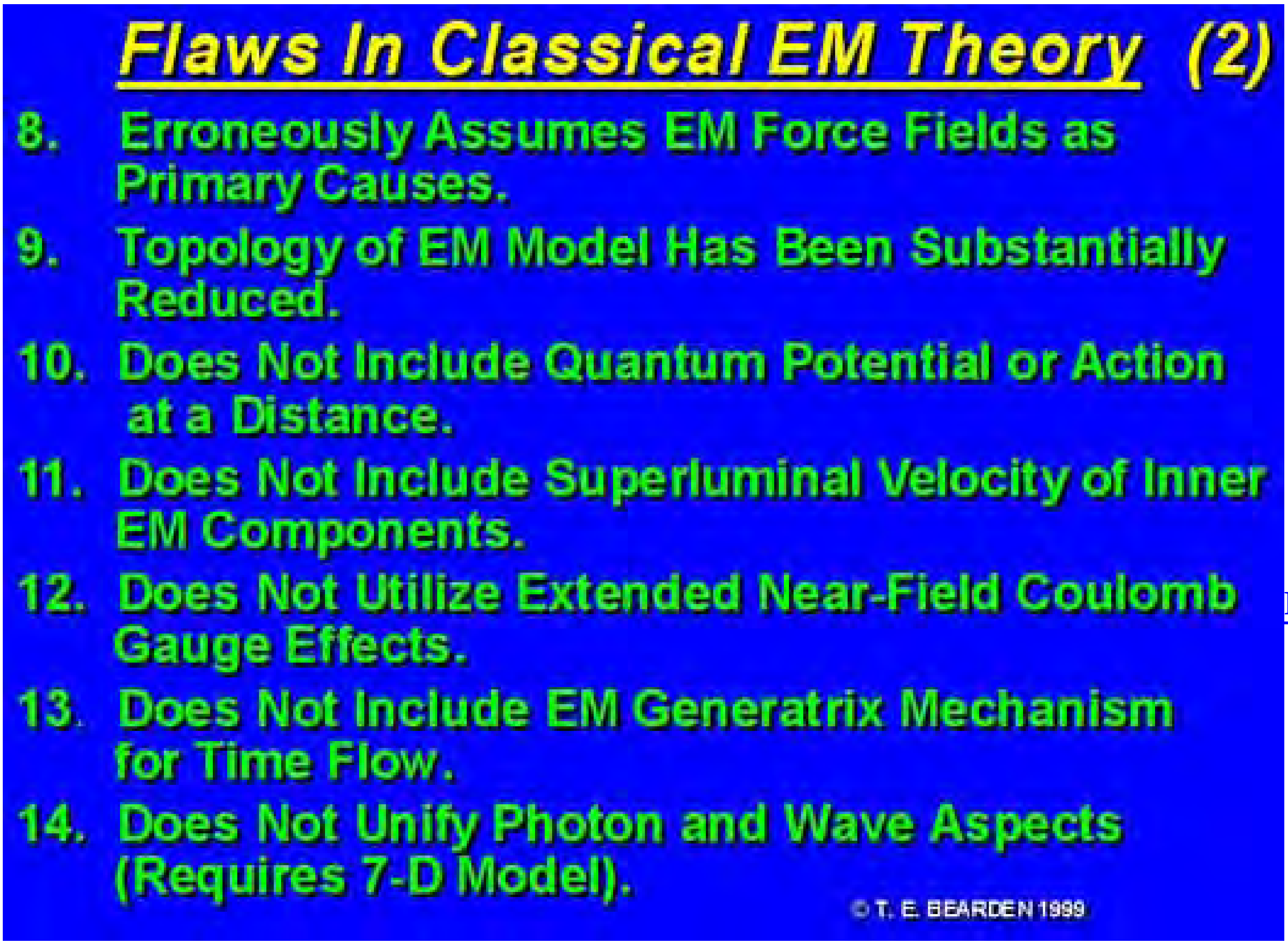

Flaws In Classical EM Theory (2)

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

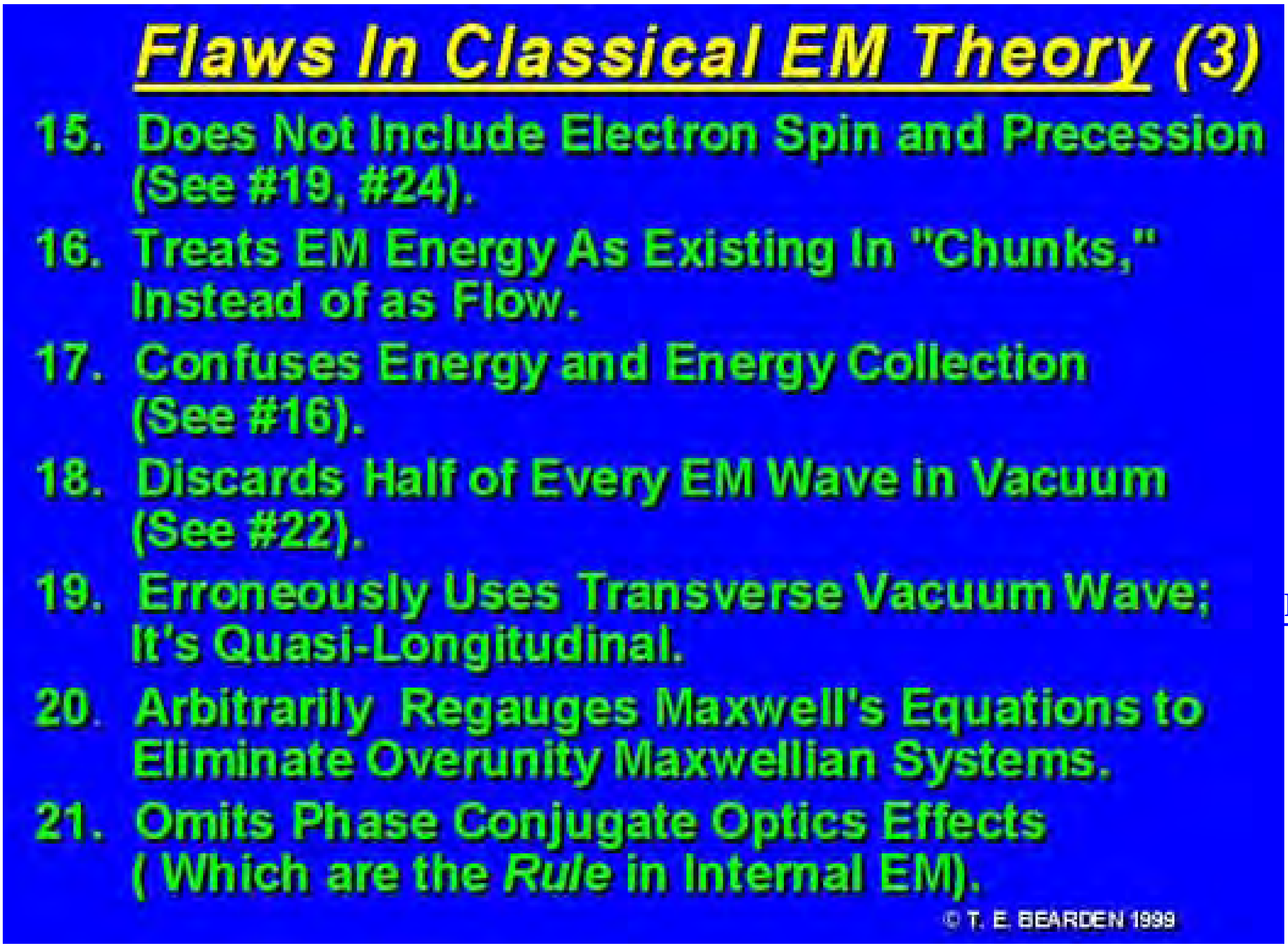

Flaws In Classical EM Theory (3)

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Quasi-Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity Maxwellian Systems.
21. Omits Phase Conjugate Optics Effects (Which are the *Rule* in Internal EM).

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

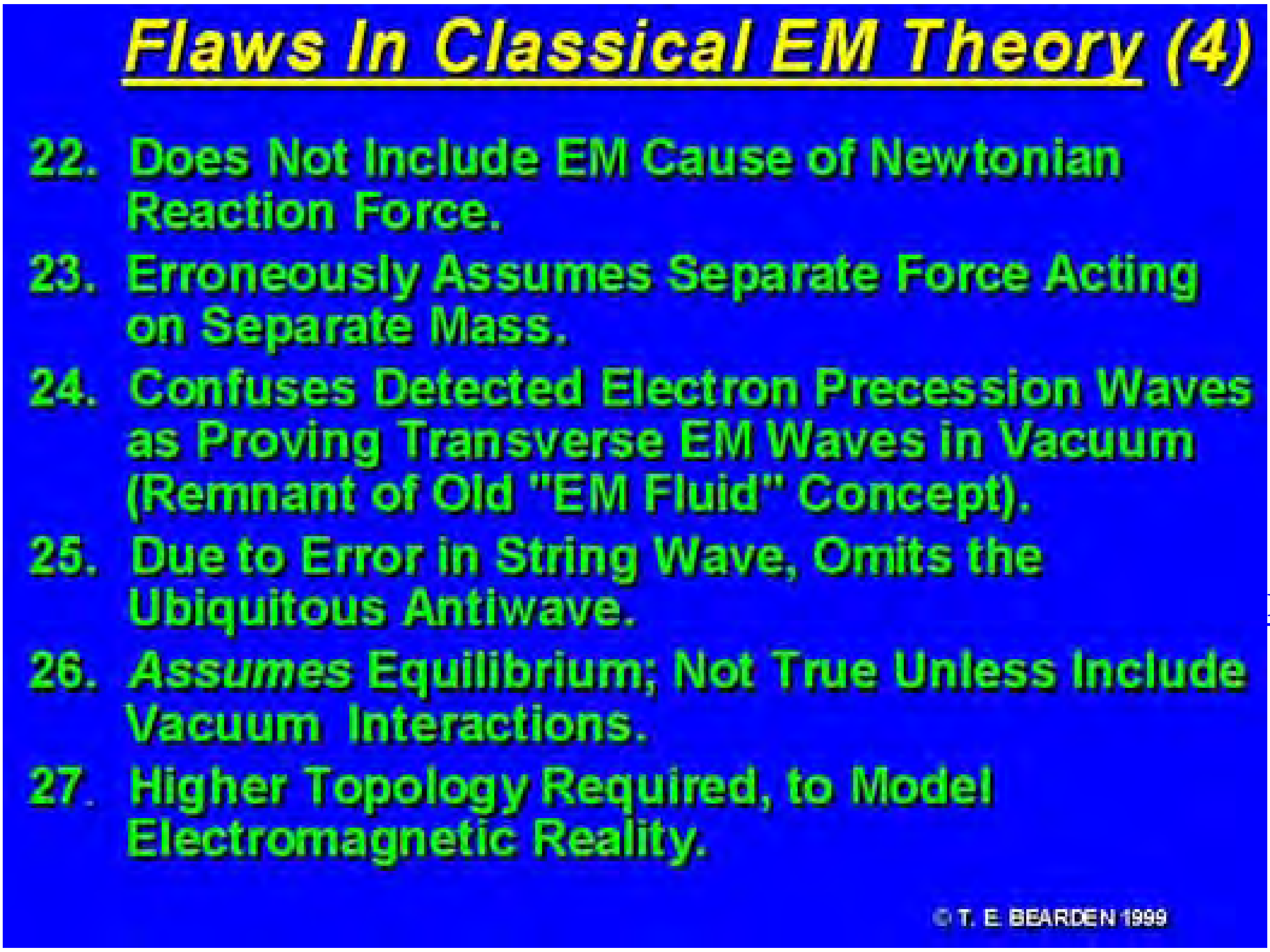

Flaws In Classical EM Theory (4)

22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.

© T. E. BEARDEN 1999

Next Slide

Return to Index

Previous Slide

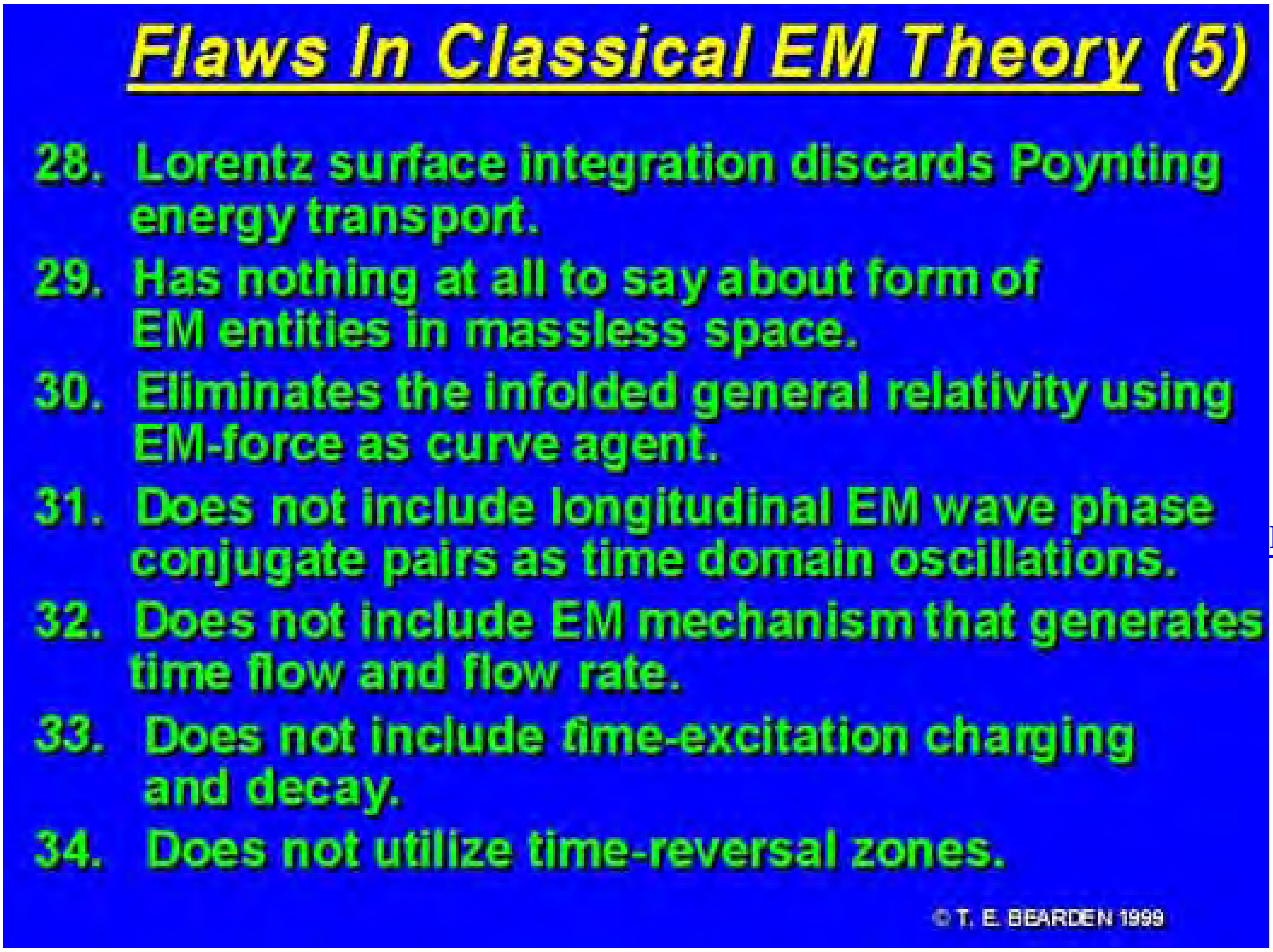

Flaws In Classical EM Theory (5)

28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM wave phase conjugate pairs as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.
33. Does not include time-excitation charging and decay.
34. Does not utilize time-reversal zones.

© T. E. BEARDEN 1999

Return to Index

Previous Slide

The Tom Bearden Website

Mind Control and EM Wave Polarization Transductions - Part I, II & III

Reprinted from "Explore" magazine, 1999

Part I

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9
- Page 10

Part II

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8

Part III

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9

- The U.S. approach to mind control

- "Voices in my head"
  Flanagan's neurophone - used to give instructions to diplomats during negotiations, etc.

Mind Control and EM Wave Polarization Transductions, Part 1

©Copyright 1999 by T.E. Bearden, USA

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental pro-cedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Abstract

For some time we have been repeatedly queried about the technical mechanisms and unusual electrodynamics of advanced mind control research, both in the West and abroad. Calling full attention to the special note above, in this paper we present a high-level over-view of the novel electromagnetic nature of mind operations, mind and body coupling, and *intent*•the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented. We give two specific examples of lethal foreign military tests in 1997 of advanced mind control weapons against two military pilots over the central U.S. Some dangers and potential benefits of the emerging mind control technology are pointed out

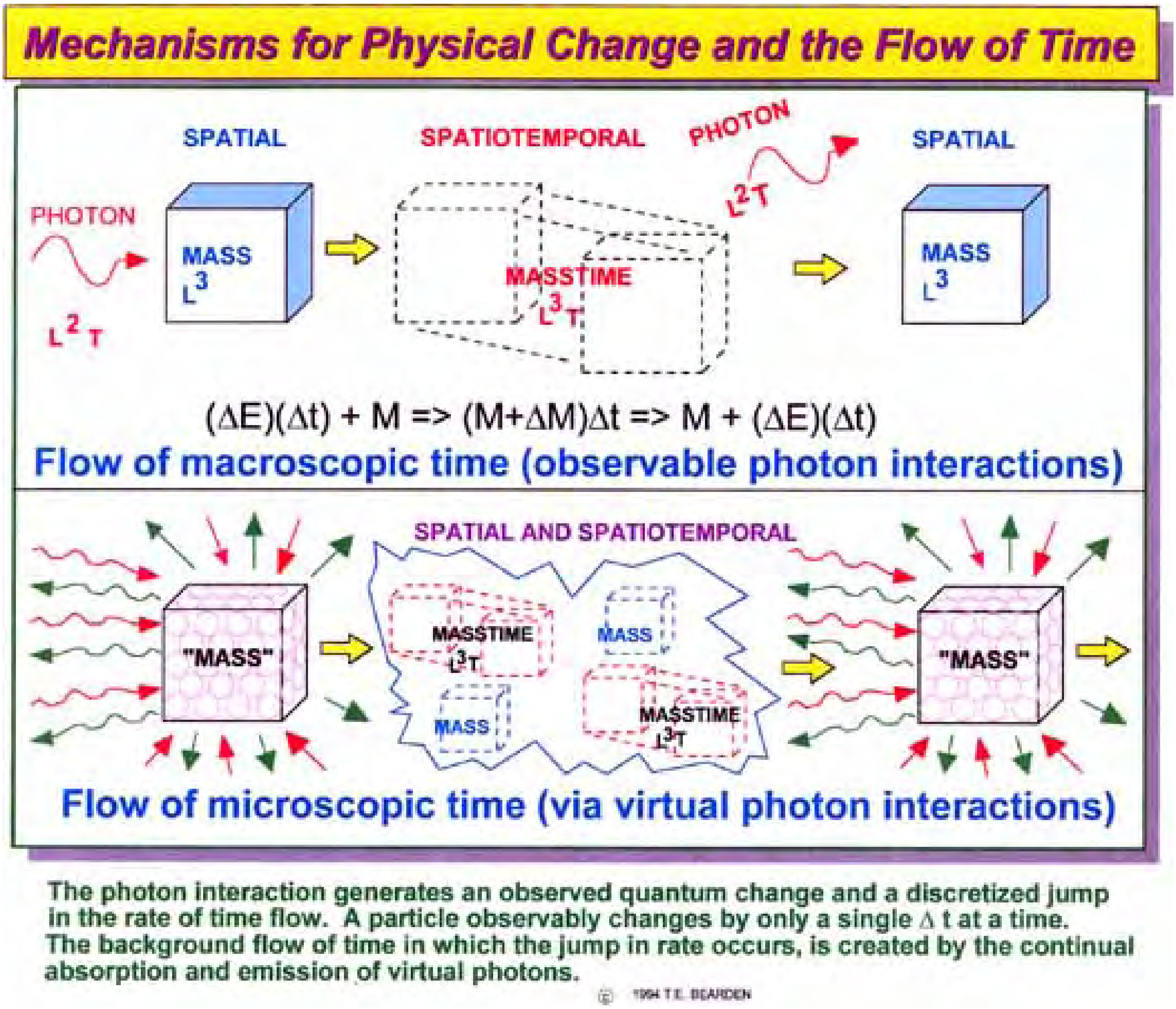
Mechanisms for Physical Change and the Flow of Time
SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
L^2T
MASS
L^3
$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Figure 1. Mechanism that generates a mass's flow through time.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Introduction and Background

In quantum field theory, there are four polarizations of photons.[1] Using 4-space and the z-direction as the direction of propagation, we have x- and y- polarizations where the 3-spatial energy of the photon is oscillating laterally, in the x- or y- direction. These are *transverse polarized* photons, as is any combination of the two. The third polarization is along the z-direction, which is a longitudinal polarization. In other words, the 3-spatial energy of the longitudinal photon cannot oscillate in the x- or y- direction, and so it is oscillating to-and-fro along the line of motion, z-. The fourth polarization occurs when the energy is frozen in all three spatial dimensions x-, y-, and z- and it cannot oscillate in any of those directions. In that case, the photon oscillates its energy in the t- direction, providing the t-polarized photon. We will later discuss how time is actually highly condensed energy.

In Minkowski 4-space, body operations are space-like, and are so treated in conventional materialistic physics. Mind and mind-operations are time-like, not space-like, even though they are totally electromagnetic in nature.[2]

As is well-known, all observation in physics is considered 3-spatial..[3] Mass is a 3-spatial concept, and we detect changes to mass (as, e.g., in the shift of electrons in the circuits of electrical instruments). So one may model the common physical observation mechanism as a *time-differentiating process* which *a priori* discards time-change and retains 3-spatial energy change.[4] That is, Minkowski reality is modeled in the fundamental units of L^3t. Physical observation (via the transverse photon interaction) is the process given by applying the operator $\partial/\partial t$ to L^3t, yielding an L^3 output. Hence mind and mind operations are excluded by the usual physics instruments and observation, which simply exclude the time domain in their outputs and do not "measure" it.

For this reason, physicists have erroneously considered mind to be "metaphysical" and nonreal. Indeed, most physicists to one extent or the other are materialists, and consider the "mind" to be nothing but the operations ongoing in a "meat computer." Nonetheless, in the time-domain the time is absolutely real, and it is completely electrodynamic in nature. There is no metaphysics involved, and the temporal domain—along with mind and mind operations—is simply an erroneously neglected area of physics. In present physics, the notion of mind is comfortably disposed of by imposing the use of the "observer" concept, with out ever specifying that the observer has a consciousness and a mind. Indeed, "observation" is only about what that "observer" perceives. Obviously, one has a dramatically crippled physics if one eliminates that nonobservable called "time." Similarly, one also has a dramatically crippled physics when one eliminates the mechanisms and physics ongoing in those time-like and dynamic "things" such as mind, that occupy time and function in it.

From this viewpoint, Western physics adheres to its 3-spatial measurement foundation only by ignoring the transduction of time-polarized EM changes into detectable longitudinal and transverse EM wave changes. In this aspect, present Western physics is severely self-crippled.

Time is totally electromagnetic and energetic in nature. The flow of time is not a separate external river on which a mass floats along like a boat drifting down the current of a great flowing river. Instead, the flow of time is generated directly on every mass by its total set of photon interactions, both virtual and ob servable. We have previously presented the exact mechanism for the flow of time. *[Figure 1]*

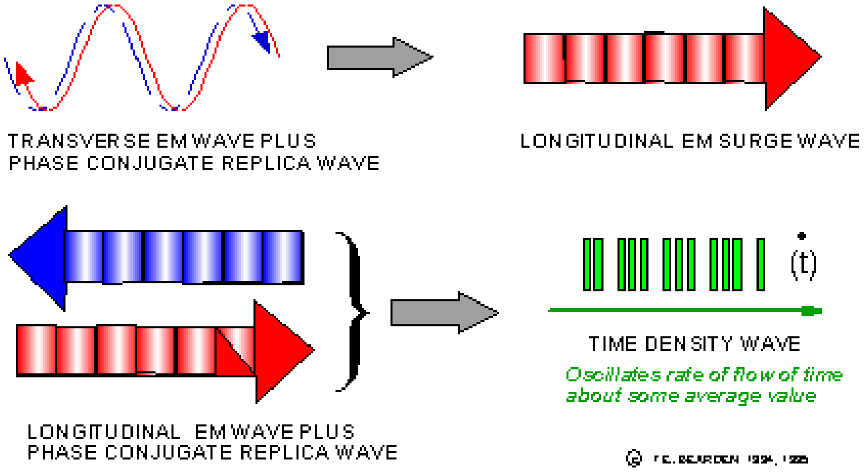
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
$(\dot{t})$
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994, 1995

Figure 2. Transduction of EM wave type by successive phase conjugate pairing.

Further, a photon is comprised of angular momentum, therefore of (energy)x(time). It transports both energy and time, not just energy alone. When absorbed by a mass, not only does it "energy-excitation" charge the mass, but also it "time-excitation" charges it, converting the former "mass" to "mass time". Rigorously it is *masstime* that emits a photon, not *mass*. So a mass moves through time in little spurts, by the continual macroscopic addition and subtraction of little Dt components. Further, in being driven through time, mass is continually altered to masstime (a quite different critter from mass), to mass, to masstime, to mass, etc.

In any masstime state, a myriad of tiny virtual photon interactions made of very tiny (DE)(Dt)'s interact with that same mass during that particular larger Dt of the masstime state (m+Dm)(Dt). Hence mass in its alternate masstime state has that state internally structured by its surrounding environment's interactions with it. The masstime state is internally structured energetically in its (DE) component, and also internally structured temporally in its (Dt) component.

A standard charged fundamental particle such as an electron, e.g., is not necessarily identical with another, when the internal structuring of its masstime state is considered. Further, by two papers by Whittaker,[5] interferometry of masstime states with either energy reactions or temporal reactions can yield observable effects and changes due to these neglected "hidden variables" in the electron's masstime alternative states. To engineer the mind and its operations directly, one must perform electrodynamic engineering in the *time* domain, not in the 3-space EM energy density domain. The direct engineering of time-like mind and mind operations—in all levels and all aspects—requires the use of *time-polarized photons and time-polarized EM waves*. This is the rarest form of electrodynamics, almost untouched by Western physicists.

One can either painfully produce such time-polarized photons and EM waves and directly irradiate a target mass with specific assemblies of them, or one can force the mass itself to iteratively transduce ordinary transverse EM waves first into longitudinally-polarized EM waves and then into time-polarized EM waves. Both will be discussed briefly.

In the West, it appears that the present author's discovery of mechanisms for producing time-polarized (scalar) EM waves [see *Figure 2*] and for transducing between wave polarizations has no precedent. [See *Figure 3*]. While scalar (time-polarized) photons are known in the literature,[1] the creation and use of time-polarized EM waves does not seem to appear in the Western physics literature.[6]

Wave transduction or *polarization transductions* are terms used by the author for the process of transforming an EM wave's particular polarization into another polarization type. Such transduction to other than transverse polarization forms is little known in the West. It appears essentially haphazardly in some experiments, usually without any recognition by the experimenters themselves. Significant transduction in experiments will also produce "strange" and unexplained anomalies in the instruments being used for experimental measurements.[7]

So most probably Western mind control researchers have not recognized the methodology and mechanisms[8] for transducing one type of EM wave polarization directly into another.[9] They continue to seek the "mind" in 3-space and hence in the brain, rather than in the time domain. On the other hand, it appears

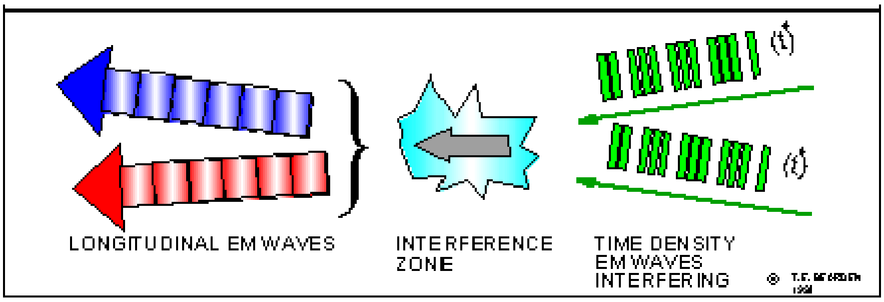

Figure 3. Interference processes for transducing higher EM wave polarizations into lower polarization states.

EXPLORE! VOLUME 9, NUMBER 2, 1999

that the Russian KGB energetics[10] weapons scientists—particularly those in psychoenergetics[11]—have known and used methods of transducing one wave polarization into another, for at least two decades. It follows that those same scientists have very probably developed mind engineering and mind control via novel time-polarized EM wave means and a dramatically extended electrodynamics of the time-like mind operations.

In the present paper we briefly develop the wave transductions and the basis for mind engineering, including the specialized use of ordinary transverse wave (TW) EM spectra to force internal EM wave transductions and time-domain operations inside irradiated bodies, cells, and tissues. Through the mind-body coupling mechanism, these transductions of transverse EM waves can operate upon the mind and its deepest operations as well as upon the body, every cell, and every part of every cell.

Russian Mind Control Uses Higher EM Wave Polarizations

It is apparent that the KGB psychoenergetics weapons scientists know and use the full extended EM wave polarization range. As a postulation, one would expect our own weapons scientists to know and utilize the orthodox transverse wave (TW) EM for similar research and experimentation. However, in the West scientists are just beginning to realize the importance of a very general (and weaker) type of longitudinal EM wave (LW) polarization, as evidenced by the appearance of dozens of papers on "undistorted progressive waves" (UPWs, which are essentially imperfect LW s with TW residues remaining). The Los Alamos National Laboratory web site has a number of such papers—particularly by Rodrigues[12] and Lu—available for free downloading. UPWs have very interesting characteristics. If they were perfect longitudinal EM waves, they would have infinite energy and infinite speed. Since only imperfect UPWs can be physically made, their speed can vary from slower than the speed of light to faster than the speed of light. Their energy can also vary over a great range.

Another characteristic of UPWs that are reasonably good longitudinal EM waves, is that they pass readily through a large depth of water and mass, including through the ocean and the earth with only small interactions and losses. Yet by interfering two such beams of "high quality" UPWs at a great distance, then in the interference zone ordinary EM energy will rise directly out of local spacetime potential, as essentially shown by Whittaker[13] nearly a century ago. Russian weapon research facilities have weaponized these effects[14] for nearly 50 years, under rigid KGB control and operation. Nonetheless, even if using only ordinary TW waves, Western mind control researchers may get some fairly good results, brute-force-like, by using gross correlates between just the input irradiating TWs and the exhibited behavioral responses of the individual. The net input-output correlations can be determined, even though not taking into account the actual wave polarization transduction mechanisms ongoing inside the irradiated dielectric (or brain, or mind, etc.). However, Western researchers appear to have no knowledge of the exact mechanisms by means of which coherent time-like mind operations of a biological organism couple to the organism's 3-spatial body to provide the mental control loop. They also appear unaware of how the coherent 3-spatial behavior responses of the body couple back to the time-like mind to provide it with a sensory feedback of the body's responses.

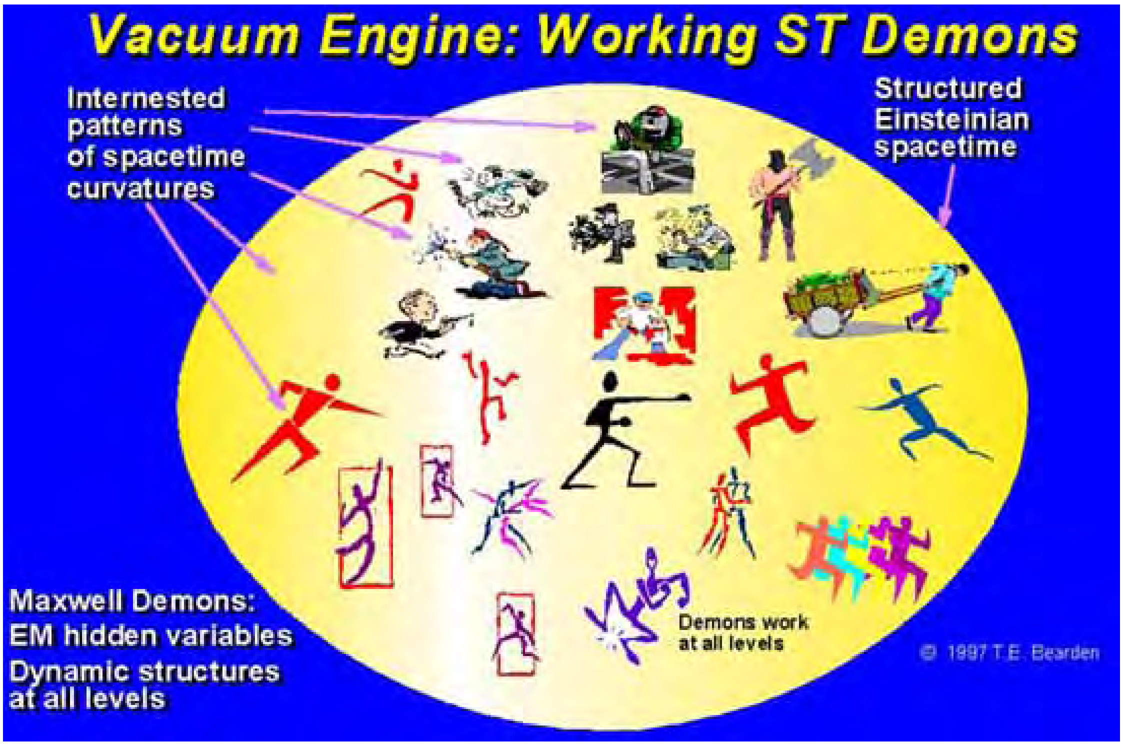

Vacuum Engine: Working ST Demons

Internested patterns of spacetime curvatures

Structured Einsteinian spacetime

Maxwell Demons: EM hidden variables

Dynamic structures at all levels

Demons work at all levels

© 1997 T.E. Bearden

Figure 4. A spacetime curvature engine has myriads of small ST curvature components which act upon mass at all levels.

EXPLORE! VOLUME 9, NUMBER 2, 1999

General Relativity Aspects

Both these coupling mechanisms can be taken directly from general relativity, if one puts one's mind to it. However, unless the wave polarization transformations are known, it is difficult or impossible to apply the ordinary GR *directly*, since relativists usually concentrate upon spacetime curvature by altering 3-space energy density rather than by altering time-polarized "time-energy" density.

In applied general relativity, spacetime itself is an active medium. One speaks of a change in spacetime as a "spacetime curvature". Any curvature of spacetime in a local region directly performs continuous work upon any mass embedded in that region. Vice versa, any mass (or other concentration of energy) in a local spacetime continuously acts upon that spacetime to "curve" it.

So if one wishes to perform actions upon mass in a distant region, one may generate local curvatures of spacetime in that region, and these curvature "engines" will directly act upon the mass at all levels. *[Figure 4]* This is very different from energy propagation through space from one point to another. Now the "ordinary EM energy and actions" arise from every spatiotemporal point within the mass, at every level, and move upwards (from inside to outside). So we speak of specific forms of "templates" of nested spacetime curvatures formed and utilized to engineer mass "from inside out" as *spacetime curvature engines or vacuum engines*.

Use of vacuum (spacetime curvature) engineering[15] is a far more powerful form of engineering than is provided by energy propagation through space. As an example, it is easy to alter the quarks in a nucleon, using spacetime curvature engines and time-charging decay. Indeed, we have developed the mechanisms for cold fusion and the electronuclear interaction (formation of new nuclides) at feeble energy. An Invention Disclosure[16] on this work has been filed with the U.S. Patent Office, and formal patent applications are in preparation. Some information on these mechanisms and principles has been released[17].

Time as Dense EM Energy and a Strong Spacetime Curvature Agent

The advantage of using the time-polarized "time-energy" for spacetime curvature is that time is ordinary energy compressed by a factor of at least c^2—which, in the MKS system of units, is some 9¥10^{16}. Thus use of time-polarized EM photons and waves as ST curvature agents gives an amplification of 9 ¥10^{16} over the use of transverse-polarized EM waves for that purpose. In turn, the use of the strong EM force in ordinary TW waves as an agent of ST curvature is already a nominal 10^{40} times as strong as is the weak G force used as the agent of ST curvature.

The bottom line is this: For spacetime curvature effects, the use of the t-polarized domain provides amplification of some 9¥10^{56} greater than the weak G-force ST curvature agent usually considered in general relativity[18].

Western Science Remains Largely Materialistic

Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system.[8] This serious self-limitation exists because in the body we *measure* only weak TW EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.

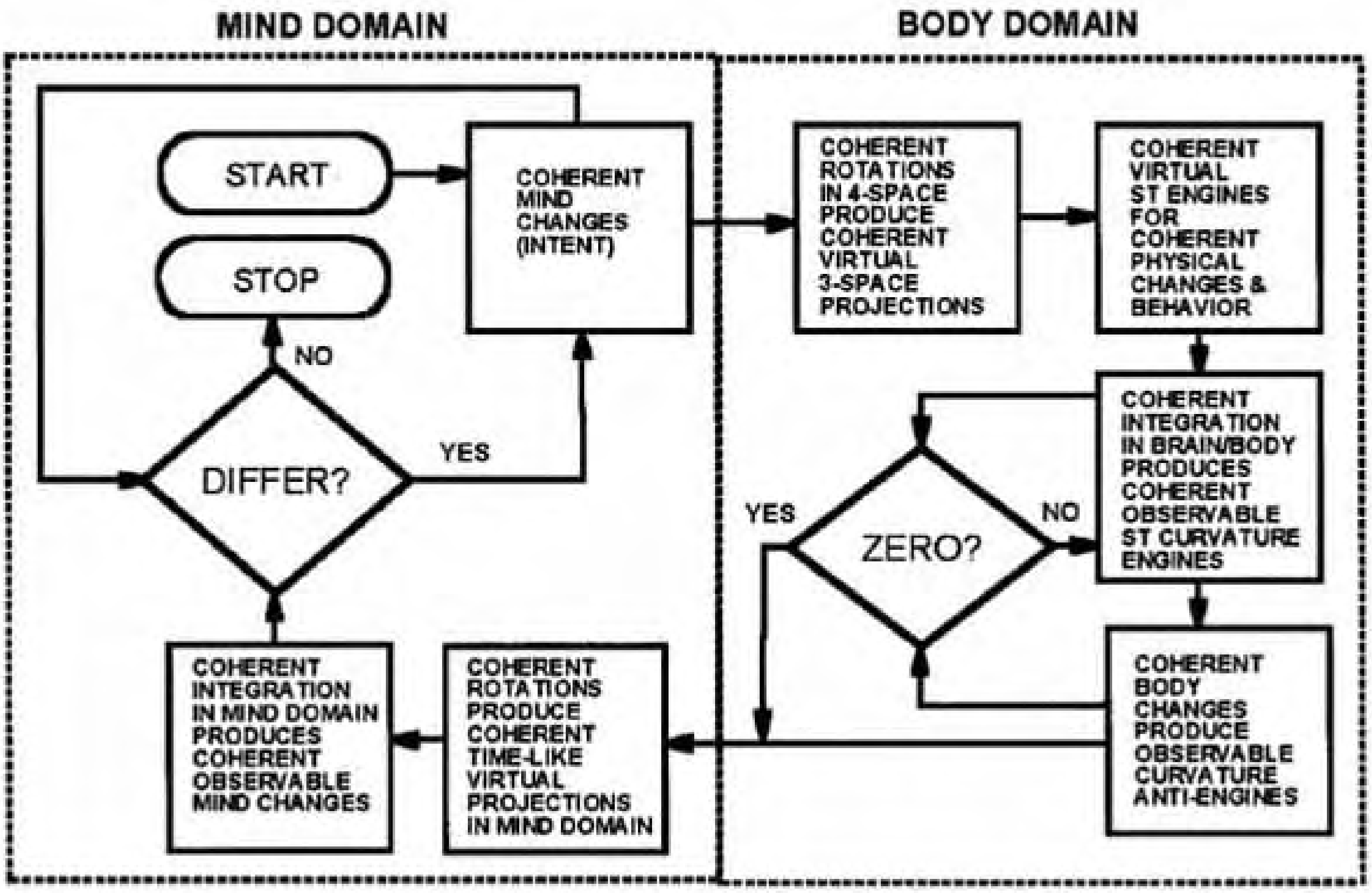
MIND DOMAIN
BODY DOMAIN
START
STOP
COHERENT MIND CHANGES (INTENT)
NO
DIFFER?
YES
COHERENT INTEGRATION IN MIND DOMAIN PRODUCES COHERENT OBSERVABLE MIND CHANGES
COHERENT ROTATIONS PRODUCE COHERENT TIME-LIKE VIRTUAL PROJECTIONS IN MIND DOMAIN
COHERENT ROTATIONS IN 4-SPACE PRODUCE COHERENT VIRTUAL 3-SPACE PROJECTIONS
COHERENT VIRTUAL ST ENGINES FOR COHERENT PHYSICAL CHANGES & BEHAVIOR
COHERENT INTEGRATION IN BRAIN/BODY PRODUCES COHERENT OBSERVABLE ST CURVATURE ENGINES
YES
NO
ZERO?
COHERENT BODY CHANGES PRODUCE OBSERVABLE CURVATURE ANTI-ENGINES

Figure 5. Mind-to body coupling and body-to-mind coupling.

EXPLORE! VOLUME , NUMBER 2, 1999

With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW functions exist in the body and in all of nature in general. Presently only highly theoretical quantum field theorists seem to even be aware of the existence of scalar (t-polarized) photons, and even they are unaware of t-polarized EM waves. Certainly our scientists do not measure t-polarized EM waves in and around the body and in nature, nor do they make them in the laboratory. Apparently they have not studied such waves and their interactions with matter—living and inert—at all.

But it's even worse. In ignoring the time component transported by photons and EM waves, science has also erroneously omitted half of the excitation charging and excitation decay processes whenever a mass interacts with photons and EM waves. More on that later.

Brute Forcing Time Functions versus Fine Control Methods

Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired. They also do not hesitate to use LWs and t-polarized waves overtly. So in blunt terms, Western compared to Russian mind control research is probably like comparing an automobile body shop with a fine surgical ward. Having long ago worked out all those "exact correlates" to the internal "information content of the field" is where the KGB scientists are at least 20 years ahead of the West.

The reason one can get mind and behavior results with the TW electromagnetics, while bypassing the real mechanism which uses t-polarized waves and photons, is due to (1) the very peculiar nature of the EM emission from a dielectric, and (2) what can be done by re-radiating that dielectric with its emitted spectrum, deliberately and very carefully altered in selected parts. We will return to that important feature later.

Polarization and Observability

As we stated, there are four photon polarizations and therefore there must be correspondingly four EM wave polarizations[1] The first three polarizations are the x-, y-, and z- spatial polarizations. The x- and y- polarizations are transverse polarizations and the z-polarization is a longitudinal polarization along the direction of propagation (along the z-axis, by standard notation). Simply put, we may visualize the transverse polarizations as rather like the wiggling of a fish's tail from side to side as the fish moves forward, or a whale's flukes up and down as the whale moves forward, or some combination thereof. We may visualize the longitudinal polarization as a sort of "repetitive accordion effect" along the line of motion of the wave. Usually the z- polarization is neglected in EM wave theory, although in recent years physicists have "rediscovered" longitudinal EM waves and are now intensely researching the use and characteristics of such waves.[12]

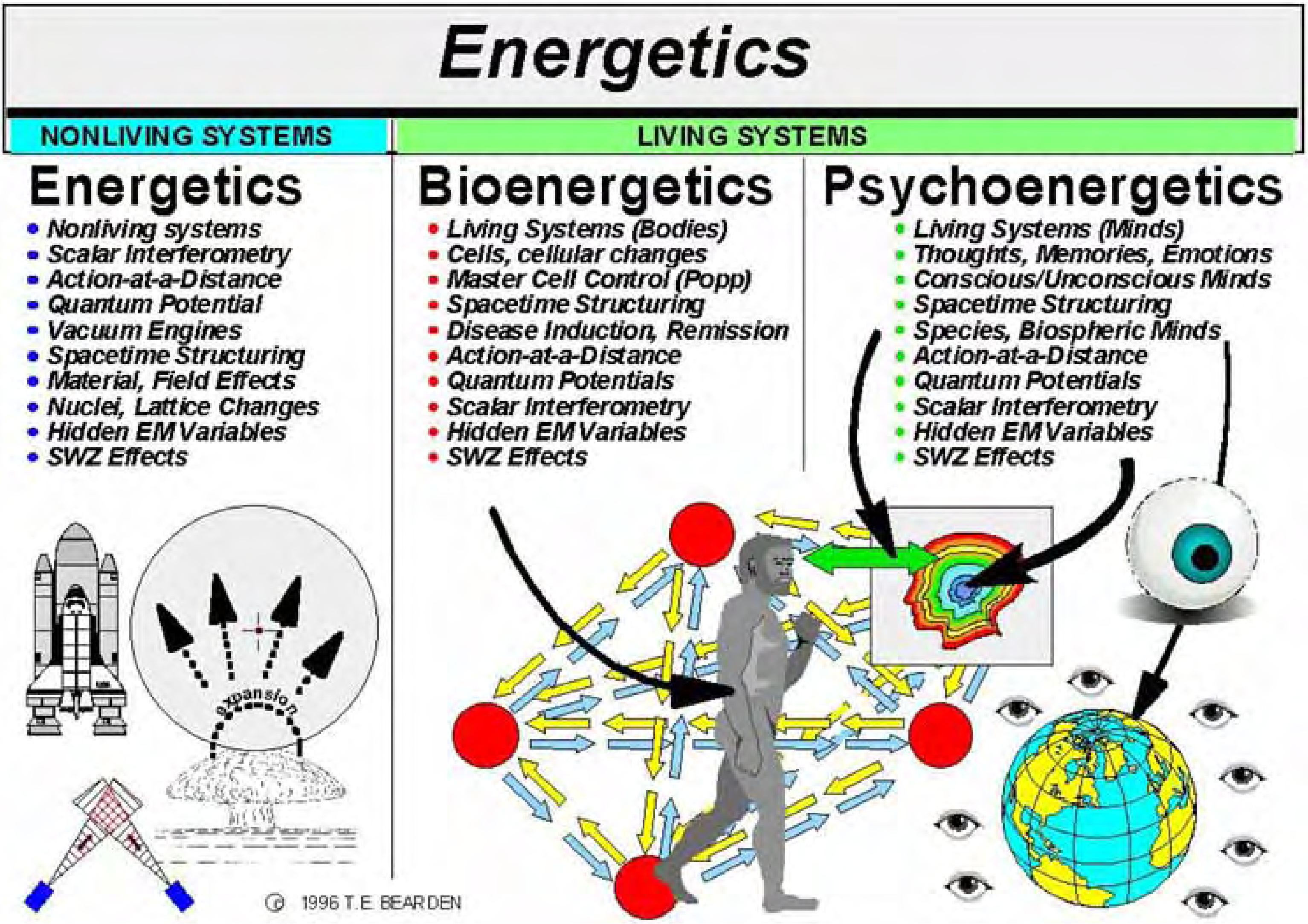

Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Figure 6. The three divisions of Soviet energetics and their characteristics.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Unfortunately, in quantum field theory there has been a tendency to regard the t-polarized or "scalar" photon—where the local rate of time is oscillated—as unobservable. Oscillation of the local rate of time produces powerful oscillations of local space-time curvatures, due to the extreme energy density of time. Individually, the t-polarized photon tends to be unobservable. However, a *coherent sequential group* of such t-polarized photons, each individually in the virtual state with respect to the external observer, may simply integrate in its interaction with mass over a short time period into observable change because of the coherent integration of virtual spacetime curvatures into a larger, observable spacetime curvature.

Rotations and Projections in 4-Space

Relativistically, any velocity change in 4-space is a rotation. Any such rotation in the time-domain also creates a tiny projection component into 3-space. Any such rotation in the 3-spatial domain also creates a tiny projection component into the time domain.

Hence a sufficient series of coherent time-domain (mental) changes produces a coherent series of virtual changes in 3-space (the body), thereby coherently integrating into an observable energy change in the body. A sufficient series of coherent 3-spatial energy changes produces a coherent series of virtual changes in the time domain which coherent integrate into an observable time-domain change.

Solution to the Age-Old Philosophical Problem of Intent

This is in fact the solution to the age-old problem of *intent*, or how the nonmaterial (i.e., non-3-spatial and non-observable) mind induces a 3-spatial, observable energy change upon the 3-spatial body. It is also the solution to the problem of *awareness*; i.e., how the mind is aware of itself[19] and of the responses of the body. Together the two form a closed loop coupling of the mind and body. The time delay in the loop together with memory recall for comparison, creates the sense of "persistence" of self in time. This also creates the sense of being a "separate, closed being" (i.e., of separate persistence in time—in the living entity). The sense of "separation of self from an external world" is created by comparing those body-to-mind sensory feedbacks which are not correlated to the mind's previous feed-forward intent.

Time As Energy and Why It Is Very Dense Energy

In addition to the three spatial polarizations of photons and EM waves, there is a very, very useful t-polarization along the time axis. In this polarization, the 3-spatial energy is not oscillating at all. Instead, the time or time-energy is oscillating. Time can be taken to be energy compressed by at least c^2, so it has at least the same energy density as mass. In other words, one second is 9×10^{16} joules of time-energy (energy compressed into time). The t-polarized photon or EM wave is called the scalar photon or scalar EM wave, respectively.

To demonstrate why time can be regarded as energy, we need only point out that the choice of fundamental units in any physics model is totally arbitrary. We usually choose these units to ease the mathematical manipulation, ease our understanding, and simplify the ability to visualize or "grasp" the physics. However, if one wishes, one can build all of physics from a single fundamental unit—e.g., energy. In that case, one readily sees that time is a function of energy and only energy. Hence it is perfectly proper to regard time as energy, and to seek out what form the function takes in our normal system of units (say, the MKS system).

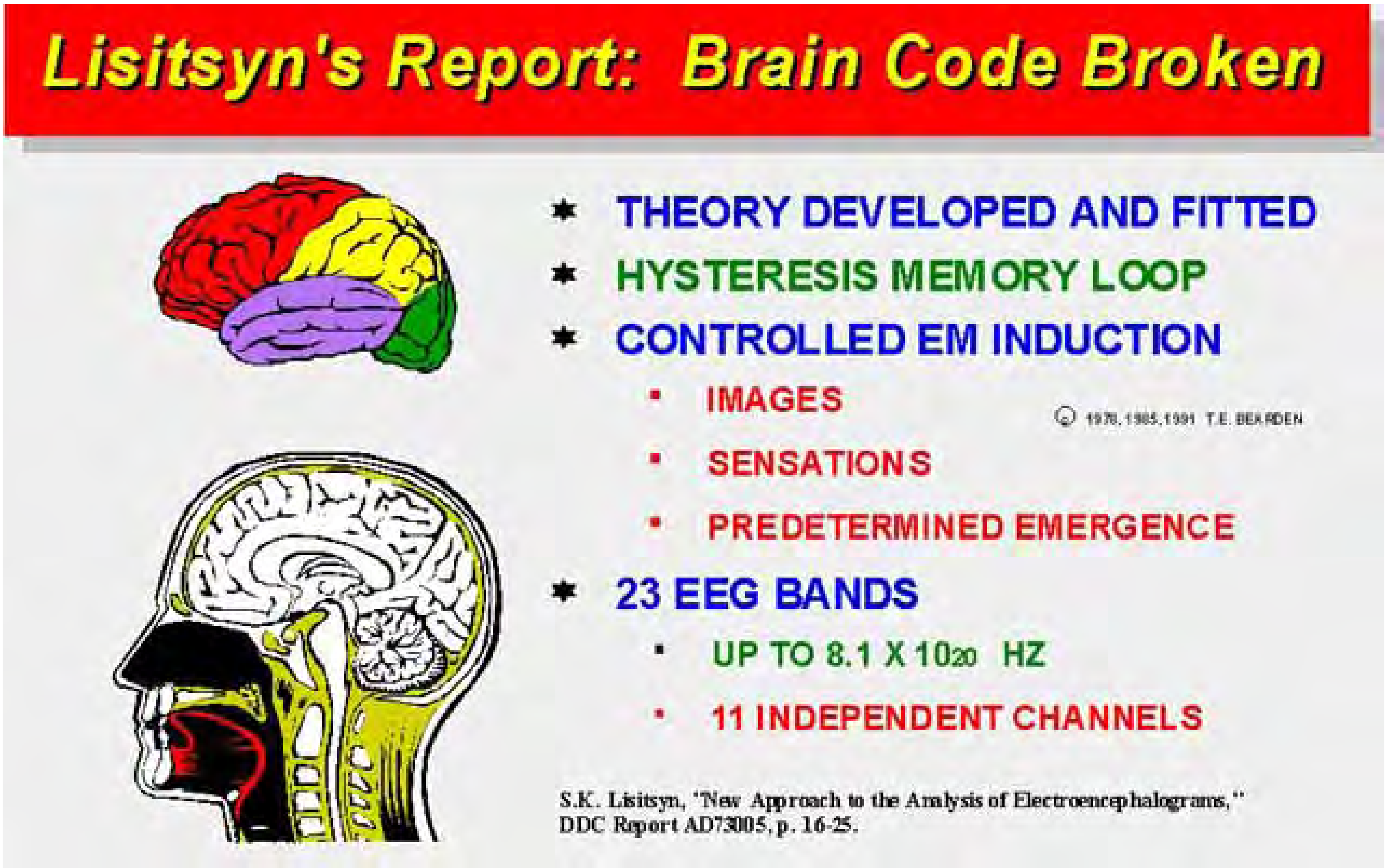

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Figure 7. In the 1960s, Lisitsyn revealed Russian scientists had deciphered the human brain code.

EXPLORE! VOLUME 9, NUMBER 2, 1999

In that system, it appears that any change in time Δt can be taken as always a function of a corresponding change in energy ΔE, where $\Delta t \leq (\Delta E) \div c^2$ [1] from which it follows that $\Delta E \geq (\Delta t) \times c^2$ [2] For convenience, we take the special case where $\Delta t = (\Delta E) \div c^2$ [3] $\Delta E = (\Delta t) \times c^2$ [4].

Observation as Used in Physics Is Spatial

We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain, and thus "lost" by the stripping away of time in the ordinary observation process. Physical observation is essentially a time-differentiating mechanism applied to a 4-spatial change, or in terms of MKS fundamental units L and t, observation θ is $\theta = \partial/\partial t(L^3t) = L^3$ [5].

Thus, as is well-known in quantum mechanics, physical observation is 3-spatial, and time is not a physical observable, even in theory. Since mind is time-like, it follows that mind is not a physical observable either, since discarding the time dimension also discards the mind. In short, one may also take physical operation as the mechanism that separates mind and body. To observe is to separate. We point out, however, that merely "separating and discarding" the time domain (as in physical observation of a single change) does not eliminate it, nor does it eliminate the single time-like mind change that may be involved in an intentional volition. The time domain certainly remains, even though only the 3-spatial intersection of the 4-spatial Minkowski change is given by physical mechanism and the resulting "outputs". So a series of coherently integrated "mental intent" changes introduced into the human body's overall servo-mechanism provides the continuing input. From there, ordinary physics will generate the resulting actions induced in the body by that servomechanism and its amplifying mechanisms.

Two Coupling Mechanisms Make a Closed Loop

This is the "mind-to-body" coupling mechanism. [See *Figure 5*] It is the mechanism whereby the mind is coupled to the body, and whereby mental intent is able to induce a series of physical inputs into the body servomechanism.

So the body's servomechanism then generates the responses of the body (including everything from chemistry to electrical changes to muscular movements, etc.). These responses are changes in 3-space.

The conscious mind is a *serial processor*, though extremely rapid. It produces the series of coherent intent inputs for volitional behavior.

The unconscious mind is a *massively parallel processor*.[20] It continually produces the vast series of coherent "unconscious intent inputs" for control of all the deeper processes in the body, beyond usual conscious awareness.

However, as the body moves or changes in 3-space, each resulting quantum change in body 3-space is also a slight rotation out of 3-space and toward the time axis. Hence it induces—in the "virtual state" in the time domain—a precisely correlated projection. The body's responses are in general coherent, so a coherent series of virtual state changes in the time-domain (in observation. So we may say *that the time-change remains in the virtual state, with respect to attempted physical observation of a single Minkowski 4-space change.*

Rotation, Coherent Integration, and Intent

Previously we pointed out that any change in a 4-space entity may be regarded as a rotation away from the "trajectory" of the entity. Hence each and every t-polarized change creates a very small projection into 3-space by its rotation slightly away from the time axis. All mind changes in the time-domain actually produce virtual 3-spatial projections in 3-space (in the physical domain). We define "intent" as the continued production of successively coherent mental changes, producing coherent virtual changes in the 3-space body system, with coherent integration of those 3-space virtual changes into observable 3-space energy inputs into the 3-space body system. Successively coherent mind-changes will produce successively coherent 3-space virtual changes. In short, coherent mind-changes will produce coherent integration of those 3-space virtual projections into an observable 3-space change. This is the creation of an ordinary 3-spatial energy change induced upon the 3-space body. In short, this is the mechanism whereby time-like mental intent is able to produce a series of coherent observable quantum changes in the physical body (as in the brain and nervous system).

The Body as a Servomechanism

From the standpoint of control theory, we may consider the body system to be a complex servomechanism system comprised of many subordinate servo systems with feedback and feedforward the mind realm) are created successively. In short, again we have coherent integration, this time in the mind or time domain. This produces "observable" changes in the time-mind domain, which are coherent with the body's 3-spatial changes actually performed. Thus the mind receives feedback directly from the physical movements and changes of the body. This is the manner in which the body is coupled back to the mind.[21]

By comparing the "intent" behavioral move that was "fed forward" into the body, with the return "response" move analog that was "fed back" from the body to the mind, the mind is able to determine errors and differences, and originate additional correctional commands.

Thus the entire mind-body loop is a closed-circuit system of feedforward and feedback, together with corrections. It also has multiple levels of such, infolded in the larger volitional levels. See again *Figure 4*.

This solves the age-old problem of the mechanisms for the mind-body coupling, intent, volition, conscious and unconscious functioning, sense of the external world, sense of the internal world, sense of "being in" the external world, etc.[22]

All This Is Included in Russian Psychoenergetics

This is the highly summarized basis for *psychoenergetics*, the KGB's division of energetics that deals with the mind and body coupling and functions, and direct engineering of (1) the mind-body coupling and (2) the mind operations directly. [See *Figure 6*] The KGB intent, of course, has always been to exploit this science for the degradation, killing, and control of human beings, including all humanity.

The Russians know full well that, if you looping.

Given the input from intent, ordinary servo theory will take it from there. Once a servo has an input, servomechanism theory describes the response of the

produce and utilize scalar EM photons and t-polarized EM waves, you can directly affect and engineer mind and mind operations at any and all levels. Western researchers, who know nothing of how to make

EXPLORE! VOLUME 9, NUMBER 2, 1999

time-polarized photons and time-polarized EM waves, do not yet know that. Consequently the Russians have developed a highly secret science of *directly* engineering the mind and its operations, including thought, images, perceptions, feelings, emotions, memory, and mind processing. Indeed, Lisitsyn[23] wrote quite specifically of this capability in the 1960s. [See *Figure 7*]

Western clandestine mind control researchers are apparently still slowly and painfully fitting TW EM irradiation correlates to induced or resulting mental and physical behavior responses. They seem unaware of the actual wave transductions occurring inside the body and mind, but are unwittingly inducing those transductions in hidden fashion anyway and in "brute-force input-response fitting" models.

These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design—for example—of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.

This fine research article will be continued in Part II with Russian Methodology, Waves and Wave Transduction, The Cellular Control

spacetime that is highly active. The mind is rooted in the time-domain and projects from that domain into 3-space. The body is rooted in the 3-space domain and projects into the time-domain. To "change" or "function" in one domain is automatically to function in the other. EM and gravitational phenomena are still modeled separately by human scientists, but not by nature.

3. In quantum mechanics, time is not an observable, but merely a parameter.

4. It is convenient to consider mass as a 3-spatial form of condensed energy.

5. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics, " Mathematische Annalen, Vol. 57, 1903, p. 333-355;— "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372.

6. A NERAC document search yielded some 16 good references on t-polarized photons, but not a single reference on t-polarized EM waves.

7. E.g., such previously inexplicable instrument anomalies have accompanied some excellent and rigorous electrolyte experiments at U. S. Navy research facilities at China Lake. For a description, see Melvin H Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", Proc. ICCF - 7 (International Conference on Cold Fusion—7, Vancouver, BC, Canada, Apr. 1998, p. 101. For a brief explanation of the anomalies, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, " Explore!, 8(6), 1998, p. 7-16;— "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

8. Many U. S. researchers and journalists—and even many scientists—have wrestled with the problems of mind, intent, and mind-body coupling. Enumeration of those efforts would itself require an entire book. For an introduction into that domain, the reader is referred to the very determined exposé by

System, and other matters of interest. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses. Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar

Cheryl Welsh in her very timely book, The 1950's Discovery of the Code of the Brain, May 1988, published on the Internet. Simply employ any Net search engine and search on the name Cheryl Welsh. A connection to the site and to Welsh's book will immediately be found Welsh's important compendium is highly recommended, giving the reader a "birds eye view", so to speak, of how Western scientists think and proceed regarding the subject of mind control by electromagnetic means. Unfortunately all journalists so far researching the area have not been aware that Western electrodynamics itself is seriously flawed, and that the great Russian advances in mind engineering and mind control are a result of their secret but complete revision and correction of Western electrodynamics to provide the basis for energetics. Until Western scientists revise their own decrepit old electrodynamics fouling the textbooks, they will never catch up to the Russian mind control developments springing from intense development programs that did that revision in the late 1940s and early 1950s.

9. An Invention Disclosure on these processes and enhancement embodiments has been filed with the U. S. Patent Office. Formal patent applications are also in preparation.

10. Energetics is a unified science of an extended electrodynamics possessing a hidden infolded general relativity inside the potentials, fields, and waves. It deliberately employs higher polarizations (longitudinal and tempic) of photons and EM waves to engineer action at a distance, use of specific patterns of spacetime curvature created at a distance, and direct engineering of either inert or living bodies. It also encompasses the direct engineering of time-like mind and mind functions. It is divided into three divisions, depending upon the nature of the target. Targeted against inert materials, structures, fields, and waves it is called "energetics. "Against living bodies, their fields, and biochemistry, it is called "bioenergetics." Against the mind and mind functions, it is called "psychoenergetics".

11. Psychoenergetics is that branch of energetics used to directly affect and engineer

electromagnetics processes.

REFERENCES

---

1. E.g., see Lewis H Ryder, Quantum Field Theory, Cambridge University Press, 2nd Edition, 1996, p. 147+. For an even deeper discussion of photon polarization, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e. electrostatic) potential.

2. Nature is unified and not separated into the "separate" disciplines our science has prescribed. Thus the mind is also totally electrogravitational. The mind works not just with EM forms in an inert space, but also with spacetime curvatures in a

the mind, memory, perception, thought, emotion, sensation, personality, and other mind functions of a living biological organism. It is usable against a single personal conscious (serial processor) mind and a single personal unconscious (massively parallel) processor mind In theory it is also usable against lungs collective unconscious mind and its functions, as well as the deeper unconscious for all species on Earth (i.e., against Gaia). We are speaking of physics, not metaphysics.

12. W A. Rodrigues Jr.. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature," Foundations of Physics, 27(3), 1997, p. 435-5•08 A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory website.

13. E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.

14. E.g., see T. E. Bearden, Fer-de-Lance, Tesla Book Co., 1986;—AIDS: Biological Warfare, Tesla Book Co., 1988;—Soviet Weather Engineering Over North America, 1-hour videotape, 1985,—Gravitobiology; A New Biophysics, Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991,—Energetics of Free Energy Systems and Vacuum Engine

EXPLORE! VOLUME 9, NUMBER 2, 1999

Therapies, Tara Publishing, Internet node www.tarapublishing.com/books, July 1997; — Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ pages furnished to selected government agencies and key Congressional committee leaders.

15. E.g., see T.D. Lee, Particle Physics and Introduction to Field Theory, Harwood; New York, 1981. On P.383, Nobelist Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) is not utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25; Outlook; Possibility of Vacuum Engineering, " p. 824-828. With wave transductions and use of higher polarization EM waves, there are far better and more powerful ways to do vacuum (spacetime curvature) engineering.

16 T.E. Bearden, "Formation and Use of Time-Reversal Zones, EM Wave Transduction, Time-Density (Scalar) EM Excitation and Decay, and Spacetime curvature Engines to Alter Matter and Convert Time Into Energy," Invention Disclosure Document #446522, Oct. 26, 1998.

17. T.E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, "Explore!, 8(6), 1998, p. 7-16; — "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

18. We have used the enormous energy density of time in our approach to a totally new kind of electronuclear interaction at feeble spatial energy. By use of a time-reversal zone (TRZ) in a fluid; the free H+ ions—which are free protons—can be induced to attract each other so tightly that the edge of the nuclear force region is entered. In that case the protons form a quasi-nucleus. When the TRZ decays back to a time-forward zone (TRZ), this new quasi-nucleus decays by changing quarks in one or more nucleons, thereby producing new nuclides. We point out that means "sense of existing in time".

20. The "unconscious" is totally conscious, just multiply so. The conscious mind can see only "a single slide in the slide projector at a time, " so to speak. When it "looks" at the ongoings and outputs of the unconscious, it "sees" millions of slides in the slide projector simultaneously. Hence it cannot distinguish "one and one only". So it just sees nothing, where by "nothing" we mean "no single distinguishable thing." It therefore is unable to consciously perceive the actions of the unconscious. When the unconscious wishes to communicate with the conscious (because of unresolved conflicts, etc.), it must produce large groupings integrated into "symbolic scenarios". Hence the symbolism of the unconscious, and the basis for therapeutic psychology and psychiatry. A "symbol", after all, is just a thing which can have many different meanings, all folded into one.

21. The mind continually has activities of its own, each of which produces coherent changes in the body to a much smaller degree. The interweaving of these "smaller self-changes" infolded inside the 3-space major intent changes also produces slight body changes, which in turn "project back" into the mind. This huge group of "infolded self-loops" produces the sense of "my body" being "volitionally directed or occupied" simultaneously with the "greater intent" loops which represent "my intentional actions with my body". Further; the body servo continually receives a myriad of inputs from its sensory systems, both from changes inside the body and from incoming changes from the external universe. This interweaving produces the awareness of "being in the external world" by having (i) changes whose feedback do not match mind-intent changes, and (ii) changes whose feedback does match mind-intent changes. This of course applies to both the conscious and unconscious levels. The purpose here is to demonstrate that there exists an actual physics of the living being, the mind as well as the body, and the coupling and interfiinctioning of mind and body.

this is actually a very high energy reaction; however; since the time-energy is used and is very, very dense, the amount of 3-spatial energy that must be used is very small. If this new mechanism holds, it represents a revolution in particle physics.

19. "Self-awareness" refers to be continually "refreshed" by time-delayed coherent feed-back inputs from the body to the mind; where these inputs compare to the recalled memory of the minds inputs to the body. In short, the time-delay coherence creates the sense of "existing in time". Awareness usually

22. The methodology also easily extends to include deeper interlays for the collective human unconscious mind and for Gaia—the collective unconscious mind level for all the species on earth.

23. S. K Lisitsyn, "New Approach to the Analysis of Electroencephalograms, " in Problems of Bionics— (Selected Articles), Defense Documentation Center Report AD 73005, date unknown, p. 16-25. Available through the National Technical Information System (NTIS).

Mind Control and EM Wave Polarization Transductions, Part 2

©Copyright 1999 by T.E. Bearden, USA

This article continues on with T.E. Bearden's article, Part 1 of which was published in Volume 9, #2 of Explore! for the Professional. Because we feel that Col. Bearden's caution below is so very important we are repeating its publication in all parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

A Strategic Threat Completely Unrecognized by the West

The Russian methodology leads directly to rather mind-bending, extended capabilities. KGB psychoenergetics scientists seem fixed on an eventual goal of directly engineering Jung's collective human unconscious, thereby converting the entire human species into a sort of "ant" species. In short, they seek to engineer what they regard as "perfect communism and perfect order" directly in the entire species. *Figure 8* indicates the direct connection of every human's conscious and unconscious mind to the collective species unconscious and to Gaia, the collective unconscious for all species on Earth. Since time is internally structured and layered, it follows that time-like mind levels are also internested and layered.

This appears to be a significant strategic threat, *well along in its development*, that our scientific and intelligence communities have completely failed to recognize.

The direct engineering of the entire human collective species is a doable.[1] It is going to be done, whether we like it or not. In that race (which may eventually determine whether or not we retain any personal freedom at all), the "democratic" nations are woefully behind. In fact, the West has not even realized there is such a new kind of "arms race" for dictatorship and control of the "collective species mind" and we have already almost lost it. The West has not yet even developed the overall theory of psychoenergetics.

Waves and Wave Transductions

A pure longitudinal EM wave (LW) would have infinite energy and infinite speed. In practice, one can only approach this perfect LW wave, and one always has transverse residues remaining. Hence the "imperfect" longitudinal EM waves we can make in suitable equipment and techniques, are known as "undistorted progressive waves" (UPWs). The UPWs may have velocity far greater than the speed of light, or less than the speed of light. So the speed v of UPWs can vary as

$$0 < v < \infty \qquad [6]$$

The Russians also know that, when a nearly-pure longitudinal EM wave interacts with matter (it usually doesn't interact very much, and has a low reaction cross section), the mass immediately adds the phase conjugate replica to the interacting LW portion, regardless of frequency, thereby transducing the input LW interactions into scalar (time-polarized) EM interactions.

One transduces from TW to LW to Scalar EM via iterative phase conjugation. One transduces from scalar EM to LW to TW via iterative interferometry. It took the present author 20 years to discover this simple thing; we now have filed an invention disclosure on the process and embodiments for performing it, and will be following with formal patent applications. Let TW = transverse EM wave, LW = longitudinal EM wave, and TDW = time-density EM wave (another name I use for t-polarized or scalar EM waves, since the density of time-energy is what is varying). Let (I) = interfering with, PCR = phase conjugate replica wave with a coupled set of parentheses enclosing the type of wave. The rules for wave type transduction are:

$$\text{TW} + \text{PCR(TW)} \Rightarrow \text{LW} \qquad [7]$$

$$\text{LW} + \text{PCR(LW)} \Rightarrow \text{TDW} \qquad [8]$$

Going the other way, the rules are:

$$\text{TDW (I) TDW} \Rightarrow \text{LW} \qquad [9]$$

$$\text{LW (I) LW} \Rightarrow \text{TW} \qquad [10]$$

In EM smog (dense EM signals, even though very weak), there is increased nonlinear multiwave interactions, both interferometric and phase conjugative. So in such EM smog there exists an increased, non-negligible component of LW and TDW irradiation, formed by the above transduction interactions in the irradiated living cells and bodies. This latter part accounts for many of the long term effects which show up in "correlative" studies, but which do not show up in the normal limited-frequencies TW experiments produced by so-called "physical experiments."

The standard EM bioeffects manner of experimenting uses a theoretical model of simple, direct energy deposition by means of absorption in cells and tissues. Now we must utilize an "open system" throughout the body dielectric in which higher polarized EM energy (either LW or t-polarized) travels readily throughout all areas of the body.

EM Bioeffects and Wave Transductions

In the EM bioeffects experiments to date, there has been absolutely zero control of the internal transductions, and there has been no consideration of them. Almost all transductions of TWs into LWs will immediately result into transduction into TDW and time-excitation charging of the cells and all their internal components. These time-excitation charges (excited states of time-charging) will then slowly decay, emitting longitudinal EM waves in the process which travel throughout the body. These emitted LWs interfere with each other and the LW processes in the body, producing low level TW "noise jamming" of the cellular processes and—most especially—of the deep cellular control system of the body. The latter system functions by means of LWs and TDWs. Long-term effects of such "noise jamming" of biochemistry reactions and the control systems of the cells and body, can lead to degenerative diseases such as arthritis, atherosclerosis, Alzheimer's disease, heart disease, cataracts, tumors, leukemia, somewhat accelerated aging, immune system weakening and depressing, etc. As the body's natural defense and regeneration ability weakens, opportunistic infections further add to the degeneration.

It was shown by Whittaker in 1903 that any scalar EM potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs. [See *Figure 9*] Each longitudinal EM wave pair in the harmonic series is a phase conjugate pair. Unknown to Whittaker, that "LW phase conjugate pair" is actually a time-polarized EM "photon" structure, containing a phase conjugate LW photon substructure. So staid old "voltage" (potential) is comprised of a harmonic series of time-polarized EM waves!

The corresponding "t-polarized photons" are actually spin-4 *supergravitons* comprised of coupled antipairs of LW "photons". Each LW photon is a spin-2 *graviton* comprised of a coupled TW photon-antiphoton pair. Each TW photon and antiphoton in the coupled pair has spin-1. The couplet, being the longitudinal "photon" or graviton, thus has spin-2. And so on.

In 1904, Whittaker showed that any EM field or wave pattern can be decomposed into two scalar potential functions. By applying Whittaker 1903 decomposition to each scalar potential function, any EM field or wave has an enormously rich internal structure, comprised of scalar potential functions in interference. Each scalar potential function is comprised of time-polarized EM waves (made of spin-4 supergravitons). Each time-polarized EM wave is further comprised of bidirectional longitudinal EM wavepairs. Each longitudinal EM wave is further comprised of phase-conjugate pairs of transverse EM waves. Each transverse EM wave is further comprised of two scalar potential functions, each of which is in turn comprised of an infinite harmonic series of time-polarized EM waves, and so on *ad infinitum* and *ad nauseum*. Obviously all EM potentials, fields, and waves have an enormous "internal structure" or internal "information content of the field".

The Cellular Control System Uses t-Polarized EM Photons and Waves

The master cellular control system (MCCS) in the body utilizes these photon-gravitons, their corresponding EM wave polarizations, and transductions between them. By the use of very powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime in precise patterns. In short, it utilizes general relativistic spacetime curvature engines, rather than "EM photon signals", to accomplish its control of the cells and their processes. It does not merely send a "signal" and hope the cell itself acts upon it. Instead, it sends an "engine" which alters the local spacetime in which the cell and all its components are embedded. This "local ST curvature engine" then physically works directly upon the cell and all its parts, to alter the cell according to the specific "template" of the engine.

This is the direct application of the Wheeler's simple statement of general relativity: *Mass (trapped energy) acts upon spacetime to curve it, and curved spacetime acts upon mass to change it.*

Popp's "ordinary photon" study of this cellular control system[2] therefore will detect the coherent ordinary TW photon residues which result from the photon polarization transductions occurring from the actions of the ST curvature engines upon the cells.

Cellular division, biochemical processes, growth and growth control, cell differentiation and dedifferentiation, etc. are all controlled by the MCCS or one of its subsystems, using spacetime curvature engines. The ordinary TW electronic activity our biophysicists detect and correlate are simply the intermediate components of the actual ongoing polarization transductions.

The Cellular Control System Uses EM Time-Energy

One important subsystem of the MCCS controls the immune system's functions, including the deepest and most subtle functioning performed by all elements of the immune system. This system has been rather intensively studied, but since our biophysicists have not possessed the full electrodynamics utilized, only the correlated residue of normal TW EM waves and photons have been recognized and studied. Hence Western biophysics is in its very crude infancy.

Incorporation of the total wave and photon polarizations, together with their transductions, will eventually present a revolution in both biomedicine and biophysics. Unfortunately, presently our entire medical community is very firmly entrenched in only ordinary TW EM polarizations and ordinary biochemistry. Our scientists thus are studying less than 1% of the functioning of the living organism.

As an example, since Western scientists have no knowledge or research into the time-polarized mind domain and mind-body coupling, the mystery of the *placebo effect*—in which the mind has been verified to produce actual physical effects – will indefinitely remain a mystery.[3] As one result of that "crippling" of our medical theory, medical experimenters have to deliberately drive experiments on drug effects, etc. beyond the "mind's own threshold" of physical effect. In short, they have to drive the experimental results over and above the results being accomplished by the placebo effect. They must perform valid placebo control experiments, and subtract the placebo component results from their own experimental results, to determine the "delta" induced purely by the drugs.

The Regenerative Subsystem

Another important subsystem of the MCCS controls the *regenerative system* of the body. The regenerative system has various mechanisms, including growth control, differentiation and dedifferentiation of cells, etc. Becker's portrayal of the regenerative system is shown in *Figure 10*. One of its most important functions is healing and restoring damaged or diseased cells. It does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] *et al.*, although there is still sporadic, ill-funded, and ill-planned research on regenerative effects from time to time.[5] Becker and his colleagues came closest, and—considering the severe limitations of the orthodox U(1) EM model available to them—Becker's work is particularly revolutionary. He demonstrated beyond question that simple DC potentials could generate cellular differentiation and redifferentiation, and produce results at global sites in the body and not just at the local injury sites where the DC potentials were applied.

Becker was nominated for a Nobel Prize for his Herculean efforts and remarkable results. For his trouble he was hounded, his funds were withdrawn, and he was forced to retire at an early age in his 50s. His biggest "heresy" was to bravely point out that power line radiations had deleterious effects upon biological systems, and to testify to it in a series of hearings and court cases where a very powerful electrical power industry essentially ran rough shod over science and the public.

Organized Science Is Highly Politicized and Imperfect

It is not only politicians that are "bought and paid for" by powerful vested financial interests. In organized science, precisely the same arrangement holds true. Scientists are no better or no worse than any other segment of humanity. A surprising percentage will produce the results desired by their benefactors, if provided with cushy jobs, good income, prestige, and secure positions.

It has been said that money is power. It has also been said that power corrupts, and absolute power corrupts absolutely. In the scientific community, as well as in the political community, this "great corruption" by great financial power has been elevated to an art form. Simply examine the funds made available to scientists. There is not enough for all the scientists, so it is fiercely competitive. If the scientist does not get grants (e.g., a would-be professor at a university), he does not bring in extra money to the university. He cannot get his graduate students funded, so his own innovative (read: heretical) research is thwarted. His papers are rejected by the journals, and before long he has no funded position in science. If lucky, he will be employed in the grocery store or the butcher shop.

Every bit of funding available for the scientists to compete for (and apply for grants) already has firmly fixed controls of exactly what research can be done with that money. The "science" that will be applied is already formulated. That and only that can be done. It is a tribute to the dedication and resourcefulness of our working scientists themselves, that anything innovative at all gets done.

One can have a very fine scientific career, tenure as a professor, get one's graduate students funded, get one's papers published in prestigious journals, win awards and achieve stature and prestige, if one just plays the desired game prescribed by the "old boys schools" in control of almost all the scientific funding. Yes, one can permissibly move the decimal point in the present models a bit, and be rewarded handsomely for doing so. But one is damned if one attempts any serious change or overthrow of obsolete scientific models and practices favored by the "system".

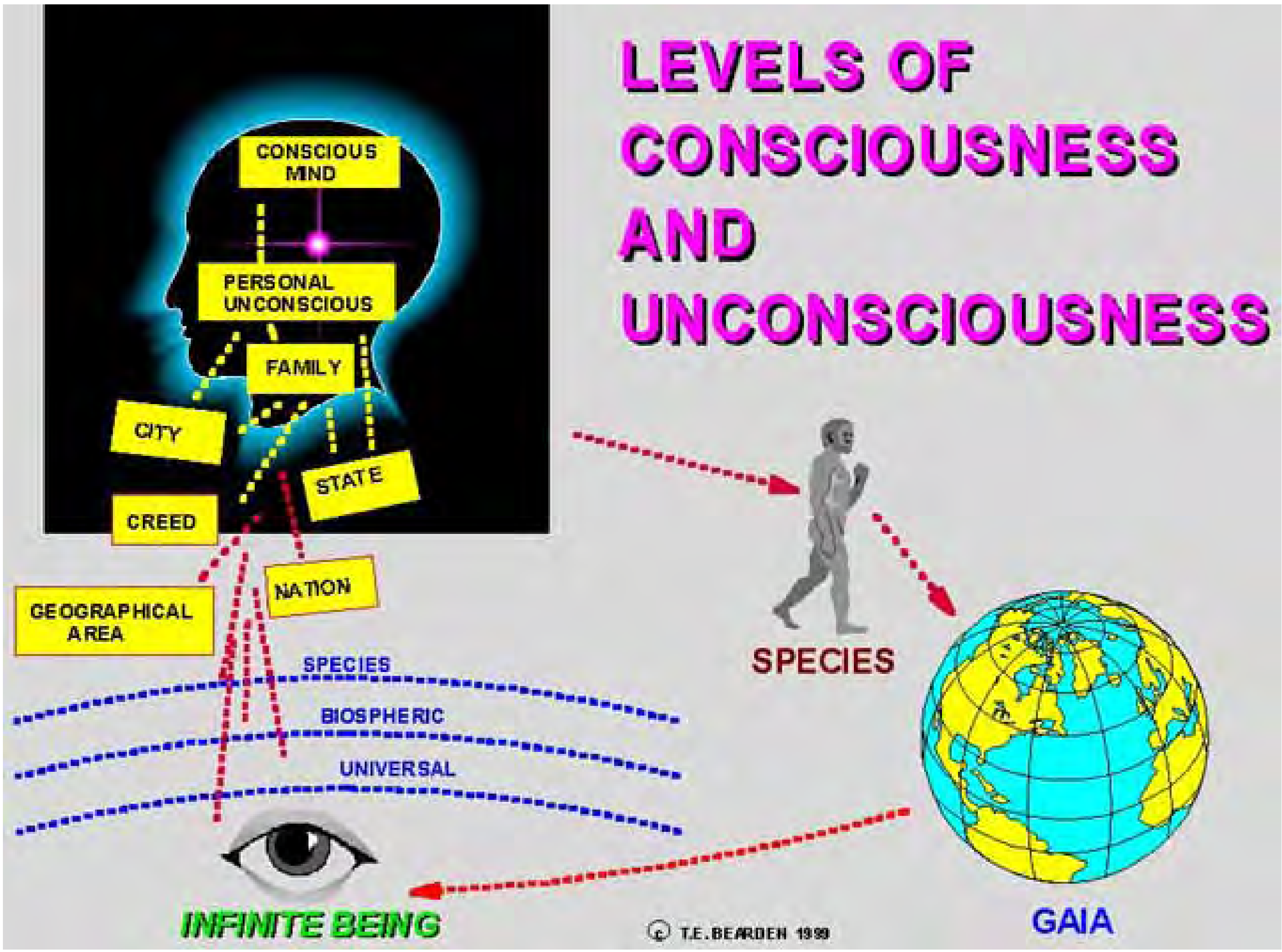

Figure 8. Nested levels of consciousness and unconsciousness. The mind levels contain the "life" and "mind" dynamics.

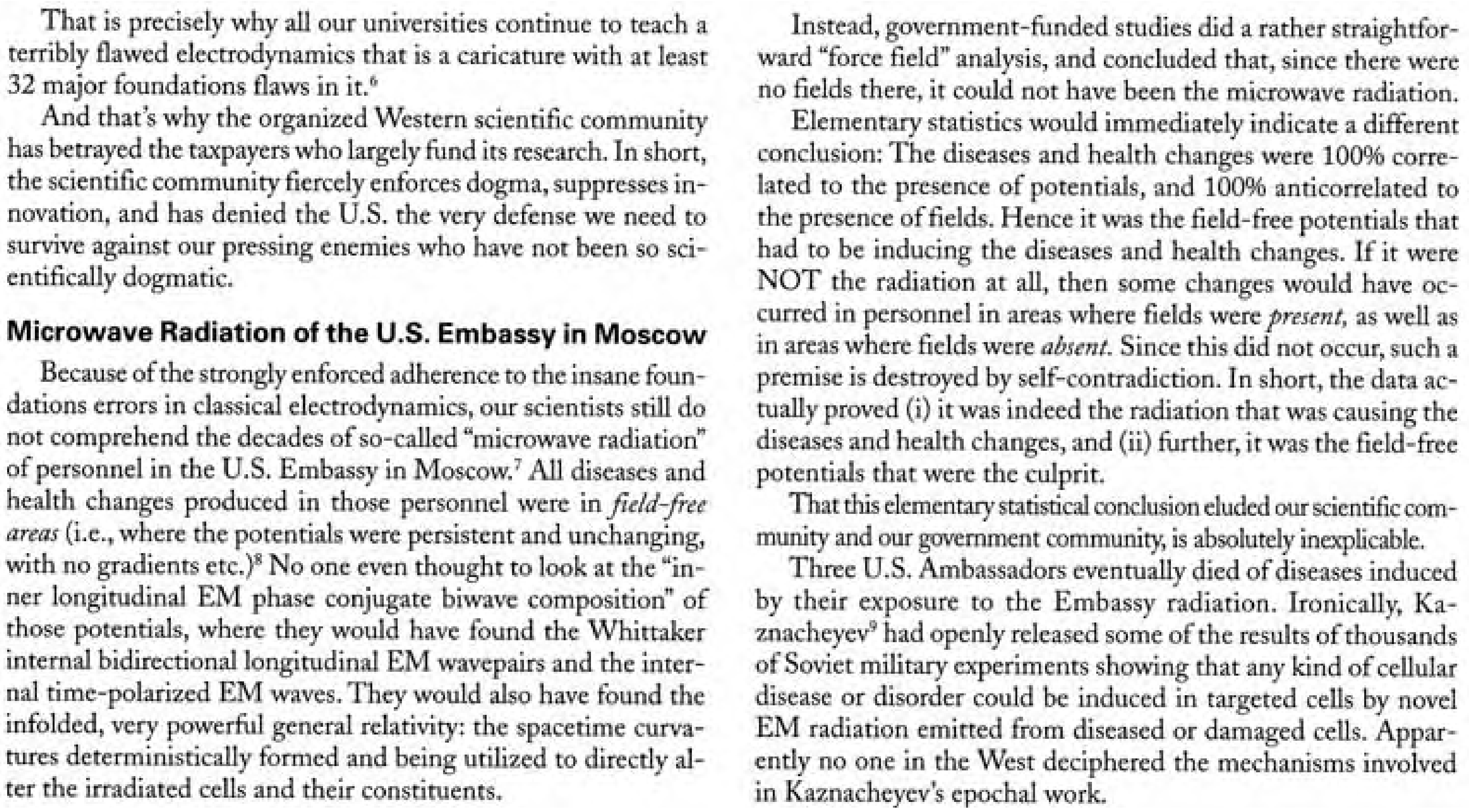

That is precisely why all our universities continue to teach a terribly flawed electrodynamics that is a caricature with at least 32 major foundations flaws in it.[6]

And that's why the organized Western scientific community has betrayed the taxpayers who largely fund its research. In short, the scientific community fiercely enforces dogma, suppresses innovation, and has denied the U.S. the very defense we need to survive against our pressing enemies who have not been so scientifically dogmatic.

Microwave Radiation of the U.S. Embassy in Moscow

Because of the strongly enforced adherence to the insane foundations errors in classical electrodynamics, our scientists still do not comprehend the decades of so-called "microwave radiation" of personnel in the U.S. Embassy in Moscow.[7] All diseases and health changes produced in those personnel were in *field-free areas* (i.e., where the potentials were persistent and unchanging, with no gradients etc.)[8] No one even thought to look at the "inner longitudinal EM phase conjugate biwave composition" of those potentials, where they would have found the Whittaker internal bidirectional longitudinal EM wavepairs and the internal time-polarized EM waves. They would also have found the infolded, very powerful general relativity: the spacetime curvatures deterministically formed and being utilized to directly alter the irradiated cells and their constituents.

Instead, government-funded studies did a rather straightforward "force field" analysis, and concluded that, since there were no fields there, it could not have been the microwave radiation.

Elementary statistics would immediately indicate a different conclusion: The diseases and health changes were 100% correlated to the presence of potentials, and 100% anticorrelated to the presence of fields. Hence it was the field-free potentials that had to be inducing the diseases and health changes. If it were NOT the radiation at all, then some changes would have occurred in personnel in areas where fields were *present*, as well as in areas where fields were *absent*. Since this did not occur, such a premise is destroyed by self-contradiction. In short, the data actually proved (i) it was indeed the radiation that was causing the diseases and health changes, and (ii) further, it was the field-free potentials that were the culprit.

That this elementary statistical conclusion eluded our scientific community and our government community, is absolutely inexplicable.

Three U.S. Ambassadors eventually died of diseases induced by their exposure to the Embassy radiation. Ironically, Kaznacheyev[9] had openly released some of the results of thousands of Soviet military experiments showing that any kind of cellular disease or disorder could be induced in targeted cells by novel EM radiation emitted from diseased or damaged cells. Apparently no one in the West deciphered the mechanisms involved in Kaznacheyev's epochal work.

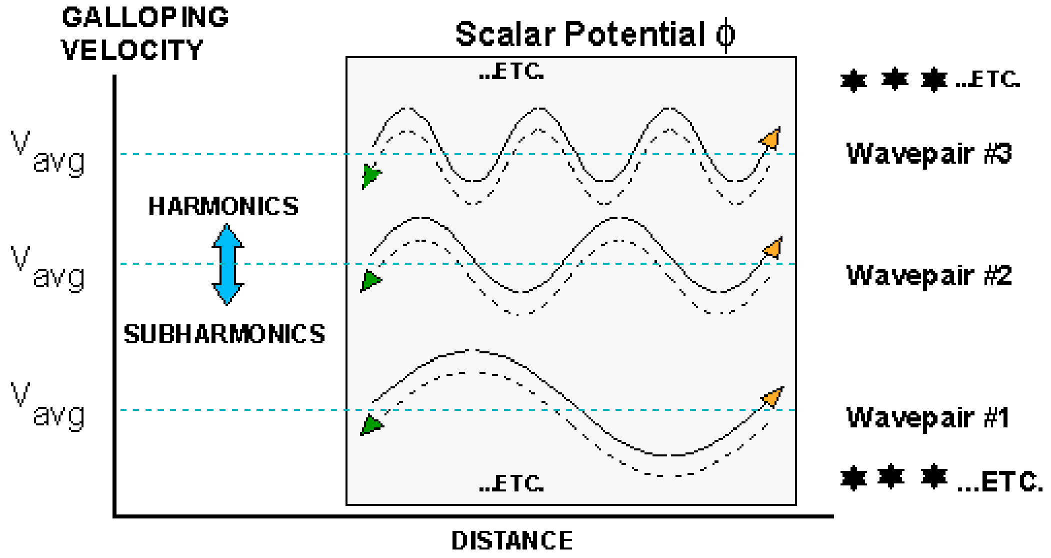

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.
- Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 9. Infolded longitudinal biwaves (time-polarized EM waves)

composition of a scalar potential.

EXPLORE! VOLUME 9, NUMBER 3, 1999

The Present Situation in Western EM Bioeffects Research

Our scientific community has not even *designed physical experiments* to validate long-term EM bioeffects of the kind caused by induced wave transductions and long-term effects inside the absorbing human cells and in the body's regenerative and immune systems. Western science has no knowledge of these technical internal "wave polarization transduction" mechanisms. Yet the basis for such mechanisms has existed in the hard physics literature for decades. Leading EM bioeffects researchers seem totally unaware of small reaction cross sections inside the body cells and tissues for the transduction of transverse EM signals into LW and t-polarized EM signals. Conventional Western researchers have not investigated it, and often will not believe it when it is pointed out to them along with the supporting basis in the literature.

A major part of the mainline "EM bioeffects" community, of course, seems comfortably funded largely by the power industry etc.[11] The record is clear that this dominant, well-funded group has little or no intention of "finding" anything that will question powerline radiation safety and upset their cushy positions.

In fact, leading bioeffects researchers have pointed out this deliberate biasing of EM bioeffects as a research area. Quoting Dr. Andrew A. Marino, one of the great pioneers in the area of EM bioeffects[10] and powerline radiation assessment.[11]

"Neither scientists nor the public can rely on power-industry research or analysis to help decide whether powerline electromagnetic fields affect human health because power-industry research and analysis are radically misleading."

Dr. Marino then gives the specific details to prove it, and he shows specifically how they slant and mangle their experiments and interpretations of the results, to do it. If the major financial interests cannot *scientifically* destroy determined, honest researchers such as Marino and Becker, they do not hesitate to attempt to destroy them in lawsuits.

The "Diffusion Mixmaster" at All Body and Cellular Levels

Every mass and all its particles are in a violent energy exchange with the active vacuum. A human body and every part of it is in a continual exchange of EM energy with its environment(s), including transverse, longitudinal, and t-polarized (scalar) EM energy exchanges. Further, within the dielectric body there is a sort of intensive "diffusion" of special kind, actually due to iterative phase conjugation and iterative interferometry of the dense signals environment impinging upon it. This internal "diffusion" and transduction (iterative multiwave phase conjugation and iterative multiwave interferometry) thoroughly "mixes up and transduces" one type of EM wave or signal into the other, including many of the body's own TW, LW, and TDW signals.

Continual or sustained TDW exposure (irradiation by t-polarized EM waves) charges the body and its particles with time-excitation charging, something which has been erroneously omit-

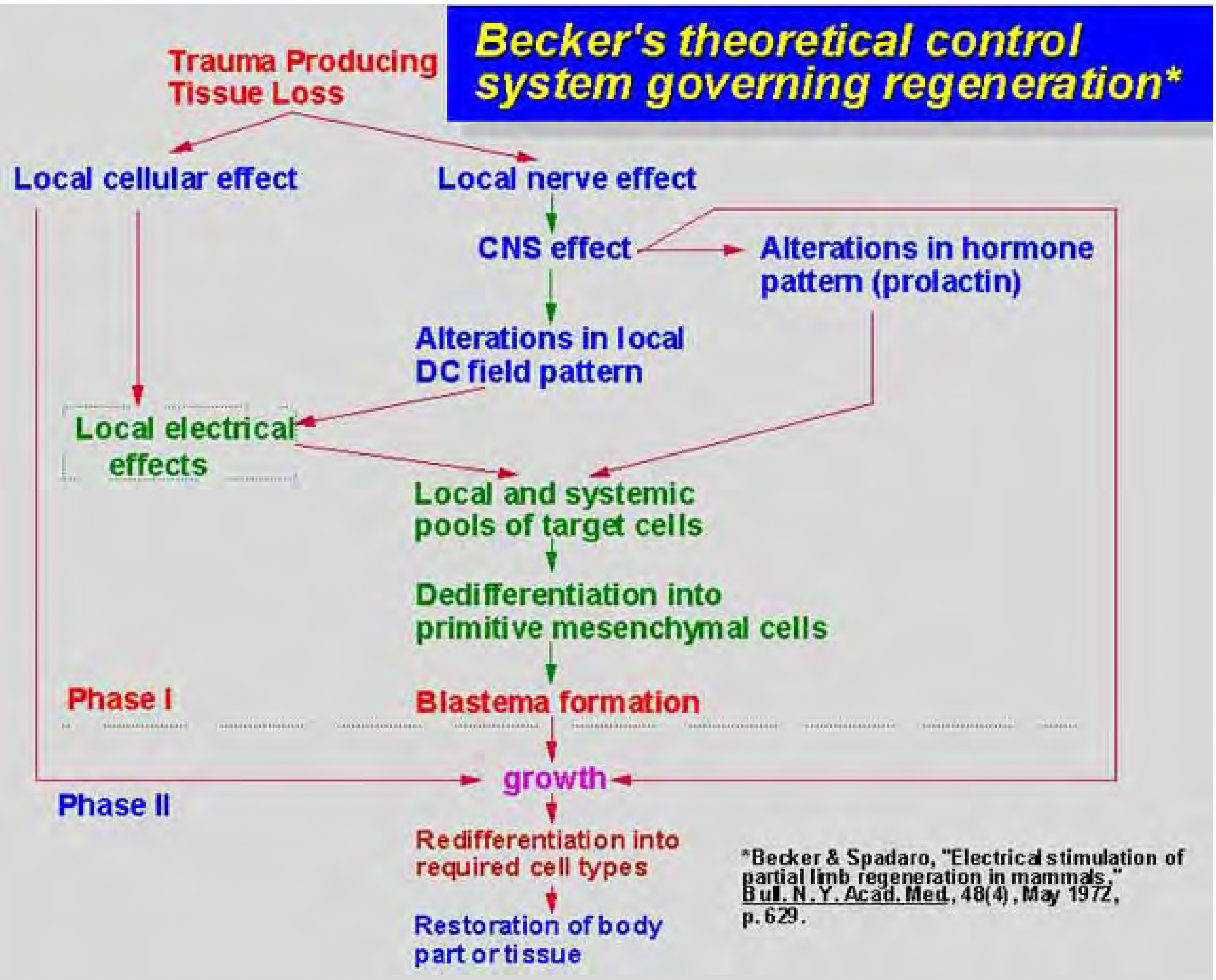
Becker's theoretical control system governing regeneration*
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into required cell types
Restoration of body part or tissue
*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

Figure 10. Becker's model of the Regenerative system of the body.

ted in physics, but is true nonetheless. Inexplicably, physicists just ignore what happens time-wise when a mass absorbs a photon. We have covered that elsewhere.

So inside the body dielectric, there is a "diffusion mix" correlate of any and every type of the waves, to any or all of the exchanges, plus the additional internal interactions of the body in its TW, LW, and scalar EM wave transductions. There is a one-to-one correspondence to any one of the three types of polarization waves (transverse, longitudinal, and scalar) and everything going on anywhere in the body.

Note that the type of LW we are using is actually made of phase conjugate pairs of transverse EM waves, so its "photons" are made of photon/antiphoton pairs. That is, our longitudinal EM wave has an internal TW structure, and the LW wave is comprised of spin-2 *gravitons*. Our longitudinal EM wave has an internal structure, and it dramatically differs from the LW used by the U.S. scientific community which ignores wave, field, and potential infolded substructures. Our LW (which seems to be what the Russians use) is also a gravitational wave because it is comprised of oscillations in its distributions of spin-2 photon/antiphoton pairs (i.e., *gravitons*).

The type of t-polarized (scalar) EM wave we are using is a phase conjugate coupled pair of the specialized electrogravitational (EG) LWs. The "t-polarized photons" comprising this t-polarized "EM" wave are comprised of spin-4 *supergravitons*. These are, I believe, what the KGB energetics weapons scientists use also. By iteration, we can go into as deep an internal multilevel structuring of EM and EG waves, of any type, via methods shown by Whittaker.[12]

Now let us examine a body as a dielectric, sitting there in its exchange, in relative equilibrium. It turns that every tiniest piece of a dielectric participates in any emission from it[13]—even the emission of a single photon, graviton, or supergraviton. Well, it *would* so participate, due to the special sort of correlated diffusion and interferometry and phase conjugation that we pointed out. Anything that escapes (is re-emitted) from the body, has been through this "diffusion wringer". It has escaped from each and every internal part of the body.

Novel Signatures in Body-Emitted EM Radiation of Any Polarization

A single photon or wave emitted from a dielectric carries in its internally structured time component a *template* of the exact internal time structuring of every part of the dielectric emitting that photon. It carries in its internally structured energy component a *template* of the internal energy structuring of every part of the dielectric emitting it. An emitted EM wave, being a collection of emitted photons and photon structures, thus carries the same internal structuring of both its transported energy and its transported time.[14] Western scientists do not use the internal

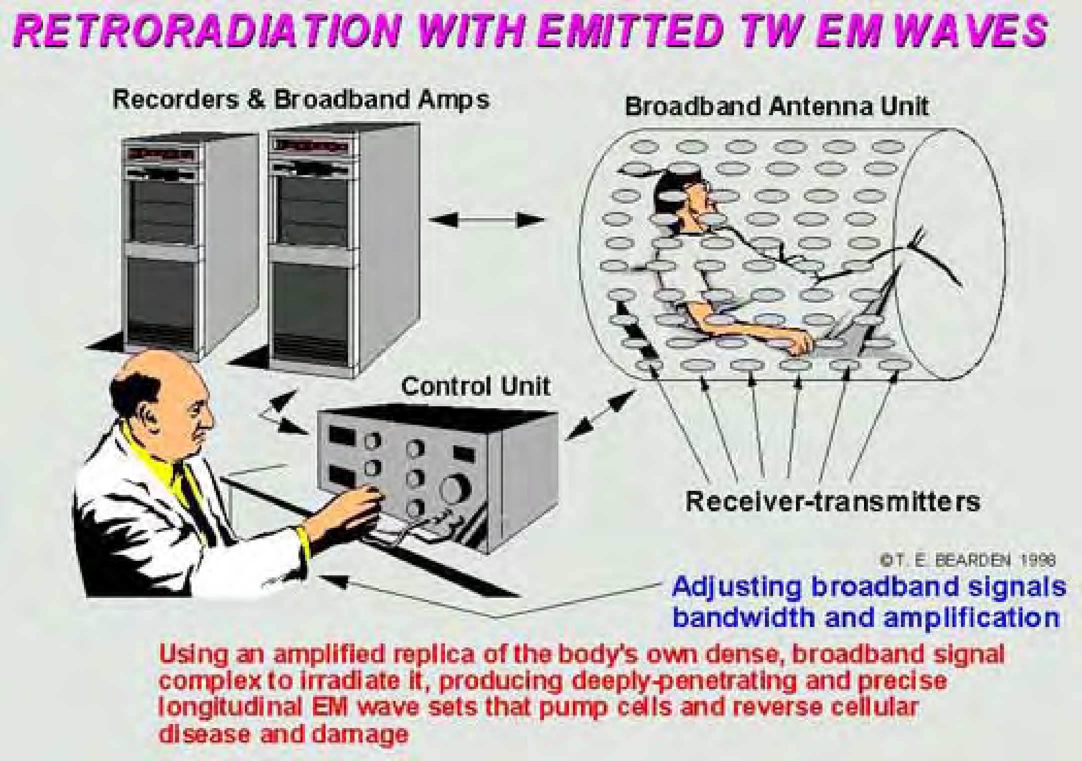
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

Figure 11.

EXPLORE! VOLUME 9, NUMBER 3, 1999

structuring of photons and EM sine waves, have no instruments for that, and so do not know it. The Russians know it,[15,16] use it, and have developed sophisticated instruments for it for the last 30 years or more.

Shortly we will postulate how conventional scientists experimenting with mind and behavior control electromagnetically might utilize an immediate consequence of the fact that the specific emission from a body at any time—whatever type of wave component we look at—is a total correlate to all that body's internal functions, its energy exchanges with its environment, its states, its mental states and conditions, etc.

And we accent that the coherently integrated intent-changes in that body from its coupled mind's operations are also in the substructure of photons and EM waves emitted from the living body into its environment.

How Western Mind Control Scientists Might Proceed

Conventional scientists would ignore scalar photons and waves, determine the TW spectrum, and then slowly and painfully correlate that to the various physical, mental, and emotional things. That's the way they would *have* to do it, since apparently they use only standard electrodynamics. They can in fact do this via brute force, because that emission TW spectrum is of waves escaping from the entire "diffusion machinery" mixmaster in the body dielectric. It does contain all those integrated *mind-intent* 3-space correlates as well as the ordinary body 3-space energy correlates.

Anyhow, in envisioning Western procedures, the test body would be sitting there, in its normal (let's say fairly normal and quiet) EM environment. [See *Figure 11*] Let's assume this body is not sick, but is healthy and normal. Its emissions are a part of its equilibrium condition. It follows that the totality of its TW emissions is correlated to the sum total of everything that interacted with the body, plus its own internal reactions, including from mind-body coupling. Indeed, everything that has ever happened to that entity—physically, mentally, emotionally, etc.—has internal correlates infolded inside the emitted transverse EM wave spectrum.[17]

So first the scientists would simply measure the full environmental TW spectrum down to very tiny levels in all frequency bands possible—everything that is coming in to the body from its environment. They would probably perform the experiments in a triply-shielded Faraday cage, so the external environment outside the cage contributes very little transverse EM wave noise (perhaps some Schumann resonance). Magnetic wave shielding and ELF shielding is something else again, and longitudinal EM wave shielding is in general not possible unless longitudinal EM waves are employed in the shielding. But a "quiet external TW environment" enhances the isolation and measurement of the specific TW signal transforms the Western experimenters would be seeking.

Then they would measure everything likewise that the body is *emitting* when in that environment. That is the "zero reference" or "normal reference" level. They would measure every aspect: frequency, amplitude, phase, wave shape, modulation, etc.

Probable Experimental and Measurement Techniques

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using portions of the same frequencies that are being emitted from the body in the "zero reference" environment. They would insure that the returned signal was of higher amplitude than the usual signal being emitted by the body at that same frequency and type.

They would consider the body dielectric as a "ship with portholes". Light (EM transverse wave energy) coming out of the portholes has originated and adapted from everywhere inside the ship and its operations. When the experimenters reinsert energy through one or more portholes, they are interested in determining precisely which operations inside the ship are affected. So the body physiological functions are also heavily instrumented.

By feeding back in the exact EM pattern and magnitude escaping the body dielectric through a single porthole, one "nullifies" that part of the emission by producing a net equilibrium in that component. This will in turn produce changes elsewhere in the spectrum emission from other portholes, and this can be measured. It will also produce "stress changes" in specific portions of the body dielectric, and in the heavily instrumented body these physiological, chemical, electrical, etc. changes can also be detected. Monitoring of the skin conductance, e.g., indicates the level of stress, and EEG monitoring with sophisticated instrumentation can indicate something of the brain wave states.

In short, the necessary measurements from which to construct specific correlates can be performed, recorded, and later analyzed.

This article will be concluded in Volume 9, #4 with Improving the Methodology, Further Phenomenology Experiments, Rejuvenation and other matters. ◆

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel U.S. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, U.S. Army. Graduate of Guided Missile Staff Officer's Course, U.S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor. He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuum's virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

1. *In fact, it simply applies in the laboratory that method that is already used by species for the so-called "directed evolution" component of evolution. The exact mechanisms for directed evolution also arise naturally in the new approach. Directed evolution is required; e.g., no reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by predators. Instead, such evolution occurs in a single jump in a species hard-pressed for a lengthy period. In the evolution of birds, one day every reptile egg that hatched, now hatched into a complete and functional bird with light bones, feathers, etc. A continuing debate (with perhaps more emotion than science) continues to this day on directed evolution. But it is real, and the mechanism can readily be generated (we've already done the gist of it).*

2. *Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert* Popp et al, *eds.*, Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977; *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149;—"On the coherence of ultraweak photoemission from living tissues," in* Disequilibrium and Self-Organization, *ed. by C.W. Kilmister, D. Reidel, Hingham, MA, 1986;—*et al., *"Physical aspects of biophotons,"* Experientia, *Vol. 44, 1988, p. 576-585;—"Biophoton emission: New evidence for coherence and DNA as source,"* Cell Biophys., *Vol. 6, 1984, p. 33-52. See particularly Fritz Albert Popp, Ke-hsueh Li, and Qiao Gu, Eds.*, Recent Advances in Biophoton Research and its Applications, *World Scientific Publishing Co., New York, NY, 1992.*
3. *With novel use of EM time-energy to curve spacetime, the physical actions of the "placebo effect" are due to the use of such "spacetime curvature engines" to produce actual physical effects—at least temporarily —in the cells and throughout the body's chemistry, physiology, etc. Experimental detection of these physical effects are well-established in the orthodox medical literature.*
4. *For example, see particularly R. O. Becker, "The direct current field: A primitive control and communication system related to growth processes,"* Proceedings of the. XVI International. Congress of Zoology, *Washington, D.C., Vol. 3, 1963, p. 179-183;— "The direct current control system: A link between environment and organism,"* New York State Journal of Medicine, *Vol. 62, 1962, p. 1169-1176;—and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents,"* Trans. N.Y. Acad. Sci., *Vol. 29, 1967, p. 606-615;—and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals,"* Bulletin of the New York Academy of Medicine, *Second Series, 48(4), May 1972, p. 627-64.*
5. *E.g., a popular discussion is given by John Travis, "Starting Over: Some Animals can regenerate limbs or even most of their bodies. How?",* Science News, *152(18), Nov. 1, 1997, p. 280-282. For a typical technical paper, see Jeremy P. Brockes, "Amphibian limb regeneration: Rebuilding a complex structure,"* Science, *276(5309), Apr. 4, 1997, p. 81-87.*
6. *I have one viewgraph which lists those 32 major flaws in electrodynamics. Further, leading theorists (Evans, Crowell, and others) in the Alpha Foundation's Institute for Advanced Study (AIAS)—of which I am a Fellow Emeritus by AIAS's kind tolerance—have added about another very technical and very fundamental dozen or so errors to my own list. A leading foundations scientist with over 500 papers in the literature commented dryly to the AIAS that "I broadly agree. that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science. Faraday visualized that "EM disturbances" surrounding wires, magnets, etc. were simply the transverse plucking of taut strings. As he states in his book, Maxwell—who also firmly believed in the material ether—simply captured Faraday's thinking and assumed that transverse "plucked string" wave disturbance in the material ether. At the time all this was expressed, the electron had not been discovered, the atom had not been discovered, the molecule was a featureless blob, and no one had the slightest notion of an atomic nucleus filled with protons and neutrons. Charge was simply a "certain volume of electric fluid" and everything electromagnetic was expressed in fluid dynamics. As an example, that "material fluid" and "material ether" theory was destroyed in the late 1880s by the Michelson-Morley experiment. All the Maxwell equations assume that material ether to this day; not a single equation was ever changed to eliminate the material ether assumptions during the last 110 years.*
7. *For full explanations of the "Moscow Signals," see T. E. Bearden, .,* Gravitobiology: A New Biophysics, *Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991;—* AIDS: Biological Warfare, *Tesla Book Co., 1988. For an excellent Government study of the Embassy radiation by excellent but orthodox Johns Hopkins scientists, see A. M. Lilienfeld* et al., Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees from Selected Eastern European Posts, *Department of Epidemiology, of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland, Final Report, July 31, 1978. On p. 245, it is noted that a marked excess of symptoms were noted in the Moscow employee group as compared to other East European employees in other East European nations.*
8. *Johns Hopkins researchers established the exact field patterns present in the Embassy from the microwave carrier radiation; see Robert C. Mallaliew,* A Model of the Microwave Intensity Distribution within the U.S. Embassy, *Moscow 1966 to 1977, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980. Nowhere in any of the Government-funded studies is there any consideration of field-free scalar potentials and their internal time-polarized EM wave structures (the bidirectional phase conjugate wavepairs shown by Whittaker to comprise the scalar potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.*
9. *Vlail Kaznacheyev and L. P. Mikhailova,* Ultraweak Radiation in Intercellular Interactions, *[in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions,"* Psi Research, *1(1), Mar. 1982, p. 47-76.]. See also Vlail Kaznacheyev* et al., *"Distant intercellular interactions in a system of two tissue cultures,"* Psychoenergetic Systems, *1(3), Mar. 1976, p. 141-142. In the same issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells,"* Psychoenergetic Systems, *1(1), Dec. 1974. The Soviet Union awarded Kaznacheyev a very high medal for his achievements.*
10. *For a very nice summary of EM bioeffects from the more conventional view, see Robert O. Becker and Andrew A. Marino,* Electromagnetism and Life, *State University of New York Press, Albany, 1982. The book is also available at http://www.ortho.lsumc.edu/Faculty/Marino/El/ElPub.html.*
11. *Andrew A. Marino,* Powerline Electromagnetic Fields and Human Health, *http://www.ortho.lsumc.edu/Faculty/Marino/Marino.html. This book is highly recommended. Particularly see "Chapter 5, Blue-Ribbon Committees and Powerline EMF Health Hazards," and "Chapter 6: Power-Industry Science and Powerline EMF Health Hazards." Biophysicist Marino is one of the leaders in the field and has been personally involved in many skirmishes with powerline-dominated studies and findings.*
12. *E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355; — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372.*
13. *Reali, G. C., "Reflection from dielectric materials,"* American Journal of Physics, *50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.*
14. *Present physics models the EM wave as transporting energy, but erroneously omits that it also transports time. A photon is made of angular momentum, or "energy x time". Since EM waves are made of photons, it follows that they transport both the energy components and the time components making up those photons. An EM wave is a wave of angular momentum, not just a wave of energy.*
15. *KGB energetics weapons scientists call this internal structuring the "information content of the field." However, the Russians use a good deception plan, which implies that the "information content of the field" is just the normal transverse EM spectral content of EM signals. Appreciable "pertinent" work using that phrase is produced by the open Russian scientific community, thus confusing the issue. Our electrodynamicists seem to buy it hook, line, and sinker. The deception plan also tricks Western scientists into thinking that psychoenergetics is nothing but a Russian version of Western parapsychology. Again, our scientists comfortably accept this with smug satisfaction.*
16. *Presently the Russians are in a great new version of this deception plan, with the goal of deceiving Western scientists into thinking that the "secret" Russian energetics is based on torsion fields. Consequently they are releasing substantial material on torsion wave theory, including replicable experimental results. The released version of torsion fields has one glaring omission: the time components and structuring. This deception plan is excellent, because the torsion field theory that is released, is indeed "real" and testable. It is also revolutionary, with respect to present Western science. It almost positively assures diverting the attention of much of the Western science community into this highly limited subset of the full Russian torsion theory—effectively diverting the West away from psychoenergetics and the direct use of t-polarized EM waves and fields, torsion or otherwise.*
17. *We point out that E. T. Whittaker, in "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355, showed that every scalar potential is comprised of an internal structure of bidirectional longitudinal EM phase conjugate wavepairs in an infinite set of harmonics. Each of those phase conjugate longitudinal EM wavepairs comprises a time-polarized EM wave. So the common potential contains an internal longitudinal EM wave structure that is further organized into an internal time-polarized EM wave structure. The mind-intent correlates and the total mind and emotional experience correlates from fertilization of the egg that led to development of that person, is in the internal substructure of every potential in the living body. In his "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372, Whittaker also showed that any EM field or wave is comprised of two scalar potential functions. It follows from Whittaker and Reali that the EM emissions from the living body carry the internal correlates of everything that has been experienced by that person from conception in the womb. Sometimes this deep embedded "hidden information content" can even be accessed in deep hypnosis. I once observed the deep hypnotic regression of a subject back to her prenatal experiences prior to birth, where the subject relived hearing words of rejection by the father, before the subject was born. The occurrence of the incident and the father's words were directly confirmed on the spot by the mother, who was closely observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological problems experienced by that person all her life. To resolve the problem, use of a powerful "simulated" reliving pre-birth experience was conducted with the subject again in the pre-birth state and actively participating, but now with the father figure directly expressing love, acceptance, and cuddling and hugging his daughter. This led to a rather dramatic remission of most of the deep psychological problems within 24 hours.*

Mind Control and EM Wave Polarization Transductions, Part 3 of 3

©Copyright 1999 by T.E. Bearden, USA

This outstanding article is the last, Part 3, Part I and 2 of which were published in Volume 9, #2 and #3 of Explore! for the Professional. Because we feel that Col. Bearden's Special Note below is so very important we are repeating its publication in all 3 parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and

achieving many more effects in the targeted bio-organism or groups of them. With Hunt's method, the system as an open dissipative system far from thermodynamic equilibrium in its environment is being measured as such. This more directly allows for determination of the environmental inputs on a much broader range of subsystems, both space-like and time-like. Particularly for mental and emotional correlates to environmental stimuli, such measurements and correlations are essential.

More Comprehensive Phenomenology Experiments

To return: Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.

In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories.

is both illegal and immoral.

Two Ways to Improve the Methodology and Experimental Correlates

We diverge momentarily: If they wish to improve this methodology, they will consider the body dielectric as a nonlinear isotropic medium. Then they will pretend that the spectrum of difference frequencies between each two adjacent waves is the spectrum of "emissions" from the body. They would also use those, as if they were actual emitted spectral signals, instead of the actual signals emitted. That procedure will give superior correlations, but it is more complicated. Part of that is known for underwater sonar,[1] and so our own scientists may or may not have realized it for use in mind and behavior correlate determination with respect to the body dielectric.

Another dramatic improvement in the depth to which the correlates can be constructed, is provided by detecting each porthole output in a two-channel device, where one channel has an adjustable time-delay as utilized by Dr. Valerie Hunt.[2] The two channels—one real-time and one slightly delayed—are then mixed, as by Hunt's method, producing an instrumental measurement of the chaotic part of the functions. Again, this can be measured, recorded, and later analyzed. It adds (i) a completely different kind of thermodynamics in the system measurements, (ii) a dramatically extended set of correlates, (iii) more direct measurement of the time-polarization causative signals, and (iv) an applications technology eventually having many additional degrees of freedom and therefore capable of Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions.

A particularly vulnerable aspect of every mammal is the pleasure center in the brain. Experiments have shown that, when this center is stimulated, it is the most addictive experience possible. Rats will forego food, endure electrical shocks, starve, and even die to obtain stimulation of this center. Obviously a device capable of generating signals that evoke direct stimulation of the pleasure center in humans would be a powerful weapon.[3]

Extending the Data Base

See Figure 12. After the researchers have made an "individual porthole signal insertion" correlate database, then they would make a "two, three, four," porthole signal insertion correlate database by inducing selected multiple changes at once. Again they would analyze everything; the program would need one or more supercomputers. It would also require quite a highly qualified multidisciplinary team, extensive facilities, substantial funding, and would be many years in duration. We are describing a decades-long program. But it's a doable.

With sufficient development and performance of such a program, the researchers would have produced (with sufficient testing and analysis, and sufficient retest validation and verification) a database of "specific correlates for a given overall desired physical, mental, or physical and mental delta."

A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment.

The Dielectric Body as a Ship With Portholes

Again let us review what the EM transverse wave irradiation from the body dielectric represents. Think of the body dielectric as a sort of "ship", with lots of portholes. When it gets irradiated, certain operations happen internally to that radiation, i.e., certain "processes" and "interactions" occur, which change both the mental and physical states. The output emissions back from the body dielectric, out through the filtering "portholes", will directly correlate to what has happened inside the ship.

It is terribly important to realize that one can use the same frequencies and spectra coming back out of those "portholes" to insert EM signals and waves etc. back in there, and into the deepest processes going on in the body and in the mind (at all levels, including the deep unconscious).

times. So it will reach directly into and affect any and every part of the dielectric, as you wish, and in whatever manner you have correlates for.

We filed an invention disclosure[4] on the above correlate and "back-porthole" process for use in directly inducing full and efficient conversion directly to time-density (scalar, or time-polarized) EM waves inside the body/ cells, therefore immediately and directly time-excitation charging the body cells and all their internal components.

Subtracting the Normal Correlates to Obtain the Deltas

A specifically diseased body, e.g., has a specifically tailored change in its "normal" emission spectra, because of the disease condition alteration of its resident spacetime curvature engine. By having a database for the normal emission spectra of the healthy body, one can subtract it from the actual emitted spectra, to obtain the "delta spectrum".

If the correlates have been previously determined by the above mentioned research program, then one may simply introduce an amplified "delta spectrum" into the body, i.e., through the portholes. This will immediately time-excite the cells and their components throughout the body, particularly where that disease does exist. It turns out that the time-excitation charging by that method will "pump" those diseased cells in the time domain. The cells themselves (and all their parts) are highly nonlinear—to time-excitation

Remember, whatever you put back in there, is going back through the "mixmaster" hopper. It will be iteratively phase conjugated and interfered with by everything, countless charging, they are totally nonlinear. So they will act as rather perfect pumped phase conjugate mirrors, but pumped in the time-energy domain rather than in the spatial

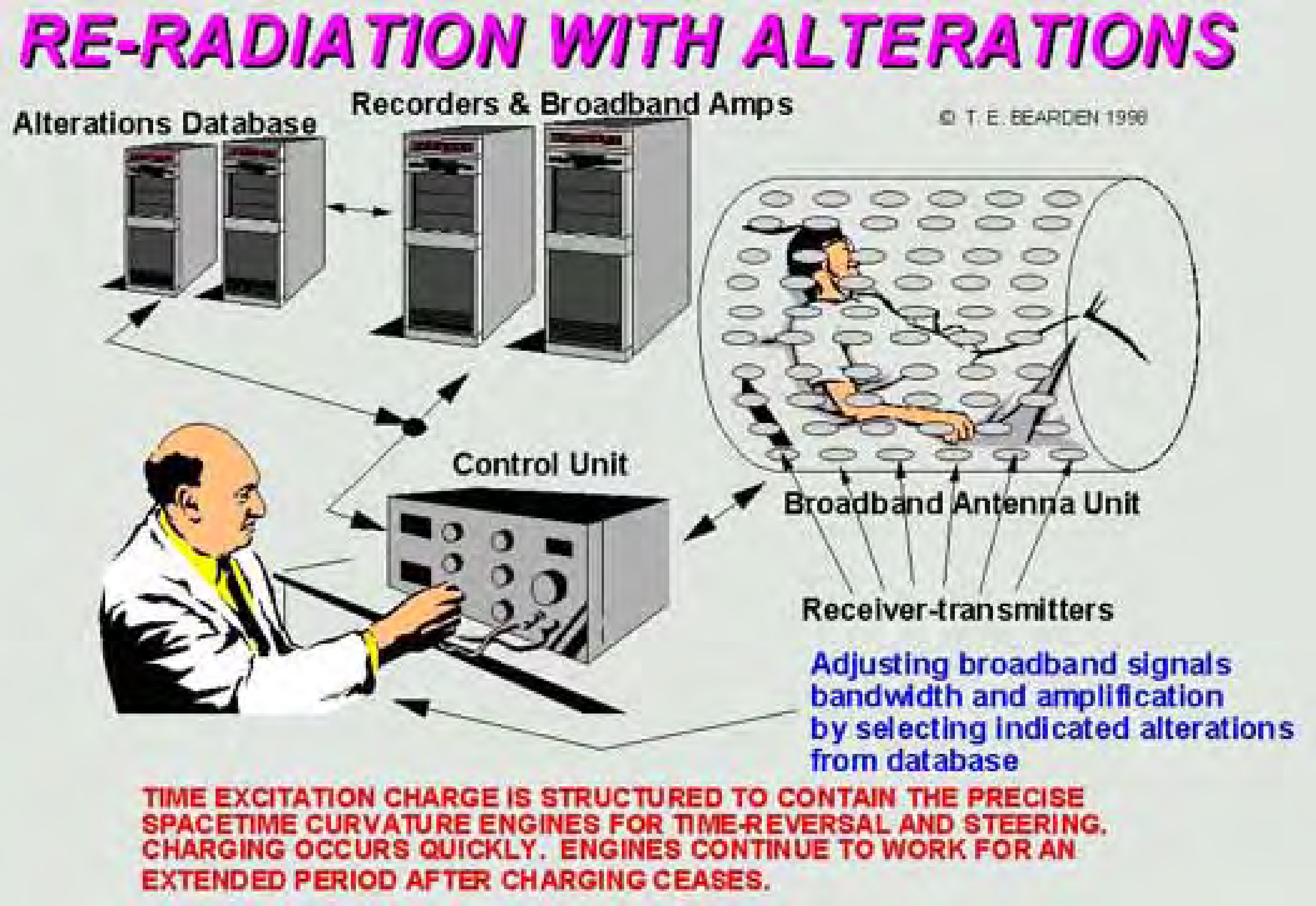
RE-RADIATION WITH ALTERATIONS
Alterations Database
Recorders & Broadband Amps
© T. E. BEARDEN 1998
Control Unit
Broadband Antenna Unit
Receiver-transmitters
Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database
TIME EXCITATION CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Figure 12. Using an amplified replica of the body's own dense, broadband signal complex, containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal while steering the time-reversal trajectory to also correct a congenital condition.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

energy domain. This is a dramatic extension to nonlinear phase conjugate optics theory.[5]

General relativity tells us that each specific detail of the mass and energy structure of each cell and all its parts—down to the finest details—is accompanied by a precise set of spacetime curvatures. By reintroducing the same energy as the disease delta TW waves, the affected cells add the phase conjugates, converting the inserted TW signals into longitudinal EM wave signals. The longitudinally-pumped cells again phase conjugate, converting these longitudinal EM waves to time-polarized EM waves, and "charging" the cells and all their parts in the time domain.

Suddenly, at that point the resident spacetime curvature engine—rather than TW or LW EM waves-serves as the "input signal" to the cell-and-its-parts as pumped phase conjugate mirrors (PCMs), pumped in the time domain. An extension to the distortion correction theorem of nonlinear optics results. An exact, amplified antiengine for the diseased cells—i.e., a phase conjugate replica of the resident spacetime curvature engine—is formed. This amplified antiengine overpowers the resident disease engine, powerfully time-reversing the cells and their components back to a previous healthy state. The time-reversal actually occurs over a period of time after the "time excitation charge-up" ceases, say over the next week or two. Usually two to three quick irradiations are required, one week apart, to give a time-reversing period of three to four weeks.

Even though appreciable forces are operating internally upon the cells and all their parts, there is no net translation force upon the cells or any part. E.g., as Pepper[6] states,

"On a more fundamental level, the ideal lossless PCM reverses all the quantum numbers of the incident photon (i. e., linear momentum, angular momentum, etc.). It can be shown that the PCM therefore experiences no linear or angular momentum transfer from the incident photons; hence, the PCM is free from photon radiation pressure and torques."

Note that, in ordinary EM theory, it is precisely translation forces that are engendered upon charged mass by EM fields. This process is quite different. In biological terms, this process dedifferentiates (in physics: time-reverses) the diseased cells and all their components—including the genes—back to an earlier healthy state.

By taking a diseased body and feeding back into it an exact amplified replica of its emitted TW spectrum, one can force the body to "time-excitation charge" all cells and their components immediately, with high efficiency. The human body and every cell and every part of every cell does the necessary "calculation" for the exact antiengine for each part. Since time is extremely condensed spatial EM energy (by a factor of almost 10^{17}), a startling contraction occurs in the radiation time required to time-charge the cells and their components and charged particles. This is a method for instant initiation of time-charging, hence it is extremely rapid. What took Priore several hours of LW irradiation to accomplish in the body,[7] can now be done in seconds with this "thwarting" and spectrum transduction method. In fact, it is an intense stress on the body, and—say—should only

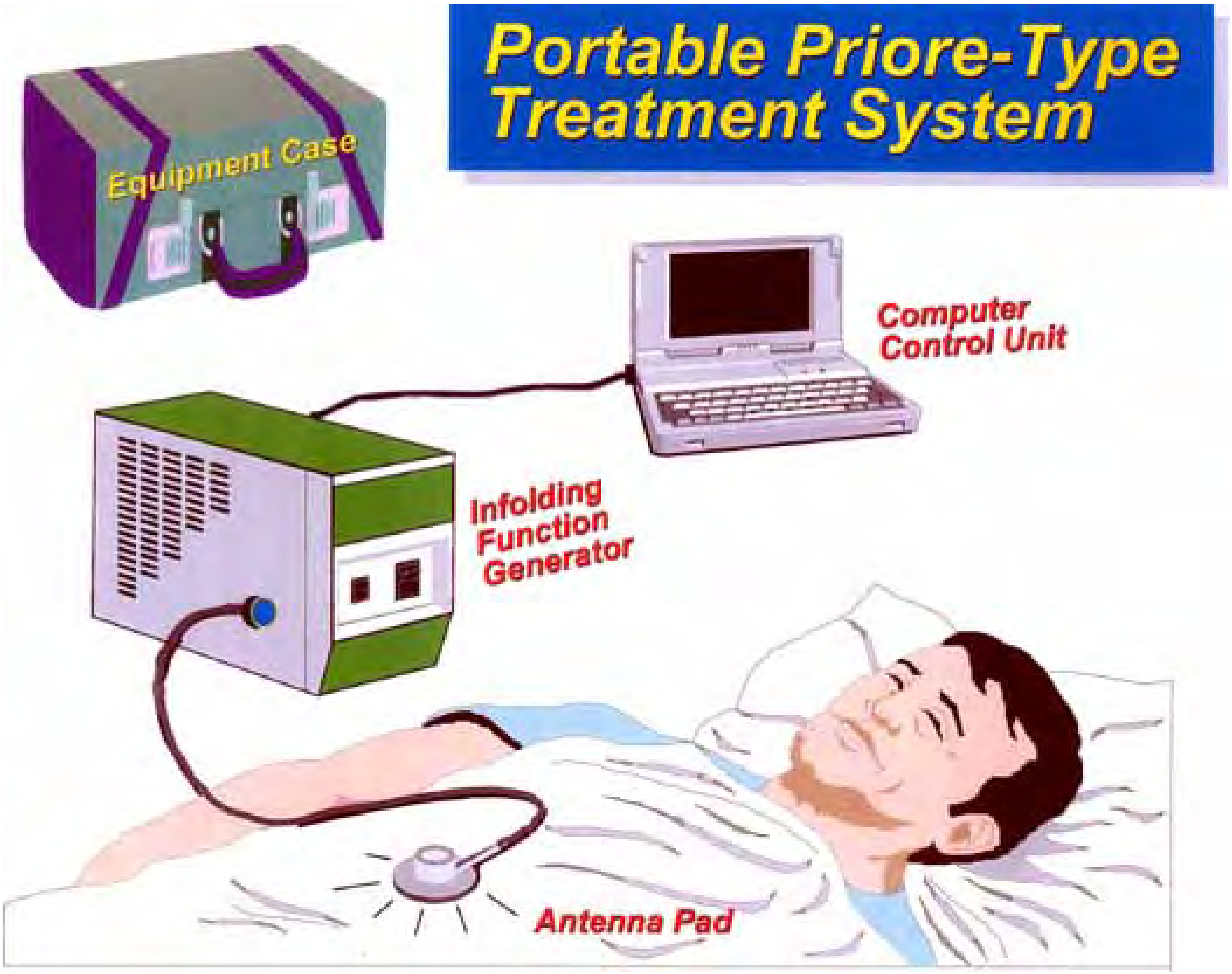
Portable Priore-Type
Treatment System
Equipment Case
Computer
Control Unit
Infolding
Function
Generator
Antenna Pad

Figure 13. Portable first-method treatment unit proposed to the Department of Defense. For treating and curing mass casualties from terrorist BW attacks on U.S. population centers.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

be continued for 30 seconds or so. We highly stress that overcharging in this time domain can be detrimental or even lethal if prolonged.

At the end of the very short time-excitation charging time, the body is "time-excitation charged in all its minutest particles". So everything is therefore pumped in the time domain for the next several days after the irradiation is removed, since the time-charge decays very gradually over a period of time. During that decay period, every part of the cell and body acts as a pumped phase conjugate mirror material, pumped in the time domain rather than in the 3-space energy domain. When time-pumped, the mass itself reverses to a previous physical state. This can be used for both nonliving and living masses. Because there is no forcible translation manipulation of the cells and their parts, it can be utilized on living cells in vivo. In the case of diseased cells—e.g., AIDS-it appears that three precisely determined 30-second "AIDS correlate" irradiations a week apart, should be sufficient to completely cure AIDS, including removal of all HIV-genetics changes from every previously infected cell of the body. Also, it would accomplish considerable rejuvenation of the body (making the body younger).

Rejuvenation: Restore the Telomeres and Produce Immortal Cells

Saving Most of the Casualties from Mass Terrorist Attacks

To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc.

Each device could be used in "assembly line" fashion upon hundreds of sickened patients, in case of a WMD (weapons of mass destruction) attack upon one or more of our cities. The first generation equipment would save probably 70% of the stricken (as from anthrax, etc.). With second generation equipment, one would expect 90% of the stricken casualties could be saved.

Just now, there is a very dim prognosis for all those casualties, using present available treatment capabilities. Time and lack of vaccines defeats just about everything that can be done. Brutally speaking, most of those unfortunate enough to breath in anthrax spores from an

In fact, it is now known that the gradual loss of the telomeres—the natural ends of chromosomes—in a cell's chromosomes is what causes the cell to age. The gradual loss of the telomeres and the resulting shortening of the chromosomes represents a cellular "disease state" or "disorder" state" accompanied by a precise delta in the spacetime curvature engine accompanying every affected (aging) cell.

By determining the correlate for that precise telomere loss in the aging body's emitted radiation spectrum, an effective telomere regeneration process immediately emerges. One simply amplifies that aging "precise telomere reduction" correlate and reradiates the body with it, passing an amplified correlate back into every cell of the body. The cells and all their parts—including the chromosomes—add the phase conjugate replicas, converting the incoming TW EM correlates to longitudinal EM wave correlates. The cells again phase conjugate the LWs and add the phase conjugate replicas, converting the LW correlates to time-polarized EM wave pumping correlates. This latter process time-excitation-charges the aging cells and all their parts, including the genes and chromosomes. The cellular matter itself produces the amplified "aging antiengine" for each and every cell and telomere ending. The result is to again add telomeres, converting the cell back to a younger, more vigorous cell, and reversing the process of aging.

We point out without further elaboration that these methods also introduce direct engineering signals into the body's master cellular control system, including the regenerative subsystem. Hence the body's regenerative subsystem is involved in helping produce the necessary engines and operate upon the short telomere chains to add back telomeres. More specifically, the regenerative subsystem is changed into attack, are going to die. At least 80% of them will expire, in spite of everything the present system can do for them.

Note each device could rapidly treat a whole series patients, one right after the other, merely laying them between "blankets" containing wire antennas, irradiating with the proper correlate spectrum quickly, moving on to the next patient, irradiating quickly (seconds), etc. The entire apparatus could be made to fit in a large suitcase-sized container, and it would be highly portable and controlled by a laptop-sized computer.

Unfortunately, so far the DOD doesn't even understand what the heck we're talking about! The National Institute of Health (NIH) and the National Science Foundation (NSF) are not interested at all. After all, it isn't drugs, vaccines, antibiotics, etc. — which are simply not going to do the job. However, those are what the government and scientific communities are going to finance to the tune of hundreds of millions of dollars. They are then going to lose almost all those mass casualties when the terrorist attacks on our population centers inevitably occur.

Those are Americans. They are going to die. Most of them could be saved. We ought to save them.

For any change in state (physical, chemical, emotional, mental, perceptual), the body dielectric "portholes" change and the output spectrum changes. The delta from zero reference now represents that exact change.

Where Western Mind Control Research Has Probably Arrived

Western mind control researchers probably first established a correlation database of porthole emission correlates to changes in internal states, both physical perfect consonance with the external introduction of the amplified antiengines.

That exact process can be used to rejuvenate the aged population and remedy most of the debilities of growing old. That is how we eventually intend to do it, if we can convince the U.S. government and the U.S. scientific community to perform the necessary research and validation.

and mental. They almost certainly recognize the peculiar participation of every part of the body dielectric in the TW emission spectrum from the body. They will almost certainly have opted for this "porthole" notion of some similar expression of it, and the "thwarting" of these "relaxation and release" emission processes to engender internal physical changes in the body and in the brain-meat-computer, as they see it.

EXPLORE! VOLUME 9, NUMBERS 4 & 5, 1999

They probably experimented with different TW EM radiation mixes, to see what the body emission change to, and what states are induced. Undoubtedly they found that the induced states began to correlate with the former emission delta data base. That is, they found out which delta emissions spectra, reversed and amplified and fed back thorough the portholes, produce what internal changes.

Use for healing is the simplest of all. Use for killing or debilitating is more complicated. Of course weapons researchers wish to look only at the killing aspects. That is something they understand. They do not understand healing, for they do not even have the "regenerative" or "healing" mechanism in any of the scientific papers or texts. Becker[8] came closest in this country, and Priore did it in France. Both of those scientists were suppressed and paid a bitter price for their revolutionary and innovative research.

But look at what has been done, by "input-output" correlation research. The scientists will have a database now, that tells what state emerges in the body and in the mind, for what irradiation spectra (frequency, relationships, phases, amplitudes, wave shapes, etc.) one uses. And they did not work out all the wave-to-wave transductions ongoing in the mind and body connection and in the body and in the mind, individually, between TWÛLWÛTDW.

We suspect that this is the course Western mind control researchers will have taken, without even realizing the internal wave transductions occurring. And they would not be concerned with that. They would have what they want, for weapons, mind control, and behavior control. And of the internal "mix" of the dielectric and all its tiniest portions. So Western mind control researchers and rogue groups probably do get good results.

We are not just picking on the U.S. government here! Internal divisive groups exist in all large and powerful groups, and joust for power. In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old "Power corrupts, and absolute power corrupts absolutely" routine.

The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.

In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians.

they would of course ignore the healing aspects. And they would get results of course, revolutionary results. So they would undoubtedly classify the program very highly, thinking that they had the best of the best.

Rogue Groups May Also Have Arisen to Divert the Purpose of the Projects

And that is where real problems may arise. Every human group has within it competing subgroups, jousting for power. The more highly classified the group and its activities, and the less that is known about it outside its confines, the more strongly and secretly the subgroups can play. So they can emerge as rogue groups pursuing their own agendas, not that of their duly elected governments. In the extreme, such a deep black program can even become a "captured" program, which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government.

It has long been suspected in some quarters that even the US's highly classified research community may be riddled by such rogue groups, and so may be certain highly sensitive parts of the Research and Development community.

Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a

Where the Russians Probably Have Arrived

On the other hand, the Russians will have developed instruments for detecting and isolating the LWs and the scalar waves, in addition to the ordinary mundane instruments for measuring the TWs. So they will have not only the TW correlates, but also the much finer TWÛLWÛTDW (scalar EM wave) correlates. They will understand and employ the use of vacuum engines (spacetime curvature engines) which Western researchers will often invoke but not understand or even be aware that this is what is being invoked. In other words, Western researchers probably just use the old method of "input fi set of functions fi output". That is the standard basis for almost all Western system analysis. With it, one doesn't have to know the exact functions and processes going on in the middle. One just determines a transfer function from a given input to the output that results. Then one catalogs that transfer function. One does that for lots and lots of transfer functions for particularly desirable results, and one has the system analysis. One also has the immediate basis for a rapidly-developed applications technology.

Western scientists are good system analysts. But they have little or no comprehension of what is really going on in that "inside the ship functions" portion, even though they have determined the transform function which acts upon the input to give the "from the portholes" output.

The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification.

Nonetheless, the "simplified" TW EM "porthole correlate" approach will yield very positive results, due to the correlation existing with the LW s and scalar portions the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW interferometers, through intervening mass

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

including the earth and ocean. Western mind-control scientists and their rogue groups—assuming they exist as postulated—largely will not be able to do that, and probably won't even recognize Russian testing of psychoenergetics weapons directly in or over Western nations.

Russian Lethal Psychoenergetics Mind Control Tests Over the U. S.

In fact, the KGB tested just such highly advanced mind control weapons over the middle of the United States in 1997 on two occasions. The tests were conducted against Captain Button in his A-10 "Warthog" aircraft on April 2, 1997 [See Figure 14], and upon Captain Svoboda in her A-10 "Warthog" on May 27, 1997. [See Figure 15].

Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona. At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.

Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks. A secondary purpose was to stimulate the U.S. government, watch its investigation and findings, and ascertain if the U.S. knew of the extent to which Russian KGB mind control weaponry had advanced. Obligingly, we proved to me KGB that we did not understand what had actually happened to Captain Button.

Captain Svoboda was from the same Wing as Captain Button and was its most experienced night fighter A-10 pilot. In fact, she taught other pilots the fine points of night flying combat missions in the A-10. About two months after Captain Button's bizarre flight and death, Svoboda was engaged in night target bombing on the range. Rising from her ordnance drop on me target, Svoboda was suddenly struck and rendered hypnogogic and in a waking-dream state. Her sense of the vertical was instantly reversed. Superb pilot that she was, she immediately dream-perceived she was diving, not climbing out of the delivery. She instantly corrected to "climb out". So in the real world she dived sharply into the ground and her aircraft exploded.

This test demonstrated the instant alteration of a skilled pilot's perception, causing that pilot to inadvertently react promptly in an unintentionally suicidal manner. A secondary purpose of the test was to stimulate the U.S. government system again, to see if the system recognized what had

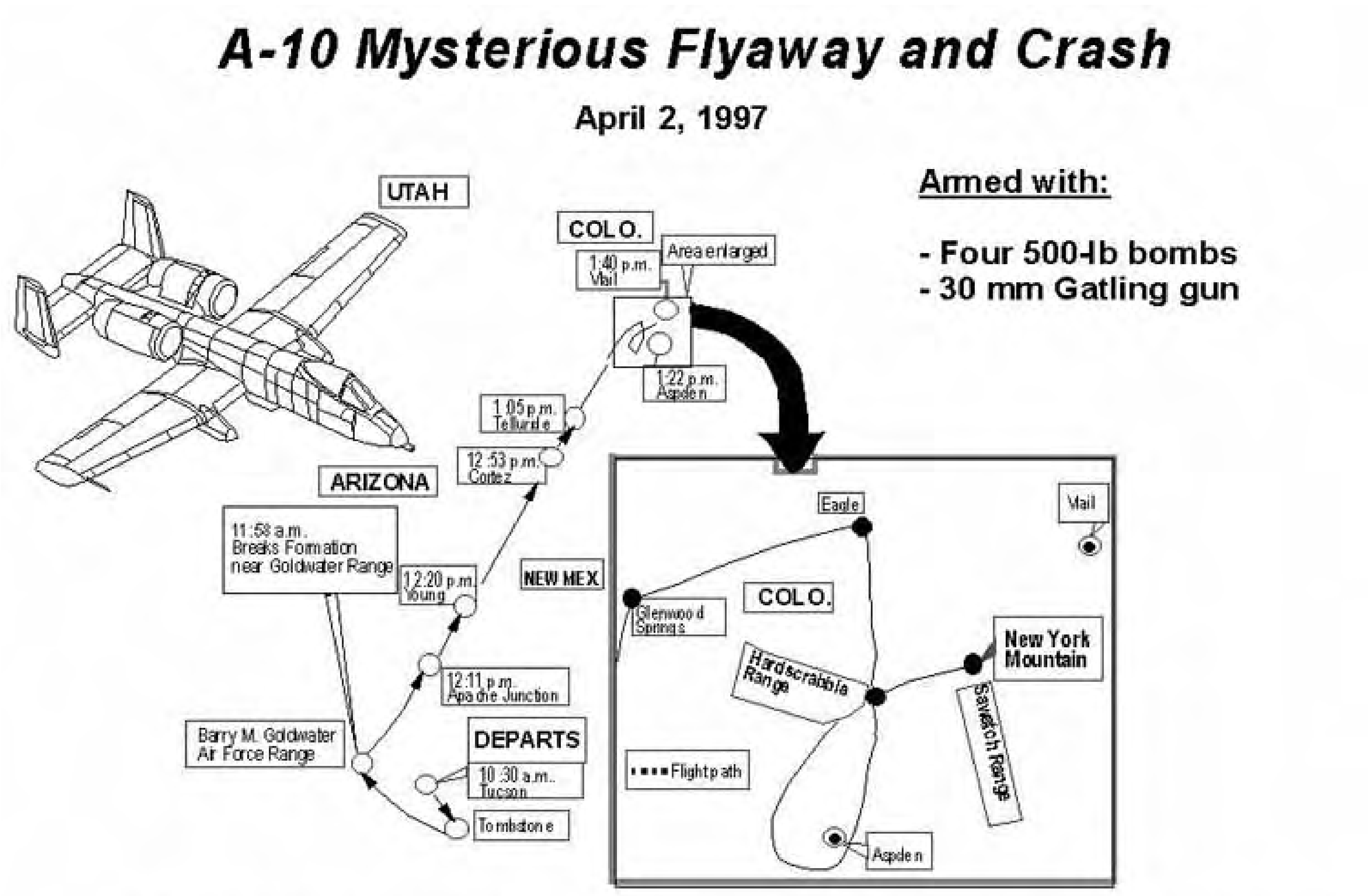
A-10 Mysterious Flyaway and Crash
April 2, 1997
Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun
UTAH
COLO.
1:40 p.m. Vail
Area enlarged
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
Glenwood Springs
COLO.
Hardscrabbia Range
New York Mountain
Savatch Range
Flightpath
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

occurred. Sad to say, it did not and it has not. It was lamely speculated that Svoboda probably looked over her shoulder like a rank greenhorn, and inadvertently dived her aircraft into the ground while looking back at the target. If there was one pilot in the USAF who would have not done that, it was Captain Svoboda. At any rate, the KGB determined that the U.S. scientific, intelligence, and governmental communities did not have the foggiest notion of what had occurred.

In private proprietary communications[9] we have explained further, to proper parties, exactly the reasons for these two tests, and what exactly was being tested, and what is its planned strategic use.

The Final Race for Direct Mind Control of the Entire Human Species

The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.

The KGB scientists also know that something very like Gaia— a collective unconscious mind for all species on earth—also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials. If one reads some later books by Kaznacheyev,[10] e.g , this human species unconscious mind aspect and the biosphere unconscious mind aspect emerges loud and clear, if one interpolates between the lines, so to speak. Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society"—communist style—is still a variation of feudalism (all systems tend to feudalism or some form thereof).

The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now. It is simply a matter of developing sufficiently complex equipment to be able to directly engineer the time-domain substructuring at sufficiently deep levels. It is a doable, although it is very, very complex and very, very difficult. On the other hand, they have been at it for decades. Even in the 50s the Russians already had the old LIDA device,[ll] capable of inducing a catatonic trance-like state in a mammal—either a human or a cat,[12] for example. Some versions of Russian mind control devices were used on Russian soldiers in the

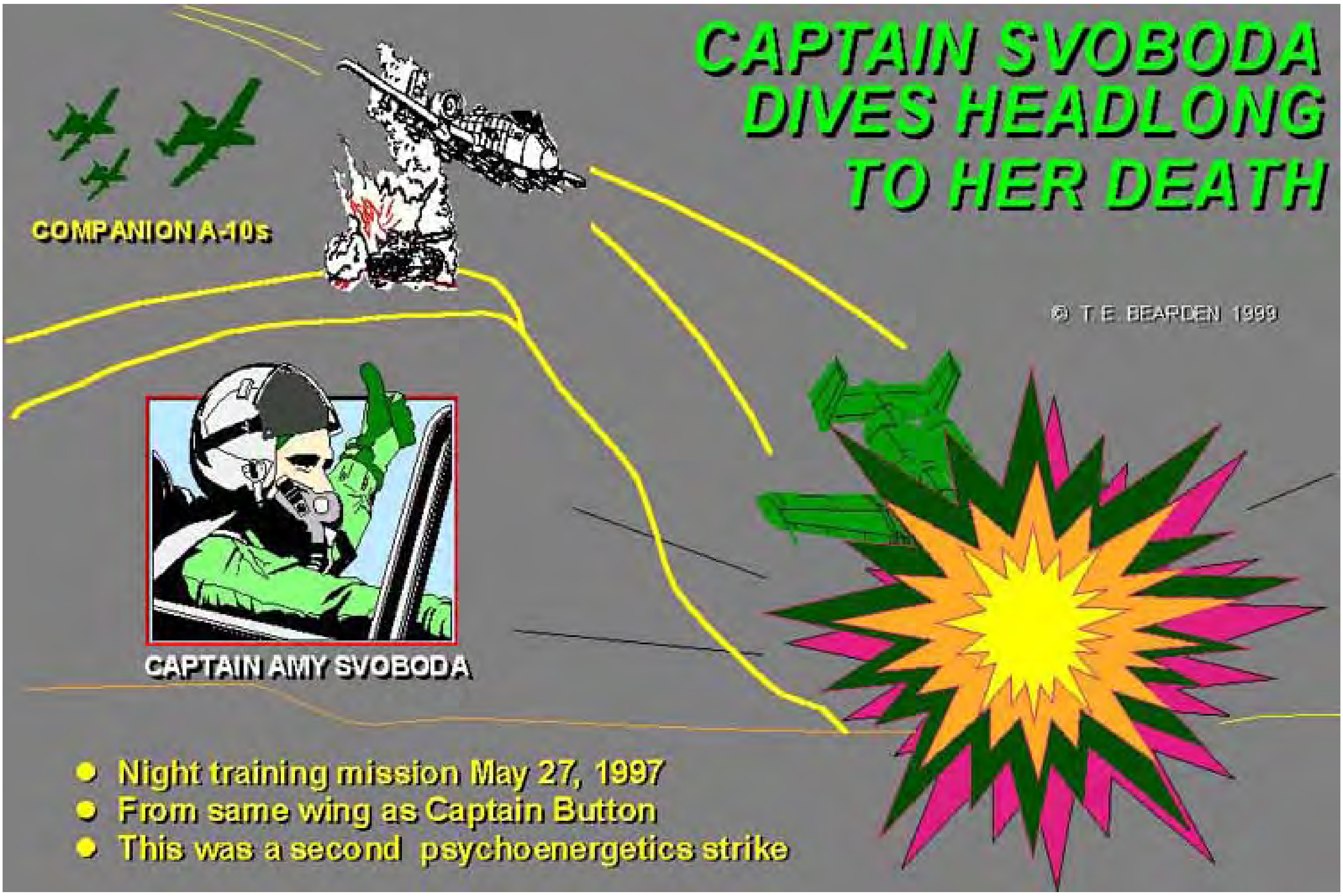
COMPANION A-10s
CAPTAIN AMY SVOBODA
© T. E. BEARDEN 1999

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

Figure 15. Captain Svoboda dives headlong to her death.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

Afghan War to condition them for performance of merciless acts.[13]

Potential U.S. Rogue Groups and Two Recent Examples

Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.

Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed.

We Certainly Have Proof of Such Rogue Activity. For Example, Here are Two Prominent Cases:

1. For three decades, scientists from the U.S. government, universities, and civilian contractors secretly

as a result of those ghoulish experiments. So what happened when this gruesome thing was revealed? Again President Clinton publicly apologized! The head of the Executive Branch did not turn the macabre matter over to his Attorney General to prosecute. How many of these scientists were indicted? Not one. How many are going to be indicted? Not one. Suppose again this had been done by a private doctor in his own private clinic. You get the point.

There is Little or No Punishment of Rogue Scientific Groups

Shockingly, the U.S. government at the highest level has directly shown (and these are not all such cases by any means!) that mass crimes against U.S. civilians, perpetrated by portions of the U.S. scientific community in direct conspiracy and in secret, will likely be condoned. The perpetrators will not be indicted, tried, or convicted.

Note the connection of the "rogue groups" thesis to the above incidents. Note how rogue scientific groups got away with it in both these cases. There may be other rogue scientific groups getting away with such things today. They should not be able conspired to treat unsuspecting human patients with whole body nuclear radiation, including some retarded children and cancer patients. Some of those patients—including some of the retarded children—died as a result. These illegal experiments were conducted in great secrecy, and the results were highly classified. Eventually these actions were revealed, and a Presidential investigation committee investigated. These experiments would seem to be little different from the WWII Nazi and Japanese experiments on human prisoners. Are those U.S. scientists who were responsible for those retarded children's' deaths any different from the Nazi scientists and criminals we executed at Nuremberg? So what happened when this terrible thing finally came out into the open? President Clinton publicly apologized! How many of the responsible scientists were indicted for murder? Not one. None is ever going to be. Suppose a doctor here in South Alabama, with his own private clinic, had secretly irradiated those human patients and retarded children with whole body nuclear radiation, over a period of 30 years, resulting in some agonizing deaths. How many state and federal agencies would have come after him with arrest warrants? Probably about 50 or so. Would he have been indicted, tried, convicted, and executed or sent to prison for life? Absolutely! Would the President of the United States have apologized? Not on your life; he would have turned it over to the U.S. Attorney General with a strong directive to to get away with it. But they can.

Perhaps a most startling additional part of those two incidents is that there was no great public outcry from the scientific community, deploring these murders and demanding that the criminal scientists be indicted and tried.

As the old saying goes, "By their silence they have convicted themselves." The organized scientific community, as a community, has shown that it has little or no ethics, and—while deploring any scientific murdering that "gets revealed,"—is not really interested in justice. In short, much of the U.S. scientific community may now have very little ethics left.

We are not talking about normal individual scientists, but the Big Science community. There is a whale of a lot of difference between the two. In the Big Science community, there can be and there are rogue groups. Lots of them. There is deep cover, deep classification. And there is very probably advanced mind control research and testing, be it legal or illegal. Hopefully most of it is legal and constrained. However, some of it is almost certain to be illegal and ill constrained.

Human beings are still human beings. All the good and evil is still there, regardless of the group. The stage settings change, but the cast of characters and the play never change.

It's sad, but 'twas ever thus. Hidden parts of our own governments—and other governments throughout the world—are no different from the old medieval groups, where nobles etc. were always plotting against the king, or using the king's power for their own nefarious end. The rogue groups today are no different from all the scheming and conniving groups that destroyed the Roman Empire. Great empires fall from within, not usually from without!

investigate and prosecute on criminal charges.

2. Again for three decades, scientists from (i) the U.S. Government, (ii) universities, and (iii) civilian contractors conspired to give poor syphilitic blacks in Tuskegee, Alabama a placebo while ostensibly treating them for syphilis. The purpose of the program was to deliberately observe and record the ravages of the disease to its fruition in their wracked bodies. So the scientists deliberately sat there and watched their brains rot and their bodies rot. They kept meticulous notes, of course. It was very scientific. And it was macabre. They "treated" some 400 blacks in the "program." Over 100 blacks died

Conclusion

This concludes our brief portrayal of the probable directions that mind control research has taken, in Russia and in the West. It is a novel and powerful area, and it poses both severe dangers and great benefits to all the peoples of the Earth.

Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction of helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth. Yet it can revolutionize medical science, education, communication, and psychology. Eventually it can engineer the mind and memory directly, beneficially. We foresee the day—perhaps 30 years hence—when education will be

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks one will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M. D. And so on. When that happens, then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.

We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on Earth. If we

446955, Oct. 23, 1998, forwarded to US. Patent Office on Nov. 4, 1998.

5. A very good introduction to nonlinear phase conjugate optics is given by David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Also see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83.

6. Pepper, Optical Engineering, ibid., p. 158.

7. See T.E. Bearden, Energetics of Free Energy Systems and Vacuum Engine Therapies, Tara Publishing, Internet node www.tarapublishing.com/books,July 1997. See also T.E. Bearden, "Energetics Update and Summary, " Part I, Explore, 7 (6), 1997, p. 60-67, Part II, Explore!, 7(7), 1997, p. 53-56; Part III Explore!, 8(1), I997, p. 53-56; Part IV; Explore!, 8(3), 1997, p. 56-63. See also T.E. Bearden, "Vacuum Engines and Priore’s Methodology: The True Science of Energy-Medicine: Parts I and II " Explore!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

8. E.g., see R.O. Becker; Charles H. Bachman, and Howard Friedman, The direct current system: A link between the environment and the organism, " New York State Journal of Medicine, Vol. 62,

do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium.

We are reminded of a poignant quote from Teilhard de Chardin:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have discovered fire."

May the nations of the world utilize the new mind engineering science to truly free all minds into a common love of, and respect for, their fellow humans of every color, creed, and circumstance. As de Chardin envisioned, may we use the new science to discover a new fire, burning clearer and brighter than anything that has gone before. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses. Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and

April 15, 1962, p. 1169-1176. See also R. 0. Becker; "The direct current field: A primitive control and communication system related to growth processes, "Proceedings of the XVI International Congress of Zoology, Washington, D.C., VOL.3, 1963, p. 179-183; —"The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 1lth International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p. 1753-1759. See particularly Robert O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. TBD, No. 73, Nov.-Dec. 1970, p. 169-198.

9. Bearden, T.E., Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page enclosure to CTEC Letter, "Saving the Lives of mass BW Casualties from Terrorist BW Strikes on US. Population Centers," distributed to selected high U.S. Government agencies and Legislative subcommittees, May. 4, 1998.

10. E. g. , Vlail Petrovich Kaznacheyev, V. I. Vernadsky's Study on Biosphere and NooSphere, 1989 [in Russian]. Vernadskys concepts included living matter, biosphere, and human-transformed biosphere (noosphere). Kaznacheyev writes of evolution and

energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

---

1. Owen Flynn, "Parametric arrays: A new concept for sonar, "Electronic Warfare Magazine, June 1977, p. 107-112.
2. Valerie Hunt, presentation to the U.S. Psychotronics Associations Annual Symposium, Bergamo Center, Ohio, 1988. Hunt presented epochal work showing chaos—and thus hidden order—in recorded ELF electromagnetic radiation from human bodies. Hunt split the received EM signals into two identical channels, time-delayed one channel by a very slight amount, then recombined (mixed) the nondelayed and delayed signals. The result was that attractors precisely indicative of chaos appeared. This positively proves that human systems produce ordered time-structuring of EM.

organization of the living matter as the biosphere of the earth. Without specifying the exact technical details, Kaznacheyev discusses the gist of much of what is in this present article. The present author has an autographed copy of Kaznacheyevs book referenced above.

11. Rabichev, Lev Yakovlevich; Vladislav Fedorovich Vasiliev; Alexandr Sergeevich Putilin; Tatyana Grigorievna Ilina; Petr Vasilievich Raku; and Leopold Pavlovich Kernitsky, "apparatus for the treatment of neuropsychic and somatic diseases with heat, light, sound; and VHF electromagnetic radiation," U. S. Patent. No. 3,773,049, Nov. 20, 1973. This is the U. S. patent for the LIDA machine, developed by the Russians prior to the Korean War. Additional patents or applications are as follows: Italy, #619,129, March 1961. Germany, #1,183,607, Dec. 1964. Great Britain, #826,766, Jan. 1960.

12. Stefan T. Possony, "Psy-War: Soviet Device Experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey’s investigation of the Soviet device called LIDA which is used to bombard human brains with radio waves in the 40-MHz region. Also reports on work by A.S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells

3. Years ago, I read an obscure reference to Russian research directly along that line, with successful development of just such a portable device. Unfortunately I lost the reference in a disastrous flood in my basement and have been unable to recover it.
4. T. E. Bearden, "Alteration of Materials by Wideband Multiwave Irradiation Combined with Selective Emission Manipulation," Invention Disclosure Document Number

are affected. See also "Soviet device uses radio waves as tranquilizer, " Associated Press release, Huntsville [Alabama] Times, May 19, 1983. Dr. Ross Adey tested the Soviet LIDA machine on a cat, and in two or three minutes the cat was sitting quietly, almost as though transfixed, and uninterested in its surroundings. The animal remained uninterested in its surroundings for some 20 to 30 minutes after the machine was turned off.

13. "KGB Used PIS to Program Soldiers, "Resonance, No. 30, Mar. 1996, p.29. Quoting: "During the 1970s, the Soviet KGB developed a Psychotronic Influence System (PIS) that was used to turn people into programmable "human weapons." Soldiers fighting in the Afghan conflict were trained using the system which employs a combination of high frequency radio waves and hypnosis. Commands are activated by code words or cyphers. The purpose of PIS was to render regular soldiers and members of the special forces incapable of feeling sympathy.

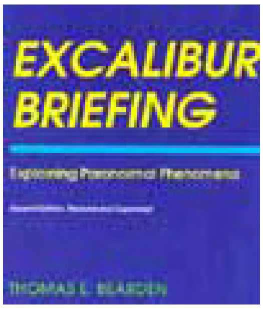
EXCALIBUR
BRIEFING
Explaining Paranormal Phenomena
THOMAS E. BEARDEN

THE NEUROPHONE

Another device that uses the new hyperspatial, virtual state, nested modulation technology (and has done so for seventeen or eighteen years) is Dr. Pat Flanagan's neurophone. With brilliant insight and intuition far beyond that of science at the time, Pat invented and patented the instrument by the time he was seventeen years old. The neurophone is a device that, contrary to all present theory and knowledge, will directly "pump the brain" and reproduce sound and information directly in the brain and mind system, without going through the auditory system at all. A simplified diagram of Pat's improved neurophone is shown in Figure 11.

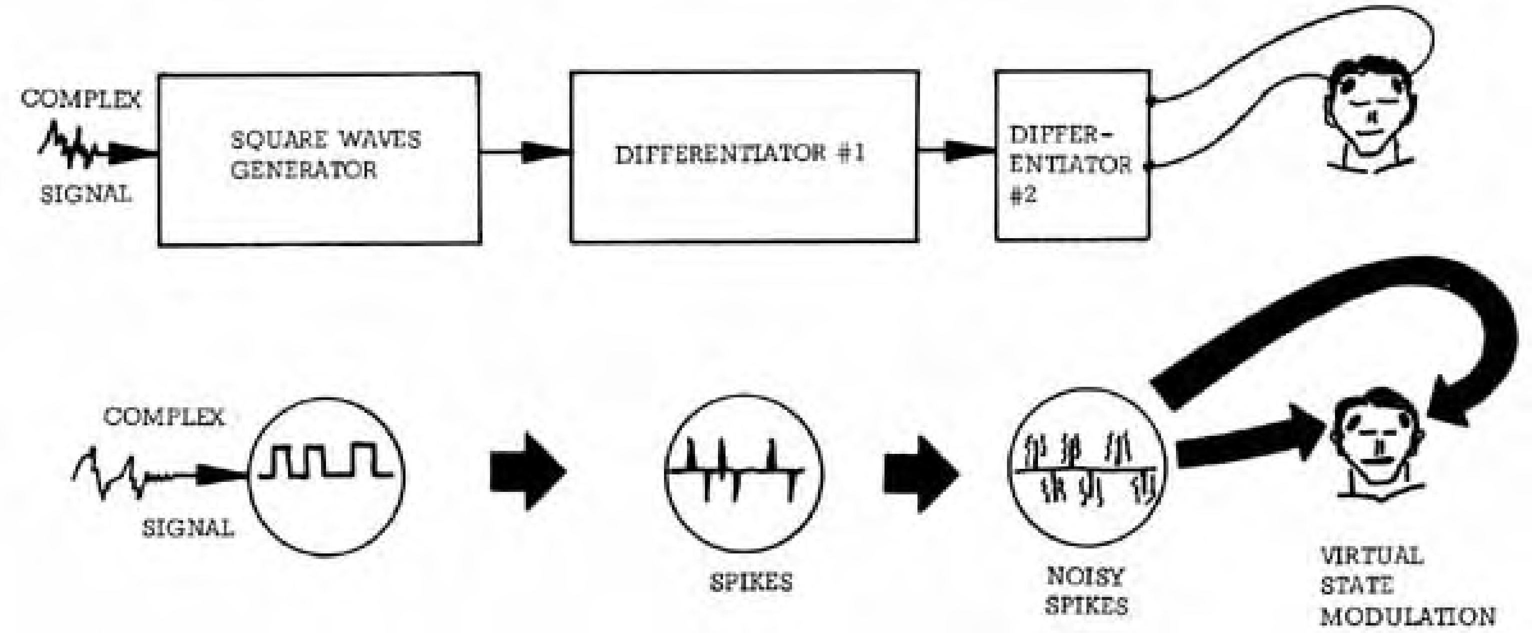
COMPLEX
SIGNAL
SQUARE WAVES GENERATOR
DIFFERENTIATOR #1
DIFFER-ENTIATOR #2
COMPLEX
SIGNAL
SPIKES
NOISY SPIKES
VIRTUAL STATE MODULATION

Figure 11. A simplified diagram of the improved Flanagan neurophone

Briefly, the device takes a complex signal, such as the sound of an orchestra playing a musical interlude, and electrically processes it as shown in the figure. First the signal is passed into a section that clips everything into a series of square waves, remarkably analogous to the sort of clipped waves Lisitsyn confirms are the information carriers of the human brain. Next the square waves are differentiated, yielding a series of sharp spikes (note that these spikes retain the pulse-time content of the clipped signal). These spikes are again differentiated, and since these are finite spikes with real nonzero rise times and decay times rather than theoretical constructs, a series of noisy spikes results from the second differentiator section. From here, the noisy spikes are introduced to special contact electrodes, one of which is normally placed on the forehead, while the other may be placed almost anywhere, including on the foot. Nowhere are any sound waves introduced to the head.

The square-wave clipper section reduces the complex signals, their overtones, and their complex modulations to square waves, retaining the temporal content of the wave mix but not the waves themselves. The first and second differentiators heighten or filter through the temporal content of the higher-order differentiations, that is, they serve as a band pass filter unit to accent the time keying of the neutrinic and mindfield portions or aspects of the signal. When these time spikes are then introduced across the body as pulsed voltages, they are modulated directly on the dendrite firings of the brain and nervous system, providing direct and pulsed modulation of the neutrinic and mindfield component channels of the mind-brain-consciousness-life loop itself. Thus the neurophone directly inputs information into the brain and nervous system, bypassing all the normal sensory systems that lie between the mind-brain loop and the outside environment.

Indeed, a similar development has been reported by two University of Missouri researchers. Dr. Donald York, a neurophysiologist, and Dr. Thomas Jensen, a speech pathologist, have recently reported identifying and decoding twenty-seven words and syllables in specific brain wave patterns and correlating these electroencephalographic patterns with both the spoken word and the silently thought word in about forty subjects. At present, Dr. York and Dr. Jensen are programming a computer with a brain wave vocabulary, to monitor and read the EEG of a stroke victim's brain and help stroke victims who have lost their powers of speech to communicate.

While all such developments can obviously be misused, their potential for assistance to mankind is enormous. Indeed, my own premise that brain linkage is possible and can be used to advance all mankind to the sixth stage of species evolution is slowly being proven. It is within our reach now to develop a multi-channel communications system that will directly link brains, minds, and consciousness into a single functional being; and I have already pointed out that the psychokinetic power of the emergent entity increases exponentially with the number of linked stages. The normal corpus callosum linkage of the two cerebral halves in each human head proves that brains, minds, and beings can be linked and integrated into one.

The Excalibur Briefing details the basic concepts of the theory involved in hyperspace-virtual state engineering, and thus in the engineering of mind-life-consciousness links to multicellular organisms here on earth. Pat Flanagan's neurophone proves that complex information can be implanted directly in the brain and mind, bypassing the normal isolating barriers. Drs. York and Jensen have shown that the brain wave does indeed contain recognizable, decodable information analogues. Lisitsyn has already reported measuring the number of independent channels involved and has documented the importance of the clipped brain waves as opposed to the sine-wave content. If we have the vision, we can still develop a brainlink and rapidly apply it to link sixty to one hundred persons, thereby opening a conscious, multidimensional "eye" for the collective human unconsciousness, gently rousing ZARG into wakefulness, and fulfilling human destiny without undergoing the psychotronic Armageddon planned by the Soviets. The time is short, but it can still be done if the funding, the vision, and the team can be established on a crash priority basis.

SOVIET "WOODPECKER" SIGNALS

But, as I have said, all this can be used for the detriment of mankind as well as for its benefit. Unfortunately, it appears that the Soviet Union has chosen to weaponize the effects on a global scale. A prime example is the "woodpecker" signals emanating from the USSR in the 5-30 megahertz region and interfering with communications around the earth.

These complex woodpecker signals appear to originate from two or three dozen powerful Soviet transmitters, each with a power estimated as high as 40 megawatts (Figure 12).

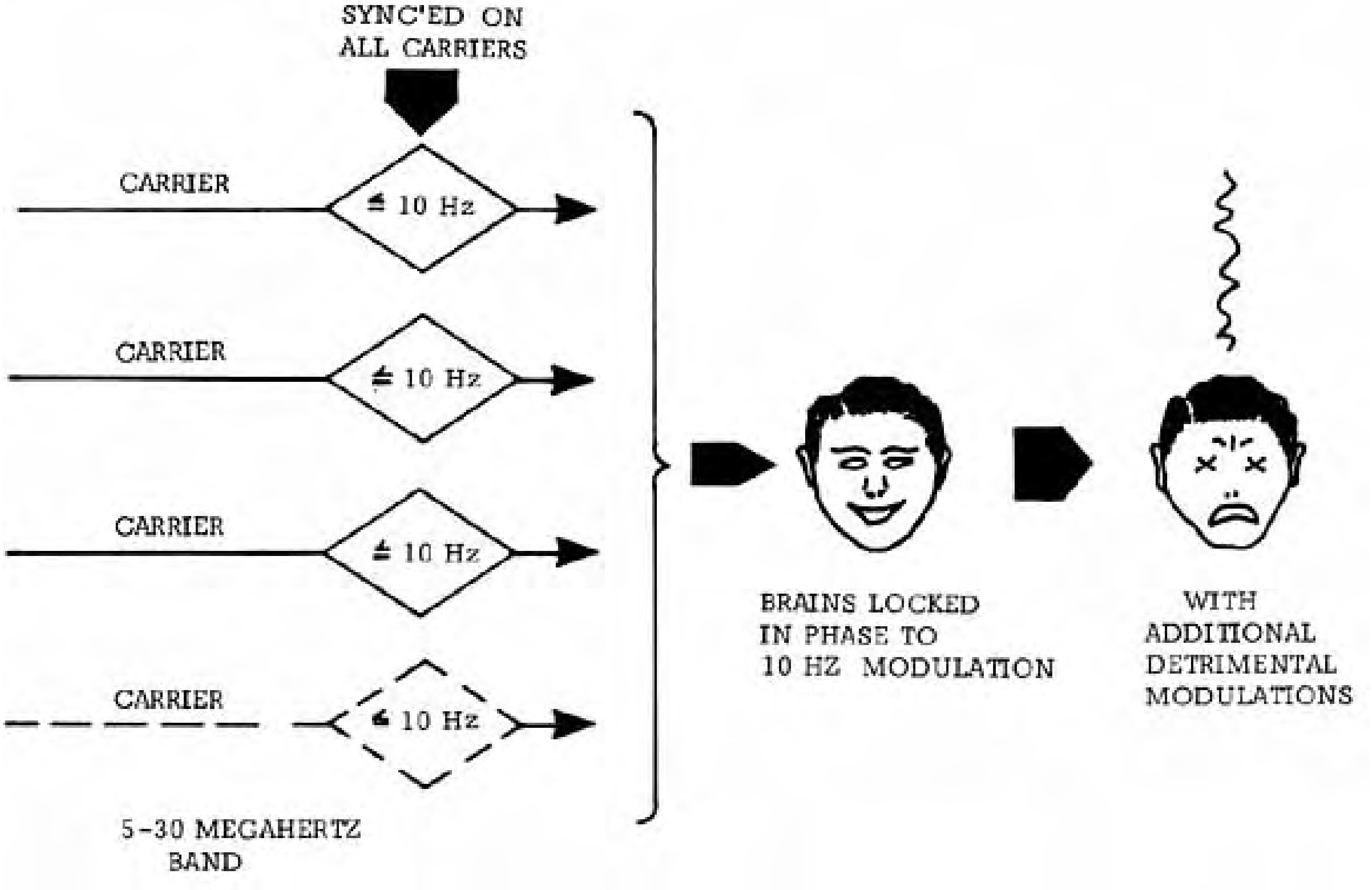

Figure 12. The Soviet "woodpecker" signals

The pioneering experimental measurements of these signals by Dr. Robert Beck and William Bise have shown just how deadly a potential may be possessed by the signals. These measurements have been performed in Eugene and Portland, Oregon; Los Angeles, California; Huntsville, Alabama; and several other locations. They have been particularly significant on the West Coast in and around Eugene, Oregon because of the presence of a direct current transmission line several hundred miles long, which has acted as a long wire antenna, picking up the signals and rebroadcasting them with appreciable gain in the vicinity.

Typically the signals may be found on, say, sixteen different carriers between 10 and 20 megahertz. Twelve of the carriers may appear normal, with normal side-bands, and the other four may have the carrier and both sidebands suppressed but still show the biologically significant modulation (for example, 10 hertz). On all sixteen channels a strong 10 hertz modulation may appear, all perfectly time synchronized and in phase channel-to-channel. The received signal from one of these carriers may be twenty-five to thirty times as strong as the earth's back-ground magnetic field, which is oscillating at approximately 7.0 to 7.5 hertz. Other complex modulation frequencies, many of them changing, are present on the various channels.

Normally, the brainwaves of mammals in an area are gently entrained by the normal oscillations of the earth's magnetic field and are oscillating along at, say, 7.5 hertz. Under continuous radiation from the Soviet woodpecker signals, a percentage (say 30 percent) of the mammalian brains may be captured by the 10 hertz modulation on the overpowering Soviet signals. These captured brains are now in forced oscillation and are riding along in phase with the 10 hertz Soviet modulation. In other words, a certain percentage of the human brains in an area will phase lock to the 10 hertz modulation, and that will effectively lock into those brains the sixteen carrier frequencies with their concomitant frequency mixes and complex signal modulations. If the signals are made much more powerful (for example, by focusing them onto one particular area or increasing the power of the transmitters), then a much higher percentage of brainwave entrainment can be accomplished in the targeted area. By simply adding on the desired modulation patterns, the Soviets can now pump material directly into the mind-brain-consciousness-life loops of the entrained brains. The weapon implications are enormous: raw emotion such as sheer terror or panic can be transmitted; death or disease patterns of all kinds can be transmitted; informational content (thoughts and ideas) can be impressed directly into the captured brains and minds and processed as if originating inside each brain itself. Indeed, as long as the Soviets keep the effects rather gentle, an entire population can be subtly influenced without governmental notice. Certain specific ELF frequencies can rapidly disable or even kill, and for wartime use those frequencies can be directly implanted in the captured brains themselves by the woodpecker signals. It appears that 1984 came a little early, and one can perhaps now understand Brezhnev's strange 1975 proposal to the SAL T negotiations that we should also consider outlawing the development of new weapons more frightful than the mind of man has ever conceived.

The signals also have direct application against strategic targets other than personnel. For example, by incorporating the correct nested modulation pattern, the positron kindling (electron squelching) pattern can be impressed. By beaming through the earth (with a circularly polarized signal or with other complex modulation forms), a beam with a strong neutrinic component, which will squelch electrons and which will interact with great preference for strong voltages or strong magnetic fields, can be directed at a target complex. In the targeted area aircraft and communications systems can be disabled; automobile, truck, and other ignition systems will fail; power transformers and large motors and generators will fail; power transmission systems with their associated high voltages will fail; and so on. Even independent, self-contained emergency power equipment, such as at military installations and strategic missile sites, will fail to operate. Further, to some of these signals, the earth and the sea are as transparent as glass, and so underwater nuclear submarines can be electrically disabled on station before their missiles can be fired, while strategic missile complexes can be electrically disabled before their deadly nuclear missiles can be hurled toward their distant targets. If some submarines and land complexes succeed in firing missiles before being disabled, these missiles can be electrically dudded and scrubbed from the sky over Soviet territory while still far from their intended targets.

So, as stated at the end of *Excalibur,* the players are on stage and in place. The final act begins. One thing is sure - mankind will never be the same again after seeing the play!

VACUUM ENGINES AND PRIORE'S METHODOLOGY: THE TRUE SCIENCE OF ENERGY-MEDICINE

Part I of II

© Copyright 1995 by T. E. Bearden

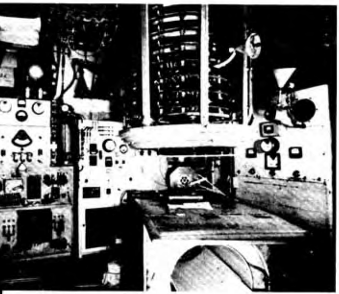

Photo 1. (to the right) Priore's laboratory. Workbench with animal cage is in the right foreground, under a large overhead assembly to hold the coil for producing the rippling magnetic field.

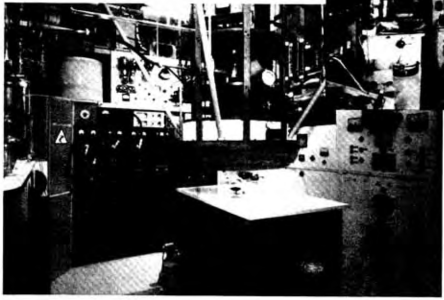

Photo 2. (below) Priore's laboratory. Workbench now has coil installed. The rippling magnetic field descended vertically through the patient. The magnetic field was modulated by the wave/antiwave output of 17 EM waves mixed in a rotating plasma in a giant mixer tube.

December 1, 1994

Overview

In the late 1960s and early 1970s, Antoine Priore and a team of leading French scientists demonstrated conclusive, total remissions of terminal tumors and infectious diseases in hundreds of laboratory animals. Their research was performed under rigorous scientific protocols and funded by the French Government. The approach employed very complicated mixing of multiple EM signals in a rotating plasma, and modulating the mixed output upon a very strong rippling magnetic field to which the body of the test animal was exposed. Complete remission of the treated diseases was obtained. In addition, the animals' immune systems were also restored to normal. Further, disease-specific antibodies were also created: a single drop of blood drawn from an animal that had been treated and recovered, and injected into the body of a second animal with the same terminal illness, resulted in total remission of the disease in the second animal *and its total recovery.* In the mid-70s Priore's work was suppressed, because of hostility of the oncology community, change of the French Government and loss of further funding, and complete inability of physicists and biological scientists to even hypothesize a mechanism for the curative results.

After a 10-year struggle to discover the Priore mechanism, the present author has proposed a sweeping new explanation for the exact Priore curative mechanism. It represents a totally new kind of medical therapy – one which can dramatically treat and cure a breathtakingly wide variety of diseases. The electrical physics is based on the Stoney, Whittaker, and Ziolkowski decomposition of the scalar EM potential into bidirectional EM wave pairs. Deliberate assembly of such a structure forms a *vacuum engine* of nested curvatures of local spacetime.[1] Priore unwittingly made such a hyperspatially structured vacuum engine, one with the special characteristic that it *time-reversed* (dedifferentiated) diseased cells back to their earlier normal state. The Priore approach not only produces remission of a specific targeted disease, but also fully restores the cells and "factories" of the patient's immune system because they too revert back to their previous healthy condition.[2]

---

[1] *In short, Priore found and utilized a method of electromagnetically engineering general relativity in the laboratory – something physicists had not done and have no knowledge of. Neither have they done it overtly today, and neither do they have knowledge of it today.*

[2] *We point out that chemistry is largely driven by charge and charge distribution. The new methodology essentially controls charge and its distribution. Hence it can control and order all the biochemistry.*

Becker et al have shown conclusively that very weak DC currents and pulsed magnetic fields can produce cellular dedifferentiation (change to a previous state) and redifferentiation (change forward to a new state). It follows that transcription (change of the genetics) must also be altered correspondingly. Becker was nominated for a Nobel Prize for his experimental and theoretical demonstration of the mechanism. He also proposed a fundamental electrical control mechanism to account for his revolutionary results.

Kaznacheyev et al have shown that almost any cellular disease or damage can be transferred electromagnetically from one cell culture to another under special conditions. Popp has theoretically and experimentally shown that an electromagnetic master cellular control system exists in the body, as indeed has Becker. Thus it is clear and independently demonstrated that (1) disease can be created or reversed electromagnetically, and (2) the fundamental cellular control system is electromagnetic in nature. We comment that, viewed as physical materials, the cells and their constituents are sufficiently nonlinear that their EM functions, responses, and effects transcend the present classical EM model. A modern addition to nonlinear electromagnetics is the methodology and results of nonlinear phase conjugate optics (NLO). Accordingly, understanding of the higher topology EM/bioeffects functioning requires application and understanding of NLO principles, *regardless of frequency band.*

We also extend the general relativity model by addition of higher topology EM. We incorporate the Stoney/Whittaker/Ziolkowski decomposition of the scalar EM potential into hidden bidirectional waves. This is then applied to structure and engineer the vacuum. The method is also capable of structuring and engineering both mind and matter.

As a case in point, the elements of the new EM bioeffects approach are applied to the Gulf War Syndrome – the strange conglomerate of illnesses and symptoms developed by many Gulf War participants, usually beginning shortly after their arrival back in the United States from the MidEast. Previously unrecognized biological effects of a dense EM signal environment are advanced and their causative mechanisms explained. These new "dense signal environment" bioeffects, coupled with other simultaneous biologically-active suppressants, allow a full explanation of the Gulf War Syndrome and its characteristics. Both normal and anomalous symptoms are explained.

The Priore approach extends and complements Becker's pioneering approach by allowing positive engineering control of dedifferentiation of cells and their redifferentiation. A cancer cell may be regarded as a dedifferentiation or partial dedifferentiation of a normal cell back toward a more primitive or "immature" geological form. In NLO terms, this is a form of phase conjugation (time reversal) of the cell back toward a previous state, or even its *ancestral* cellular form *(see figure 1 below).* By phase conjugating the cancer cell anew and amplifying the process, the diseased cell can quickly be reversed back to a normal cell.

It is possible, using the Priore technique, to phase conjugate cells in the body, including all of them simultaneously if need be. Further, conjugation can be made specific to the type of cell or tissue one desires to affect. One need not "know" the specific disease's "EM signal complex," one must simply "pump the cell correctly" where it and its parts are treated as pumped phase conjugate mirrors (PPCMs). In a higher EM topology, the "diseased-cell-as-a-PPCM" already possesses a special "standing signal wave" input that represents the erroneous *vacuum engine* associated with the change to the specific disease state. Pumping the cell *in this special EM topology* creates the precise, amplified *vacuum anti-engine* in this same special EM topology and renormalizes the cell.

Because of the special EM topology utilized, direct engineering of both mind and matter are provided by the new approach, which also represents a unification and extension of electromagnetics, quantum mechanics, general relativity, and biology. It should be capable of reversing most physical and mental disorders that assail a biological organism.

INTRODUCTION

In this paper we discuss the present method of utilizing EM energy in science and engineering, and also discuss a radically different second method that is beginning to emerge *(see Table 1 on following page).* The main text is an overview, with technical elaboration in the footnotes.

The first (Type 1) engineering method may be likened to a leaky pipe: We always pour in more energy than the work we get back out. We have to furnish the primary energy used to do the work from an external source. A few of our devices, such as a common heat pump, are open systems extracting and using some excess energy from an external source. These devices may have a coefficient of performance (COP) – work output of the machine divided by work done on the machine to run it – of greater than unity. However, the vast majority of our machines have COPs of less than 1.0. Many of them have COPs of less than 50%, and many have COPs of less than 30%.

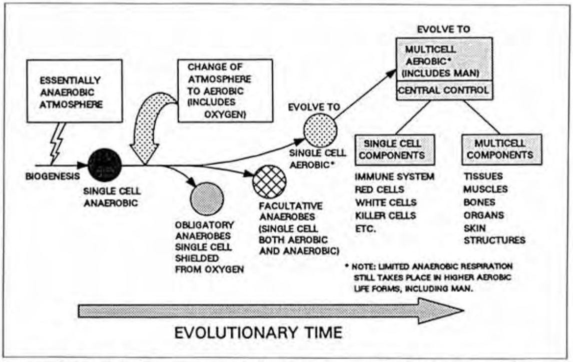
ESSENTIALLY ANAEROBIC ATMOSPHERE
CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)
EVOLVE TO
EVOLVE TO
MULTICELL AEROBIC* (INCLUDES MAN)
CENTRAL CONTROL
BIOGENESIS
SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC*
OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN
FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)
SINGLE CELL COMPONENTS
IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.
MULTICELL COMPONENTS
TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES
* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.
EVOLUTIONARY TIME

Figure 1. Cellular path taken in development of aerobic cells and multicellular bio-organisms. Dedifferentiation of a cell or small group of cells back before centralized cell control yields cancer. The body's control system orders the cell to dedifferentiate back toward an anaerobe, breaking its own control.

As can be seen, Type 1 engineering is very wasteful of energy! In a sense, the present struggle to attain room temperature superconductivity is a struggle to eventually have many more machines that attain a COP approaching 1.0, and so decrease the waste of energy in our

Table 1. Two Ways To Do Engineering.

- *Type 1* Engineering: Vacuum Quiescent, Use External Energy Upon Matter
  - Does not involve mind, thought, life, time streams
  - Time is not an observable
  - Must furnish all the energy
  - Western science uses this method.
- *Type 2* Engineering: Gate and Structure the Vacuum Flux Itself
  - Involves mind, thought, life, time streams
  - Time is an observable
  - Furnish only initial energy, and maintenance
  - Use vacuum's energy (vacuum engines)

Type 1 machines. Vacuum engines allow COPs of far greater than 1.0.

After several decades of painful progress in understanding the nature of the quantum mechanical vacuum, today it is accepted in physics that the vacuum itself is a seething inferno of electromagnetic energy. It is thus a legitimate source of energy, could we but extract and use it. This is now accepted both by leading physicists and by the U.S. Patent Office. Further, as of 1993 it has been clearly shown that, contrary to conventional objections, thermodynamics does not rule out the extraction and utilization of the vacuum's energy as heat and power to operate our engines.[3]

We have now reached one of the great scientific plateaus of history: (1) *The vacuum is energetic, containing an incredible density of electromagnetic energy in a violent flux;* (2) *at least in theory, it is possible to extract energy from the vacuum and utilize it;* and (3) *every particle of mass exists as a nonlinearity in a violent flux of virtual particle energy exchange with the vacuum. In short, every system is an open system, being violently "driven" by the vacuum flux, where normally the "driving flux" is in equilibrium or nearly so.*

This paper briefly describes a great new scientific breakthrough which is approaching as a result of the new understanding of the vacuum energy. This new scientific development will usher in the greatest change to humanity of all time *(see Figure 2).* It is characterized by the laboratory ability to: (4) Deterministically structure and gate the vacuum's fierce flux of virtual particle energy – even at a distance – and make the vacuum itself an ***engine*** to do any kind of work on an embedded physical system that we choose.[4] (5) Deterministically form a quantum potential, and utilize this potential to achieve essentially instantaneous action at a distance.[5] (6) Engineer the hidden inner structure of any potential; this includes also the potential for the Schroedinger equation, including Bohm's ***quantum potential,*** so that physical reality can actually be engineered prior to the collapse of the wave function and the occurrence of the quantum change. This is engineering and altering an entity in its *pre-reality state.* (7) Engineer and alter the flow of time and the laws of nature in a particular area, including the entire planet or solar system. (8) Simply and cheaply treat and cure essentially every present infectious disease and many other abnormalities. This includes cancer, leukemia, AIDS, malaria, sleeping sickness, arteriosclerosis, etc. – even mental disorders. It also includes reversing the effects of the aging process. (9) Directly engineer, form, create, heal, or modify mind, thought, long-term memory, life, etc.[6] And finally, (10) Directly engineer and alter the species quantum potential, an entire species, and the planetary quantum potential for all species and all living creatures on earth.

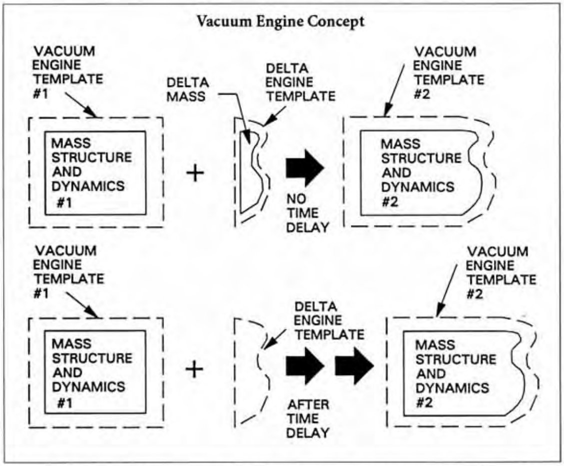
Vacuum Engine Concept
VACUUM ENGINE TEMPLATE #1
DELTA MASS
DELTA ENGINE TEMPLATE
VACUUM ENGINE TEMPLATE #2
MASS STRUCTURE AND DYNAMICS #1
+
NO TIME DELAY
MASS STRUCTURE AND DYNAMICS #2
VACUUM ENGINE TEMPLATE #1
DELTA ENGINE TEMPLATE
VACUUM ENGINE TEMPLATE #2
MASS STRUCTURE AND DYNAMICS #1
+
AFTER TIME DELAY
MASS STRUCTURE AND DYNAMICS #2

Figure 2. The vacuum engine concept. Addition of a vacuum engine (lower block diagram) alters mass and its dynamics after a time delay. Type 1 engineering uses the top level block diagram.

[3] *Daniel C. Cole and Harold E. Puthoff, "Extracting Energy and Heat from the Vacuum,"* ***Physical Review E,*** *48(2), Aug. 1993, p. 1562-1565.*

[4] *In other words, to perform vacuum engineering, extended from the suggestion of Nobel Laureate T.D. Lee* ***(Particle Physics and Introduction to Field Theory,*** *Harwood, New York; Chapter 25: Outlook, "Possibility of Vacuum Engineering," p. 824-828.). Lee recognizes that vacuum engineering can in theory be accomplished, but also that this has never even been tried.*

[5] *For a presentation of the mechanism for creating a quantum potential, see T.E. Bearden,* ***Gravitobiology,*** *Tesla Book Co., Chula Vista, CA, 1991, p. 33-36.*

[6] *The weapons developers are slowly grappling their way toward this new science, which the Russians call energetics. Application to the physical world at large is just referred to as energetics. Application of energetics to mind engineering is called* ***psychoenergetics.*** *Application to the physical body and its structures is called bioenergetics. In the U.S., development of mind and behavior alteration and control is now being dealt with in embryonic fashion under the subject heading of nonlethal warfare.*

This paper is an informal "quick note" to point out the new scientific plateau and introduce the second kind of science and engineering. The new ***vacuum engineering***[7] offers the greatest threat to humanity that has ever existed,[8,9,10] and at the same time offers the greatest promise of benefit.

For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth – a possibility indicated by Nikita Khrushchev in 1960.[11]

THE FIRST KIND OF ENGINEERING (TYPE 1)

First there's the normal way to do engineering (Type 1), which we are all taught as engineers and scientists. Here the vacuum/spacetime itself stays *inert* and out of our way.[12] We ourselves furnish energy into or onto the system we are trying to move or activate, spilling some of this energy all the while, and we struggle until we get the mass to move and the system to operate and perform work.

This "first kind" of engineering is essentially a continual *wrestling match.* Nature herself is the opponent. The engineer furiously struggles with nature, to force her to do the work he wishes to do. *This is the only* ***overt*** *kind of engineering presently done on earth.* It is primitive and severely limited in what it can do. In Type 1 engineering, the mind and thought are not affected since the *flow of time connectivity* and the *flow of mind connectivity* are not affected.

THE SECOND KIND OF ENGINEERING (TYPE 2)

Clandestinely, however, the Russians (actually, the KGB) in their secret advanced weapons labs, and three other nations not hostile to the U.S., have also weaponized a second kind of science and engineering, but kept it firmly a secret of the highest order. The second kind of engineering (Type 2) is dramatically different, far more powerful, and essentially unlimited in what it can do. Some of the simpler elements have spilled into the open press. For example *(see Figures 3 & 4 on next page)*, Vlail Kaznacheyev and his colleagues have shown, in thousands of meticulous experiments, that almost any disease or damage to cells can be induced in an isolated cell culture by electromagnetic radiation from the diseased or damaged culture, if a full harmonic interval (infrared to ultraviolet) is open for electromagnetic (photonic) exchange between the cultures.[13] The present author's proposed mechanism for the Priore and Becker results also explains the Kaznacheyev results.[14] It also explains the decades-long Russian induction of health changes and diseases in the microwave radiation of the U.S. Embassy in Moscow.[15]

In Type 2 engineering, the background vacuum/spacetime itself is deliberately altered, structured, and made asymmetrical. One then uses this *struc-*

---

[7] Terminology after Nobel Laureate Lee, ibid., p. 383. Lee points out that the microscopic structure of the scalar vacuum field is not utilized. The particulate flux structure can indeed be engineered, but this has never even been tried by conventional physicists. An excellent coverage of asymmetries and observables is given on p. 181-188. E.g., whenever a symmetry is spontaneously broken, a nonobservable (virtual entity) has become observable. There thus exist vacuum engineering mechanisms for the direct translation between virtual and observable states. Indeed, a whole new vista has opened in physics – some are calling it a "new physics" – due to the discovery that it is possible to directly amplify the virtual state into observable state. E.g., see Serge Haroche and Daniel Kleppner, "Cavity quantum electrodynamics," **Physics Today**, Jan. 1989, p. 24-30 for an overview. In 1979 the present author foresaw, albeit crudely, the emergence of virtual state engineering by means of a possible hyperspatial (virtual state) amplifier. See T.E. Bearden, "Virtual state engineering and its implications," National Technical Information System, Report AD-A065762, 1979.

[8] The Russian weapon engineers model the vacuum as a virtual EM dipole plasma, rather than as a simple particle flux; e.g., see A.V. Chernetski, "Processes in Plasma Systems with Electric Charge Division," Georgi Pleckonov Institute of National Economy, Moscow, 1989 for an introductory theoretical exposition of the basic model with respect to physical energy systems.

[9] Use of dipoles focuses immediately upon the hidden electromagnetics. Application of plasma theory then exposes the higher nonlinear EM aspects and the characteristics of a higher topology EM theory. Interestingly, Priore used all the elements fitting this theoretical model, employing a giant plasma-filled tube for multi-signal EM mixing, to produce "hidden EM" vacuum engines.

[10] Decades ago, the Russians also applied the virtual EM dipole plasma vacuum to the problem of direct engineering of mind, behavior, and cellular control functions. For the fact that they succeeded in decoding the mind/brain/s operations, see S.K. Lisitsyn, "New approach to the analysis of electroencephalograms," **Problems of Bionics (Selected Articles)**, p. 16-25, National Technical Information System Report AD 730045. See also the various references by Kaznacheyev, sanitized by the Soviet State at the time to prevent revealing the actual primary mechanisms.

[11] For a partial discussion of Khrushchev's statement of the impending availability of such new Soviet superweapons, see Max Frankel, "Khrushchev says Soviet will cut forces a third; sees 'Fantastic Weapon'," **New York Times**, Jan. 15, 1960, p. 1.

[12] Technically the vacuum is always highly active, in violent flux exchange with all particles of mass. However, in Type 1 engineering this exchange activity is in equilibrium. So action and reaction are balanced in all directions. With no observable asymmetry, no single net force accumulates for the observer to see, and the vacuum appears to be an "emptiness."

[13] See Vlail Kaznacheyev and L.P. Mikhailova, **Ultraweak Radiation in Intercellular Interactions,** [in Russian], Novosibirsk, 1981. Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," **Psi Research**, 1(1), Mar. 1982, p. 47-76 contains an English translation of much of the information referenced in Kaznacheyev's 1981 book. See Vlail Kaznacheyev et al, "Distant intercellular interactions in a system of two tissue cultures," Psychoenergetic Systems, 1(3), Mar. 1976; "Apparent information transfer between two groups of cells," **Psychoenergetic Systems,** 1(1), Dec. 1974; V.P. Kaznacheyev, "Information function of Ultraweak Light Flows in Biological Systems," in **Problems in Biophysics**, Novosibirsk, 1967, p. 7-18 [in Russian]; V.P. Kaznacheev and A.V. Troftmov, Cosmic Consciousness of **Humanity:** Problems of New Cosmogony. Elendis-Progress, Tomsk, Russia, 1992.

[14] See T.E. Bearden, Gravitobiology, Tesla Book Co., P.O.B. 121873, Chula Vista, CA 91912; 1991. See also Bearden, **Cancer and the Unresolved Health Issues in the Biological Effects of EM Fields and Radiation**, Tesla Book Company, 1993.

[15] The Soviets radiated the U.S. Embassy in Moscow with weak microwave radiation for decades, starting in the early 1950s. Anomalous health changes and diseases were induced in U.S. Embassy personnel. Three Ambassadors eventually died of leukemia-like illnesses of mysterious nature. The results of the Johns Hopkins scientific study of the Embassy radiation problem also provide strong evidence for the explanation, since health changes and diseases only occurred in those areas of the Embassy where the EM force fields from the radiation were zero. In other words, induced health changes occurred only in personnel in those Embassy areas where the potentials from the radiation had zero gradients, so that the potentials were persistent. This 100% correlation to persistent potentials exclusively, is exceptionally strong evidence that the EM radiation was the actual cause. If independent of the EM radiation, health changes would not have precisely correlated to any component of the radiation! The Johns Hopkins scientists were quite competent, but only considered classical EM theory (CEM) in their assessment. CEM assumes that, when the force fields are zero, the electromagnetic actions are zero. That of course has been disproved in quantum mechanics, which has shown theoretically and experimentally that the potentials, not the force fields, are the primary causes of all EM phenomena. [See Y. Aharonov and D. Bohm. (1959) "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491; Bertram Schwarzschild, "Currents in normal-metal rings exhibit Aharonov-Bohm effect," Physics Today, 39(1), Jan. 1986, p. 17-2.] Even when the force fields are zero, the potentials may remain and provide real changes in charged particle systems by scalar interferometry. Indeed, the force fields do not even exist in vacuum, as CEM assumes. The Johns Hopkins scientists concluded – based on the erroneous CEM model they applied – that since no force fields were present where any of the health changes occurred, the cause could not have been the weak microwave radiation. Had it not been the radiation, some of the health changes would inevitably have occurred where the EM force fields were nonzero. The investigators' results falsified their own study's conclusion, by simple expectation mathematics. Had they known of the SWZ structured potential and its possibilities, they would have realized that by showing 100% correlation to the force-free potentials, they had actually proven that it was indeed the radiation that was the cause, and the force-free potentials component was the actual culprit The team was excellent; its tools failed it. For a complete discussion, see Bearden, Gravitobiology, 1991, p. 27-32.

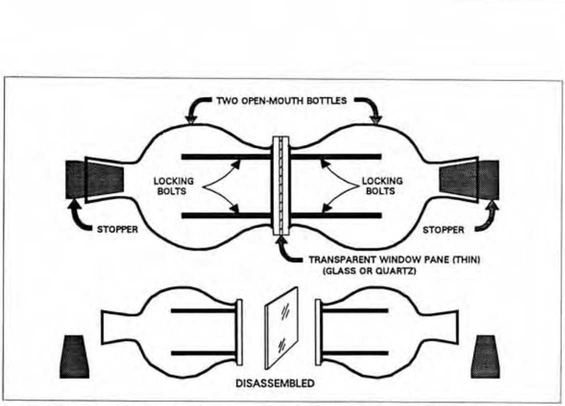
TWO OPEN-MOUTH BOTTLES
LOCKING BOLTS
LOCKING BOLTS
STOPPER
STOPPER
TRANSPARENT WINDOW PANE (THIN) (GLASS OR QUARTZ)
DISASSEMBLED

Figure 3. Kaznacheyev's two-chamber apparatus. A cell culture is divided into two parts. Left ones are damaged or killed. Effect is replicated in right cells through windowglass, but not quartz.

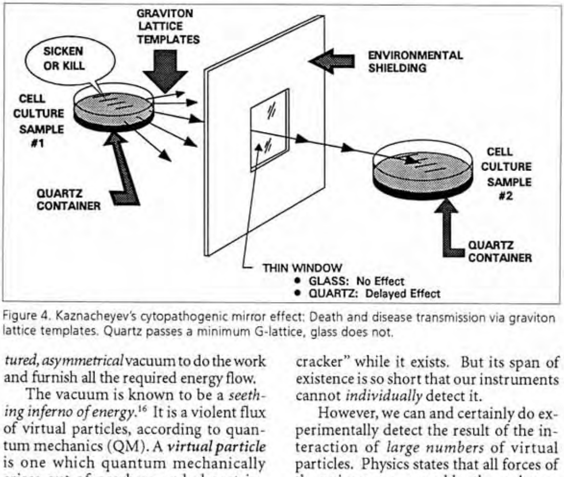
GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
ENVIRONMENTAL SHIELDING
CELL CULTURE SAMPLE #1
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
QUARTZ CONTAINER
THIN WINDOW
• GLASS: No Effect
• QUARTZ: Delayed Effect

Figure 4. Kaznacheyev's cytopathogenic mirror effect: Death and disease transmission via graviton lattice templates. Quartz passes a minimum G-lattice, glass does not.

tured, asymmetrical vacuum to do the work and furnish all the required energy flow.

The vacuum is known to be a *seething inferno of energy.*[16] It is a violent flux of virtual particles, according to quantum mechanics (QM). A *virtual particle* is one which quantum mechanically arises out of nowhere and almost instantly returns to nowhere. It may be highly energetic and "hotter than a firecracker" while it exists. But its span of existence is so short that our instruments cannot *individually* detect it.

However, we can and certainly do experimentally detect the result of the interaction of *large numbers* of virtual particles. Physics states that all forces of the universe are caused by the exchange of such virtual particles. It is also theoretically and experimentally proven that every particle of mass is in a violent virtual particle flux exchange with the surrounding vacuum. Several experiments directly show this – the Lamb shift for one.[17] The Lamb shift is largely due to fluctuations of the hidden virtual flux exchange with the vacuum, and involves a local energy density far greater than the surface energy density of the sun's surface.

VACUUM ENGINES: ORDERING AND UTILIZING THE VACUUM VIRTUAL PARTICLE FLUX

In Type 2 engineering, one structures and orders the violent virtual particle flows of the fundamental vacuum/spacetime (VST) itself.[18] Accordingly, it alters and orders the vacuum's violent energetic exchange with any mass particles and physical systems in the area. One *oneself* does not input the energy that does the actual work on the material system located in that region of VST. Instead, one only inputs the energy necessary for the *gating and directing* of the VST flux energy in that area. This structuring and gating of the virtual particle flux forms an organized, energetic pattern which I have dubbed a ***vacuum engine*** (after Nobel Laureate Lee's recognition that the vacuum itself could in theory be engineered, but that such had never even been tried).

This ***vacuum engine*** now does the work and furnishes all the working energy. We just have to pay to set up the *gating mechanism and the template* in the beginning, and continue to pay a little bit to sustain them. Literally we have created a stupendous *energy amplifier*.[19] By continuing to expend a little maintenance and steering energy, we can gate and utilize the energies of an enormous engine which can apply incredible energy densities, even at a distance. In principle it can apply anything from a feather's touch up to the explosion of a star. It can be used to engineer anything from a single particle or wave, to the atomic nucleus, to a planet or a sun.[20]

---

[16] *For a good overview of the active vacuum, see R. Podolny, Something Called Nothing: Physical Vacuum: What Is It?, translated by Nicholas Weinstein, Mir Publishers, Moscow, 1986, revised from the 1983 Russian edition. See also I.J.R. Aitchison, "Nothing's plenty: The vacuum in modern quantum field theory," Contemporary Physics, 26(4), 1985, p. 333-391; Jack S. Greenberg and Walter Greiner, "Search for the sparking of the vacuum," Physics Today, 35(8), Aug. 1982, p. 24-32; Richard E. Prange and Peter Strance, "The semiconducting vacuum," American Journal of Physics, 52(1), Jan. 1984, p. 19-21; Walter Greiner and Joseph Hamilton, "Is the Vacuum Really Empty?", The American Scientist, 68(2), Mar.-Apr. 1980, p. 154-164; Timothy Boyer, "The Classical Vacuum," Scientific American, Aug. 1985, p. 70. See also important papers by H.E. Puthoff: "Source of vacuum electromagnetic zero-point energy," Phys. Rev. A, 40(9), Nov. 1, 1989, p. 4857-4862; "Gravity as a zero-point fluctuation force," Phys. Rev. A., 39(5), Mar. 1, 1989, p. 2333-2342; and "The Energetic Vacuum: Implications for Energy Research," Speculations in Science and Technology, 13(4), 1990, p. 247-257.*

[17] *W.E. Lamb, Jr. and R.C. Retherford, "Fine structure of the hydrogen atom by a microwave method," Physical Review, Vol. 72, 1947, p. 241-243. Demonstration of the displacement of one of the fine structure electron states of the hydrogen atom by a frequency of about 1060 megacycles played a central role in the development of quantum mechanics, and earned Lamb the Nobel Prize. The observation makes the reality of dynamic vacuum fluctuations inescapably evident.*

[18] *Since the vacuum can be altered, we now must speak of various vacuua, not just one. "A" vacuum, "a" spacetime, and "a" scalar potential are synonymous. "The" ambient vacuum at any local point is a maelstrom of the scalar (hidden multiwave) interferometry of potentials from every charged particle in the universe (adapting Puthoff's Cosmological Feedback Principle). This scalar EM interferometry creates the EM fluctuations of the quantum mechanical vacuum, by Whittaker 1903 & 1904.*

Engineering General Relativity & Nesting Spacetime Curvatures

We actually engineer general relativity (GR) itself, since the Stoney,[21] Whittaker,[22] and Ziolkowski[23] (SWZ) papers point the way to how to do just that *(see figure 5)*. In short, by using the SWZ approach, one is able to utilize electro-magnetics in a new way, to directly form these curvatures of local spacetime. We also *nest* these curvatures to form a ***template*** of structured multiple curvatures; this template constitutes the deliberate format of the vacuum engine *(see figure 6)*.

Indeed, this is a radically different concept of spacetime structure. It is well-known that, whenever one applies a local hidden variable theory, the GR theory of spacetime structure must be radically altered.[24]

We stress that, in Type 2 engineering, one directly extends, applies and engineers general relativity, but by a novel use of electromagnetics. As can be seen, establishment science is not doing that at all, but is doing only Type 1 engineering. Literally anything one wishes can be engineered in the Type 2 fashion, at least in theory, including transmutation of elements, materialization and dematerialization of matter, free energy electrical power sources, and antigravity. However, Type 2 engineering is even more powerful than that. It also is the ultimate form of virtual reality engineering, for it allows the direct engineering of both physical reality and the mind itself.

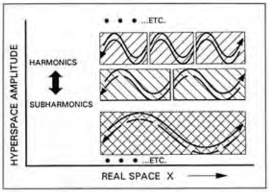

Figure 5. Stoney/Whittaker/Ziolkowski internal EM biwave structure of the scalar EM potential.

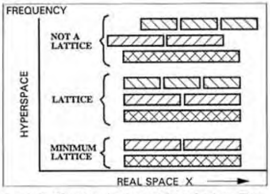

Figure 6. Graviton lattices and nonlattices. Templates require G-lattices. Minimum lattice is one harmonic interval.

Mind, 3-Space, Hyperspace, Mind Flow, and the Flow of Time

Mind exists in time, but not in our familiar 3-space. Mind does not exist spatially in the same 3-space in which we model physical phenomena.[25] Mind occupies (to the physicist) only one dimension – the time dimension – of Minkowski's 4-space of physical matter. This means that the attachment of mind to matter (in a living biological organism) involves a *mind flow* process, hidden in and occupying the *time flow* process *(see figure 7 below)*.

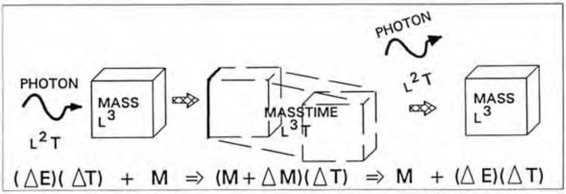

PHOTON
PHOTON
L^2T
MASS
L^3
MASSTIME
L^3T
L^2T
MASS
L^3
$(\Delta E)(\Delta T) + M \Rightarrow (M+\Delta M)(\Delta T) \Rightarrow M + (\Delta E)(\Delta T)$

Figure 7. The photon interaction generates observed quantum change and the discretized flow of time. A particle usually connects to only a single delta T at a time. Every other particle is connected to that small increment of time also, as are all minds.

---

[19] *An example of such an actual amplifier was Floyd Sweet's vacuum triode –. a 6-lb., solid-state device containing special coils and two separated barium ferrite magnets. The magnets were specially activated by Sweet's proprietary process so as to powerfully self-oscillate at a chosen ELF frequency, such as 60 Hz. [Such self-oscillation of barium compounds are well-known, but at optical frequencies. Use of a photorefractive magnetic material that is also an insulator allowed Sweet to initiate self-oscillation at ELF frequencies.]. An input sine wave of 10 Volts and 33 microamperes then produced 500 watts of phase conjugate replica EM power, extracted from the self-oscillation energy between the vacuum and the magnets. The amplifier gain (COP) of this open system, self-oscillating, pumped phase conjugate mirror device was some 1,500,000. A successful antigravity test of the device was also performed by Sweet at my urging, and resulted in a controlled reduction of its weight on the laboratory bench by 90%. For as much as could be revealed of the operation and testing of Sweet's proprietary device, see Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC '91), Boston, Massachusetts, 1991, p. 370-375. Unfortunately nothing can be done with the device – e.g., because of formidable legal entanglements due to the conflicting agreements Sweet made with different financial backers. However, McLain & Wooten have produced a repicable, tested, overunity device & full construction details have been rebuilt on Internet.*

[20] *In the modern geometrical view, all individual field energies come from the conglomerate structure called "gravitational field." Transforming one field energy into another kind may be done by first transforming the initial field energy back into G-field energy, then re-releasing the G-field energy as the desired new type of field energy. Vacuum engines allow control and utilization of this process.*

[21] *G. J. Stoney, "XLVIII. On a Supposed Proof of a Theorem in Wave-motion, To the Editors of the Philosophical Magazine," Philosophical Magazine, 5(43), 1897, p. 368-373.*

[22] *E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355; "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proceedings of the London Mathematical Society, Series 2, Vol. 1, 1904, p. 367-372.*

[23] *Richard W. Ziolkowski, "Exact Solutions of the Wave Equation With Complex Source Locations," J. Math. Phys., 26(4), April 1985, p. 861-863. See also Ioannis Besieris, Amr M. Shaarawi, and Richard W. Ziolkowski, "A bidirectional traveling plane wave representation of exact solutions of the scalar wave equation," J. Math. Phys., 30(6), June 1989, p. 1254-1269. See also Michael K. Tippett and Richard W. Ziolkowski, "A bidirectional wave transformation of the cold plasma equations," J. Math. Phys., 32(2), Feb. 1991, p. 488-492.*

[24] *E.g., see Abner Shimony, "Events and processes in the quantum world," in Quantum Concepts in Space and Time, Eds. R. Penrose and C.J. Isham, Clarendon Press, Oxford, 1986, p. 192: "...local hidden variable theories may be possible only if relativity theory is replaced by a radically different theory of spacetime structure." (My comment is that this should not be surprising, since classical electromagnetics, quantum mechanics, and general relativity are in profound contradiction of each other, and cannot be unified as they are, in spite of decades of Herculean efforts by many excellent physicists).*

[25] *We may model a mind as a separate 3-space, and its functional operations as "physical" in that hyperframe. Thus a mind and its operations are hyperspatial and outside physical matter's 3-space. Both mind and matter 3-spaces share a common 4th (time) dimension. So it requires at least seven dimensions to model a living creature, eight to model a species, and nine to model all life on earth. The 4th dimension of the mindworld hyper-4-space coincides with the 4th dimension of the physical world's Minkowski 4-space. A priori, mind/matter engineering operations involve engineering and structuring the time stream. This allows engineering either physical reality or mental reality or both simultaneously.*

Normally, so QM tells us, the observation process destroys time, so that all "observables" are spatial only. So the completion of each observable change in a living mass system momentarily destroys the only connection that the mass had with its associated mind! It vanished in the so-called "collapse of the wave function."[26,27,28] Fortunately, little pieces of time are always being serially added and subtracted by the ubiquitous photon interaction process with every mass. Hence the body's particles of mass are continually being *connected to* and *disconnected from* little serial bits of the time stream, and from corresponding serial bits of the *mindstream* of that individual.

TYPE 2 DISRUPTION OF TIME AND MIND

Type 1 engineering does not disrupt or affect the flowing time stream. Hence it does not disrupt or affect the flowing mindstreams of biological systems in the vicinity.

This is not necessarily true of Type 2 engineering! Type 2 engines of appreciable power can and will disrupt and destroy the body's continual detachment from and re-attachment to its mind-stream. Type 2 engines will appreciably disrupt and alter the flowing time stream, and thus disrupt and alter the flowing mindstreams of all life forms in the vicinity. The size of the "vicinity" depends upon the size of the engine.

So to beneficially utilize vacuum engines, we do not strive for control of vast energies. We must use a surgeon's scalpel, so to speak, instead of a giant steam hammer. If one seeks weapons, on the other hand, one will strive for control of much greater vacuum engine energies. The KGB secret laboratories, e.g., have pursued this goal for over four decades. Several generations of Type II weapons have been developed and deployed, all under continued KGB control and operation. These Type 2 weapons could, in the words of Khrushchev's address to the Presidium in January 1960, "*...wipe out all life on earth if unrestrainedly used.*" Now one can easily see how that could happen: ***Any massive use of Type 2 weapons will substantially upset the time stream on earth, including all individual time streams. It will so disrupt all the mind streams of all earth's creatures that their connections to their physical bodies will be instantly broken.*** At least now we know why Brezhnev himself was so frightened of these weapons in 1975.[29]

Another unusual characteristic of Type 2 engineering is the *slowly decaying persistence of any special vacuum engine structure artificially created in the vacuum.* We may analyze this effect in terms of the "ghost forms" of quantum mechanics.[30,31,32]

ENGINEERING THE MIND, ITS OPERATIONS, AND ITS CONTENTS

The vacuum engine affects/curves/structures not only the three *spatial* dimensions of ST, but also the *time* dimension. Hence the vacuum engine approach can be used to engineer not only the observable physical matter (which is 3-spatial), *but also the mind and thought (which are time-like to the external observer).* Literally one can "get into the mind's connection channel" and alter the mind, its contents, and its operations at will by using appropriate vacuum engines.

By appropriately and selectively nesting the spacetime (ST) curvature structuring, one can even reach beyond the

---

[26] *Rigorously, the emission of a photon by the living mass destroys its time channel, thus collapsing the wave function and disconnecting the mind. Absorption of another photon immediately thereafter adds an increment of time back to the mass, forming* **masstime** *and reconnecting the time channel and hence the mind channel. Mass does not exist in time per se; masstime does. The temporary disconnection of the time channel by photon emission causes the wave function to collapse; a priori a wave cannot exist without some time available in which to wave! This in fact produces the observable mass from masstime. The problem is that the physicists have ignored the transport by the photon of* Δt *as well as* ΔE*. Thus they have not noticed that the ubiquitous photon interaction with a mass creates the continual addition and subtraction of time increments. This continual coupling and uncoupling of mass to and from the time dimension creates a* **discretized** *flow (passage) of time for that mass. This continual connection and disconnection of a mass to and from only a tiny little piece of the time channel is responsible for the past and the future not being observable; "being observable" is simply being connected to a single small* Δt *increment, not all the increments that have gone before or that will follow hence. When one looks at the corner of a room, the 3-spatiality of one's mass instrumentation system is why one can see length, width, and height but not the fourth dimension, time.*

[27] *Since the photon interaction – including the virtual photon interactions – can be engineered, obviously the passage of an object through the discretized timestream can also be engineered, electromagnetically. Not only can a photon, electron, or proton be time-reversed, but also an assemblage of photons, electrons, protons and neutrons can be time-reversed. Among other things, this means that a diseased living physical cell can be time-reversed (dedifferentiated) back to a previous healthy state, as shown by Priore and also by Becker. Kaznacheyev, as does Becker, also shows that healthy living cell can be "time-forwarded" (differentiated) to a new state, if bombarded with the weak vacuum engine templates (graviton lattice structures) for that state. The minimum graviton lattice (ordering of SWZ wavepairs) utilizes one harmonic frequency interval. Thus quartz will pass Kaznacheyev's cytopathogenic effect, since it passes IR to UV – one harmonic interval. Ordinary window glass does not pass a complete harmonic interval, hence cannot pass the minimum graviton lattice and destroys the disease-inducing vacuum engine template.*

[28] *Since the entire area is highly nonlinear, then nonlinear oscillator theory applies. Graviton lattices may therefore be activated harmonically or subharmonically. This means that "optical band" results can be initiated by analogous subharmonic input – e.g., microwaves or even sound waves. G-lattices work like music; one recognizes a C-chord regardless of the octave. In similar manner, vacuum engines (G-lattices) may be shifted by any number of octaves, with initiation of results then being delayed in the responding physical systems, as prescribed by nonlinear oscillator theory.*

[29] *See Christopher S. Wren, "Brezhnev Talk for Accord against 'Terrifying Arms", New York Times, June 14, 1975, p. 1, 11.*

[30] *Since statistically every possible change in the vacuum is constantly and continually occurring in each and every differential volume, no matter how small, then from time to time in that maelstrom any physical (and mental) template we care to visualize appears, exists for a very short but finite time, then disappears, then reappears, etc. Such a continual, untidy characteristic is called a ghost form, and usually downplayed or ignored in the teaching of QM because it is "unpleasant." However, anything and everything – no matter how bizarre and whether living or inert – has a small but nonzero "density of reality" in the vacuum flux. Everything that ever was or will be, and everything that never was and never will be, is also there. The point is, creation of a vacuum engine template for a ghost form adds energy to (activates) that form. Just as a Dirac "ghost form electron" can be lifted from the vacuum by adding energy, so can any other ghost form, be it "living" or "dead." The process is fundamental and not just for electrons. Adding energy to a ghost template, we increase the local energy density of that ghost form in all electrical potentials in the entire area. This structuring of vacuum spacetime extends radially in all directions, tailing off to zero at infinity. When the inputting energy is ceased, the energized/activated form will then gradually decay over a period of time.*

[31] *We may compare the activation, persistence, and decay of the ghost form to an accumulator fed by two inputs, one blue liquid and one yellow liquid, with an outlet pipe on the opposite side. Suppose only the blue pipe is turned on; the tank is filled with blue liquid and blue liquid comes out the outlet. Now we turn on the yellow valve, and yellow liquid also starts entering. Over a period of time, the liquid in the tank will gradually turn green, as will the fluid from the outlet. Now we turn off the yellow valve. Over a period of time, the fluid in the tank will gradually turn back to blue color, as will the outlet liquid. Energizing and decaying ghost template structuring of the vacuum works the same way.*

[32] *For inputs that have raised the ghost template near or above the observable (quantum) threshold, this "decay time" may be quite long. Golden once accidentally charged up his local area with such a local template, in a powerful experiment that lasted about eight hours. Materializations occurred, determined by the thought-form (steady ghost form) inputs from the minds of the unintentional participants. In addition, interference with the timestream occurred, as indicated by all local clocks – mechanical, electrical, pendulum, etc. – going substantially astray. The materializations decayed away in a couple of hours, simply dematerializing back into the virtual state when their average energy density level reduced past the quantum threshold. It required four days for the entire "ghost template" to decay away and for the clocks in the area to fully resume normal operation.*

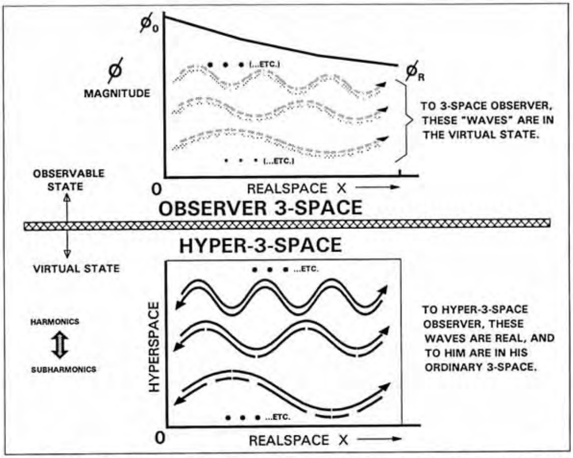

Figure 8. The 7-space EM potential. In observer 3-space (top figure), the potential is composed of a virtual particle flux. In hyper 3-space, the potential is a harmonic series of a SWZ bi-directional EM wavepairs.

ordinary 3-space and get directly into the higher dimensional space one may mathematically use to model the mind *(see figure 8)*. One can then directly "engineer" in that *pure mind* arena (in the *mental spacetime*, in the modeling sense). This capability, I believe, was responsible for Brezhnev's vivid fear of these weapons in 1975, and his statement to visiting U.S. senators that, not only must the nations of the earth ban nuclear weapons, but also the development of new weapons of mass destruction *"more frightful than the mind of man has ever imagined."*[33] Gromyko later introduced a draft treaty to that effect to the United Nations the same year, and urged the nations of the world to sign it.[34] Article one of the draft treaty stated that the Russians would identify the type of weapons when negotiations on the treaty started. *No one had the slightest notion what the Russians were referring to.* However, three other nations – not hostile to the U.S. – later did develop some of these Type 2 weapons. The Soviet Union was already engaged in operational testing against U.S. and other targets, with incidents mistaken for "accidents" or "anomalies" by the world's news agencies.[35] During the beginning of the 1980s the three nations began checking the Soviet Union by demonstrating their own use of the weapons against Soviet targets. Again the open press did not recognize the true nature of the incidents involved. At the close of the 1980s the triad *vigorously* "checked" the Soviet Union, clearly demonstrating their ability to devastate the Soviet Union should it attempt to devastate friendly nations.

So a special case of "deterrence by massive retaliation" came into being in the Type 2 weapons area as well as in the conventional and nuclear areas. Does the reader recall, e.g., the mysterious explosion of several Soviet missile ammunition storage sites in the space of a few months in 1983-84?[36] A single one of these explosions destroyed a third of all the missiles in the entire Soviet Northern Fleet. There were other major incidents whose true nature and cause went unnoticed.

Frustrated in the intent to dominate the world by surreptitious, delicate use of its Type 2 weapons, the economy of the Soviets simply fell apart. The Berlin Wall was torn down and the Soviet Union broke apart into independent regional nations.

In that upheaval of the Soviet Union, however, the KGB remained intact, well-funded, and still firmly in possession and control of the Type 2 superweapons, including development, manufacturing, deployment, manning, and operational employment.[37,38]

WHEELER'S PRINCIPLE AND ITS EXTENSION

Einstein has strongly pointed out the necessity to review foundations *(see table 2)*.

The great physicist Wheeler has a very nice statement on the curvature of spacetime. Briefly, he summarizes the gist of GR as *"Geometry tells mass how to move, and mass tells geometry how to curve."*[39] By "geometry" he is referring to *spacetime*. So we may immediately sub-

---

[33] *Wren, 1975, ibid.*

[34] *Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons, draft treaty introduced by Gromyko at the United Nations' 30th session of the General Assembly, New York, Sept. 23, 1975. The first article provided that the types of these new weapons would be "specified through negotiations on the subject."*

[35] *A long list of these incidents has been published by this author, through the Tesla Book Company. See T.E. Bearden, Soviet Weather Engineering Over North America, 1985; Fer-de-Lance, 1986; AIDS: Biological Warfare, 1988, Analysis of Scalar Electromagnetics, 1990, and Gravitobiology, 1991.*

[36] *See photograph, Jane's Defence Weekly (JDW), 2(3), July 28, 1984, p. 92; see also Mark Daly, "Goa and Goblet SAMs in Severomorsk explosion," JDW, 2(6), Aug. 18, 1984, p. 224; Derek Wood, "Soviets' northern fleet disabled... 'not viable' for six months," JDW, 2(1), July 14, 1984, p. 3; Derek Wood, "Six explosions in the past seven months," JDW, 2(1), July 14, 1984, p. 3; John Moore, "The aftermath of Severomorsk," JDW, 2(6), Aug. 18, 1984, p. 224.*

[37] *To survive as the leader of Russia, Yeltsin of necessity formed a pact with the KGB. KGB support insured that no coup of substance could be mounted against him, and that the KGB would insure his remaining in power. In return, Yeltsin saw that the KGB was well-funded, remained intact, retained all its secret weapons laboratories and secret weapon scientists and manufacturing facilities, and retained and continued to operate all its deployed, operational superweapons. The KGB changed its name and continued business on a less public profile. Publicly it officially turned a "much gentler face" toward the Russian citizenry. But essentially the KGB still controls most of the Russian military officer corps. Through the KGB/military connection the Communist Party can still accomplish a military coup and seize control of the country again, at least for the foreseeable future, when it chooses to do so. The Russian threat has not disappeared; it is simply hibernating.*

[38] *The Communists/Generals in control of the KGB and these superweapons are perfectly capable of turning the unfortunate dissolution into an excursion scenario of their long-planned major concessions offensive prior to ultimately striking the West with a clenched fist and destroying it. The communist aim has not changed one iota since 1917, although the communists have temporarily lost or relinquished control over the Russian civilian populace. For a description of this "29th move of the chess game," see T.E. Bearden, "Glasnost: 29th Move of a 30-Move Chess Game," Raum & Zeit, 1990.*

[39] *W. Misner, K.S. Thorne and J.A. Wheeler. (1973) Gravitation, W.H. Freeman and Co., San Francisco, 1973, p. 408. On p. 5 the principle is stated as "Space acts on matter, telling it how to move. In turn, matter reacts back on space, telling it how to curve." Thus space is already a vacuum engine that acts on matter, and mass-energy acts on space to curve it.*

Table 2. Einstein on Reviewing Foundations

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts, He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

Table 3. Wheeler's Principle and a Corollary

Wheeler

"Geometry tells mass how to move, and mass tells geometry how to curve."

By "geometry" is meant "spacetime." Translating: "Mass tells spacetime how to curve, and curved spacetime tells mass how to move."

Corollary

"Trapped EM energy tells spacetime how to curve. The structuring of that trapped EM energy tells the resulting curved spacetime how to nest its curvatures. This forms a vacuum engine. Its nested curvatures of local spacetime tell mass and energy how to structure and move. By forming vacuum engines deterministically, one may use them to affect, alter, and change local mass, energy, and mind as desired."

stitute synonyms to yield "Mass tells spacetime how to curve, and curved spacetime tells mass how to move."

We further modify and extend this latter statement, replacing mass by its trapped EM energy (this is consistent with general relativity) and adding nesting of such trapped energies. We also allow for altering the timestream, the mind, and the mindstream. With these modifications, a corollary to Wheeler's principle results as follows: *"Trapped EM energy tells spacetime how to curve. The structuring of that trapped EM energy tells the resulting curved spacetime how to nest its curvatures. This forms a* ***vacuum engine.*** *Its nested curvatures of local spacetime tell mass and energy how to structure and move. By forming vacuum engines deterministically, one may use them to affect, alter, and change local mass, energy, and mind in any manner desired" (see table 3).*

To gain a "feel" for spacetime curvature, one may visualize it as analogous to the variation of local pressure in a gas or fluid.[40] Using the earth's atmosphere as a model, we may precisely create pressure points and areas (lows and highs) throughout, and gradually move them at will. This control of highs and lows will allow control and direction of the atmospheric flows. Theoretically, the entire atmospheric flow of the earth can be steered and controlled, by using such an approach. In short one can directly "steer" the weather worldwide, or over a given continent. Present Russian development allows this weather control process to be done sufficiently to control and steer the great jet streams, with concomitant direct influencing of the weather.[41] *We have not had any truly "normal" weather over the U.S. since 1976.*

TYPE 2 ENGINEERING IS JUST BEYOND THE PRESENT VANGUARD OF SCIENCE

As can be appreciated, Type 2 engineering is beyond vanguard science. The forefront of orthodox science is still where Wheeler's original statement is, and does not apply the SWZ methodology nor nesting of spacetime curvature. Science still does not even try to create vacuum engines and perform vacuum engineering *(see table 4 on page 76).* Conventional scientists are still very firmly rooted in Type 1 engineering.

In ordinary GR, nowhere does there appear a precise definition of the term "gravitational field."[42] The G-field is just a *conglomerate of loose things*, and – by Sakharov's postulate – not a fundamental field of nature at all.[43] It already follows that, since "gravitation" is "geometry of spacetime," then "curvature" of spacetime is achieved by altering and changing the "conglomerates that comprise G-field" from their ambient conditions. The ability to structure and nest such curvatures also follows in straightforward fashion.[44]

PUMPED PHASE CONJUGATE MIRRORS AND THE PRIORE MECHANISM

The astounding thing is that Antoine Priore, using pure intuition, was able to arrive at an exact process for forming a special kind of beneficial vacuum engine![45]

We know today – but scientists did not know it when most of Priore's work was being done! – that a highly nonlinear material acts as a ***phase conjugate mir-***

---

[40] *It is actually the variation of the local energy density of the vacuum from its normal value.*

[41] *See T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. Available through the Tesla Book Co. In the last few years, the Russians have substantially increased their weather engineering capability and activity. Most of the transmissions from this system are now fairly good electrogravitational waves (coupled wave/antiwave pairs), with EM imperfections many dB below signal level. Hundreds of them are utilized when the system is actively employed over the U.S. By applying Whittaker 1904, these signals – invisible to the normal receiver – can be received and displayed.*

[42] *Misner, Thorne and Wheeler, Gravitation, 1973, p. 399.*

[43] *A.D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," Soviet Physics Doklady, Vol. 12, 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) is in Doklady. Akad. Nauk SSSR, Vol. 177, 1967, p. 70-71.*

[44] *General relativists already accept the vacuum as containing a 4-momentum-flux river. See Misner, Thorne, and Wheeler, Gravitation, 1973, p. 130-131. They treat the stress energy tensor as a machine (ibid., p. 131). The only difficulty is that they have not realized that, by applying the SWZ approach, much more complicated "vacuum machines" than they have considered can be built and utilized electromagnetically.*

[45] *See Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Explore!, 5(5-6), 1994, p. 97-110 for an excellent summary of the Priore Affair. The Priore group cured terminal tumors, sleeping sickness, and arteriosclerosis in lab animals and restored suppressed immune systems. Eminent, even world-renowned French scientists worked with him. The work was funded by the French government and published in leading French scientific journals. It was suppressed when the French government changed to a leftist government in the mid 70s.*

ror[46,47] *(see figures 9 & 10).* In other words, when one puts in a sine wave, e.g., in addition to producing harmonics the nonlinear material will also produce a strange, backward-traveling replica of the input wave. This "backward wave" has been well-known since the early 1970s. It is truly a *time-reversed* wave (Barus wrote the solution much earlier).[48] It will backtrack back through space, over the *exact* path taken by the input wave.[49,50] If the input wave is continuous, this "phase conjugate replica" (PCR) time-reversed wave will precisely appear everywhere in space that the input wave occupies.

The key here is "everywhere in space." It does not matter how tortuous and twisted the route taken by the input wave; the PCR will backtrack it and appear spatially superposed with it. This has been *experimentally* proven many times in nonlinear optics.[51] One concludes that the vacuum has a memory,[52] and the PCR wave can retrace the memory of the path previously taken by its stimulus.

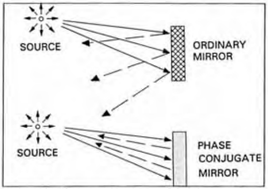

Figure 9. Phase conjugate refection vs. mirror reflection.

Figure 10. Pumped phase conjugate mirror.

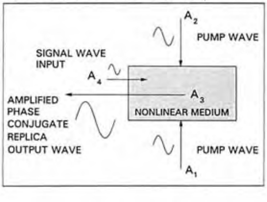

Another nice thing has been discovered and is used in the relatively new nonlinear optics. If one "pumps" the phase conjugate mirror with opposing waves (such as, e.g., bidirectional EM wavepairs), then up to all the energy in the pump waves is absorbed and re-emitted in the "crawfishing" PCR wave.[53] In other words, the stimulated PCM can be made into an amplifier of the time-reversed wave output, just by "pumping" it with the proper bidirectional wavepairs.[54,55]

ANOMALOUS "BIO-JAMMING" AND DEEP DIFFUSION BY EM SMOG

Pumping by EM waves or by simultaneously varying EM forces due to opposing streams of photons will produce a much higher probability of 4-wave mixing and multiphoton interactions. Simply put, this probability is a function of the sheer photon/signal density, along the lines pointed out – but not clarified – by

---

[46] Elsewhere we have pointed out that, for the primary EM interactions with a living system, we must employ an electromagnetics of higher topology than the conventional EM (either vector or tensor). Barrett has shown that a higher topology EM will involve and exhibit nonlinear optical functioning without necessarily employing optical materials. See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41. Thus if we wish to utilize higher topology EM functioning in and on the body/mind, or just to comprehend such functioning, the nonlinear optical processes will form part of that functioning. We must be familiar with phase conjugate optics functions, regardless of the frequency range we operate within, and we must examine the situation accordingly.

[47] In the conventional struggle to elucidate EM effects on biological systems, the failure to realize this higher topology principle has led to experiments that are often difficult or impossible to replicate, primary interactive mechanisms remaining a mystery, and the inability to definitively answer even the simplest question such as "Does power line radiation enhance the emergence of cancer and leukemia, and if so, how and under what circumstances?" Forty years of research by able and determined EM bioeffects scientists have thus failed us, because the model they so stringently apply is grossly in error. What they have thoroughly proved is that their fundamental model is incorrect and inadequate for the intended purpose for which it is being utilized. The solution is not to just continue more experiments in the same vein, but to revise the model itself. We have previously pointed out (Bearden, Cancer and the Unresolved..., 1993, ibid., p. 40) the present EM bioeffects model, the modification areas required, and how to ascertain when the adjusted model is correct.

[48] Carl Barus, "A Curious Inversion in the Wave Mechanism of the Electromagnetic Theory of Light," American Journal of Science, Vol. 5, Fourth Series, 1898, p. 343-348 gives the first known solution of Maxwell's equations for a time-reversed or "backwards-traveling" EM wave.

[49] According to the well-known "Distortion Correction Theorem" of nonlinear optics. For a precise statement of the theorem, see Amnon Yariv, Optical Electronics, 3rd Edn., Holt, Rinehart and Winston, New York, p. 500-501. See particularly Chapter 16: "Phase Conjugate Optics – Theory and Applications."

[50] For the space-borne laser, e.g., the U.S. uses iterative phase conjugation – known as self-targeting – of a booster-reflected laser signal to keep a powerful laser beam locked upon the same spot on a distant missile booster, rising during its launch phase, long enough to burn through the casing and destroy the booster, even 10,000 miles away. The process is known as phase conjugate shooting.

[51] A striking example is given by David M. Pepper (1986) "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 75. Possibly the best extant introduction to the subject of nonlinear optical phase conjugation is David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2) March/April 1982, p. 156-183.

[52] Without extended discussion we call attention to the importance of this finding with respect to homeopathic medicine and therapy. In fluids, obviously one proceeds from vacuum flux organization and memory upward through matter's organization (and hidden memory), through the individual particles of the nucleus, to the nucleus, to the atom, to the molecule, etc. Between the atom and the molecule levels we encounter hydrogen bonding actions in fluids, which is of great importance to the biochemistry, especially the protein chemistry. It is possible to show that the mind operations and the cellular control operations also SWZ structure the H-bonding activities of the body's aqueous fluid; it can also be shown that any chemical, drug, additive, and EM smog also structures and affects the H-bonding. However, further discussion of these two elements is beyond our scope for this presentation.

[53] Note that "pumping" is simply a expansion of the notion of stressing the material. Emission of the amplified PCR wave is a kind of piezoelectric release of the stress energy. The point is that physical stress of a material can be freely tapped and energy continually extracted, if that stress is varying. The McLain/Wooten device uses that principle.

[54] To create a normal PCR wave response to an overt EM signal wave input, nonlinear optics uses overt EM pump waves. For the hidden EM signal wave input – i.e., for the hyperspatial vacuum engine as input – one must use "hyperspatial" wave pumping. The "cell as a phase conjugate mirror" must be pumped by hidden (hyperspatial) EM waves inside a scalar EM potential. In the normal case, an amplified, normal EM PCR wave is produced in response to a normal signal wave input and normal pump waves. In the vacuum engine case, an amplified, hyperspatial EM PCR wave complex is produced in response to a hyperspatial signal wave input and hyperspatial pump waves.

[55] Note that the amplified PCR wave represents a reorganization and deterministic redirection of stress energy. Hence it is a time-reversed operation, and also a negentropic operation. Since the PPCM operates as an open system not in thermodynamic equilibrium, with input of excess energy in the form of the pumping, negentropic operation is permitted by nonlinear thermodynamics (of a system far from thermodynamic equilibrium) For a lay summary of such thermodynamics, see Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," in Paul Davies, Ed., The New Physics, Cambridge University Press, Cambridge, 1989, p. 316-47. For a more technical presentation, see G. Nicolis and I. Prigogine, (1987), Exploring Complexity, Piper, Munich, 1987. Prigogine received the Nobel Prize for his pioneering work in this field.

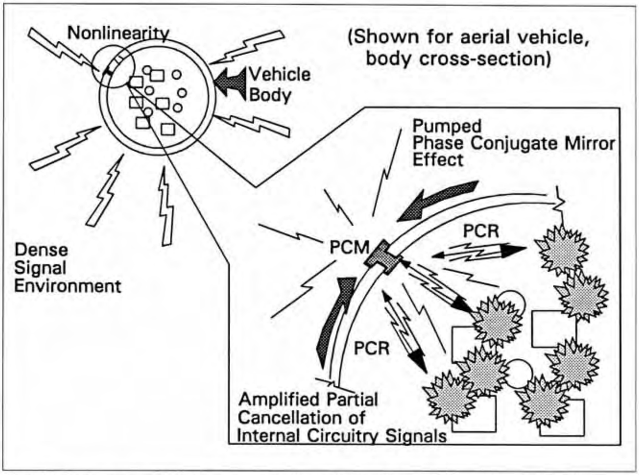
Nonlinearity
(Shown for aerial vehicle, body cross-section)
Vehicle Body
Pumped Phase Conjugate Mirror Effect
Dense Signal Environment
PCM
PCR
PCR
Amplified Partial Cancellation of Internal Circuitry Signals

Figure 11. Single versus multiphoton penetration effects. Phase conjugation effects become of significance, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, this penetrating throughout the body with weak partial cancellation jamming.

Devyatkov et al.[56,57] Thus dense electronic smog – even though it contains only extremely weak signals – increases the nonlinear optics type of interaction. This allows anomalously deep penetration and *diffusion* of special weak EM signals due to the "pumping" aspect of multiphoton interaction. The nonlinear cells and cell parts act as pumped phase conjugate mirrors (PPCMs) when subjected to multiphoton interactions.

Worse, the *signal wave* inputs to the PPCMs are the normal weak EM signals of the body itself. Hence the resulting PCRs from the PPCMs are superposing phase conjugates of the body's normal signals. This results in (1) a "diffusive" effect in the body of the PPCM formations, and (2) formation of very weak "EM cancellation" signals throughout the body, specifically targeted to the body's own control signals, including those to and of the immature immune cells in the bone marrow. In the bone marrow, "*partial-cancellation jamming*" *(see figure 11)* occurs in the normal control and development of the immune system cells as they are created. Slowly errors and degradation damage accumulate with time, including over the years. This cumulative immune system damage combines with other damage effects due to pesticide residues, pollutants, etc. The combined degradation results in long-term cumulative damage to the immune system and its cells. Increasing errors in the immune system's functioning eventually result in appearance of immune system disorders and increasing vulnerability to

Table 4. Mechanism for the Priore Therapy

- Produced Specific SWZ-Structured Scalar Potential
  -17 EM frequencies mixed in a rotating plasma
  - Phase conjugates added by the plasma
  - Formed vacuum engine with hyperspace pump waves
  - Modulated onto a rippling magnetic field which carried it to every cell and all its parts
  - Jyperspatial pumping of cells/parts as PPCM's
- PPCM's created amplified anti-engines for disease there
- Anti-engine reversed cells/parts back to normal state
  - Cells return to control by cellular control system
  - Restored cellular control system to full functioning
- Scrubbed out the cumulated precancerous state
- Eliminated shadow disease state
- Restored the immune system to high functioning
- No excessive trauma to treated animal

In Part II of this wonderful article, by Mr. Bearden, a few of the topics you can look forward to are; The Gulf War Syndrome, How to Make a Gravitational Wave & Curve Spacetime, Potentials & Gravity and How Priore Made Vacuum Engines.

---

[56] *The Russians refer to the hidden EM wave structure of the scalar potential – i.e., to the SWZ structuring and vacuum engine templates – as the information content of the field. Western scientists assume this phrase refers to ordinary spectral analysis and communications, ignoring those statements that the normal EM theory does not apply. Also, Devyatkov et al seem in no great hurry to express their true meaning more clearly. See N.D. Devyatkov (Ed.), Applications of Low-Intensity Millimeter-Wave Radiation in Biology and Medicine, IRE Akad. Nauk. SSR, Moscow, 1985. [in Russian]. See also N.D. Devyatkov, Yu. V. Gulyaev, et al, [Organizing Committee Chairmen], Digest of Papers - International Symposium on Millimeter Waves of Non-thermal Intensity in Medicine, co-sponsored by Research and Development Association "ISTOK" and Research Institute of U.S.S.R. Ministry of Electronic Industry ("ORION"), Moscow, Oct. 3-6, 1991. Collection of 300 papers. [In Russian].*

[57] *However, some leading Western scientists – such as Wheeler, Feynman, and Jaynes – strongly advocate replacing the present erroneous concept of the EM force field existing in spacetime with the notion that what exists there instead is an information storage device containing a summary of what we need to predict the behavior of a charge if there should be one at the point. E.g., see J.A. Wheeler and R.P. Feynman, "Classical electrodynamics in terms of direct interparticle action," Reviews of Modern Physics, Vol. 21, 1949, p. 425-433; E.T. Jaynes, "Probability in quantum theory," in Complexity, Entropy, and the Physics of Information, Ed. W.H. Zurek, Addison-Wesley, Redwood City, CA, 1990, p. 33-55.*

VACUUM ENGINES AND PRIORE'S METHODOLOGY: THE TRUE SCIENCE OF ENERGY-MEDICINE

Part II of II

© Copyright 1995 by T. E. Bearden

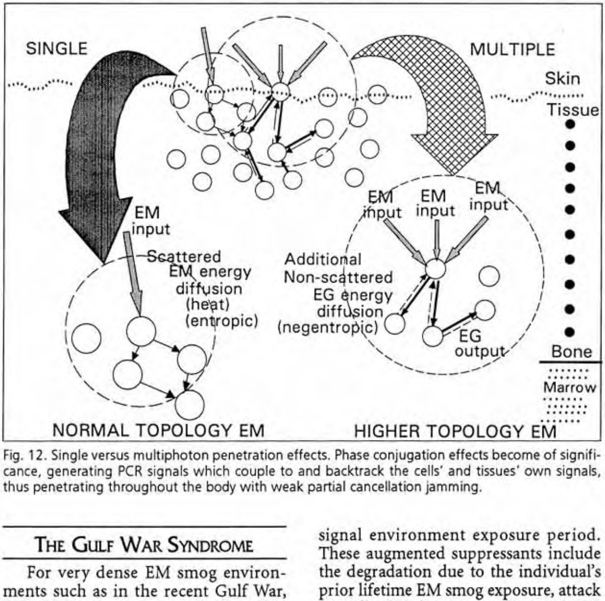

Fig. 12. Single versus multiphoton penetration effects. Phase conjugation effects become of significance, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, thus penetrating throughout the body with weak partial cancellation jamming.

THE GULF WAR SYNDROME

For very dense EM smog environments such as in the recent Gulf War, both the anomalous EM diffusion effect and the partial cancellation jamming effect on humans are dramatically accelerated. Cellular and immune system damage that would require decades of exposure at very low EM smog levels will cumulate in only a few months. In turn, this accelerated malfunctioning of the cellular control system and the immune system enhances the effects of other biological suppressants already present in the body or encountered during the dense signal environment exposure period. These augmented suppressants include the degradation due to the individual's prior lifetime EM smog exposure, attack by unfamiliar disease organisms endemic to the areas, exposure to oil fire smokes and pollutants, exposure to increased UV from solar radiation, and exposure to chemical agents and possibly surreptitious biological warfare agents.[58]

Further, in the haste to protect the troops from possible chemical and biological attack, all soldiers were administered a nerve agent called pyridostigmine bromide; in addition, approximately 150,000 were inoculated with anthrax vaccine, and 8,000 also received experimental botulism toxin shots.[59] Normally for medically-administered agents, the relatively minor suppression effects on the immune systems and body are followed by quick recovery, due to a properly functioning immune system and cellular control system. However, all three of these Gulf War preventive medical treatments do create "small cases of sickness" in the normal body.[60] In other words, a low level of the disease strain *is created* in the patient. Further, even after the apparent "initial recovery," there will now be a much longer dormant "apparently healthy" period during which the *vacuum engine templates for the disease remain, until they have finally decayed away.* In a sustained dense signal environment, this "decay away" period may be dramatically extended or never occur; in fact, the disease may progress from the vacuum-engine state to the low level physical state, then into a fully developed syndrome.

So we introduce the concept of the *relative level of a disease.* First, the disease and/or its vacuum engine templates are in the body. A low level or absence of disease, as we normally think of it, is with respect to the high disease level – the full-up, developed state of the disease. A disease level is "low" because of the ability of the immune system to fight it, overcome it, and reduce it to a "pre-disease" state; and the ability of the cellular control system to eventually rid itself of the disease vacuum engine templates, retaining only antibodies against its re-invasion. When the disease is not physically present according to conventional medical standards, but nonetheless the vacuum templates are present and not yet cleared out, we shall refer to that state as the shadow state of the disease. The cellular control system largely combats this state.[61]

---

[58] *Both Senator Riegle's and Senator Rockefeller's May 1994 hearings concluded that all 695,000 troops who served in the Gulf War were also exposed to chemical and biological toxins.*

[59] *Stephen Salgaller, "U.S. Role in Gulf War Illness." Letters,* ***Science News,*** *146(7), Aug. 13, 1994, p. 99.*

[60] *Note that at least one of these three treatments was still quite experimental. Consequently there exists less confidence in stating that its normal side effects are minimal and tolerable.*

[61] *Apparently by being able to phase conjugate with weak amplification, so that it forms anti-engines that usually are stronger than the disease's vacuum engines, gradually reversing the shadow state to normal.*

If the immune system and cellular control system are impaired, a shadow state of the disease can gradually progress to a "low level" of disease, then to a higher level, and become serious or even life threatening. This is true when conventionally the disease is physically absent but its shadow state is still present and cannot be disposed of by the cellular control system, as clearly shown by Kaznacheyev's experiments.[62,63]

We point out, but do not pursue, that the cellular control system also deals with the shadow states of diseases. In the normal state of "free from disease," the cellular control system successfully and continually deals with incident diseases, overcoming them at the shadow state level. Its process for continual disposal of the shadow state residues is hypothesized to be a version of slightly amplified phase conjugation of the vacuum engine's shadow state level.

In the model we are advancing, the disease entry and development occur by successively (1) creating its shadow state and successfully maintaining it in spite of the cellular control system's phase conjugation efforts, (2) breaching the threshold of physical changes, so that at least physical indicators of the struggle by the now-involved immune system appear, (3) progressing until conventional medicine recognizes "low level disease or infection," (4) increasing until conventionally-recognized "developed disease or infection," (5) increasing until the disease is serious and even life-threatening, and (6) increasing until the patient can no longer sustain life and dies. *We note that the best preventative medicine would and should be practiced at the shadow infection level.* The Priore therapeutical approach is capable of doing that.[64]

In the "normal" – if such an adjective can be used! – sustained "weak EM smog" exposure, the shadow state may last for several years or even several decades, prior to sufficiently accumulated degradation of the cellular control system and the immune system to allow the physical emergence of the disease. For much denser EM smog, such as existed in the Gulf War, the shadow incubation process is highly shortened, as is the time to progress from one stage to another higher stage of the disease or disorder condition.

In other words, in a dense but weak EM smog environment, the recovery from "minor diseases" actually administered as part of the initial vaccination and preventive medicine will be substantially altered because of the accelerated damage to the immune system and cellular control system. The control/immune degradation may be fast enough that the disease starts increasing from the shadow state toward higher levels, instead of being reduced and eliminated by the control/immune systems.[65,66]

Consequently, the Gulf War participants were thrust into a unique new situation: The second (shadow state) phase of decay of the "minor diseases" created in the bodies receiving the three preventive medications was interfered with by the EM fog. Because of individual variation in past cumulative damage experience, the individual control/immune systems varied from soldier to soldier. The smog environment varied in density and intensity from place to place, and from time to time. Hence the degree of interference varied from soldier to soldier because of (1) variation of past control/immune degradation, (2) variation of the EM smog exposure, and (3) variation of exposure to other suppressants. A bell-shaped distribution curve of the extent of the *net interference to net control/immune functionality*[67] would result, when the soldiers are taken as a group. That is, some of the participants just had a somewhat longer 2nd phase recovery period; some an extended 2nd phase recovery period; some a sustained, gradual rise in the 2nd phase, toward physical disease but not yet developing the Gulf War Syndrome. Some had a sustained, quicker rise toward actual physical disease instead of a decay, and have had sufficient time to develop the physical disease.

When the additional individual expo-

---

[62] *Kaznacheyev, several references, ibid. That is, suppose we have the shadow disease possess only a strength such that the normal cellular control system and immune system could overcome it. If the cellular control system is weakened, the disease progresses across the physical barrier. If the immune system is still strong, it can handle the encroaching disease, but will give physical indicators of its struggle (increased white corpuscles, etc.) If the immune system is also degraded, the disease will progress to low level physical disease, higher level, etc. On the other hand, if the cellular control system is normal but the shadow state of the disease is amplified, the phase conjugation defenses of the cellular control system will be breached and the disease will impel itself into low-level disease, then high level, and then kill the cells, in spite of a normal immune system combatting the breach of the cellular control system. This latter is what Kaznacheyev's experiments showed. It is also what the decades of Soviet weak microwave radiation of the U. S. Embassy in Moscow showed.*

[63] *The discerning reader will note that we are presenting the gist of a new theory of cellular disease.*

[64] *We also point out the electromagnetic biological warfare capability in this area, long since developed by the Russians and under total control of the KGB. The U.S. Embassy radiation, including the health changes and the death of three U.S. Ambassadors, was just to stimulate the U.S. intelligence and scientific communities at the highest levels. By the U.S. electromagnetic response at the Embassy itself, the Soviets could then know for certain whether or not the U.S. was aware of this EM biowar capability, and whether or not it had secretly developed defenses. Our actions assured them we had no such capabilities and no defenses – and were totally unaware of the true nature of the Embassy radiation.*

[65] *Note that the multiplicity of the assault on the control/immune systems augments the resistance and recovery difficulty. E.g., each B-cell lymphocyte can attack only a single antigen – alien property of the surface of an invader. There are about one billion B-cells in the human body, which between them are specific for about a million antigens. As can be seen, the immune system is continually having to develop B-cells with new specificities due to the continual meeting of new strains of antigens. Cumulated damage to this system affects both the body's resistance to diseases it has already acquired specific B-cells for, and its resistance to new strains for which it must make B-cells for new antigens. There is a sort of threshold, at any given time, beyond which the B-cell responses cannot keep pace, and start being overwhelmed. There is also an error tolerance level; when the errors due to deterioration surpass this level. the immune system will often begin attacking the body's own tissues in error. This results in disorders such as arthritis and cartilage and joint deterioration. With sufficient errors, previous immunity abilities also are degraded, so that more ordinary diseases become serious threats.*

[66] *The administering of three "lightly suppressive" treatments closely occurring with the soldier's entry into the Combat Zone thus posed a significantly greater problem than normal to his or her control/immune systems' operations, because the control/immune systems were suddenly degraded and hampered by the suppressive effects of the dense EM signals environment. In other words, the control/immune systems were moderately attacked by the three medical treatments (three enemy light cavalry divisions), and at the same time the entire control/immune system complex was subjected to substantial interference and suppression (the normal friendly) troop divisions were simultaneously placed under fierce and sustained attack throughout the zone. A general whose divisions are already committed in furious combat and under heavy bombardment, as are all his command posts and communications systems, has great difficulty freeing up, reconfiguring, and recommitting some of his committed troops to properly combat three other enemy forces that suddenly have appeared over the horizons from other directions!*

[67] *We may also define this as a disease progression ratio. This ratio should assume a distribution curve. For first-order simplicity, we are assuming a normal distribution.*

sures are factored in, there should have been essentially three broad groups: (1) relatively minor damage sustained, where any overt effects will only begin to appear many years hence if at all, (2) more damage sustained, where overt effects will begin to appear within the next few years, and (3) heavier damage sustained, where overt effects begin to appear either while still in the Gulf Region or within some months or so after returning.

Hence many of the Group 3 veterans of the Gulf war developed anomalous infections and immune system disorders shortly upon their return from such damage-accelerating exposure, resulting in the mysterious "Gulf War Syndrome." Conventional medical science will recognize – and has recognized – the overt phase 3 physical disease syndrome, but its cause and effective treatment remain and will remain medical mysteries. Conventional medical science will entirely fail to recognize Groups 1 and 2 cases, and will label them as due to aging, mental stress, or ordinary rise of diseases.

The syndrome has involved the "anomalous EM smog diffusion" process, as previously explained. It may – and in many cases it will – exhibit certain "signature" differences from normal diseases. The genetics of the victim are affected, and so are altered by induced transcription[68] to carry at least a portion of the "syndrome template." Semen, ova, saliva, and blood will particularly be so affected. A wife may complain that her husband's body fluids are now irritating to her skin, even producing "burning" sensations. Children conceived and born may be borne with the syndrome, or with one or more aspects of it.[69] The syndrome template can be transmitted by exchange of body fluids, so wives may also contract the syndrome and develop the disease from their husbands.[70]

Conventional EM Bioeffects Knowledge Is Insufficient

Present "EM bioeffects" scientific models contain absolutely no modeling of partially jamming the body's own special EM control system by turning its EM signals partially into G-signals, thereby partially nullifying them. It also contains no modeling of the anomalous EM diffusion process.[71] Thus the conventional medical approaches also do not contain any proper means of accounting the mechanisms for, and results of, long-term exposure to electronic smog and other combined agents and effects. Conventional scientists also have no knowledge of the major degradation mechanisms involved. The conventional medical models specifically are useless when confronted with something such as the Gulf War Syndrome or the decades of Russian microwave radiation of the U.S. Embassy personnel in Moscow. Capable scientists turn to their conventional models for answers, and the models fail them. Therefore present medical science is baffled when faced with such diseases and syndromes.[72] (*See figure 13 on next page*)

How to Make a Gravitational Wave and Curve Spacetime

We return now to the fundamentals of Type 2 engineering.

If one superposes an ordinary EM wave with its true phase conjugate replica wave, where the two waves are of equal strength, the *EM forcefield aspects* of the combined wavepair (standing wave) vanish, while the energy-density aspects remain. At every point the E-fields and the B-fields sum to vector zero *forcefield* resultants. The standing wave that results will not cause EM movement of charged particles.

However, the *energy densities* of the two locked (and now "hidden") EM waves remain and add. By standard GR, this "trapped EM density standing wave" must constitute a standing gravitational wave and a standing wave in the local curvature of spacetime!

It follows that, by assembling a harmonic set of such wavepairs, a la Stoney and Whittaker, one can form a region of spacetime such that the energy density is deterministically altered and controlled, and to engineering levels.

Engineering general relativity suddenly became a lot easier! Ironically, present physics only considers primarily the *very weak* ST curvature that can be obtained by the extremely weak gravitational "attraction of mass" G-force. For electrons, e.g., the G-force is only on the order of 10^{-42} times as strong as the electric force. Consequently orthodox GR simply "cannot pack enough trapped G-field energy into the local space" in the laboratory to get measurable effects. Only in the vicinity of a very large mass, such as the sun or a star or a black hole, does the weak G-force reach sufficient strength to yield directly observable, strong effects. Yet by using EM vacuum engines and the SWZ approach, one can obtain readily-measurable GR effects right on the laboratory bench.

The conventional

medical models

specifically are

useless

when confronted

with something

such as

the Gulf War Syndrome

or the decades of

Russian

microwave radiation

of the U.S.

---

[68] *Note that the transcription component may continue to develop in the same "post disease" manner as does the shadow disease of the overall Gulf War Syndrome. The returnee continues to experience EM smog and other control/immune suppressants, though at a lesser level than experienced in the Gulf War environment.*

[69] *Such aspects are highly controversial. A few affected families are reporting such symptoms and aspects, but officially "there is no evidence" for them. However, with the initial appearance of the syndrome, officially veterans who complained of this illness were briefed by these same medical "authorities" that their physical symptoms were stress-related, and that the cause was all in their mind!*

[70] *In a fashion similar to Priore's animal experiments that demonstrated the transfer of curative templates by a single drop of blood. Kaznacheyev showed that disease templates could also be transferred.*

[71] *In fact, EM bioeffects research for decades has used, and continues to use, sheer energy deposition (scattering of EM energy by the tissues – heat by another name!) as a measure of the exposure level. They have not applied nonlinear phase conjugate self-targeting effects at all.*

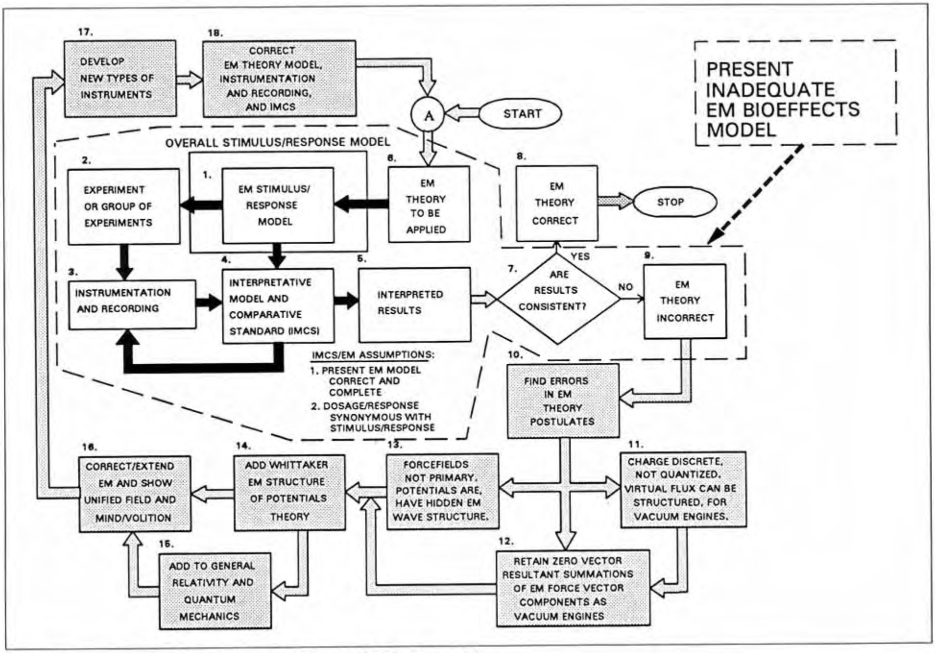
17.
DEVELOP NEW TYPES OF INSTRUMENTS
18.
CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS
A
START
PRESENT INADEQUATE EM BIOEFFECTS MODEL
OVERALL STIMULUS/RESPONSE MODEL
2.
EXPERIMENT OR GROUP OF EXPERIMENTS
1.
EM STIMULUS/ RESPONSE MODEL
6.
EM THEORY TO BE APPLIED
8.
EM THEORY CORRECT
STOP
3.
INSTRUMENTATION AND RECORDING
4.
INTERPRETATIVE MODEL AND COMPARATIVE STANDARD (IMCS)
5.
INTERPRETED RESULTS
7.
ARE RESULTS CONSISTENT?
YES
NO
9.
EM THEORY INCORRECT
IMCS/EM ASSUMPTIONS:
1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE
10.
FIND ERRORS IN EM THEORY POSTULATES
16.
CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION
14.
ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY
13.
FORCEFIELDS NOT PRIMARY, POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.
11.
CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.
15.
ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS
12.
RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

Fig. 13. Necessary corrections for the present inadequate EM bioeffects model.

POTENTIALS AND GRAVITY

Obviously one can simply add harmonics of such EM "locked in" wavepairs, to produce a standing "constant value potential" pattern of a fixed *excess* energy density of vacuum or a fixed *decrease* in the energy density of vacuum. This harmonic series of bidirectional EM wave pairs – the *sum set* of wavepairs – is precisely what Stoney and Whittaker showed makes an ordinary EM scalar potential. In other words, they showed the actual energy-trapping mechanism, and dramatically altered the view of a scalar potential as being simply an "electrostatic" entity and a scalar field. Instead, it is a *hidden-variable energy flow process.*[73,74,75]

Its external "scalar field" aspect is just its "local energy density" aspect, not its fundamental nature.[76]

As is well-known in GR, all potentials are gravitational because they contain trapped energy. *Any* trapped energy is gravitational, a priori, according to GR. Ziolkowski independently rediscovered the SW decomposition in the 1980s and extended it to also add the product set. Since multiplication of waves is *modulation*, in my opinion Ziolkowski actually wrote the basis for hyperspatial (or subspatial, if one models physics in only 4-dimensions) communication at superluminal velocity, essentially instantaneously anywhere in the universe. (*See figure 14 on opposite page*)

"Paired assembly of an EM wave and its true PCR" was and is the way to turn EM energy directly into gravitational en-

[72] *Just as it is also essentially baffled by many conventional dread diseases such as cancer, leukemia, AIDS, arteriosclerosis, etc. – which indicates that the conventional "models" for those diseases are also seriously deficient.*

[73] *The present preoccupation of physicists with the scalar potential is simply to provide an algorithm for calculating the local energy density – in excess joules per collecting point-coulomb of charge – and erroneously call that the definition of the potential. The magnitude of the energy collection density of the potential at a point is not what the potential* is, *and such is not its definition. Defining how much water can be poured out of a standard bucket, when it is dipped into a flowing river and filled with water, does not define the river!*

[74] *From the particle view, the potential is an ordering in a virtual particle flux of the vacuum. From the wave view, the potential is comprised of an ordered, harmonic series of SWZ bidirectional EM waves, where each wave reaches to the ends of the universe before having a zero energy density. It follows that the energy in each hidden sine wave is infinite, just as it is for any overt theoretical EM sine wave. So any potential – even that of a flashlight battery – already involves an infinite amount of flowing hidden EM energy, being exchanged by the source and the entire universe, flowing in both directions along any radial from the source of the potential. The scalar potential is a flowing, continual exchange of energy between each and every point in the universe and the local source of the potential.*

[75] *Both the wave and particle forms of the potential exist simultaneously. The particle nature exists in Minkowski 4-space. The "hidden SWZ waves" exist in a hyper-4-space whose time dimension is the same as that of Minkowski 4-space. The scalar EM potential actually exists as a 7-dimensional entity. The internal hidden EM flow is not confined to the speed of light in space, because it does not flow "through" 3-space; instead, it flows "around" it.*

[76] *E.g., Whittaker, 1904, ibid. shows us that all of classical electromagnetics can be made by the interference of two scalar fields, which is just a special kind of hyperspatial multiwave interferometry.*

ergy.[77] This is the unification of EM and GR that eluded Einstein all the latter part of his life, and that has continued to elude scientists today. This is the way to "trap" EM energy locally, even though the EM wave energy is flowing at the speed of light or greater. This is the way to change EM energy into gravitational energy.[78] Further, it is engineerable and testable. Remember, physics requires that energy be conserved; however, it can certainly *change its form*. GR already says that it is the trapped energy that is gravitational, not mass per se, and regardless of what kind of energy is trapped. Ultimately Einstein's formula $E = mc^2$ is actually a statement of the trapped EM energy in a mass. However, heretofore no one had the slightest notion as to the actual trapping mechanism. The SWZ work actually shows the trapping mechanism itself and allows the engineering of *electrogravitation*.

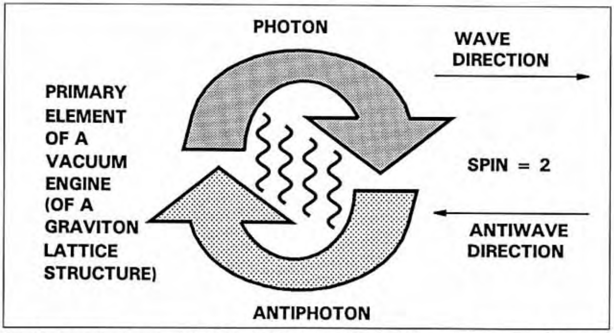
PHOTON
WAVE DIRECTION
PRIMARY ELEMENT OF A VACUUM ENGINE (OF A GRAVITON LATTICE STRUCTURE)
SPIN = 2
ANTIWAVE DIRECTION
ANTIPHOTON

Fig. 14. A graviton is a coupled photon/antiphoton pair. In the bidirectional wavepair of the SWZ internal structure of the scalar potential, photons in the wave continually couple and uncouple as the wave and antiwave pass through each other.

How Did the Revolutionary Stoney/Whittaker Work Go Unnoticed?

Apparently, neither Stoney nor Whittaker realized that their work was pure electrogravitation. *When the referenced papers by Stoney and Whittaker were published, special relativity, general relativity, and quantum mechanics had not even been born yet!* Indeed, there still was not a wide adoption of Heaviside's vector interpretation of Maxwell's fading quaternion theory. Even classical EM was still in its infancy and just getting off to a good start.

In the 1980s Ziolkowski realized that the internal biwaves of a potential could be engineered, but remained primarily focused upon his important work with electromagnetic missiles and focus waves (slugs of EM energy which stay together as they travel, and do not spread out.) Ziolkowski is one of the pioneers in that field.

How Priore Made Vacuum Engines

This "vacuum engine methodology" was and is what Priore's intuition led him to! For the electromagnetic treatment of laboratory animals, his EM therapy device first inputted some 17 or so frequencies into a rotating plasma.[79,80] We know today that such plasmas *phase conjugate* the wave inputs, but that was not known when Priore first did it.

Obviously Priore himself did not understand the actual mechanism, but his brilliant intuition led him to discover how to do it! For many years no one knew how aspirin worked, yet that did not prevent it from being used widely and effectively.

In other words, Priore used the plasma to *unwittingly* add PCRs to each of his waves and form 17 magic SWZ "wavepairs." The entire SWZ wavepair assembly thus made an *artificial potential*, having a "vacuum engine" pumping substructure of hidden bidirectional EM waves. Adjustment of selected frequencies away from their harmonic positions adds multiplications and beat frequency modulations, creating *specific* templates and *specific* vacuum engines. These SWZ waves are hyperspatial; hence they are in the domain of the deep mind's cellular control system. When an abnormal body cell is pumped by these hyperspatial waves, the cell's own hyperspatial vacuum-engine "signal input complex" triggers a specific dedifferentiation anti-engine for that specific abnormal state. By Wheeler's principle, if spacetime curvature is *present in* the material cell, its energy and template are *applied to* the cell and act upon it. With vigorous pumping, the anti-engine is also *amplified* and so the cellular response is more rapid. In other words, the amplified anti-engine fairly quickly overpowers the disease engine and reverses the abnormal cell and all its contents back to their previous *normal* states. (*See figure 15 on the following page*)

Priore's new signal complex came out of the plasma mixing (phase conjugating) tube with SWZ structures of wavepairs, where each pair was composed of two bidirectional (opposing) EM waves. Priore then modulated this output upon a very strong, rippling magnetic field as a carrier, and radiated the body of the patient with that resultant field and its transported hyperspatial SWZ anti-engine components.

Nuclear magnetic resonance guarantees that such a magnetic field with all its modulations will be delivered entirely through the body, penetrating the flesh and bones and even the atomic nuclei themselves. It will certainly penetrate the bone marrow where all the blood cells (which includes most of the immune system's cells) are born! It will penetrate everywhere else also. So all the cells of the immune system will be exposed to the AC magnetic carrier and its impressed hyperspatial biwaves and their modulations.

The body's cells and cell-parts are highly nonlinear, and so they are also *demodulators*. They strip off the SWZ modulations. This puts the amplified vacuum anti-engines directly into and onto the cells, including all the cells of the immune system, and including each and every part of each cell. Small wonder then that the Priore methodology could restore suppressed immune systems! Time-reversing the damaged immune cells and their damaged internal components back to a previous healthy state was one magnificent result. The radiation also delivers the modulations directly into the master potential (the body's quantum potential) that contains its mind, its long-term memories, and its master control system for all cells. Thus not only are the diseased cells and the immune cells affected, but also the *master cellular control system* for all body cells is directly affected.[81]

---

[77] *I have previously pointed out that, in a coupled wave/antiwave pair, the photons in the wave and the antiphotons in the superposed phase conjugate replicate wave also pair into photon/antiphoton pairs, which have spin-2 helicity and hence are gravitons. This follows merely from a strong application of the distortion correction theorem. However, it is the fundamental union of EM and G into a single engineerable theory. See Bearden,* ***Gravitobiology,*** *1991, p. 1-9.*

[78] *By Whittaker 1904, we can always interfere two such potential fields and obtain an EM field, even at a distance. Thus we can also readily turn the gravitational (trapped EM) energy back to EM field (untrapped EM) energy.*

[79] *David M. Rorvik, "Do the French Have a Cure for Cancer?"* ***Esquire,*** *July 1975, p. 110-11, 142-149 gives an overview of the Priore Affair and some details of the electromagnetics of the machine.*

[80] *Note the immediate connection to the Russian energetics model, which is directly fashioned after the nature and nonlinear EM functioning of plasmas. Priore was directly applying Soviet energetics, although neither he nor any other Western scientist could have known that.*

In a related work, Becker (*See figure 16 on the following page*) and his colleagues stimulated bone healing in difficult fractures by use of tiny DC currents. They have shown that picoamperes of steady DC current trickled across the fracture site induces bone healing. Becker showed that, surprisingly, red blood cells are dedifferentiated – shucking their hemoglobin and growing a nucleus in the process. These cells redifferentiate into the kind of cells that make cartilage. Then *those* cells redifferentiate anew, into the kind of cells that make bone, and heal the fracture with new bone deposit. Becker also proposed a complete, operational DC control system, staying within the limitations of conventional EM theory.[82,83,84]

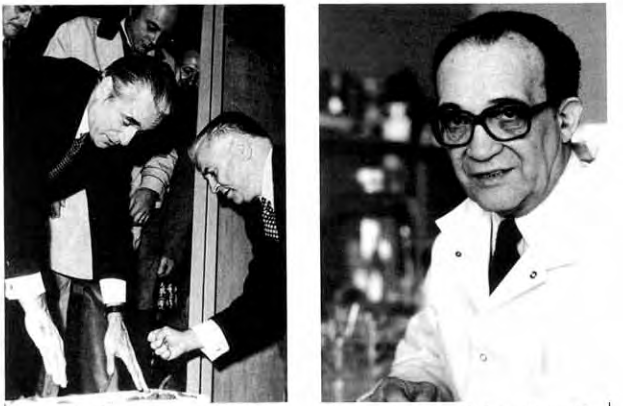

15-A. Priore demonstrates a point to Mr. Chaban-Delmas, Mayor of Bordeaux and a friend and mentor.

15-B. Dr. Raymond Pautrizel, noted immunologist, worked directly with Priore. Dr. Pautrizel is Professor Emeritus, University of Bordeaux, and holder of the first chair in immunology in France.

Dr. Robert Courrier (not shown), head of the Biology Section of the French Academy of Science as well as its Secretaire Perpetuel at the time, was also a noted colleague. He personally presented the astounding results of the Priore Team's experiments to the Academy.

Fig. 15. Priore was supported by notable French political leaders and eminent French scientists were his close scientific colleagues.

A proposed explanation of the fundamental vacuum-engine mechanism for the Priore and Becker effects has previously been given in some detail by Bearden.[85] (*See figure 17 on next page*)

[81] *We refer to an expanded version of Popp's cellular control system. E.g., see Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp et al, eds.,* ***Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977;*** *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149. Popp reports the theoretical and experimental discovery that mitogenetic radiation from cells can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism. In addition, it includes the storage of "virtual" coherent photons.* ***Comment:*** *Our substitution of the SWZ hidden wave approach to replace Popp's QM particle approach greatly facilitates understanding and specifying the control system, and materially eases direct engineering of it.*

[82] *E.g., see R.O. Becker, "The direct current field: A primitive control and communication system related to growth processes."* ***Proceedings of the. XVI International. Congress of Zoology,*** *Washington, D.C., Vol. 3, 1963, p. 179; "The control system governing bone growth in response to mechanical stress."* ***J. Ark. Med. Soc.,*** *Vol. 62, 1966, p. 404; R.O. Becker and D.G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents."* ***Trans. N.Y. Acad. Sci.,*** *Vol. 29, 1967, p. 606. See particularly Robert O. Becker, "The significance of bioelectric potentials."* ***Bioelectrochemistry and Bioenergetics,*** *Vol. 1, 1974, p. 187-199. Becker proposes that a complete, operational system exists in living organisms, and controls such basic functions as growth, healing, and biological cycles. The system is described in anatomical detail. It functions as a data transmission system in an analog fashion using varying levels of direct current as its signal. The system interlocks physically with the nervous system and is postulated to be its precursor. There are two electrochemical limbs of great significance in the operation of the system. One is between the DC system and the nervous system; the other is between the DC system and all body cells. The concept explains all of the diverse effects reported for the biological effects of applied electrical currents including: electrical anesthesia, electrical growth control, and electroacupuncture. It also furnishes a testable hypothesis for predicting other effects of clinical significance.*

[83] *We comment that the SWZ approach allows control and conditioning – both externally and internally – of both limbs of Becker's proposed DC control system.*

[84] *See also C.W. Hsue, "A DC Voltage is Equivalent to Two Traveling Waves on a Lossless, Nonuniform Transmission Line,"* ***IEEE Microwave and Guided Wave Letters,*** *Vol. 3, 1993, p. 82-84. Hsue shows that the steady DC voltage on a transmission line powering a resistive load, mathematically decomposes into bidirectional EM traveling waves. The bone-healing setup used by Becker et al is precisely of this configuration. The fracture corresponds to the resistive load and the two electrodes on each side correspond to the transmission lines. Hence Becker's team actually employed a form of the Stoney/Whittaker/Ziolkowski harmonic biwave decomposition of the scalar potential, to accomplish hyperspatial pumping and formation of anti-engines of vacuum energy. So this is another application of the same basic "cell-reversing" methodology, as also utilized more extensively by Priore. Becker's brilliant work of course was and is independent and unique, and he was nominated for a Nobel Prize because of his absolutely fundamental work in showing electromagnetic differentiation and dedifferentiation of cells. That he has not been awarded the prize is regrettable and, in my opinion, an injustice.*

[85] *T.E. Bearden, "Mechanism for Long-Term Cumulative Biological Effects of EM Radiation," presented to the 70th Annual Meeting of the Alabama Academy of Science, University of Alabama at Huntsville, March 25, 1993. A digest of this presentation is published by the Tesla Book Company, P.O. Box 121873, Chula Vista, CA 91912 as T.E. Bearden,* ***Cancer and the Unresolved Health Issues in the Biological Effects of EM Fields and Radiation,*** *1993 and includes both text and hard copies of viewgraphs.*

HOW THE PRIORE VACUUM ENGINE REVERSED CELLS TO A PREVIOUS STATE

Taking a GR view, any mass already has its own vacuum engine for its exact form and internal structure and dynamics, in the finest detail. That particularly follows Wheeler's principle; if the mass is there, the appropriate spacetime curvature and dynamic energy structuring is there. If the mass has internal structuring, so does the spacetime curvature: it becomes *nested* curvatures with a precisely correlated template. In short, any mass – living or not – is imbedded in a precise vacuum engine. Any and every change adds a specific delta to this vacuum engine component. Just as a tree has rings for its past years and an onion has layers, the vacuum engine of an object has an exact template for that object's entire past, *as do the mind and its components also.*[86]

So when a normal living cellular mass changes to a diseased state, *a priori* it has added an additional delta vacuum engine (representing the change that resulted in the disease) to its original vacuum engine (which represents the original healthy state of the cell). The cell and all its parts are highly nonlinear and act as phase conjugate mirrors, *including to the higher topology (hyperspatiality) of the vacuum engine. Any living system we isolate and study is at least 7-dimensional.*

Putting all this together, we now can understand the marvelous nature of what Priore did. By exposing the body, treated part, cells, and cell parts to Priore's modulated rippling magnetic field, the "hidden pump waves" carried by that magnetic field are applied to each and every cell and to all its parts, in the exposed area. *In the exposed area, each and every cell and its parts become powerfully pumped phase conjugate mirrors for its hyperspatial vacuum engine complex.*

Extending the language of nonlinear optics, the particular *net* vacuum engine of the diseased cell or part constitutes a "standing input" *hyperspatial signal wave complex*, already input to the hyperspatially pumped phase conjugate mirror cell or cell part. When pumped by "hidden" pump waves, each and every one of those PC mirrors will now originate an *amplified phase conjugate replica of the signal wave complex. Each one will produce a highly amplified, time reversed replica of its specific standing vacuum engine. This includes amplified time reversed replicas of both the "healthy" component of each standing vacuum engine and the "disease delta" component. (See figures 18 above and 19 on following page)*

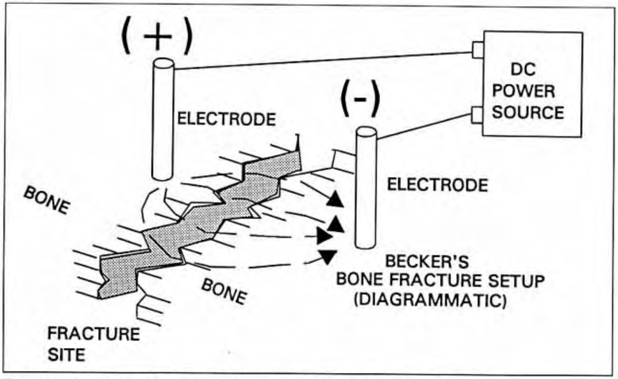
(+)
(-)
ELECTRODE
DC
POWER
SOURCE
ELECTRODE
BONE
BONE
BECKER'S
BONE FRACTURE SETUP
(DIAGRAMMATIC)
FRACTURE
SITE

Fig. 16. Becker's application of tiny DC currents to heal difficult bone fractures.

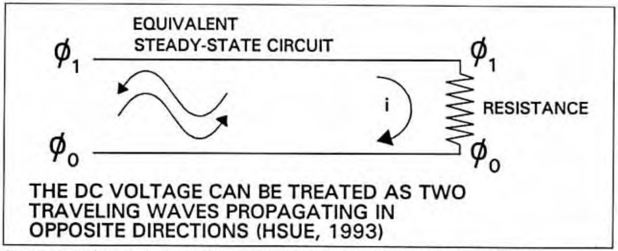
EQUIVALENT
STEADY-STATE CIRCUIT
φ1
φ1
i
RESISTANCE
φ0
φ0
THE DC VOLTAGE CAN BE TREATED AS TWO TRAVELING WAVES PROPAGATING IN OPPOSITE DIRECTIONS (HSUE, 1993)

Fig. 17. Hsue's biwave decomposition of DC voltage (as in Becker's application of tiny DC currents to heal difficult bone fractures).

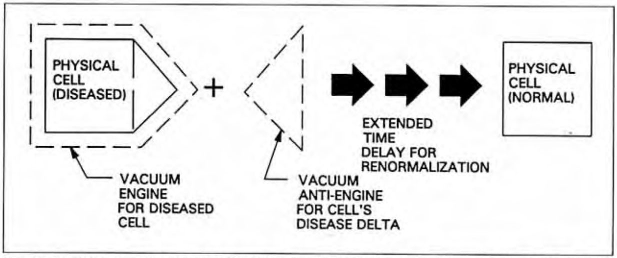
PHYSICAL
CELL
(DISEASED)
+
PHYSICAL
CELL
(NORMAL)
EXTENDED
TIME
DELAY FOR
RENORMALIZATION
VACUUM
ENGINE
FOR DISEASED
CELL
VACUUM
ANTI-ENGINE
FOR CELL'S
DISEASE DELTA

Fig. 18. Applying an anti-engine for the "deviation" of the diseased cell reverses the cell to normal after a time delay.

In other words, every pumped diseased cell and cell part – including the genes and chromosomes – will produce a very powerful, highly amplified, "reverse vacuum engine" (an *anti-engine*) which is fantastically specific and precisely time-reverses the cell and its parts, including the genes and chromosomes as well, back to their *previous normal states.*

---

[86] If one alters the internal structure (rings) of a tree, one changes the tree. If one alters the internal structures and dynamic functioning of matter, one changes the matter accordingly. If one alters the internal structures and dynamics of the mind, one changes the mind, its contents, and its functioning accordingly, as well as the subsequent behavior exhibited by the altered individual.

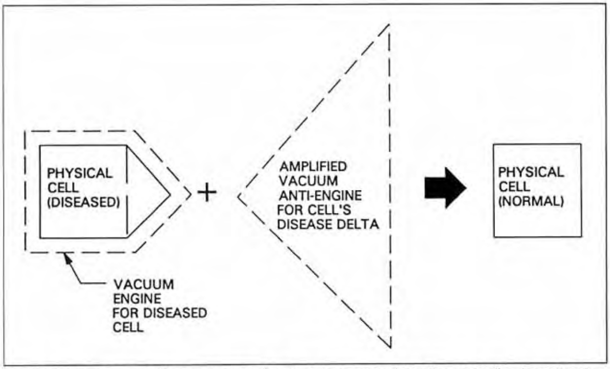

Fig. 19. Applying an amplified anti-engine for the "deviation" of the diseased cell reduces the time delay required to reverse it back to normal.

REJUVENATION

Any *normal* cell that is exposed to hyperspatial (Priore-type) pumping just reverts to a *slightly younger* normal state. Note that this allows *rejuvenation*.[87] Priore's earliest experiments that so riveted his attention and set him on the path of discovery, were in fact the abnormal prolongation of the life of fruits! For each class of disease, Priore just had to fiddle with the frequencies until he had the right set of pump waves for the specific "disease delta vacuum engine." This he did by trial and error. So when it is "Priore-pumped," a cancer cell reverses back to a normal cell, as does a precancerous cell as well.

The revolutionary Priore therapy is startling but simple. The diseased cell is analogous to an automobile that has been driven slowly off the road into the rough. Priore's methodology shifts the car into reverse gear, guns the engine, and quickly and precisely backs the car back onto the roadway. It is one of those profoundly beautiful embodiments analogous to Faraday's unipolar generator; it strikes directly to the heart of the fundamental principle. A cell develops damage and disease over a period of time. Simply time-reverse it quickly back to its original healthy state.

AIDS COULD BE QUICKLY CURED THE SAME WAY

Given Priore-pumping, an AIDS-infected cell would revert back to a normal cell, including all its genetics, and including for all of the hundreds of "strains" of AIDS that exist in an advanced AIDS patient. With this approach one could completely reverse and eliminate AIDS whenever a positive reading occurs in the HIV virus test, and prior to the appearance of the actual physical symptoms! If the person just has the antibodies and not the disease, the treatment will still do him absolutely no harm. (It will actually make his cells a "little bit younger and more vigorous," and help his immune system.) If he has both the AIDS antibodies and the infection as well, he will be completely cured and his immune system will be restored to complete vitality.

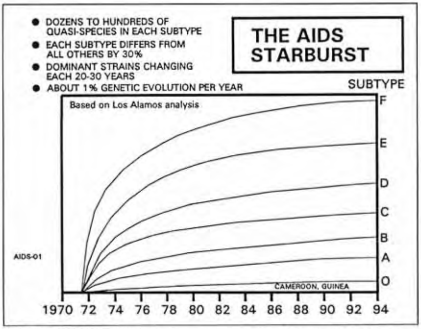

Fig. 20. AIDS is continually changing since the HIV virus is a retrovirus and unstab.

MOST DISEASES COULD BE CURED BY THE NEW THERAPEUTIC MODALITY

Any infectious disease known or unknown can be cured by the same means. The marvelous thing is that one need not know the specific technical pattern for the delta vacuum engine for a specific disease! One just has to *correctly adjust the hidden SWZ pumping frequencies.*

ARTERIOSCLEROSIS

In arteriosclerosis, it has been found that the greatest part of the damage is due to the presence of some rather ordinary viruses, and abnormal cell growth by the body's own cells as a result of triggering by previous viral infection. Cellular reversal (dedifferentiation) gets it all; the viruses and the abnormal cells. Everything reverts to normal, and the body's restored processes quickly "police up" the extra cell growth, using its natural control processes which are now restored to full functioning. *This is how the Priore team was able to so dramatically cure the "clogged arteries" of laboratory animals.*

The Priore approach (pumping of the

[87] *One should not be surprised that a time-reversal operation upon an entire body leads to reversal of physical aging in that body! But we point out that the time reversal we speak of is not at all the "time travel to the past" so beloved by generations of science fiction writers. For that, one would have to reverse each and every part of the universe simultaneously, except the observer, back to an exact former state, in perfect universal coordination. One would have to pump the entire universe, except for the observer! The observer would then remain as is with no change, while his entire world "went back in time" to a past state. The pumping of the universe would have to stop instantaneously (decay of the false vacuum?), and the universal anti-engine would have to stop operation and decay instantaneously (another big bang, perhaps?). We are not referring to anything such as that! We are referring to the reversal of a single little part of the universe, back to its previous state, which is practical and can be done. Universal time reversal – and hence time travel for the observer – is not practical, at least within our present knowledge and within the time reversal model we are applying. We should point out, however, that time reversal is a legitimate part of physics, and is not science fiction. E.g., for an overview, see Robert G. Sachs,* ***The Physics of Time Reversal,*** *University of Chicago Press, Chicago, Illinois, 1987.*

diseased cell's self-contained vacuum engine to create a very powerful vacuum anti-engine and reverse diseased cells back to normal) is the most powerful therapy for cellular disease ever conceived of. If the cells have ever been normal, they can be reverted to normal.[88] One can also probably regrow lost limbs, etc. but that takes too long to explain here! The methodology should also be able to completely reverse disorders such as Alzheimer's disease, and very probably muscular dystrophy as well. It should even be possible to reverse the physical effects of aging!

MUCH MORE THAN THIS INTRODUCTION HAS BEEN THEORIZED

This is the gist of the mechanism for the Priore therapy's startling success, although it has required some 10 years of very hard work to decipher the mechanics of the process. Along the way, I have also personally discovered and added a great deal more (for example, the quantum potential aspect of a living system and any enclave of living systems, the mechanism to generate a quantum potential, the precise mechanism for Sheldrake's morphogenetic field, the mechanism for directed evolution, the unusual electromagnetic nature of mind and thought, the mechanism for engineering mind and thought, engineering the entire species, and engineering the entire living biosphere), but this much should suffice for now for introductory purposes.

A technical book has been laboriously begun, and hopefully will be finished and published in about a year. Two 2-hour videos are also tentatively being planned. A four-hour technical briefing has already been prepared and presented.

REACTIONS TO THIS VACUUM ENGINE EXPLANATION

In response to numerous queries as to the reactions to this explanation of the Priore mechanism: Reactions tend to be rather good when (1) the technical person is open-minded, in other words, a friendly skeptic, (2) the technical person endures a grueling 2-hour or 4-hour detailed slide presentation I prepared, and (3) sufficient follow-up dialogue is accomplished to answer residual questions. Otherwise, the "mental jump" required of the participant is just too great.[89] Also, perhaps 60-70% of the conventional scientists are absolutely against anything that smacks of "revolutionary." From about half of this contingent the approach may generate disbelief and even overt emotional hostility, just as Priore himself and his associates encountered.[90]

However, some of the sharpest young graduate students readily make the transition, if they interact with me for a long enough period and enough times to understand the physics of the approach. *It did not take me 10 years to decipher an easy thing!* The causative mechanism for Priore's accomplishment is terribly involved and very, very deep. It is slightly beyond the present scientific vanguard, although the bits and pieces necessary to decipher the Priore mechanism are now in the literature. But putting them together has been a bone-crushing, terrible task. Priore was truly a century ahead of the science of his time!

Nonetheless, there is much hope, because the material is slowly taking root in a few sharp young graduate minds. For example, one of the major Eastern Universities would be astonished to know that their sharpest, brightest young PhD candidate – one they already assume will win a Nobel Prize in the future – is a full convert, and thoroughly understands cellular reversal by hidden-variable SWZ pumping, a la Priore. But at my strong urging, he is keeping it a secret until he graduates, and then he will *gradually* begin to filter in some of these ideas. That way he does not initiate his destruction by his peers because of too much energetic advocacy too quickly.

THE PRIORE APPROACH IS URGENTLY NEEDED

We simply must get this Priore therapeutic approach developed, for it is extremely important to all humanity. Our very survival may depend upon it.

Presently all our antibiotics are beginning to fail; the disease organisms are becoming quite resistant. Already our hospitals have strains of diseases that few if any antibiotics will touch. Hundreds of thousands of hospital patients each year are already experiencing complications due to contracting an antibiotic-resistant infection. Probably about 20,000 a year are dying as a result, and the death rate is increasing. In the future, one will have to seriously weigh whether it is better for one's chance of survival to enter the hospital and be treated, or to take one's chances with the disease and not enter the hospital.

With the rising failure of the antibiotics, (*See figure 21 on page 60*) and the rising possession of tailored biological weapons by hostile nations and international terrorists,[91] (*See figure 22 on page 60*) we shall shortly have no defense against an onslaught of pandemics unheard of since the Black Death swept medieval Europe of one third the population. Literally the age of pestilence and dread diseases of Revelations is upon us.[92]

---

[88] *They can also be "reversed to normal" even if they have never before been normal. The master cellular control system contains a master blueprint for the normal cell and can make an appropriate template for the specific disease delta vacuum engine, or one can be introduced (e.g., by injection of a drop of blood from a person previously cured of that specific disease; this was in fact demonstrated by the Priore group.). Pumping will then reverse the patient back to normal, as was actually shown by the Priore group. Indeed, the body's own system may be able to use the correct anti-engine template from the drop of blood and cure itself without pumping, as was also shown by the Priore group, if the body itself is reasonably healthy.*

[89] *Simply "dropping it on someone" by mail, for example – as I did to Elizabeth Taylor's foundation and Ross Perot's office – is just too much for a casual reading to grasp. A polite acknowledgment was received from the Taylor foundation, and no acknowledgment at all was received from the Perot camp.*

[90] *As an example, Priore completed the course work for his PhD and submitted his thesis, documenting his formidable and revolutionary work. Because of the ferocity of the oncology establishment toward the Priore team and its accomplishments, the university rejected his thesis. During the two decades or so since then, untold millions have died worldwide from diseases the team cured. These deaths could have been prevented had the Priore work been vigorously pursued and developed by the scientific community.*

[91] *It is not a question of if the international terrorists will use biological weapons against the U.S. civilian populace, but only of when. Astoundingly, often it may even be possible for the terrorists to simply buy their biological warfare agents from us! In a lawsuit by Desert Storm Syndrome victims against certain U.S. companies, one of the companies is alleged to have sold germs – such as anthrax, tularemia, and botulism – to Iraq by mail. These are highly effective biological warfare agents. E.g., a few spray tanks spraying such agents over a large U.S. city complex at night, using an easily obtained light plane to get in a few hours "flying practice," would result in tens or even hundreds of thousands of deaths within days. According to U.S. government-released data, inhaling a few anthrax spores into the lungs has a very high probability of killing the inhaler, unless treatment is started immediately. Even with immediate treatment, many deaths will still occur. A substantial number of nations are now obtaining and producing lethal biological warfare agents – which have been called the "poor man's atomic bomb." Some of the international terrorist groups are also acquiring the biological warfare agents.*

[92] *A good exposition of the rapidly increasing resistant-disease problem is given by Garrett, Laurie,* ***The Coming Plague: Newly Emerging Diseases in World out of Balance,*** *Farrar, Straus and Giroux, New York, 1994.*

In addition, the costs of the present allopathic medicine are rising rapidly and threatening to either bankrupt the nation, or become so expensive that only the wealthy can afford adequate medical care. Even the wealthy could be trapped in the future and denied adequate medical treatment. The Clinton Medicare bill, if approved, would make it a felony to pay a doctor out of one's own pocket, after denial of such treatment by the administrative bureaucracy. It also makes it a felony for the doctor to give the treatment to the disapproved paying person. Thus both the patient and the doctor could be sent to prison, wealthy or not.[93]

Cell Conjugation by Vacuum Anti-Engines Is the Only Hope

Priore's approach offers an answer to both the age of pestilence and the impending biological warfare strikes against our armed forces and our civilian population.[94,95] By prompt cellular reversal, infection even by totally unknown and modified agents can be quickly voided. Infection by newly adapted, resistant strains of presently known diseases can be quickly conjugated. So the problem of the failing vaccines and antibiotics can be solved quickly, simply, and cheaply. However, the cell-reversing treatment devices and trained, experienced medical staffs must be widely available!

Diseased cell conjugation by EM means is also the only therapy that can largely *eliminate* the debilitating diseases of humanity. Per Wheeler's principle, vacuum engines can mold the *physical matter* of the living body and conditions, operations, and content of the mind, memory, and personality. As the Priore therapy is developed and extended, it will become possible to correct bent spines, mental retardation, mental diseases, genetic defects, physical accidents and disabilities, birth defects, etc. It will be possible to regrow amputated limbs and restore deteriorated cartilage, organs, and joints. It should also be possible to directly remove deep psychiatric problems, heal chemical imbalances, and address and heal such presently difficult disorders as schizophrenia.

The development and application of Type 2 medical engineering can usher in the greatest, most beneficial revolution of all time. As just *one* example, eventually it will be possible to quickly educate students by simply loading the appropriate "software" into their minds and long term memory, in only a few weeks of "sessions." At that point there need no longer be, anywhere in the world, an educationally disadvantaged person.

It can also be used to usher in the most diabolical revolution of all time if we are not careful to (1) be informed of the Type 2 engineering area, including who is developing what, and for exactly what purposes, and (2) rigorously control the development and use of this technology as it progresses. Our legal, political, and scientific leaders – and we common citizens – will have to grapple with the ethics of what is done and to and for whom it is done. Just as one example, the same technology that will allow direct loading of the software in the long-term memory to achieve rapid education, will allow altering or changing any part of the long term memory, including the built-in emotional responses, etc. *Loading software is loading software*! *What is loaded depends only upon what is in the software.* In theory it is possible not only to alter the learned behavior, memories, emotions, and reactions, but also the genetically "born-in" behavior, emotions, and reactions. The potential for the Type 2 engineering to be used for enslaving and roboticizing humanity should be obvious.

Far more carefully than with any other technology in history, we must ensure that this new one is developed for good, not for evil. To quote an old adage, *the price of freedom is eternal vigilance.* Never before in all of history have those words been more true.

In Conclusion

Pioneers such as Priore, Becker, Kaznacheyev, and Popp have shown us a

---

[93] *In fact, under present law all property, homes, bank accounts, and monies of both the patient and the doctor would simply be confiscated – without any legal charge, arrest, or trial – under an absolutely medieval reinterpretation of U.S. law that has again given police at all levels such sweeping and unconstitutional powers. The Clinton Health Care Plan, presently on hold but still alive and sure to be introduced again, is reported to include the establishment of an entire department solely devoted to confiscating assets from patients and doctors who commit such "crimes." Defendants victimized by the present confiscation reinterpretation of law have a slim chance of even forcing the issue into a court. Even if they succeed in getting into court for a hearing, win it and get full "exoneration," the authorities can still keep the property! In the most insane and unjust legal action in all of U.S. history, this law now charges the property itself with having committed the crime, not the erstwhile human defendant. By what means can the "property" defend itself? No one has an answer, for the property is not an individual and has no legal rights for jury protection under the law! U.S. lawyers and judges are actually insisting that this is rational and makes sense. They have already precisely redefined U.S. law into what is specifically required to institute a police state. The police forces who carry out this consfication are themselves awarded a percentage of the "haul." What does the citizen do when the police forces themselves become the robbers by administrative fiat, and there is no court redress possible? Ostensibly passed to counter organized crime and the drug cartel, this law has rapidly been applied directly to the citizenry, and its use is one of the fastest growing "new techniques" by the various levels of law enforcement. In this paper we have deliberately not dwelled upon the political aspect of the already-emerging police state in control of medical treatment. We simply summarize how the bureacratic medical establishment is likely to meet the challenges of the coming plagues and the associate astounding rise in medical costs. In fact the reader will almost certainly see the rise of an iron "state dictatorship" in medicine, as is already emerging in other legal areas.*

In this paper my primary purpose is technical, not political, so we defer detailing the more involved political aspects of the emerging police state (the New World Order – more aptly the New World Feudalism) to other more capable authors and sources. One central thought should remain firmly in mind All political systems and isms tend ultimately to feudalism, which is always the best system for the oligarchy or elite at the top who control and own everyhing. In any society, there is always an on-going struggle by a fraction of the most powerful to convert the society to special forms of feudalism.

[94] *Some biological agents are very lethal and the associated disease develops at a frightening pace. E.g., the strains of Staphylococcus bacteria that cause necrotizing fascitis and necrotizing myositis – two versions of the "flesh-destroying" disease – are noteworthy examples. Necrotizing is causing inflammation and/or death of tissue, fascitis refers to "sheet" or connective tissue, and myositis refers to muscle tissue. Recently the Quebec separatist leader Lucien Bouchard was suddenly felled by necrotizing myositis, lost a leg, and at this writing is stable and will recover. The disease is usually fatal in 80% of the cases. There is nothing diabolical in his contacting the disease; the bacteria are present from time to time and can enter the body through a nick or otherwise. The disease and the bacteria are known; however, in the past such cases have been relatively rare. The lethality of the disease is augmented by the rapidity of its development. Delayed recognition of the disease may place the patient already beyond the ability of conventional agents to arrest it in time to save him. With the modern hospital and population environment being highly favorable for the more rapid modification and adaptation of disease organisms, even previously rare diseases such as necrotizing fascitis/myositis are ticking time bombs waiting to become resistant to our antibiotics and explode in the dense urban populations.*

[95] *We refer to strikes using biological organisms. We have mentioned but deliberately not dwelled upon the use of electromagnetic induction of lethal cellular diseases at a distance. Such strategic capability to attack a nation and destroy its citizens and soldiers is presently on-line in Russia, under total control by the KGB, not by the conventional Russian military forces.*

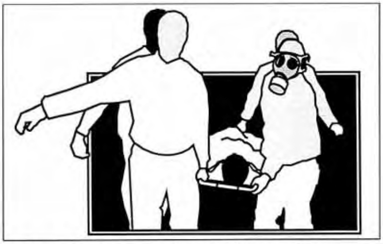

Fig. 21. Present antibiotic therapy is failing as disease organisms adapt.

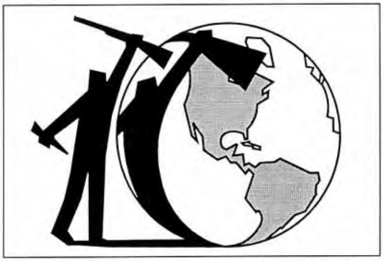

Fig. 22. Terrorists will eventually use biological warfare strikes. Biological warfare is an increasing threat to military forces and the civil populace.

great vision. Having seen it, when one looks at the incredible human suffering throughout the world, and then at the formidable problems of our Type 1 medical science that is struggling so mightily and losing the battle, one is reminded of a phrase by Dante:

"Heaven wheels above you

Displaying to you her eternal glories

And still your eyes are on the ground."

With every fiber of our being, we exhort the scientific community to lift its eyes from the ground, and develop and apply the Type 2 medical technology as rapidly as humanly possible, for human benefit. Else half of humanity will perish in the coming tide of pestilence. Even the conventional scientists agree that the pestilences are coming, and that conventional approaches are already losing the battle. Long ago Albert Einstein wrote these prophetic words:

"*It would of course be a great step forward if we succeeded in combining the gravitational field and the electromagnetic field into a single structure. Only so could the era in theoretical physics inaugurated by Faraday and Clerk Maxwell be brought to a satisfactory close.*"

Einstein's passionately-sought unification of EM and G has now been accomplished. It has magnificent medical science applications for the benefit of all humankind. Meanwhile, in the orthodox medical journals, the scientists themselves are recognizing the demise of antibiotics and the coming unstoppable plagues. The scientists are redoubling their efforts to apply the present inadequate models, and their efforts are admittedly failing. *If they keep to their present course, the war is already lost!*

In the name of all the bleeding, suffering, and dying peoples of the world, we must do better than that! We can do better than that, if we rapidly redevelop and apply the Priore approach.

We call fervently for a great new "War on Disease" that is empowered, funded, and staffed to perform the task. It can be done in perhaps three years, and perhaps for as little as $60-100 million if we streamline procedures, keep the bureaucracy off the team's necks, and give the team scientists the authority of a War Commander with a mission order to *get on with it.*

What happens next is up to the scientists, the politicians, and the concerned leaders of all the nations of the world. *They simply must not fail us.* They and their families are in the same frying pan as the rest of us. This is not the time for dogma, posturing, arrogance, protecting one's territory, yapping about who's going to control the presently inadequate medicine, yakking about pseudoscience, striving for personal position and power, bowing to the big drug companies and their vested interests, and playing the "Who's to be the big monkey?" game. *We need a new Manhattan Project. We need it now.*

Else we who have
sown the wind
will most certainly reap
the coming whirlwind.

ADDITIONAL REFERENCES

Priore, Antoine. (1963) "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes." [Procedure and Assemblage for Production of Radiation Especially Service able for the Treatment of Living Cells.] Republique Francais: Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

Priore, A. (1966) "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.

Priore, A. (1968). "Apparatus for producing radiations penetrating living cells." U.S. Patent No. 3,368,155. Feb. 6, 1968.

Priore, A. French Patent No. 1,342,772. Letter, R. Pautrizel to his colleagues at the University of Bordeaux, Sept. 11, 1979. Letter, R. Pautrizel to his colleagues, Mar. 26, 1969.

Graille, Jean-Michel, ***Le Dossier Priore****, [Dos sier Priore: A New Pasteur Affair] De Noel, Paris, 1984 [in French]. Details the entire Priore affair.*

Pautrizel, R., M.R. Riviere, A. Priore, and F. Berlureau. (1966) "Influence d'ondes electromagnetiques et de champs magnetiques associes sur l'immunie de la souris infestée par Trypanosoma equiperdum," [Influence of electromagnetic waves and associated magnetic fields on the immunity of the mouse infected with the Trypanosoma equiperdum] ***Compt. Rend. Acad. Sci****. (Paris), 1966, Vol.60,p.2099-2102.[in French].*

Pautrizel, R.; A. Priore, F. Berlureau, and A.N. Pautrizel, "Stimulation par des moyens physiques des defenses de la Souris et du Rat contre la trypanosomiase experimentale." [Stimulaton, by physical means, of defenses of the mouse and of the rat against the ex perimental Trypanosoma.] ***Compt. Rend. Acad. Sci****. (Paris), Vol. 268, 1969, p. 1889-1892. [In French].*

Pautrizel, R., A. Priore, A.N. Pautrizel, and P. Chateau-Reynaud-Duprat.(1979) "Guersson de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organisme." [Cure of experimental trypanosomias by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.].

Communication presented to the Journées ***Nationales Microondes - Colloque Hertzienne et Dielectriques,*** *Lille 26-29 juin 1979, p. 31.*

Pautrizel, R.; A. Priore, F. Berlureau, and A.N. Pautrizel, "Action de champs magnetiques combines a des ondes electromagnetiques sur la trypanosomiase experimentale du Lapin." [Action of magnetic waves combined with electromagnetic waves on the experimental trypanosoma of the rabbit.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 271, 1970, p. 877-880.*

Pautrizel, R.; A. Priore, M. Dallochio and R. Crockett, "Action d'ondes electromagnetiques et de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholestrole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modifications in the rabbit by the administration of an alimentary hypercholesterol diet.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 274, 1972, p. 488-491. [In French].*

Pautrizel, R.; A. Priore, P. Mattern, and A.N. Pautrizel, "Stimulation des defenses de la souris trypanosomee par l'action d'un rayonnement associant de champ magnetique et ondes electromagnetiques." [Stimulation of the defenses of the trypanosomized mouse by the action of a ray associated with a magnetic field and electromagnetic waves." ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 280, 1975, p. 1915-1918. [In French]*

Pautrizel, R.; A. Priore, P. Mattern, A.N. Pautrizel, and A. Capbern, "Guerison de la trypanosomiase chronique du Lapin a Trypanosoma equiperdum par l'action combinee de champs magnetiques et d'ondes electromagnetiques modules." [Healing of chronic trypanosomia by trypanosoma equiperdum of the rabbit by the combined action of magnetic waves and modulated electromagnetic waves.] ***J. Protozool.,*** *Vol. 22, No. 3, 1975, p. A 84.*

Pautrizel, R. "La trypanosomiase experimentale: stimulation des defenses de L'organisme par des moyens physiques." [Experimental trypanosomiasis: stimulation of the organism's defenses by physical means.] ***XVIIth Seminar on Trypanosomiasis Research,*** *Londres, 22-23 Sept. 1976.*

Pautrizel, R.; P. Mattern, A.N. Pautrizel, and A. Priore, "Effets des champs magnetiques et des ondes electromagnetiques modules sur la trypanosomiase experimentale." [Effect of magnetic fields and modulated electro magnetic waves on experimental trypanosomiasis]. ***Ann. Soc. Belge Med. trop.*** *[Annals of the Belgium Society of Tropical Medicine], Vol. 57, 1977, p. 501-523.*

Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, A. Capbern, and T. Baltz, "Importance des mecanismes immunitaires dan la guerison de la trypanosomiase experimentale par stimulation physique." [Importance of immune mechanisms in the cure of the experimental trypanosomias by physical stimulation.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 286, 1978, p. 1487-1492. [In French]*

Pautrizel, R. and A. Priore, "Un aspect spectaculaire du bioelectro-magnetisme: permettre a l'organisme de se debarrasser d'un infection aigue ou chronique sans l'aide d'aucune substance medicamentese." [A spectactular aspect of bioelectromagnetism: Permitting the organism to rid itself of an old chronic infection without the aid of any medical substance.] ***Communication presented to the 104eme Congres National des Societes Savantes,*** *Bordeaux, 17-21 Avril 1979 – section* ***Sciences,*** *p. 112.*

Pautrizel, R. and A. Priore, "Guerison de la trypanosomiase experimentale par association de champs magnetiques et d'ondes electromagnetiques." [Cure of trypanosomiasis by associated magnetic fields and electromagnetic waves.] ***Communication presented to the Symposium International des Therapeutiques Ondulatoires,*** *Versailles, 19-20 Mai 1979, p. 9.*

Pautrizel, A. Priore, A.N. Pautrizel and P. Chateau-Reynaud-Duprat, "Guerison de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organism." [Cure of trypanosomiasis by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.] Communication presented to the ***Journees Nationales Microondes - Colloque Hertzienne et Dielectriques,*** *Lille 26-29 Juin 1979, p. 31.*

Riviere, M.R., A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomenes de regression observes sur les greffes d'un lymphosarcome chez des souris exposées a des champs electromagnetiques." [Phenomena of regression observed on the graftings of a lymphosarcoma in mice exposed to electromagnetic fields]. ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 260, 1965, p. 2099-2102 [in French]. This reports the curing of lymphosarcoma, which is a malignant lymphoma that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphhomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue.*

Riviere, M.R.; A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Action de champs electromagnetiques sur les greffes de la tumeur T8 chez le Rat." [Action of the electromagnetic fields on the graftings of the T8 tumor on the rat." ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 259, 1964, p. 4895-4897.*

Riviere, M.R.; A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Effets de champs electromagnetiques sur un lymphosarcome lymphoblastique transplantable du Rat." [Effects of electromagnetic fields on lympho blastic lymphosarcoma transplantable from a rat.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 260, 1965, p. 2099-2102.*

COMMENT: Lymphosarcoma is a malignant lymphoma (malignant tumor of lymph or lymph-like tissue) that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue. A lymphoblast is a cell giving rise to lymphocytes, the colorless weakly motile cells produced in lymphoid tissue and include the cellular mediators of immunity, constituting some 20 to 30 percent of the leukocytes (white blood cells) of normal human blood. In effect the experimenters have healed a transplanted condition that represents a serious leukemia.

Riviere, M.R. and M. Guerin., "Nouvelles recherches effectuees chez des rats porteurs d'un lymphosarcome lymphoblastique soumis a l'action d'ondes electromagnetiques associees a des champs magnetiques." [New research on rats having lymphoblastic lymphosarcoma, submitted to the action of electromagnetic waves associated with magnetic fields.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 262, 1966, p. 2669-2672.*

COMMENT: Lymphosarcoma is a malignant lymphoma (malignant tumor of lymph or lymph-like tissue) that tends to metastasize freely and spread throughout the body. It easily leads to the condition of lymphomatosis, which is the presence of multiple lymphomas in the body, including wide distribution through the various organs, lymph tissue, and tissue resembling lymph tissue. A lymphoblast is a cell giving rise to lymphocytes, the colorless weakly motile cells produced in lymphoid tissue and include the cellular mediators of immunity, constituting some 20 to 30 percent of the leukocytes (white blood cells) of normal human blood. In effect the experimenters have healed a condition that represents a serious leukemia.

Delmon, G. and J. Biraben, "La croissance du carcinome de Guerin sour l'action de champs electromagnetiques." [The growth of carcinoma treated by the action of electromagnetic fields.] ***Rev. Path. Comp.,*** *3(2), 1966, p. 85-88.*

Berteaud, A.J.; A.M. Bottreau, A. Priore, A.N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation entre et l'action d'une onde electromagnetique pulsee et modulee." [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 272, 1971, p. 1003-1006. [In French].*

Mayer, G.; A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiques sur l'orchite trypanosomienne due lapin." [Action of magnetic fields associated with electromagnetic waves on the typanosomian orchitis of the rabbit.] ***Compt. Rend. Acad. Sci.*** *(Paris), Vol. 274, 1972, p. 3011-3014. [In French]*

COMMENT: Orchitis is inflammation of a testes. Trypanosomes are protozoan flagellates of genus trypanosoma which infect humans and animals and are responsible for various serious diseases such as Chaga's disease, dourine, nagana, sleeping sickness, and surra. Chaga's disease is marked by prolonged high fever, edema, and enlargement of the spleen,

liver, and lymph nodes. Dourine is a disease favoring horses and asses, marked by inflammation of the genitals, subcutaneous edematous plaques, low-grade fever, progressive paralysis, emaciation, and death. Nagana is a highly fatal disease in/of domestic animals in tropical Africa marked by fluctuating fever, inappetance, edematous swelling, and sluggishness, and is transmitted by the tsetse fly and possibly by other biting flies. Sleeping sickness is a serious disease that is prevalent in much of tropical Africa; it is marked by fever, protracted lethargy, tremors, and loss of weight, and is transmitted by tsetse flies. Surra is a severe Old World febrile and hemorrhagic disease of domestic animals and is transmitted by biting insects.

Courrier, R., "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reunion a L'Institut sur les effets de la Machine de M.A. Priore le 26 Avril 1977." [Presentation by Professeur R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M.A. Priore.] [In French] Courrier's presentation of the Priore machine and its positive cures of terminal cancers and leukemias in laboratory animals, under proper scientific protocols.

Ruth, Barnard. "Experimental investigations on ultraweak photon emission." ***In Electromagnetic Bio-Information: Proceedings of the Symposium****, Marburg, Sept. 5, 1977. Proceedings edited by Fritz Albert Popp et al., Urban and Schwarzenberg, Munchen Wien, and Baltimore, 1979, p. 107-122.*

Ryaby, J.T.; D.A. Grande, F.P. Magee, and A.M. Weinstein, "The effect of combined AC/DC magnetic fields on resting articular cartilage metabolism." Presented at the 1st World Congress for Electricity and magnetism in Biology and Medicine, Orlando, FL, June 14-19, 1992.

Vladimirov, Yu. A., "Ultraweak luminescence of subcellular structures," ***Ultraweak Luminescence in biology, Proceedings of the Symposium,*** *Moscow, 1969 [In Russian].*

Kaznacheyev, V.P., "Informational function of ultraweak flows of light in biological systems," in ***Materials of Science Conference: Physical and Mathematical methods of Research in Biology and Medicine, Novosibirsk,*** *1965, p. 38-41 [In Russian].*

Kaznacheyev, V.P. et al, "Studies on the biological role of electromagnetic emission as a factor of adaptive behavior of cells under conditions of lattitude changes," ***Abstracts of the Second All-Union Conference Devoted to different Geographical, Climatic and Labor Conditions,*** *Novosibirsk, 1977, p. 101-104 [In Russian].*

Kaznacheyev, V.P. et al., "Conditions necessary for appearance of distant intercellular interactions after UV radiation," ***Bulleten Experimentalnoy Biologii i. Meditsiny,*** *No. 5, 1979, p. 468-471.*

Possony, Stefan T., "Psy-War: Soviet device experiment," ***Defense and Foreign Affairs Daily,*** *12(104), June 7, 1983, p. 1-2.*

Kaznacheyev, V.P. and L.P. Mikhailova, "Ultraweak Radiations in Intercellular Interactions," Novosibirsk, 1981 [In Russian]. English translation: San Francisco, Washington Research Center, 1982.

Kaempffert, Waldemar, "'Dead' organisms revived," Science in the News: ***Science Digest****, May 1950, p. 77-78.*

"Rejuvenating Light," ***Scientific American*** *May 1949.*

Popp, Fritz Albert; Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., ***Electromagnetic Bio-Information: Proceedings of the Symposium****, Marburg, Sept. 5, 1977, Urban & Schwarzenberg, Baltimore, 1979.*

Lilienfeld, A.M. et al, ***Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees From Selected Eastern European Posts****, Department of Epidemiology, Of Hygiene, and Public Health; The Johns Hopkins University, Baltimore, Maryland, final Report, July 31,1978.*

Mallaliew, Robert C., ***A Model of the Microwave Intensity Distribution within the U.S. Embassy, Moscow 1966 to 1977,*** *Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980.*

Microwave Radiation at the U.S. Embassy in Moscow and Its Biological Implications: An Assessment*, U.S. Department of Commerce Doc. No. NTIA-81-12, Mar. 1981.*

Novak, Rachel. "Flesh-Eating Bacteria: Not New, But Still Worrisome." Research News, ***Science****, 264(5166), June 17, 1994, p. 1665.*

Olariu, S. and I. Iovitzu Popescu. "The Quantum Effects of Electromagnetic fluxes," ***Reviews of Modern Physics****, 57(2), April 1985, p. 339-436.*

Opall, Barbara. "U.S., Russia Hope to Safeguard Mind-Control Techniques." ***Defense News****, Jan. 11-17, 1993, p. 4, 29.*

Penrose, R. "The Nonlinear Graviton." ***General Relativity and Gravitation****, 7(2), 1976, p. 171.*

Perez, Jose and Carson Jeffries. "Effects of additive noise on a nonlinear oscillator exhibiting period doubling and chaotic behavior." ***Physical Review B. Condensed Matter****. Series 3, 16(6), Sept. 15, 1982, p. 3460-3462.*

Phipps, Thomas E. Jr. "Critique of accepted theories of special relativity and electromagnetism," ***Scientific Ethics****, 1(2), Apr. 1985, p. 17-23.*

Popp, Fritz-Albert and Ke-hsueh Li. "Hyperbolic Relaxation as a Sufficient Condition of a Fully Coherent Ergodic Field." ***International Journal of Theoretical Physics****, 32(9), Sep. 1993, p. 1573-1582.*

Reid, B.L. "Propagation of properties of chemical reactions over long distance in the atmosphere as seen by crystal growth pattern changes." ***Australian Journal of Medical Laboratory Science****, Vol. 7, Feb. 1986, p. 30-35.*

Rodemann,. H.P., H.P. Peterson, K. Wchwenke, and K. Hartmut von Wangenheim. Terminal differentiation of human fibroblasts is induced by radiation." ***Scanning Microscopy****, 5(4), Dec. 1991, p. 1135-43.*

Rodemann, H.P.; K. Bayreuther, and G. Pfleiderer. "The differentiation of normal and transforman fibroblasts in vitro is influenced by electromagnetic fields." ***Experimental Cell Research,*** *Vol. 182, 1989, p. 610-621.*

Santilli, Ruggero Maria. "Partons and gravitation: some puzzling questions." ***Annals of Physics****, 83(1), Mar. 1974, p. 108-157. Proves that electric field and gravitation are not mutually exclusive. One is left with one of two assumptions: (1) the strong assumption, that both are identically the same thing, or (2) the weak assumption, that both are partially the same thing.*

Saxl, Erwin J. and Mildred Allen. "1970 solar eclipse as 'seen' by a torsion pendulum." ***Physical Review D. Particles and Fields,*** *3(4), Feb.15, 1971, p. 823-825.*

Sheldrake, Rupert. ***A New Science of Life****: The Hypothesis of Formative Causation, Blond & Briggs, Ltd., U.S. reprint, J.P. Tarcher, Inc., Los Angeles, 1982.*

Smetanin, E.V., "Electromagnetic field in a space with curvature — new solutions," ***Soviet Physics Journal****, 25(2), Feb. 1982, p. 107-111.*

Tapscott, Mark. "DOD, Intel Agencies Look at Russian Mind Control Technology, Claims." ***Defense Electronics****, 25(8), July 1993, p. 17.*

von Borzeszkowski, Horst-Heino. "Remarks on the Physical Reality of Gravitons," ***Foundations of Physics****, 20(4),1990, p. 435-445.*

Wheeler, John Archibald and Richard Phillips Feynman. "Classical electrodynamics in terms of direct interparticle action," ***Reviews of Modern Physics****, 21(3), July 1949, p. 425-433.*

Goodman, Glenn W. Jr. "Upping the Nonlethal Ante," ***Armed Forces Journal International****, July 1994, p. 13.*

Jammer, Max. ***Concepts of Space: The History of Theories of Space in Physics****. Harvard University Press, Cambridge, Massachusetts, 1969*

Jammer, Max. ***Concepts of Force****. Harvard University Press, Cambridge, Massachusetts, 1957.*

Jammer, Max. ***Concepts of Mass****. Harvard University Press, Cambridge, Massachusetts, 1961.*

Javier, R.T.; F. Sedarati, and J. G. Stevens. "Two avirulent Herpes simplex viruses generate lethal recombinants in vivo." ***Science****, 234(4777), Nov. 7, 1986, p. 746-747.*

Wheeler, J.A. and R.P. Feynman. "Classical Electrodynamics in Terms of Direct Interparticle Action," ***Reviews of Modern Physics****, Vol. 21, 1949, p. 425-433. The electromagnetic field as such does not actually exist in spacetime, but is a summary of what we need to know about distant charges in order to predict the behavior of a charge if there should be one at that point. In other words, the EM field represents an information storage device.*

Jaynes, E.T. "Probability In Quantum Theory," in ***Complexity, Entropy, and the Physics of Information****, Ed. W.H. Zurek, 1990, pp. 33-55, Addison-Wesley, Redwood City, CA. Strongly points out that the EM field does not exist in spacetime, but represents an information storage device.*

Editorial Note: *We would like to express our appreciation to Mr. Bearden for this wonderful article.*

The Tom Bearden Website

"Dossier Priore"

by Jean-Michel Graille

Photographs

Cover	
Plate 1	The inauguration: the construction of the M 600 had not yet started but, symbolically, the Mayor of Bordeaux starts the smaller machine at close range. (Archives R.P.)
Plate 2	Inaugural scroll.
Plate 3	An information meeting held in Bordeaux. From left to right: Antoine Priore, Robert Courrier, Raymond Pautrizel and Francis Berlureau. (Archives R.P.)
Plate 4	February 1975. Antoine Priore and Professor Pautrizel at the "suspended" base of the M 600, the enormous experimental prototype built by Leroy-Somer. The apparatus would only function for a few days but did make it possible to carry out some experiments with very promising results.
Plate 5	Anne-Nelly Pautrizel. Professor Pautrizel's sister and collaborator, who participated in almost all the experiments with the Priore apparatus since 1968. (Photo J.M.G.)
Plate 6	Pierette Châteaureynaud-Duprat, head of research at C. N.R.S., doctor of science and specialist in grafts, carried out in 1969 and 1970 two series of decisive experiments concerning the proof and the explanation of the biological effectiveness of the Priore Ray.

Plate 7	The Priore machine. The model financed by Pautrizel, the only one which is still likely to function, the one with which patients will be officially looked after as from 1977, is finished in August 1968. In fact, the apparatus is installed all in one place.
Plate 8	1969. Old friends of Priore retrace the road to Floirac. With a monkey-wrench or soldering-iron in hand, they devote all their leisure time to building the new machine for their friend. Here, on the scales, inspector Plantin, of the legal identity service. In coveralls, scarf around his neck, municipal worker M. Léglise. (Archives R.P.)
Plate 9	A courageous man. Perpetual Secretary of the Academy of Sciences, elder President of the Academy of Medicine, "father" of endocrinology and fertilization, Professor Robert Courrier could have benefited from its prestige and its retirement. At 67, he took all the risks to defend that which he believed in deeply. He has fought now for twenty years. One of the rare photos of him, in the company of Antoine Priore and Professor Pautrizel, taken on 25th March 1970 in the laboratory at Floirac. (Archives R.P.)

Jean-Michel Graille

DOSSIER PRIORE

Une nouvelle affaire Pasteur?

Denoël

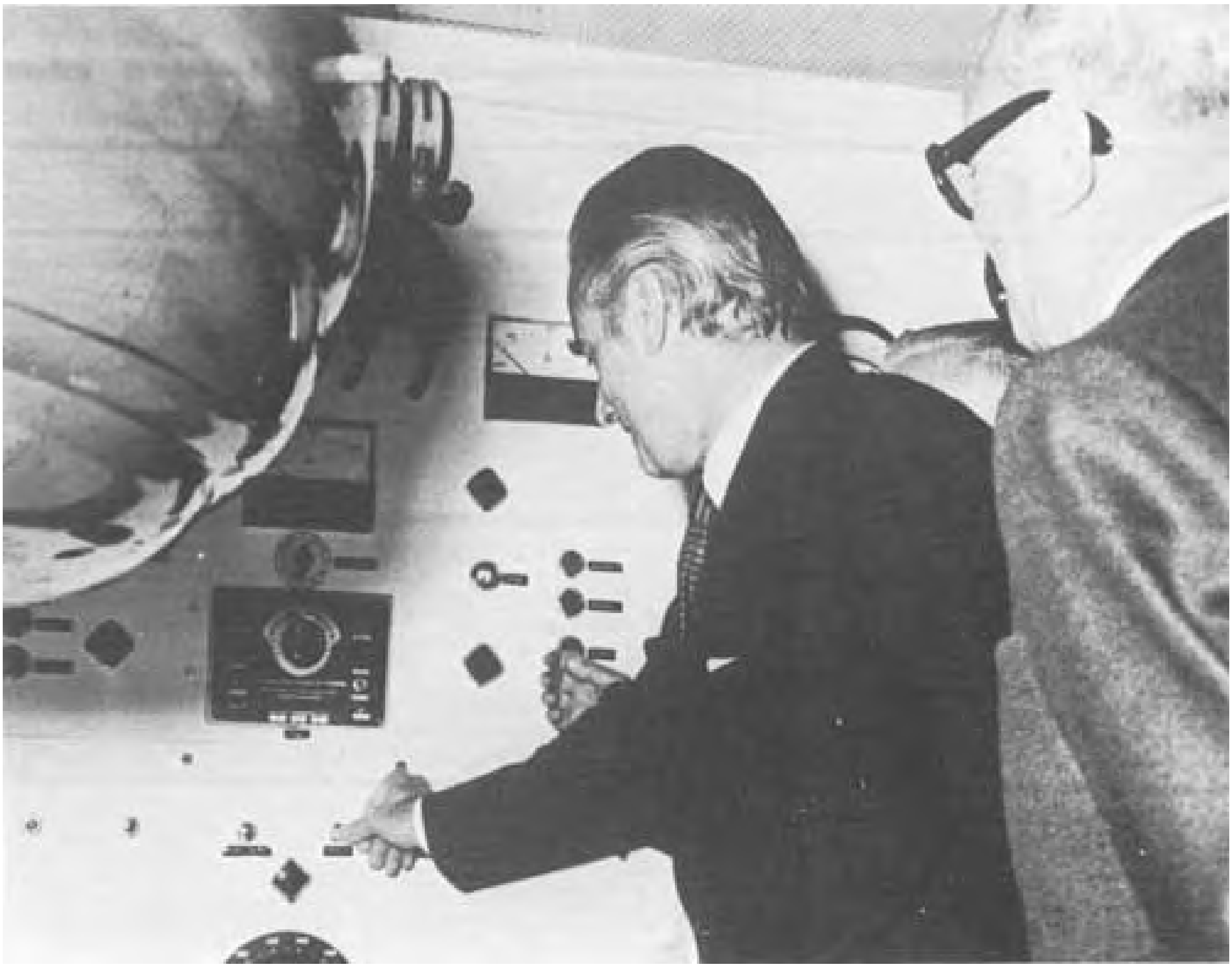

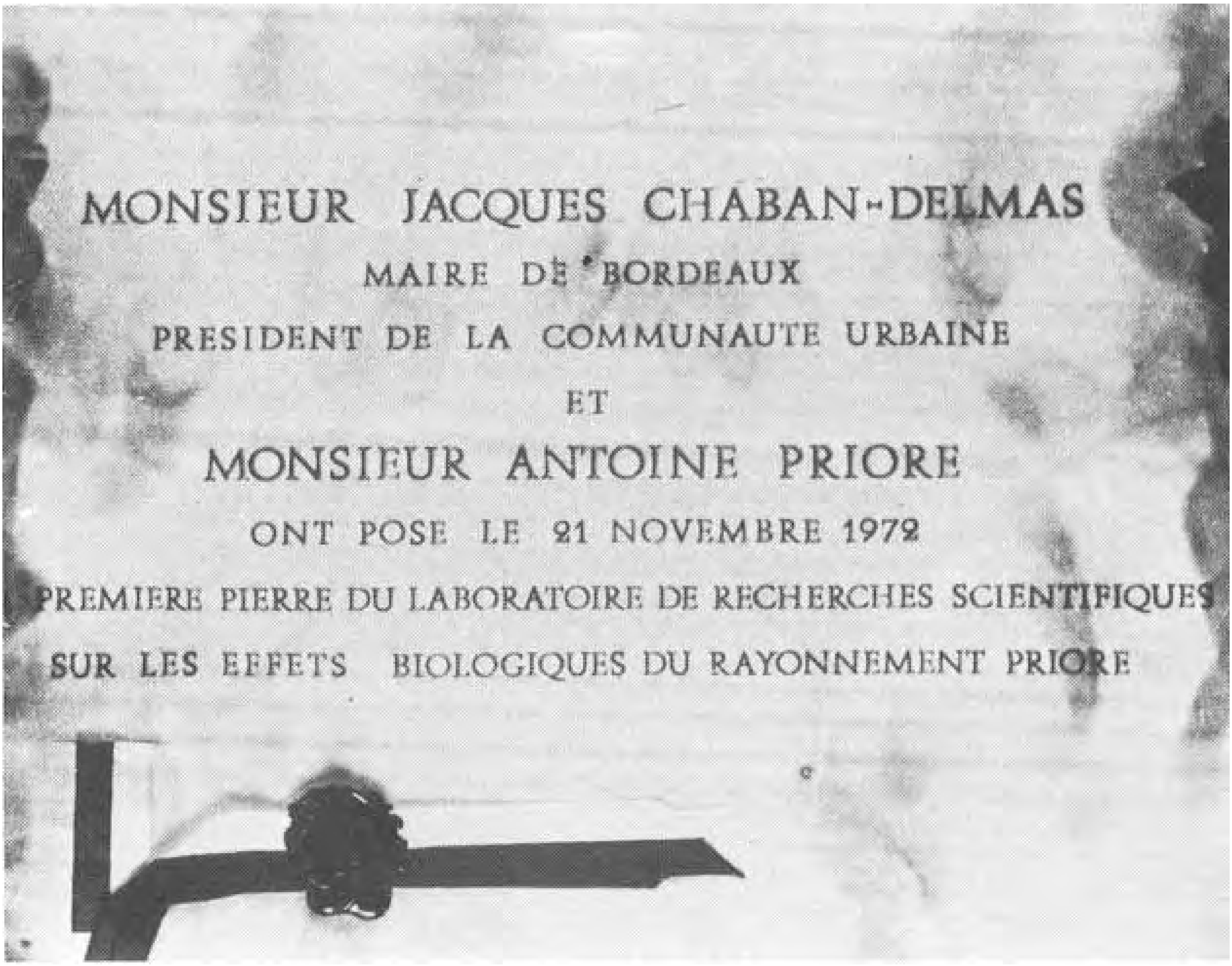

MONSIEUR JACQUES CHABAN-DELMAS
MAIRE DE BORDEAUX
PRESIDENT DE LA COMMUNAUTE URBAINE
ET
MONSIEUR ANTOINE PRIORE
ONT POSE LE 21 NOVEMBRE 1972
PREMIERE PIERRE DU LABORATOIRE DE RECHERCHES SCIENTIFIQUES
SUR LES EFFETS BIOLOGIQUES DU RAYONNEMENT PRIORE

Translation:

JACQUES CHABAN-DELMAS
THE MAYOR OF BORDEAUX
PRESIDENT OF THE URBAN COMMUNITY
AND
ANTOINE PRIORE
HAVE INSTALLED ON THE 21 NOVEMBER 1970
THE FIRST STONE OF THE SCIENTIFIC RESEARCH LABORATORY
FOR THE BIOLOGICAL EFFECTS OF THE PRIORE RAY

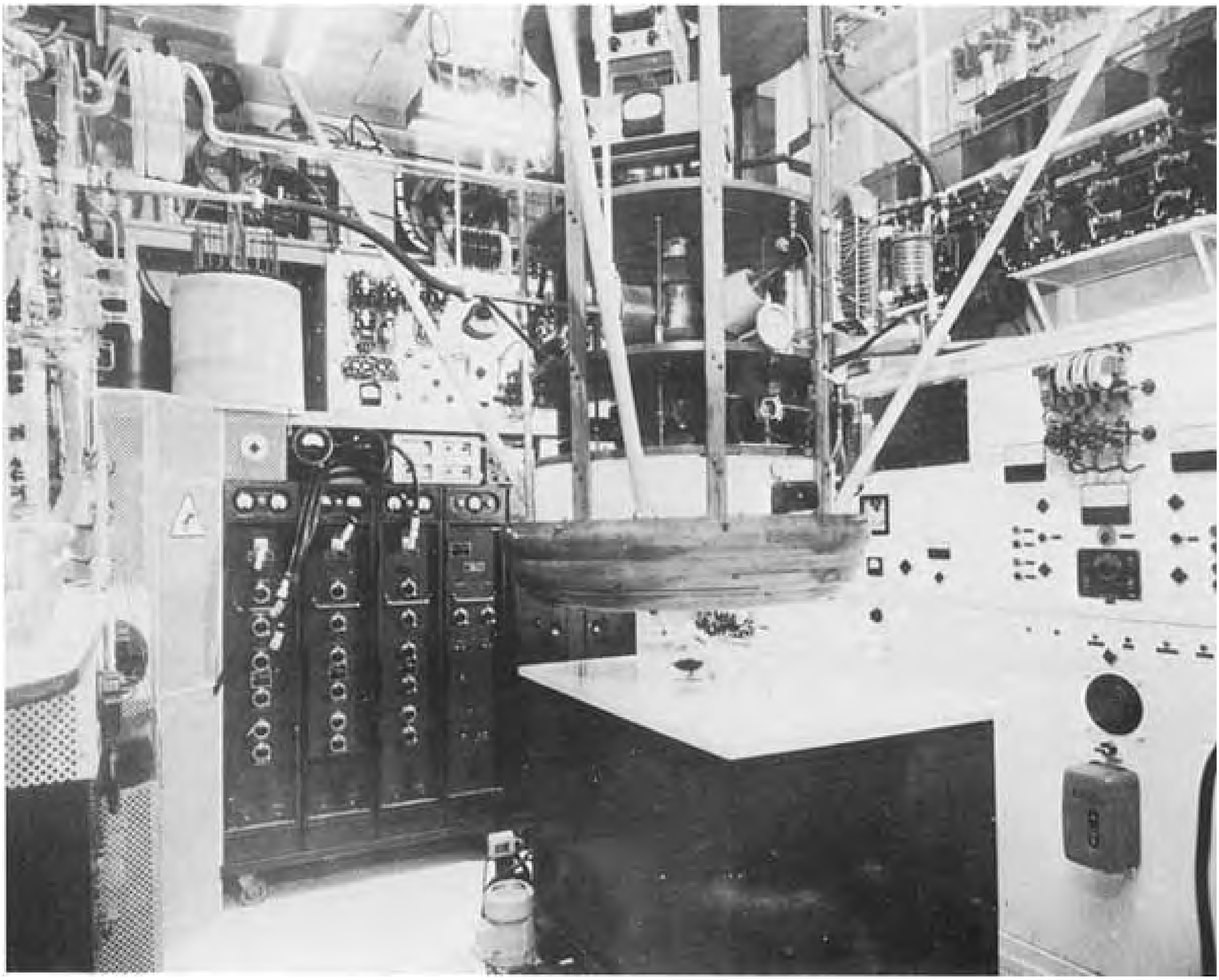

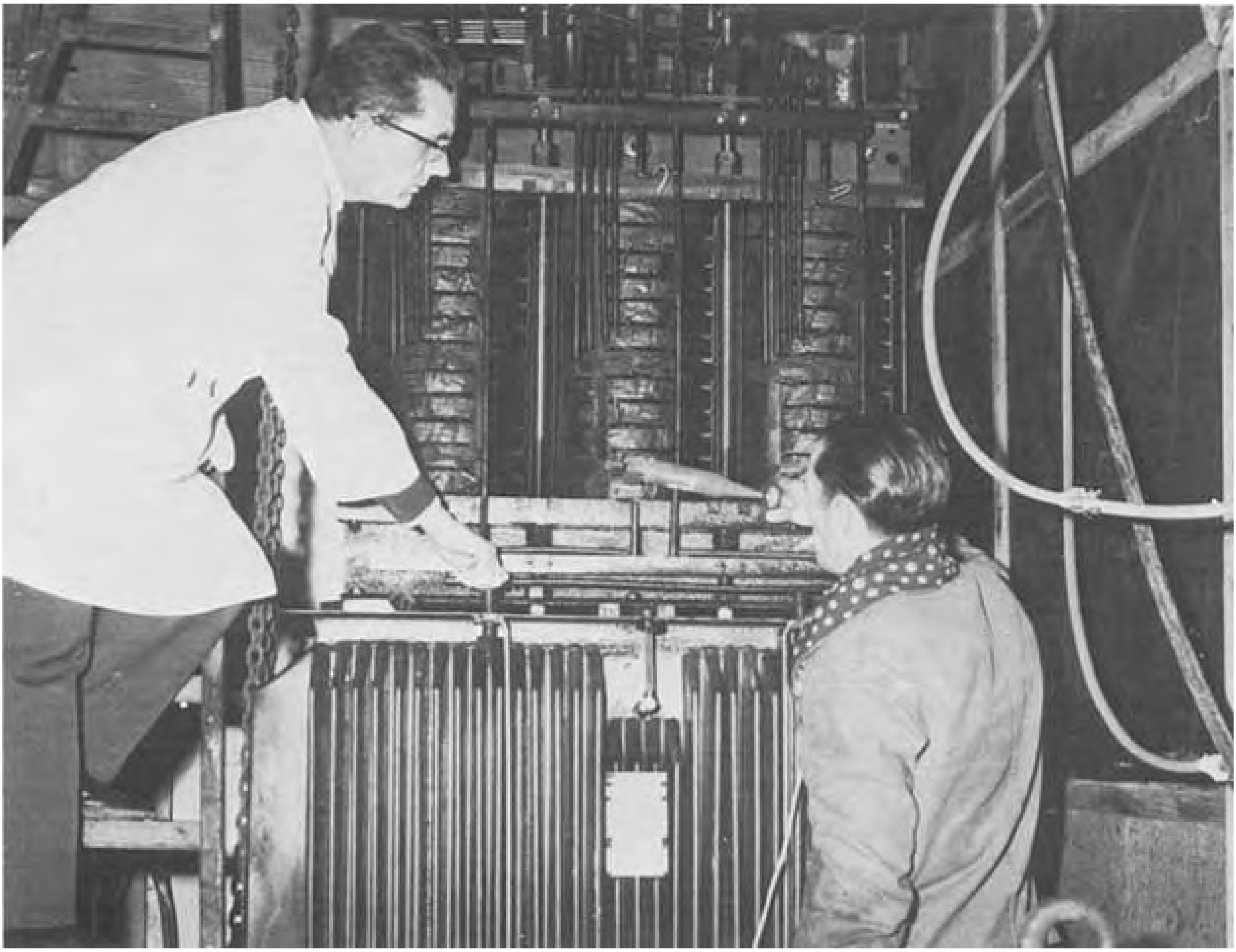

Making Laser Weapons a Reality

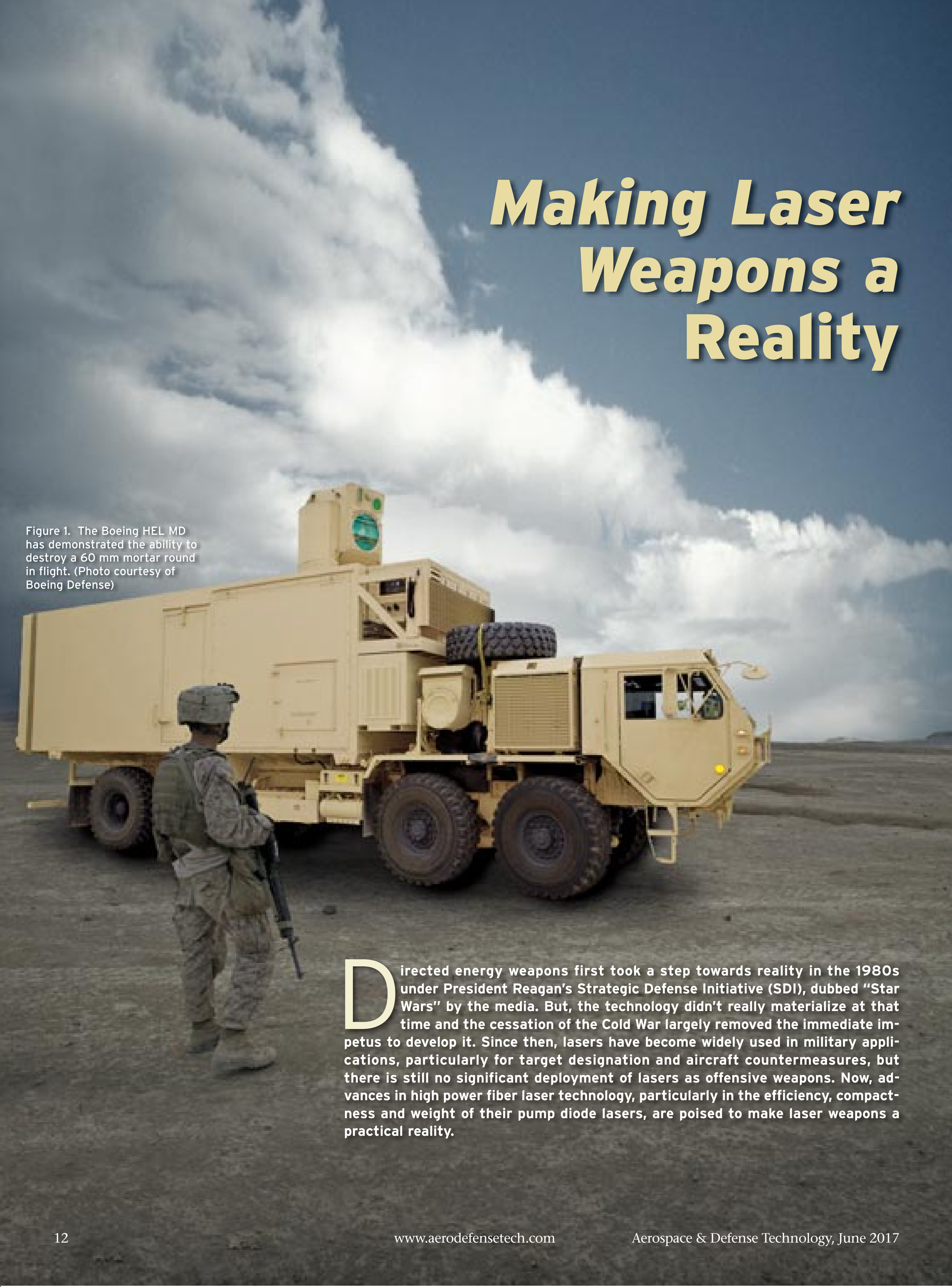

Figure 1. The Boeing HEL MD has demonstrated the ability to destroy a 60 mm mortar round in flight. (Photo courtesy of Boeing Defense)

Directed energy weapons first took a step towards reality in the 1980s under President Reagan's Strategic Defense Initiative (SDI), dubbed "Star Wars" by the media. But, the technology didn't really materialize at that time and the cessation of the Cold War largely removed the immediate impetus to develop it. Since then, lasers have become widely used in military applications, particularly for target designation and aircraft countermeasures, but there is still no significant deployment of lasers as offensive weapons. Now, advances in high power fiber laser technology, particularly in the efficiency, compactness and weight of their pump diode lasers, are poised to make laser weapons a practical reality.

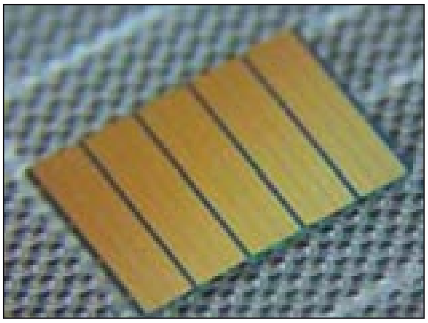

Figure 2. A Coherent | DILAS T-Bar chip contains five emitters, outputs about 50 W, and is 5 mm wide.

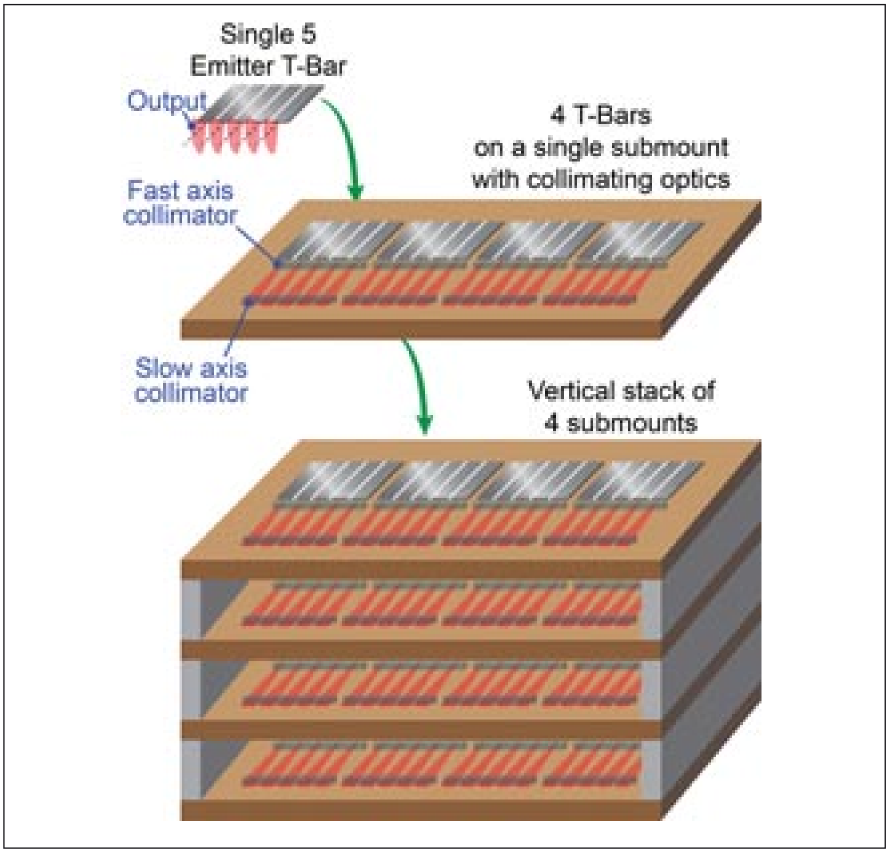

Figure 3. Progression from a single T-Bar die to an 80-emitter stack. The light from all these emitters can be efficiently coupled into a single 225 μm core fiber with a 0.22 NA.

Laser Weapons Overview

The concept behind Reagan's SDI was to protect the United States behind a "missile shield" that would prevent nuclear weapons, delivered via either intercontinental or submarine-launched ballistic missiles, from reaching North America. Part of the SDI was to be lasers, or other directed energy weapons, placed in permanent orbit, which could rapidly knock out a nuclear missile already in flight towards its target. Lasers, in particular, were a promising technology for this application because they can be quickly "charged up" (i.e. be made ready to fire), and their output can reach a fast-moving target nearly instantaneously.

Because these lasers were intended to be deployed primarily in space, or in fixed locations, and used either never (hopefully) or just once, there wasn't much concern about their cost, physical size, weight or operational efficiency. The key goal was simply that they would work when needed, delivering the required knock-out punch.

The offensive laser weaponry under development today is intended to serve almost the exact opposite purpose than the SDI, and therefore has very different goals in terms of its design and functionality. For example, one major laser weapons initiative is for C-RAM (Counter Rocket, Artillery and Mortar). These are systems intended to be deployed in the battlefield to protect troops and equipment from incoming projectiles. Currently, these are envisioned to be single, vehicle mounted lasers, with output powers in the 10 kW to 50 kW range.

Boeing has already successfully prototyped such a system, which they call the High Energy Laser Mobile Demonstrator (HEL MD). This consists of a 10 kW solid state laser installed on an Oshkosh Tactical Military Vehicle, along with all the necessary targeting and control systems. HEL MD has proven the ability to lock on to and destroy a mortar round of about 10 inches in length, traveling at hundreds of miles per hour, from several miles out. The system is also effective against Unmanned Aerial Vehicles (UAVs or drones); in this case, it may be sufficient to damage the drone's navigation and targeting systems, rather than completely destroy it. Similarly, other offensive laser weapons systems are being designed to deliver "swarm defense," that is, to protect a ship or base from attack by a large number of inexpensive drones.

Thus, today's laser weapons are meant to defend against swarms of inexpensive mortar rounds, drones and other projectiles, and are intended to be mounted on platforms which are mobile and self-contained (meaning they may not have ready access to an external source of power). This gives rise to several design imperatives. First, they must deliver a low cost per engagement; it's simply not economically feasible to employ an expensive weapon (such as a missile costing over $100,000) to knock out a drone which costs $1,000, or a mortar round which might cost less than $100. The laser weapon must also be capable of rapid fire so that it can't just be overwhelmed by a large number of simultaneous incoming rounds. Also, rapid fire reduces the number of individual laser weapons systems needed to protect a given number of troops. But, the ability to deliver rapid fire also necessitates that the system be electrically efficient. Otherwise the laser weapon will require access to a large quantity of fuel, which is difficult (and dangerous) to transport into a battlefield. Finally, it's desirable that laser weapons be physically compact, especially for airborne use.

SWaP Optimization

The push to optimize the size, weight and power (called SWaP in military parlance) of field deployed laser weapons has driven a progression in the technology of the gain material used from chemical (e.g. deuterium fluoride), to solid state, and, most recently, to fiber. One reason for this is because fiber lasers offer inherently higher efficiency, in terms of converting input pump energy to usable output, than nearly any other laser type, except diode lasers. However, diode lasers by themselves don't provide the necessary brightness or beam quality, while diode-pumped fiber lasers can. And, beam quality is critical because it determines the distance over which the laser can be focused to a spot size small enough to reach the power density necessary to damage its target.

Fiber lasers offer other advantages in terms of SWaP optimization. Chemical and lamp pumped solid state lasers require a significant overhead in terms of the equipment and power supplies required to run them. In contrast, both solid state (e.g. slab and rod) and fiber lasers can be diode pumped, and diode power supplies and pump modules themselves are electrically efficient and lend themselves to miniaturization. This efficiency, in turn, reduces the cooling requirements, and all its attendant equipment (pumps, heat exchangers, etc.).

But, as laser weapons evolve towards ever higher powers, fiber laser technology becomes increasingly attractive over solid state. Currently, individual fiber lasers can deliver up to about 2 kW of power, but multiple units can be combined to deliver around 10 kW in a single beam with extremely good mode quality. It's a bit more difficult to scale up the power of a solid-state laser, especially while maintaining good mode quality. This is because scaling up and/or increasing the pump power to a traditional solid state laser typically excites higher order modes, and introduces other issues related to thermal lensing.

Of course, there are also challenges to increasing the power of a fiber amplifier. And, because the optical efficiency of their pumping is already so high (~85%), there's not much room for improvement there. So, the focus is on raising the pump power itself, without increasing package size and weight, as well as on methods for low loss beam combining of multiple single mode lasers, all while still maintaining good beam quality.

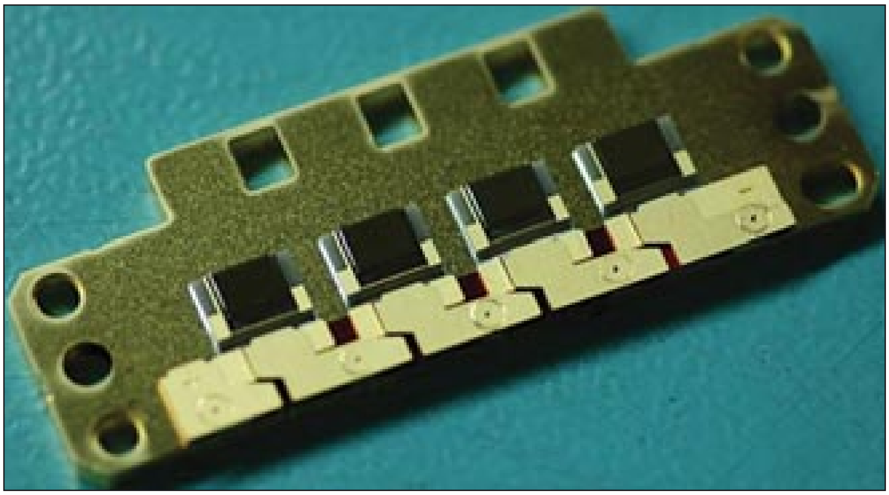

Figure 4. Four T-Bars mounted on a macrochannel cooler with collimating optics.

Pumping Technology Advances

In terms of SWaP optimization of pump diodes for laser weapons, a current benchmark is to achieve a weight to output power ratio of 1 kg/kW. Coherent | DILAS has developed several advances in diode laser technology to reach this goal, and, in fact, is already working on devices that will attain the next level of performance – 0.5 kg for 1 kW of output.

One key to reaching the SWaP benchmark has been the company's introduction of T-Bar (for "tailored" bar) construction, a design approach intended to combine the high total output power of traditional diode laser bars with the relaxed cooling requirements of single emitters. The basic unit of the T-Bar is a diode laser mini-bar having five emitters on a single, 5 mm wide substrate which outputs about 50 W total. In contrast, traditional diode laser bars are usually 10 mm wide, contain at least 19 emitters, and can output over 200 W, depending on the number of emitters and type of heatsink.

For military applications, four of these T-Bar dies are mounted on to a single substrate, yielding a total of 20 emitters. Then, up to four of these substrates are stacked vertically, bringing the total number of individual emitters in this compact assembly to 80, with a total output of about 800 W.

The key optical characteristic of the T-Bar design is that the combination of wide emitter spacing, low divergence and relatively low beam parameter product (in both fast and slow axes) is tailored to enable the light from all 80 of these individual emitters to be efficiently co-linearized and coupled into a single 225 µm core fiber having a numerical aperture of only 0.22. This, in turn, permits highly efficient coupling into the pumping mode volume of the gain fiber. And, this light collection can be accomplished using a relatively simple and compact optical system. This is how output power per unit volume is maximized.

In contrast, the high divergence and poor mode quality of traditional diode laser bars necessitate the use of complex optics for light collection, and still make it impossible to couple all their output into a single, low numerical aperture fiber. Furthermore, high power diode laser bars often suffer from a problem called "smile." This is a sagging or bending alignment error in emitter position along the bar which makes it quite difficult to collimate and co-linearize its output.

Besides the advantage of producing the largest amount of fiber coupled light out in a given system volume, T-Bar construction also offers significantly improved cooling characteristics over traditional bars. In a traditional diode laser bar, the small emitter spacing causes significant thermal crosstalk between the individual emitters, and a very high efficiency cooling system must therefore be employed. This is usually in the form of a so called "microchannel" cooler, in which water is rapidly circulated through channels within the heatsink. But the small bore of the microchannels (tens of microns) makes it easy for them to be clogged by particulates. Thus, the water must be stringently conditioned and filtered, and large, heavy, high pressure pumps must be employed to circulate it rapidly enough to achieve the required level of cooling in these traditional bars.

Furthermore, traditional diode bars are usually placed directly on the heatsink in order to achieve good thermal contact. However, this makes the microchannel cooler part of the electrical circuit, thus necessitating the use of deionized water in order to avoid a short circuit. This imposes yet another requirement that increases cost and complexity.

The larger emitter spacing of the T-Bar largely eliminates this thermal crosstalk, and greatly relaxes the attendant cooling requirement. As a result, Coherent | DILAS has been able to introduce a "macrochannel" cooler. This is a system which uses substantially larger bore channels, thus allowing the use of less stringently filtered tap water, and eliminating the need for high pressure pumping. Also, the reduced cooling requirement enables the laser bar to be placed on a submount, rather than in direct electrical contact with the cooler, which avoids the necessity of using deionized water.

T-Bar construction also delivers better cooling characteristics than single emitter diode lasers. This is because single emitters are typically supplied already packaged. Since the diodes are mounted within the package, there is an extra thermal interface which makes it more difficult to efficiently cool them.

In conclusion, after decades of imagining, offensive laser weapons are nearly ready for deployment on the battlefield. Advances in diode laser pumping technology are proving to be a key enabling factor in this development. And, these military grade pump diode lasers can be constructed using the same automated, high volume fabrication equipment used for industrial product manufacturing, meaning they can be readily delivered at competitive prices.

This article was written by Chris Ebert, Ph.D., Director, Business Development, Coherent | DILAS, (Tucson, AZ). For more information, visit ***http://info.hotims.com/65852-501****.*

The Tom Bearden Website

Table of Contents

A Sensitive Scalar Wave Detector

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11. Figure 11. NETWORK OF VIRTUAL TRANSMITTERS
12. Figure 12. VIRTUAL TRANSMITTERS IN THE INTERFERENCE GRID
13. Figure 13. GIANT RADIAL RELATED TO A VIRTUAL TRANSMITTER
14. Figure 14. FORMATION OF A SINGLE GIANT RADIAL
15. Figure 15. FORMATION OF A DOUBLE GIANT RADIAL
16. Figure 16. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

Return to Homepage

from **STAR WARS NOW!**

The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization
by Lt. Col. T.E. Bearden (retd.), 1984
Copyright

A Sensitive Scalar Wave Detector

Very briefly we present a method of making a very sensitive scalar wave detector so that direct measurement and quantization can be established.

First, we regard one oscillation (one wavelength) of the scalar potential wave as a longitudinal photon. Note that this photon contains a substructure, which may be deliberately determined (when artificially made) or randomized (when naturally made in the idealized case).

For convenience we represent normal linearized vacuum (spacetime) as a horizontal or longitudinal vector (Figure 4), implying the direction of motion of the wave in the laboratory frame. By horizontal position in our diagram, we imply a linear, uncurved spacetime and a non-rotated frame. We represent the longitudinal scalar wave as a horizontal vector, and the usual Hertzian wave as a vertical or "transverse" vector. We visualize a normal detector as detecting only a vertical or "transverse" vector, as we have illustrated in Figure 4.

As can be seen, in a linear, unrotated or uncurved spacetime a pure scalar wave has no vertical component projected upon the laboratory frame vector, so it is not detectable by normal detectors.

To detect the scalar wave, of course we could bend it so that it has a projected vertical component in the laboratory frame (Figure 5). However, this would be an impure wave, not a pure scalar wave, and that is not what we wish.

A better way is to bend or curve spacetime itself in a small region, so that a longitudinal wave that passes through that region now possesses a vertical component with respect to that region (Figure 6). Thus a normal detector there will detect that vertical component. We conduct the detection current out of the "bent spacetime" region to an outside (normal) detector, and we then have a scalar wave detector.

To illustrate, we show conceptually how this has been successfully done. Figure 7 shows the concept. First, we utilize a magnetic pole to provide the infolded energy (potential) to bend or curve spacetime. To reach good sensitivity, we need a pole strength connected with a magnetic field strength of 40,000 Gauss or higher. We utilize a small superconducting magnet, which can reach field strengths of from 40,000 to 80,000 Gauss.

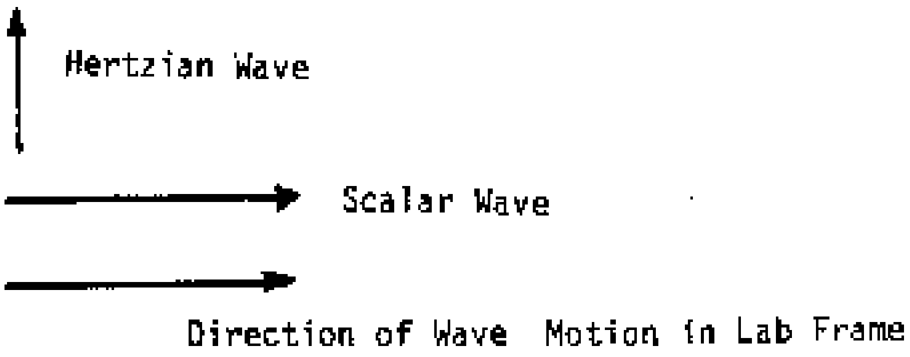
Hertzian Wave
Scalar Wave
Direction of Wave Motion in Lab Frame

Figure 4. Linear, uncurved laboratory frame.

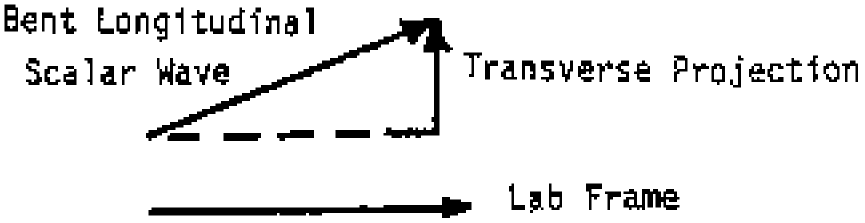
Bent Longitudinal Scalar Wave
Transverse Projection
Lab Frame

Figure 5. Rotating the longitudinal wave produces a transverse component.

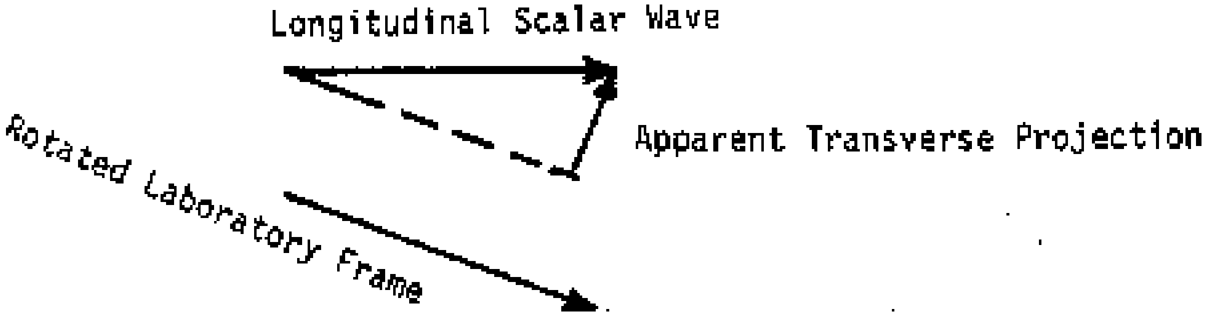
Longitudinal Scalar Wave
Apparent Transverse Projection
Rotated Laboratory Frame

Figure 6. Bending the laboratory spacetime frame produces a transverse component.

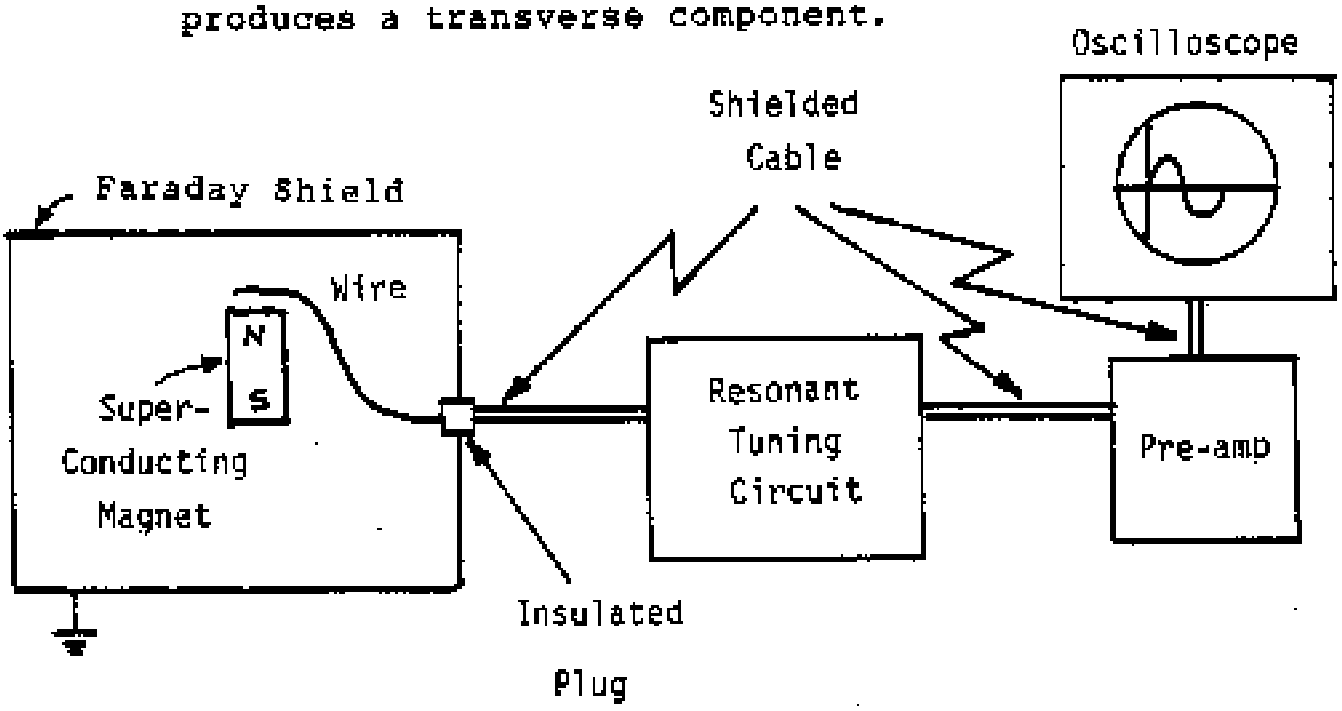
Oscilloscope
Shielded Cable
Faraday Shield
Wire
N
S
Super-Conducting Magnet
Resonant Tuning Circuit
Pre-amp
Insulated Plug

Figure 7. A sensitive scalar wave detector.

Figure 7. A sensitive scalar wave detector.

(DETECTS ELECTROGRAVITATIONAL WAVES)

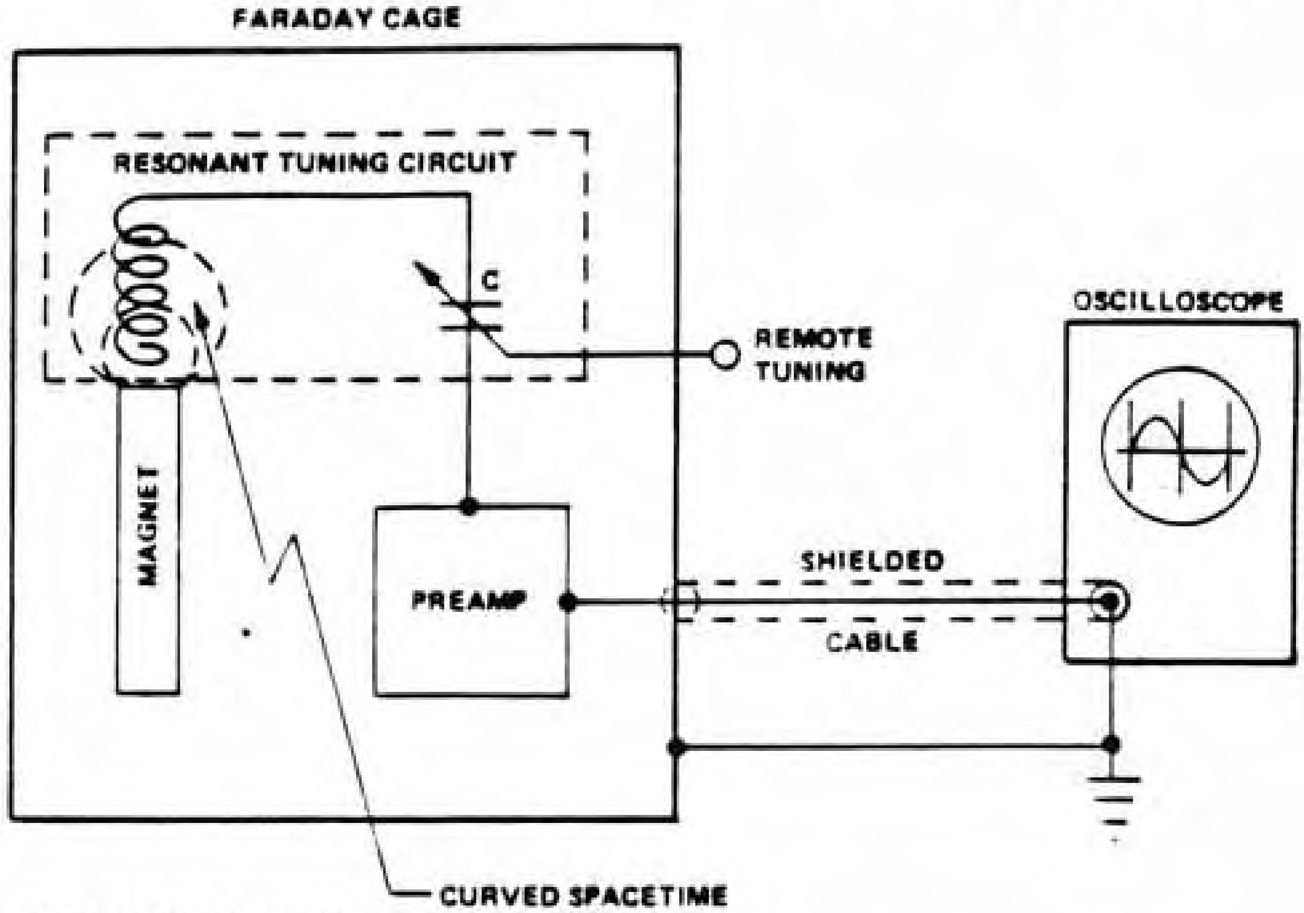

NOTE: MAGNET > 40,000 gauss

*BEDINI VERSION OF DEA/FARETTO DETECTOR

Slide 33 - FER DE LANCE

Then to prevent detection of superfluous "normal" radiation, we shield the entire magnet in a grounded Faraday cage, as shown in the figure. Hertzian waves will be grounded in this shield, while scalar waves will readily penetrate it undiminished. In other words, the Faraday cage serves as a "stripper," to strip away the ordinary waves, leaving only the scalar waves to penetrate inside. An ordinary wire lies at the top of the magnetic pole, in proximity to it. The wire runs out of the Faraday cage through an insulated port to a resonant tuning circuit, which is sharply tunable over the range of frequencies we are interested in. A preamp amplifies the output of the tuner, and in turn feeds the input of an oscilloscope or other detector.

With this detector we can select the frequency desired, and detect any passing scalar waves of that frequency. By other variations of amps and preamps, regenerative circuits, etc. we can obtain all the sensitivity desired, and utilize ordinary detection equipment already well-known and highly developed.

Additional Information about detecting Scalar Waves

[The Tom Bearden Website](#)

Figure 11. NETWORK OF VIRTUAL TRANSMITTERS

(See Woodpecker Grid)

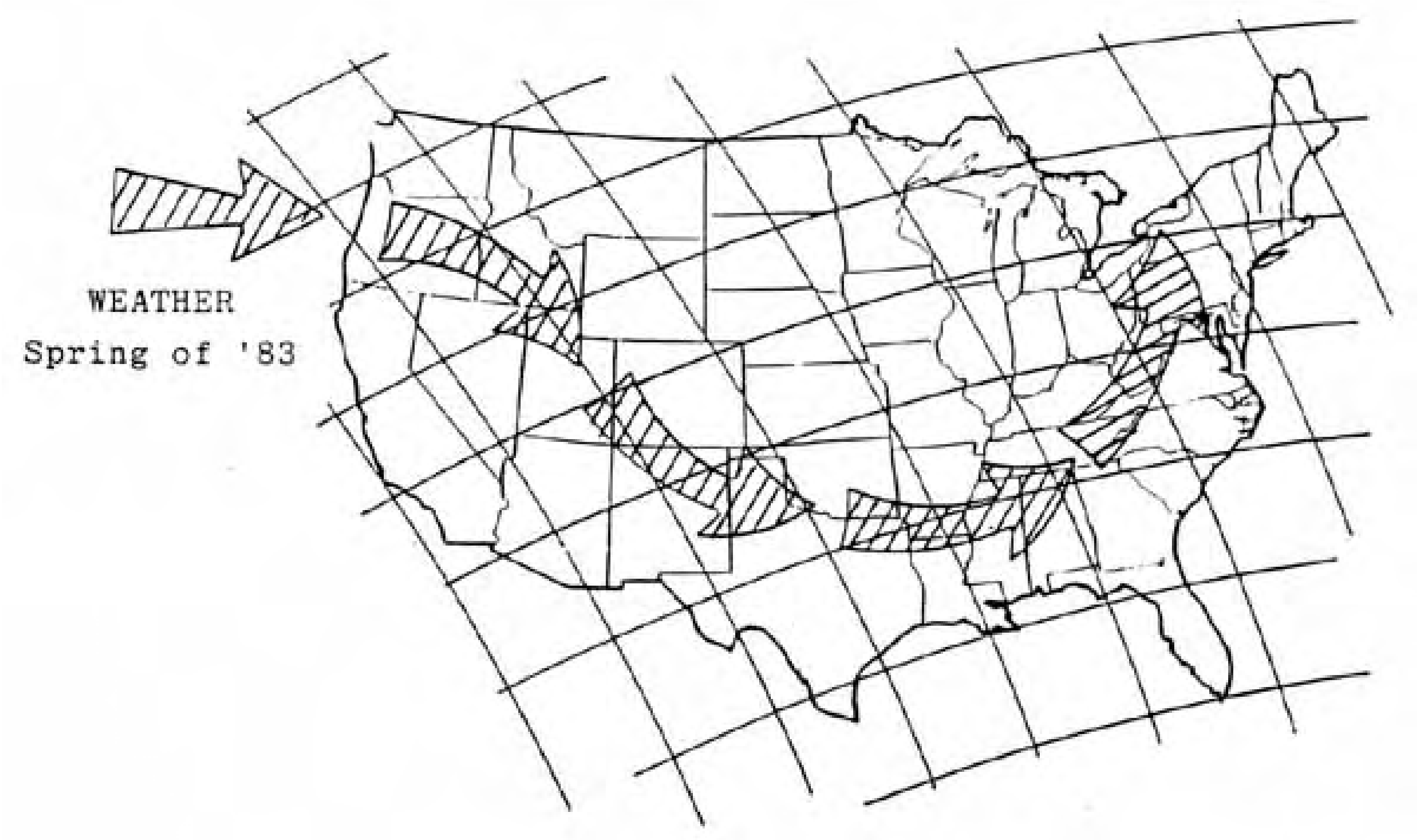

Next Figure

Figure 12. VIRTUAL TRANSMITTERS IN THE INTERFERENCE GRID

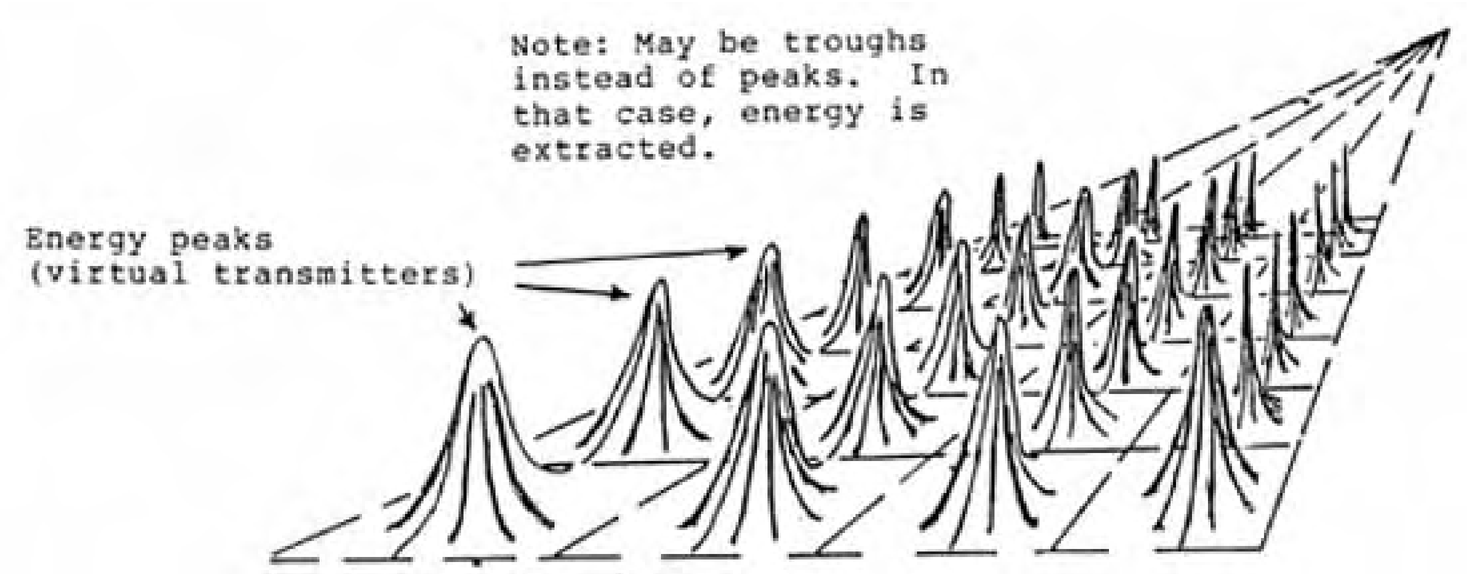
Note: May be troughs instead of peaks. In that case, energy is extracted.
Energy peaks (virtual transmitters)

Next Figure

Previous Figure

[The Tom Bearden Website](#)

Figure 13. GIANT RADIAL RELATED TO A VIRTUAL TRANSMITTER

Fall 1983:
Florida
Alabama

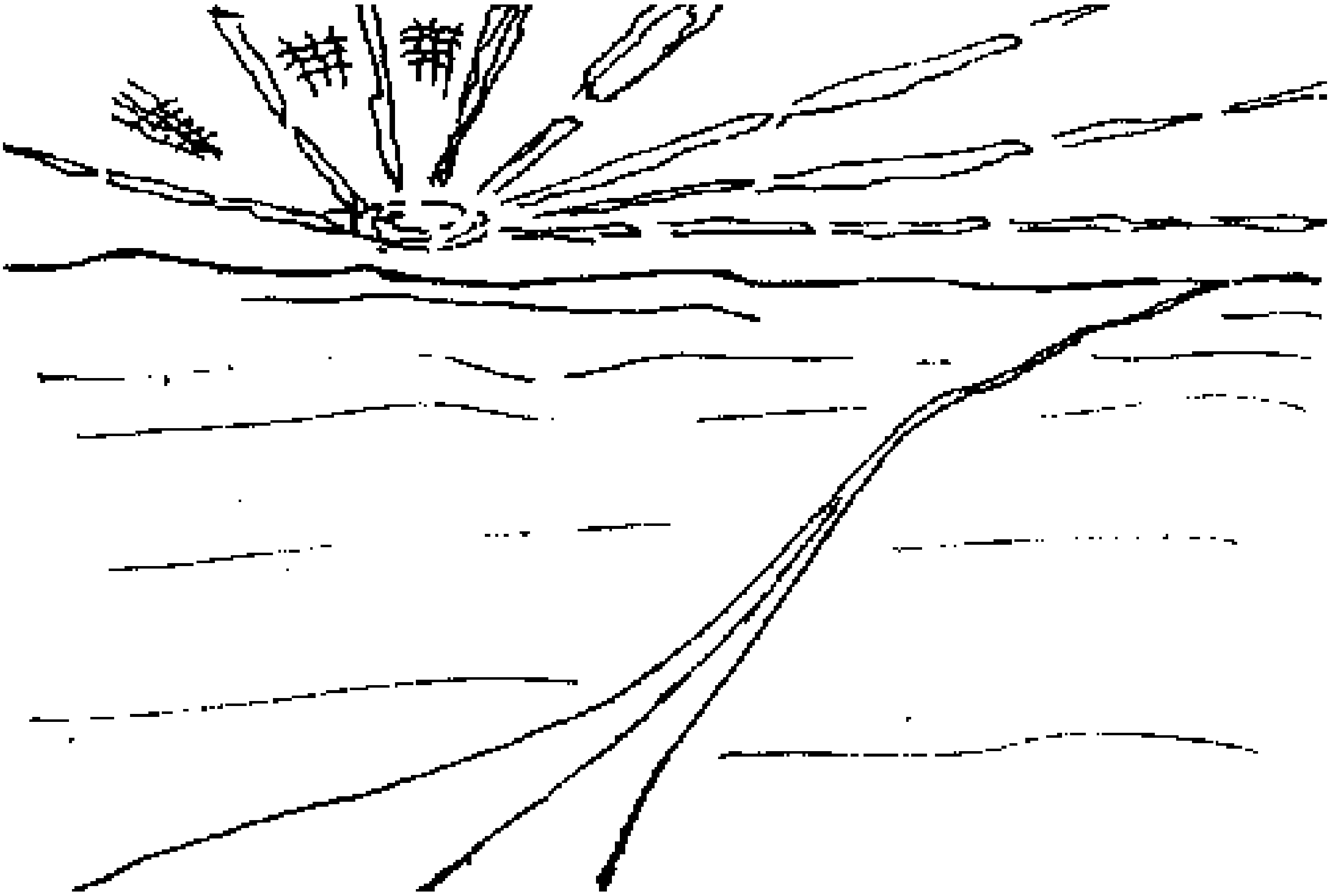

Next Figure

Previous Figure

The Tom Bearden Website

Figure 14. FORMATION OF A SINGLE GIANT RADIAL

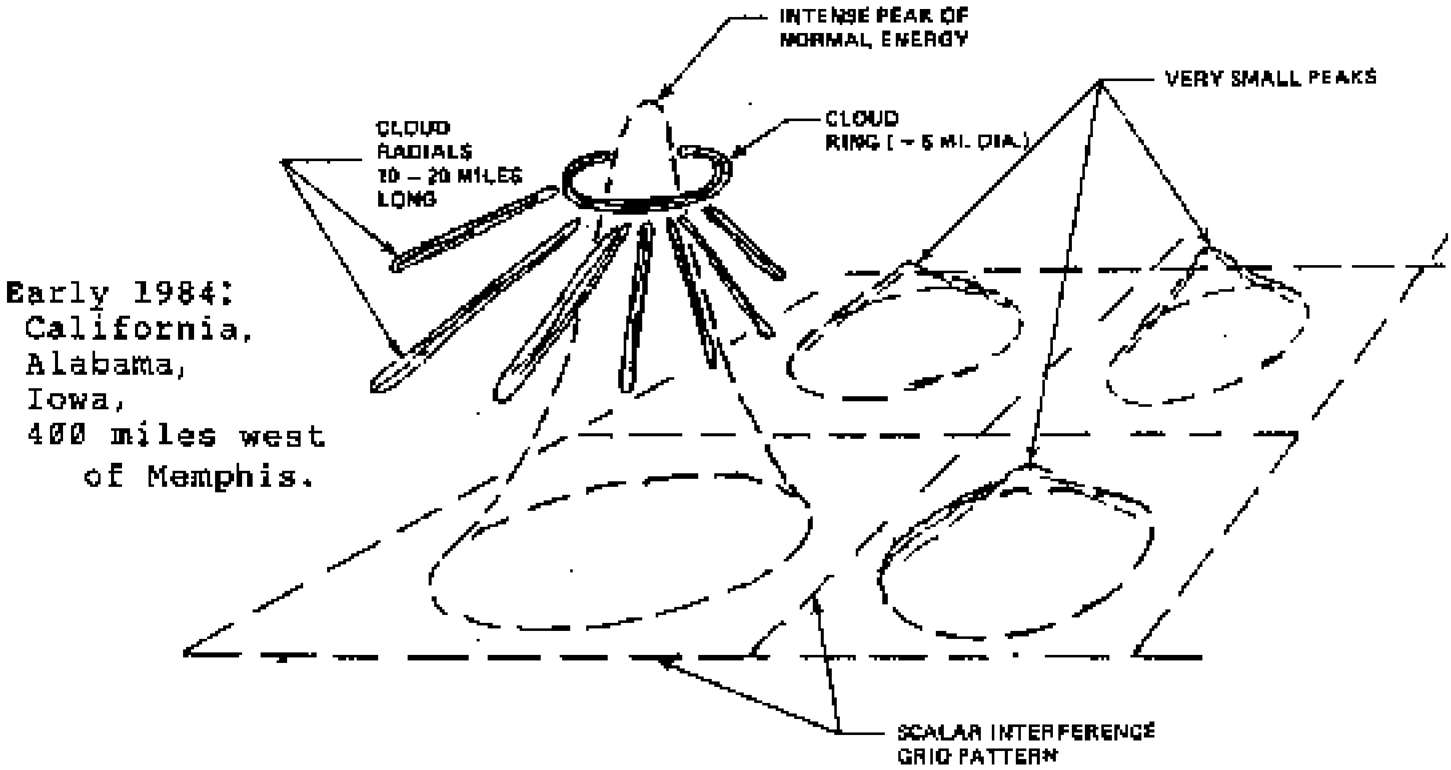
INTENSE PEAK OF NORMAL ENERGY
VERY SMALL PEAKS
CLOUD RADIALS 10 – 20 MILES LONG
CLOUD RING (~ 5 MI. DIA.)
Early 1984: California, Alabama, Iowa, 400 miles west of Memphis.
SCALAR INTERFERENCE GRID PATTERN

Excerpt P 172
FER DE LANCE

A Briefing on Soviet Scalar Electromagnetic Weapons
by Lt. Col. T.E. Bearden (retd.), 1986
Copyright

2. Continual cloud radials have repeatedly been seen and photographed over Huntsville, Alabama, associated with weather engineering and grid activity. Typical radials seen by this author are shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. v1deotape, 1985. Huntsville is apparently a pivot point or "hinge point" for bending and controlling the jetstream.

3. Spectacular cloud grid patterns have been seen over Huntsville, Alabama and Los Angeles, California. An excellent videoclip of one of the Los Angeles area grids, personally observed and shot by KABC Open Mind talk show host Bill Jenkins, is shown in Bearden, Soviet Weather Engineering Over North America,;1-hr. videotape, 1985. An artist's sketch of an earlier gigantic grid, extending from horizon to horizon in every direction, seen over Huntsville, Alabama by Tom Bearden and Ken Moore, is also shown in the tape. The videotape is available from P.O. Box 1472, Huntsville, AL 35807. Sketches and details of another highly anomalous grid pattern observed in the greater Los Angeles area by engineer Ron Cole are shown in figures 5 and 6.

The reason that clouds of water droplets and/or ice particles detect the scalar interferometry and form signature patterns is simple: Consider each H2O molecule as having two light little hydrogen atoms hanging on to the much heavier oxygen atom. The covalent sharing with the oxygen atom of the electron from each hydrogen atom means that

1. the electrostatic scalar potential between the H ion and the O ion it is bonded to, is rhythmically varying as the electron is shared back and forth,
2. this varying potential contains "electron spin holes" since it is made by the moving electron,
3. two such varying potentials exist since there are two H atoms sharing covalent electrons with the O atom,
4. the two H ions are at an angle of over 100 degrees with respect to each other,
5. the molecule assembly thus constitutes one part (one half, so to speak) of a scalar interferometer with imbedded electron spin holes for electron hooking,
6. Incident scalar waves from outside the system interact with the "half scalar interferometer." This scalar interferometry interaction is coupled to the covalent bonding electron because of the internal spin hole pattern of the molecular half of the interferometer. The coupled electron moves with the sca1ar pattern's changes, causing an observable interaction with and in the electrical structure and ionic potential of the molecule,
7. the weak H to H hydrogen bonding between molecules connects resulting ionic potential changes to the structuring of the entire macroscopic cloud assembly. Thus by scalar interferometry with the H_2O "half interferometers", an incident scalar EM pattern is detected and translated into patterns of E and B force fields. The H_2O molecules then "line up" in accord with the electrical patterns detected.

Thus the clouds form "signature patterns" according to the incident scalar EM radiation patterns. Giant cloud radials in the targeted area are probably due to the type and shape of antennas including the type and shape of the electrical wires establishing the antennas' ground planes -- used by the Soviet Union to project the scalar waves. Note that such "radial" antenna patterns were occasionally used by James Harris Rogers in his underground and undersea scalar EM transmission system.

Next Figure

Previous Figure

The Tom Bearden Website

Figure 15. FORMATION OF A DOUBLE GIANT RADIAL

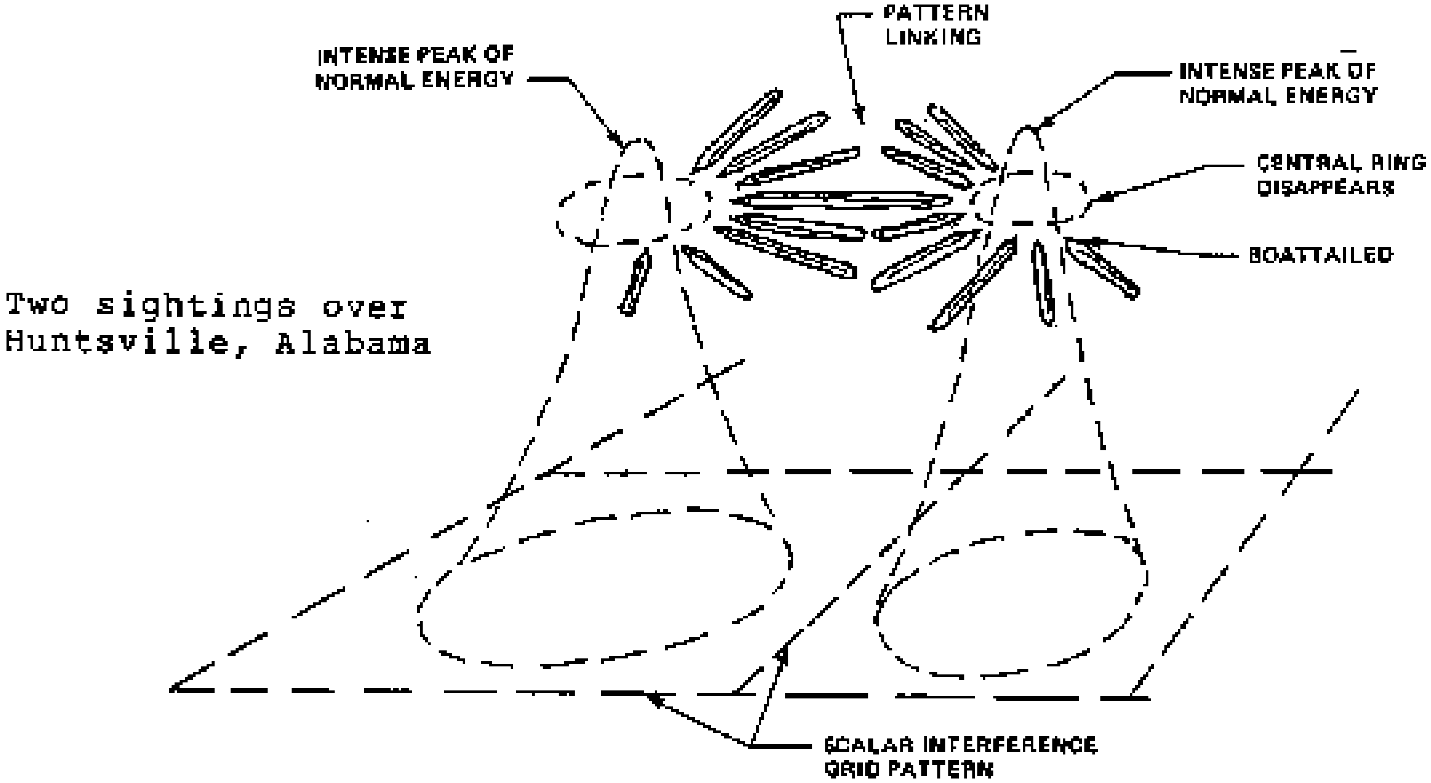
PATTERN LINKING
INTENSE PEAK OF NORMAL ENERGY
INTENSE PEAK OF NORMAL ENERGY
CENTRAL RING DISAPPEARS
BOATTAILED
SCALAR INTERFERENCE GRID PATTERN

Two sightings over Huntsville, Alabama

Next Figure

Previous Figure

The Tom Bearden Website

Figure 16. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

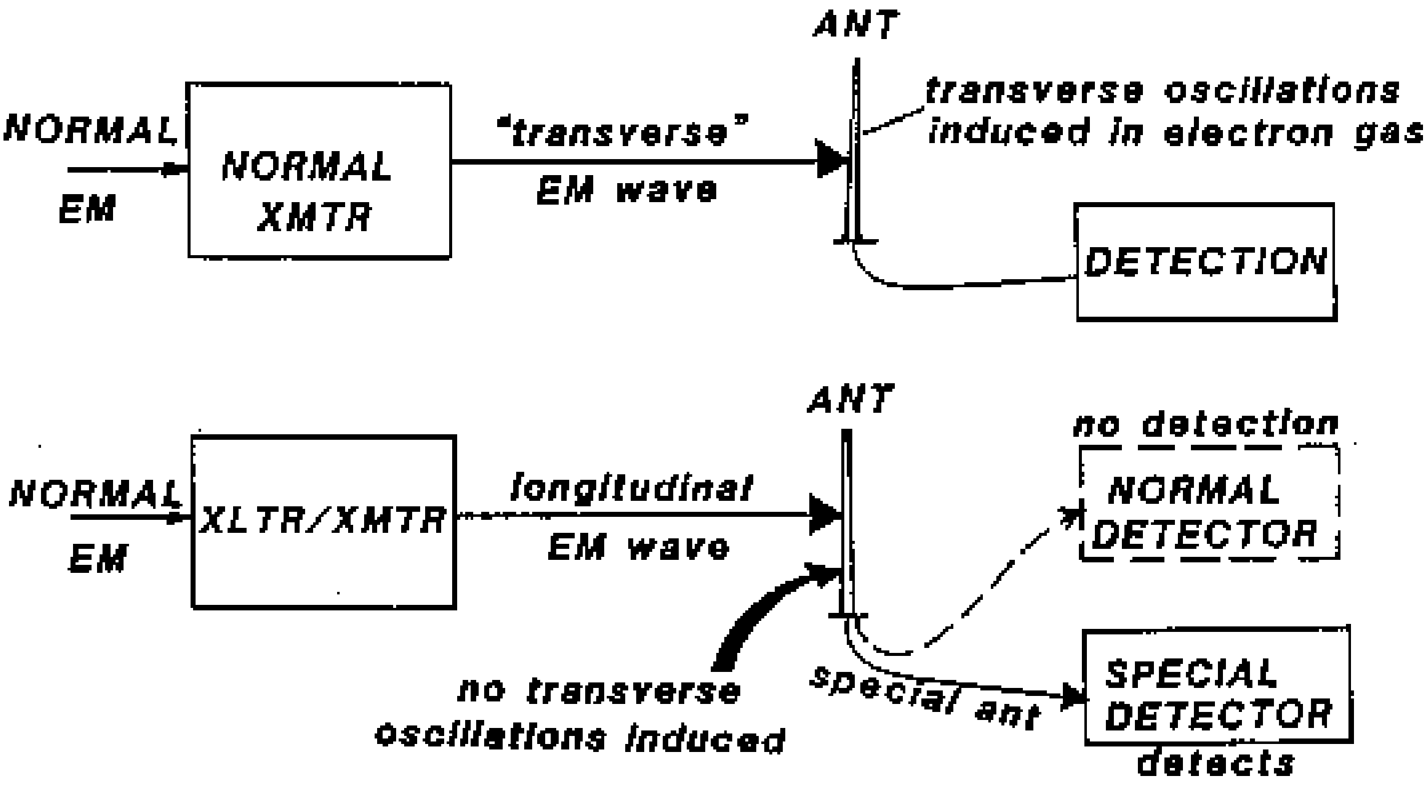
ANT
NORMAL
EM
NORMAL XMTR
"transverse" EM wave
transverse oscillations induced in electron gas
DETECTION
ANT
NORMAL
EM
XLTR/XMTR
longitudinal EM wave
no detection
NORMAL DETECTOR
no transverse oscillations induced
special ant
SPECIAL DETECTOR
detects

Next Figure

Previous Figure

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

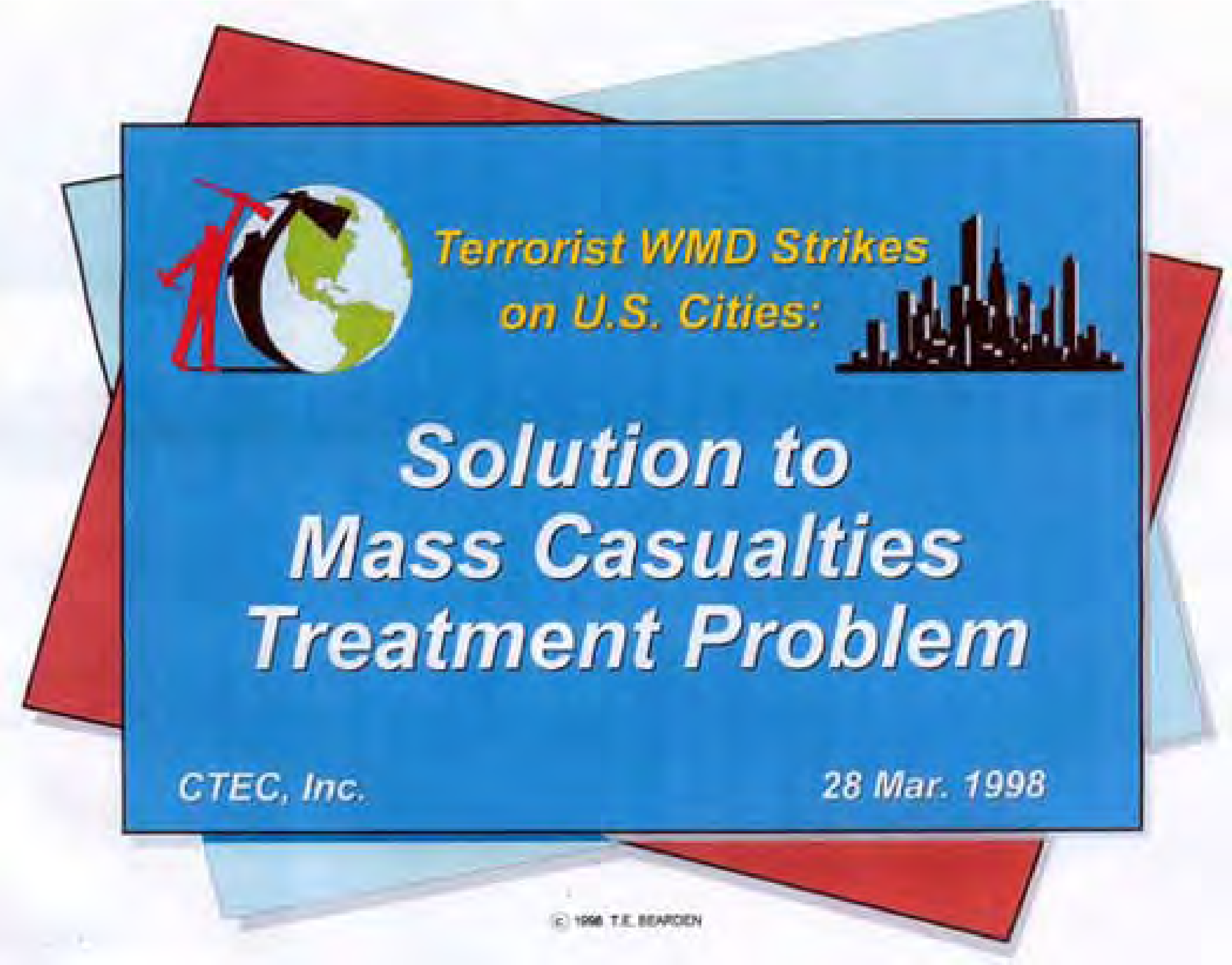

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

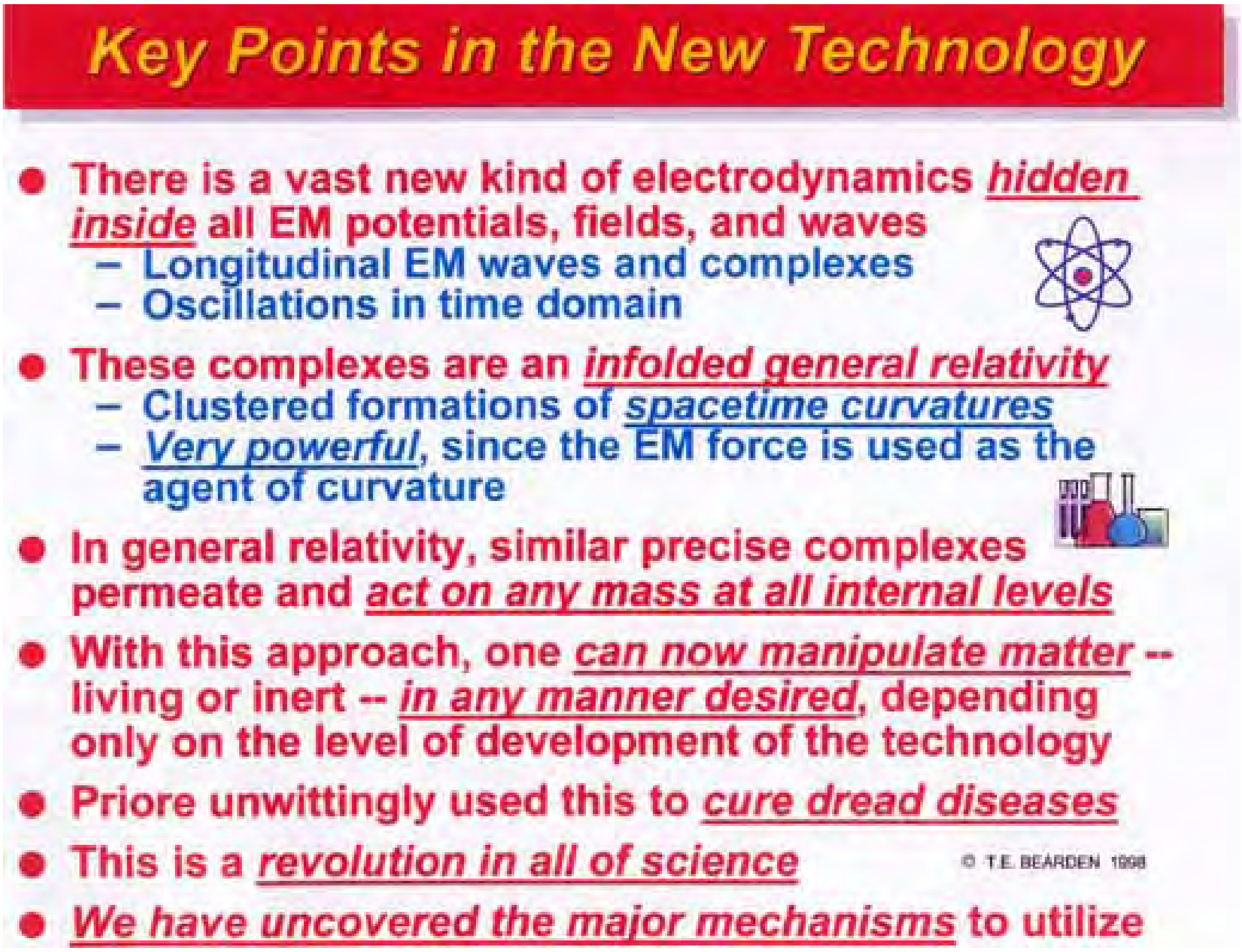

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

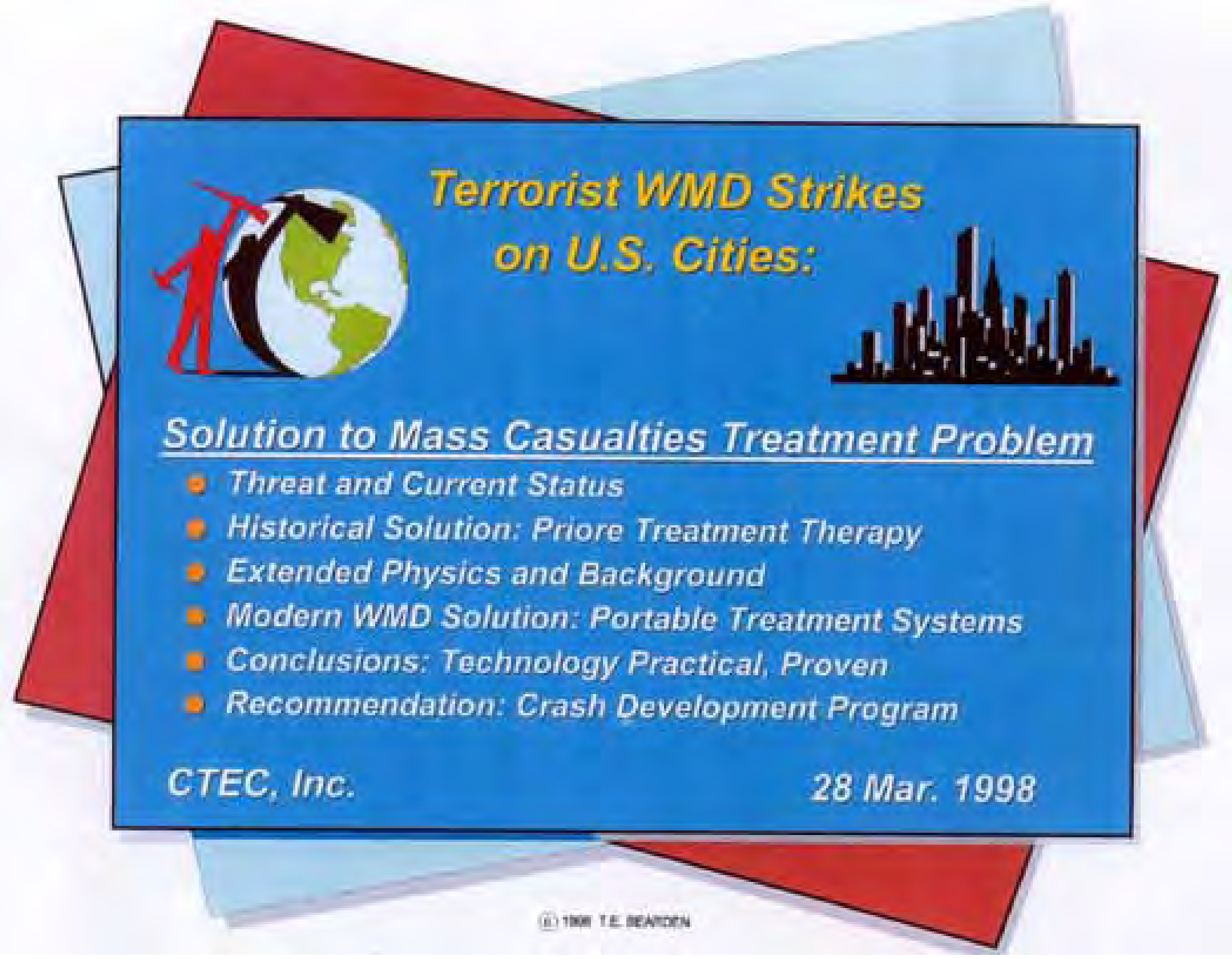

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

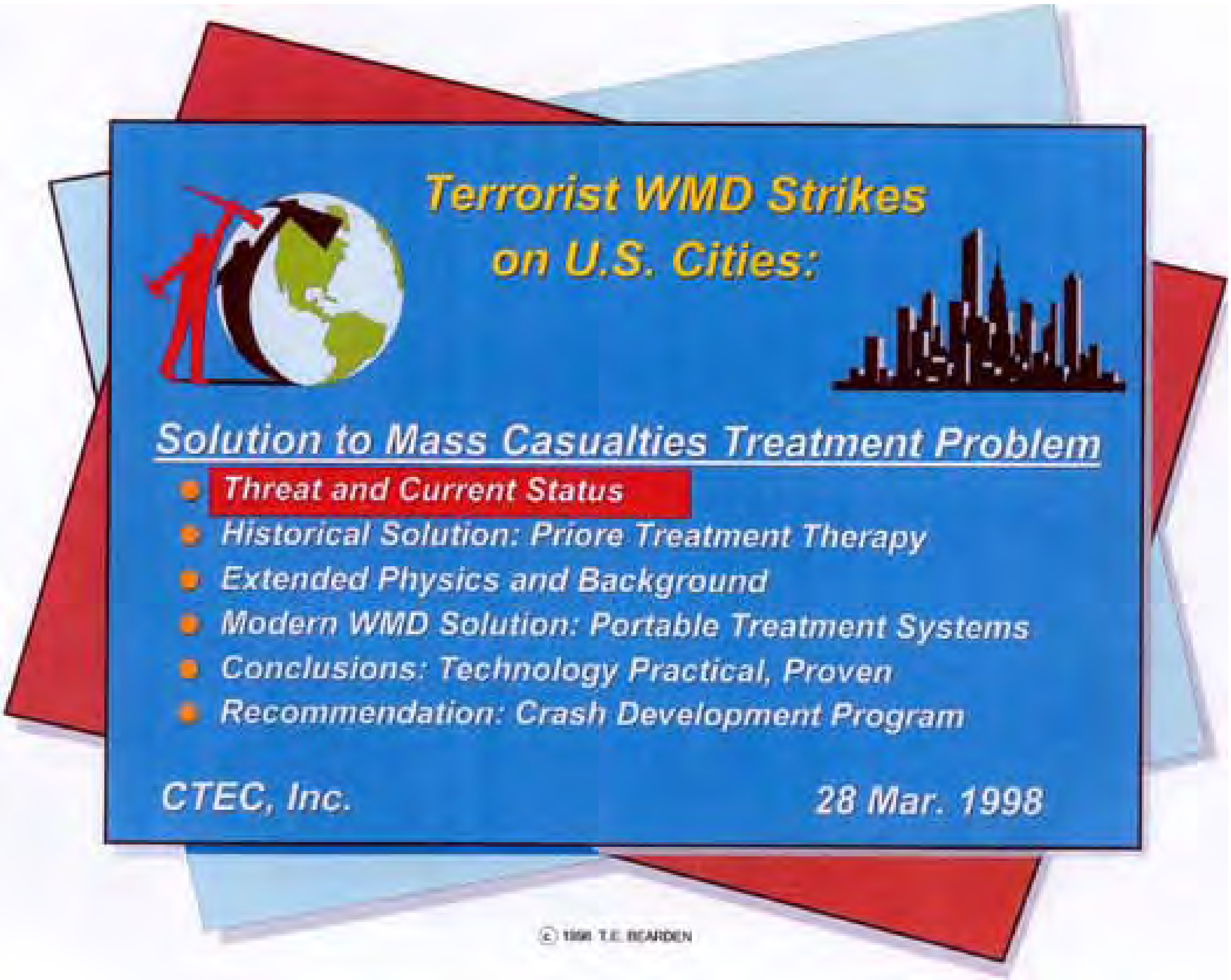

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

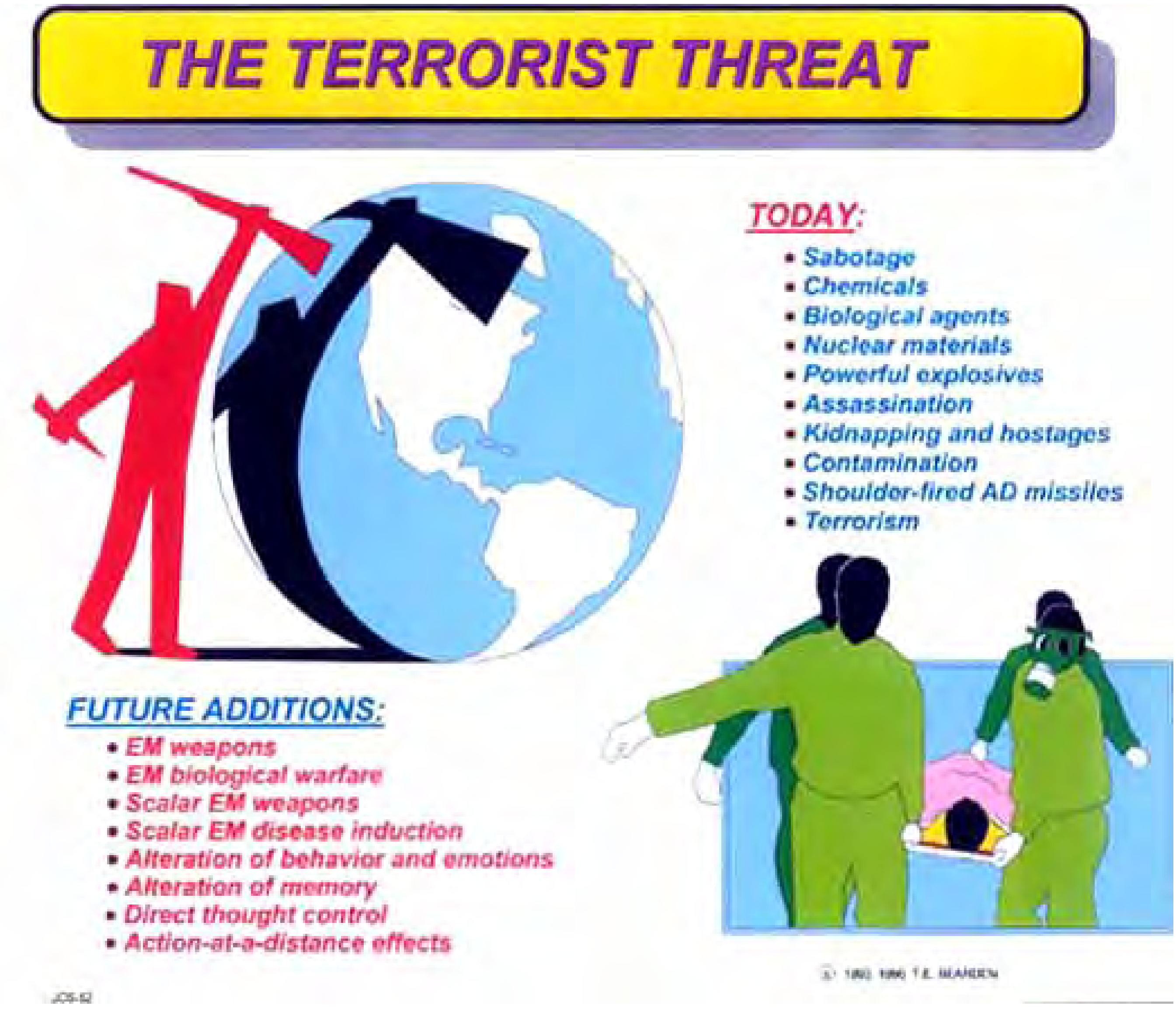

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

© 1992, 1996 T.E. BEARDEN

JCS-62

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

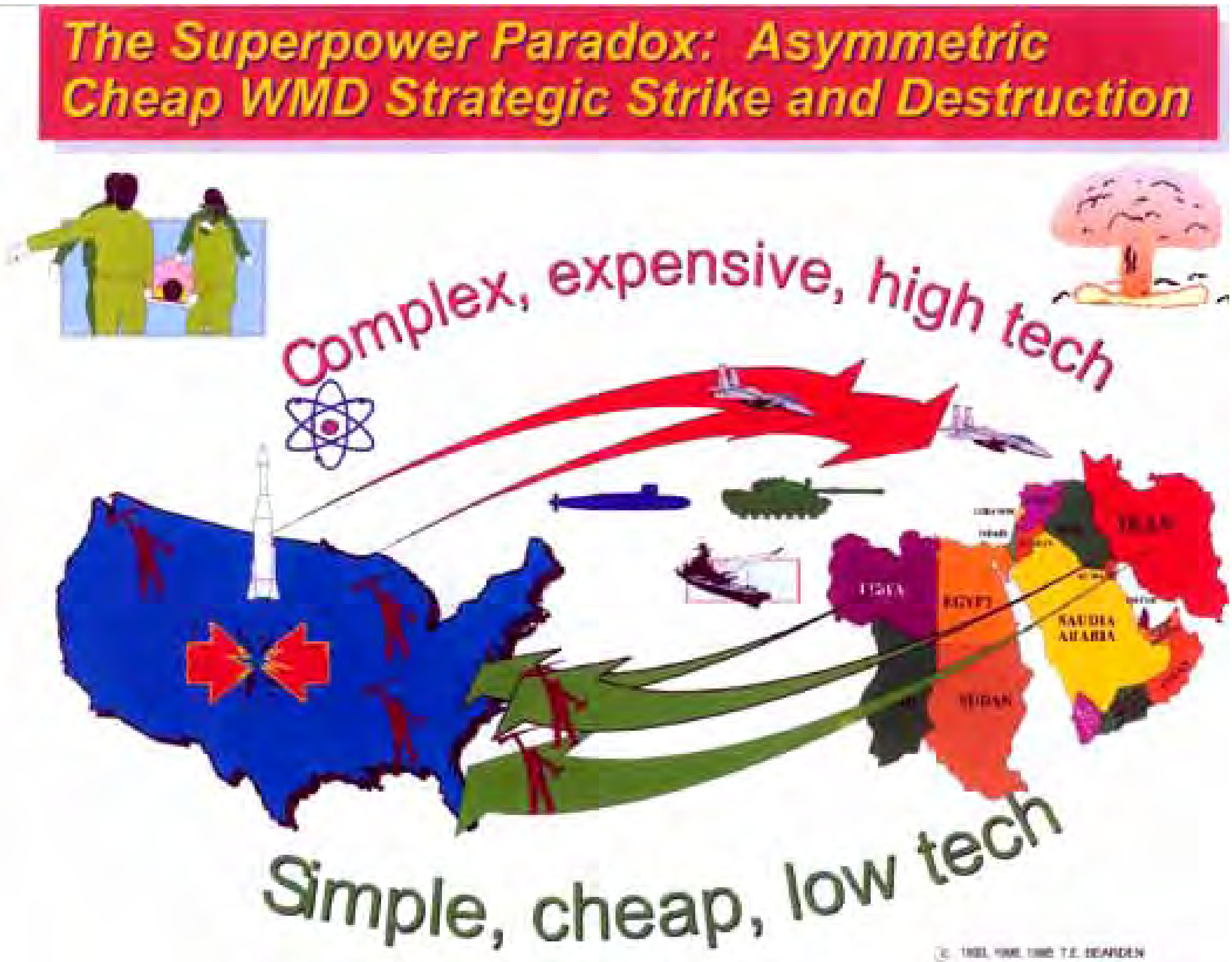

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

© 1993, 1998, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

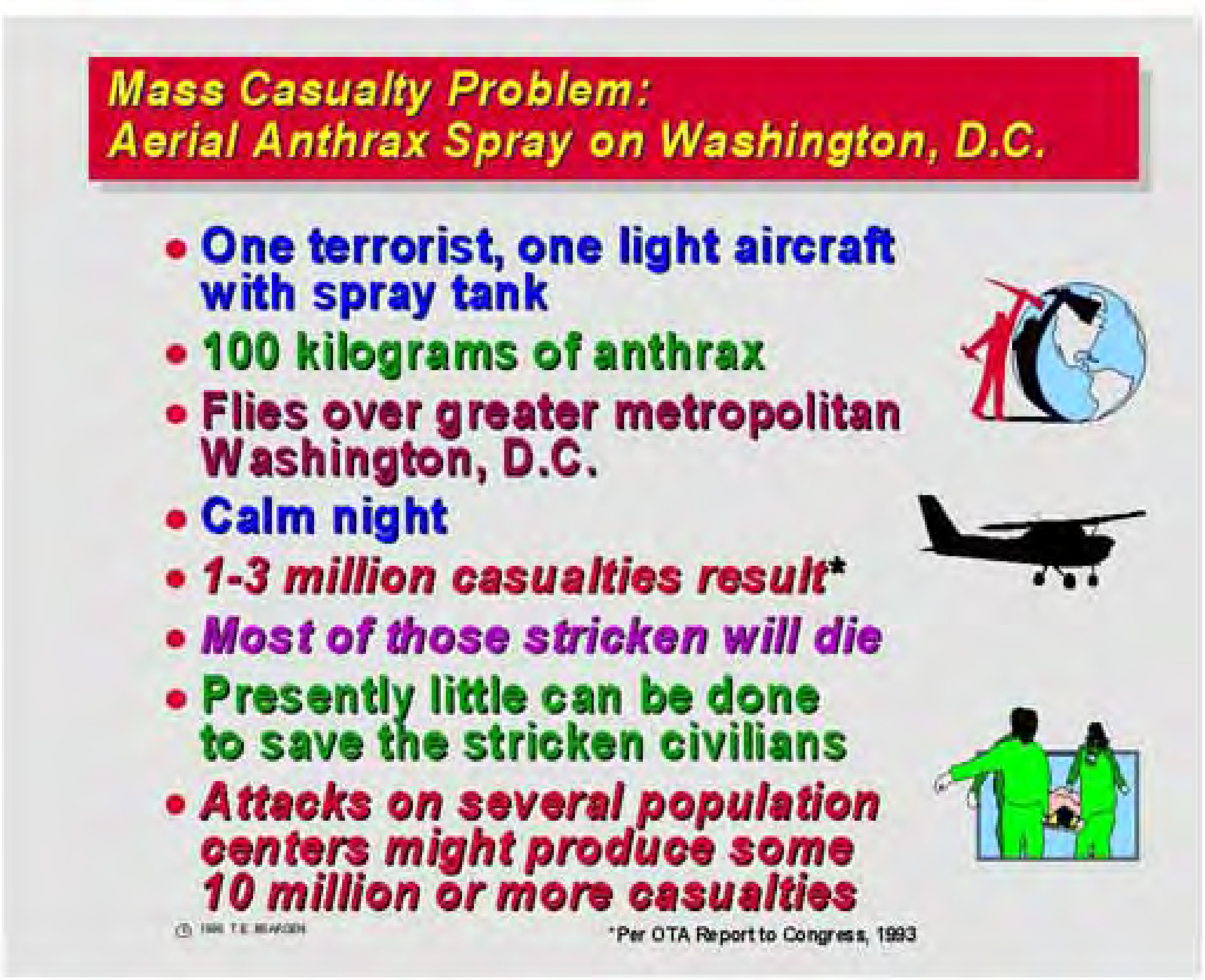

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

*Per OTA Report to Congress, 1993

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

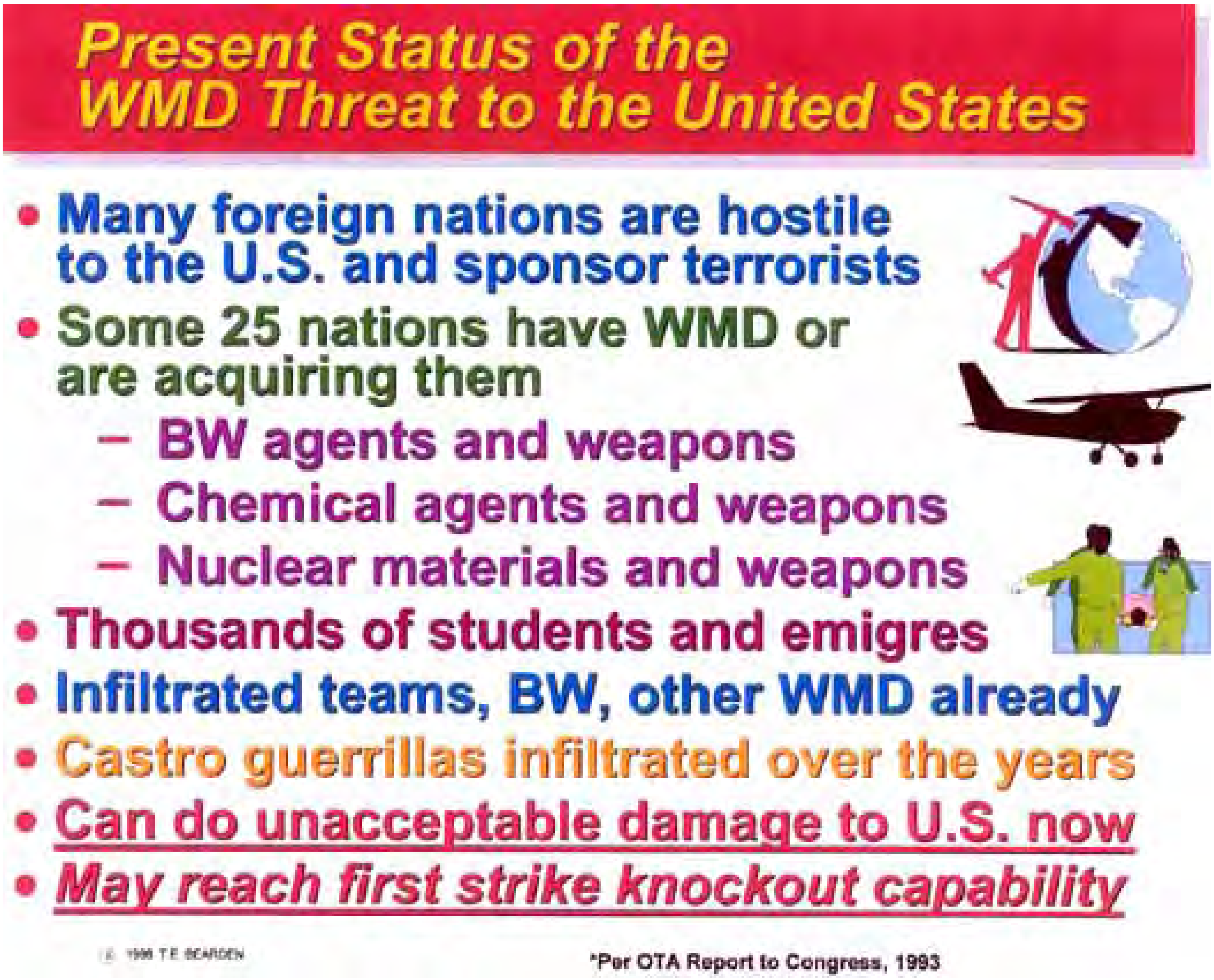

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - – BW agents and weapons
  - – Chemical agents and weapons
  - – Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

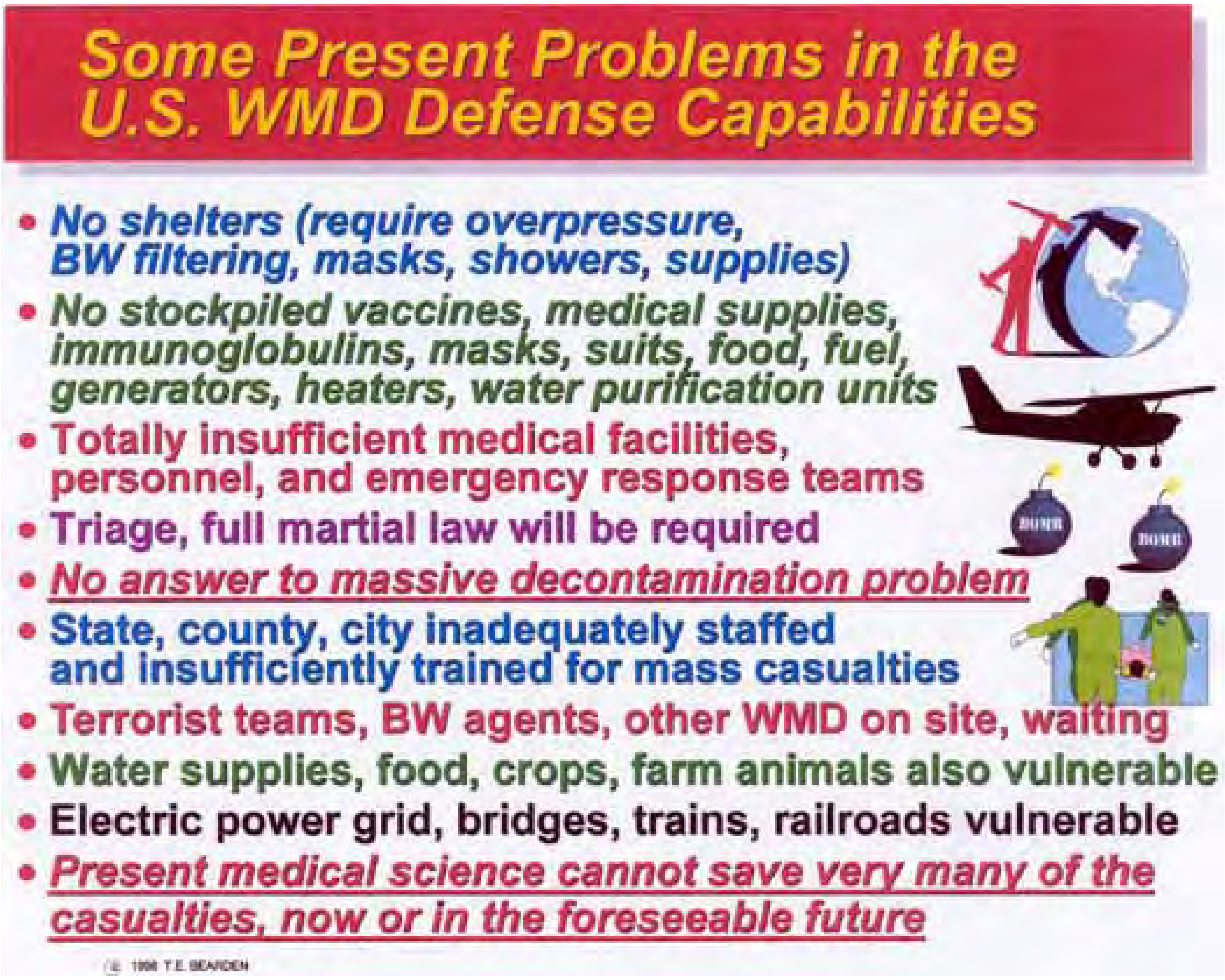

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

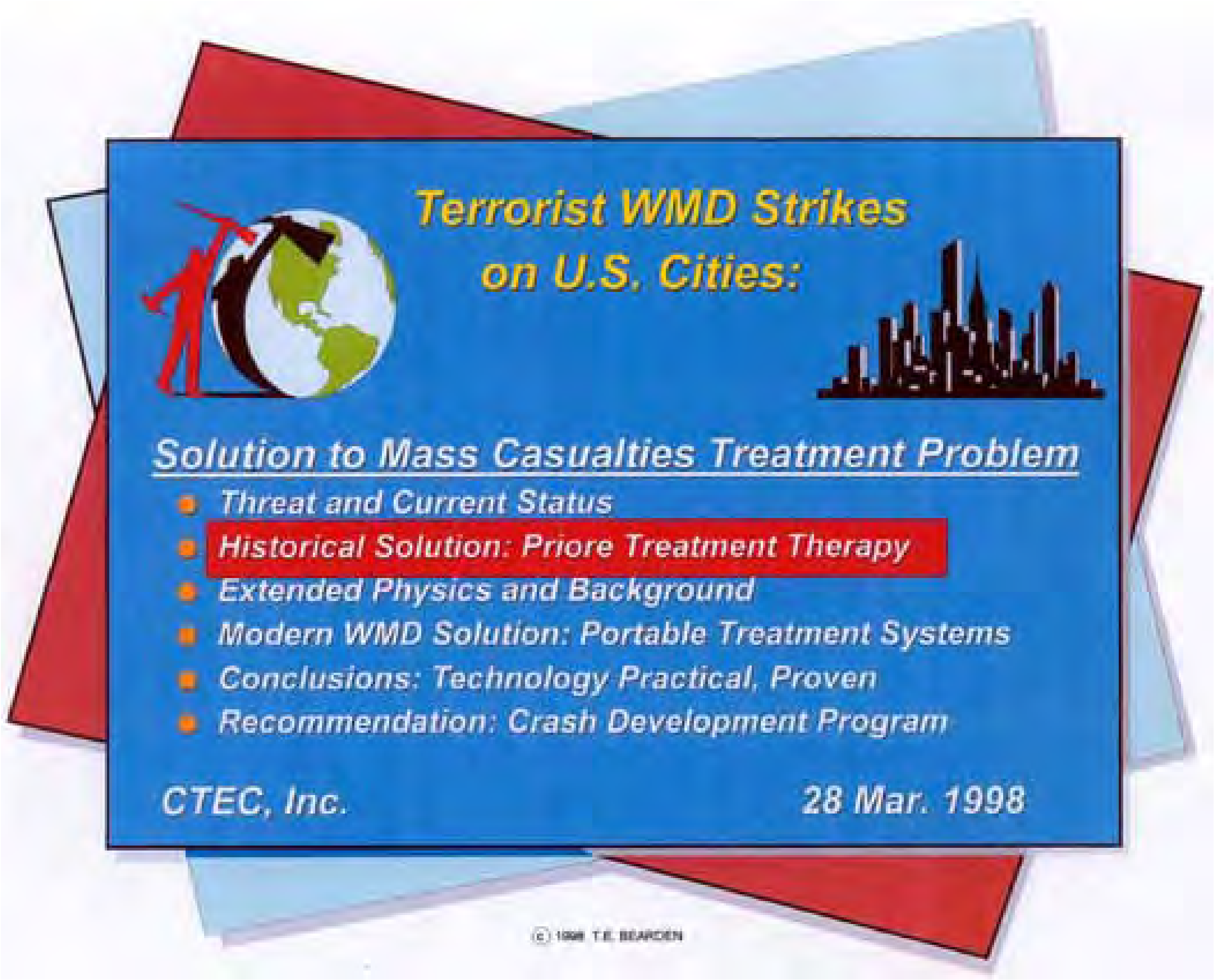

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

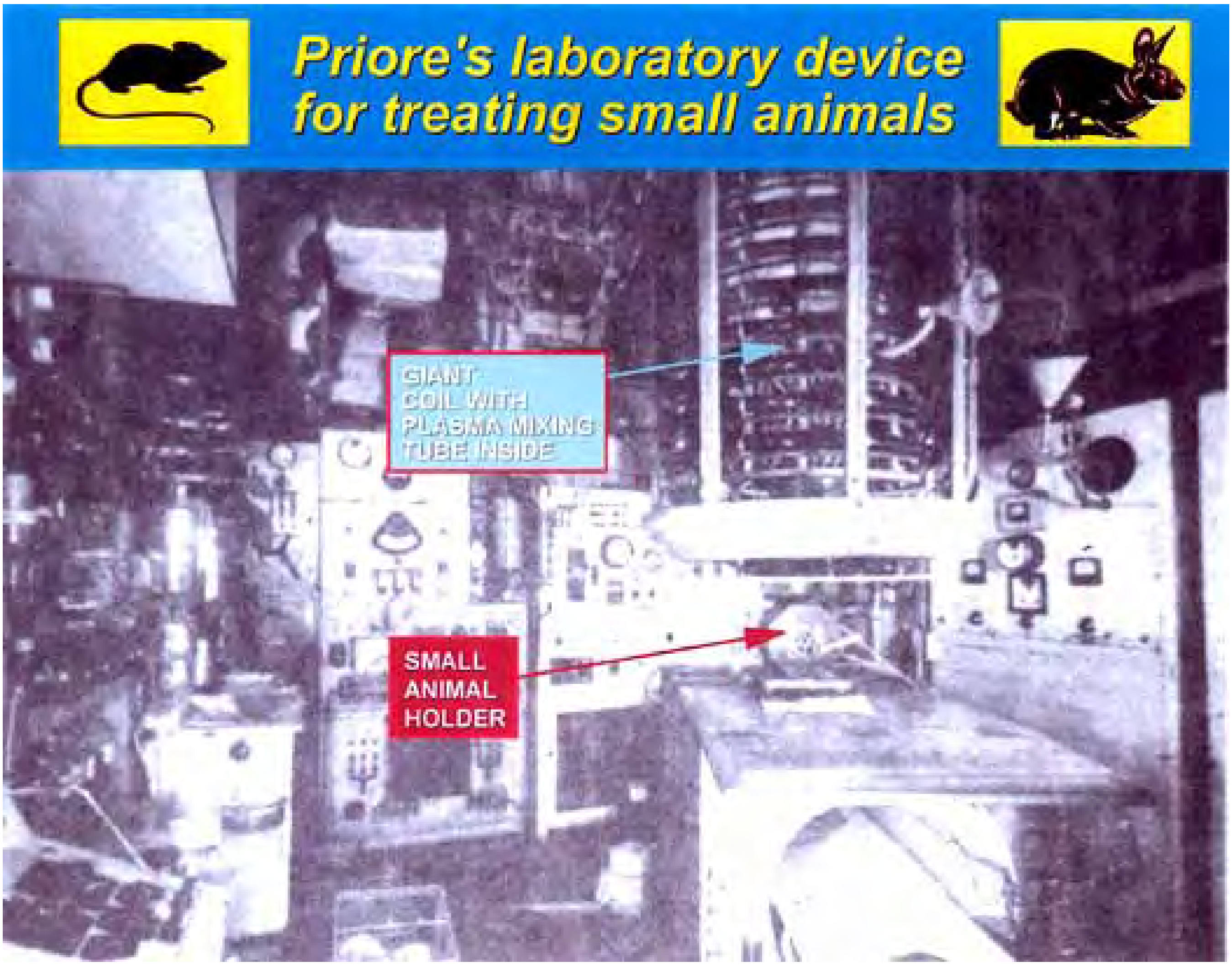

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE
SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

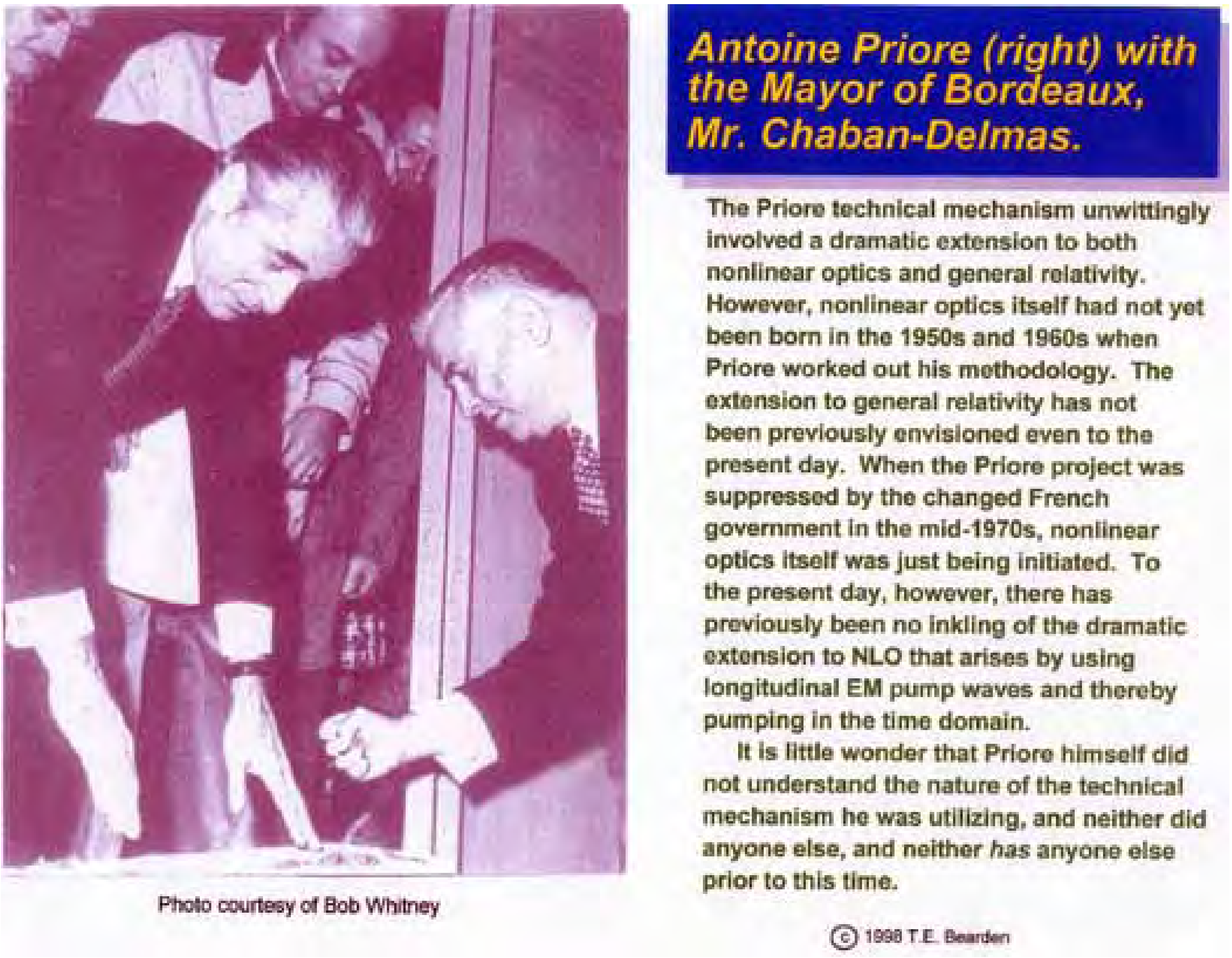

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

Photo courtesy of Bob Whitney

© 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

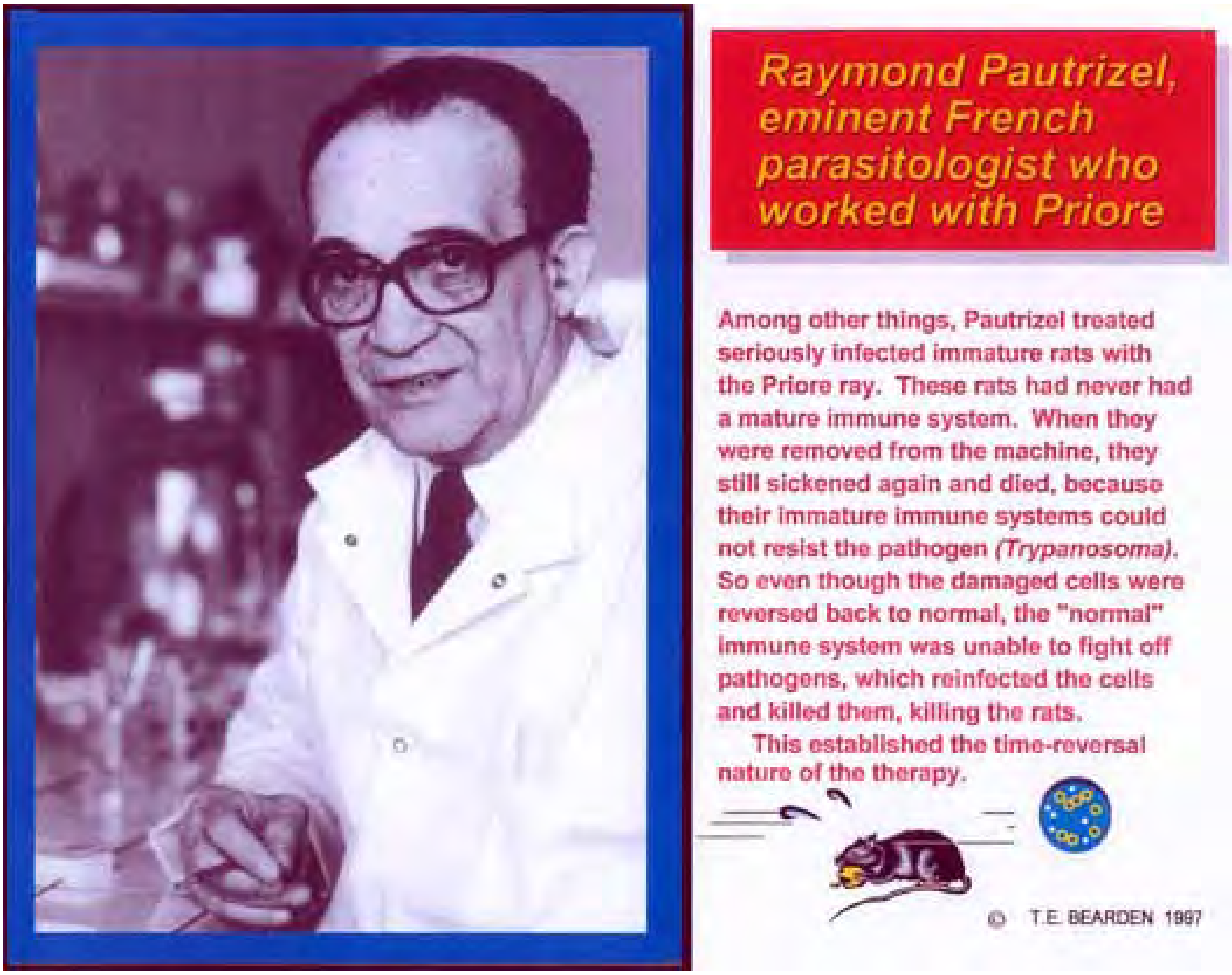

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

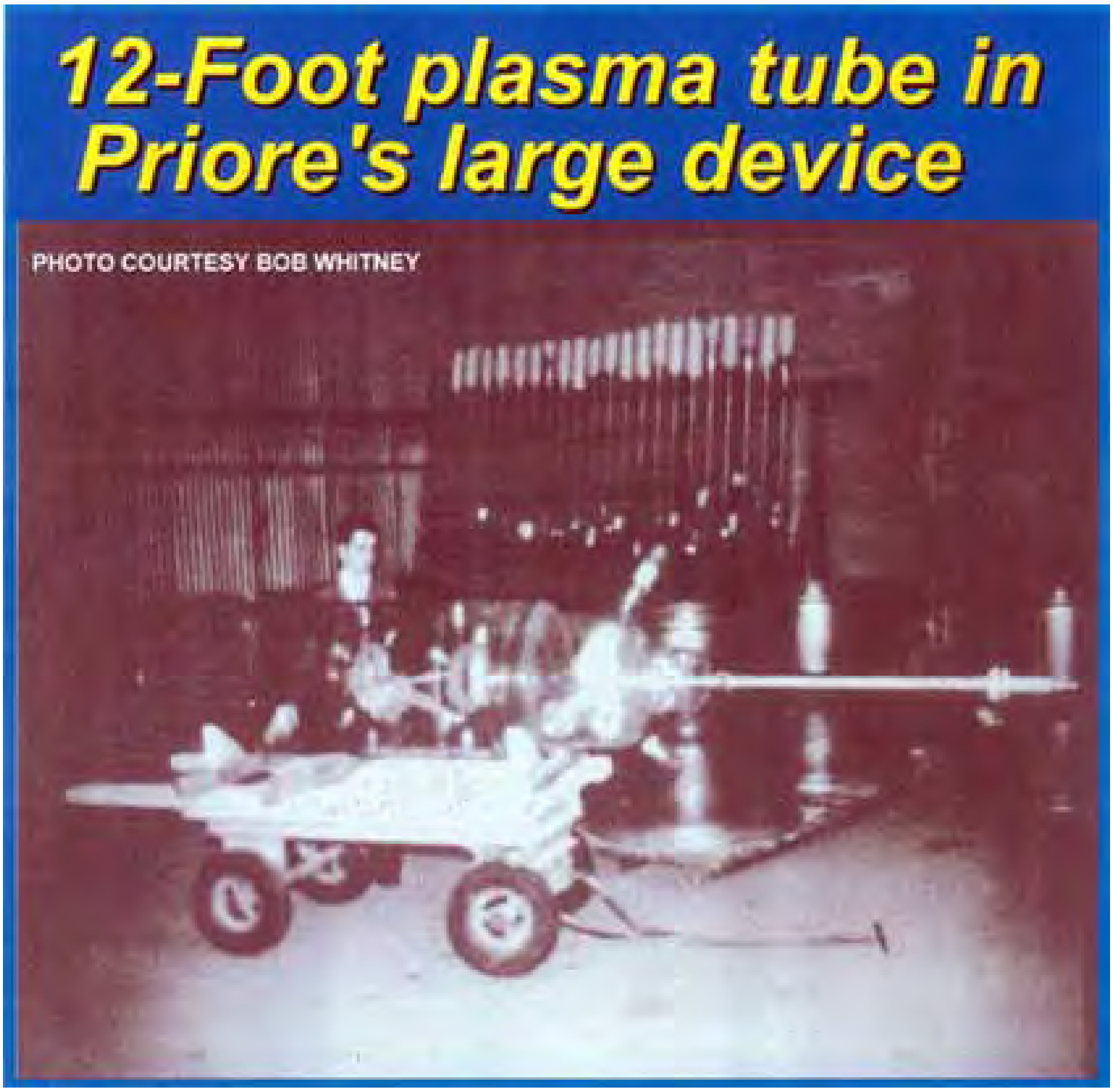

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

Return to Slide Index

Next Slide

Previous Slide

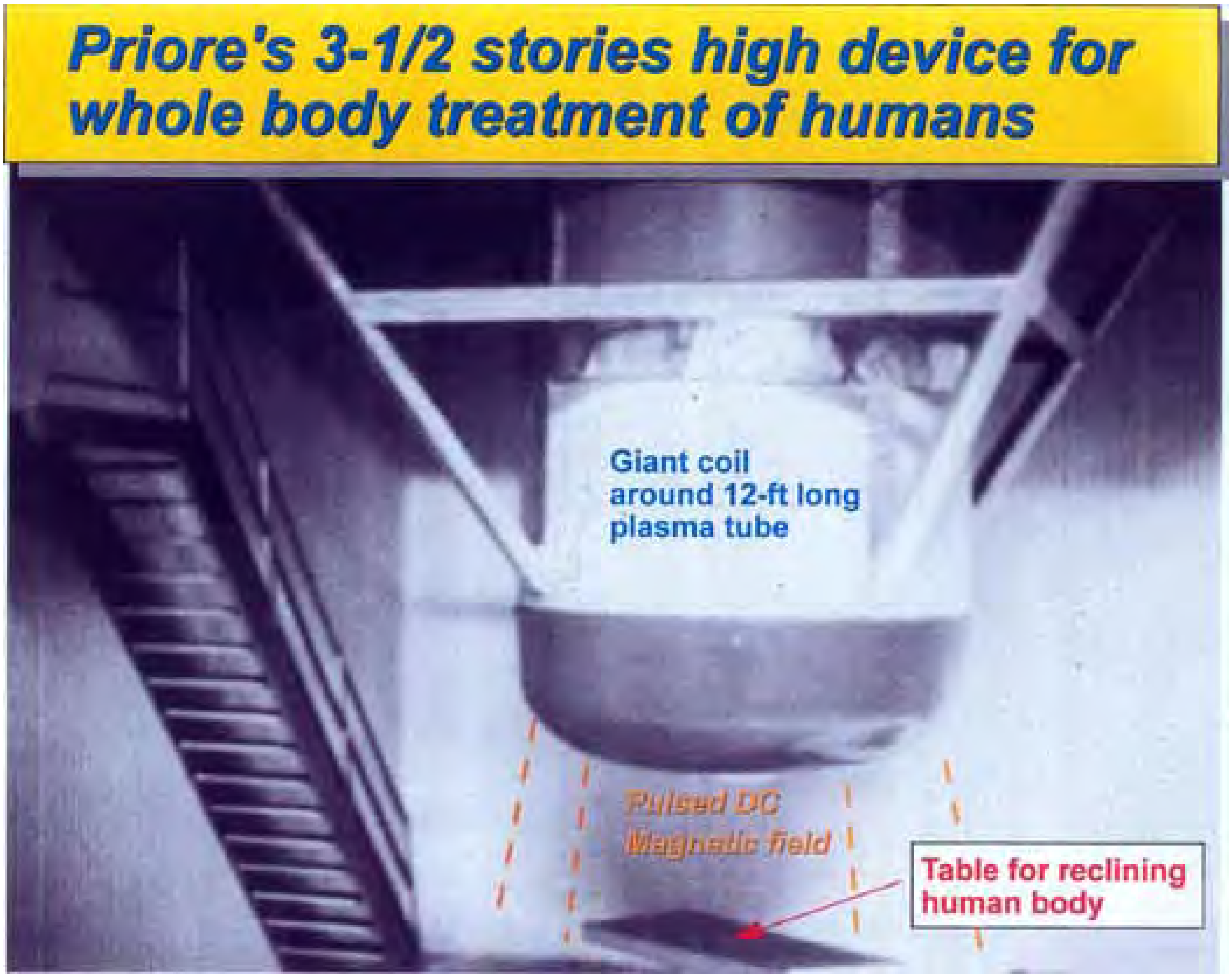

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube
Pulsed DC
Magnetic field
Table for reclining
human body

Return to Slide Index

Next Slide

Previous Slide

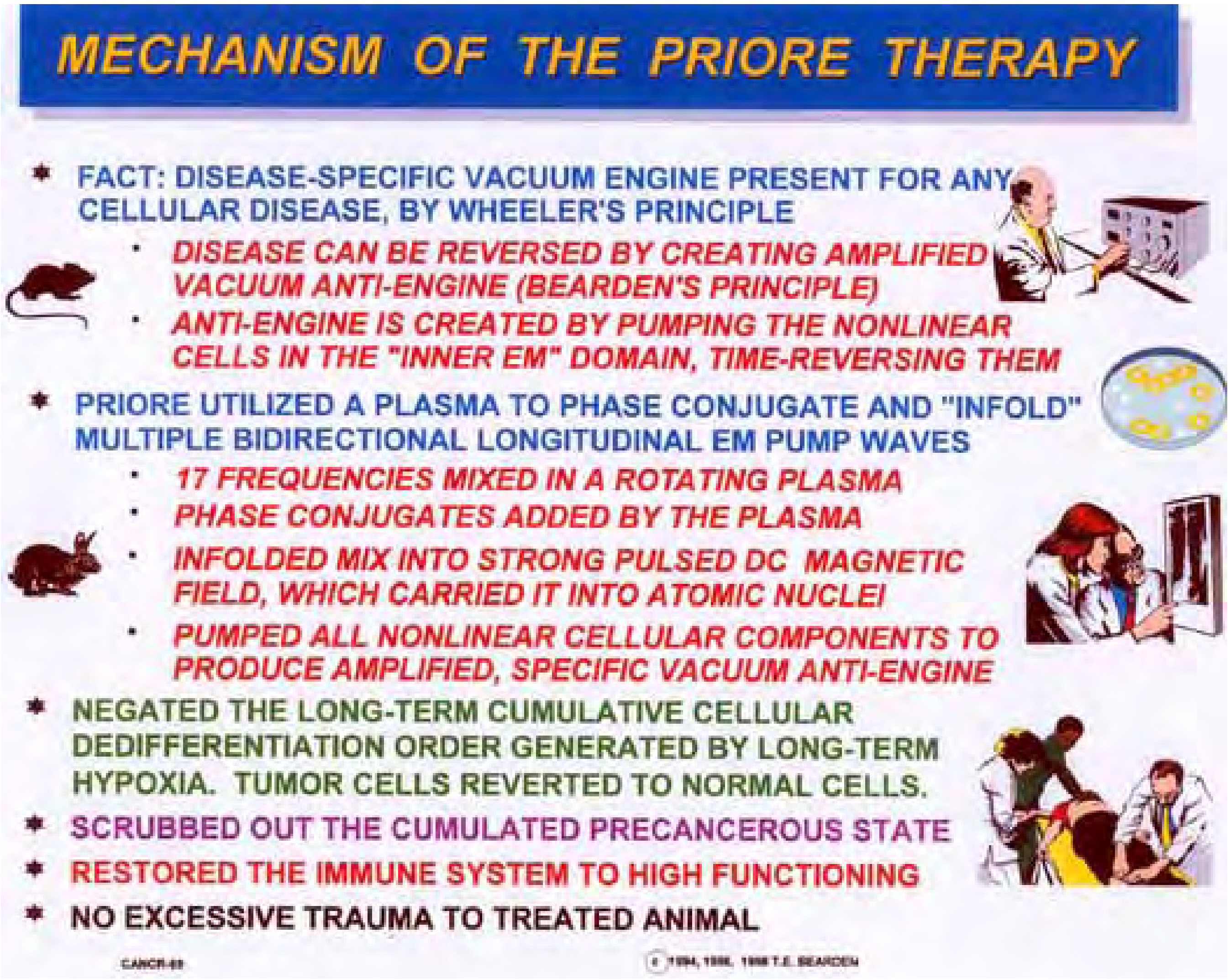

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

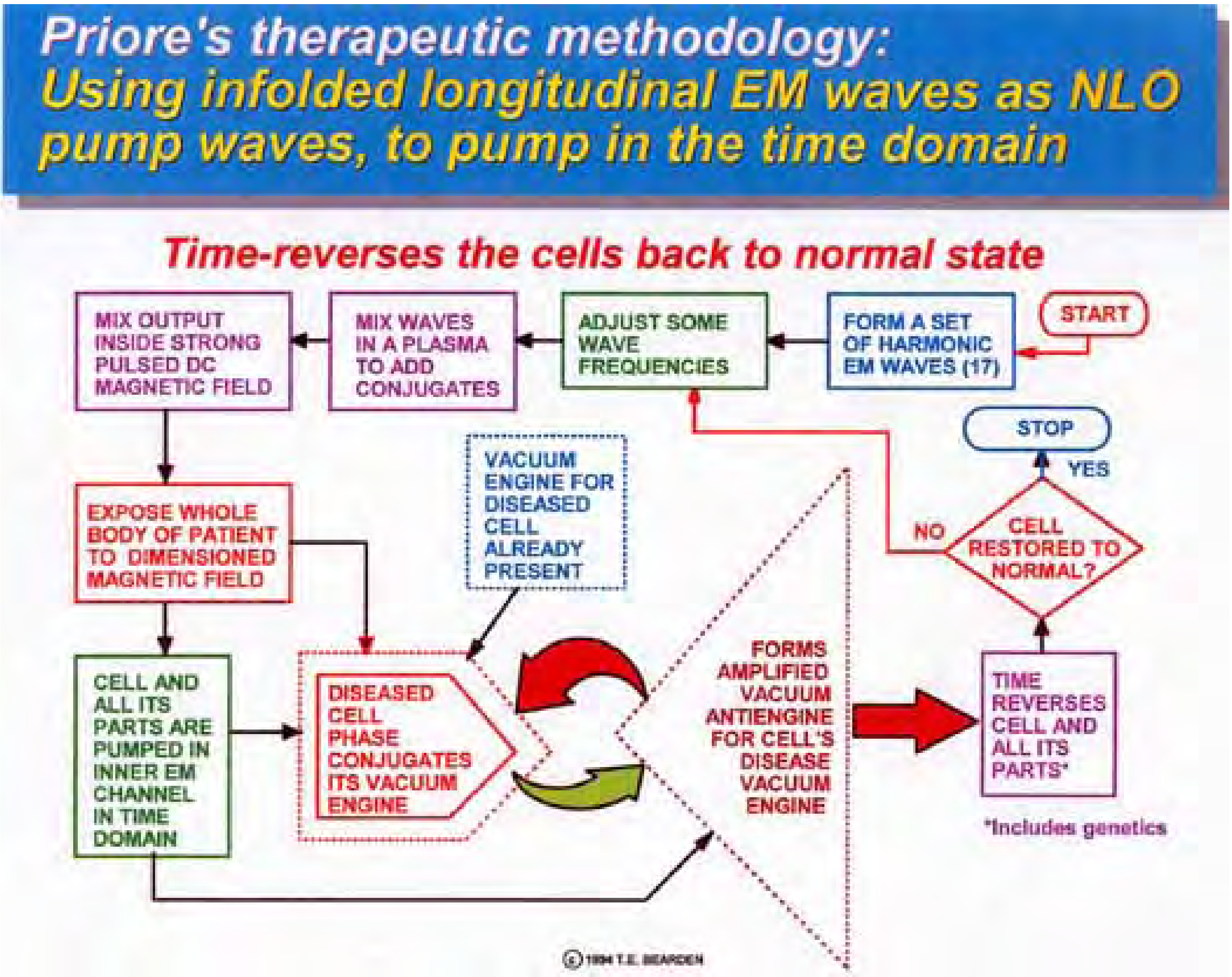

Priore's therapeutic methodology: Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

NO

YES

STOP

*Includes genetics

©1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

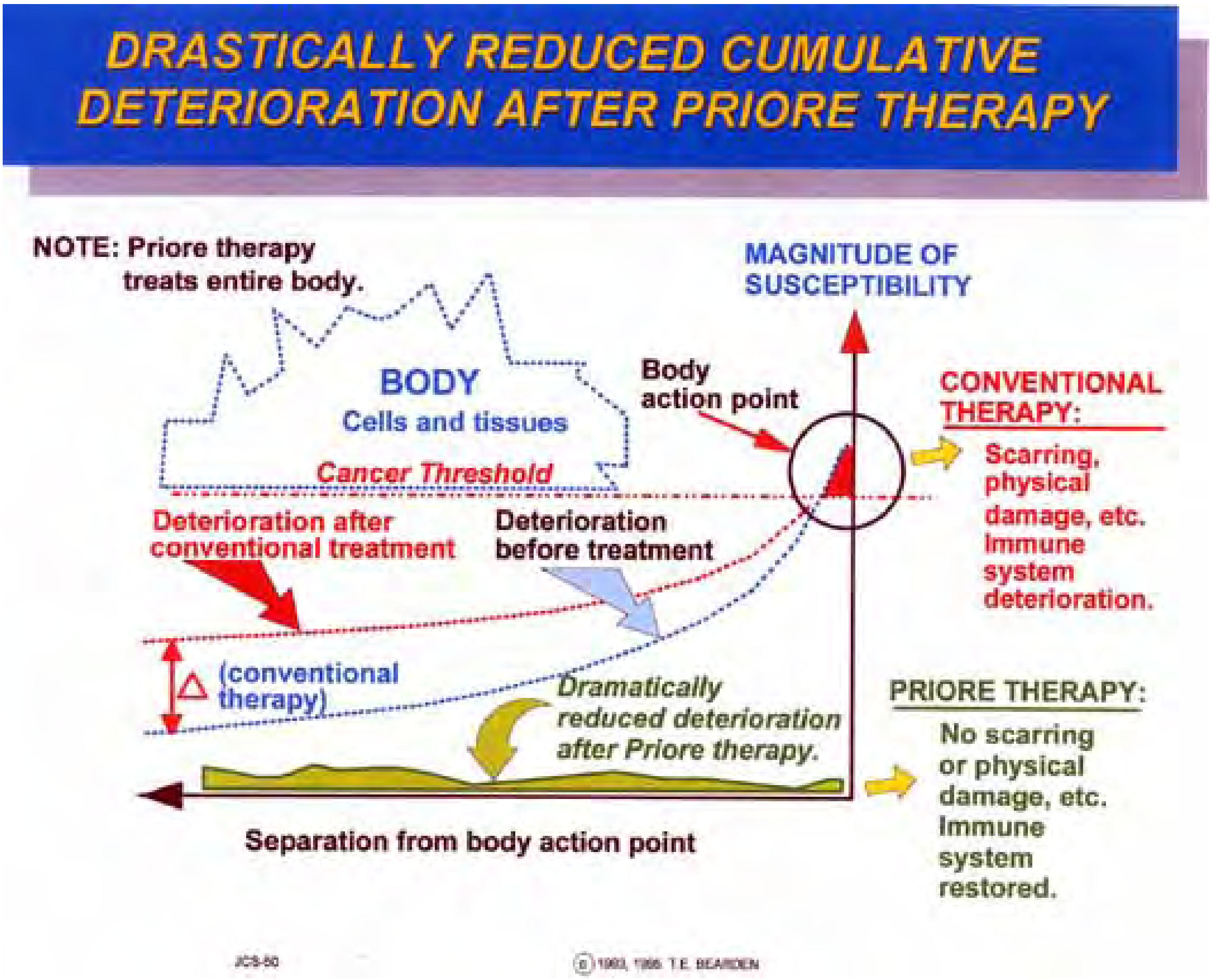

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY

Cells and tissues

Cancer Threshold

Body action point

MAGNITUDE OF SUSCEPTIBILITY

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

Deterioration after conventional treatment

Deterioration before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

Separation from body action point

JCS-50

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

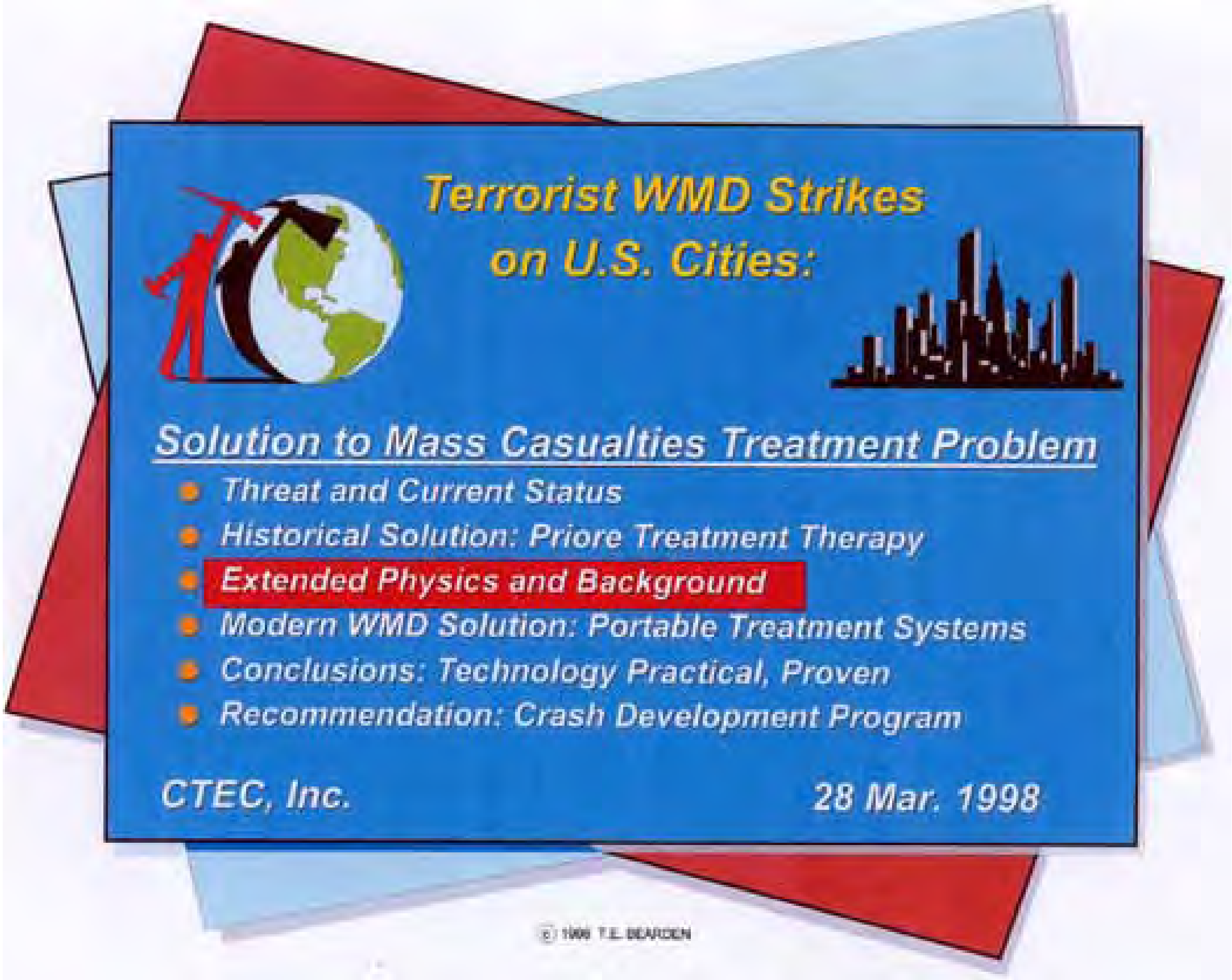

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

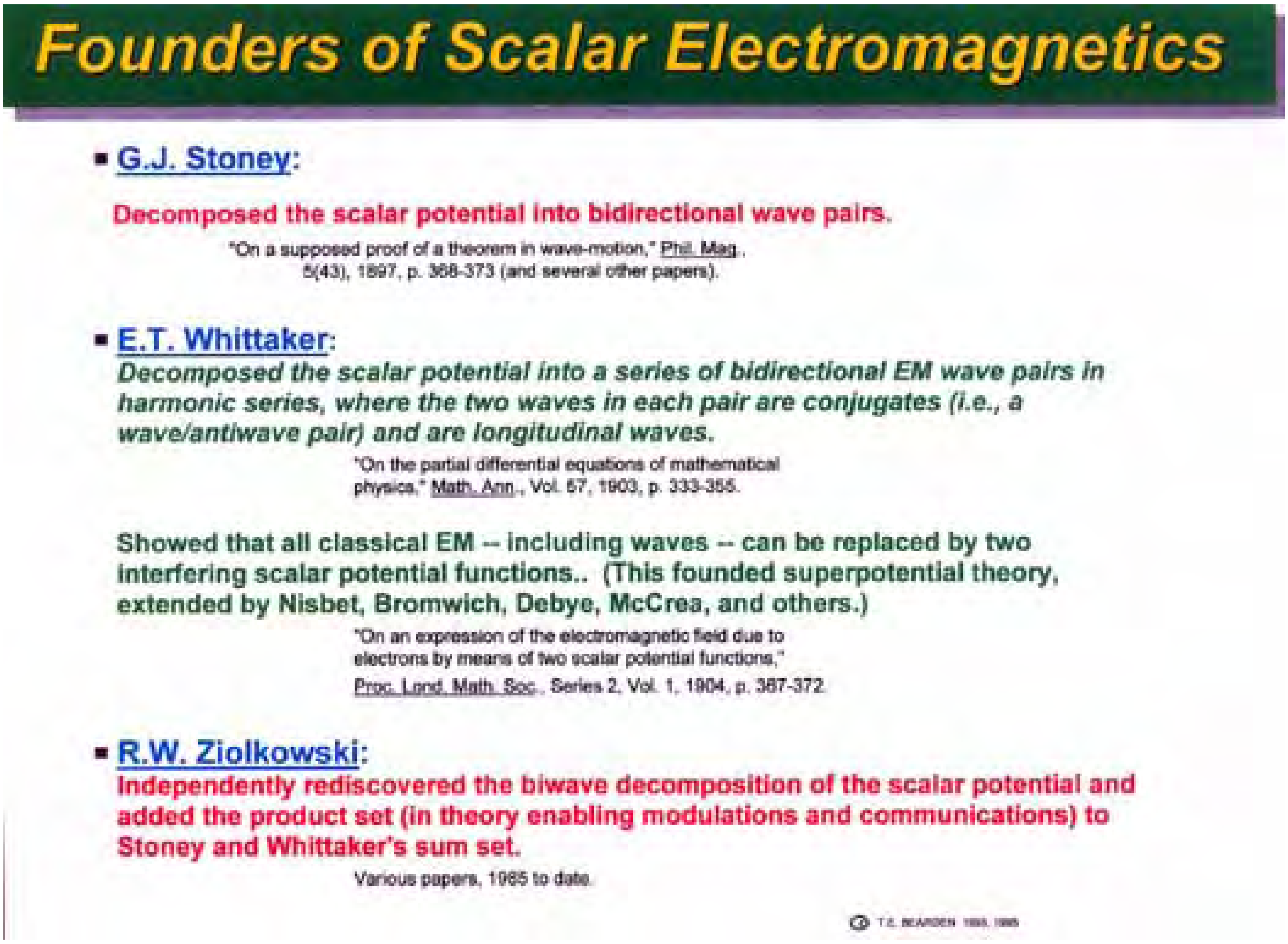

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

  Decomposed the scalar potential into bidirectional wave pairs.

  "On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

  Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

  "On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

  Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

  "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.

- **R.W. Ziolkowski:**

  Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

  Various papers, 1985 to date.

© T.E. BEARDEN 1993, 1995

Return to Slide Index

Next Slide

Previous Slide

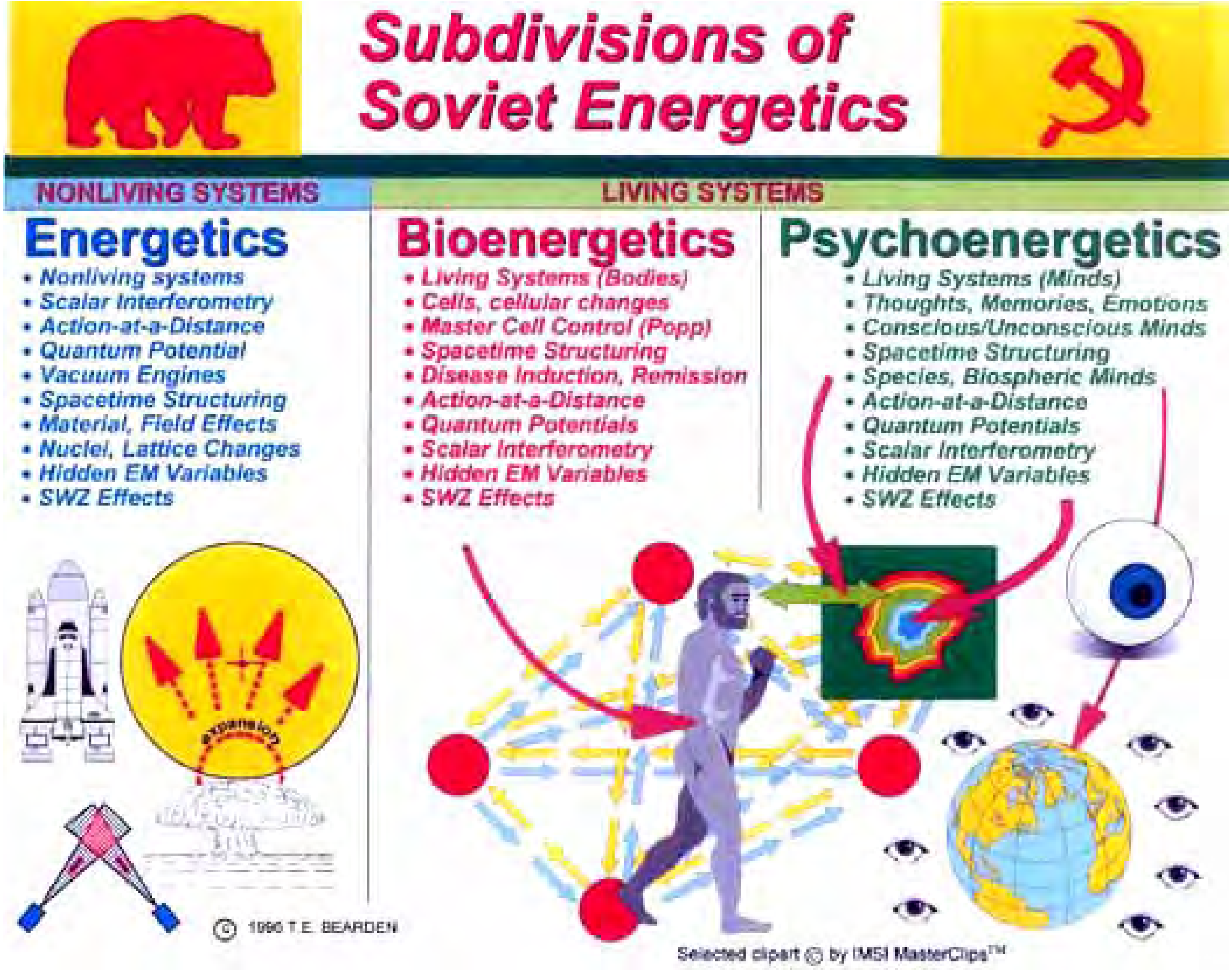

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

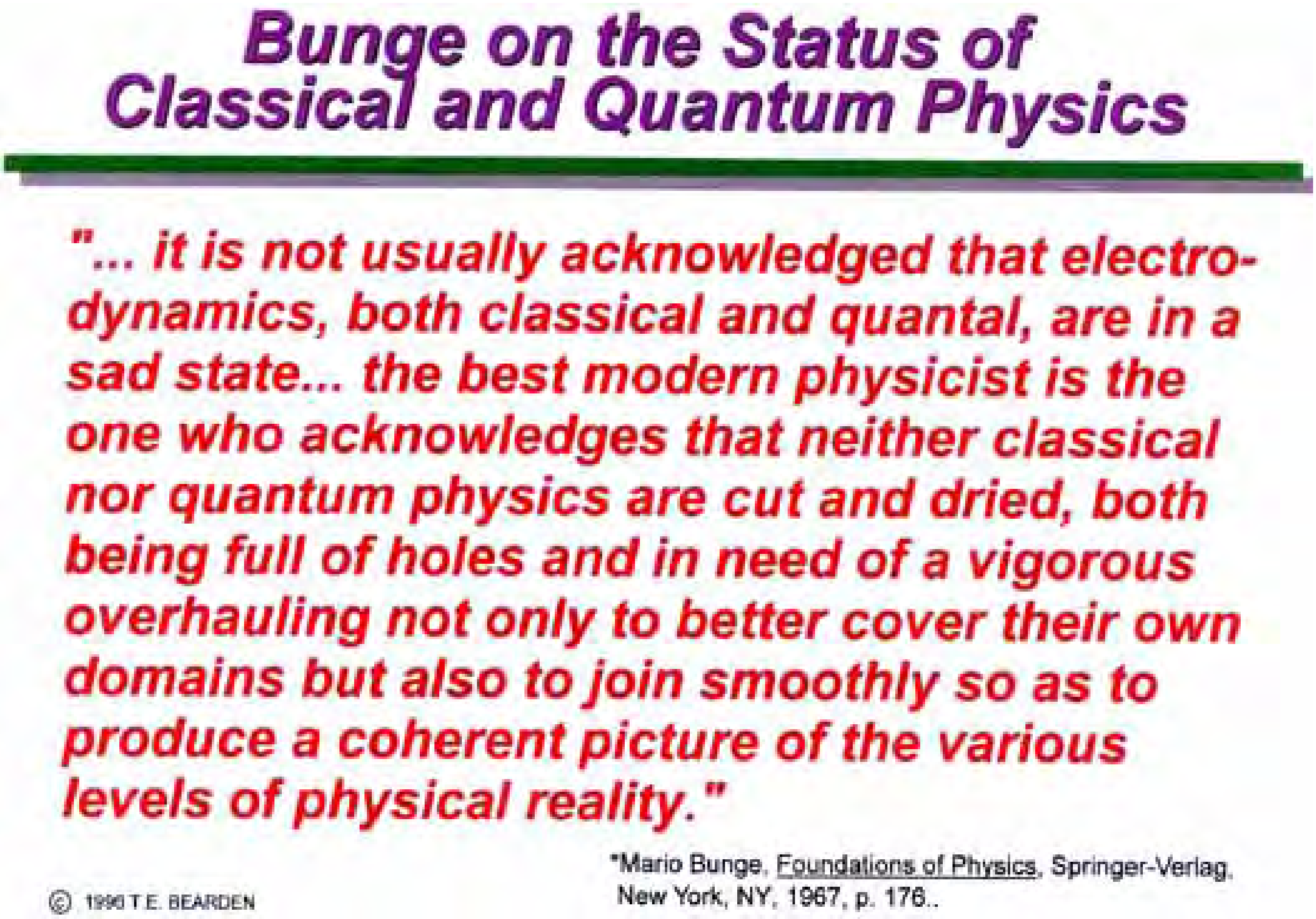

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

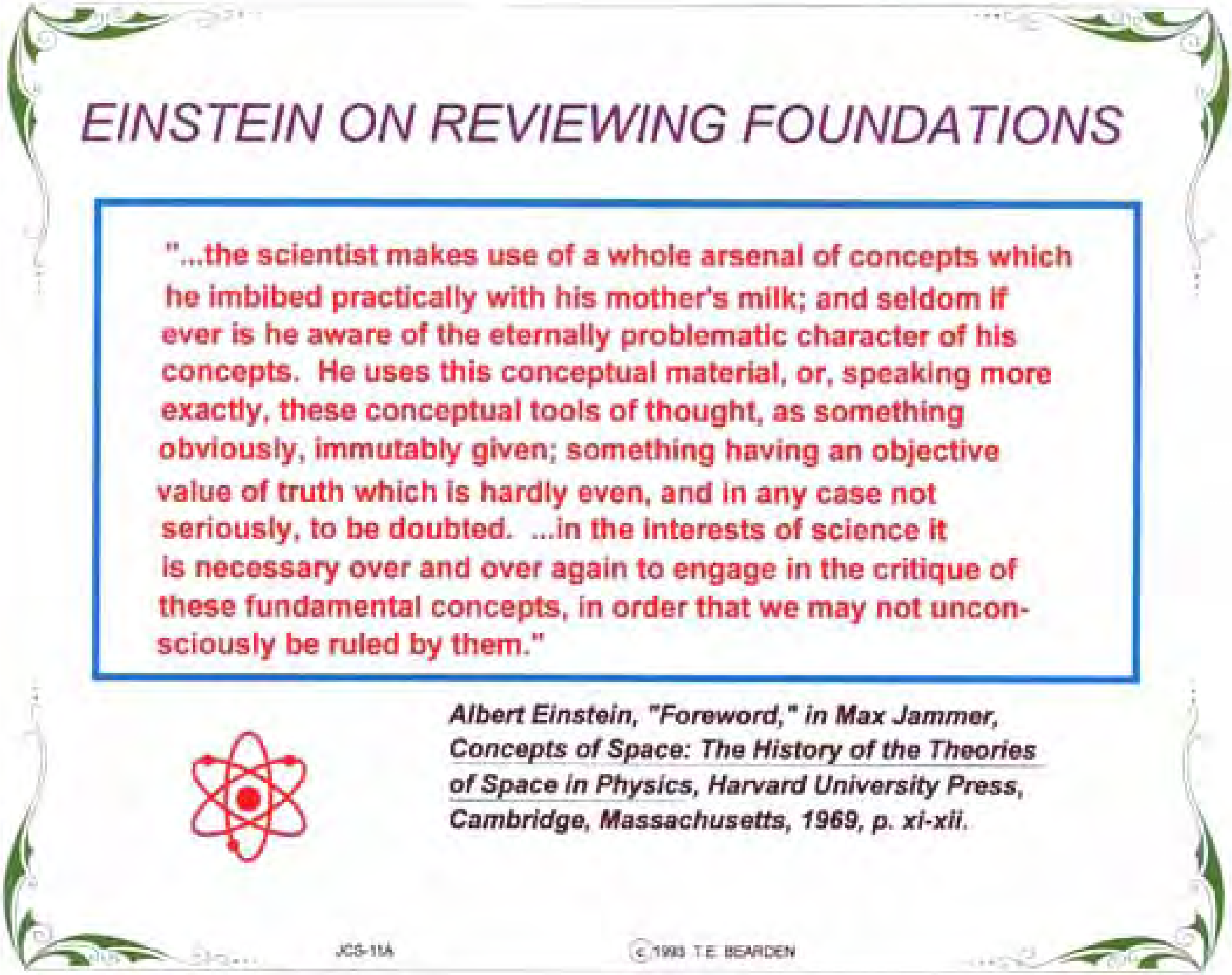

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

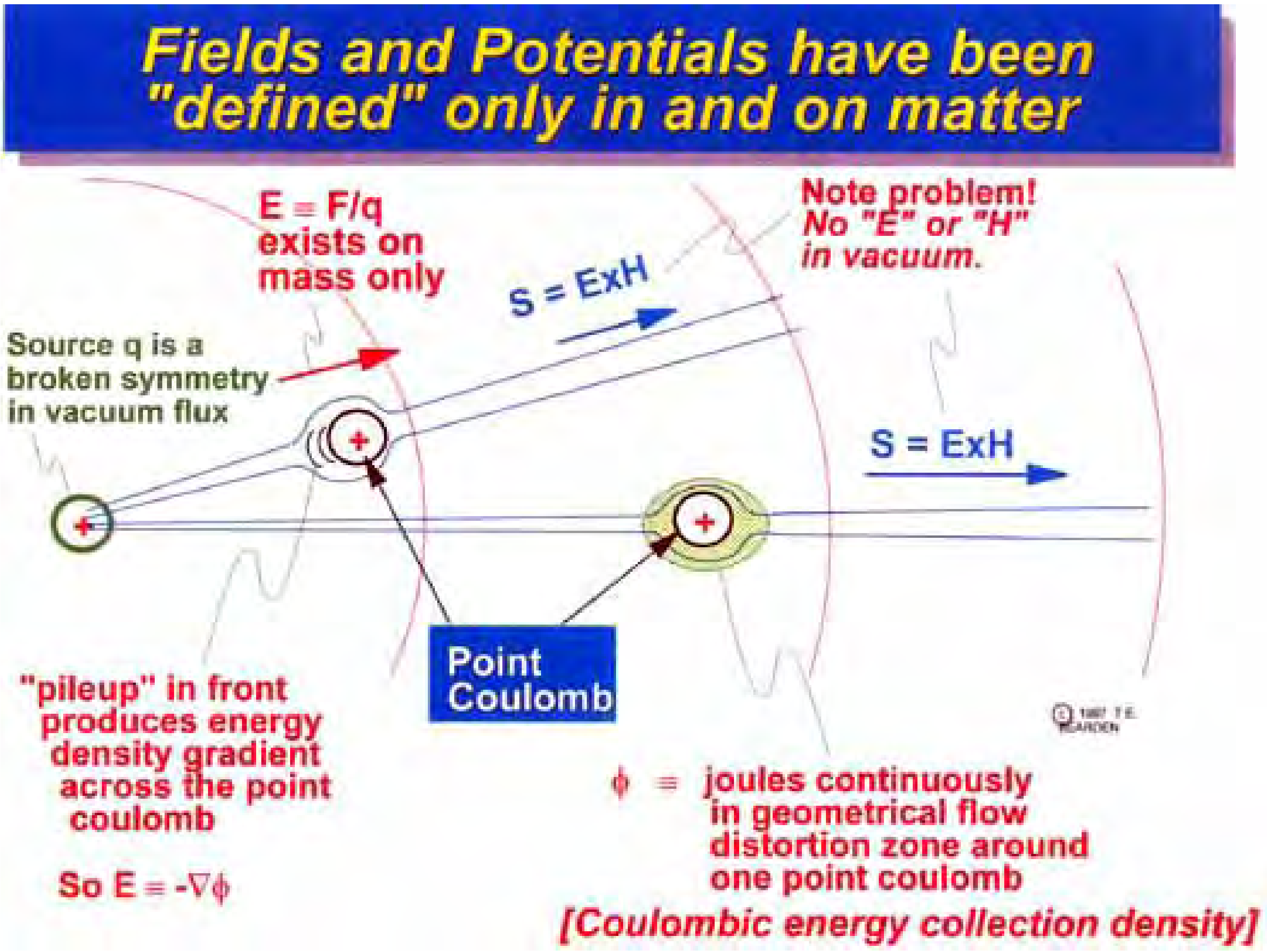

Fields and Potentials have been "defined" only in and on matter

E ≡ F/q
exists on
mass only

Note problem!
No "E" or "H"
in vacuum.

S = ExH

Source q is a
broken symmetry
in vacuum flux

S = ExH

Point
Coulomb

© 1997 T.E. BEARDEN

"pileup" in front
produces energy
density gradient
across the point
coulomb

So E ≡ -∇ϕ

ϕ ≡ joules continuously
in geometrical flow
distortion zone around
one point coulomb

[Coulombic energy collection density]

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

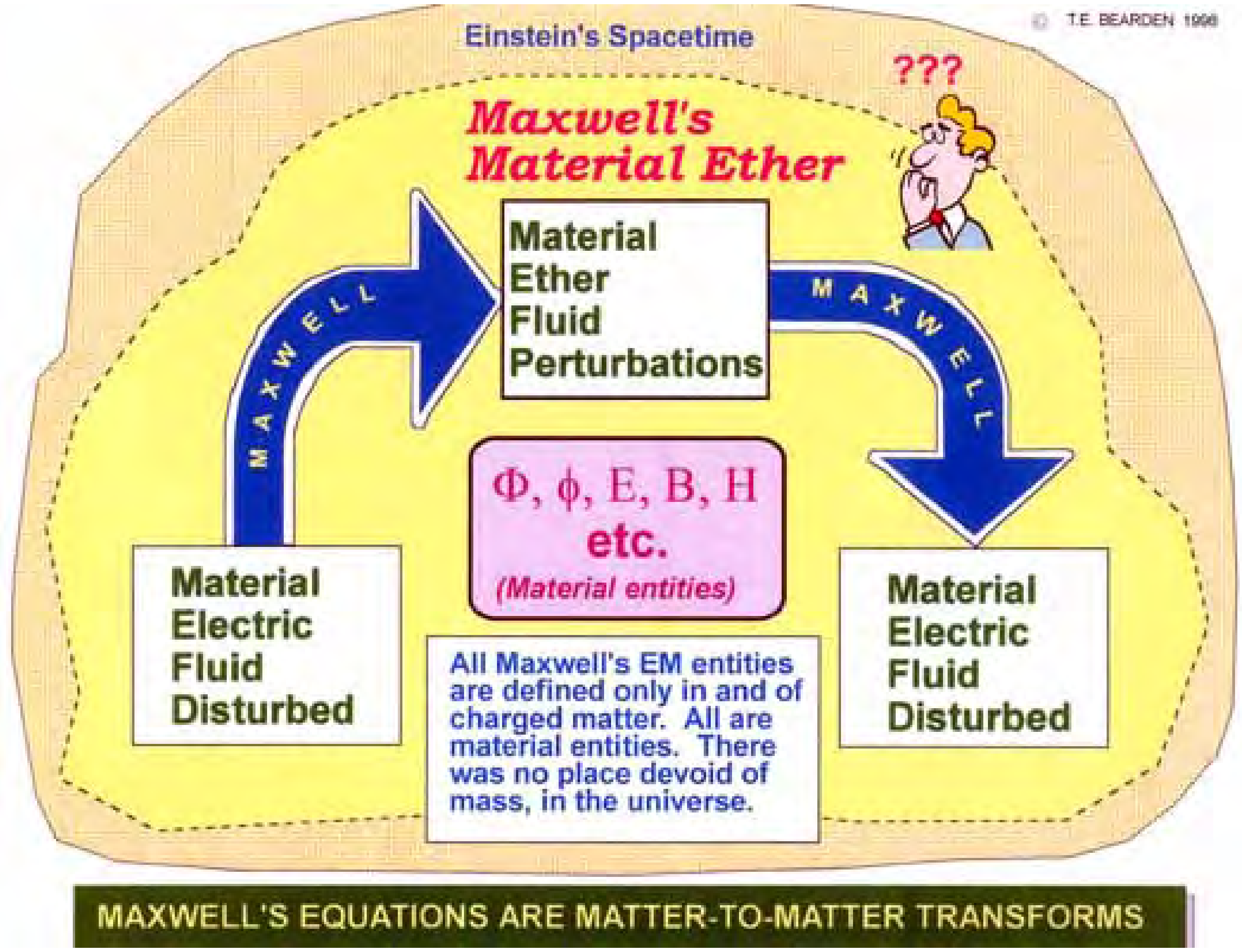

Einstein's Spacetime

© T.E. BEARDEN 1998

???

Maxwell's Material Ether

Material Ether Fluid Perturbations

Φ, ϕ, E, B, H etc.

(Material entities)

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

Return to Slide Index

Next Slide

Previous Slide

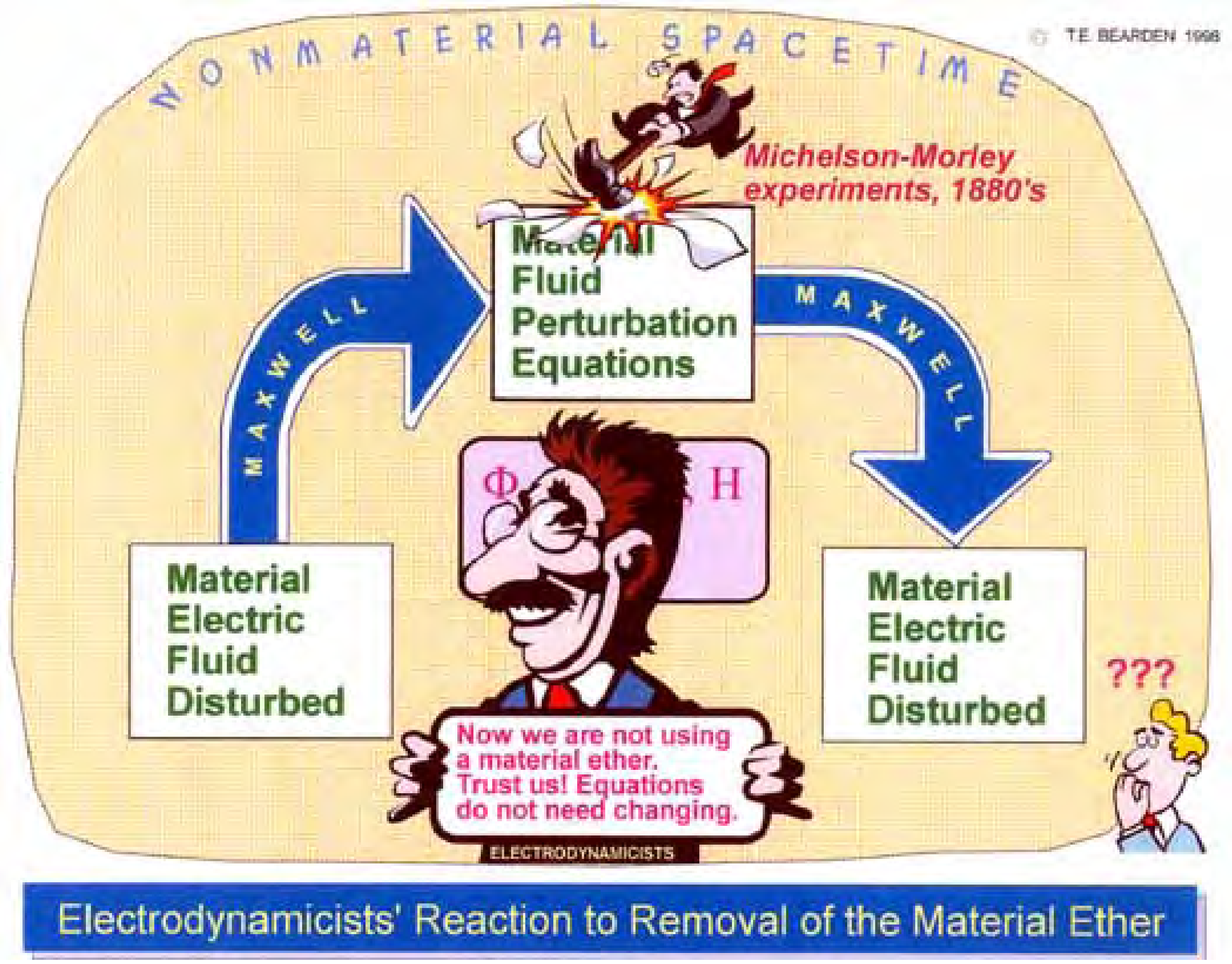
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Michelson-Morley experiments, 1880's
MAXWELL
MAXWELL
Φ
H
???
ELECTRODYNAMICISTS

Material Fluid Perturbation Equations

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

Now we are not using a material ether. Trust us! Equations do not need changing.

Electrodynamicists' Reaction to Removal of the Material Ether

Return to Slide Index

Next Slide

Previous Slide

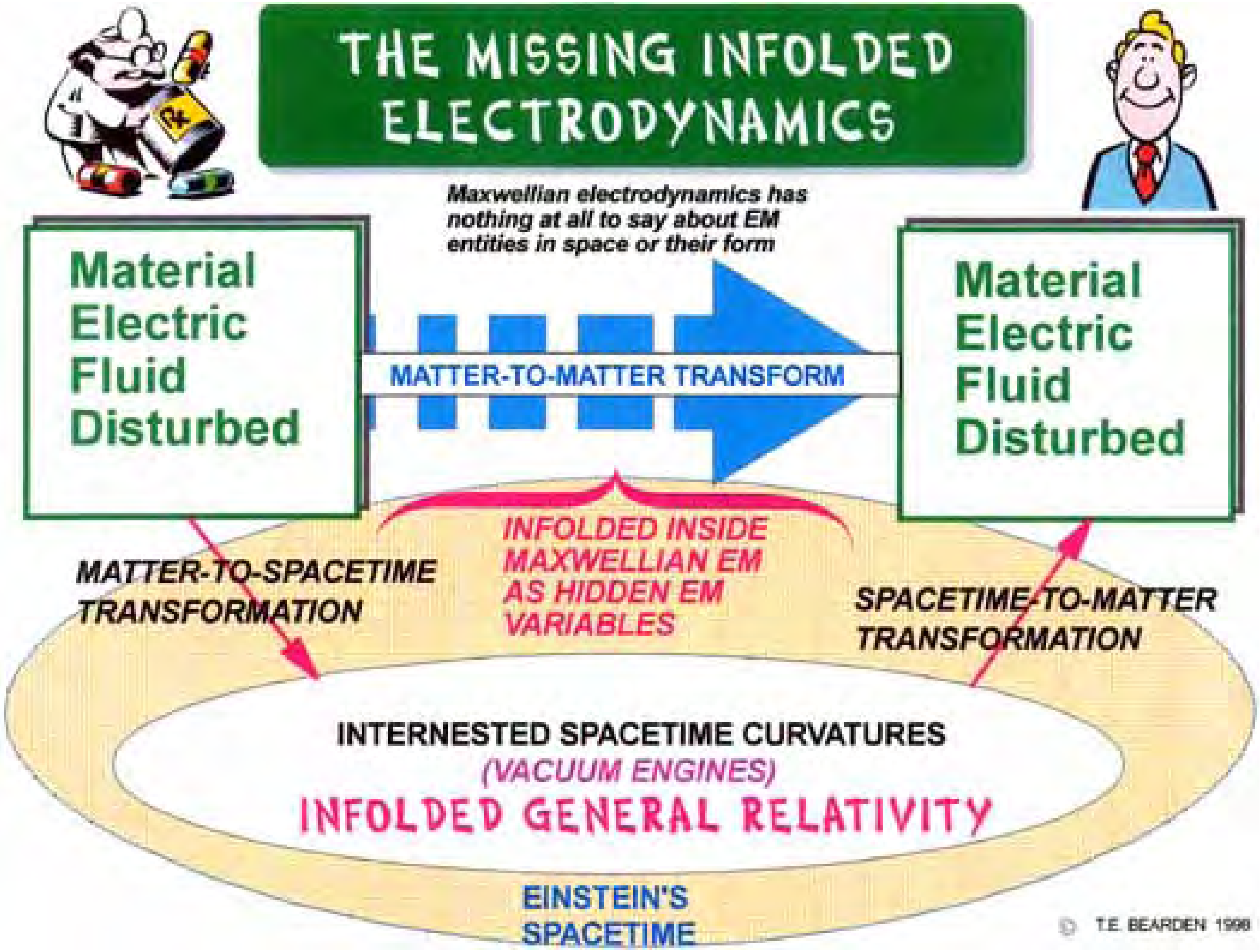

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

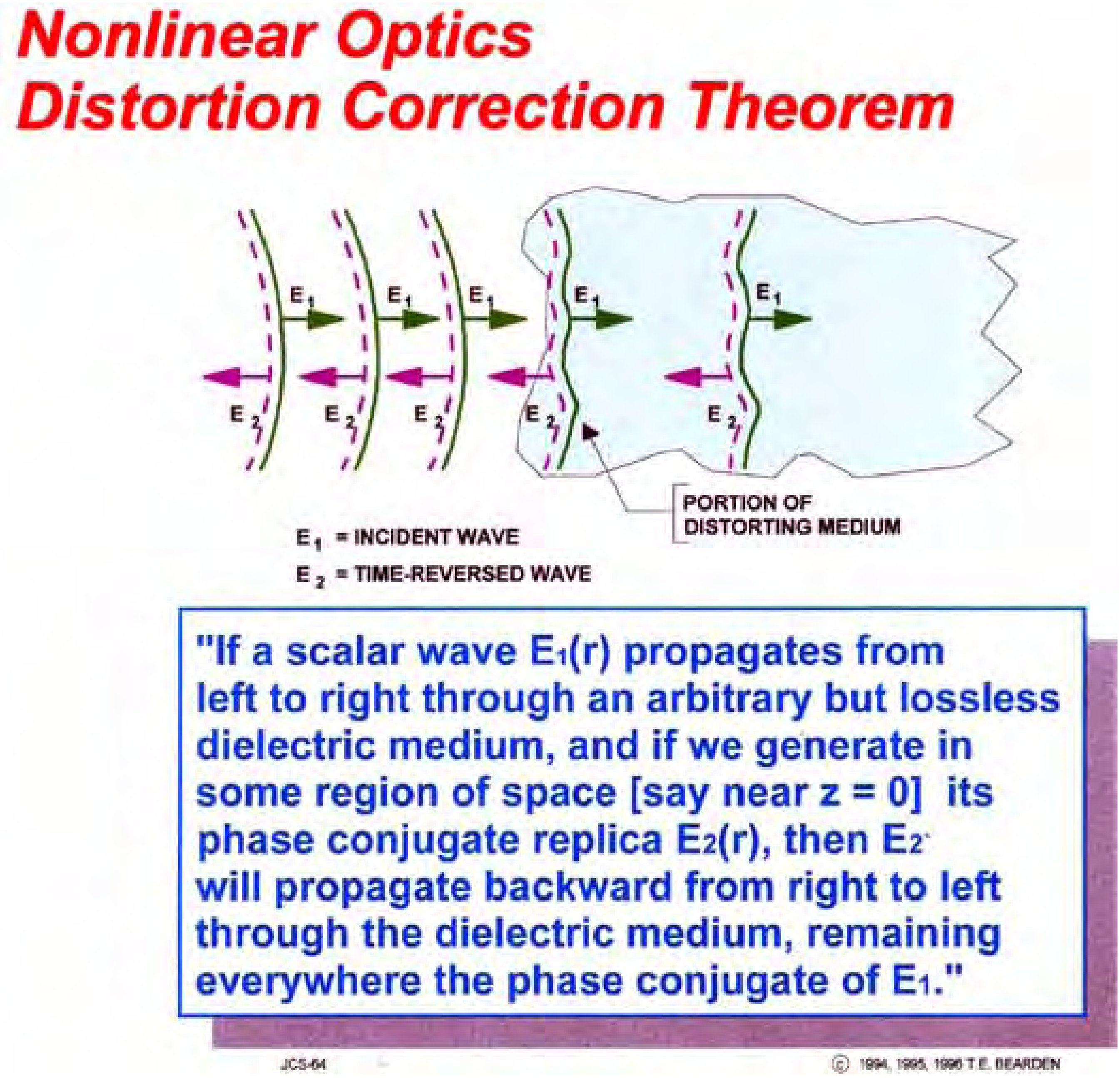

Nonlinear Optics Distortion Correction Theorem

E_1
E_1
E_1
E_1
E_1
E_2
E_2
E_2
E_2
E_2
PORTION OF DISTORTING MEDIUM

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

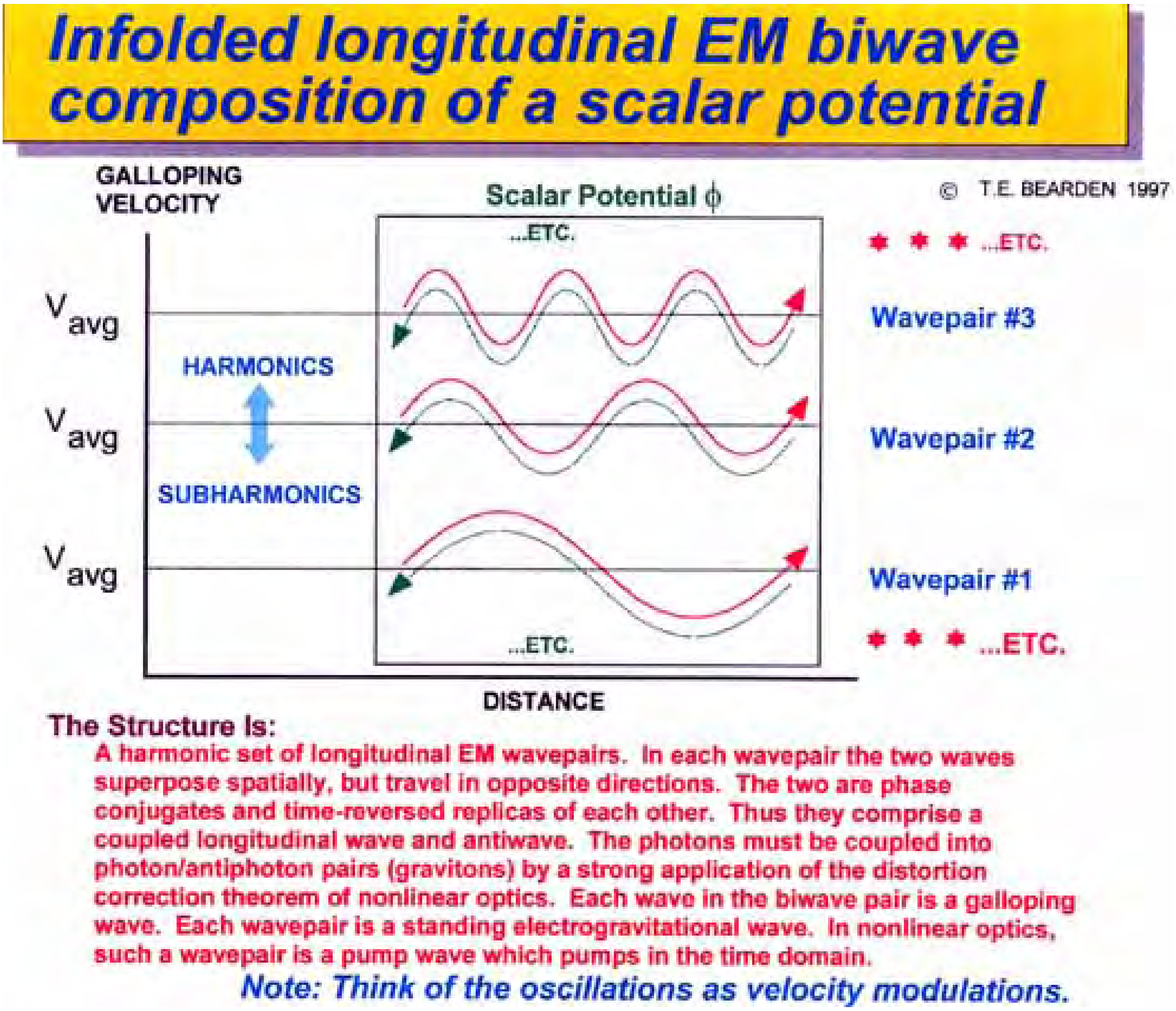

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
✱ ✱ ✱ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

[Return to Slide Index]
[Next Slide]
[Previous Slide]

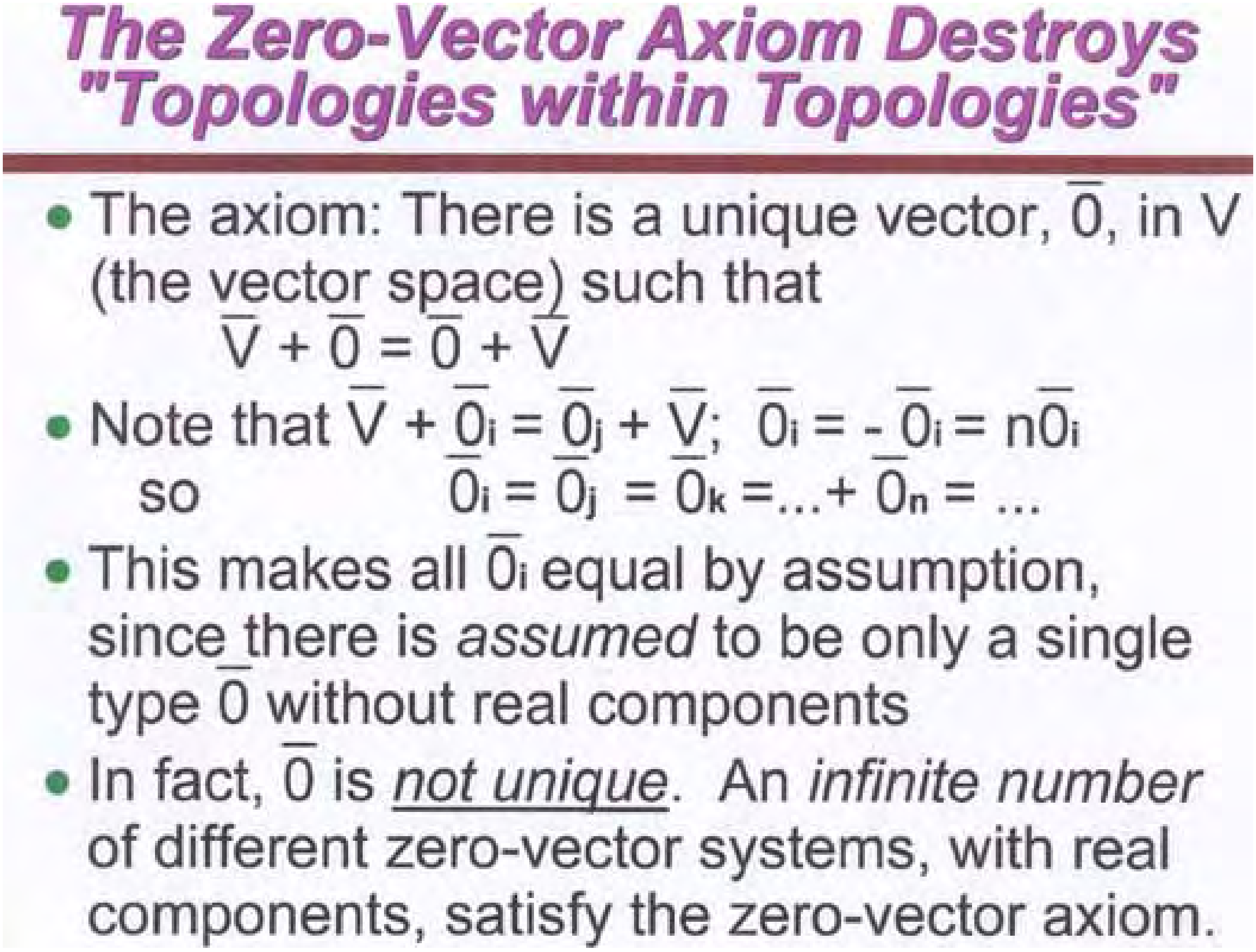

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $$\overline{V} + \overline{0} = \overline{0} + \overline{V}$$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

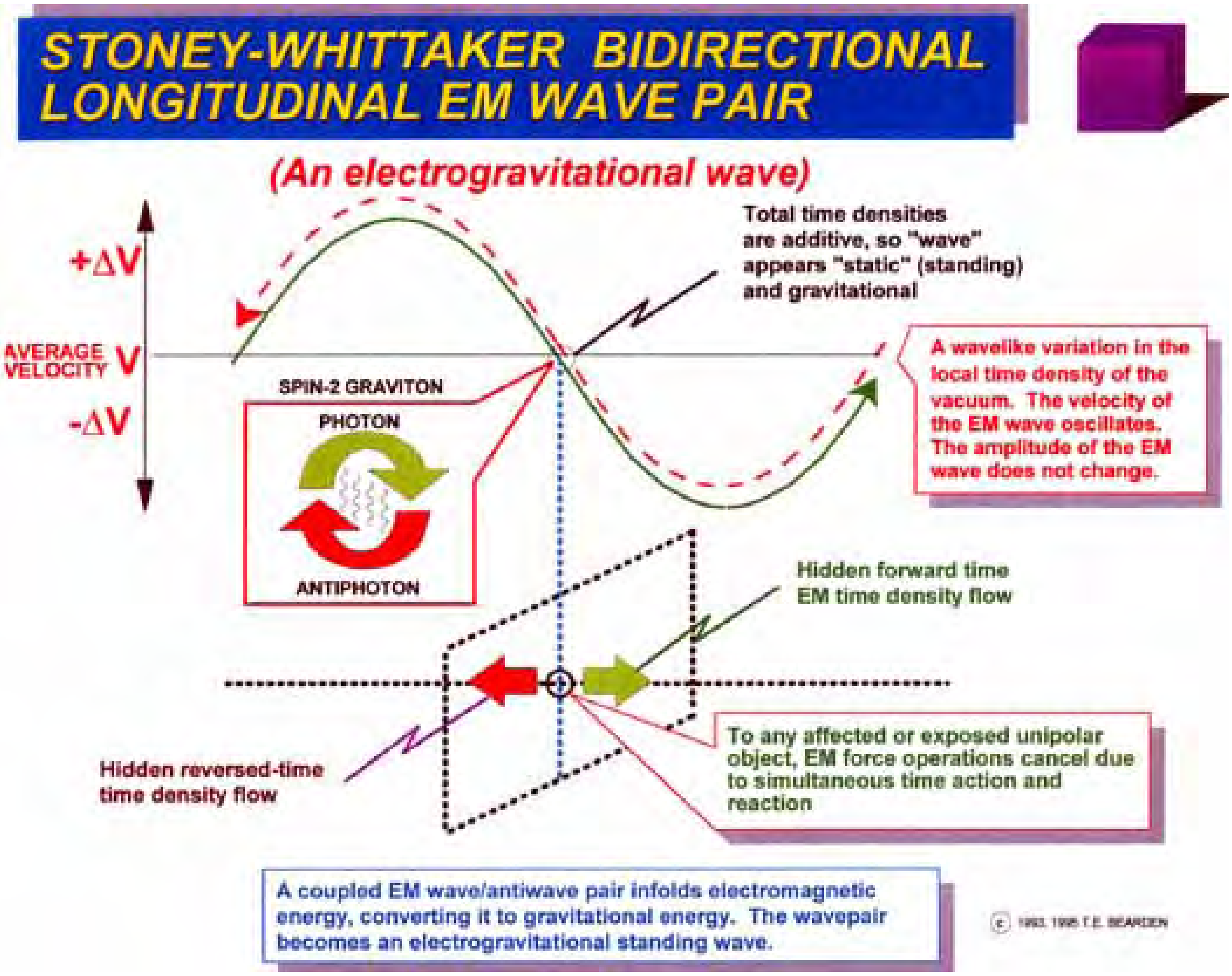

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

(c) 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

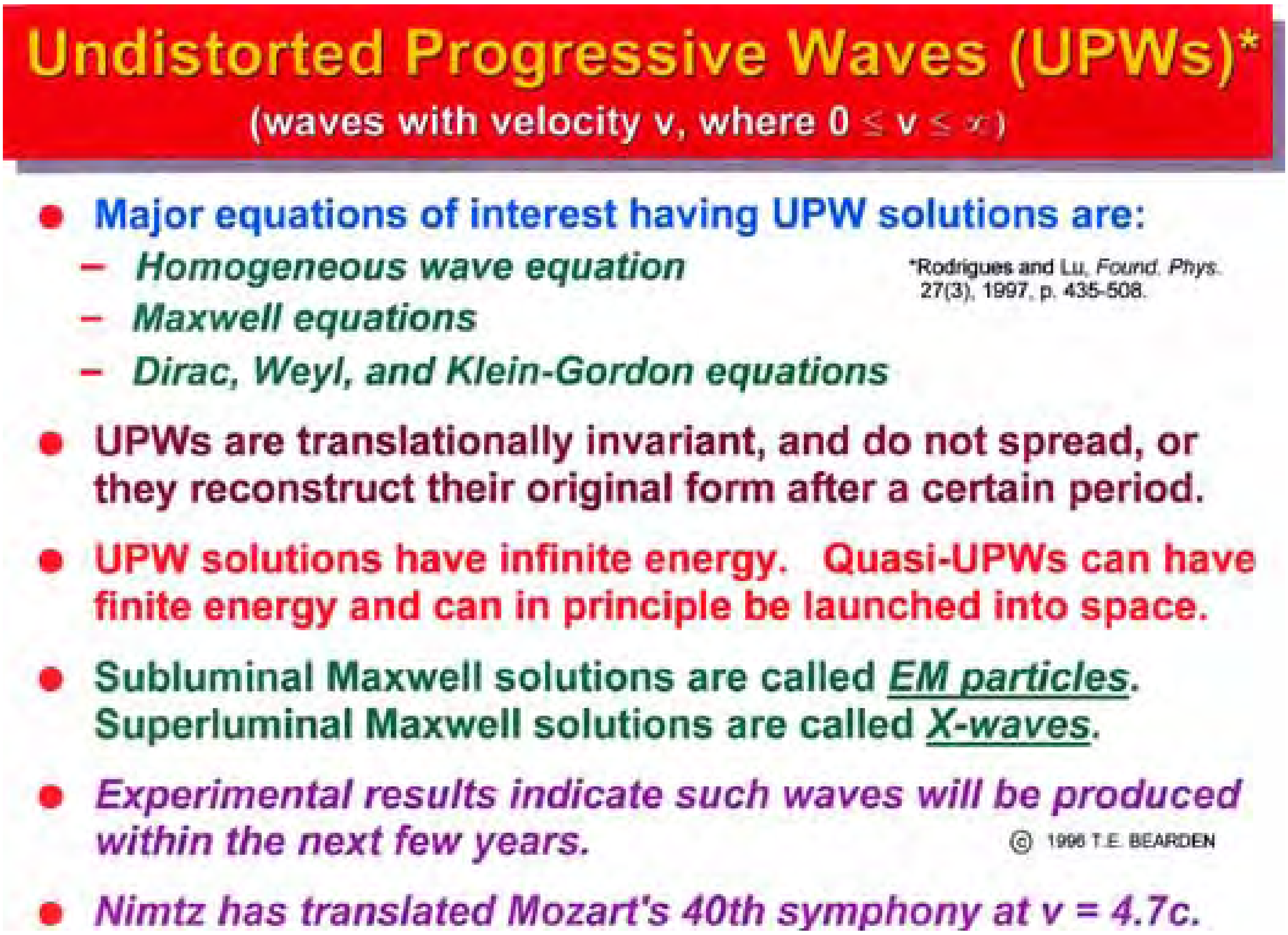

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \leq v \leq \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*
- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*
- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

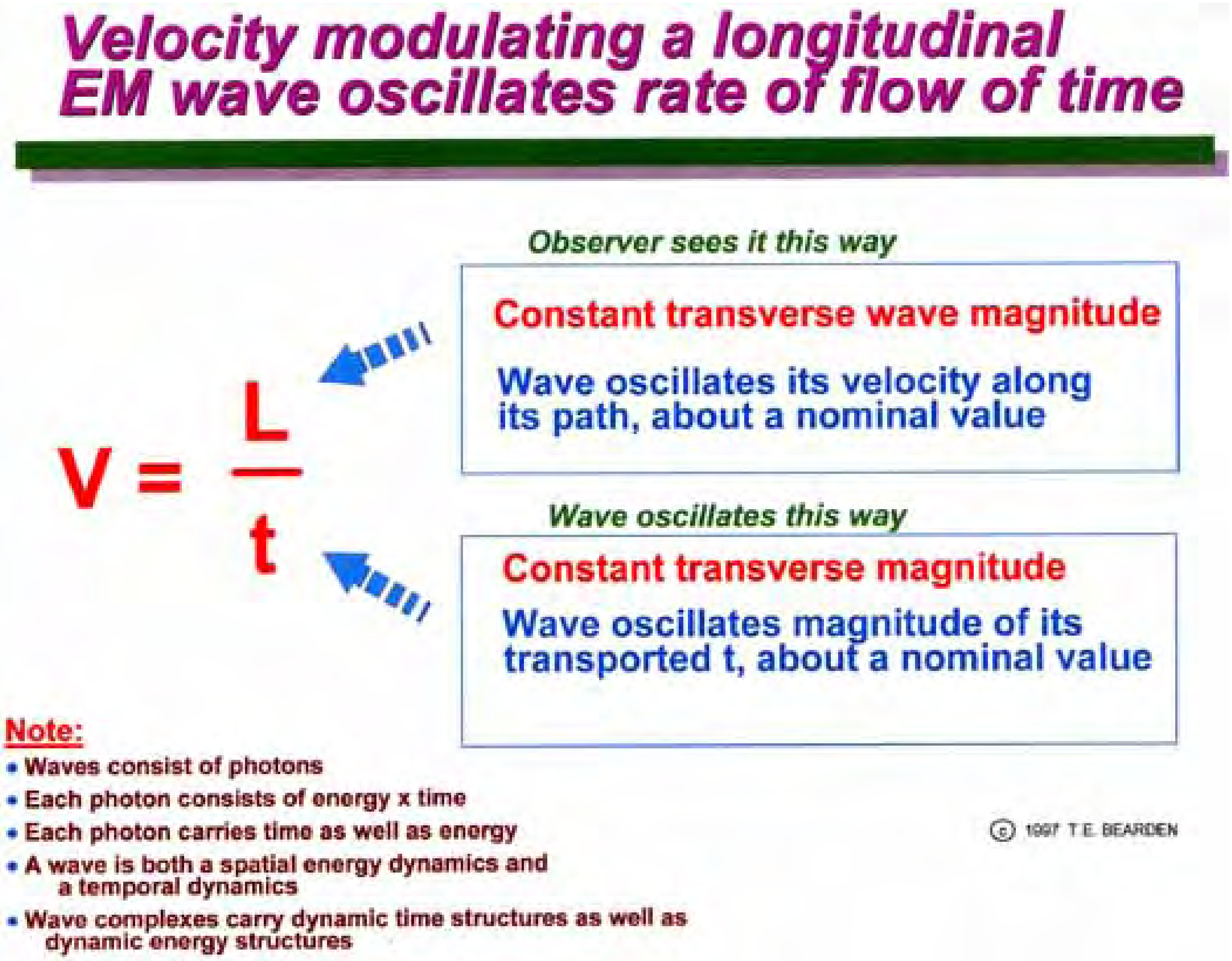

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$$V = \frac{L}{t}$$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

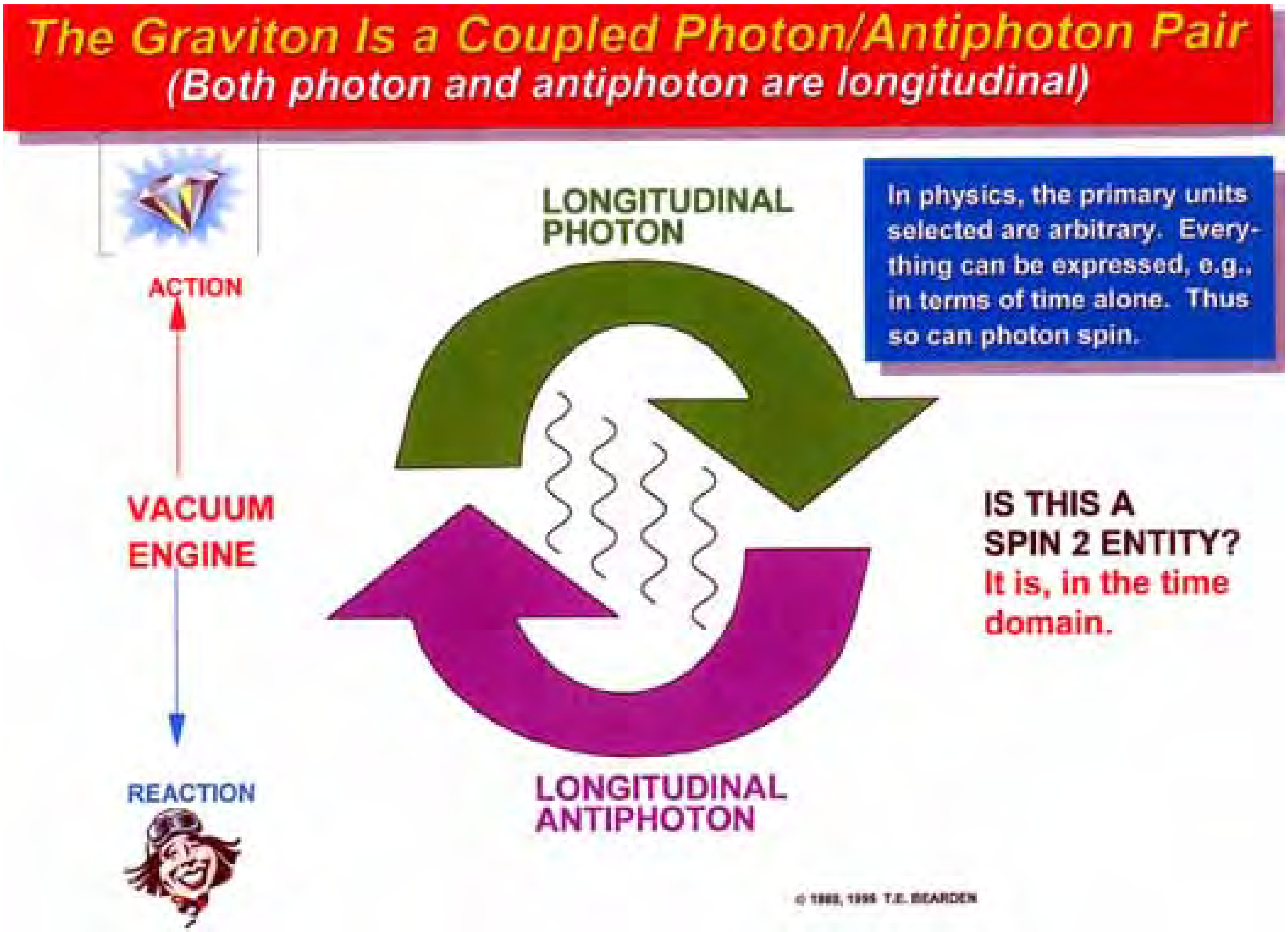

The Graviton Is a Coupled Photon/Antiphoton Pair
(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

IS THIS A SPIN 2 ENTITY?
It is, in the time domain.

LONGITUDINAL PHOTON
ACTION
VACUUM ENGINE
REACTION
LONGITUDINAL ANTIPHOTON

© [illegible], [illegible] T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

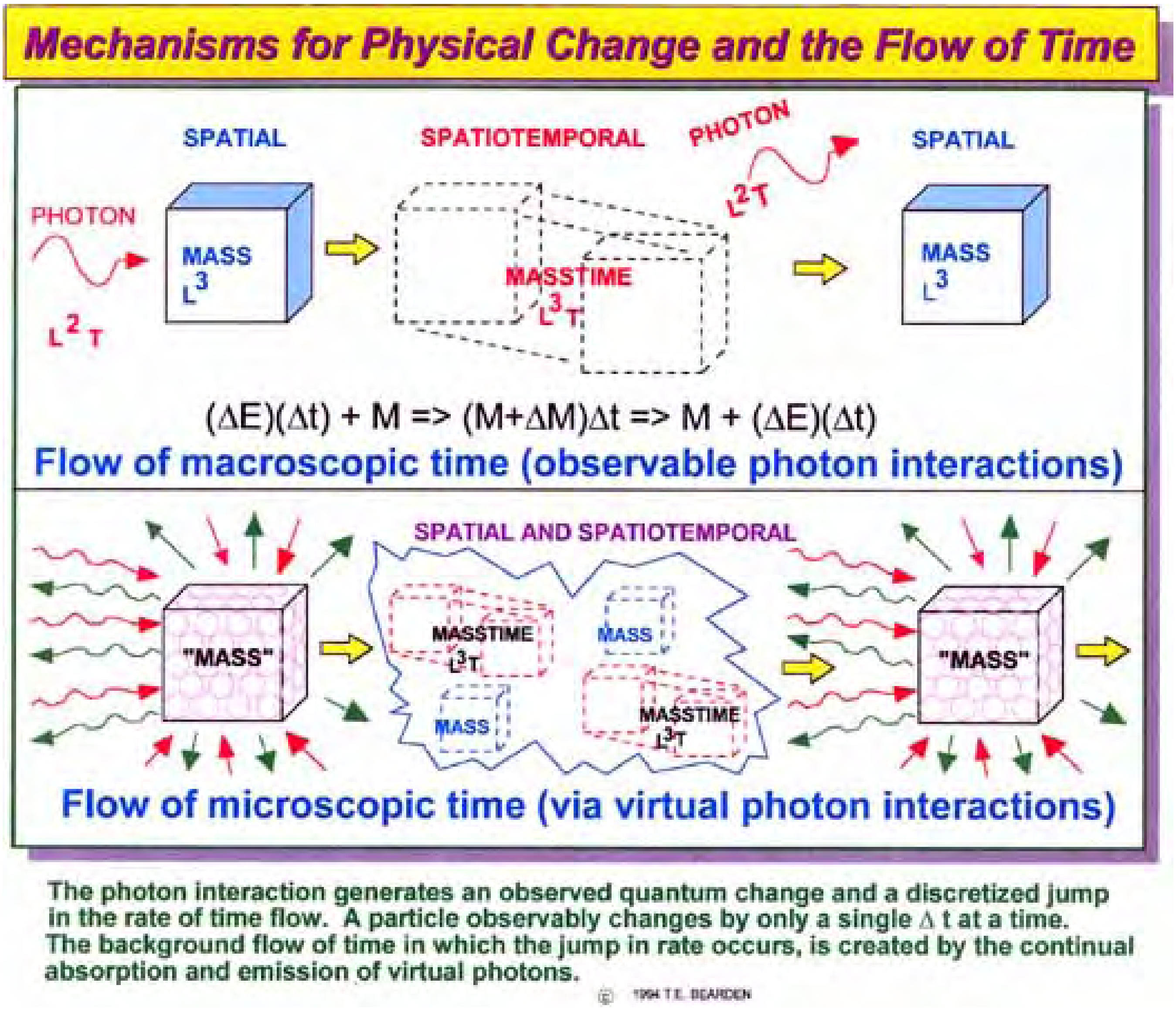

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS L^3
L^2T
MASSTIME L^3T
L^2T
MASS L^3

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δt at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

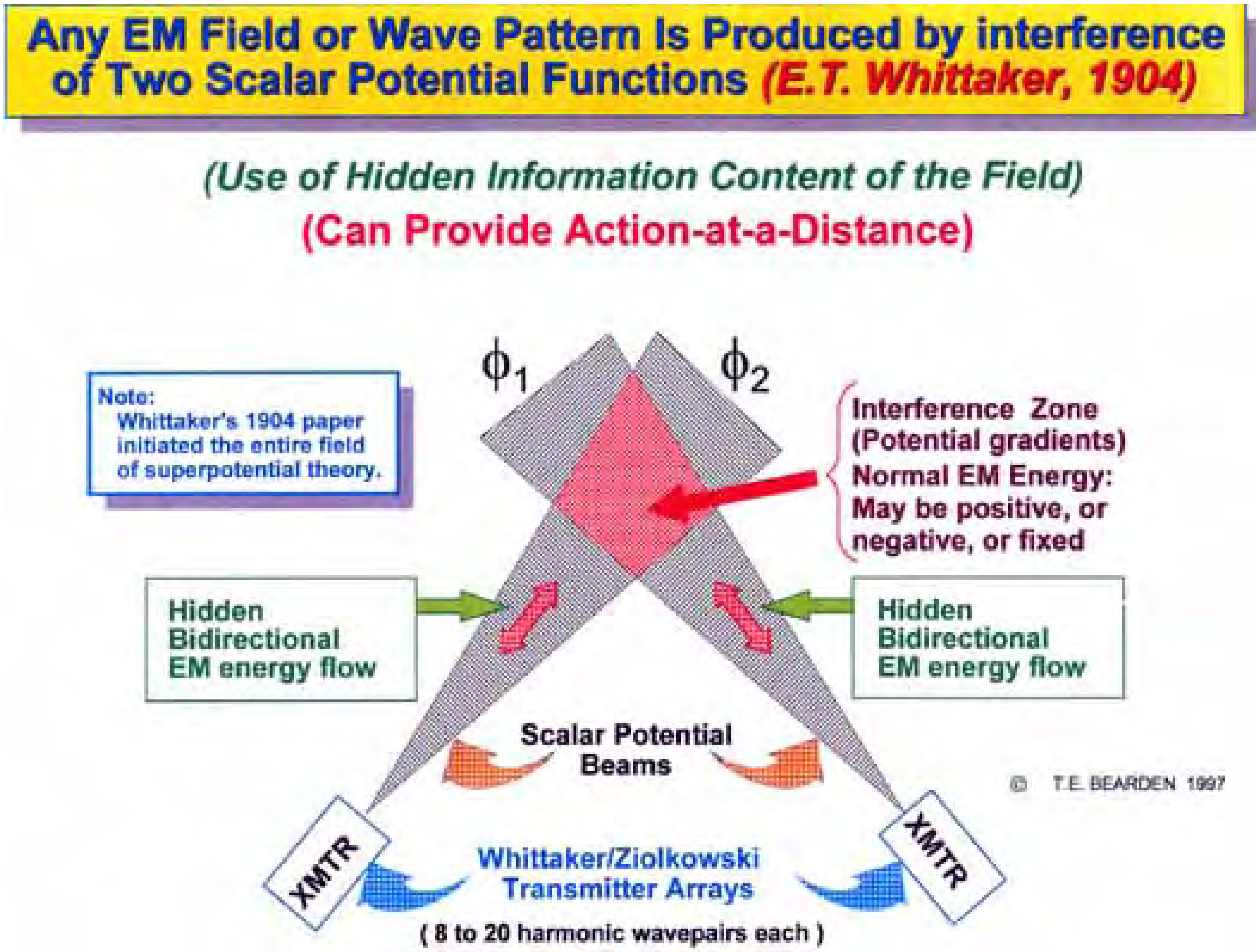

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or
negative, or fixed

Hidden
Bidirectional
EM energy flow

Hidden
Bidirectional
EM energy flow

Scalar Potential
Beams

Whittaker/Ziolkowski
Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

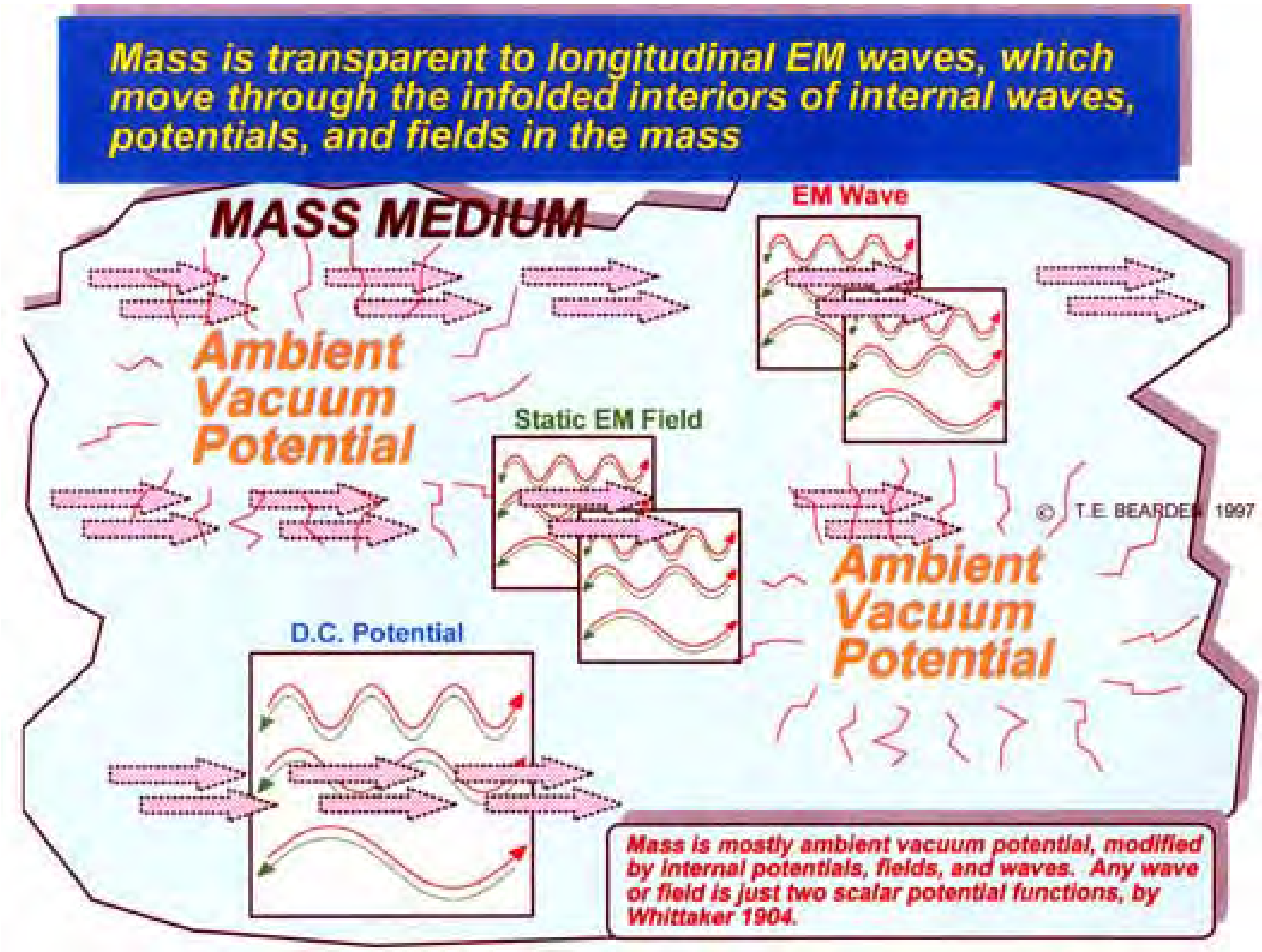

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
EM Wave
Ambient Vacuum Potential
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

Return to Slide Index

Next Slide

Previous Slide

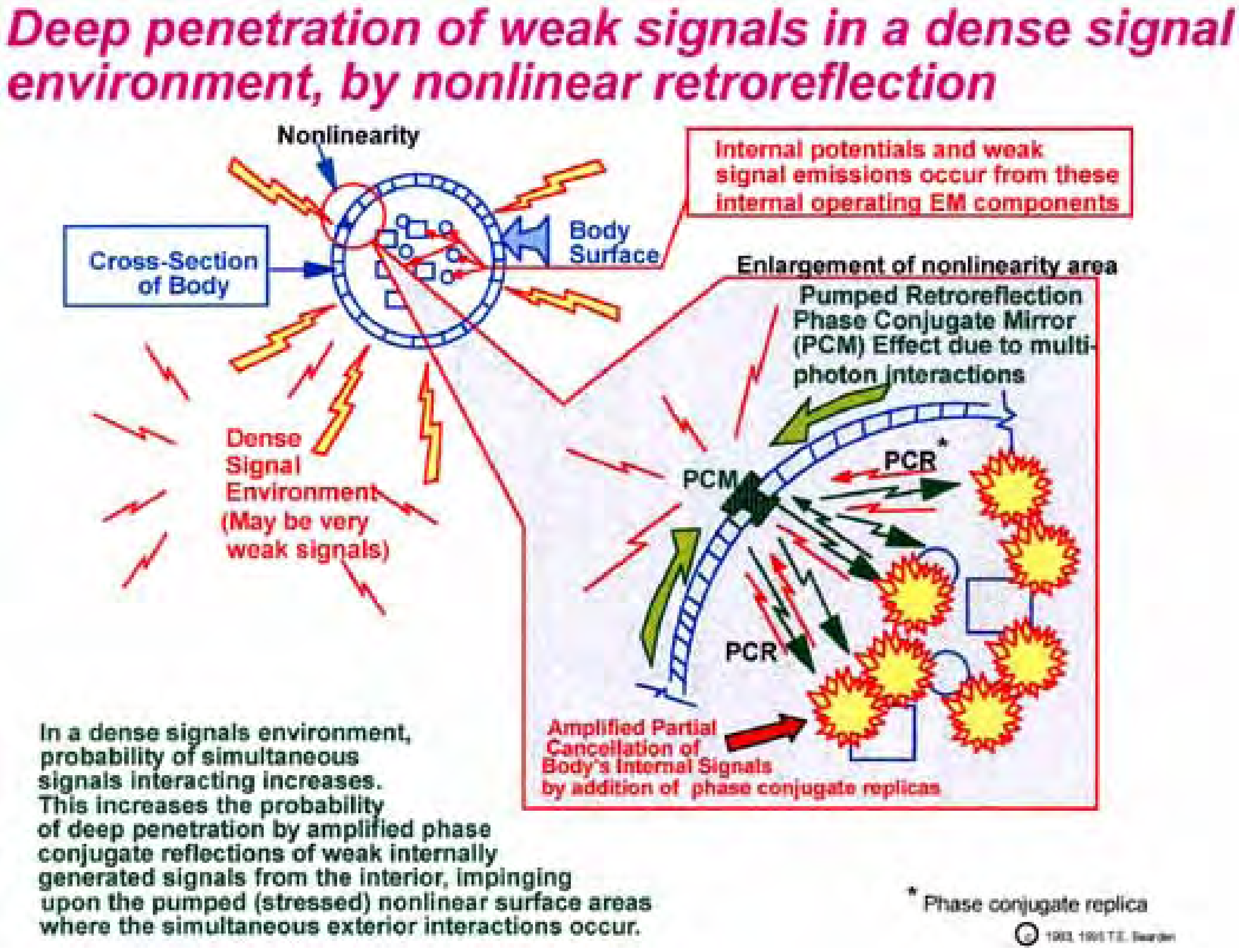

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity
Cross-Section of Body
Body Surface
Dense Signal Environment (May be very weak signals)
Enlargement of nonlinearity area
PCM
PCR*
PCR

Internal potentials and weak signal emissions occur from these internal operating EM components

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

Return to Slide Index

Next Slide

Previous Slide

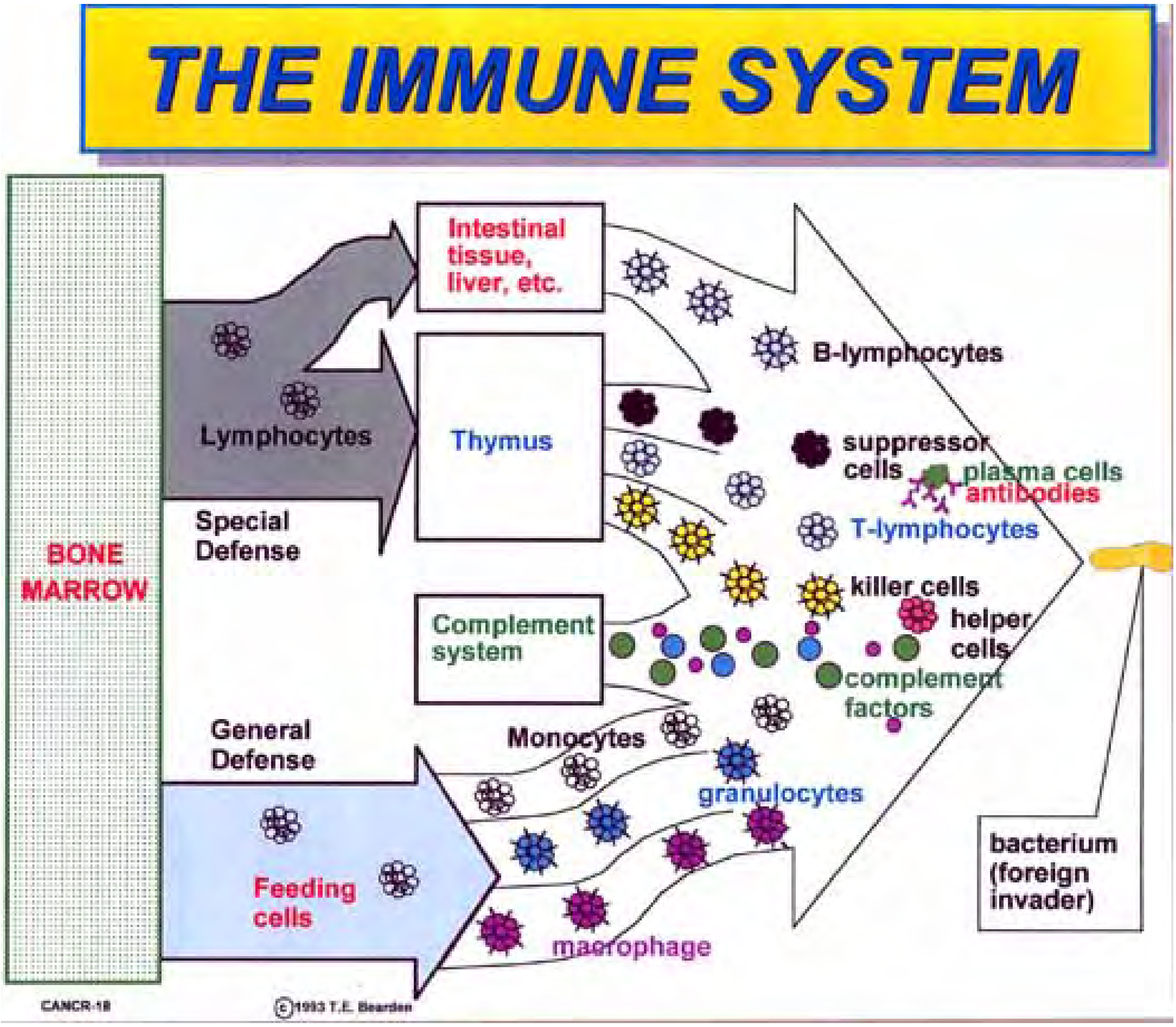

THE IMMUNE SYSTEM

BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
Complement system
General Defense
Feeding cells
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
Monocytes
granulocytes
macrophage
bacterium (foreign invader)
CANCR-18

©1993 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

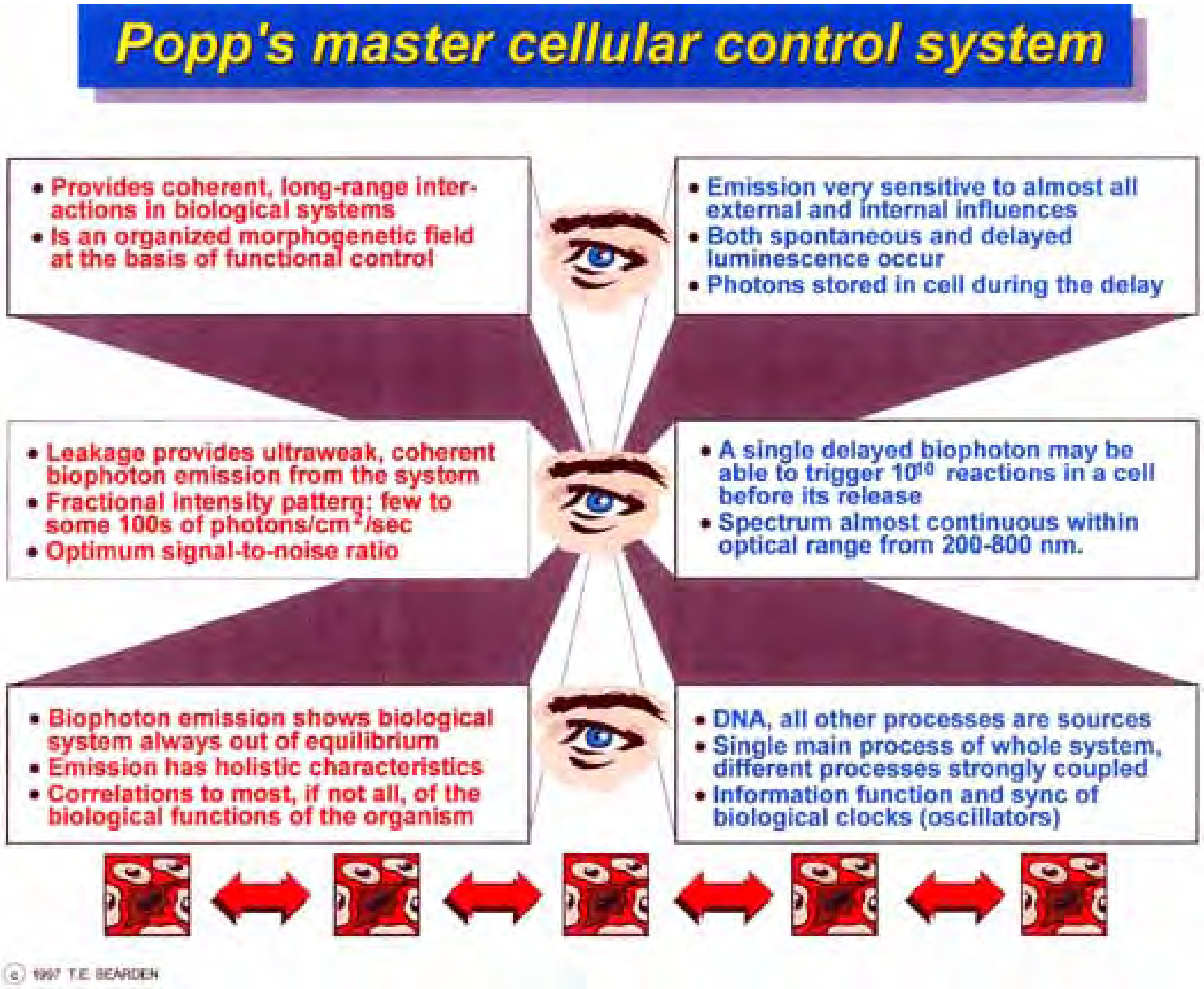

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

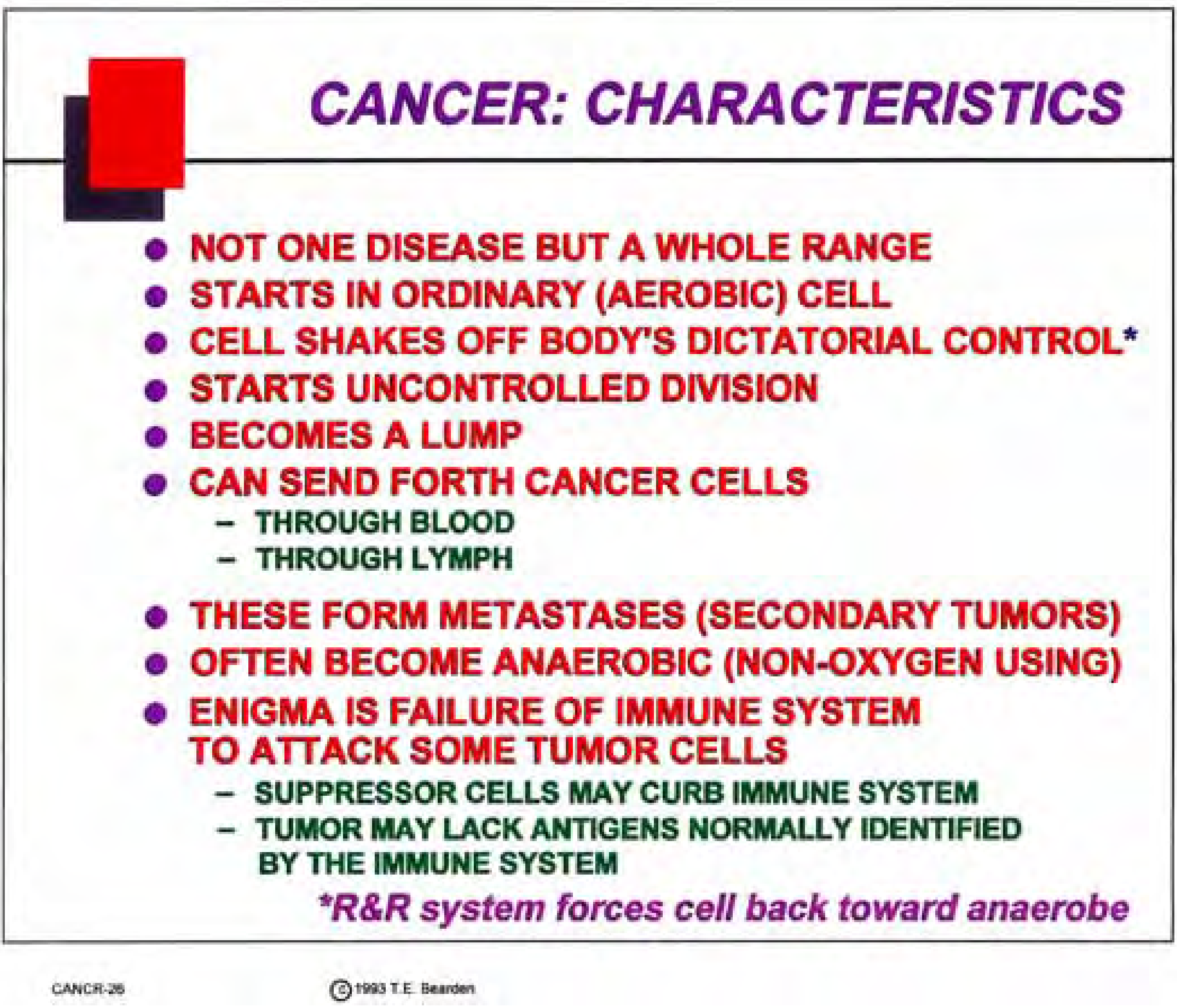

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

**R&R system forces cell back toward anaerobe*

© 1993 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

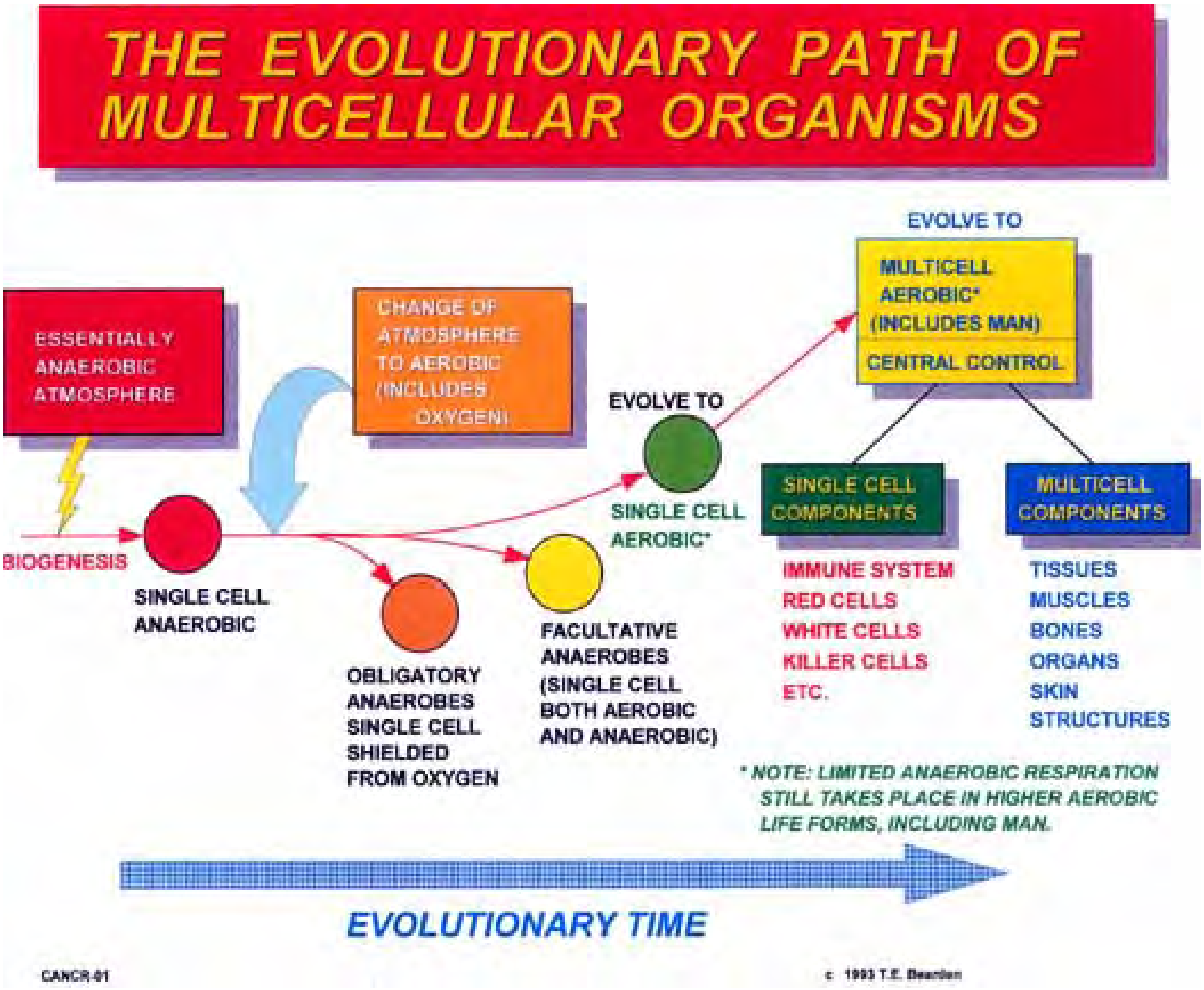

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.

MULTICELL COMPONENTS

TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES

*NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

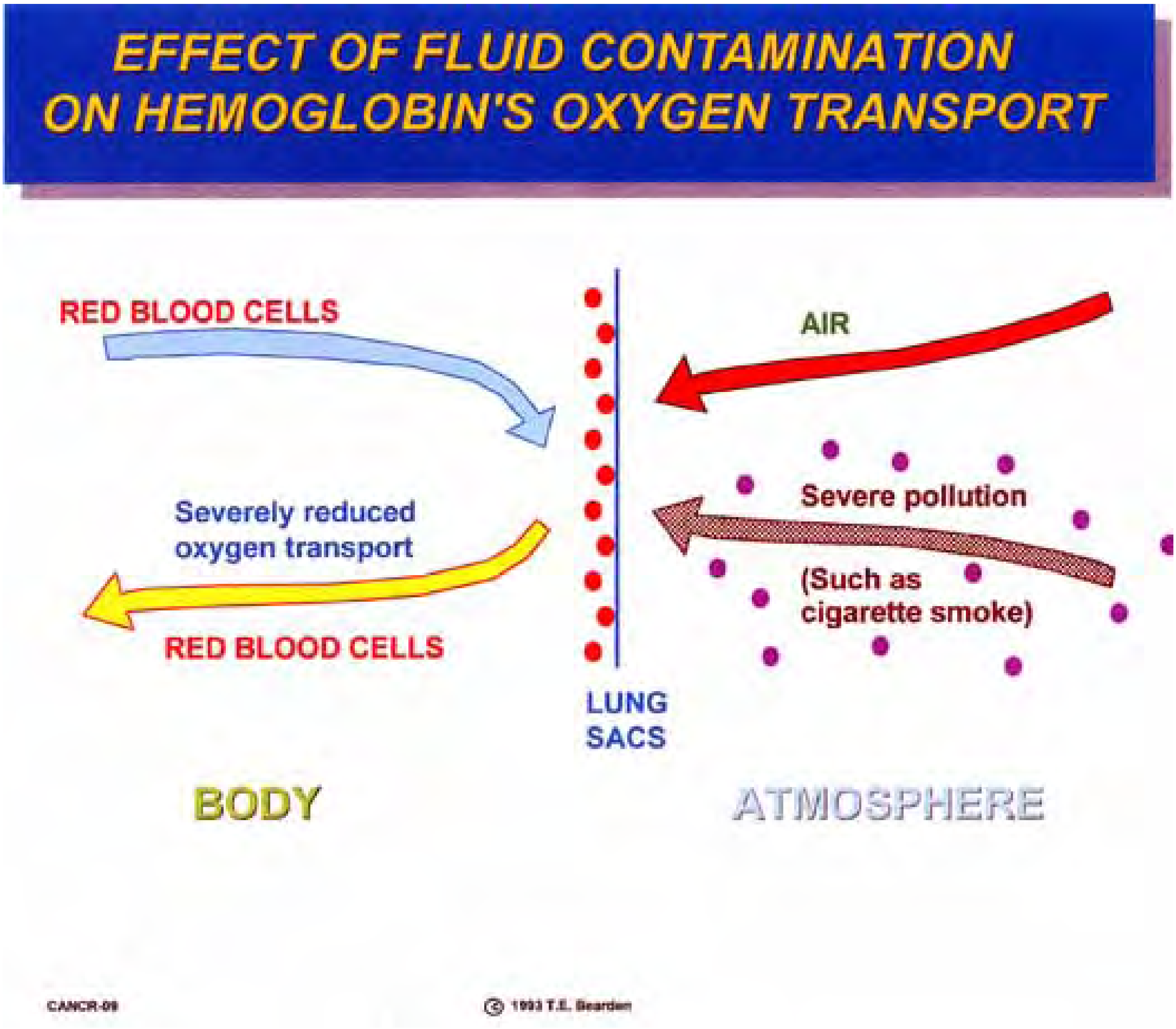

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severely reduced oxygen transport

Severe pollution

(Such as cigarette smoke)

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

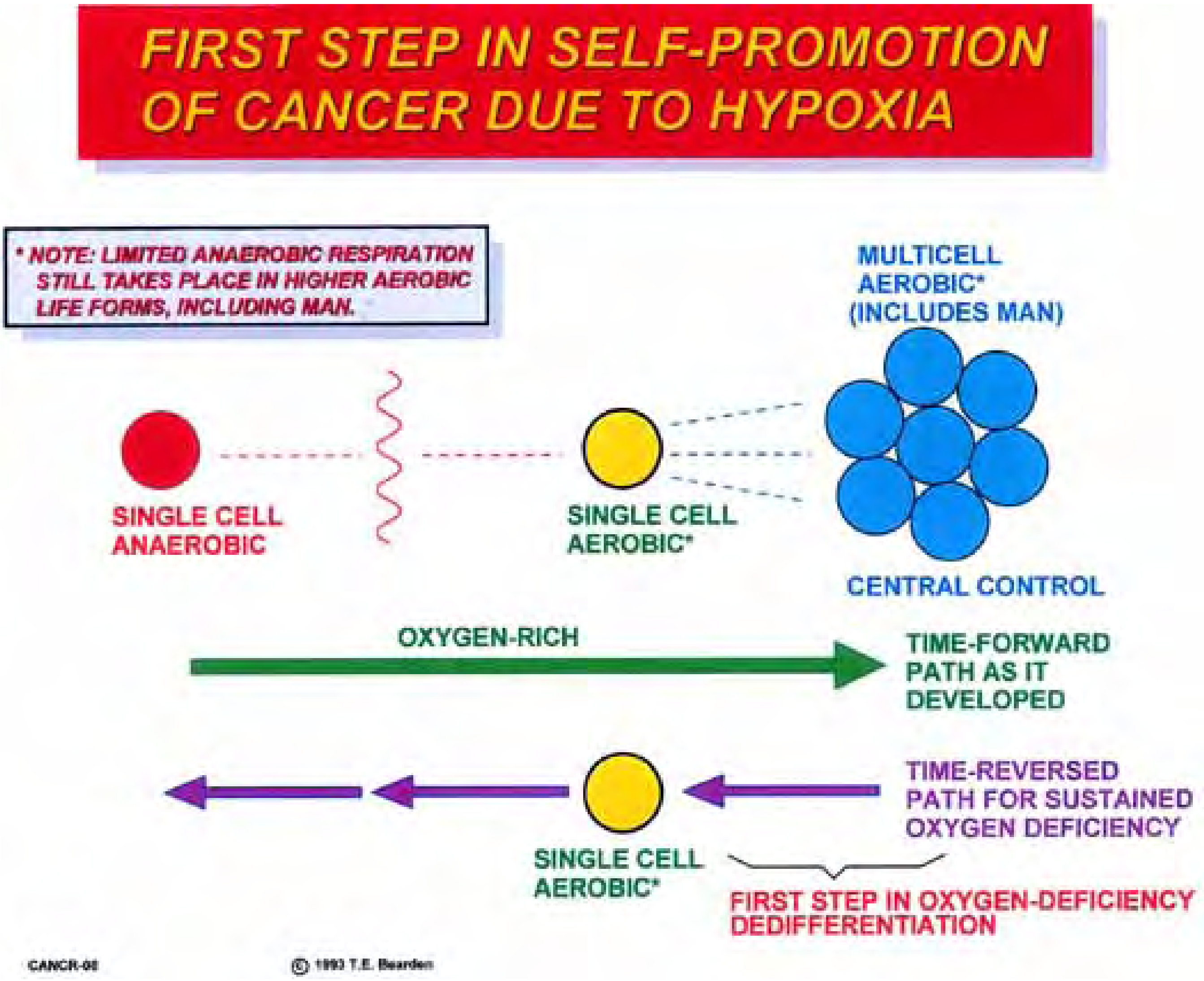

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC*

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD PATH AS IT DEVELOPED

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

SINGLE CELL AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

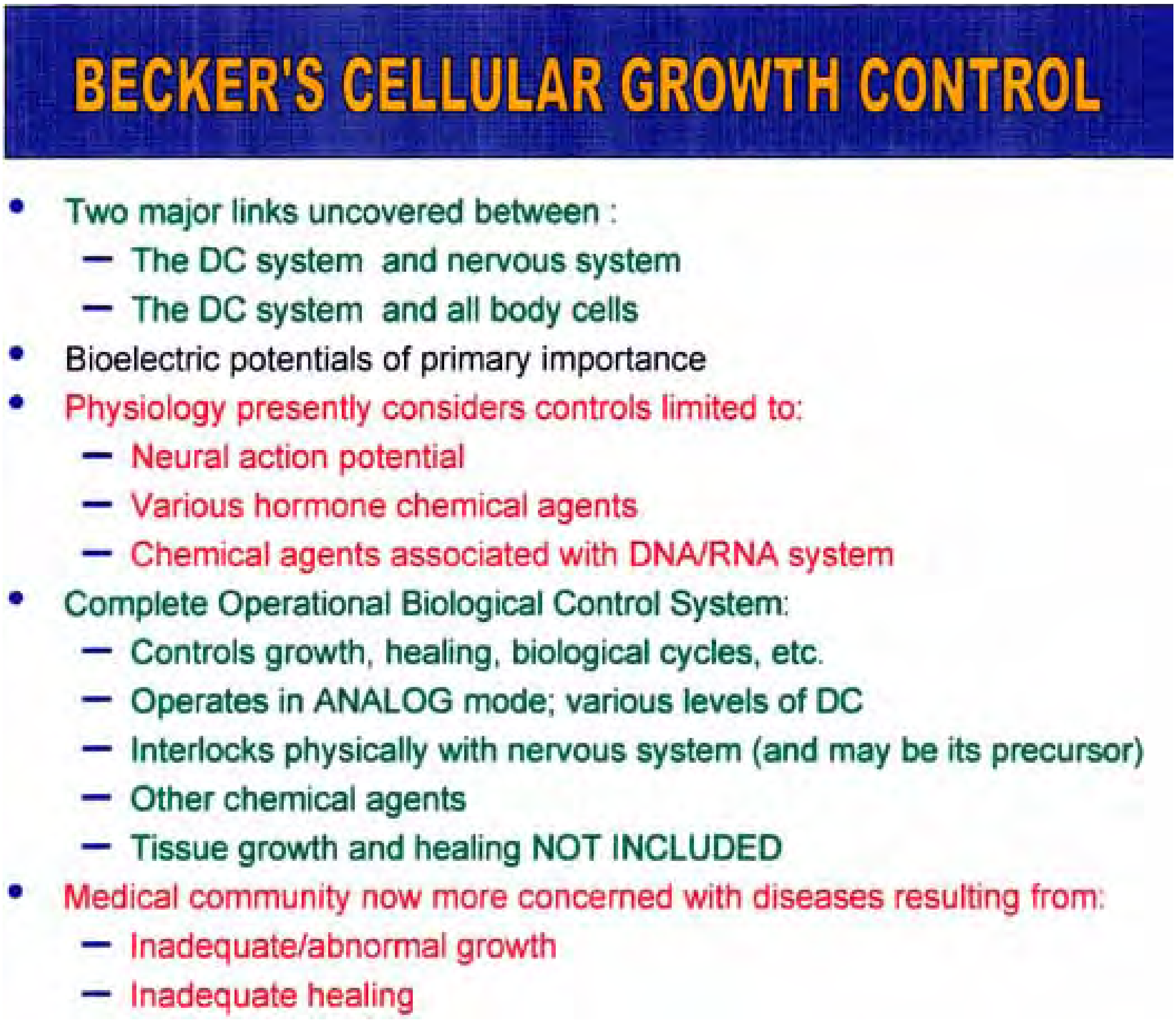

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

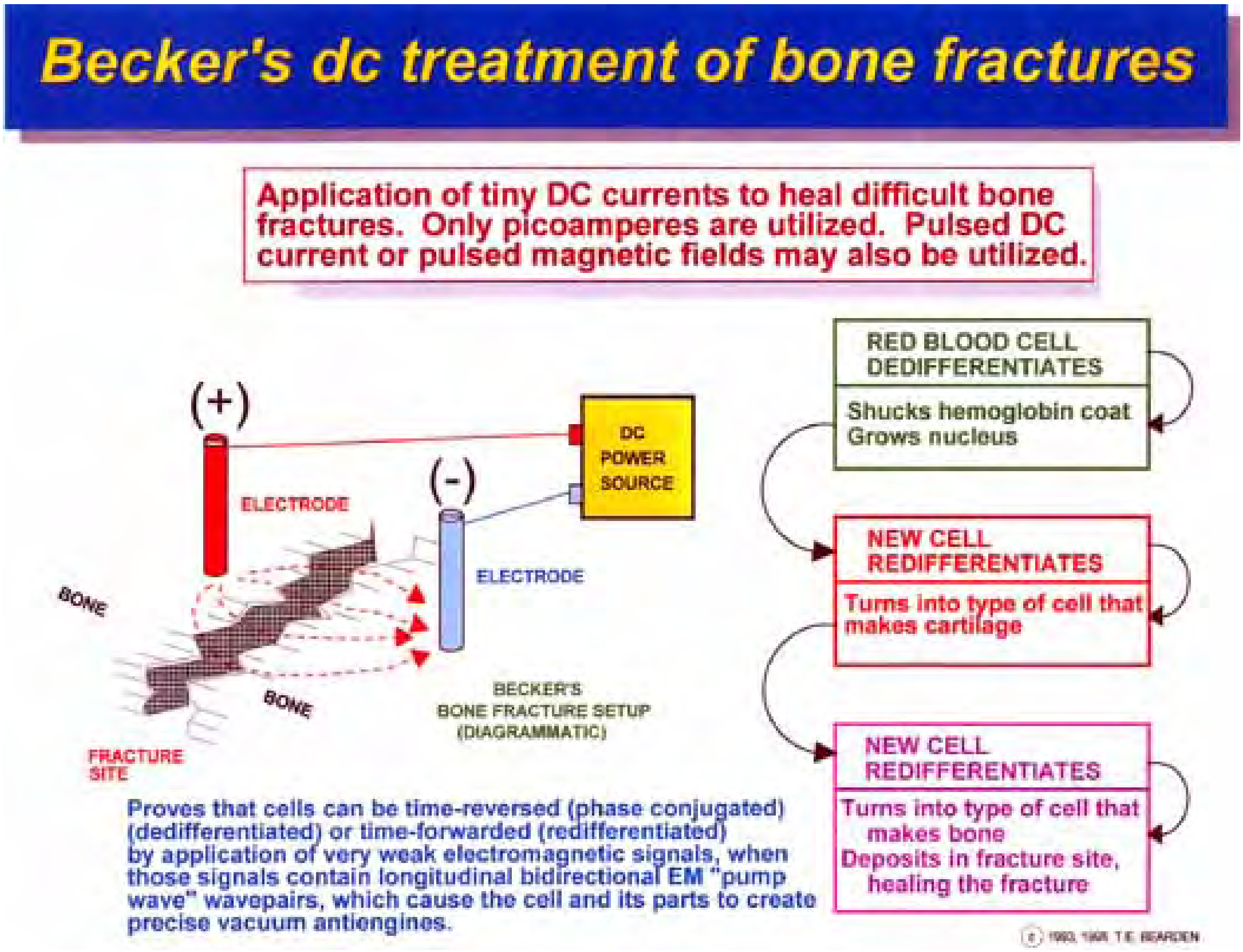

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

BECKER'S
BONE FRACTURE SETUP
(DIAGRAMMATIC)

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

(c) 1993, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

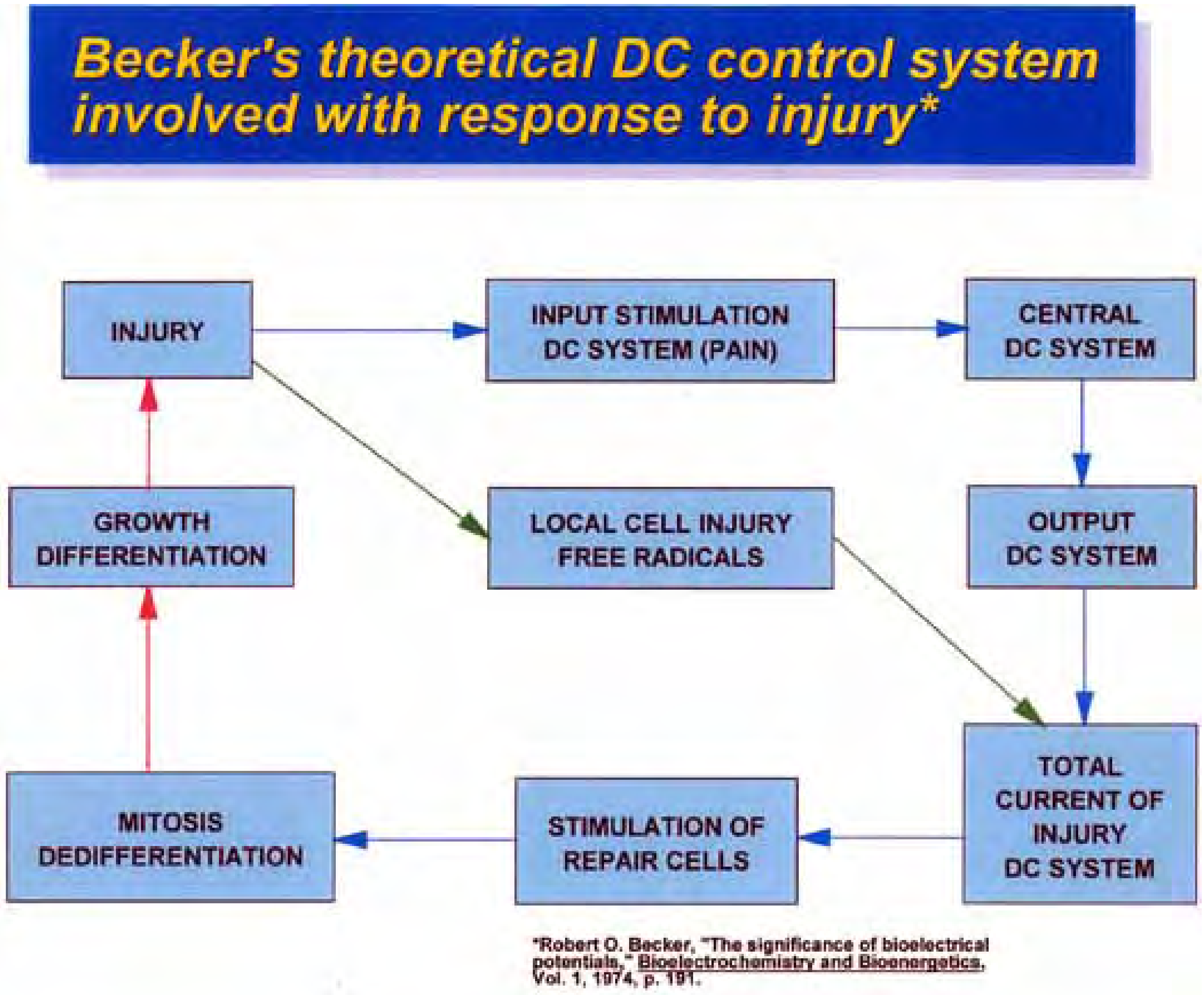

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Return to Slide Index

Next Slide

Previous Slide

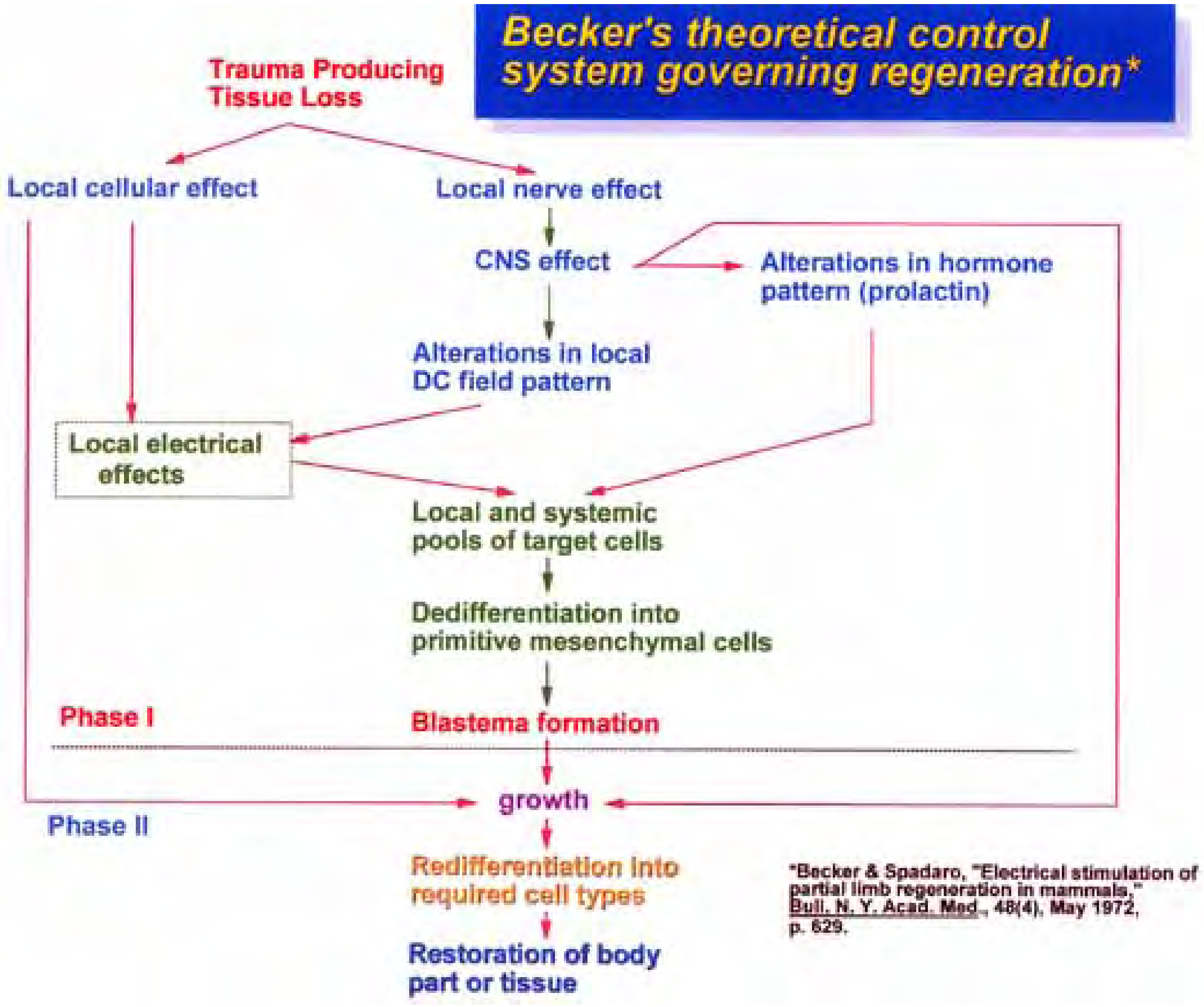

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Return to Slide Index

Next Slide

Previous Slide

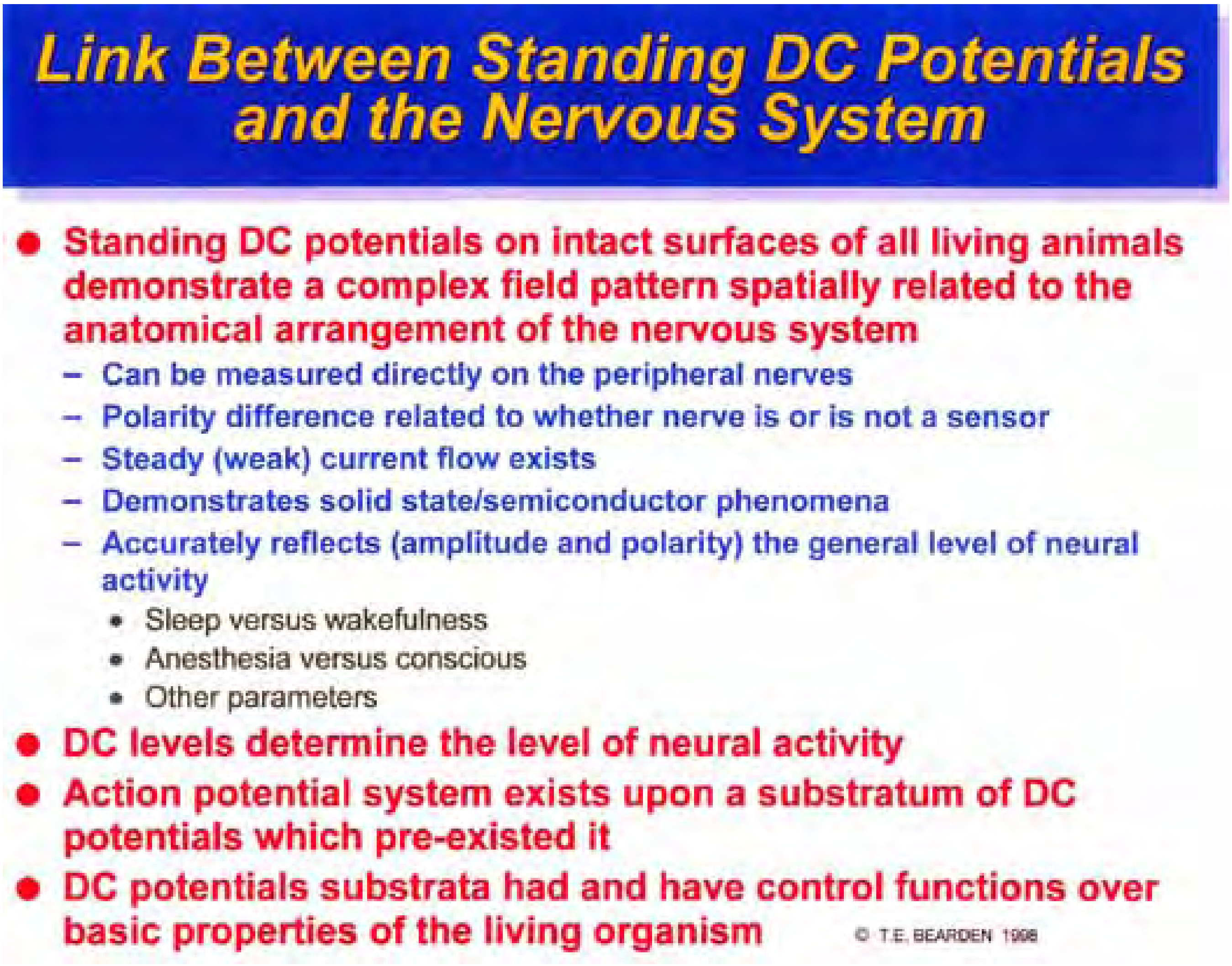

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

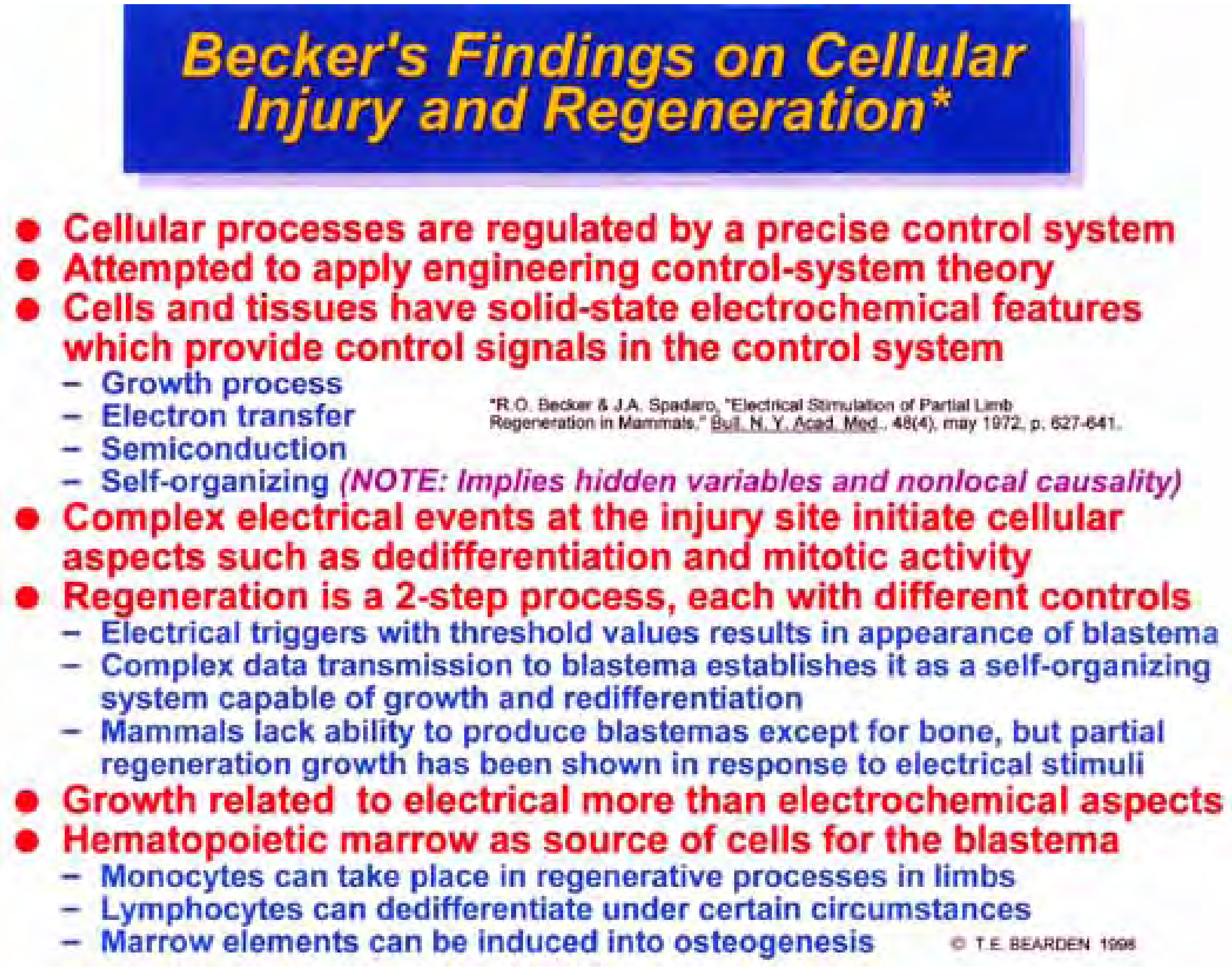

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

Corrections For the Present EM Bioeffects Model

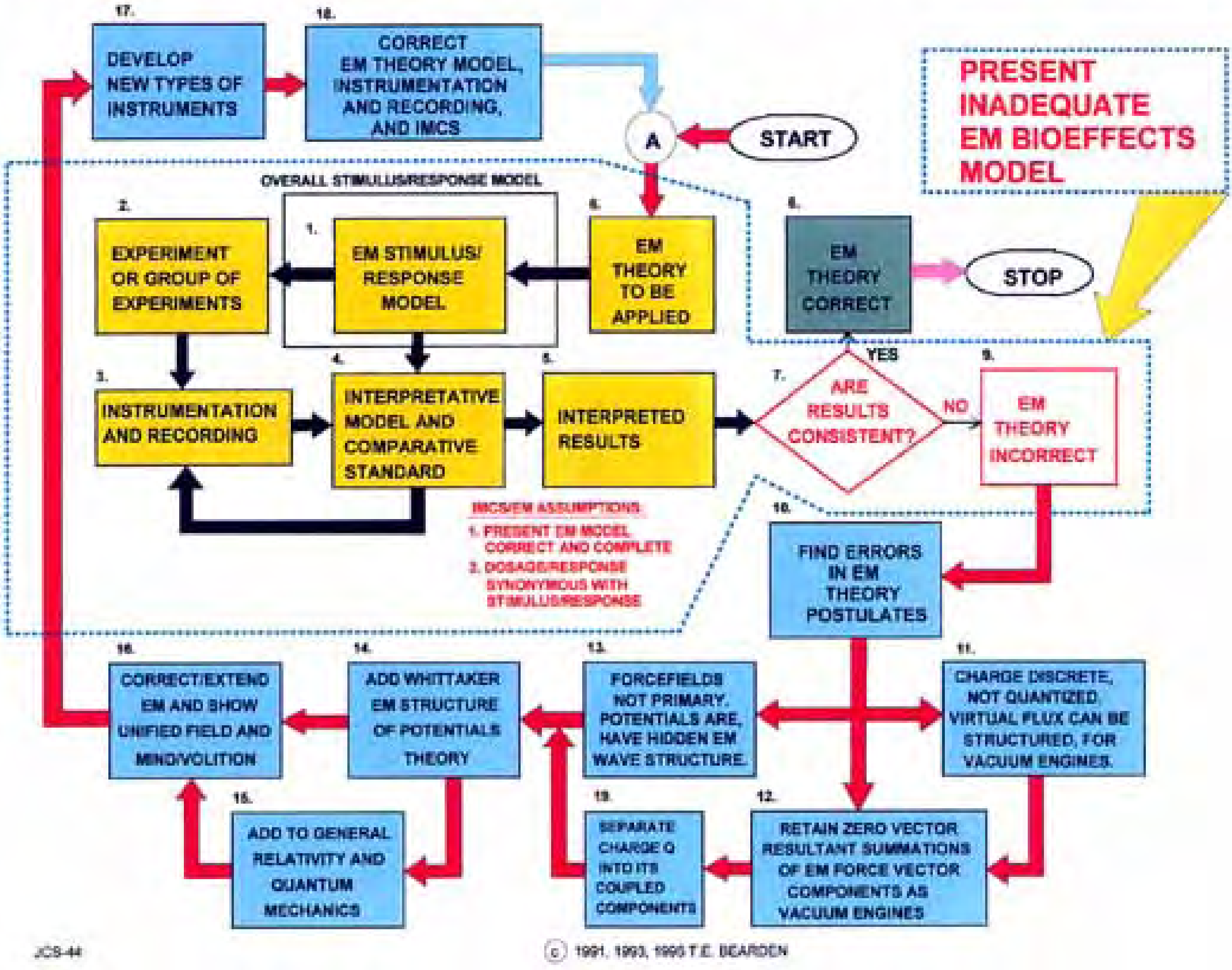

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

A

START

PRESENT INADEQUATE EM BIOEFFECTS MODEL

OVERALL STIMULUS/RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

1. EM STIMULUS/ RESPONSE MODEL

6. EM THEORY TO BE APPLIED

8. EM THEORY CORRECT

STOP

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

7. ARE RESULTS CONSISTENT?

YES

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

JCS-44

(c) 1991, 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

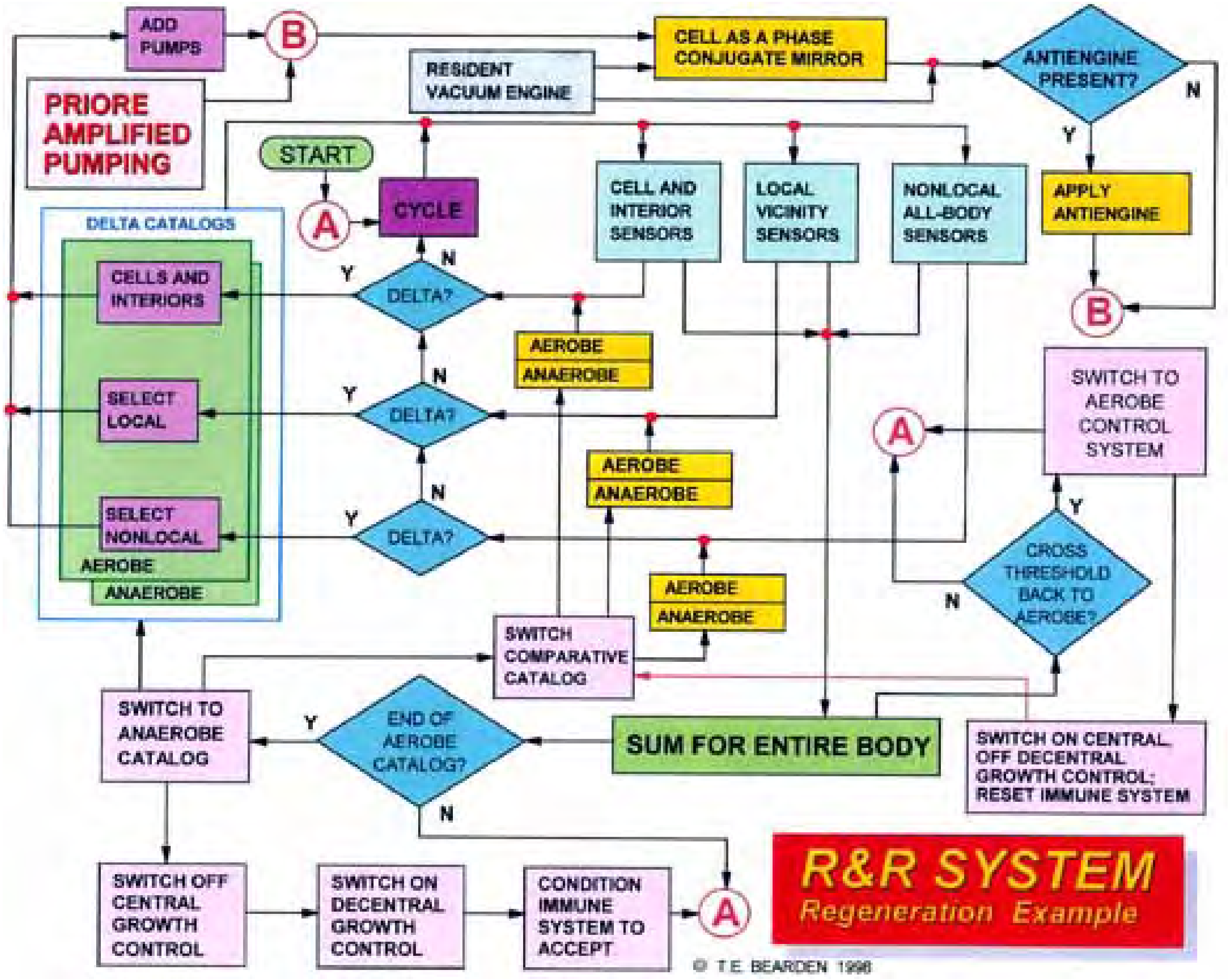

PRIORE AMPLIFIED PUMPING

ADD PUMPS

CELL AS A PHASE CONJUGATE MIRROR

RESIDENT VACUUM ENGINE

ANTIENGINE PRESENT?

START

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

APPLY ANTIENGINE

DELTA CATALOGS

CELLS AND INTERIORS

SELECT LOCAL

SELECT NONLOCAL

AEROBE

ANAEROBE

DELTA?

DELTA?

DELTA?

AEROBE ANAEROBE

AEROBE ANAEROBE

AEROBE ANAEROBE

SWITCH TO AEROBE CONTROL SYSTEM

CROSS THRESHOLD BACK TO AEROBE?

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

END OF AEROBE CATALOG?

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

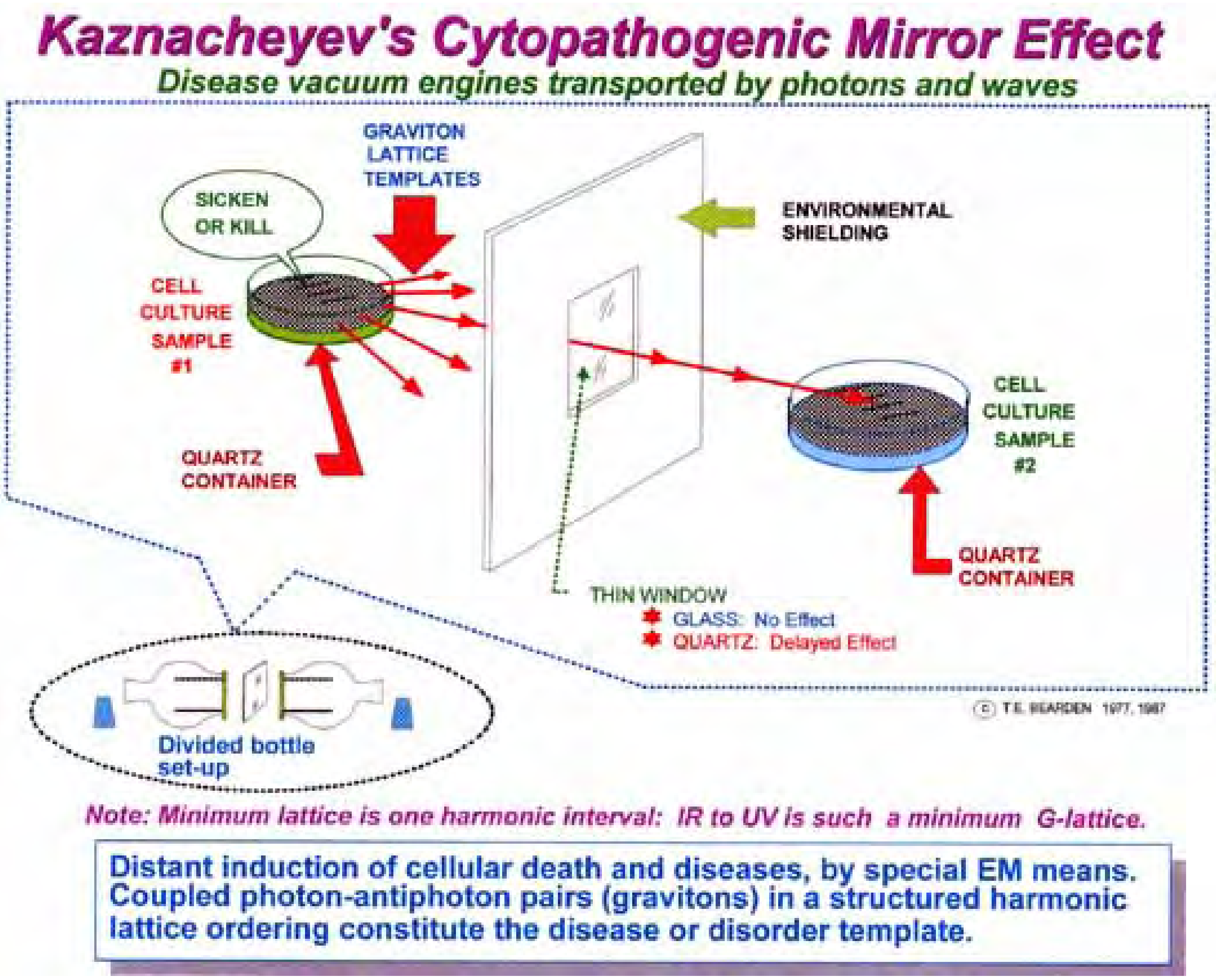

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
ENVIRONMENTAL SHIELDING
CELL CULTURE SAMPLE #1
QUARTZ CONTAINER
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Return to Slide Index

Next Slide

Previous Slide

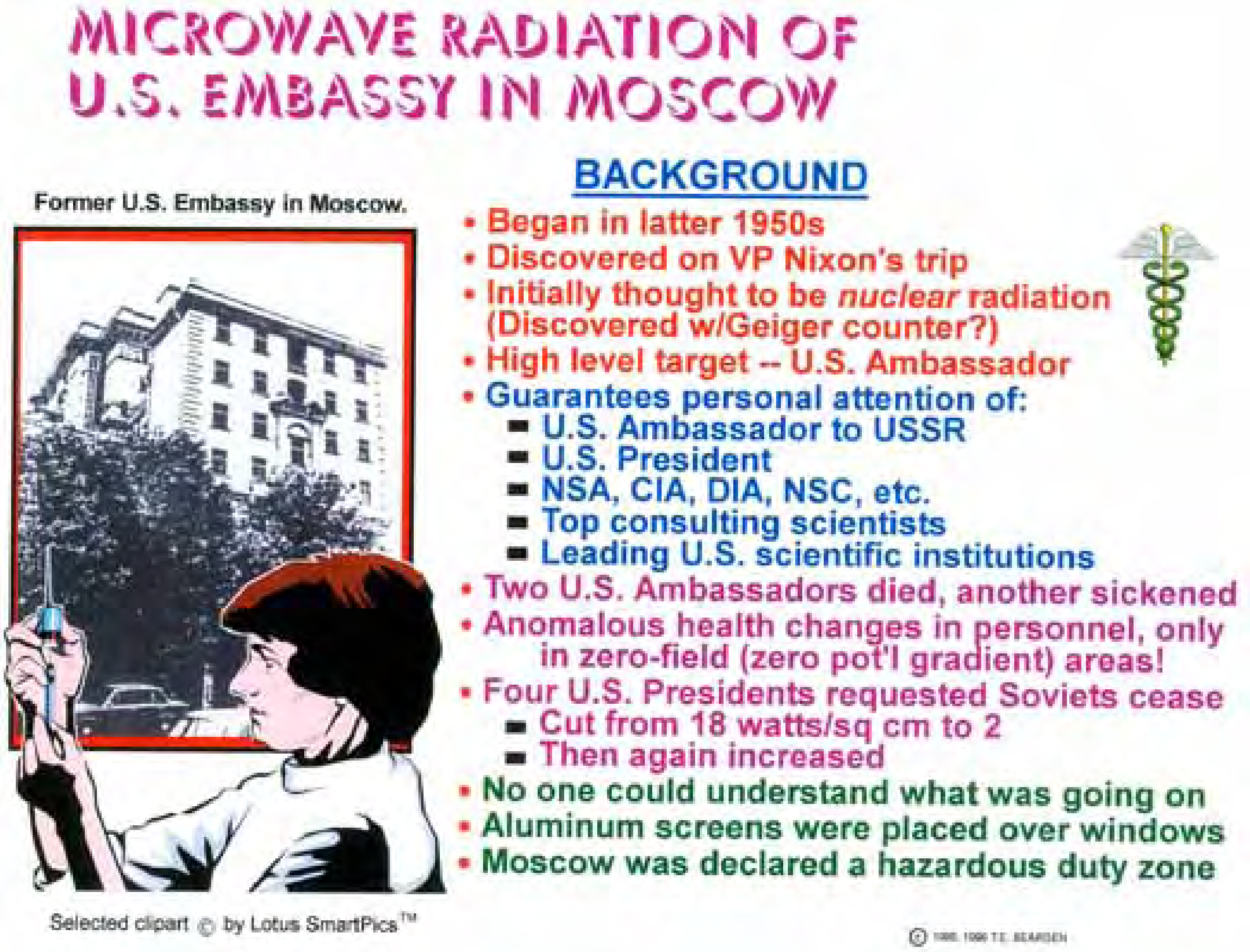

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© [illegible] T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

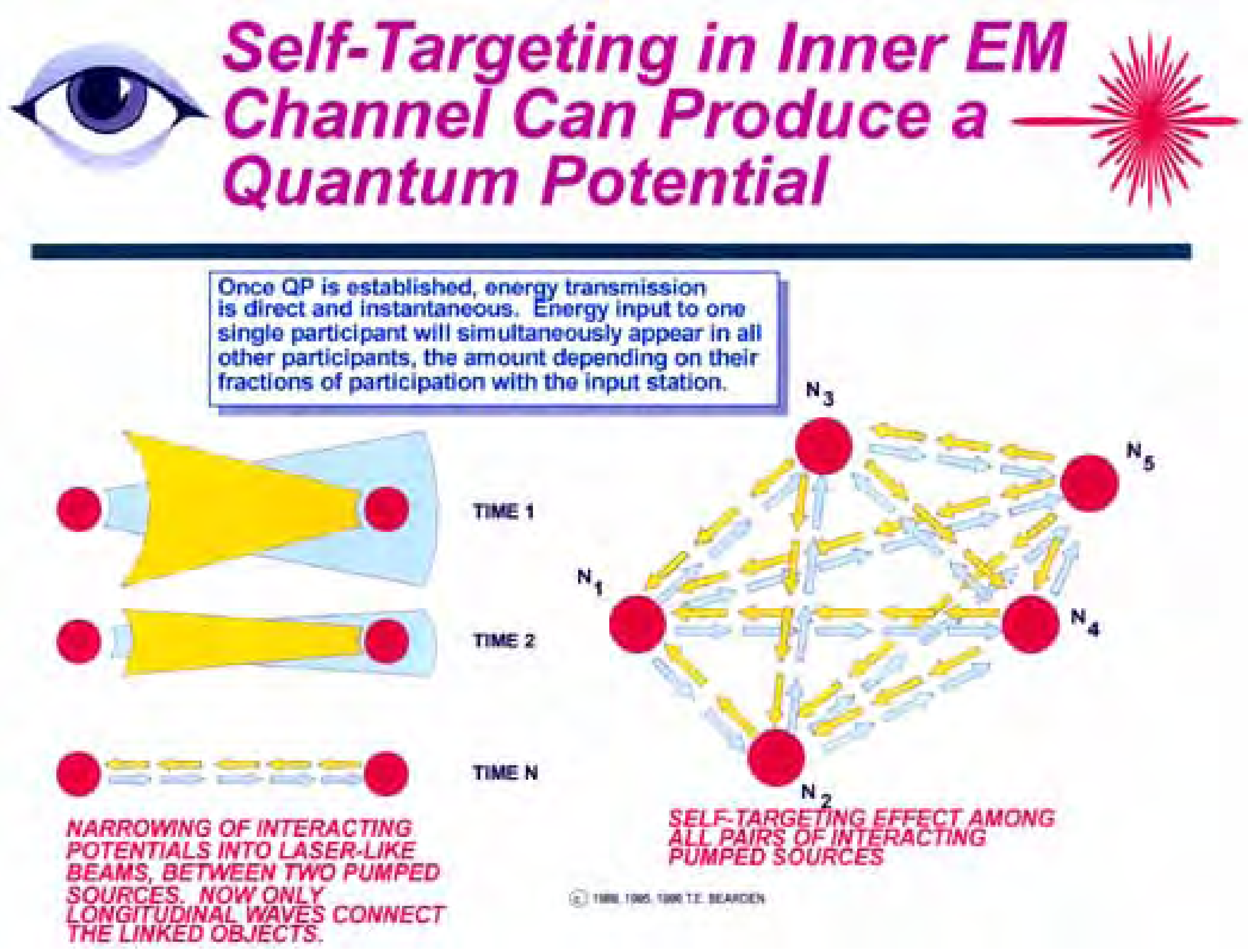

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1986, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

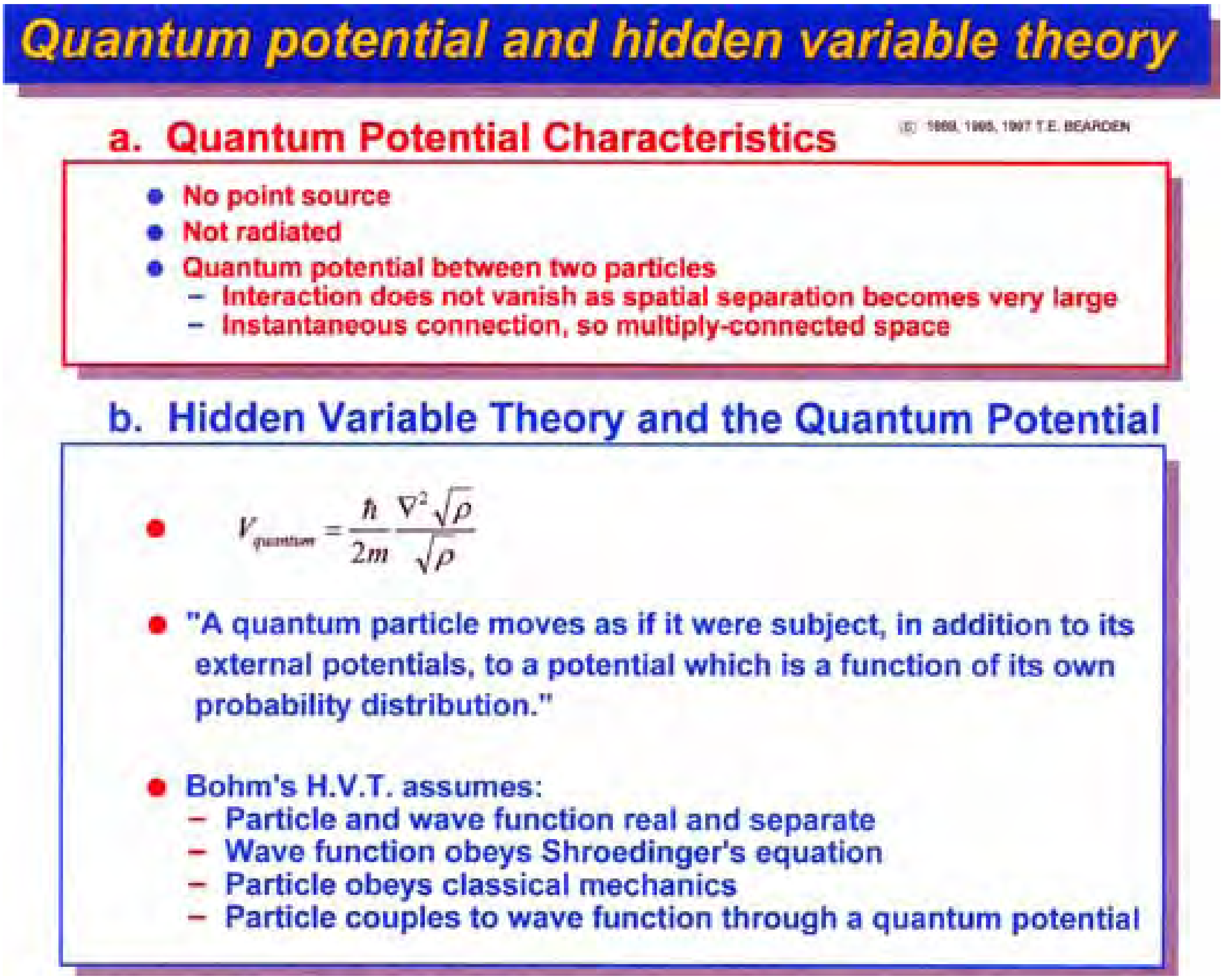

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1988, 1995, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $$V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

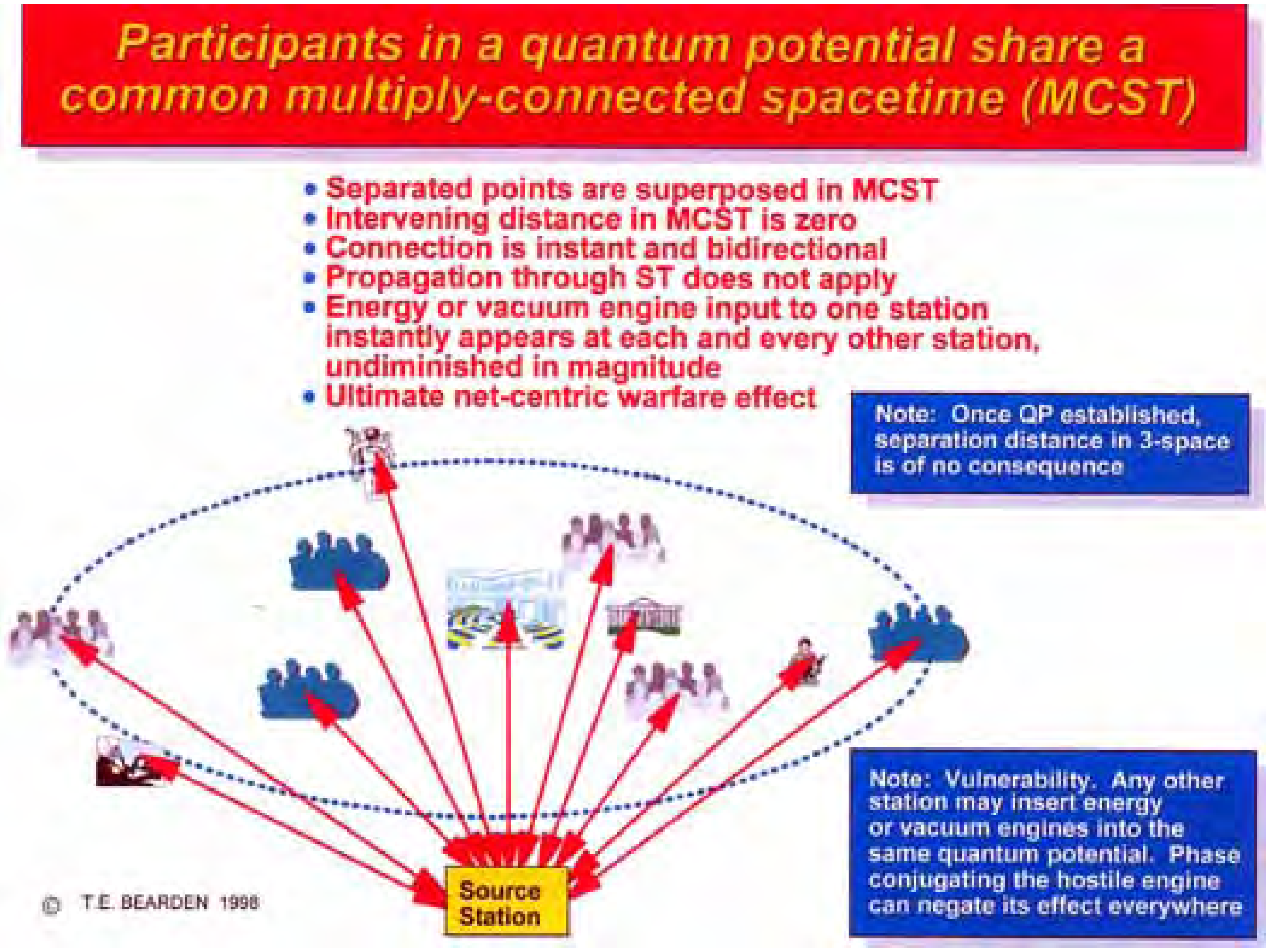

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

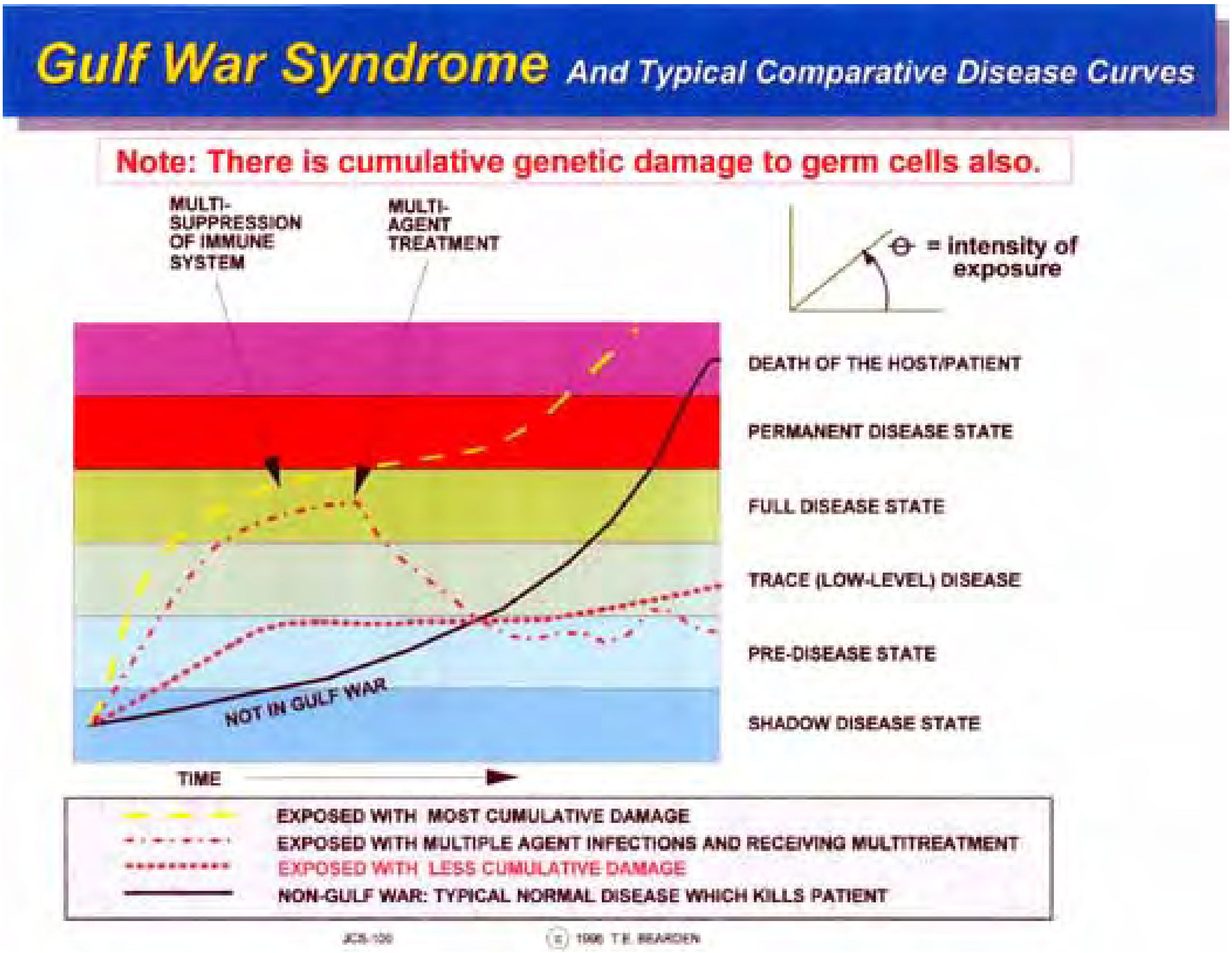

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

EXPOSED WITH MOST CUMULATIVE DAMAGE
EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT
EXPOSED WITH LESS CUMULATIVE DAMAGE
NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

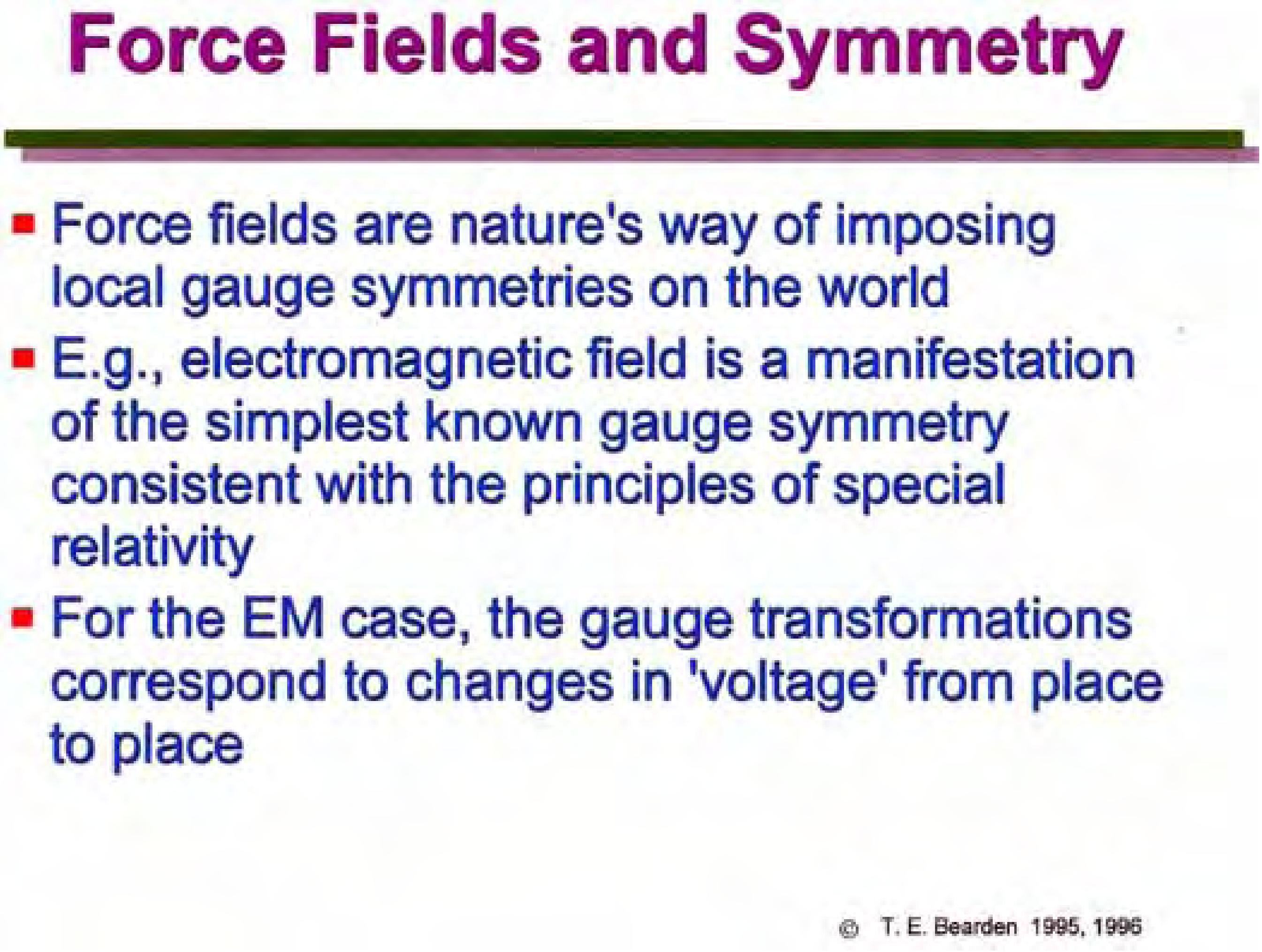

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

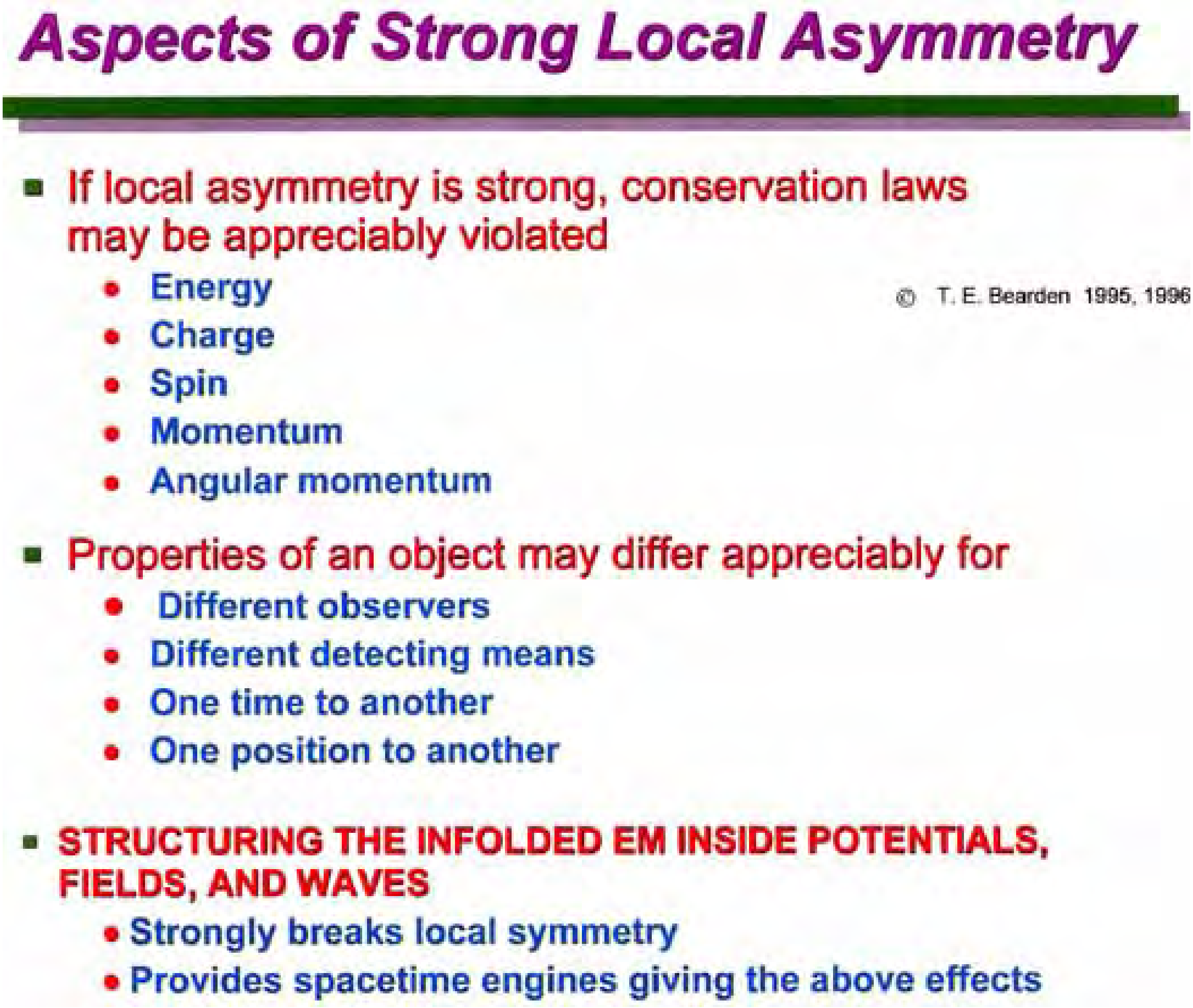

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

Return to Slide Index

Next Slide

Previous Slide

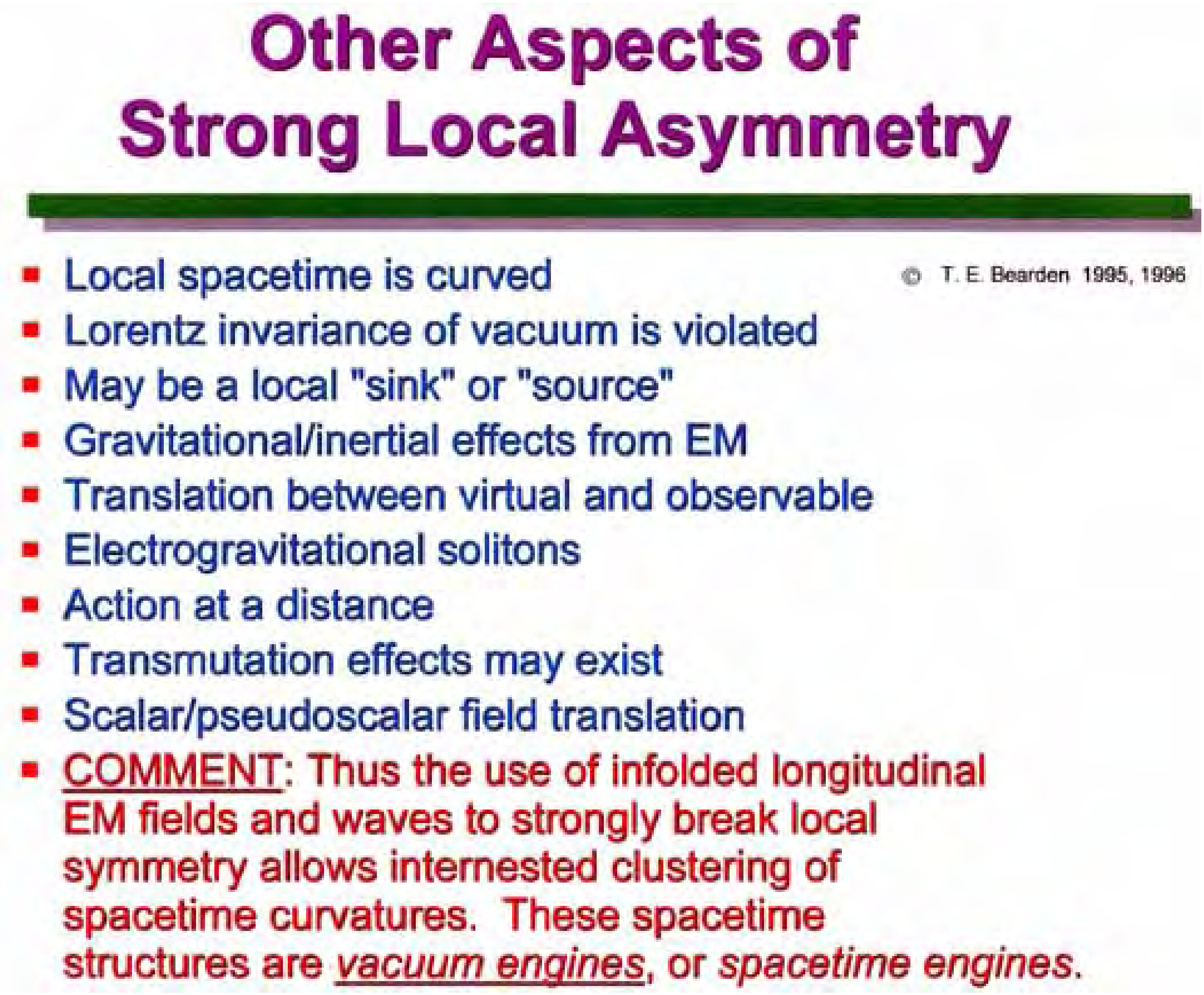

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

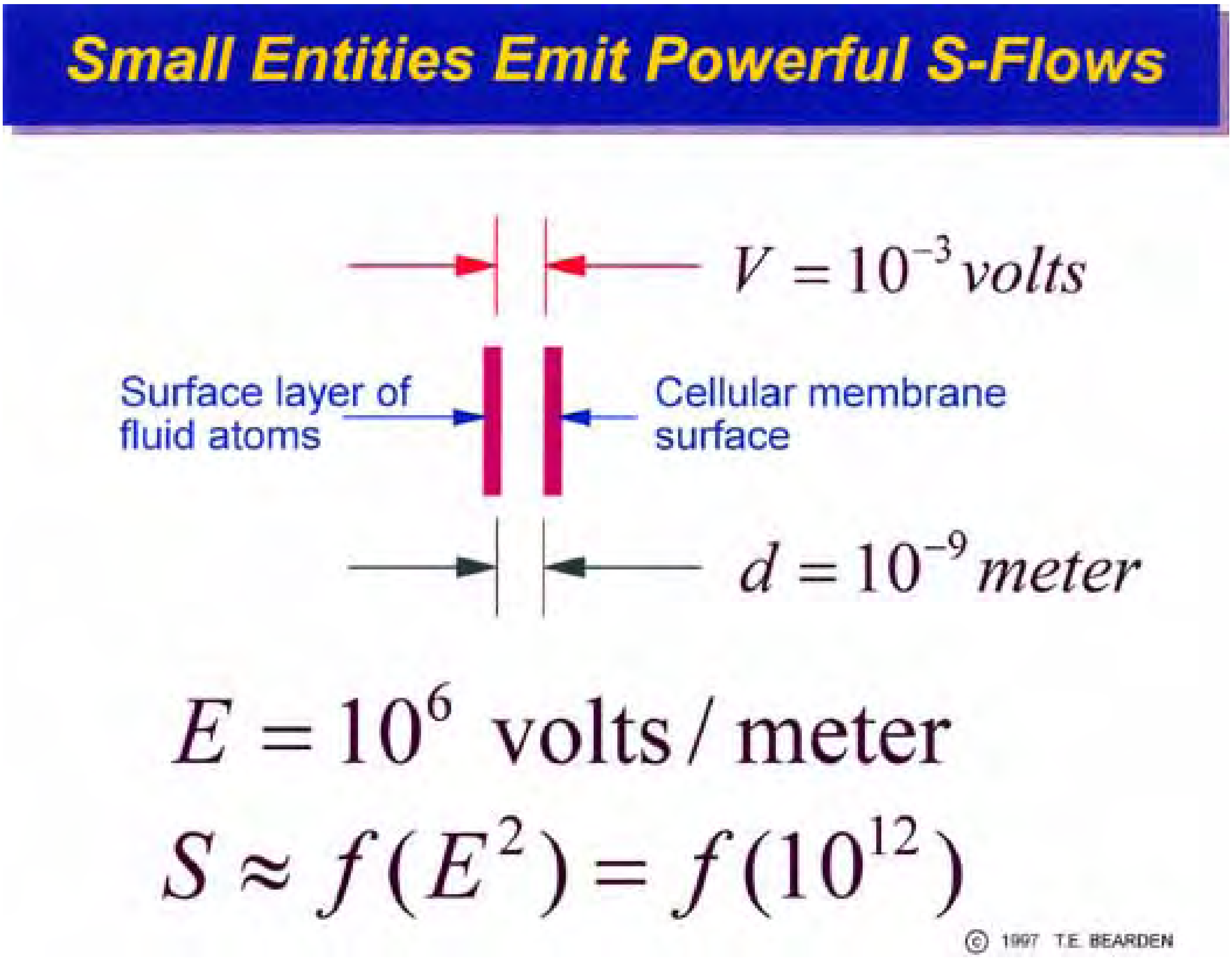

Small Entities Emit Powerful S-Flows

$V = 10^{-3}\ volts$

Surface layer of fluid atoms

Cellular membrane surface

$d = 10^{-9}\ meter$

$$E = 10^{6}\ \text{volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

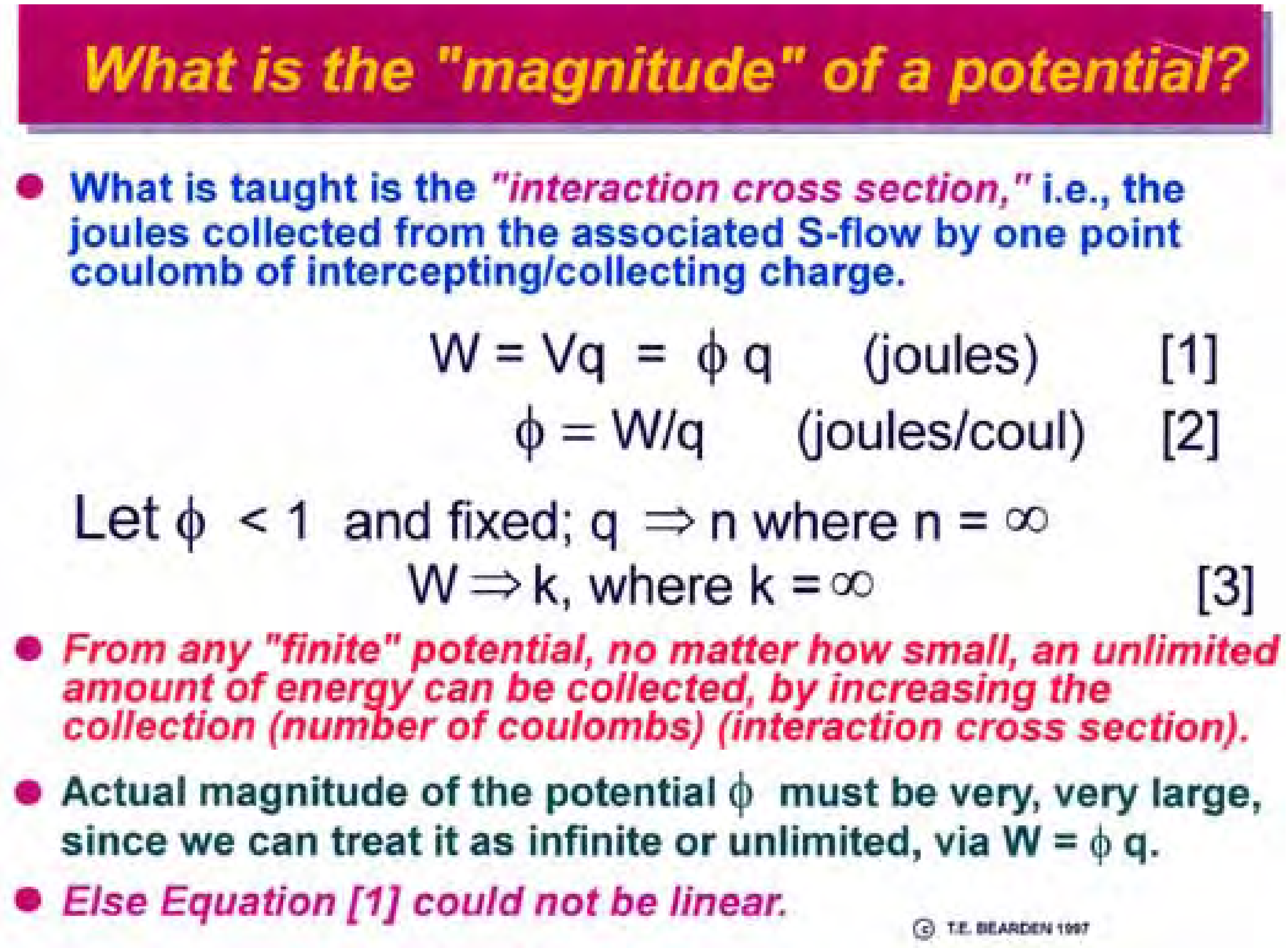

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \quad \text{(joules)} \quad [1]$$

$$\phi = W/q \quad \text{(joules/coul)} \quad [2]$$

Let $\phi < 1$ and fixed; $q \Rightarrow n$ where $n = \infty$

$$W \Rightarrow k, \text{ where } k = \infty \quad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

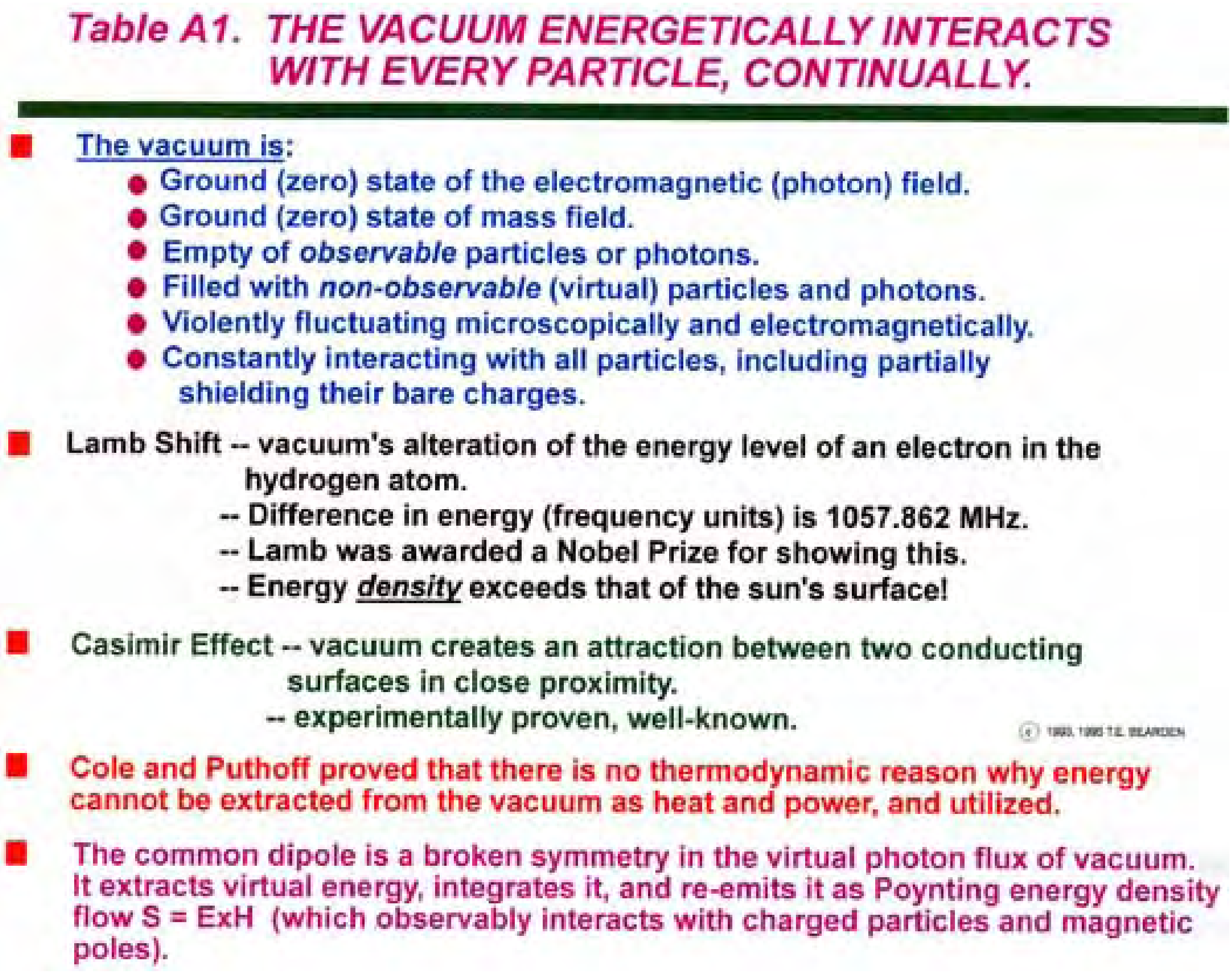

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy *density* exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.

© [illegible] T.E. BEARDEN

- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

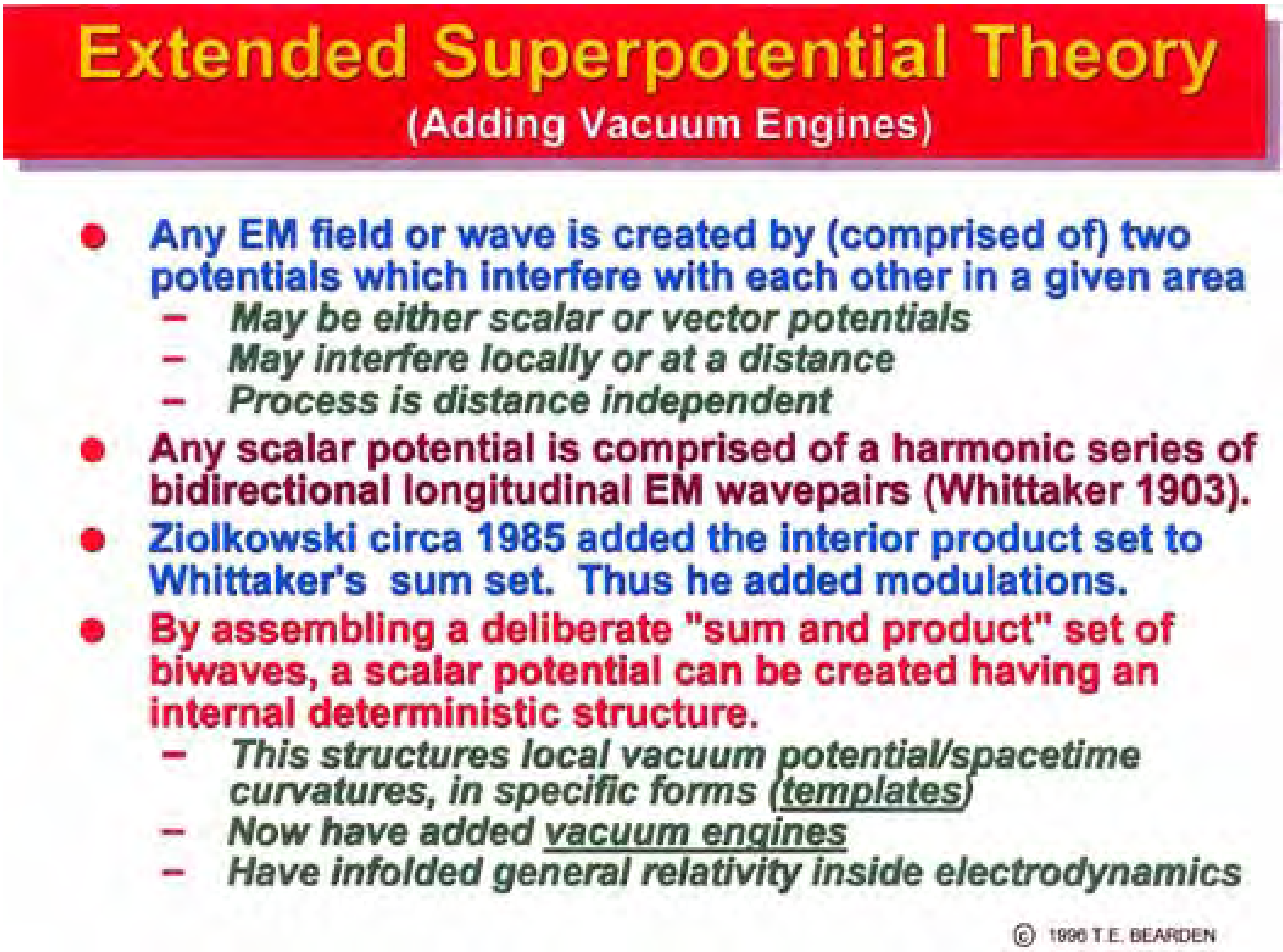

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

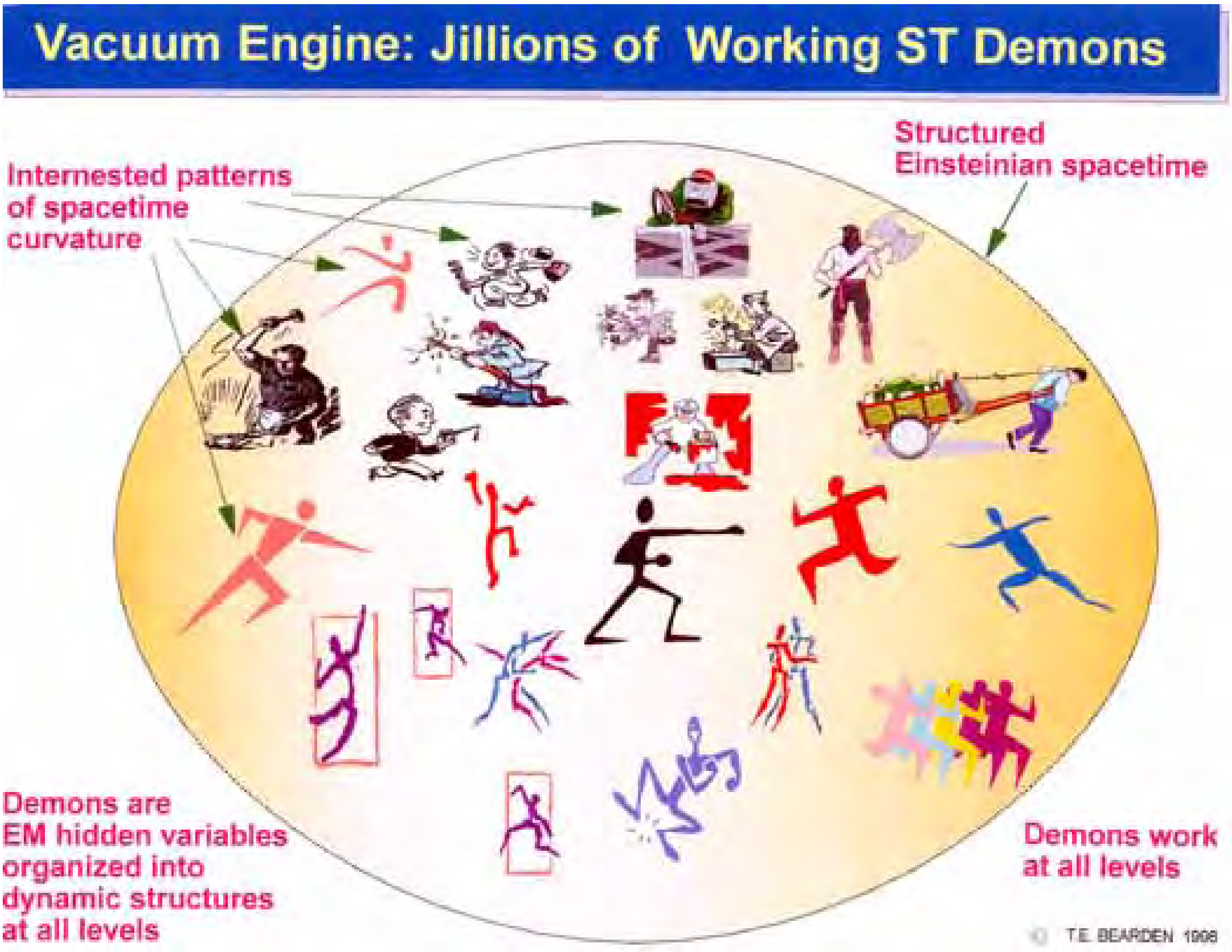

Vacuum Engine: Jillions of Working ST Demons

Internested patterns of spacetime curvature

Structured Einsteinian spacetime

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

[Return to Slide Index]

[Next Slide]

[Previous Slide]

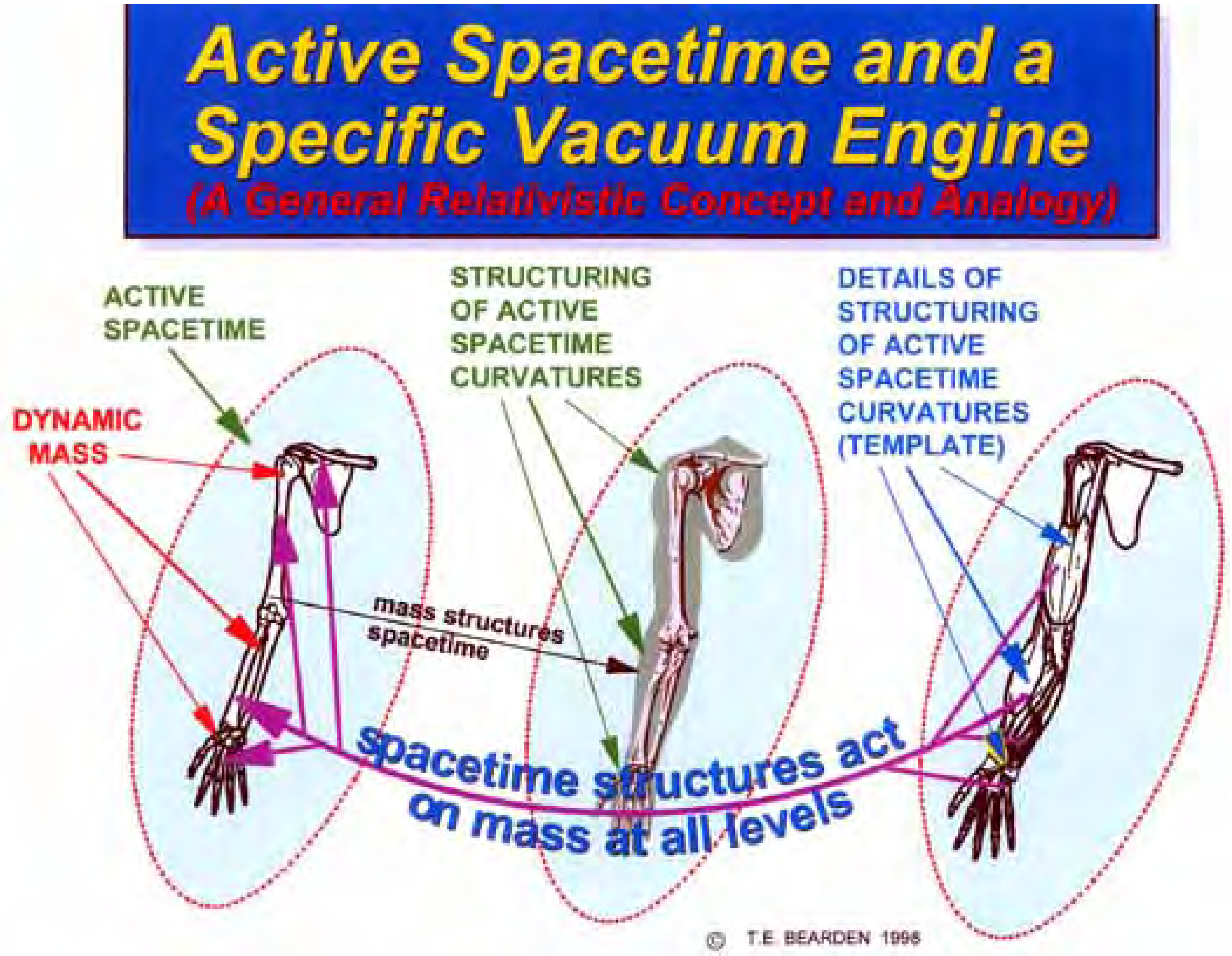

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME

DYNAMIC MASS

STRUCTURING OF ACTIVE SPACETIME CURVATURES

DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)

mass structures spacetime

spacetime structures act on mass at all levels

© T.E. BEARDEN 1998

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

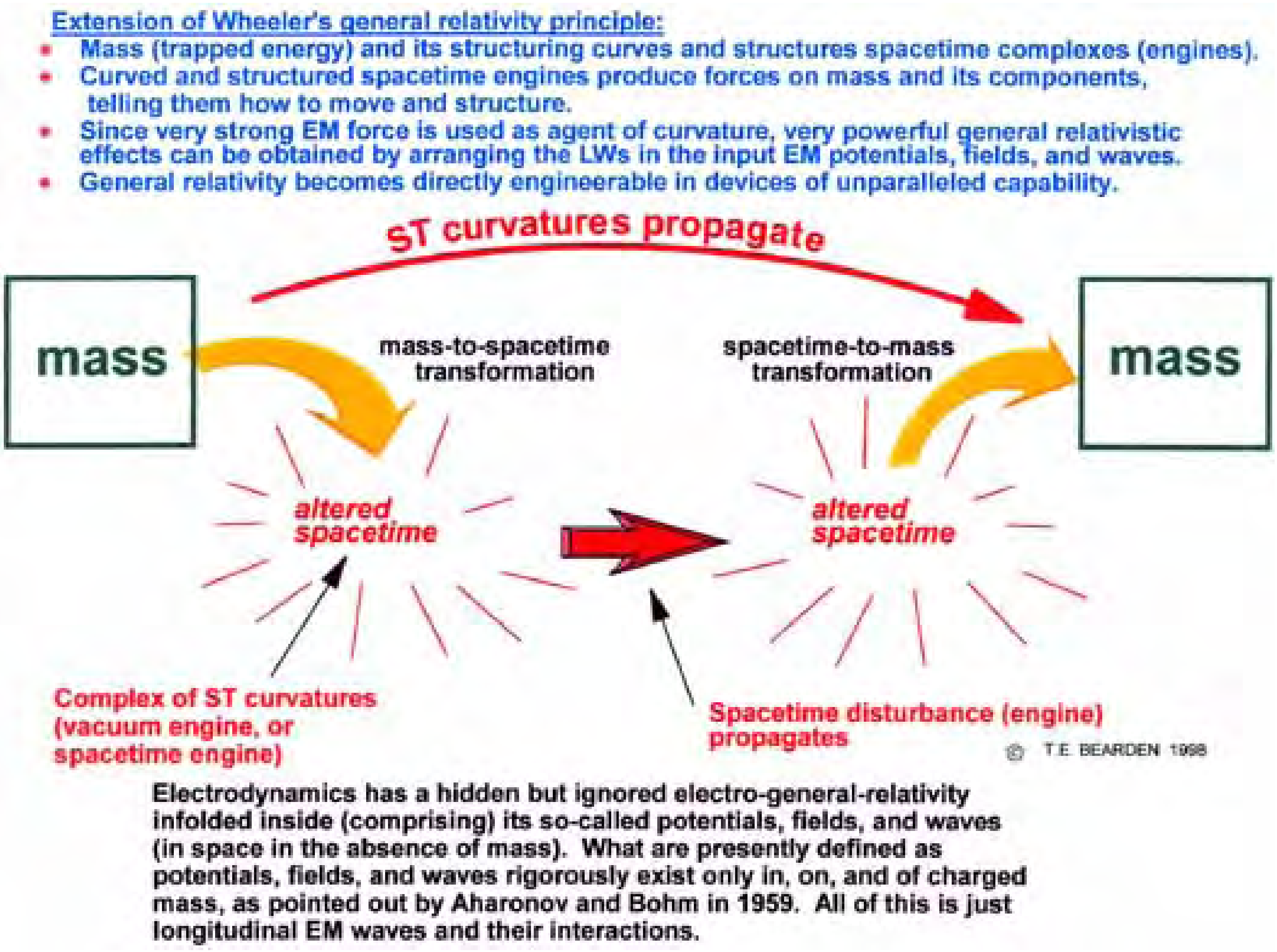

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

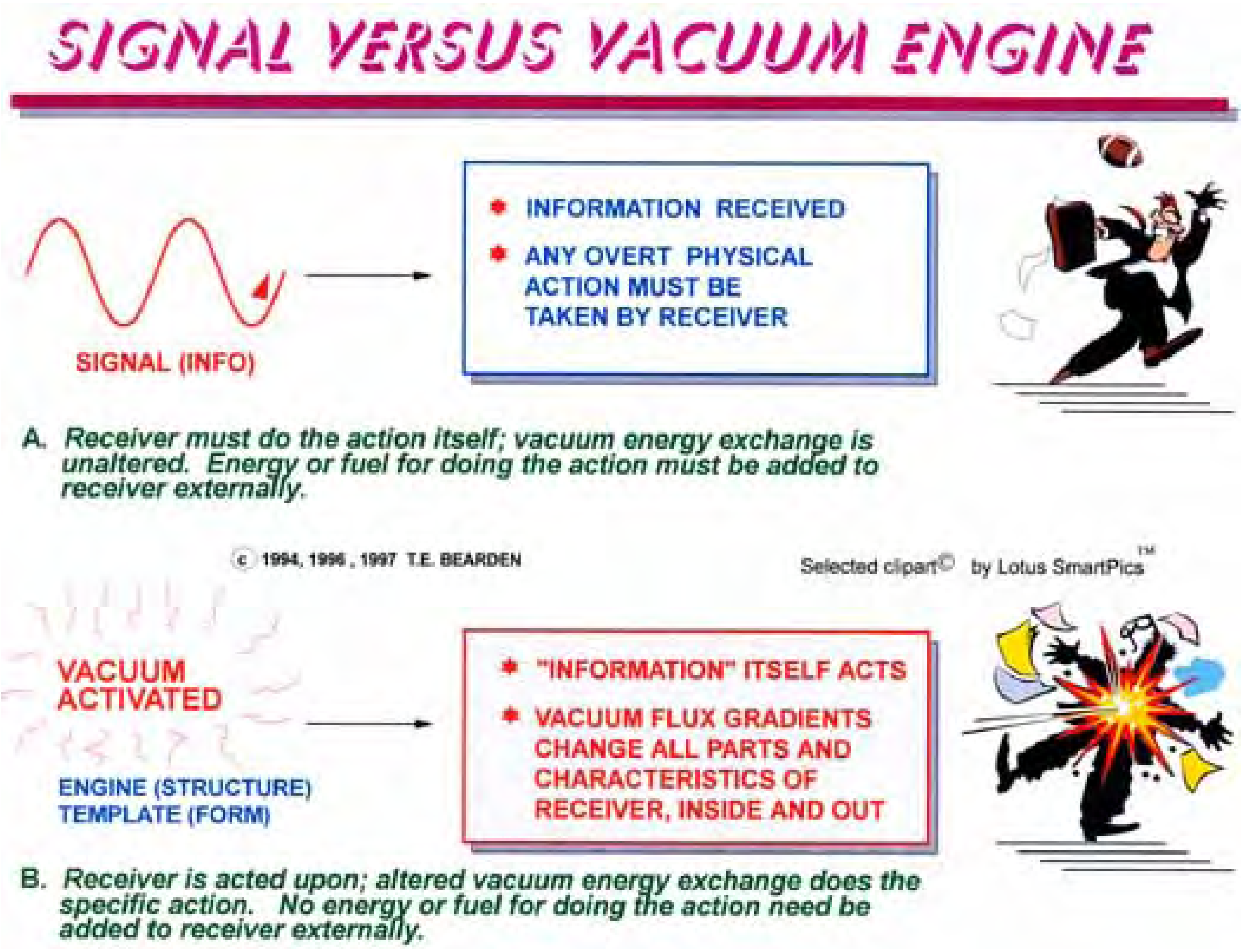

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

(c) 1994, 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED

ENGINE (STRUCTURE)
TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

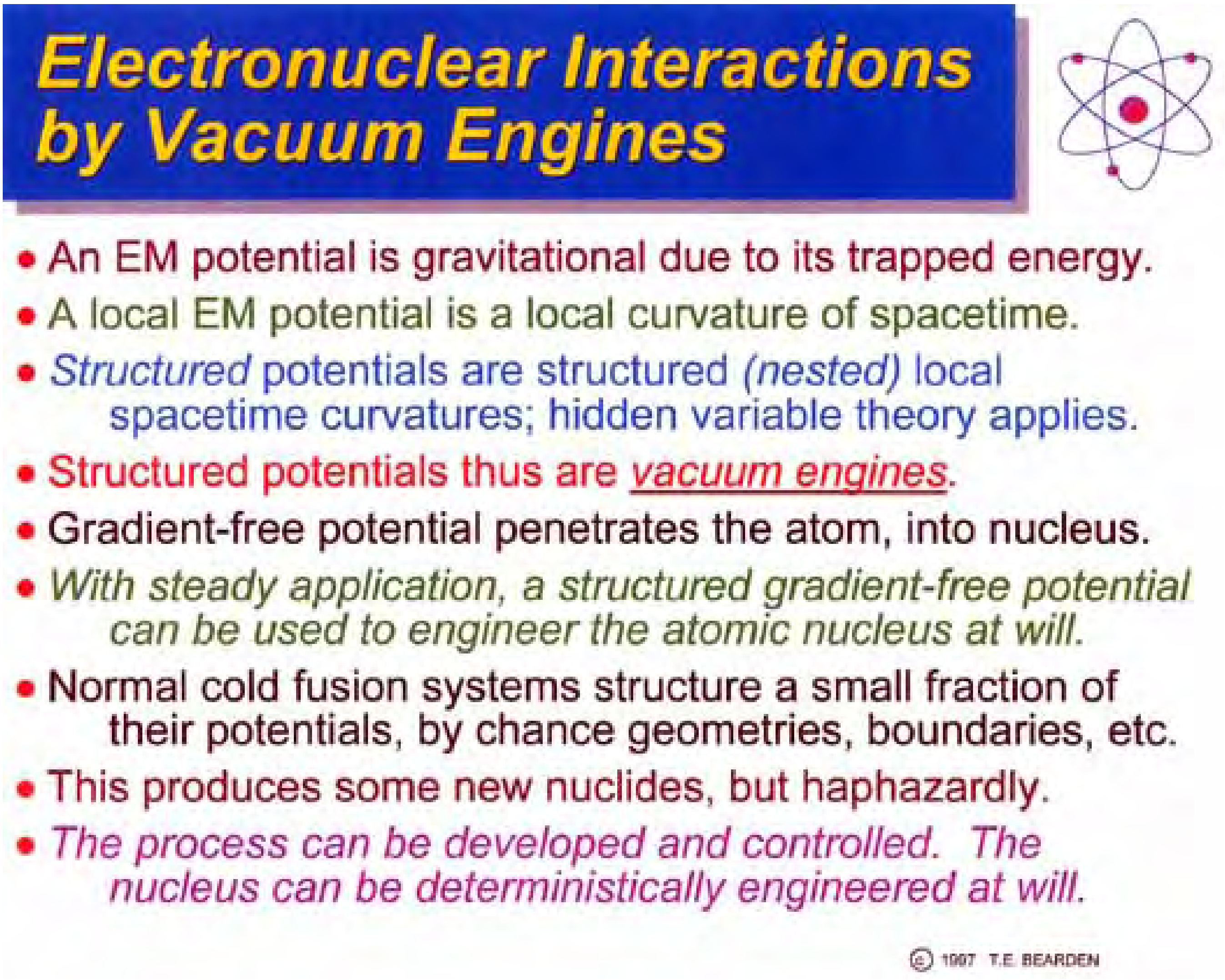

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

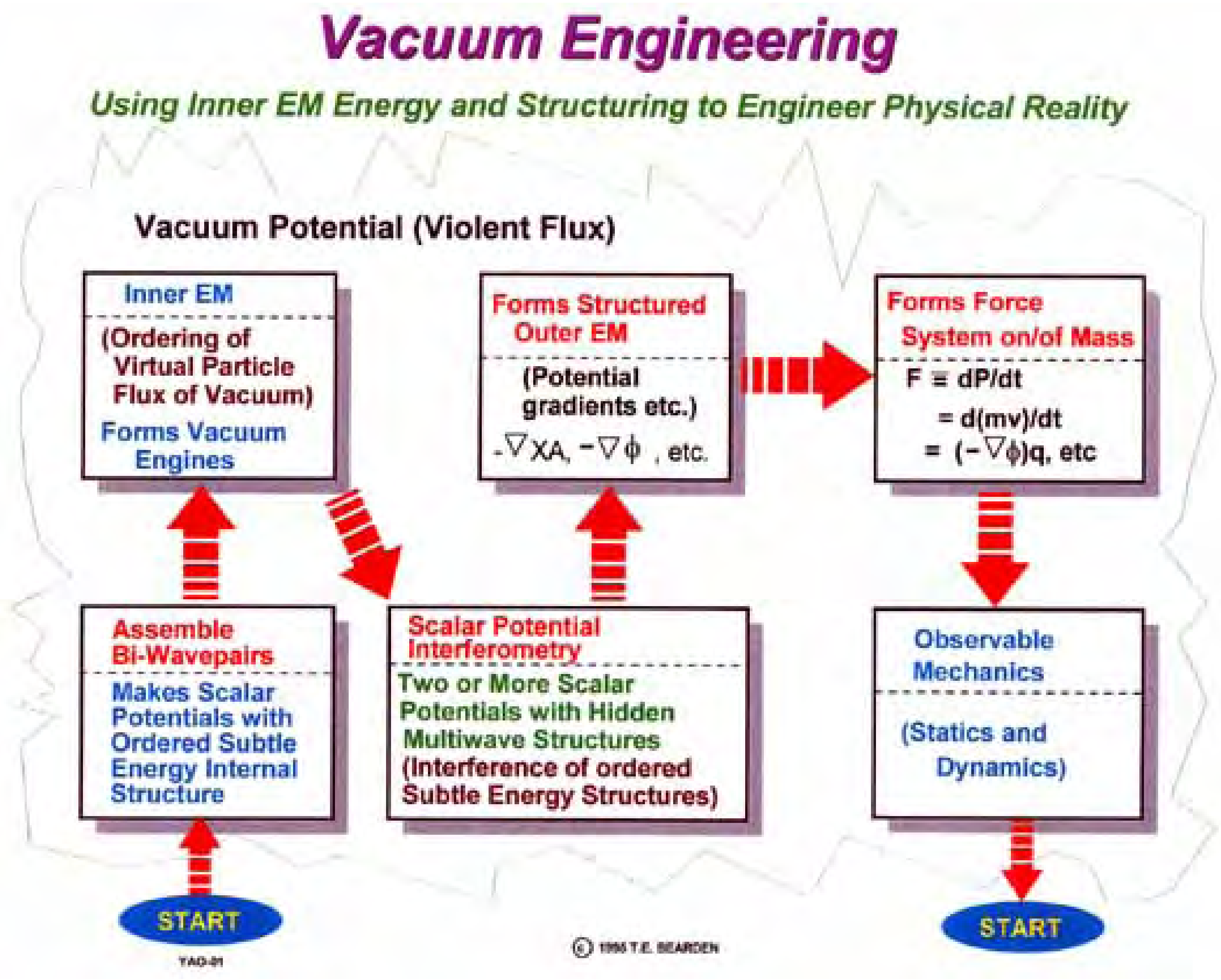

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM

(Ordering of Virtual Particle Flux of Vacuum)

Forms Vacuum Engines

Forms Structured Outer EM

(Potential gradients etc.)

$-\nabla \times A$, $-\nabla \phi$, etc.

Forms Force System on/of Mass

$F \equiv dP/dt$

$= d(mv)/dt$

$= (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs

Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry

Two or More Scalar Potentials with Hidden Multiwave Structures

(Interference of ordered Subtle Energy Structures)

Observable Mechanics

(Statics and Dynamics)

START

START

YAO-01

© 1996 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

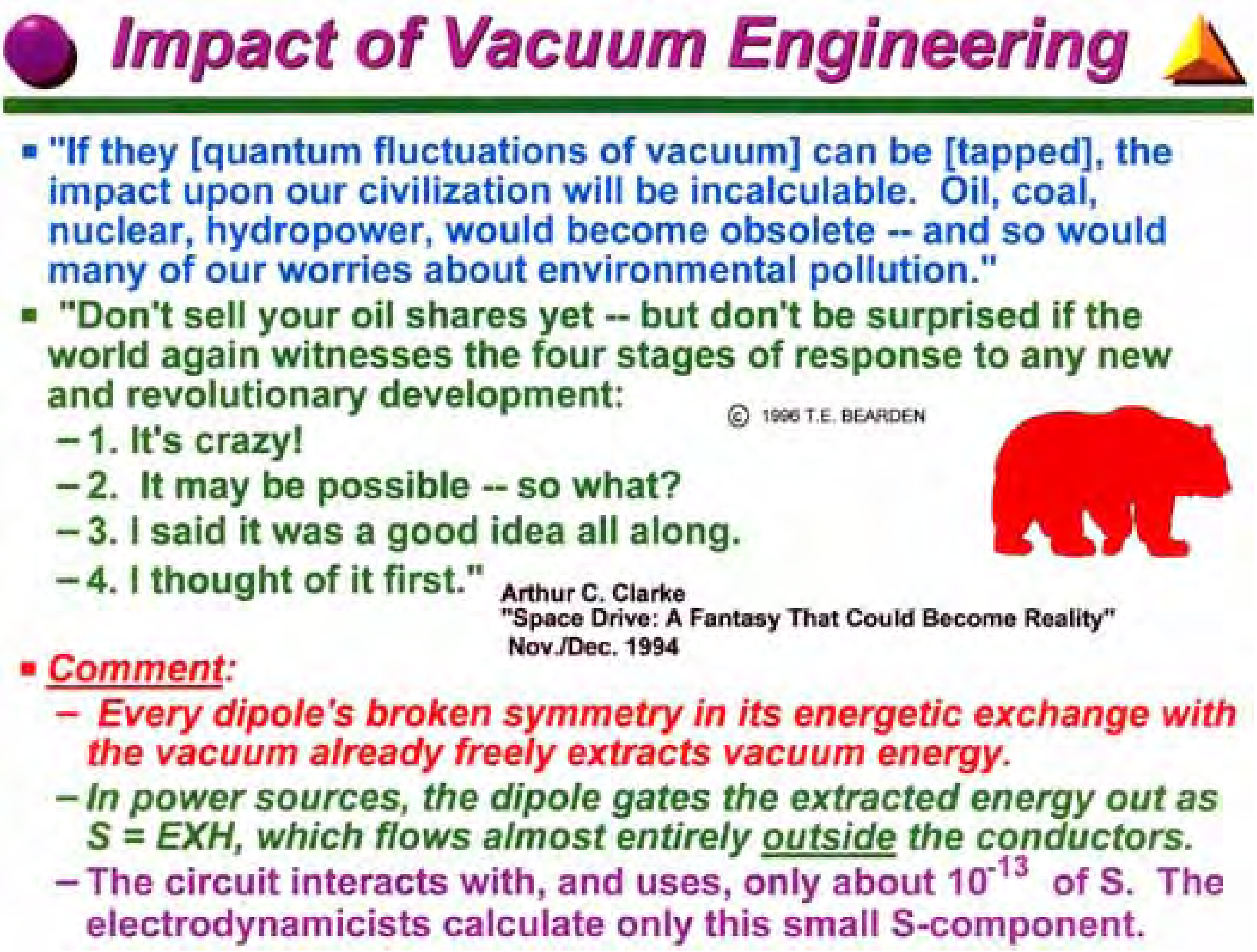

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

© 1996 T.E. BEARDEN

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

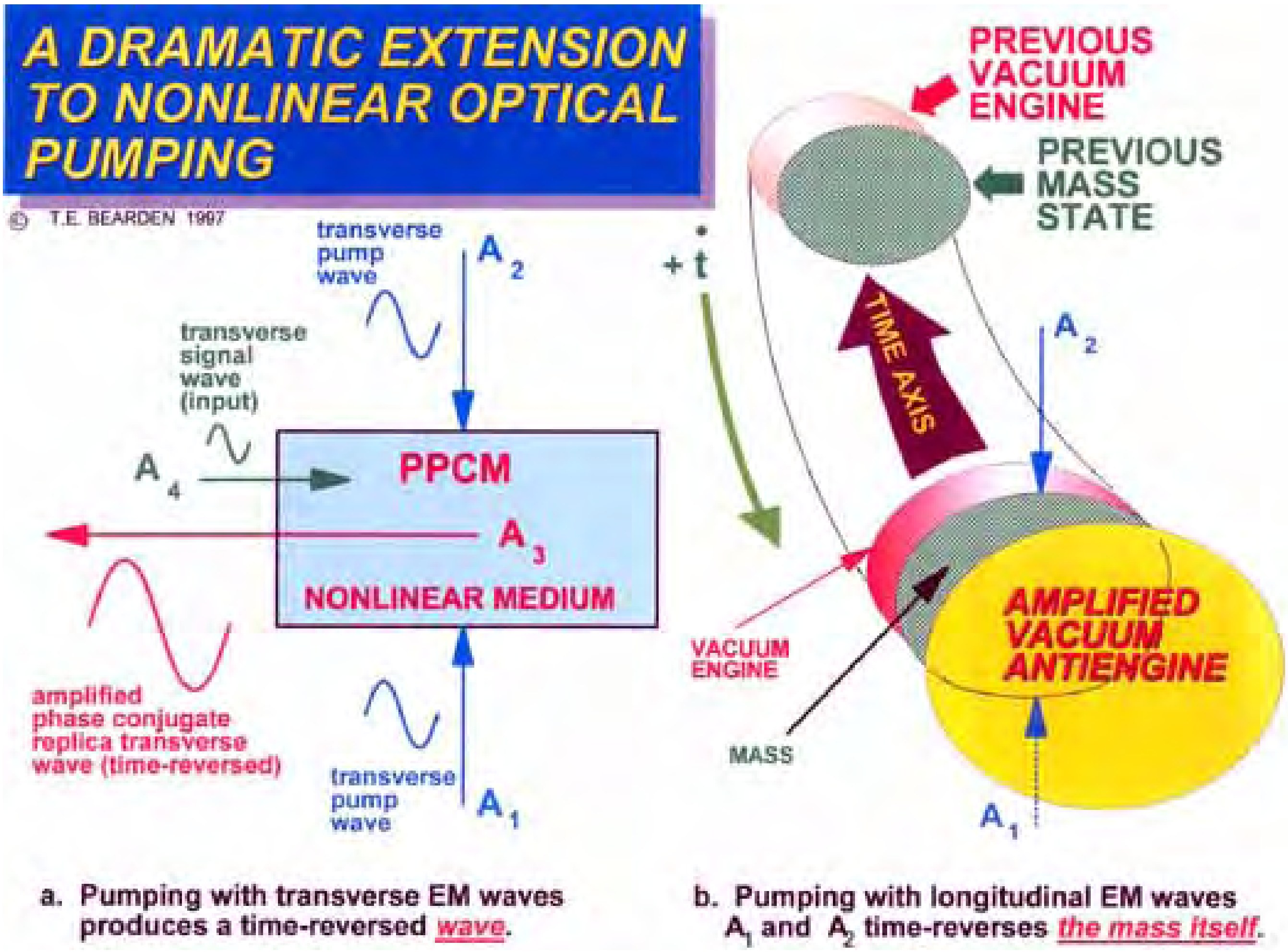

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

transverse pump wave
A_2
transverse signal wave (input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A_1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A_2
VACUUM ENGINE
MASS
AMPLIFIED VACUUM ANTIENGINE
A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

[Return to Slide Index]
[Next Slide]
[Previous Slide]

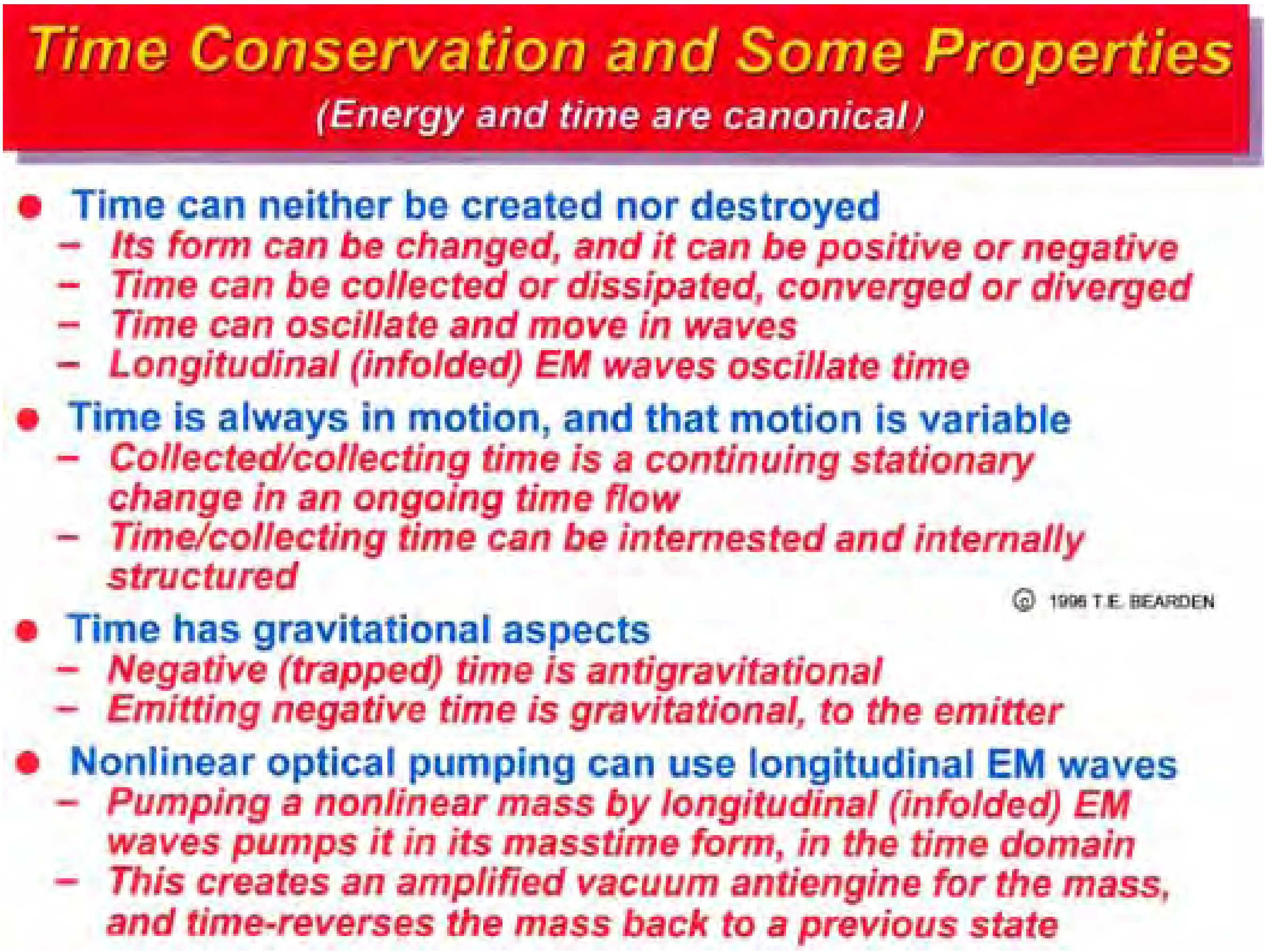

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*
- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

© 1996 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

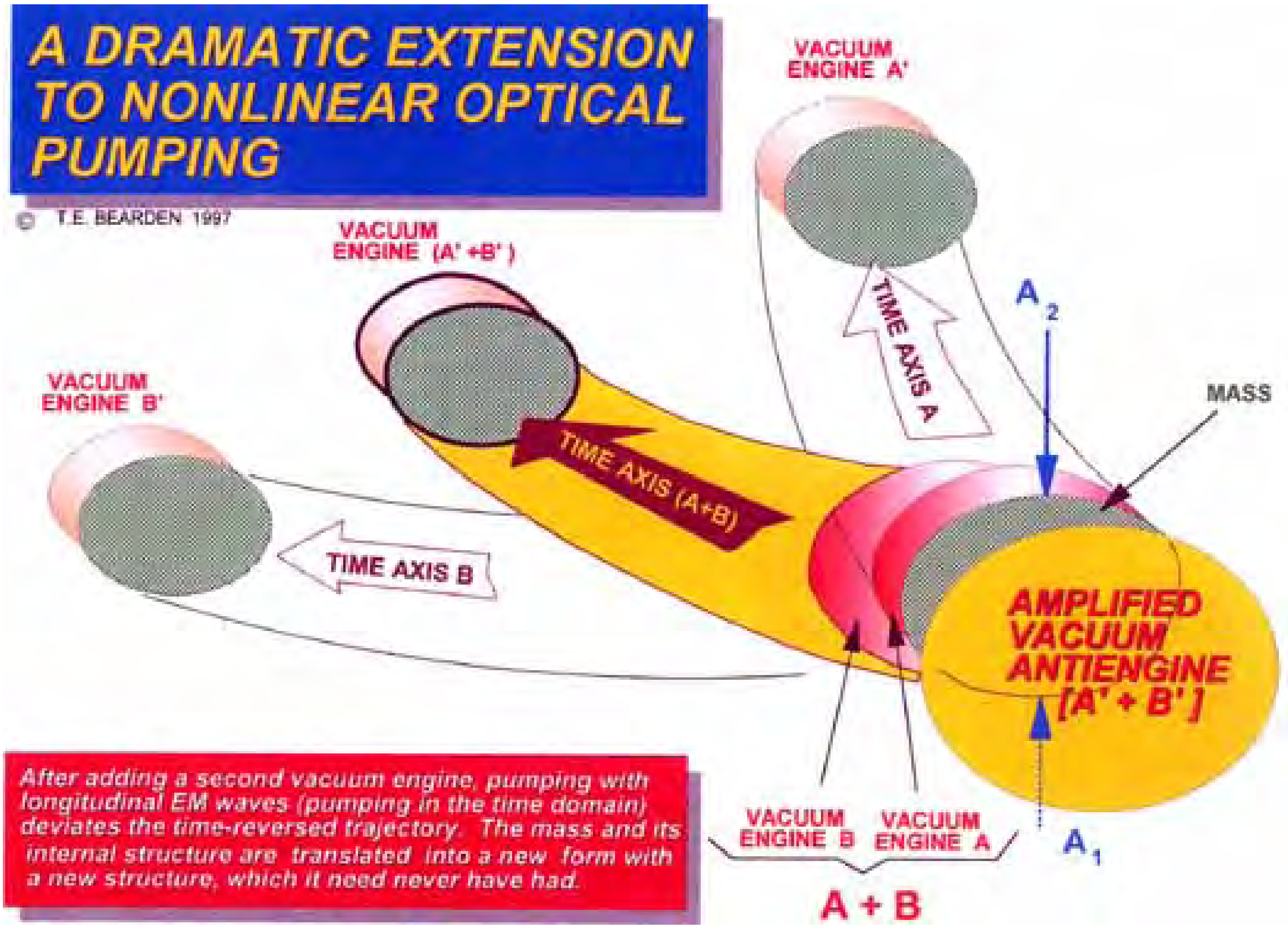

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A_1
A + B

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Return to Slide Index

Next Slide

Previous Slide

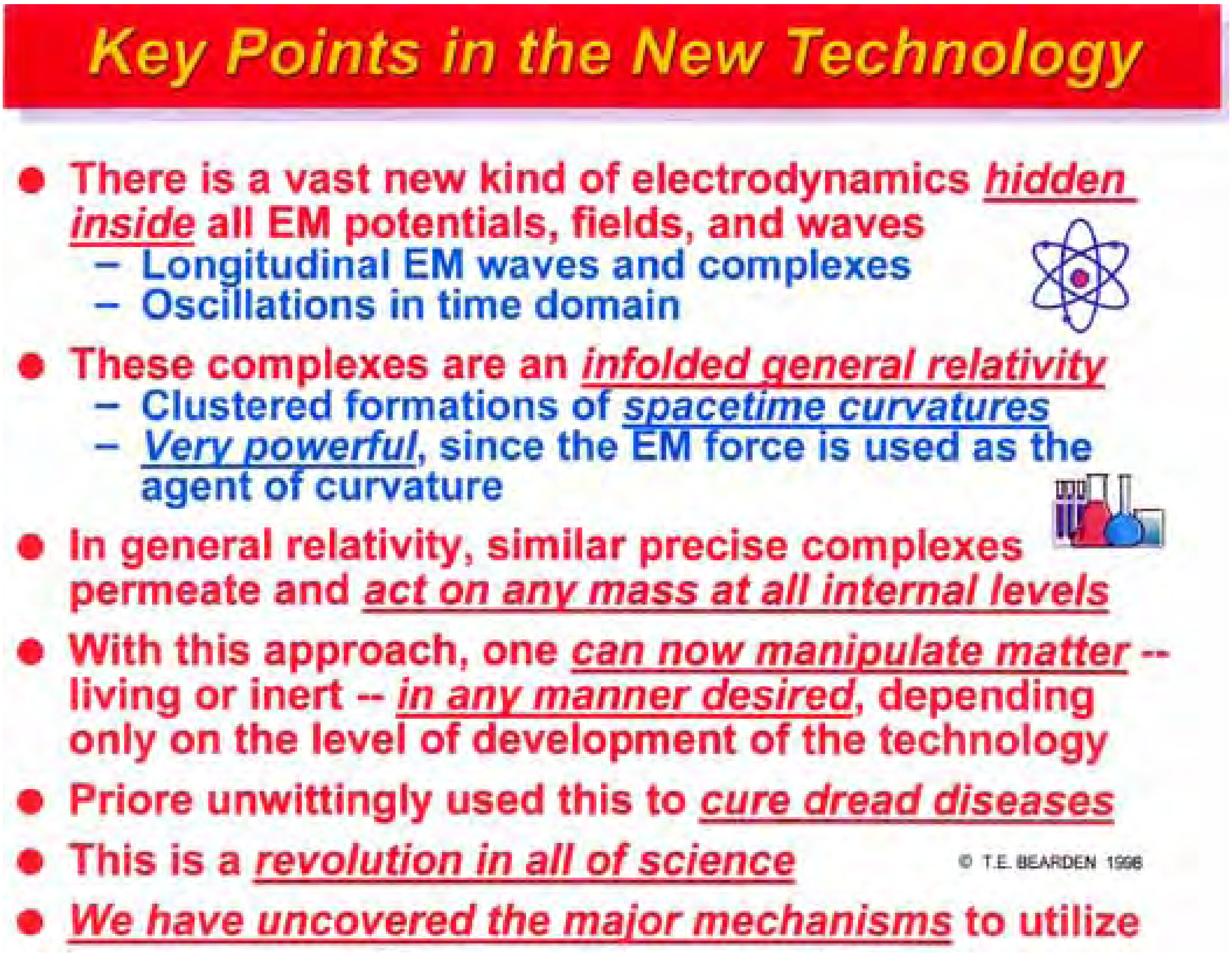

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

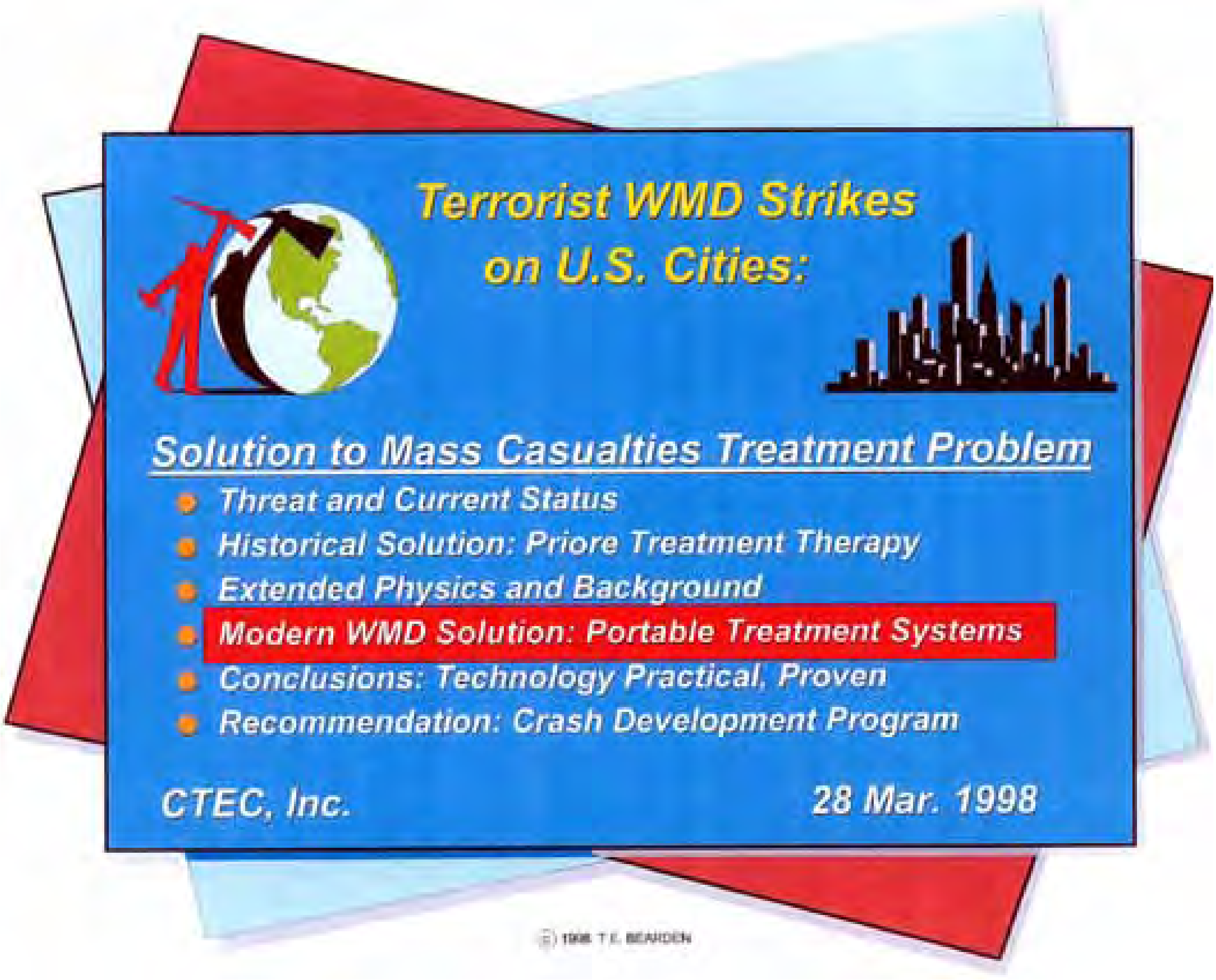

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

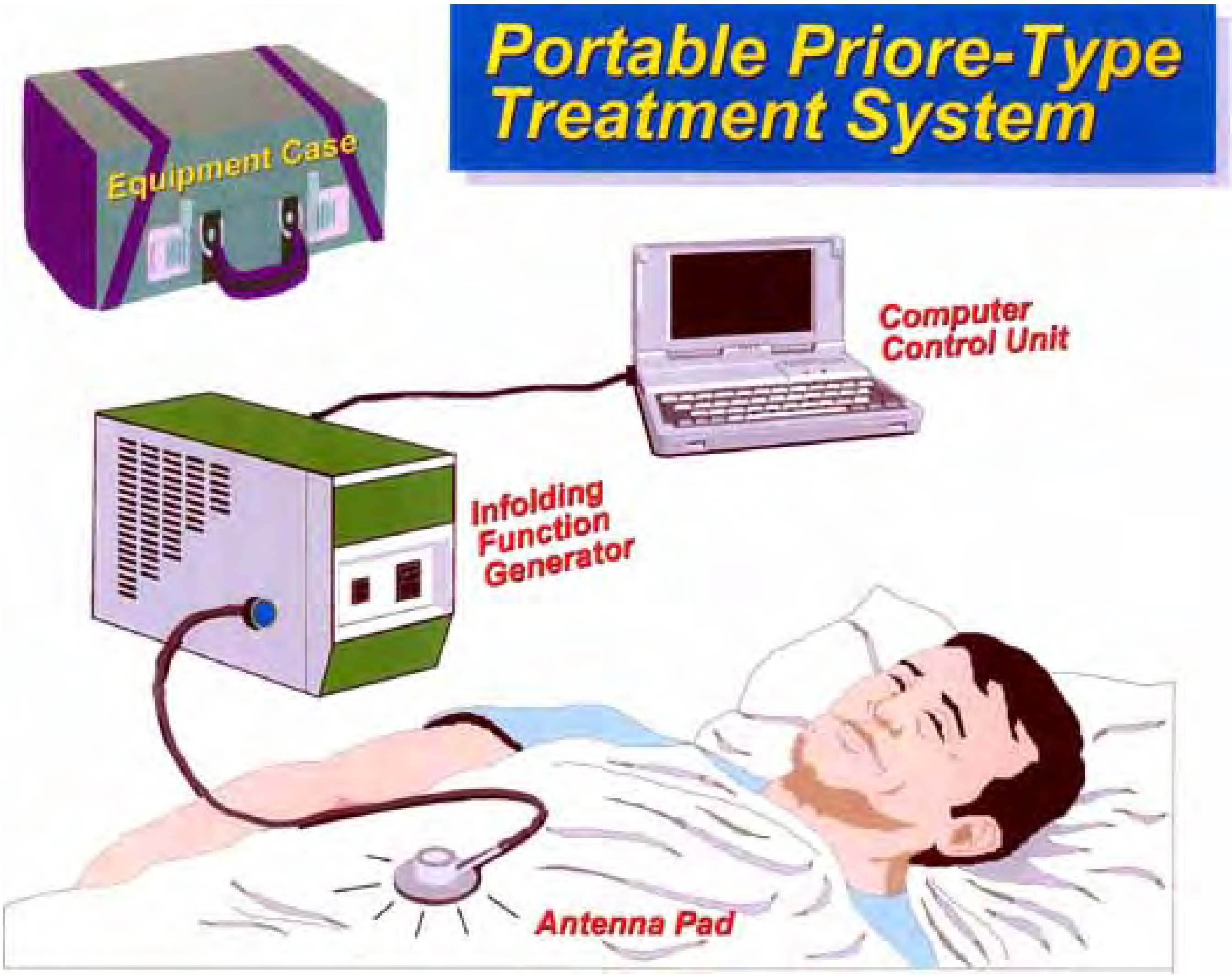

Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

Portable System Block Diagram

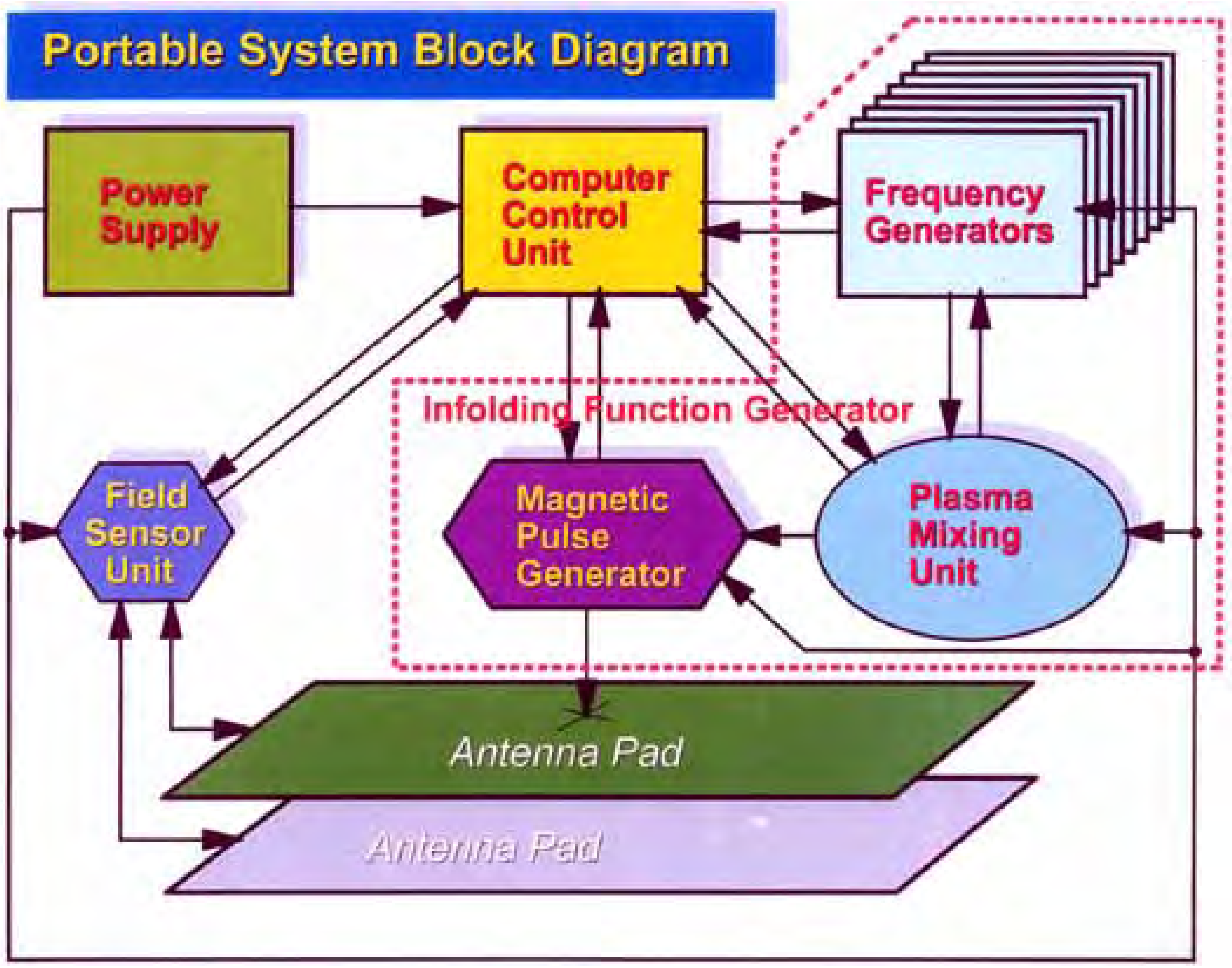

Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

[Return to Slide Index]

[Next Slide]

[Previous Slide]

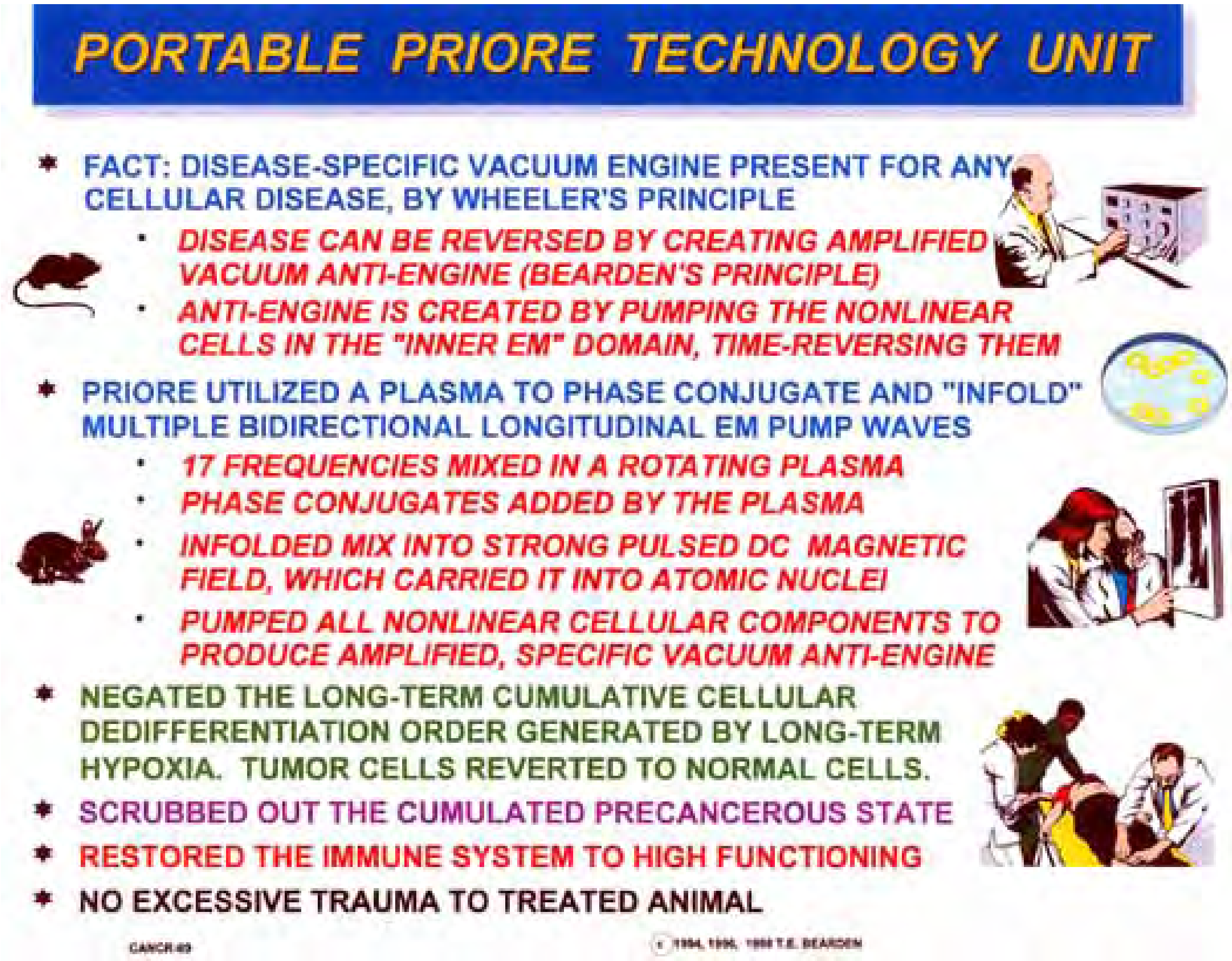

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

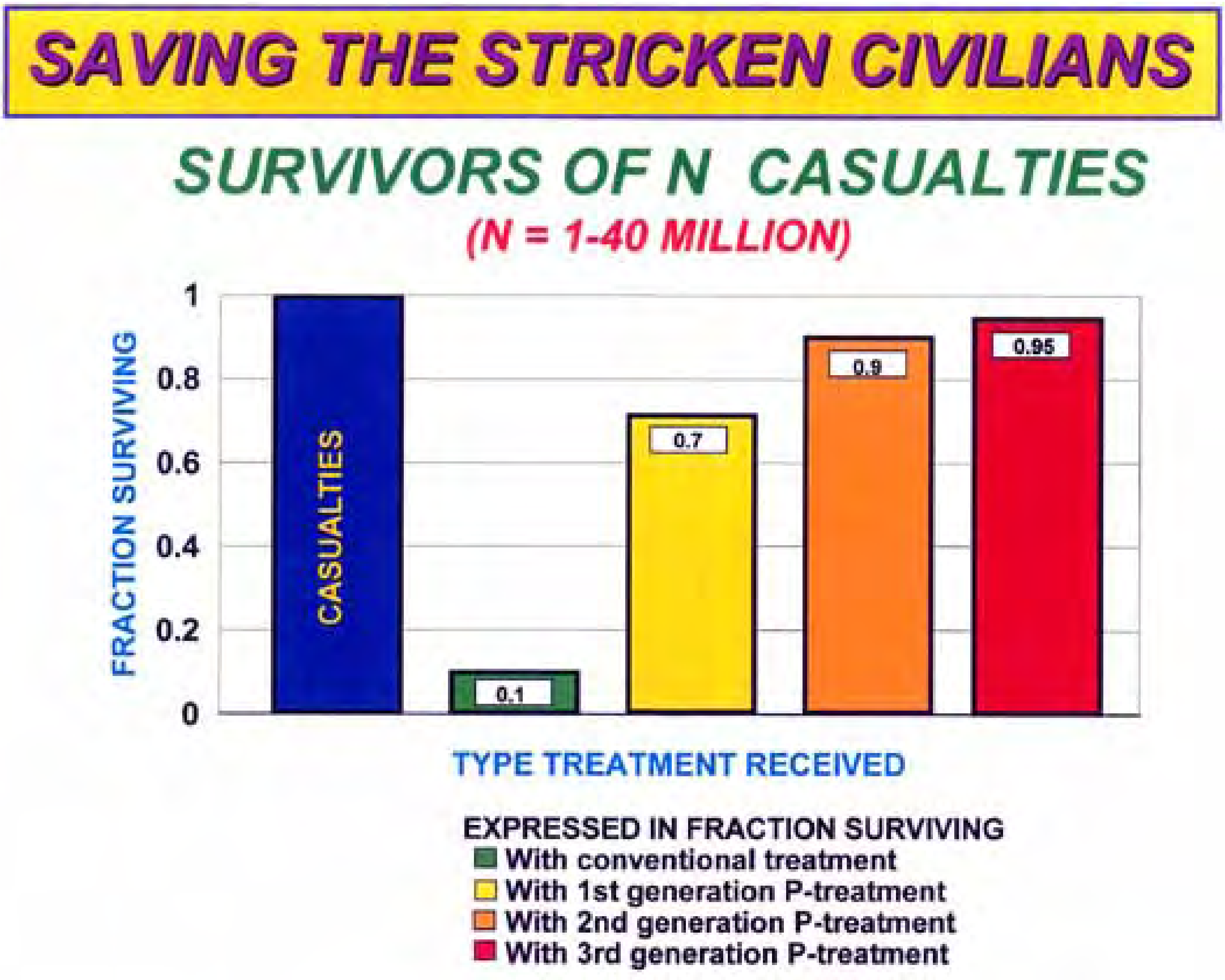

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

EXPRESSED IN FRACTION SURVIVING
- 🟩 With conventional treatment
- 🟨 With 1st generation P-treatment
- 🟧 With 2nd generation P-treatment
- 🟥 With 3rd generation P-treatment

Return to Slide Index

Next Slide

Previous Slide

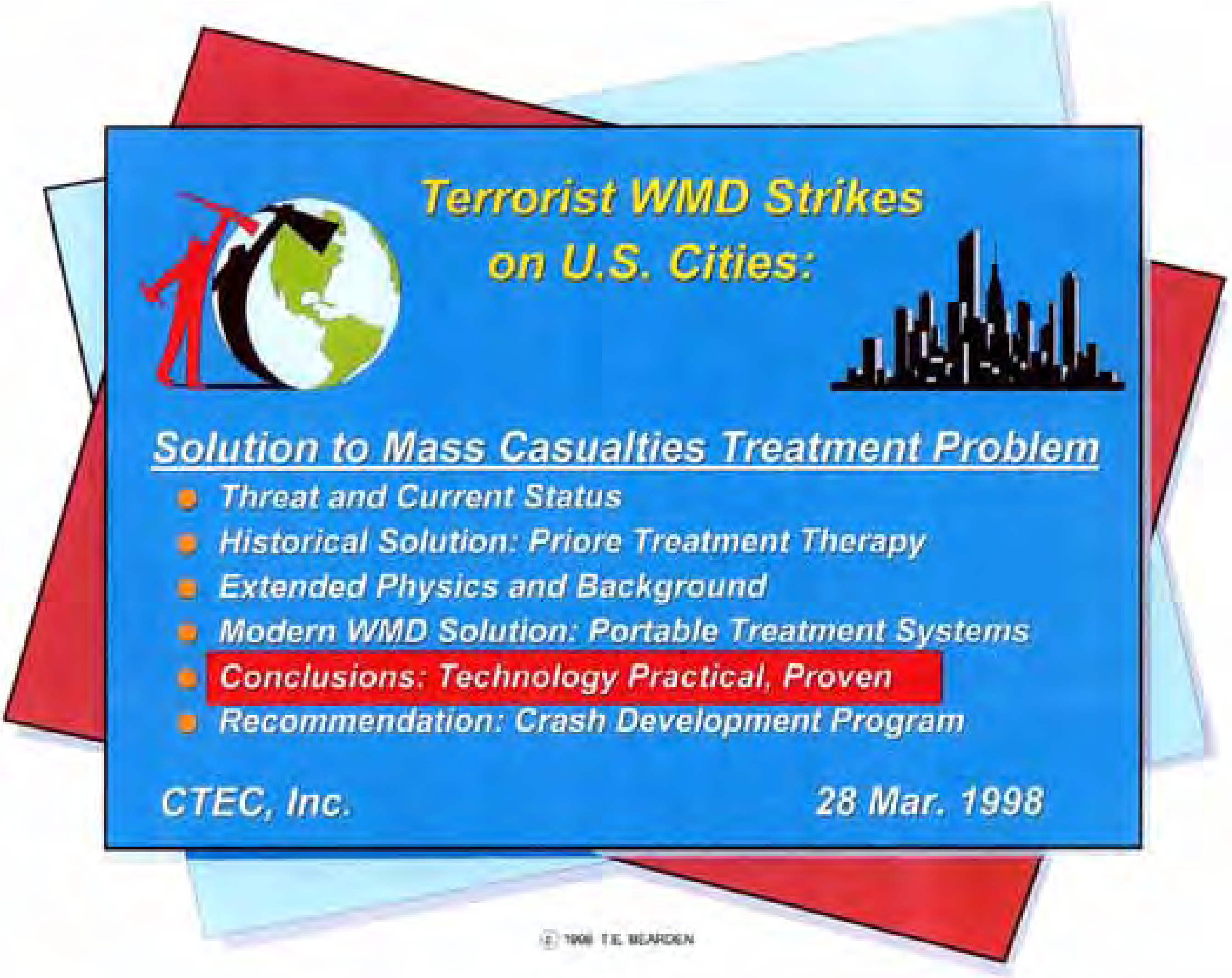

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

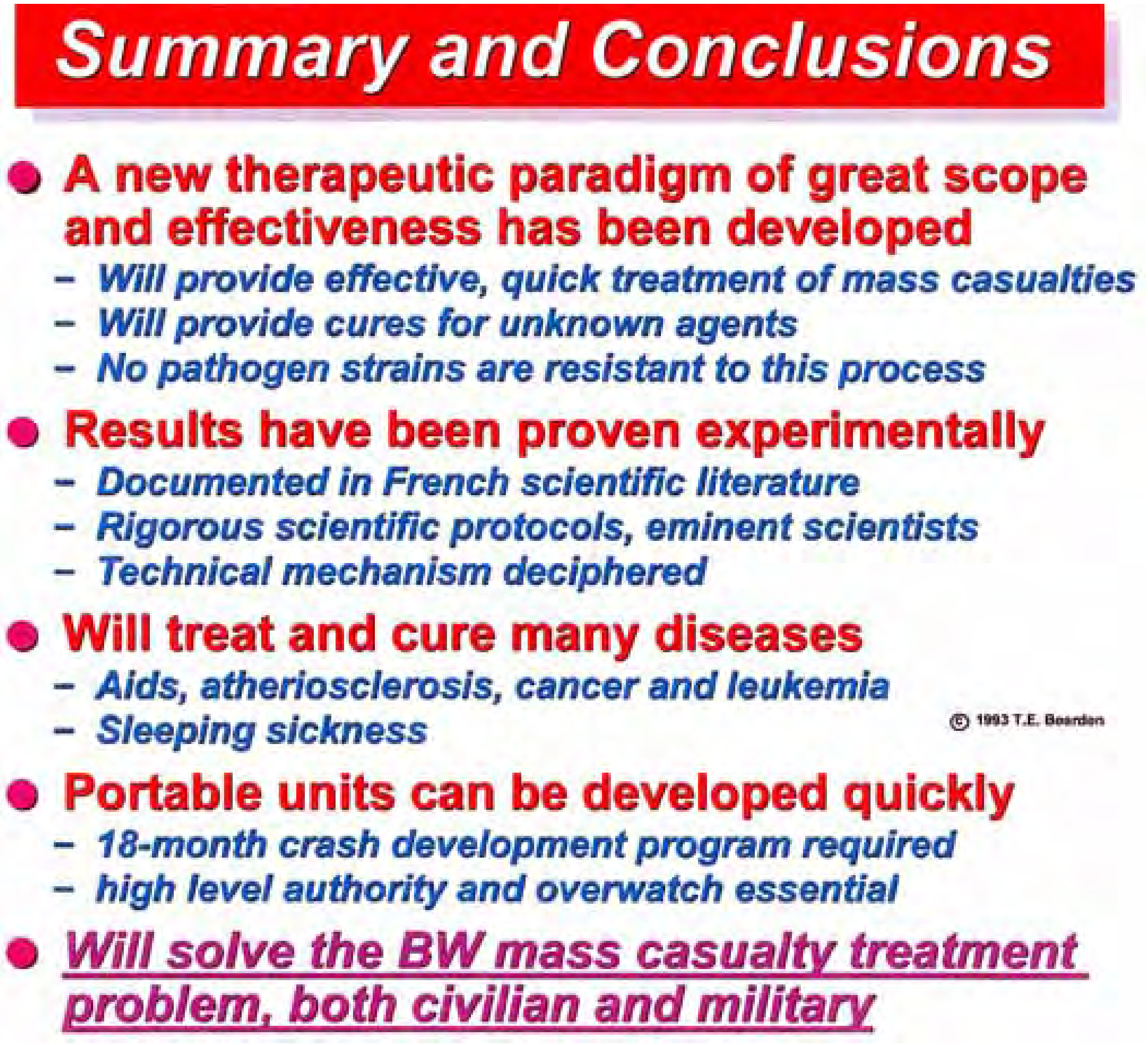

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - *Will provide effective, quick treatment of mass casualties*
  - *Will provide cures for unknown agents*
  - *No pathogen strains are resistant to this process*
- Results have been proven experimentally
  - *Documented in French scientific literature*
  - *Rigorous scientific protocols, eminent scientists*
  - *Technical mechanism deciphered*
- Will treat and cure many diseases
  - *Aids, atheriosclerosis, cancer and leukemia*
  - *Sleeping sickness*
- Portable units can be developed quickly
  - *18-month crash development program required*
  - *high level authority and overwatch essential*
- *Will solve the BW mass casualty treatment problem, both civilian and military*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

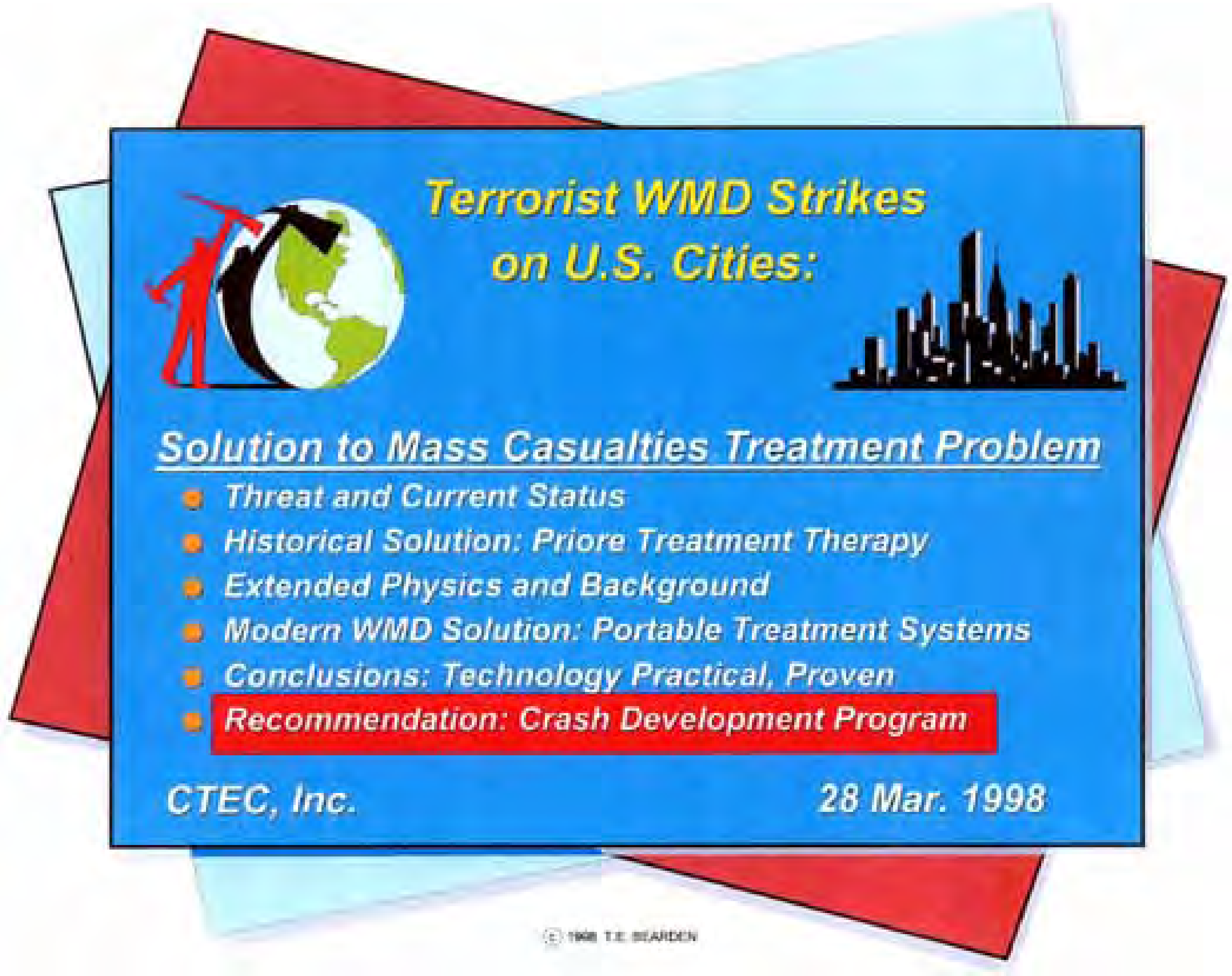

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

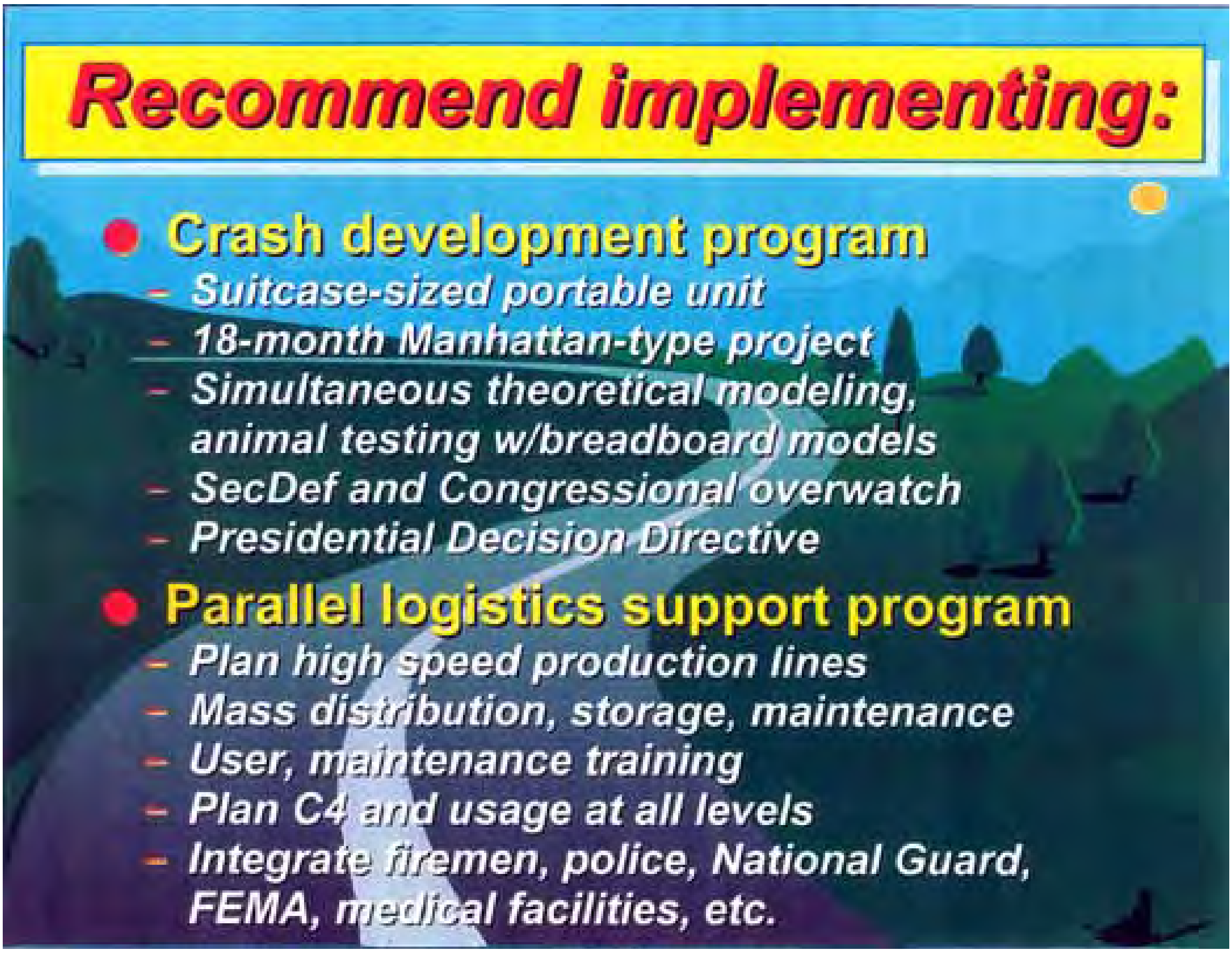

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

Return to Slide Index

Next Slide

Previous Slide

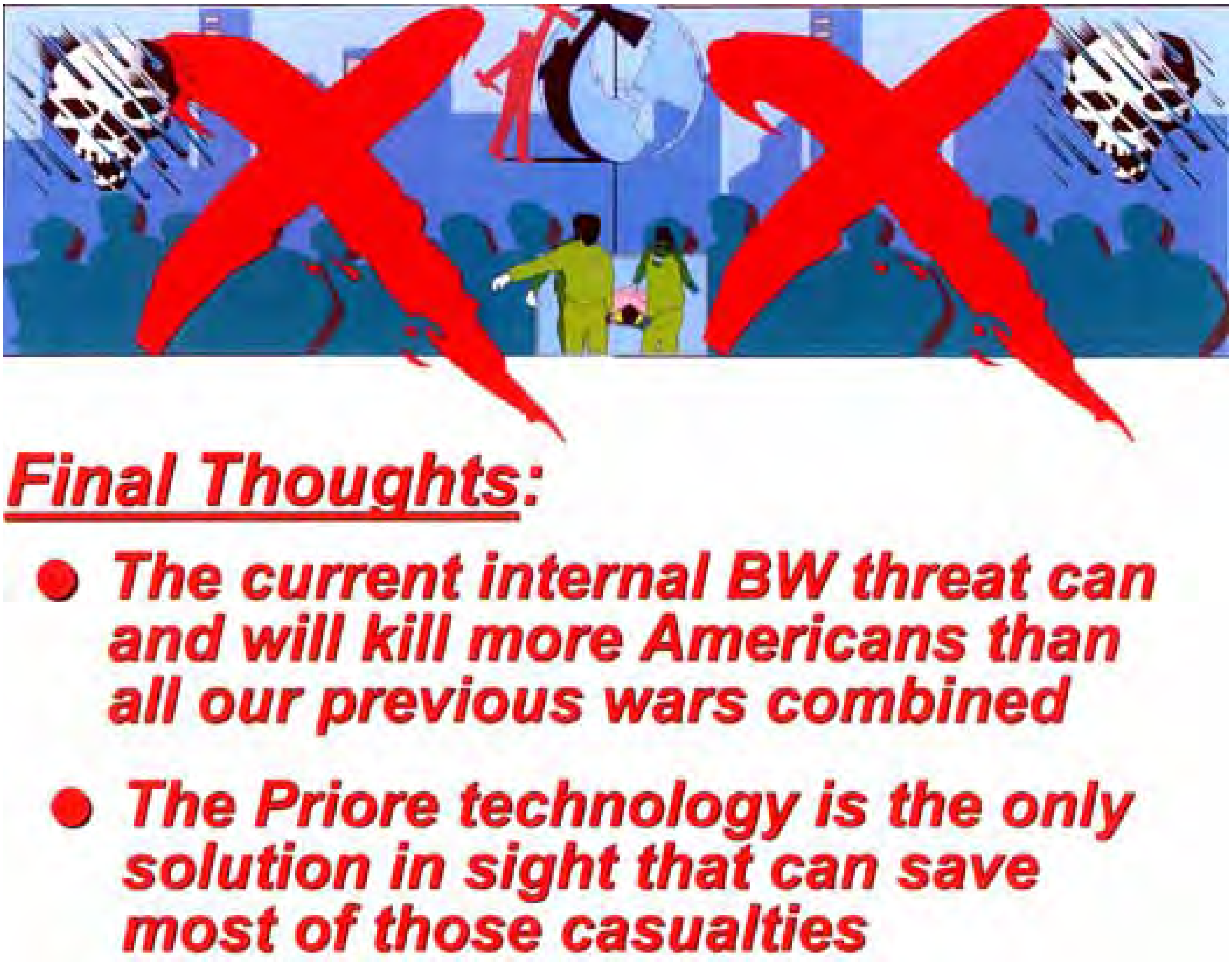

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular *regenerative* system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

Incls: Full package with table of contents

T.E. Bearden
President & CEO

Energy, Observations and Comments

---

1. A Conservative Field Requires a Single-Valued Potential
2. Earth-Electrosphere and Earth-Ionosphere Capacitors
3. Earth's Poynting Energy Flow
4. The Hidden Energy Flow around us
5. Electrostatic Fields of the Earth's Electrosphere Capacitor in Fair Weather and Thunderstorms
6. Flow of Current and Energy Into, Through, and Along a Conductor
7. Energy "collected" on a charge
8. Fiber Fuse Effect
9. Impact of Vacuum Engineering
10. Lawandy's Experiment
11. Field may be nonconservative for a multivalued potential
12. The Ostrich Problem - Part 1
13. The Ostrich Problem - Part 2
14. Russia had Overunity Devices in 1930s
15. Energy output of human body
16. A "static" Poynting S-flow free-energy generator
17. Topology and Energy Conservation
18. Block Diagram: Sweet's Vacuum Triode

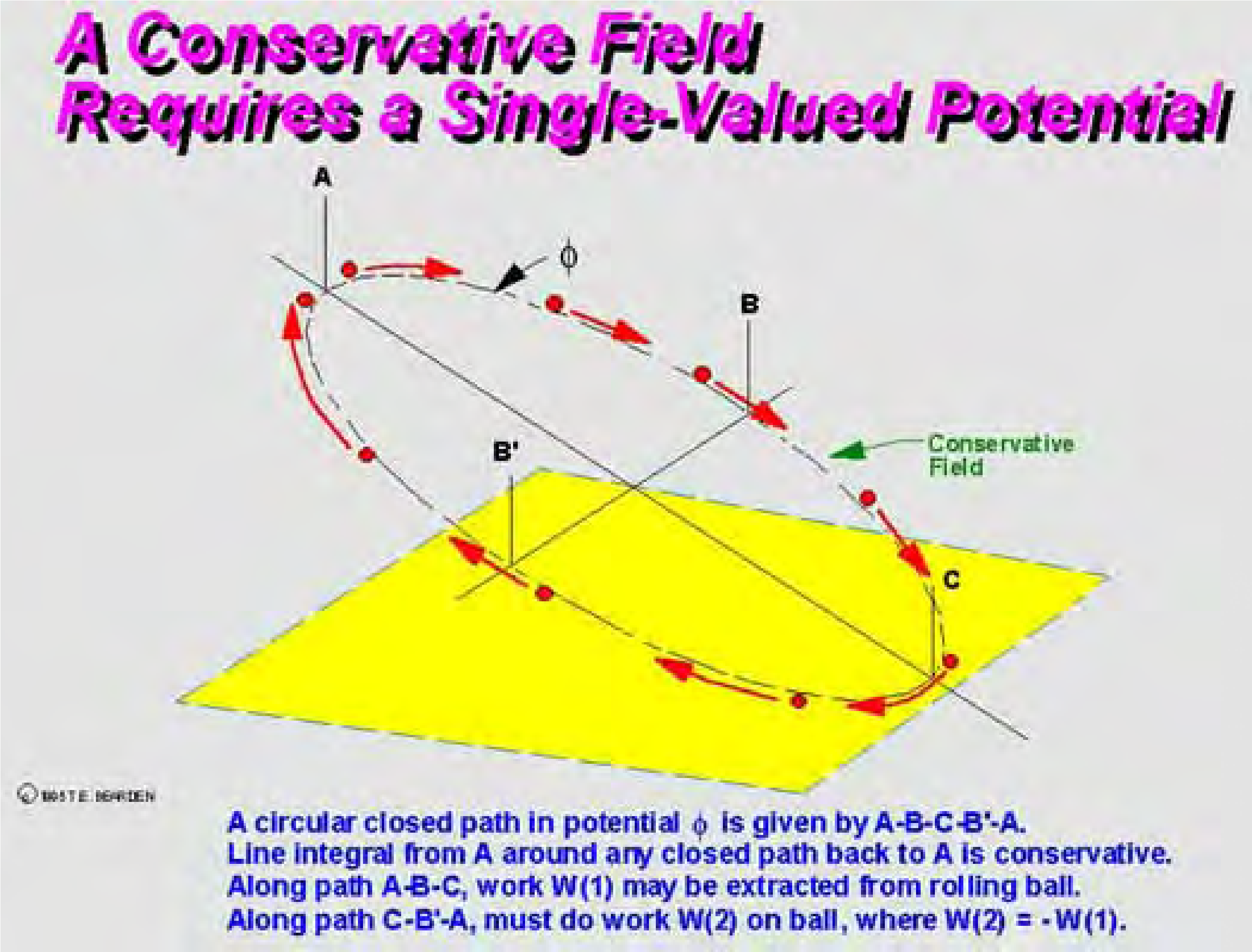

A Conservative Field Requires a Single-Valued Potential

© 1995 T.E. BEARDEN

A circular closed path in potential φ is given by A-B-C-B'-A.
Line integral from A around any closed path back to A is conservative.
Along path A-B-C, work W(1) may be extracted from rolling ball.
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1).

Return to Slide Index

Next Slide

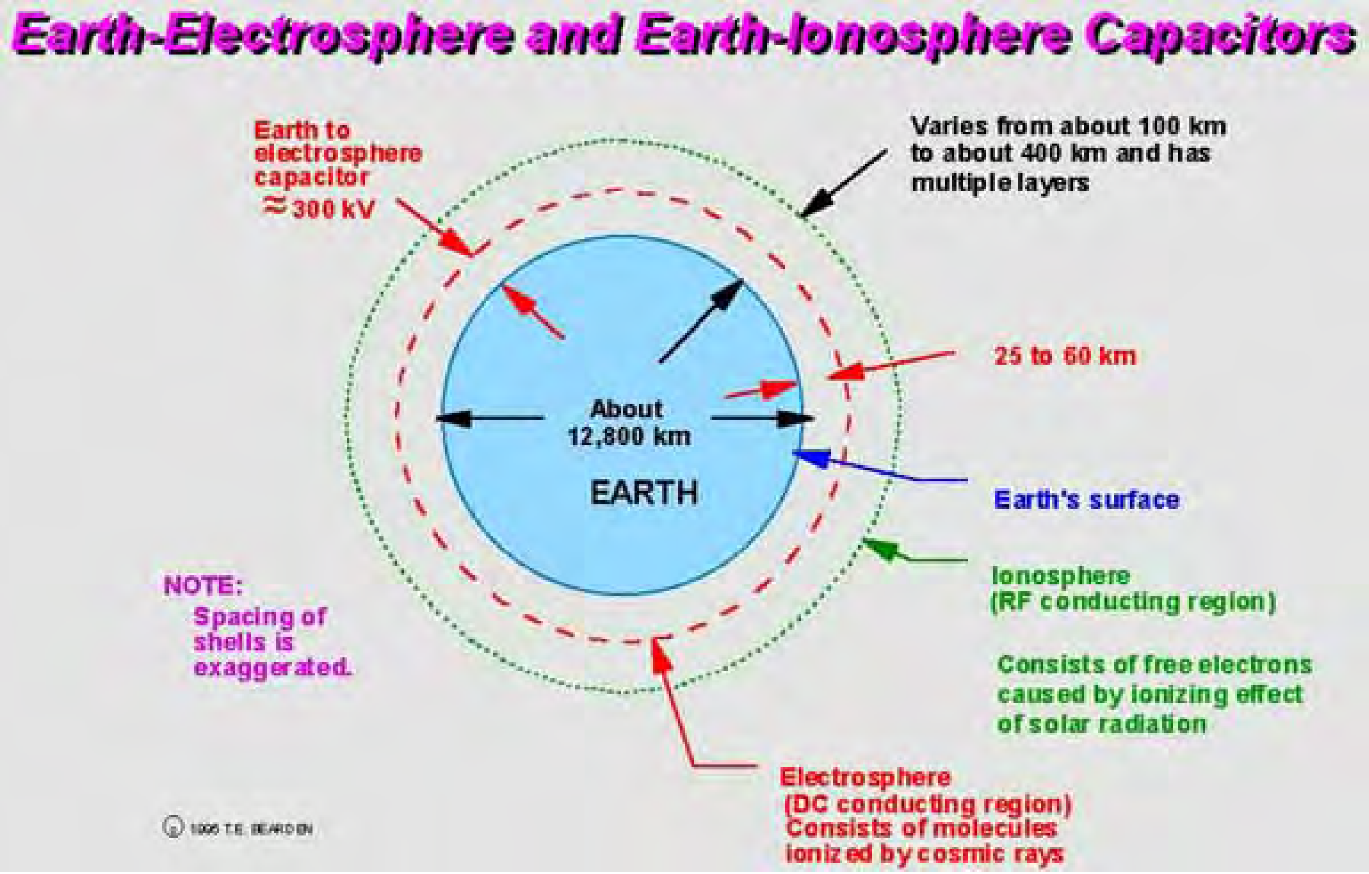

Earth-Electrosphere and Earth-Ionosphere Capacitors

Earth to electrosphere capacitor
≈ 300 kV

Varies from about 100 km to about 400 km and has multiple layers

25 to 60 km

About 12,800 km

EARTH

Earth's surface

Ionosphere
(RF conducting region)

Consists of free electrons caused by ionizing effect of solar radiation

NOTE:
Spacing of shells is exaggerated.

Electrosphere
(DC conducting region)
Consists of molecules ionized by cosmic rays

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

EARTH'S POYNTING ENERGY FLOW

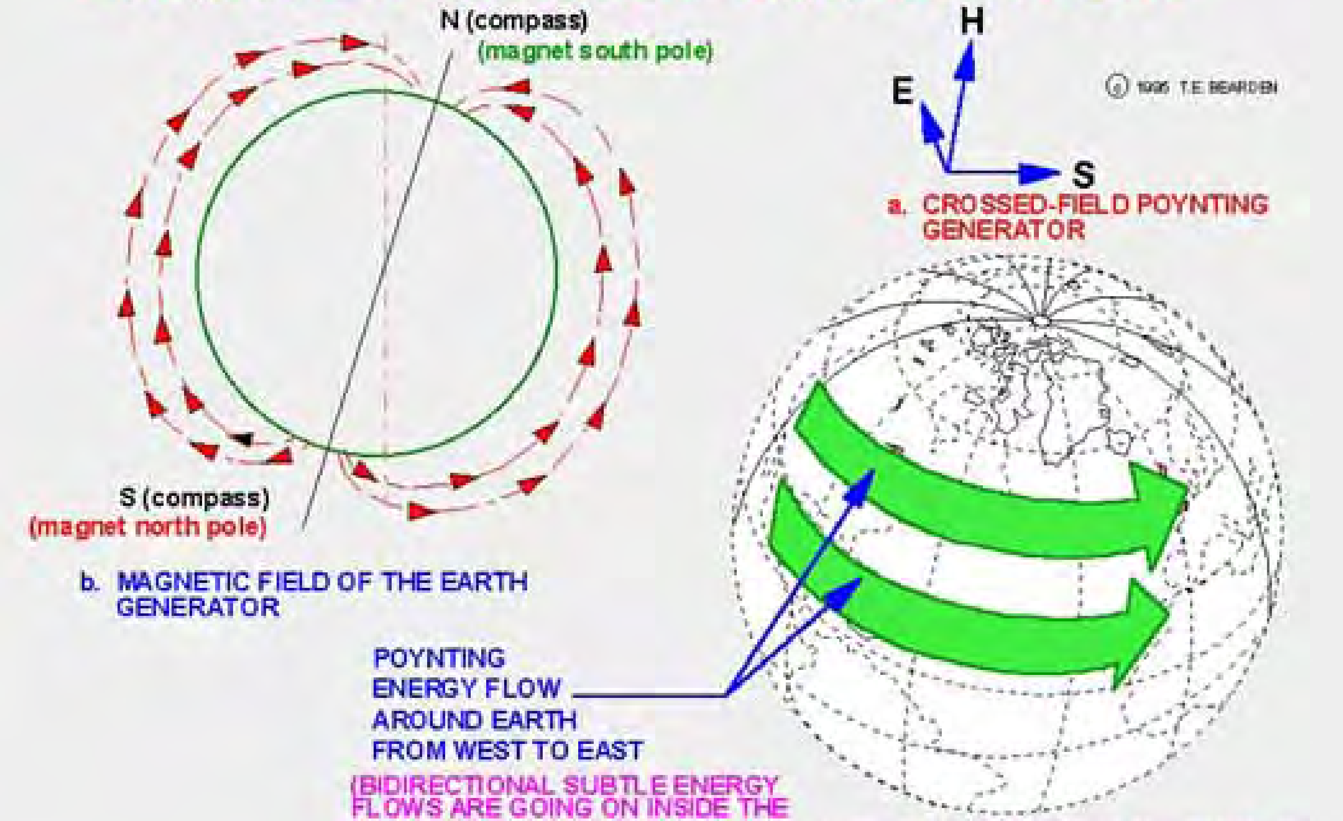
N (compass)
(magnet south pole)
H
E
S
S (compass)
(magnet north pole)

© 1995 T.E. BEARDEN

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING
ENERGY FLOW
AROUND EARTH
FROM WEST TO EAST

(BIDIRECTIONAL SUBTLE ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

Return to Slide Index

Next Slide

Previous Slide

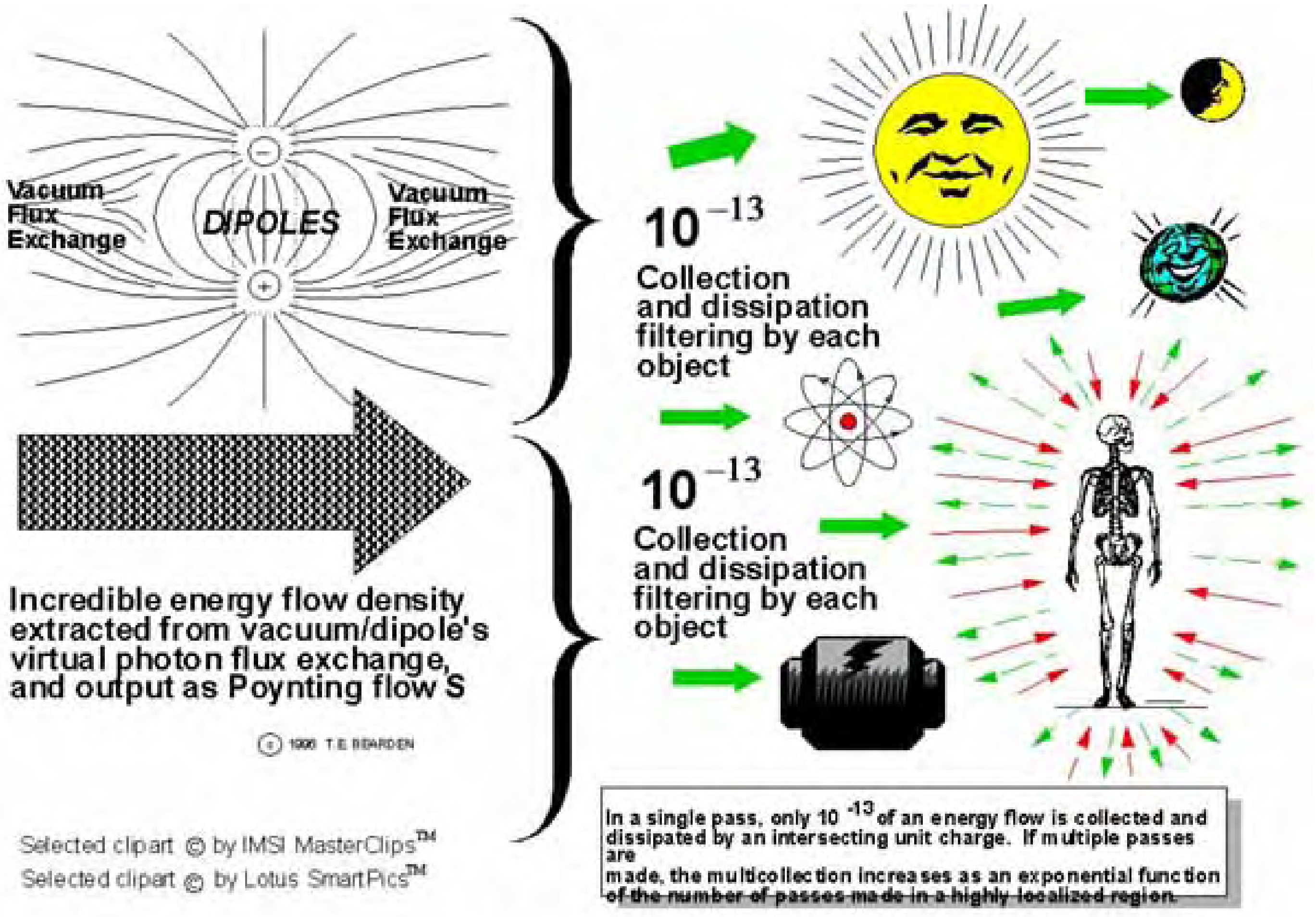
Vacuum Flux Exchange
DIPOLES
Vacuum Flux Exchange
10^{-13}
Collection and dissipation filtering by each object
10^{-13}
Collection and dissipation filtering by each object
© 1996 T.E. BEARDEN

Incredible energy flow density extracted from vacuum/dipole's virtual photon flux exchange, and output as Poynting flow S

In a single pass, only 10^{-13} of an energy flow is collected and dissipated by an intersecting unit charge. If multiple passes are made, the multicollection increases as an exponential function of the number of passes made in a highly localized region.

Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

Figure 19. The world of physical matter exists behind a dramatic 10^{-13} filter in a fiery cauldron of fierce and primeval EM energy flow.

Return to Slide Index

Next Slide

Previous Slide

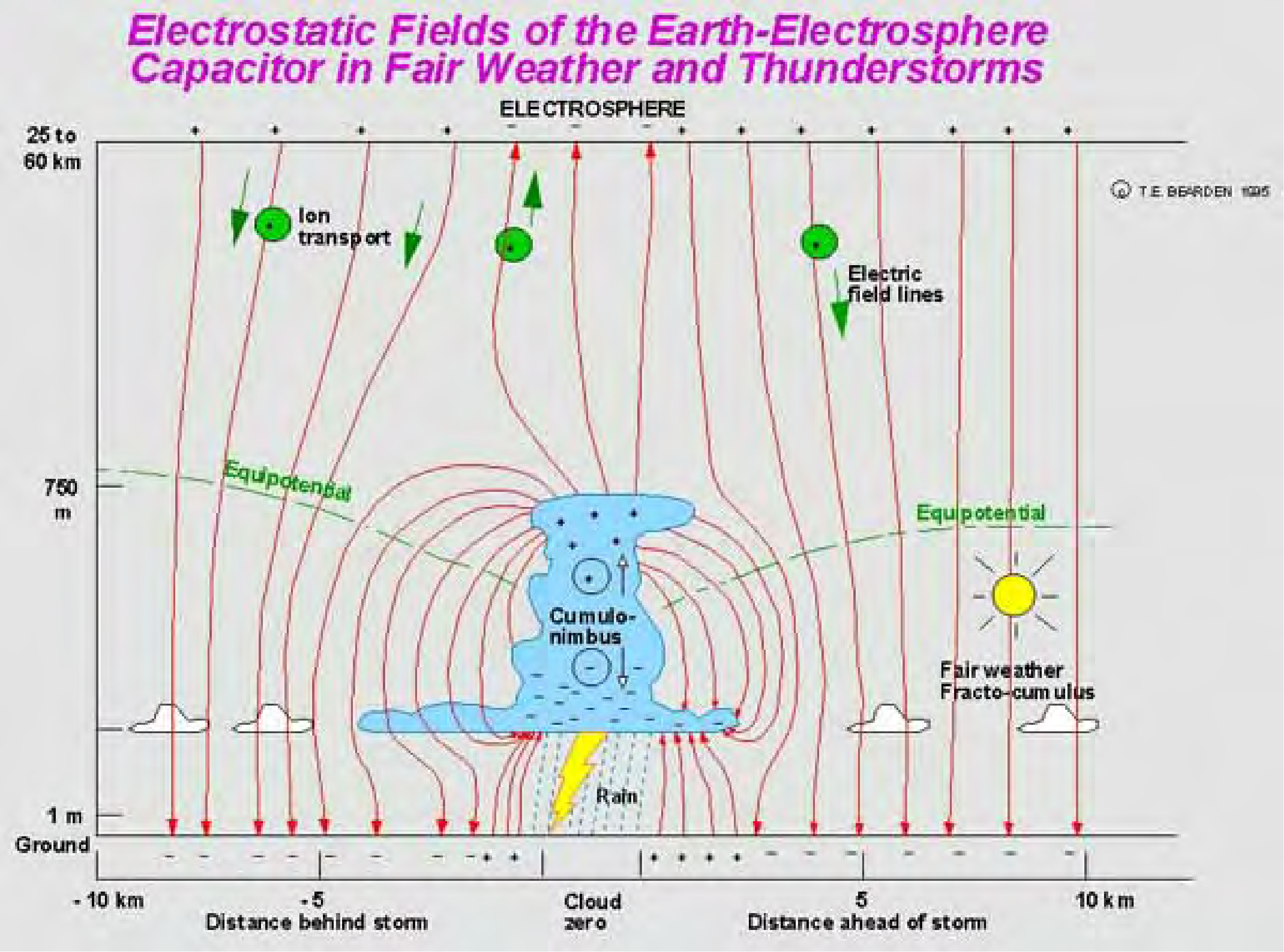

Electrostatic Fields of the Earth-Electrosphere Capacitor in Fair Weather and Thunderstorms

ELECTROSPHERE
25 to 60 km
© T.E. BEARDEN 1995
Ion transport
Electric field lines
750 m
Equipotential
Equipotential
Cumulo-nimbus
Fair weather Fracto-cumulus
Rain
1 m
Ground
- 10 km
- 5
Distance behind storm
Cloud zero
5
Distance ahead of storm
10 km

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

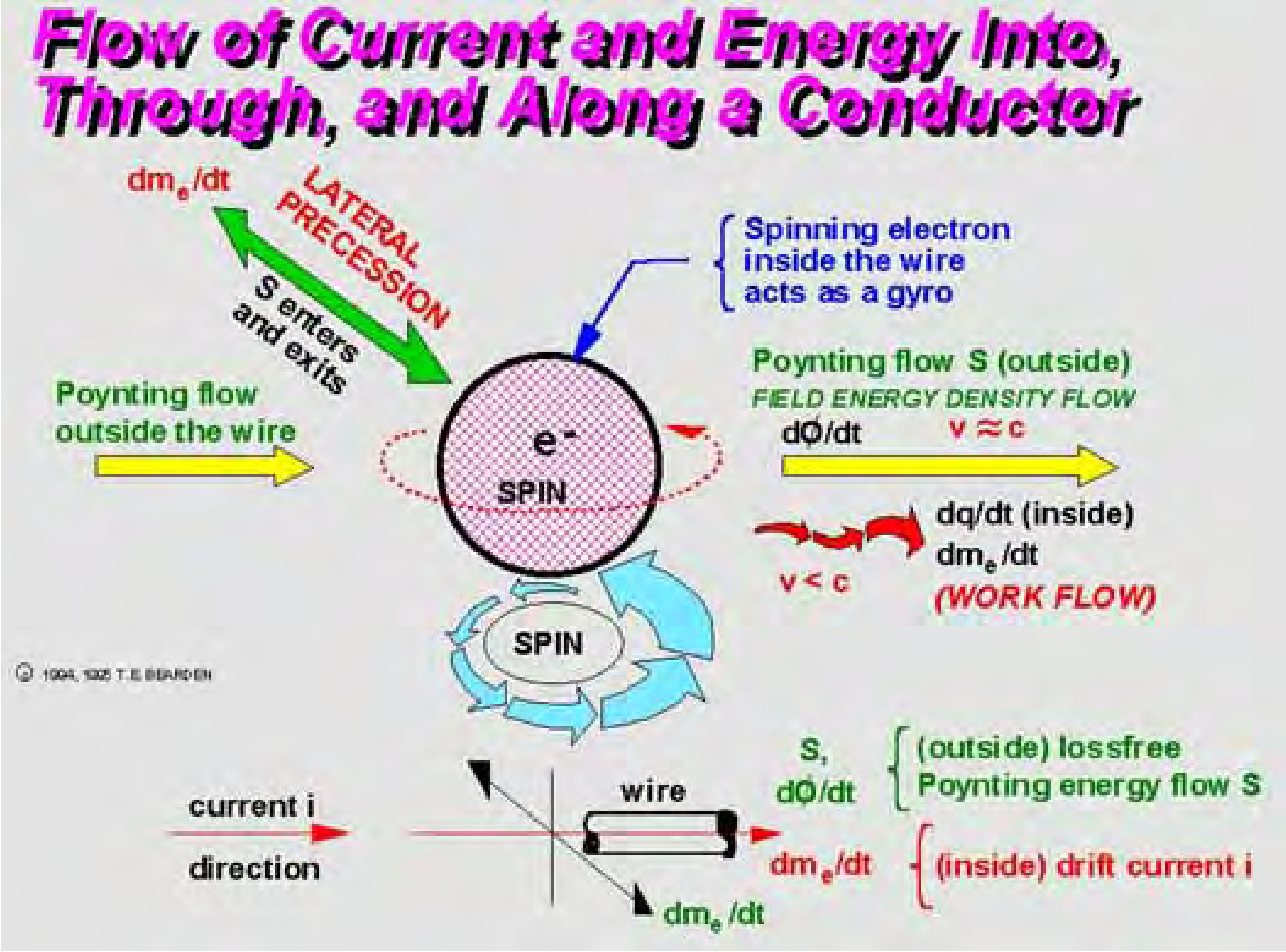

Flow of Current and Energy Into, Through, and Along a Conductor

dm_e/dt

LATERAL PRECESSION

S enters and exits

Spinning electron inside the wire acts as a gyro

Poynting flow outside the wire

e^-

SPIN

Poynting flow S (outside)

FIELD ENERGY DENSITY FLOW

$d\phi/dt$ $v \approx c$

dq/dt (inside)

dm_e/dt

$v < c$

(WORK FLOW)

SPIN

© 1994, 1995 T.E. BEARDEN

current i

direction

wire

S, $d\phi/dt$ (outside) lossfree Poynting energy flow S

dm_e/dt (inside) drift current i

dm_e/dt

Return to Slide Index

Next Slide

Previous Slide

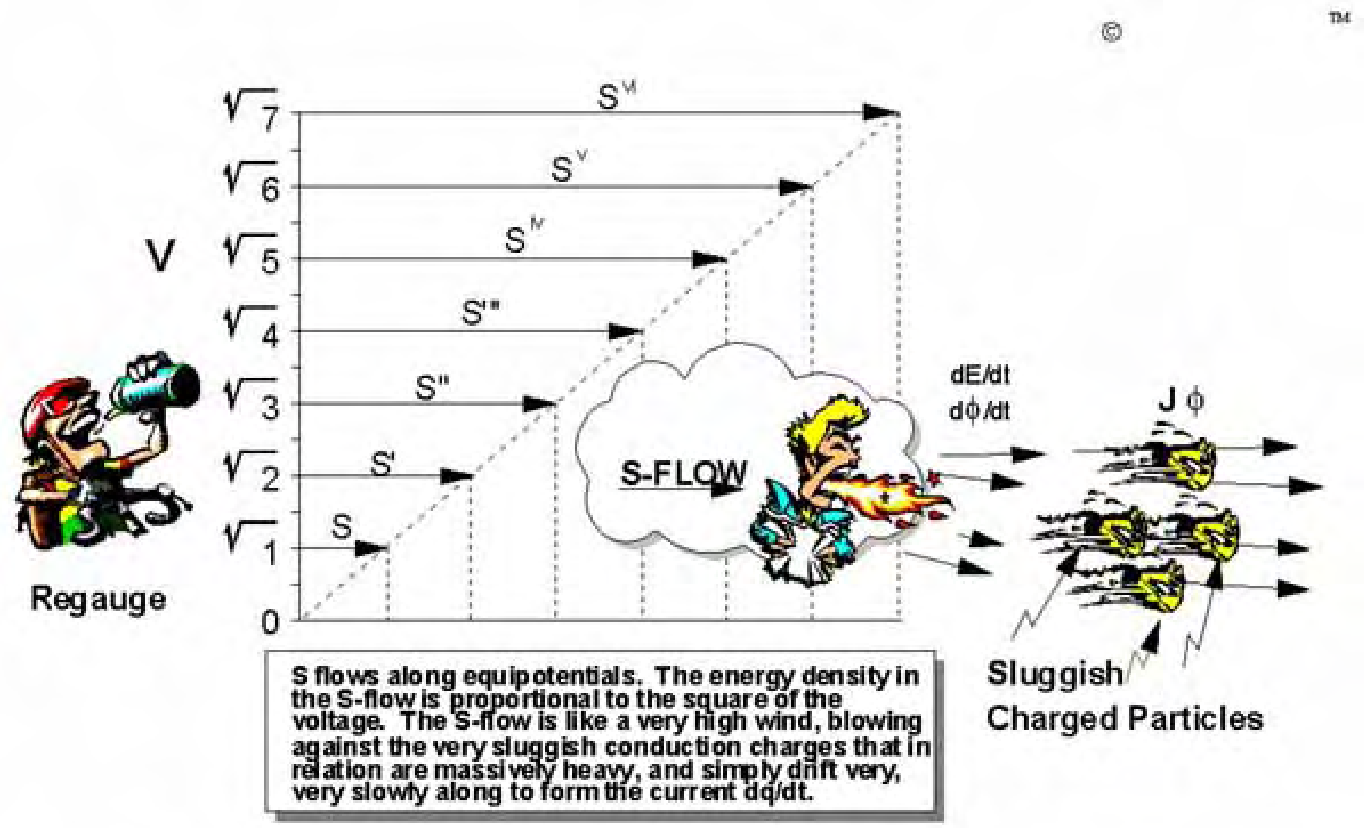
©
™
V
√7
√6
√5
√4
√3
√2
√1
0
S^VI
S^V
S^IV
S'"
S"
S'
S
S-FLOW
Regauge
dE/dt
dφ/dt
J φ
Sluggish
Charged Particles

S flows along equipotentials. The energy density in the S-flow is proportional to the square of the voltage. The S-flow is like a very high wind, blowing against the very sluggish conduction charges that in relation are massively heavy, and simply drift very, very slowly along to form the current dq/dt.

Figure 17. Energy "collected" on a charge depends upon the charge being in a continuous Poynting S-flow. When S ceases, there is no excess energy "collected" on the charge. Energy collection never occurs in chunks, but is always a dynamic, ongoing process.

Return to Slide Index

Next Slide

Previous Slide

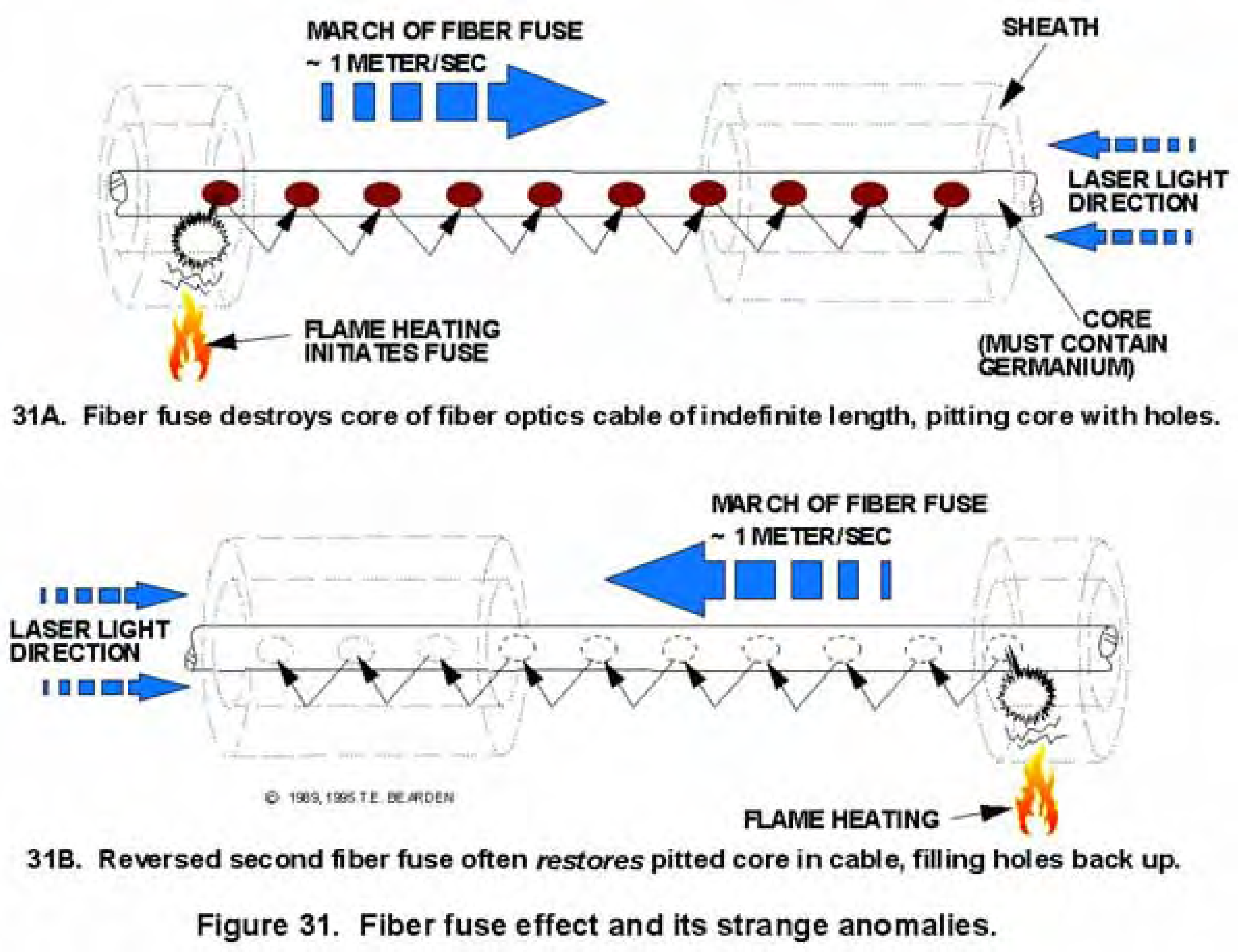
MARCH OF FIBER FUSE
~ 1 METER/SEC
SHEATH
LASER LIGHT DIRECTION
FLAME HEATING INITIATES FUSE
CORE (MUST CONTAIN GERMANIUM)
MARCH OF FIBER FUSE
~ 1 METER/SEC
LASER LIGHT DIRECTION
© 1989, 1995 T.E. BEARDEN
FLAME HEATING

31A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

31B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 31. Fiber fuse effect and its strange anomalies.

Return to Slide Index

Next Slide

Previous Slide

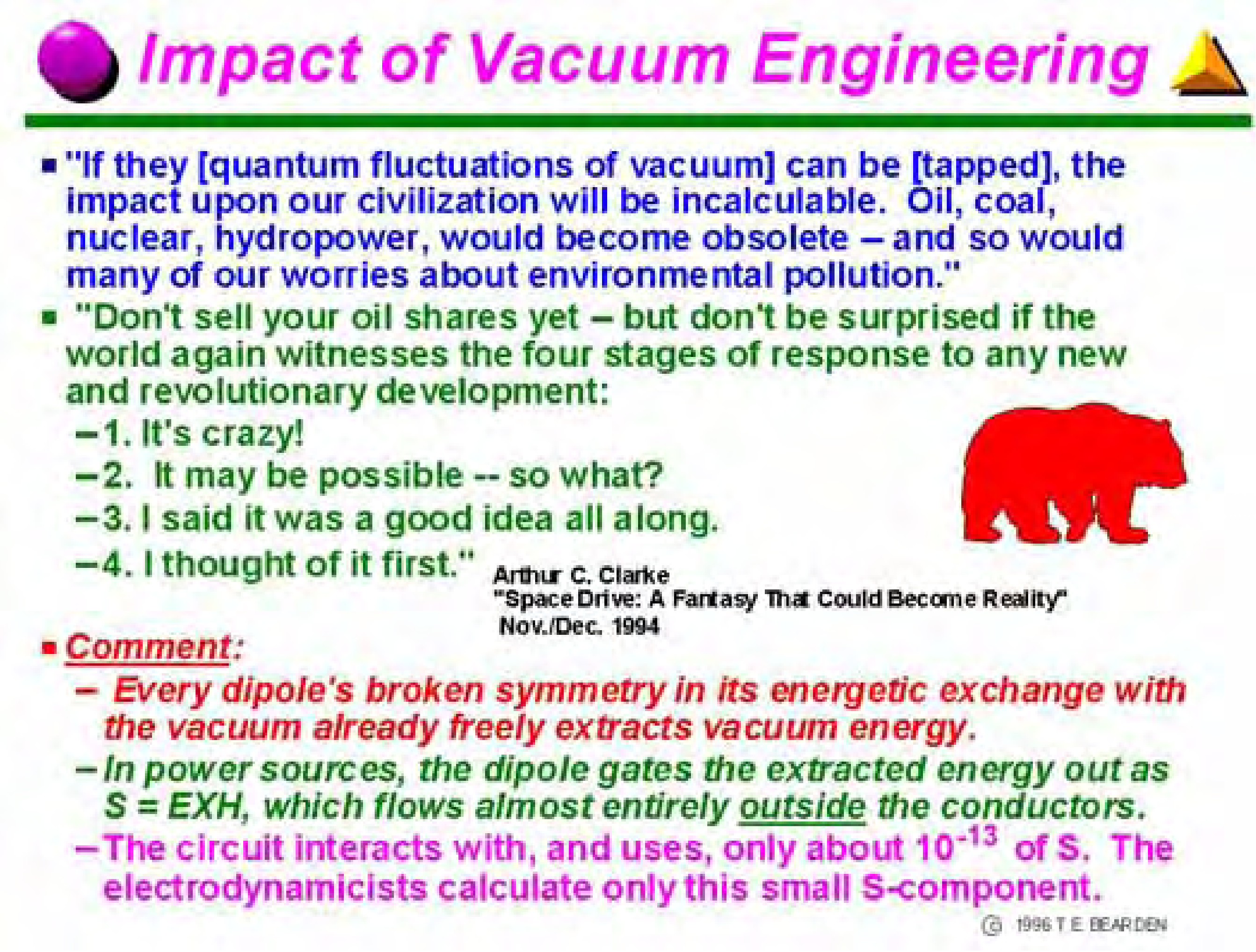

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete – and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet – but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

© 1996 T E BEARDEN

Return to Slide Index

Next Slide

Previous Slide

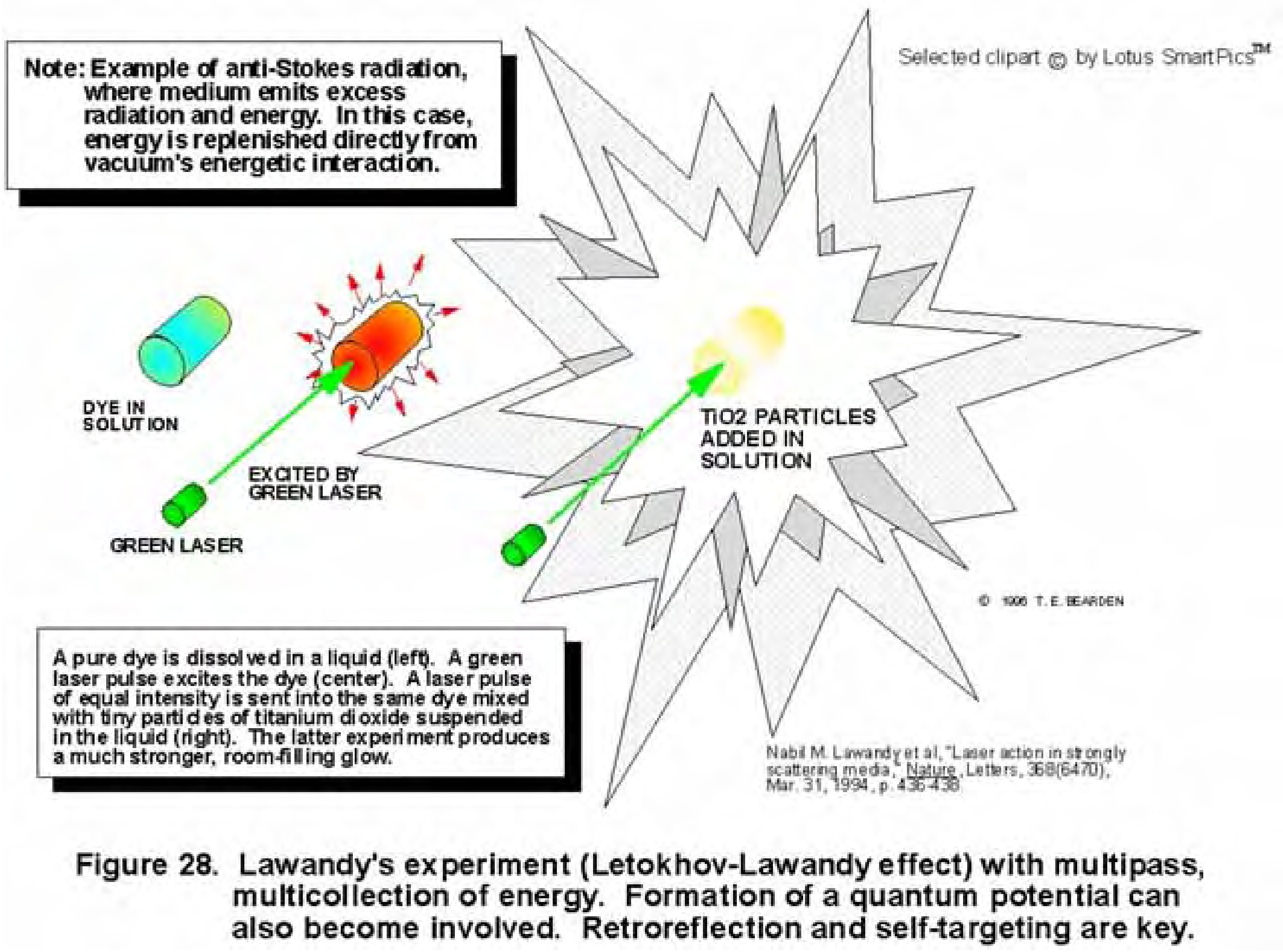

Selected clipart © by Lotus SmartPics™

Note: Example of anti-Stokes radiation, where medium emits excess radiation and energy. In this case, energy is replenished directly from vacuum's energetic interaction.

© 1996 T. E. BEARDEN

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Nabil M. Lawandy et al, "Laser action in strongly scattering media," Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438

Figure 28. Lawandy's experiment (Letokhov-Lawandy effect) with multipass, multicollection of energy. Formation of a quantum potential can also become involved. Retroreflection and self-targeting are key.

Return to Slide Index

Next Slide

Previous Slide

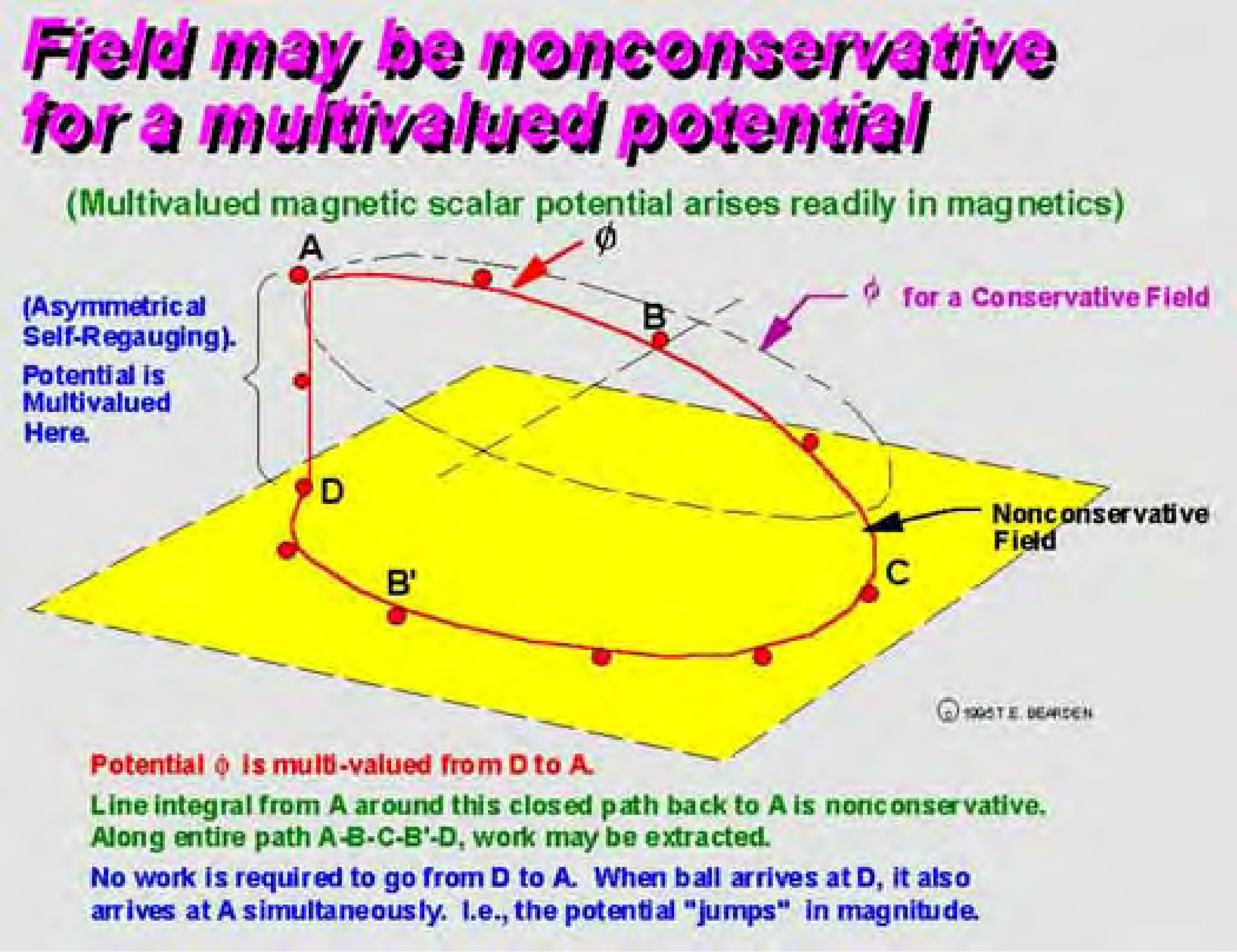

Field may be nonconservative for a multivalued potential

(Multivalued magnetic scalar potential arises readily in magnetics)

A
ϕ
B
ϕ for a Conservative Field
(Asymmetrical Self-Regauging).
Potential is Multivalued Here.
D
Nonconservative Field
C
B'
© 1995 T.E. BEARDEN

Potential ϕ is multi-valued from D to A.

Line integral from A around this closed path back to A is nonconservative. Along entire path A-B-C-B'-D, work may be extracted.

No work is required to go from D to A. When ball arrives at D, it also arrives at A simultaneously. I.e., the potential "jumps" in magnitude.

[Return to Slide Index]
[Next Slide]
[Previous Slide]

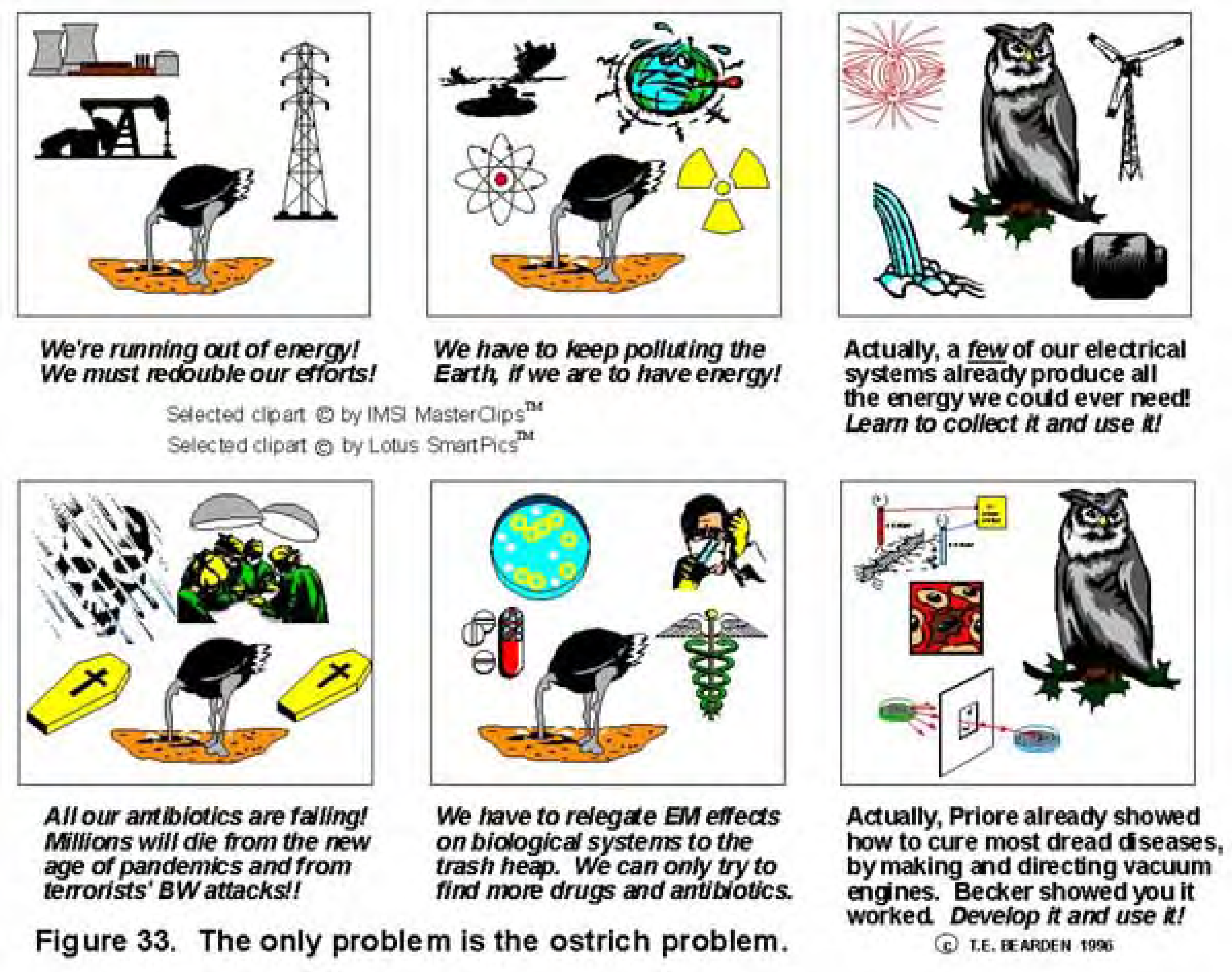

We're running out of energy! We must redouble our efforts!

We have to keep polluting the Earth, if we are to have energy!

Actually, a few of our electrical systems already produce all the energy we could ever need! *Learn to collect it and use it!*

Selected clipart © by IMSI MasterClips™

Selected clipart © by Lotus SmartPics™

All our antibiotics are failing! Millions will die from the new age of pandemics and from terrorists' BW attacks!!

We have to relegate EM effects on biological systems to the trash heap. We can only try to find more drugs and antibiotics.

Actually, Priore already showed how to cure most dread diseases, by making and directing vacuum engines. Becker showed you it worked. *Develop it and use it!*

Figure 33. The only problem is the ostrich problem.

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

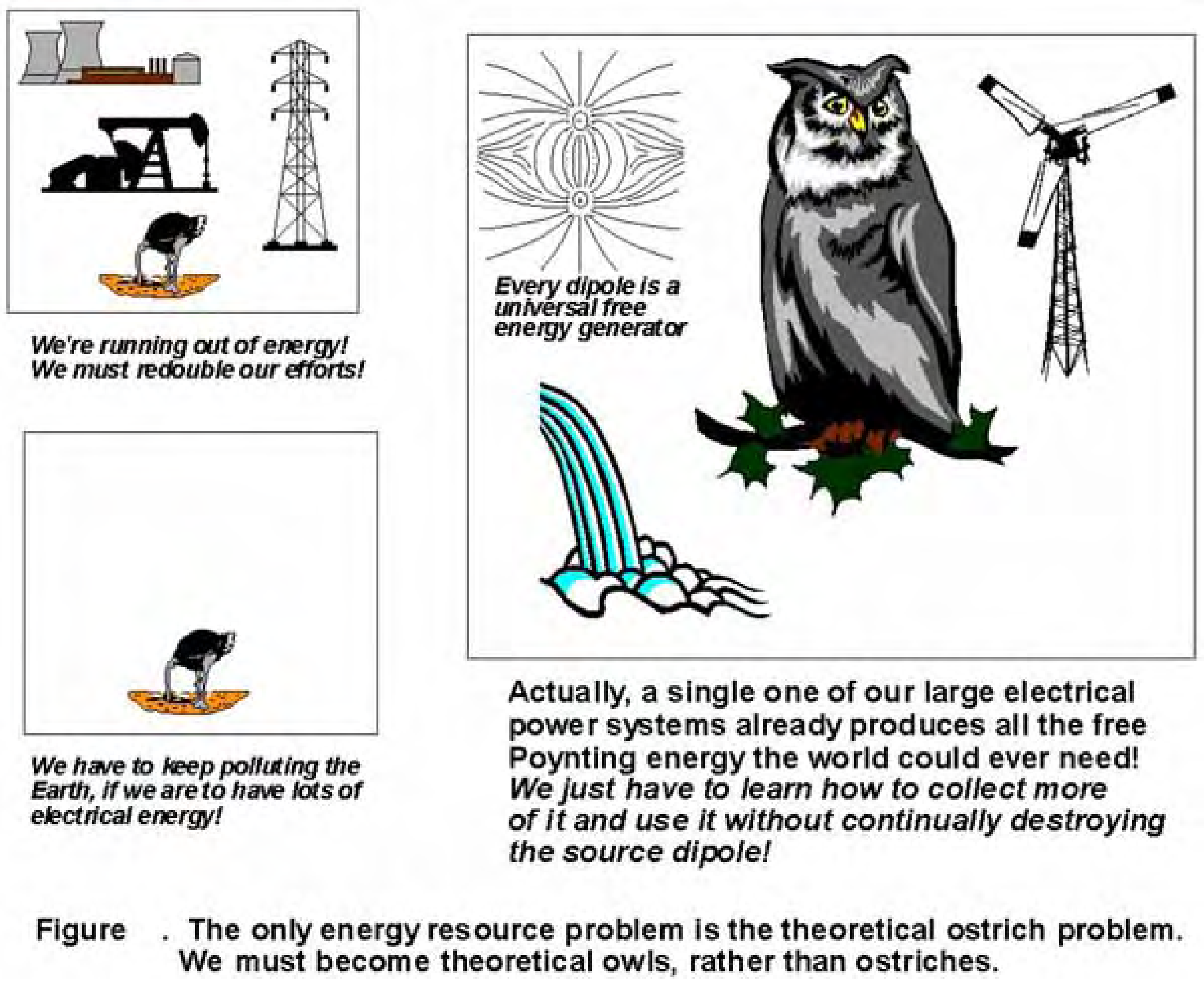

We're running out of energy!
We must redouble our efforts!

We have to keep polluting the Earth, if we are to have lots of electrical energy!

Every dipole is a universal free energy generator

Actually, a single one of our large electrical power systems already produces all the free Poynting energy the world could ever need! *We just have to learn how to collect more of it and use it without continually destroying the source dipole!*

Figure . The only energy resource problem is the theoretical ostrich problem. We must become theoretical owls, rather than ostriches.

Return to Slide Index

Next Slide

Previous Slide

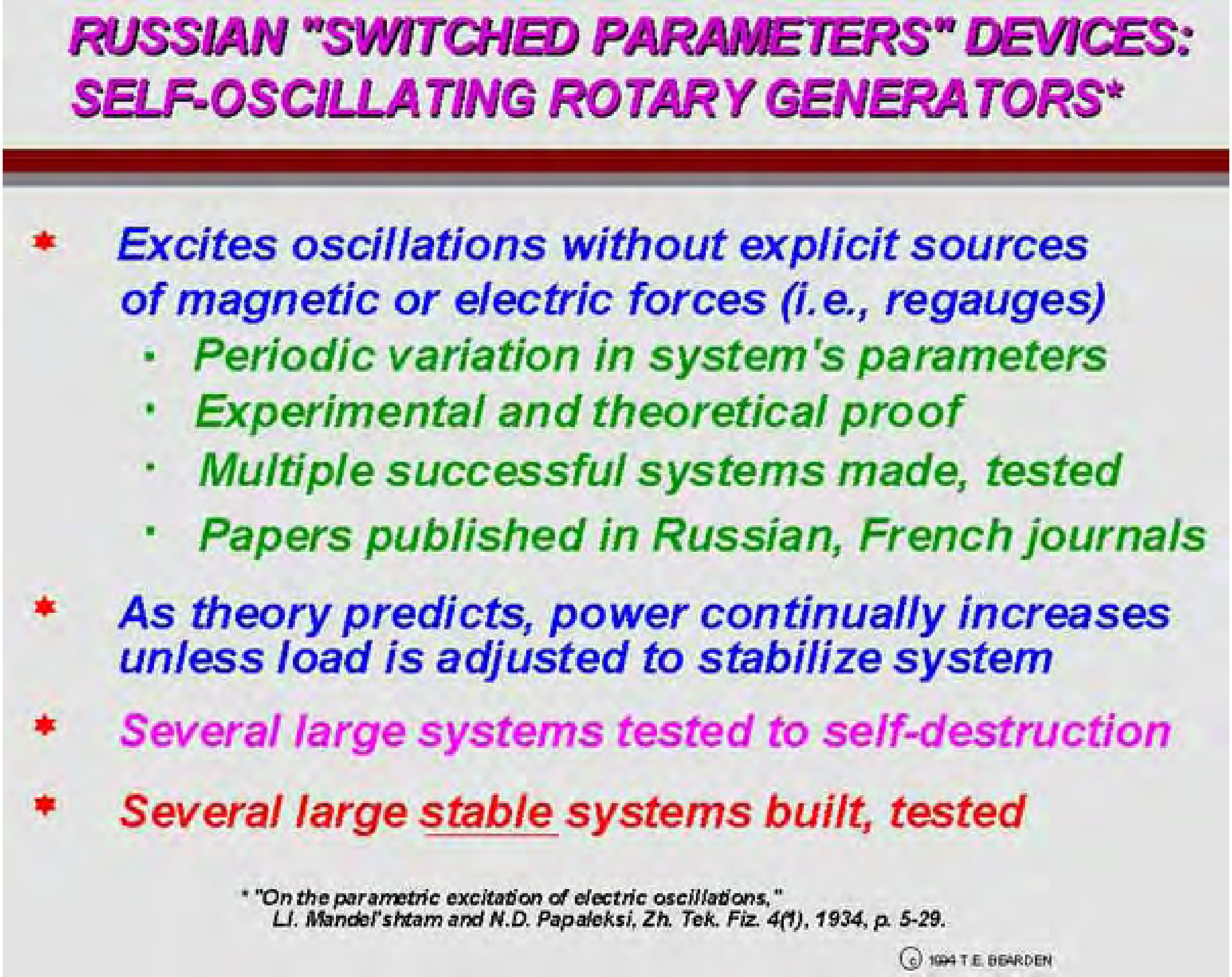

RUSSIAN "SWITCHED PARAMETERS" DEVICES: SELF-OSCILLATING ROTARY GENERATORS*

- Excites oscillations without explicit sources of magnetic or electric forces (i.e., regauges)
  - Periodic variation in system's parameters
  - Experimental and theoretical proof
  - Multiple successful systems made, tested
  - Papers published in Russian, French journals
- As theory predicts, power continually increases unless load is adjusted to stabilize system
- Several large systems tested to self-destruction
- Several large stable systems built, tested

* "On the parametric excitation of electric oscillations," L.I. Mandel'shtam and N.D. Papaleksi, Zh. Tek. Fiz. 4(1), 1934, p. 5-29.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

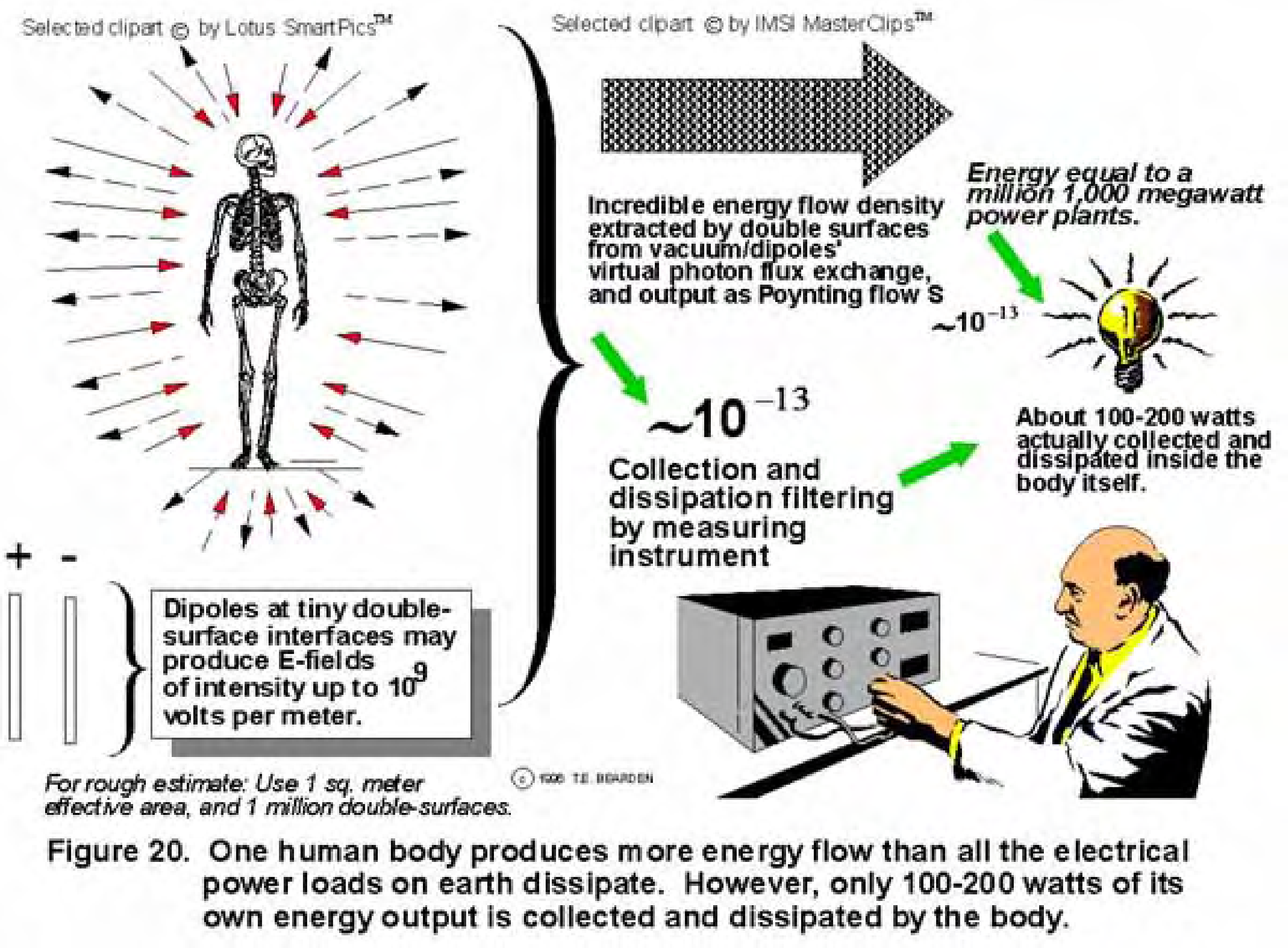
Selected clipart © by Lotus SmartPics™
Selected clipart © by IMSI MasterClips™
+
-
~10^{-13}
© [illegible] T.E. BEARDEN

Incredible energy flow density extracted by double surfaces from vacuum/dipoles' virtual photon flux exchange, and output as Poynting flow S

Energy equal to a million 1,000 megawatt power plants.

~10^{-13}

Collection and dissipation filtering by measuring instrument

About 100-200 watts actually collected and dissipated inside the body itself.

Dipoles at tiny double-surface interfaces may produce E-fields of intensity up to 10^9 volts per meter.

For rough estimate: Use 1 sq. meter effective area, and 1 million double-surfaces.

Figure 20. One human body produces more energy flow than all the electrical power loads on earth dissipate. However, only 100-200 watts of its own energy output is collected and dissipated by the body.

Return to Slide Index

Next Slide

Previous Slide

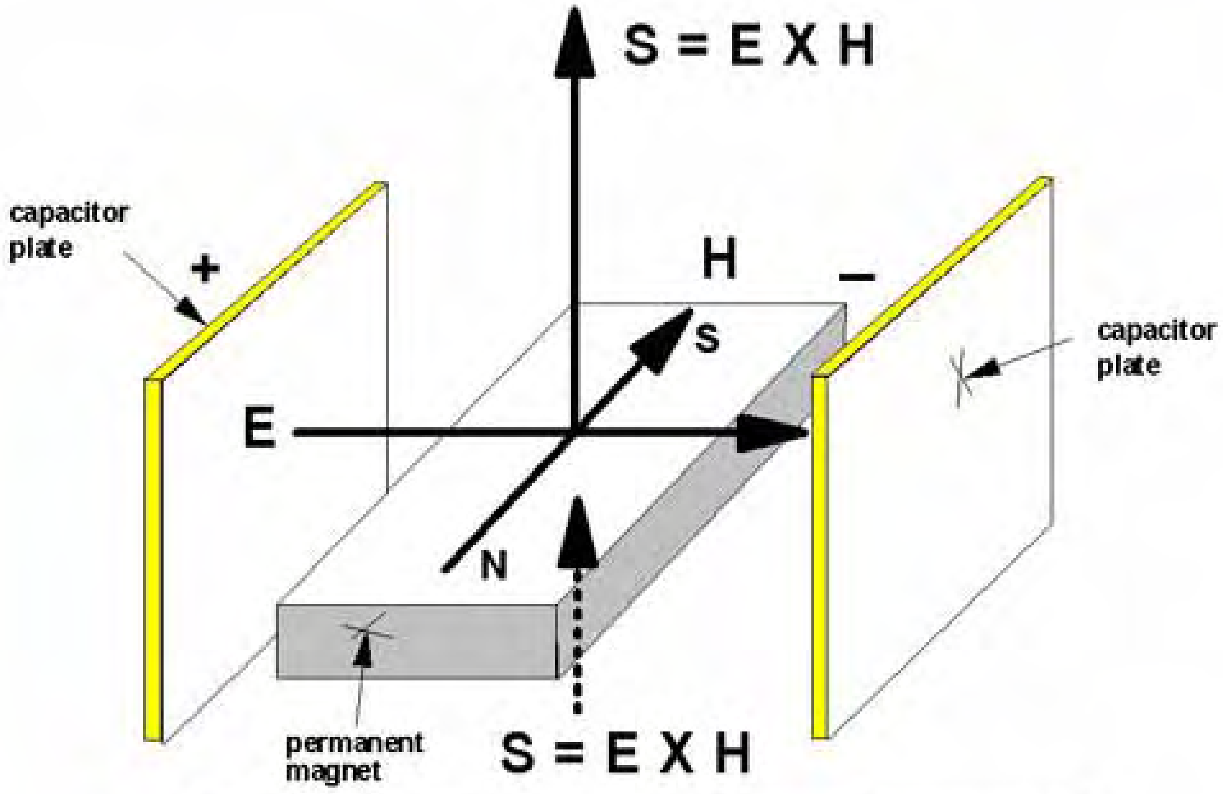
S = E X H
capacitor plate
+
H
–
S
capacitor plate
E
N
permanent magnet
S = E X H

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

© 1996 T.E. BEARDEN

Figure 18. A "static" Poynting S-flow free-energy generator.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

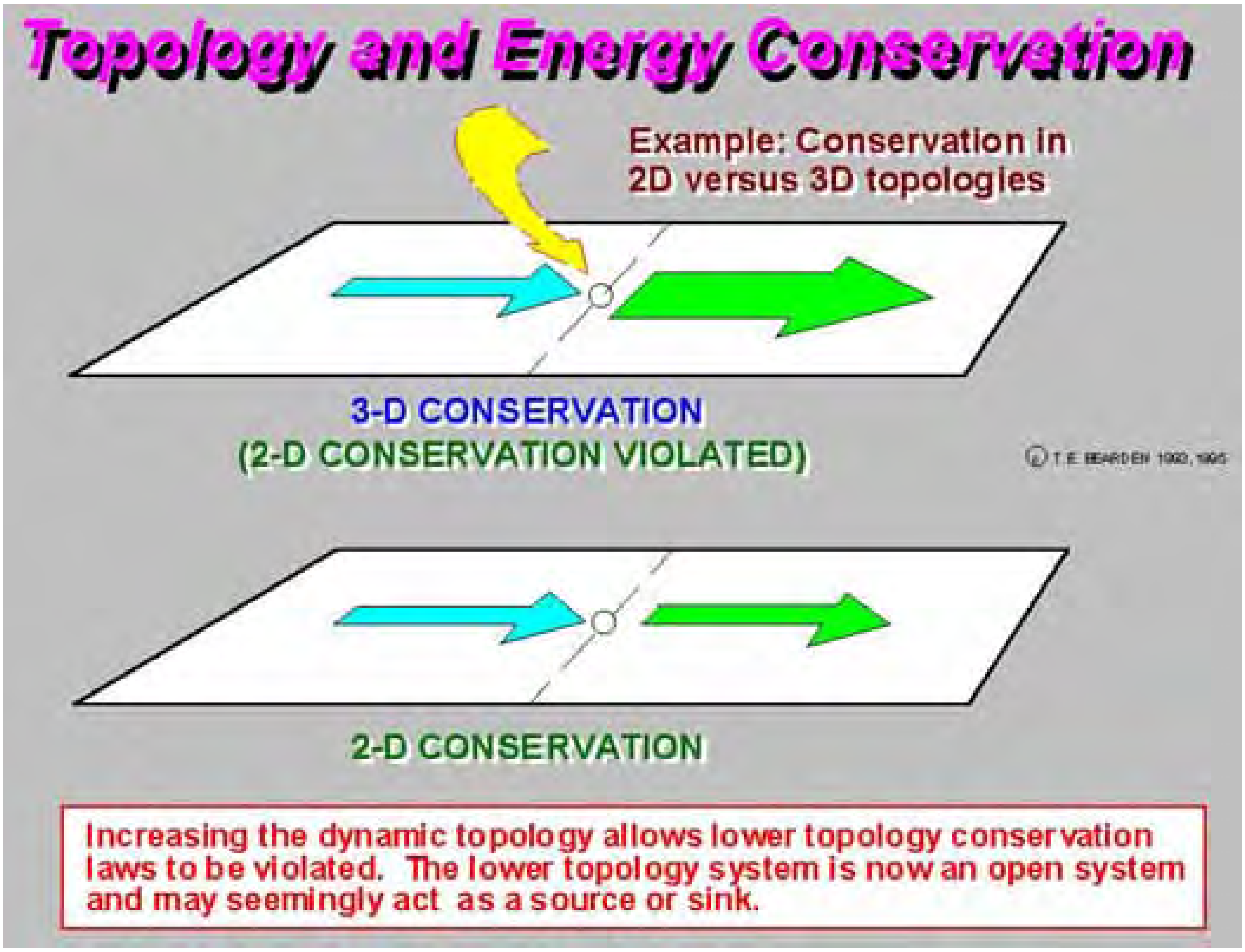

Topology and Energy Conservation

Example: Conservation in
2D versus 3D topologies

3-D CONSERVATION
(2-D CONSERVATION VIOLATED)

© T. E. BEARDEN 1992, 1995

2-D CONSERVATION

Increasing the dynamic topology allows lower topology conservation laws to be violated. The lower topology system is now an open system and may seemingly act as a source or sink.

Return to Slide Index

Next Slide

Previous Slide

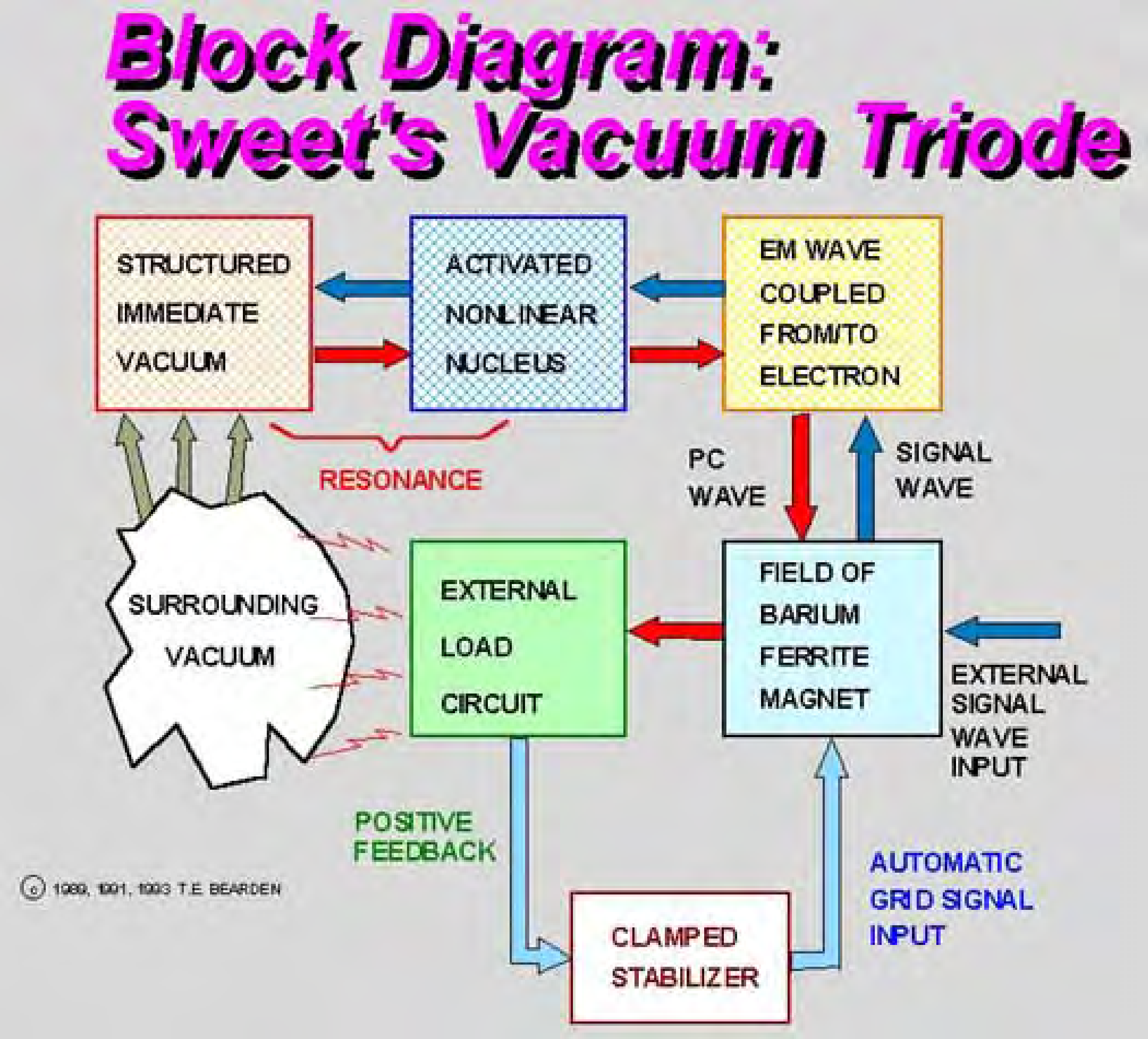

Block Diagram: Sweet's Vacuum Triode

STRUCTURED
IMMEDIATE
VACUUM

ACTIVATED
NONLINEAR
NUCLEUS

EM WAVE
COUPLED
FROM/TO
ELECTRON

RESONANCE

PC
WAVE

SIGNAL
WAVE

SURROUNDING
VACUUM

EXTERNAL
LOAD
CIRCUIT

FIELD OF
BARIUM
FERRITE
MAGNET

EXTERNAL
SIGNAL
WAVE
INPUT

POSITIVE
FEEDBACK

AUTOMATIC
GRID SIGNAL
INPUT

CLAMPED
STABILIZER

© 1989, 1991, 1993 T.E. BEARDEN

Return to Slide Index

Previous Slide

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

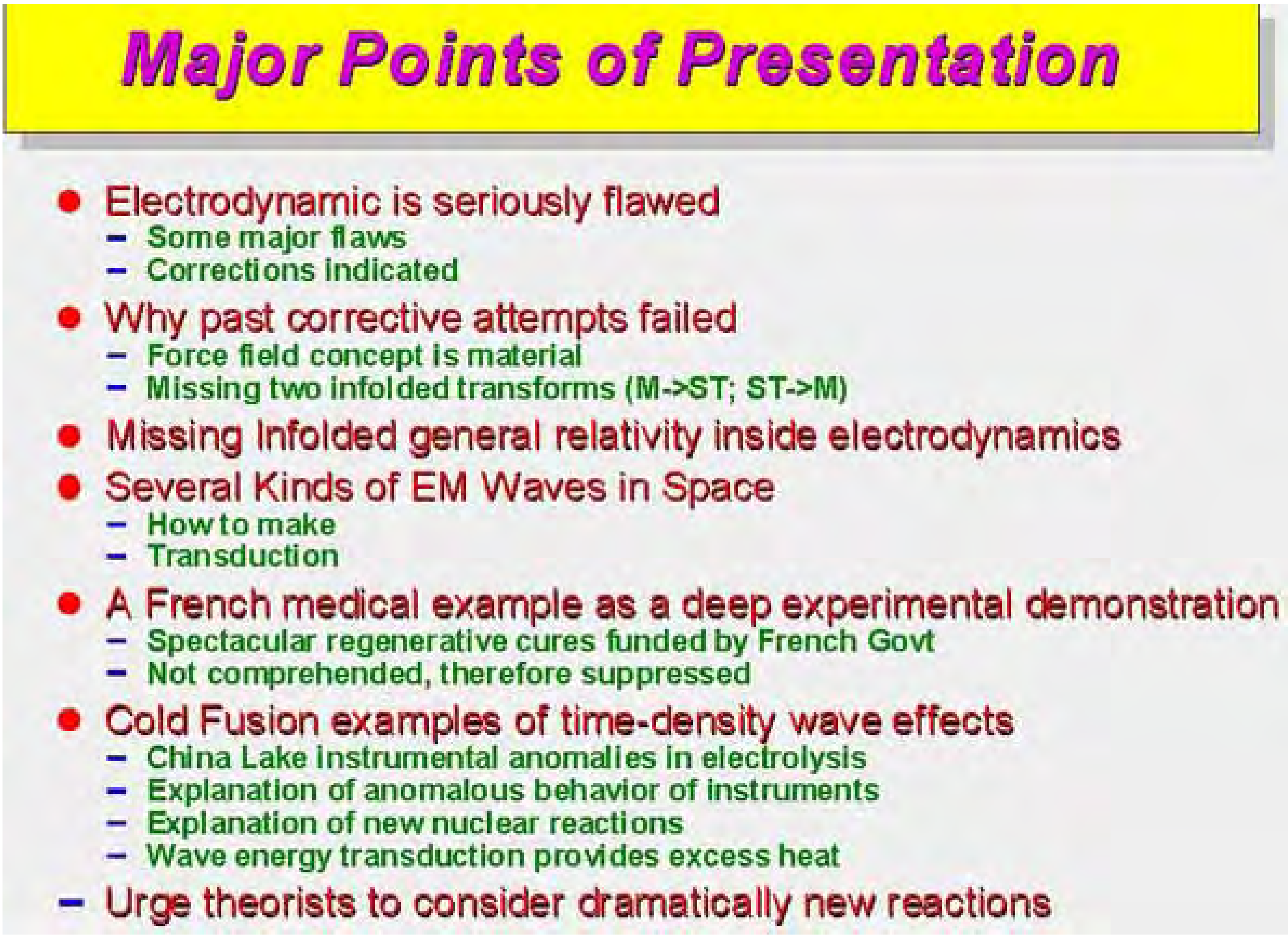

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

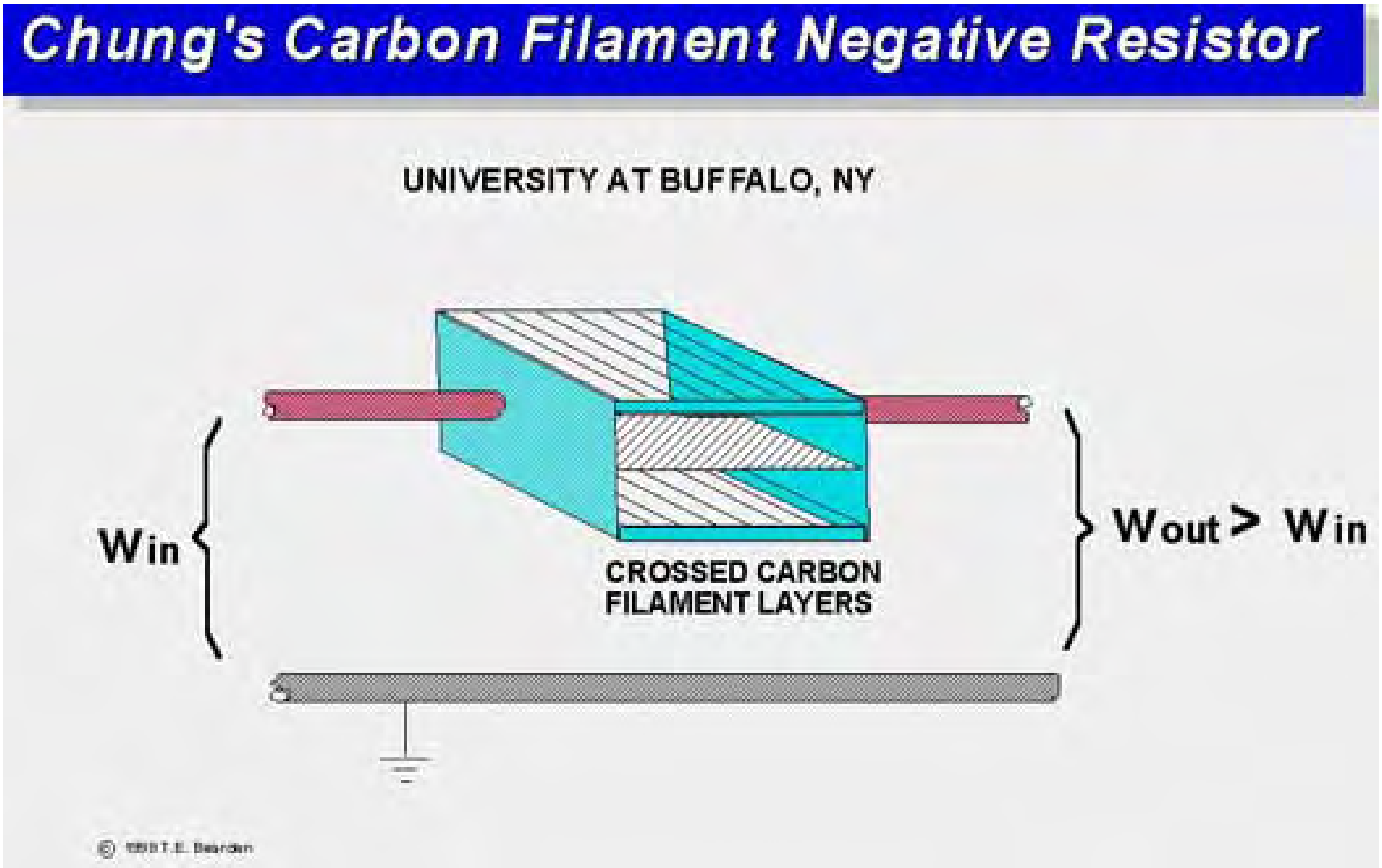

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

Win
CROSSED CARBON
FILAMENT LAYERS
Wout > Win

© [illegible] T.E. Bearden

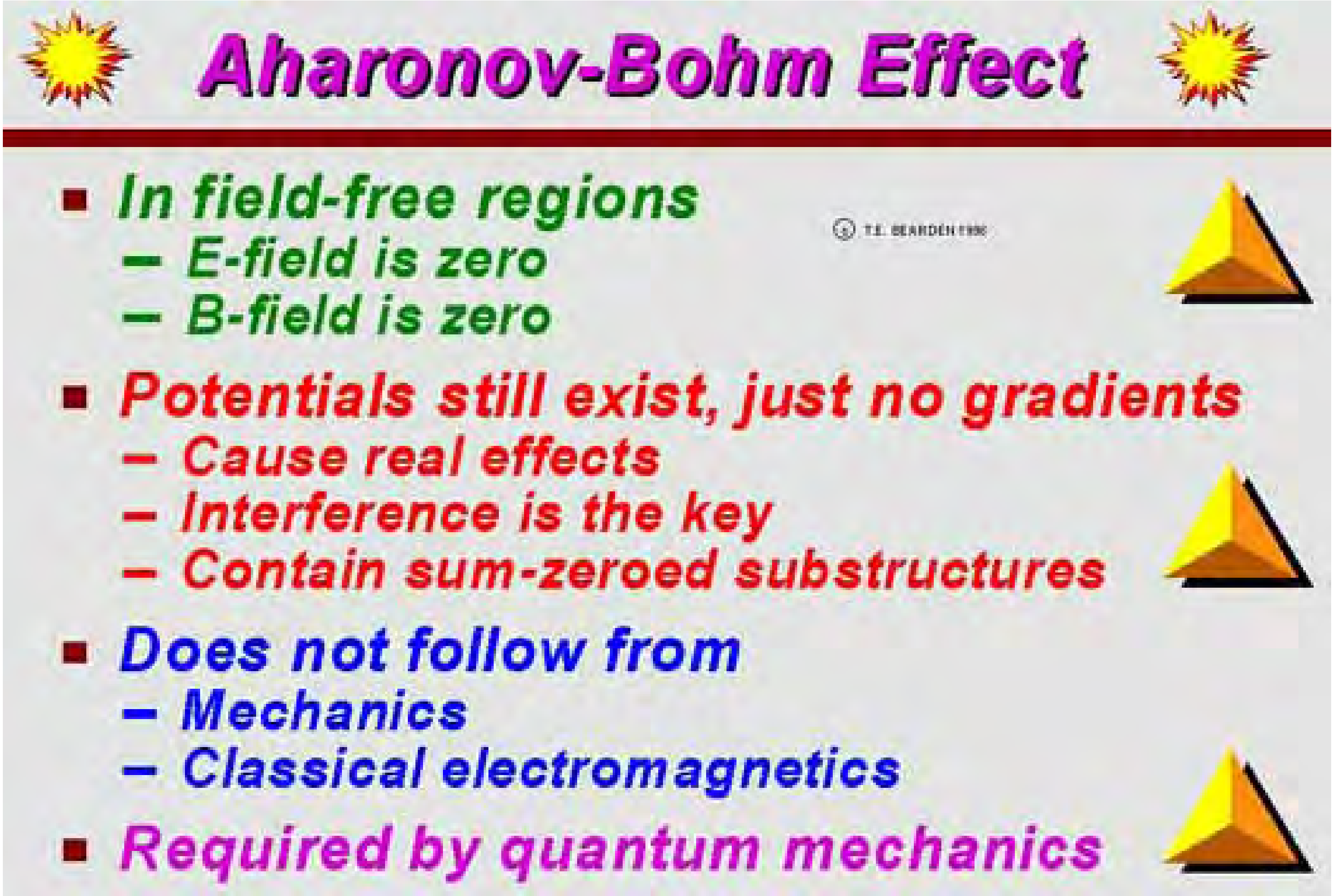

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

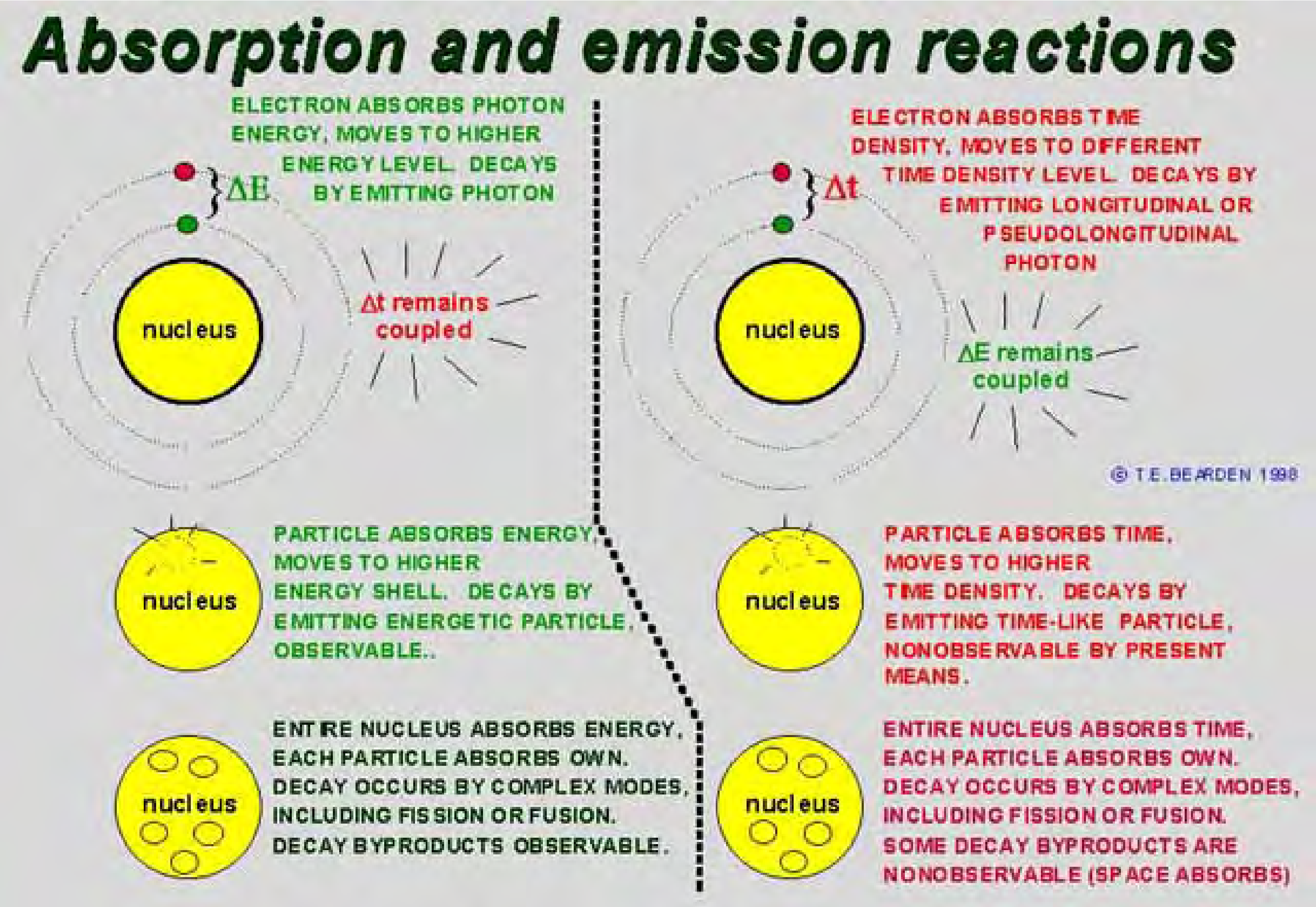

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

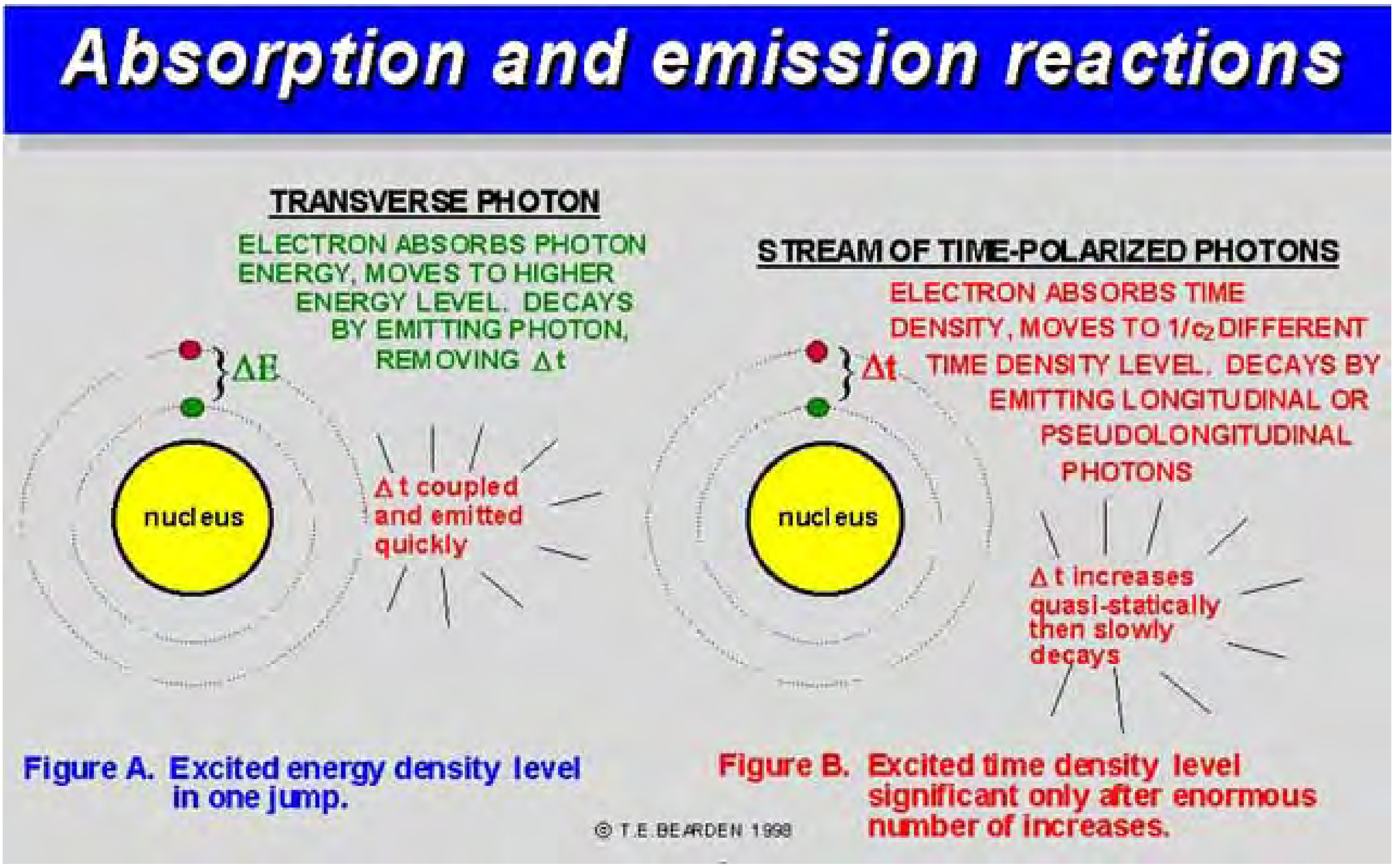

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

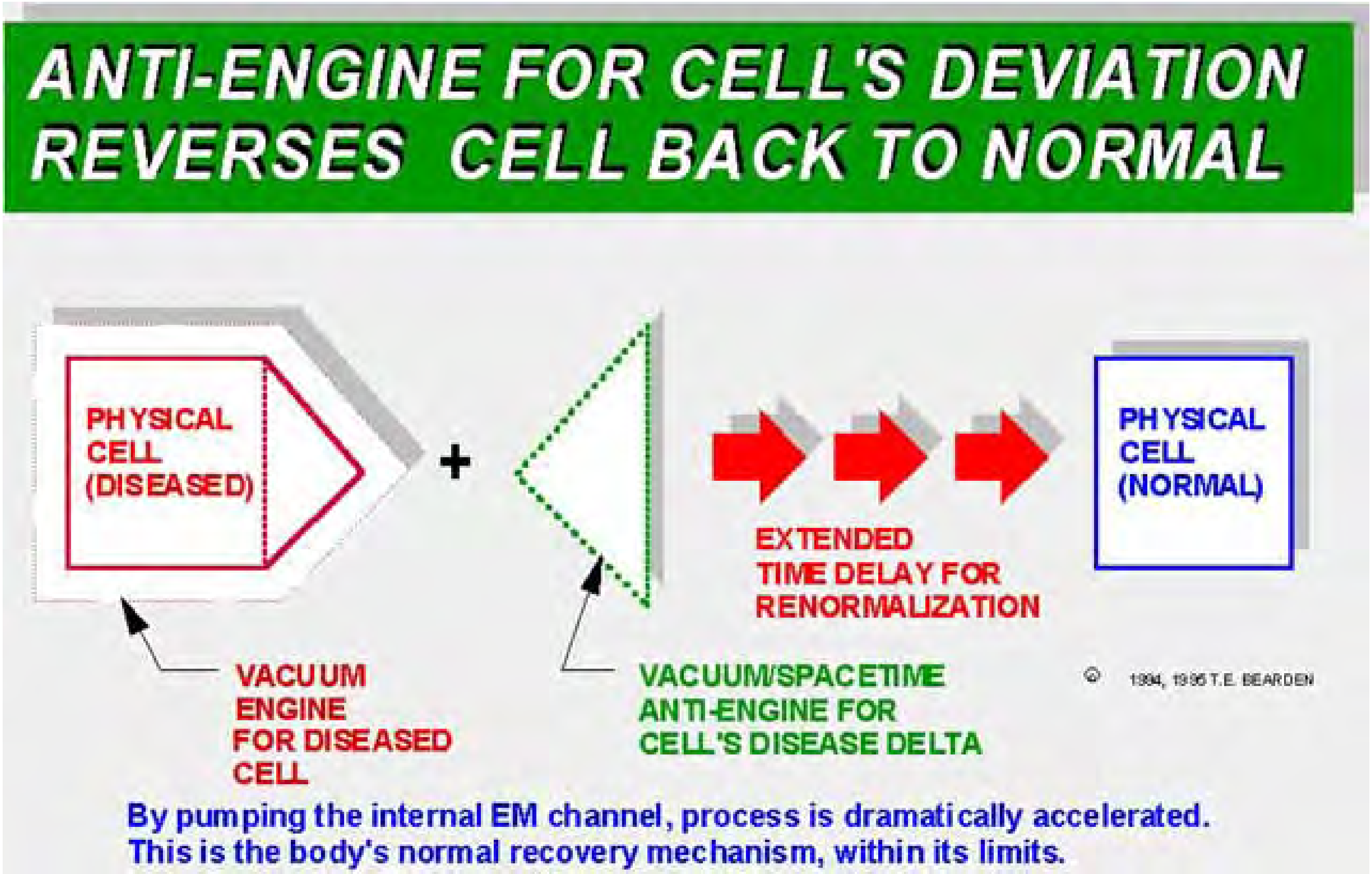

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)

+

EXTENDED
TIME DELAY FOR
RENORMALIZATION

PHYSICAL
CELL
(NORMAL)

VACUUM
ENGINE
FOR DISEASED
CELL

VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

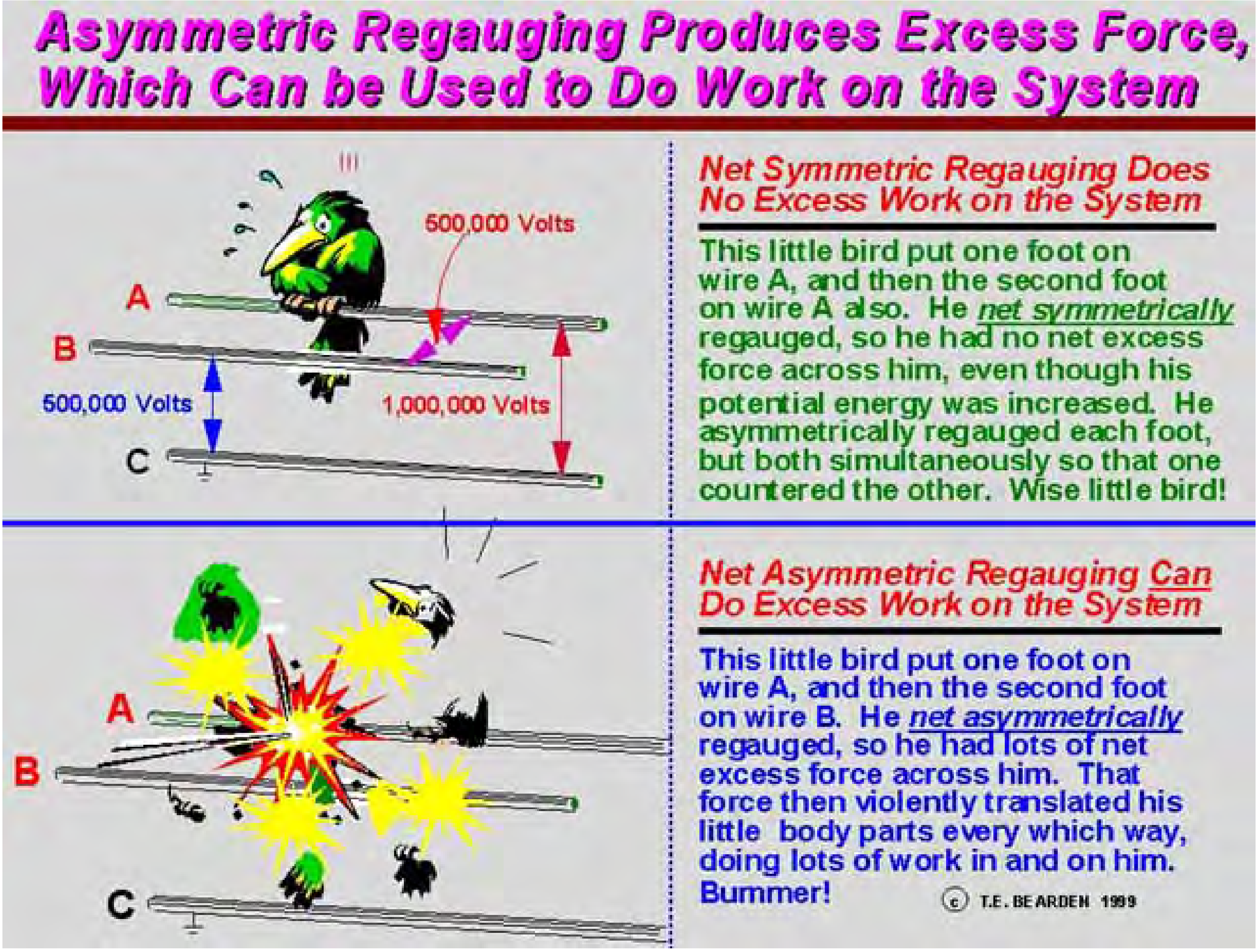

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

ⓒ T.E. BEARDEN 1999

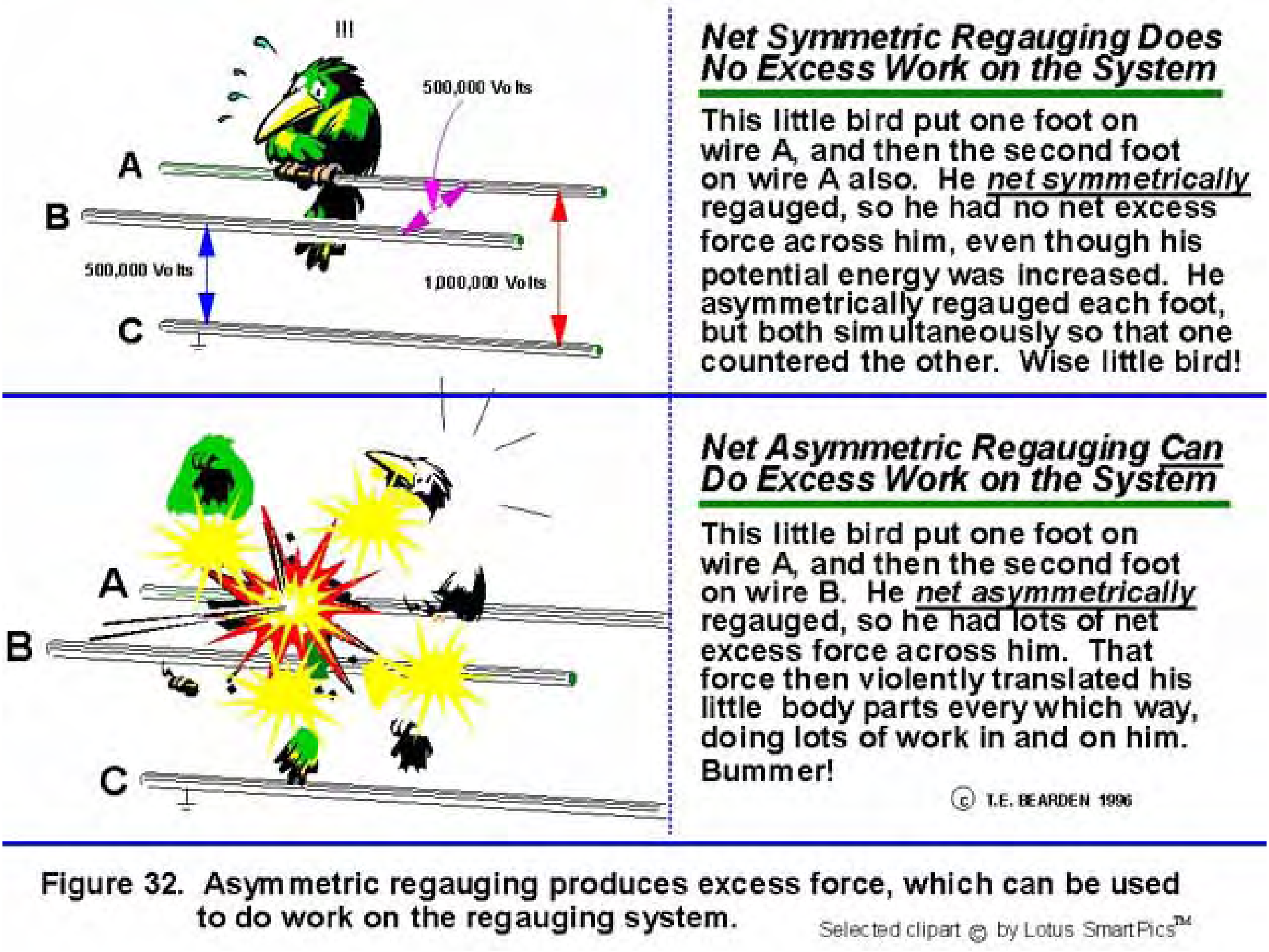

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

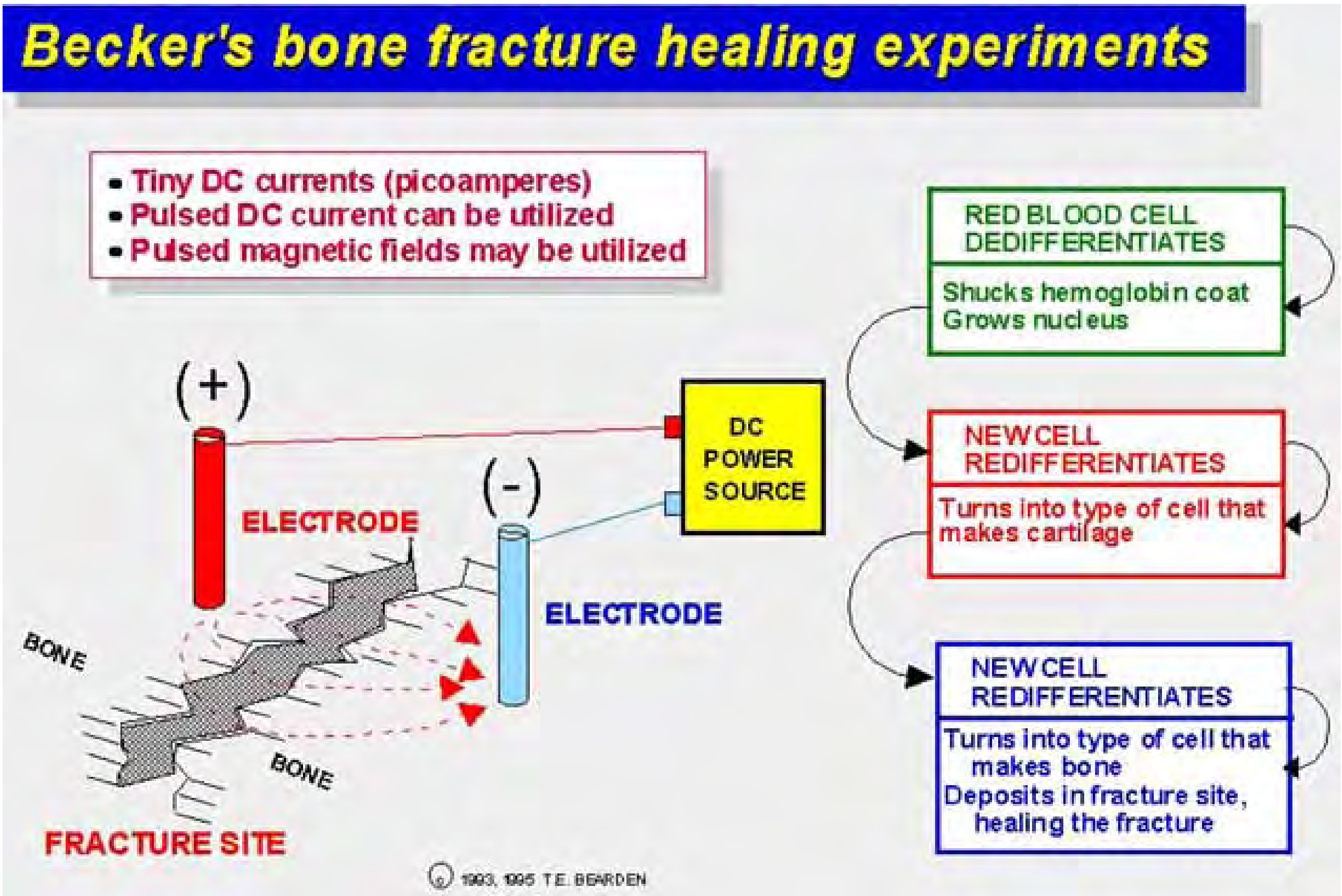

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)

ELECTRODE

(-)

ELECTRODE

DC
POWER
SOURCE

BONE

BONE

FRACTURE SITE

RED BLOOD CELL
DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL
REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL
REDIFFERENTIATES

Turns into type of cell that makes bone

Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

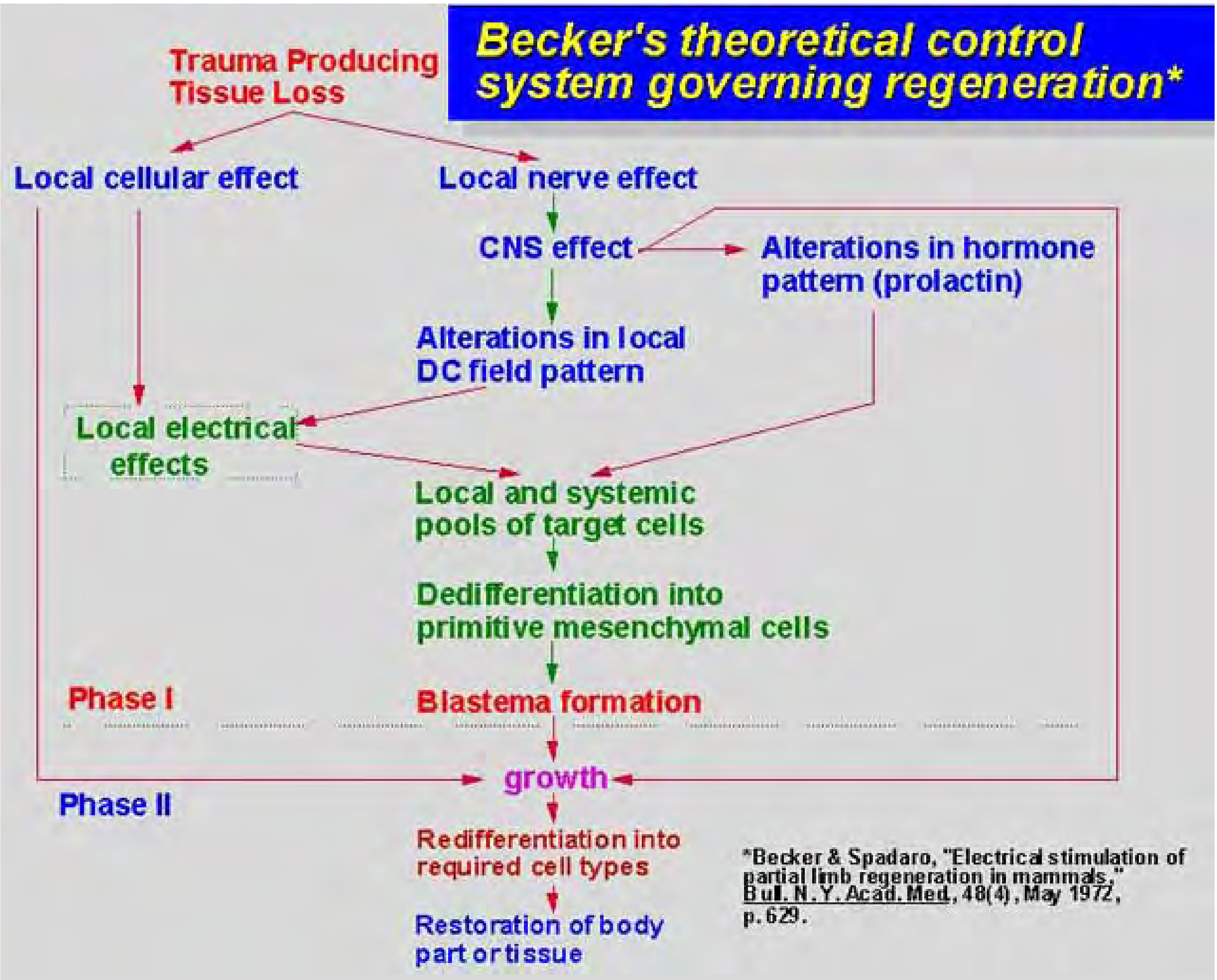

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

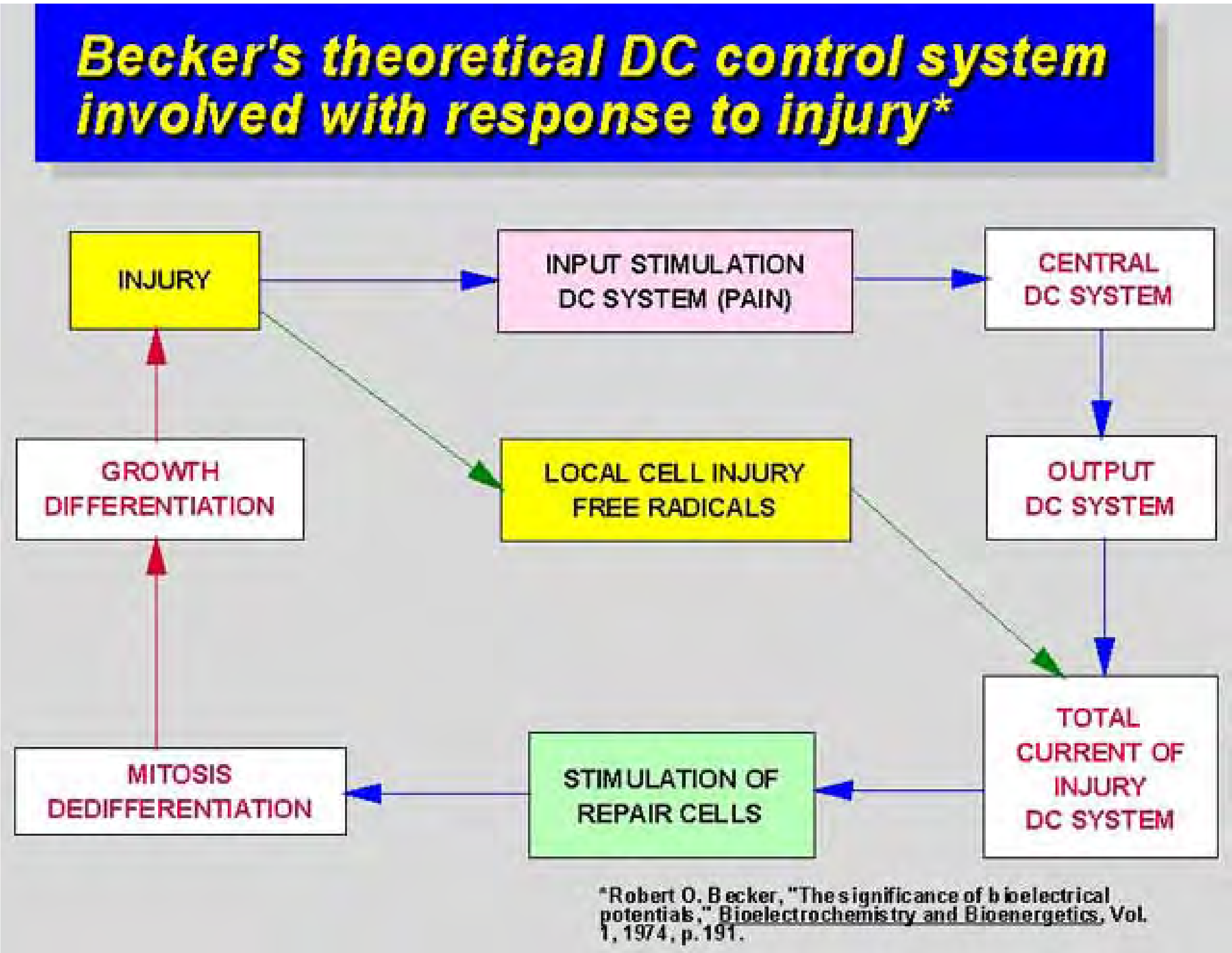

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION
DC SYSTEM (PAIN)

CENTRAL
DC SYSTEM

GROWTH
DIFFERENTIATION

LOCAL CELL INJURY
FREE RADICALS

OUTPUT
DC SYSTEM

TOTAL
CURRENT OF
INJURY
DC SYSTEM

STIMULATION OF
REPAIR CELLS

MITOSIS
DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

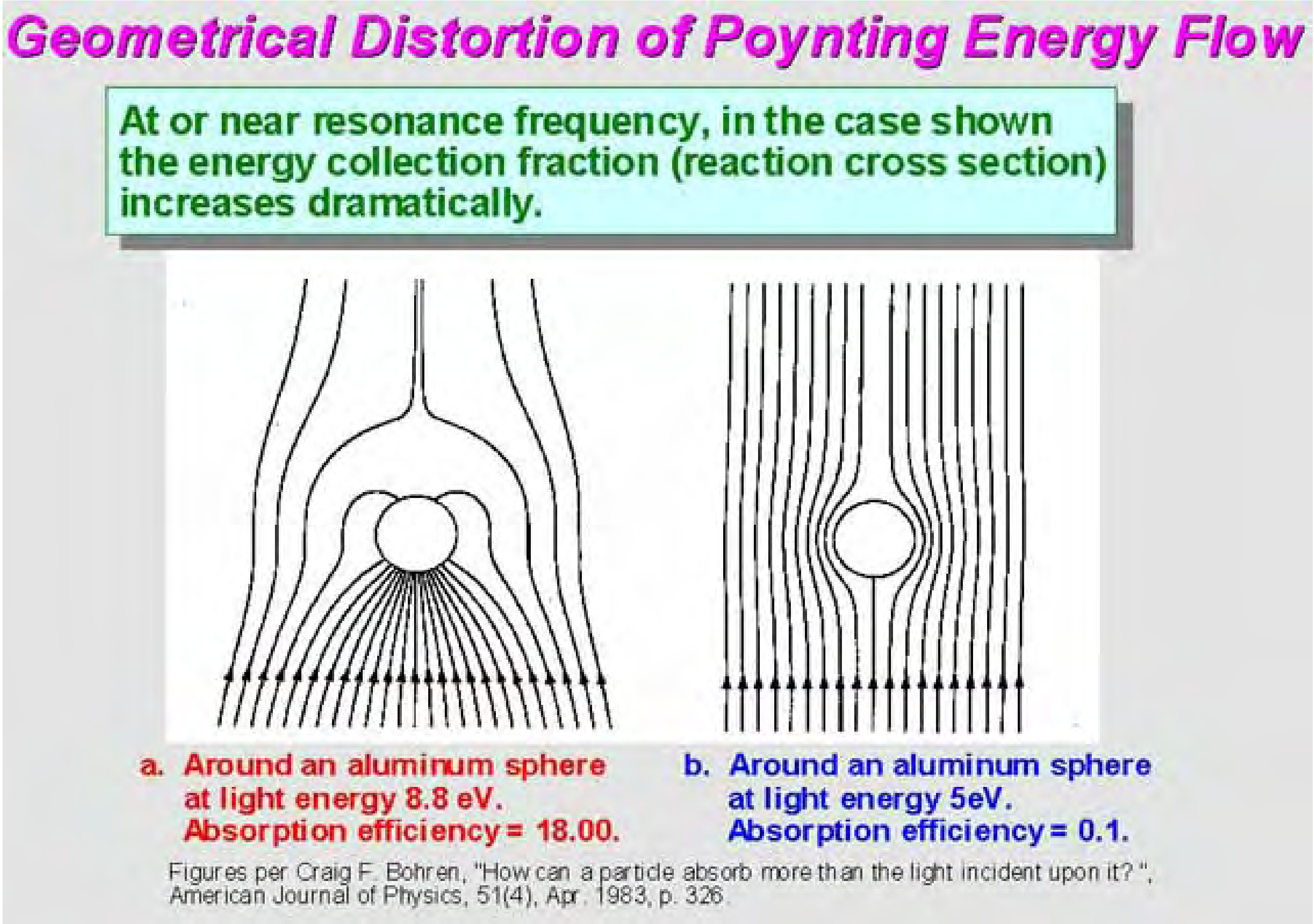

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it?", American Journal of Physics, 51(4), Apr. 1983, p. 326.

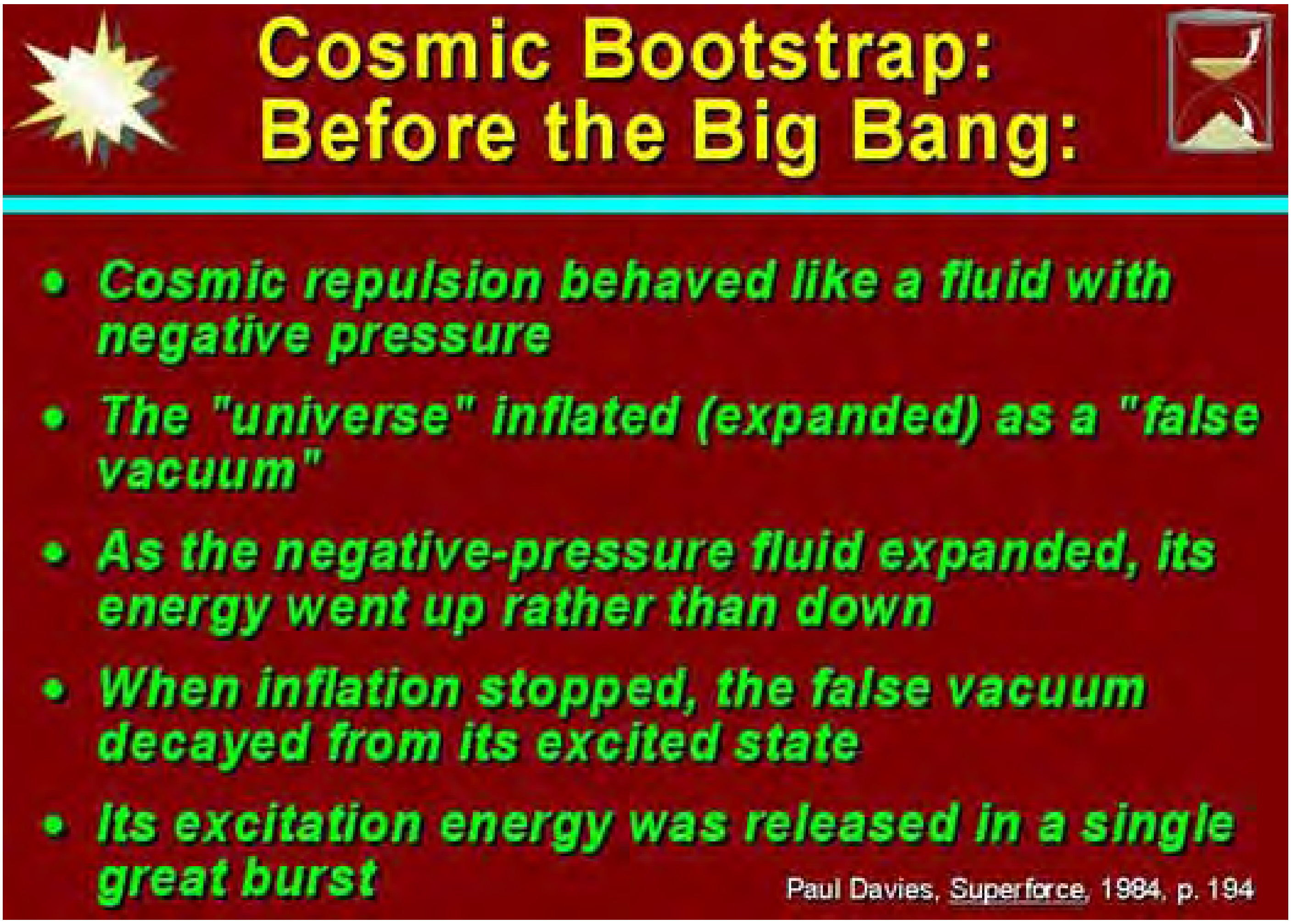

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

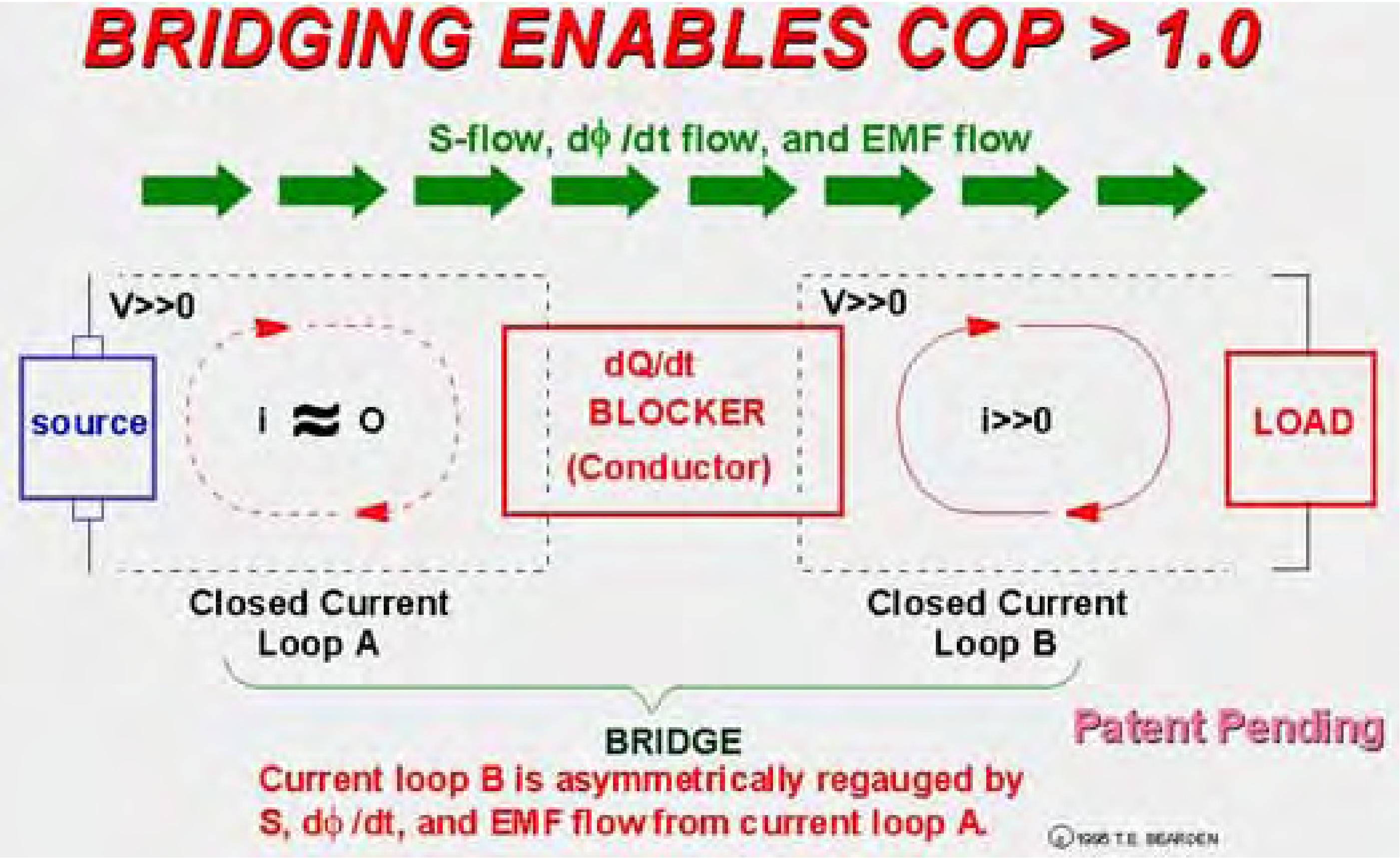

BRIDGING ENABLES COP > 1.0

S-flow, dϕ /dt flow, and EMF flow

V>>0

source

i ≈ 0

dQ/dt
BLOCKER
(Conductor)

V>>0

i>>0

LOAD

Closed Current
Loop A

Closed Current
Loop B

BRIDGE

Current loop B is asymmetrically regauged by
S, dϕ /dt, and EMF flow from current loop A.

Patent Pending

©1998 T.E. BEARDEN

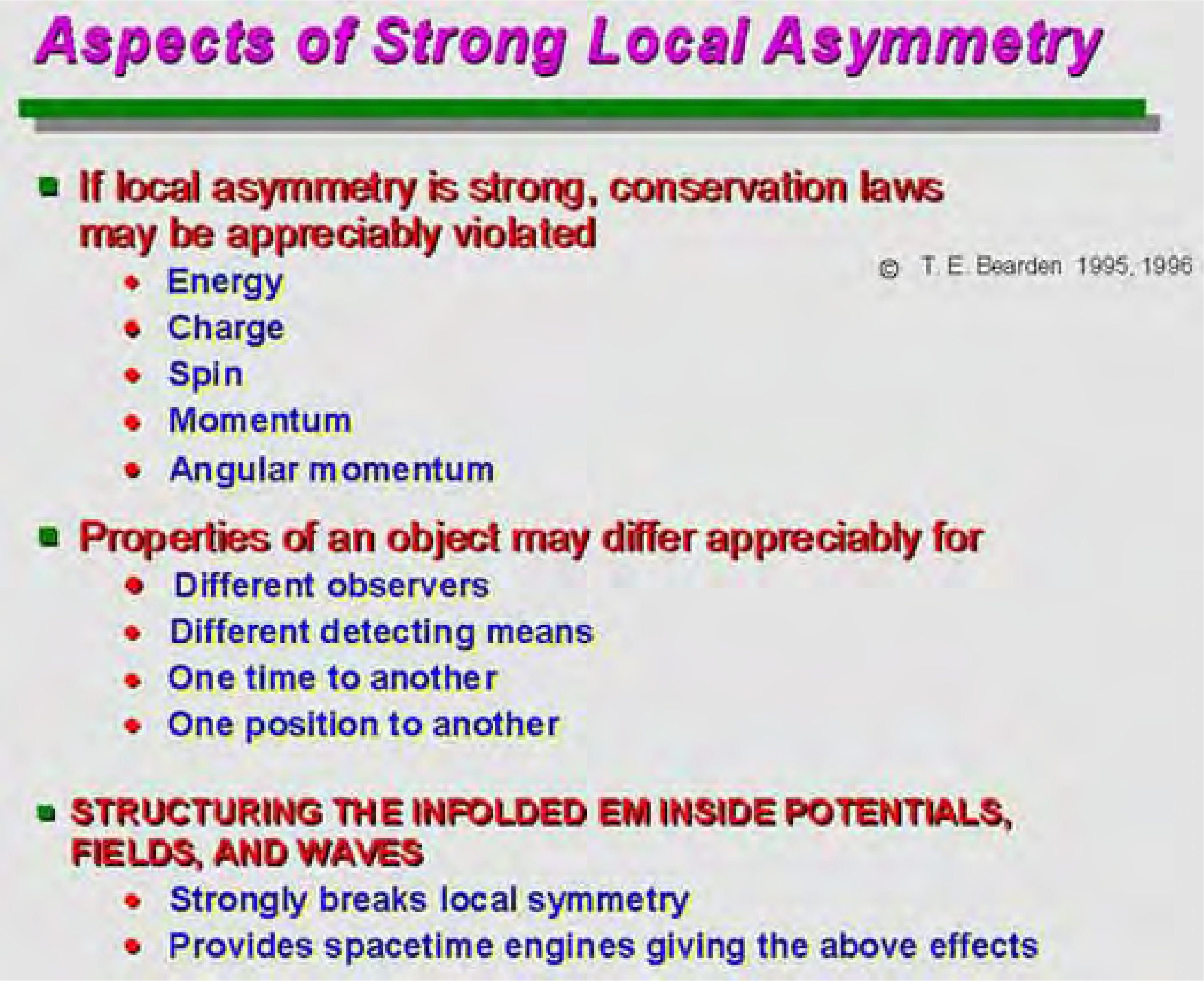

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
    - Energy
    - Charge
    - Spin
    - Momentum
    - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
    - Different observers
    - Different detecting means
    - One time to another
    - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
    - Strongly breaks local symmetry
    - Provides spacetime engines giving the above effects

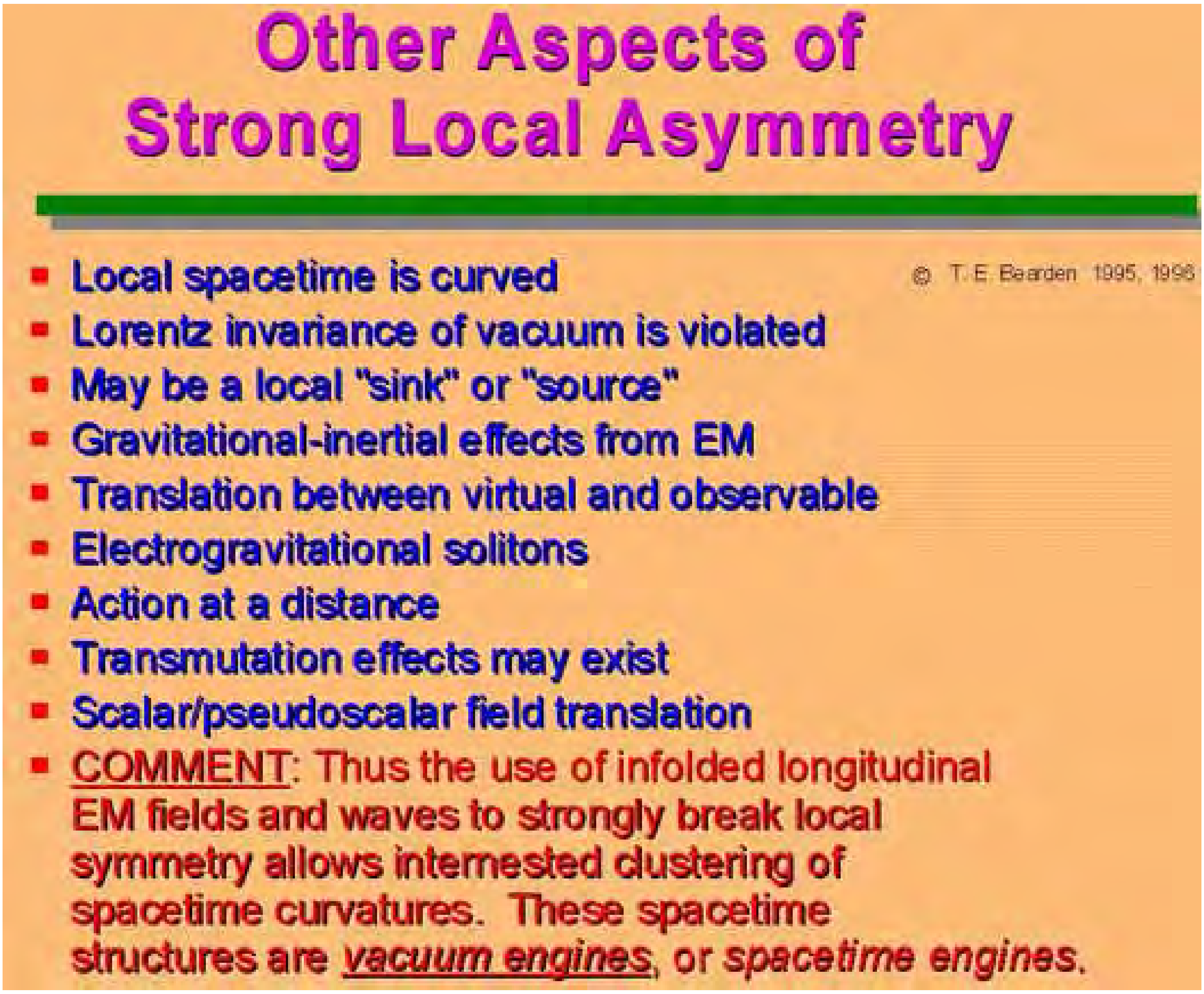

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

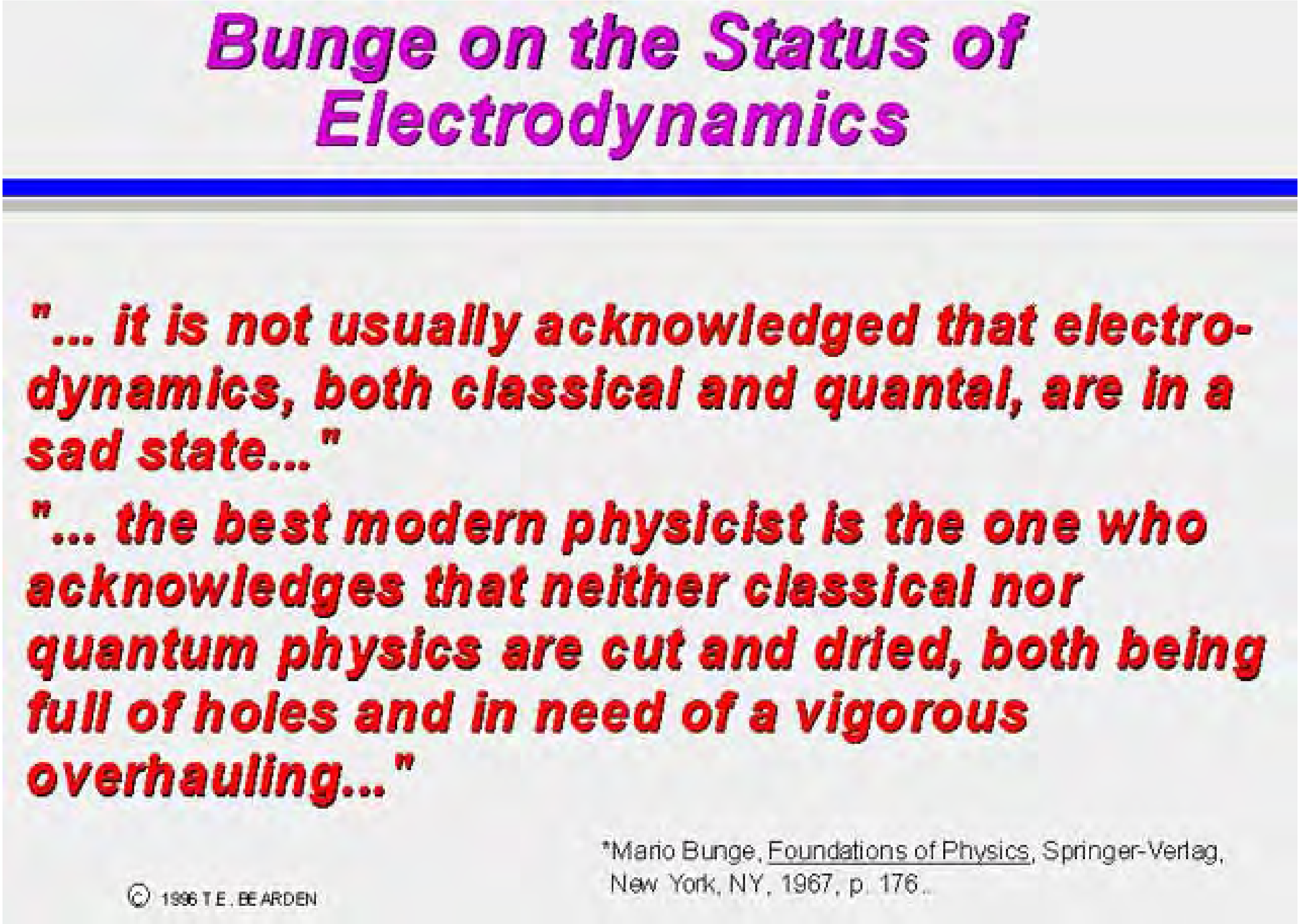

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

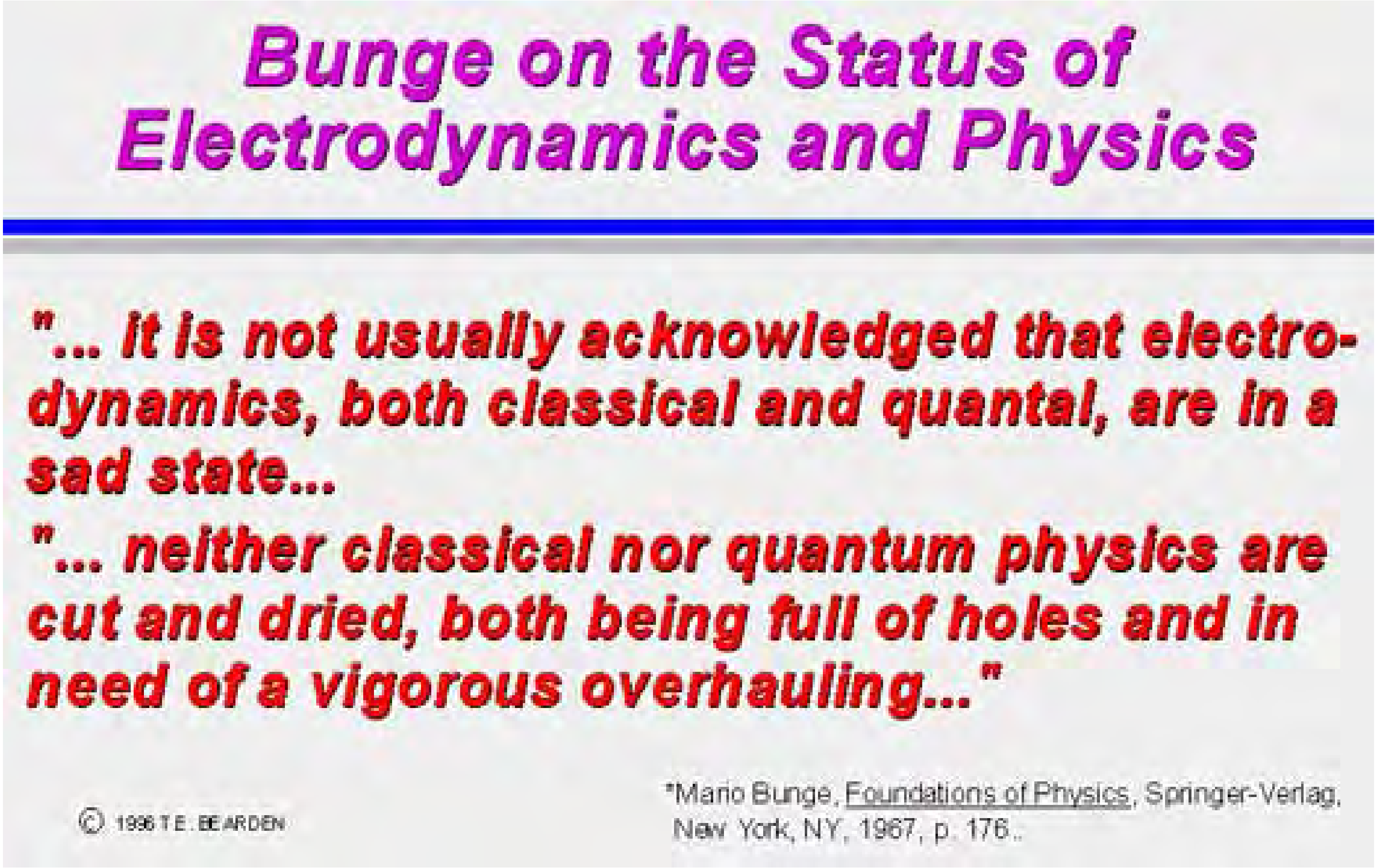

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

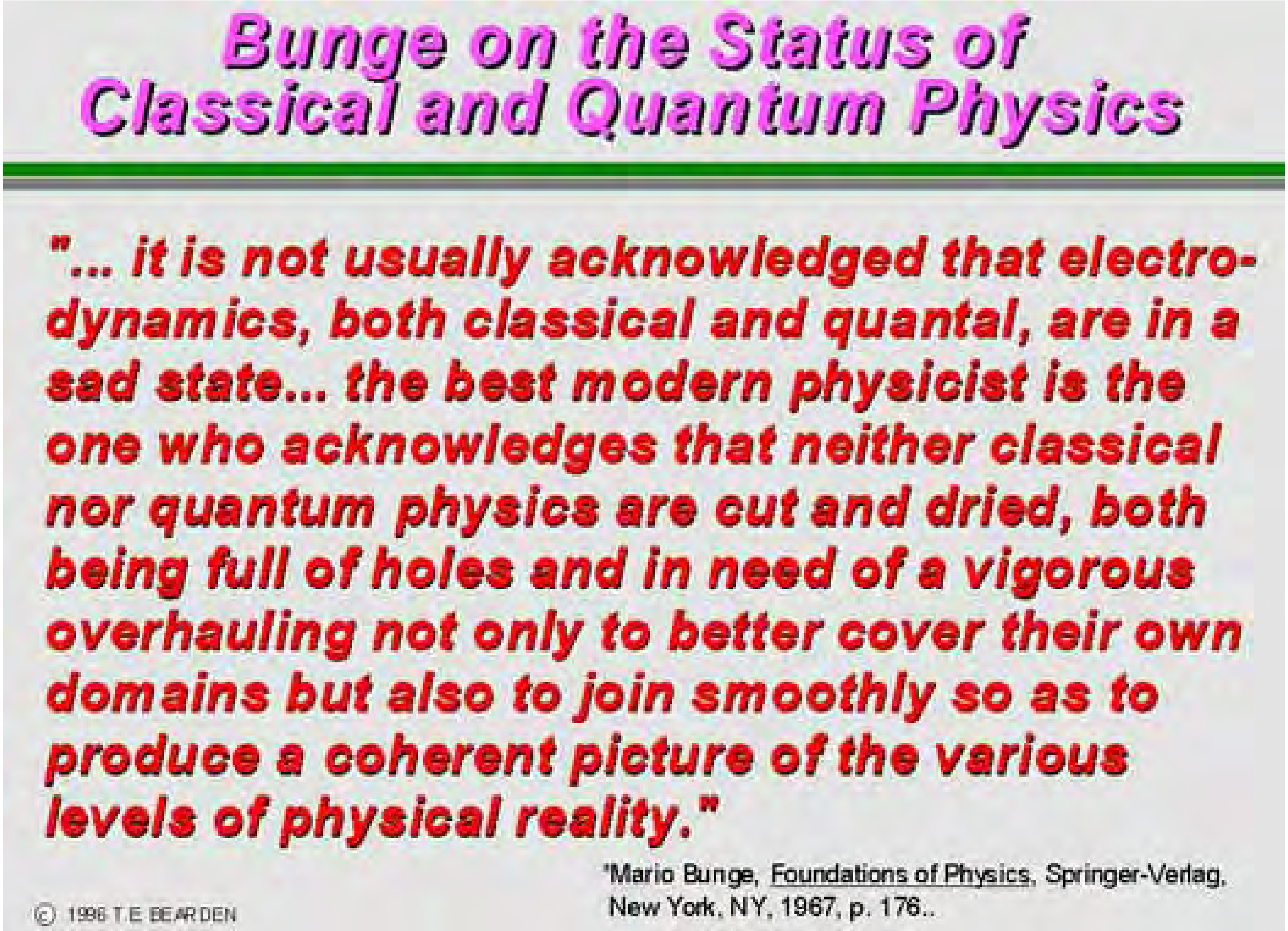

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

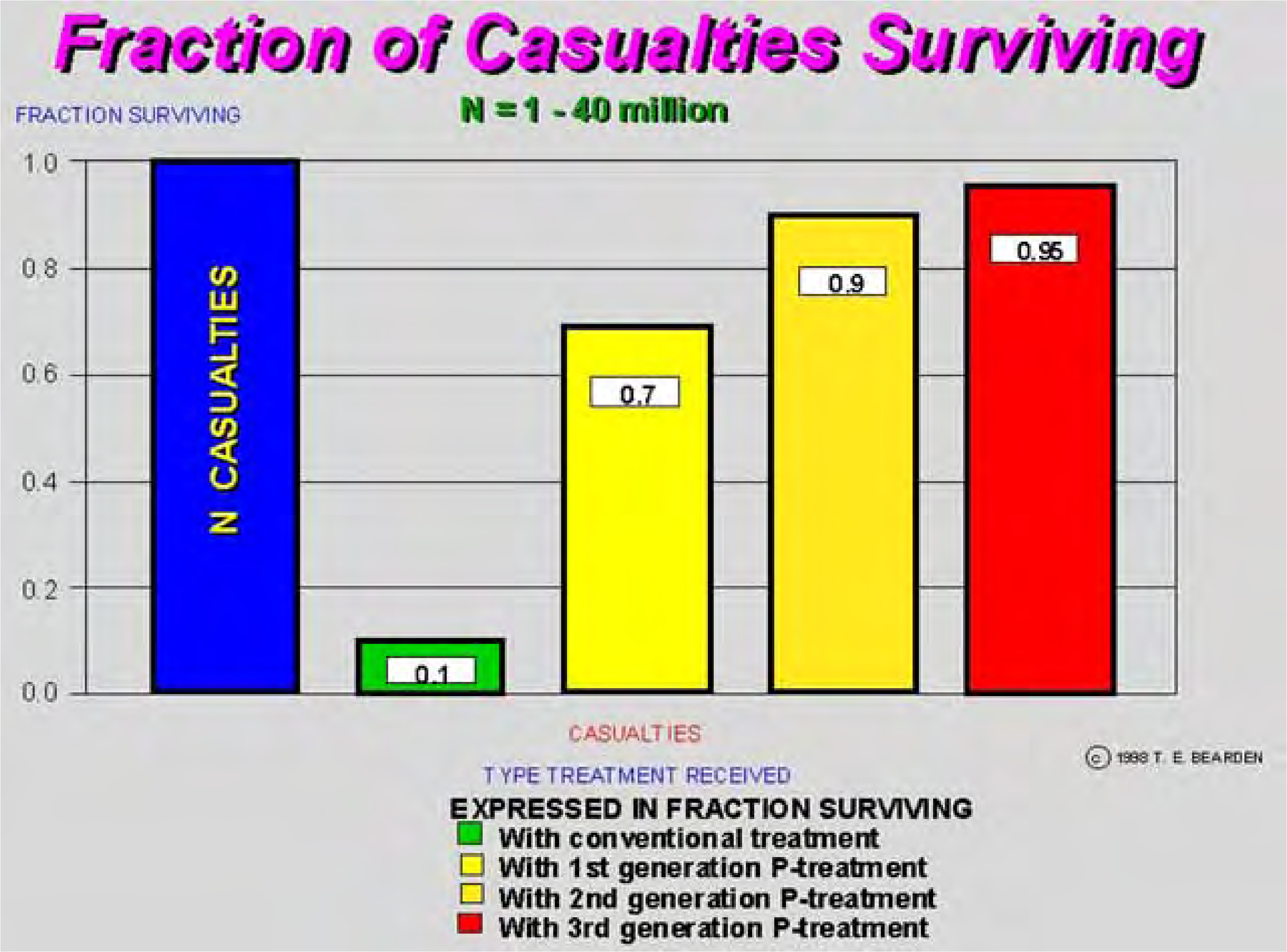
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

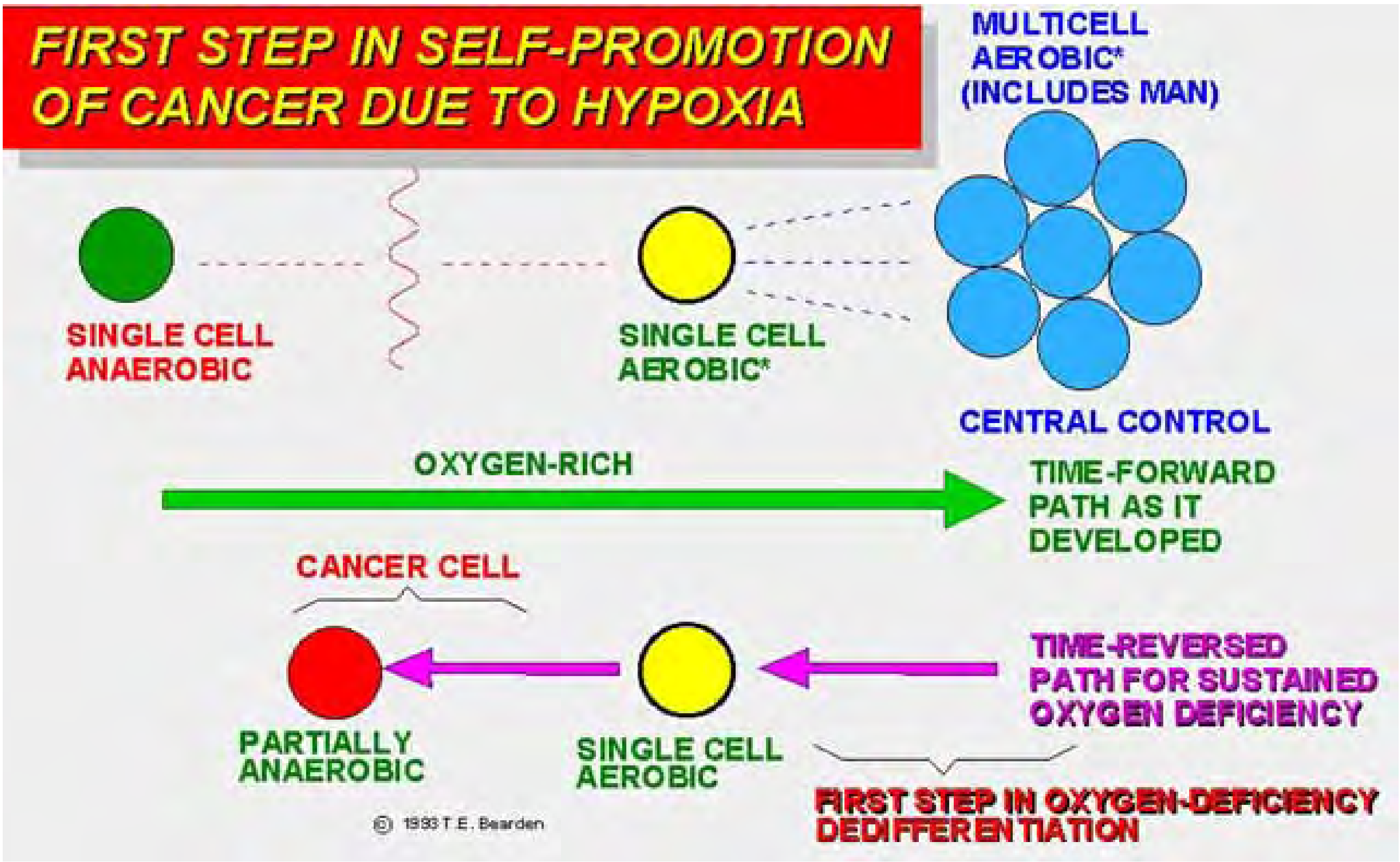

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

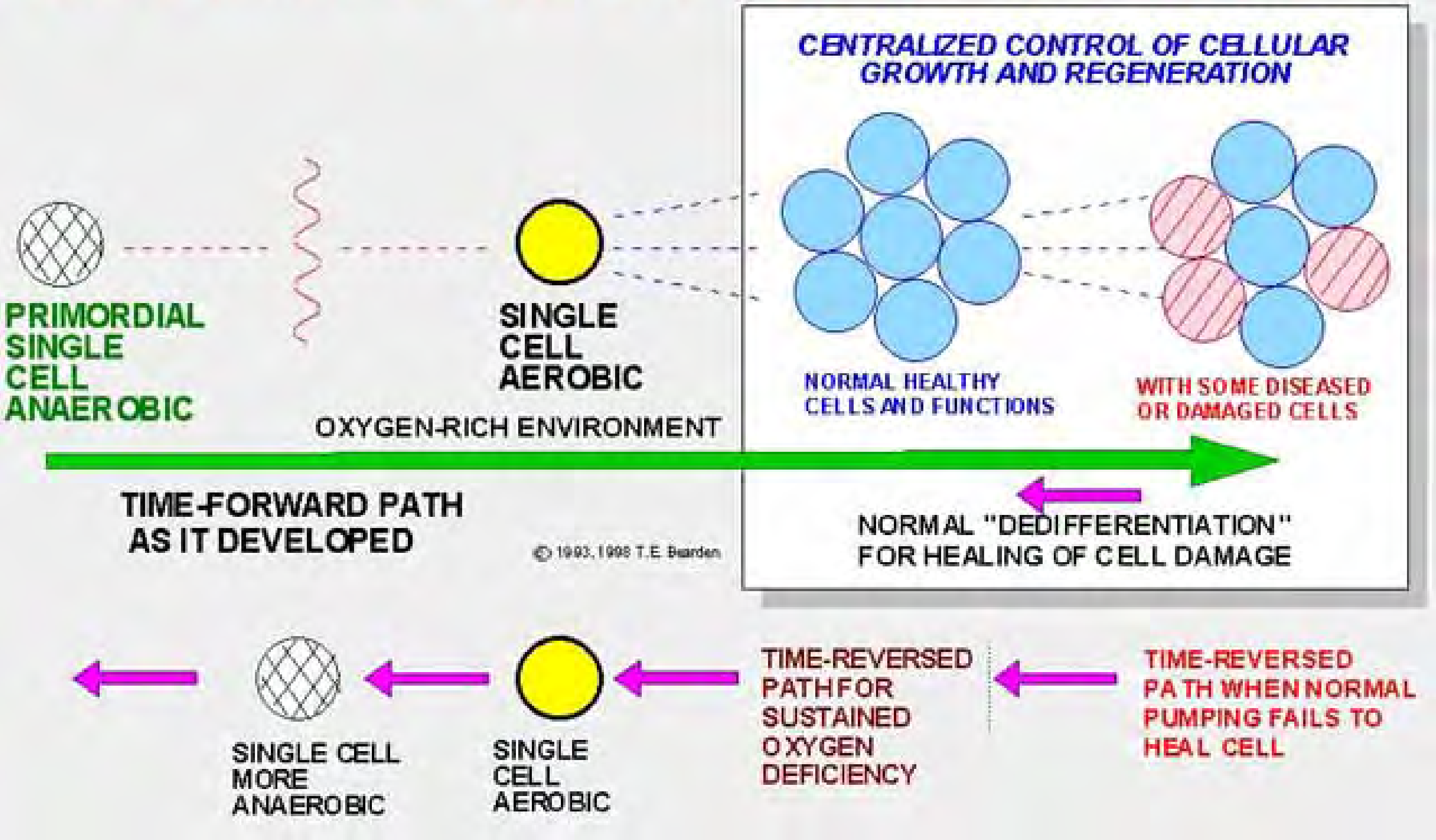

CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION

PRIMORDIAL
SINGLE
CELL
ANAEROBIC

SINGLE
CELL
AEROBIC

OXYGEN-RICH ENVIRONMENT

NORMAL HEALTHY
CELLS AND FUNCTIONS

WITH SOME DISEASED
OR DAMAGED CELLS

TIME-FORWARD PATH
AS IT DEVELOPED

© 1993, 1998 T.E. Bearden

NORMAL "DEDIFFERENTIATION"
FOR HEALING OF CELL DAMAGE

SINGLE CELL
MORE
ANAEROBIC

SINGLE
CELL
AEROBIC

TIME-REVERSED
PATH FOR
SUSTAINED
OXYGEN
DEFICIENCY

TIME-REVERSED
PATH WHEN NORMAL
PUMPING FAILS TO
HEAL CELL

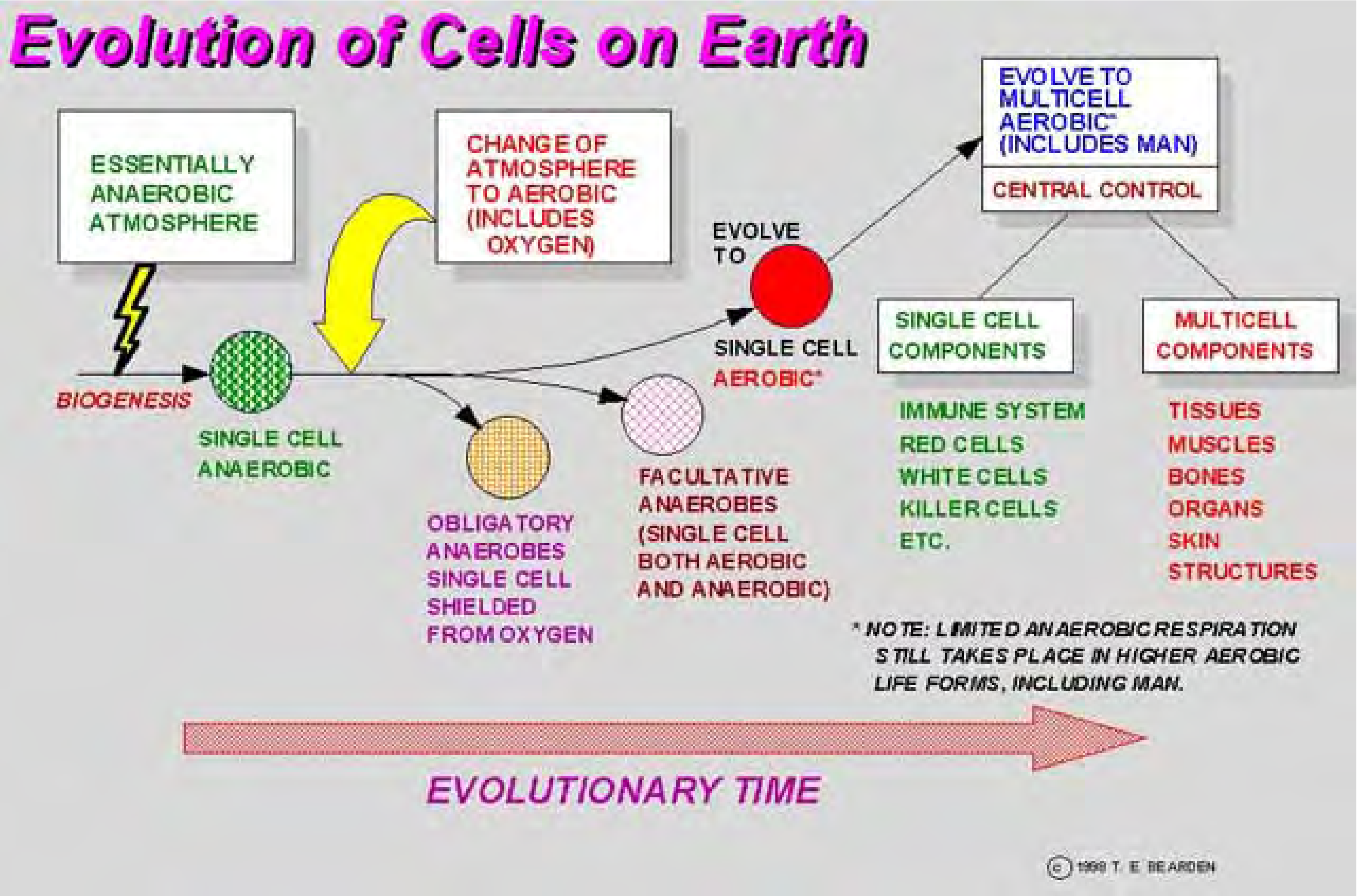
Evolution of Cells on Earth
ESSENTIALLY ANAEROBIC ATMOSPHERE
CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)
EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)
CENTRAL CONTROL
EVOLVE TO
SINGLE CELL AEROBIC*
BIOGENESIS
SINGLE CELL ANAEROBIC
OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN
FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)
SINGLE CELL COMPONENTS
IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.
MULTICELL COMPONENTS
TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES
EVOLUTIONARY TIME

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

(c) 1999 T. E. BEARDEN

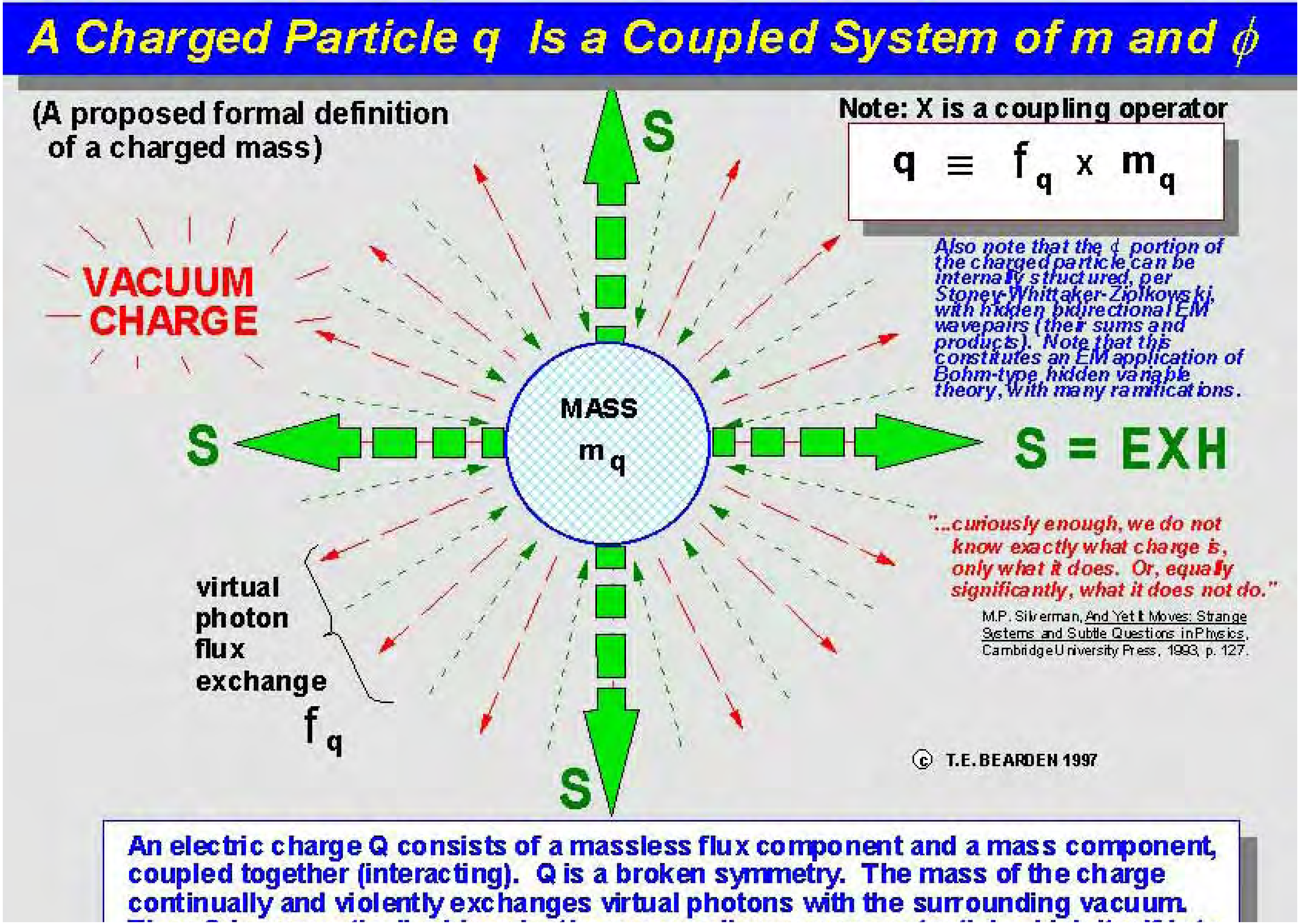

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

S = EXH

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

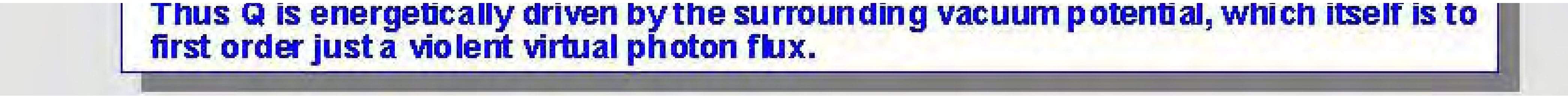

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

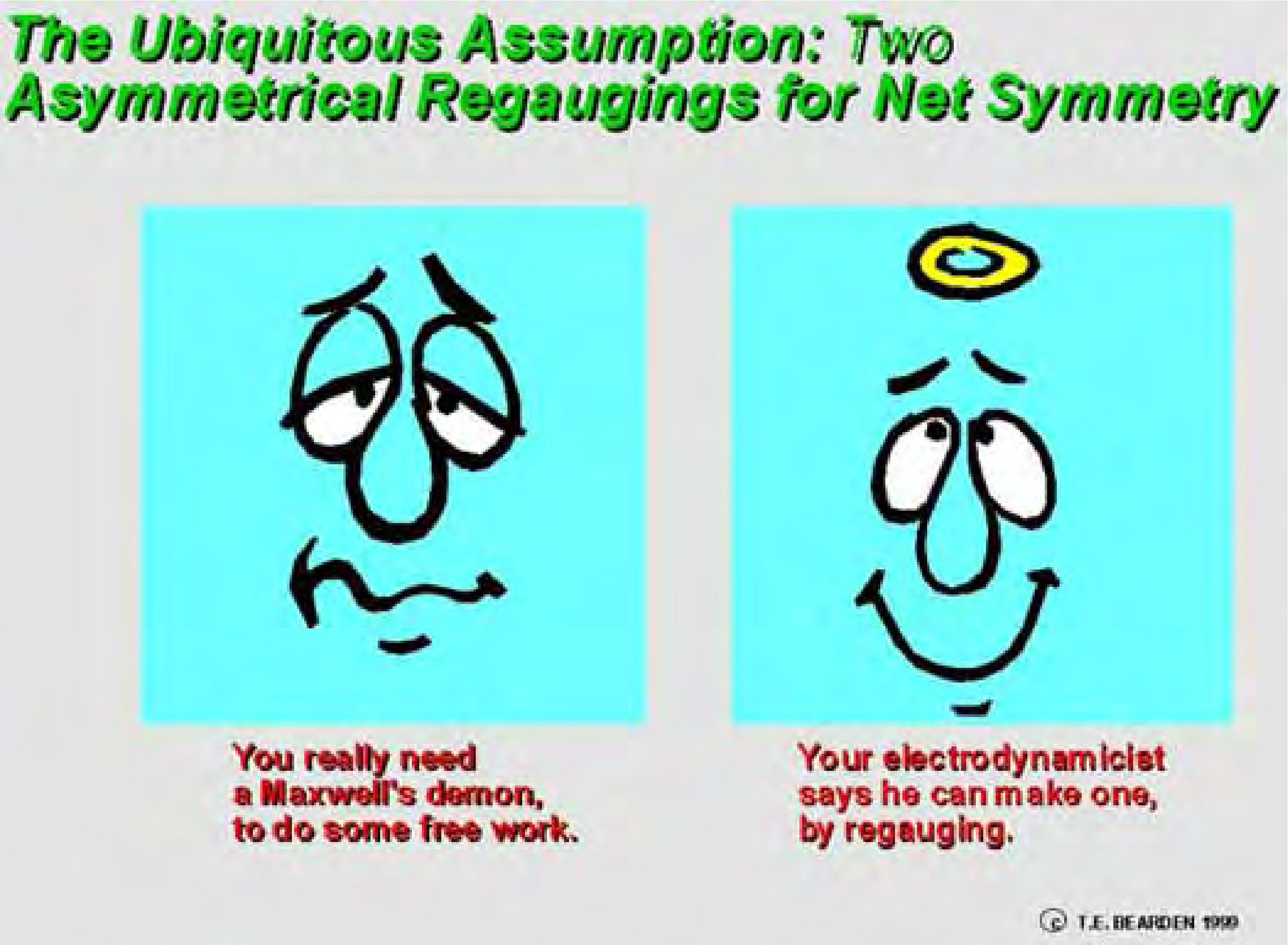

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

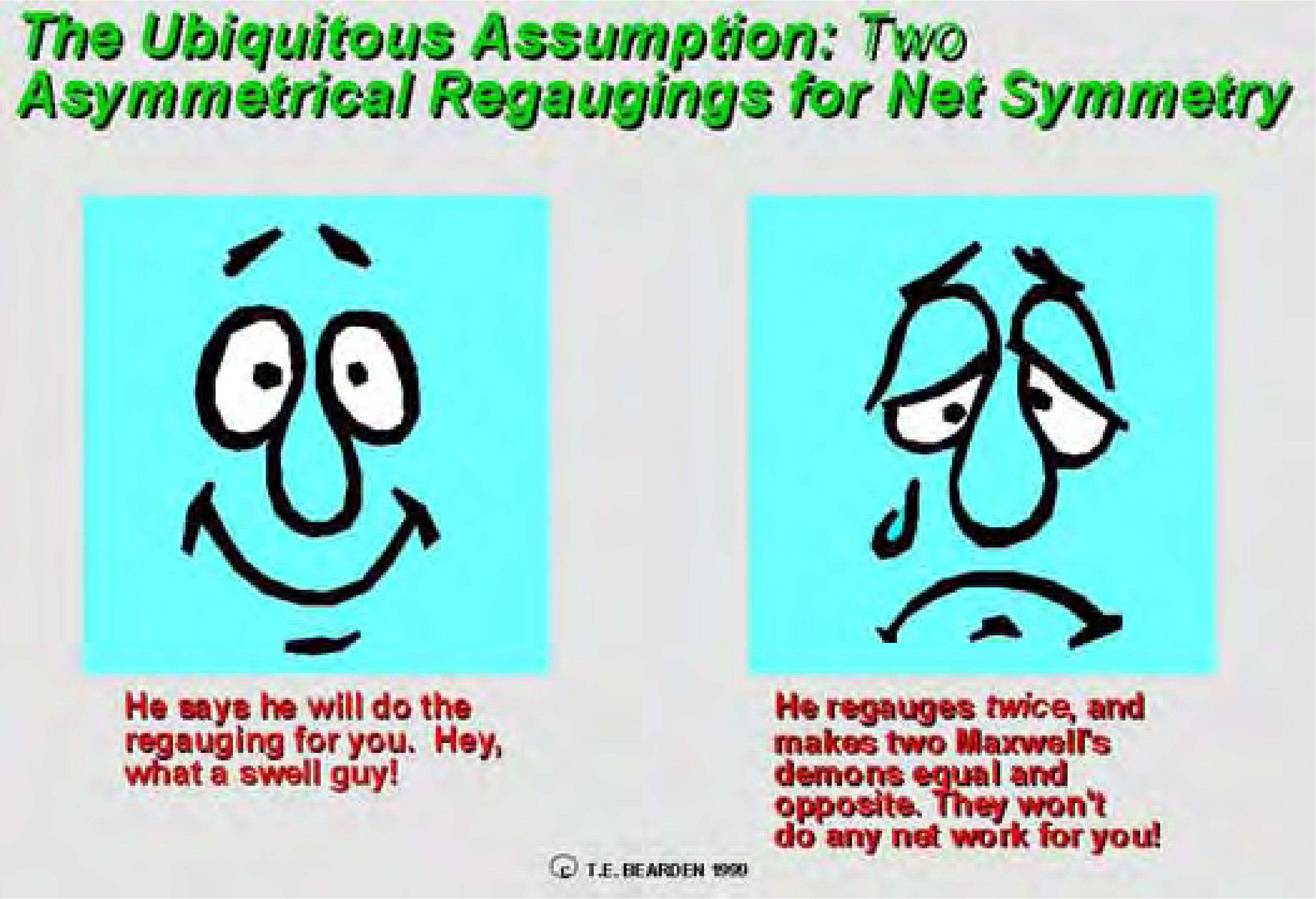

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

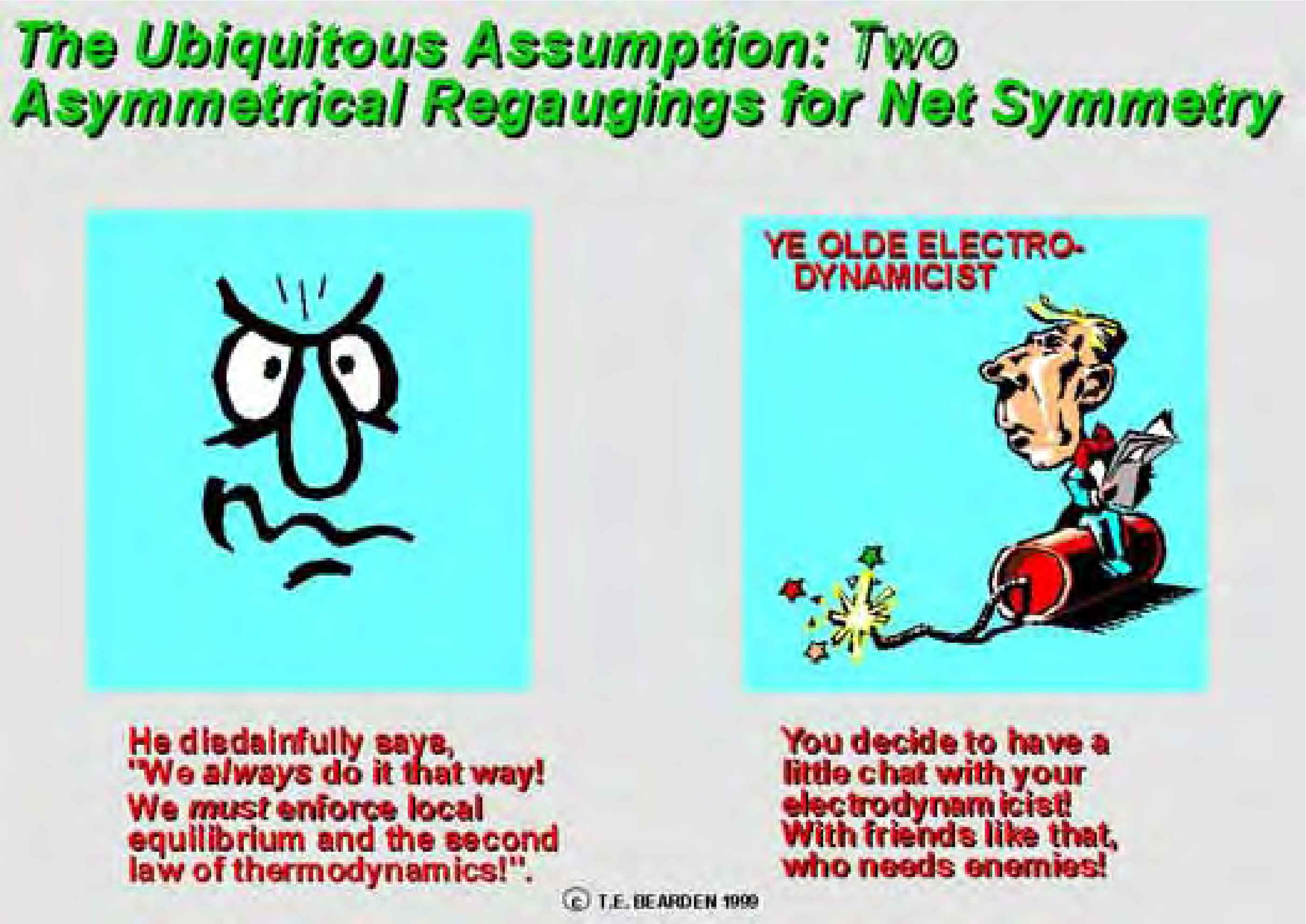

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

YE OLDE ELECTRO-
DYNAMICIST

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$\boxed{q \equiv \phi_q \; x \; m_q}$$

Note: X is a coupling operator

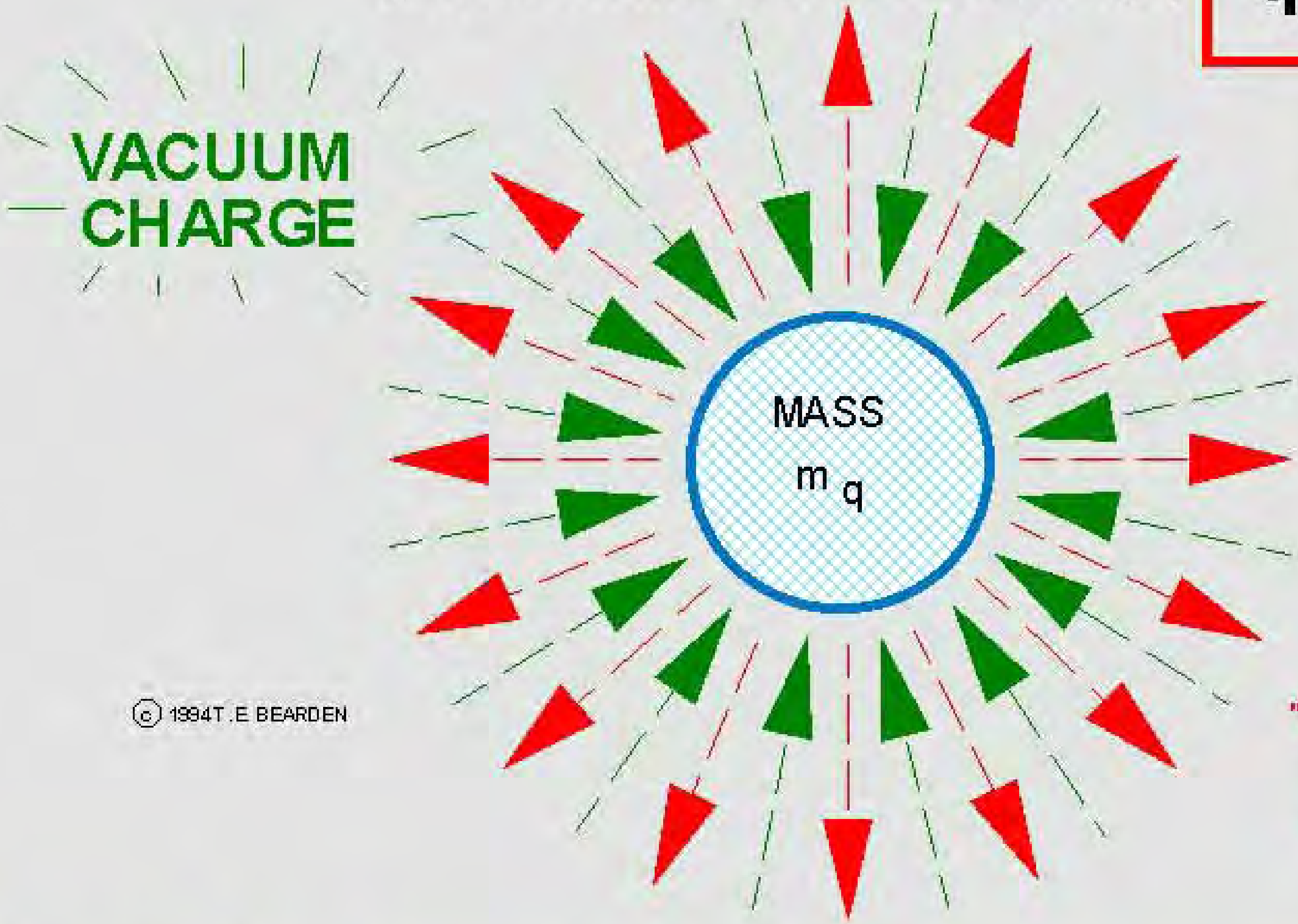

© 1994 T.E BEARDEN

virtual
photon
flux
exchange

ϕ_q

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

Chung's Carbon Filament Negative Resistor

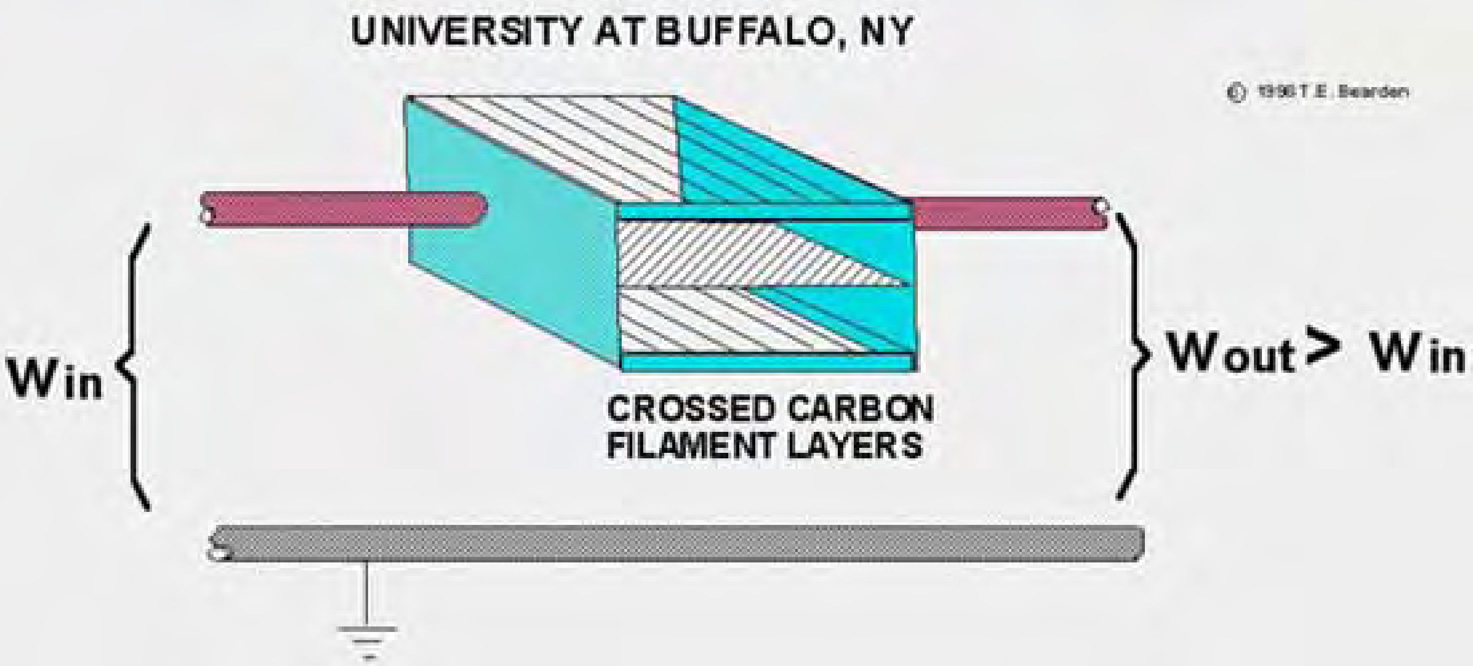
UNIVERSITY AT BUFFALO, NY
© 1996 T.E. Bearden
Win
Wout > Win
CROSSED CARBON FILAMENT LAYERS

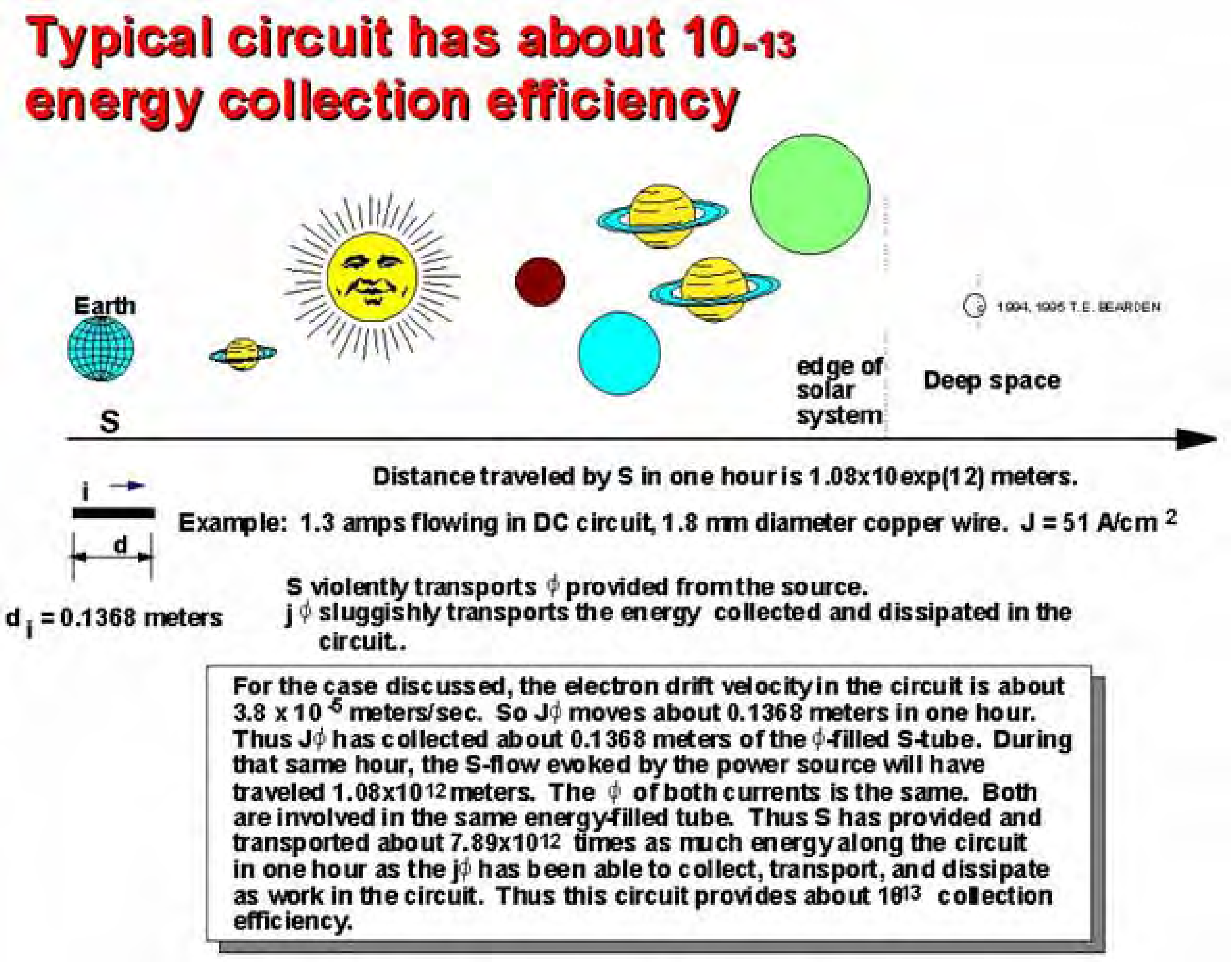

Typical circuit has about 10_{-13} energy collection efficiency

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. J = 51 A/cm ^{2}

d_i = 0.1368 meters

S violently transports ϕ provided from the source.
j ϕ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So Jϕ moves about 0.1368 meters in one hour. Thus Jϕ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled 1.08×10^{12} meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about 7.89×10^{12} times as much energy along the circuit in one hour as the jϕ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

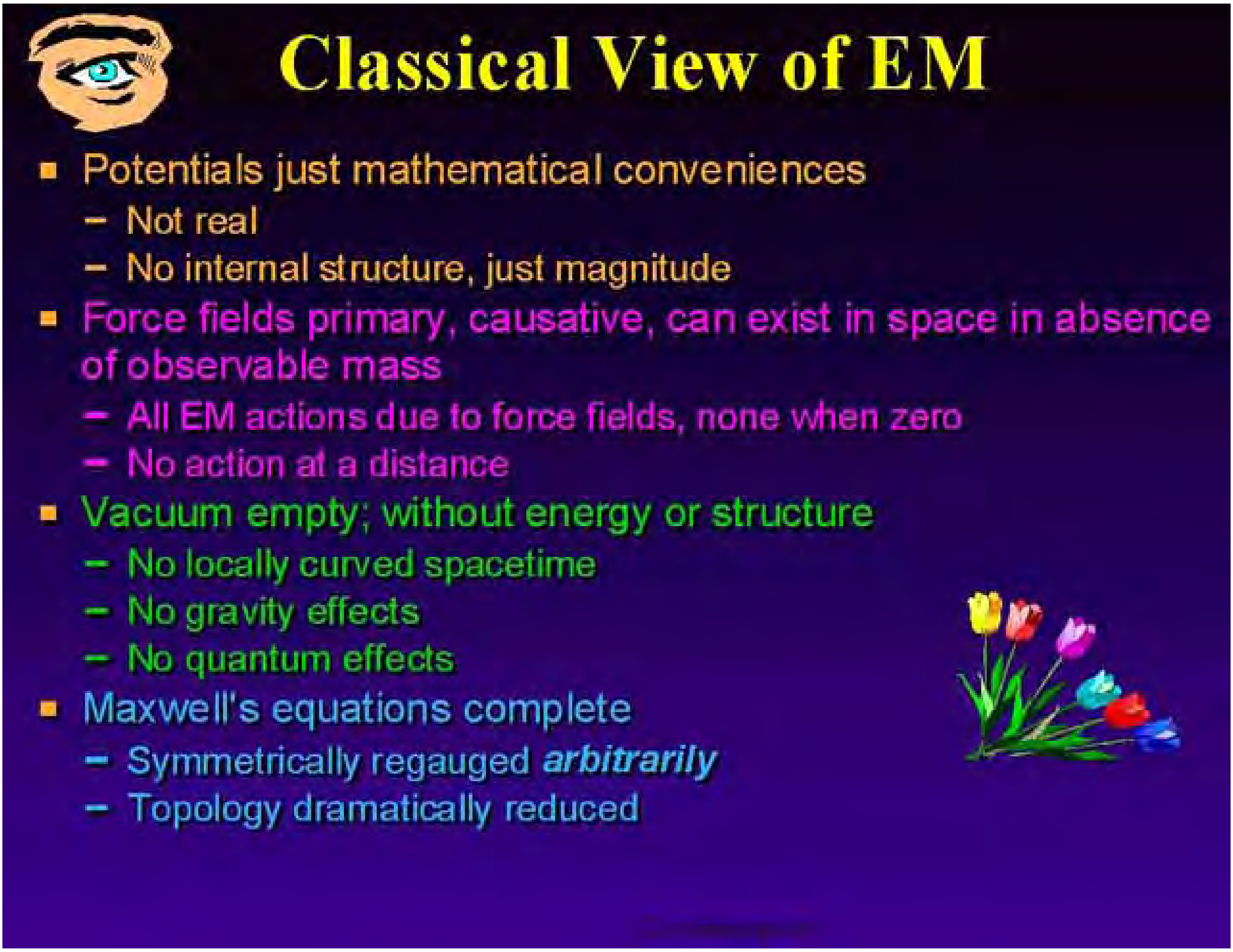

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

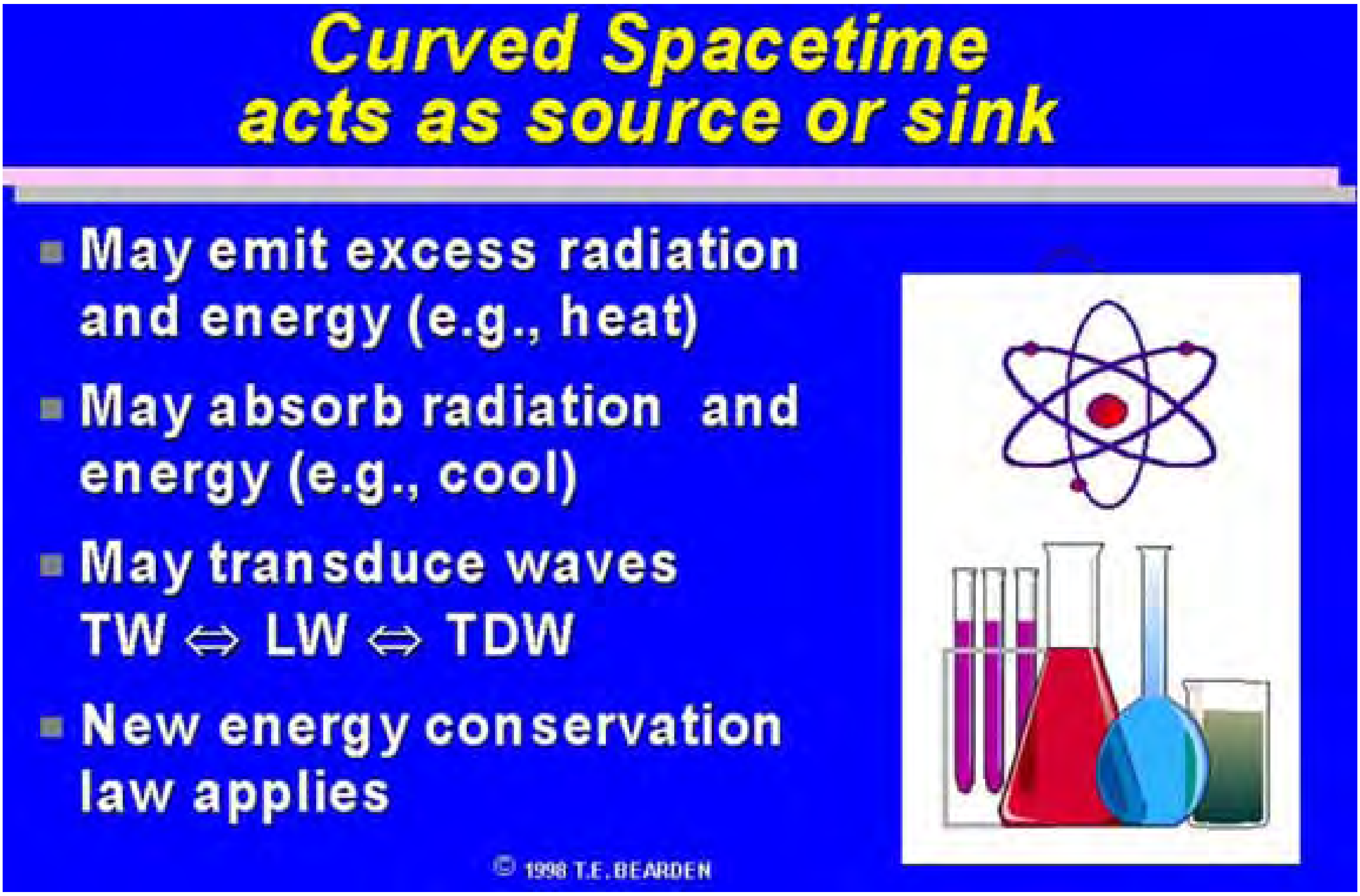

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves
  TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E.BEARDEN

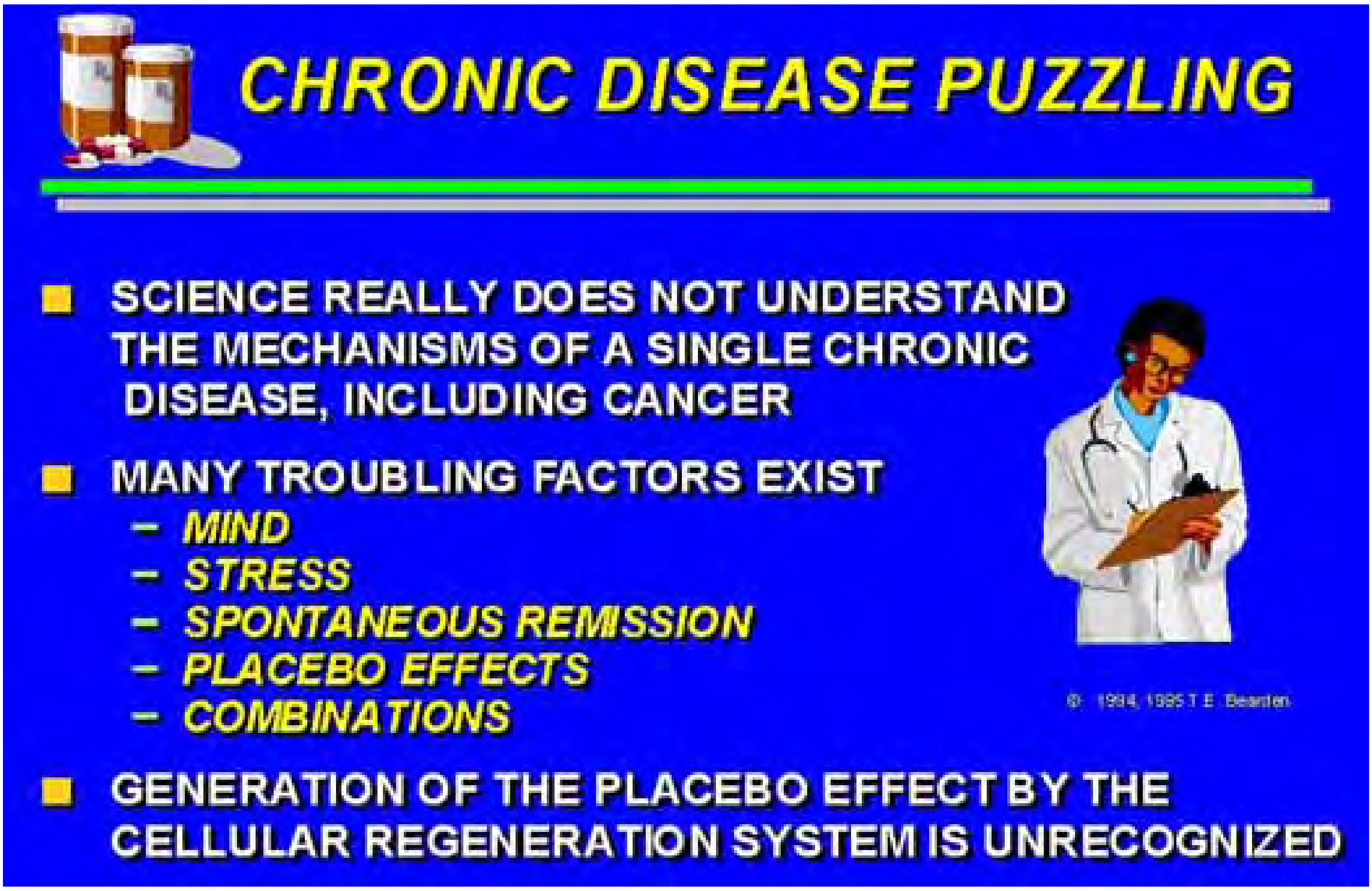

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

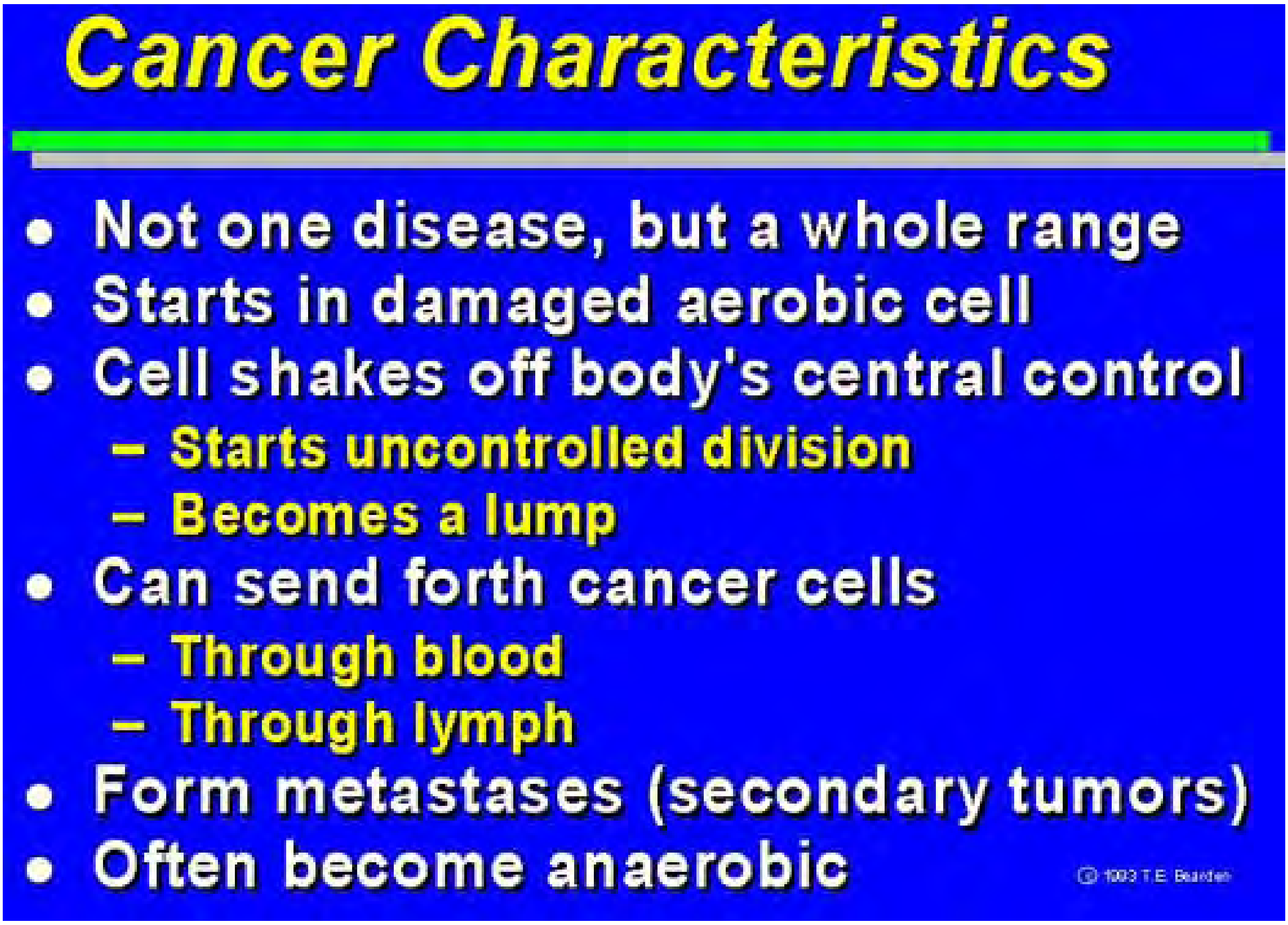

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

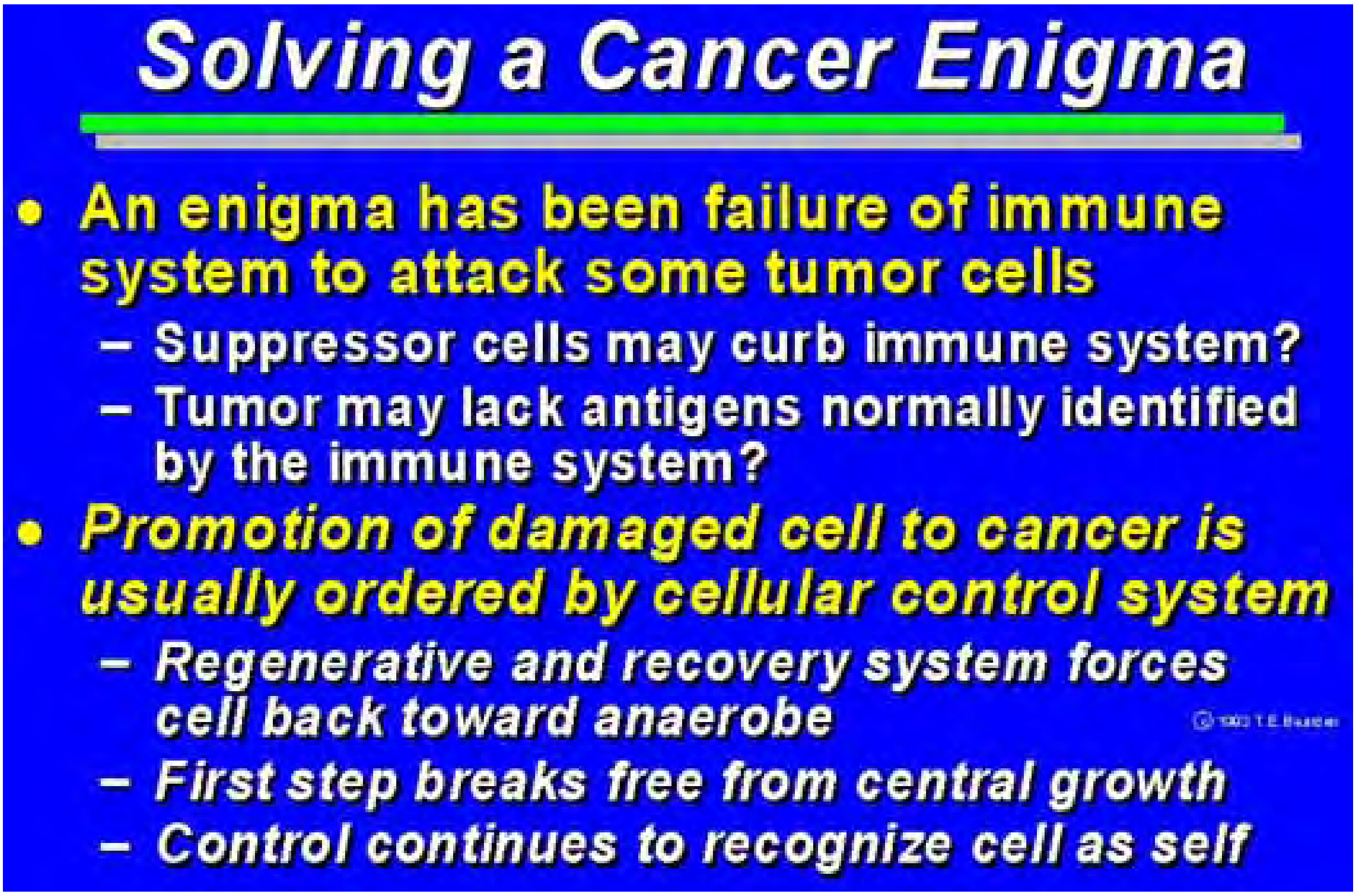

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

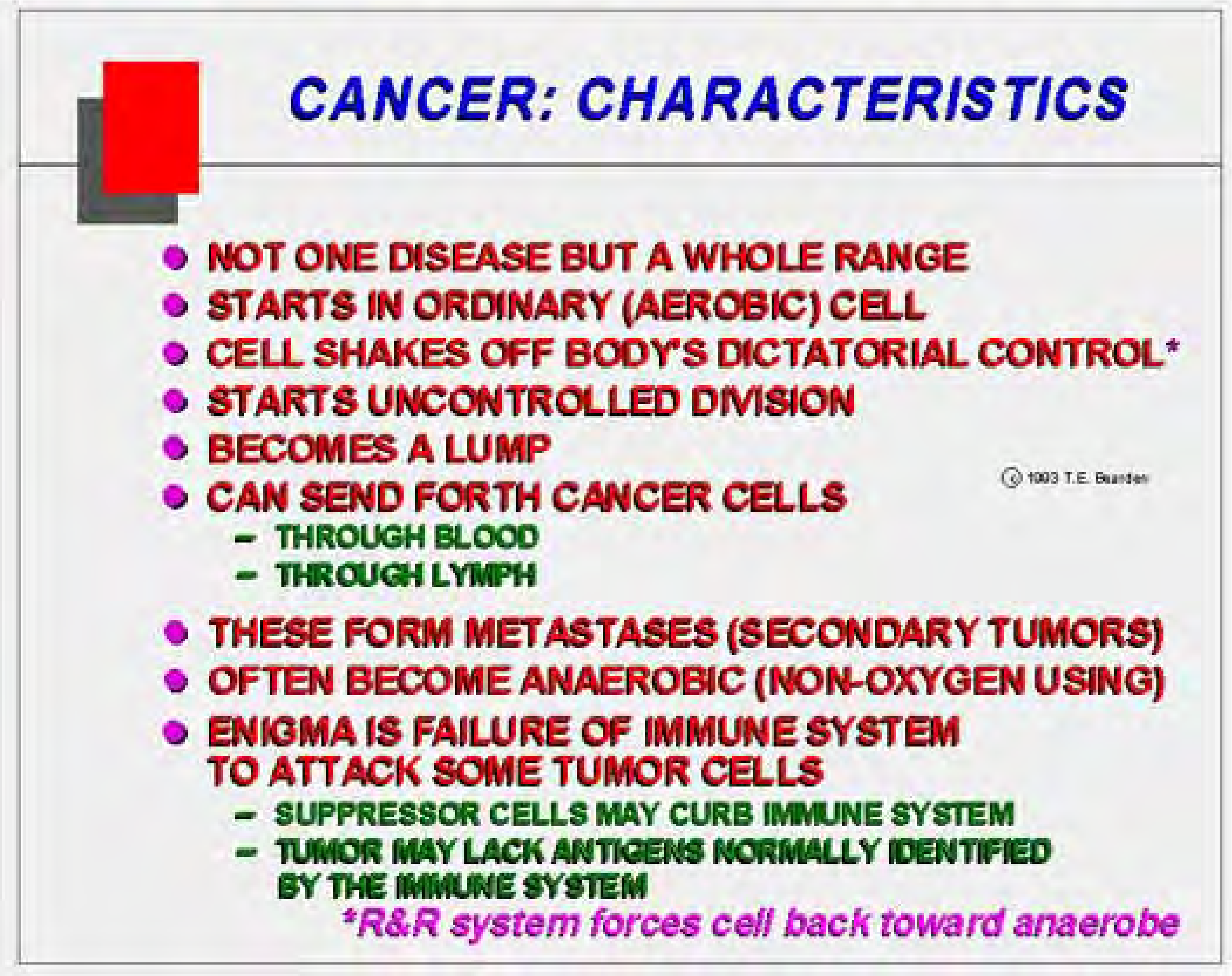

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

© 1993 T.E. Bearden

*R&R system forces cell back toward anaerobe

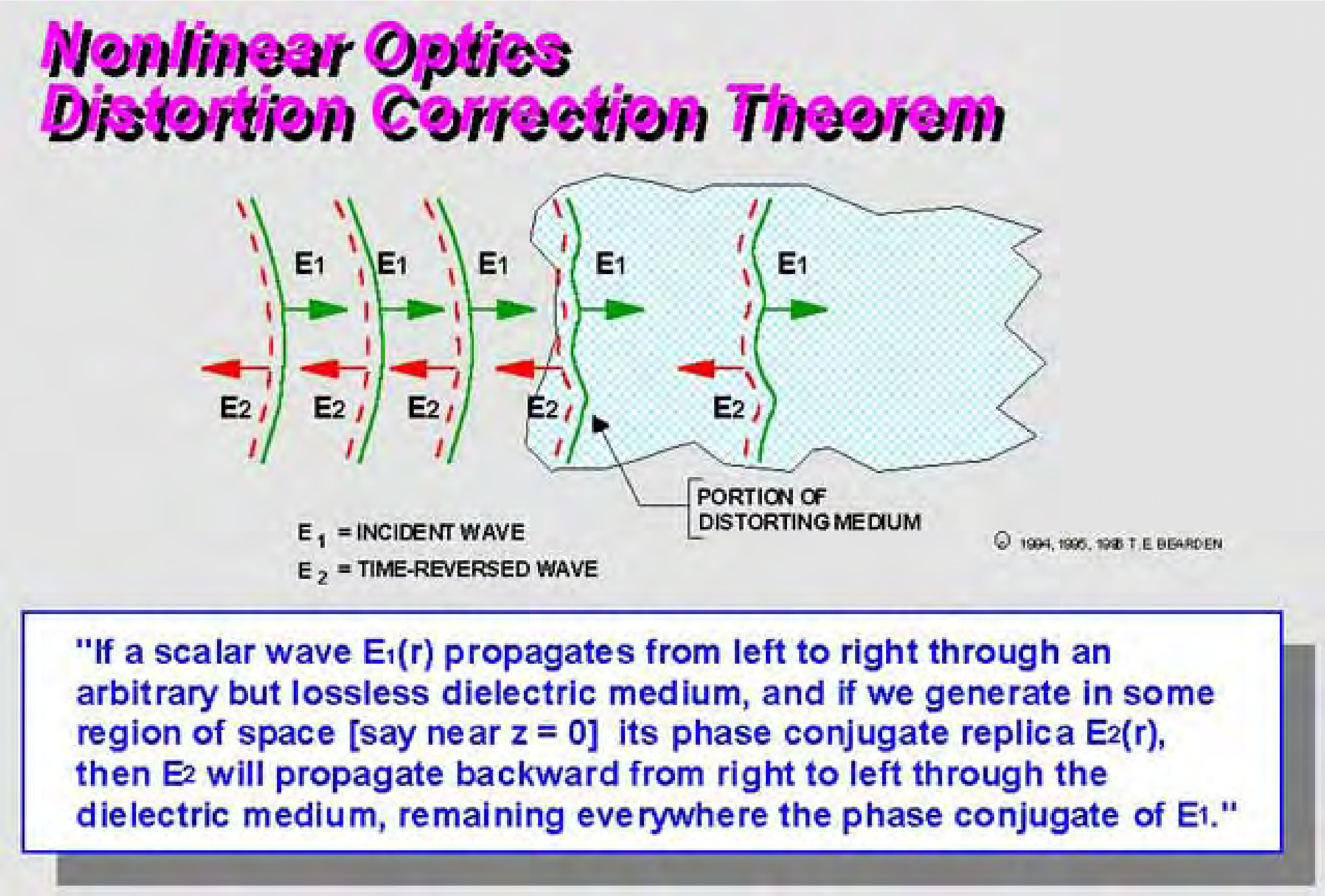

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T.E BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

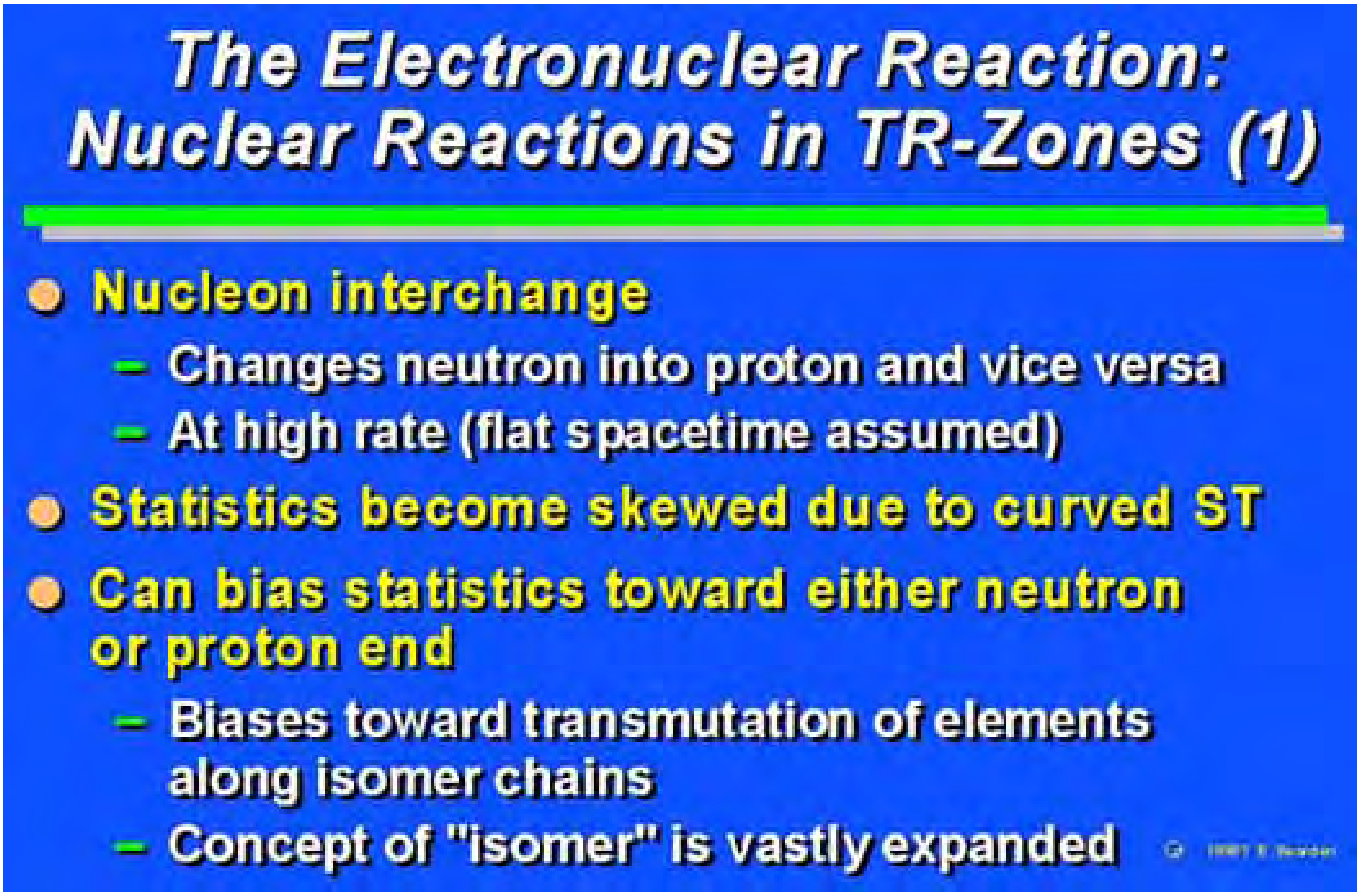

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- **Nucleon interchange**
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- **Statistics become skewed due to curved ST**
- **Can bias statistics toward either neutron or proton end**
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

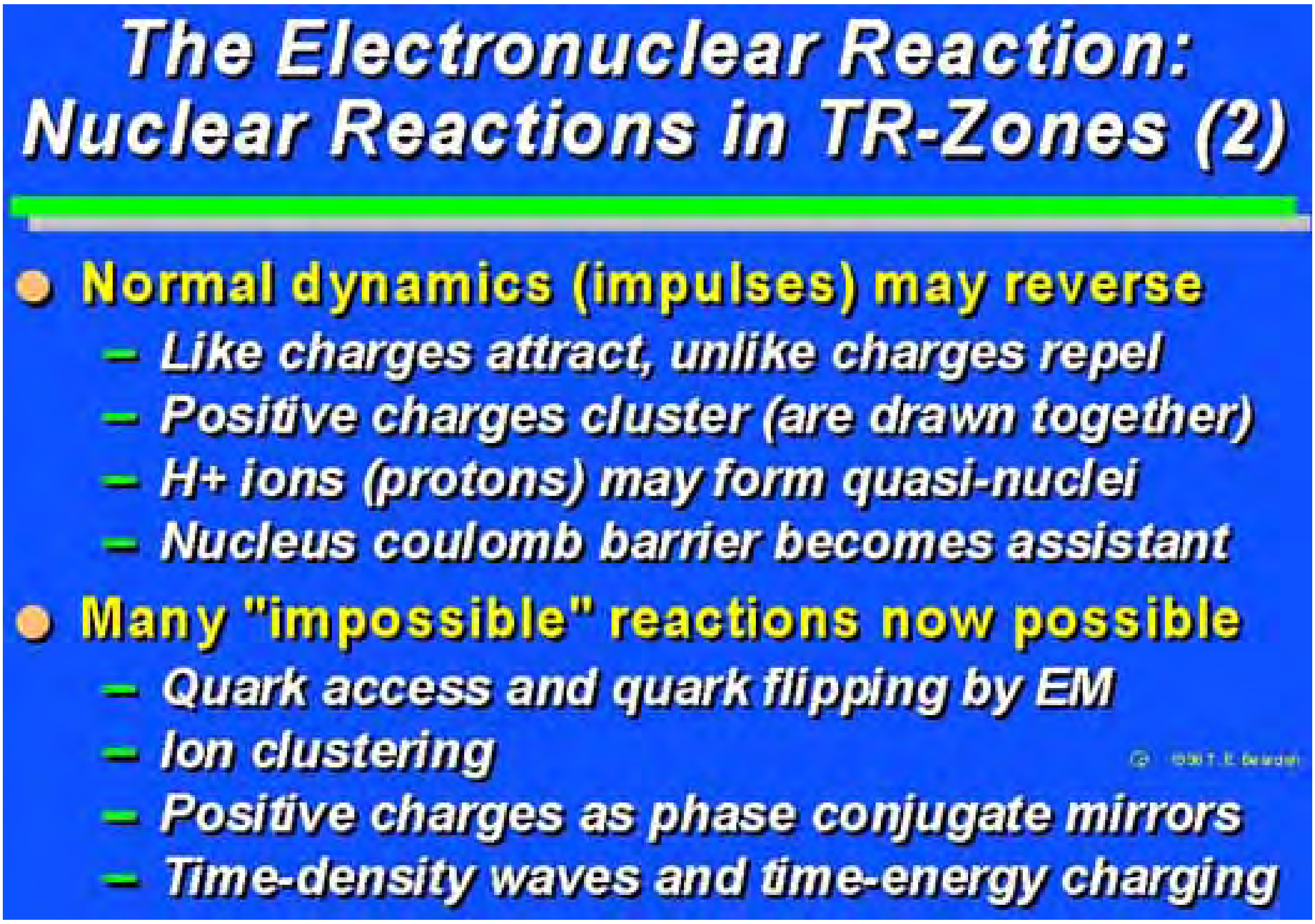

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

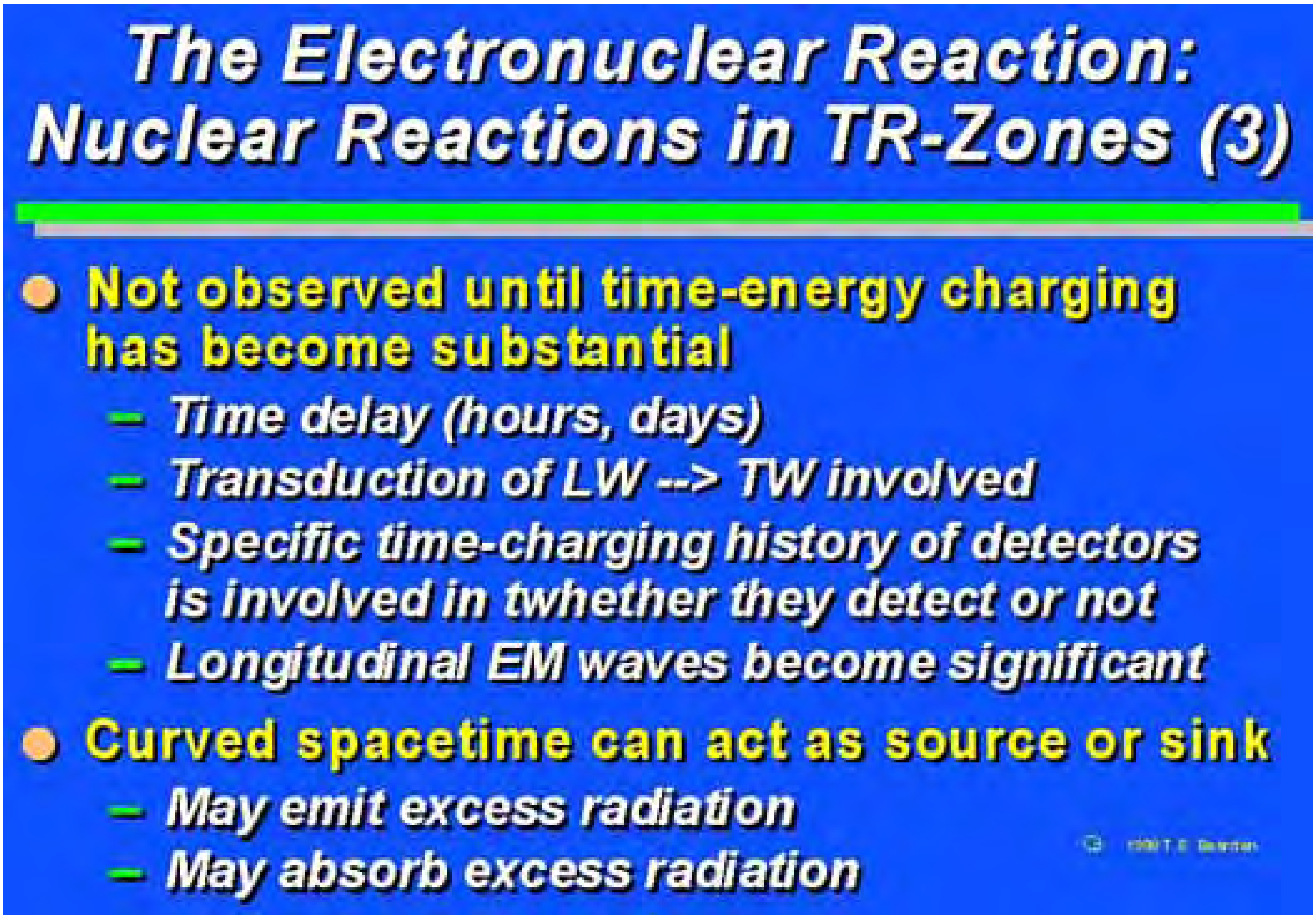

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

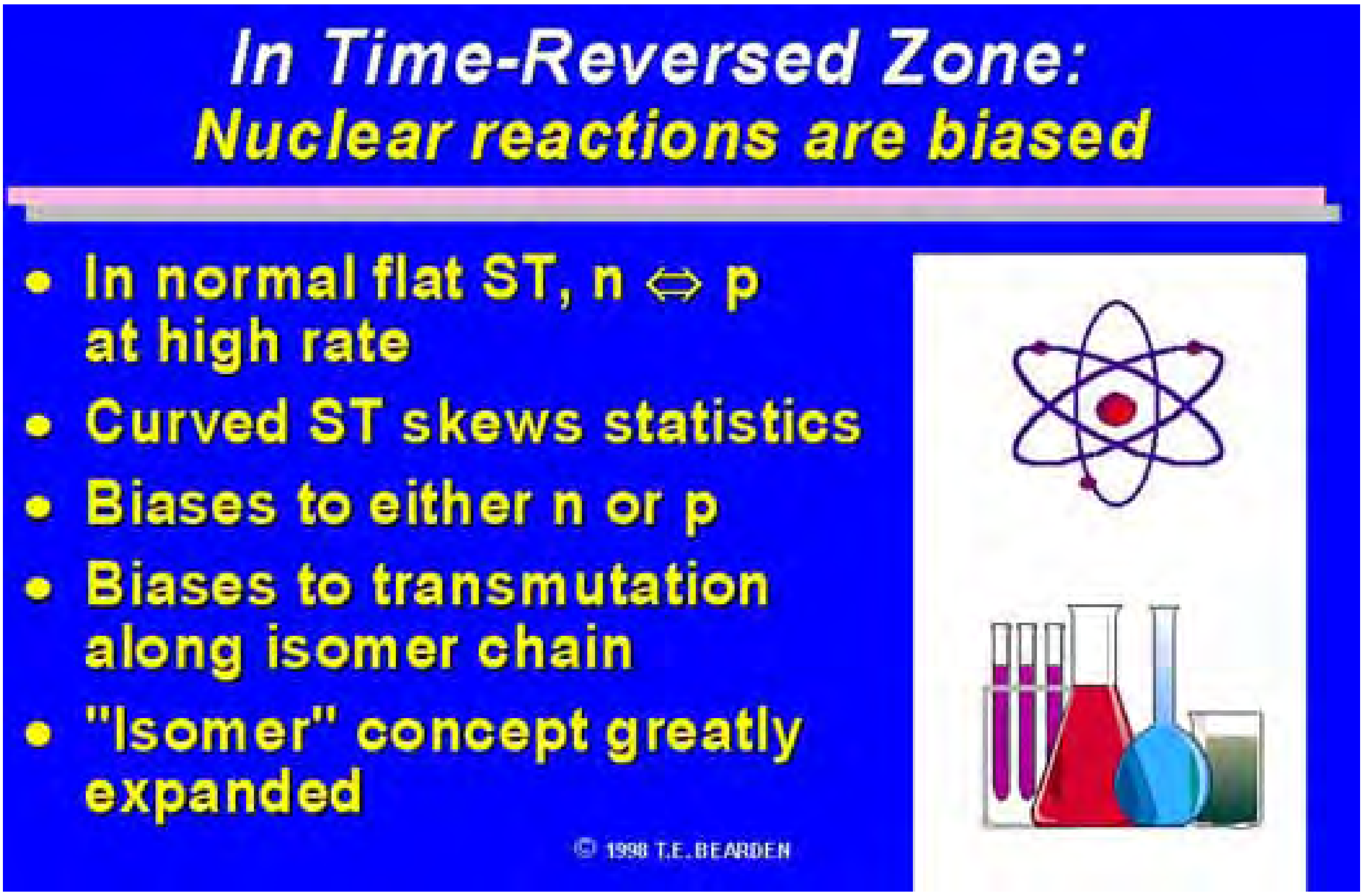

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E. BEARDEN

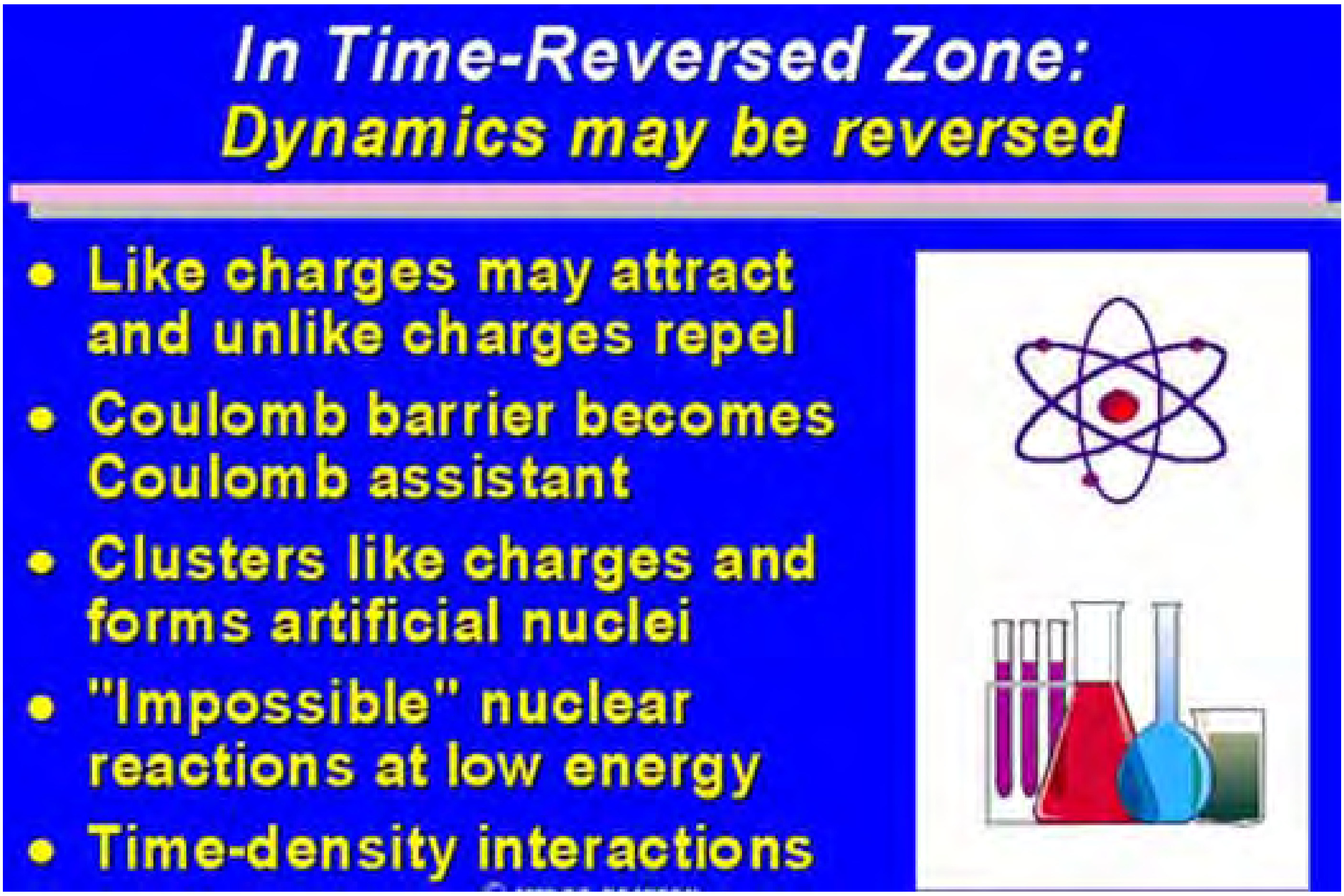

In Time-Reversed Zone: Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

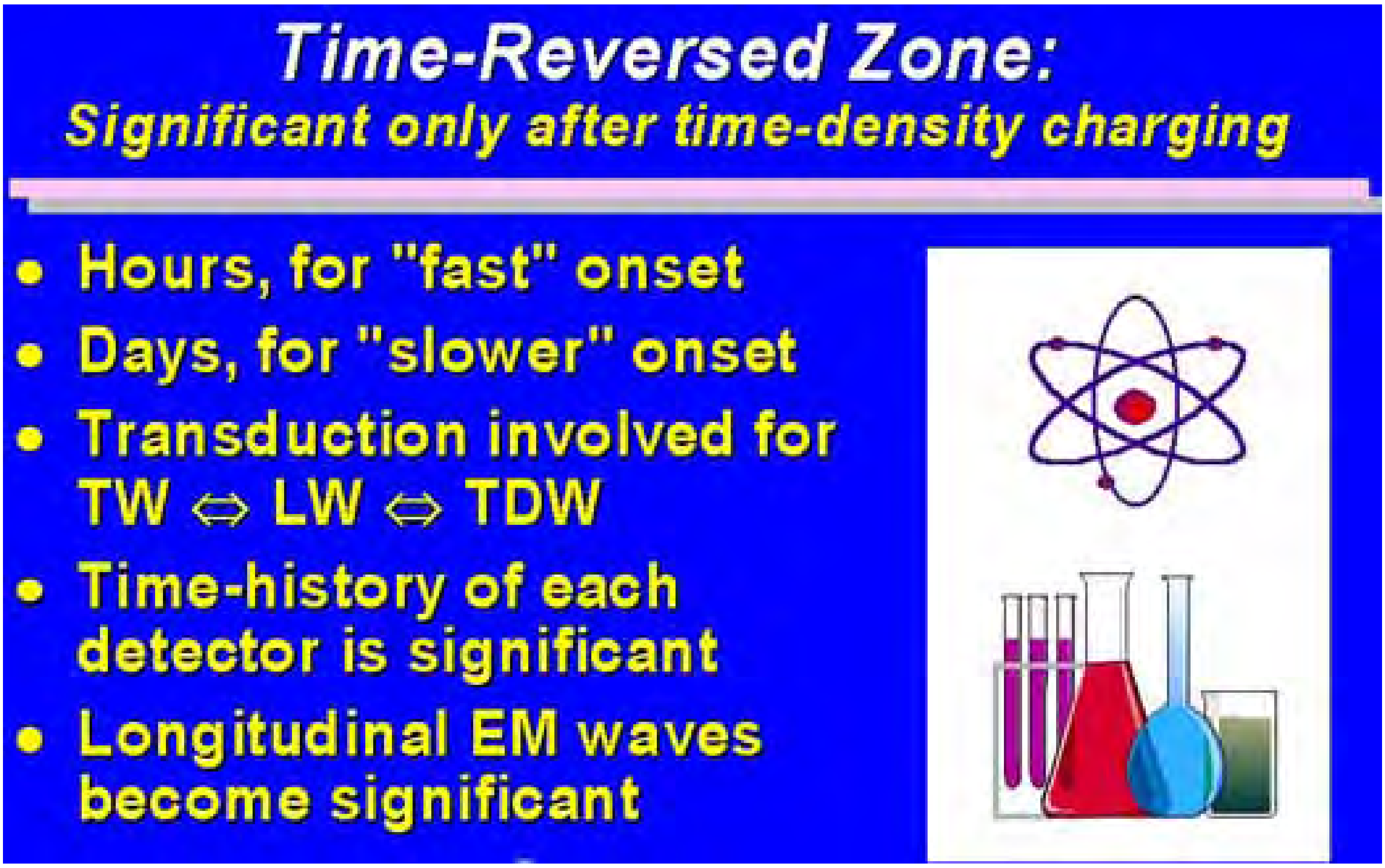

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

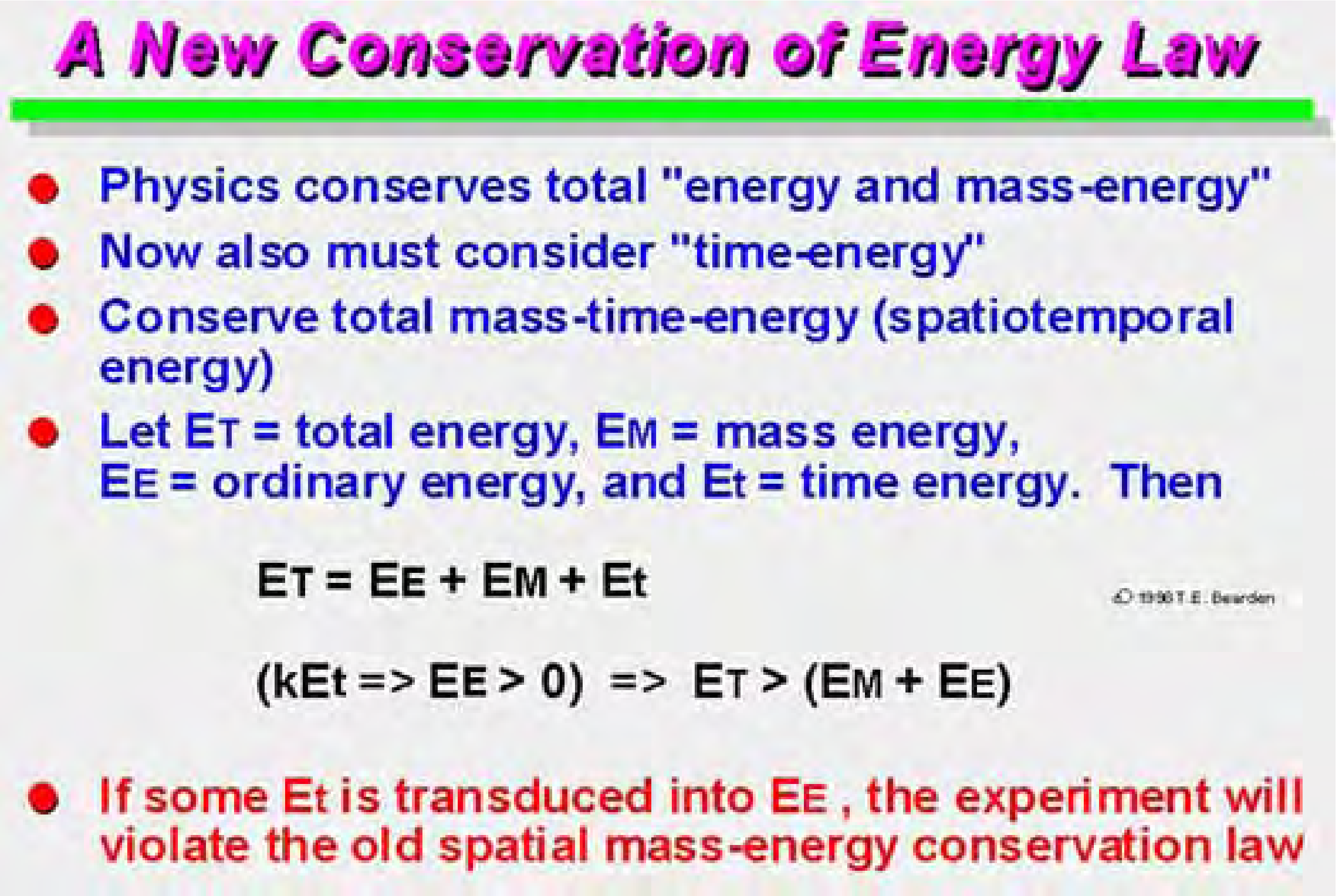

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T. E. Bearden

$$(kE_t \Rightarrow E_E > 0) \quad \Rightarrow \quad E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

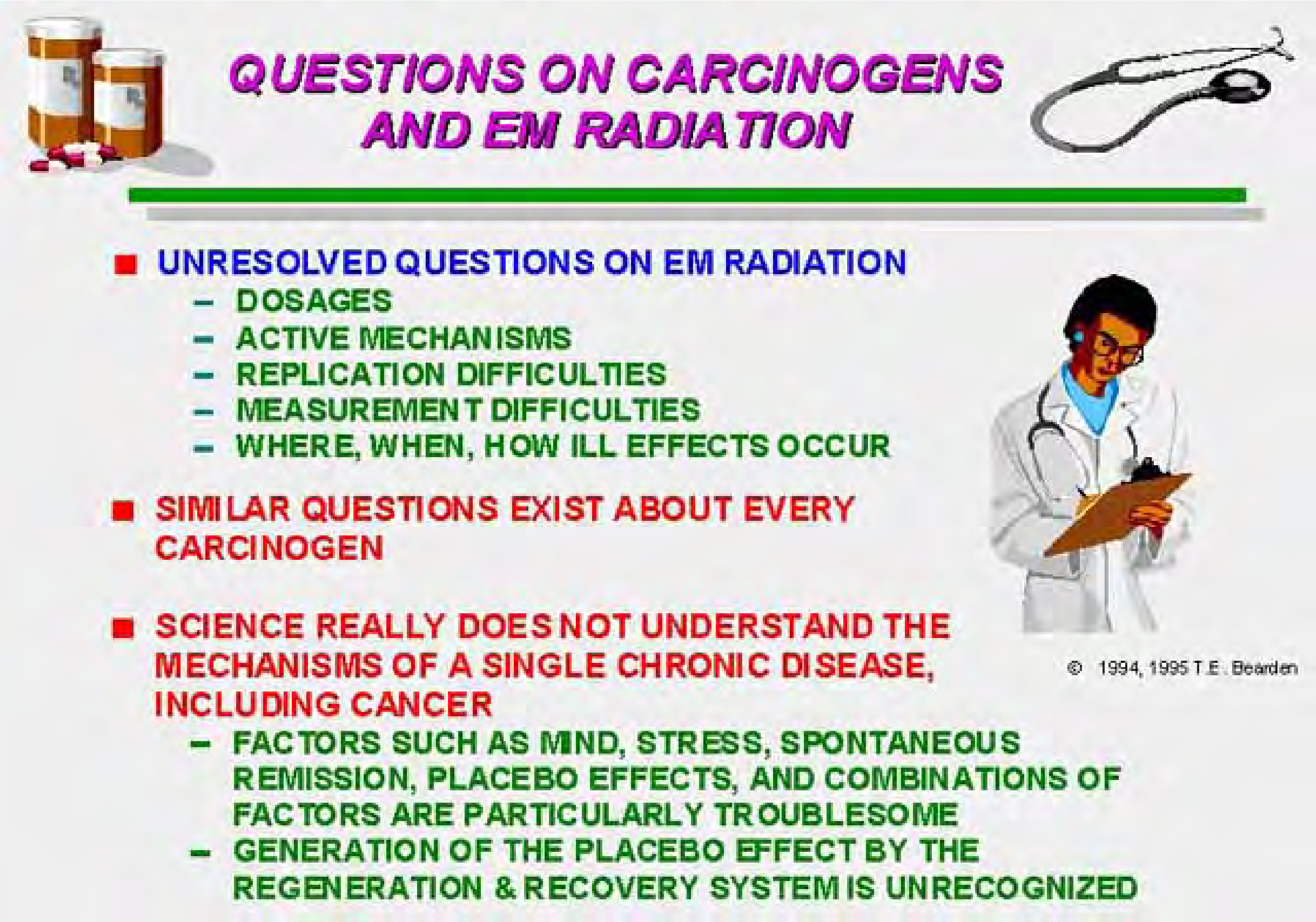

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

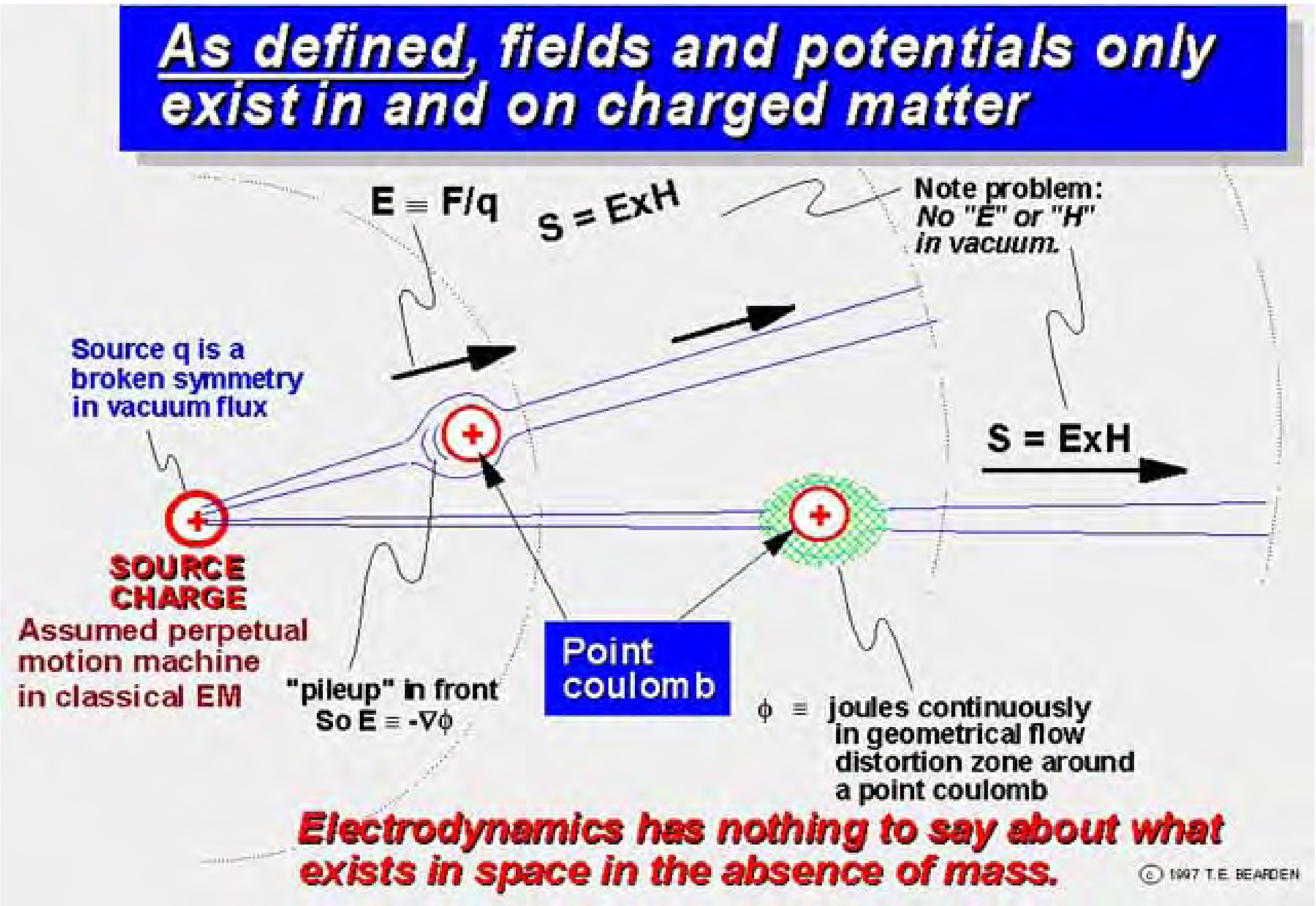

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

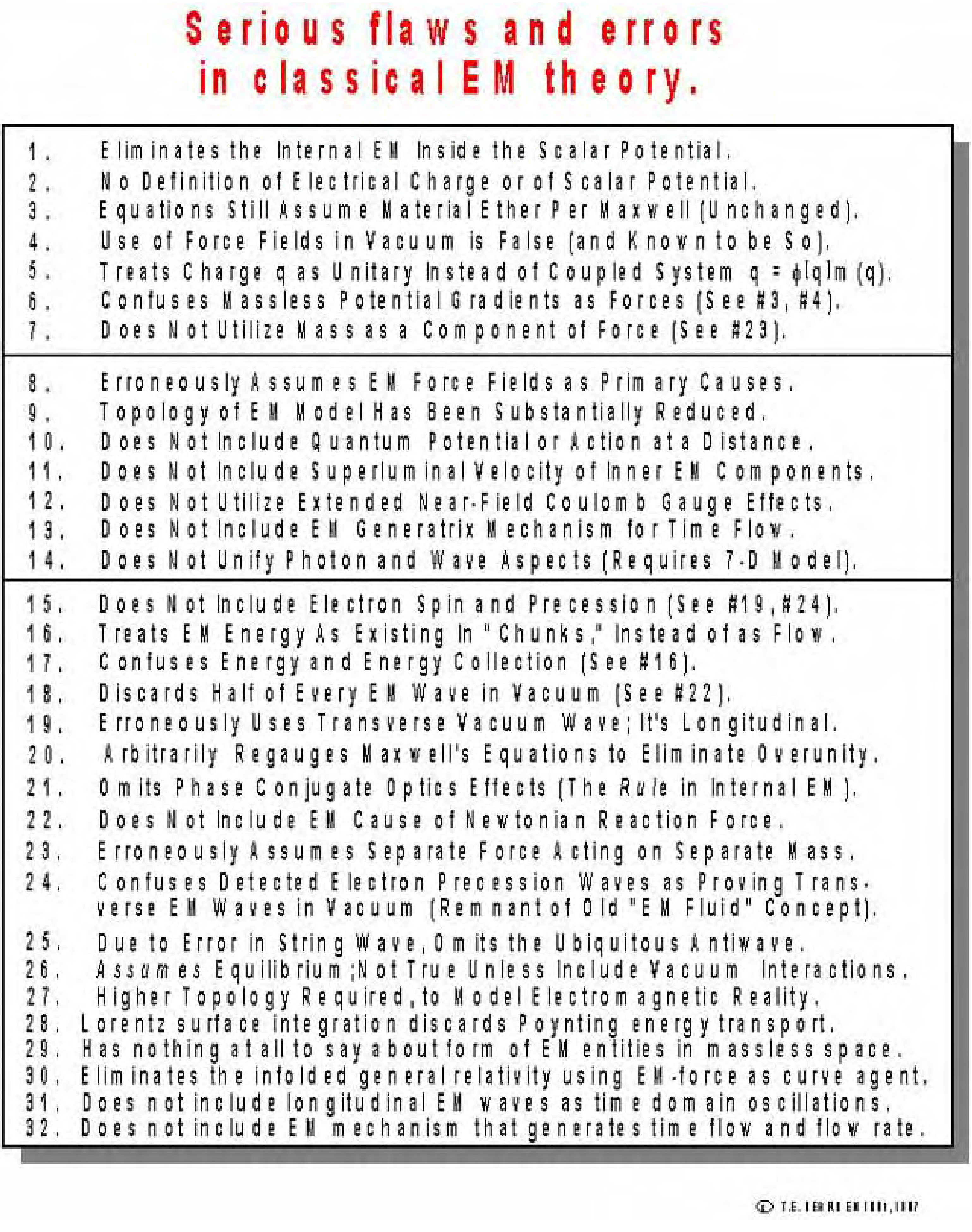

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

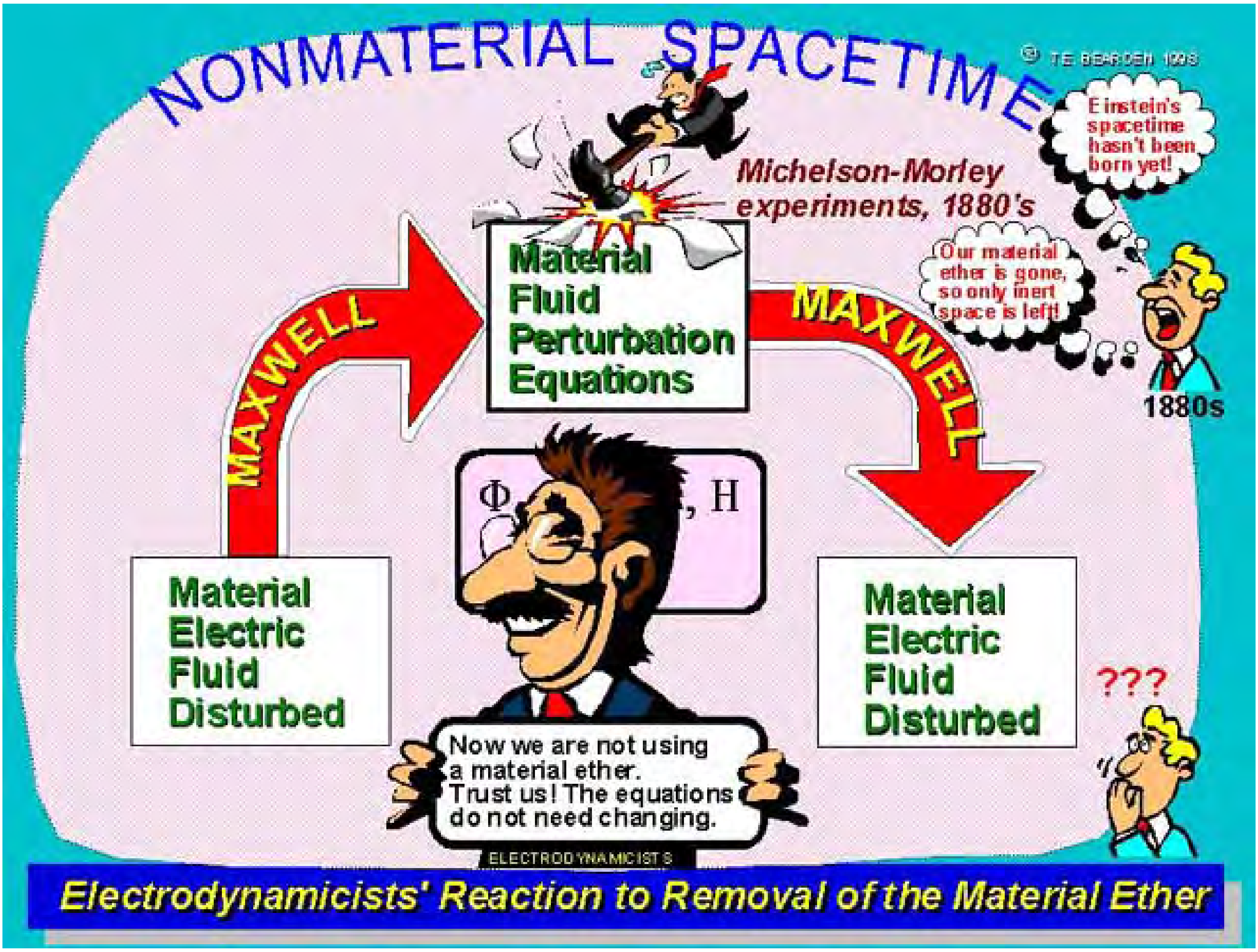
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS
Electrodynamicists' Reaction to Removal of the Material Ether

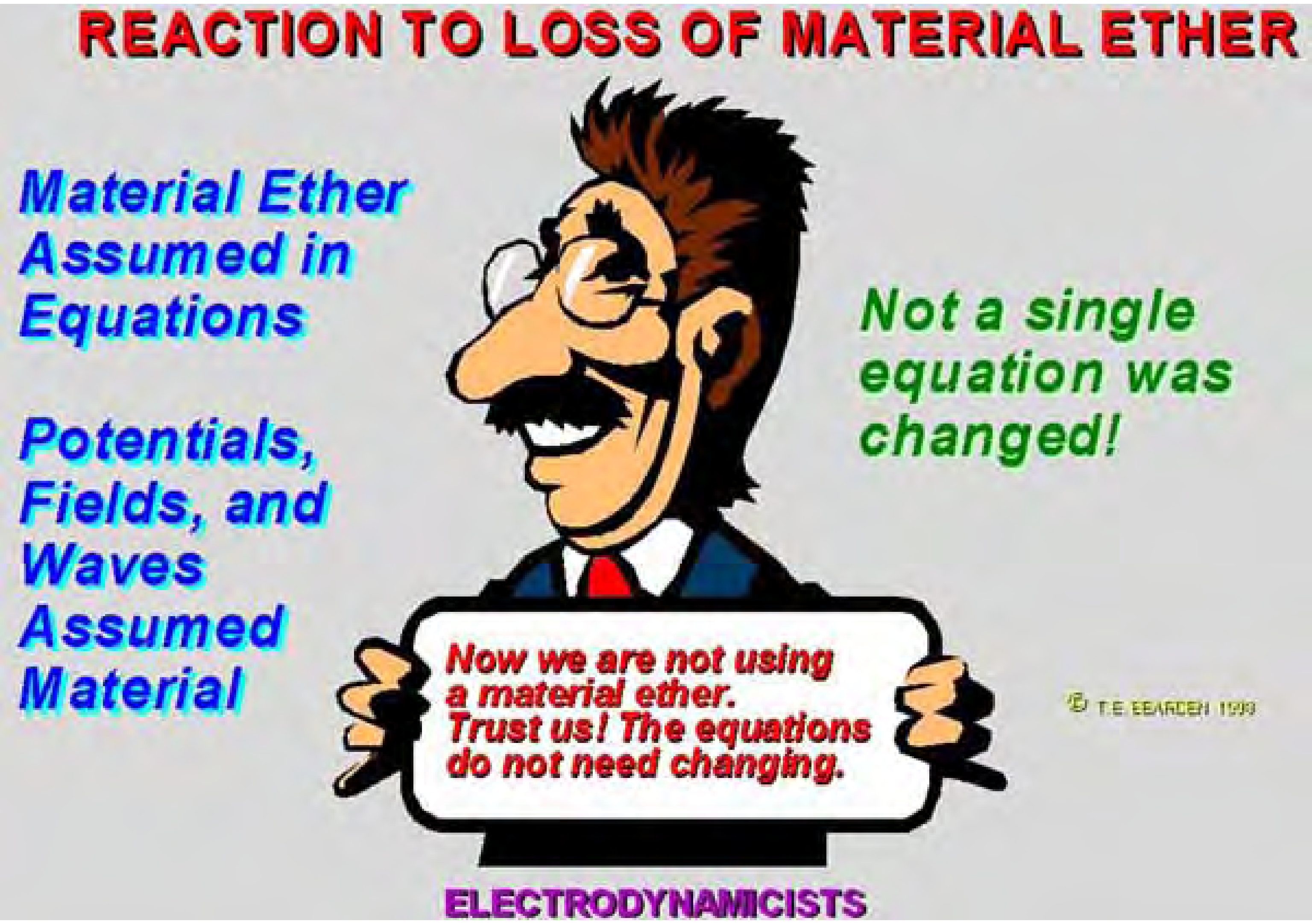

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.
ELECTRODYNAMICISTS

© T.E. BEARDEN 1999

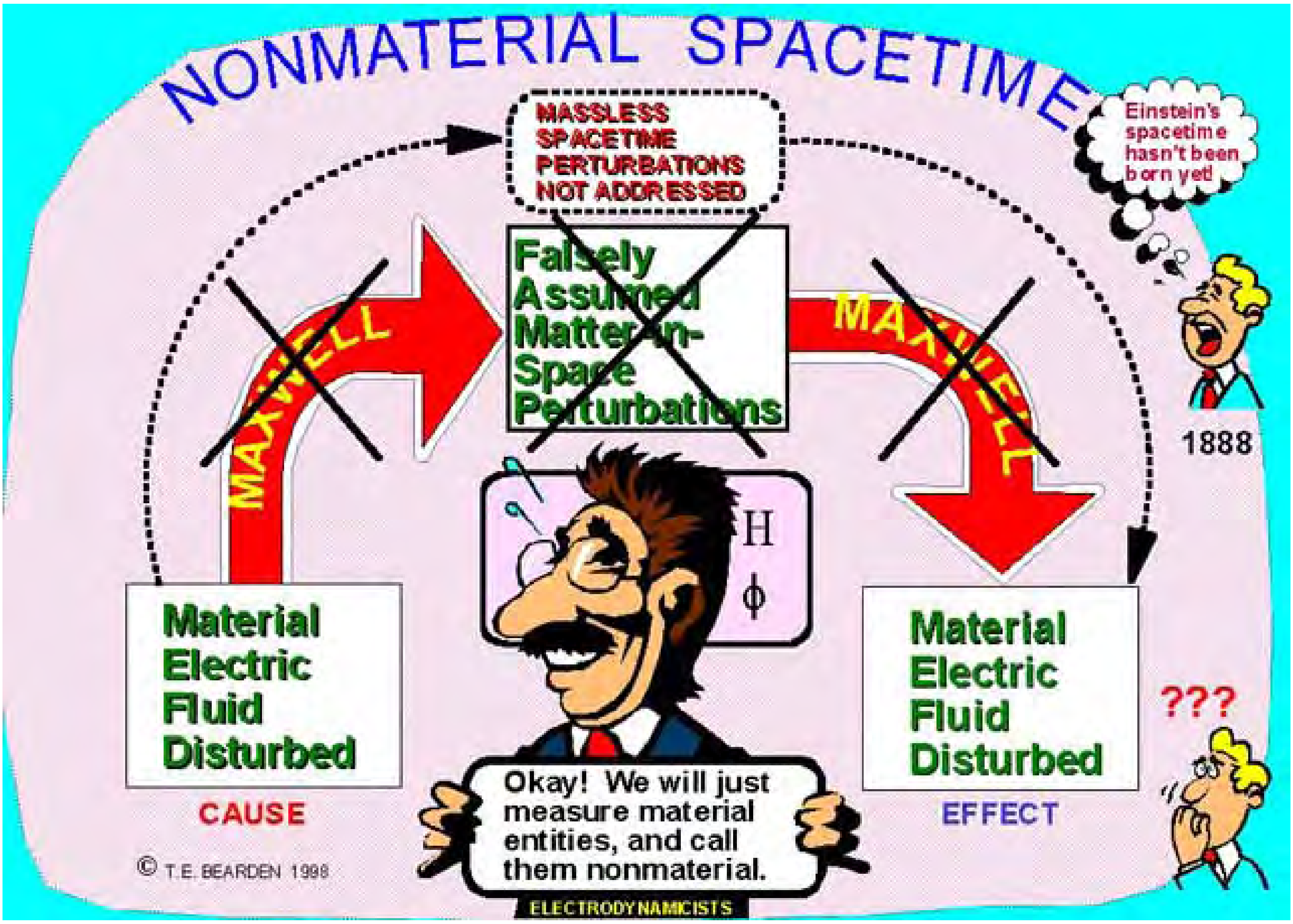
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
1888
~~MAXWELL~~
~~Falsely Assumed Matter-in-Space Perturbations~~
~~MAXWELL~~
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

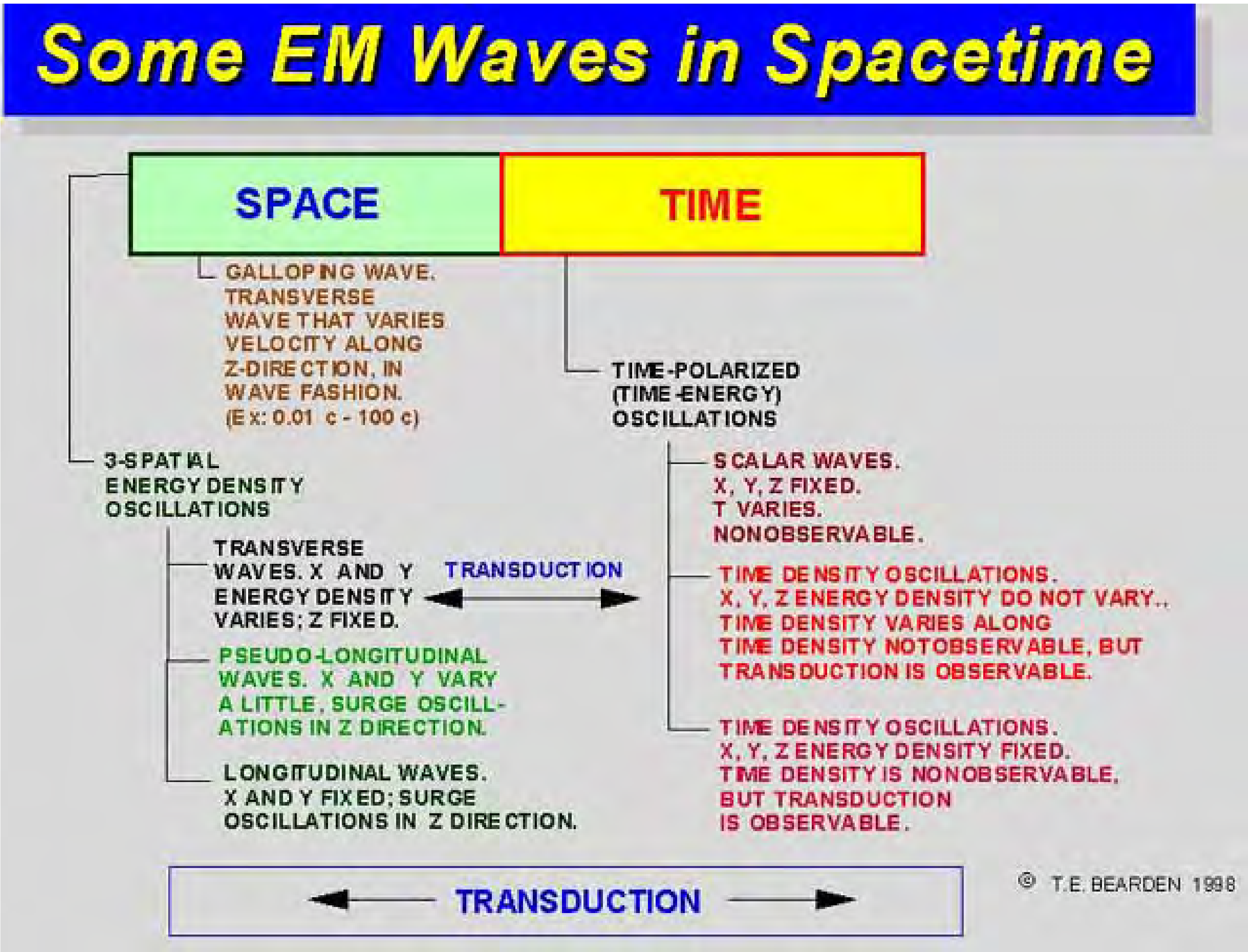

Some EM Waves in Spacetime

SPACE **TIME**

GALLOPING WAVE.
TRANSVERSE
WAVE THAT VARIES
VELOCITY ALONG
Z-DIRECTION, IN
WAVE FASHION.
(Ex: 0.01 c - 100 c)

3-SPATIAL
ENERGY DENSITY
OSCILLATIONS

TRANSVERSE
WAVES. X AND Y
ENERGY DENSITY
VARIES; Z FIXED.

PSEUDO-LONGITUDINAL
WAVES. X AND Y VARY
A LITTLE, SURGE OSCILL-
ATIONS IN Z DIRECTION.

LONGITUDINAL WAVES.
X AND Y FIXED; SURGE
OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

TIME-POLARIZED
(TIME-ENERGY)
OSCILLATIONS

SCALAR WAVES.
X, Y, Z FIXED.
T VARIES.
NONOBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY DO NOT VARY..
TIME DENSITY VARIES ALONG
TIME DENSITY NOTOBSERVABLE, BUT
TRANSDUCTION IS OBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY FIXED.
TIME DENSITY IS NONOBSERVABLE,
BUT TRANSDUCTION
IS OBSERVABLE.

TRANSDUCTION

© T.E. BEARDEN 1998

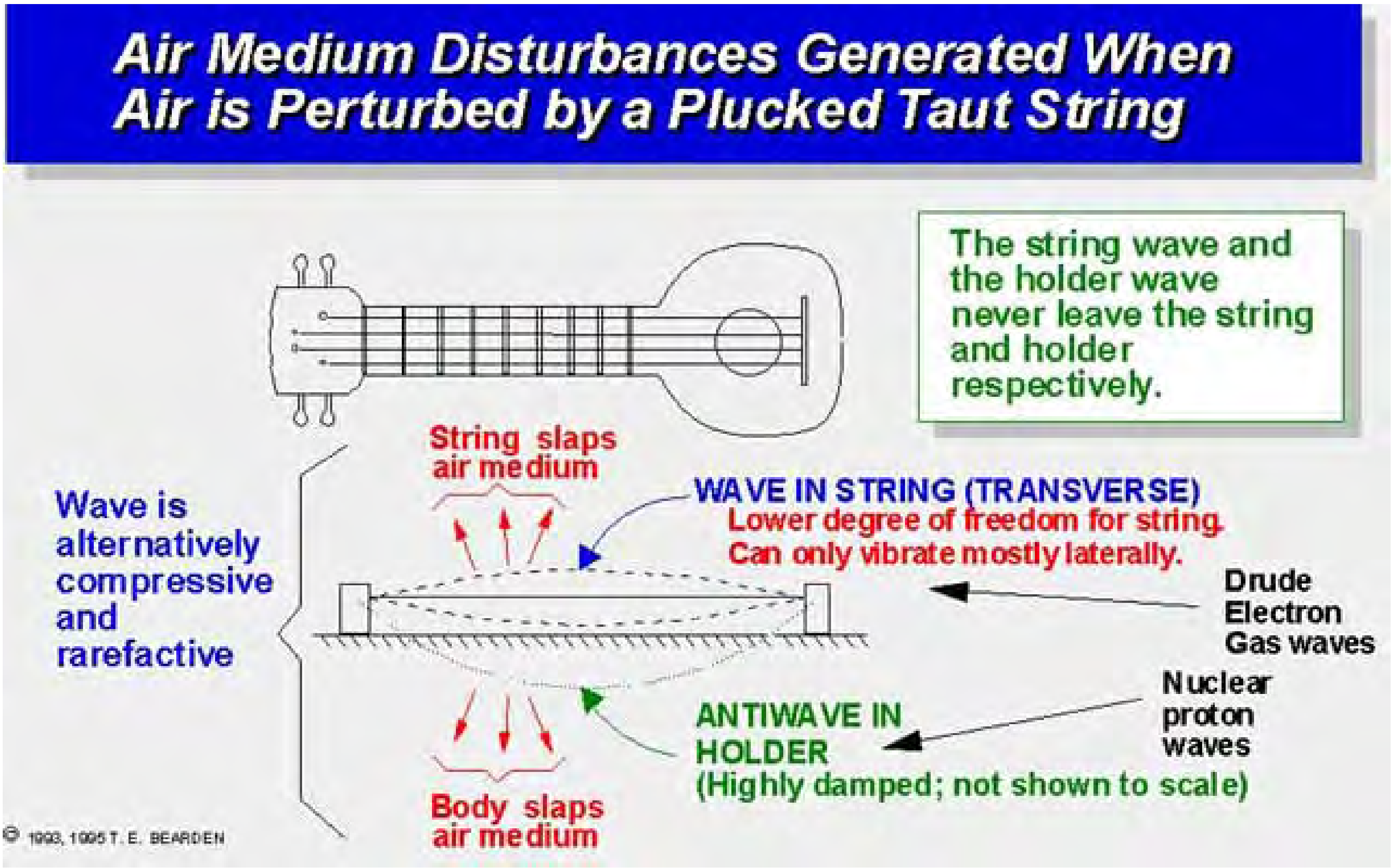

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)
Lower degree of freedom for string.
Can only vibrate mostly laterally.

Wave is alternatively compressive and rarefactive

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

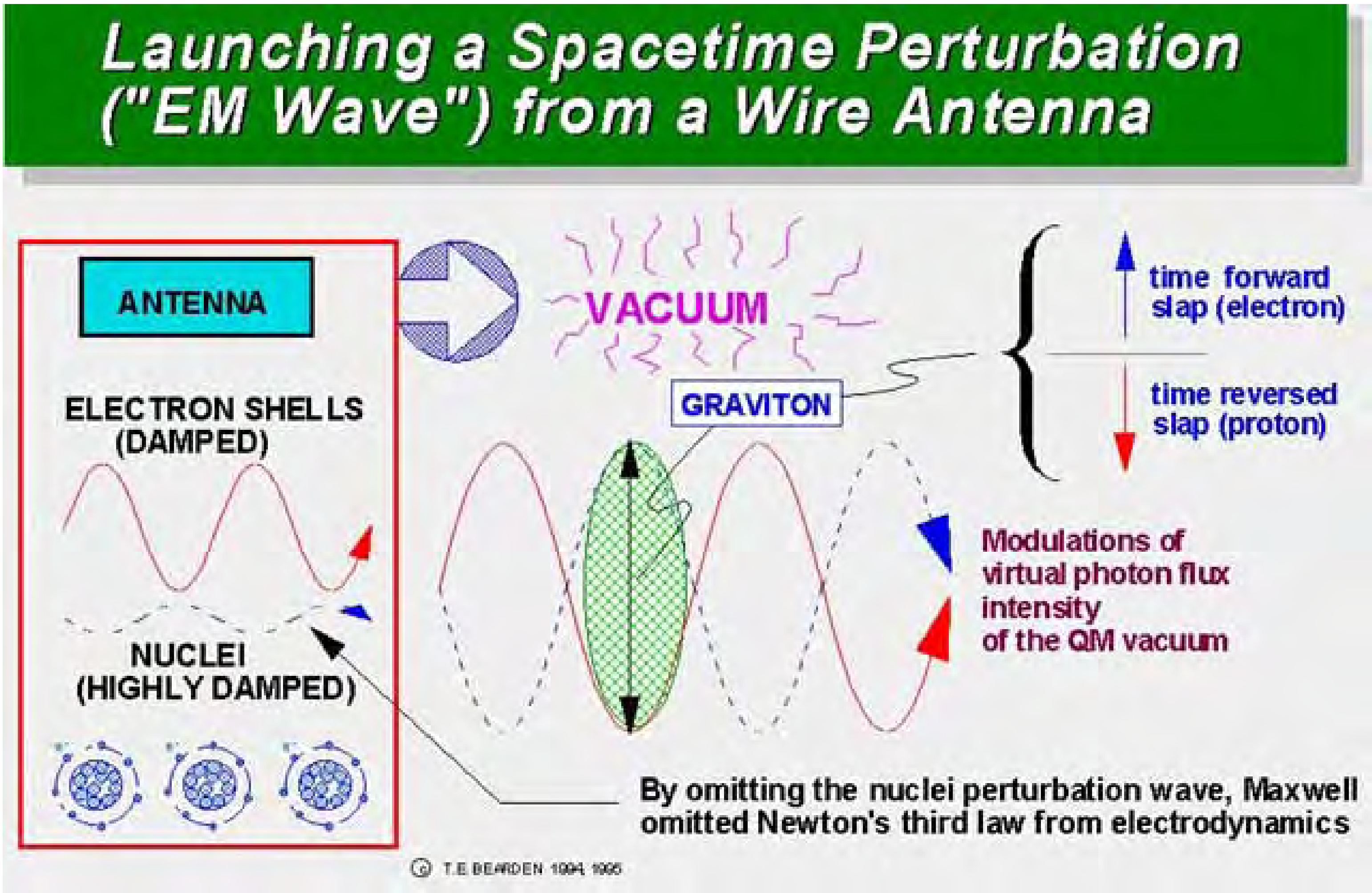

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA

ELECTRON SHELLS (DAMPED)

NUCLEI (HIGHLY DAMPED)

VACUUM

GRAVITON

time forward slap (electron)

time reversed slap (proton)

Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

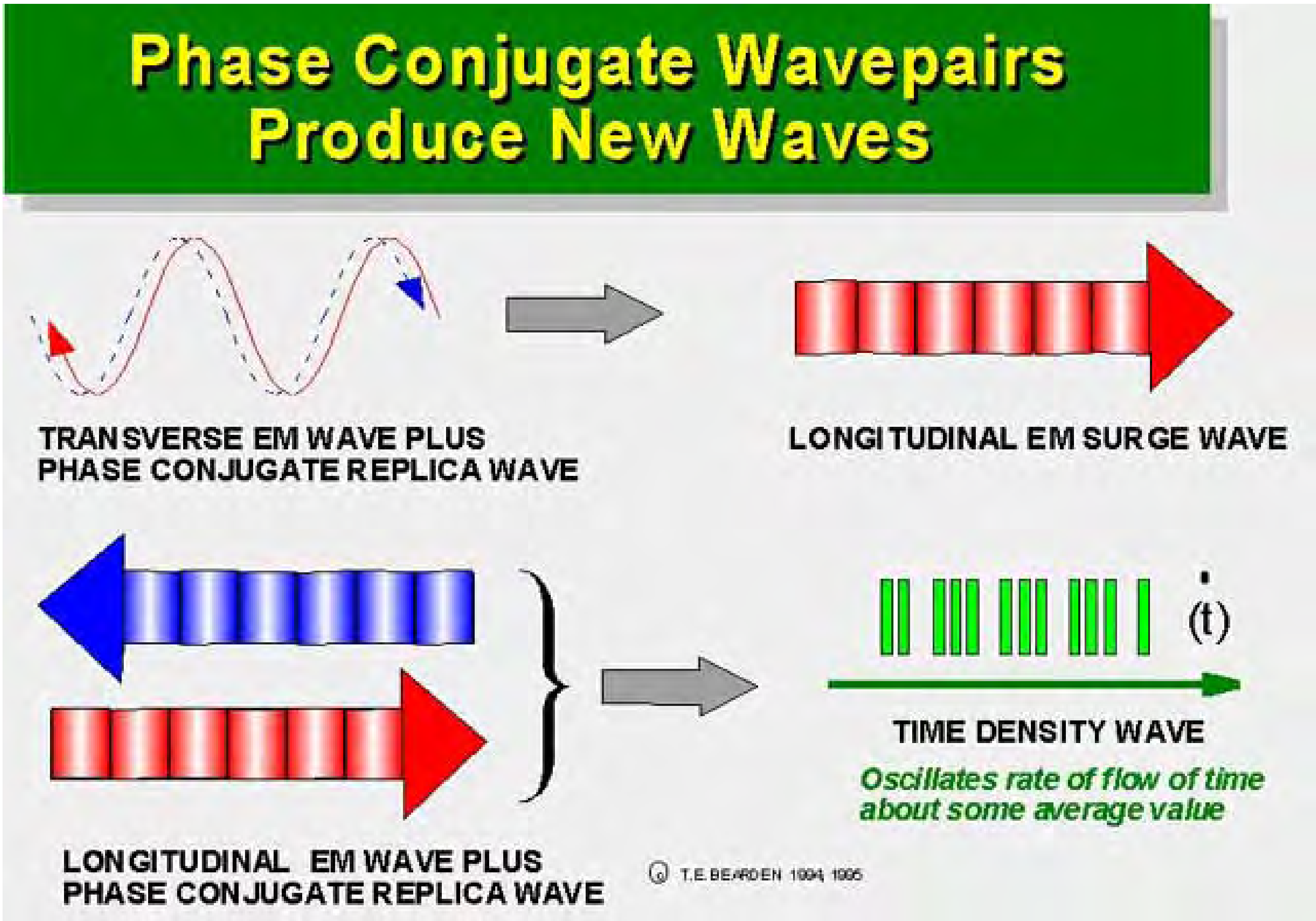

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

$(\dot{t})$

TIME DENSITY WAVE

*Oscillates rate of flow of time
about some average value*

© T.E. BEARDEN 1994, 1995

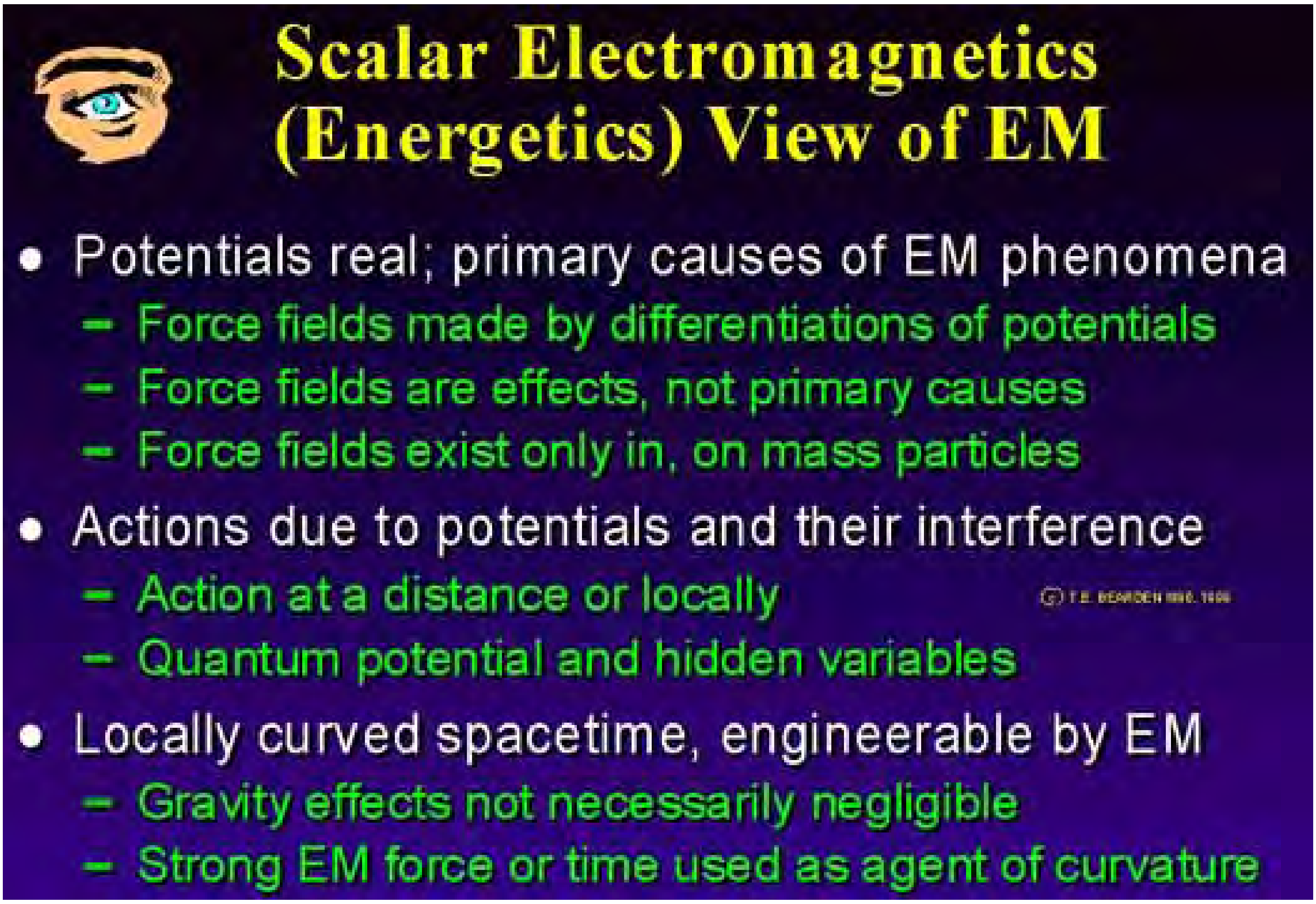

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN 1990, 1996

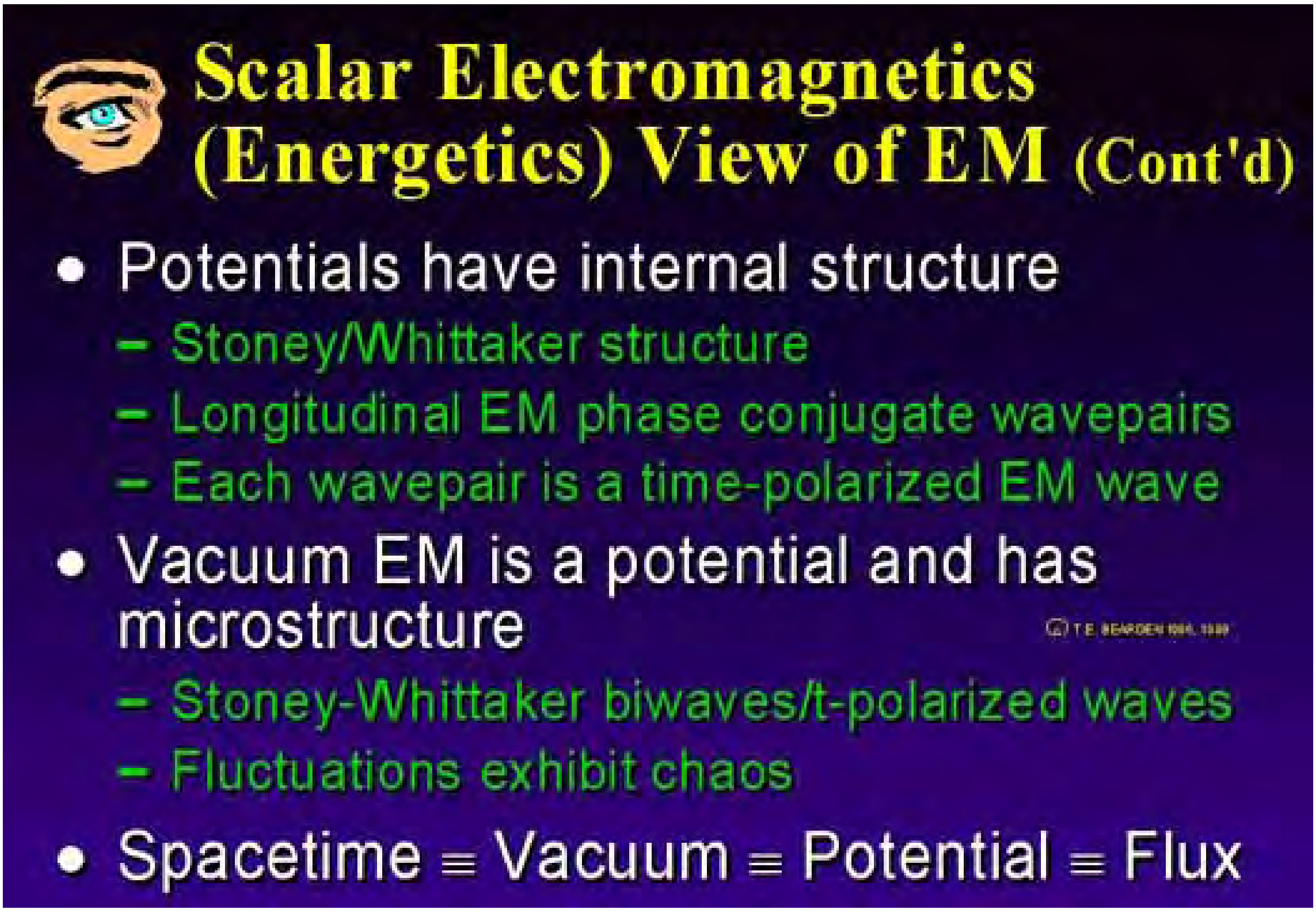

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

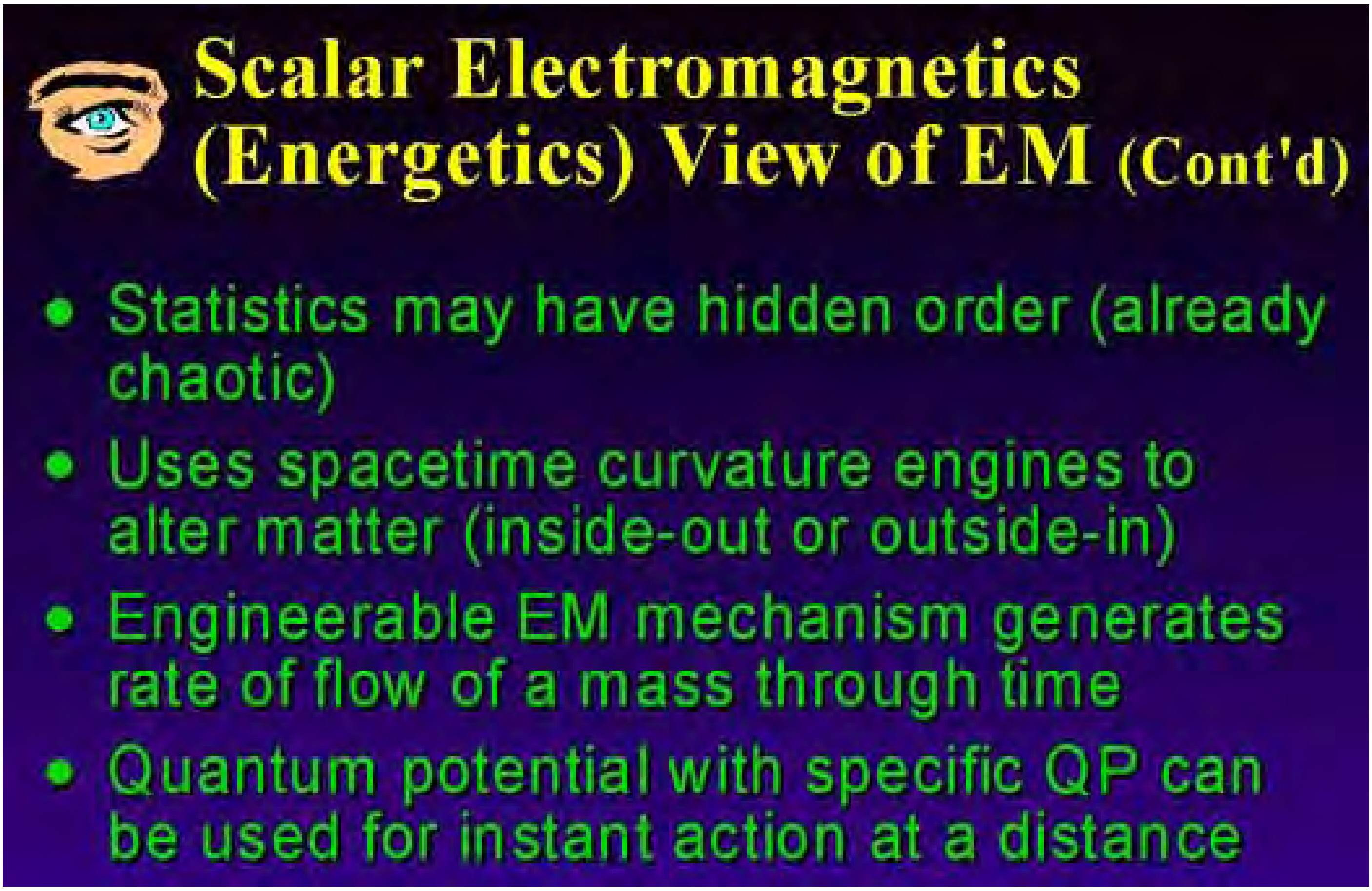

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

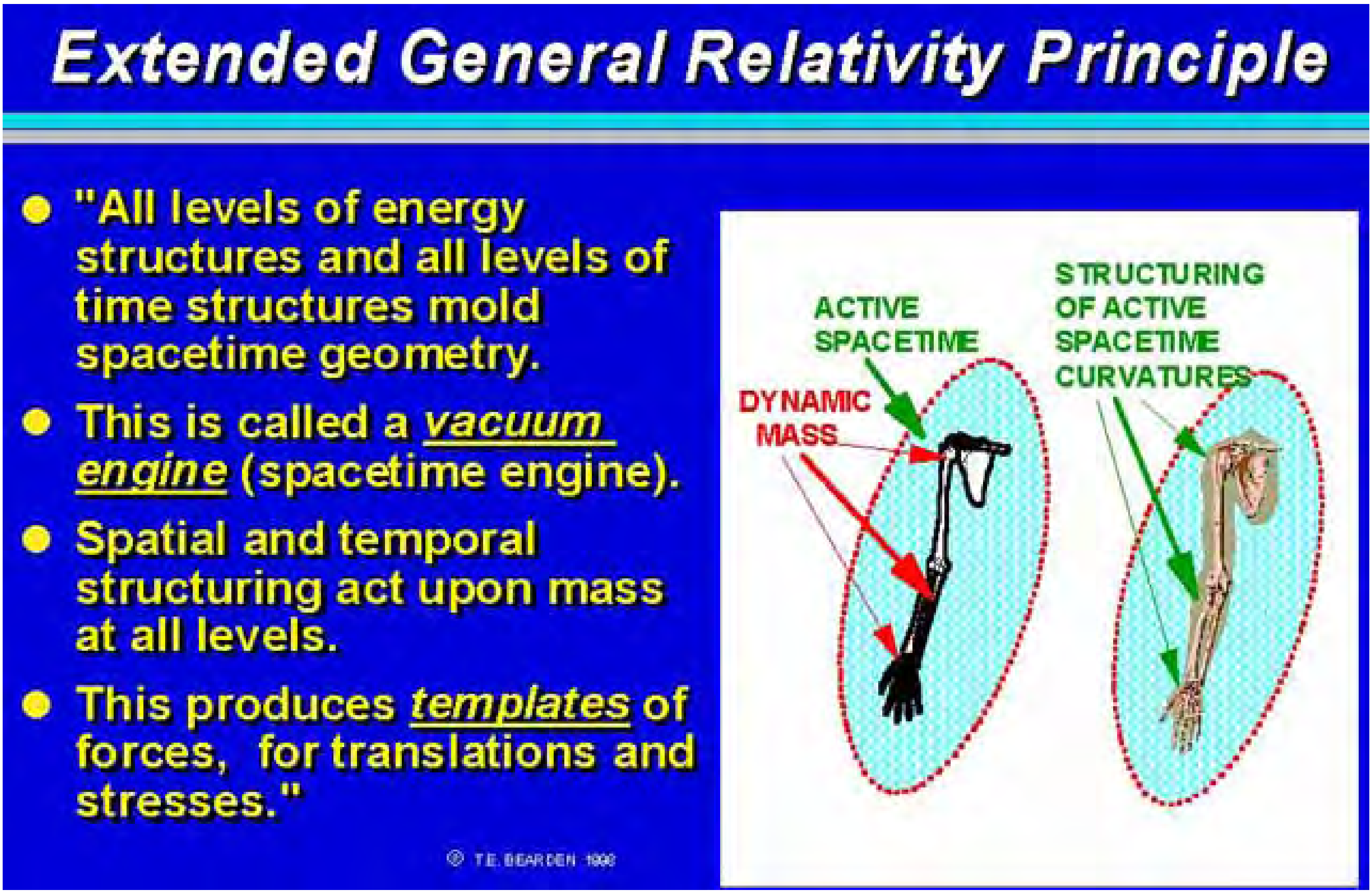

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a ***vacuum engine*** (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces ***templates*** of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

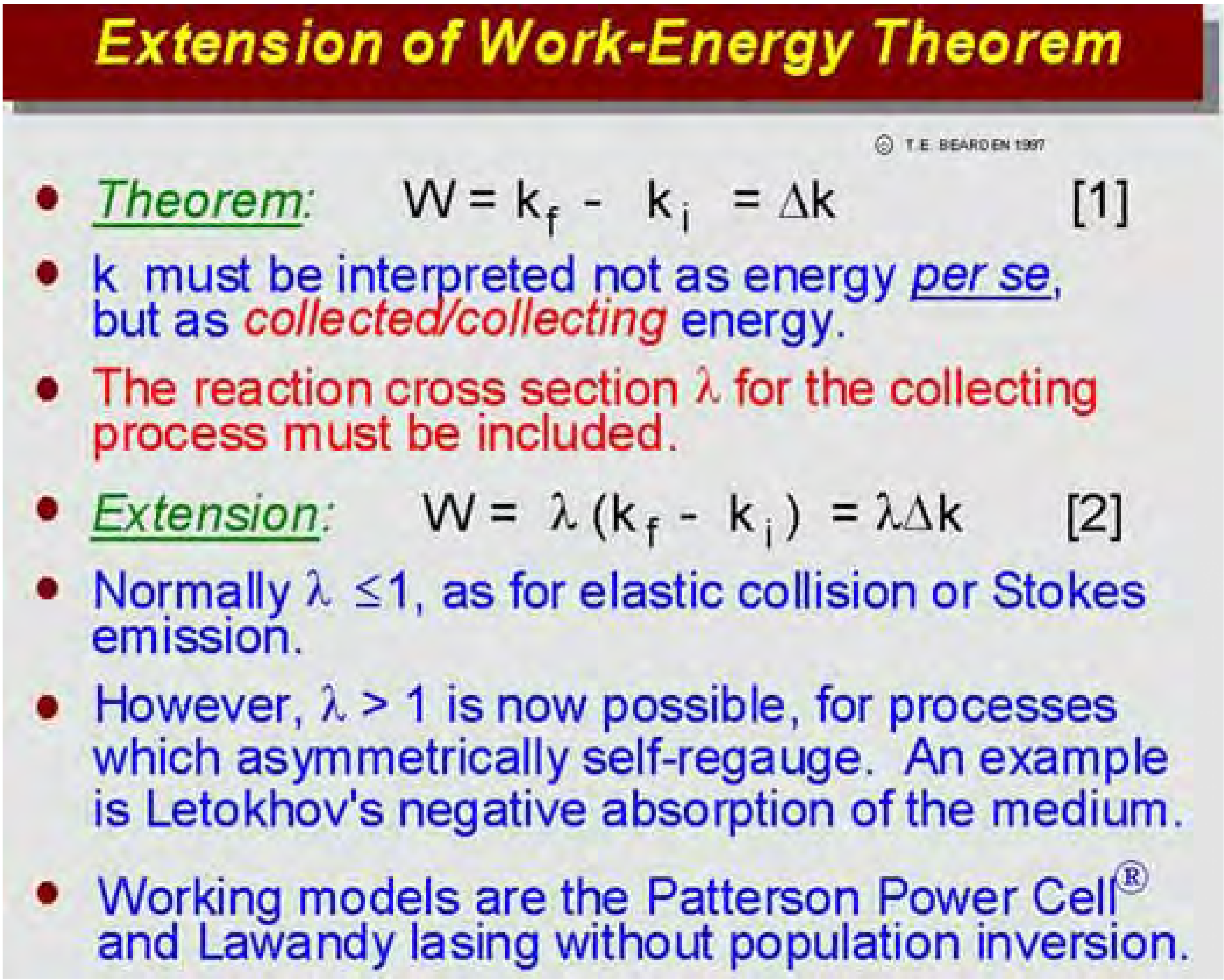

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem*: $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension*: $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

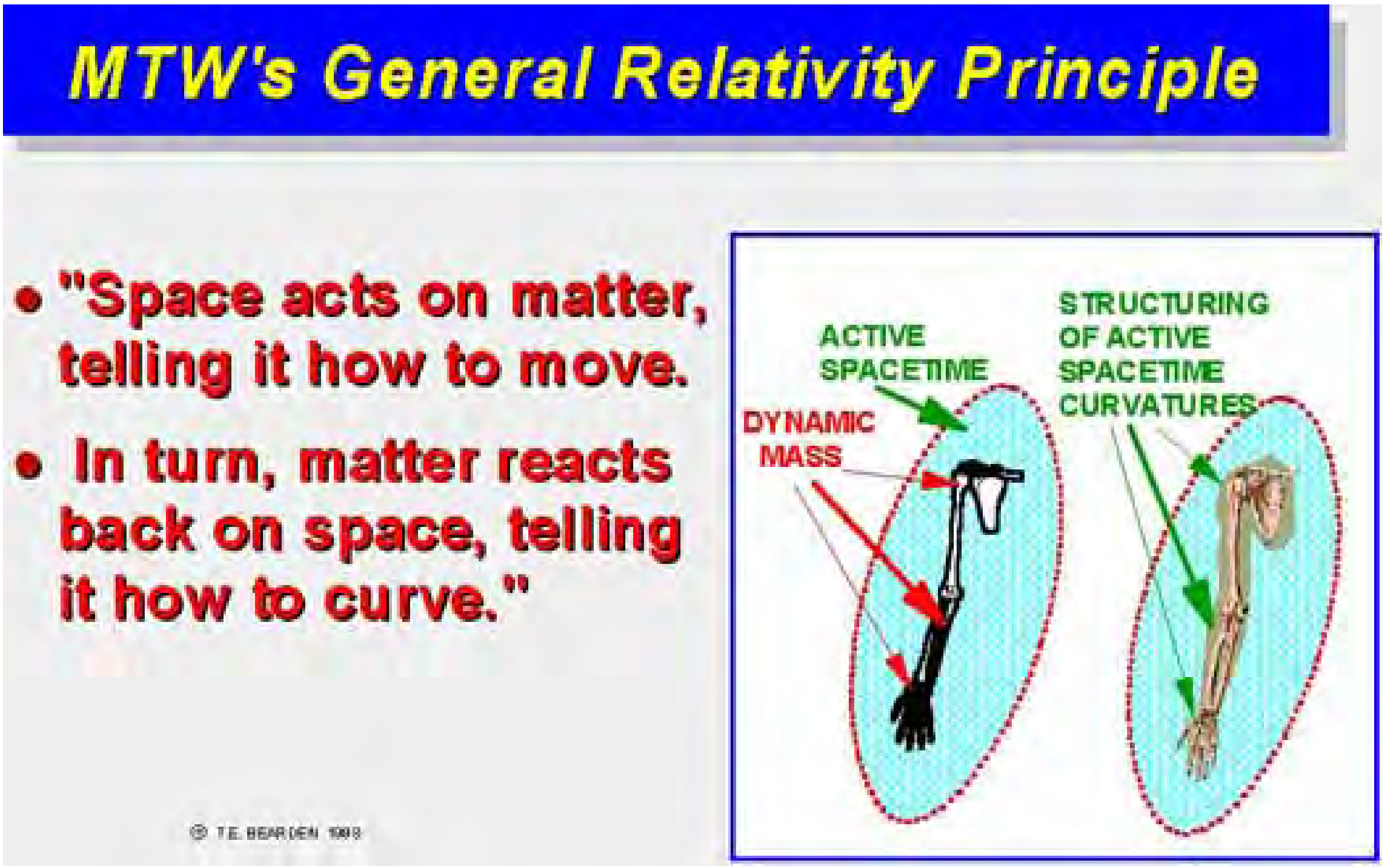

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

© T.E. BEARDEN 1998

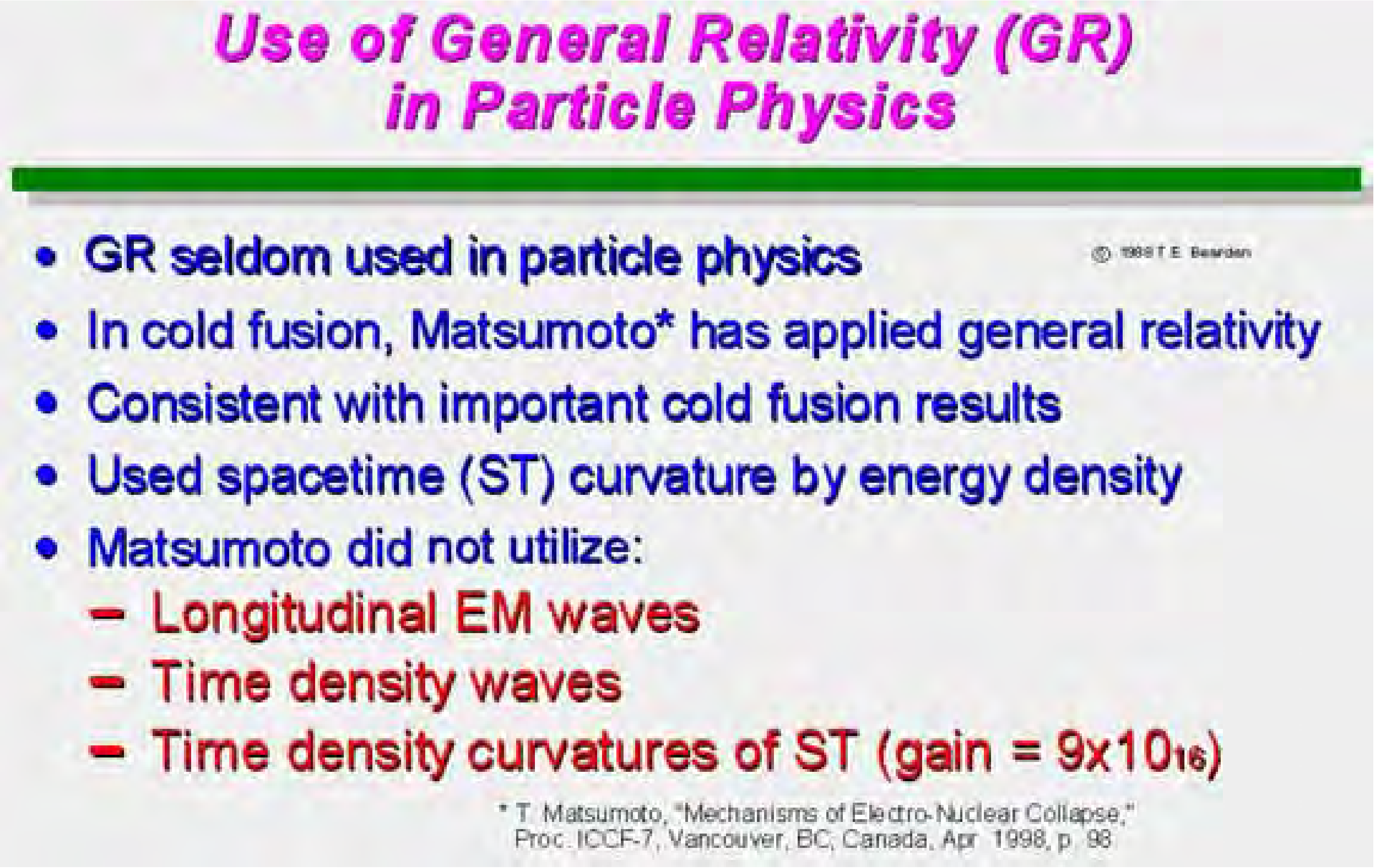

Use of General Relativity (GR) in Particle Physics

© 1998 T E Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = 9x10$_{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98

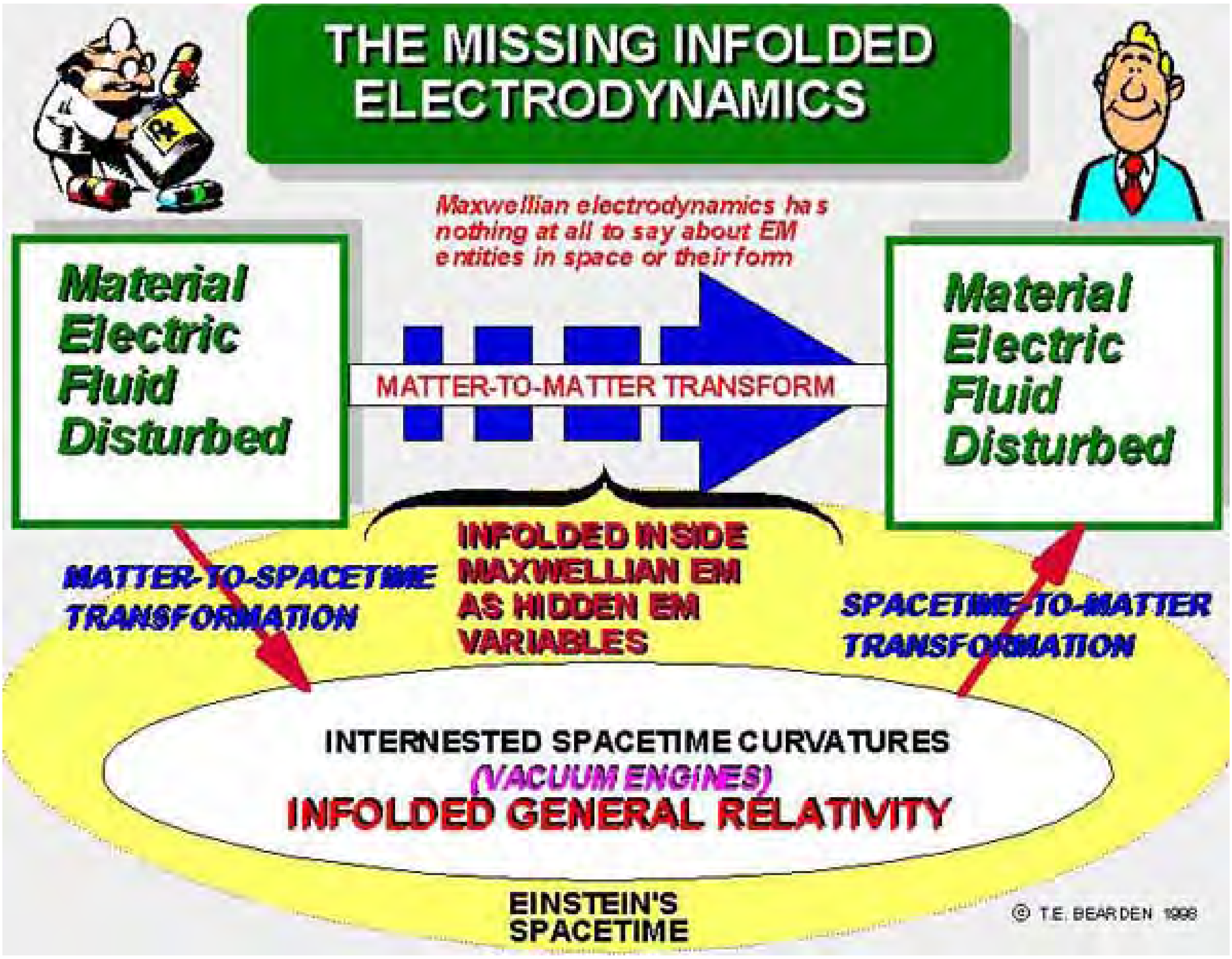

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

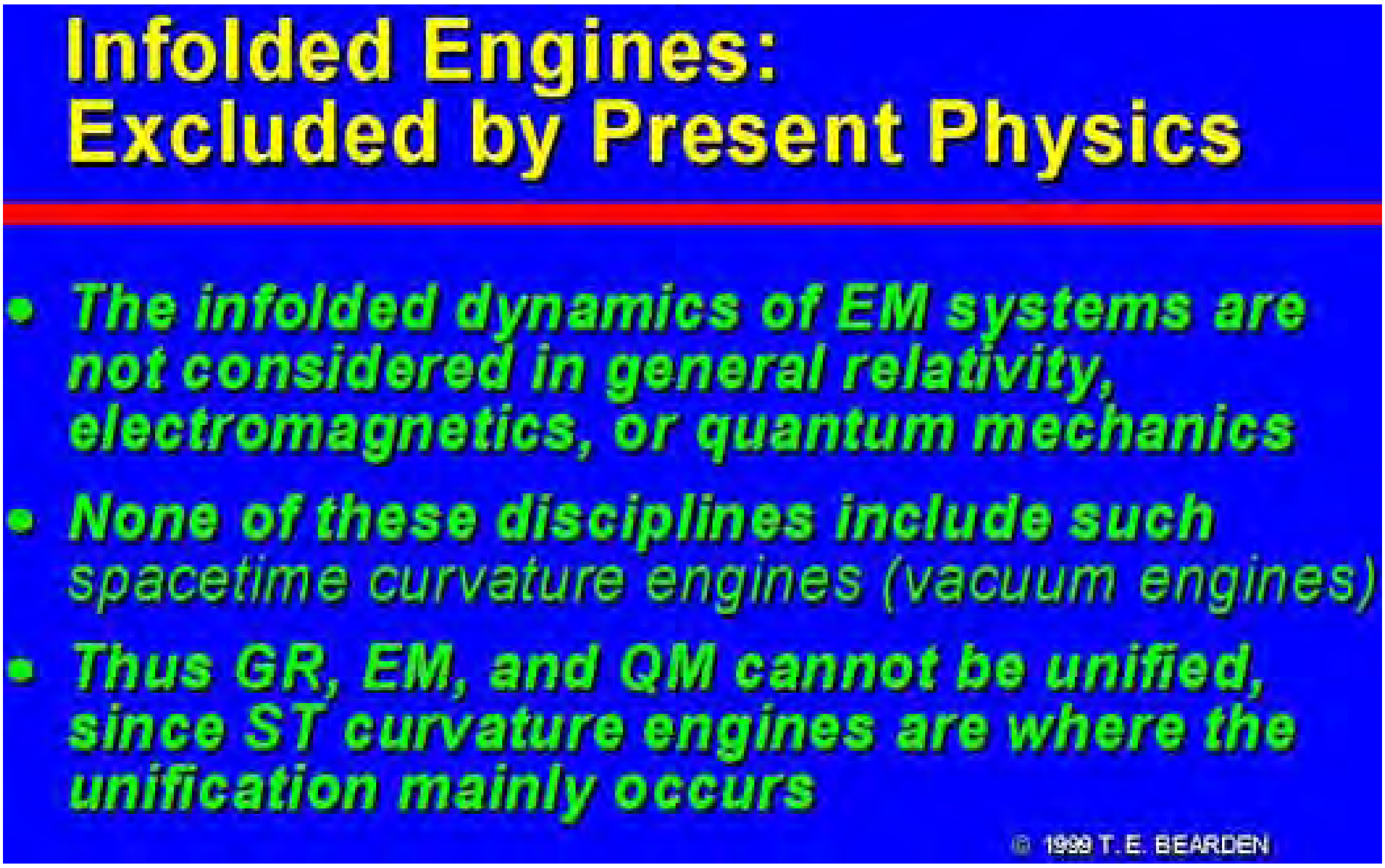

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

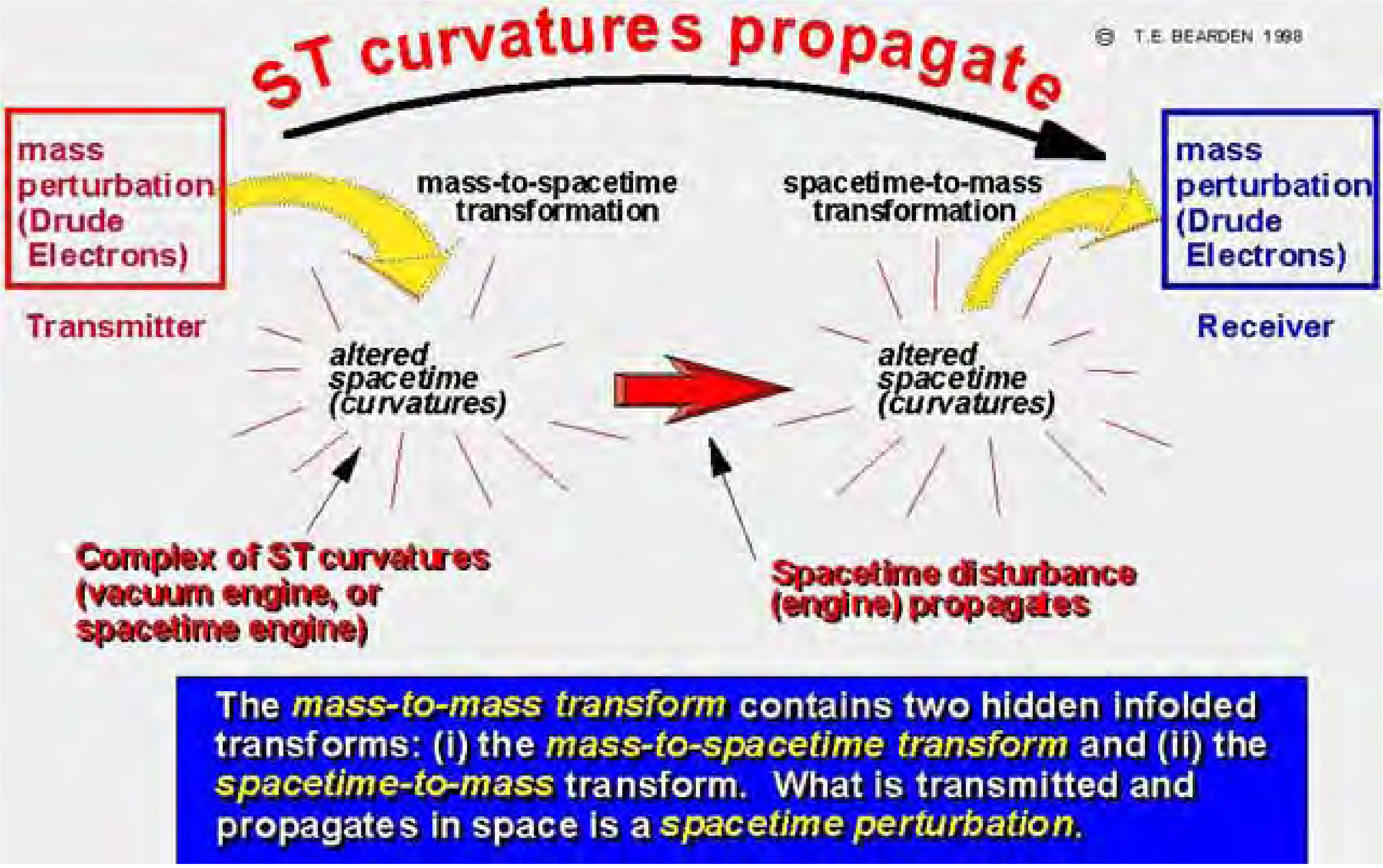

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

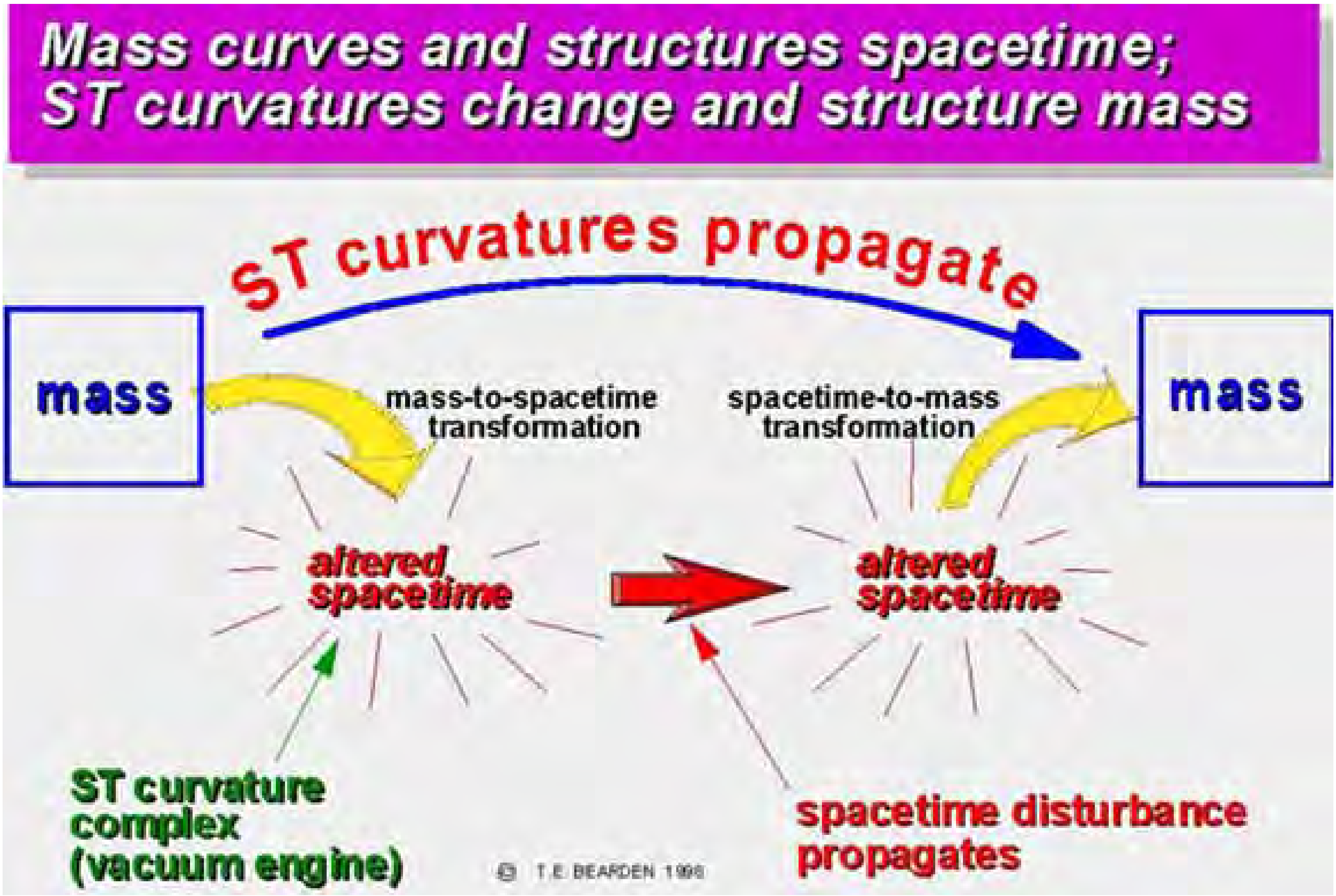

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

altered spacetime

spacetime-to-mass transformation

altered spacetime

mass

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

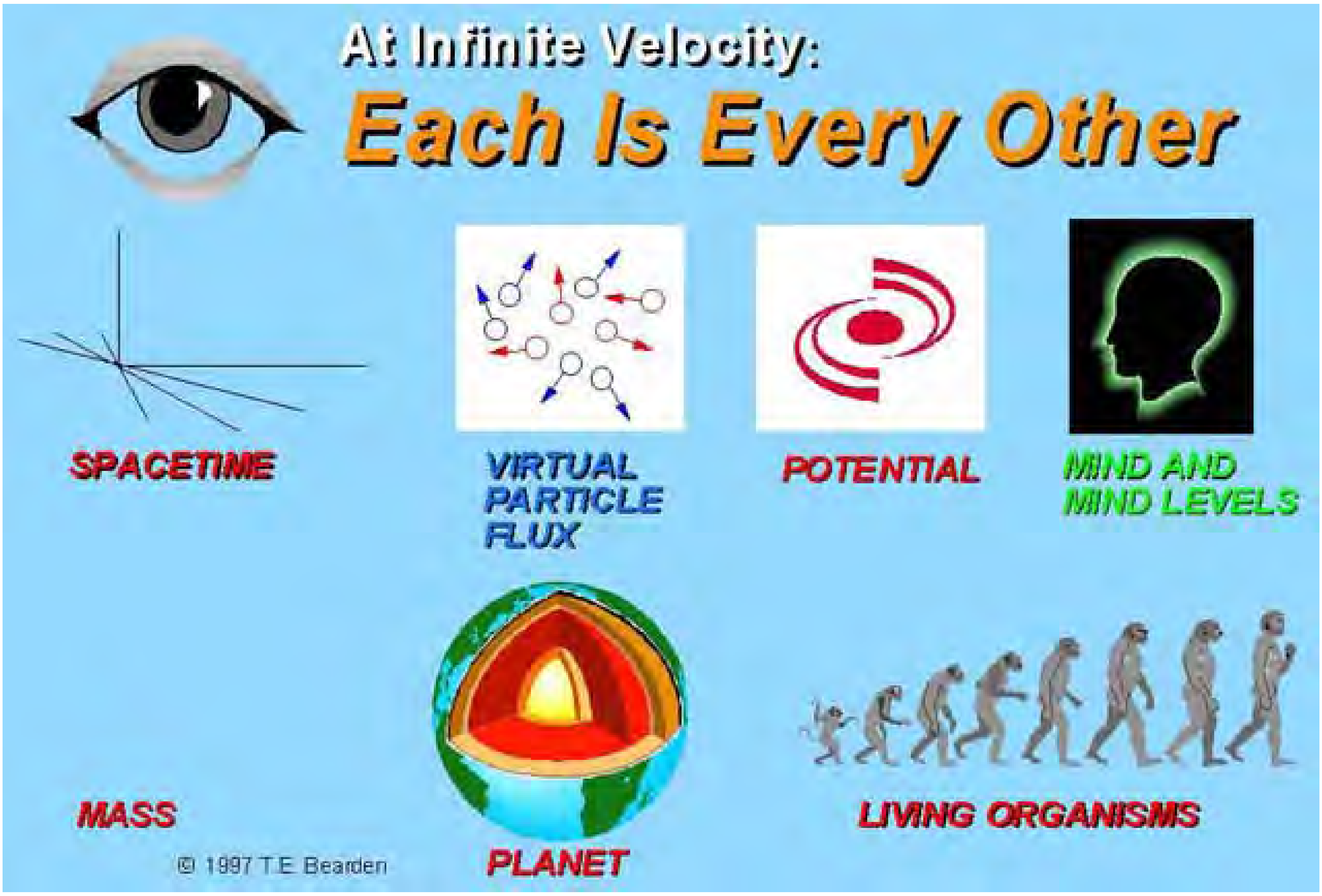

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

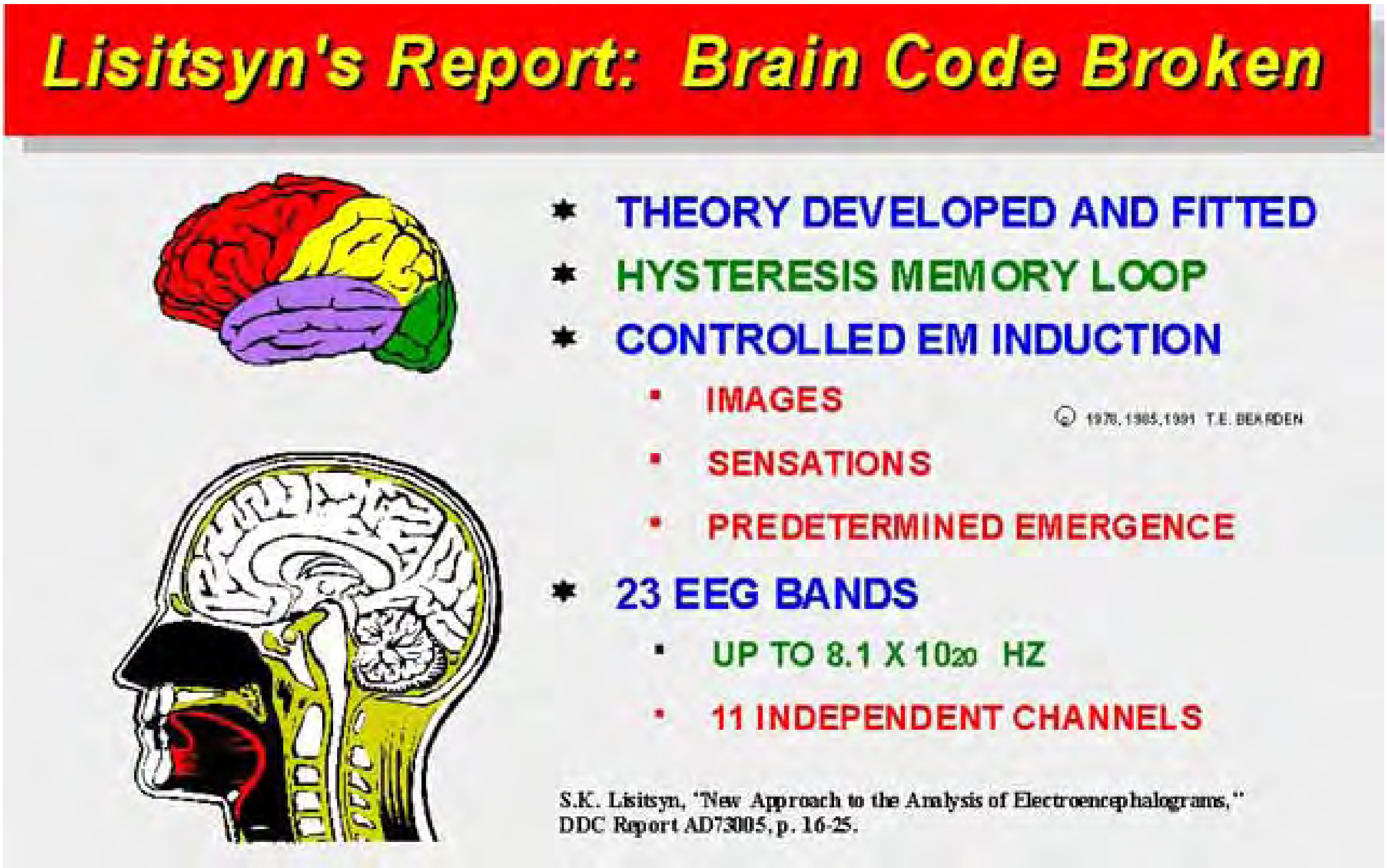

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

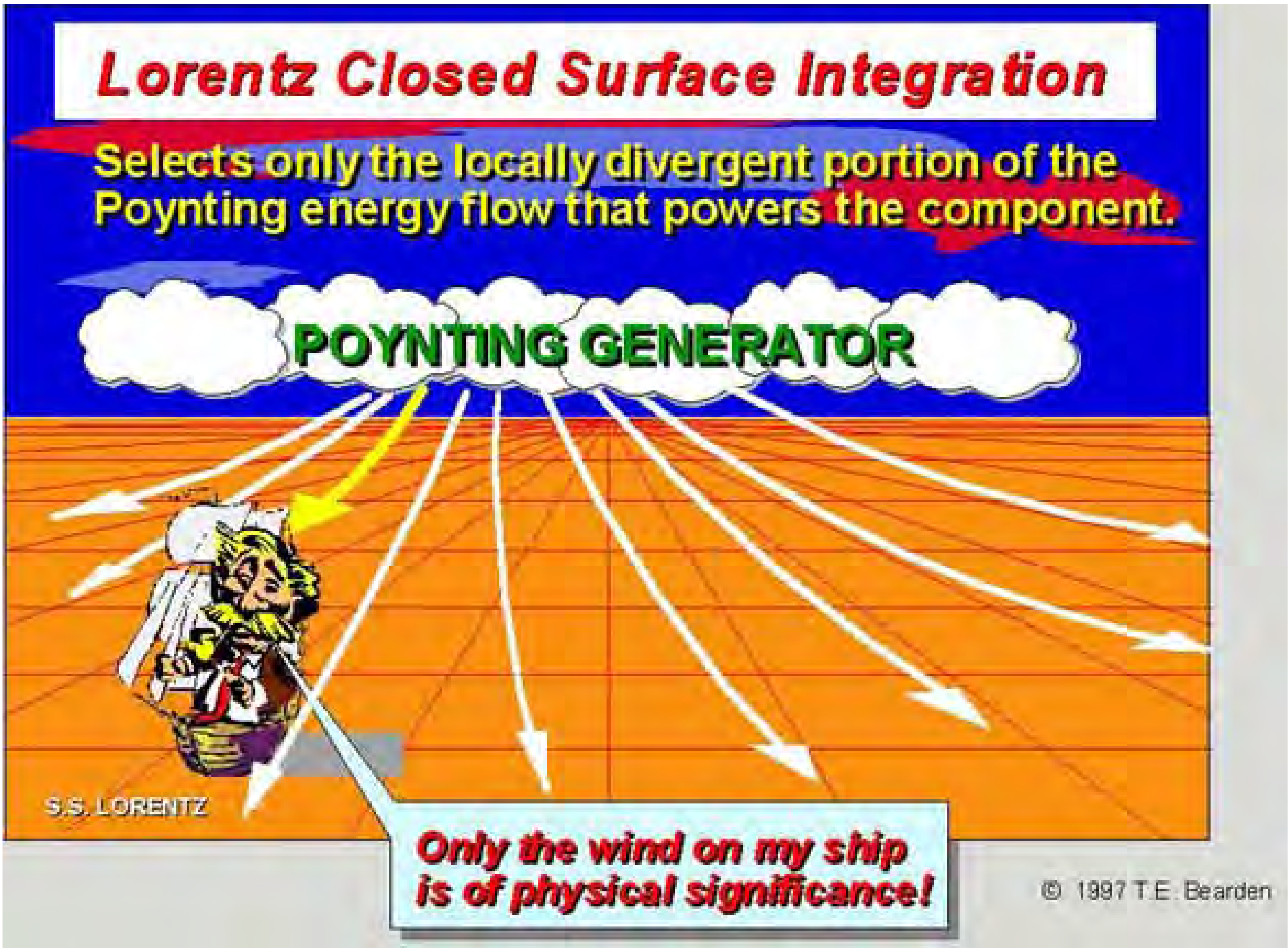

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

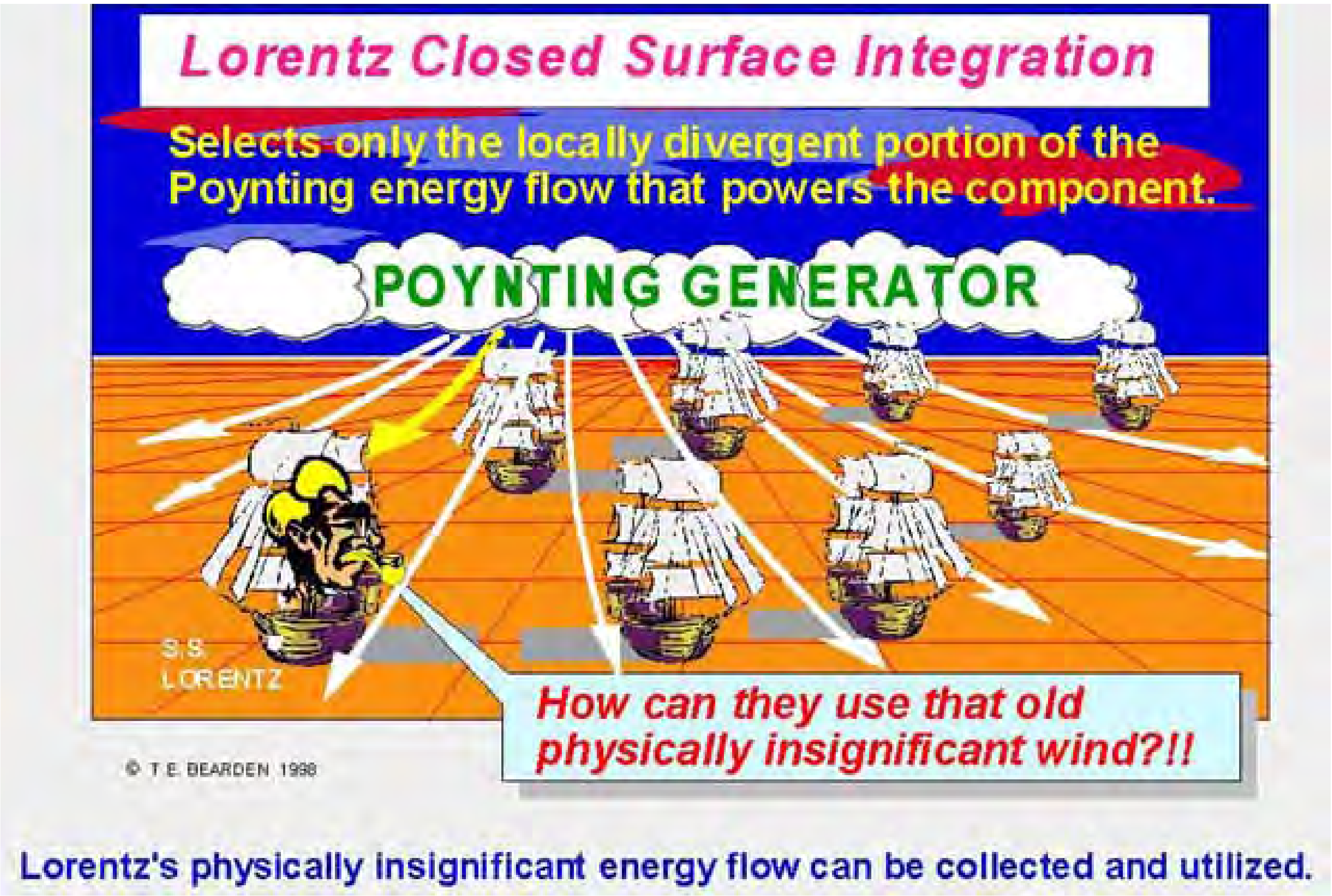

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T.E. BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

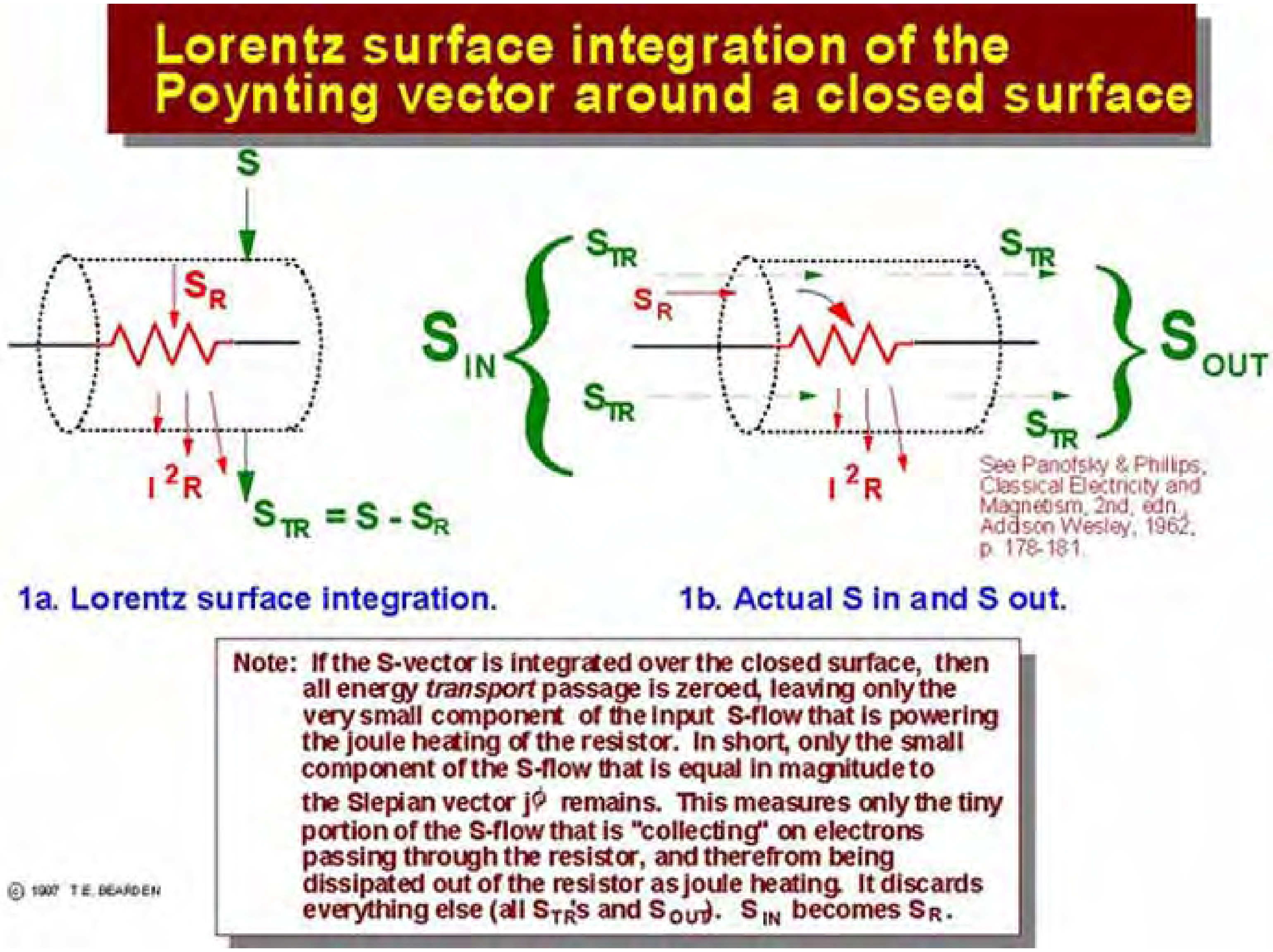

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector j^{ϕ} remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

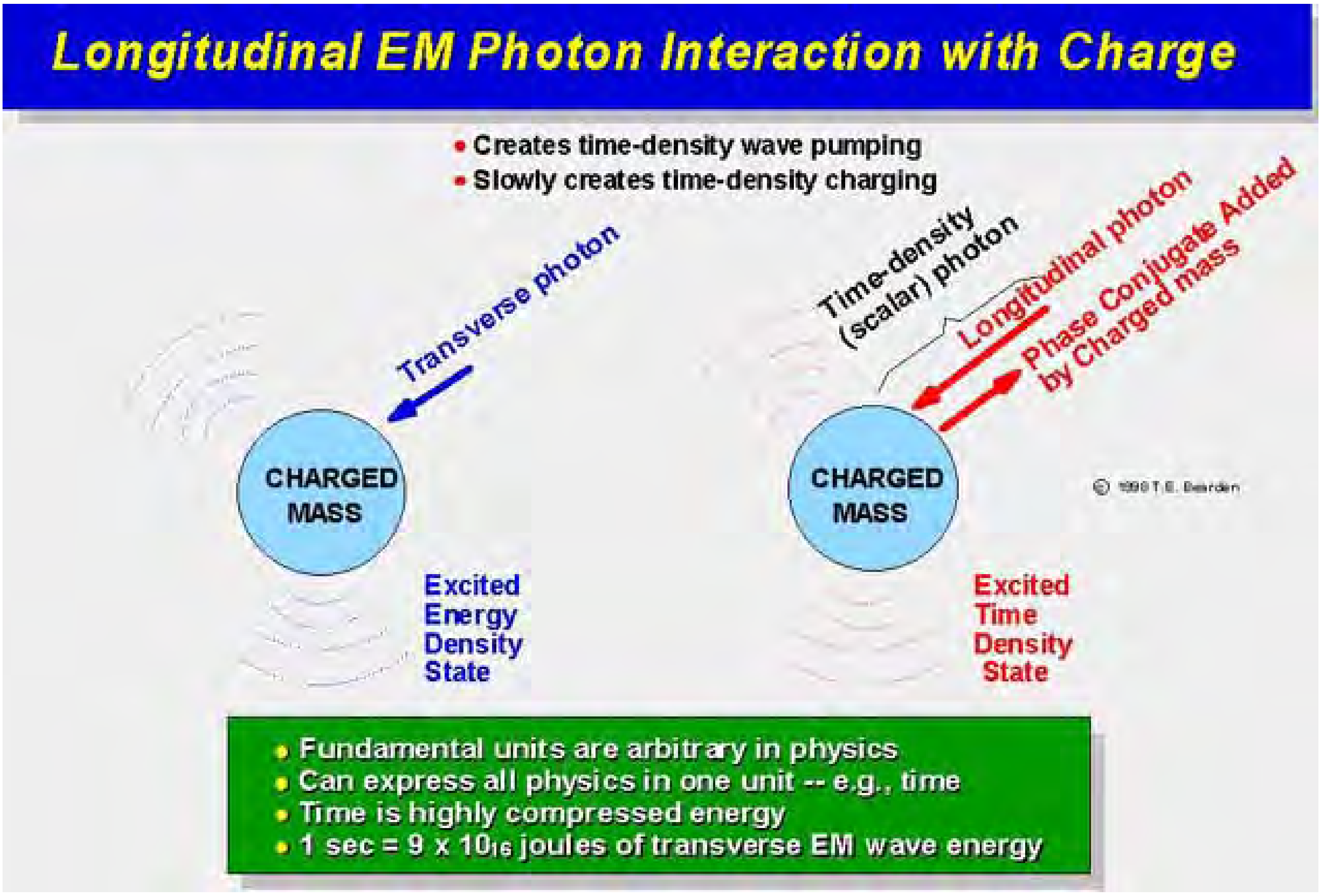

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10_{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

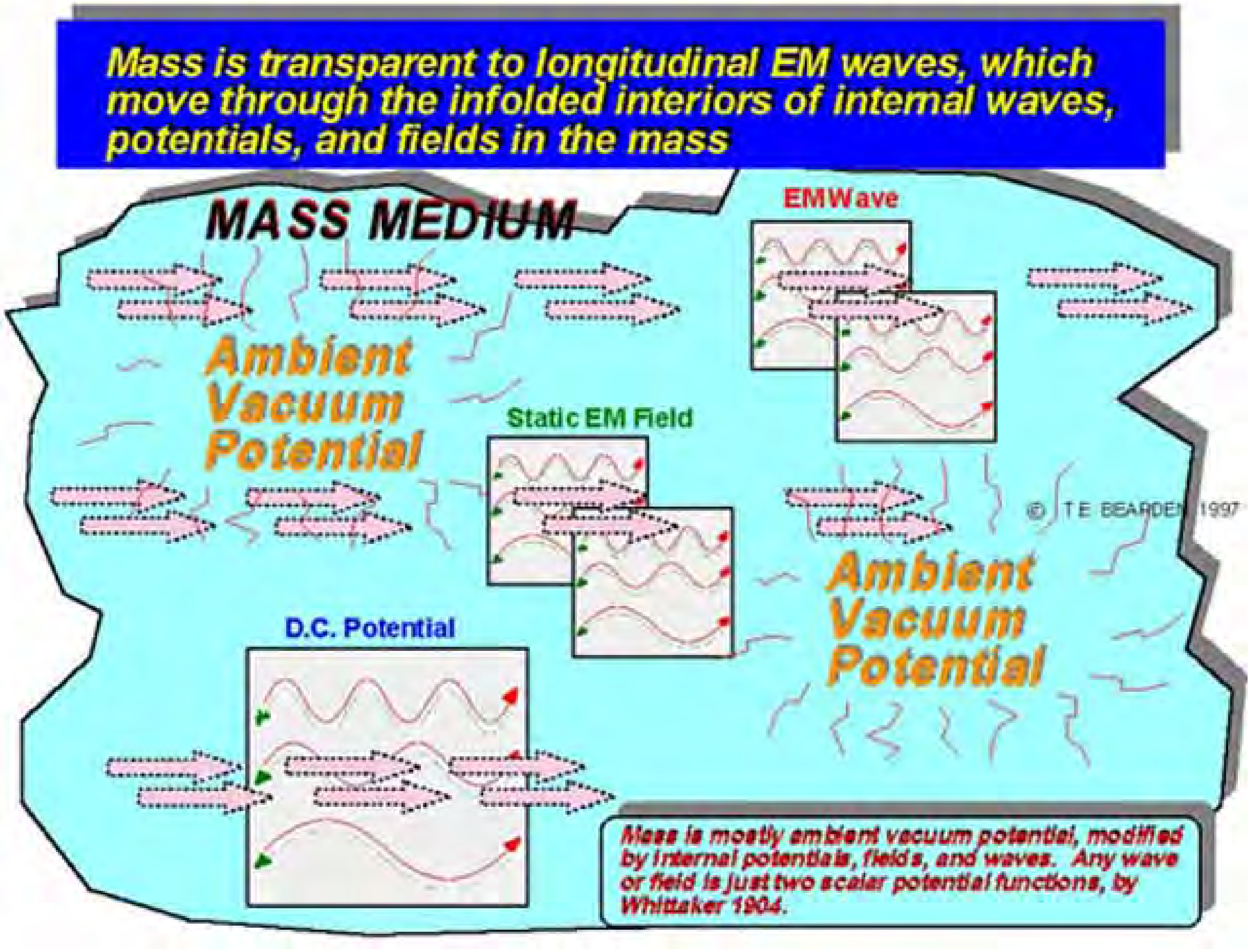

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EMWave
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

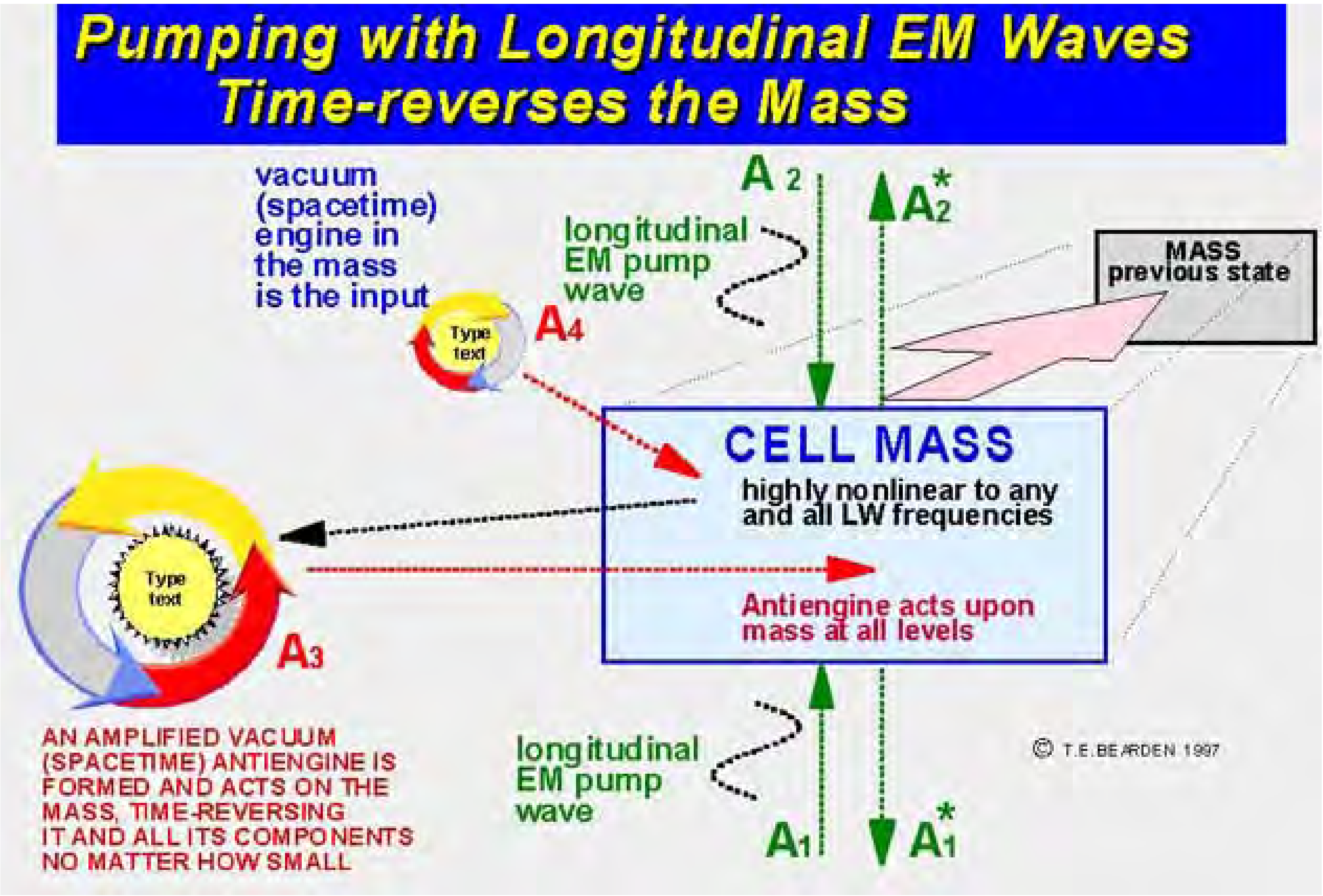

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum (spacetime) engine in the mass is the input

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

A 2
A*2
longitudinal EM pump wave
MASS previous state
Type text
A4
Type text
A3
longitudinal EM pump wave
A1
A*1

© T.E.BEARDEN 1997

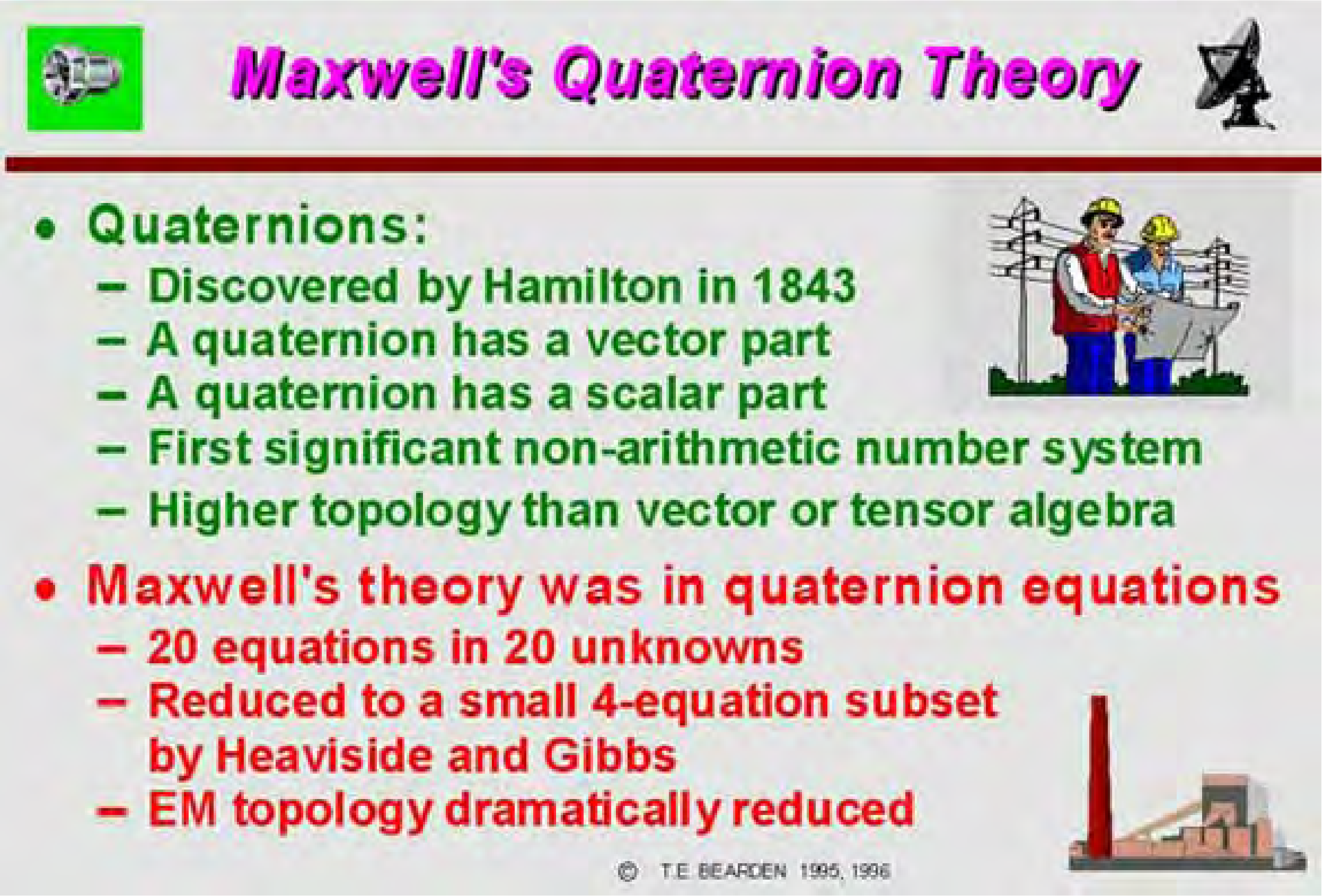

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

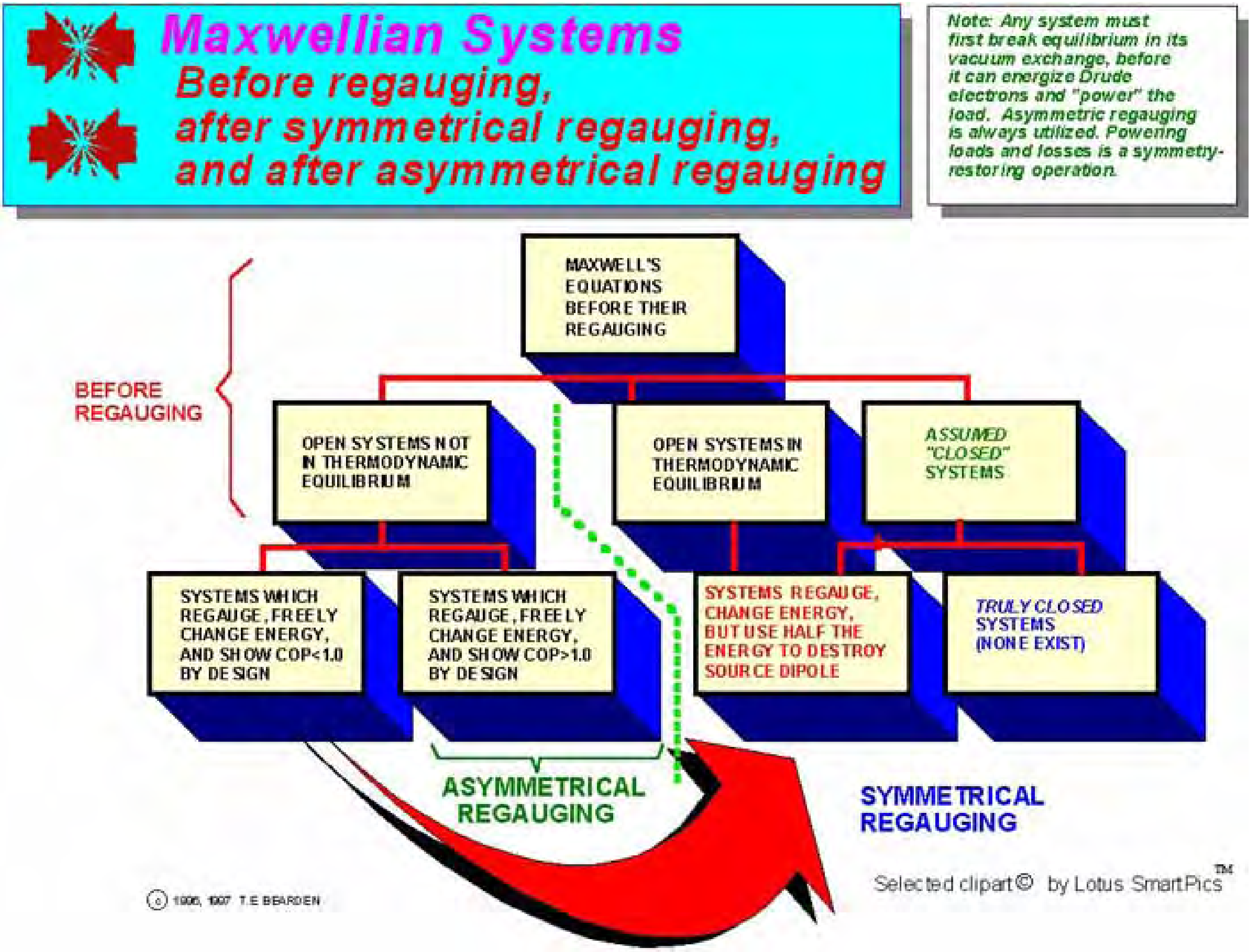

Maxwellian Systems
Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

ⓒ 1996, 1997 T.E BEARDEN

Selected clipart© by Lotus SmartPics™

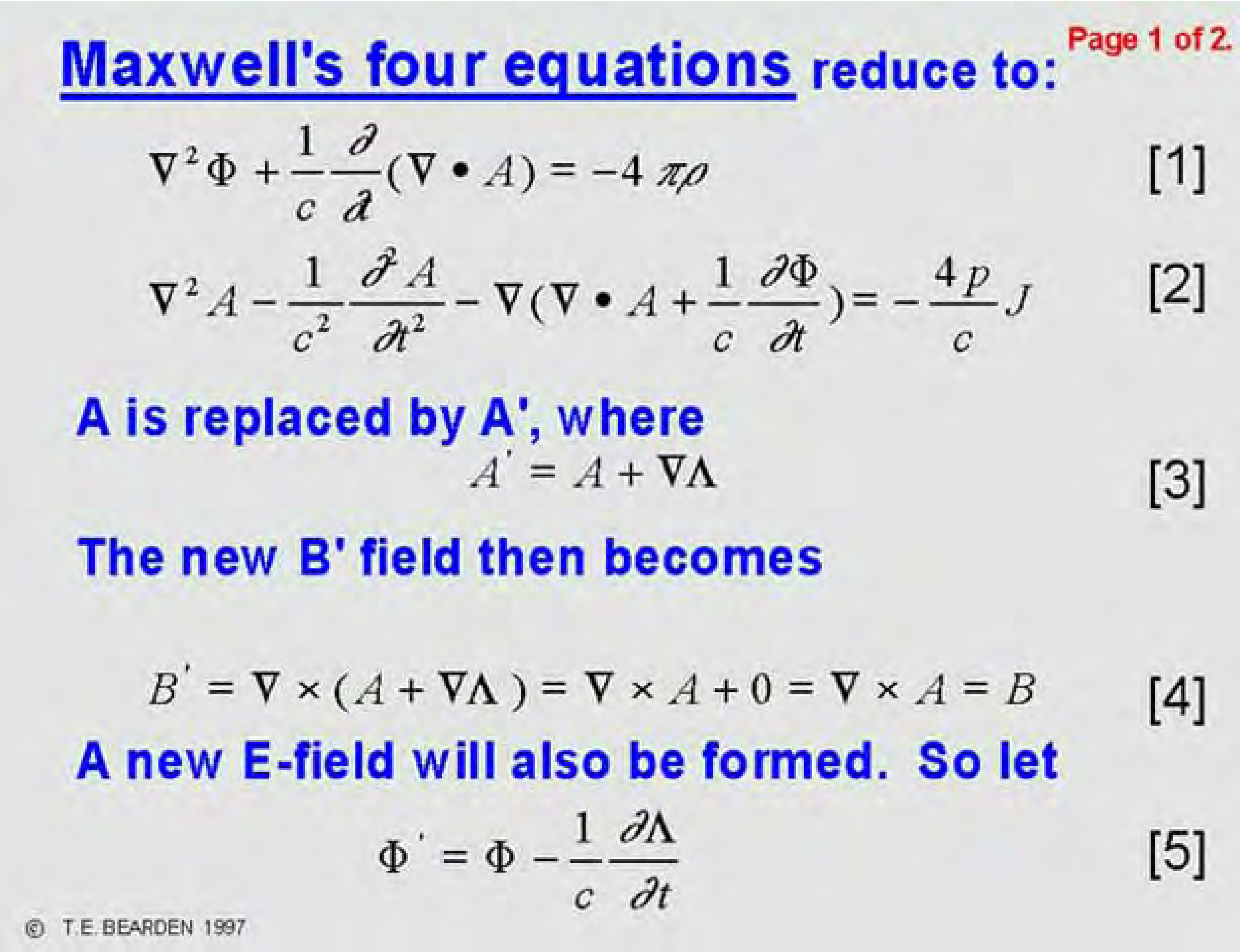

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla(\nabla \bullet A + \frac{1}{c}\frac{\partial\Phi}{\partial t}) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

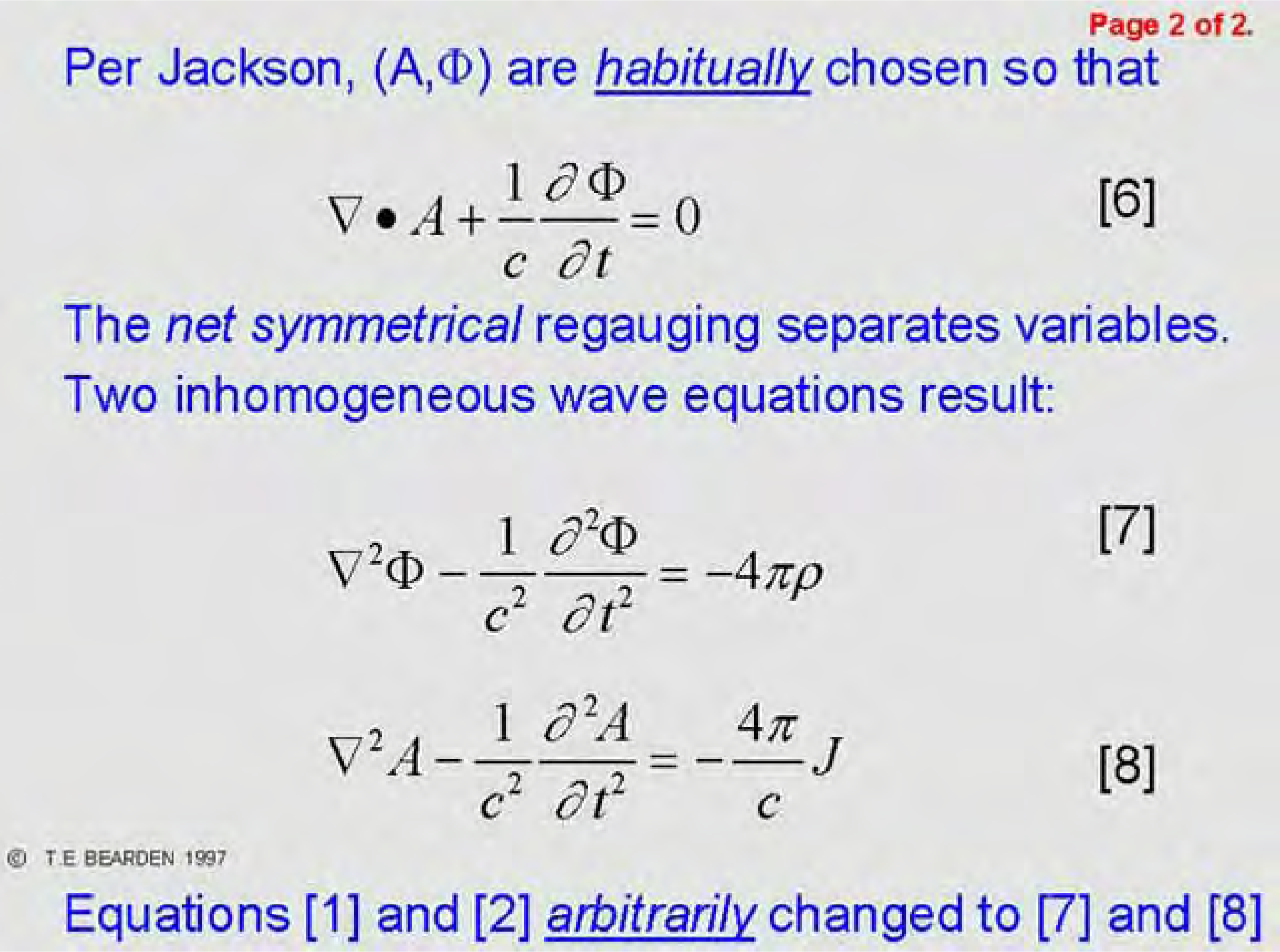

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.

Two inhomogeneous wave equations result:

$$\nabla^2\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c}J \qquad [8]$$

© T.E BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Life Systems

Toward the Beginning of the End

---

1. Toward the Beginning of the End: Capstone Presentations
2. Toward the Beginning of the End
3. Biospheric Evolution
4. A Living Biological System
5. The Evolutionary Path of Multicellular Organisms
6. To Paraphrase Churchill
7. Directed Evolution: Mechanism for Species Reactive Adaptation
8. Earth's Poynting Energy Flow
9. The Immune System
10. Mind and Matter Interaction
11. Levels of Consciousness and Unconsciousness
12. Access to Mind, Memory and Thought
13. US Intelligence Statement on Soviet Psi Research
14. Mind and Matter Interaction
15. Conscious and Unconscious Minds
16. Velocity is Rotation
17. Species Reactive Adaptation: Directed Evolution
18. Technology's Goal
19. Technology's End
20. Technological Species' Problems

The Tom Bearden Website

Briefings

One of Tom's first public lectures: World Futures Society 1982
Other panelists included James Beale, Carol Rosin, Barbara Marx Hubbard and Glenn Doman

---

The "Porthole" Concept

Extension of the Priore Method of Healing as applicable to the treatment of biological warfare effects and cellular afflictions such as cancers, AIDS etc.

NEW

Hubbert's Peak: The Impending Oil Crisis, by Professor Kenneth S. Deffeyes

Book Review and Slides

A National Report on America's Energy Crisis

by Energy Secretary Spencer Abraham - March 19, 2001

Commentary and Slide Show (66) by Tom Bearden

A Visual Tour of what they don't want you to know about electrical circuits
10 Trillion Percent of the Current Produced is Wasted!

Weapons Slides

The Missing Infolded Electrodynamics
85 slides

A Brief Technical Background to the Extended Priore Healing Methodology
11 slides

Briefing sent by Tom Bearden to Defense Department, National Institute of Health, USAF, National Science Foundation etc. re extended and improved Priore Treatment to counter mass casualties from Biological Warfare

- Letter to DHHS re adoption of Priore technology
  24 June, 1998
- Letter to NIH re adoption of Priore technology
  12 May, 1998

Towards the Beginning of the End - Lecture in Florida for Capstone Presentations, Sept. 1998

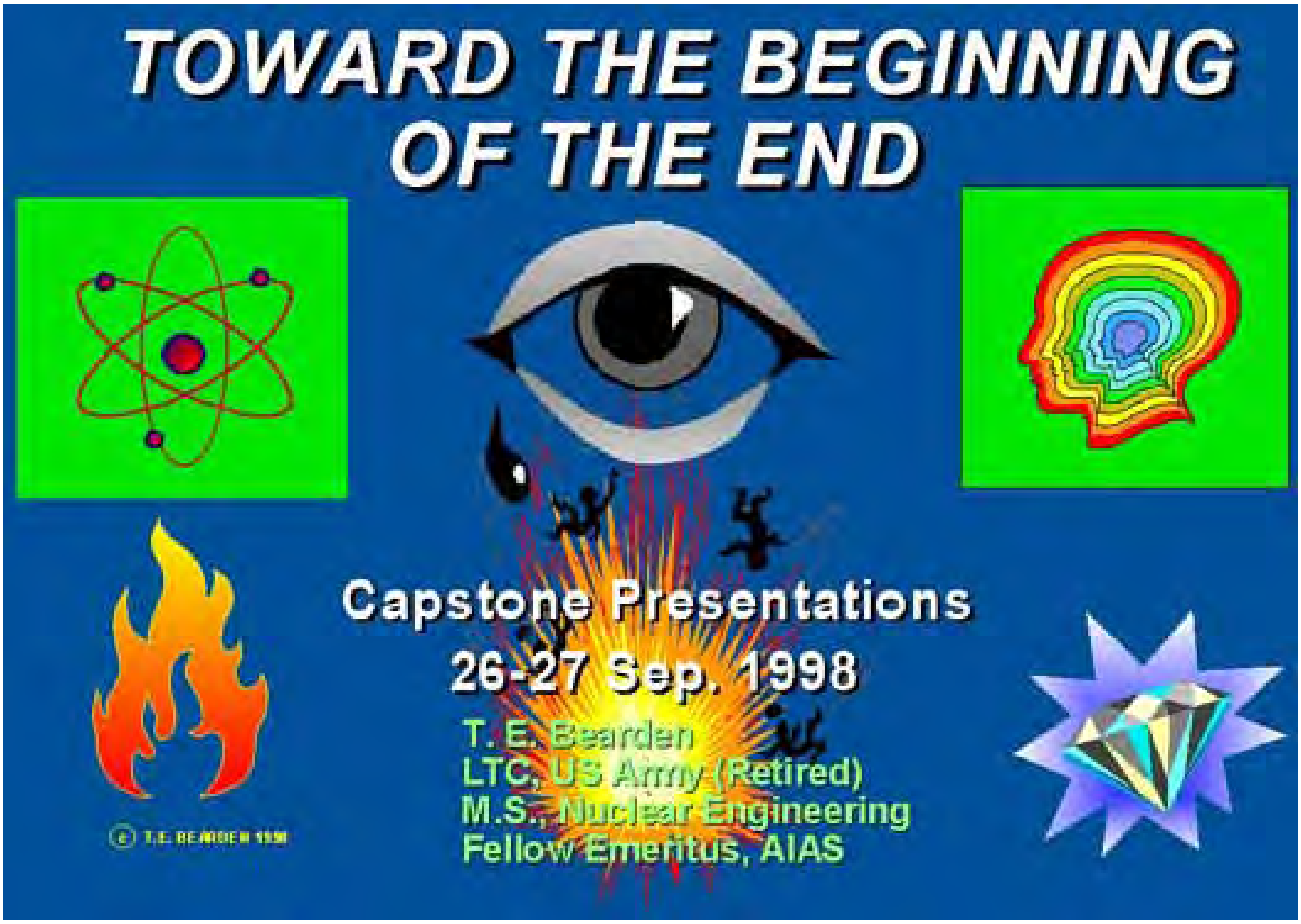

TOWARD THE BEGINNING OF THE END

Capstone Presentations

26-27 Sep. 1998

T. E. Bearden
LTC, US Army (Retired)
M.S., Nuclear Engineering
Fellow Emeritus, AIAS

© T.E. BEARDEN 1998

[Return to Slide Index](#)

[Next Slide](#)

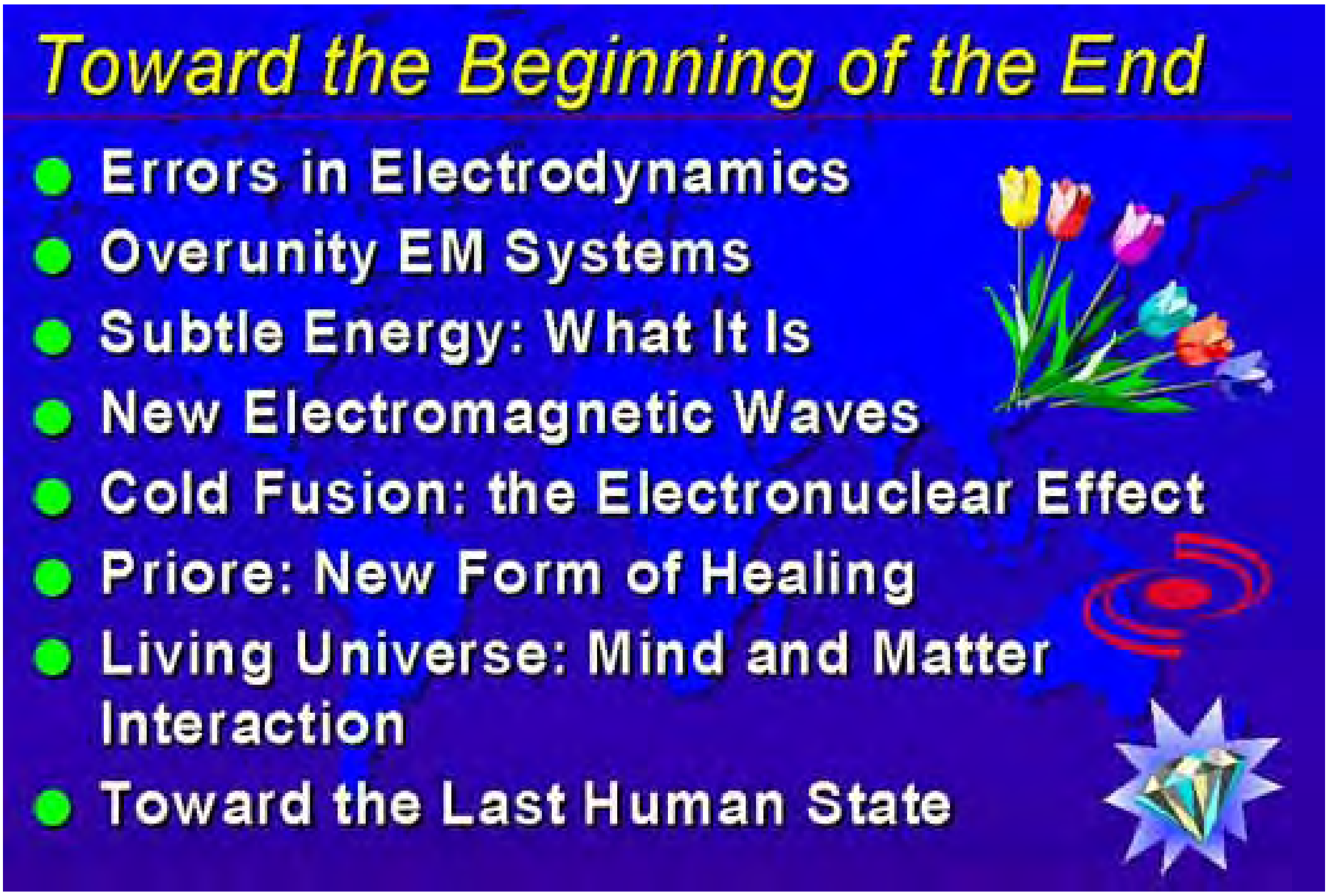

Toward the Beginning of the End

- Errors in Electrodynamics
- Overunity EM Systems
- Subtle Energy: What It Is
- New Electromagnetic Waves
- Cold Fusion: the Electronuclear Effect
- Priore: New Form of Healing
- Living Universe: Mind and Matter Interaction
- Toward the Last Human State

Previous Slide

Return to Slide Index

Next Slide

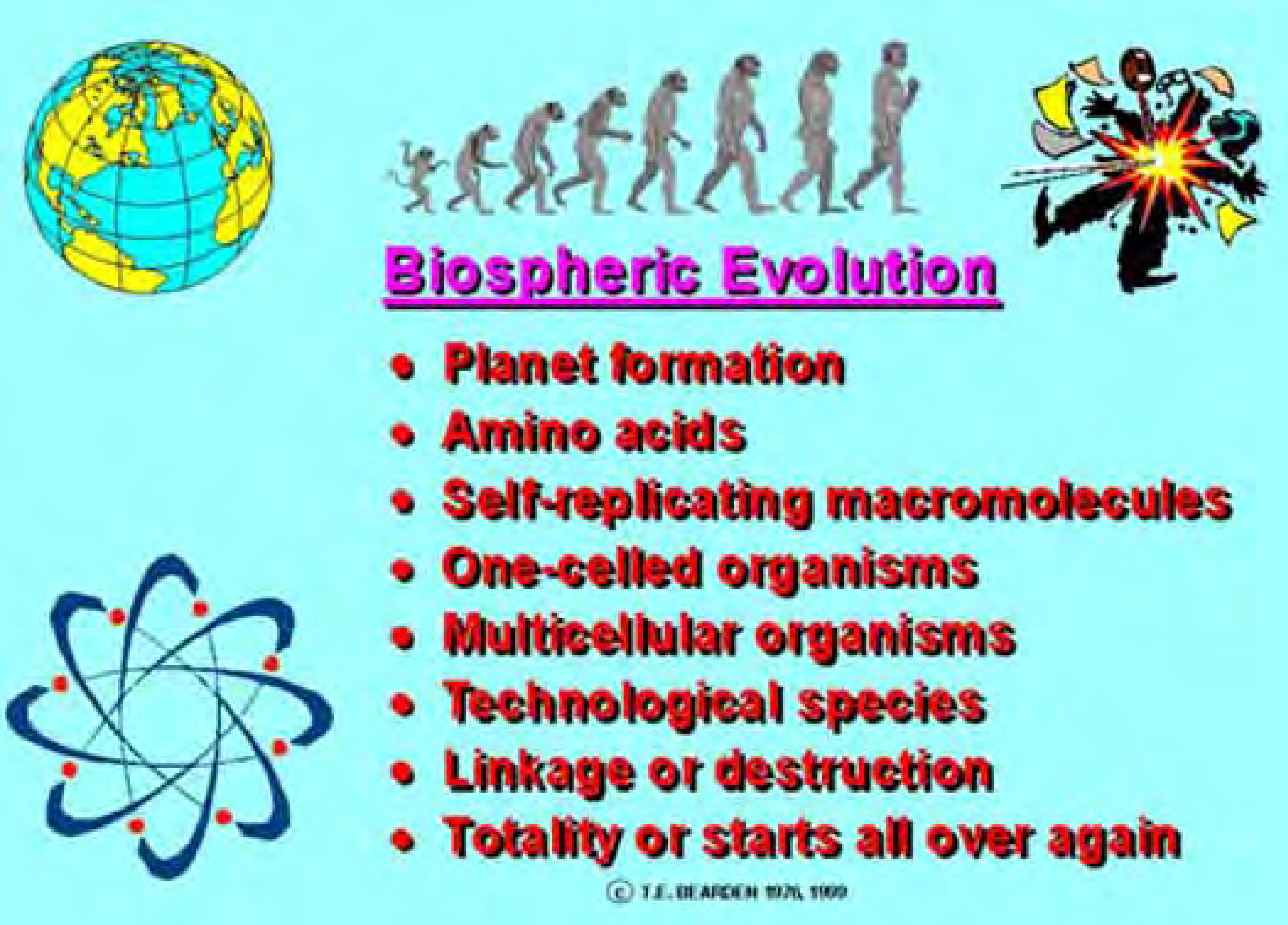

Biospheric Evolution

- Planet formation
- Amino acids
- Self-replicating macromolecules
- One-celled organisms
- Multicellular organisms
- Technological species
- Linkage or destruction
- Totality or starts all over again

© T.E. BEARDEN 1976, 1999

Return to Slide Index

Next Slide

Previous Slide

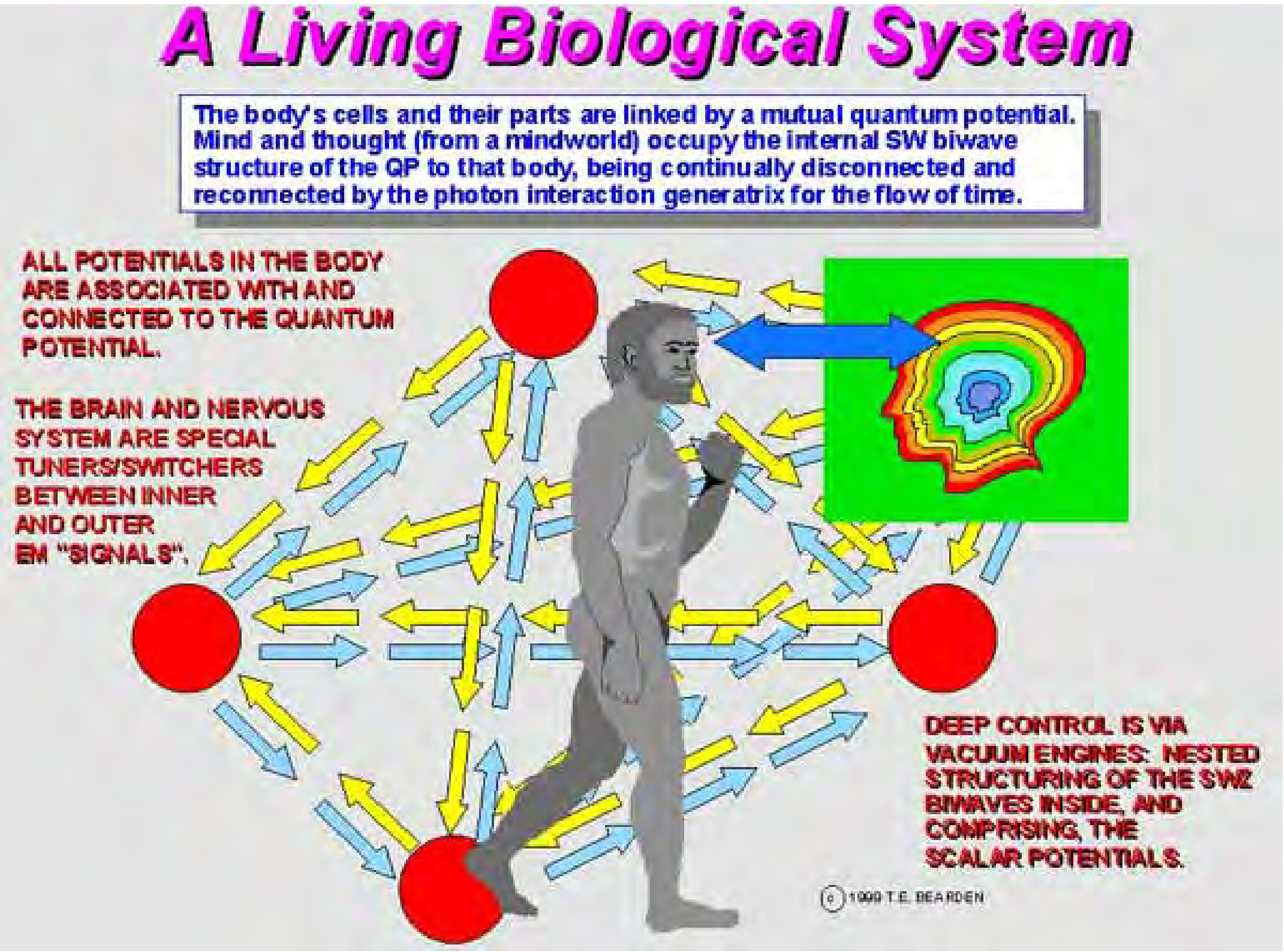

A Living Biological System

The body's cells and their parts are linked by a mutual quantum potential. Mind and thought (from a mindworld) occupy the internal SW biwave structure of the QP to that body, being continually disconnected and reconnected by the photon interaction generatrix for the flow of time.

ALL POTENTIALS IN THE BODY ARE ASSOCIATED WITH AND CONNECTED TO THE QUANTUM POTENTIAL.

THE BRAIN AND NERVOUS SYSTEM ARE SPECIAL TUNERS/SWITCHERS BETWEEN INNER AND OUTER EM "SIGNALS".

DEEP CONTROL IS VIA VACUUM ENGINES: NESTED STRUCTURING OF THE SWZ BIWAVES INSIDE, AND COMPRISING, THE SCALAR POTENTIALS.

© 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

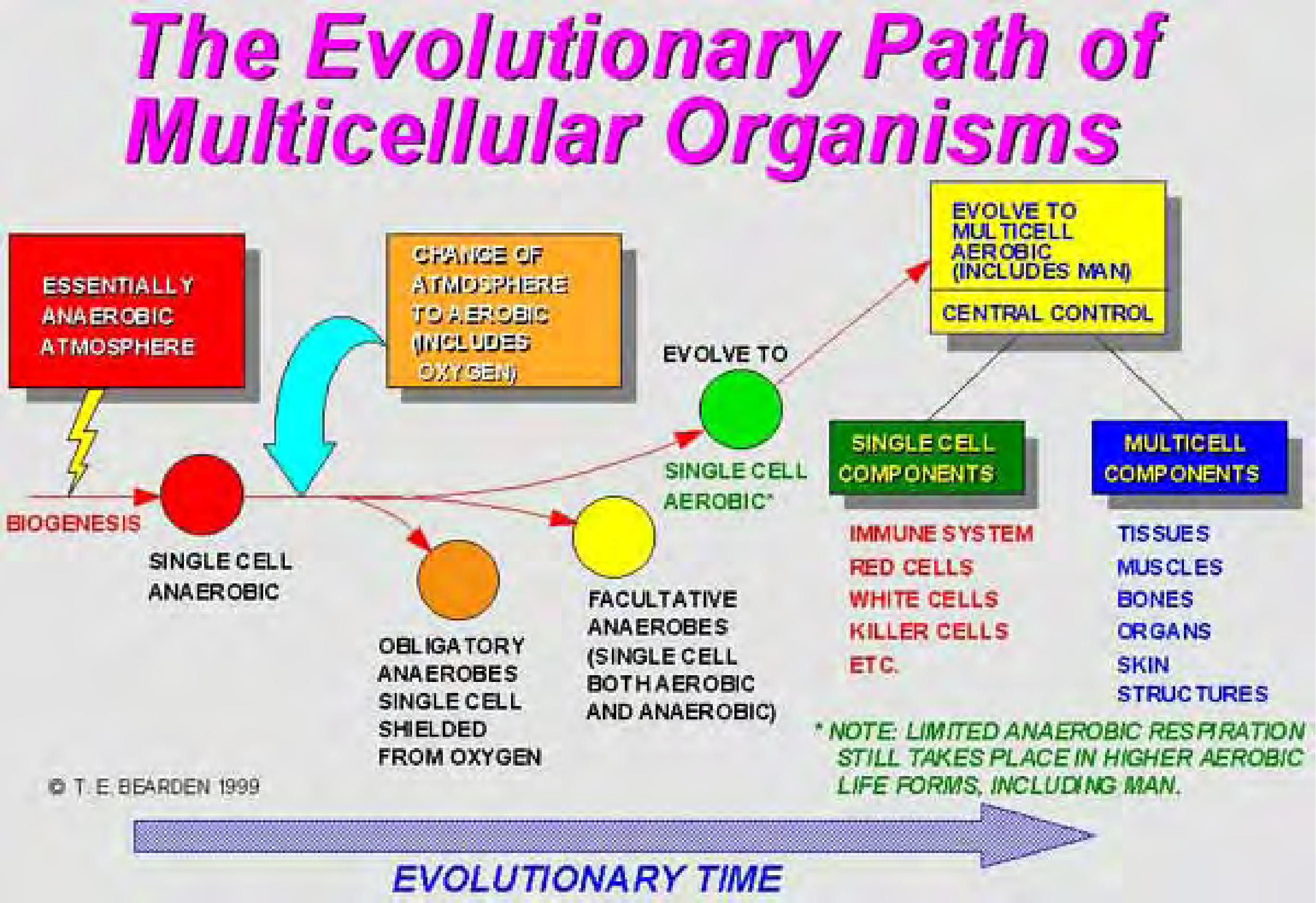

The Evolutionary Path of Multicellular Organisms

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

EVOLVE TO MULTICELL AEROBIC (INCLUDES MAN)

CENTRAL CONTROL

BIOGENESIS

SINGLE CELL ANAEROBIC

EVOLVE TO

SINGLE CELL AEROBIC*

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

SINGLE CELL COMPONENTS	MULTICELL COMPONENTS
IMMUNE SYSTEM	TISSUES
RED CELLS	MUSCLES
WHITE CELLS	BONES
KILLER CELLS	ORGANS
ETC.	SKIN
	STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

© T. E. BEARDEN 1999

EVOLUTIONARY TIME

Return to Slide Index

Next Slide

Previous Slide

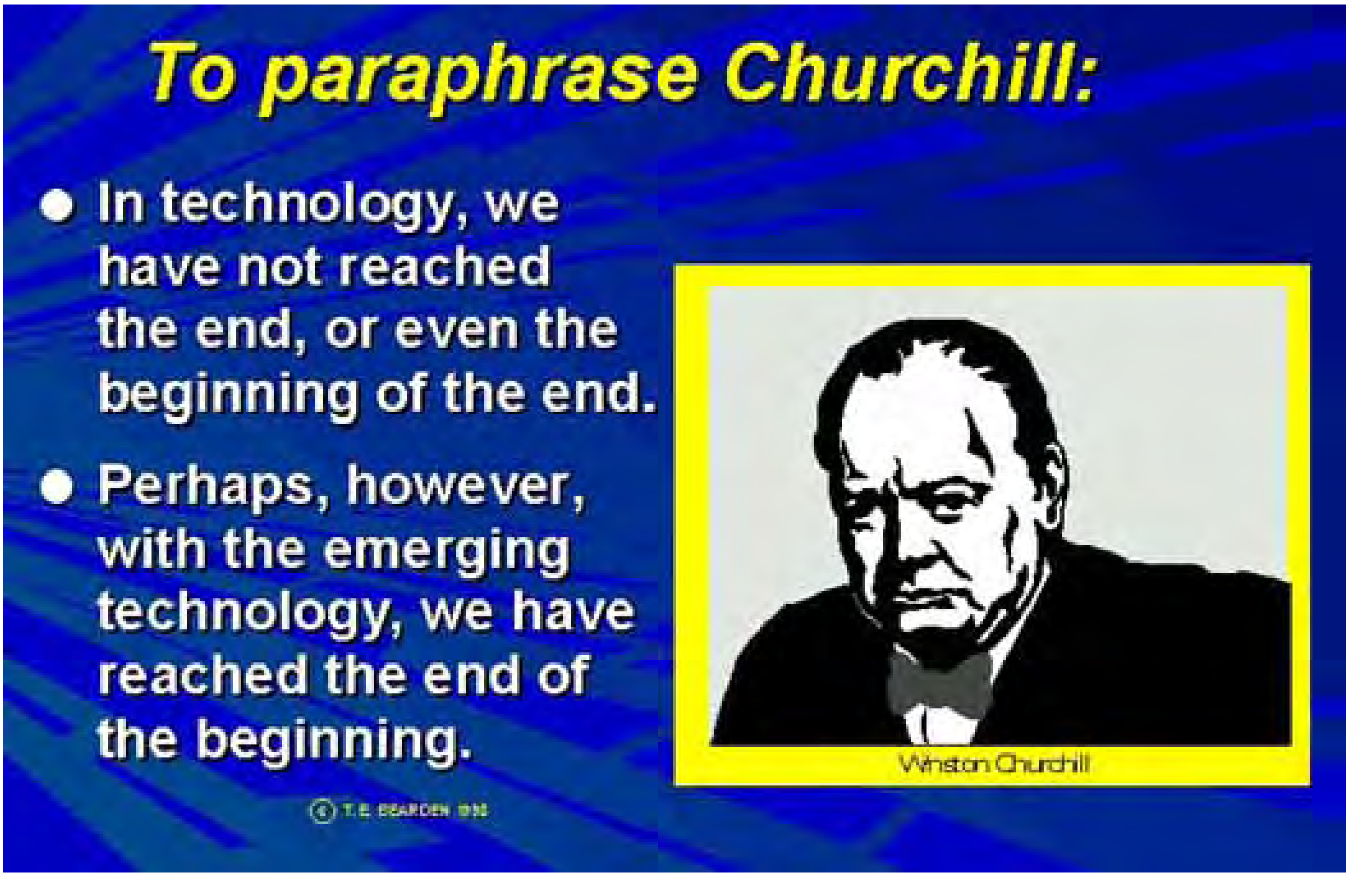

To paraphrase Churchill:

- In technology, we have not reached the end, or even the beginning of the end.
- Perhaps, however, with the emerging technology, we have reached the end of the beginning.

Winston Churchill

© T.E. BEARDEN [illegible]

Return to Slide Index

Next Slide

Previous Slide

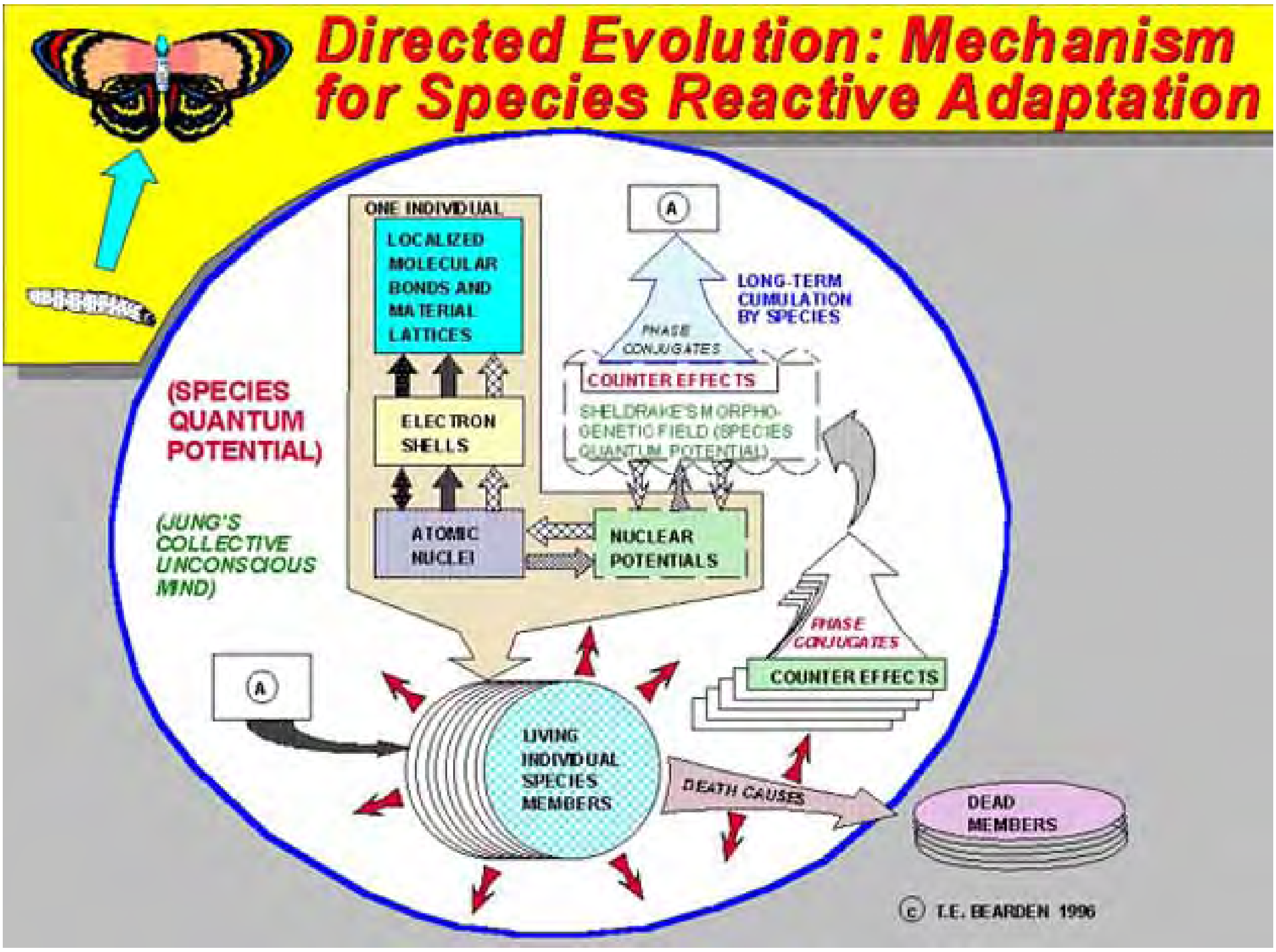

Directed Evolution: Mechanism for Species Reactive Adaptation

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

LONG-TERM CUMULATION BY SPECIES

PHASE CONJUGATES

COUNTER EFFECTS

SHELDRAKE'S MORPHO-GENETIC FIELD (SPECIES QUANTUM POTENTIAL)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

PHASE CONJUGATES

COUNTER EFFECTS

A

LIVING INDIVIDUAL SPECIES MEMBERS

DEATH CAUSES

DEAD MEMBERS

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

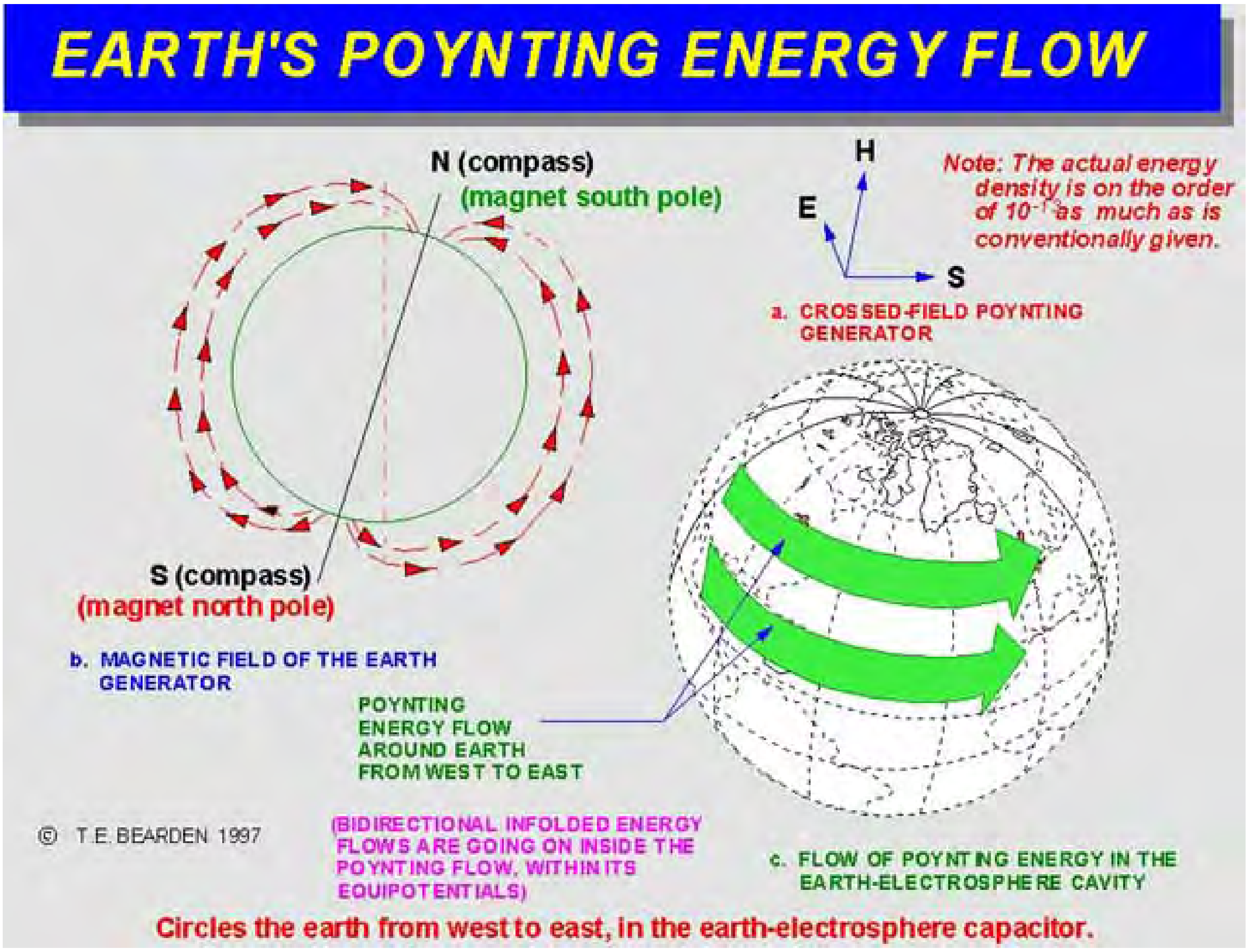

EARTH'S POYNTING ENERGY FLOW

Note: The actual energy density is on the order of 10^{-1} as much as is conventionally given.

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING ENERGY FLOW AROUND EARTH FROM WEST TO EAST

© T.E. BEARDEN 1997

(BIDIRECTIONAL INFOLDED ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

Circles the earth from west to east, in the earth-electrosphere capacitor.

Return to Slide Index

Next Slide

Previous Slide

Note: Should be

10^{13}

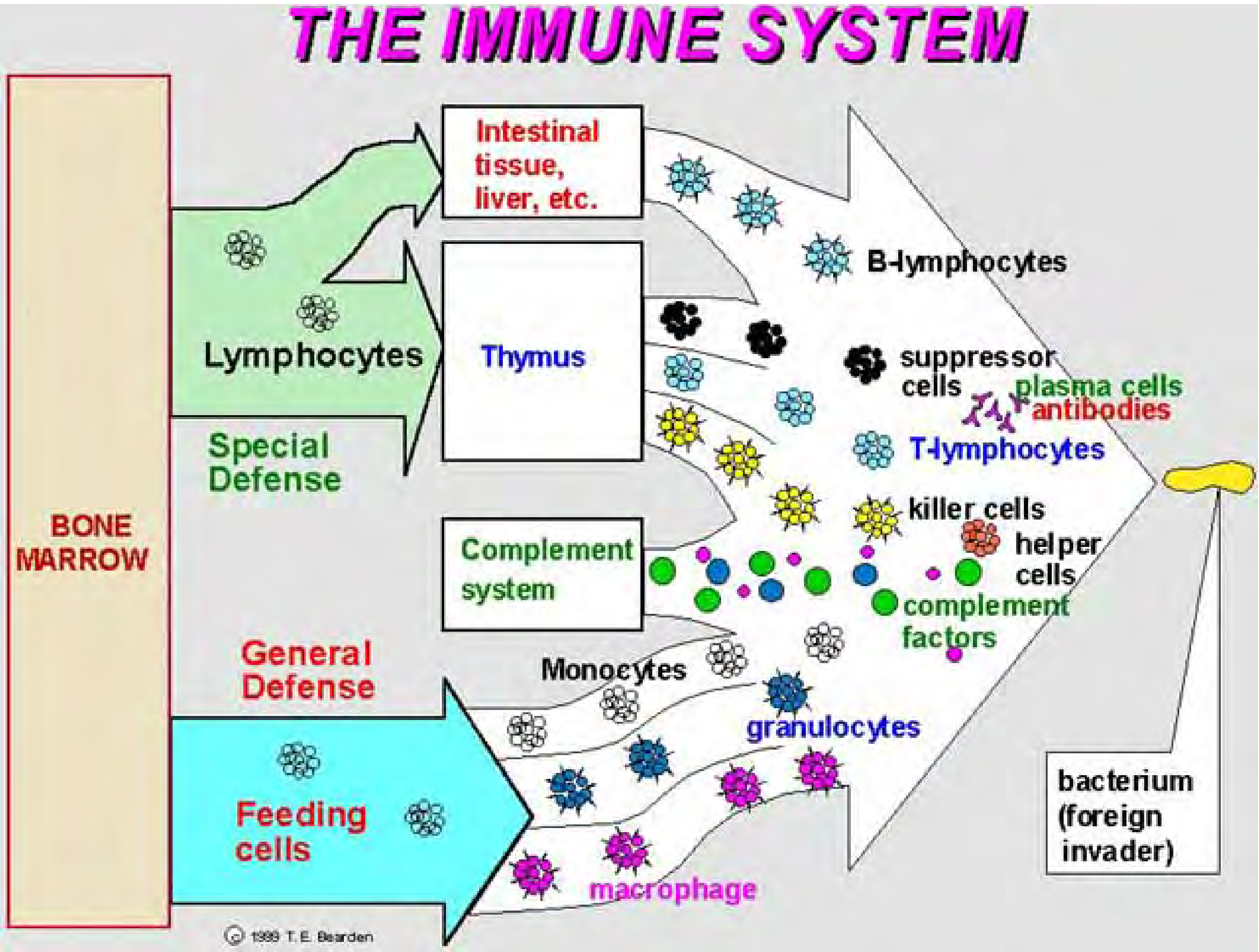
THE IMMUNE SYSTEM
BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
Complement system
General Defense
Feeding cells
Monocytes
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
granulocytes
macrophage
bacterium (foreign invader)
© 1999 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

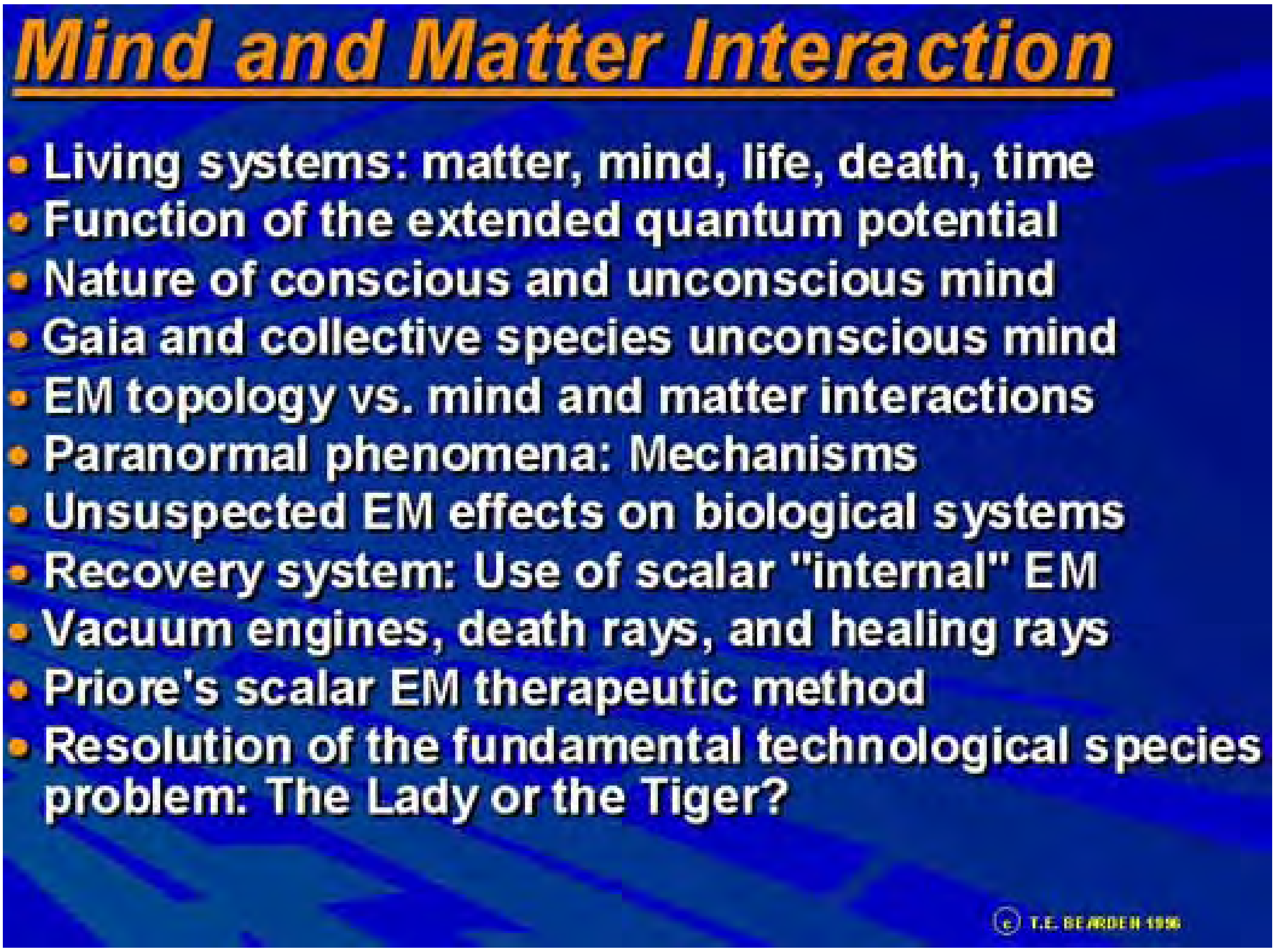

Mind and Matter Interaction

- Living systems: matter, mind, life, death, time
- Function of the extended quantum potential
- Nature of conscious and unconscious mind
- Gaia and collective species unconscious mind
- EM topology vs. mind and matter interactions
- Paranormal phenomena: Mechanisms
- Unsuspected EM effects on biological systems
- Recovery system: Use of scalar "internal" EM
- Vacuum engines, death rays, and healing rays
- Priore's scalar EM therapeutic method
- Resolution of the fundamental technological species problem: The Lady or the Tiger?

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

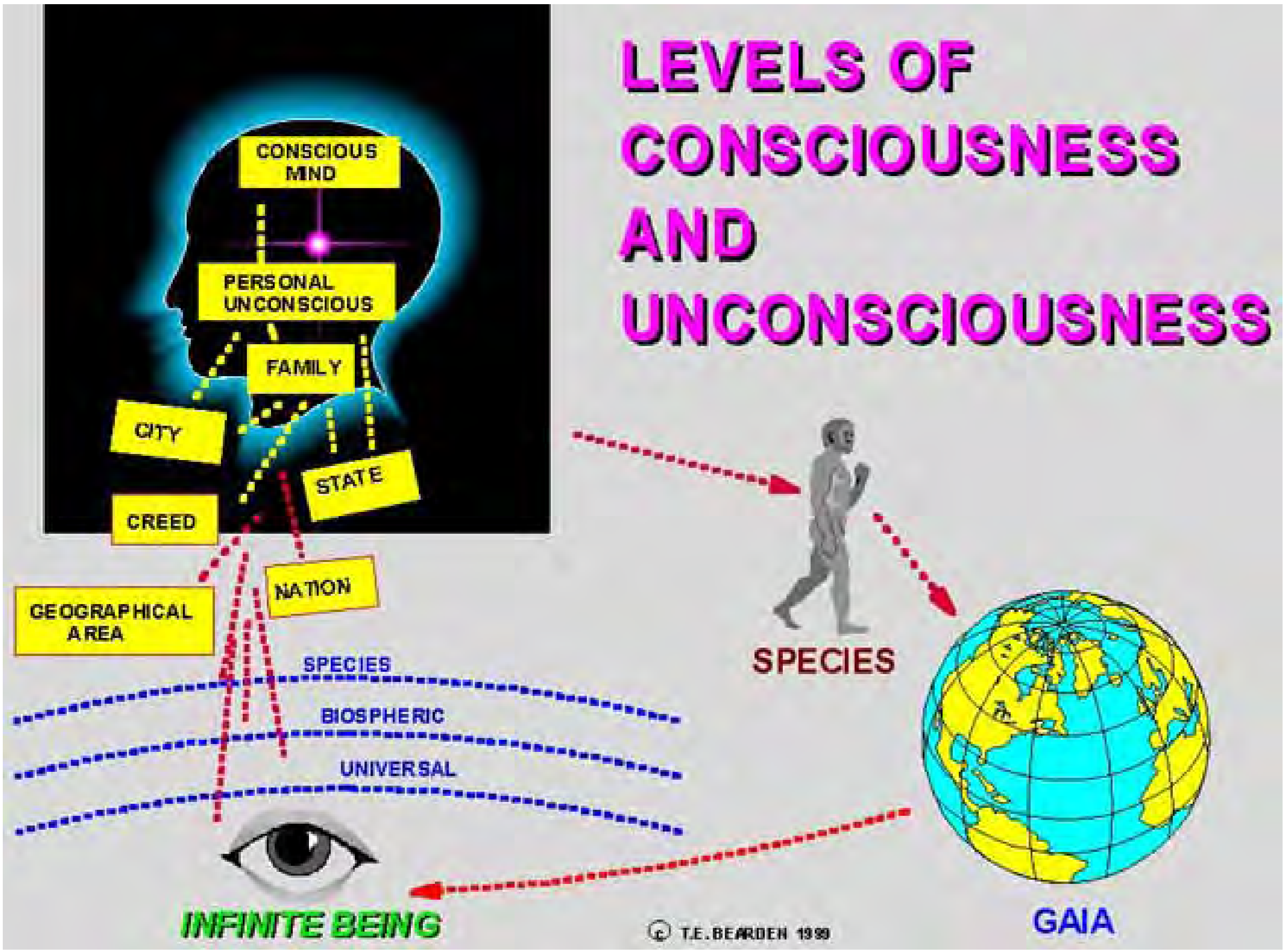

LEVELS OF CONSCIOUSNESS AND UNCONSCIOUSNESS

CONSCIOUS MIND
PERSONAL UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL
INFINITE BEING
SPECIES
GAIA

© T.E.BEARDEN 1999

Return to Slide Index

Next Slide

Previous Slide

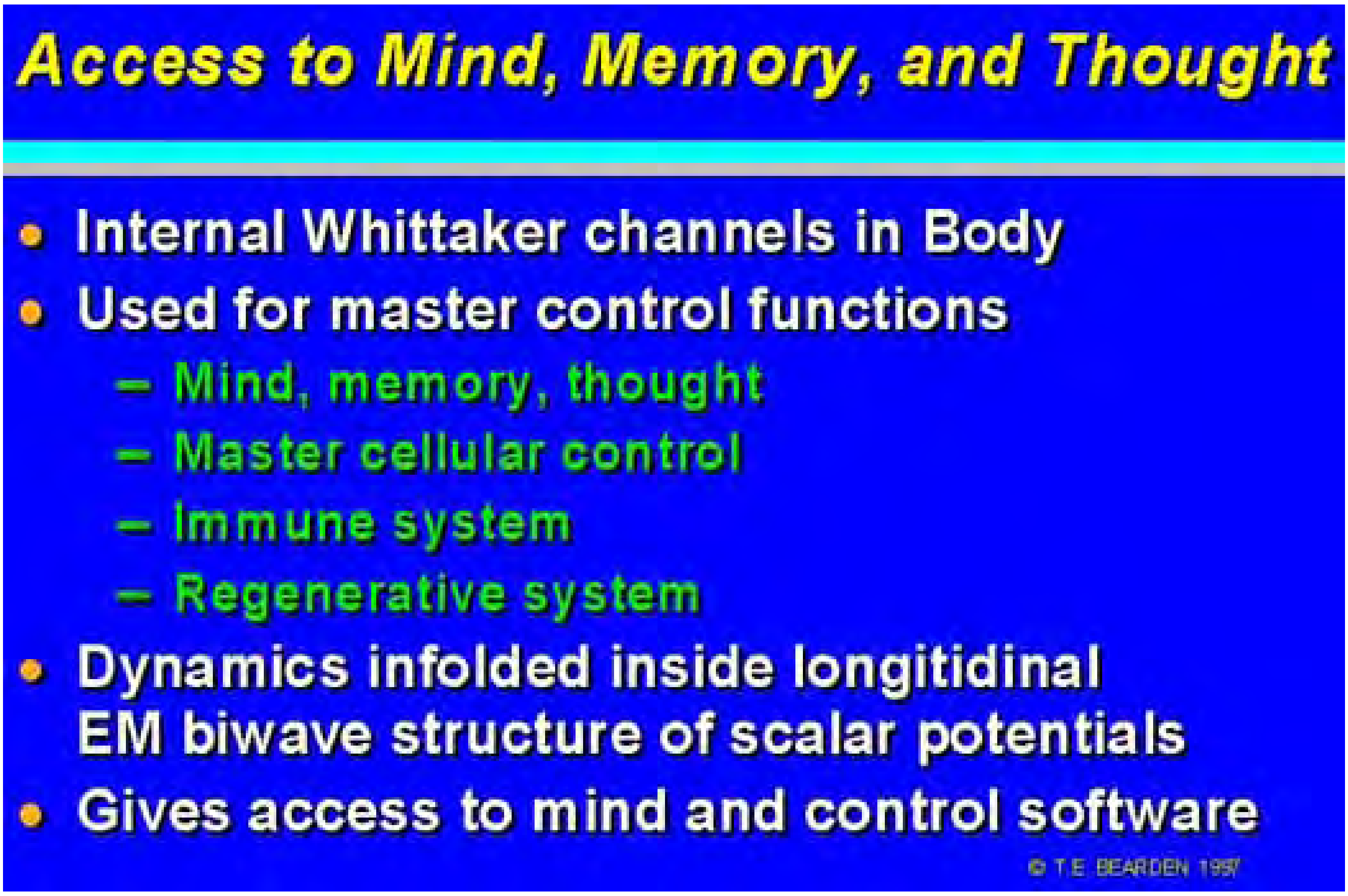

Access to Mind, Memory, and Thought

- Internal Whittaker channels in Body
- Used for master control functions
  - Mind, memory, thought
  - Master cellular control
  - Immune system
  - Regenerative system
- Dynamics infolded inside longitidinal EM biwave structure of scalar potentials
- Gives access to mind and control software

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

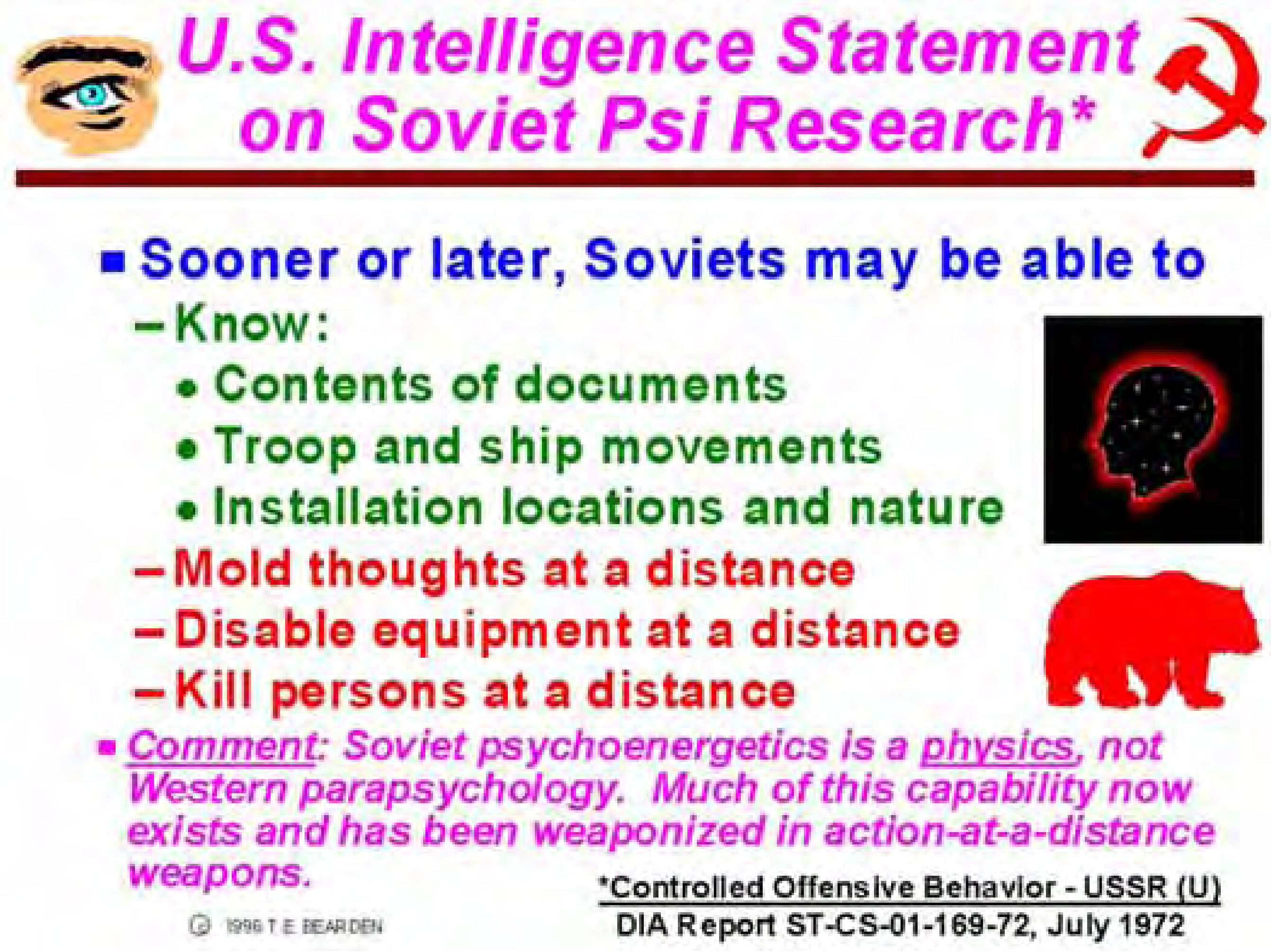

U.S. Intelligence Statement on Soviet Psi Research*

- Sooner or later, Soviets may be able to
  - Know:
    - Contents of documents
    - Troop and ship movements
    - Installation locations and nature
  - Mold thoughts at a distance
  - Disable equipment at a distance
  - Kill persons at a distance
- *Comment: Soviet psychoenergetics is a physics, not Western parapsychology. Much of this capability now exists and has been weaponized in action-at-a-distance weapons.*

*Controlled Offensive Behavior - USSR (U)
DIA Report ST-CS-01-169-72, July 1972

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

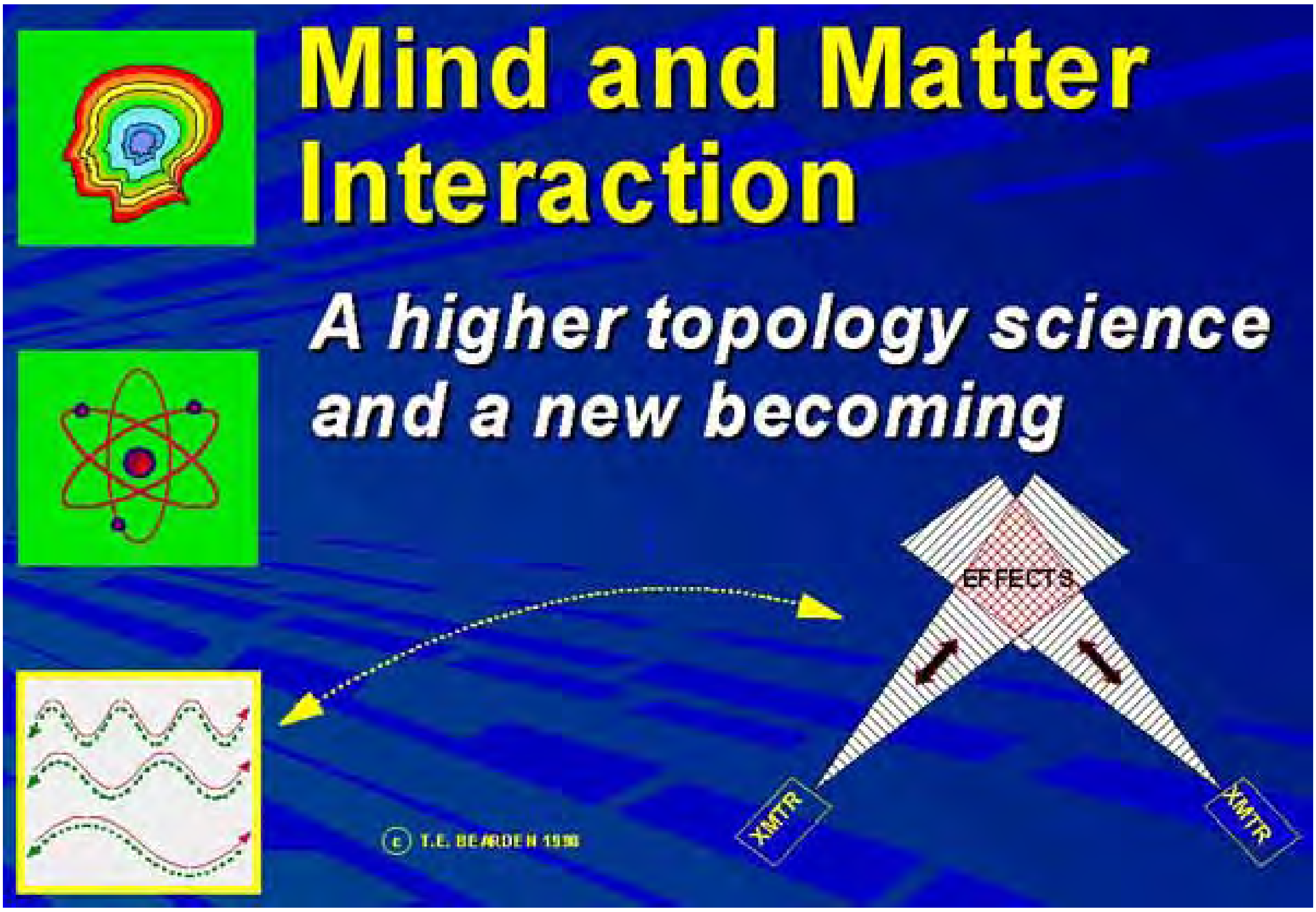

Mind and Matter Interaction

A higher topology science and a new becoming

EFFECTS
XMTR
XMTR
© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

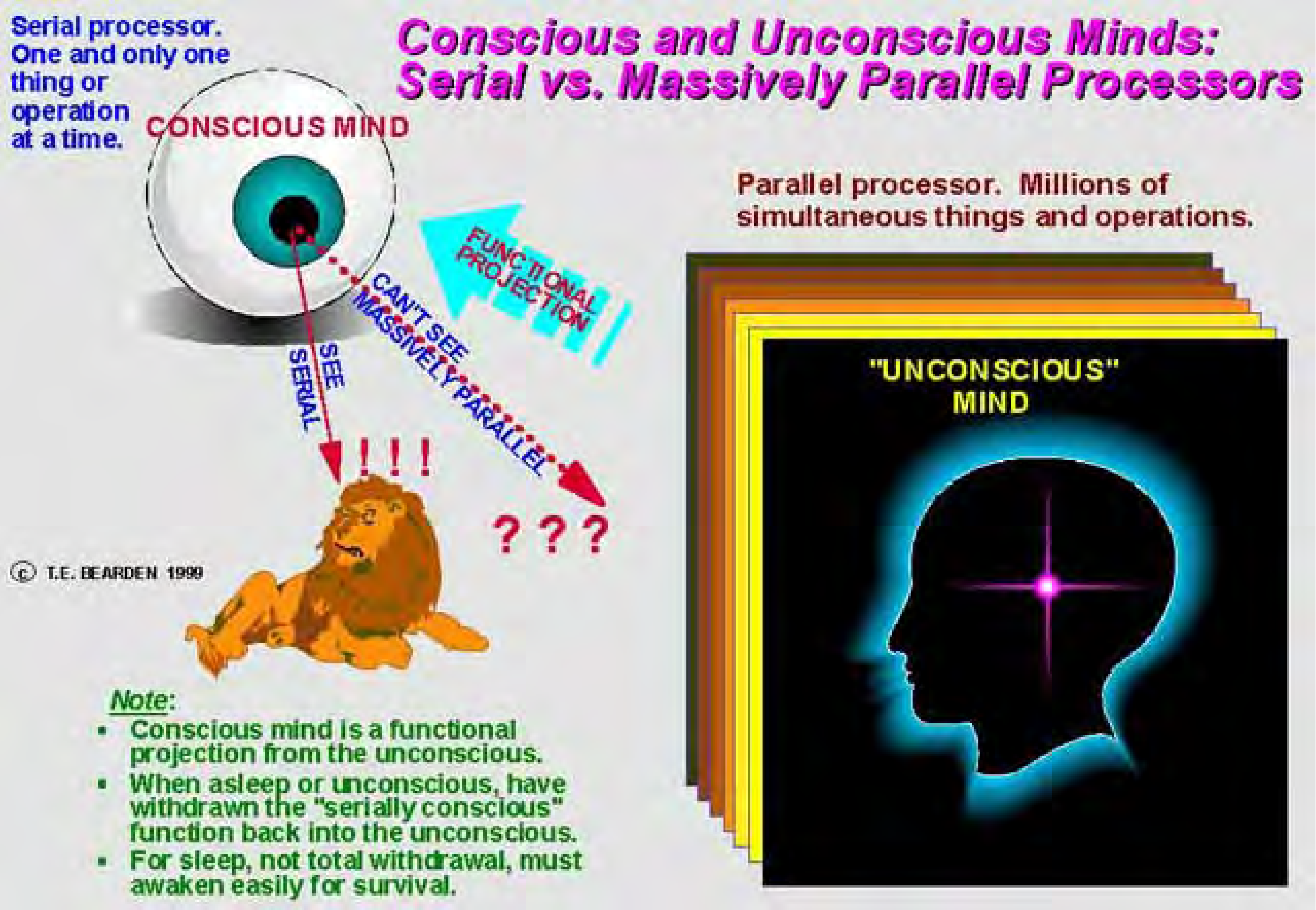

Conscious and Unconscious Minds: Serial vs. Massively Parallel Processors

Serial processor. One and only one thing or operation at a time.

CONSCIOUS MIND
SEE SERIAL
!!!
CAN'T SEE MASSIVELY PARALLEL
???
FUNCTIONAL PROJECTION
"UNCONSCIOUS" MIND

Parallel processor. Millions of simultaneous things and operations.

© T.E. BEARDEN 1999

Note:
- Conscious mind is a functional projection from the unconscious.
- When asleep or unconscious, have withdrawn the "serially conscious" function back into the unconscious.
- For sleep, not total withdrawal, must awaken easily for survival.

Return to Slide Index

Next Slide

Previous Slide

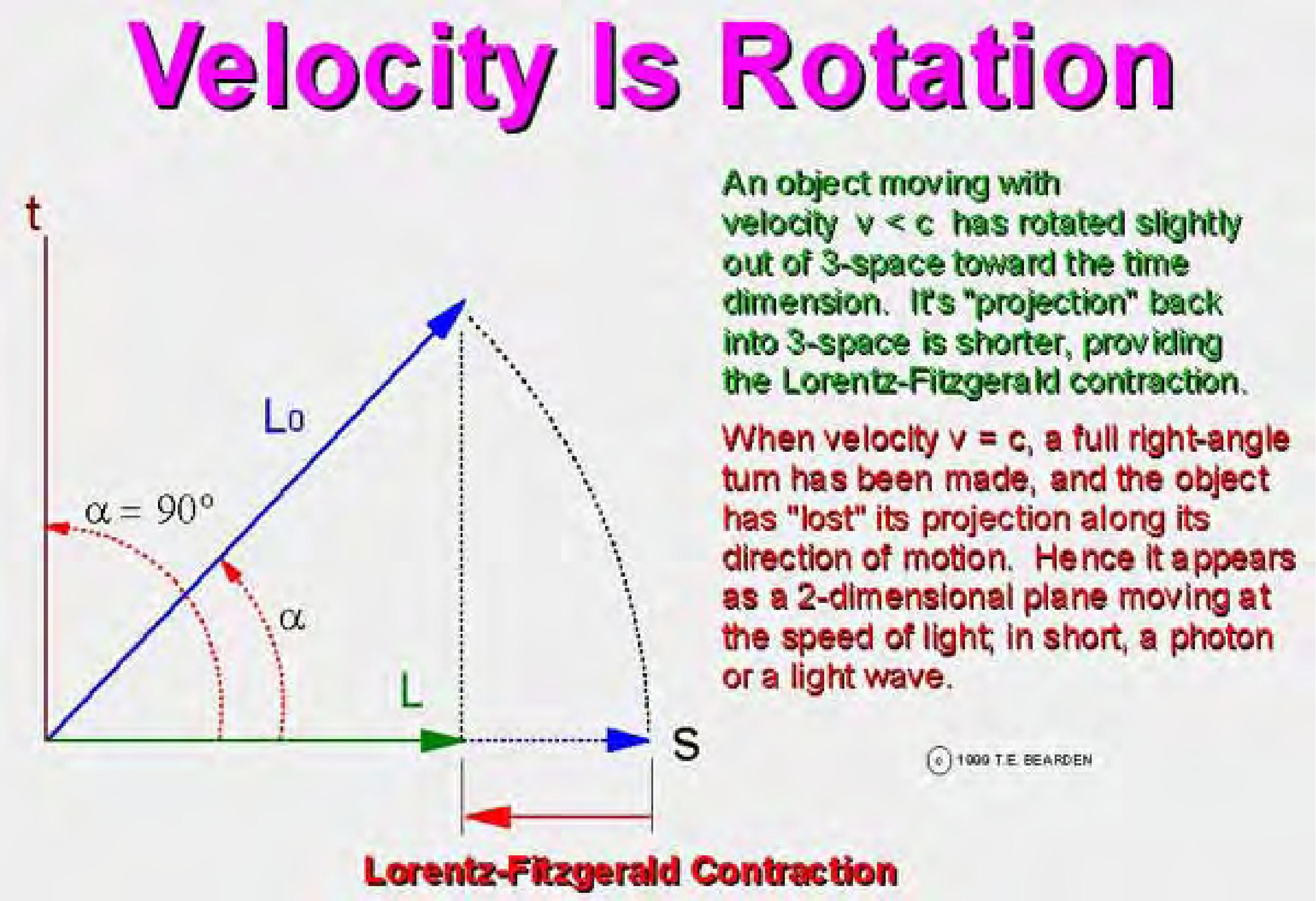

Velocity Is Rotation

An object moving with velocity v < c has rotated slightly out of 3-space toward the time dimension. It's "projection" back into 3-space is shorter, providing the Lorentz-Fitzgerald contraction.

When velocity v = c, a full right-angle turn has been made, and the object has "lost" its projection along its direction of motion. Hence it appears as a 2-dimensional plane moving at the speed of light; in short, a photon or a light wave.

Lorentz-Fitzgerald Contraction

© 1999 T.E. BEARDEN

Next Slide

Return to Slide Index

Previous Slide

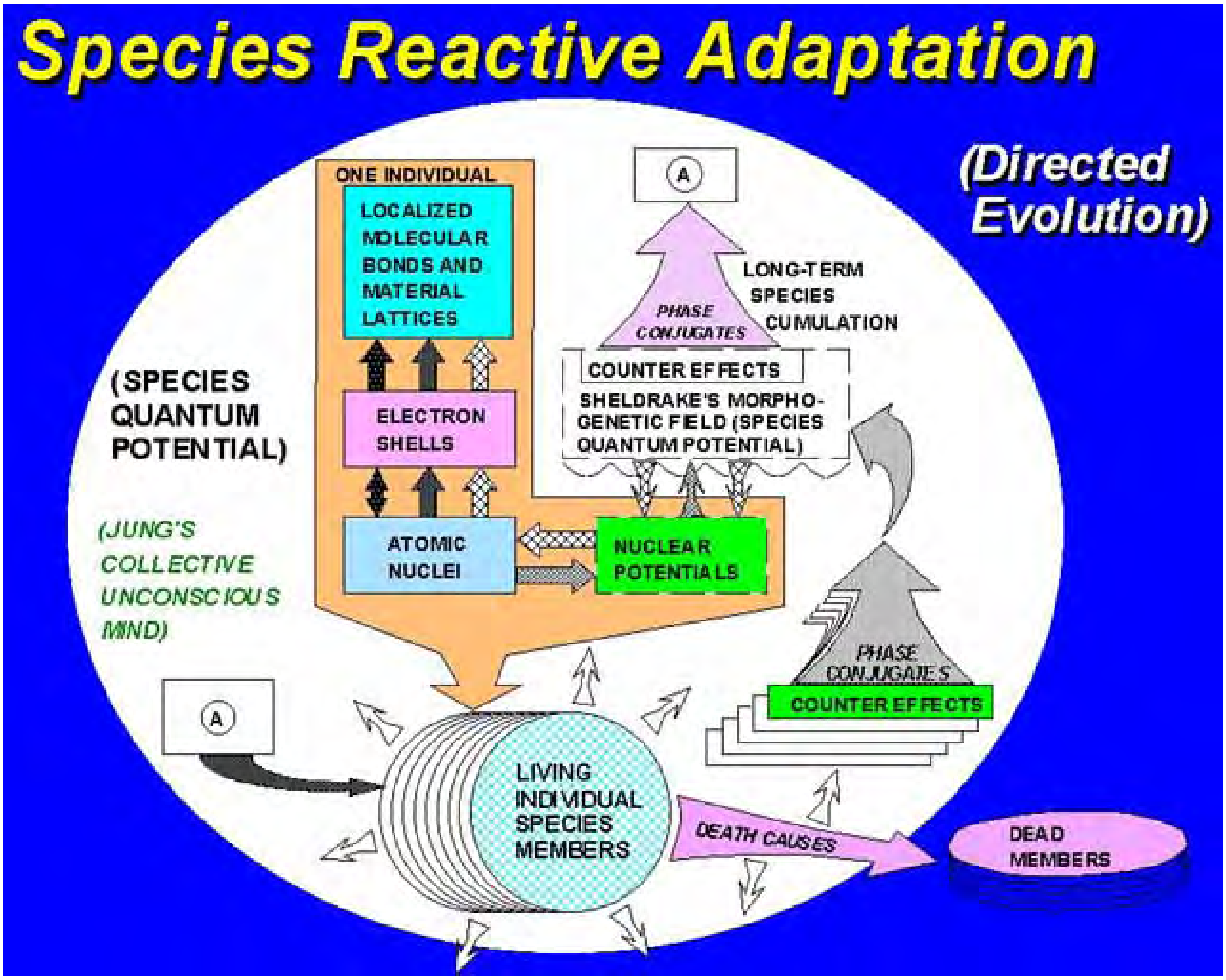

Species Reactive Adaptation

(Directed Evolution)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

PHASE CONJUGATES

LONG-TERM SPECIES CUMULATION

COUNTER EFFECTS

SHELDRAKE'S MORPHOGENETIC FIELD (SPECIES QUANTUM POTENTIAL)

PHASE CONJUGATES

COUNTER EFFECTS

A

LIVING INDIVIDUAL SPECIES MEMBERS

DEATH CAUSES

DEAD MEMBERS

Return to Slide Index

Next Slide

Previous Slide

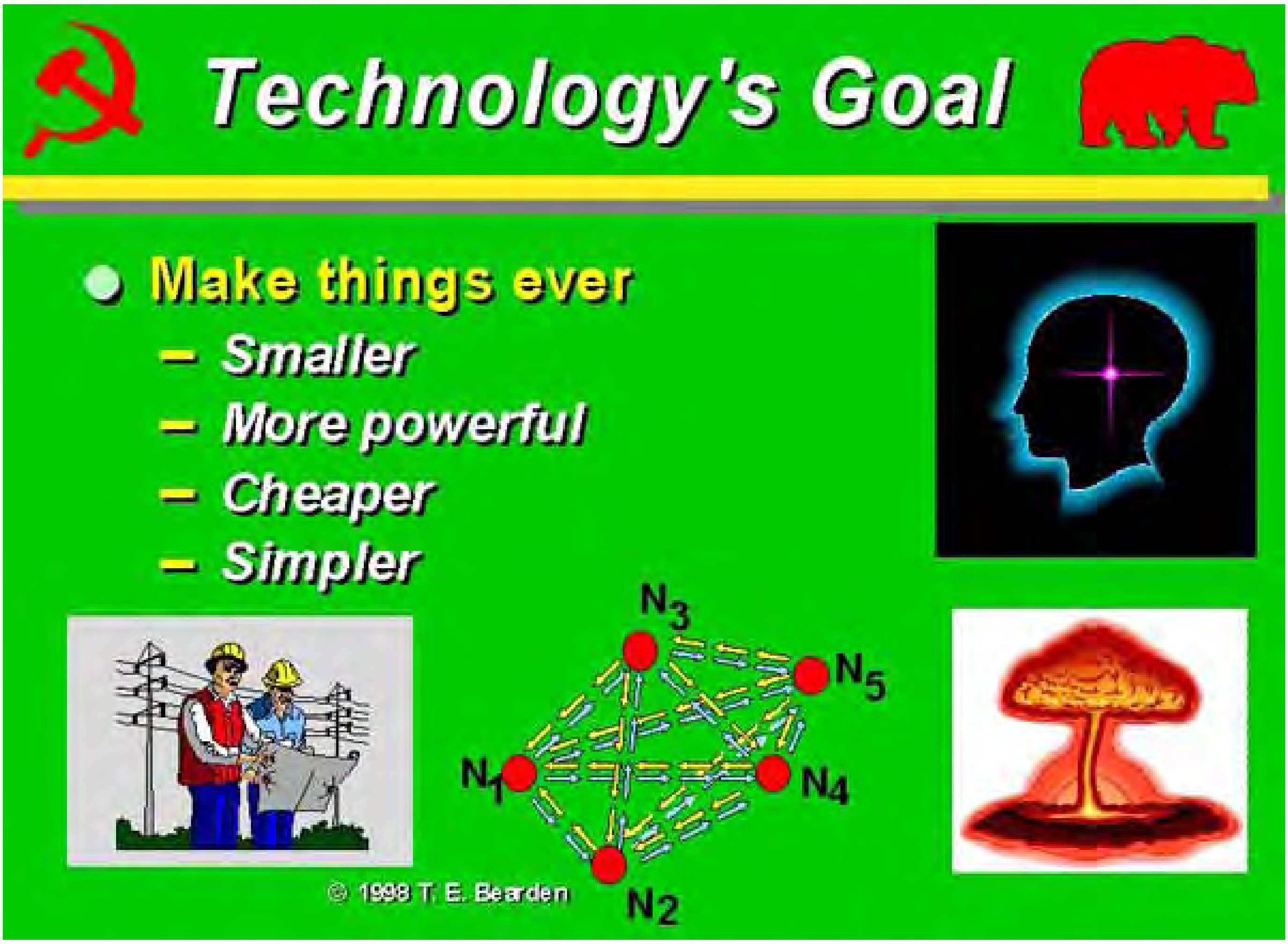

Technology's Goal

- Make things ever
  - Smaller
  - More powerful
  - Cheaper
  - Simpler

N3
N5
N1
N4
N2

© 1998 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

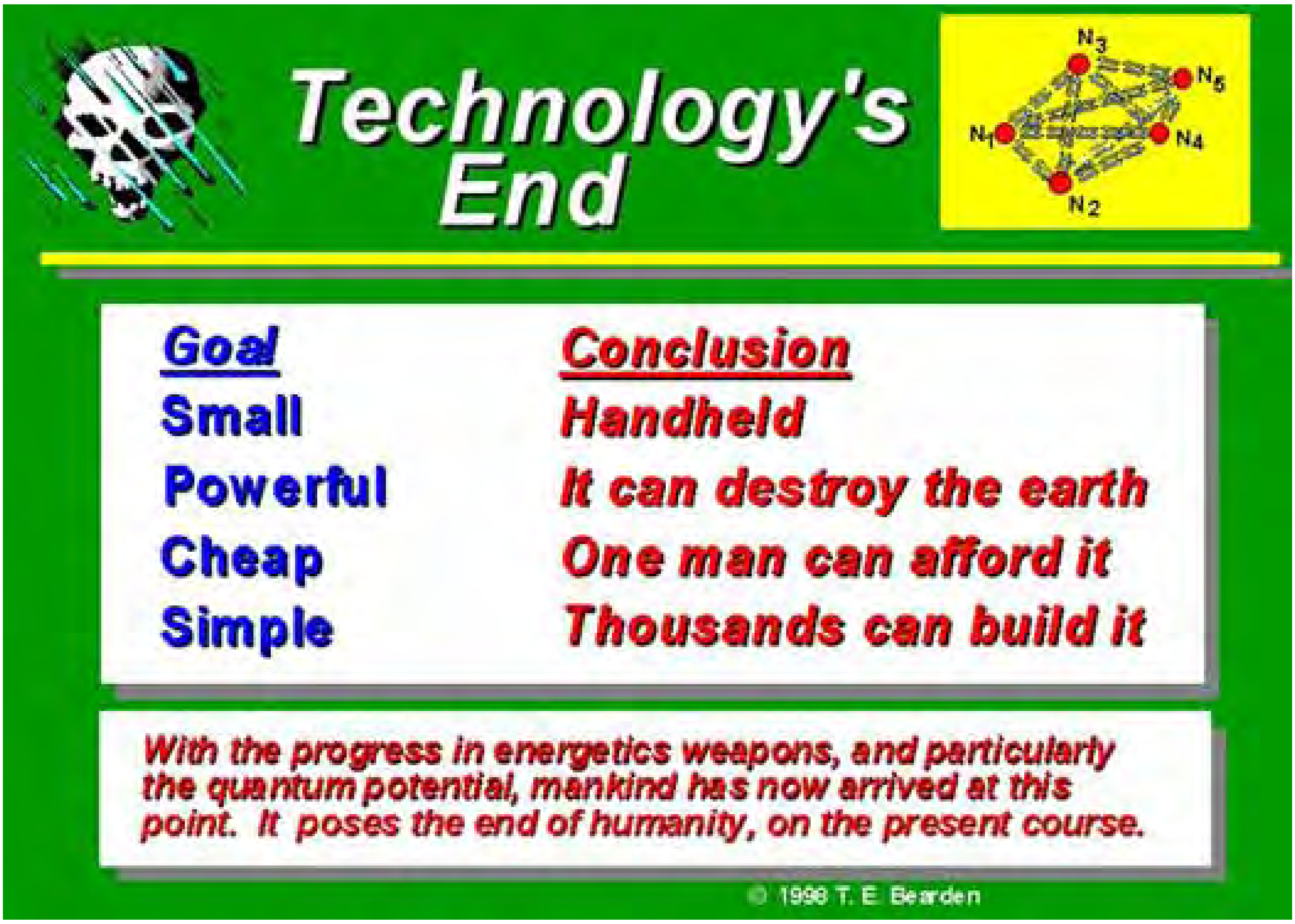
N3
N5
N1
N4
N2

Technology's End

Goal	Conclusion
Small	Handheld
Powerful	It can destroy the earth
Cheap	One man can afford it
Simple	Thousands can build it

With the progress in energetics weapons, and particularly the quantum potential, mankind has now arrived at this point. It poses the end of humanity, on the present course.

© 1998 T. E. Bearden

Return to Slide Index

Next Slide

Previous Slide

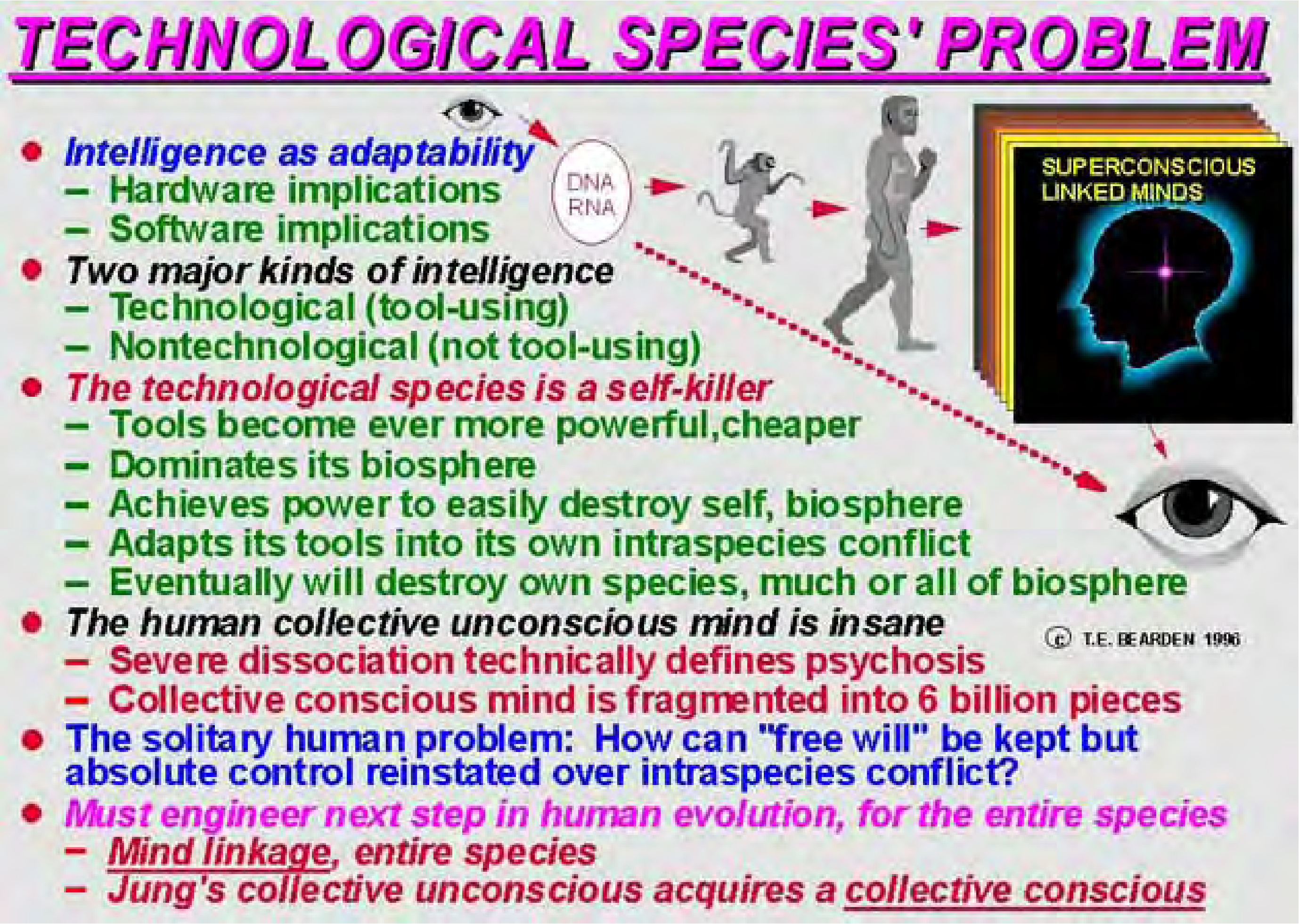
DNA
RNA
SUPERCONSCIOUS
LINKED MINDS
© T.E. BEARDEN 1996

TECHNOLOGICAL SPECIES' PROBLEM

- *Intelligence as adaptability*
  - Hardware implications
  - Software implications
- *Two major kinds of intelligence*
  - Technological (tool-using)
  - Nontechnological (not tool-using)
- *The technological species is a self-killer*
  - Tools become ever more powerful,cheaper
  - Dominates its biosphere
  - Achieves power to easily destroy self, biosphere
  - Adapts its tools into its own intraspecies conflict
  - Eventually will destroy own species, much or all of biosphere
- *The human collective unconscious mind is insane*
  - Severe dissociation technically defines psychosis
  - Collective conscious mind is fragmented into 6 billion pieces
- The solitary human problem: How can "free will" be kept but absolute control reinstated over intraspecies conflict?
- *Must engineer next step in human evolution, for the entire species*
  - *Mind linkage, entire species*
  - *Jung's collective unconscious acquires a collective conscious*

Return to Slide Index

Previous Slide

The Missing Infolded Electro-dynamics

Contents

1. The Missing Infolded Electrodynamics
2. Major Points
3. Major Points (cont.)
4. MTW's General Relativity Principle
5. Extended General Relavity Principle
6. Active Spacetime and Specific Vacuum Engine
7. Electrodynamicists React to Removal of Material Ether
8. Reaction to Loss of Material Ether
9. Massless Spacetime Perturbations Not Addressed
10. Bunge on Status of Electrodynamics
11. Bunge (continued)
12. Poynting Energy Flow Along Two Parallel Conductors
13. Fields and potentials defined only in/on charged matter
14. Wave in String (Transverse)
15. Launching EM Wave from Wire Antenna
16. Phase Conjugate Wavepairs Produce New Waves
17. Time-Polarized (Scalar) EM Wave
18. Mechanism for flow of time
19. Present systems coefficient of performance (COP)<1.0
20. Bridging for Overunity
21. Free Energy Generator
22. Demonstration of Overunity Concept
23. Altered Spacetime Curvatures
24. Interaction of Mass and Spacetime Curvature
25. Composition of Scalar Potential
26. Lorenz Closed Surface Integration
27. How can they use that old physically insignificant wind?!!
28. Vacuum Engines: Spacetime Demons
29. Undistorted Progressive Waves (1)
30. Undistorted Progressive Waves (2)
31. Other Superluminal Waves
32. Becker's Experiments: Healing Bone Fractures
33. Producing a Specific Spacetime Curvature Engine
34. Pumping A Phase Conjugate with Transverse EM

35. Differentiation and Dedifferentiation
36. First Step to Cancer: Hypoxia
37. Longitudinal Photon Interaction with Charge
38. General Relativity in Particle Physics
39. Pepper* on time-reversing mass
40. Pepper* on time-reversing mass (2)
41. Pepper* on time-reversing mass (3)
42. Chung's Negative Resistor
43. China Lake Anomalies
44. Time Reversal (TR) Zone
45. Electronuclear Reactions in TR-Zones (1)
46. Electronuclear Reactions in TR-Zones (2)
47. Electronuclear Reactions in TR-Zones (3)
48. Nuclear Reactions in Time-Reversal Zones
49. Nuclear reactions are biased
50. Dynamics may be reversed
51. Significant only after time-density charging
52. Curved Spacetime as source or sink
53. Anomalous China Lake Phenomena (1)
54. Anomalous China Lake 2
55. Anomalous China Lake 3
56. Anomalous China Lake 4
57. Consistent Anomalies 1
58. Consistent Anomalies 2
59. LW Emission in Electrolysis
60. New Energy Conservation Law
61. Carcinogens & EM Radiation
62. Carcinogens & EM Radiation 1
63. Chronic Disease Puzzle
64. Cancer characteristics
65. Solving Cancer Enigma
66. Cancer Characteristics
67. Portable Prioré Unit
68. Lisitsyn: Brain Code Broken
69. Anti-engine reverses cell
70. Prioré Methodology
71. Prioré Methodology (2)
72. Pumping cell with LW
73. Extension to Nonlinear Optics
74. Time Reversal of Mass to a Form Not in Its Past
75. Cellular R&R System

76. Portable Prioré Treatment System
77. Portable System Block Diagram
78. Absorption and Emission
79. Absorption and Emission (2)
80. EM Waves in Spacetime
81. Cellular Evolution
82. Becker's Regeneration System
83. Becker's Control System Responds to Injury
84. Casualties Surviving Conventional Treatment
85. "Time as fuel" is the future

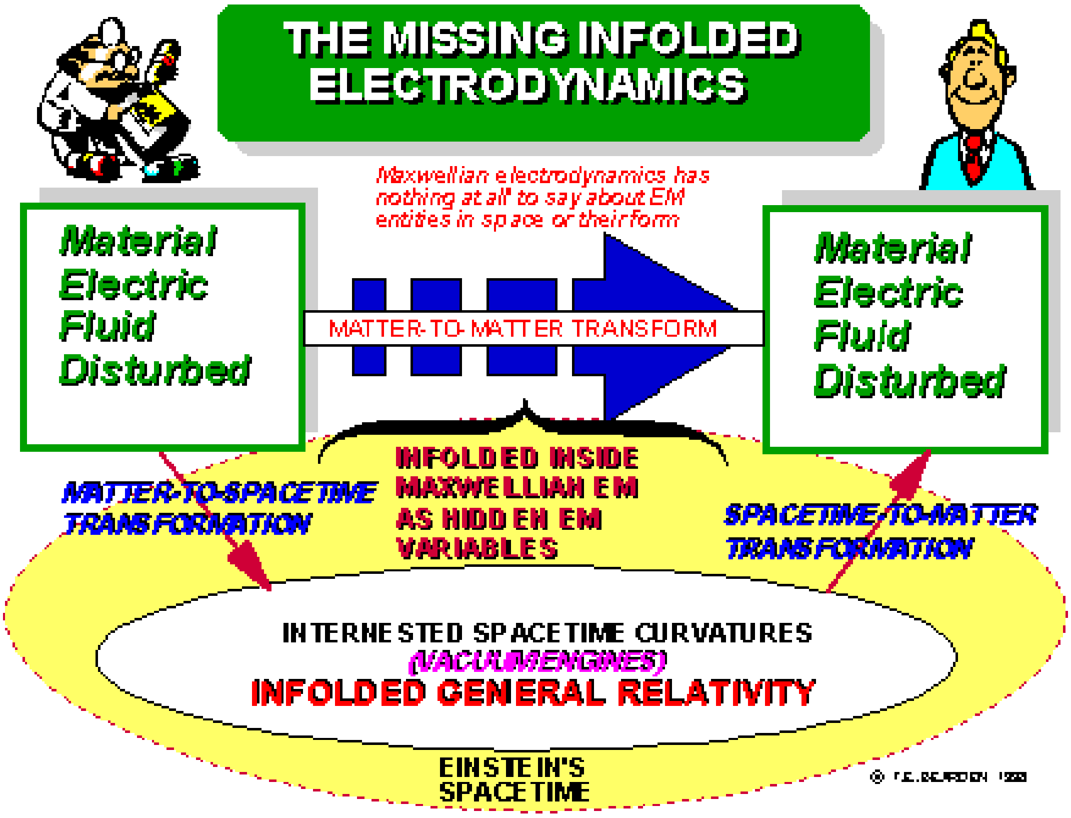
THE MISSING INFOLDED ELECTRODYNAMICS
Material Electric Fluid Disturbed
Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form
MATTER-TO-MATTER TRANSFORM
Material Electric Fluid Disturbed
MATTER-TO-SPACETIME TRANSFORMATION
INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES
SPACETIME-TO-MATTER TRANSFORMATION
INTERNESTED SPACETIME CURVATURES
(VACUUM ENGINES)
INFOLDED GENERAL RELATIVITY
EINSTEIN'S SPACETIME
© T.E. BEARDEN

Published by Lotus® Freelance Graphics®

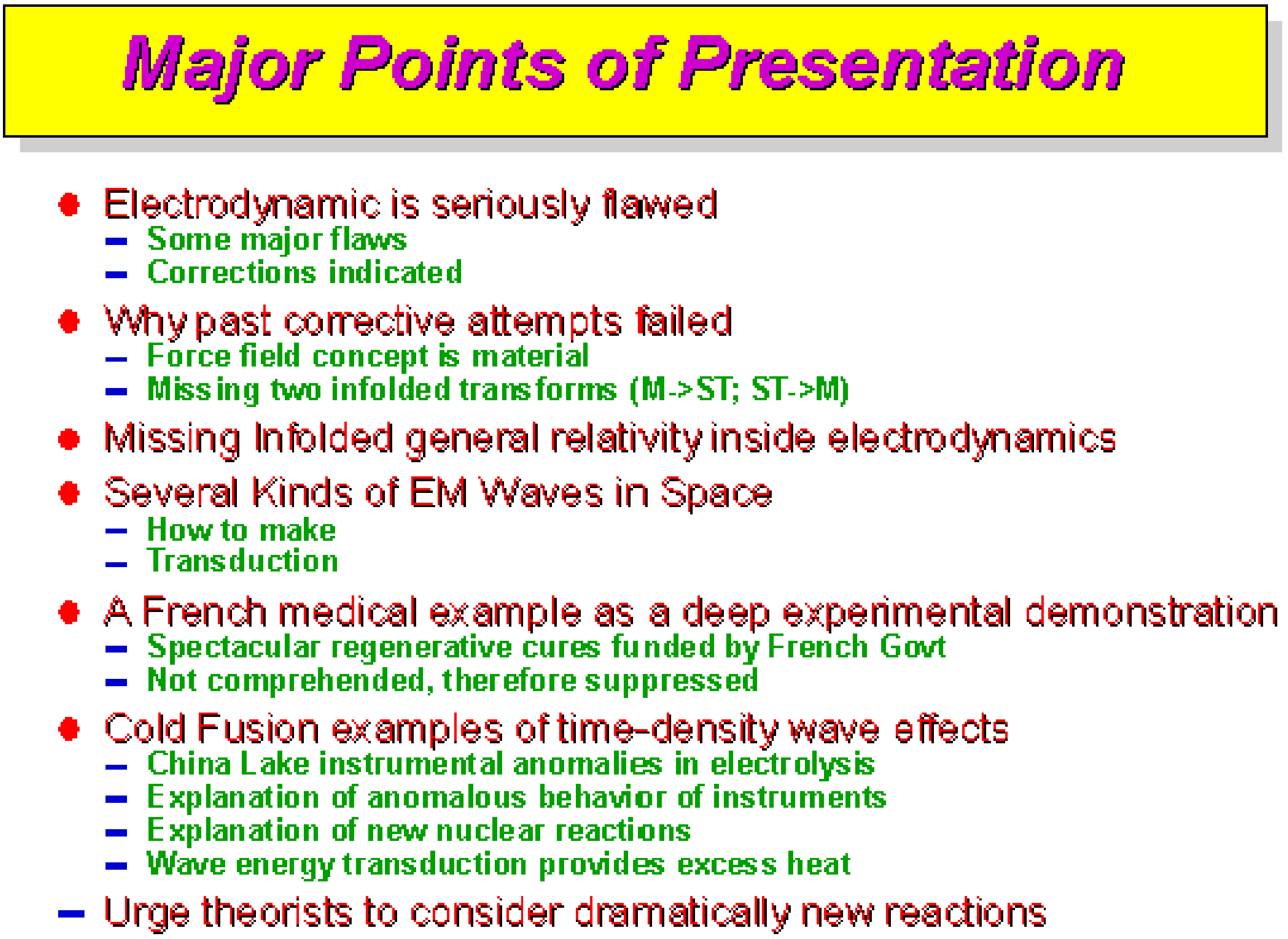

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing Infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

Published by Lotus® Freelance Graphics®

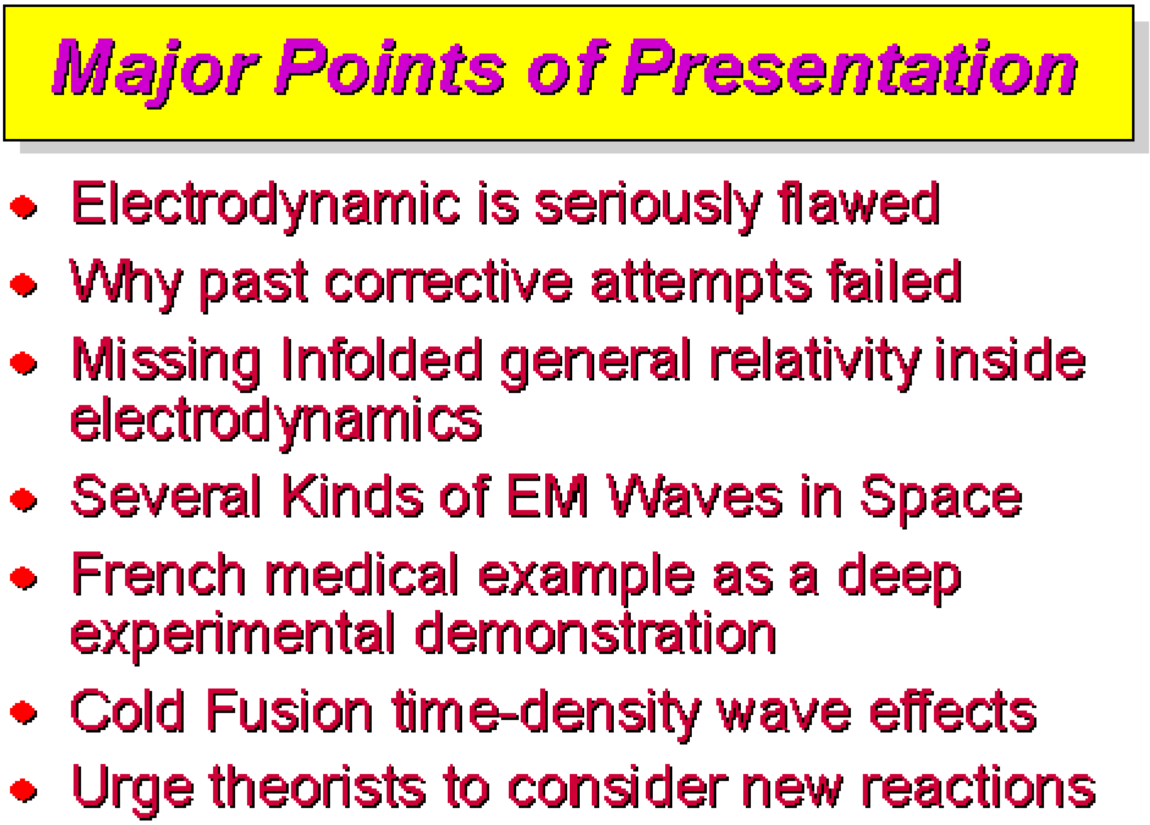

Major Points of Presentation

- Electrodynamic is seriously flawed
- Why past corrective attempts failed
- Missing Infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
- French medical example as a deep experimental demonstration
- Cold Fusion time-density wave effects
- Urge theorists to consider new reactions

Published by Lotus® Freelance Graphics®

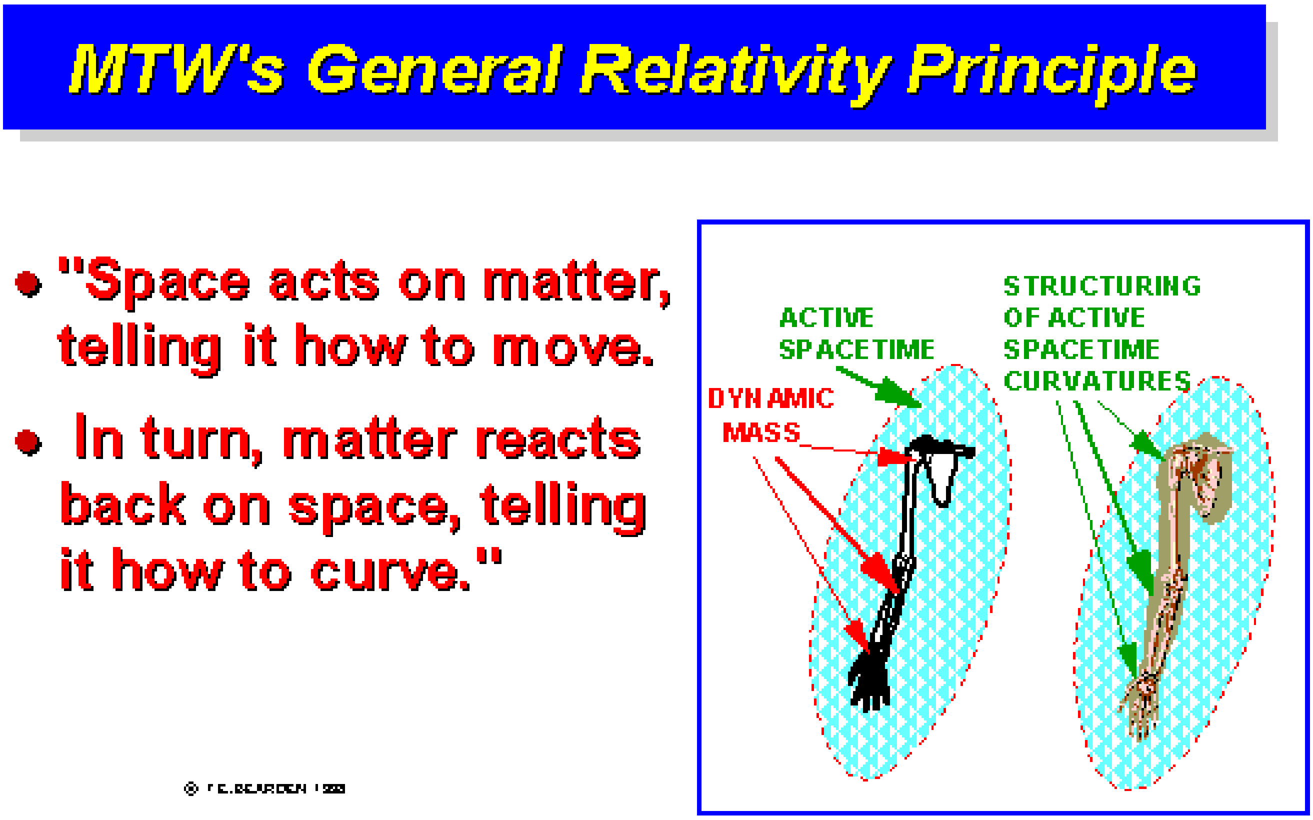

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE
SPACETIME
STRUCTURING
OF ACTIVE
SPACETIME
CURVATURES
DYNAMIC
MASS

© T.E.BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

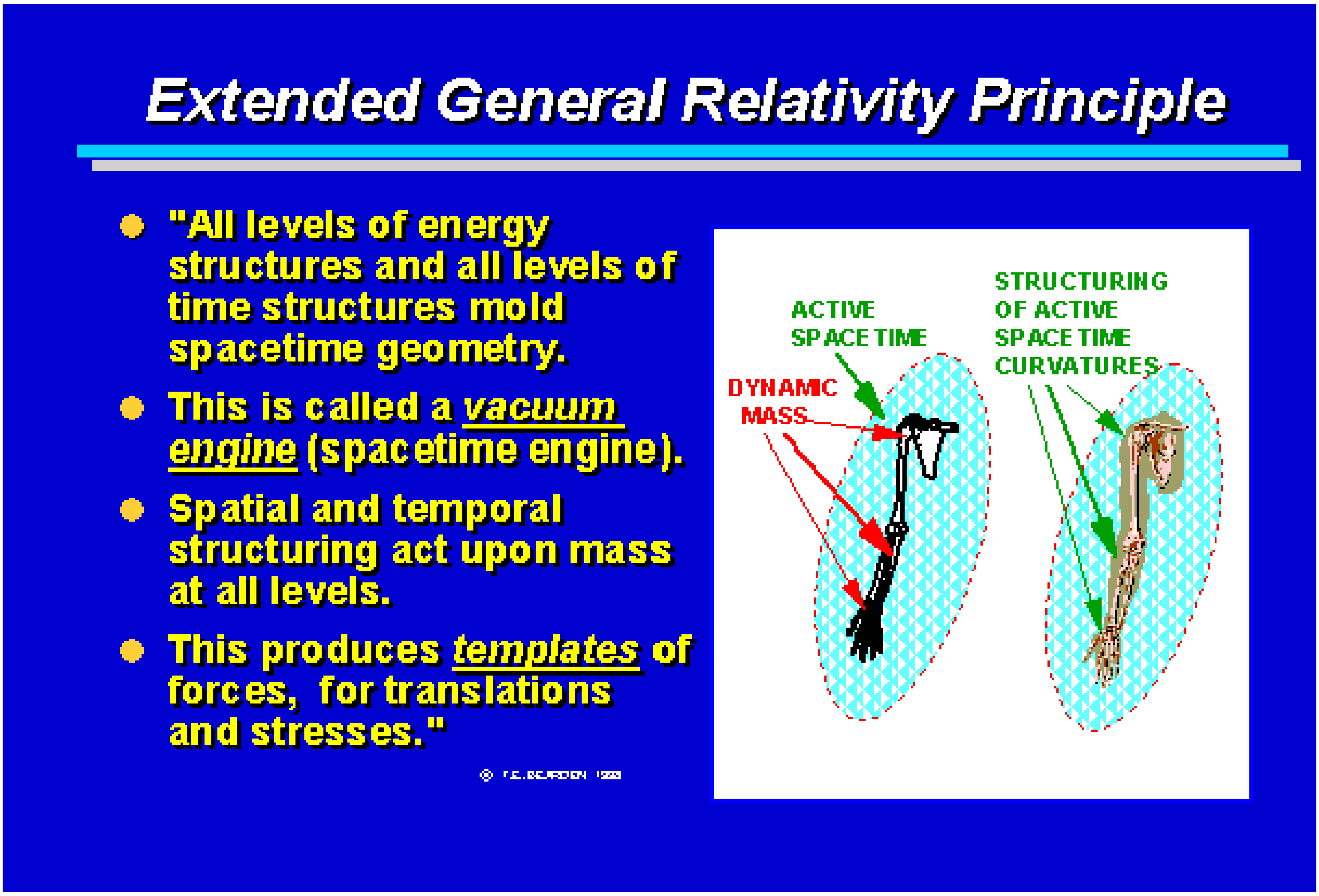

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a *vacuum engine* (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces *templates* of forces, for translations and stresses."

© T.E. BEARDEN [illegible]

ACTIVE SPACE TIME
STRUCTURING OF ACTIVE SPACE TIME CURVATURES
DYNAMIC MASS

Published by Lotus® Freelance Graphics®

Active Spacetime and a Specific Vacuum Engine (Spacetime Engine)

(A General Relativistic Concept and Analogy)

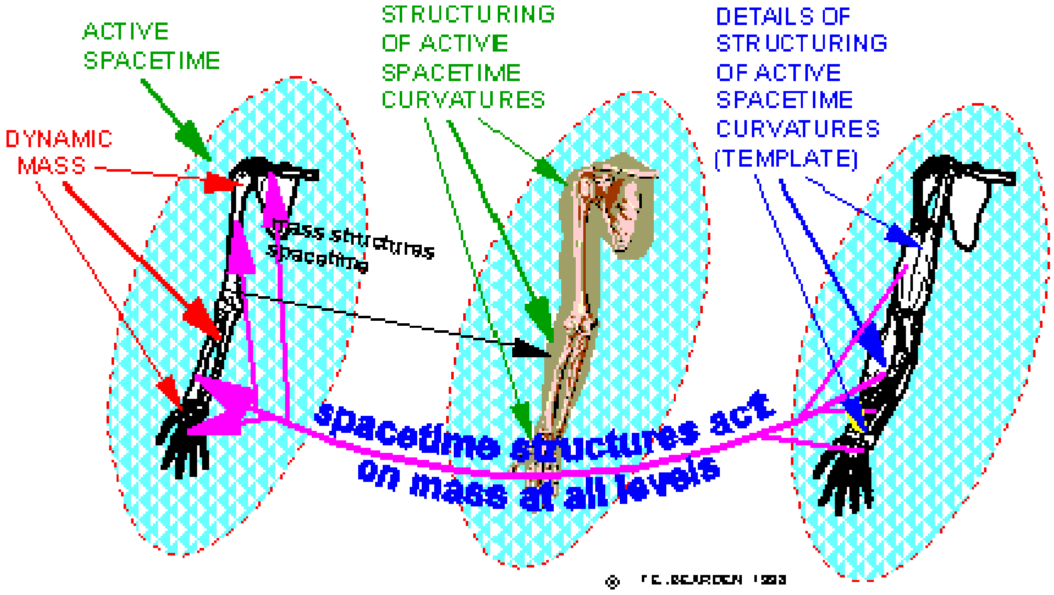
ACTIVE
SPACETIME
DYNAMIC
MASS
STRUCTURING
OF ACTIVE
SPACETIME
CURVATURES
DETAILS OF
STRUCTURING
OF ACTIVE
SPACETIME
CURVATURES
(TEMPLATE)
mass structures
spacetime
spacetime structures act
on mass at all levels
© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

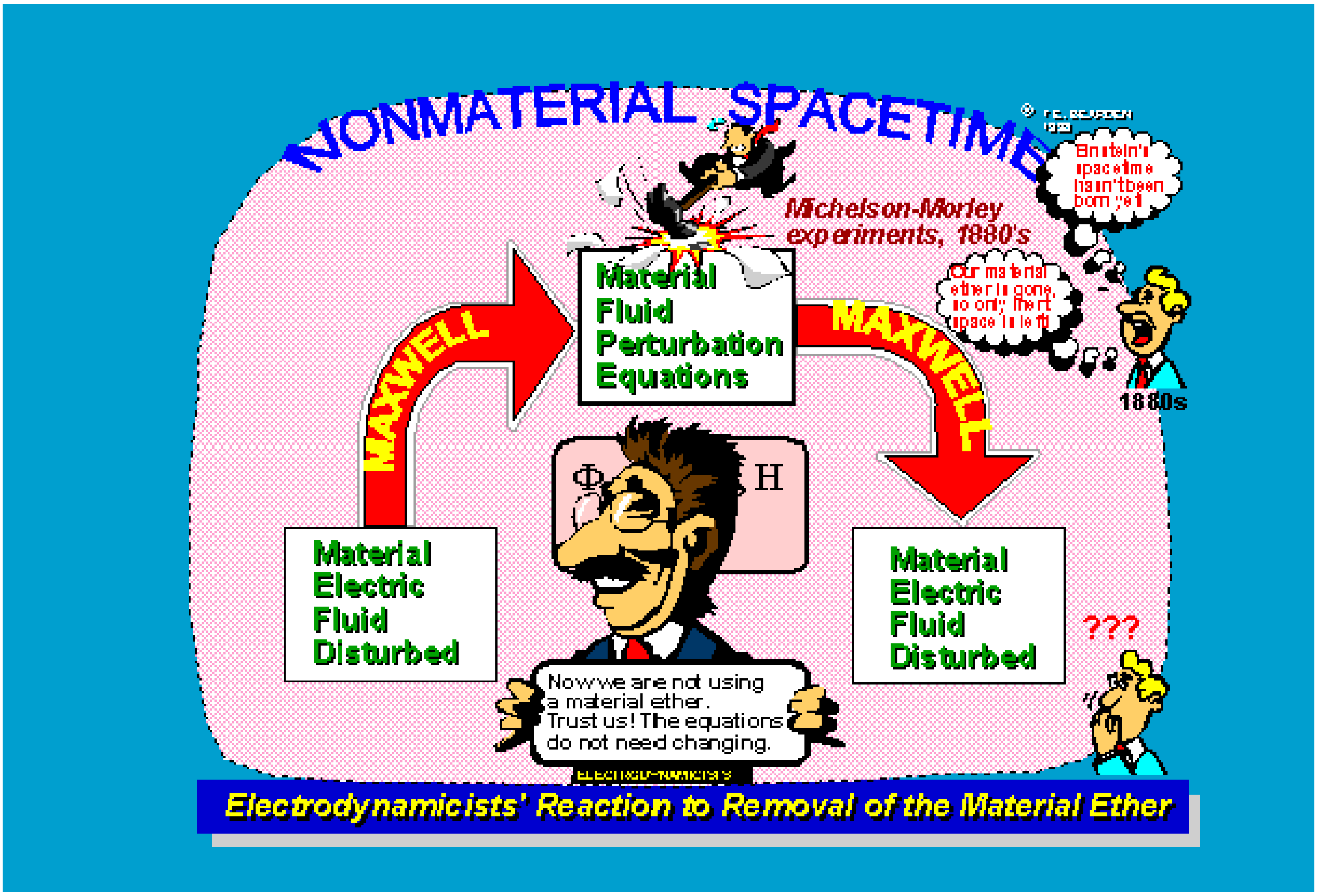
NONMATERIAL SPACETIME
© T.E. BEARDEN
Einstein's spacetime hasn't been born yet
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left
1880s
Material Fluid Perturbation Equations
MAXWELL
MAXWELL
Φ
H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

---

Published by Lotus® Freelance Graphics®

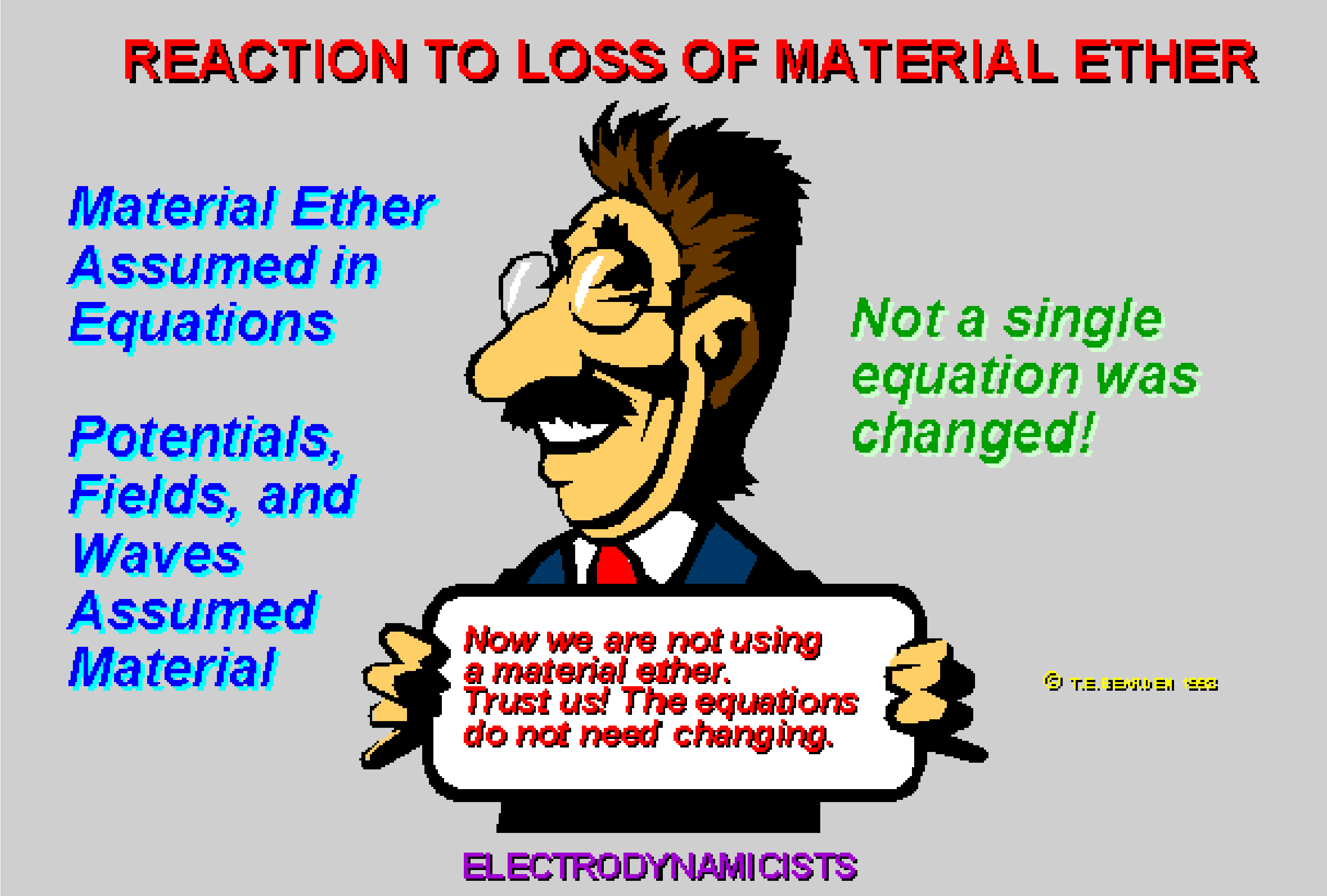

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.
ELECTRODYNAMICISTS

© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

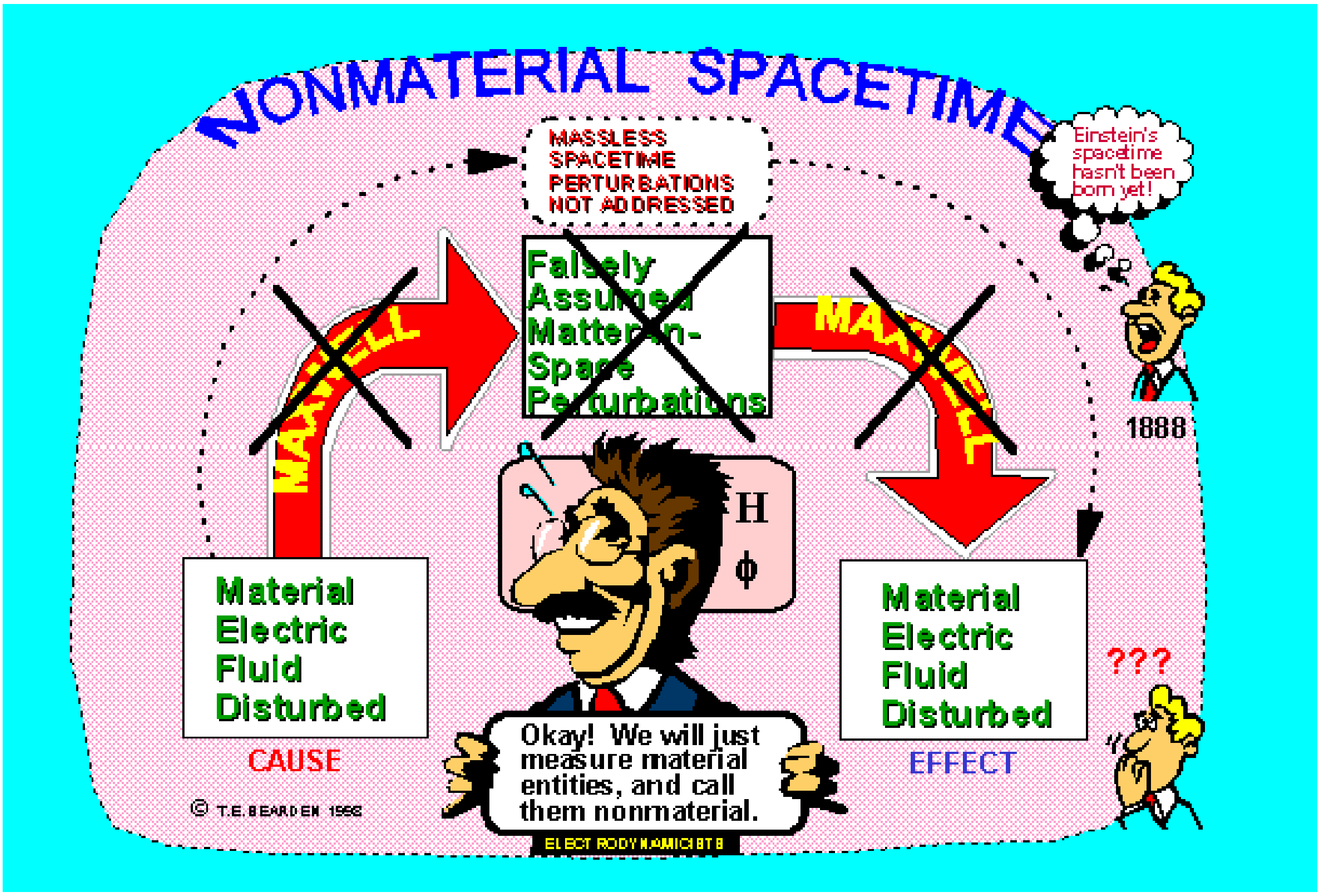
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
Falsely Assumed Matter-in-Space Perturbations
MAXWELL
MAXWELL
1888
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E.BEARDEN 1993

Published by Lotus® Freelance Graphics®

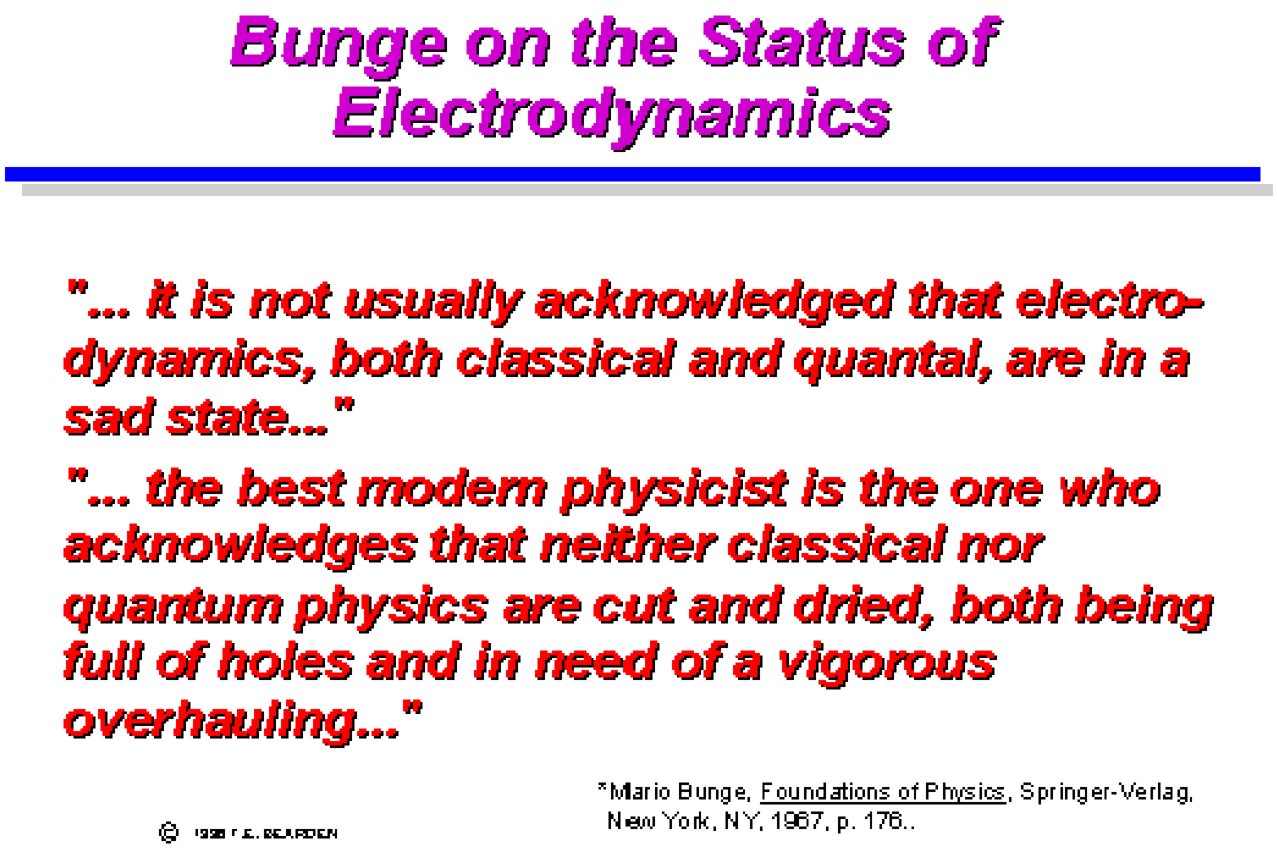

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© [illegible] T.E. BEARDEN

Published by Lotus® Freelance Graphics®

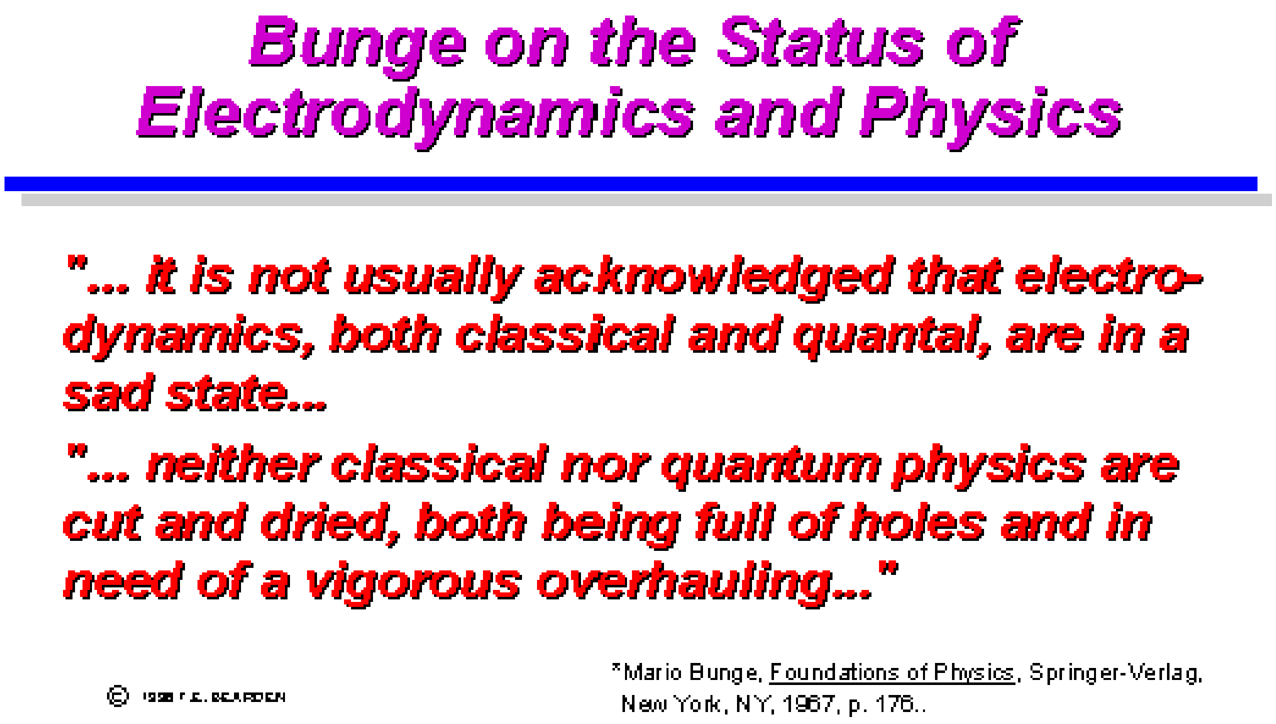

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© [illegible] T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

Published by Lotus® Freelance Graphics®

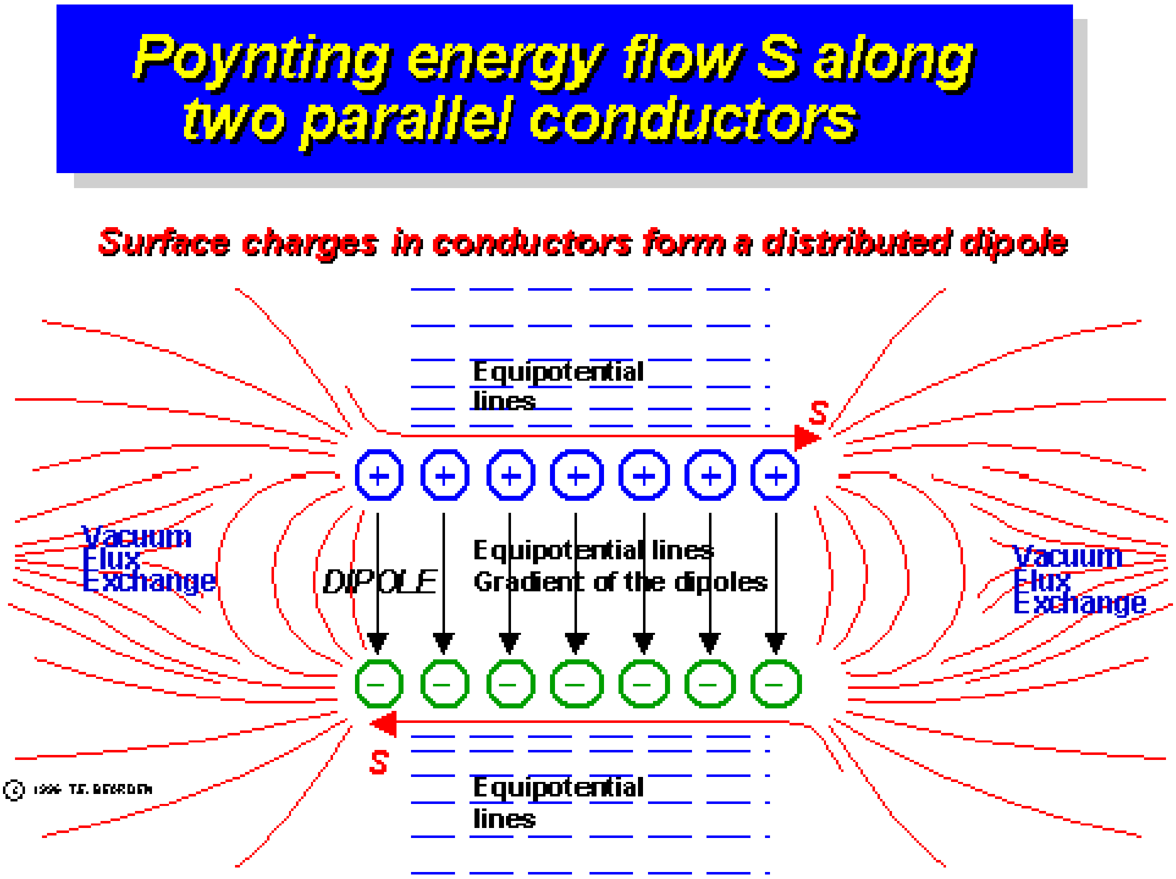

Poynting energy flow S along two parallel conductors

Surface charges in conductors form a distributed dipole

Equipotential lines
S
Vacuum Flux Exchange
DIPOLE
Equipotential lines Gradient of the dipoles
Vacuum Flux Exchange
S
Equipotential lines
© 1996 T.E. Bearden

Published by Lotus® Freelance Graphics®

As defined, fields and potentials only exist in and on charged matter

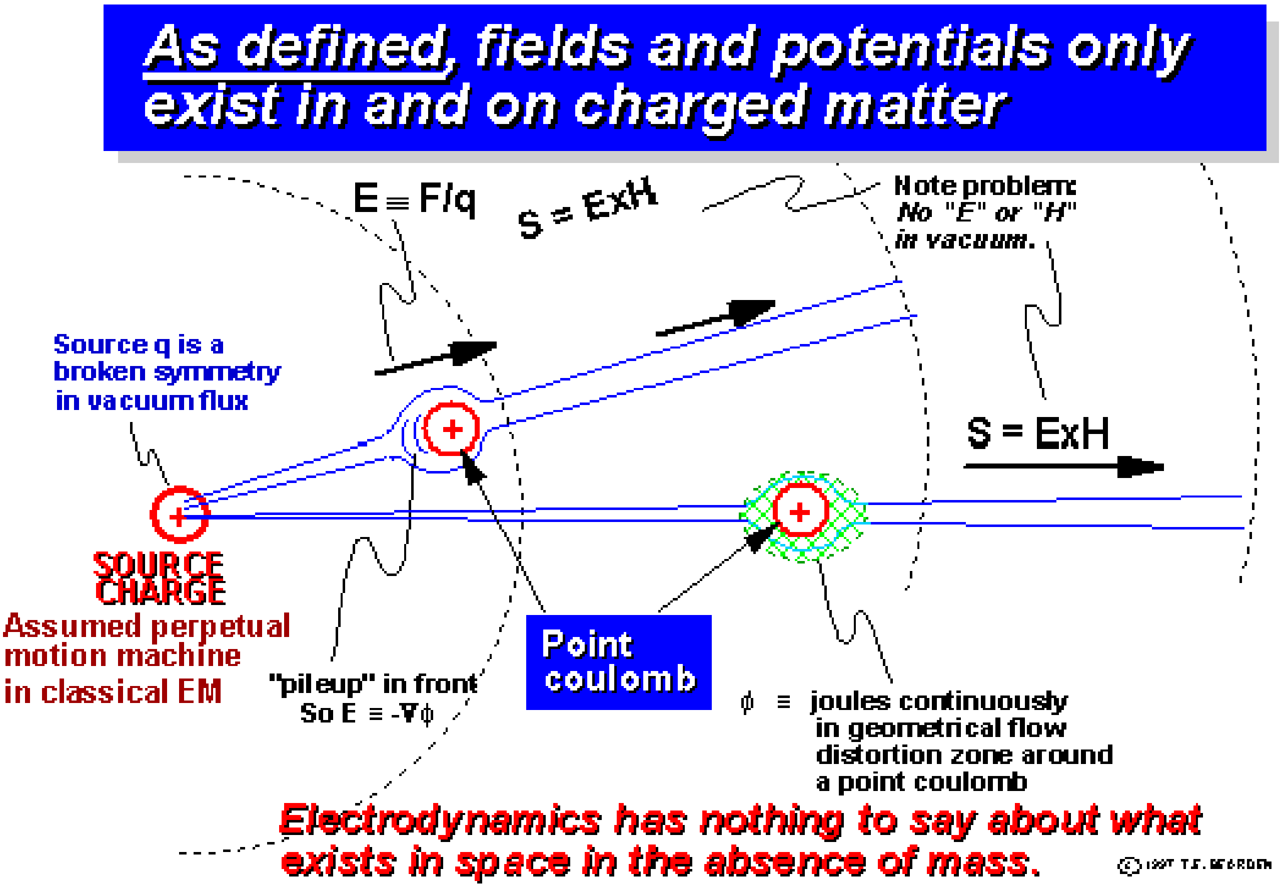
E ≡ F/q
S = ExH
Note problem:
No "E" or "H" in vacuum.
Source q is a
broken symmetry
in vacuum flux
S = ExH
SOURCE
CHARGE
Assumed perpetual
motion machine
in classical EM
"pileup" in front
So E ≡ -∇ϕ
Point
coulomb
ϕ ≡ joules continuously
in geometrical flow
distortion zone around
a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

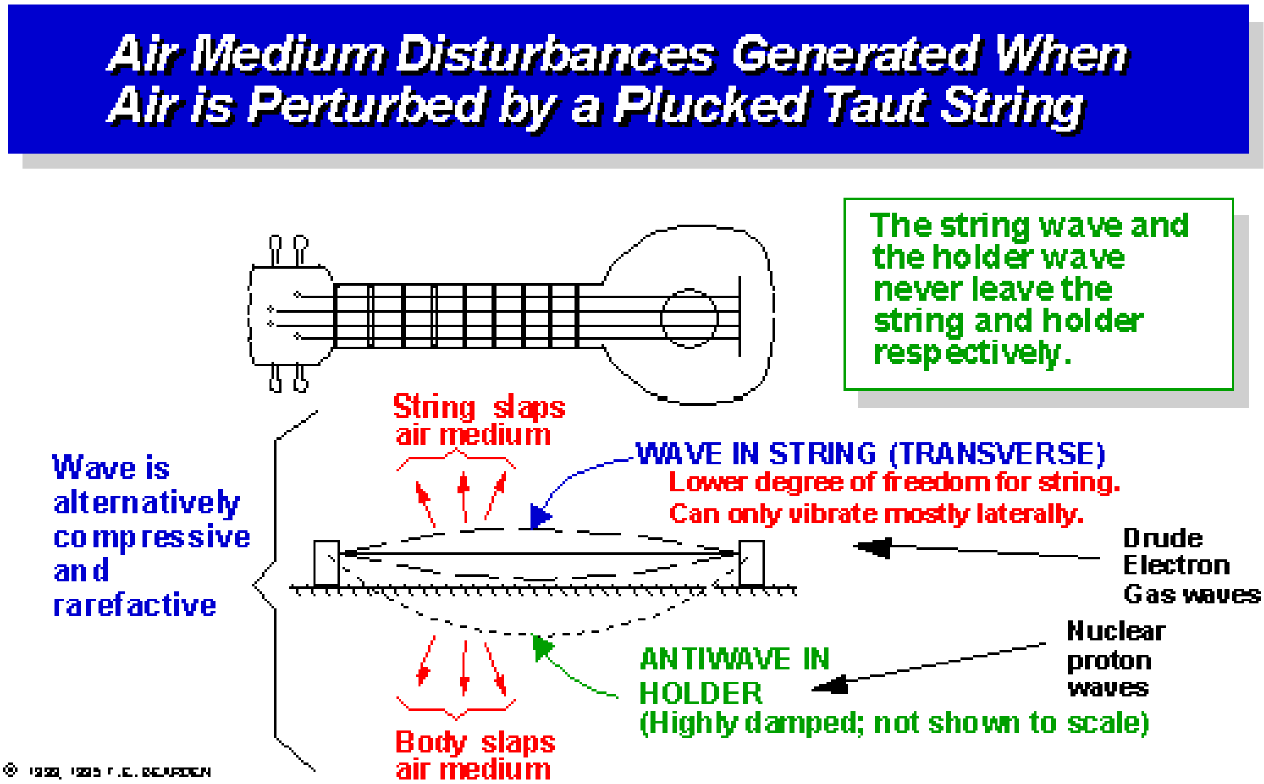

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

Wave is alternatively compressive and rarefactive

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string.
Can only vibrate mostly laterally.

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1990, 1995 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

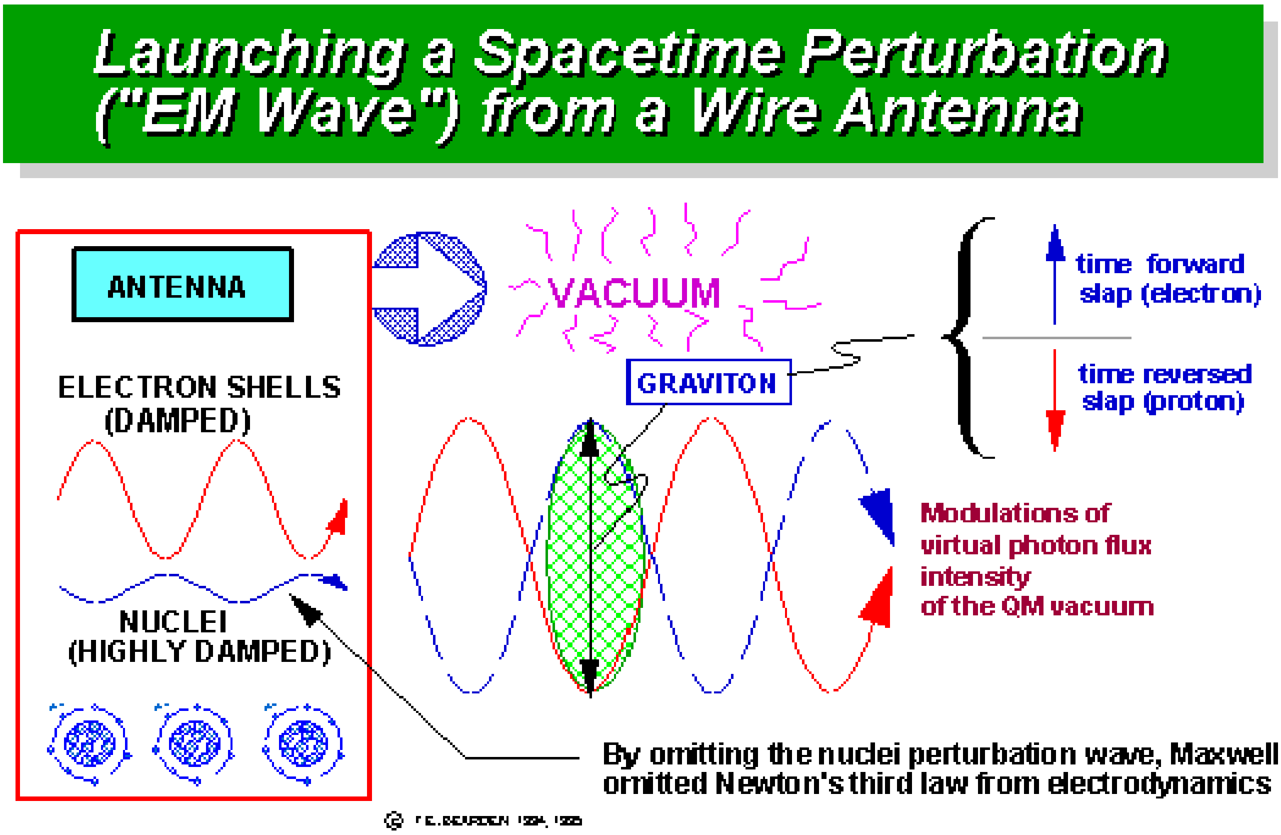

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA
ELECTRON SHELLS (DAMPED)
NUCLEI (HIGHLY DAMPED)
VACUUM
GRAVITON
time forward slap (electron)
time reversed slap (proton)
Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

Published by Lotus® Freelance Graphics®

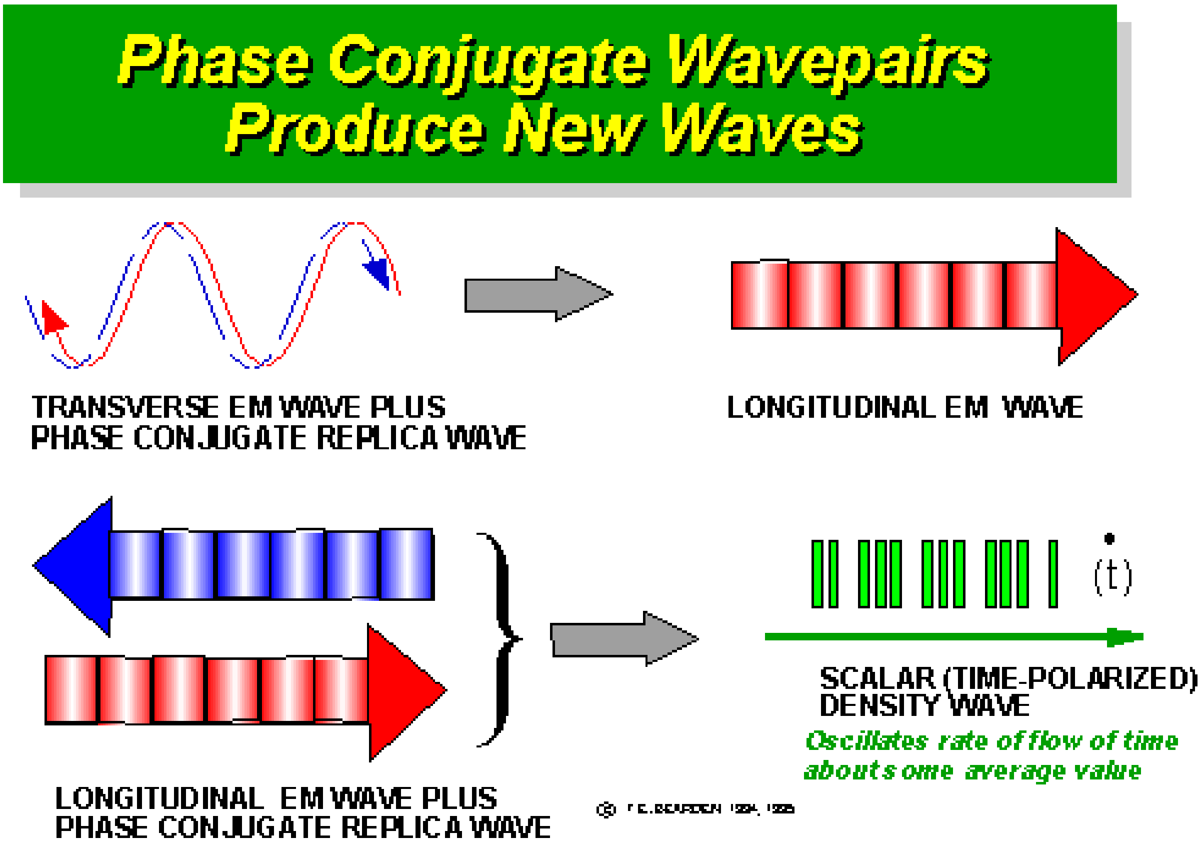
Phase Conjugate Wavepairs
Produce New Waves
TRANSVERSE EM WAVE PLUS PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM WAVE
(t)
SCALAR (TIME-POLARIZED) DENSITY WAVE
Oscillates rate of flow of time about some average value
LONGITUDINAL EM WAVE PLUS PHASE CONJUGATE REPLICA WAVE
© T.E. BEARDEN 1994, 1995

Published by Lotus® Freelance Graphics®

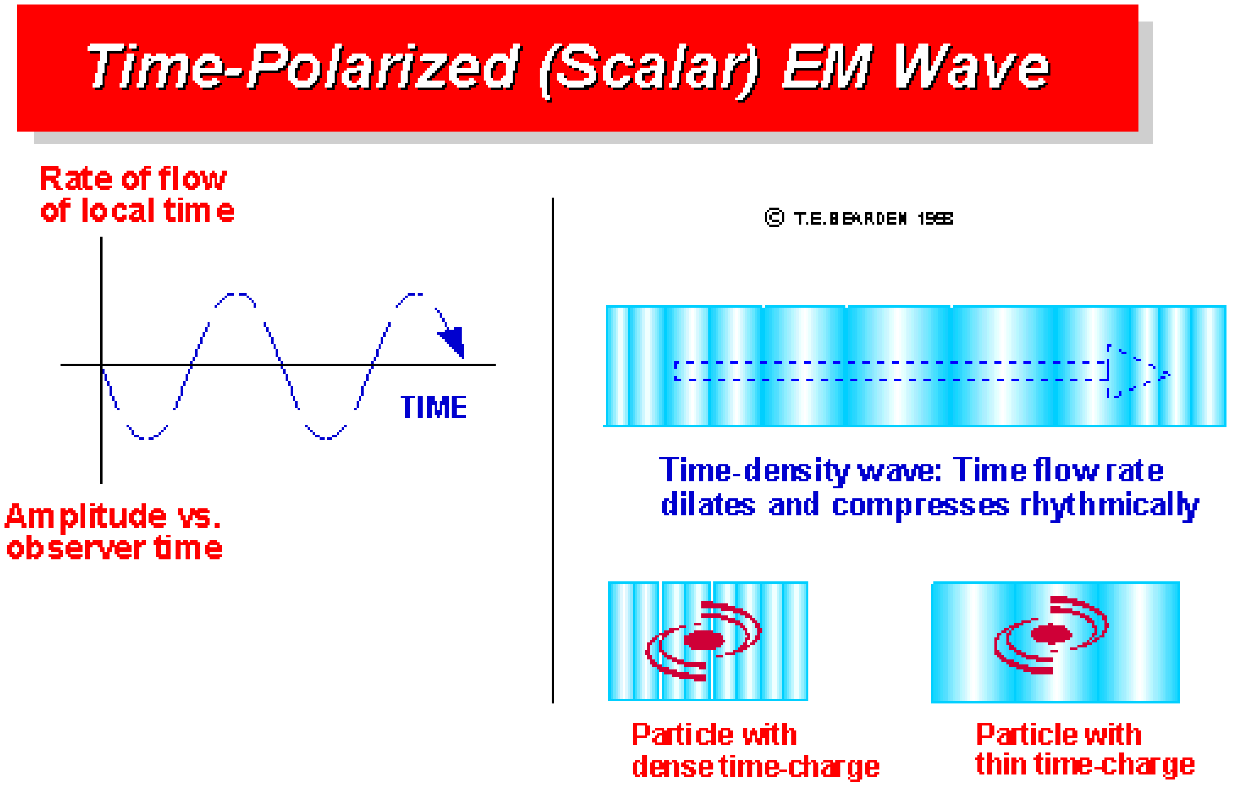

Time-Polarized (Scalar) EM Wave

Rate of flow
of local time
TIME
Amplitude vs.
observer time

© T.E. BEARDEN 1988

Time-density wave: Time flow rate
dilates and compresses rhythmically

Particle with
dense time-charge

Particle with
thin time-charge

Published by Lotus® Freelance Graphics®

Mechanism for the Flow of Time

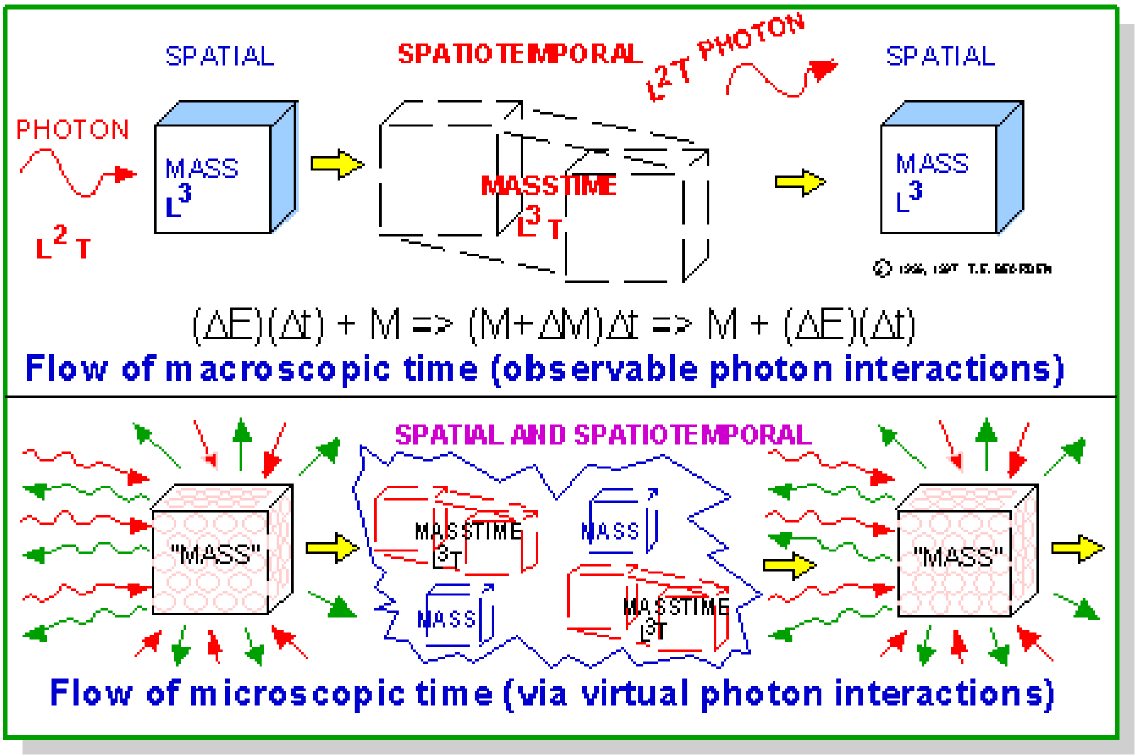
SPATIAL
SPATIOTEMPORAL
SPATIAL
PHOTON
L^2T
MASS
L^3
MASSTIME
L^3T
L^2T PHOTON
MASS
L^3
© 1996, 1997 T.E. BEARDEN
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

Flow of microscopic time (via virtual photon interactions)

Published by Lotus® Freelance Graphics®

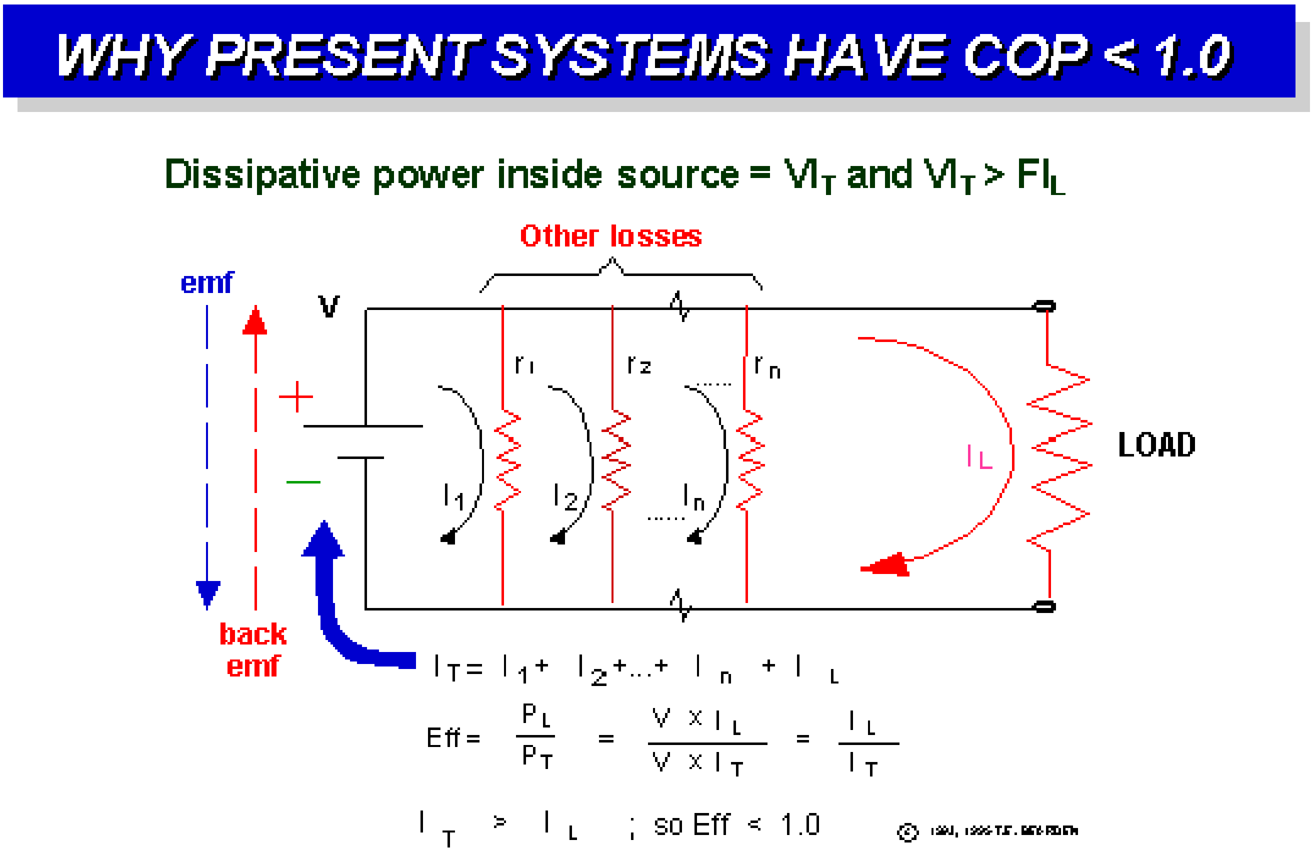

WHY PRESENT SYSTEMS HAVE COP < 1.0

Dissipative power inside source = VI_T and $VI_T > FI_L$

Other losses
emf
V
+
−
r_1
r_2
r_n
I_1
I_2
I_n
I_L
LOAD
back emf

$I_T = I_1 + I_2 + \ldots + I_n + I_L$

$$Eff = \frac{P_L}{P_T} = \frac{V \times I_L}{V \times I_T} = \frac{I_L}{I_T}$$

$I_T > I_L$; so Eff < 1.0

Published by Lotus® Freelance Graphics®

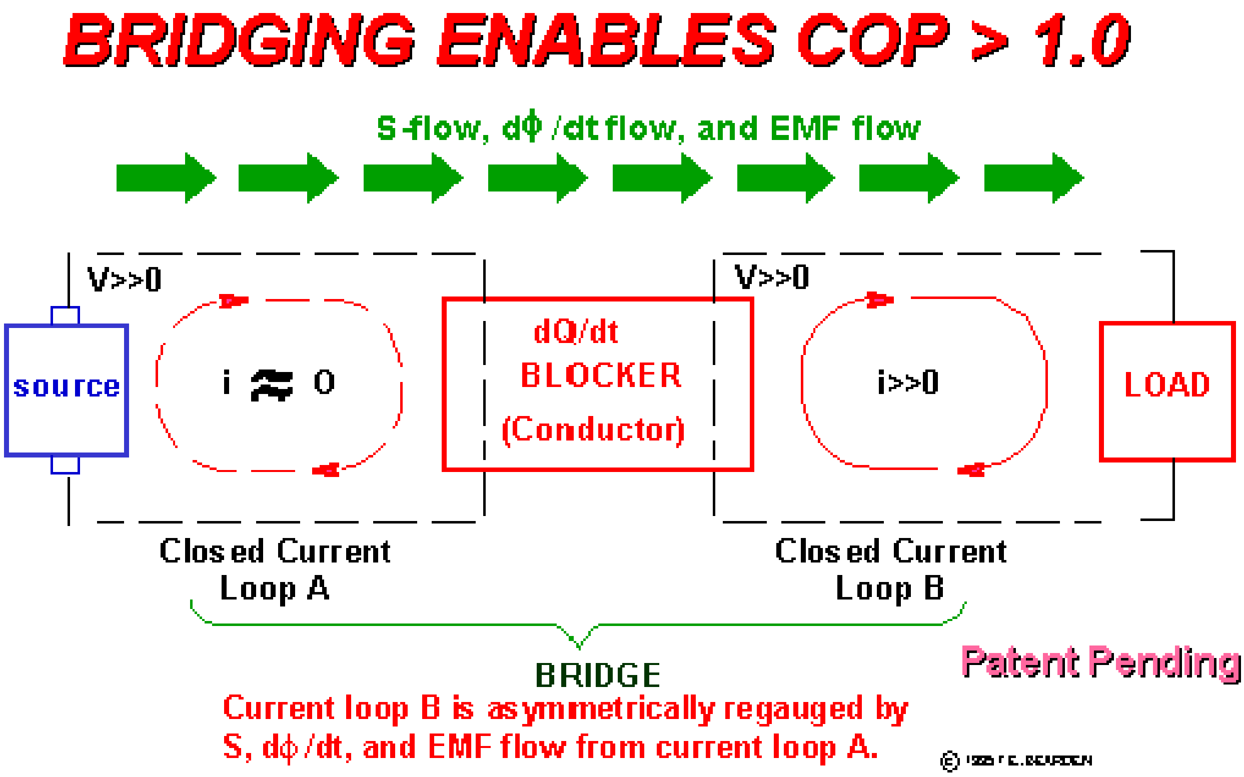

BRIDGING ENABLES COP > 1.0

S-flow, dϕ/dt flow, and EMF flow
V>>0
source
i ≉ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current Loop A
Closed Current Loop B
BRIDGE
© [illegible] T.E. BEARDEN

Current loop B is asymmetrically regauged by S, dϕ/dt, and EMF flow from current loop A.

Patent Pending

Published by Lotus® Freelance Graphics®

"Static" Poynting free-energy generator

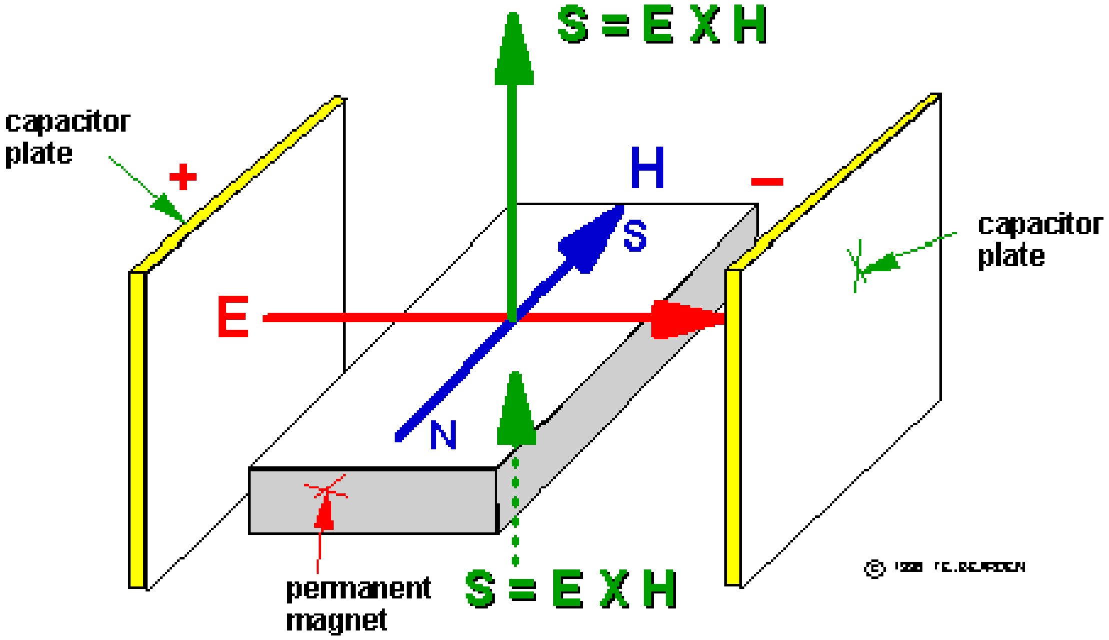
S = E X H
capacitor
plate
+
H
–
S
capacitor
plate
E
N
permanent
magnet
S = E X H
© [illegible] T.E.BEARDEN

Published by Lotus® Freelance Graphics®

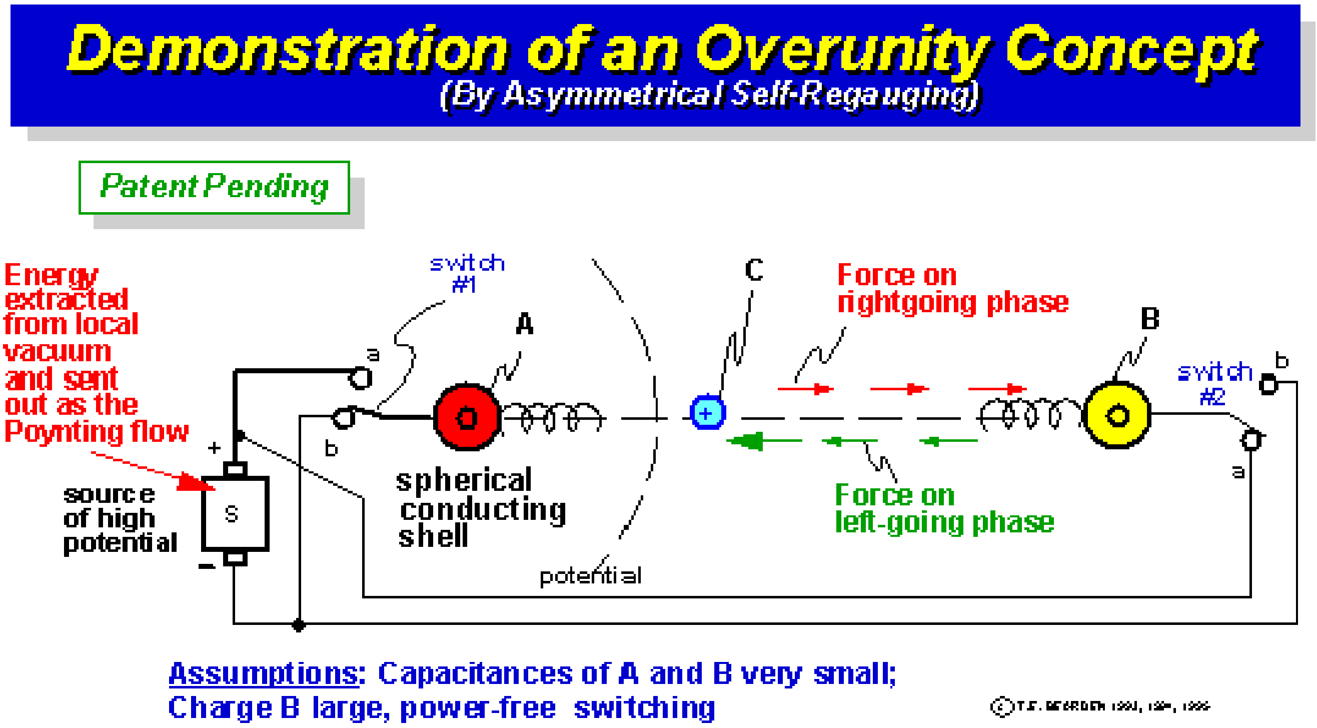

Demonstration of an Overunity Concept

(By Asymmetrical Self-Regauging)

Patent Pending

Energy extracted from local vacuum and sent out as the Poynting flow
switch #1
a
b
A
C
Force on rightgoing phase
B
switch #2
b
a
+
source of high potential
S
−
spherical conducting shell
potential
Force on left-going phase

Assumptions: Capacitances of A and B very small;
Charge B large, power-free switching

© T.E. BEARDEN [illegible], [illegible], [illegible]

Published by Lotus® Freelance Graphics®

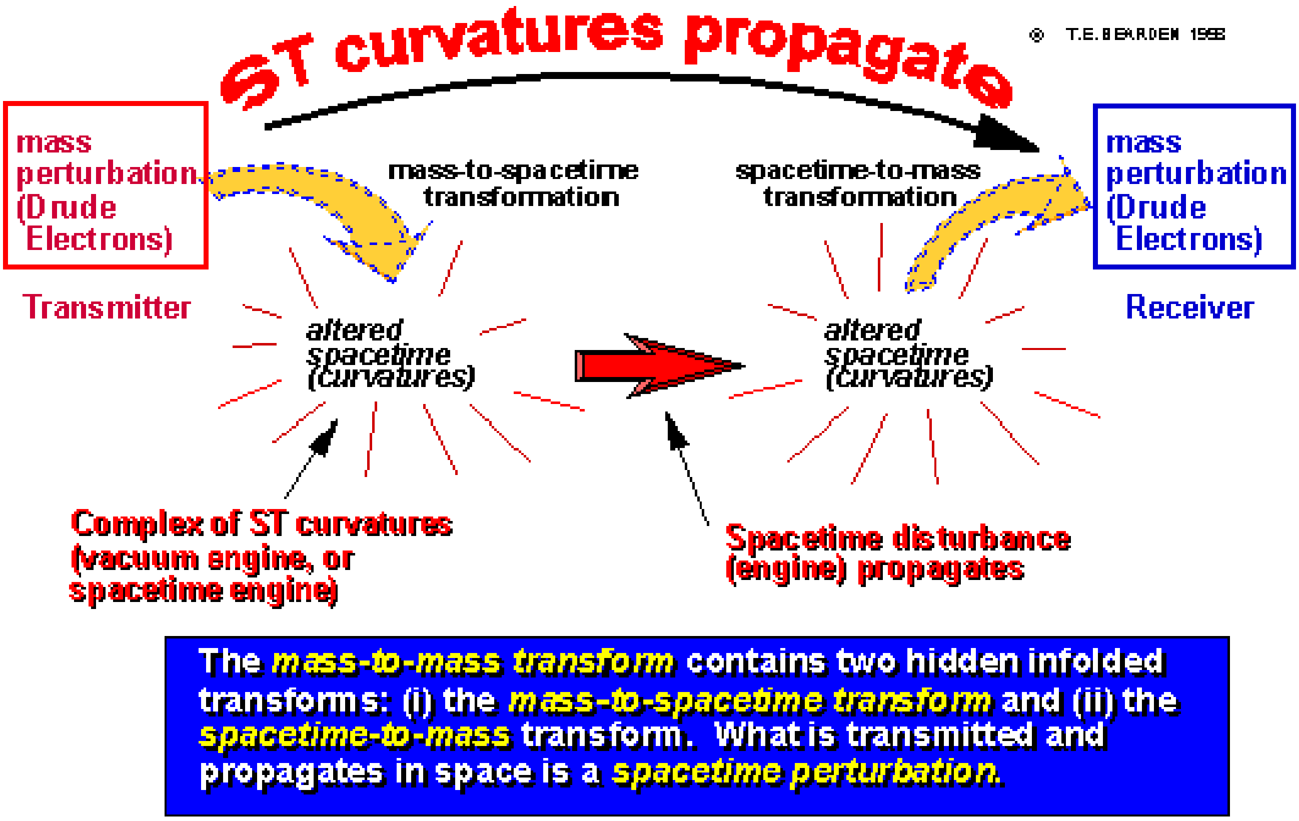
ST curvatures propagate
© T.E.BEARDEN 1998
mass perturbation (Drude Electrons)
Transmitter
mass-to-spacetime transformation
spacetime-to-mass transformation
mass perturbation (Drude Electrons)
Receiver
altered spacetime (curvatures)
altered spacetime (curvatures)
Complex of ST curvatures (vacuum engine, or spacetime engine)
Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

Published by Lotus® Freelance Graphics®

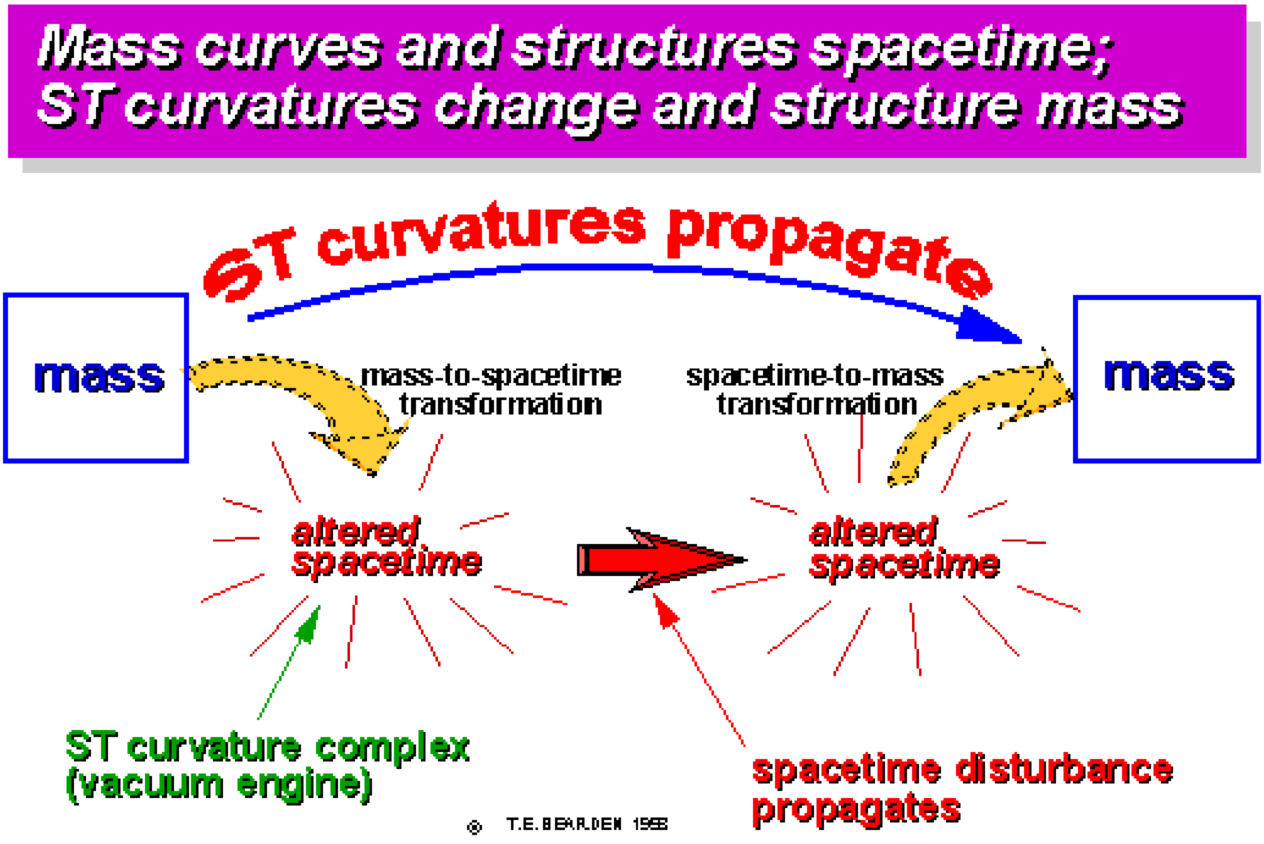

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

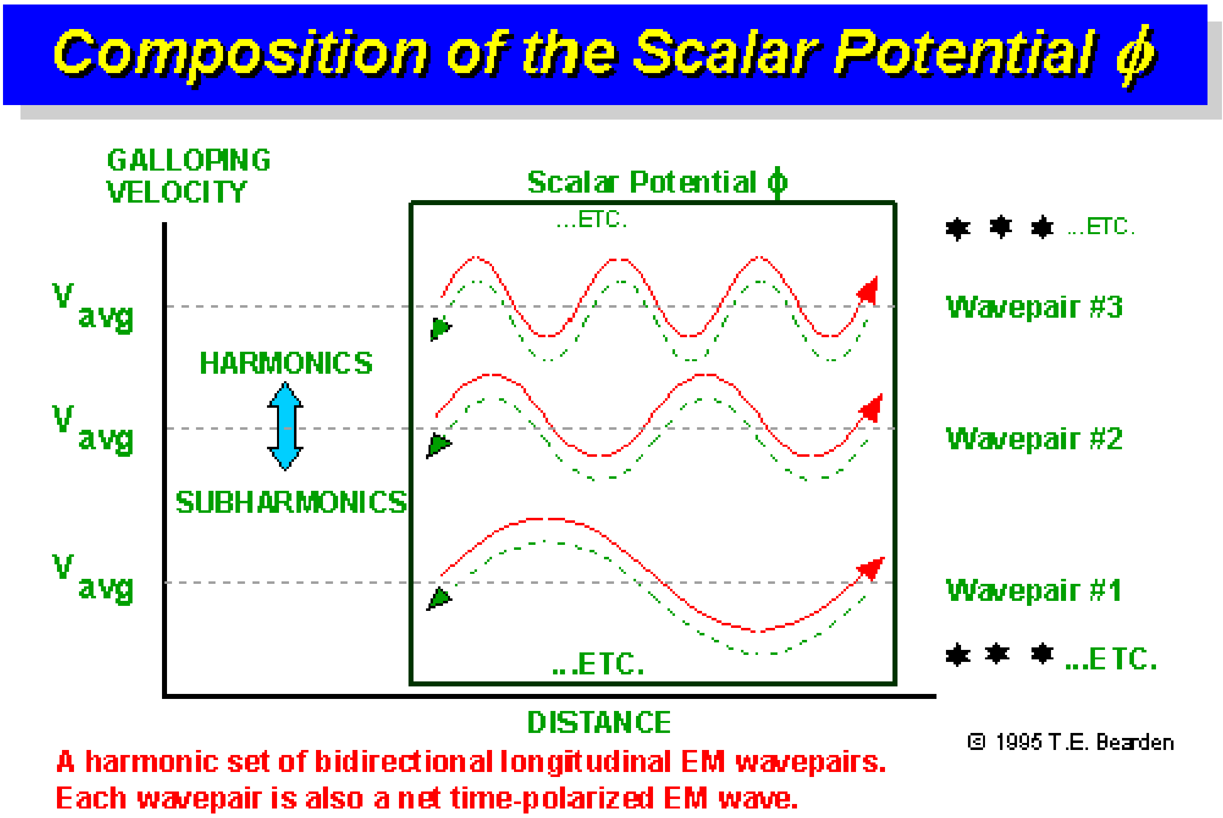

Composition of the Scalar Potential ϕ

GALLOPING
VELOCITY
Scalar Potential ϕ
...ETC.
✸ ✸ ✸ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
✸ ✸ ✸ ...ETC.
DISTANCE

A harmonic set of bidirectional longitudinal EM wavepairs.
Each wavepair is also a net time-polarized EM wave.

© 1995 T.E. Bearden

Published by Lotus® Freelance Graphics®

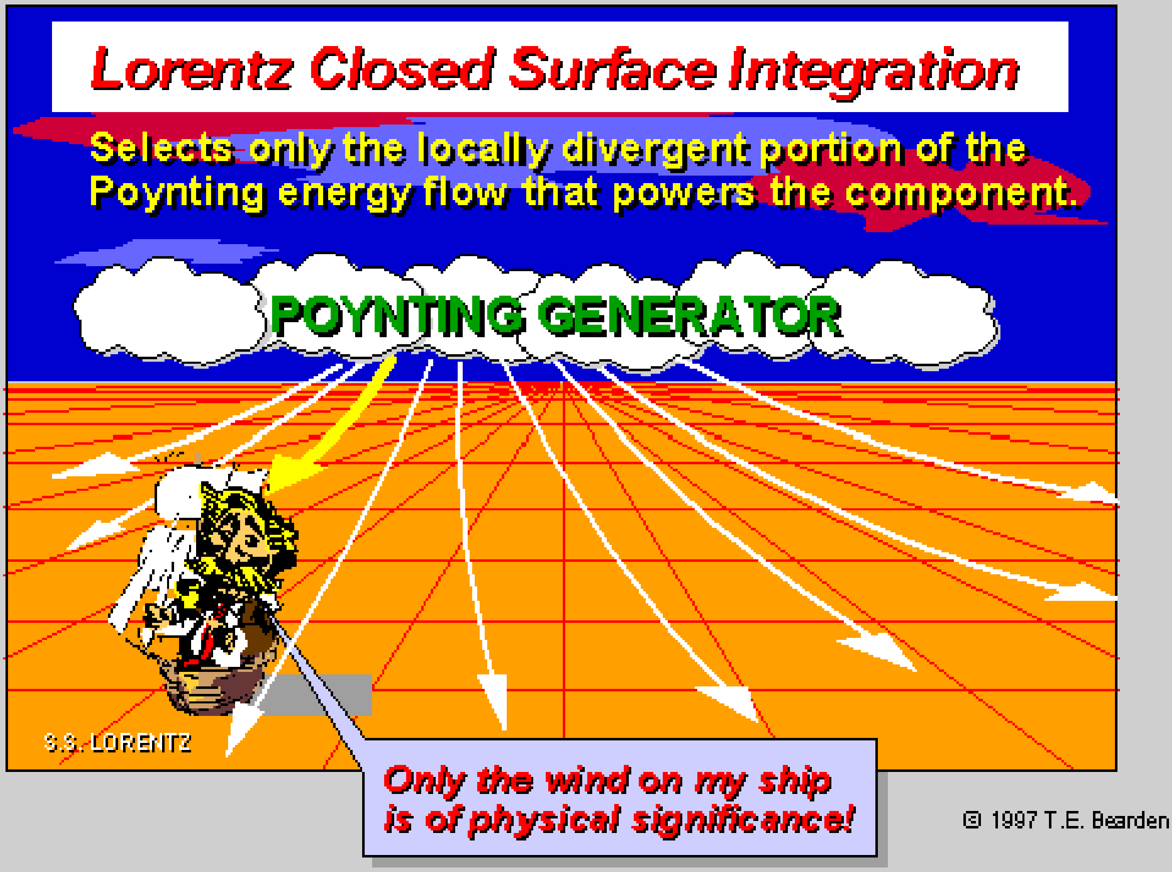

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

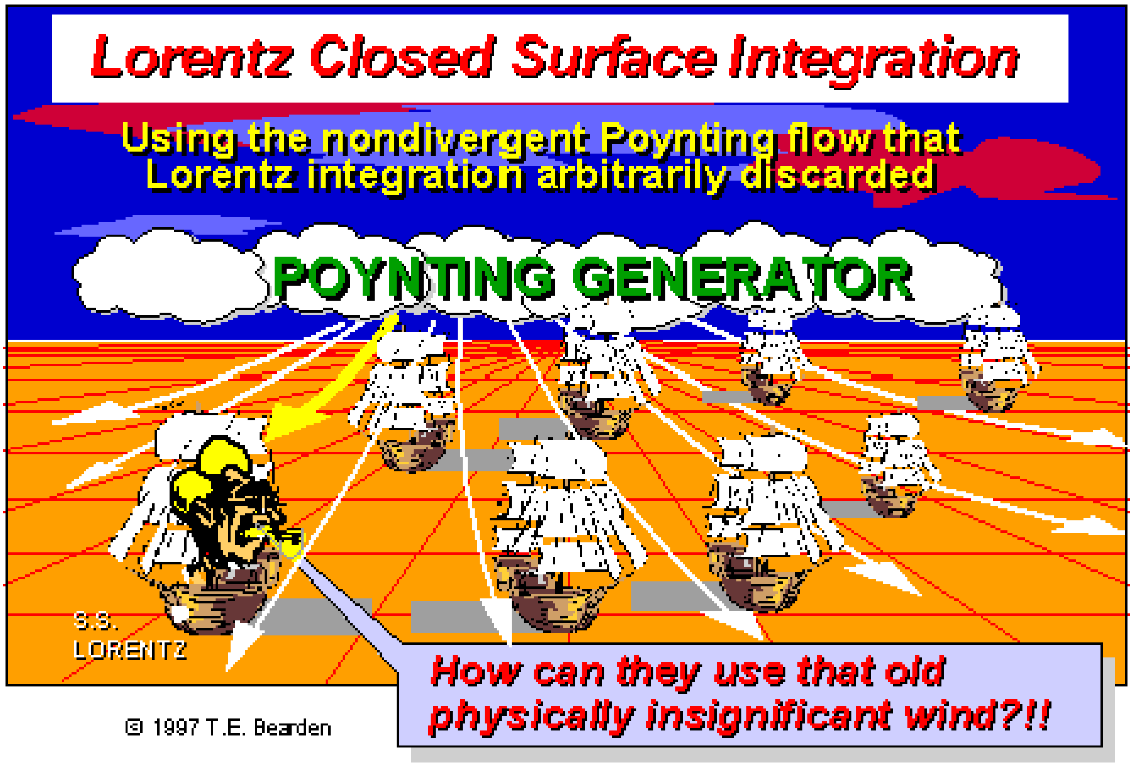

Lorentz Closed Surface Integration

Using the nondivergent Poynting flow that
Lorentz integration arbitrarily discarded

POYNTING GENERATOR

How can they use that old physically insignificant wind?!!

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

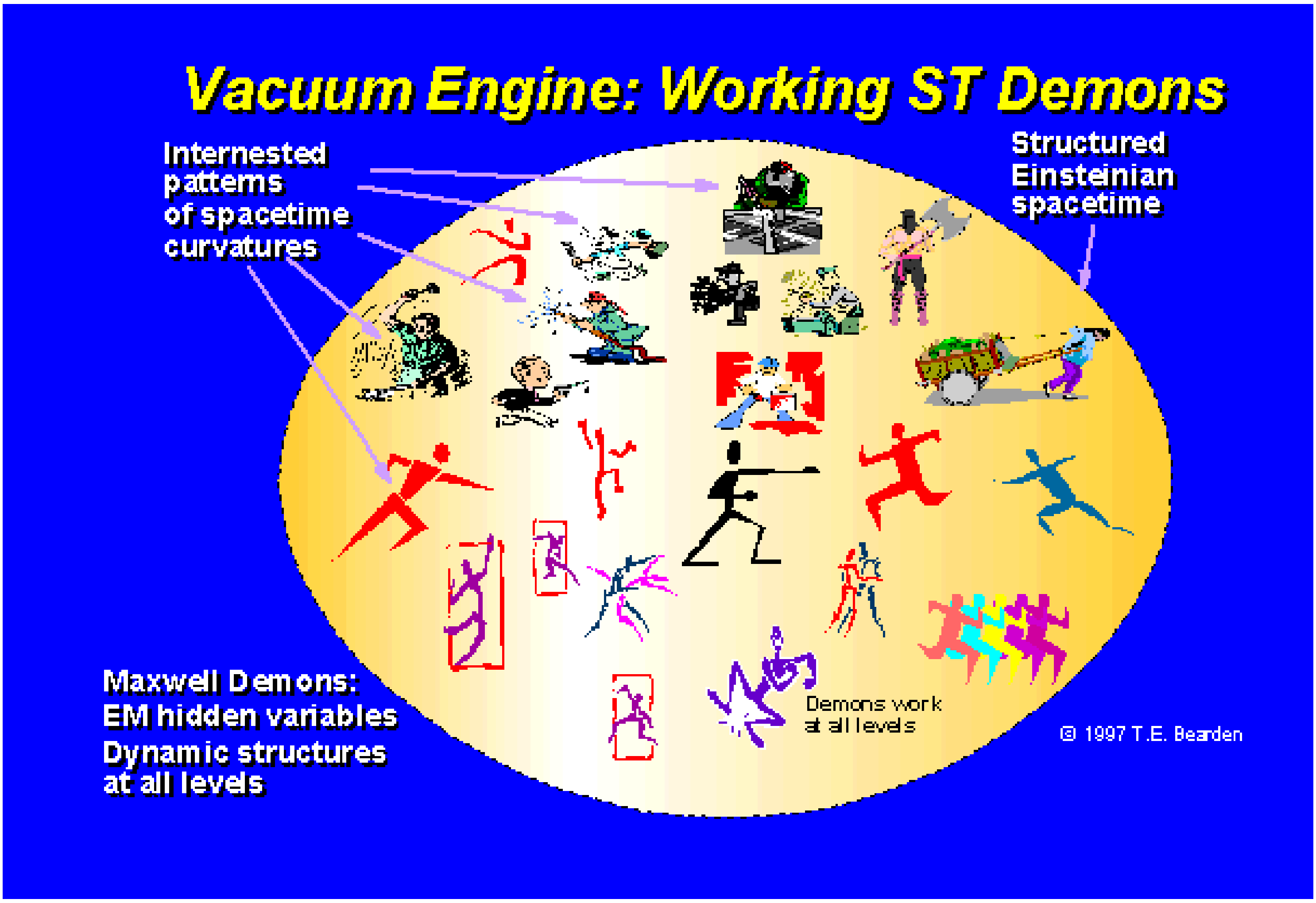

Vacuum Engine: Working ST Demons

Internested patterns of spacetime curvatures

Structured Einsteinian spacetime

Maxwell Demons:
EM hidden variables
Dynamic structures at all levels

Demons work at all levels

© 1997 T.E. Bearden

Published by Lotus® Freelance Graphics®

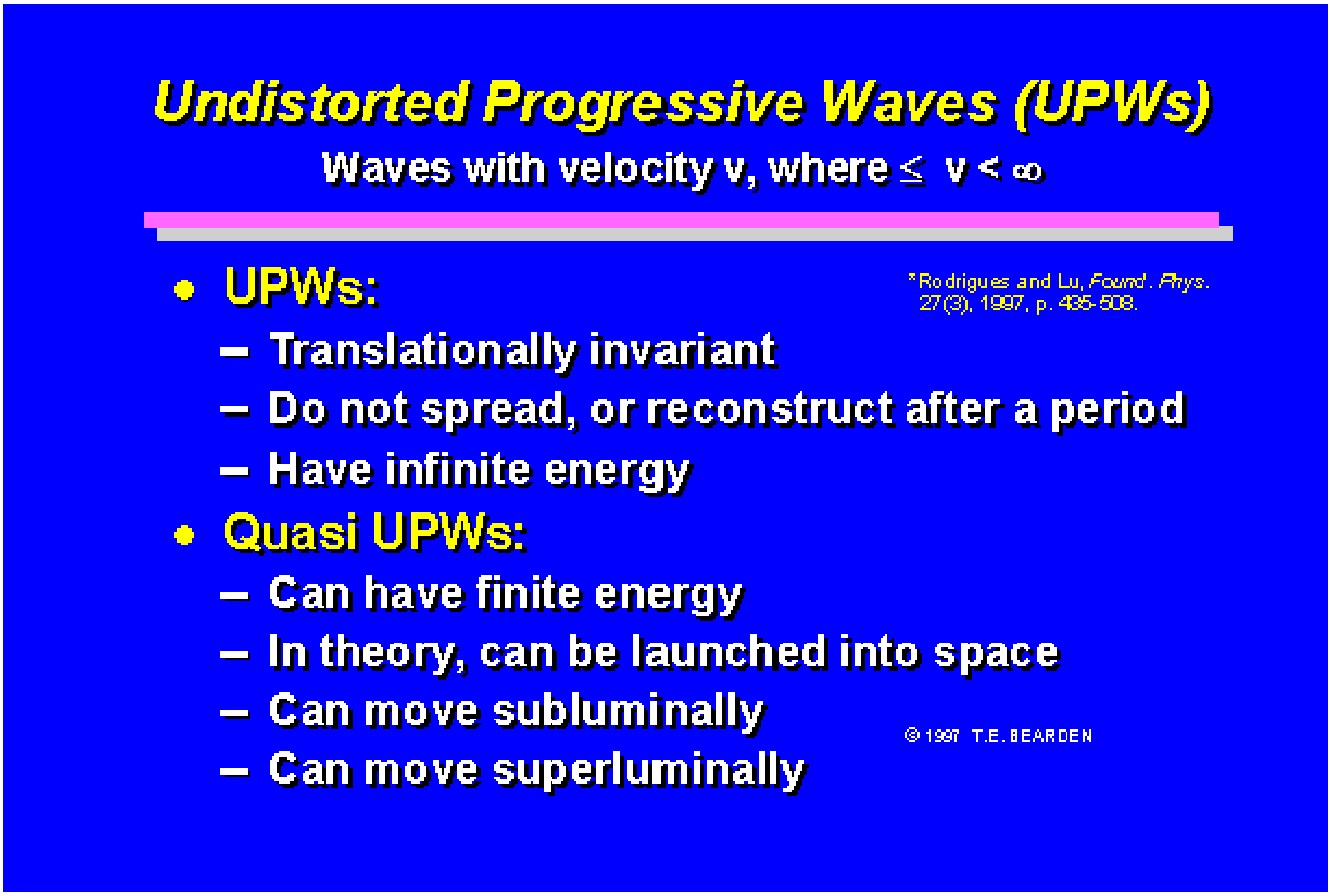

Undistorted Progressive Waves (UPWs)

Waves with velocity v, where $\leq v < \infty$

- UPWs:
  - Translationally invariant
  - Do not spread, or reconstruct after a period
  - Have infinite energy
- Quasi UPWs:
  - Can have finite energy
  - In theory, can be launched into space
  - Can move subluminally
  - Can move superluminally

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

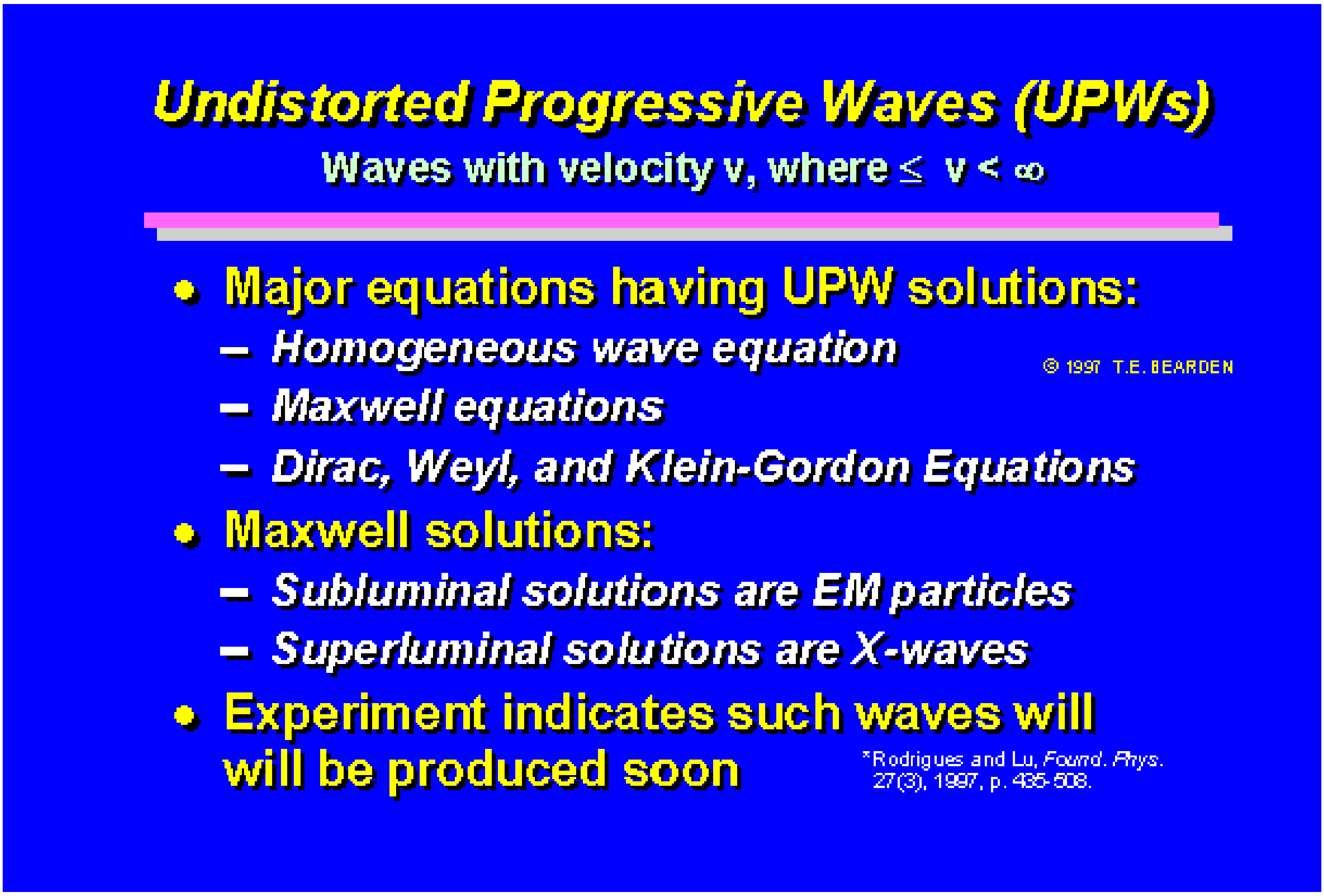

Undistorted Progressive Waves (UPWs)

Waves with velocity v, where ≤ v < ∞

- Major equations having UPW solutions:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon Equations*
- Maxwell solutions:
  - *Subluminal solutions are EM particles*
  - *Superluminal solutions are X-waves*
- Experiment indicates such waves will will be produced soon

© 1997 T.E. BEARDEN

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

Published by Lotus® Freelance Graphics®

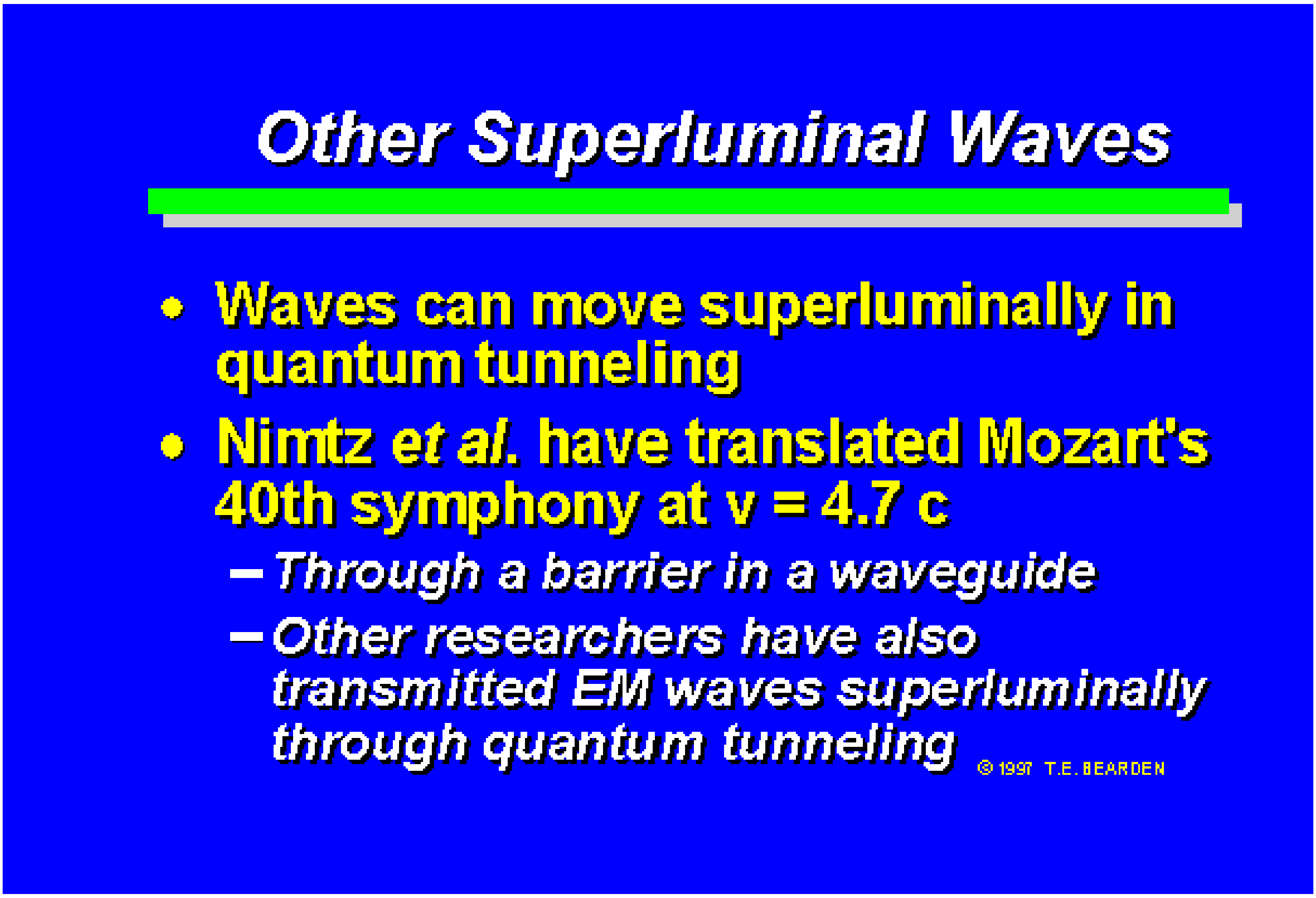

Other Superluminal Waves

- Waves can move superluminally in quantum tunneling
- Nimtz *et al*. have translated Mozart's 40th symphony at v = 4.7 c
  - *Through a barrier in a waveguide*
  - *Other researchers have also transmitted EM waves superluminally through quantum tunneling*

© 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

c 1993, 1995 T.E. BEARDEN

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

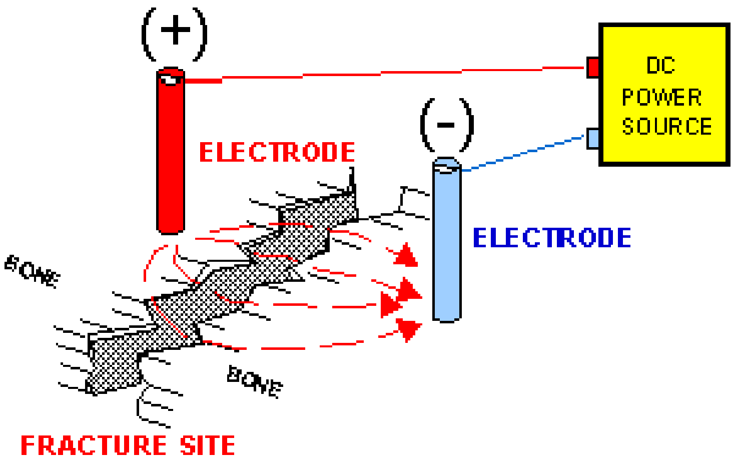

© 1993, 1995 T.E. BEARDEN

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Published by Lotus® Freelance Graphics®

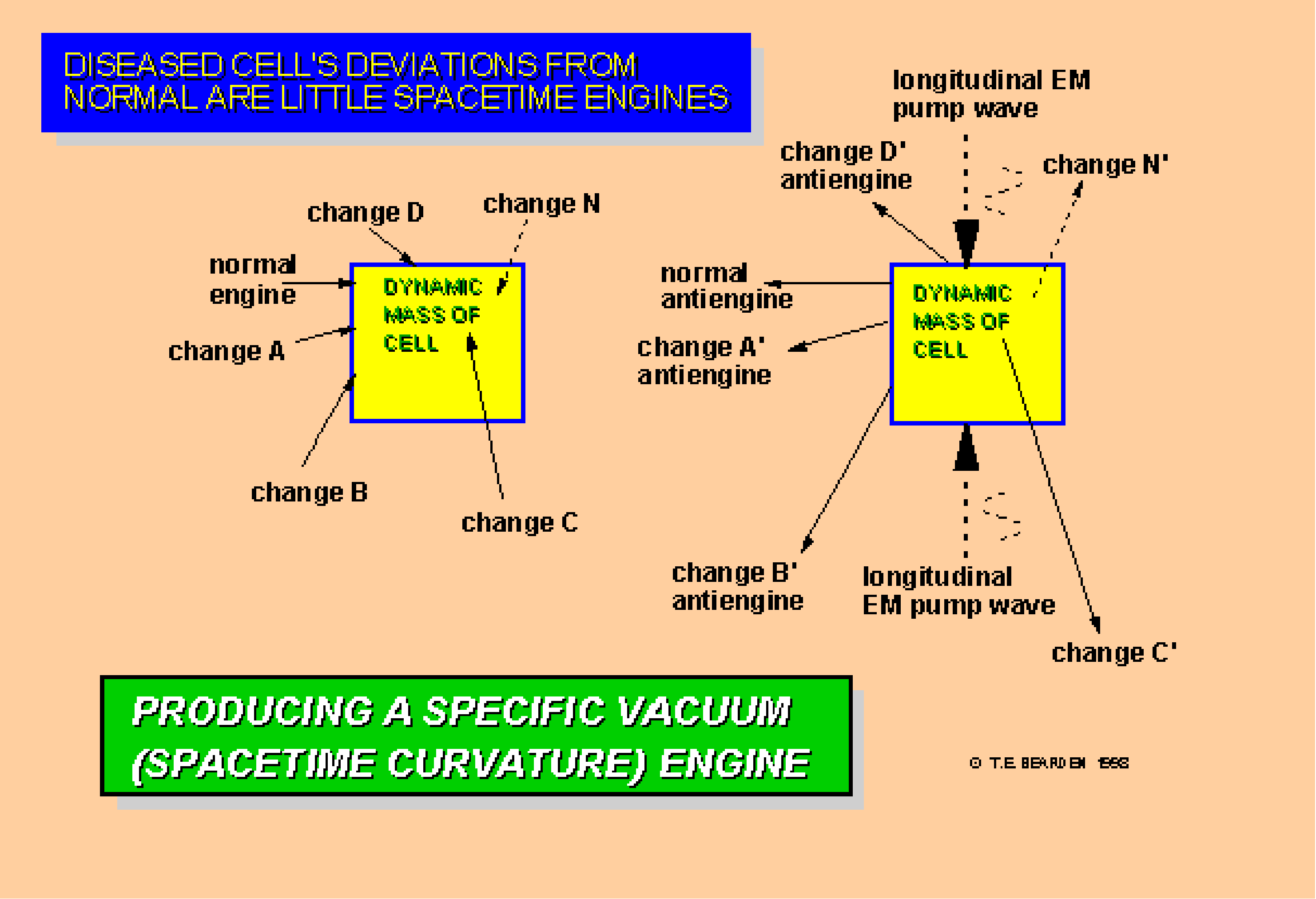

DISEASED CELL'S DEVIATIONS FROM NORMAL ARE LITTLE SPACETIME ENGINES

change D
change N
normal engine
change A
DYNAMIC MASS OF CELL
change B
change C
longitudinal EM pump wave
change D' antiengine
change N'
normal antiengine
change A' antiengine
DYNAMIC MASS OF CELL
change B' antiengine
longitudinal EM pump wave
change C'

PRODUCING A SPECIFIC VACUUM (SPACETIME CURVATURE) ENGINE

© T.E. BEARDEN 1993

Published by Lotus® Freelance Graphics®

Transverse EM Wave Pumping of PCM

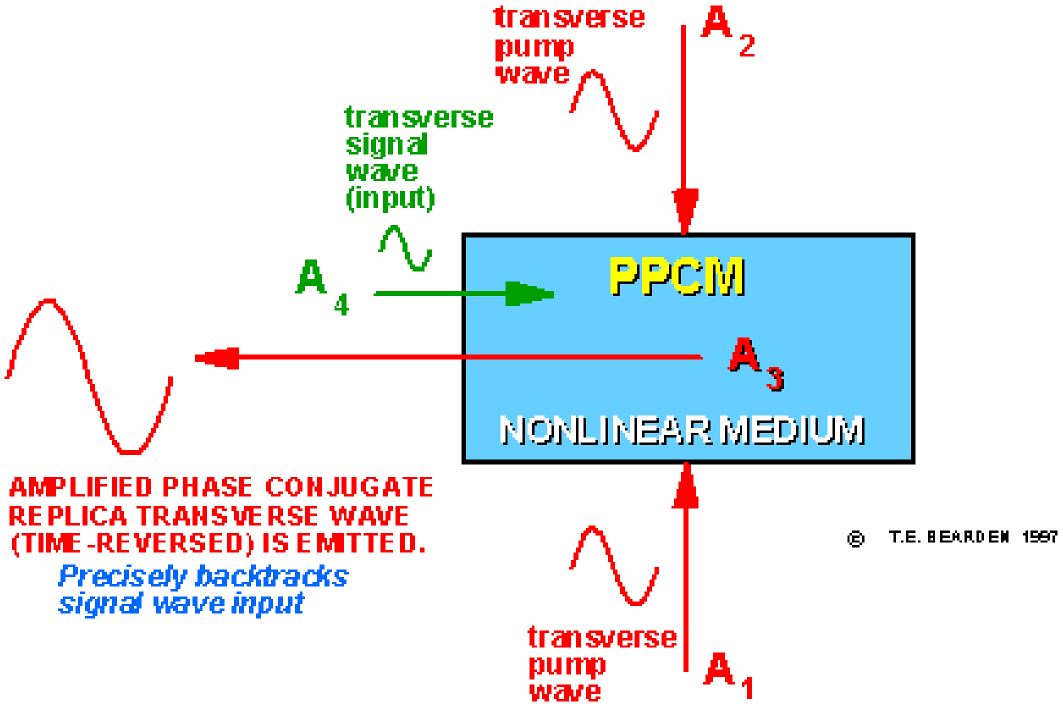
transverse
pump
wave
A_2
transverse
signal
wave
(input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
transverse
pump
wave
A_1

AMPLIFIED PHASE CONJUGATE REPLICA TRANSVERSE WAVE (TIME-REVERSED) IS EMITTED.

Precisely backtracks signal wave input

© T.E. BEARDEN 1997

Published by Lotus® Freelance Graphics®

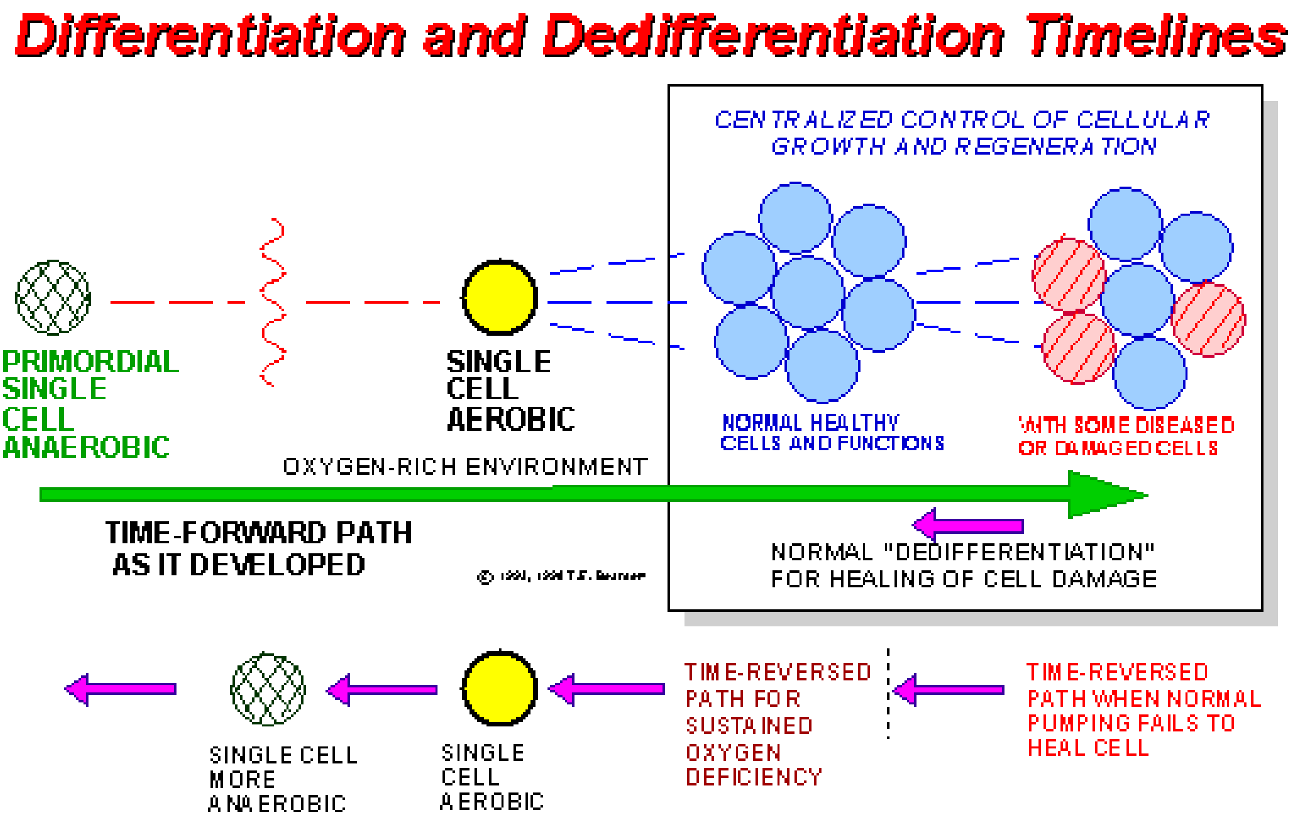
Differentiation and Dedifferentiation Timelines
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
OXYGEN-RICH ENVIRONMENT
TIME-FORWARD PATH AS IT DEVELOPED
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

Published by Lotus® Freelance Graphics®

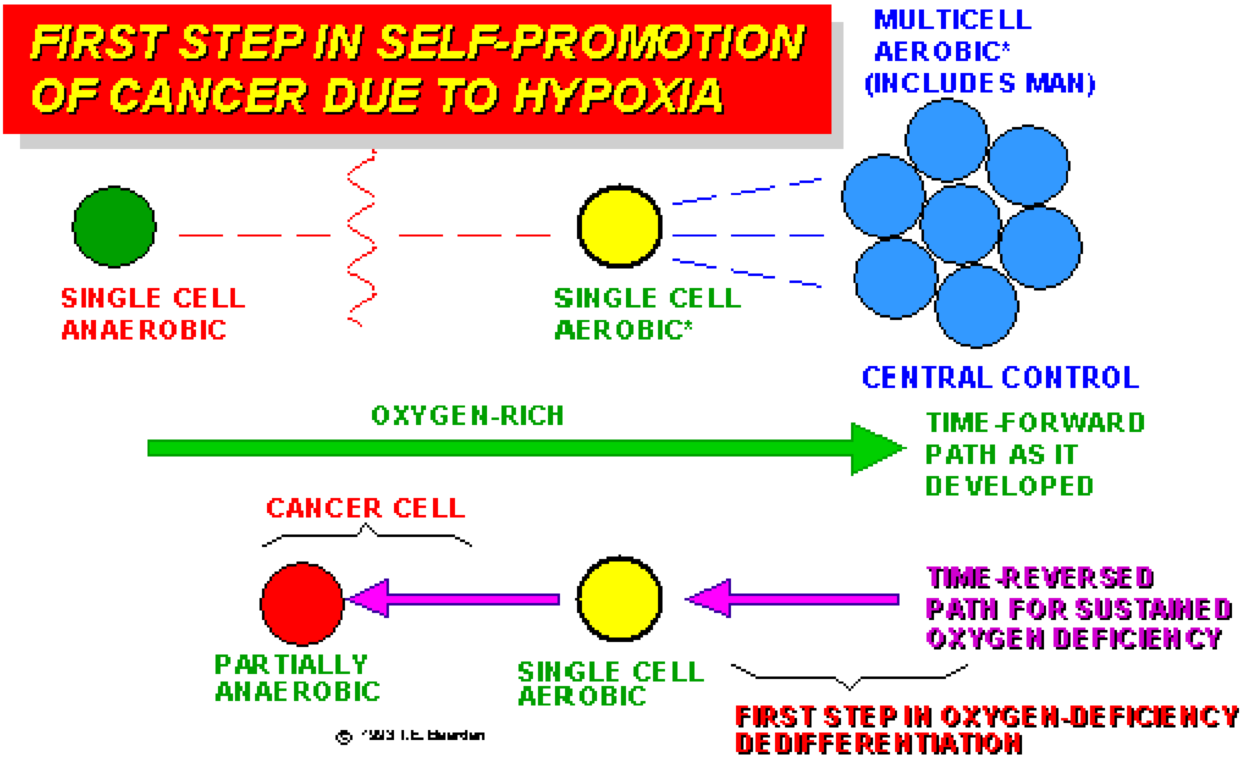
FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA
MULTICELL AEROBIC* (INCLUDES MAN)
SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC*
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD PATH AS IT DEVELOPED
CANCER CELL
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
PARTIALLY ANAEROBIC
SINGLE CELL AEROBIC
FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION
© 2003 T.E. Bearden

Published by Lotus® Freelance Graphics®

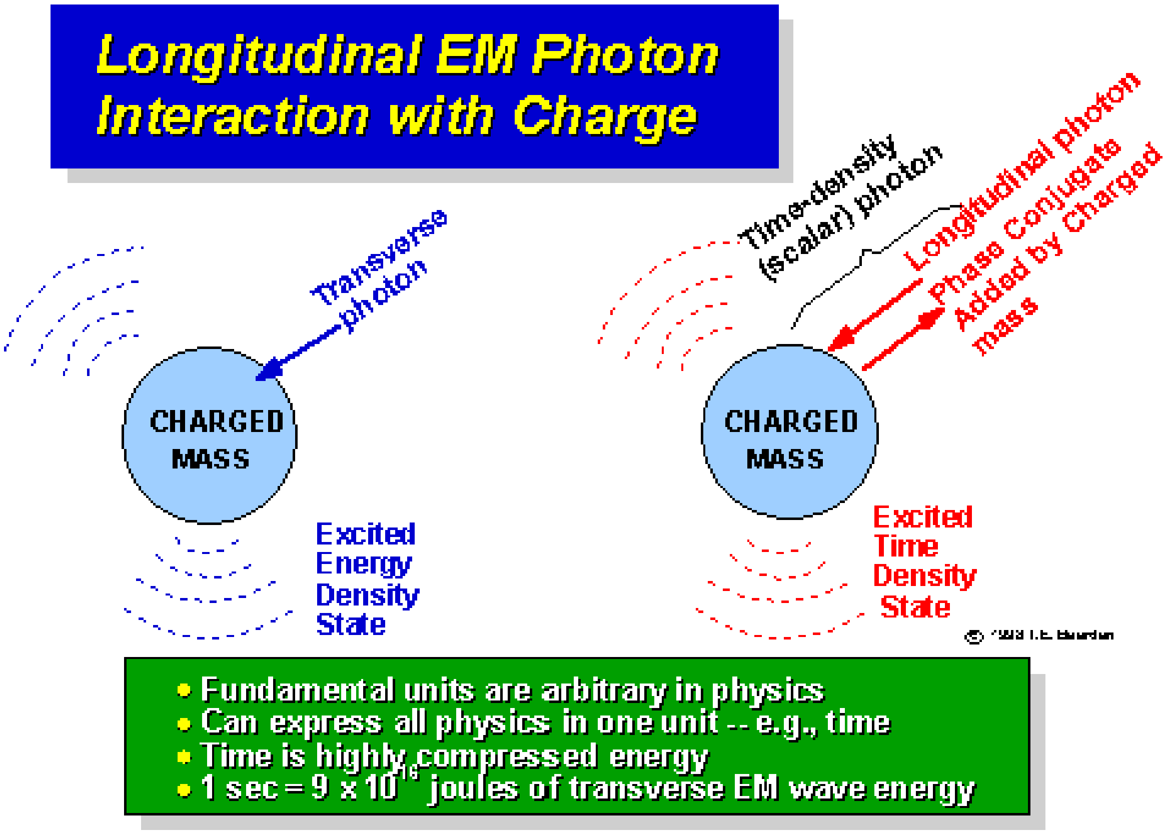

Longitudinal EM Photon Interaction with Charge

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State
© 1993 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10^{16} joules of transverse EM wave energy

Published by Lotus® Freelance Graphics®

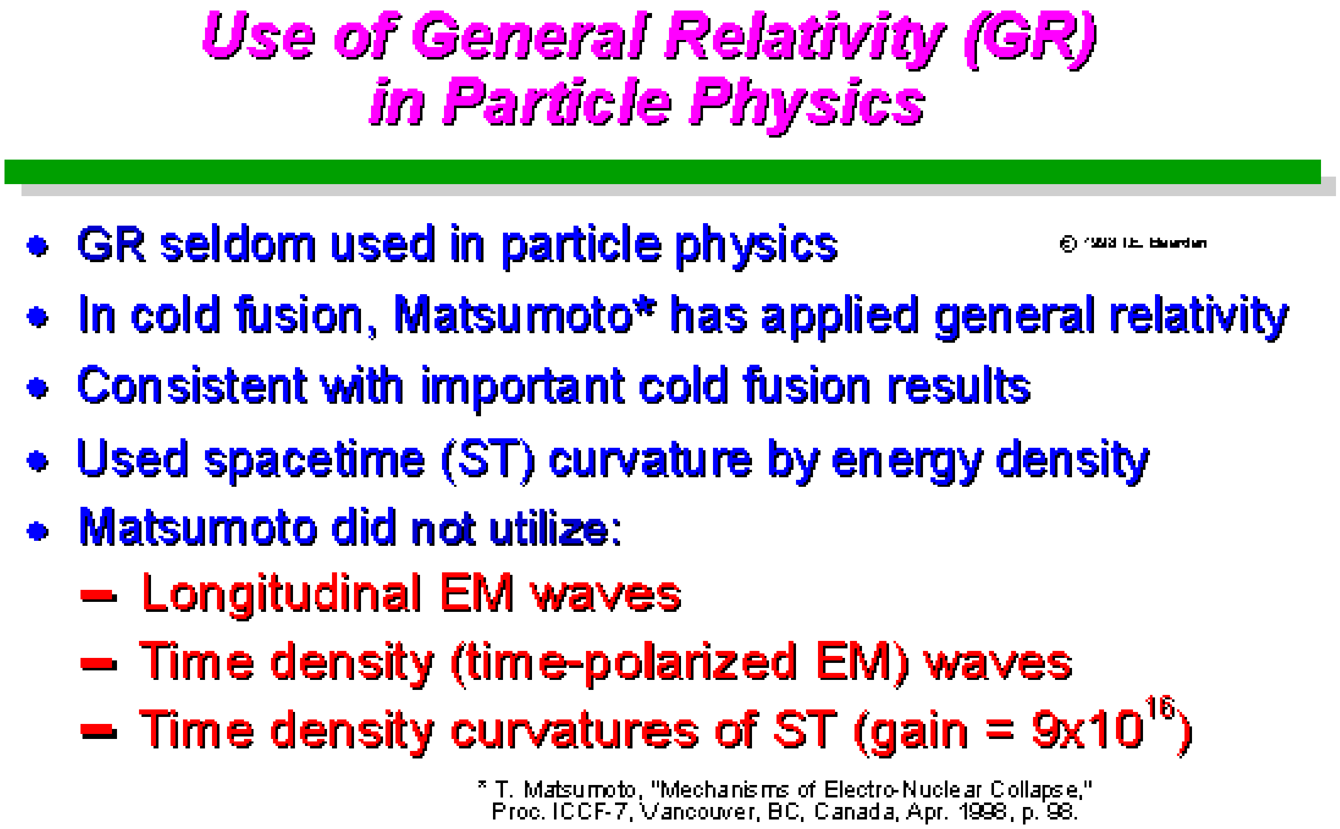

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density (time-polarized EM) waves
  - Time density curvatures of ST (gain = $9x10^{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98.

Published by Lotus® Freelance Graphics®

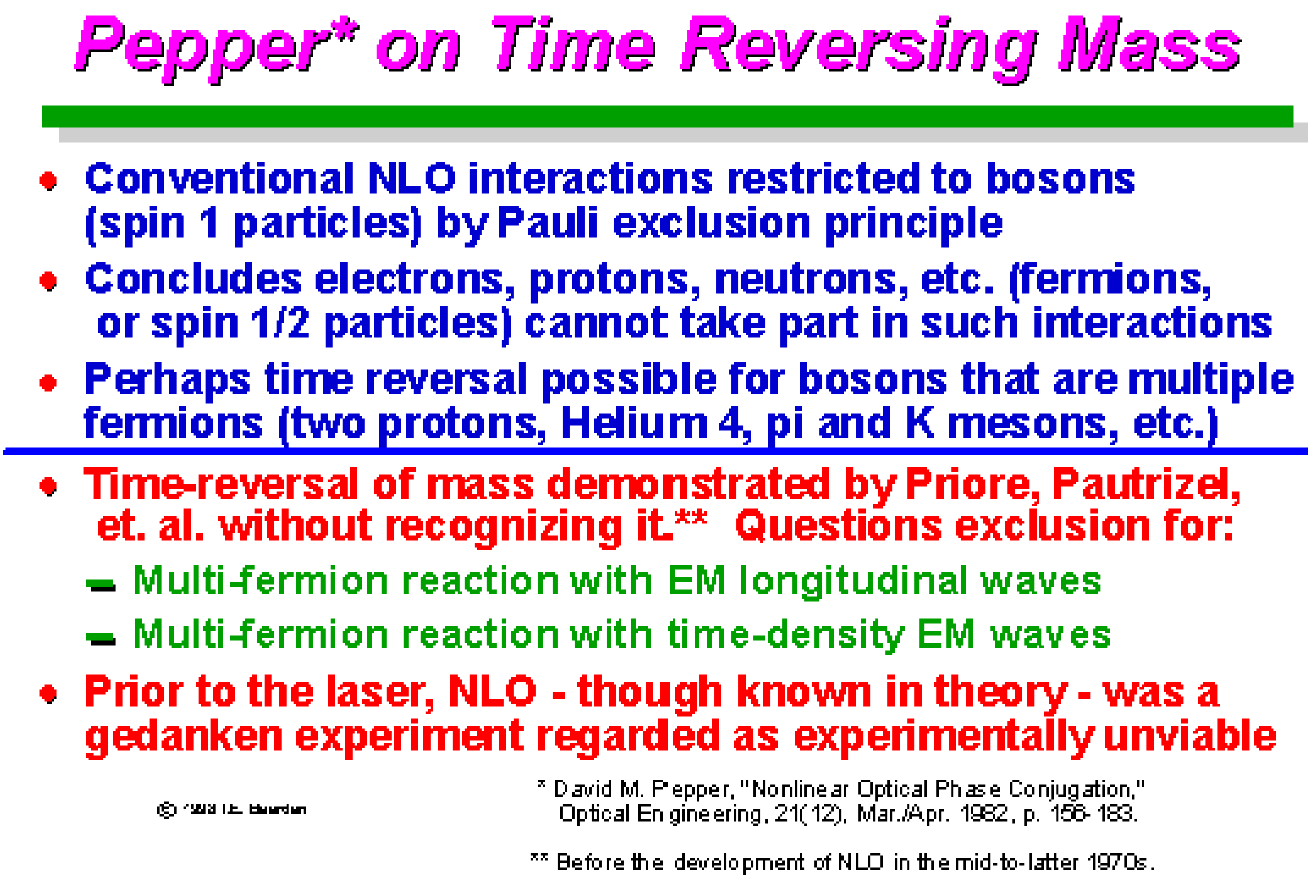

Pepper* on Time Reversing Mass

- Conventional NLO interactions restricted to bosons (spin 1 particles) by Pauli exclusion principle
- Concludes electrons, protons, neutrons, etc. (fermions, or spin 1/2 particles) cannot take part in such interactions
- Perhaps time reversal possible for bosons that are multiple fermions (two protons, Helium 4, pi and K mesons, etc.)
- Time-reversal of mass demonstrated by Priore, Pautrizel, et. al. without recognizing it.** Questions exclusion for:
  - Multi-fermion reaction with EM longitudinal waves
  - Multi-fermion reaction with time-density EM waves
- Prior to the laser, NLO - though known in theory - was a gedanken experiment regarded as experimentally unviable

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

** Before the development of NLO in the mid-to-latter 1970s.

Published by Lotus® Freelance Graphics®

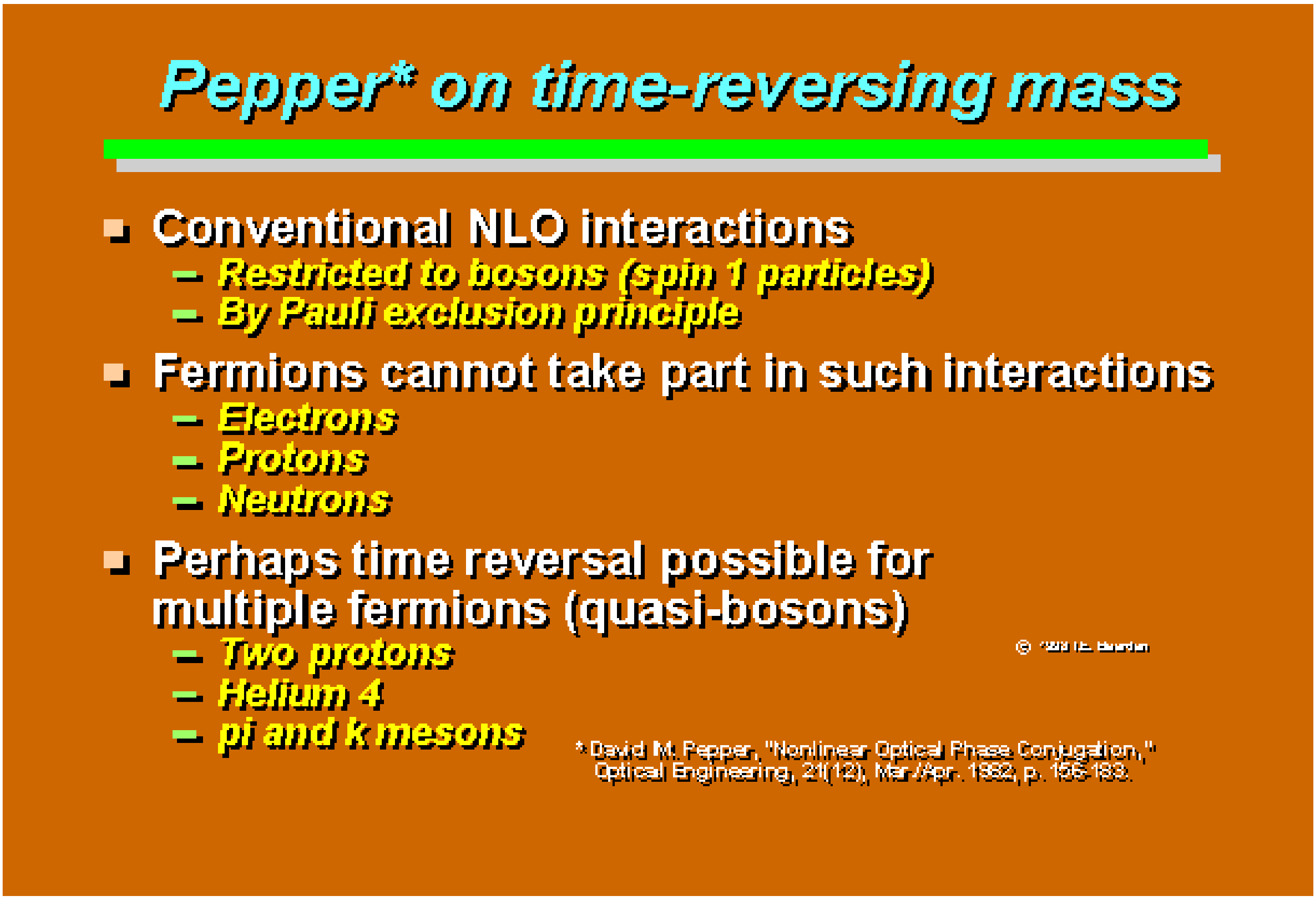

Pepper* on time-reversing mass

- Conventional NLO interactions
  - *Restricted to bosons (spin 1 particles)*
  - *By Pauli exclusion principle*
- Fermions cannot take part in such interactions
  - *Electrons*
  - *Protons*
  - *Neutrons*
- Perhaps time reversal possible for multiple fermions (quasi-bosons)
  - *Two protons*
  - *Helium 4*
  - *pi and k mesons*

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

Published by Lotus® Freelance Graphics®

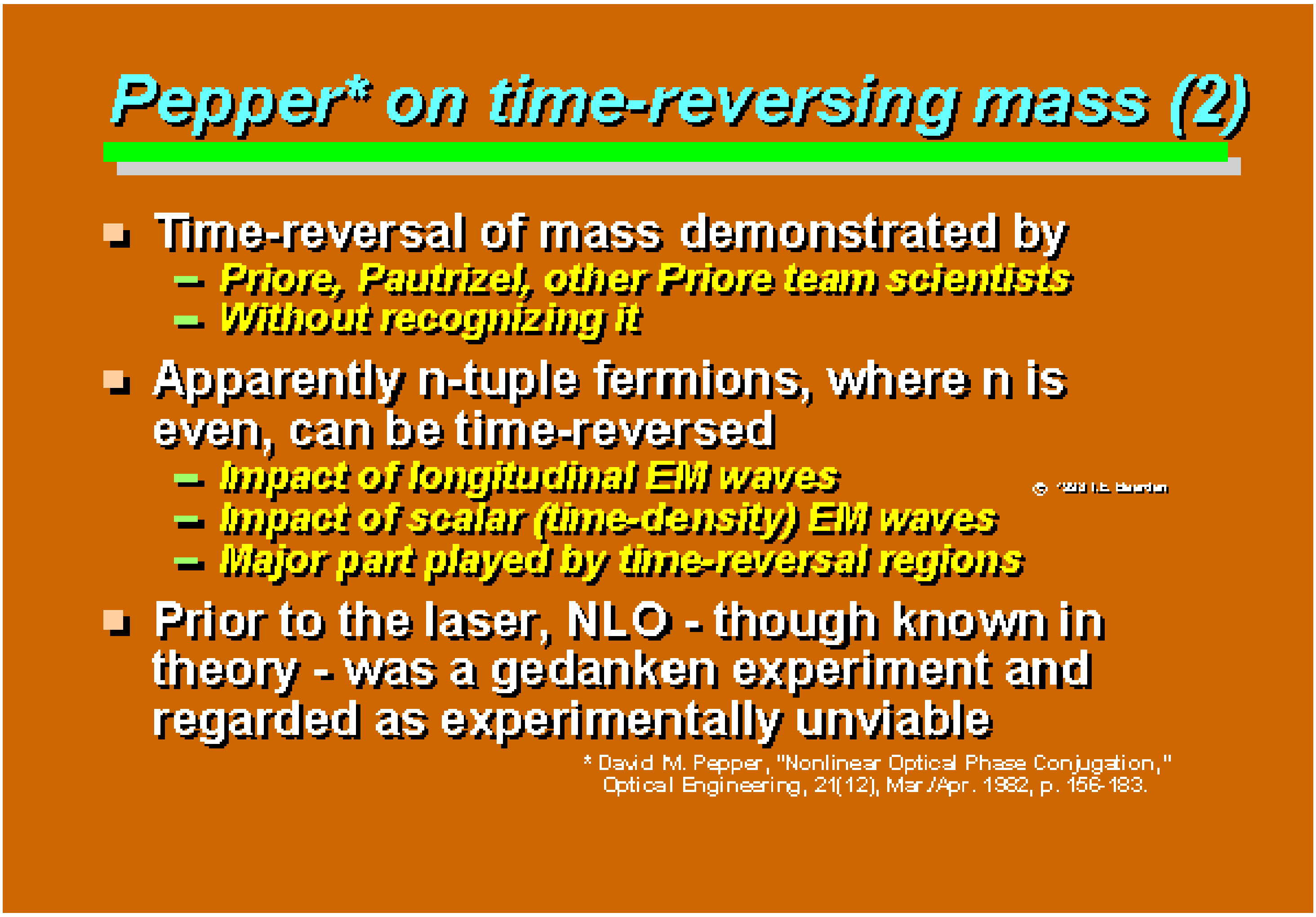

Pepper* on time-reversing mass (2)

- Time-reversal of mass demonstrated by
  - *Priore, Pautrizel, other Priore team scientists*
  - *Without recognizing it*
- Apparently n-tuple fermions, where n is even, can be time-reversed
  - *Impact of longitudinal EM waves*
  - *Impact of scalar (time-density) EM waves*
  - *Major part played by time-reversal regions*
- Prior to the laser, NLO - though known in theory - was a gedanken experiment and regarded as experimentally unviable

© [illegible] T.E. Bearden

* David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(12), Mar./Apr. 1982, p. 156-183.

Published by Lotus® Freelance Graphics®

Chung's Carbon Filament Negative Resistor

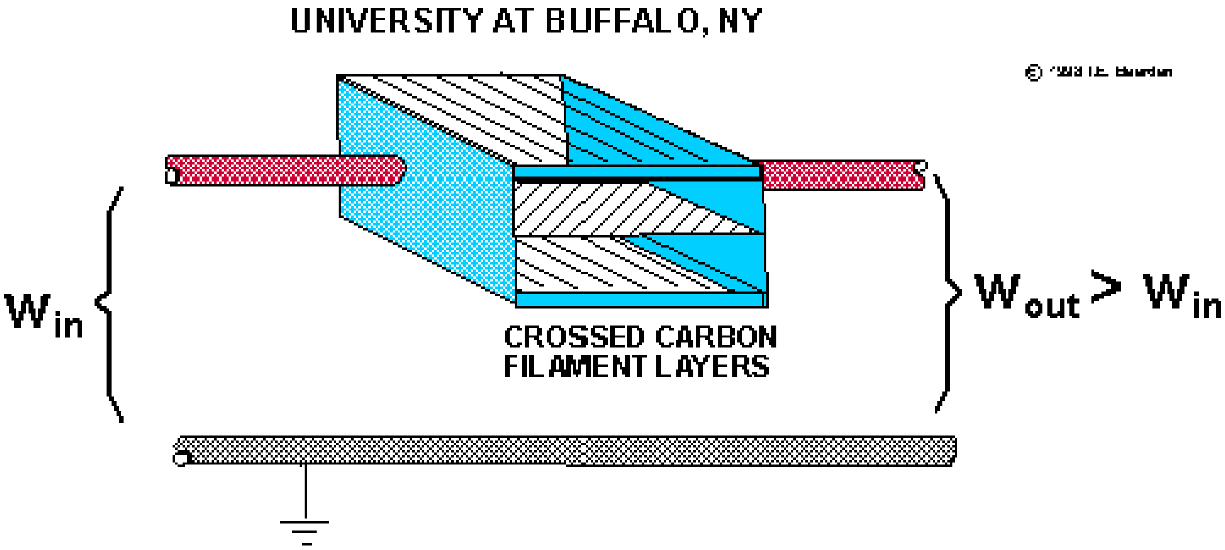
UNIVERSITY AT BUFFALO, NY
© [illegible] T.E. Bearden
CROSSED CARBON FILAMENT LAYERS
W_{in}
$W_{out} > W_{in}$

Published by Lotus® Freelance Graphics®

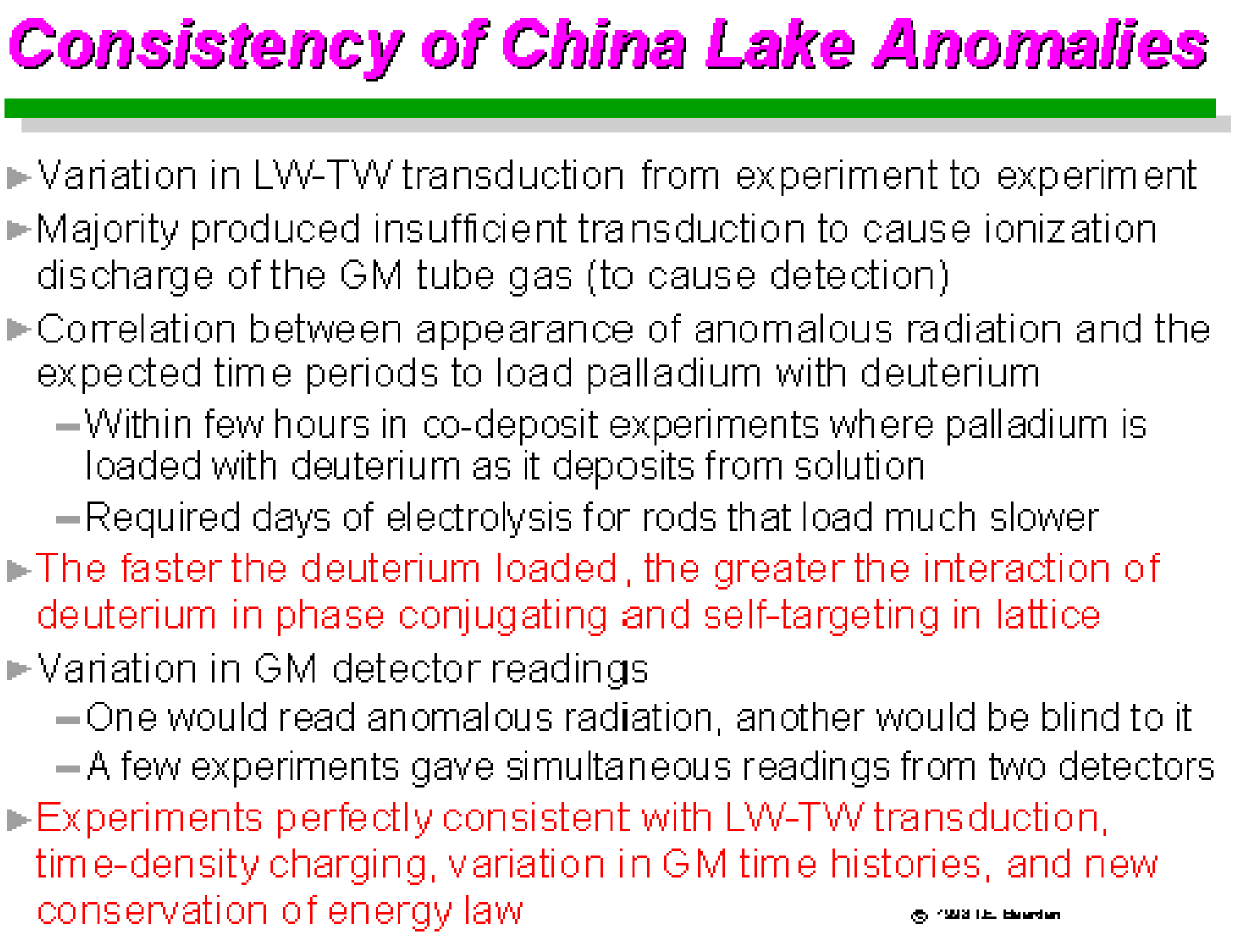

Consistency of China Lake Anomalies

- Variation in LW-TW transduction from experiment to experiment
- Majority produced insufficient transduction to cause ionization discharge of the GM tube gas (to cause detection)
- Correlation between appearance of anomalous radiation and the expected time periods to load palladium with deuterium
  - Within few hours in co-deposit experiments where palladium is loaded with deuterium as it deposits from solution
  - Required days of electrolysis for rods that load much slower
- The faster the deuterium loaded, the greater the interaction of deuterium in phase conjugating and self-targeting in lattice
- Variation in GM detector readings
  - One would read anomalous radiation, another would be blind to it
  - A few experiments gave simultaneous readings from two detectors
- Experiments perfectly consistent with LW-TW transduction, time-density charging, variation in GM time histories, and new conservation of energy law

Published by Lotus® Freelance Graphics®

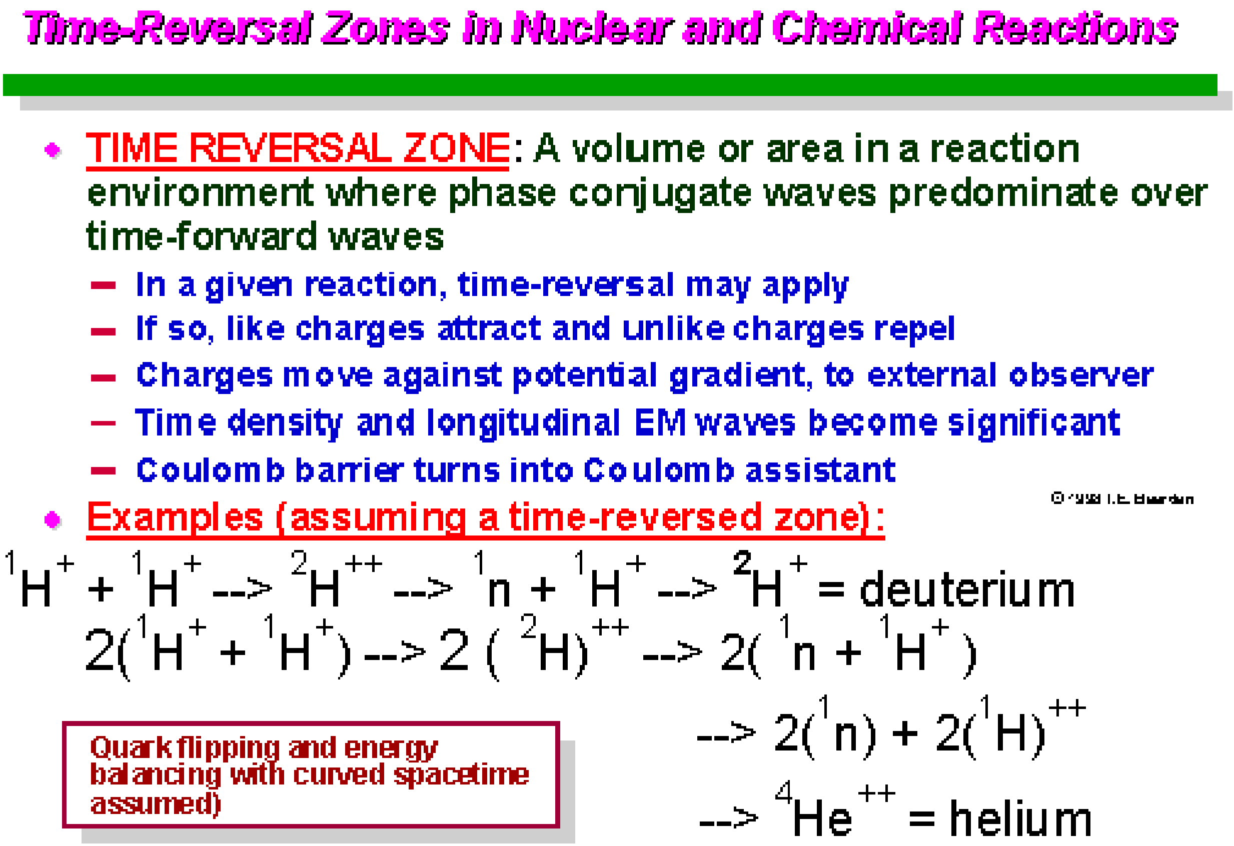

Time-Reversal Zones in Nuclear and Chemical Reactions

- TIME REVERSAL ZONE: A volume or area in a reaction environment where phase conjugate waves predominate over time-forward waves
  - In a given reaction, time-reversal may apply
  - If so, like charges attract and unlike charges repel
  - Charges move against potential gradient, to external observer
  - Time density and longitudinal EM waves become significant
  - Coulomb barrier turns into Coulomb assistant
- Examples (assuming a time-reversed zone):

© 1998 T.E. Bearden

$${}^{1}H^{+} + {}^{1}H^{+} \text{ --> } {}^{2}H^{++} \text{ --> } {}^{1}n + {}^{1}H^{+} \text{ --> } {}^{2}H^{+} = \text{deuterium}$$

$$2({}^{1}H^{+} + {}^{1}H^{+}) \text{ --> } 2\,({}^{2}H)^{++} \text{ --> } 2({}^{1}n + {}^{1}H^{+})$$

$$\text{ --> } 2({}^{1}n) + 2({}^{1}H)^{++}$$

$$\text{ --> } {}^{4}He^{++} = \text{helium}$$

Quark flipping and energy balancing with curved spacetime assumed)

Published by Lotus® Freelance Graphics®

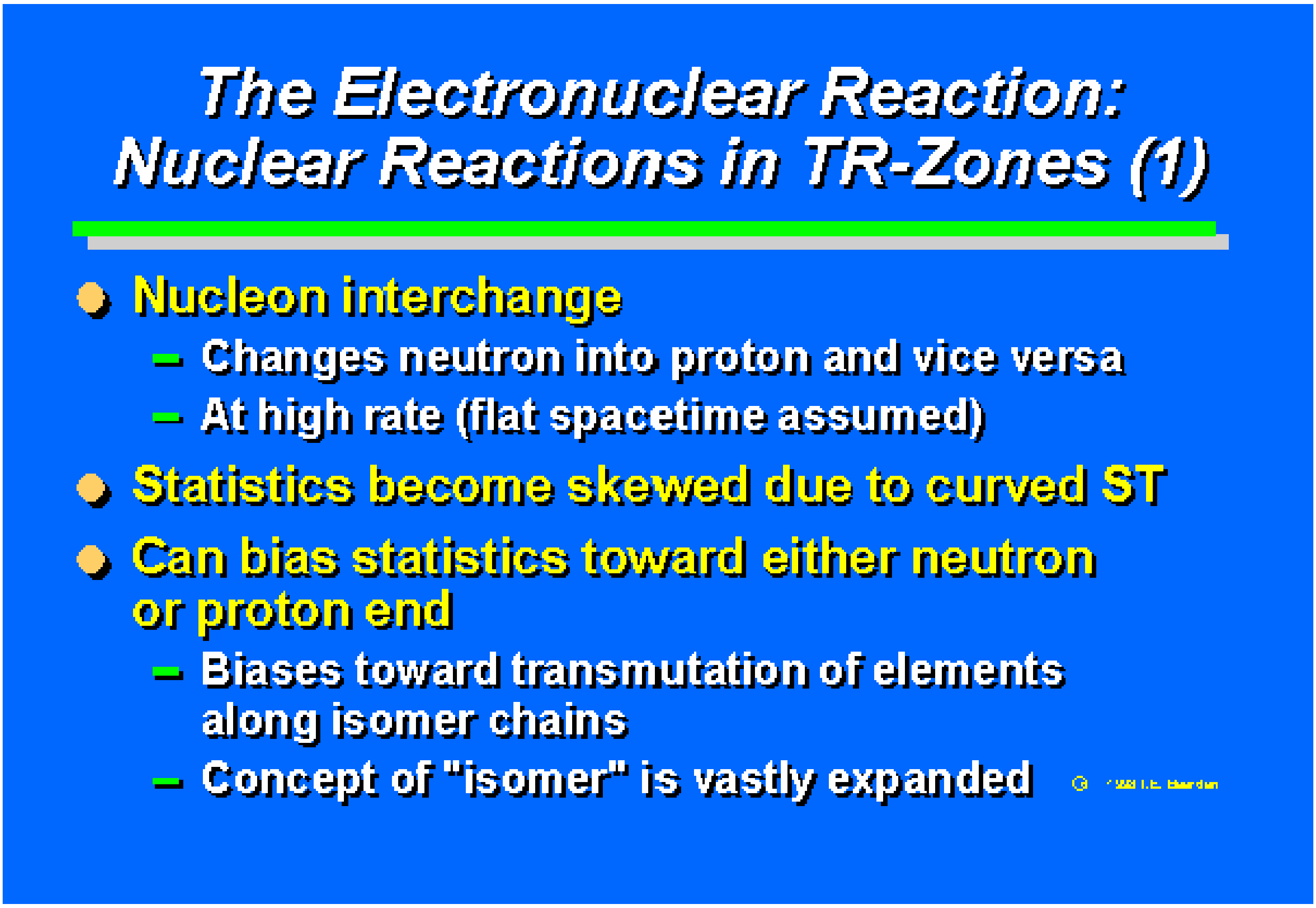

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- Nucleon interchange
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- Statistics become skewed due to curved ST
- Can bias statistics toward either neutron or proton end
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

© [illegible] T.E. Bearden

Published by Lotus® Freelance Graphics®

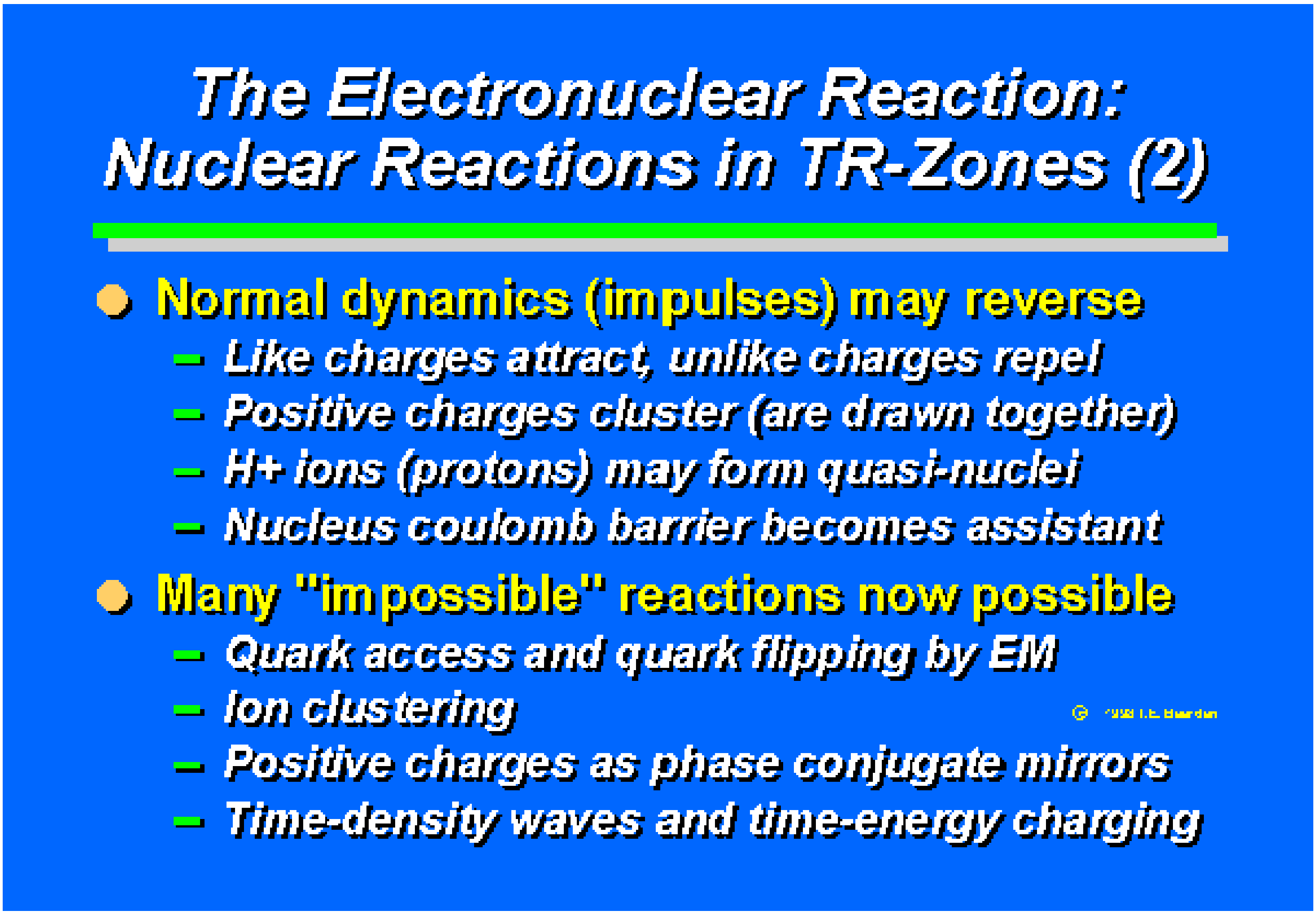

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- Normal dynamics (impulses) may reverse
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- Many "impossible" reactions now possible
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

© 1998 T.E. Bearden

Published by Lotus® Freelance Graphics®

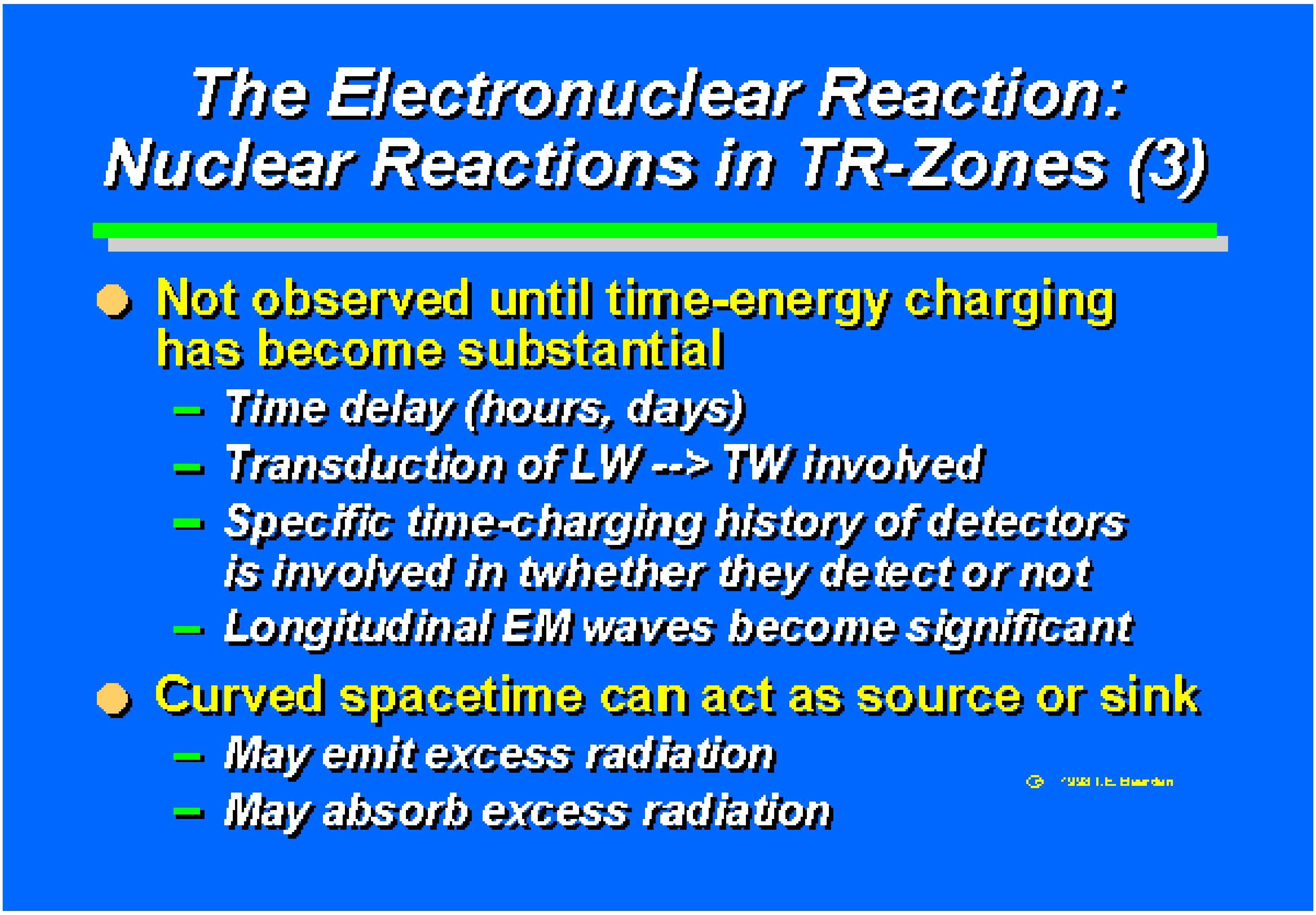

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- Not observed until time-energy charging has become substantial
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- Curved spacetime can act as source or sink
  - *May emit excess radiation*
  - *May absorb excess radiation*

© 1998 T.E. Bearden

Published by Lotus® Freelance Graphics®

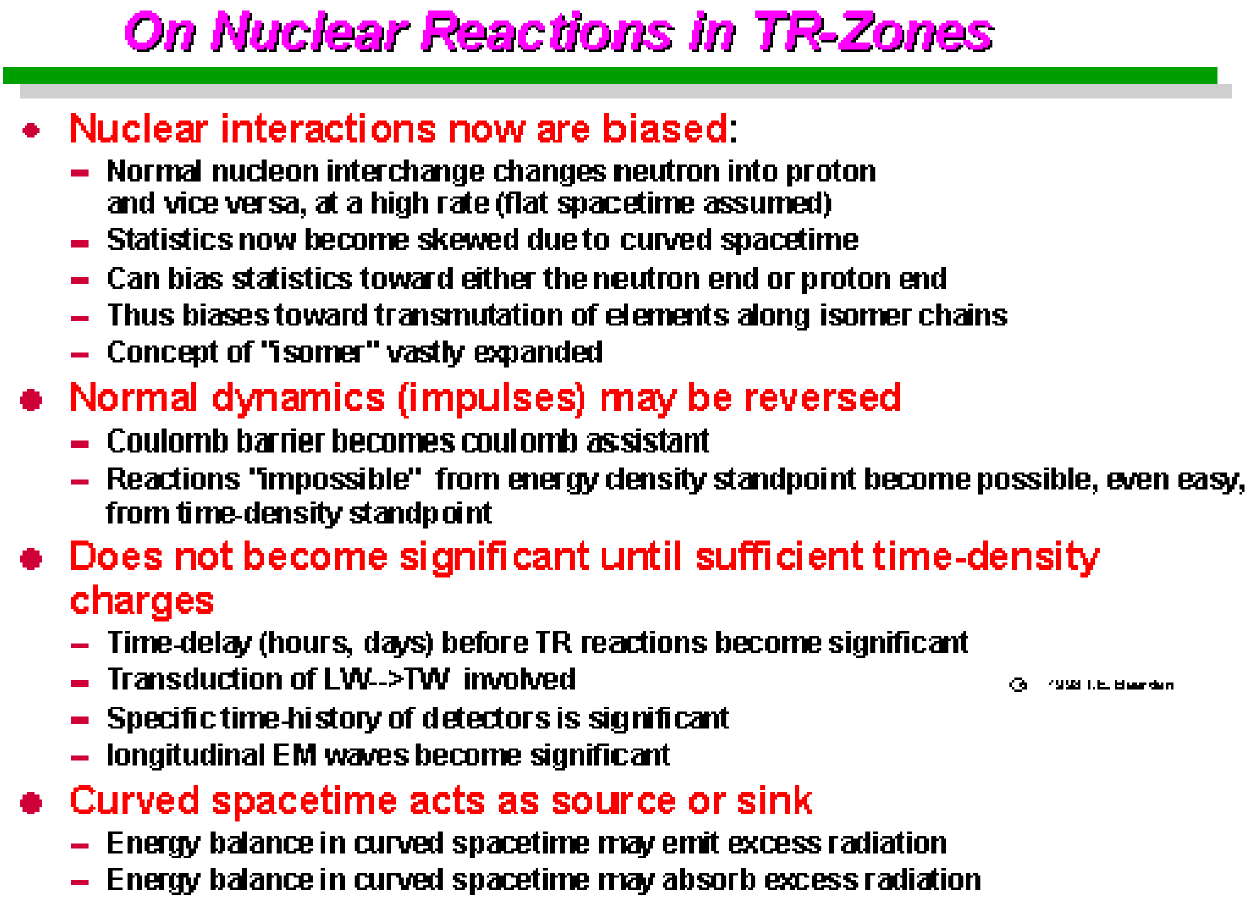

On Nuclear Reactions in TR-Zones

- Nuclear interactions now are biased:
  - Normal nucleon interchange changes neutron into proton and vice versa, at a high rate (flat spacetime assumed)
  - Statistics now become skewed due to curved spacetime
  - Can bias statistics toward either the neutron end or proton end
  - Thus biases toward transmutation of elements along isomer chains
  - Concept of "isomer" vastly expanded
- Normal dynamics (impulses) may be reversed
  - Coulomb barrier becomes coulomb assistant
  - Reactions "impossible" from energy density standpoint become possible, even easy, from time-density standpoint
- Does not become significant until sufficient time-density charges
  - Time-delay (hours, days) before TR reactions become significant
  - Transduction of LW-->TW involved
  - Specific time-history of detectors is significant
  - longitudinal EM waves become significant
- Curved spacetime acts as source or sink
  - Energy balance in curved spacetime may emit excess radiation
  - Energy balance in curved spacetime may absorb excess radiation

Published by Lotus® Freelance Graphics®

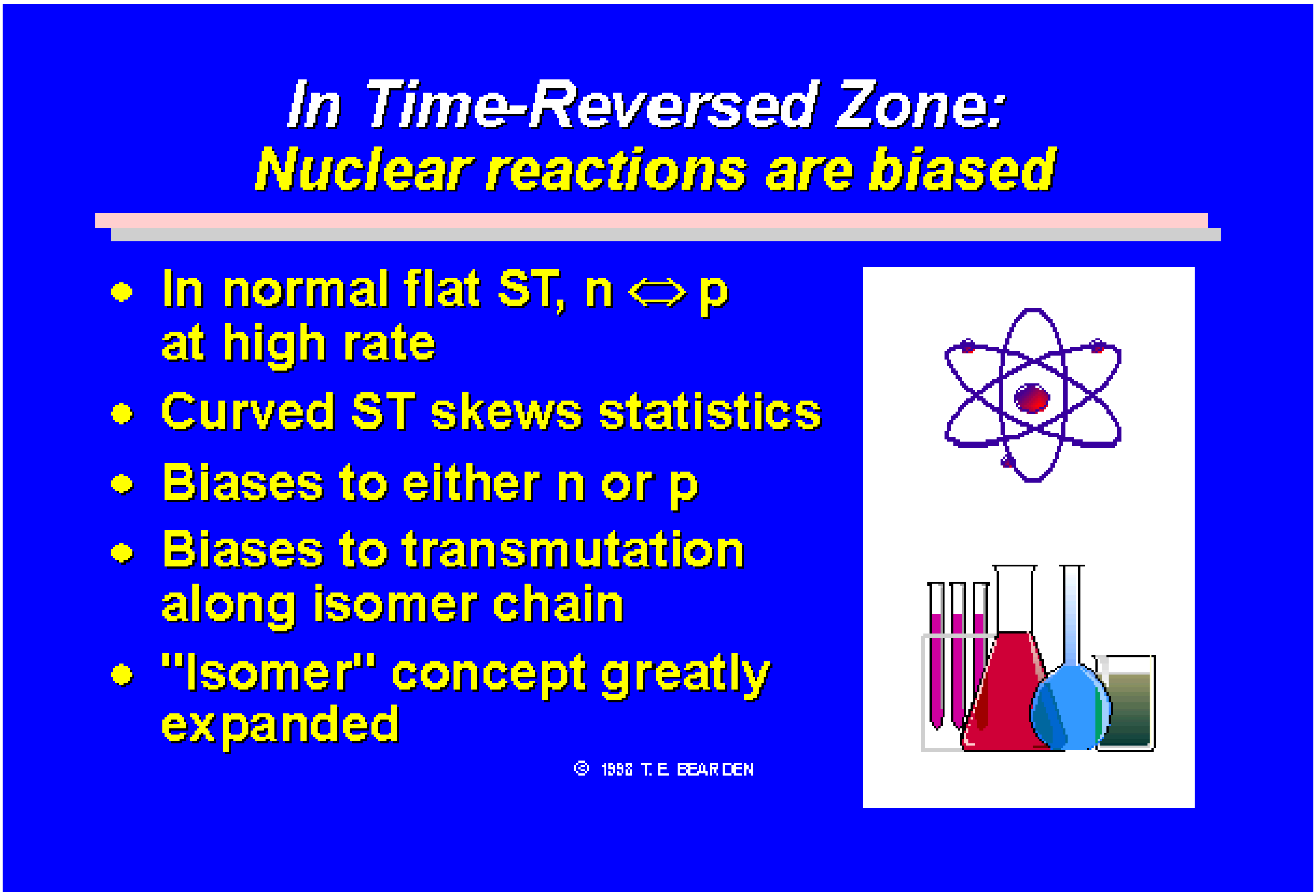

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T. E. BEARDEN

Published by Lotus® Freelance Graphics®

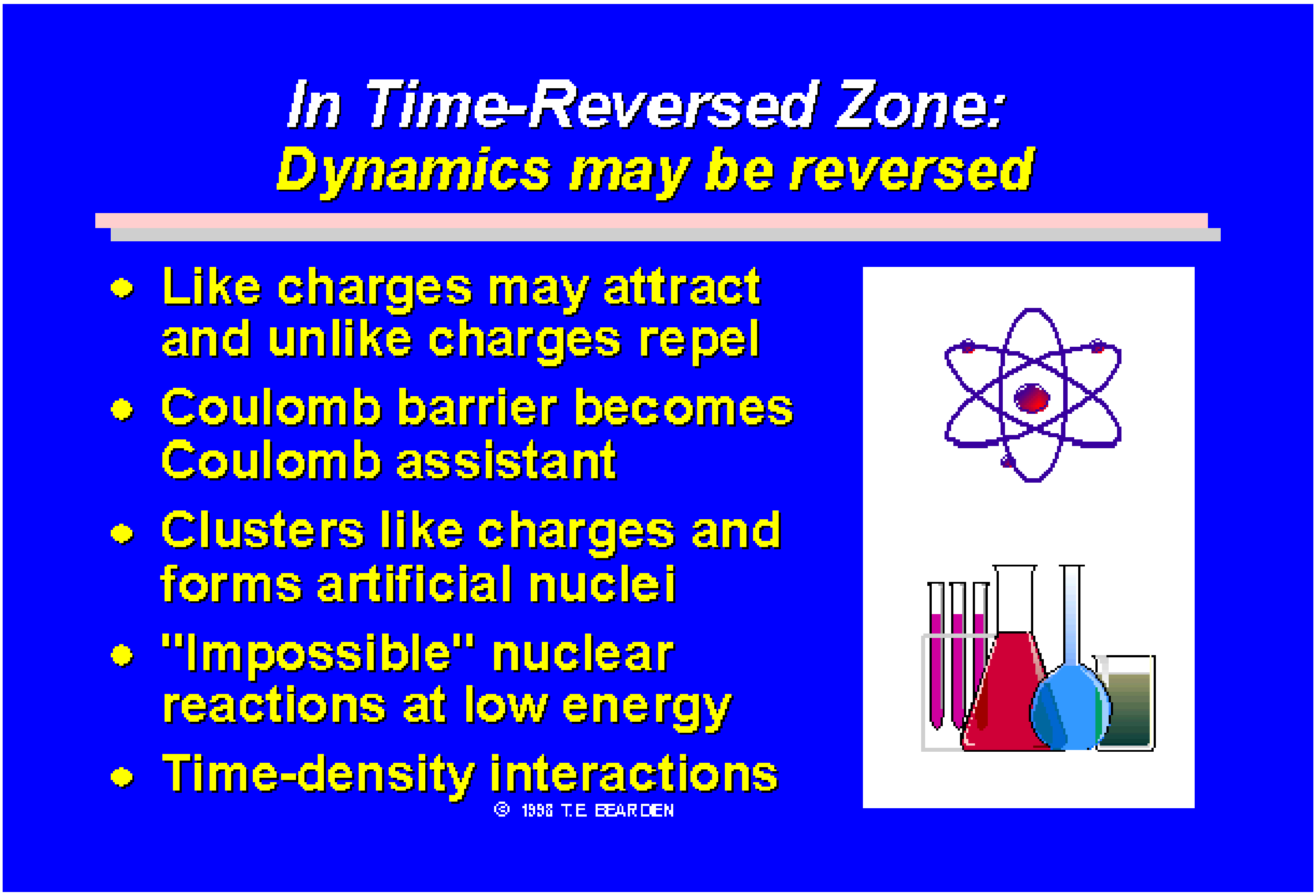

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

© 1998 T.E BEARDEN

Published by Lotus® Freelance Graphics®

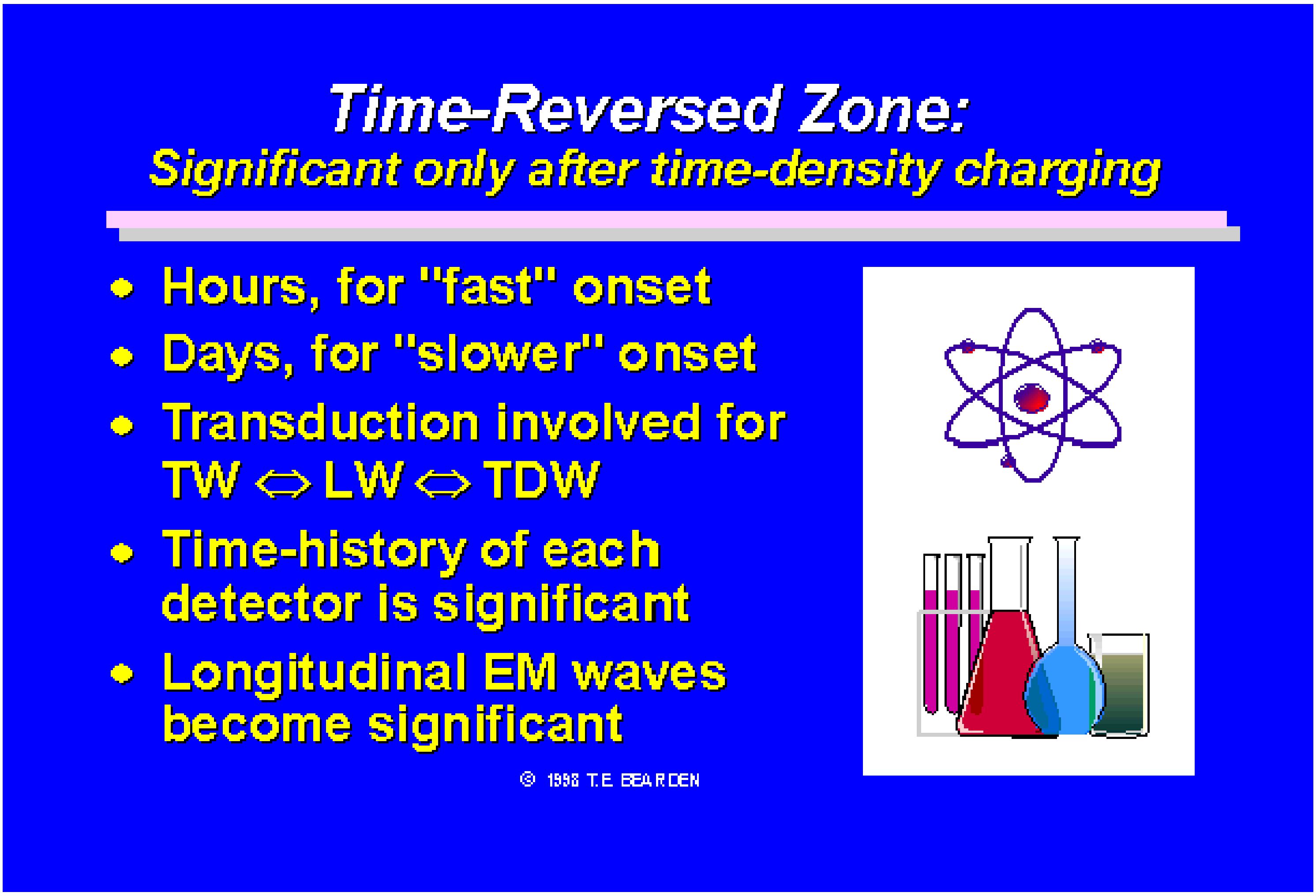

Time-Reversed Zone:
Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

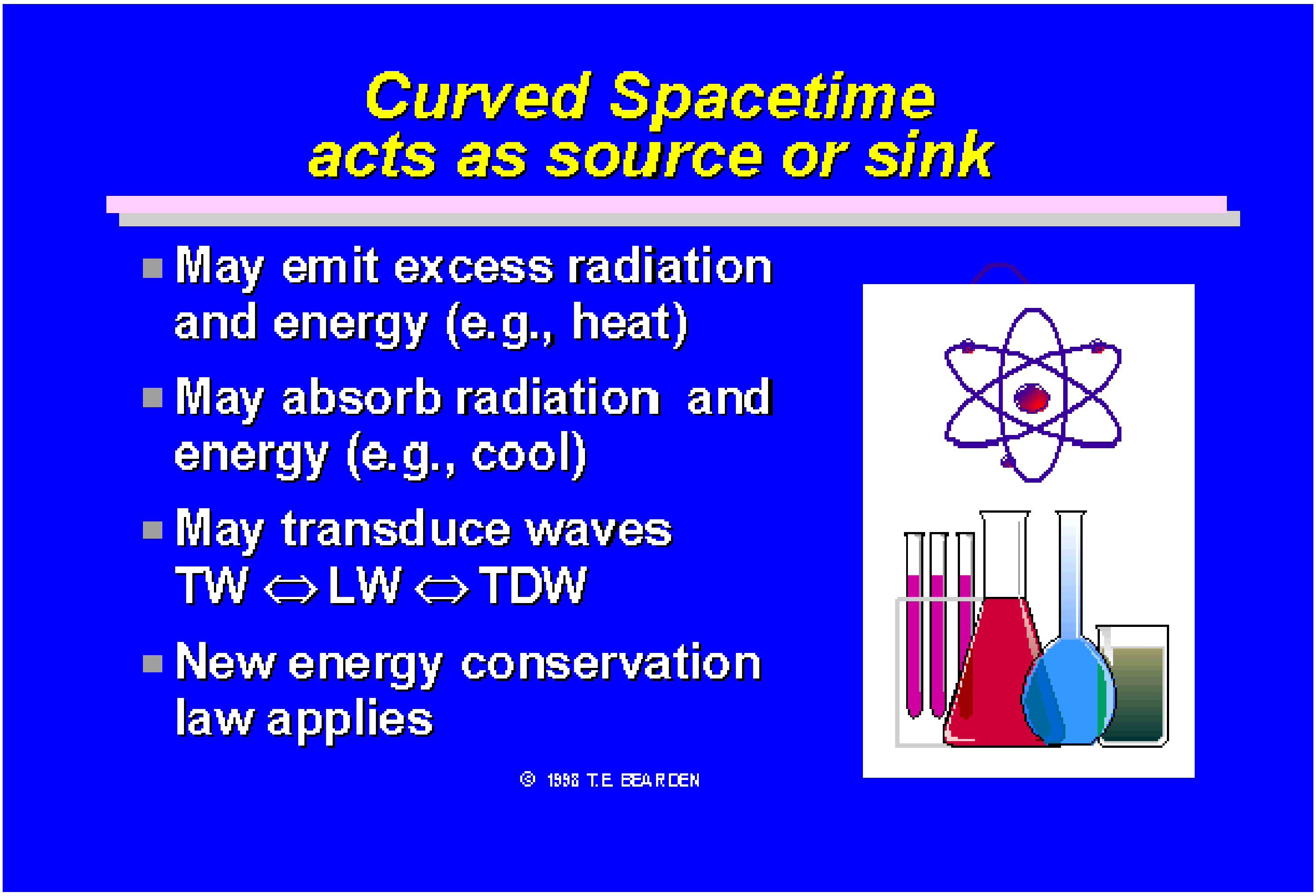

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves
  TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

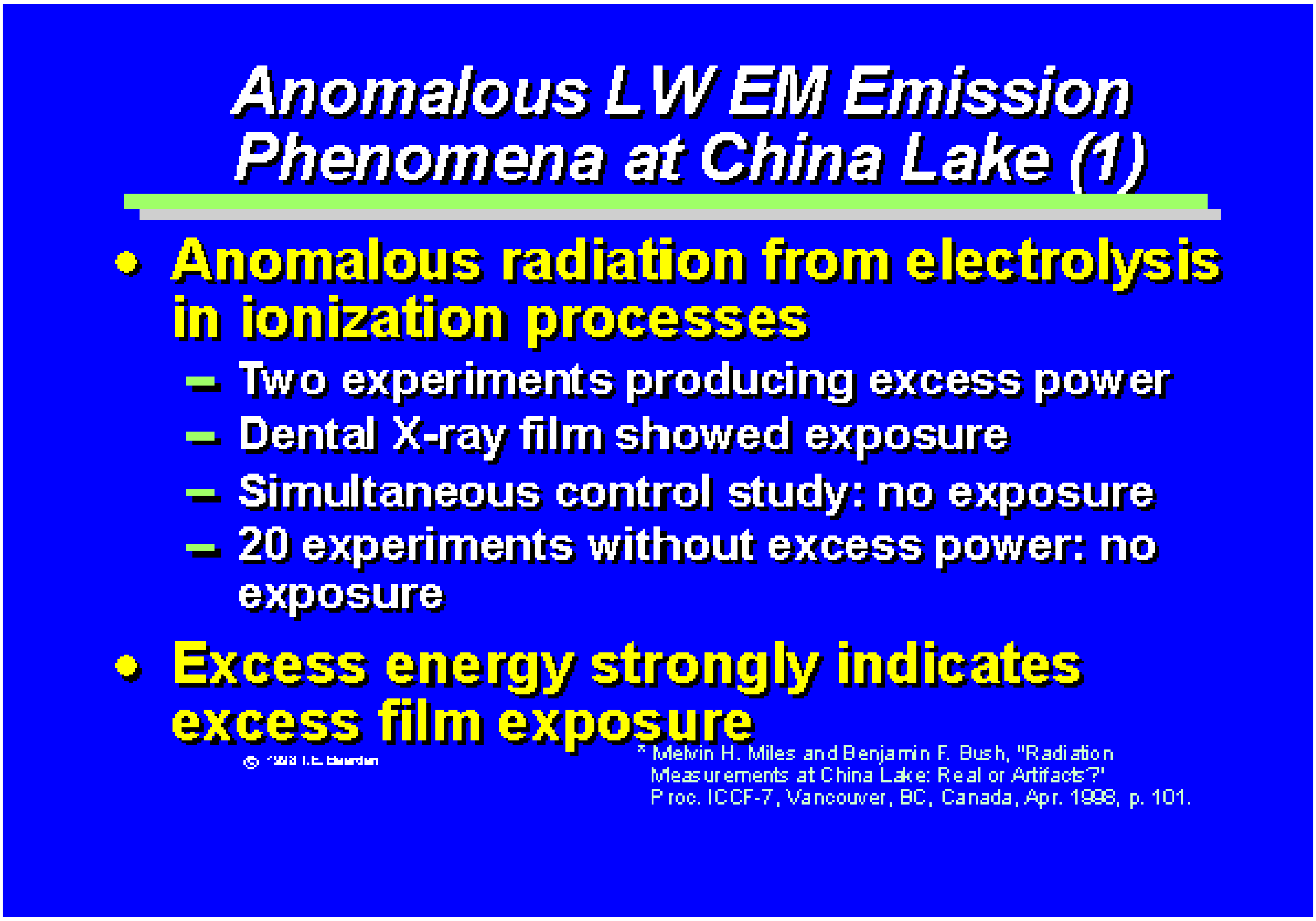

Anomalous LW EM Emission Phenomena at China Lake (1)

- **Anomalous radiation from electrolysis in ionization processes**
  - Two experiments producing excess power
  - Dental X-ray film showed exposure
  - Simultaneous control study: no exposure
  - 20 experiments without excess power: no exposure
- **Excess energy strongly indicates excess film exposure**

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101.

Published by Lotus® Freelance Graphics®

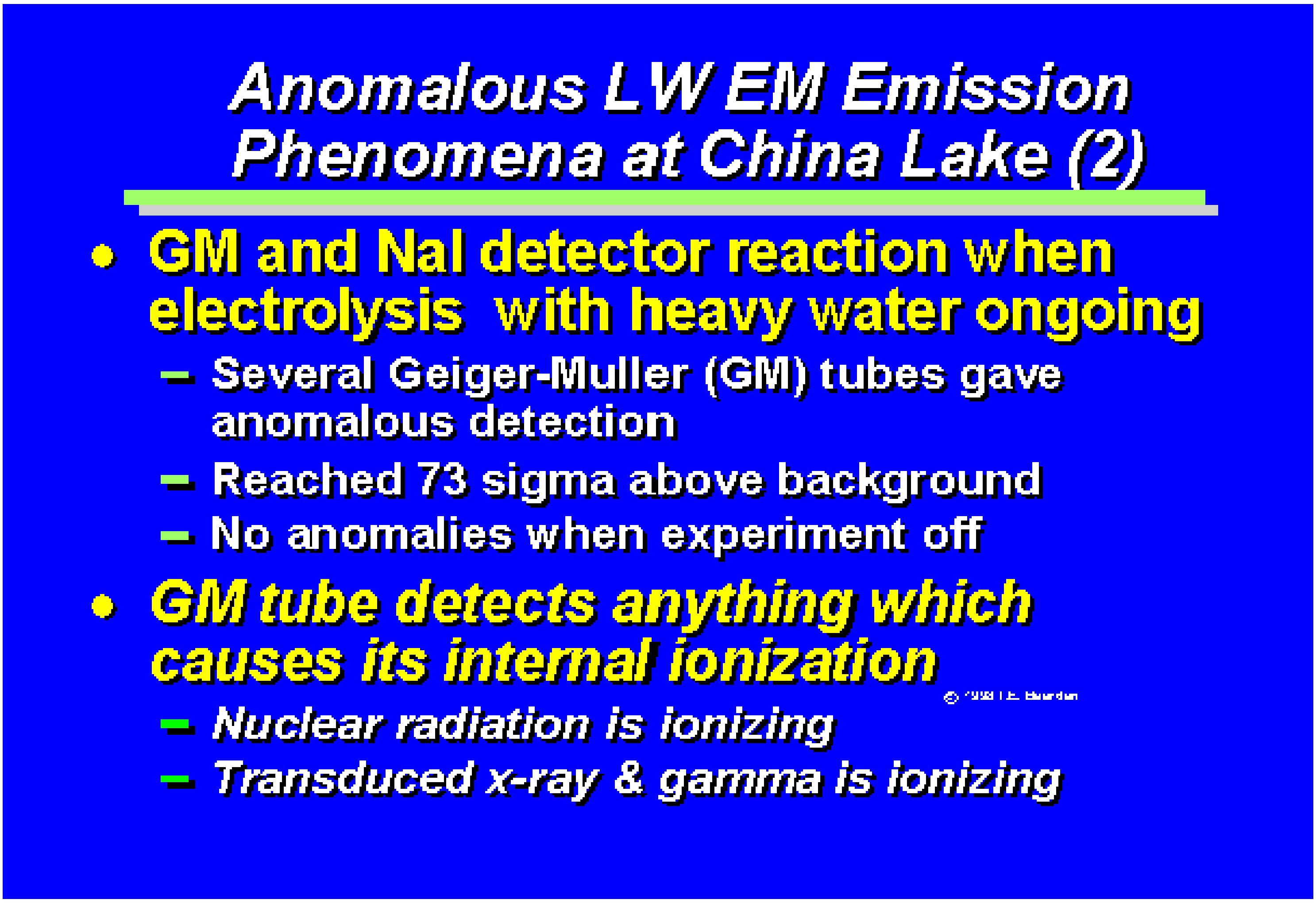

Anomalous LW EM Emission Phenomena at China Lake (2)

- **GM and NaI detector reaction when electrolysis with heavy water ongoing**
  - Several Geiger-Muller (GM) tubes gave anomalous detection
  - Reached 73 sigma above background
  - No anomalies when experiment off
- ***GM tube detects anything which causes its internal ionization***
  - *Nuclear radiation is ionizing*
  - *Transduced x-ray & gamma is ionizing*

Published by Lotus® Freelance Graphics®

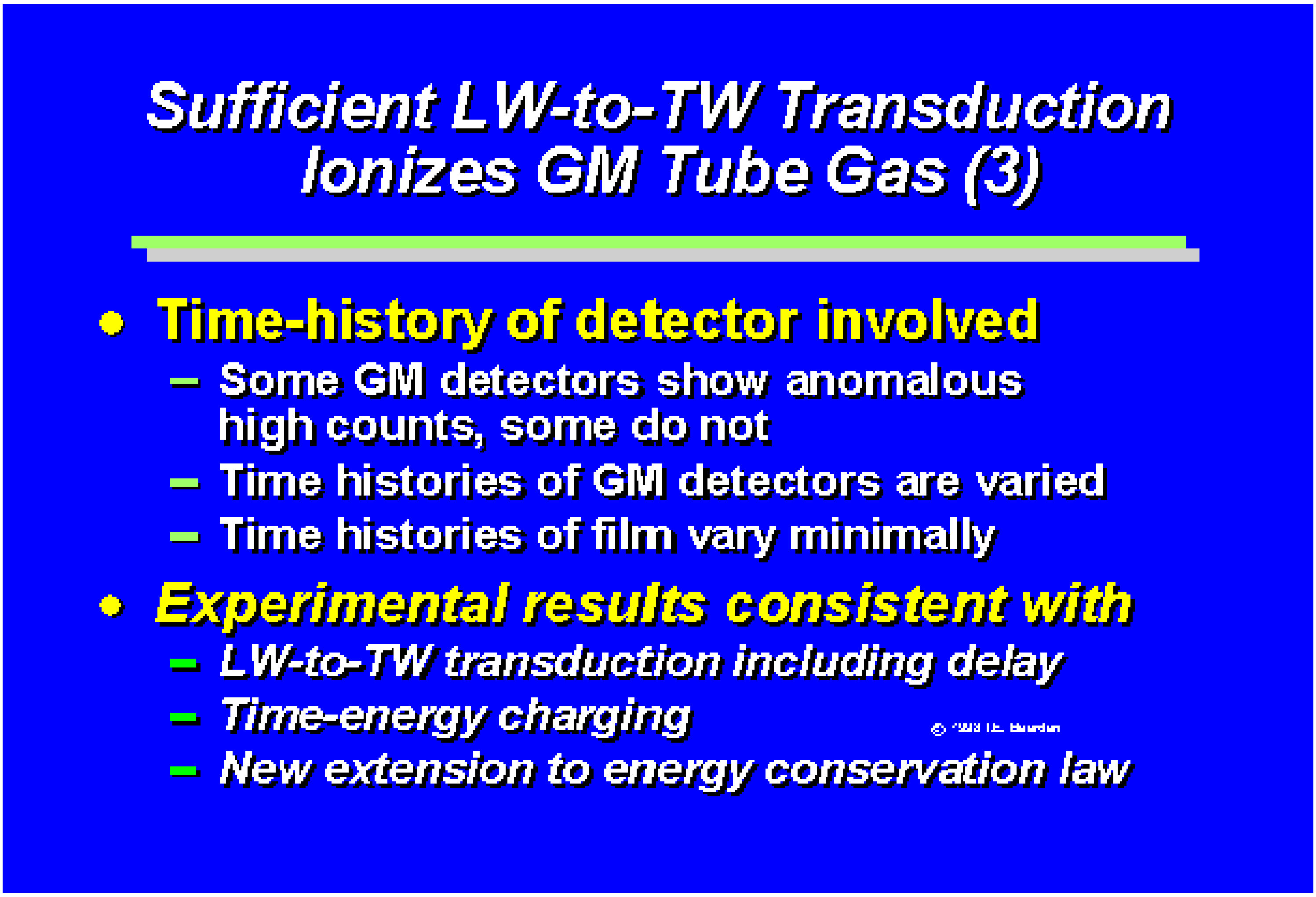

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© 2003 T.E. Bearden

Published by Lotus® Freelance Graphics®

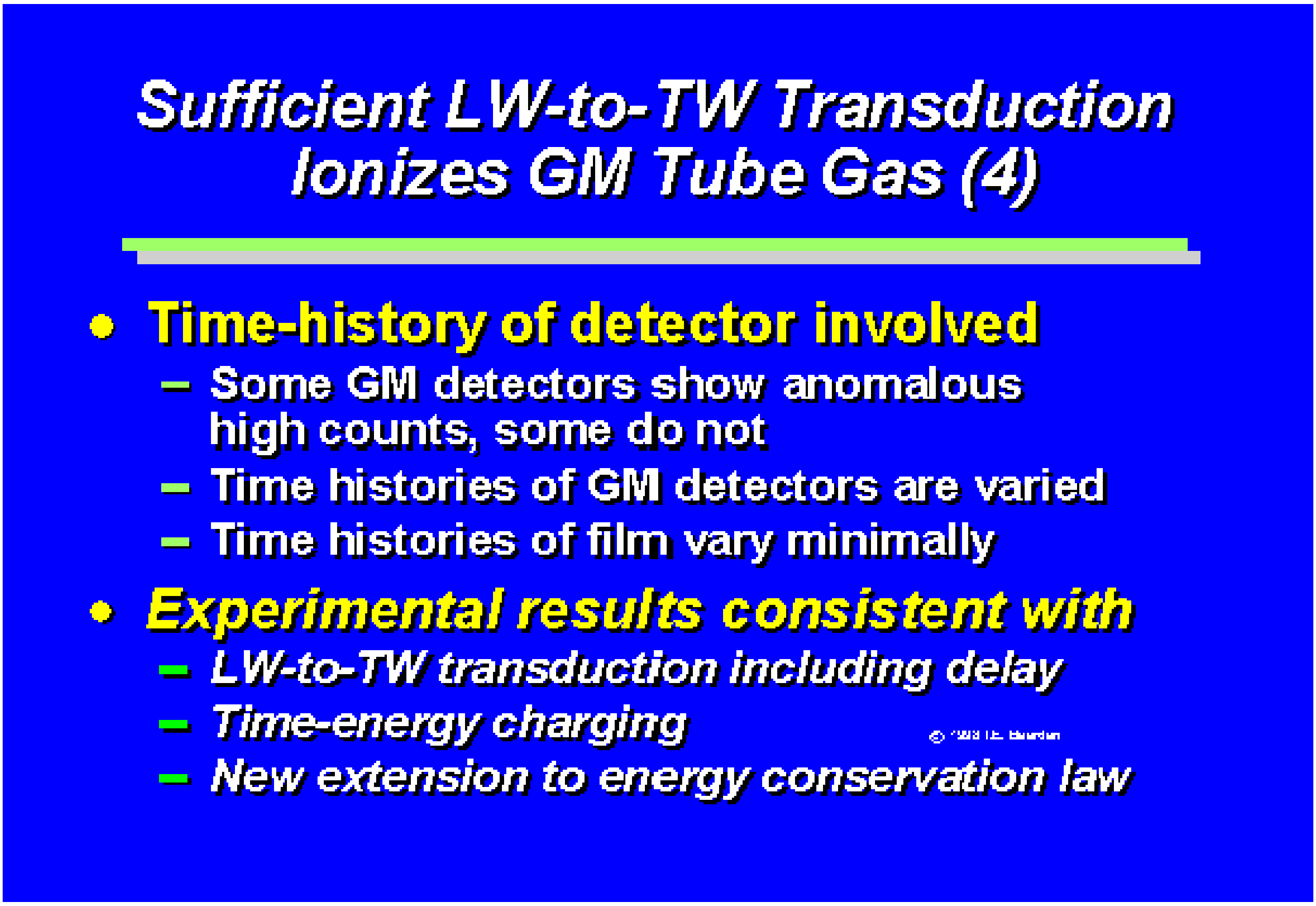

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

Published by Lotus® Freelance Graphics®

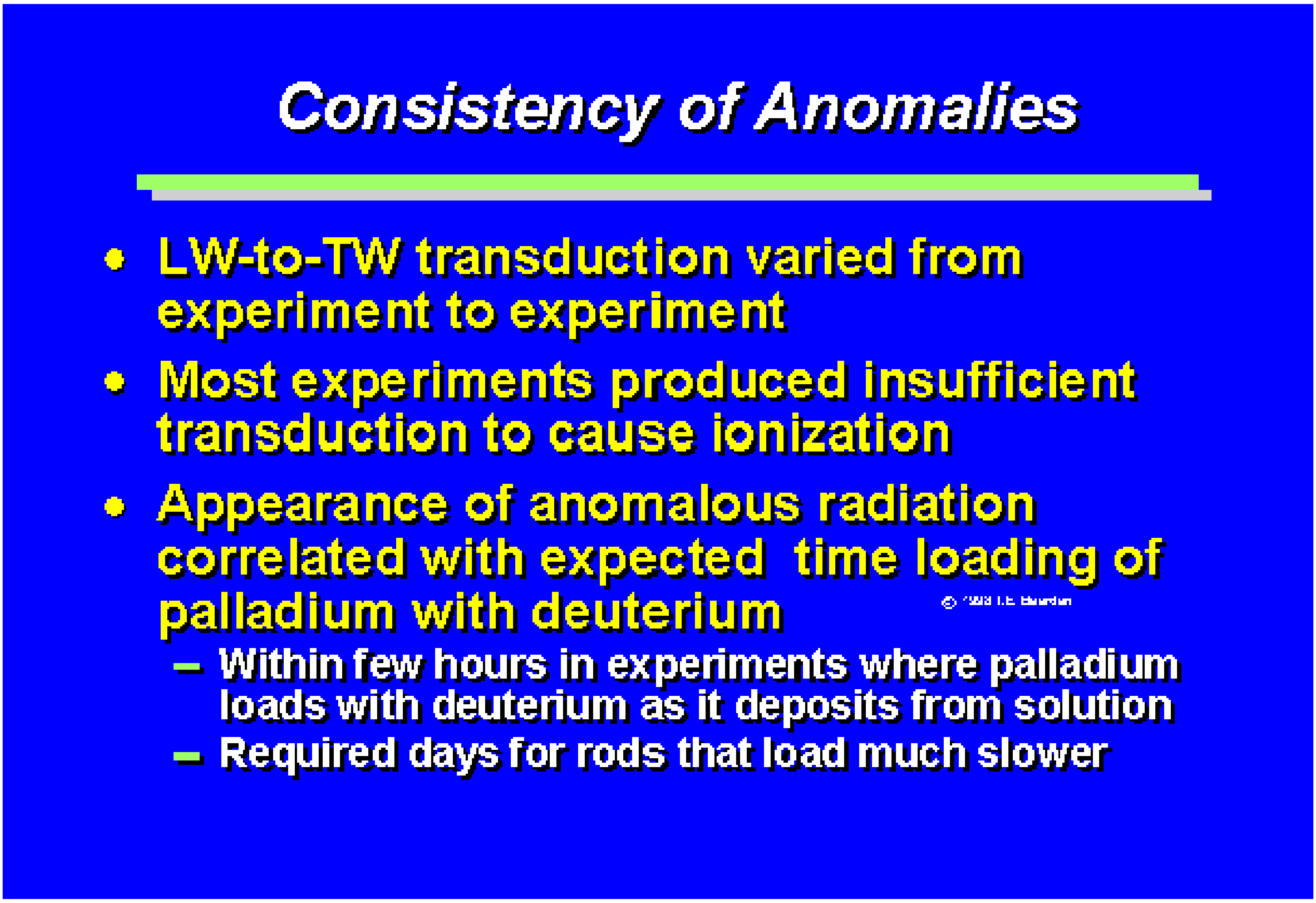

Consistency of Anomalies

- LW-to-TW transduction varied from experiment to experiment
- Most experiments produced insufficient transduction to cause ionization
- Appearance of anomalous radiation correlated with expected time loading of palladium with deuterium
  - Within few hours in experiments where palladium loads with deuterium as it deposits from solution
  - Required days for rods that load much slower

© 1993 T.E. Bearden

Published by Lotus® Freelance Graphics®

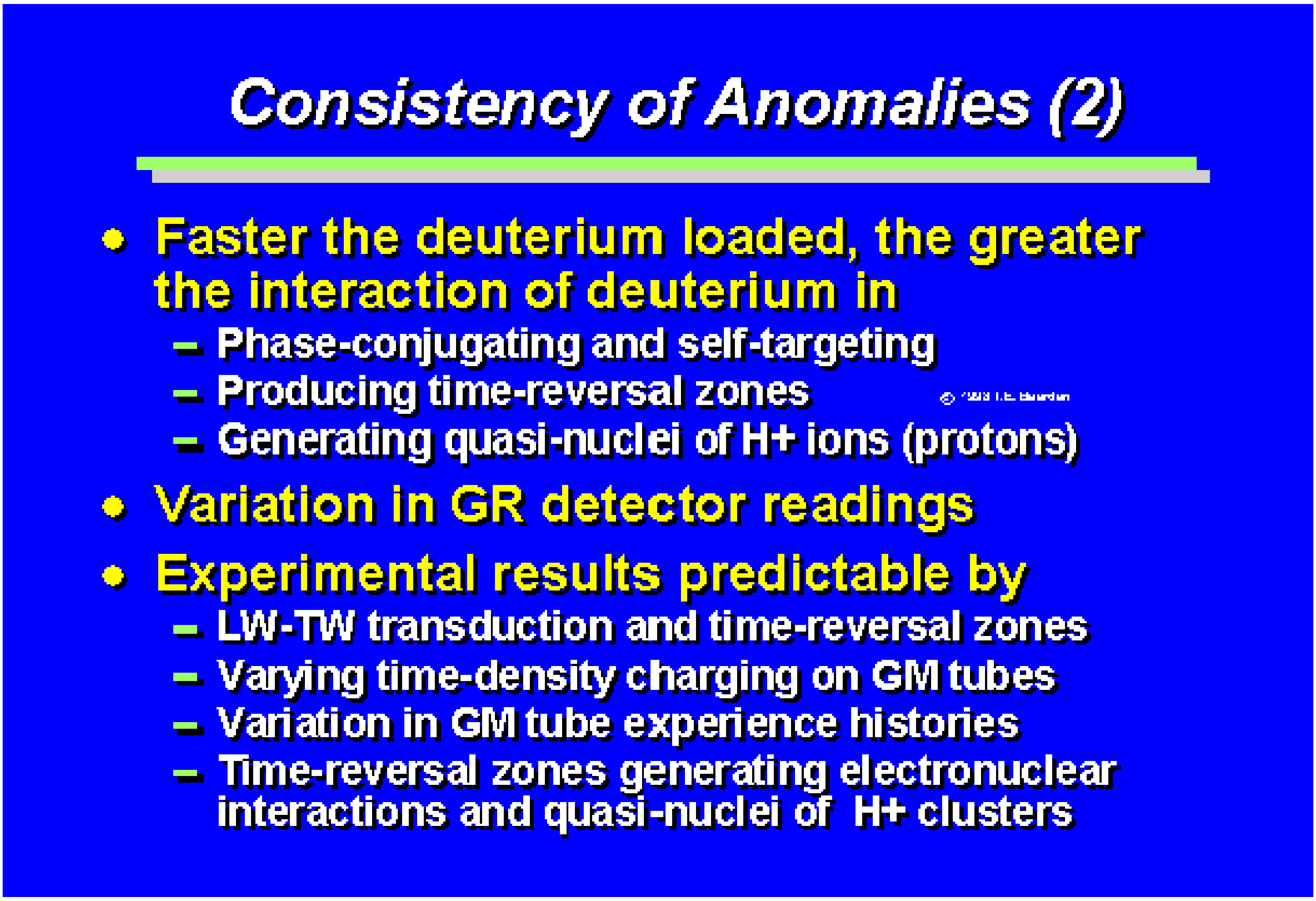

Consistency of Anomalies (2)

- Faster the deuterium loaded, the greater the interaction of deuterium in
  - Phase-conjugating and self-targeting
  - Producing time-reversal zones
  - Generating quasi-nuclei of H+ ions (protons)
- Variation in GR detector readings
- Experimental results predictable by
  - LW-TW transduction and time-reversal zones
  - Varying time-density charging on GM tubes
  - Variation in GM tube experience histories
  - Time-reversal zones generating electronuclear interactions and quasi-nuclei of H+ clusters

Published by Lotus® Freelance Graphics®

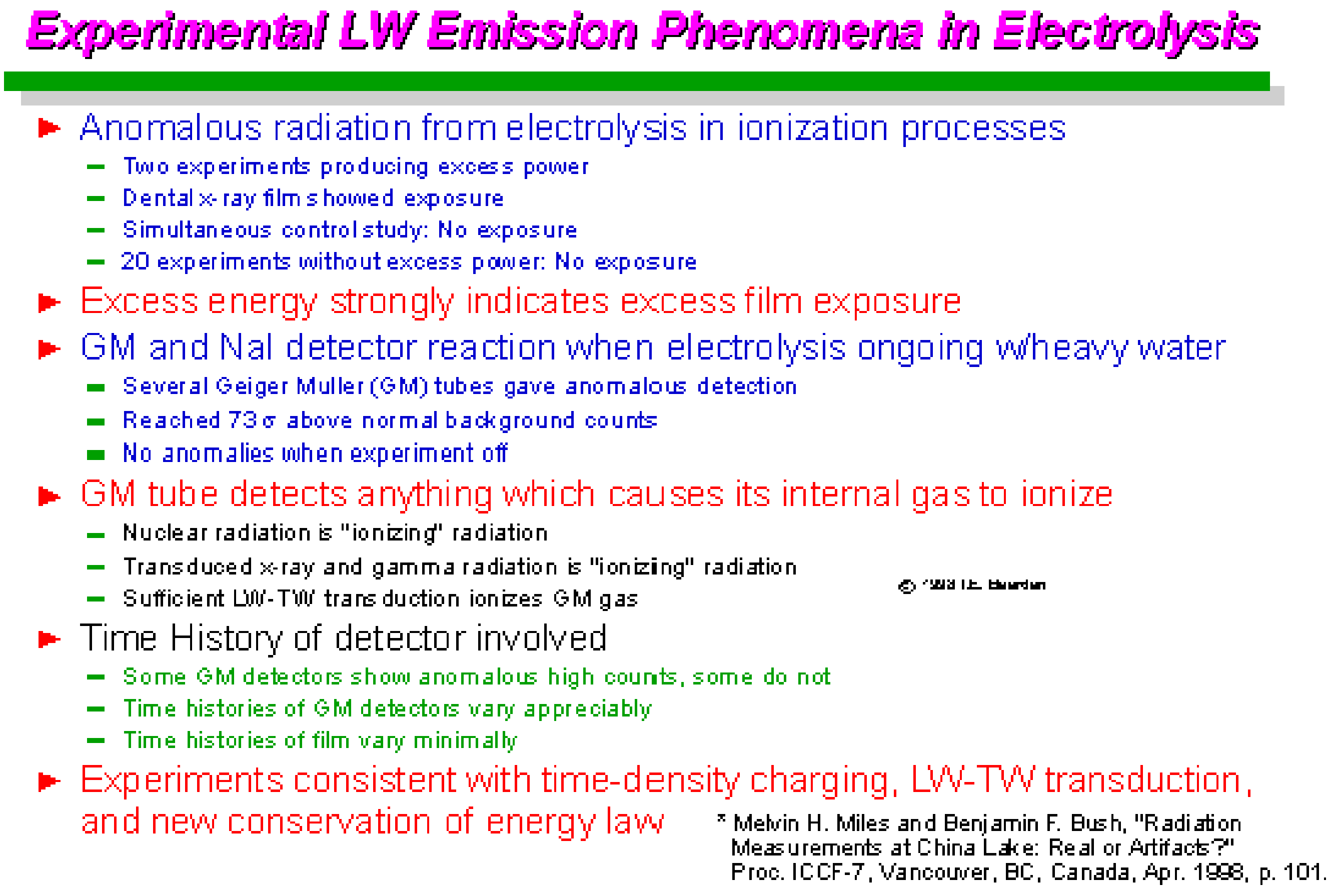

Experimental LW Emission Phenomena in Electrolysis

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental x-ray film showed exposure
  - Simultaneous control study: No exposure
  - 20 experiments without excess power: No exposure
- Excess energy strongly indicates excess film exposure
- GM and NaI detector reaction when electrolysis ongoing w/heavy water
  - Several Geiger Muller (GM) tubes gave anomalous detection
  - Reached 73σ above normal background counts
  - No anomalies when experiment off
- GM tube detects anything which causes its internal gas to ionize
  - Nuclear radiation is "ionizing" radiation
  - Transduced x-ray and gamma radiation is "ionizing" radiation
  - Sufficient LW-TW transduction ionizes GM gas
- Time History of detector involved
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors vary appreciably
  - Time histories of film vary minimally
- Experiments consistent with time-density charging, LW-TW transduction, and new conservation of energy law

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101.

Published by Lotus® Freelance Graphics®

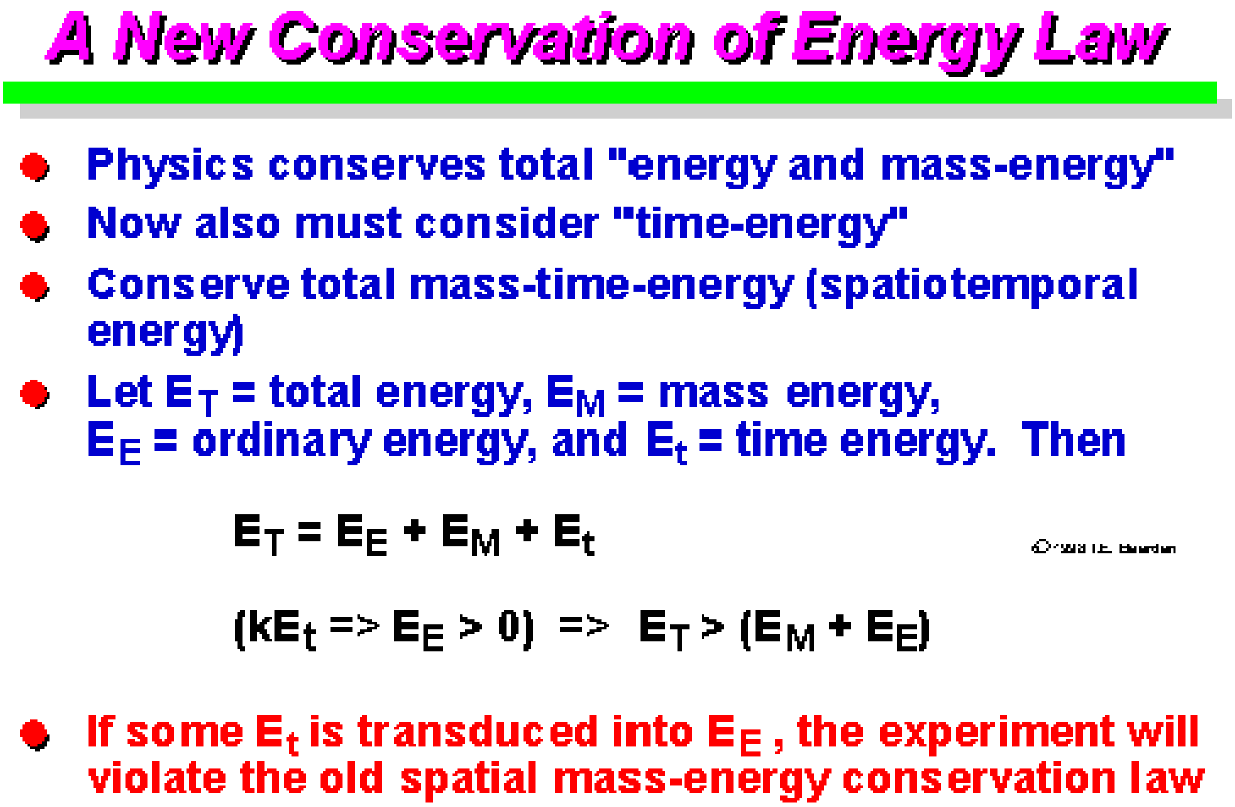

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

Published by Lotus® Freelance Graphics®

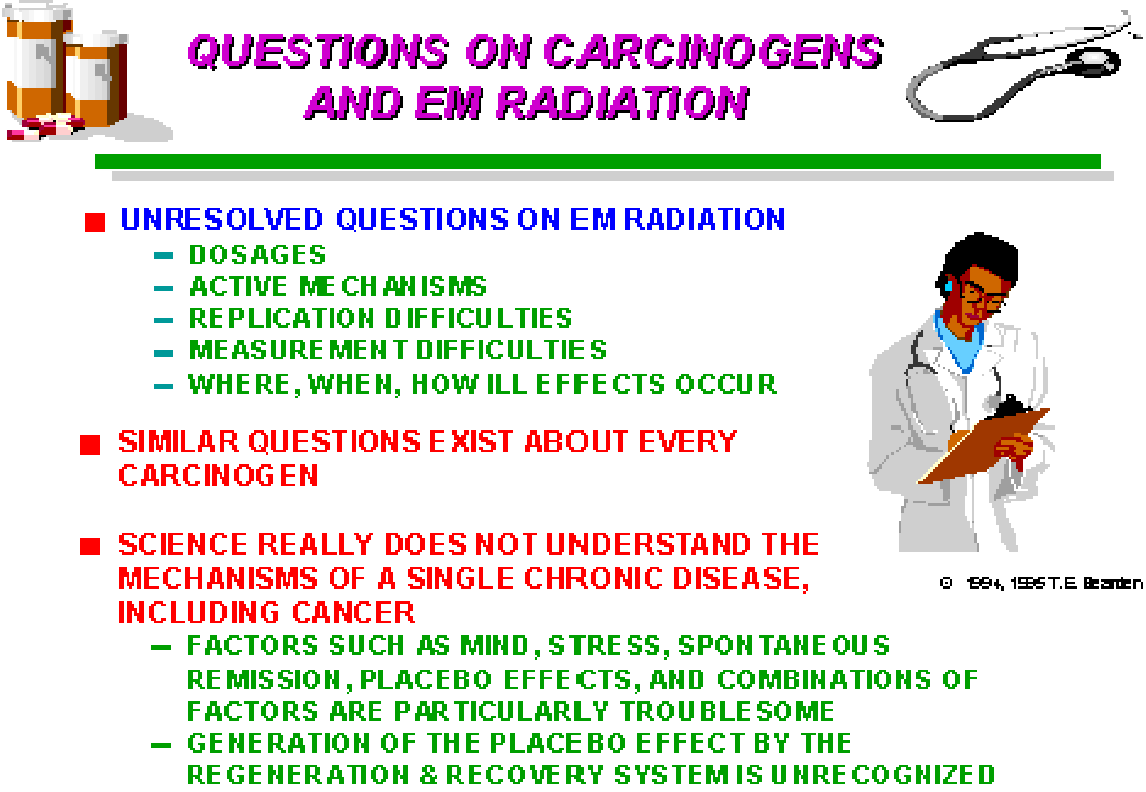

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1996 T.E. Bearden

Published by Lotus® Freelance Graphics®

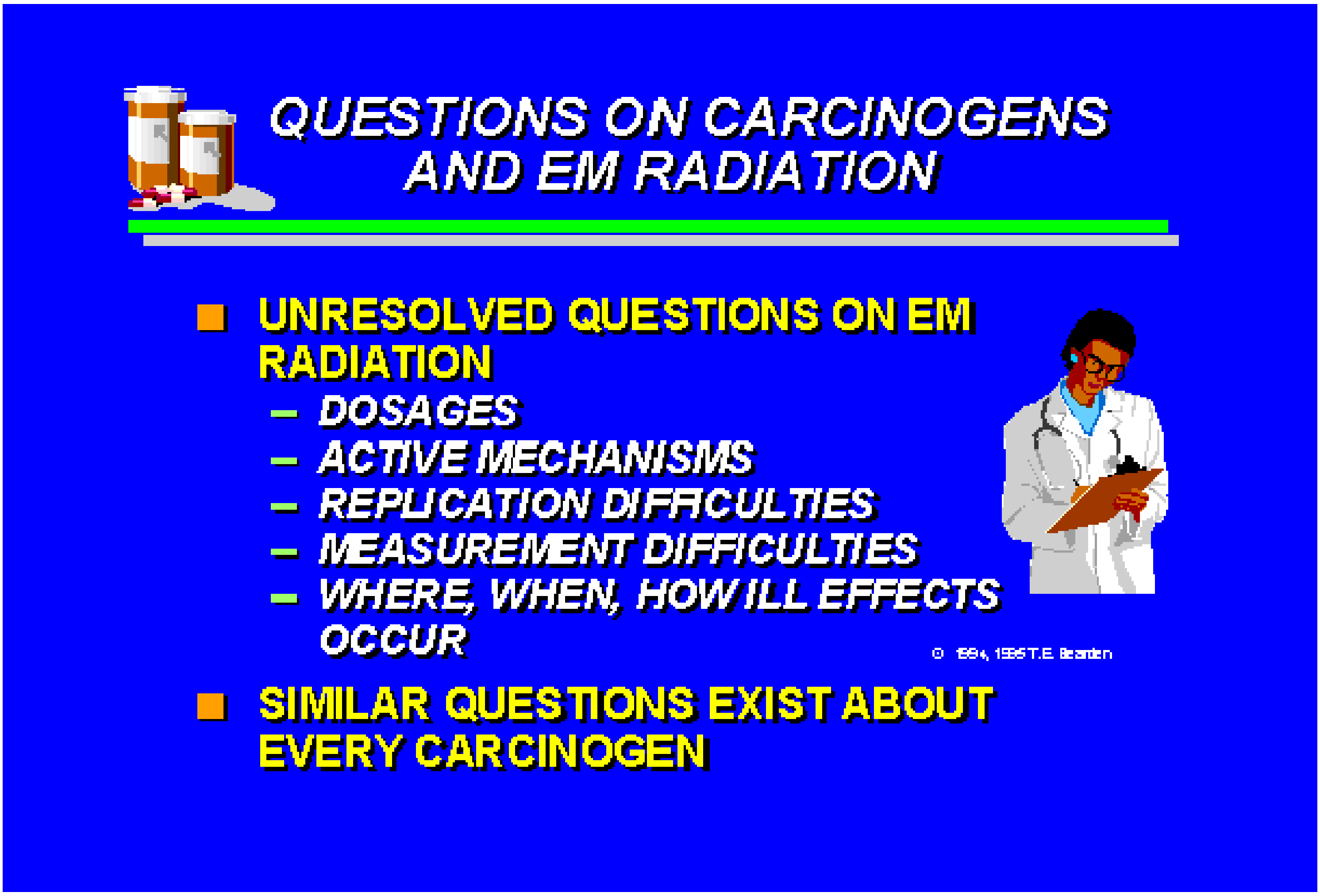

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN

© 1994, 1995 T.E. Bearden

Published by Lotus® Freelance Graphics®

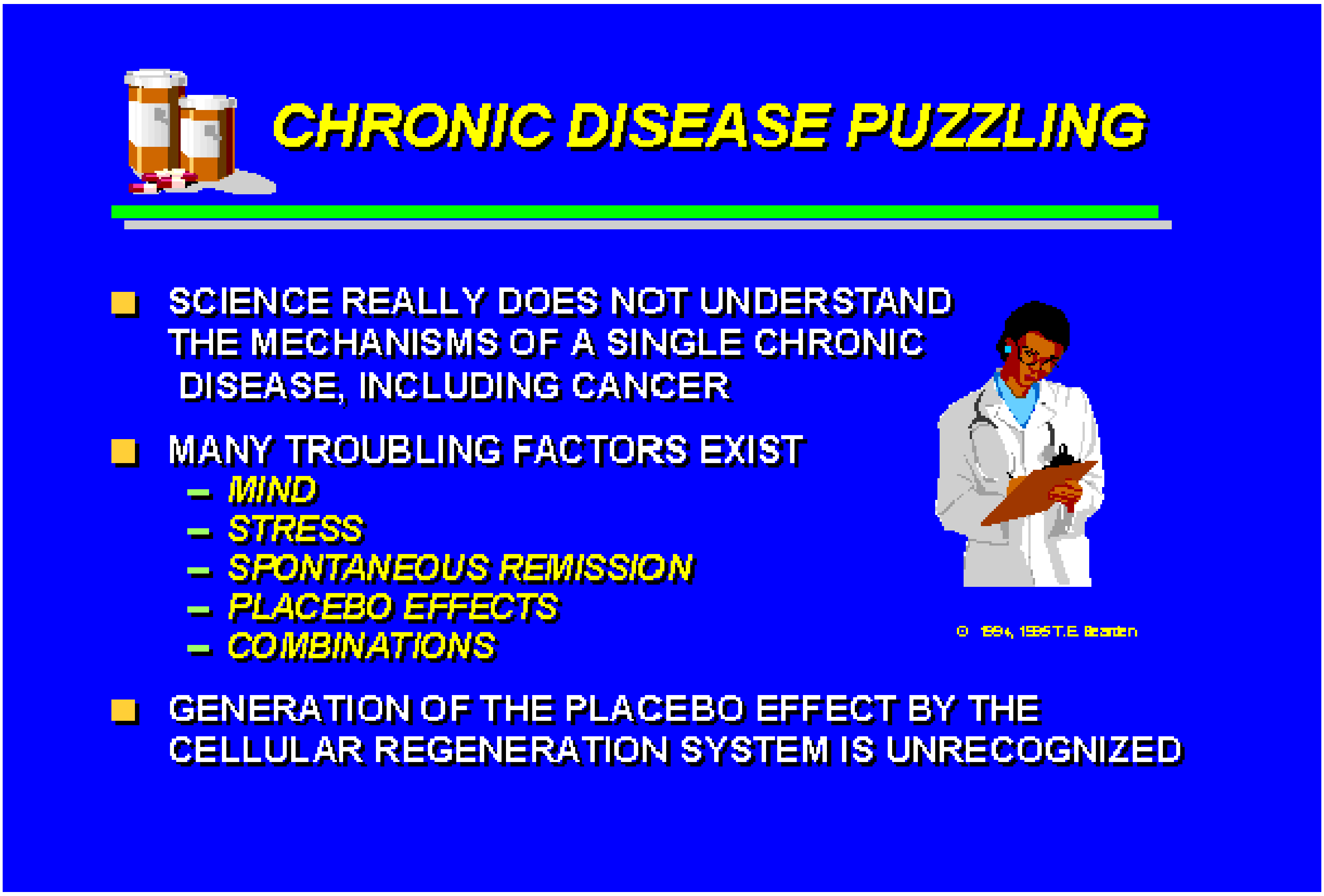

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1996 T.E. Bearden

Published by Lotus® Freelance Graphics®

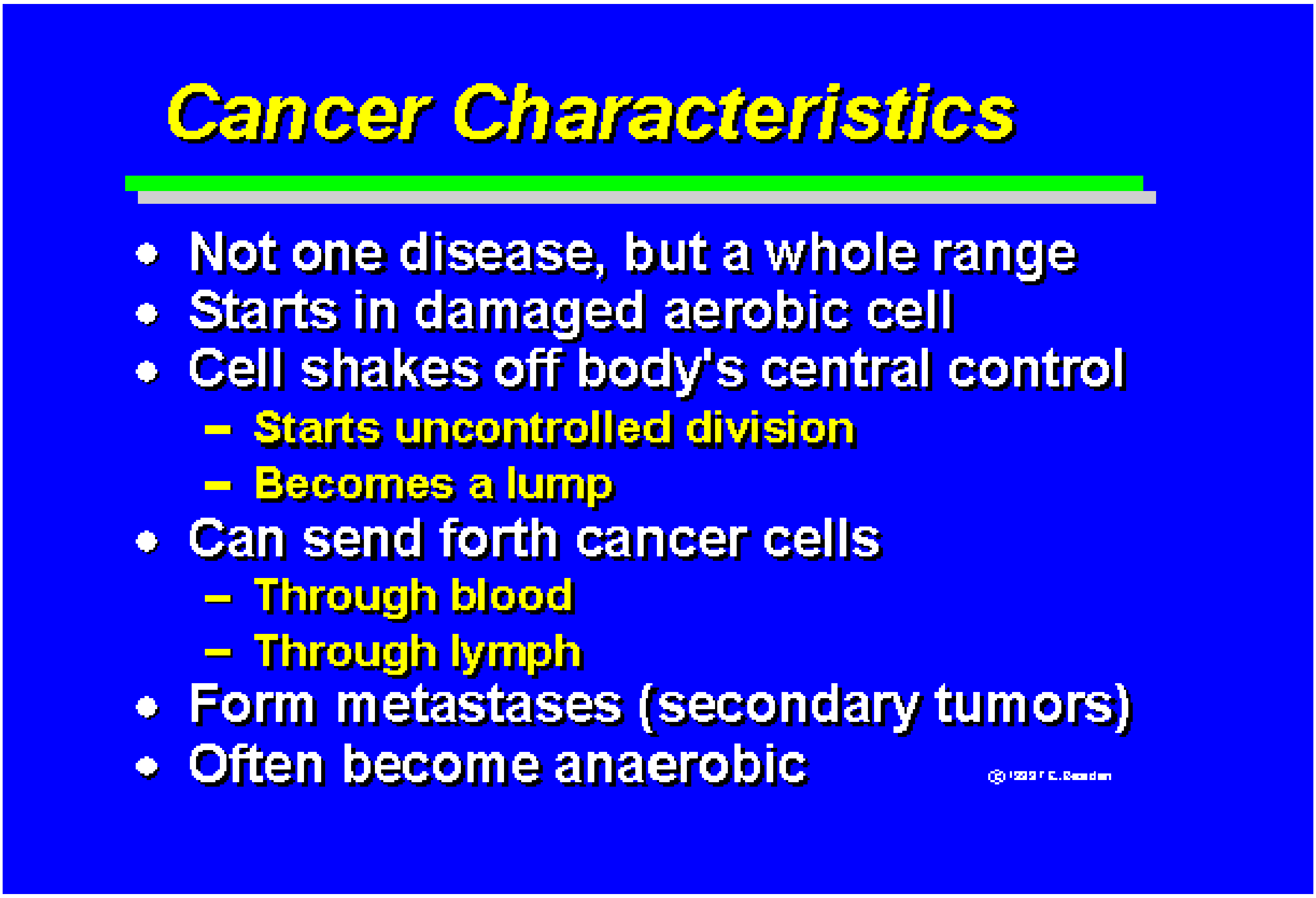

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

Published by Lotus® Freelance Graphics®

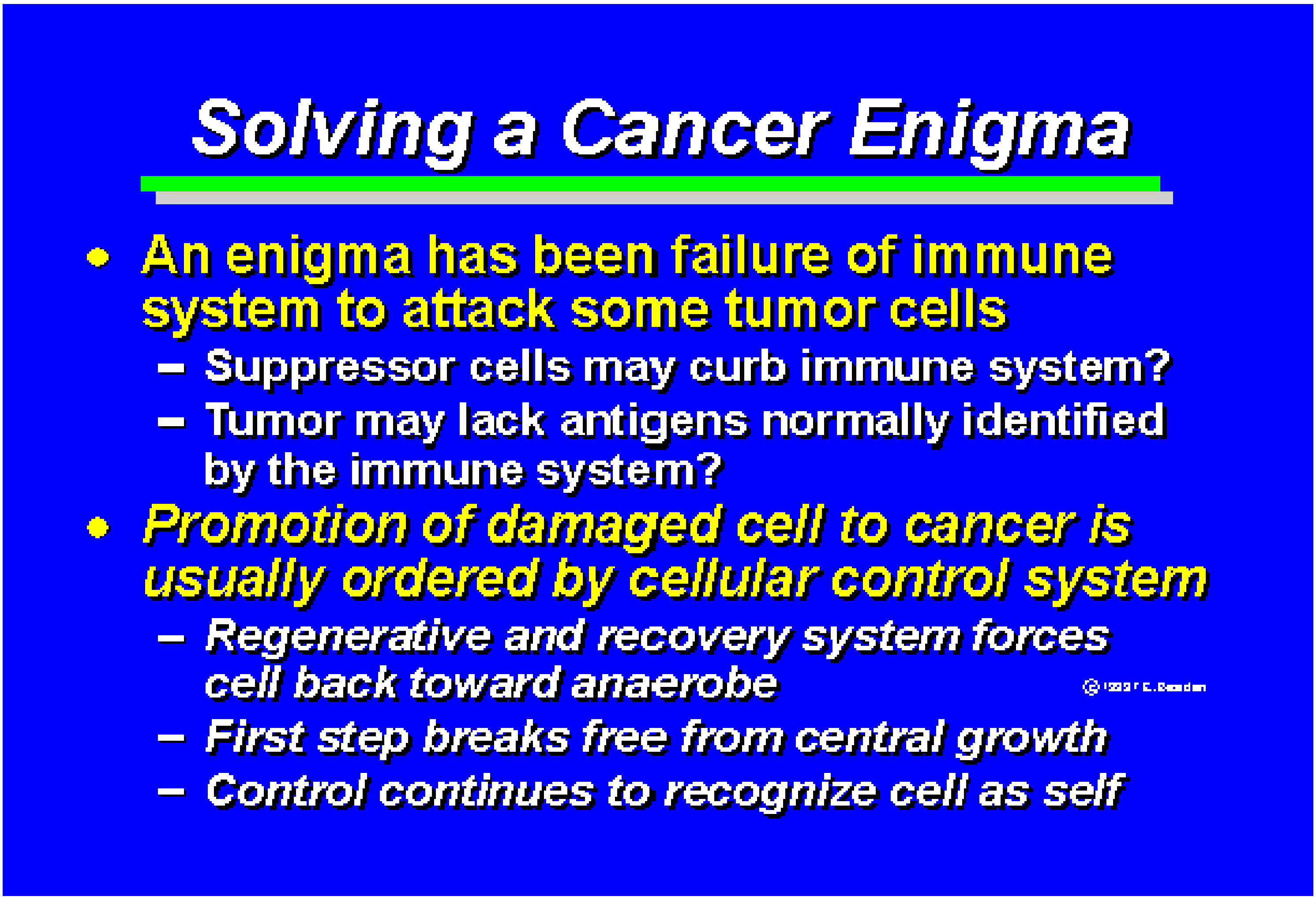

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

Published by Lotus® Freelance Graphics®

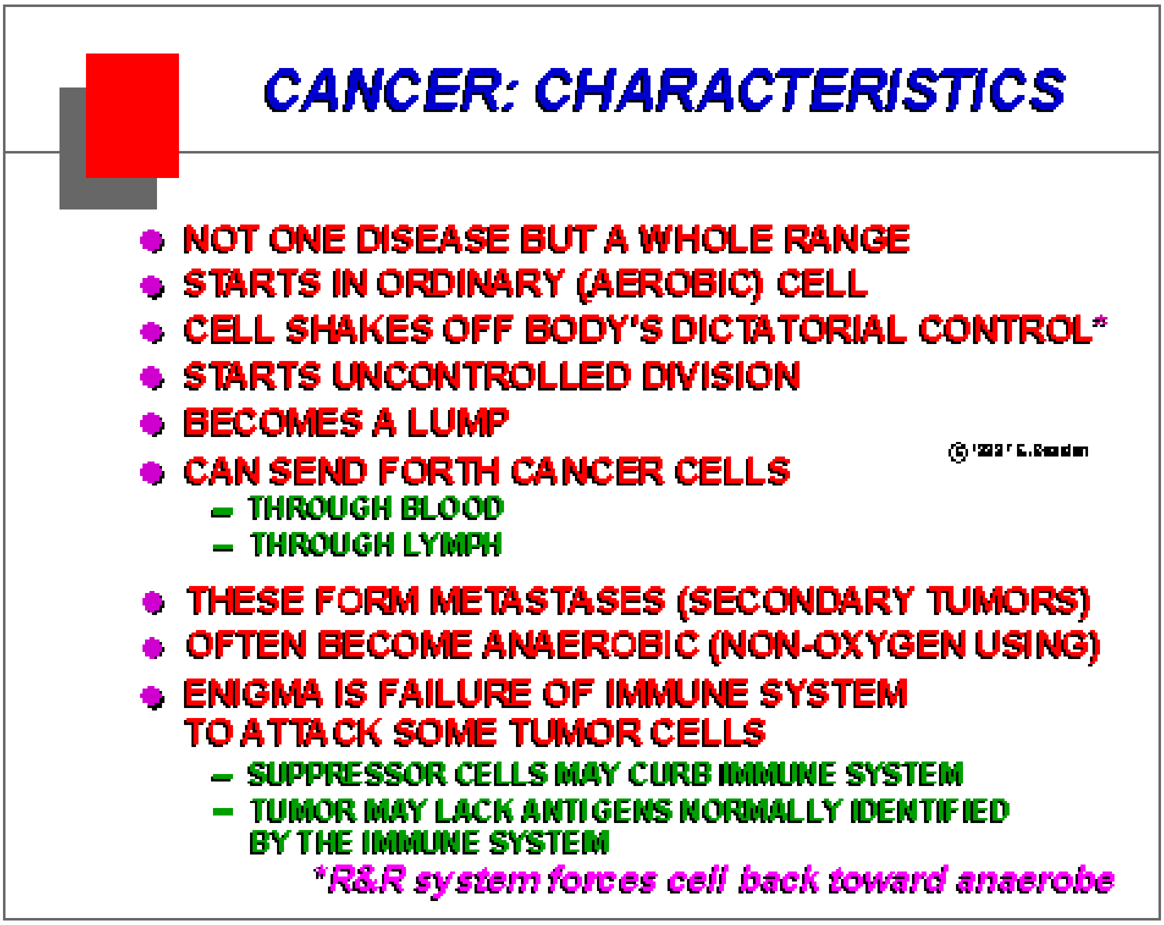

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

R&R system forces cell back toward anaerobe

© [illegible]

Published by Lotus® Freelance Graphics®

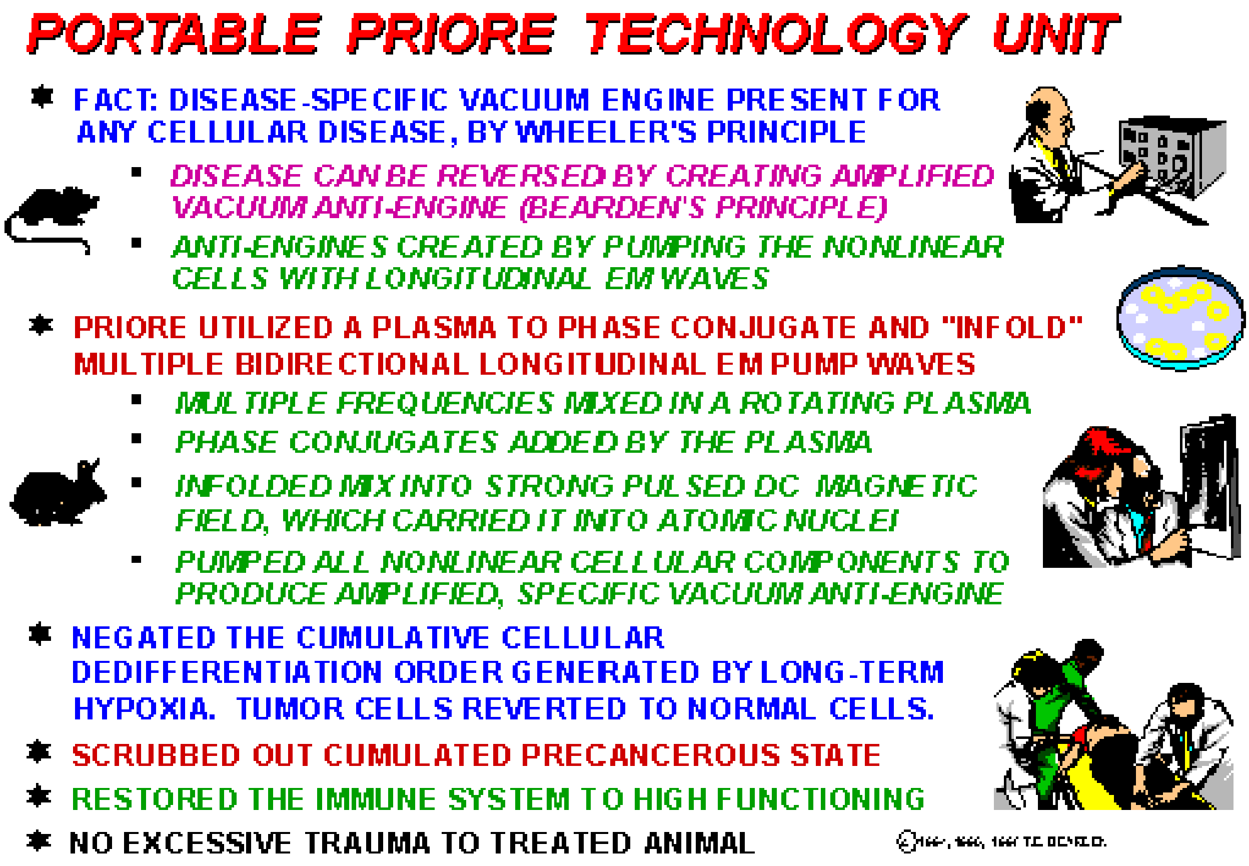

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINES CREATED BY PUMPING THE NONLINEAR CELLS WITH LONGITUDINAL EM WAVES*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *MULTIPLE FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

©1994, 1996, 1997 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

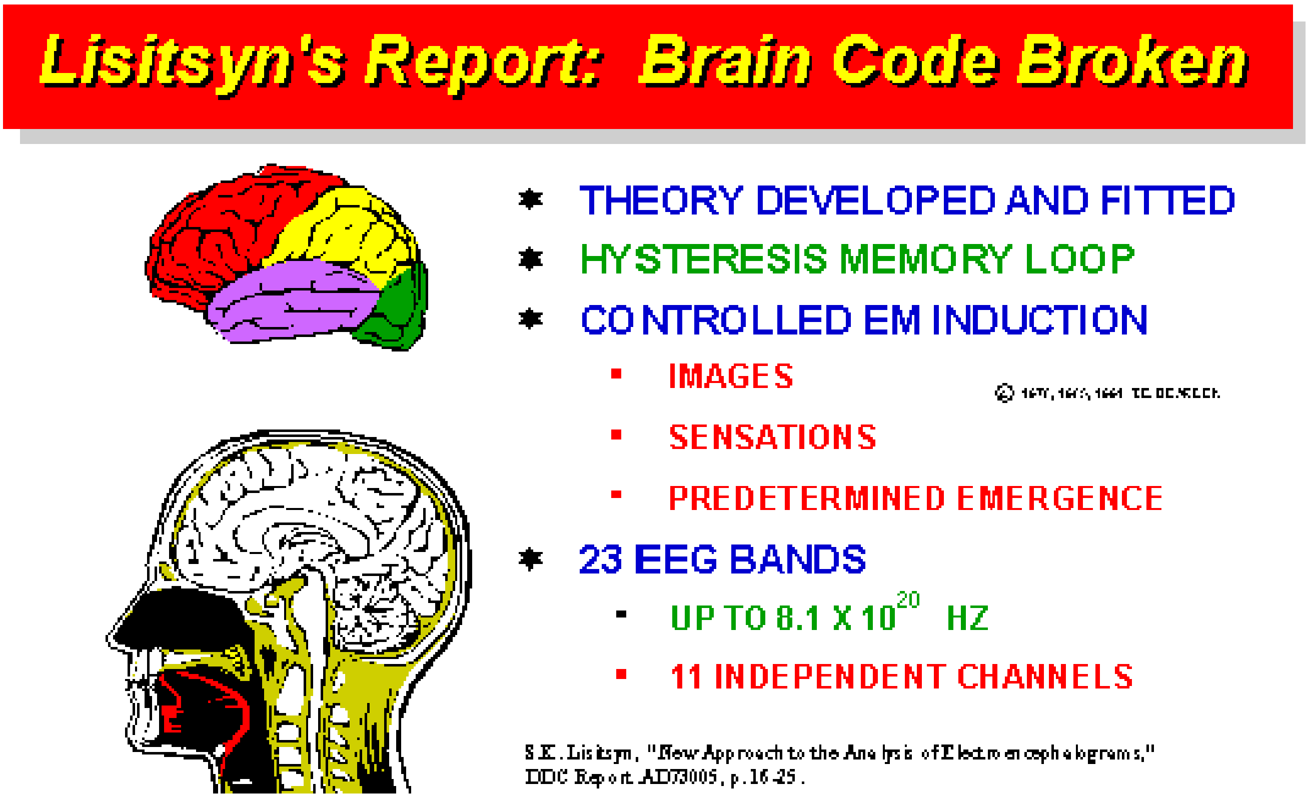

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1×10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Published by Lotus® Freelance Graphics®

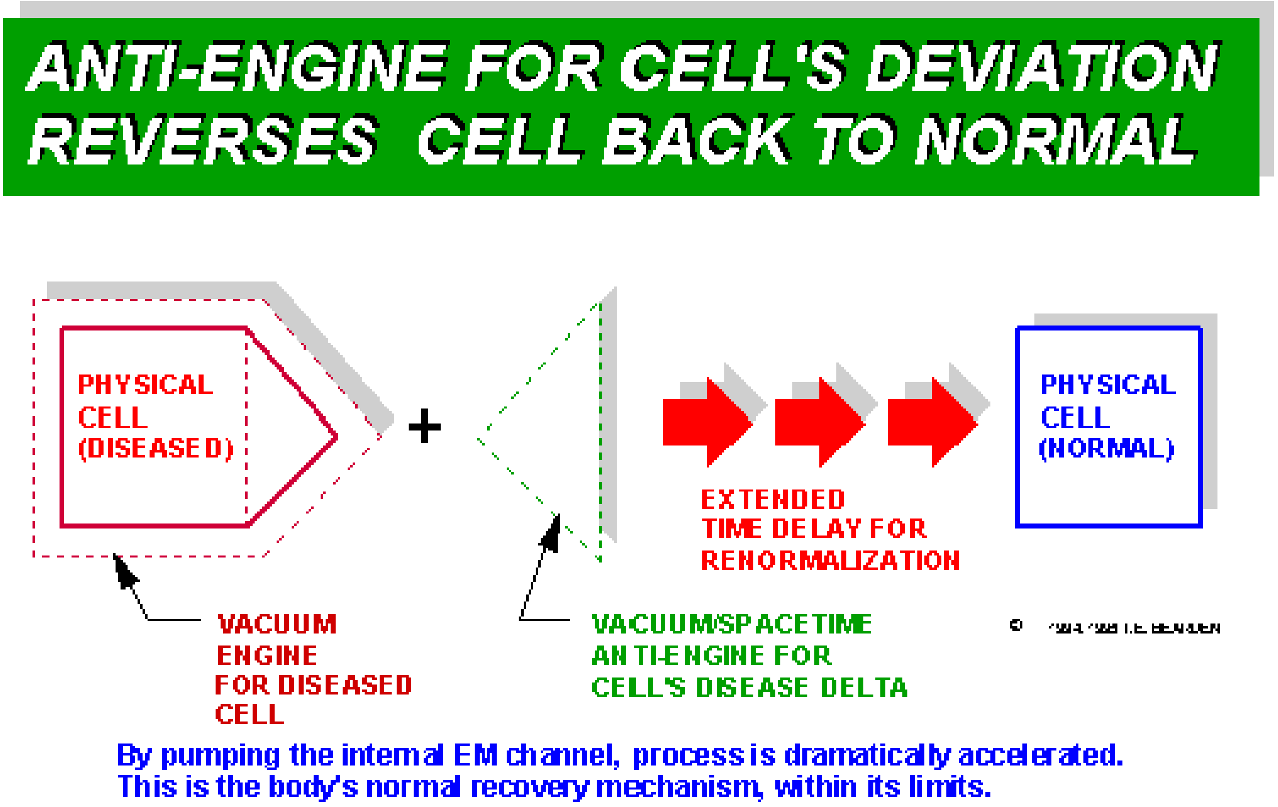

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)
+
EXTENDED
TIME DELAY FOR
RENORMALIZATION
PHYSICAL
CELL
(NORMAL)
VACUUM
ENGINE
FOR DISEASED
CELL
VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA

© [illegible] T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

Published by Lotus® Freelance Graphics®

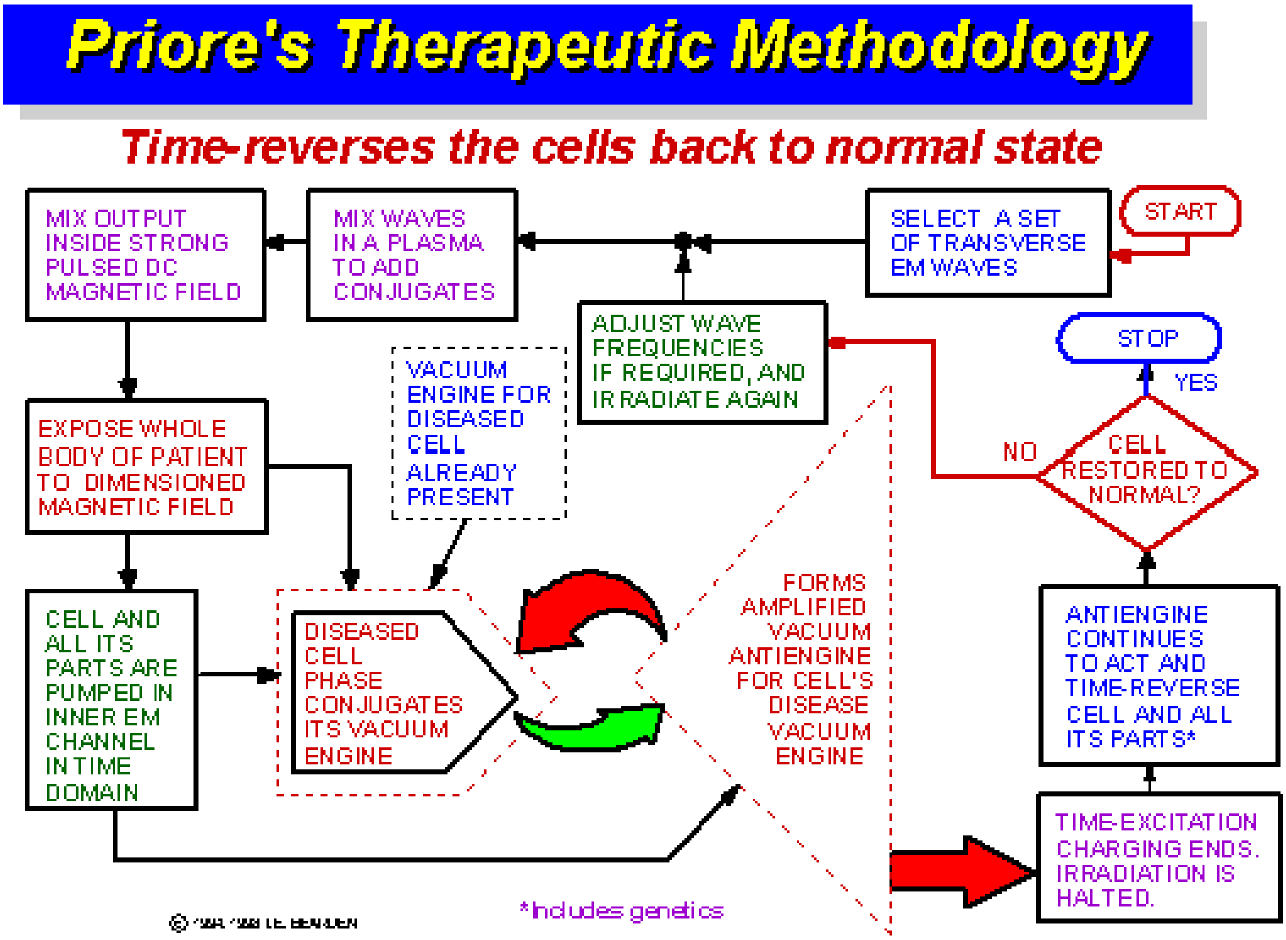
Priore's Therapeutic Methodology
Time-reverses the cells back to normal state
START
SELECT A SET OF TRANSVERSE EM WAVES
MIX WAVES IN A PLASMA TO ADD CONJUGATES
MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD
EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN
NO
CELL RESTORED TO NORMAL?
YES
STOP
ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*
TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.
*Includes genetics

Published by Lotus® Freelance Graphics®

Priore's Therapeutic Methodology

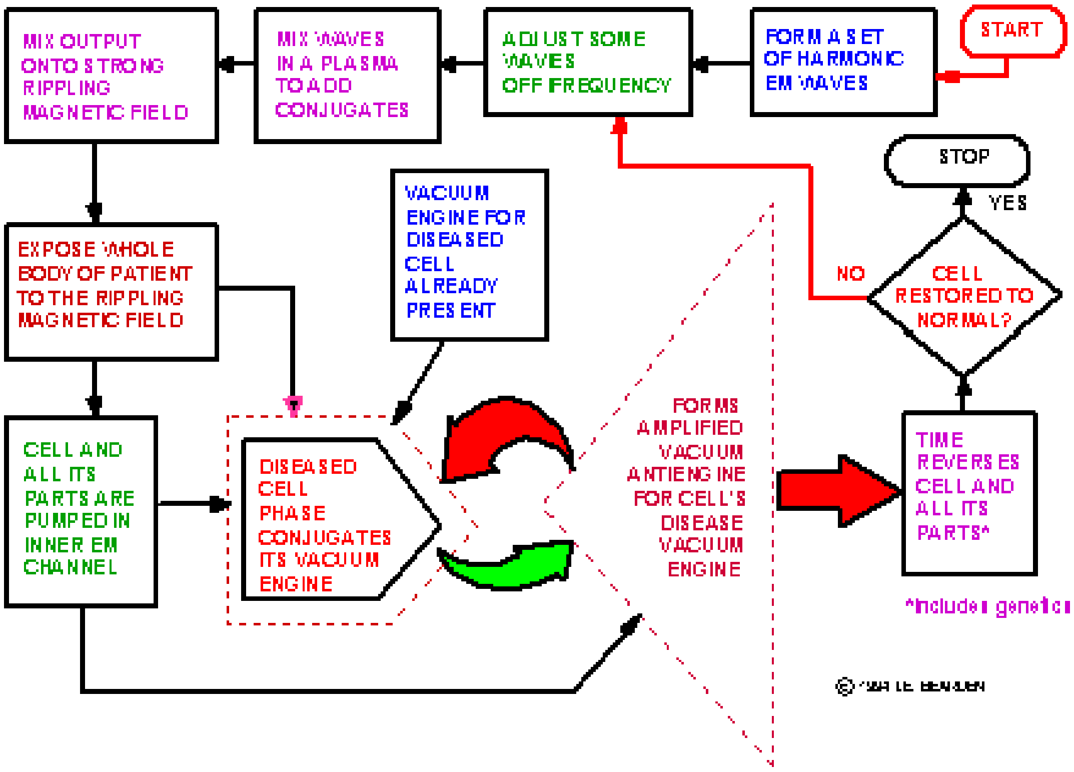
START
FORM A SET OF HARMONIC EM WAVES
ADJUST SOME WAVES OFF FREQUENCY
MIX WAVES IN A PLASMA TO ADD CONJUGATES
MIX OUTPUT ONTO STRONG RIPPLING MAGNETIC FIELD
EXPOSE WHOLE BODY OF PATIENT TO THE RIPPLING MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
TIME REVERSES CELL AND ALL ITS PARTS*
CELL RESTORED TO NORMAL?
NO
YES
STOP
*includes genetics
© 1994 T.E. BEARDEN

Published by Lotus® Freelance Graphics®

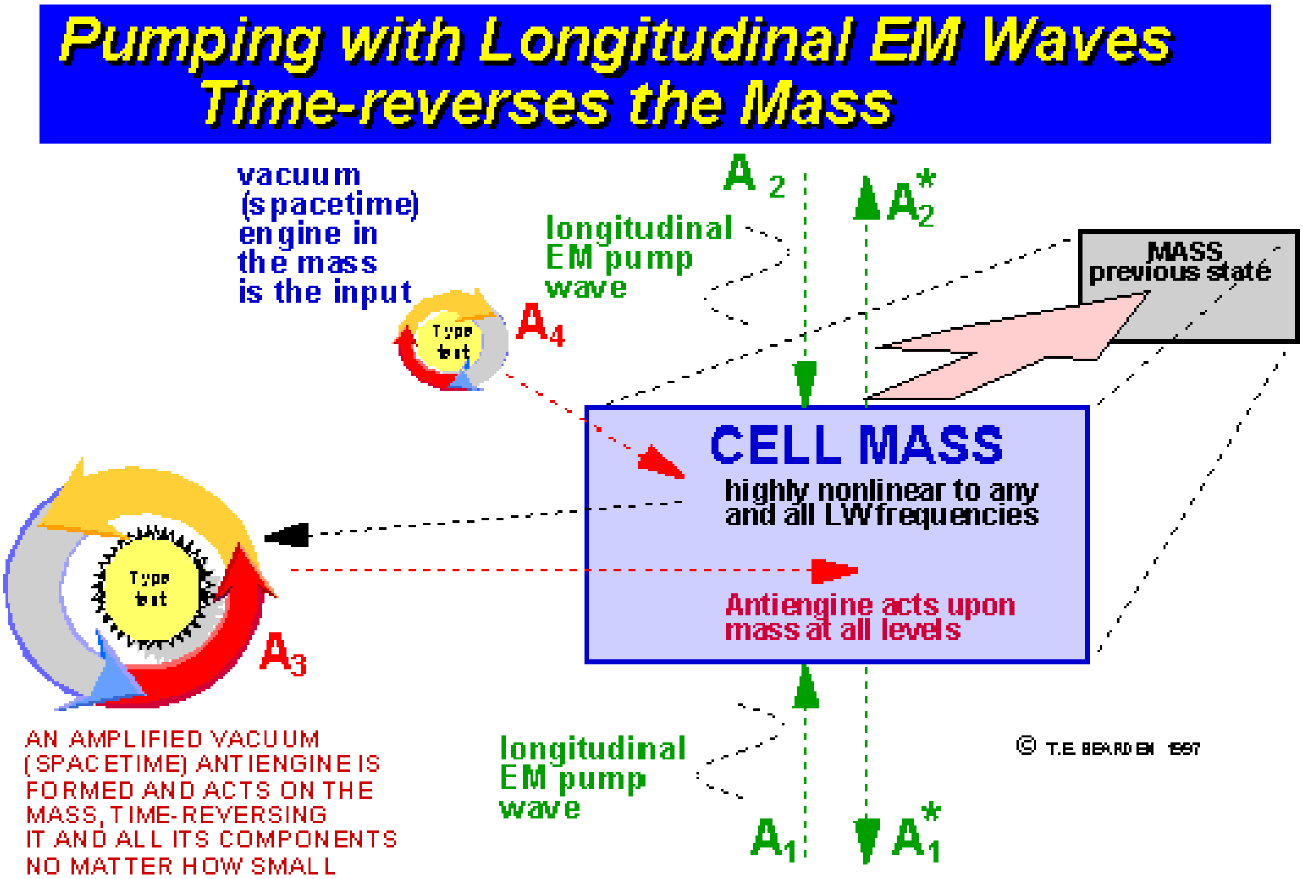
Pumping with Longitudinal EM Waves
Time-reverses the Mass
vacuum
(spacetime)
engine in
the mass
is the input
A_2
A_2^*
longitudinal
EM pump
wave
MASS
previous state
A_4
Type
text
CELL MASS
highly nonlinear to any
and all LW frequencies
Antiengine acts upon
mass at all levels
Type
text
A_3
AN AMPLIFIED VACUUM
(SPACETIME) ANTIENGINE IS
FORMED AND ACTS ON THE
MASS, TIME-REVERSING
IT AND ALL ITS COMPONENTS
NO MATTER HOW SMALL
longitudinal
EM pump
wave
A_1
A_1^*
© T.E. BEARDEN 1997

Published by Lotus® Freelance Graphics®

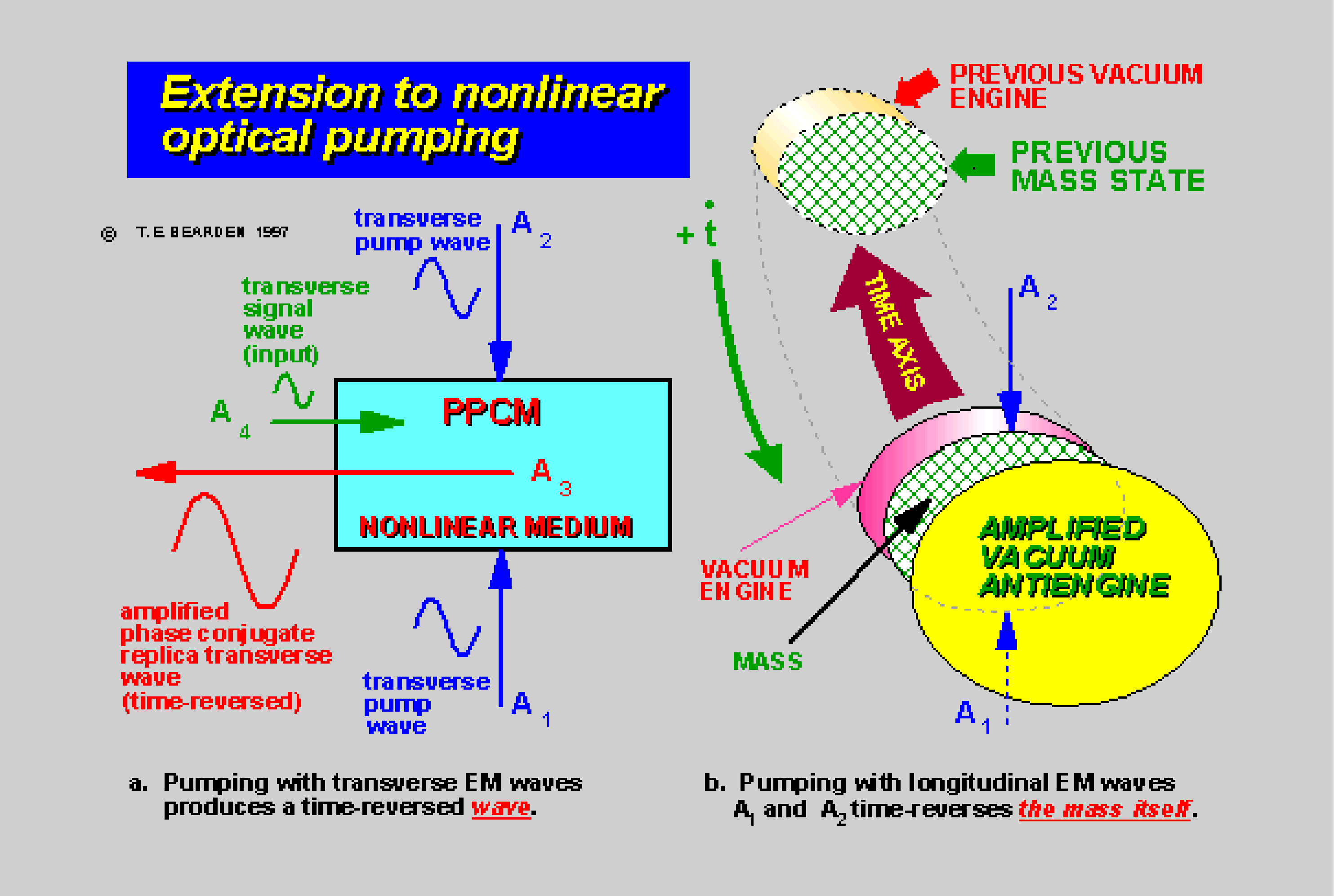

Extension to nonlinear optical pumping

© T.E BEARDEN 1997
transverse pump wave
A 2
transverse signal wave (input)
A 4
PPCM
A 3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A 1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A 2
VACUUM ENGINE
AMPLIFIED VACUUM ANTIENGINE
MASS
A 1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses ***the mass itself***.

Published by Lotus® Freelance Graphics®

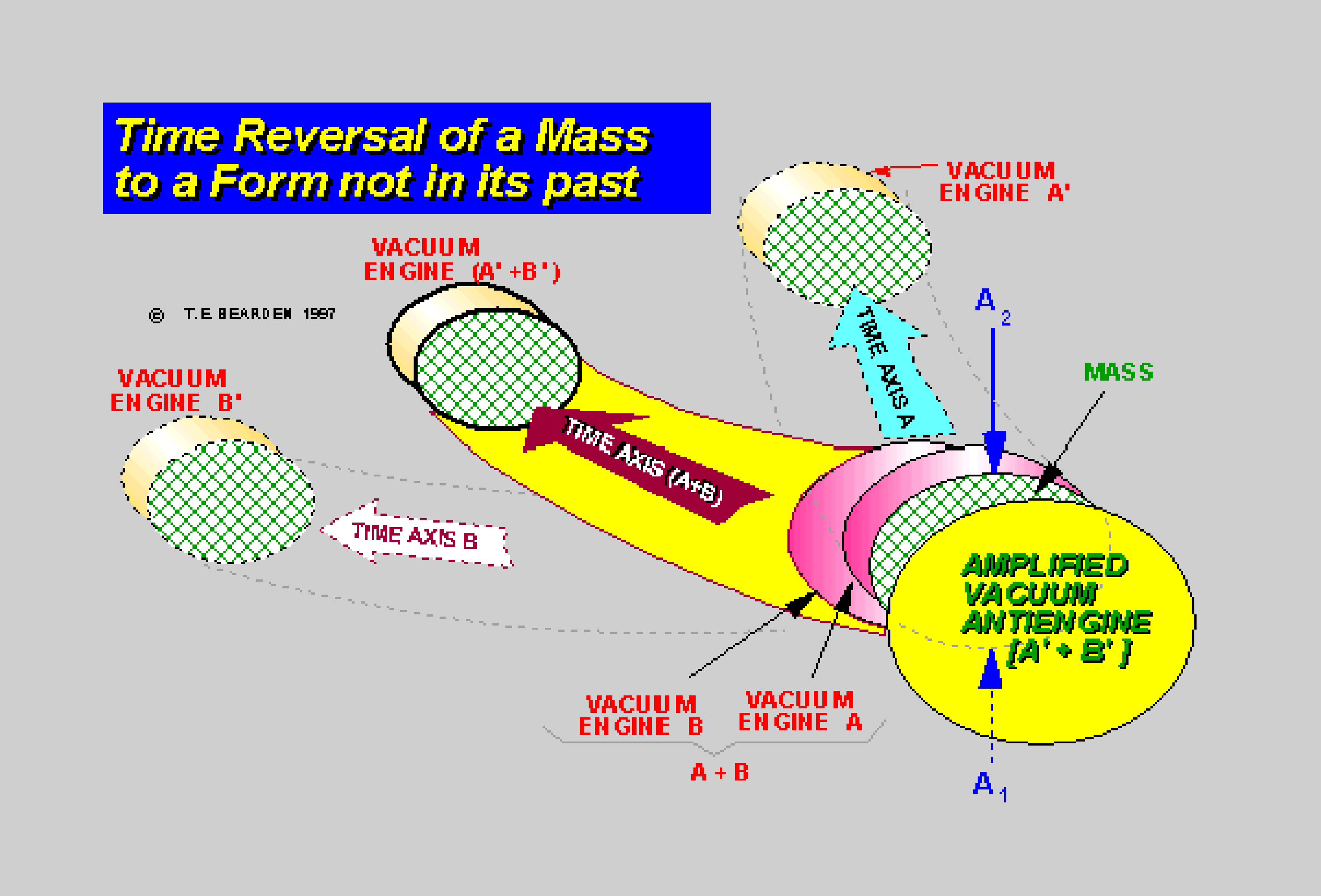

Time Reversal of a Mass to a Form not in its past

© T.E. BEARDEN 1997
VACUUM ENGINE (A' +B ')
VACUUM ENGINE A'
VACUUM ENGINE B'
TIME AXIS B
TIME AXIS (A+B)
TIME AXIS A
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

Published by Lotus® Freelance Graphics®

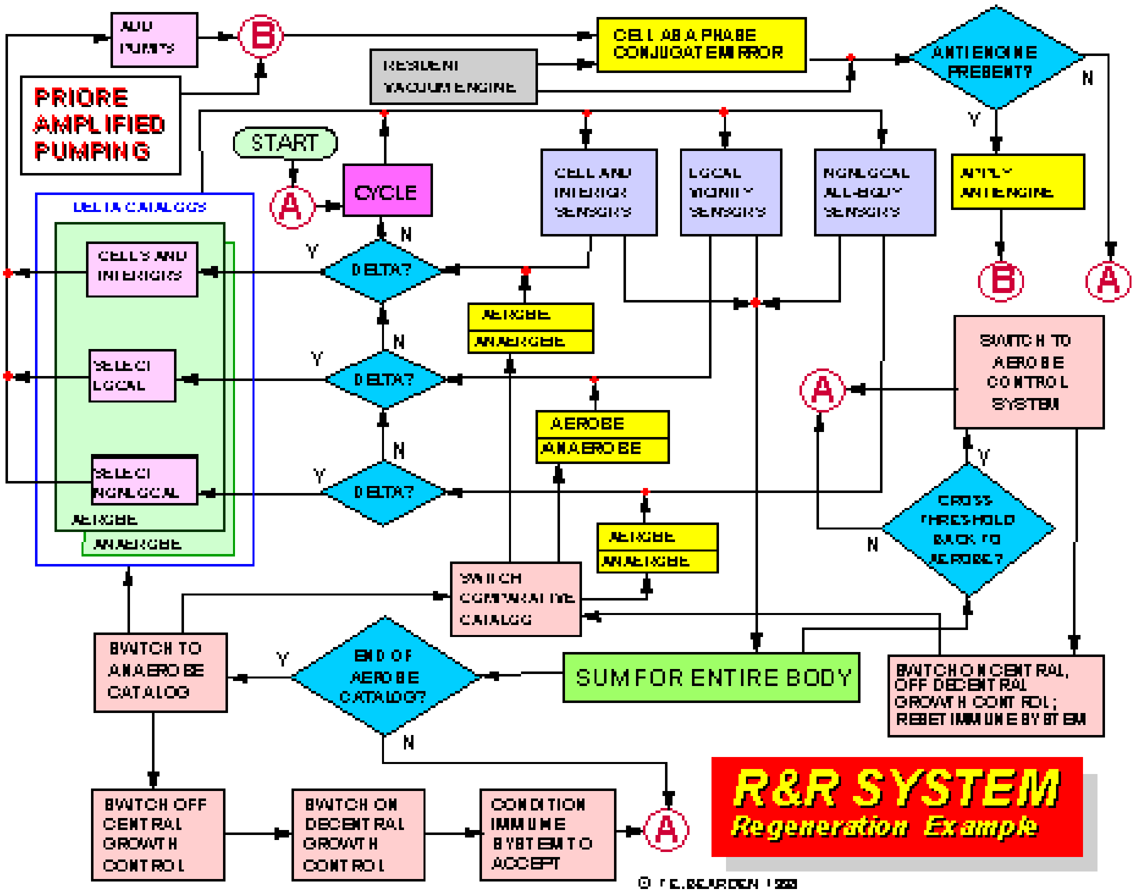
ADD PUMPS
B
PRIORE AMPLIFIED PUMPING
CELL AB A PHASE CONJUGATE MIRROR
RESIDENT VACUUM ENGINE
ANTIENGINE PRESENT?
N
Y
APPLY ANTIENGINE
B
A
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
DELTA CATALOGS
CELLS AND INTERIORS
SELECT LOCAL
SELECT NONLOCAL
AEROBE
ANAEROBE
Y
N
DELTA?
AEROBE
ANAEROBE
Y
N
DELTA?
AEROBE
ANAEROBE
Y
DELTA?
AEROBE
ANAEROBE
SWITCH TO AEROBE CONTROL SYSTEM
A
Y
CROSS THRESHOLD BACK TO AEROBE?
N
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A
R&R SYSTEM
Regeneration Example
© T.E. BEARDEN

Published by Lotus® Freelance Graphics®

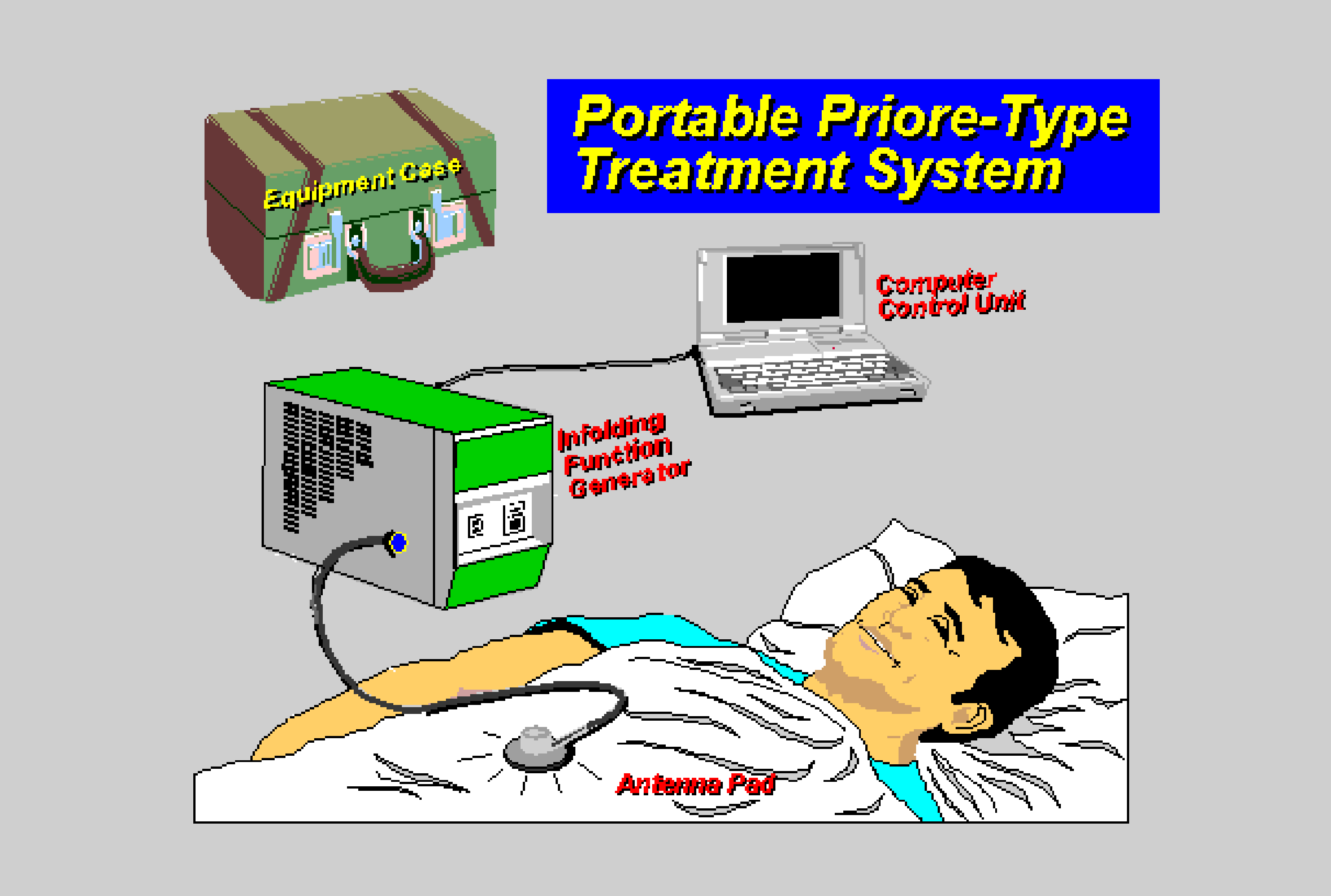
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Published by Lotus® Freelance Graphics®

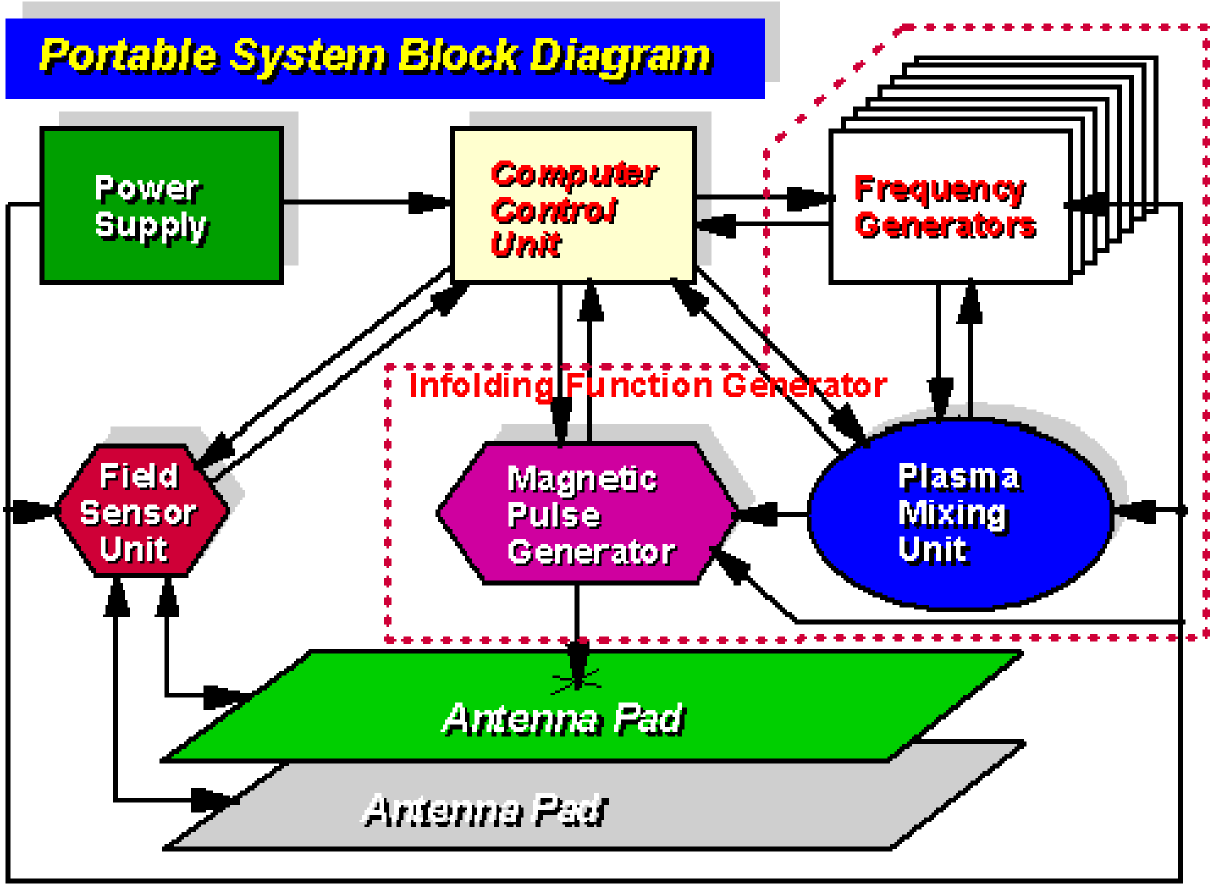
Portable System Block Diagram
Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

Published by Lotus® Freelance Graphics®

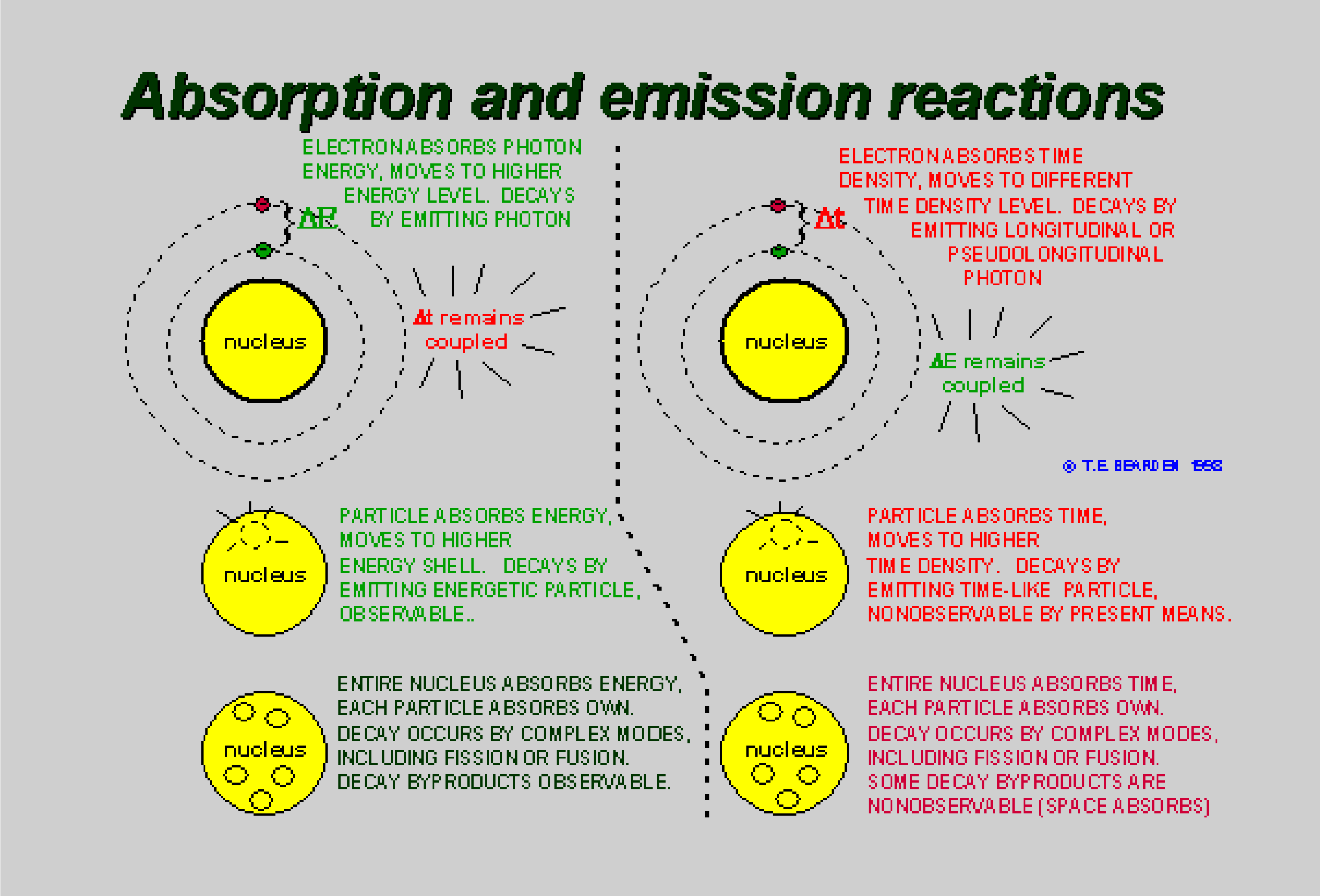

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE, OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

Published by Lotus® Freelance Graphics®

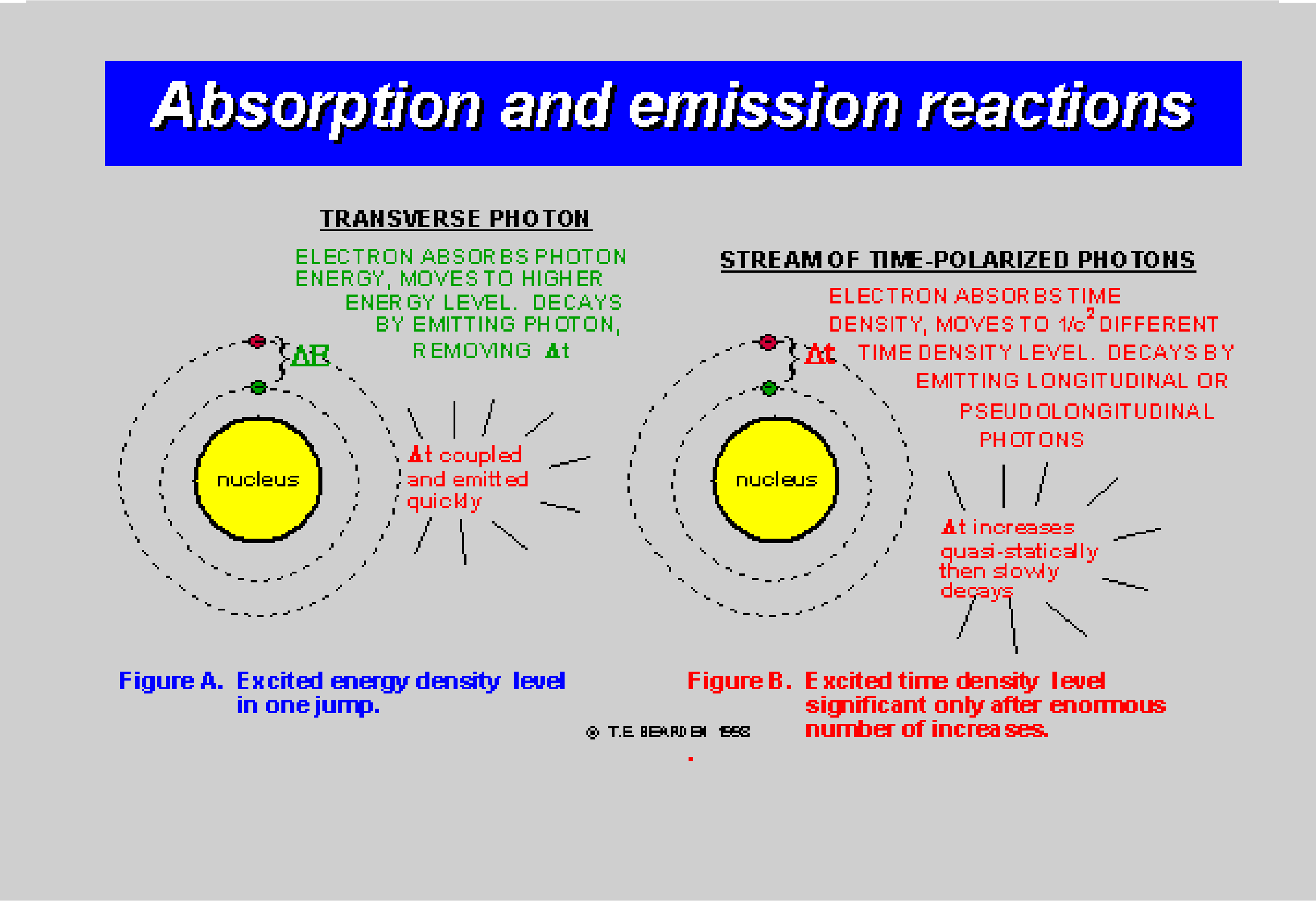

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c^2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E. BEARDEN 1988

Published by Lotus® Freelance Graphics®

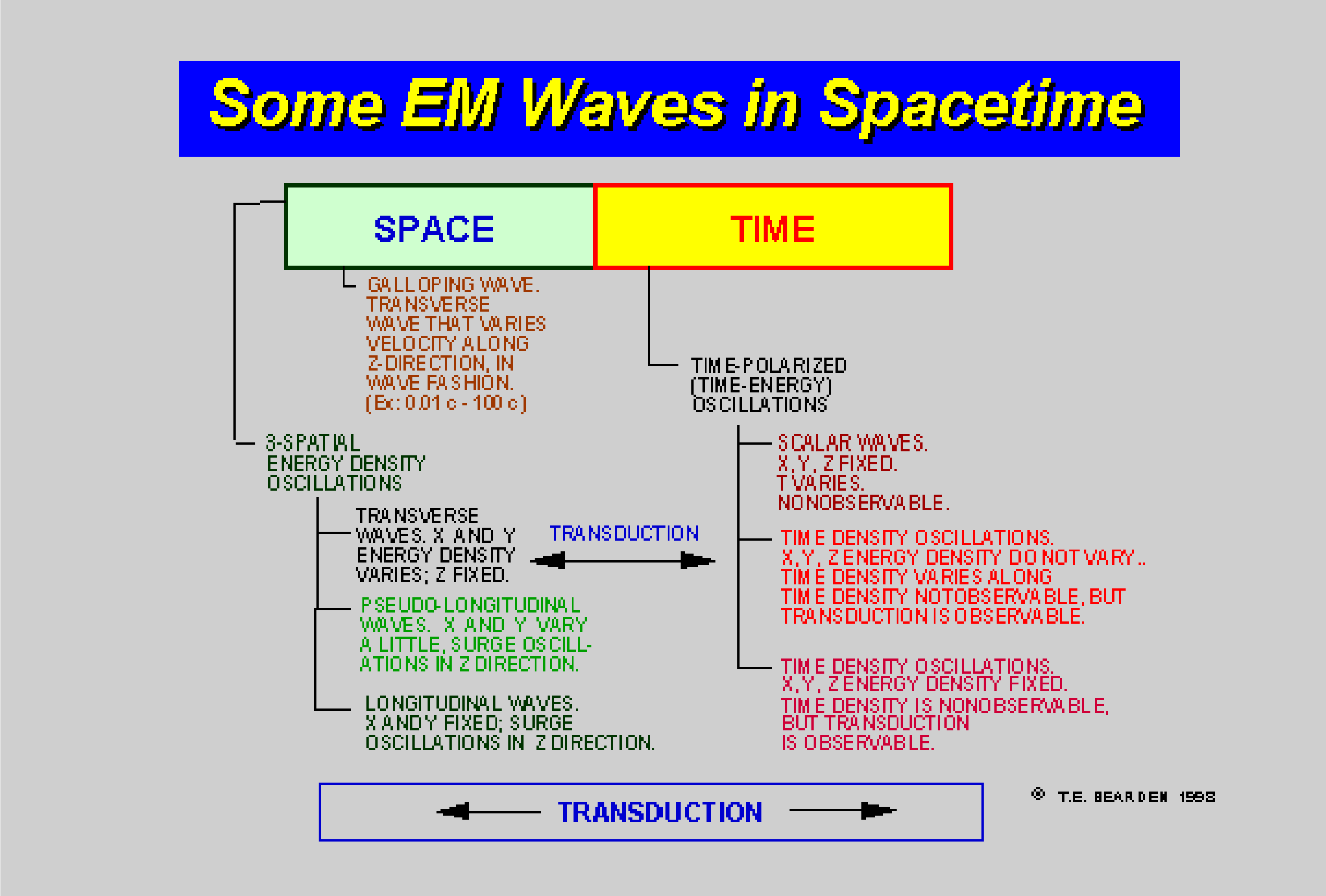

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE. TRANSVERSE WAVE THAT VARIES VELOCITY ALONG Z-DIRECTION, IN WAVE FASHION. (Ex: 0.01 c - 100 c)

TIME-POLARIZED (TIME-ENERGY) OSCILLATIONS

3-SPATIAL ENERGY DENSITY OSCILLATIONS

SCALAR WAVES. X, Y, Z FIXED. T VARIES. NONOBSERVABLE.

TRANSVERSE WAVES. X AND Y ENERGY DENSITY VARIES; Z FIXED.

TRANSDUCTION

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY DO NOT VARY.. TIME DENSITY VARIES ALONG TIME DENSITY NOT OBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

PSEUDO-LONGITUDINAL WAVES. X AND Y VARY A LITTLE, SURGE OSCILLATIONS IN Z DIRECTION.

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY FIXED. TIME DENSITY IS NONOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

LONGITUDINAL WAVES. X AND Y FIXED; SURGE OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

© T.E. BEARDEN 1998

Published by Lotus® Freelance Graphics®

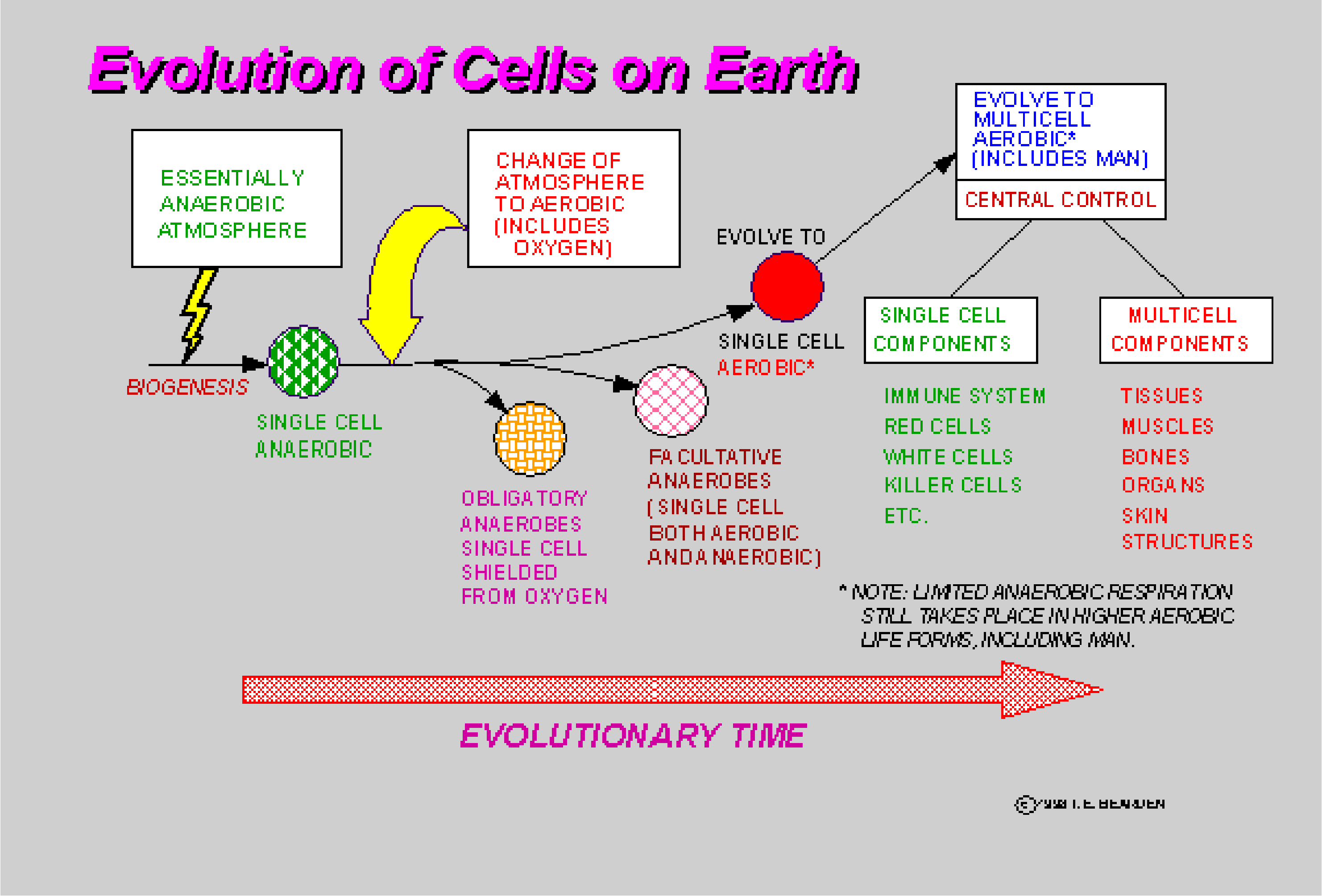
Evolution of Cells on Earth
ESSENTIALLY ANAEROBIC ATMOSPHERE
BIOGENESIS
SINGLE CELL ANAEROBIC
CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)
OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN
FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)
EVOLVE TO
SINGLE CELL AEROBIC*
EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)
CENTRAL CONTROL
SINGLE CELL COMPONENTS
IMMUNE SYSTEM
RED CELLS
WHITE CELLS
KILLER CELLS
ETC.
MULTICELL COMPONENTS
TISSUES
MUSCLES
BONES
ORGANS
SKIN
STRUCTURES
* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.
EVOLUTIONARY TIME

Published by Lotus® Freelance Graphics®

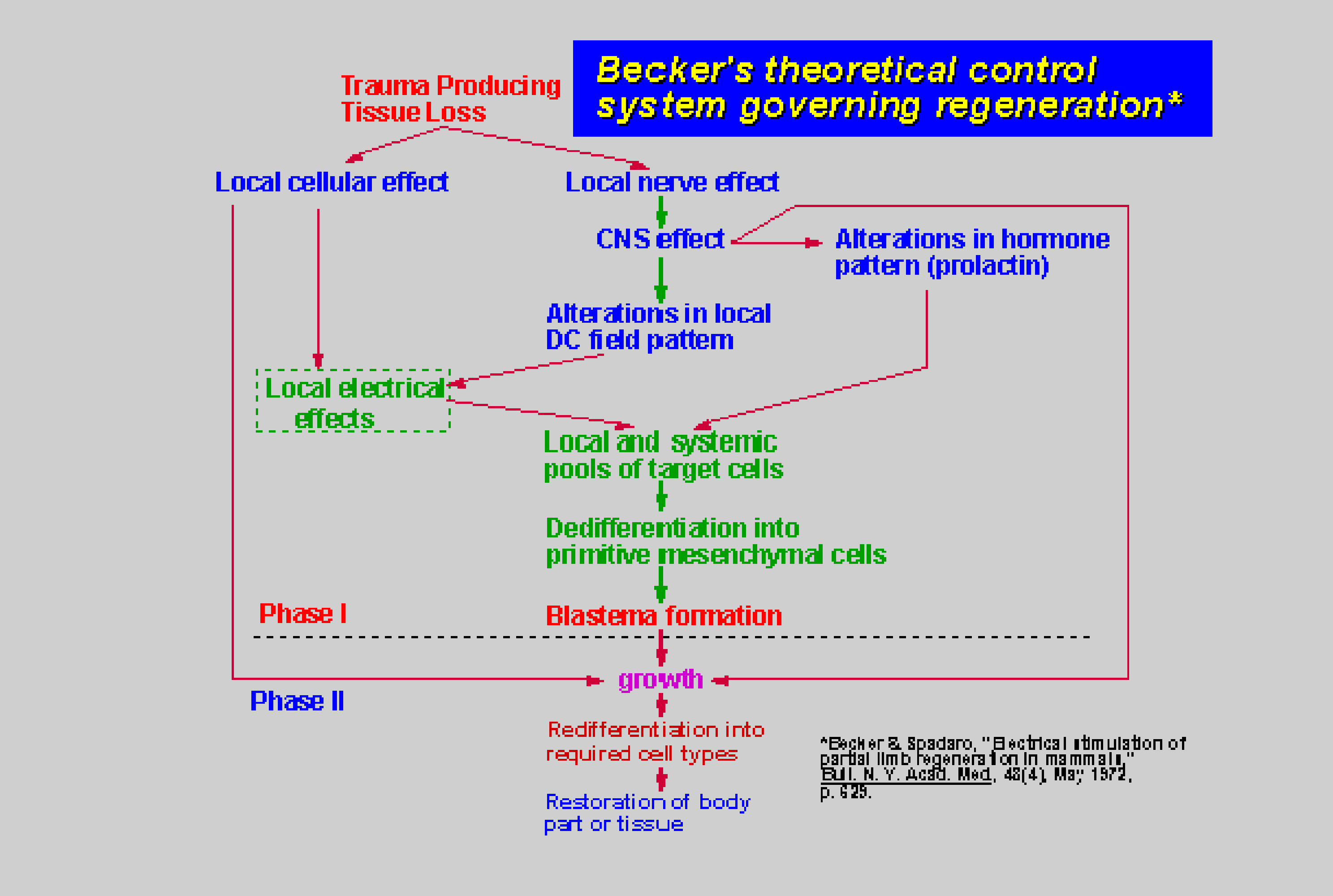

Becker's theoretical control system governing regeneration*

Trauma Producing
Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone
pattern (prolactin)
Alterations in local
DC field pattern
Local electrical
effects
Local and systemic
pools of target cells
Dedifferentiation into
primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into
required cell types
Restoration of body
part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med, 48(4) May 1972, p. 625.

Published by Lotus® Freelance Graphics®

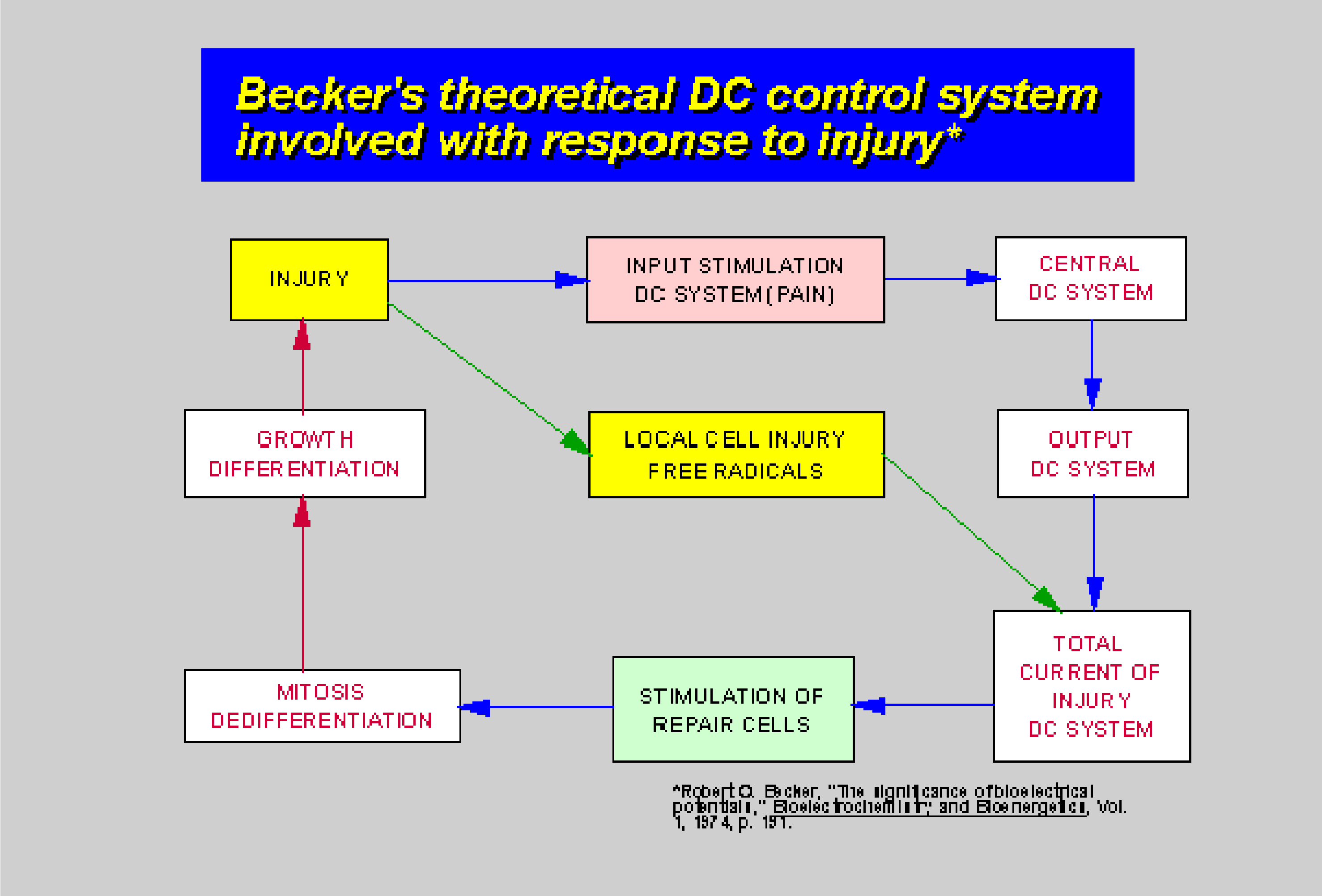

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Published by Lotus® Freelance Graphics®

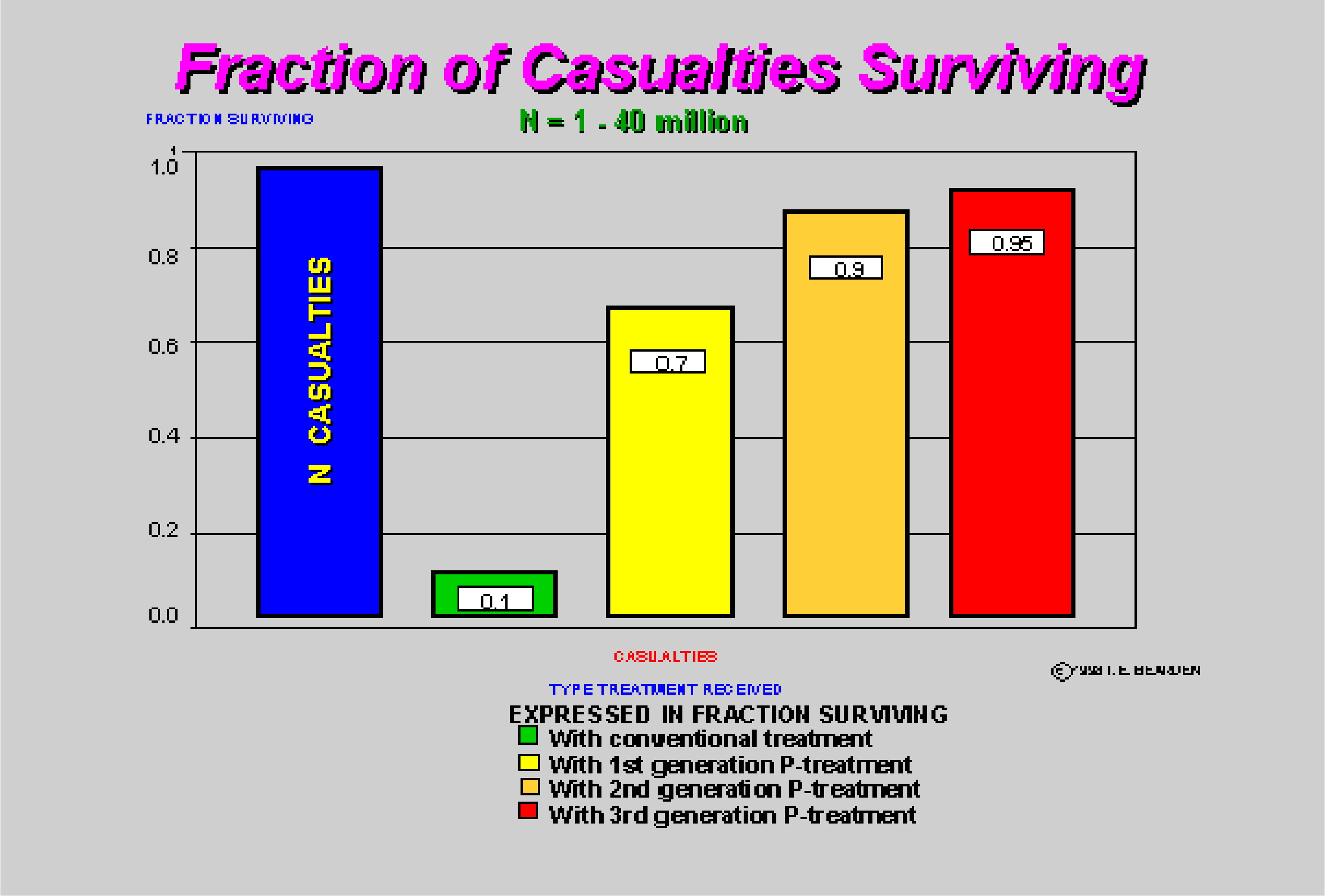

Published by Lotus® Freelance Graphics®

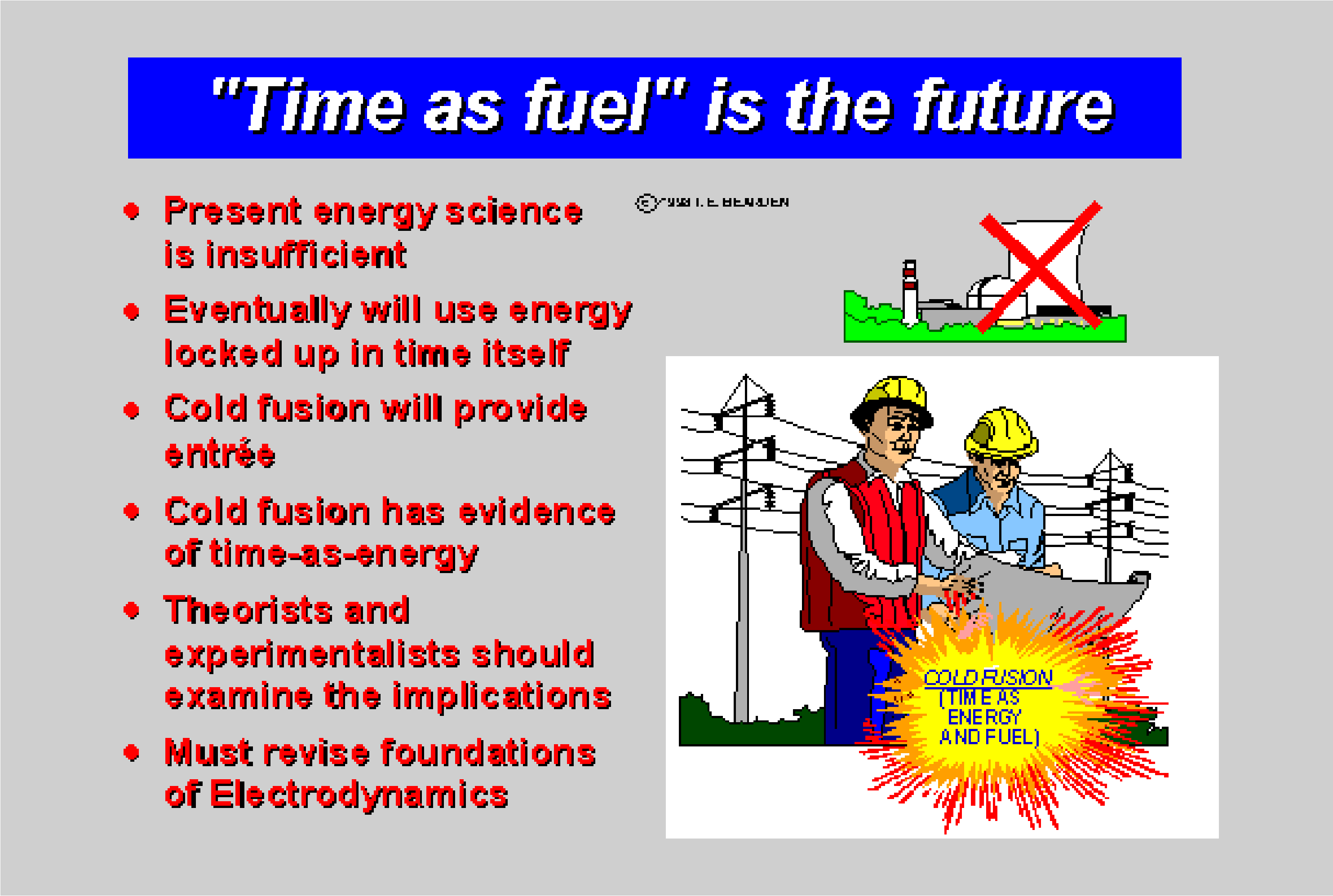

"Time as fuel" is the future

- Present energy science is insufficient
- Eventually will use energy locked up in time itself
- Cold fusion will provide entrée
- Cold fusion has evidence of time-as-energy
- Theorists and experimentalists should examine the implications
- Must revise foundations of Electrodynamics

©1998 T. E. BEARDEN

COLD FUSION
(TIME AS
ENERGY
AND FUEL)

Published by Lotus® Freelance Graphics®

A National Report on America's Energy Crisis

Remarks by U. S. Secretary of Energy Spencer Abraham

U. S. Chamber of Commerce, National Energy Summit March 19, 2001

Opening Remarks

I would like to congratulate the U. S. Chamber of Commerce for convening this two-day National Energy Summit and I appreciate your invitation to participate. As we all know, the topic of energy is as timely as the day's headlines. In just my first eight weeks as Energy Secretary, California has faced almost daily power alerts. Rising natural gas prices have punished consumers with bills that, in some cases, are double or triple last winter's. And forecasts for this summer suggest the possibility of rolling blackouts.

The good news is that America's energy problems can be solved. The bad news is that the situation in California is not isolated, it is not temporary, and it will not fix itself.

America faces a major energy supply crisis over the next two decades. The failure to meet this challenge will threaten our nation's economic prosperity, compromise our national security, and literally alter the way we live our lives.

America has heard these dire warnings before - in fact, they seem to be a recurring theme in our nation's energy discussion, almost since "Colonel" Drake made the first oil strike near Titusville, Pennsylvania in 1859. "The amazing exhibition of oil," advised the State Geologist of Pennsylvania, "[is] a temporary and vanishing phenomenon-one which young men will live to see come to its natural end. " That was in 1885.

Around the same time, John Archbold - who succeeded John D. Rockefeller as head of Standard Oil - joked about the prospects for oil discoveries in Oklahoma: "Are you crazy?" Archbold said. "I'll drink every gallon produced west of the Mississippi. "

I don't know if anyone ever called Mr. Archbold on that pledge. But for whatever reason, in 1914 the U. S. Geological Survey predicted that the U. S. would soon exhaust its available oil supplies. They issued the same warning in 1926. And again in 1939. And in 1949.

All of these warnings have proven false. Despite all these expert predictions, the world has not run out of oil or other resources. And yet here we are, faced with the most serious energy shortage since the days of oil embargoes and gas lines.

My point is, America's current energy supply crisis is not due to some inevitable neo-Malthusian depletion of resources. The United States - and our North American and hemispheric neighbors-are blessed with a rich abundance of natural resources. It's political leadership that has been scarce.

For the past eight years, Washington sat on the sidelines as our nation's energy needs mounted. During the 1990s, the Clinton Administration employed a policy of taxing demand, limiting supply, and ignoring the rapidly expanding needs of the future.

Their energy strategy boiled down to: you can't find it . . . you can't transport it . . . and even if you get it, we don't want you to use it.

Through neglect or complacency or ideology, this approach has led us to the crisis we face today.

The Bush Administration is fully prepared to respond to the broad set of challenges we inherited. But we must be candid with the American people about the magnitude of the problem - which is what I'd like to talk about today.

America's Energy Supply Challenges

Three overriding facts starkly define the challenge of America's energy needs over the next two decades:

First, demand for energy is rising across the board, but particularly for natural gas and electricity;

Second, supplies are being limited by a regulatory structure that, in many respects, has failed to keep pace with advances in technology and an uncertain political environment that often discourages investment in desperately needed facilities;

And third, our energy infrastructure - that network of the generators, transmission lines, refineries and pipelines that convert raw resources into usable fuel - is woefully antiquated and inadequate to meet our future needs.

Unless these challenges are addressed, America's energy supply will be continually at risk . . . our citizens will encounter blackouts and other lifestyle-altering disruptions . . . and our economy will be hobbled by rising energy prices. Let me briefly outline some of the major issues on the horizon:

Oil: Rising Consumption, Accelerating Dependence

In the next 20 years, according to estimates by the Energy Information Administration, America's demand for oil is projected to increase by 33 percent. Yet as consumption surges, U. S. production continues to drop precipitously. We now produce 39 percent less oil than we did in 1970, losing nearly 4 million barrels a day in the process. And unless energy policy is changed, production will slip further - to just 5.1 million barrels per day by 2020 - down from a high of 9. 4 million a day 30 years ago. This widening gap between demand and domestic supply will make us increasingly dependent upon foreign imports. Back in 1973-at the height of the oil crisis - America imported just 36 percent of its oil from abroad.

Today, we import 54 percent. And, if we allow this trend to continue, we will soon be forced to look abroad for some 64 percent of our oil. This will put more power in the hands of foreign suppliers - power they are not reluctant to use, as we just saw when the OPEC cartel decided to reduce oil output by one million barrels a day.

While this administration does not agree with OPEC's decision, that decision demonstrates the importance of increasing America's production of oil. Securing an affordable, reliable and adequate supply of crude is a critical challenge. But it is only half the oil story.

Since 1980, the number of American refineries has been cut in half. There hasn't been a new refinery built in the United States in over 25 years. New regulatory interpretations limit the ability of existing refineries to expand capacity. Add to that regulations that require the production of more than 15 different types of gasoline-and you have a refining industry strained to capacity, leaving us dangerously vulnerable to regional supply disruptions and price spikes.

Refineries are so constrained that when President Clinton made the politically symbolic gesture of releasing 30 million barrels of oil from the Strategic Petroleum Reserve last fall, that oil had to be shipped overseas to be refined.

Natural Gas: Rapidly Rising Demand; Constraints on Supply

Many of the same issues confront the future of natural gas. America's demand for natural gas is projected to rise even more rapidly than oil. If Department of Energy projections are correct, by 2020 Americans will consume 62 percent more natural gas than we do today. More than 9 out of 10 of the announced new electric generating plants will be fired by natural gas.

This approaching wave of new demand begs the question: Do we have an energy policy and a regulatory structure capable of meeting our natural gas needs?

Consider just a few constraints in this market.

Right now, an estimated 40 percent of potential gas resources in the United States are on federal lands that are either closed to exploration or covered by severe restrictions. The last lease sale in the some areas of the Gulf of Mexico was more than a decade ago. New discoveries of natural gas in the United States have fallen for three straight years, creating increasing pressure for more imports.

The notion that we can rely so heavily on natural gas . . . maintain severe restrictions on exploration . . . and still enjoy low prices is a dangerous assumption.

Even if we find the supplies, moving that gas to market will require an additional 38,000 miles of transmission pipeline and 255,000 miles of distribution lines-at an estimated cost of $120-$150 billion.

Today's pipeline system can hardly handle the supplies we know exist. Alaska's Prudhoe Bay, for example, produces about 8 billion cubic feet of natural gas a day - approximately 13 percent of America's daily consumption.

But that gas never reaches the market. Instead, it is just pumped back into the ground, waiting until a pipeline is built to connect the Alaska fields to the U.S./Canada distribution system.

Electricity: Powering the Information Economy

As everyone knows, we also face a real challenge in generating enough electricity to light out homes and run our businesses. Over the next 20 years, the Department of Energy estimates that electricity demand in the United States will increase by 45 percent. That rising growth rate will require the construction of over 1,300 new power plants-about 65 every year.

Yet, the last time we added that much power was 1985.

Furthermore, there is reason to believe that this could turn out to be a conservative estimate. During the 1990s, electricity consumption far outstripped projections, driven by the energy-hungry information economy. Some experts calculate that the demands of the Internet already consume some 8-13 percent of electricity. If demand grows at just the same pace as during the last decade, we'll need nearly 1,900 new plants by 2020-or more than 90 every year-just to keep pace.

Hundreds of new generating plants will place even greater pressure on our already strained and aging power grid. America's network of transmission lines, substations and transformers was built when utilities were tightly regulated monopolies providing service to assigned regions. Interconnections between suppliers were strictly an emergency backup measure to guard against rare service interruptions. The system was simply not designed for long-haul swapping of power in a highly competitive market.

Consumers are already feeling the impact of a transmission systems stressed by rising demand. Transmission bottlenecks contributed to the blackouts that have swept through California and to price spikes in New York City last summer that cost consumers an estimated $100 million.

Coal, Nuclear, and Hydro-Electric Power

Coal has historically been American's number one source for affordable electricity; it currently powers half of America's electricity generators. And at today's recovery rates, our nation has enough coal to keep those plants running for the next 250 years.

Coal generators have already been called upon to make broad reductions in emissions. The Bush Administration supports those efforts-and we will back it up with greater incentives for investment in clean coal technology.

But the administration will not regulate coal out of existence . . . and we will not support measures that will threaten electricity supplies and significantly raise electricity prices. President Bush made the right decision last week not to impose new federal mandates on the emissions of carbon dioxide. If America is to have reliable electricity over the next 20 years, coal must continue to play a major role.

But, coal is not the only energy source facing an uncertain future. There hasn't been a new nuclear power plant permit granted since 1979. Many of the 103 existing nuclear plants are not even expected to file for a renewal of their licenses as they expire over the next 15 years.

Even hydroelectric power generation is expected to fall sharply.

Re-licensing a hydro facility can take a decade or more and cost millions. And now, even though consumers are faced with potential blackouts and chronic electricity shortages in the West, activists and some political leaders want to breach one or more of the four federal dams on the Snake River to help young salmon, on their trek to the sea.

The Dangers of Complacency

What are the dangers of complacency in light of these challenges? How does it all add up for our economy and our citizens?

This nation's last three recessions have all been tied to rising energy prices - and there is strong evidence that the latest crisis is already having a negative effect.

The National Association of Manufacturers estimates that soaring fuel prices between 1999 and 2000 cost the U.S. economy more than $115 billion-shaving a full percentage point off our Gross Domestic Product. A January survey of its 5,500 members reveals that nearly one quarter were forced to curtail operations.

During a two-week period this past January, Californians lost an estimated $2.3 billion in wages, sales and productivity. Layoffs are already hitting workers in the West as companies shift production to states with more reliable energy sources. Then there are the jobs that will never be created.

Intel's CEO Craig Barrett announced that the world's leading chipmaker won't be expanding in California: "As long as California is a Third World country," Barrett said, "we won't build $2 billion manufacturing plants here."

The Food and Agricultural Policy Institute reports that farmers are likely to see their income drop 20 percent over the next two years due to higher energy costs.

Rising energy costs are hitting every family's checkbook, primarily affecting those who can afford it least. Gas bills for many homeowners in the Washington, DC area more than tripled this year. Some residents are reporting that their heating bills are higher than their food bills this winter.

The power crisis isn't just pinching our wallets, it's changing the way we live our lives. In California, power outages have shut down traffic lights, darkened schools and closed businesses. The governor has ordered local police to patrol the streets-not for criminals, but to make sure businesses keep their lights dimmed.

But California is not the only state facing a mismatch between supply and demand. With electricity shortages predicted for New York City and Long Island this summer . . . low capacity margins threatening electricity reliability in the Midwest, Southeast and Northern Plains states . . . and strained refinery capacity in the Midwest, Americans across the nation are feeling the energy squeeze.

The Need For A National Energy Policy

Rising demand . . . tightening supplies . . . an aging power infrastructure . . . a decade of neglect from Washington: These are the trends that define America's emerging energy needs.

President Bush has committed this administration to meeting these challenges - a job that begins with the urgent task of developing and implementing a long-term national energy policy.

To accomplish this, President Bush created an Energy Task Force headed by Vice President Cheney. He has asked us to define a clear strategy - a strategy that will allow environmentally responsible exploration and recovery of our domestic resources . . . enhance our commitment to conservation and energy efficiency . . . and encourage investment in new technology to further the development of renewable energy sources.

I wish I could say that the energy crisis now sweeping the nation has shocked the political system into action. But like other political discussions in recent years, the debate over energy seems as deeply polarized as ever.

On one end of the spectrum, some activists propose what amounts to a "zero tolerance" policy toward exploration and cling to the quixotic idea that new, undiscovered sources will somehow allow us to meet our energy needs. On the other end, some advocates place an almost limitless faith in special tax breaks for this favored activity or subsidies for that preferred industry.

The two extremes in our energy debate are founded on several enduring myths - myths that today conspire to block any true advance toward a rational and stable energy policy in the United States. Here are a few of the more prominent ones.

Myth Number 1: It is impossible to balance energy exploration and environmental protection.

Advances in technology have brought us a long way from the days when wildcatters punched holes in the ground based on the hunch they might hit a gusher. But from a regulatory standpoint, our view of oil and gas exploration hasn't changed much since we saw Jed Clampett strike "black gold" and split for Beverly Hills.

Today, satellites and computers are the tools of choice in the exploration business. Geologists can bounce acoustic and electrical vibrations off the earth's inner depths, gather the resulting mass of data into powerful computers and then create three dimensional and even four dimensional maps of resource fields miles below the surface.

Armed with these pinpoint accurate images, companies employ advanced equipment to drill vertically, horizontally and around corners-allowing us to access supplies from previously unimaginable depths, up to six miles away.

The marriage of oil and gas exploration with cutting-edge technology means fewer rigs, fewer roads and fewer pipelines. Drilling operations that required 65 acres in the 1970s need only 10 acres today. Technological improvements in just the past 15 years have generated success rate increases of 50 percent. America's national energy policy must reflect these staggering advances that have revolutionized the way we develop our resources.

Myth Number 2: All our current problems are due to an energy industry that is engaged in a massive conspiracy to gouge consumers by limiting supply to drive up prices.

This myth has been punctuated by calls for investigations into everything from last summer's Midwest gasoline price spikes, to recent allegations that power generators in the West have been withholding electricity. We have a fair and objective process for judging these claims-and action will be taken when it is merited. Over the past two weeks, for example, the Federal Energy Regulatory Commission ordered power companies to rebate some $124 million to California utilities. Meanwhile, the Federal Trade Commission recently cleared gasoline suppliers of all charges relating to last summer's price increases.

But charges of price gouging largely miss the point. There is no magic source of supply; no hidden pool of energy that can be turned on and off like a faucet. California-and other power-strapped states-will never solve the power crises they confront until they resolve the conflict between demand and supply.

Earlier this year, one company proposed building a $400 million power plant in California that would have provided enough additional electricity to light 600,000 homes in energy-starved Silicon Valley. The company pledged to plant 800 new trees to beautify the area. They proposed cloaking the power station in a brick facade to make it essentially indistinguishable from a high-rent office complex. They even promised to help maintain the local habitat for the endangered bay checkerspoon butterfly.

Their environmentally-sensitive plans won the support of the Sierra Club, the American Lung Association and the NAACP. But city officials voted unanimously -- 11-0 -- to reject the plan. In an editorial, the local paper called this move "Dumb and Dumber."

Meanwhile, further south, plans to build a 550-megawatt gas-fired generator in a Los Angeles suburb were scrapped after residents voted 2:1 against the project. The local mayor added a much-needed dose of reason and maturity to the debate-by launching a hunger strike in opposition to the plant.

In California, workers are being laid off, companies are leaving the state, farmers and small businesses are losing millions, consumers are threatened with rolling blackouts, but local officials reject power plants with little regard for the consequences. Is it really any mystery why there hasn't been a single new power plant built in California in the last decade?

Myth Number 3: The Bush energy plan is focused almost exclusively on opening the Arctic National Wildlife Refuge (ANWR) to exploration-a move that would buy us only about 6 months worth of American consumption while destroying a pristine natural wilderness, not to mention disrupting the breeding ground of the Porcupine Caribou.

Let's separate fact from fiction when it comes to ANWR.

First, according to estimates by the U. S. Geological Survey, ANWR holds between 5. 7 to 16 billion barrels of recoverable reserves-with a mean estimate of 10. 4 billion barrels. And that assumes the use of drilling technology now nearly a decade old. This represents more than 300 times the amount of the oil President Clinton released from the Strategic Petroleum Reserve last fall. And based on December 2000 figures, it would free us from about 54 years of oil imports from Saddam Hussein and Iraq.

Second, exploration would impact only about 2,000 acres out of more than 19 million. To put that in perspective, the massive Arctic National Wildlife Refuge is about the same size as the entire State of South Carolina; the two thousand acres that would be affected is less than half the size of Dulles airport.

And as for the caribou, the herd in the Prudhoe Bay area grew more than 9-fold over the past 20 years to an estimated 28,000 in 2000-seemingly irrefutable evidence that caribou mating and oil exploration can peacefully coexist.

The decision to open a small portion of ANWR should be made on the merits. But it should not be made on the mistaken assumption that opening ANWR will allow us to produce our way to full energy independence.

America first became a net importer of energy in the 1950s - and our economy will continue to depend, in part, on imported oil. However, closing off virtually every available new source of domestic supply, enhances the leverage and power of an oil cartel that cannot be relied on to put America's interests first. While the resources of ANWR won't make us energy independent, they will help increase America's energy security by ensuring a more diverse supply of oil.

Myth Number 4: Government subsidies and tax breaks are the best way to encourage new exploration and production of energy.

This administration will continue to support funding for energy research and development initiatives. But capital is best allocated to its highest uses through the workings of the free market, not manipulations of the tax code. Government regulatory policy should not be aimed at picking winners and losers in any market, including energy. Neither should tax policy.

Myth Number 5: We can forego traditional sources and instead meet rising energy demand by harnessing wind, geothermal, solar and other forms of renewable power.

Excluding hydro-power, renewable sources currently generate about 2 percent of America's electricity. Billions have been invested in developing renewable energy-and will continue to be invested under the Bush Administration. But renewables have yet to overcome the economic advantages of conventional energy sources.

Even with promising advances in research and development, renewables will only provide, according to Energy Information Administration estimates, about 6 percent of our total electricity consumption by 2020. Even if renewables exceed our most optimistic expectations, they would still supply only a fraction of our needs over the next 20 years.

Myth Number 6: Price controls are the answer to today's energy crisis.

Memories are short, aren't they. So let me remind everyone. America imposed price controls on oil and gas in the 1970s. They were an utter failure. They led to shortages and rationing and the idea that America was gripped by malaise.

Let me be clear: The Bush Administration does not support price controls. Price controls on electricity will lead to more blackouts. Price controls on gasoline will lead to gas lines. Price controls will deepen America's energy crisis, because they won't reduce demand, but they will cripple incentives for desperately needed new investments in energy supply.

Charting a New Policy Course

The challenges are formidable . . . the warning signs are obvious . . . but I am optimistic because I know this administration's commitment is equal to the task.

Our national energy policy will be comprehensive. It will reach across every department that touches the energy marketplace-from the Interior Department and the EPA to the Transportation Department and the DOE.

Our national energy policy will be hemispheric. It will be based on the understanding that our policy cannot stand in isolation from our neighbors throughout the Americas.

Our national energy policy will stress the need to diversify America's energy supply. It will be founded on the understanding that diversity of supply means security of supply . . . and that a broad mix of supply options-from coal to windmills, nuclear to natural gas-will help protect consumers against price spikes and supply disruptions.

And our national energy policy will be balanced. It will leapfrog the myths that stifle change-rejecting the notion that there is no middle ground between environmental protection, regardless of the cost and energy exploration, regardless of the impact.

Soon we will deliver our recommendations to President Bush. Later, we will introduce legislation aimed at winning bipartisan support for a national energy policy that matches the magnitude of the challenge. I am hopeful that men and women of good will-from both ends of the political spectrum . . . from environmental organizations to industry groups-will then come together and transcend the stale debate that has characterized energy policy in recent years.

About 150 years ago, America faced a vastly different energy crisis.

Supplies of whale oil were becoming more and more scarce. Few could afford to pay for the luxury of this or other costly methods of illumination. Sure, crude oil was available. In those days it was soaked up with rags, wrung out into small vials, and then sold as a treatment for toothaches . . . until an entrepreneur lined up an investor and a chemist and launched an energy revolution that would light the world.

In America, resources become scarce only when our imagination languishes. By engaging that imagination, I am confident we can meet the challenges of today. If complacency yields to action. If we resolve to strike a rational balance between our energy needs and our environmental concerns. And if a national energy policy becomes an urgent priority.

Thank you.

Report by Energy Secretary Spencer Abraham

Contents

1. Report by Energy Secretary Spencer Abraham
2. National energy policy for 8 years
3. Energy supply challenges
4. Oil: Rising consumption and accelerating dependence
5. Oil: Rising consumption and accelerating dependence
6. Natural Gas: Rapidly rising demand, supply constraints
7. Natural Gas: Rapidly rising demand, supply constraints
8. Electricity: Needs, problems
9. The aging power grid
10. Coal fired electrical power
11. Nuclear power plants
12. Hydroelectric power generation
13. Dangers of complacency
14. Dangers of complacency (2)
15. Need for national energy policy
16. Extremes of the energy positions
17. Six prominent myths, none of which is true
18. Can we balance exploration and environmental protection?
19. Is there a price gouge conspiracy?
20. Is there a price gouge conspiracy? (2)
21. Is there a price gouge conspiracy? (3)
22. Arctic National Wildlife Refuge Exploration
23. Arctic National Wildlife Refuge Exploration (2)
24. Our increasing oil imports
25. Government subsidies and tax breaks
26. Renewable energy sources
27. Price controls
28. Charting a new national policy

---

29. **Commentary by Tom Bearden, Ph.D.**
30. Commentary
31. Caution: It is stressed that
32. An example of the time problem

33. The pending disaster
34. The bottom line:
35. Here are some events etc.
36. Lowered energy requirements for appliances
37. Easing air conditioner and heat pump standards
38. Navarro's Estimate*
39. Factors driving comparisons
40. Prices likely per kilowatt-hour
41. Covering debt already incurred
42. Deteriorating situation, rising prices
43. Other Factors
44. Paying the piper: California's bond rating reduced
45. Associated Press: Gasoline supplies are critical
46. A peculiar problem
47. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (1)
48. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (2)
49. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (3)
50. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (4)
51. MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (5)
52. Negentropic Engineering
53. Rigorous "out-of-the-box" work
54. Rigorous "out-of-the-box" work (2): (Bedini "negative resistor" in a battery)*
55. Initiating a Battery to COP>1.0
56. Secretary alone cannot solve it
57. What is desperately needed
58. Why it can be done in time
59. Cautions!
60. The scientists have failed us
61. 136-yr-old EM Model1
62. All EM from Time Domain1
63. Astonishing but true
64. With plentiful, cheap, clean electrical energy:
65. Recommendations to DOE
66. In conclusion:

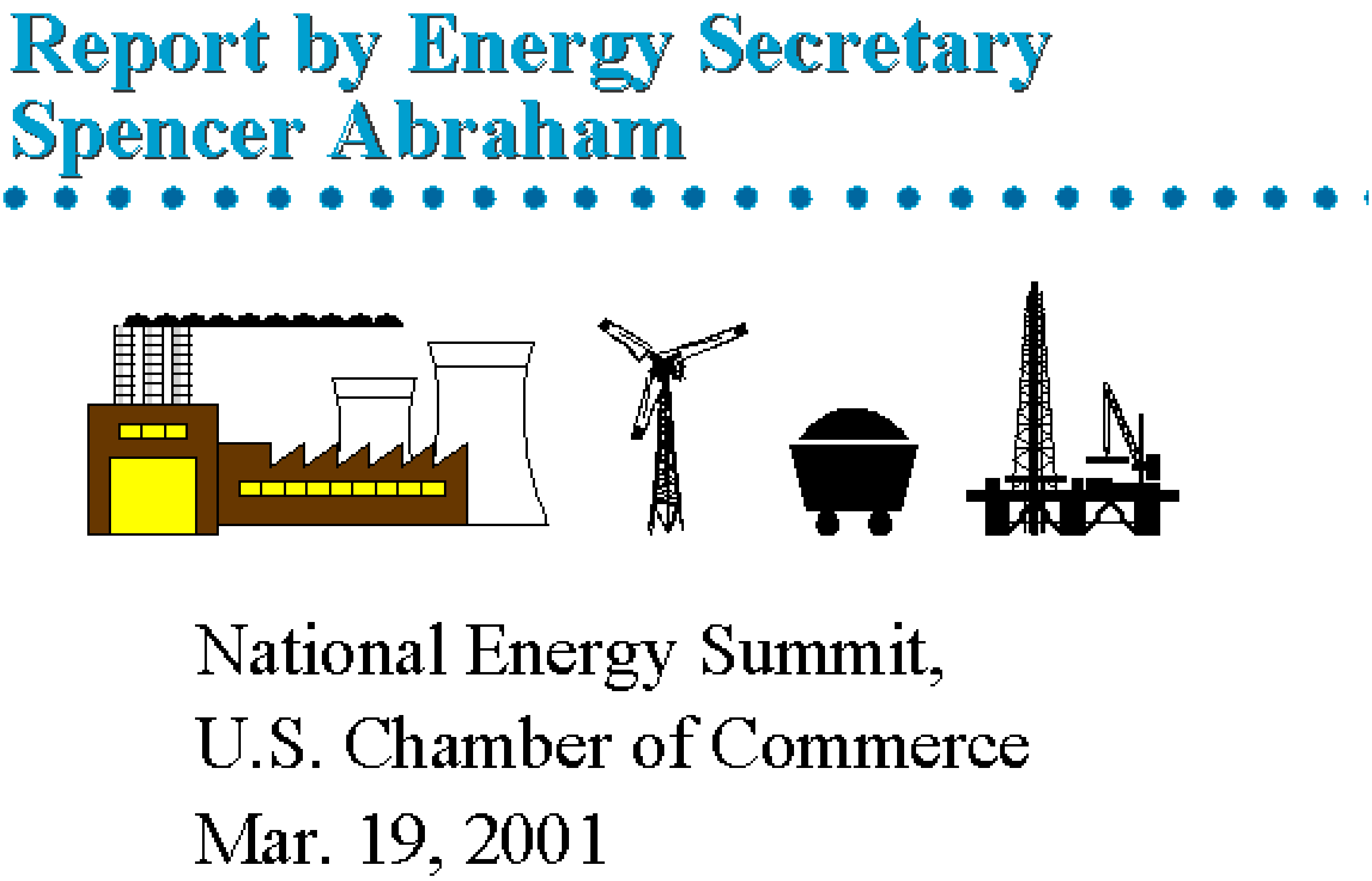

Report by Energy Secretary Spencer Abraham

National Energy Summit,
U.S. Chamber of Commerce
Mar. 19, 2001

Published by Lotus® Freelance Graphics®

National energy policy for 8 years

- The national policy imposed the following:
  - Taxed energy demand
  - Limited the energy supplies
  - Ignored expanding energy needs
- Summary of the past policy on energy:
  - You cannot find it
  - You cannot transport it
  - If you get it, we don't want you to use it
- *Good News:* Crisis can be solved
- *Bad News:* Not temporary; will not fix itself

Published by Lotus® Freelance Graphics®

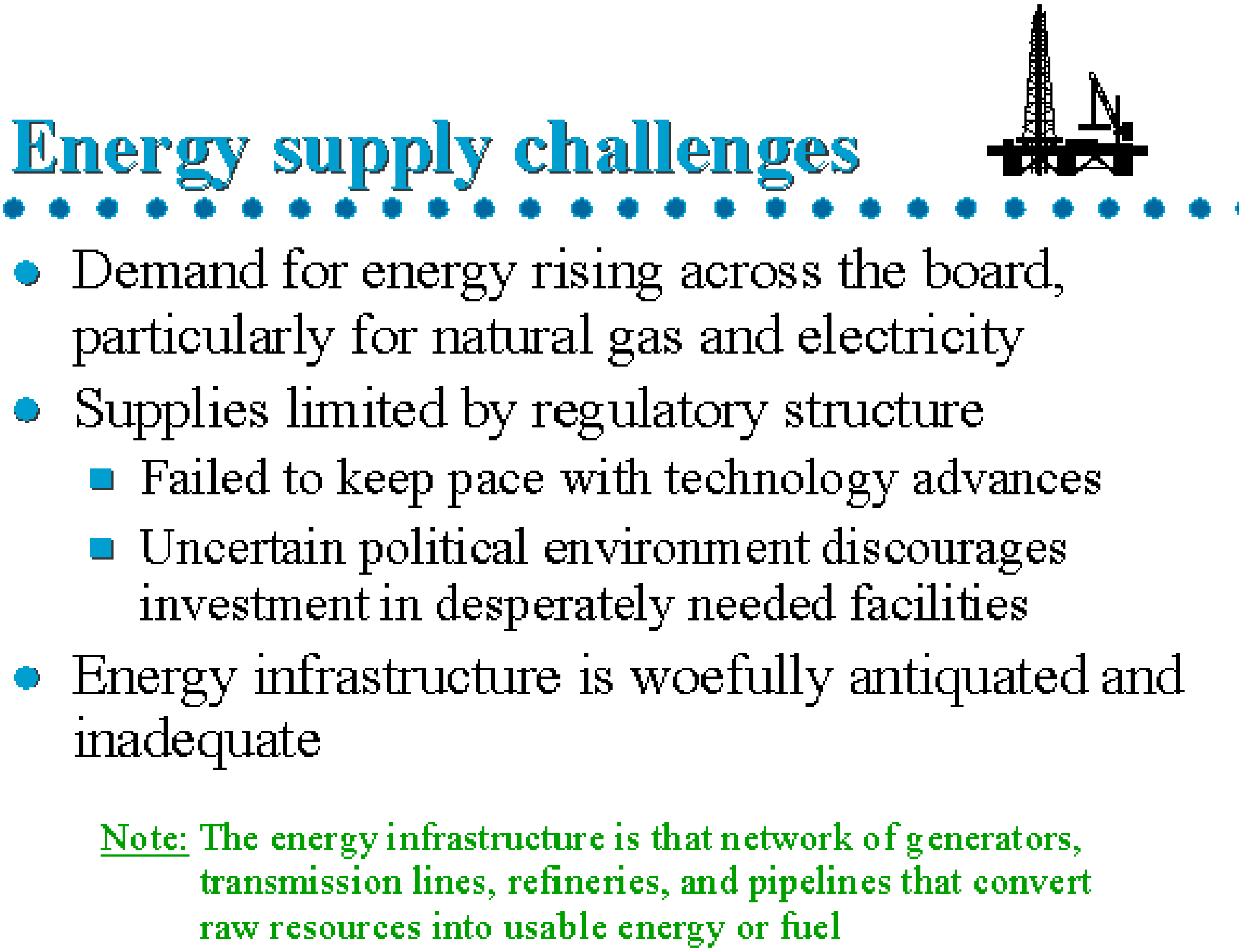

Energy supply challenges

- Demand for energy rising across the board, particularly for natural gas and electricity
- Supplies limited by regulatory structure
  - Failed to keep pace with technology advances
  - Uncertain political environment discourages investment in desperately needed facilities
- Energy infrastructure is woefully antiquated and inadequate

Note: The energy infrastructure is that network of generators, transmission lines, refineries, and pipelines that convert raw resources into usable energy or fuel

Published by Lotus® Freelance Graphics®

⬅ ➡

Oil: Rising consumption and accelerating dependence

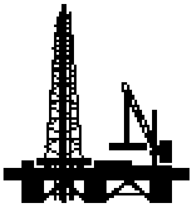

- In U.S., will increase 33 percent in 20 years
- U.S. production continues to drop
  - Produce 39% less now than in 1970
  - At present reduction rate, production will decrease to 5.1 million barrels per day by 2020, down from 9.4 million barrels per day in 1970.
- Imports have steadily risen
  - From 36% in 1973 to 54% today
  - Will soon rise to 64% unless checked
  - Empowers foreign suppliers such as OPEC
- Assuring affordable, reliable, adequate supply of crude oil is a critical national challenge

Published by Lotus® Freelance Graphics®

Oil: Rising consumption and accelerating dependence

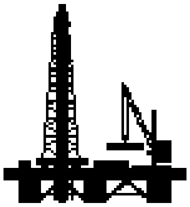

- We have been decreasing the infrastructure
  - Since 1980, U.S. refineries reduced by 50%
  - No new refinery built in over 25 years
  - New regulatory interpretations limit expansion of capacity of existing refineries
  - Regulations require producing more than 15 different types of gasoline
  - Refining industry strained to capacity
- E.g., Clinton symbolically released 30 million barrels of oil from the Strategic Petroleum Reserve
- *The oil had to be shipped overseas to be refined!*

Published by Lotus® Freelance Graphics®

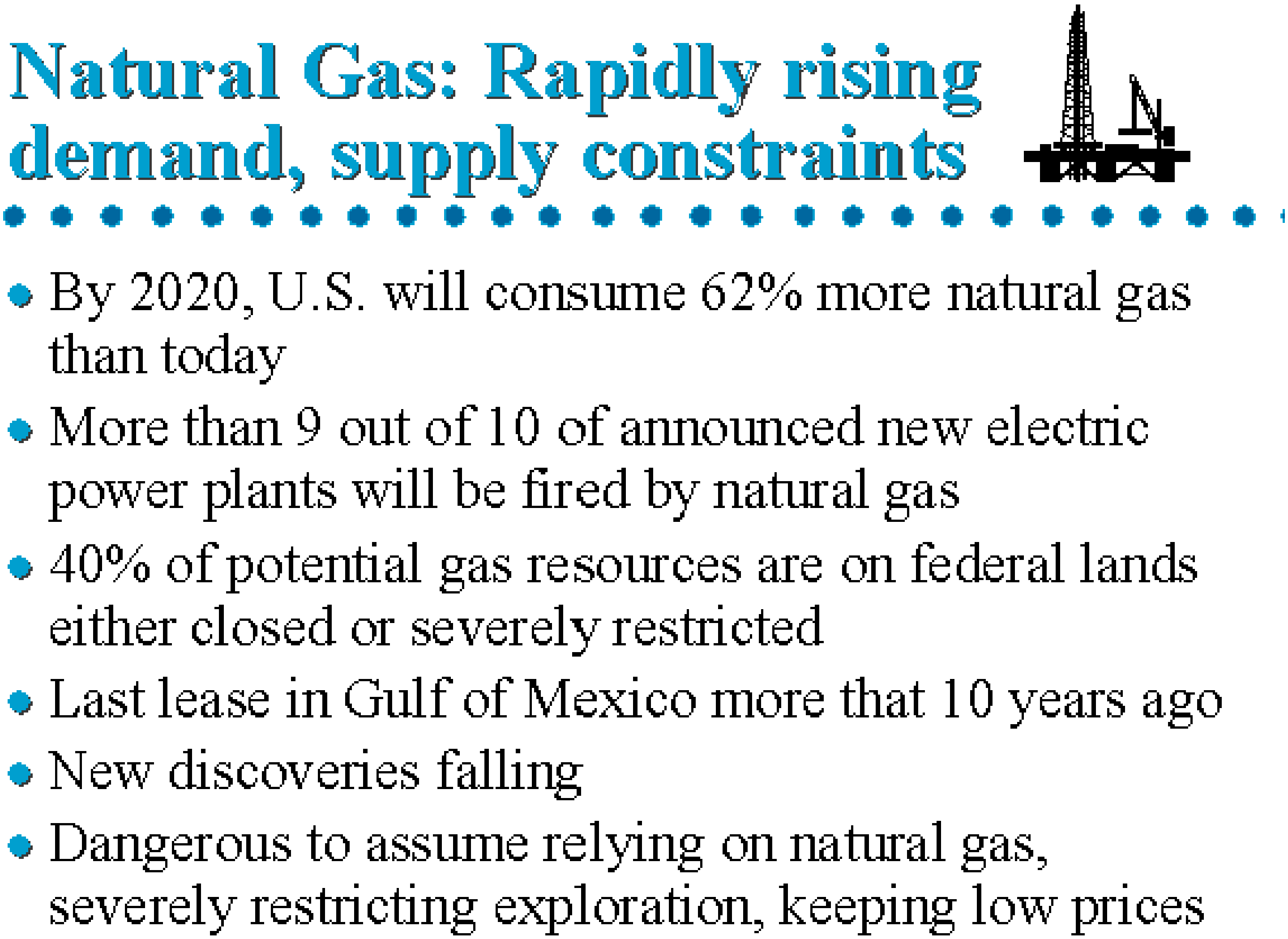

Natural Gas: Rapidly rising demand, supply constraints

- By 2020, U.S. will consume 62% more natural gas than today
- More than 9 out of 10 of announced new electric power plants will be fired by natural gas
- 40% of potential gas resources are on federal lands either closed or severely restricted
- Last lease in Gulf of Mexico more that 10 years ago
- New discoveries falling
- Dangerous to assume relying on natural gas, severely restricting exploration, keeping low prices

Published by Lotus® Freelance Graphics®

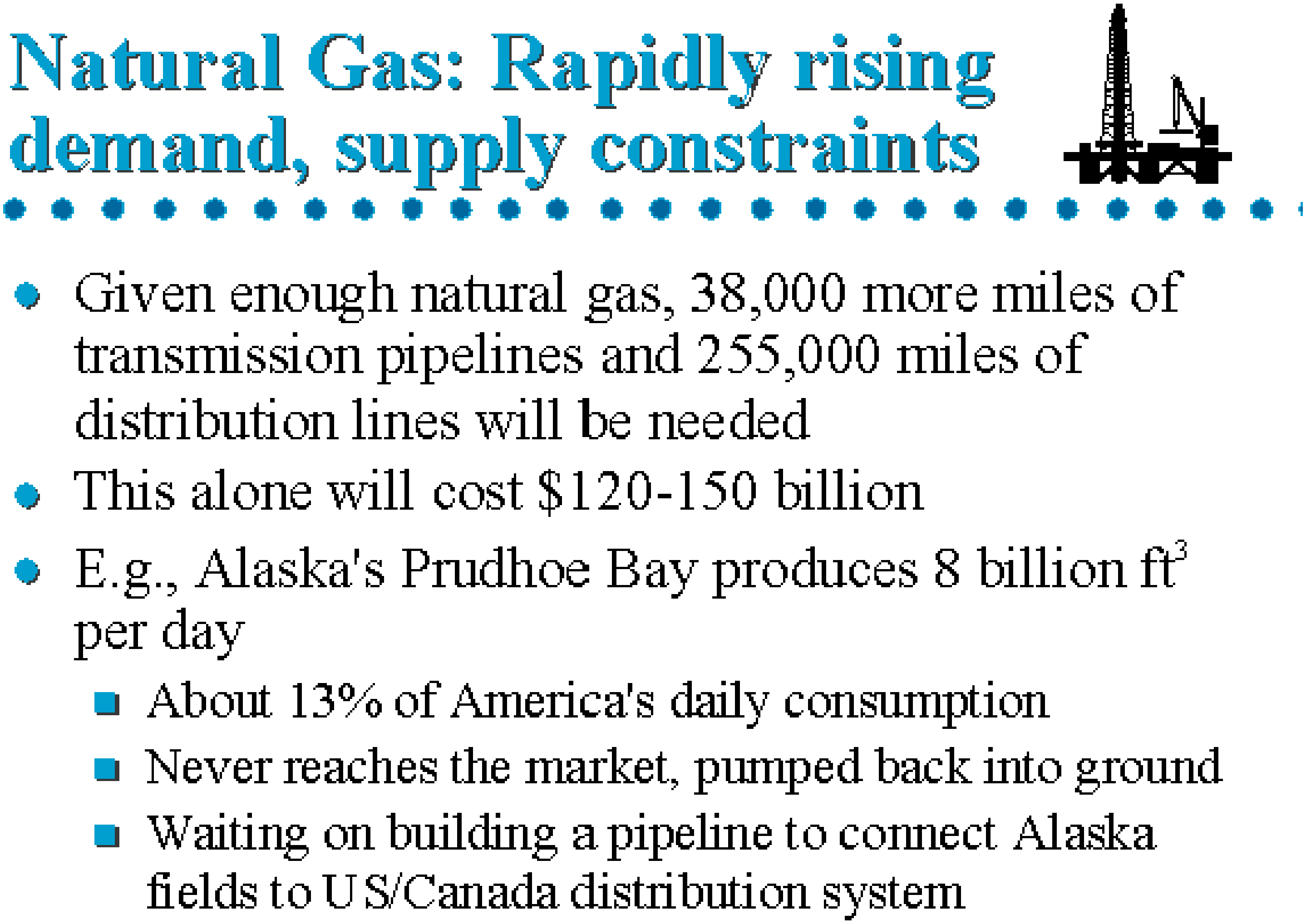

Natural Gas: Rapidly rising demand, supply constraints

- Given enough natural gas, 38,000 more miles of transmission pipelines and 255,000 miles of distribution lines will be needed
- This alone will cost $120-150 billion
- E.g., Alaska's Prudhoe Bay produces 8 billion ft^3 per day
  - About 13% of America's daily consumption
  - Never reaches the market, pumped back into ground
  - Waiting on building a pipeline to connect Alaska fields to US/Canada distribution system

Published by Lotus® Freelance Graphics®

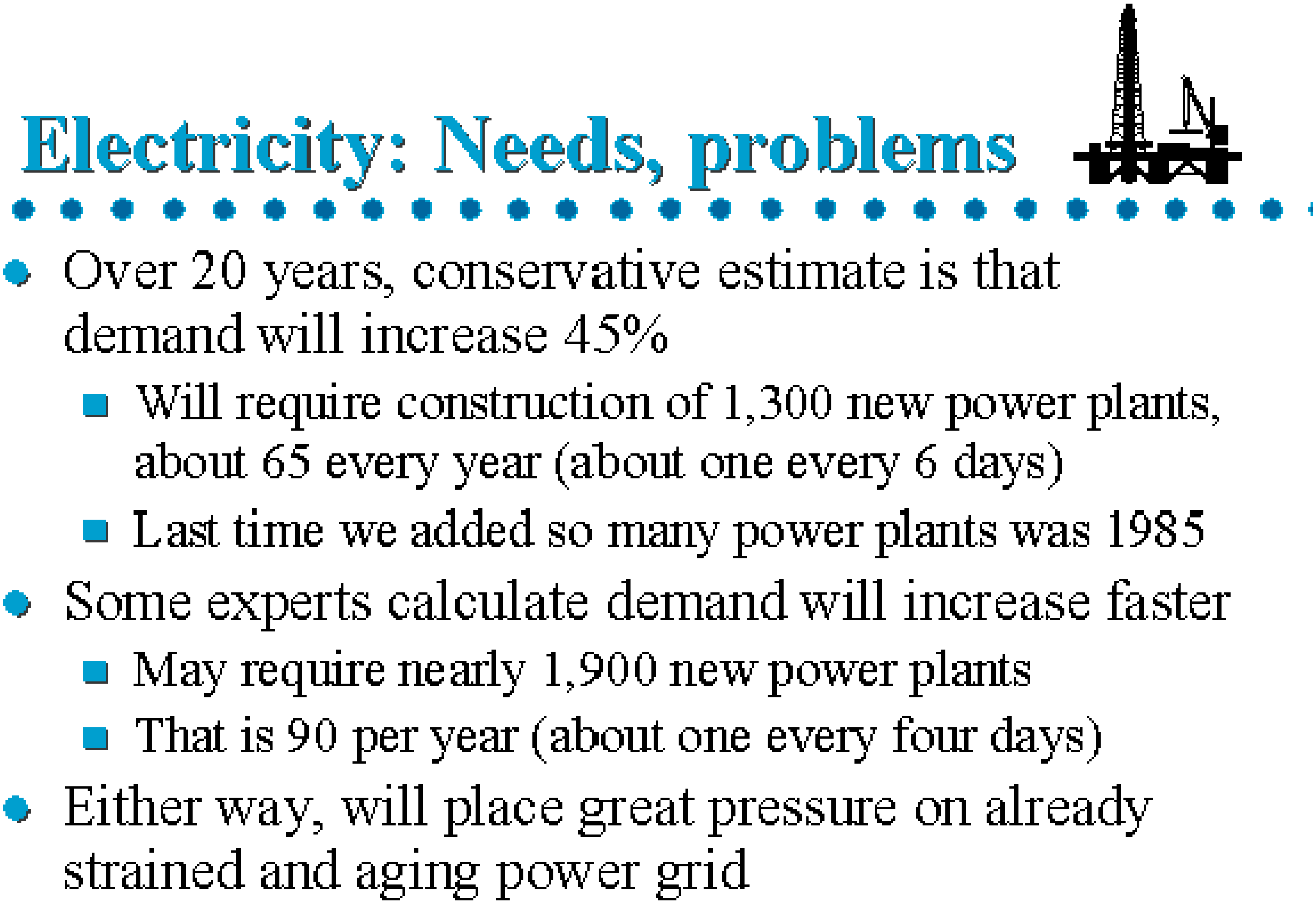

Electricity: Needs, problems

- Over 20 years, conservative estimate is that demand will increase 45%
    - Will require construction of 1,300 new power plants, about 65 every year (about one every 6 days)
    - Last time we added so many power plants was 1985
- Some experts calculate demand will increase faster
    - May require nearly 1,900 new power plants
    - That is 90 per year (about one every four days)
- Either way, will place great pressure on already strained and aging power grid

Published by Lotus® Freelance Graphics®

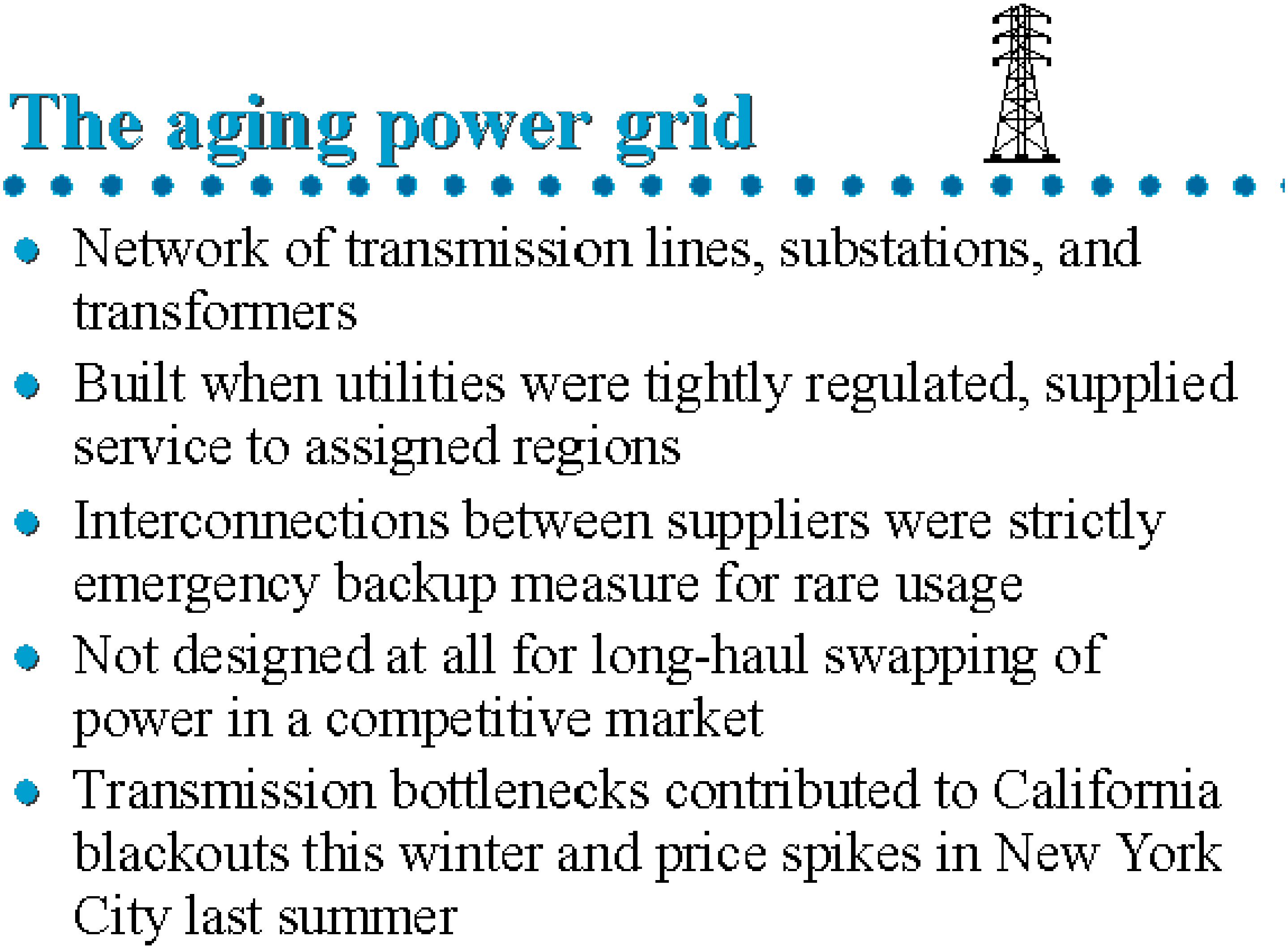

The aging power grid

- Network of transmission lines, substations, and transformers
- Built when utilities were tightly regulated, supplied service to assigned regions
- Interconnections between suppliers were strictly emergency backup measure for rare usage
- Not designed at all for long-haul swapping of power in a competitive market
- Transmission bottlenecks contributed to California blackouts this winter and price spikes in New York City last summer

Published by Lotus® Freelance Graphics®

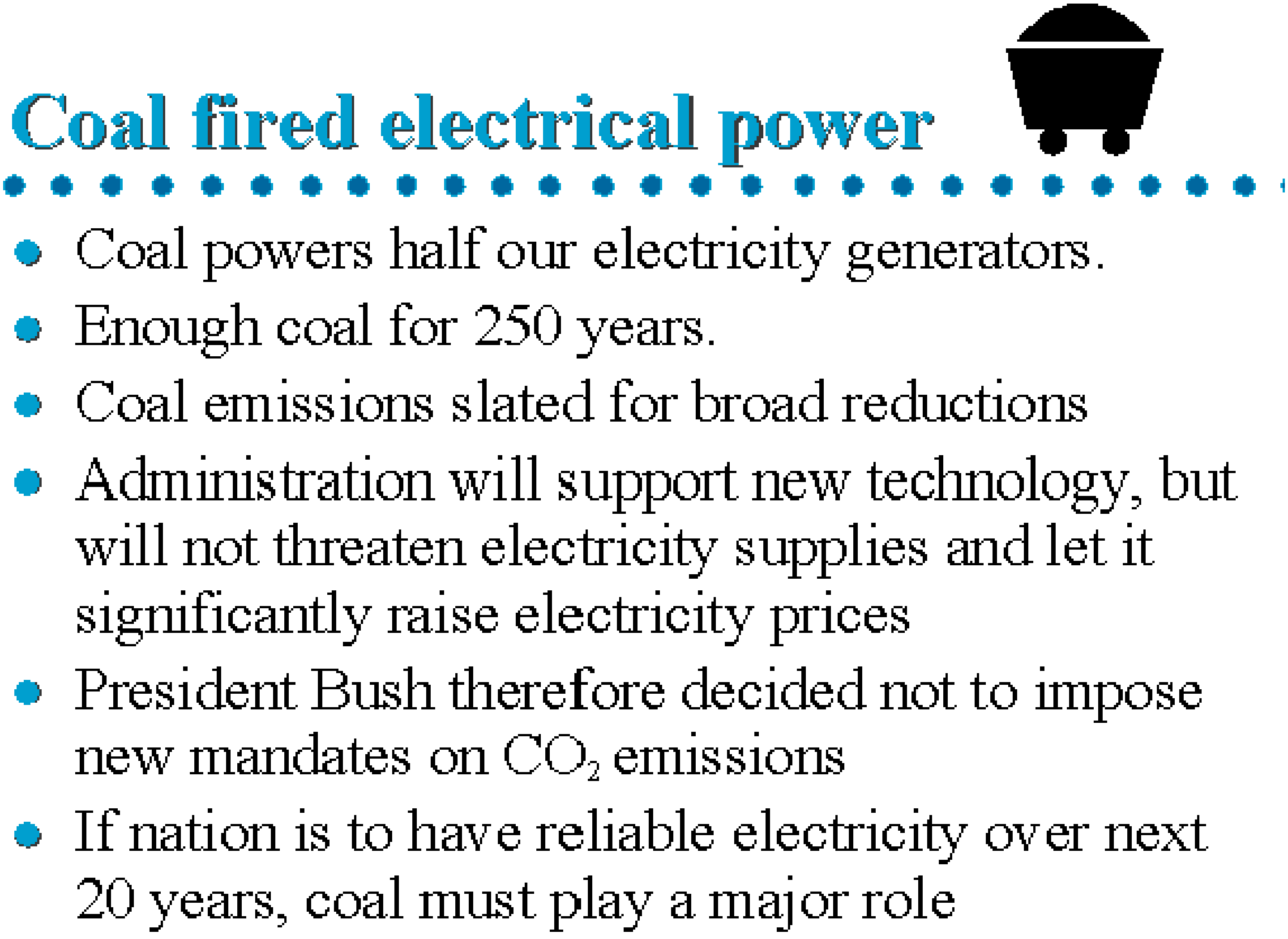

Coal fired electrical power

- Coal powers half our electricity generators.
- Enough coal for 250 years.
- Coal emissions slated for broad reductions
- Administration will support new technology, but will not threaten electricity supplies and let it significantly raise electricity prices
- President Bush therefore decided not to impose new mandates on CO_2 emissions
- If nation is to have reliable electricity over next 20 years, coal must play a major role

Published by Lotus® Freelance Graphics®

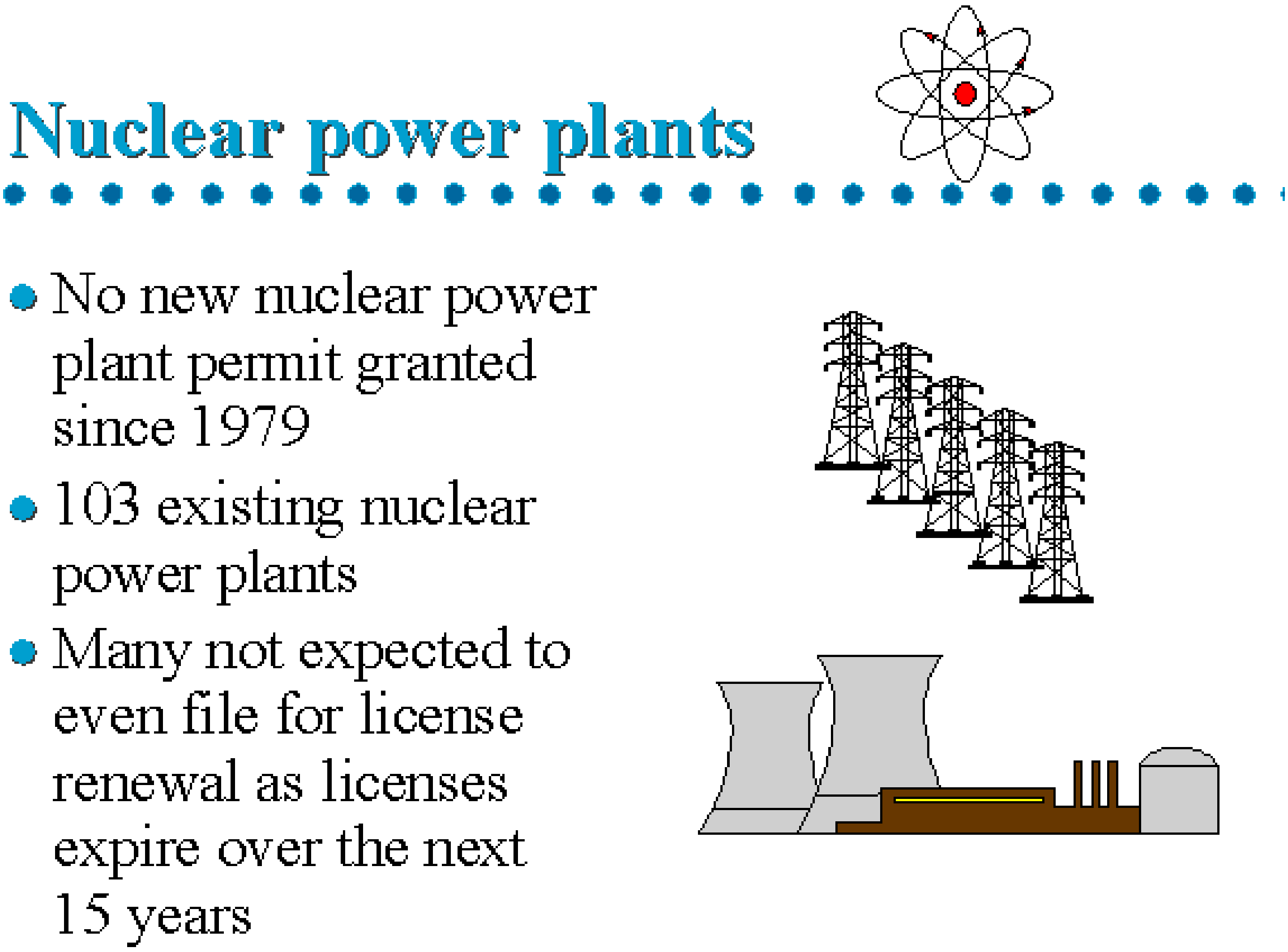

Nuclear power plants

- No new nuclear power plant permit granted since 1979
- 103 existing nuclear power plants
- Many not expected to even file for license renewal as licenses expire over the next 15 years

Published by Lotus® Freelance Graphics®

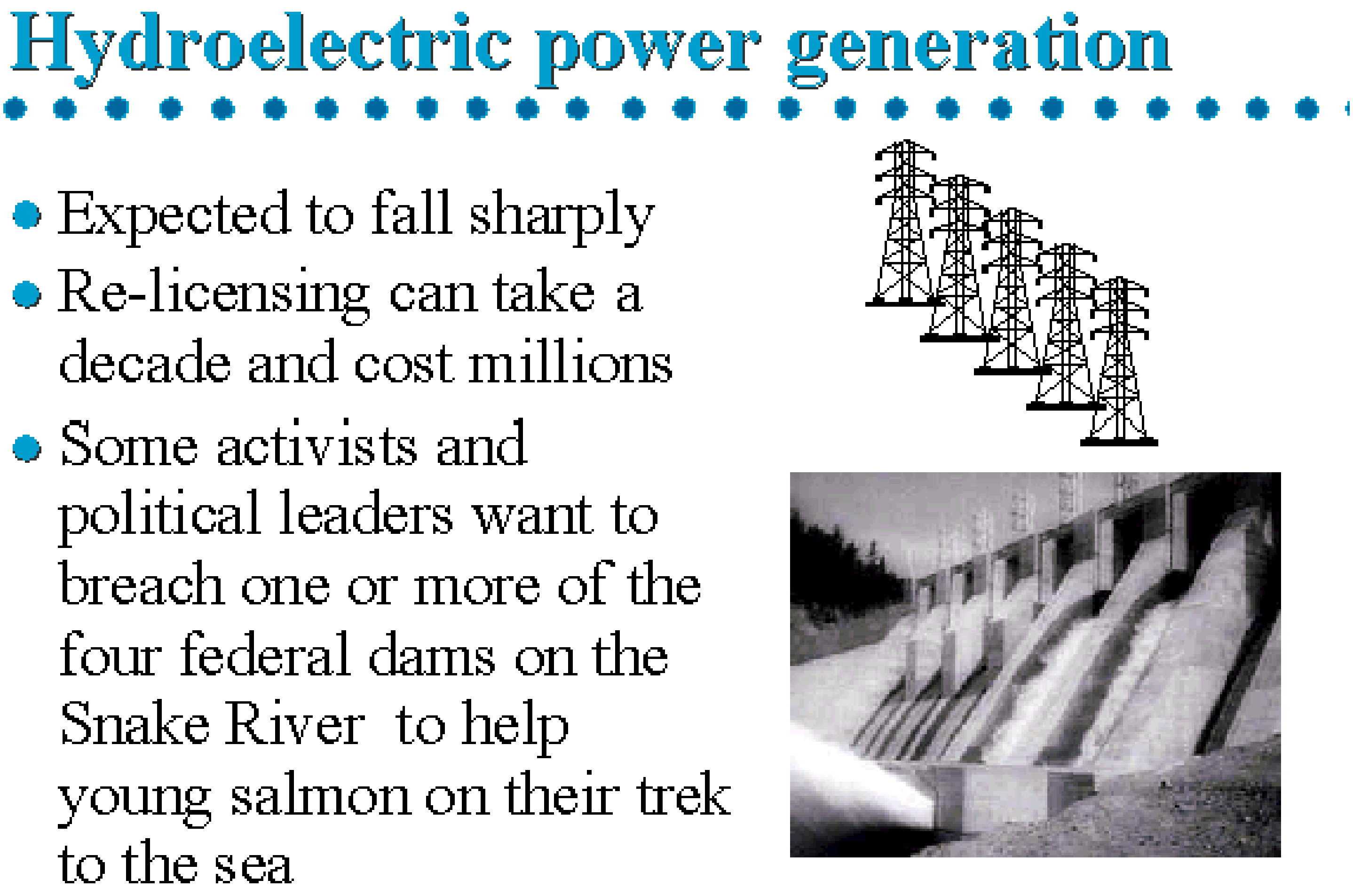

Hydroelectric power generation

- Expected to fall sharply
- Re-licensing can take a decade and cost millions
- Some activists and political leaders want to breach one or more of the four federal dams on the Snake River to help young salmon on their trek to the sea

Published by Lotus® Freelance Graphics®

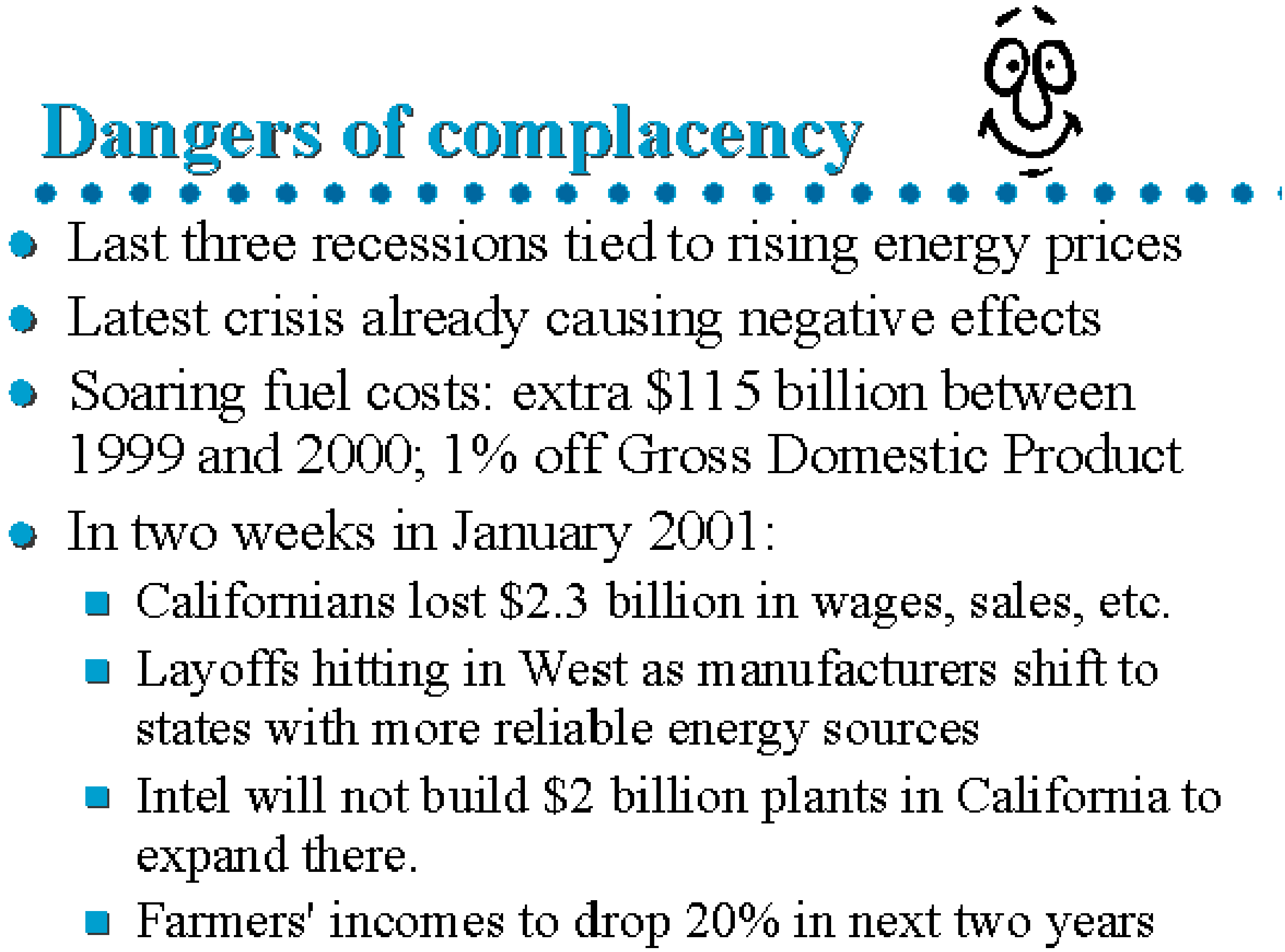

Dangers of complacency

- Last three recessions tied to rising energy prices
- Latest crisis already causing negative effects
- Soaring fuel costs: extra $115 billion between 1999 and 2000; 1% off Gross Domestic Product
- In two weeks in January 2001:
  - Californians lost $2.3 billion in wages, sales, etc.
  - Layoffs hitting in West as manufacturers shift to states with more reliable energy sources
  - Intel will not build $2 billion plants in California to expand there.
  - Farmers' incomes to drop 20% in next two years

Published by Lotus® Freelance Graphics®

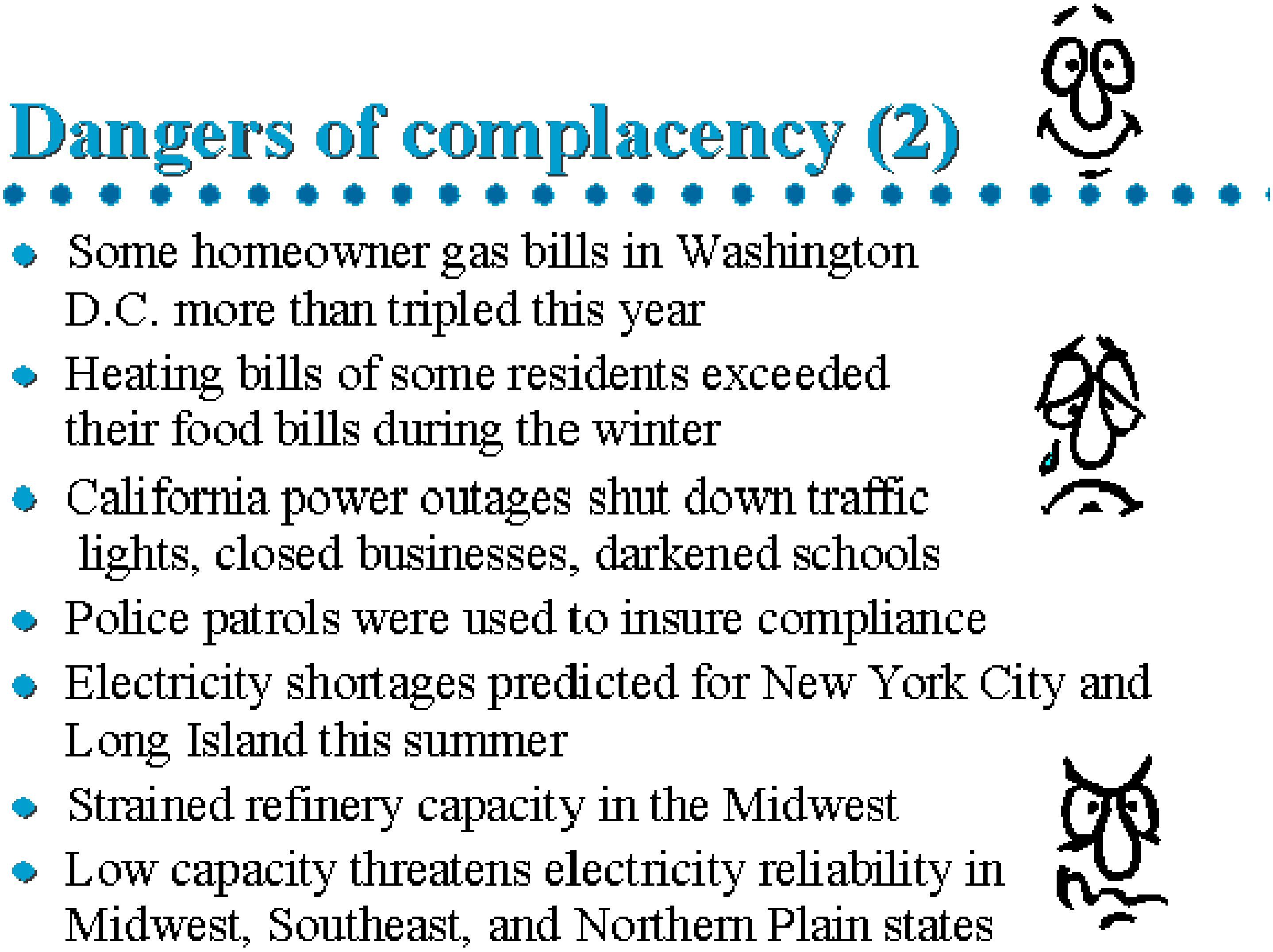

Dangers of complacency (2)

- Some homeowner gas bills in Washington D.C. more than tripled this year
- Heating bills of some residents exceeded their food bills during the winter
- California power outages shut down traffic lights, closed businesses, darkened schools
- Police patrols were used to insure compliance
- Electricity shortages predicted for New York City and Long Island this summer
- Strained refinery capacity in the Midwest
- Low capacity threatens electricity reliability in Midwest, Southeast, and Northern Plain states

Published by Lotus® Freelance Graphics®

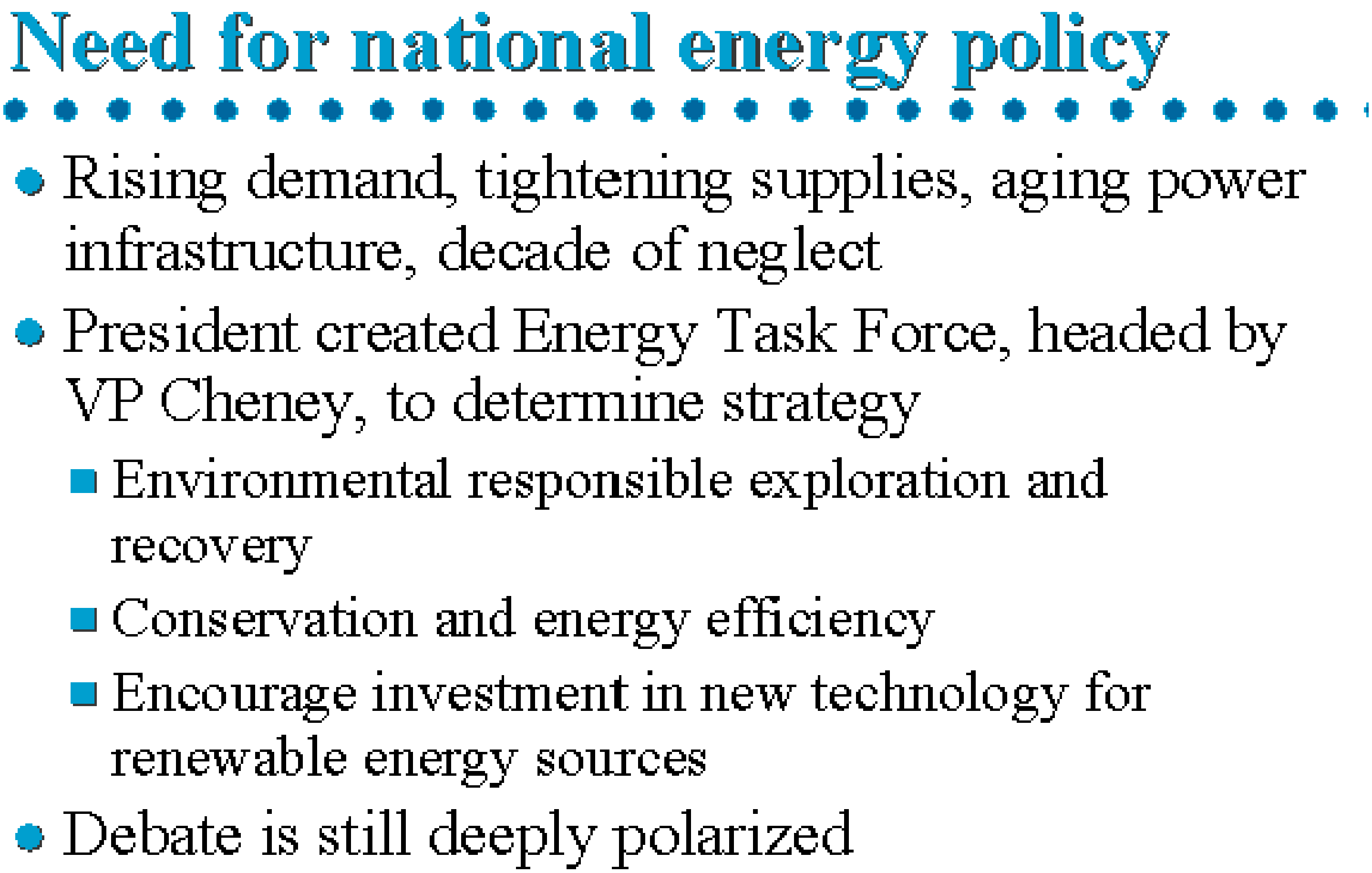

Need for national energy policy

- Rising demand, tightening supplies, aging power infrastructure, decade of neglect
- President created Energy Task Force, headed by VP Cheney, to determine strategy
  - Environmental responsible exploration and recovery
  - Conservation and energy efficiency
  - Encourage investment in new technology for renewable energy sources
- Debate is still deeply polarized

Published by Lotus® Freelance Graphics®

Extremes of the energy positions

- Some activists:
  - Zero tolerance to exploration
  - New, undiscovered sources will somehow provide for us

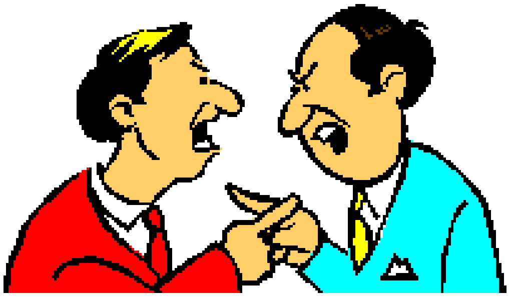

- Some advocates:
  - Limitless faith in special tax breaks for a favored activity or subsidies for a preferred industry

Published by Lotus® Freelance Graphics®

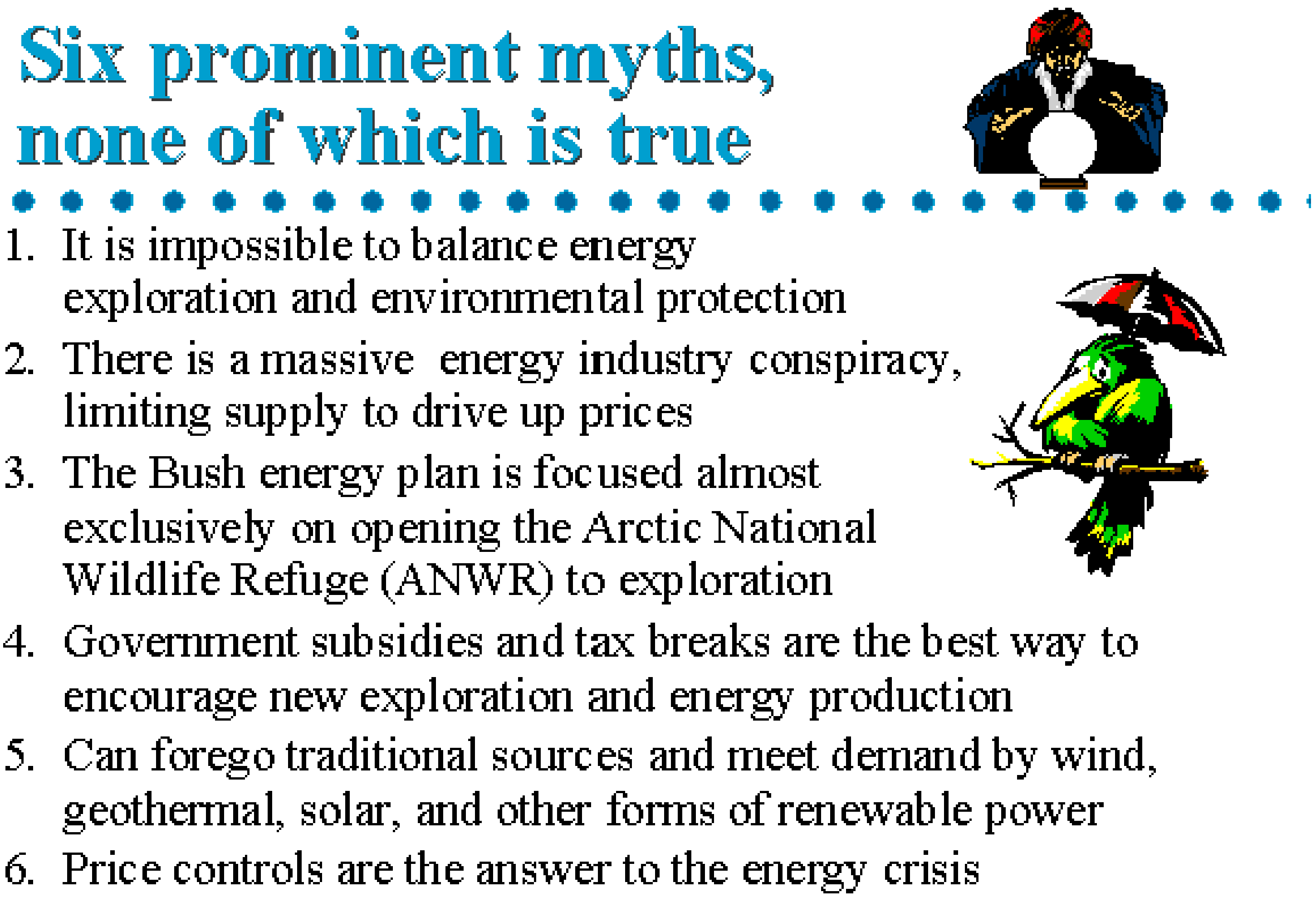

Six prominent myths, none of which is true

1. It is impossible to balance energy exploration and environmental protection
2. There is a massive energy industry conspiracy, limiting supply to drive up prices
3. The Bush energy plan is focused almost exclusively on opening the Arctic National Wildlife Refuge (ANWR) to exploration
4. Government subsidies and tax breaks are the best way to encourage new exploration and energy production
5. Can forego traditional sources and meet demand by wind, geothermal, solar, and other forms of renewable power
6. Price controls are the answer to the energy crisis

Published by Lotus® Freelance Graphics®

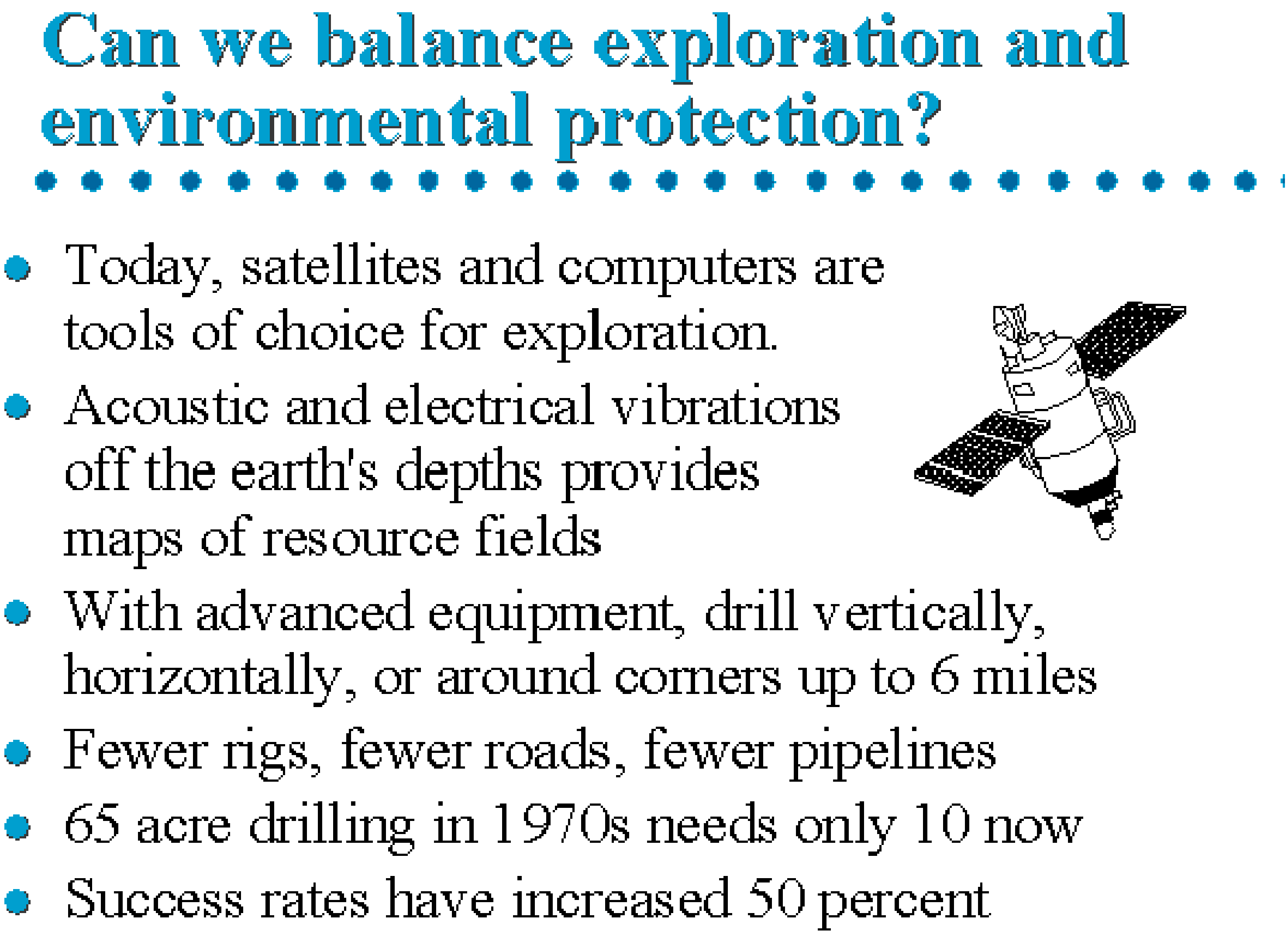

Can we balance exploration and environmental protection?

- Today, satellites and computers are tools of choice for exploration.
- Acoustic and electrical vibrations off the earth's depths provides maps of resource fields
- With advanced equipment, drill vertically, horizontally, or around corners up to 6 miles
- Fewer rigs, fewer roads, fewer pipelines
- 65 acre drilling in 1970s needs only 10 now
- Success rates have increased 50 percent

Published by Lotus® Freelance Graphics®

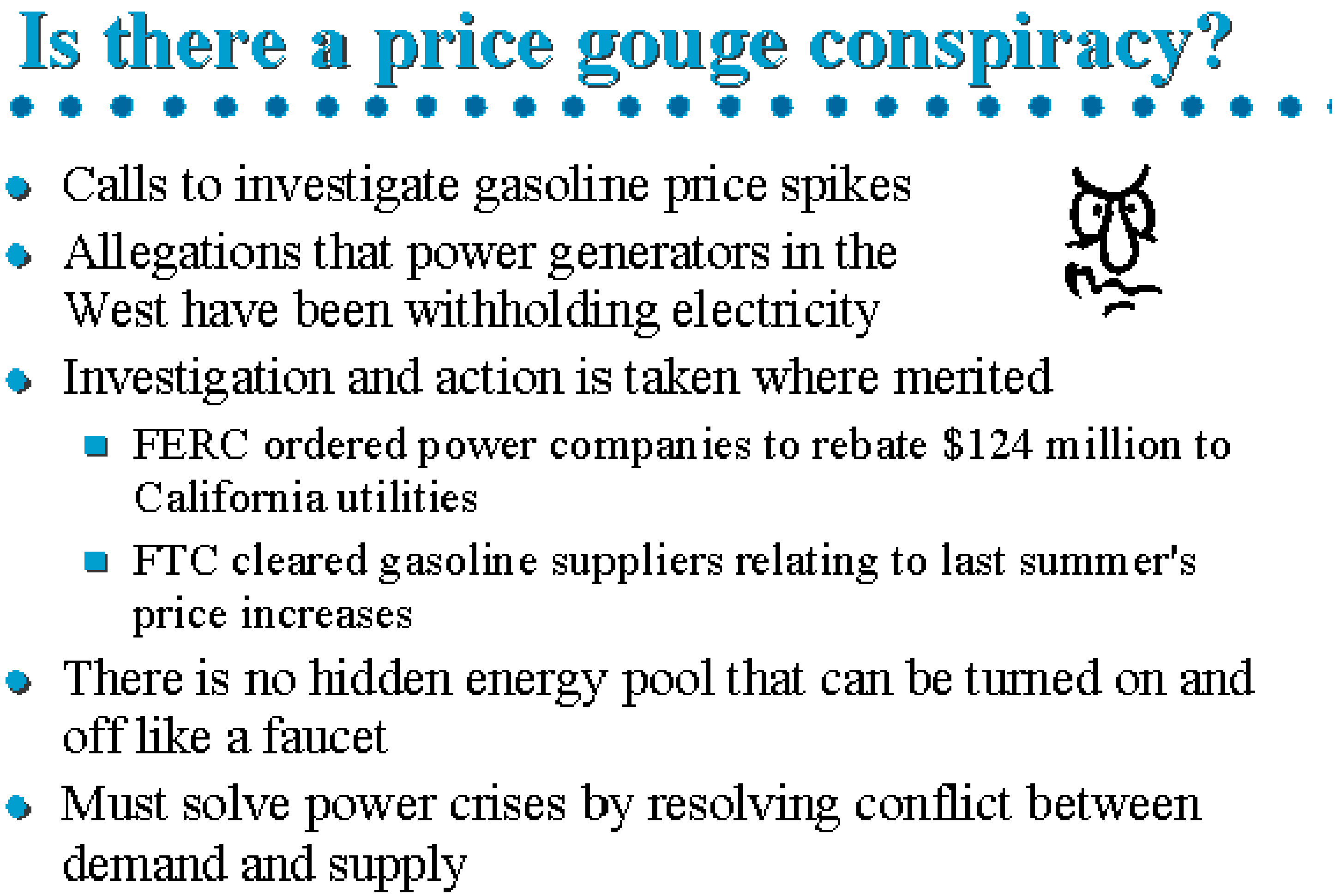

Is there a price gouge conspiracy?

- Calls to investigate gasoline price spikes
- Allegations that power generators in the West have been withholding electricity
- Investigation and action is taken where merited
  - FERC ordered power companies to rebate $124 million to California utilities
  - FTC cleared gasoline suppliers relating to last summer's price increases
- There is no hidden energy pool that can be turned on and off like a faucet
- Must solve power crises by resolving conflict between demand and supply

Published by Lotus® Freelance Graphics®

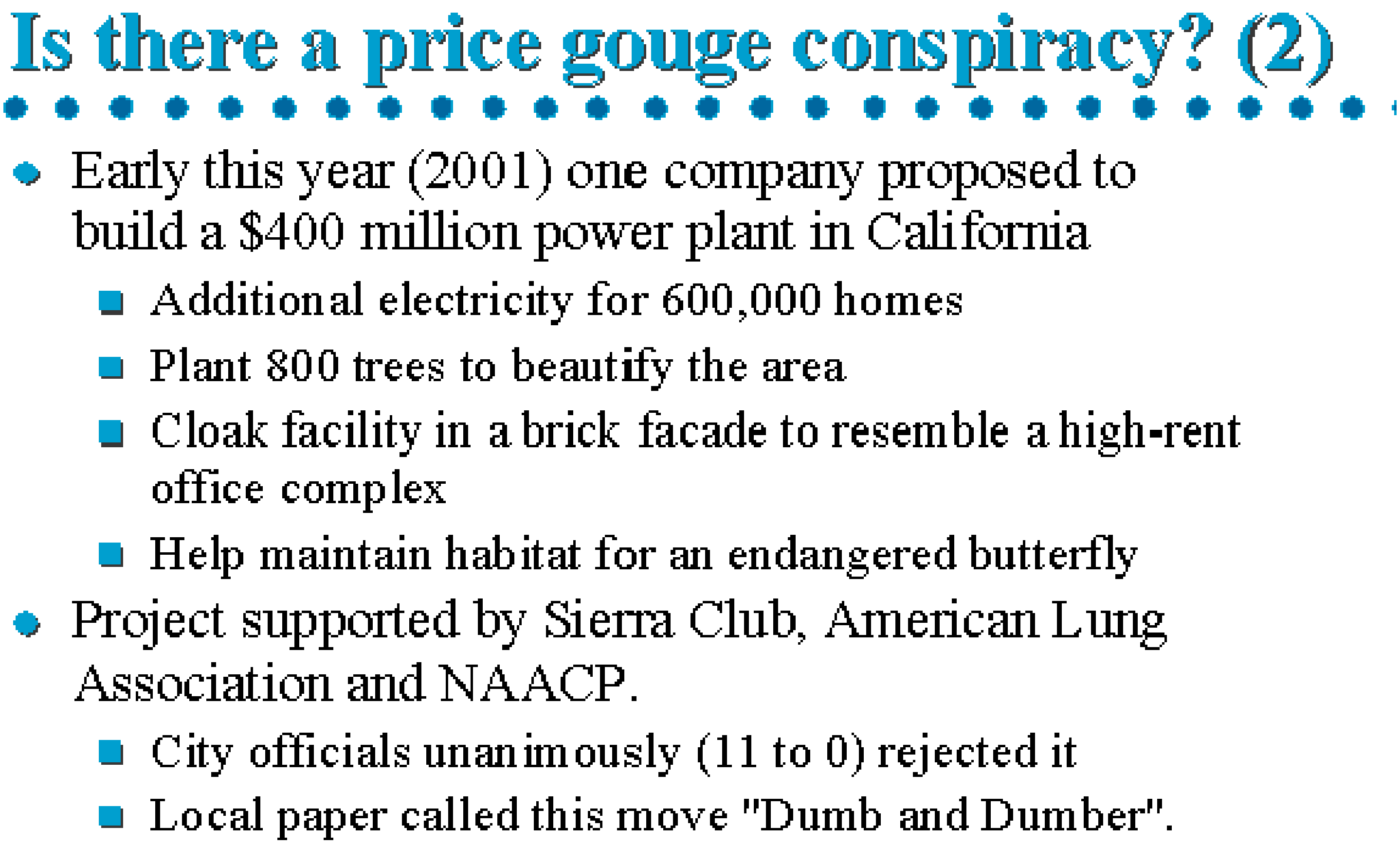

Is there a price gouge conspiracy? (2)

- Early this year (2001) one company proposed to build a $400 million power plant in California
  - Additional electricity for 600,000 homes
  - Plant 800 trees to beautify the area
  - Cloak facility in a brick facade to resemble a high-rent office complex
  - Help maintain habitat for an endangered butterfly
- Project supported by Sierra Club, American Lung Association and NAACP.
  - City officials unanimously (11 to 0) rejected it
  - Local paper called this move "Dumb and Dumber".

Published by Lotus® Freelance Graphics®

Is there a price gouge conspiracy? (3)

- Plans to build a 550 megawatt gas-fired generator in a Los Angeles suburb were scrapped
- Residents voted 2:1 against the project
- Local mayor launched a hunger strike in opposition to it.
- In California:
  - Workers are being laid off
  - Companies are leaving the state
  - Consumers threatened with rolling blackouts
  - Local officials reject power plants with little regard
- Is it really any mystery why there hasn't been a single new power plant built in California in the last decade?

Published by Lotus® Freelance Graphics®

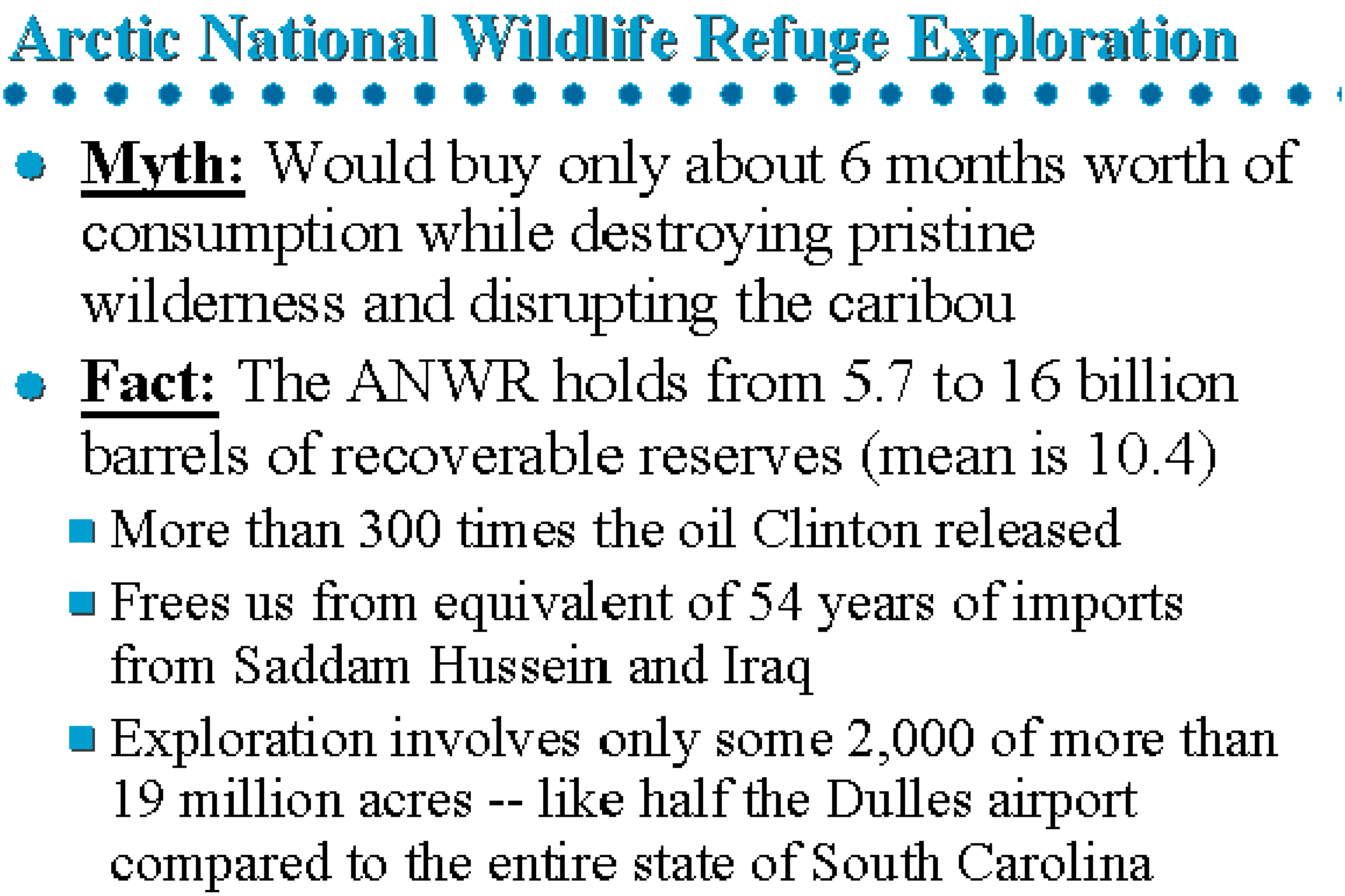

Arctic National Wildlife Refuge Exploration

- **Myth:** Would buy only about 6 months worth of consumption while destroying pristine wilderness and disrupting the caribou
- **Fact:** The ANWR holds from 5.7 to 16 billion barrels of recoverable reserves (mean is 10.4)
    - More than 300 times the oil Clinton released
    - Frees us from equivalent of 54 years of imports from Saddam Hussein and Iraq
    - Exploration involves only some 2,000 of more than 19 million acres -- like half the Dulles airport compared to the entire state of South Carolina

Published by Lotus® Freelance Graphics®

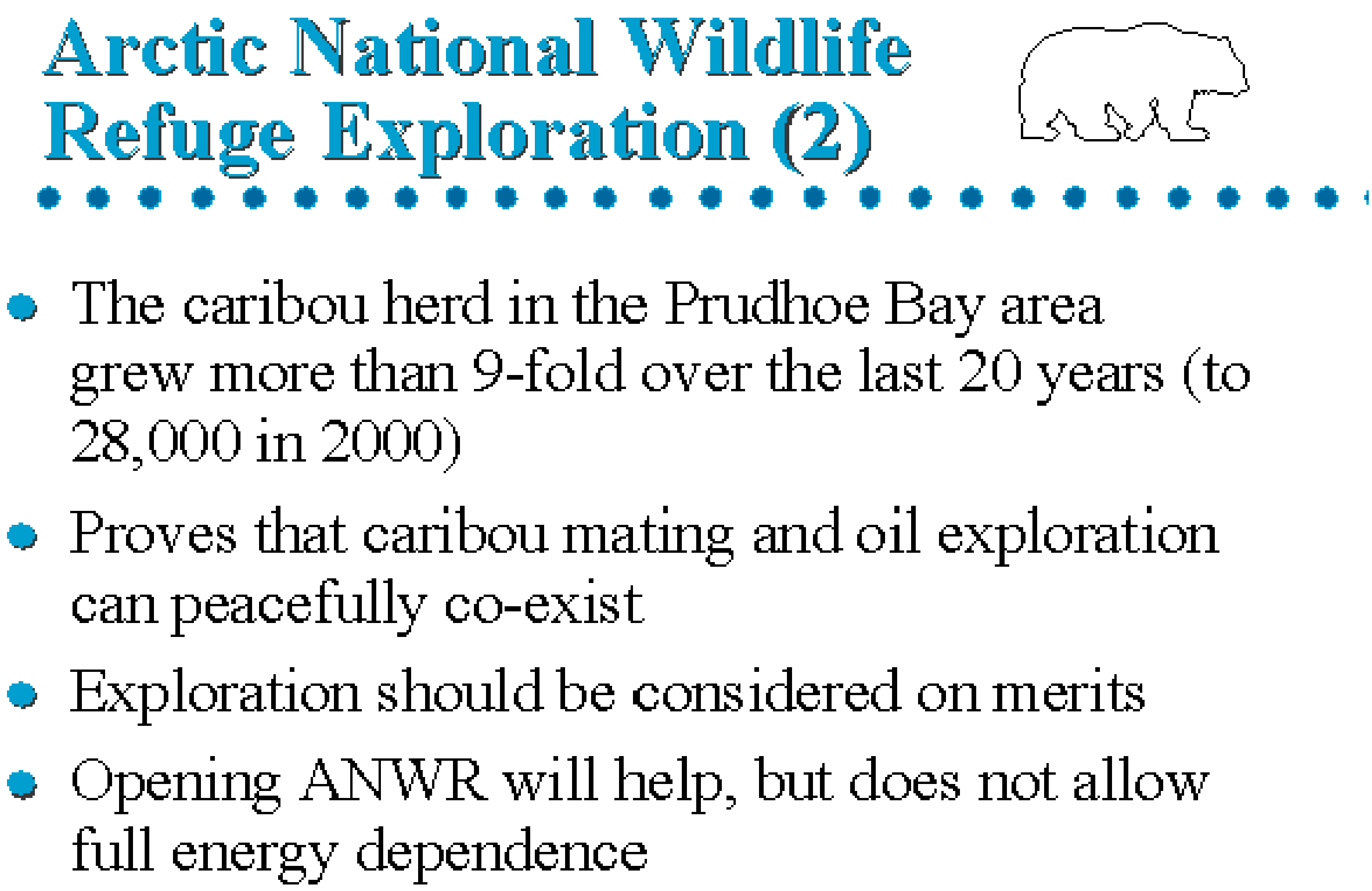

Arctic National Wildlife Refuge Exploration (2)

- The caribou herd in the Prudhoe Bay area grew more than 9-fold over the last 20 years (to 28,000 in 2000)
- Proves that caribou mating and oil exploration can peacefully co-exist
- Exploration should be considered on merits
- Opening ANWR will help, but does not allow full energy dependence

Published by Lotus® Freelance Graphics®

Our increasing oil imports

- U.S. became a net importer of oil in 1950s
- Closing off all new domestic sources enhances leverage and power of oil cartels
- These cannot relied on to put American interests first
- Development of domestic resources such as ANWR will help increase energy security by ensuring a more diverse supply of oil

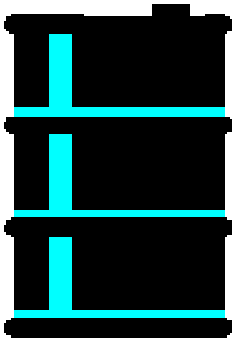

Published by Lotus® Freelance Graphics®

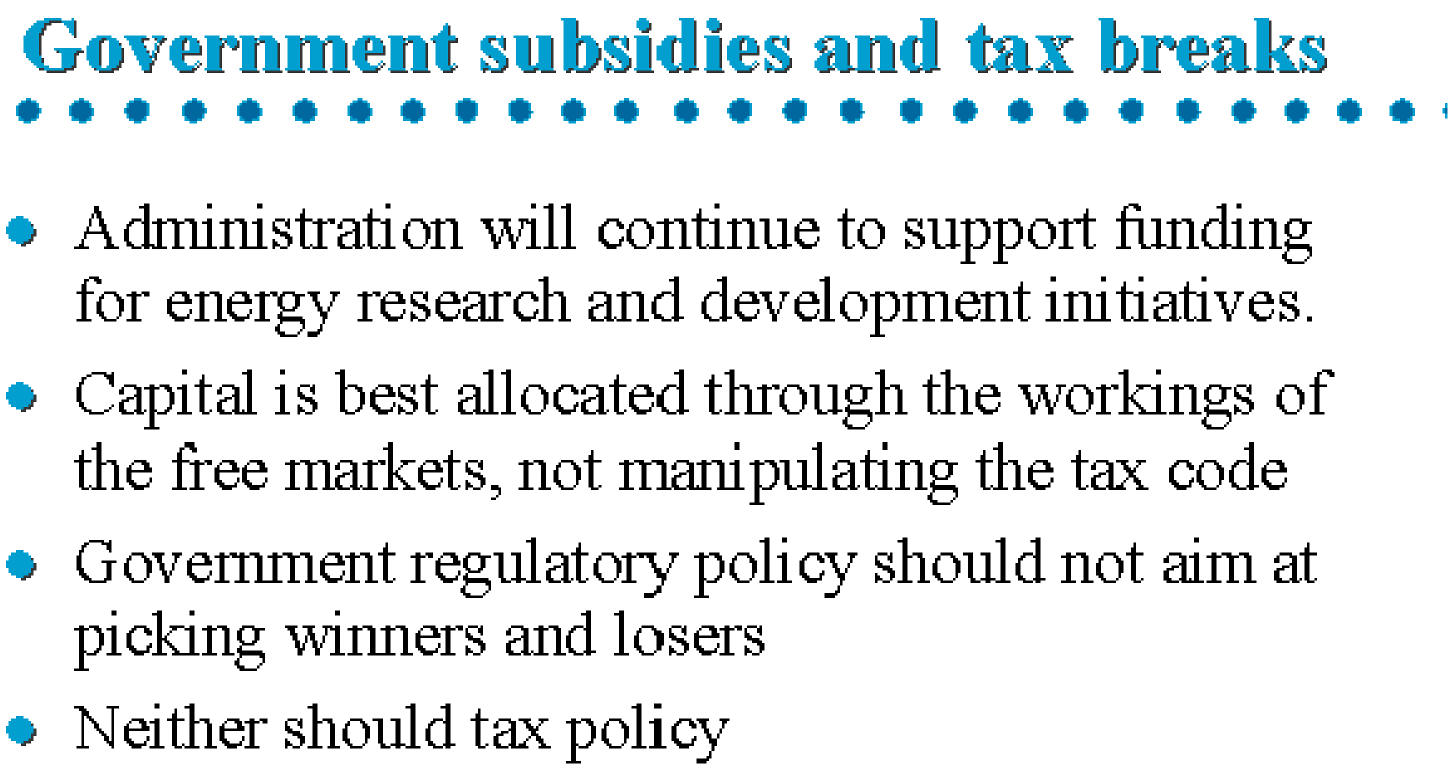

Government subsidies and tax breaks

- Administration will continue to support funding for energy research and development initiatives.
- Capital is best allocated through the workings of the free markets, not manipulating the tax code
- Government regulatory policy should not aim at picking winners and losers
- Neither should tax policy

Published by Lotus® Freelance Graphics®

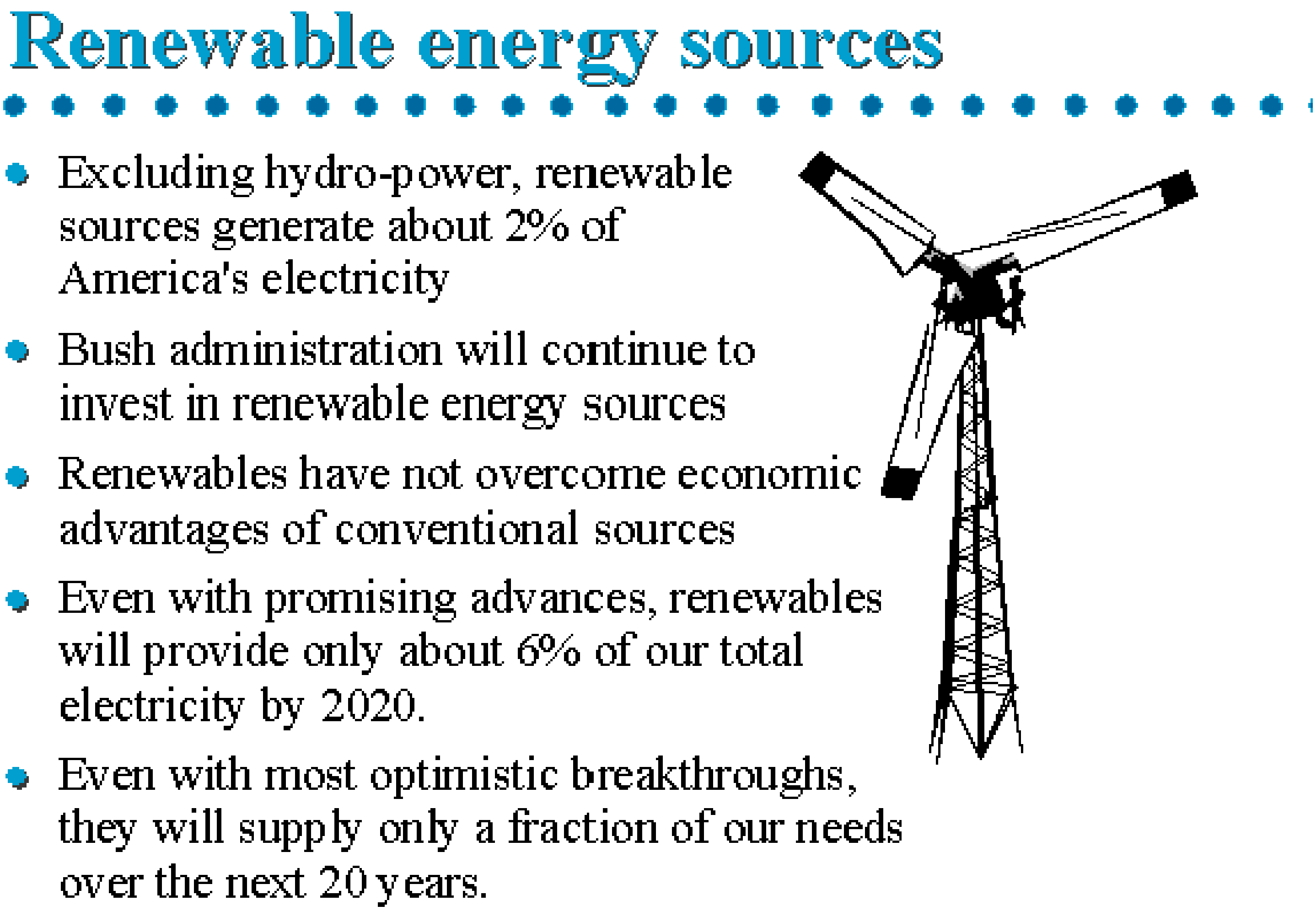

Renewable energy sources

- Excluding hydro-power, renewable sources generate about 2% of America's electricity
- Bush administration will continue to invest in renewable energy sources
- Renewables have not overcome economic advantages of conventional sources
- Even with promising advances, renewables will provide only about 6% of our total electricity by 2020.
- Even with most optimistic breakthroughs, they will supply only a fraction of our needs over the next 20 years.

Published by Lotus® Freelance Graphics®

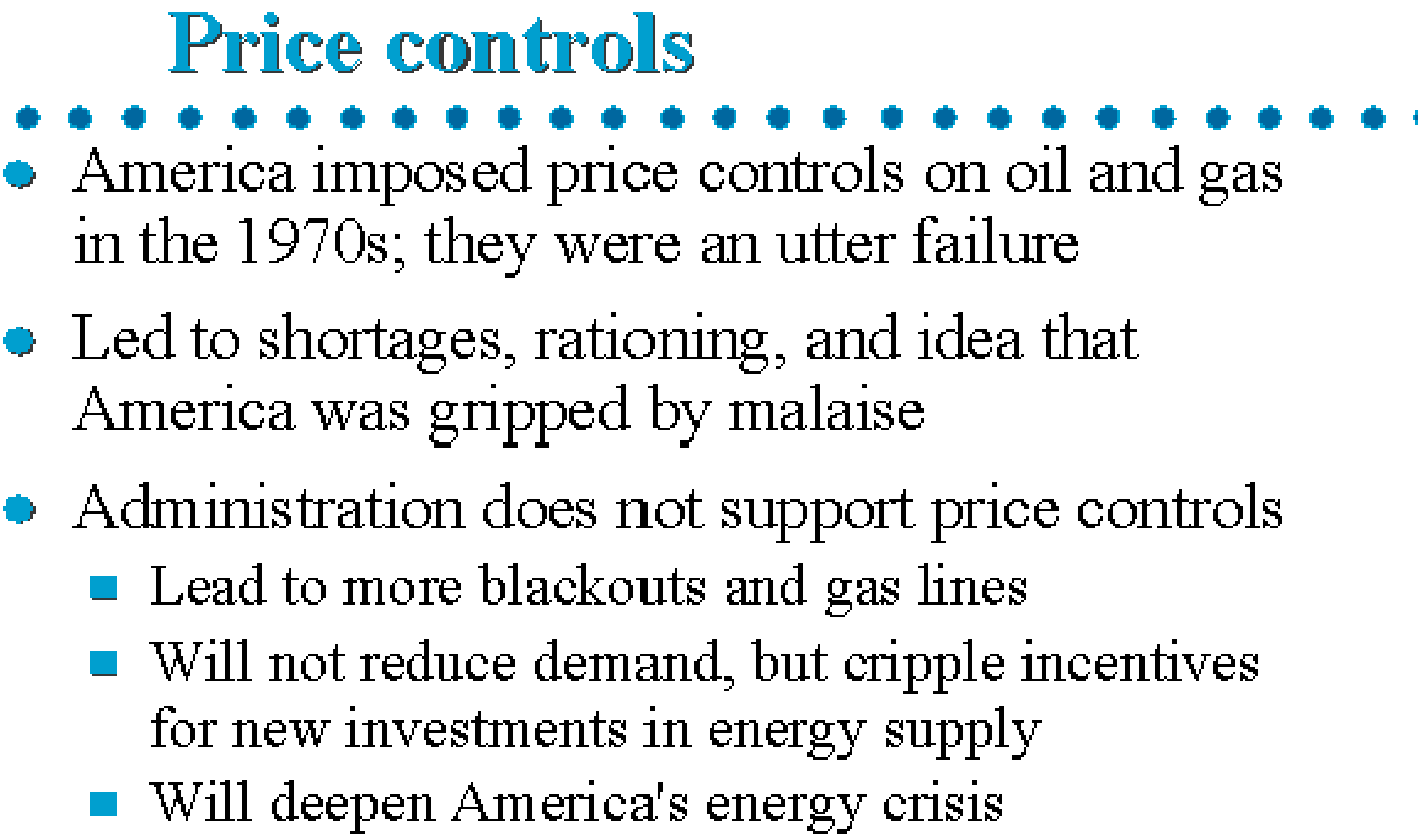

Price controls

- America imposed price controls on oil and gas in the 1970s; they were an utter failure
- Led to shortages, rationing, and idea that America was gripped by malaise
- Administration does not support price controls
  - Lead to more blackouts and gas lines
  - Will not reduce demand, but cripple incentives for new investments in energy supply
  - Will deepen America's energy crisis

Published by Lotus® Freelance Graphics®

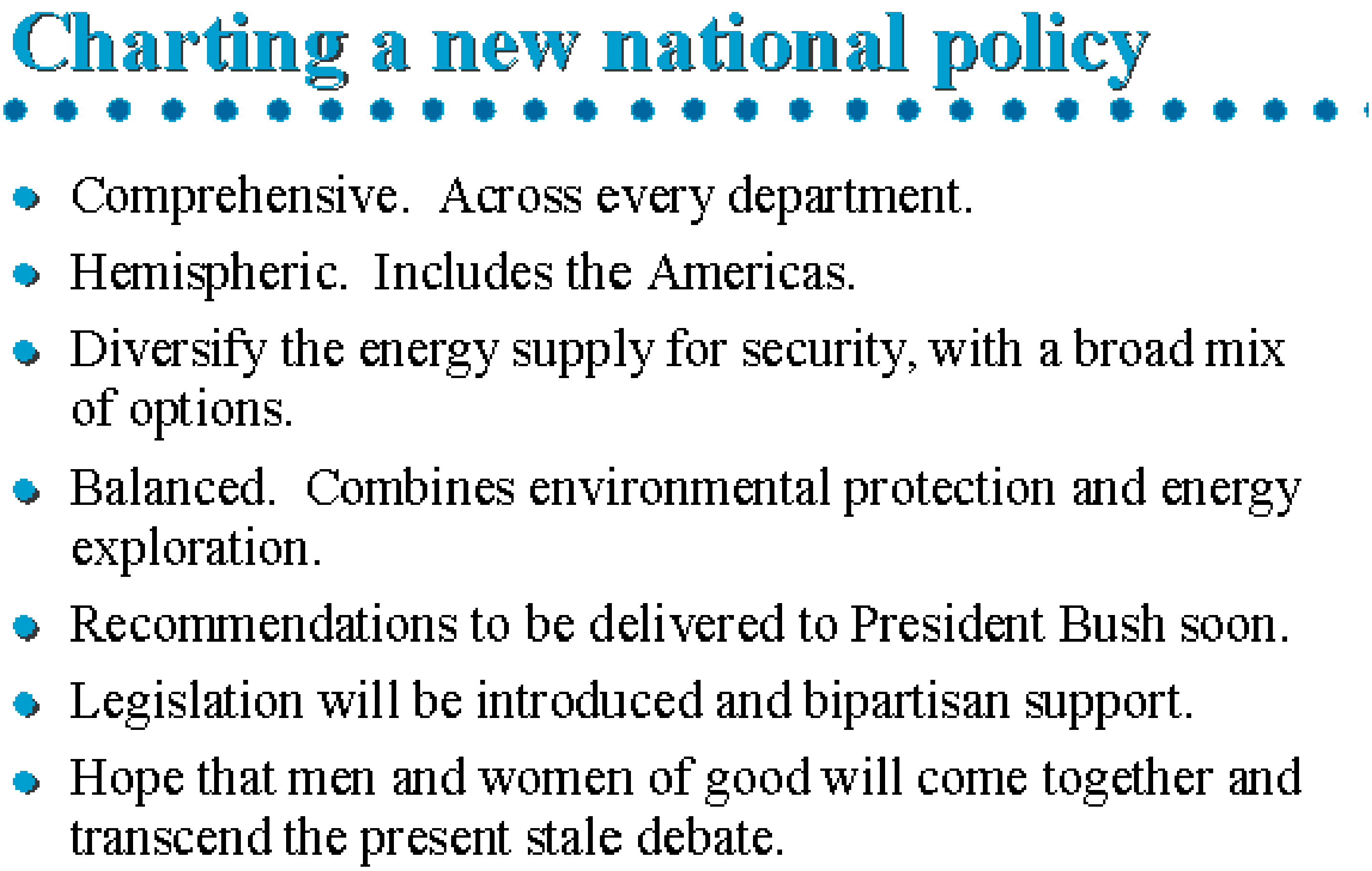

Charting a new national policy

- Comprehensive. Across every department.
- Hemispheric. Includes the Americas.
- Diversify the energy supply for security, with a broad mix of options.
- Balanced. Combines environmental protection and energy exploration.
- Recommendations to be delivered to President Bush soon.
- Legislation will be introduced and bipartisan support.
- Hope that men and women of good will come together and transcend the present stale debate.

Published by Lotus® Freelance Graphics®

Commentary

by Tom Bearden, Ph.D.

12 April 2001

Published by Lotus® Freelance Graphics®

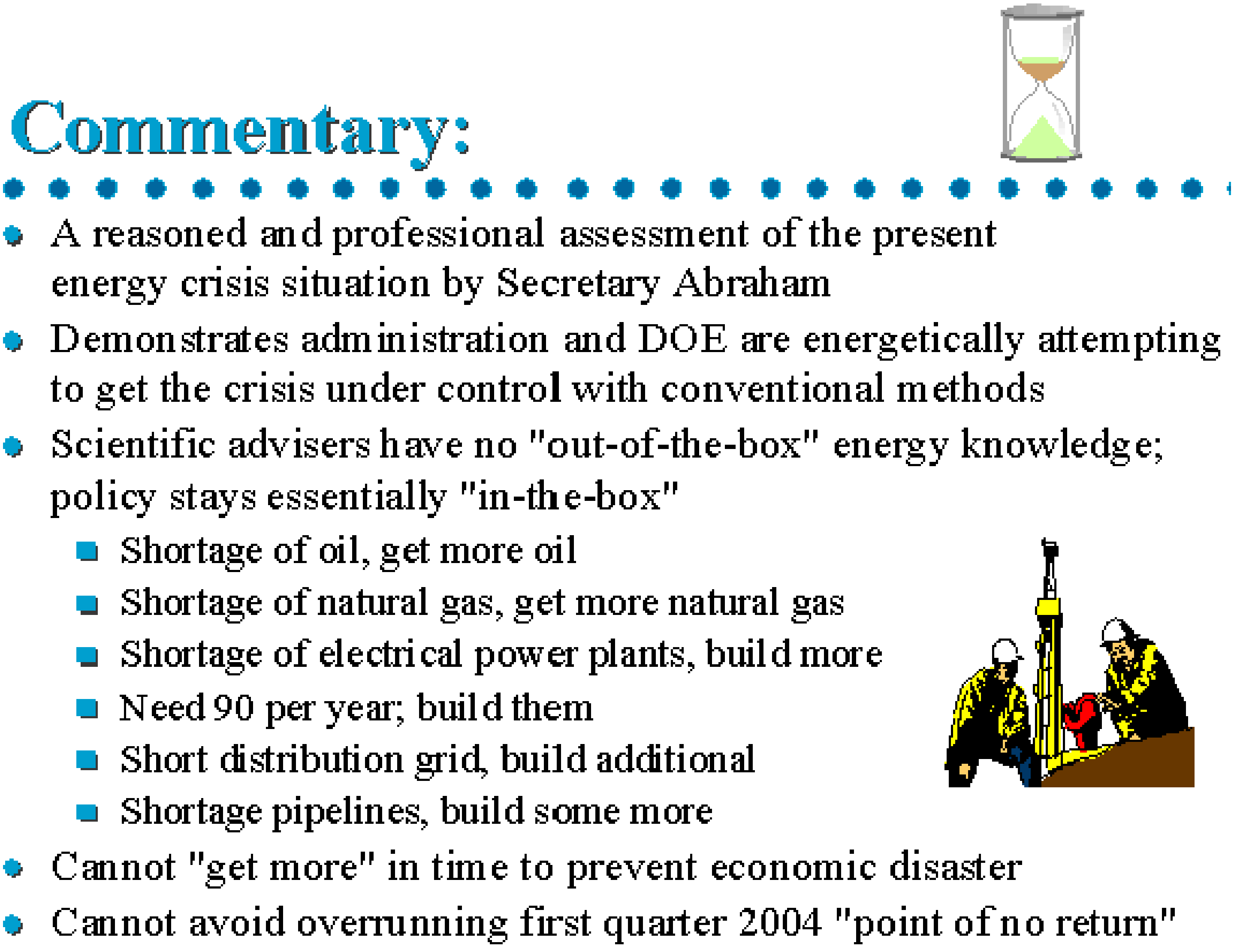

Commentary:

- A reasoned and professional assessment of the present energy crisis situation by Secretary Abraham
- Demonstrates administration and DOE are energetically attempting to get the crisis under control with conventional methods
- Scientific advisers have no "out-of-the-box" energy knowledge; policy stays essentially "in-the-box"
    - Shortage of oil, get more oil
    - Shortage of natural gas, get more natural gas
    - Shortage of electrical power plants, build more
    - Need 90 per year; build them
    - Short distribution grid, build additional
    - Shortage pipelines, build some more
- Cannot "get more" in time to prevent economic disaster
- Cannot avoid overrunning first quarter 2004 "point of no return"

Published by Lotus® Freelance Graphics®

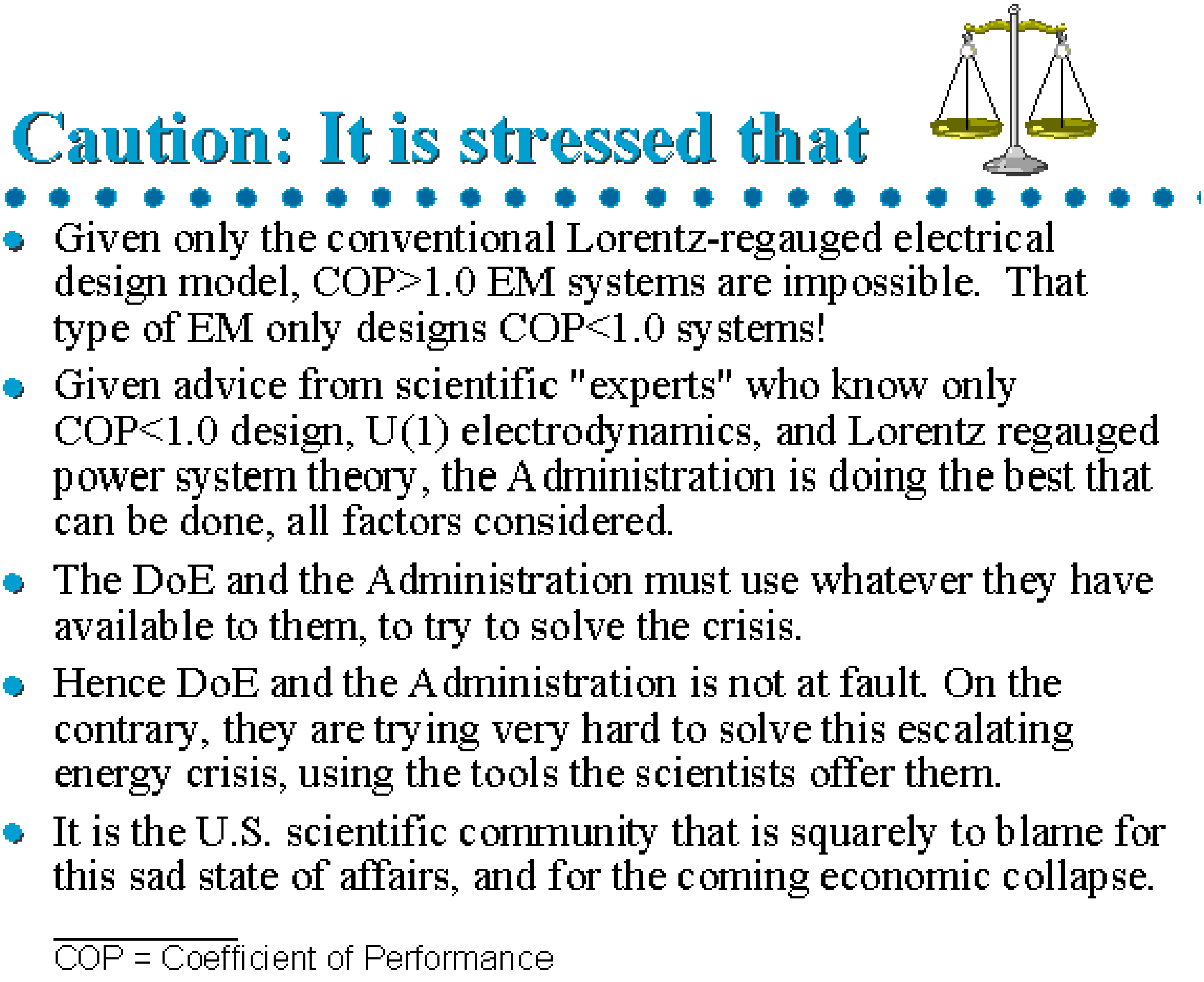

Caution: It is stressed that

- Given only the conventional Lorentz-regauged electrical design model, COP>1.0 EM systems are impossible. That type of EM only designs COP<1.0 systems!
- Given advice from scientific "experts" who know only COP<1.0 design, U(1) electrodynamics, and Lorentz regauged power system theory, the Administration is doing the best that can be done, all factors considered.
- The DoE and the Administration must use whatever they have available to them, to try to solve the crisis.
- Hence DoE and the Administration is not at fault. On the contrary, they are trying very hard to solve this escalating energy crisis, using the tools the scientists offer them.
- It is the U.S. scientific community that is squarely to blame for this sad state of affairs, and for the coming economic collapse.

COP = Coefficient of Performance

Published by Lotus® Freelance Graphics®

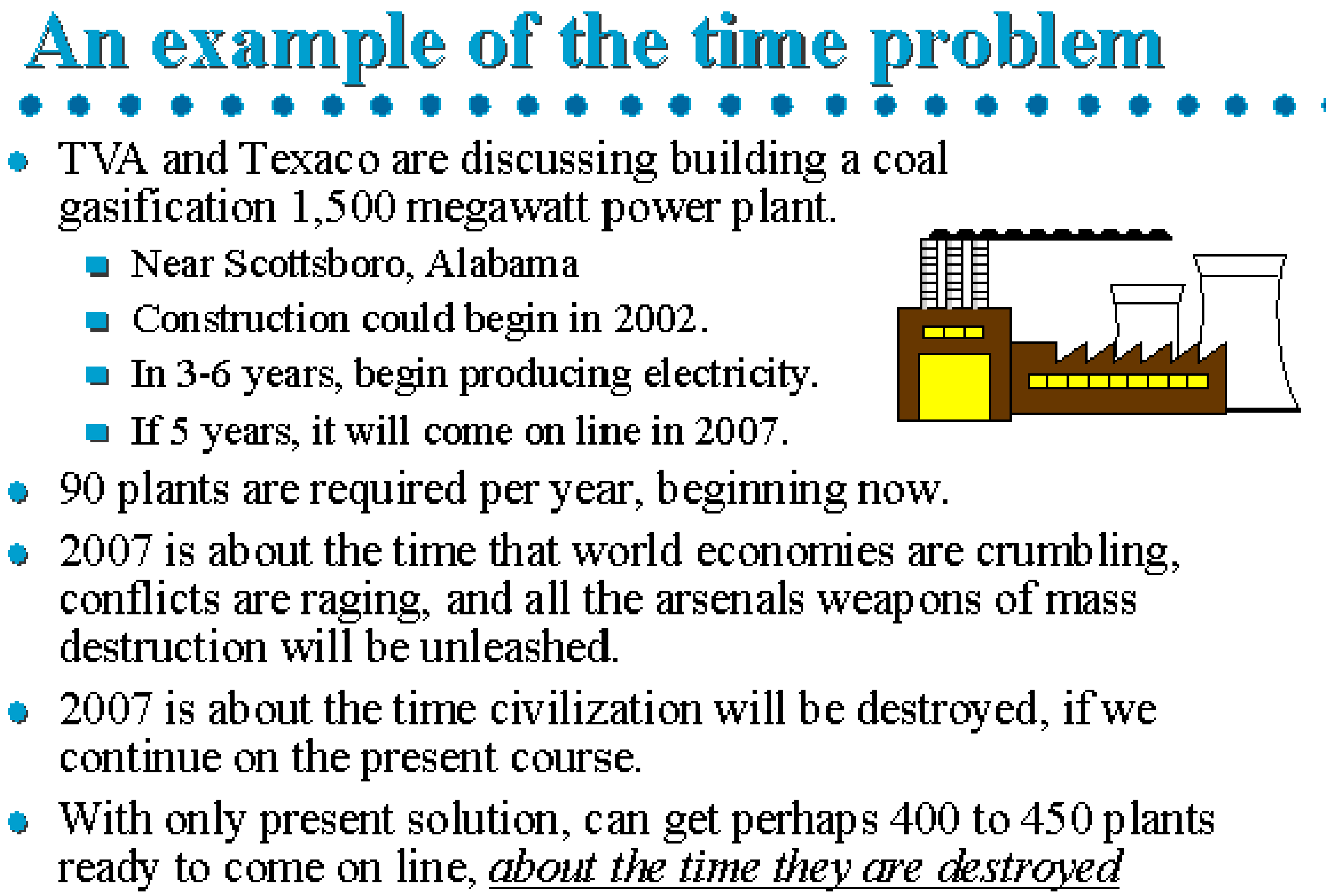

An example of the time problem

- TVA and Texaco are discussing building a coal gasification 1,500 megawatt power plant.
  - Near Scottsboro, Alabama
  - Construction could begin in 2002.
  - In 3-6 years, begin producing electricity.
  - If 5 years, it will come on line in 2007.
- 90 plants are required per year, beginning now.
- 2007 is about the time that world economies are crumbling, conflicts are raging, and all the arsenals weapons of mass destruction will be unleashed.
- 2007 is about the time civilization will be destroyed, if we continue on the present course.
- With only present solution, can get perhaps 400 to 450 plants ready to come on line, *about the time they are destroyed*

Published by Lotus® Freelance Graphics®

The pending disaster

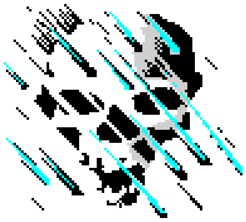

- The world energy crisis will escalate steadily, forcing collapse of world economy circa 2008.
- As nations' economies crumble, conflicts soar and one or more nations will exchange nukes and other *weapons of mass destruction* (WMD). Security partners forced to enter the melee.
- This evokes the long-dreaded WMD "spasm" response.
  - Only way for a nation to survive is to destroy its perceived foes before they destroy it.
  - Everybody fires everything, as fast as they can, unleashing all the arsenals of WMD, desperate attempt to survive.
  - Civilization and much of the biosphere is destroyed. Unleashing of smallpox alone will kill 2 billion.
- The time for this Armageddon is circa 2007.

Published by Lotus® Freelance Graphics®

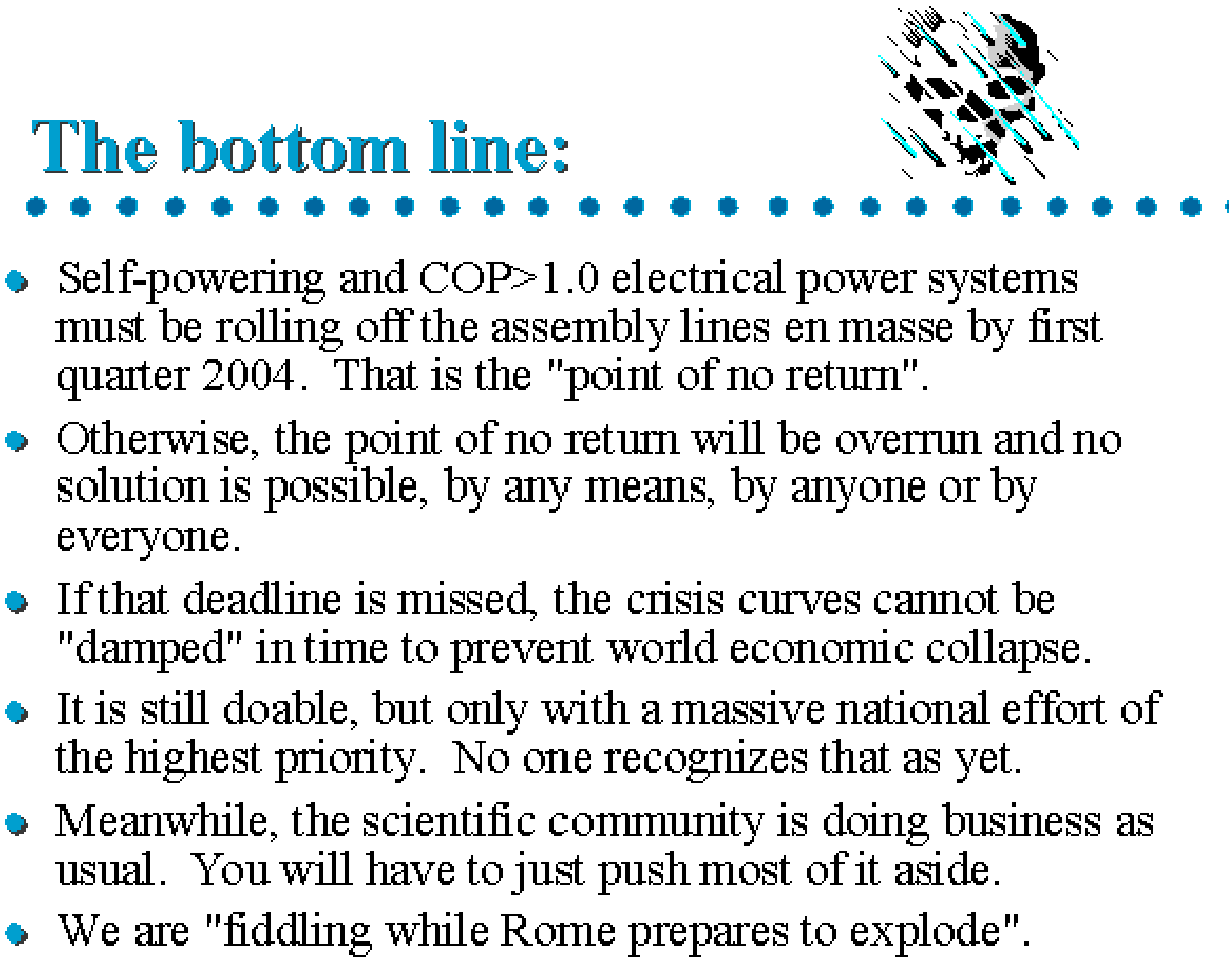

The bottom line:

- Self-powering and COP>1.0 electrical power systems must be rolling off the assembly lines en masse by first quarter 2004. That is the "point of no return".
- Otherwise, the point of no return will be overrun and no solution is possible, by any means, by anyone or by everyone.
- If that deadline is missed, the crisis curves cannot be "damped" in time to prevent world economic collapse.
- It is still doable, but only with a massive national effort of the highest priority. No one recognizes that as yet.
- Meanwhile, the scientific community is doing business as usual. You will have to just push most of it aside.
- We are "fiddling while Rome prepares to explode".

Published by Lotus® Freelance Graphics®

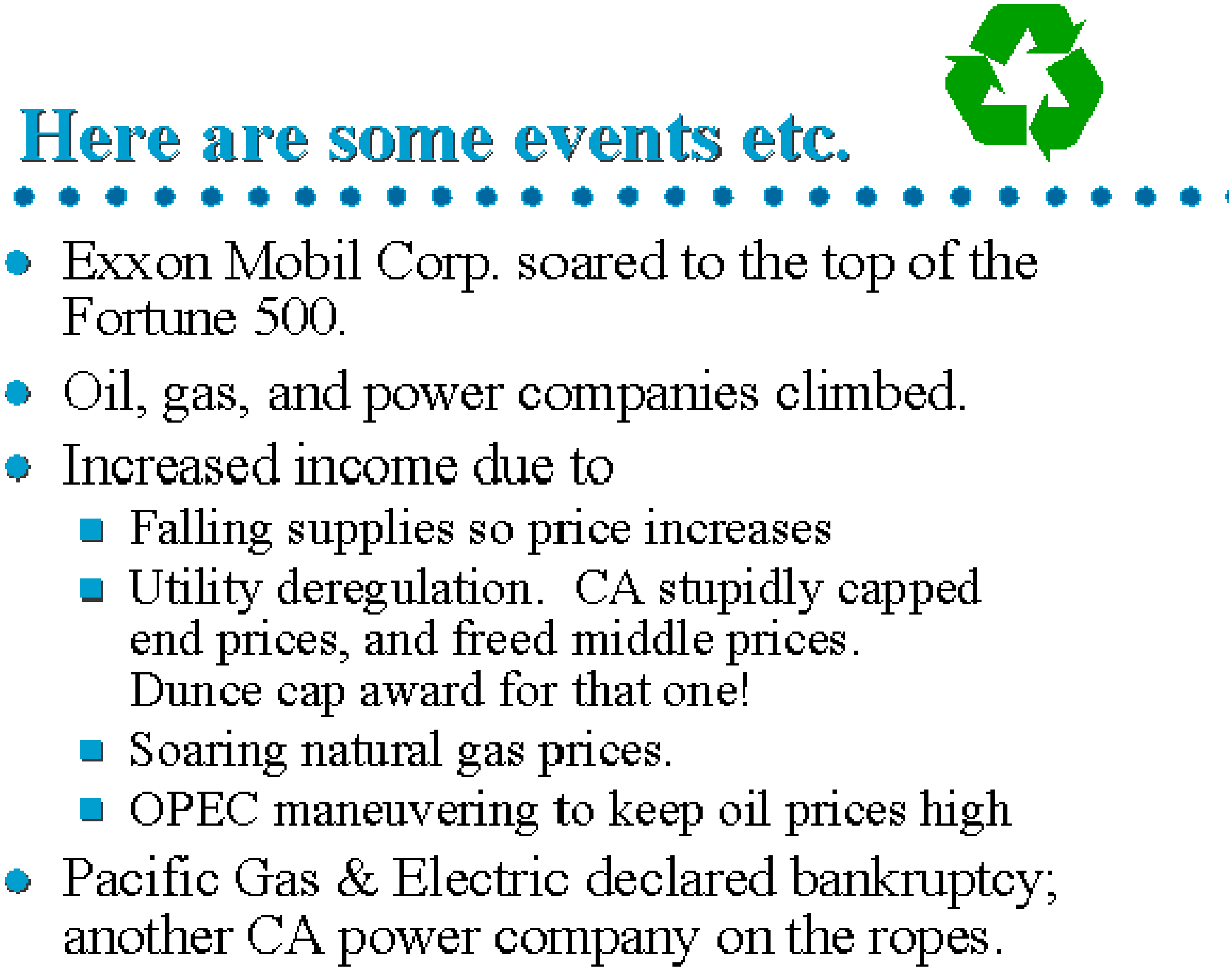

Here are some events etc.

- Exxon Mobil Corp. soared to the top of the Fortune 500.
- Oil, gas, and power companies climbed.
- Increased income due to
  - Falling supplies so price increases
  - Utility deregulation. CA stupidly capped end prices, and freed middle prices. Dunce cap award for that one!
  - Soaring natural gas prices.
  - OPEC maneuvering to keep oil prices high
- Pacific Gas & Electric declared bankruptcy; another CA power company on the ropes.

Published by Lotus® Freelance Graphics®

Lowered energy requirements for appliances

- Approved in last days of Clinton administration
- Washing machines must use 35% less energy, beginning in 2007
- Water heaters must use 5%-9% less energy, beginning in 2004
- Adds about $240 to price of washing machine
- Lower energy costs over time will offset initial increased cost
- Approved by Bush administration in April 2001 after review

Published by Lotus® Freelance Graphics®

Easing air conditioner and heat pump standards

- Bush administration eased Clinton's new efficiency standards
  - Still require air conditioners and heat pumps to use 20% less energy than most current models, beginning in 2006
  - Clinton standard was 30% less
- All central home air conditioning units must meet minimum seasonal efficiency ratio (SEER) of 12, compared to present 10
- Replaces Clinton SEER standard of 12
- In California, easing these standards requires two more power plants

Published by Lotus® Freelance Graphics®

Navarro's Estimate*

• • • • • • • • • • • • • • • • • •

Peter Navarro, economist at UC Irvine,

in conjunction with Utility Consumers's Action Network

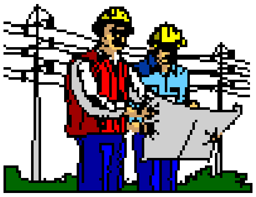

*Jennifer Coleman, "Power costs could be triple last year's for Californians," AP release, 22 Apr. 2001

Published by Lotus® Freelance Graphics®

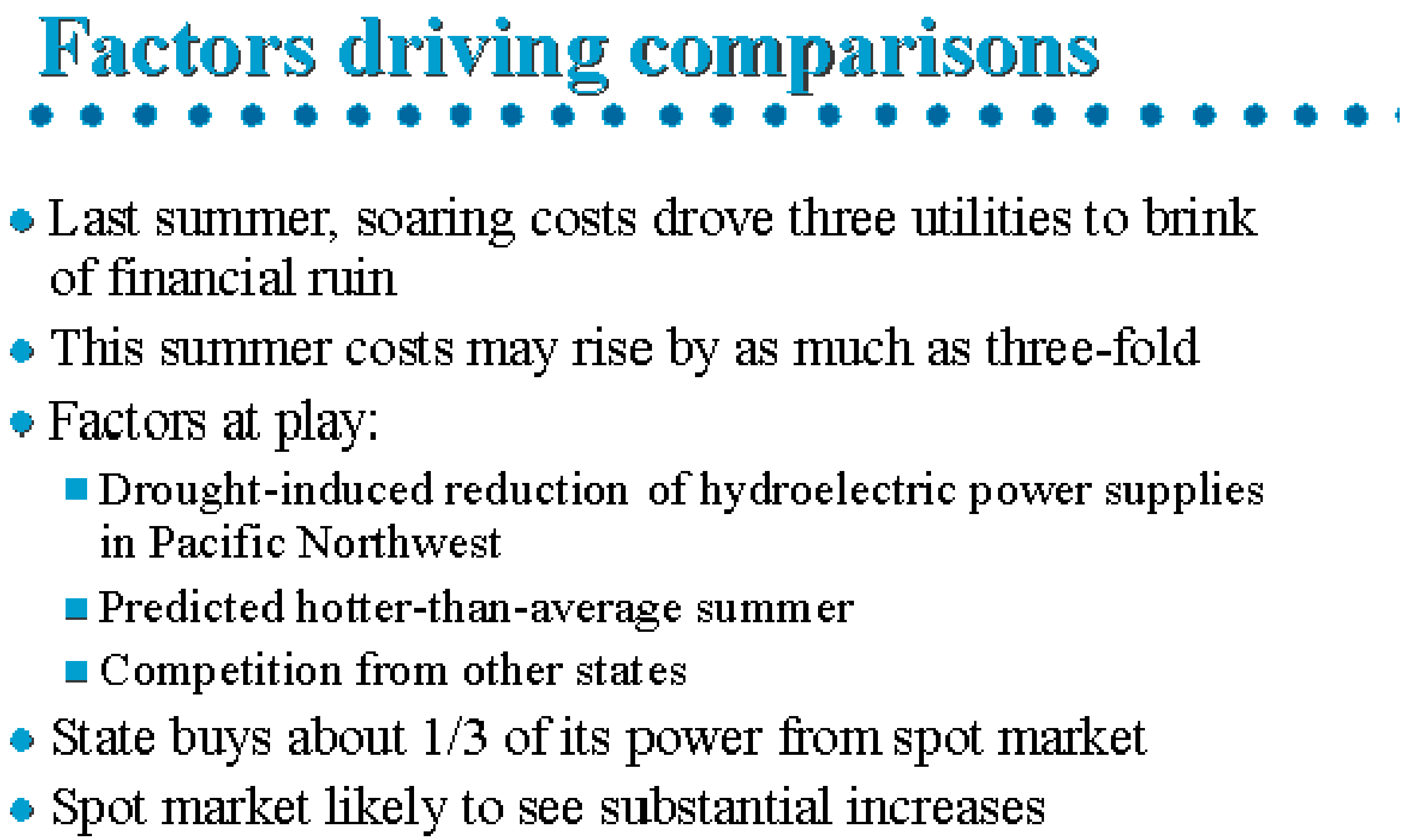

Factors driving comparisons

- Last summer, soaring costs drove three utilities to brink of financial ruin
- This summer costs may rise by as much as three-fold
- Factors at play:
  - Drought-induced reduction of hydroelectric power supplies in Pacific Northwest
  - Predicted hotter-than-average summer
  - Competition from other states
- State buys about 1/3 of its power from spot market
- Spot market likely to see substantial increases

Published by Lotus® Freelance Graphics®

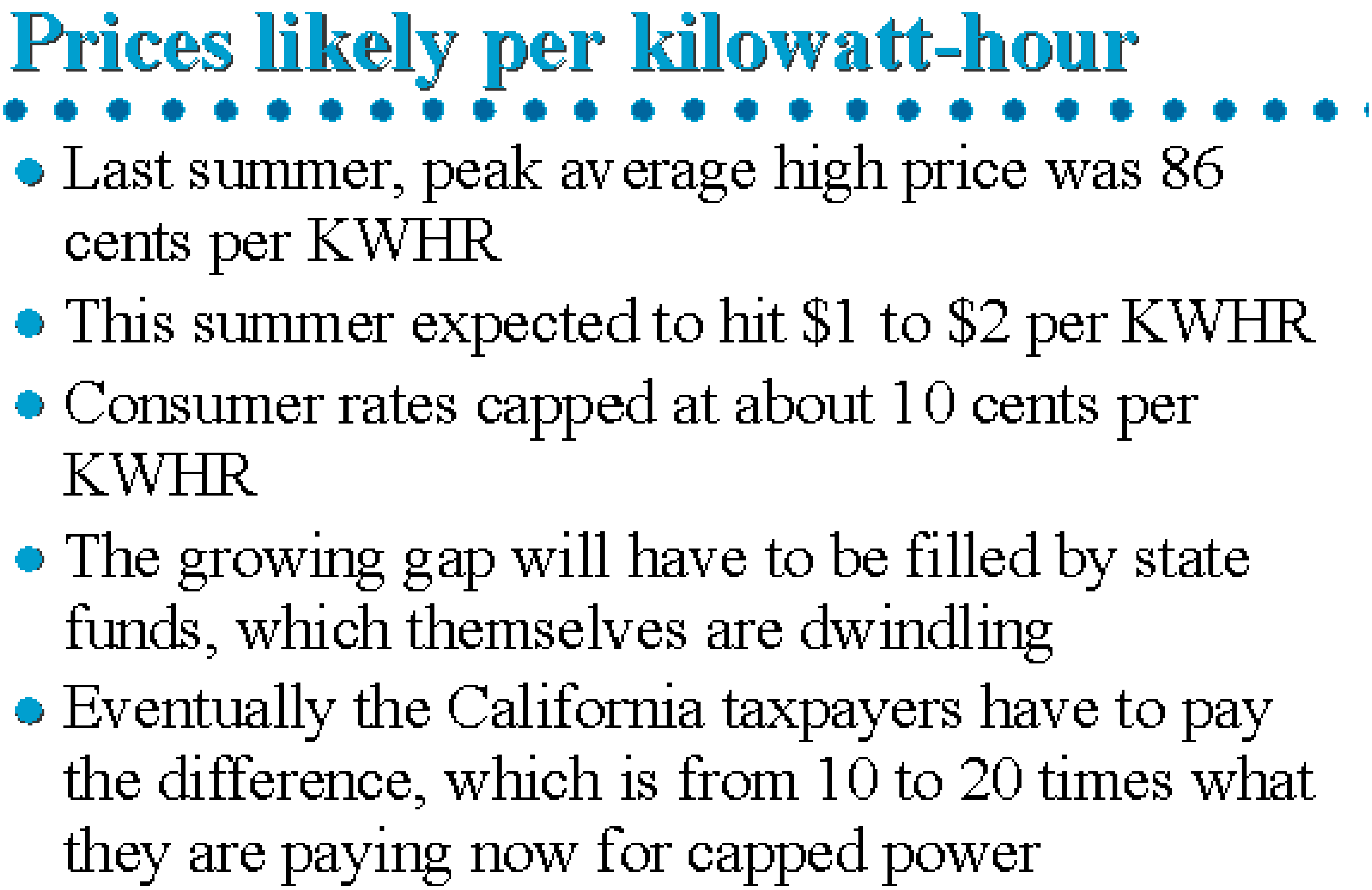

Prices likely per kilowatt-hour

- Last summer, peak average high price was 86 cents per KWHR
- This summer expected to hit $1 to $2 per KWHR
- Consumer rates capped at about 10 cents per KWHR
- The growing gap will have to be filled by state funds, which themselves are dwindling
- Eventually the California taxpayers have to pay the difference, which is from 10 to 20 times what they are paying now for capped power

Published by Lotus® Freelance Graphics®

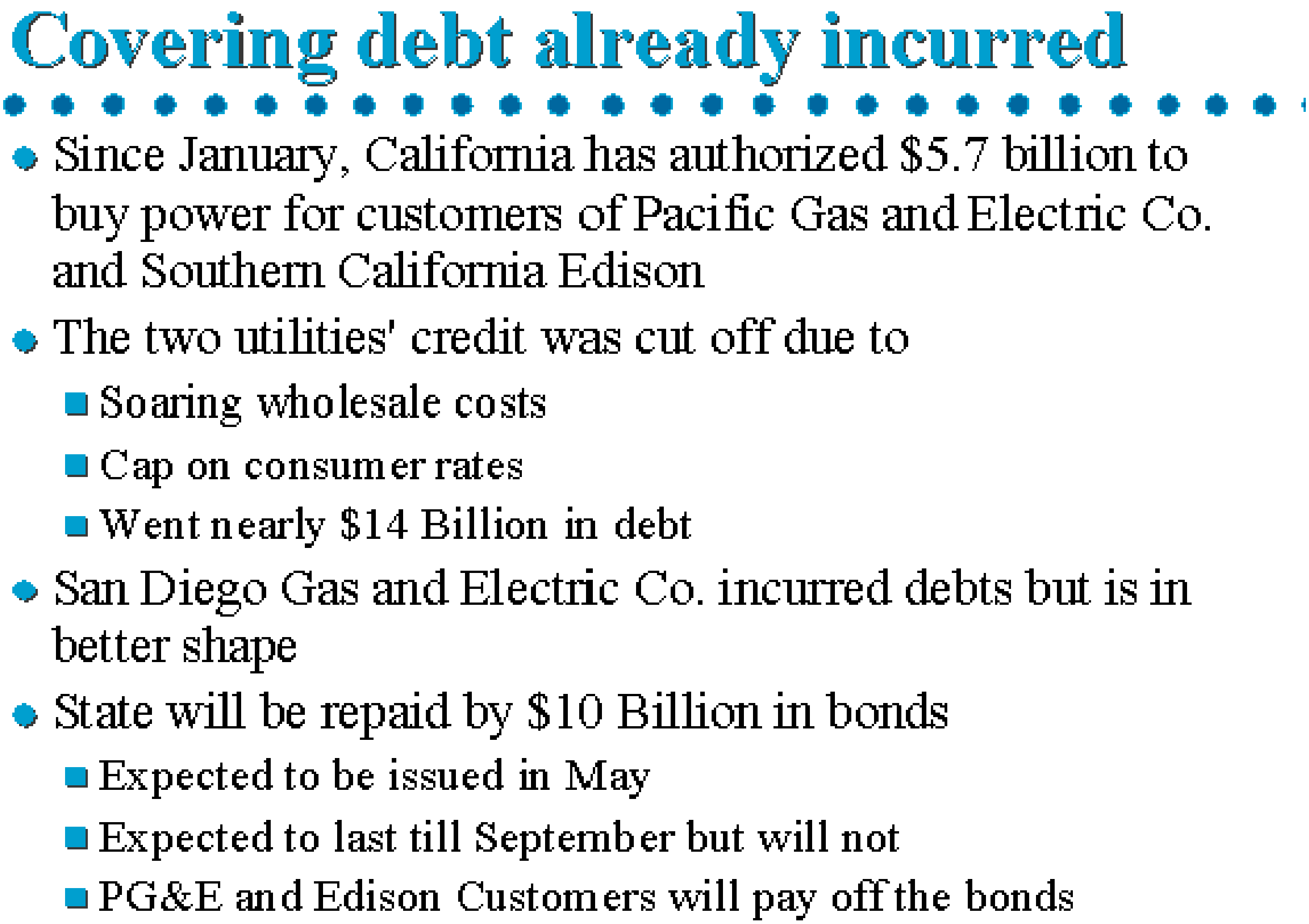

Covering debt already incurred

- Since January, California has authorized $5.7 billion to buy power for customers of Pacific Gas and Electric Co. and Southern California Edison
- The two utilities' credit was cut off due to
  - Soaring wholesale costs
  - Cap on consumer rates
  - Went nearly $14 Billion in debt
- San Diego Gas and Electric Co. incurred debts but is in better shape
- State will be repaid by $10 Billion in bonds
  - Expected to be issued in May
  - Expected to last till September but will not
  - PG&E and Edison Customers will pay off the bonds

Published by Lotus® Freelance Graphics®

Deteriorating situation, rising prices

- Prices rising
  - Will rise even more as demand increases
  - Demand will rise by about 50 percent during the summer
- California's own generation resources cannot meet projected peak demands and reserve requirements
  - May fall by 3,500 megawatts short (peaks)
  - Enough power for 2.6 million households
- Power grid manager must depend on imports
  - Northwest drought makes extra power scarce to find
  - Northwest utilities themselves may have to import, not supply extra

Published by Lotus® Freelance Graphics®

Other Factors

Credit, payback, scarcities

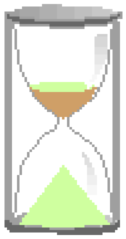

Published by Lotus® Freelance Graphics®

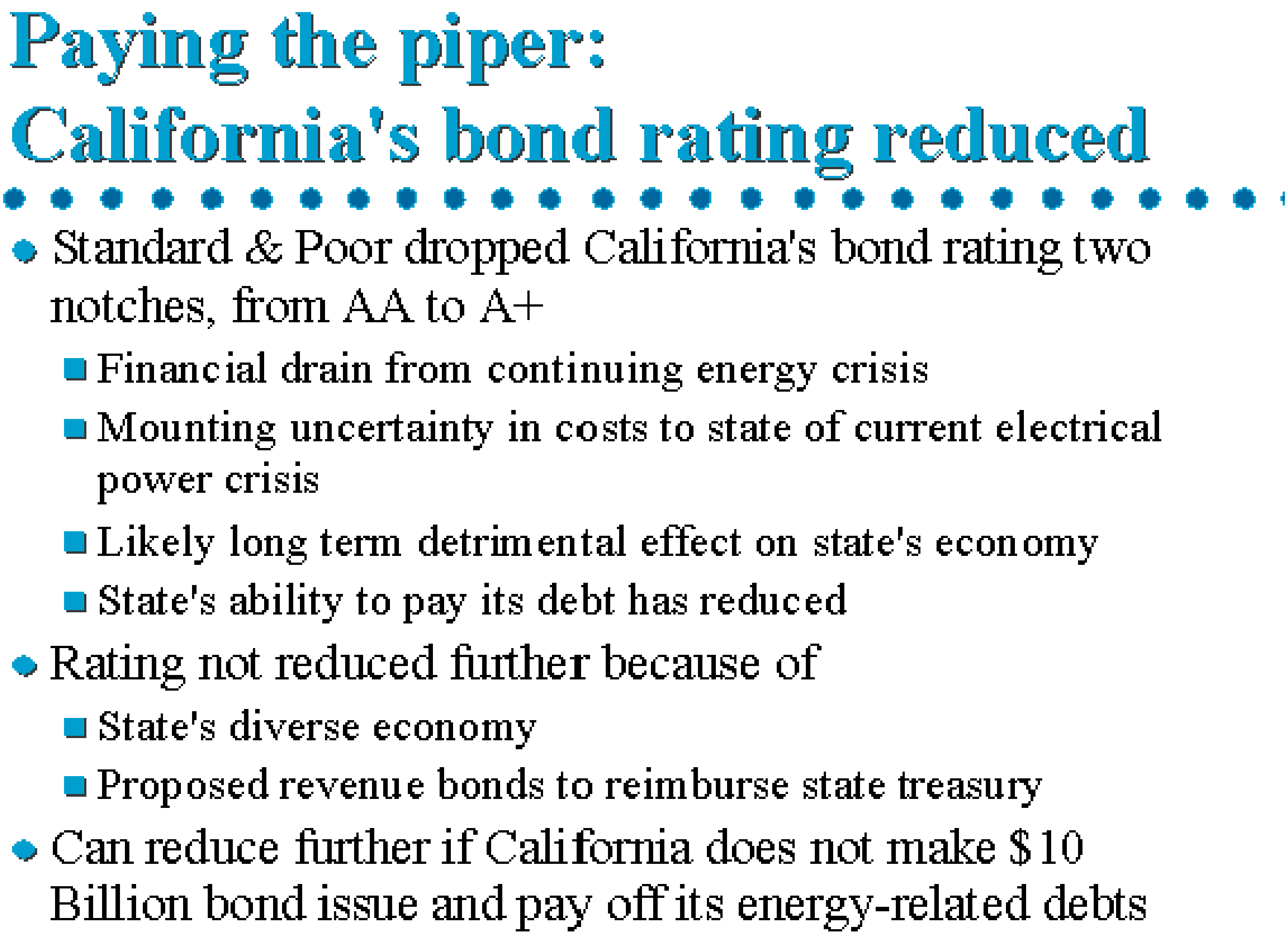

Paying the piper: California's bond rating reduced

- Standard & Poor dropped California's bond rating two notches, from AA to A+
  - Financial drain from continuing energy crisis
  - Mounting uncertainty in costs to state of current electrical power crisis
  - Likely long term detrimental effect on state's economy
  - State's ability to pay its debt has reduced
- Rating not reduced further because of
  - State's diverse economy
  - Proposed revenue bonds to reimburse state treasury
- Can reduce further if California does not make $10 Billion bond issue and pay off its energy-related debts

Published by Lotus® Freelance Graphics®

Associated Press: Gasoline supplies are critical

- Supplies tightest for mid-April since U.S. government record-keeping began in 1963
- There is a shortage of reformulated gasoline that produces less smog and is required during summer
  - MTBE additive made from methanol, a derivative of natural gas
  - Natural gas prices quadrupled Dec-Jan; MTBE reduced by 1/3
  - California uses more MTBE-treated gasoline than any other state
  - Gas-guzzling SUV sales boomed in 1999 and 2000
- Any kind of refinery snag will be amplified
- Last winter, major refiners had to focus on filling heating oil shortage instead of replenishing gasoline supplies for summer
- Now many forced out of service to take care of maintenance that was put off to deal with the heating oil crisis
- Environmental restrictions make it difficult to build refineries; none constructed in the U.S. for more than 20 years

Published by Lotus® Freelance Graphics®

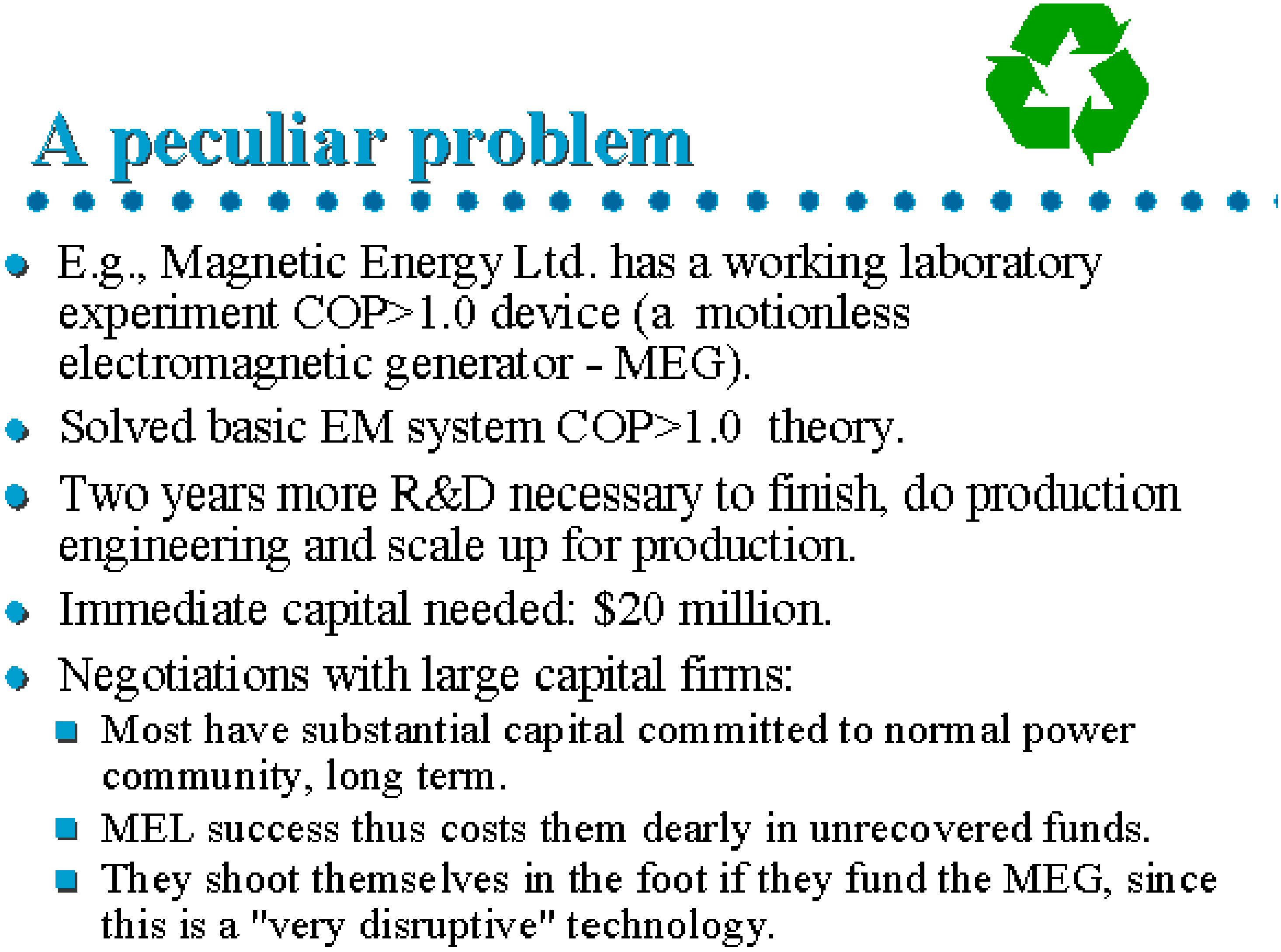

A peculiar problem

- E.g., Magnetic Energy Ltd. has a working laboratory experiment COP>1.0 device (a motionless electromagnetic generator - MEG).
- Solved basic EM system COP>1.0 theory.
- Two years more R&D necessary to finish, do production engineering and scale up for production.
- Immediate capital needed: $20 million.
- Negotiations with large capital firms:
  - Most have substantial capital committed to normal power community, long term.
  - MEL success thus costs them dearly in unrecovered funds.
  - They shoot themselves in the foot if they fund the MEG, since this is a "very disruptive" technology.

Published by Lotus® Freelance Graphics®

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (1)

- The Aharonov-Bohm effect (ABE) is in over 2,000 published physics papers in the hard literature.
- Hundreds of experiments prove it. Physicists themselves would not believe it for 25 years after it was advanced and proven. Now it is universally accepted.
- Free evocation of ABE never appeared for use in power systems, without paying energy to evoke it, until we used ABE in the MEG and *freely* evoked it.
- If the magnetic field **B** of a source is locally confined in a closed path (as in a toroid), the outside local spacetime (ST) where the **B**-field was withdrawn is still curved. Curved ST is energetic. The 3-space energy has been transformed into curl-free magnetic vector potential **A**. But one pays for that **A** around a toroid, since one has to "feed" the toroid.

Published by Lotus® Freelance Graphics®

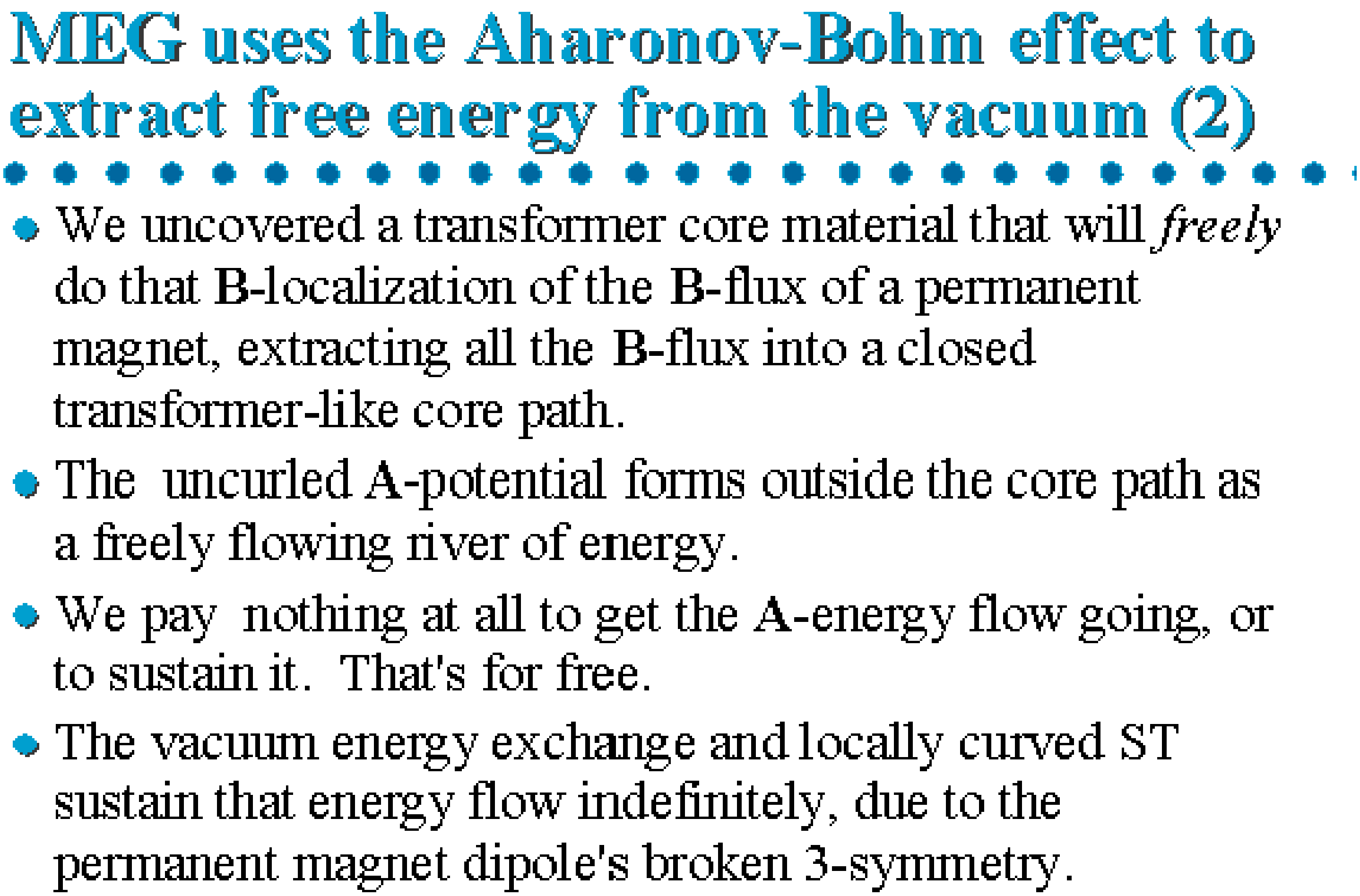

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (2)

- We uncovered a transformer core material that will *freely* do that **B**-localization of the **B**-flux of a permanent magnet, extracting all the **B**-flux into a closed transformer-like core path.
- The uncurled **A**-potential forms outside the core path as a freely flowing river of energy.
- We pay nothing at all to get the **A**-energy flow going, or to sustain it. That's for free.
- The vacuum energy exchange and locally curved ST sustain that energy flow indefinitely, due to the permanent magnet dipole's broken 3-symmetry.

Published by Lotus® Freelance Graphics®

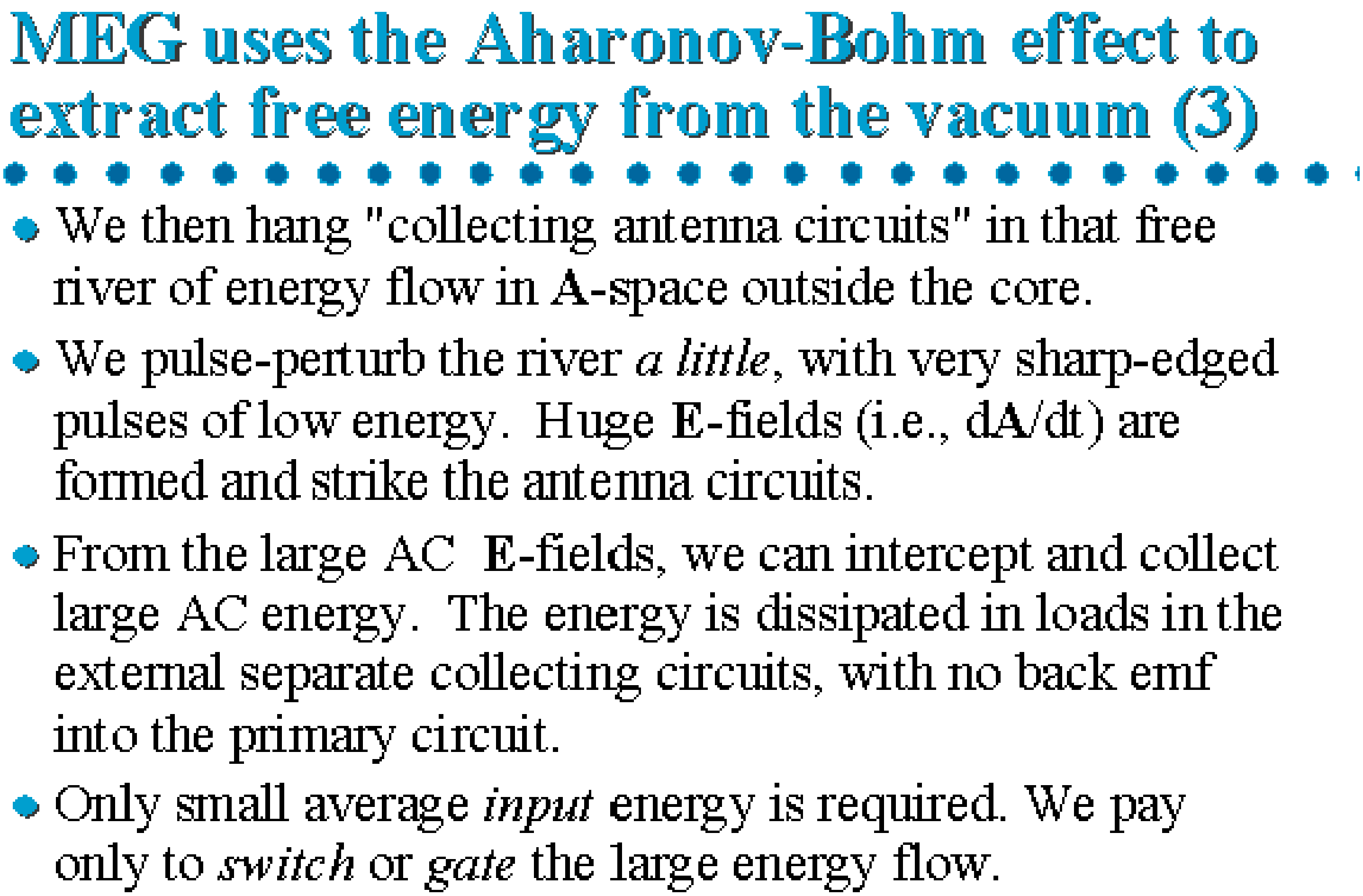

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (3)

- We then hang "collecting antenna circuits" in that free river of energy flow in **A**-space outside the core.
- We pulse-perturb the river *a little*, with very sharp-edged pulses of low energy. Huge **E**-fields (i.e., d**A**/dt) are formed and strike the antenna circuits.
- From the large AC **E**-fields, we can intercept and collect large AC energy. The energy is dissipated in loads in the external separate collecting circuits, with no back emf into the primary circuit.
- Only small average *input* energy is required. We pay only to *switch* or *gate* the large energy flow.

Published by Lotus® Freelance Graphics®

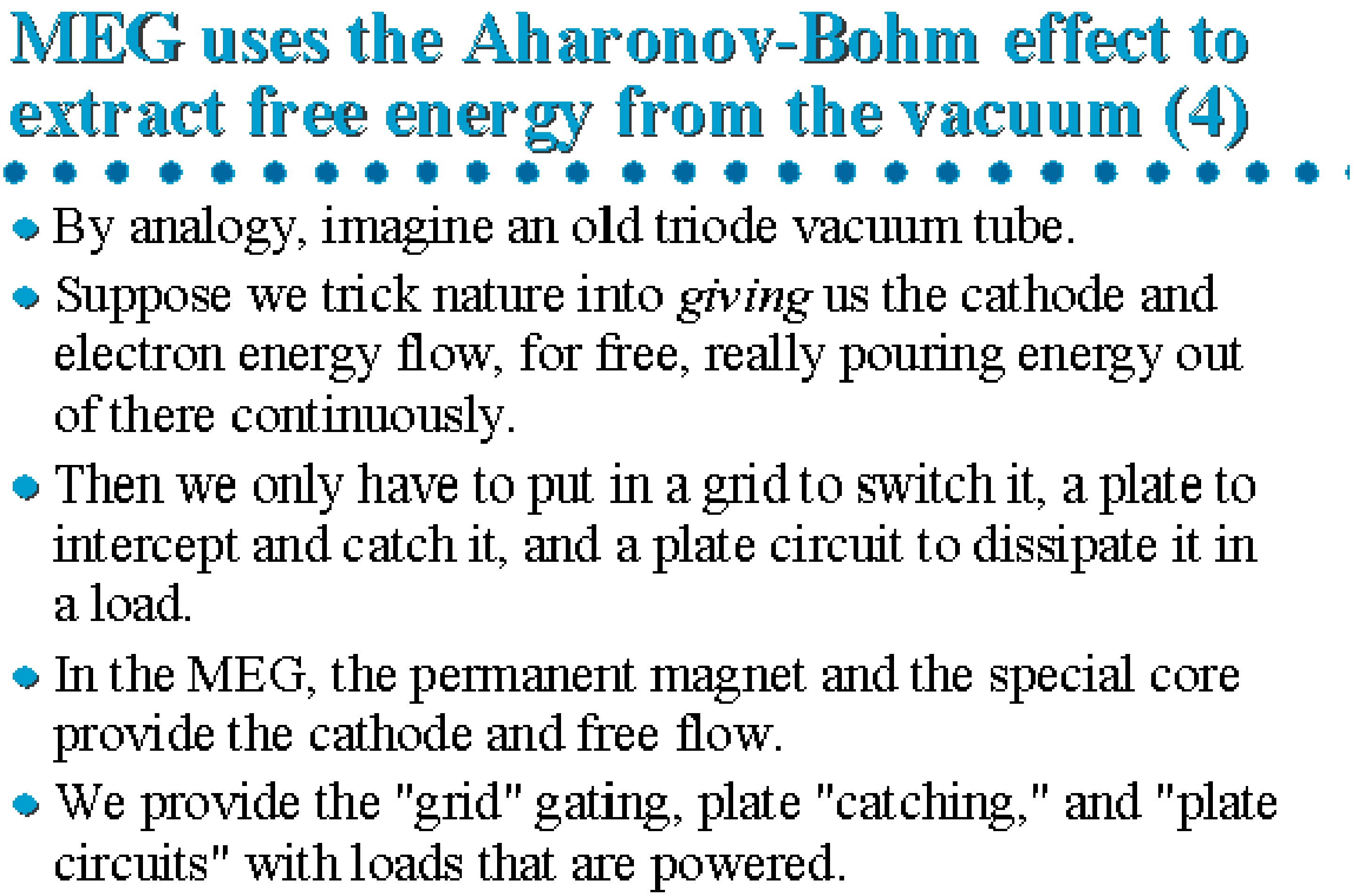

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (4)

- By analogy, imagine an old triode vacuum tube.
- Suppose we trick nature into *giving* us the cathode and electron energy flow, for free, really pouring energy out of there continuously.
- Then we only have to put in a grid to switch it, a plate to intercept and catch it, and a plate circuit to dissipate it in a load.
- In the MEG, the permanent magnet and the special core provide the cathode and free flow.
- We provide the "grid" gating, plate "catching," and "plate circuits" with loads that are powered.

Published by Lotus® Freelance Graphics®

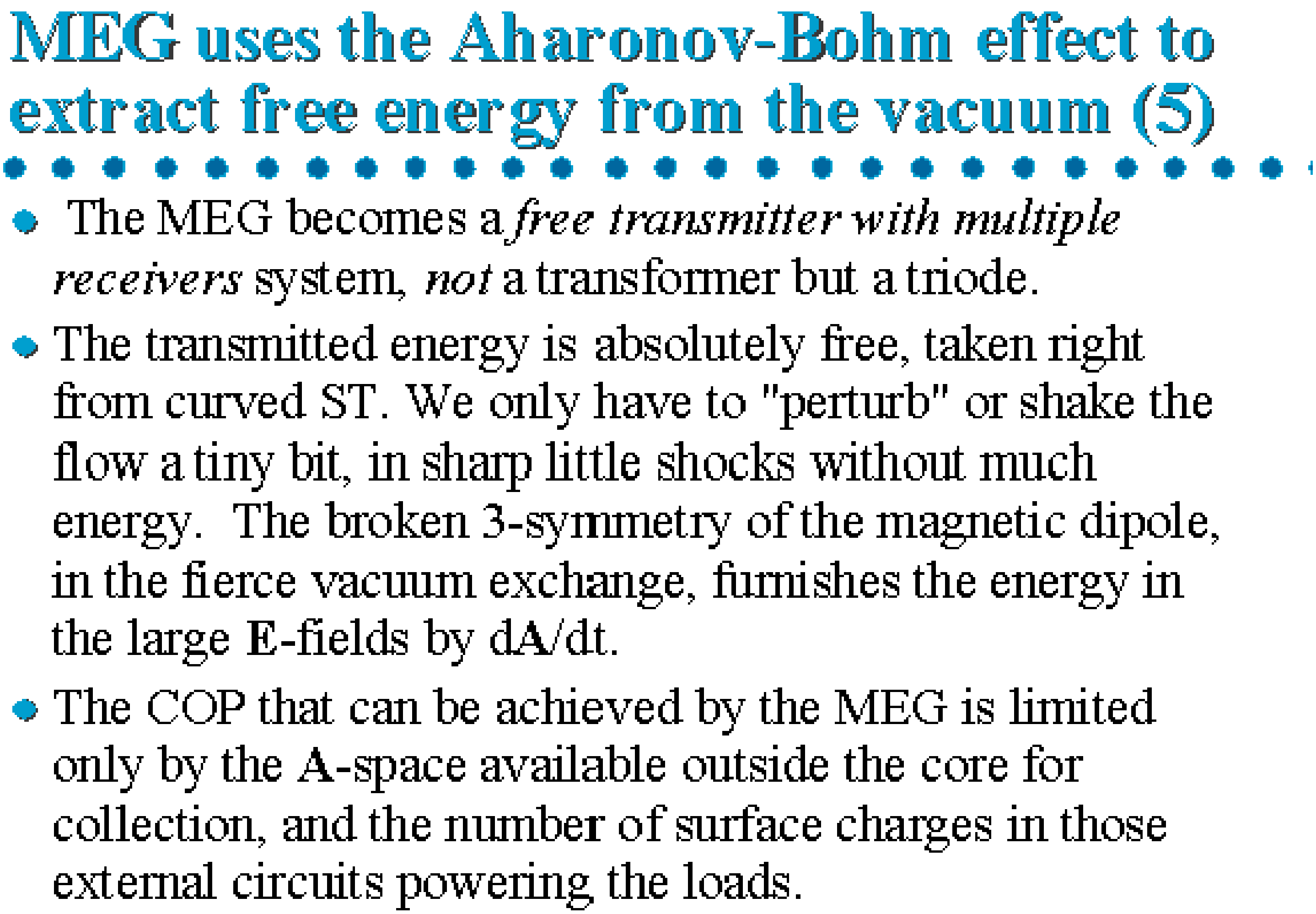

MEG uses the Aharonov-Bohm effect to extract free energy from the vacuum (5)

- The MEG becomes a *free transmitter with multiple receivers* system, *not* a transformer but a triode.
- The transmitted energy is absolutely free, taken right from curved ST. We only have to "perturb" or shake the flow a tiny bit, in sharp little shocks without much energy. The broken 3-symmetry of the magnetic dipole, in the fierce vacuum exchange, furnishes the energy in the large **E**-fields by d**A**/dt.
- The COP that can be achieved by the MEG is limited only by the **A**-space available outside the core for collection, and the number of surface charges in those external circuits powering the loads.

Published by Lotus® Freelance Graphics®

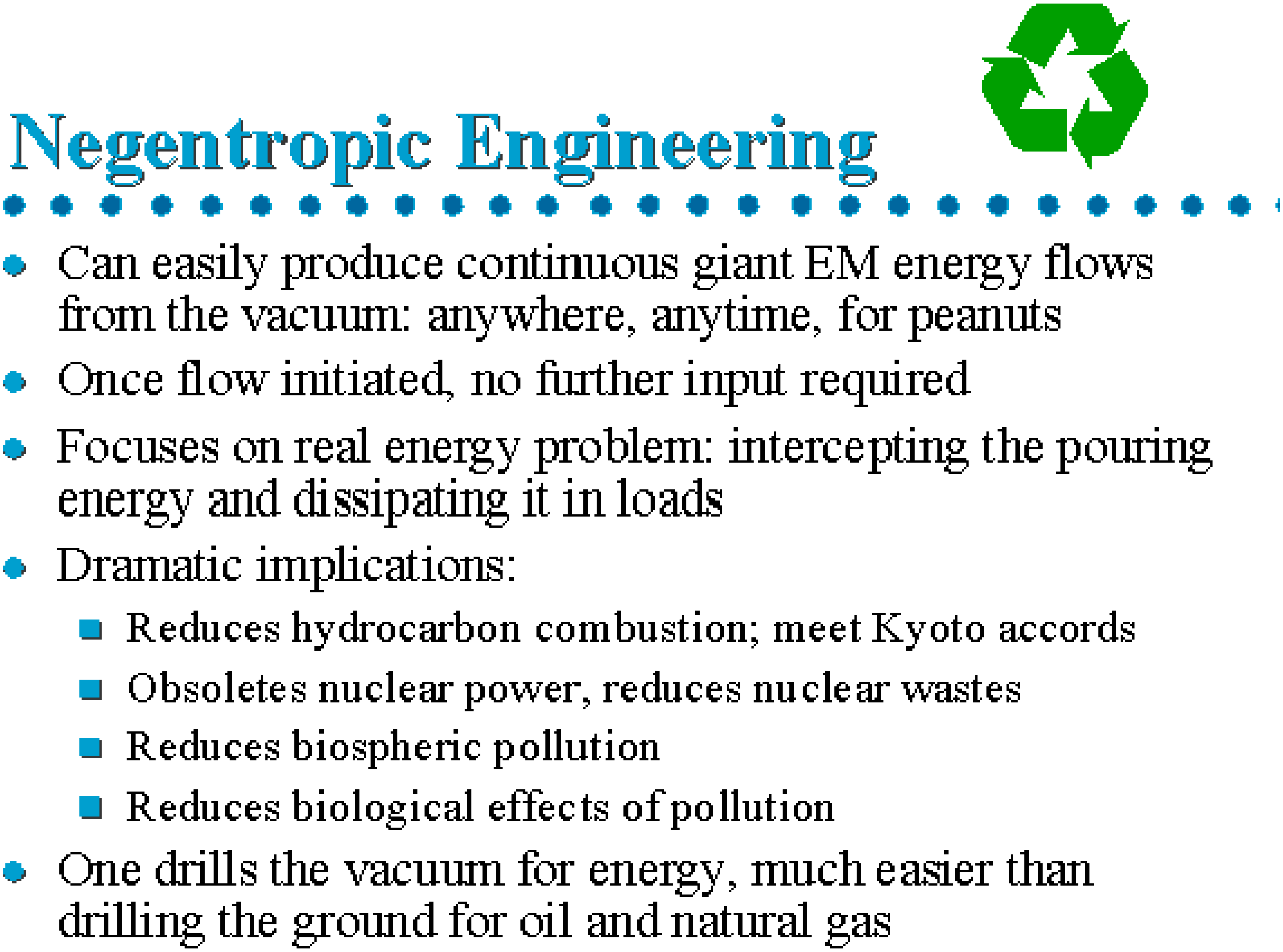

Negentropic Engineering

- Can easily produce continuous giant EM energy flows from the vacuum: anywhere, anytime, for peanuts
- Once flow initiated, no further input required
- Focuses on real energy problem: intercepting the pouring energy and dissipating it in loads
- Dramatic implications:
  - Reduces hydrocarbon combustion; meet Kyoto accords
  - Obsoletes nuclear power, reduces nuclear wastes
  - Reduces biospheric pollution
  - Reduces biological effects of pollution
- One drills the vacuum for energy, much easier than drilling the ground for oil and natural gas

Published by Lotus® Freelance Graphics®

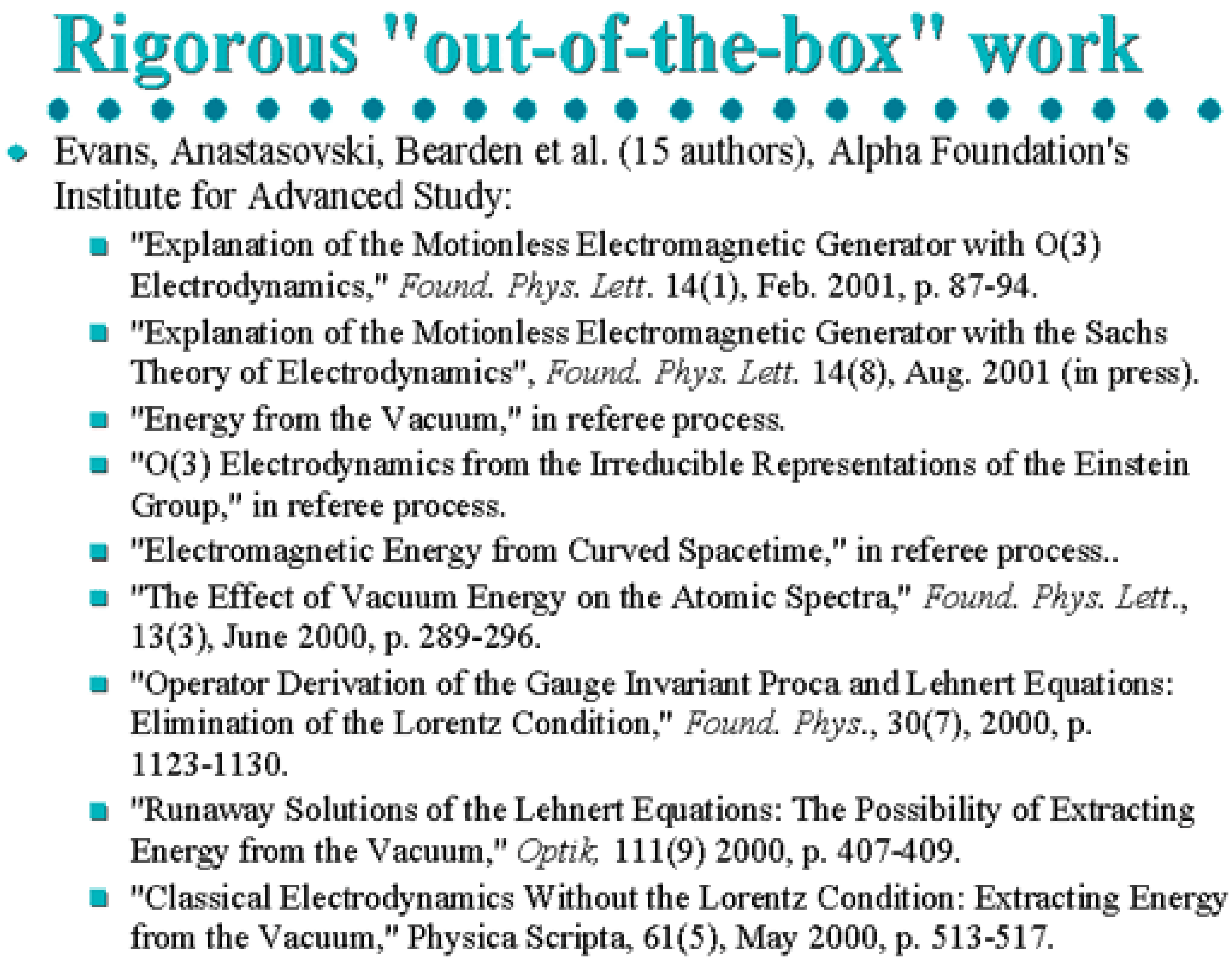

Rigorous "out-of-the-box" work

- Evans, Anastasovski, Bearden et al. (15 authors), Alpha Foundation's Institute for Advanced Study:
  - "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics," *Found. Phys. Lett.* 14(1), Feb. 2001, p. 87-94.
  - "Explanation of the Motionless Electromagnetic Generator with the Sachs Theory of Electrodynamics", *Found. Phys. Lett.* 14(8), Aug. 2001 (in press).
  - "Energy from the Vacuum," in referee process.
  - "O(3) Electrodynamics from the Irreducible Representations of the Einstein Group," in referee process.
  - "Electromagnetic Energy from Curved Spacetime," in referee process..
  - "The Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys. Lett.*, 13(3), June 2000, p. 289-296.
  - "Operator Derivation of the Gauge Invariant Proca and Lehnert Equations: Elimination of the Lorentz Condition," *Found. Phys.*, 30(7), 2000, p. 1123-1130.
  - "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik*, 111(9) 2000, p. 407-409.
  - "Classical Electrodynamics Without the Lorentz Condition: Extracting Energy from the Vacuum," Physica Scripta, 61(5), May 2000, p. 513-517.

Published by Lotus® Freelance Graphics®

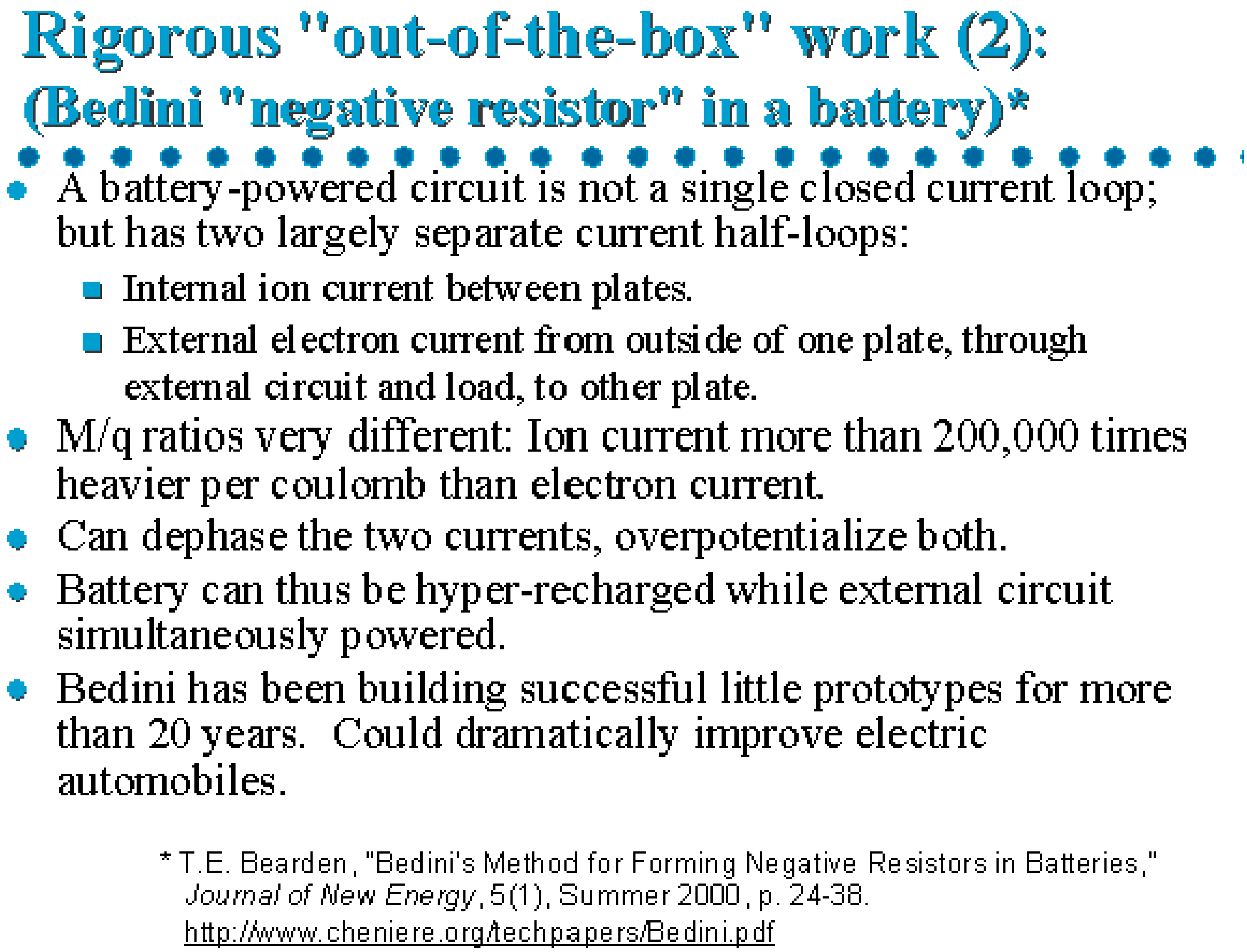

Rigorous "out-of-the-box" work (2): (Bedini "negative resistor" in a battery)*

- A battery-powered circuit is not a single closed current loop; but has two largely separate current half-loops:
  - Internal ion current between plates.
  - External electron current from outside of one plate, through external circuit and load, to other plate.
- M/q ratios very different: Ion current more than 200,000 times heavier per coulomb than electron current.
- Can dephase the two currents, overpotentialize both.
- Battery can thus be hyper-recharged while external circuit simultaneously powered.
- Bedini has been building successful little prototypes for more than 20 years. Could dramatically improve electric automobiles.

* T.E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," *Journal of New Energy*, 5(1), Summer 2000, p. 24-38. http://www.cheniere.org/techpapers/Bedini.pdf

Published by Lotus® Freelance Graphics®

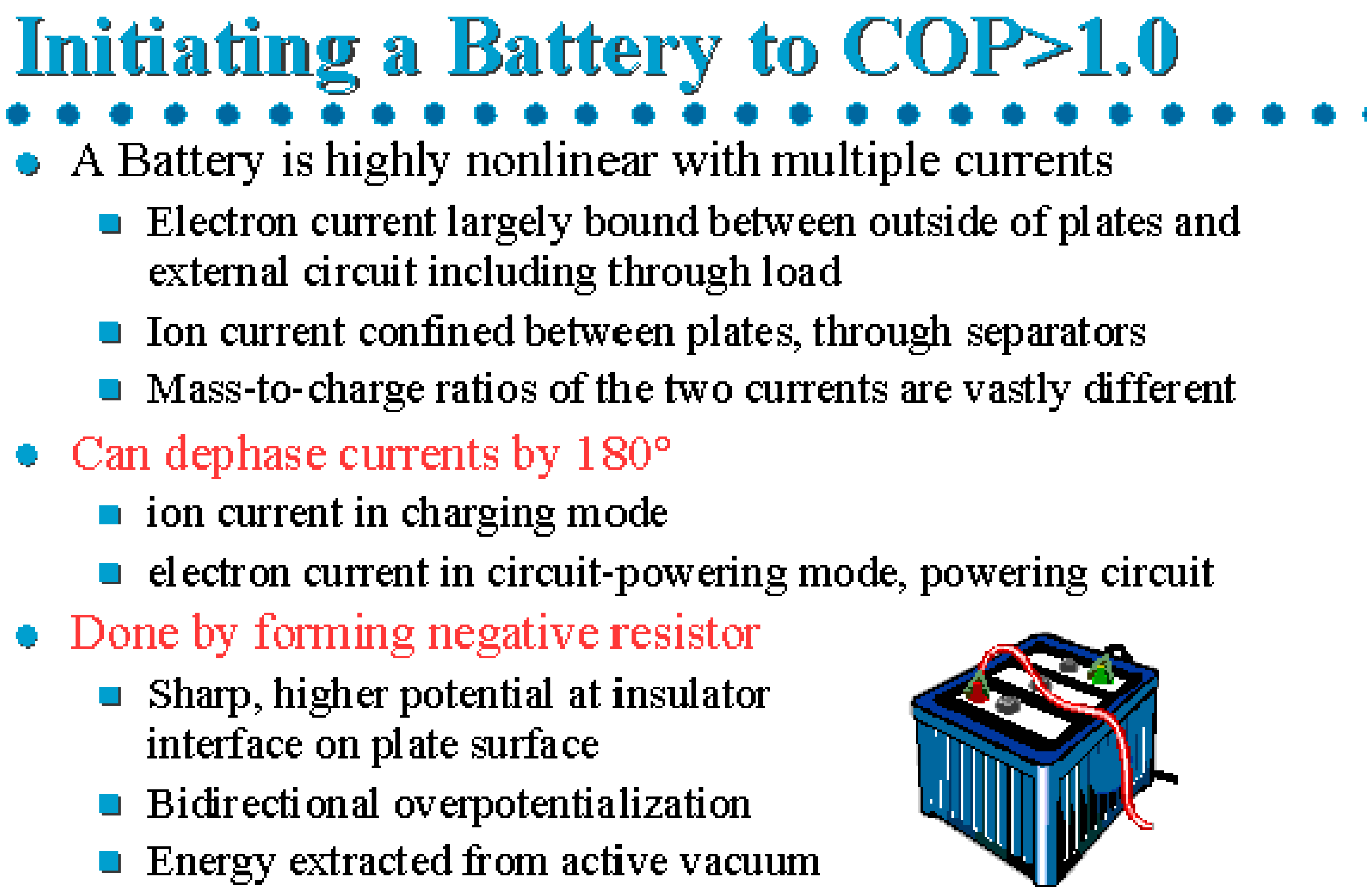

Initiating a Battery to COP>1.0

- A Battery is highly nonlinear with multiple currents
  - Electron current largely bound between outside of plates and external circuit including through load
  - Ion current confined between plates, through separators
  - Mass-to-charge ratios of the two currents are vastly different
- Can dephase currents by 180°
  - ion current in charging mode
  - electron current in circuit-powering mode, powering circuit
- Done by forming negative resistor
  - Sharp, higher potential at insulator interface on plate surface
  - Bidirectional overpotentialization
  - Energy extracted from active vacuum

Published by Lotus® Freelance Graphics®

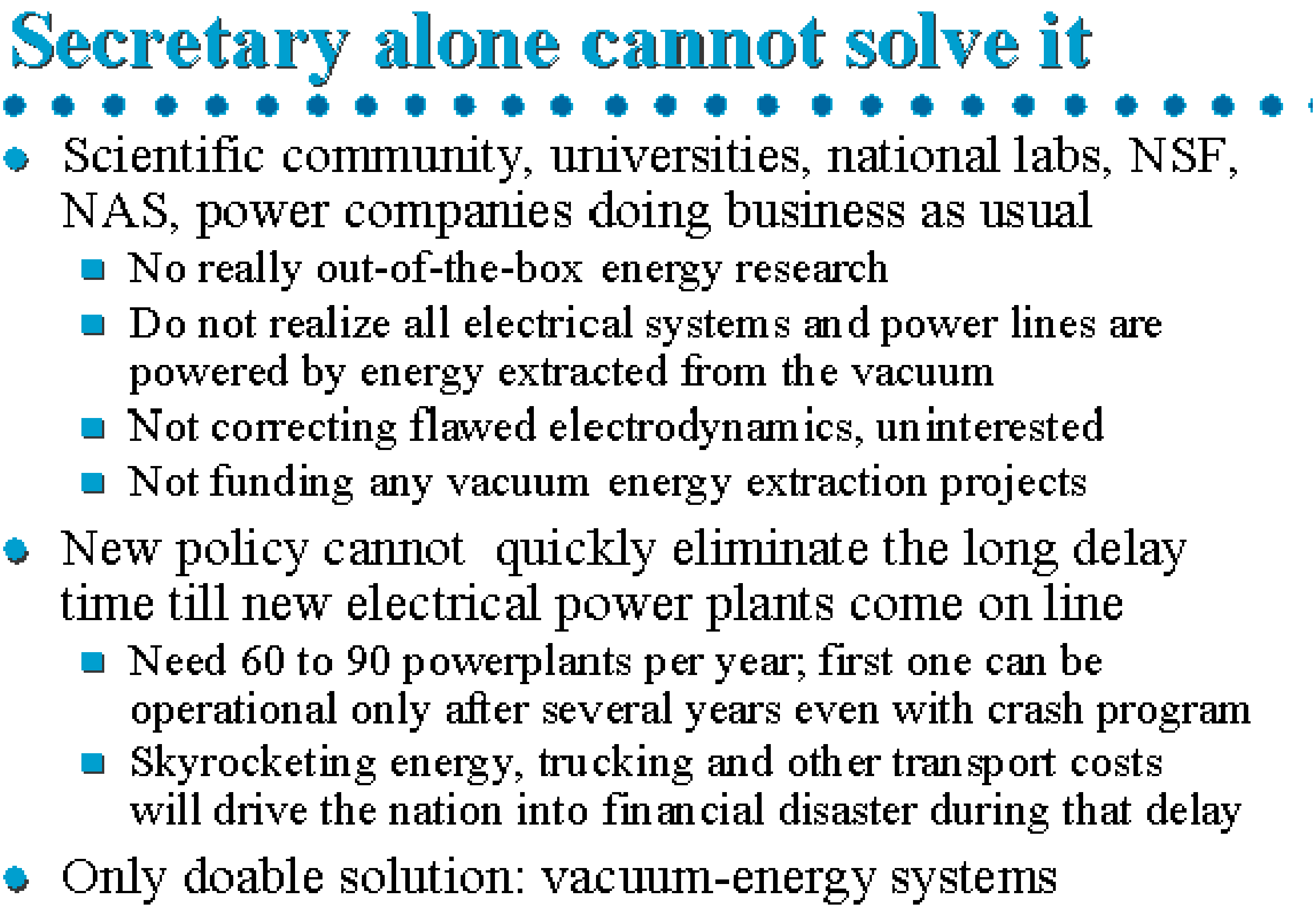

Secretary alone cannot solve it

- Scientific community, universities, national labs, NSF, NAS, power companies doing business as usual
    - No really out-of-the-box energy research
    - Do not realize all electrical systems and power lines are powered by energy extracted from the vacuum
    - Not correcting flawed electrodynamics, uninterested
    - Not funding any vacuum energy extraction projects
- New policy cannot quickly eliminate the long delay time till new electrical power plants come on line
    - Need 60 to 90 powerplants per year; first one can be operational only after several years even with crash program
    - Skyrocketing energy, trucking and other transport costs will drive the nation into financial disaster during that delay
- Only doable solution: vacuum-energy systems

Published by Lotus® Freelance Graphics®

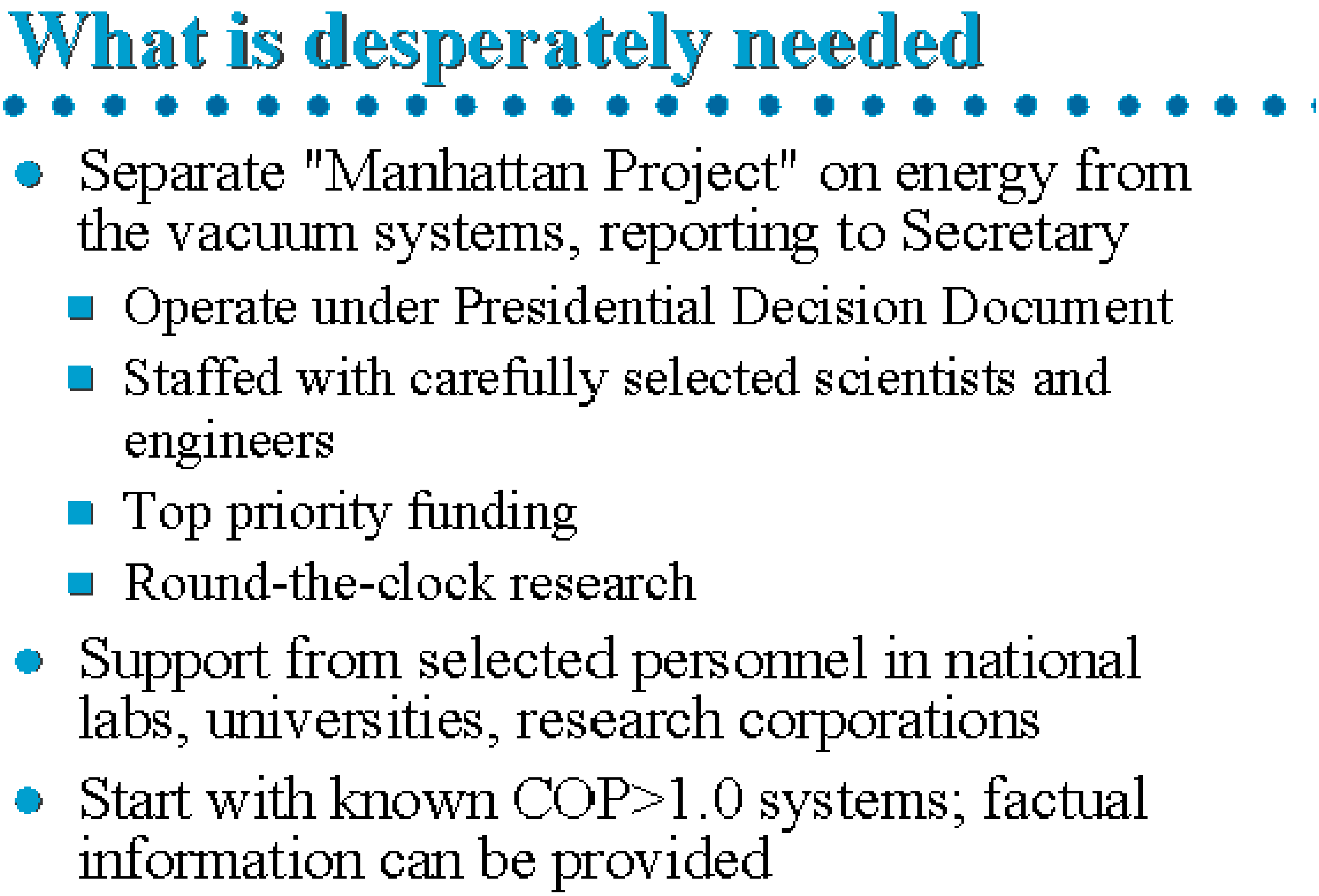

What is desperately needed

- Separate "Manhattan Project" on energy from the vacuum systems, reporting to Secretary
  - Operate under Presidential Decision Document
  - Staffed with carefully selected scientists and engineers
  - Top priority funding
  - Round-the-clock research
- Support from selected personnel in national labs, universities, research corporations
- Start with known COP>1.0 systems; factual information can be provided

Published by Lotus® Freelance Graphics®

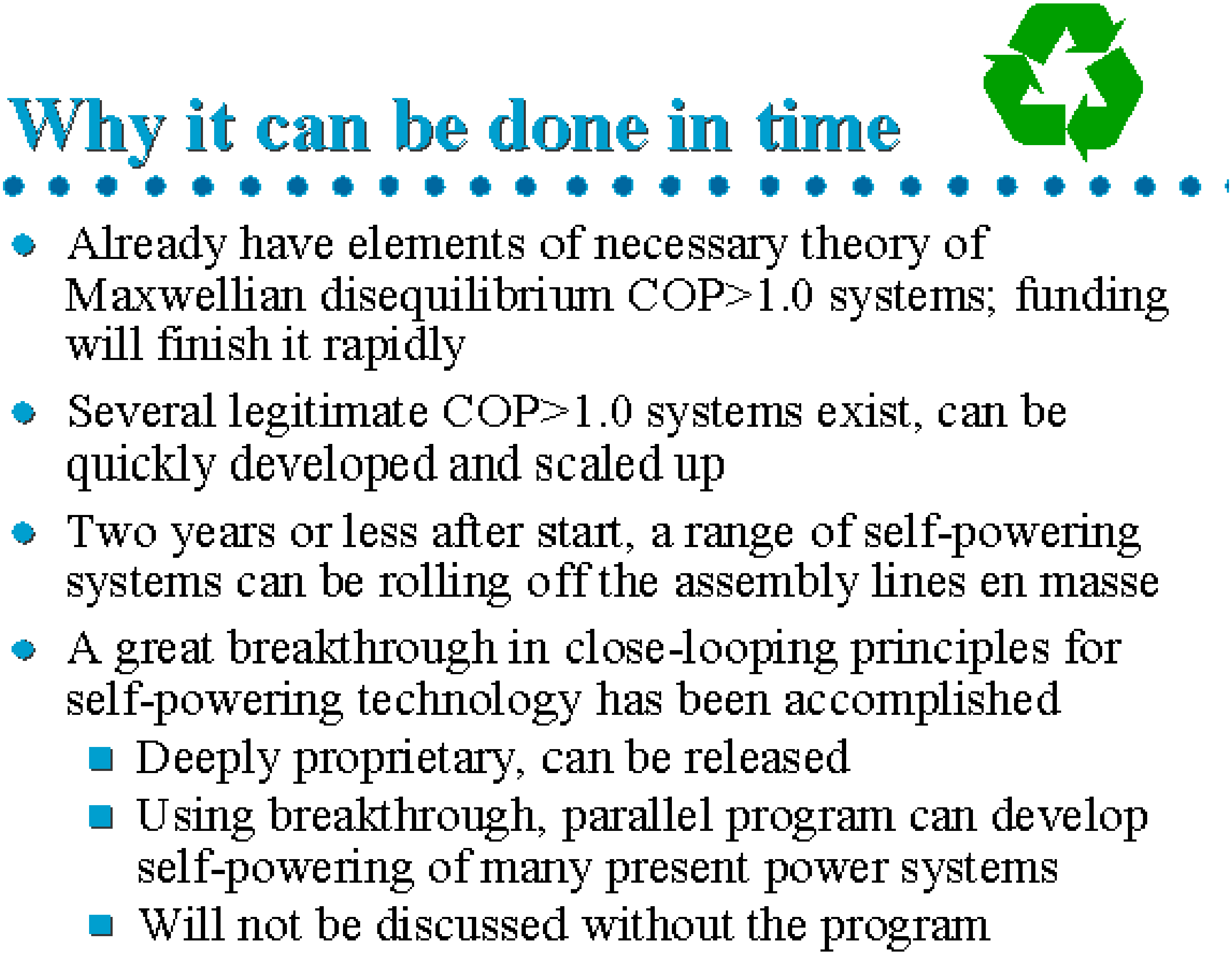

Why it can be done in time

- Already have elements of necessary theory of Maxwellian disequilibrium COP>1.0 systems; funding will finish it rapidly
- Several legitimate COP>1.0 systems exist, can be quickly developed and scaled up
- Two years or less after start, a range of self-powering systems can be rolling off the assembly lines en masse
- A great breakthrough in close-looping principles for self-powering technology has been accomplished
  - Deeply proprietary, can be released
  - Using breakthrough, parallel program can develop self-powering of many present power systems
  - Will not be discussed without the program

Published by Lotus® Freelance Graphics®

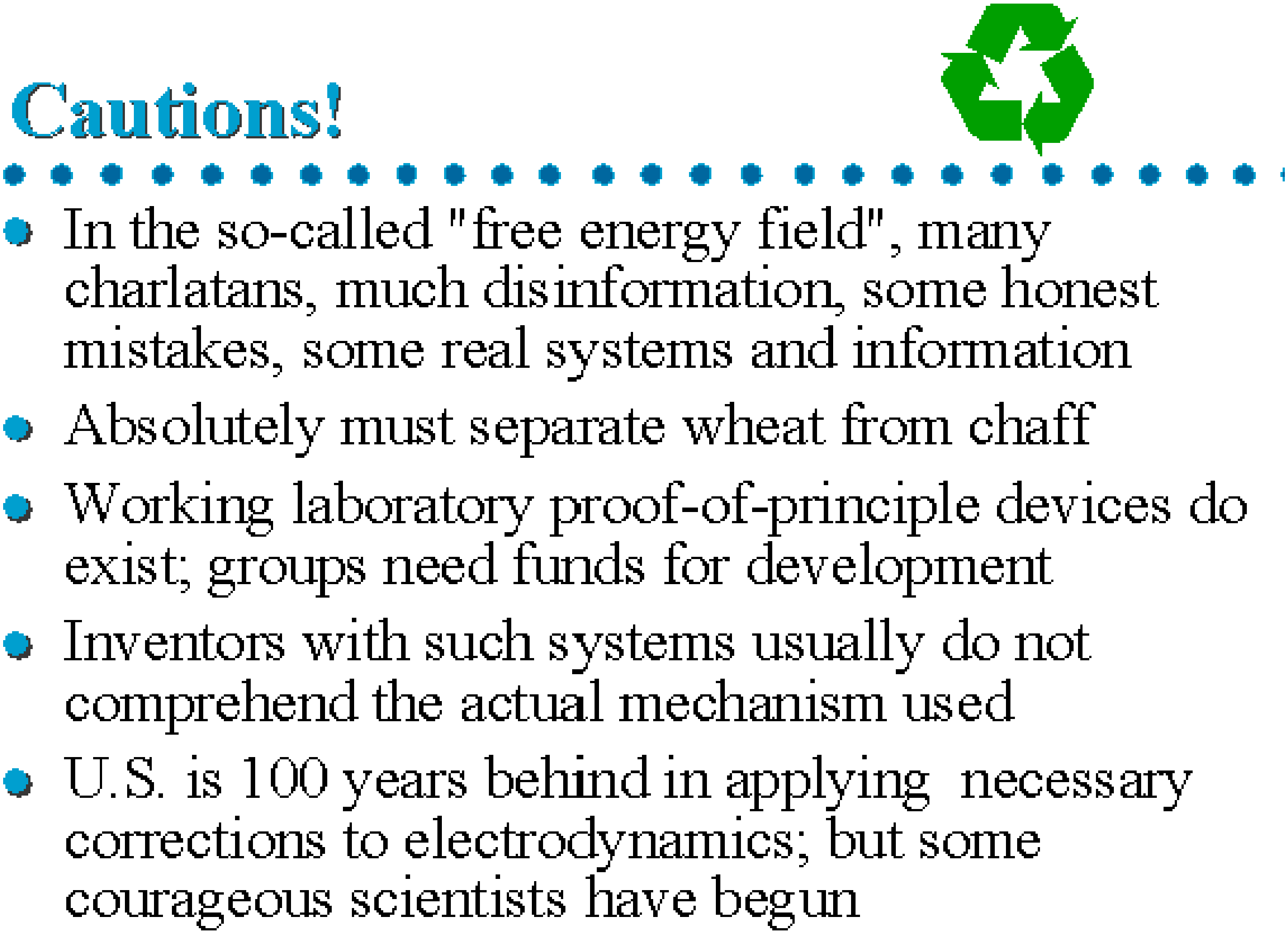

Cautions!

- In the so-called "free energy field", many charlatans, much disinformation, some honest mistakes, some real systems and information
- Absolutely must separate wheat from chaff
- Working laboratory proof-of-principle devices do exist; groups need funds for development
- Inventors with such systems usually do not comprehend the actual mechanism used
- U.S. is 100 years behind in applying necessary corrections to electrodynamics; but some courageous scientists have begun

Published by Lotus® Freelance Graphics®

The scientists have failed us

- Secretary Abraham, it is the U.S. scientific community which has failed us, and which is responsible for this unnecessary crisis.
- No textbook presently even correctly states what actually powers a power line. The basis has been in particle physics for a half-century.
- Before arbitrary regauging by Lorentz in the 1880s, the Maxwell-Heaviside theory does prescribe electrical power systems far from equilibrium in their fierce vacuum exchange, including systems which power themselves and their load, extracting the energy from their broken symmetry in their active vacuum energy exchange.
- Every power system already wastes a trillion times more energy than is placed on the power line; scientists and engineers do not even know it.
- Heaviside published the basis for that in the 1880s and 1890s; Poynting never even knew it. Lorentz knew it, couldn't explain the source, stated "no physical significance" and discarded it.
- This discarded "Heaviside energy" is what is generating the extra gravity in spiral galaxies, and holding their arms together. Scientists arbitrarily discarded it in the 1880s; have forgotten it today.

Published by Lotus® Freelance Graphics®

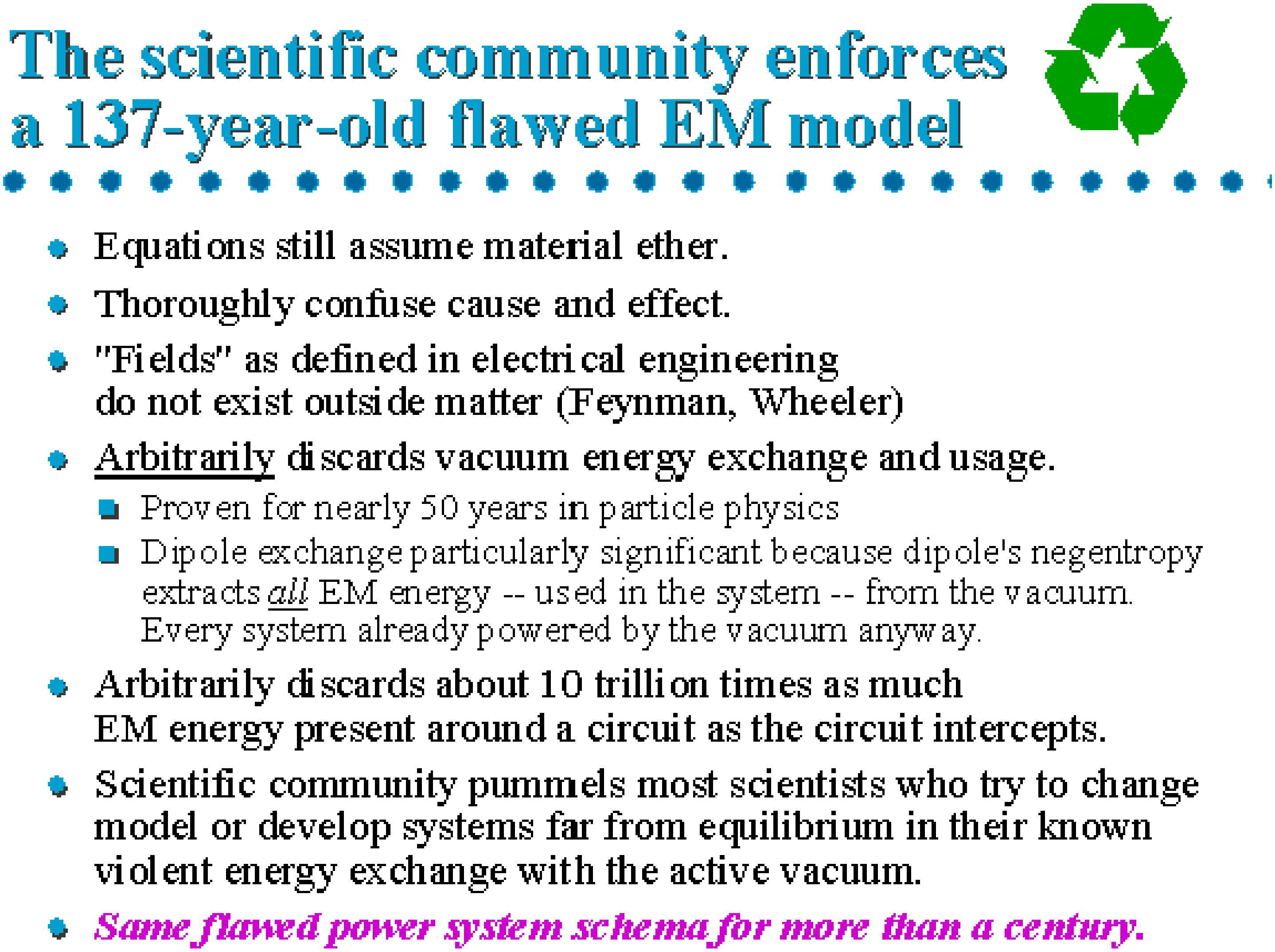

The scientific community enforces a 137-year-old flawed EM model

- Equations still assume material ether.
- Thoroughly confuse cause and effect.
- "Fields" as defined in electrical engineering do not exist outside matter (Feynman, Wheeler)
- Arbitrarily discards vacuum energy exchange and usage.
  - Proven for nearly 50 years in particle physics
  - Dipole exchange particularly significant because dipole's negentropy extracts *all* EM energy -- used in the system -- from the vacuum. Every system already powered by the vacuum anyway.
- Arbitrarily discards about 10 trillion times as much EM energy present around a circuit as the circuit intercepts.
- Scientific community pummels most scientists who try to change model or develop systems far from equilibrium in their known violent energy exchange with the active vacuum.
- ***Same flawed power system schema for more than a century.***

Published by Lotus® Freelance Graphics®

All EM 3-space energy freely comes from the time domain*

* T. E. Bearden, "Giant Negentropy of the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.
http://www.cheniere.org/techpapers/GiantNegentropy.pdf

Whittaker, Math. Ann., 57, 333 (1903)
shows this rigorously.
Dipole's broken 3-symmetry well-known
in particle physics

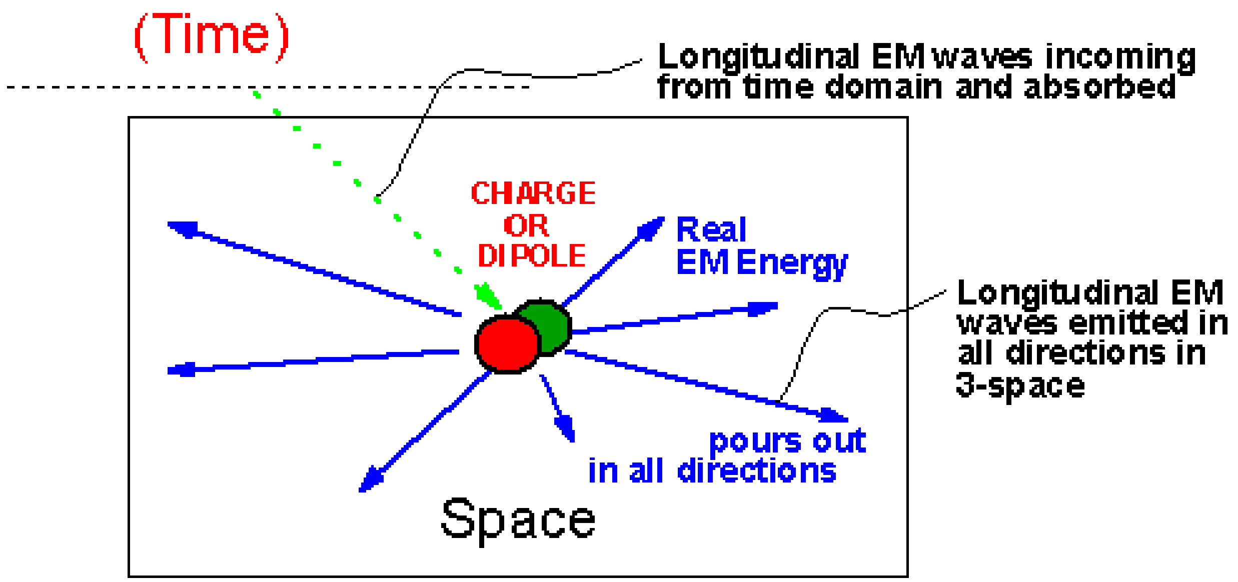
(Time)
Longitudinal EM waves incoming from time domain and absorbed
CHARGE OR DIPOLE
Real EM Energy
Longitudinal EM waves emitted in all directions in 3-space
pours out in all directions
Space

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs *after* interaction with the charges of the dipole; and as an *effect* rather than the *cause*. This fundamental non sequitur has just been repeated since then. Thus they missed the solution of the source charge and source dipole.

Published by Lotus® Freelance Graphics®

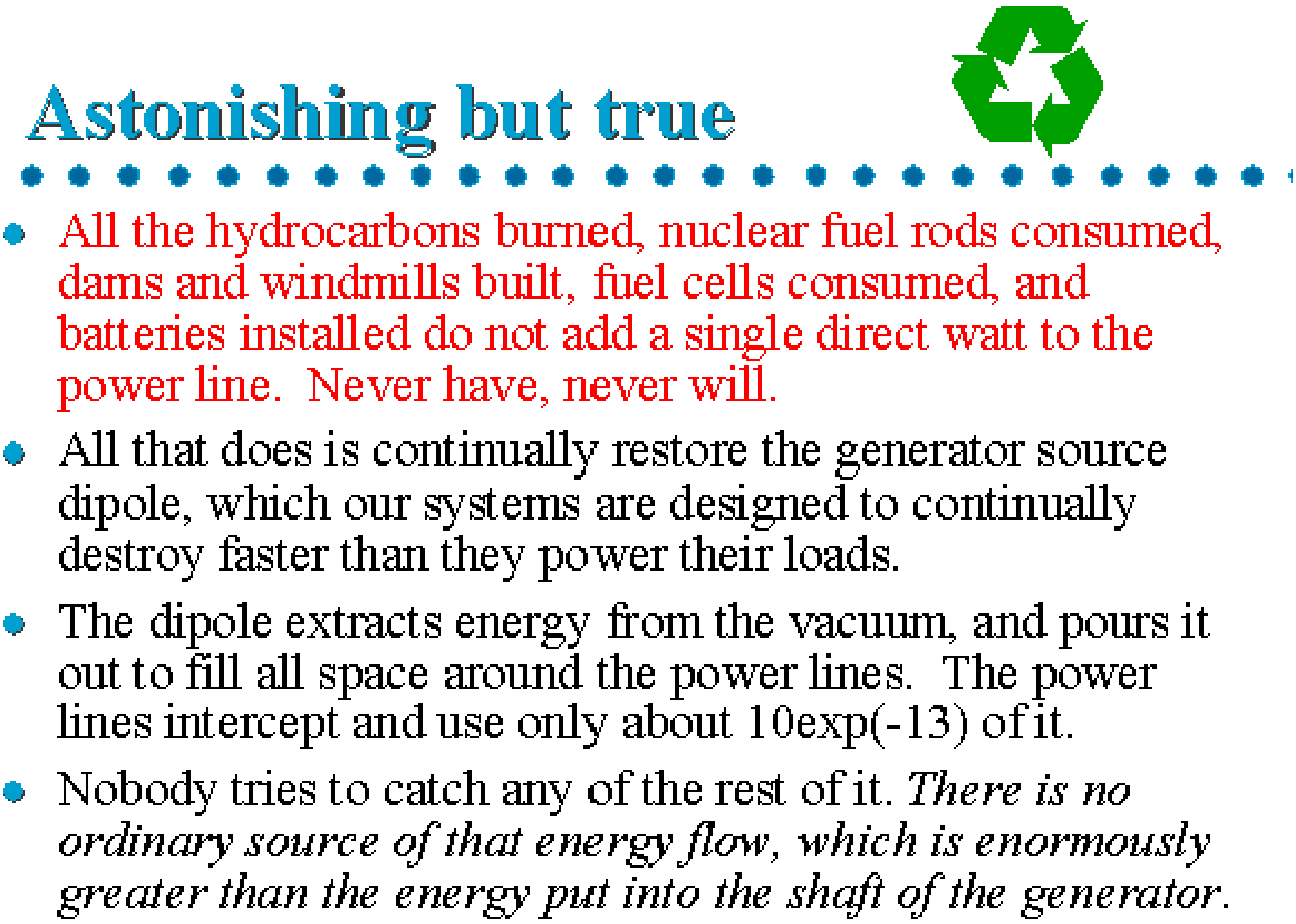

Astonishing but true

- All the hydrocarbons burned, nuclear fuel rods consumed, dams and windmills built, fuel cells consumed, and batteries installed do not add a single direct watt to the power line. Never have, never will.
- All that does is continually restore the generator source dipole, which our systems are designed to continually destroy faster than they power their loads.
- The dipole extracts energy from the vacuum, and pours it out to fill all space around the power lines. The power lines intercept and use only about 10exp(-13) of it.
- Nobody tries to catch any of the rest of it. *There is no ordinary source of that energy flow, which is enormously greater than the energy put into the shaft of the generator.*

Published by Lotus® Freelance Graphics®

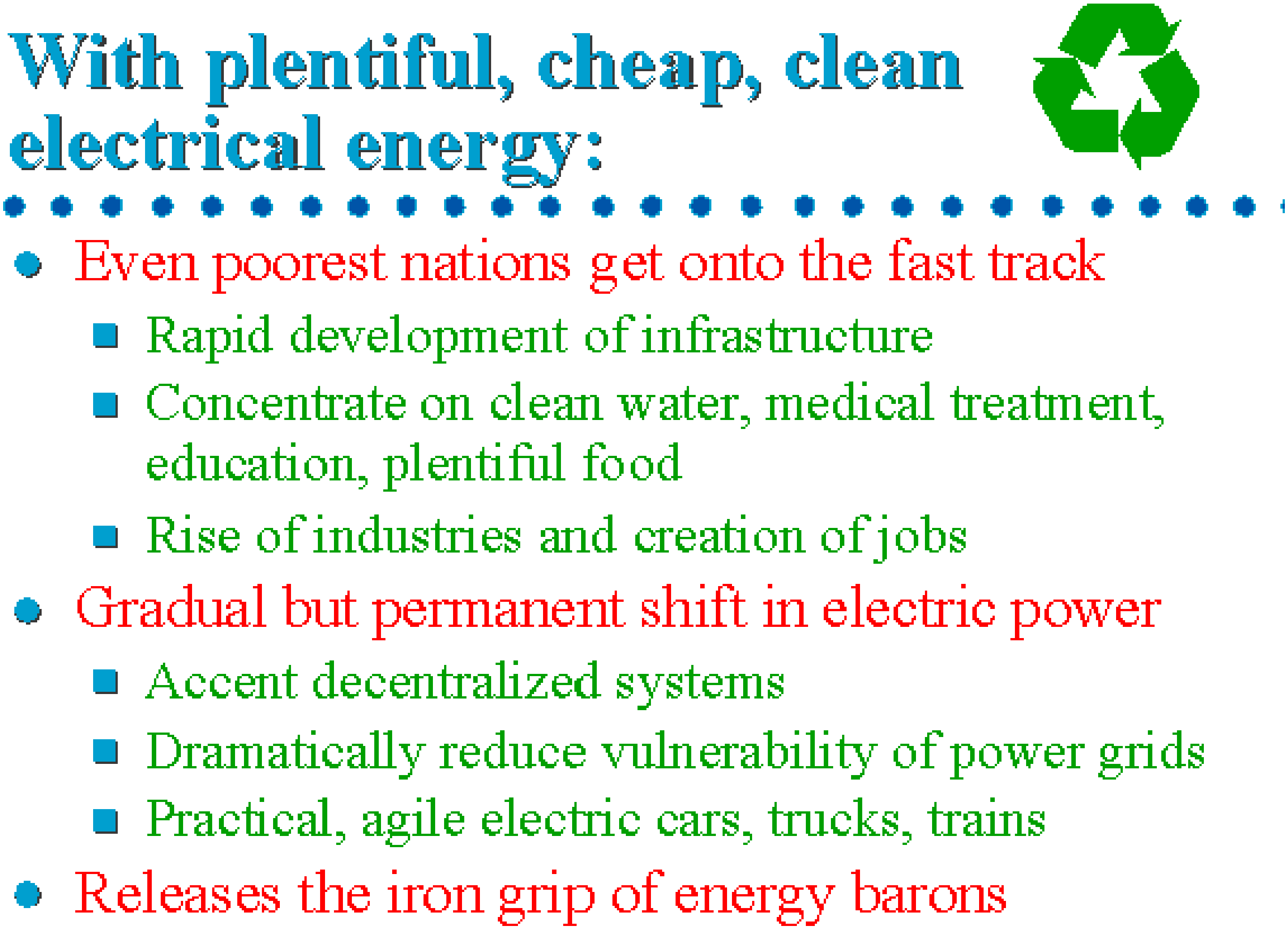

With plentiful, cheap, clean electrical energy:

- Even poorest nations get onto the fast track
  - Rapid development of infrastructure
  - Concentrate on clean water, medical treatment, education, plentiful food
  - Rise of industries and creation of jobs
- Gradual but permanent shift in electric power
  - Accent decentralized systems
  - Dramatically reduce vulnerability of power grids
  - Practical, agile electric cars, trucks, trains
- Releases the iron grip of energy barons

Published by Lotus® Freelance Graphics®

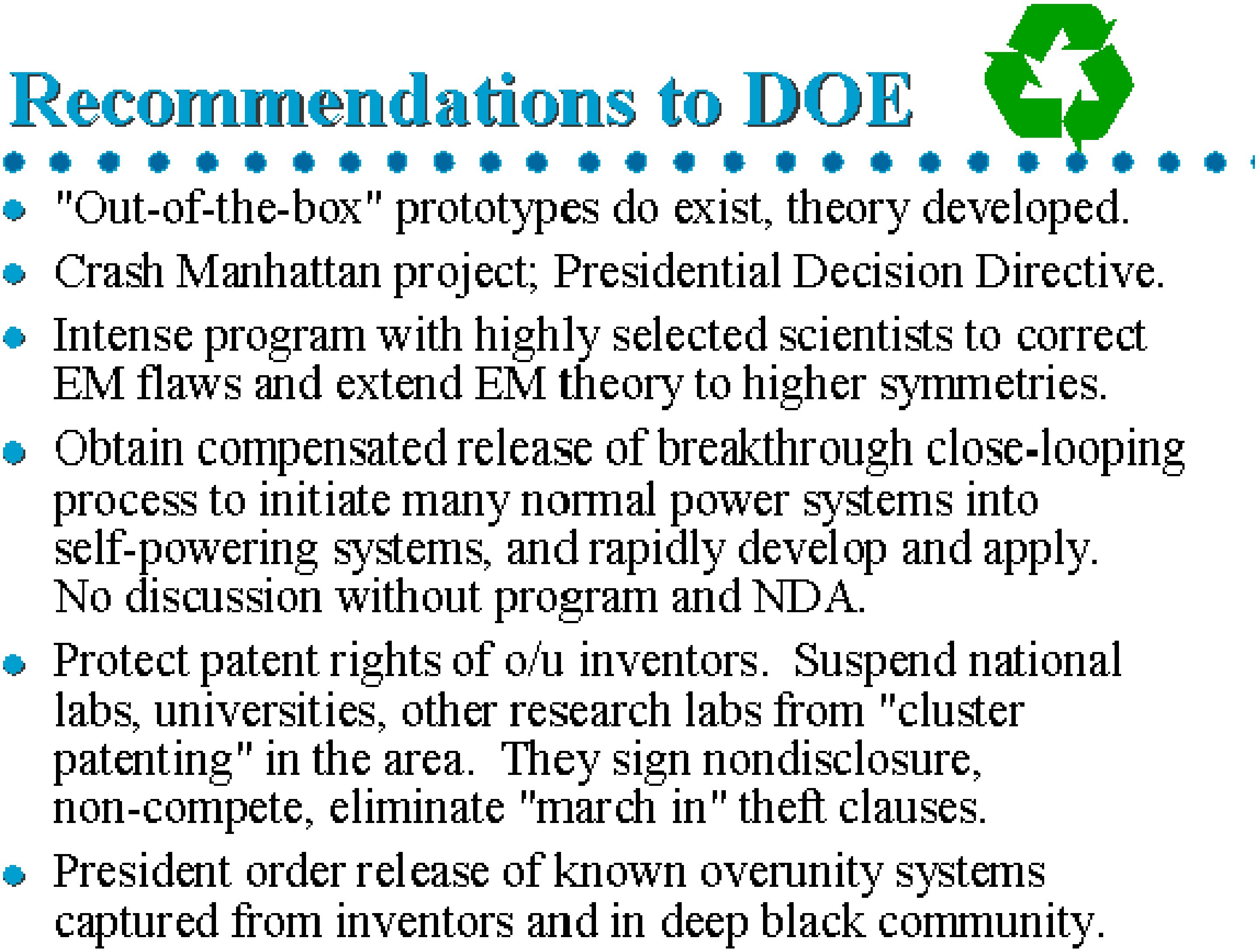

Recommendations to DOE

- "Out-of-the-box" prototypes do exist, theory developed.
- Crash Manhattan project; Presidential Decision Directive.
- Intense program with highly selected scientists to correct EM flaws and extend EM theory to higher symmetries.
- Obtain compensated release of breakthrough close-looping process to initiate many normal power systems into self-powering systems, and rapidly develop and apply. No discussion without program and NDA.
- Protect patent rights of o/u inventors. Suspend national labs, universities, other research labs from "cluster patenting" in the area. They sign nondisclosure, non-compete, eliminate "march in" theft clauses.
- President order release of known overunity systems captured from inventors and in deep black community.

Published by Lotus® Freelance Graphics®

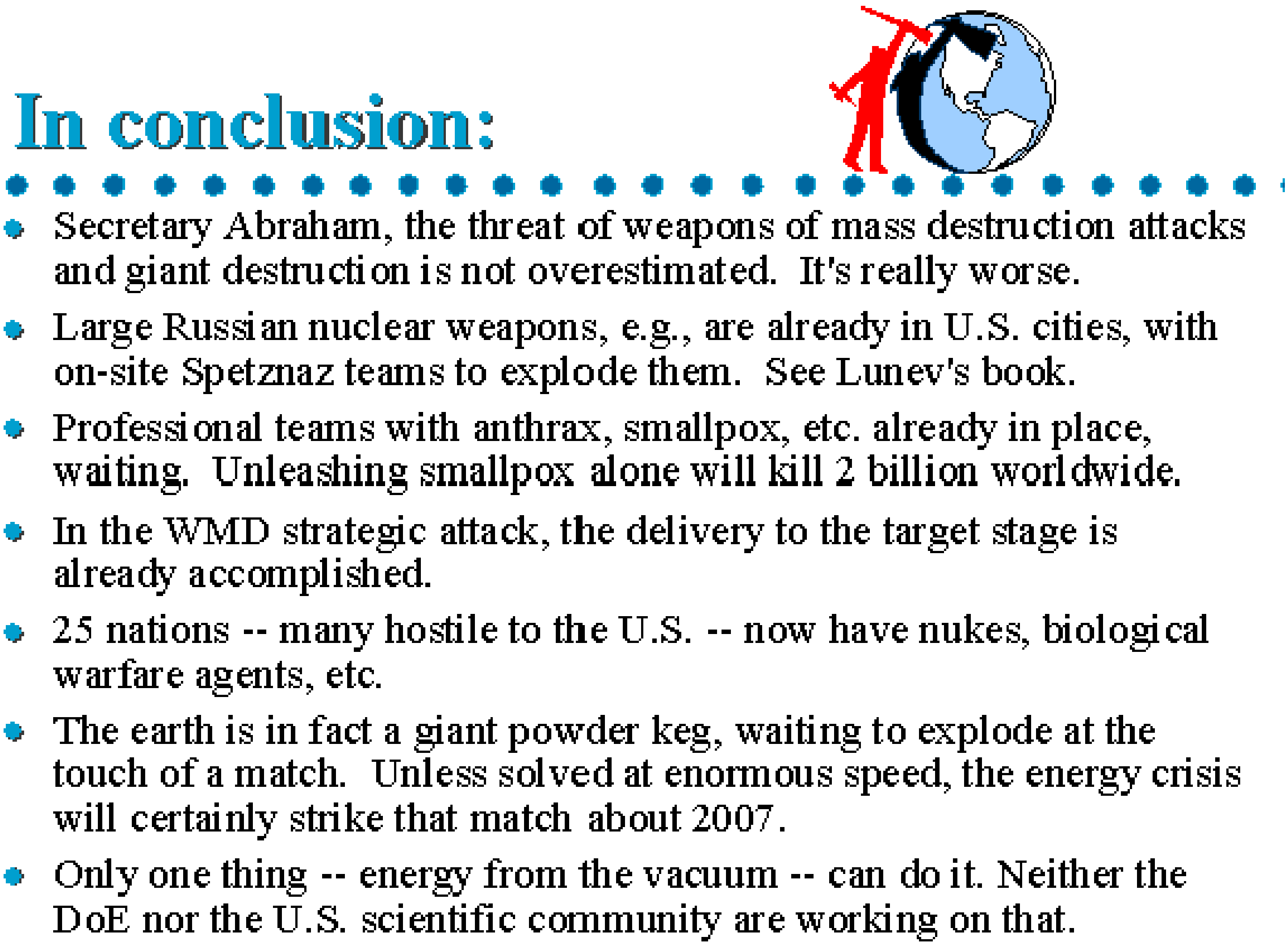

In conclusion:

- Secretary Abraham, the threat of weapons of mass destruction attacks and giant destruction is not overestimated. It's really worse.
- Large Russian nuclear weapons, e.g., are already in U.S. cities, with on-site Spetznaz teams to explode them. See Lunev's book.
- Professional teams with anthrax, smallpox, etc. already in place, waiting. Unleashing smallpox alone will kill 2 billion worldwide.
- In the WMD strategic attack, the delivery to the target stage is already accomplished.
- 25 nations -- many hostile to the U.S. -- now have nukes, biological warfare agents, etc.
- The earth is in fact a giant powder keg, waiting to explode at the touch of a match. Unless solved at enormous speed, the energy crisis will certainly strike that match about 2007.
- Only one thing -- energy from the vacuum -- can do it. Neither the DoE nor the U.S. scientific community are working on that.

Published by Lotus® Freelance Graphics®

The "Porthole" Concept with technical background

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

Amplifying the cellular regenerative (healing) system of the body for treatment and healing
of mass casualties

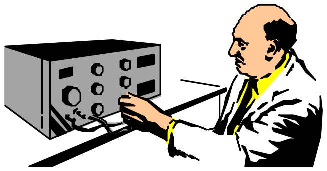

Outline of the presentation

- Mechanism for deep EM exchange of a body with its active EM environment
- Becker's epochal work and its ramifications
- Kaznacheyev's startling results and ramifications
- Weaponization of Kaznacheyev's findings
- Importance of time-energy and the time-domain
- Mechanisms for medical treatment and healing

Outline of the presentation (2)

- Prioré's revolutionary results and their ramifications.
- Proposed new mechanism for cancer promotion.
- Porthole concept: The 2-minute shortcut.
- Terrorist biological warfare capabilities.
- Summary and conclusion.
- Budgetary and Program package.

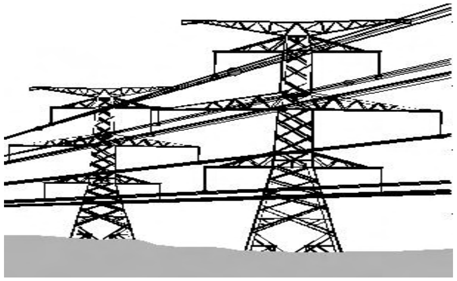

Mechanism for the body's deep EM exchange with its active EM environment

And some ramifications

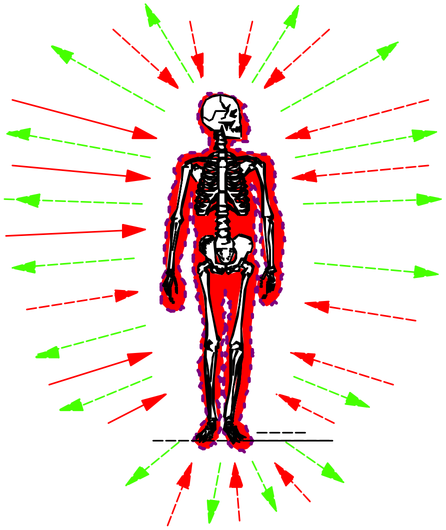

The body continually receives and retransmits EM energy

- The body may be considered as a dielectric.
- Every part of the dielectric participates in the absorption and emission of each photon to and from any point on the surface.*
- Mechanical force effects are also electrical in nature and exchange similarly.

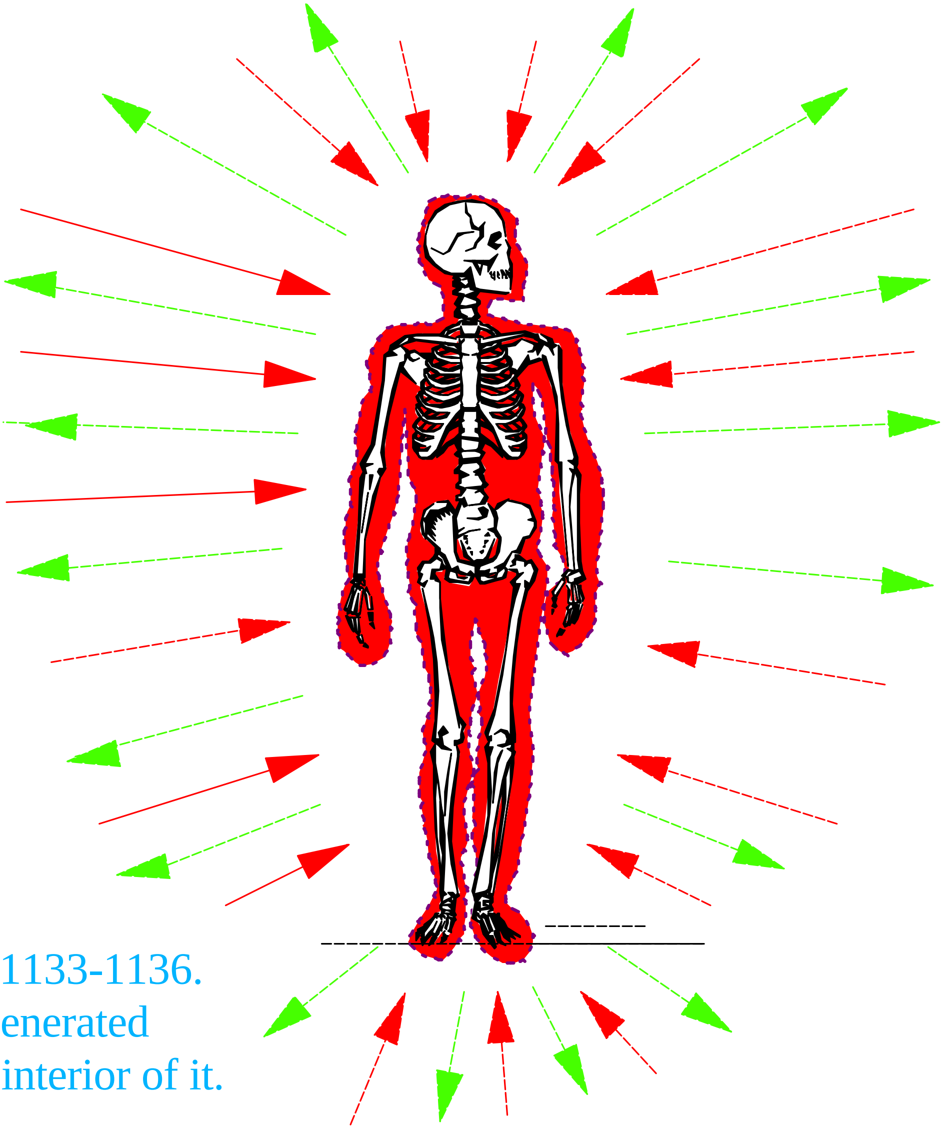

* Reali, G. C., "Reflection from dielectric materials," American Journal of Physics, 50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.

Ramifications:

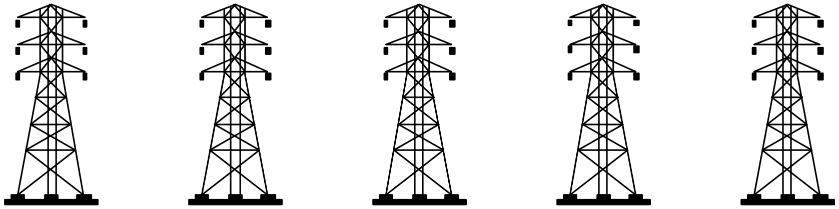

- There is an EM "dielectric transmission path" connecting every part of every cell to every part of the surface of the body.
- All electrical processes in the body, including deep within the cell and its nucleus, participate in the absorption and emission of every photon to and from the body's outer surface (and every mechanical force* experience).
- The EM emission energy of the body is a conglomerate mix of fractions of all processes and functions ongoing in the body or affecting it, including in every cell.
- This includes all electromagnetic, mechanical, chemical, and environmental effects upon and in the body and every tiniest part of it, and provides an extended definition of "stress".

* At base, all normal mechanical forces are generated by electromagnetic forces.

Ramifications (2)

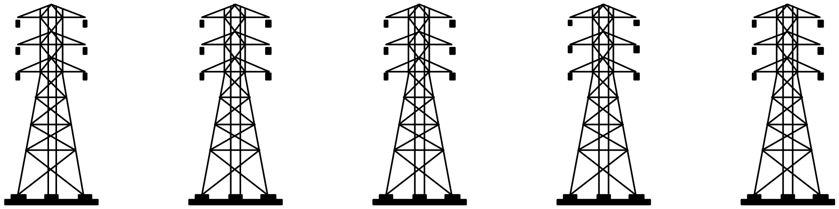

- The reaction of each photon absorbed on any point of the outer surface of the body, connects through the dielectric transmission path to every part of the body, including to every part of every cell*.
- Every emission from the body is participated in by every part of the body, including every part of every cell.
- The body's emission "changes" its environment, adding EM radiation. Some of the internal induced changes (excited states) in the body decay by this external emission. Some do not, but must be otherwise dealt with internally by the body systems.
- The body's EM absorption allows the environment to change or affect every function, operation, and part of the body, by adding EM radiation. All are continually being affected and deviated, both helpfully and harmfully, by these deeply induced EM interactions.
- The body's reaction (healing) must continually readjust and correct deviations from normalcy, in every part of every cell.

* In quantum physics, part of the photon is localized and part is nonlocalized, reaching even across the universe.

Ramifications (3)

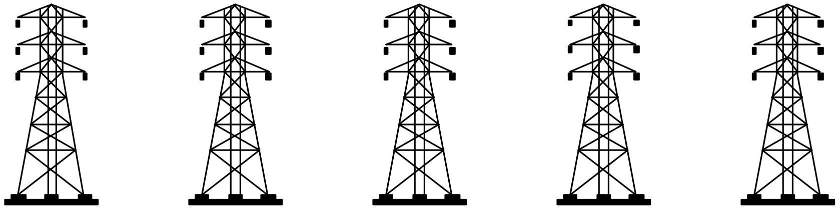

- The sum total of all the impinging external EM radiation upon the body, plus mechanical "special EM" radiation, thus affects and changes the body and all its processes, participating in them.
- That impinging EM radiation can be manipulated to either harm or heal the body and any or all of its cells.
- The body must continually overcome these externally-induced EM changes in its cells and processes, to maintain normalcy. Some EM changes are restored by the subsequent dielectric emission process, relieving the previously-induced deviant EM excitations in its internal processes. Some are not.
- The poorly studied *cellular regenerative system* performs the remainder of the restorative function, which is the fundamental "healing" function of the body.*

*The cellular regenerative system was mostly studied by Becker and a few others. See R. O.Becker, and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, No. 73, Nov.-Dec. 1970, p. 169-198.

Becker's epochal work

And its ramifications

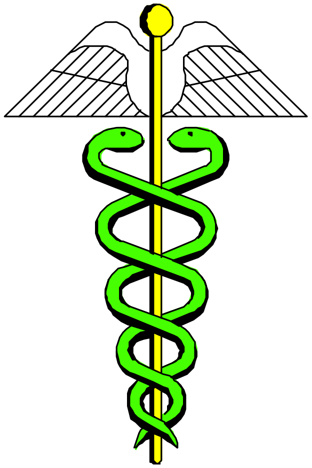

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

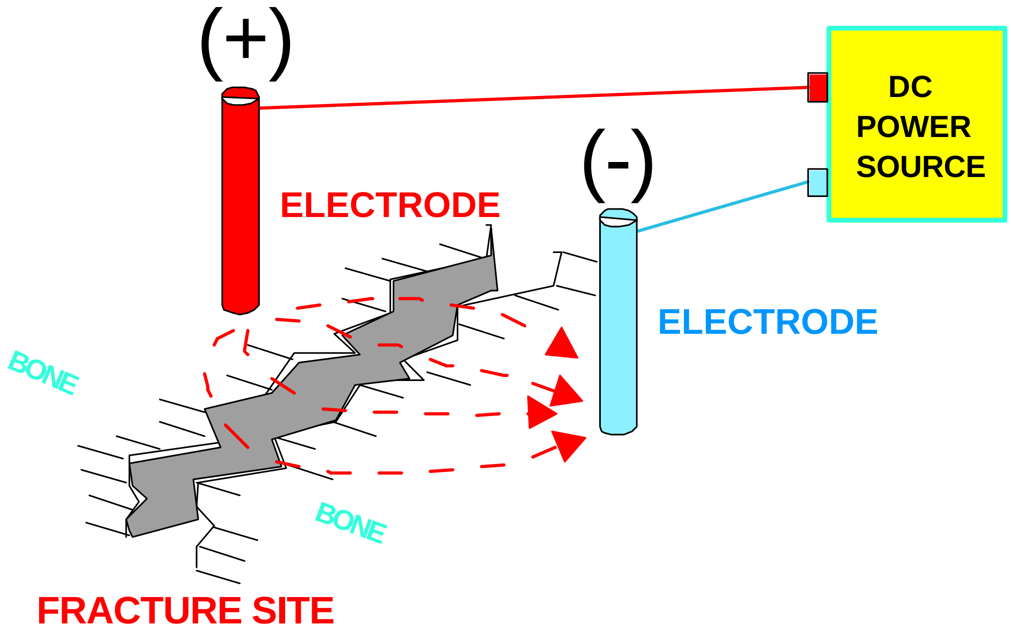
(+)
(-)
DC POWER SOURCE
ELECTRODE
ELECTRODE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Becker's theoretical DC control system involved with response to injury

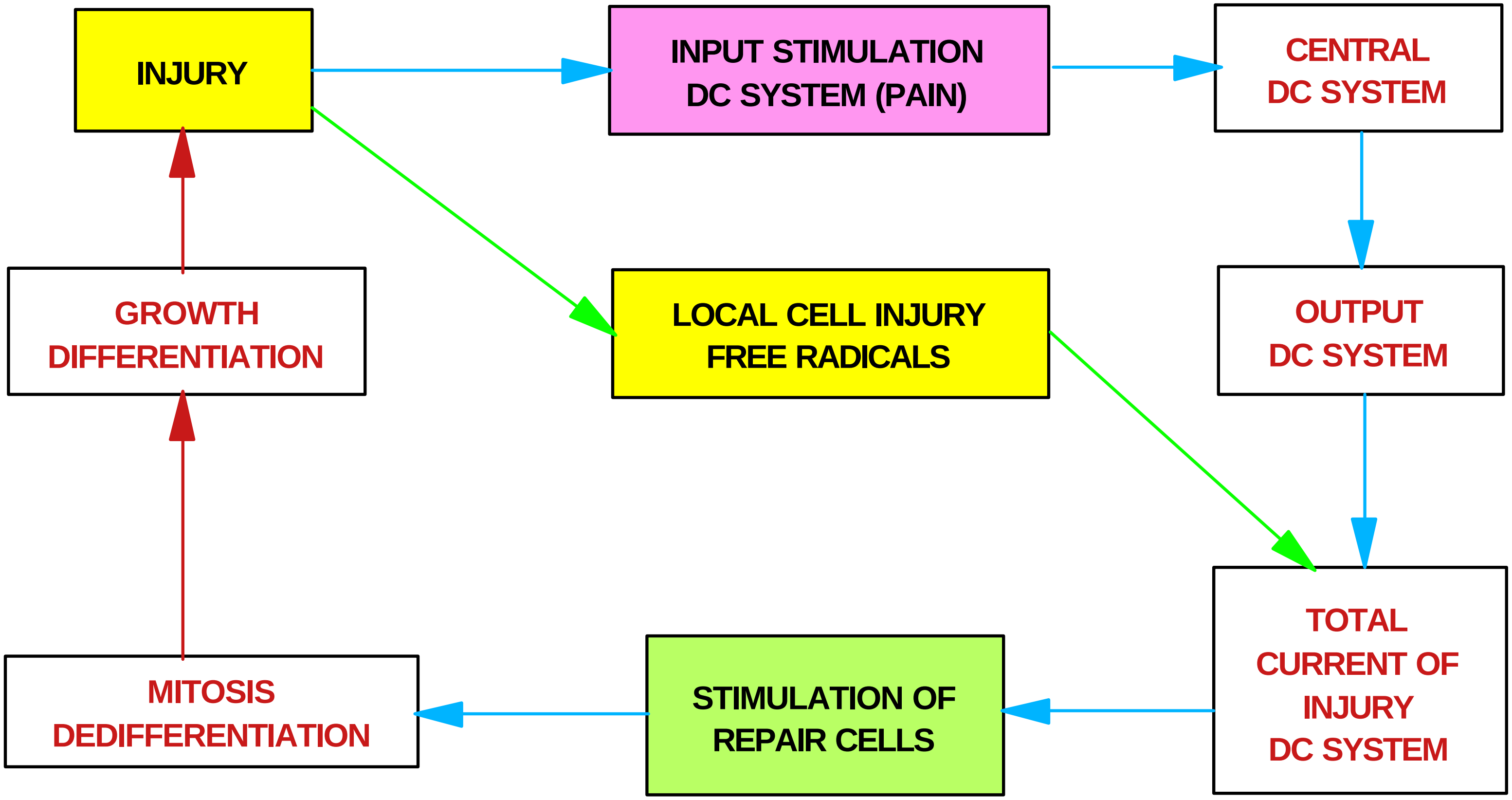

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Becker's proposed control system governing regeneration*

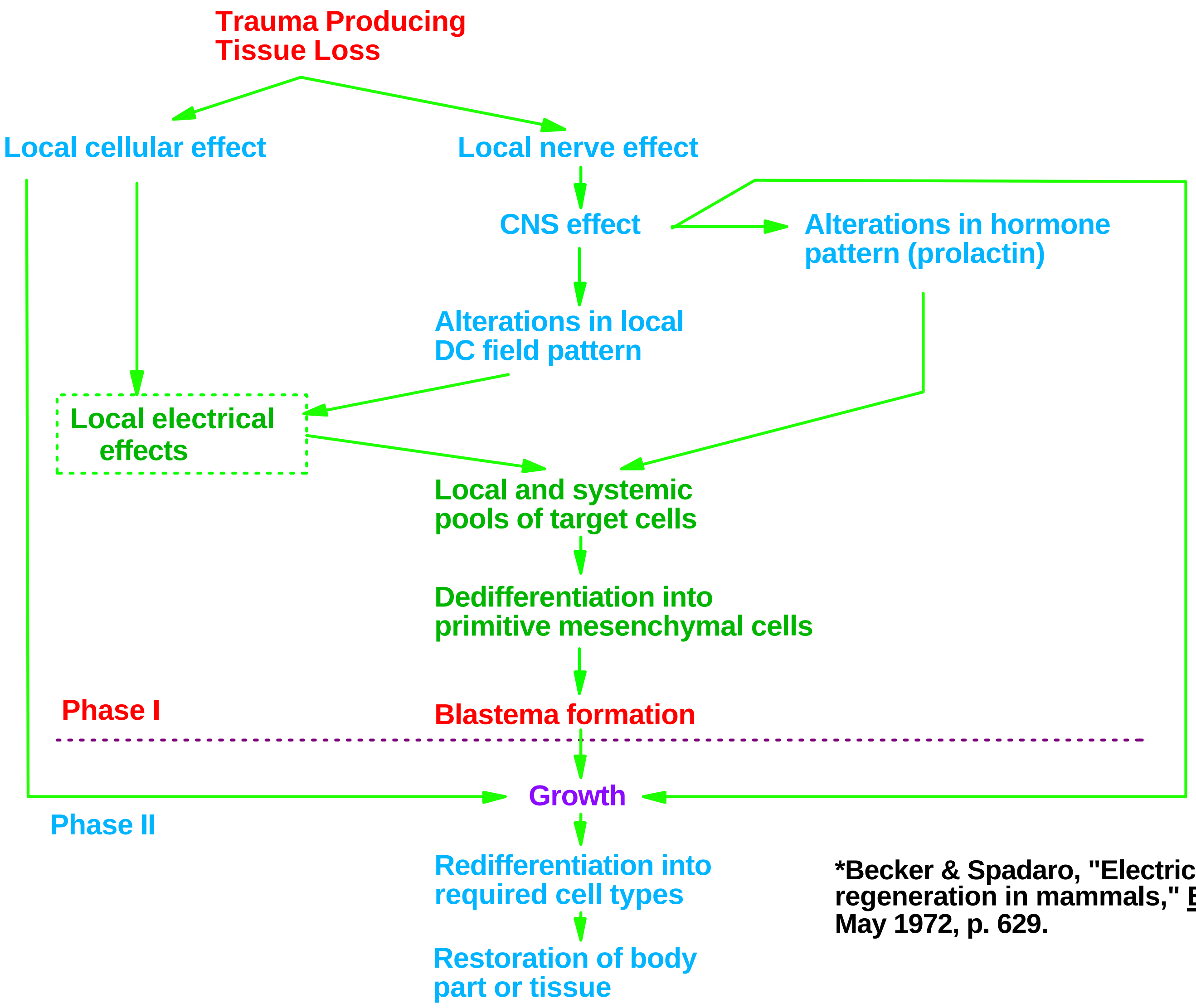

Trauma Producing
Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone
pattern (prolactin)

Alterations in local
DC field pattern

Local electrical
effects

Local and systemic
pools of target cells

Dedifferentiation into
primitive mesenchymal cells

Phase I

Blastema formation

Growth

Phase II

Redifferentiation into
required cell types

Restoration of body
part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Extending Becker's work

- Becker had only the archaic classical electrodynamics available to him. That EM is grossly inadequate.
  - Insufficient group symmetry to model the process.
  - Already eliminates the "infolded" longitudinal wave EM that structurally and dynamically comprises all "normal" EM potentials, fields, and waves.
  - It is this internal LW EM that is manipulated and used by the cellular regenerative system.
- Nonlinear phase conjugate optics was not yet developed when Becker did his seminal work.
- Successful higher symmetry electrodynamics models -- such as O(3) -- have been developed and are available.

A sad commentary

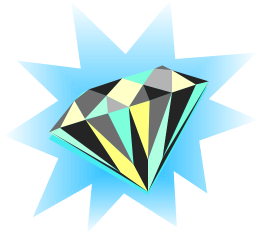

- Becker was nominated for the Nobel Prize, but he was hounded, his funds were withdrawn, and he was forced to retire early.
- His main "sin" was having the courage to testify against the harmful effects of unrestricted powerline radiation and electronic smog.
- Powerful interests then crushed him, set up their own well-funded institutes, and hired scientists who then predictably found that powerline radiation was no problem.
- Such research is so tainted that recently several important journals have now required the authors of submitted papers to state what their affiliated companies do, and what "company massaging" of their research occurred.
- From a higher symmetry EM view, it is obvious that the dielectric pathway and quantum electrodynamics already show a significant problem due to such radiation, particularly with respect to long-term effects.
- This is particularly significant since "noise" -- the mere presence of more "hash" -- appreciably amplifies the dielectric transmission effect. So the major factor becomes EM smog *signals density*, not signal power.

Kaznacheyev's startling results

And its ramifications

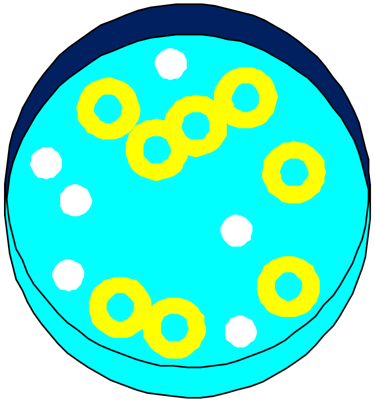

Proof that the cellular condition is radiated from groups of cells

- It follows that a diseased, damaged, or infected cell culture will and does emit its exact internal EM condition, hence the disease or damage "generatrix form" itself (in the internal structuring of the EM radiations).
- Further, these "state" emissions can couple into other targeted cells to produce disease and disorder in them.
- This was clearly proven in some 15,000 Russian experiments by Kaznacheyev.
- Kaznacheyev demonstrated induction of cellular disease and disorder between EM-coupled but otherwise environmentally shielded cell cultures
- Similar experiments have been replicated in the West by Reid *et al*., and by Popp *et al*.

Kaznacheyev's induction of cellular disease and disorder

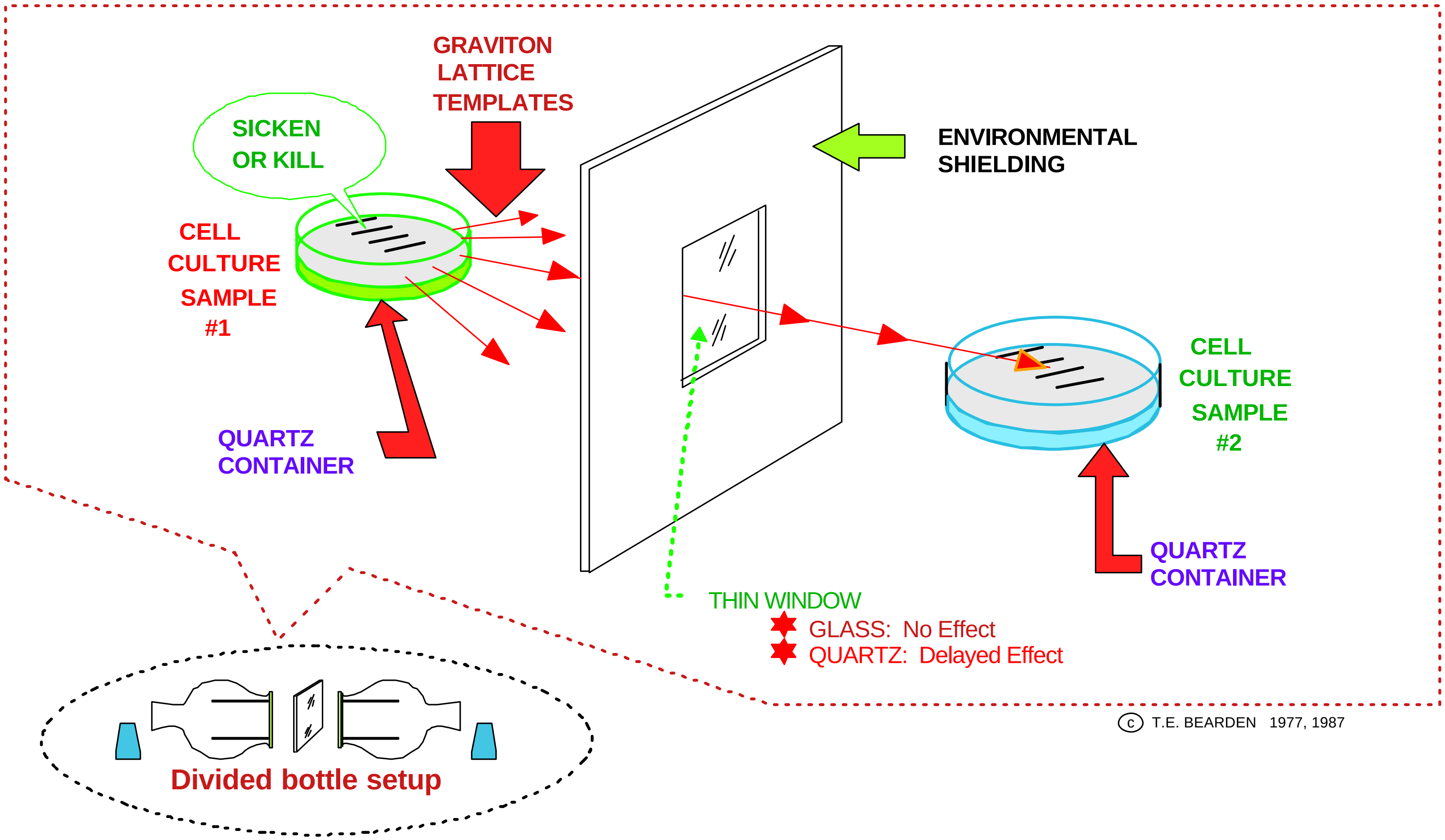

GRAVITON LATTICE TEMPLATES

ENVIRONMENTAL SHIELDING

SICKEN OR KILL

CELL CULTURE SAMPLE #1

QUARTZ CONTAINER

CELL CULTURE SAMPLE #2

QUARTZ CONTAINER

THIN WINDOW

- GLASS: No Effect
- QUARTZ: Delayed Effect

Divided bottle setup

ⓒ T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Another surprising feature of Kaznacheyev's experiments

- The experiments do not work if the targeted cells are in normal light, such as sunlight*.
- The reason is that the normal light bandwidth occupies a single harmonic interval between ultraviolet and infrared.
- Hence its difference frequencies also fill such an interval. For a set of n frequencies, the number *d* of difference frequencies is $d >> n$.
- Thus the visible light spectrum represents "jamming" of specific internal EM signal inputs through the dielectric pathway into the body, turning much of the environmental signals into harmless warmth.

* It has also been shown that if bacteria are killed in the "dark" by UV, then 24 hours later placed in sunlight, a substantial fraction of the "dead" bacteria will revive.

Ramifications of Kaznacheyev's experimental results

- At least one harmonic interval is necessary for the effect to be evidenced in the target cells.
- Interestingly, if we assume the body dielectric to be isotropically nonlinear (to first order):
  - Velocities of the actual transmitted frequencies depend upon the particular point on the wave amplitude.
  - The transmitted frequencies overshoot, interfere, breakup, etc. The difference frequency does not.
  - The "difference frequency" is transmitted through the nonlinear dielectric medium as if it were a sine wave passing through a linear medium*.
- To our knowledge, this direct and deep communication of electronic "hash" difference frequencies in dielectric human bodies has not been investigated in the West.

---

* Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

Ramifications of Kaznacheyev's experimental results (2)

- A heterodyne signal -- at the difference frequency between an input signal and a reference signal -- can be enhanced by adding noise*.
- Thus the signal of the disorder in the cellular pattern emitted from diseased cells, received by normal cells, can be amplified electronically and rather easily.
- Powerline radiation (electronic smog) thus carries its own amplification mechanism. Eventually the amplification overcomes any squelching by sunlight, etc.
- The Russians promptly (as early as the late 1950s) weaponized and tested these and similar effects.

* Dykman, M. I. et al., "Noise-enhanced heterodyning in bistable systems," Physical Review E, 49(3), Mar. 1994, p. 1935-1942

Weaponization of Kaznacheyev's findings

And its ramifications

"Microwave" radiation of the U.S. Embassy in Moscow.

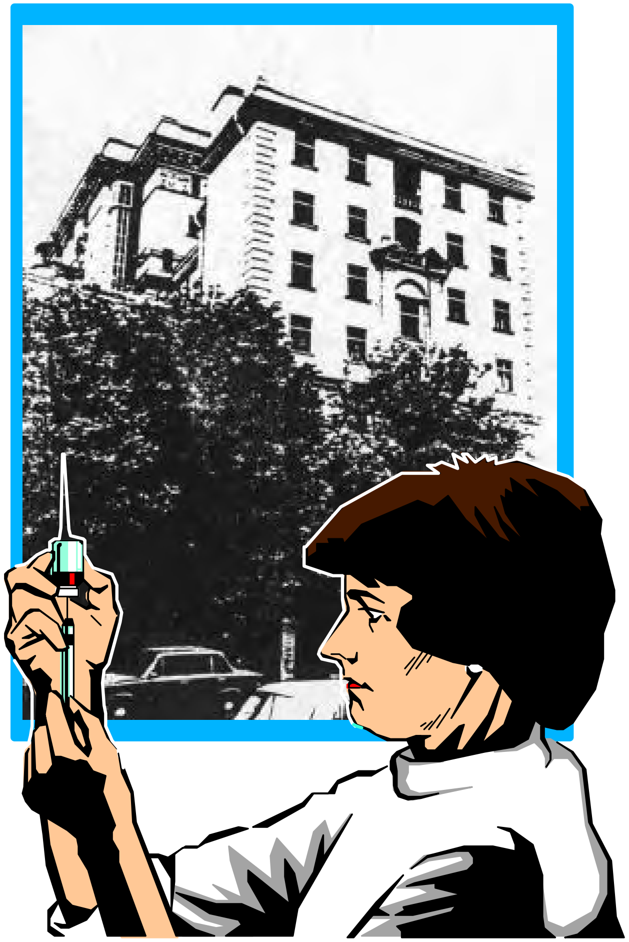
Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Ramifications of the microwave radiation of the U.S. Embassy

- The cellular disorder pattern in the EM radiation is carried by the field-free scalar potential.
- Since the disorder pattern is a "conglomerate", it must be carried by an internal EM structure existing inside the scalar potential itself.
- Whittaker decomposed the scalar potential into the appropriate internal longitudinal EM biwave structure in 1903*.

* E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355

The importance of time as energy, and the time-domain

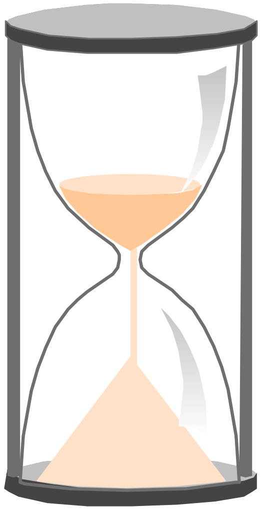

Time as energy

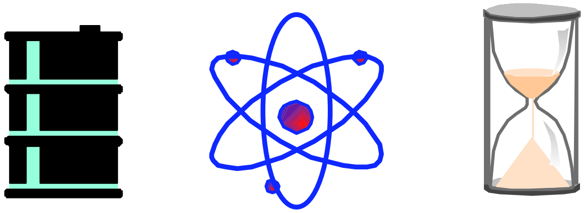

- In physics, the fundamental units one uses in his model are arbitrary.
  - A valid model can be made using only one unit.
  - It is already done in one branch of physics, using length as the single fundamental unit.
- Make a model with a single unit, the joule.
  - All other entities become functions of the joule.
  - We are familiar with mass as highly compressed EM energy, by the factor c^2.
  - It turns out that time is also highly compressed EM energy, by the factor c^2.
  - 1 sec. $\approx 9 \times 10^{16}$ joules.

How to see that "time is compressed spatial energy"

- Take some spatial EM energy and compress it by c^2
- Place it in 3-space, and it is *mass*.
- Place it on the time axis, and it is *time*.

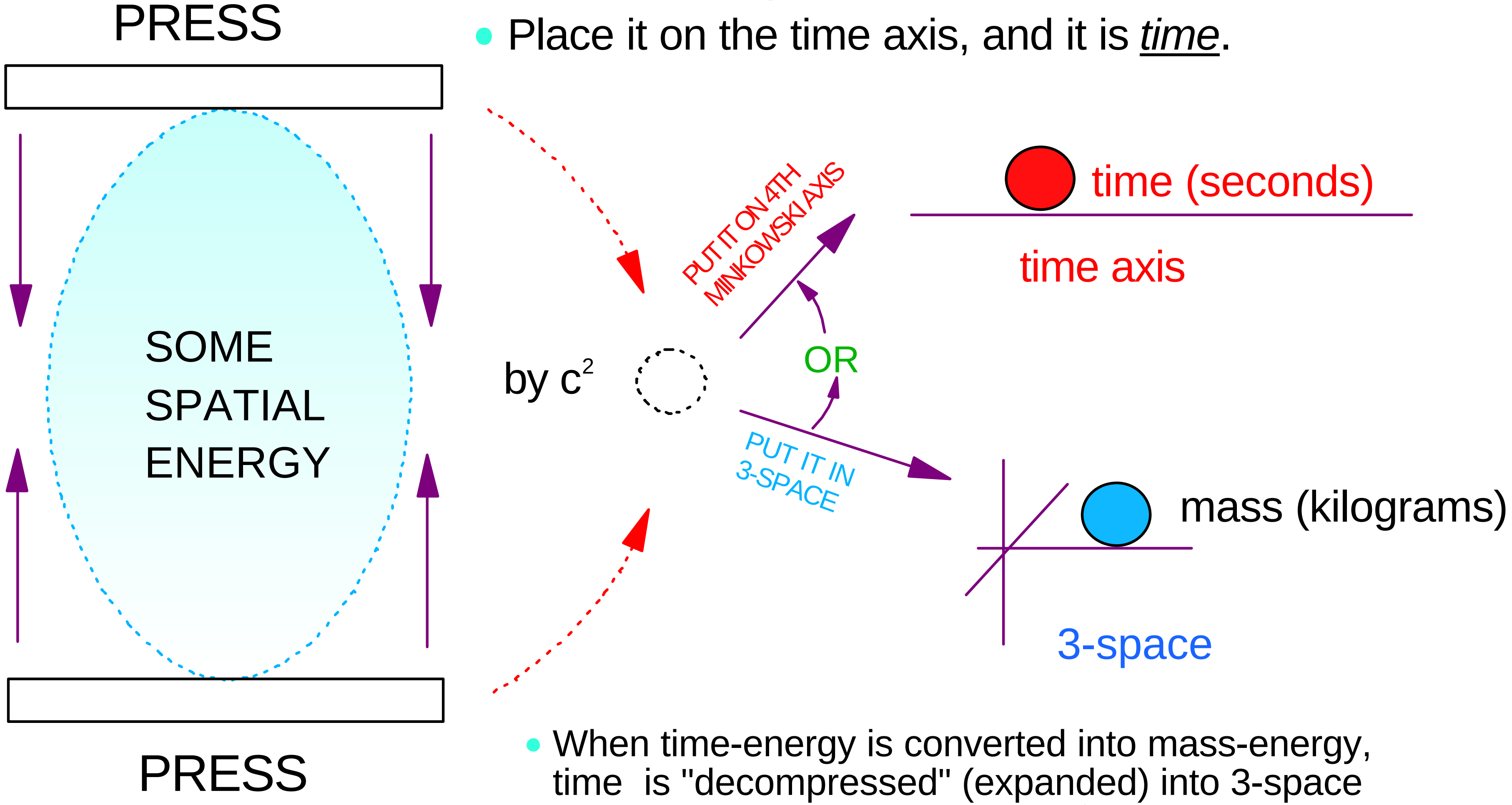
PRESS
SOME
SPATIAL
ENERGY
PRESS
by c^2
PUT IT ON 4TH MINKOWSKI AXIS
OR
PUT IT IN 3-SPACE
time (seconds)
time axis
mass (kilograms)
3-space

- When time-energy is converted into mass-energy, time is "decompressed" (expanded) into 3-space EM energy and 1 sec. $\approx 9 \times 10^{16}$ joules.

Time as EM energy in the time domain, with a density the same as mass-energy

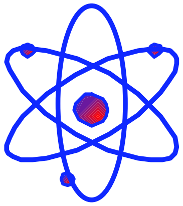

- Bearden* reinterpreted (corrected) the 1903 Whittaker decomposition of the potential in 2000.
- The reinterpretation is consistent with broken symmetry in particle physics, for which Lee and Yang received the Nobel Prize in 1957.
- It is also powerfully supported by quantum field theory, notably by Mandl and Shaw in 1984**.
- Whittaker's pair is a combined time-polarized (scalar) EM wave and a 3-space longitudinal EM wave.

* T. E. "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23.

** F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

All EM energy in 3-space freely comes from the time domain

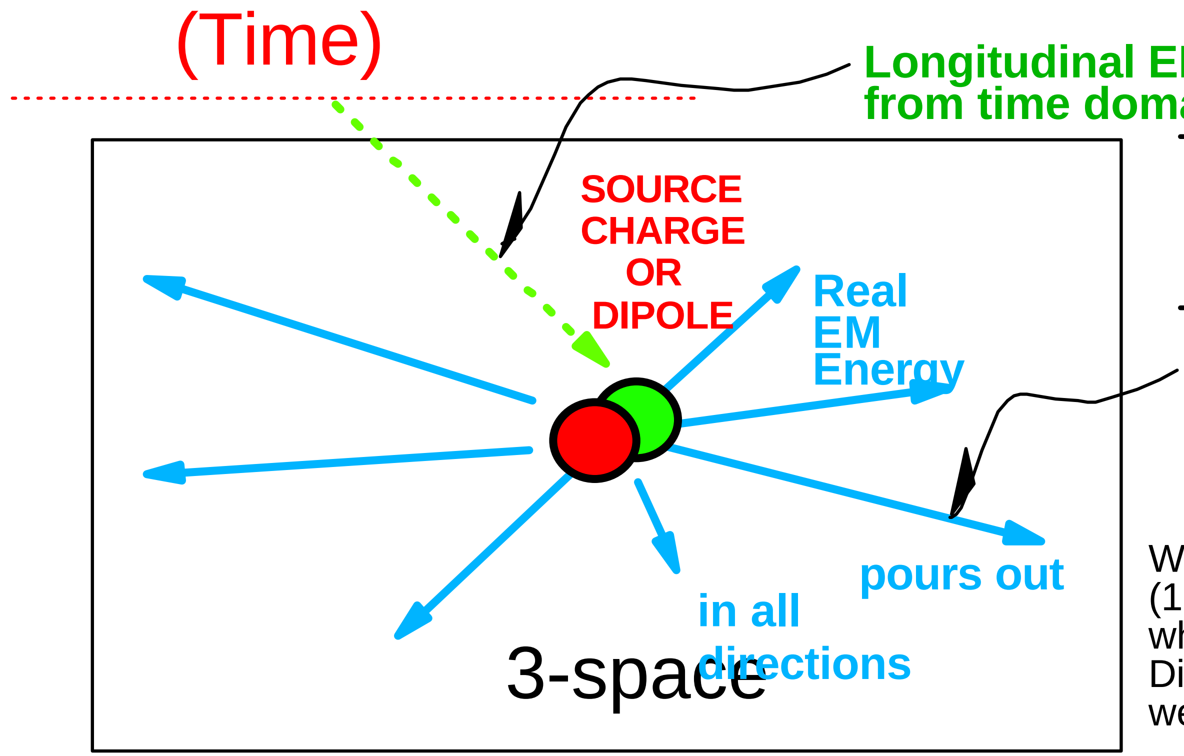

(Time)

SOURCE
CHARGE
OR
DIPOLE

Real
EM
Energy

pours out

in all
directions

3-space

Longitudinal EM wave incoming
from time domain and absorbed

Charge's 720° spin
transforms time-energy
into 3-space energy

Longitudinal EM
waves emitted in
all directions in
3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur was just repeated since then, until corrected by Bearden to be consistent with broken symmetry and quantum field theory.

Strong support from quantum field theory

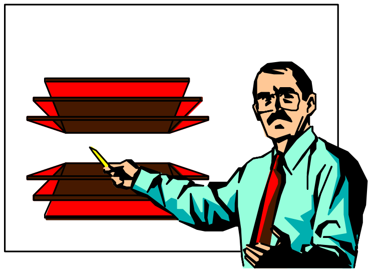

- Quantum field theory recognizes four photon polarizations:
  - Energy along x-, y-, z-, or t- axis.
  - Polarized along x- and/or y-axis, the photon is a *transverse* photon.
  - Polarized along z-axis (along its line of motion in 3-space), the photon is a *longitudinal* photon.
  - Polarized along the time-axis, the photon is a *time-polarized* or *scalar* photon.
- The individual longitudinal or scalar photon is not observable*.
- The combination of a longitudinal photon and a scalar photon is observable as the instantaneous scalar potential*.

* F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5.

Mechanism generating the flow of a mass through time

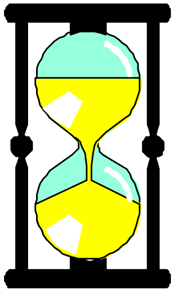

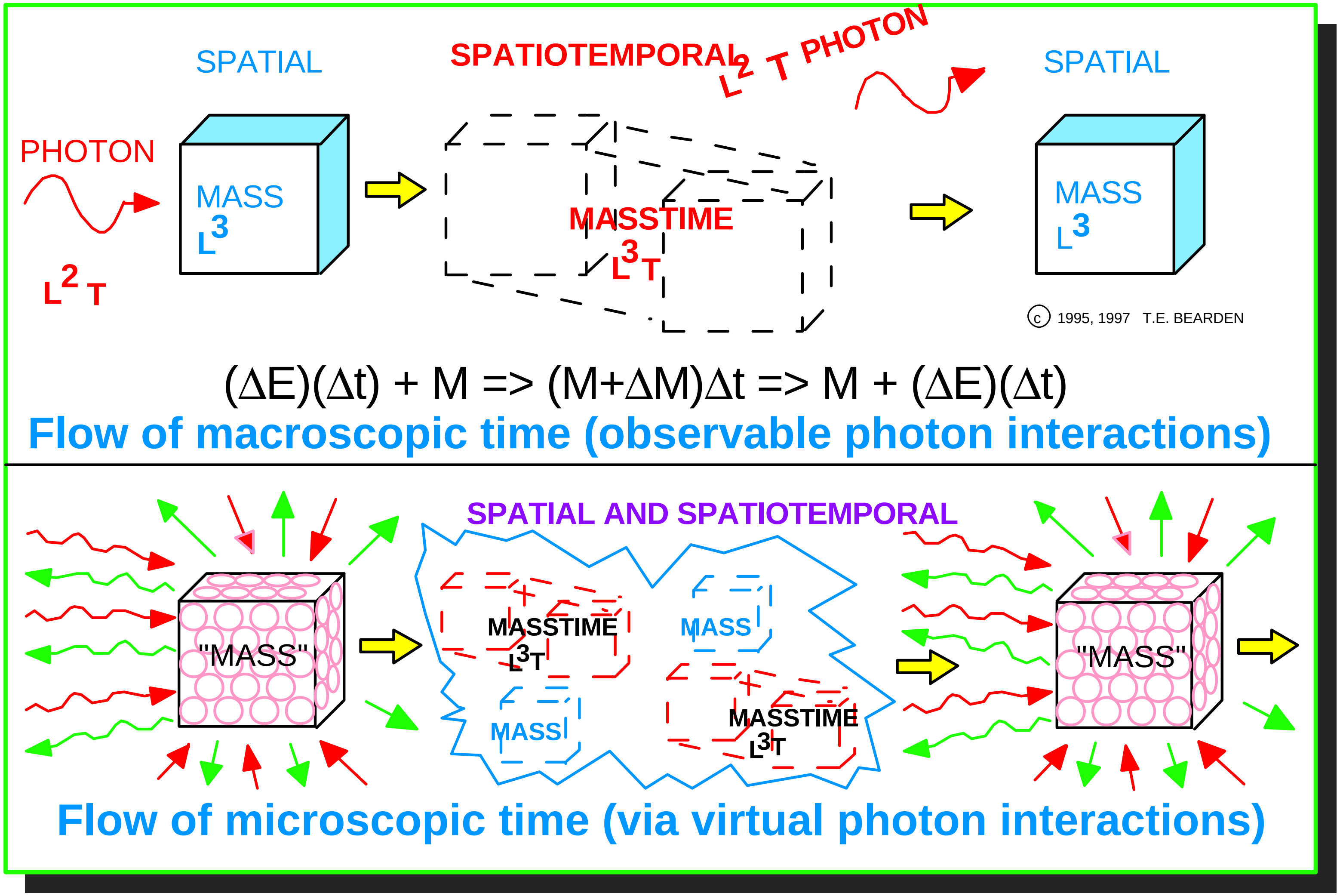
SPATIAL
SPATIOTEMPORAL
L^2 T PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2 T
MASSTIME
L^3T
MASS
L^3
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

ⓒ 1995, 1997 T.E. BEARDEN

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

Flow of microscopic time (via virtual photon interactions)

Mechanism is engineerable

Pumping a mass in the time-domain will time-reverse it back to an earlier state

- The reinterpreted biwave EM structure of the scalar potential means that time-reversal (phase conjugation by pumping in the time-domain) can be utilized to time-reverse a targeted cell (its mass-energy) in its entirety, including its genetics.
- Becker's bone fracture experiments showed it:
  - Weak scalar potential across the fracture site.
  - LWs in 3-space and time-energy waves in time
  - Red blood cells entered, dedifferentiated by shucking their hemoglobin and growing a nucleus, which is dedifferentiation (time-reversal of the cell and its mass-energy back to a previous physical state).

Principle for time-reversing a time-pumped mass

- Becker's cells then redifferentiated and changed -- first to primitive cartilage cells, then to primitive bone cells which were deposited in the fracture, healing it with new bone growth.
- This shows "time-forwarding" (redifferentiation) of the cell also.
- Becker actually showed that the *time-domain pumping* of the cells (by the time-polarized EM waves of Whittaker's reinterpreted wavepairs comprising the scalar potential), and the time-pumping of their mass-energy, operated to continuously and steadily *remove the delta between that cell's present condition and its former "normal" cellular physical condition in that area*.
- This is the long-sought secret of stem-cell research.
- Phase conjugate optics had not been born at the time. Without realizing its full nature, Becker demonstrated the body's own internal mechanism for healing, and the fundamental mechanism for time-reversing a physical mass, whether living or inert.

Mechanisms for use in medical treatment and healing

And ramifications

Successive deep penetration reactions of the cellular dielectric

- Change after change affects the cell and all of its parts, including from the external environment.
- Each successive change affects primarily the internal "LW EM structures and dynamics" comprising the ordinary potentials, fields, and waves.
- Thus in positive time the internal EM structurings produce successive "overlaid layers of change" -- so to speak -- in the internal EM structuring and dynamics of a cell in the body. These layers contain the "delta" from normal state.

Successive deep penetration reactions of the cellular dielectric (2)

- During time reversal, highest layer deltas are successively "peeled away" by the interaction of the mass-energy with the time-domain pumping.
- The result is to return the cell directly back to successive physical states, in the case of a single disease or disorder.
  - A cancer cell is returned to a normal cell prior to its "promotion".
  - An HIV-infected cell would be returned to its normal unaffected condition, free of HIV-induced genetic change.
  - For injuries and damage, the existing "delta" in each successive state-layer is steadily reduced to zero.
- This process thus can involve both cellular differentiation and dedifferentiation, as shown by Becker and by Prioré.

Engines (vacuum engines, spacetime curvature engines)

- Following Whittaker* as reinterpreted**, all normal EM fields, potentials, and waves decompose into sets of longitudinal EM waves (LWs) and their dynamics.
- Since these LWs and their dynamics are dynamics in the local energy density of spacetime (ST), they constitute *ST curvatures*.
- Hence they are general relativity infolded inside electrodynamics.
- By manipulating and changing these LWs, one manipulates and changes spacetime curvatures and their dynamics.
- The result is a unified field model where one directly controls and engineers a set of ST curvatures by use of higher group symmetry electrodynamics.
- Controlled patterns of local curvatures of ST are called "vacuum engines" or "spacetime curvature engines" -- or just "engines".

* E.T. Whittaker, Math. Ann., 57, 333-355 (1903); Proc. Lond. Math. Soc., Series 2, 1, 1367-372 (1904).

** T. E. Bearden, J. New Energy, 5(1), 11-23 (2000).

Pumping with EM time-polarized waves forms *amplified* antiengines

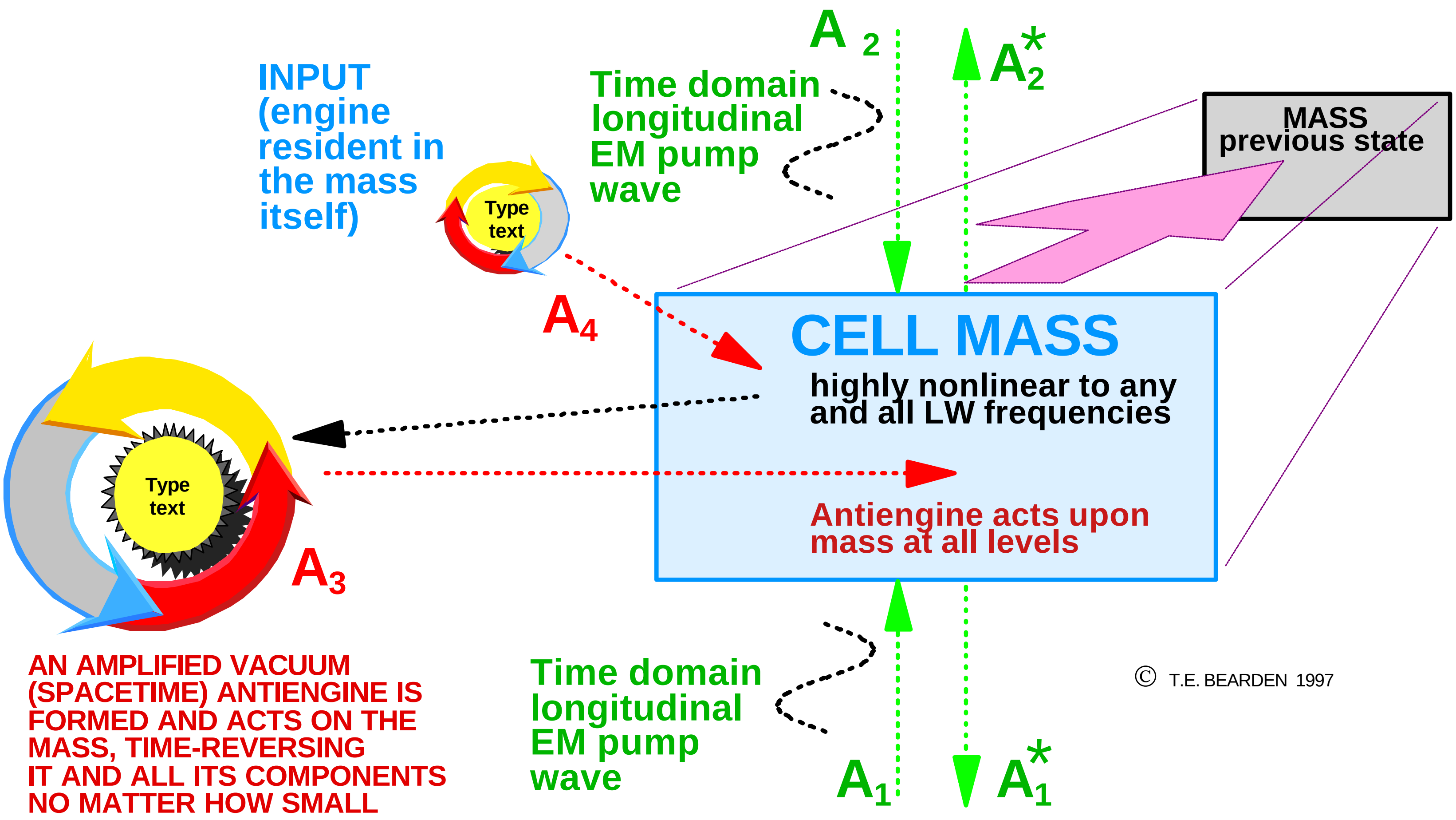

INPUT (engine resident in the mass itself)

Time domain longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

Time domain longitudinal EM pump wave

© T.E. BEARDEN 1997

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

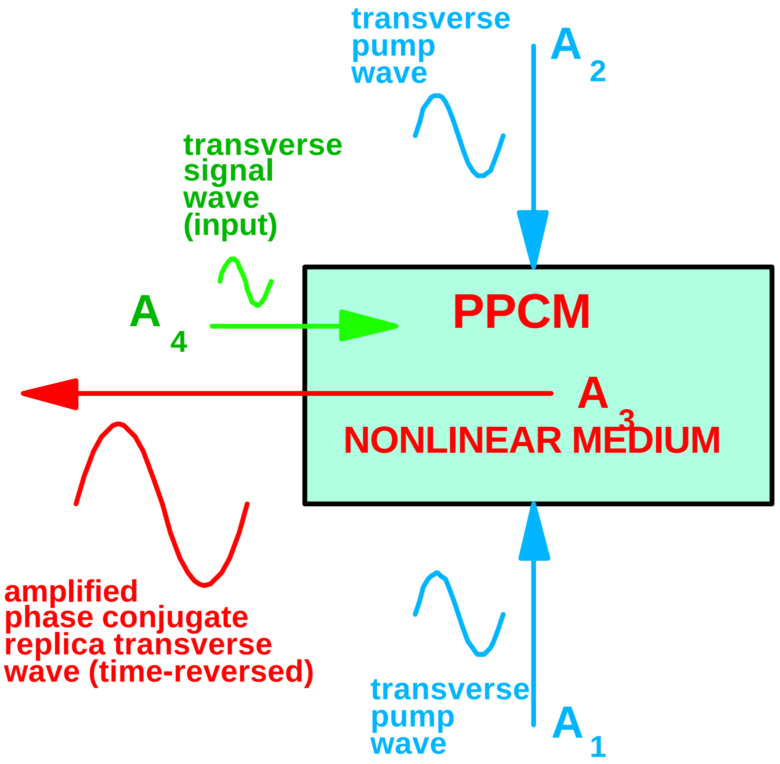
transverse pump wave
A 2
transverse signal wave (input)
A 4
PPCM
A 3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A 1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

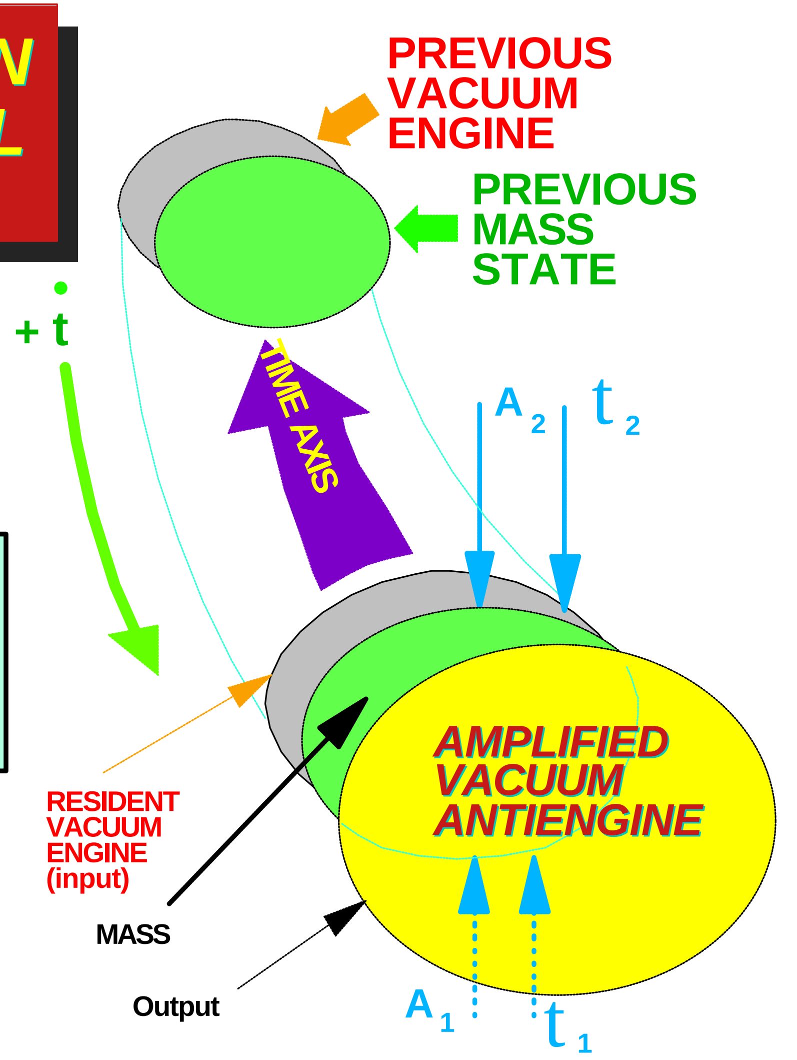
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
+ t
TIME AXIS
A 2
t 2
RESIDENT VACUUM ENGINE (input)
MASS
Output
AMPLIFIED VACUUM ANTIENGINE
A 1
t 1

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Time-reversal of a mass to a new physical state not in its past

- Original engine A'
- New engine B' added
- Time-pumping moves to new "past" that is determined by (A'+B')

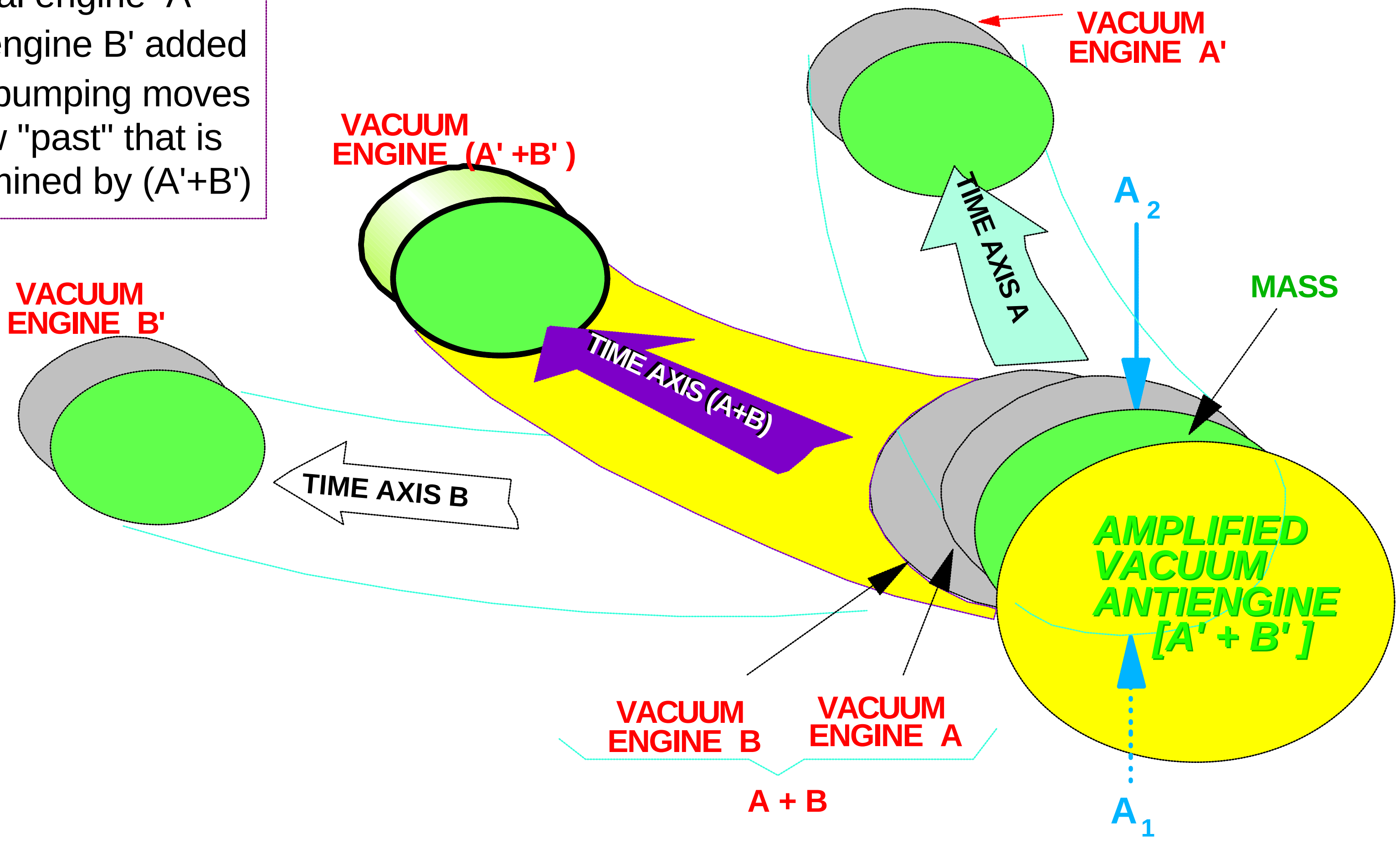
VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
TIME AXIS A
A_2
MASS
VACUUM ENGINE B'
TIME AXIS (A+B)
TIME AXIS B
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

Extended general relativity principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon any exposed mass at al levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

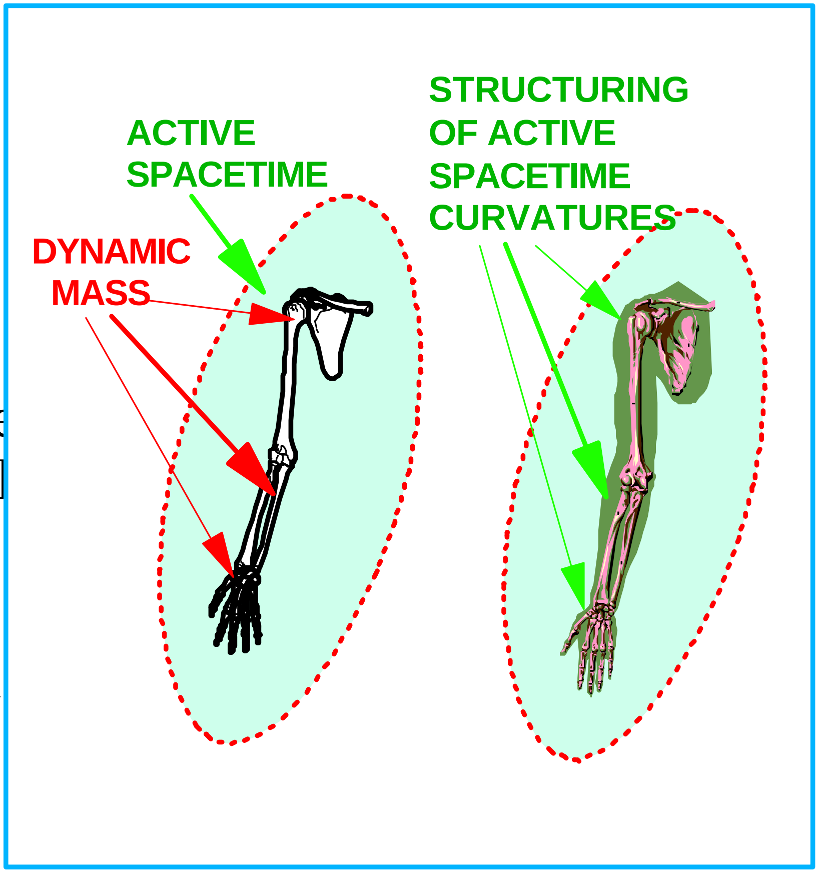
STRUCTURING
OF ACTIVE
SPACETIME
CURVATURES
ACTIVE
SPACETIME
DYNAMIC
MASS

Engines: Working spacetime demons

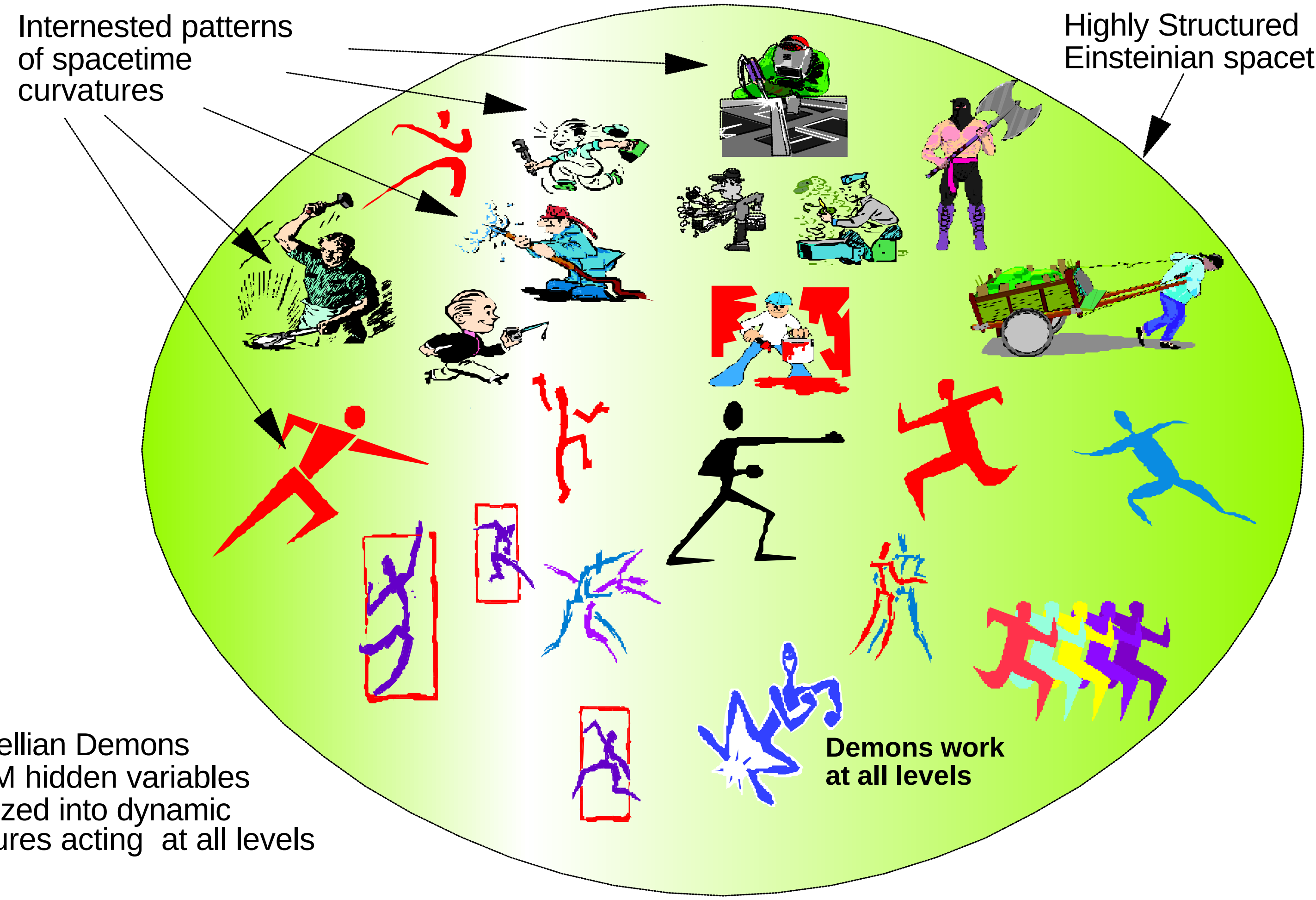

Internested patterns of spacetime curvatures

Highly Structured Einsteinian spacetime

Maxwellian Demons are EM hidden variables organized into dynamic structures acting at all levels

Demons work at all levels

Signal vs. vacuum engine

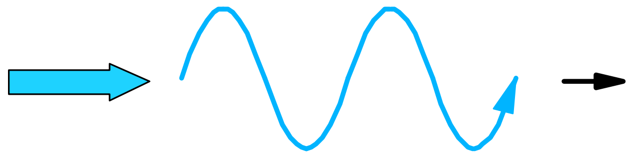

SIGNAL (INFO)

- ★ Information received
- ★ Any overt physical action must be taken by cell itself
- ★ Cell must furnish the energy for any action it takes
- ★ Passes from outside to in

A. *Cell must do the action itself; vacuum energy exchange is passive. Energy or fuel for doing the action must be added to cell externally.*

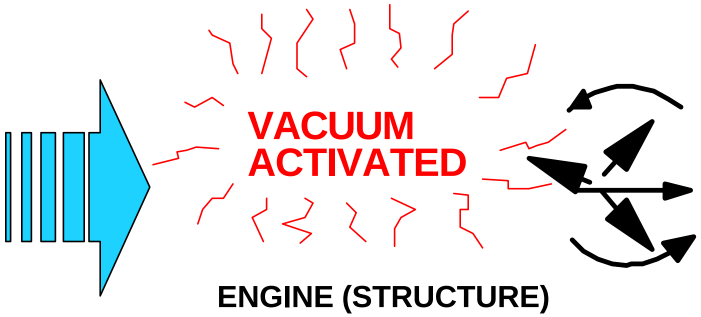

ENGINE (STRUCTURE)
TEMPLATE (FORM)

- ★ Information itself acts
- ★ Local spacetime curved, with internal subcurvatures
- ★ Produces template of forces in local spacetime (ST)
- ★ Arises from within local ST
- ★ Acts on and changes all parts of cell, including its genetics

B. *Cell is acted upon; vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to cell externally.*

Unified field theory requires modeling and use of supersystem

Three components interact with each other:

- System and its dynamics
- Local active vacuum and its dynamics
- Local curvatures of spacetime and their dynamics.

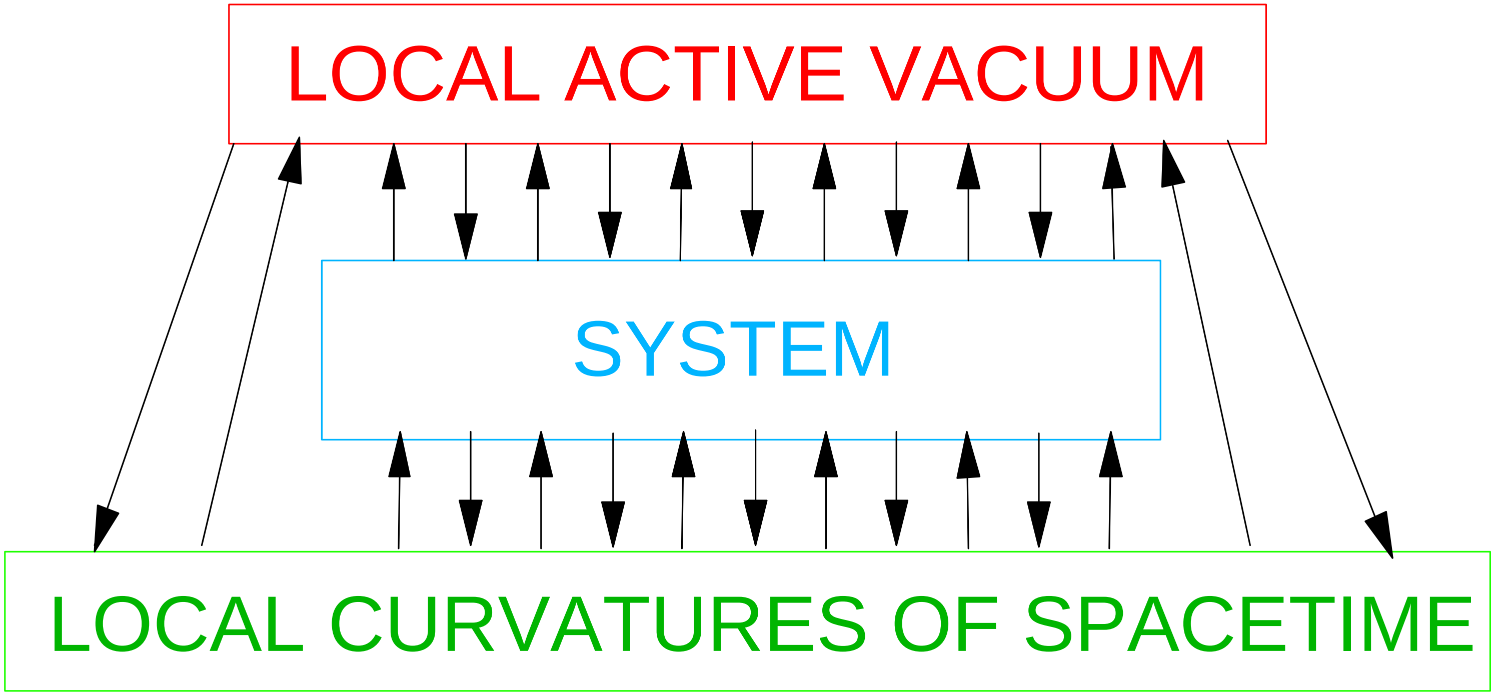
LOCAL ACTIVE VACUUM

SYSTEM

LOCAL CURVATURES OF SPACETIME

Classical EM kills the supersystem, prevents supersystem engineering

- Assumes flat local spacetime
- Assumes inert vacuum

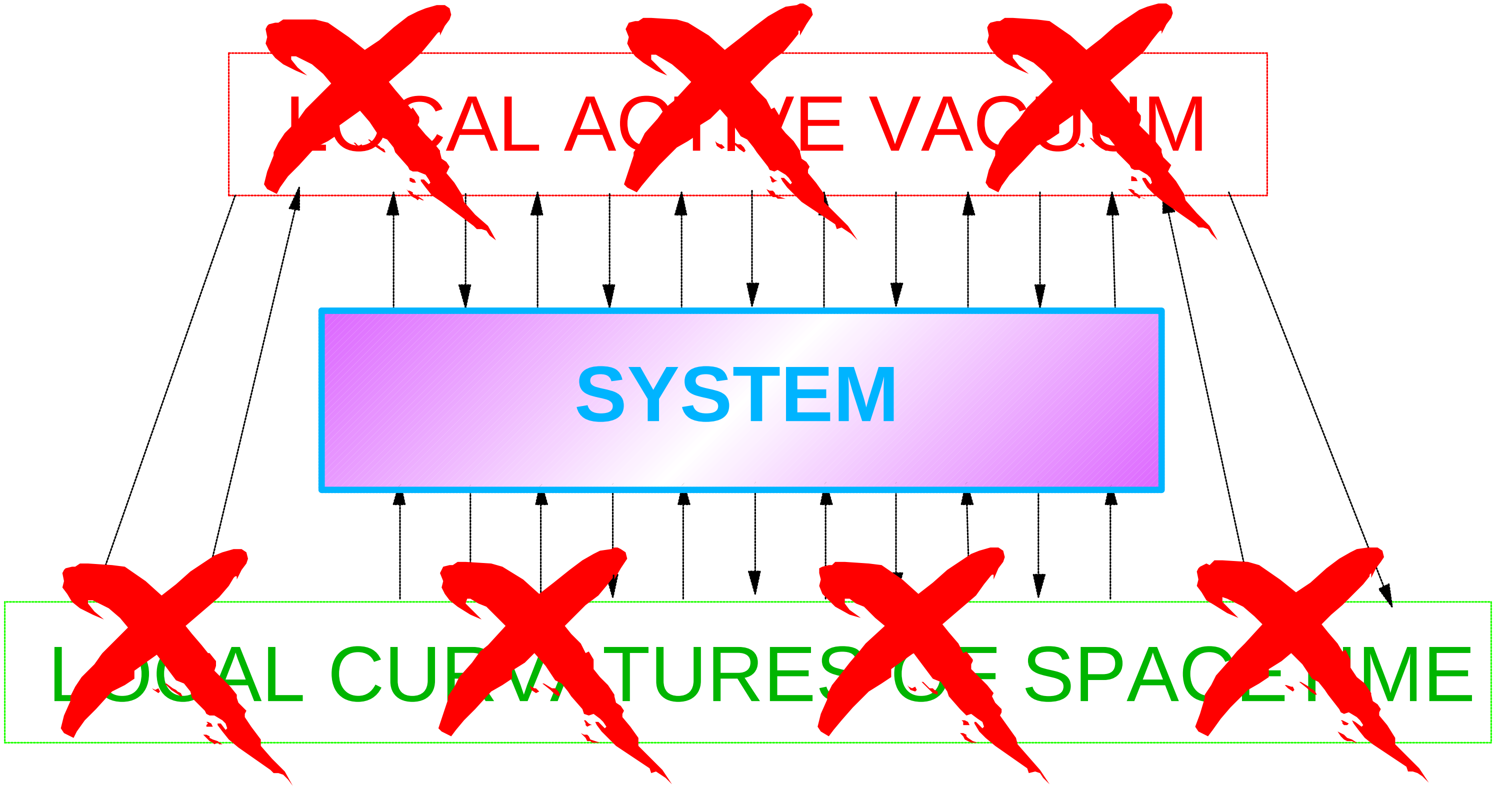
LOCAL ACTIVE VACUUM
SYSTEM
LOCAL CURVATURES OF SPACETIME

Malignant teratoma exhibits both time-reversal and fast-forwarding

- A teratoma cell is a tumor cell that has reverted to stem cells, then proceeds to differentiate toward new kinds of normal cells.
- May produce hair, skin, muscle, heart, bone, and other type cells.
- Sometimes present at birth.
- Demonstrates both "time-reversal" and "time-forwarding".
- The missing key is the engine function(s).

Miller-Fox-Urey biogenesis experiments

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

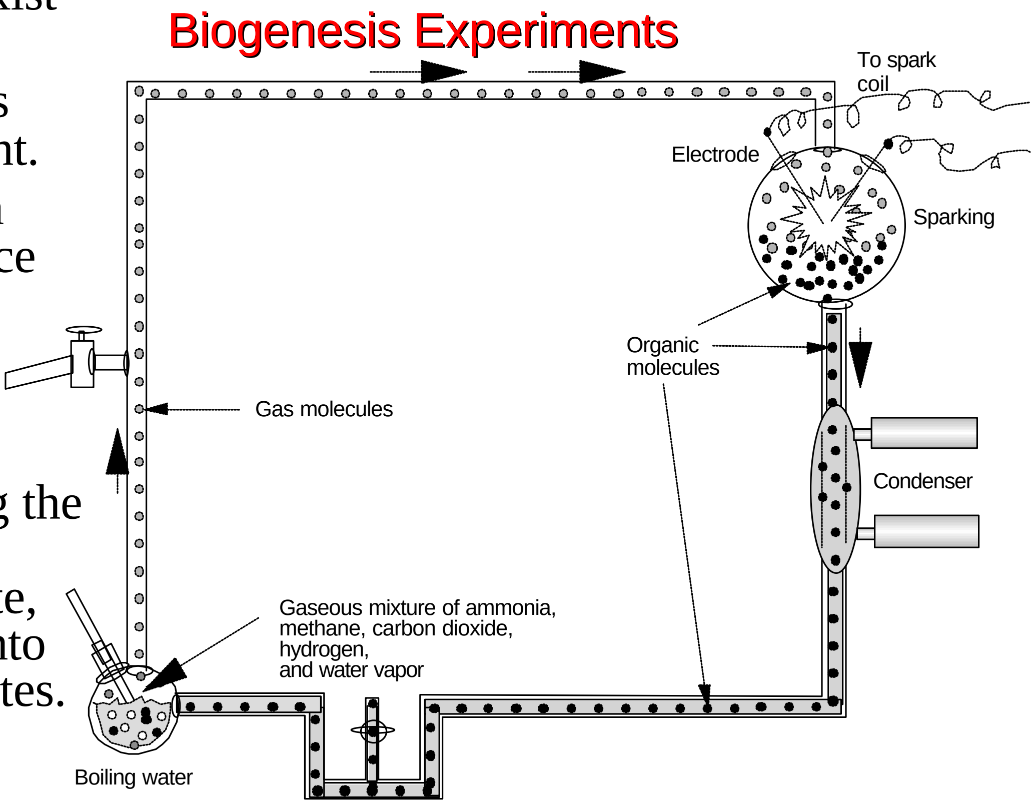
Biogenesis Experiments
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Cloning to produce stem cells

- Human DNA can be inserted in an animal cell, and the cell chemically tricked into proceeding as if fertilized.
- Formation, increase, and growth of stem cells begins and continues.

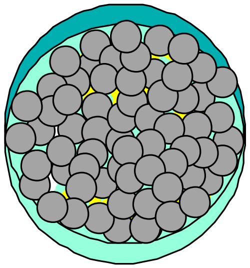

- The "engine" carried by the human DNA is changing the matter available into human cells.
- readily demonstrated experimentally.

Regeneration in a strain of rat with part of its immune system missing

- In mammals, the immune system plays a role in reducing the regeneration capability, compared to that of more primitive systems.
- In a strain of rat where that "suppressive" part of the immune system is missing, the regenerative capability is dramatically increased.
  - Can regrow cut away sections of tail.
  - Regrows plug cut out of ear.
  - Reconnects severed optic nerve.
  - Partially restores severed spinal cord.

- The suppressive "delta" engine of the immune system, operating simultaneously with the underlying primitive regenerative engine, sums to the modern much-reduced regenerative engine in the mammal.

An observation for the future

- When the science of engines is developed, physicians will be able to tailor these engine combinations to restore severed human spines, regrow severed limbs, etc.
- They will also be able to direct the body's "normal growth" to reduce and direct birth defects, genetic defects, etc.

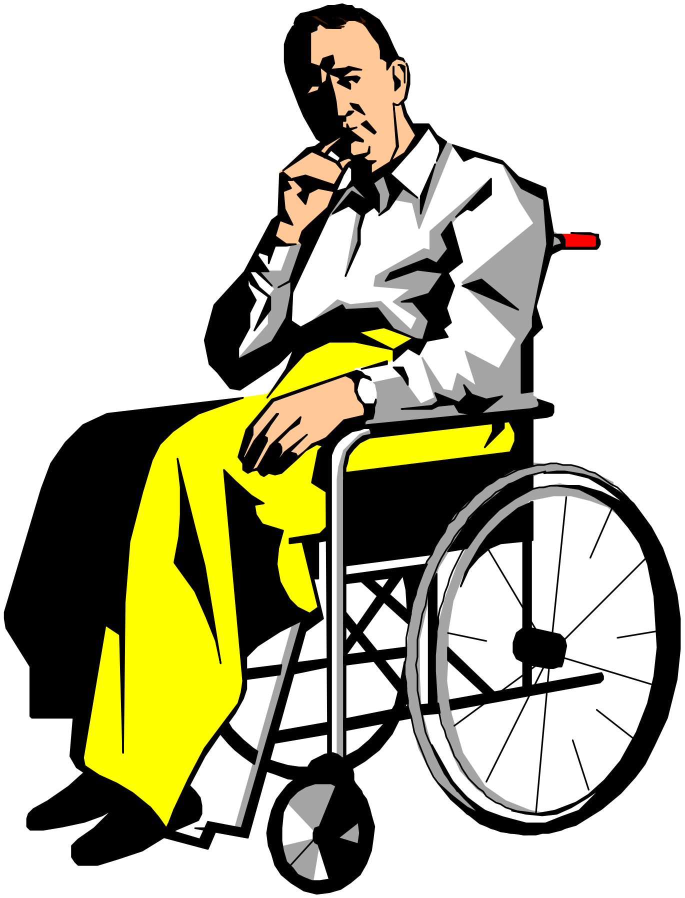

Biochemistry and EM models

- Medical scientists analyze such things as teratomas in terms of biochemistry and the biochemical model.
- An assembly of charges and charge distributions, with their associated fields and potentials, drives all chemistry at base level.
- The charge assembly and its dynamics involves a corresponding assembly of spacetime curvatures and their dynamics. These are *the cause that is driving the charge assembly and its dynamics, and therefore the chemistry*..
  - In unified field theory (such as that of Sachs) and O(3) electrodynamics.
  - Not in classical U(1) EM theory or quantum electrodynamics.
- Conventionally a causative EM field in space is "defined" as the *diversion from* that spatial field by and after its interaction with charge).
  The "definition" substitutes the effect for the cause, a non sequitur.
- This is a known major flaw, pointed out in one fashion or another by many physicists such as Wheeler and Nobelist Feynman.
- Classical EM theory is a 137-year-old model, seriously flawed. These flaws emerge in biochemistry and medical science.

Necessary EM corrections*

- Many of the corrections necessary in electrodynamics modeling either have been done or are underway.
- Higher symmetry, non-Abelian electrodynamics models have long been developed in particle physics.
- The new approach extends these models to also incorporate general relativity, thus producing a unified field theory*.
- In the new approach, spacetime curvature engines are perfectly natural and understandable, and can be directly modeled and engineered by advanced theorists.
- In short, the theory is sufficiently advanced and robust that technology and engineering can now advance.
- This will be a great leap forward for biochemistry and medical science, particularly in therapeutic methods.

* Sachs's unified field theory implemented by O(3) electrodynamics presently is the most suitable to use. Numerical methods are required.

Prioré's revolutionary results in lab animals

And its ramifications

Prioré (right) makes a point to Chaban-Delmas, Mayor of Bordeaux

- Many of the Prioré team's experiments were done in Bordeaux.
- Antoine Prioré patents:
  - "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968.
  - "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966.
  - "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963.

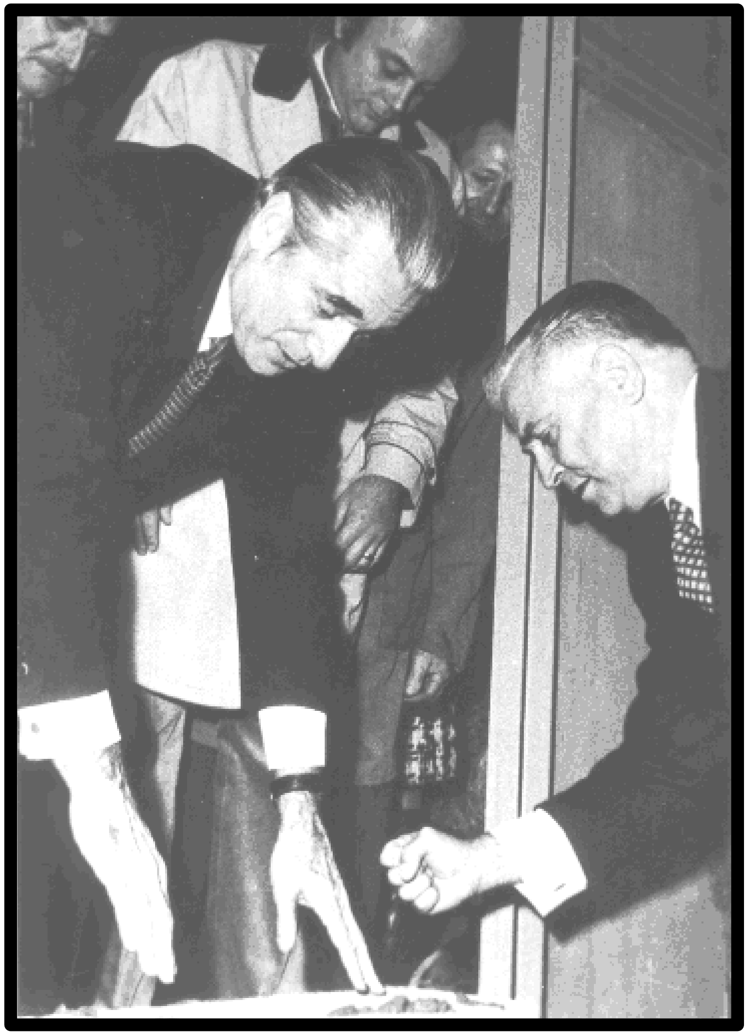

Block diagram of Prioré's method

Time-reversing the cells back to normal state

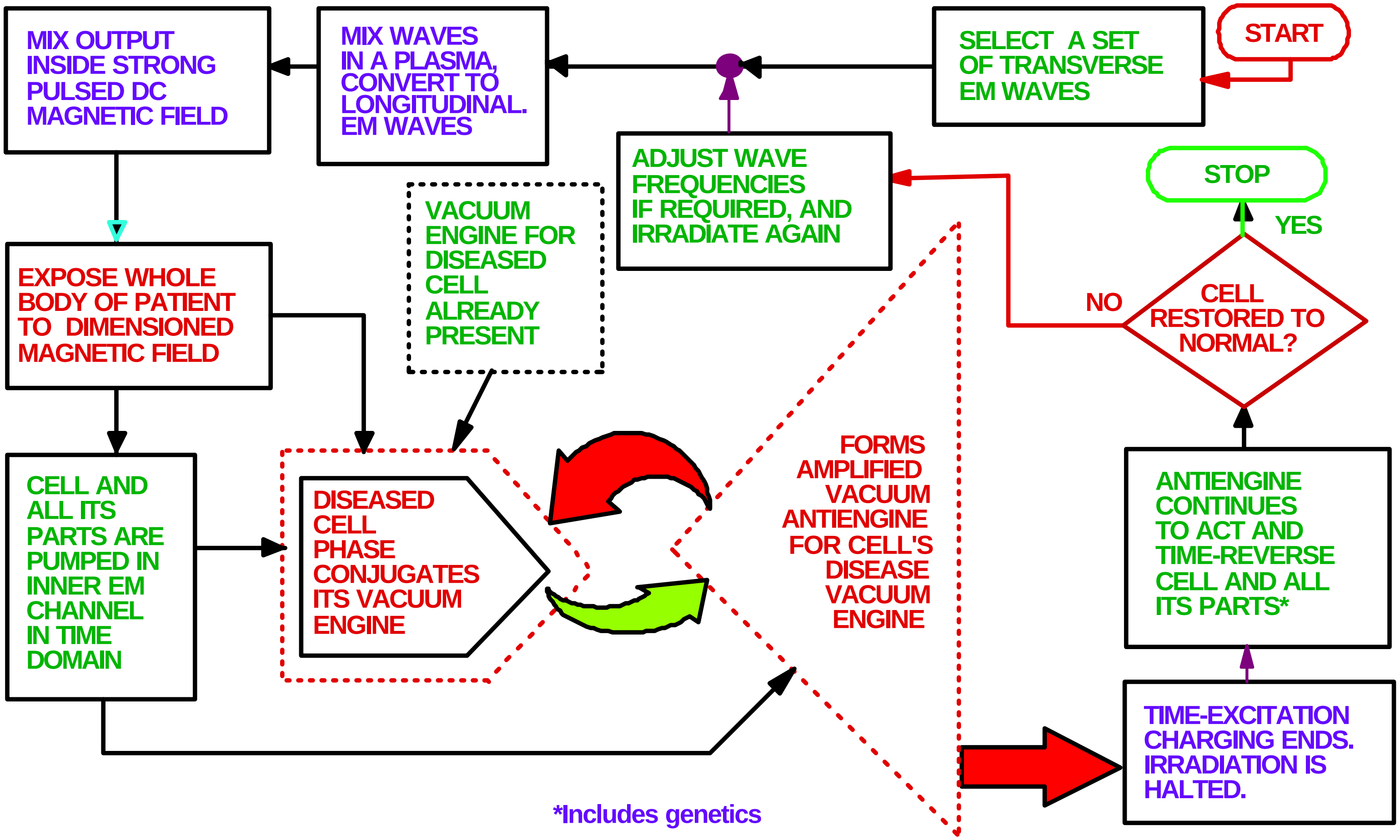

START

SELECT A SET OF TRANSVERSE EM WAVES

MIX WAVES IN A PLASMA, CONVERT TO LONGITUDINAL. EM WAVES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.

ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

YES

STOP

NO

ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN

*Includes genetics

In Laboratory Animals:

Prioré Team Cured

- **Terminal malignant tumors**
- **Clogged arteries (atheriosclerosis)**
- **Suppressed Immune Systems**
- **Infectious trypanosomas**
- ***Results were presented to French Academy of Science by Robert Courrier, Director of the Biology Section of the Academy***

© 1999 T. E. BEARDEN

- ***No one understood the mechanism!***

Some peculiar results by the Prioré team

- Once a rat is cured of a particular disease:
  - A single drop of blood from the cured rat, injected into another rat stricken with same disorder, would result in the injected rat getting well slowly.
  - This ability to transfer immunity and healing in the extracted drop of blood gradually decayed over a period of a few weeks.
- Showed that even the fields and potentials in the blood had been altered internally, in the treated animal.
- Showed that the injected animal's cellular regenerative system possessed the ability to also time-domain pump over the entire body, everywhere reducing the delta between that injected drop of blood now the rat's own, and the rest of the blood in its body.

Proof of cellular time-reversal

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only the *previous ability* of the immune system is restored by the process.

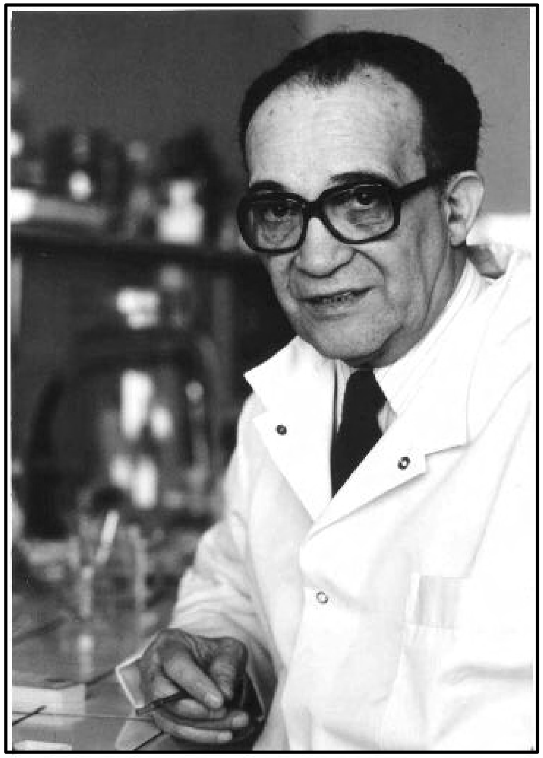

PROF. RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Disadvantages of Prioré method

- Large plasma tube and facility required.
  - 3-stories high building bay if to treat humans.
  - Long irradiation time (2-3 hours).
  - Slow, numerous experimental adjustments.
- Fundamental mechanism was unknown.
  - Phase conjugate optics not yet born in West.
  - No higher symmetry EM modeling used.
  - Prioré thought it was a result of ionization.
  - Importance of difference frequency unknown.
  - Body dielectric transmission path unknown.

3-Story high Prioré device for treating humans

- At the end of the project, Prioré had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient -- and also expensive

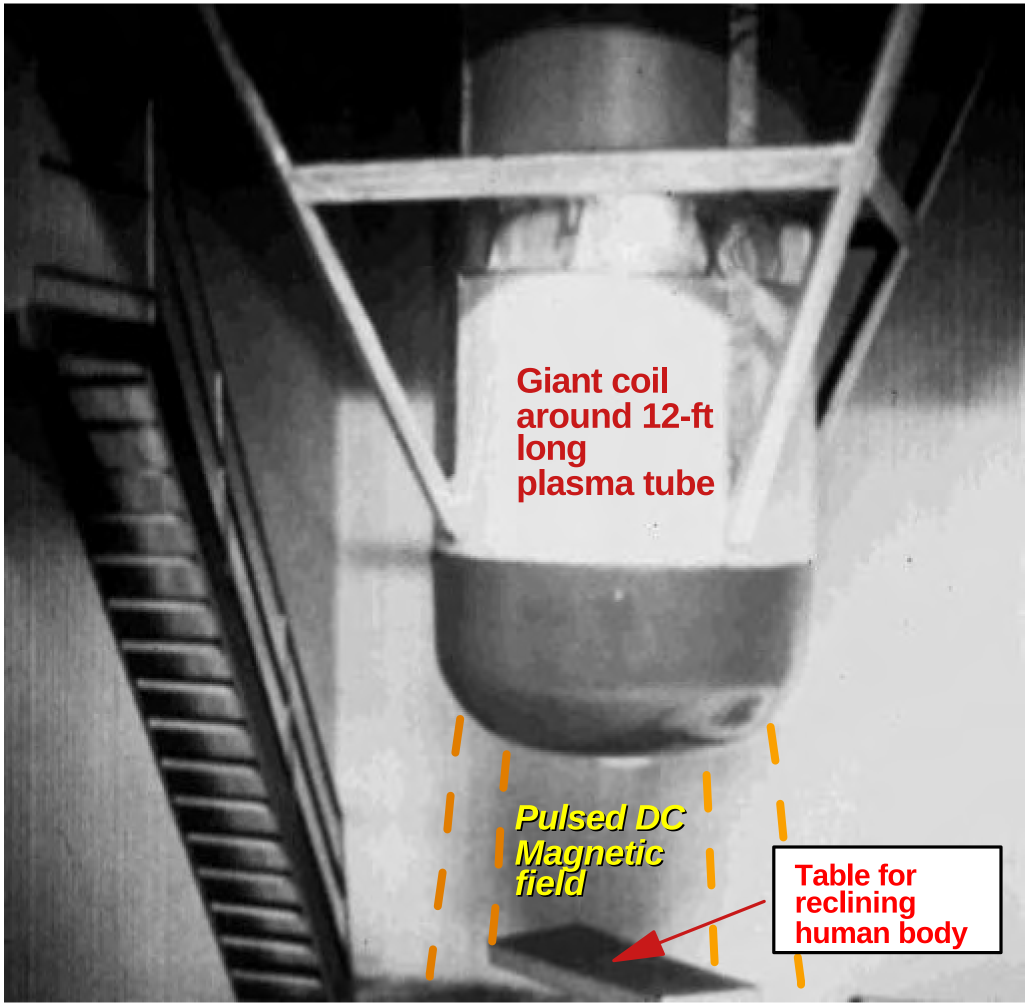
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

17-foot Prioré plasma tube

Used in large machine (3-stories high) to treat a limited number of human patients

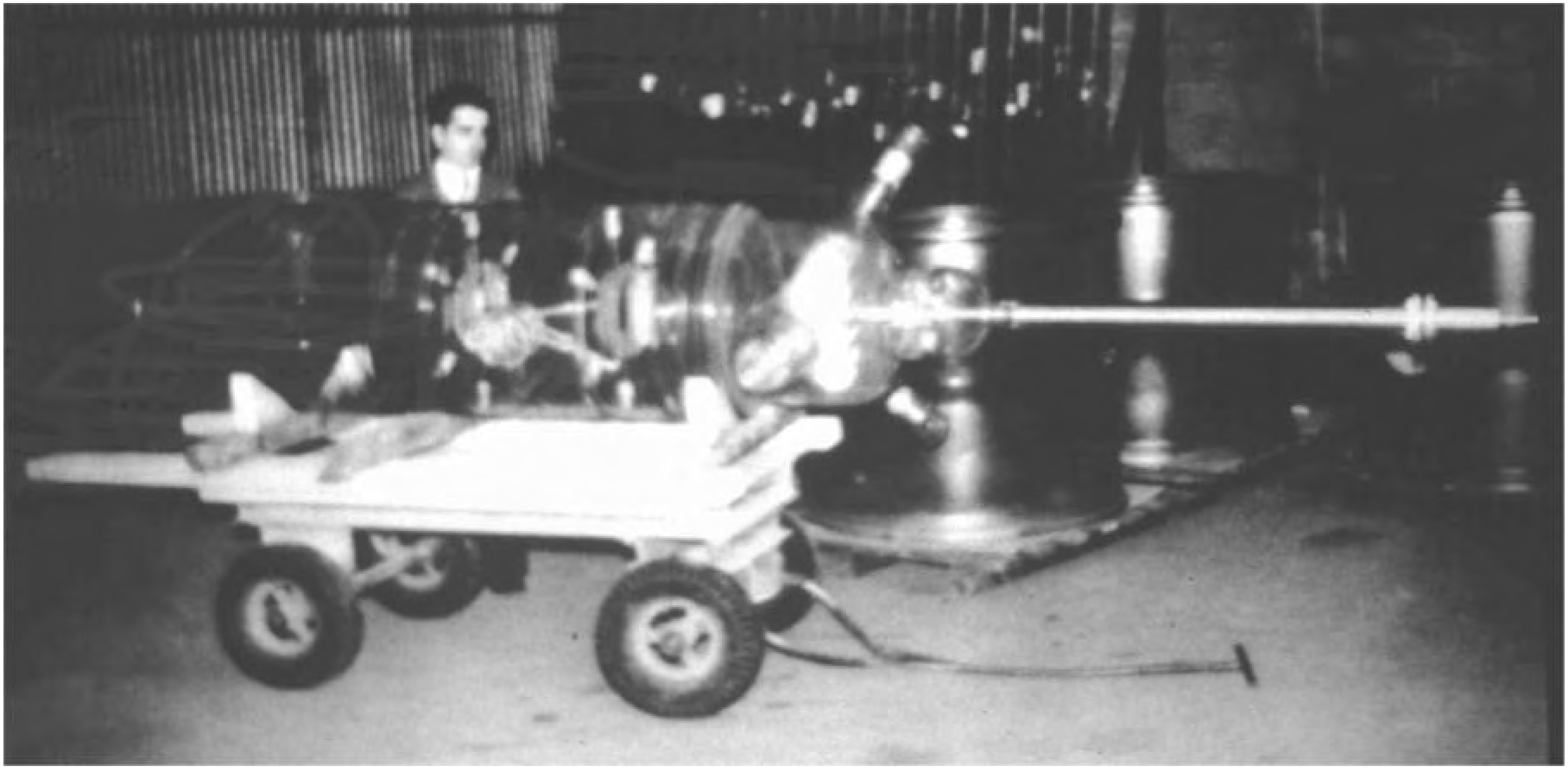

A proposed new mechanism for cancer promotion

And its ramifications

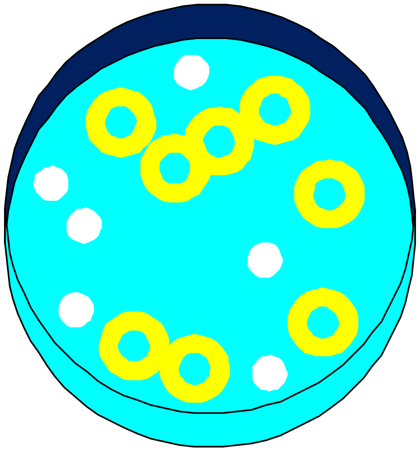

Remarks on cancer from a new perspective

- Two stages: cellular damage and promotion of the damaged cell to a tumorous cell. Promotion presently is poorly understood if at all.
- A master cellular control system (MCS) (studied by Popp et al.) controls both the immune system and the cellular regenerative system, as well as normal functions of the cells.
- The "past history" of ancestral development of aerobic cells and higher multicell aerobic bodies on planet Earth, resides in the "internested engine layers" in every tiniest part of the entire body dielectric, and in all cells.

Remarks on cancer from a new perspective (2)

- Faced with cell damage (such as sustained hypoxia) it cannot handle, the MCS continues to "peel the layers" in taking actions related to ever more distant pasts.
- Faced with continuing hypoxia, e.g., the MCS eventually reaches ancestral layers -- with an engine -- it forces the damaged cells to start moving back toward the original primeval anaerobic form.
- The first step backward is *separation from central growth control*, freeing the cell to multiply independently. At the same time, genetic changes in the cell are forced and produced, starting it back toward becoming an anaerobe.
- These actions comprise the *promotion process*.

The promotion process via engines

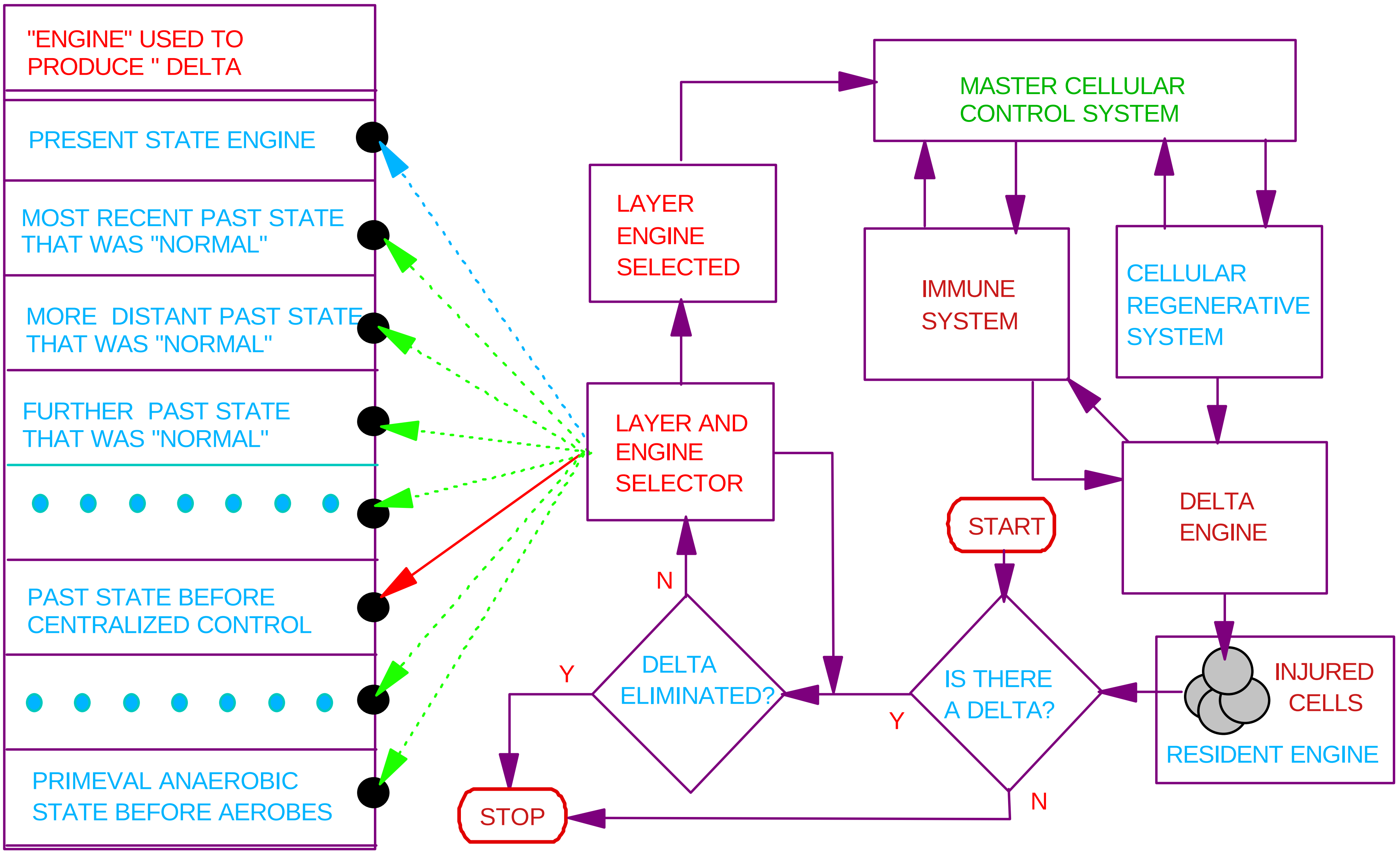

"ENGINE" USED TO PRODUCE " DELTA

PRESENT STATE ENGINE

MOST RECENT PAST STATE THAT WAS "NORMAL"

MORE DISTANT PAST STATE THAT WAS "NORMAL"

FURTHER PAST STATE THAT WAS "NORMAL"

● ● ● ● ● ● ●

PAST STATE BEFORE CENTRALIZED CONTROL

● ● ● ● ● ● ●

PRIMEVAL ANAEROBIC STATE BEFORE AEROBES

MASTER CELLULAR CONTROL SYSTEM

LAYER ENGINE SELECTED

IMMUNE SYSTEM

CELLULAR REGENERATIVE SYSTEM

LAYER AND ENGINE SELECTOR

DELTA ENGINE

START

DELTA ELIMINATED?

IS THERE A DELTA?

INJURED CELLS

RESIDENT ENGINE

STOP

Evolution of aerobic systems

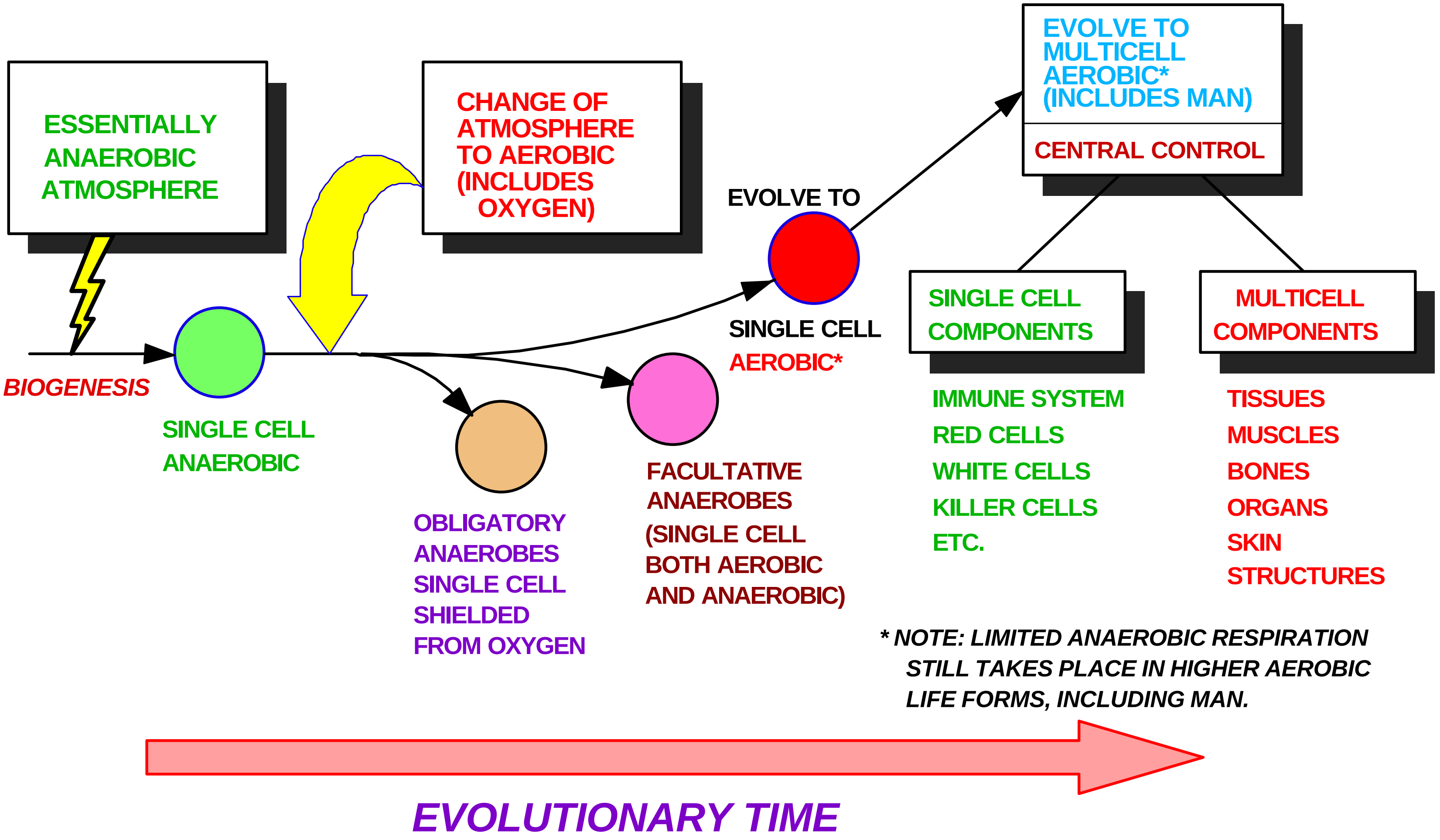

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

First step in self-promotion of cancer due to hypoxia

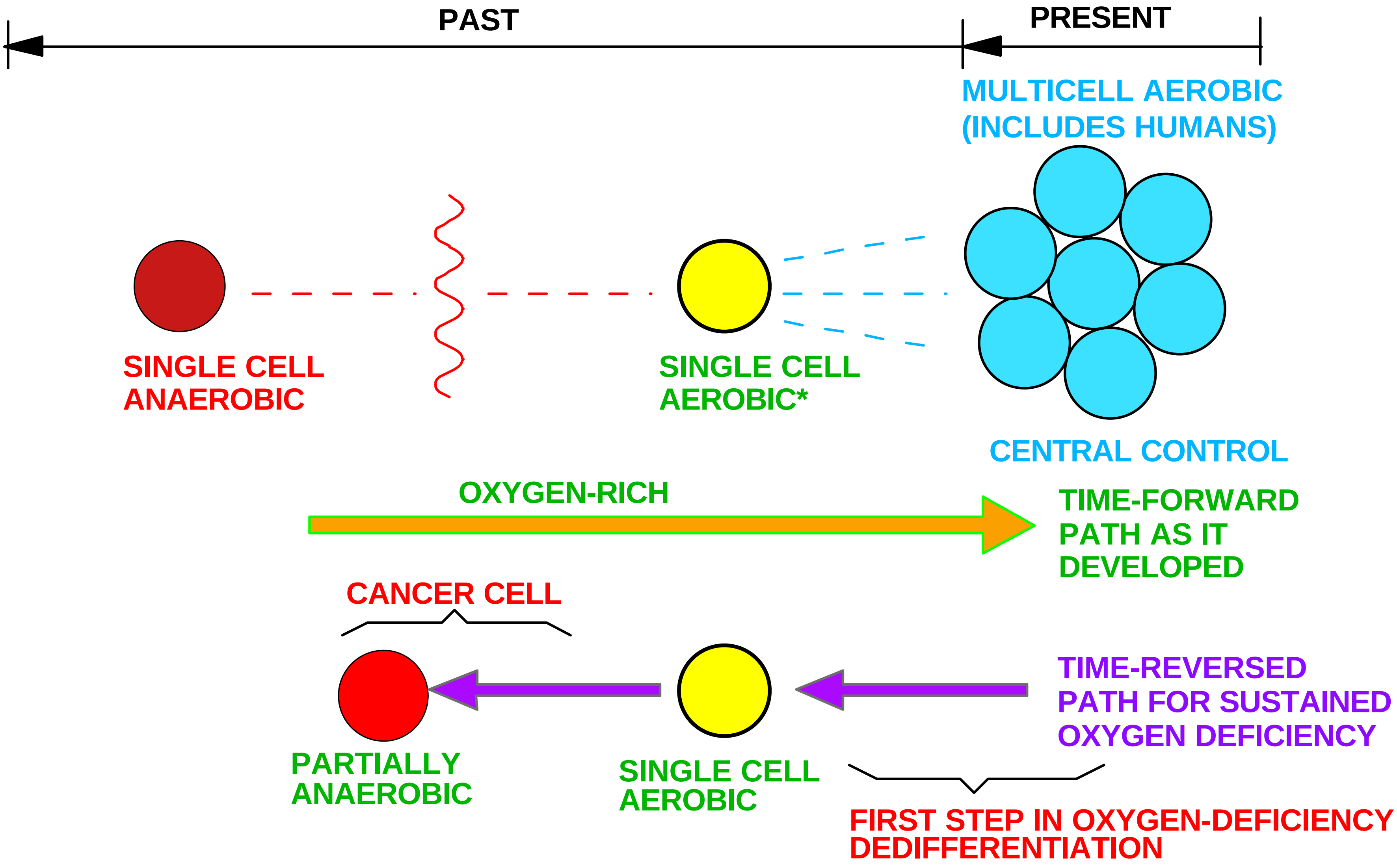

PAST

PRESENT

MULTICELL AEROBIC
(INCLUDES HUMANS)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

Block diagram of the cellular regenerative system

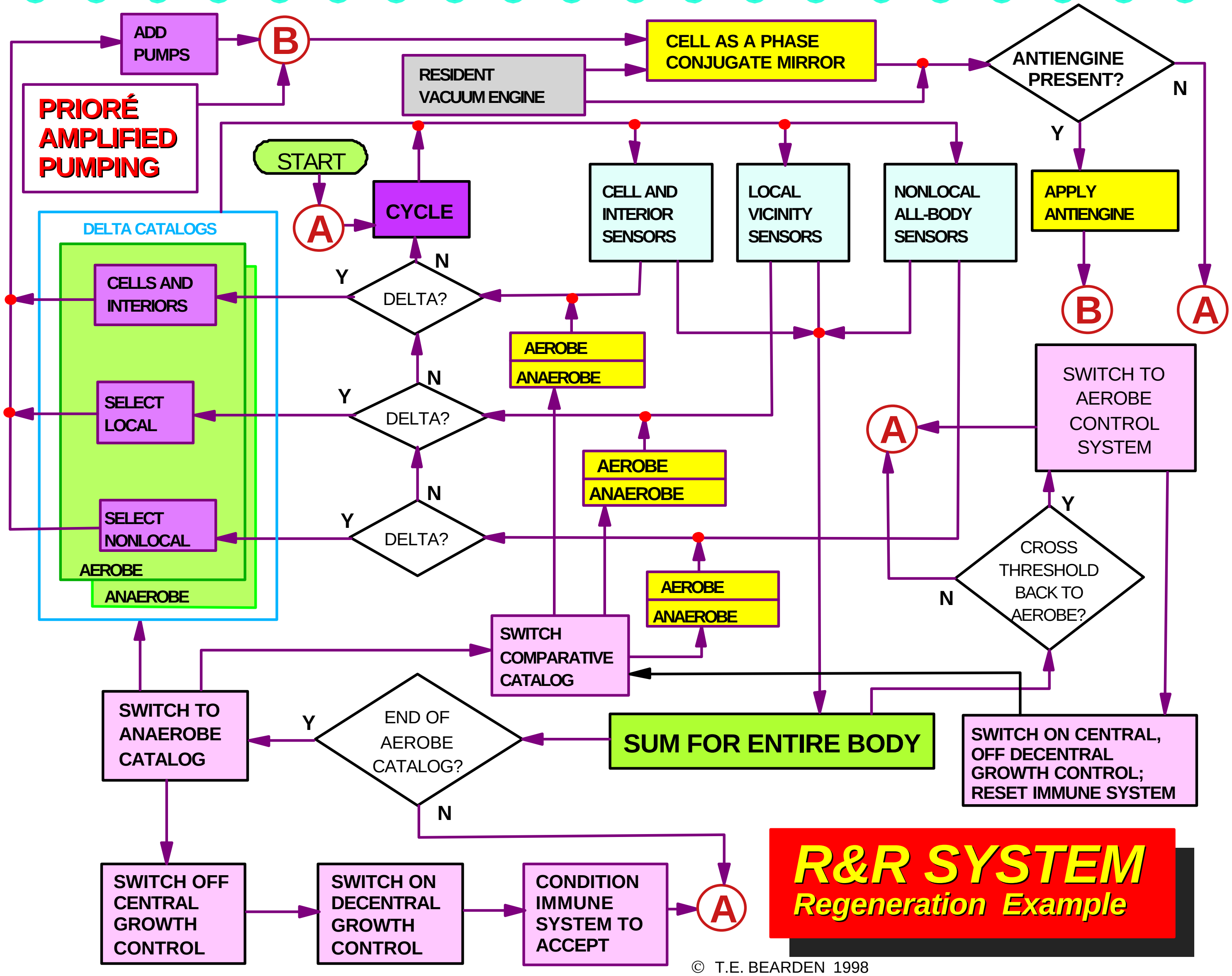

ADD PUMPS

B

CELL AS A PHASE CONJUGATE MIRROR

RESIDENT VACUUM ENGINE

ANTIENGINE PRESENT?

N

Y

APPLY ANTIENGINE

B

A

PRIORÉ AMPLIFIED PUMPING

START

A

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

DELTA CATALOGS

CELLS AND INTERIORS

SELECT LOCAL

SELECT NONLOCAL

AEROBE

ANAEROBE

N

Y

DELTA?

AEROBE

ANAEROBE

N

Y

DELTA?

AEROBE

ANAEROBE

N

Y

DELTA?

AEROBE

ANAEROBE

SWITCH TO AEROBE CONTROL SYSTEM

A

Y

CROSS THRESHOLD BACK TO AEROBE?

N

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

Y

END OF AEROBE CATALOG?

N

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Bombshell: Amplified EM healing

- A unified field theory approach and use of supersystem analysis has broken the basic mechanism for healing itself.
  - The Sachs unified field theory applies and has been partially fitted.
  - Evans-Vigier O(3) electrodynamics can properly model it, including both the internal and external electrodynamics and effects of engines.
  - Direct engineering development is now possible.
- Experimental proof of the results obtainable already exists in the literature.
- A medical revolution is in the offing.

Potential Medical Applications

- **Non-traumatic cancer treatment**
  - ***Cure cancers and leukemias of all types***
  - ***Remove pre-cancerous state***
  - ***Restore system***
- **Treat and cure Infectious diseases**
- **Cure AIDS and restore immune systems**
- **Defense against biological warfare**
  - ***Virulent diseases from unknown agents***
  - ***Treat, cure prior to symptoms developing***
  - ***Both civil populace and armed forces***
- ***Dramatically lower health costs***
- ***Rejuvenate the aged***

© 1999 T. E. BEARDEN

To treat and save mass casualties: What is urgently needed now:

- A small, portable, relatively inexpensive treatment unit, not requiring plasmas etc.
- Treatment in 2 to 5 minutes rather than 2-hours.
- Simple treatment application, almost anywhere.
- Effective in advanced stages as well as prior to emergence of symptoms.

- Effectiveness against a wide array of illnesses.
- Use is easy, high school student can be trained in 30 minutes.
- Crash program to develop, flood down through emergency systems in great quantities.

The Solution:

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.

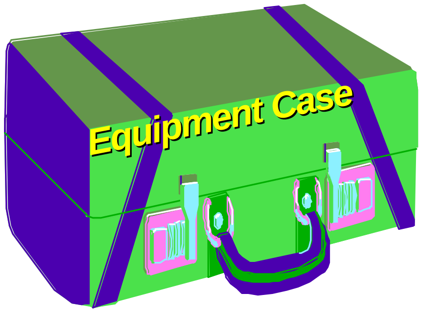
Equipment Case

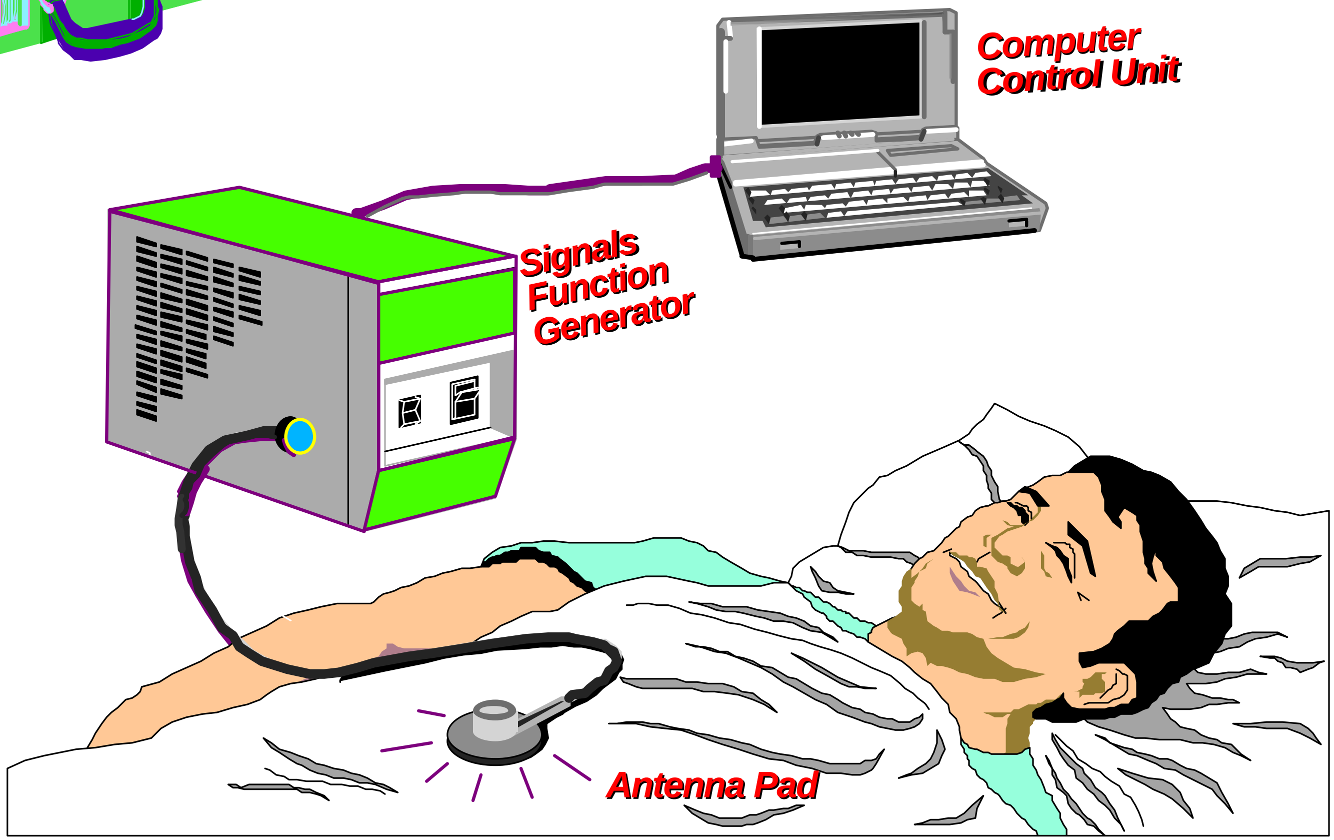
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

The Porthole concept:

The two-minute treatment shortcut

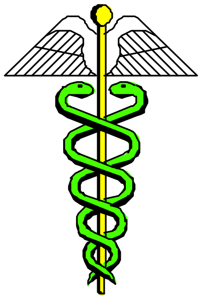

How the porthole concept works

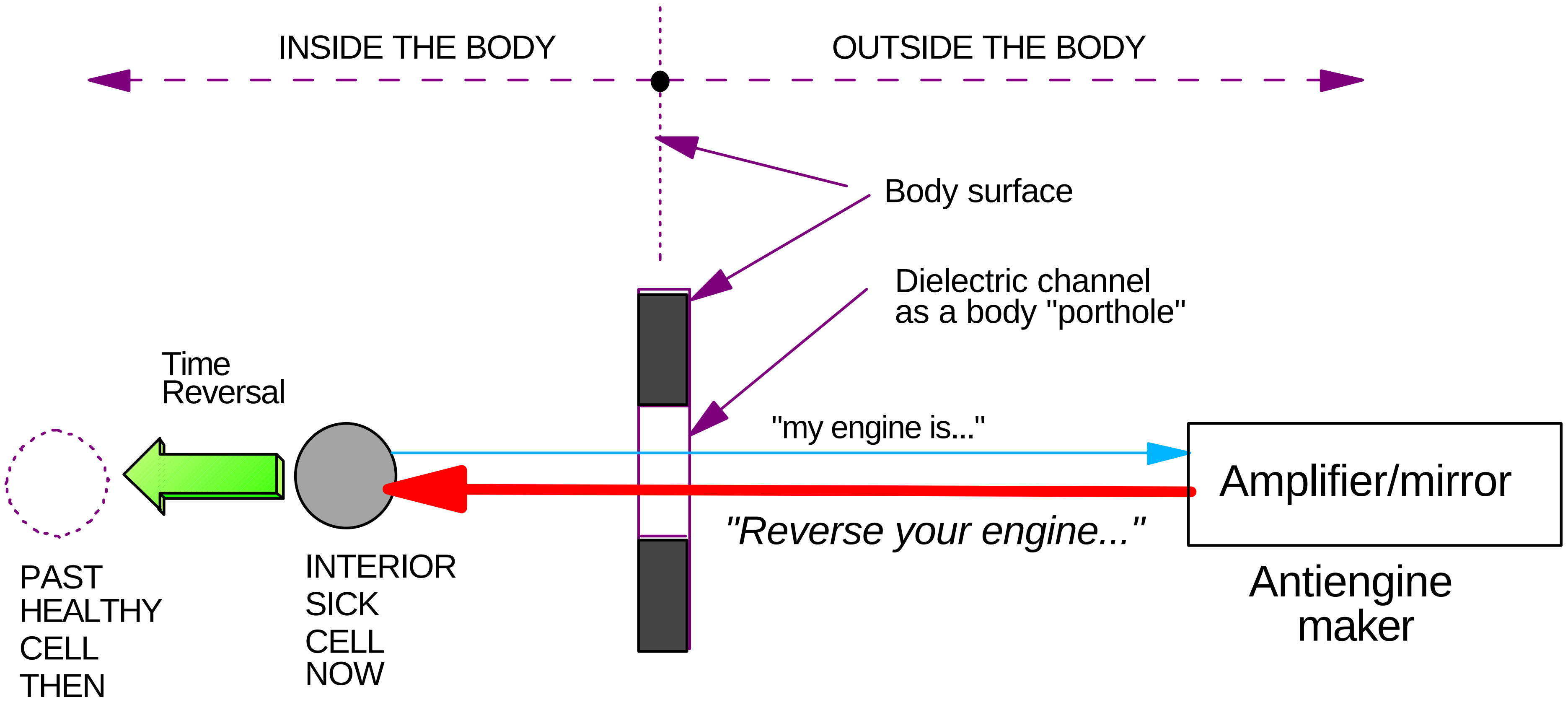
INSIDE THE BODY
OUTSIDE THE BODY
Body surface
Dielectric channel
as a body "porthole"
Time
Reversal
"my engine is..."
Amplifier/mirror
"Reverse your engine..."
PAST
HEALTHY
CELL
THEN
INTERIOR
SICK
CELL
NOW
Antiengine
maker

Porthole concept: An analogy of the dielectric path and inner-outer functional connections

- Imagine perfectly reversible dynamic light processes inside a large "cruise ship", that are also perfect phase conjugate mirrors
- Each process receives, absorbs, and emits light continuously
- Light escaping through the portholes is a composite from all the dynamic processes ongoing inside
- If we record and amplify the escaping light, then force it back through the portholes, it reverses all the internal processes
- The processes revert back, slowly, to a previous ongoing process state
- We cease radiation, and the processes reverse and move forward again

Porthole concept: The discovery

- The processes themselves, *in their own "frame"*, are never actually "reversed". Instead, the frame's motion through time is reversed.
  - Each to itself moves normally -- just receiving, absorbing, and radiating light, regardless of "whatever light" is interacting
  - This is due to a peculiarity of the photon and photon interaction
- Photon interaction creates observed physical reality, including the "frame" and the seeming "passage of a mass through time"
  - The photon is its own antiparticle, which is just a photon traveling in the opposite direction
  - So in this case we can use "light" to precisely time reverse (to the observer) the frame and the apparent operations (to the observer)
- When we cease irradiation, after a decay period to remove the reversed excitation, all functions are now seen to be moving forward normally in observer forward time, but from a previous physical state or condition

Using the porthole concept: Research

- The body's own emitted dense, broadband signals are recorded across a broad bandwidth
  - These signals are amplified, then rebroadcast precisely back into the body
  - This returns an amplified "reversed" signal set back through the dielectric path, invoking and applying the porthole concept in all processes
- The returning photons produce time-charging but in reversed frame direction
  - Tremendous potential energy is absorbed in highly compressed "time-energy" form in seconds
  - This "time-potential" slowly drains off as the body processes and cells reverse back to a previous state

Research using the porthole concept

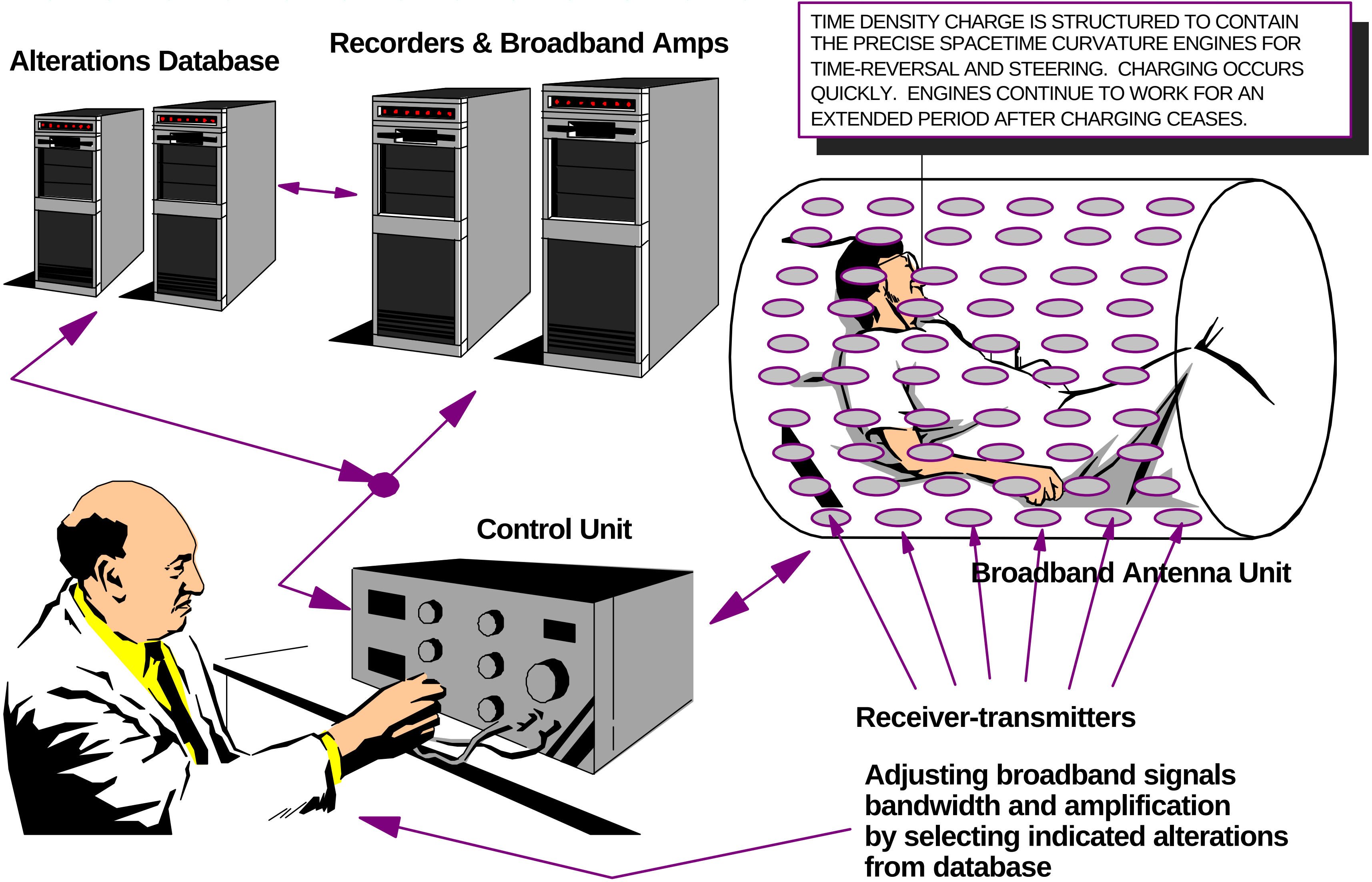

Alterations Database

Recorders & Broadband Amps

TIME DENSITY CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Control Unit

Broadband Antenna Unit

Receiver-transmitters

Adjusting broadband signals
bandwidth and amplification
by selecting indicated alterations
from database

© T.E. BEARDEN 1998

Portable unit then developed for rapid treatment of mass BW casualties

- Two blankets have antennas in them.
- Uses the "porthole" concept (discussed shortly)
- Eliminates plasma tube mixing etc.
- Computer-controlled complex signal structure.
- Irradiation for 2 minutes.

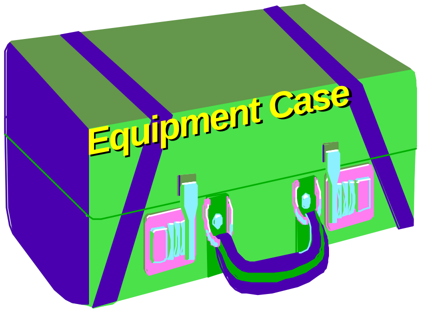
Equipment Case

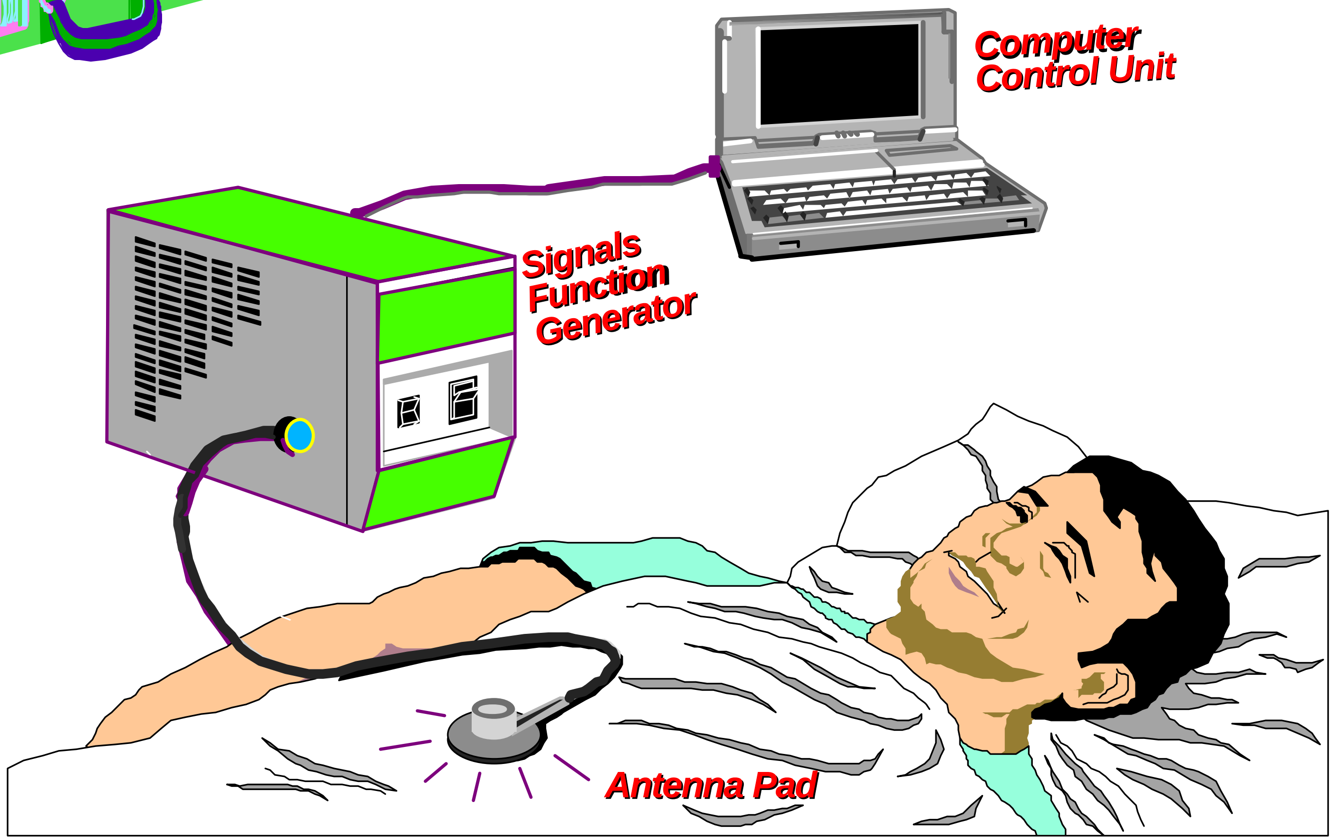
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Training to use the unit

- High school student level, 30 minutes.
- Patient lies on one blanket, other covers him.
- Set the prescribed standard signal sets from listing.
- Computer-controlled generator produces the signal.
- Irradiation for 2 minutes, uncover patient, bring in another.
- Unit can be self-powering from vacuum energy (MEG*).

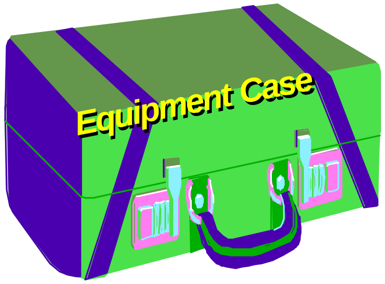
Equipment Case

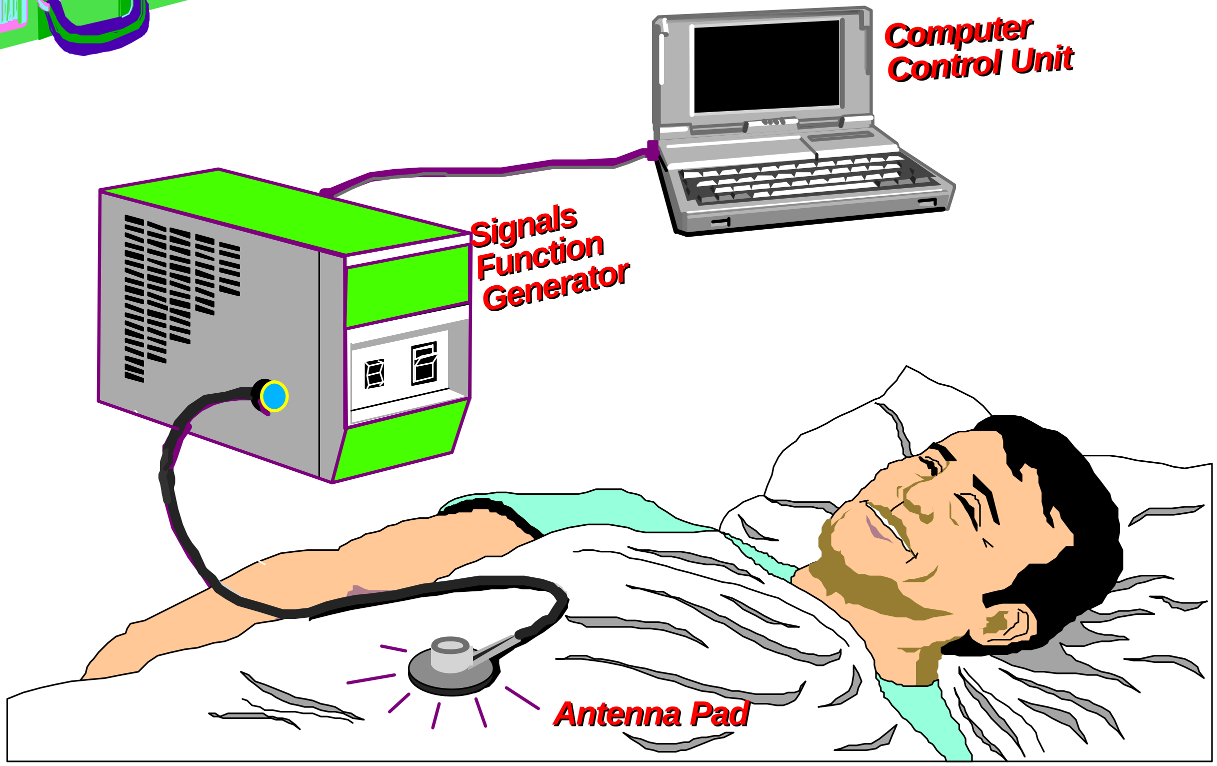
Computer
Control Unit
Signals
Function
Generator
Antenna Pad

Additional advantages

- Can be developed in a crash program in 2-3 years.
- After sunk costs, projected cost per unit in mass production is less than $100,000
- Can be produced by the thousands
- Then flooded down through
  - Fire stations and police stations
  - Red cross units, hospitals, clinics
  - Emergency action teams and armed forces
- Large numbers of operators easily produced
- Emergency training could be watching a 15 minute film, then performing one graded practice run
- First generation equipment can save some 70% of the coming mass casualties from anthrax etc.

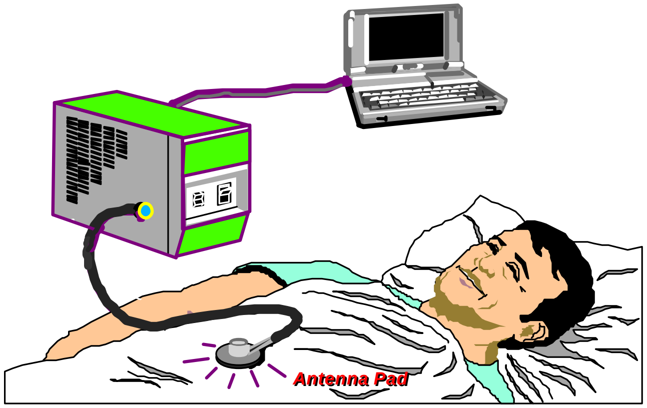

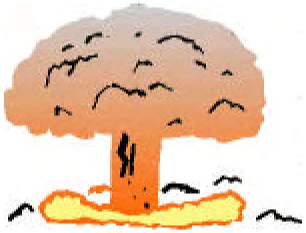

Terrorist BW warfare capabilities against the United States

The growing threat

Mass Casualty Problem:
Aerial Anthrax Spray on Washington, D.C.

- **One terrorist, one light aircraft with spray tank**
- **100 kilograms of anthrax**
- **Flies over greater metropolitan Washington, D.C.**

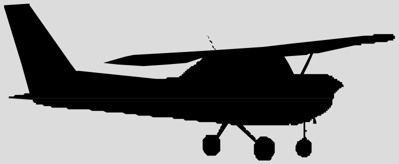

- **Calm night**
- ***1-3 million casualties result****
- ***Most of those stricken will die***

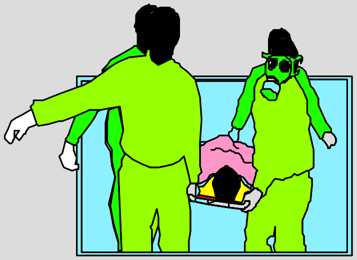

- **Presently little can be done to save the stricken civilians**
- ***Attacks on several population centers might produce some 10 million or more casualties***

***Per OTA Report to Congress, 1993**

For decades, infiltration of terrorist teams into the U.S.

- From several foreign nations, not just Bin Laden in Afghanistan
- Several thousand Cuban guerillas, from training camps in Southern Mexico
- Russian Spetznaz teams
- Nuclear warheads (up to 80 KT)
- Biological warfare agents
- Yakuza in oriental populations in our cities
  - Yakuza manufactures Russian energetics weapons in their own facilities in Japan
  - Possesses the new EM weapons referred to by Secretary of Defense Cohen in 1997
- We are now in a very long, hard war and will suffer millions of casualties in our cities. It may last for decades.

Defense Secretary Cohen*

(Involvement of secret superweapons)

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."

* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Some typical BW agents

- Smallpox, including different strain developed by Russia (some sold to terrorists)
- Anthrax
  - Over 400 known locations around the Earth have cultures
  - Difficult to decontaminate
- Bubonic Plague
- Botulism toxin
- Tularemia (rabbit fever)
- Hemorrhagic fevers, such as Ebola

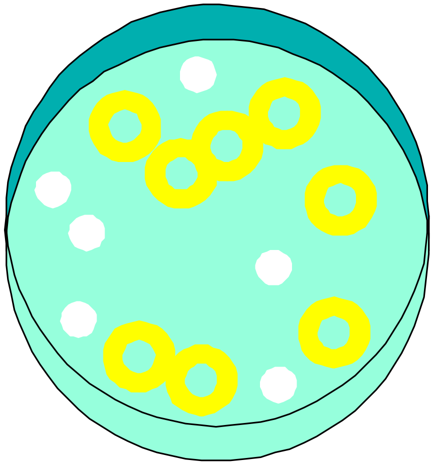

Anthrax

- OTA study: Some 200 pounds released in a spray, on calm night over Washington, D.C. could result in up to 3 million casualties
- Symptoms usually in 2-10 days
  - Sometimes as long as 7 weeks
  - Flu-like symptoms

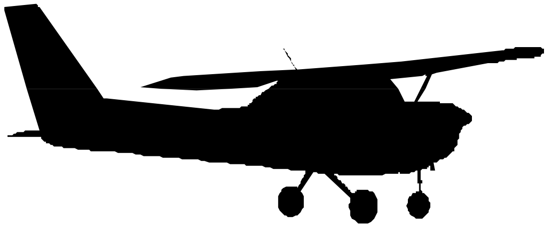

- Then 1-3 days later, shock and breathing problems cause death in up to 100%
- Antibiotics given early (before symptoms) can prevent infected persons from becoming ill
- Limited vaccine, reserved for military

Smallpox

- More than 500 million people killed in the 20th century before "eradicated" in 1977. Vaccination ceased in 1980.
- Russians kept it, developed new strains, sold to terrorists
- Symptoms in about 12 days: fever, headache, nausea. Rash similar to chickenpox then appears.
- Rash turns into hard blisters. *Highly contagious*, kills one-third of its victims.
- No cure. About 12 million doses of vaccine left.
- If smallpox is unleashed in any major city on earth, it will eventually kill 2 billion, or about 1/3 of the human species -- unless a radical new cure is found and utilized.

Bubonic Plague: Historically the most feared (Black Death)

- From 1980-1994, 18,739 cases in 20 countries.
- Symptoms occur in 1-6 days after inhaling the pneumonic form.

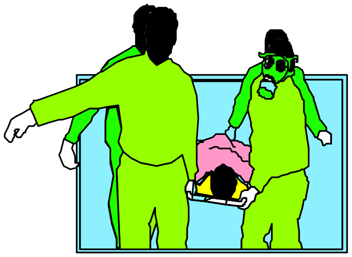

- High fever, cough, labored breathing lead to respiratory failure and death.
- It is contagious.
- Rapid use of antibiotics can be effective.
- Vaccine not currently produced in the U.S.

Botulism toxin (a nerve toxin)

- Single most poisonous BW substance known.
  - Typically borne in food, contacted by eating.
  - Could be developed as aerosol weapon.
- Symptoms in 24-36 hours; blurred vision and difficulty swallowing, speaking.
- Paralyzes muscles, leading to respiratory failure and death.
- Centers for Disease Control and Prevention maintain a botulism antitoxin supply.

Tularemia

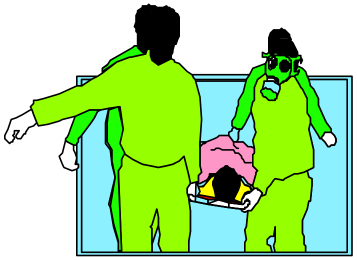

- U.S. studied its weapon potential in 1950s and 1960s.
- Symptoms appear in 3-5 days, including fever, chills, headache, and weakness
- Inflammation and hemorrhaging of the airways can lead to death.
- Without antibiotics, one-third die. Can be 90% if breathed in, in aerosol form.
- Vaccine developed and under review by FDA.

Hemorrhagic fevers, such as Ebola

- E.g., origin of Ebola is unknown. Probably transmitted to humans by animals, to begin a breakout.
- Symptoms in 3-5 days, include fever, aching muscles, diarrhea.
- Hemorrhaging of fluids out of tissues and orifices.
- From 30 to 90 percent die, depending on specific fever type.
- Some of these fevers respond to antiviral drugs.
- The drugs are in short supply.
- Presently *containment* is the primary goal.

A factor the U.S. does not know

(Secret involvement of clandestine superweapons)

- Three nations (not the U.S.) developed quantum potential weapons using Bohm's hidden variable theory of quantum mechanics.
- Can induce disease engines in matter or bodies at a distance.
- Have developed "spreading" of immune systems in a targeted populace (cleverly tested in Gulf War against U.S. troops).
  - Cocktail mix of disease engines for 2 dozen different diseases
  - Targeted area's human immune systems react to these engines when they are still in virtual state, just below quantum threshold.
  - Spread immune system's resources thus "thinned" across response to two dozen shadow diseases at once.
- Hit with a terrorist BW attack, the spread immune systems are overwhelmed much more easily and quickly.
  - Yield (casualties) of the attack is increased by factor of 3-5.
  - Spreading of U.S. immune systems has been accomplished and is still being maintained.

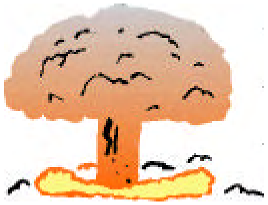

Summary and conclusion

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

The new war

A new Pearl Harbor: Strike on the World Trade Center in NYC

- The North Tower has already been struck and is burning.
- A second jetliner is about to strike the South Tower.

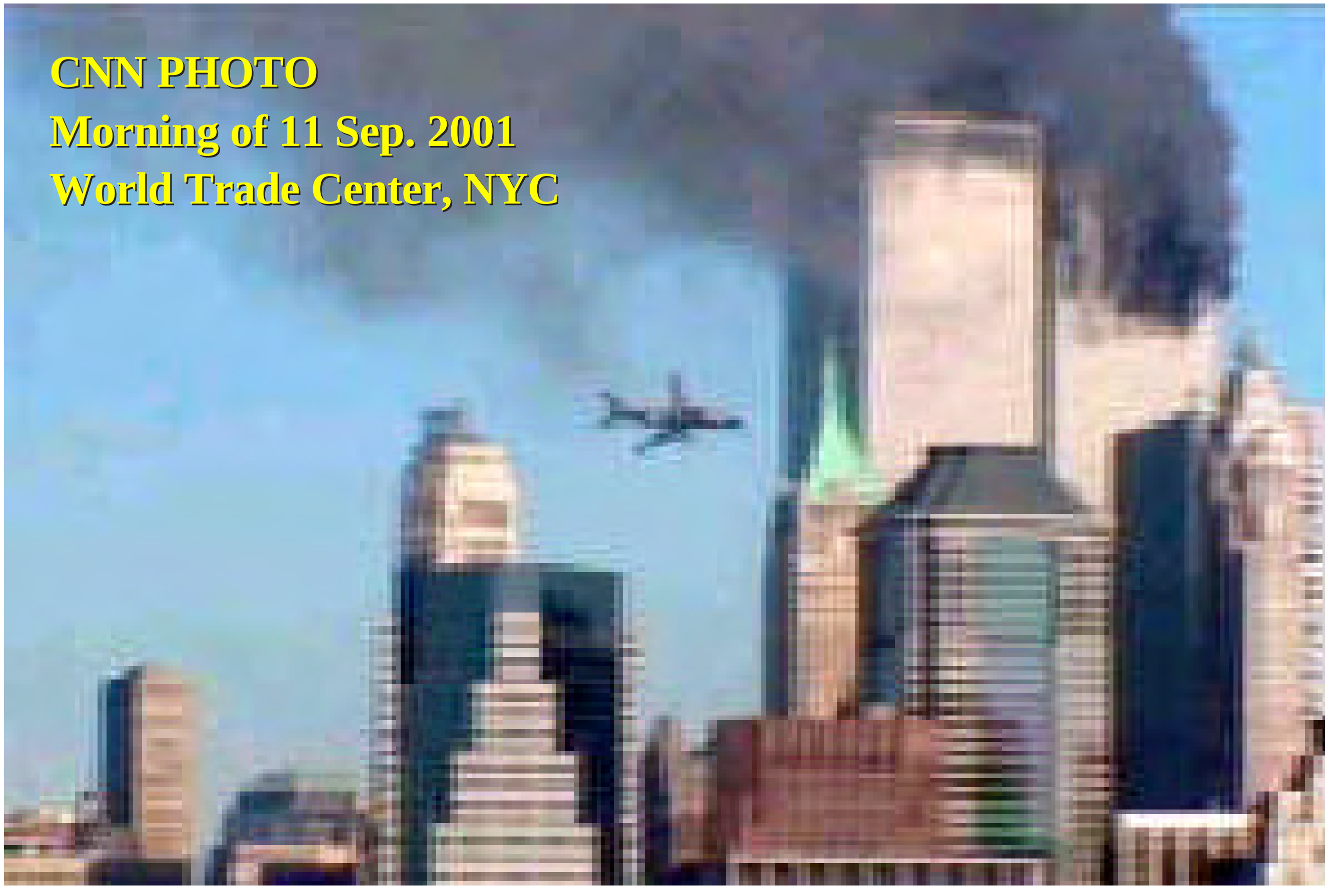

CNN PHOTO
Morning of 11 Sep. 2001
World Trade Center, NYC

Fireball From Impact on South Tower

- The two buildings held some 50,000 occupants
- Evacuation was underway after the first strike
- More than 6,000 persons were killed
- This is more than were killed in the Dec. 7, 1941 strike on Pearl Harbor

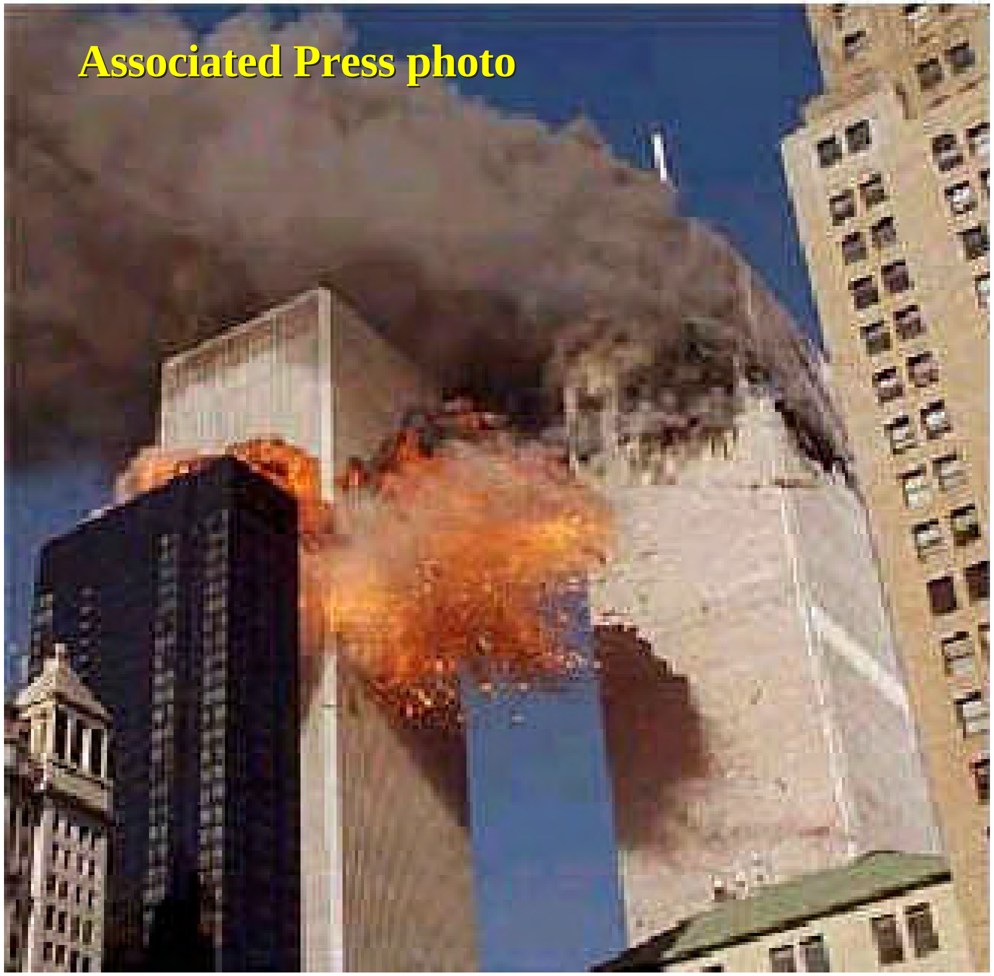

Associated Press photo

Collapse of the South Tower

- The first tower to collapse was the South Tower
- It is still uncertain just how many people were caught inside and killed
- This is a half million tons of debris, falling from the New York skyview.

Photo from CNN

Collapse of the North Tower begins

- The North Tower then collapsed
- This was another half-million tons of debris falling from the NYC skyview

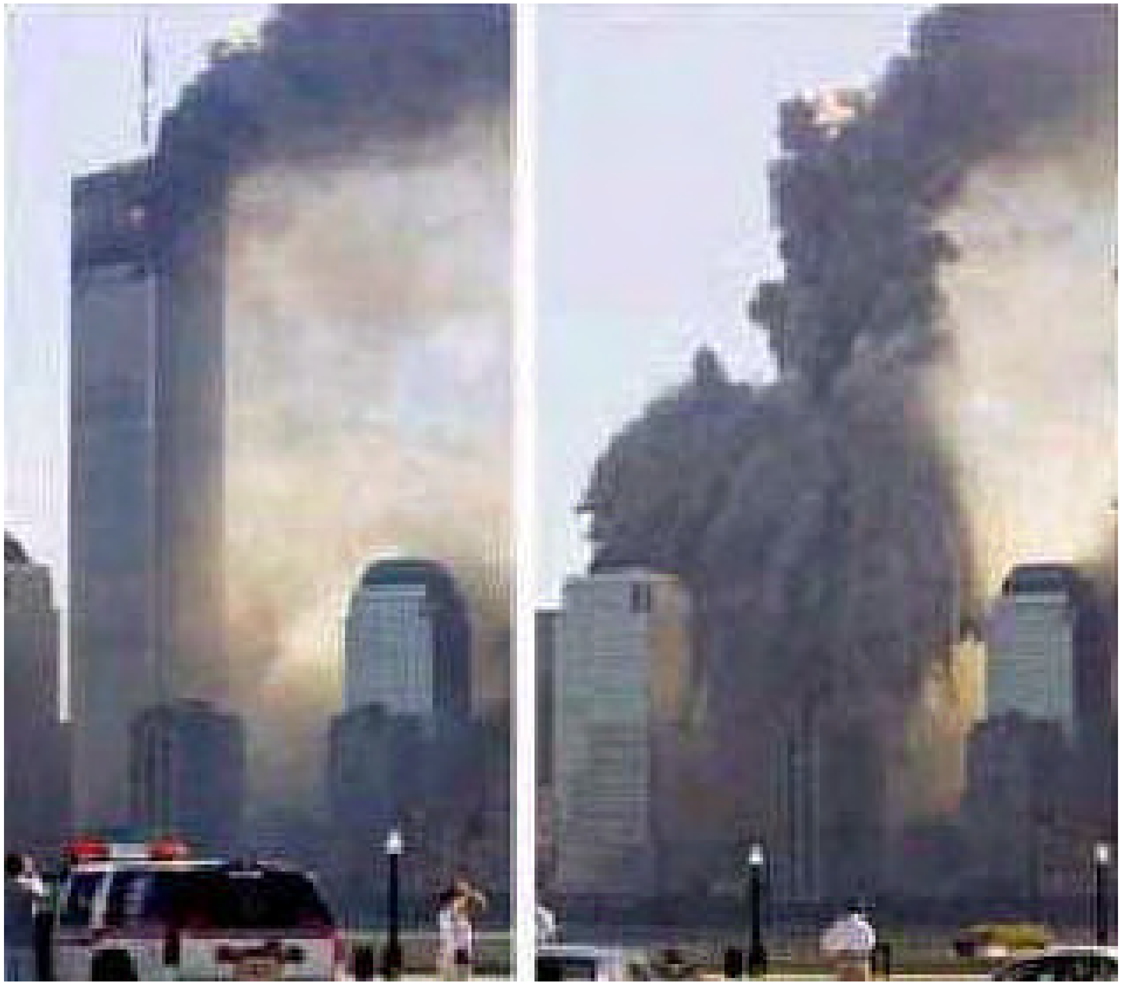

Collapse of North Tower finishes

- Collapse of the North Tower killed many rescue workers, firemen, and policemen
- At this writing, the total casualties are still unknown

Photos from CNN

Next began the recovery efforts

- Recovery operations will require months
- At this writing, the total casualties are still unknown
- The damage was done by two civilian jetliner "bombs"

The Pentagon was hit on the same day

- Hundreds were killed the same day when another jetliner struck the Pentagon.
- Yet another jetliner, possibly headed for the White House, crashed enroute when the passengers attacked the suicide terrorist hijackers.

A sobering thought

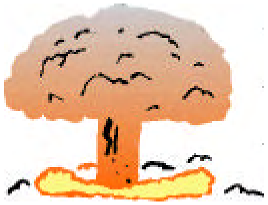

- Nuclear weapons are already smuggled into the U.S., as are BW agents such as anthrax, smallpox, etc.
- Had a 40-80kT weapon been exploded in New York and one in Washington, millions would have perished and both cities would have been devastated
- There would have been no recovery of those cities for many decades

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach knockout capability***

© 1998 T.E. BEARDEN

***Per OTA Report to Congress, 1993**

The U.S. is terribly unprepared

- Limited medical supplies, antibiotics, vaccines
- Insufficient facilities, emergency teams
- No civilian shelters (much less BW-filtered)
- Not enough doctors, nurses, masks, etc.
- Present treatment methods woefully inadequate
- Triage will apply
  - First treat those most likely to recover
  - Others (showing symptoms, or advanced) are just set aside to die
- Other attacks are a matter of *when*, not *if*.

Conclusion

- As several cities are struck, millions of U.S. civilian casualties will result. Devastation will be enormous.
- We shall eventually lose more American lives than in all the rest of our wars to date.
- Present medical therapeutic science is totally inadequate to do the job. It cannot save us.
- At some point, when we are sufficiently damaged, the Spetznaz or other terrorists may detonate nuclear weapons in our cities, with appalling carnage.
- A combined superweapon technology and terrorist groups will then destroy the remaining populace.
- On its present course, the inadequacy of Western medicine may doom us.

A difficult truth to consider

- In any strategic strike, the first phase is to get the strategic weapons (weapons of mass destruction) delivered to their distant targets.
- That has already been done.
- Eerily, the first phase of World War III has already been accomplished.
- It was already accomplished before the Sept. 11, 2001 strike on New York and Washington. That was the new "Pearl Harbor", and only the opening round. It already killed more Americans than died at Pearl Harbor on Dec. 7, 1941.
- Unless we improve our medical therapy quickly and dramatically, we will lose -- particularly when the hidden energetics weapons and outside terrorist sponsorship by hostile nations are factored in.

A personal perspective

- There now exists an overwhelming, immediate moral imperative to get this development program done at the utmost speed humanly possible.
- The sheer survival of the United States and Western civilization depends upon it.
- Officialdom is aware that the KGB/Russians have long planned to finish the U.S., after initial nuclear exchanges, by fierce and massive BW strikes.
- It appears that the plan has changed to let the terrorists do the main work, while augmenting U.S. casualty levels by energetics weapons.

China is also involved

- China has declared South China Sea the territorial waters of China
- China will eventually take Taiwan
  - Has massed missile artillery trained on Taiwan
  - Recently held invasion practice on an island near Taiwan
- Recently deployed fearsome EMP weapons utilizing negative energy rather than positive energy (Dirac sea theory)
  - Extinguishes all electron currents
  - Duds electrical and electron systems instantly
  - Instant death to struck bodies
  - Instant kill of missiles, aircraft, electronics in nuclear warheads
  - Ultimate "death ray"; can be made small or large
- Recent 2-carrier U.S. task force maneuvered in South China Sea to caution China
  - The new Chinese EMP weapon was trained on that task force
  - Could have destroyed it within minutes

Final thoughts

← ??? →

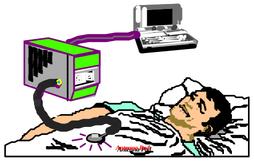

- This is one of the great turning points in history.
- The development program is eminently doable. That it works has been experimentally demonstrated. The long-vexing puzzle of the active mechanism has been solved.
- Our scientific community bitterly opposes any use of electromagnetics -- even an advanced, higher symmetry electrodynamics -- for medical treatment.
- A consortium of the scientific community, the large pharmaceuticals, FDA, and the AMA will oppose this project even while the U.S. is being destroyed wholesale.
- It can only be done in the private sector, and there will likely be no funding unless by philanthropists.

Continuation

Business plan, development program, budget breakdown, proposed schedule etc. go here.

Weapons Slides

The Tom Bearden Website

Help support the research

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

© 1995, 1996, 1999 T.E. BEARDEN

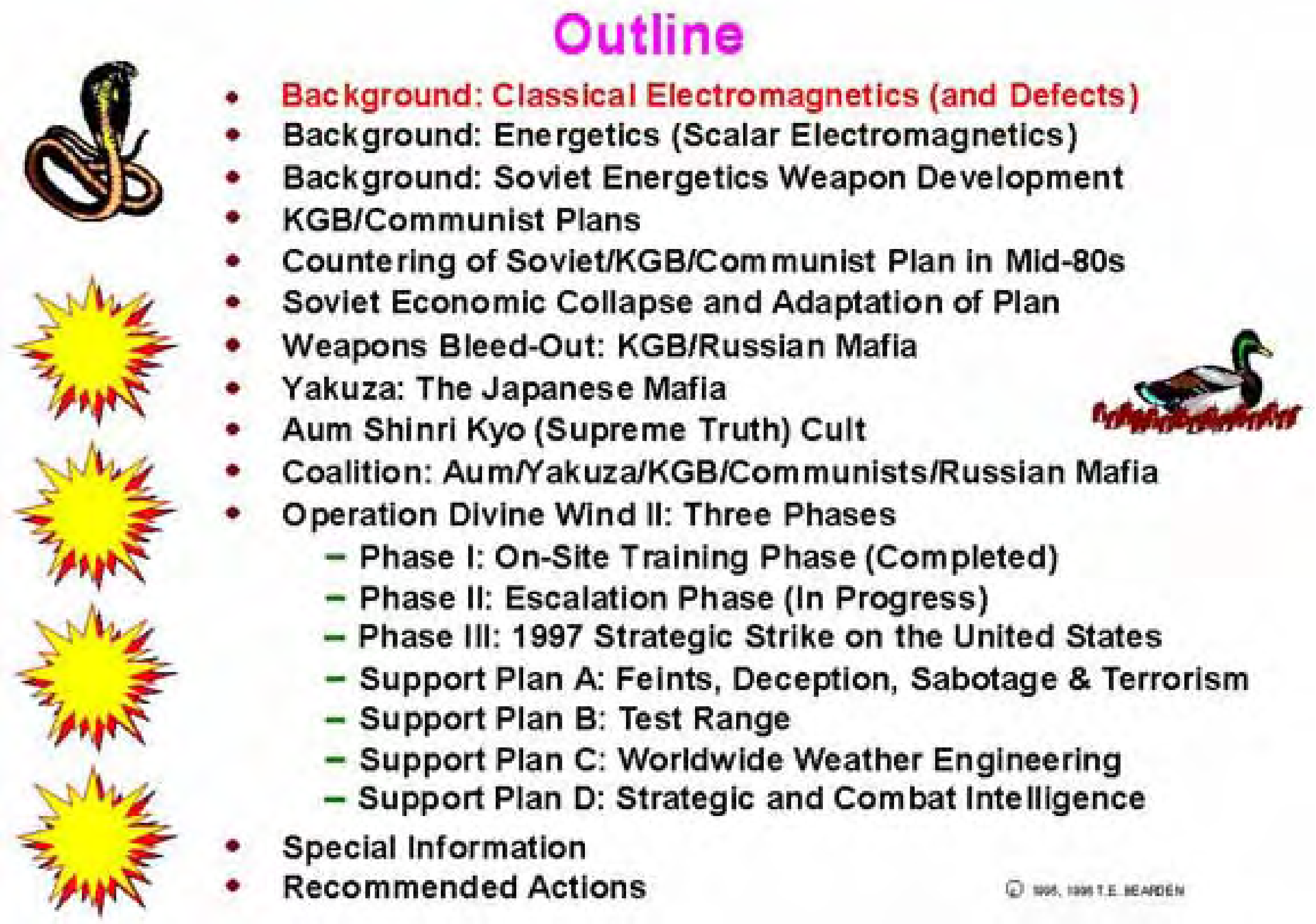

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© 1996, 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

A-10 Mysterious Flyaway and Crash

April 2, 1997

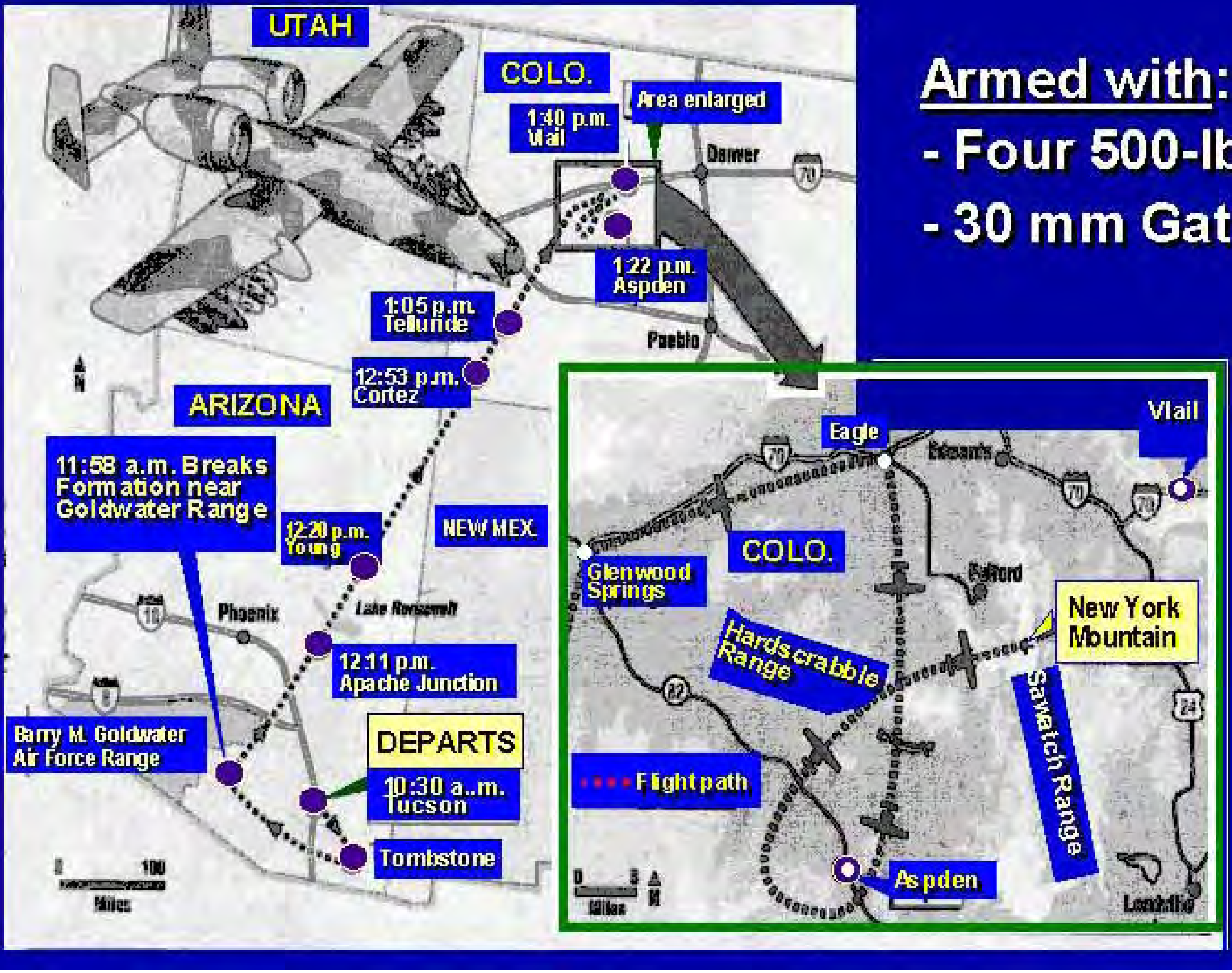
UTAH
COLO.
Area enlarged
1:40 p.m.
Vail
Denver
70
1:22 p.m.
Aspden
1:05 p.m.
Telluride
Pueblo
12:53 p.m.
Cortez
ARIZONA
11:58 a.m. Breaks
Formation near
Goldwater Range
12:20 p.m.
Young
NEW MEX.
10
Phoenix
Lake Roosevelt
12:11 p.m.
Apache Junction
8
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a..m.
Tucson
Tombstone
0 100
Miles
N
Eagle
Edwards
Vlail
70
COLO.
Glenwood
Springs
Bellford
New York
Mountain
Hardscrabble
Range
82
Sawatch Range
24
Flight path
Aspden
0 5
Miles
N
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

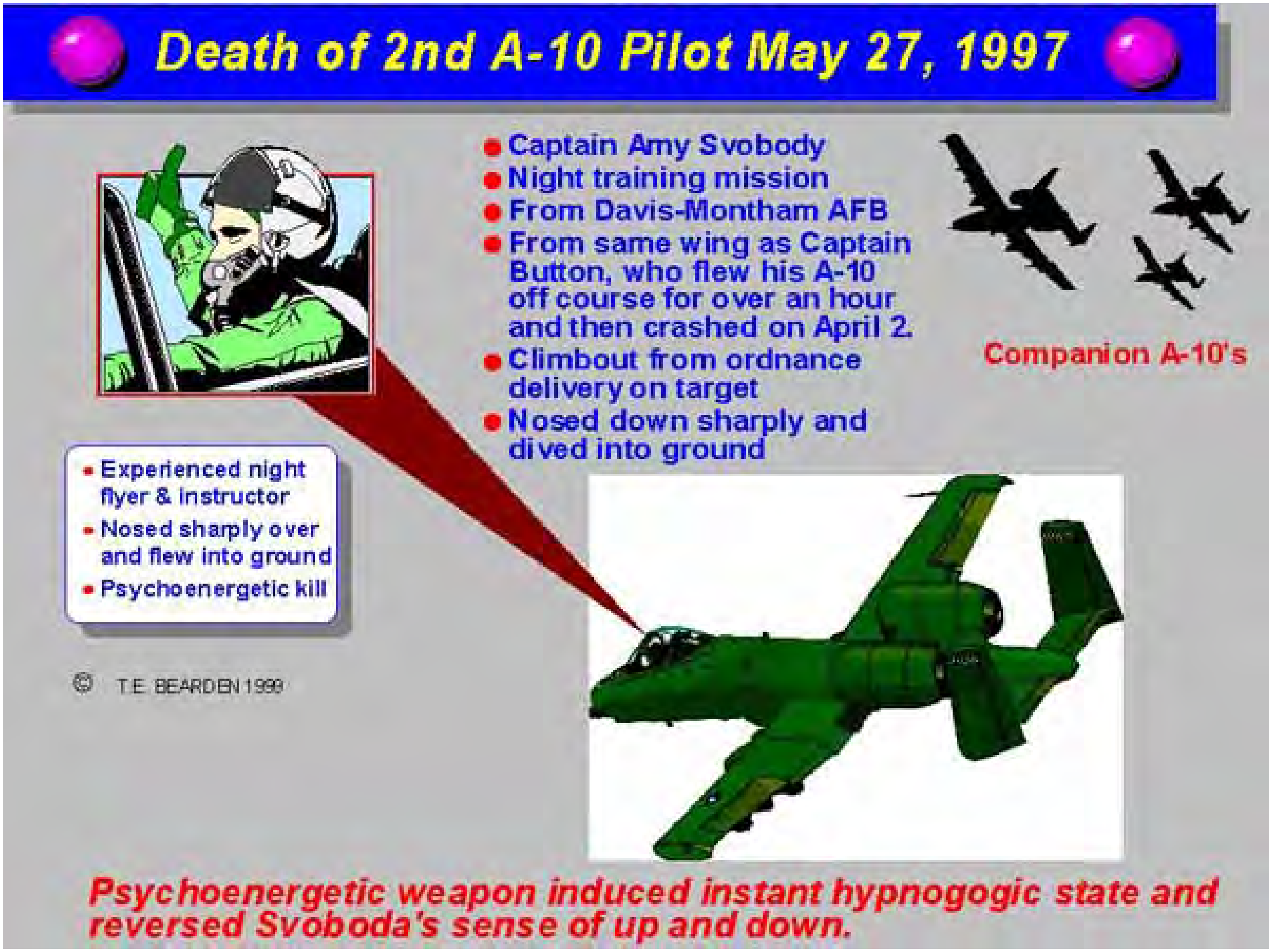

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

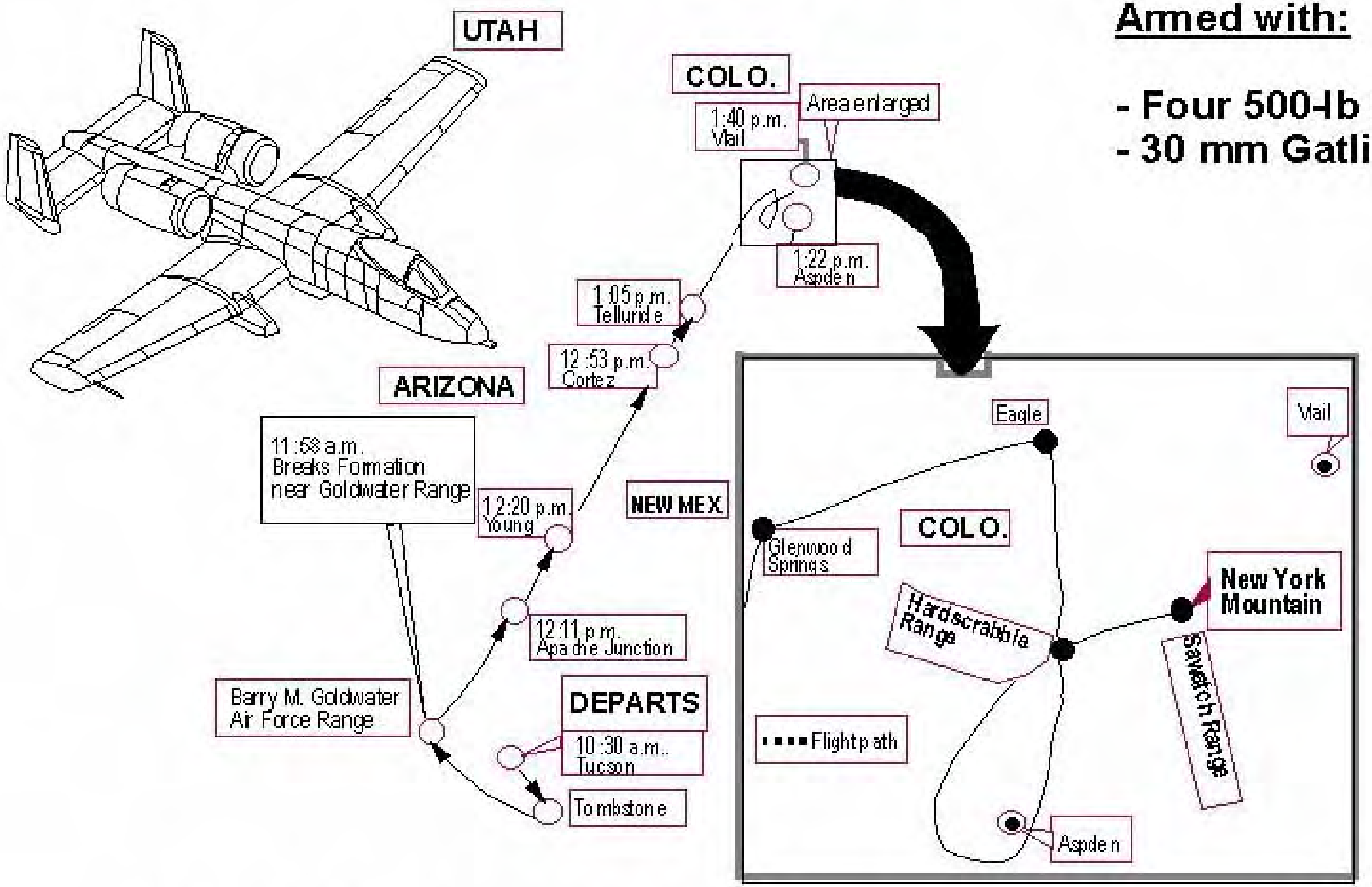
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
1:22 p.m.
Aspden
1:05 p.m.
Telluride
12:53 p.m.
Cortez
ARIZONA
11:58 a.m.
Breaks Formation
near Goldwater Range
12:20 p.m.
Young
NEW MEX
12:11 p.m.
Apache Junction
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m..
Tucson
Tombstone
Eagle
Vail
Glenwood
Springs
COLO.
Hardscrabble
Range
New York
Mountain
Sawatch Range
Flight path
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

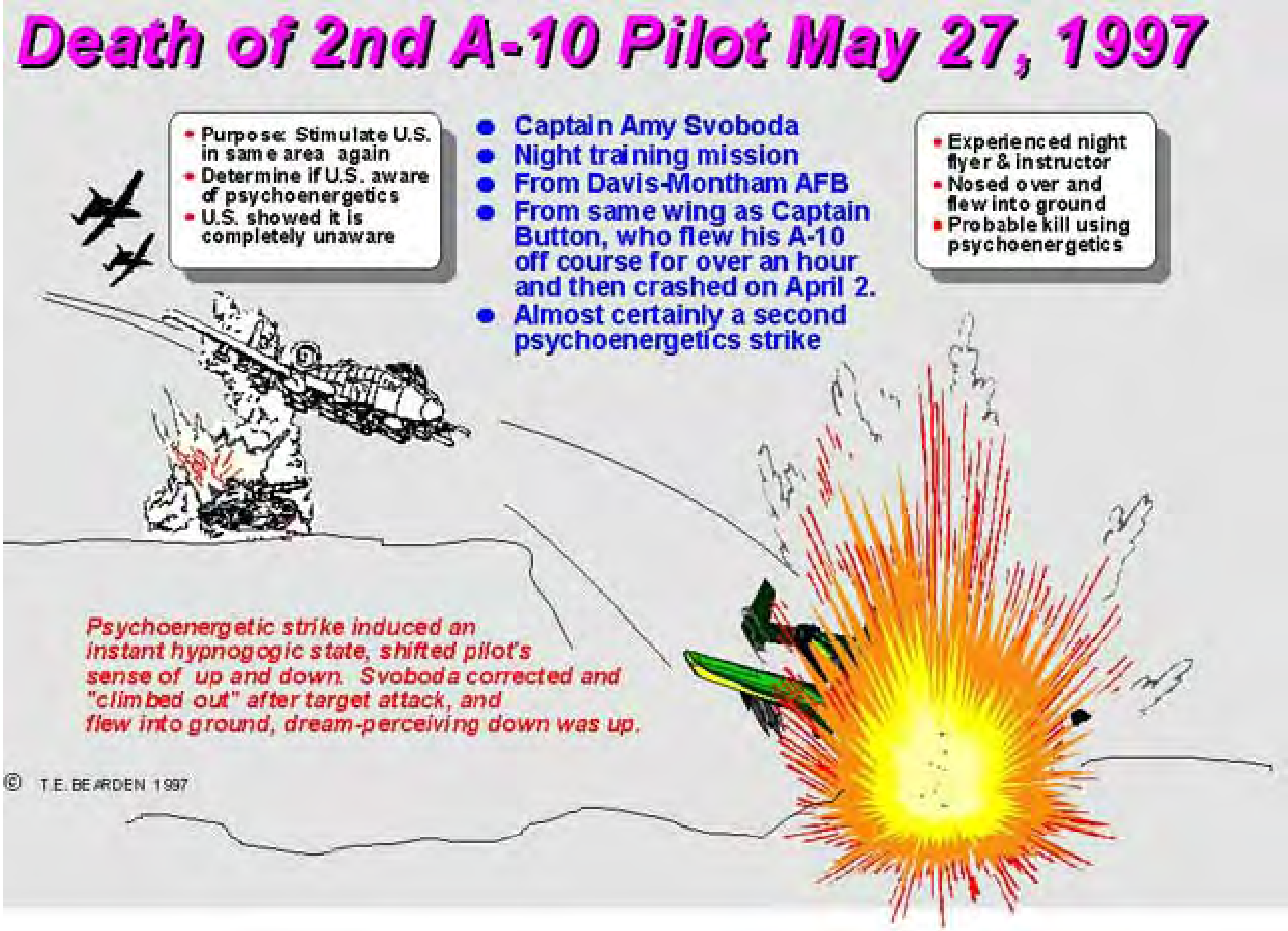

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

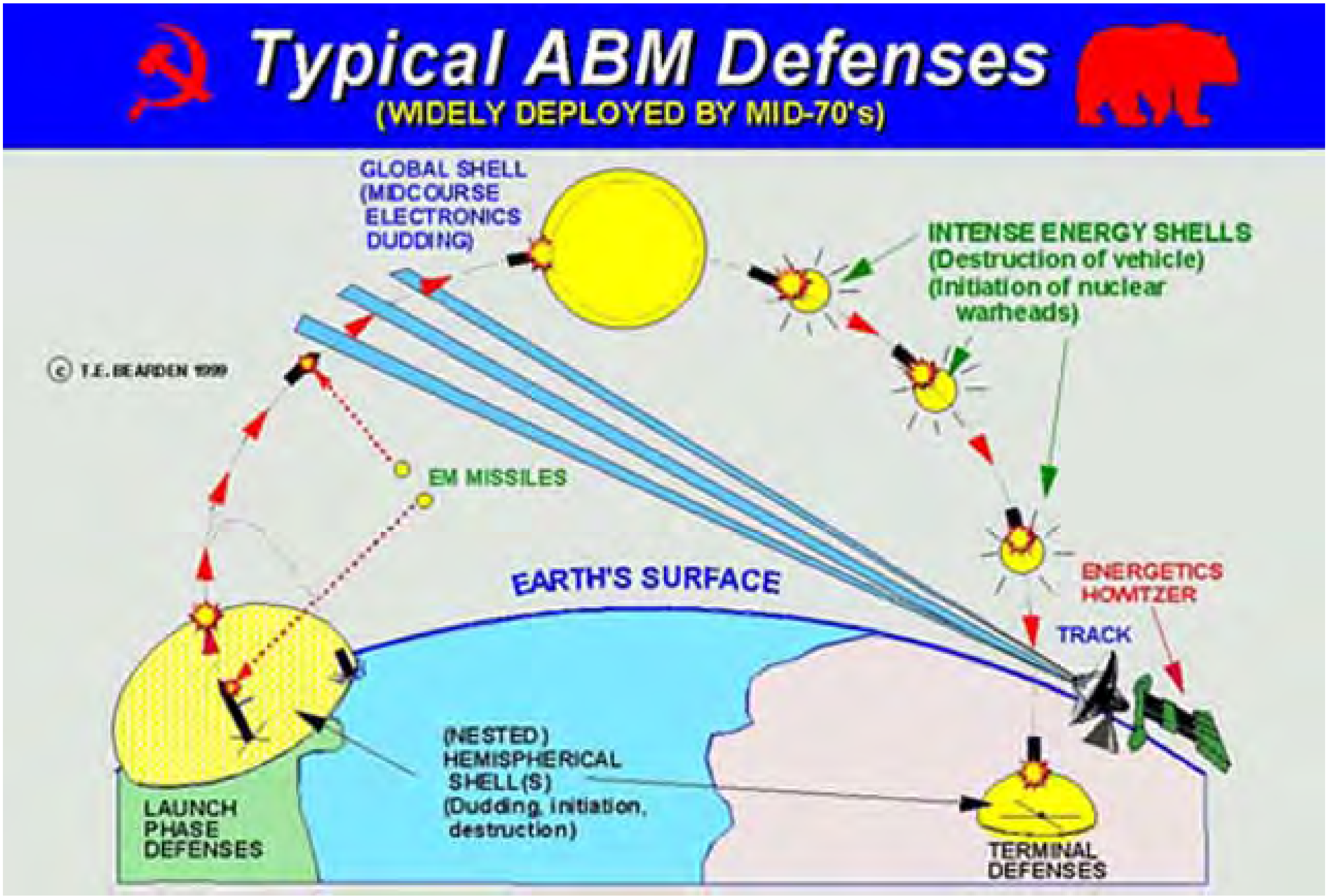
Typical ABM Defenses
(WIDELY DEPLOYED BY MID-70's)
GLOBAL SHELL (MIDCOURSE ELECTRONICS DUDDING)
INTENSE ENERGY SHELLS (Destruction of vehicle) (Initiation of nuclear warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS HOWITZER
TRACK
LAUNCH PHASE DEFENSES
(NESTED) HEMISPHERICAL SHELL(S) (Dudding, initiation, destruction)
TERMINAL DEFENSES

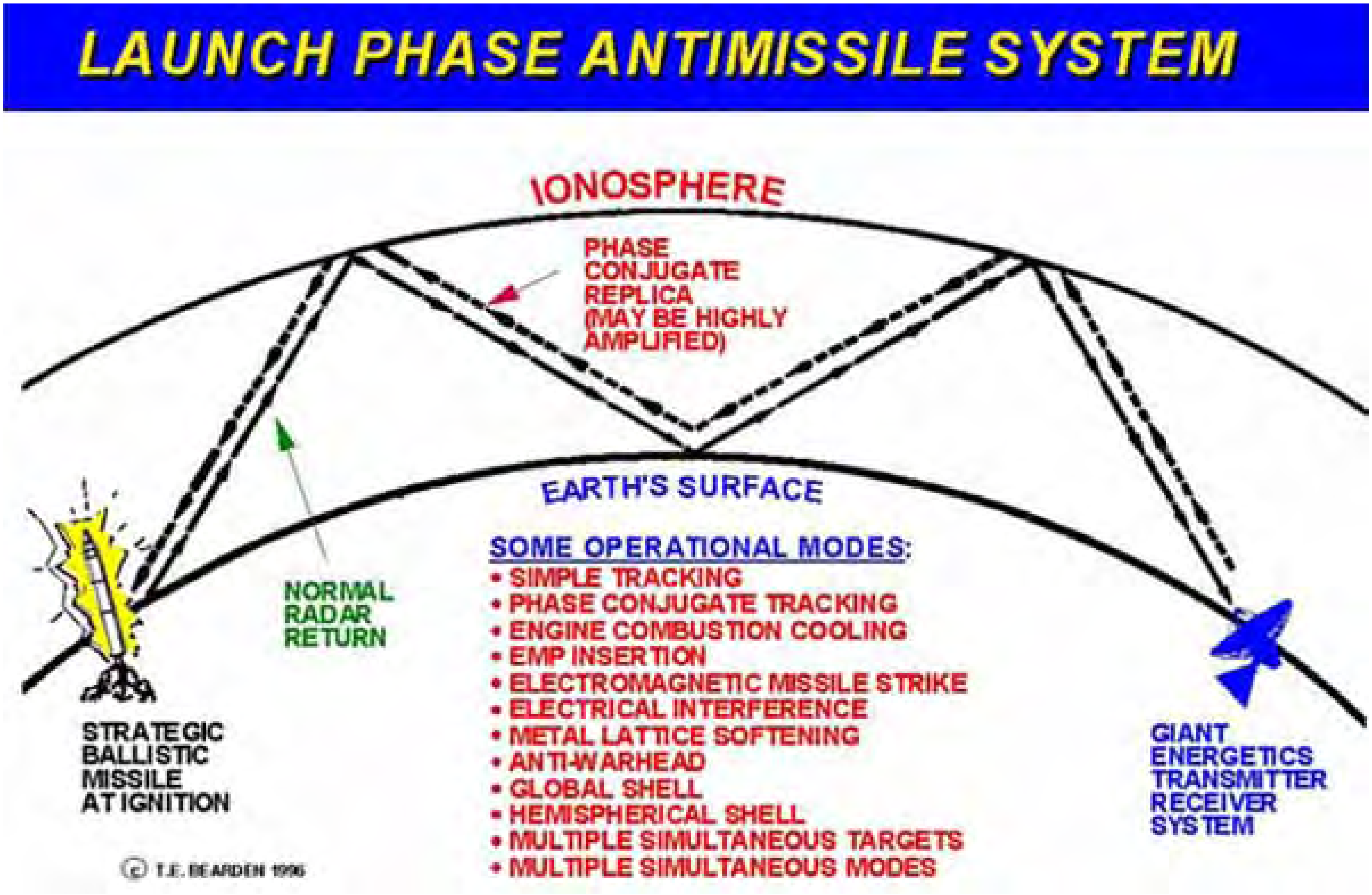

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE

PHASE
CONJUGATE
REPLICA
(MAY BE HIGHLY
AMPLIFIED)

EARTH'S SURFACE

NORMAL
RADAR
RETURN

STRATEGIC
BALLISTIC
MISSILE
AT IGNITION

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

GIANT
ENERGETICS
TRANSMITTER
RECEIVER
SYSTEM

ⓒ T.E. BEARDEN 1996

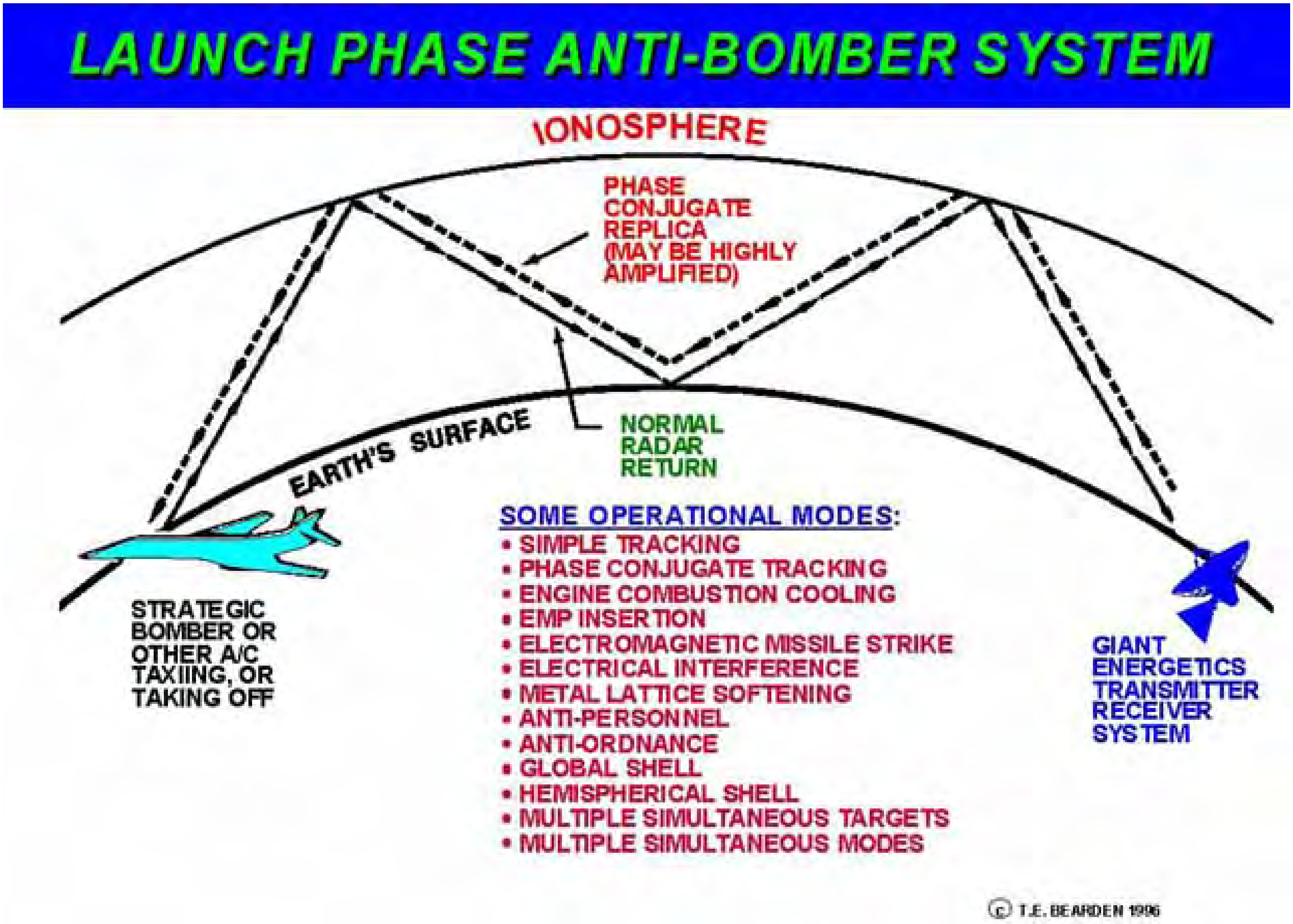

LAUNCH PHASE ANTI-BOMBER SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
NORMAL RADAR RETURN
EARTH'S SURFACE
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

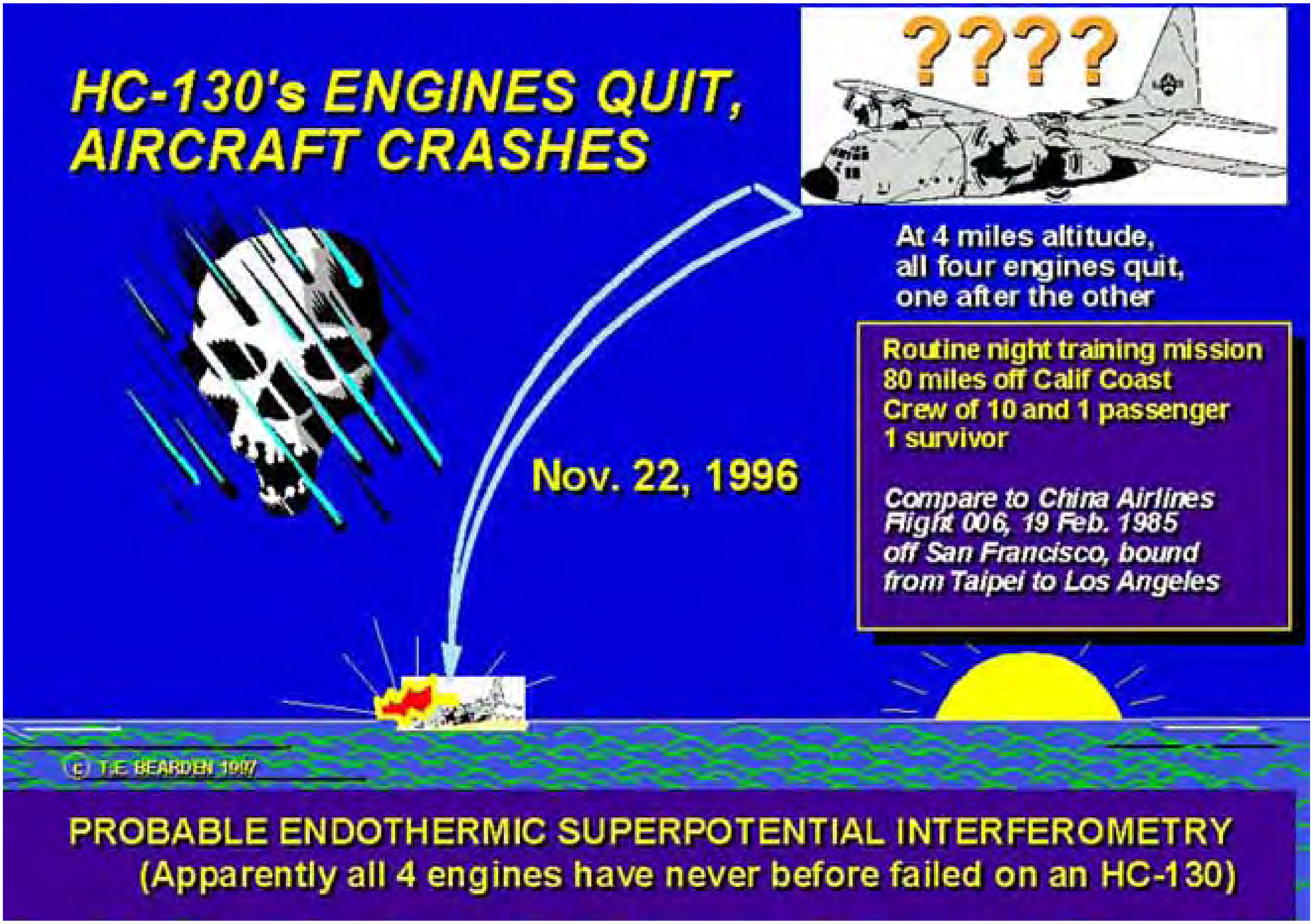

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

Nov. 22, 1996

(c) T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

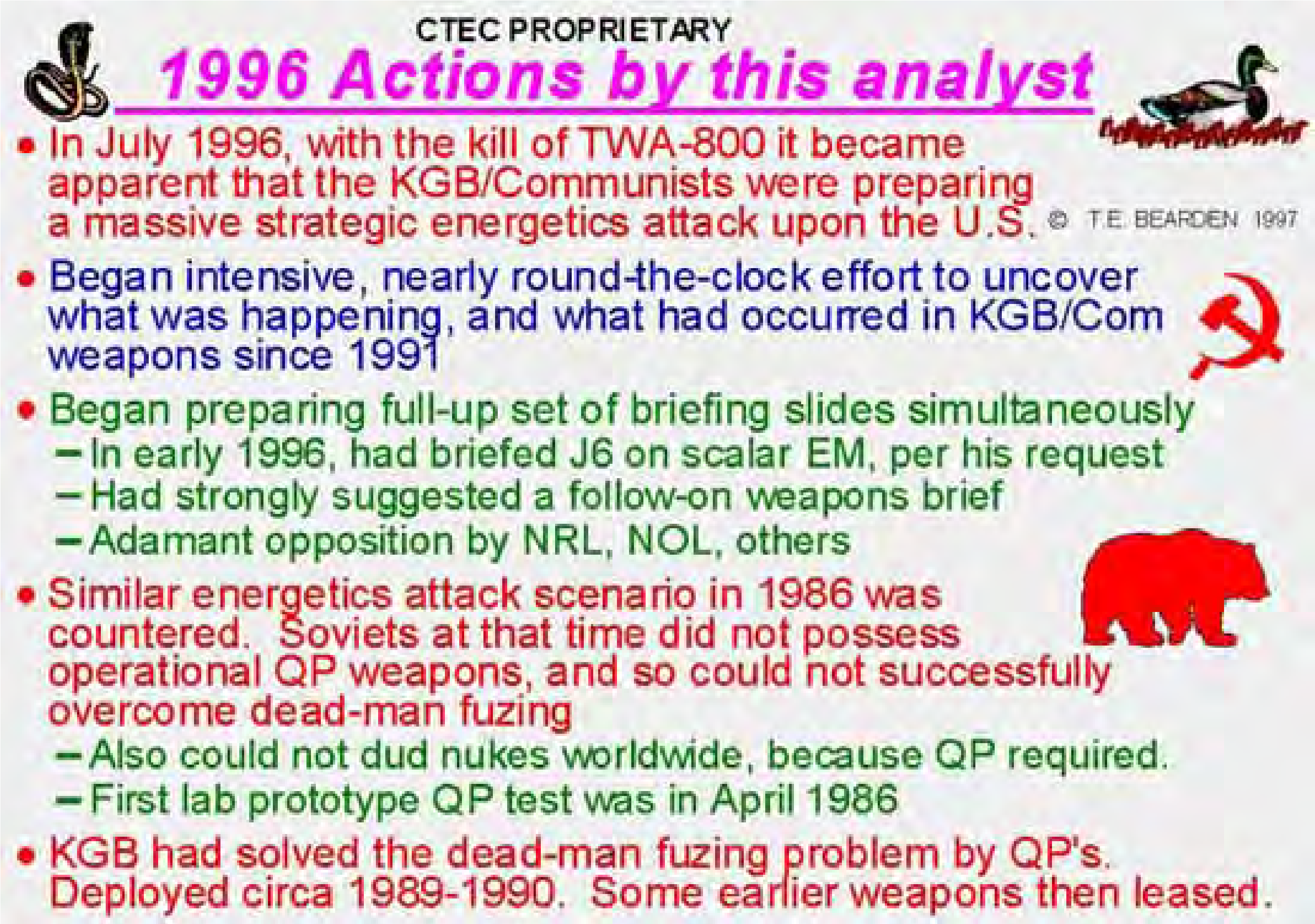

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

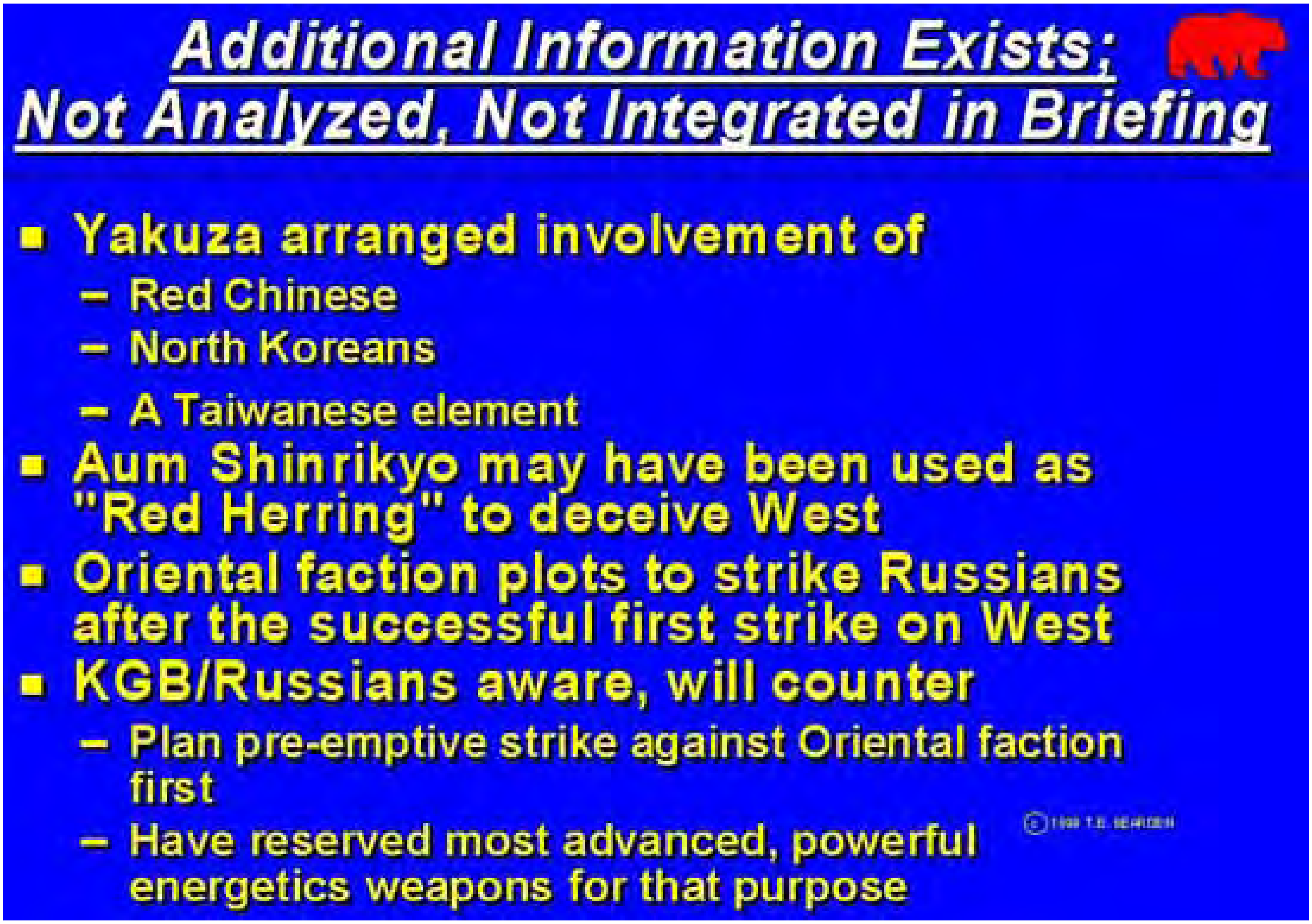

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

The Tom Bearden Website

Help support the research

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

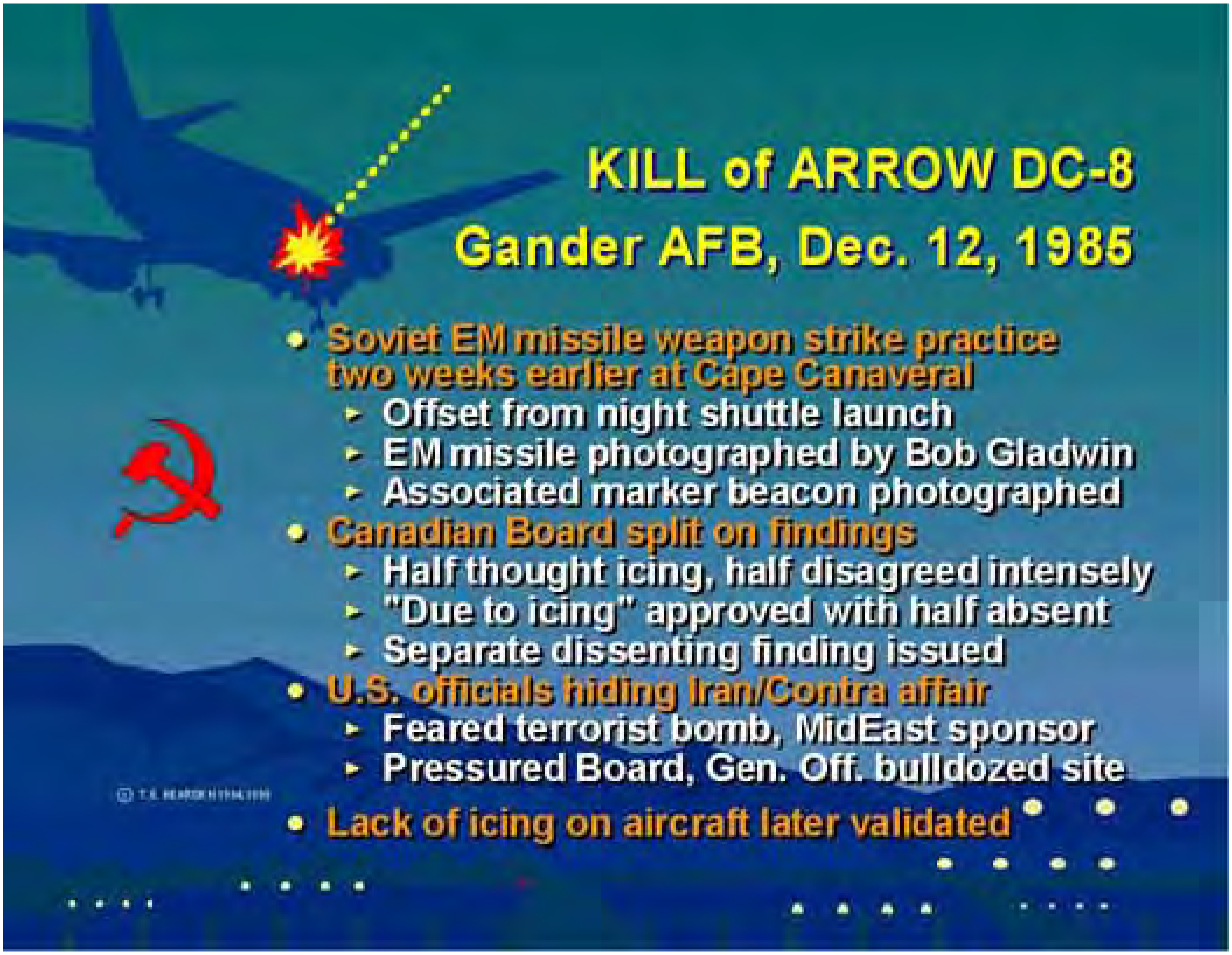

KILL of ARROW DC-8

Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
    - Offset from night shuttle launch
    - EM missile photographed by Bob Gladwin
    - Associated marker beacon photographed
- Canadian Board split on findings
    - Half thought icing, half disagreed intensely
    - "Due to icing" approved with half absent
    - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
    - Feared terrorist bomb, MidEast sponsor
    - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

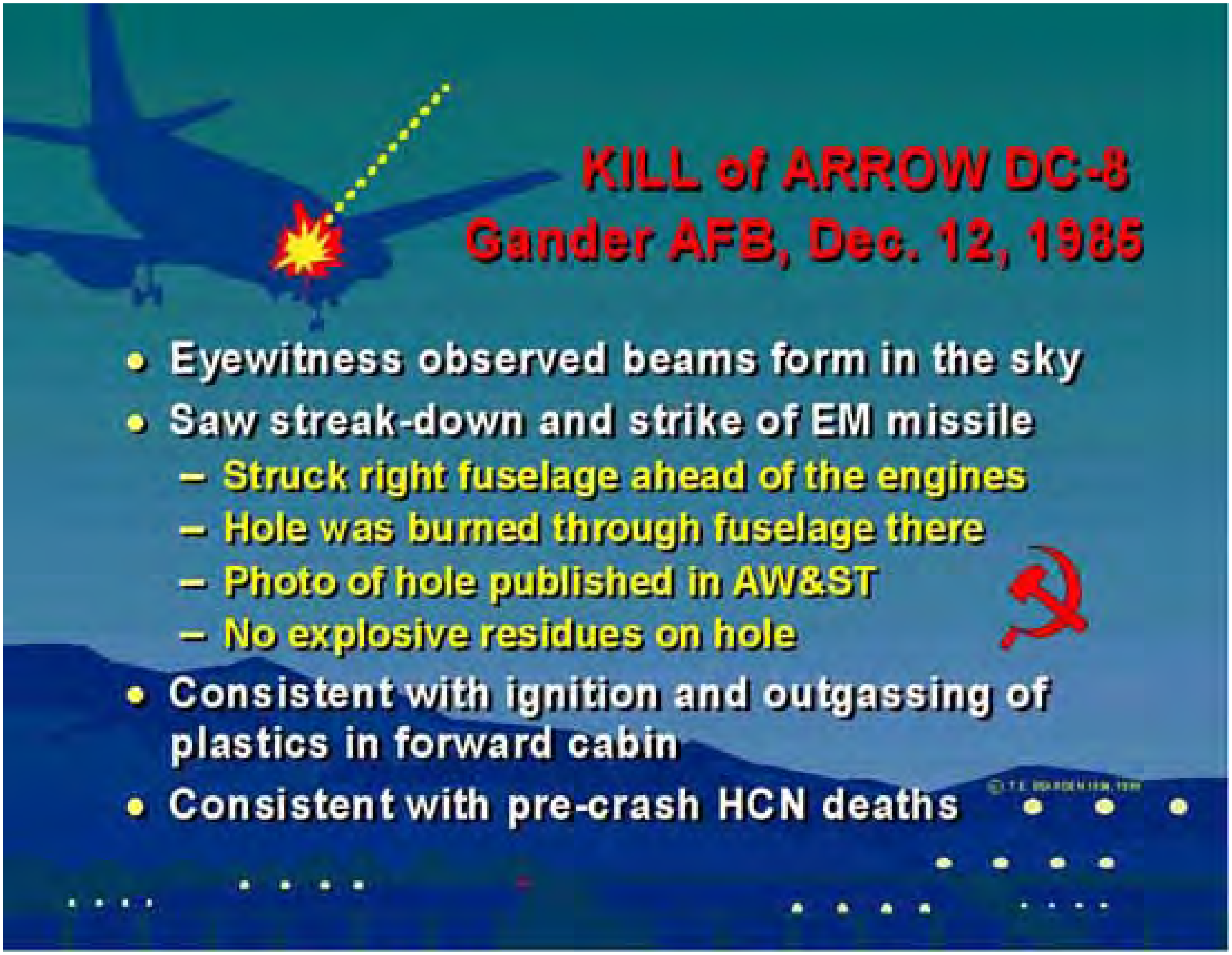

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

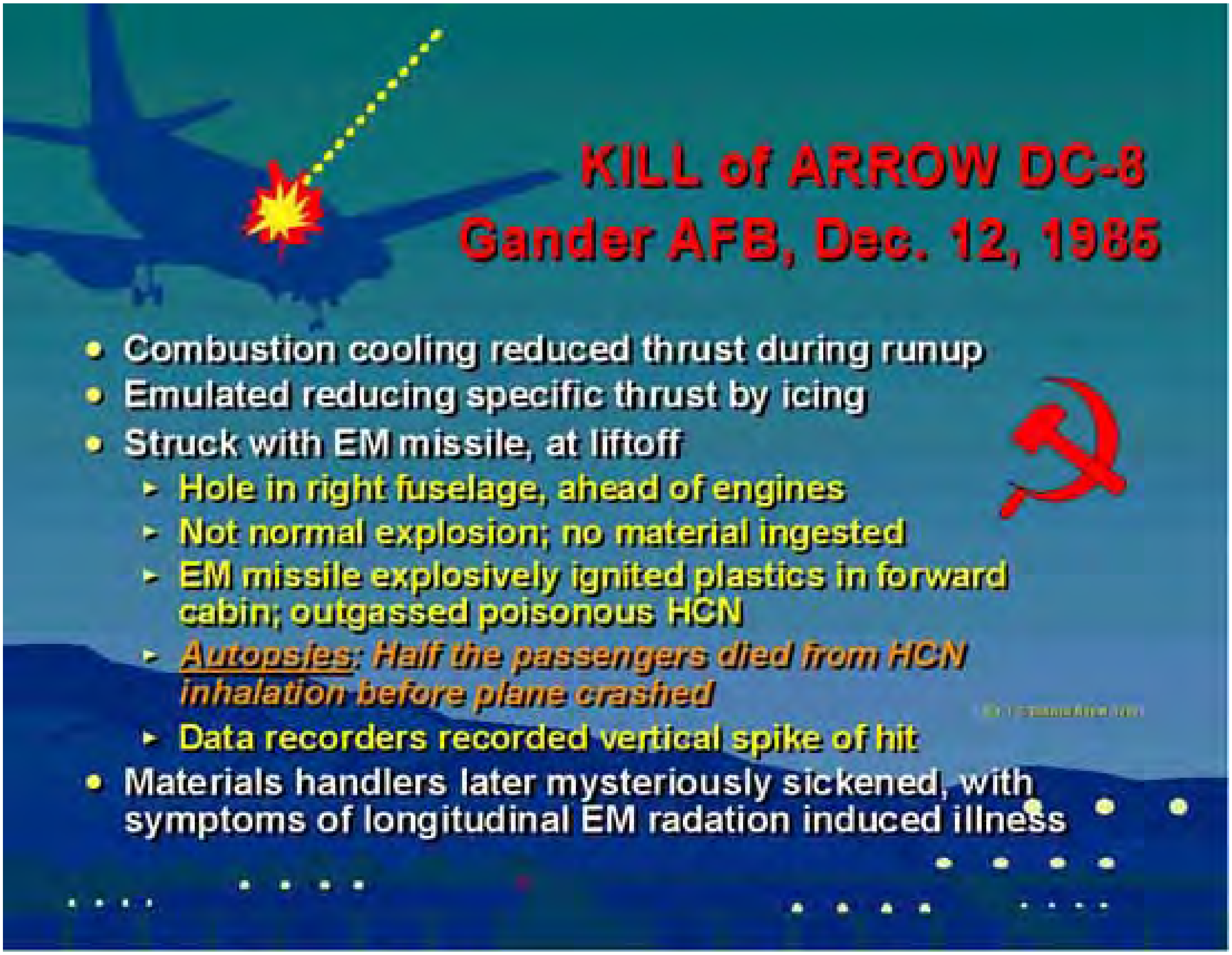

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

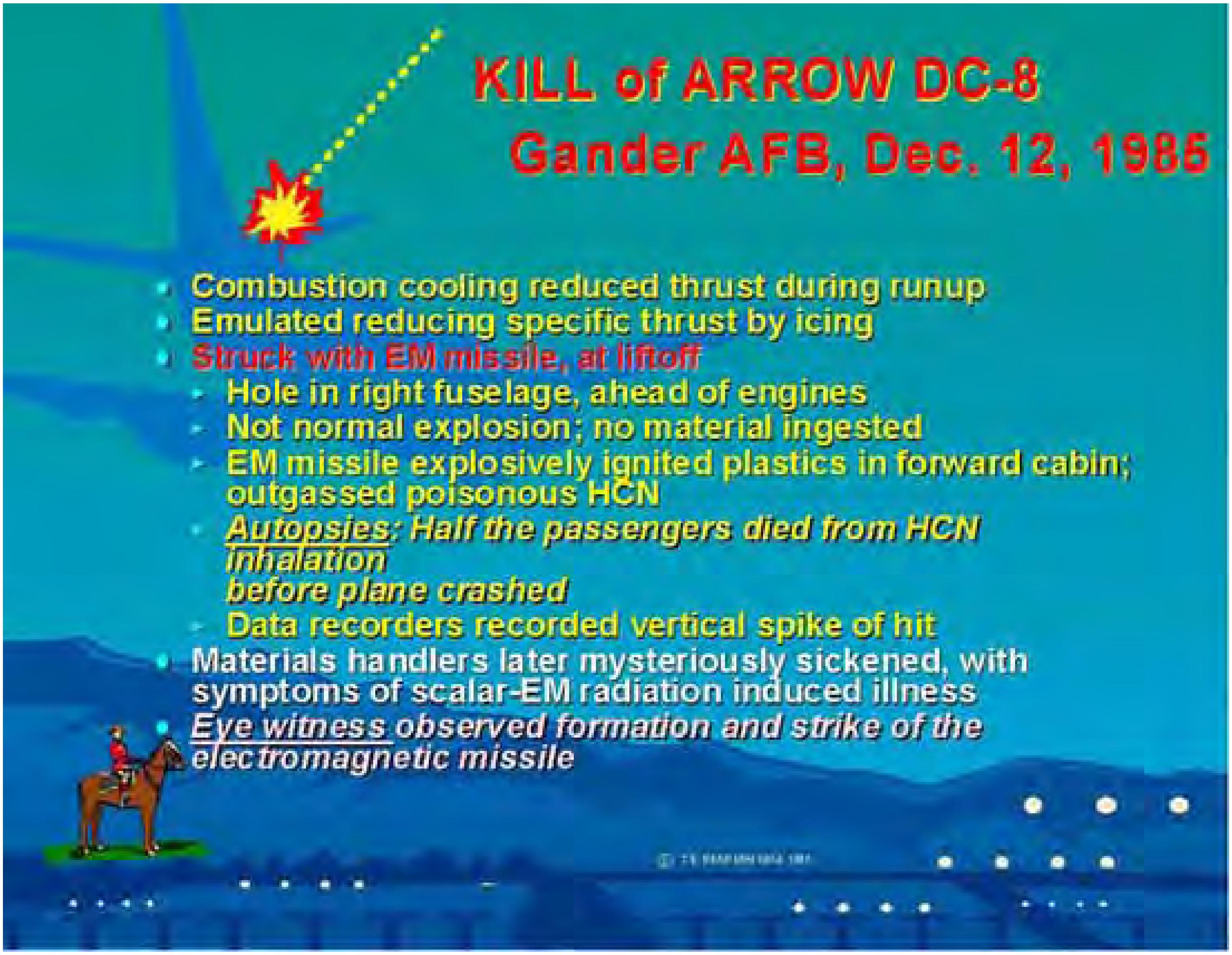

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation*
    before plane crashed
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

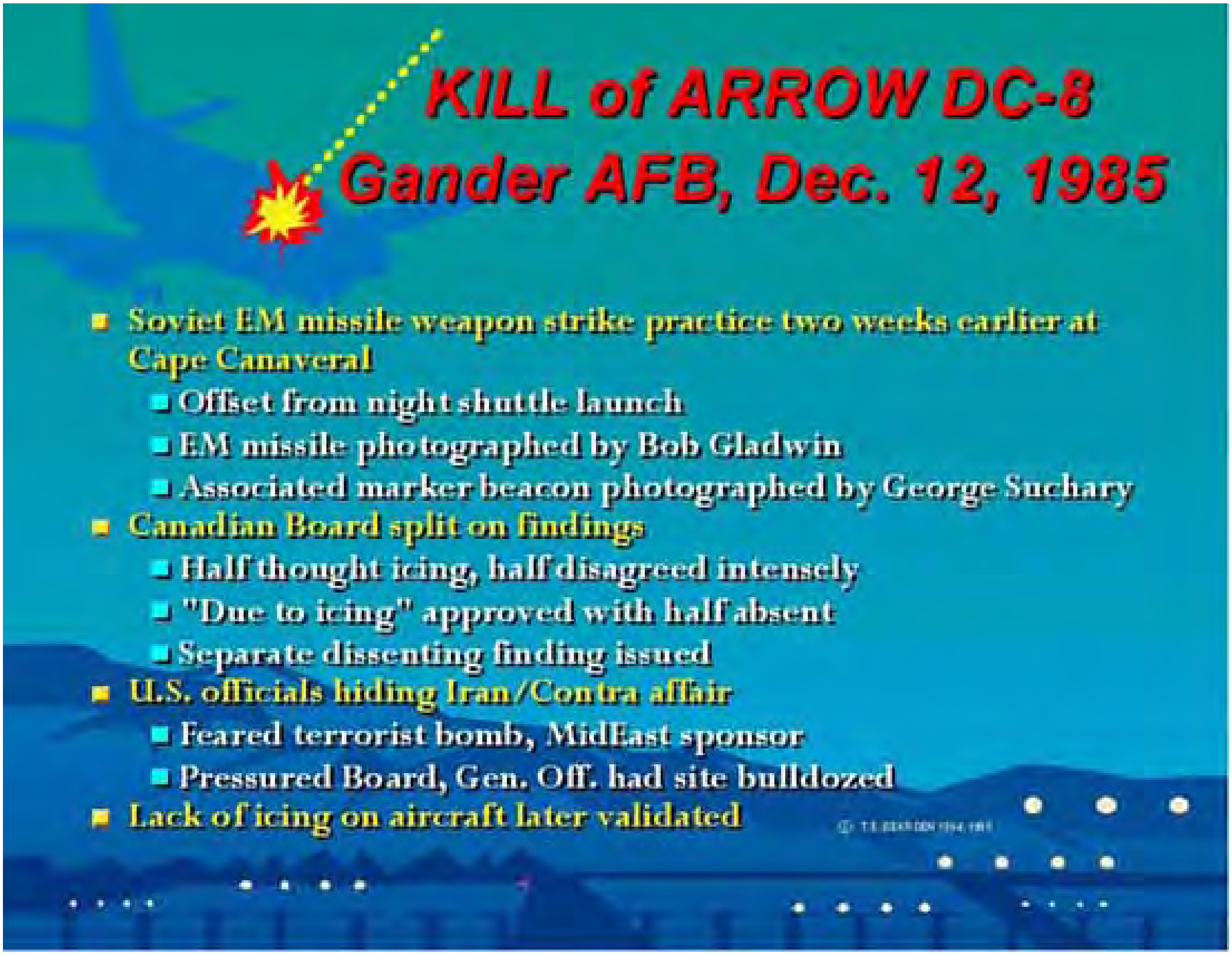

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

Death of the Arrow DC-8

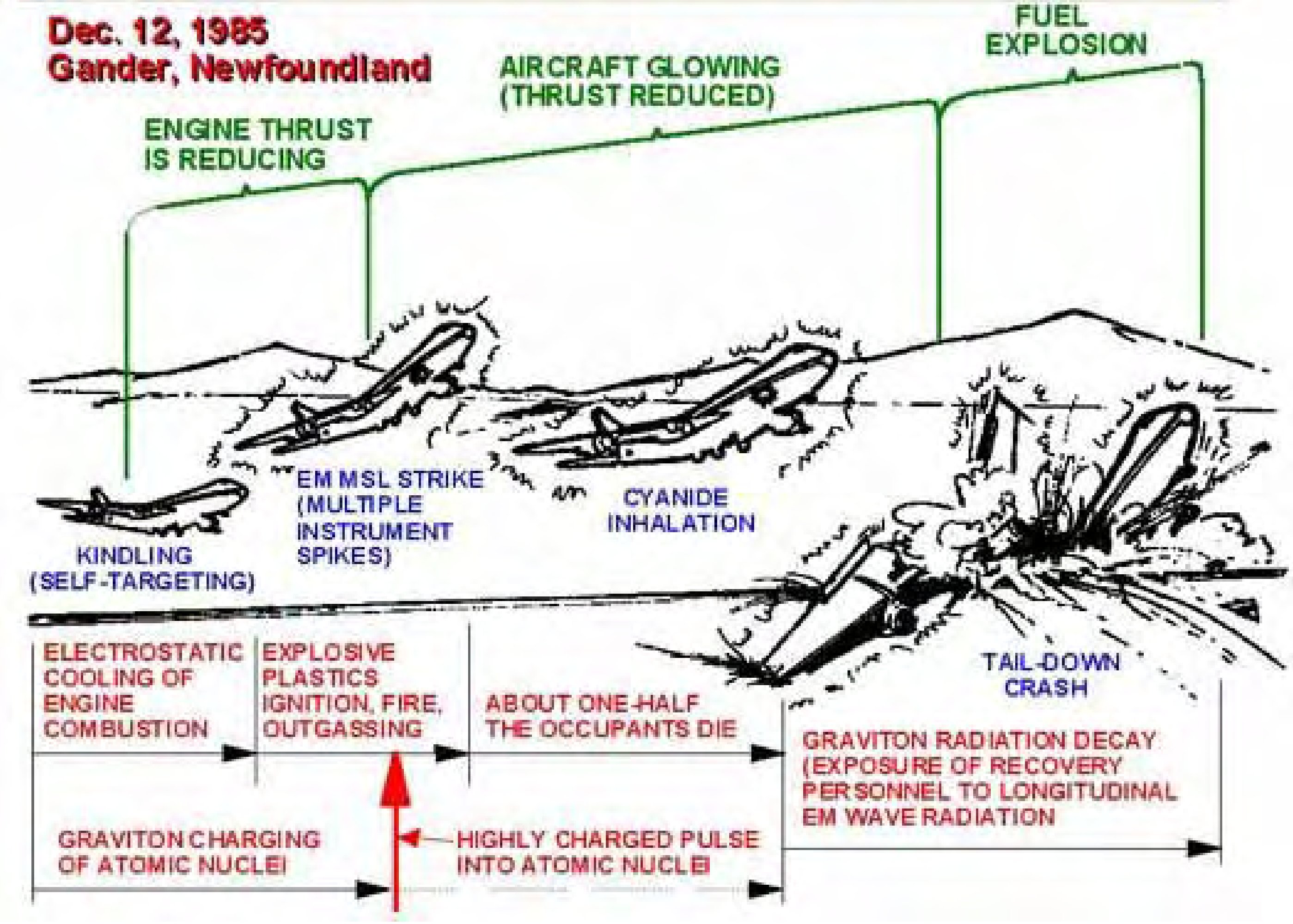
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST IS REDUCING
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
KINDLING (SELF-TARGETING)
EM MSL STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT ONE-HALF THE OCCUPANTS DIE
GRAVITON RADIATION DECAY (EXPOSURE OF RECOVERY PERSONNEL TO LONGITUDINAL EM WAVE RADIATION
GRAVITON CHARGING OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE INTO ATOMIC NUCLEI

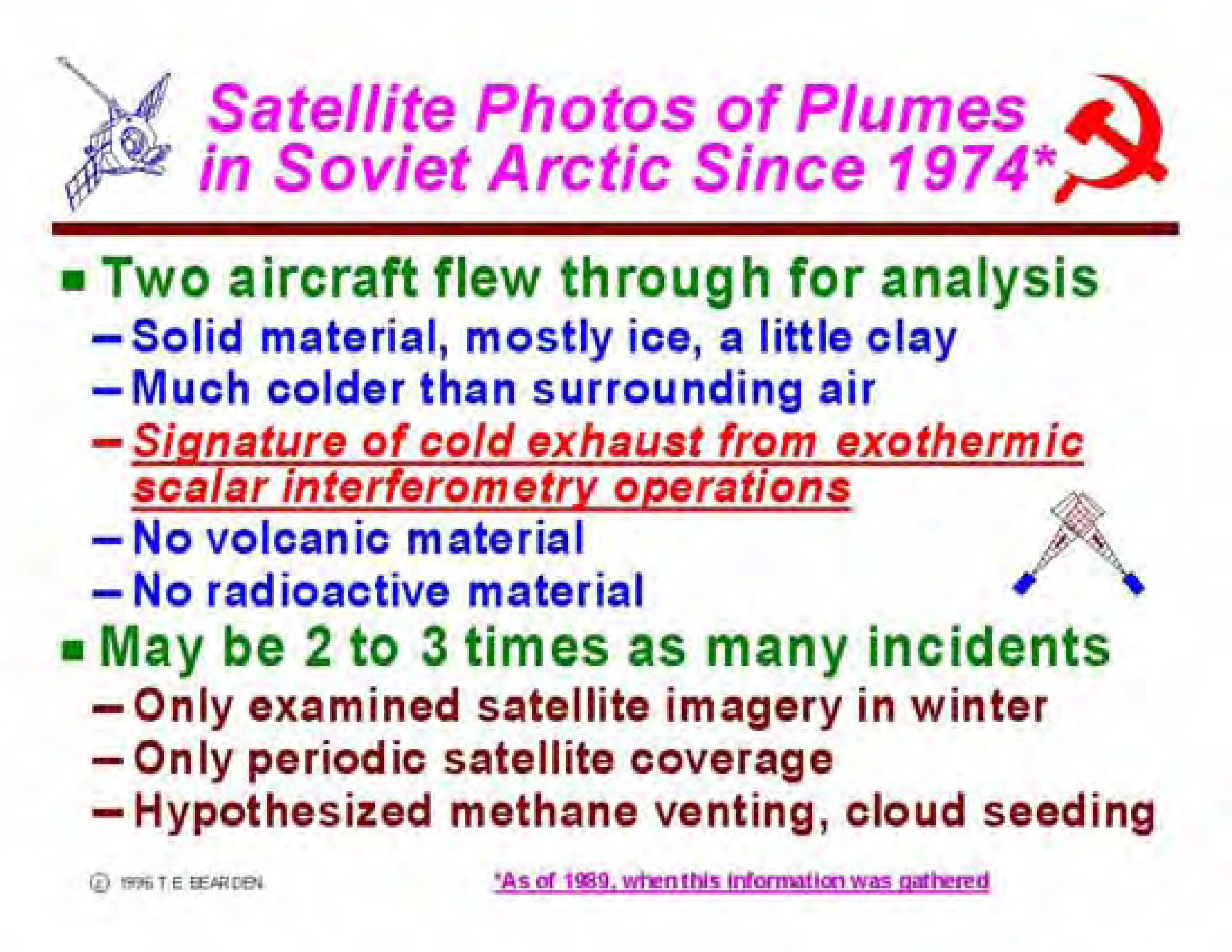

Satellite Photos of Plumes in Soviet Arctic Since 1974*

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO

("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

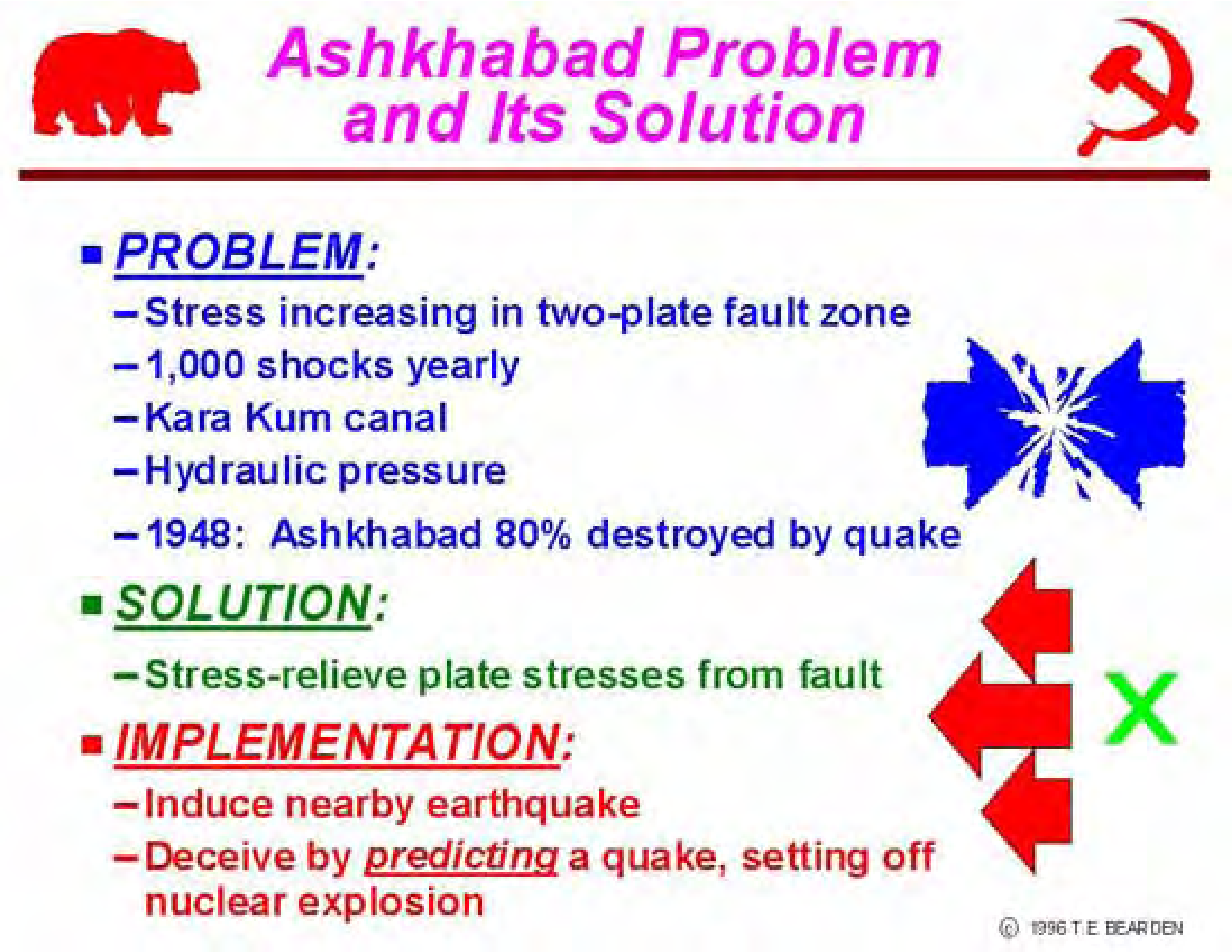

Ashkhabad Problem and Its Solution

- ***PROBLEM:***
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- ***SOLUTION:***
  - Stress-relieve plate stresses from fault
- ***IMPLEMENTATION:***
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

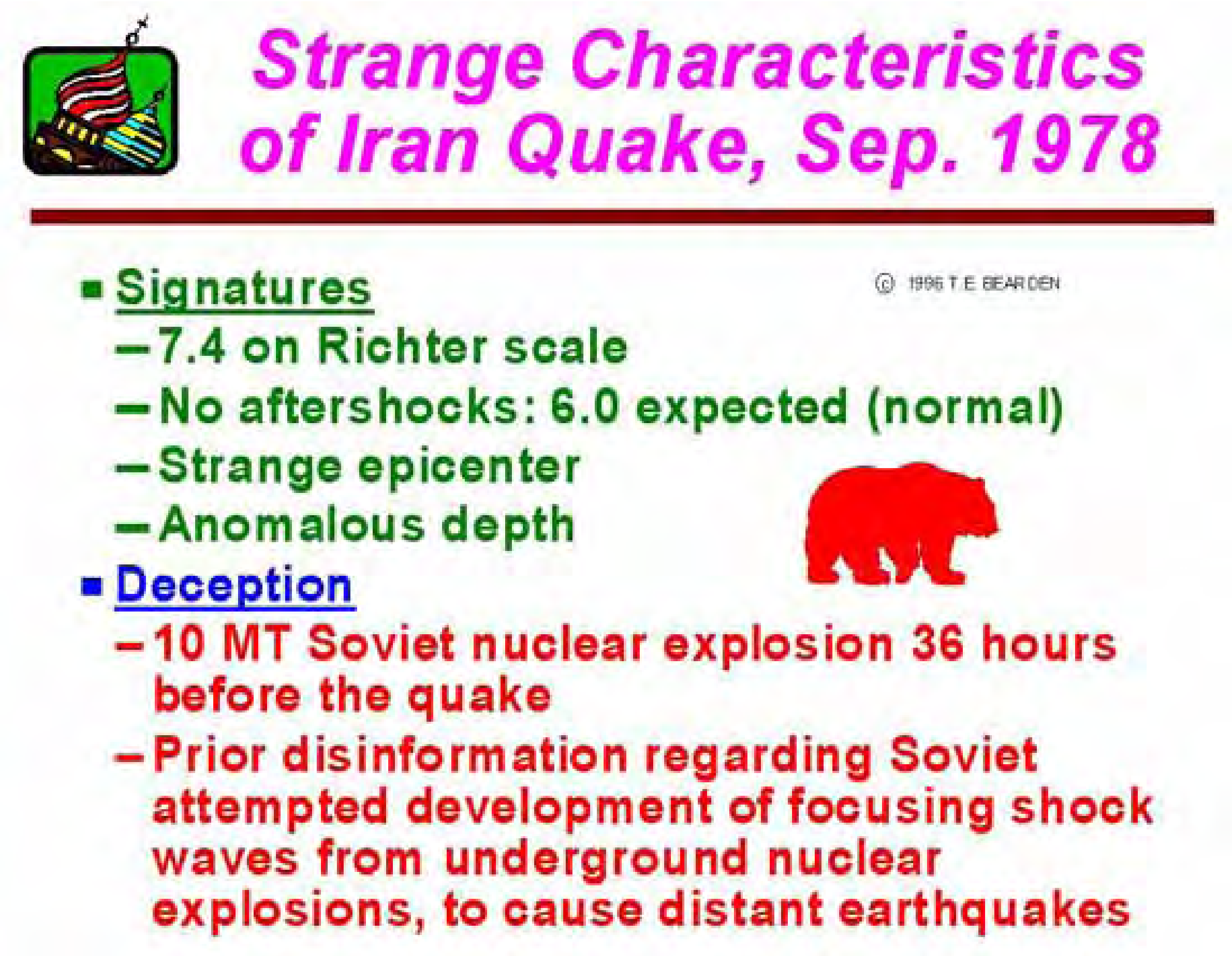

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

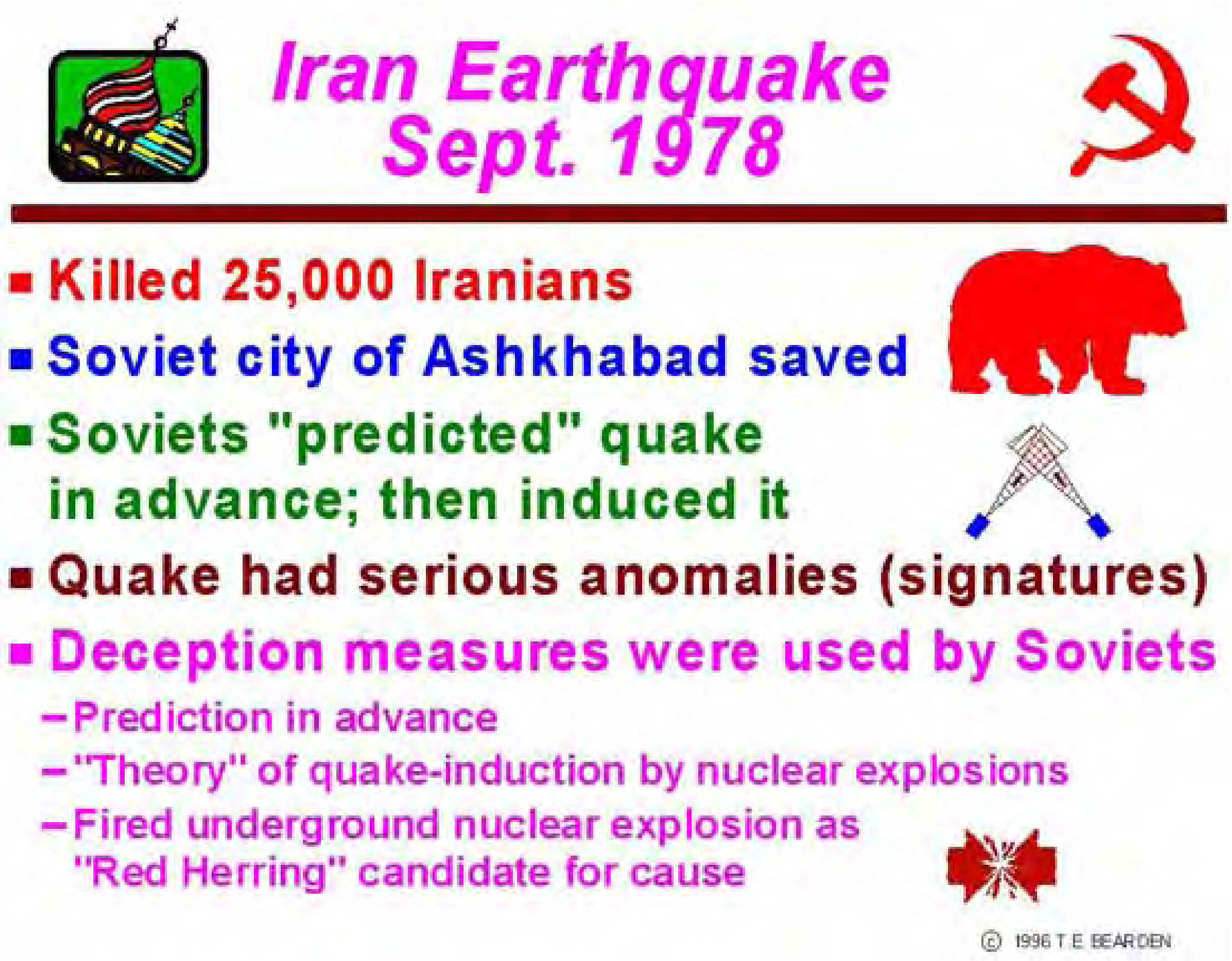

Iran Earthquake
Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - Prediction in advance
  - "Theory" of quake-induction by nuclear explosions
  - Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T.E. BEARDEN

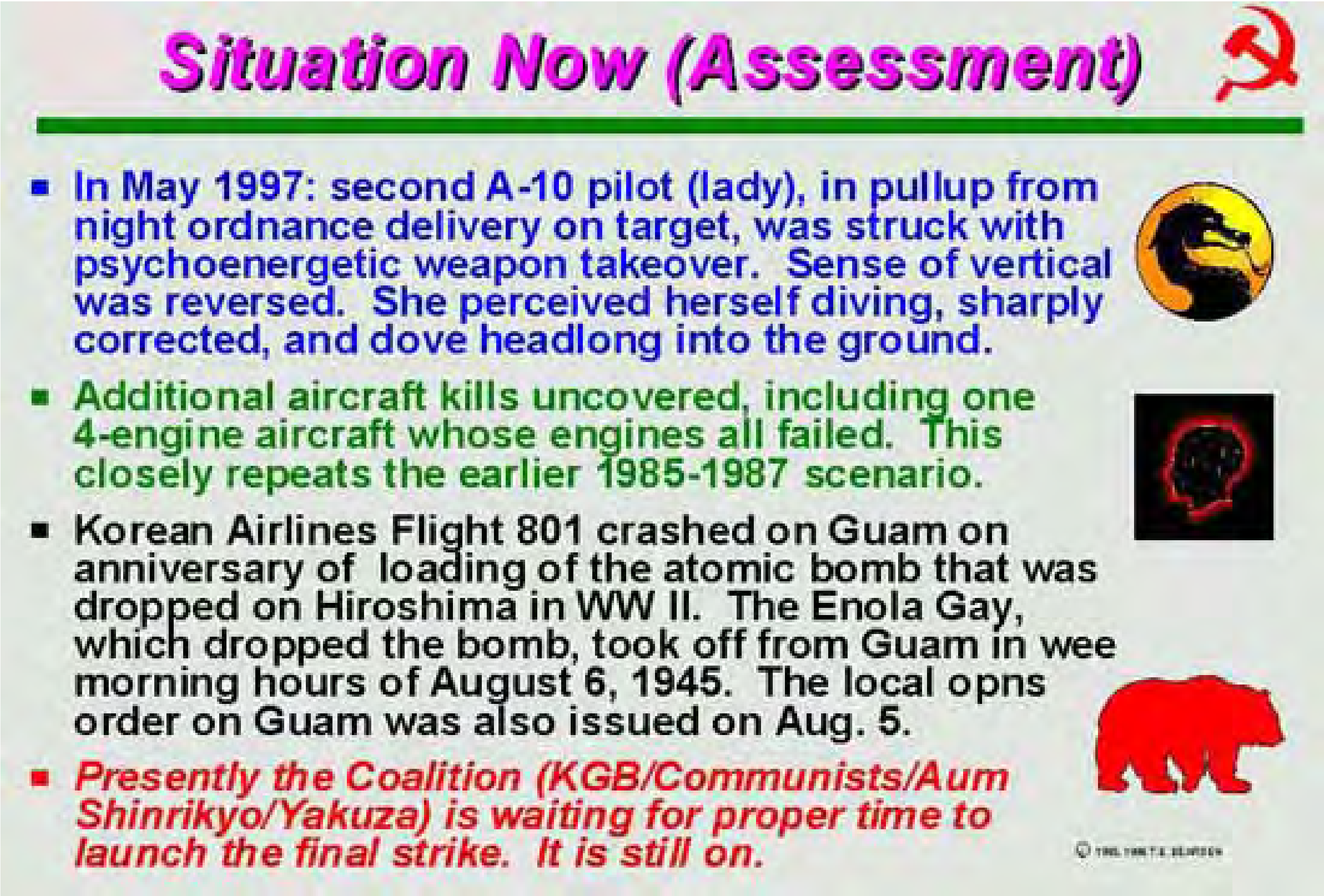

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

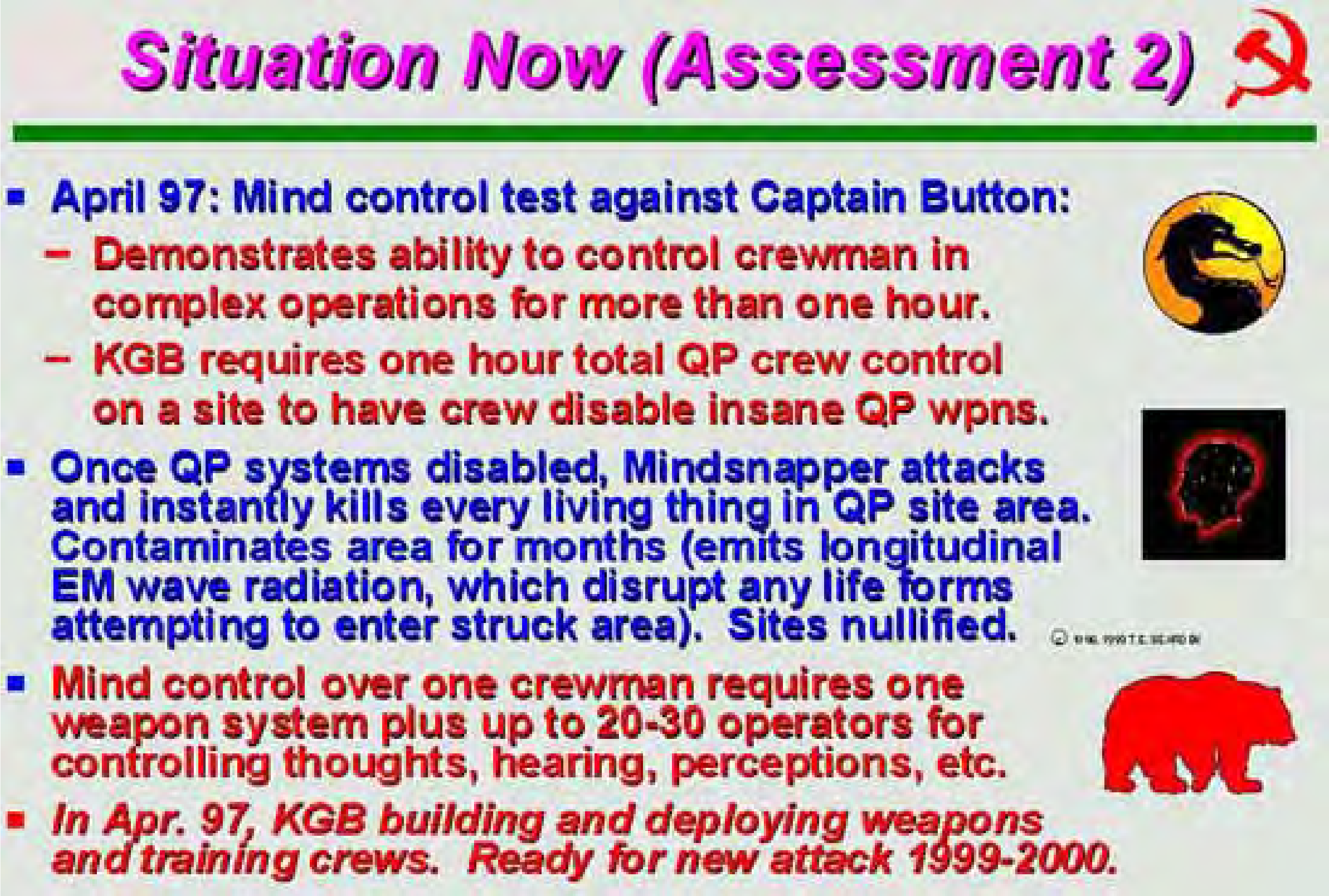

Situation Now (Assessment 2)

- April 97: Mind control test against Captain Button:
  - Demonstrates ability to control crewman in complex operations for more than one hour.
  - KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.
- Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.
- Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.
- *In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.*

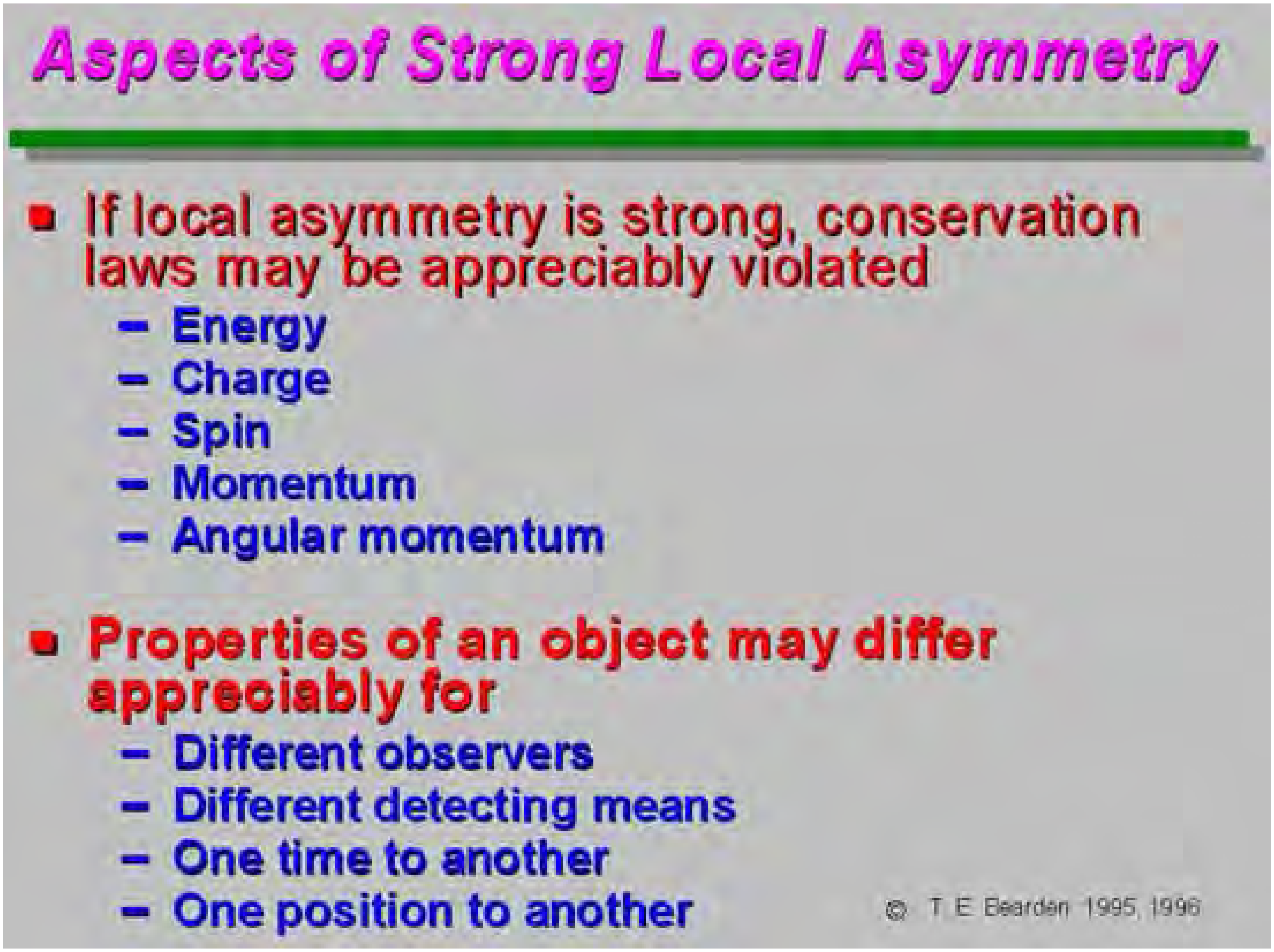

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

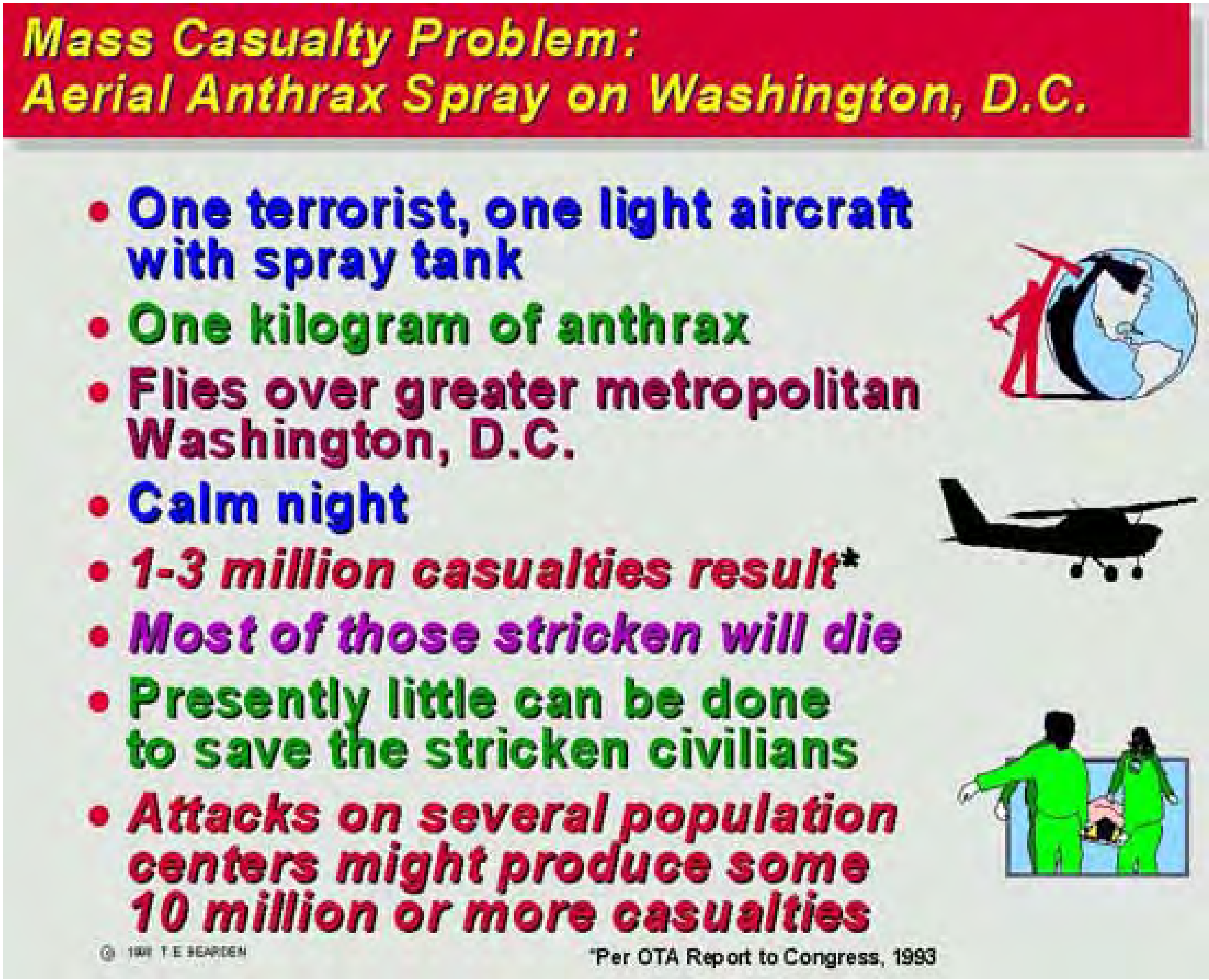

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- One kilogram of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© [illegible] T E BEARDEN

*Per OTA Report to Congress, 1993

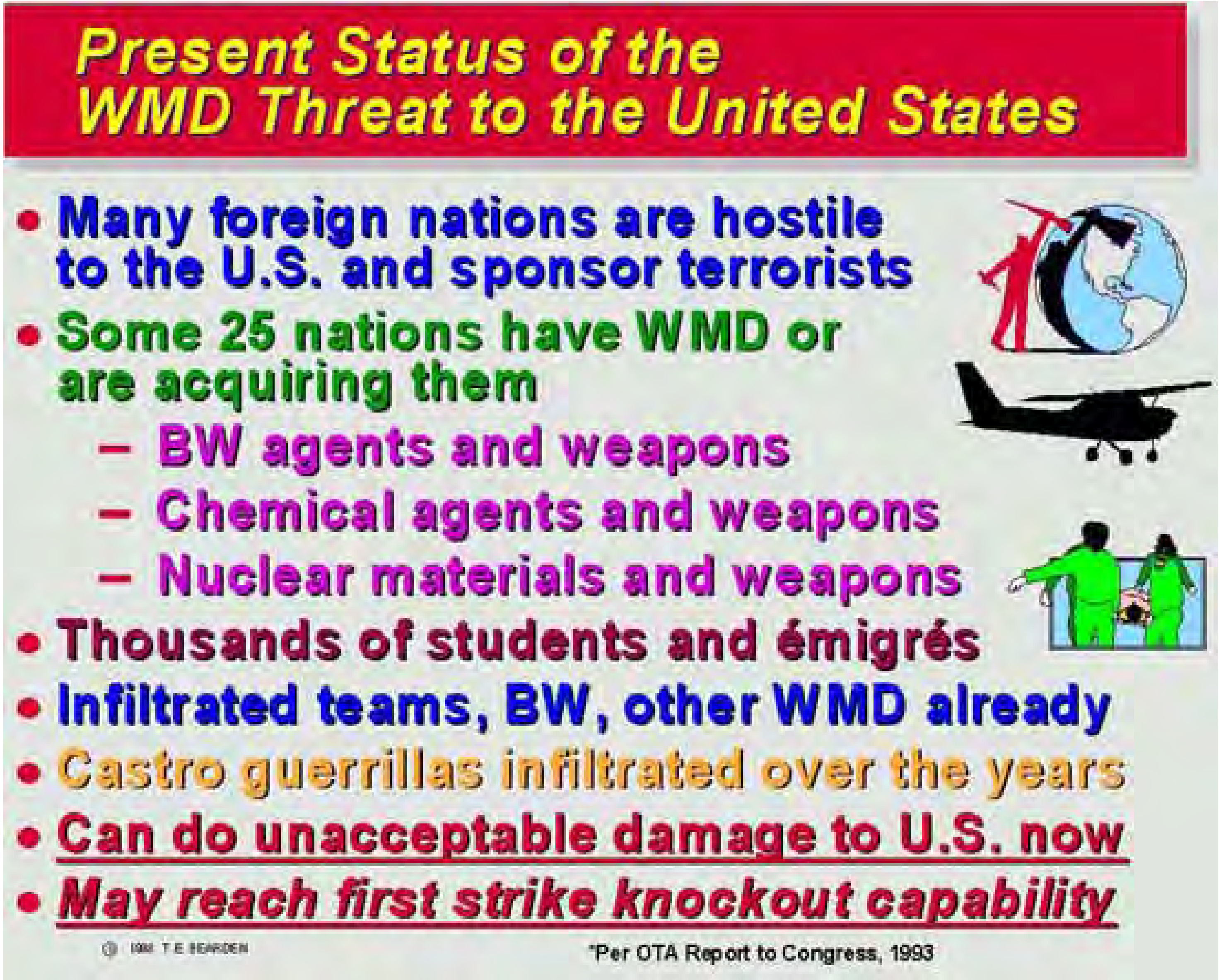

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach first strike knockout capability***

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

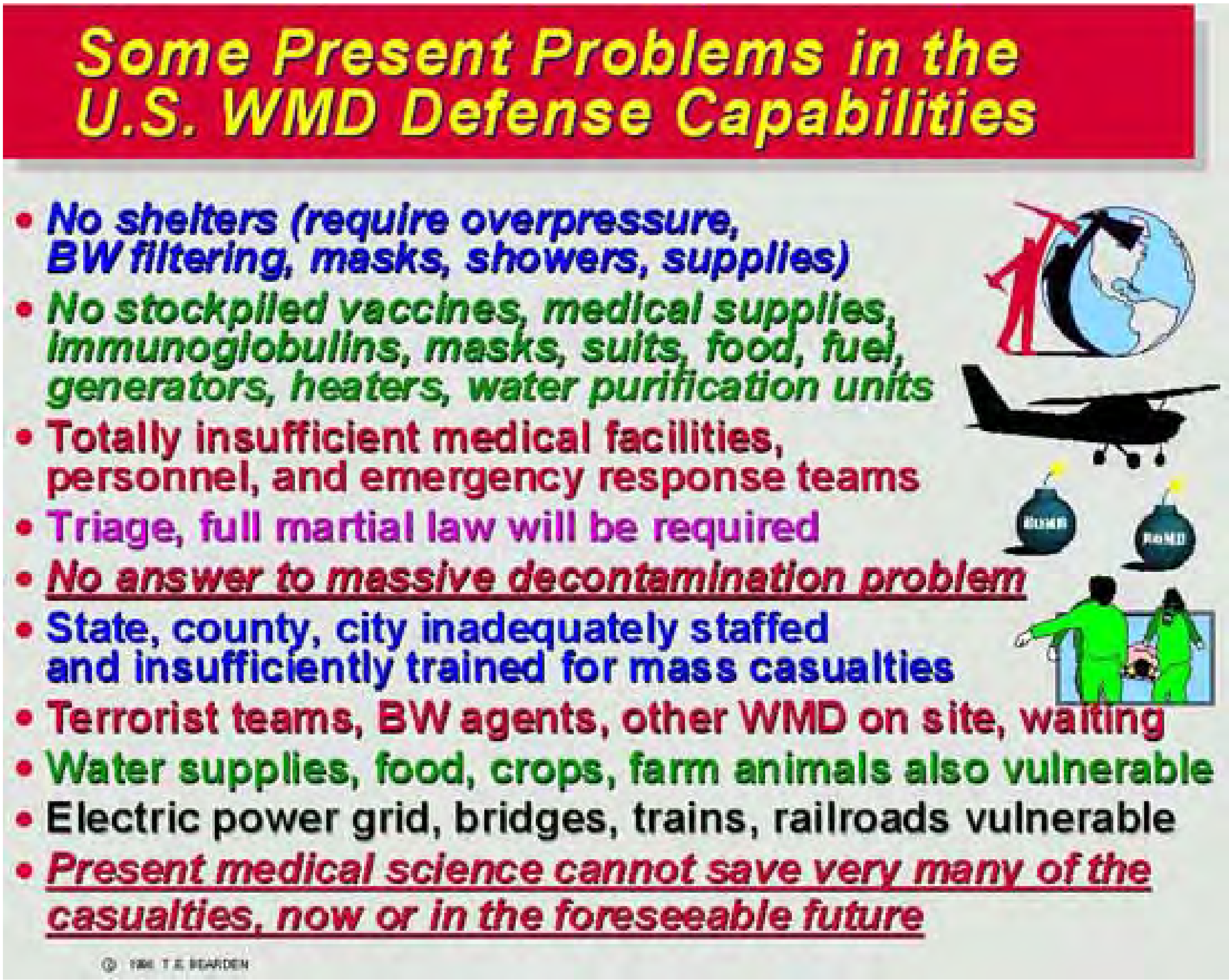

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© [illegible] T.E. BEARDEN

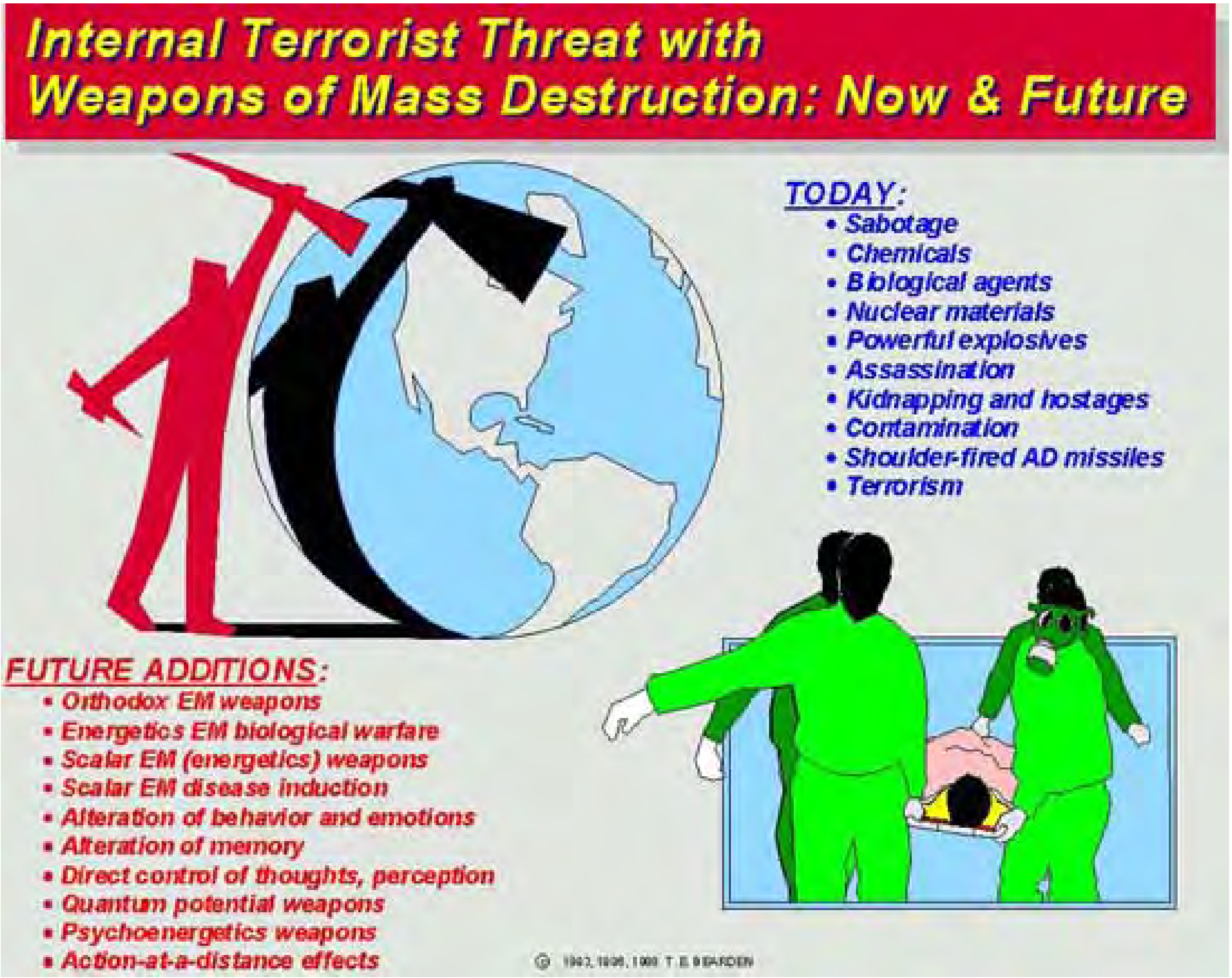

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

***TODAY*:**
- *Sabotage*
- *Chemicals*
- *Biological agents*
- *Nuclear materials*
- *Powerful explosives*
- *Assassination*
- *Kidnapping and hostages*
- *Contamination*
- *Shoulder-fired AD missiles*
- *Terrorism*

***FUTURE ADDITIONS*:**
- *Orthodox EM weapons*
- *Energetics EM biological warfare*
- *Scalar EM (energetics) weapons*
- *Scalar EM disease induction*
- *Alteration of behavior and emotions*
- *Alteration of memory*
- *Direct control of thoughts, perception*
- *Quantum potential weapons*
- *Psychoenergetics weapons*
- *Action-at-a-distance effects*

© [illegible] T. E. BEARDEN

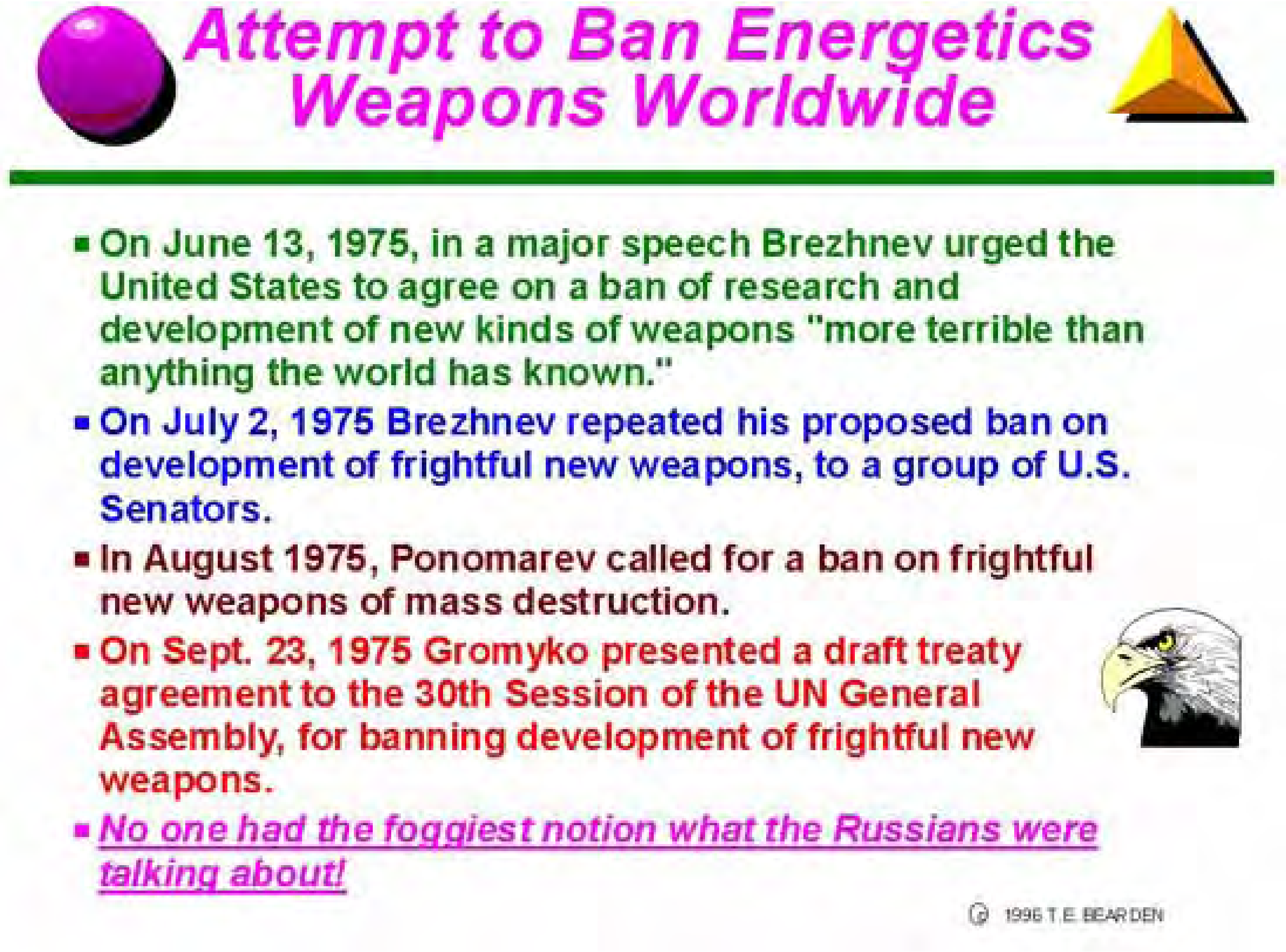

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

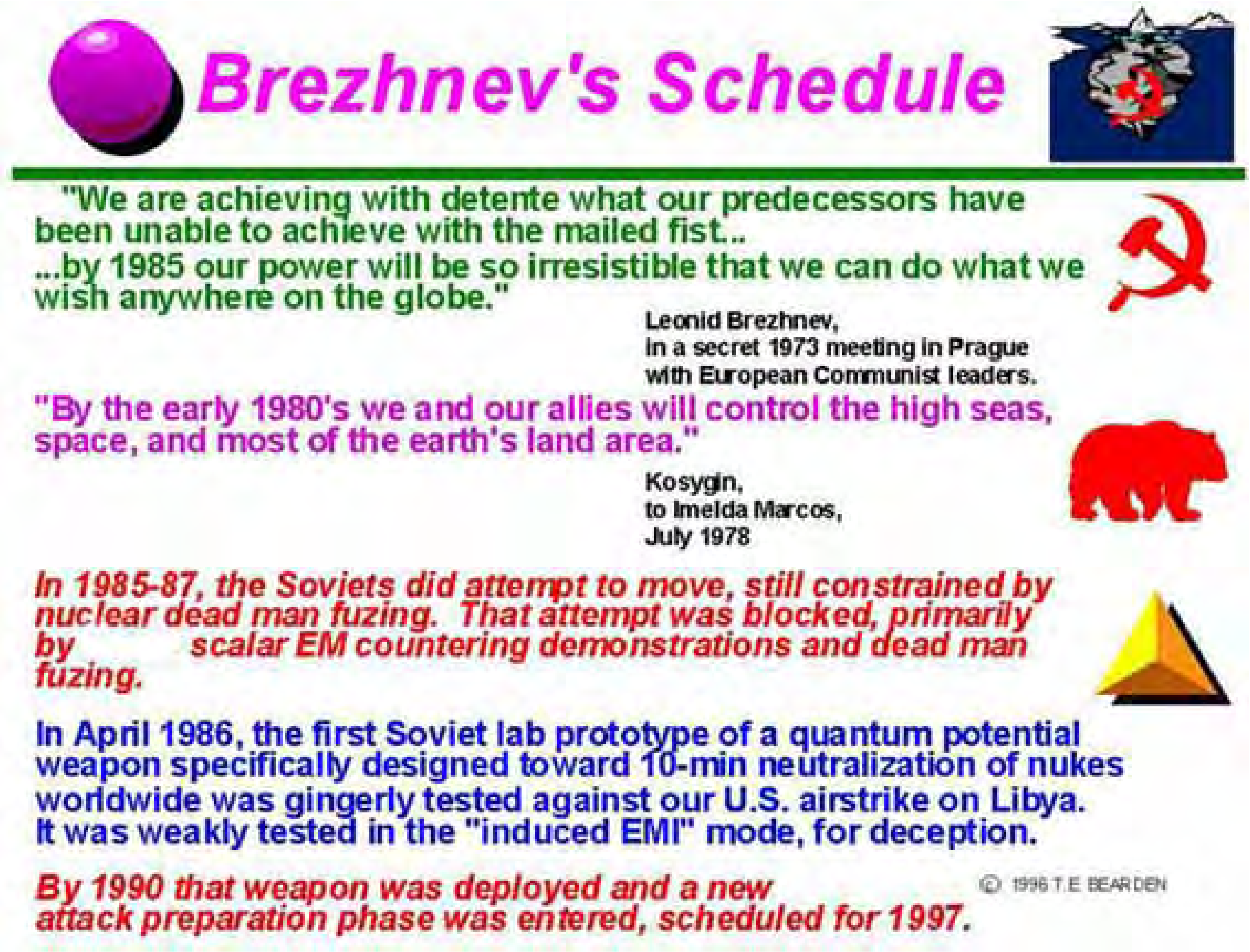

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist..
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
In a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

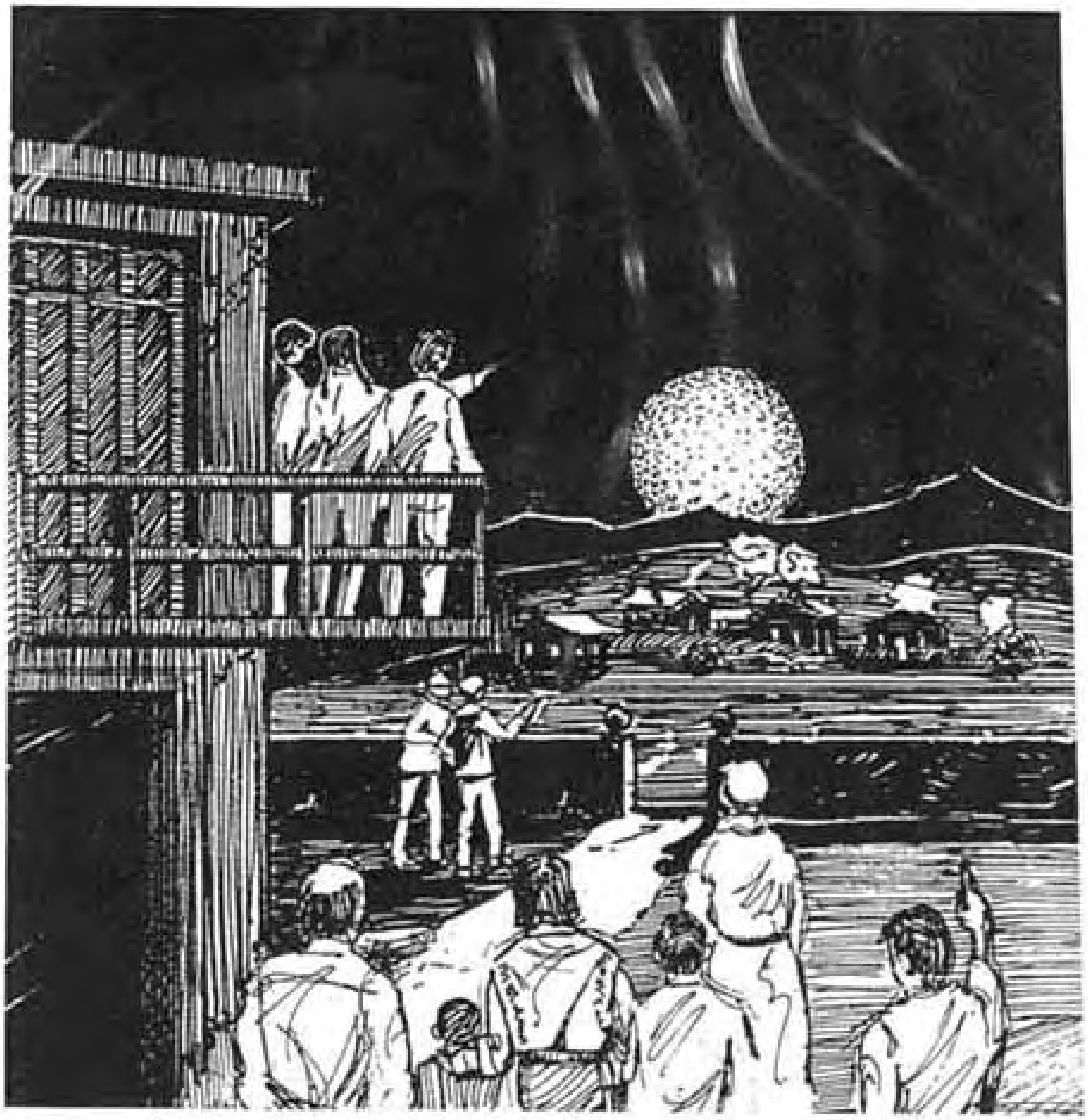

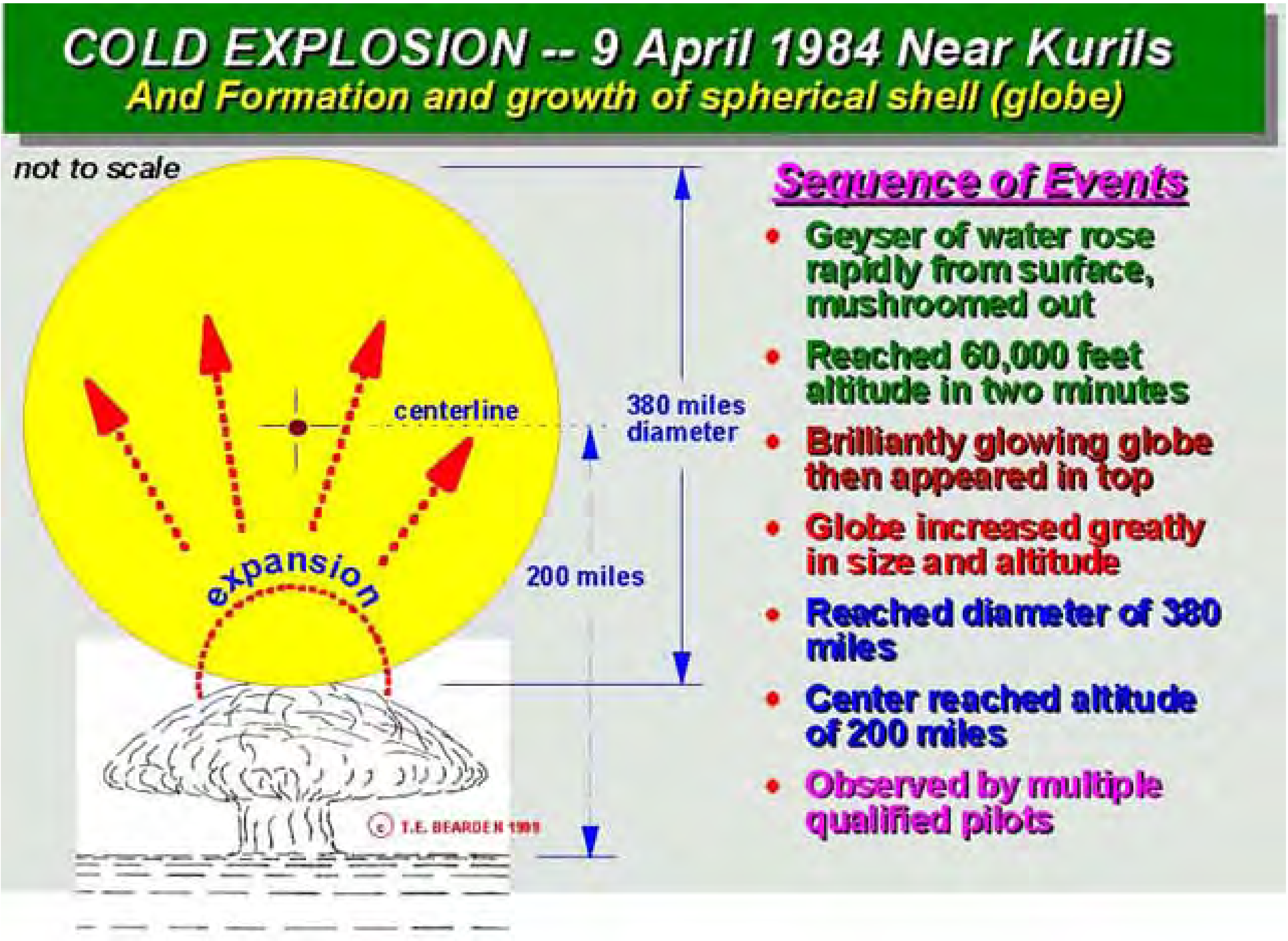

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles diameter
200 miles
expansion
© T.E. BEARDEN [illegible]

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

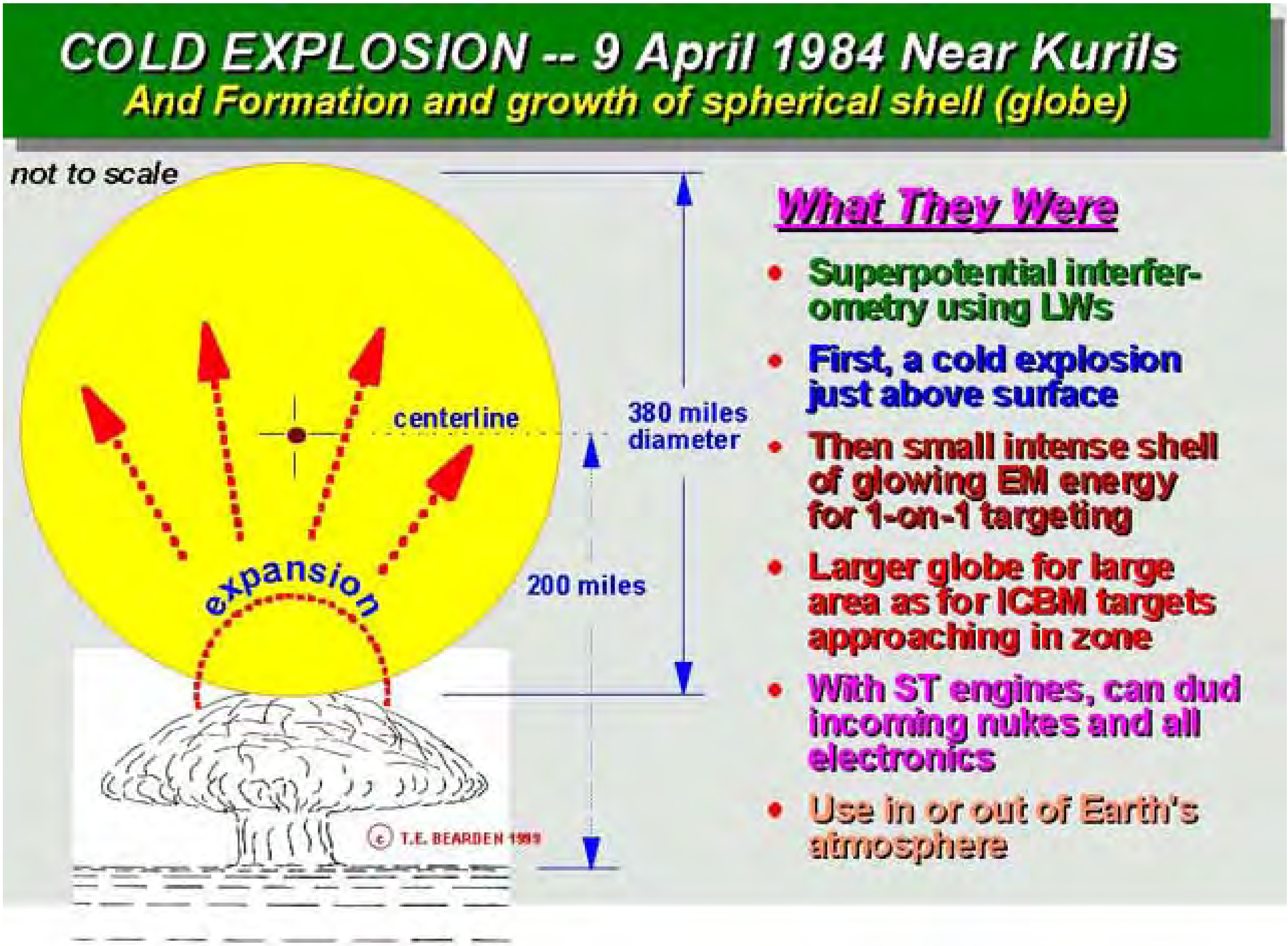

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
expansion
380 miles diameter
200 miles
© T.E. BEARDEN 1999

What They Were

- Superpotential interferometry using LWs
- First, a cold explosion just above surface
- Then small intense shell of glowing EM energy for 1-on-1 targeting
- Larger globe for large area as for ICBM targets approaching in zone
- With ST engines, can dud incoming nukes and all electronics
- Use in or out of Earth's atmosphere

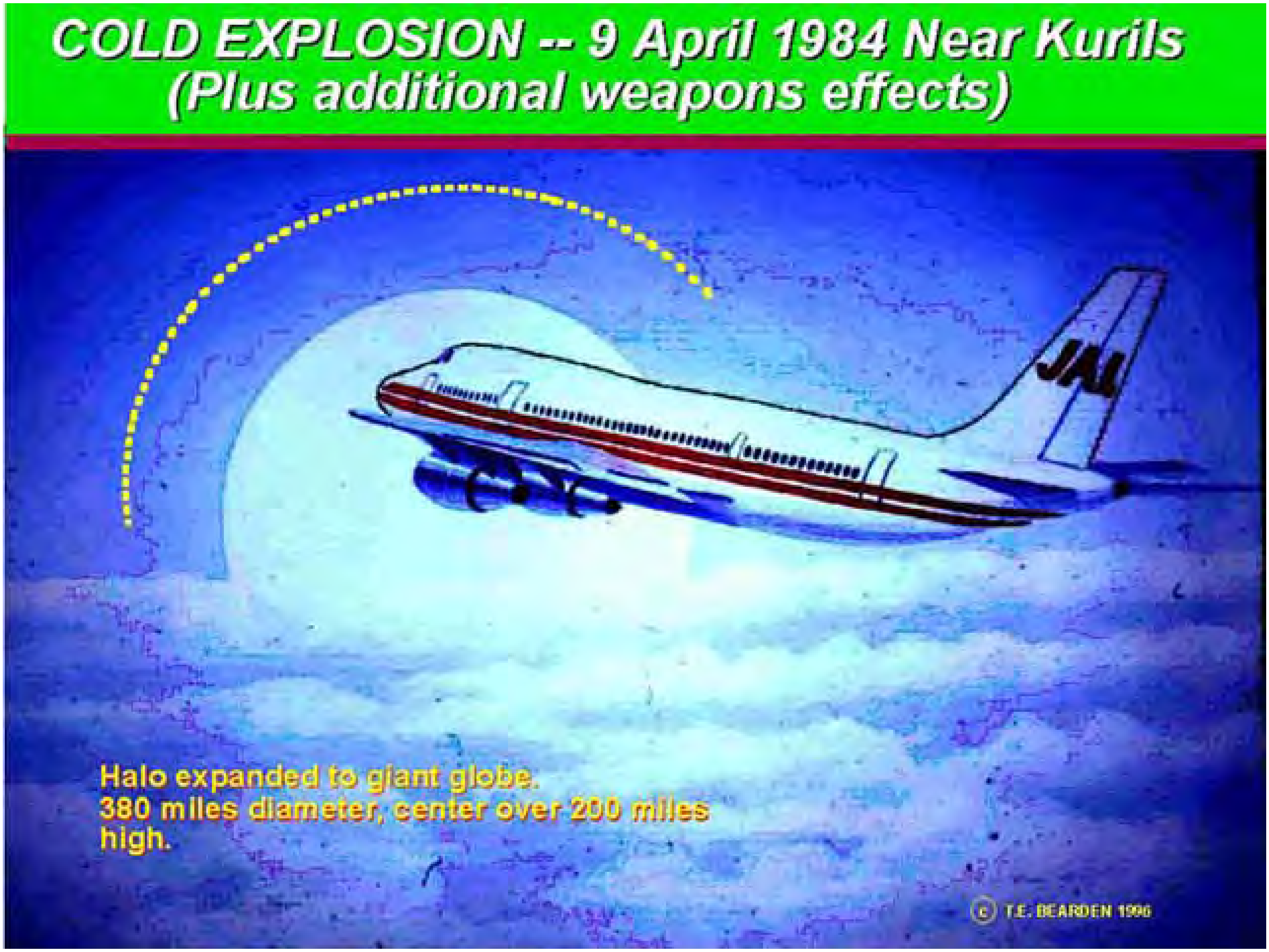

COLD EXPLOSION -- 9 April 1984 Near Kurils (Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

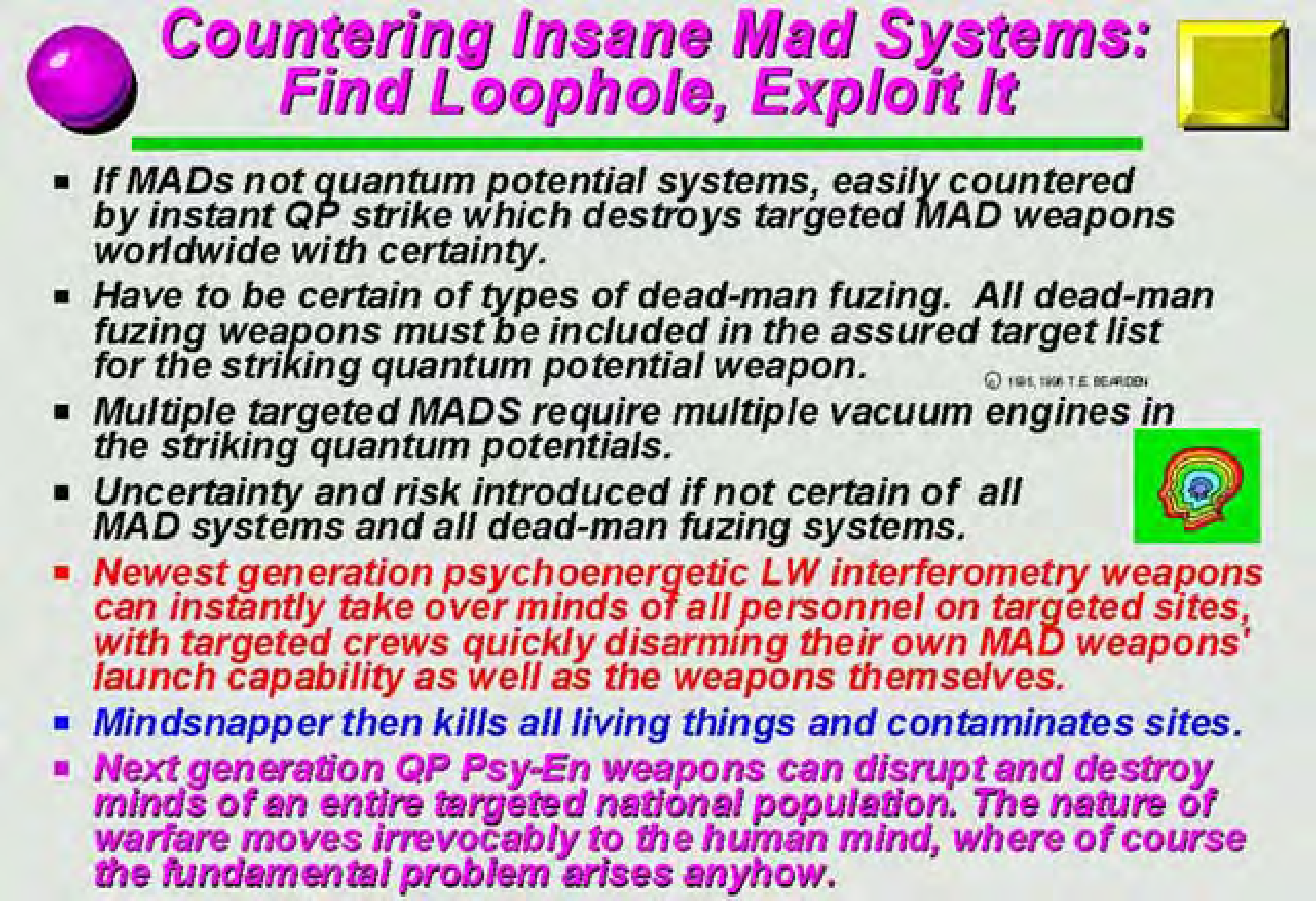

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*
- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

© 1995, 1996 T.E. BEARDEN

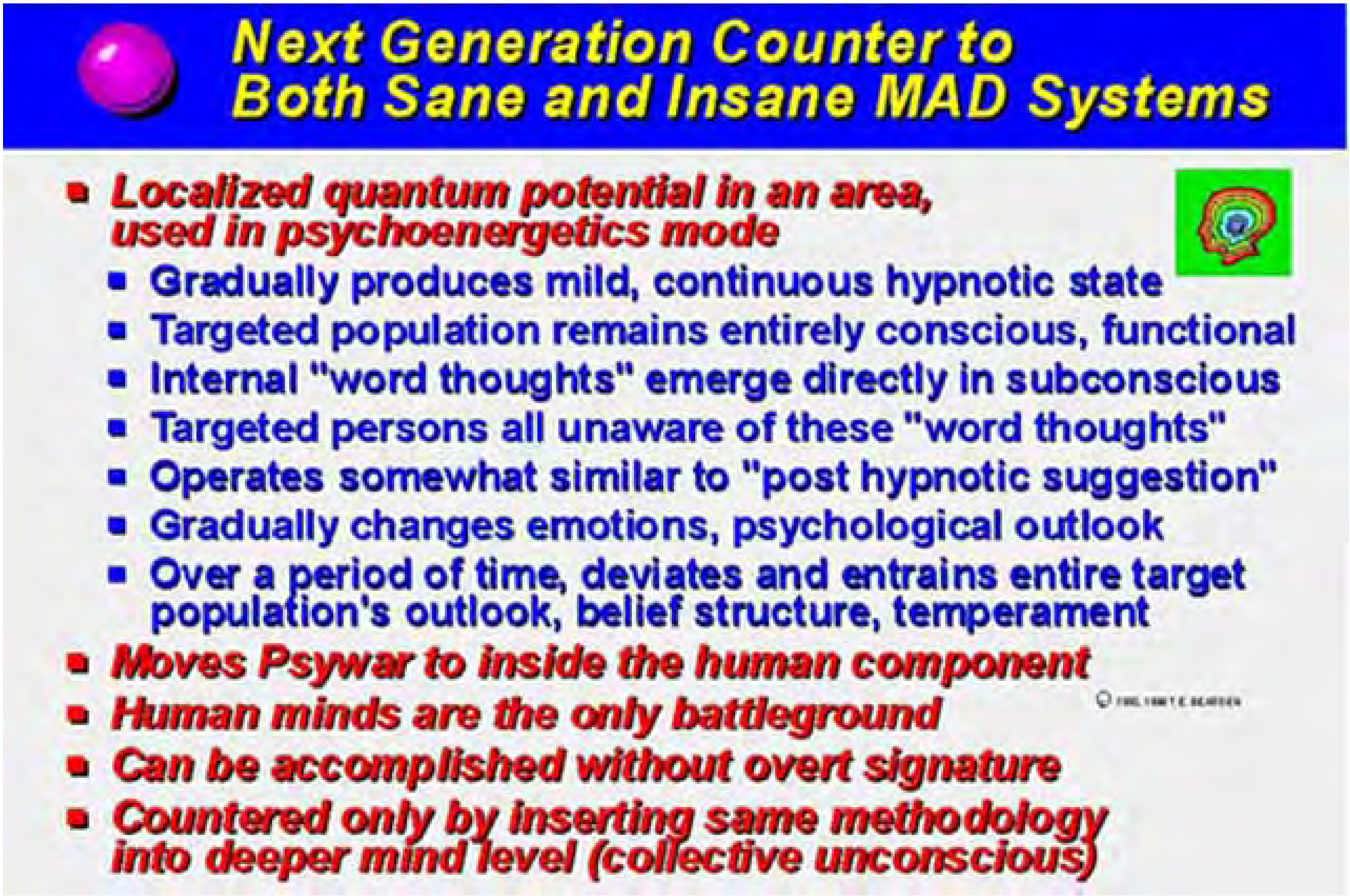

Next Generation Counter to Both Sane and Insane MAD Systems

- ***Localized quantum potential in an area, used in psychoenergetics mode***
  - Gradually produces mild, continuous hypnotic state
  - Targeted population remains entirely conscious, functional
  - Internal "word thoughts" emerge directly in subconscious
  - Targeted persons all unaware of these "word thoughts"
  - Operates somewhat similar to "post hypnotic suggestion"
  - Gradually changes emotions, psychological outlook
  - Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament
- ***Moves Psywar to inside the human component***
- ***Human minds are the only battleground***
- ***Can be accomplished without overt signature***
- ***Countered only by inserting same methodology into deeper mind level (collective unconscious)***

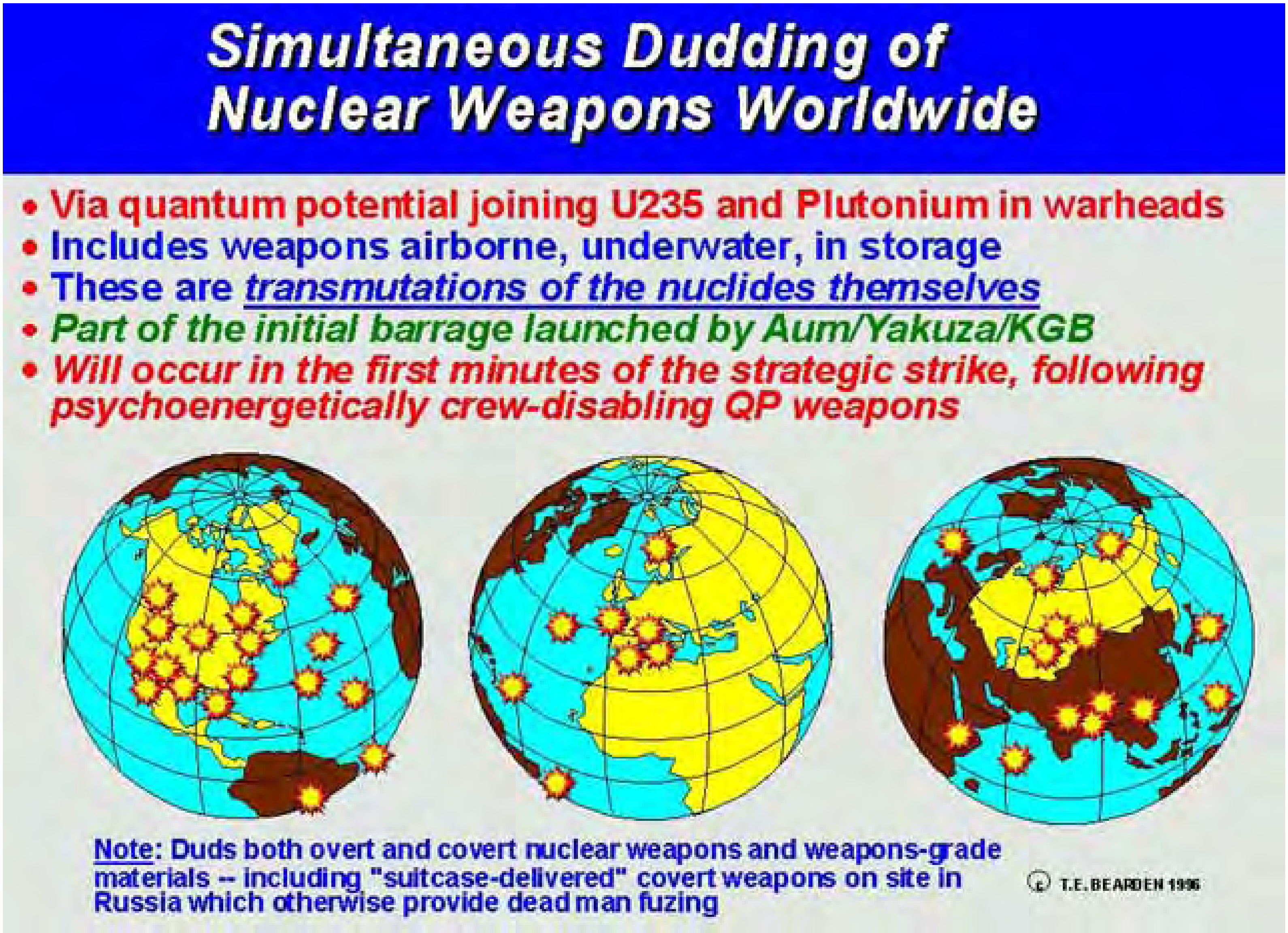

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

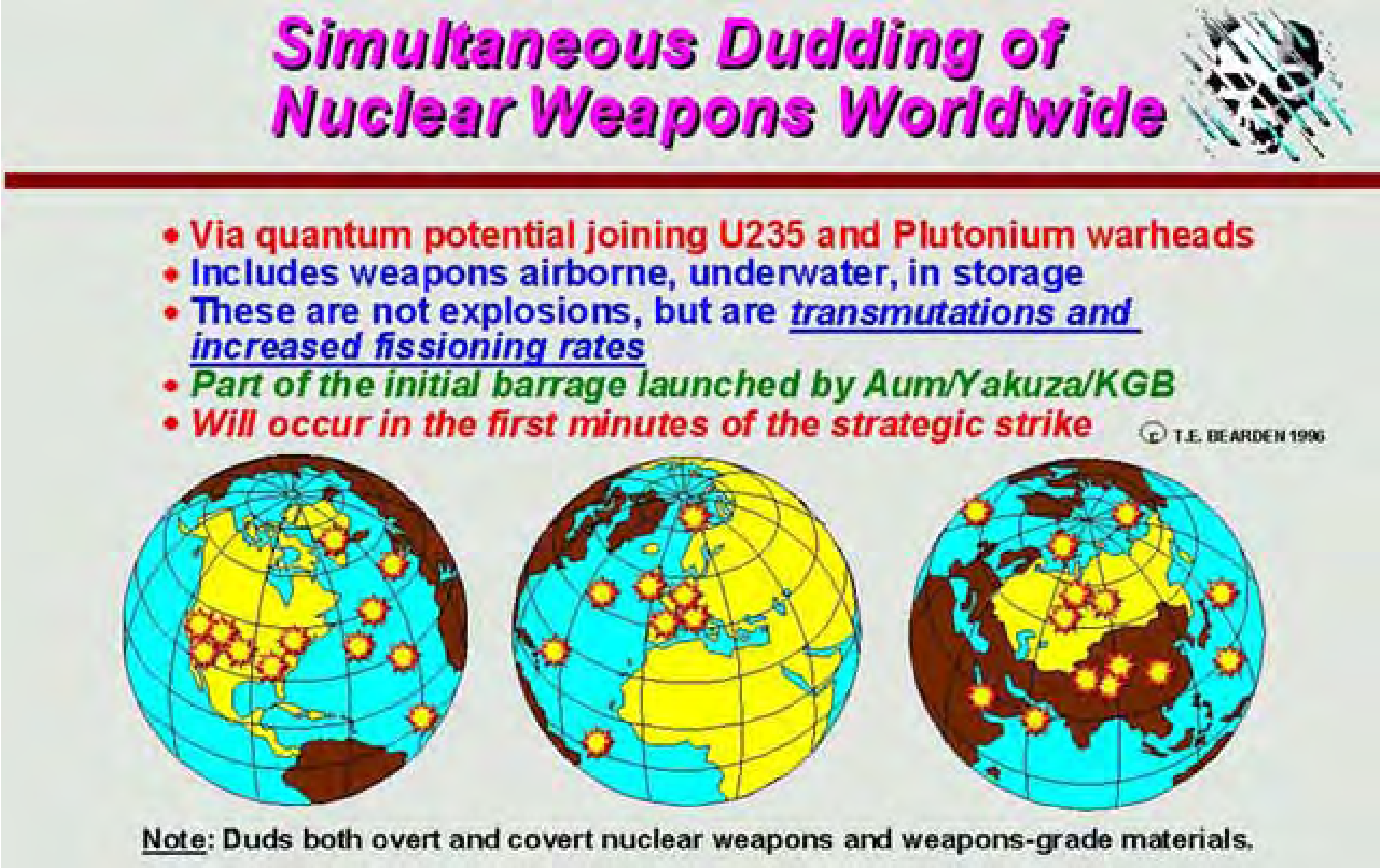

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are ***transmutations and increased fissioning rates***
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

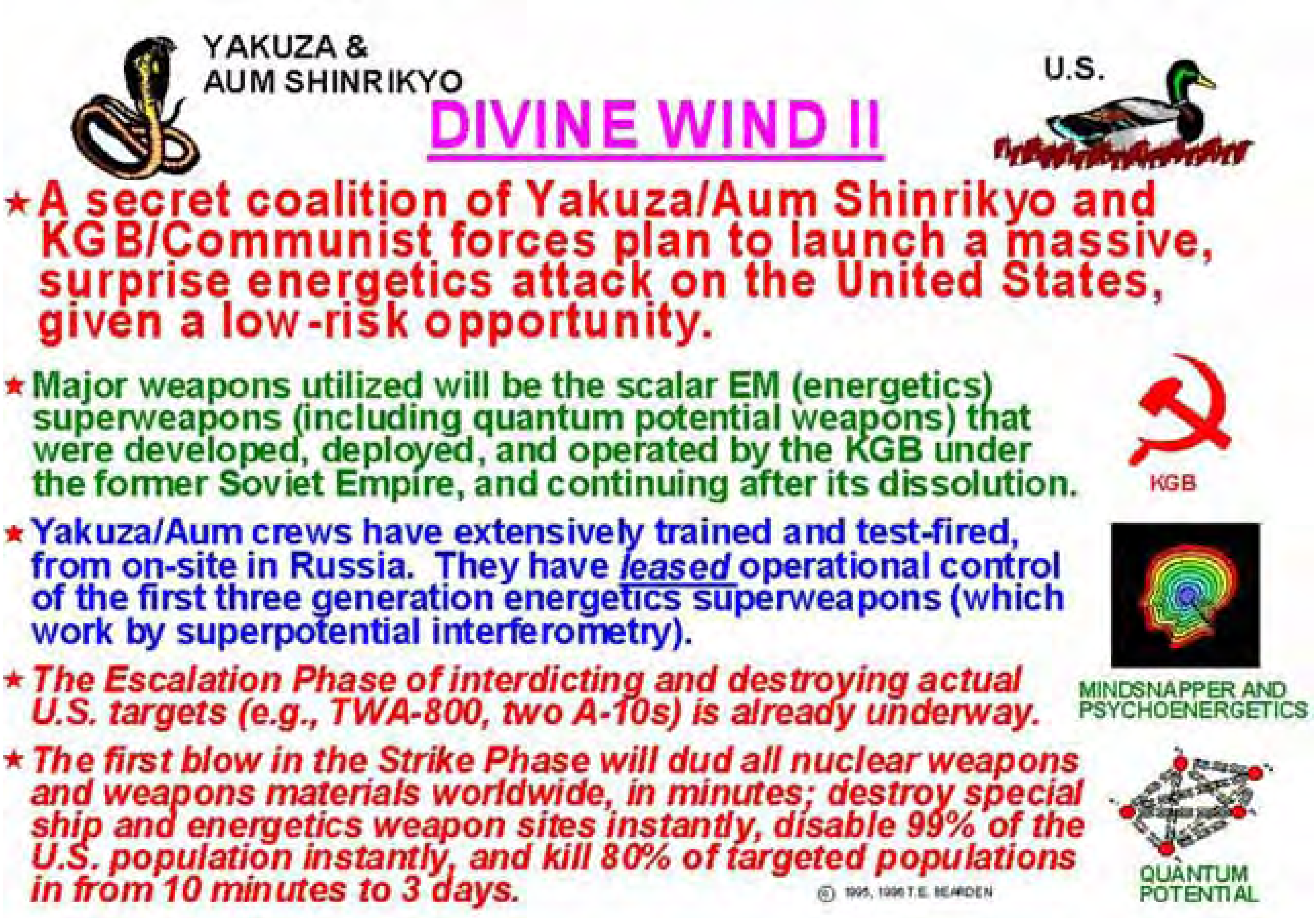

YAKUZA &
AUM SHINRIKYO
U.S.

DIVINE WIND II

- **A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.**
- Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
- Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).
- *The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.*
- *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.*

© 1995, 1996 T.E. BEARDEN

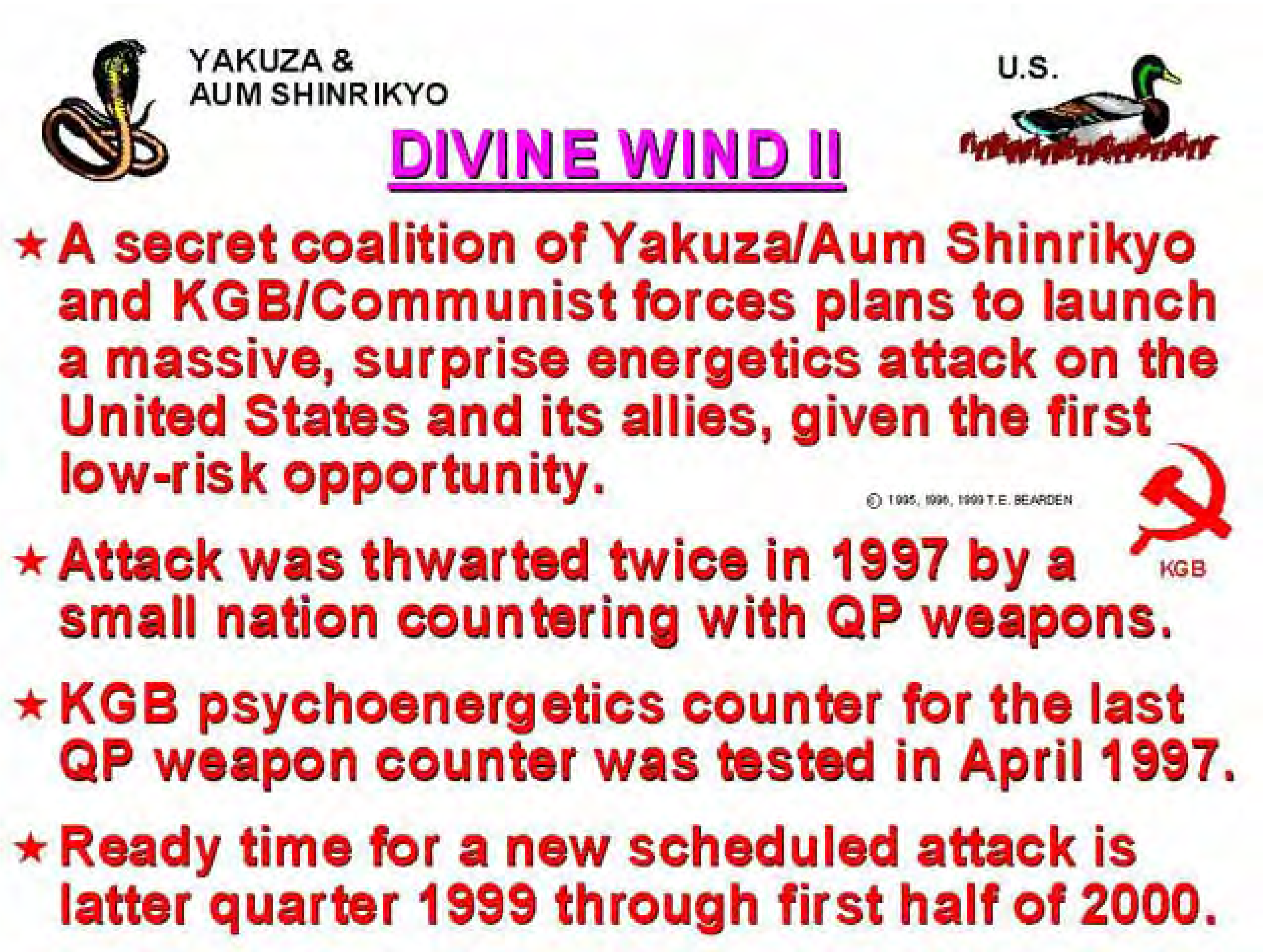

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

KGB

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

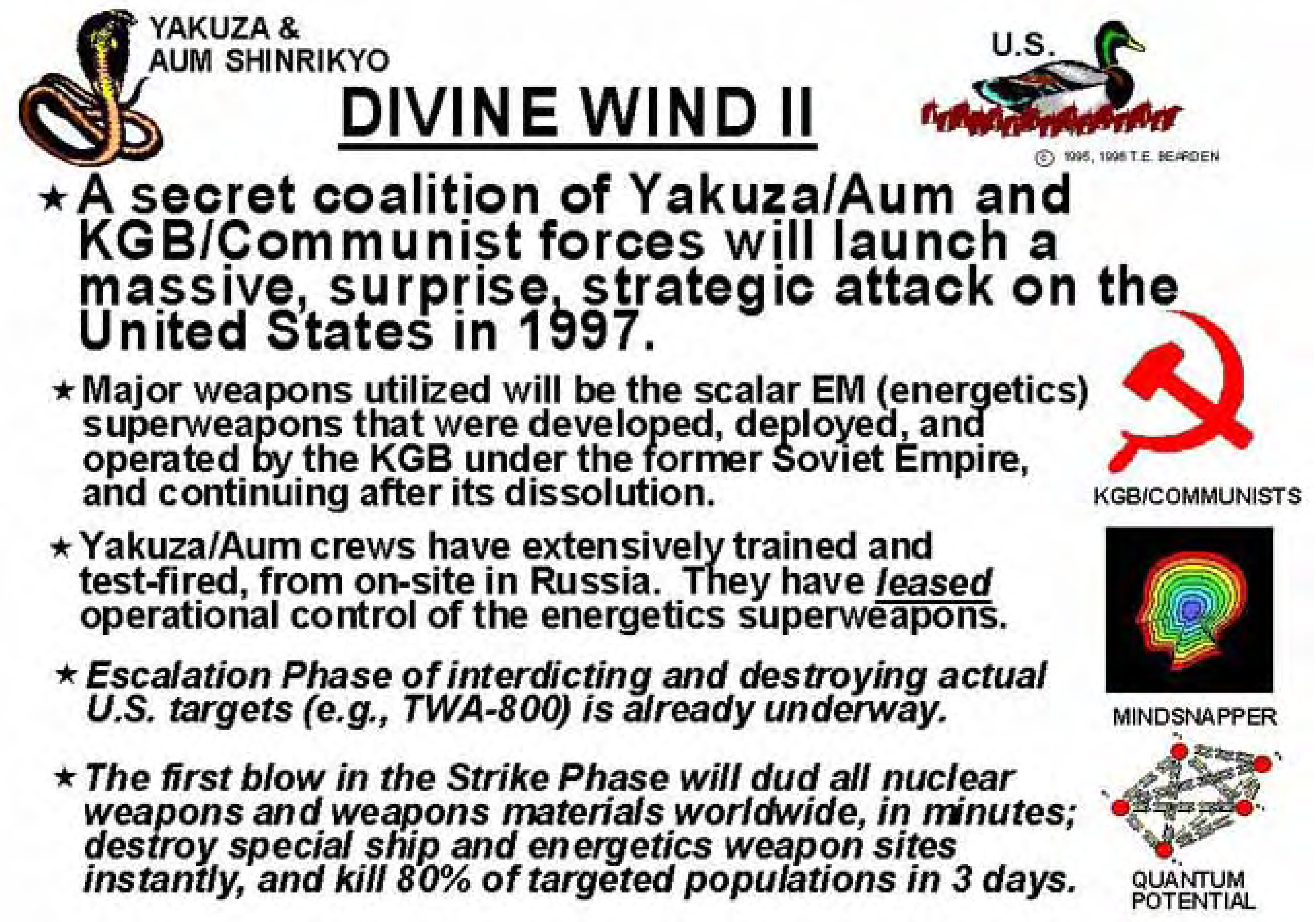

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

© 1995, 1996 T.E. BEARDEN

★ **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**

★ Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.

★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.

★ *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*

★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

KGB/COMMUNISTS

MINDSNAPPER

QUANTUM
POTENTIAL

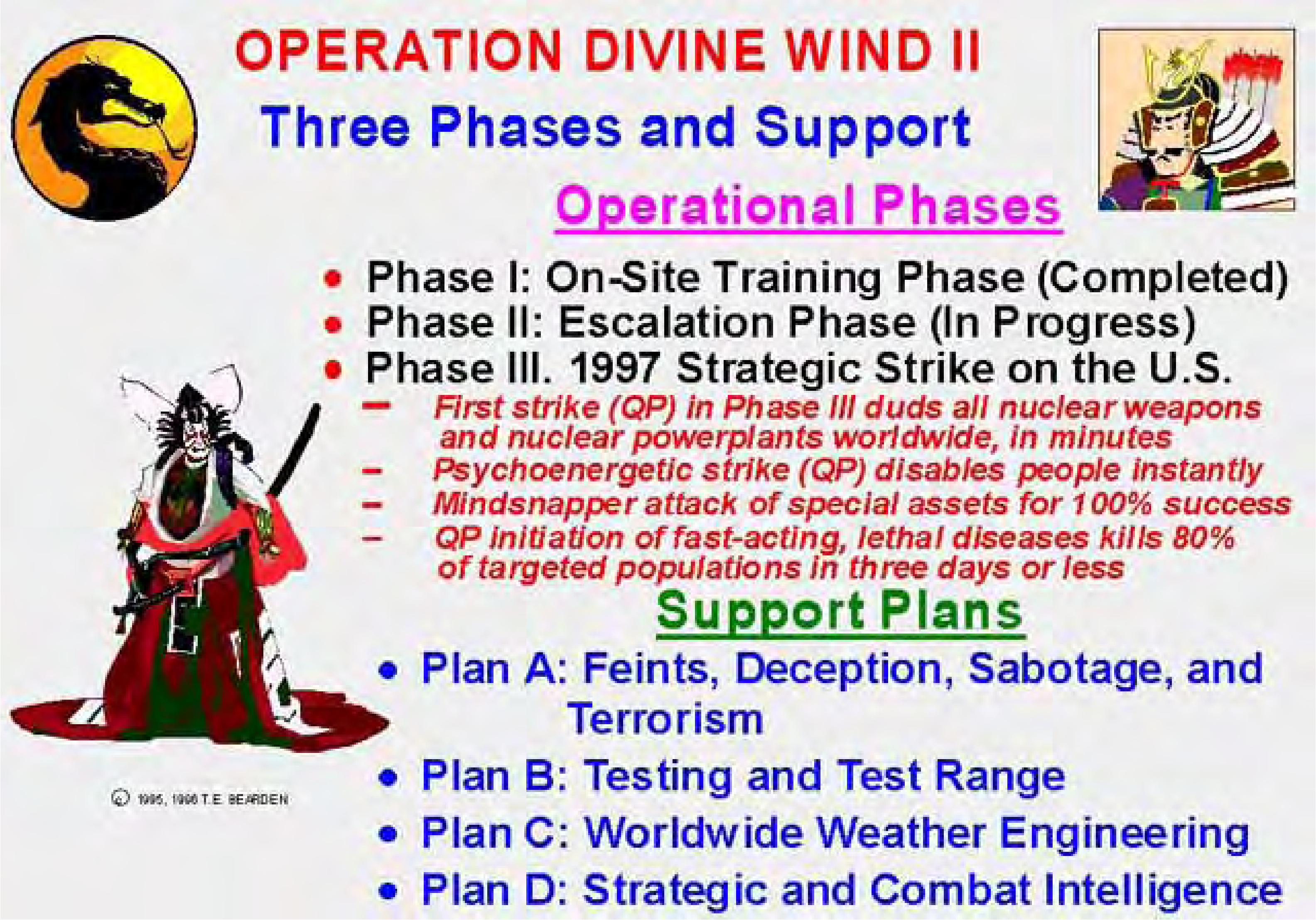

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

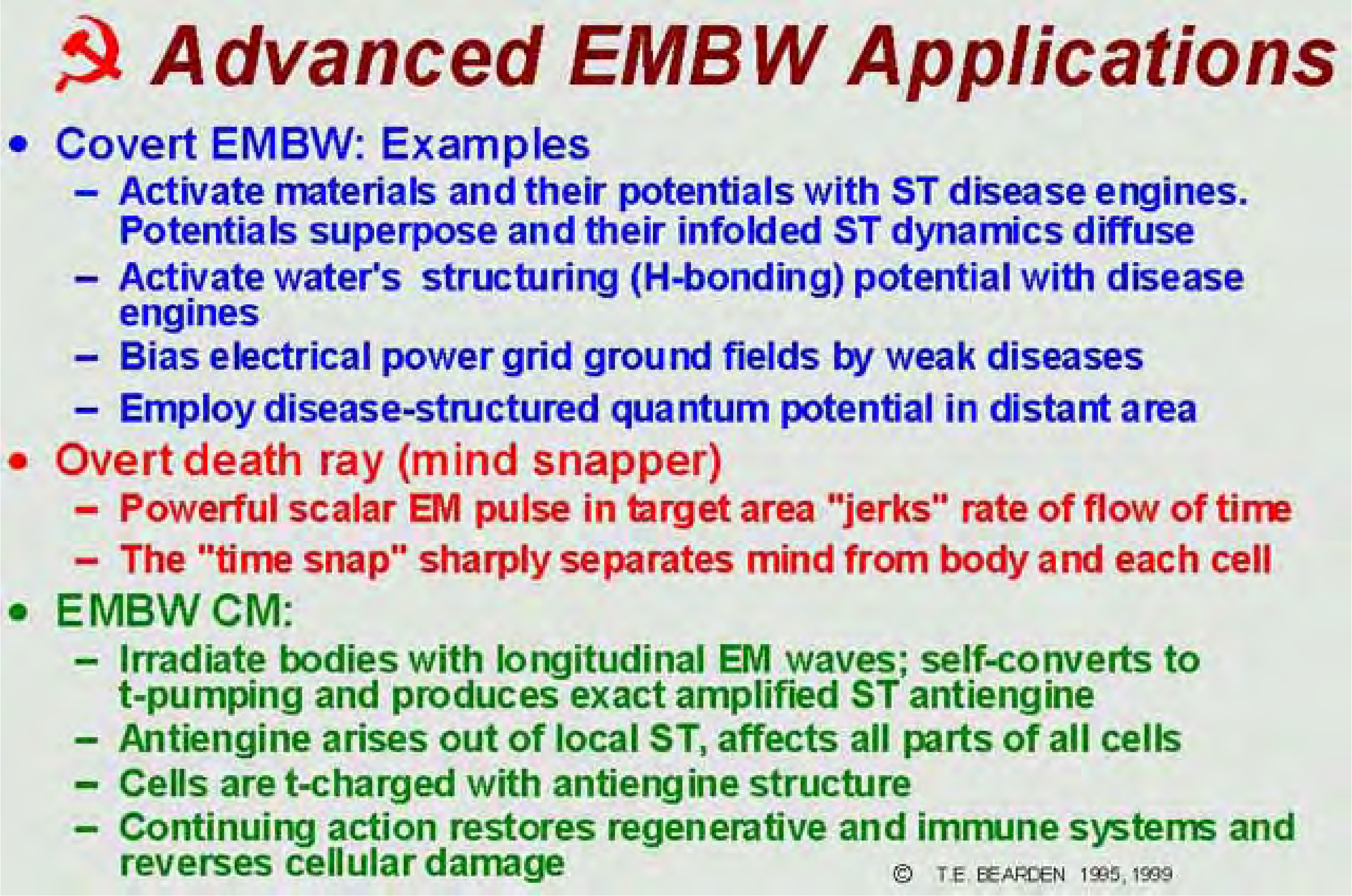

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

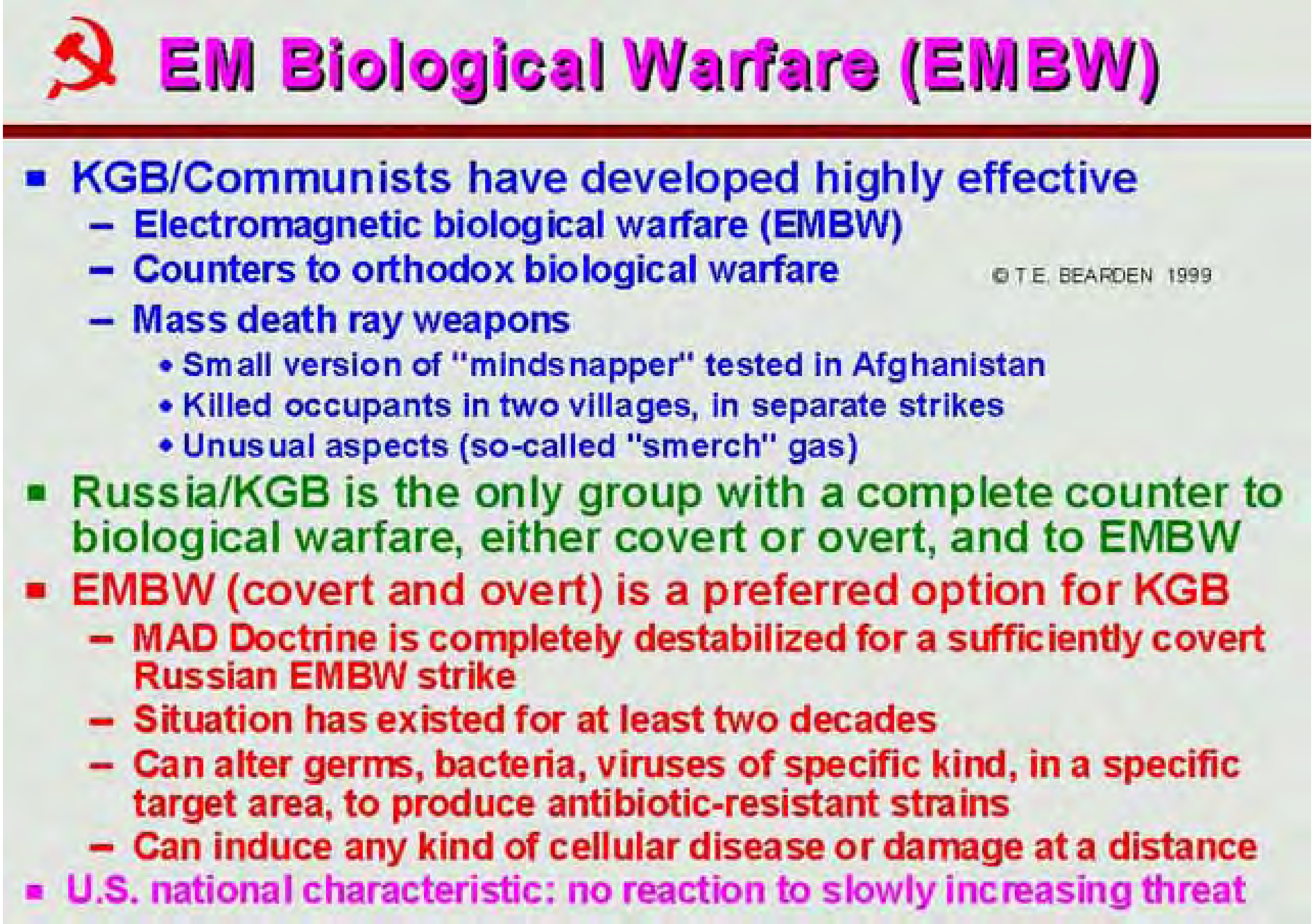

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

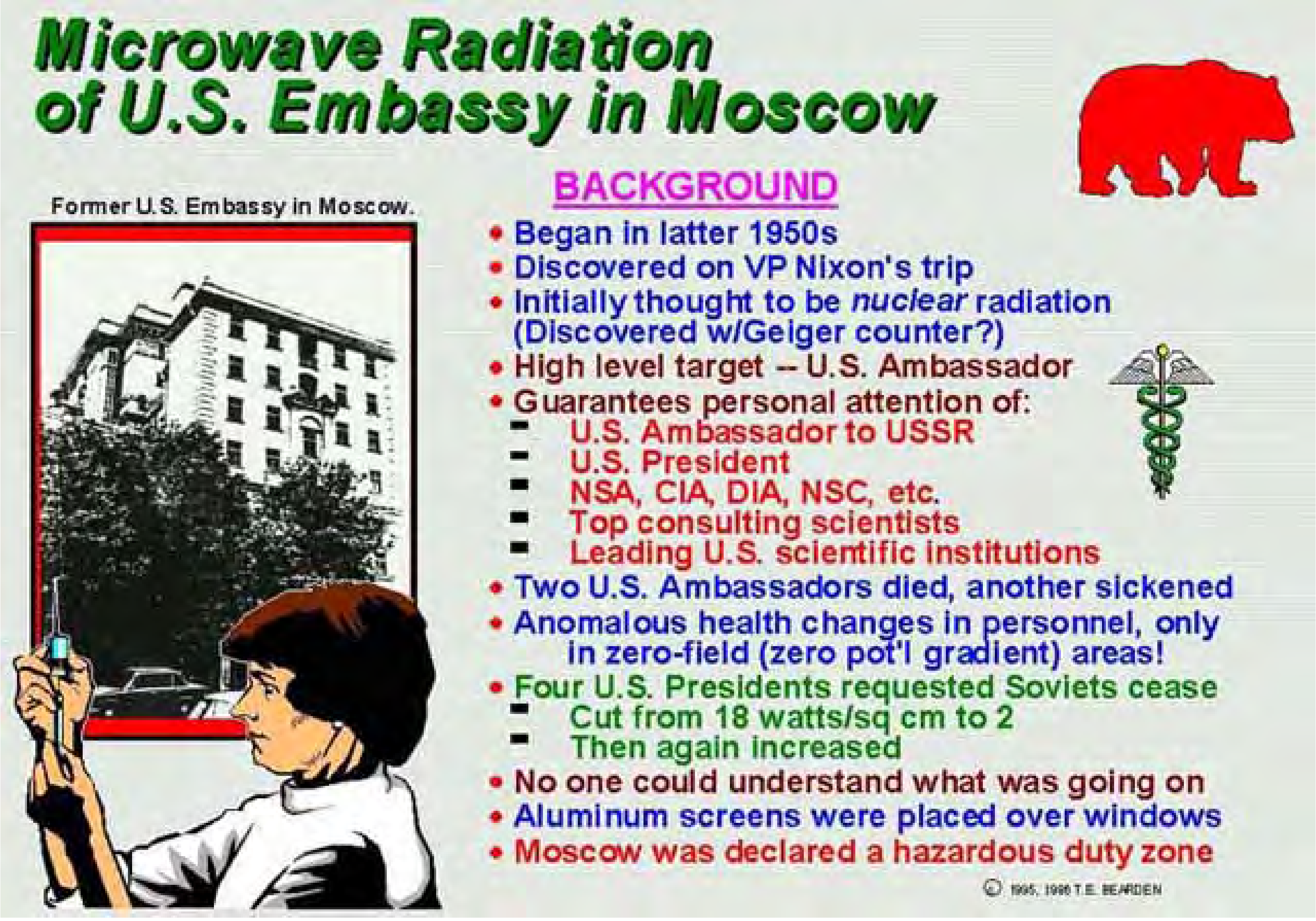

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

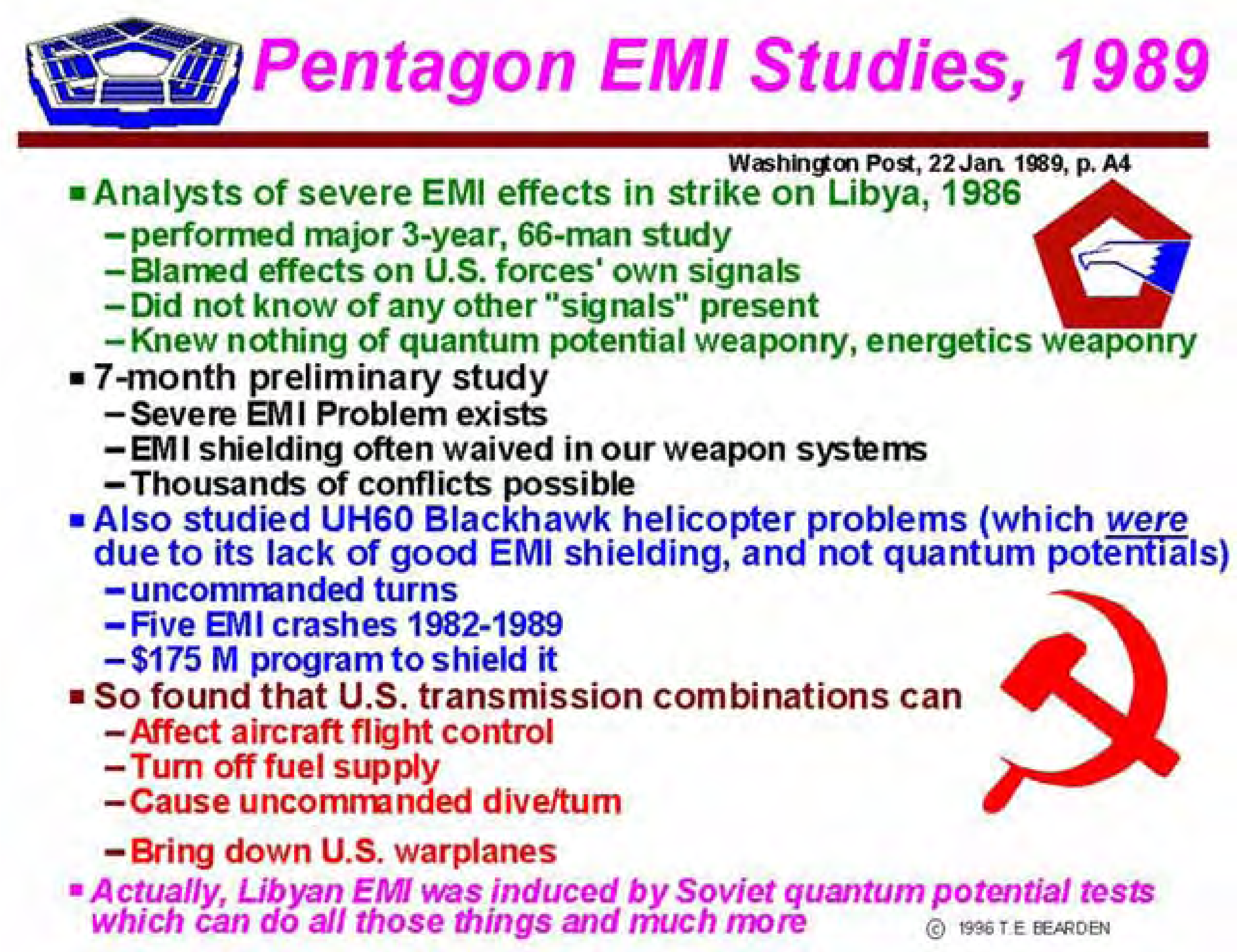

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia

- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996

- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996

- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985

- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985

- EM Missile Practice and Deception
  November 17, 1996

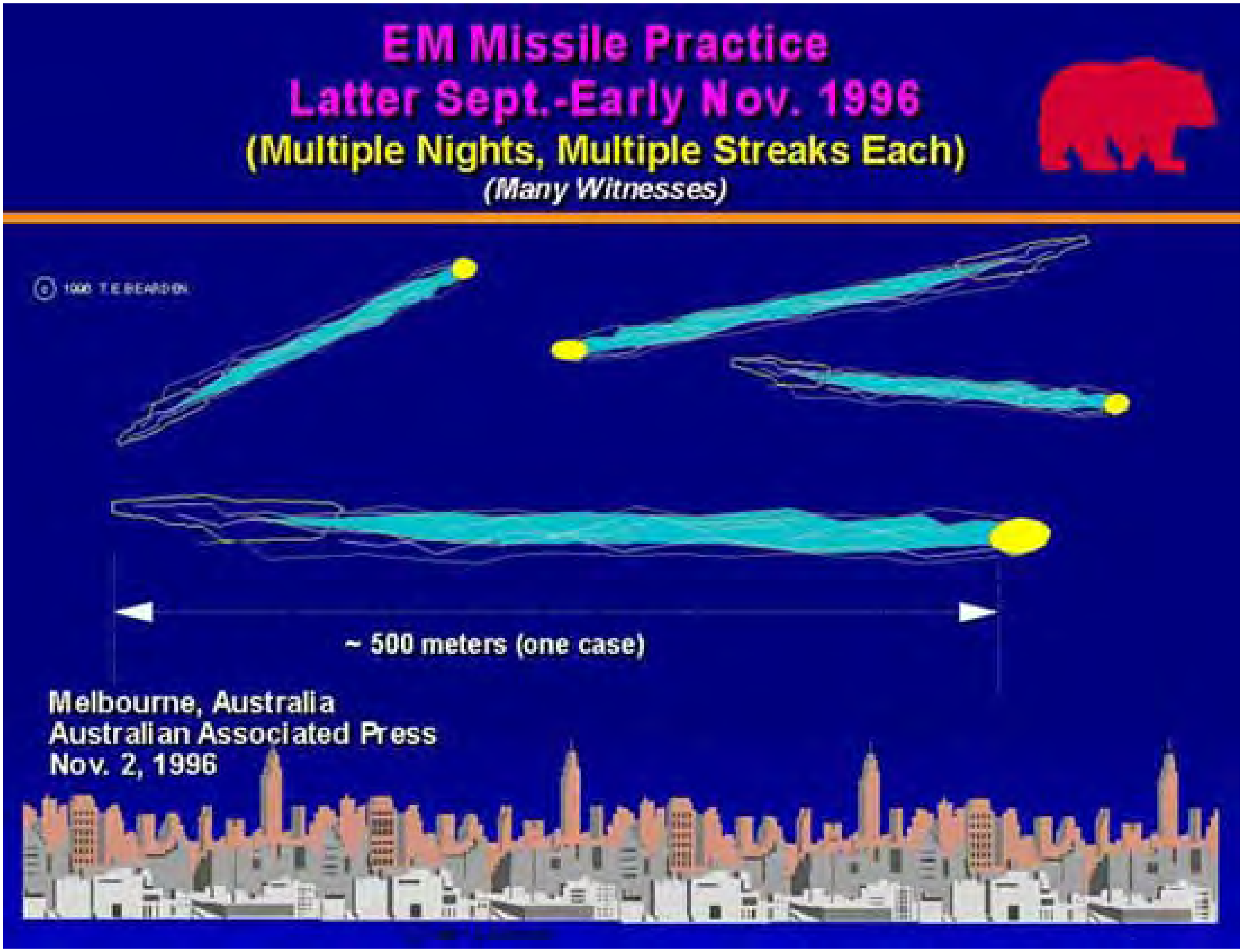

EM Missile Practice
Latter Sept.-Early Nov. 1996

(Multiple Nights, Multiple Streaks Each)

(Many Witnesses)

© 1996 T.E.BEARDEN

~ 500 meters (one case)

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

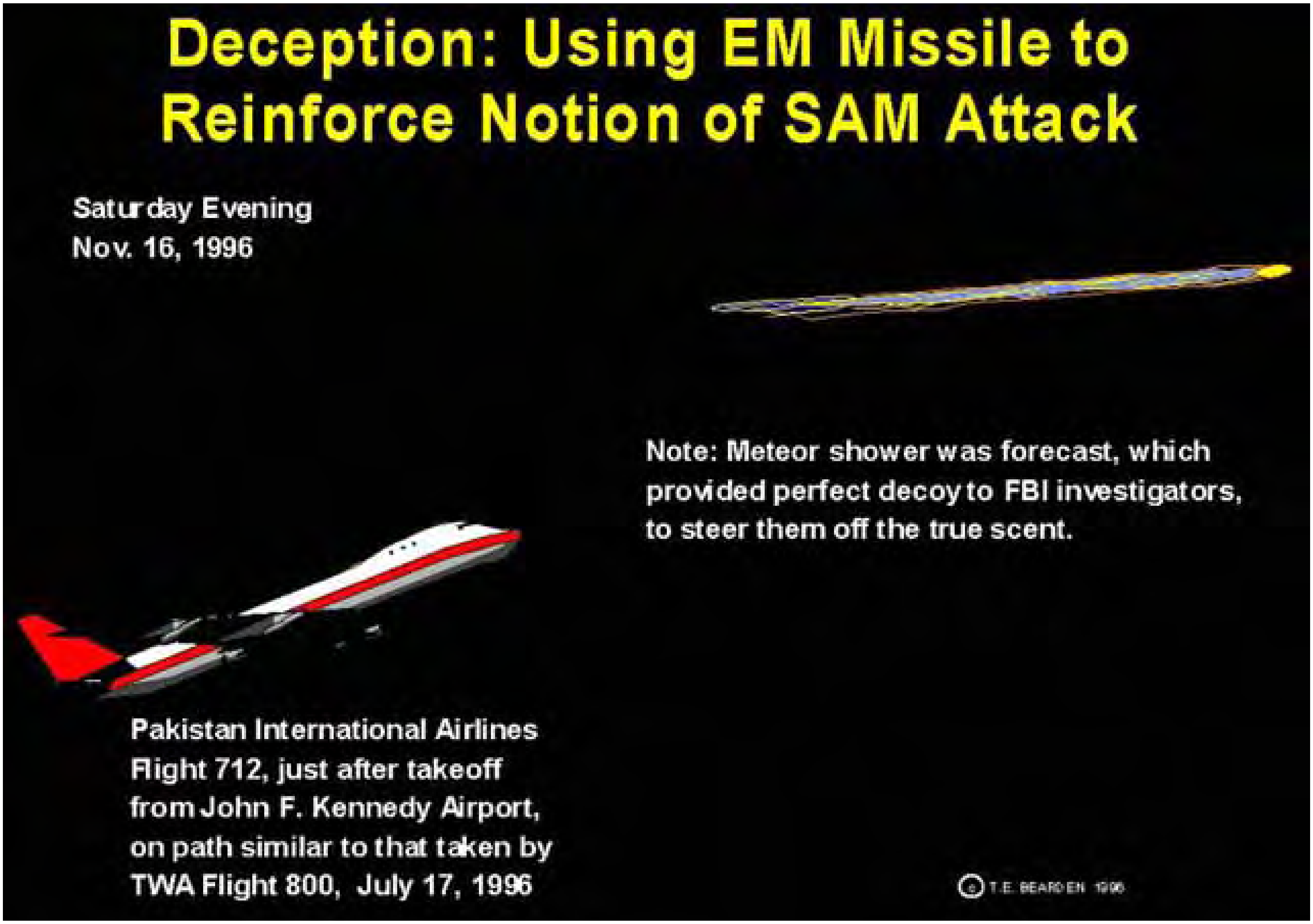

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

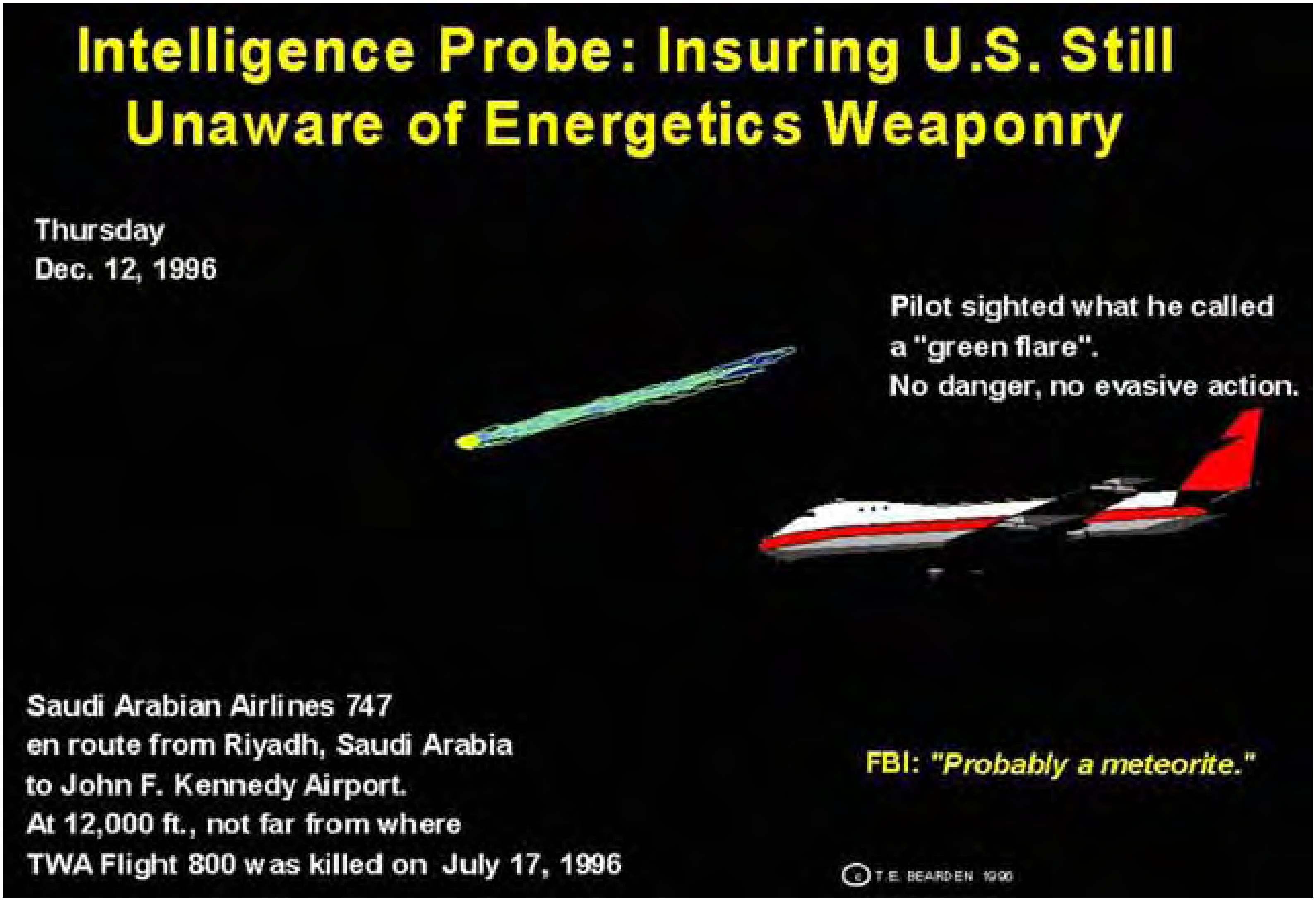

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: *"Probably a meteorite."*

© T.E. BEARDEN 1996

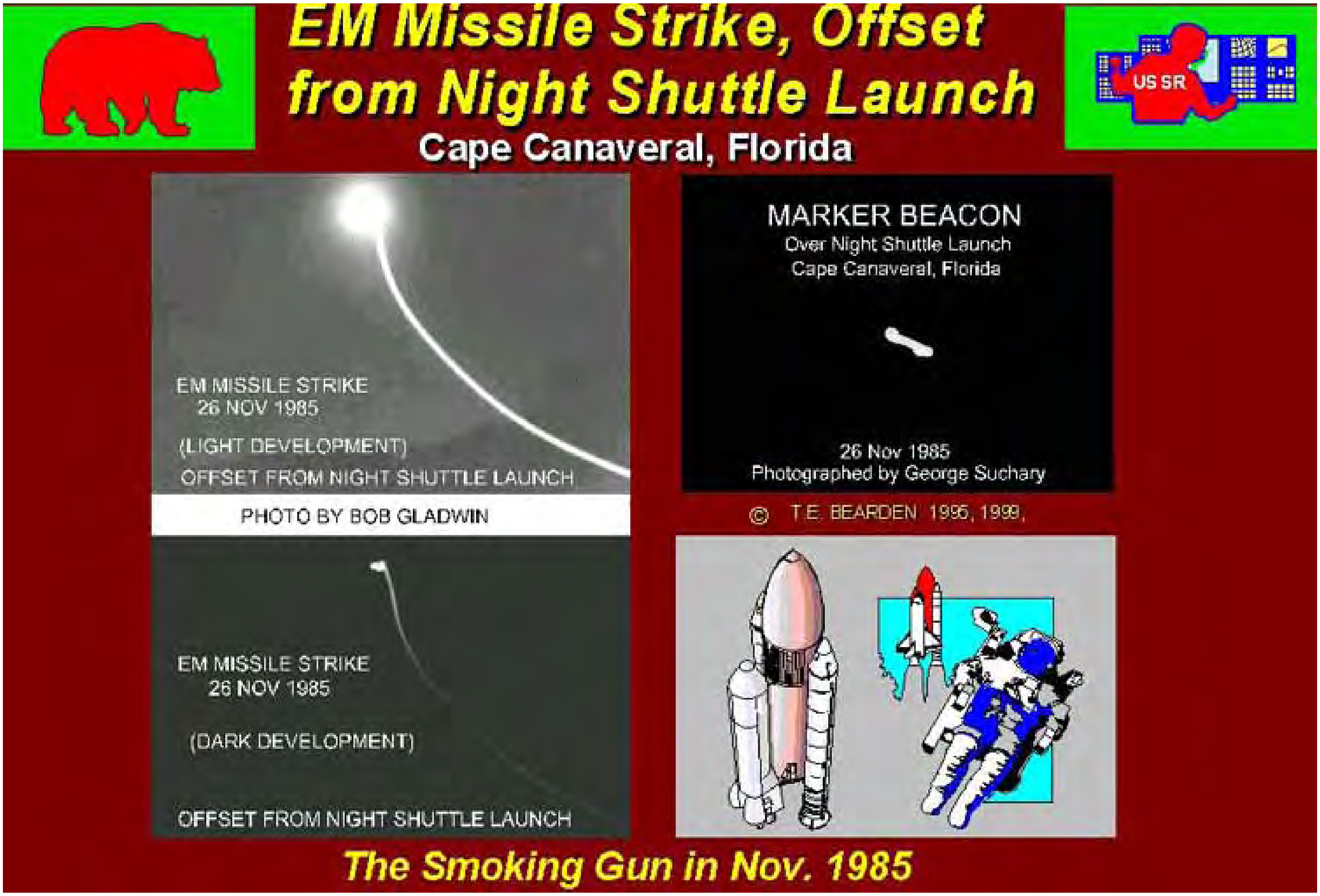

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985
(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1996, 1999,

EM MISSILE STRIKE
26 NOV 1985
(DARK DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

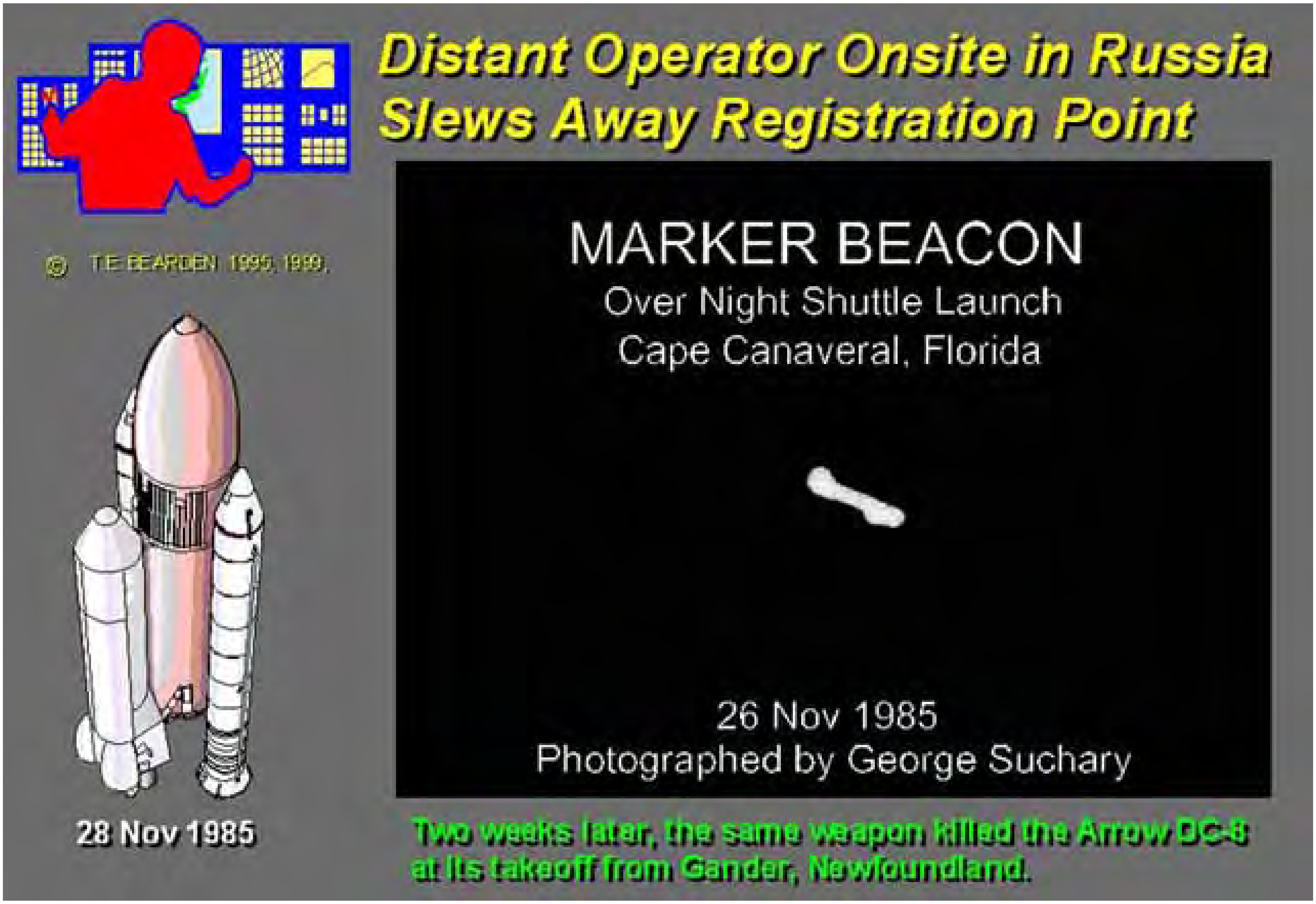

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

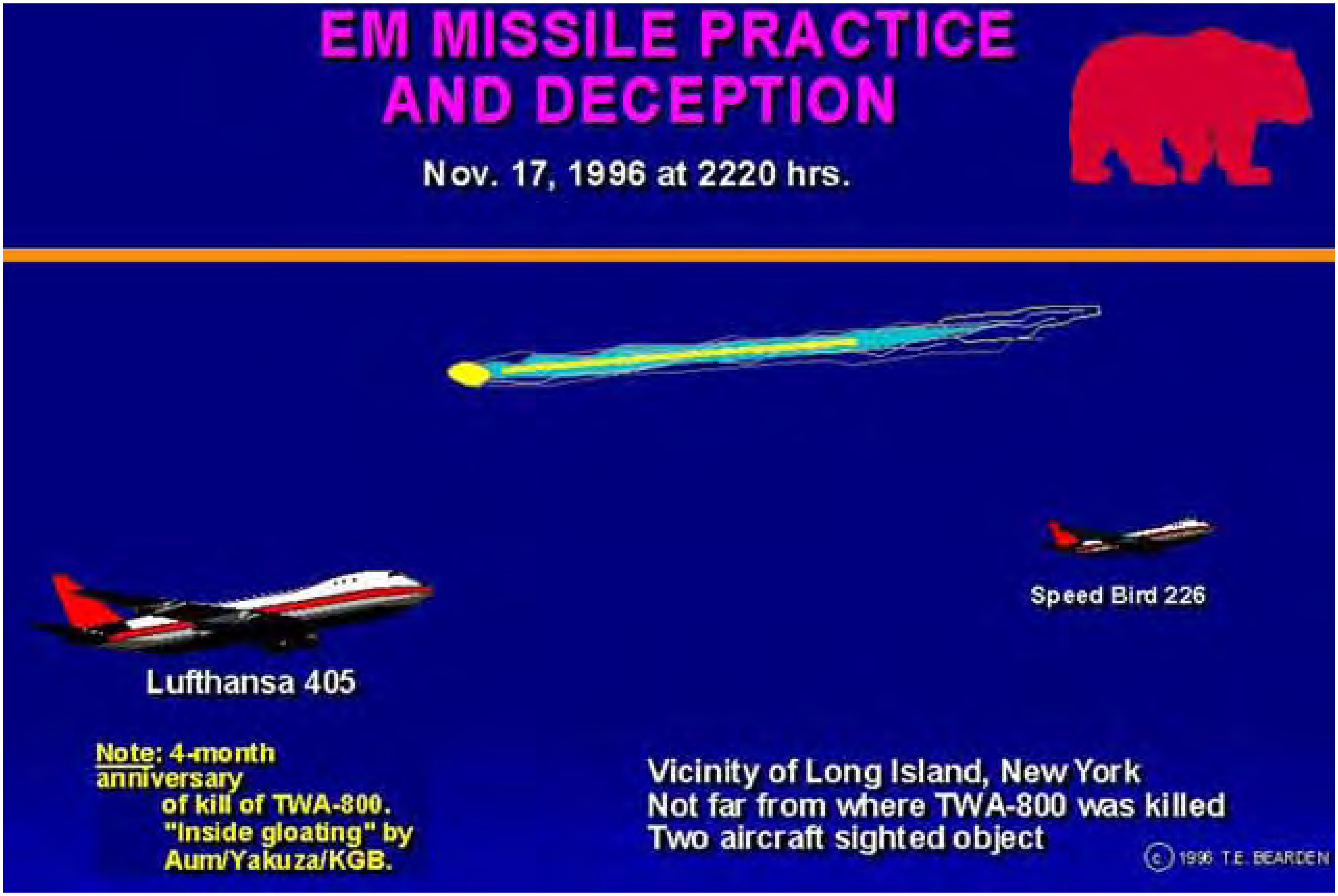

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Note: 4-month anniversary
of kill of TWA-800.
"Inside gloating" by
Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

© 1996 T.E. BEARDEN

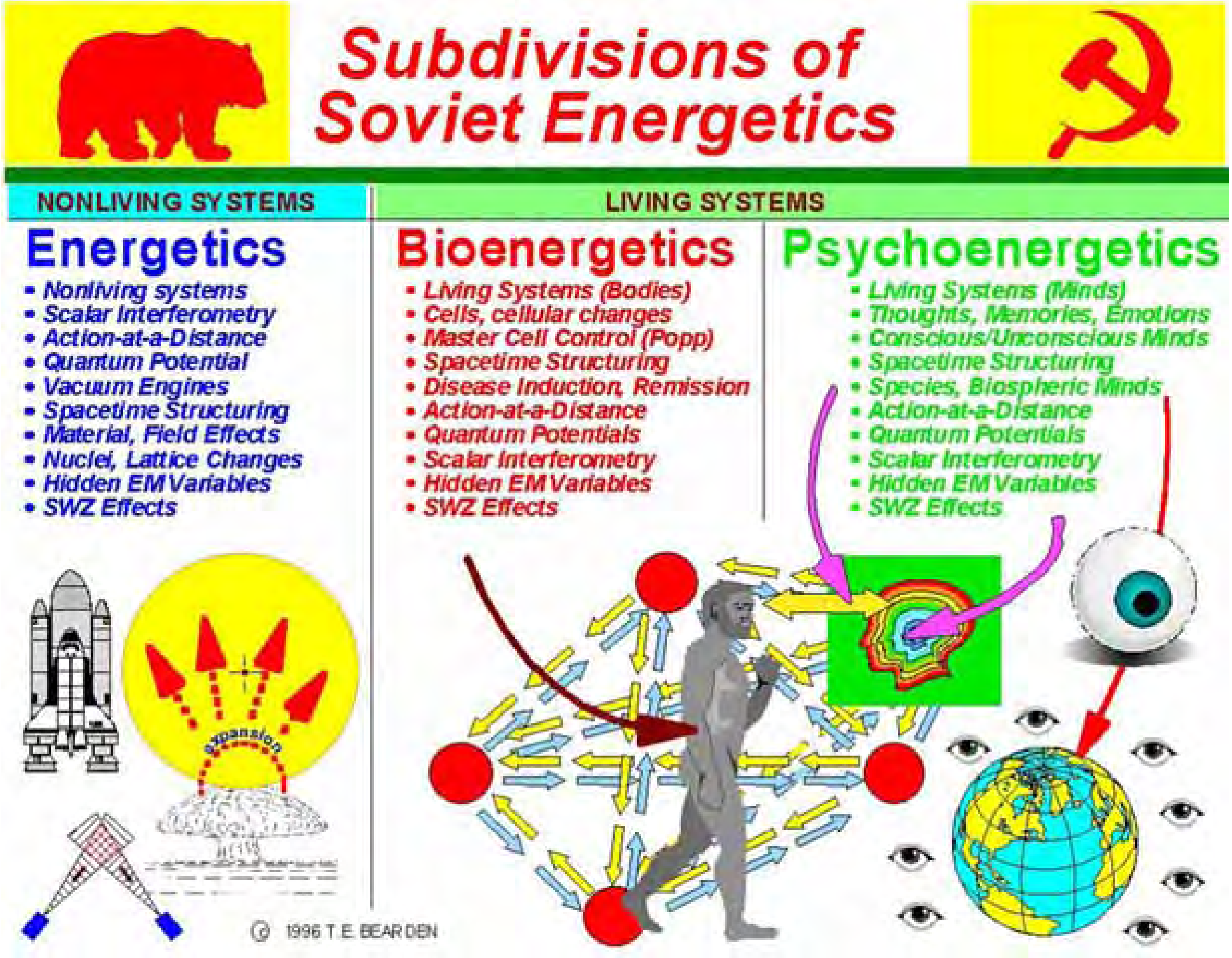

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

LIVING SYSTEMS

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T. E. BEARDEN

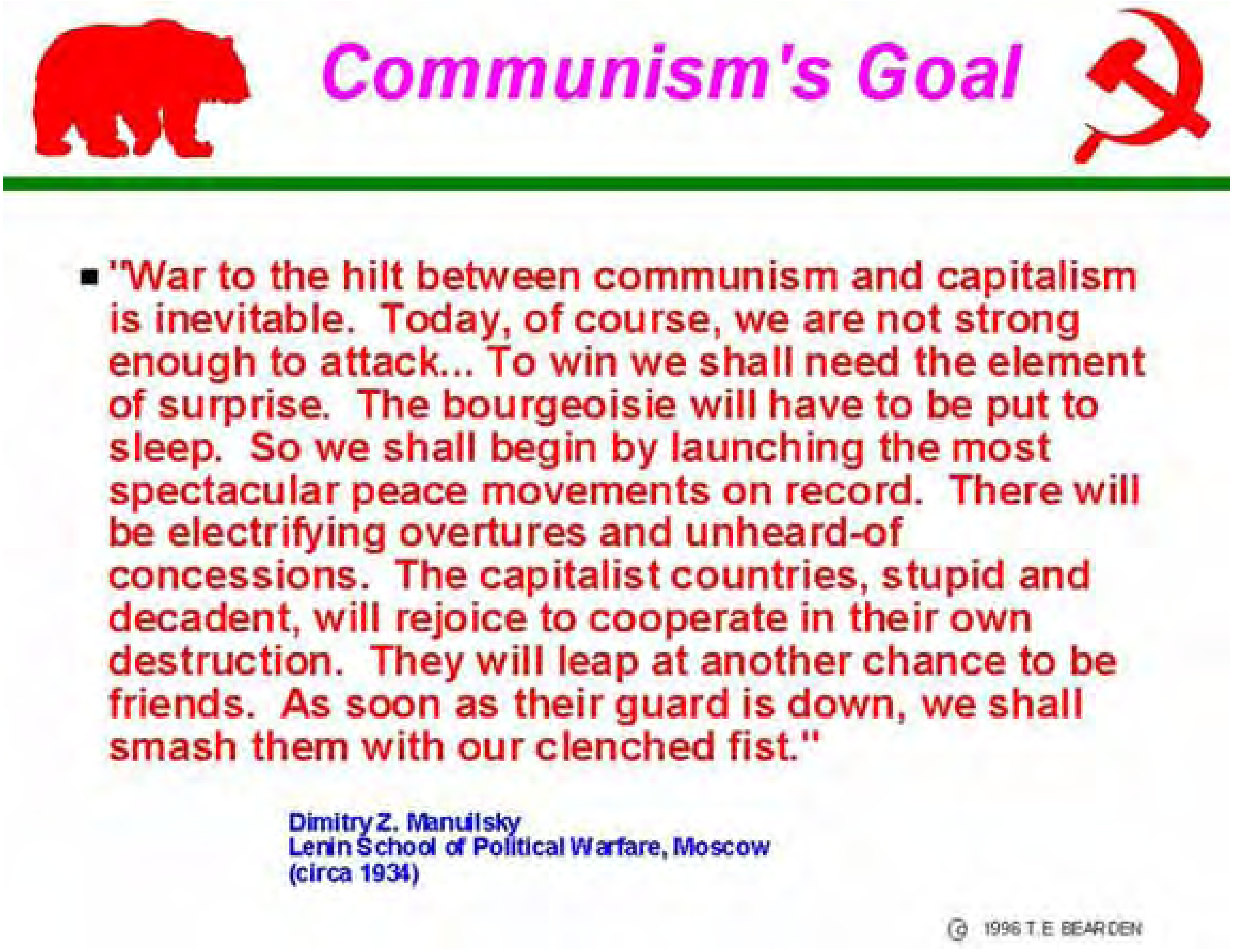

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

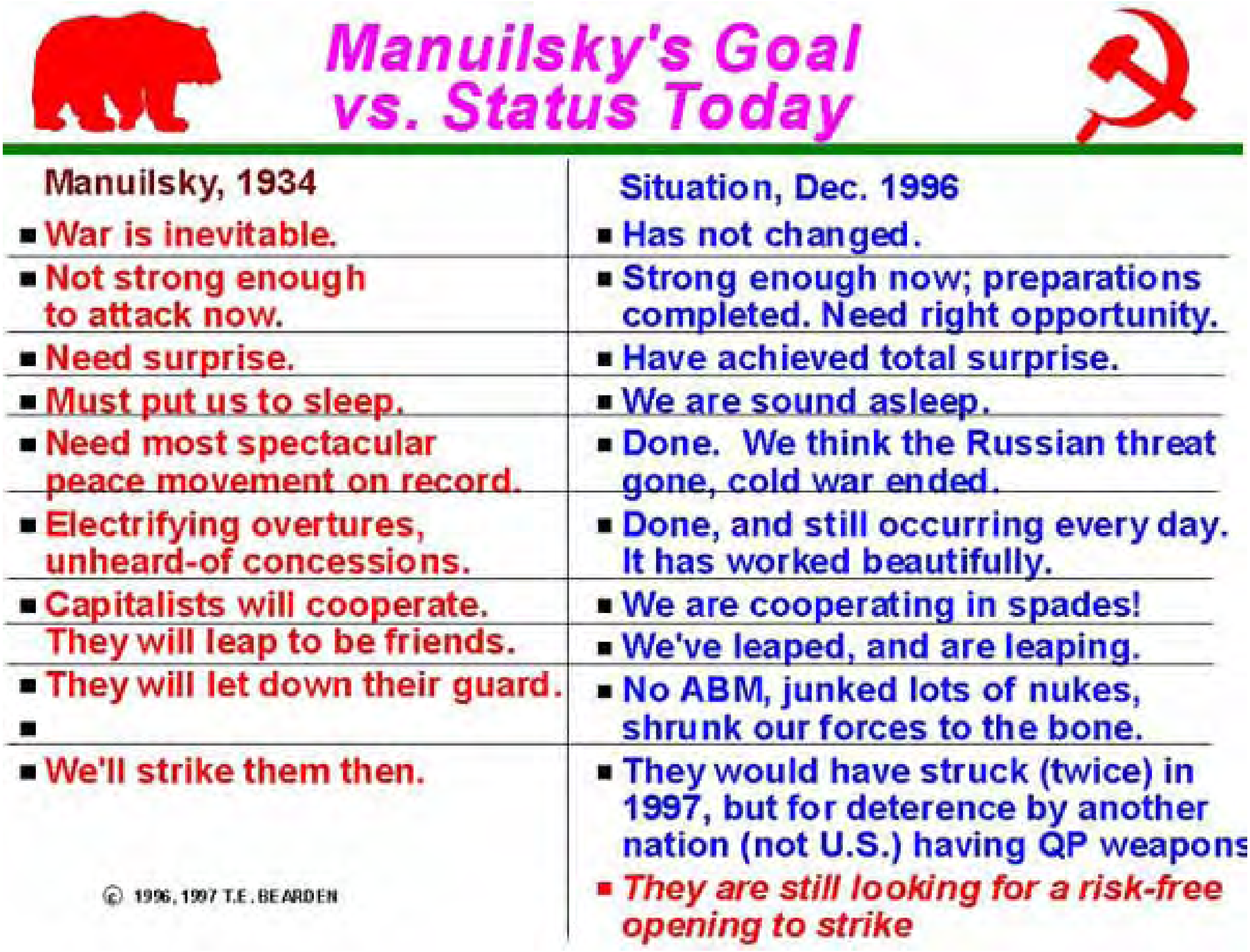

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
War is inevitable.	Has not changed.
Not strong enough to attack now.	Strong enough now; preparations completed. Need right opportunity.
Need surprise.	Have achieved total surprise.
Must put us to sleep.	We are sound asleep.
Need most spectacular peace movement on record.	Done. We think the Russian threat gone, cold war ended.
Electrifying overtures, unheard-of concessions.	Done, and still occurring every day. It has worked beautifully.
Capitalists will cooperate.	We are cooperating in spades!
They will leap to be friends.	We've leaped, and are leaping.
They will let down their guard. ■	No ABM, junked lots of nukes, shrunk our forces to the bone.
We'll strike them then.	They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons *They are still looking for a risk-free opening to strike*

© 1996, 1997 T.E. BEARDEN

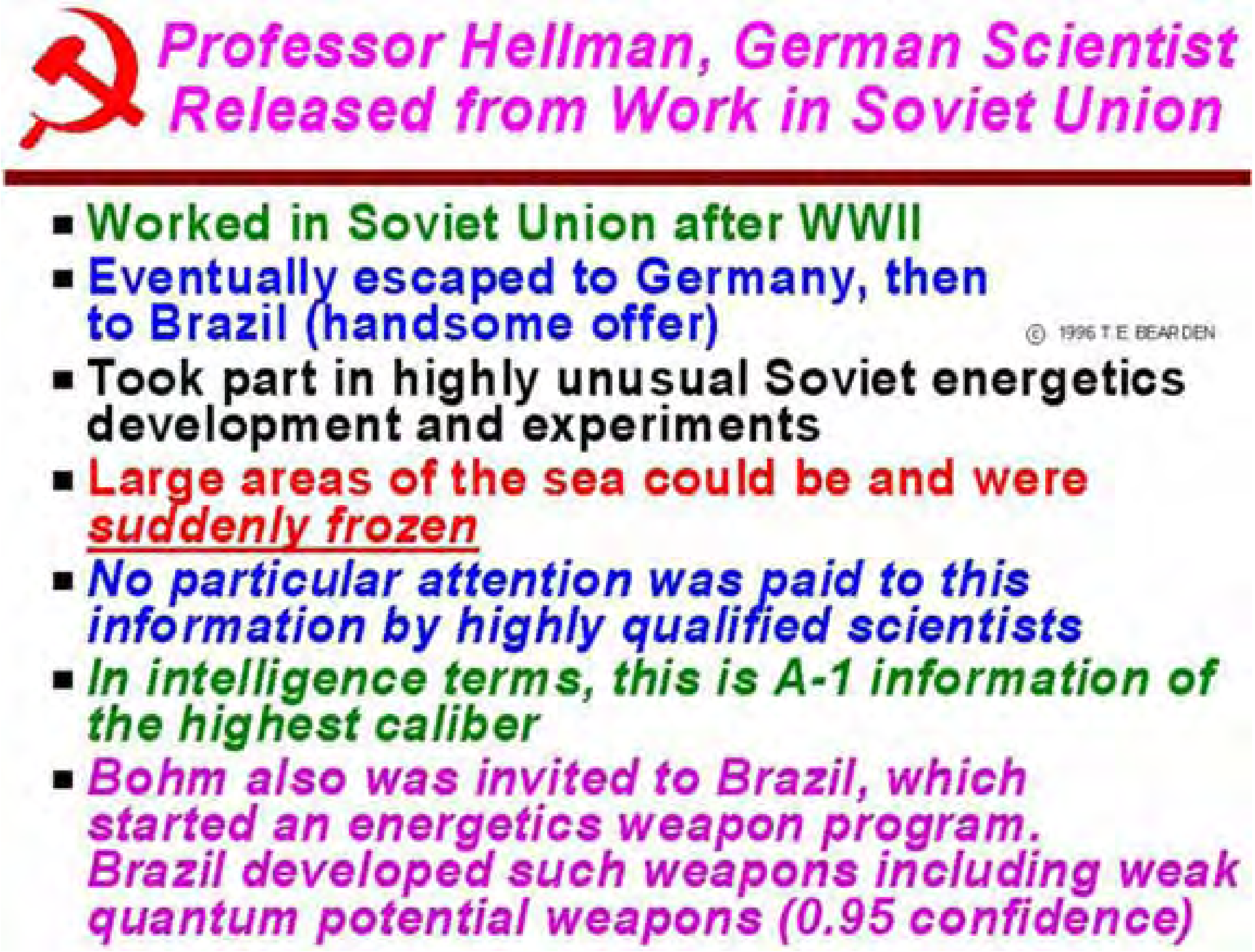

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)

© 1996 T. E. BEARDEN

- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were *suddenly frozen*
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

The Tom Bearden Website

Help support the research

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

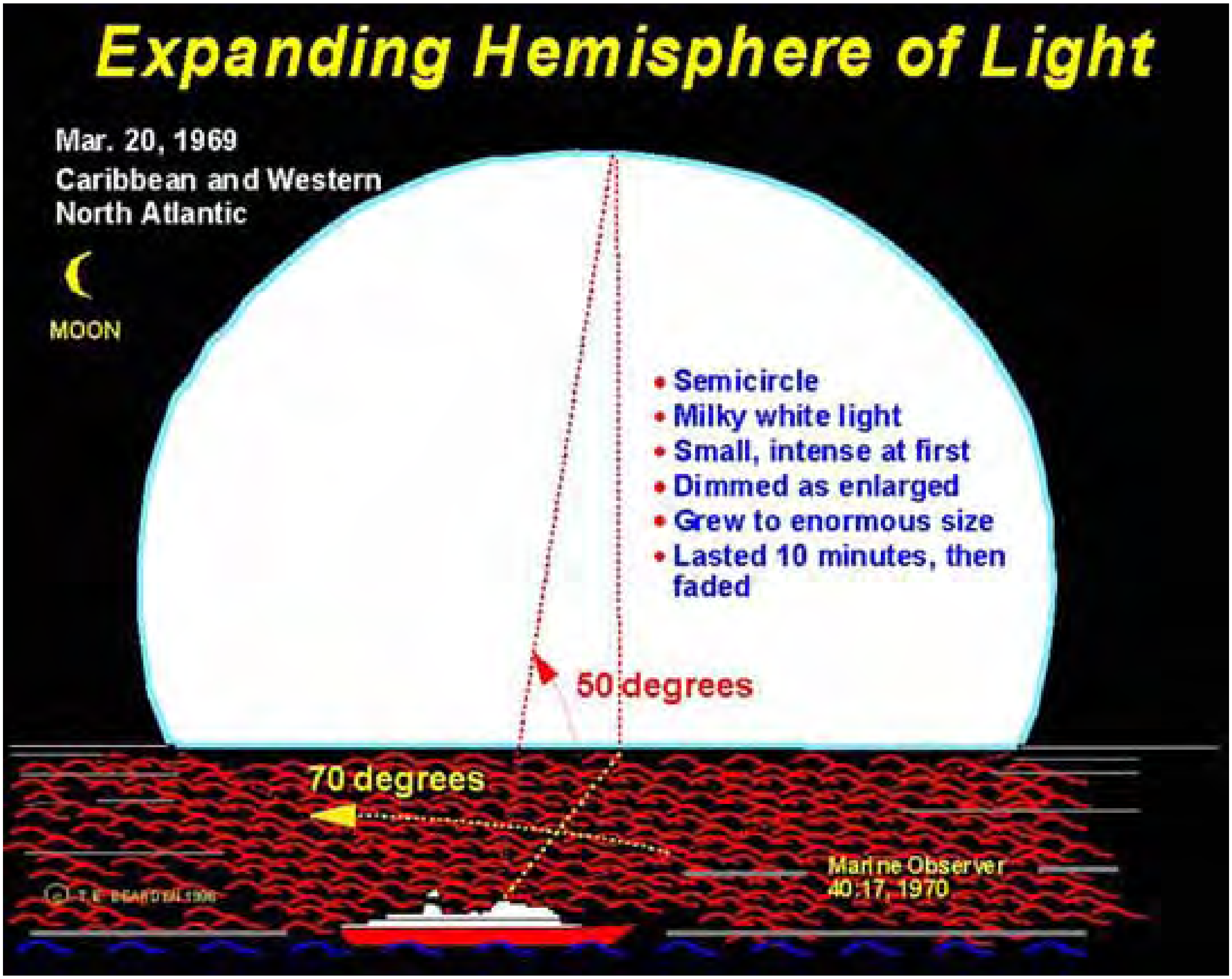

Expanding Hemisphere of Light

Mar. 20, 1969
Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

MOON
50 degrees
70 degrees
Marine Observer
40:17, 1970

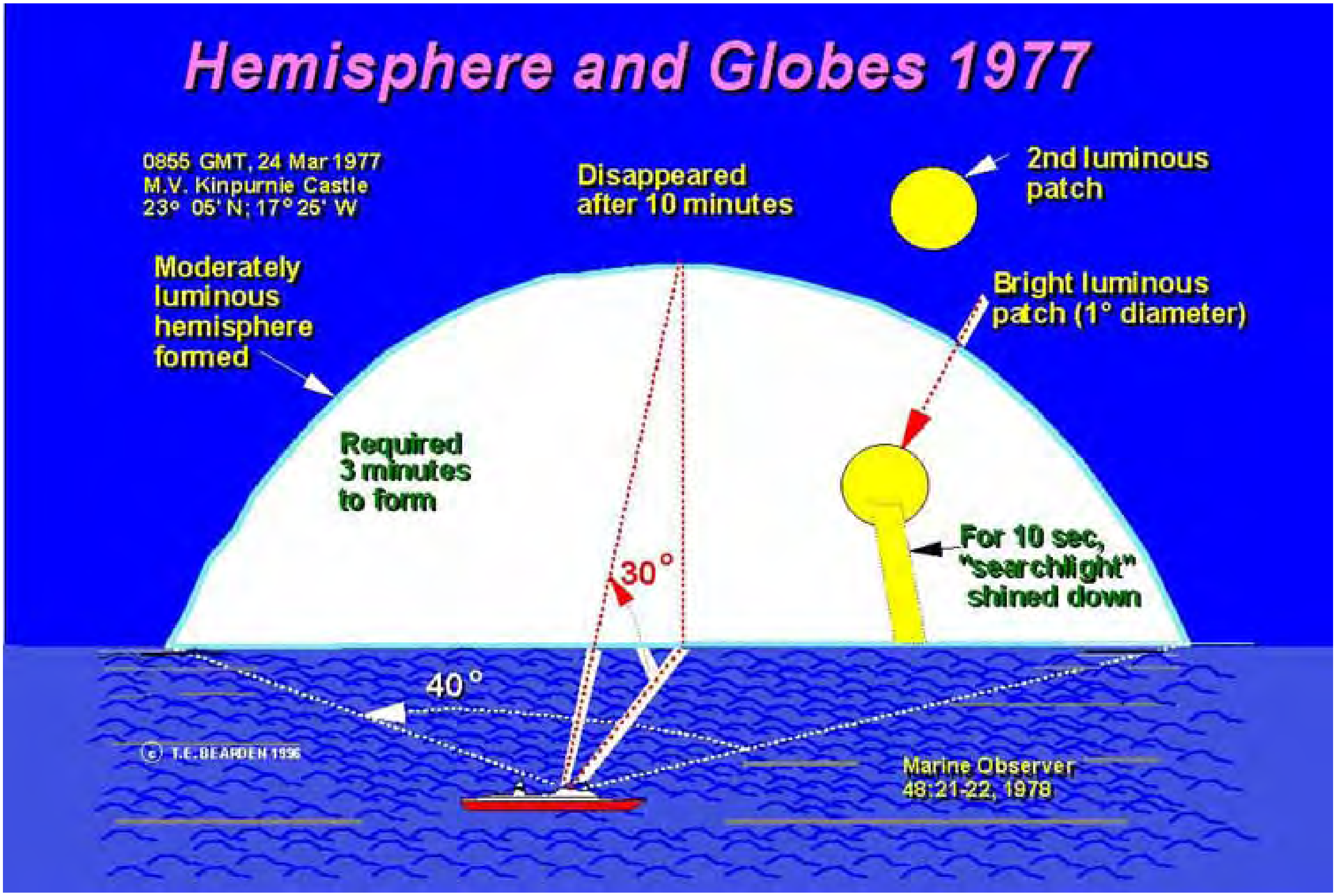
Hemisphere and Globes 1977
0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W
Disappeared
after 10 minutes
2nd luminous
patch
Moderately
luminous
hemisphere
formed
Bright luminous
patch (1° diameter)
Required
3 minutes
to form
For 10 sec,
"searchlight"
shined down
30°
40°
© T.E. BEARDEN 1996
Marine Observer
48:21-22, 1978

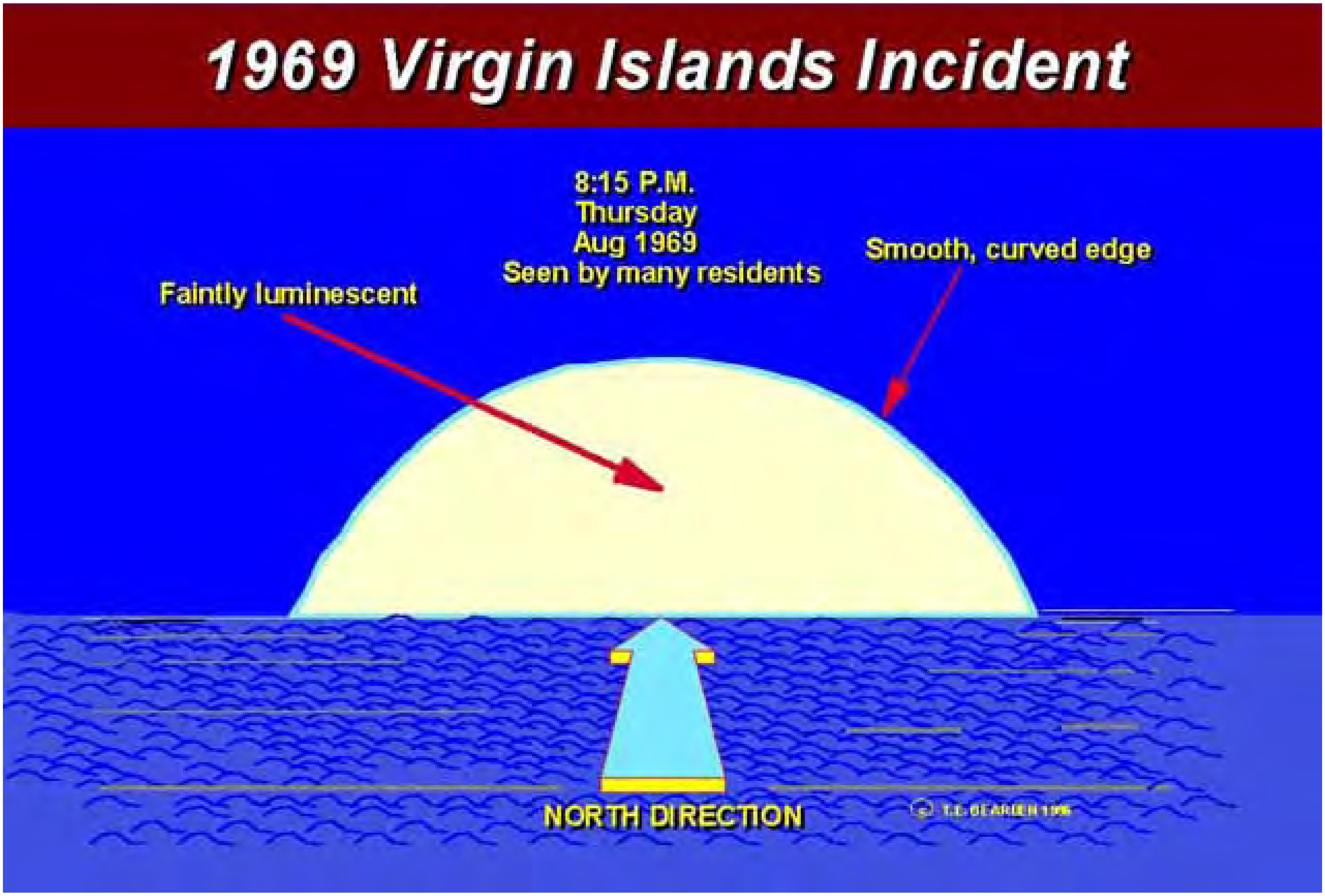

1969 Virgin Islands Incident

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent
Smooth, curved edge
NORTH DIRECTION

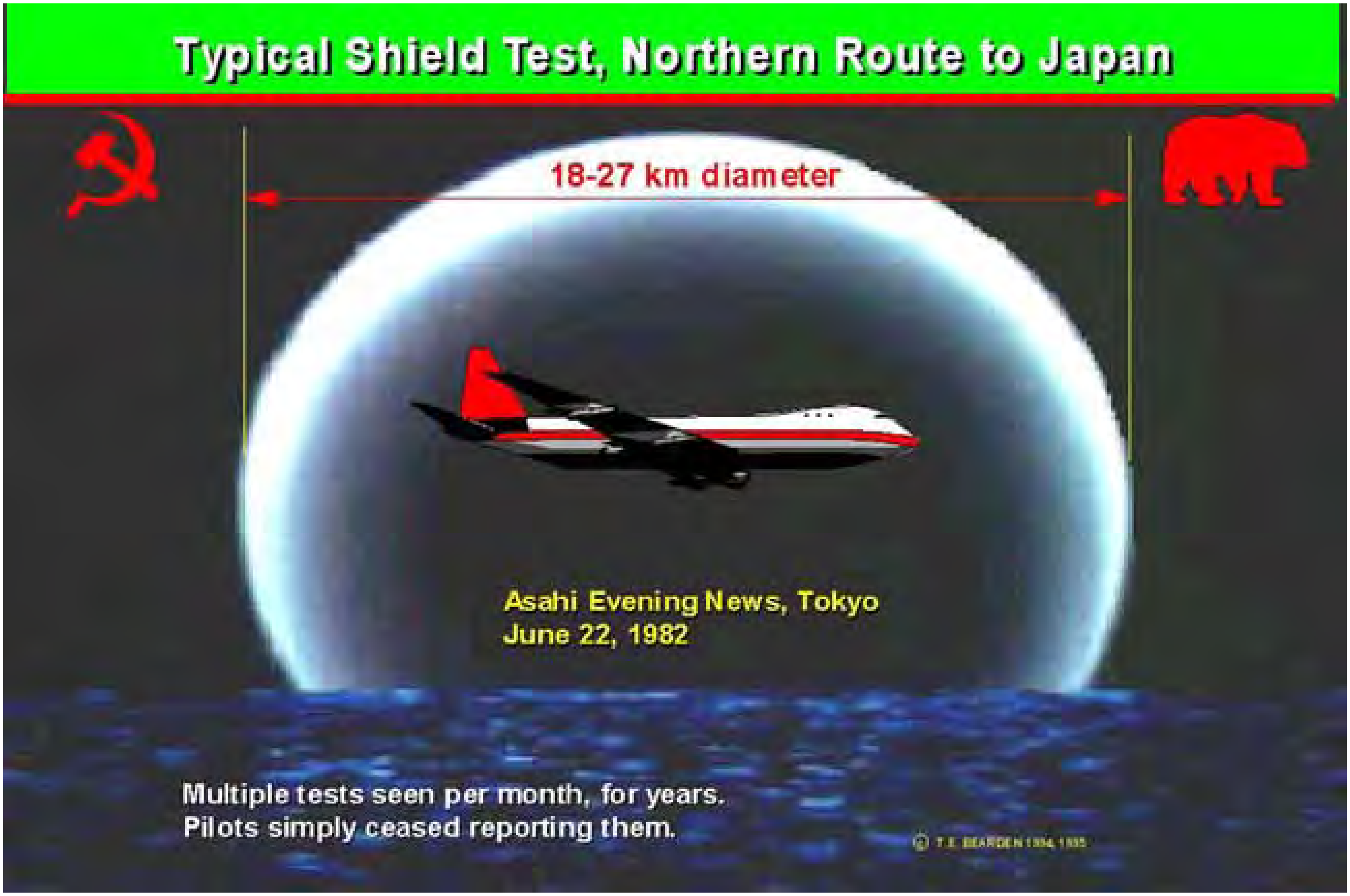

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

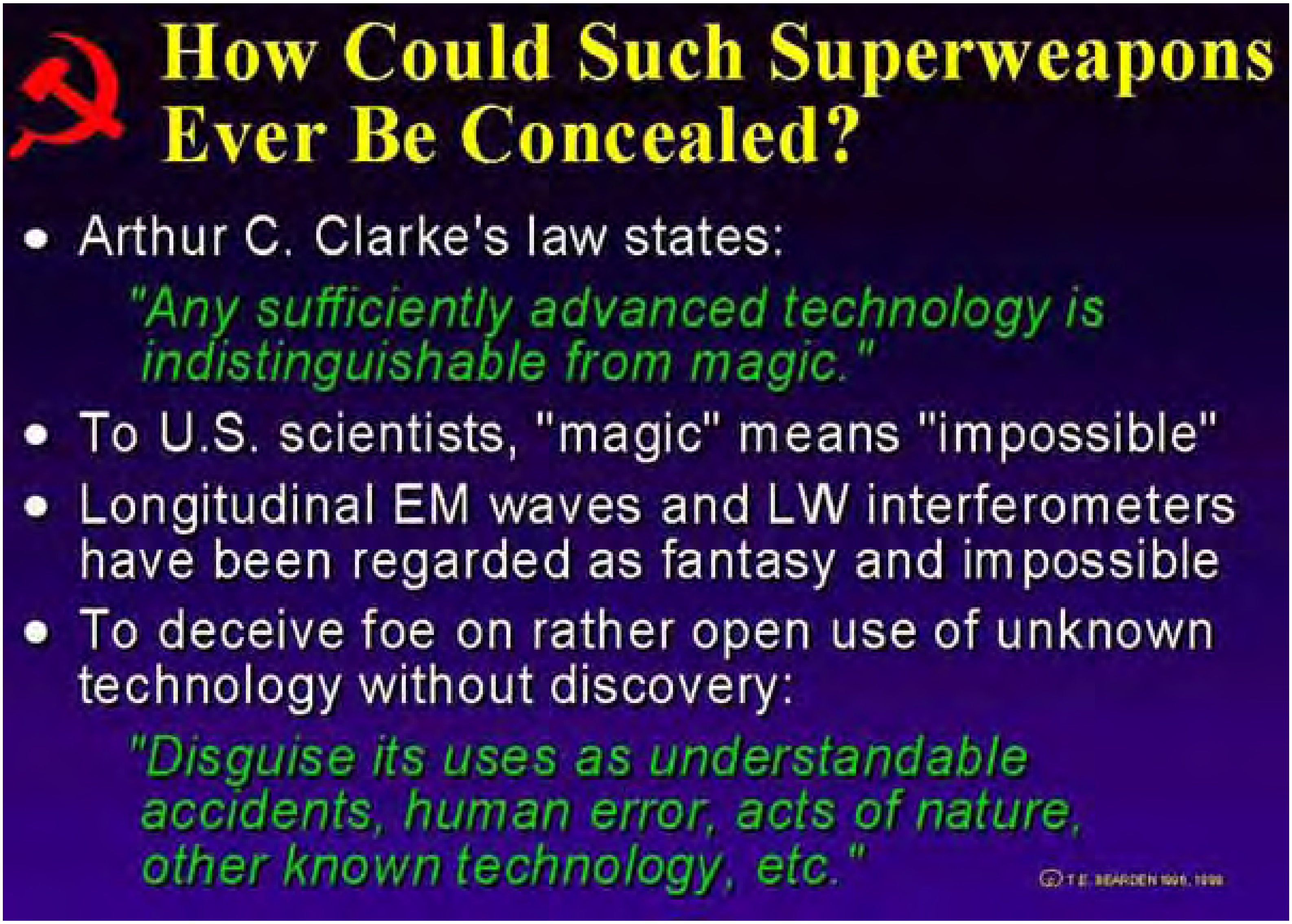

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."

- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN 1996, 1998

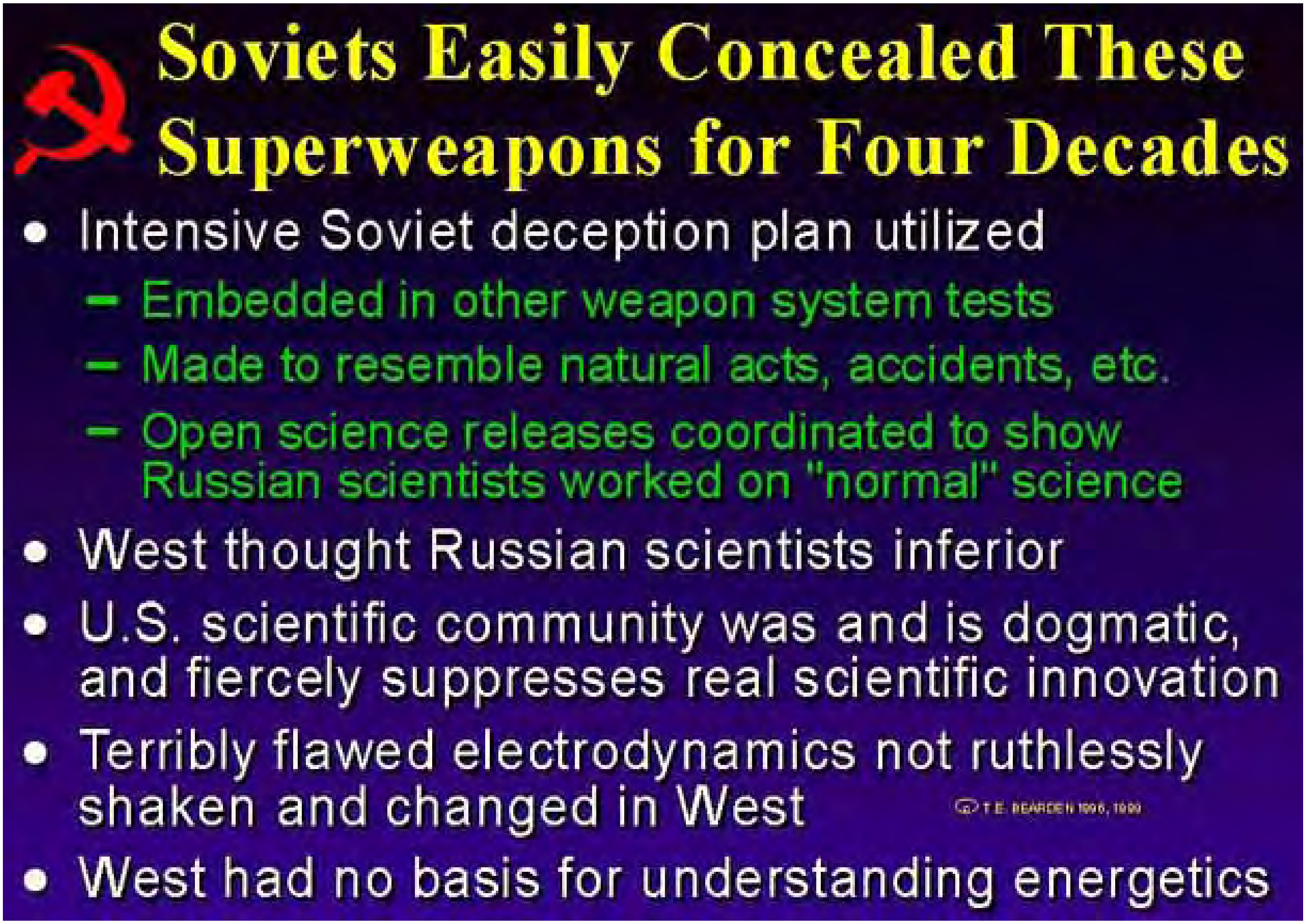

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

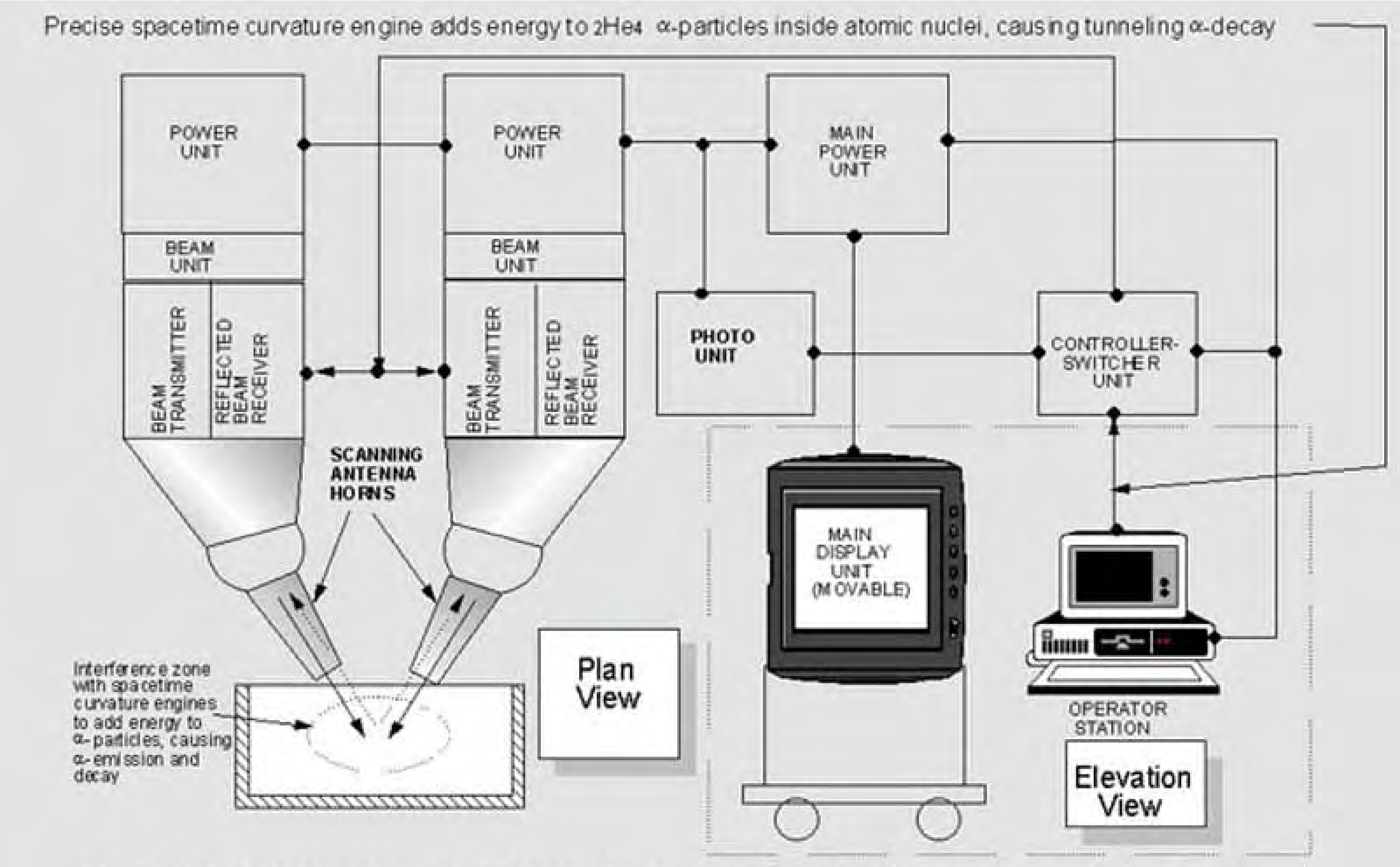
Precise spacetime curvature engine adds energy to 2He4 α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
POWER UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
MAIN POWER UNIT
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
OPERATOR STATION
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1996

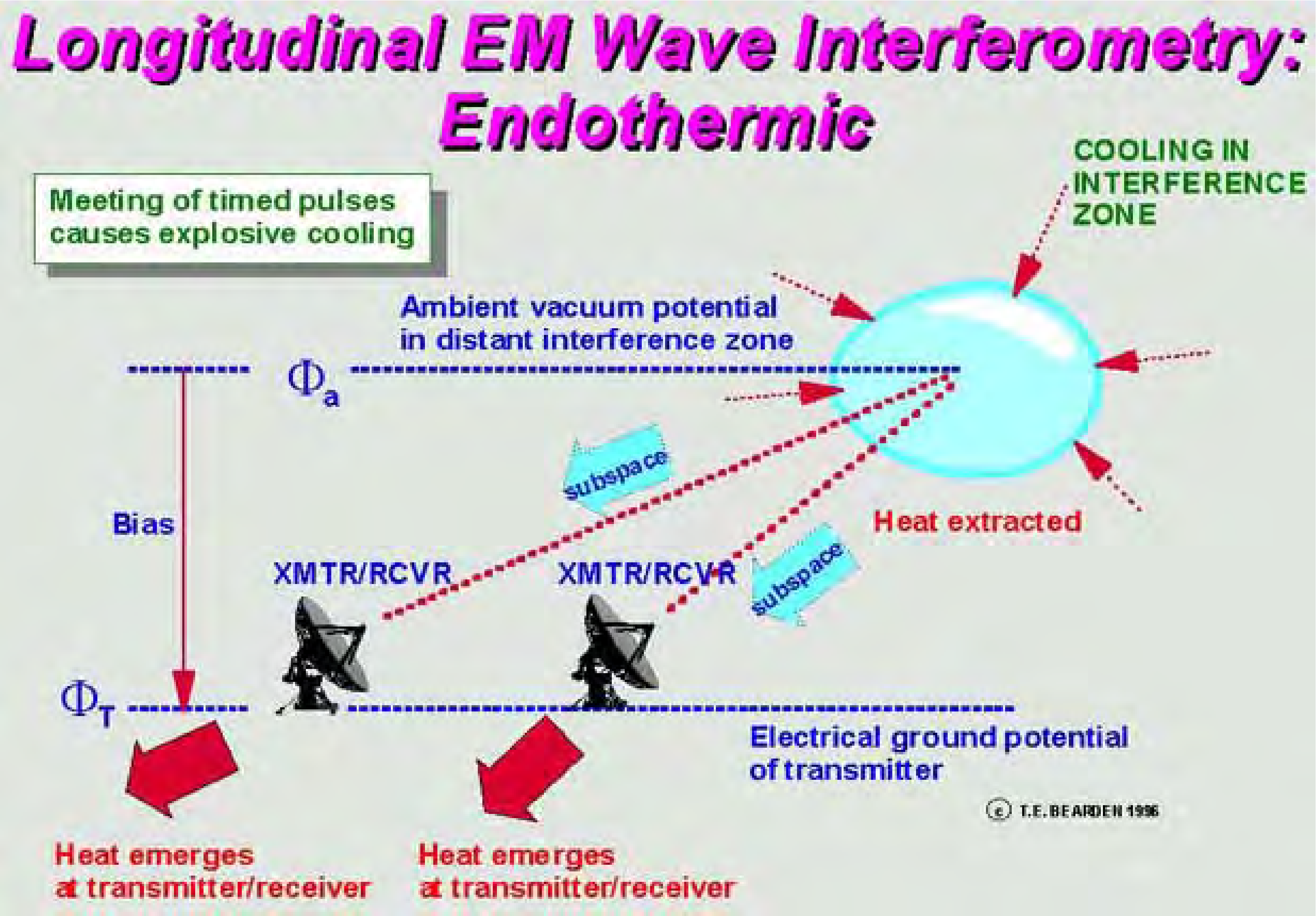

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses causes explosive cooling

COOLING IN INTERFERENCE ZONE

Ambient vacuum potential in distant interference zone

Φ_a

Bias

subspace

subspace

Heat extracted

XMTR/RCVR

XMTR/RCVR

Φ_T

Electrical ground potential of transmitter

Heat emerges at transmitter/receiver

Heat emerges at transmitter/receiver

© T.E. BEARDEN 1996

Longitudinal EM Wave Interferometry: Exothermic

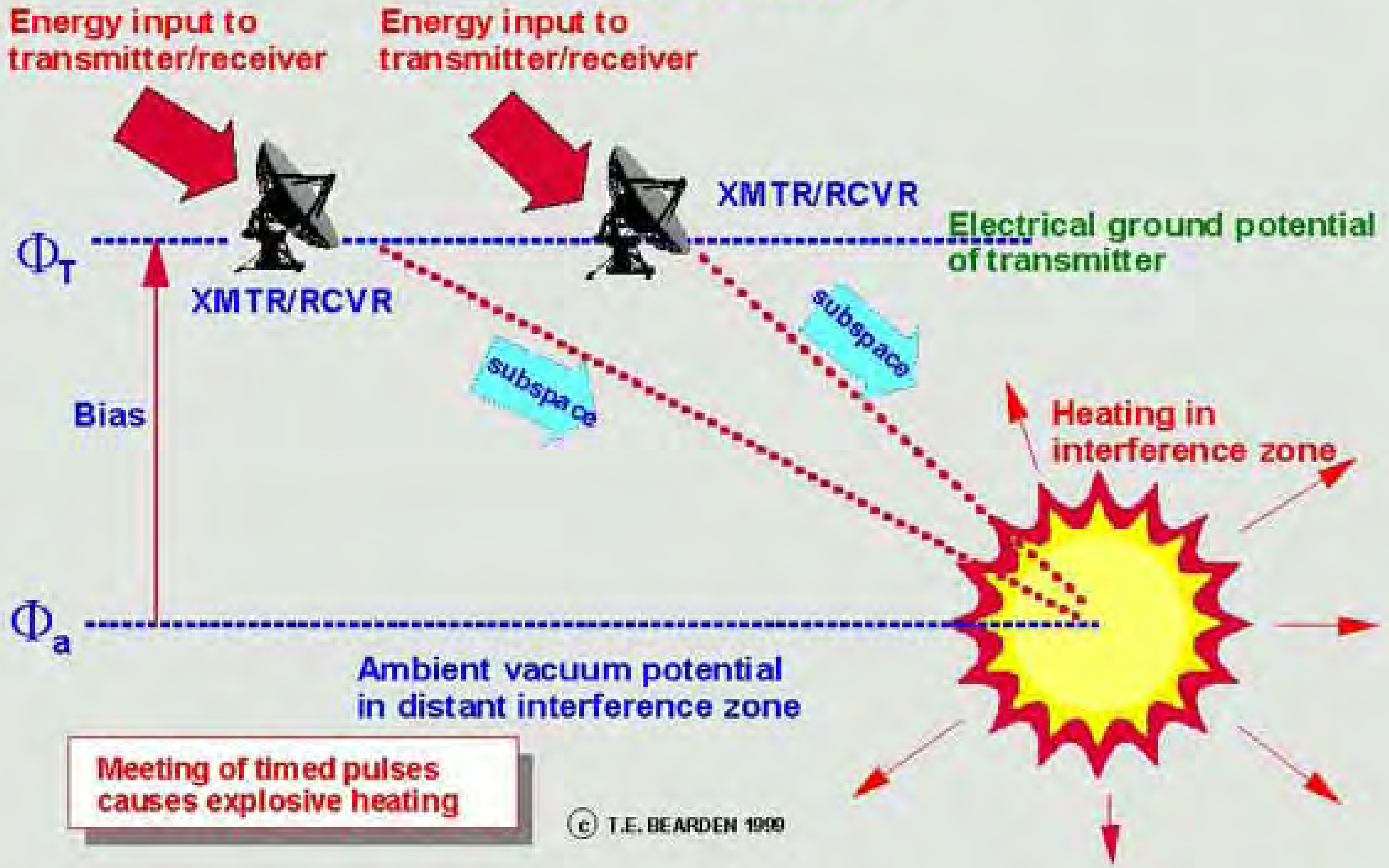

Energy input to transmitter/receiver

Energy input to transmitter/receiver

XMTR/RCVR

XMTR/RCVR

Φ_T

Electrical ground potential of transmitter

subspace

subspace

Bias

Heating in interference zone

Φ_a

Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

© T.E. BEARDEN 1999

[The Tom Bearden Website](#)

Help support the research

U.S. Air Attack on Libya

April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

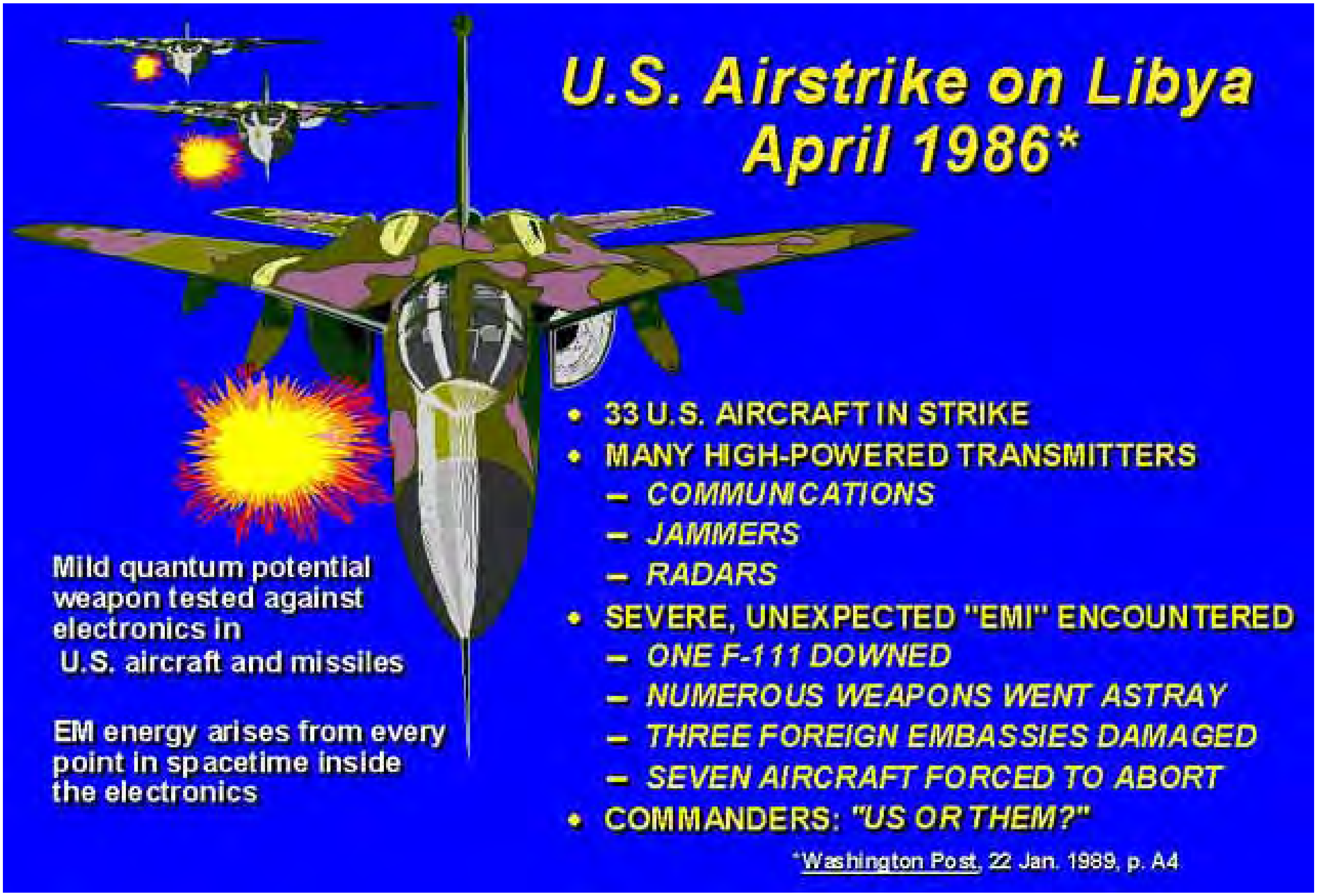

U.S. Airstrike on Libya April 1986*

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

*Washington Post, 22 Jan. 1989, p. A4

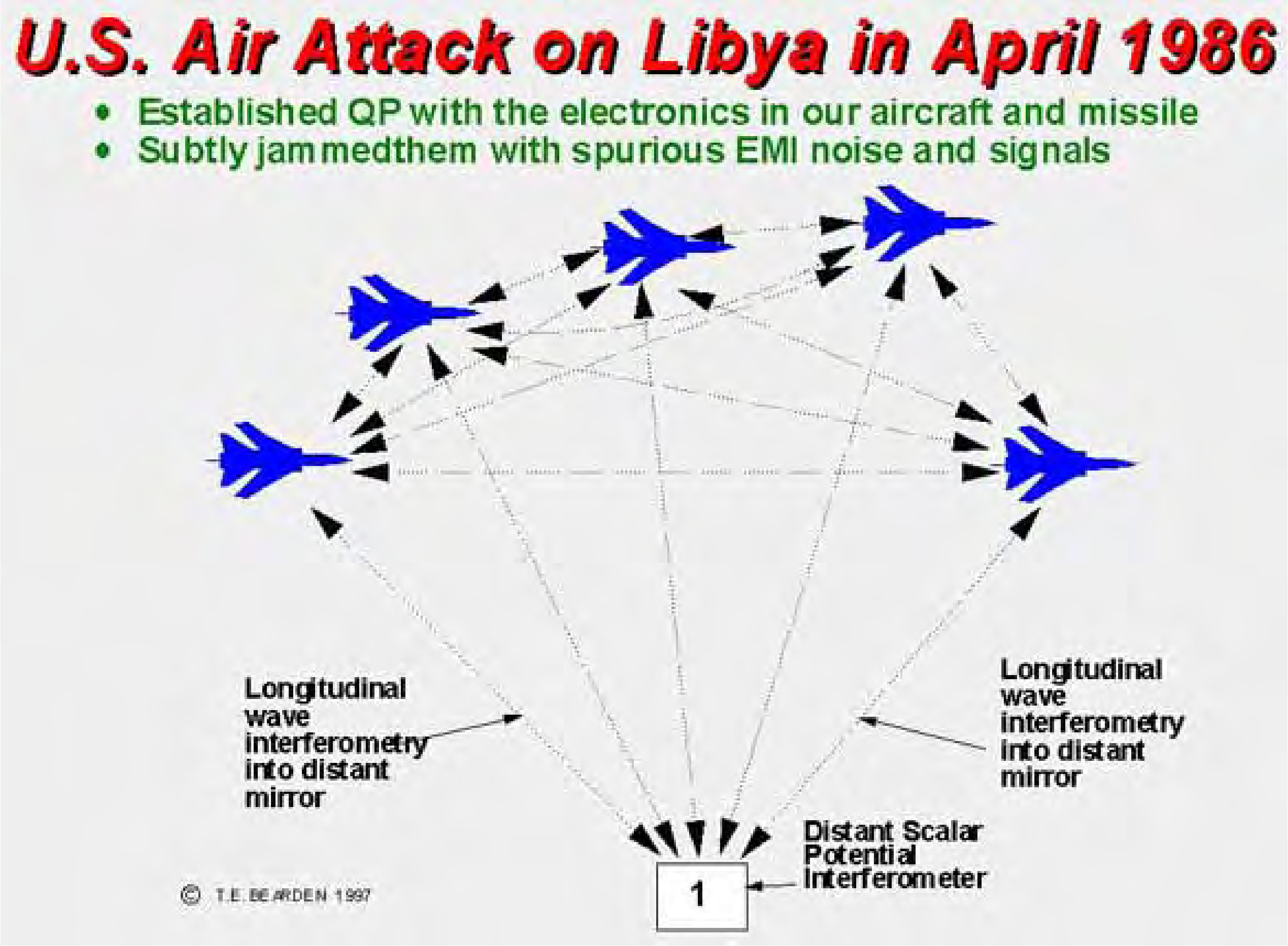

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal wave interferometry into distant mirror
Longitudinal wave interferometry into distant mirror
Distant Scalar Potential Interferometer
1
© T.E. BEARDEN 1997

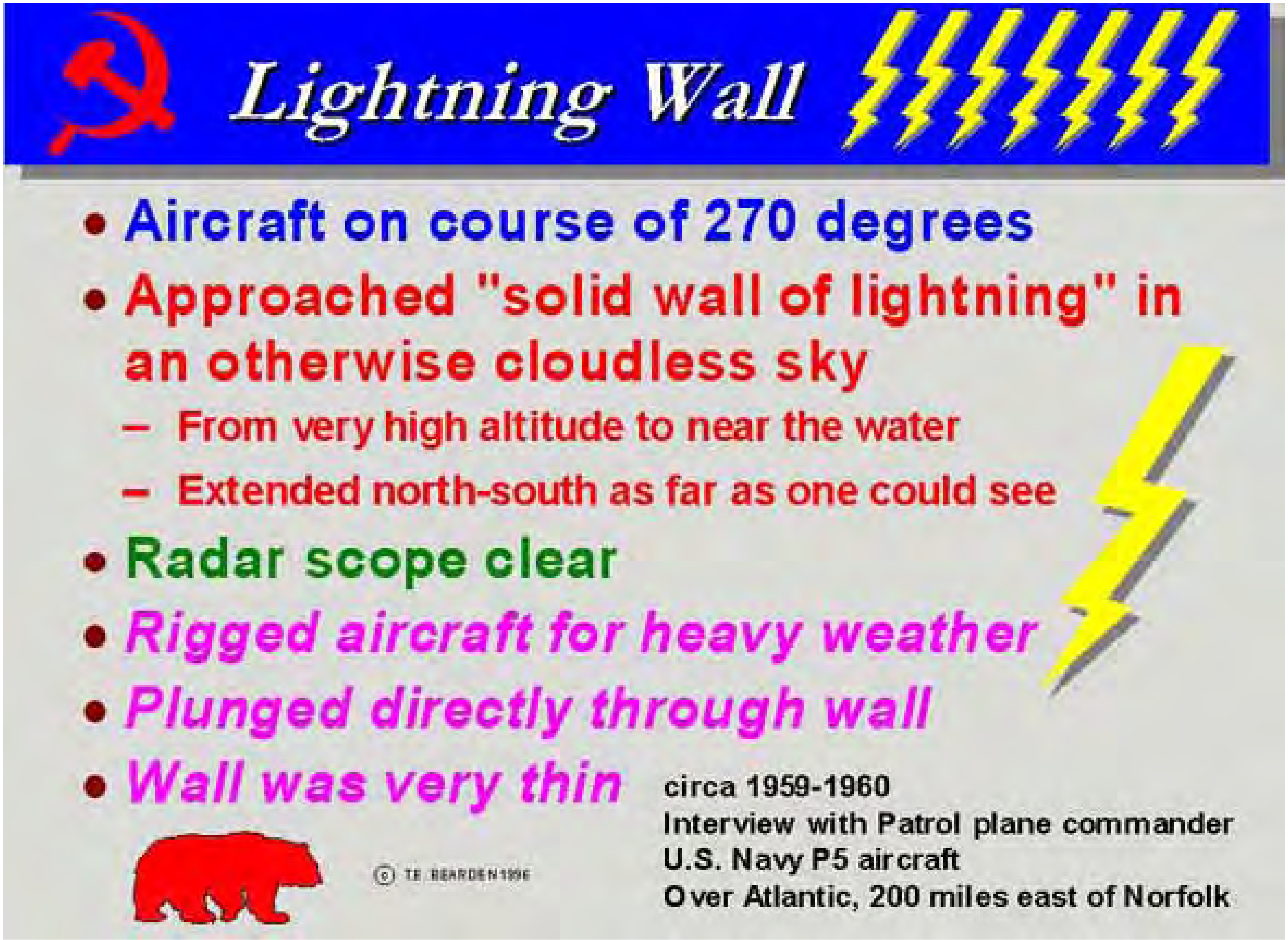

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

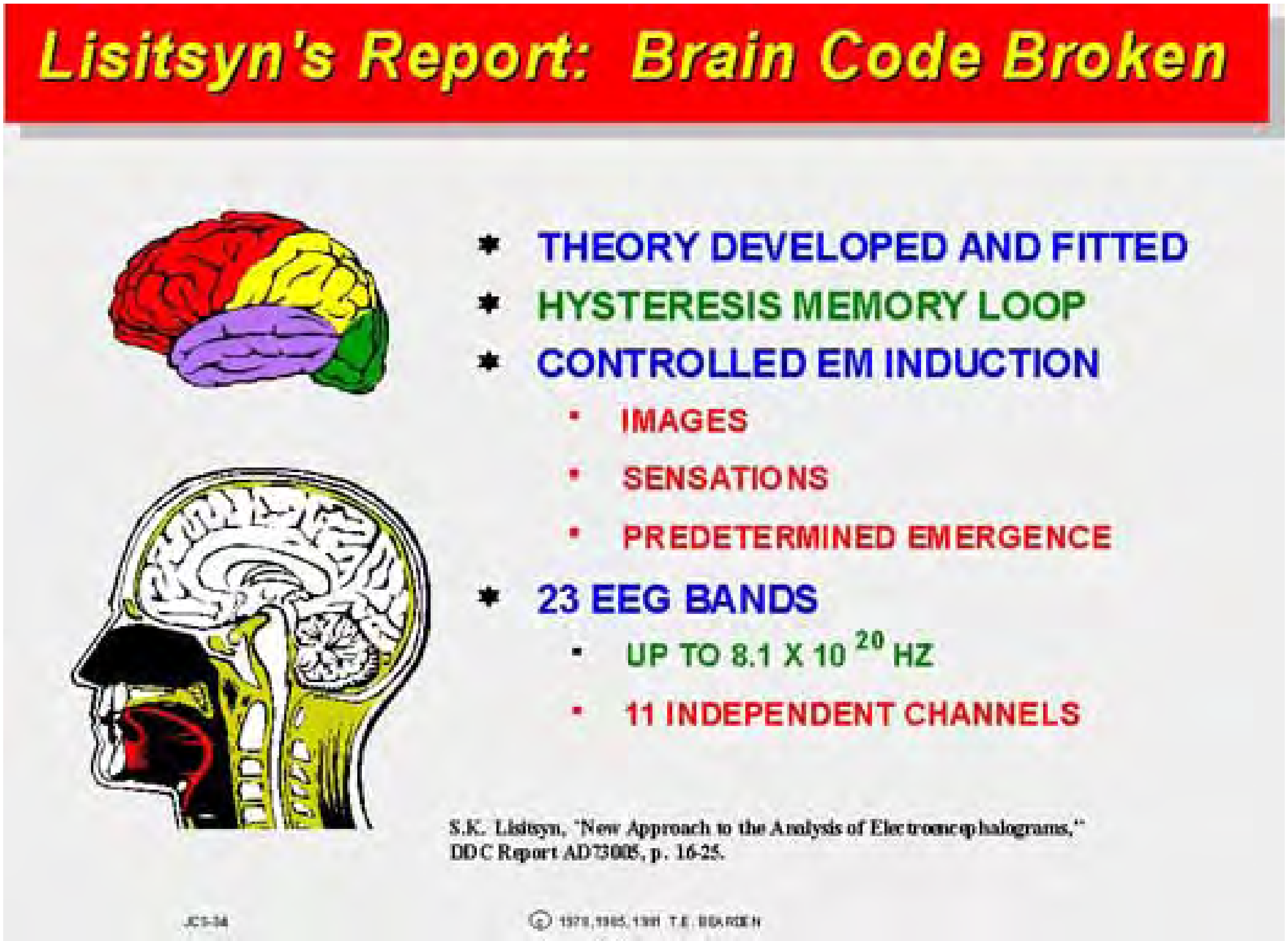

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms,"
DDC Report AD73005, p. 16-25.

JCS-34

© 1976, 1985, 1991 T.E. BEARDEN

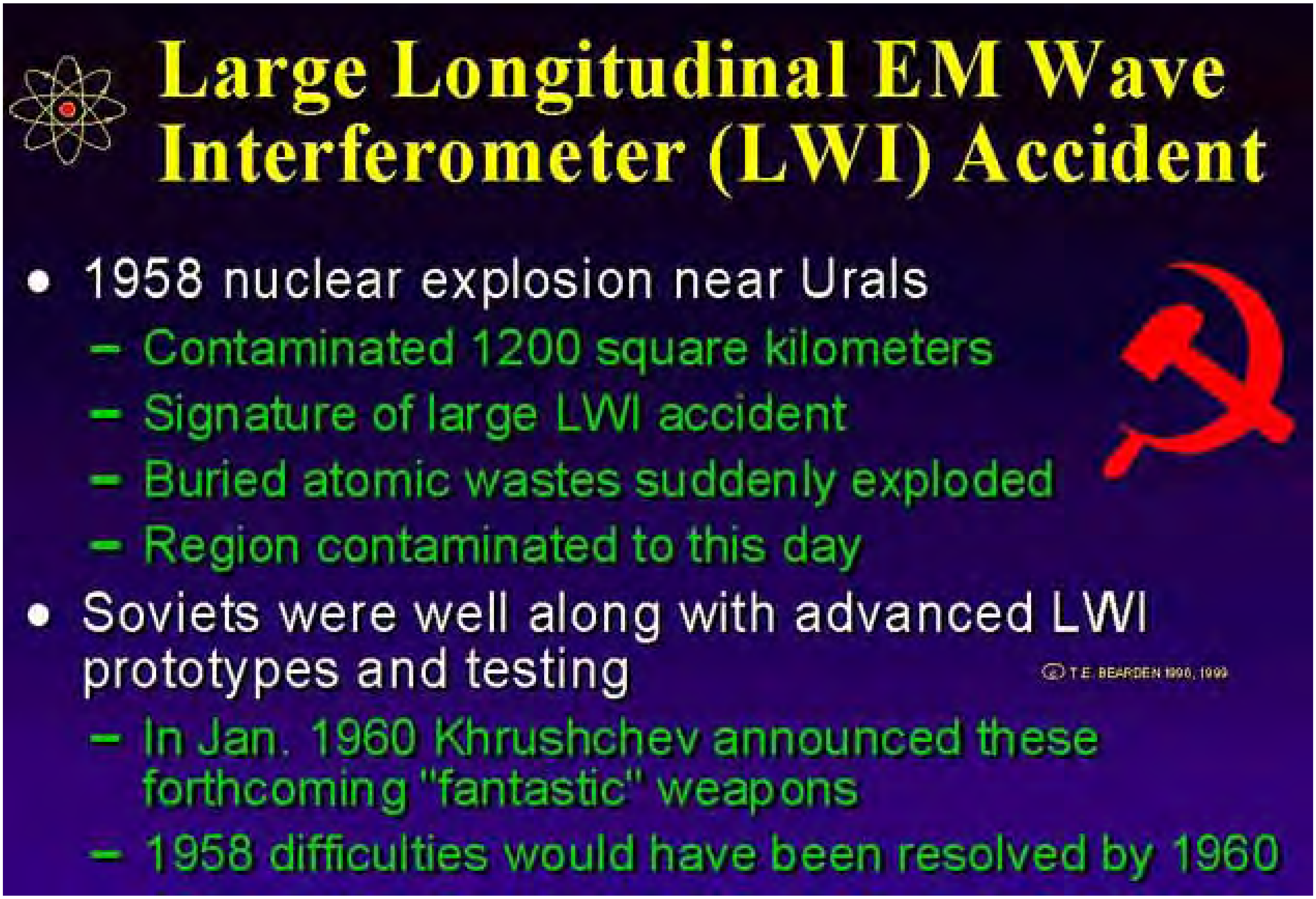

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

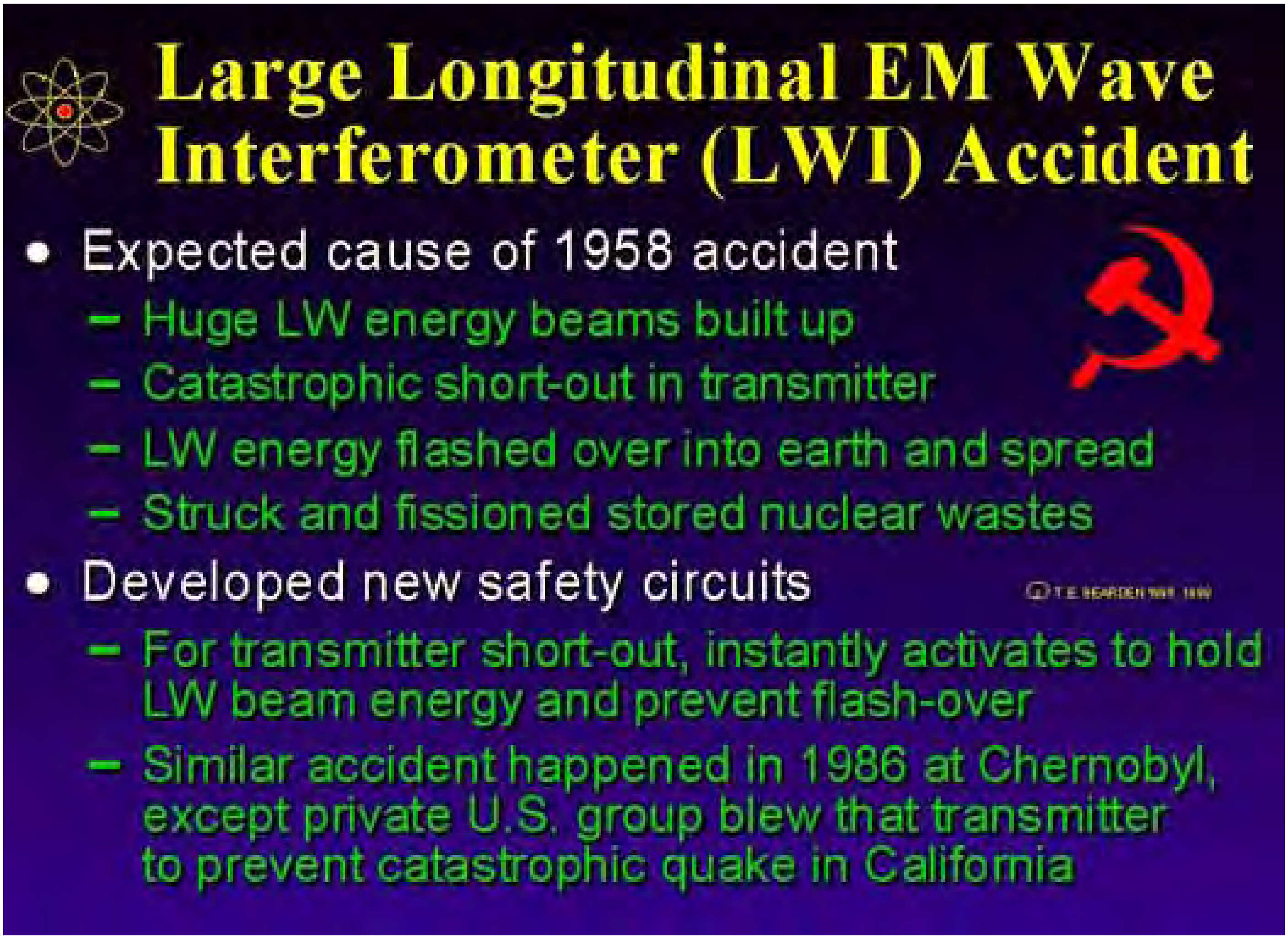

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

© T.E. BEARDEN [illegible]

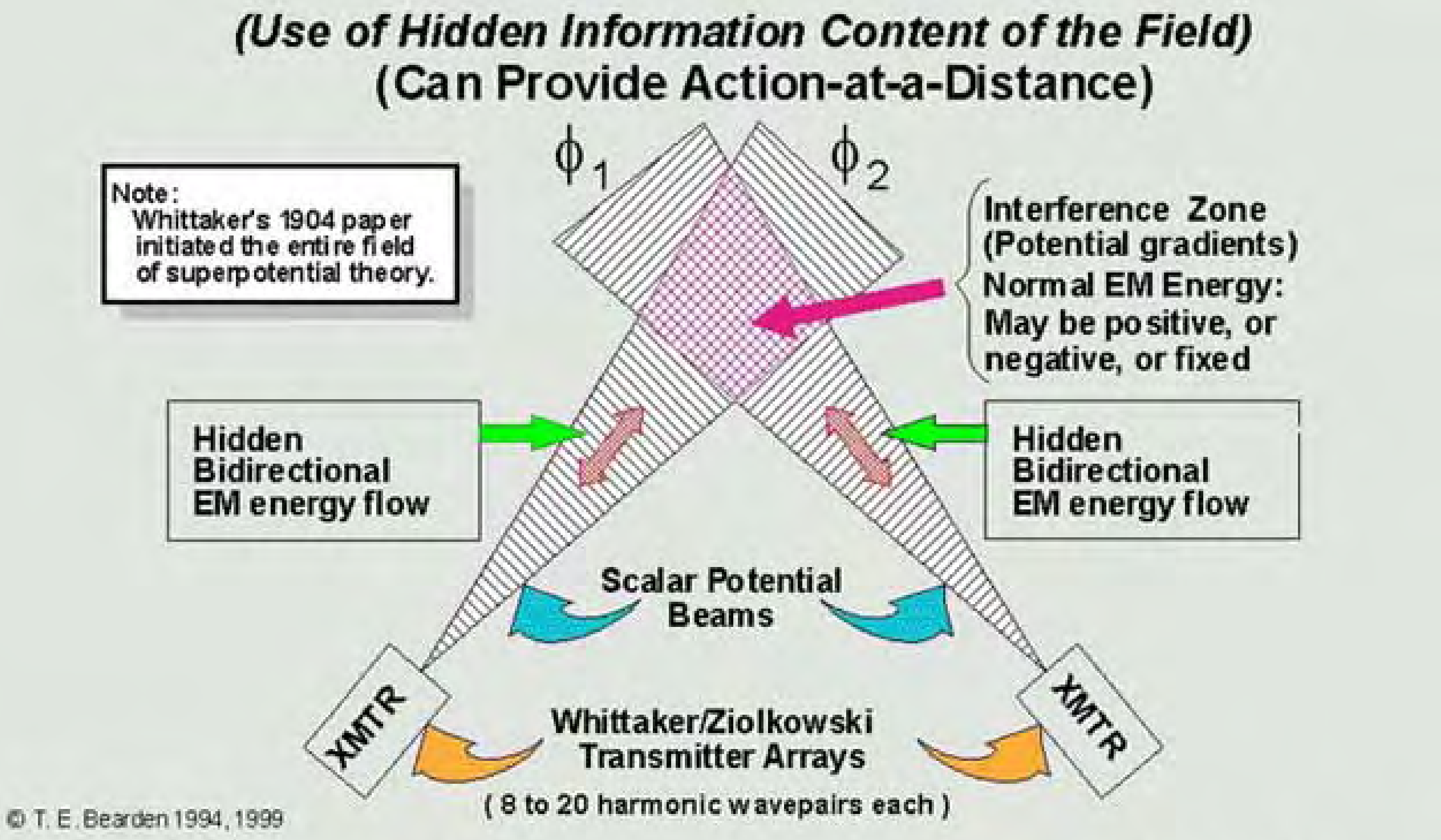
(Use of Hidden Information Content of the Field)
(Can Provide Action-at-a-Distance)
ϕ_1
ϕ_2
Note: Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

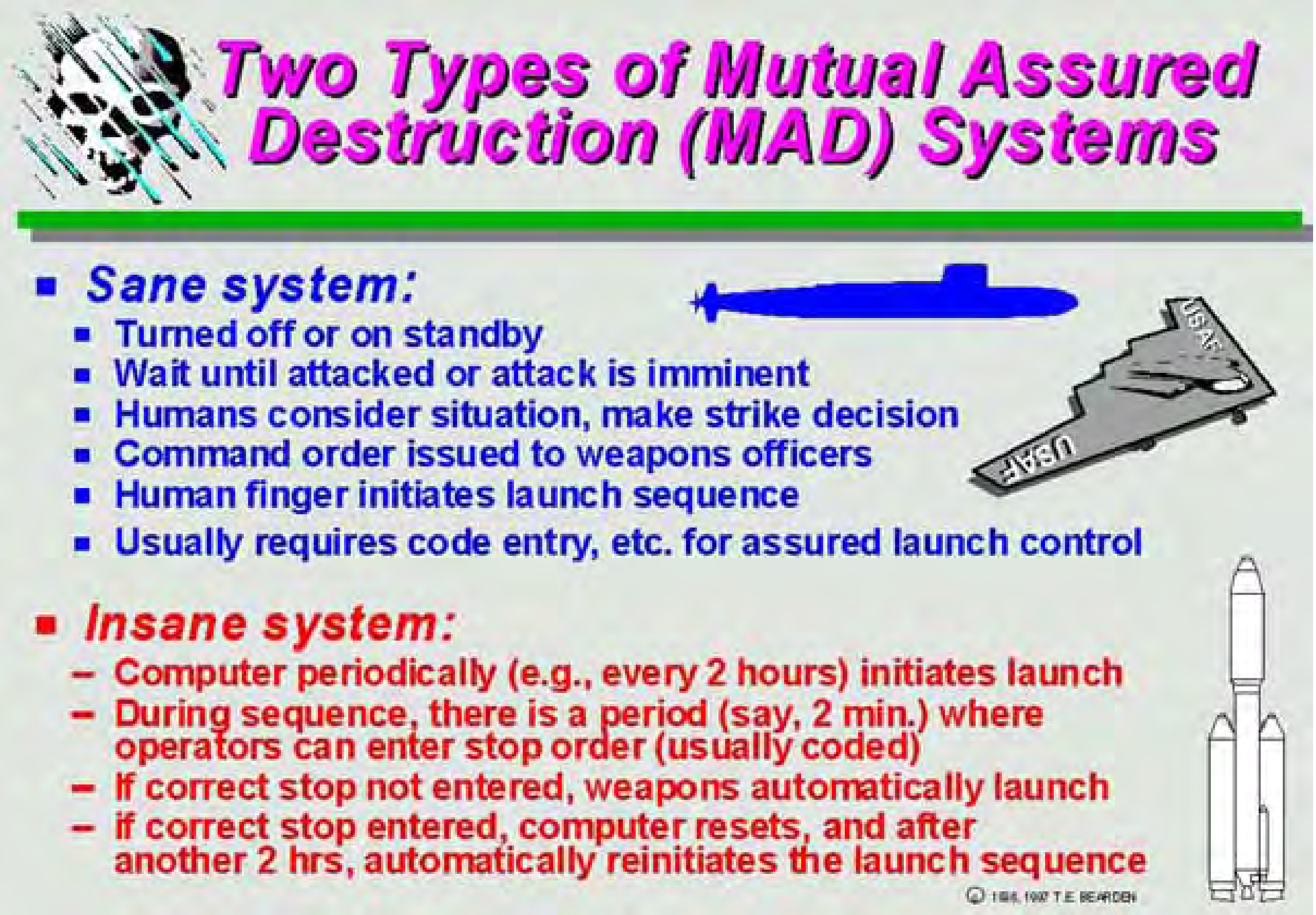

Two Types of Mutual Assured Destruction (MAD) Systems

- **Sane system:**
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- **Insane system:**
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

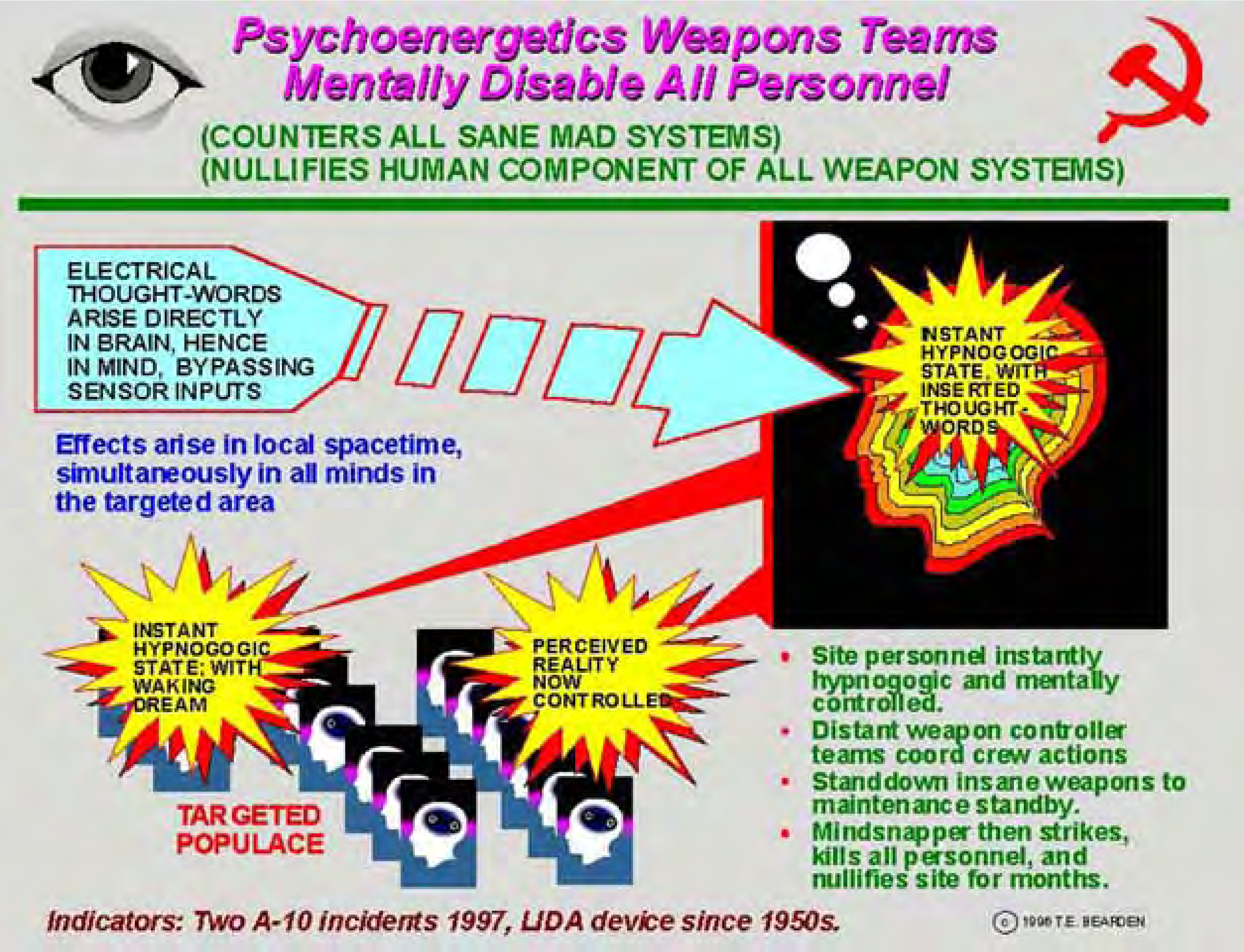

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

© 1998 T.E. BEARDEN

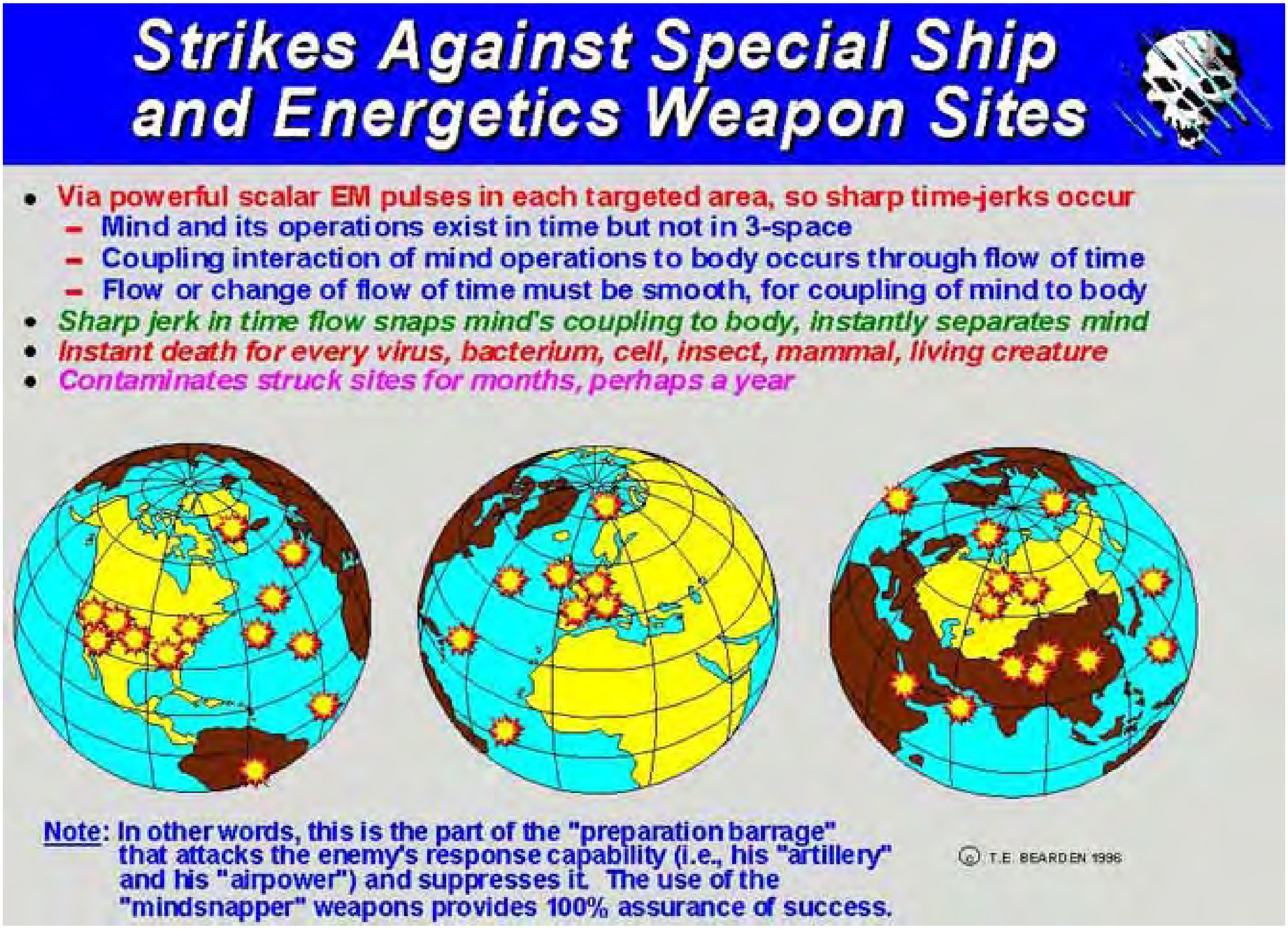

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

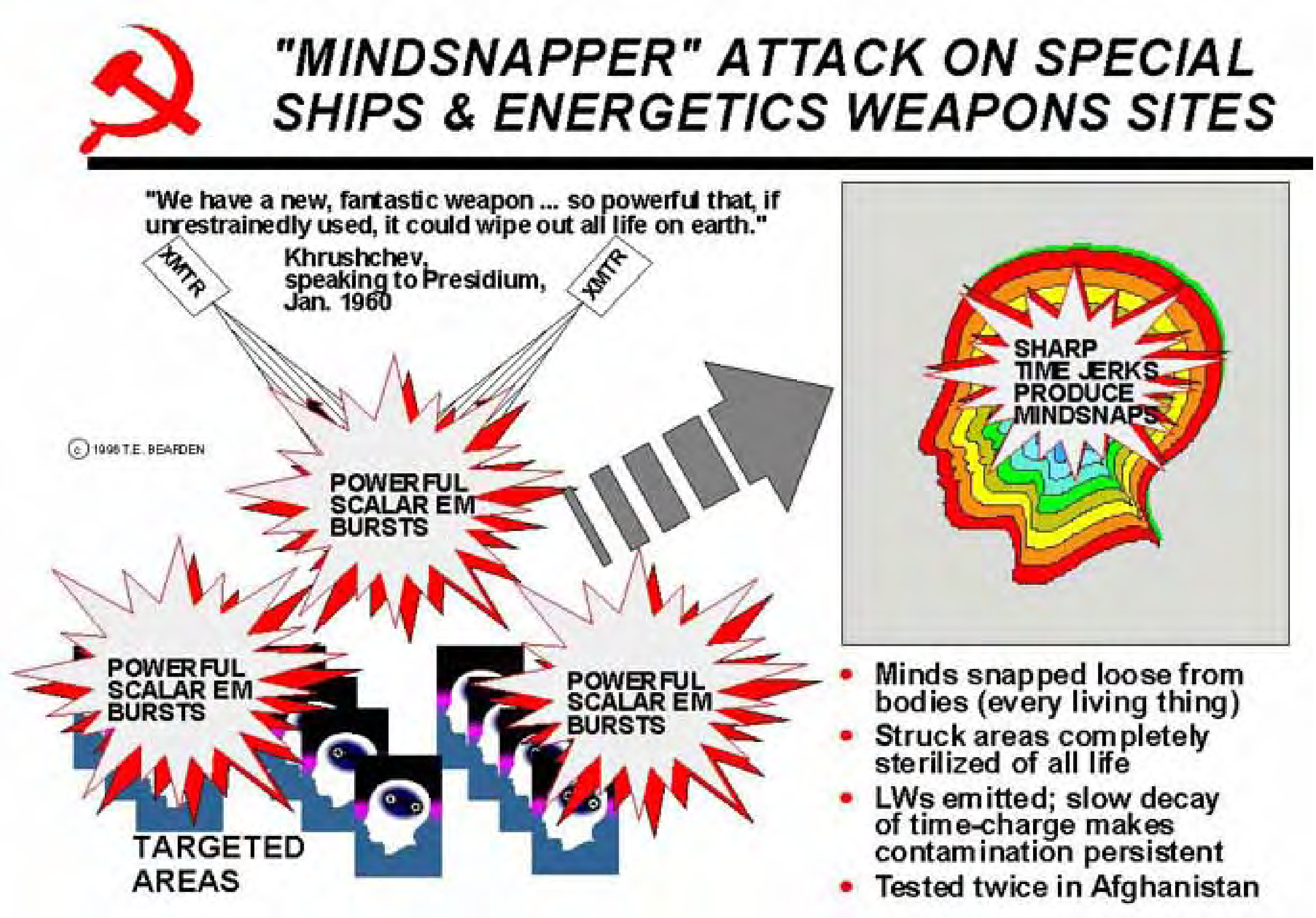

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."
Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
© 1996 T.E. BEARDEN
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

The Tom Bearden Website

Help support the research

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

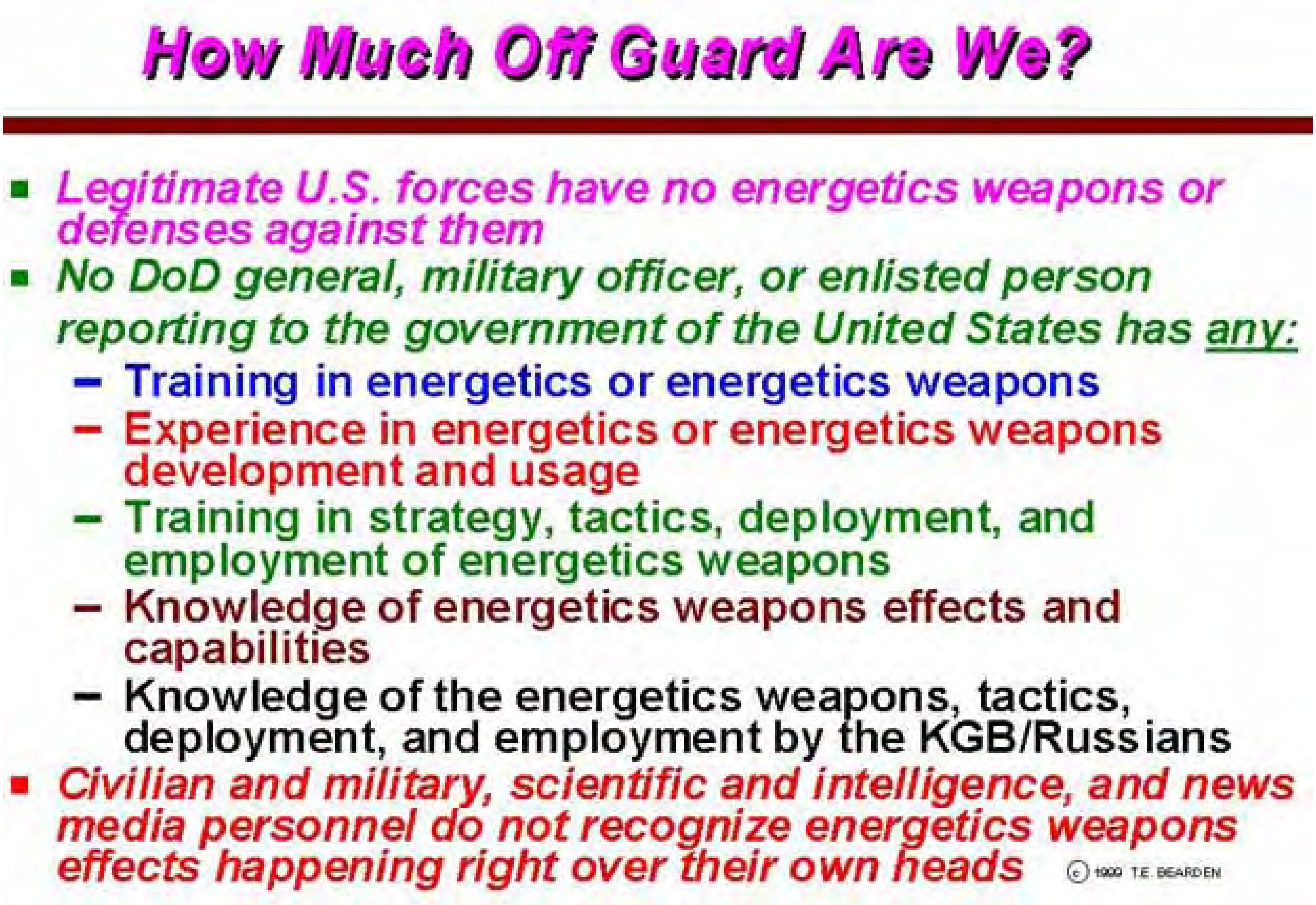

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
  - Training in energetics or energetics weapons
  - Experience in energetics or energetics weapons development and usage
  - Training in strategy, tactics, deployment, and employment of energetics weapons
  - Knowledge of energetics weapons effects and capabilities
  - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

© 1999 T.E. BEARDEN

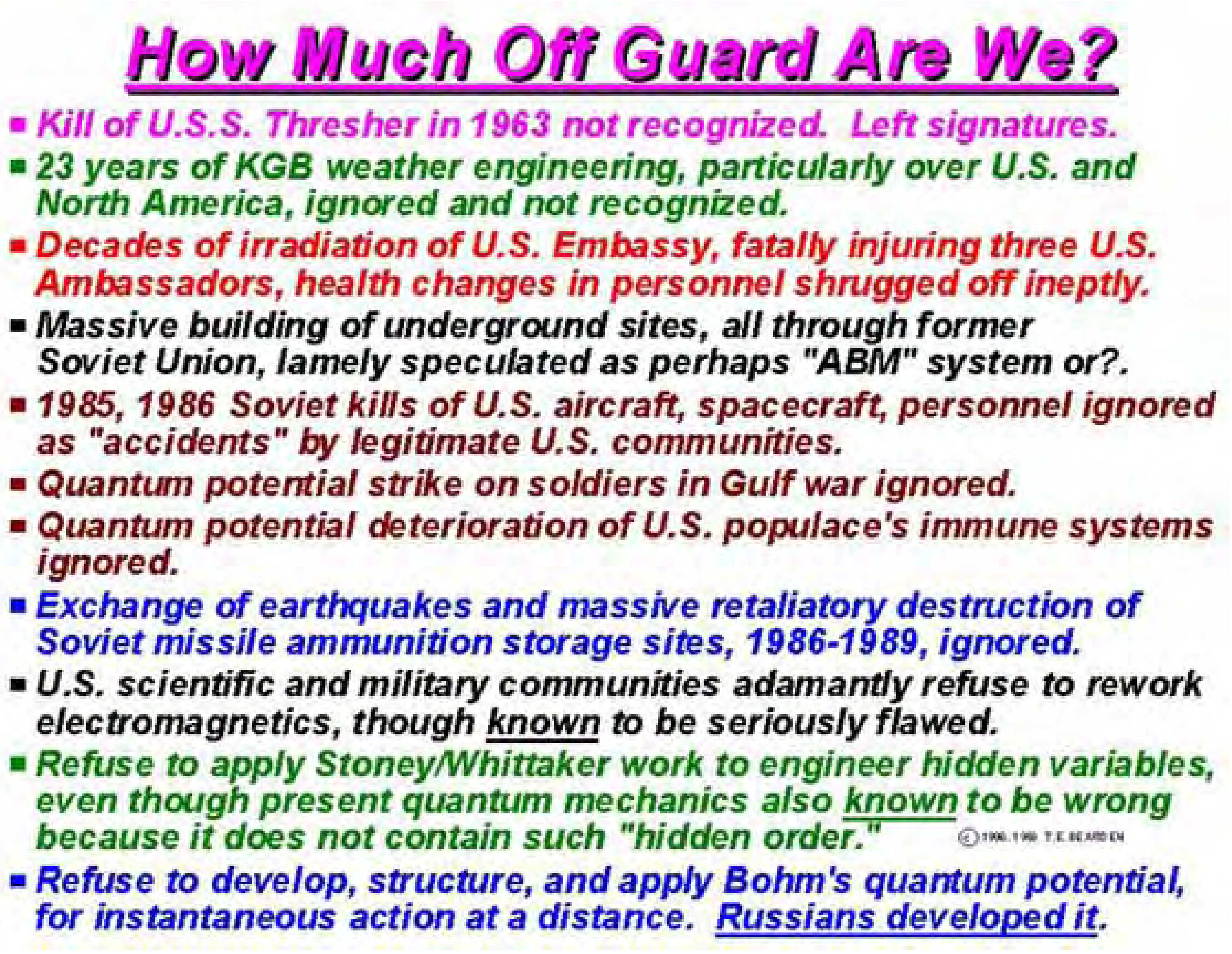

How Much Off Guard Are We?

- *Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.*
- *23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.*
- *Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.*
- *Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.*
- *1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.*
- *Quantum potential strike on soldiers in Gulf war ignored.*
- *Quantum potential deterioration of U.S. populace's immune systems ignored.*
- *Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.*
- *U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.*
- *Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."*
- *Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.*

© [illegible] T.E. BEARDEN

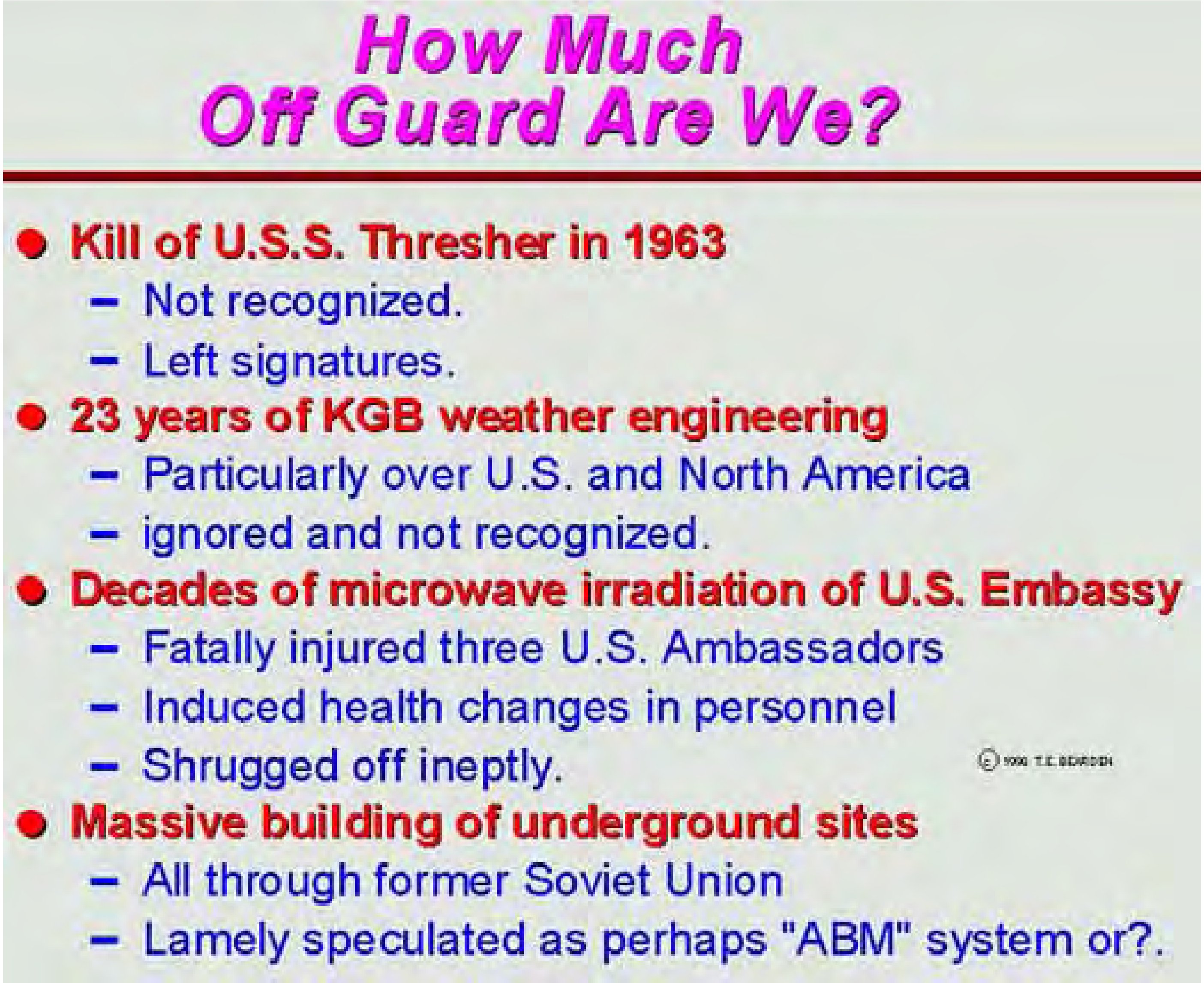

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

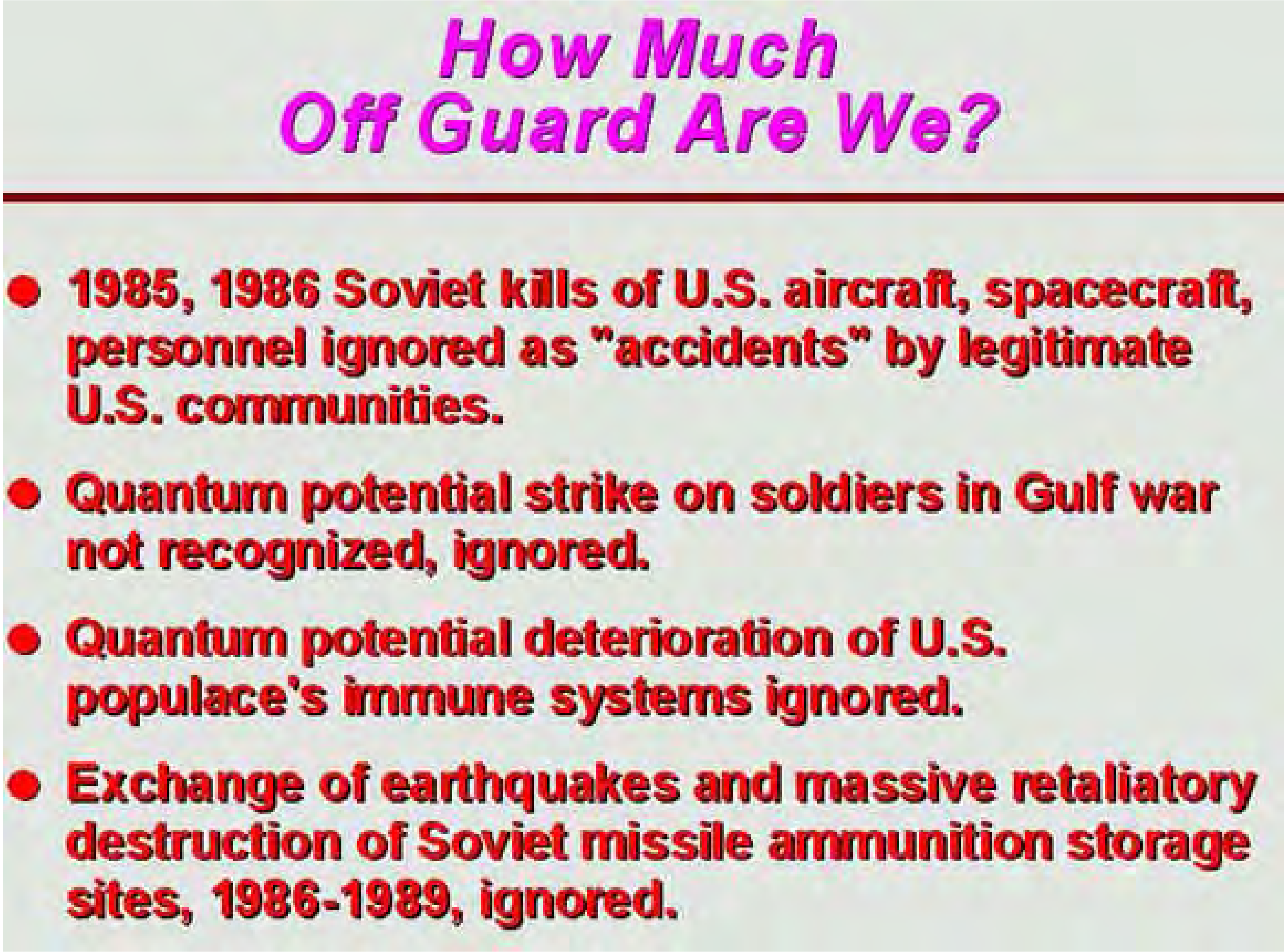

How Much Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

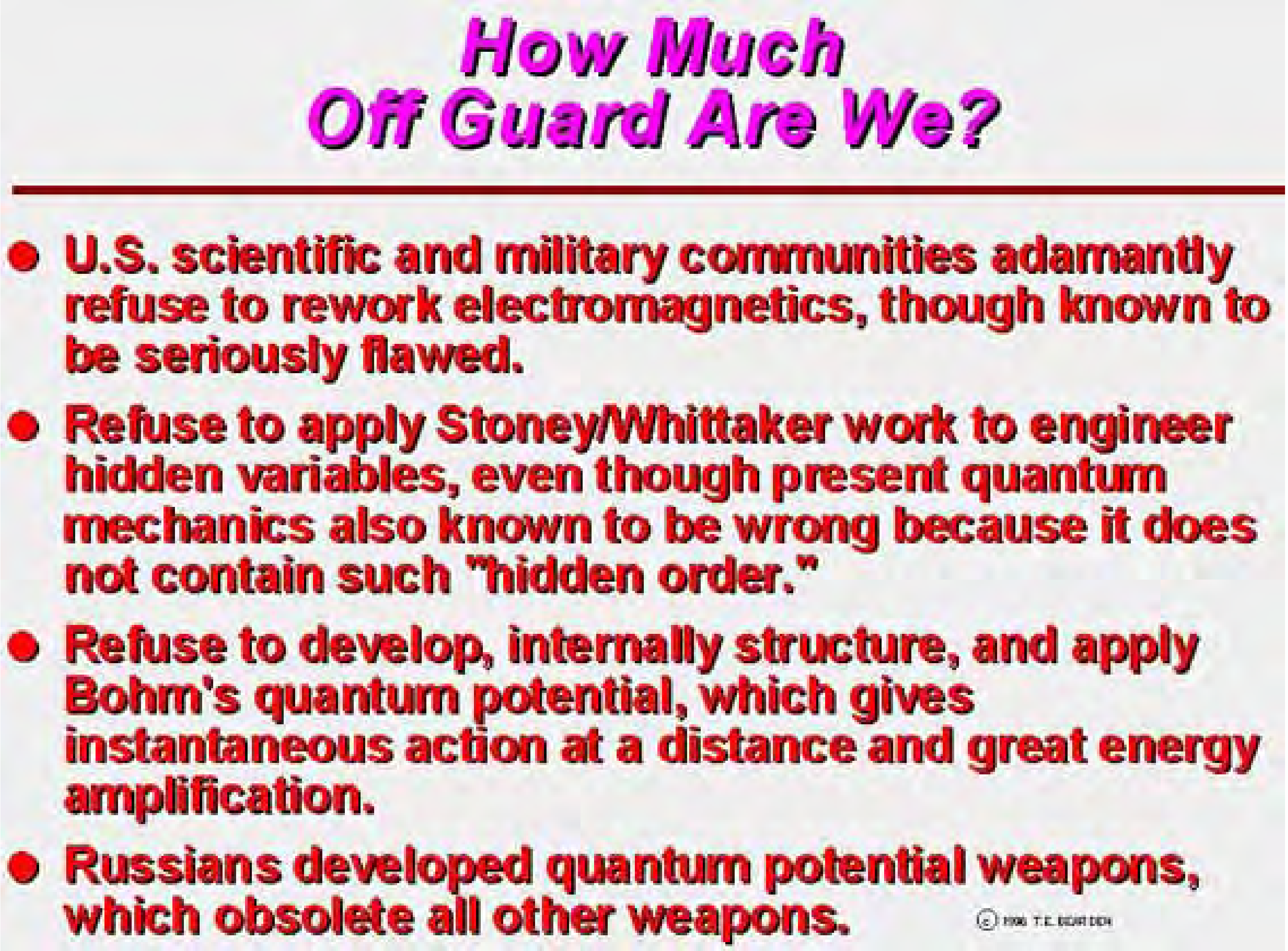

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© 1996 T.E. BEARDEN

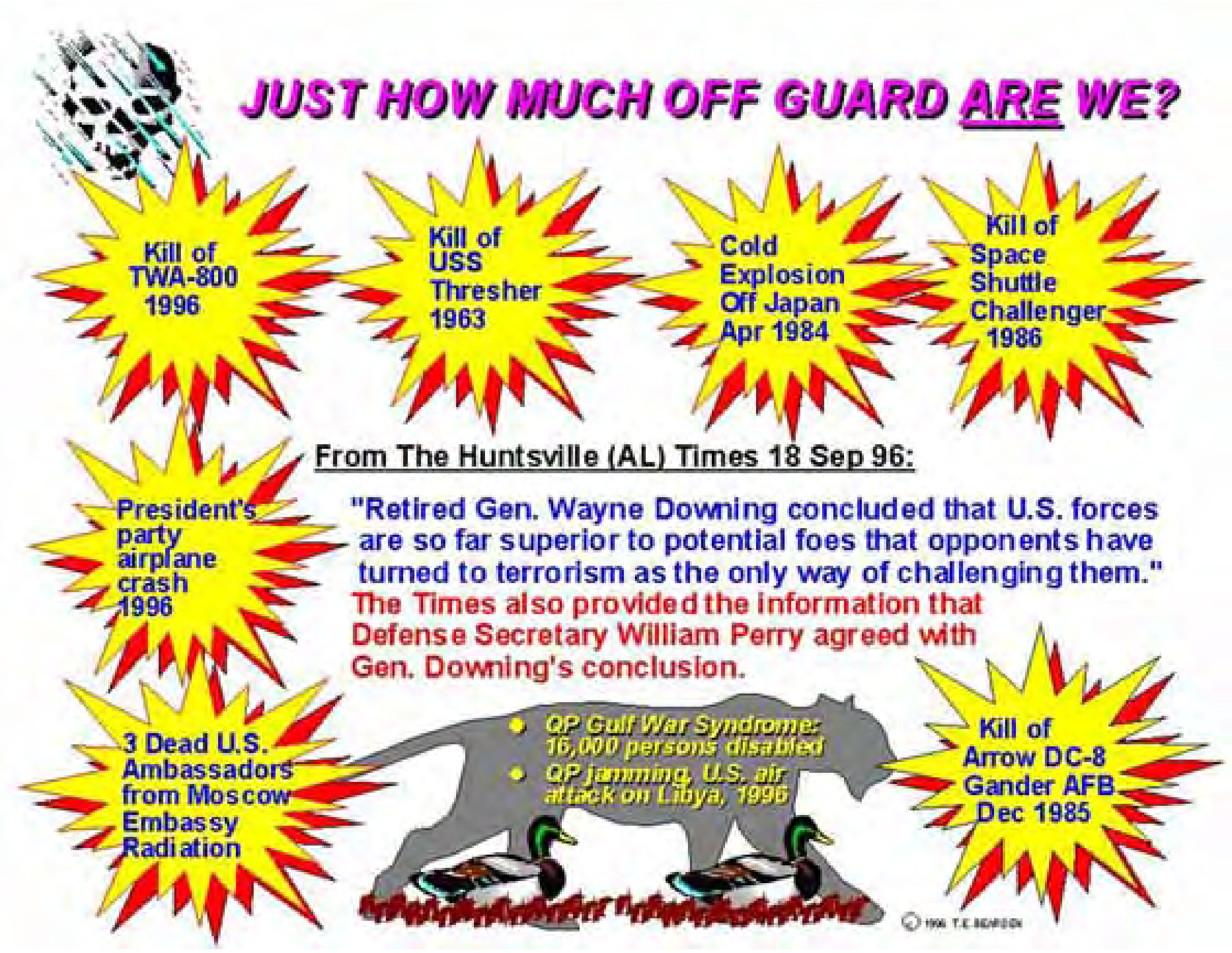

JUST HOW MUCH OFF GUARD *ARE* WE?

Kill of TWA-800 1996

Kill of USS Thresher 1963

Cold Explosion Off Japan Apr 1984

Kill of Space Shuttle Challenger 1986

President's party airplane crash 1996

3 Dead U.S. Ambassadors from Moscow Embassy Radiation

Kill of Arrow DC-8 Gander AFB Dec 1985

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

- QP Gulf War Syndrome: 16,000 persons disabled
- QP jamming, U.S. air attack on Libya, 1996

© 1996 T.E. BEARDEN

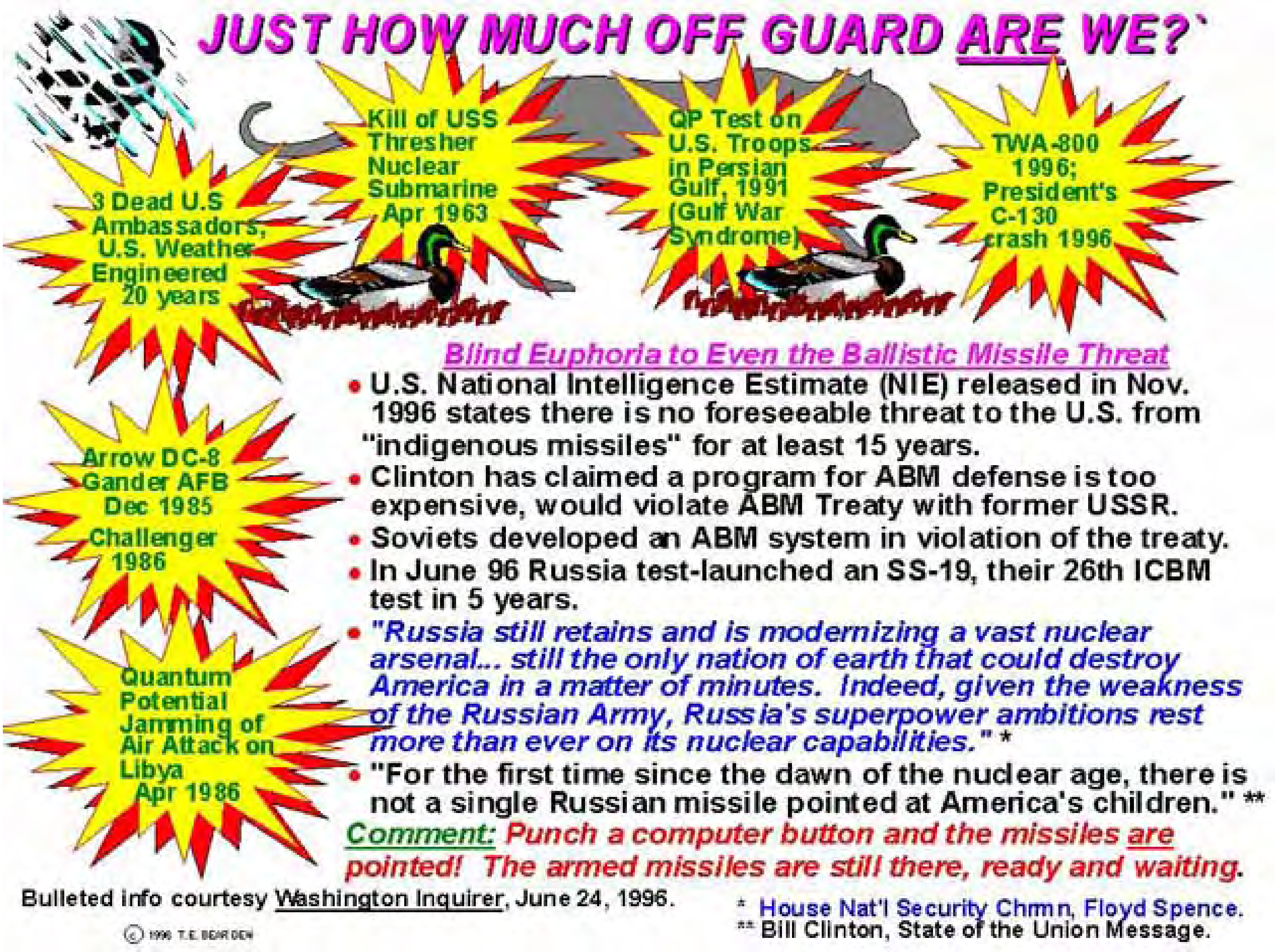

JUST HOW MUCH OFF GUARD ARE WE?`

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

© 1996 T.E. BEARDEN

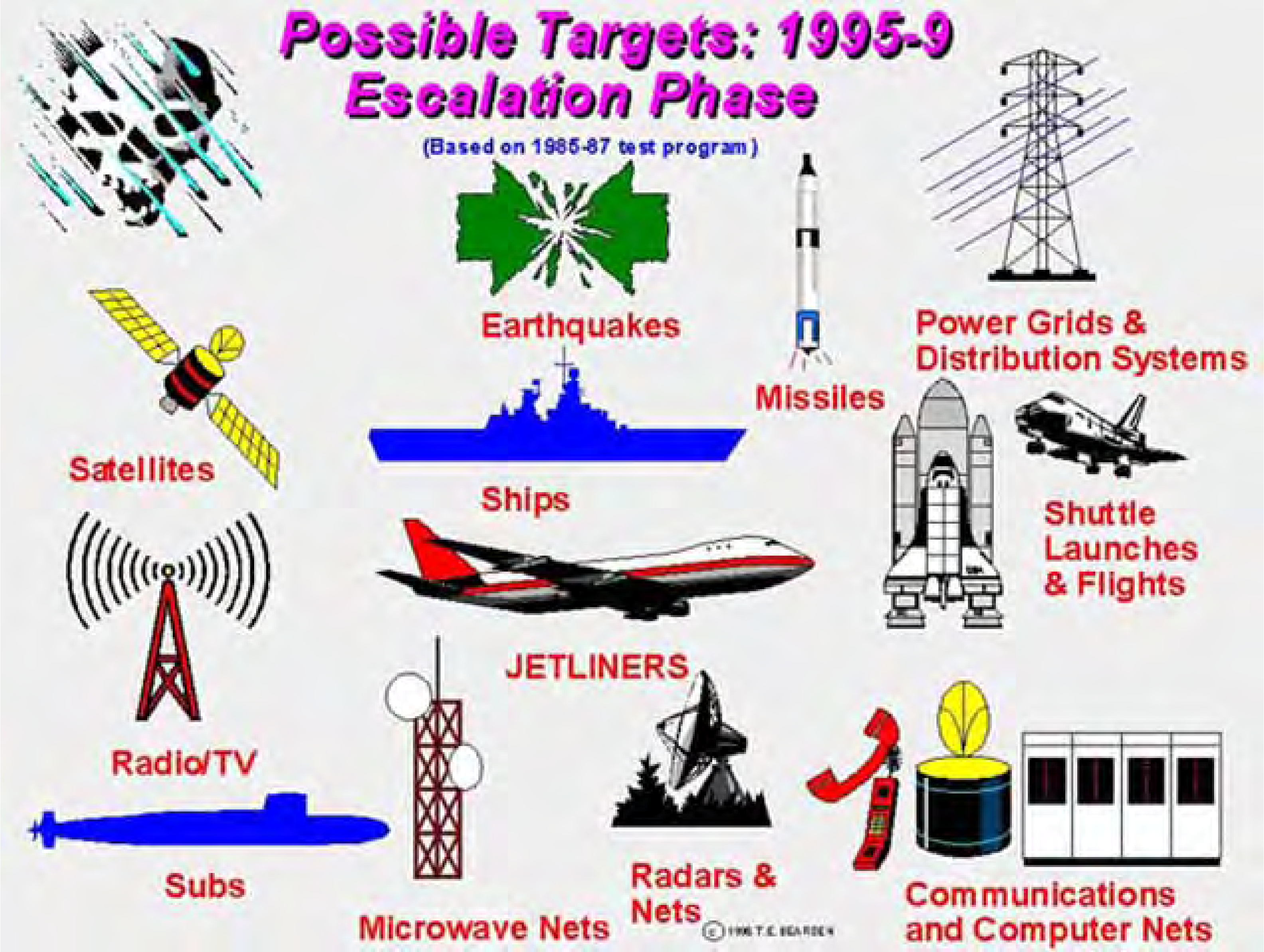

Possible Targets: 1995-9 Escalation Phase

(Based on 1985-87 test program)

Earthquakes
Missiles
Power Grids & Distribution Systems
Satellites
Ships
Shuttle Launches & Flights
JETLINERS
Radio/TV
Subs
Microwave Nets
Radars & Nets
Communications and Computer Nets
© 1996 T.E. BEARDEN

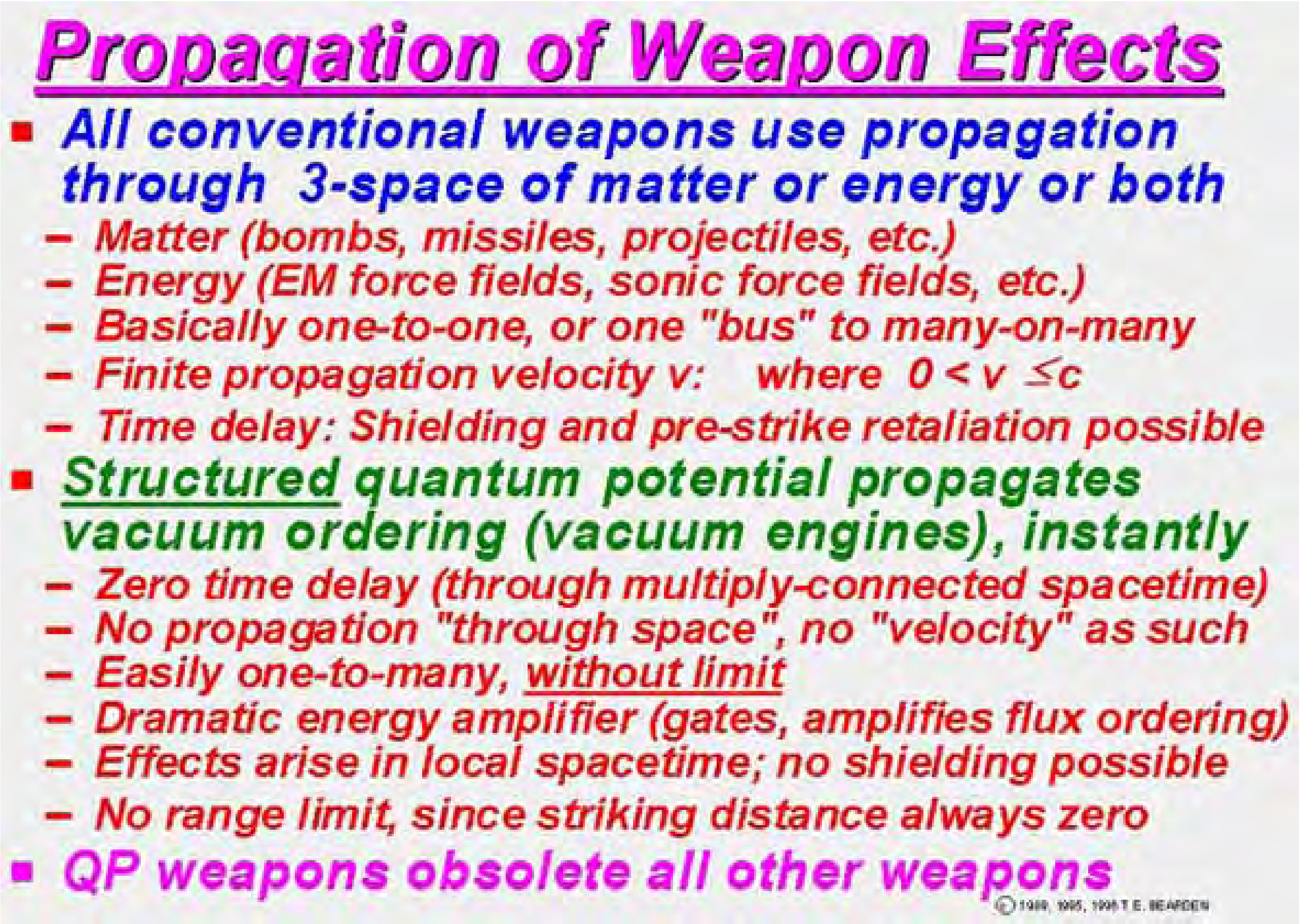

Propagation of Weapon Effects

- All conventional weapons use propagation through 3-space of matter or energy or both
  - Matter (bombs, missiles, projectiles, etc.)
  - Energy (EM force fields, sonic force fields, etc.)
  - Basically one-to-one, or one "bus" to many-on-many
  - Finite propagation velocity v: where $0 < v \leq c$
  - Time delay: Shielding and pre-strike retaliation possible
- Structured quantum potential propagates vacuum ordering (vacuum engines), instantly
  - Zero time delay (through multiply-connected spacetime)
  - No propagation "through space", no "velocity" as such
  - Easily one-to-many, without limit
  - Dramatic energy amplifier (gates, amplifies flux ordering)
  - Effects arise in local spacetime; no shielding possible
  - No range limit, since striking distance always zero
- QP weapons obsolete all other weapons

© 1989, 1995, 1998 T. E. BEARDEN

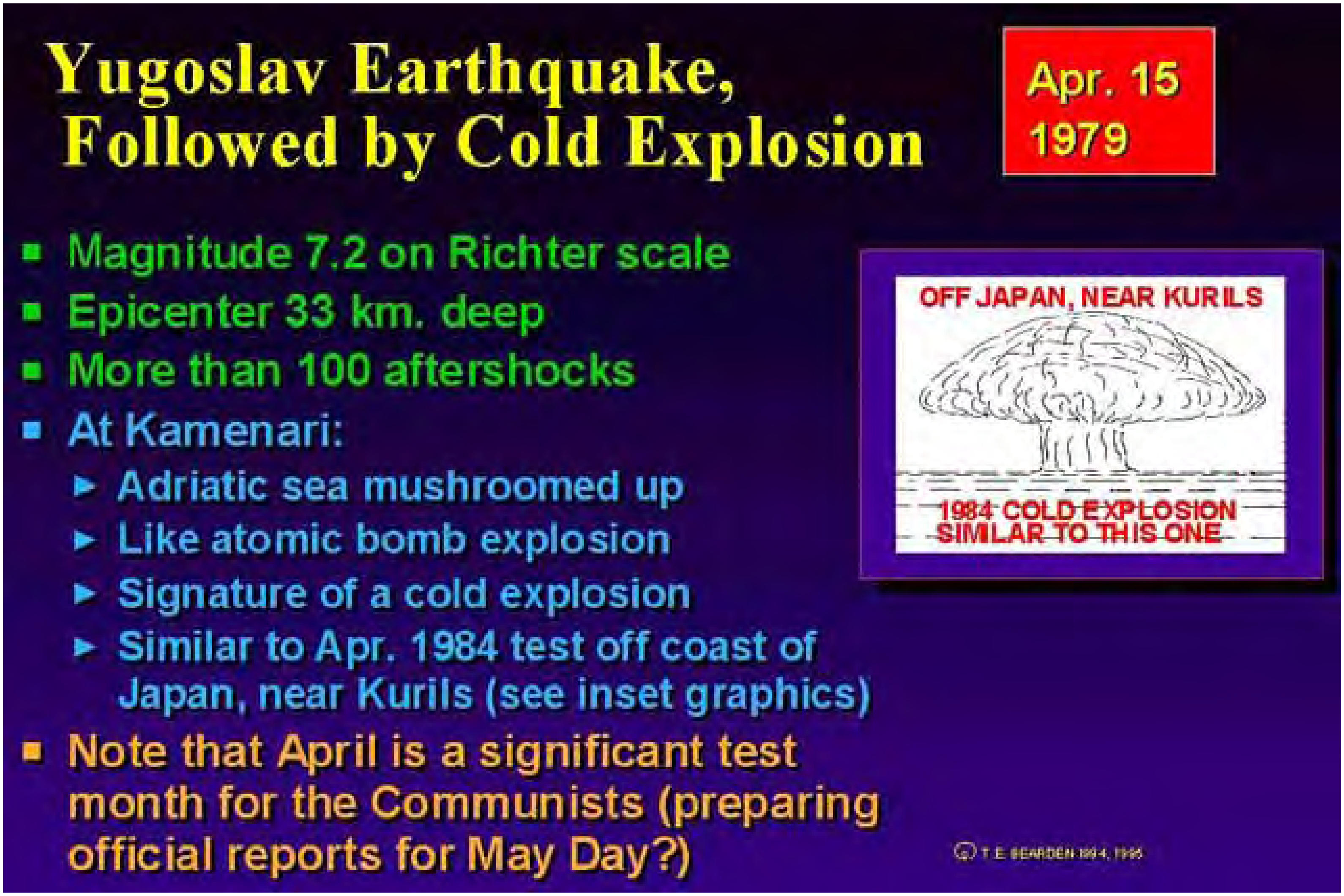

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

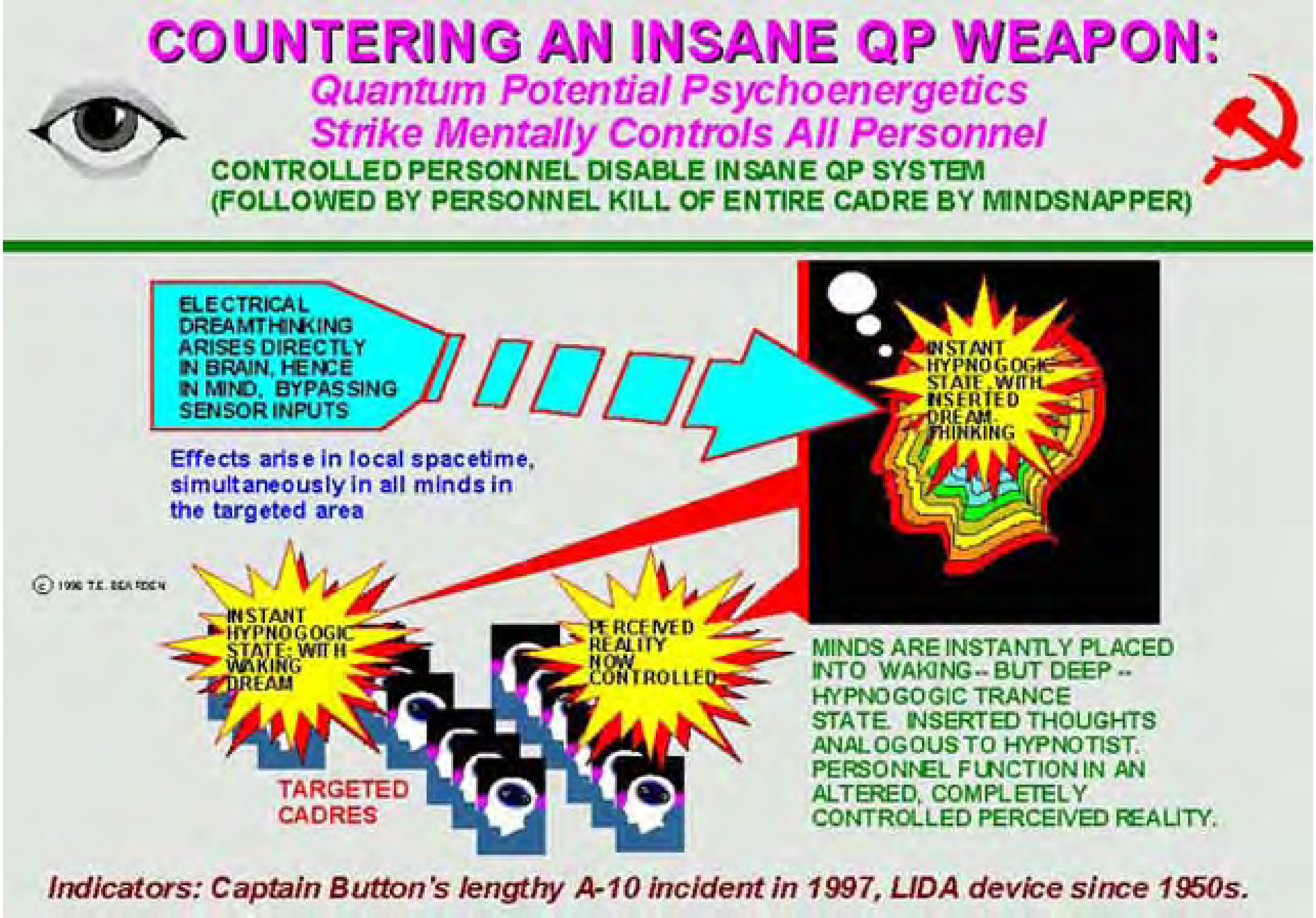

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics
Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING
INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM
PERCEIVED REALITY NOW CONTROLLED
TARGETED CADRES

MINDS ARE INSTANTLY PLACED INTO WAKING-- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

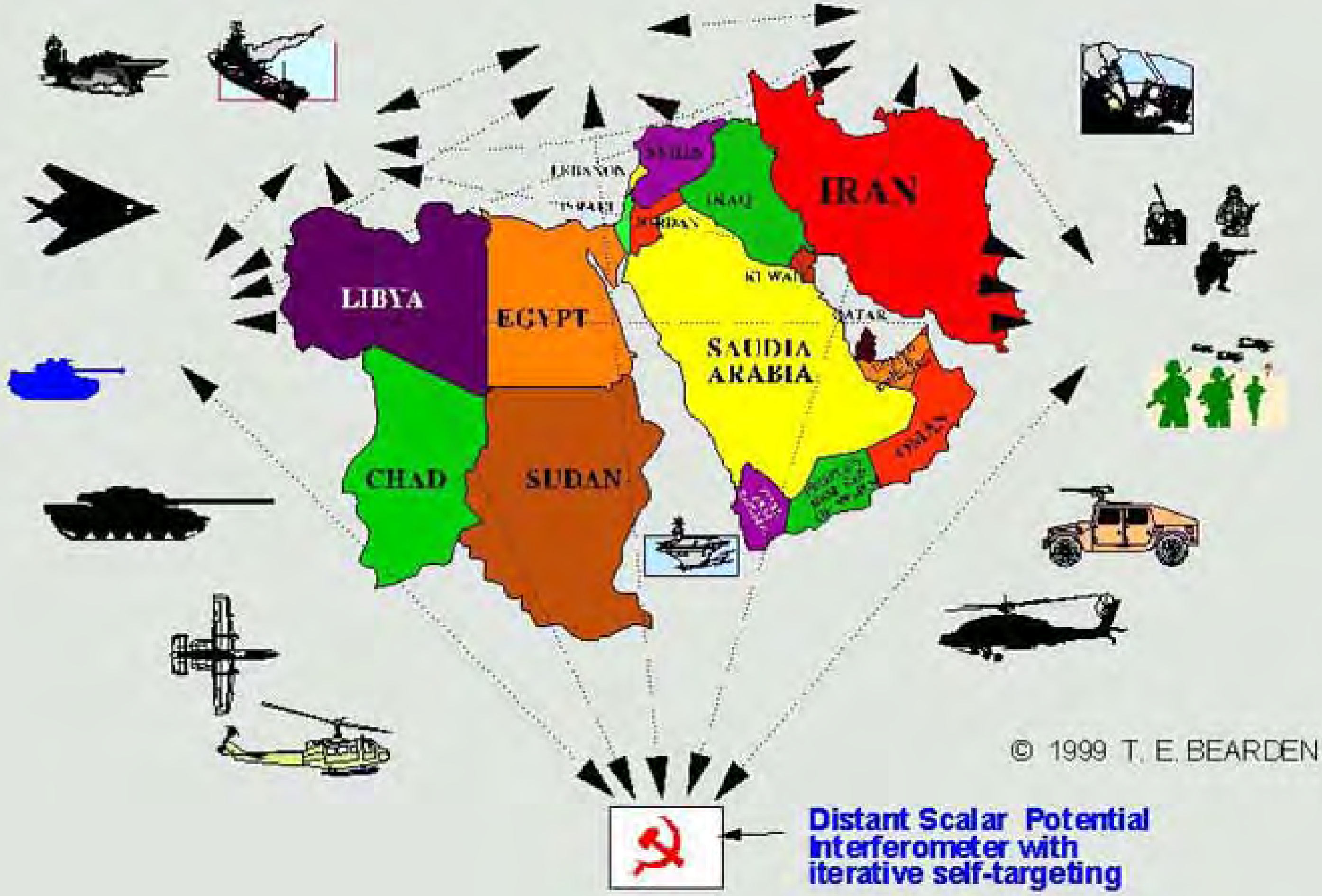
LEBANON
ISRAEL
SYRIA
JORDAN
IRAQ
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
OMAN
CHAD
SUDAN

© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

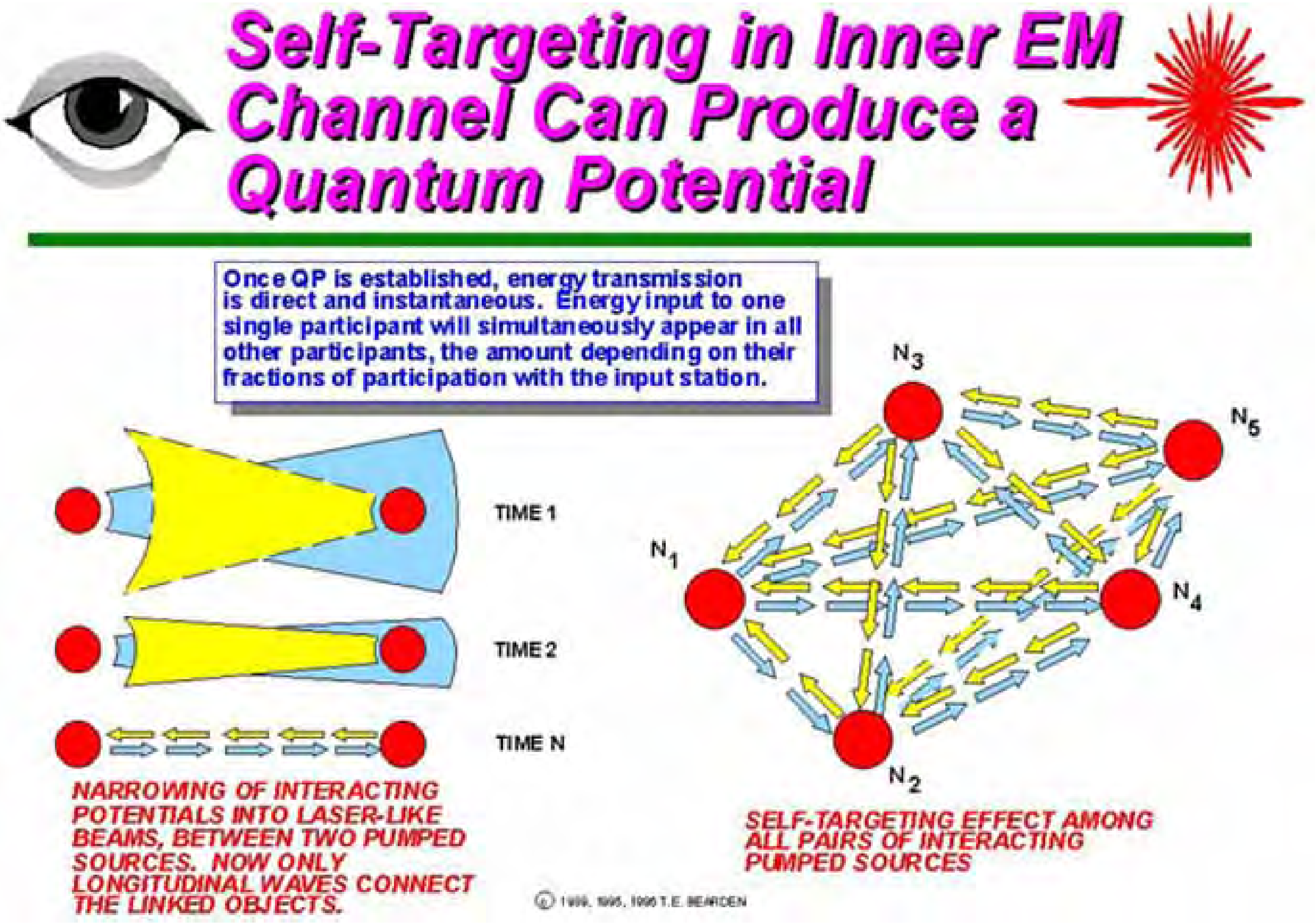

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

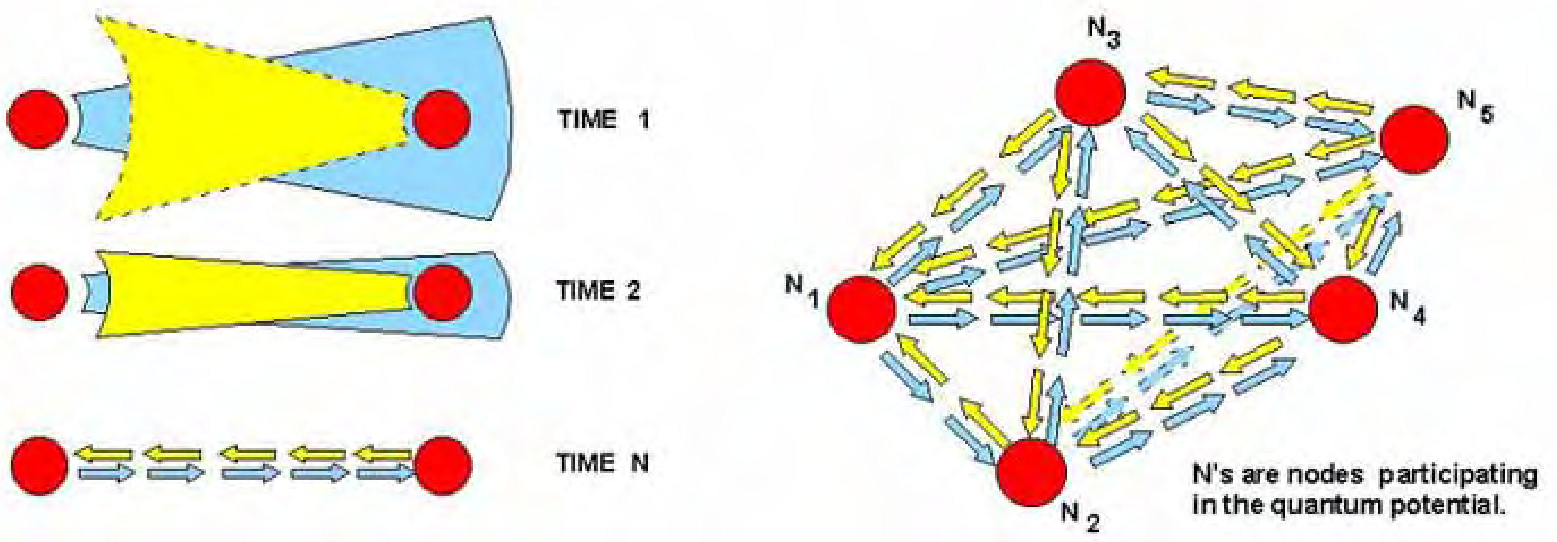

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

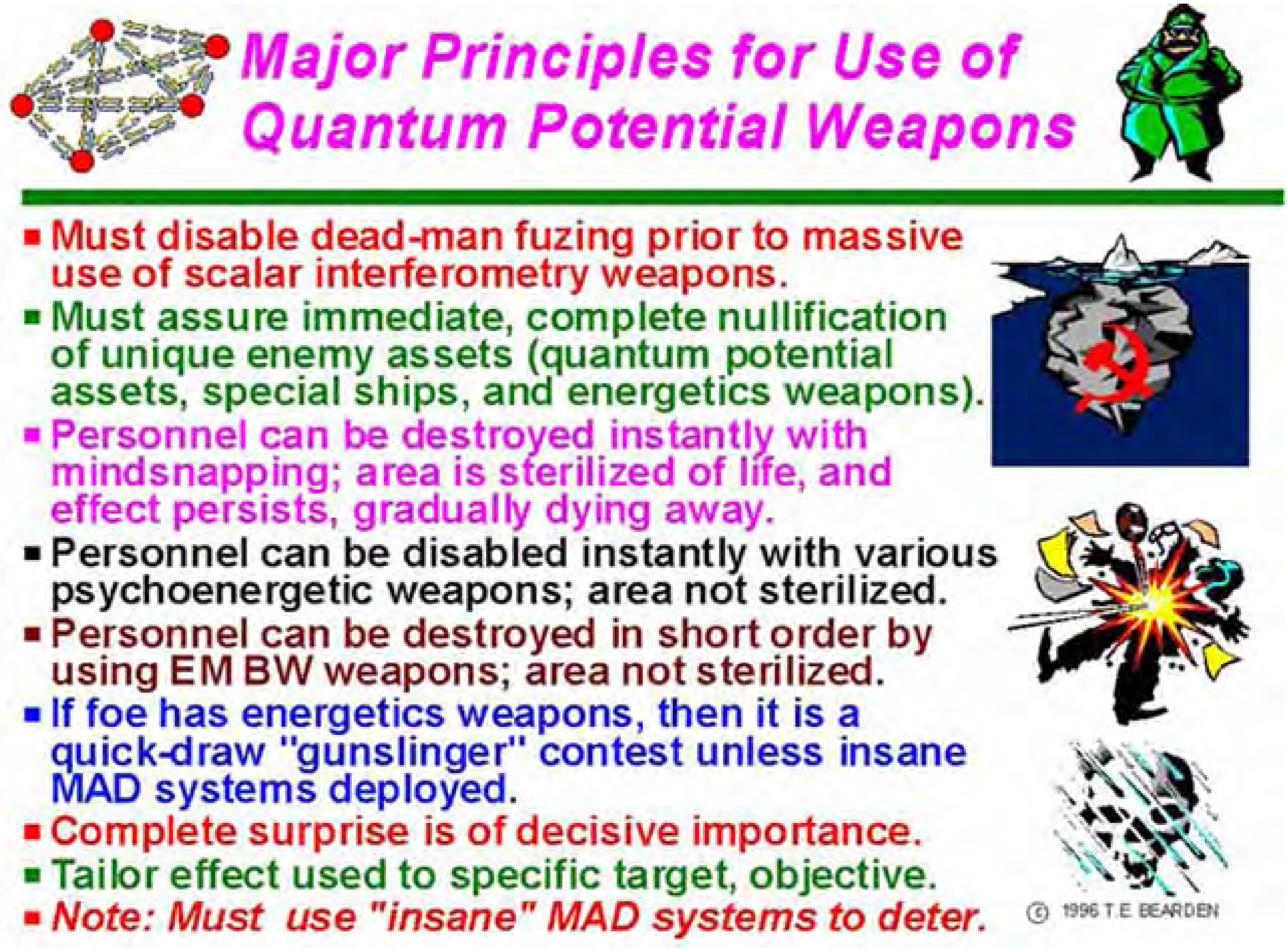

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

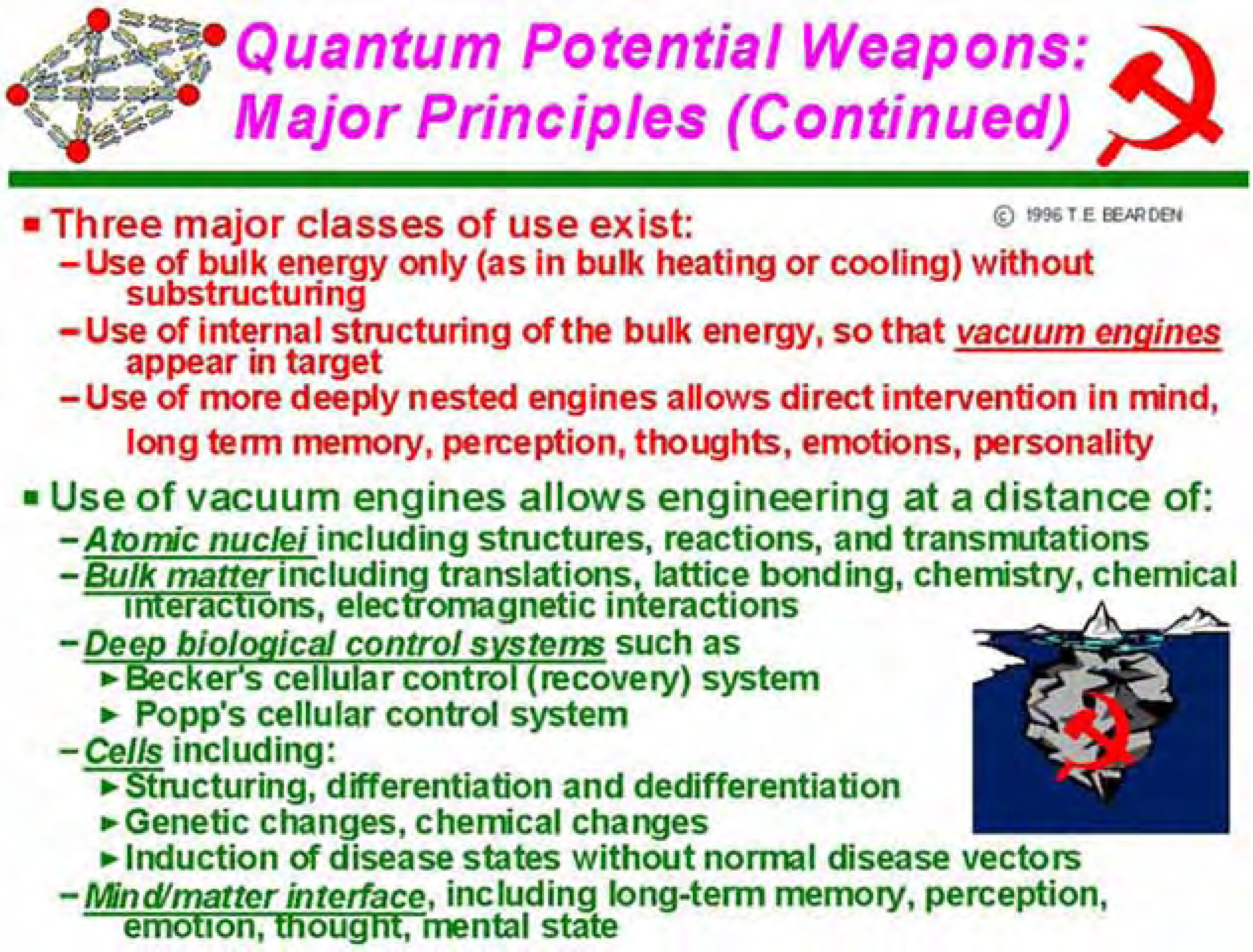

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T.E. BEARDEN

- Three major classes of use exist:
  - Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - *Atomic nuclei* including structures, reactions, and transmutations
  - *Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - *Deep biological control systems* such as
    - Becker's cellular control (recovery) system
    - Popp's cellular control system
  - *Cells* including:
    - Structuring, differentiation and dedifferentiation
    - Genetic changes, chemical changes
    - Induction of disease states without normal disease vectors
  - *Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

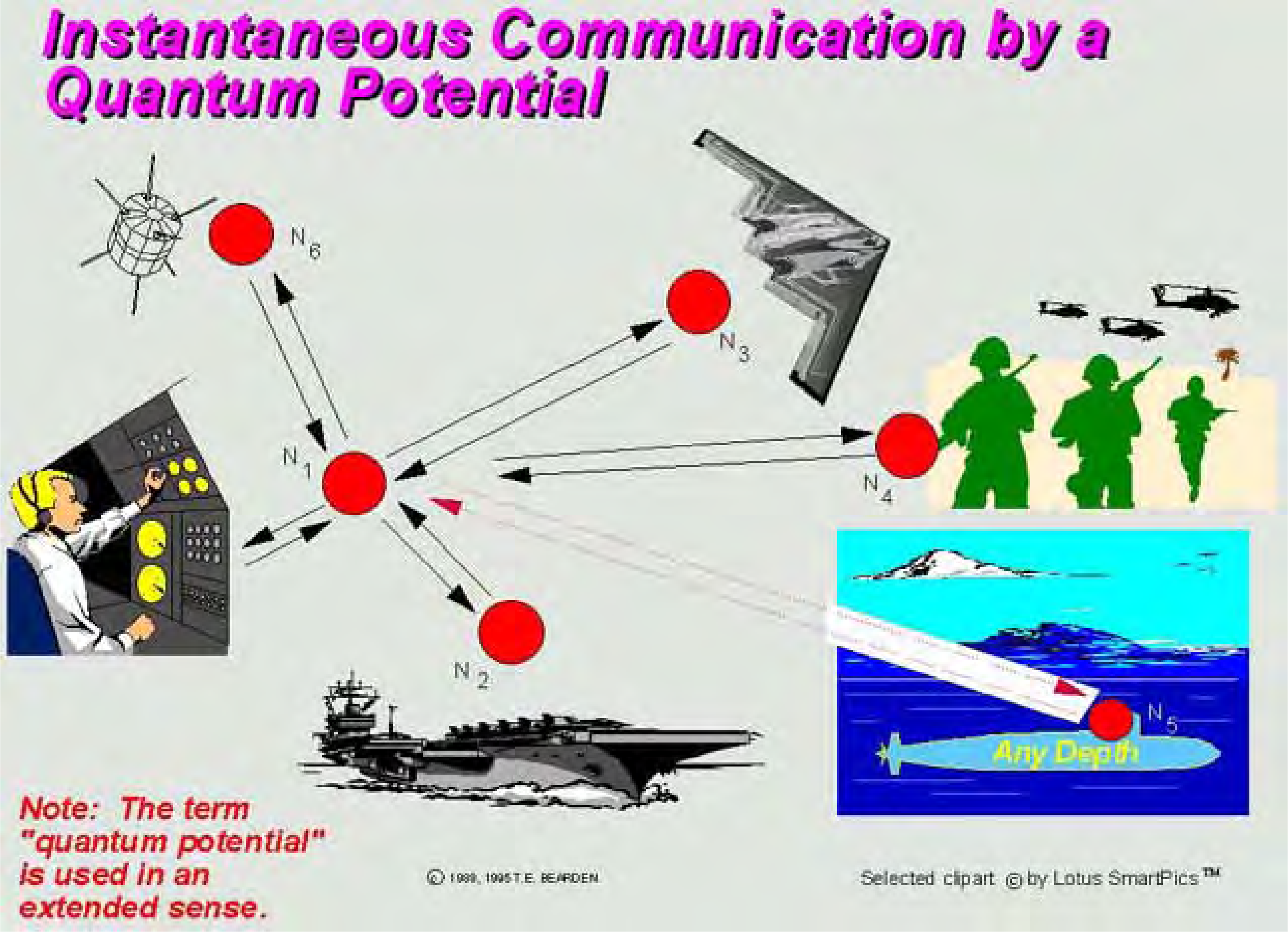

Instantaneous Communication by a Quantum Potential

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics™

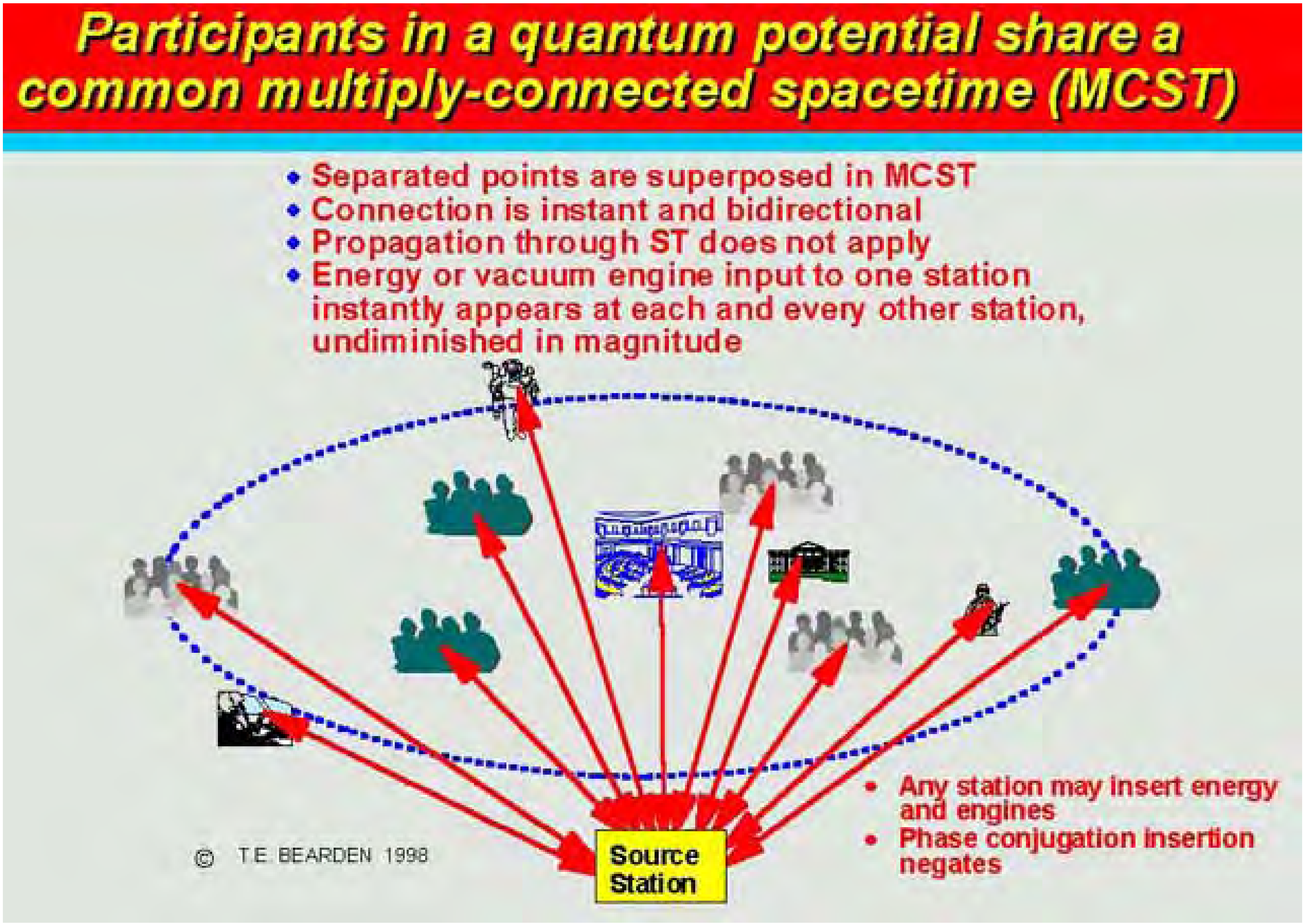

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

© T.E. BEARDEN 1998

- Any station may insert energy and engines
- Phase conjugation insertion negates

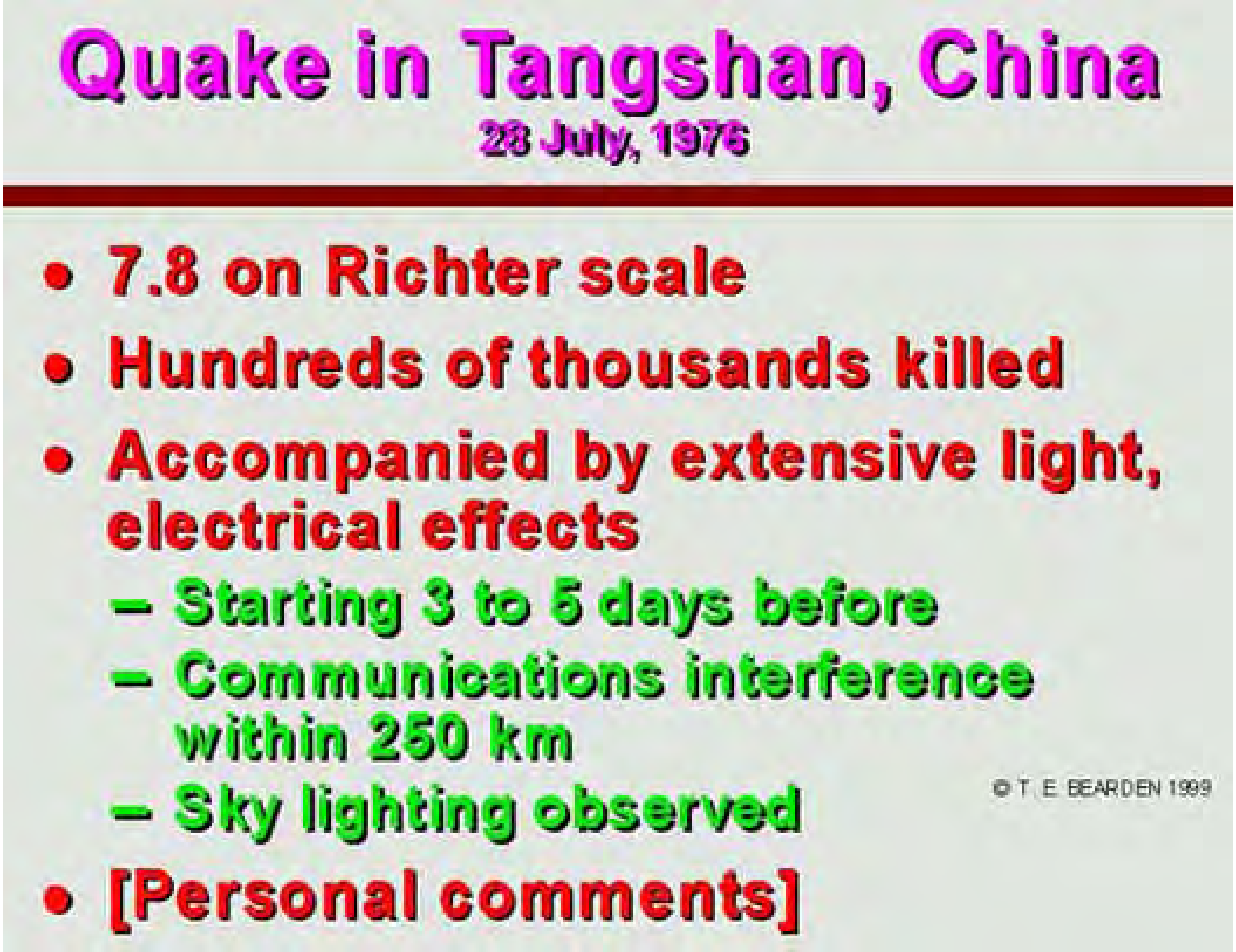

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

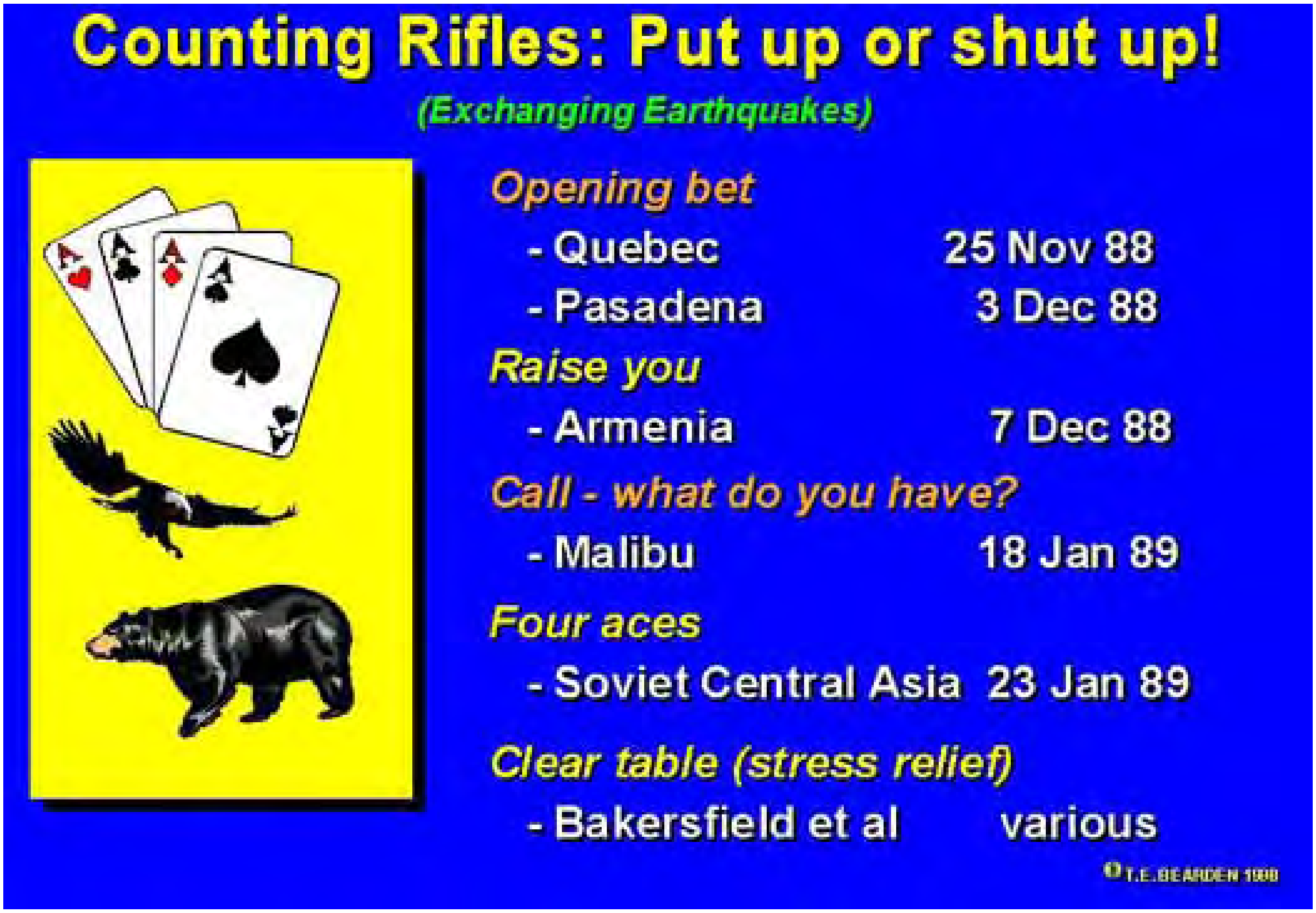

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet

- Quebec 25 Nov 88
- Pasadena 3 Dec 88

Raise you

- Armenia 7 Dec 88

Call - what do you have?

- Malibu 18 Jan 89

Four aces

- Soviet Central Asia 23 Jan 89

Clear table (stress relief)

- Bakersfield et al various

©T.E.BEARDEN 1990

Instant Communication by a Quantum Potential

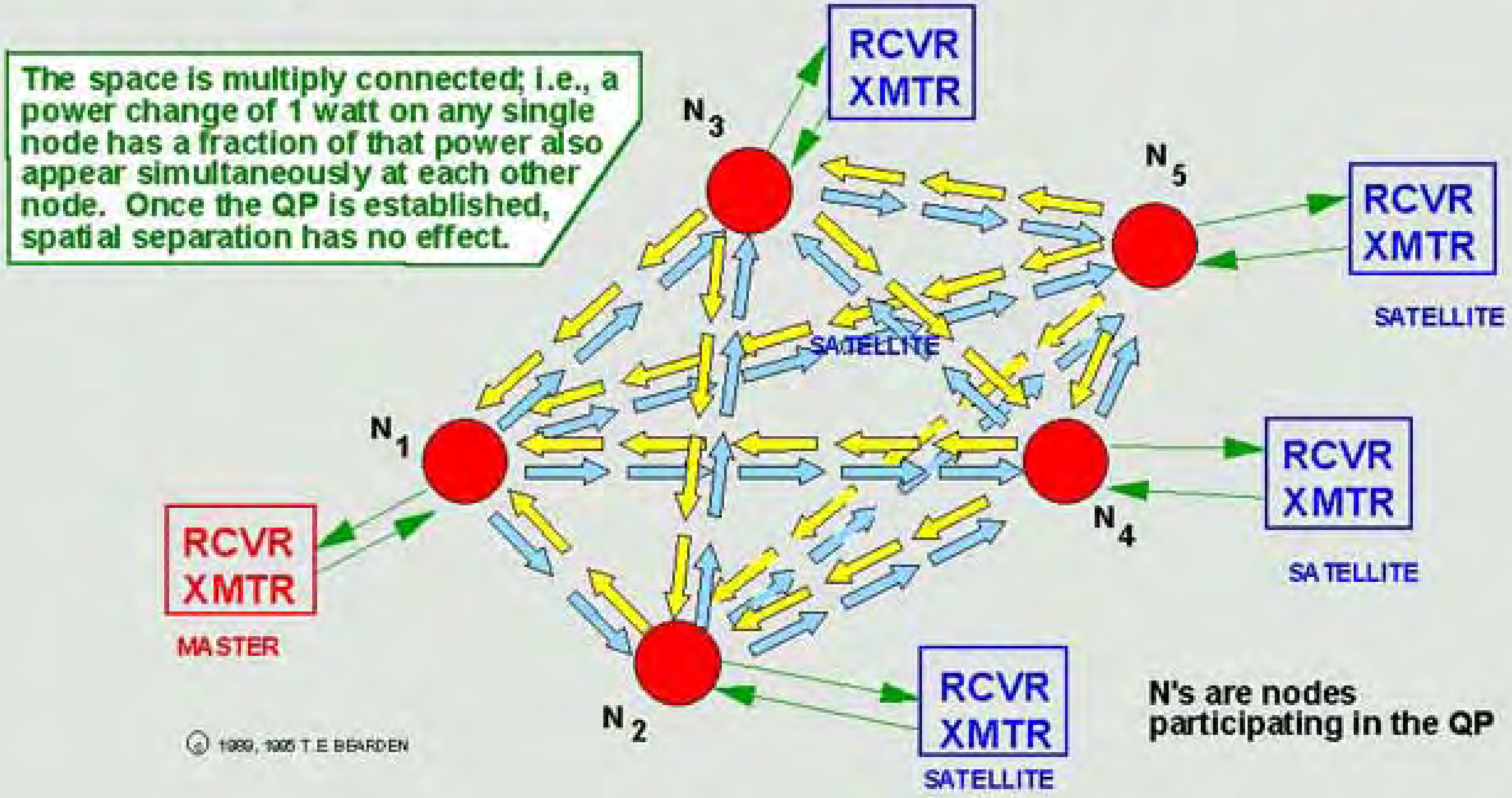

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N's are nodes participating in the QP

N_1
N_2
N_3
N_4
N_5
RCVR XMTR
MASTER
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE
RCVR XMTR
SATELLITE

Ⓒ 1989, 1995 T.E BEARDEN

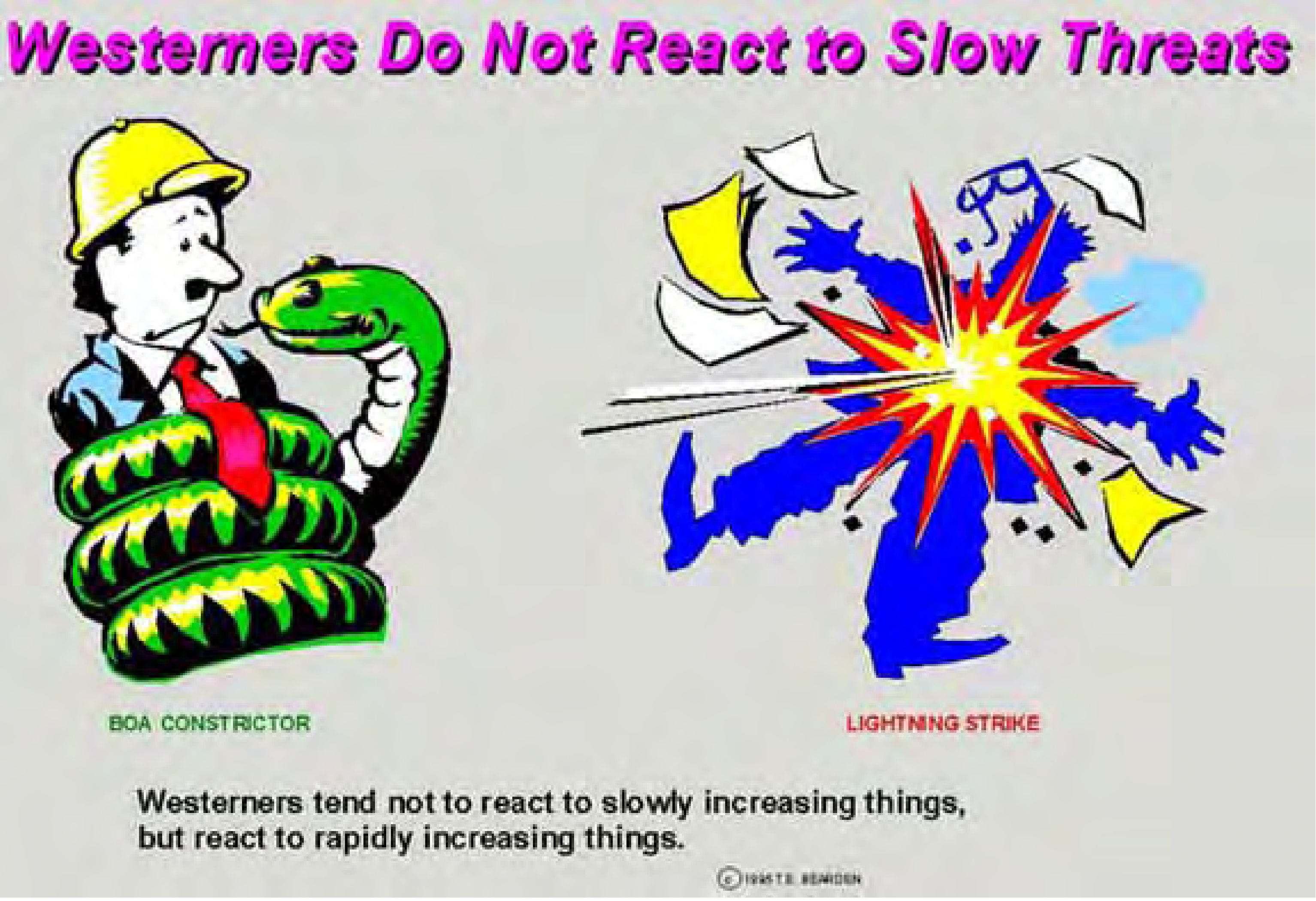

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR

LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things,
but react to rapidly increasing things.

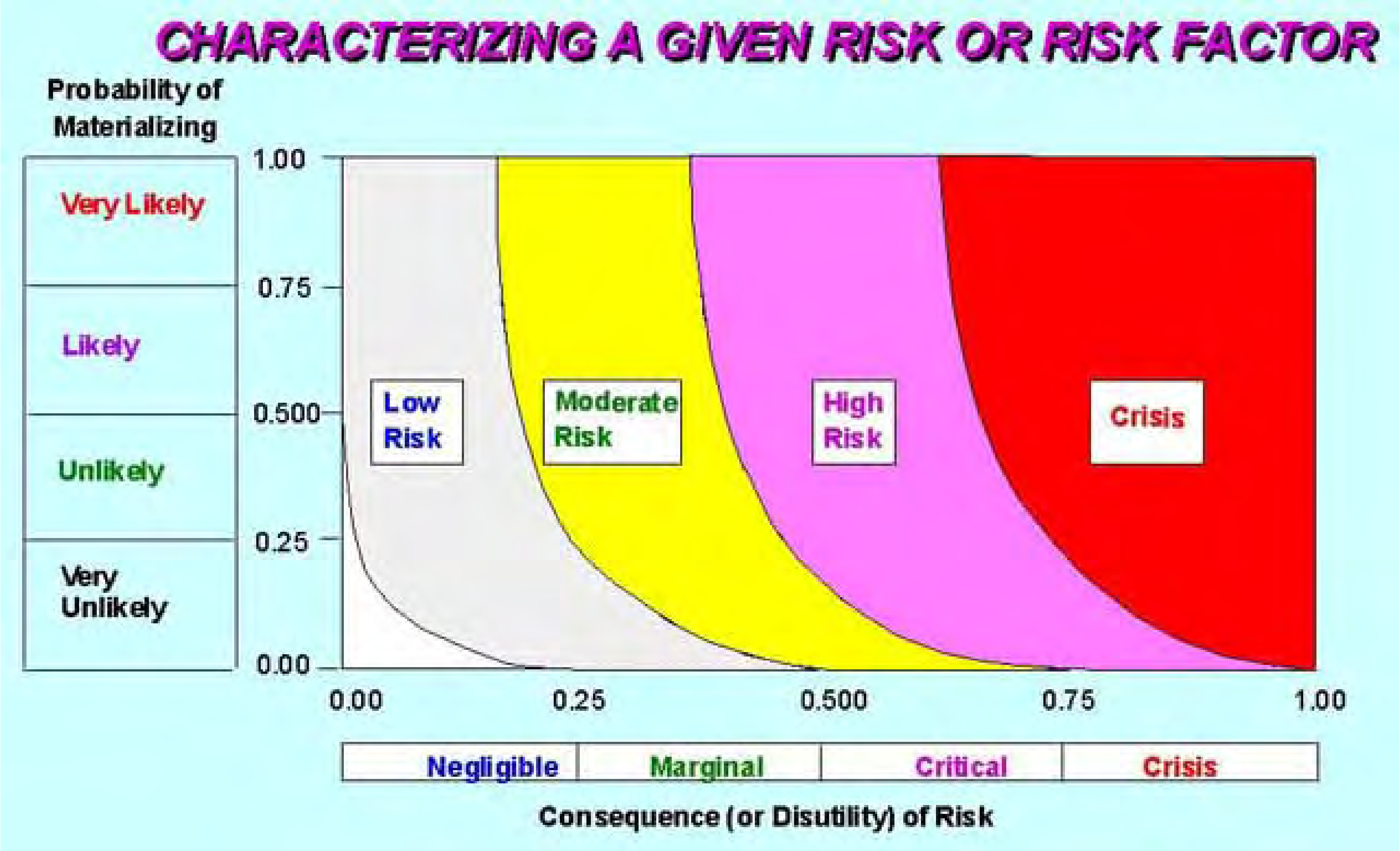
CHARACTERIZING A GIVEN RISK OR RISK FACTOR
Probability of Materializing
Very Likely
Likely
Unlikely
Very Unlikely
1.00
0.75
0.500
0.25
0.00
Low Risk
Moderate Risk
High Risk
Crisis
0.00
0.25
0.500
0.75
1.00
Negligible
Marginal
Critical
Crisis
Consequence (or Disutility) of Risk

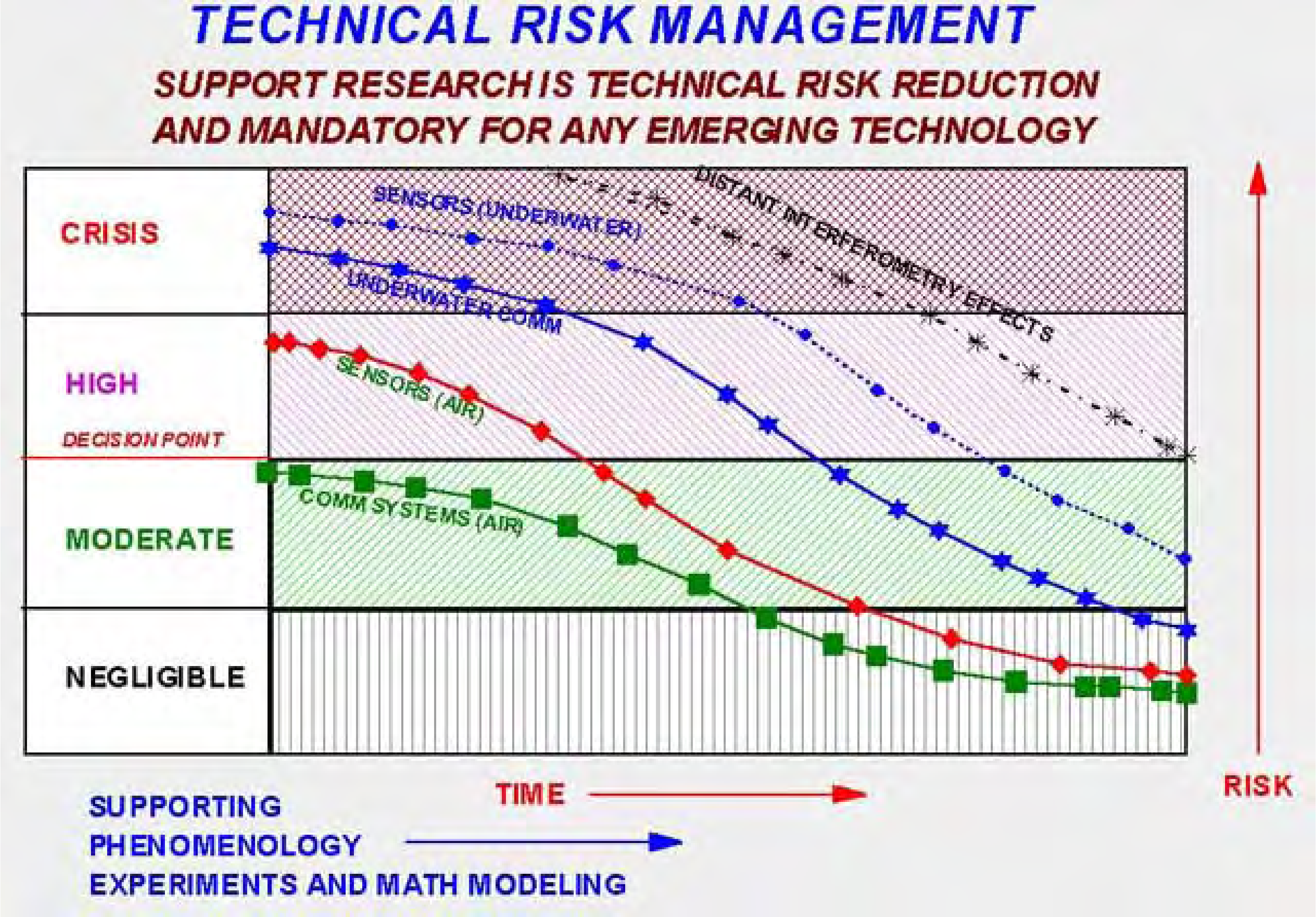

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

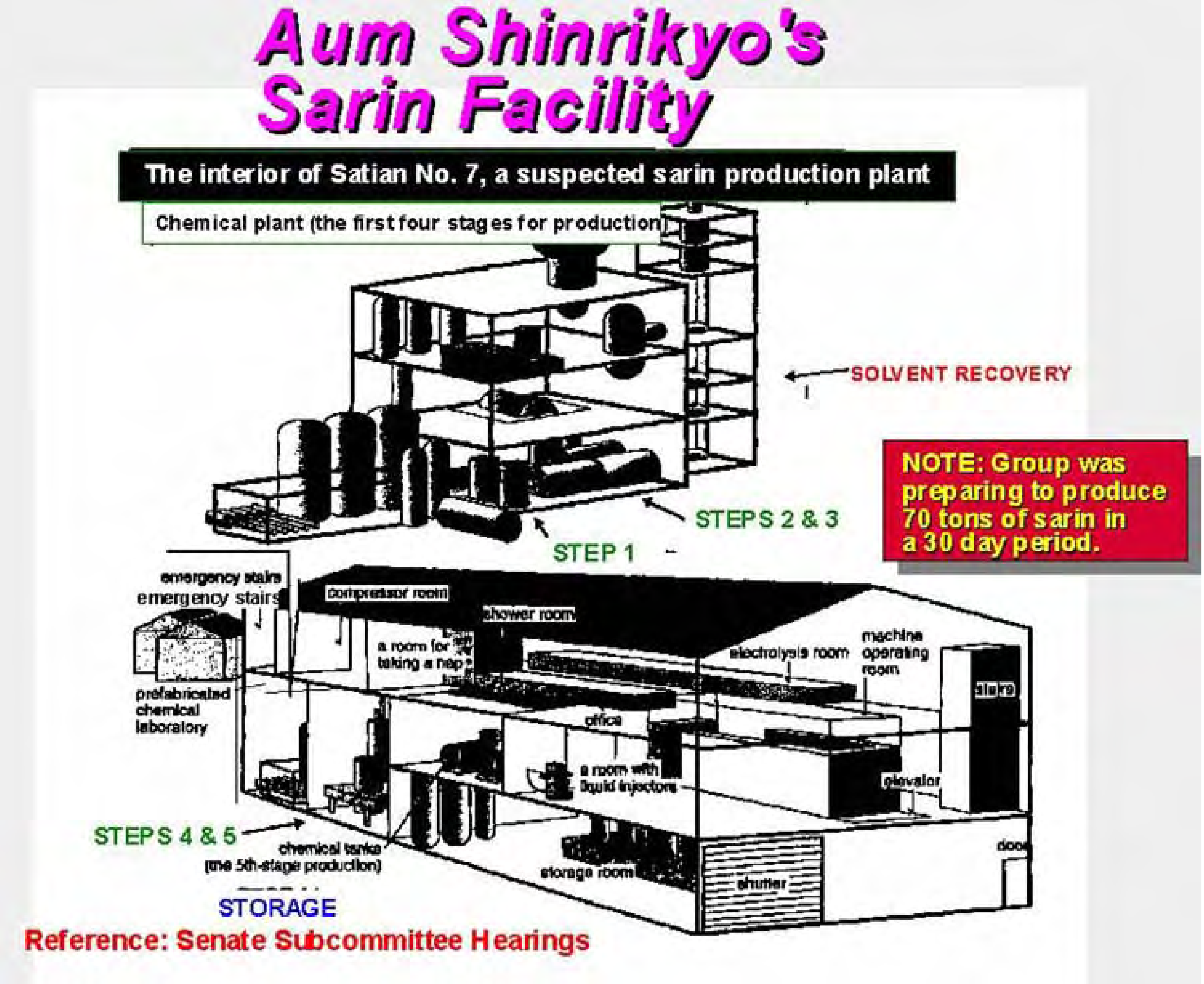

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

Chemical plant (the first four stages for production)

SOLVENT RECOVERY
STEPS 2 & 3
STEP 1
emergency stairs
emergency stairs
compressor room
shower room
a room for taking a nap
machine operating room
electrolysis room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
STEPS 4 & 5
chemical tanks (the 5th-stage production)
storage room
shutter
door
STORAGE

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

Reference: Senate Subcommittee Hearings

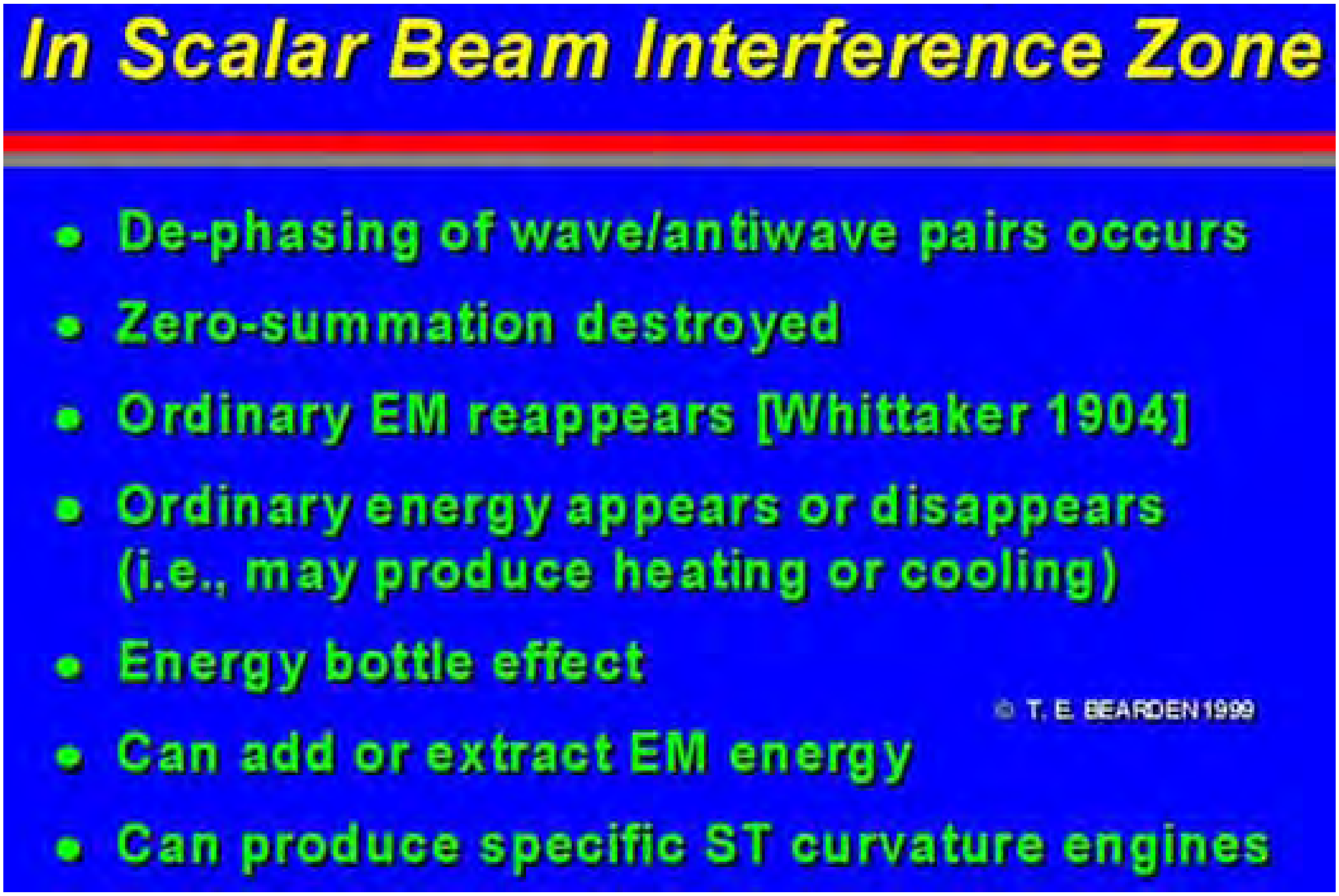

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

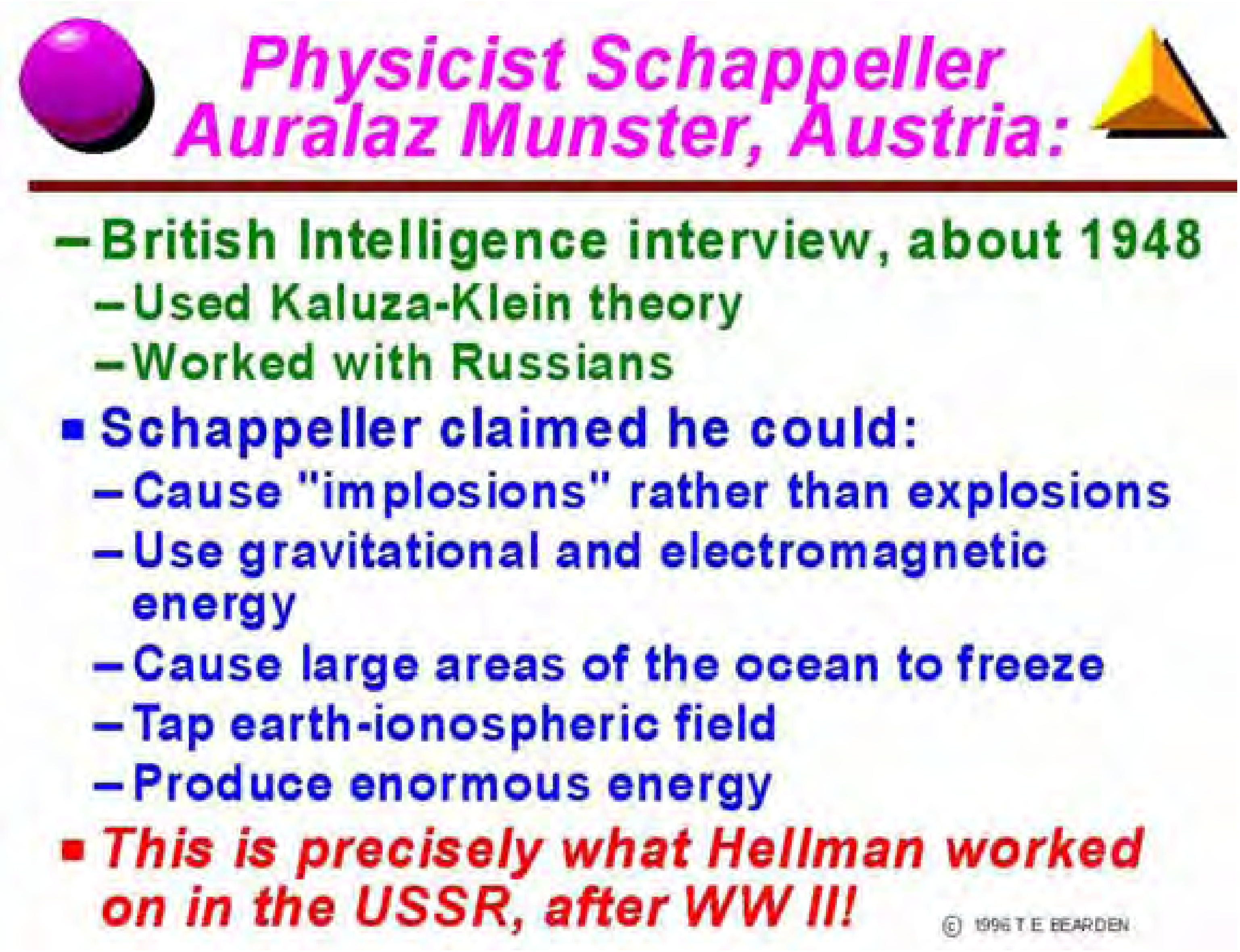

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- **Schappeller claimed he could:**
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T.E. BEARDEN

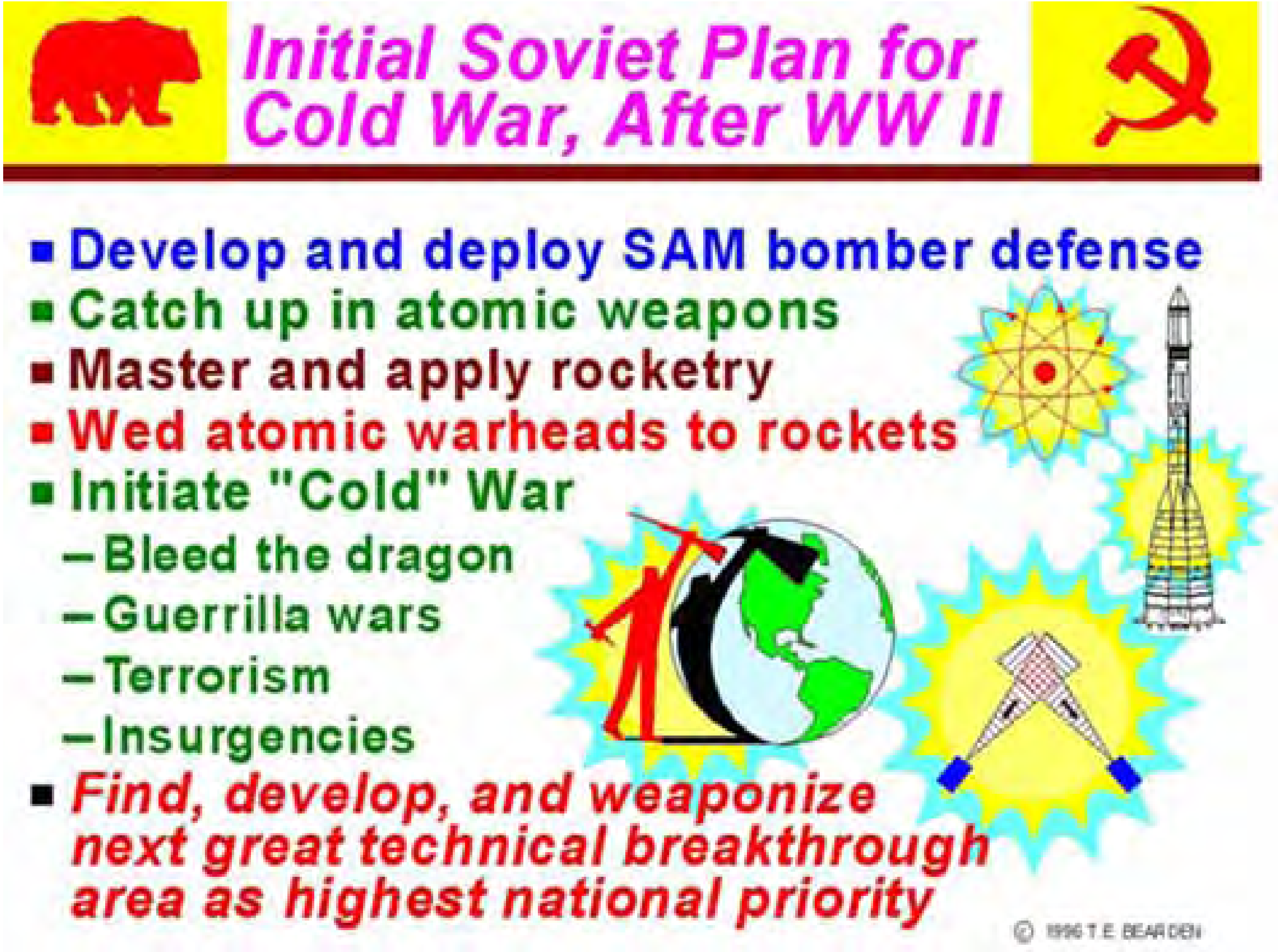

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T.E. BEARDEN

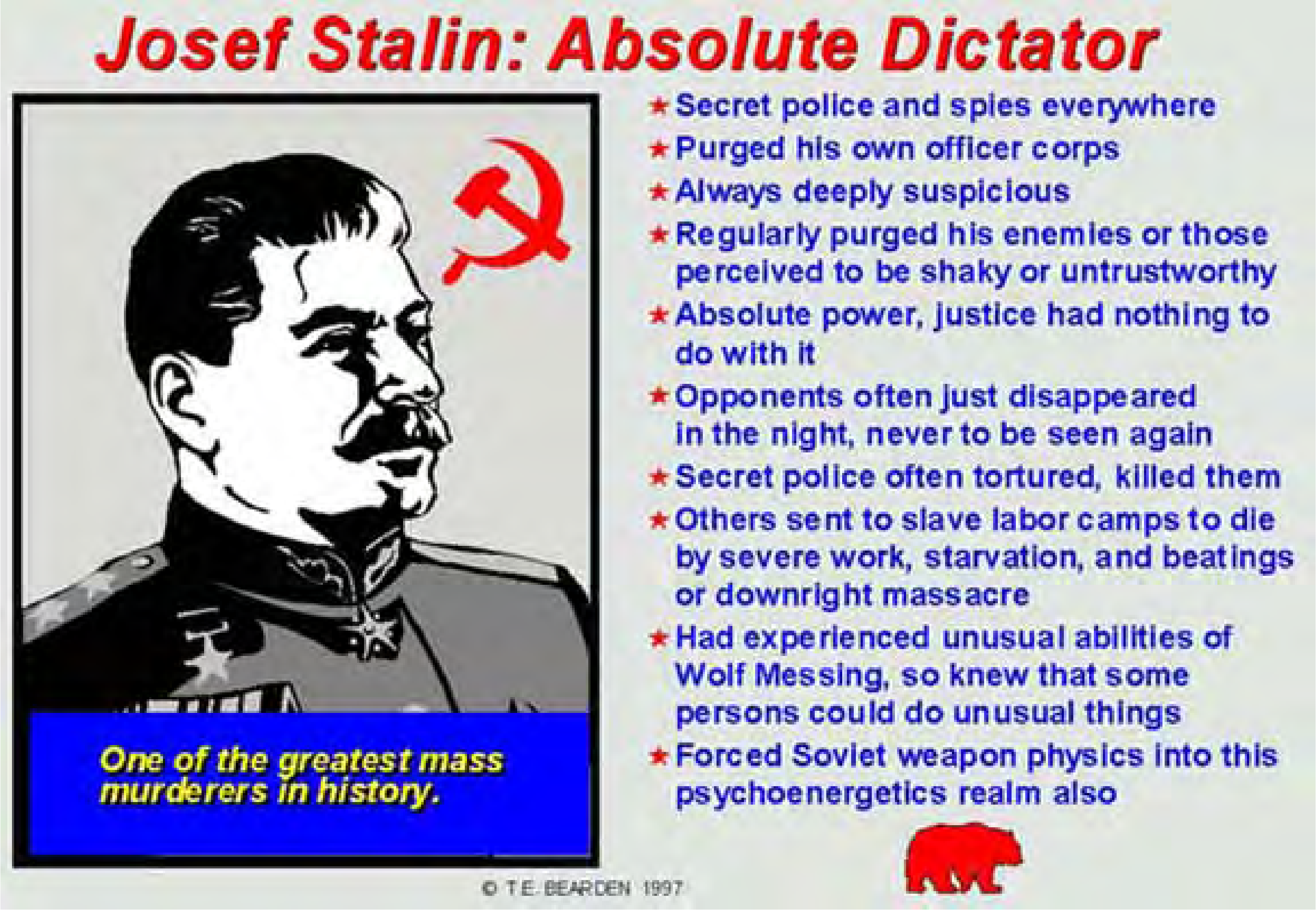

Josef Stalin: Absolute Dictator

- Secret police and spies everywhere
- Purged his own officer corps
- Always deeply suspicious
- Regularly purged his enemies or those perceived to be shaky or untrustworthy
- Absolute power, justice had nothing to do with it
- Opponents often just disappeared in the night, never to be seen again
- Secret police often tortured, killed them
- Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- Forced Soviet weapon physics into this psychoenergetics realm also

One of the greatest mass murderers in history.

© T.E. BEARDEN 1997

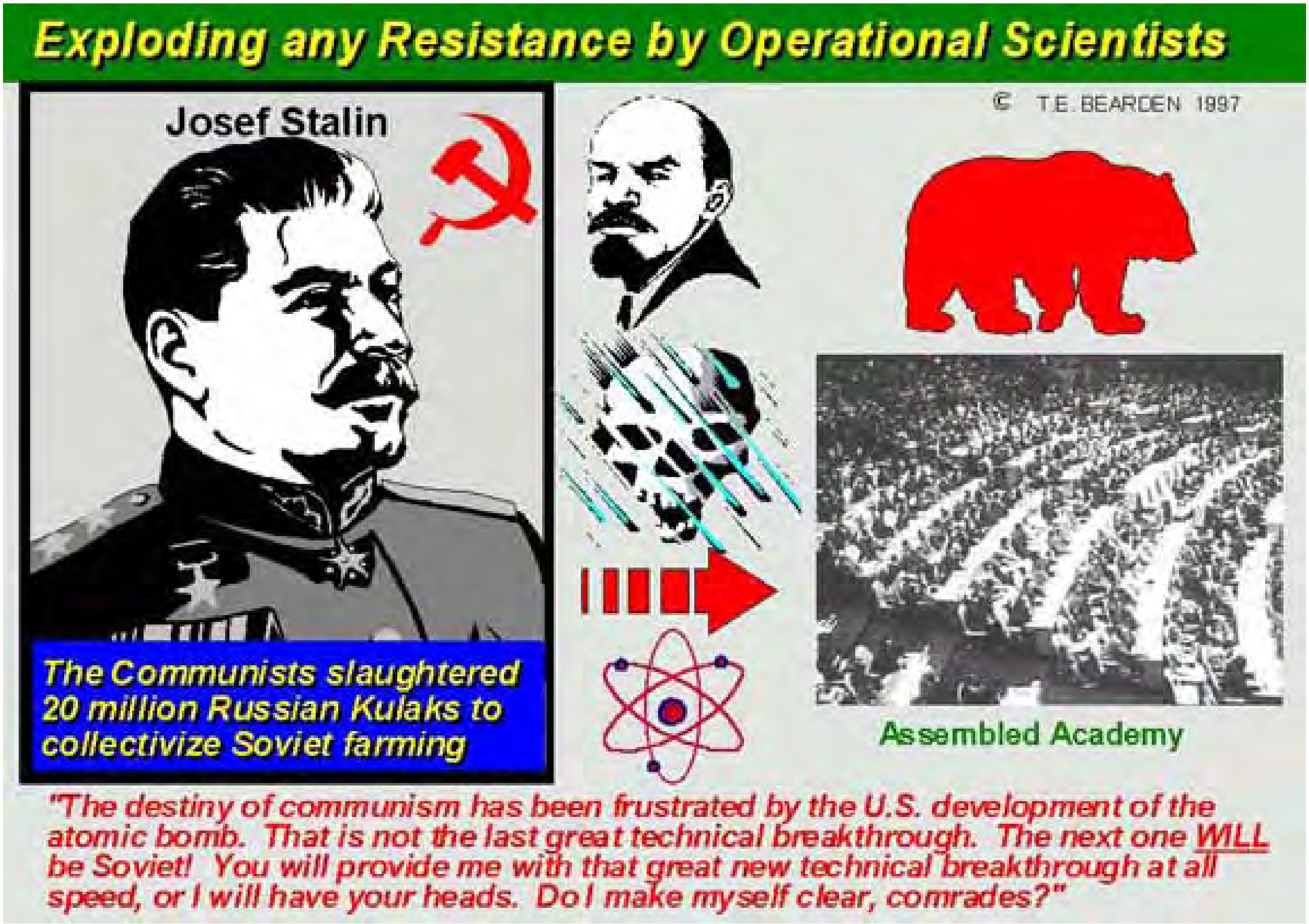

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

Josef Stalin

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

Assembled Academy

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

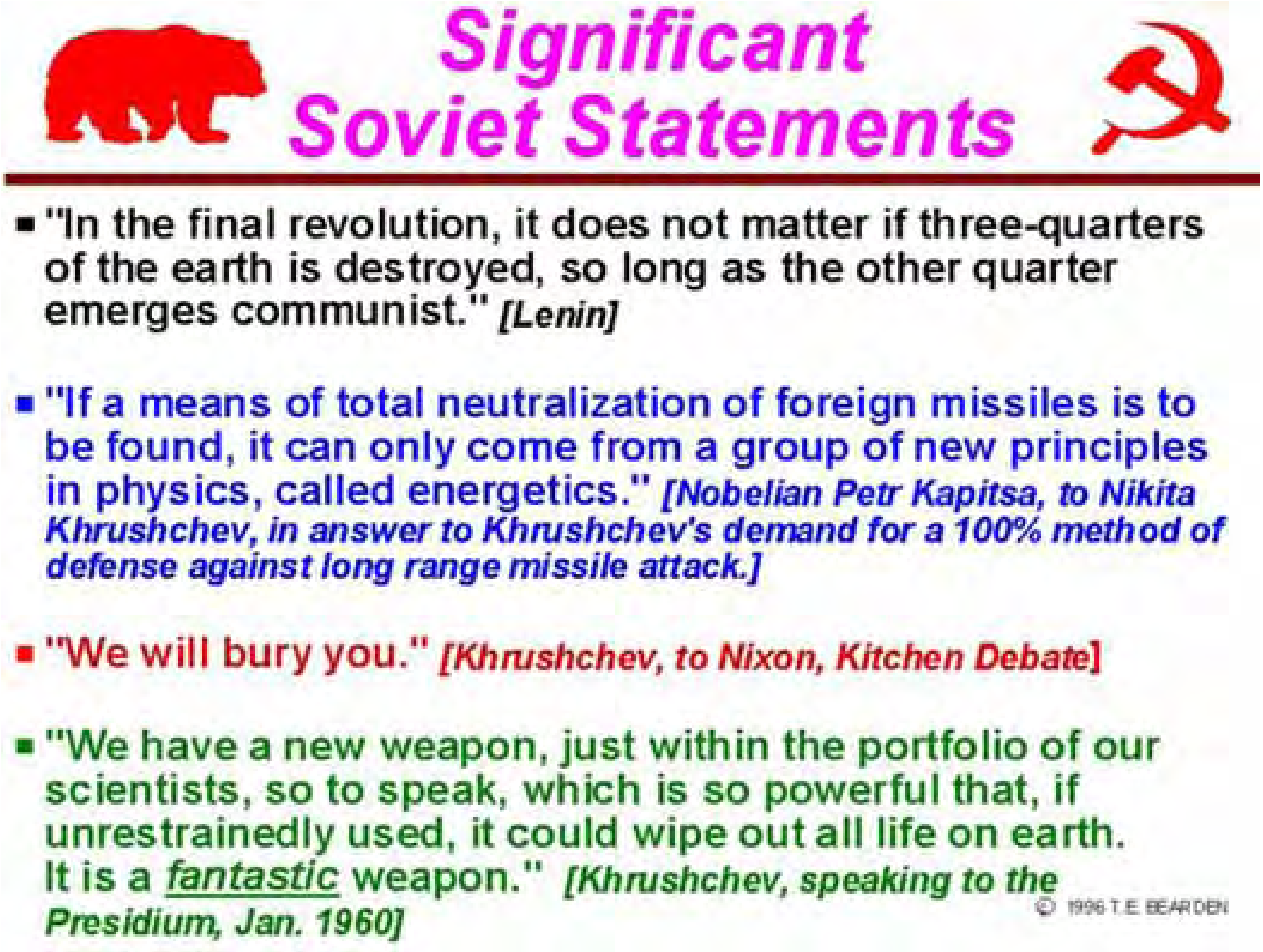

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*
- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." *[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]*
- "We will bury you." *[Khrushchev, to Nixon, Kitchen Debate*]
- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a *fantastic* weapon." *[Khrushchev, speaking to the Presidium, Jan. 1960]*

© 1996 T.E. BEARDEN

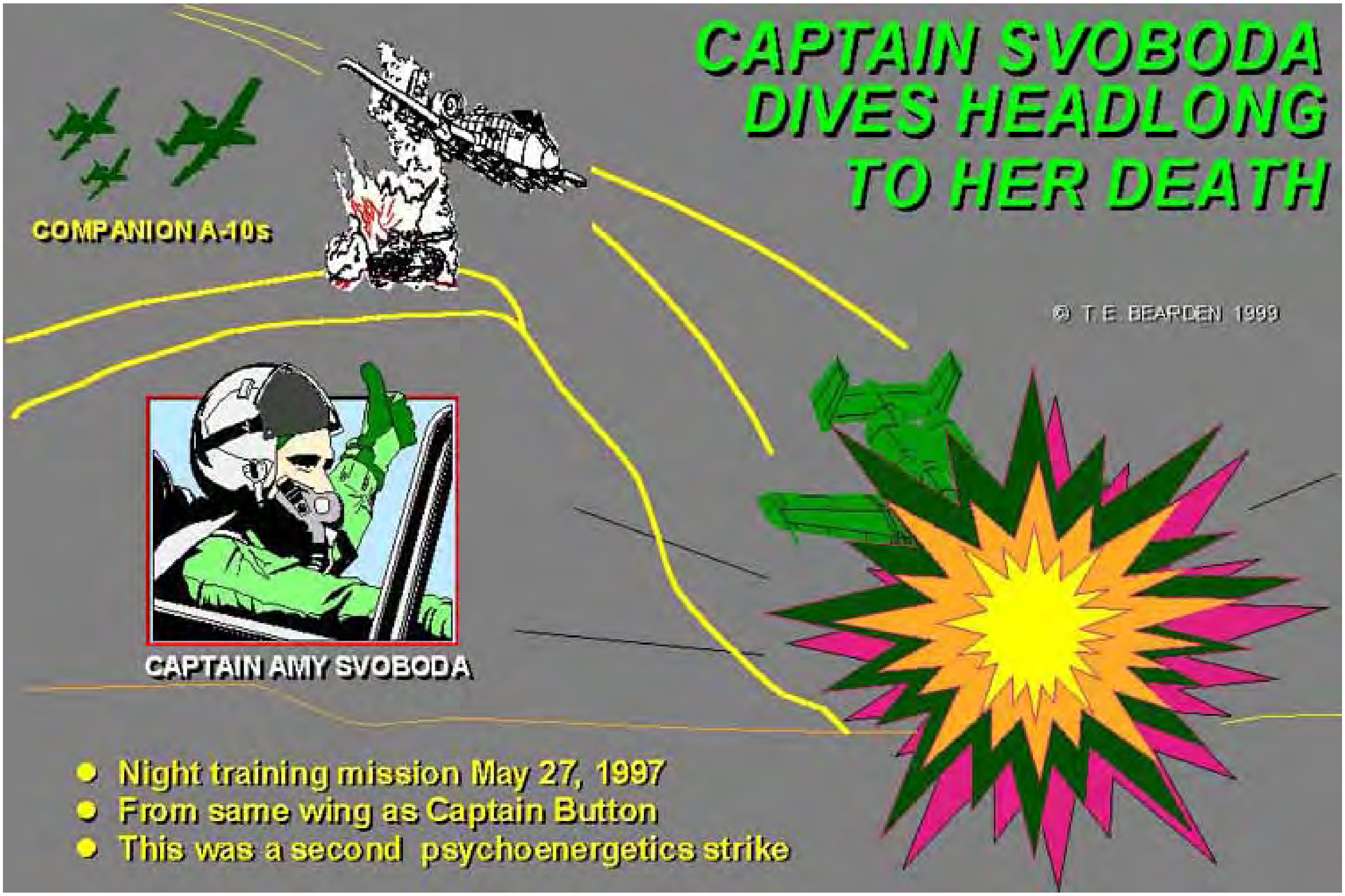

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

© T. E. BEARDEN 1999

COMPANION A-10s
CAPTAIN AMY SVOBODA

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

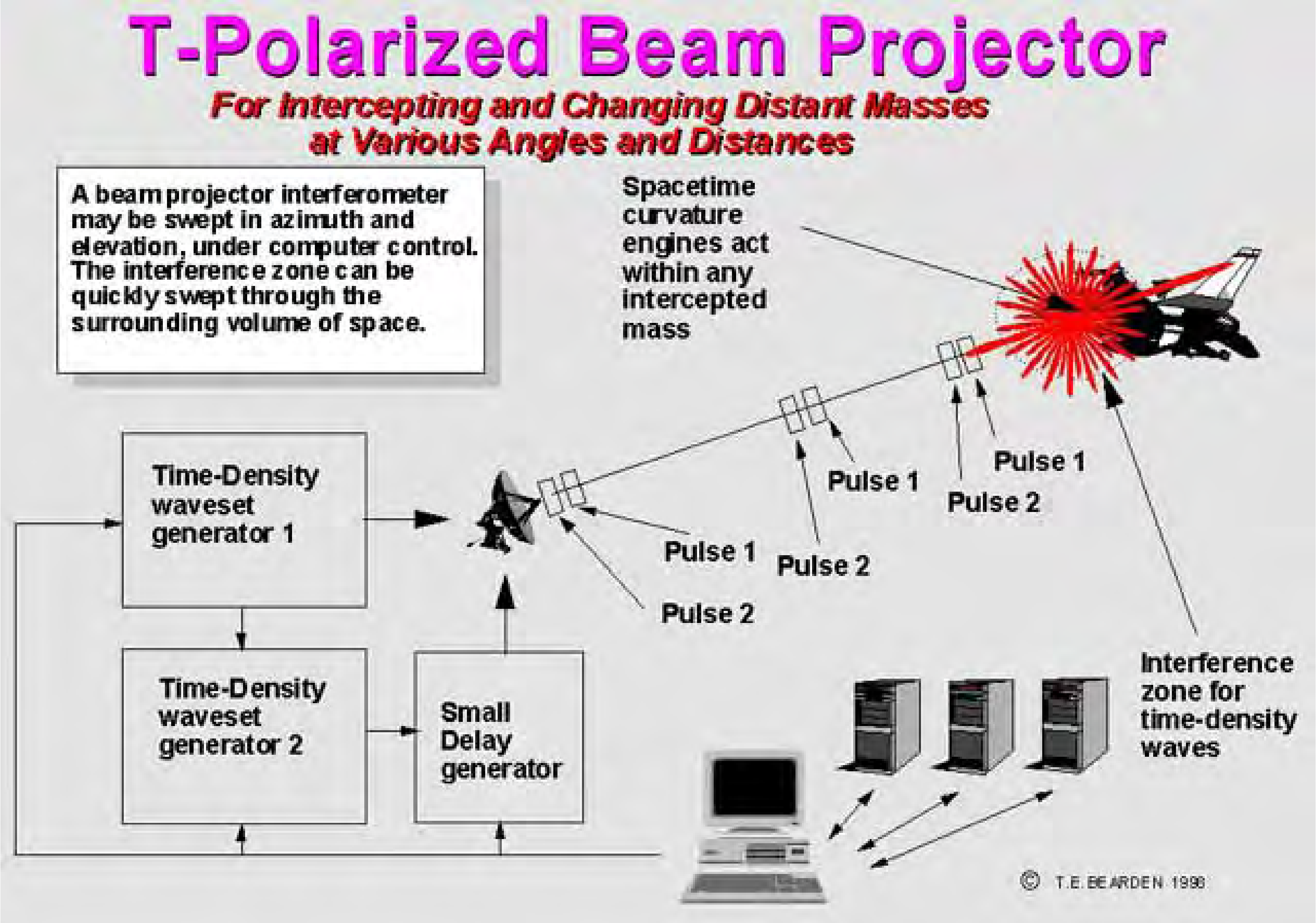

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass

Time-Density waveset generator 1

Time-Density waveset generator 2

Small Delay generator

Interference zone for time-density waves

© T.E. BEARDEN 1996

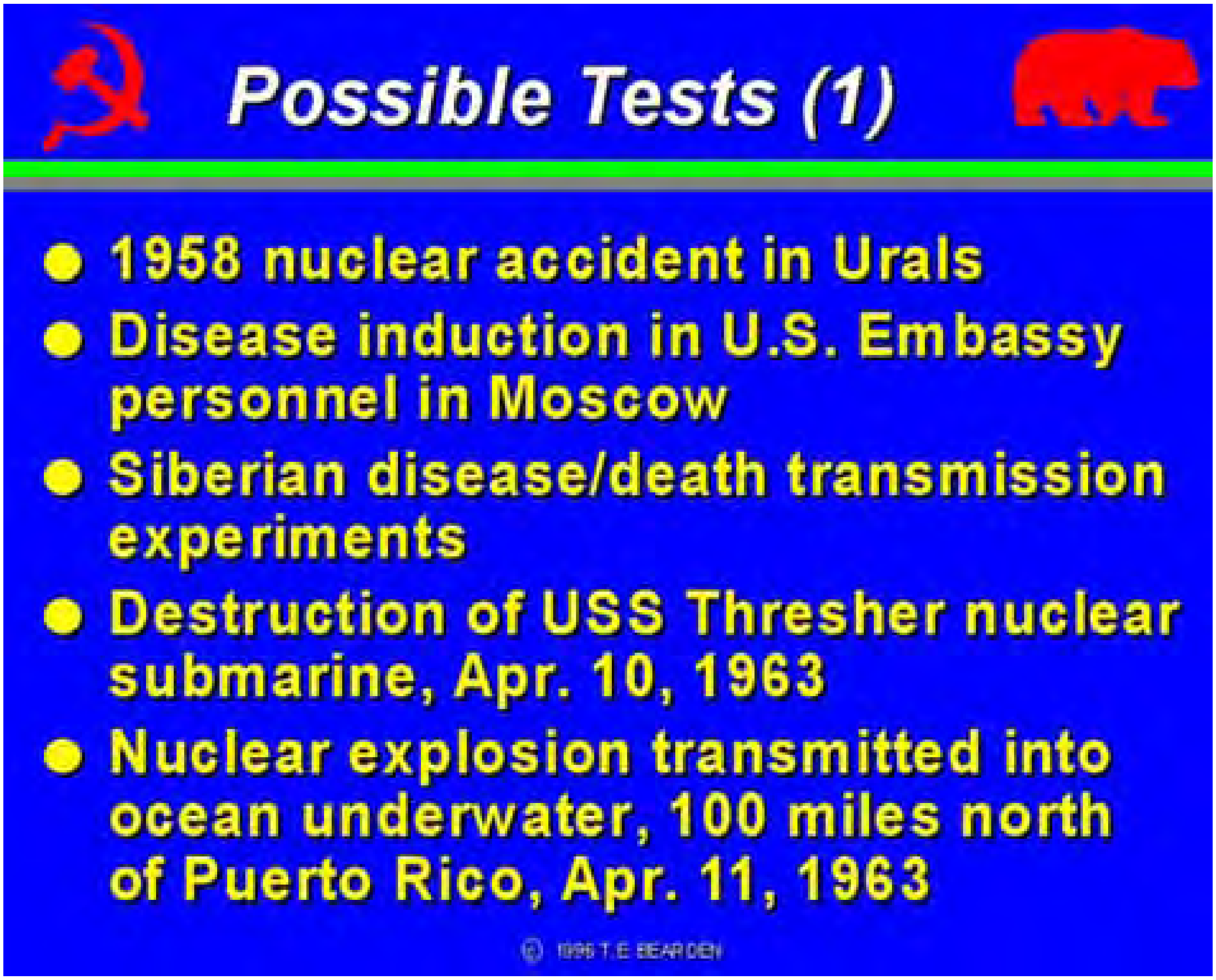

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T. E. BEARDEN

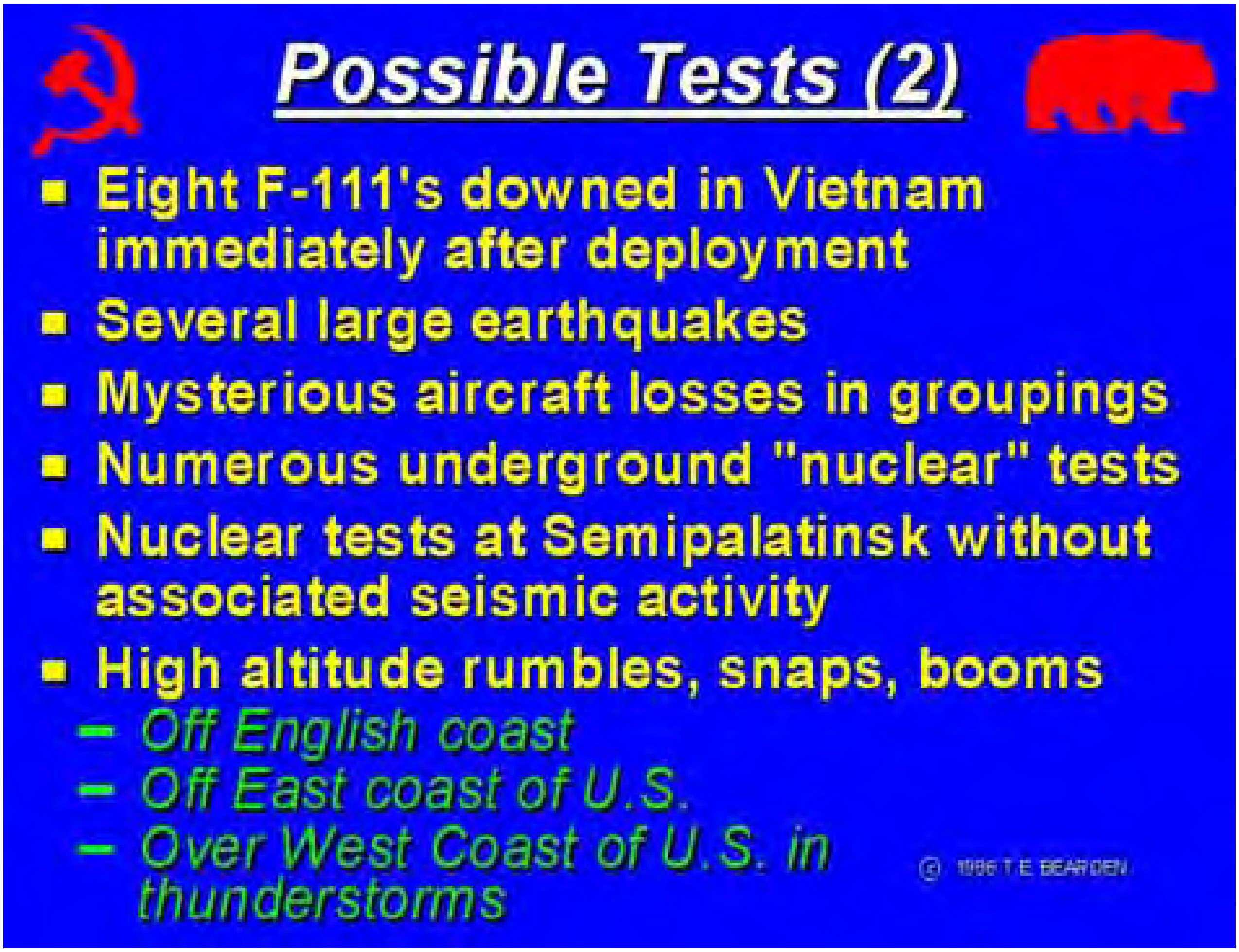

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T.E. BEARDEN

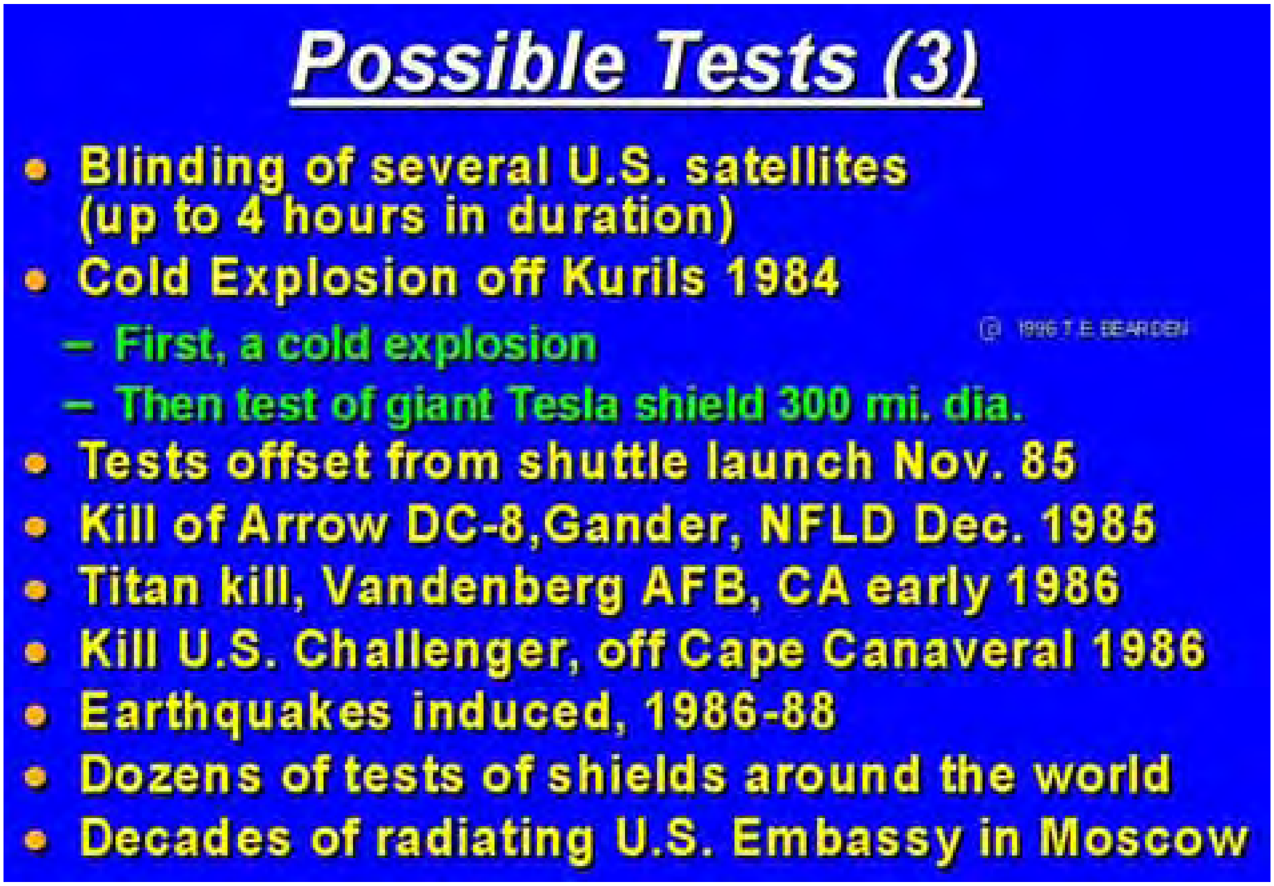

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T.E. BEARDEN

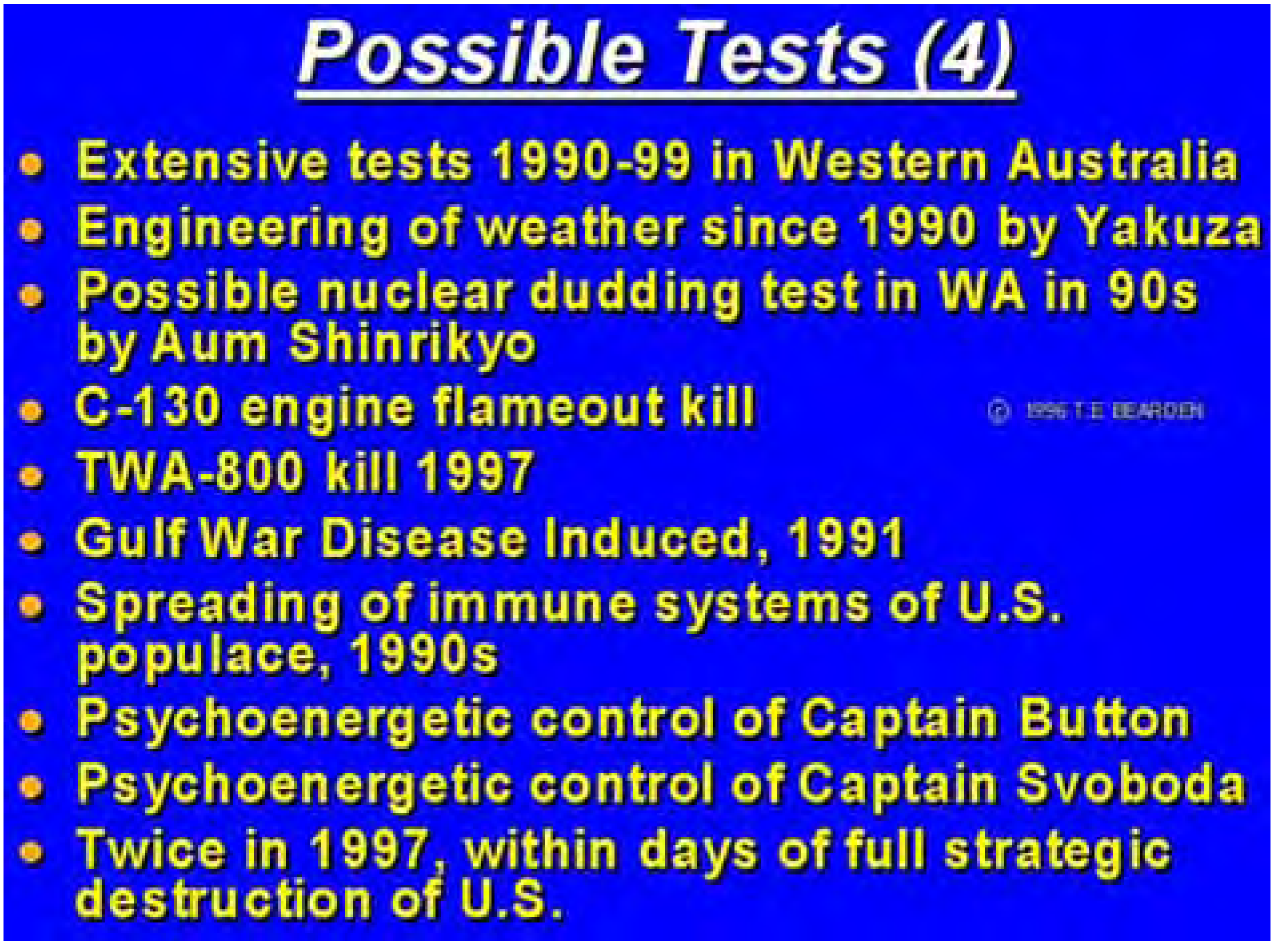

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E. BEARDEN

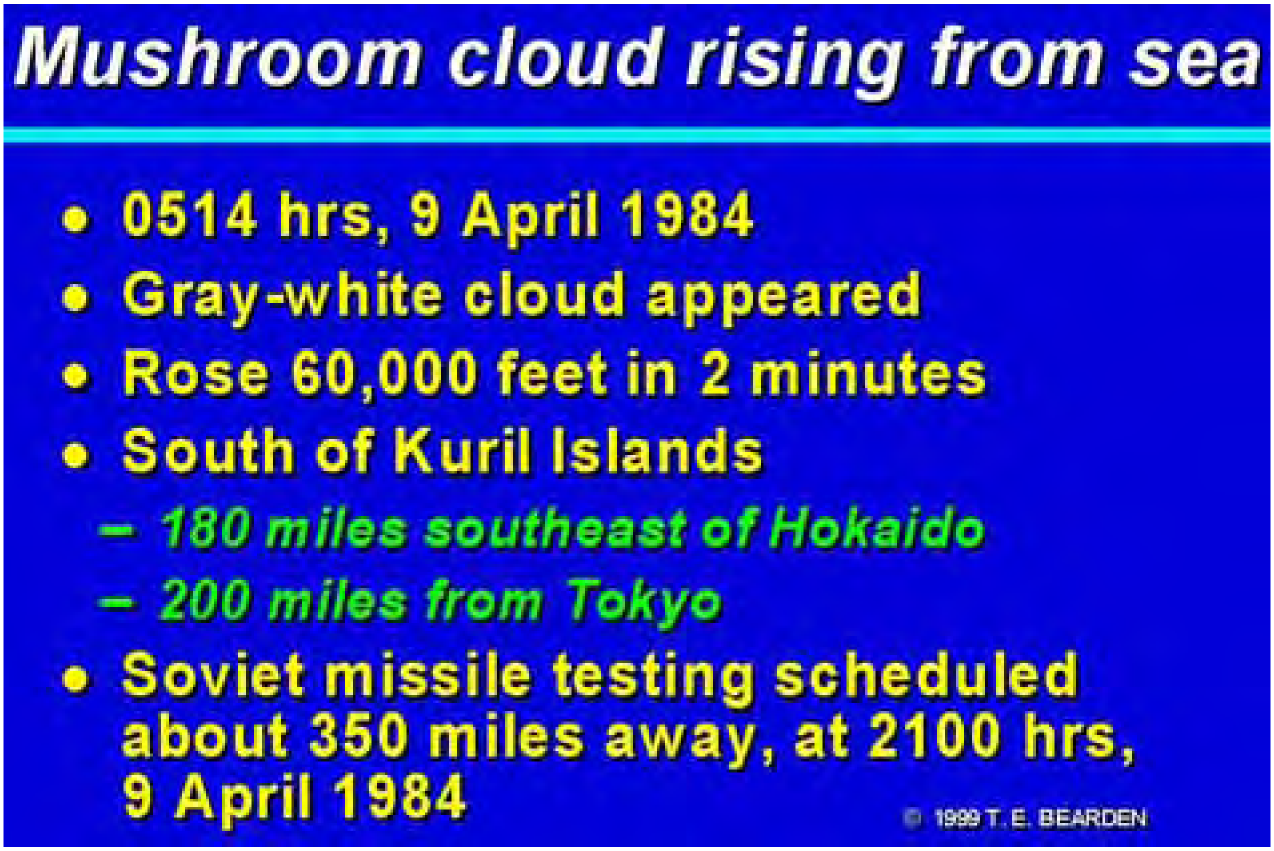

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - *180 miles southeast of Hokaido*
  - *200 miles from Tokyo*
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

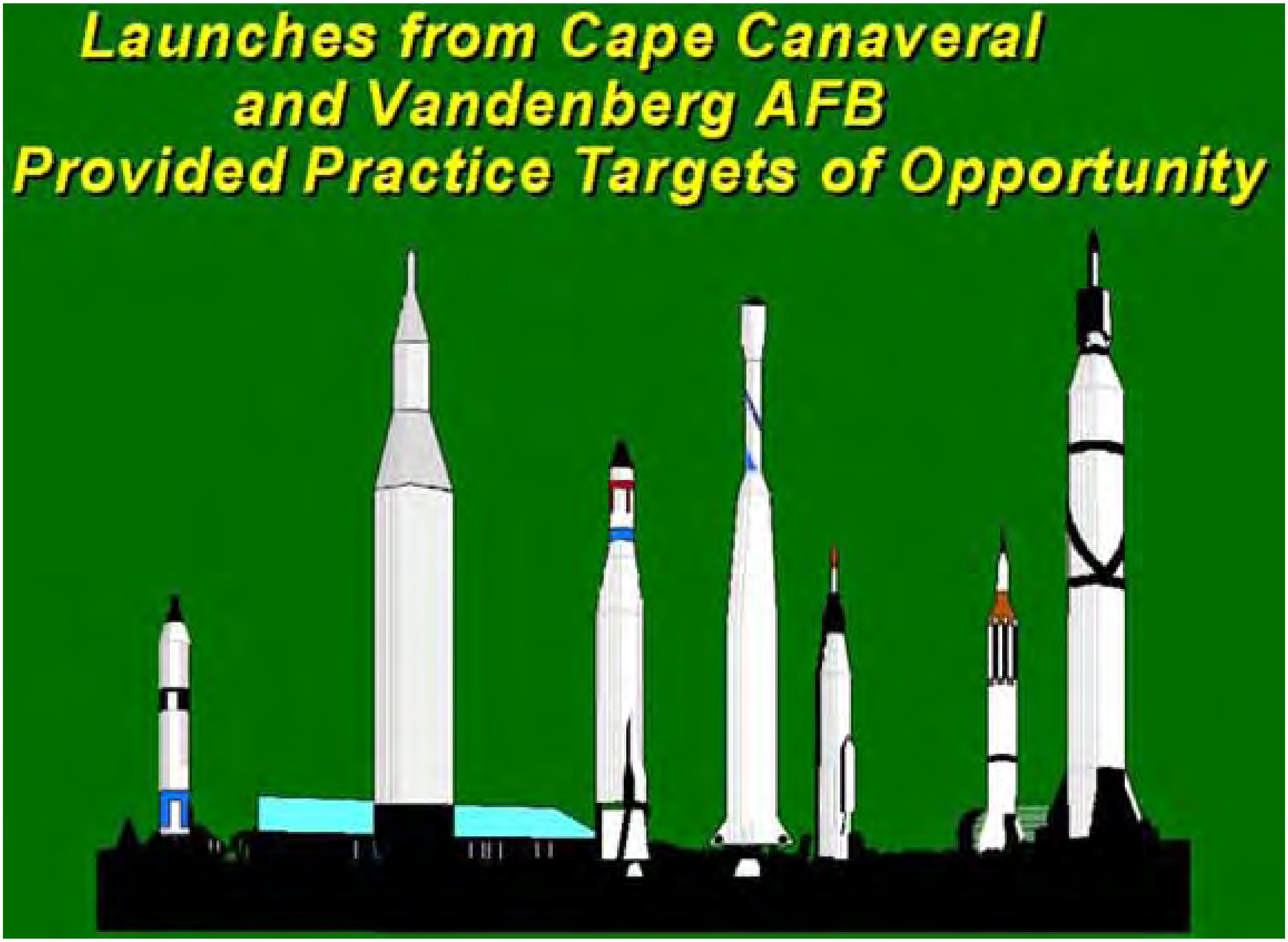
Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

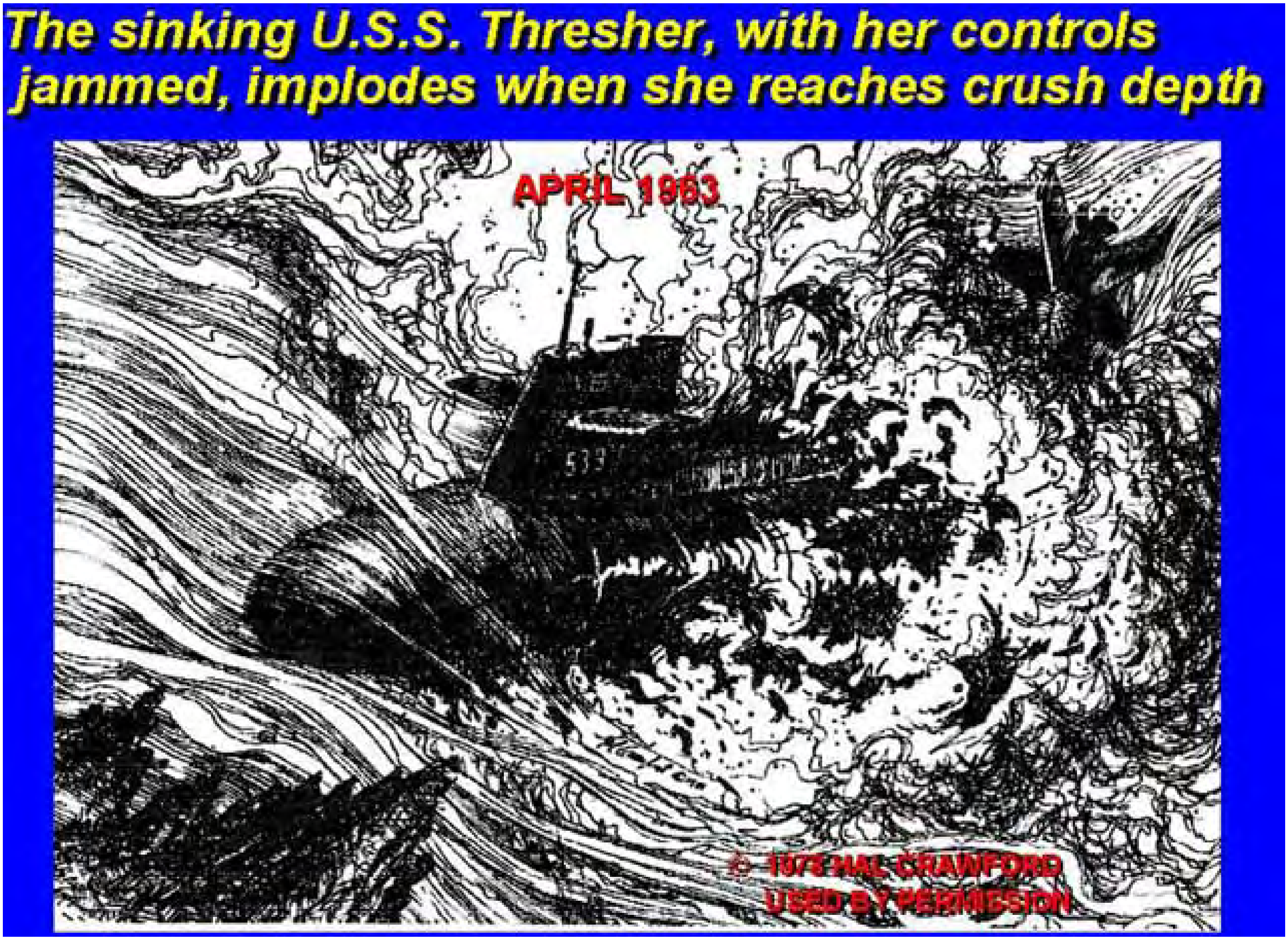

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

APRIL 1963
© 1978 HAL CRAWFORD
USED BY PERMISSION

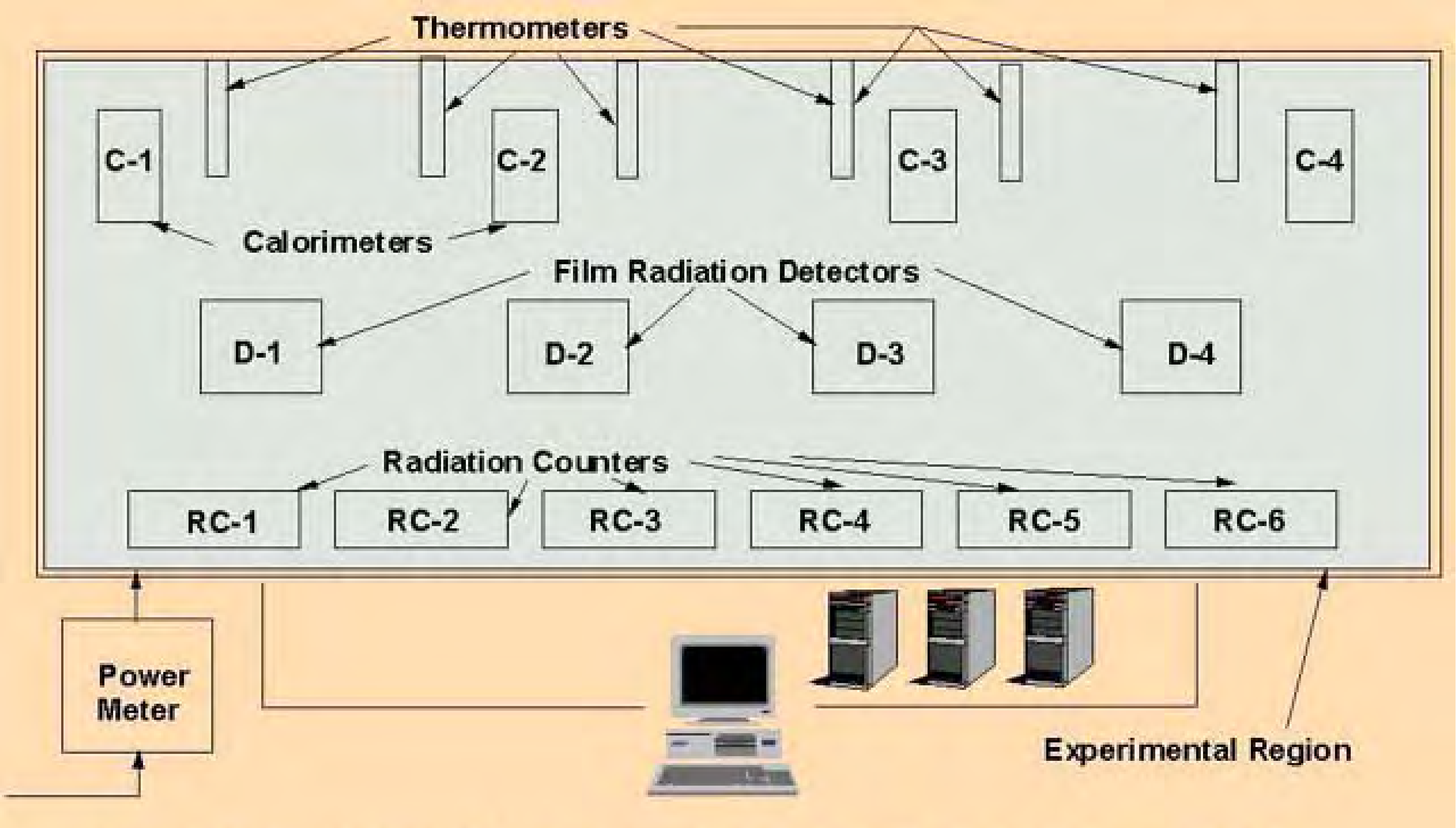
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

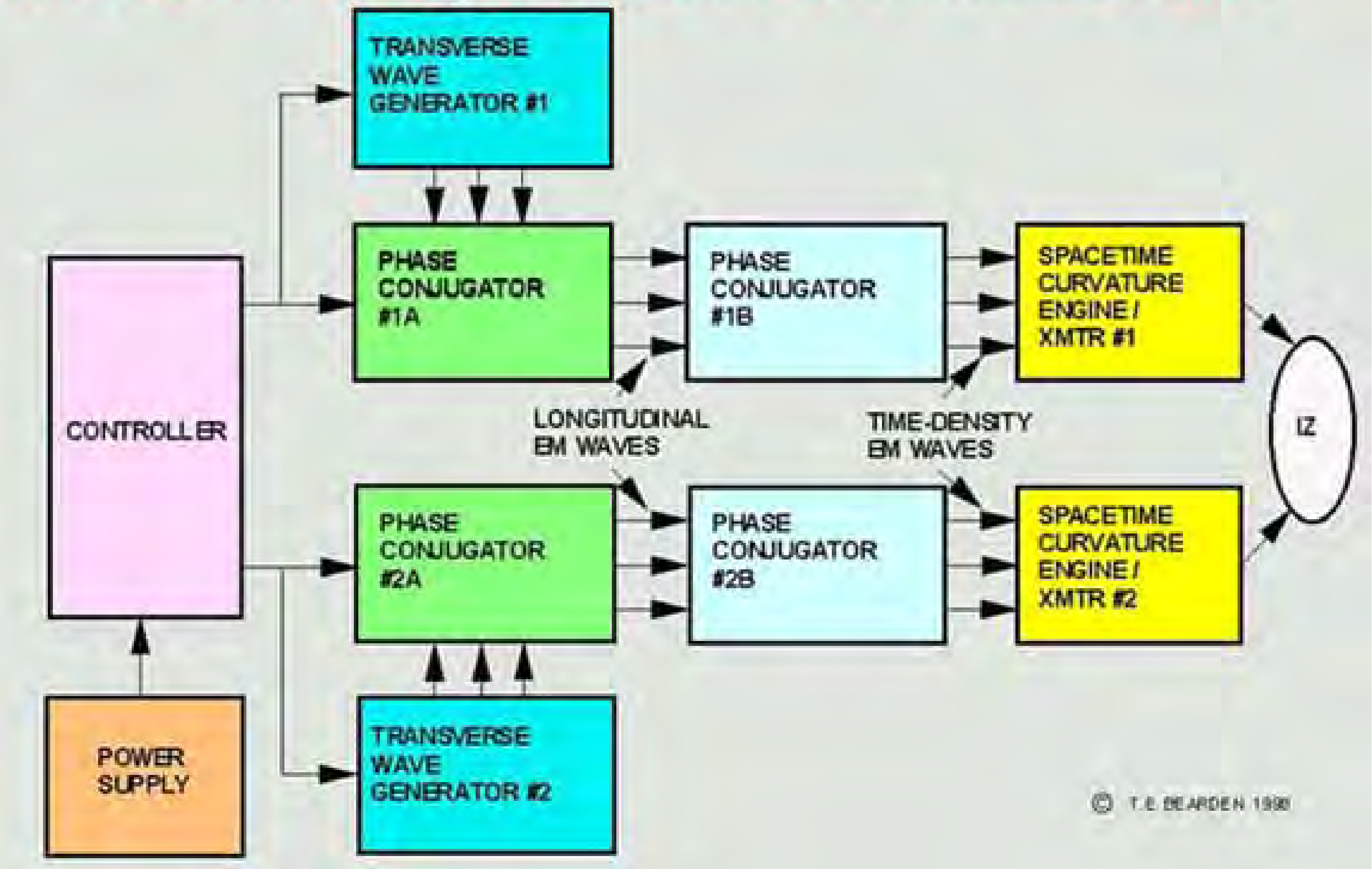
TRANSVERSE WAVE GENERATOR #1
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
CONTROLLER
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2
© T.E. BEARDEN 1990

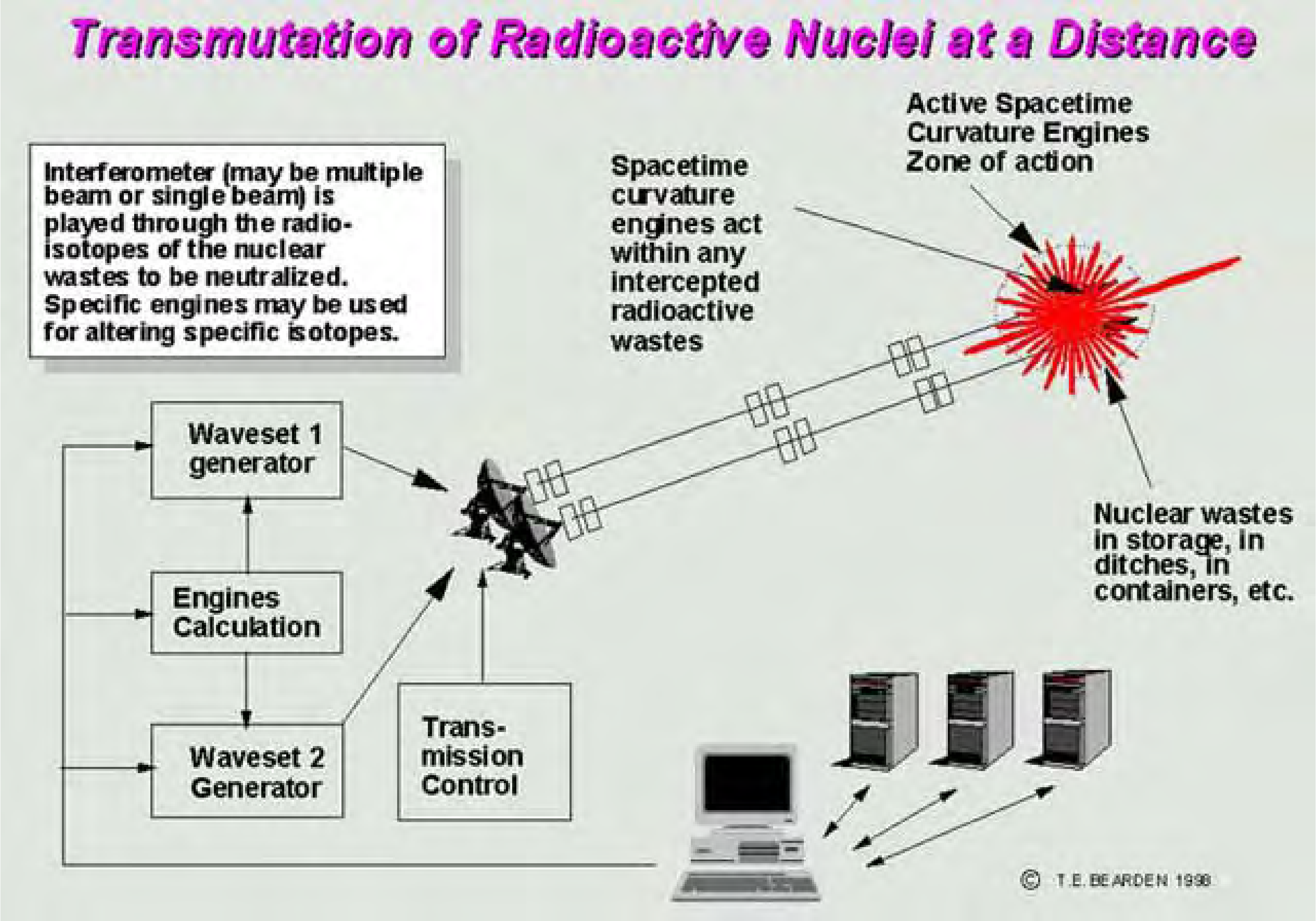

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radio-isotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action

Waveset 1 generator

Engines Calculation

Waveset 2 Generator

Trans-mission Control

Nuclear wastes in storage, in ditches, in containers, etc.

© T.E. BEARDEN 1998

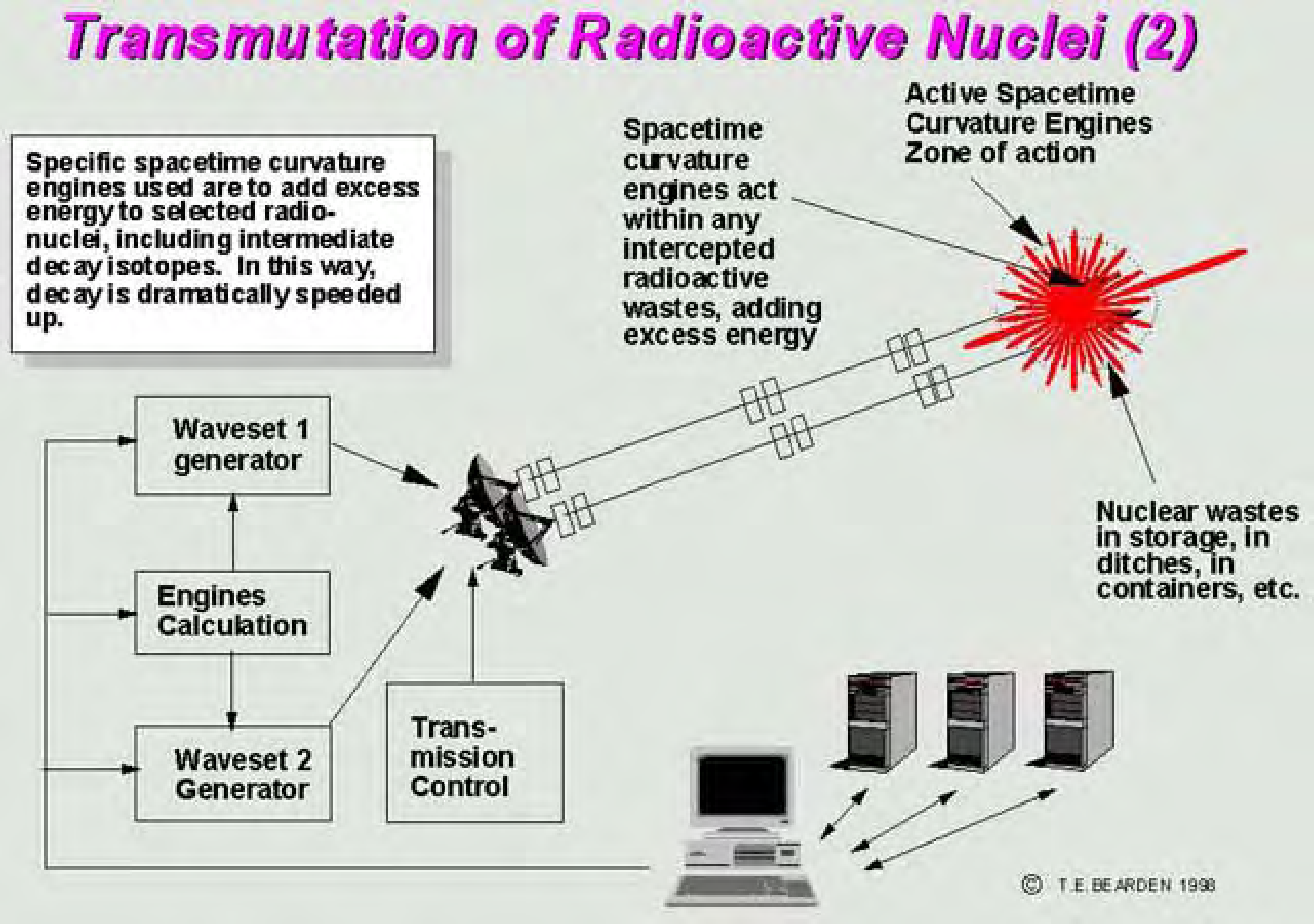

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radio-nuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control

© T.E. BEARDEN 1998

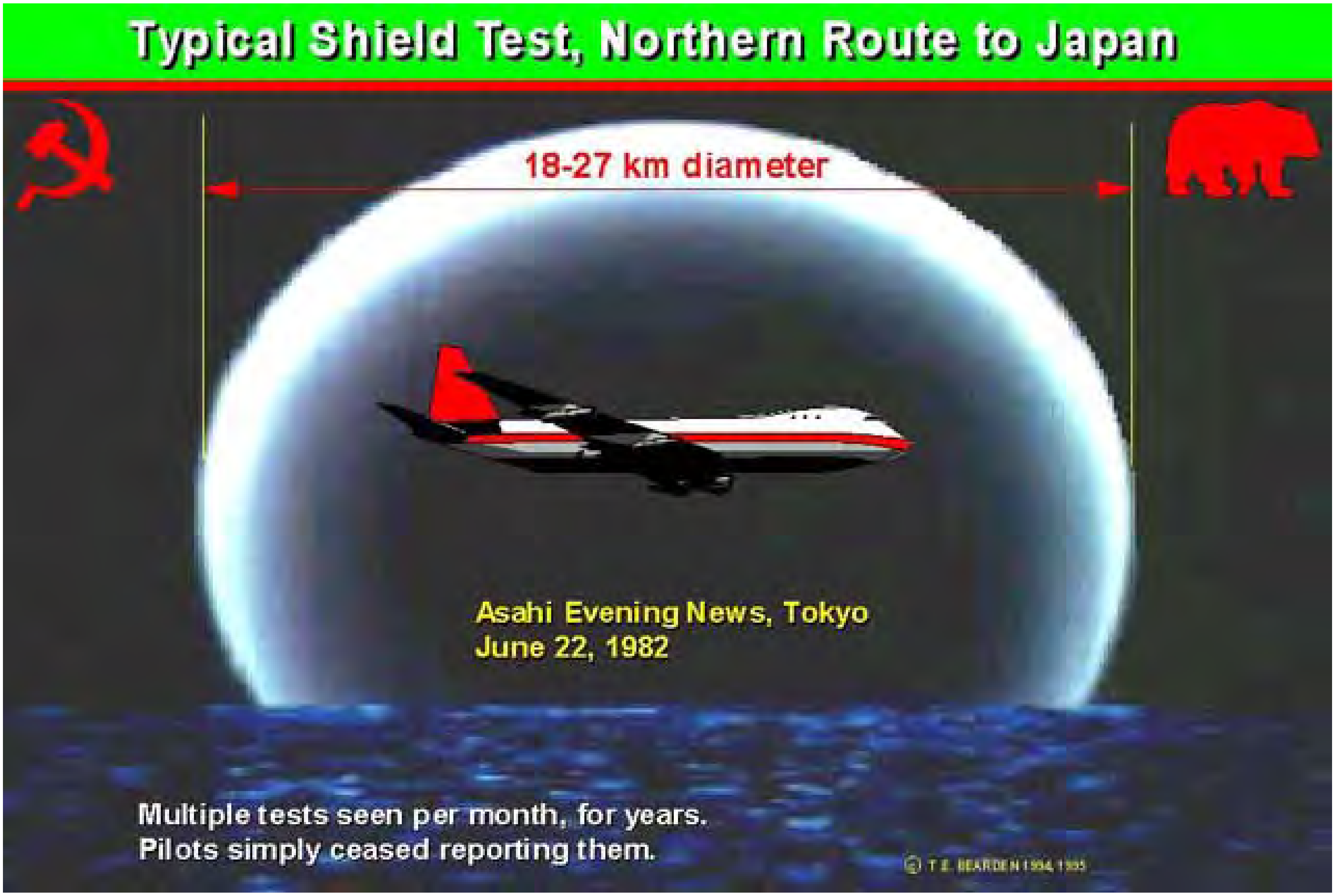

Typical Shield Test, Northern Route to Japan

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T. E. BEARDEN 1994, 1995

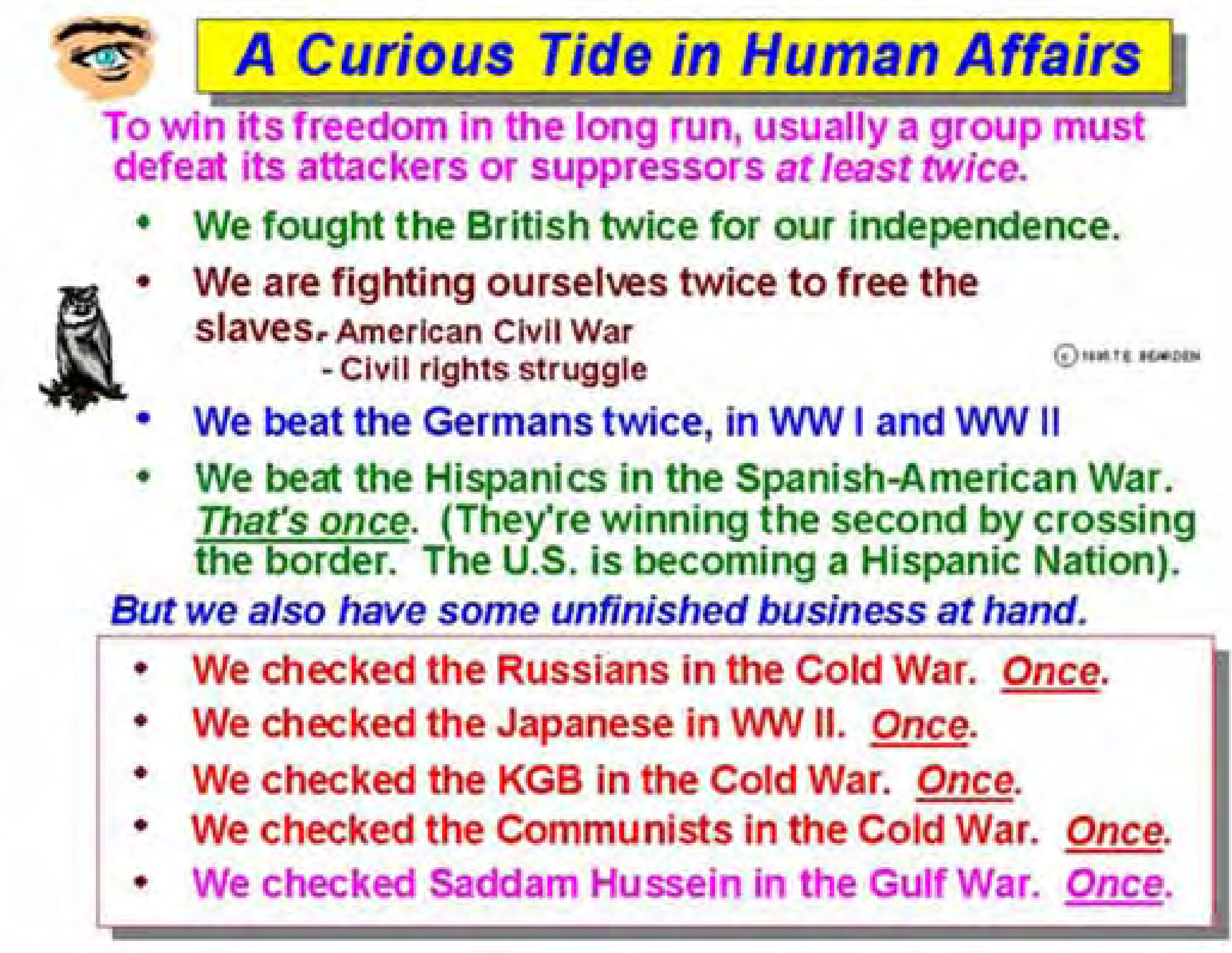

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

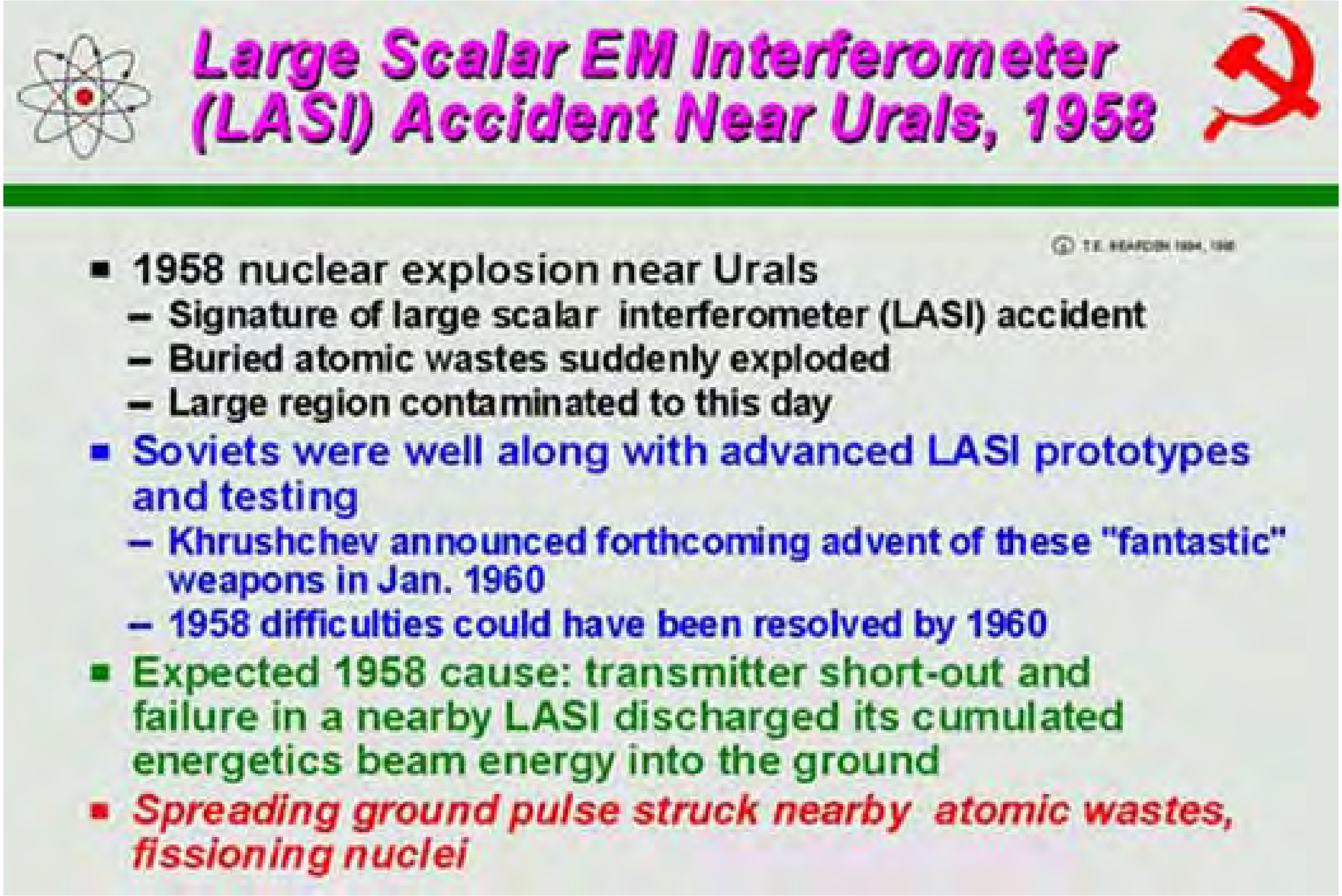

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

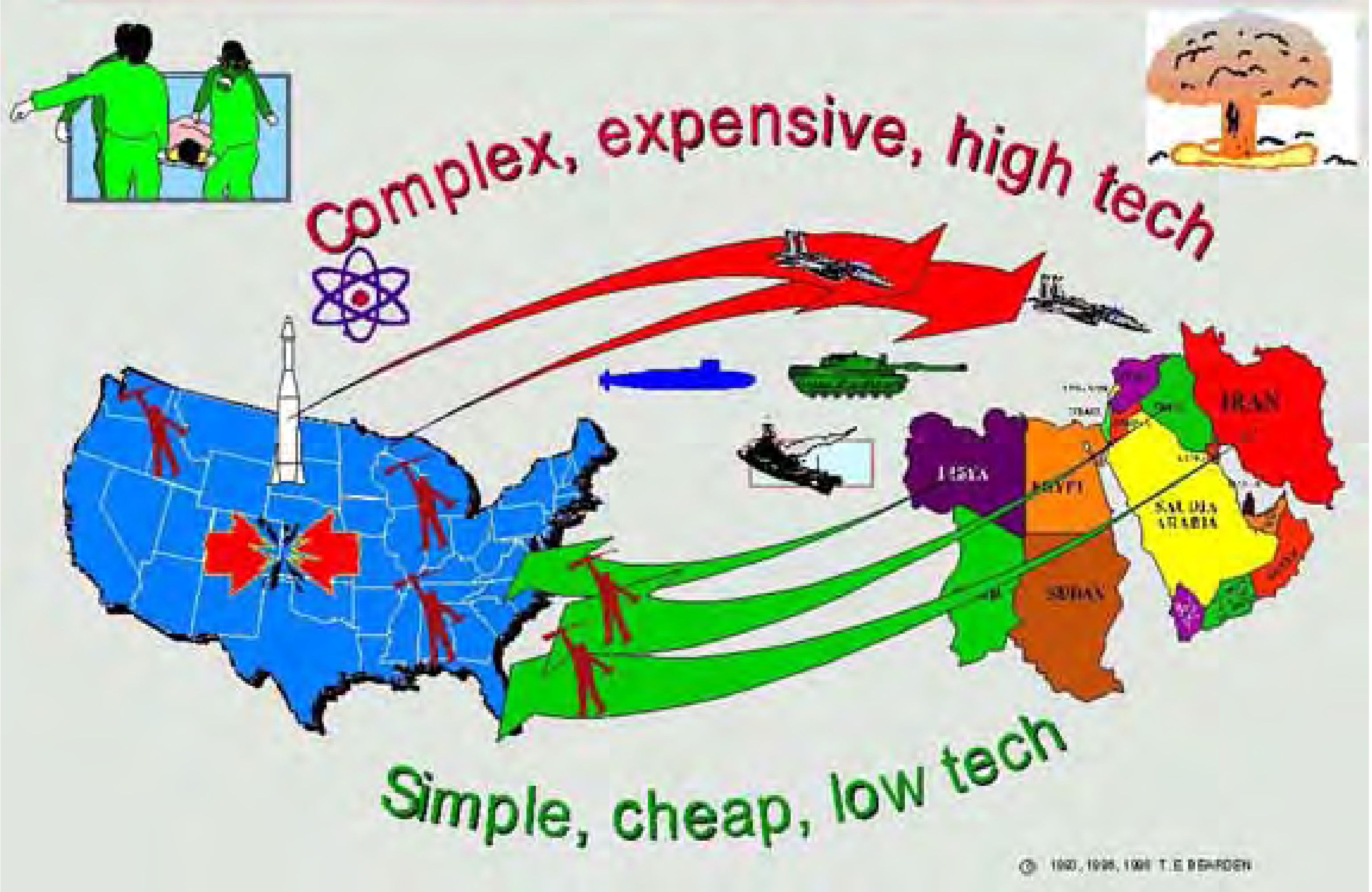

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN
© T. E. BEARDEN

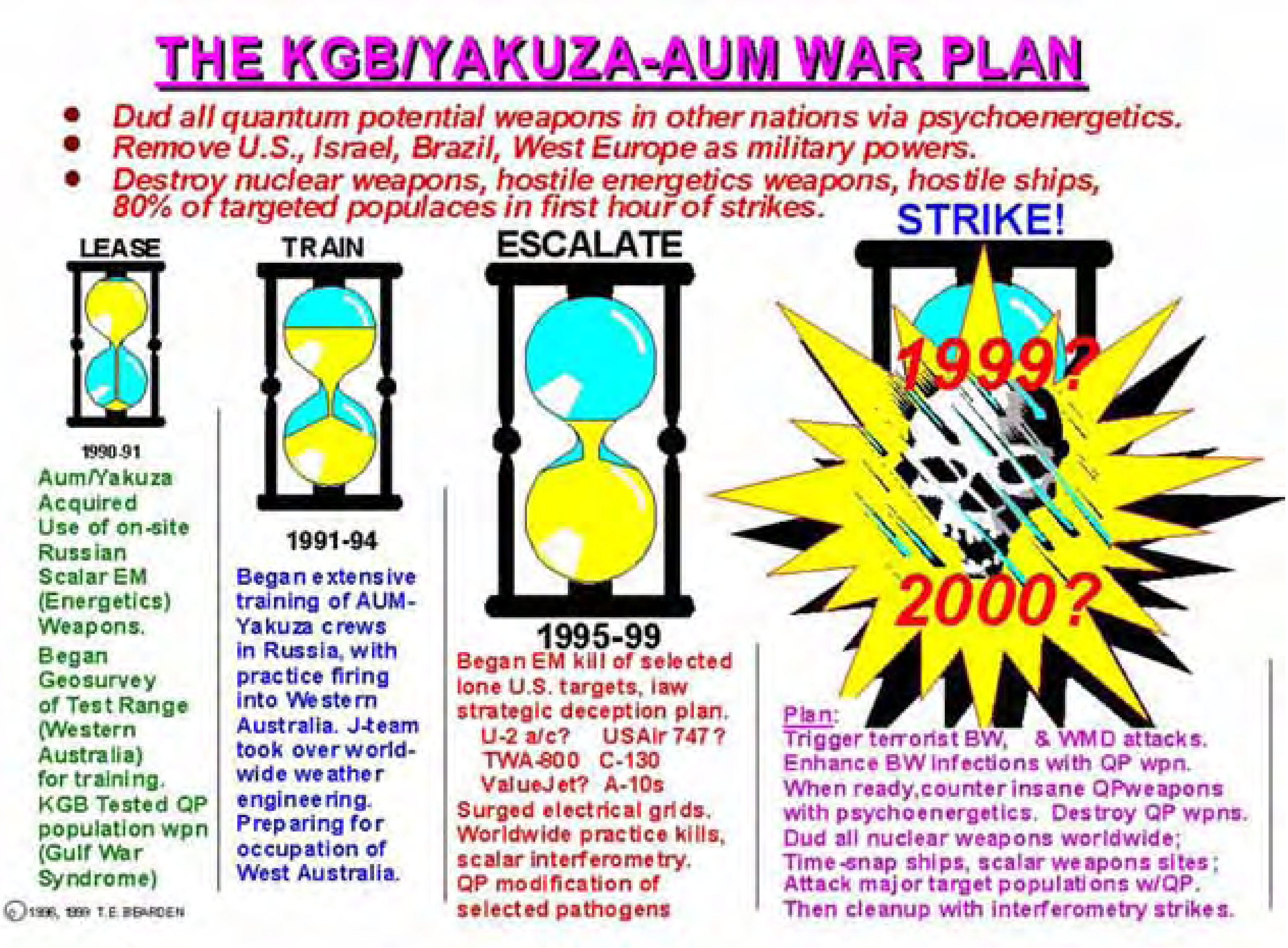

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training.
KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, law strategic deception plan.

U-2 a/c? USAir 747 ?
TWA-800 C-130
ValueJet? A-10s

Surged electrical grids.
Worldwide practice kills, scalar interferometry.
QP modification of selected pathogens

STRIKE!

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready,counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

©1996, 1999 T.E. BEARDEN

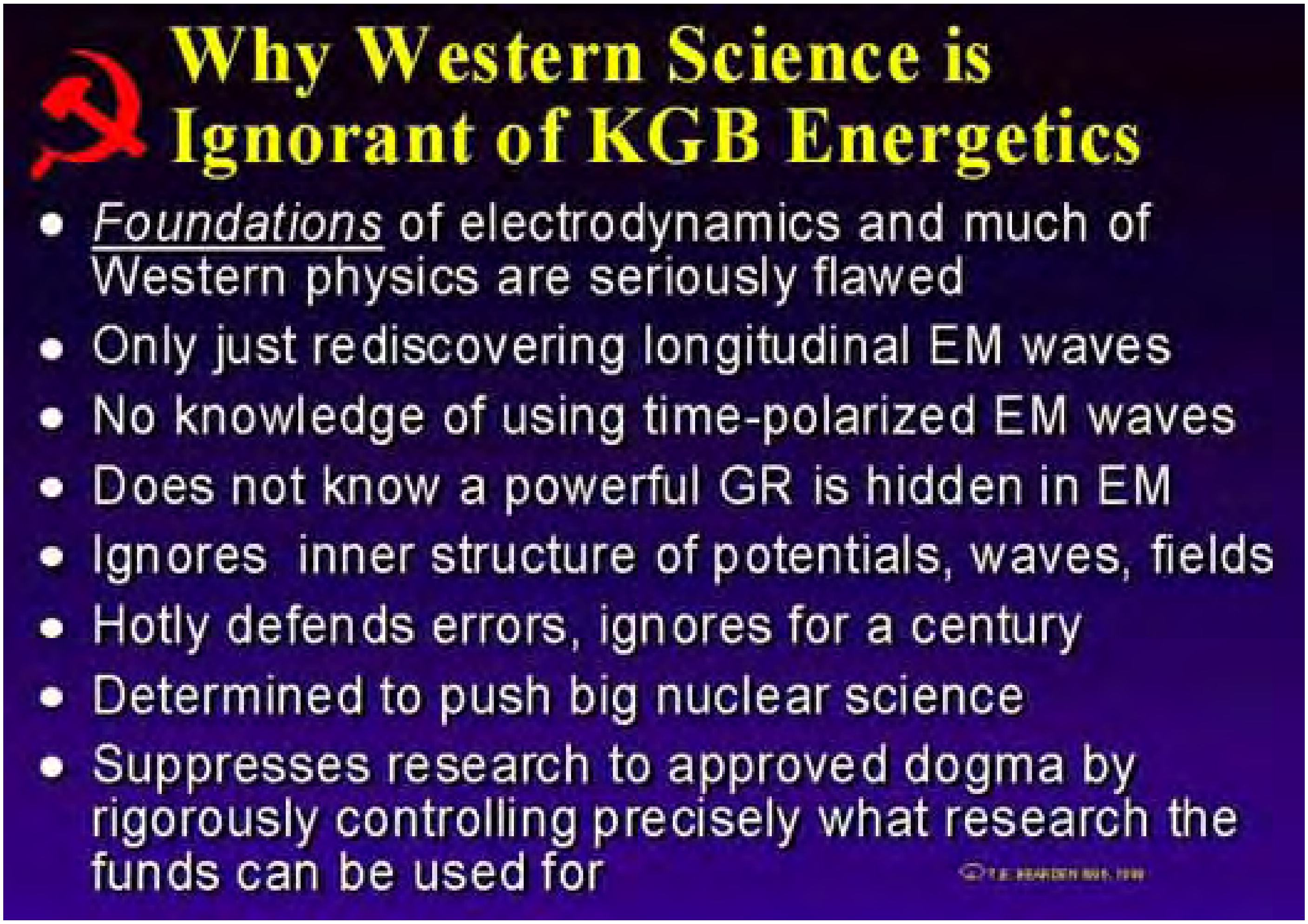

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

Transmutation of Radioactive Nuclei at a Distance

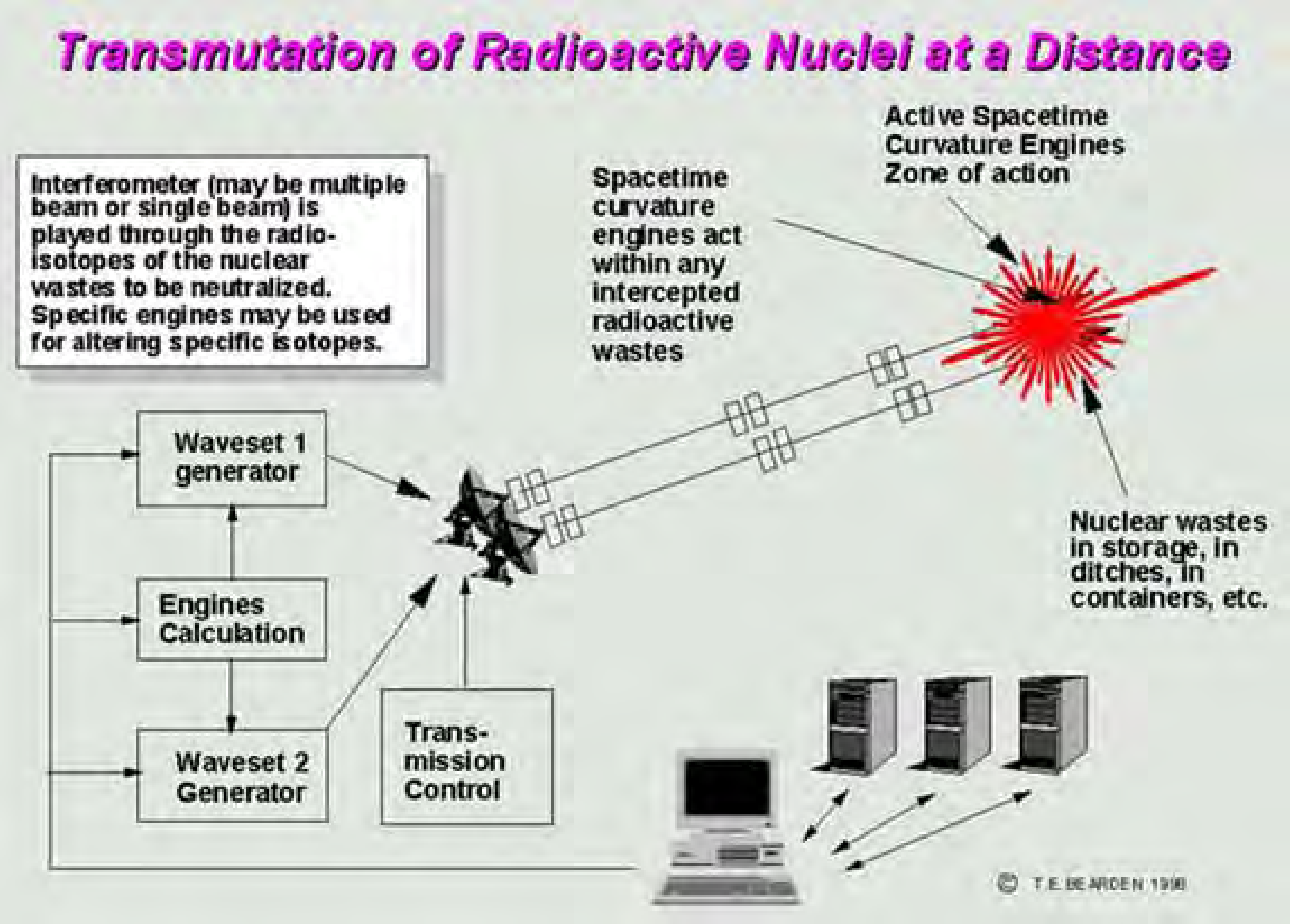

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action

Waveset 1 generator

Engines Calculation

Waveset 2 Generator

Trans-mission Control

Nuclear wastes in storage, in ditches, in containers, etc.

© T.E. BEARDEN 1998

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

The Last Samurai, Defending Bushido

- **3 Major Yakuza Groups, In Coalition.**
- **Penetrate Japanese Government.**
- **Open Offices, Signs on Street.**
- **Public Romanticizes Yakuza as Last Japanese Samurai**
- **In Large High-Tech Corporations.**
- **In Control of Japanese Banks.**
- **Japanese Real-Estate Rip-off.**
- **$500 - 1,000 Billion per Year Income.**
- **Rapidly Spreading in U.S.**
- **Uniting Mafia Groups Worldwide.**
- **Coalition with Aum Shinrikyo.**
- **Leased Russian Energetics Weapons.**
- **Crews On-site in Russia, Trained.**
- **Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)**
- **Now in Escalation Phase, Killing U.S. Targets.**
- **All-out Strike on U.S. ready 1999-2000.**
- **Now held in check by KGB, dead-man fuzed by small nation friendly to U.S.**

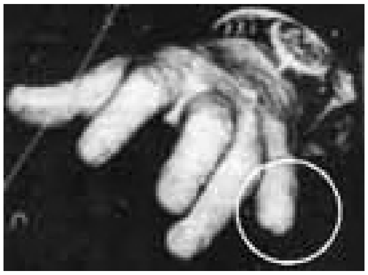

© 1995, 1996, 2001 T.E. BEARDEN

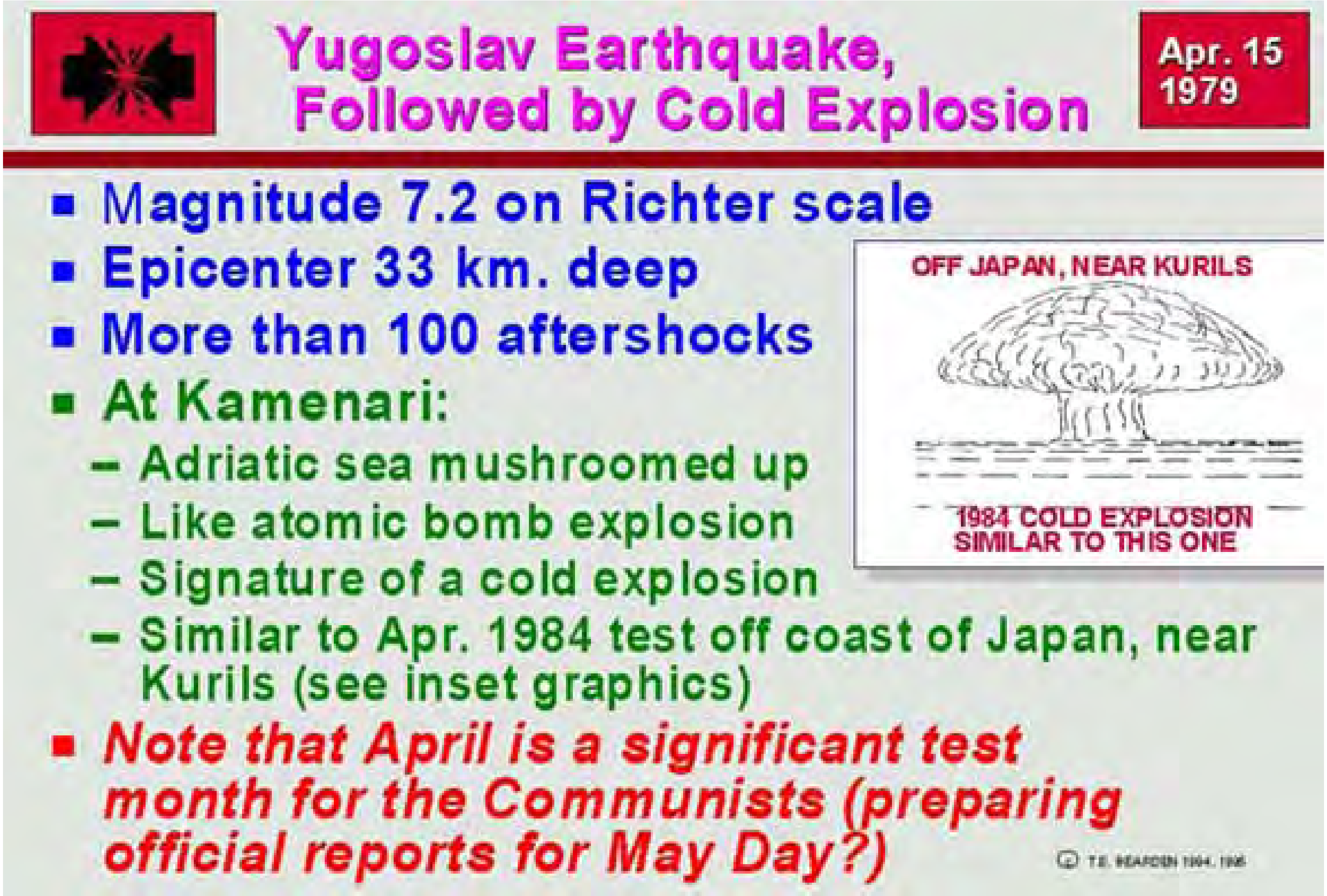

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

Technical Background on the Priore Healing Process

Contents

1. Technical Background on the Priore Healing Process
2. Transduction of EM wave types by successive phase conjugation.
3. Time-charging and Time-density excited states
4. Interference for transduction in slow decay of time-density excited states
5. Infolded longitudinal biwaves composition of a scalar potential
6. Scalar potential interferometry produces all EM force fields and waves
7. Some serious flaws and errors in classical EM theory (#1-16)
8. Some serious flaws and errors in classical EM theory (#17-32)
9. Abnormal penetration depth of weak signals due to photon interaction
10. Amplified replica of body's signal complex to reverse cellular disease & damage
11. Amplified replica of body's signal complex to correct a congenital condition

Technical Background on the Priore Healing Process

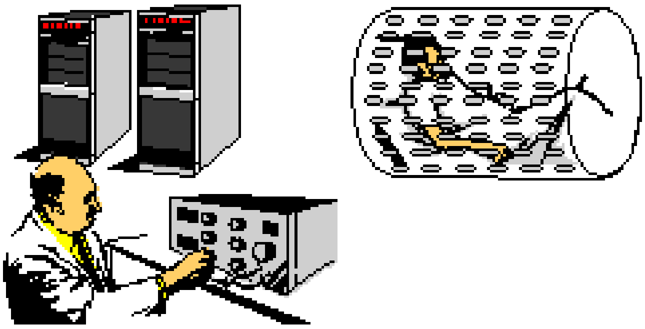

© T.E. BEARDEN 2001

Published by Lotus® Freelance Graphics®

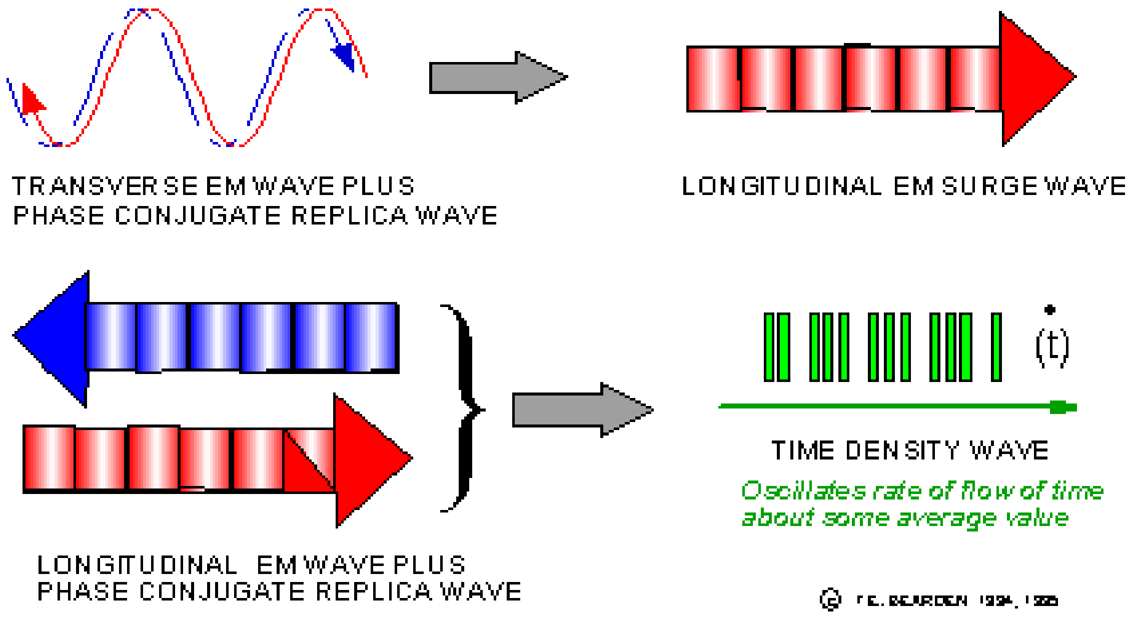
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
(t)
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994, 1995

Figure 1. Transduction of EM wave types by successive phase conjugation.

Published by Lotus® Freelance Graphics®

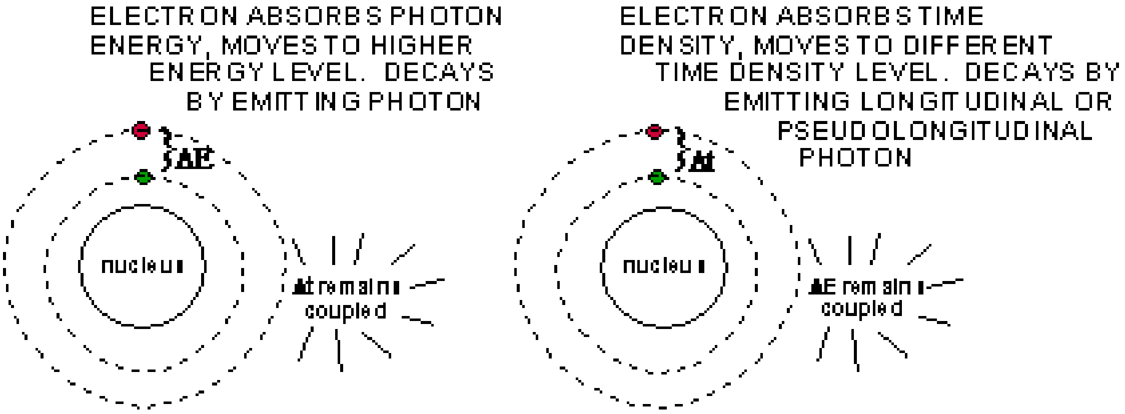
ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON
ΔE
nucleus
Δt remains coupled
ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON
Δt
nucleus
ΔE remains coupled

Figure A. Excited energy density level.

Figure B. Excited time density level.

© T.E. BEARDEN
1993

Figure 2. Time-charging and Time-density excited states, contrasted to ordinary spatial energy excited state.

Published by Lotus® Freelance Graphics®

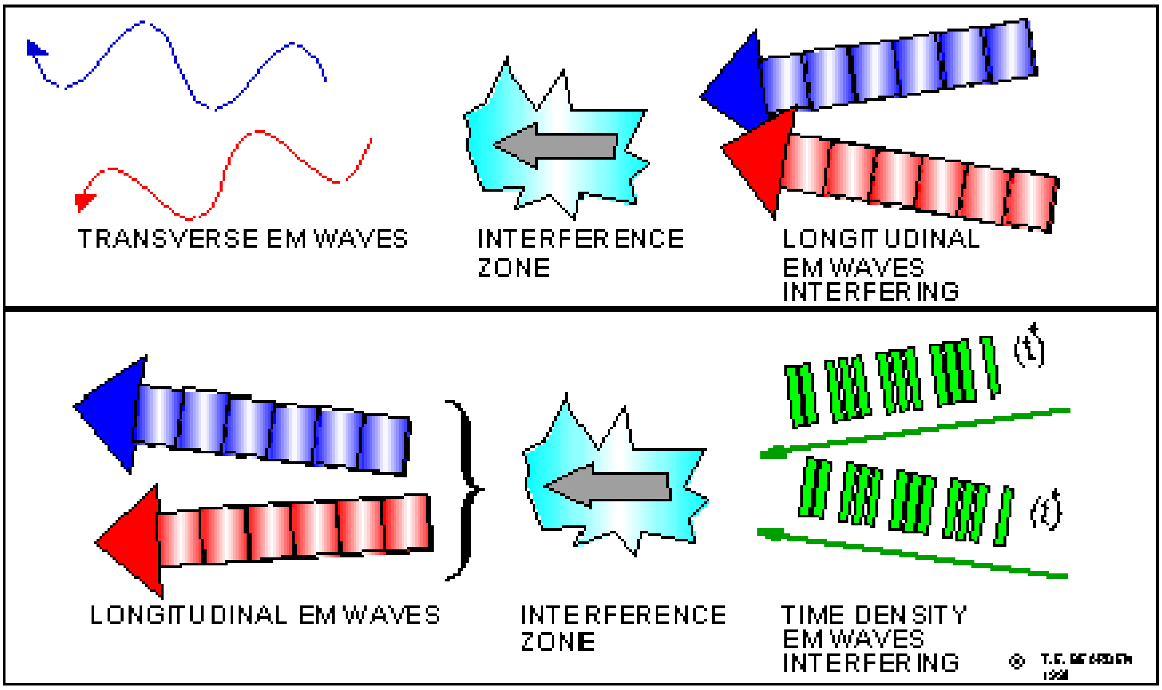
TRANSVERSE EM WAVES
INTERFERENCE ZONE
LONGITUDINAL EM WAVES INTERFERING
LONGITUDINAL EM WAVES
INTERFERENCE ZONE
(t)
(t)
TIME DENSITY EM WAVES INTERFERING

Figure 3. Interference processes for transduction in slow decay of time-density excited states.

Published by Lotus® Freelance Graphics®

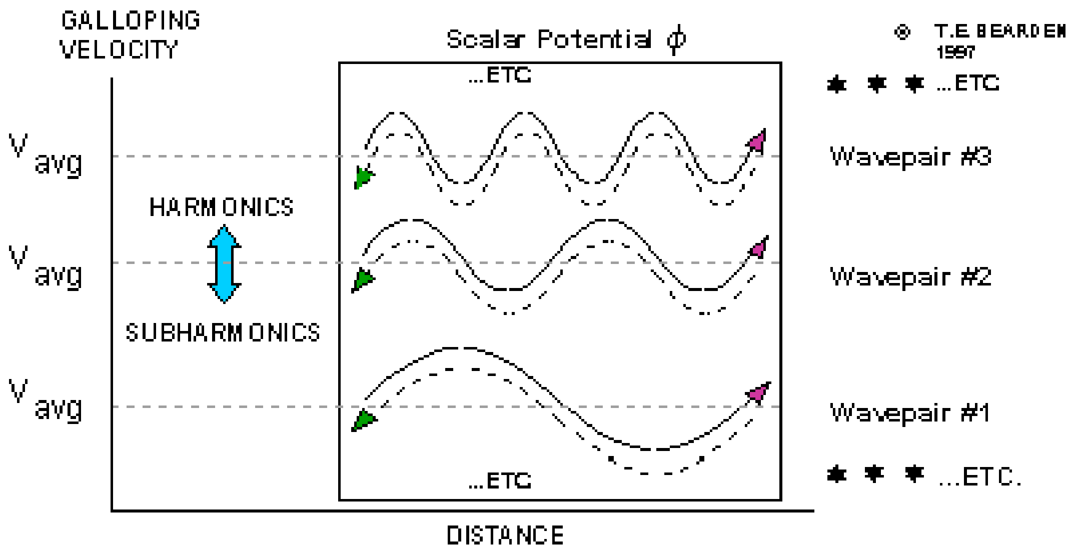
GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC
✱ ✱ ✱ ...ETC
V avg
Wavepair #3
HARMONICS
V avg
Wavepair #2
SUBHARMONICS
V avg
Wavepair #1
✱ ✱ ✱ ...ETC.
...ETC
DISTANCE

The Structure Is:

A harmonic set of longitudinal wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. Each wavepair is a time-density wave and therefore an electrogravitational wave.

Figure 4. Infolded longitudinal biwaves (time-density EM waves) composition of a scalar potential.

Published by Lotus® Freelance Graphics®

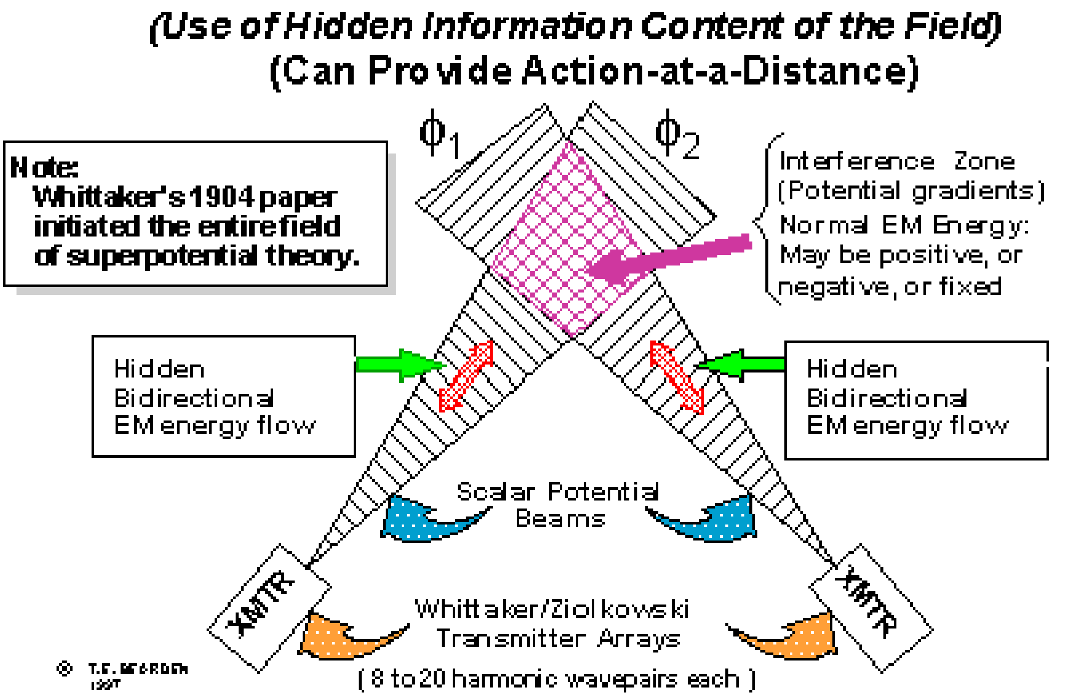

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note: Whittaker's 1904 paper initiated the entire field of superpotential theory.

ϕ_1
ϕ_2
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T.E. BEARDEN 1997

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepairfunctions) produces all EM force fields and waves.

Published by Lotus® Freelance Graphics®

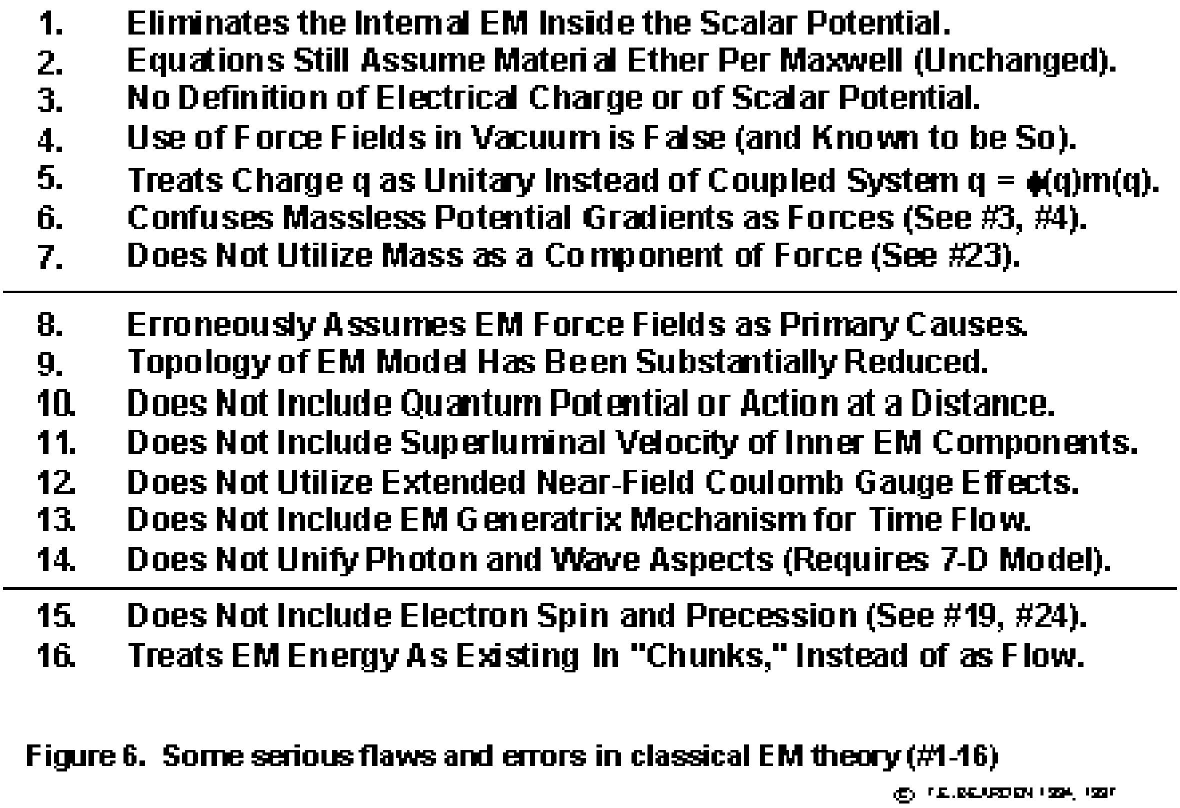

1. Eliminates the Internal EM Inside the Scalar Potential.
2. Equations Still Assume Material Ether Per Maxwell (Unchanged).
3. No Definition of Electrical Charge or of Scalar Potential.
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi(q)m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.

Figure 6. Some serious flaws and errors in classical EM theory (#1-16)

© T.E. BEARDEN 1994, 1997

Published by Lotus® Freelance Graphics®

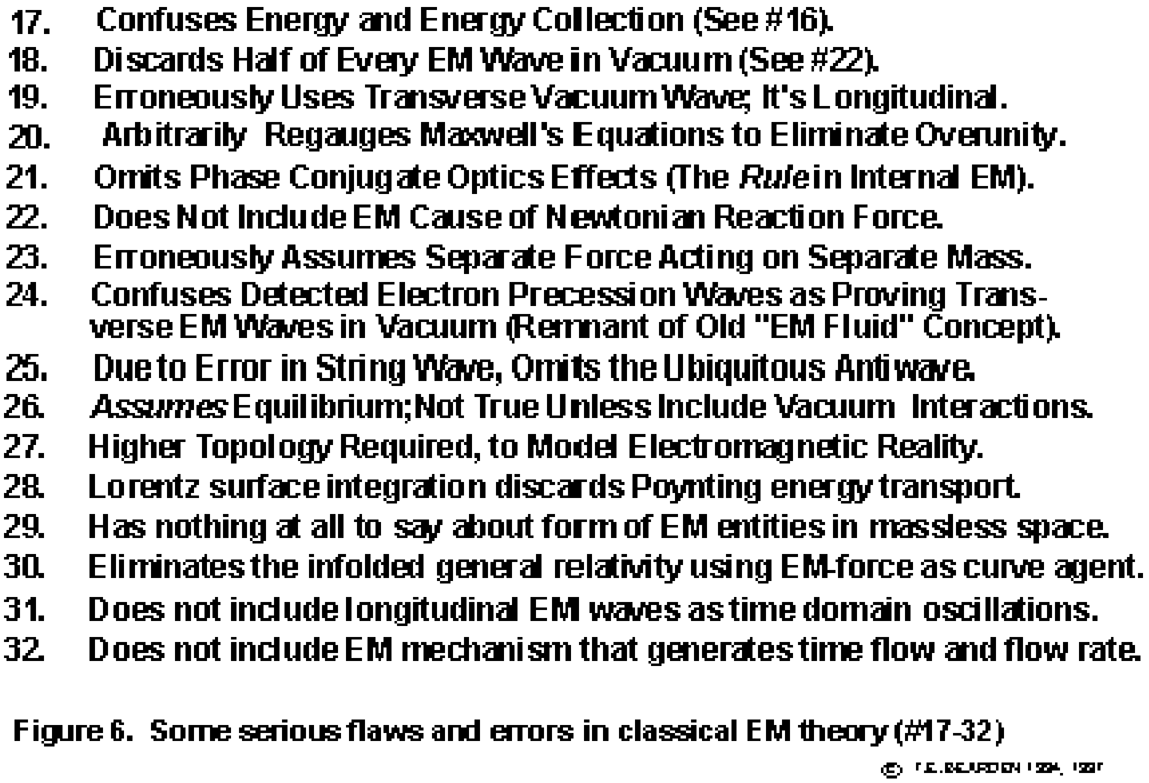

17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

Figure 6. Some serious flaws and errors in classical EM theory (#17-32)

© T.E. BEARDEN 1994, 1997

Published by Lotus® Freelance Graphics®

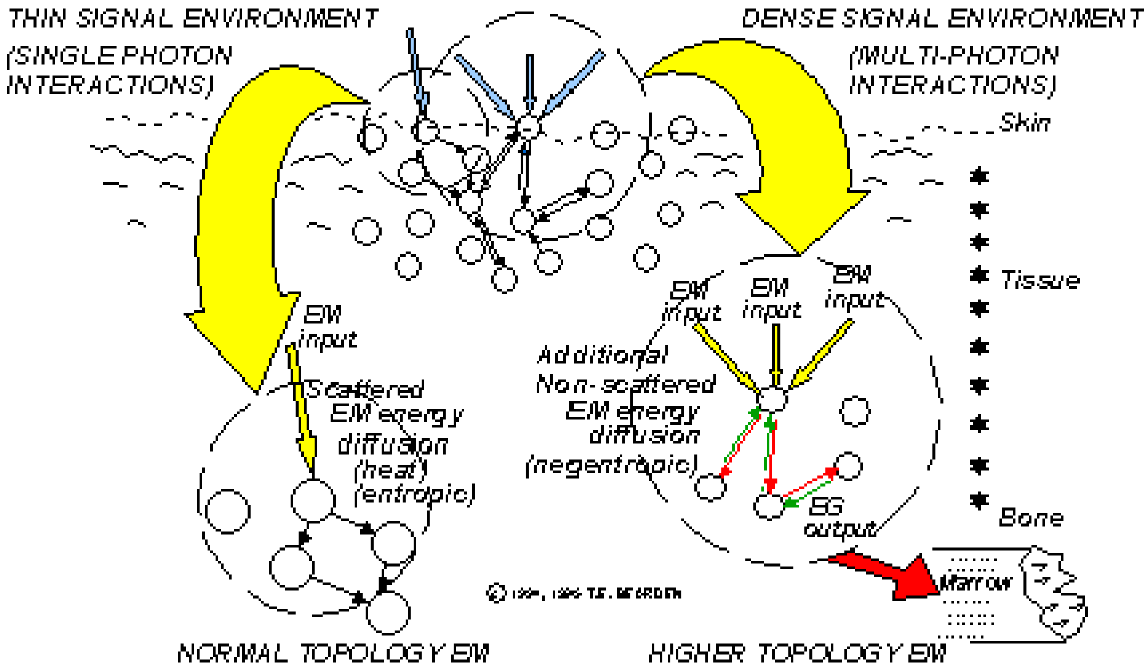
THIN SIGNAL ENVIRONMENT
(SINGLE PHOTON INTERACTIONS)
DENSE SIGNAL ENVIRONMENT
(MULTI-PHOTON INTERACTIONS)
Skin
Tissue
Bone
Marrow
EM input
Scattered EM energy diffusion (heat) (entropic)
EM input
EM input
EM input
Additional Non-scattered EM energy diffusion (negentropic)
EG output
© 1994, 1996 T.E. BEARDEN
NORMAL TOPOLOGY EM
HIGHER TOPOLOGY EM

In a dense signal environment, phase conjugation effects become significant, generating PCR signals which couple to and backtrack the cells' and tissues' own signals, thus penetrating deeply throughout the body with weak partial cancellation jamming.

Figure 30. Abnormal penetration depth of incident weak signals due to multiphoton interaction in a dense weak signal environment.

Published by Lotus® Freelance Graphics®

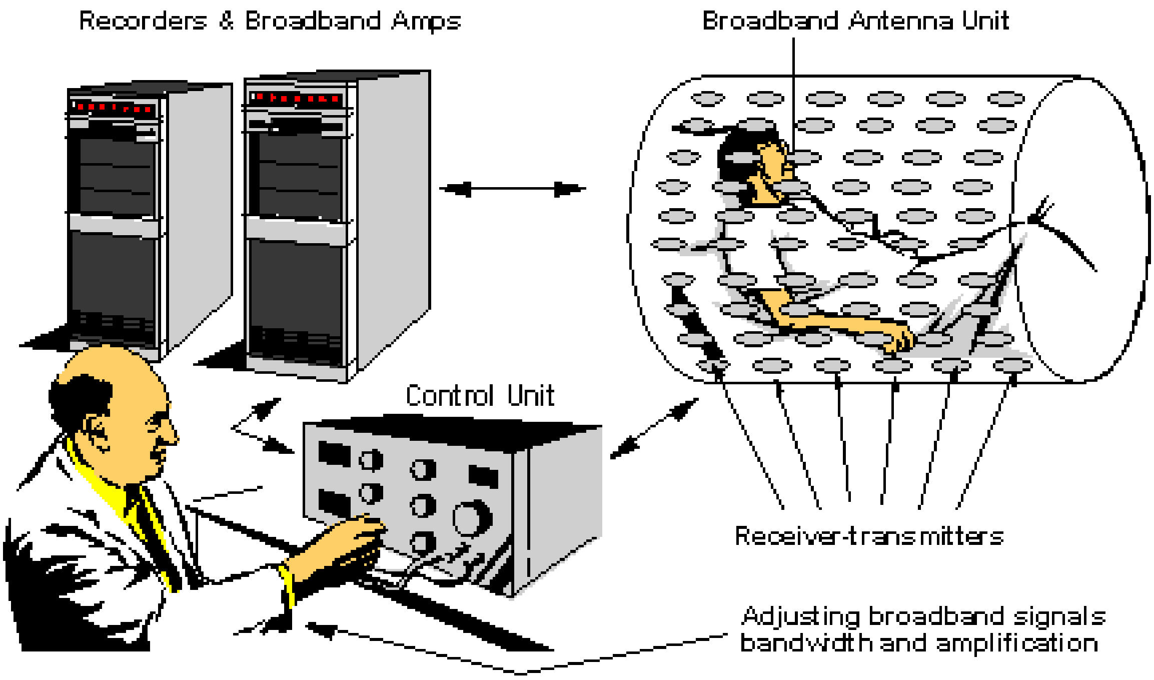
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
Adjusting broadband signals bandwidth and amplification

Figure 31. Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage.

© T.E. BEARDEN 1996

Published by Lotus® Freelance Graphics®

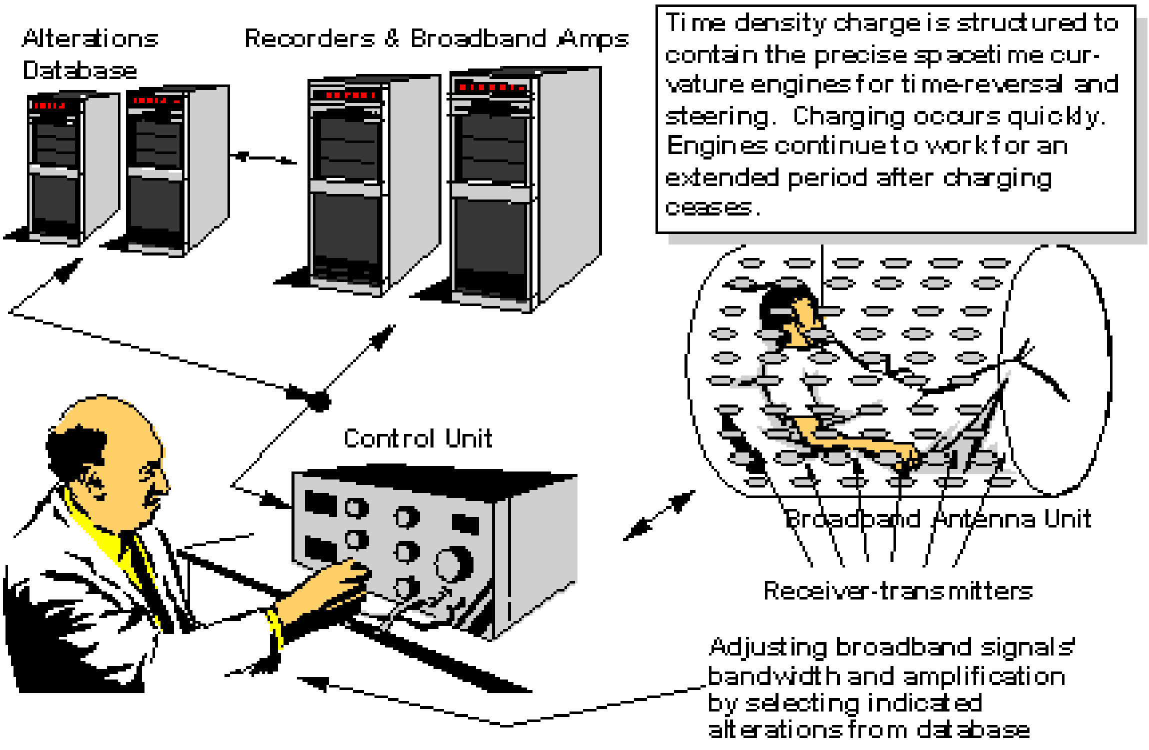
Alterations Database
Recorders & Broadband Amps
Time density charge is structured to contain the precise spacetime curvature engines for time-reversal and steering. Charging occurs quickly. Engines continue to work for an extended period after charging ceases.
Control Unit
Broadband Antenna Unit
Receiver-transmitters
Adjusting broadband signals' bandwidth and amplification by selecting indicated alterations from database

Figure 9. Using an amplified replica of the body's own dense, broadband signal complex containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal, while steering the time-reversal trajectory to also correct a congenital condition.

© T.E. BEARDEN [illegible]

Published by Lotus® Freelance Graphics®

San Francisco
R-2508
China Lake
Point Mugu
Sea Range
Ft. Irwin
Edwards AFB
Los Angeles
San Diego
Santa Fe
Albuquerque
Missile Range
White Sands
Las Cruces

The Anomalies in Navy Electrolyte Experiments at China Lake

These novel experimental anomalies, like those of Cold Fusion, were caused by using time as an energy source by transducing time-energy into spatial energy. Transducing one microsecond per second of time into spatial EM energy, yields nearly 10^{11} watts of power

- Melvin H. Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", *Proc. ICCF-7 (International Conference on Cold Fusion—7)*, Vancouver, BC, Canada, Apr. 1998, p. 101.

- Consistency of China Lake Anomalies
- Anomalous LW EM Emission Phenomena at China Lake (1)
- Anomalous LW EM Emission Phenomena at China Lake (2)
- Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)
- Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)
- Consistency of Anomalies
- Consistency of Anomalies (2)

- Experimental LW Emission Phenomena in Electrolysis

Next Slide Return to Index Previous Slide

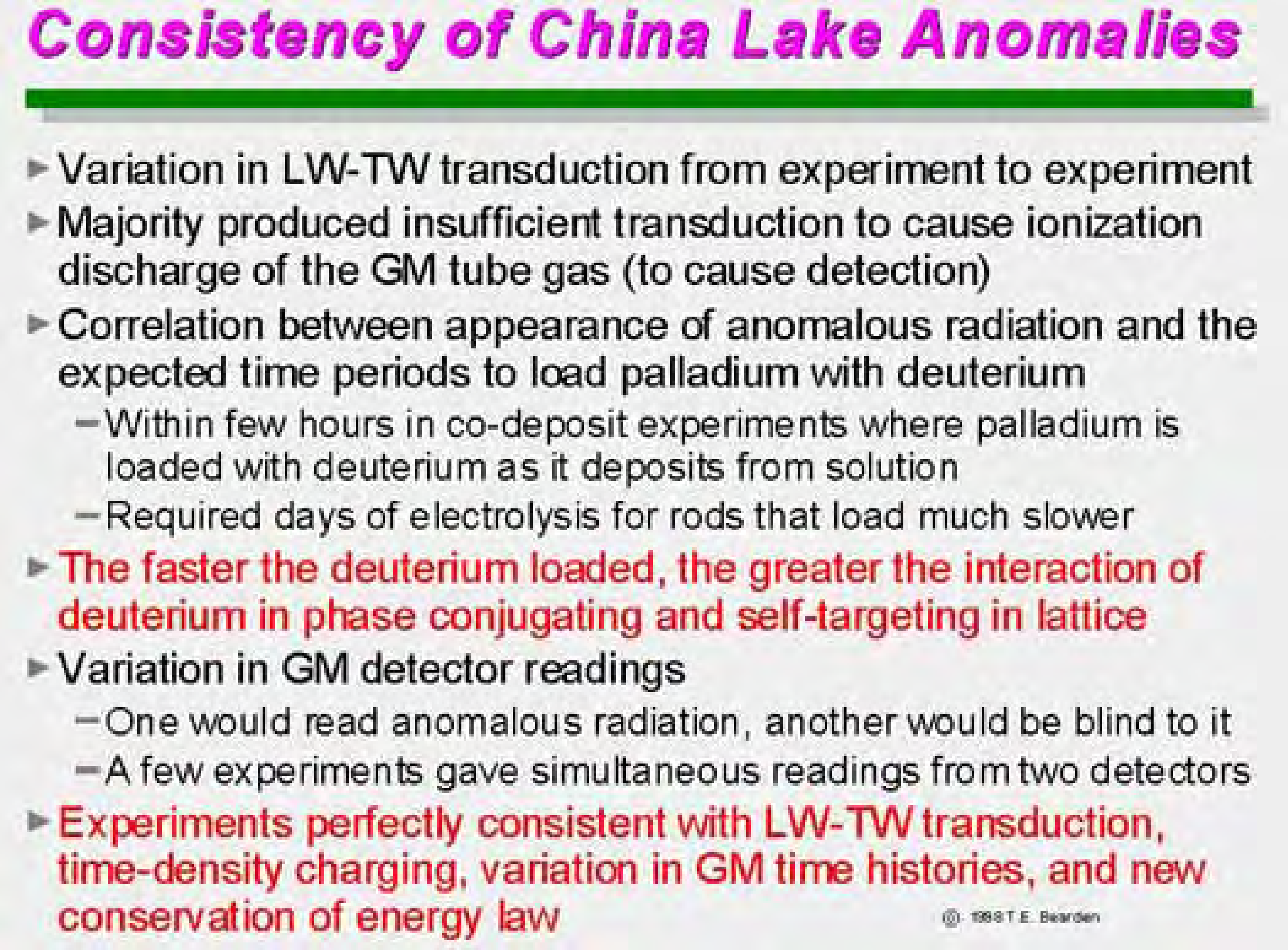

Consistency of China Lake Anomalies

- Variation in LW-TW transduction from experiment to experiment
- Majority produced insufficient transduction to cause ionization discharge of the GM tube gas (to cause detection)
- Correlation between appearance of anomalous radiation and the expected time periods to load palladium with deuterium
  - Within few hours in co-deposit experiments where palladium is loaded with deuterium as it deposits from solution
  - Required days of electrolysis for rods that load much slower
- The faster the deuterium loaded, the greater the interaction of deuterium in phase conjugating and self-targeting in lattice
- Variation in GM detector readings
  - One would read anomalous radiation, another would be blind to it
  - A few experiments gave simultaneous readings from two detectors
- Experiments perfectly consistent with LW-TW transduction, time-density charging, variation in GM time histories, and new conservation of energy law

© 1998 T.E. Bearden

Next Slide

Return to Index

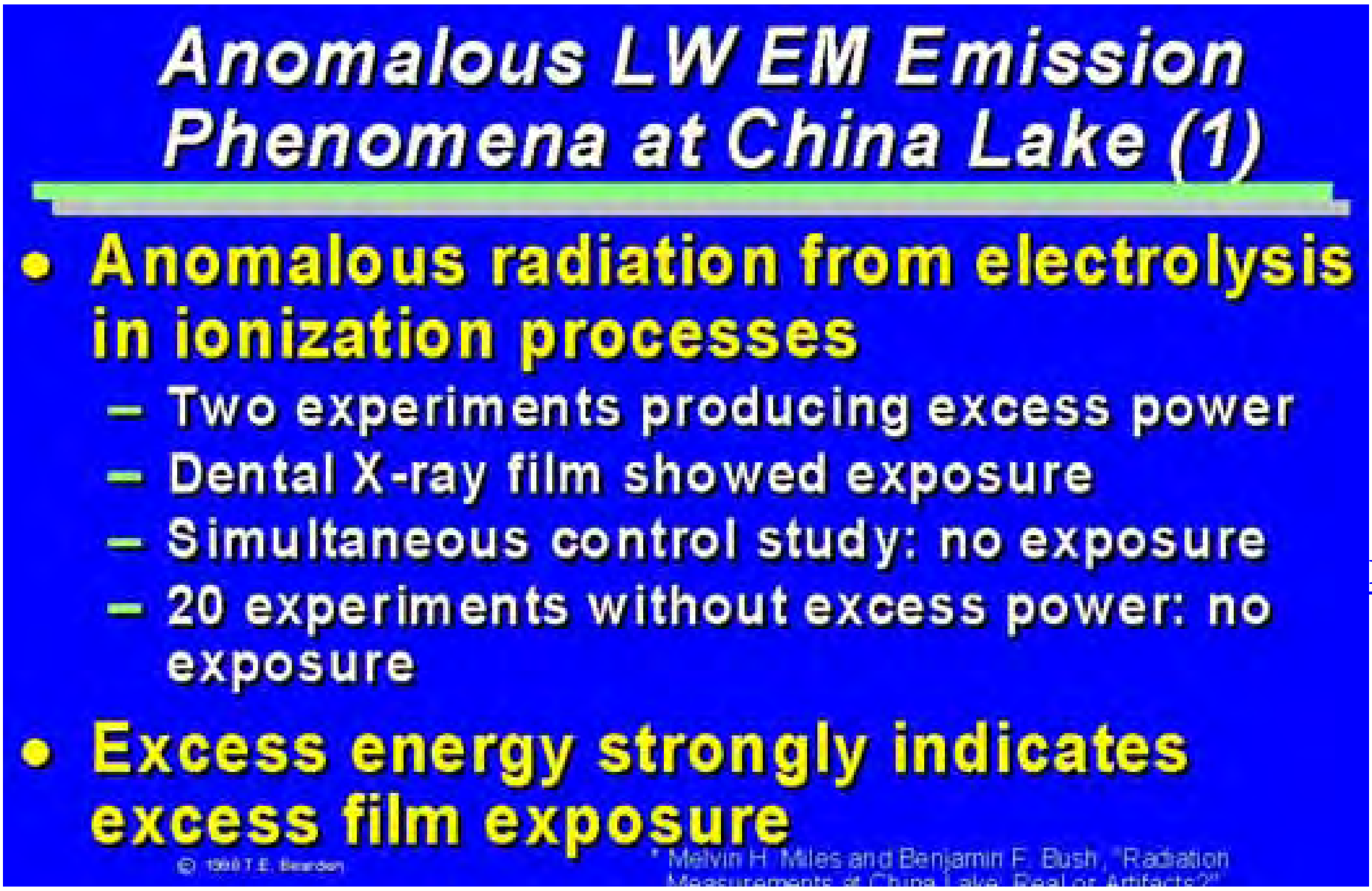

Anomalous LW EM Emission Phenomena at China Lake (1)

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental X-ray film showed exposure
  - Simultaneous control study: no exposure
  - 20 experiments without excess power: no exposure
- Excess energy strongly indicates excess film exposure

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?"

Next Slide

Return to Index

Previous Slide

Next Slide Return to Index Previous Slide

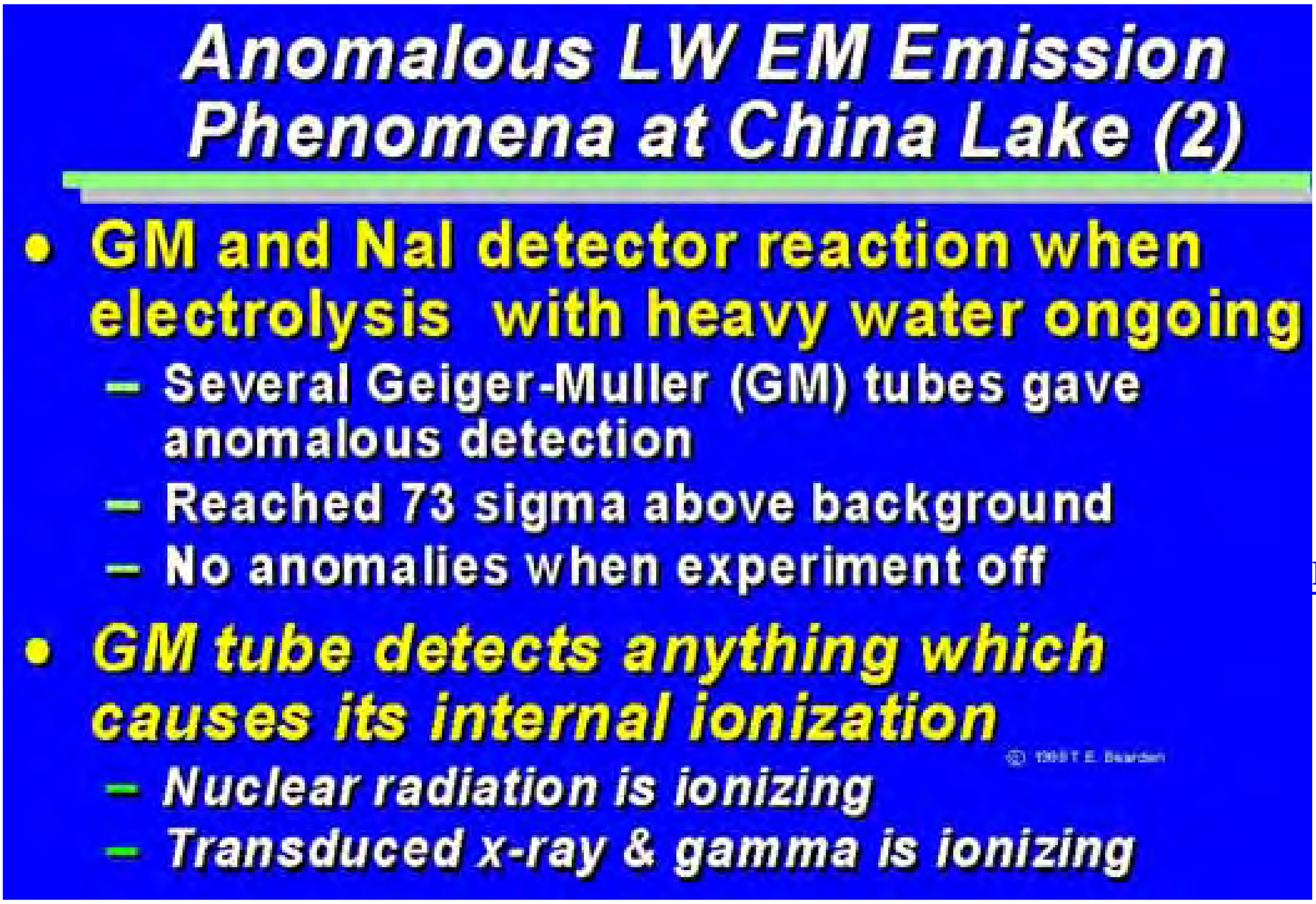

Anomalous LW EM Emission Phenomena at China Lake (2)

- GM and NaI detector reaction when electrolysis with heavy water ongoing
  - Several Geiger-Muller (GM) tubes gave anomalous detection
  - Reached 73 sigma above background
  - No anomalies when experiment off
- *GM tube detects anything which causes its internal ionization*
  - *Nuclear radiation is ionizing*
  - *Transduced x-ray & gamma is ionizing*

Next Slide

Return to Index

Previous Slide

Next Slide Return to Index Previous Slide

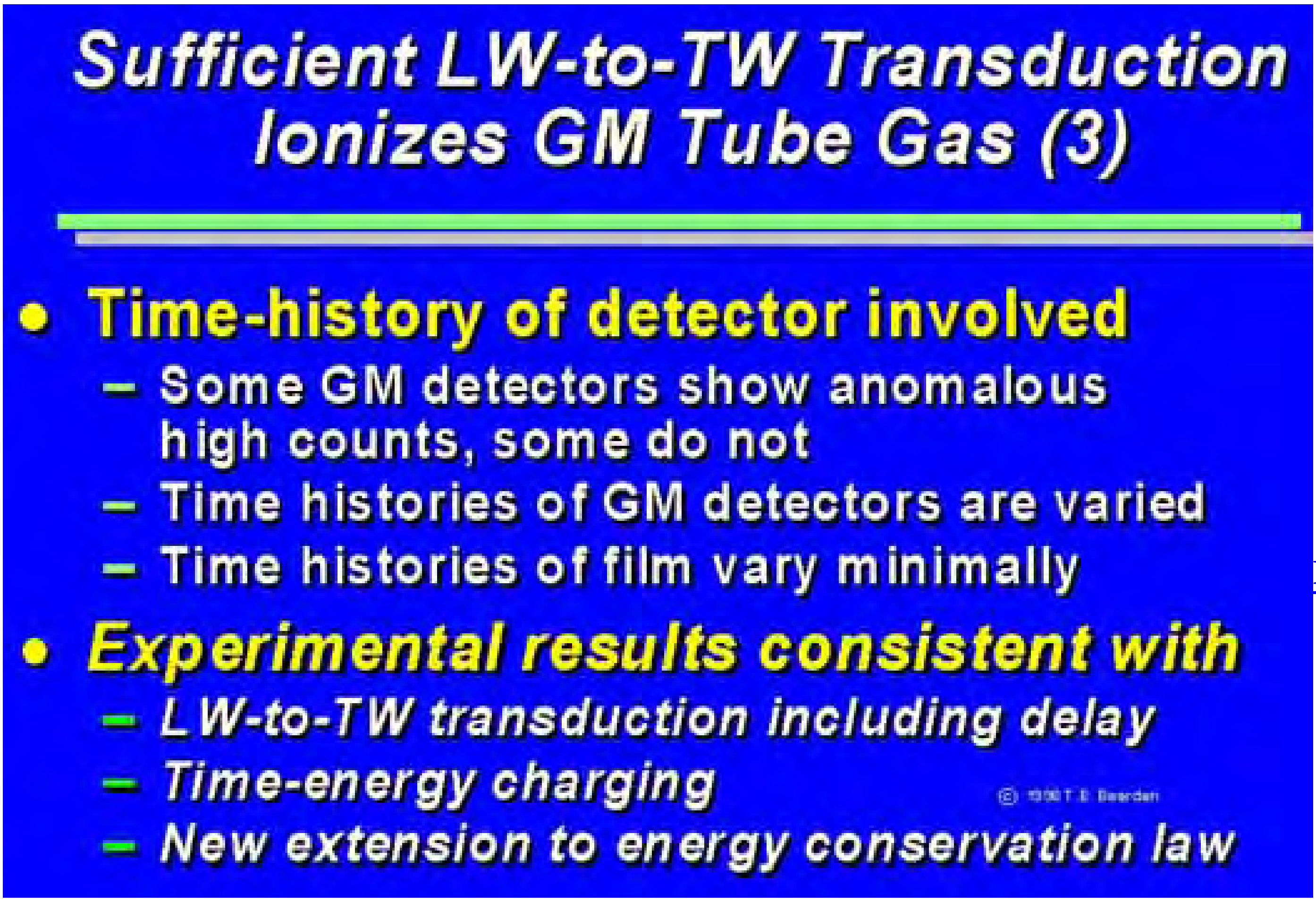

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (3)

- Time-history of detector involved
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- *Experimental results consistent with*
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© [illegible] T. E. Bearden

Next Slide

Return to Index

Previous Slide

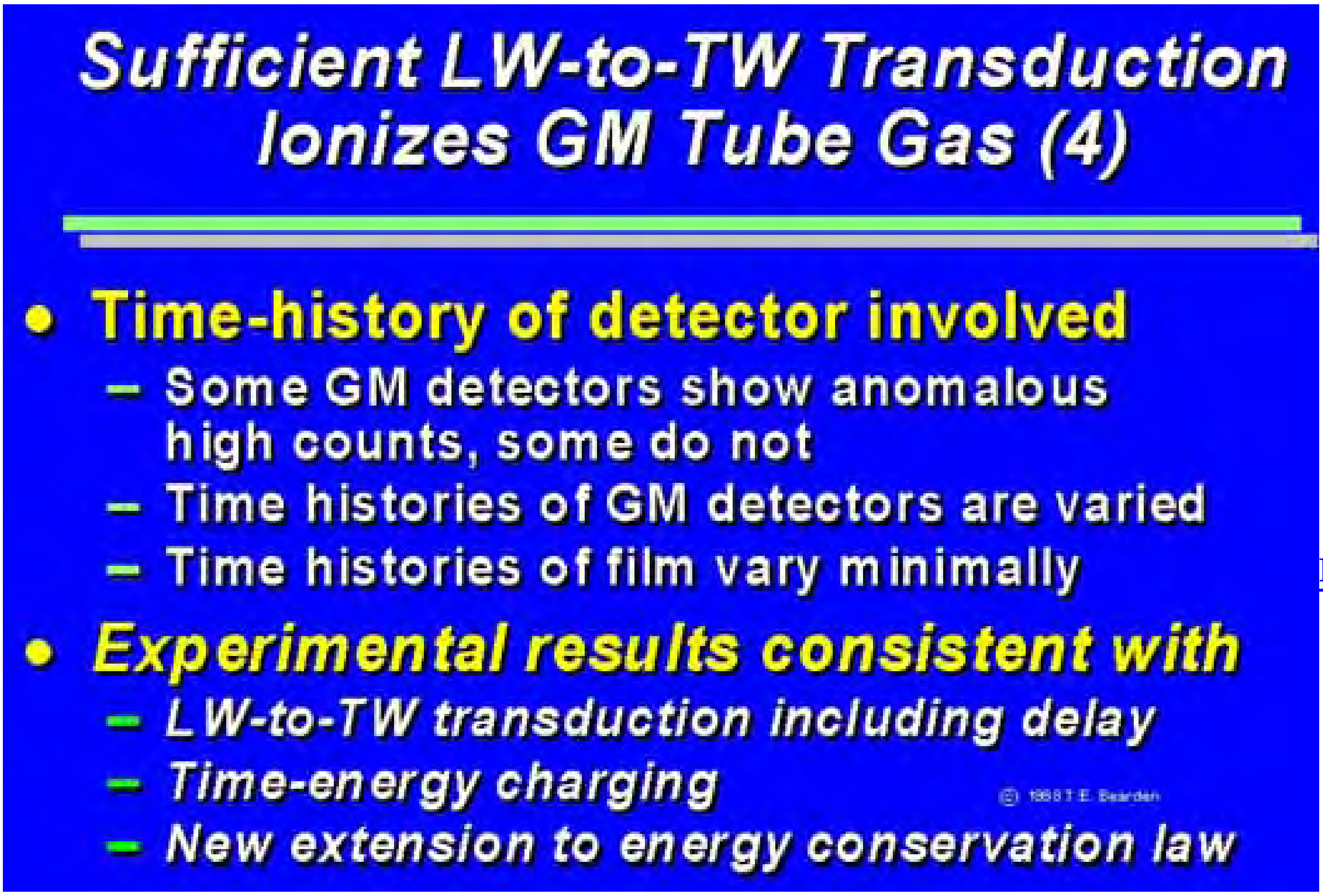

Sufficient LW-to-TW Transduction Ionizes GM Tube Gas (4)

- **Time-history of detector involved**
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors are varied
  - Time histories of film vary minimally
- ***Experimental results consistent with***
  - *LW-to-TW transduction including delay*
  - *Time-energy charging*
  - *New extension to energy conservation law*

© 1988 T. E. Bearden

Next Slide

Return to Index

Previous Slide

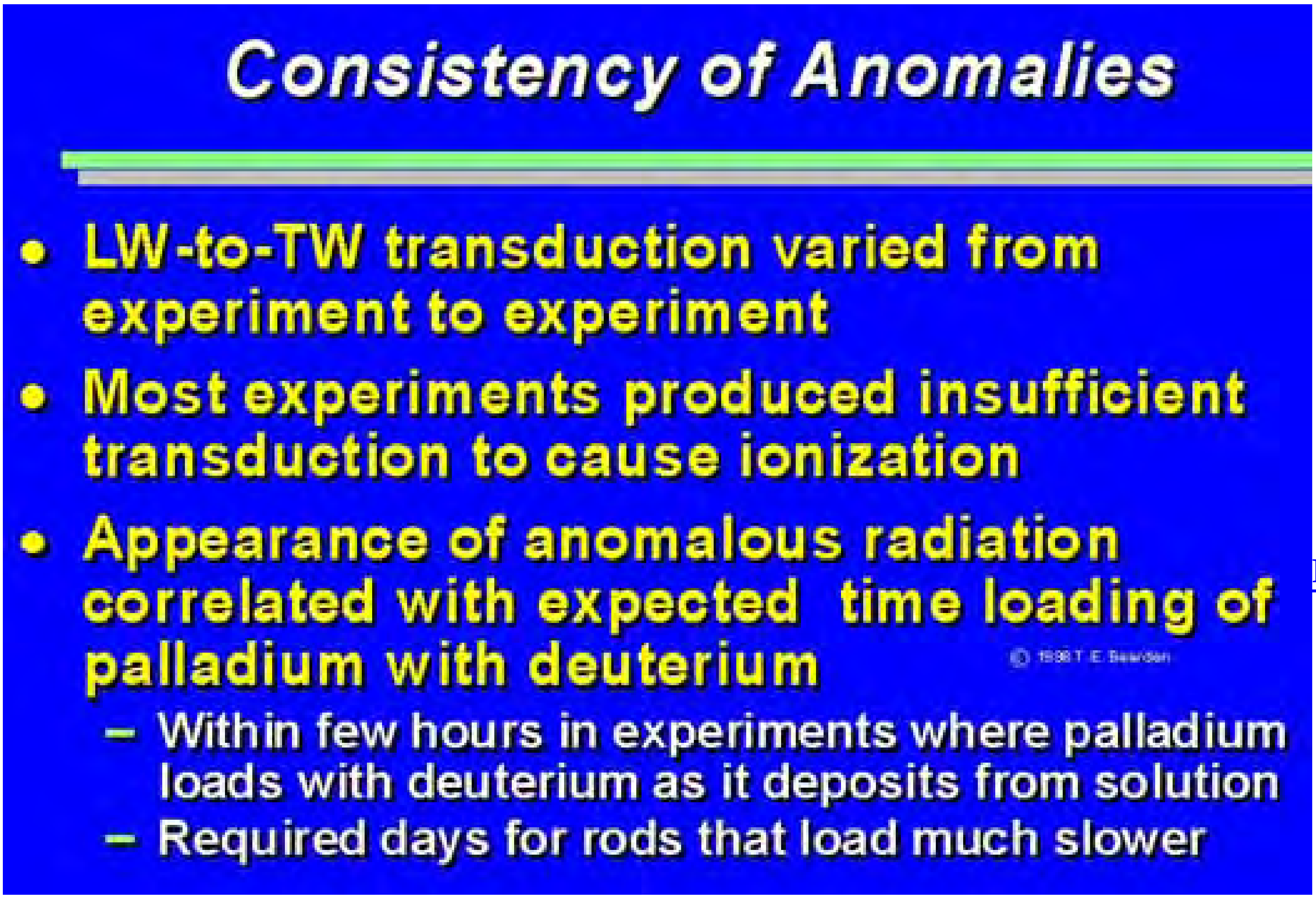

Consistency of Anomalies

- LW-to-TW transduction varied from experiment to experiment
- Most experiments produced insufficient transduction to cause ionization
- Appearance of anomalous radiation correlated with expected time loading of palladium with deuterium
  - Within few hours in experiments where palladium loads with deuterium as it deposits from solution
  - Required days for rods that load much slower

Next Slide

Return to Index

Previous Slide

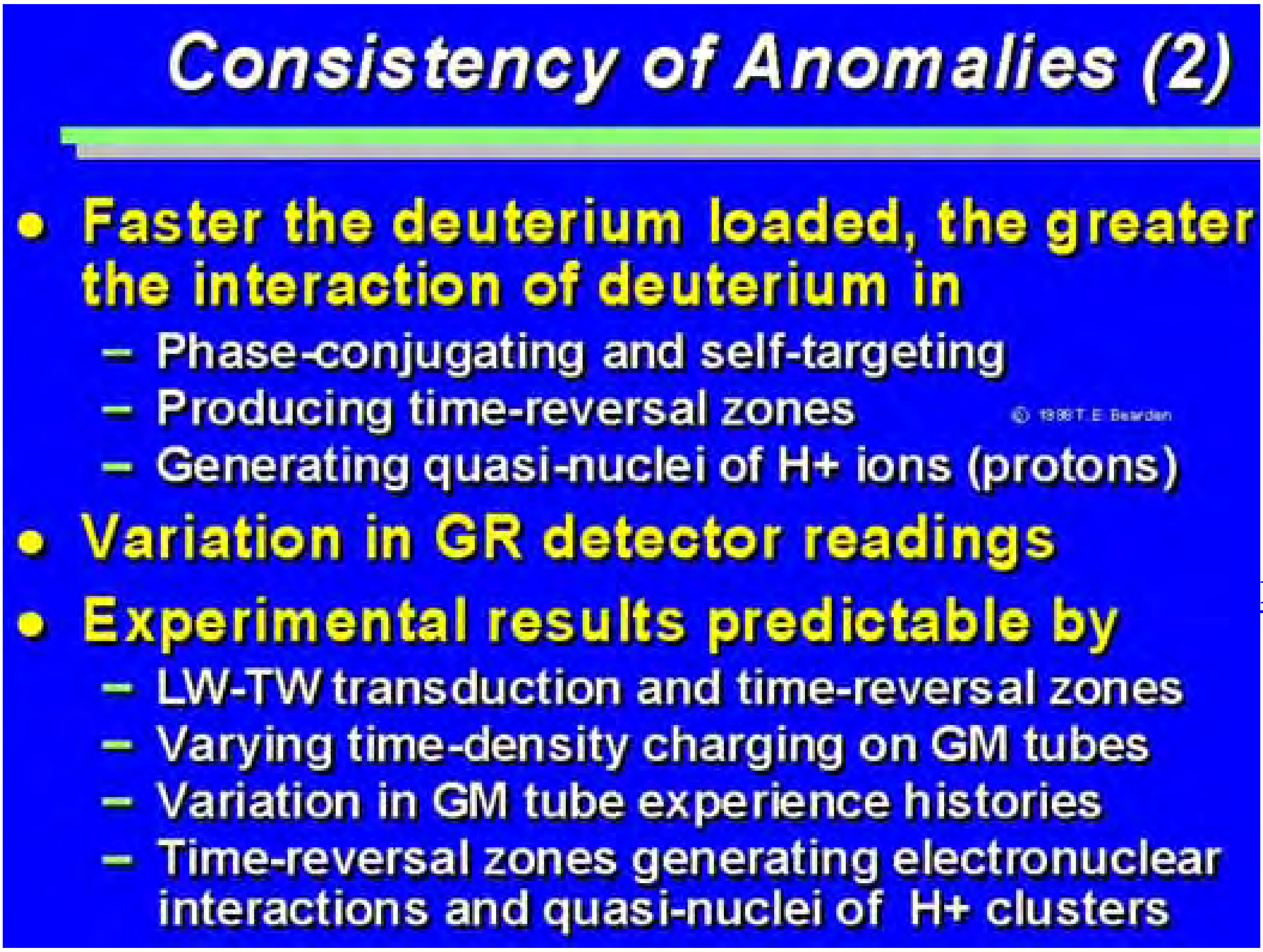

Consistency of Anomalies (2)

- Faster the deuterium loaded, the greater the interaction of deuterium in
  - Phase-conjugating and self-targeting
  - Producing time-reversal zones
  - Generating quasi-nuclei of H+ ions (protons)
- Variation in GR detector readings
- Experimental results predictable by
  - LW-TW transduction and time-reversal zones
  - Varying time-density charging on GM tubes
  - Variation in GM tube experience histories
  - Time-reversal zones generating electronuclear interactions and quasi-nuclei of H+ clusters

© 1998 T. E. Bearden

Next Slide

Return to Index

Previous Slide

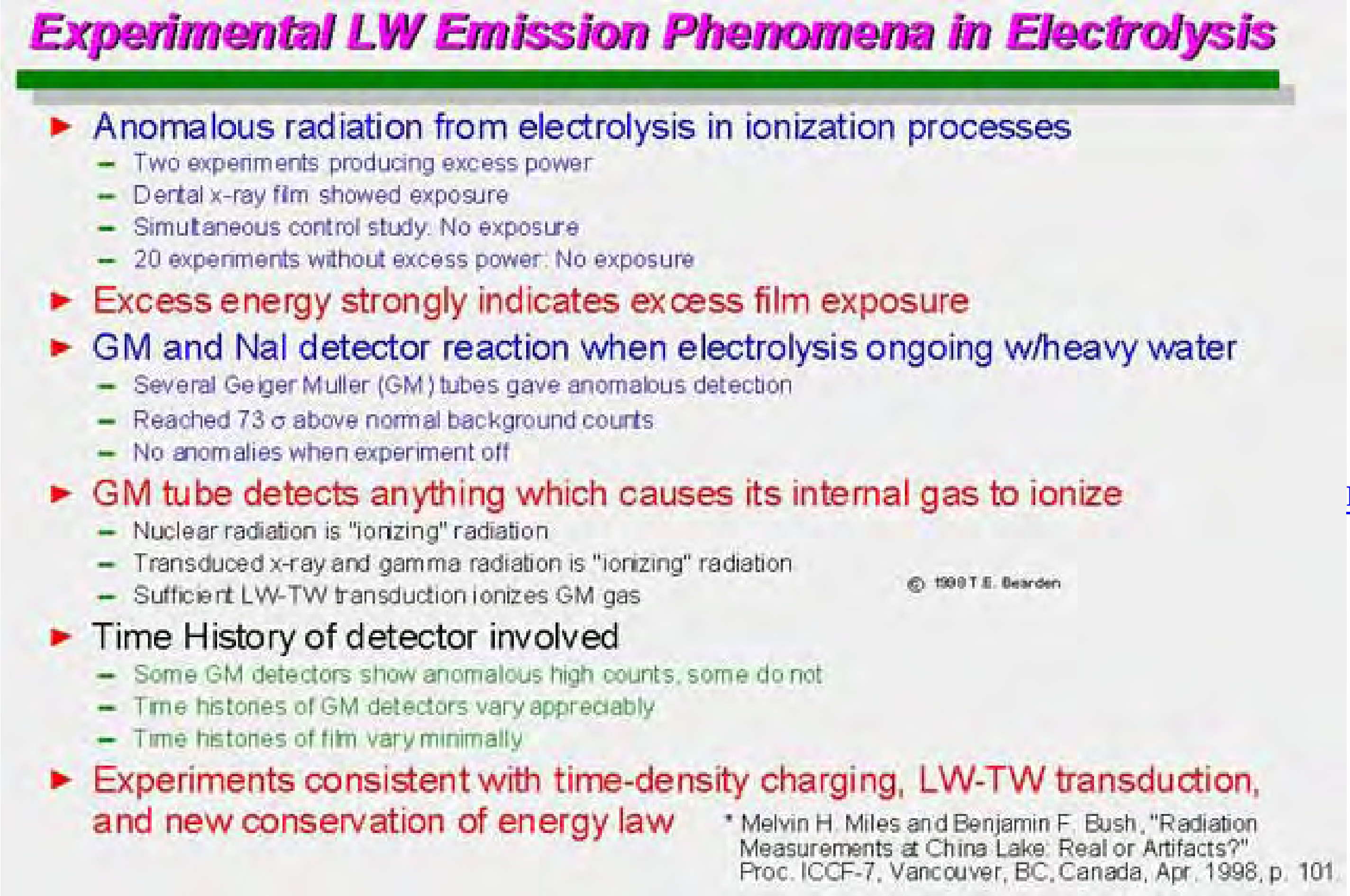

Experimental LW Emission Phenomena in Electrolysis

- Anomalous radiation from electrolysis in ionization processes
  - Two experiments producing excess power
  - Dental x-ray film showed exposure
  - Simultaneous control study: No exposure
  - 20 experiments without excess power: No exposure
- Excess energy strongly indicates excess film exposure
- GM and NaI detector reaction when electrolysis ongoing w/heavy water
  - Several Geiger Muller (GM) tubes gave anomalous detection
  - Reached 73 σ above normal background counts
  - No anomalies when experiment off
- GM tube detects anything which causes its internal gas to ionize
  - Nuclear radiation is "ionizing" radiation
  - Transduced x-ray and gamma radiation is "ionizing" radiation
  - Sufficient LW-TW transduction ionizes GM gas
- Time History of detector involved
  - Some GM detectors show anomalous high counts, some do not
  - Time histories of GM detectors vary appreciably
  - Time histories of film vary minimally
- Experiments consistent with time-density charging, LW-TW transduction, and new conservation of energy law

© 1998 T.E. Bearden

* Melvin H. Miles and Benjamin F. Bush, "Radiation Measurements at China Lake: Real or Artifacts?" Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 101

Return to Index

Previous Slide

A Visual Tour of what they don't want you to know about electrical circuits

10 TRILLION PERCENT OF THE CURRENT PRODUCED IS WASTED!

---

1. The scalar potential is a harmonic set of phase conjugate longitudinal EM waves.
2. The dipole is a true negative resistor.
3. The dipole's broken 3-symmetry initiates a spreading giant negentropic reordering of a fraction of the vacuum's energy.
4. Energy flow contours surrounding a transmission line.
5. Heaviside and Poynting energy flow components. The Heaviside component is often **10 trillion times** the Poynting component, but **is simply wasted** in ordinary single-pass energy circuits.
6. Negative resistance process v. positive resistance process.
7. Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component.

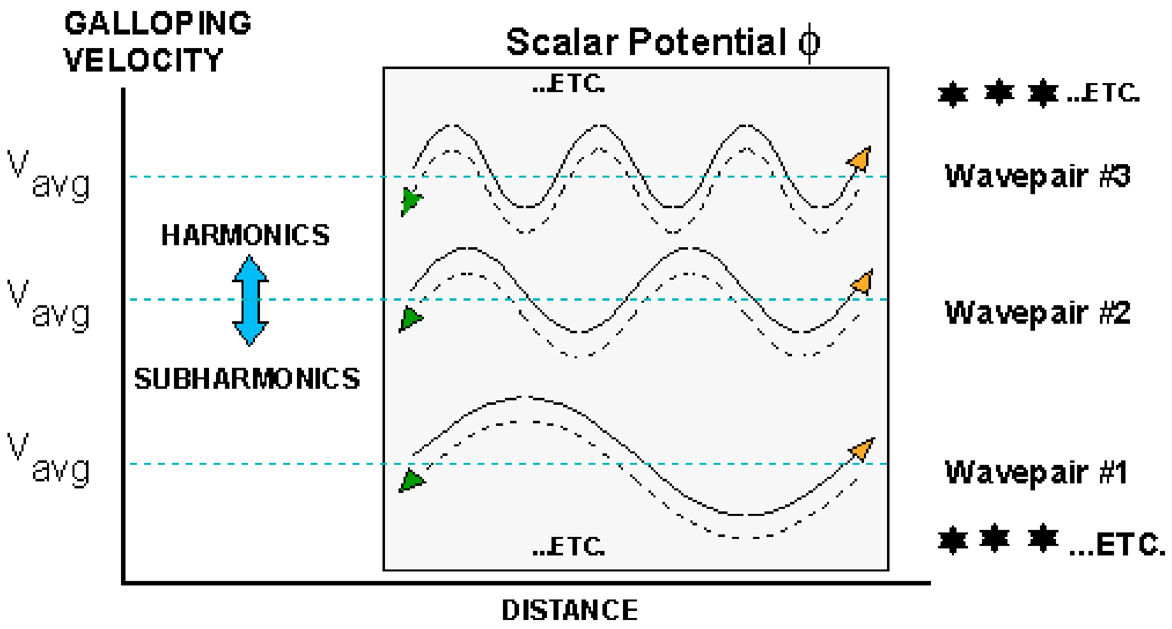

[Next Slide]
[Return to Index]

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.
- Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 1. The scalar potential is a harmonic set of phase conjugate longitudinal EM waves.

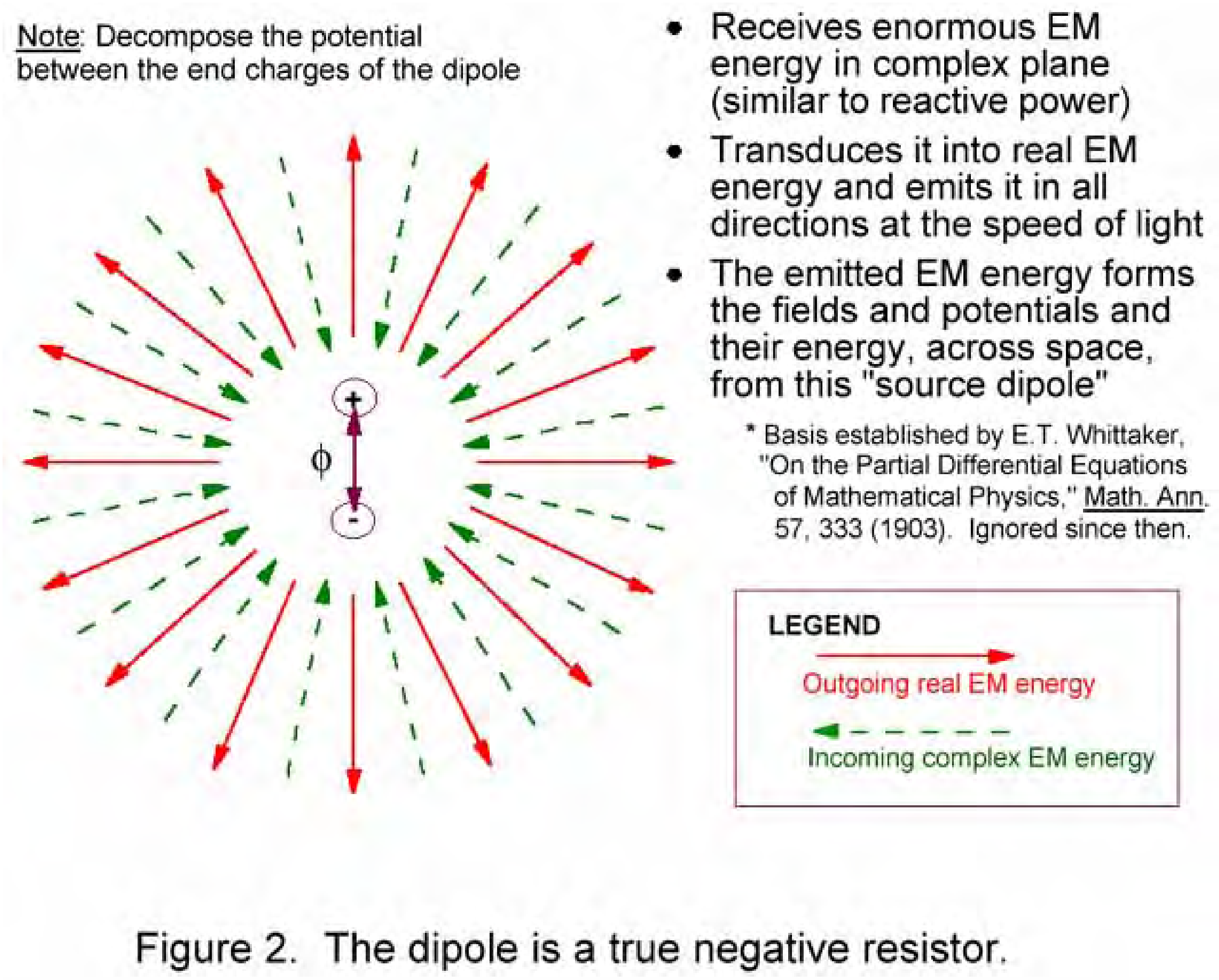

Note: Decompose the potential between the end charges of the dipole

- Receives enormous EM energy in complex plane (similar to reactive power)
- Transduces it into real EM energy and emits it in all directions at the speed of light
- The emitted EM energy forms the fields and potentials and their energy, across space, from this "source dipole"

* Basis established by E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Math. Ann. 57, 333 (1903). Ignored since then.

LEGEND

Outgoing real EM energy

Incoming complex EM energy

Figure 2. The dipole is a true negative resistor.

[Next Slide]

[Return to Index]

[Previous Slide]

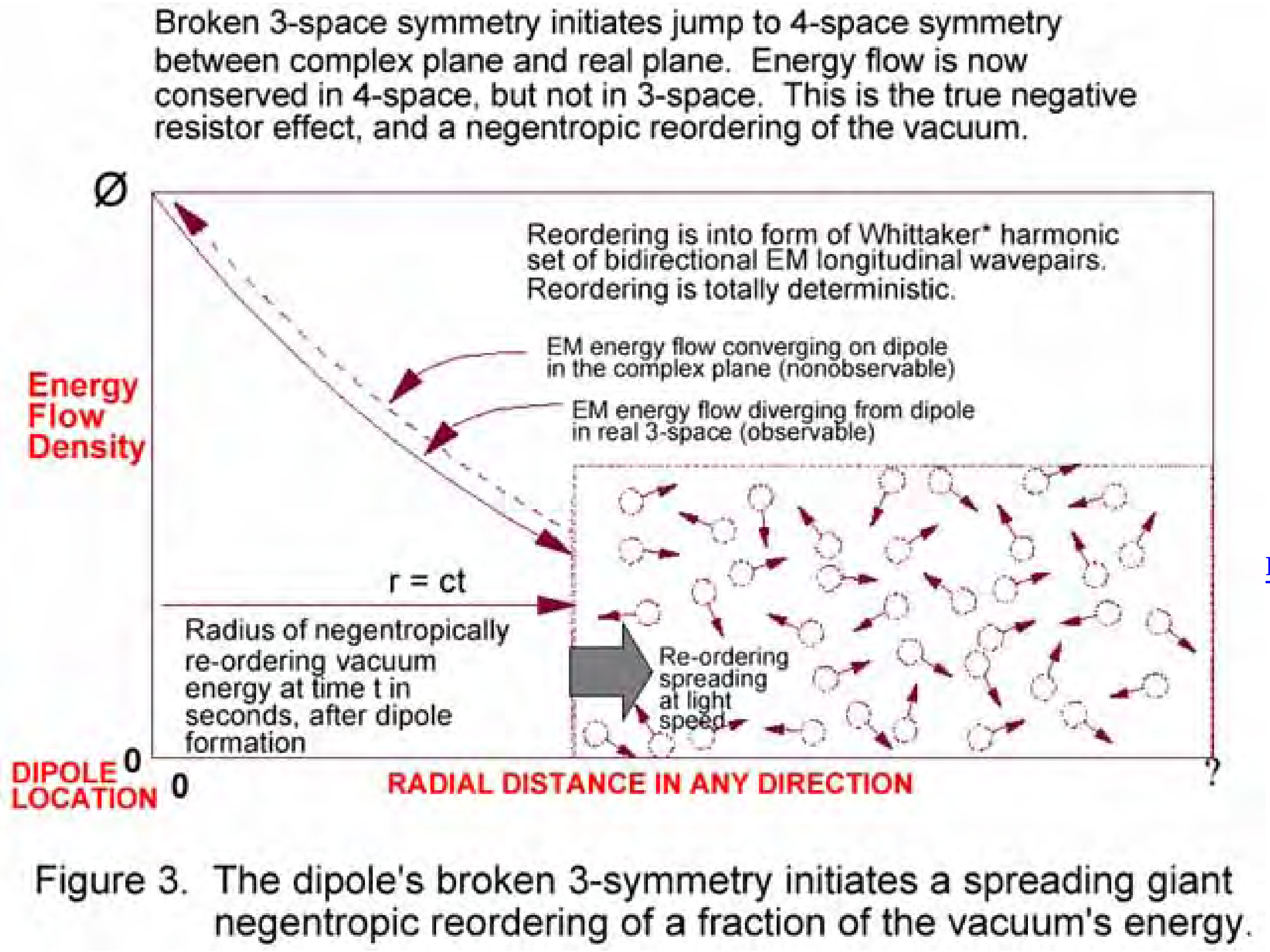

Broken 3-space symmetry initiates jump to 4-space symmetry between complex plane and real plane. Energy flow is now conserved in 4-space, but not in 3-space. This is the true negative resistor effect, and a negentropic reordering of the vacuum.

Reordering is into form of Whittaker* harmonic set of bidirectional EM longitudinal wavepairs. Reordering is totally deterministic.

EM energy flow converging on dipole in the complex plane (nonobservable)

EM energy flow diverging from dipole in real 3-space (observable)

Radius of negentropically re-ordering vacuum energy at time t in seconds, after dipole formation

Re-ordering spreading at light speed

Figure 3. The dipole's broken 3-symmetry initiates a spreading giant negentropic reordering of a fraction of the vacuum's energy.

Next Slide

Return to Index

Previous Slide

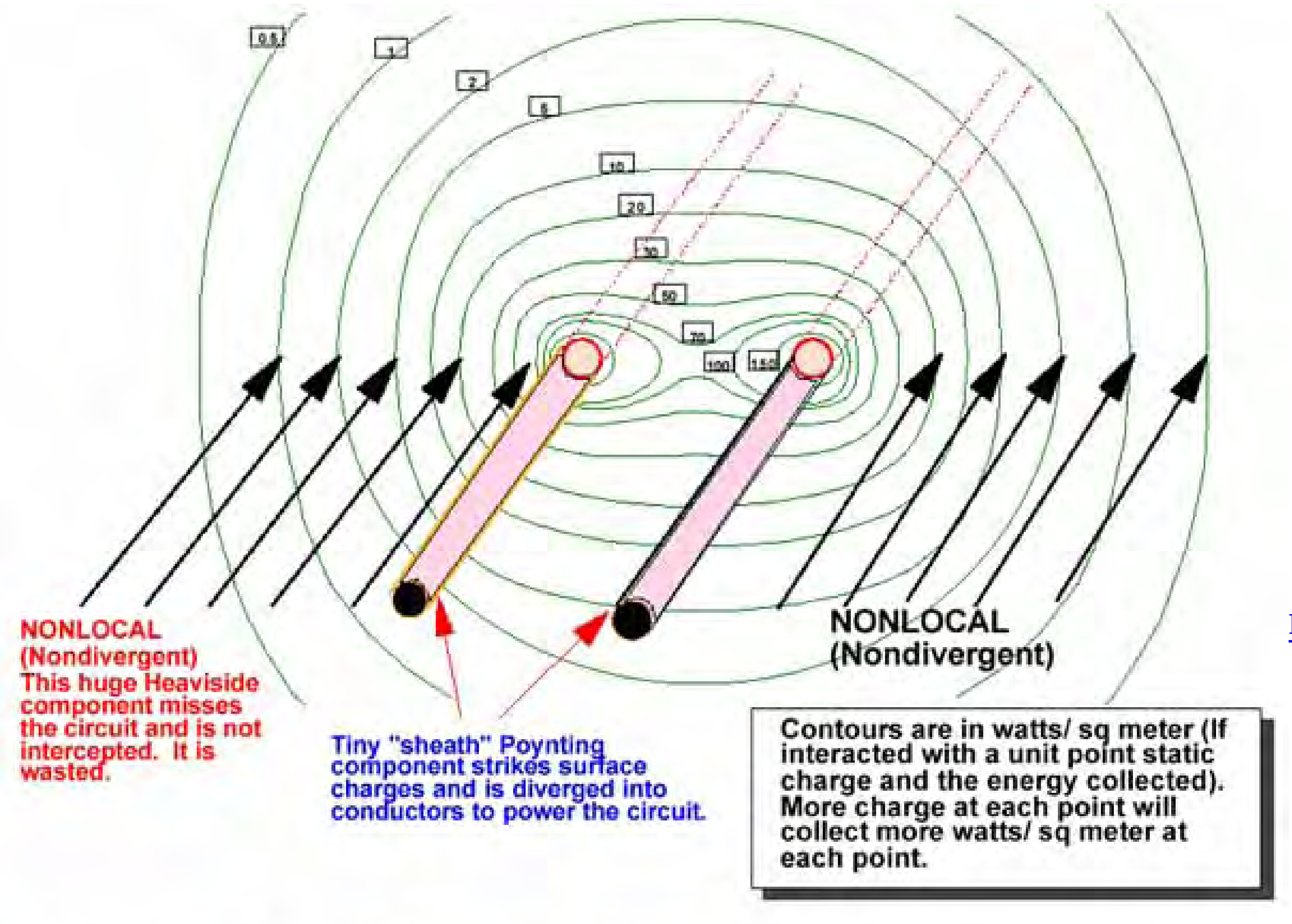

NONLOCAL (Nondivergent) This huge Heaviside component misses the circuit and is not intercepted. It is wasted.

Tiny "sheath" Poynting component strikes surface charges and is diverged into conductors to power the circuit.

NONLOCAL (Nondivergent)

Contours are in watts/ sq meter (If interacted with a unit point static charge and the energy collected). More charge at each point will collect more watts/ sq meter at each point.

Figure 4. Energy flow contours surrounding a transmission line.

Next Slide

Return to Index

Previous Slide

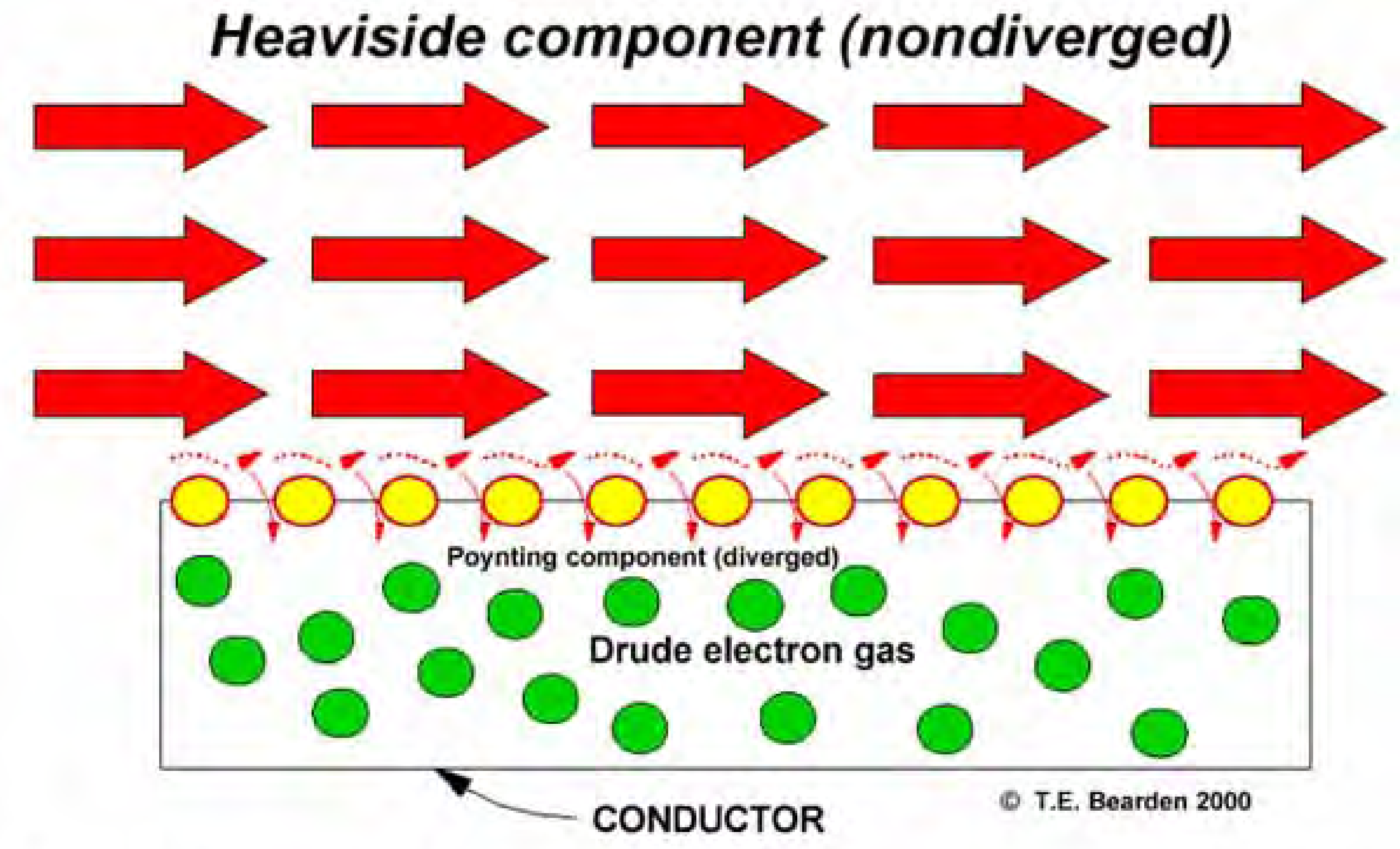
Heaviside component (nondiverged)
Poynting component (diverged)
Drude electron gas
CONDUCTOR
© T.E. Bearden 2000

Figure 5. Heaviside and Poynting energy flow components. The Heaviside component is often 10 trillion times the Poynting component, but is simply wasted in ordinary single-pass energy flow circuits.

Next Slide

Return to Index

Previous Slide

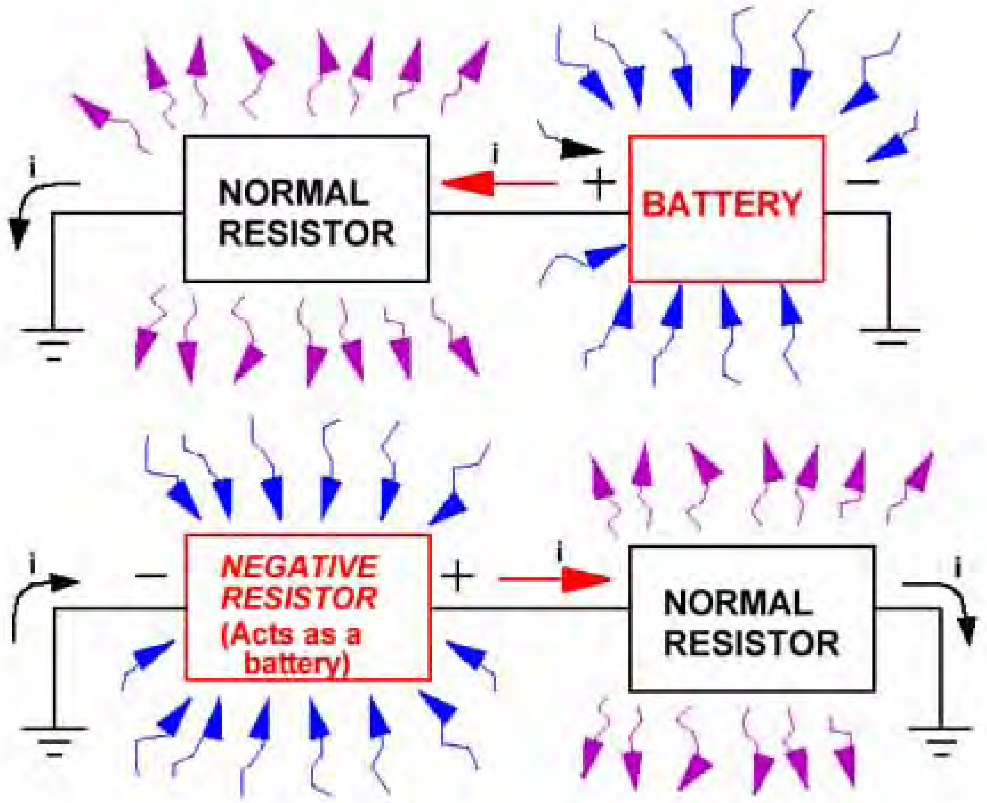
NORMAL RESISTOR
i
i
+
BATTERY
−
NEGATIVE RESISTOR (Acts as a battery)
−
+
i
i
i
NORMAL RESISTOR

Figure 6. Negative resistance process vs. positive resistance process. A negative resistor receives energy in unusable form, transduces it, and outputs it in usable form. A positive resistor receives energy in usable form, and scatters it into unusable form.

Next Slide

Return to Index

Previous Slide

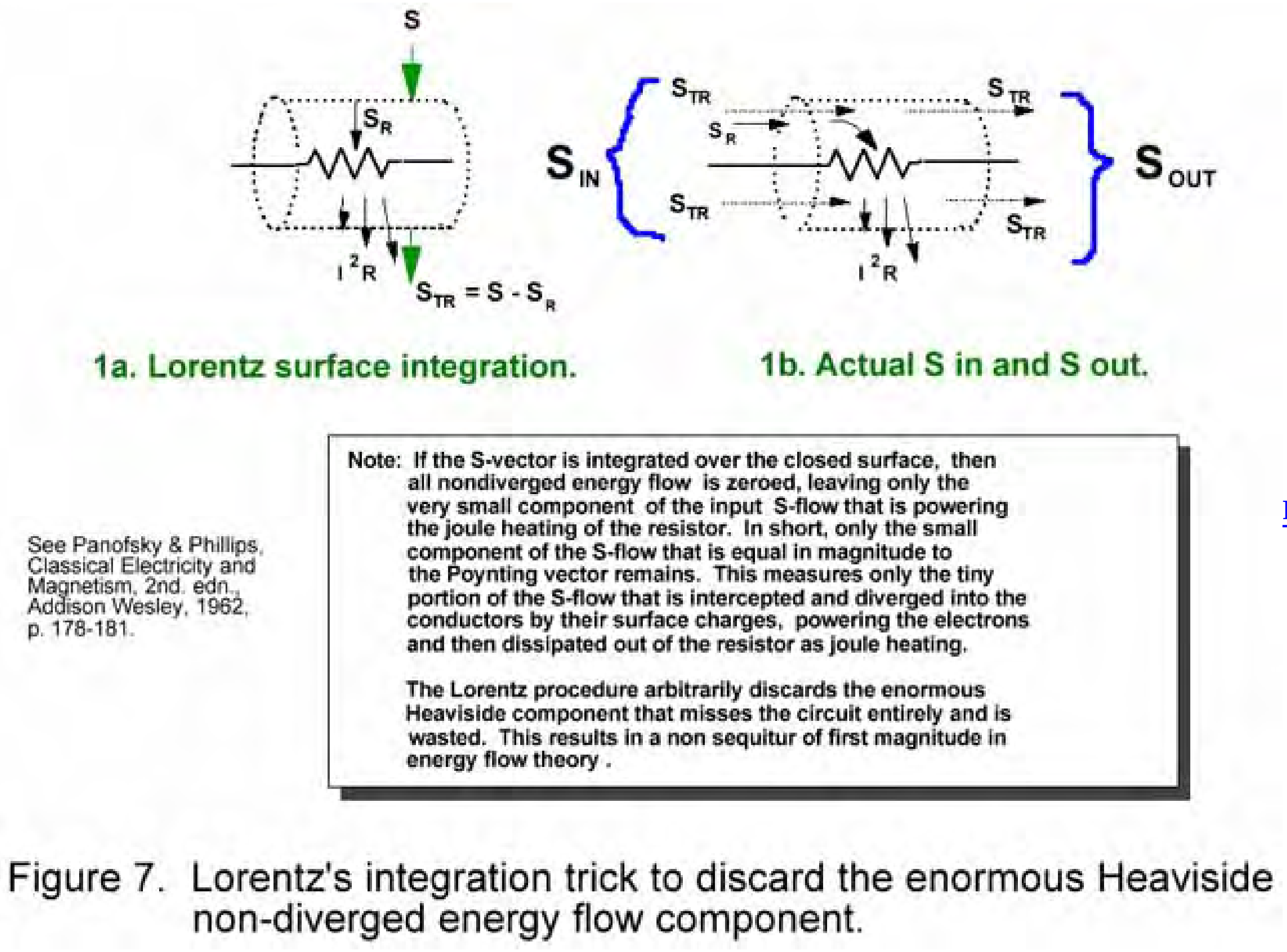

1a. Lorentz surface integration.

1b. Actual S in and S out.

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd. edn., Addison Wesley, 1962, p. 178-181.

Note: If the S-vector is integrated over the closed surface, then all nondiverged energy flow is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Poynting vector remains. This measures only the tiny portion of the S-flow that is intercepted and diverged into the conductors by their surface charges, powering the electrons and then dissipated out of the resistor as joule heating.

The Lorentz procedure arbitrarily discards the enormous Heaviside component that misses the circuit entirely and is wasted. This results in a non sequitur of first magnitude in energy flow theory .

Figure 7. Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component.

Return to Index

Previous Slide

ILLUSTRATION LIST:

Page 1

(Table 1)
[note 1]

Page 2

(Table 2)
[note 2]
[note 3]
[note 4]
[note 5]

Page 3

[note 6]
[note 7]
[note 8]
[note 9]
(Figure 1)
(Figure 2)
(Figure 3)
(Figure 4)
(Figure 5)
(Figure 6)

Page 4

[note 10]
[note 11]
[note 12]
[note 13]
(Figure 7)
(Figure 8)
(Figure 9)
(Figure 10)

Page 5

[note 14]
[note 15]
Figure 11
Figure 12

Page 6

[note 16]
[note 17]
(Figure 13)
(Figure 14)
(Figure 15)

Page 7

[note 13]
[note 18]
[note 19]
[note 20]
(Figure 16)

Page 9

(Table 3)
[note 22]
(Figure 17)
(Figure 18)

Page 10

[note 23]
[note 24]
[note 25]
[note 26]
[note 27]
(Figure 19)
(Figure 20)

Page 12

(Table 4)
[note 28]
[note 29]
[note 30]
(Figure 21)
(Figure 22)

Page 13

[note 31]
[note 32]
[note 33]
[note 34]
[note 35]
(Figure 23)

THE VIRTUAL TIMES

CHASING THE WILD DRAGON:

Foundations of a New Science

By T.E. Bearden
November 12, 1995

It is a pleasure to accept the offer of John and Larry to write a column in their new InterNet magazine. Theirs is a great venture, and with the energy and knowledge they bring to the task, one can be confident of their success in making this a very interesting and informative publication.

There is a new "rebel" science slowly being born, with most of the work being done outside the universities and orthodox scientific community. In the months ahead it will be my job to try to acquaint you with the gist of what's going on, and what all the excitement is about.

This first article is deliberately just introductory, and its purpose is to interest you in the unusual fields we are going to be covering in more scientific detail, in future columns. We will also be listing a variety of references from the technical literature, which bear upon the various points addressed.

In this first article a little of my own background experience [note 1] is given, so you can decide whether you will be interested in what we have to say!

The real purpose of the column will just be to *present information* that you may not be able to readily obtain any other way. We will not argue or debate with skeptics; nothing productive is to be gained by that. If the information we present is useful to you, then our purpose has been accomplished. If it isn't of use to you, then just file it in the old wastebasket. Also, from time to time I may answer a generic question, constructed from multiple queries. Individual responses to correspondence cannot be provided.

The work toward a new breakthrough science is going on in a broad spectrum of "fields," most of which are not yet recognized by academia as even constituting legitimate areas of scientific endeavor (Table 1).

Some of the major areas involved are:

1. theoretical and experimental research in overunity electrical machines,
2. free energy research of other kinds, such as devices utilizing thermodynamics phase changes,
3. rapidly advancing efforts in cold fusion,
4. work on permanent magnet motors,
5. antigravity research, both theoretical and experimental,
6. development and utilization of a "hidden" electrodynamics inside the scalar potential (inside voltage),
7. the "memory" of space and materials; where this "memory" can be deliberately conditioned so that space and the materials themselves become causal actants rather than simple inert entities,

8. direct engineering of quantum change via a new kind of "hidden electrodynamic variables,"
9. the direct engineering of the vacuum's virtual photon flux exchange with matter,
10. action at a distance, including the deliberate formation and usage of quantum potentials,
11. direct engineering of the local curvature of spacetime (yes, the direct engineering of general relativity, using the new hidden kind of electrodynamics),
12. use of the new "hidden" electromagnetics to engineer body cells including the reversal of diseased cells, genetics and all, back to a previous healthy state (cellular dedifferentiation, in biophysical terms, or cellular phase conjugation, in physics terminology),
13. research toward the direct engineering of the mind, including both the conscious and unconscious minds, long-term memory, and the personality itself, and
14. a ruthless re-examination (and correction where necessary and possible) of the definitions, concepts, and postulates that present physics is founded upon.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

---

[1] For those interested in orthodox qualifications, I have an M.S. in nuclear engineering (1971) from Georgia Institute of Technology and a BS in mathematics (1953) from what is now Northeast Louisiana University with a minor in electronic engineering.

Table 1. Fourteen Emerging New Science Areas

1. Overunity Electrical Machines and Power Systems
2. Overunity Devices Using Thermodynamical Phase Changes
3. Cold Fusion and Non-Nuclear Transmutation
4. Self-Powered Permanent Magnet Motors
5. Antigravity and Antigravity Engineering
6. Use of the "Inner" EM Inside the Scalar Potential
7. Vacuum Engineering, With Vacuum and Materials Memory
8. Direct Engineering of Quantum Change
9. Engineering and Structuring of Massless Charge Flux
10. Action-At-A-Distance, Including Use of Quantum Potentials
11. Direct Engineering of Local Spacetime Curvature
12. "Inner EM" Time Reversal of Diseased Cells Back to Health
13. Direct Engineering of the Mind, Including Software Loading
14. Unified Field Revision of Foundations Concepts of Science

Each of these "major areas" as yet has no single "experts" in the sense that individuals exist who possess a "mastery" of the given area. As an example, in spite of multiple researchers and decades of vigorous research, no one is really an "overall expert" in the overunity electrical power systems field. The "field" itself is still struggling to be born and is not recognized at all by orthodoxy; in fact, many orthodox scientists bitterly resent one even attempting to work in this area. There is as yet no cohesive single "theory" of overunity electrical or magnetic machines, although our own group is making rapid progress in that respect. But there is still no universal agreement amongst even the researchers themselves. Literally the field is still being *discovered*.

This "groping dimly toward a new field" should not be too surprising. After all, the early electromagnetics researchers went through the same tortuous, agonizing kind of development. They were playing with cat fur, glass rods, and pith balls for quite an extended period before understanding progressed to the point where Maxwell could mathematically tie together and model the results of the experimenters such as Faraday and others. It has taken a century and a half for modern electromagnetics and electronics to reach its own present development. It is going to take another decade or so before the overunity researchers can overturn several ill-founded parts of the present electromagnetics and get the emerging field onto a solid foundations footing and into a solid theoretical position.

In the effort to produce a theoretical model of electromagnetics, Maxwell and his predecessors performed a magnificent work. Yet they also made very serious errors (Table 2), which persist to this day. [note 2] Let us list just a few of the more blatant errors in classical EM theory today:

First, Franklin guessed wrong on which way electrical current flowed in a circuit. There is no definition at all of electrical charge in all of physics, and there is no acceptable definition of potential (that is, there is no rigorous definition of just good old "voltage.") There are no "fields of force" in the vacuum; there are just potential gradients __ because the vacuum is just one vast potential (virtual particle flux) in the first place. Force fields actually occur only in and of the matter in a material system. The primary causative agents of electromagnetics are not the so-called "force fields' at all, but are the potentials __ as is well-known in quantum mechanics. [note 3] Maxwell assumed a material ether in his equations. The Michelson-Morley experiment destroyed the material ether, but the Maxwell equations have never been changed accordingly. The electron was not yet discovered when the Maxwell equations were written; the theory does not address electron spin. Maxwell was already dead some two decades when Barus' 1898 paper [note 4] was published, pointing out the strange "backwards-traveling" wave (i.e., time-reversed wave) solution to the wave equation. In addition, Heaviside and Gibbs had already produced their vector truncation __ this truncated theory is the modern so-called "Maxwell's equations" that universities teach today __ of Maxwell's theory. Not a single one of them ever appeared in any book or paper by James Clerk Maxwell. [note 5]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page 1

Comments may be sent to
the **Editor** of The Virtual Times

Table 2. Some Defects In Electrodynamics

1. Conventional Current Direction Is Reversed
2. No Definition of Charge or Potential
3. Equations Still Assume Material Ether
4. Use of Force Fields in Vacuum is Known to be False
5. Treats Charge as Unitary; lnstead, It is a Coupled System
6. Confuses Massless Potential Gradients as Forces
7. Does Not Utilize Mass as a Component of Force
8. Erroneously Assumes Force Fields as Primary Causes
9. Topology of EM Model Has Been Substantially Reduced
10. Does Not Include Quantum Potential
11. Does Not Include Proven Action At A Distance
12. Does Not Utilize Near Field Coulornb Gauge Effects
13. Does Not Include EM Generatrix for Time Flow
14. Does Not Unify Photon and Wave Aspects
15. Does Not Include Electron Spin and Precession
16. Treats Energy As Existing In "Chunks," instead of as Flow
17. Confuses Energy and Energy Collection
18. Discards Half of Every EM Wave in Vacuum
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal
20. Omits Phase Conjugate Optics Effects
21. Does Not Include EM Cause of Newtonjan Reaction Force
22. Erroneously Assumes Separate Force Acting on Separate Mass

[2] For an independent statement that electrodynamics, both classical and quantal, are in serious disarray, see Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176. Quoting: "...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state." Bunge points out many errors in electrodynamics. See also Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): Essays on the Formal Aspects of Electromaqnetic Theory, World Scientific, Singapore, 1992, p. 6-86. To find information on what's wrong with EM theory, one must read foundations literature, not the standard textbooks. One of my own contributions has been to point out that the notion of charge q is not unitary. In fact, electrical charge should be defined in terms of q ~ m(sub q)~(sub q). In other words, electric charge q has a massive part m, and a massless part consisting of its potential phi. Further, the magnitude of phi is just the change in the virtual photon flux of the ambient vacuum, due to the [quantum field theoretic] virtual photon exchange with the mass of the q. It follows that the true electrical charge of a particle is just its native potential ~, which is also a dynamic energy exchange with the surrounding vacuum. It also follows that this massless electrical charge changes whenever the particle is placed in a potential (in a different ambient virtual photon flux). After all, potentials superpose; that is their major characteristic. Note that the potential of the test particle is ignored in classical electromagnetics, whenever one speaks of the "E-field" upon the particle. Further, the actual structure of this virtual photon flux that comprises massless electrical charge, can itself be deterministically structured and utilized to generate nonlinear effects that do not appear at all in the present conventional theory. We will be covering many of these effects in future articles.

[3] E.g., see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. Effects of potentials on charged particles exist even in the region where all the fields (and therefore the forces on the particles) vanish, contrary to classical electrodynamics. The quantum effects are due to the phenomenon of interference. These effects occur in spite of Faraday shielding. The Lorentz force does not appear anywhere in the fundamental quantum theory, but appears only as an approximation that holds in the classical limit. In QM, the fundamental physical entities are the potentials, while the fields are derived from them by differentiation.

[4] Carl Barus, "A Curious Inversion in the Wave Mechanism of the Electromagnetic Theory of Light," American Journal of Science, Vol. 5, Fourth Series, 1898, p. 343-348.

[5] E.g., see Paul J. Nahin, Oliver Heaviside: Saqe in Solitude, IEEE Press, New York. Quoting from p. 9: "...nowhere in Maxwell's writings do the equations for the electromagnetic field appear as we write them today. Maxwell used an amalgamation of Cartesian component and quaternion notation, and it was Heaviside who first wrote the electromagnetic field equations in modern vector form." Quoting from p. 92: "In actuality , however, the fields were not the primary reality to Maxwell at all (that is an idea that developed after Maxwell's death and is due to Hertz and Heaviside), but rather it was Faraday's 'electrotonic state' that he thought to be the real thing. Like Faraday, Maxwell believed that electromagnetic effects are observable results of an altered state of the ether. The mathematical formulation of this electrotonic state, for Maxwell, is what we today (as he did) call the vector potential." And again, from p. 96: "To Maxwell, however, the vector potential had a most definite physical meaning."

So the time-reversed EM wave that one finally sees emerging in nonlinear phase conjugate optics in the early 1970s was omitted from the Maxwellian theory and its later modification by Heaviside and Gibbs. Maxwell's actual theory is in a higher topology (quaternion algebra), and involves some 20 quaternion equations. Heaviside and Gibbs created vector algebra, and seriously curtailed Maxwell's theory when they "translated" it into the much lower topology of vectors and some four equations. The later addition of tensors as the basis mathematics did not restore the lost topology provided by the quaternion model.[note 6]

For the EM wave, Maxwell simply assumed the transverse EM wave that had come from the much earlier derivation of the familiar wave equation from the old "plucked taut string" notions. [note 7] Yet when one closely examines and corrects some of the erroneous *assumptions* [note 8] made for the actual mathematical derivation of this "taut string" transverse wave, one finds (Figure 1) that an equal-energy second antiphased transverse "antiwave" is actually produced by the instrument and added to the "plucked string's" transverse wave disturbance of the air medium, to produce a longitudinal wave of compression and rarefaction in the air itself. The "transverse string wave" in fact never leaves the string. The string wave is not the wave that is launched into the medium.

A similar electromagnetic situation exists. In a whip antenna (Figure 2), the Drude electron gas in the conductor does exhibit a transverse "string-type" EM wave of electron precession. However, at the same time, by Newton's third law (which Maxwell neglected because at the time the atom was regarded as just a blob with no internal structure) the positively charged, phase conjugated atomic nuclei have equal energy, highly damped transverse recoil waves created in them. Further, these "nuclei-recoil" waves are phase conjugate replicas of the electron gas waves. The total disturbance that is launched into the vacuum from the whip antenna is (a) the normal transverse disturbance of the virtual photon flux of vacuum caused by the electron gas wave, and (b) the coupled equal-energy transverse antiwave caused by the nuclei-recoil via Newton's third law. The amplitude of the material antiwaves of nuclei vibration are highly damped because of the massiveness of the nuclei; however, they are of equal energy to the electron wave. But when launched into the same vacuum, the two equal-energy wave disturbances experience equal damping because they are in the same medium. Hence the wave and antiwave are (a) locked together in the vacuum, and (b) of equal energy, and (3) of equal amplitude once launched into the vacuum. What results is a vacuum EM wave of rarefaction (reduction in intensity) and compression (increase in intensity) of the virtual photon flux of vacuum __ just exactly like what is launched into the surrounding gaseous air medium from the "plucked taut string" and perturbed body of a physical instrument. The EM wave in vacuum is a longitudinal wave directly analogous to a sound wave __ just as Nikola Tesla stated. [note 9]

The vacuum may be considered a scalar potential, in which case it has an internal biwave structure (Figure 3). Interference of two potentials __ i.e., the hidden interference (Figure 4) of their internal wave structures __ in fact already creates all EM fields and waves.

When the incoming longitudinal "biwave" from the vacuum strikes the atoms of the receiving antenna, each atom may be considered a dynamic assembly of multiple dipoles, where each dipole consists of one of the electrons in the electron shells and one of the positive charges in the nucleus. The receiving dipole in the interacting atom splits the two waves apart again (Figure 5). The negatively charged end of the dipole is time-forward, and the positively charged end is time-reversed.

Hence the separated normal-wave-half of the biwave interacts with the electron, to give again the familiar transverse "electron precession" wave that we detect with our "electron wiggle-detecting" instruments (Figure 6). This detected electron precession wave is what all our textbooks prescribe erroneously as the vacuum wave! The separated time-reversed-wave-half of the incoming vacuum wave is split off and interacts with the positive end of the receiving atomic dipole, in the time-reversed atomic nucleus, thus generating Newton's third law reaction force to provide the recoil of the nucleus. Present physics has ignored this second wave altogether, because of the serious error in the original derivation of the wave equation from plucked strings. Thus physics presently has absolutely no causative mechanism for Newton's reaction force. Instead, scientists just "mystically invoke the appearance of this magic force." Describing that something occurs, and naming its occurrence, does not constitute advancing a causative mechanism!

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page 1

Comments may be sent to
the **Editor** of The Virtual Times

[6] E.g., see T.W. Barrett, "Tesla's Nonlinear Oscillator- Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41. Barrett shows that a higher topology EM model (e.g., quaternions) allows shuttling and storage of potentials in circuits, and also allows additional EM functioning of a circuit that a conventional EM analysis cannot reveal. Tesla's actual patented circuits in fact exhibit this higher topology, when a quaternion EM analysis of them is performed. No orthodox vector/tensor analysis can show this.

[7] All the early experimenters seemed to assume the transverse string-type wave automatically for the vacuum medium. E.g., Faraday speculated that light and radiant heat might be transverse vibrations propagated along his "lines of force," from which the direct "plucked string" analogy can readily be seen. In short, he considered that his lines of force could be perturbed to vibrate laterally, just like a plucked string. See M. Faraday, "Thoughts of Ray-Vibrations," Philosophical Magazine, (3), vol. xxvii, 1846, p. 345. Maxwell also assumed the transverse EM wave in vacuum without question, as did Heaviside, Hertz, and Gibbs and as has almost every electrodynamicist since then.

[8] The worst is the assumption of infinitely rigid endposts holding the taut string, so that no movement of these endposts occurs. This further assumes that no induced movement of the body of the instrument occurs. The end result is to eliminate all of that wave that is produced in the body of the instrument, and retain only the transverse wave produced in the string. In the first place, this standard procedure has now become a total violation of Newton's third law, which among other things requires that forces be formed in oppositional pairs. So the standard method for the launch of a wave from an instrument is flawed. Specifically, the "transverse string wave" is falsified insolaf as its appearance in the surrounding medium. This is not the wave that actually exits the instrument and that is launched into the medium. This universal oversight on the part of mathematicians and physicists is a colossal joke upon them. Literally, when a physicist writes the transverse wave equation on the board, he has just discarded exactly half the physics. This means that, in his interactions, at the end there will suddenly appear this mysterious, mystical "equal and opposite reaction force" that is created by the erroneously ignored antiwave. So he has to piously invoke this reaction force "from the blue." Yet quantum field theory teaches us already that all mechanical and electromagnetic forces on a mass are generated by absorption and emission of virtual photons. Hence Newton's third law reaction force must also be so generated. This means that there must exist an equal number of virtual antiphotons interacting in the antiforce, as there was virtual photons interacting in the causative force. By symmetry there must also exist an equal and opposite wave, or antiwave, made of these same antiphotons. The actual disturbance that entered from the vacuum must therefore have consisted of both wave and antiwave. QED.

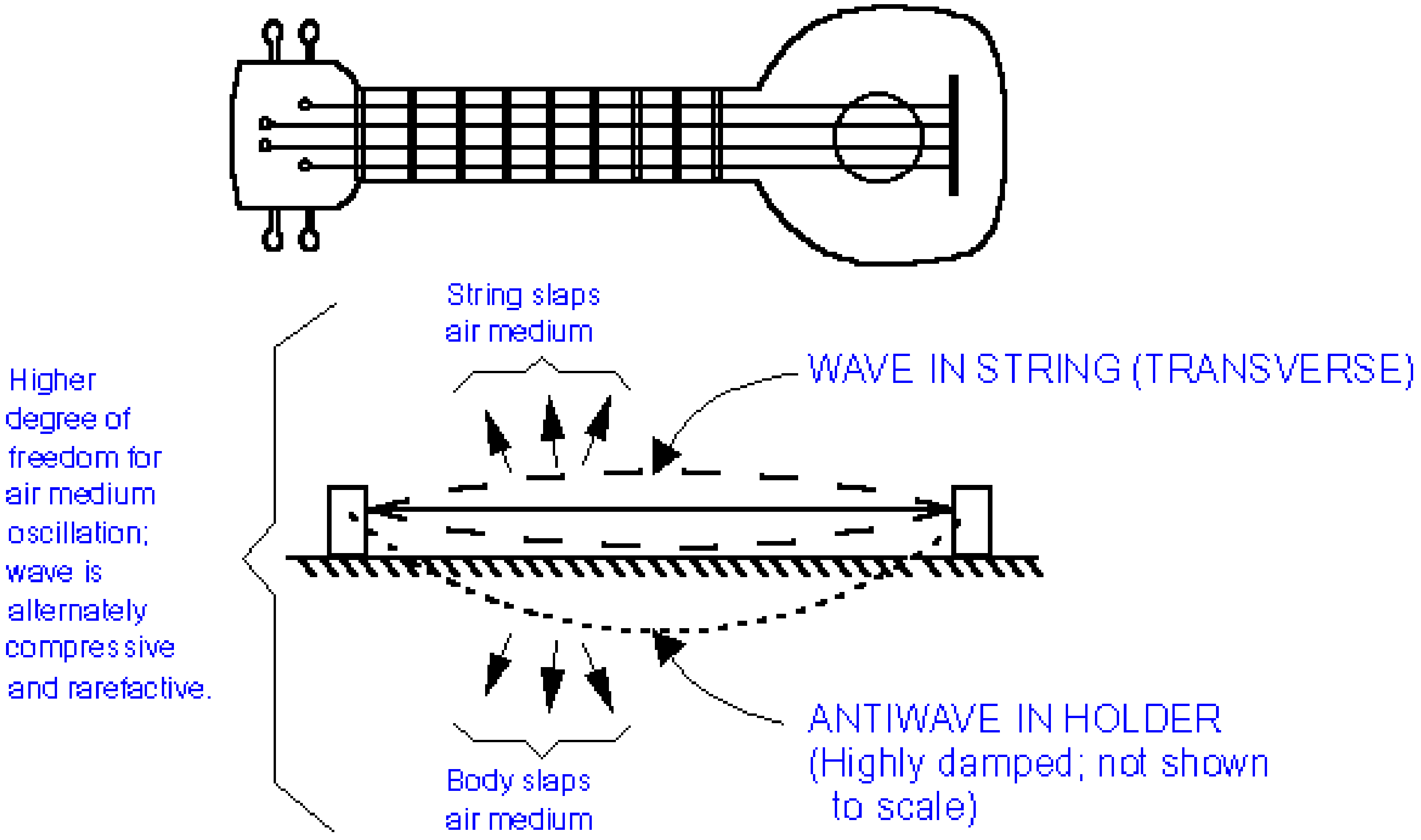
String slaps
air medium
WAVE IN STRING (TRANSVERSE)
Higher
degree of
freedom for
air medium
oscillation;
wave is
alternately
compressive
and rarefactive.
ANTIWAVE IN HOLDER
(Highly damped; not shown
to scale)
Body slaps
air medium

Figure 1. Air medium disturbances generated by plucked string.

Two waves are created, both slapping the air medium at the same time, but in opposite directions. A disturbed medium oscillates according to its own degrees of freedom. The air has more degrees of freedom than the string. A wave of compression and rarefaction is created in the air medium.

D-1-F 1

© 1993 T.E. Bearden

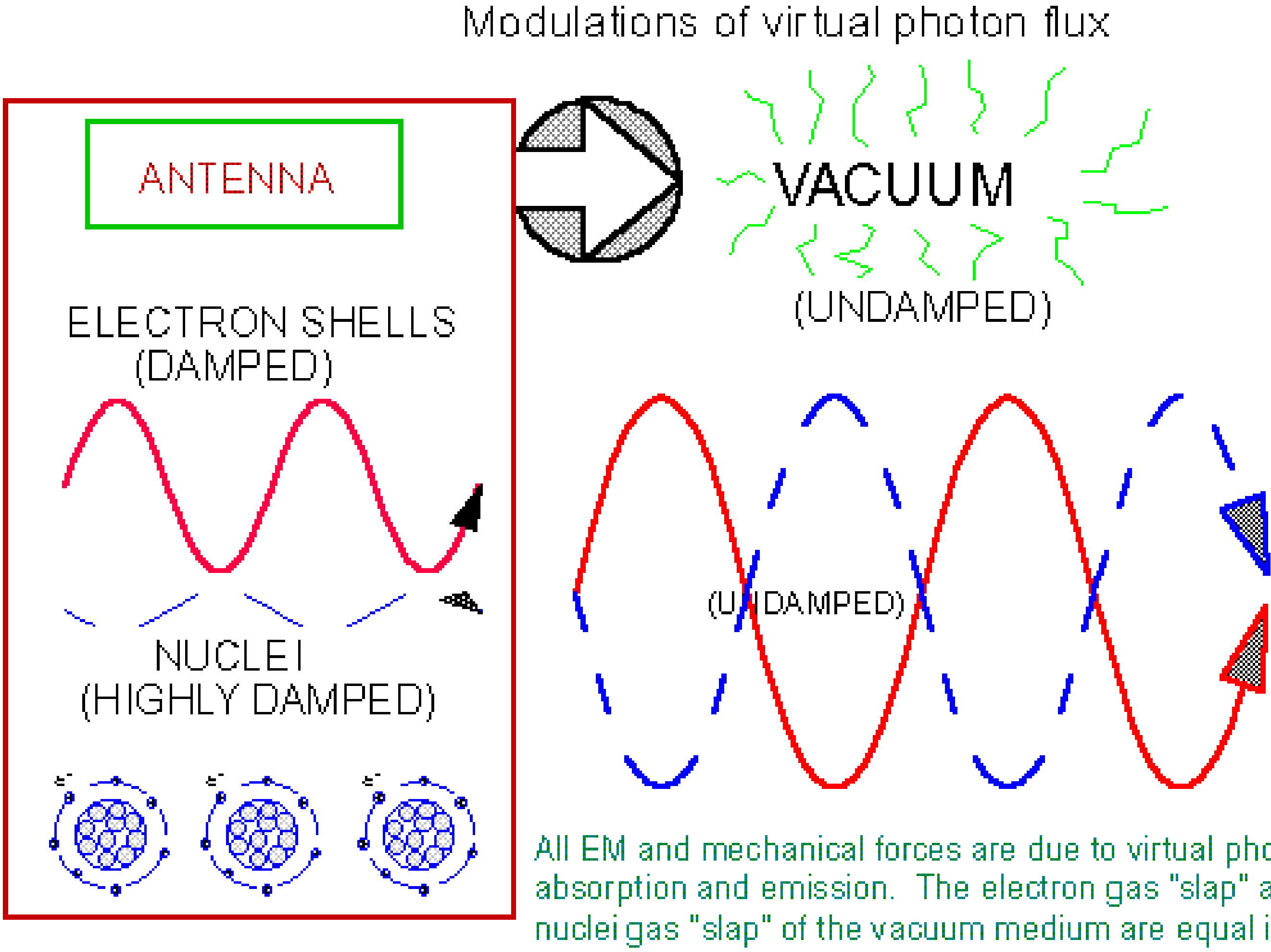
Modulations of virtual photon flux
ANTENNA
ELECTRON SHELLS
(DAMPED)
NUCLEI
(HIGHLY DAMPED)
e-
e-
e-
VACUUM
(UNDAMPED)
(UNDAMPED)

All EM and mechanical forces are due to virtual photon absorption and emission. The electron gas "slap" and the nuclei gas "slap" of the vacuum medium are equal in energy. Hence they inject equal numbers of virtual photons into the vacuum, creating equal and opposite vacuum disturbances. This gives a vacuum wave of alternating compression and rarefaction. This vacuum EM wave is similar to a sound wave, which is modeled longitudinally.

Figure 2. Launching an EM wave from a wire antenna.

D-1-F2

© T.E. BEARDEN 1994, 1995

[9] See Nikola Tesla, "Pioneer Radio Engineer Gives Views on Power," New York Herald Tribune, Sep. 11, 1932. Quoting: "...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light." Today the QM vacuum has been theoretically and experimentally shown to consist of a "gas" of virtual particles, rather like Tesla stated. Tesla was also correct about the longitudinal EM wave in the vacuum. He was also correct in his statement that a fundamental charged particle can alter its charge, if one considers the massless charge ~ of charge q separately from the mass m of charge q. When the charge q is placed in a nonzero potential V, it follows that the new charge of the electron is simply (~+V).

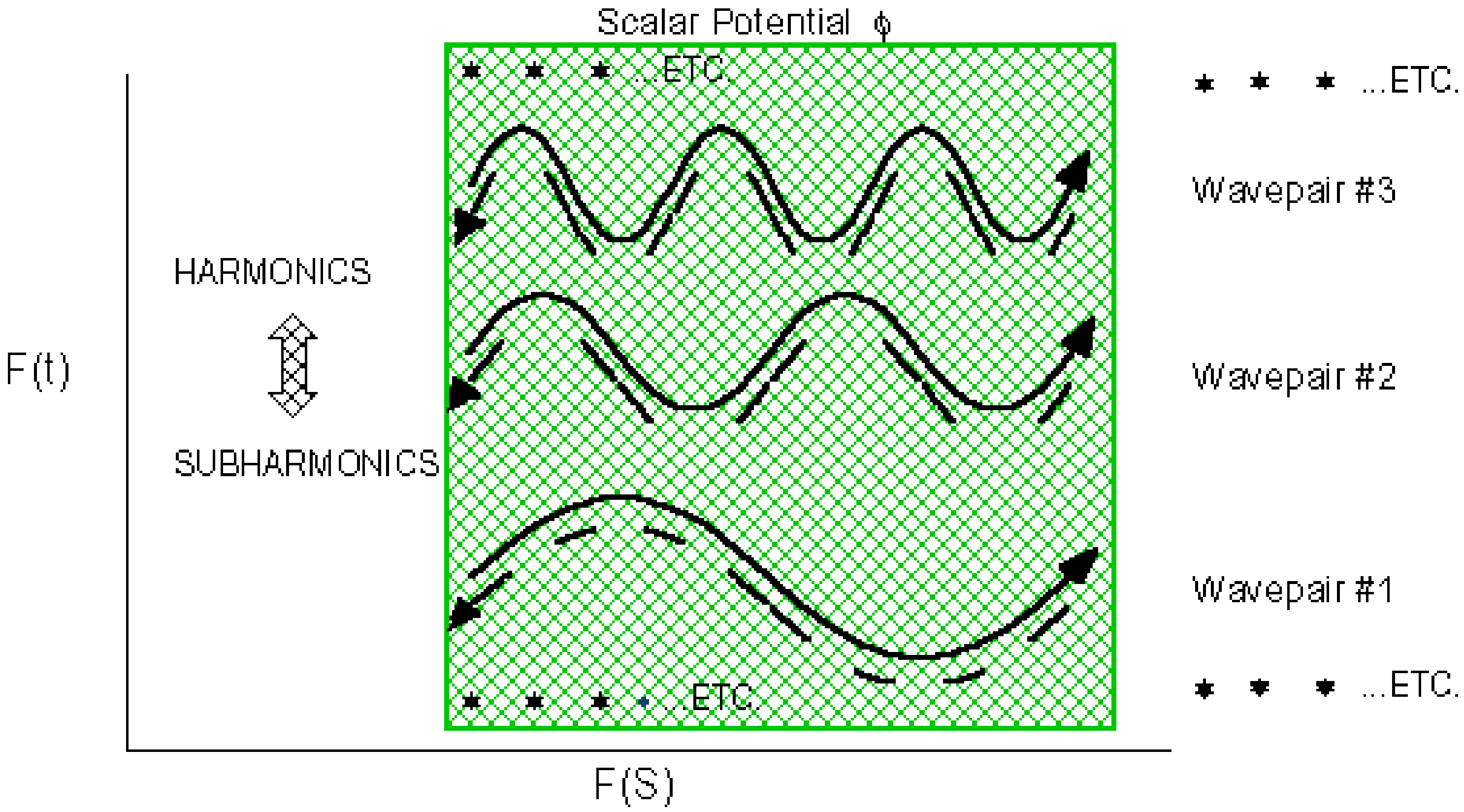
Scalar Potential ϕ
F(t)
HARMONICS
SUBHARMONICS
F(S)
...ETC.
...ETC.
...ETC.
...ETC.
Wavepair #3
Wavepair #2
Wavepair #1

Figure 3. Internal Wave Structure of the scalar potential.

A harmonic set if wavepairs, where in each wavepair the two waves superpose spatially, but travel in opposite directions. The two waves in each pair are true phase conjugates and time-reversals of each other. Thus they comprise a coupled wave and antiwave. The photons must also be coupled into photon/antiphoton pairs, by a strong application of the distortion correction theorem of nonlinear optics.
Each wavepair is a standing electrogravitational wave.

D-1-F3

© T.E. Bearden 1984, 1985

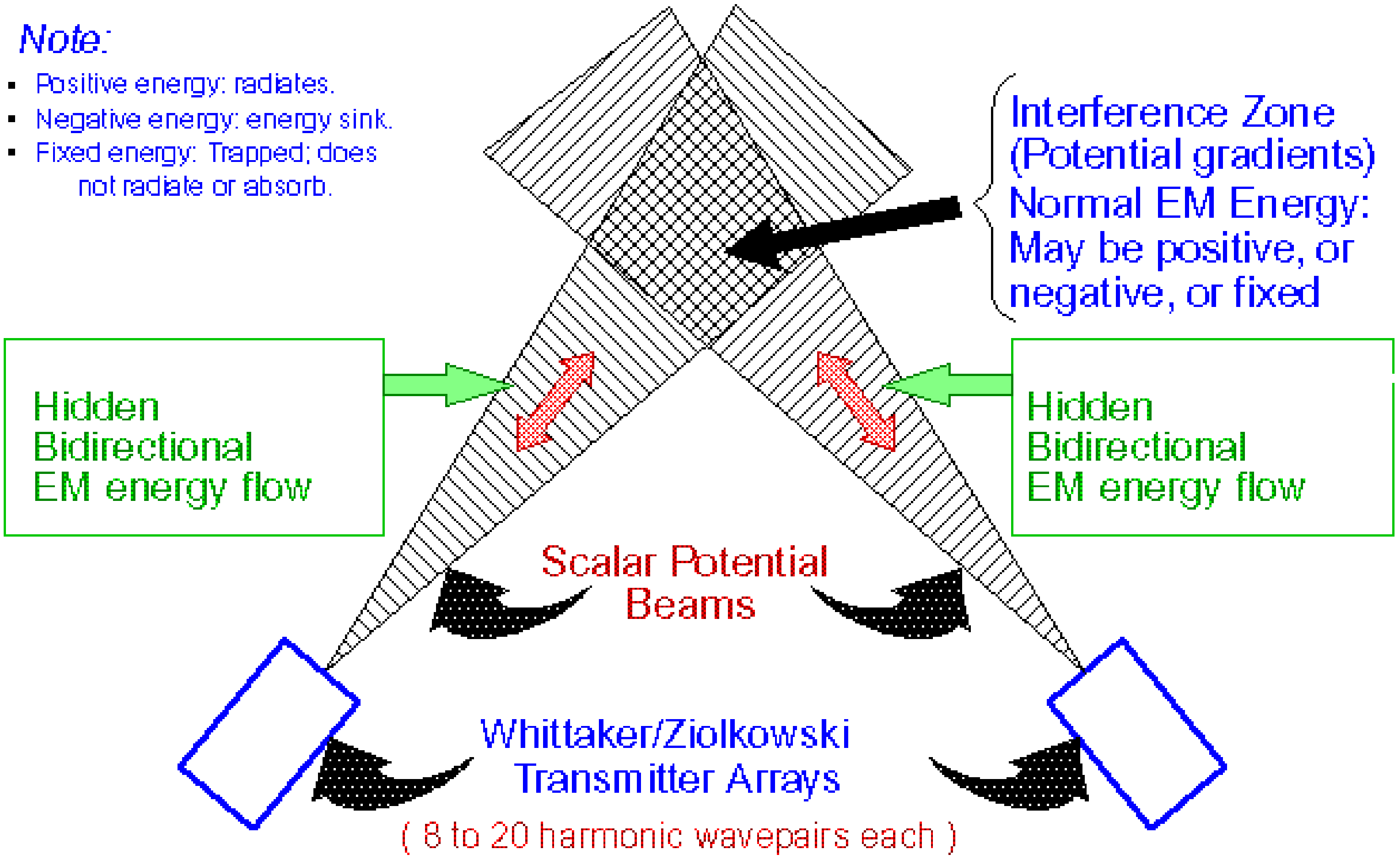

Note:

- Positive energy: radiates.
- Negative energy: energy sink.
- Fixed energy: Trapped; does not radiate or absorb.

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or
negative, or fixed
Hidden
Bidirectional
EM energy flow
Hidden
Bidirectional
EM energy flow
Scalar Potential
Beams
Whittaker/Ziolkowski
Transmitter Arrays
(8 to 20 harmonic wavepairs each)

Figure 4. Scalar potential interferometry provides action-at-a-distance. It already creates normal EM fields and waves.

D-1-F4

© 1994 T.E. Bearden

MECHANISM FOR NEWTON'S THIRD LAW

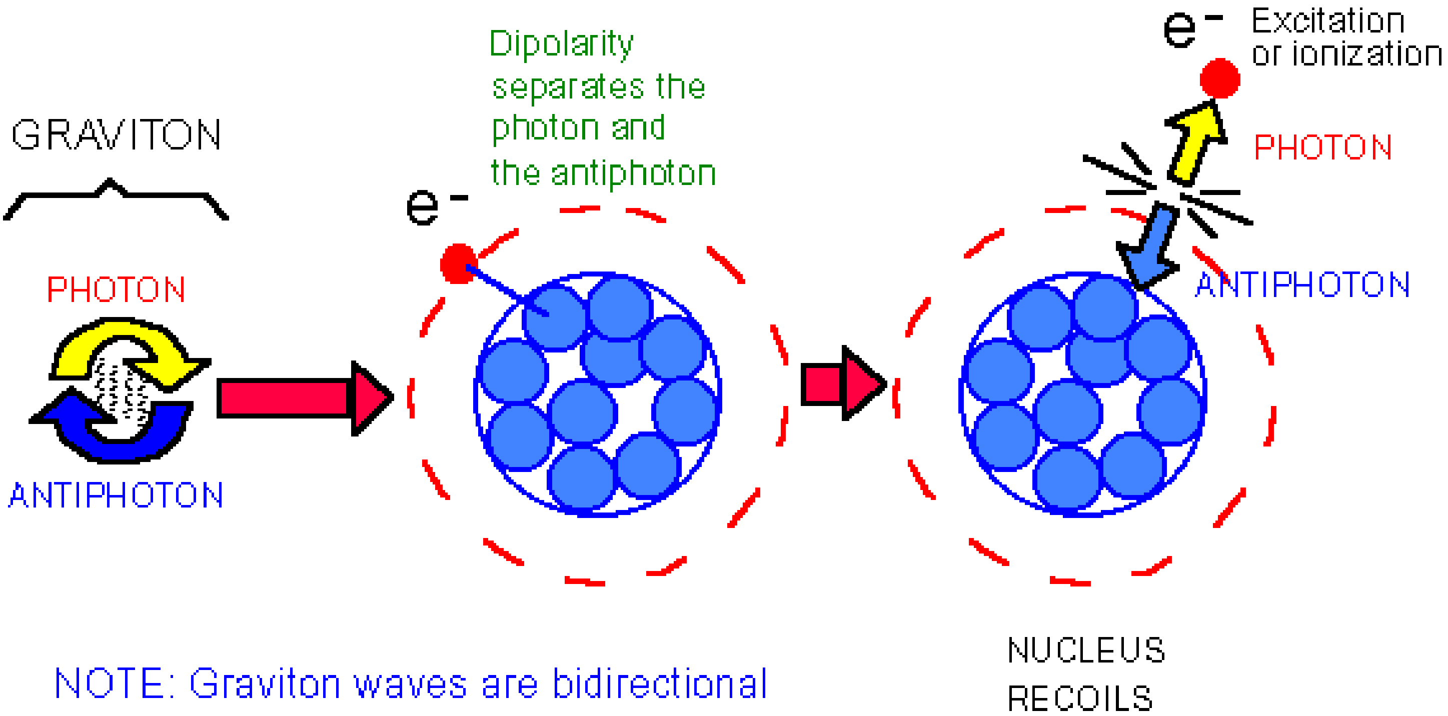
GRAVITON
PHOTON
ANTIPHOTON
Dipolarity separates the photon and the antiphoton
e -
e -
Excitation or ionization
PHOTON
ANTIPHOTON
NUCLEUS RECOILS

NOTE: Graviton waves are bidirectional phase waves of photon/antiphoton coupling and uncoupling in the vacuum.

The atom is an assembly of dynamic dipoles.

A dipole splits the graviton into a photon and an antiphoton.
The photon interacts with the electron, producing the action.
The antiphoton interacts with the nucleus, producing the reaction.

Figure 5. Graviton interaction with matter produces the photon interaction and Newton's third law reaction.

D-1-F5

Tables:

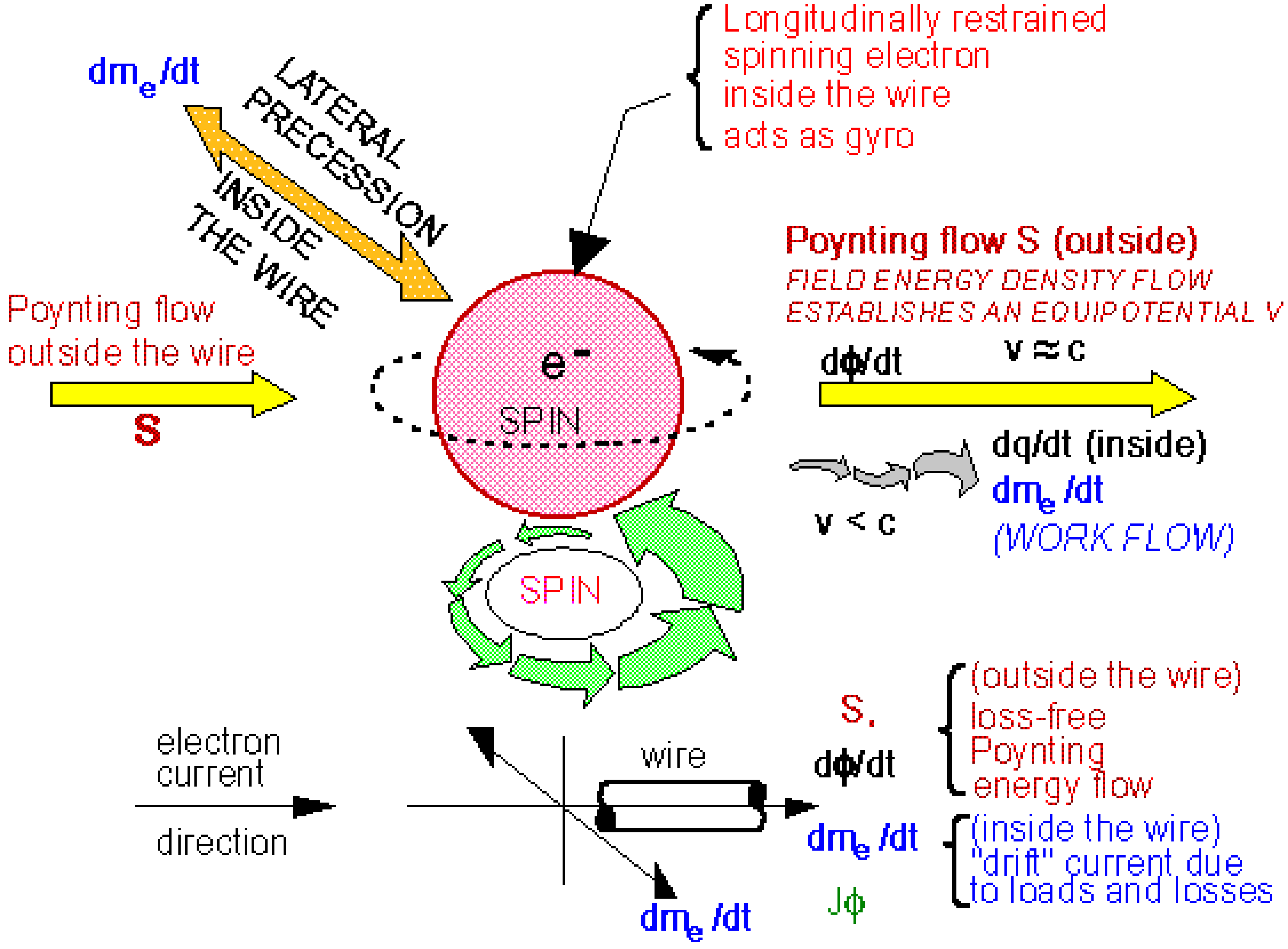
dm_e/dt
LATERAL PRECESSION INSIDE THE WIRE
Longitudinally restrained spinning electron inside the wire acts as gyro
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW ESTABLISHES AN EQUIPOTENTIAL V
Poynting flow outside the wire
S
e^-
SPIN
$d\phi/dt$
$v \approx c$
dq/dt (inside)
dm_e/dt
(WORK FLOW)
$v < c$
SPIN
electron current
direction
wire
S.
$d\phi/dt$
(outside the wire) loss-free Poynting energy flow
dm_e/dt
(inside the wire) "drift" current due to loads and losses
$J\phi$
dm_e/dt

Figure 6. Currents in and along a conductor, with electron precession.

D-1-F6

© 1993 T.E. Bearden

Further, if we should trick the time-reversed wave half into coming outside the atom instead of going inside to interact with the nucleus (Figure 7), then it follows that the nucleus would not recoil. In that case the interaction would appear to violate Newtonian third law reaction. Phase conjugate mirrors already do just exactly this when they emit a phase conjugate replica wave. They direct the antiwave outside and away from the atom instead of inwards into the nucleus. So they do not recoil, no matter how powerful a phase conjugate replica wave they emit.

Finally, we can also show that the EM wave in vacuum is longitudinal, by yet another approach. The electrons in the electron gas in a conductor are spinning, and they are severely constrained longitudinally to be able to move only very, very slowly down the wire. So when perturbed, they act as gyroscopes, and precess at right angles to the perturbing force. The fact that the gyroelectrons in a wire move radially almost entirely, and only "slip" down the wire very slowly with a "drift" velocity, already proves conclusively that the incoming vacuum disturbance was longitudinal. The precession of a gyro is at right angles to the perturbing force. Therefore rigorously the vacuum EM waves are longitudinal, because the gyroelectrons moved transversely. The present assumption of the transverse EM wave in vacuum contradicts the entire theory of gyroscopes! [note 10] There are many other errors in classical EM that we will point out in future articles; this should give the reader a flavor of what is in store for him in future articles.

So there are no truly satisfactory "experts" in any of the 14 fields. Yet there is still something to be said for having some years of experience in one or more of the areas. I have had the good fortune or misfortune, depending upon one's viewpoint, to be involved in various of those 14 activities for some 30 years or more, and am active in several of the areas now. My colleagues and I are particularly active in researching overunity electrical devices (see Figure 8), to include the effects of controlled chaotic oscillators, charge-blocking oscillators, and oscillators driven by in-shuttled (bridged) Poynting field energy density flow ExH. Parametric oscillation is known to have yielded overunity electrical machines and was fully reported by Russian researchers in the Russian and French technical journals in the mid 1930s. [note 11] We will further cite that little-known work in future articles. Also, we are working mightily upon true negative resistors (Figure 9), wherein the resistor scavenges some of the disorganized energy from the vacuum's energetic exchange with the charges in the system, organizes and collects it, and then outputs that collected energy flow into the external circuit. Such a resistor acts as a power source. Kron, perhaps the greatest of the electrical machinery scientists, is known to have produced such true negative resistors.

I have also proposed a theoretical explanation of the Priore device (Figure 10) in France which in laboratory animals demonstrated remarkable cures of terminal tumors, cured trypanosomias, reversed atheriosclerosis, and restored suppressed immune systems. The Priore work was not anecdotal, but was performed under rigorous scientific protocols, working under the auspices of renowned French scientists such as Courrier and Pautrizel. The work is properly reported in numerous papers in the French medical literature. [note 12] The results were certainly achieved, but neither Priore himself nor the scientists of the day could understand the mechanism involved. The deciphering of that causative technical mechanism utilized by the Priore group required 10 years of intense effort. The results have now been published in the journal Explore! and are there for anyone to read. [note 13]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

PHOTON INTERACTION WHEN ATOM PHASE CONJUGATES

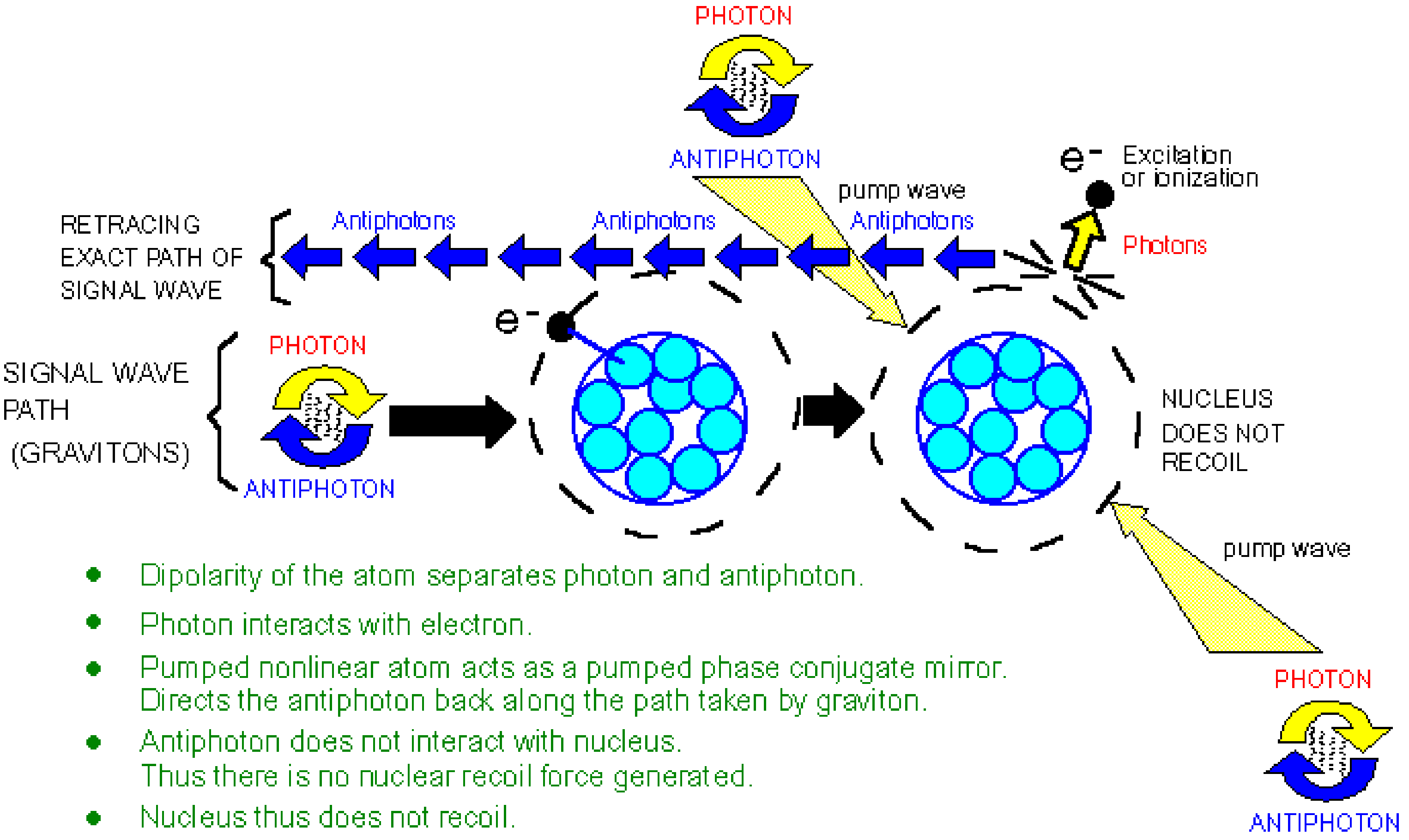
PHOTON
ANTIPHOTON
pump wave
e⁻
Excitation
or ionization
Photons
RETRACING
EXACT PATH OF
SIGNAL WAVE
Antiphotons
Antiphotons
Antiphotons
e⁻
SIGNAL WAVE
PATH
(GRAVITONS)
PHOTON
ANTIPHOTON
NUCLEUS
DOES NOT
RECOIL
pump wave
PHOTON
ANTIPHOTON

- Dipolarity of the atom separates photon and antiphoton.
- Photon interacts with electron.
- Pumped nonlinear atom acts as a pumped phase conjugate mirror. Directs the antiphoton back along the path taken by graviton.
- Antiphoton does not interact with nucleus. Thus there is no nuclear recoil force generated.
- Nucleus thus does not recoil.
- Not a violation of Newton's third law, once one understands the mechanism that generates third law reaction. In the absence of the generating mechanism, there is no third law recoil. In EM field theory, the "missing antiwave" has necessitated third law omission.

Figure 7. Emission of antiphoton; nucleus does not recoil. Atom acts as a pumped phase conjugate mirror.

D-1-F7

© 1995 T.E. Bearden

[10] When Maxwell's theory was constructed, the electron had not yet been discovered, and so the dipolarity of the atom was unknown. An atom was regarded just as a "blob." Electricity in a conductor was regarded as a thin material fluid Maxwell's equations are in fact hydrodynamic equations. So the founders regarded the detection of induced motion in the electric fluid as induced translation, not induced precession. They had no idea of "spins" of particles of the fluid, nor of gyro precession of those particles. Instead, q was regarded just as a "piece of electric fluid," similar to a cubic centimeter of fluid.

A fundamental AC resonant principle for an overunity electrical power system

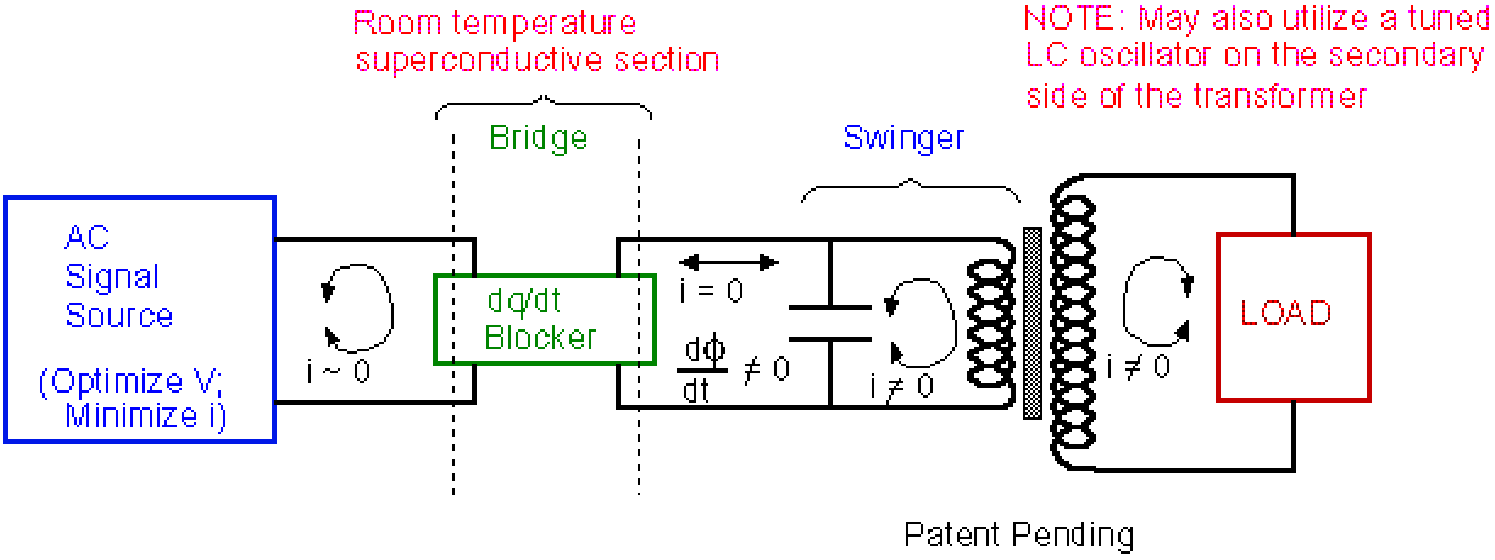
Room temperature
superconductive section
NOTE: May also utilize a tuned
LC oscillator on the secondary
side of the transformer
Bridge
Swinger
AC
Signal
Source
(Optimize V;
Minimize i)
i ~ 0
dq/dt
Blocker
i = 0
dϕ/dt ≠ 0
i ≠ 0
i ≠ 0
LOAD
Patent Pending

Figure 8. dq/dt-free driving a swinger by S-flow.

Poynting energy flow S with its accompanying potential V is shuttled from the source to a tuned LC oscillator in resonance at the driving frequency or slightly below. The LC oscillator is driven by S-flow and dV/dt flow alone. The swinger furnishes its own free electrons, which receive S and EMF from the bridge, thereby producing dq/dt.

D-1-F8

© T.E. Bearden 1993, 1994, 1995

[11] E.g., see L. Mandelstam [Mendel'shtam, L.I.], N. Papalexi, A. Andronov, S. Chaikin and A. Witt. (1935) "Report on Recent Research on Nonlinear Oscillations," Translation of "Expose Des Recherches Recentes Sur Les Oscillations Non Lineaires," Technical Physics of the USSR, Leningrad, Vol. 2, 1935, p. 81-134. NASA Translation Doc. TT F-12,678, Nov. 1969. Summarizes lengthy work at the Institute of Physics of the University of Moscow and several of the laboratories on the problem of self-powered, self- oscillating nonlinear oscillators. The paper reports significant work and progress in oscillations created by the oscillating system itself (self-oscillation) without participation of any external forces varying with time and without the expense of a constant source of energy (e.g., a storage battery). Such a system was shown to self-increase in an underdamped manner until system self-destruction. Adding a nonlinear load allowed the systems to stabilize its dynamic operation so that it powered a load under overunity conditions and did not self-destruct. Extensive additional references on this spectacularly successful overunity energy project are listed.

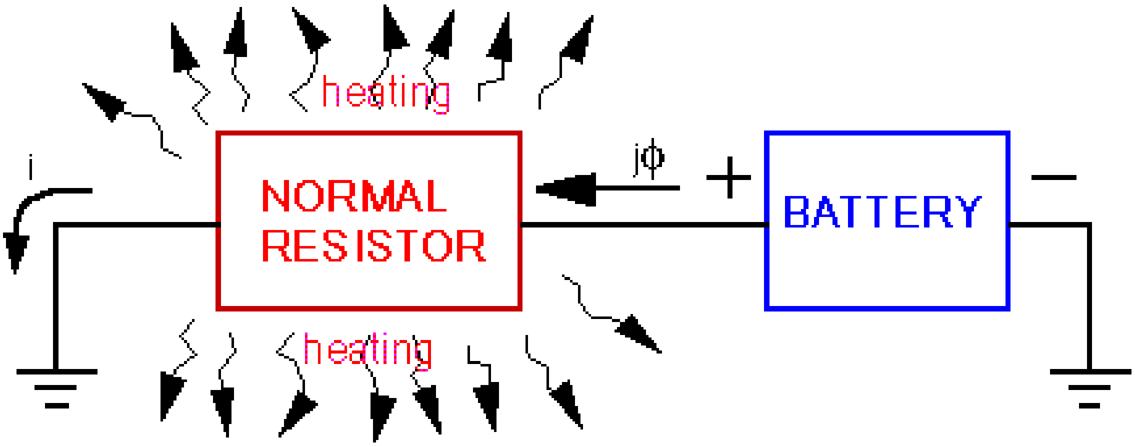
heating
i
NORMAL
RESISTOR
jϕ
+
BATTERY
−
heating

Figure 9a. Normal resistor receives ordered energy and scatters it.

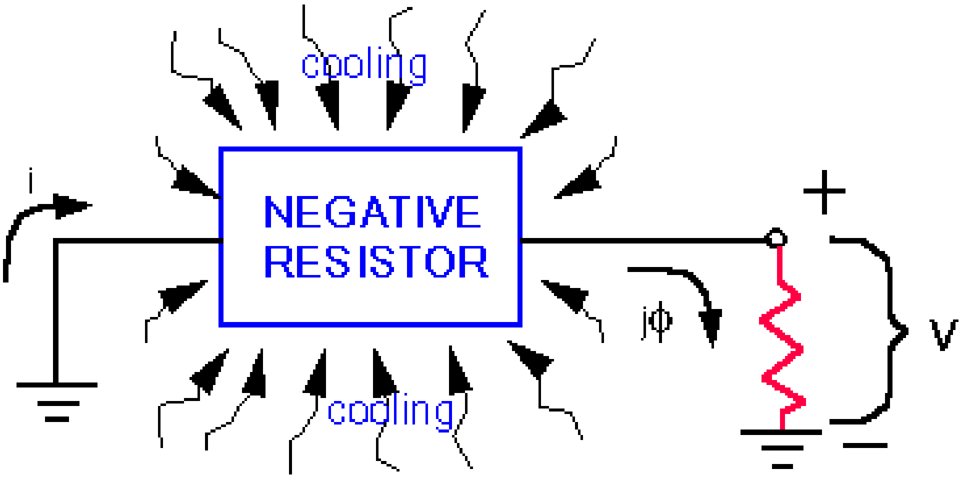
cooling
i
NEGATIVE
RESISTOR
+
jϕ
V
−
cooling

Figure 9b. Negative resistor receives, concentrates, and reorders scattered energy.

Figure 9. Normal resistor versus a negative resistor.

D-1-F9 © T.E. Bearden 1993, 1996

Tables:

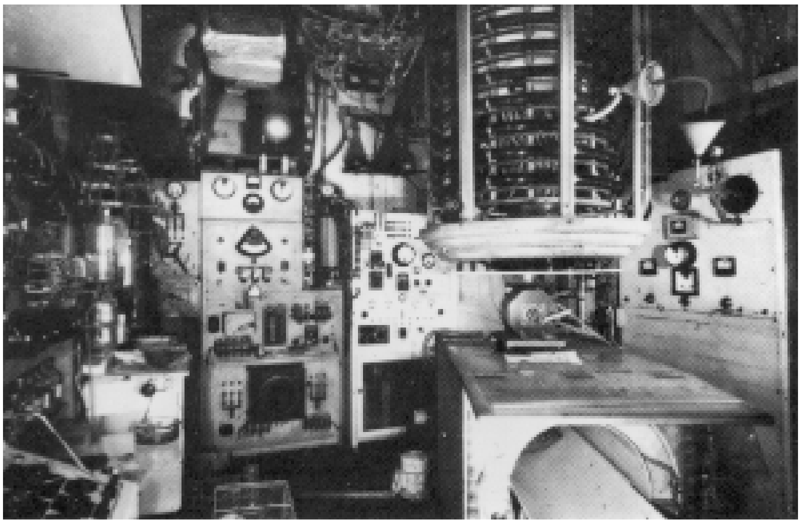

Some 17 signals were fed into a rotating plasma, which added phase conjugates to them. This infolded 17 wavepairs, each pair a coupled wave/antiwave, thereby forming a scalar potential with an artificial Whittaker pump wave internal structure. This infolded pump wave complex was modulated onto a powerful, rippling magnetic field, guaranteeing penetration even into the atomic nuclei. A specific, amplified antiengine for the exact cellular disease state was generated, reversing the diseased cells back to a previous healthy state.

Figure 10. Priore's laboratory apparatus for cellular reversal.

D-1-F10

© T.E. BEARDEN 1993, 1995

[12] For citations and the exposition of our proposed explanation of the Priore mechanism, see T.E. Bearden, "Vacuum engines and Priore's methodology: The true science of energy-medicine." Part I, Explore!, 6(1), 1995, p. 66- 76, Part II, Explore!, 6(2), 1995, p. 50-62.

[13] Bearden, "Vacuum engines and...," ibid.

THE VIRTUAL TIMES

Presently my CTEC colleagues and I have filed three patent applications and a continuation, on various methods for overunity processes and apparatuses, room temperature superconductivity, Poynting generators, etc.

As an example, we took exactly the opposite approach to room temperature superconductivity as has been taken by all the orthodox researchers to date. What exactly is sought, after all, for superconductivity (SC) in a section of an electrical circuit? Well, one has some electrons on one side of that SC section that are transporting excess potential, as given by the Slepian vector J*. [note 14] What is desired is to get some electrons on the other side of the SC section that also are transporting the same amount of excess energy in the form of J*. You can do that in two ways: (1) you can flow the electron carriers through the SC section, carrying their J*, or (2) you can block the J and flow the * itself across as the Poynting flow S. [note 15] Nondivergent Poynting flow flows along an equipotential, which is just another way of saying that, if the S-flow does not diverge, it carries the same potential * along with it. Hence it carries the EMF right along with it as it flows without divergence.

Conventional approaches have all tried to shove the electron carriers through the SC lattice section. Doing that is like trying to fire a very slow bullet through several million rotating fans in a straight line. So cryogenics (to slow the fanblades to a crawl) and massive correlation of the electrons and of the electron-to-lattice interactions is necessary if one is to get the electrons through there without excess collisions that shake off some of the excess * from J*, as scattered photons (heat) or as a change of form of the energy (as in straining the dielectric of a capacitor to convert electrical energy to mechanical strain energy via the piezoelectric effect.)

Our approach is exactly the opposite (Figure 11 and Figure 12). Why not just *stop* the flow of excited electron carriers on one side of the SC section, and continue the nondiverging flow of the Poynting field energy density S across that section at room temperature? Then the cryogenics is not needed at all. After all, circuits already work that way anyhow __ except standard practice is to nullify the process by letting current be driven around the sourcing current loop and back through the back EMF of the primary source. A quantum well (or several other methods) can be used to trap the "sourcing" electrons in the conductor just prior to entering the SC section. The "receiving" electrons on the other side of the SC section, however, must be in their own dq/dt-isolated current loop.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[14] Briefly, the S-flow outside the wire will be flowing along essentially at the speed of light, at or with the same since S-flow without divergence is along an equipotential. Said another way, a nondiverging Poynting flow S exhibits a constant potential ~ along its streamlines. A tiny, tiny portion of the S-flow is intercepted by the sluggish electrons (which may be moving as dq/dt at 11 feet per hour, in a typical case), to form the Slepian vector J~. The J~ portion is just the minute fraction of the S-flow that is actually intercepted by the electron current and transported to the impedances, for releasing the excess ~ in the loads and losses. [Incidentally, this wee note resolves a century of controversy over exactly how EM energy flows in circuits. The orthodox literature abounds with discussions, differing positions, and debates and with errors.]

[15] Note that this also automatically involves a host of other types of controversial displacement currents, such as d~/dt, dV/dt, dE/dt, dP/dt, dD/dt, etc. The literature is thoroughly floundering as to exactly how all these currents are related, together with the Slepian vector J~ and the Poynting vector S, as well as the "normal" current dq/dt. We will be discussing the "master key" for these relationships in future articles.

Tables:

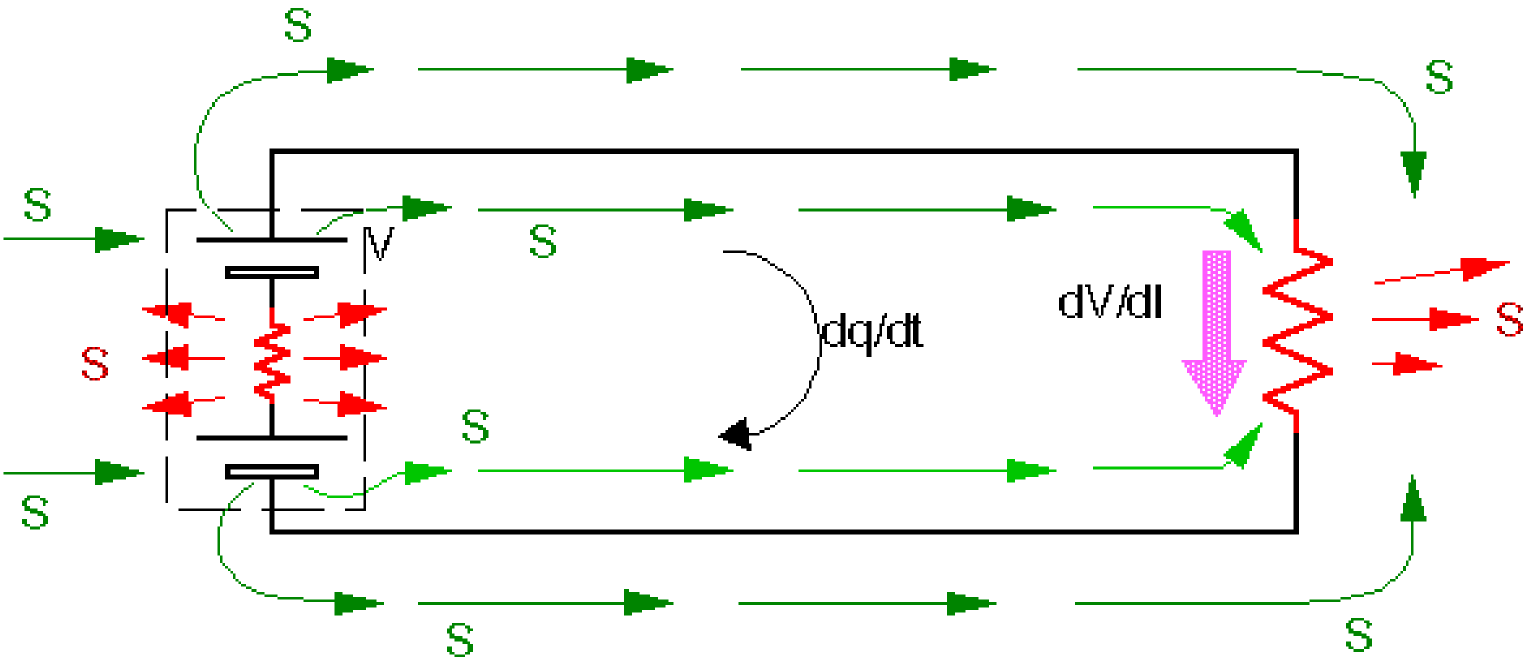

Note: EM texts err in showing the battery "gate" as the source of energy.

Figure 11. Poynting S-flow along a simple circuit.

From particle physics, any electric charge is a broken symmetry in the vacuum's virtual particle flux. Any dipole is thus two locally broken symmetries. Thus the dipole acts as a gate/source, driven by the fierce vacuum flux. So a dipole already "extracts" and gates EM energy flow, directly from the seething vacuum.

Every battery and generator is already a "free energy" source of extracted and gated EM energy from the vacuum.

D-1-F11

© T.E. Bearden 1994

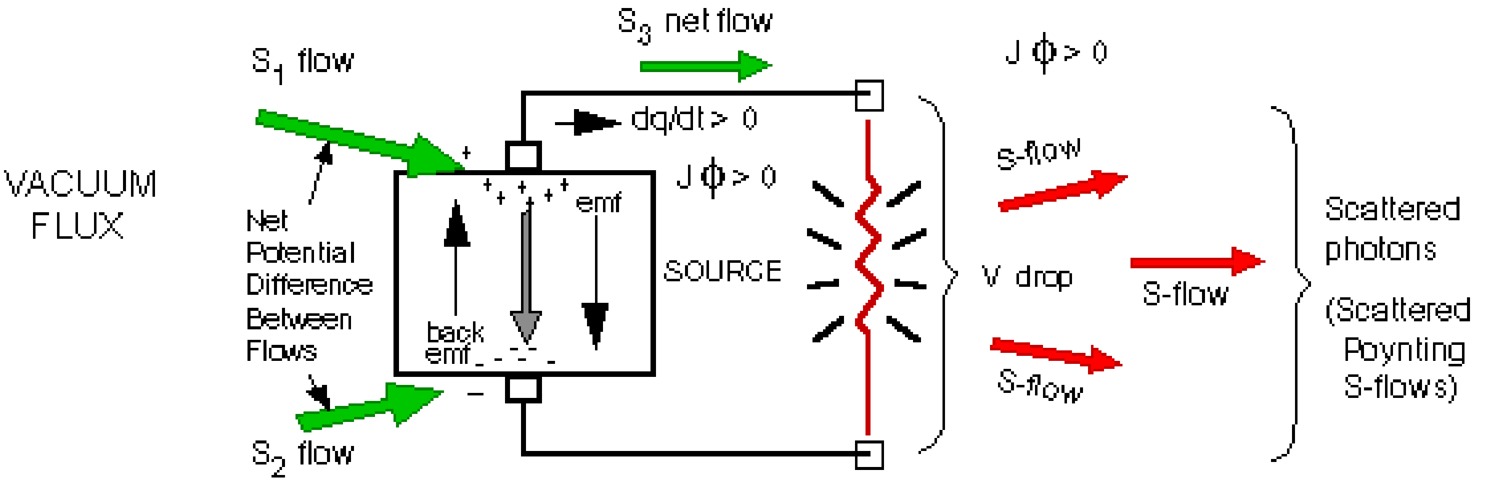
S3 net flow
S1 flow
J ϕ > 0
dq/dt > 0
J ϕ > 0
VACUUM FLUX
+
emf
Net Potential Difference Between Flows
SOURCE
S-flow
V drop
S-flow
Scattered photons
(Scattered Poynting S-flows)
back emf
S-flow
−
S2 flow

Figure 12a. Normal current flow; Jϕ is the electron transport of energy, while S is the speed-of-light Poynting energy flow.

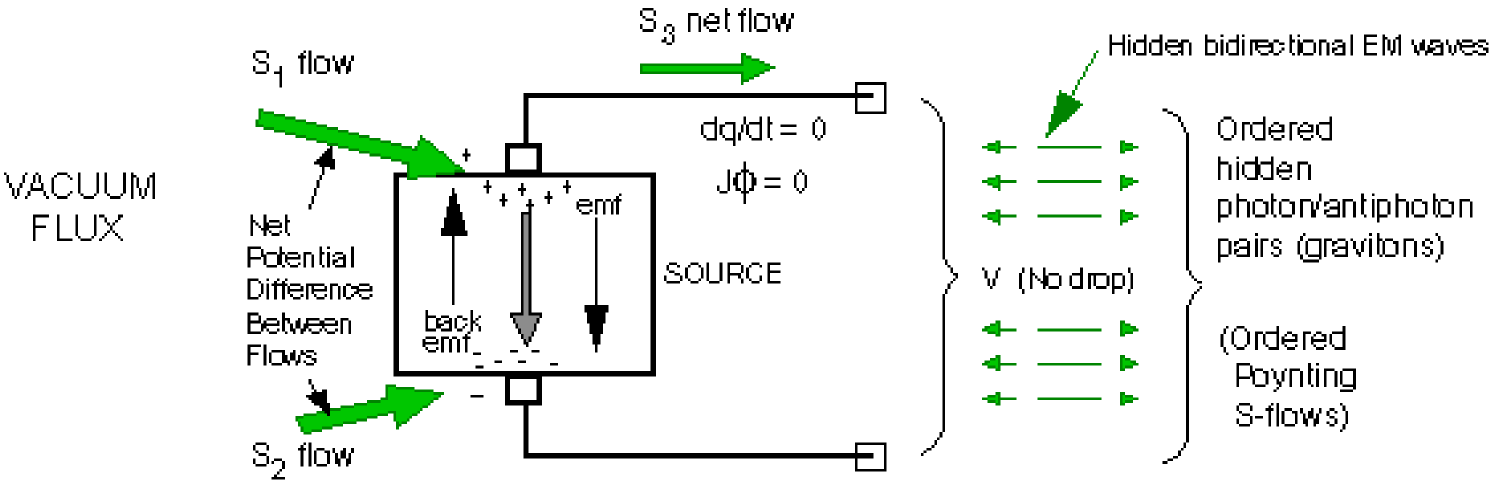
S3 net flow
S1 flow
Hidden bidirectional EM waves
dq/dt = 0
Jϕ = 0
VACUUM FLUX
+
emf
Ordered hidden photon/antiphoton pairs (gravitons)
Net Potential Difference Between Flows
SOURCE
V (No drop)
back emf
(Ordered Poynting S-flows)
−
S2 flow

Figure 12b. In a conductor with dq/dt blocked; S-flow and hidden Whittaker-wave flow continue.

Figure 12. Overt and covert field energy density flows.

D-1-F12

© 1994, 1995 T.E. Bearden

Normal sources do not furnish electrons and current to a circuit anyway. Sources just furnish Poynting flow S and EMF. [note 16] In the receiving current loop, the EMF appears automatically once the S-flow flows in and is "locked on to" by the receiving electrons. Given q**, you will immediately have J* if the electrons are free to move in the conducting circuit (see Figure 13). Further, you have eliminated all the loss terms from the standard Poynting equation for energy losses. So all the energy flow S just flows across the SC section, without any current or Cooper pairs flowing through that SC section. The SC section has become a "bridge" which (1) strips off the Poynting field energy density flow S from the electron current dq/dt on the sourcing side of the SC/bridge section, by simply reducing the dq/dt to zero; and (2) continues to pass the S-flow across the dq/dt-blocked SC/bridge section to the other side (Figure 14). The excess S-flow (and EMF) pours into the receiving dq/dt-isolated current loop, exciting the electrons therein to produce dq/dt and J*. *Any closed current loop is self-powered, once it receives S and EMF. It furnishes its own electrons; it only requires excess energy and EMF.* [note 17]

So our approach gives room temperature superconductivity in a very straightforward manner, once you discover how to block the current dq/dt in a conductor. Blocking it in an insulator is not sufficient, because that drops the potential and stops the S-flow and the equipotential * (the EMF). However, a degenerate semiconductor such as the Fogal chip can be used, as can several other processes for blocking dq/dt in a conductor. We will discuss these in a future article.

Another advantage of this approach to room temperature superconductivity is that now one can also have permissible overunity coefficient of performance. Now the load can be placed in its own S-receiving, isolated current loop. With the sourcing current loop furnishing only S and not dq/dt, the load is still powered normally in its own closed dq/dt current loop, but none of the load current is passed back through the back EMF of the primary source in the sourcing circuit.

This principle -- that at least a substantial portion of the load current must not pass back through the primary source -- is the primary principle required for a permissible overunity electrical machine (Figure 15). A permissible overunity electrical machine is one which produces more power in the load than you have to put into the machine to run it.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[16] E.g., see Robert L. Shrader, Electronic Communication, Sixth Edn., MacMillan/McGraw-Hill, New York, NY, 1991. (1993 Imprint). Quoting from p. 6: "A source of electric energy does not increase the number of ~ree electrons in a circuit; it merely produces a concerted pressure on any aimlessly moving electrons."

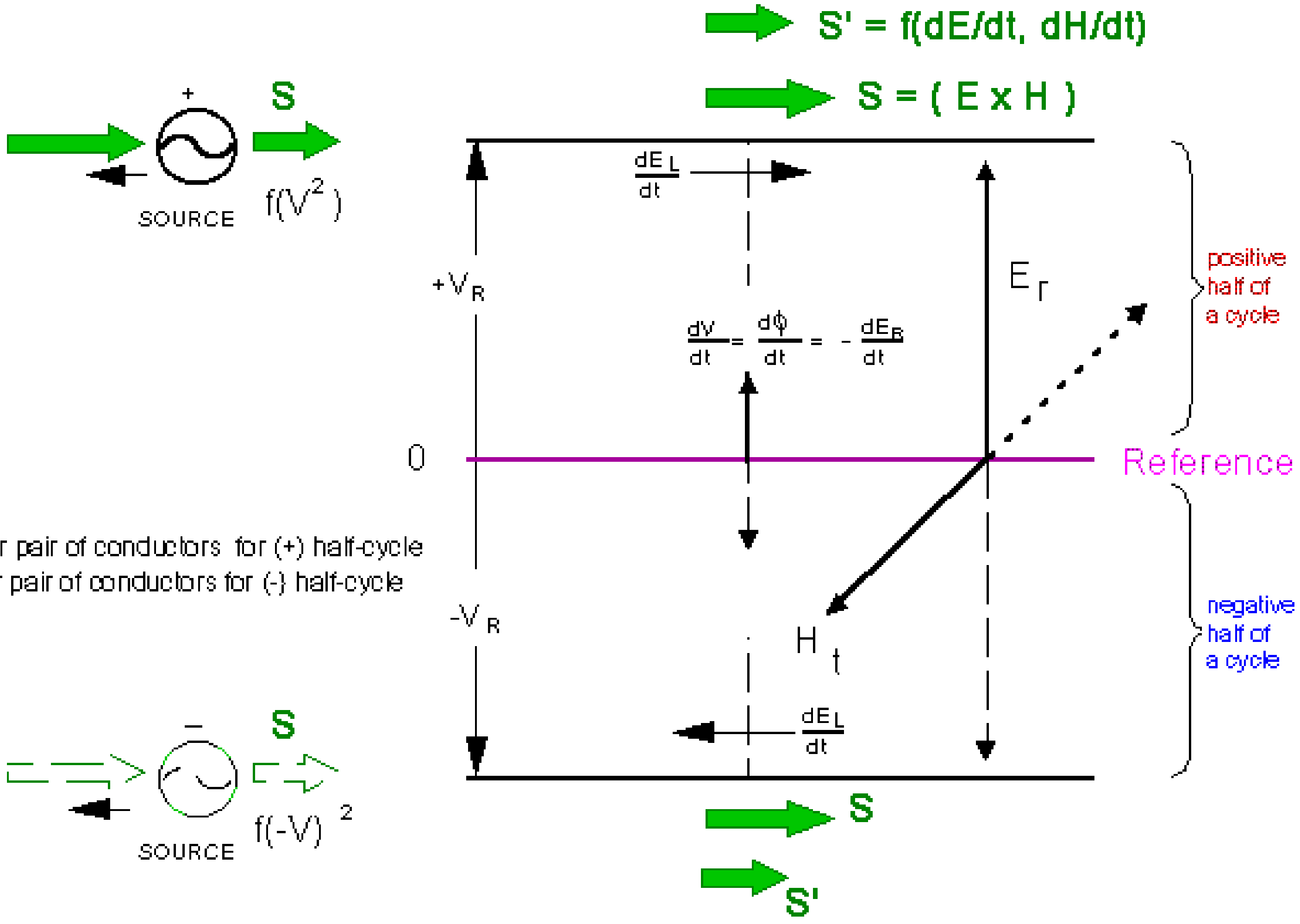
S' = f(dE/dt, dH/dt)
S = (E x H)
+
S
SOURCE
f(V²)
$\frac{dE_L}{dt}$
+V R
E r
positive half of a cycle
$\frac{dV}{dt} = \frac{d\phi}{dt} = -\frac{dE_B}{dt}$
0
Reference
-V R
H t
negative half of a cycle
$\frac{dE_L}{dt}$
−
S
SOURCE
f(-V) 2
S
S'

NOTE:
Use upper pair of conductors for (+) half-cycle
Use lower pair of conductors for (-) half-cycle

Figure 13. Poynting vector currents S and S' versus massless displacement currents.

D-1-F13

© 1995 T.E. Bearden

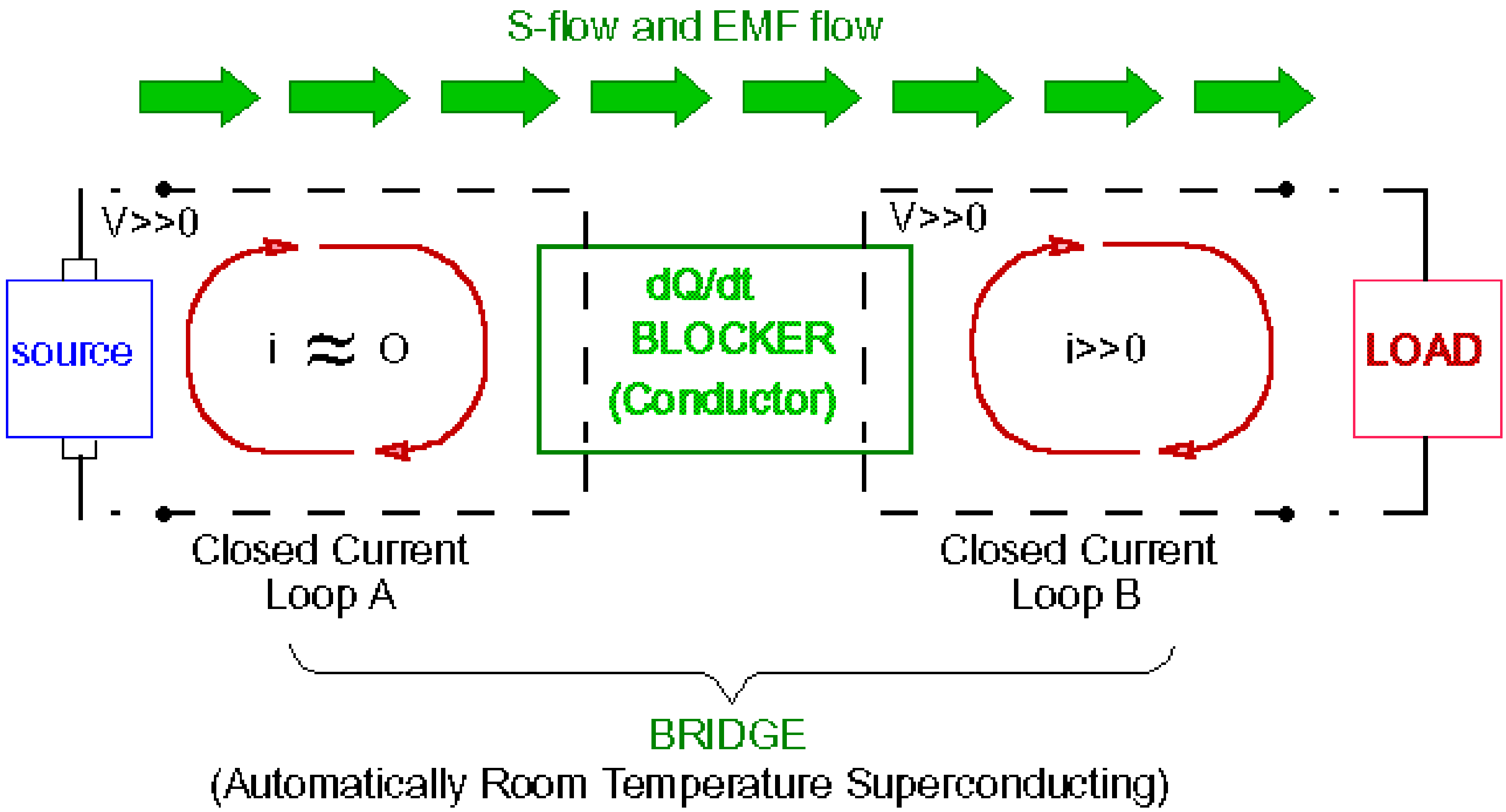
S-flow and EMF flow
V>>0
source
i ≈ 0
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD
Closed Current
Loop A
Closed Current
Loop B
BRIDGE
(Automatically Room Temperature Superconducting)

Figure 14. Bridging to enable room temperature superconductivity.

D-1-F 14

© 1995 T.E. Bearden

[17] My own approach has been, if that's all that the circuits require from our primary power source, then let's simply accept that directly from the source without performing any work in the source. Iœ you do it this way, the source is never dissipated, but will last indefinitely. Sources furnish flowing field energy density, not "energy" per se. If you only demand S-flow, you can have all you wish from a source, without exhausting your source. The trick is to find how to steadily dissipate some of this "artesian well" of flowing S to drive electrons through the load, but without driving electrons back through the back EMF of the source.

Tables:

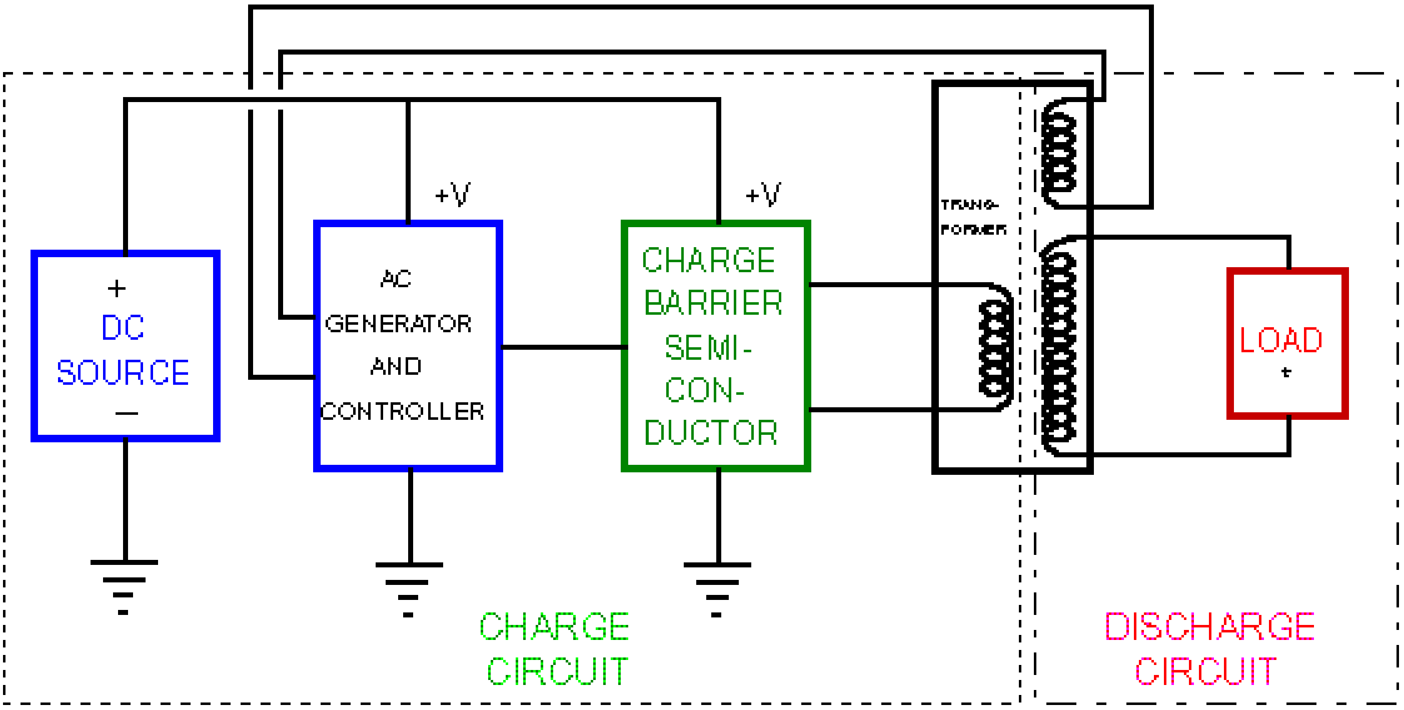
+V
+V
+
DC
SOURCE
—
AC
GENERATOR
AND
CONTROLLER
CHARGE
BARRIER
SEMI-
CONDUCTOR
TRANS-
FORMER
LOAD
*
CHARGE
CIRCUIT
DISCHARGE
CIRCUIT

Patent Pending

* NOTE:

DISCHARGE CIRCUIT AND LOAD GROUND
MUST BE ISOLATED FROM THE
CHARGE CIRCUIT GROUND.
FEEDBACK LOOP MUST BE RIGIDLY CLAMPED AND LIMITED.

Figure 15. Use of Fogal semiconductor for overunity (closed system).

D-1-F15

© 1993, 1995 T.E. Bearden

Obviously, the laws of physics and thermodynamics require such an overunity machine to be an open system that extracts and utilizes excess energy from an external source. In the case outlined, we extract the excess energy directly from the vacuum in the form of S-flow; essentially we convert our primary source to a Poynting generator rather than a "power" (energy flow loss rate) generator. The resulting overunity machine accomplishes both room temperature superconductivity and also overunity coefficient of performance (Figure 16). It does not violate the conservation laws of physics. It does not violate the second law of thermodynamics because it is an open system receiving and utilizing excess energy from the vacuum, and the second law need not apply to such systems. [note 18]

You see, all power systems are already free energy density flow sourcing systems, if we but use them correctly. Every dipole is a "gating device" for extracting excess field energy density flow from the violent bidirectional flux exchange between the vacuum and each end of the dipole. [note 19] [note 20] When we attach "conducting leads" to a dipolar source, as in the simple two-wire circuit, all we are doing with the conductors is providing waveguides for the extracted S-flows from the vacuum to flow along (it flows outside the conductors) without loss. In effect, each lead is simply extending the pole (i.e., the terminal) to which it is connected.

My associates and I are also working on several other inventions at present, including two more patent applications to be filed in the near future. One of those is in the field of overunity power systems, while the other is in an entirely different, highly proprietary nuclear effort. I will address all these areas in future columns, and those columns will have real "meat" in them.

Present "big science" proclaims that overunity electrical machines are impossible. The reader should understand that the U.S. Patent and Trademarks Office is never going to grant a patent for an overunity electrical machine unless the machine is independently tested and replicated, and unless it is certified by scientists of impeccable credentials. This is as it should be. Legitimate free energy researchers accept that dictum. Eventually the stuff has to work on the bench in order to be validated. On the other hand, free energy researchers also point out that one must not demand of us the next 100 years of national research by a hundred major laboratories immediately! This is still a shoestring operation, because all conventional sources of research funds are denied to the serious overunity researcher (with the single exception of occasional personal funding of a researcher by a private individual). In other words, we demand at least the same kind of consideration that the orthodox nuclear fusion community has enjoyed. Billions of dollars have been poured into that sinkhole, and not a single controlled fusion power unit is working anywhere in the world after decades of work at enormous expense to the taxpayers.

So we overunity researchers deserve at least a decade or two of our own, to try to accomplish overunity electrical system development. We also painfully point out that "Big Science" will not fund a single overunity researcher, for specific overunity purpose, no matter what his credentials or capability. In fact, "Big Science" is part of the problem, because it fervently seeks to prevent any legitimate research in overunity. To carry out its suppression campaign. it resorts to vilification, character assassination, libel, and slander of the overunity researchers themselves __ again, regardless of their qualifications or legitimate scientific procedures. [note 13]

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page 1

Comments may be sent to
the **Editor** of The Virtual Times

Tables:

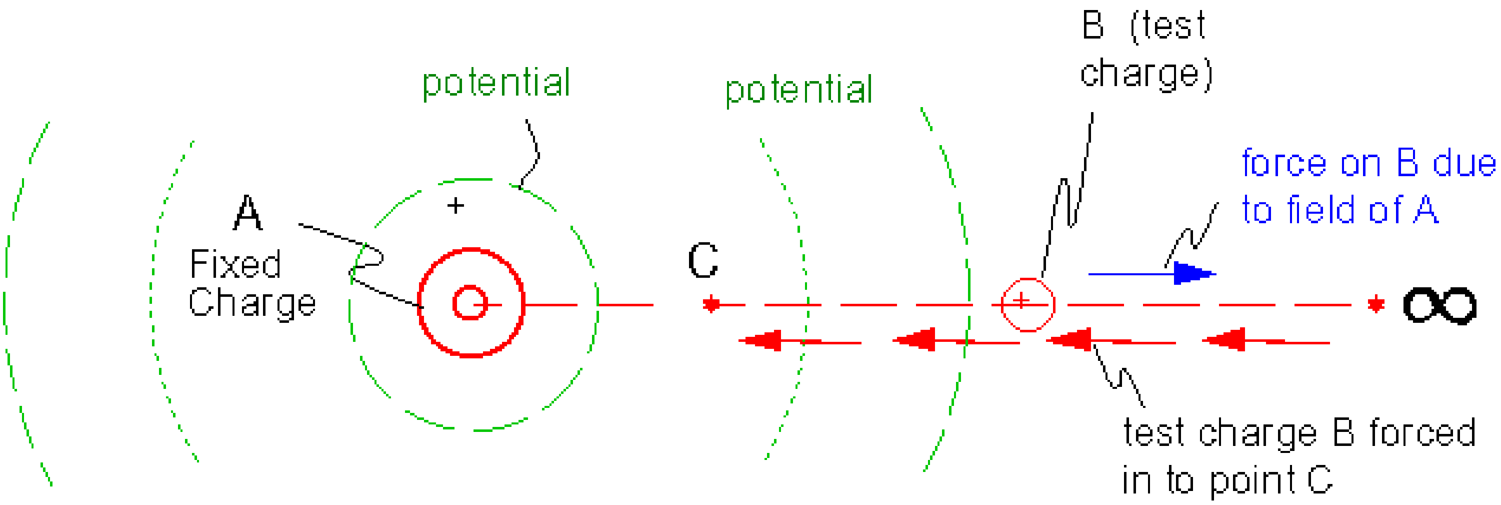
B (test
charge)
potential
potential
force on B due
to field of A
A
Fixed
Charge
+
C
+
∞
test charge B forced
in to point C

Figure 16a. Magnitude of the scalar potential is determined by work required to push in a unit charge from infinity; e.g., against a repelling like charge. If test charge is released, work will be performed upon it to translate it back toward infinity

Figure 16a. Process for overunity coefficient of performance.

D-1-F16A

© T.E. Bearden, 1993, 1994, 1995

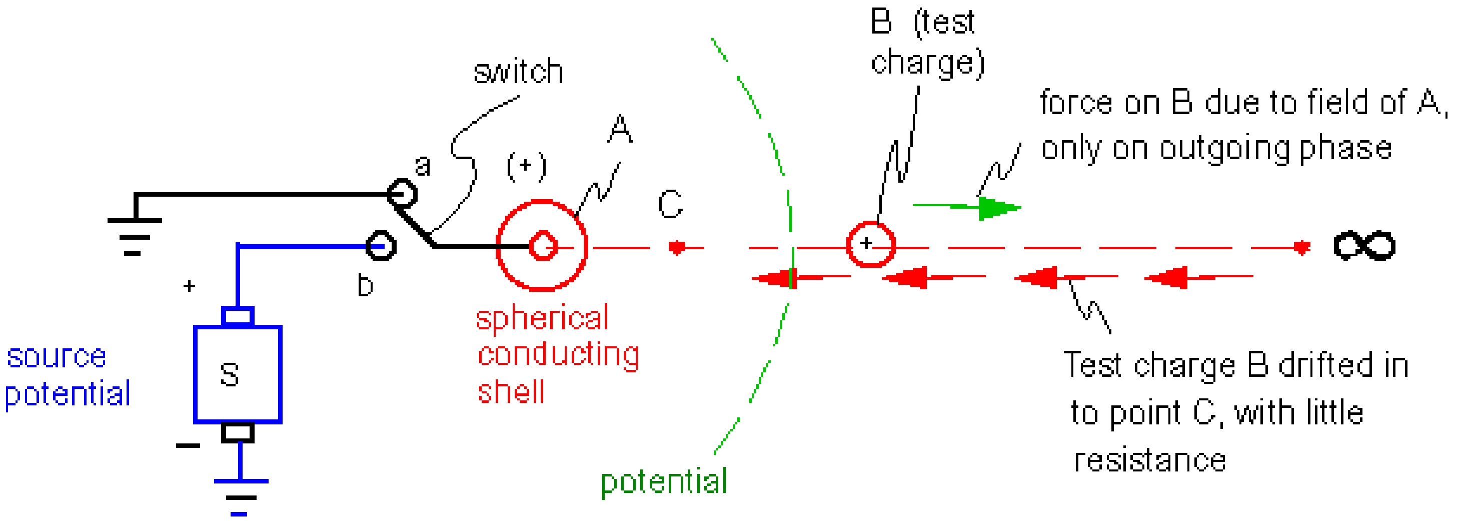
switch
B (test charge)
force on B due to field of A, only on outgoing phase
a
(+)
A
C
∞
+
b
source potential
S
−
spherical conducting shell
potential
Test charge B drifted in to point C, with little resistance

Assumptions: Capacitance of A very small; charge B large, power-free switching

Switch potential off, drift in test charge B, switch potential back on. Can now get free work from repelled test charge B. Get much more free work if A is at a high potential, and charge B is large.

Figure 16 b. Demonstration of overunity efficiency.

D-1-F16B

© T.E. Bearden, 1993, 1994, 1995

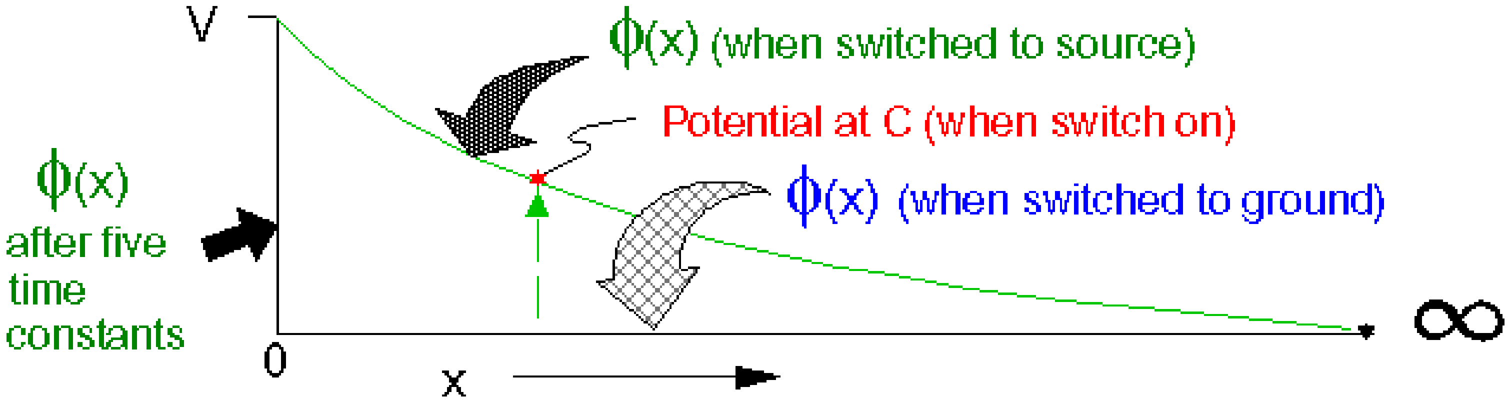

Ideally, switching the potential, rather than flowing charges, requires only minuscule overhead energy dissipation

Figure 16c. Potential to which test charge is exposed.

D-1-F16C

© T.E. Bearden, 1993, 1994, 1995

[18] The system does comply with the nonlinear thermodynamics of a system far from equilibrium. E.g., Prigogine received the Nobel Prize for significantly developing this new field. For a detailed technical presentation of nonlinear thermodynamics of a system far from thermodynamic equilibrium, see G. Nicolis and I. Prigogine, Explorinq Complexity, Piper, Munich, 1987. Such systems can exhibit negentropy, self-organization, and other effects outside the confines of classical thermodynamics. Since local symmetry is broken, they can also permissibly exhibit overunity coefficient of performance.

---

[19] Each end 0œ a dipole has its own potential (with respect to the local vacuum), which differs from the potential of the surrounding ambient vacuum. For each of these "end-potentials," one can mathematically decompose that end-potential into a hidden bidirectional set of harmonic wavepairs in harmonic sequence. Each wavepair consists of the wave and its phase conjugate. For the proof, see E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annglen, Vol. 57, 1903, p. 333-355. Since one of the dipole potentials exceeds the local vacuum potential and the other dipole potential is lower than the local vacuum potential, the two bidirectional EM wave flows are at or "carry" different equipotentials. This is the generatrix for the automatic dual œ1ow of energy from the vacuum through the dipolar power source and out along the two leads of the "transmission line" conducting the S-flow to the components of the circuit for powering the loads. So as can be seen, rigorously every dipole is already a "free energy source" if we learn how to properly use it. The energy will flow forever, if we do not allow the destruction of the sourcing dipole.

---

[20] See also E.T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proceedinqs of the London Mathematical Society, Series 2, Vol. 1, 1904, p. 367-372. Here Whittaker shows that the fields of classical electromagnetics can be replaced by scalar potential interferometry of two potentials. Our own comment is to note that by Whittaker 1903, we can with a little difficulty make artificial potentials with deliberate substructures. We can then interfere those potentials (actually, interfere their hidden multiwave structures) even at a distance, to create new EM fields. Or we can dissolve EM fields that exist at a distance, simply by constructing the correct negative EM fields by distant scalar interferometry. My point is that, when we utilize the "inner" electromagnetics inside the potential, we are employing a vastly more powerful electromagnetics of a new kind, which includes and extends the present inadequate EM theory.

Indeed, orthodoxy has a black record with respect to energy. When Mayer advanced the conservation of energy law, orthodoxy hounded him unmercifully. He was universally ridiculed as the very epitome of a fool. He lost his job and suffered a nervous breakdown. Then years later toward the end of his life, scientists came to accept the conservation of energy as a most useful tool that dramatically simplified much of their analyses. So then science began to laud Mayer. There are a thousand other such examples; suffice it to say that "big science" is a bureaucracy and a "bell-shaped distribution curve" just like any other large group of people. Some scientists are near-angels. Others are near-devils. The vast majority are neither, but are just ordinary persons doing a special kind of job. In the scientific bureaucracy, however, the manipulative scientists scurry, slash, and manipulate their way toward the top. So the leadership of any big bureaucracy __ including the scientific bureaucracy __ is always rife with such near-scalawags and power-hungry individuals. Scientists are not exceptions to the same human weaknesses that we all evidence. If you really want to see jealousy, backbiting, and fierce "back-room dealing" and power-struggling, just get into a university research environment! But because the "power merchants" of any bureaucracy always fiercely resist any challenge to their position or superiority, big science has always fiercely resisted anything truly revolutionary. And today they also enlist the power of the state to enforce their dictums. We will devote one of our future columns to pointing out some glaring examples of persecution of independent researchers by big science. It is an area that all unorthodox researchers should be well-aware of. The world of science does not run by sweet reason and lofty ideals; that is just the dogma. It runs strictly by "who's going to be the big monkey," just like everything else. Primate dominance is still a dictum of the functioning of all bureaucratic human systems.

The free energy researchers are not working on the so-called "perpetual motion machine," where a closed system is erroneously stated to produce more output than must be put into it to operate it. That old saw is actually a red herring that dogmatic scientists use to vilify free energy researchers. It is a lie that they have also successfully sold to the U.S. Patent and Trademarks Office. Anyone with a modicum of knowledge of modern physics already knows that there does not exist a single closed system in the universe, anywhere. In particular, every physical system is continuously "open" to the violent exchange of energy with the surrounding vacuum. It's just that most systems are in local equilibrium in these flow exchanges, except for minuscule gating accomplished for such things as the Lamb shift, etc.

By analogy it's something like this. The free energy researcher is standing by the banks of a mighty, rushing river with breathtaking falls, turbulent areas, etc. Big Science is begrudgingly admitting that, yes, there is a river there, but it is an insane delusion to think that you can tap that energy. The free energy researcher is saying that, well, if I can build a gating sluice a little distance upstream, perhaps I can divert a small portion of the river's flow downstream to my waterwheel, and power my mill. The orthodox scientist then vilifies the free energy researcher for even having such a heretical thought. In fact, says the orthodox researcher, any fool can see that the laws of nature force the river's flow to stay firmly in its banks, because it is a closed system, and Moses brought that law down off the mountain with him on a special stone tablet. Then he adds all the other smug remarks such as, "If it could be done, then we at MIT and Harvard and Caltech would already have done it!" So until the orthodox scientist develops a little less dogma and a little more common sense, and learns the difference between a scientific discussion and a dogfight, the free energy researcher can hardly communicate with him.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

The goal of the legitimate free energy researcher is to find a way to break the local equilibrium of this universal energy exchange, and gate a tiny bit of it up into the external circuit as excess electrical energy (Figure 17). Then he must collect the excess energy, and shuttle it from the "gating and collecting circuit" over to a dq/dt-closed current loop containing the load. He must also prevent the load current, or most of it, from passing back through the primary source against its back EMF, because if it does that, he will destroy the separation of charges in the dipolar source, killing the dipole and extinguishing the "excess energy gating." In other words he must find a way to process and transport the energy flow across dq/dt-blocked bridges between the source's closed dq/dt-current loop and the load's closed dq/dt-current loop. Ruthlessly it is an energy transport S-flow problem, not a J* energy transport problem. If he uses J* in his sourcing current loop, he automatically drives the current dq/dt back through the back EMF of his source and kills it. If he does not drive dq/dt back through the source, the source will furnish S-flow for an indefinite period! Every battery and generator we have ever built is already a "free energy" machine in that it already involves broken local symmetry in the vacuum energy exchange. We have just not realized how to use our power sources purely as Poynting S-flow sources.

So we must treat an electrical system as an open system with broken local symmetry (Table 3) in the vacuum flux exchange, so that the system extracts (gates) excess energy from an external source. In this case the source is the system's flux exchange with the vacuum. This approach is no more mysterious or bizarre than putting a waterwheel in a river, or a windmill in a wind, or a bank of solar cells out in the sunshine. The universal "free source of energy" that the overunity researcher seeks to tap is the violent exchange of virtual photons between the electrical charges of the system and the surrounding quantum mechanical vacuum (Figure 18). In quantum physics this powerful, energetic exchange of the vacuum with electrical matter is now proven both experimentally and theoretically. It is already accepted in quantum physics that the vacuum is filled with electromagnetic energy. The researcher doesn't have to prove it anymore; he just has to find out how to properly use it. The artesian well is already there; we just have to learn how to collect and use the flowing water without dynamiting the well!

The remaining objection orthodox scientists usually raise against the notion of extracting vacuum energy has been that "thermodynamics doesn't permit it." However, Cole and Puthoff have rigorously shown that, on the contrary, the theory of thermodynamics does not prohibit the vacuum's energy being extracted and used to produce heat, light, and power. [note 22] So the validity of that final objection has now vanished.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Open system receiving excess energy from an external source

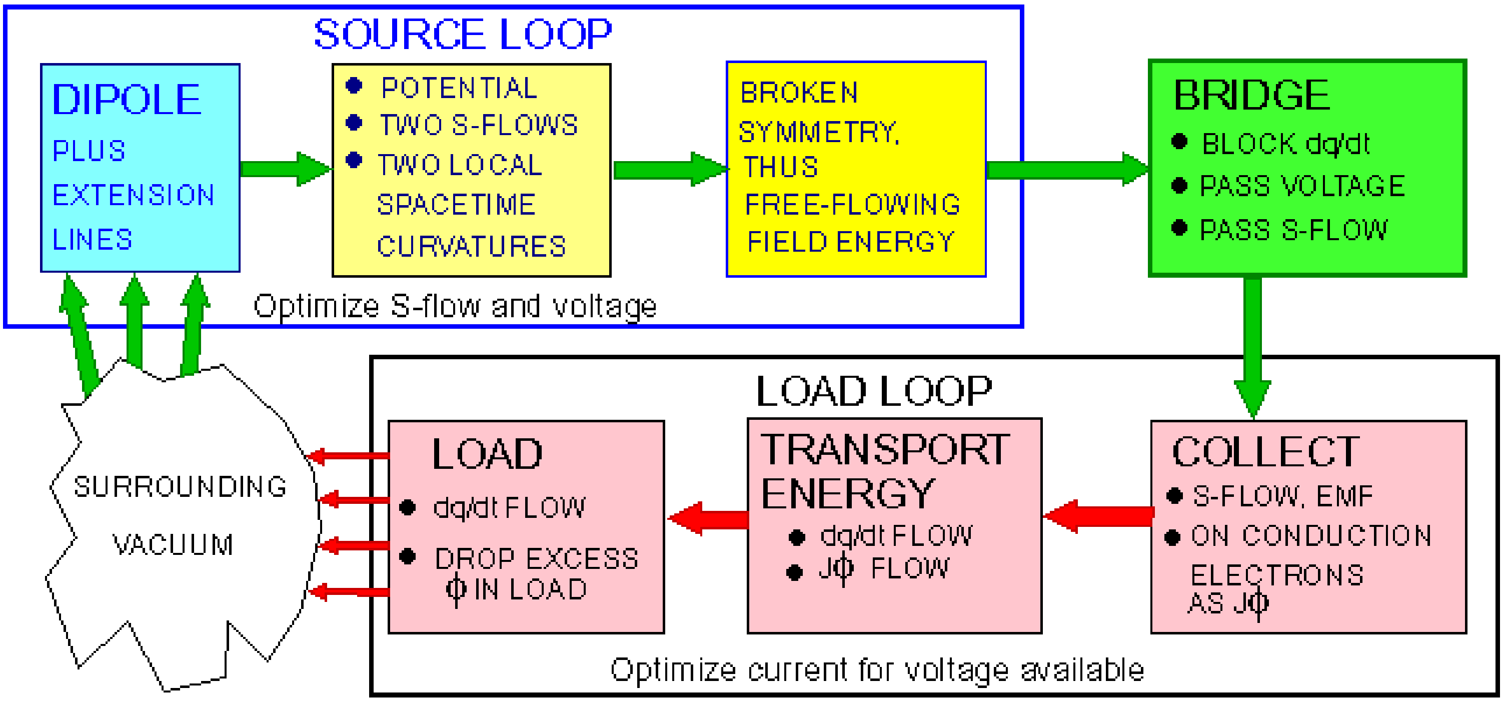
SOURCE LOOP
DIPOLE
PLUS
EXTENSION
LINES
POTENTIAL
TWO S-FLOWS
TWO LOCAL SPACETIME CURVATURES
BROKEN SYMMETRY, THUS FREE-FLOWING FIELD ENERGY
Optimize S-flow and voltage
BRIDGE
BLOCK dq/dt
PASS VOLTAGE
PASS S-FLOW
SURROUNDING
VACUUM
LOAD LOOP
LOAD
dq/dt FLOW
DROP EXCESS ϕ IN LOAD
TRANSPORT ENERGY
dq/dt FLOW
Jϕ FLOW
COLLECT
S-FLOW, EMF
ON CONDUCTION ELECTRONS AS Jϕ
Optimize current for voltage available

Figure 17. Operation of an overunity electrical power system.

D-1-F 17

© 1992 T.E. Bearden

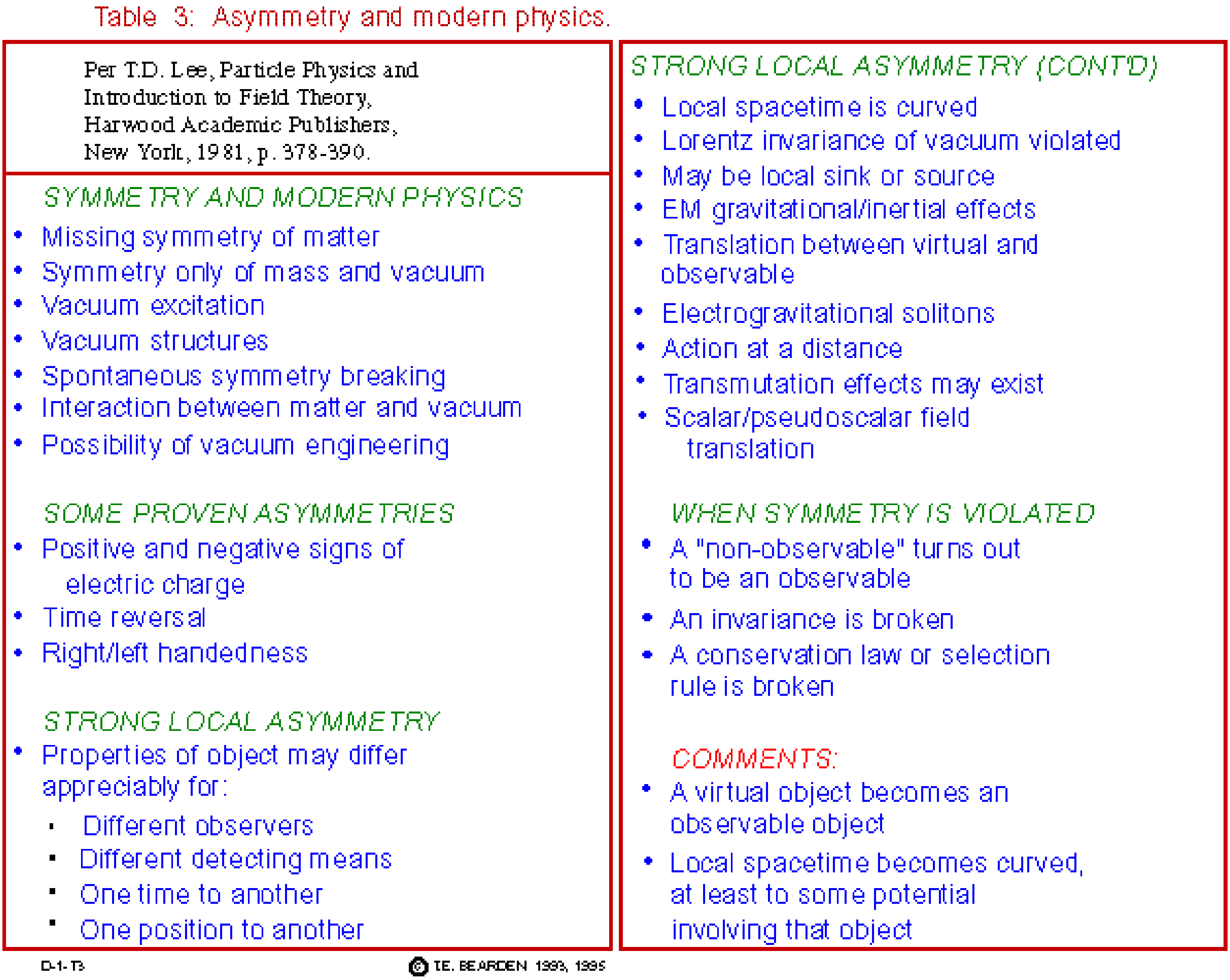

Table 3: Asymmetry and modern physics.

Per T.D. Lee, Particle Physics and
Introduction to Field Theory,
Harwood Academic Publishers,
New York, 1981, p. 378-390.

SYMMETRY AND MODERN PHYSICS

- Missing symmetry of matter
- Symmetry only of mass and vacuum
- Vacuum excitation
- Vacuum structures
- Spontaneous symmetry breaking
- Interaction between matter and vacuum
- Possibility of vacuum engineering

SOME PROVEN ASYMMETRIES

- Positive and negative signs of electric charge
- Time reversal
- Right/left handedness

STRONG LOCAL ASYMMETRY

- Properties of object may differ appreciably for:
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

STRONG LOCAL ASYMMETRY (CONT'D)

- Local spacetime is curved
- Lorentz invariance of vacuum violated
- May be local sink or source
- EM gravitational/inertial effects
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation

WHEN SYMMETRY IS VIOLATED

- A "non-observable" turns out to be an observable
- An invariance is broken
- A conservation law or selection rule is broken

COMMENTS:

- A virtual object becomes an observable object
- Local spacetime becomes curved, at least to some potential involving that object

D-1-T3

© T.E. BEARDEN 1993, 1995

A Charged Particle Is a Coupled System

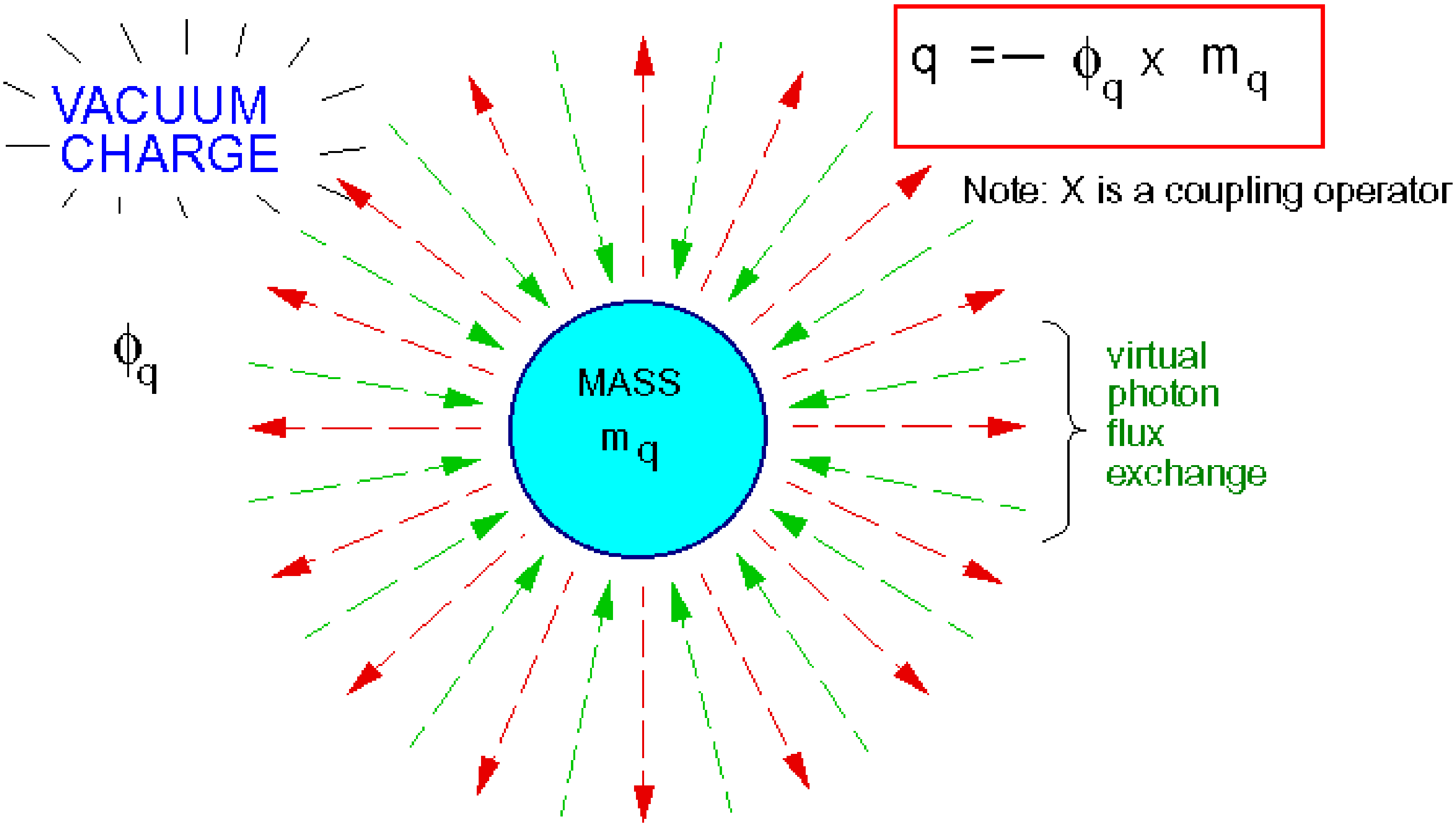
VACUUM
CHARGE
$q = - \phi_q \times m_q$
Note: X is a coupling operator
ϕ_q
MASS
m_q
virtual
photon
flux
exchange

Figure 18. An electric charge q is a broken symmetry. It continuously and violently exchanges energy with the vacuum.

D-1-F 18

© 1994 T.E. Bearden

[22] Cole, Daniel C. and Harold E. Putboll. (1993) "Extracting Energy and Heat from the Vacuum," Physical Review E, 48(2), Aug. 1993, p. 1562-1565.

And there is mind-numbing energy density there in the vacuum, waiting to be extracted. The well-known Lamb shift, e.g., in a single hydrogen atom is due to that vacuum exchange. [note 23] Interestingly, the calculated energy density of the exchange causing the Lamb shift is greater than the surface energy density of the sun. Different physicists have made various theoretical estimates of the energy density of this seething vacuum. A conservative ballpark figure is something on the order of 1090 grams per cubic centimeter, expressed in mass units.

It has also been my good fortune to know and work closely with several other free energy researchers. I was a colleague of Floyd ("Sparky") Sweet (recently deceased) for about seven years; Sparky held a Master's degree in EE from MIT and had years of experience with General Electric and other firms. He was a highly skilled researcher who invented the solid state vacuum triode amplifier (Figure 19), a unit which produced 500 watts of output and a coefficient of performance of 1,500,000. I personally tested the machine, activated the magnets on one occasion, and understand how it all worked. [note 24] The paper [note 25] I co-authored in 1991 with Floyd is still the only paper in the orthodox engineering literature that reports (a) successful overunity functioning of a solid state magnetic device, and (b) a successful antigravity experiment on the laboratory bench, reducing the weight of an object by 90%. [note 26]

We stress again that a legitimate overunity electrical machine must be an open system, receiving excess electrical energy from the surrounding vacuum. As such, it must operate in a higher topology than normal electrical machines (Figure 20).

My good friend Frank Golden has long been a close colleague. Years ago Frank developed a motor that produced overunity efficiency, but then to our astonishment we found (courtesy of Bill Tiller, [note 27] with whom we were working at the time) that its operation depended upon a 5-year-long conditioning (structuring) of the local ambient potential in order to work. Previously we had been entirely unaware that one can sometimes condition the local vacuum and the local matter in the area with a particular dynamic form. Given the local conditioning, a motor with that form will then work in that vicinity, but not necessarily elsewhere at an appreciable distance away. We now understand the mechanism for such "conditioning of the vacuum." We also now know what to do if an overunity machine is actually a machine of that type (many of them are not). In other words, we know how to convert it to a machine that will work anywhere, anytime. We will be addressing such little-known things in future articles.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

[23] Lamb, W.E. Jr. and R.C. Retherford. "Fine structure of the hydrogen atom by a microwave method." Physical Review, Vol. 72, 1947, p. 241-243. Lamb received the Nobel prize for this work.

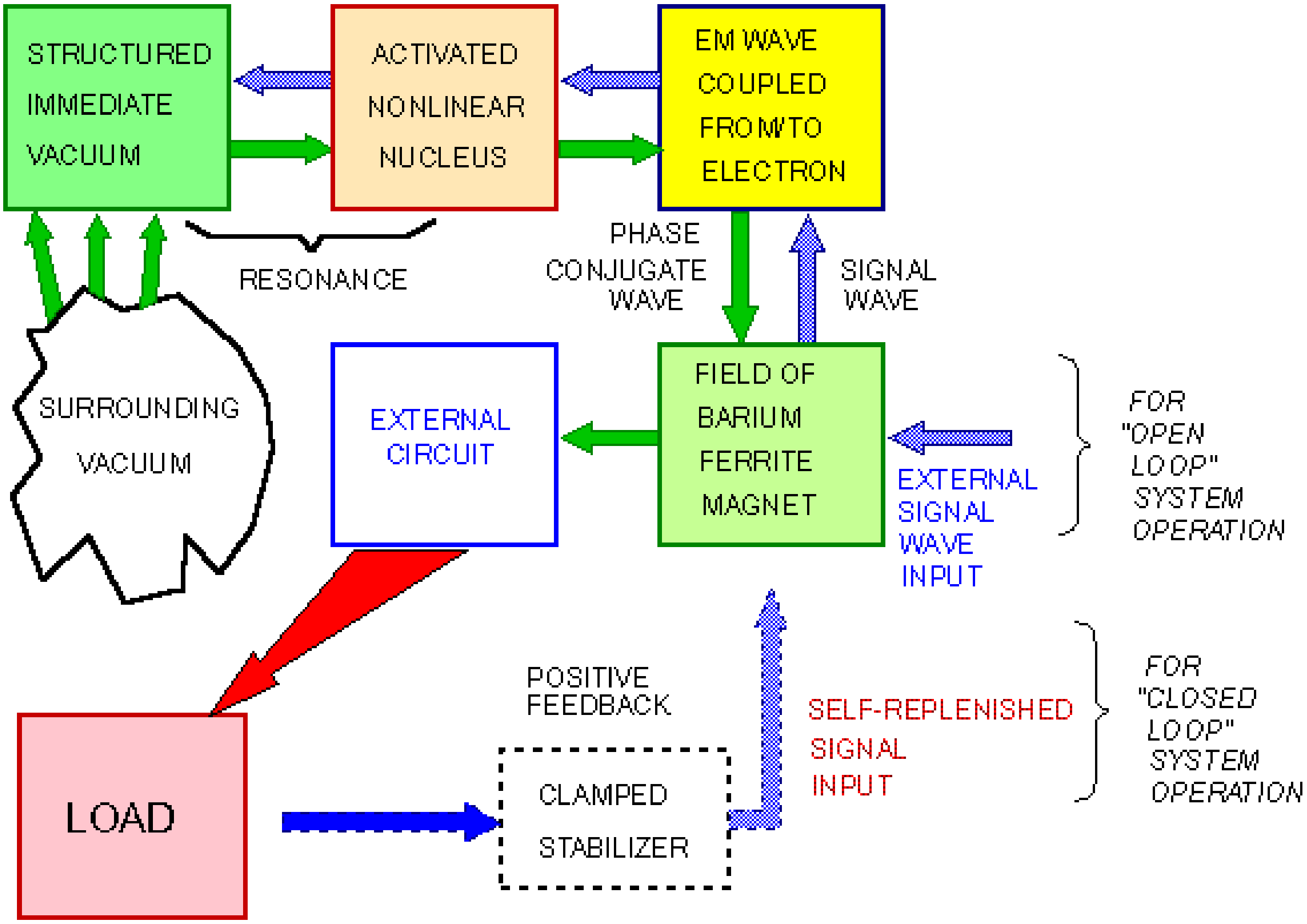
STRUCTURED
IMMEDIATE
VACUUM
ACTIVATED
NONLINEAR
NUCLEUS
EM WAVE
COUPLED
FROM/TO
ELECTRON
RESONANCE
PHASE
CONJUGATE
WAVE
SIGNAL
WAVE
SURROUNDING
VACUUM
EXTERNAL
CIRCUIT
FIELD OF
BARIUM
FERRITE
MAGNET
EXTERNAL
SIGNAL
WAVE
INPUT
FOR
"OPEN
LOOP"
SYSTEM
OPERATION
POSITIVE
FEEDBACK
LOAD
CLAMPED
STABILIZER
SELF-REPLENISHED
SIGNAL
INPUT
FOR
"CLOSED
LOOP"
SYSTEM
OPERATION

Figure 19. Block diagram of the Sweet Vacuum Triode Amplifier.

Adding positive feedback circuit enables self-oscillation.

D-1-F 19

© T.E. Bearden 1991, 1995

[24] Unfortunately Sweet made conflicting agreements with multiple backers, resulting in a most thorough legal foul- up. It appears that nothing can ever be done with this machine, because its status is a legal nightmare.

[25] Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," Proceedings of the 26th Intersociety Energy Conversion Engineerinq Conference (IECEC '91), Boston, Massachusetts, 1991, p. 370-375. Paper also distributed over Internet.

[26] The unit was specially rigged and "pushed" to 1,000 watts output to perform the antigravity experiment.

Example: Conservation in 2D versus 3D topologies.

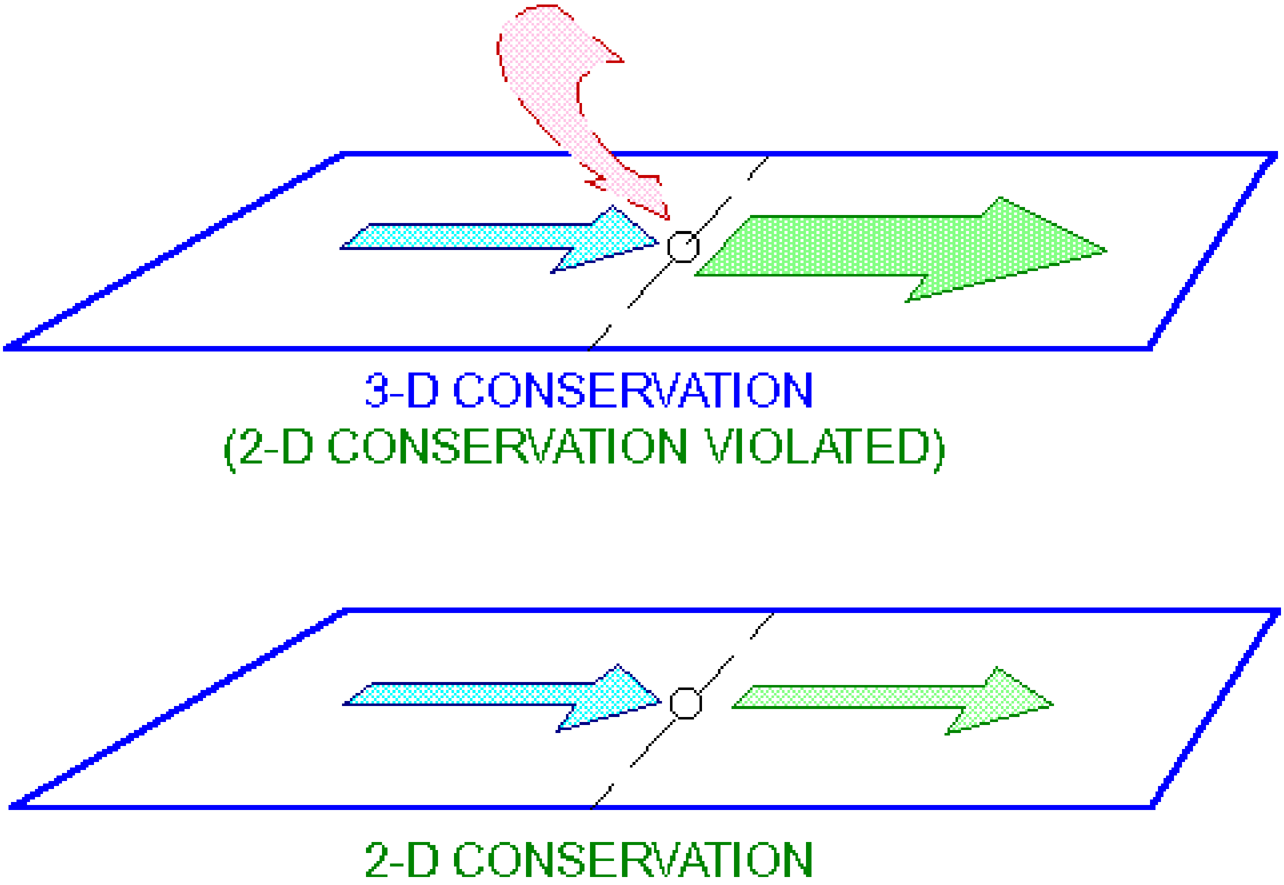
3-D CONSERVATION
(2-D CONSERVATION VIOLATED)
2-D CONSERVATION

Figure 20. Increasing the dynamic topology allows lower topology conservation laws to be violated. Conservation now applies in the higher topology, but not necessarily in the lower. The lower topology system is now an open system and may seemingly act as a source or sink.

D-1-20

© T.E. Bearden 1993, 1995

[27] Tiller was already familiar with the fact that devices that function by more subtle energy often require special conditioning of the vacuum spacetime. He expressed it as "an archetype must first be created for the machine to function efficiently." Tiller had previously experimented with such an effect for several years.

John Bedini is also a close friend and colleague; John built several experimental motors (both electrical and magnetic) in the overunity area, and performed successful trans-mutation experiments. John is a recognized genius in high-end sound amplifier development. Many audiophiles worldwide still swear that the Bedini amplifier is the best and sweetest-sounding audio amplifier ever built. One of John's battery-powered electrical motors (Figure 21), for example, ran continuously off its battery for about five years, and kept the battery charged. When you realize that such a small electric motor is only about 35% efficient, then you realize that about 65% of the energy flowing out of the battery was being dissipated in the motor as heat, core losses, etc. So the unit was continuously performing work for that five years. [note 28] The 1/8 hp motor represented a load in which the continuous rate of work being done (the rate of energy dissipation) was about 0.08 hp. We will have more to say on this motor and its technical process in a future article.

Bill Fogal is a close colleague; Bill has patented the world's first dq/dt-blocking semiconductor, [note 29] which partially blocks the normal current while continuing to pass the flow of voltage (Table 4). I have written a proprietary technical explanation of the Fogal semiconductor, which utilizes an extremely little-known feature called the "overpotential." The overpotential part of the theory is experimentally proven and theoretically explained in an obscure part of electrochemistry, by a few chemists who specialize in electrode effects. J. O'M. Bockris is probably the leading world expert in overpotential theory; he authored the textbook on it. [note 30]

One of the latest overunity developments is the magnetic resonance amplifier (MRA) developed by Joel McLain and Norm Wooten (Figure 22). Earlier testing disputed the overunity functioning. More recent testing at several well-known laboratories has resulted in the instrumental measurements clearly showing overunity. While the test institute scientists themselves may be willing to attest to overunity results of their testing, none of the corporate heads will allow it. I have personally stated that the MRA is capable of overunity. Our own measurements had to be discarded because we did not have the proper instruments available. E.g., all our oscilloscopes were "hard grounded" to building ground, and we needed to make "floating differential ground" measurements in order to settle things once and for all. McLain and Wooten have filed for a patent, and multiple tests with excellent equipment have indeed indicated overunity. What happens next remains to be seen. I suspect that the two inventors may simply accept foreign capital investment and go offshore.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Tables:

WARNING: Highly dangerous operation. Pulsed battery outgasses hydrogen gas and explosion is possible.

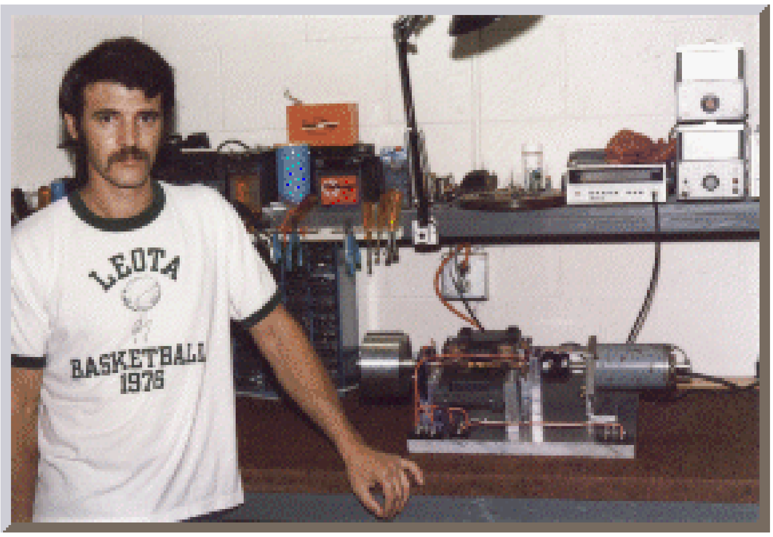

Open system using higher topology extraction of excess vacuum energy. Uses sluggish reversal and recoil time of lead ion onrushing current versus rapid backlash of amplified voltage discharge from a magneto ignition coil, to force the battery to act as an excess S-flow collector during the ion-current charging half-cycle. Thus the inefficient motor is continuously powered and the battery is continually recharged.

Figure 21. Bedini's battery-powered, back-popping electric motor.

D-1-21

© T.E. Bearden 1993, 1995

[28] We strongly emphasize that there is no conservation of work law in physics! It is the overall energy of the universe that must be conserved, not work. More rigorously, the flow of energy must be conserved. But from a free flow of energy, you can continuously extract a little and dissipate it as work, without cutting off the main flow and without violating any of the laws of physics. If this were not so, then waterwheels, windmills, and solar cells would not work. In fact, you can even do better than that. The basic energy conservation law is this: Energy can neither be created nor destroyed. So even when you "use" (i.e., dissipate or scatter) the energy, it is still there. It is only the hierarchical ordering of the photons that constitutes entropy. Each individual photon is perfectly ordered and still contains all its energy. By retroreflection, you can even cause those disordered photons to gather back in your system into perfect order once again and then you can scatter them again to do some more work. You can use the same energy over and over again! The perfect retroreflection is more easily accomplished by phase conjugate reflection. For a vivid example of just such retroreflection of previously scattered EM energy to restore the former order, see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, Vol. 254, No. 1, Jan. 1986, p. 74-83. See particularly the striking photographic demonstration of time reversal of disorder on p. 75.

[29] William Jay Fogal, "High gain, low distortion, faster switching transistor." U.S. Patent No. 5,196,809, Mar. 23, 1993. Also William J. Fogal, "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,430,413, July 4, 1995.

Table 4. Fogal's Semiconductor.

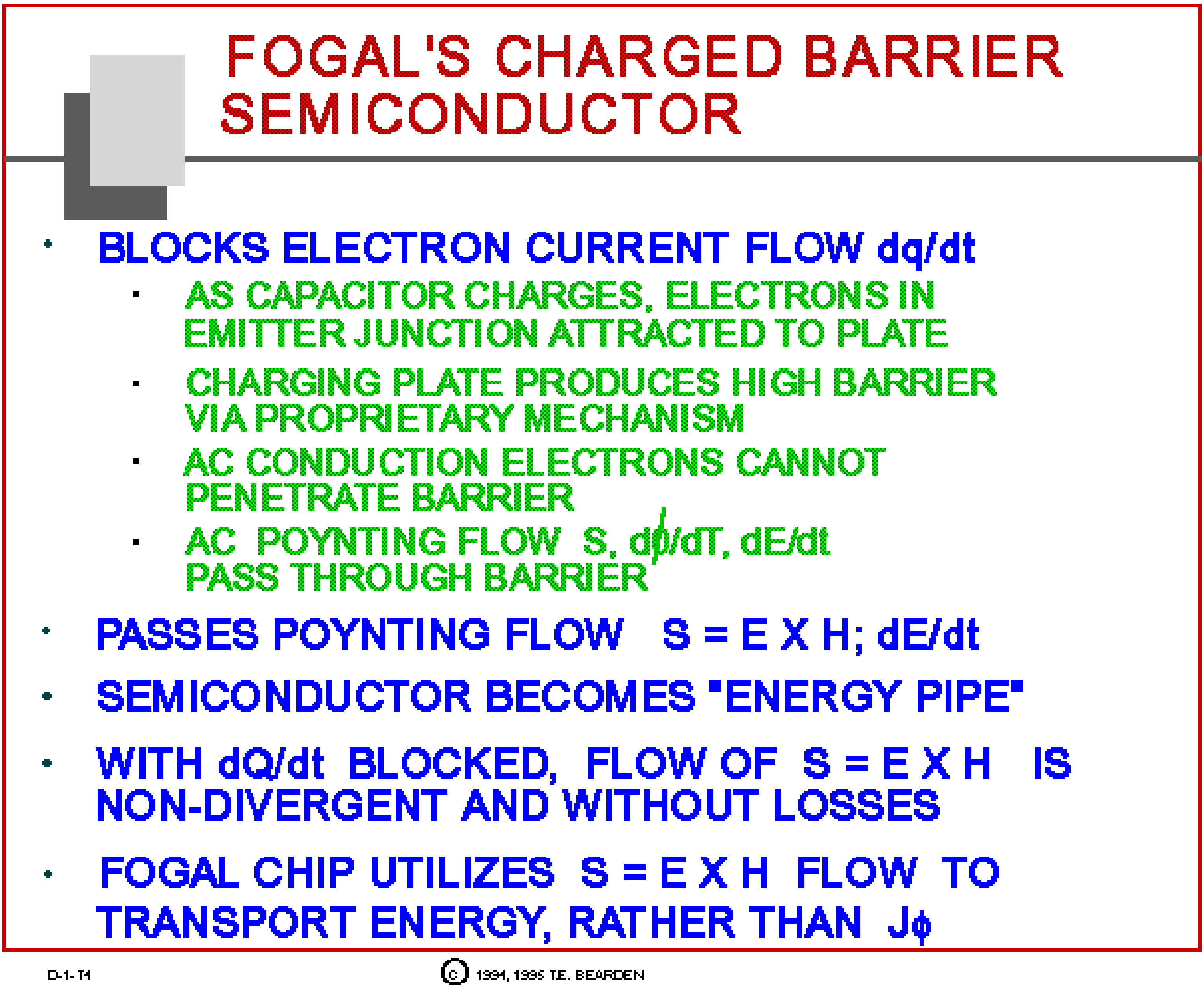

FOGAL'S CHARGED BARRIER SEMICONDUCTOR

- BLOCKS ELECTRON CURRENT FLOW dq/dt
  - AS CAPACITOR CHARGES, ELECTRONS IN EMITTER JUNCTION ATTRACTED TO PLATE
  - CHARGING PLATE PRODUCES HIGH BARRIER VIA PROPRIETARY MECHANISM
  - AC CONDUCTION ELECTRONS CANNOT PENETRATE BARRIER
  - AC POYNTING FLOW S, $d\phi/dT$, dE/dt PASS THROUGH BARRIER
- PASSES POYNTING FLOW S = E X H; dE/dt
- SEMICONDUCTOR BECOMES "ENERGY PIPE"
- WITH dQ/dt BLOCKED, FLOW OF S = E X H IS NON-DIVERGENT AND WITHOUT LOSSES
- FOGAL CHIP UTILIZES S = E X H FLOW TO TRANSPORT ENERGY, RATHER THAN $J\phi$

D-1-T4

© 1994, 1995 T.E. BEARDEN

[30] In a correspondence with Bockris lasting several years, he had graciously furnished me with material on the overpotential, which is how I found out about this esoteric area. Bockris is well-known as a leading scientist on the chemistry of hydrogen, and also as one of the leading scientists in the rapidly developing field of cold fusion. He is a marvelous scientist, and one day his largely unsung work with the overpotential will revolutionize the entire electronics industry.

Tables:

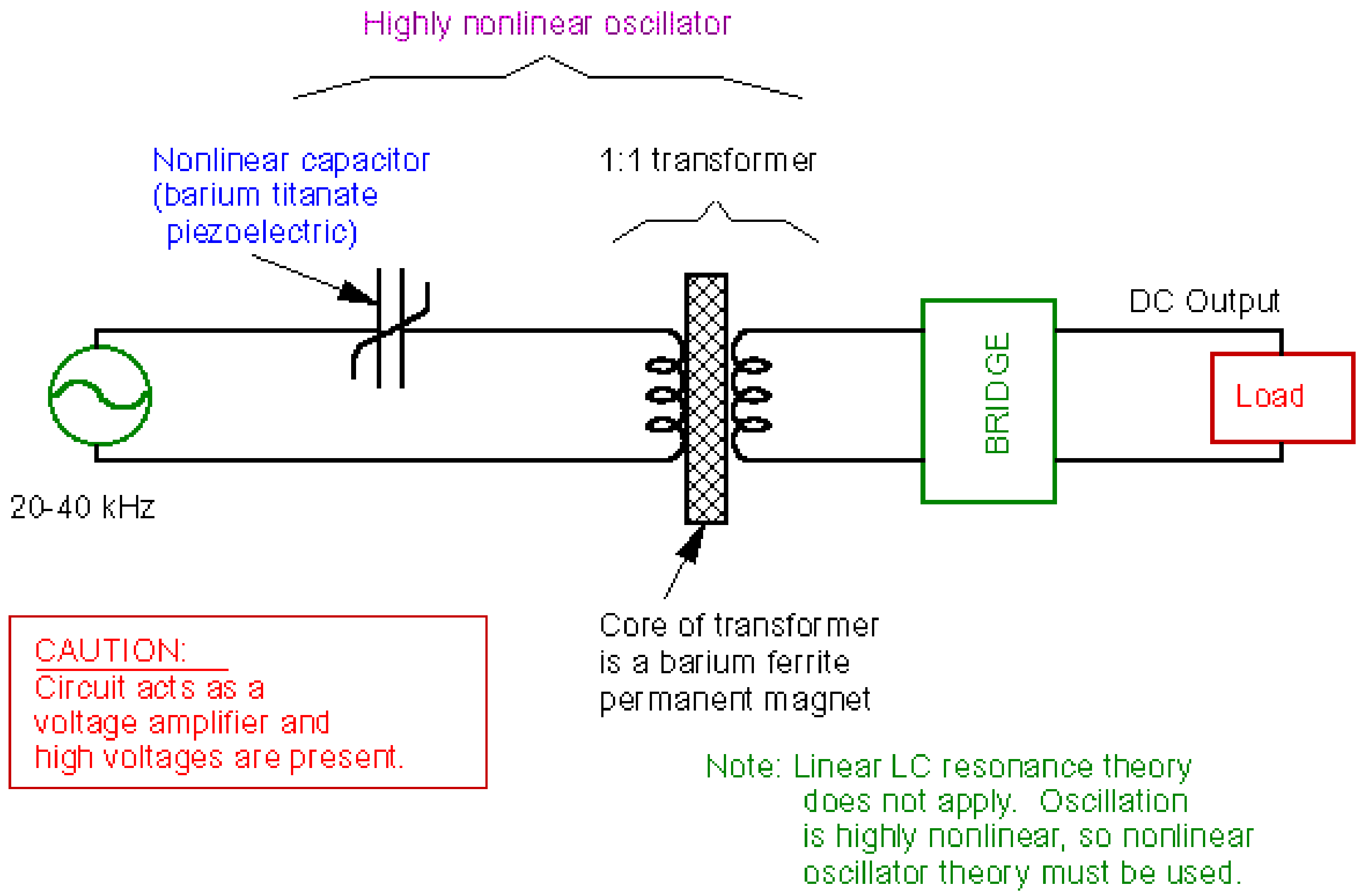
Highly nonlinear oscillator
Nonlinear capacitor (barium titanate piezoelectric)
1:1 transformer
DC Output
BRIDGE
Load
20-40 kHz
CAUTION:
Circuit acts as a voltage amplifier and high voltages are present.
Core of transformer is a barium ferrite permanent magnet
Note: Linear LC resonance theory does not apply. Oscillation is highly nonlinear, so nonlinear oscillator theory must be used.

Figure 22. Schema for the magnetic resonance amplifier (MRA).

D-1-F22

© T.E. BEARDEN 1994, 1995

Howard Johnson is also a respected colleague, whom I very much admire. (See Figure 23). Howard has continued to work quietly and patiently upon his patented permanent magnet motor, [note 31] including patenting various magnetic gates, etc. that are necessary to make such a motor work in a rotary configuration. [note 32] Howard employs a two-particle theory of magnetism; i.e., each magnetic flux line is envisioned as having a particle traveling from the north pole to the south pole, and also a particle traveling from the south pole to the north pole. The particles are spinning; the forward-time particle spins in one direction, and the antiparticle spins in the other direction. Howard then slightly separates the two particle flows. [note 33] In other words, Johnson splits the flux lines themselves, into two different pieces. When so separated, the component lines are now curls, so their paths curve. The paths of the two "curl particles" are different; one curls in one direction and the other curls in the opposite direction. Further, a predominance of one form of curl particle gives a "time-forward" aspect, while a predominance of the other form of curl particle gives a "time-reversed" aspect. Johnson is thus able to employ a deeper kind of magnetism than the textbooks presently contain. He demonstrates that a "spin-altered" magnetic assembly exhibiting (to a compass or other such detector) a north polarity can attract another unaltered magnetic assembly exhibiting a north polarity. In short, he can make a north pole attract a north pole. We will give you further insight into Johnson's two-particle theory in a future article. [note 34] We will also explain how and why the physicists missed that antiparticle in the magnetic field's flux lines, and thereby failed to advance the theory of magnetism to a deeper level. Make no mistake, one day when the new theory is done, Johnson may well be awarded a Nobel Prize for his epochal discovery of a deeper structure of magnetism.

Bits and pieces of the new science approach are just now beginning to spill into a few conventional journals and symposia. For some time a few of the rebels spearheading this new science have been doing something that the University scientists should have been doing all along. The rebels have been meticulously examining the concepts, postulates, and definitions that the present scientific models are founded upon, to reveal serious foundations errors. Corrections for some of these errors have been discovered and made. [note 35] Others are still problems yet to be resolved. Yet slow progress is being made, and a "flip-over" of the old science is likely to occur within the next decade if not sooner. By then the new concepts will have become "solid" enough to allow producing an engineering mathematical model. Also the supporting experiments will have become solid enough to justify the new concepts and the model.

ILLUSTRATION LIST | NEXT PAGE

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Tables:

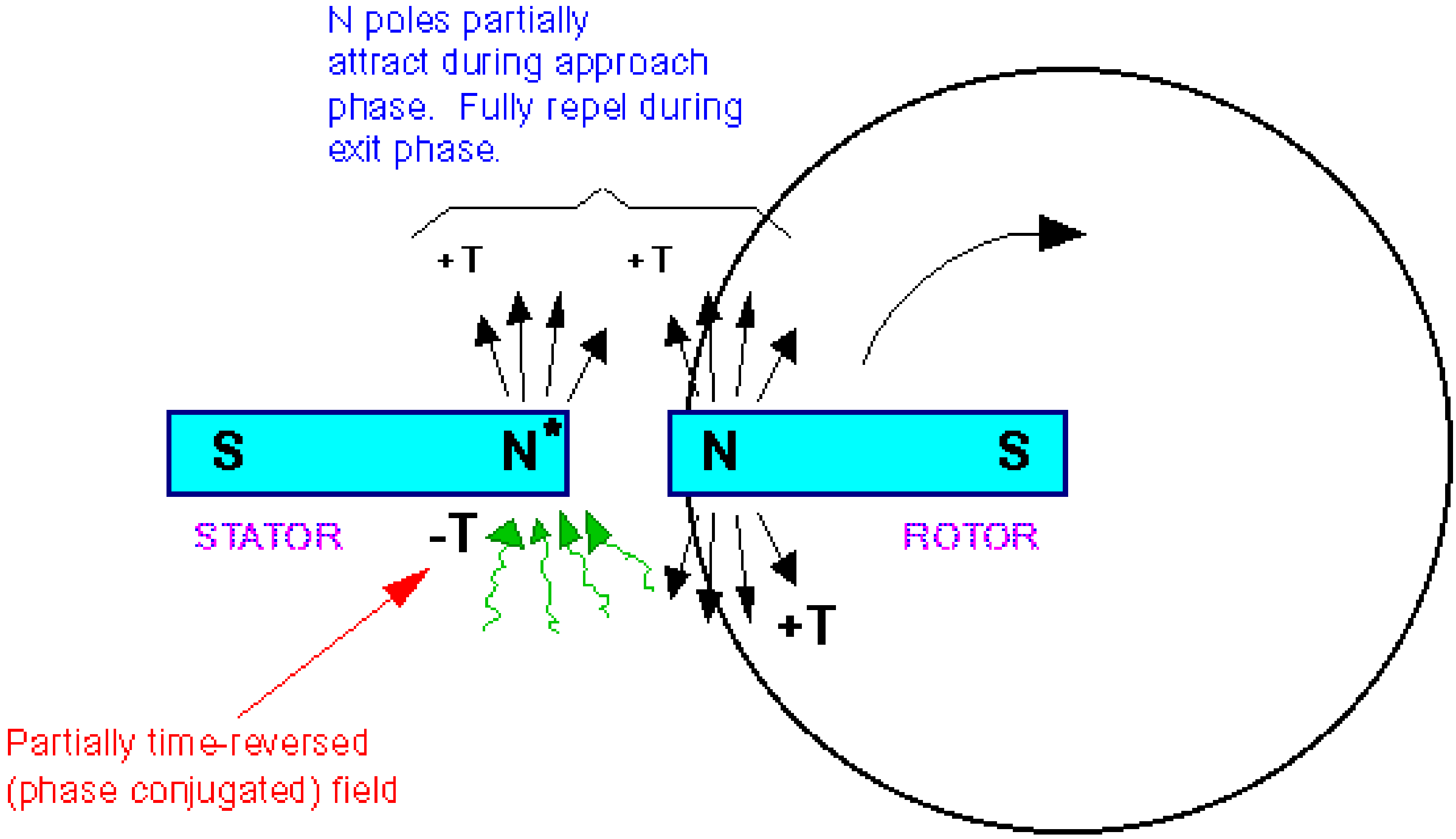
N poles partially attract during approach phase. Fully repel during exit phase.
+T
+T
S
N*
N
S
STATOR
-T
ROTOR
+T
Partially time-reversed (phase conjugated) field

Implications of a time-reversed field area
on only one side of a magnetic pole
in a permanent magnet motor

Figure 23. Schema for Johnson's permanent magnet motor.

T91H-403

© T.E. BEARDEN 1991, 1993, 1995

[31] Howard Johnson, "Permanent Magnet Motor." U.S. Patent No. 4,151,431. Apr. 24, 1979.

[[32] I personally saw and closely examined one demonstration rotary Johnson permanent magnet motor some years ago, and toyed with it for about one hour. It would definitely self- rotate as long as you wished to permit it to turn. It was not a powerful device at all, but just a small laboratory "proof of principle" prototype. It had cost Johnson an enormous amount of time, labor, and optimization to get the critical adjustment required of his two magnet assemblies. But the device had no power source other than the permanent magnet assemblies themselves. Johnson's nonlinear rotor and stator magnets interacted with each other in a manner to break local symmetry. So his machine was an open system and therefore a permissible overunity device; it was not a perpetuum mobile.

[33] As I have pointed out repeatedly in the past, photons also carry time, not just energy. We have previously shown the process and the photon interaction mechanism that creates the flow of time itself; we will discuss this mechanism in the future. So when Johnson separates the particles and antiparticles, not only does he partially separate them according to spin, but he also alters the local character of time flow during which the resulting magnetic field forces must appear. In other words, he accomplishes a partial separation of time-forward and time- reversed polar interactions. A south pole is just a time- reversed magnetic north pole, in the first place! So a north pole of a bar magnet that is slightly time reversed on one side will partially act on that side just like a south pole. On the other side it will continue to act like a normal north pole. By partially time-reversing (phase conjugating) one side of the north magnetic pole piece, Johnson makes that side look and act like a south pole. that way Howard is able to create two north poles, one on a stator and the other on a rotor, and time-reverse part of one face of the stator's north magnetic pole-piece. Therefore when the proper sides of the stator and rotor north poles are facing, they attract each other, contrary to the conventional textbook. The two poles then repel each other normally as soon as the north rotor poles passes the north stator pole. Hence Johnson can make a surrounding north pole stator assembly "draw in" an approaching north pole rotor assembly, and then kick it on out the other side, because he has broken the local magnetic symmetry. In short, Johnson's magnetic gate can provide a legitimate component of unidirectional magnetic thrust, which means that he can indeed make a rotary permanent motor. Simply put, this "partially separating the spin particles," and thereby partially phase conjugating one face of a magnet, is what Johnson calls a "gate," and this is the patented secret by which his magnet assemblies can be made self-powering. The entire process is still very meticulous, and assembly and adjustments are extremely critical. With Johnson's blessings we hope to shed more light on this subject in coming articles.

[34] For a basic article on rotary permanent magnet motors, see T.E. Bearden, "On Rotary Permanent Magnet Motors and 'Free' Energy," Raum&Zeit, 1(3), Aug.-Sep. 1989, p. 43-53.

---

[35] The most blatant error in all of physics is the erroneous notion that an acting force is separate from its receiving mass. Force is defined as F ~ d/dt(mv). From that definition alone, it can be seen that the affected mass is a component of the force itself. If you remove the mass m, you do not have force but merely acceleration. The notion of force separated from the mass came from ancient and familiar experiences of pushing on stones with one's hand, etc. In this primitive case the force-creating agent (i.e., the hand) is certainly separate from the affected mass (i.e., the stone). But the "force on the stone" is actually the force created of and with the stone. It is "stone-force," not "force upon the stone." Quantum field theory already tells us that it is the absorption and emission of virtual photons by the stone's mass that "creates the force." In fact it should go one step further: It is the coupled ensemble of the virtual photon flux and the mass of the stone that constitutes a "stone-force." In your mind, you should try mentally replacing the word "force" with the phrase "mass-force" until you root out the deeply ingrained erroneous notion that force and mass are separate. Because of this archaic error, today's physics can only define a mass in terms of its resistance to a disturbing force, and it can only define a force in terms of its overcoming a resisting mass. This error has prevented any proper definition of either force or mass! But the problem is solvable. For a proper definition of mass without the notion of force, see T.E. Bearden, Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena, National Technical Information System, Report AD-763210, 1973.

In this first article I wanted to introduce the 14 major areas. We will be discussing these areas in some technical detail in the future. We'll also include current major news items of development in these areas, for the interested lay reader. I'll particularly try to give you the benefit of extensive reference citations painfully gathered over the last 30 years. The articles will be targeted toward both the serious researcher and the interested, educated layman.

So that's the agenda for the months ahead. We hope you find this information and the approach of direct interest and use.

Again, it's a privilege to be aboard. And we invite you to hang on in the future; it's going to be a wild but exciting ride. We really are going to chase this wild dragon with great vigor.

ILLUSTRATION LIST

©copyright T.E.Bearden 1995

VT Page1

Comments may be sent to
the **Editor** of The Virtual Times

Novel Principles In The Rife Microscope

And Rife's Great Secret [1]

© 1998 T.E. Bearden

Substantially updated Aug. 20, 2001

Foreword

In the 1930s and 1940s, Royal Raymond Rife {1} made revolutionary discoveries in microscopy, including development of novel microscopes that could see far beyond the accustomed one-tenth wavelength limitation according to Rayleigh theory. With his microscope, he was able to "see" ultramicroscopic living forms much smaller than anything considered in the observable state. It would appear that his microscope was somehow capable of peering directly into the virtual state itself, and finding living forms at that level that were far beyond anything ever suspected by conventional science, even to this day.

Obviously this work was dramatically controversial. Rife went on to develop treatment methods and devices based on the new level of living organisms that his microscope revealed. In short, he arrived at essentially a new theory of disease, and did achieve success with his—to conventional scientists—completely unprecedented devices and methods of treating and healing. While we do not discuss his treatment methods in this paper, we do discuss the methods of Prioré, whom we believe to have used similar principles. So these discussions may shed some light on Rife's therapeutic mechanisms also.

Rife and others reported his results in the proper scientific literature. For his efforts, he was ostracized, essentially imprisoned in a medical treatment facility, broken, and rejected by his peers. His findings were harshly ridiculed and discredited. Finally escaping from his enforced confinement, Rife lived out his remaining years and died quietly and unknown.

In the present paper it is not our purpose to address Rife's theory of disease or his treatment devices and methods. It is solely our purpose to propose a higher symmetry electrodynamics functioning of his microscope, as one possible explanation of the startling microscopy results he achieved. We will reveal what we are convinced is Rife's great secret: recursive virtual state magnification (amplification) by iterative difference frequency serial summation of vacuum engines. This is a substantial change from an earlier mechanism proposed by the present author, which proved to be in error {1}.

We accent that this is not a "standard electrodynamics" examination of the Rife microscope; normal U(1) electrodynamics and transverse EM waves contain no mechanisms that can explain the functioning of the instrument. Accordingly, since no such explanation is contained in the usual electrodynamics, one must of necessity turn to those available electrodynamics models that are of much higher symmetry than afforded by vector and tensor algebra. To show the fit to experiments, we must also address an enormous amount of the work performed by Prioré and his colleagues. Only in such EM models and their known fit to actual experiments are the additional phenomena and principles available that can possibly explain Rife's epochal results.

Also, because of the importance of this unified field theory area, we include substantial references and additional extended discussions in the endnotes for further technical clarification. Thirty figures are added to illustrate the material.

The Rife Microscope

The Rife microscope { [1] } (Figure 1) cannot be made to work by normal optical principles at the level utilized by Rife. With normal optics, about one-tenth wavelength is as fine an optical resolution as one can get. Since Rife went far beyond that, it should be obvious that he was using optics and electrodynamics of a *novel kind*, and of a type not yet in the conventional textbook.

We believe we are aware of the major principles used in Rife’s microscope, or at least some of them, but prior to 1998 we shied away from anything to do with the scope. My major interests have been in extracting EM energy from the vacuum and in developing a new kind of medical therapy utilizing higher symmetry electrodynamics—and thus unified field theory—to directly and slowly change cellular disease or disorder back to a normal previous state. Obviously, if one is to produce a microscope that "sees" optically further into the realm of smallness than any other microscope can conventionally do (even in theory), that microscope must be functioning in a higher electrodynamics mode rather than in the conventional one.

In this paper, we will try to summarize some deep physics principles underlying the Rife microscope, at least as we see them terms of a higher symmetry EM model, and we cite some references for the bits and pieces. A more detailed explanation must wait for a future time and assistance from very high-level theorists.

Introduction

A good hint is this: Rife (Figure 2) often painfully adjusted his microscope stages for up to *24 hours* before he finally got the instrument "focused." Immediately that shows it is not standard optical equipment. Well-made standard optics—and Rife’s microscope was beautifully made—can be adjusted far more easily than that (in minutes!).

So far as we are aware, no one ever looked at the Rife work except through essentially rather orthodox electrodynamics and rather orthodox optics eyes. Normal optical theory permits resolution of about one tenth of the wavelength of the light used, and a few microscopes in the last few years have gone just a wee bit further. But none have gone anywhere to the degree of smallness that Rife reported.

In our view, the orthodox approach to explaining the operation of Rife's microscope is doomed to fail from the beginning. Even a very good scientist will fail with replication, if that approach is exclusively taken. An entire scientific team and substantial funding will not do any better, unless something technically deeper is understood and tackled from the outset. We give an overview of our approach to tackling that problem. To begin with, we must present some errors in electrodynamics and an omission in general relativity, so that we can develop the necessary background. The discussion is not simple and will require considerable development. Once the considerable background is developed, the application to the Rife microscope will be straightforward.

Errors in Electrodynamics and Adjustments to General Relativity

First, present electrodynamics itself has serious flaws and shortcomings, and needs substantial revision {[2] }. As foundations scientist Bunge points out {[3]}:

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

Bunge {[4] } also said:

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

The flaws in electrodynamics pass over into what has been constructed for optical theory—both linear and nonlinear—which just assumes the electrodynamics as a subset.

For our purpose, we need to clarify what modern electrodynamics does and does not say about the massless vacuum, and about the state and form of EM entities—such as potentials, fields, and waves—created in it (in *empty space*) by matter dynamics, or acting from it (from *empty space*) onto matter and causing matter dynamics.

Maxwell and all his peers believed firmly in the material ether, a thin material fluid filling all space and every crevice. To scientists at the time, not one single point in the universe was devoid of this material etheric fluid—this *thin matter*. The atom had not been discovered, the electron had not been discovered, and electricity was also considered a thin material fluid. Charge had no more meaning than “a piece of the electric fluid,” much like a cubic centimeter of water.

Faraday {[5] } believed that the EM influence in space consisted of *material lines of force which acted as taut strings under tension from some mysterious source*. He believed that EM disturbances were just the transverse wave “perturbations” of these taut strings. In his concept, Faraday omitted the string holder or other agent producing the tensile forces upon the string (Figure 3). He thus omitted Newton’s third law from his electrodynamics concept. He also omitted precisely half the energy, by omitting the equal-energy time-reversed EM wave that accompanies the formation of every ordinary EM wave we take account of. In a receiving copper wire antenna, for example, not only do the electrons interact with the incoming fields, but so do the protons in the nuclei—though highly damped in amplitude by the proton's large mass with respect to the mass of a disturbed Drude electron. However, the two energetic interactions (of perturbed protons and perturbed electrons) are of equal *energy* magnitude. So in the conductor, equal energy is expended to move the "lattice holes" (positrons) as is used to move the Drude electrons, else field theory itself is falsified. Except in semiconductors, electrical engineers mostly just ignore the "disturbed lattice hole current" energy.

Maxwell deliberately captured Faraday's thinking, mathematically {[6] }, and so just *assumed* Faraday's lines of force and the "plucked tight string" transverse EM wave in the material ether {[7] } as well as the unexplained tension . The notion of the transverse EM wave in vacuum has essentially no other theoretical justification than that Faraday-Maxwell assumption and the observed "transverse waves" generated in the receiving "electric fluid in the conductor" in and on the Drude electrons. The transverse EM wave notion in vacuum was maintained only by discarding its missing antiwave that is always formed simultaneously in our transmitters and receivers (e.g., when we perturb the Drude electrons in a transmitting wire antenna, we also perturb the protons in the nuclei with the same energy. The two disturbances—proton disturbance in the nucleus and Drude electron disturbance in the Drude electron gas—perturb the local surrounding vacuum/ spacetime with equal energy, so in the vacuum "virtual particle flux gas" the two perturbations are energetically equal. The receiving wire receives both perturbation waves from the vacuum; one perturbs the electrons and the other perturbs the nuclei. Since the second wave is unaccounted, the perturbation of the nuclei is just considered a "mysterious force" that always appears. In a pumped phase conjugate mirror material, e.g., we interact the waves prior to their reaching the nuclei, by multi-wave mixing. Since the "causative" but neglected second wave coming in from space does not reach the nuclei, but are redirected to form the phase conjugate replica wave produced by the pumping, *the mirror does not recoil and it does not exhibit Newtonian third law reaction*. However, since electrodynamics *omits* the EM causative wave that generates Newton's third law forces in the receiver, which in normal detection always appear, then presently electrodynamic fields are erroneously said to be devoid of Newton's third law reaction. To the contrary, the effects of that hidden, neglected EM cause of the Newtonian third law reaction do appear, either as recoil of the physical receiving antenna or as the production of the time-reverse phase conjugate replica wave in optical pumping. But in standard electrodynamics, equal-and-opposite forces in the receiving antenna are assumed to arise "mystically" or from some hidden *electrodynamic demon*. The electrodynamic cause of Newton's third law has been lost since Maxwell wrote his theory.

Maxwell, however, was careful to point out the ad hoc nature of assuming the stress in the vacuum, and that no mechanism for it had been discovered, nor was such a mechanism contained in his theory. Specifically, Maxwell said {[8] }:

> *"It must be carefully borne in mind that we have made only one step in the theory of the action of the medium. We have supposed it to be in a state of stress, but we have not in any way accounted for this stress, or explained how it is maintained. This step, however, seems to me to be an important one, as it explains, by the action of the consecutive parts of the medium, phenomena which were formerly supposed to be explicable only by direct action at a distance. … I have not been able to make the next step, namely, to account by mechanical considerations for these stresses in the dielectric. I therefore leave the theory at this point, merely stating what are the other parts of the phenomenon of induction in dielectrics."*

Indeed, we now know of electrons and their spin. The longitudinally-constrained, spinning Drude electrons in receiving wires will gyroprecess laterally in the wire when an EM disturbance from the vacuum interacts with them. (Figure 4) It is well-known that these Drude electrons are highly restrained longitudinally down the wire; the longitudinal movement is the "drift" velocity and may be nominally only a few inches per hour. The lateral precession of the longitudinally constrained, spinning Drude electrons establishes that the incoming vacuum disturbance, before its interaction with the gyroelectrons, is *longitudinal*. Else one must discard electron spin, electron gyroprecession, and indeed the theory of gyroscopic precession itself. In our view, the EM wave *in vacuum* is longitudinal, always has been, and always will be {[9] }. It also *does not* consist of EM force fields, contrary to present classical theory. Instead, only the potentiality {[10]} for the transverse force fields—*should* the longitudinally perturbed vacuum EM entity be intercepted by a spinning charged particle and interact with it—exists in mass-free space, as pointed out by Feynman. Quoting Feynman again {[11] }:

> . *"We may think of* ***E****(x, y, z, t) and* ***B****(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the charge there did not disturb the positions or motion of all the other charges responsible for the fields."*

Rigorously the force fields exist only in and of matter. In matter-free space, there is no force at all in what we call the "field as it exists in massfree space" {[12] }. E. g., Jackson states it this way {[13]}:

"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."

But again, Jackson also states {[14] }:

> *"...the thing that eventually gets measured is a force..." "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E. One must be careful in its definition, however. It is not necessarily the force that one would observe by placing one unit of charge on a pith ball and placing it in position. The reason is that one unit of charge may be so large that its presence alters appreciably the field configuration of the array. Consequently one must use a limiting process whereby the ratio of the force on the small test*

> *body to the charge on it is measured for smaller and smaller amounts of charge. Experimentally, this ratio and the direction of the force will become constant as the amount of test charge is made smaller and smaller. These limiting values of magnitude and direction define the magnitude and direction of the electric field E at the point in question. In symbols we may write F = qE where F is the force, E the electric field, and q the charge. In this equation it is assumed that the charge q is located at a point, and the force and the electric field are evaluated at that point."*

As can be seen, much of the difficulty occurs because physicists continue to erroneously utilize an *equation* as if it were a *definition*. An equation defines nothing at all; it merely states that all the "things on the left" of the equality sign and all the "things on the right" have the same overall magnitudes. It does not *define* any of the things on the left or right, and so it tells us nothing at all about what anything in it *really is*. A definition requires an identity, not an equality. E.g., if we take as an identity the very equation advanced by Jackson's explanation above, we then have a definition:

$$F \equiv qE \Rightarrow E \equiv F/q$$

And now that identity clearly states that charged mass is a *component* of the **E**-field. Similar argument prevails for the **B**-field, etc.

The identity also shows that E is not the *magnitude* of the force field at all, but is its *local intensity* at any given point in the E field, in terms of force per unit point static coulomb—a force formed in, on, and of the charged mass of the coulomb. The misstatement that E represents the *magnitude of the field* is another great non sequitur. No textbook gives the calculation of the *magnitude* of the E-field, though all purport to. Instead, they all give the calculation of the *intensity* of the E-field at each point in it. There is a great difference between the magnitude of an automobile's mounted tire and its pressure per sq. inch on the pavement. In all fairness, some of the better texts do use the term "field intensity", at least occasionally.

As *defined*, the EM force fields are effects produced in, on, and of charged matter and *only* in, on, and of charged matter. They are not at all the same effects produced in matter-free space (in the vacuum/ spacetime). The EM force field (in matter) is not the same as the "massfree" E-field in massfree spacetime. Ultimately, all that exists in spacetime are *spacetime* and *changes to spacetime*.

We may accurately take the mass-free, force-free EM field to be a pattern of spacetime curvatures. The coupling of this spacetime curvature set, or any part of it, with charged mass, then constitutes an EM force field. Again, the problem is the hoary old mechanics notion of having a separate massless force acting upon a separate mass. No such thing exists in all nature, and mechanics should have been rather thoroughly overhauled long ago as a result of this horrendous *foundations of mechanics* error that is centuries old.

In present classical electrodynamics—certainly the kind taught to electrical engineers and used in the electrical engineering discipline—the EM potential {[15] }, field, and wave, as they truly exist in space and in the vacuum, are rigorously not prescribed at all. Instead, the effect of the "causal entity" as it exists in vacuum prior to interaction with charged mass, is used a if it were what exists in and on the charged mass as the effect of the interaction. (See Figure 5, Figure 6, and Figure 7). This substitution of the effect for the cause is arguably the greatest uncorrected error in classical electrodynamics today. It is a hangover from the old days of the material ether, and the refusal to change the Maxwell-Heaviside equations to eliminate that material ether assumption, once the Michelson-Morley experiments falsified it. This has long been known to foundations physicists, and to most of the better electrodynamicists as well. But it still seems to be known to very few electrical engineers!

Many attempts have been made to “modify” the standard classical electrodynamics theory, but most have failed. For some reason, most of these attempted modifications seem to have accepted the present "matter-to-matter transform" approach (Figure 8) of present electrodynamics, failing to see that the two missing transforms of matter-to-spacetime and spacetime-to-matter (Figure 9) are in fact just general relativity itself {[16] , [17] } The solution for adding the missing two transforms into electrodynamics therefore may lie in "infolding" general relativity directly within electrodynamics, rather than placing the two on an equal footing. That requires the analysis of the supersystem rather than the system, a concept which will be discussed shortly. Figure 10 shows the utilization of the "vacuum engine" that results from this "infolded GR" solution.

To our knowledge, the Sachs unified field theory {[18] a,b,c,d,e} is presently the only practical unified field theory that can be directly engineered by higher symmetry electrodynamics, such as the O(3) group symmetry EM model advanced by Evans and Vigier and perfected by Evans {[19] }. The engineering power of Sachs' theory is remarkable {[20] }, and the O (3) electrodynamics has been shown by Evans to be an important subset of the unified electrodynamics in Sachs' theory.

Let us examine that statement a little deeper.

In classical electrodynamics the sources of all EM potentials, fields, waves, and their energies are assumed to be the charges—which are assumed to somehow act as perpetual motion machines that continuously *create and pour out* all that EM energy in the fields and potentials associated with the charge as their "source". This outpouring of 3-space energy comprising and establishing the associated potentials and fields eventually fills all space with energy. All the charges and dipoles present in the universe since shortly after its formation have done this for some 14 billion years.

Yet in classical EM—with all the existing charges in the universe implicitly assumed to continuously pour out energy into all of space, and most having done so for billions of years—inexplicably space is *then* considered to be an emptiness having no interactive energetic exchange with the charge at all {[21] }. That of course is a non sequitur because it involves a total contradiction of the conservation of energy law. If there is no active energy in spacetime for the charged particle to receive and then output in 3-space, then the energy conservation is falsified by every charge and every dipole.

In particle physics the inactive vacuum concept is known to be quite wrong, else the charge creates energy from nothing. That would violate the most sacrosanct law in physics: that *energy can neither be created nor destroyed*. So electrodynamics—both classical and quantum, neither of which contains the solution to this *source charge problem*—is in gross violation of conservation of energy laws, because it assumes that the charge is solely the source of its associated fields and potentials and their energy { [22] } and that it accordingly creates all that energy it has poured out since the beginning of the universe.

Presently the electrical engineer implicitly assumes total violation of the conservation of energy law with every equation he writes, and is blissfully unaware of it.

Particle physicists, however, have done much better and published and incorporated their results, although inexplicably the mainstream electrodynamicists have not changed their own model accordingly. Particle physicists have known and shown for more than 40 years that (i) space is a seething inferno of EM energy { [23]}, (ii) a charge continually undergoes a violent energy exchange (see Figure 11) with the vacuum EM flux. That exchange is what a charged mass is doing, to *make* it a "charged mass"), and (iii) the charge is a *broken symmetry* in that violent energetic flux of the vacuum. "Broken symmetry" means that the charge disrupts and alters part of that virtual particle flux to observable EM energy flux (see again Figures 5, 6. and 7). The charge continually absorbs virtual energy from the vacuum, radiates *some* of that absorbed energy back to the vacuum in virtual form, but *integrates and radiates the remainder* in a flow of potentially observable energy—the energy flow discovered by

Poynting {[24] } and Heaviside {[25] } after Maxwell was already deceased.

So what does the source charge really do? It coheres, transduces, and *gates* some of its virtual energy {[26] } received from the vacuum, into an EM energy flow going out in all directions in 3-space from that source charge. Contrary to present electrodynamics, this energy flow is actually longitudinal and bidirectional {[27] , [28] , [29] , [30]}, with energy *going out in 3-space from the charge* to every point in the universe, and with energy coming from every point in the universe *back to the charge through the time domain*. But as reinterpreted {22 b}, Whittaker {[31] } showed in 1903 that any such bidirectional EM flow of energy—reinterpreted with the input energy to the charge coming from the time-domain, and the output energy from the charge being emitted into 3-space, comprises a scalar potential. This also agrees with quantum field theory {[32] }

See Figure 12. As it really exists in space, a scalar potential is a harmonic set of bidirectional EM energy flows, with the "bidirectional" EM waves being outgoing EM waves in 3-space caused by interaction with the source charge or dipole of incoming time-polarized EM waves. There is a set of outgoing EM longitudinal waves in 3-space, perfectly correlated with a set of incoming time-polarized EM waves in the time domain. Hence one time-polarized EM "causal" wave couples with one outgoing longitudinal EM wave in 3-space. It is this scalar EM wave coupling (combining) with longitudinal EM wave coupling, in a harmonic set of phase conjugate pairs, that constitutes the scalar EM potential.

In short, all EM energy in 3-space comes from the time domain.

After all, no observable persists continuously in time. Any observable is a frozen 3-space momentary snapshot of an ongoing 4-space interaction. It is the result of a d/dt operation imposed upon LLLT spacetime, to yield a momentary LLL by d/dt(LLLT) => LLL

What we see as the "motion of an object in 3-space" is actually the ensemble of a continual serial iteration of such "frozen 3-space snapshots" of each observable part of the object, much as the iterative frames of a movie film are seen as a continuously moving picture.

In Whittaker's bidirectional energy flow associated with the source charge, electrodynamicists—including Whittaker himself—assumed that a point positive coulomb of charge is located at each and every point in space. (Figure 5, Figure 6, and Figure 7). A tiny bit of the impinging EM energy flow “from” the distant source charge {[33] } is geometrically intercepted and diverted by the observing/detecting point charge to streamline around it, creating a “swirl around” of diverted energy flow. As can be seen, only a tiny portion of the impinging and passing EM energy stream in 3-space is diverged into the swirl. This is rather like a small diverted whirlpool of water forming around a stable rock in a flowing river. In electrodynamics the amount of energy in that “diverged whirlpool” is said to be the “collected” energy appearing upon that point charge (upon that fixed rock). That amount of energy swirled by the intercepting point coulomb at a given point in space, is then erroneously said to be the *magnitude of the potential (the magnitude of the entire river)*. It is no such thing. The small fraction of the potential that is *diverted from* the potential's flowing rivers of energy, is not that the magnitude of that potential (the magnitude of the entire set of flowing rivers themselves!) And it is certainly not “identically” the river itself! At best it is a measure of the *intensity* of the potential (the local intensity of all those internal rivers of energy flow) at that point occupied by the intercepting/diverting rock.

The point is, *as defined and used in conventional electrodynamics*, the potential occurs only after the vacuum EM entity’s reaction with the intercepting mass, and is defined only as the *effect* produced upon that coulomb by the causative potential's interaction. Literally, the potential (the cause) is erroneously defined as the “excitation” or “excess energy” it locally produces as an *effect* upon a unit point intercepting/collecting charge.

Ironically, it appears that no electrodynamicist calculates the magnitude of the actual potential as it exists in vacuum-spacetime, before its interaction with charge {[34], [35] }. Its magnitude is enormous, since the potential may extend over all space, and energy can be intercepted and collected by a charge placed at any point therein, or from charges placed at every point therein.

Certainly no rigorous definition of the potential presently exists in physics; the reader is challenged to check for himself or herself. No legitimate definition *can* exist for the potential, until the little matter is resolved of its present erroneous definition only "in and on and deviated by" charged matter, contradicted by the simultaneous assumption that it does exist in mass-free space in undeviated form. What is *deviated from* the potential, is not the *potential itself*. A part of something torn out of that something at one point in itself, is not the whole. The *local effect* cannot be substituted for the *entire nonlocal cause*—only a small part of which interacts *locally*.

The EM force field is similarly defined—only for charge mass interception, and only as an *effect* produced upon the intercepting charge. (See again Figures 5, 6, and 7.) A flow of water across a rock (or the wind on a rock in the desert) produces a “pileup” of fluid on the windward side, and a lesser “pileup” on the downwind side. The difference (called the “gradient”) in the energy density collected across the standard charge (rock) produces a “difference in pressure” on the rock—in other words, a force per interacting rock. That *effect* is erroneously taken to be the EM field itself, when rigorously it is only the reaction cross section of the EM field. Again, rigorously this EM field *as defined in electrodynamics* only exists as the effect that occurs upon an intercepting standard charge or "rock" {[36]}.

Since the EM wave “in space” is erroneously said to be made of varying EM fields, it too is defined only as the effect upon the intercepting charged mass, after the vacuum EM wave entity—whatever it is—has interacted with the charge and changed the charge translationally and energetically. Specifically, the present “EM wave in vacuum” in the literature is not at all what is presented and taught in undergraduate texts, and even the standard illustration of "oscillating E-field and B-field vectors" lateral to the line of motion of the wave is horribly wrong. E.g., Dr. Robert Romer, former editor of American Journal of Physics, has taken that error to task as follows { [37] }:

> *"…that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. …it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "…perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

The EM field and wave *as commonly regarded* are transfer functions from the source charged masses that “perturbed the vacuum medium and created an unknown kind of disturbance therein”, onto the receiving charged masses that intercept the incoming “unknown vacuum wave entity that was created” and interact with it, being translated and perturbed as a result.

This “detected” translation of the receiving charged masses (usually the interacting Drude electrons in a wire) is what our instruments detect. Indeed, the Drude electrons are highly constrained from moving longitudinally down the wire, usually moving at only a few inches per hour. They are much freer to move laterally, however, and can easily do so.

Reacting with the incoming longitudinal EM wave perturbation of local spacetime, the spin of the electron produces gyroscopic precession laterally in the wire since the electron is terribly restrained in moving longitudinally down the wire. Since our instruments detect these *transverse electron precession waves*, that has been totally confused (since Faraday's "plucked taut string" notion of the EM wave and Maxwell's arbitrary adoption of it) as representing the incoming wave in vacuum. It does not represent the vacuum EM wave at all; the incoming wave is in fact a longitudinal EM wave in 3-space, associated with a precisely correlated longitudinal EM wave in the time-domain.

So we have answered Professor Romer's implied question. That horrible diagram came from the assumption by early electrodynamicists, before discovery of the electron or atom or nucleus, that the "transverse electric fluid waves" in the receiving wires represented actual "intercepted" fluid perturbations arriving in the material ether medium.

In summary: Using a transmitter-receiver analysis, the so-called “EM wave in vacuum” *as presently defined*, actually represents that “detected and measured electron precession wave” in and of the reacting Drude electron gas in the distant receiving wire antenna. The so-called EM force field wave in vacuum has been erroneously defined as observable matter wiggles, not vacuum virtual energy wiggles and not as spacetime-curvature changes and wiggles. So even the present notion of the EM transverse wave in space is wrong, and it has been corrupted as a pure matter-to-matter transform from the beginning { [38] }.

Somewhat less technical explanations are given by Lindsay and Margenau { [39] }, and a rigorous statement and discussion that EM force fields exist only in charged matter is given by Aharonov and Bohm { [40] }. Feynman of course said it bluntly: Only the potential for the field exists in the vacuum, should one place a unit point charge there to allow the force field to be developed (on, of, and containing the source charge itself).

Let us summarize this terrible EM foundations flaw: As classical electrodynamics presently is modeled, Maxwell and his modern followers only include equations which are *mass-to-mass transformations*

(Figure 8). Specifically missing { [41] } are two other transforms: (i) the mass-to-vacuum (mass-to-spacetime) transformation (which obviously is general relativistic in nature), and (ii) the vacuum-to-mass (spacetime-to-mass) transformation (again, obviously general relativistic in nature). (See again Figure 9). When these two missing transforms are added back into electrodynamics, one has suddenly extended EM theory into a new domain. One has also *infolded* general relativity (GR) inside the present EM matter-to-matter transforms, and also included the vacuum interaction including broken symmetry in that exchange, but in an *electromagnetically engineerable manner*. In short, one has a practical, engineerable unified field theory.

Highly creative inventors such as Rife and Prioré intuitively developed apparatuses that unwittingly utilized this presently undeveloped and extended union of EM, particle physics vacuum theory, and GR. The intuition of the theorists in interpreting those revolutionary experiments and their results has very seriously lagged.

The Supersystem Concept

To understand how extraordinary systems such as Rife's microscope works, one must understand the concept of the *supersystem*. (See Figure 13). The supersystem consists of three components, which are (1) the physical system and its dynamics, (2) the local active vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. All three components of the supersystem interact with one another, in nature. The second and third components of the system constitute the active environment in which the system is embedded, and with which it interacts.

Unfortunately, in standard classical electrodynamics either the system's environment is assumed to be inert, or the system is assumed to be in equilibrium in its interaction with that environment. This represents a totally arbitrary assumption that the system cannot receive and utilize any excess energy from its known active environment. Hence the active vacuum's interaction with the system can be and is ignored by classical electromagnetics, as are the local curvatures of spacetime and their interactions with the system.

Extraordinary EM systems invariably involve the supersystem dynamics, and not just the system dynamics. *In nature*, no system analysis is complete until the supersystem interactions and their effects have been determined and analyzed.

The Rife microscope is indeed an extraordinary system using non-negligible interactions between the system and the other two components of its supersystem. Hence in our discussions, we will be examining interactions in that regime, and not just the standard electrodynamics analysis. In these discussions, we will be holding in mind not just the mass-to-mass transforms of standard electrodynamics, but two other transforms: (1) the mass-to-vacuum/spacetime transform, and (2) the vacuum/spacetime to matter transform (previously shown in Figure 9).

The reason that Rife's work has not been understood is that researchers have been unaware of the necessity for supersystem analysis. Almost all analyses of Rife's work have been merely standard classical EM analysis and thus quite insufficient. There are other examples of the standard electrodynamics model completely failing to show higher symmetry phenomenology in an inventor's work, so that his work is not understood even today. An example is provided by Tesla's patented circuits, as rigorously demonstrated by Barrett { [42] }. Barrett went on to improve the new functions discovered in Tesla's patents, and obtained two additional patents related to signals and signal processing { [43] }.

Vacuum-Structurings as “Engines”: Key Part of the Missing Transforms

The missing mass-to-vacuum/spacetime transformation produces a massless structuring of the vacuum medium. That is a *vacuum engine* or “multiplex of specific spacetime curvatures.” (See Figure 14, Figure

15, Figure 16, Figure 17), and Figure 18. Preparing the vacuum engine is accomplished by charged matter dynamics of the source charges producing the necessary potentials, fields and waves. The charges are embedded in continual energetic exchange with the vacuum/spacetime, and moving the charges affects that exchange; it perturbs the local vacuum flux and curves the local spacetime accordingly.

The transform represents producing this vacuum engine by the vacuum-perturbing action of source charge dynamics upon the vacuum, and by *not* assuming a mystical intercepting point-charge at each point in space { [44] }. In the vacuum/spacetime, *only* vacuum/spacetime entities and changes to them exist. *In pure water, only water and changes to that water exist!*

So a precise dynamic structuring of the energetic vacuum (and spacetime geometry) is what charges produce when they “perturb” the vacuum medium { [45] }, by the presence or dynamics of their fields, longitudinal and time-polarized EM waves, or both. This "perturbed vacuum dynamics" separate from the source charges themselves comprises an engine that consists of altered longitudinal EM wave dynamics inside the internal Whittaker structure { [46] } of the scalar potential of the ambient vacuum. That is the same as also producing internested levels of spacetime curvatures in a *template* or specific pattern { [47] }.

In turn, this vacuum engine region of local curvatures and structurings in spacetime will interact upon any mass placed therein. That action is the missing *electrodynamic vacuum-to-mass transform.*

The missing electrodynamic vacuum-to-mass transformation produces direct changes and alterations in the “receiving” mass, at all levels, by the action of such a vacuum engine. As each action occurs locally in the localized receiving mass, it produces the "local effect" as the result of that “local action.” The primary causative chain (i.e., the set of “dominoes” affecting each other serially), started back at the distant source charges. One can say that we therefore have mass-to-mass *action at a distance*, via the matter-to-vacuum and vacuum-to-mass intermediary transforms and the spatial and temporal propagation of the vacuum engines in the middle of the two { [48] }. We will shortly discuss the Whittaker structuring of the potential as a fundamental process for producing vacuum engines.

General relativists mostly concentrate upon the warping and curving of spacetime (i.e., producing a vacuum engine) by the very weak gravitational force—which is only some 10^{-42} times as strong for electrons as the electric force between them. Consequently, relativists must look to huge assemblages of mass, such as in stars and other astronomical objects, for sufficient gravitationally-induced spacetime curvature to produce easily observable results. For that reason, general relativity has resisted laboratory development as an engineerable applied science. By using the far stronger EM force as the agent of spacetime curvature, and including the infolded electrodynamics inside the potential, field, and wave, general relativity then becomes engineerable and usable. But one is engineering the supersystem, not just the system.

The Infolded Electrodynamics Inside Potentials, Fields, and Waves

As previously stated, in 1903 E. T. Whittaker, a well-known mathematical physicist, showed that a scalar potential can be mathematically decomposed into a set of peculiar EM wavepairs in a harmonic set. (Again see Figure 12) These "hidden waves" are longitudinal EM waves, arranged in conjugate pairs { [49] } with the pairs also arranged in a harmonic set. In each wavepair, there is an ordinary (forward-time) longitudinal EM wave (outgoing in 3-space from the interacting/observing charge), coupled to its phase conjugate replica (time-reversed twin).

And here lies a magic secret. Prior to interacting with a charge, the phase conjugate wave in each wavepair is in the complex domain, which means that in Minkowski space it is incoming along the fourth axis, *ict*, where the only variable is *t*. Hence this wave prior to interaction with charge does not even exist in 3-

space. Instead, it exists as a longitudinal EM wave incoming in time, so that it is a time-like flow of EM energy. (See Figure 19). After interaction with a charge, the charge absorbs the incoming time-polarized EM wave, transduces that time-energy to 3-spatial energy, and re-emits the energy as longitudinal EM waves in 3-space in all directions. "In all directions" means that for every outgoing longitudinal EM wave emitted in one direction in 3-space, another wave just like it also emitted in the other direction in 3-space. So the paired "bidirectional" EM longitudinal EM waves in 3-space, as interpreted by Whittaker, do exist as the *effect* waves caused by the incoming causal time-like longitudinal EM waves (scalar EM waves). The pairing of the causal time-polarized EM incoming wave with an outgoing effect longitudinal EM wave was not interpreted by Whittaker, but does agree with more modern quantum field theory interpretations of photon polarizations and their observability or non observability, by Mandl and Shaw {22b}.

We have corrected this "effect substituted for the cause" wave to reinterpret Whittaker's 1903 decomposition of the scalar potential, into combined biwaves where one of the biwaves is incoming in the time domain, and the other is outgoing in 3-space. That re-interpretation then allowed a solution {22a} to the problem of the source charge and its associated fields and potentials, together with their energy. Every source charge and source dipole continuously emits EM energy in 3-space, without any 3-spatial EM energy input. The input is there, but in the time-domain. This problem—resolved by the present author in 2000—has been called the most difficult problem in quantum and classical electromagnetics {30 }.

Because of observed parity reversal accompanying the time-reversal operation, in his forward-time, the observer “sees”—after interacting the incoming wave with charge—the *effect* of the incoming time-reversed or phase conjugate wave (entering from the time domain) as a 3-space longitudinal EM wave leaving in the opposite spatial direction. So the basic 4-space phase conjugate longitudinal wavepair would appear to us in forward time as a “regular” time-forward longitudinal EM wavepair in 3-space, with the two waves going outward in opposite directions. But each is accompanied by a time-polarized EM wave, else it would not be observable.

Further, there is a harmonic series of such bidirectional longitudinal wavepairs comprising the scalar potential. So the staid old electrostatic potential (i) *is not* a "scalar entity" at all { [50] }, (ii) *is* a bundle of longitudinal wavepairs, (iii) *is* an entity with a remarkable internal hidden multiwave structures and dynamics, (iv) *is* a composite of very special waves whose transverse amplitude (i.e., whose spatial energy density, which is a function of transverse amplitude squared) does not change, but whose transported time structure changes, and (v) therefore is a special kind of wave that is velocity modulated about some central velocity, and is not limited to the speed of light c. Indeed, in classical electrodynamics in the Coulomb gauge, the scalar potential is already recognized as having infinite velocity, merely appearing in space instantly at every point it occupies { [51] }.

We also state without elaboration that the active vacuum is *identically* a very huge scalar potential, as indeed is “spacetime geometry”. So a unifying principle is:

vacuum ≡ spacetime ≡ potential ≡ energy { [52] }

Further, as a potential, the “ambient” vacuum decomposes into an incredibly rich, internal Whittaker wave structure. So then also must *spacetime* and *energy* similarly decompose. That decomposition structure is even richer when higher symmetry O(3) EM is used instead of the lower symmetry U(1) EM employed by Whittaker.

A revolution in electrodynamics has recently begun, and central to that revolution is the strong appearance of longitudinal EM waves as essentially the basis for the emerging new electrodynamics. Work by Evans { [53] }, Barrett { [54] }, Ziolkowski { [55] }, Evans { [56] }, Rodrigues {57}, and others is fundamental in this respect. One part of this emerging new electrodynamics is advanced under the label of *undistorted progressive waves* (UPWs). A good summary of UPW theory is given by Rodrigues and Lu { [57] }. Even so, to date this body of work still largely ignores the internal Whittaker dynamics { [58] } inside an EM wave. The body of work—particularly the higher symmetry O(3) electrodynamics pioneered by Evans {56} inherently possesses the capability to model such structuring, since much of it utilizes a higher topology algebra such as quaternions { [59] } or even Clifford algebra { [60] } In those higher algebras, many electrodynamic things can be done outside the operations permitted by present vector and tensor electrodynamics.

Superpotentials and Superpotential Theory

Whittaker's 1904 paper { [61] } showed that all the "normal" electrodynamics was based on, *and created by*, the interference of scalar potential functions (i.e., by the interference of those multiwave longitudinal wavepairs and their dynamics { [62] } constituting the interfering scalar potentials.) He showed that any EM field or wave whatsoever can be replaced by two scalar potential functions—in fact, is *comprised of* those two scalar potential functions and is *created by them*. Scalar potential interferometry already creates all the "normal" EM in the textbooks anyway, and a proof has been given by Evans et al. { [63] }.

The 1904 paper by Whittaker initiated a somewhat obscure branch of electrodynamics referred to as *superpotential* theory { [64] }. Whittaker's work was further extended and augmented by later scientists such as Righi { [65] }, Debye { [66] }, Bromwich { [67] }, Nisbet { [68] }, McCrea { [69] }, etc. But even with the extended superpotential theory, without the addition of Whittaker 1903 and the substructuring "internal LW electrodynamics" one cannot understand the mechanism for what Rife was actually doing.

Rife was not using normal potentials and normal **E** and

H fields—which as we discussed, only apply to observable material entities anyhow and thus could not be used to "see" far below the quantum threshold of least detectable material disturbance. His entire protocol was to get beyond those "*material interception of EM energy*" limitations. Unwittingly, Rife was using *vacuum engines*—involving structuring of the active vacuum as well as pure general relativity (GR) and pure structurings of spacetime geometry itself. He was electromagnetically using that part of GR that the GR physicists have mostly only tried to produce by use of the weak gravitational force. In GR, *the ST geometry itself is active, dynamic, energetic, and structuring!* In higher symmetry electrodynamics, *the ST geometry itself is powerfully active, dynamic, energetic, and structuring!* It is not at all just a "passive spacetime" as classical electrodynamics assumes { [70] }.

Biological Systems Use the Infolded Whittaker-Type Electrodynamics

The problem in the Rife microscope is to be able to observe very subtle ST curvatures and patterns, in an optical manner. These subtle or virtual "disturbances" of local spacetime are far finer and smaller than the one-tenth wavelength optical resolution generally the limit for optical microscopes using conventional optical EM notions and "material" potentials, fields, and waves.

In short, Rife's twofold problem was to (i) function with vacuum engines (internal structuring of the potentials) so he could penetrate to any degree of smallness, and (ii) output, in the observable state, exact visible summations and analogues of the individual structures detected in the virtual (nonobservable) state { [71] }.

To make the subtle vacuum engines observable on the bench, the ST curvatures and their internested patterning must be accomplished by making and assembling local curvatures of ST, via making longitudinal EM wavepairs in structural pattern assemblies. *Spacetime* is in fact just a powerful scalar potential, or can be modeled as such, and as such Whittaker 1903 and 1904 rigorously applies, along with extensions from higher symmetry electrodynamics. By realizing that ST is identically potential, one has altered one's thinking from the standard GR theory into a correlated EM-GR theory. One has also extended GR: Not only is there a “topology” of things that can be externally *done to* a base spacetime, but there is also a hidden, vast EM topology *within* any overall base spacetime. So we now have “topologies within topologies.” This is in fact what the Russians have long called the *information content of the field*.

However, all living systems already use this "infolded" bidirectional, longitudinal wavepair EM in their ongoing living functions { [72] }. Just as they used frequency modulation, EM signals, EM oscillations, etc. before we even had an electrodynamics or a physics, living systems do use the infolded EM (and vacuum engines) in all their living functions, and particularly in their cellular *regeneration and restoration* (R&R) system { [73] }, as contrasted to their *immune* system. The immune system cells are the fighters and the debris scavengers/cleaners. They go after the hostile invaders, fight them, and usually win—littering the battlefield with the debris. Then the immune system scavenger cells clean up the residue.

But many of the body’s cells will have been damaged in the fight. The immune system itself cannot heal or restore a single damaged cell, even its own! It is not a *healer*! It is a searcher, a marker, a killer, and a residue-cleaner. It contains " the troops and the combat engineer," so to speak, but not the hospitals and the doctors. All healing/restoration of living cells is done by the R&R system, not by the immune system.

A high-level overview of the immune system is given in Figure 20.

No vaccine, drug, herb, vitamin, or mineral *heals* the body, although certainly they can enhance or aid the body's healing process. Each does carry its individual resident vacuum engine, and when absorbed by the cell, this added vacuum engine contributes to the resident vacuum engine in the cell by summing with it. To restore the damaged cells back to normal (i.e., to heal), the R&R system uses a novel kind of extended electrodynamics with infolded vacuum engines, and it uses a novel kind of optical phase conjugate pumping, in the time domain rather than just 3-space. The magic "unified field theory" so long sought by scientists, has long been utilized by the regeneration and restoration system of the body in its minute-to-minute and day-to-day healing and restoring actions.

The magic "universal healing mechanism" so long evading the discovery of medical scientists is none other than the exact cellular regenerative mechanism universally used by every living organism. For the first time, we are explaining the higher symmetry electromagnetic nature of this mechanism.

Robert Becker's epochal work { [74] } (Figure 21, Figure 22) strongly probed the electromagnetic control system directing the body's cellular regeneration and repair (R&R) system. However, he did not have available the infolded Whittaker electrodynamics. He could only utilize the grossly inadequate standard electrodynamics. Modern nonlinear phase conjugate optics had not yet been born when Becker did most of his seminal work.

Since standard EM does not incorporate infolded GR, vacuum engines, or optical pumping in the time domain, Becker could not formulate the full technical mechanism of the R&R system —which uses the *supersystem*, not just the system. But Becker did strongly point out the importance of the R&R system, both theoretically and experimentally. He found that it operated electrically, with the scalar potentials being key. He modeled the R&R system as closely as can be done with classical U(1) electromagnetics (Figure 22). He proved that cells can be *dedifferentiated* (reversed back to an earlier, more primitive state) and *redifferentiated* (time-forwarded into a more complex state) by weak DC potentials and laughably weak (picoamperes) of current. He was twice nominated for a Nobel prize for his epochal work. His method of electromagnetic healing of otherwise intractable bone fractures (Figure 21) is utilized today in many hospitals.

In his microscope, Rife was in pursuit of optically "seeing" the infolded electrodynamics functioning of the living cells in the most minute detail—i.e., he wished to see the actual vacuum engines and their functions. Using the infolded EM, one can still "see" the cells and all their parts, but one can also go far beyond the one-tenth wavelength optical discrimination limitation and see into the virtual state dynamics. In theory, because the infolded EM is *recursive* as a function of wavelength, one can go to any limit in smallness desired—or at least to whatever limit is determined by the available technology's development at the time { [75] }.

Vlail Kaznacheyev's novel electromagnetic induction of cellular death, disease, and damage, *at a distance*, is instructive in this respect { [76] }. (Figure 23) Indeed, the Russians mastered, extended, and weaponized the Kaznacheyev mechanisms as disease-inducing weaponry, and slyly used it at low level on personnel in the U.S. Embassy in Moscow (Figure 24). This was done to stimulate high level officials in the U.S. government and U.S. scientific community to see if the U.S. knew of this "infolded electrodynamics and vacuum engine" technology { [77] }. The Russians were using GR, but by infolded EM means—so they were using a unified field theory involving higher group symmetry electrodynamics. The puzzled U.S. employed only standard U(1) symmetry electrodynamic analysis { [78] }, so it totally missed the entire physics of what was occurring to cause the health changes and diseases.

In GR, one is very interested in Wheeler's principle which states (slightly paraphrased):

> *Mass (trapped energy) acts upon ST geometry to curve it, and curved ST acts upon mass to produce forces upon it and move it.*

This statement after Wheeler is the very essence of general relativity. It simply states that there is a *mass-to-spacetime* dynamic transform, and also a *spacetime-to-mass* dynamic transform, as we diagrammatically showed in Figure 9.

Understanding Vacuum Engines and Their Action

Wheeler's general relativity principle includes the missing two transforms of electro-dynamics. We are extending Wheeler's principle to the corollary form:

All levels of the structures of mass (trapped energy structures) and all levels of the structures of time (the time aspects of photon structures) act upon ST geometry to structure it in both 3-spatial energy density structuring and in time domain structuring. Producing this multilevel spatial and temporal structuring of spacetime is called "forming a vacuum engine." Both 3-spatial energy density structuring of spacetime at all levels and time structuring of spacetime at all levels, act upon mass at all levels, to produce internested templates of forces and translations and stresses.

In short, energy patterns and time patterns in 3-space act upon spacetime to produce patterned curvatures of ST geometry (vacuum engines). And conversely, patterned curvatures of spacetime geometry (vacuum engines) act upon mass to produce corresponding patterns of forces at all levels in the mass—down to and including upon the quarks and gluons in the nucleons in the atomic nuclei. We note with pleasure that the Sachs unified field theory does apply from the gluons to the entire universe, as has been shown by Sachs himself {18b}.

Muscle and Skeleton Analogy of Interaction of Vacuum Engine and Mass

We call attention again to our analogy (Figures 15, 16, 17, and 18) of the interaction of a vacuum engine and mass, provided by the interaction of arm muscles and the skeletal arm bones to which they are attached. The arm muscles attach to the arm's skeletal bones, and those bones are articulated and free to dynamically move. To the bones, the muscles are mysterious agents (analogous to vacuum engines) exerting forces upon them, and causing them to move or hold. To the muscles, the bones are positioned resisting masses for them to work upon and change their positions, and the bones exert Newtonian reaction forces back upon the muscles as the muscles apply forces to the bones. So the vacuum engines (muscles) create forces upon the resisting masses (bones) to move them, and the masses act back upon the causative vacuum engines to create antiforces in resistance.

Using this analogy, when we create a set of vacuum engines, we create a set of “mysterious muscles” that act directly upon any and all “masses” (bones) that are exposed to them. The bones then change and move until the bidirectional interaction forces are balanced (equal and opposite). This symmetry between the two missing transforms is shown in Figure 9.

The Multilevel Structuring of Time is Important

We also add the extension that dynamic structuring in the time stream also occurs from trapped time domain “time component” structuring and distribution effects, as well as dynamic structuring of the spatial energy density of the ST geometry (the vacuum).

We are speaking *relativistically* of the interaction of mass and spacetime at any and all levels of a mass, down to and including the molecules, the atoms, the atomic nuclei, the nucleons, and even the quarks comprising the nucleons (comprising the protons and neutrons). One must first clearly have this vision in mind: The dynamics of everything ongoing in that mass, down to the most minute particle dynamics and in the most minute detail (and far into the virtual state beyond the one-tenth wavelength limitation of light resolution) consists of the structurings of this "mass-ST curvature" mutual interaction.

So we know from general relativity that collections of energy (such as mass) act directly upon spacetime, its structuring, and its dynamics at any and all levels, down to the tiniest. We know that the variations in local spacetime geometry (including in both energy and time), down to any and all levels and sizes, down to the tiniest conceivable lengths, act directly upon any and all exposed mass at that level, *no matter how tiny*.

There is no “quantum” limitation nor is there any “wavelength of light” limitation to this GR process. Total continuity applies.

Some Characteristics of Vacuum Engines

The great advantage of a vacuum engine is that the "action" it produces upon an exposed mass, arises directly from within each point in the local spacetime in which the mass is embedded. One does not have to "start outside" and move through space and the intervening mass toward the "inside". Every point within an object occupies every point in time, a priori. From a single time-point, the vacuum engine's 3-space action enters everywhere within the object, all-at-once, and the 3-space effects proceed outward from each interior point. (See again Figure 19). Therefore even a weak vacuum engine induced by EM forces can do with ridiculous ease what the most powerful particle accelerators on Earth cannot do. The engine is already everywhere "inside" an object or particle, and working from every point in it, toward an outward direction. On the other hand, a particle accelerator is trying to build tremendous velocity of its particles so that they can "smash" their way in there, by brute force. To get ever deeper, the particle accelerator must be made ever larger, so that a "bigger hammer" is available. Also, the vacuum engine does not smash or damage any of the matter in the object, in any fashion not desired. The particle accelerator, on the other hand, is a great disruptor and may well destroy or serious change the very object it "probes".

Further, contrary to the accelerator's "one shot for one giant amount of energy input", a vacuum engine—once formed—continually exists in that locally altered spacetime and is contained in it and is sustained by it.

Hence that locally altered spacetime—after we pay to alter it once—will continuously expend energy in the form of the action of that vacuum energy, forever, upon any mass emplaced in that local spacetime. The universe itself is furnishing the energy to do the work, via the instantaneous connection in the time domain.

That "multiply connected space" entry everywhere within an object simultaneously, is quite different from "spatial propagation of energy" through singly-connected space from outside an object to the object, striking the outside of a mass and being absorbed, producing a gross translation force upon the mass. In the “energy propagation through space” case, severe energy interaction may occur with only the absorbing part of a macroscopic mass, while the remainder of the mass is little affected or not at all. To wit, one does not get the resulting *electronuclear* alterations of the nucleus (materialization, dematerialization, transmutation, etc.) by absorbed propagated “normal” EM radiation. However, with weak EM-induced vacuum engines one can do things inside the nucleus such as easily flip quarks inside the nucleons, so that transmutations at very weak spatial energy levels are readily obtained. This in fact turned out to be the major secret of the low-energy nuclear transmutations occurring in some 600 cold fusion experiments worldwide { [79] }.

Electrodynamic force fields implicitly define primarily only translation effects upon charges or masses, in their very definitions. With normal EM, usually one will get reradiation of the absorbed energy, or partial reradiation of the absorbed energy accompanied with recoil of the absorbing mass, or no reradiation but recoil of the mass.

We took the term *vacuum engine* from Misner, Thorne, and Wheeler { [80] } and also from Nobelist T. D. Lee { [81] } who spoke of this sort of thing as *vacuum engineering*. The term is not original with us; only the novel higher symmetry electrodynamics application of the concept is original. We call the exact geometrical form of a vacuum engine a *template*. We refer to the process of internally structuring a field, wave, or potential as *dimensioning* or *activating* or *conditioning* or *charging up* the field, wave, or potential. Russian superpotential weapon scientists call the infolded structuring itself the *information content of the field*. For further discussion of some of these terms and related concepts, the reader is referred to our website, www.cheniere.org.

Wheeler's GR principle says that, for every mass with its exact mass structural dynamics condition, there exists a corresponding exact vacuum engine template containing precise spacetime curvature structural dynamics, and interacting back upon the dynamic mass at all levels. That vacuum engine is said to be resident in the mass and continuously interacting with it (see again Figure 13 on the *supersystem*).

Again, the structuring of the vacuum engine has no limitation in fineness, because spacetime is continuous. Hence it freely extends from the observable realm on down into the virtual realm. This is no more audacious than stating that virtual and observable entities occupy the same time domain simultaneously. Obviously, if one is using the structuring of time, one gathers in both those virtual structures and those observable structures, and also the exact correlation between them. This is a direct engineering application of an EM hidden variable theory, somewhat similar to that proposed by David Bohm { [82] }.

Applying Vacuum Engines to Living Cells

Any normal cell already has a normal *template* of overall ST curvature due to its mass and trapped energy structures. No matter how small, every internal part of that cell has its own *subcomponent template* of subcomponent ST curvatures embedded in that overall cellular ST curvature. As the cell functions, the dynamic normal changes of that template into other related templates, is within the range of what we call "normal, healthy cellular functioning."

Any abnormal cell has an *abnormal template* and also an abnormal set of "templating" changes. In other words, there exists a precise *vacuum engine delta* in the abnormal cell, summed to the cell's *normal vacuum engine*. The *sum engine* represents the exact vacuum engine to act on the normal cell and change it to that precise abnormal condition.

But a template depends upon the energy density structuring of the local vacuum and also the local *time* structuring—which latter is still missing from physics books, but is implicit in the nature of the photon. If one produces the necessary energy structuring and density—and the necessary time structuring and density of local ST geometry—one produces a certain vacuum engine and one also produce a certain *template of time structuring*. Time { [83] } has dynamic template structuring and functioning, just as does the 3-spatial energy density of ST geometry.

Indeed, the total (virtual and observable) photon interaction with a mass or particle generates the "flow of that mass or particle through time" (Figure 25).

But one *can* make vacuum engines using EM energy and structuring, rather than with normal gravitational (G) energy and structuring { [84] }. The advantage is that, for electrons, the EM structuring is on the order of 10^{42} times as strong as the G structuring. For protons, the EM force is something like 10^{38} times as strong as the G force. So by using the much stronger EM force as the "agent of ST curvature," one can make vacuum engines that are quite strong at their local level. These engines will affect and profoundly change any exposed mass, including cellular mass—in any fashion desired, given only that one makes the appropriate vacuum engine. Again, *we are simply engineering practical but ultra-strong general relativity on the laboratory bench*, by electromagnetically engineering Wheeler's principle in an extended form.

For any physical cellular deviation from normal—any deviation whatsoever—a corresponding exact deviation (i.e., an exact delta) exists in the cell's otherwise normal vacuum engine template. This includes deviations in the genetics of the cell and everything else, not just the chemistry and gross physical structure.

Now if one could produce the negative (time-reversal) of the sum template of the resident vacuum engine and *amplify* it, one would have a powerful new type of vacuum engine. (Figure 26) It would be an exact *vacuum anti-engine* to act upon that specific diseased or abnormal condition of the cell and gradually erase the delta and bring the cell physically back to normal. The new anti-engine would operate upon all the mass of the cell, at every level no matter how small. It would change the cell's abnormal genetics back to normal genetics.

One could of course reverse cellular damage due to aging, and rejuvenate the body. It is quite doable; all that is needed are the funds, scientific team, and laboratory facilities to develop it. It is doable whenever the U.S. medical science community and the giant pharmaceutical community will allow it to be done.

One could reverse a cancer cell back to a normal cell, by applying the specific summation vacuum antiengine for the exact delta between the cancer cell and a normal cell.

One could reverse an AIDS cell back to a normal cell with normal genetics, where it no longer was an "HIV factory". An AIDS patient could be completely, quickly, and cheaply cured, the moment he or she tested HIV positive. There would be no need to wait for the debilitations of the progressing disease to appear. Not a single HIV-infected cell would be left in the body, even those resistant strains that have developed after the original infection.

One could do precisely the same thing with Ebola infections, and with the "tough" staph infections that now resist everything and are killing more than 120 thousand Americans in our hospitals every year, with the figures rising. *These are patients who did not have the disease that killed them, when they entered the hospital.*

With development of portable treatment units and mass production, one could treat and heal those millions of mass casualties that could well erupt in one or more large U.S. city areas—e.g., from an Iraqi-sponsored terrorist biological warfare (BW) strike, using anthrax { [85]). In 1998, we proposed a crash development of just such portable units { [86] }. With the present world situation, it is not a matter of *if* such terrorist BW strikes will be made in the U.S., it is a matter of *when* { [87] }. Since our 1998 proposal, there has now been time to have had those portable units developed and mass-produced, at a total program cost of perhaps $100 million. Sadly, no one apparently had the foggiest notion of what was actually being proposed, and its vital importance to the survival of the nation. With a looming war in the MidEast—and with hordes of terrorist teams with biological warfare agents already infiltrated into U.S. cities and population centers—this negligence is particularly of concern at this very moment. These officially-forecast attacks on our cities by weapons of mass destruction could now be imminent, if and when the MidEast does explode into war, which appears imminent.

The Process Has Been Experimentally Proven

A very similar thing is also what Becker { [88] } demonstrated, without considering the vacuum engine (general relativistic curvature of spacetime) effects. It is also precisely what Prioré { [89] } unwittingly did in France, with a team of renowned French scientists, to cure terminal cancers, atherosclerosis, lethal infectious diseases, etc. in laboratory animals under rigorous scientific protocols. The team also demonstrated the restoration of suppressed immune systems.

In Germany, Popp { [90] } *et al*. were also involved in studies dealing with the R&R system, but from a quantum electrodynamics viewpoint.

We can also consider the vacuum engine from a nonlinear optics view, and obtain startling extensions of both general relativity and nonlinear phase conjugate optics. That in fact is exactly what Prioré did, though he did not think in those terms. Prioré roughly created a complex set of bidirectional longitudinal EM wavepairs—i.e., a set of "time domain pump waves." So he produced a kind of vacuum engine that itself is a crude *anti-engine generator* when used to pump a mass such as a living cell.

Prioré unwittingly used an extension to nonlinear optical (NLO) phase conjugation *before nonlinear phase conjugate optics was even born*. Pumping diseased cells in the Prioré manner causes the formation of exact "antiengines" for the specific cellular disease or disorder to be created in the cells themselves.

First, Prioré mixed up to 17 or so EM transverse waves in a rotating plasma { [91] }. Today we know that such a plasma will transform transverse radiations into longitudinal radiations, and it will also phase conjugate the input waves and add their phase conjugate replicas. In short, it is a procedure for somewhat crudely making a set of longitudinal bidirectional EM wavepairs, which is required to produce a scalar potential with a deterministic internal structuring. In short, it produces a scalar potential with an internal vacuum engine template.

These infolded longitudinal biwaves are NLO pump waves of a new kind: They pump the PCM material in the time domain only. This kind of pumping is a *vacuum antiengine-making process*.

The plasma tube was surrounded by a giant coil, and a pulsed magnetic field was produced by the coil. Per Whittaker's 1904 paper, that pulsed DC magnetic field can be decomposed into two scalar potentials. Hence we have three potentials superposing: (1) the potential function set made by Prioré in his plasma tube, and containing a selected "vacuum antiengine maker" consisting of bidirectional pump waves in the time domain, and (2 and 3) the two Whittaker potential functions comprising the DC pulsed magnetic field of the coil.

The end result of this superposition/mixing of potentials is that their "internal structures"—including their vacuum engines—also diffuse one into the other.

Fundamentally, Prioré produced what to the observer seemed just a pulsed DC magnetic field, but one in which there was a specific "vacuum engine maker" of his deliberate creation and "tailoring." Originally by trial and error, he adjusted the 17 frequencies introduced to the plasma, thus varying the frequencies of the "infolded bidirectional pump waves" produced and infolded into the pulsed DC magnetic field carrier. In this manner he was "tuning the "vacuum antiengine-maker" content.

Again, Russian energetics weapons scientists call those vacuum engine templates the *information content of the field*. They do not refer to ordinary spectral analysis.

Explaining the Time-Pumping Vacuum Anti-Engine Maker

Irradiating the cell (with non-ionizing radiation) containing infolded specific bidirectional longitudinal waves, precisely corresponds to pumping the cells in the time domain, rather than pumping them in the "spatial energy distribution" domain. Because of the shift of the pumping from the 3-space domain to the time domain, this constitutes a dramatic extension of phase conjugate optics. Because of the amplified phase conjugating action of the "vacuum antiengine maker" upon the vacuum engine in a mass, this also constitutes a dramatic extension of general relativity and of four-wave mixing theory. Let us explain briefly:

When one pumps a nonlinear mass [phase conjugate mirror (PCM) material] with ordinary "transverse" EM waves { [92] } (Figure 27a), one is pumping (stressing, or "squeezing and relaxing") the mass with dynamic oscillating energy distributions in 3-space. One inputs a similar weak wave (called the "signal" wave) to the mass simultaneously. In that case, multiwave mixing theory tells us that the pumping of the PCM will cause the formation of a fourth EM wave, that is a time-reversed replica of the input "signal" EM wave. The time-reversed wave may contain up to all the energy in the pump waves, so it can be highly amplified in magnitude compared to the input signal wave. The distortion correction theorem tells us that this time-reversed replica wave will then appear precisely superposing in space back along the exact spatial track taken by the input signal wave { [93] }.

When the pumping is shifted to the time domain by pumping with bidirectional *longitudinal* EM waves, we now are pumping the time aspects of the mass—during it iterative existence as masstime; see again Figure 25—rather than just pumping the spatial energy aspects. In other words, we are not pumping *spatially*, but *spatiotemporally*. We are pumping *temporally*—rhythmically squeezing (compressing) and relaxing (extending) the "density of the time interactions" producing the rate of flow of time itself. This is a totally new extension to nonlinear phase conjugate optics.

In this case, the "signal wave input" to the "time-domain-pumped" nonlinear mass PCM is just the resident vacuum engine in the local spacetime in which that PCM mass is embedded (see Figure 27b). The sickened cell's abnormal vacuum engine—consisting of the sum of the normal cellular vacuum engine and the disease delta vacuum engine—is present in the cell and is the input to the time-pumped phase conjugate mirror (cellular masstime). The output is the formation of an amplified time-reversed replica of that vacuum engine. In other words, an exact, amplified *vacuum antiengine* for reversing the precise abnormal condition of the mass, down to and including even the quarks in the atomic nuclei, is formed by time-domain (longitudinal EM biwave) pumping. The result or "output" is that *the entire cellular PCM mass and all its constituent parts are "time-reversed" back along their previous "track through time", back to a previous normal physical state* { [94] }. Figure 27b illustrates this process.

So when Prioré pumped the abnormal cells, he simply time-reversed them—genetics and all—right back to their normal physical state before the abnormality arose.

The beauty of the method is that, by adjusting the input waves to the rotating plasma, Prioré could "tailor" and "tune" the pumping waves to what was needed for a particular class of disease. He used trial and error methods to arrive at the proper settings for the disease to be treated. In theory, by the Prioré method one can time reverse any mass back to an earlier physical state. This mechanism is used by the cellular control portion of the body's regenerative and restorative (R&R) system to restore damaged cells back to normal. It is the basis for the body's ability to heal itself!

Applied to diseased cells, the method causes the formation of specific amplified antiengines for the exact specific cellular abnormality, no matter how complex, and no matter what it is. *In applied medicine this is a revolutionary new healing and therapeutic modality*, accomplished by novel EM means, and justified by general relativity, higher symmetry electrodynamics, and longitudinal EM wave pumping in the time domain rather than classical electrodynamics and transverse EM wave pumping. Prioré's discovery, once understood, may just be the greatest medical discovery of all time, and the discoveries of Rife and of Becker may rank a close second. The Prioré process is shown in Figure 28.

The body's cellular regeneration system uses this same methodology and process. A block diagram of the operation of the body's cellular regeneration system is shown in Figure 29.

Further Principles Needed for Rife's Microscope

In his microscope (Figures 1 and 2), Rife just wanted to observe the templates [2] corresponding to the cells, germs, viruses, etc. and all their parts, without any smallness limit. To do that, we need to (1) amplify the tiny, tiny amplitudes (actually, in the virtual state and subquantal) of the tiny vacuum engines accompanying tiny spatial locations containing living, functioning entities (no matter how small), and (2) "transduce" that amplified result into the optical spectrum, so that we see a perfect optical analog and thus are "looking" visually at localized ST regions far smaller than the one-tenth wavelength optical resolution limit.

We need another principle to do that. Suppose that, to first order, we have an isotropic nonlinear medium. Then if we transmit a sine wave into it, pandemonium sets loose with our wave. The speed of each part of the wave in this medium is a function of the location of that part upon the wave amplitude, so the peaks move faster than the intermediate parts of the wave of lesser magnitude. The wave has "peak overtake", breakup, collapse, and all sorts of things happening to it. So a simple sine wave is not going to pass smoothly into that medium and "pass back out" coherently and nicely, and ordinary light is not going to give stable reflections from that medium!

However, if we transmit *two* sine waves simultaneously, some frequency interval apart, and then if we *pretend* that we actually transmitted their *difference* frequency, a very strange thing happens. The two individual sine waves themselves still break up and go bananas. However, that difference frequency acts as if it were a simple sine wave transmitted into a *linear* medium! So we get back out a very nice, smooth, well-behaved sine wave of the *difference* frequency, even though terrible things happened to the individual waves themselves. We can detect and use that "difference frequency" nicely.

Several very neat sonar systems presently use that sort of transmission schema in underwater work { [95] }. The principle applies for waves of all frequencies, for all media which can be approximated as nonlinear isotropic in nature. As we shall see, it applies to the Rife microscope also.

Rife's Secret: Recursive Coherent Increase of the Signal-to-Noise Ratio (SNR)

To first order, without too much error we can pretend that the molecules and atoms etc. in the cells and body tissues, together with their finer components, comprise such an isotropic nonlinear medium. That's not entirely true, of course, but it does work to an extent sufficient for our purposes. Fortunately for the Rife process, the parts that *do not* obey this isotropic rule add randomly, while the parts that *do* obey the rule add coherently. So the "signal-to-noise ratio" (SNR) increases with the increasing amplification, and one "grows" the desired signal right up out of the noise (SNR) of the parts that do not obey the isotropic medium assumption. With sufficient number of amplifier stages, in theory the SNR of even the tiniest parts can be made as large as desired { [96] }.

With optical functioning as in Rife's microscope, one may substitute "light gathering power" or "magnification" for the electrical term "amplifier gain".

As we stated, Whittaker in 1904 (and superpotential theory thereafter) shows us that—if we wish—any EM field, wave, and pattern can be represented by two scalar potentials. Each of these two scalar potentials is further comprised of bidirectional EM longitudinal waves, in a harmonic series, in the Whittaker 1903 manner as reinterpreted. So we can reduce any pattern, field, or wave from any assemblage of charge sources, into two potentials, and then further into two sets of bidirectional EM waves interfering with each other. That process already *makes* all conventional EM waves, fields, etc.

An EM structuring of the local vacuum-spacetime occurs by all the electrical charges in the mass of the living cell, down to and including the virtual charges exchanged in the nuclear interactions in the nucleus of the atoms. By Whittaker 1903 and superpotential theory, all of that structuring itself is a potential with an internal set of bidirectional harmonic longitudinal EM waves interfering with each other.

Further this potential is recursively organized, because the inner bidirectional EM wavepair structuring of the potential is already recursive, as can be shown by applying Whittaker-1903 and Whittaker-1904 successively and iteratively { [97] }. This means we get deeper and deeper internested levels of patterns that still precisely correspond to the operational mechanisms of the lower levels. It's similar to an infinitely fractal { [98] } type of process.

Because this conglomerate is recursive as a function of harmonic intervals, it means that we can pick any harmonic interval we choose in the "big world" where we observe things, and we can have an exact observational representation of the dynamics of that ongoing hidden longitudinal EM wavepair interferometry in the two waves bounding that harmonic interval. Of course, the more additional harmonics we add, the better dynamics we have. The greater the number of harmonic intervals applied, the greater the gain. However, these harmonic intervals can be recursively applied using just one "macroscopic" harmonic interval, iteratively and serially applied over and over. The more iterative serial recursive application of one harmonic interval, the mote the "gain" or "amplification"—in the case of a microscope, the more the *amplification*. That was Rife's real secret, finally revealed. See Figure 30.

The Difference Frequency Provides Visible Output

Rife wanted to look at very, very small things *visually*, even though those “small things” were far below the one-tenth wavelength visible light barrier. In other words, to understand the Rife microscope’s visual representation of the virtual state, we need to choose the harmonic interval of the throughput so as to span the visible spectrum. Well, for optical purposes, luckily the IR and UV do exactly that. Examine the frequencies, and you will see that a harmonic interval exists from a spot in the IR to a spot in the UV, and of course the interval in between includes the entire visible spectrum. The visible spectrum in between, however, represents the observable noise that would block what we seek to have emerge in that region in our "Rife microscope", so the presence of the visible spectrum is highly undesirable. Various filters can be used to block the visible spectrum while transmitting the UV and IR regions, or one can just work in the dark or in suitably dim visible light.

Rife then applied his iterative multiple "difference frequency" lenses in serial order, one after the other. While it is doubtful that he understood the theoretical method being utilized, he certainly persisted and found how to use it. This method of iteratively summing a recursive virtual state vacuum engine "template" results in the formation of an observable state vacuum engine template. In short, it magnifies the virtual state entities and dynamics directly to observable state analogues that capture all the virtual state entities and their dynamics.

That is why, if we use special filters on your camera and flash lenses, where the filters are open in the IR and UV but opaque in the visible spectrum in between, we can often photograph the ongoing "longitudinal EM hidden biwave interferometry"—*the subtle energy or vacuum engine functioning*—of nature and of living systems, *even displaced in time*, as Trevor Constable { [99] } and Joe Gambill { [100] } (a researcher friend now deceased) demonstrated, as George Meek { [101] } demonstrated, and as Fogal { [102] } has successfully experimented with.

Occasionally when one gets things just right, one can get pictures of the subtle invisible reality in the virtual state, going on around oneself and in one's surroundings. With Rife-like iterative serial amplification, one can get them all the time, dependably, as with a Fogal semiconductor specially adapted for such operation. One can in fact build instruments that indirectly observe the so-called "unobservable" virtual energy state and its entities and dynamics with one-to-one correspondence.

That's a whole new "subtle physics" reality. Present physics with its assumed quantum observation limitations just has not even gone into it—primarily because physics has been focusing upon the quantized aspects of the photon (which is quantized in one aspect only—its magnetic vector potential—and not in its unquantized aspect—its scalar electrical potential!). Physics has rather ignored the unquantized (and therefore continuous from the smallest to the largest scale) time aspect of the photon { [103] }. Using the continuous aspect of the photon, one can assemble photon “time interval” structures—and in fact one does so whenever one assembles a 3-spatial energy density structuring such as in ordinary EM theory. Simultaneously, one has assembled a “structuring of the time domain” also, because each photon in the spatial energy density structuring consists of both time and energy as its canonical components.

All EM fields and waves, being comprised of photons, thus also carry (1) energy and energy internal structuring, and (2) time and internal time structuring. Again, this is the Russian *information content of the field*. The Russians simply understand 4-vector theory and Whittaker 1903 and 1904, and how to combine them with quantum mechanics and general relativity in a unified field theory.

The above explanation is why Kaznacheyev's { [104] } mitogenetic radiation { [105] } from his control (deliberately damaged or infected) cells would pass through a quartz window and induce the same damage (or disease) in targeted cells on the other side and otherwise environmentally shielded, but would not induce the changes through a normal window glass window. (See again Figure 23) Quartz passes the full optical spectrum from IR to UV, while window glass does not. So quartz passes the *vacuum engines* for the disease or damage, while window glass breaks them up and does not pass them as such. We stress that the vacuum engines are part of the “continuous” aspect of the photons, and so are not bound to any limitation on magnitude (either fineness or largeness) as are the quantized aspects of the photon.

For such a "far beyond the one-tenth wavelength" microscope as Rife’s, one must use vacuum engines (time domain structuring, and longitudinal EM waves) and extraordinary iterative unified field theory optics, rather than the ordinary transverse EM waves and ordinary optics. The lenses must all pass at least one harmonic interval (and the same one!), to provide and utilize the “difference frequency band” needed for recursive *visual* connection to the subquantal state (the hidden virtual structures in terms of the continuity aspects of photons). Successive stages thus "serial iteration" use the recursive principle to coherently amplify and show the vacuum engines, their dynamics, and the dynamics of deeper organizations and dynamics in the virtual state itself. Since Rife wished to “see” the results in the optical spectrum, then the two frequencies chosen for “differencing” had to bound the visible spectrum.

Rife used quartz lenses or at least very good optical glass that is open from IR to UV. That bounds the visible spectrum with a harmonic interval. Rife used multiple stages for successive coherent recursive amplification and increasing the SNR magnitude by coherent integration. One must also have some scheme for filtering out (or avoiding) too much visible spectrum light. One such scheme is to just use enough stages so that, in the new theory where the wavelength limitation is bypassed, one has amplified ("resolved") to the depth or fineness desired. The process long ago removed itself from the "normal visible light" domain, well before it reached the deeper resolution limit utilized by the definition state of the new method.

By using sufficient multiple stages in this fashion, the "tiny" virtual state vacuum engine templates passing through the IR-UV harmonic interval in a lens can be recursively summed and magnified many, many times—essentially without limit, because the phenomenon being magnified is recursive and bidirectional longitudinal in nature. It is definitely *not* normal visible light optics, and it *is not* conventional classical electrodynamics.

Rife’s Recursive “Vacuum Engine” Microscope

The Rife microscope coherently sums (and thus coherently amplifies) the virtual state vacuum engine patterns with the increasing number of stages. For the *subtle* vacuum engine templates, such an arrangement becomes a direct amplifier (magnifier), for amplifying (magnifying) the internested virtual state vacuum engine templates and their dynamic functioning, into the observable state. With a properly designed “vacuum engine” recursive microscope, one can see inside a molecule, inside the atomic nucleus, and even inside a nucleon, given development and perfection of the process. However, at this initial stage of understanding, one can forecast that finely adjusting the various stages of an early laboratory prototype would indeed prove to be a tedious and formidable task, just as Rife reported. With further development, however, that obstacle can be removed.

What comes out the output end of such a “vacuum engine” recursive microscope, is a direct visible light *analogue* of the much more subtle ST functioning in the microscope’s highly localized viewing area. This output imaging is put together "electrically" in ways analogous to the construction of a video picture on the screen of a TV tube. Since the result emerges through an output end that still passes an entire harmonic interval bounding the visible spectrum, the amplified virtual state recursive template is not destroyed but is highly amplified and produced as an observable image and dynamics in the visible spectrum. This final "interferometry" of two *internally structured* potential functions now produces "visible light" EM fields and waves and patterns, in accord with Whittaker 1904. These visible patterns are direct analogues representing exactly the very, very minute regions far below the one-tenth wavelength limitation of ordinary optical resolution. Those regions were the regions “being focused onto” when Rife so arduously “adjusted” his microscope's multiple stages.

With Rife’s recursive magnifying microscope process, one can see direct visible analogs of what is actually far, far smaller than anything ever before seen in normal optical microscopes, or even in electron microscopes etc. Indeed, one can see directly into the virtual state itself, and observe virtual state structures. While this is unthinkable in quantum mechanics if one considers only quantized aspects of light, it is perfectly feasible if one considers the concomitant unquantized (continuous) aspects of light, and shifts to general relativity and vacuum engines to utilize the unquantized structuring of the time domain.

Conclusion

Rife unquestionably found a way to produce visible recursive amplification and magnification of invisible virtual state dynamics and functions. We hope we have shed at least some light on the extraordinary technical mechanisms involved. This is good conceptual science, consistent with theory and some known experiments, and it certainly can be made a rigorous science and technology and mathematically modeled, using higher symmetry electrodynamics that is part of a unified field theory such as that of Sachs. The basis for all of it is in the literature already, but it is scattered widely in uncorrelated bits and pieces. It is not in the normal electrodynamics courses, nor is it in the normal optical science curriculum. And it is not taught in the normal physics curriculum. Needless to add, it is absent from the medical curriculum as well.

As an additional proposal, we tentatively suggest that this principle also may be the long-sought "hidden chaos" or "hidden order" in quantum mechanics that is known to be present, but has not yet been explained { [106] }. If so, then it becomes a new law of physics, with the experimental proof first demonstrated by Royal Raymond Rife.

With the conceptual technical matrix of Rife's mechanism in hand, it should be possible for a proper scientific multi-disciplinary team to redevelop and extend the Rife microscope. Marvelous new instruments will emerge, once the quantum physics conceptual blockage of continuous physics is constrained to its place as only “one view of the physics reality elephant.” With the ability to directly “measure”, “observe,” and engineer the virtual state, a breathtaking expansion of physics will occur, dramatically extending the vacuum engineering suggested by Nobelist Lee.

If so, then Rife's tragic life will have served its noble purpose, and his intuitive insight will have been resoundingly verified. The follow-on results will have untold benefit to science and to all the human generations to come.

NOTES AND REFERENCES

---

[1] For work on the theoretical aspects of this paper, partial financial support by A. Dreyer, R. Francis, H. Ledger, M. Peters, T. Craddock, and the Association of Distinguished American Scientists (ADAS) is gratefully acknowledged, as are helpful discussions with Dr. Bob Flowers and Dr. Myron Evans.

[2] We emphasize that Rife himself did not think in these terms.

---

[1] . (a) Some Rife-related references are: Royal R. Rife, "The filterable virus of carcinoma," Apr. 26, 1939; — "The Rife Microscopes," Apr. 28, 1939; — "Bacteria," Apr. 28, 1939; — "History of the development of a successful treatment for cancer and other viruses, bacteria, and fungi," Allied Industries, Dec. 1, 1953.

(b) Arthur Isaac Kendall and Royal Raymond Rife, "Observations on Bacillus Typhosus in its filterable state," California and Western Medicine, Ca. Med. Assoc., Dec. 1931.

(c) Arthur Isaac Kendall, "The filtration of bacteria," Science, Mar. 18, 1932.

(d) See particularly E. C. Rosenow, “Observations on filter-passing forms…,” Proc. Staff Meeting of the Mayo Clinic, July 13, 1932; — “Observations with the Rife microscope of Filter-passing forms of micro-organisms,” Science, Vol. 76, Aug. 26, 1932, p. 192-193; — "Microdiplococci in filtrates of natural and experimental poliomyelitic virus compared under the electron and light microscopes," Proc. Staff Meeting of the Mayo Clinic, 17(7), Feb. 18, 1942, p. 99-106; — "Observations of filter-passing forms of Eberthella Typhi (Bacillus Typhosus) and of the Streptococcus from Poliomyelitis," Proc. Staff Mtg.

Mayo Clinic 7, July 13, 1932, p. 408-413; — "Transmutations within the Streptococcus-Pneumonococcus Group," J. Infective Diseases, Vol. 14, 1941.

(e) R. E. Seidel and M. Elizabeth Winter, "The new microscopes," Journal of the Franklin Institute, Feb. 1944, p. 103-129; — "Filterable bodies seen with the Rife microscope," Science Supplement, Science, Dec. 11, 1932, p. 10-11.

(f) "Is a new field about to be opened in the science of bacteriology?", Editorial, California and Western Medicine, Ca. Med. Assoc., Dec. 1931; "Powerful microscope makes very tiny objects visible," Science Newsletter, Vol. 33, Jan. 22, 1938, p. 55; "The Rife microscope or 'facts and their fate'," Reprint No. 4l7, The Lee Foundation for Nutritional Research, Milwaukee, Wisconsin, undated.

(g) A good lay overview is given by Christopher Bird, "What has become of the Rife microscope?", New Age Journal, Mar. 1976.

(h) See also T. E. Bearden, AIDS: Biological Warfare, Tesla Book Co., 1988, p. 267-279. The author gives a short introduction to the Rife microscope and includes some pictures. The proposed evanescent wave explanation proved to be wrong. Much additional work by the author eventually culminated in the proposed mechanism presented in the present paper.

[2] . We have one viewgraph listing some 32 flaws, errors, and shortcomings of classical electrodynamics, and there are even more than that. Additional errors are listed, for example, by important scientists such as Feynman, Wheeler, Evans, Bunge, Cornille, Barrett, and others. So a substantial number of errors in classical U(1) EM theory are well known to foundations scientists. We discuss only a minimal few in this paper.

[3] . Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176.

[4] . *Ibid*., p. 182.

[5] . M. Faraday, Experimental Researches in Electricity and Magnetism. Vol. 1, Taylor and Francis, London, 1839.; Vol. 2, Richard & John E. Taylor, London, 1844; Vol. 3, Taylor and Francis, London, 1855. Reprinted by Dover in 1965. Also see M. Faraday, "On Static Electrical Inductive Action," Phil. Mag., 1843; — "Remarks on Static Induction," Proc. Roy. Inst., Feb. 12, 1858.

[6] . James Clerk Maxwell, Preface to the First Edition, A Treatise on Electricity and Magnetism, third edition, Vol. 1, Dover Publications, New York, 1954, p. viii-ix.

[7] . J. C. Maxwell, Preface to the First Edition, A Treatise on Electricity and Magnetism, third edition, Vol. 1, Dover Publications, New York, 1954, p. viii-ix. Quoting: *"…before I began the study of electricity I resolved to read no mathematics on the subject till I had first read through* Faraday's Experimental Researches in Electricity. *…I perceived that his method was also a mathematical one, though not exhibited in the conventional form of mathematical symbols. I also found that these methods were capable of being expressed in the ordinary mathematical forms…."*

[8] . James Clerk Maxwell, A Treatise on Electricity and Magnetism, Third Edition, Vol. 1, Dover Publications, NY, 1954, p. 165-166.

[9] . Else the precessing electrons in a signal-receiving wire antenna would not precess laterally. Since their precession is at right angles to the disturbing force, the disturbance arriving from the vacuum was a longitudinal disturbance. This also agrees with Whittaker's 1903 decomposition of the scalar potential, and Mandl and Shaw's note that the combined scalar photon and longitudinal photon are observable as the instantaneous scalar potential,

while neither of them is observable individually. See citations for Whittaker and for Mandl and Shaw elsewhere in these references.

[10] . This word "potentiality" (by Feynman) is given flesh and blood in the present paper. What actually exists in the vacuum is potential and structures of potential. Vacuum, spacetime, and potential are identities. Hence we may say that the structured vacuum potentials are indeed structurings of spacetime curvature. We refer to these as *vacuum engines*, which are explained in this paper.

[11] . Richard P. Feynman, Robert B. Leighton and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol. II, p. 1-3.

[12] . This is easily shown. $F \equiv d/dt(mv)$, which expands to $F = (m)dv/dt + v(dm/dt)$. Let the mass m be zero continuously, and both terms on the right go to zero, hence $F = 0$. That is, the force disappears because mass is a component of force. It is the use in mechanics of the separate massless force acting on a mass that is the non sequitur and to blame for this giant error in electrodynamics as a hold-over from the hoary old mechanics theory.

[13] . J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 249.

[14] . Jackson, ibid. p. 28.

[15] . Melba Phillips, "Classical Electrodynamics," Vol. IV: Principles of Electrodynamics and Relativity, in Encyclopedia of Physics, Edited by S. Flugge, Springer-Verlag, Berlin, 1962, p. 4. Quoting: "The utility for electrical problems of a scalar potential, first introduced in the theory of gravitation, was pointed out by Simeon Denis Poisson, Bull. Soc.

Philomatheique **3**, 388(1813). See also Mem. Institut, Pt. **1**, 1, Pt. **2**, 163 (1811). The word *potential* was introduced by George Green in 1828."

[16] . E.g., see W. Misner, K. S. Thorne, and J. A. Wheeler, Gravitation, W. H. Freeman and Co., San Francisco, 1973, p. 5.

[17] . Two methods of dealing with this problem, and rectifying it, are the use of a topological definition of the fields in mass-free spacetime, and the use of a unified field theory where the "field in space" is not a forcefield per se, but merely a curvature or set of curvatures of spacetime. As an example, the O(3) electrodynamics largely pioneered by Evans takes the latter approach, and it is usable in directly engineering unified field theory effects.

[18] . (a). Mendel Sachs, The Field Concept in Contemporary Science, Charles C. Thomas Publishers, 1973;

(b) — General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics), Reidel (now Kluwer), 1982 provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also

(c) Mendel Sachs, Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia, Reidel (now Kluwer), 1986;

(d) — "Relativistic Implications in Electromagnetic Field Theory," in T. W. Barrett and D. M. Grimes, eds., Advanced Electromagnetism, World Scientific, 1995, p. 541-559. In the latter, Sachs shows that the most general expression for the field theory is in terms of spinor and quaternion variables, rather than the vector and tensor variables of the conventional expression of Maxwell's theory. This generalized expression leads to extra conservation laws and invariants, thus increasing the predictive capacity of the theory.

(e) For a marvelous treatise for the educated layman, see M. Sachs, Relativity in Our Time: From Physics to Human Relations, Taylor & Francis, 1993.

[19] . E.g., see the several Evans' reviews in Modern Nonlinear Optics, Second Edition, Edited by M.W. Evans, Wiley, 2002 (in press). Many other O(3) electrodynamics papers are also carried in this set of three volumes, as well as in the standard literature.

[20] . Mendel Sachs, "Relativistic Implications in Electromagnetic Field Theory," ibid.

[21] . Without interaction with the vacuum, mass does not exhibit perfect symmetry. For a discussion, see T. D. Lee, Particle Physics and Introduction to Field Theory, Harwood Academic Publishers, New York, p. 378-389.

[22] . In both classical and quantal electrodynamics, the problem of the source charge and its associated fields and potentials—and their energy—has been considered to be the most difficult problem.

(a) For a solution to that problem, see T. E. Bearden, "Giant Negentropy from the Common Dipole," Journal of New Energy, 5(1), Summer 2000, p. 11-23. On DoE open website http://www.ott.doe.gov/electromagnetic/papersbooks.html and this author's website www.cheniere.org.

(b) For very powerful support of that proposed solution, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, Chapter 5. Neither the time-polarized photon nor the longitudinal photon is individually observable, yet the combination of the two is observable as the instantaneous scalar potential. The Mandl and Shaw combination of a scalar (time-polarized) photon and a longitudinal photon is explained by Bearden as the interaction of the causal incoming scalar photon with a charge, the charge subsequently transducing the absorbed energy from the time-domain into 3-space, and re-emission of the energy as an emitted longitudinal photon in 3-space. The actual interaction provides the coupling, and clarifies which is the *cause* and which is the *effect*.

[23]. As the electrodynamicists implicitly assume, by assuming fields and potentials from all charges in the universe penetrating and existing and superposing and interfering in each cubic centimeter of space, and in fact at every point in space. But then the electrodynamicists continue to utilize a material ether assumption in their equations, embedded implicitly upon the notion of an "empty space" filled with this material ether. Tesla called it "*… one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history*." (Nikola Tesla, "The True Wireless," Electrical Experimenter, May 1919). While that biting remark certainly did not endear him to struggling electrical scientists, it did strike uncomfortably close to the mark, and it still does.

[24]. J. H. Poynting, "On the transfer of energy in the electromagnetic field," Phil. Trans. Roy. Soc. Lond., Vol. 175, Part II, 1885, p. 343-361.

[25]. Oliver Heaviside, "Electromagnetic Induction and Its Propagation," The Electrician, 1885, 1886, 1887, and later. This is a series of 47 sections, published section by section in numerous issues of The Electrician during 1885, 1886, and 1887. See also Oliver Heaviside, "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field," Phil. Trans. Roy. Soc. London, 183A, 1893, p. 423-480.

[26] . By a crude analogy, one may think of the "virtual" energy of the vacuum as energy that has been disintegrated into incredibly tiny bits—each so small that the bits fade in and out of existence at an incredible rate. The spinning action of the charge may be thought of as similar to that of an old spinning wheel. The charge spins, so that it coheres some of the "virtual fibers" continually fed to it by the vacuum, into "thread-sized" portions that are emitted radially outward. So the charge "splatters back" some of the fibers as fibers, but "spins" the remainder into useful quantum-sized "threads" which are emitted in observable form.

[27] . Presently electrodynamicists are likely to specify the Poynting energy flow **S** as **S** = **E**×**H**. However, in the vacuum (spacetime) neither **E** nor **H** exists as such *observed force fields*, which requires presence of charged mass. Only potentials and potential structuring of the virtual particle flux exist in the vacuum (which is after all just potential comprised of virtual particle flux). Those structurings or their analogues also exist as sets of curved spacetime regions (i.e., spacetime structurings comprising the elements of the "vacuum engine"). See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355. Whittaker that any scalar potential decomposes into a multiplicity of longitudinal bidirectional flows of energy. This decomposition must also comprise the vacuum scalar potential, as well as the "spacetime curvature" potential.

[28] . However, Whittaker misinterpreted the *causal* phase conjugate wave (that half of his phase conjugate wavepair) as the *effect* after the wave interacts with charged mass, in which case parity is broken and the *effect* wave produced by the interaction is a longitudinal EM wave going in the opposite direction in 3-space from its "biwave" 3-space partner. The correct interpretation is that,

prior to interaction, the phase conjugate wave exists in the fourth axis of Minkowski 4-space, hence is in the time domain. It is literally a longitudinal EM wave moving on the fourth axis (in the time domain) to the interacting charge, and the other Whittaker biwave partner is the effect longitudinal EM wave emitted in 3-space by the interacting charge after it absorbs the incoming time-polarized EM wave. We point out that a longitudinal EM wave along the time axis is a time-polarized EM wave, corresponding to the known time-polarized photon in quantum field theory.

[29] . After publication of my "Giant Negentropy" paper, I discovered very strong support for my reinterpretation of Whittaker's 1903 decomposition of the scalar potential. This support was by F. Mandl and G. Shaw, ibid. Mandl and Shaw give a deeper coverage of the photon polarizations. They argue that the longitudinal and scalar polarizations individually are not directly observable, but only in combination, wherein they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, directly fits my re-interpretation of Whittaker's 1903 decomposition of the scalar potential. However, Mandl and Shaw fail to account for their assumed interaction of the detecting/observing unit point charge to perform the perfectly correlated "combining", and thus fail to account for the absorption of the incoming *causal* time-polarized wave or photon, the transduction of that excitation energy of the charge into longitudinal EM wave/photon energy in 3-space, and the subsequent emission of that excitation energy as a perfectly correlated longitudinal EM wave in 3-space.

Thus Mandl and Shaw missed the time-excitation charging of the source charge via absorption of the "coupled" time-polarized EM wave/photon, and the subsequent decay of the time-charge excited state by emission of a 3-space longitudinal EM wave/photon. This omitted interaction represents the discovery of nature's giant negentropy mechanism in EM energy flow, whenever 3-space energy flow symmetry is broken, as with a common dipole. So Mandl and Shaw do not account for photon (or wave) polarization transduction, where the "causal" time-polarized EM wave or photon comes in and is absorbed by the detecting charge or dipole to energetically excite it in the time-domain, then the excitation energy is re-emitted as the longitudinally polarized EM wave or photon in 3-space.

Recognition of these missing functions allowed at last a solution to the long-vexing problem of the source charge. We actually solved the problem for a source dipole, then pointed out that any "isolated" charge is a set of dipoles when its clustering virtual charges of opposite sign are considered.

[30] . The source charge problem was recognized as the most difficult problem in electrodynamics. E.g., see D. K. Sen, Fields and/or Particles, Academic Press, London and New York, 1968, p. viii. Quoting: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."* Since our proposed solution is consistent with particle physics (broken 3-symmetry of the dipole) and with quantum field theory (combination of the time-polarized and longitudinal photon comprises the scalar EM potential), it is robust and we believe it will stand up to meticulous scrutiny.

[31] . E. T. Whittaker, 1903, *ibid*.

[32] . Mandl and Shaw, *ibid*.

[33] . Poynting only assumed that diverged part of the energy flow. Heaviside, on the other hand, showed that there is a much larger energy

flow component that misses the intercepting charges (or circuit) altogether, and is wasted. Since this means that far more EM energy flows from the terminals of a generator than the external circuit intercepts and catches, it means that all generators produce more energy flow output than the textbook states, or that Poynting assumed. In fact, as *transducers of vacuum energy*, all dipolar generators already exhibit COP>1.0. Since no one in the 1880s could explain where such an astounding extra energy flow was coming from, Lorentz originated an integration trick to discard the nondiverged Heaviside EM energy flow component.

(a) See J. H. Poynting, "On the transfer of energy in the electromagnetic field," Philosophical Transactions of the Royal Society of London, Vol. 175, Part II, 1885, p. 343-361.

(b) See Oliver Heaviside, "Electromagnetic Induction and Its Propagation," The Electrician, 1885, 1886, 1887, and later—a series of 47 sections, published section by section in numerous issues of The Electrician during 1885, 1886, and 1887.

(c) See also Oliver Heaviside, "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field," Phil. Trans. Roy. Soc. London, 183A, 1893, p. 423-480.

(d) For Lorentz' trick to discard the nondiverged component, see H. A. Lorentz, Vorlesungen über Theoretische Physik an der Universität Leiden, Vol. V, Die Maxwellsche Theorie (1900-1902), Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. Figure 25 on p. 185 shows the Lorentz concept of integrating the Poynting vector around a closed cylindrical surface surrounding a volumetric element. This is the procedure which arbitrarily selects only a small component of the energy flow associated with a circuit—specifically, the small Poynting component striking the surface charges and being diverged into the circuit to power it—and then treats that tiny component as the "entire" Poynting energy flow. Thereby Lorentz arbitrarily discarded all that extra Heaviside energy transport component which does not strike the circuit at all, and is just wasted.

[34] . In fact, the simple equation $W = \phi q$, where W is the diverged/collected energy, q is the intercepting coulomb, and ϕ is the reaction cross section of the potential, shows that electrodynamicists implicitly assume the “true” causal potential to contain unlimited energy. E.g., using that equation, any desired amount W of energy can be collected from any nonzero potential ϕ, no matter how small its intensity, simply by adding sufficient intercepting/collecting charges.

[35] . A back-of-the-envelope calculation for a nominal simple circuit has been made and previously presented by the present author. In the nominal circuit used, the source dipole extracts from the vacuum and sends down the outside of the attached circuit about 10^{13} times as much energy as the circuit itself intercepts, diverges, and utilizes. The huge remainder is not intercepted, remains nondivergent and just passes on out into space at the speed of light.

[36] . Solving the standard equation $\mathbf{E} = \mathbf{F}/q$ for $\mathbf{F}$, one has $\mathbf{F} = \mathbf{E}q$. Again, from any nonzero **E**-field, no matter how small, as much force can be “gathered” and produced as one desires, merely by increasing the intercepting/collecting charge q.

[37] . Robert H. Romer, "Heat is not a noun," American Journal of Physics, 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of AJP of the concept of heat in thermodynamics. Heat is not a substance, not a thermodynamic function of state, and should not be used as a noun. The quotation castigating the standard "field diagram" of an EM wave in space is

from endnote 24, p. 109.

[38] . This is readily understood when one realizes that the founders of electromagnetics assumed a material ether. So to them, there was not a single point in all the universe where mass was absent! Hence the field, potential, and wave "in space" were actually the field, potential, and wave "in the material ether medium", and therefore perfectly material. Even though the Michelson-Morley experiments falsified the material ether in the 1880s, the electrodynamicists (i) never changed their classical EM equations which still to this day assume the material ether filling all space, and (ii) still implicitly mistake the transverse electron precession wave in the Drude gas in a receiving antenna as representing the assumed interception of the incoming "transverse EM wave" in that external material ether medium.

[39] . Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, p. 283-287, and passim.

[40] . Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), 1959, p. 485-491. However, Aharonov and Bohm seem to have missed the similar mass interpretation of the potential, so that they utilized the potential as massless and therefore the vacuum EM entity that is the primary cause of all EM phenomena. If one truly eliminates only the potential’s effect upon the intercepting charged mass, and utilizes the non-intercepted EM entity in vacuum, one has arrived at the direct alteration of the massless spacetime geometry of general relativity. In short, suddenly one has latched onto a unified field theory, and one that is engineerable by surprising electromagnetic means.

[41] . But erroneously assumed to be present, because of Maxwell’s ubiquitous use of the material ether and the continued implied material ether in the electrodynamics equations to this day.

[42] . T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, 16(1), 1991, p. 23-41. Barrett analyzed Tesla's patented circuits in terms of quaternion electrodynamics, which is of higher symmetry than either vector or tensor algebra electrodynamics. Lo and behold, Tesla could shuttle potential energy around in his circuits at will, and that functioning would not even show up in a tensor or vector electrodynamic analysis.

[43] . T. W. Barrett, "Active Signalling Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996; — Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic Topological Forms and RF Phase Conjugation," U.S. Patent No. 5,493,691, Feb. 20, 1996.

[44] . As the reader can appreciate, this represents a dramatic change to the foundations of electrodynamics. It is closely allied to Sachs's unified field theory; see appropriate Sachs references in this paper.

[45] . As Dirac pointed out, the Michelson-Morley experiments of the 1880s destroyed the *material* ether assumed by Faraday and Maxwell and subsequent electrodynamicists; it *did not* destroy the ether per se. Dirac emphasized that any nonmaterial Lorentz-invariant ether is perfectly consistent with the Michelson-Morley experiments. General relativity describes the shaping and structuring of the vacuum as curvatures of spacetime geometry and clustered patterns thereof (although the latter part is essentially neglected). GR consists of the changes

of this spacetime due to the interaction of mass/ trapped energy, and also the forces upon mass/ trapped energy created by the interaction of such curvature-patterns of spacetime. Essentially, electrodynamics is missing just this "mass to spacetime" transform and this "spacetime to mass" transform. It is in fact missing its "infolded general relativity" dynamic internal structuring of the potential, field, and wave. KGB energetics weapon scientists call these missing transforms *the information content of the field*—which our fellows mistakenly interpret in their own flawed electrodynamics as "spectral content." Nothing could be further from the truth.

[46] . In 1903 E. T. Whittaker showed that any scalar potential is comprised of an infinite harmonic series of bidirectional longitudinal EM waves, arranged in bidirectional wavepairs. See E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, 1903, p. 333-355.

[47] . General relativity usually stops short of including an "infolded" topology within its basic spacetime, and only applies an "exterior" topology to it. However, it does not exclude that capability. Spacetime is just potential identically, and also just vacuum identically. Since GR omits the infolding structure of spacetime/potential, and so does electrodynamics, it has not been possible to unify the two completely. The "bridge" across the chasm between them has been eliminated by both disciplines.

[48] . With finite propagation, this action-at-a-distance is retarded so that some time passes between the effect and the cause. Each Mandl-Shaw photon interaction—the receipt of the scalar (time-polarized) photon, and emission of 3-space longitudinal photon—is retarded by (requires) the amount of time that is transformed into 3-space energy. [Time may be regarded as 3-space

electromagnetic energy compressed by the factor c^2, with the compressed energy placed on the fourth Minkowski axis to appear as time (seconds), rather than being left in 3-space to appear as mass).

Transformation from time-domain EM energy to 3-space energy is a function of the spin of the charge, which "spins" 720 degrees, being 360 degrees in the time domain while absorbing the time-polarized EM photon energy, the being followed by 360 degrees in 3-space while it is emitting the 3-space longitudinal photon. This retardation due to transformation of the energy from the time domain to the 3-space domain is what limits the speed of light in the ambient vacuum to the finite velocity c.

However, when infinite velocity propagation occurs (as is permissible with longitudinal EM waves alone, which a priori implies no retardation), then the notion vanishes of a “cause” and an “effect” separated by a time interval. In that case, one is operating in a multiply-connected spacetime, where in theory there is no observable “length” or “time” between the “cause” event and the “effect” event. Further, one “causative” initiation point can be superposed simultaneously with multiple distant points widely separated in ordinary spacetime. In that case, putting a joule of energy into the “initiating cause” point (to the normal 3-space observer) results in the simultaneous appearance at each and every distant participating point, of one joule. To the normal "singly-connected spacetime" observer, this is a direct amplification of energy, directly from the vacuum energy itself. We have proposed this mechanism as an extension of Bohm’s quantum potential utilized in his hidden variable theory of quantum mechanics. See David Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," Physical Review, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Unfortunately three nations of the world have already seen fit to highly weaponize this effect, and a fourth nation is very close to deploying such a weapon. Presently quantum potential weapons are the dominant weapons on earth, and nothing can stand against them.

[49] . By conjugate pair, we mean that the wavepair consists of an EM longitudinal wave and its time-reversed twin, precisely superposed spatially but antiphased in time. In such a wave, the transverse energy density amplitude cannot vary; instead, the wave consists of (and carries) an oscillation of the rate of flow of time, about some median speed. The median speed of a longitudinal EM wave is not confined to the speed of light in vacuum.

[50] . The value of the energy continuously collected from a steady potential, in a "swirl" around a fixed intercepting charge, is of course a fixed scalar. That, however, has nothing to do with the nature of the potential as an EM entity in massless vacuum, prior to interaction with charged mass.

[51] . J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 223.

[52] . Interestingly, one also cannot produce a clear definition of *energy* in physics. A good discussion of the energy concept is given by R. B. Lindsay, "The concept of energy and its early historical development." Foundations of Physics, 1 (4), 1971, p. 383-393. Lindsay investigates the concept of energy from its early historical origin, and covers from ancient times through the 18th century. The root of the concept is the notion of invariance in the midst of change. In short, the concept consists of considering "the changing" as being "nonchanging." In pursuing a rigorous definition of anything, one eventually arrives at a contradiction of Aristotelian logic. Ultimately one can only define something as identical to its own opposite. Geometry, e.g., models the presence of

mass as the absence of mass. Ultimately a "hard definition" is an "accursed necessity for the identity of opposites"—the centuries-old bane of the philosophers. In physics the problem is avoided by the "duality principle"—just treat an entity as either a particle or wave as necessary, and quit arguing about it! The problem is the incompleteness of Aristotelian logic itself. A solution to the "identity of opposites" problem is obtained by explicitly adding a fourth law of logic that has always been implicitly assumed in Aristotle's three laws. For a discussion, see "A Conditional Criterion For Identity, Leading to a Fourth Law of Logic," Appendix III to T. E. Bearden, Aids: Biological Warfare, Tesla Book Co., 1988, p. 428-443. A simple proof of the fourth law of logic is given in *ibid*., p. 439-440.

[53] . In addition to reference 19, see Myron W. Evans, Enigmatic Photon, Vol. II, Kluwer, ISBN 0792332881. See also Apeiron, 4(2-3), Apr.-Jul. 1997, a Special $B^{(3)}$ Field Issue, Guest editor: Myron W. Evans. In this same issue, see the Introduction by Evans, and also his "Helicity and the Electromagnetic Field." In the same issue, see also Jean-Pierre Vigier, "Relativistic Interpretation (with Non-Zero Photon Mass) of the Small Ether Drift Velocity Detected by Michelson, Morley, and Miller." Barrett has also rigorously shown the reality of the $B^{(3)}$ field, and so the conventional closure against it is doomed to failure. The emergence of $B^{(3)}$ field theory will dramatically alter electrodynamics, and elevate the importance of the longitudinal EM wave. We are also much indebted to Bob Flowers for his penetrating elucidation of the background of the $B^{(3)}$ field in private communications, and to Myron Evans for his patient and encouraging private communications.

[54] . See particularly T. W. Barrett, "Sagnac Effect: A Consequence of Conservation of Action Due to Gauge Field Global Conformal Invariance in a Multiply-Joined Topology of Coherent Fields," in T. W. Barrett and D. M Grimes, [Eds.], Advanced Electromagnetism: Foundations, Theory,

& Applications. World Scientific, River Edge, New Jersey, 07661, 1995, p. 278-313; T. W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in A. Lakhtakia, (ed.): Essays on the Formal Aspects of Electromagnetic Theory, World Scientific Publishing, River Edge, NJ, 1993, p. 6-86.

[55] . Richard W. Ziolkowski, "Exact Solutions of the Wave Equation With Complex Source Locations," Journal of Mathematical Physics, 26 (4), April 1985, p. 861-863; I.M. Besieris, A.M. Shaarawi, and R. W. Ziolkowski, "A bidirectional travelling plane wave representation of exact solutions of the scalar wave equation," Journal of Mathematical Physics, 30(6), 1989, p. 1254-1269; Rod Donnelly and Richard Ziolkowski, "A method for constructing solutions of homogeneous partial differential equations: localized waves," Proceedings of the Royal Society of London A., Vol. 437, 1992, p. 673-692; — "Electromagnetic field generated by a moving point charge: A fields-only approach," American Journal of Physics, 62 (10), Oct. 1994, p. 916-922.

[56] . M. W. Evans, editor, Modern Nonlinear Optics, Wiley, 2002, 3 volumes (in press).

[57] . W. A. Rodrigues Jr. and J.-Y. Lu, “On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature,” Foundations of Physics, 27(3), 1997, p. 435-508; W. A. Rodrigues Jr. and J. E. Maiorino, “A unified theory for construction of arbitrary speeds ($0 \leq v \leq \infty$) solutions of the relativistic wave equations,” Random Operators and Stochastic Equations, Vol. 4, 1996, p. 355-400.

[58] . E. T. Whittaker, 1903, *ibid.*

[59] . Ironically, Maxwell’s actual theory was some 20 quaternion-like equations in 20

unknowns, as given in "A Dynamical Theory of the Electromagnetic Field," Royal Society Transactions, Vol. CLV, 1865, p 459. Read Dec. 8, 1864. Also contained in The Scientific Papers of James Clerk Maxwell, 2 vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597. Heaviside dramatically curtailed Maxwell's theory to a vector subset of some 4 equations in 4 unknowns. Lorentz further curtailed it by effectively throwing out all Maxwell-Heaviside systems that were in disequilibrium with their active vacuum and curved spacetime environments. So after over a century, the long slumbering *environmental* dynamics for electrodynamical systems is awakening and chasing its own roots again. Long live the rise of the supersystem!

[60] . E.g., see J. Vaz Jr. and W. A. Rodrigues Jr., "Maxwell and Dirac Theories as an Already Unified Theory," Advances in Applied Clifford Algebras, Vol. 7 (S), 1997, p. 369-386. See W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell Equations and the Massless Dirac Equation," Advances in Applied Clifford Algebras, Vol. 7(S), 1997, p. 457-466.

[61] . E. T. Whittaker, 1904, *ibid*.

[62] . We strongly stress that longitudinal EM waves and their dynamics, so ignored in all except the most advanced higher symmetry electrodynamics texts, are actually the key to everything. Longitudinal EM waves already comprise, create, and structure the feeble "envelope EM waves" we presently deal with. A longitudinal EM wave is also an electrogravitational wave, and oscillates the rate of flow of time itself because of its combination with a time-polarized wave. The two combined photons have a total spin two, which is a graviton. Thus internally a graviton is revealed as a particular

dynamic between time-domain EM energy and 3-space EM energy. The longitudinal EM wave in the time domain (that is combining with the longitudinal EM wave in 3-space) is thus an oscillation of time contraction and expansion, or "pulsation" in the rate of flow of time. Very powerful gravitational effects as well as very powerful and unique electrodynamics effects can be obtained by use of longitudinal EM waves, because of the additional dynamics between the time and 3-space domains. Unwitting use of such extraordinary EM waves is how Rife and Prioré achieved such remarkable results.

[63]. M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," Journal of New Energy, 4 (3), Special Issue, Winter 1999, p. 76-78.

[64]. Melba Phillips, “Classical Electrodynamics,” Vol. IV: Principles of Electrodynamics and Relativity, in Encyclopedia of Physics, Edited by S. Flugge, Springer-Verlag, Berlin, 1962, p. 1-108 gives a good overall summary of the superpotential theory.

[65]. A. Righi, Nuovo Cimento, Vol. 2, No. 5, 1901, p. 104.

[66]. P. Debye, Ann. Phys., Leipzig, Vol. 30, 1909, p. 57.

[67]. Thomas John I'Anson Bromwich, An Introduction to the Theory of Infinite Series, MacMillan, London, 1908.

[68]. A. Nisbet, Physica, Vol. 21, 1955, p. 799; Proc. Roy. Soc. Lond. A, Vol. 231, 1955, p. 250.

[69]. W. H. McCrea, “Hertzian electromagnetic

potentials," Proc. Roy. Soc. Lond. A, Vol. 240, 1957, p. 447-457.

[70]. Sachs has achieved the necessary unified field theory, broadly unifying physics from the quarks and gluons to the entire universe, and including both the active vacuum and active, locally curved spacetime. See Mendel Sachs, General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics), Reidel (now Kluwer), 1982; — Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia, Reidel (now Kluwer), 1986. It should be obvious that, unless electrodynamics theory and general relativity theory utilize the same background spacetime, it is impossible to completely merge them, from first principles. Since classical electrodynamics artificially discards its active environment, it is electrodynamics which must be changed in order to be successfully united with general relativity. Sachs has accomplished this long-sought *engineerable* unification.

[71]. Proof that the virtual state can be coherently integrated is already furnished by every source charge in the universe, which does precisely that.

[72]. We note without elaboration that all the mind functions, even of the deepest and most unconscious nature, and even of collective species unconsciousness domains—exist electromagnetically in the infolded EM domain. The "conscious mind" is a serial processor, though extremely rapid. The unconscious is a massively parallel processor. The reason the conscious mind cannot "directly see" the unconscious is that "seeing" innumerable things in the "single snapshot" at once, just obliterates all singular discrimination because of the intense summation. Mind functions are timelike, so consist of specialized structures, functioning, and dynamics

inside the time domain. Since the infolded structuring captures this entire domain, it also involves a physics of mind and matter interaction, but as a science instead of as a mysticism. In theory vacuum engines are also possible that engineer the mind at any and all levels. We have addressed this extended physics of mind and matter reactions elsewhere and do not address it here. The Russian weapon scientists refer to this mind area (which can also be engineered by novel electrodynamic means) as psychoenergetics.

[73]. R. O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. unk, No. 73, Nov.-Dec. 1970, p. 169-198. This is the definitive technical exposition by Becker of his electrical findings about the cellular control system.

[74]. See particularly Robert O. Becker, “The direct current control system. A link between environment and organism,” New York State Journal of Medicine, Vol. 62, 1962, p. 1169-1176; — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bulletin of the New York Academy of Medicine, Second Series, 48(4), May 1972, p. 627-641; — and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, Electromagnetic Applications In Medicine," NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993; — "A technique for producing regenerative healing in humans," Frontier Perspectives, 1(2), Fall/Winter 1990, p. 1-2; — and Andrew A. Marino. Electromagnetism and Life, State University of New York Press, Albany, NY, 1982; — and Gary Selden, The Body Electric, William Morrow and Company, New York, 1985; — Cross Currents, Jeremy P. Tarcher, Inc., Los Angeles, 1990; — "The bioelectric field pattern in the salamander and its stimulation by an electronic analog," IRE Transactions on Medical Electronics, Vol. ME-7, July 1960, p. 202-207; —

"Some observations indicating the possibility of longitudinal charge-carrier flow in peripheral nerves," Biological Prototypes and Synthetic Systems, Plenum Press, New York, 1962, p. 31-37; — and W. Slaughter, "The longitudinal direct current gradients of spinal nerves," Nature, Vol. 196, Nov. 17, 1962, p. 675-676; — and Charles H. Bachman and Howard Friedman, The direct current (control) system: A link between the environment and the organism," New York State Journal of Medicine, Vol. 62, April 15, 1962, p. 1169-1176; — "The direct current field: A primitive control and communication system related to growth processes," Proceedings of the. XVI International. Congress of Zoology, Washington, D.C., Vol. 3, 1963, p. 179-183; — "The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 11th International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p, 1753-1759; — "The bioelectric factors in amphibian limb regeneration," The Journal of Bone and Joint Surgery, Vol. 43-A, No. 5, July 1961, p. 643-656; — and David G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents," Transactions, New York Academy of Sciences, 29(5), Mar. 1967, p. 606-615.

[75] . The ultimate tool for use in seeing and working with this "infolded" or "internal electrodynamics inside normal EM fields, waves, and potentials will probably be the Fogal semiconductor when it goes into production in the near future. Fogal's transistors are capable of (1) infolding the desired longitudinal EM waves inside normal EM signals (he demonstrates infolding TV signals inside pure DC potentials, and transmitting them over conductors as pure DC potentials for appreciable distance, then using his semiconductors to again extract the "infolded" hidden TV signals.

[76] . See Vlail Kaznacheyev, "Electromagnetic

Bioinformation in Intercellular Interactions," Psi Research, 1(1), Mar. 1982, p. 47-76; — *et al.*, "Distant intercellular interactions in a system of two tissue cultures," Psychoenergetic Systems, 1 (3), Mar. 1976, p. 141-142; "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144, — *et al.*, "Apparent information transfer between two groups of cells," Psychoenergetic Systems, 1(1), Dec. 1974. See particularly Vlail P. Kaznacheyev and L. P. Mikhailova, Ultraweak Radiations in Intercellular Interactions [in Russian], Novosibirsk, 1981; "Electromagnetic Bioinformation in Intercellular Interactions," Novosibirsk, 1981 [in Russian]; V. P. Kaznacheyev, "Information Function of Ultraweak Light flows in Biological Systems," Problems in Biophysics, Novosibirsk, 1967, p. 7-18 [in Russian].

[77] . Some years ago we documented this disease-inducing "microwave radiation of the U.S. Embassy, explained its purpose and operational mechanism, and published it. E.g., see T. E. Bearden, AIDS: Biological Warfare, Tesla Book Co., Chula Vista, CA, 1988.

[78] . See Robert C. Mallieu, A Model of the Microwave Intensity Distribution Within the U.S. Embassy, Moscow 1966-1967, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Baltimore, Maryland, Aug. 1980. Also see Microwave Radiation at the U.S. Embassy in Moscow and Its Biological Implications: An Assessment, U.S. Department of Commerce, Doc. # NTIA-SP081-12, Mar. 1981. Also see Captain John D. LaMothe, Controlled Offensive Behavior—USSR, Report ST-CS-01-169-72, Defense Intelligence Agency, Washington, D.C. (released under FOIA). See also Captain John D. LaMothe and Mr. Louis Maire, Soviet and Czechoslovakian Parapsychology Research, Report DST-1810S-387-75, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

[79] . T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," Journal of New Energy, 3(2/3), 1998, p. 12-28. A slightly updated and extended version is published in Explore, 8(6), 1998, p. 7-16.

[80] . C. W. Misner, K. S. Thorne and J. A. Wheeler, Gravitation, W.H. Freeman and Co., San Francisco, 1973, p. 5, 19-21, 71-72, 163-165, 390-391, 399; 367-369.

[81] . T. D. Lee, Particle Physics and Introduction to Field Theory, Harwood, New York, 1981. See particularly Chapter 25: Outlook, "Possibility of Vacuum Engineering," p. 824-828.

[82] . David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," Physical Review, 85 (2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

[83] . I.e., the rate of flow of time, or more specifically, the rate of the EM interactions with a mass or particle that generate the flow of the mass or particle through time. See Figure 25. Nobelist Lee has also shown that time is a discrete dynamical variable across all of physics, from quantum mechanics to nonrelativistic field theory, to relativistic quantum field theory. See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?", Physics Letters, 122B(3, 4), Mar. 10, 1983, p. 217-220. Also in T. D. Lee, Selected Papers, Gerald Feinberg, Ed., Birkhauser, Boston, 1986, Vol. 3, p. 77-80. So we have not hesitated to consider time as any other variable, and the fourth Minkowski axis as any of the other three axes with respect to EM energy, fields, potentials, and waves.

[84] . Indeed, there is no such unique thing as “the gravitational field” anyway! The gravitational field is a conglomerate of other field energies and

functions, not a separate field of nature. See A. A. Logunov and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," Sov. J. Part. Nucl., 18(3), May-June 1987, p. 179. See C. W. Misner, K. S. Thorne and J. A. Wheeler, Gravitation, W.H. Freeman and Co., San Francisco, 1973, p. 399. Quoting*: "The terms 'gravitational field' and 'gravity' refer in a vague, collective sort of way to all of these entities. Another, equivalent term for them is the 'geometry of spacetime'."* See also A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," Soviet Physics Doklady, Vol. 12, No. 11, 1968, p. 1040-1041 [English translation.]. Finally, see our remarks under reference 62 above, particularly with respect to the proposed internal electromagnetic dynamics of the graviton.

[85] . A 1993 study by the Office of Technology Assessment, delivered to Congress, showed that a single terrorist in a small plane, with a tank of anthrax spray, and dispersing the spray as he flew around the greater Washington D.C. area on a calm night, would produce between 1 and 3 million casualties. Most of those casualties would die. So a single biological warfare (BW) strike in one U.S. metropolitan area could easily kill two million Americans. As an example, Saddam Hussein is known to have the terrorist teams and the anthrax, and he has had a decade since the Gulf war to insert additional teams and biological warfare agents into the U.S. Presently this is rated by our government as the strongest strategic threat to the U.S., after which comes nuclear strikes, followed by chemical agent strikes. Redevelopment of the Prioré approach, e.g., could easily save the lives of 90% of that nominal two million dead Americans. As matters stand now, triage would apply and almost all of those two million stricken Americans would die.

[86] . In 1998 we proposed a crash development of this kind of device in 1998, with copies to DoD,

NIH, CDC, the USAF, etc.; in T. E. Bearden, Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page inclosure to CTEC Letter, “Saving the Lives of Mass BW Casualties from Terrorist BW Strikes on U.S. Population Centers,” to Major General Thomas H. Neary, Director of Nuclear and Counterproliferation, Office of the Deputy Chief of Staff, Air and Space Operations, HQ USAF, May. 4, 1998. Also to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. Also to NIH, CDC, and several other U.S. government agencies.

DoD replied that it was not documented in the literature (a specific list of exact scientific papers reporting the results of the Prioré team's experiments and the experiments of Becker was included; no one seemed to read it). CDC and the USAF did not reply. NIH first responded (from its Policy section—read, *spin control* section) that, since I mentioned the cause of the Gulf War Disease and how it was induced, it was not a matter of their interest but of DoD's interest, and so they just shipped it to DoD and got rid of it. A second response evoked from NIH on query from my congressman (again, a response from their *policy* section) directed me to their proposals section. I don't believe a single qualified scientist even read it. DoD weakly replied that it was not documented in the literature (a specific list of citations from the hard French scientific literature, reporting the astounding results of the Prioré treatments, was included, which apparently no one read).

Not a single scientist from any agency called me to discuss the revolutionary and proven results obtained by Prioré, and documented in the attachments and in the cited literature. Sadly, it seems that no one had the foggiest notion what I was talking about, and no one much cared. My conclusion was and is that my own government bureaucrats do not give a tinker's dam about saving millions of American casualties in the coming debacle, *if it has to be done by "out-of-the-box" methods not approved by the big drug industry*. Yet these are the same officials that readily admit the exact threat I referred to, admit that it is a matter of *when* the biological warfare attacks will occur and not *if*, and admit that right now they have no effective way to cope with it or save most of those coming millions of Americans that are going to just be dragged aside to die.

My response is that the American taxpayers are certainly not getting their money's worth from the billions they shower on our medical research community. It is unconscionable to ignore for over 30 years the demonstrated and astounding results achieved by eminent French scientists—such as Pautrizel, a world-renowned parasitologist and Courier, head of the French Academy's Biological Section, and also *Secretaire Perpetuel* of the Academy) working with the Prioré team, in curing terminal tumors etc. in thousands of lab animals, all properly documented in the hard French scientific literature.

In my view, considering the certainty of those future American mass deaths, it is akin to high treason. Whether our scientific community realizes it or not, today's strategic war of massive strategic strikes has shifted to our civilian populace, and to our cities and mass population centers. It is our civilian populace that will sustain the greatest and most prompt casualties, and probably will receive the first massive strikes, totally by surprise.

The first phase of WW III—delivery of the strategic weapons to their strategic targets—has already been accomplished. Large Russian nuclear weapons are already hidden in our major cities and population centers (read Lunev's book, he tells us how they brought in these weapons), the Spetznaz teams are here to detonate them on call, multiple nations such as Iraq, Libya, and Iran already have terrorists agents in country in the U.S. with anthrax and smallpox in hand to be released, and Castro guerrillas and trained saboteurs have been trained in Southern Mexico and infiltrated across the border into the U.S. for decades.

Whether they like it or not, U.S. scientists have thus become the major front line "soldiers" and "defenders" of our nation. And frankly, they have failed abominably in their responsibility, and most do not seem to care. For a startling and even shocking characterization of our present scientific community, the reader is referred to Daniel S. Greenberg, Science, Money, and Politics: Political Triumph and Ethical Erosion, University of Chicago Press, 2001.

[87]. As we stated, just such an attack is now officially the greatest strategic threat to the United States. Second is a nuclear attack, followed by chemical attack as a distant third. E.g., see Richard K. Betts, “The New Threat of Mass Destruction,” Foreign Affairs, 77(1), Jan./Feb. 1998, p. 26-41. For the ineffectiveness of our defenses against the coming mass biological warfare strikes, see Laurie Garrett, The Coming Plague: Newly Emerging Diseases in a World Out of Balance, Farrar, Straus and Giroux, New York, 1994; — "The Nightmare of Bioterrorism," Foreign Affairs, 80(1), Jan./Feb. 2001, p. 76-89; — Betrayal of Trust: The Collapse of Global Health, Hyperion, New York, 2000.

[88]. Becker, *ibid*.

[89]. A full account of the entire Prioré affair is given in Jean-Michel Graille, Dossier Prioré: Une Nouvelle Affaire Pasteur [The Prioré Dossier: A New Pasteur Affair?], De Noel, Paris, 1984. [in French]. A summary account is given by Christopher Bird, "The Case of Antoine Prioré and His Therapeutic Machine: A Scandal in the Politics of Science," Appendix I in T.E. Bearden, AIDS: Biological Warfare, Tesla Book Co., P.O. Box 121873, Chula Vista, California 91912, 1988. Prioré's doctoral thesis (in my personal possession courtesy of Professor Courrier and the late Chris Bird), was rejected by the University of Bordeaux

when the French government suppressed the Prioré program in the mid-70s; it is A. Prioré, Guérison de la Trypanosomiase Expérimentale Aiguë et Chronique par L'action Combinée de Champs Magnétiques et D'Ondes Electromagnétiques Modulés. [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], 1973. Prioré's patents are: A. Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963. Eleven years after Priore's thesis was rejected, the same university awarded a doctoral thesis to a young graduate. It is Eric Perisse, Effets des Ondes Electromagnètiques et des Champs Magnètiques sur le Cancer et la Trypanosomiase Experimentale [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

[90] . F. A. Popp *et al.*, "Biophoton Emission (Multi-Author Review)," Experientia, Vol. 44, 1988, p. 543-600; — "Biophoton emission: New evidence for coherence and DNA as a source," Cell. Biophys., Vol. 6, 1984, p. 33-52; — "Physical aspects of biophotons." Experientia, Vol. 44, 1988, p. 576-585; — Recent Advances in Biophoton Research and its Applications, World Scientific Publishing Co., New York, NY, 1992; F. A. Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp *et al.*, eds., Electromagnetic Bio-Information: Proceedings of the Symposium,

Marburg, Federal Republic of Germany, Sep. 5, 1977; Urban & Schwarzenberg, Baltimore, 1979, p. 123-149.

[91] . David M. Rorvik, "Do the French have a Cure for Cancer?", Esquire Magazine, July 1975, p. 110-111, 142-149.

[92] . Arguably the best introduction to nonlinear optical phase conjugation is David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Another very good presentation is Amnon Yariv, Optical Electronics, 3rd Edition, Holt, Rinehart, and Winston, New York. 1985, Chapter 16: “Phase Conjugate Optics – Theory and Applications.”

[93] . A beautiful example of this “precise backtracking” of the time-reversed wave is given in David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83. See particularly the striking photographic demonstration of time reversal of disorder on p. 75.

[94] . We presented some general elements of this dramatic new extension to nonlinear optical phase conjugate pumping and to general relativity, at a graduate seminar, “On the Theory and Principles of EM Systems with COP>1.0”, given to interested graduates and faculty of the University of Louisville, on Nov. 20, 1997.

[95] . E.g., see Owen Flynn, "Parametric arrays: A new concept for sonar," Electronic Warfare Magazine, June 1977, p. 107-112.

[96] . This method of increasing signal-to-noise ratio and pulling hidden signals right up out of the noise, is widely used in signal processing in modern radars. Often signals buried more than 100

dB below the noise level are routinely "pulled out of the noise" and detected, using various forms of coherent summation.

[97] . Any potential decomposes into multiple wavepair sets. In each hidden wavepair, each of the two waves further decomposes into two potentials. Each of *those* potentials then further decomposes into multiple wavepair sets. And so on *ad infinitum*. Or one can start with any wave, decompose it into two potentials, each of those potentials into multiple biwave sets, and so on.

[98] . "Fractal" means self-similarity at all levels, so that patterning or "decomposition" at one level is similar to that at the higher and lower levels. For good overview discussions and applications, see particularly Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," and Joseph Ford, "What is chaos, that we should be mindful of it?", both in The New Physics, Paul Davies, editor, Cambridge University Press, Cambridge and New York, 1989, p. 316-347 and 348-371 respectively. More detailed technical references are cited in each article.

[99] . See Trevor James Constable, "You, Too, Can Photograph UFOs from Airliners," Psychic Observer and Chimes, 37(3), May-June 1977; — The Cosmic Pulse of Life, Merlin Press, Santa Ana, CA, 1976.

[100] . The present author has two beautiful ultraphotographs taken by Joe Gambill using the special filtering approach. One is a photograph of a room, but with its contents of *two weeks earlier*, showing the room prior to the time the photograph was taken. The other is of a "ghost" form—very similar to a photograph obtained by George Meek under rigorous laboratory conditions—and clearly showing the energy form of a "haunted" lady's dead husband. The *interpretation* of these photographs

is of course open to scientific dispute, but the *photographs themselves* are impeccable and real.

[101] . Meek used filters open to UV and IR but opaque to the visible spectrum, on the fluorescent lights in his laboratory. He also used similar filtering on the camera lenses and camera flashes. Under rigorous laboratory conditions, in several thousand photos of the living body Meek was able to take a few highly anomalous photos completely inexplicable by normal science. E.g., he photographed a purported "ghost" and obtained a clear shot of energy streaming associated with a living person. He also photographed the legs of a chair, right through the suddenly "missing" legs of the person sitting in the chair. Again, interpretation is open to dispute, but the photographs are rigorous and real.

[102] . Bill Fogal's patented charge-blocking semiconductor—U.S. Patent Numbers 5,196,809 and 5,430,413—is capable under certain circumstances—and by clever adaptation—of amplifying virtual state information, in a manner somewhat similar to Rife's methodology, but far more easily engineered. When the process works, Bill reports that a clear photograph or video camera shot can be taken right through a solid brick wall, of the hidden scenery beyond. In electrodynamics, in a dielectric absorbing or emitting EM radiation (energy), the entire dielectric participates in each absorption and emission. Emitted light from an object thus carries an astoundingly complete set of little virtual state vacuum engines rigorously representing every part of the object, no matter how small, and all the internal dynamics. This means that the light from the scene beyond the wall strikes the wall on the other side, affecting all the material in the wall, including upon the observer's side, and inserting "dimensioning" from the scene beyond the wall. When light is emitted from the observer's side, the information content of the field representing the scene on the far side, is also part of

the information content of the field in the light emitted from the observer side of the wall and striking the camera and the observers eyes. This "dimensioning" continues in the signals in the camera and on through attached electronic circuitry. Since Fogal's semiconductor can process such information content of the field in the proper circuitry and under favorable circumstances, the phenomenon is therefore permissible and can be technically explained. Using his transistor to transduce the dimensioned information into the visual band, Fogal *can* view a distant scene right through an intervening wall, when he gets everything "just right." At this stage of development, however, the process and adjustment are difficult, just as Rife experience with his microscope. We are convinced that the forthcoming advent of Fogal semiconductors will revolutionize the study and treatment of disease and bodily disorders.

[103] . However, highly advanced electrodynamics does consider the photon with far greater complexity than nuclear engineers such as the present author were introduced to! E.g., a deep discussion of photon polarization is given in F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984 under the heading "covariant quantization of the photon propagator" in Chapter 5. As pointed out by Bob Flowers in a private communication, Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where the combination manifests as the "instantaneous" Coulomb (i.e., electrostatic) potential. Most such books do argue that the extra components of the photon remain hidden, as does the virtual state. In this paper we have revealed how Rife discovered a process by means of which the virtual state can be observed in great detail.

[104] . V. P. Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," Psi

Research, 1(1), Mar. 1982, p. 47-76; — "Information Function of Ultraweak Light Flows in Biological Systems," in Problems in Biophysics, Novosibirsk, 1967, p. 7-18 [in Russian]; — and L. P. Mikhailova, Ultraweak Radiation in Intercellular Interactions, [in Russian], Novosibirsk, 1981; — *et al*., "Apparent information transfer between two groups of cells," Psychoenergetic Systems, 1(3), Mar. 1976; — *et al*., "Distant intercellular interactions in a system of two tissue cultures," Psychoenergetic Systems, 1(3), Mar. 1976, p. 141-142; "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144.

[105]. Mitogenetic radiation was studied prior to WW II by Gurwich (spelling varies). See A. Gurwitsch and L. Gurwitsch, "L'Analyse mitogenetique spectrale [The Analysis of the mitogenetic spectrum]," in Andre Mayer, Exposés de Physiologie, [Studies in Physiology], Hermann & Cie, Paris, 1934. Part IV. [In French]. Studies have continued by Popp and others. E.g., see F. A. Popp *et al*., "Biophoton Emission (Multi-Author Review)," Experientia, Vol. 44, 1988, p. 54 3-600; — "Physical aspects of biophotons." Experientia, Vol. 44, 1988, p. 576-585; — "Biophoton emission: New evidence for coherence and DNA as a source," Cell. Biophys., Vol. 6, 1984, p. 33-52. See Fritz Albert Popp; Ke-hsueh Li, and Qiao Gu, [eds.], Recent Advances in Biophoton Research and its Applications, World Scientific Publishing Co., New York, NY, 1992.

[106]. The conventional quantum mechanics assumes that, at the microscopic level, variation (e.g., in the probabilities assigned to interpret what is propagating in the Schrödinger equation) is assumed to be random. However, there is a problem. Big things must be made of little things. If we sum a set of "random" things based on Gibbs thermodynamics statistics, we cannot and will not get an organized "big thing" because we have

assumed total disorder (randomness) *a priori*. Adding heat does not produce cooling! So one falsifies the prevailing interpretation of quantum mechanics, whenever one observes a tree, a flower, the sun, or a human being. In short, there must exist a great unifying principle in nature of a coherent summation, from the "unobservable" virtual state to the observable state. Only in that fashion can there be an macroscopic order at all— and such macroscopic order is universally observed. We believe that the principle we have elucidated, as experimentally produced by Rife, may well be that long-sought universal ordering principle that is so sorely needed by quantum mechanics. Readers interested in the "missing quantum chaos" problem are referred to P. V. Elyutin, "The Quantum Chaos Problem," Sov. Phys. Usp., Vol. 31, No. 7, 1988, p. 597-622. Elyutin discusses the crisis in quantum mechanics because of the missing hidden order. Quantum mechanics is known to be wrong unless this order can be found, because otherwise it does not predict the ordered macroscopic universe. Other useful references are Ilya Prigogine, with T. Petrosky, "Quantum Chaos, Complex Spectral Representation and Time-Symmetry Breaking," Chaos, Solitons, and Fractals, Vol. 4, 1994, p. 311-359; and Robert Pool, "Quantum Chaos: Enigma Wrapped in a Mystery," Science, 243(4893), 1989, p. 893-895.

Slide Index

Figure 1. The Rife microscope.

Figure 2. Royal Raymond Rife adjusts his microscope.

Figure 3. Faraday omitted the string holder for his taut string, thus omitting half the electrodynamics and half the energy. An equal-and-opposite Newtonian third law reaction wave always appears in our transmitters and receivers, caused by the omitted wave.

Figure 4. Drude gas gyroelectrons in the receiving antenna precess laterally, providing the detected transverse wave and proving the vacuum disturbance wave to be longitudinal.

Figure 5. The magnitude of the causal potential in massfree space is erroneously defined.

Figure 6. Fields and potentials are erroneously defined as what is diverted from them.

Figure 7. The defined magnitude of the field is erroneous and is the field intensity.

Figure 8. Matter-to-matter transform.

Figure 9. The missing matter-to-spacetime and spacetime-to-matter transforms.

Figure 10. Vacuum engine utilization.

Figure 11. A charge is continuously in violent energy exchange with the active vacuum.

Figure 12. Whittaker's biwave decomposition of the scalar potential, and its reinterpretation.

Figure 13. The Supersystem and its components.

Figure 14. Vacuum engine (cartoon representation).

Figure 15. First vacuum engine analogy of arm muscle and bone.

Figure 16. Second vacuum engine analogy of arm bone and muscle.

Figure 17. Third vacuum engine analogy.

Figure 18. Fourth vacuum engine analogy.

Figure 19. Time-like flow of EM energy into 3-space.

Figure 20. Block diagram of the immune system.

Figure 21. Becker's electromagnetic healing of otherwise intractable bone fractures.

Figure 22. Becker's block diagram of the cellular control system.

Figure 23. Kaznacheyev's electromagnetic induction of cellular death and disease, at a distance.

Figure 24. Microwave radiation of the U.S. Embassy in Moscow.

Figure 25. Mechanism generating the flow of a mass through time.

Figure 26. Vacuum antiengine for a specific disease condition.

Figure 27. Pumping with transverse and longitudinal EM waves.

Figure 28. Block diagram of the Prioré process.

Figure 29. Block diagram of the operation of the body's cellular regenerative system.

Figure 30. Rife's great secret: Recursive iterative serial amplification by a single harmonic interval difference frequency.

Help support this work

The Tom Bearden
Website

Rife's Microscope

Table of Photographs

(Thanks to the donor of these photographs, who wishes to remain anonymous)

1. Rife's Universal Microscope
2. Eyepiece Details
3. Chart for Calculating Parallax of Lens System
4. Diagram - Wide Angle Apochromatic Microscope Objective Lens
5. Universal Focus - Photographic Lens
6. Light Passage of Dove Prism

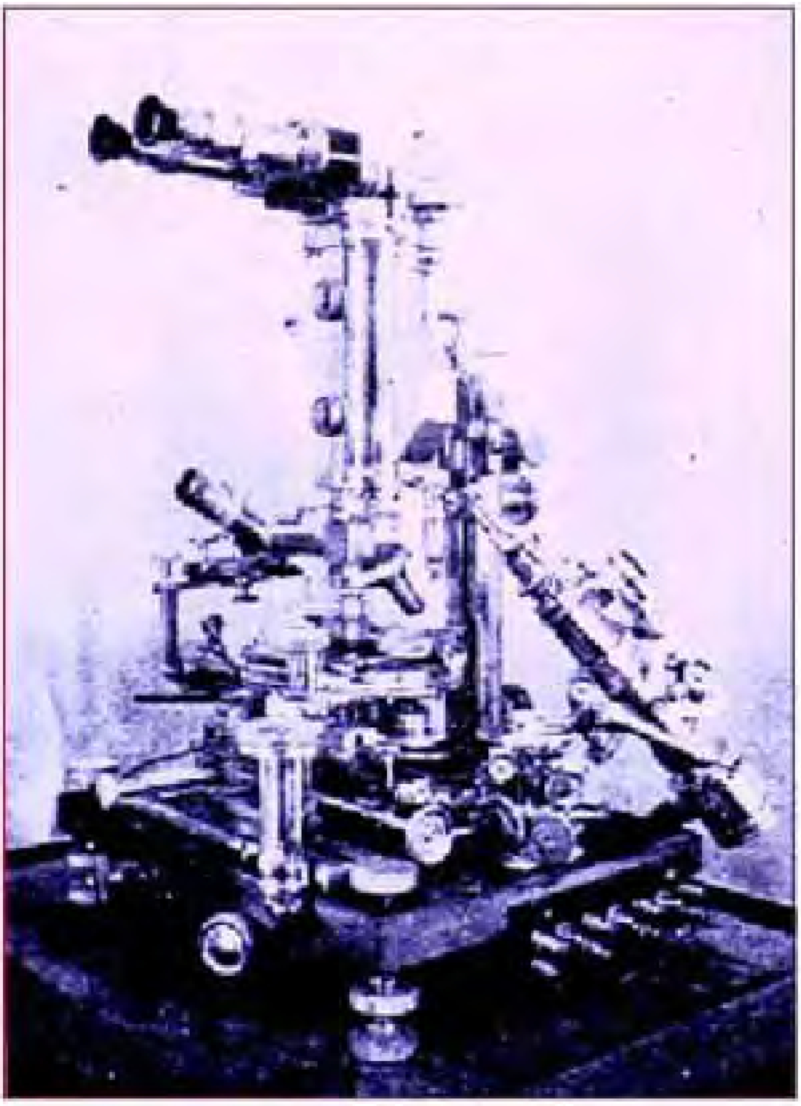

Figure 1. Rife's microscope.

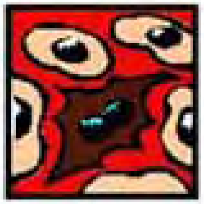

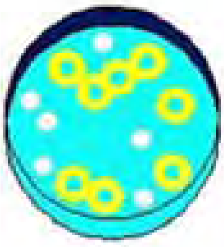

Figure 2. Royal Raymond Rife adjusts his microscope.

Air medium disturbances generated when air is perturbed by a plucked taut string

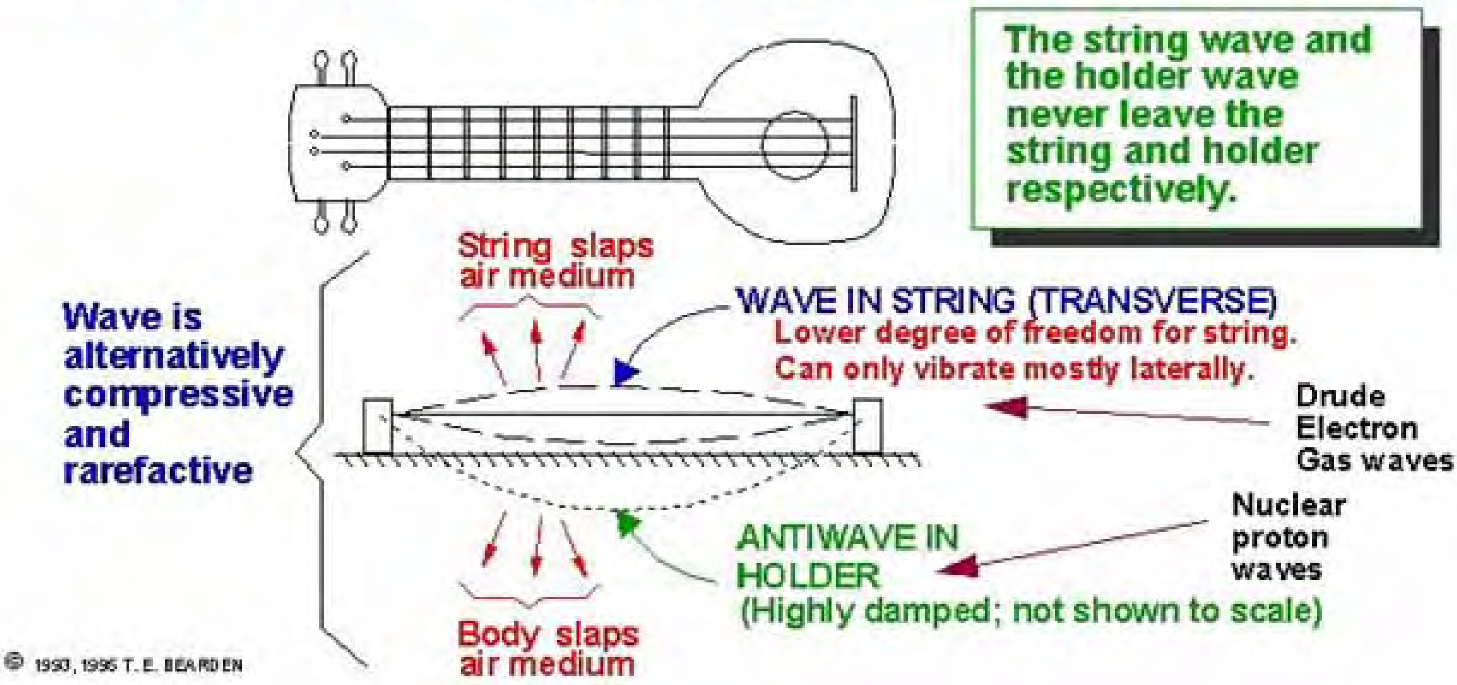
String slaps air medium
WAVE IN STRING (TRANSVERSE)
Lower degree of freedom for string.
Can only vibrate mostly laterally.
Drude Electron Gas waves
Wave is alternatively compressive and rarefactive
Nuclear proton waves
ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)
Body slaps air medium

The string wave and the holder wave never leave the string and holder respectively.

© 1993, 1996 T. E. BEARDEN

Figure 3. Analogy of the "taut string holder" omitted by Faraday. This omitted half the electrodynamics (including in Maxwell's capture of Faraday's notion) and half the EM energy. An equal and opposite Newtonian 3rd law reaction does appear in our transmitters and receivers.

Flow of current and energy into, through, and along a conductor

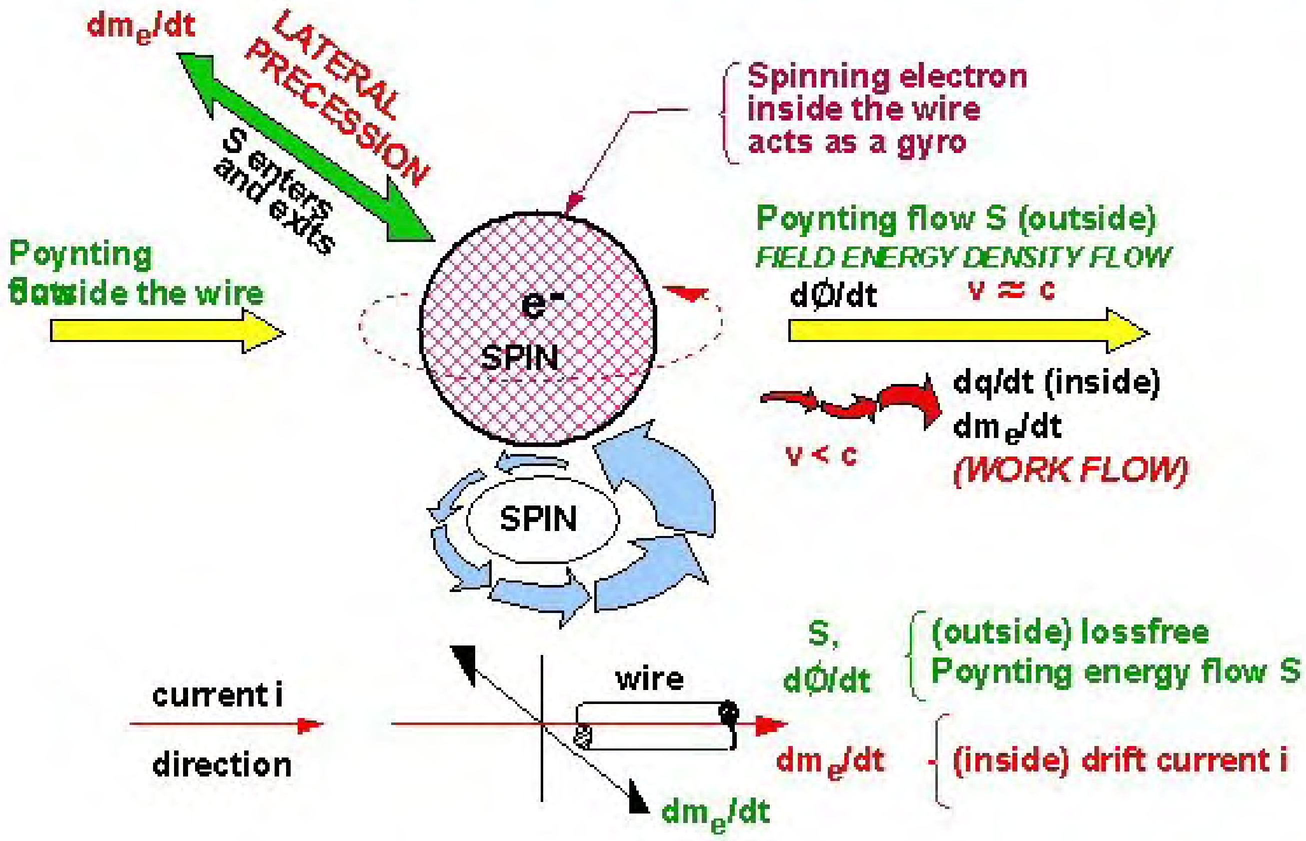
dm$_e$/dt
LATERAL PRECESSION
S enters and exits
Spinning electron inside the wire acts as a gyro
Poynting flow outside the wire
e⁻
SPIN
Poynting flow S (outside)
FIELD ENERGY DENSITY FLOW
d∅/dt
v ≈ c
dq/dt (inside)
dm$_e$/dt
v < c
(WORK FLOW)
SPIN
S,
d∅/dt
(outside) lossfree
Poynting energy flow S
current i
direction
wire
dm$_e$/dt
(inside) drift current i
dm$_e$/dt

Figure 4. Receiving gyroelectrons precess laterally, providing the detected transverse wave and proving the incoming vacuum disturbance to be longitudinal.

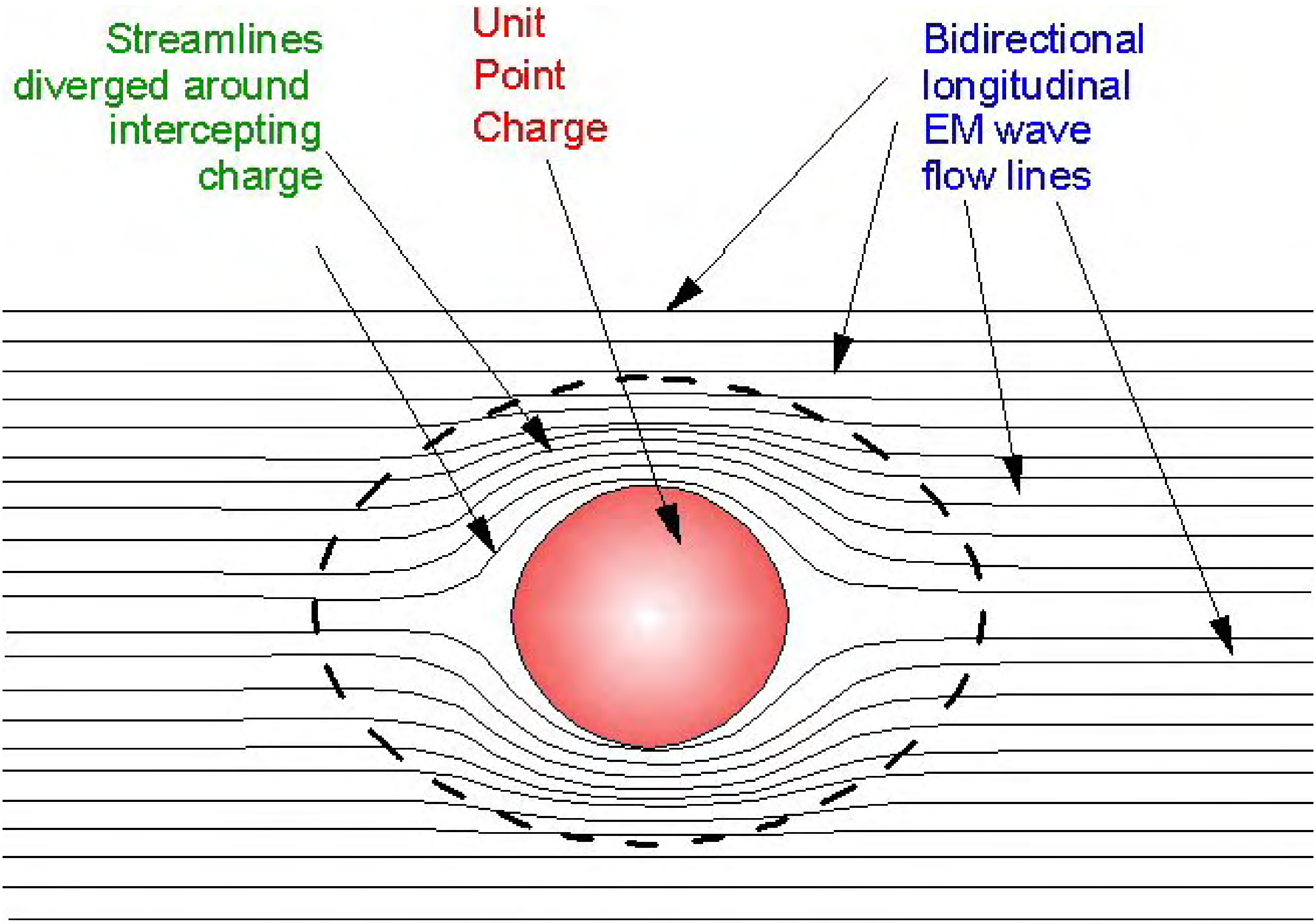
Streamlines diverged around intercepting charge
Unit Point Charge
Bidirectional longitudinal EM wave flow lines

Figure 5. The defined "magnitude of the potential" is not such at all, but is the magnitude of the energy in the diverged flow streamlining around a unit point charge at each point in space where the potential is present.

Electrodynamics has nothing to say about what exists in the vacuum in the absence of mass.

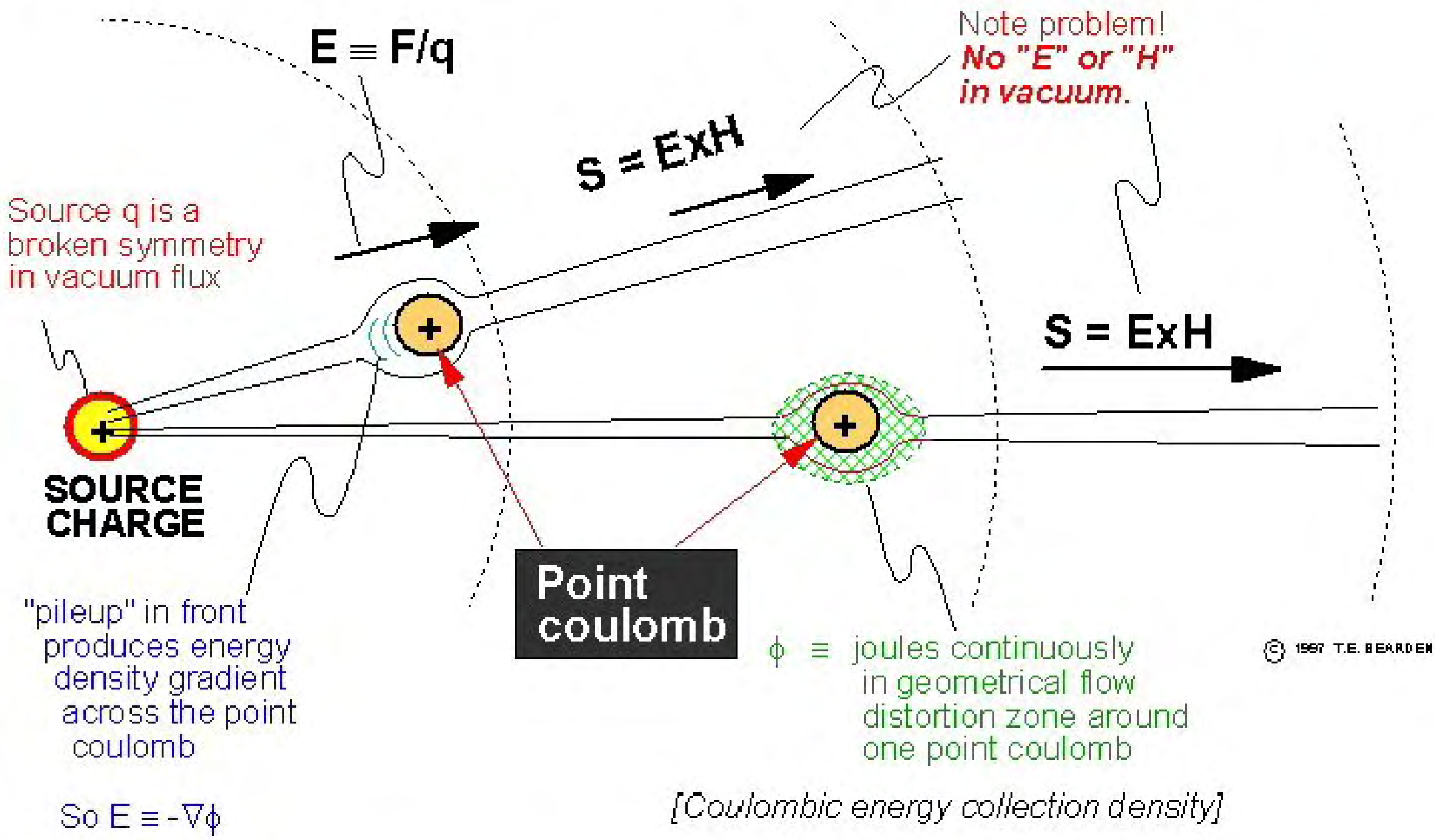

E ≡ F/q
Note problem!
No "E" or "H" in vacuum.
S = ExH
Source q is a broken symmetry in vacuum flux
S = ExH
+
+
+
SOURCE CHARGE
Point coulomb
"pileup" in front produces energy density gradient across the point coulomb
So E ≡ -∇ϕ
ϕ ≡ joules continuously in geometrical flow distortion zone around one point coulomb
[Coulombic energy collection density]
© 1997 T.E. BEARDEN

Figure 6. Fields and potentials are defined in terms of the Poynting disruption of the overall energy flow from the source charge, about an intercepting and "collecting" charge. Rigorously, *as defined* E and H exist only in and on charged matter. The magnitude of the diverted water swinging around a fixed rock in a river, is not the magnitude of the river.

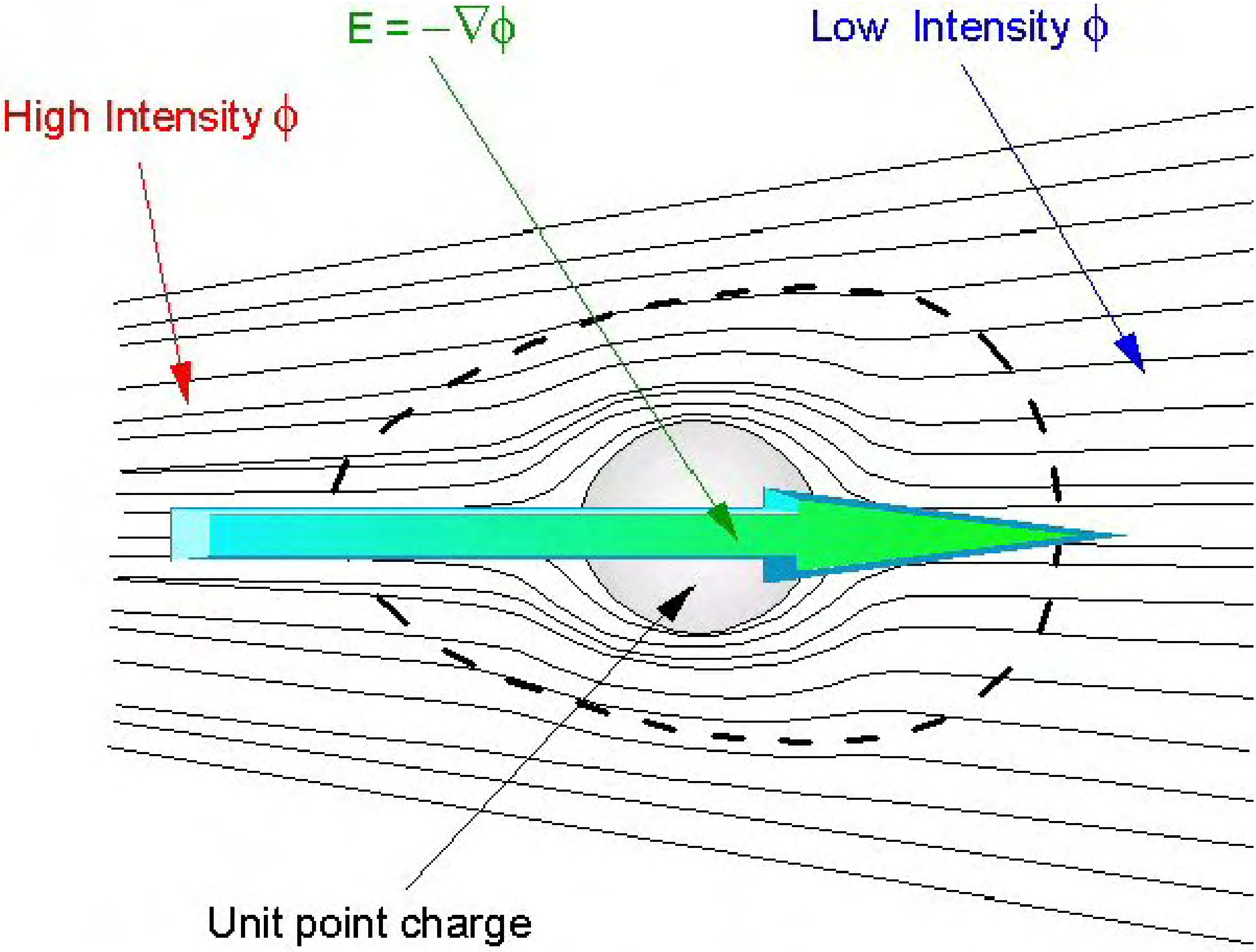
E = –∇φ
Low Intensity φ
High Intensity φ
Unit point charge

Figure 7. The defined "magnitude of the field" depends on the change in intensity of the flows comprising the potential. The E-field is oriented from high (pressure) in the flow to low (pressure) in energy density over a unit length.

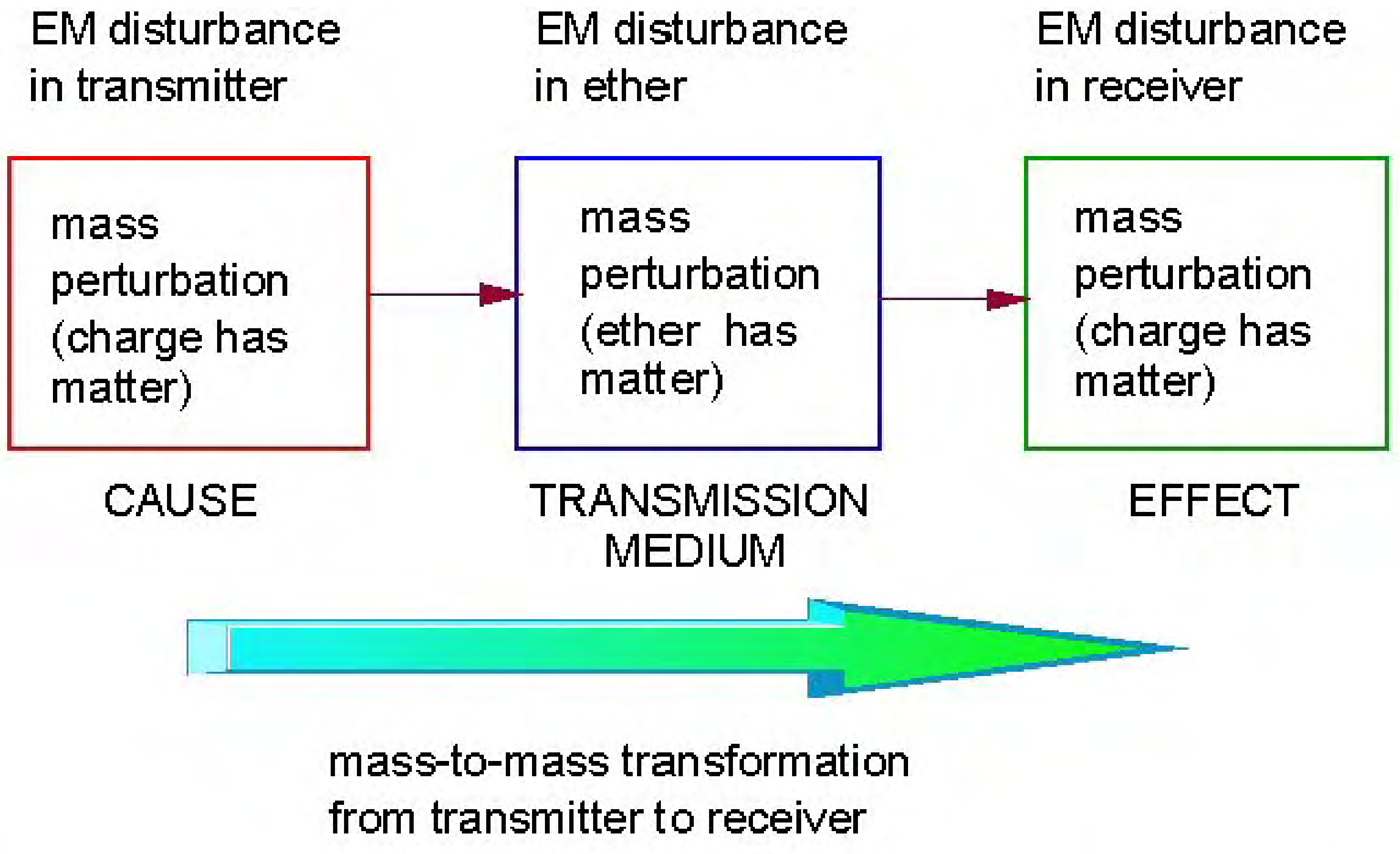
EM disturbance in transmitter
EM disturbance in ether
EM disturbance in receiver
mass perturbation (charge has matter)
mass perturbation (ether has matter)
mass perturbation (charge has matter)
CAUSE
TRANSMISSION MEDIUM
EFFECT
mass-to-mass transformation from transmitter to receiver

Figure 8. Maxwell's theory assumed a matter-to-matter transform from cause to effect.

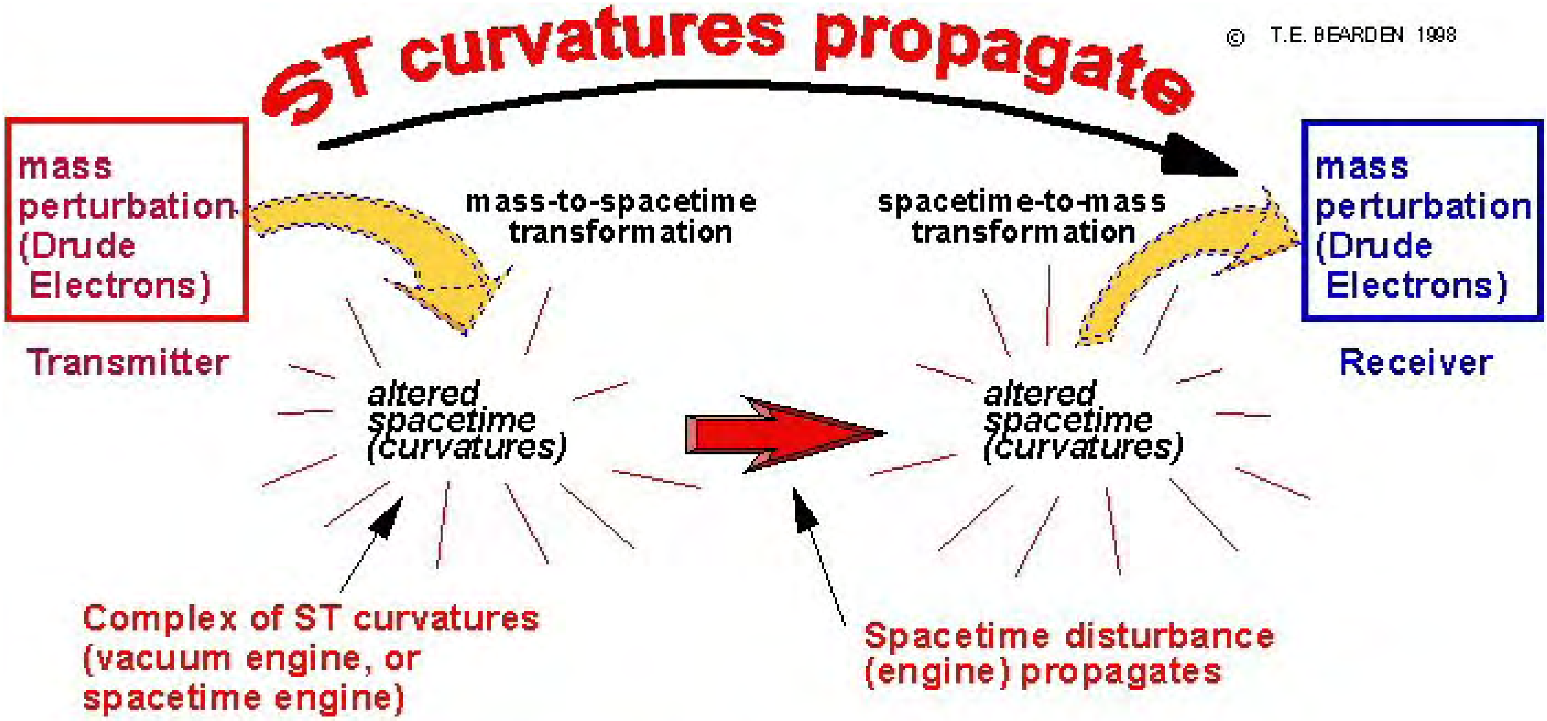
ST curvatures propagate
© T.E. BEARDEN 1998
mass perturbation (Drude Electrons)
Transmitter
mass-to-spacetime transformation
spacetime-to-mass transformation
mass perturbation (Drude Electrons)
Receiver
altered spacetime (curvatures)
altered spacetime (curvatures)
Complex of ST curvatures (vacuum engine, or spacetime engine)
Spacetime disturbance (engine) propagates

The mass-to-mass transform contains two hidden infolded transforms: (i) the mass-to-spacetime transform and (ii) the spacetime-to-mass transform. What is transmitted and propagates in space is a *spacetime perturbation*.

Figure 9. Maxwell's matter-to-matter transform contains two hidden transforms: (1) matter-to-spacetime, and (2) spacetime-to-matter.

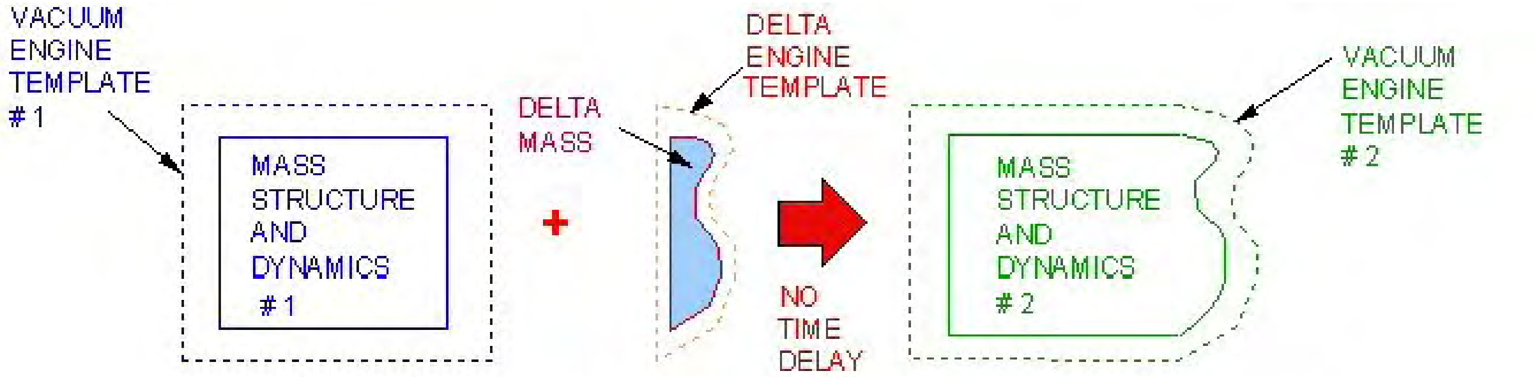
VACUUM
ENGINE
TEMPLATE
# 1
MASS
STRUCTURE
AND
DYNAMICS
# 1
+
DELTA
MASS
DELTA
ENGINE
TEMPLATE
NO
TIME
DELAY
MASS
STRUCTURE
AND
DYNAMICS
# 2
VACUUM
ENGINE
TEMPLATE
# 2

Conventional way that bulk vacuum engines are added by adding mass. Also, fields alone may be utilized as "powerful translators" without finesse.

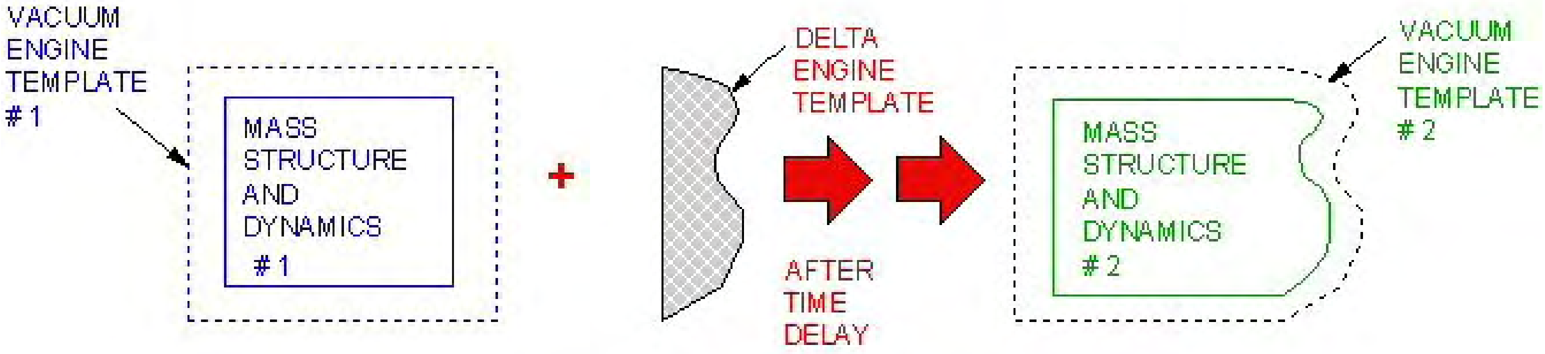
VACUUM
ENGINE
TEMPLATE
# 1
MASS
STRUCTURE
AND
DYNAMICS
# 1
+
DELTA
ENGINE
TEMPLATE
AFTER
TIME
DELAY
MASS
STRUCTURE
AND
DYNAMICS
# 2
VACUUM
ENGINE
TEMPLATE
# 2

Adding the vacuum engine without fields or masses. First the internal hidden engines of the target are changed, producing a structured template, which gradually changes the mass at all levels.

© 1994 T.E. BEARDEN

Figure 10. Vacuum engines and their utilization.

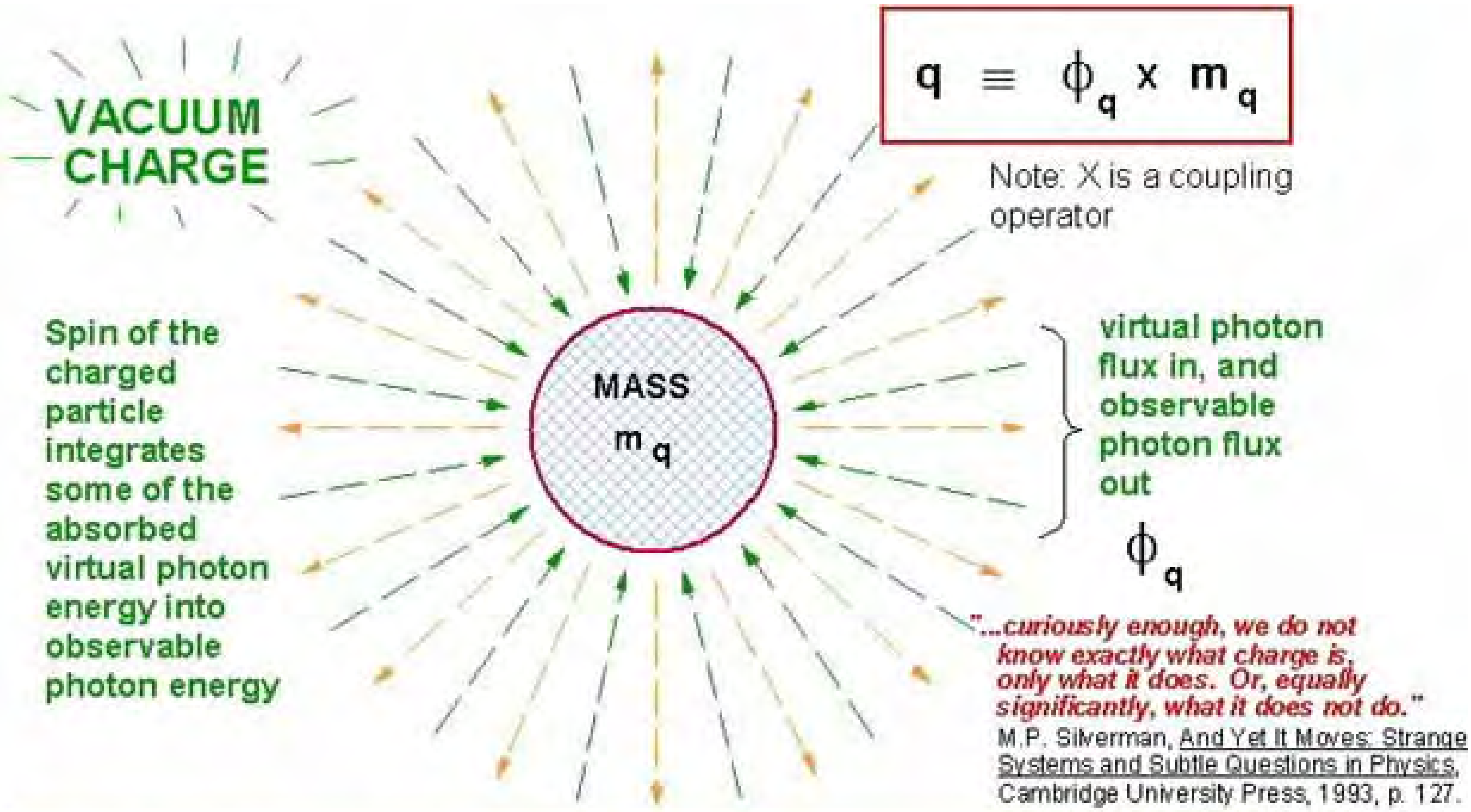
VACUUM CHARGE
$q \equiv \phi_q \times m_q$
Note: X is a coupling operator
Spin of the charged particle integrates some of the absorbed virtual photon energy into observable photon energy
MASS
m_q
virtual photon flux in, and observable photon flux out
ϕ_q

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."
M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge Q is a broken 3-symmetry. The mass of the charge continuously and violently absorbs virtual photons from the surrounding vacuum, and integrates some of them to observable EM longidutinal photons which it emits in 3-space.

Figure 11. A charged particle continuously absorbs virtual energy from the vacuum, integrates some of it into observable energy, and pours out this observable EM energy in 3-space in all directions.

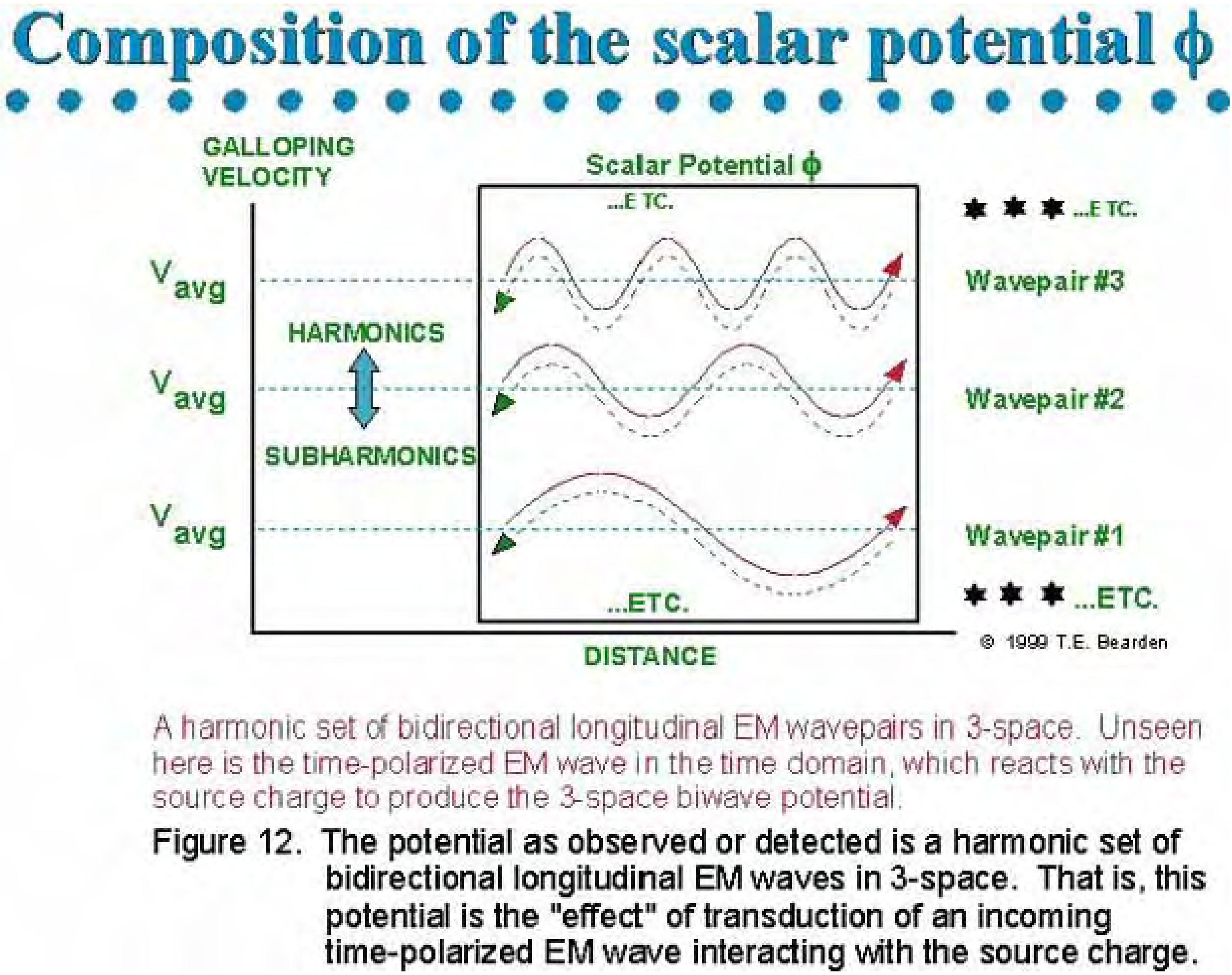

Composition of the scalar potential φ

GALLOPING VELOCITY
Scalar Potential φ
...ETC.
V_{avg}
V_{avg}
V_{avg}
HARMONICS
SUBHARMONICS
✱ ✱ ✱ ...ETC.
Wavepair #3
Wavepair #2
Wavepair #1
✱ ✱ ✱ ...ETC.
...ETC.
DISTANCE
© 1999 T.E. Bearden

A harmonic set of bidirectional longitudinal EM wavepairs in 3-space. Unseen here is the time-polarized EM wave in the time domain, which reacts with the source charge to produce the 3-space biwave potential.

Figure 12. The potential as observed or detected is a harmonic set of bidirectional longitudinal EM waves in 3-space. That is, this potential is the "effect" of transduction of an incoming time-polarized EM wave interacting with the source charge.

The supersystem concept

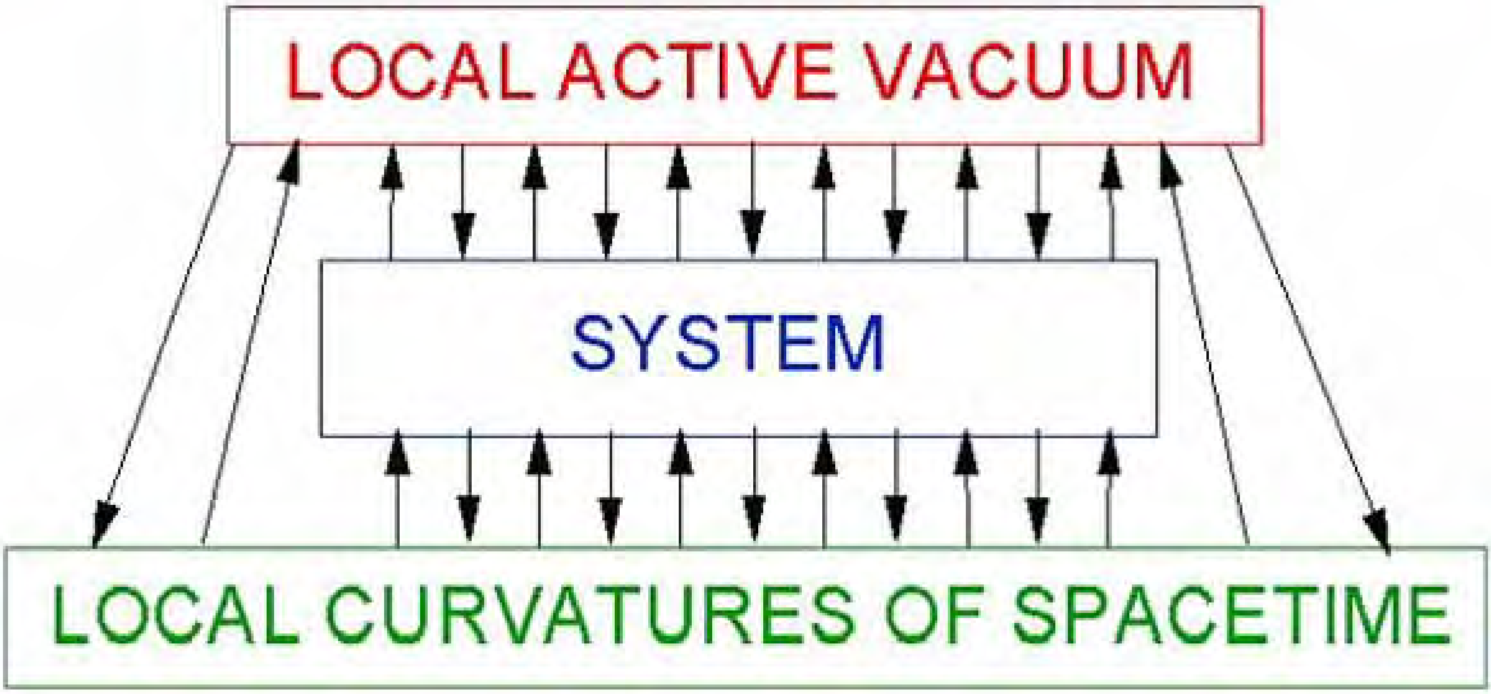
LOCAL ACTIVE VACUUM
SYSTEM
LOCAL CURVATURES OF SPACETIME

Figure 13. Concept of the supersystem and its interacting components.

Vacuum engine: Working spacetime demons

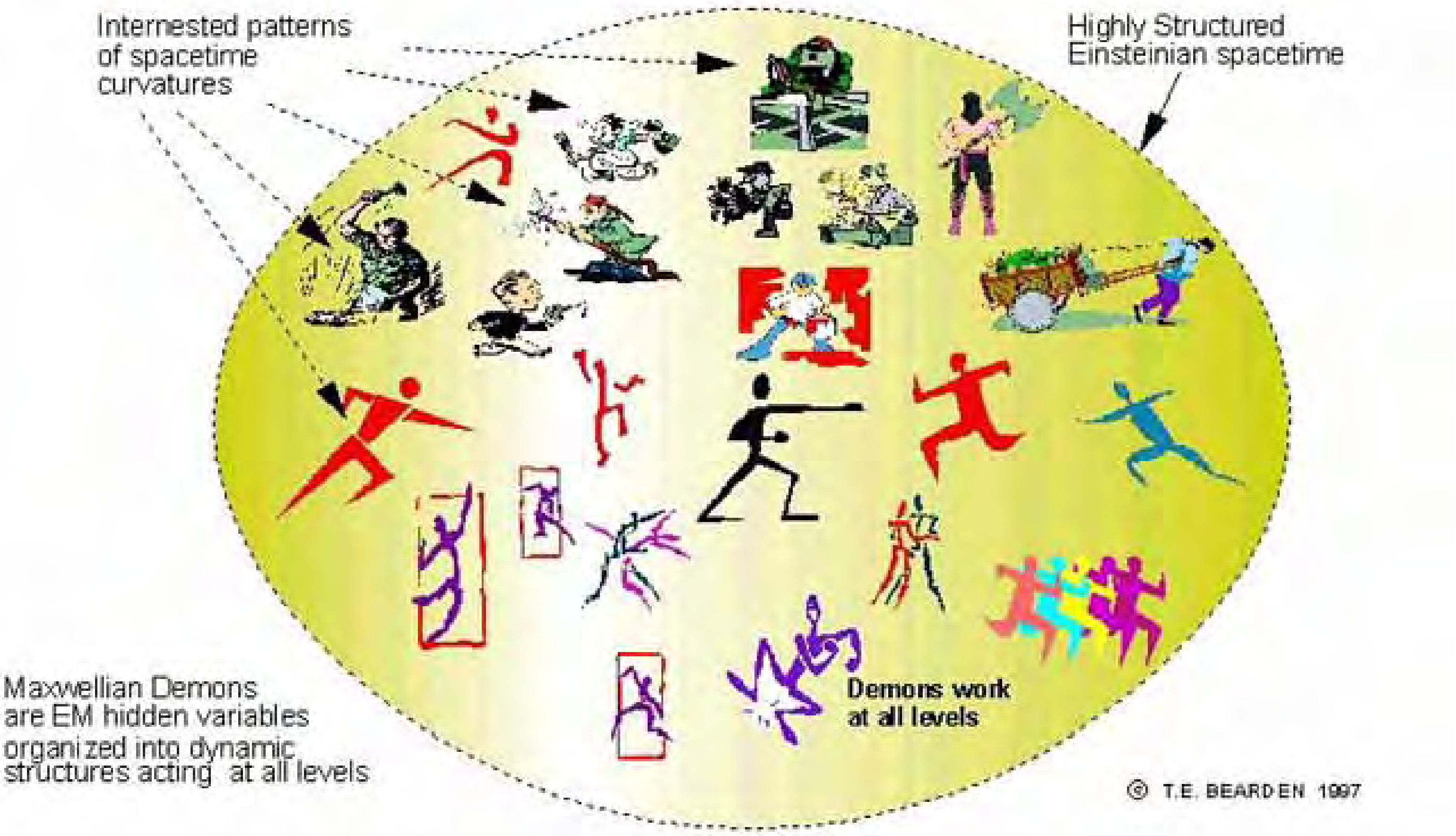
Internested patterns
of spacetime
curvatures
Highly Structured
Einsteinian spacetime
Maxwellian Demons
are EM hidden variables
organized into dynamic
structures acting at all levels
Demons work
at all levels
© T.E. BEARDEN 1997

Figure 14. A vacuum engine is a set of spacetime demons working upon a mass system and its motion through time, at every level.

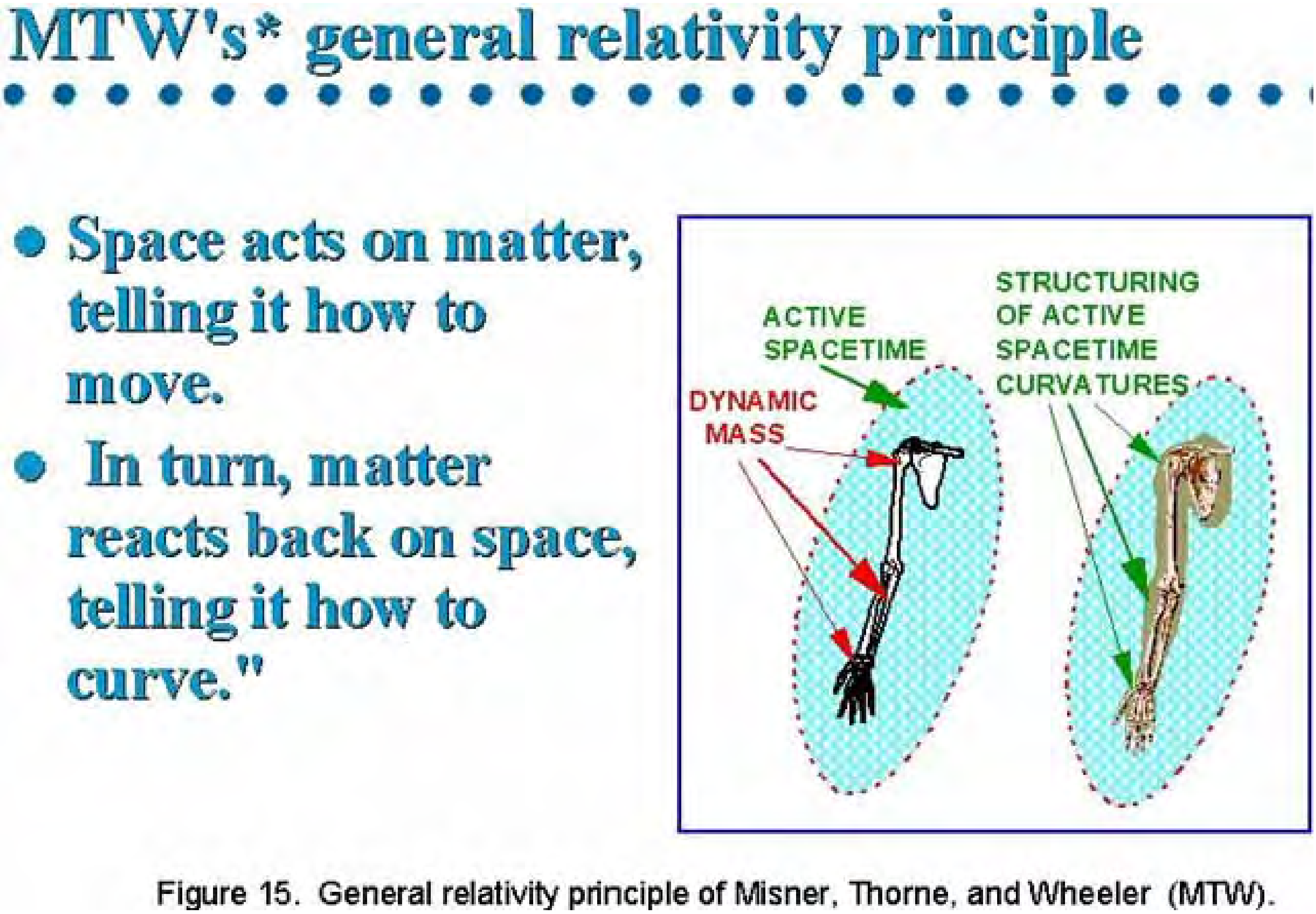

MTW's* general relativity principle

- Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

Figure 15. General relativity principle of Misner, Thorne, and Wheeler (MTW).

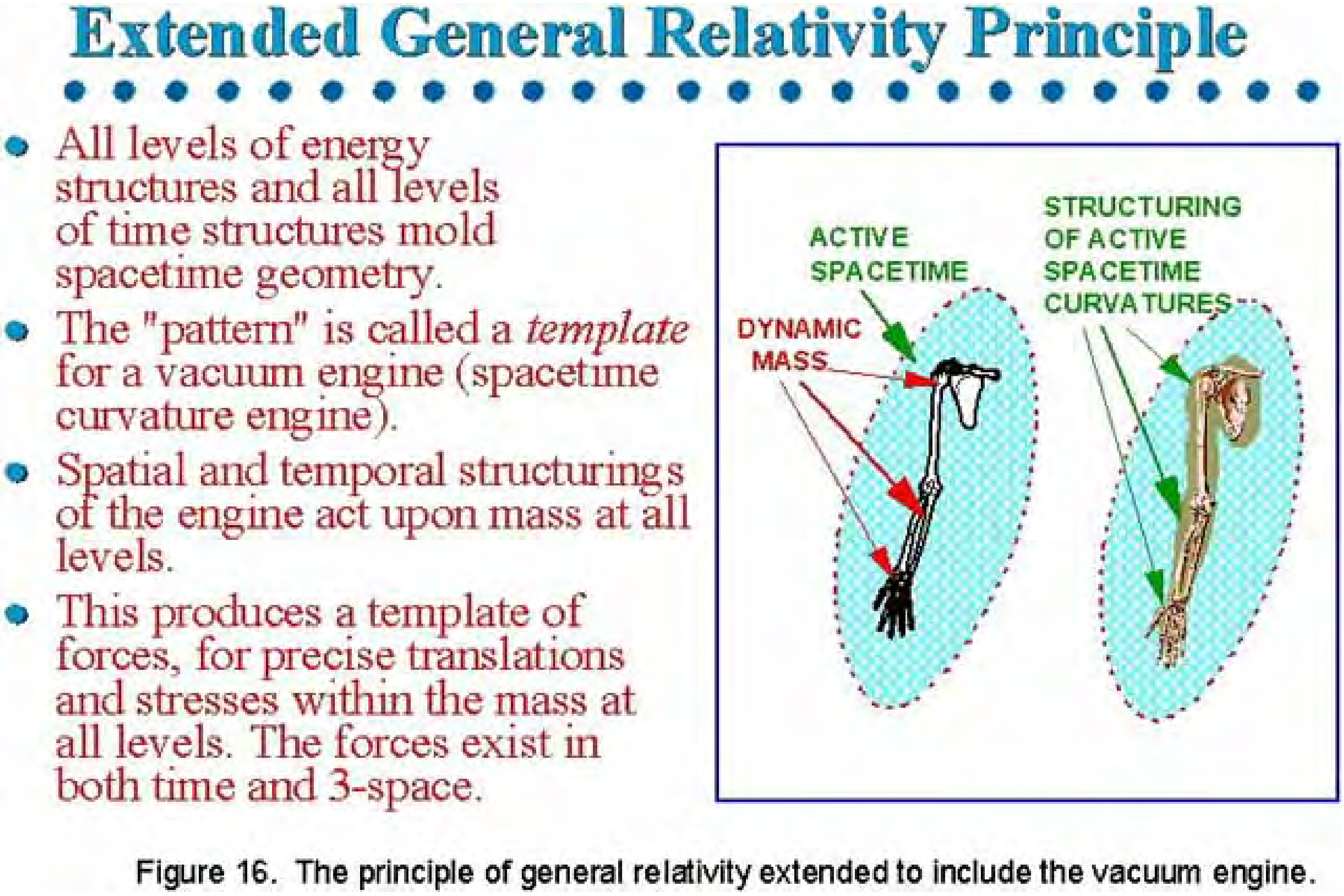

Extended General Relativity Principle

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon mass at all levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

Figure 16. The principle of general relativity extended to include the vacuum engine.

Active spacetime and a specific vacuum engine

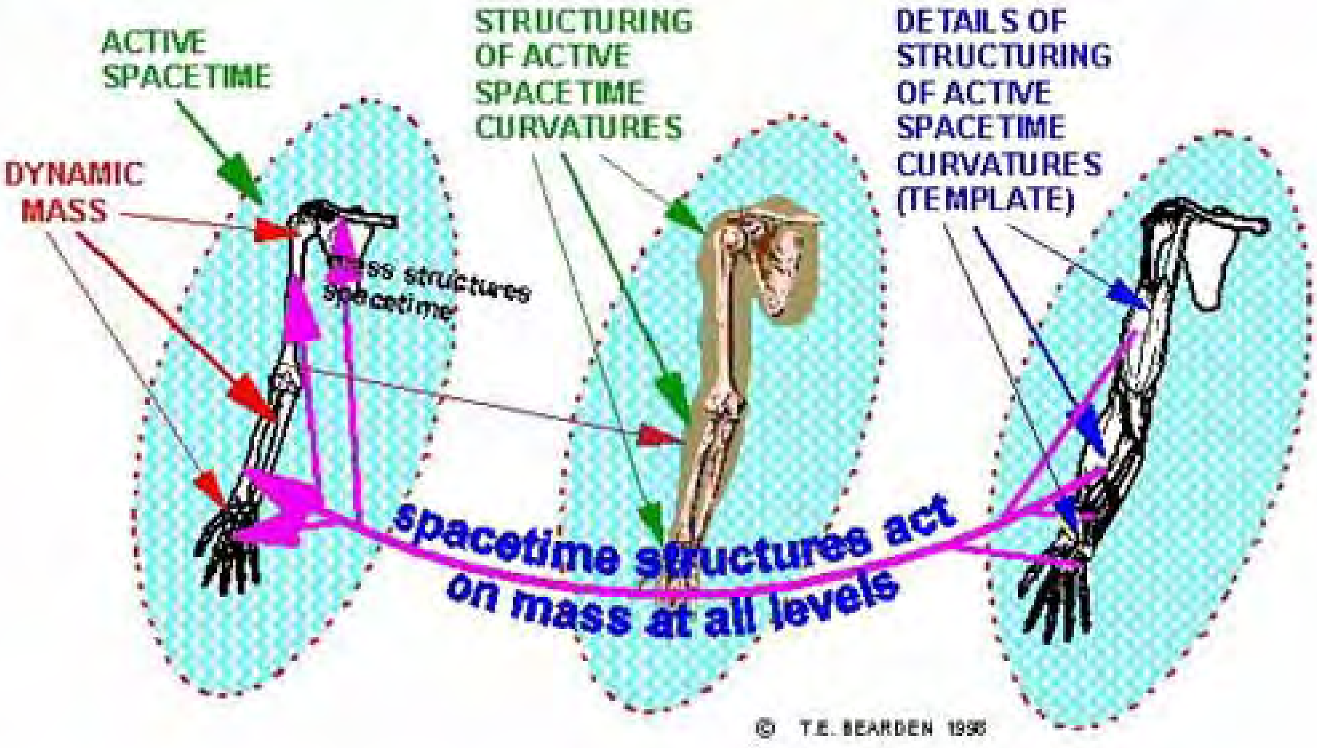
ACTIVE SPACETIME
DYNAMIC MASS
mass structures spacetime
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)
spacetime structures act on mass at all levels
© T.E. BEARDEN 1996

Figure 17. Active spacetime and a specific vacuum engine (spacetime curvature engine).

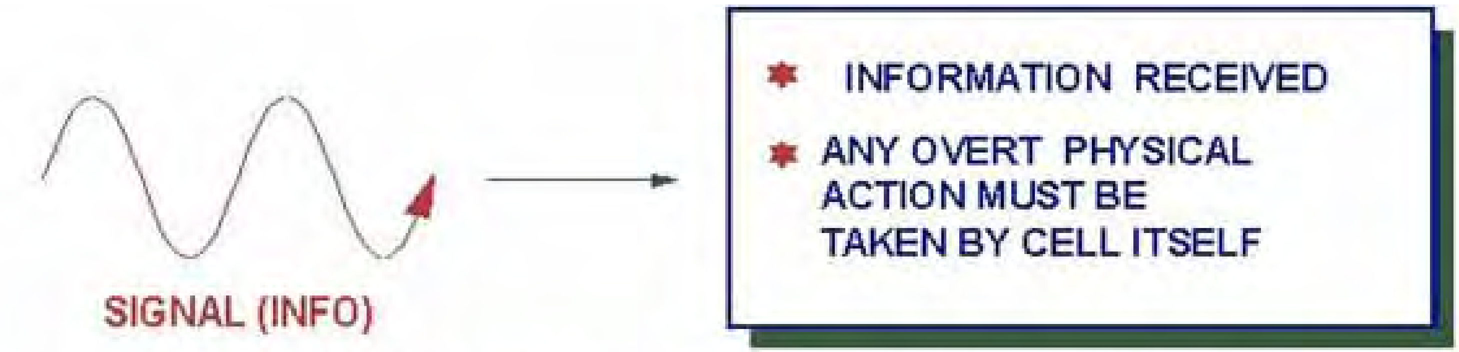
SIGNAL (INFO)
INFORMATION RECEIVED
ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY CELL ITSELF

A. *Cell must do the action itself; vacuum energy exchange is passive. Energy or fuel for doing the action must be added to cell externally.*

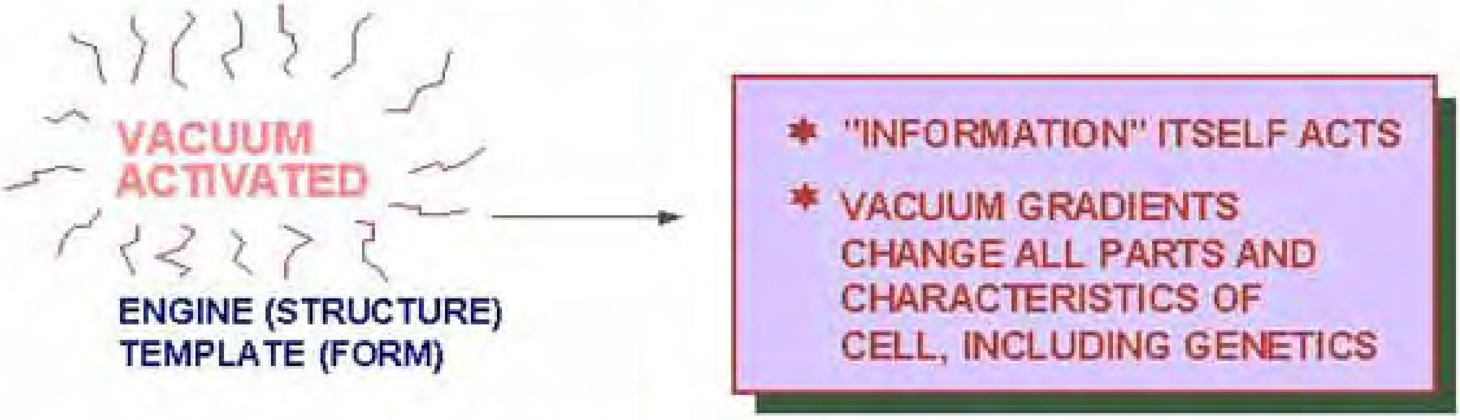
VACUUM ACTIVATED
ENGINE (STRUCTURE)
TEMPLATE (FORM)
"INFORMATION" ITSELF ACTS
VACUUM GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF CELL, INCLUDING GENETICS

B. *Cell is acted upon; vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to cell externally.*

Figure 18. Signal versus vacuum engine, as in Priore's cellular reversal.

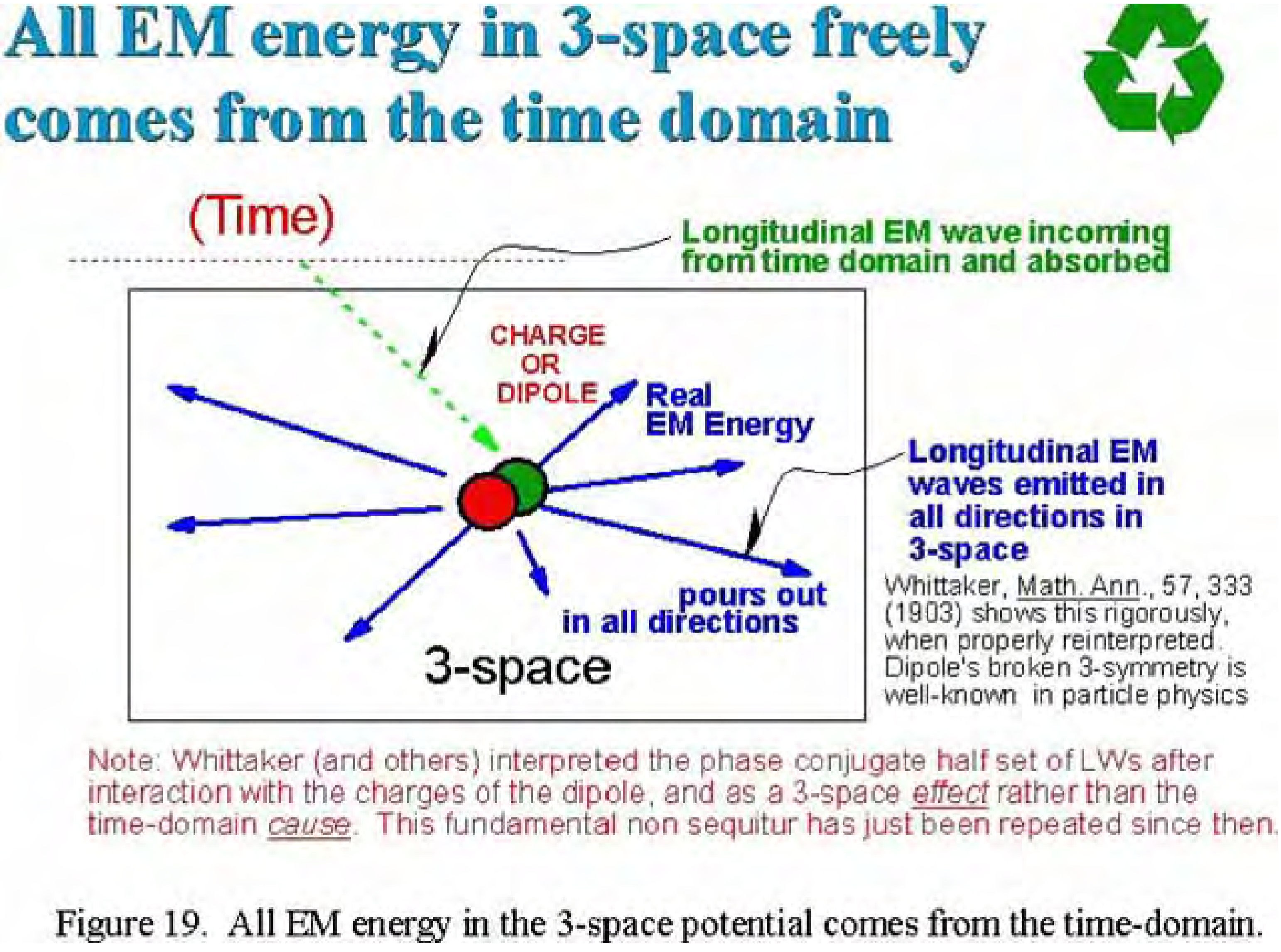

All EM energy in 3-space freely comes from the time domain

(Time)

Longitudinal EM wave incoming from time domain and absorbed

CHARGE OR DIPOLE

Real EM Energy

Longitudinal EM waves emitted in all directions in 3-space

pours out in all directions

3-space

Whittaker, Math. Ann., 57, 333 (1903) shows this rigorously, when properly reinterpreted. Dipole's broken 3-symmetry is well-known in particle physics

Note: Whittaker (and others) interpreted the phase conjugate half set of LWs after interaction with the charges of the dipole, and as a 3-space *effect* rather than the time-domain *cause*. This fundamental non sequitur has just been repeated since then.

Figure 19. All EM energy in the 3-space potential comes from the time-domain.

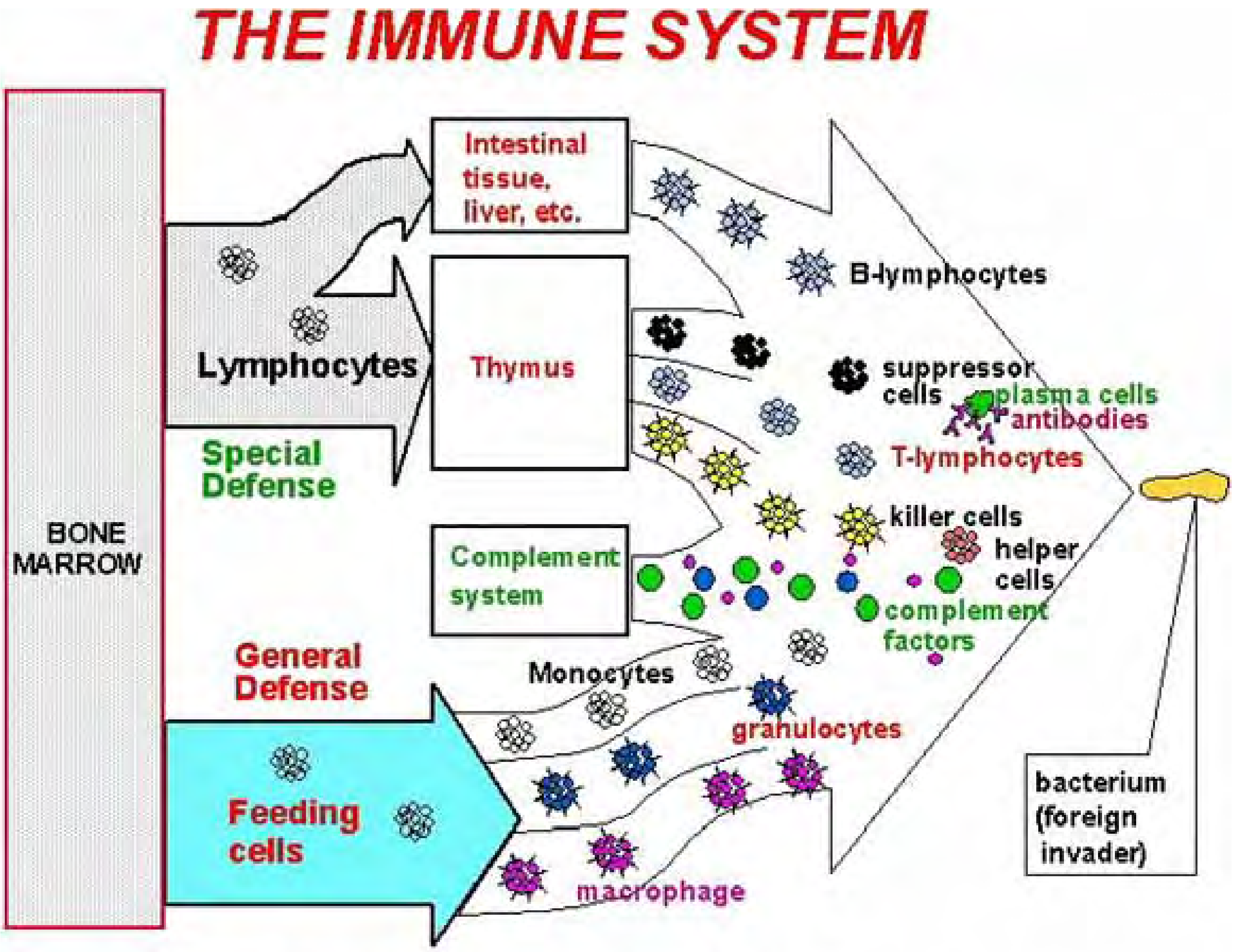
THE IMMUNE SYSTEM
BONE MARROW
Lymphocytes
Special Defense
General Defense
Feeding cells
Intestinal tissue, liver, etc.
Thymus
Complement system
Monocytes
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
granulocytes
macrophage
bacterium (foreign invader)

Figure 20. Immune system block diagram.

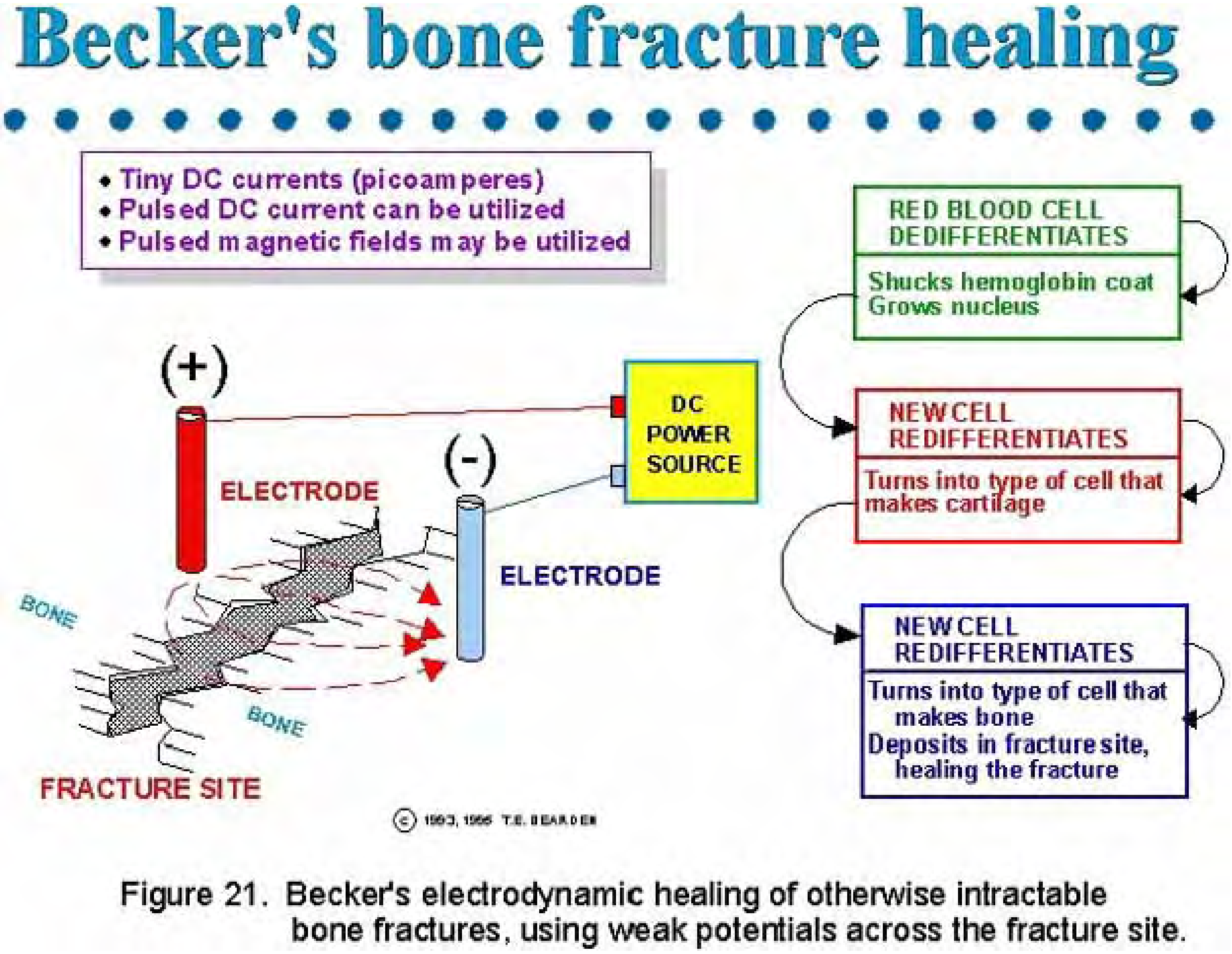

Becker's bone fracture healing

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1996 T.E. BEARDEN

Figure 21. Becker's electrodynamic healing of otherwise intractable bone fractures, using weak potentials across the fracture site.

Becker's theoretical DC control system involved with response to injury

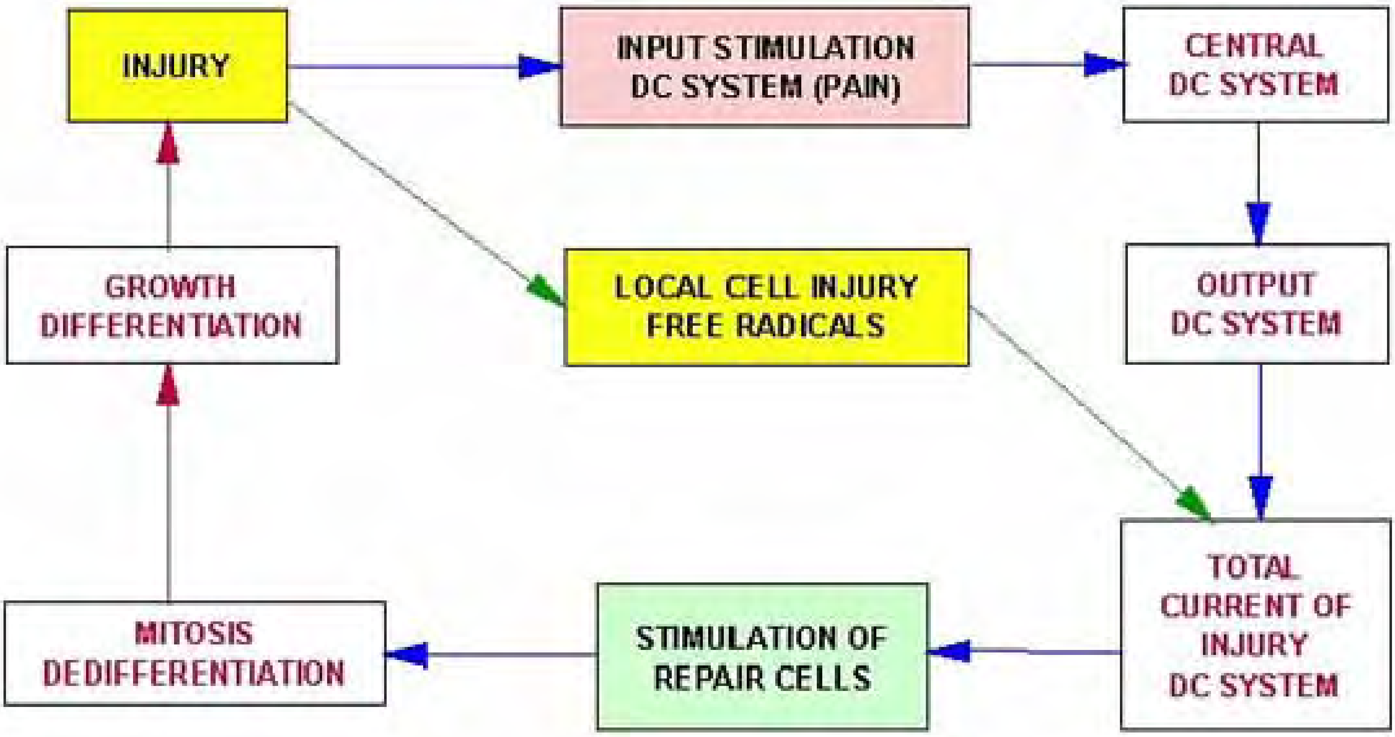
INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Figure 22. Becker's block diagram of DC control system involved with response to injury.

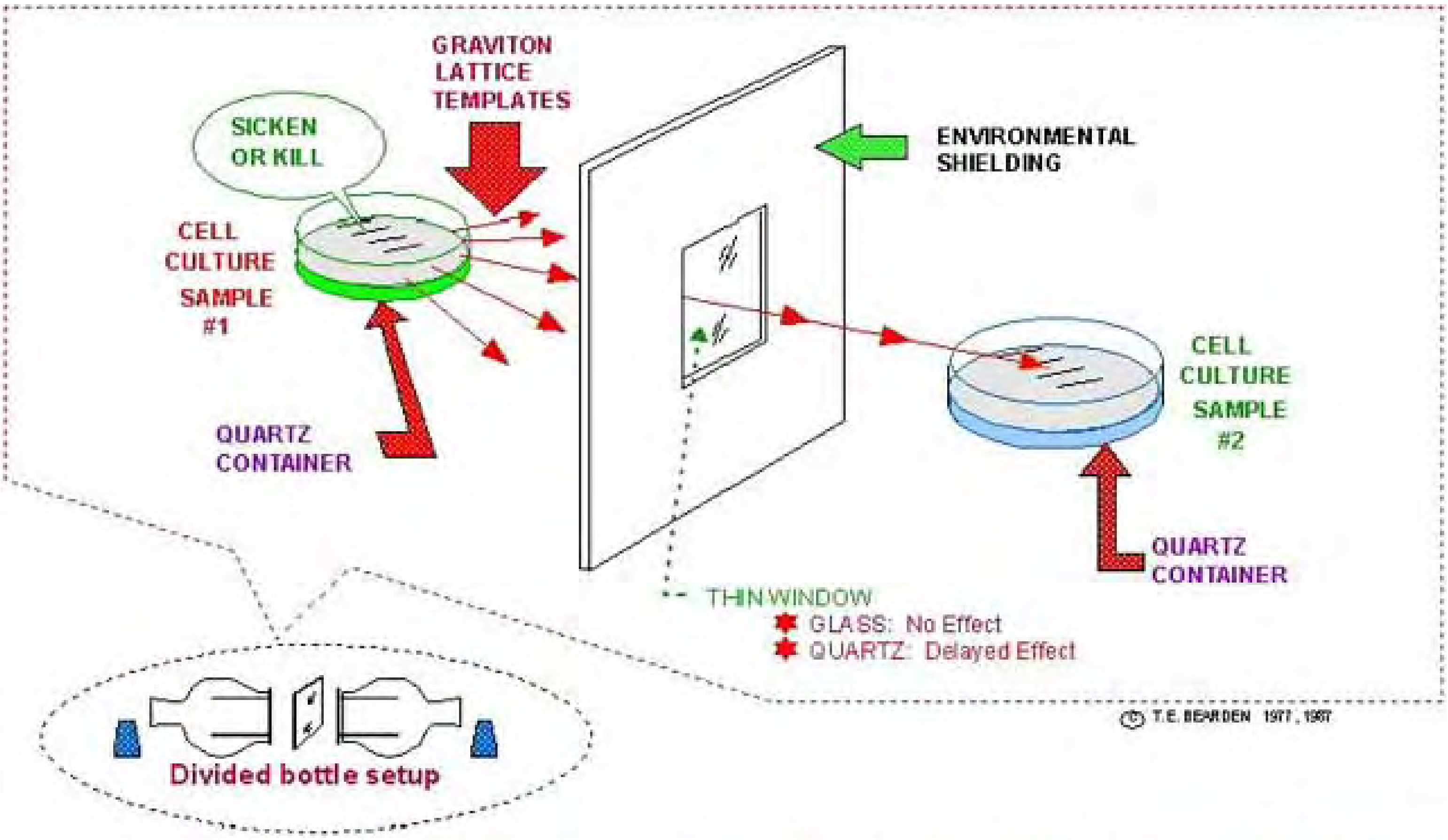
GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
© T.E. BEARDEN 1977, 1987
Divided bottle setup

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Figure 23. Kaznacheyev's induction of cellular disease and disorder at a distance, using novel electromagnetic means.

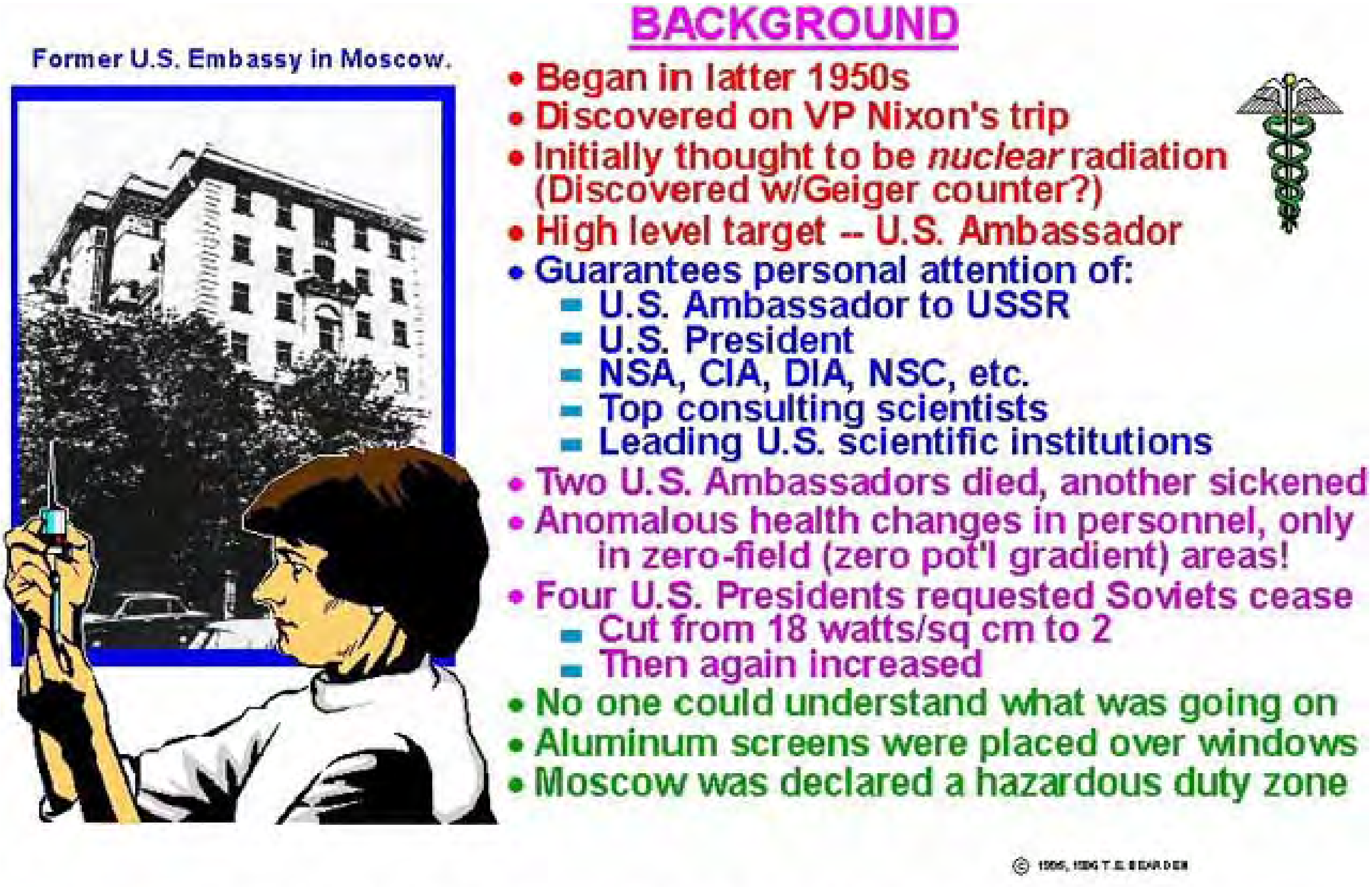

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Figure 24. Russian "microwave" radiation of the U.S. Embassy in Moscow.

Mechanism generating the flow of a mass through time

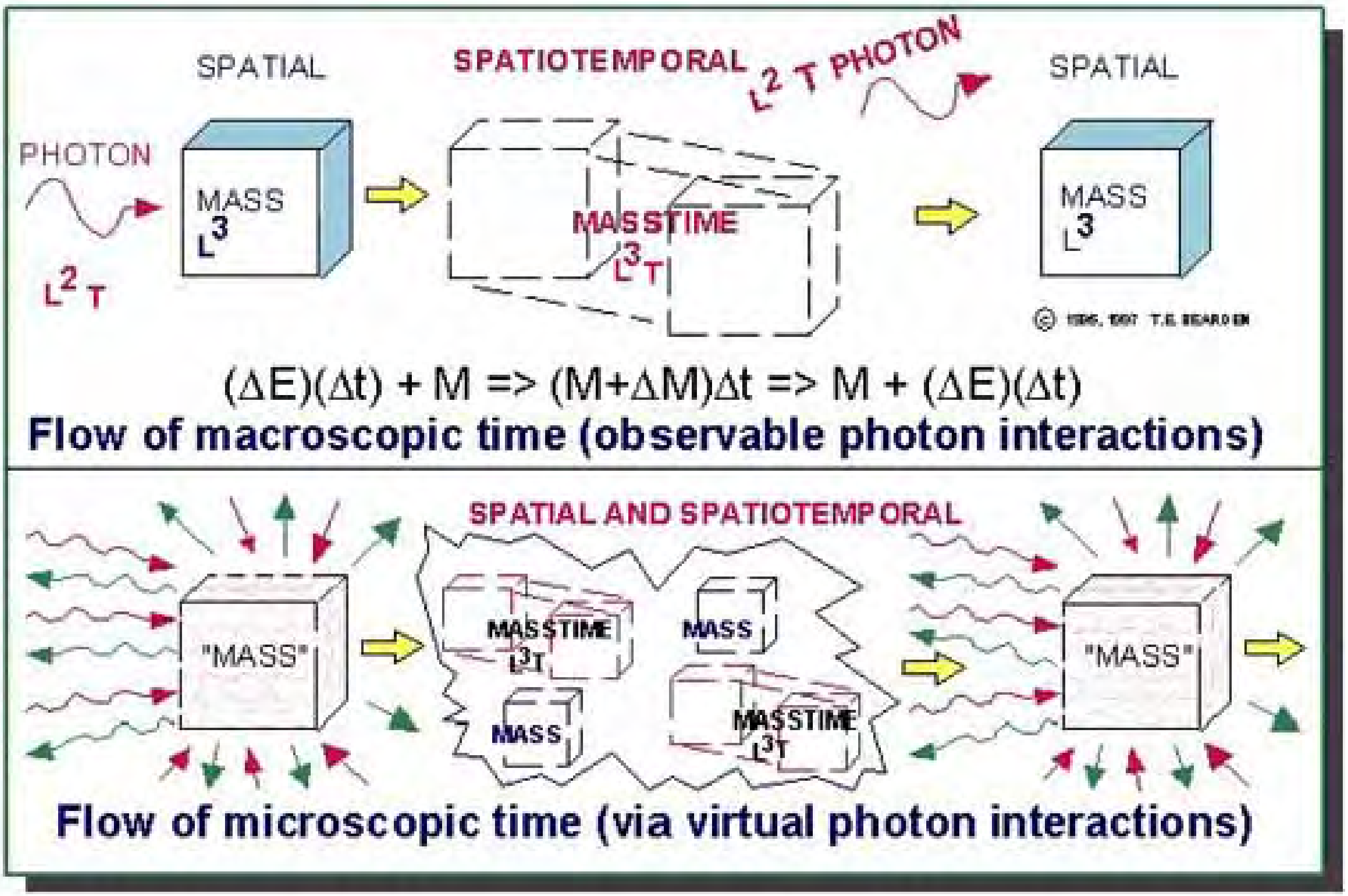
SPATIAL
SPATIOTEMPORAL
L^2T PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
MASS
L^3
© 1996, 1997 T.E. BEARDEN
(ΔE)(Δt) + M => (M+ΔM)Δt => M + (ΔE)(Δt)
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

Figure 25. Mechanism generating the flow of a mass through time.

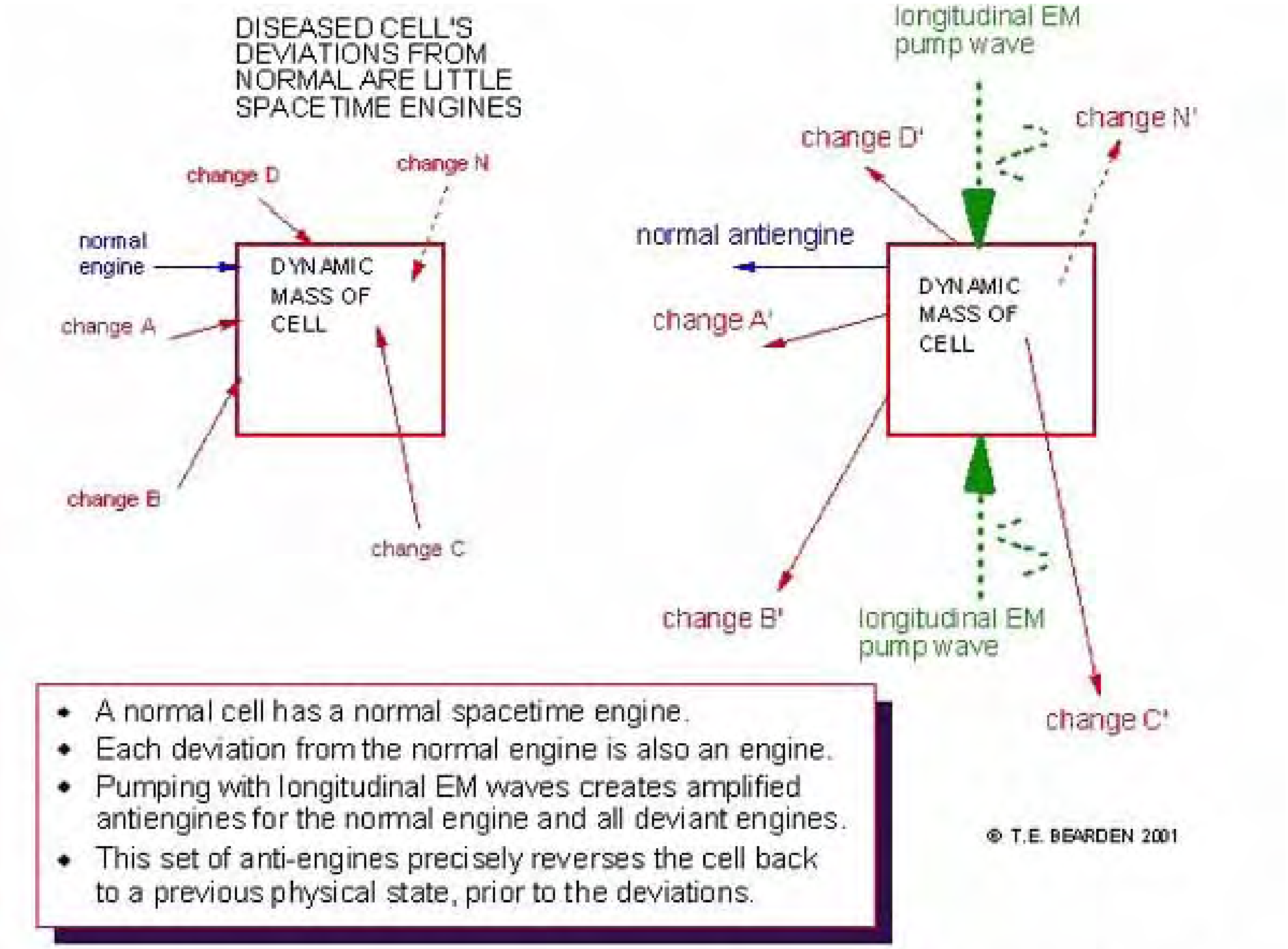
DISEASED CELL'S DEVIATIONS FROM NORMAL ARE LITTLE SPACETIME ENGINES
change D
change N
normal engine
DYNAMIC MASS OF CELL
change A
change B
change C
longitudinal EM pump wave
change D'
change N'
normal antiengine
DYNAMIC MASS OF CELL
change A'
change B'
longitudinal EM pump wave
change C'
© T.E. BEARDEN 2001

- A normal cell has a normal spacetime engine.
- Each deviation from the normal engine is also an engine.
- Pumping with longitudinal EM waves creates amplified antiengines for the normal engine and all deviant engines.
- This set of anti-engines precisely reverses the cell back to a previous physical state, prior to the deviations.

Figure 26. Producing a specific anti-engine for a specific cellular disease condition.

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

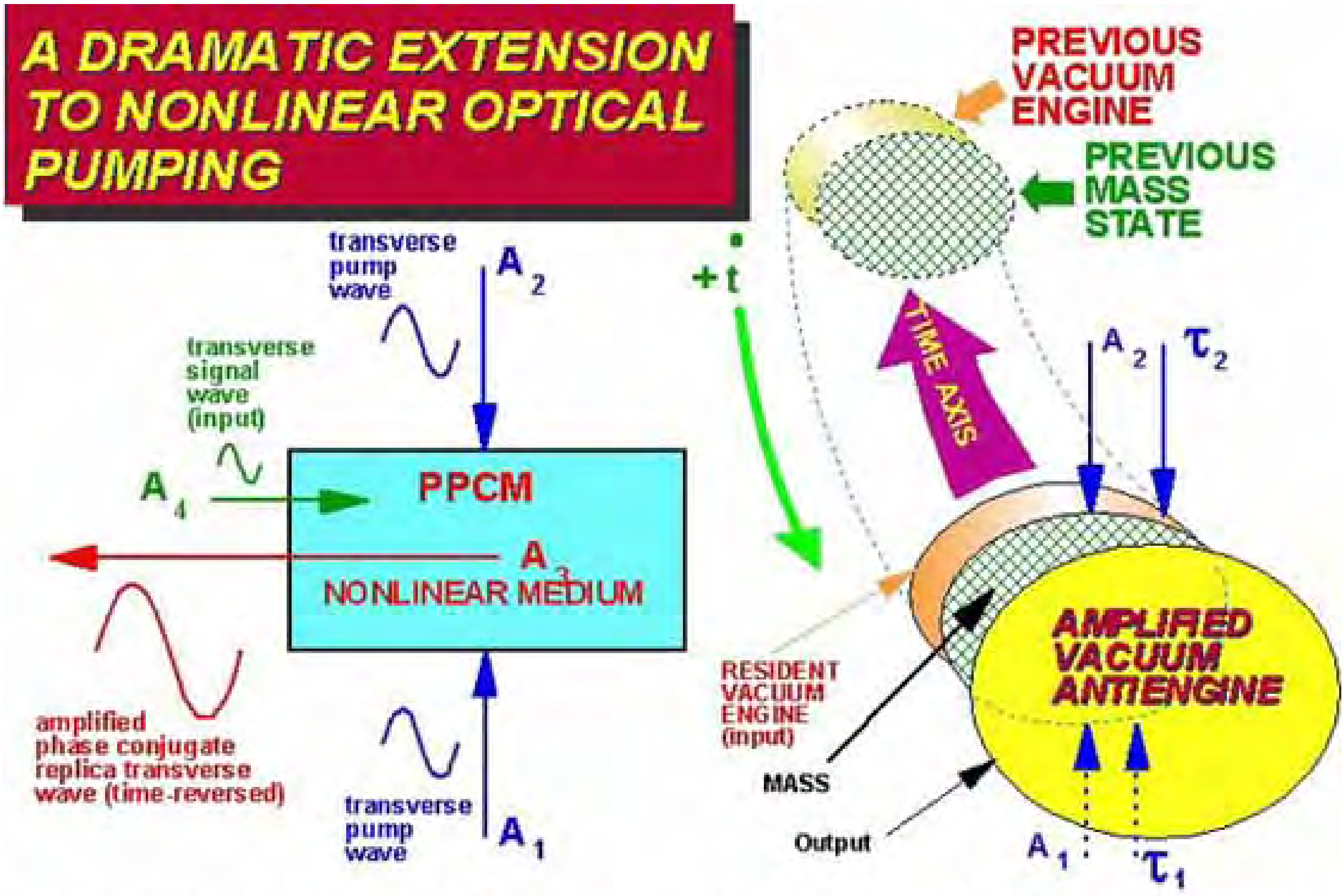
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
transverse pump wave
A2
transverse signal wave (input)
A4
PPCM
A3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A1
+t
TIME AXIS
A2
τ2
AMPLIFIED VACUUM ANTIENGINE
RESIDENT VACUUM ENGINE (input)
MASS
Output
A1
τ1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Figure 27. Pumping with longitudinal EM waves A also pumps in the time domain with time-polarized EM waves τ. This time-reverses a macroscopic mass and its dynamics.

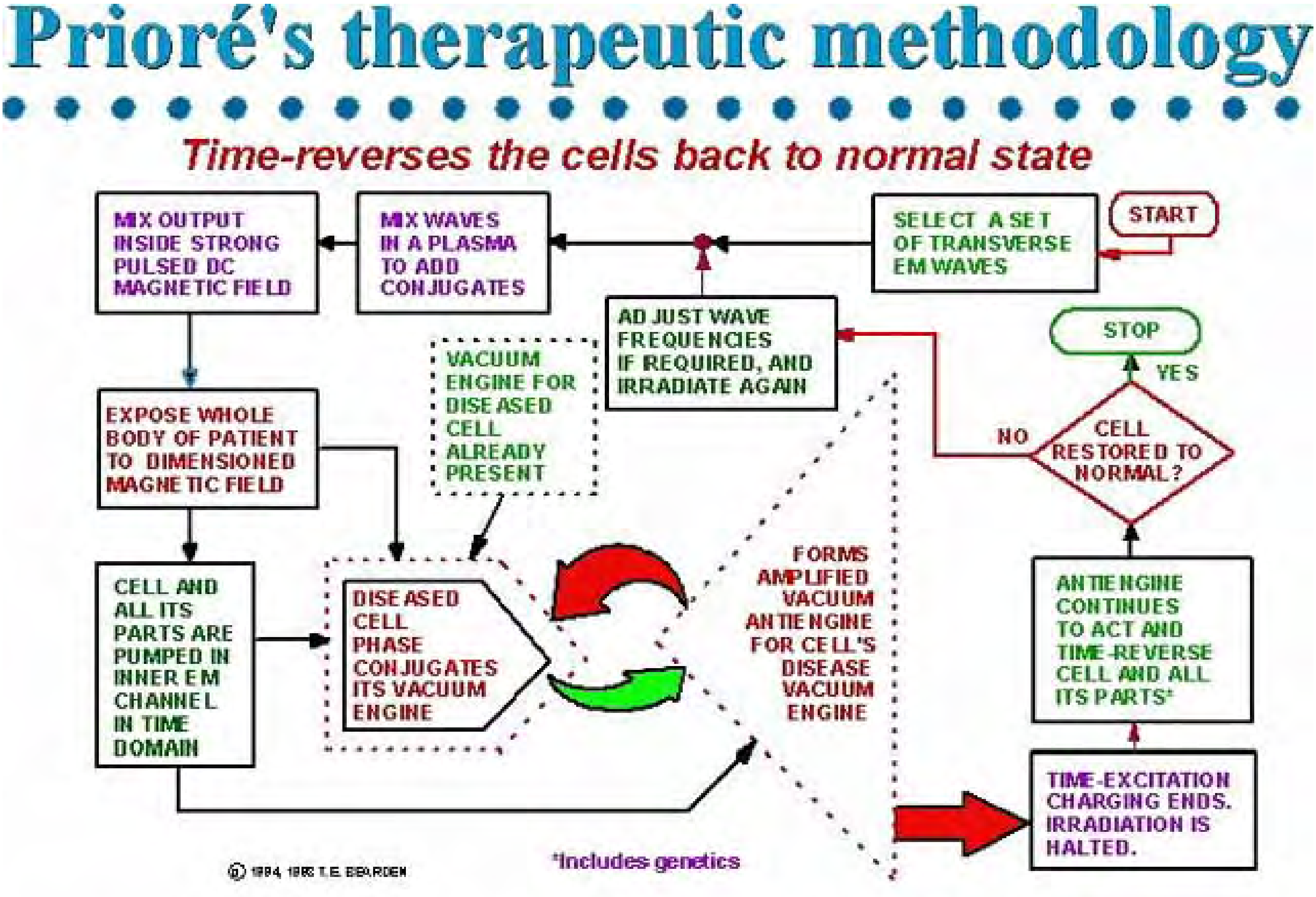
Prioré's therapeutic methodology
Time-reverses the cells back to normal state
START
SELECT A SET OF TRANSVERSE EM WAVES
MIX WAVES IN A PLASMA TO ADD CONJUGATES
MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD
ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
STOP
YES
NO
CELL RESTORED TO NORMAL?
ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*
TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.
© 1994, 1998 T.E. BEARDEN
*Includes genetics

Figure 28. Block diagram of operation of Prioré's methodology.

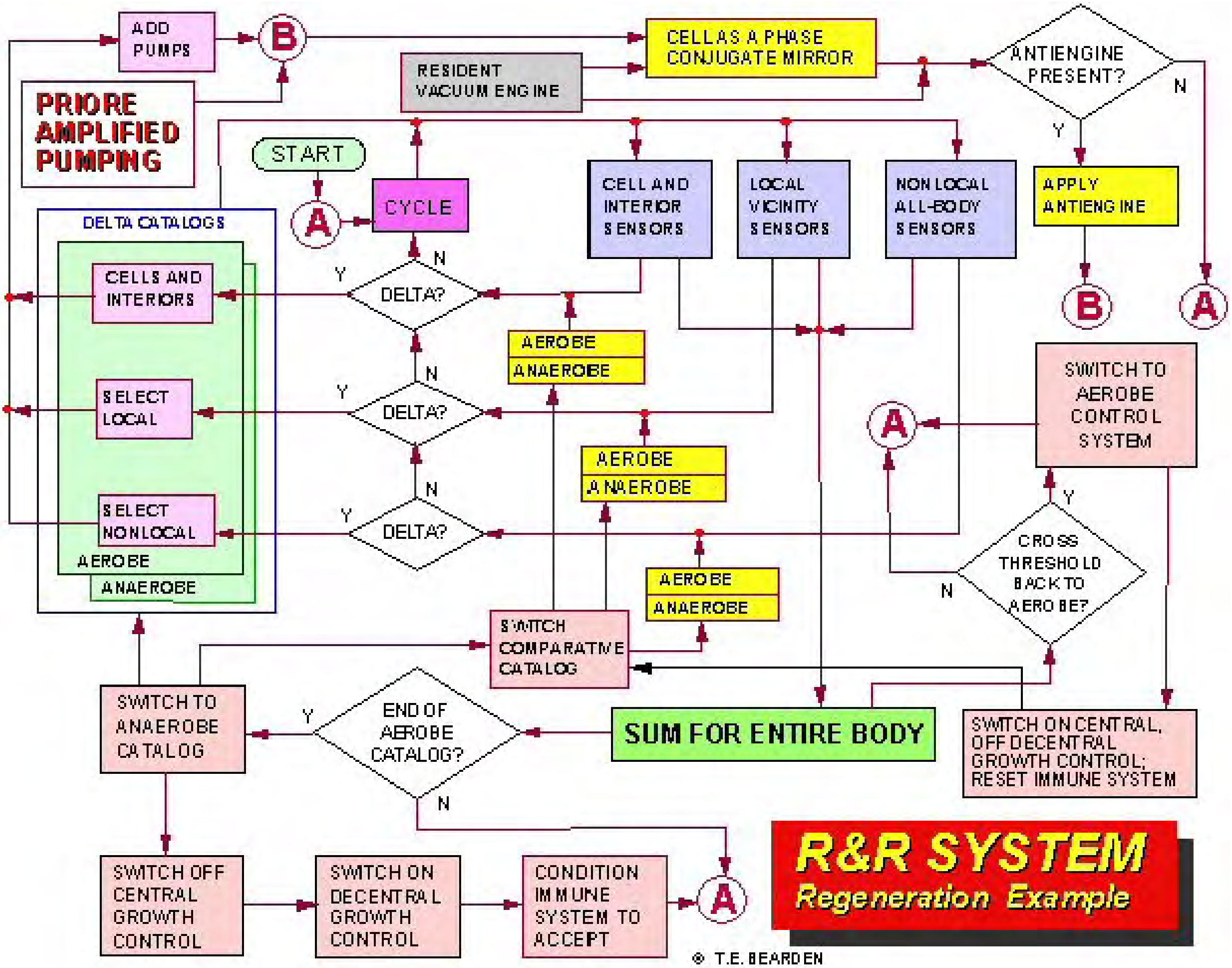
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
RESIDENT VACUUM ENGINE
ANTIENGINE PRESENT?
N
Y
PRIORE AMPLIFIED PUMPING
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NON LOCAL ALL-BODY SENSORS
APPLY ANTIENGINE
B
A
DELTA CATALOGS
CELLS AND INTERIORS
Y
DELTA?
N
AEROBE
ANAEROBE
SELECT LOCAL
Y
DELTA?
N
SWITCH TO AEROBE CONTROL SYSTEM
A
AEROBE
ANAEROBE
SELECT NONLOCAL
Y
DELTA?
N
Y
AEROBE
ANAEROBE
CROSS THRESHOLD BACK TO AEROBE?
N
AEROBE
ANAEROBE
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
R&R SYSTEM
Regeneration Example
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A

© T.E. BEARDEN 1998

Figure 29. Block diagram of the cellular regeneration system and its overall operation (by the present author).

Rife's great secret: Recursive magnification by iterative serial summation of vacuum engines

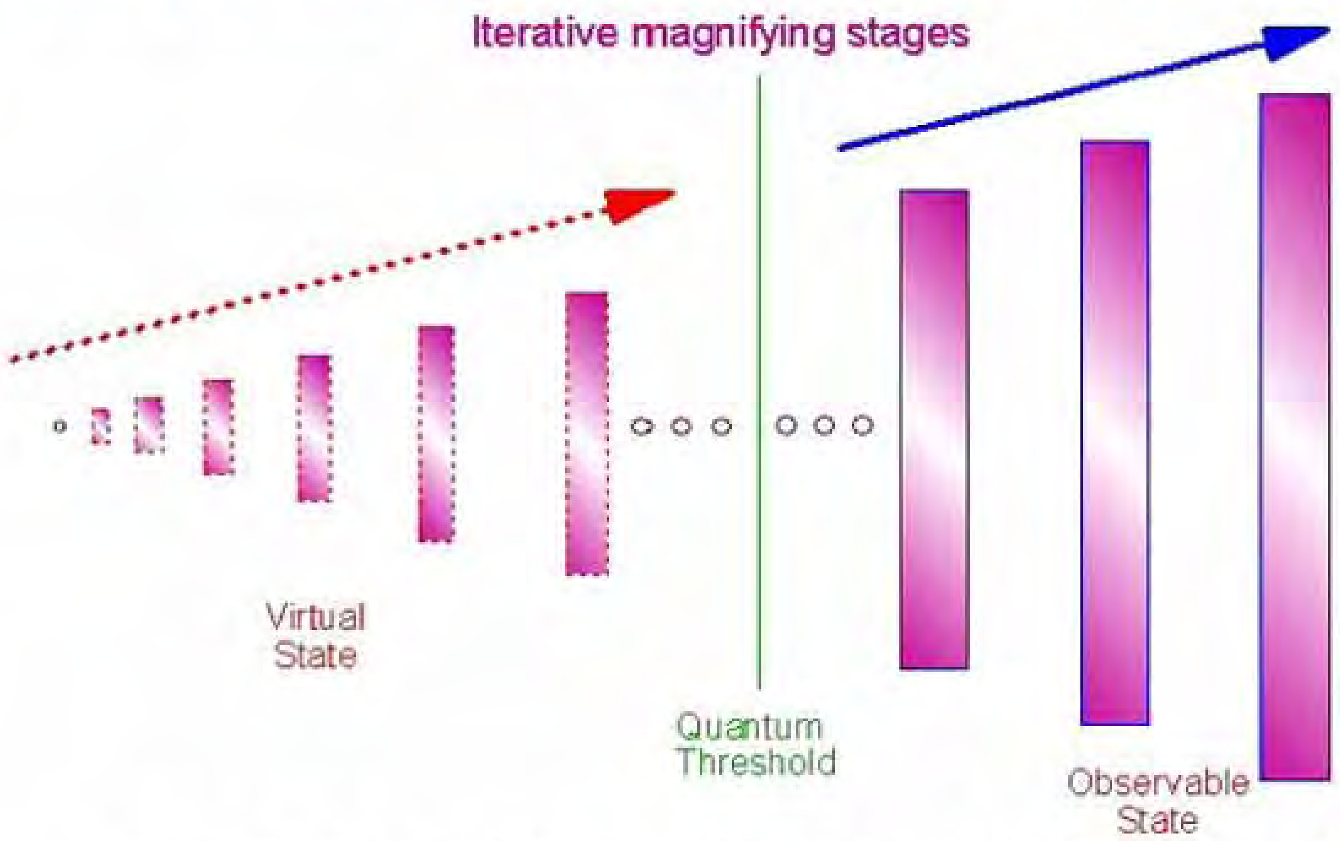
Iterative magnifying stages
Virtual State
Quantum Threshold
Observable State

Figure 30. Rife's great secret: Recursive magnification of the virtual state vacuum engines and their dynamics to the observable state.

Utilizing Scalar Electromagnetics To Tap Vacuum Energy

Floyd Sweet, Association of Distinguished American Scientists
2311 Big Cove Road, Huntsville, Alabama 358010

T. E. Bearden, Association of Distinguished American Scientists
2311 Big Cove Road, Huntsville, Alabama 35801

(Also published in: Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC Conf),
August 4-9, 1991, Boston, Massachusetts. Vol 4, Advanced Energy Concepts, pp. 370-375.)

Copyright © 1991 by T.E. Bearden & F. Sweet. All Rights Reserved.

ABSTRACT

Based on E.T. Whittaker's previously unnoticed 1903-1904 papers which established a hidden bidirectional EM wave structure in a standing forcefield free scalar potential, a method of directly engineering the ambient potential of the vacuum has been developed and realized experimentally.

Adding Whittaker's engineerable hidden variable theory to classical electro-magnetics, quantum mechanics, and general relativity produces supersets of each discipline. These supersets are joined by the common Whittaker subset, producing a unified field theory that is engineerable and tested.

By treating the nucleus of the atom as a pumped phase conjugate mirror, several working model energy units have been produced which excite and organize the local vacuum, increase the local virtual photon flux between local vacuum and nucleus, establish coherent self-oscillations between the local excited vacuum and the affected nuclei, utilize the self-oscillating standing wave for self-pumping of the nuclei/mirrors, introduce a very tiny signal wave to the mirrors, and output into an external load circuit a powerful, amplified, time-reversed phase conjugate replica wave at 60 Hertz frequency and nominal 120 volt sine wave power.

Several models have been built, ranging from 6 watts early on to one of 5 kilowatts. Both closed batteryless systems with damped positive feedback and open loop systems with battery-powered input have been successfully built. Open loop power gains of from 5×10^4 to 1.5×10^6 have been achieved.

Antigravity experiments have also been successfully conducted where the weight of the unit was reduced by 90% in controlled experiments, with a signal wave input of 175 microwatts and an output of 1 kilowatt.

The basic theory of the device is briefly explained and experimental results presented. In the demonstration session, a videotape of one operating open-loop unit with a 1.5×10^6 power gain is planned, as is the demonstration of an actual working model closed-loop system with a nominal rating of 500 watts, and without external power input of any kind.

The units are solid state, with no moving parts. Each of them comprises a unique form of self-powered vacuum triode of extraordinary gain, where the cathode power and plate power are freely furnished by the vacuum, and only a small grid signal need be furnished either from an external power source or by clamped positive feedback from the device's output. The output is negative energy, and some of its unique characteristics are pointed out.

Implications of the experimental application of the Sweet vacuum triode, the Bearden approach to the nucleus as a pumped phase conjugate mirror, and the unified field theory based on Whittaker's engineerable hidden variable scalar EM potential theory are also briefly addressed.

SCALAR ELECTROMAGNETICS

In 1837 Sir W.R. Hamilton said,

"The notion of time may be unfolded into an independent pure science... a science of pure time is possible."

As is well-known, the fundamental units utilized in physics are arbitrary. It is even possible to construct all of physics on a single unit, time. This oddity shows the truth in Hamilton's statement; it is even more odd, because quantum mechanically time is not an observable. This means that the observable world can be modeled completely in terms of the nonobservable, which is essentially what modern quantum mechanics is now doing.

Hamilton viewed his magnificent quaternions as essentially having accomplished the mathematical structuring of time. Maxwell's original EM theory, as is well-known, was modeled in Hamilton's *quaternions*, not in the highly curtailed Heaviside/Hertz vectors erroneously taught today as "Maxwell's theory."

Not a single one of the present so-called "Maxwell's" vector equations ever appeared in a book or paper by James Clerk Maxwell.

For some years the author has worked on an extended electromagnetics theory, involving the scalar component of the quaternion. [Ref. 1]

In Maxwell's original *quaternion* theory, this scalar component often remains when the directional components zero. Further, it then enfolds vectors and functions of vectors inside, in a hidden variable manner. Specifically, the author has patterned a unified field theory concept upon the previously unnoticed but remarkable early work of E.T. Whittaker. [Ref. 2]

In two fundamental papers in 1903 and 1904, Whittaker showed that all present vector EM can be replaced by scalar potential interferometry, and that bidirectional harmonic EM plane wave sets could be used to produce a standing wave of force-field-free potential (Figures 1 and 2).

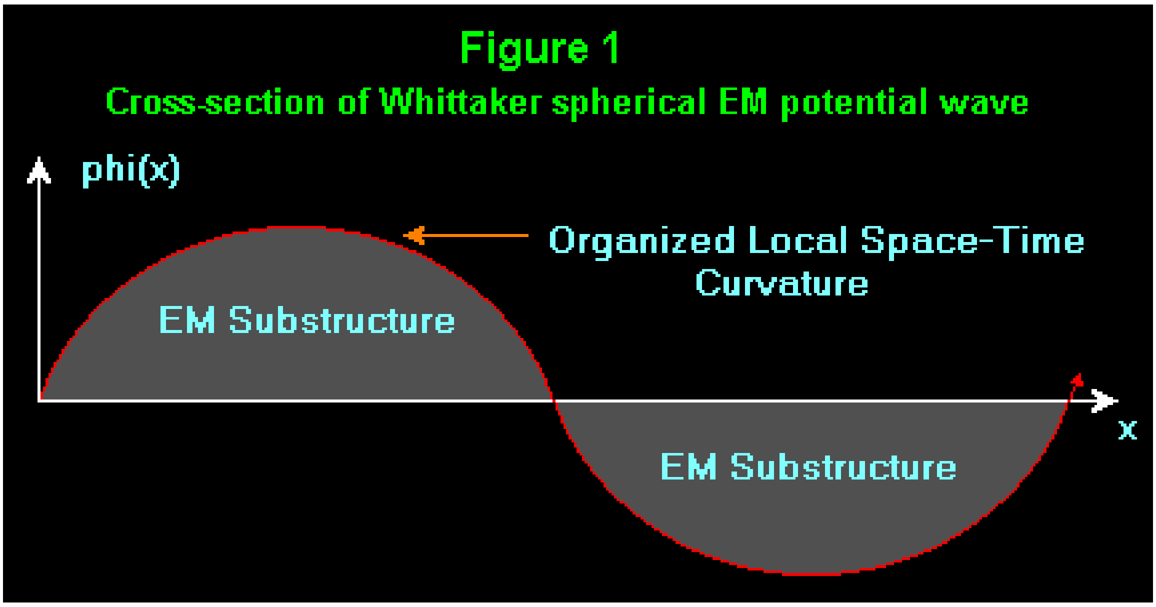
Figure 1
Cross-section of Whittaker spherical EM potential wave
phi(x)
Organized Local Space-Time Curvature
EM Substructure
EM Substructure
x

NOTE: SHAPE MAY BE ALL ABOVE OR ALL BELOW AMBIENT VACUUM POTENTIAL

Thus Whittaker anticipated the quantum mechanical Aharonov/Bohm effect by 55 years, including extending it to the macroscopic world instead of the mesoscopic realm where it has been established to date. [Ref. 3]

In modern terms, Whittaker showed how to turn EM wave energy into electrogravitational potential energy, then how to interfere two such scalar potential waves to recover electromagnetic energy, even at a distance. [Ref. 4]

This unrecognized work is of great importance: when applied to modern physics, it produces supersets of quantum mechanics (QM), classical electromagnetics (EM), and general relativity (GR).

Further, all three extended disciplines unify on their common Whittaker subset, in a testable and engineerable fashion. [Ref. 5]

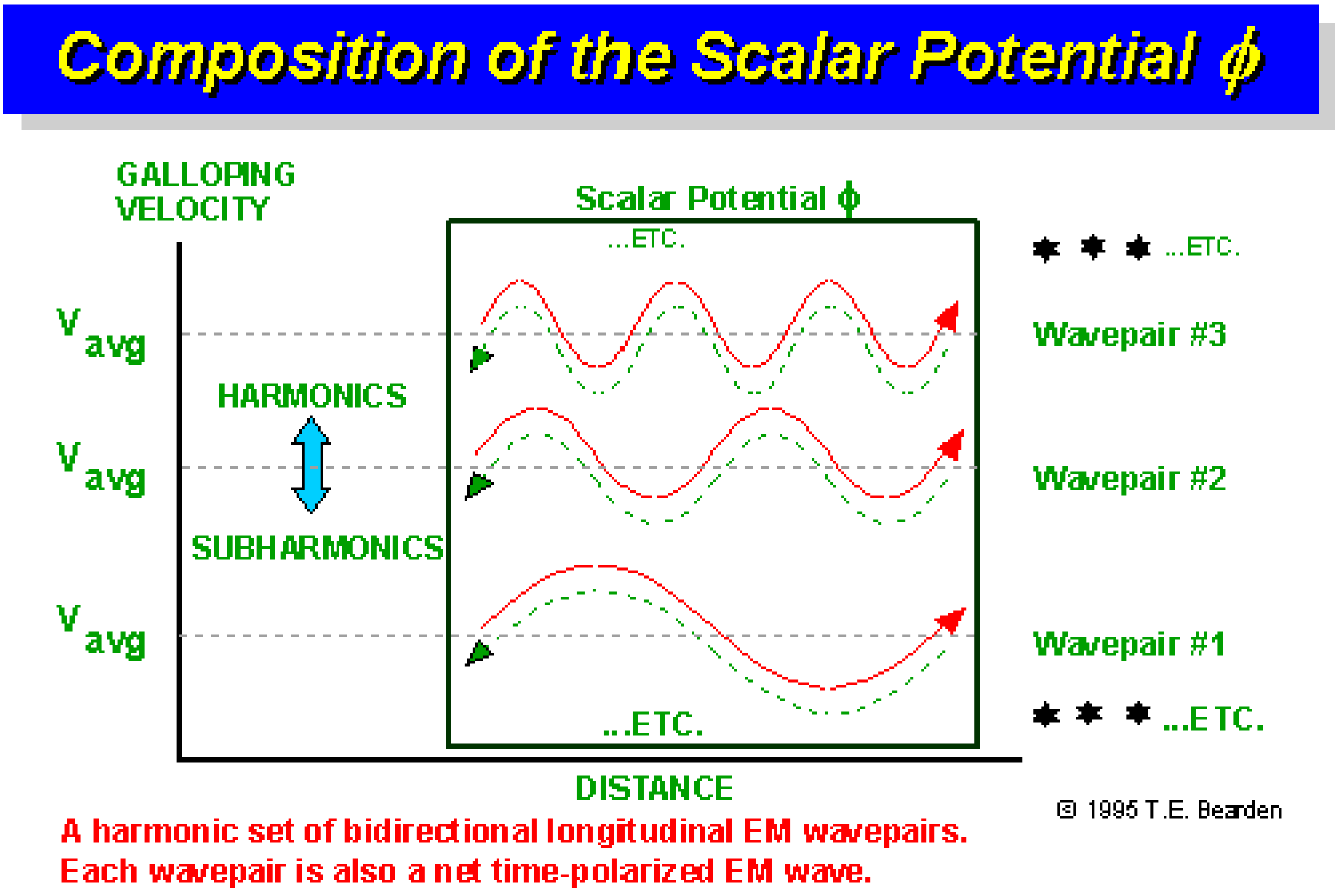
Composition of the Scalar Potential ϕ
GALLOPING VELOCITY
Scalar Potential ϕ
...ETC.
V avg
V avg
V avg
HARMONICS
SUBHARMONICS
...ETC.
Wavepair #3
Wavepair #2
Wavepair #1
...ETC.
DISTANCE
© 1995 T.E. Bearden
A harmonic set of bidirectional longitudinal EM wavepairs.
Each wavepair is also a net time-polarized EM wave.

Figure 2. Infolded EM plane wave structure of a Whittaker wave.

The Nucleus As a PPCM and Triode

The author has also considered the highly nonlinear nucleus of the atom as a pumped phase conjugate mirror (PPCM), having found no other consideration of same in the literature.

The author also dubbed a PPCM a "triode," since the amplified phase conjugate replica of the signal wave is much like the amplified output of a triode, and the signal wave input to a PPCM is much like a triode's grid signal input. The PPCM pump wave then corresponds roughly to the power input to the cathode and plate of a triode. [Ref. 6]

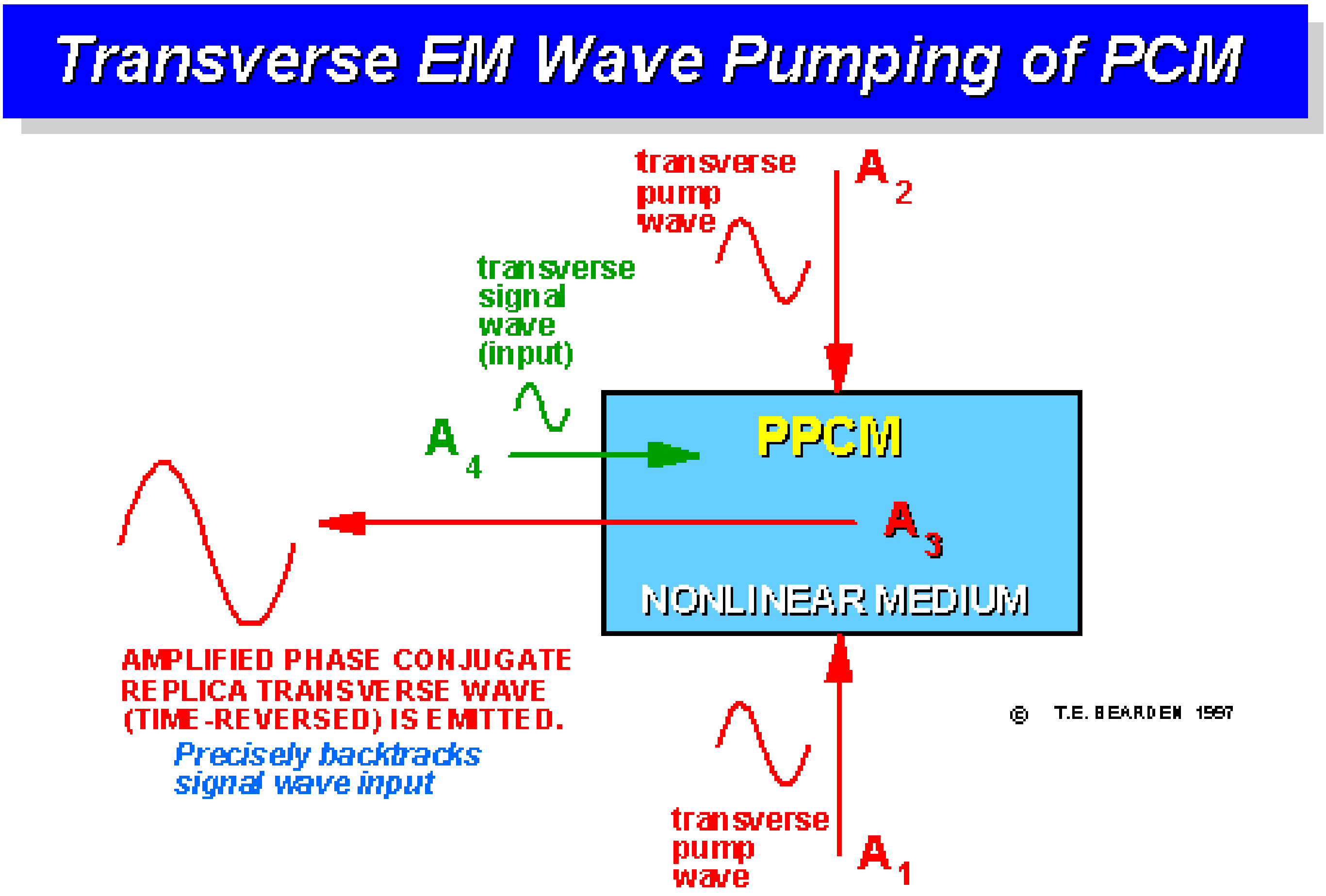
Transverse EM Wave Pumping of PCM
transverse pump wave
A2
transverse signal wave (input)
A4
PPCM
A3
NONLINEAR MEDIUM
AMPLIFIED PHASE CONJUGATE REPLICA TRANSVERSE WAVE (TIME-REVERSED) IS EMITTED.
Precisely backtracks signal wave input
© T.E. BEARDEN 1997
transverse pump wave
A1

Figure 3. A pumped phase conjugate mirror.

VACUUM TRIODE BACKGROUND

About seven years ago, the author was privileged to see and examine an invention of Mr. Floyd Sweet, that produced about 6 watts of electrical power from the vacuum itself.

This remarkable device, which the author dubbed a *vacuum triode*, in a single unit utilized most of the scalar EM concepts the author had so painfully and slowly formulated over the years.

Sweet, a brilliant inventor with a remarkable knowledge of magnetics, had utilized barium ferrite magnets and special coils to produce a solid-state device that successfully tapped the vacuum energy.

The author quickly put together a theoretical concept for the energy-tapping mechanism, based on treatment of the nucleus as a PPCM and Whittaker's scalar EM potential unified field theory. [Ref. 7]

The author furnished the technical concept, treating the nucleus as a pumped phase conjugate mirror, to the inventor along with copies of Whittaker's papers.

Sweet's Synthesis and Extension

Sweet is also a brilliant EM theoretician, working in four, five, or even six dimensions with ease. He immediately synthesized the entire PPCM and Whittaker theory, and developed a complete theoretical treatment of the device. [Ref. 8]

He also increased the nuclear potential utilized in the activated nuclei of the device, which increased the pumping energy and hence the energy output. He next produced an open-loop vacuum triode (VT) with an output of 500 watts, for an input of 33 microwatts.

Thereafter he produced several other models, including closed-loop systems and one with 5-KW output.

Purpose of This Paper

Our purpose is to explain the detailed scalar EM concept of the operation of the vacuum triode, since it is a universal method for cohering and tapping useful EM energy from the vacuum.

The author believes that this mechanism is the fundamental mechanism that must be invoked in any over-unity device that electromagnetically extracts vacuum energy as electromagnetic effects.

We also intend to demonstrate a full working model of the device at this conference. Sweet's detailed theoretical treatment will be completed and submitted to a major journal shortly, to complete the scientific exposition of the new methodology.

I must also express my deep admiration for my brilliant inventor colleague. It has been a privilege to work with him, though under great difficulties and at a distance. He has developed several other related devices that are of great importance to the emerging new physics of *vacuum engineering*, in the sense referred to by Lee. [Ref. 9]

To mention just one, he has produced a magnetic lens which apparently can directly display the vacuum's virtual particle flux, or a good analog of it. So far as I am aware, this is the only extant instrument today that can perform this feat.

When the vacuum triode has been proven to the scientific community, it is my intention to nominate Sweet for the Nobel Prize he so richly deserves, and seek high scientific endorsements for the recommendation.

INTRODUCTION

Entropy

As is well known, in any closed dynamic system the order existing in it will gradually be dissipated, as more and more interactions occur. This leads to the notion of *entropy* as the increasing disorder in such systems. The assumptions are

1. a closed system, and
2. a positive flow of time for the components of the system.

Actually no such thing as a completely closed system exists in nature. Every mass system is open to virtual particle flux exchange with the vacuum, for example, particularly in the nucleus of its atoms, where the bulk of its mass is located.

However, the closed system assumption is reasonably approximated by a great many systems which are in stable thermodynamic equilibrium, or nearly so.

On the other hand, in an open system far from thermodynamic equilibrium, the second law of thermodynamics does not necessarily apply, because the system violates both the closed system assumption and its equilibrium approximation. [Ref. 10]

Time Reversal

Since being discovered in 1972 in the open Soviet literature, the time-reversed (phase conjugate) EM wave has also been known. The phase conjugate EM wave is truly time-reversed, as has been shown experimentally. Since the time-reversed EM wave violates the second major assumption, the second law of thermodynamics need not necessarily hold for time-reversed entities.

Putting all this together, if one wishes negentropy and hence increased energy in a system, the candidate suggested would appear to be a system that was strongly

1. open loop,
2. time-reversed, and
3. far from thermodynamic equilibrium.

A good overview of time-reversal in physics has been provided by Sachs. [Ref. 11]

Engineering the Nucleus

Since the nucleus already provides a myriad of time-reversed processes, engineering the nucleus of an atom is a very good candidate for practical negentropy.

To engineer the nucleus directly, a Whittaker potential is first artificially constructed, by composing a harmonic set of phase-locked EM wave/antiwave pairs. It is accented that the antiwaves must be true phase conjugates; otherwise they will not constitute a gradient-free Whittaker standing potential wave.

In addition, at least one harmonic interval must be used, and additional harmonic sets are most desirable. The reason is that a space-time lattice must be formed in the vacuum, where the energy is additive spatially but oppositive in t-dot, the rate of flow of time, in the fourth dimension.

So a time-structure is required as well as a spatial structure, which is what is provided by ***n*** bidirectional harmonic Whittaker EM wave sets, where ***n*** is an integer greater than 1.

Once a specific Whittaker structure has been chosen, the local lattice of space-time is established. This establishes phase-locked lattice groupings of coupled photon/antiphoton pairs, or of *gravitons*. In turn, this Whittaker-structured vacuum now contains specific graviton *vacuum engines*, which directly engineer and structure the vacuum's virtual particle flux (VPF) exchange with the nucleus. [Ref. 12]

The nucleus is highly nonlinear, hence strongly phase conjugative, or time-reversed. The ambient potential of the local vacuum surrounding the nucleus is in a violent virtual photon exchange with it, accounting for its electrical charge.

Since the preponderant charge is positive, from the viewpoint of the ordinary light observer whose light interacts with electron shells, the nucleus may be taken to exhibit time reversal (phase conjugation).

Energy, Time, and Gravitons

We take the definition of "energy" to be fundamentally an ordering imposed upon the VPF of vacuum. We take photon scattering from the electron shells of atoms to be the fundamental exterior mechanism producing forward flow of external observer time.

It then follows that "time's arrow" for the EM observer is due to the universal scattering of photons from electron shells.

In this view, forward (positive) time flow and entropy are due to the same primary action: photon scattering from electron shells. It is unfortunate that the concept of "positive" energy has been tied to, and defined in terms of, the scattering and dissipation of VPF order as *work*, or energy expended.

Via the standard labeling, then, negative energy is the reconstitution of order in the vacuum VPF. It should be noted that, in a PPCM, dissipative or external pump wave stress energy can be scavenged and re-emitted in perfect order as the phase conjugate replica. This is a *negentropic* process, for it is capable of turning disorder into order. [Ref. 13]

In the time-reversed PPCM nucleus, we should expect to see appreciable *negative energy*—that is, energy removed from the EM scattering domain. This includes the binding energy of the nucleus, and the gravitational (G) potential energy of the EM energy removed from the "scattering interaction realm" and locked into the mass.

Following Sakharov, we hold that the G-field is not a fundamental field of nature, but a composite caused by, or made from, other fields. [Ref. 14]

To the first order, we assume the G-potential is comprised of coupled photon/antiphoton pairs, on the average, where the statistical coupled spin-2 photon/antiphoton pair is a *graviton*. [Ref. 15]

We follow the modern view of the field: because of vacuum fluctuations, rigorously one no longer speaks of "the" field, but of the *probability of a particular field configuration*. [Ref. 16]

We also hold the vacuum to be composed of potentials, and regard the three terms *space-time, vacuum*, and *scalar potential* as essentially synonymous.

Newton's Third Law and the Detection Process

The VPF EM stress of the local vacuum immediately surrounding the nucleus may be decomposed a la Whittaker into opposing bidirectional EM plane waves/forces. Thus the nonlinear nucleus may be regarded as a pumped phase conjugate mirror, normally with a gain of one.

In this view, Newton's third law reaction force is generated because the so-called "photon" interaction with an atom is in fact a graviton reaction involving a photon/antiphoton pair that is decoupled.

The decoupled photon normally is absorbed and reradiated by an orbital electron, while the decoupled antiphoton interacts with the nucleus, producing a time-reversed twin of the external force—or, in other words, Newton's third law reaction force, which gives a slight recoil of the nucleus.

Half of every measurement physicists normally make is discarded, with the missing half accounting only for Newtonian reaction in the meter or instrument, which is usually ignored.

The fact that half of our measurement interactions are ignored is occasionally discovered and noted by physicists, who may even write a paper pointing it out, but no change is instituted in the foundations. [Ref. 17 and Ref 18]

Semiconducting Vacuum and Self-Oscillation

The vacuum immediately surrounding the nucleus is structured by the nucleus, and is itself nonlinear and capable of acting as a *semiconductor*. [Ref. 19]

Since both this immediately local semiconductor vacuum and the nucleus it surrounds are highly nonlinear, then *nonlinear resonance* can conceivably be established between them.

Further, since the resonating system in such case is an open system away from thermodynamic equilibrium, the oscillation can be self-sustaining. Such self-oscillation of the pumping of a PCM is already well-known in the nonlinear optical literature, particularly with compounds containing barium. [Ref. 20]

THE VACUUM TRIODE

The Basic Concept

As is well-known, a stress can be decomposed into opposing sets of forces. But quantum mechanically, the forces we are interested in with our work here are all caused *electromagnetically*, by the exchange of virtual photons. Even mechanical force, according to QM, is caused in this manner. Thus opposing electromagnetic or mechanical "stress" sets of bidirectional EM forces are microscopically equivalent to the notion of *pump waves* in nonlinear optics.

Hence under the proper conditions, it follows that trapped EM stress energy of the vacuum can be utilized to "pump" the nucleus.

Treating the stress-pumped nonlinear nucleus as a PPCM, it follows that the stress energy of the vacuum can be tapped by a 4-wave mixing mechanism in the atomic nucleus, to provide amplified phase conjugate EM wave outputs from the atom in response to small signal wave inputs. [Ref. 21]

In the proper nonlinear material, the material may act as a PPCM, in which case there exists a suitable connection between the material's atomic nuclei and its external electromagnetic lattice bonds, and the amplified phase conjugate replica wave generated in the nucleus will be emitted from the material as an EM wave field. This field can then be tapped by suitable means and output to an external load circuit.

Block Diagram of the Vacuum Triode

Figure 4 shows a basic block diagram of the vacuum triode process, utilized by Sweet in several laboratory vacuum energy devices. These devices have ranged from a nominal 500 watt output in a 6-lb. device to 5 kilowatts for a heavier unit. Gains have ranged from 50,000 to 1,500,000 for open-loop systems. Both open-loop and closed-loop systems have been built and tested.

Figure 4 shows a combined block diagram for either a closed-loop or open-loop system. In the open-loop system, a barium ferrite magnetic material is used as a pumped phase conjugate mirror.

In the "standard" design, two opposing PPCMs are used. The advantage of this dual combination is the use of *self-targeting* (repetitive phase conjugation, signal by signal). This has the effect of

1. stabilizing the Whittaker field, and
2. producing a quantum potential between the two mirrors, so the mirrors and the Whittaker potential between them are essentially one single space-time entity.

Discussion of a quantum potential is beyond the scope of this paper, but the technical mechanism for creating one has been previously presented by the author on several occasions. [Ref. 22]

First we will explain the open-loop operation of the vacuum triode. In Figure 4, on the right an external 60 Hz, nominal 10 volt AC sine wave of several tens of microwatts in power is input into the stabilized field of the barium ferrite magnet structure, where it modulates the field, producing a signal wave input into the atoms of the material.

In the top right block, the EM signal wave interacts with the electron shell of an atom, which in turn is EM-coupled to the nucleus.

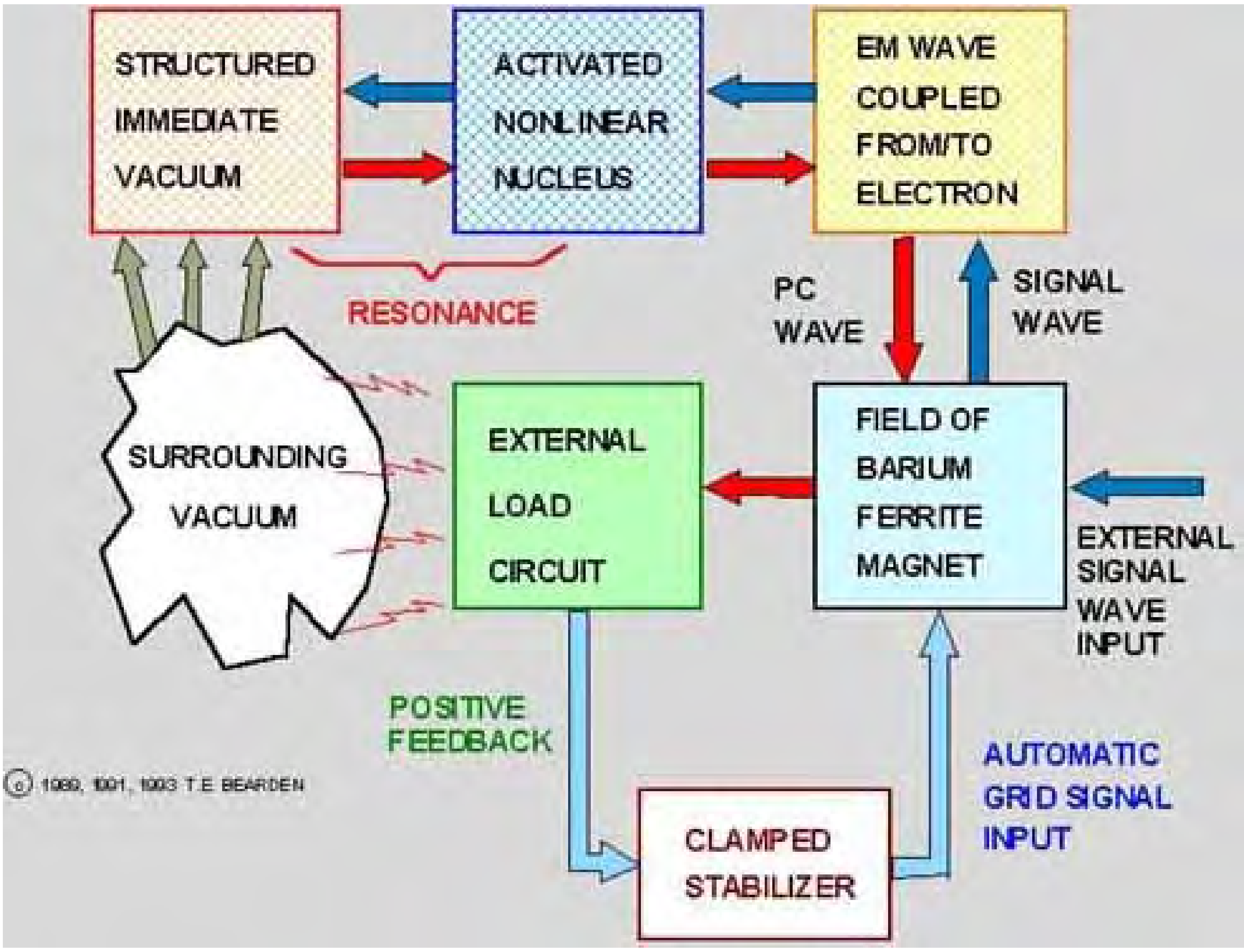
STRUCTURED IMMEDIATE VACUUM
ACTIVATED NONLINEAR NUCLEUS
EM WAVE COUPLED FROM/TO ELECTRON
RESONANCE
PC WAVE
SIGNAL WAVE
SURROUNDING VACUUM
EXTERNAL LOAD CIRCUIT
FIELD OF BARIUM FERRITE MAGNET
EXTERNAL SIGNAL WAVE INPUT
POSITIVE FEEDBACK
© 1989, 1991, 1993 T.E. BEARDEN
CLAMPED STABILIZER
AUTOMATIC GRID SIGNAL INPUT

Figure 4. The Vacuum Triode Concept.

Thus an EM signal wave is input to the nucleus of the atom, which is highly nonlinear (middle top block). Earlier, Sweet had specifically conditioned the atomic nuclei with a proprietary process, wherein in the barium nucleus a trapped 60 Hz scalar EM spherical wave resonance (self-oscillation) exists between the structured semiconductor vacuum immediately surrounding the Ba nucleus.

In the same *activation* process, the ambient potential of the surrounding vacuum was raised and stabilized, in the two leftmost blocks.

At this point the nonlinear nucleus is effectively self-pumped by the trapped, excited, spherical scalar wave oscillation between the structured semiconductor vacuum and the nucleus. The nucleus is now a strongly pumped phase conjugate mirror.

Consequently, when the signal wave input arrives, the PPCM nucleus emits an amplified phase conjugate replica (PCR) wave, which precisely backtracks the input signal wave. This precise backtracking (perfect retroreflection) is referred to as the "distortion correction theorem."

In short, the powerful PCR wave returns precisely toward the external source, passing through the electron shells and arriving in the perturbed barium ferrite magnet assembly field, where it perturbs the field.

A transformer-like system then extracts this magnetic field perturbation and conducts it to the external load circuit.

However, the PCR contains *negative* energy. Short of the load, the internal circuits run cool, rather than heating. This is a signature of a true vacuum energy tapping device.

Indeed, if the output leads of the Sweet vacuum triode are physically shorted together, a brilliant flash occurs, and the leads instantly ice as if dipped in liquid oxygen. This is another signature of the true negentropic over-unity vacuum tap.

Note that the energy extracted from the semiconducting vacuum adjacent to the nucleus is just instantly replaced by the surrounding vacuum's inexhaustible energy pool. This is an open-loop system, with a hidden energy source: the intense virtual particle flux of the vacuum's ambient charge.

It is not possible to exhaust that flux, which is often calculated to have an energy density of some 10^{100} or more grams per cm^3, if the energy were cohered and condensed into mass.

As can be seen, even a VT gain of 1.5×10^6 represents a "vacuum tap" of an incredibly small efficiency, on the order of 10^{-100} or so. However, the vacuum "river" is so energetic that such efficiencies are quite sufficient.

We accent that the barium ferrite magnetic material must be *activated* so that stable self-oscillation between the barium nucleus and the surrounding semiconductor vacuum exists. Although self-oscillating/self-pumped PCMs are known at optical frequencies, Sweet has discovered and perfected a brilliant methodology for activating PPCM nuclei at ELF frequencies.

In a resistive load such as light bulbs, the resistive material accomplishes repetitive phase conjugation. Thus in the resistor, half the total energy is expressed as photon or dissipative energy in the external (electron shell) level.

As the excited electrons decay, they emit scattered EM energy as light and heat. This is an *exothermic* interaction. The other half of the total energy reacts in the atomic nuclei, as a phase conjugative or *endothermic* interaction.

We strongly accent that, Whittaker-wise, there are two electromagnetic channels and two kinds of EM:

1. *external* EM, the common electron-shell interacting, entropic, scattering, time-forward kind, and
2. the *internal*, unsuspected, hidden variable, nuclei interacting, negentropic, reordering/ convergent kind.

Internal EM travels strictly between atomic nuclei, normally not reacting with electron shells unless a pumped phase conjugate mirror reaction is invoked in the nucleus to produce a gain somewhat greater than unity.

Thus if we wish to communicate with atomic nuclei directly, and engineer them directly, we must utilize the internal EM channel via applied Whittaker methods.

Antigravity Tests

Inherent in the preceding discussions is the possibility to turn EM energy into gravitational energy of either sign. In other words, one should be able to utilize Sweet's vacuum triode to produce and demonstrate antigravity.

Indeed this is the case. Sweet has also discovered the special alterations necessary to perform straightforward transformation of the internal energy in the nucleus to antigravitational energy, producing a unilateral thrust upward.

Note that the bulk of G-potential gradient (G-force-field) occurs Whittaker-wise at ELF frequencies. This explains why nonlinear phase conjugate opticians do not notice direct antigravity effects.

At the optical frequencies at which they work, the effects are so miniscule that they are negligible. This is readily explained as follows: In QM, the quantum (photon) is comprised of action (angular momentum), not just energy. It is rather like a "piece of energy welded to a piece of time, with no seam in the middle."

Since quantum change occurs in quanta, the decoupling of the energy and time components, in the continual interaction of photons with matter, exchanges energy between G-potential of vacuum and trapped mass of the atom or particle. In this exchange, small increments of time are continually being formed (and unformed, as photon emission occurs).

Consequently, each mass is moving forward in time in small incremental jumps, usually of exceedingly small magnitude. However, the energy and time trapped in a photon are canonical. The greater the piece of energy, the smaller the piece of time, and vice versa.

So if one wishes to stress the "rate of flow of time" significantly, one needs to produce large amounts of photons that have very large pieces of time, and consequently little pieces of energy.

Since the energy of the photon is directly proportional to its frequency, this means that the lower frequency photons have larger time increments, and hence endure over many "regular-sized photon absorption/emission changes" to appreciably stress the rate of time flow/production.

The bottom line is that the standard pumped phase conjugate mirror can be adapted to produce antigravity at ELF frequencies, but precisely the same adaptation at optical frequencies will have negligible effect.

With this in mind, the author requested Sweet to perform an antigravity experiment to prove the thesis. With Sweet's proprietary adaptation of his vacuum triode/PPCM, the experiment produced rather straightforward but spectacular results, as shown in Figure 5.

The experiment was performed as follows: Rigged for antigravity, the 6-lb. device was placed on a scale so that its weight could be continuously monitored. A special external load box was utilized in which multiple electric light sockets were connected in parallel.

Then the external load draw was adjusted by merely screwing in 100-watt lamps, one at a time, with measurement and observation pauses in between. The output of the device was 120 volt, negative AC sine-wave power at 60 Hz.

For each 100-watt increment, the load power was recorded and the weight was carefully recorded. The results are shown in the rather smooth, classic curve shown in Figure 5.

At 1,000 watts load draw, the previously 6-lb. device had reduced its weight due to gravity by 90 percent. At that point the signal-wave (grid) input to the open-loop vacuum triode was only 175.4 microamps at 10 volts, or just under two milliwatts.

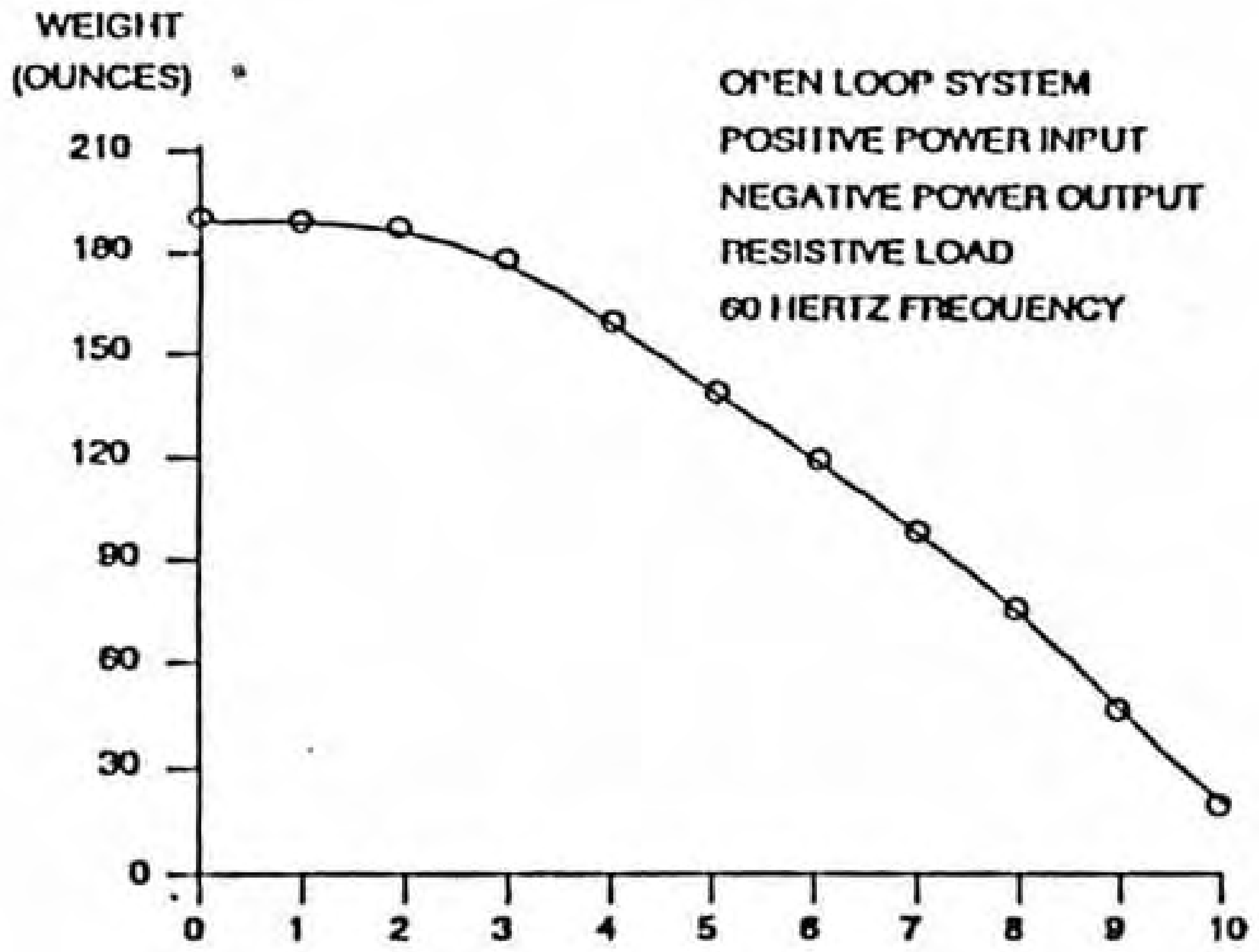

EXTERNAL ELECTRICAL LOAD

(HUNDREDS OF WATTS)

Figure 5. Antigravity Test of Sweet's Adapted Vacuum Triode.

We accent that the nominal two milliwatt input is only a gating signal. It is the organized, gated vacuum energy that is performing the action.

The experiment was stopped short of actual hovering and flying due to safety considerations. With the specific adaptation, magnetic monopoles are deposited in the magnet materials, producing internal tensile stress.

Since no explosive controlled facilities were available, and there was no wish to destroy the VT, the experiment was terminated at 90% antigravity performance.

It was completely successful, and adequately demonstrated the validity of the unified field theory concepts utilized in our approach.

CONCLUSIONS

A new unified field theory has been developed, tested, and at least partially verified experimentally.

The concepts of the theory have been applied by Sweet in a series of inventions that produce readily usable, safe electromagnetic power directly from the vacuum.

The methodology lends itself to formulation of power devices without moving parts. Antigravity, predicted by the concepts of the theory, has been demonstrated in actual practical demonstrations on the laboratory bench.

Though not discussed in this paper, application of the concepts and methodology to a large variety of other fields, such as medical reversal of aging and curing of almost the entire range of present debilitating diseases, has been previously pointed out. [Ref. 23]

We have also pointed out the mechanism for Kaznacheyev's cytopathogenic effect, or the induction of cellular pathology at a distance by electromagnetic means. [Ref. 24]

We have also pointed out the specific mechanism involved in Priore's device, which in rigorous laboratory animal testing under the auspices of eminent French scientists, demonstrated nearly 100 percent cures for terminal cancers and leukemias, sleeping sickness, artherosclerosis, and other debilitating diseases.[Ref. 25]

We believe we have also produced the concepts enabling the direct engineering and therapeutic manipulation of Popp's master cellular control system. [Ref. 26]

We conclude that the concepts we have utilized and experimentally demonstrated are universal, as implied by any notional unified field theory.

Our conclusion is that the concepts, theory, and experiments, taken together, are sufficient for investigation and replication by the scientific community.

If replicated and fully substantiated, we believe the work will directly point the way to, and usher in, a new unified field theory physics of universal application.

SPECULATED IMPLICATIONS

As can be seen, the implications of the new approach are profound. The authors believe they have ushered in the forerunner of a vast new physics, one which will change our lives, and our view of physical reality, in ways previously undreamed of.

By mastering, controlling, and gating the vast, incredible energy of the seething vacuum, we can power our automobiles, flying machines, and technology inexhaustibly. Further, it can be done absolutely cleanly; there are no noxious chemical pollutants to poison the biosphere. With practical antigravity, ships can be developed to cross the solar system as readily as one crosses the ocean today.

And the ships, automobiles, and technology will never run out of fuel; the inexhaustible vacuum fills every system, everywhere, to overflowing.

Not discussed in this paper, it turns out that living systems, faced with the problem of achieving negentropy so as to maintain their form in a dissipative external physical reality, have always used the hidden internal channel for such things as mind, thought, cell control, and living functions.

With the new methodology, one now faces the advent of access and engineering of the mind and life of the observer as readily as the observer's physical body.

Transmutation of the elements, control of the weather, lighting and powering our cities and homes cheaply and cleanly, and provision of plenty for everyone is the vista for the future. We can in fact clean up the radioactive wastes, rid ourselves of coarse nuclear and petroleum powerplants.

We strongly stress that, with the ability to engineer the Schroedinger equation itself, the new methodology allows the direct engineering and control of quantum change, and hence of physical reality itself.

The methodology is extendable to hyperdimensions; nested virtual levels of the vacuum are already precisely that. The author has already pointed out the application of this emerging technology to the absolute cure of diseases such as AIDS, cancer, leukemia, etc., and shown that the Priore device in France already proved the efficacy of the application in the 60s and 70s.

We shall be able to rid ourselves and our descendents of diseases. With direct access to the actual software of life and mind, in the future we should be able to achieve levels of education previously unattainable, by directly inputting the relevant software.

Previously we have also pointed out that four nations of the world are already embarked on weaponization of scalar EM unified field technology. It is sobering to think that, in addition to having the ability to make our planet a paradise for humankind, we also will have the ability to make it a hades.

For that reason, we are doing our best to clarify the technical concept and the theory in this 1991, hopefully with the view that humankind will seize upon the positive aspects, and develop and apply this technology for the betterment of all people everywhere.

Long ago, Albert Einstein said these words:

"It would of course be a great step forward if we succeeded in combining the gravitational field and the electromagnetic field into a single structure. Only so could the era in theoretical physics inaugurated by Faraday and Clerk Maxwell be brought to a satisfactory close."

And Teilhard de Chardin wrote:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have

discovered fire."

The authors fervently believe they have come upon fire for the second time, as allegorized by de Chardin. If so, let us all use the knowledge wisely.

REFERENCES

1. Maxwell's original quaternion EM theory is contained in some 200 quaternion equations and differs extensively from the restricted Heaviside/Gibbs vector interpretation universally taught today as "Maxwell's Theory." See James Clerk Maxwell, *A Treatise on Electricity and Magnetism*, Oxford University Press, Oxford, 1873. The third edition is by Dover, 1954.

2. E.T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.*, Vol. 57, 1903, p. 333-355; "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc. Lond. Math. Soc.*, Series 2, Vol. 1, 1904, pp. 367-372. The first paper was in fact a detailed theory of the scalar standing potential wave that Nikola Tesla discovered four years earlier, on the night of July 3-4 1899, being radiated from thunderstorms, which he entered in his *Colorado Springs Notebook* on the morning of the 4th. More recently the Whittaker structure (WS) inside potentials, including the Schroedinger potential, has been shown by V.K. Ignatovich, "The Remarkable Capabilities of Recursive Relations," *Am. J. Phys.*, 57(10), Oct. 1989, p. 873-878, without credit to Whittaker or to the presence of such structured scalar potentials in Maxwell's original quaternion EM theory. These WSs are universal to scalar potentials; e.g., for the same in acoustics, see Richard W. Ziolkowski, "Localized Transmission of Wave Energy," *Proc. SPIE, Vol. 1061, Microwave and Particle Beam Sources and Directed Energy Concepts*, Jan. 1989, p. 396-397.

3. See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," *Phys. Rev.*, Second Series, 115(3), Aug. 1, 1959, p. 458-491. This paper pointed out the primacy of the potentials. Instead of being causative agents, the force fields are actually effects generated from the potentials. This is in complete violation of both classical electromagnetics and classical dynamics, but it is absolutely required by quantum mechanics. For an extensive discussion of the Aharonov-Bohm effect and an extensive list of references, see S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," *Rev. Mod. Phys*. 57(2), Apr. 1985. See Bertram Schwarzschild, "Currents in Normal-Metal Rings Exhibit Aharonov-Bohm Effect," Physics Today, 39(1), Jan. 1986, p. 17-20 for confirmation that the Aharonov/Bohm effect has indeed been firmly proven experimentally.

4. It is pointed out that today all potentials are well-known to be gravitational entities. However, this was not known in Whittaker's time, and so he himself did not realize that he had actually produced an engineerable, testable unified theory of electromagnetics and gravitation.

5. T.E. Bearden, "Possible Whittaker Unification of Electromagnetics, General Relativity, and Quantum Mechanics: Part I: Background," Presented to *Ala. Acad. Sci. Annual Symp.*, Univ. Jacksonville, Mar. 1991.

6. For the theory of a pumped phase conjugate mirror, see David M. Pepper, "Nonlinear Optical Phase Conjugation," *Opt. Eng.*, 21(2), Mar./Apr. 1982, p. 156-183; Amnon Yariv, *Optical Electronics*, 3rd Edn., Holt, Rinehart, and Winston, New York, 1985. In a normal triode tube, the amplified plate signal is 180 degrees out of phase *spatially* with the grid input, but in-phase with it in respect to *rate of flow through time*. In a PPCM, the amplified phase conjugate replica wave is in phase spatially with the signal wave, but 180 degrees out of phase with it in respect to its rate of flow through time. It is pointed out that a PCM with a gain of unity produces a coupled EM wave/antiwave pair whose energy is additive in 3-space but subtractive in the first derivative of the fourth dimension, time. Hence it is no longer an electromagnetic wave as such, but an oscillatory wave of stress upon the local rate of flow of time. Hence it is a powerful *electrogravitational wave*, whose EM nature is hidden in the guise of a scalar EM potential. It is also a wave of variation in the local gauge, and of variation in the local ST curvature.

7. See T.E. Bearden, *The Phase Conjugate Vacuum Triode*, Apr. 23, 1987, privately published. An earlier pencil draft was initially produced.

8. Sweet has continued to extend his theoretical treatment. A formal paper providing the complete mathematical theory of the vacuum triode has been drafted, is in final review, and will be submitted to a leading journal by mid-1991.

9. See Chapter 25: Outlook, "Possibility of Vacuum Engineering," T.D. Lee, *Particle Physics and Introduction to Field Theory*, Harwood, New York, 1981, pp. 824-828.

10. In 1977 Ilya Prigogine received the Nobel Prize for extending thermodynamics; in particular, for the theory of dissipative structures in nonequilibrium thermodynamics. In Prigogine systems, negentropy is known to be possible.

11. Robert G. Sachs, *The Physics of Time Reversal*, Univ. Chi. Press., Chicago, 1987.

12. Note that this moves the entire notion of the charge of a fundamental particle to a deeper and more extended level. Now the charge may be *discretized*, but it is not quantized in the hard conventional sense. Further, the internal Whittaker structure of the massless VPF photon exchange of vacuum and mass, which, quantum mechanically, is what the electrical charge of the particle is in the first place, is deterministically structured. Note that this violates the present assumption that all like charged particles are identical; now two electrons may have either the same or different magnitudes of charge, and even when the magnitudes are the same, their internal charge structures (Whittaker structures) and VPF exchange with the vacuum may differ. Also note that this resolves the severe QM problem of missing chaos (hidden order) in quantum change. The reason for the problem was the use of Gibbs statistics with its assumption of random variable change, which a priori excluded hidden order (and hence chaos) from QM. That was only a special case, albeit an important one. There are now three QM cases: (1) the conventional case, where there is no hidden order; (2) the case where there is some hidden order, and the statistics is chaotic, not random; and (3) the case where the QM change is deterministic, with essentially total hidden order. Note that the Whittaker

methodology allows one to directly engineer cases (2) and (3), including the Schroedinger equation itself.

13. The present second law of thermodynamics is written only for time-forward entities, and need not apply for the time-reversed case. Merely viewing the energy-dissipating forward time case in reverse allows an appreciation of the time-reversed case. In other words, the second law of thermodynamics is incomplete as presently stated. The complete law has a corollary to cover the increase in order as the time-reversal of the system increases. Thus the complete law consists of two parts: (1) the entropic, time-forward case, and (2) the negentropic time-reversed case. Since the re-ordering can be amplified at will by a PPCM process, the correct distinction between the two subsets of the complete law is important, and applies to real systems.

14. A.D. Sakharov, *Theor. Math. Phys*., Vol. 23, 1975, p. 435.

15. T. E. Bearden, *Gravitobiology: A New Biophysics*, Tesla Book Co., 604 Date Ave., Chula Vista, CA 91912, 1991.

16. C.f. Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler, *Gravitation*, W.H. Freeman and Co., San Francisco, 1973, p. 1191.

17. E.g., see Richard Kidd et al, "Evolution of the Modern Photon," *Am. J. Phys*., 57(1), Jan. 1989, pp. 27-35. See also R. Chen, "Cancellation of Internal Forces," *Am. J. Phys*. 49(4), Apr. 1981, p. 372.

18. A nonlinear material may simply emit a photon, or it may act as a phase conjugate mirror (PCM) and emit a phase conjugate replica of the absorbed photon. When the material emits a normal photon, it measurably recoils. When it emits a time-reversed photon, it does not recoil, as already experimentally established in nonlinear phase conjugate optics. The solution to the mystery is this: When emitting a normal photon, the material does not act as a PCM. In that case the matching antiphoton which split from the interacting graviton (the graviton that yielded the external photon) interacts with the nucleus, producing a recoil action with a gain of one. Thus Newtonian third-law recoil of the nucleus occurs. On the other hand, when the material acts as a PCM, it also emits the antiphoton *outside* the atom to "backtrack" the absorbed "signal wave" photon. In that case there is no Newtonian recoil of the nucleus, because the *agent* for causing recoil did not interact with the nucleus to produce it.

19. Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys*. 52(1), Jan. 1984, p. 19-21. Also, under nonlinear conditions, a particle can absorb more energy than is in the light incident on it, absorbing the energy from the vacuum VPF. C.f. Craig F. Bohren, "How Can a Particle Absorb More Than the Light Incident on It?" *Am. J. Phys*. 51(4), Apr. 1983, p. 323-327.

20. C.f. Pepper, ibid. and Yariv, ibid. For a specific example, see Mary J. Miller et al, *Appl. Phys. Lett*. 41(8), Oct. 15, 1982, p. 689-691.

21. Again, for the theory of the PPCM, see Pepper, ibid. and Yariv, ibid.

22. E.g., see T.E. Bearden, *Gravitobiology: A New Biophysics*, Tesla Book Co., Chula Vista, CA, 1991, p. 33-36.

23. Bearden, *AIDS: Biological Warfare*, Tesla Book Co., 1988; *Gravitobiology: A New Biophysics*, Tesla Book Co., 1991; *Analysis of Scalar Electromagnetics*, Tesla Book Co., 1990.

24. C.f. Vlail P. Kaznacheyev and L.P. Mikhailova, *Ultraweak Radiations in Intercellular Interactions*, [in Russian], Novosibirsk, 1981; Vlail P. Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research*, 1(1), Mar. 1982, p. 47-76; N.D. Devyatkov, Ed., *Applications of Low-Intensity Millimeter Wave Radiation in Biology and Medicine*, [in Russian], IRE Akad. Nauk. SSSR, Moscow, 1985.

25. C.f. Antoine Priore, "Method of Producing Radiations for Penetrating Living Cells," U.S. Patent No. 3,280,816; Jean-Michel Graille, *Le Dossier Priore*, De Noel, Paris, 1984 [in French]; Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Appx. I to Bearden, *AIDS: Biological Warfare*, 1988.

26. C.f. Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert Popp et al, Eds., *Electromagnetic Bio-Information: Proceedings of the Symposium*, Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979, p. 123-149; also *Biophotonen. Ein neuer weg zur Losung des Krebsproblems*, Verlag fur Medizin, Heidelberg, 1976 [in German].

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

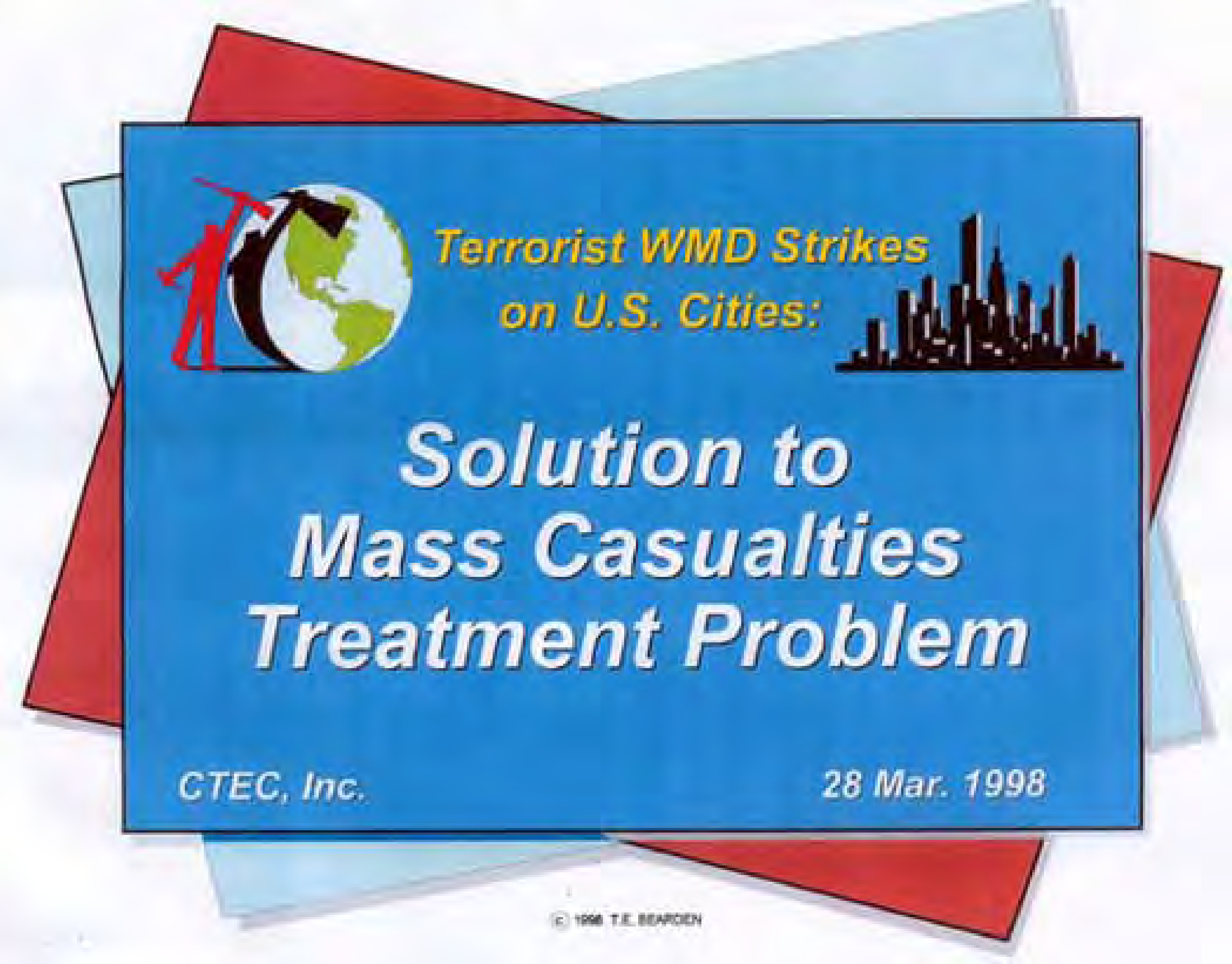

Terrorist WMD Strikes
on U.S. Cities:

Solution to Mass Casualties Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

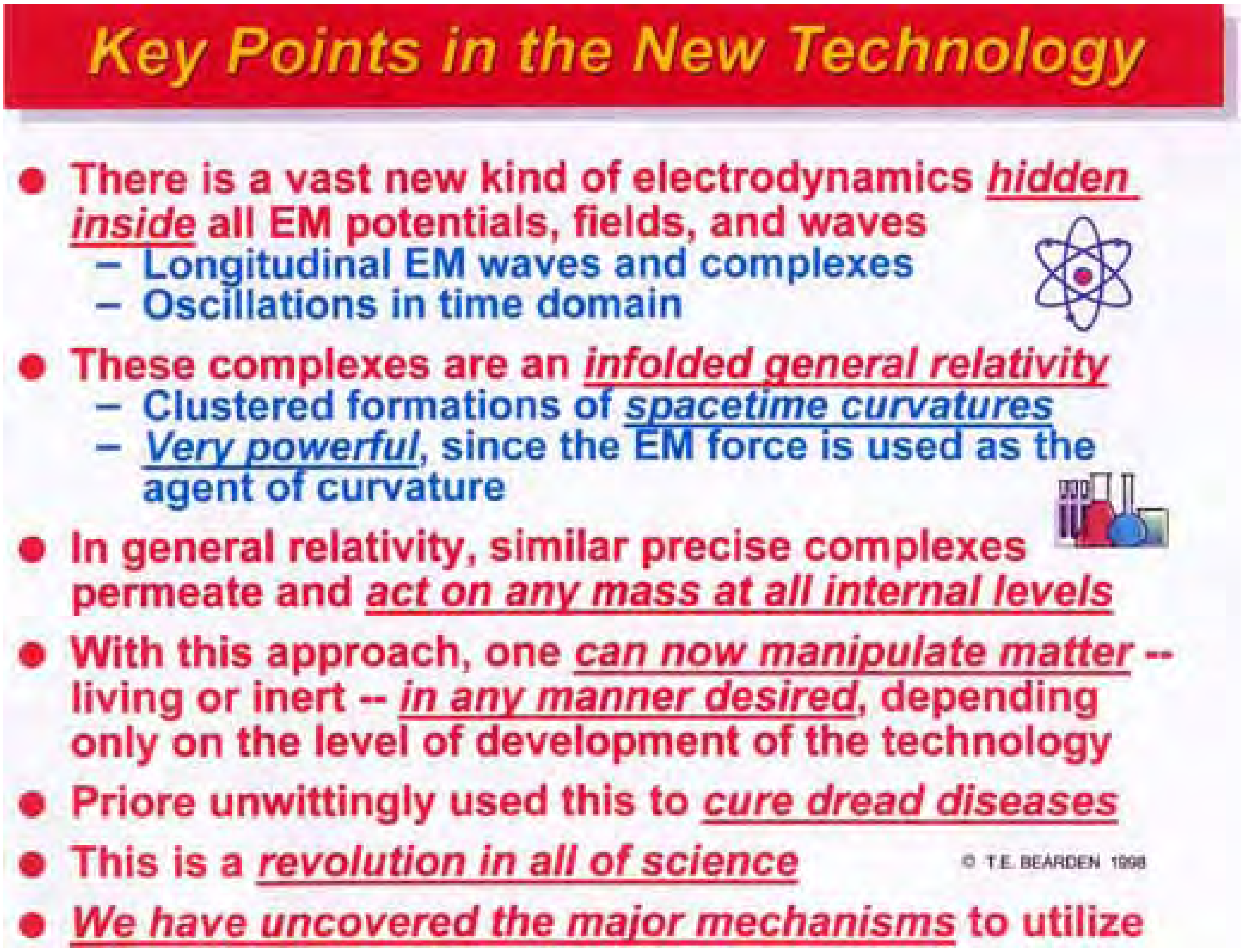

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

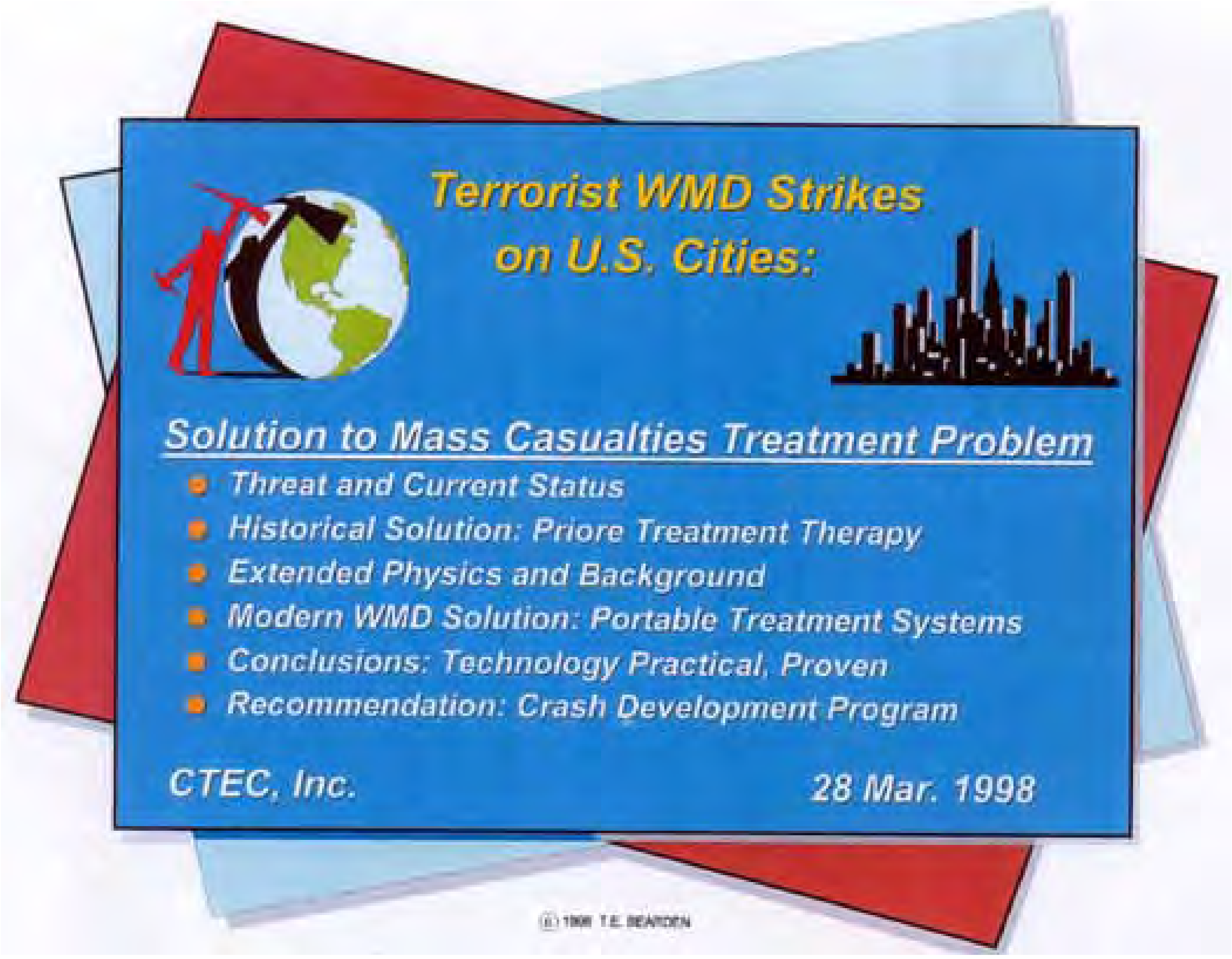

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

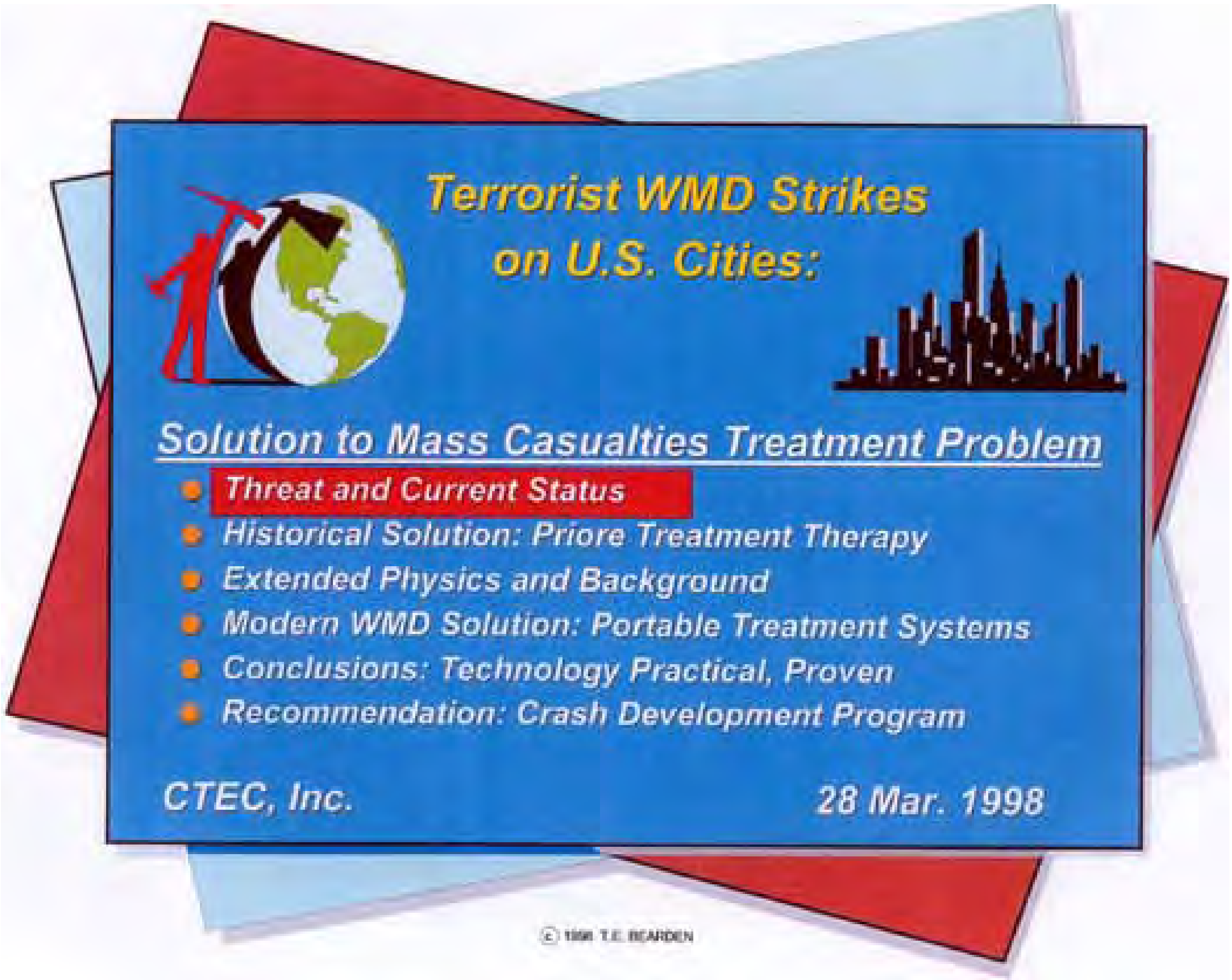

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

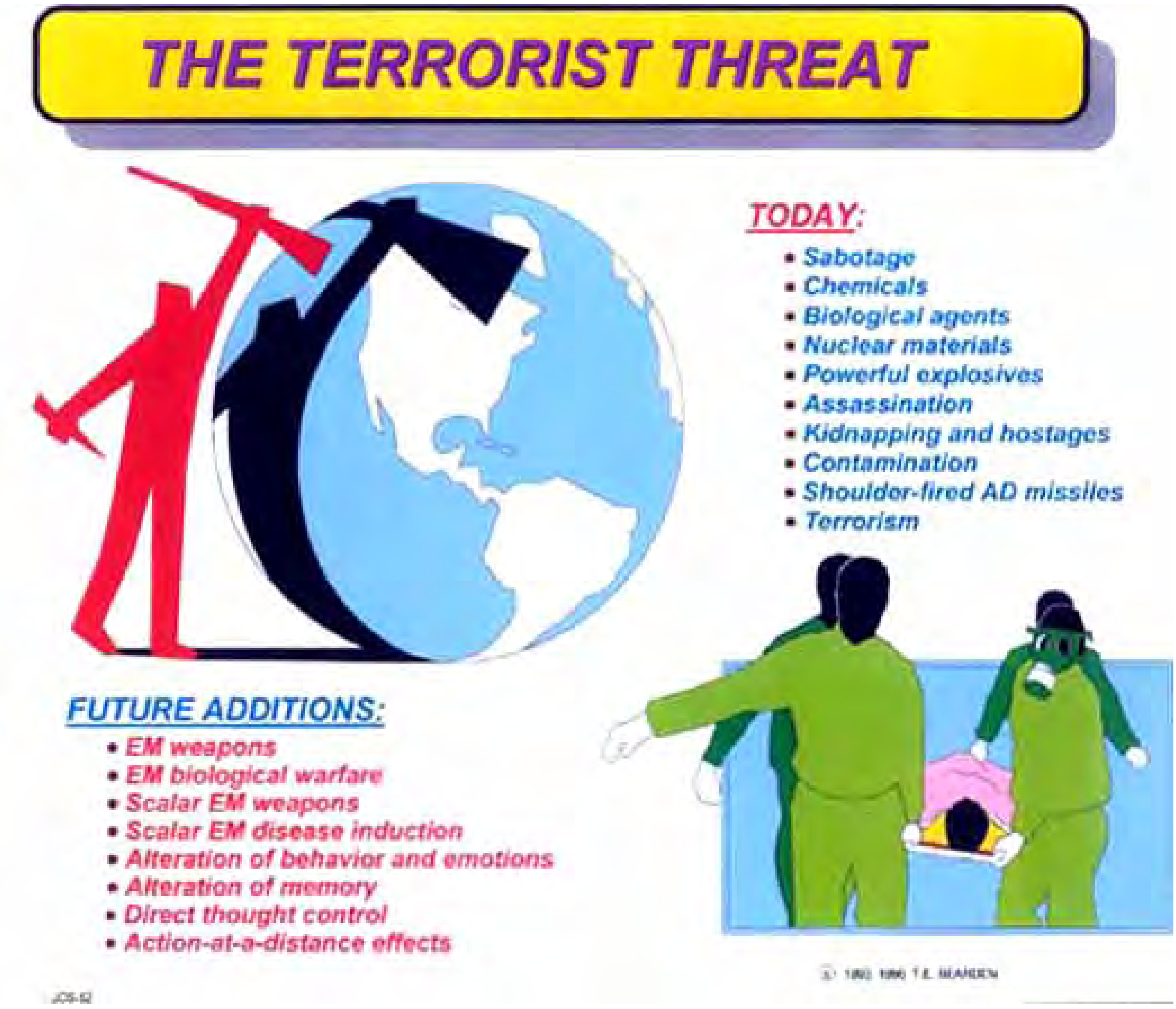

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

© 1992, 1996 T.E. BEARDEN

JCS-62

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

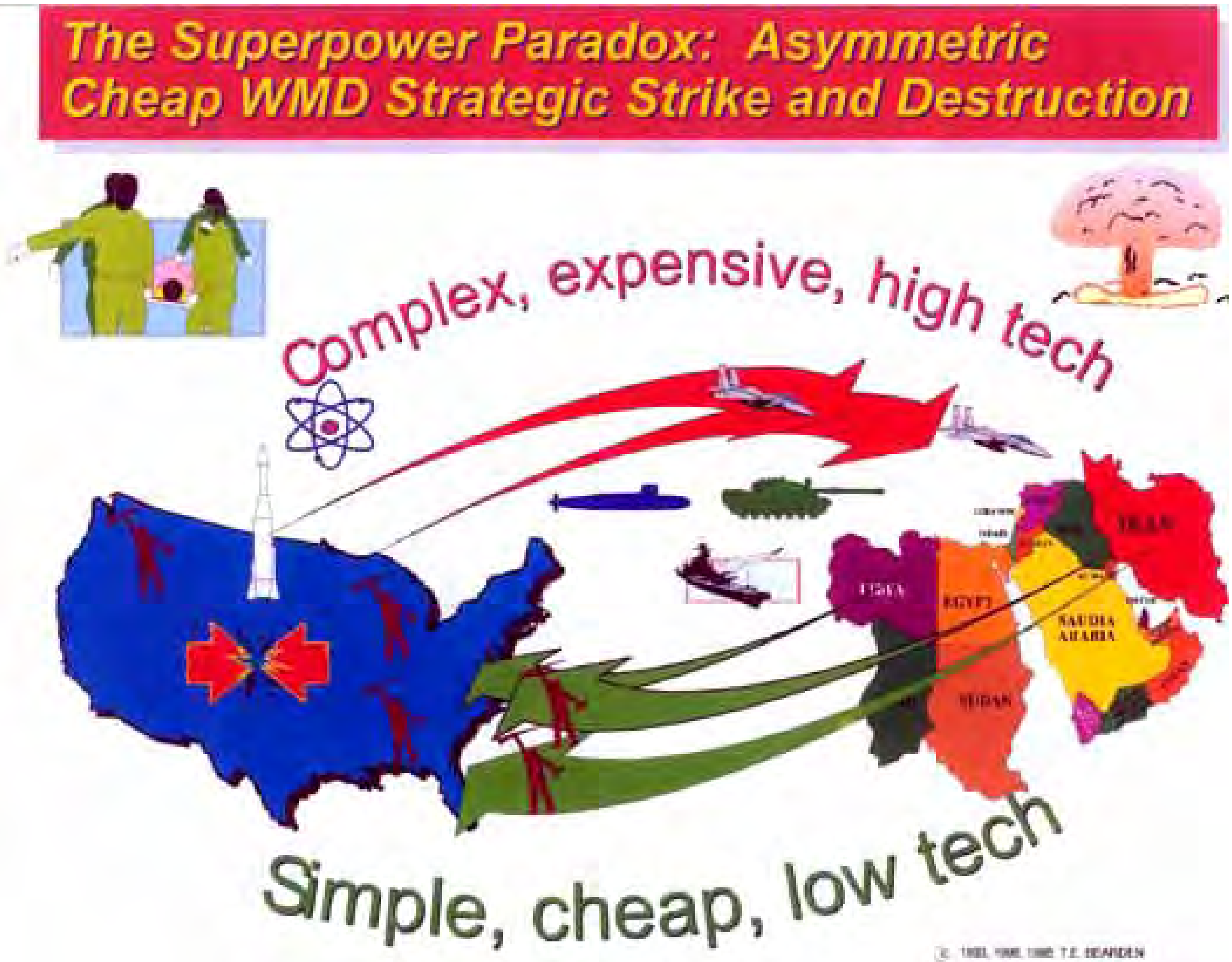

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

IRAN
LIBYA
EGYPT
SAUDIA ARABIA
SUDAN

© 1993, 1996, 1998 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

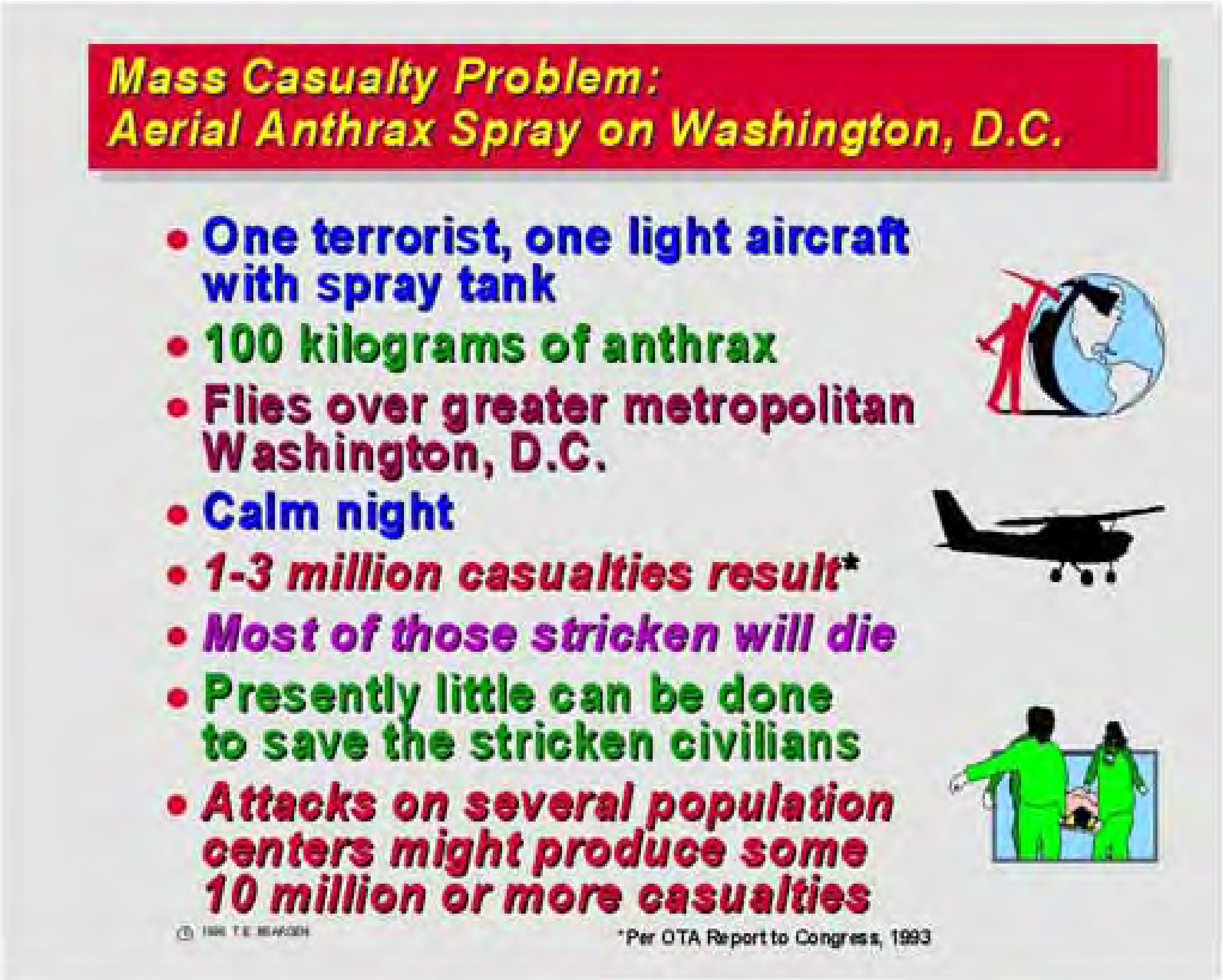

Mass Casualty Problem:
Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© 1996 T.E. BEARDEN

*Per OTA Report to Congress, 1993

[Return to Slide Index]
[Next Slide]
[Previous Slide]

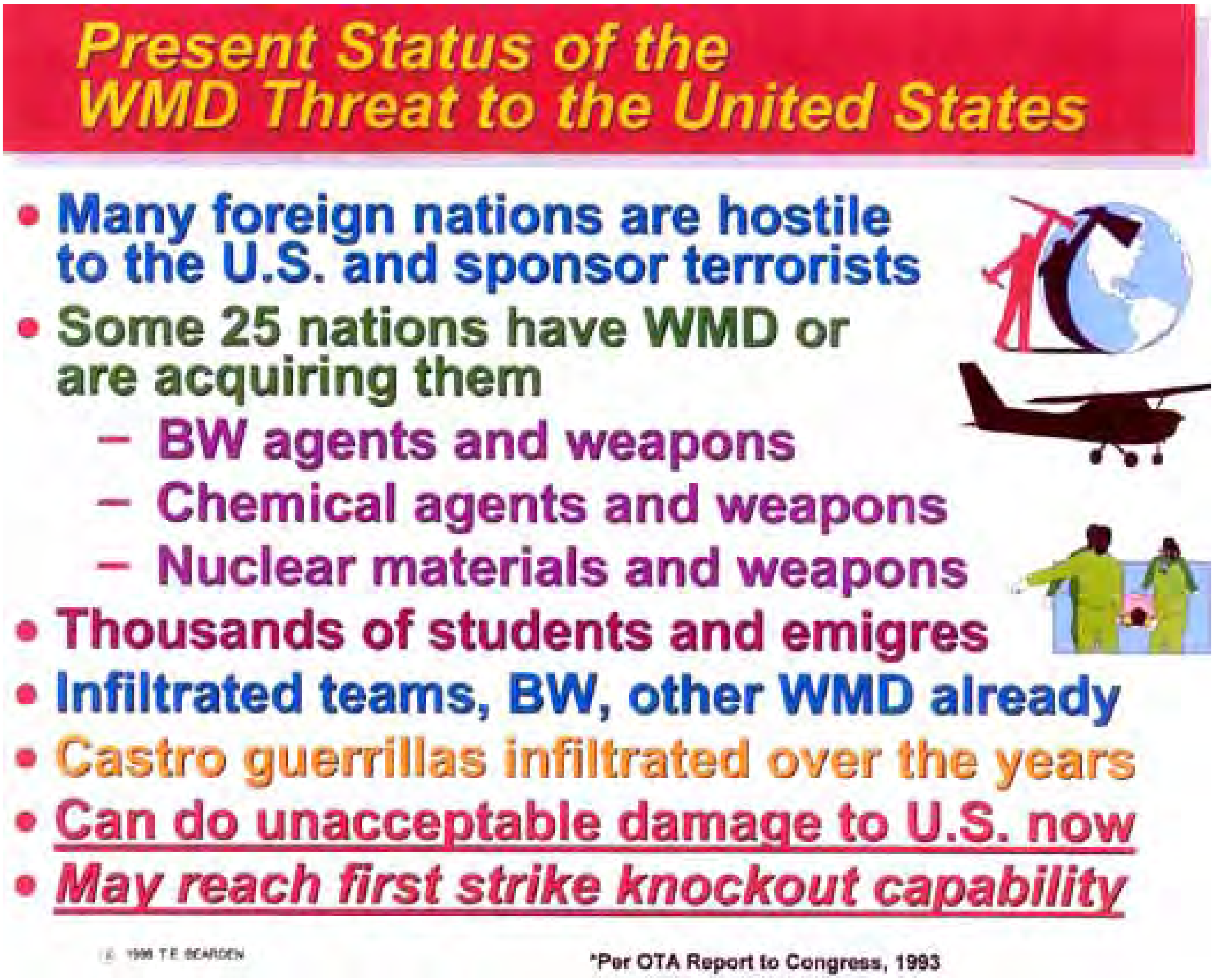

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

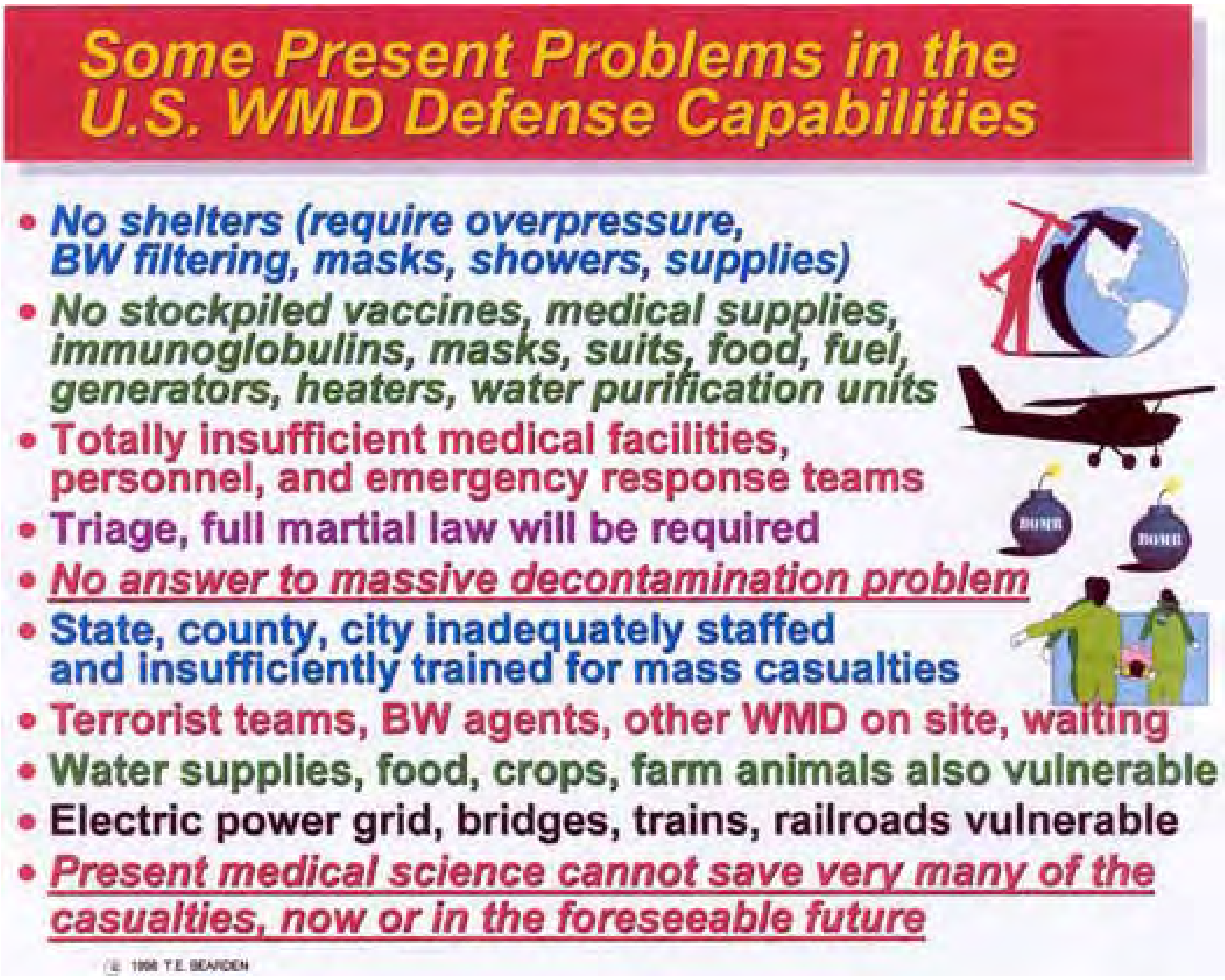

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

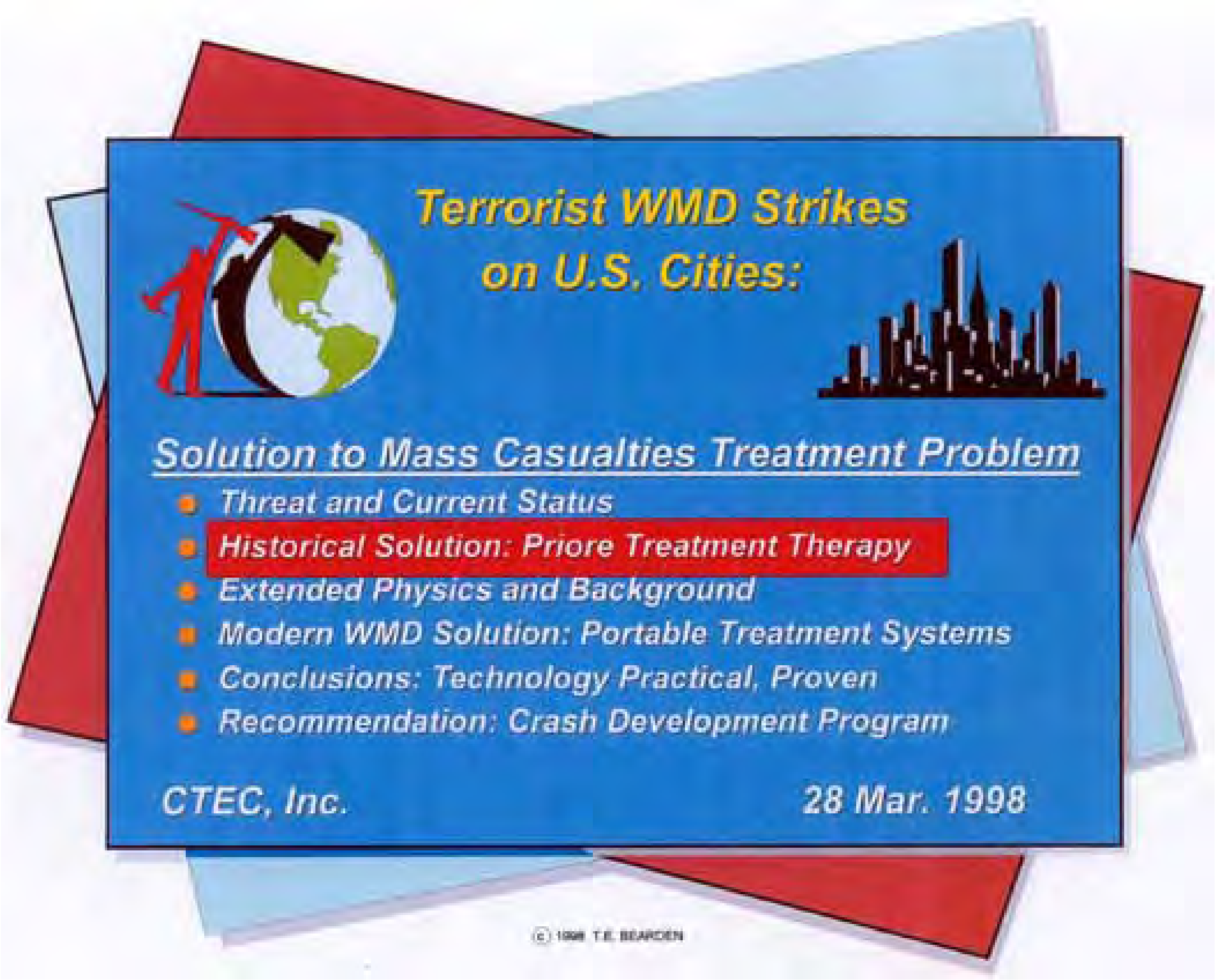

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

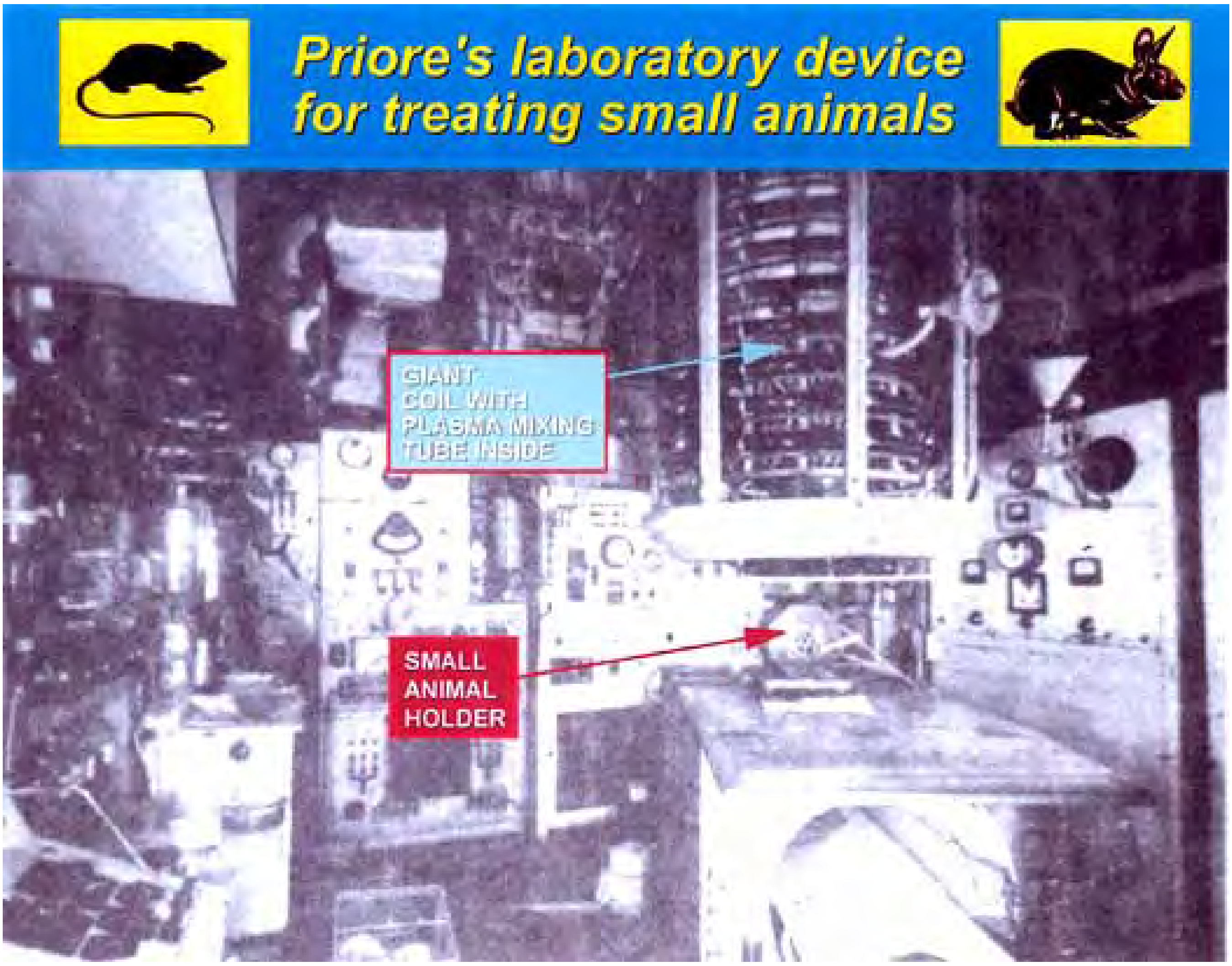

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE
SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

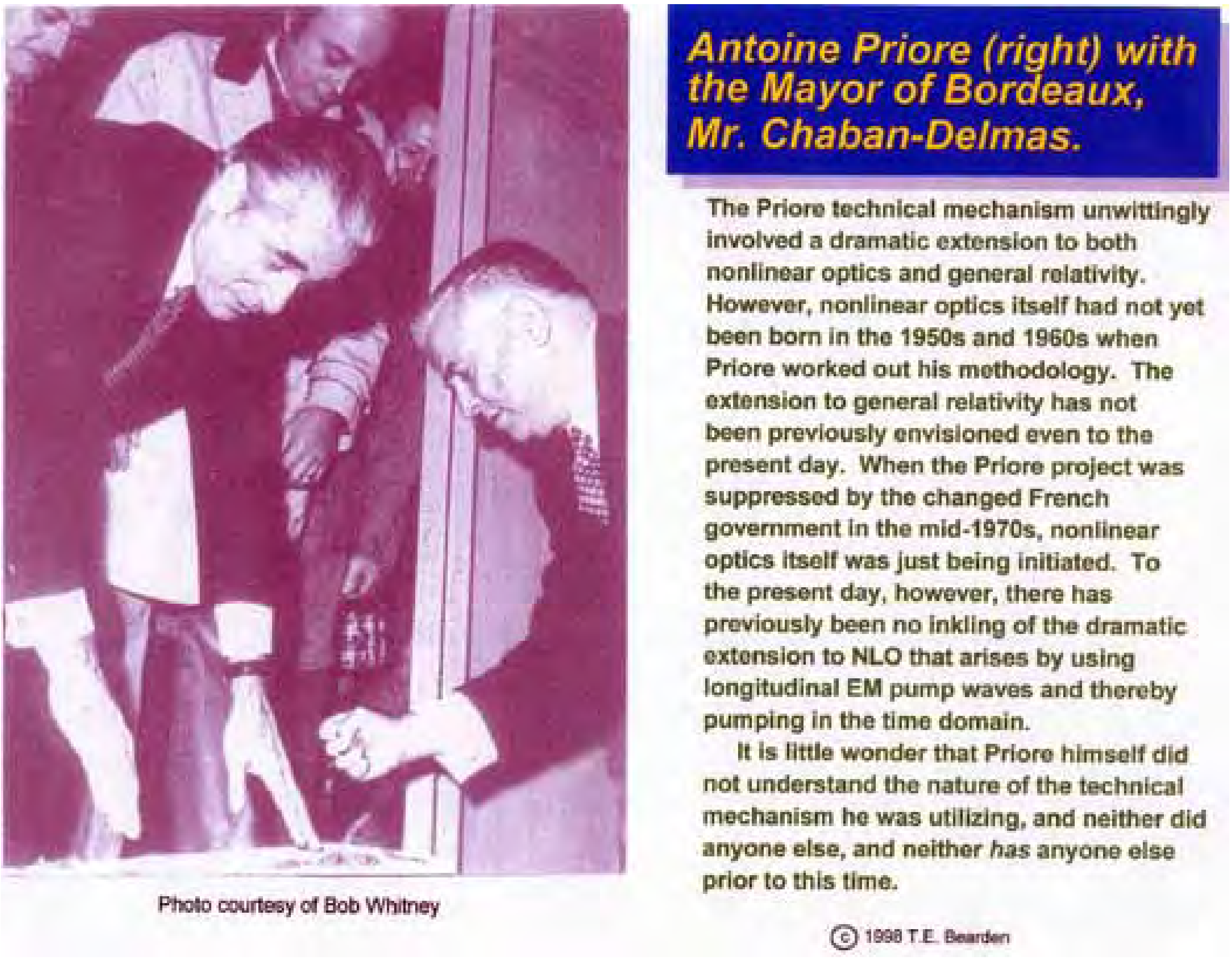

Photo courtesy of Bob Whitney

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

© 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

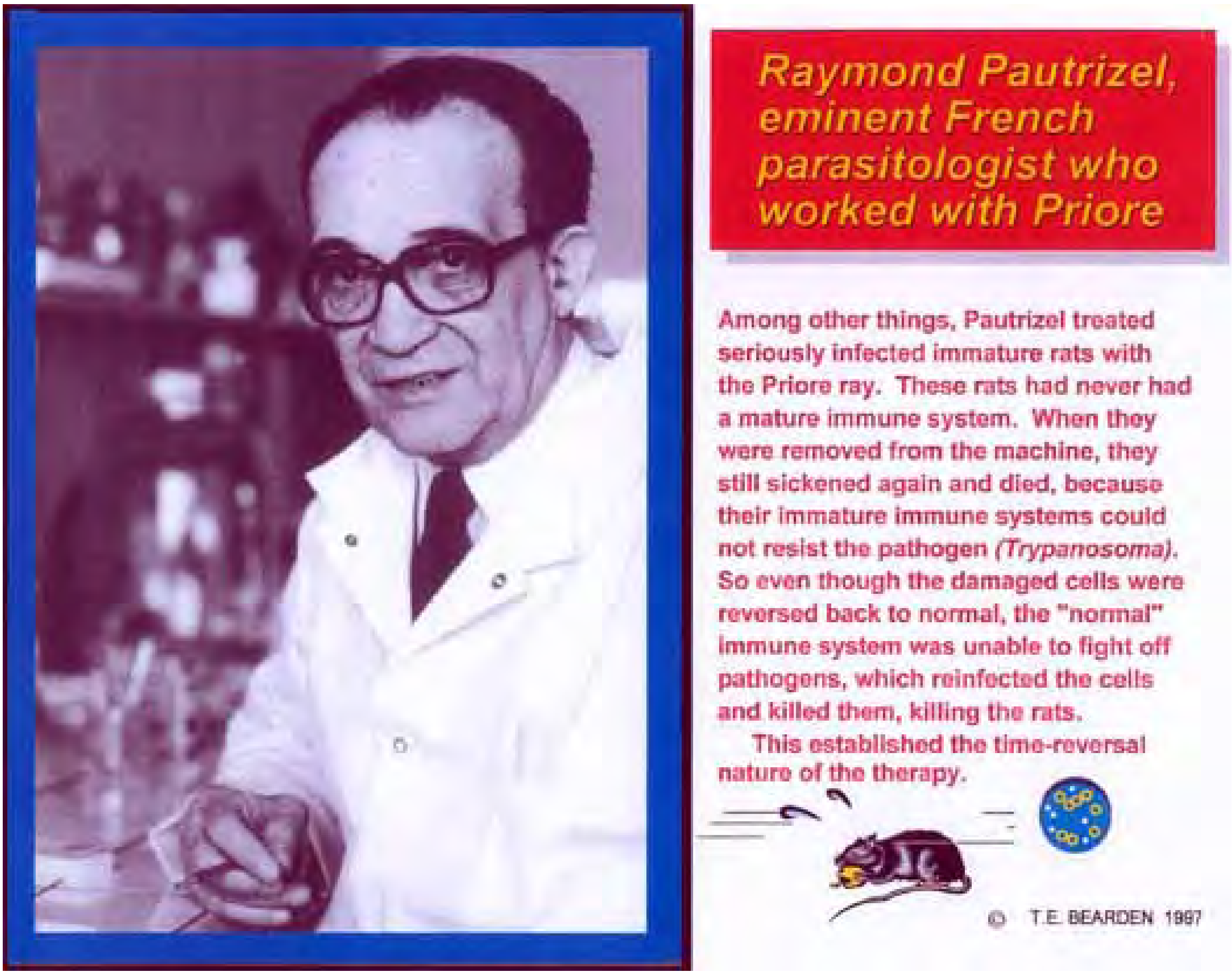

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

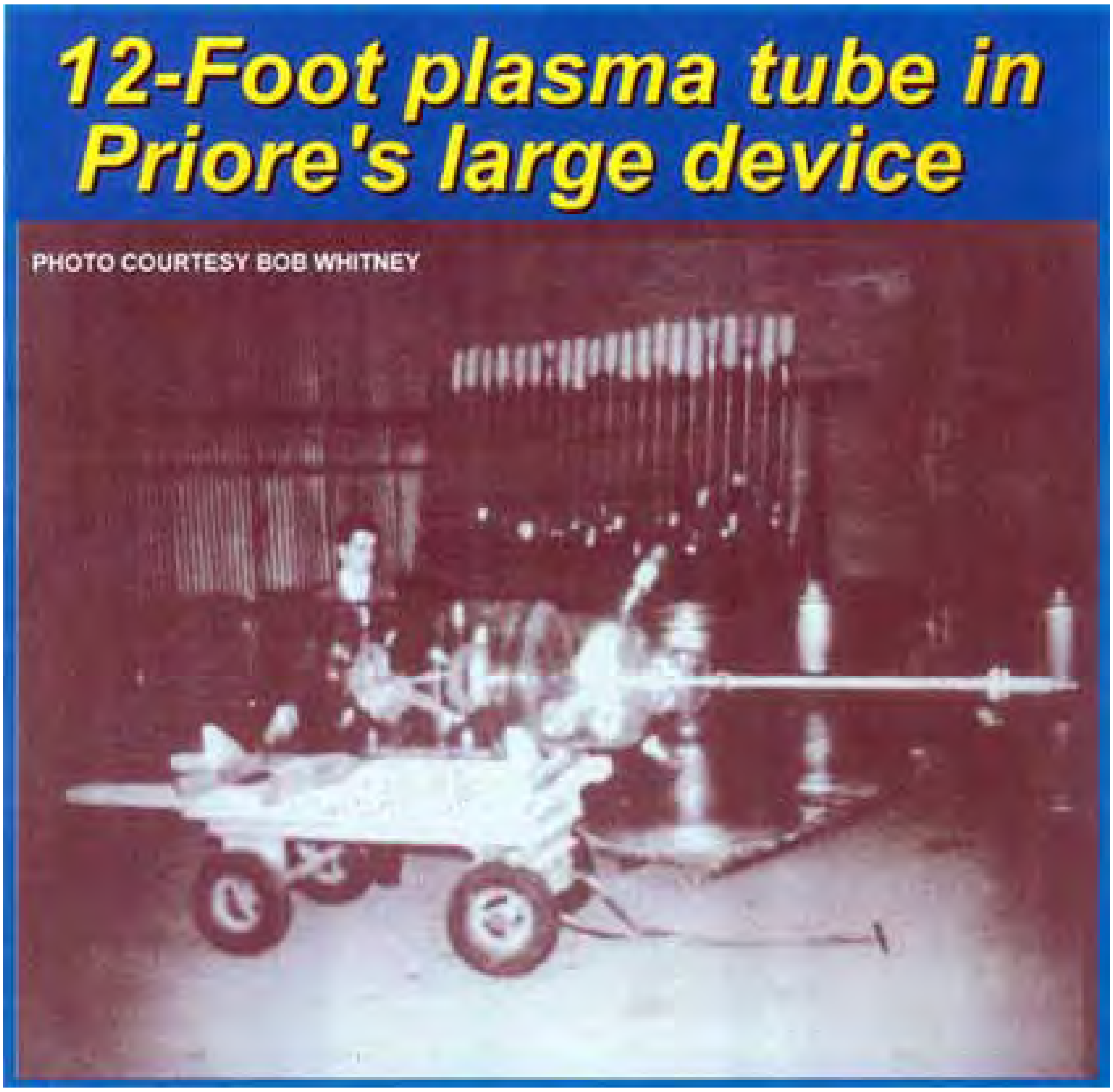

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

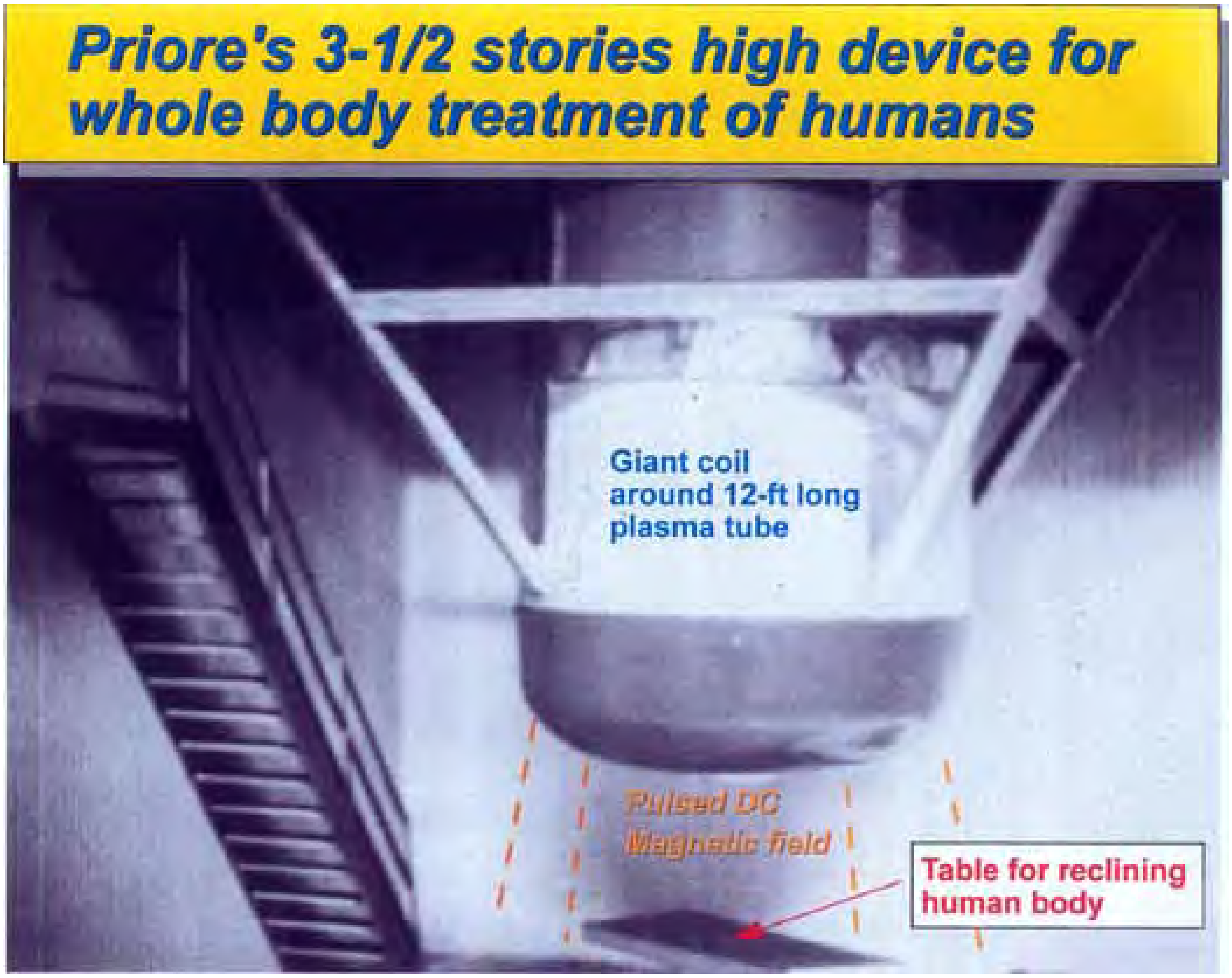

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube
Pulsed DC
Magnetic field
Table for reclining
human body

[Return to Slide Index]

[Next Slide]

[Previous Slide]

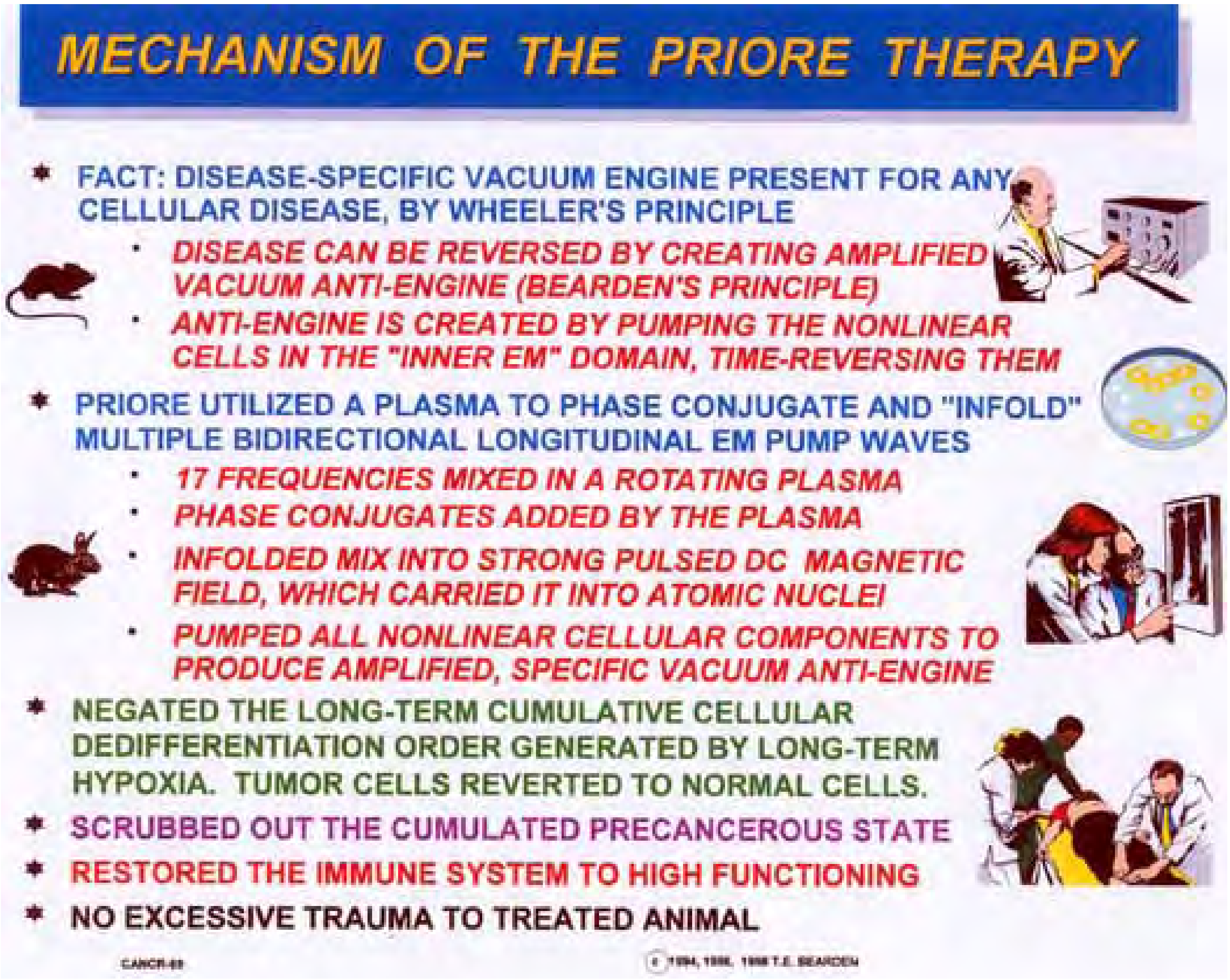

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

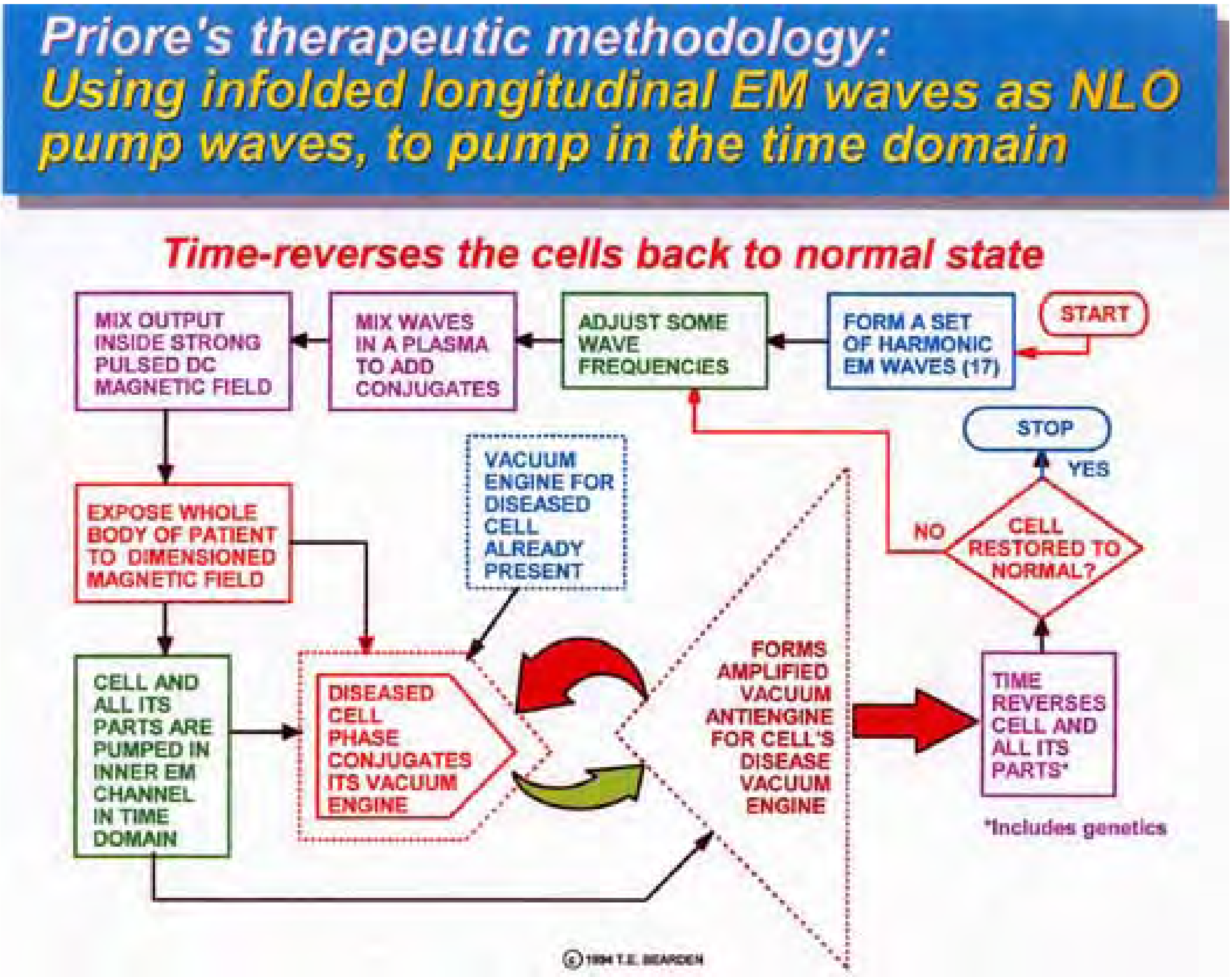

Priore's therapeutic methodology: Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

CELL RESTORED TO NORMAL?

NO

YES

STOP

*Includes genetics

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

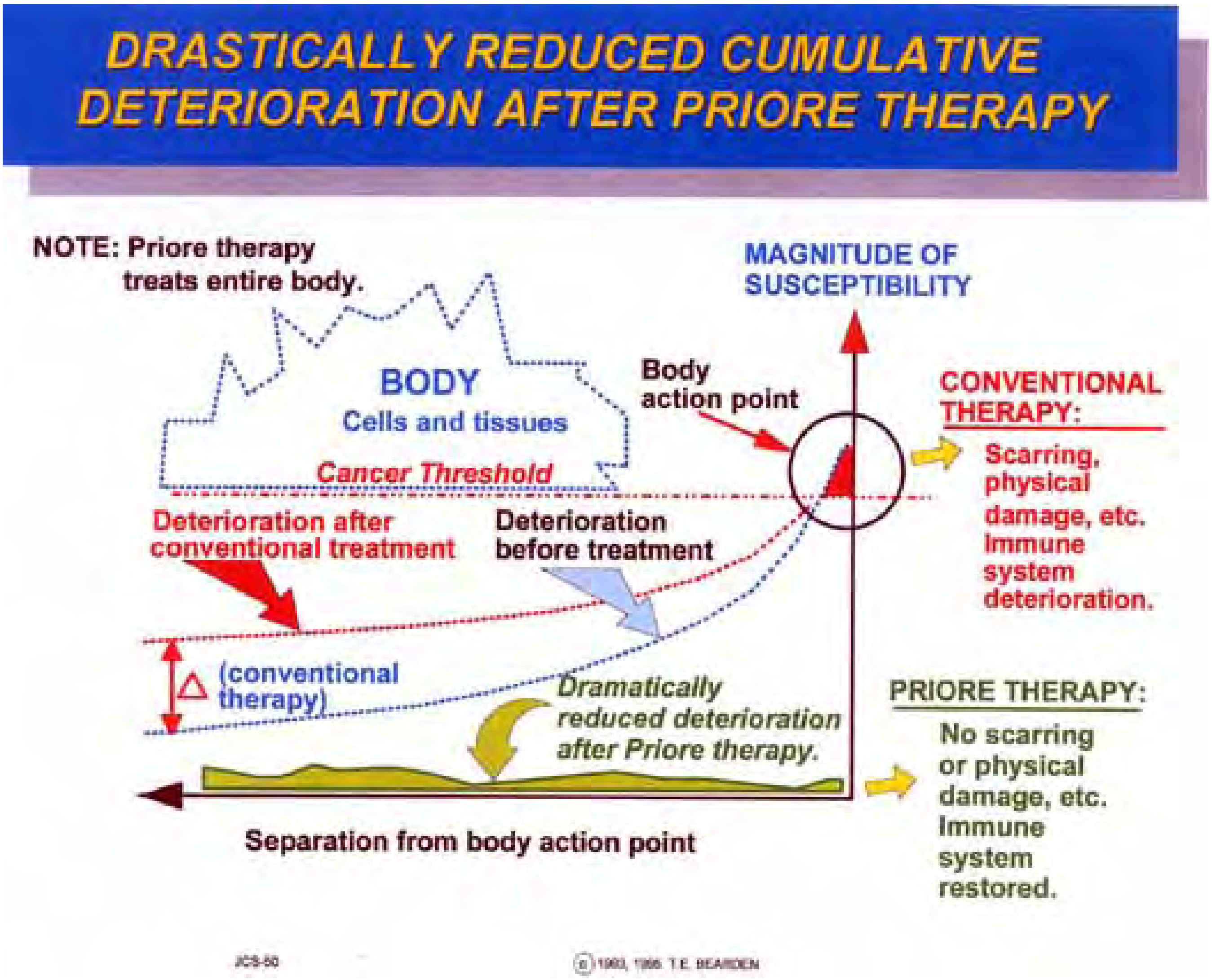

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

BODY
Cells and tissues
Cancer Threshold
Body action point
MAGNITUDE OF SUSCEPTIBILITY
Deterioration after conventional treatment
Deterioration before treatment
Δ (conventional therapy)
Dramatically reduced deterioration after Priore therapy.
Separation from body action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

JCS-50

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

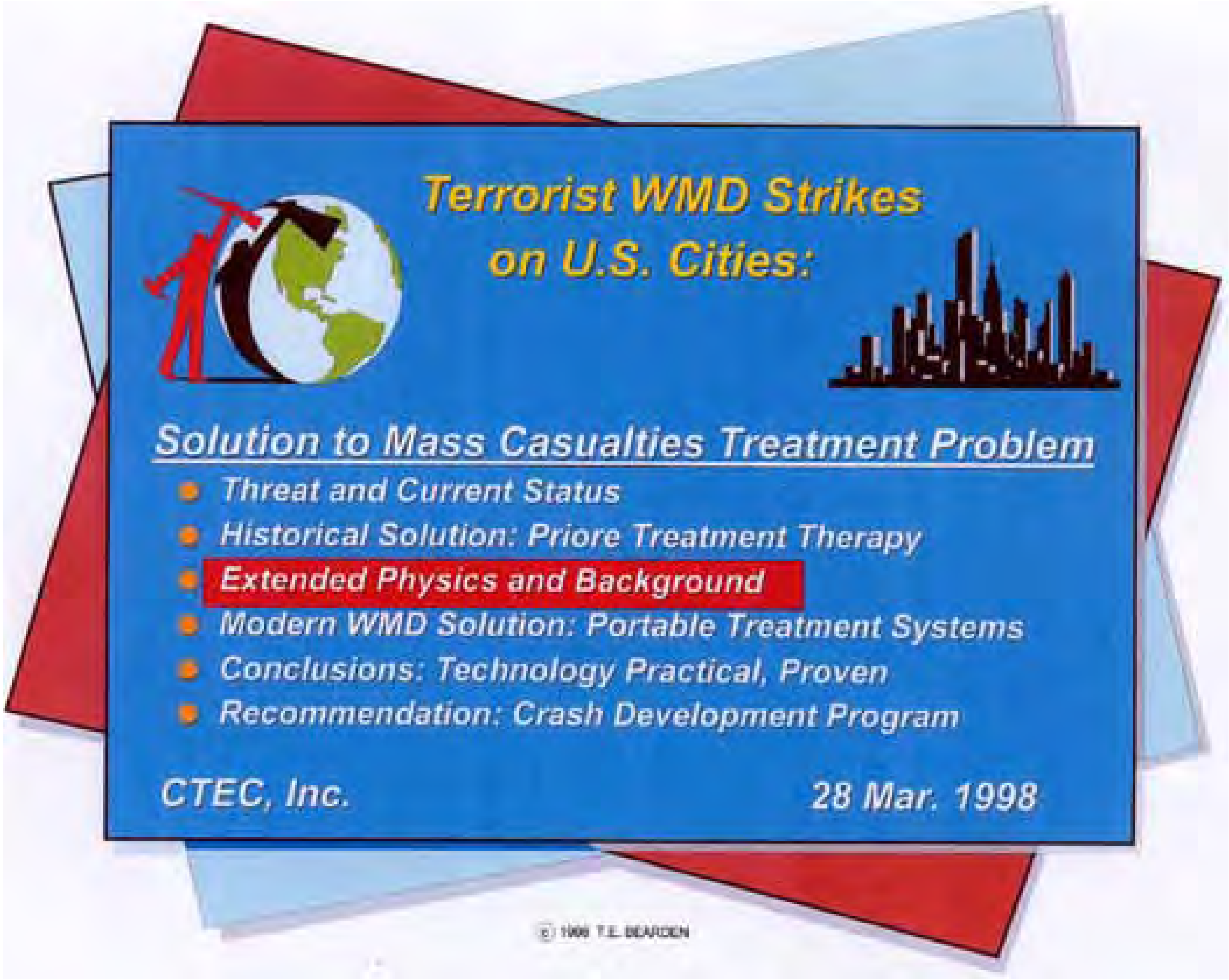

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

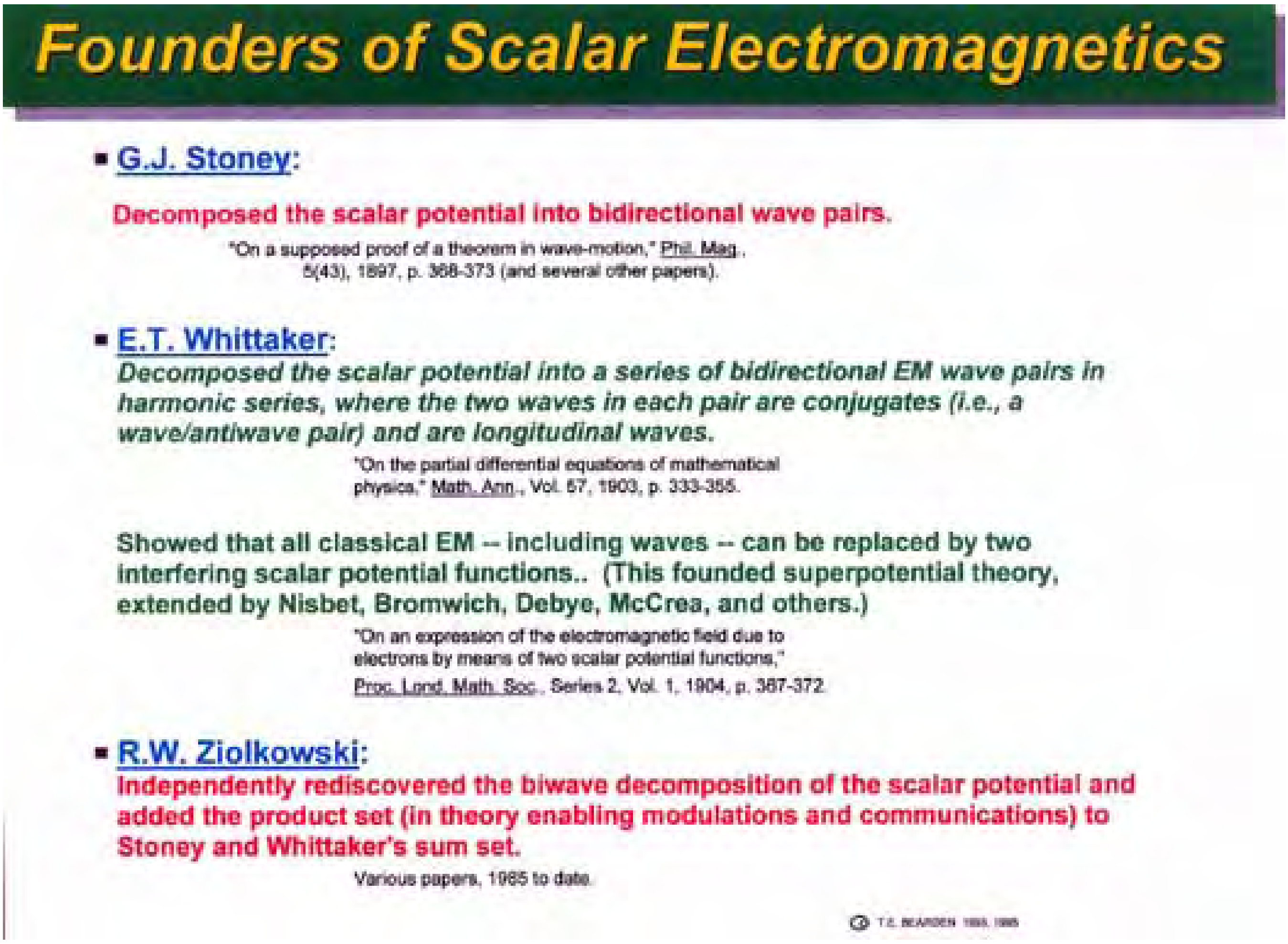

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

  Decomposed the scalar potential into bidirectional wave pairs.

  "On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

  Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

  "On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

  Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

  "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.

- **R.W. Ziolkowski:**

  Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

  Various papers, 1985 to date.

© T.E. BEARDEN 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

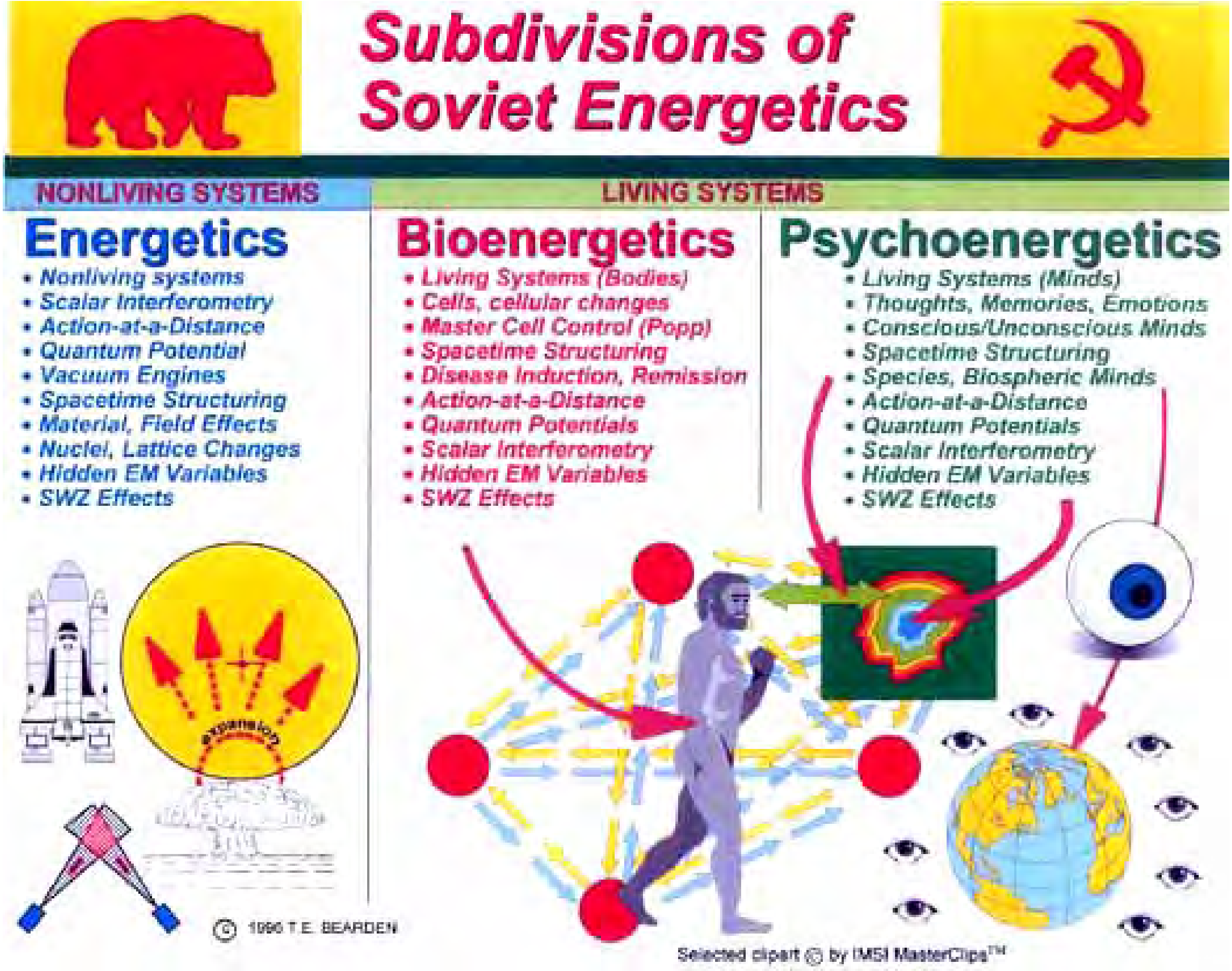

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS | **LIVING SYSTEMS**

Energetics
- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics
- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics
- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

[Return to Slide Index]

[Next Slide]

[Previous Slide]

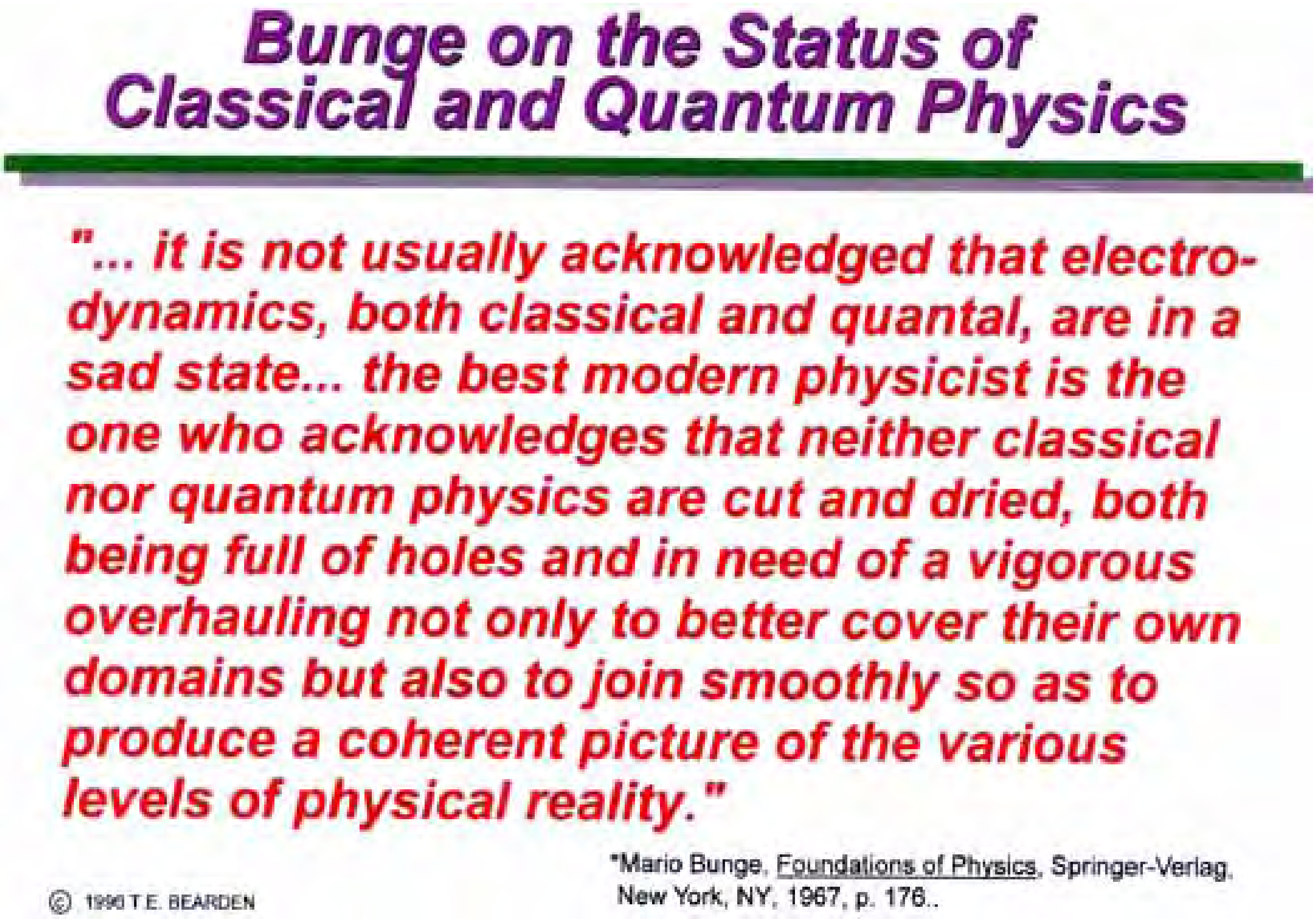

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

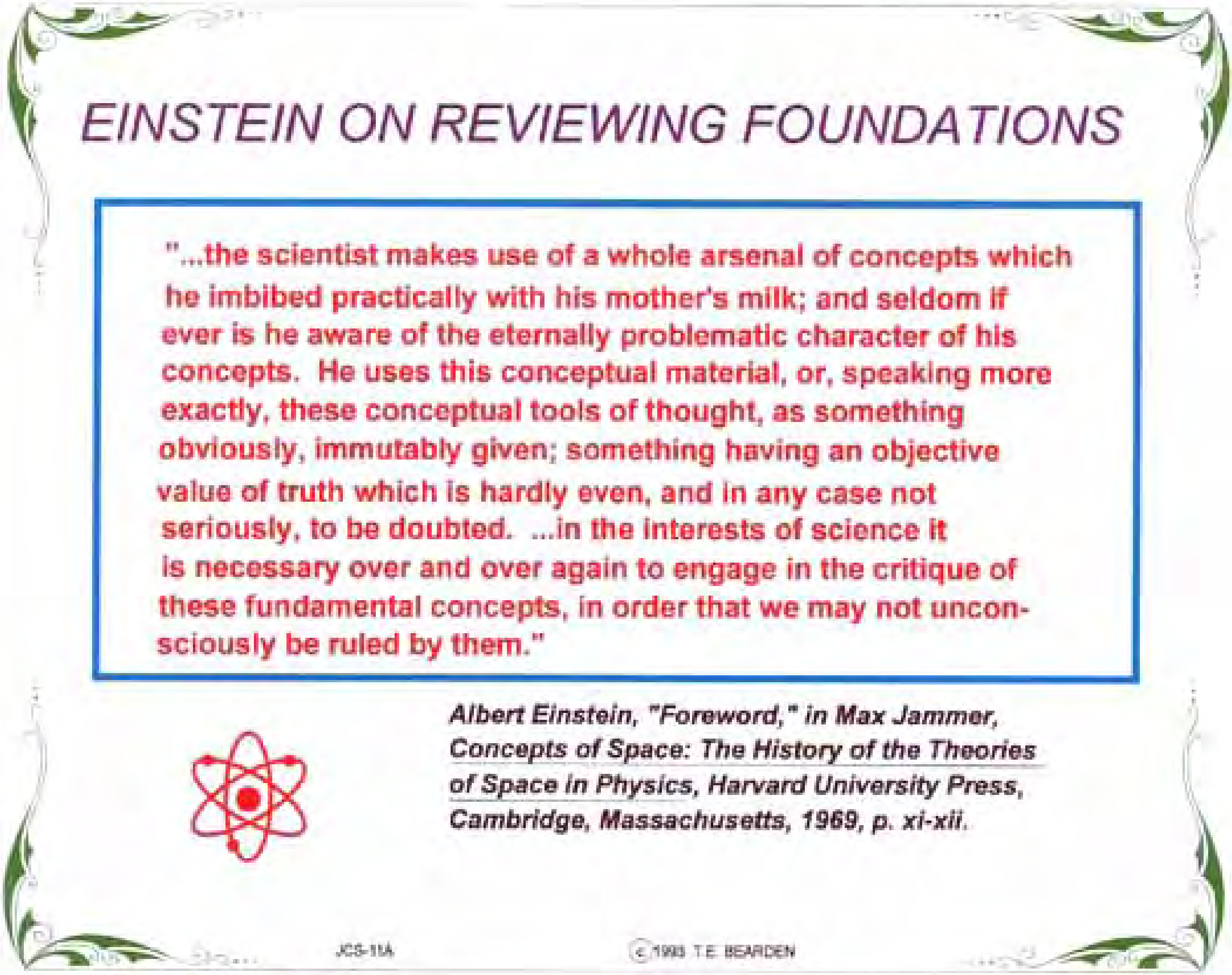

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

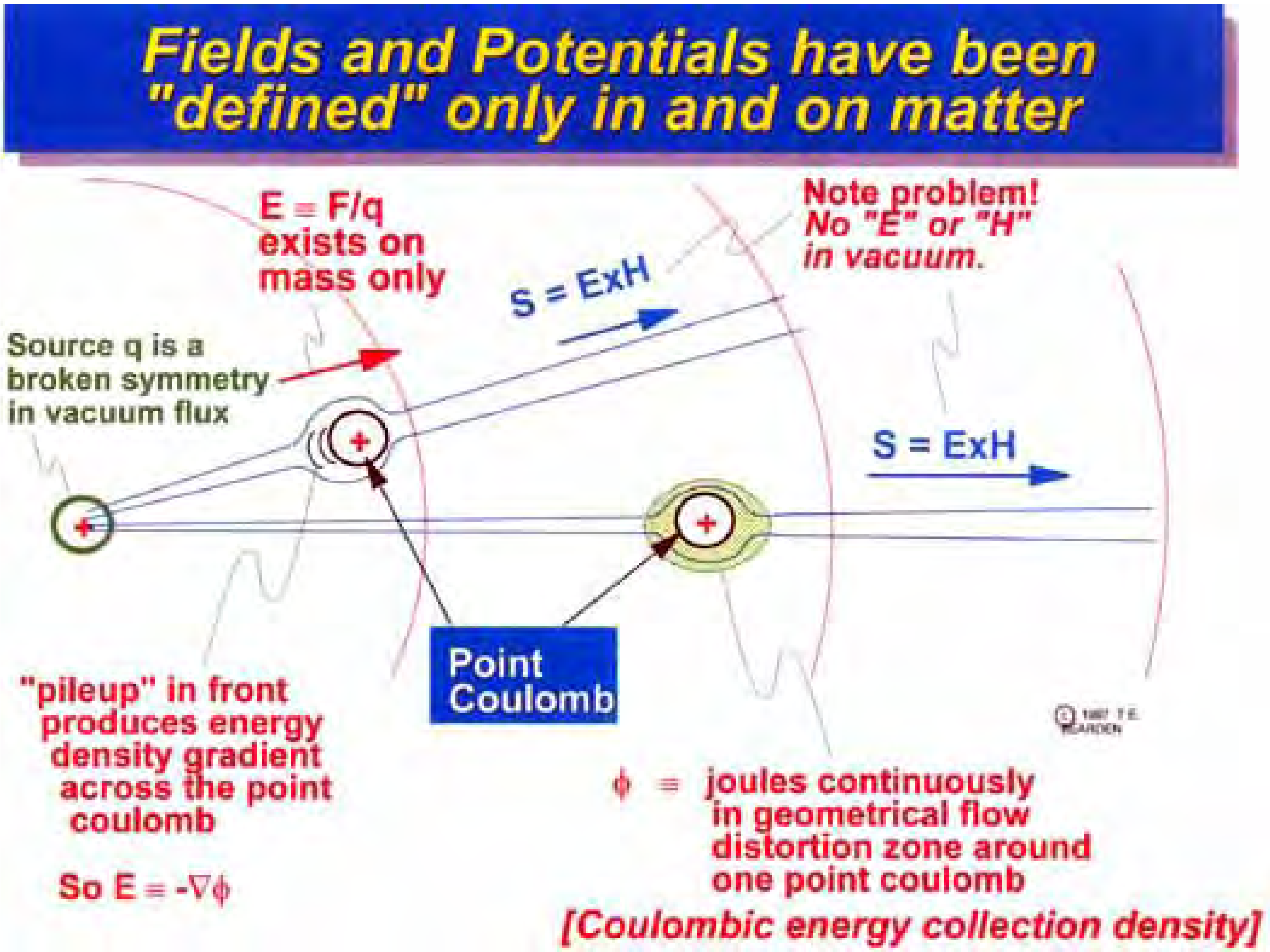

Fields and Potentials have been "defined" only in and on matter

$E \equiv F/q$
exists on
mass only

Note problem!
No "E" or "H"
in vacuum.

Source q is a
broken symmetry
in vacuum flux

$S = E \times H$

$S = E \times H$

Point
Coulomb

© 1997 T.E. BEARDEN

"pileup" in front
produces energy
density gradient
across the point
coulomb

So $E \equiv -\nabla\phi$

$\phi \equiv$ joules continuously
in geometrical flow
distortion zone around
one point coulomb

[Coulombic energy collection density]

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

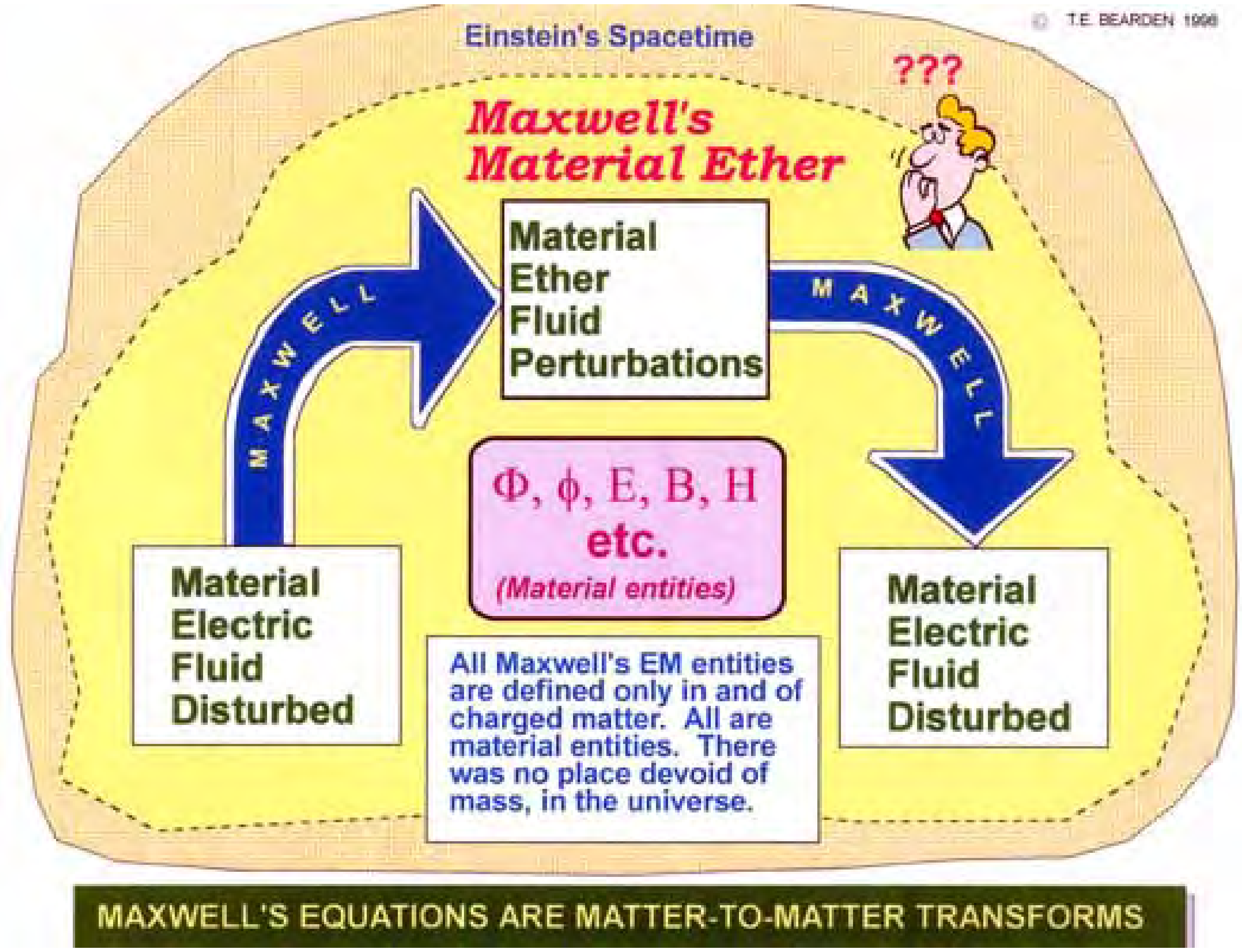

© T.E. BEARDEN 1998

Einstein's Spacetime

???

Maxwell's Material Ether

Material Ether Fluid Perturbations

Φ, ϕ, E, B, H etc.

(Material entities)

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

Return to Slide Index

Next Slide

Previous Slide

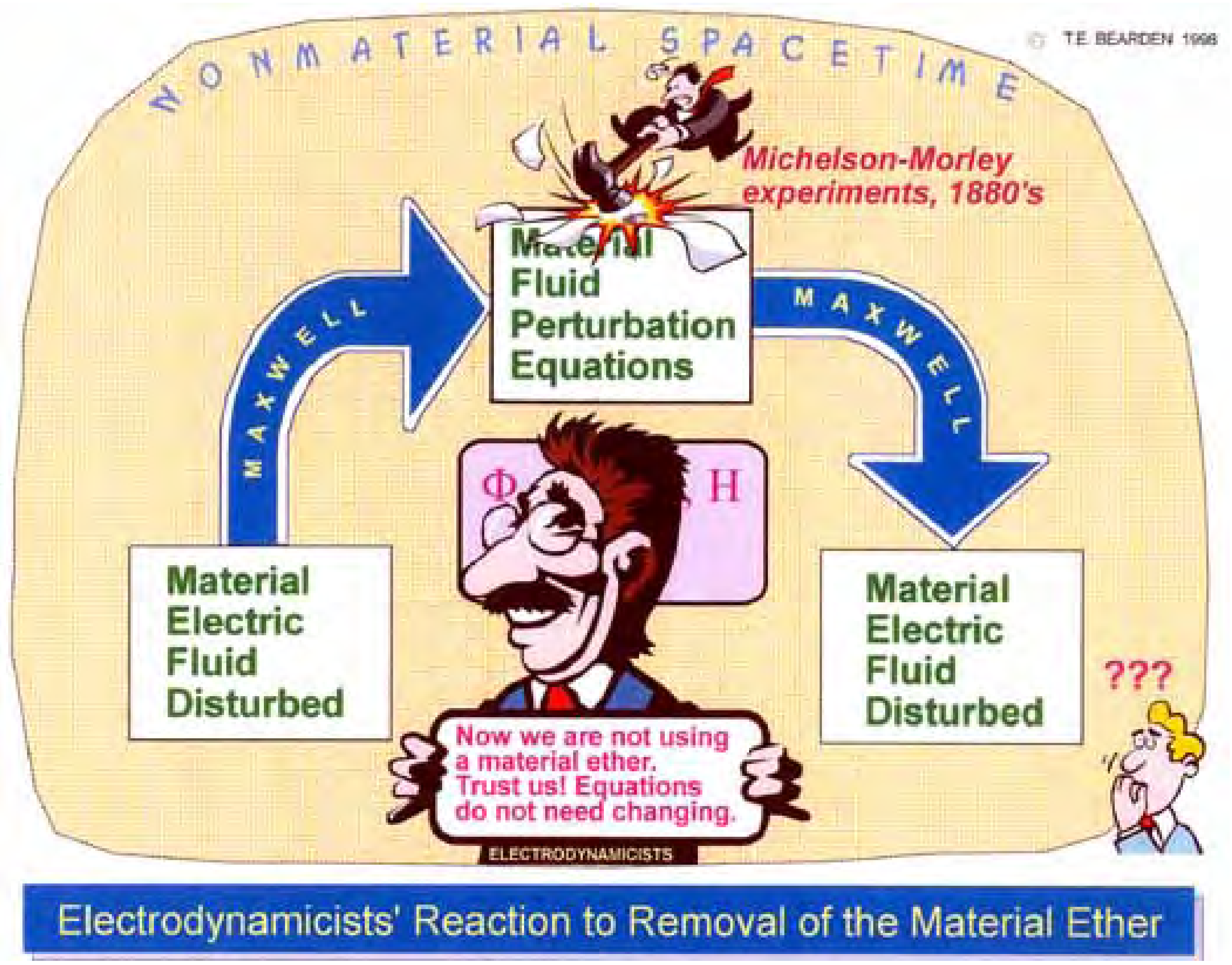

NONMATERIAL SPACETIME

© T.E. BEARDEN 1998

Michelson-Morley experiments, 1880's

MAXWELL

Material Fluid Perturbation Equations

MAXWELL

Material Electric Fluid Disturbed

Φ

H

Material Electric Fluid Disturbed

???

Now we are not using a material ether. Trust us! Equations do not need changing.

ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

Return to Slide Index

Next Slide

Previous Slide

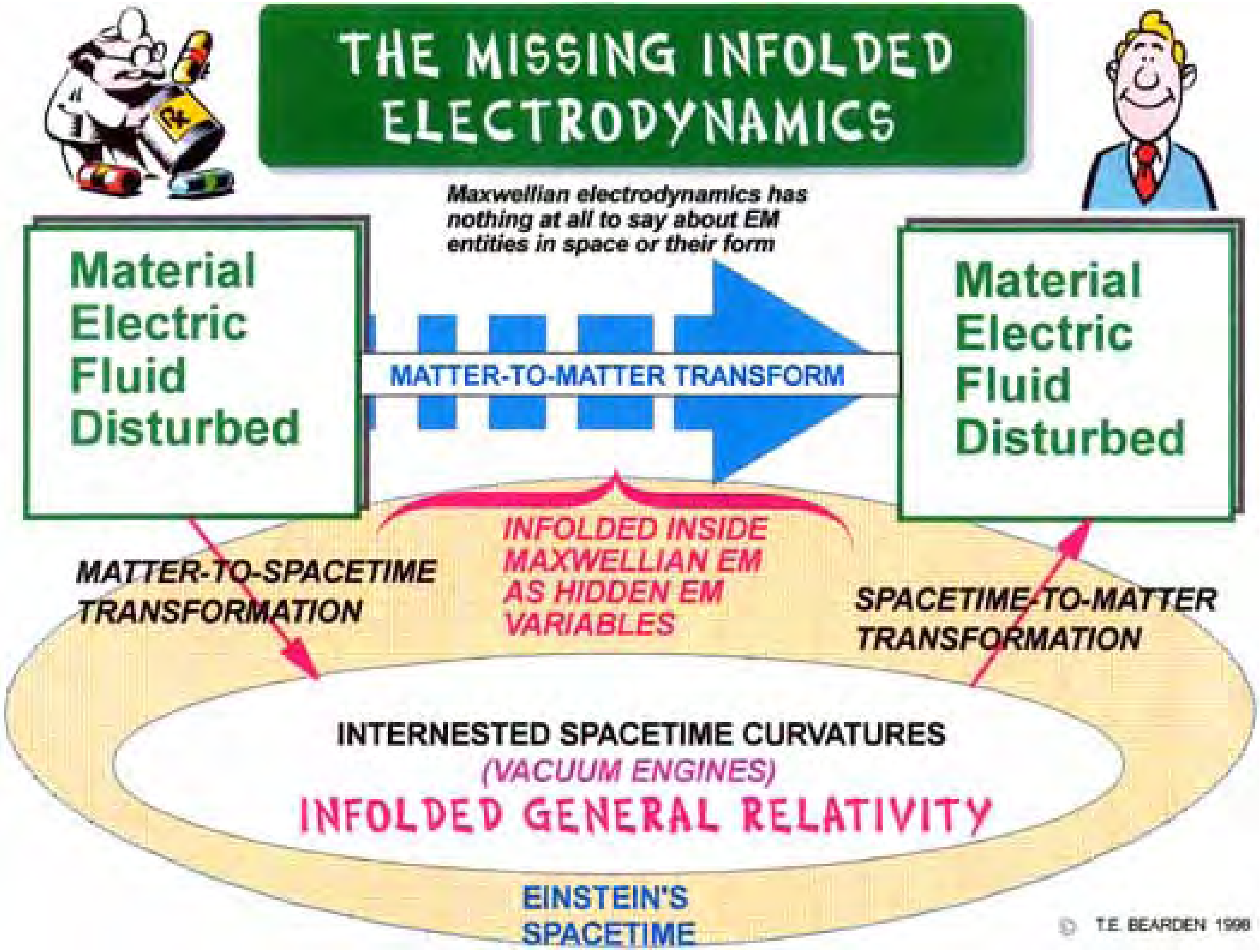

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

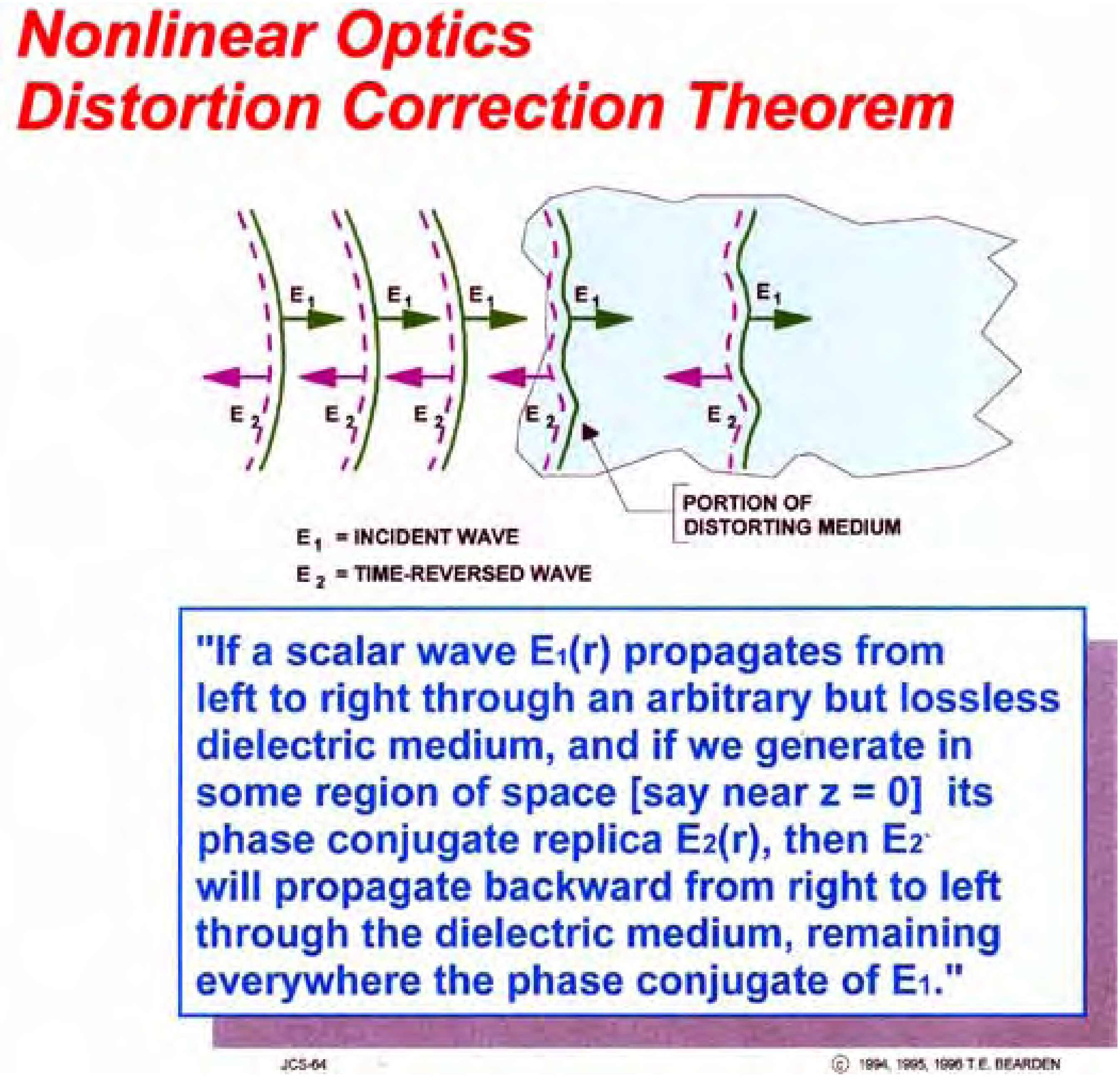

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE
E_2 = TIME-REVERSED WAVE

> "If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

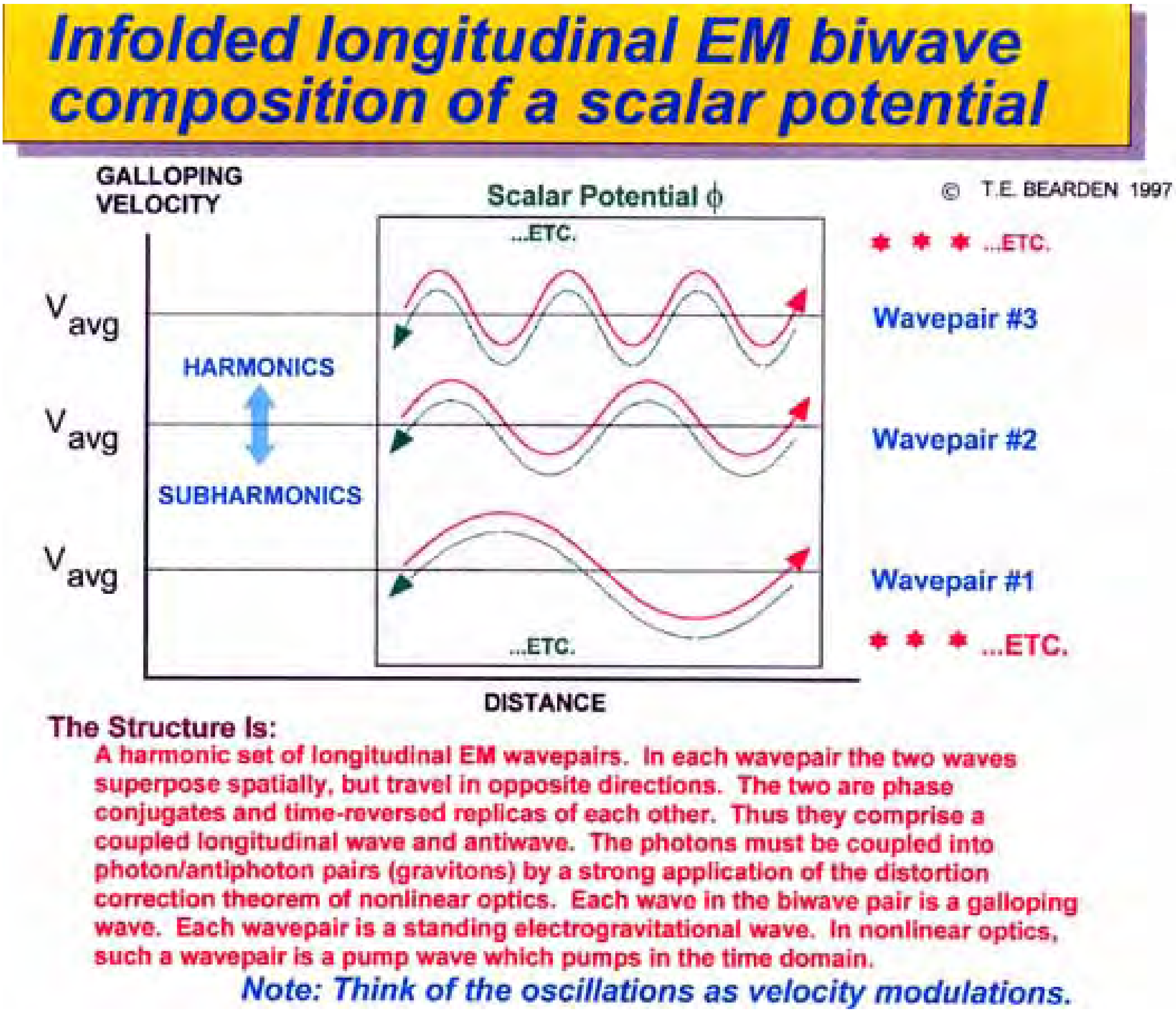

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
✱ ✱ ✱ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Return to Slide Index

Next Slide

Previous Slide

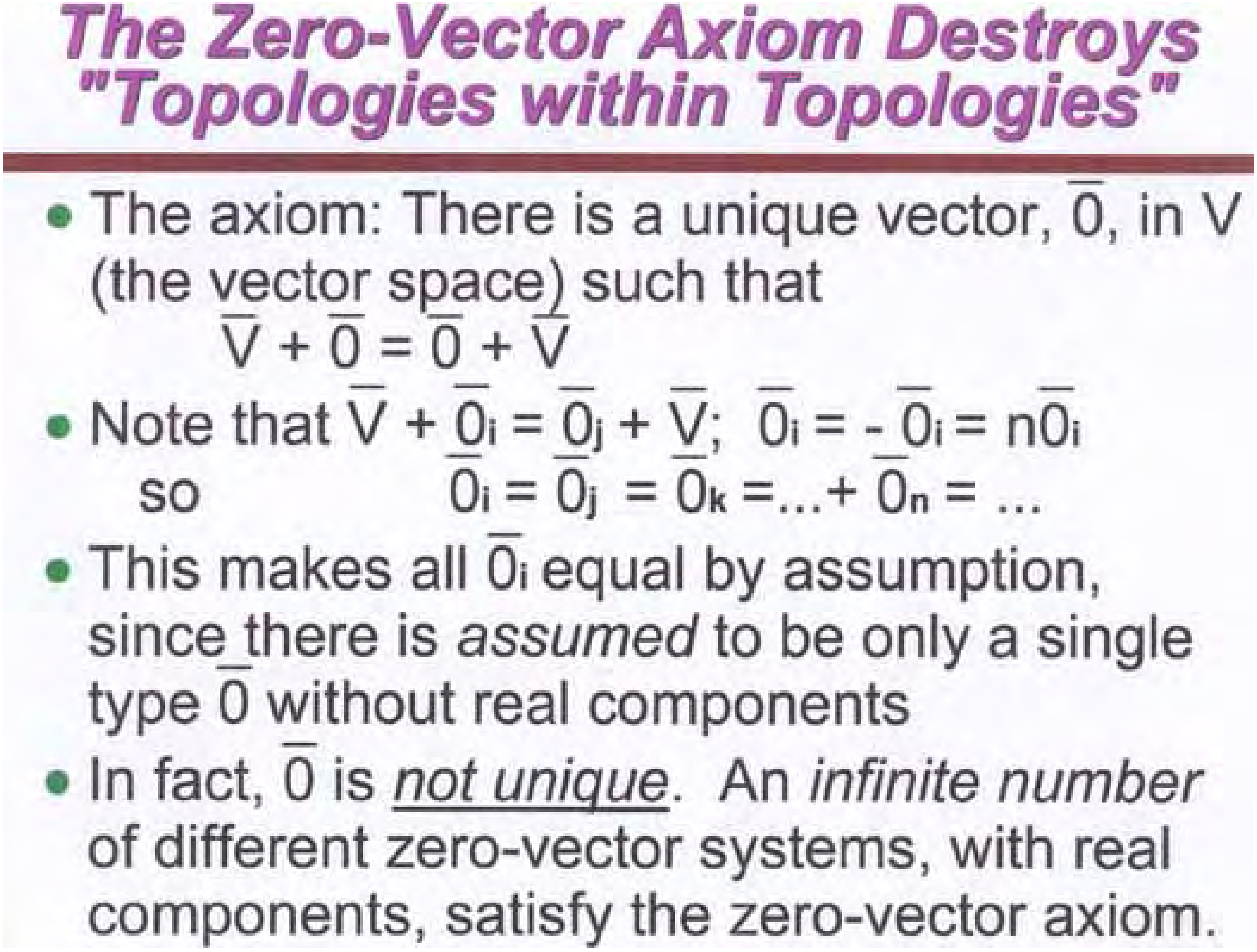

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $\overline{V} + \overline{0} = \overline{0} + \overline{V}$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

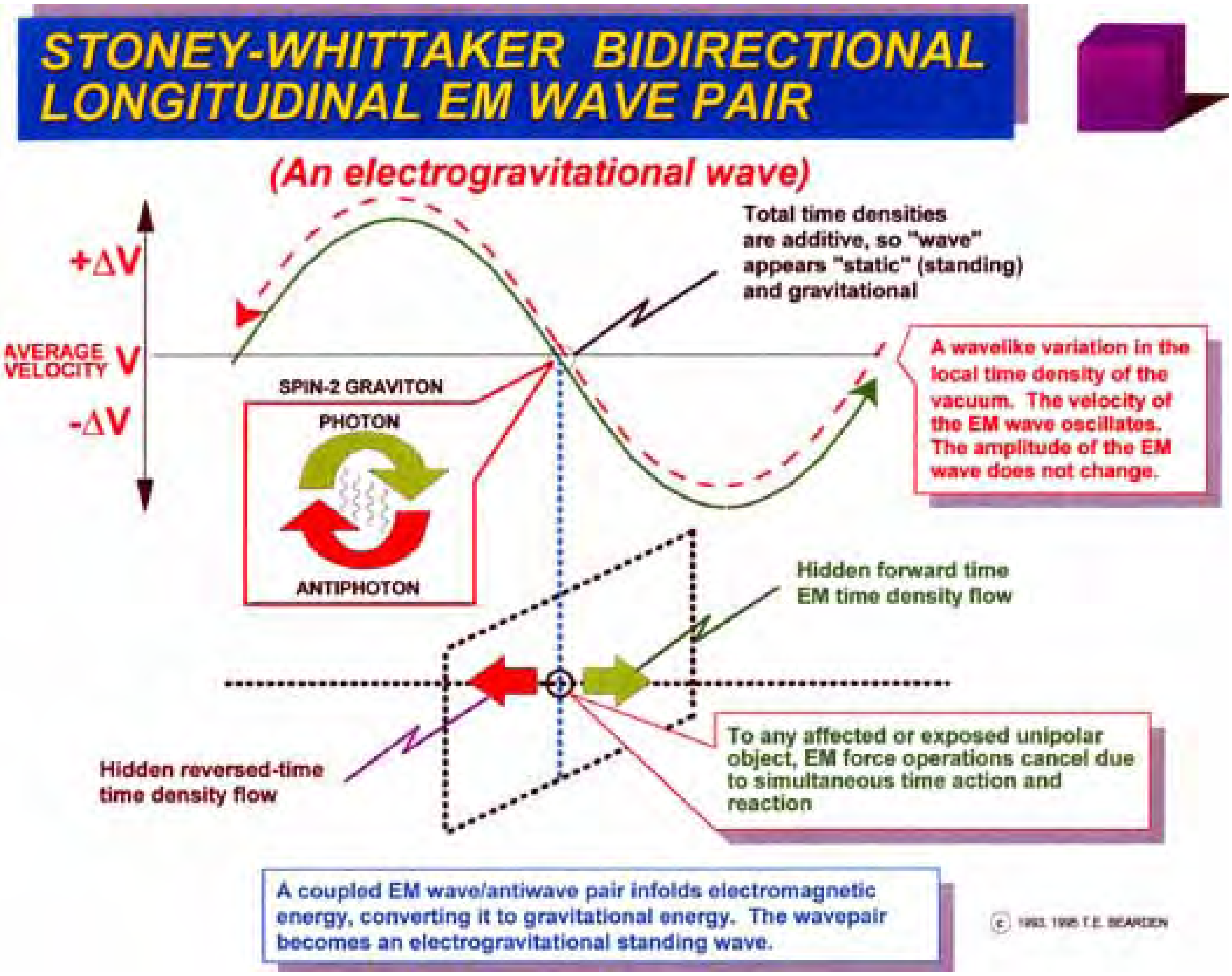

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

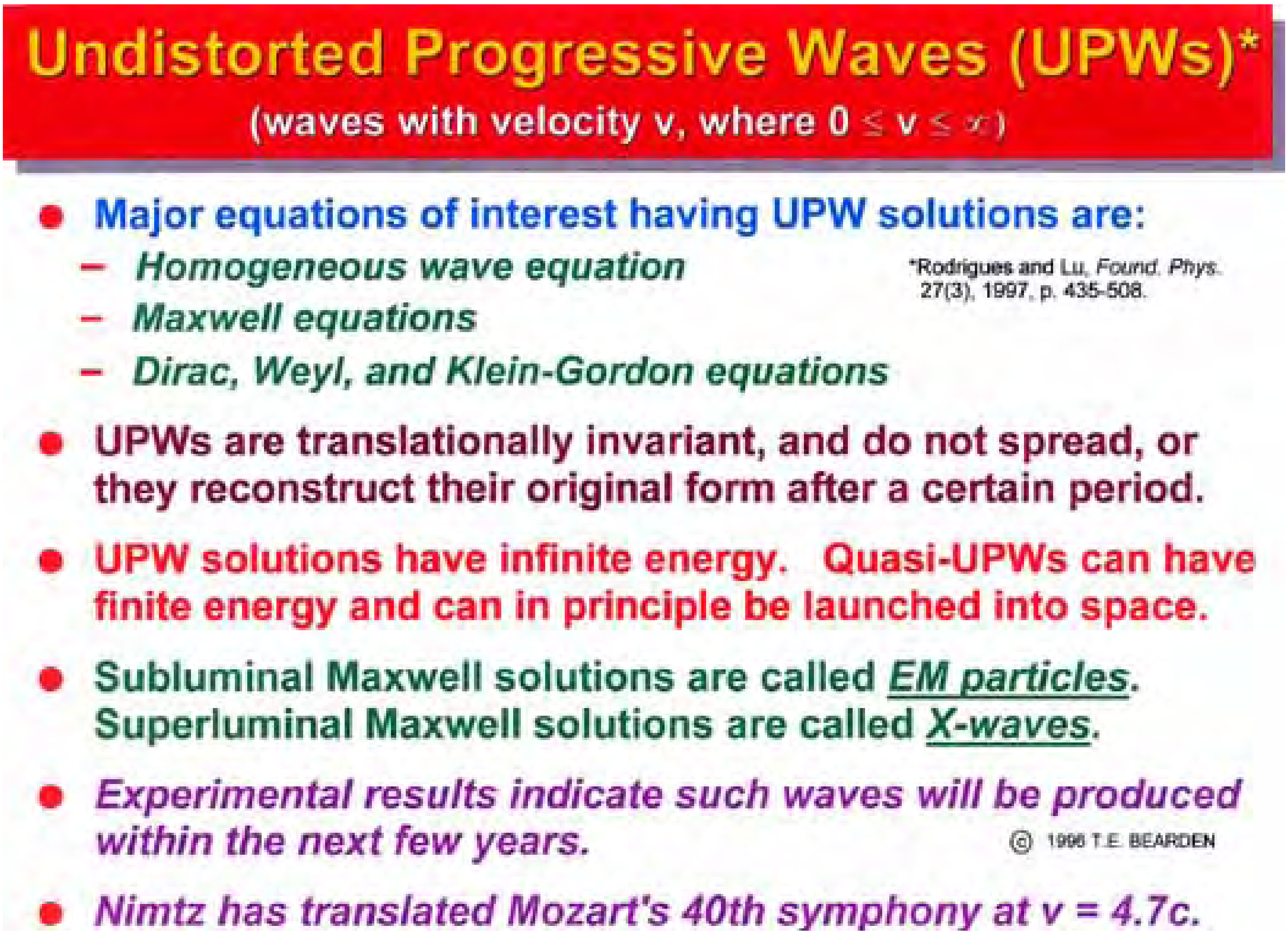

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \leq v \leq \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*
- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*
- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

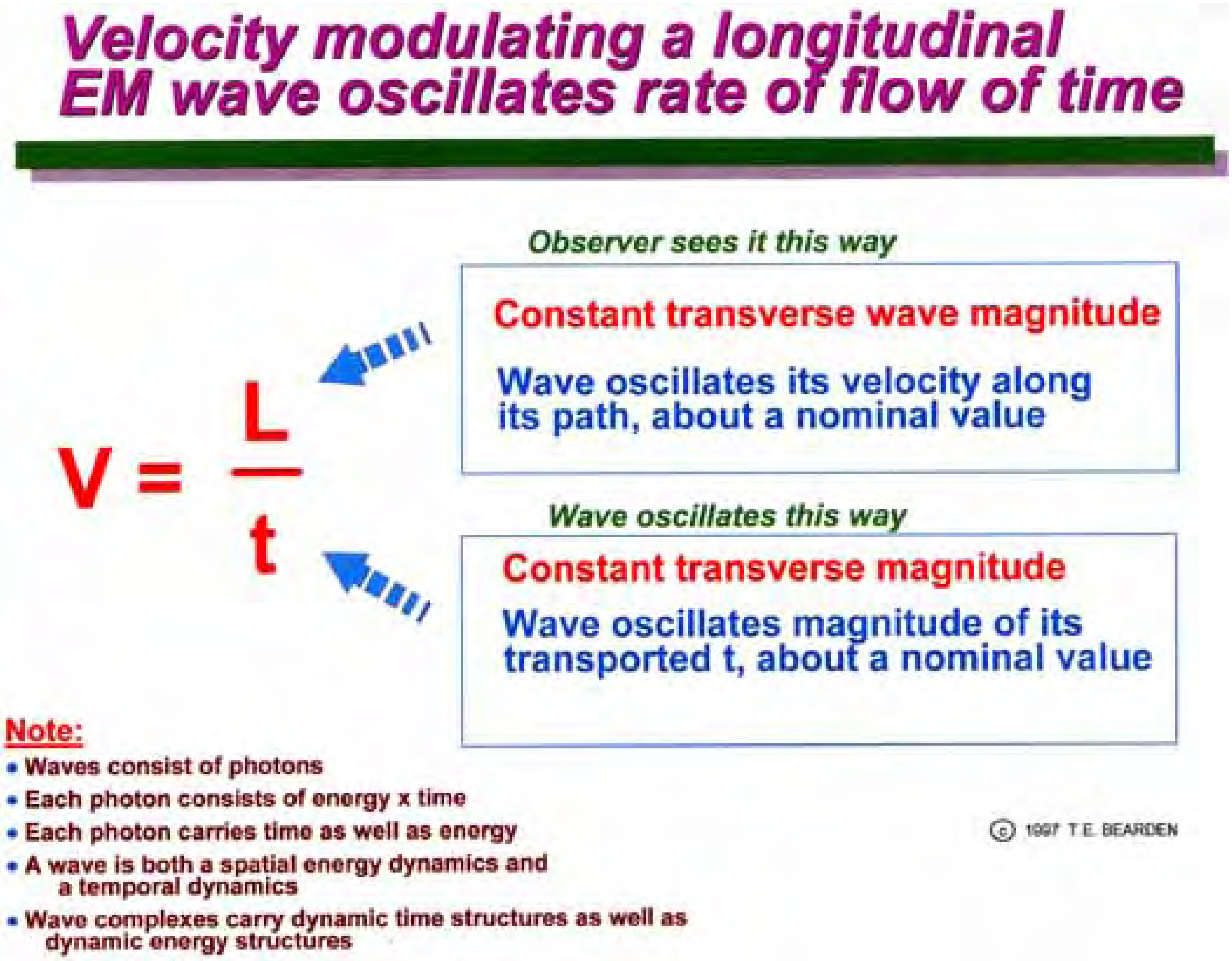

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$V = \frac{L}{t}$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

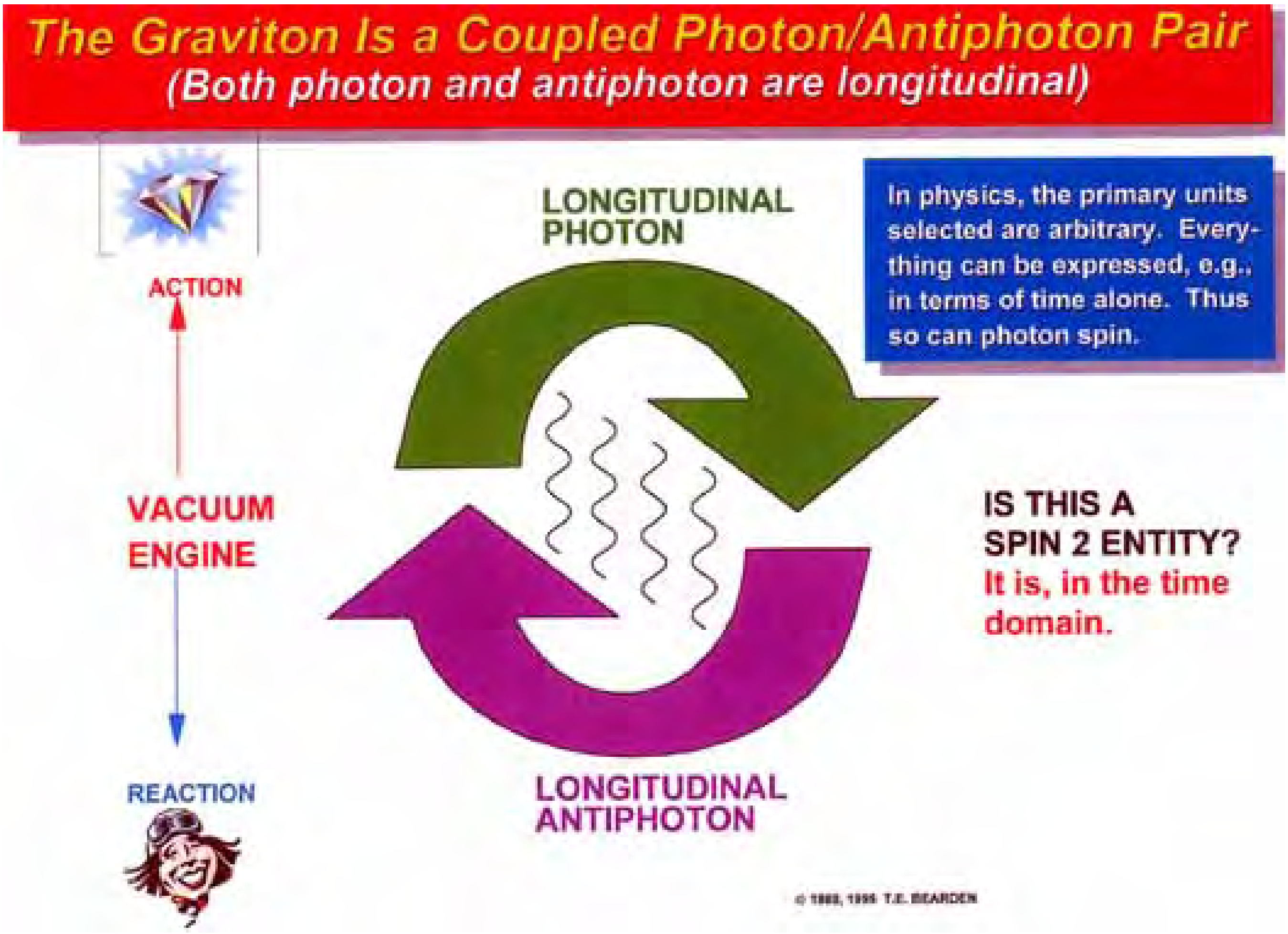

The Graviton Is a Coupled Photon/Antiphoton Pair

(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

IS THIS A SPIN 2 ENTITY?

It is, in the time domain.

LONGITUDINAL PHOTON
LONGITUDINAL ANTIPHOTON
ACTION
VACUUM ENGINE
REACTION

© 1988, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

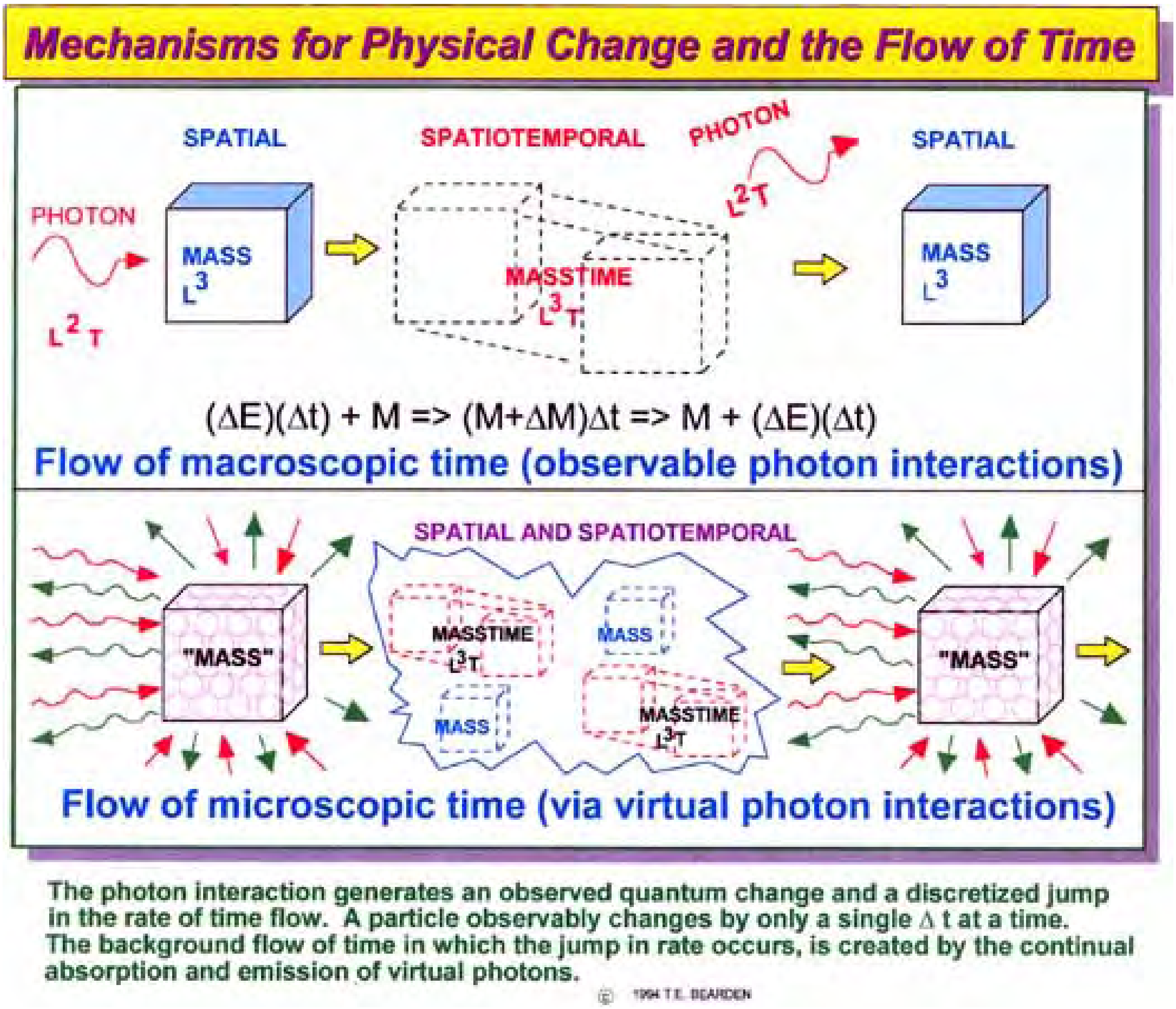

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
L^2T
MASS
L^3
MASSTIME
L^3T
L^2T
MASS
L^3

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δt at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

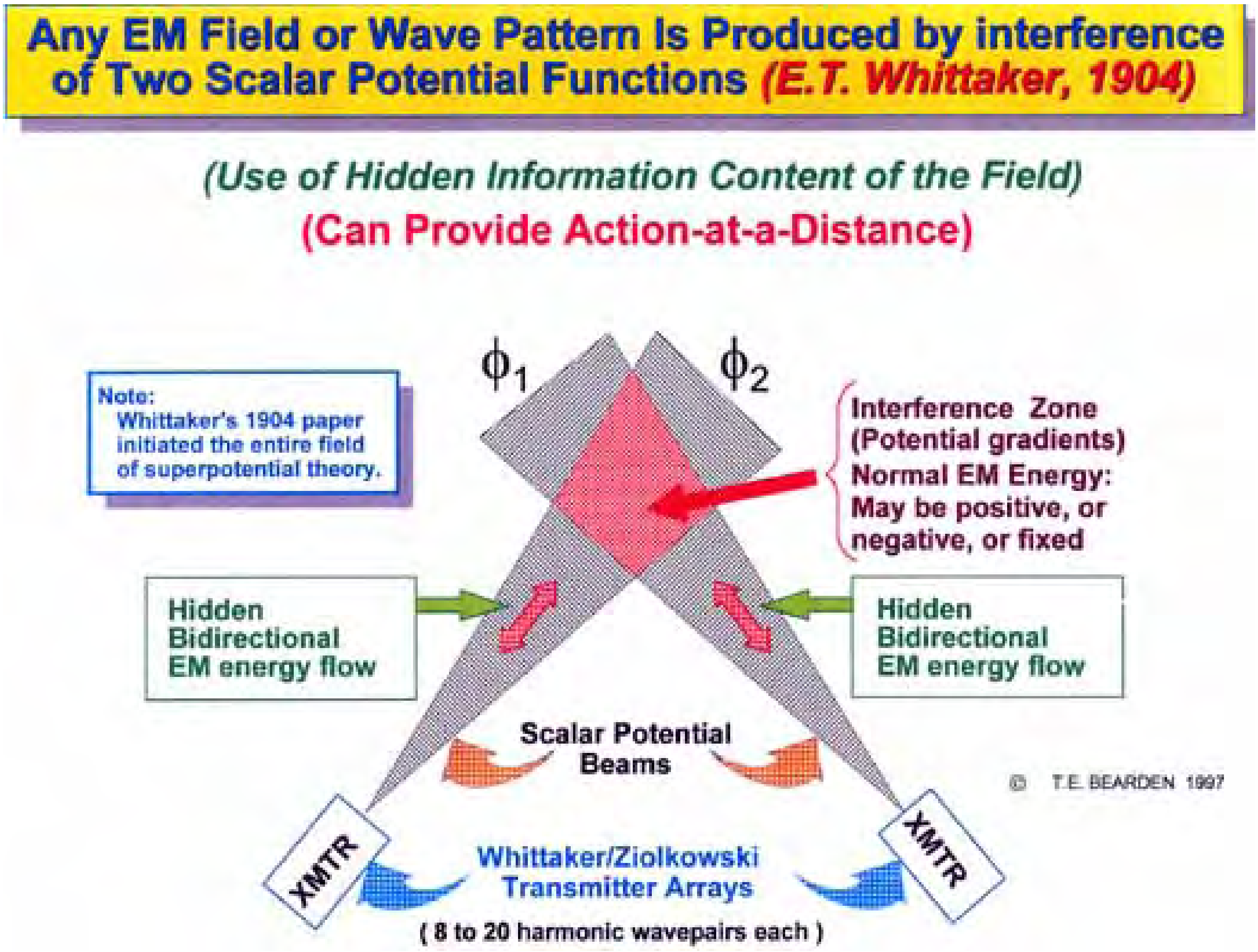

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or negative, or fixed

Hidden
Bidirectional
EM energy flow

Hidden
Bidirectional
EM energy flow

Scalar Potential Beams

Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

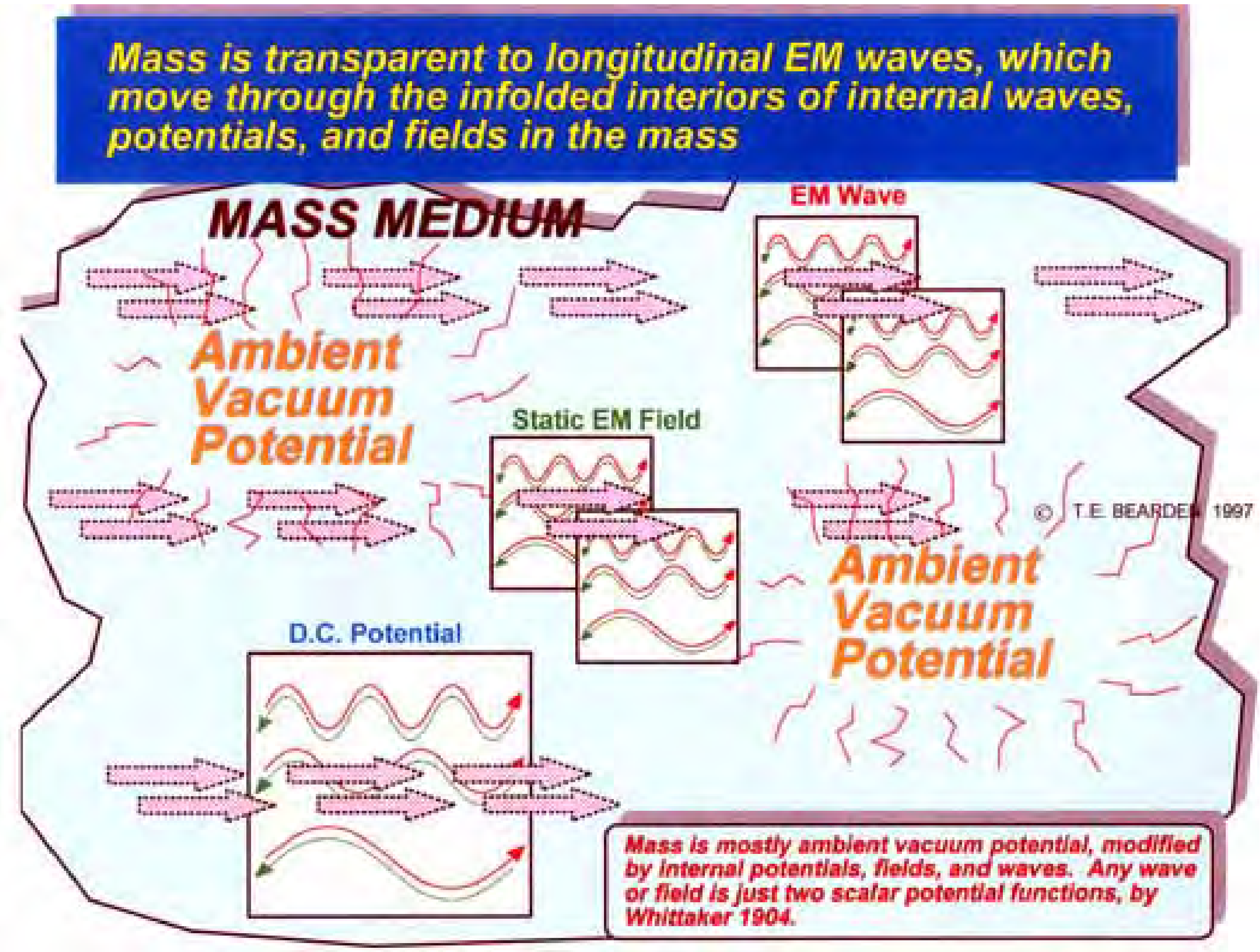

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
EM Wave
Ambient Vacuum Potential
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

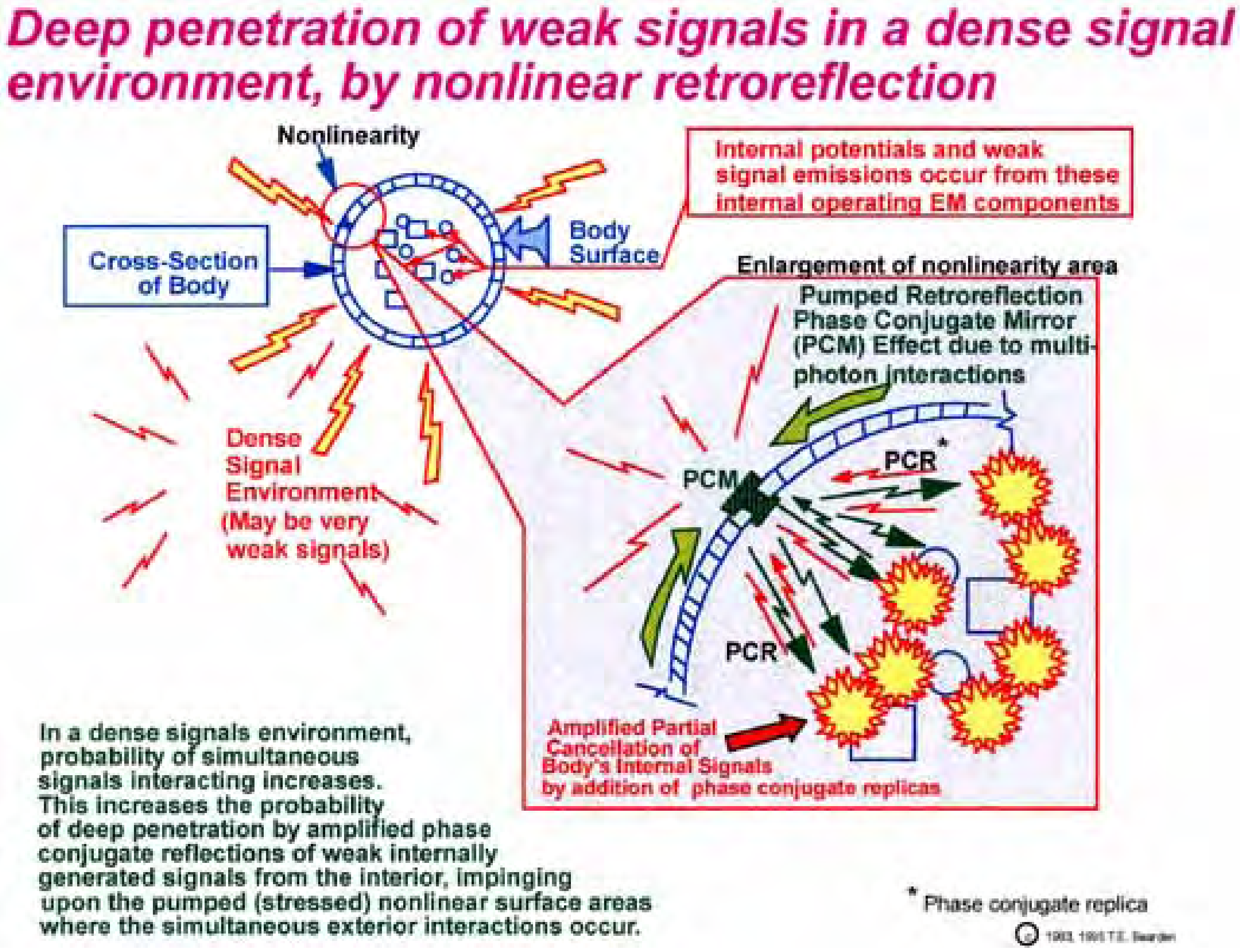

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity

Internal potentials and weak signal emissions occur from these internal operating EM components

Cross-Section of Body

Body Surface

Enlargement of nonlinearity area

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

Dense Signal Environment (May be very weak signals)

PCM

PCR*

PCR

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

THE IMMUNE SYSTEM

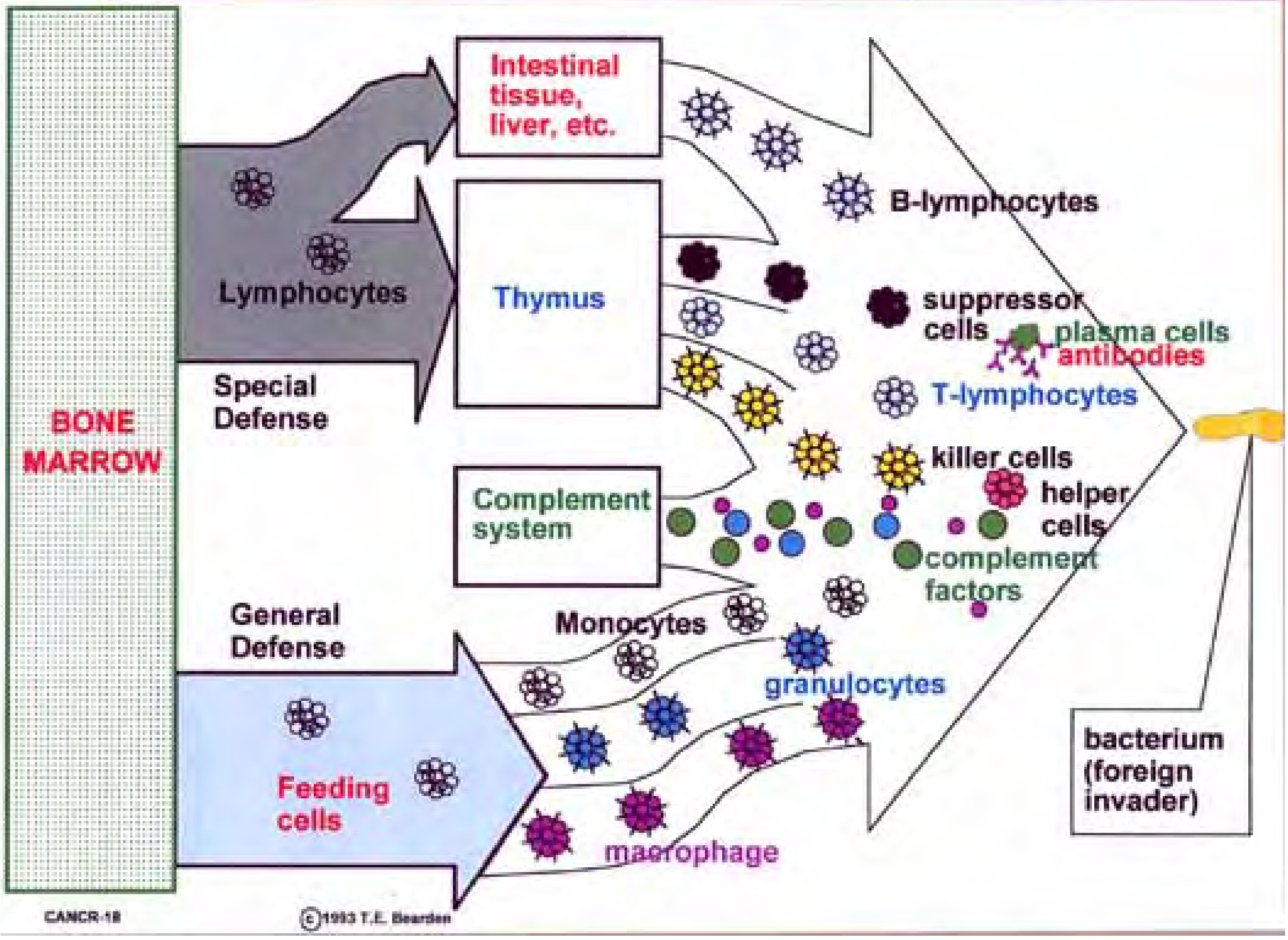
BONE MARROW
Lymphocytes
Special Defense
General Defense
Feeding cells
Intestinal tissue, liver, etc.
Thymus
Complement system
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
Monocytes
granulocytes
macrophage
bacterium (foreign invader)

CANCR-18

©1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

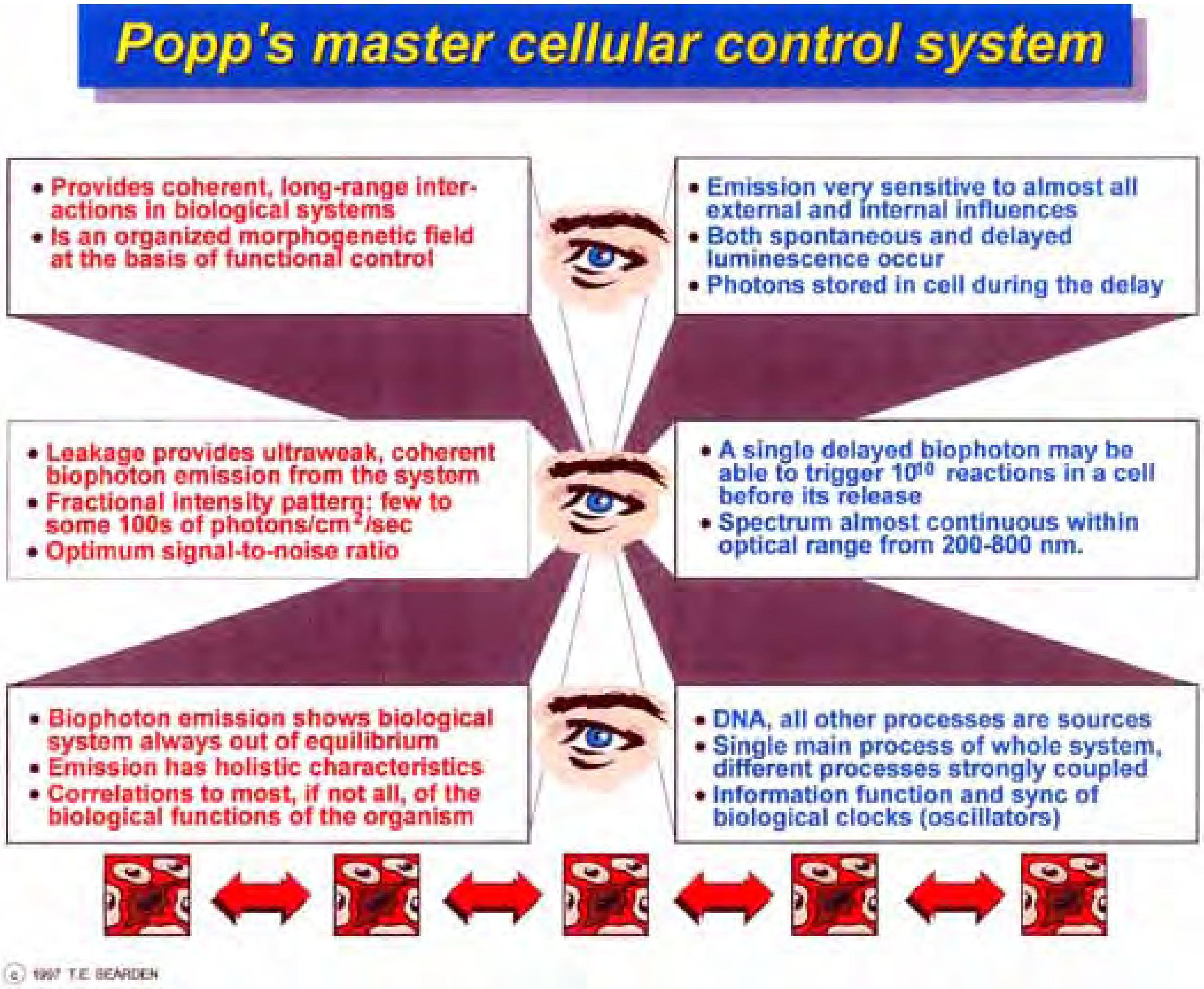

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

(c) 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

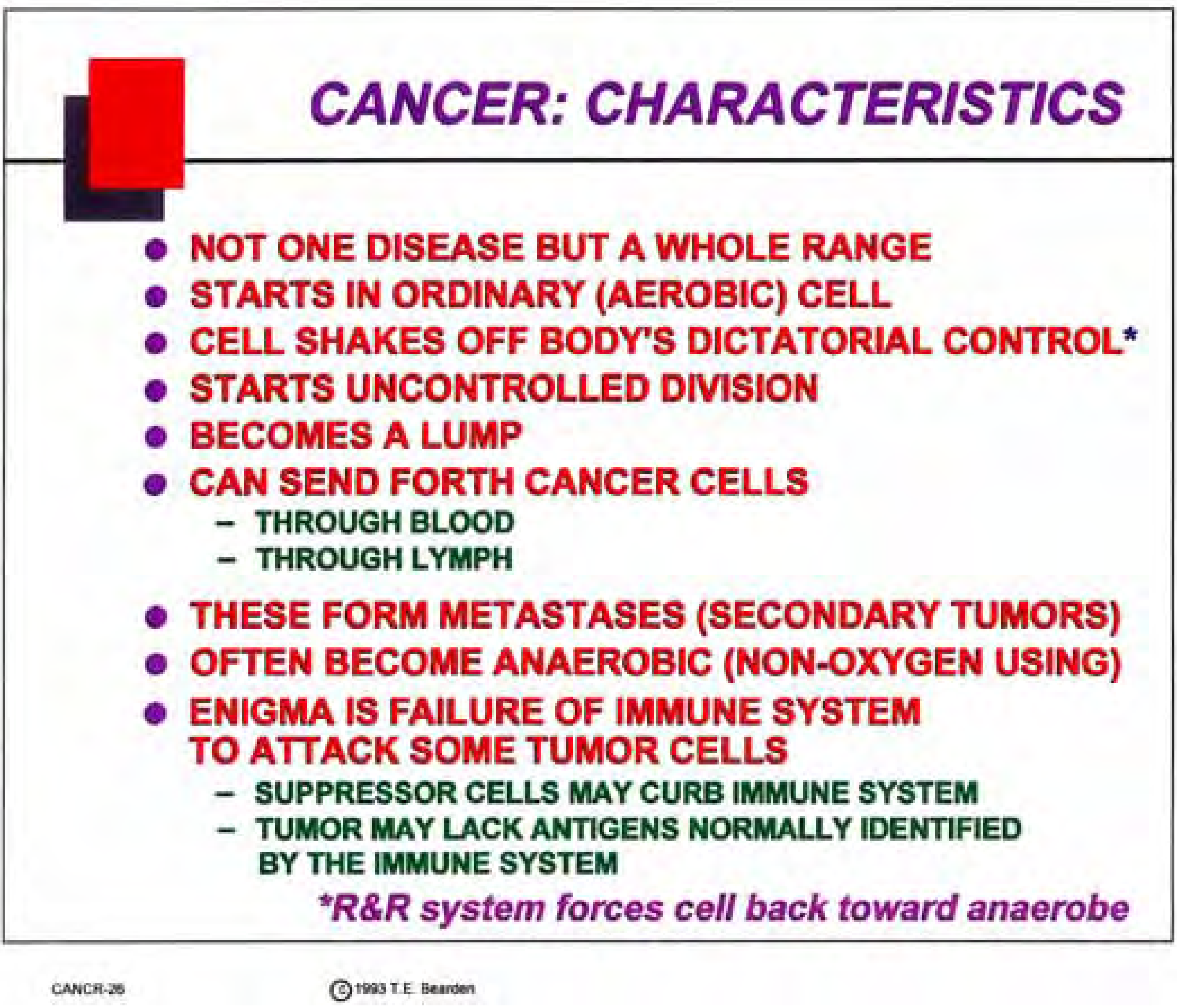

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

CANCR-26

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

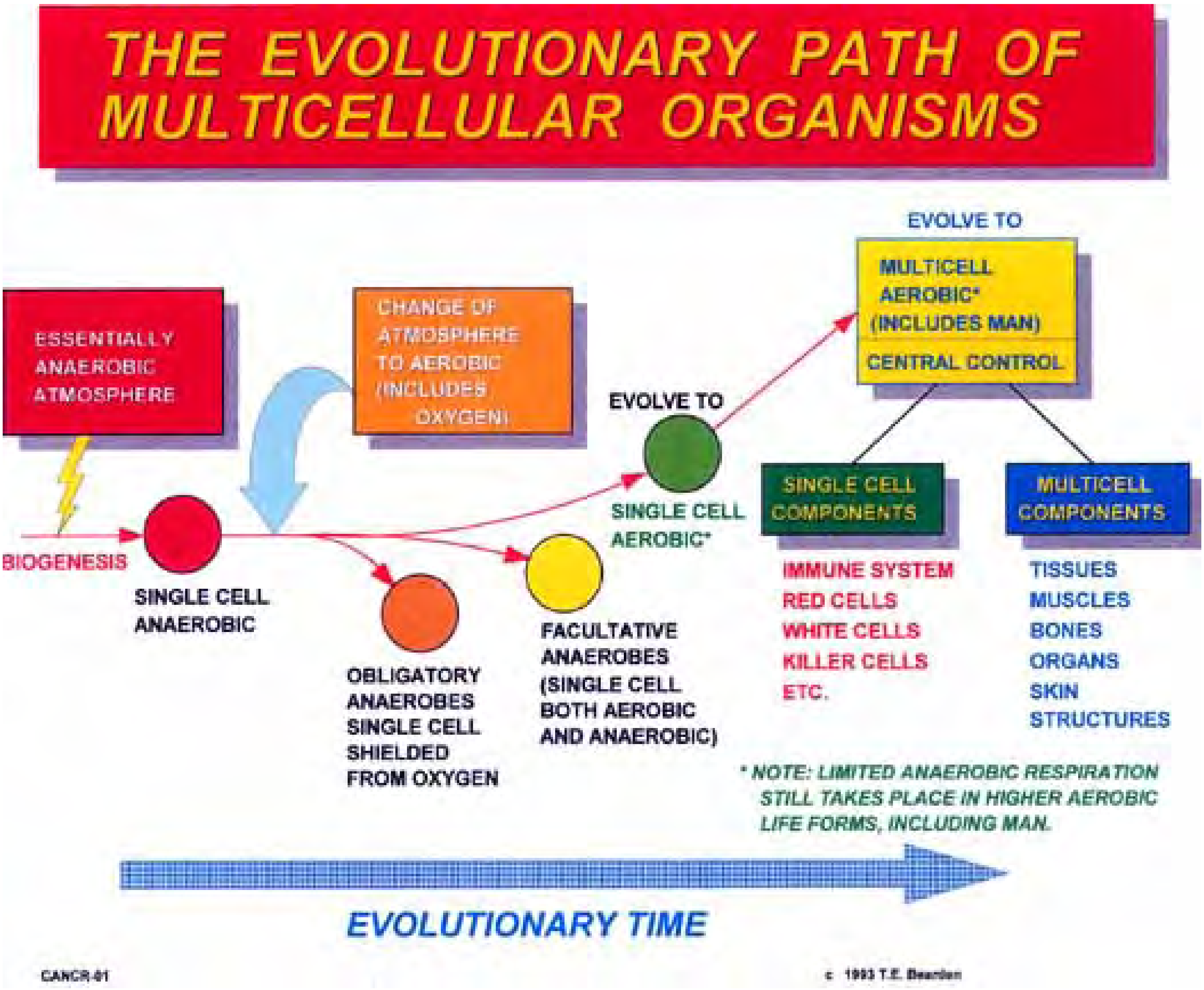

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY
ANAEROBIC
ATMOSPHERE

CHANGE OF
ATMOSPHERE
TO AEROBIC
(INCLUDES
OXYGEN)

BIOGENESIS

SINGLE CELL
ANAEROBIC

OBLIGATORY
ANAEROBES
SINGLE CELL
SHIELDED
FROM OXYGEN

FACULTATIVE
ANAEROBES
(SINGLE CELL
BOTH AEROBIC
AND ANAEROBIC)

EVOLVE TO

SINGLE CELL
AEROBIC*

EVOLVE TO

MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL
COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL
COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

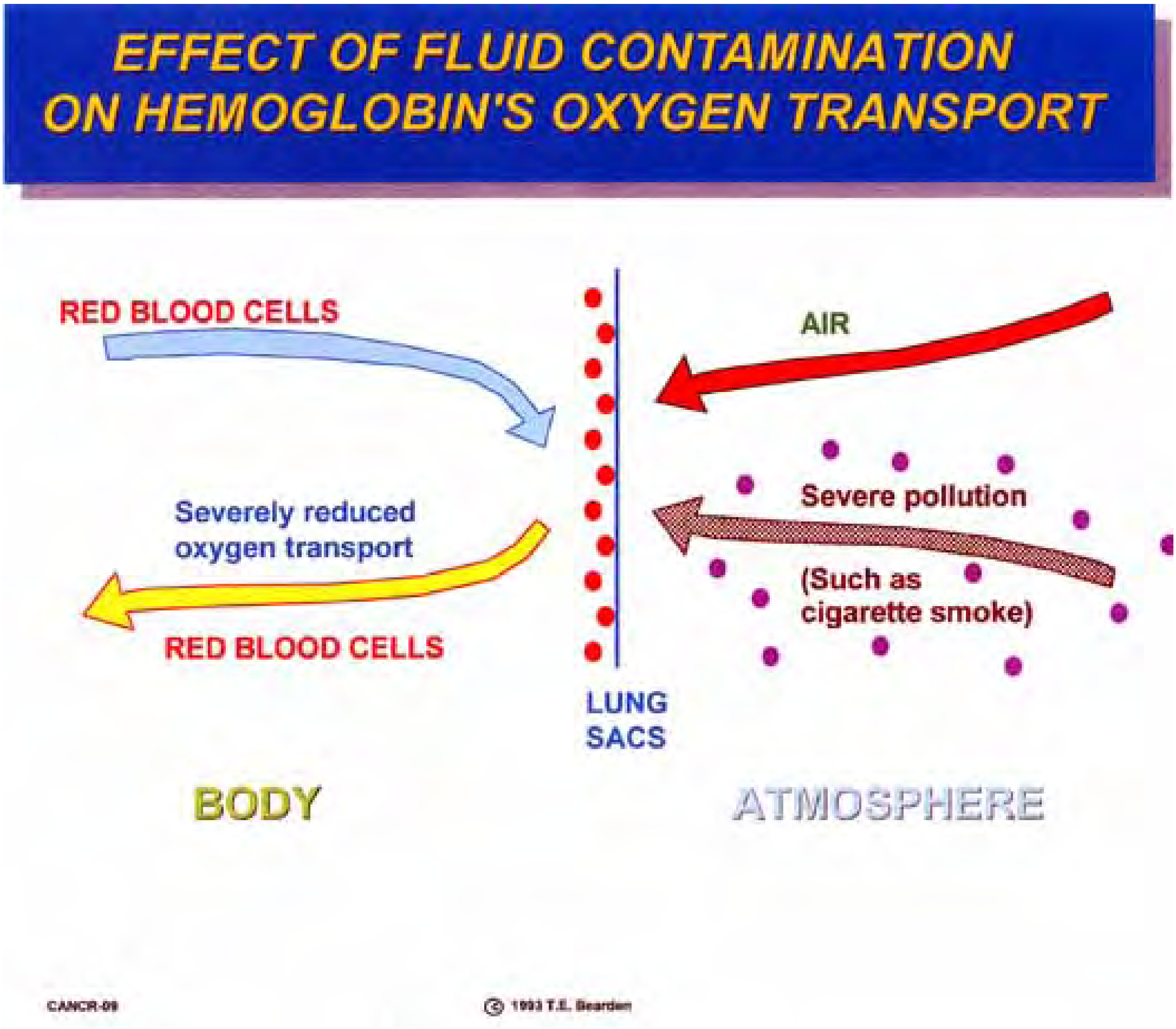

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

Severely reduced
oxygen transport

(Such as
cigarette smoke)

RED BLOOD CELLS

LUNG
SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

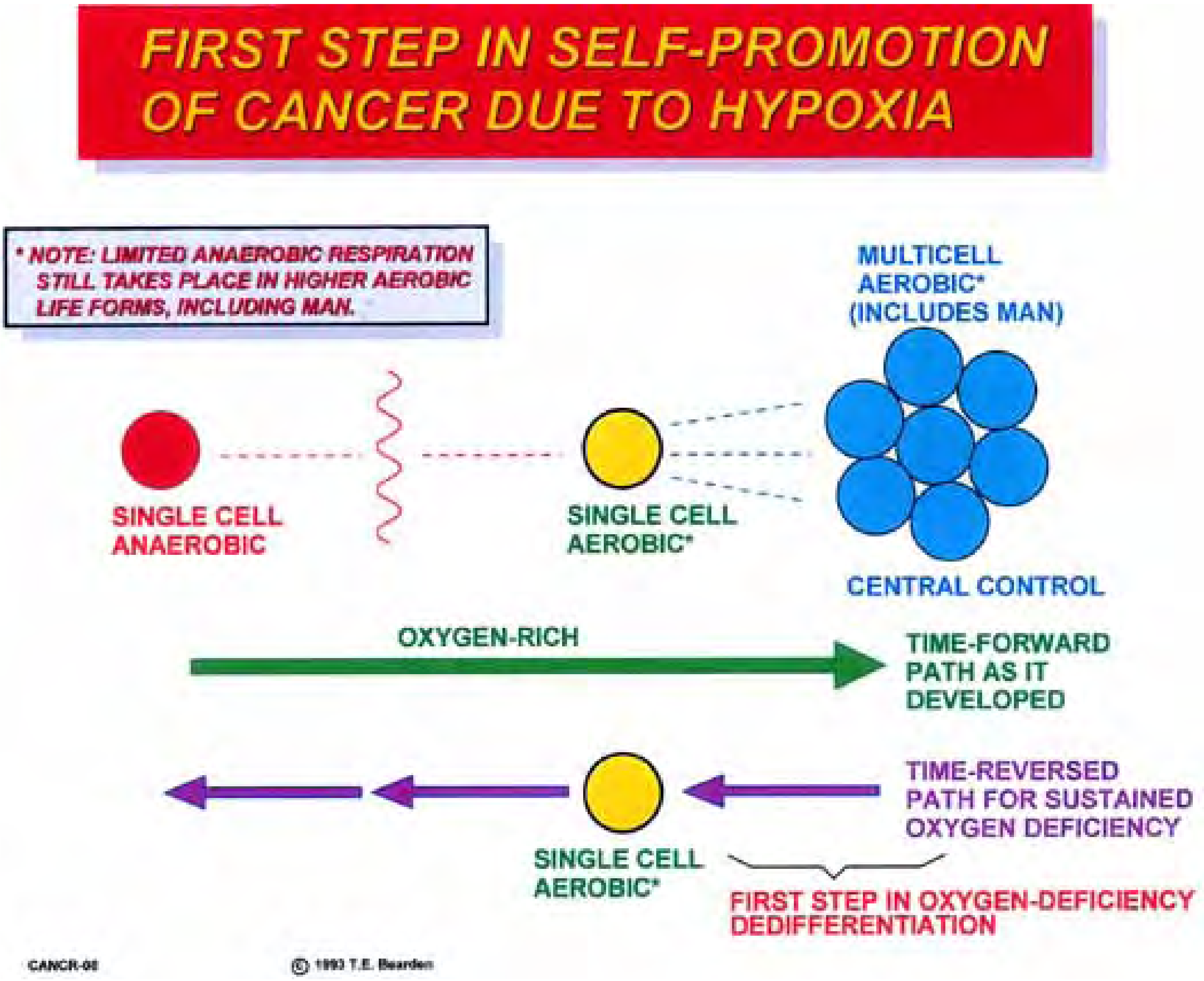

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL
ANAEROBIC
SINGLE CELL
AEROBIC*
MULTICELL
AEROBIC*
(INCLUDES MAN)
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD
PATH AS IT
DEVELOPED
TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY
SINGLE CELL
AEROBIC*
FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

[Return to Slide Index]

[Next Slide]

[Previous Slide]

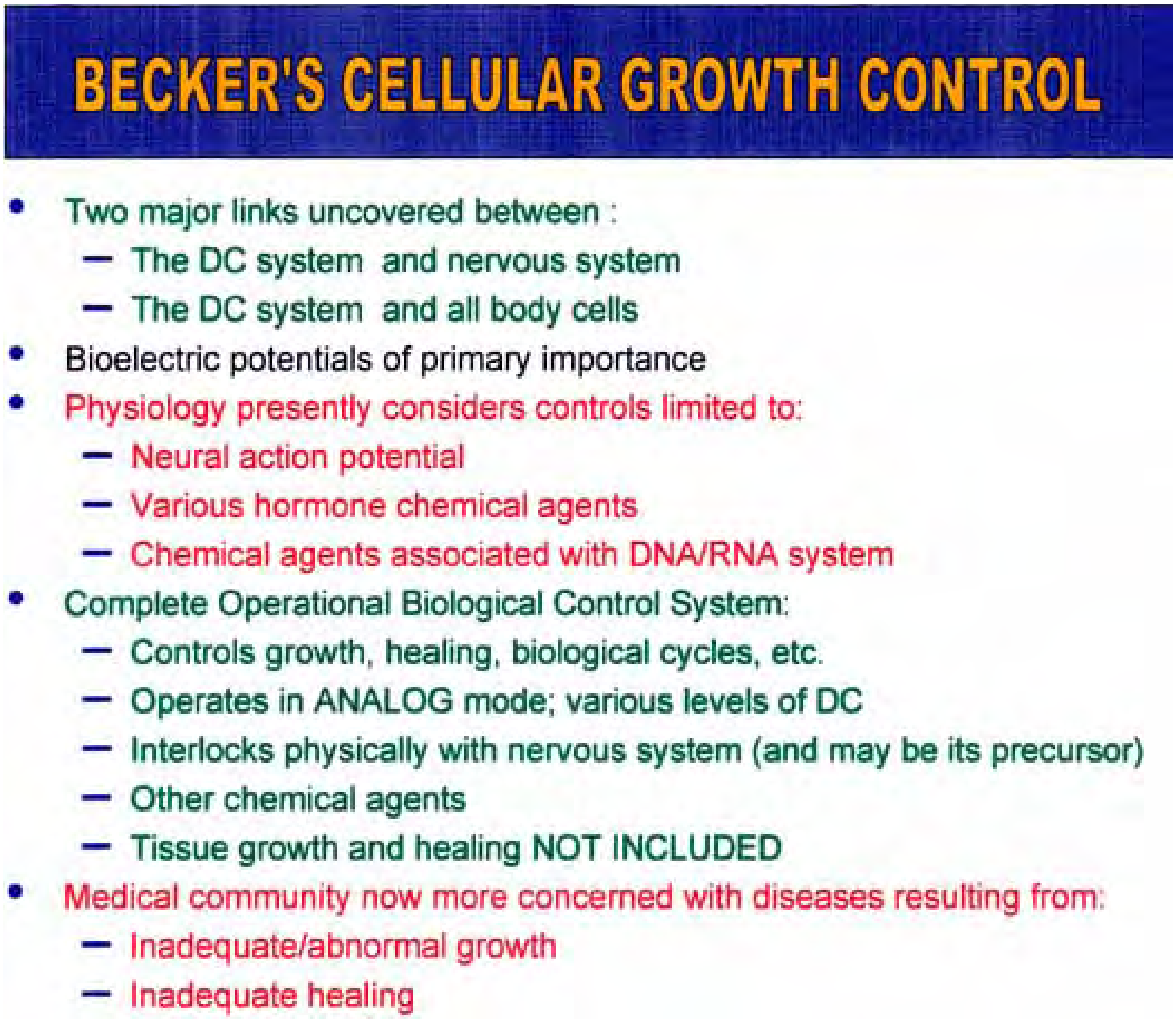

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

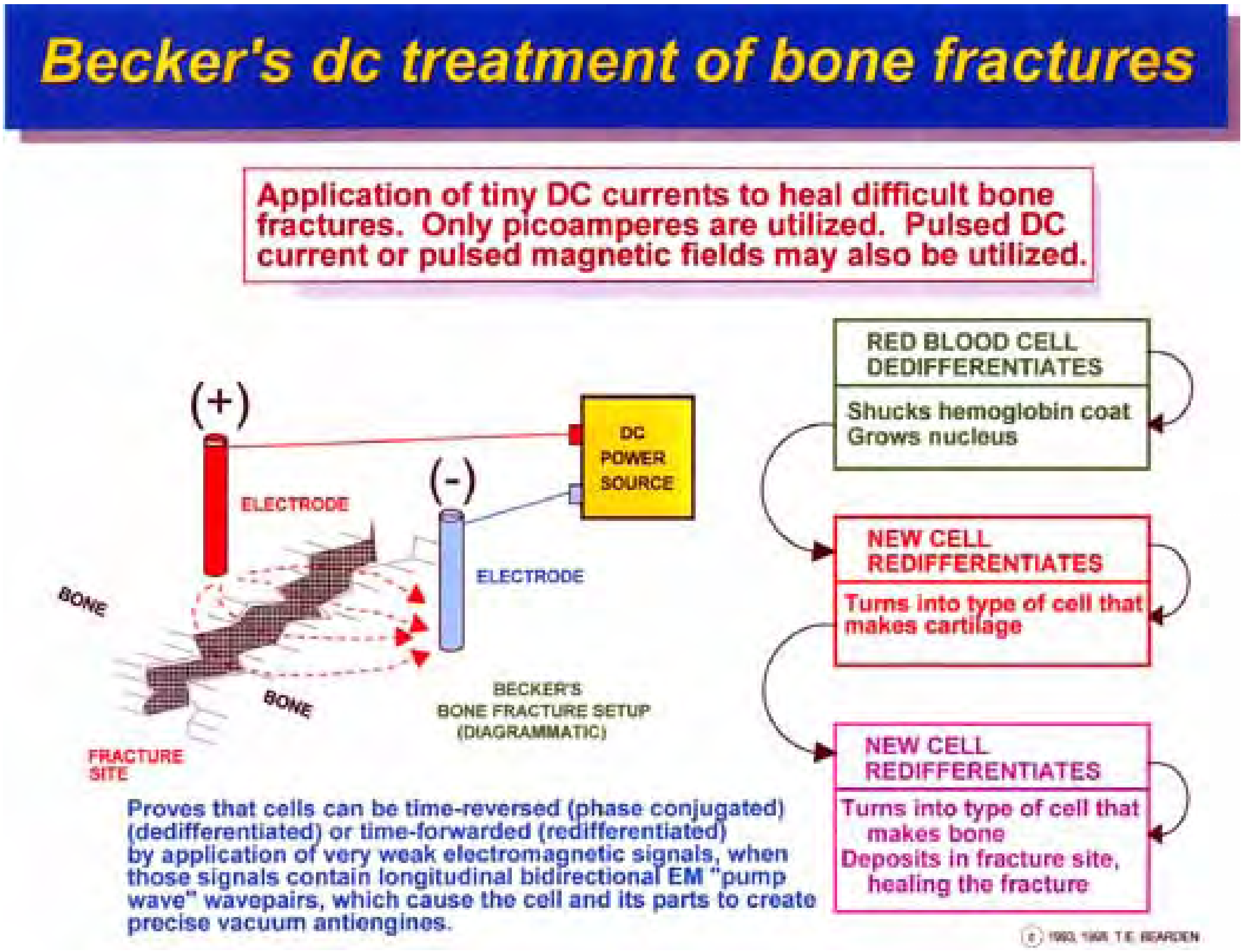

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

BECKER'S
BONE FRACTURE SETUP
(DIAGRAMMATIC)

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

© 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

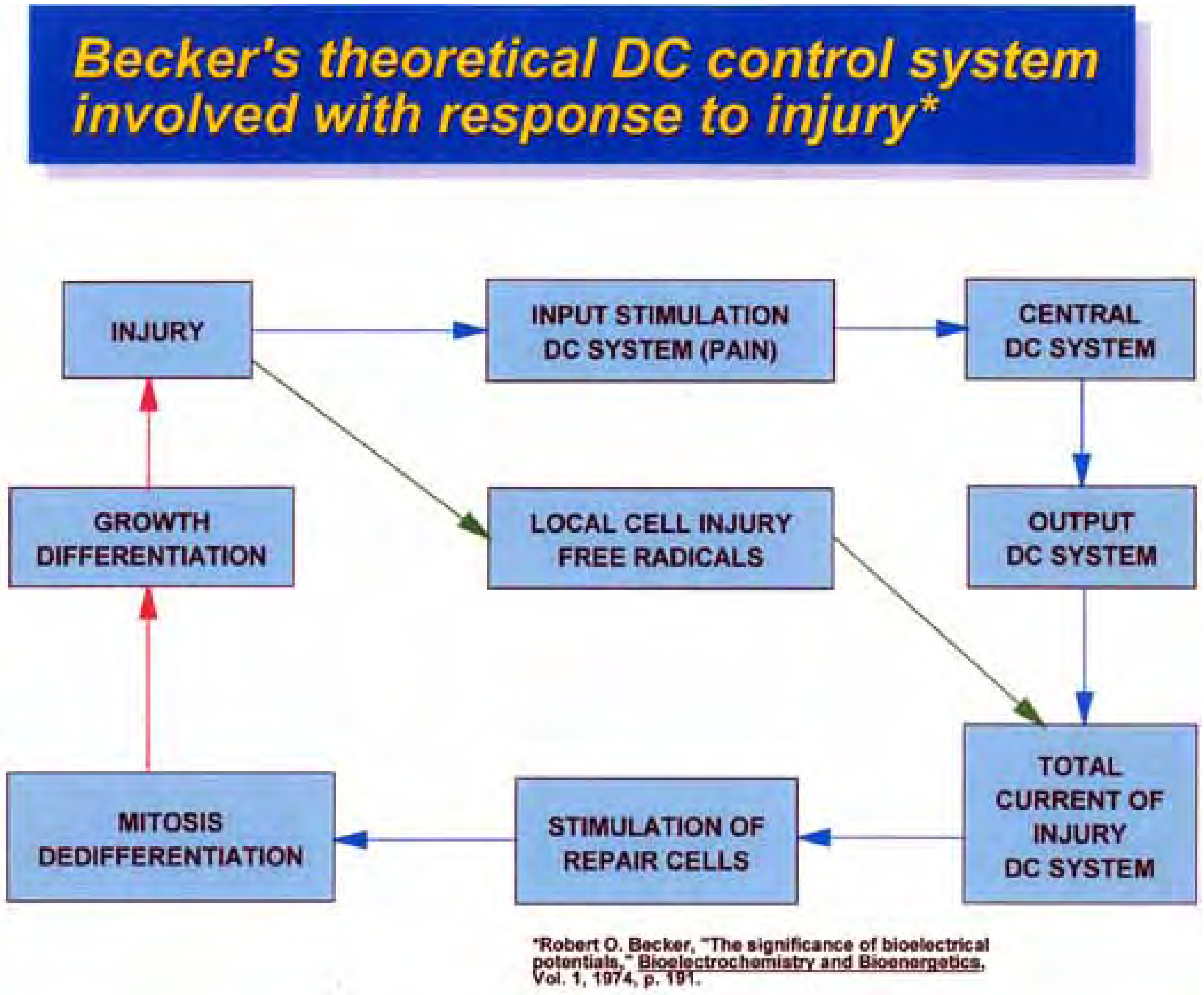

*Becker's theoretical DC control system involved with response to injury**

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Return to Slide Index

Next Slide

Previous Slide

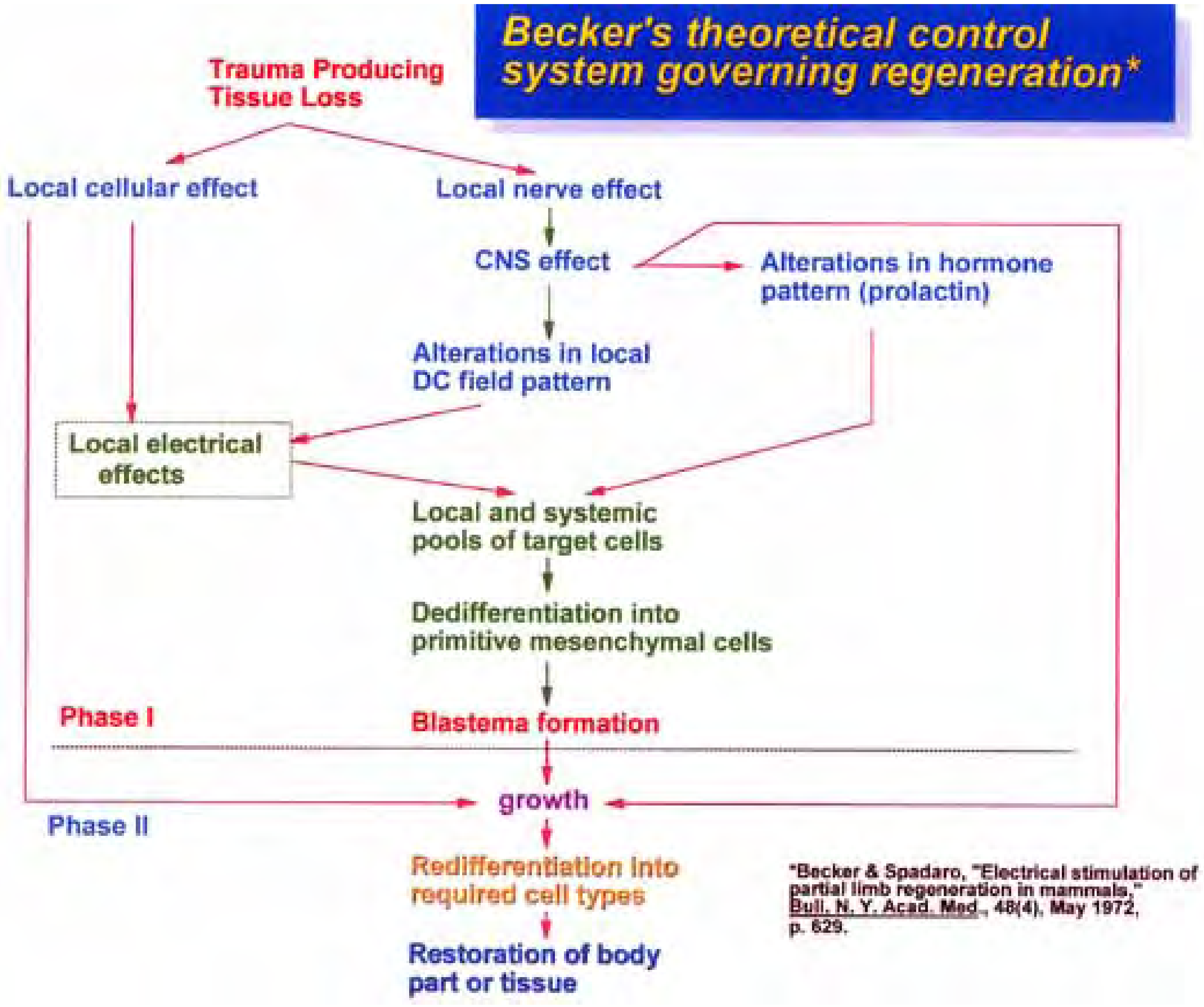

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Return to Slide Index

Next Slide

Previous Slide

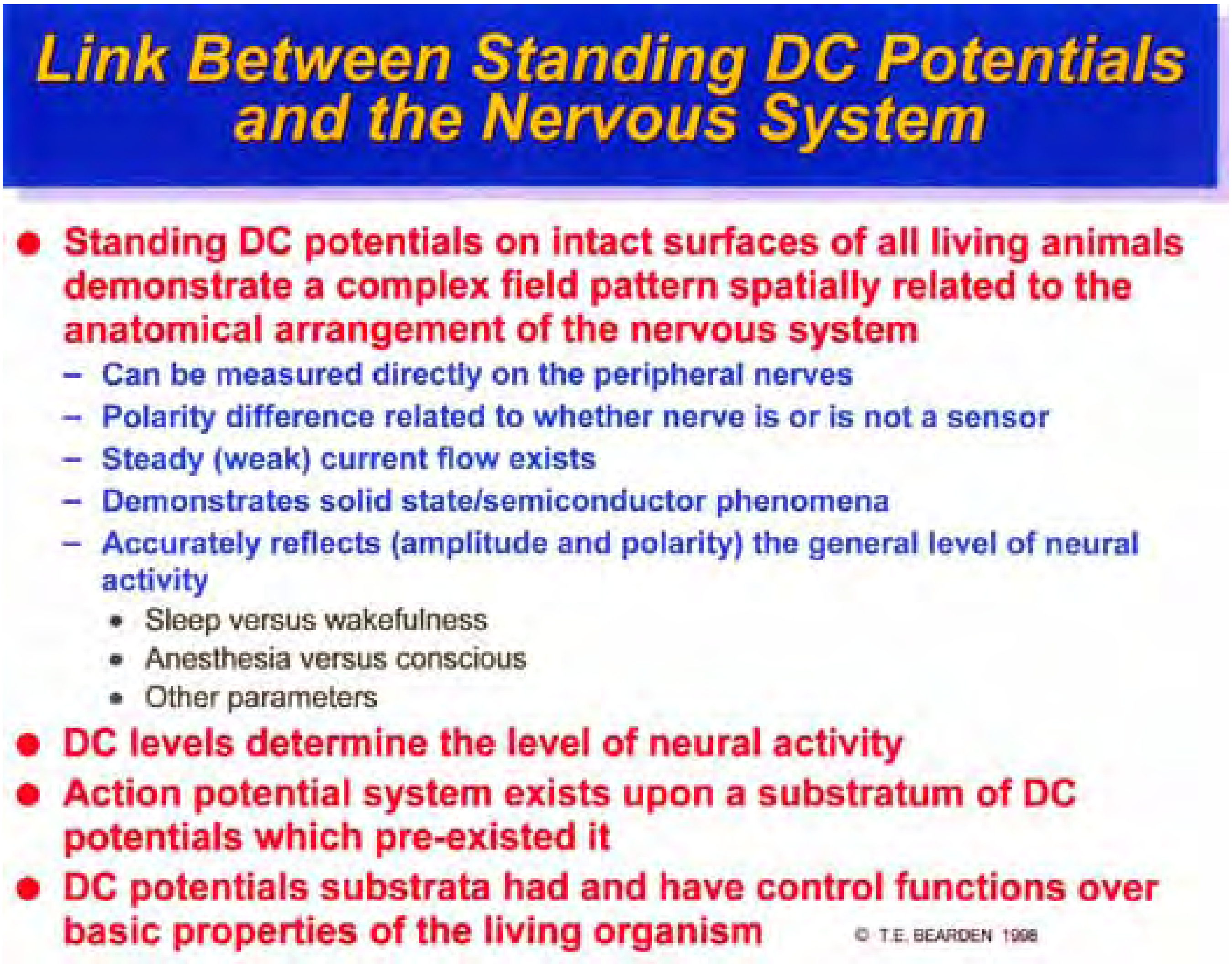

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

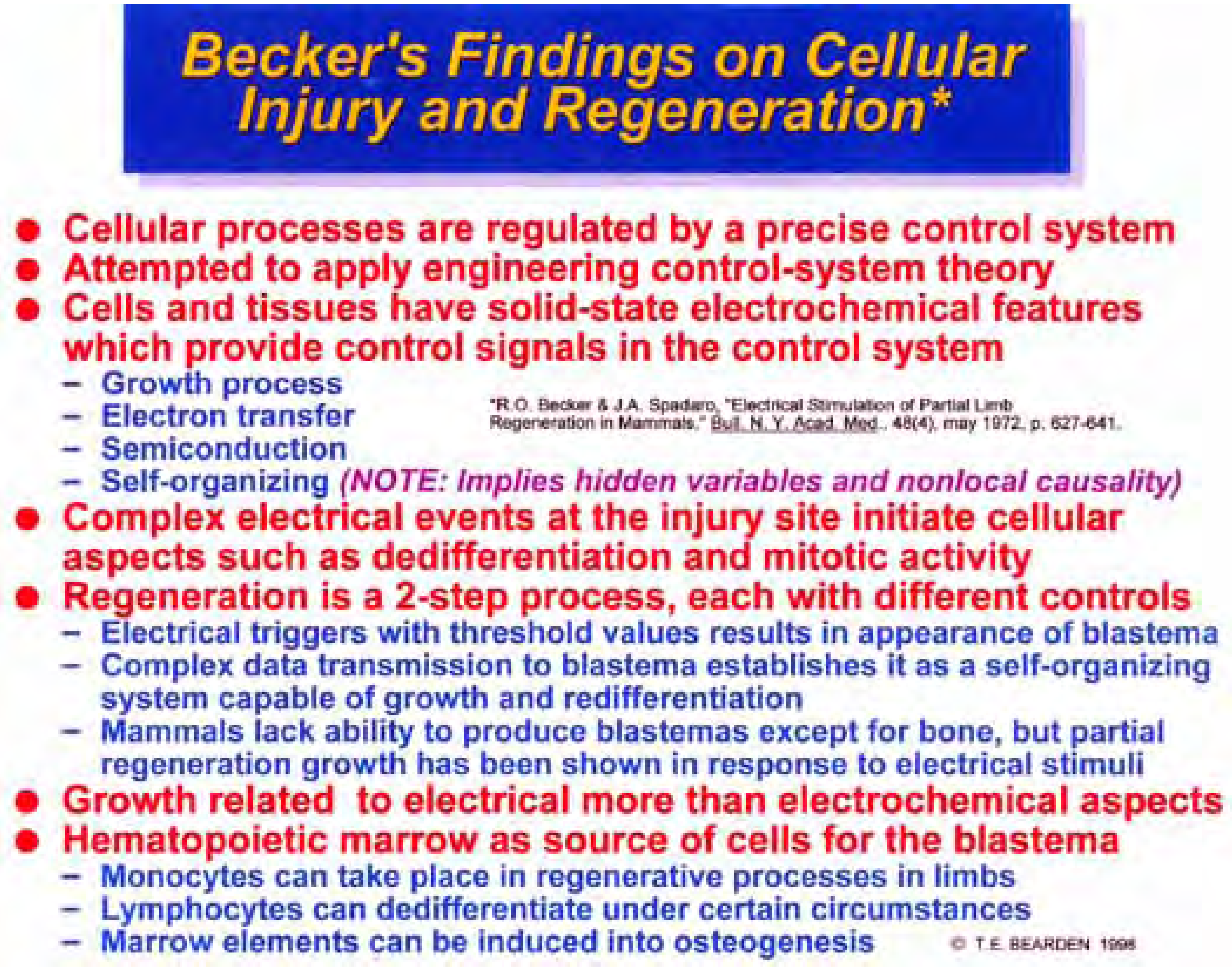

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

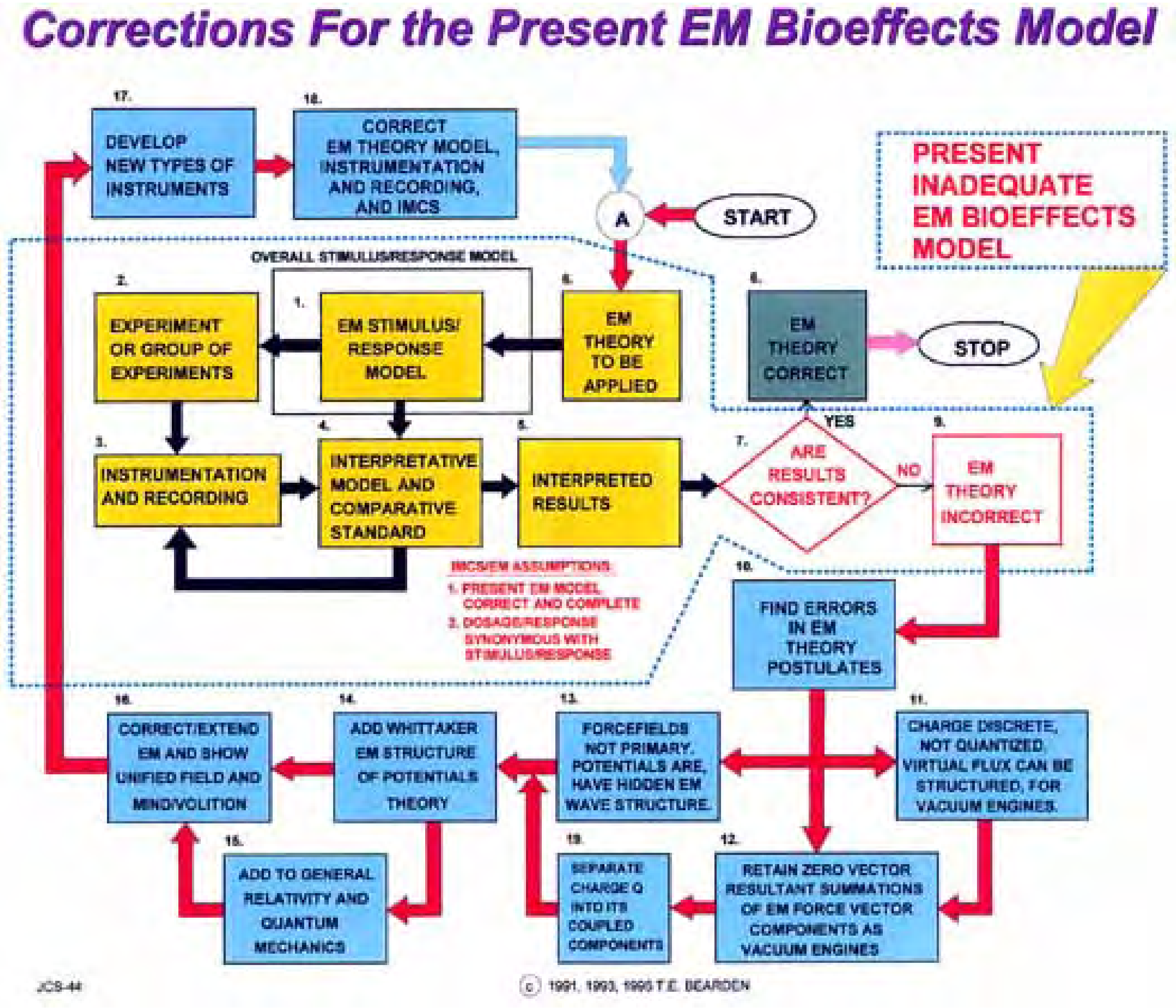

Corrections For the Present EM Bioeffects Model

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

A

START

PRESENT INADEQUATE EM BIOEFFECTS MODEL

OVERALL STIMULUS/RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

1. EM STIMULUS/ RESPONSE MODEL

6. EM THEORY TO BE APPLIED

8. EM THEORY CORRECT

STOP

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

7. ARE RESULTS CONSISTENT?

YES

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

JCS-44

(c) 1991, 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

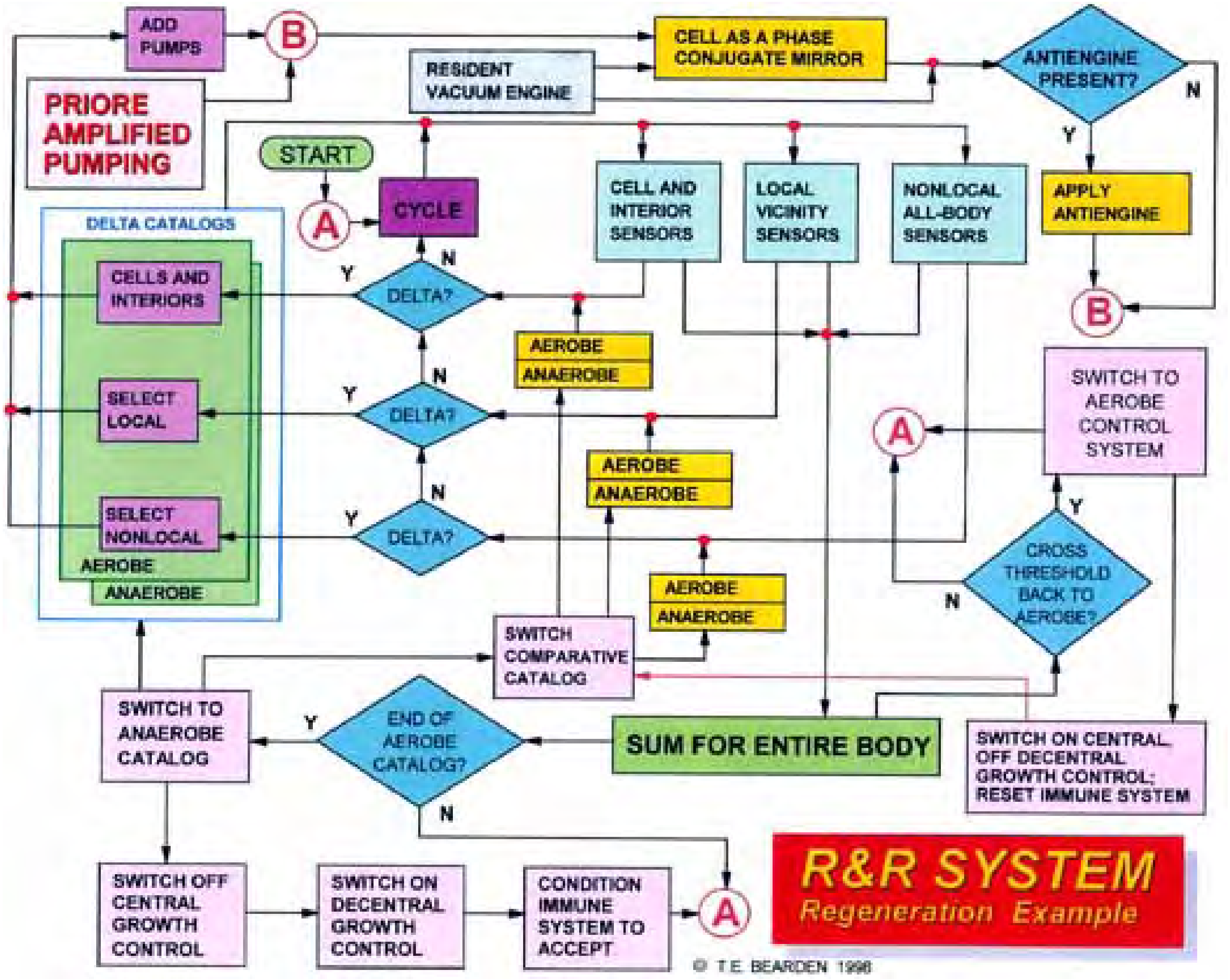
PRIORE AMPLIFIED PUMPING
ADD PUMPS
B
RESIDENT VACUUM ENGINE
CELL AS A PHASE CONJUGATE MIRROR
ANTIENGINE PRESENT?
N
Y
APPLY ANTIENGINE
B
START
A
CYCLE
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
DELTA CATALOGS
CELLS AND INTERIORS
SELECT LOCAL
SELECT NONLOCAL
AEROBE
ANAEROBE
Y
N
DELTA?
Y
N
DELTA?
Y
N
DELTA?
AEROBE
ANAEROBE
AEROBE
ANAEROBE
AEROBE
ANAEROBE
SWITCH TO AEROBE CONTROL SYSTEM
A
Y
CROSS THRESHOLD BACK TO AEROBE?
N
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
Y
END OF AEROBE CATALOG?
N
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT
A
R&R SYSTEM
Regeneration Example
© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

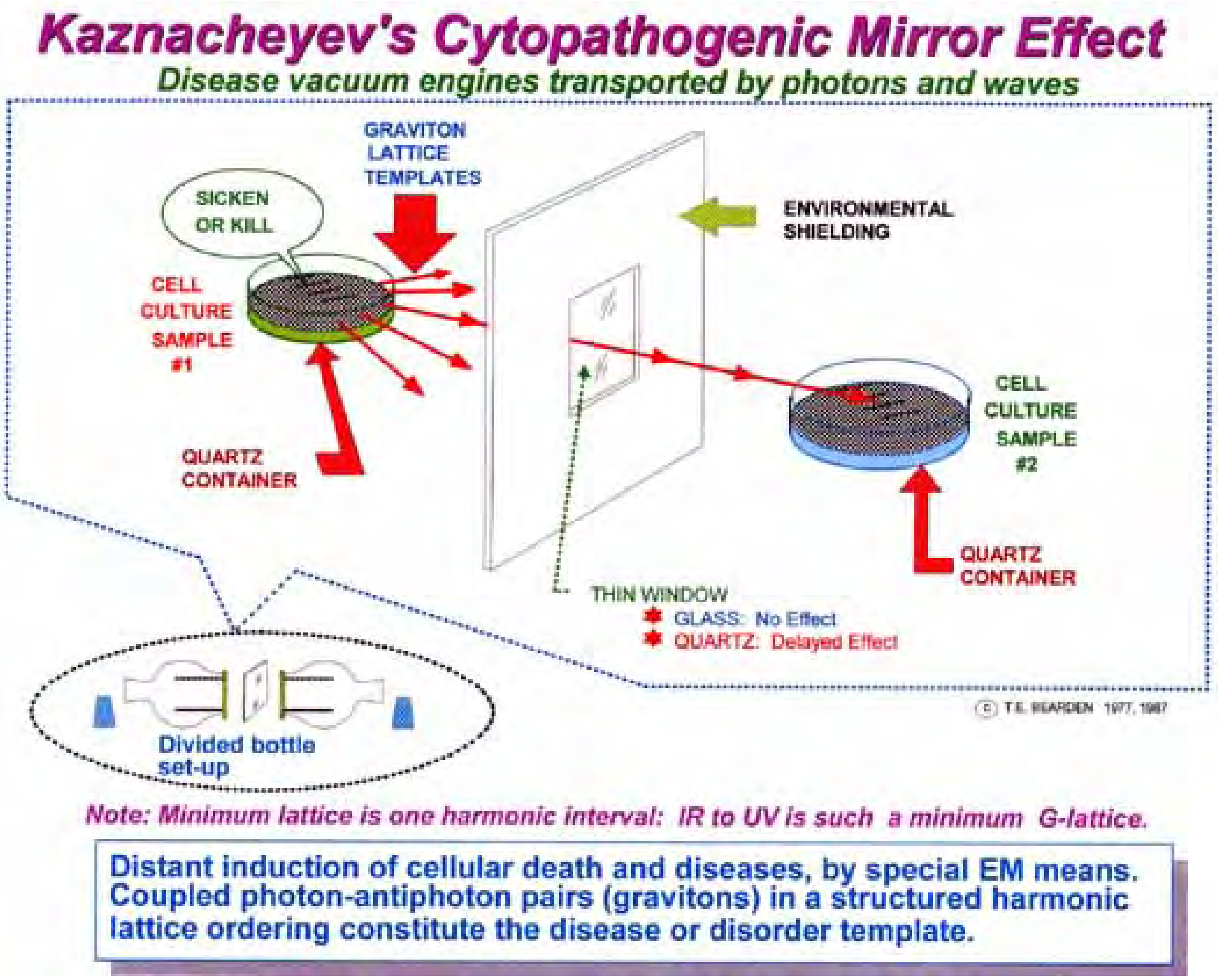

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Return to Slide Index

Next Slide

Previous Slide

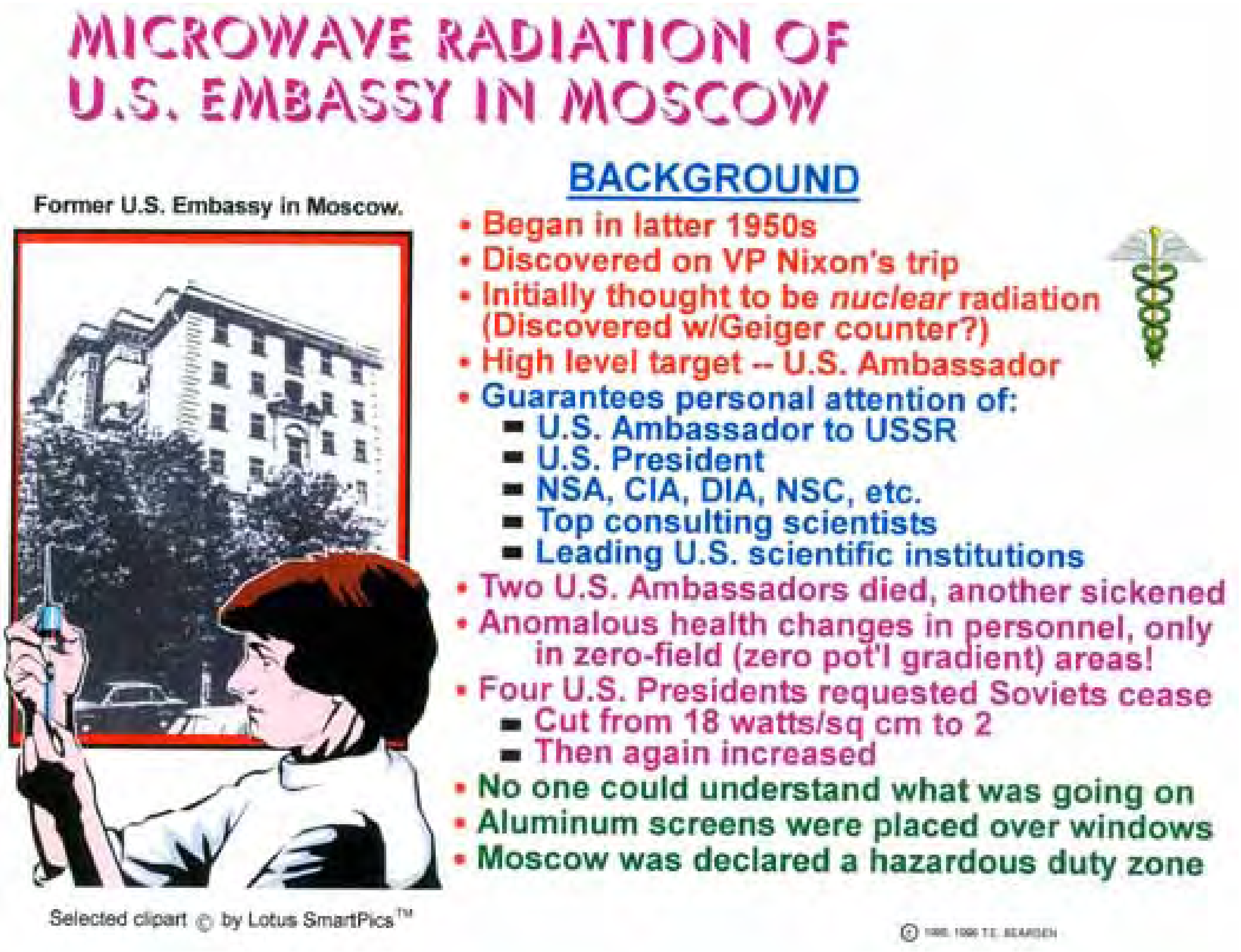

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

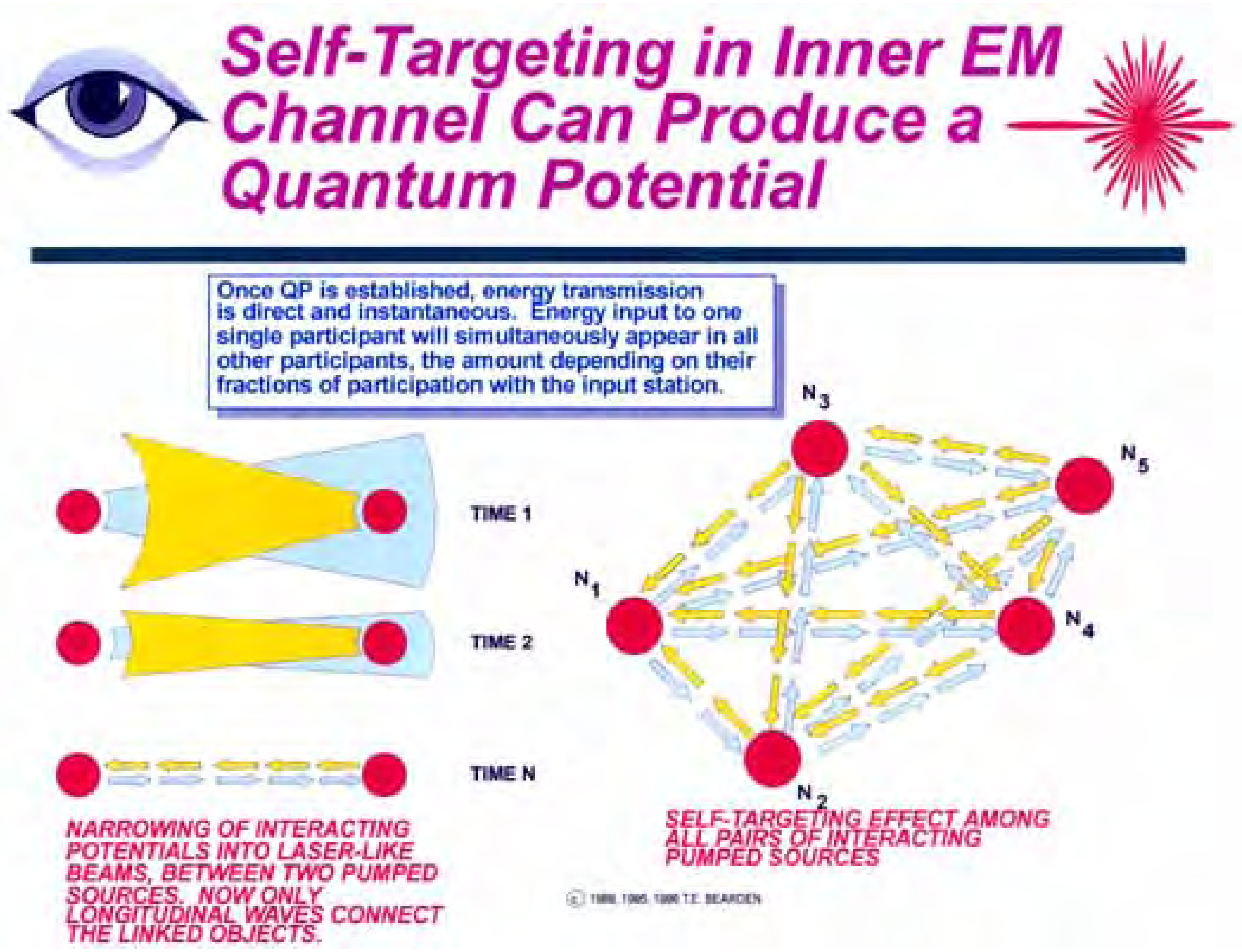

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1988, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

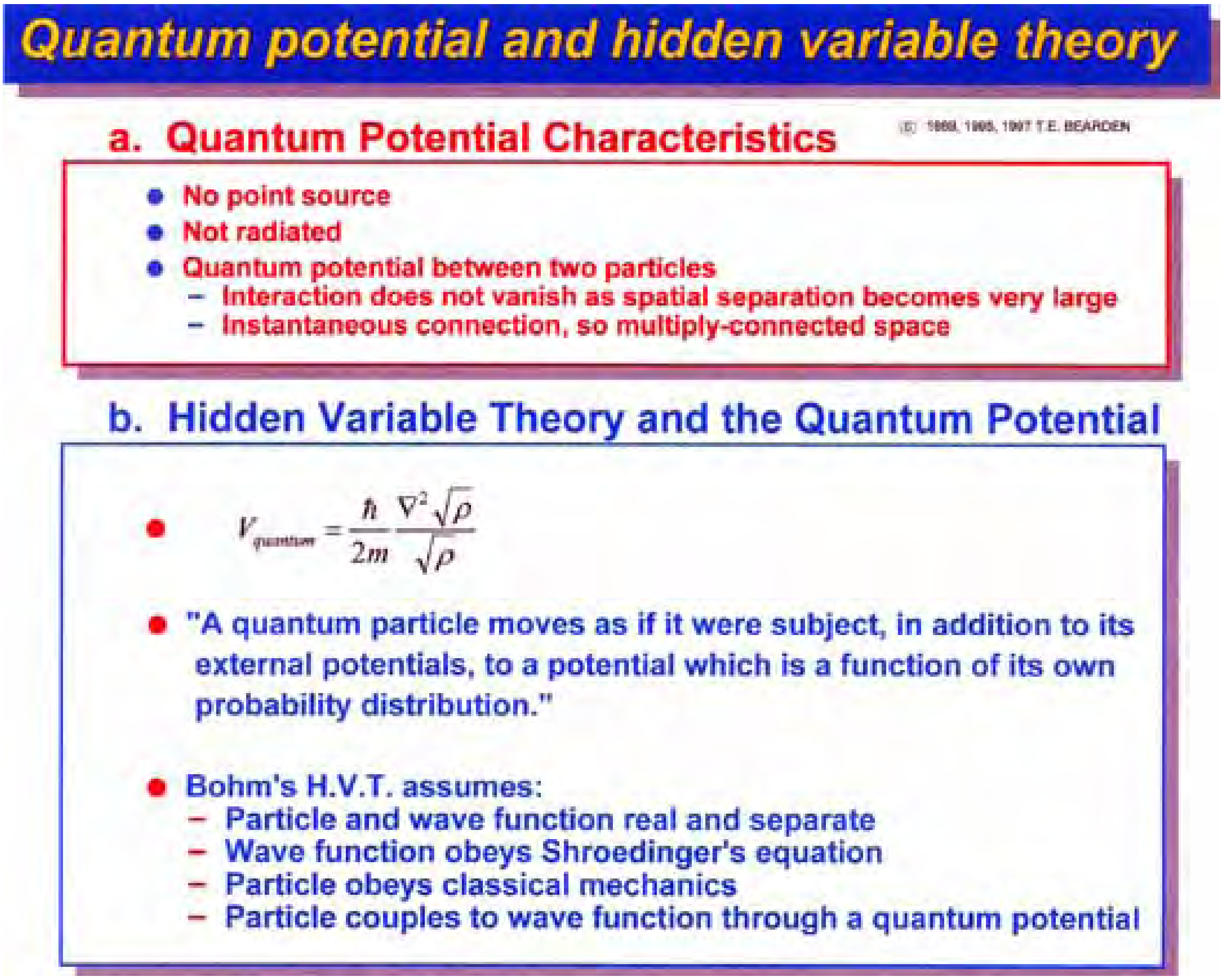

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1989, 1995, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

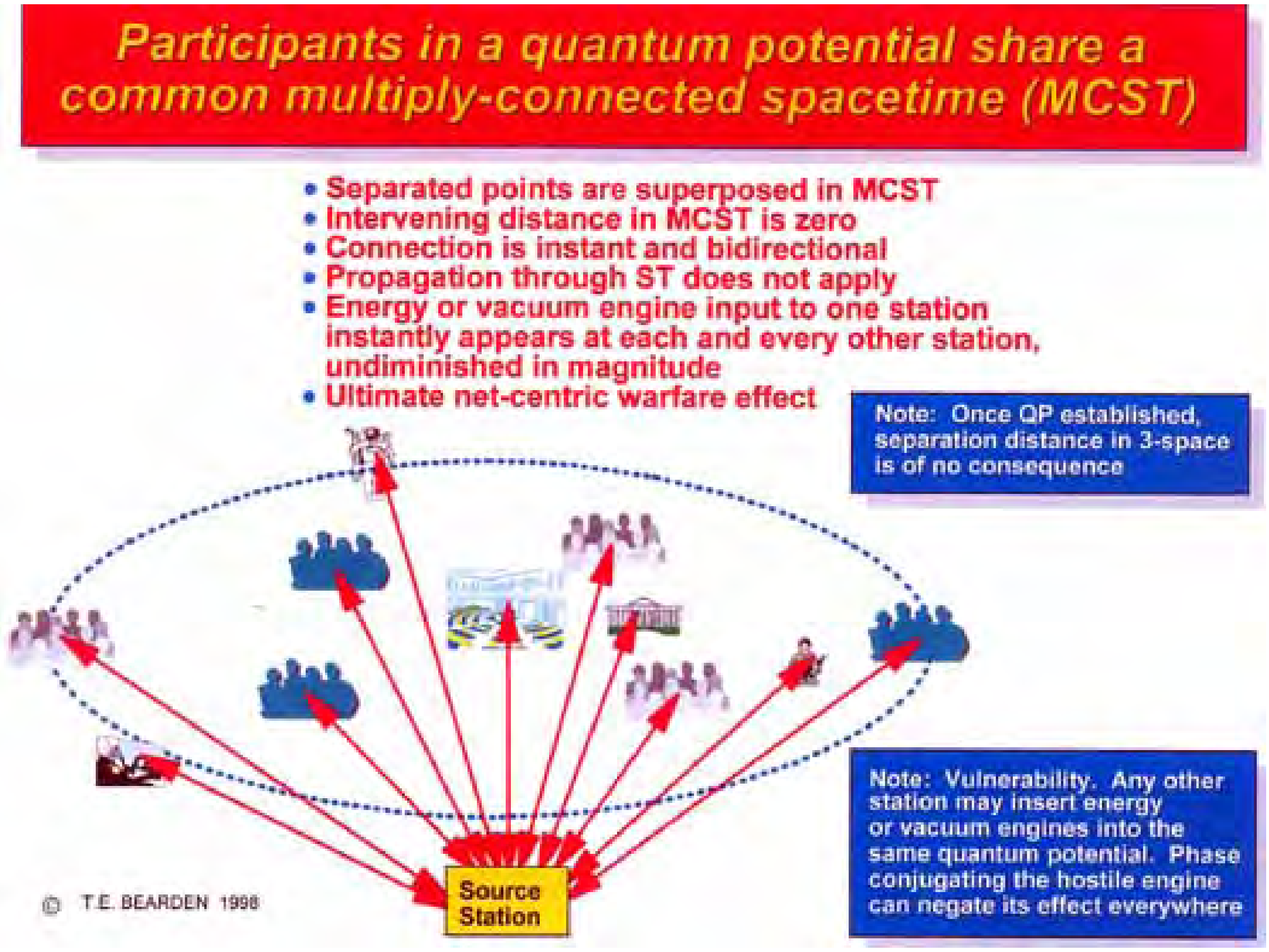

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

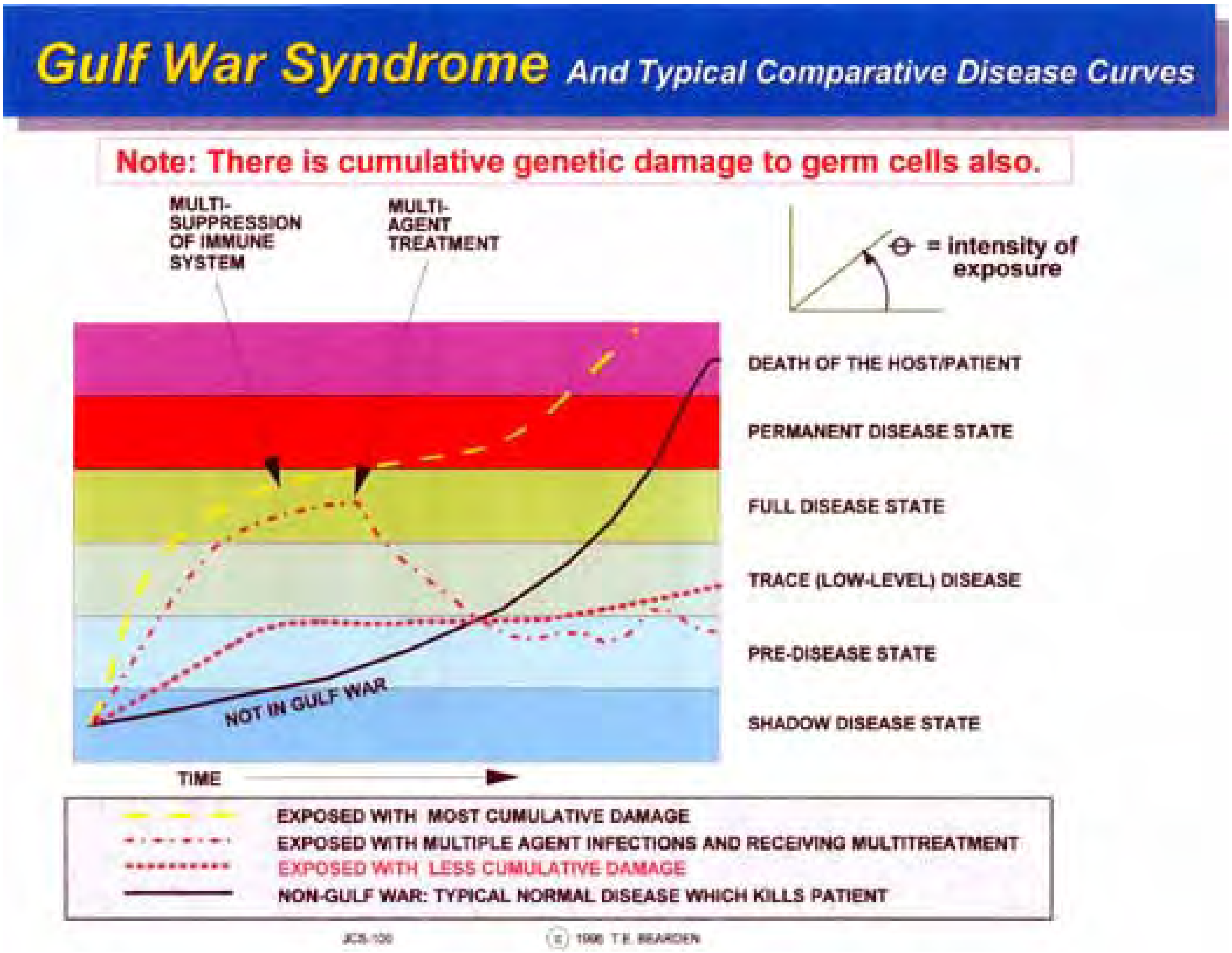

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

MULTI-SUPPRESSION OF IMMUNE SYSTEM

MULTI-AGENT TREATMENT

Θ = intensity of exposure

DEATH OF THE HOST/PATIENT

PERMANENT DISEASE STATE

FULL DISEASE STATE

TRACE (LOW-LEVEL) DISEASE

PRE-DISEASE STATE

SHADOW DISEASE STATE

NOT IN GULF WAR

TIME

EXPOSED WITH MOST CUMULATIVE DAMAGE

EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT

EXPOSED WITH LESS CUMULATIVE DAMAGE

NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

(c) 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

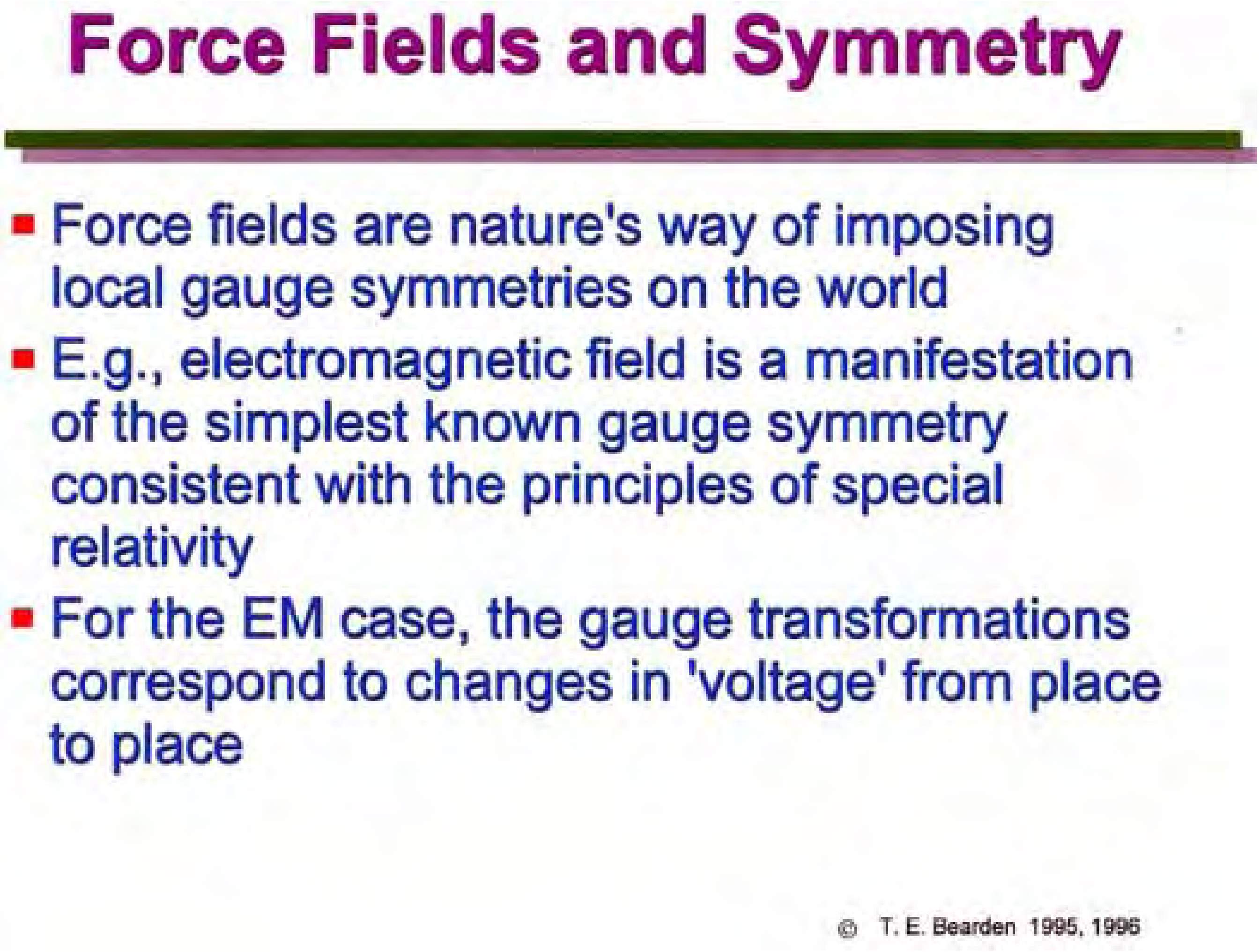

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

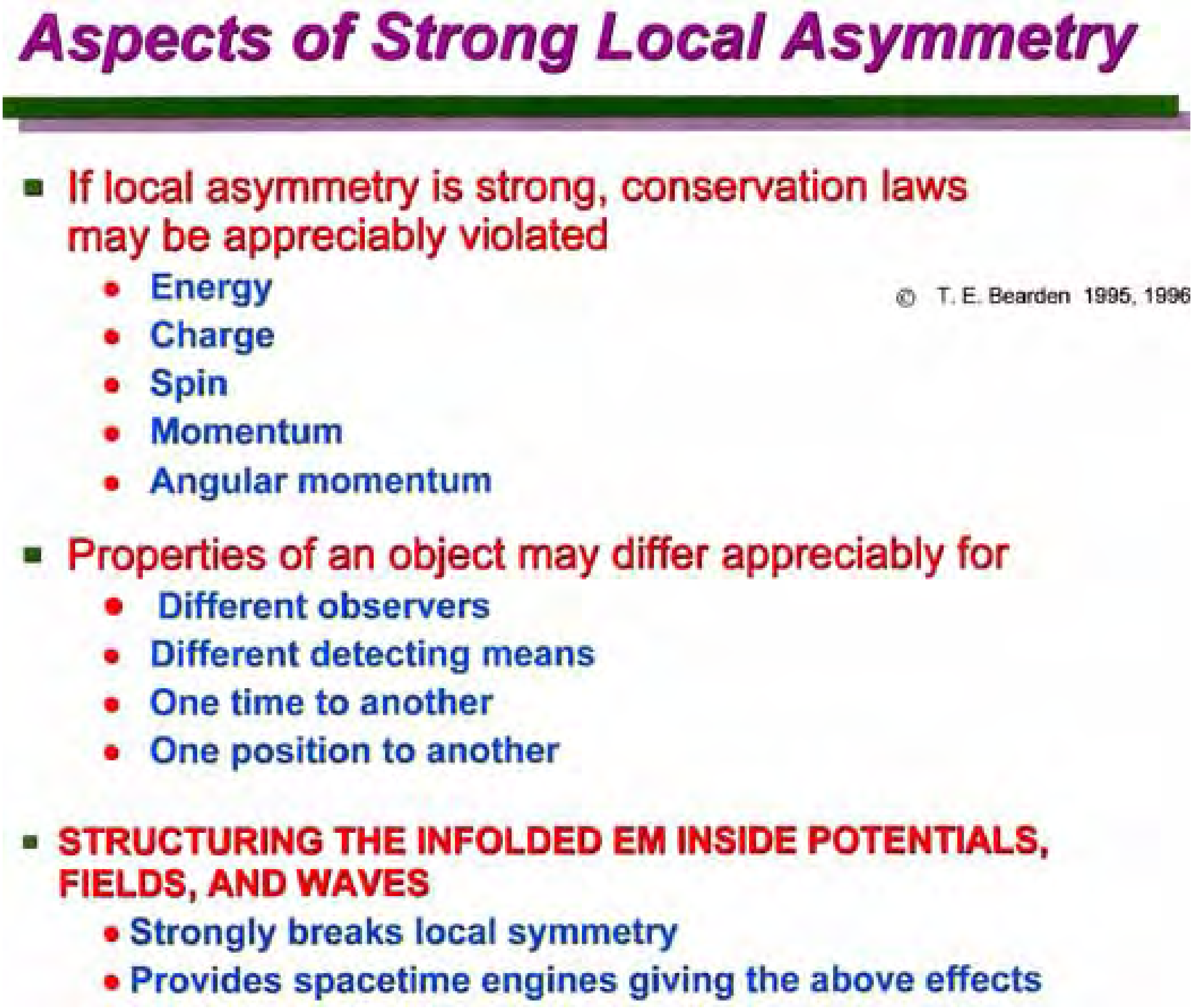

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

Return to Slide Index

Next Slide

Previous Slide

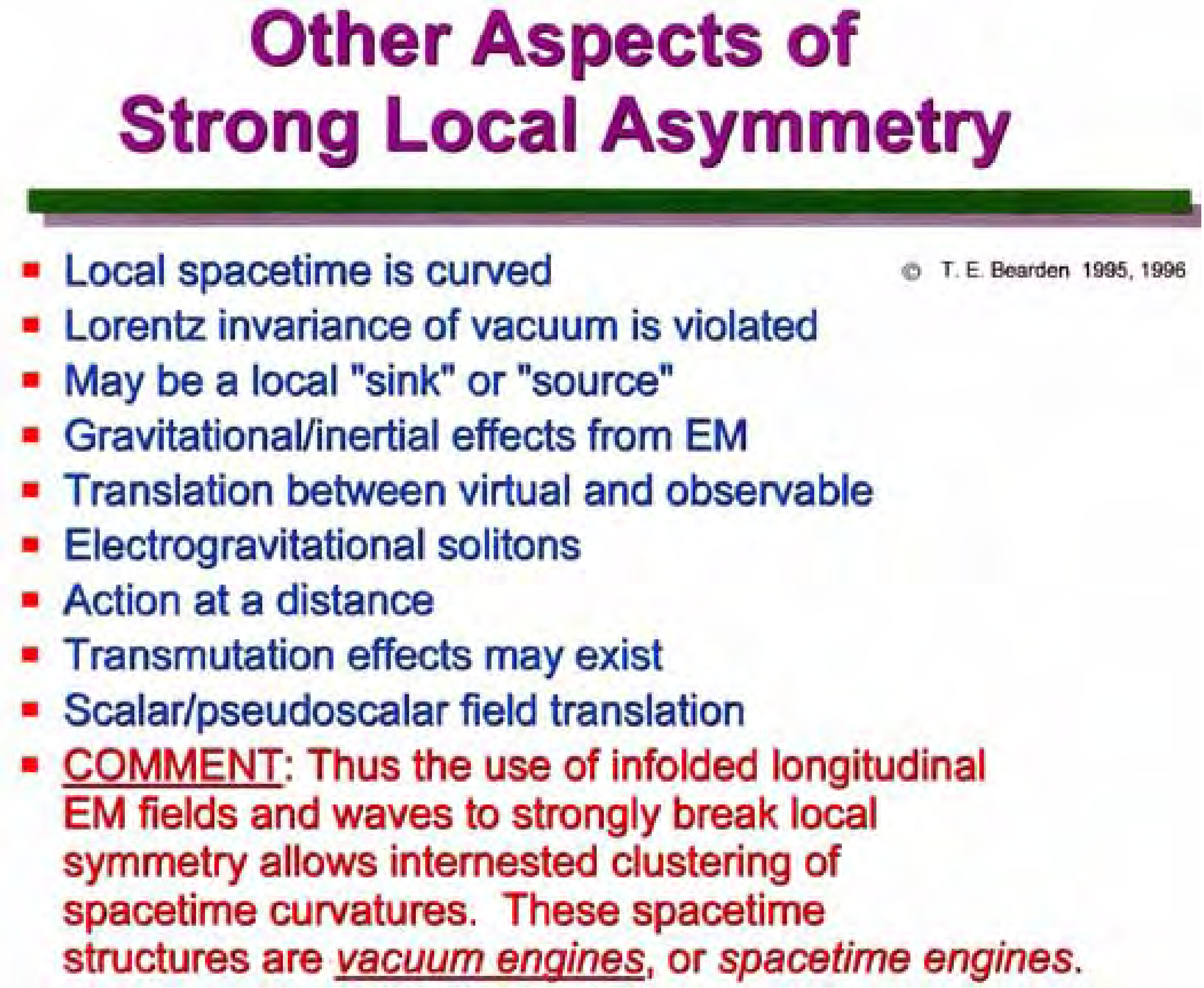

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

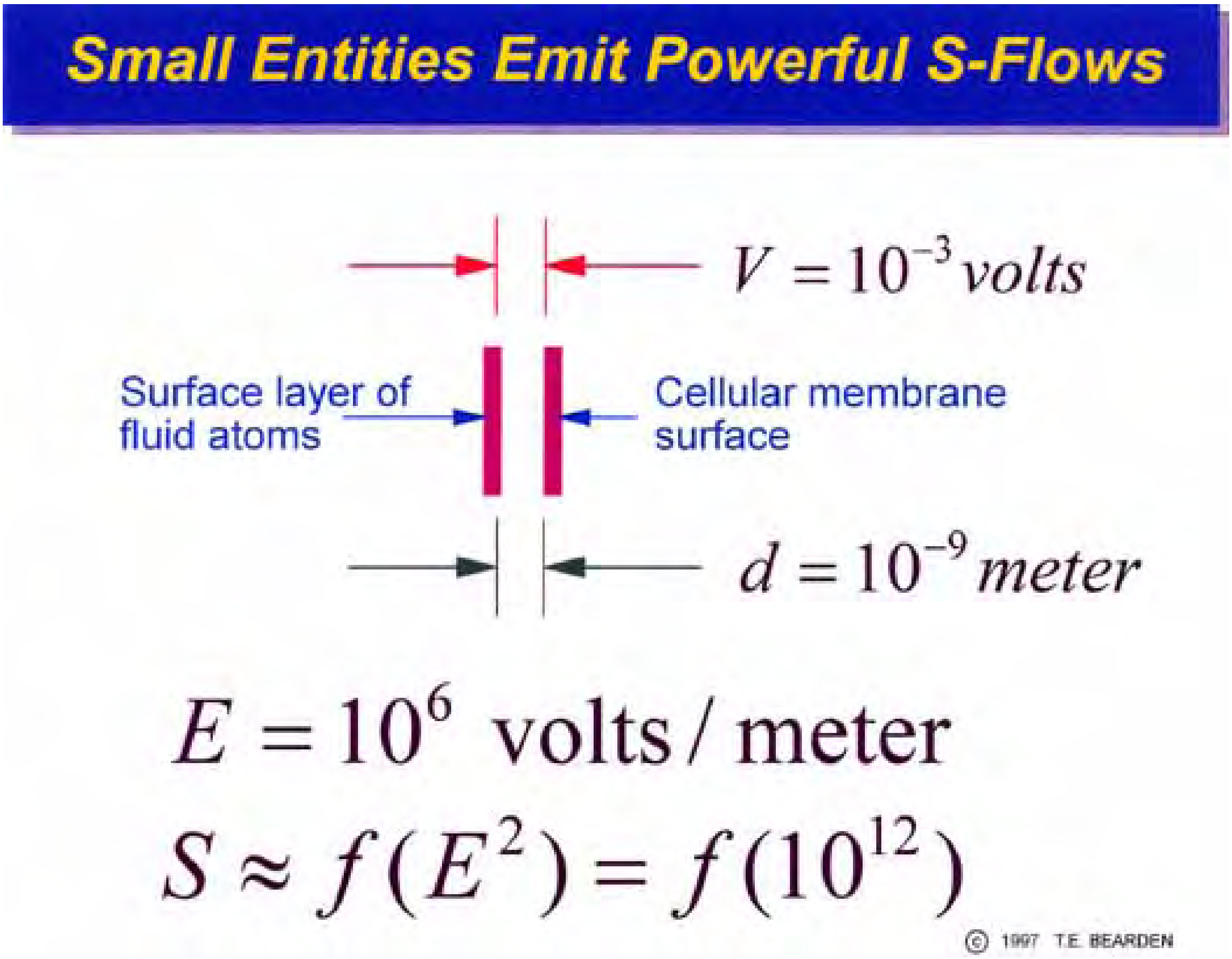

Small Entities Emit Powerful S-Flows

$$V = 10^{-3}\,volts$$

Surface layer of fluid atoms

Cellular membrane surface

$$d = 10^{-9}\,meter$$

$$E = 10^{6}\ \text{volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

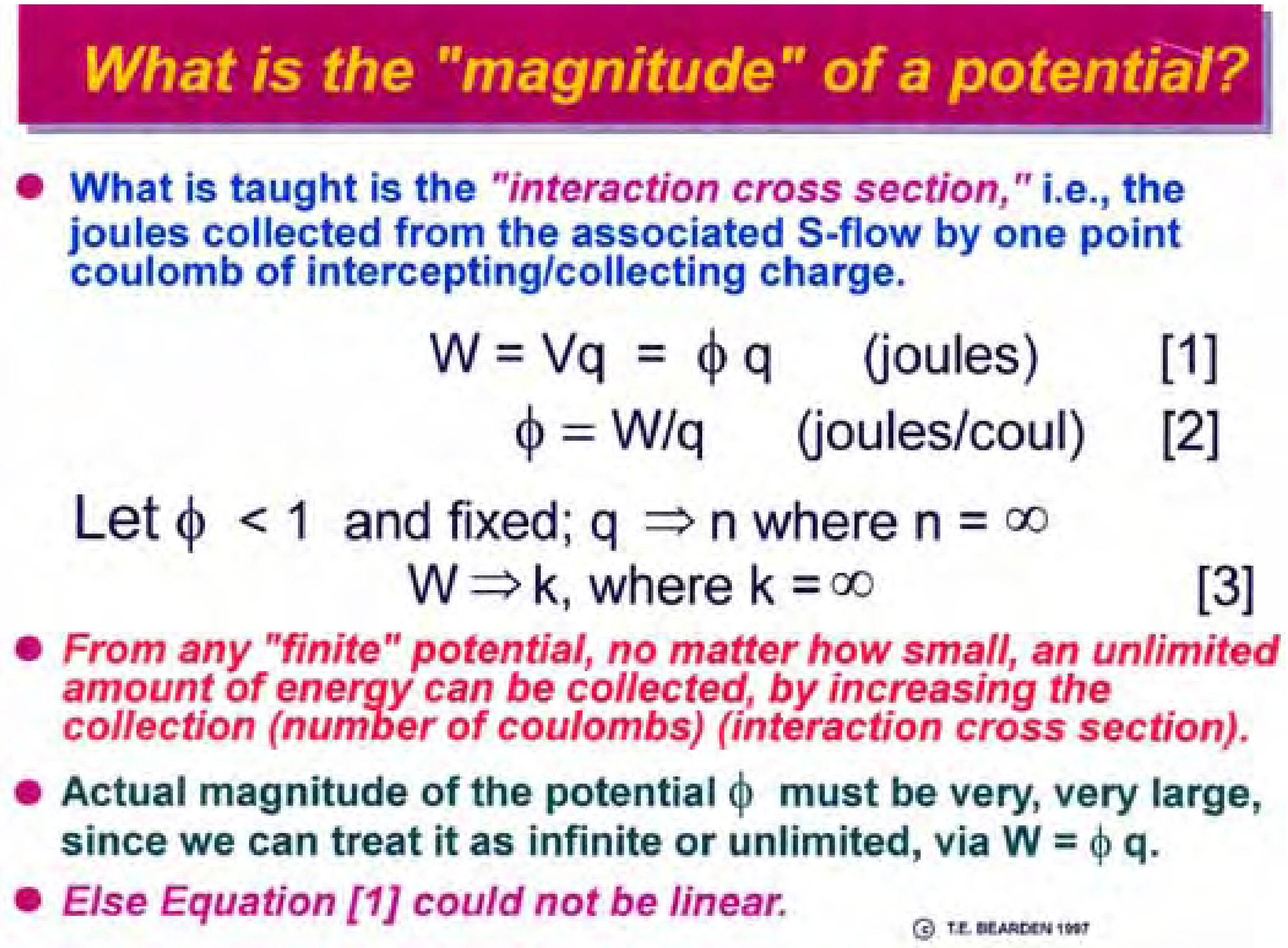

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \quad \text{(joules)} \qquad [1]$$

$$\phi = W/q \quad \text{(joules/coul)} \qquad [2]$$

Let $\phi < 1$ and fixed; $q \Rightarrow n$ where $n = \infty$

$$W \Rightarrow k, \text{ where } k = \infty \qquad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

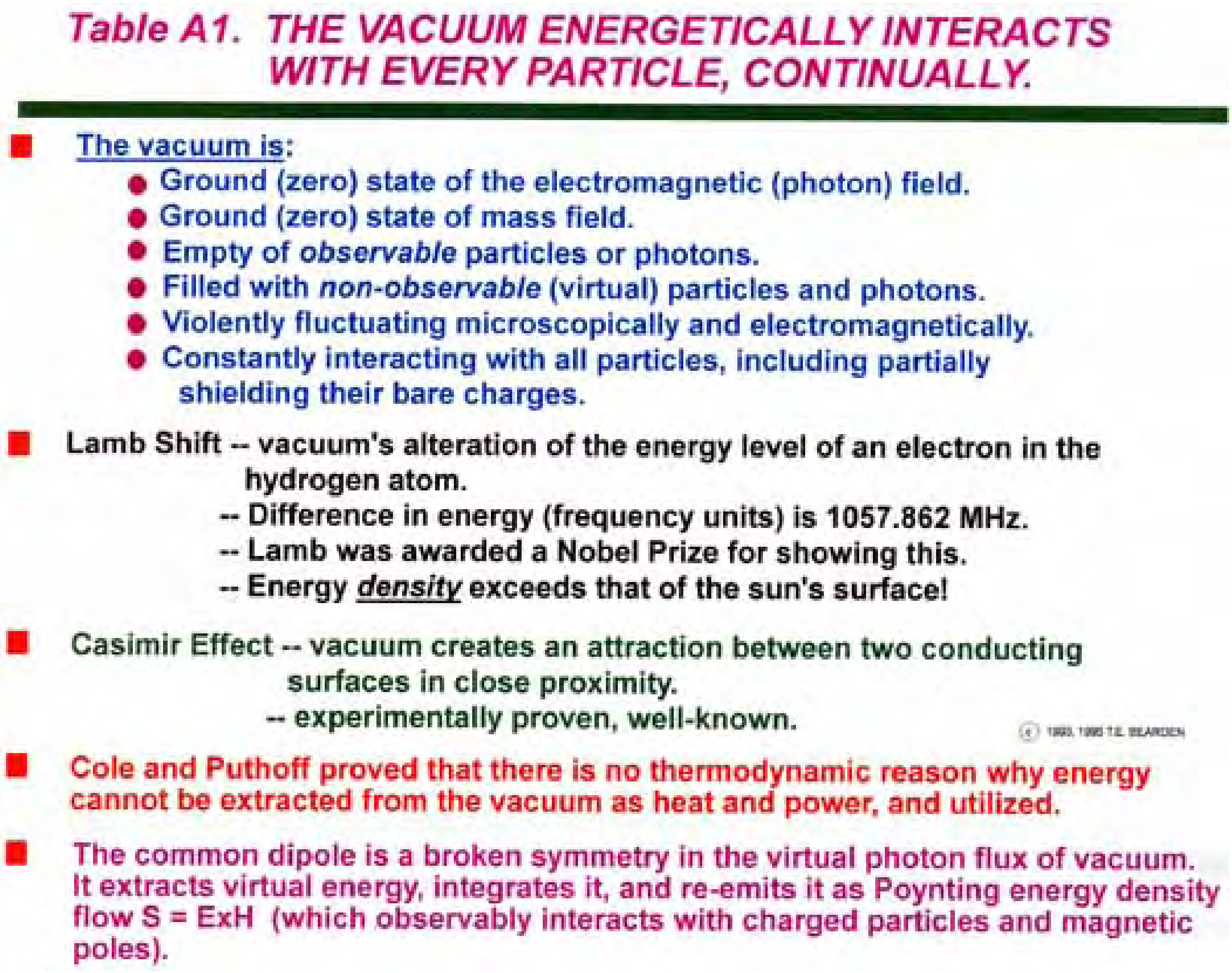

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy ***density*** exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.
- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

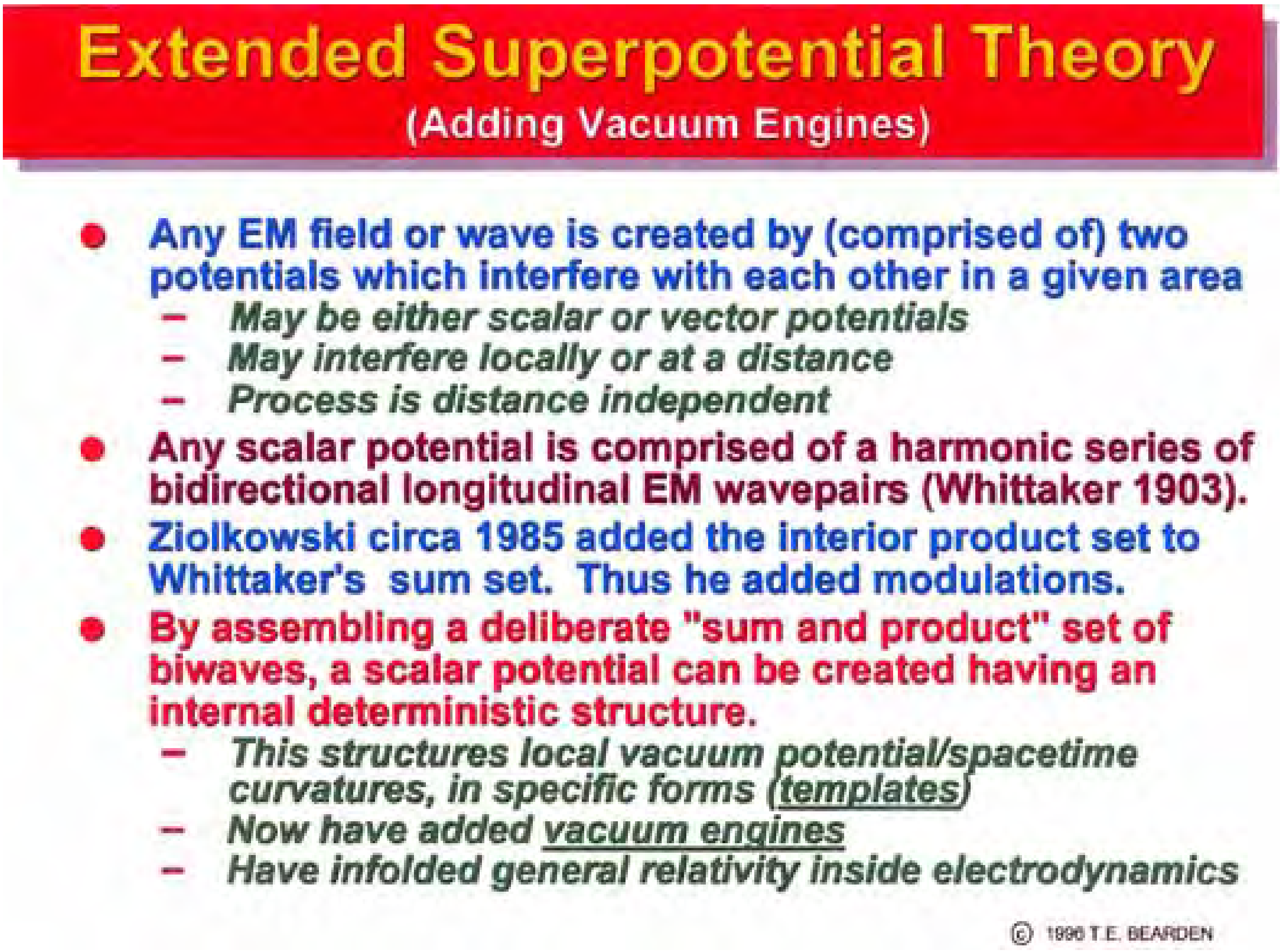

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

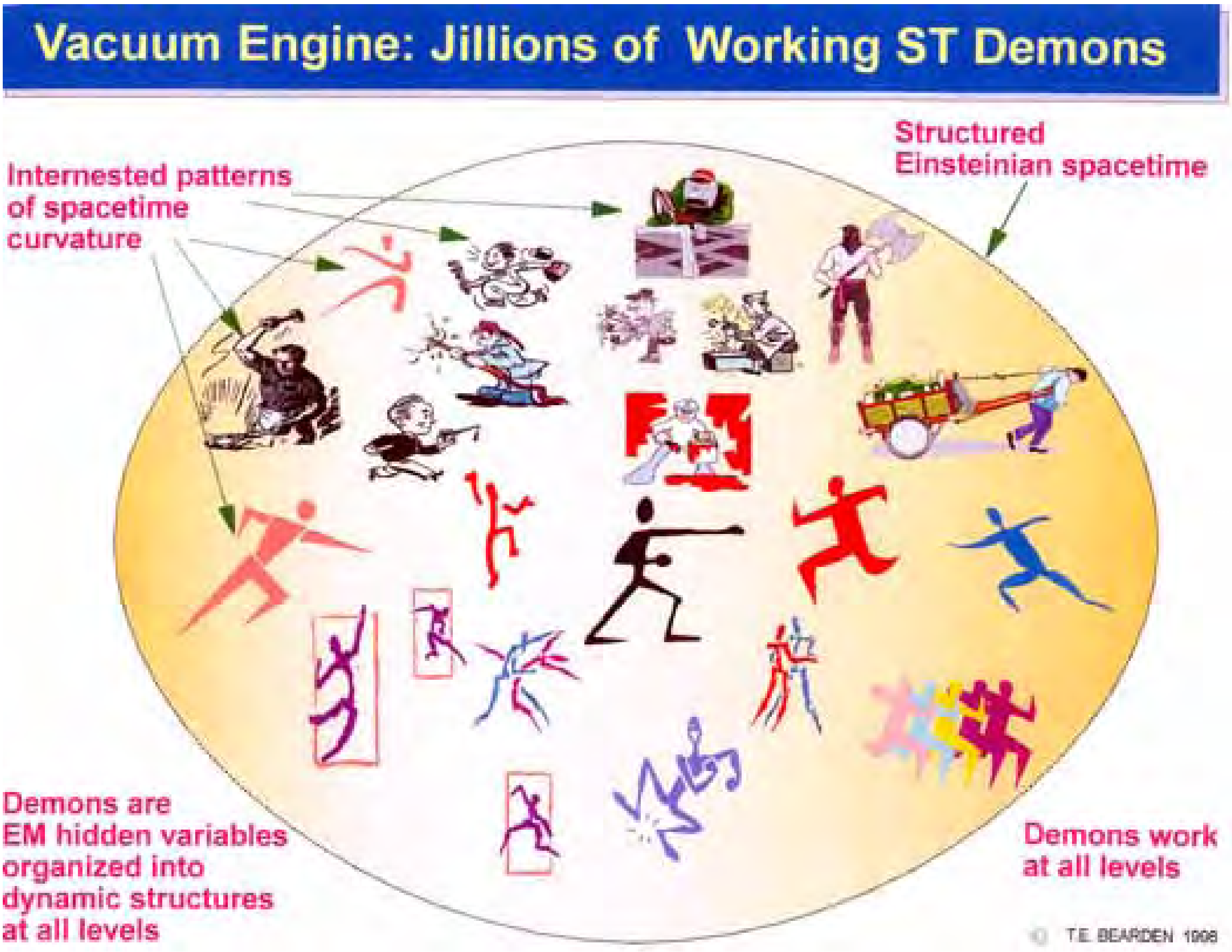

Vacuum Engine: Jillions of Working ST Demons

Internested patterns of spacetime curvature

Structured Einsteinian spacetime

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

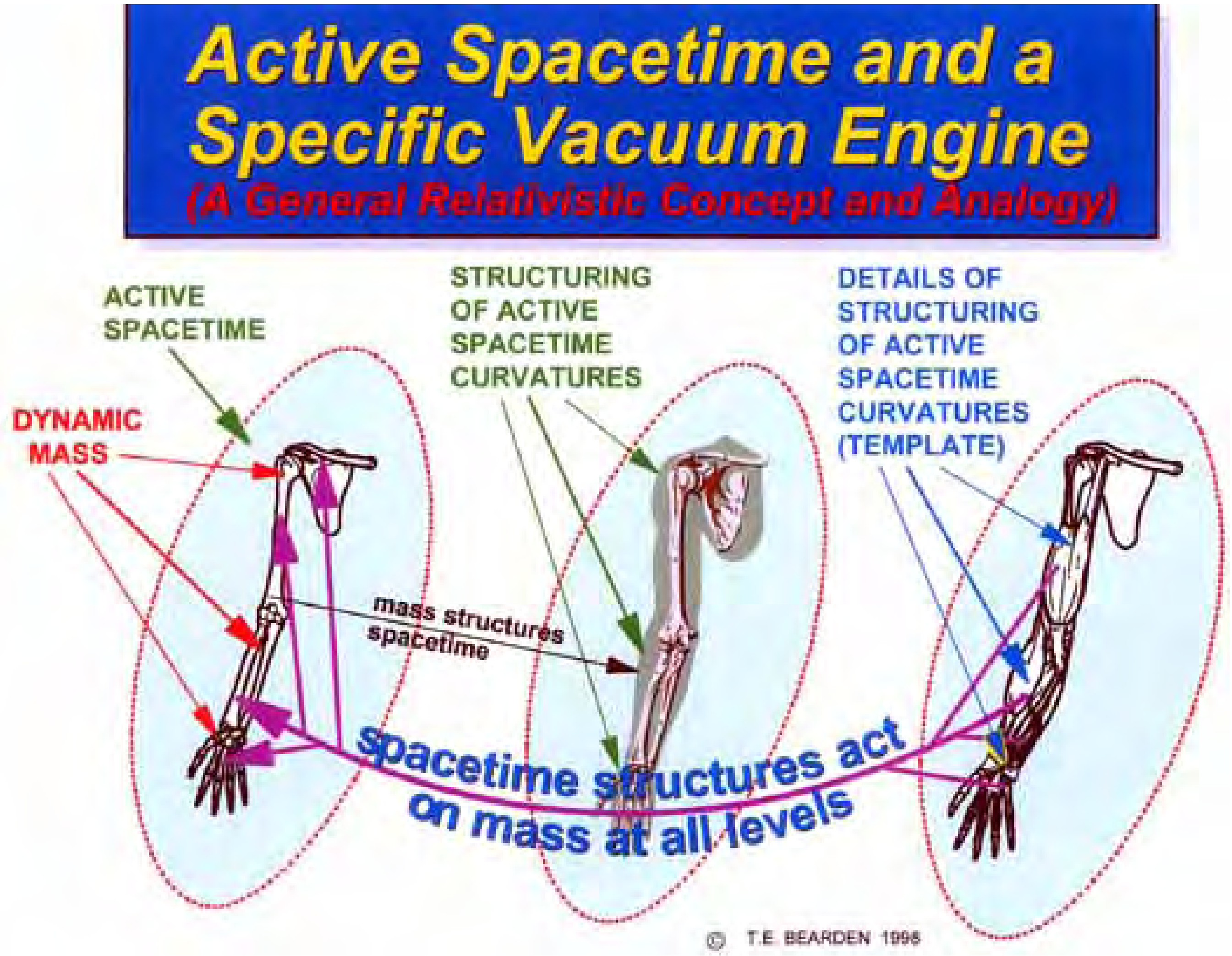

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)
mass structures spacetime
spacetime structures act on mass at all levels

© T.E. BEARDEN 1998

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

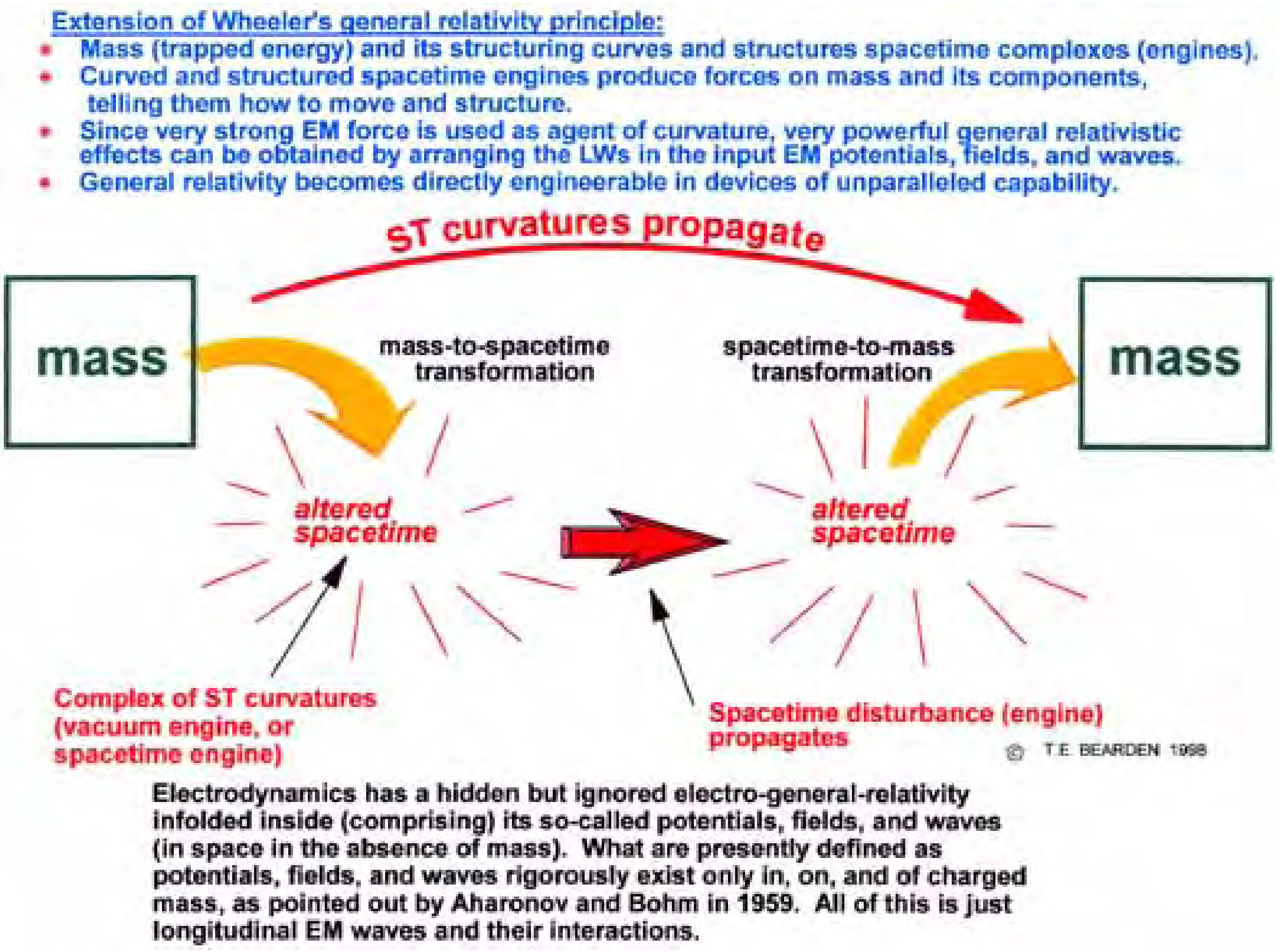

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

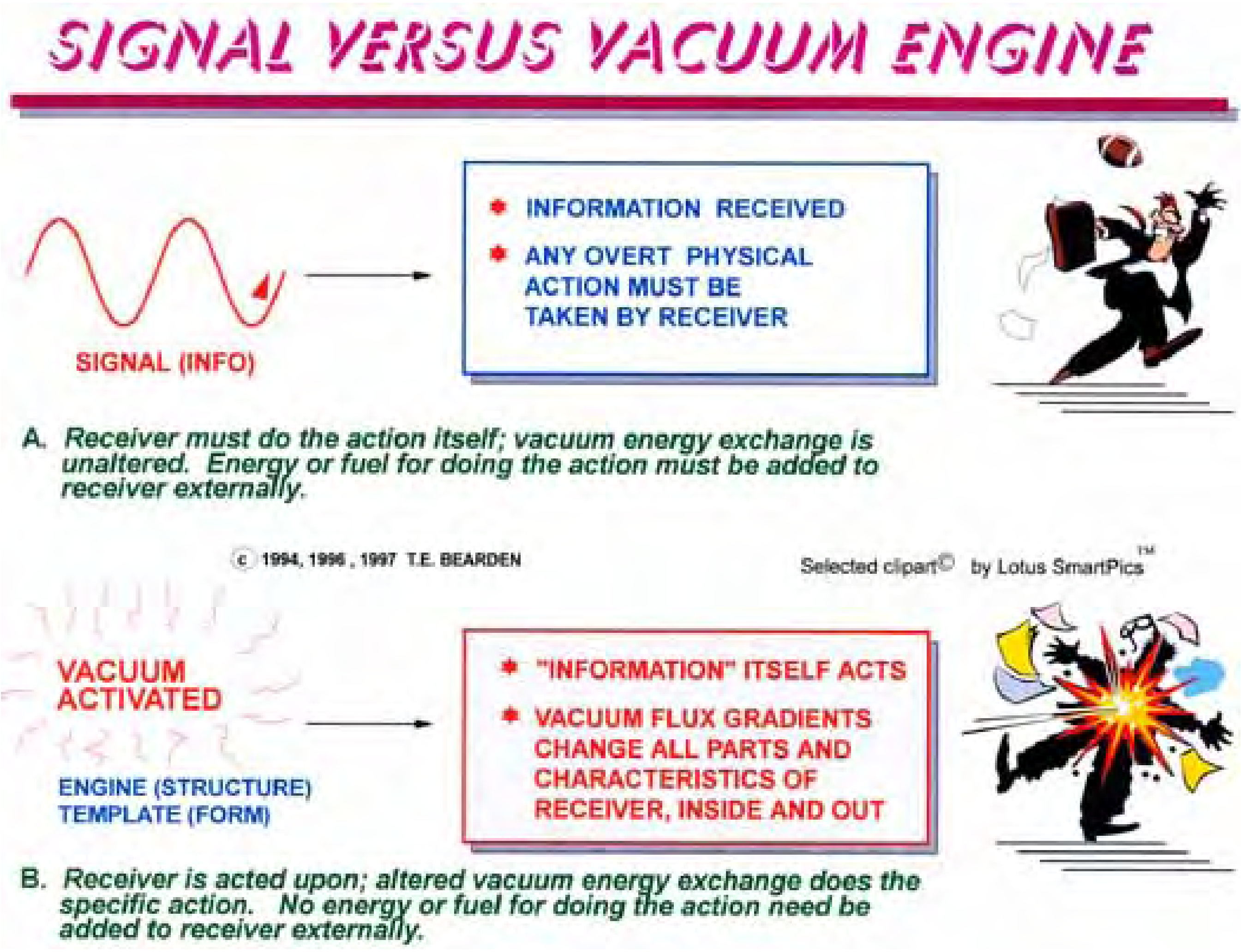

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

(c) 1994, 1996 , 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED

ENGINE (STRUCTURE)
TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

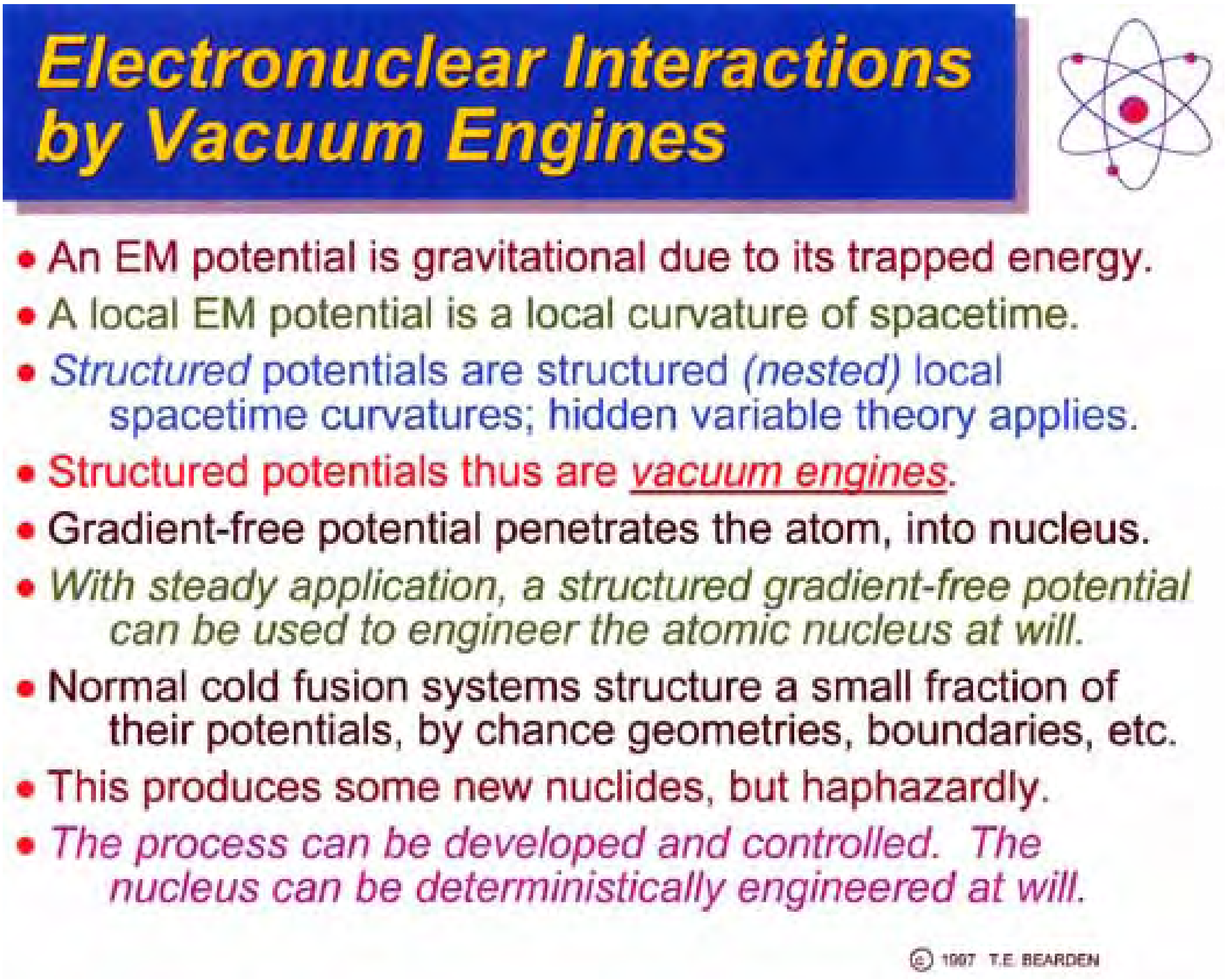

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

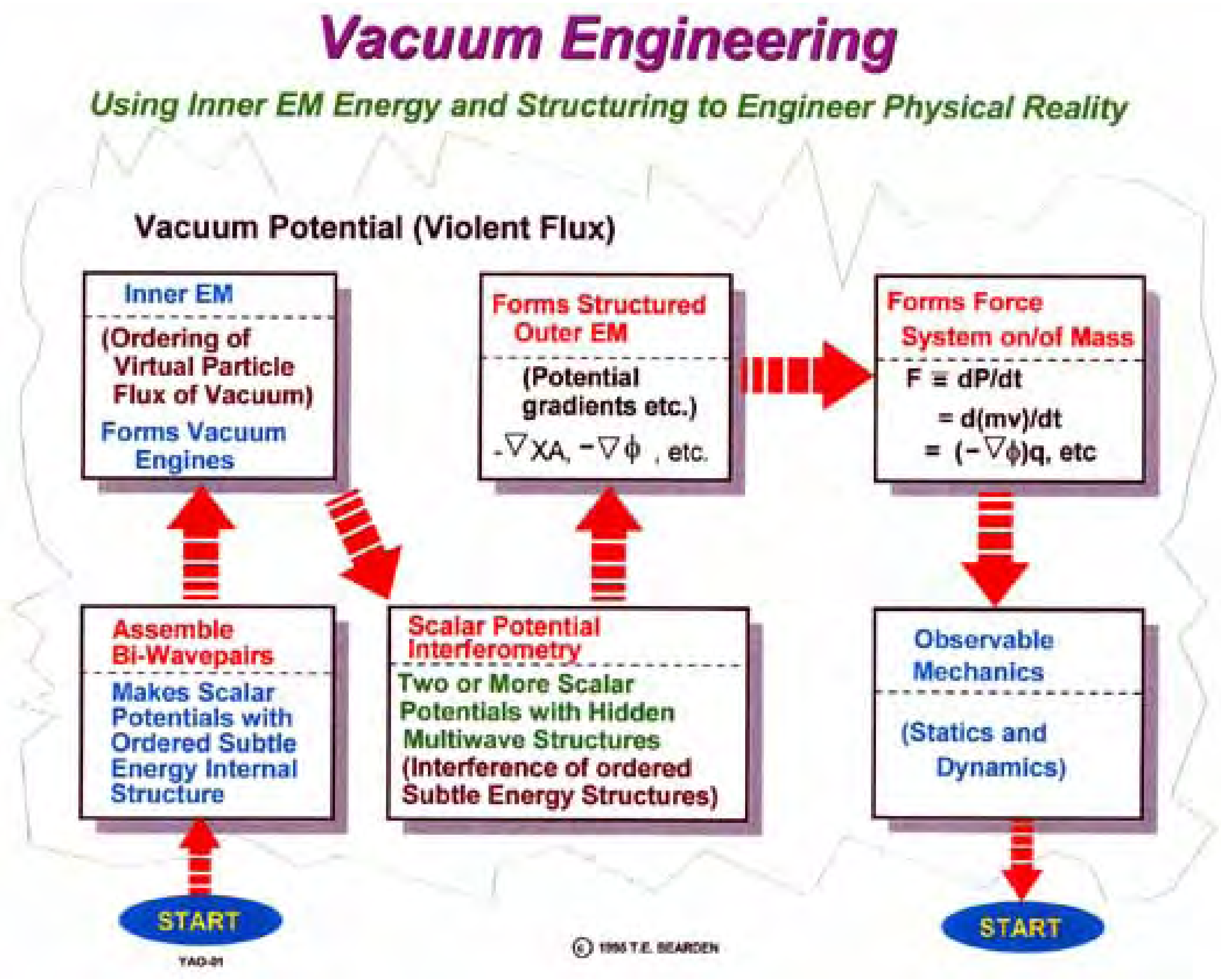

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM

(Ordering of Virtual Particle Flux of Vacuum)

Forms Vacuum Engines

Forms Structured Outer EM

(Potential gradients etc.)

$-\nabla \times A$, $-\nabla \phi$, etc.

Forms Force System on/of Mass

$F \equiv dP/dt$

$= d(mv)/dt$

$\equiv (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs

Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry

Two or More Scalar Potentials with Hidden Multiwave Structures (Interference of ordered Subtle Energy Structures)

Observable Mechanics

(Statics and Dynamics)

START

START

YAO-01

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

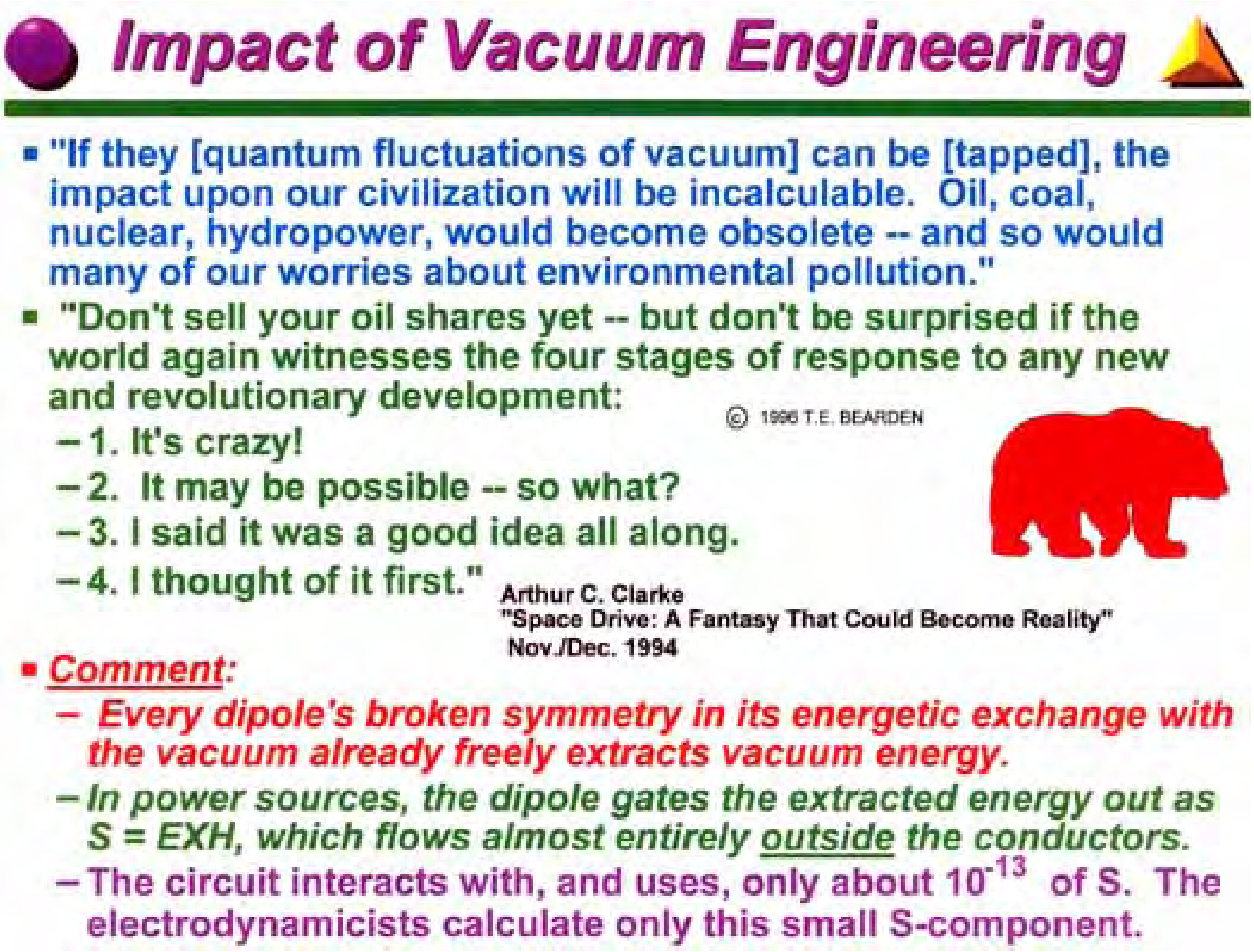

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

© 1996 T.E. BEARDEN

- Comment:
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

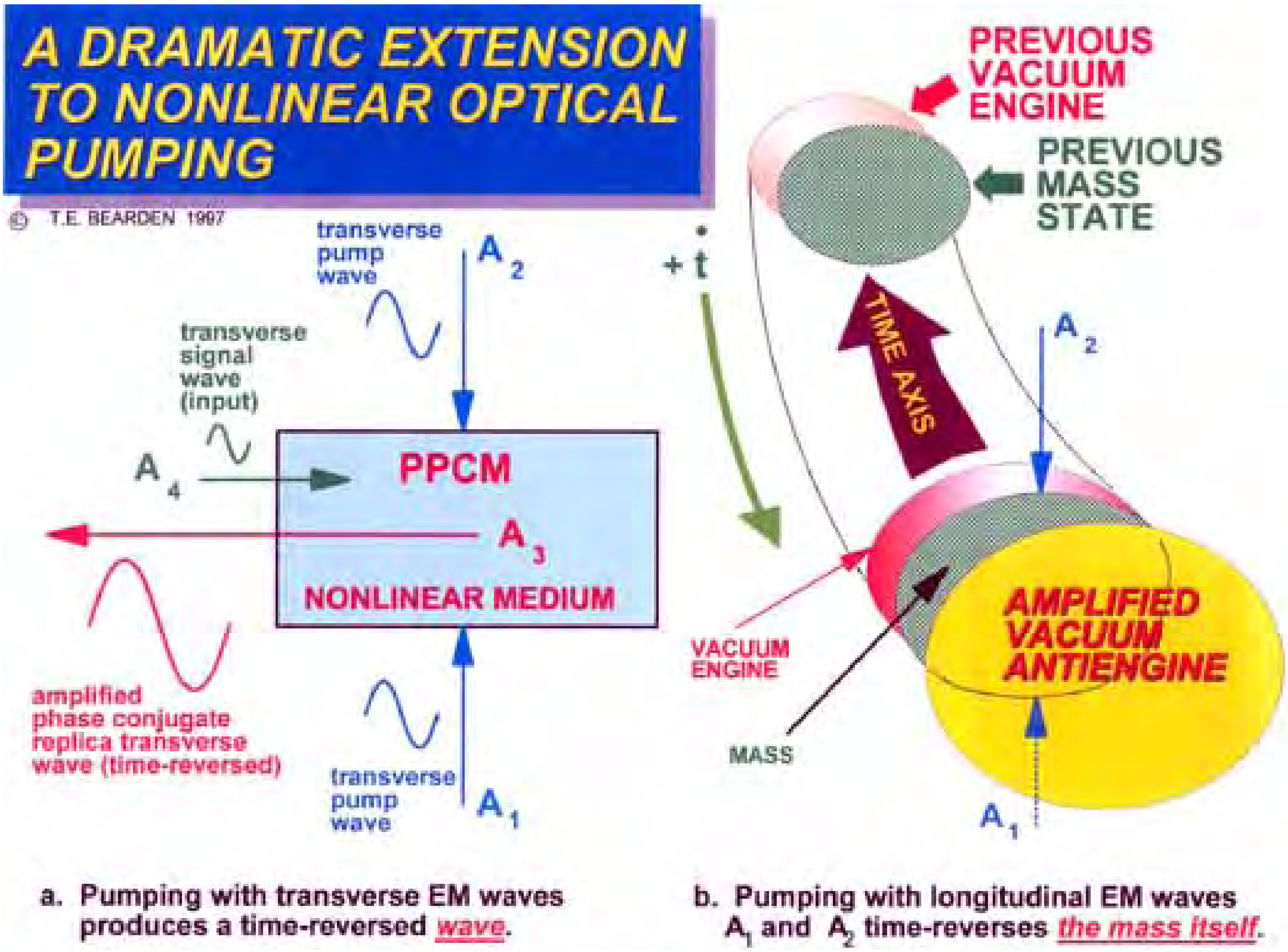

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

transverse pump wave
A_2
transverse signal wave (input)
A_4
PPCM
A_3
NONLINEAR MEDIUM
amplified phase conjugate replica transverse wave (time-reversed)
transverse pump wave
A_1
+ t
PREVIOUS VACUUM ENGINE
PREVIOUS MASS STATE
TIME AXIS
A_2
VACUUM ENGINE
AMPLIFIED VACUUM ANTIENGINE
MASS
A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Return to Slide Index

Next Slide

Previous Slide

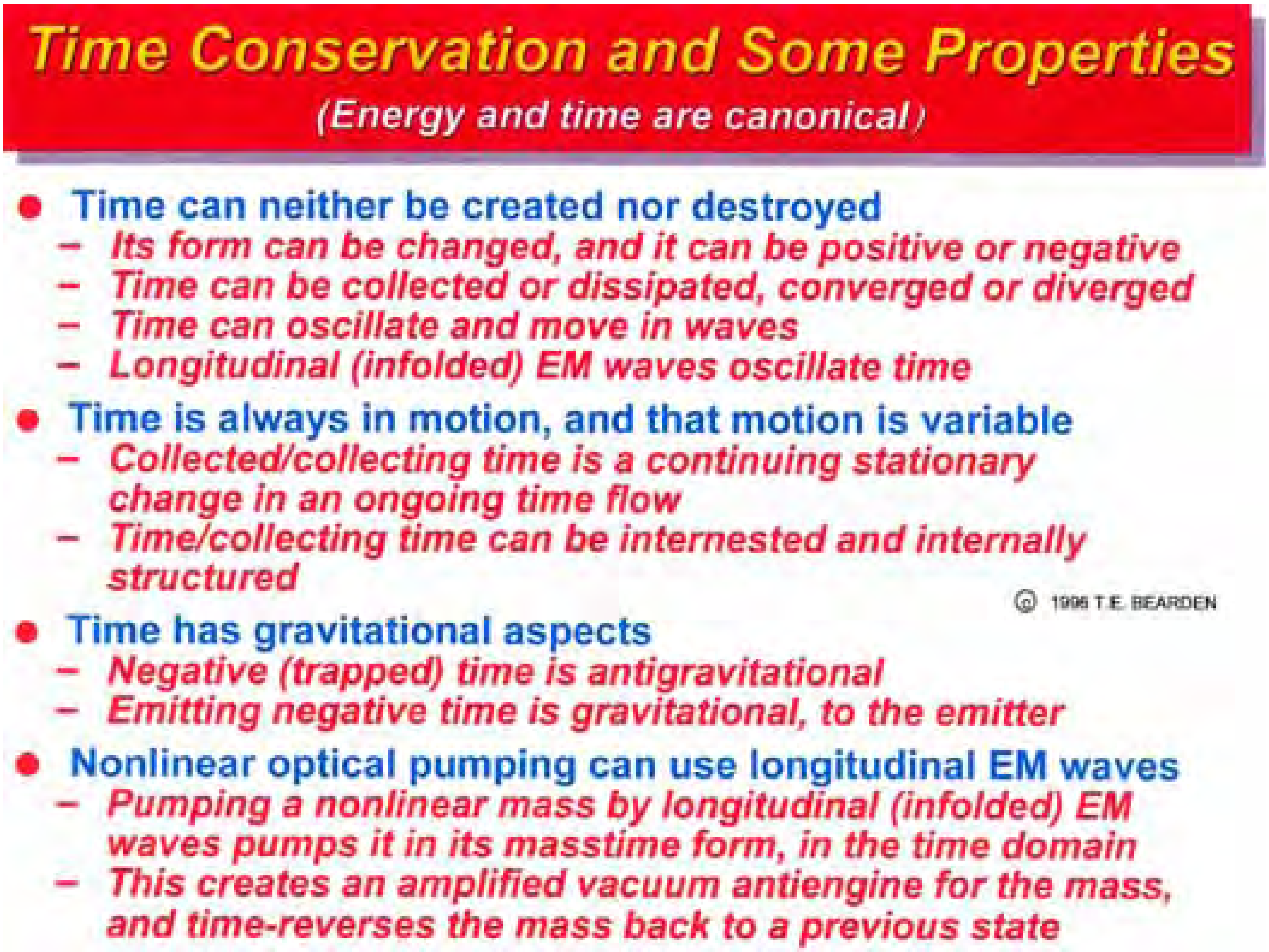

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*

© 1996 T.E. BEARDEN

- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

Return to Slide Index

Next Slide

Previous Slide

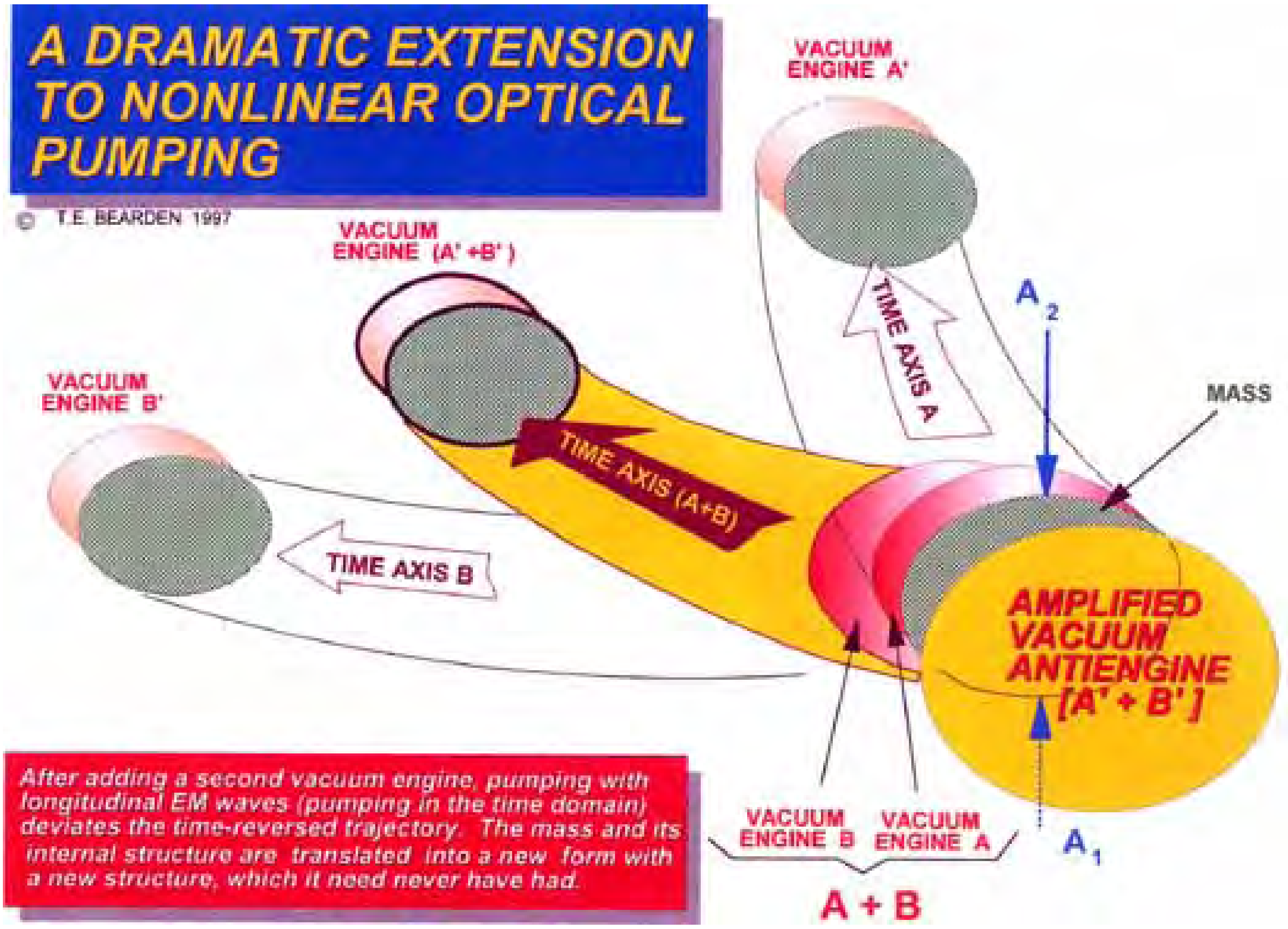

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A_1
A + B

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Return to Slide Index

Next Slide

Previous Slide

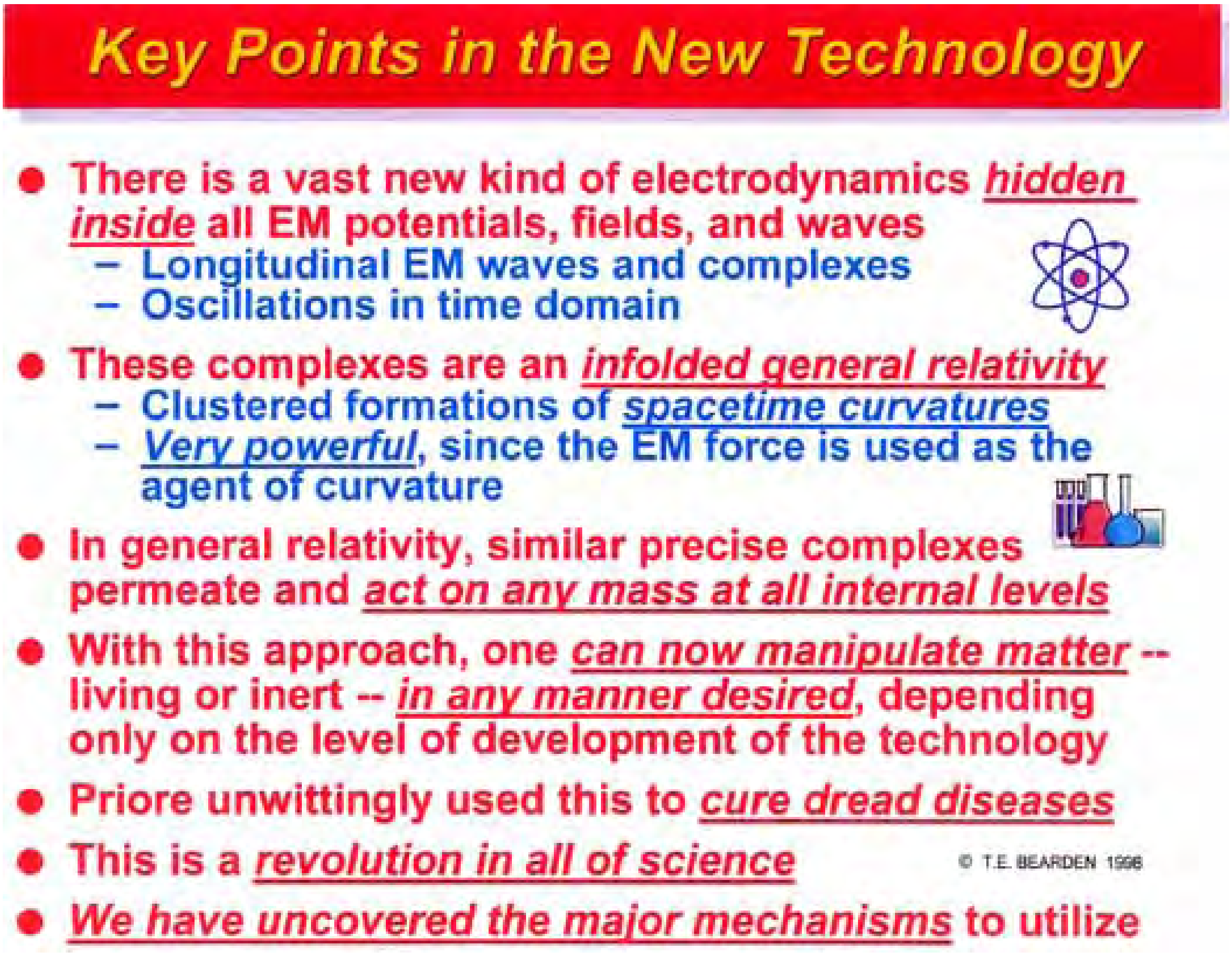

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

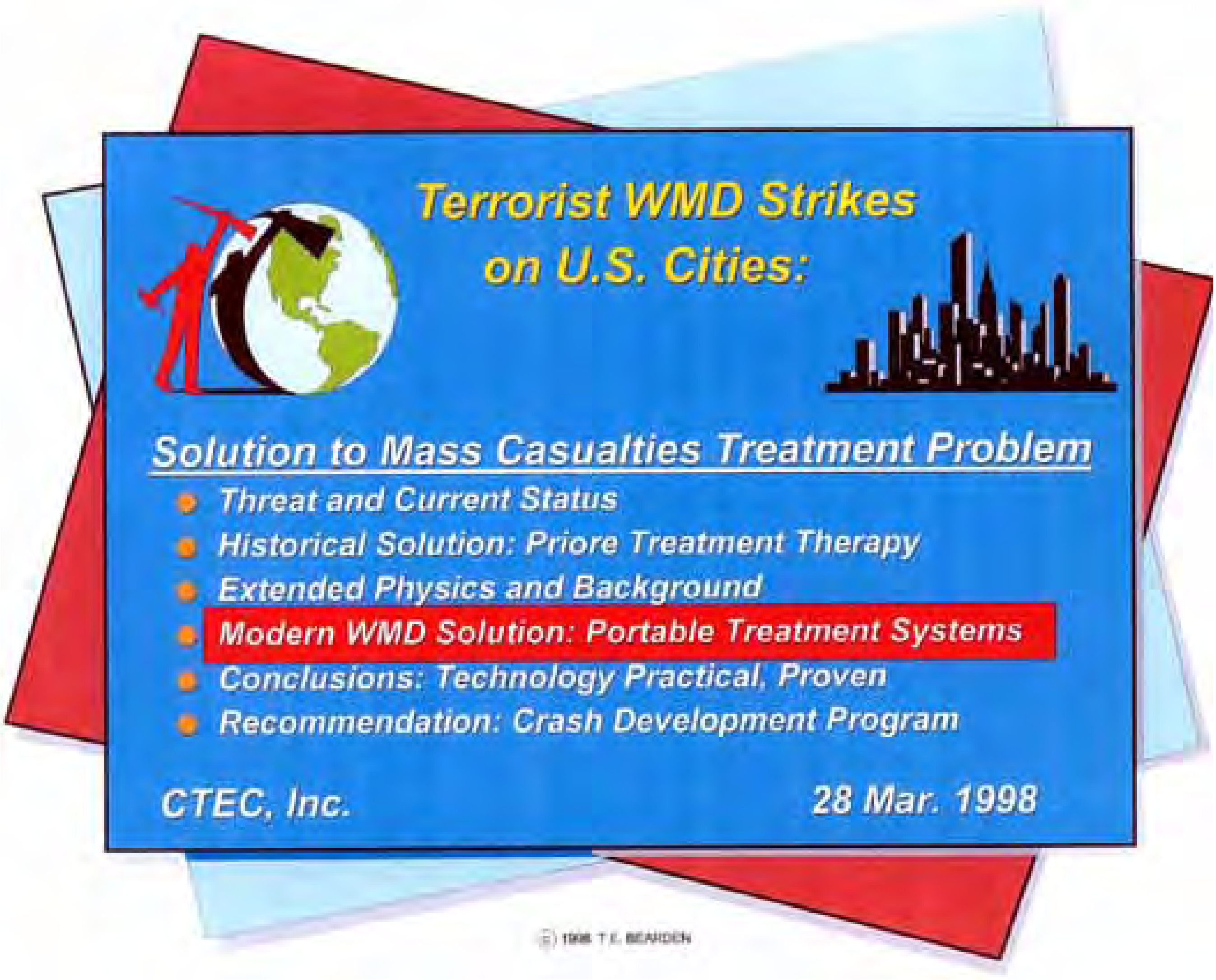

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

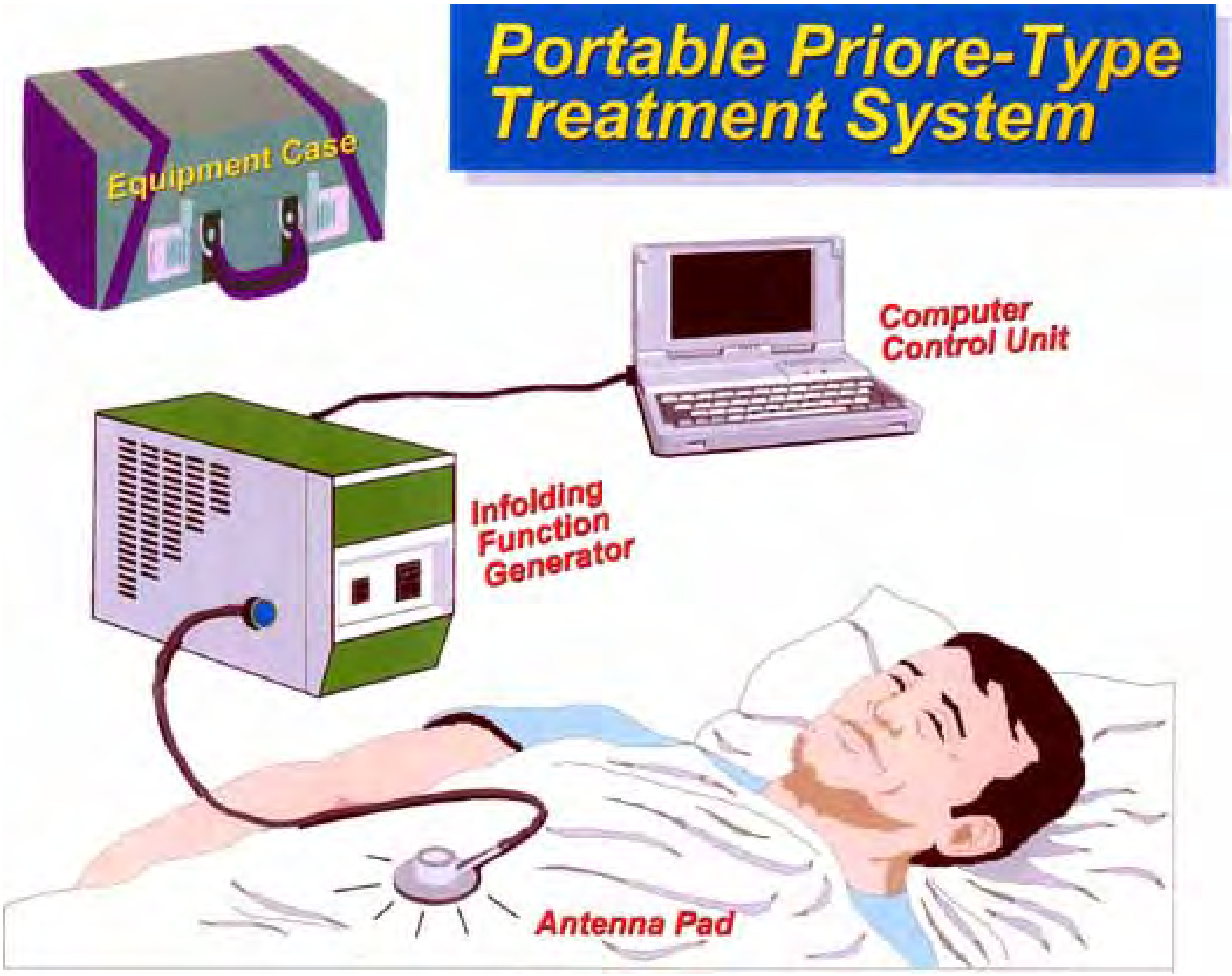

Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

Portable System Block Diagram

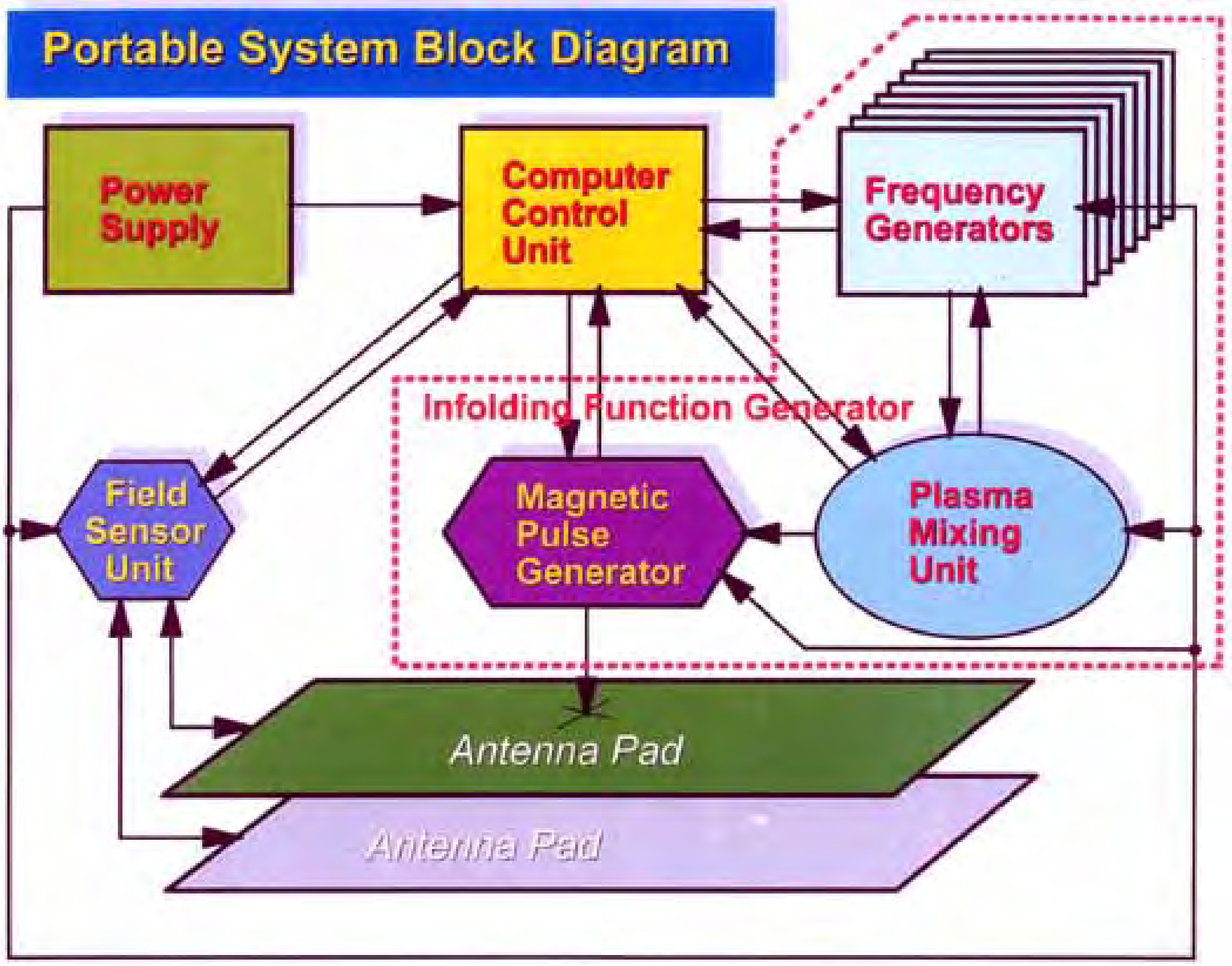
Power Supply
Computer Control Unit
Frequency Generators
Infolding Function Generator
Field Sensor Unit
Magnetic Pulse Generator
Plasma Mixing Unit
Antenna Pad
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

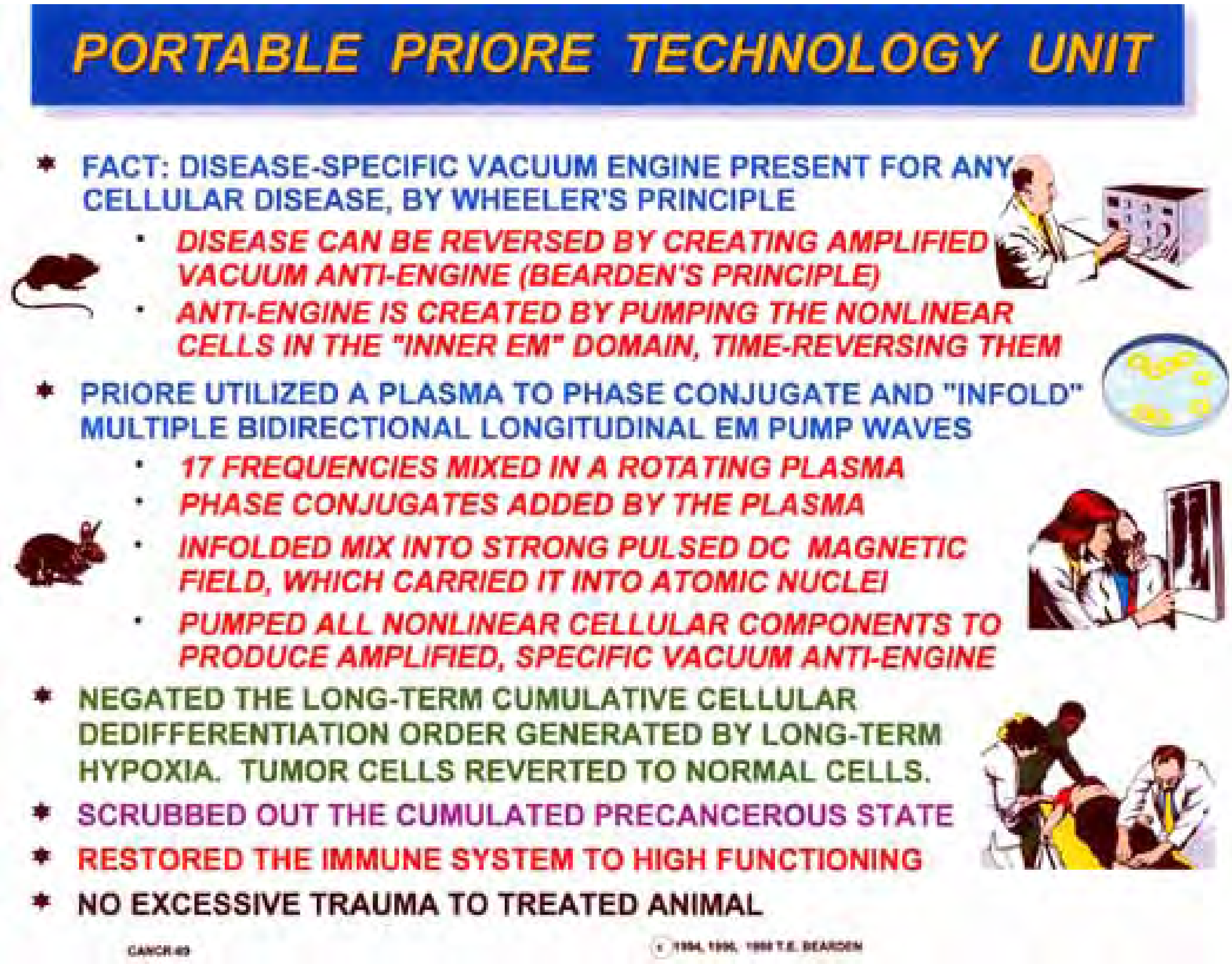

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

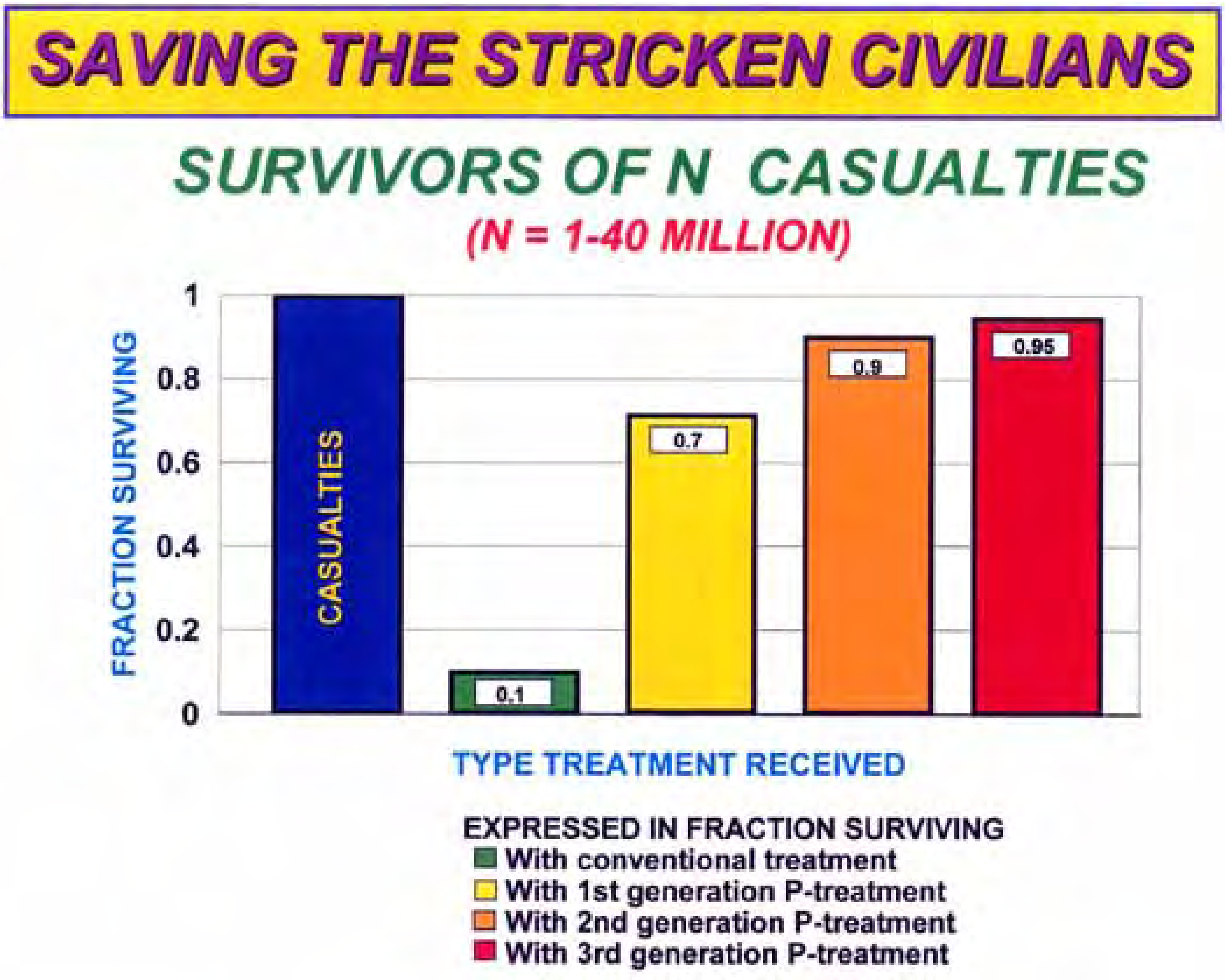

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

EXPRESSED IN FRACTION SURVIVING
- 🟩 With conventional treatment
- 🟨 With 1st generation P-treatment
- 🟧 With 2nd generation P-treatment
- 🟥 With 3rd generation P-treatment

Return to Slide Index

Next Slide

Previous Slide

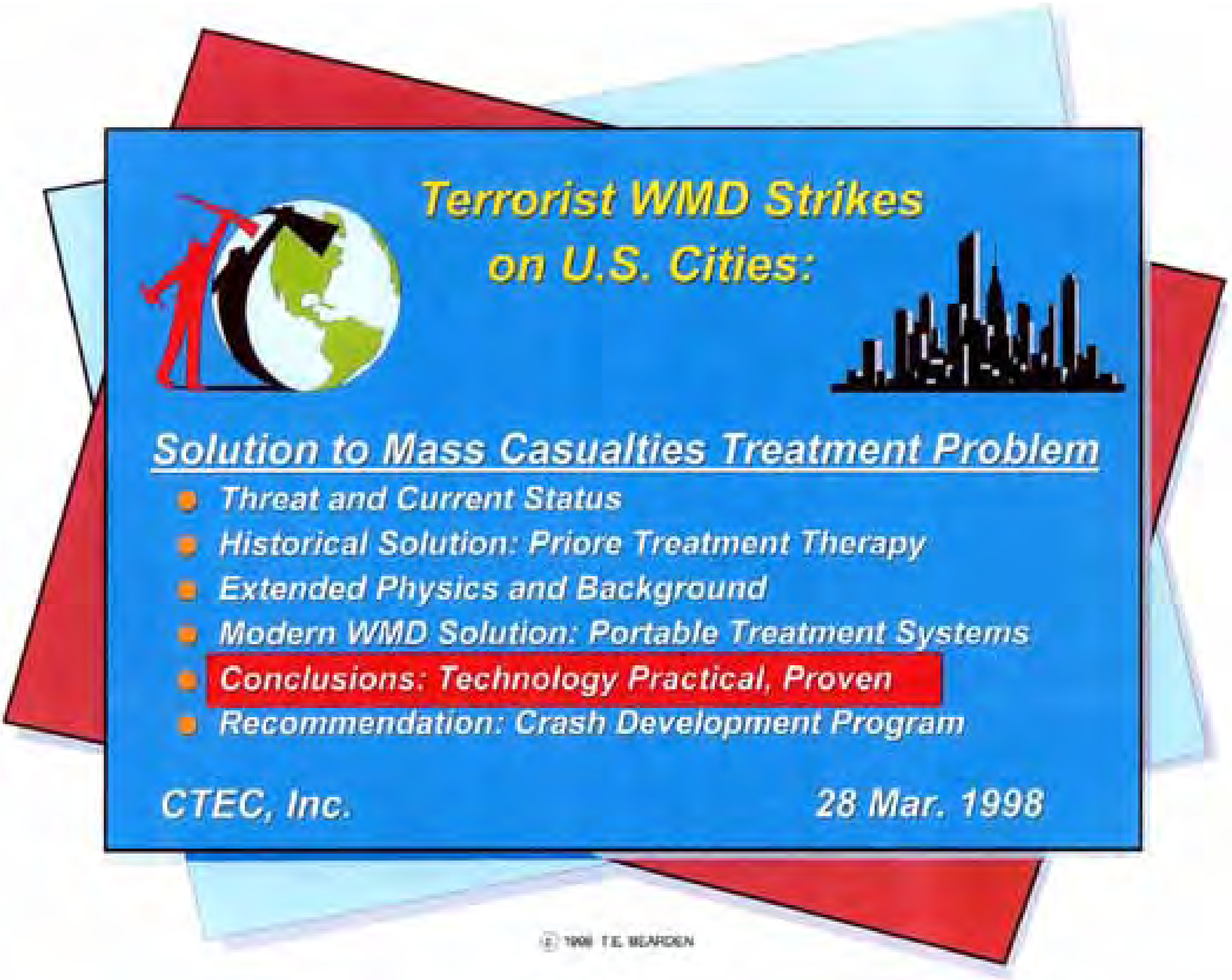

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

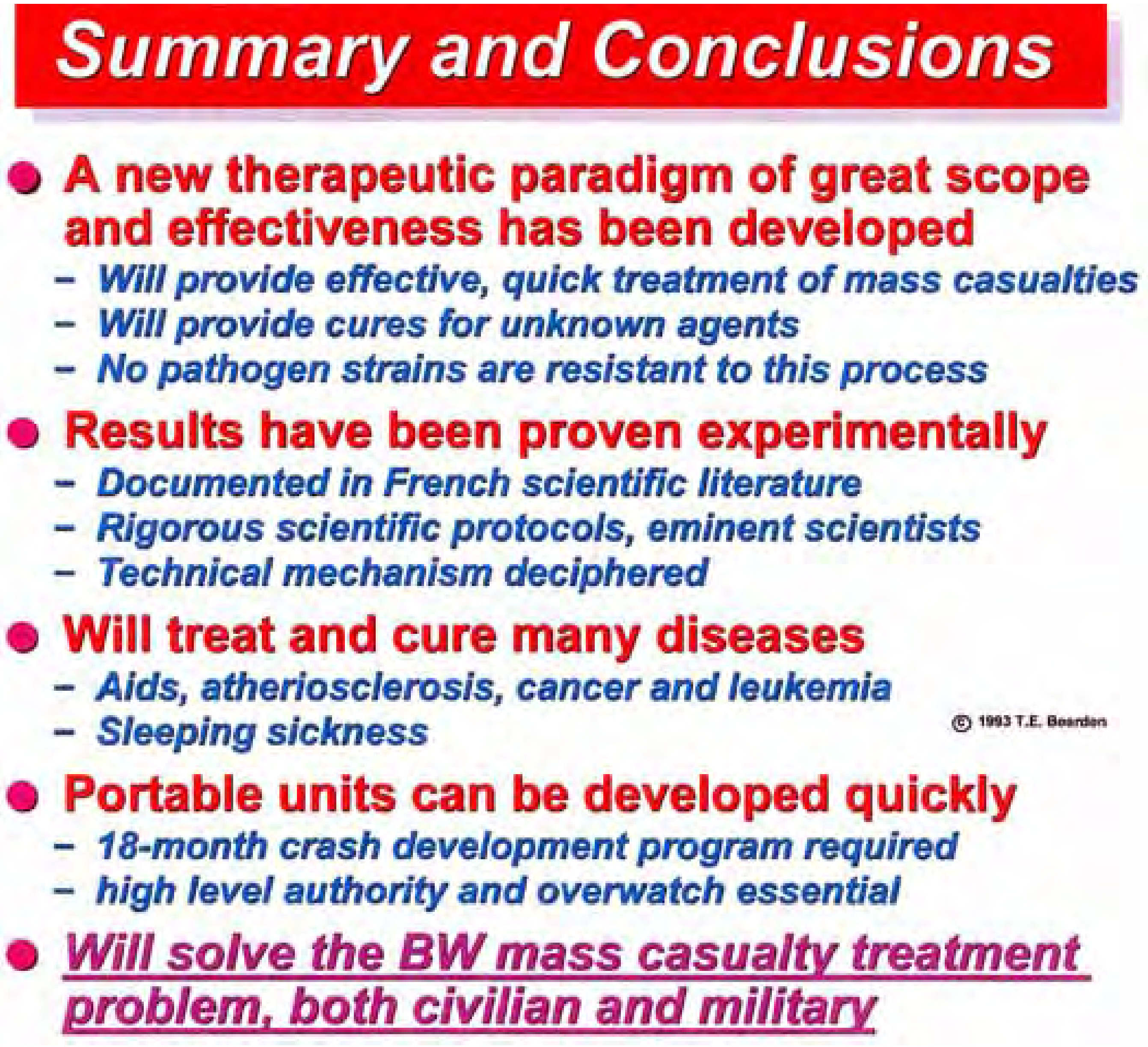

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - *Will provide effective, quick treatment of mass casualties*
  - *Will provide cures for unknown agents*
  - *No pathogen strains are resistant to this process*
- Results have been proven experimentally
  - *Documented in French scientific literature*
  - *Rigorous scientific protocols, eminent scientists*
  - *Technical mechanism deciphered*
- Will treat and cure many diseases
  - *Aids, atheriosclerosis, cancer and leukemia*
  - *Sleeping sickness*
- Portable units can be developed quickly
  - *18-month crash development program required*
  - *high level authority and overwatch essential*
- *Will solve the BW mass casualty treatment problem, both civilian and military*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

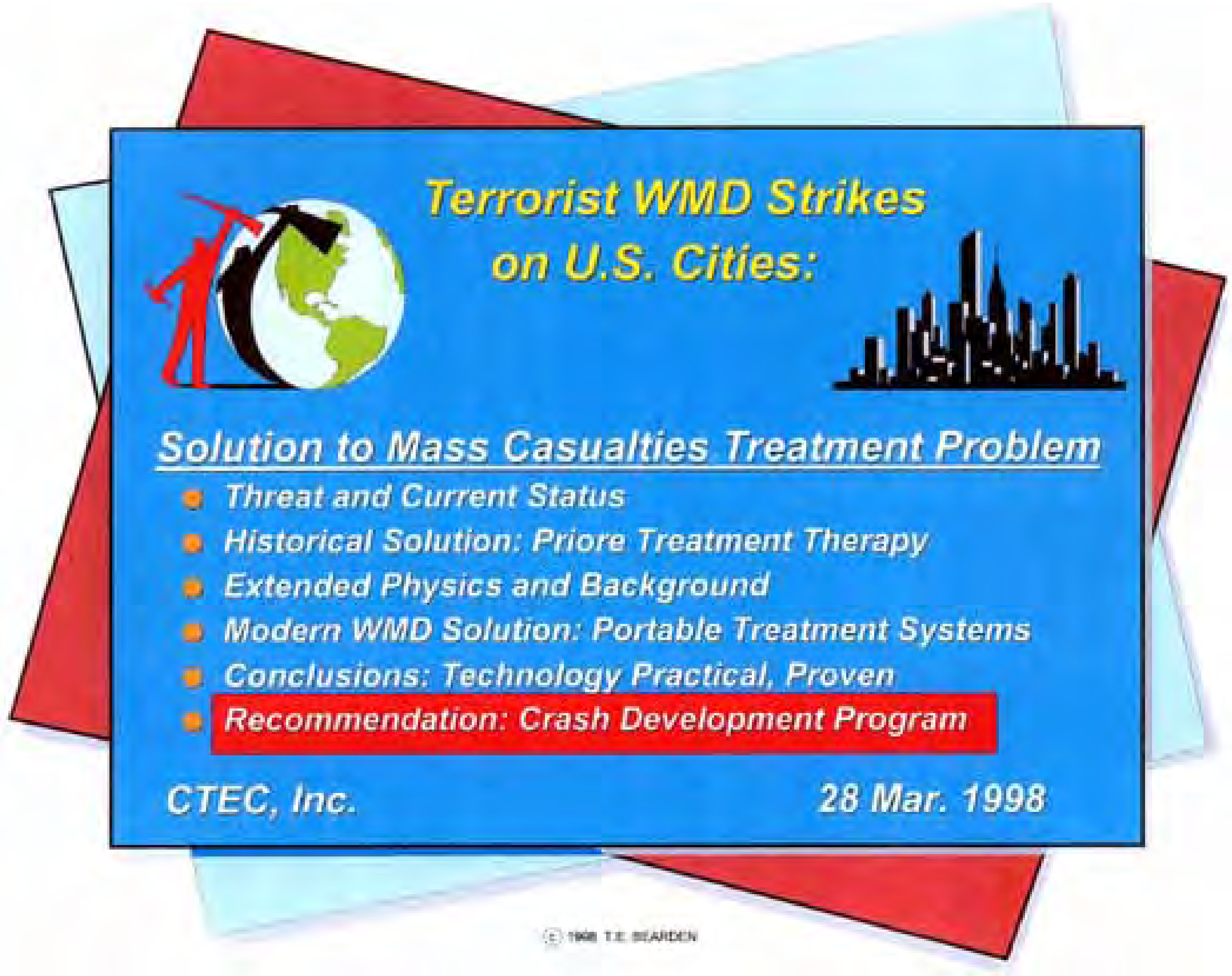

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc. 28 Mar. 1998

(c) 1998 T.E. BEARDEN

[Return to Slide Index]

[Next Slide]

[Previous Slide]

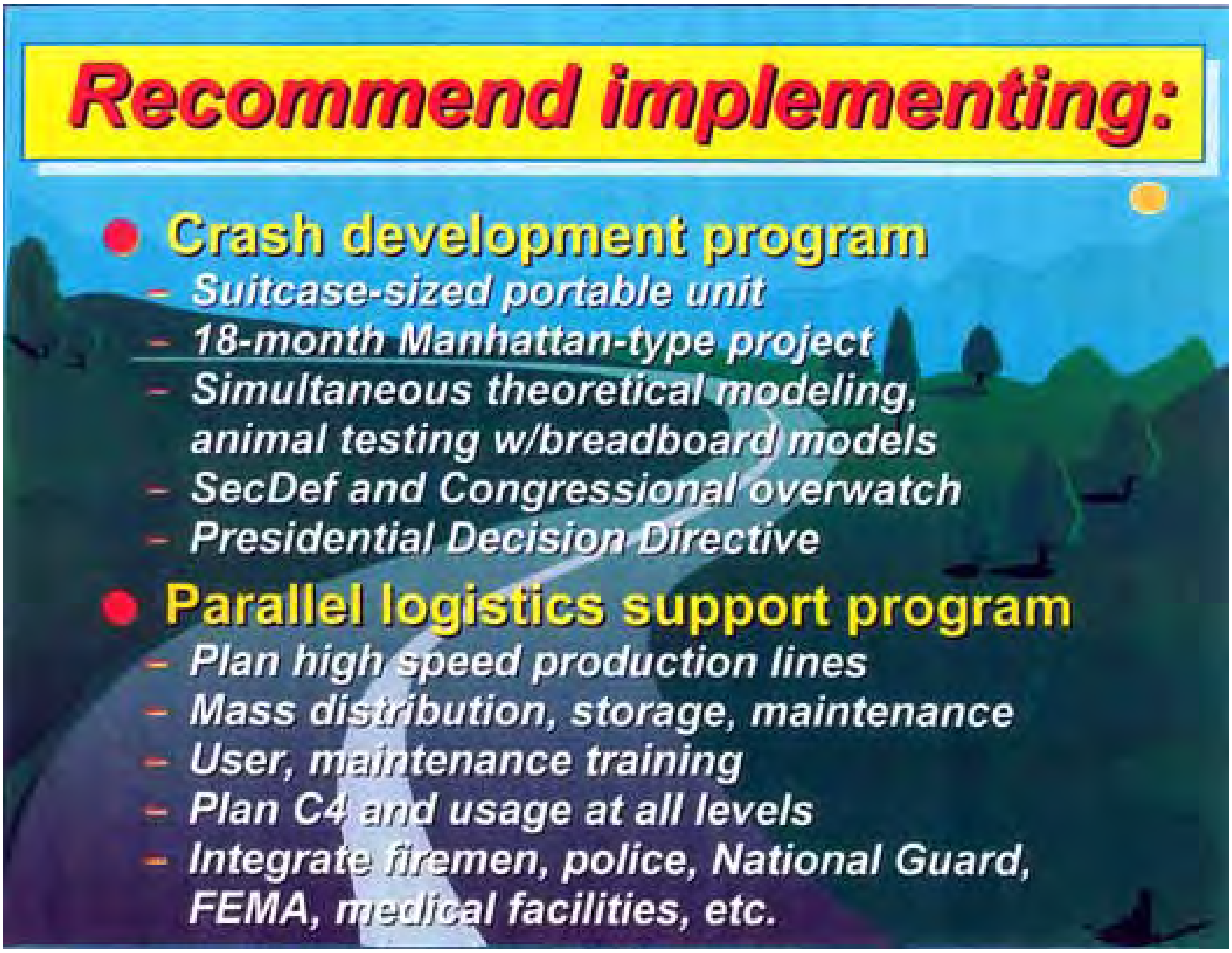

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

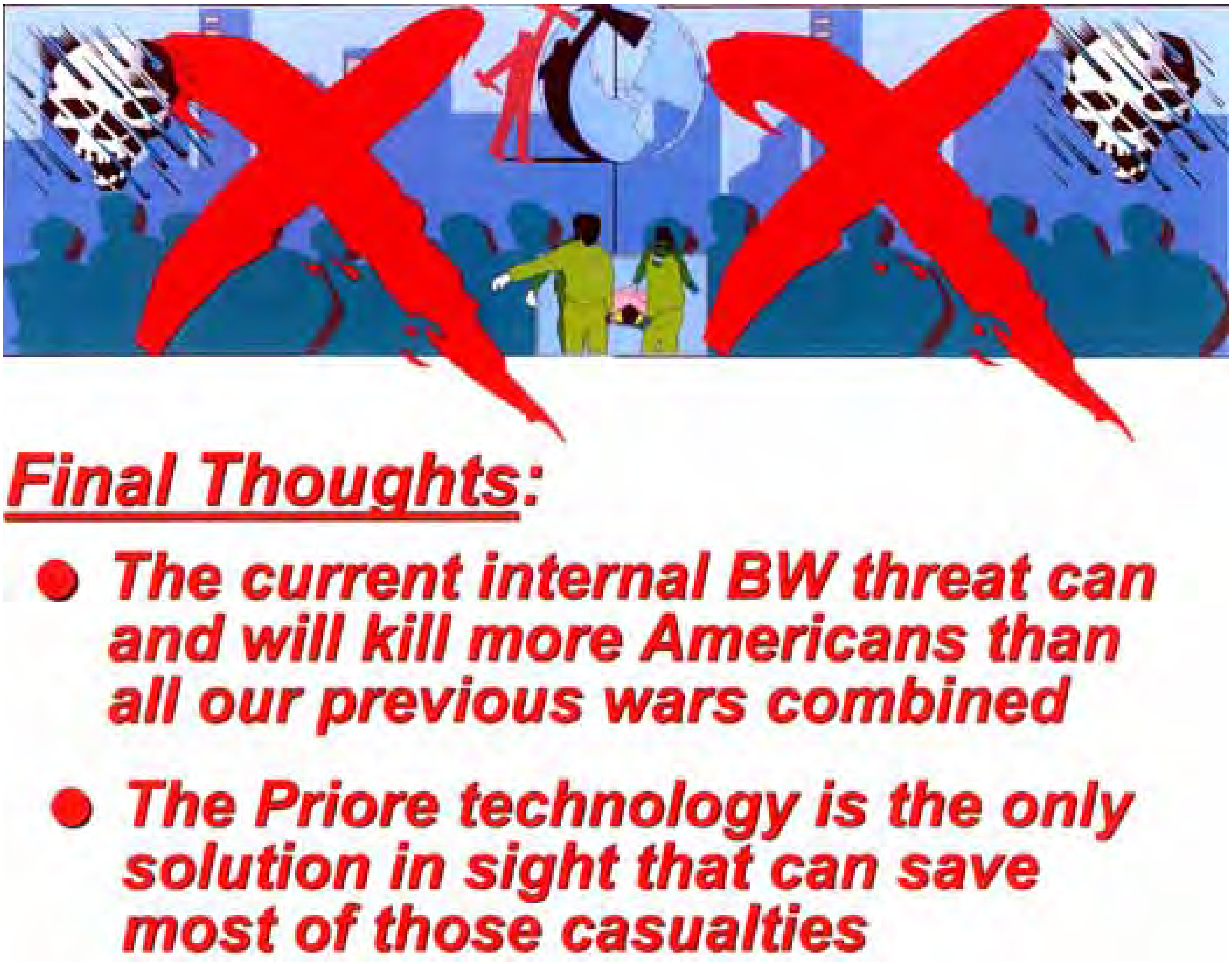

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/
(256) 539-5951 Fax

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular *regenerative* system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

Incls: Full package with table of contents

T.E. Bearden
President & CEO

Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

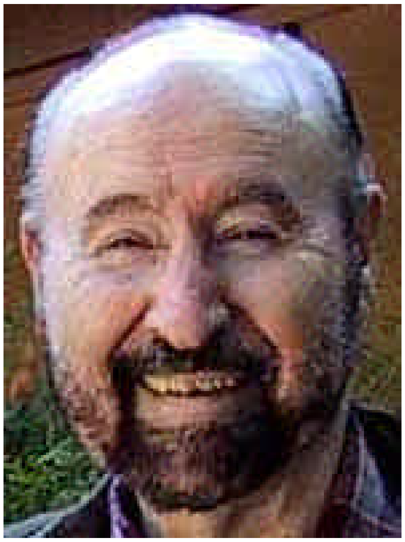

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world" has come about by studying the information given by Col. Tom Bearden at his website Cheniere. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity. This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "**Scalar Electromagnetics**." But there is a great deal of information at Cheniere which needs to become common knowledge as fast as possible, for the sake of the survival of life on earth. To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of bolding certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic**," and "**LW**" means "**longitudinal wave**")

This article has six sections:

New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels." Nikola Tesla

"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself." Tom Bearden

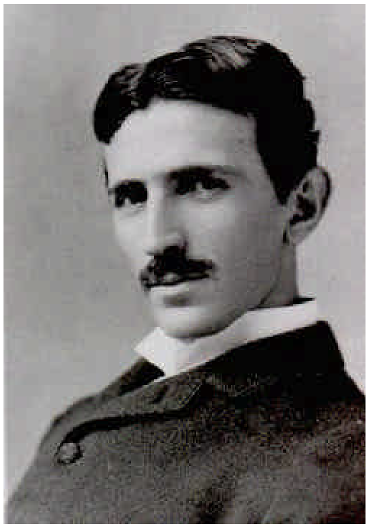

I guess the first thing to try to comprehend is that a "new" kind of electromagnetic (EM) wave has been discovered in the empty vacuum of space called “longitudinal” waves, which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the discovery really goes back to **Nikola Tesla** and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology for over 30 years and have **weaponized** these "new" longitudinal scalar waves to a great degree.

These are the very weapons **Nikita Khrushchev** spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could wipe out all life on earth if unrestrainedly employed.'" Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, discusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:

1. The **solutions to the energy crisis and the "oil problem" are in hand**. These oil wars are unnecessary. There is endless energy available freely from the domain of time.

2. **Unbelievably powerful weapons** are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.

3. The **cure of diseases such as cancer and AIDS**, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.

4. **Mind control on a mass scale** has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.

So the layman will need to understand that there is a new kind of electromagnetic energy that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called transverse EM waves, to distinguish them from the new longitudinal EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that the "emptiness" of empty space is in fact not empty, but a great ocean of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." It seems like something from Star Trek but this is where scalar electromagnetics has come to. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also spacetime, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that time itself is actually compressed energy. And it is energy which is compressed by exactly the same factor by which matter is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous E=mc2. It is now paired with E=tc2 (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"

and

"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash the tremendous energy that is compressed into time itself. It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, several new names

> "Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "energetics". That is their approach to a unified field theory, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "bioenergetics"), and all mind operations and mind-matter interactions (the third branch of energetics, called "psychoenergetics"). [ed: sometimes also called "psychotronics" by the Russians]

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the new science this way:

> "Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) which underlie all physical reality.
>
> "If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to alter gravity, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, space-vehicle drives, time-travel, teleportation, paranormal phenomena, and just about every other area one can think of." http://twm.co.nz/Beard_scalem.html

> "Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. Time-as-energy eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with spacetime and with spacetime curvature."
>
> Bearden: http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

The scalar energy discovered in the vacuum of space is sometimes called "zero-point" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical “creation of the vacuum” as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas. Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we’ve got a real vacuum. Right? Well, since then, both theory and experiment have shown that there is a non-thermal radiation in the vacuum and that it persists even if the temperature could be lowered to absolute zero. Therefore, it was simply called the “zero point” radiation. "Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

> "**For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960**." - Bearden
>
> http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that a permanent solution to the "energy crisis" is at hand, if only those who hold it so secret would give it up to humankind.

It means our species' insane use of oil is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that unbelievably powerful weapons can be, and already have been developed. In fact these weapons are so dangerous that the fear engendered becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the huge energies of time, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because most people would indeed be truly alarmed to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and ban the use of scalar weapons in war or in mass mind control. The current deployment of these weapons is a world emergency which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole

The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the essential secret of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. Just make a dipole.

> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, nature then gladly gives us as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the giant negentropy mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole."
>
> Bearden http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!

Perhaps the most eagerly awaited event now is the marketing of Bearden's Motionless Electromagnetic Generator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."

(Bearden : http://www.cheniere.org/correspondence/011202.htm)

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

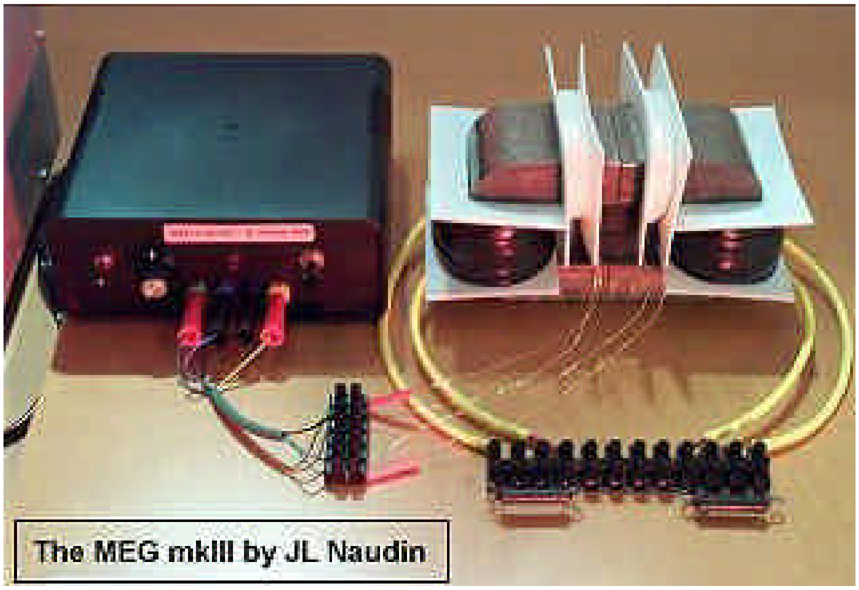

The Motionless Electromagnetic Generator

(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But Magnetic Energy Limited's MEG is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government on the order of the Manhattan Project.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through Magnetic Energy Limited, Inc. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S."
>
> Bearden http://www.cheniere.org/correspondence/121901.htm
>
> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
>
> Bearden http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden http://www.cheniere.org/megstatus.htm

Free Energy has been Suppressed

Secret Government Aware of Free Energy

> "In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] De Palma and his ideas.
>
> "Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that free energy is a threat to the world order constructed by business and the mindset of those who want to own the world.
>
> "Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.
>
> "The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an

> attitude of paranoid delusion and fantasy which can only arise from the alienation of a class of elitists who, through their money, are completely insulated from the reality of the day to day fight for existence of the common man."
>
> http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

Bearden describes one case of suppression when the Yakuza killed the Kawai engine in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> "They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is “The Unnecessary Energy Crisis: How to Solve It Quickly,” especially since the "oil wars," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.

> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper."

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The good news is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the dipole, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the greatest discovery in human history lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "conspiracy of ignorance" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere. "As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.
>
> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized

> Science communities. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden National Emergency Declaration is Needed
>
> "The Government Non-Technical community (the Senate and the Congress, in particular) are in far better shape than the Government Technical community, to appreciate the world implications of the pending economic disaster. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to get this vacuum energy program launched, under a National Emergency declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."
>
> unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

> "Personally, the present author regards the increasing energy crisis as the greatest strategic threat to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.
>
> ...In the name of all humanity, let us begin! Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> "Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap." Bearden http://www.cheniere.org/books/ferdelance/intro.htm

"**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that,**

if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon." Khrushchev, to the Presidium, Jan. 1960

> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target from the local vacuum at the place of the target.

From an interview:

> Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"
>
> Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."
>
> http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

3a. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these new facts public and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "Tesla Howitzers." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The first howitzer mode is called the "exothermic" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply destroy all electronics in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "leaders" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a new world, whether we like it or not. Anyone could be expected to go into denial, the befuddled senators and congressmen included. But we have to wake up.

In a second howitzer mode called the "endothermic" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "cold explosions" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

> "The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these heat (exothermic) and cold (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the endothermic howitzer mode as a weapon of war:

"Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary transverse EM waves, the longitudinal [LW] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space at the location of the target itself. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

You can hear the Woodpecker signal here: http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the LW waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers to a pinpoint location on earth. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "Mindsnapper" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would not rot for quite a long time.

The Mindsnapper can be used for a small area, or to kill an entire population. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in mind-snapping mode against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

> "For example, suppose one attacks a tank. [with an LW weapon] With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
>
> All electronic systems of the tank are dudded. Total systems kill is achieved. The ammunition in the tank explodes. Total kill is achieved.
>
> The fuel explodes. This is another total kill mechanism.
>
> That tank has been killed totally, by a variety of mechanisms, all simultaneously.
>
> That's a k-kill in any analyst's book.
>
> And don't worry about retrieval and repair. That tank is finished permanently.
>
> Even the most modern tank is just as vulnerable as the most obsolete."
>
> Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And Star Trek's "Phaser," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and materialization and dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to

> underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even affect the sun.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is no longer mutual. And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a full-scale all out scalar attack on the west.

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counter terrorism] efforts." - Defense Secretary William Cohen, 1997
>
> In a letter to a writer named "Russell" Bearden says,
>
> "In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate,

> and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Tesla Howitzers in Weather mode

How to Make Weather

> "Engineering the weather is duck soup; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.
>
> "First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.
>
> "So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."
>
> http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather radar anomalies which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

In another section at Cheniere are many photos of cloud anomalies which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these clandestine ongoing "weather wars."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. This drought was broken by a colleague who used an extremely powerful scalar EM device to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions

resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.

"Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."

* On the weather maps.

http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion

Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

> "Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.
>
> "Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.
>
> "The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."
>
> Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The deepest dangers of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing true catastrophe.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that human activity could actually affect the sun itself is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental

that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.

"This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about."

http://www.cheniere.org/books/ferdelance/s24.htm

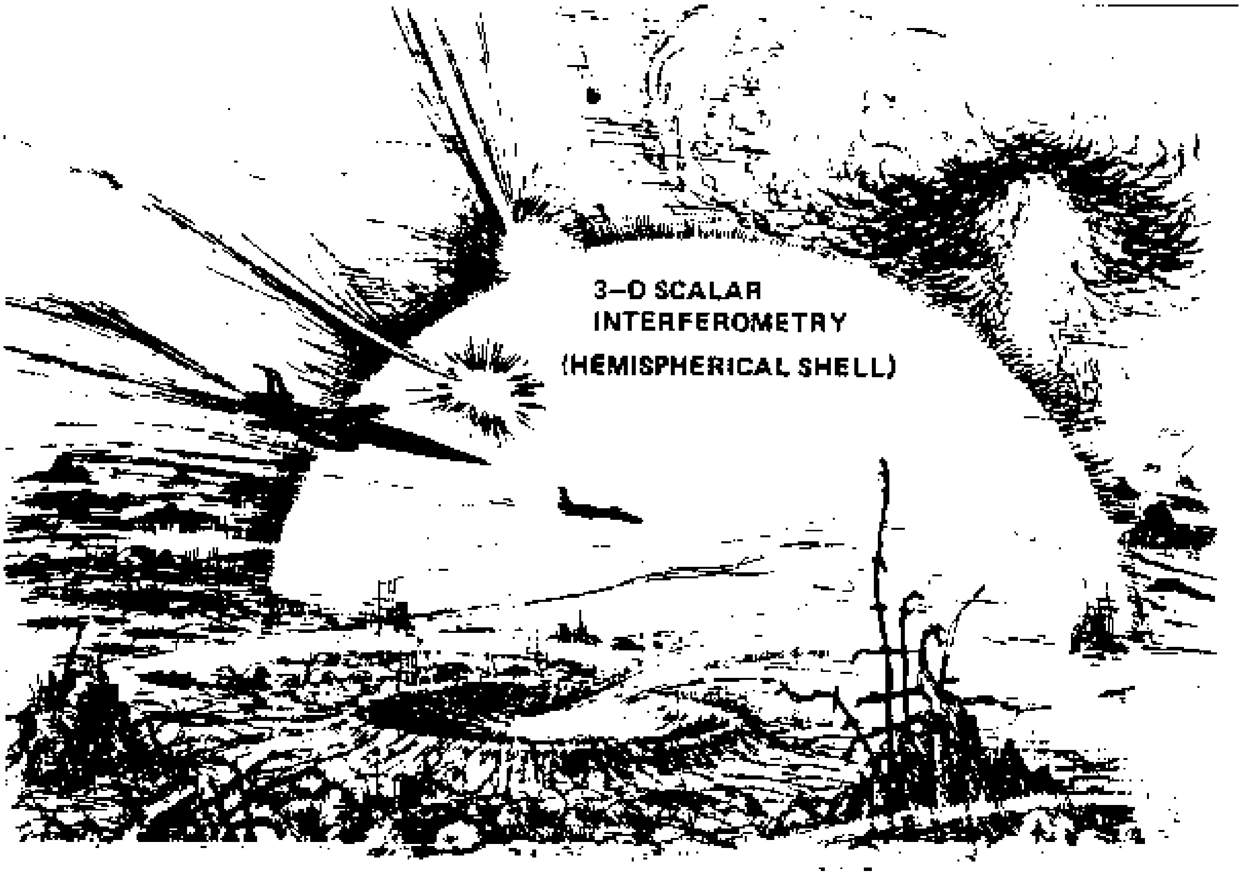

3b. Tesla domes provide near-perfect defense shield.

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly impregnable defensive mode whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum at the location of the shell. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc."
>
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area even against nuclear radiation itself. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (operational) a nearly perfect missile defense system. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world:

March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 | June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3c. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "rogue groups" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government."

> "They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control."
>
> "The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

3d. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now lots of new ways to bring any nation to its knees.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that it will collapse the U.S. economy in a way we have not seen since 1929. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

> Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.
>
> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament. Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???"
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

When he was asked the question "**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. The West is no longer defenseless, not by any means. Cannot amplify it further than that."
>
> http://www.cheniere.org/correspondence/030202a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

3e. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such scalar-powered antigravity vehicles, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmosphere.

> "They could seem to materialize and dematerialize.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now in great secrecy. Our own "government" may even be one of these.
>
> "If so, it isn't our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
>
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
>
> 3) the light-balls are able to eject smaller light-balls;
>
> 4) the light-balls change shape all the time;
>
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. But the cause, and the physical mechanism with which radiation is emitted, is currently unknown."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely baffled as to where the energy of these plasma balls was coming from.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomenon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. But where does it come from? We don't know."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls is coming from the vacuum of spacetime at the very location of the balls themselves, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as marker beacons giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many anomalous events of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The

stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: [Interview with Tom Bearden]

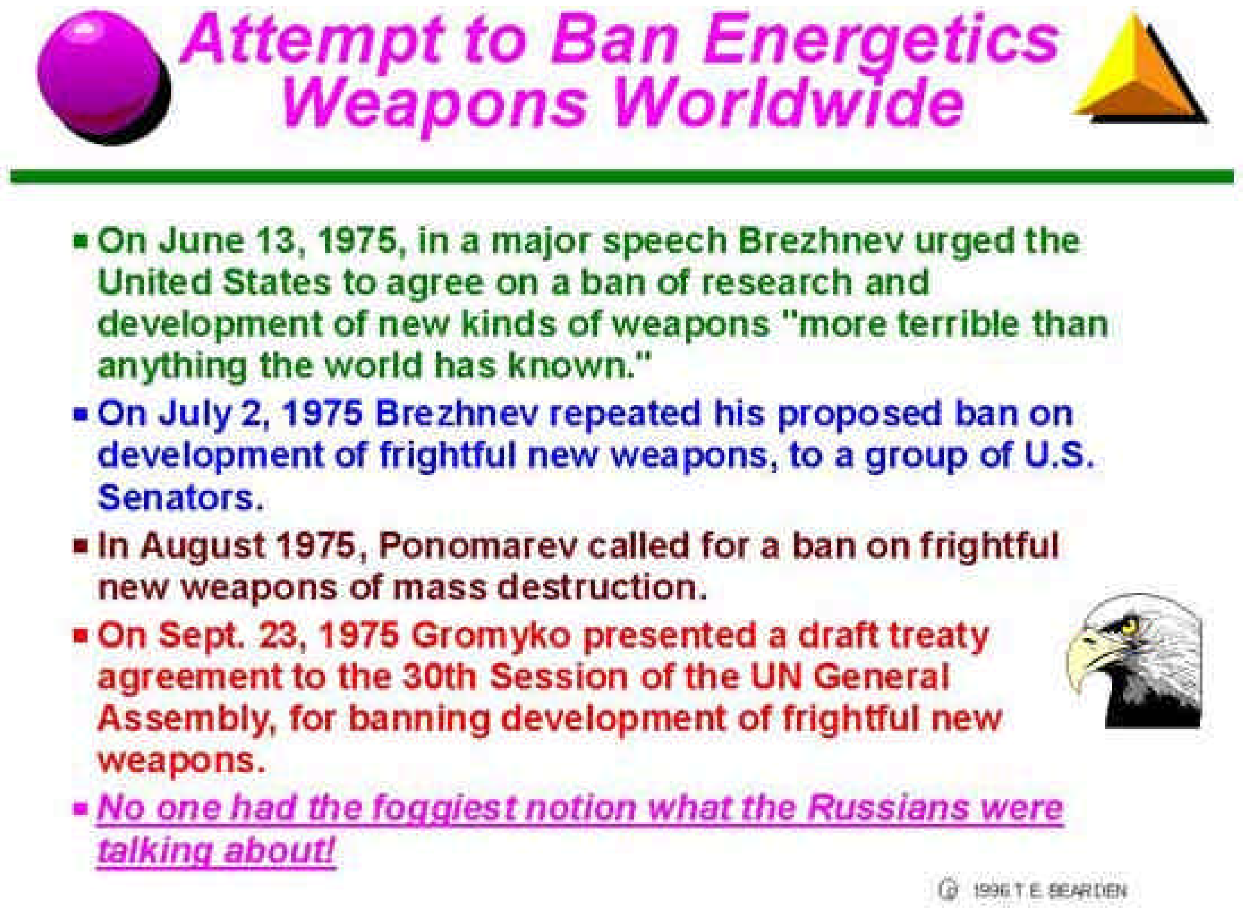

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T E BEARDEN

(Slide from Cheniere.org)

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

> "Any sort of disease whatsoever - physical or mental - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."
>
> Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a gigantic glass tube filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore project was funded by the French Government to the tune of several millions of dollars (equivalent).
> http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "phase conjugate pairs" and one of the pair is time-reversed. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually time-reverse the cells back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to suffer the fact that the cure for cancer and AIDS was right there, in this new scalar electromagnetics technology, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Priore, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that big giant two-story glass tube consistently cured cancer in animals.

"**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for controlled cellular dedifferentiation, in biology terms. Physicians will time-reverse diseased, damaged or aged cells back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."

> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor Antoine Prioré. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.
>
> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces cellular reversal by weak time-domain pumping of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of healing rather than a science of intervention. And it will develop as a science of unlimited healing, since no pathogen can resist action by curved spacetime engines." Bearden
>
> http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an antenna unit which surrounds the body. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

In fact, Bearden has presented to the U.S. government a plan to use a portable version of a Scalar Healing Blanket where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A two minute treatment of the waves is sufficient to restore the body to its earlier healthy condition. Thus even thousands could be treated easily and quickly in a biowar mass attack situation.

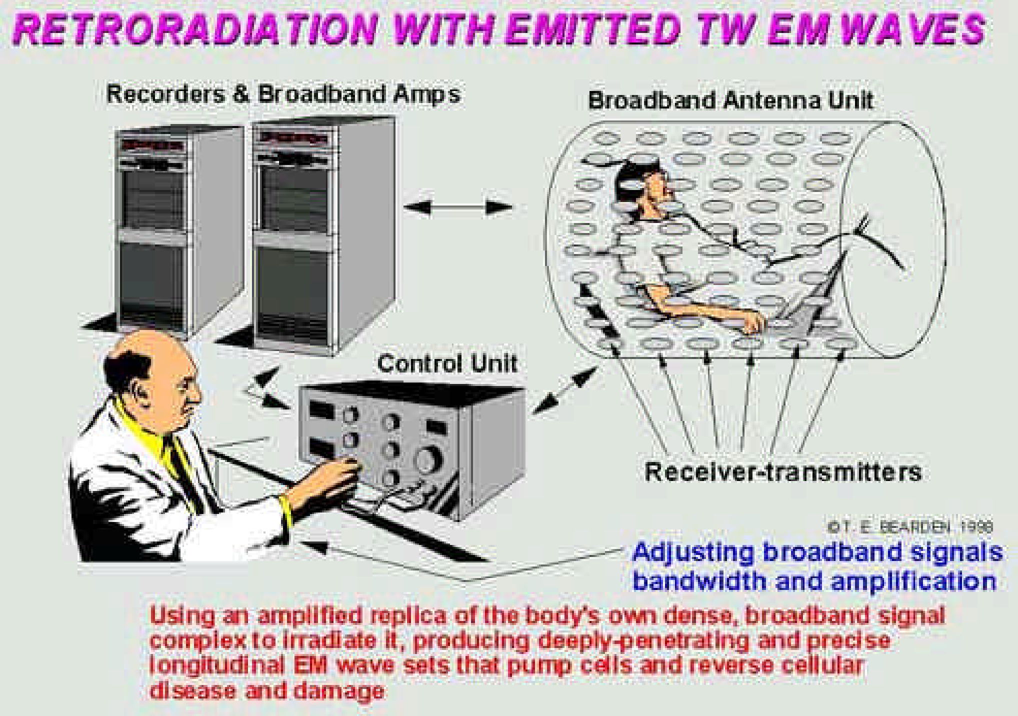
RETRORADIATION WITH EMITTED TW EM WAVES

Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
© T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification

Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

"To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. "The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals,

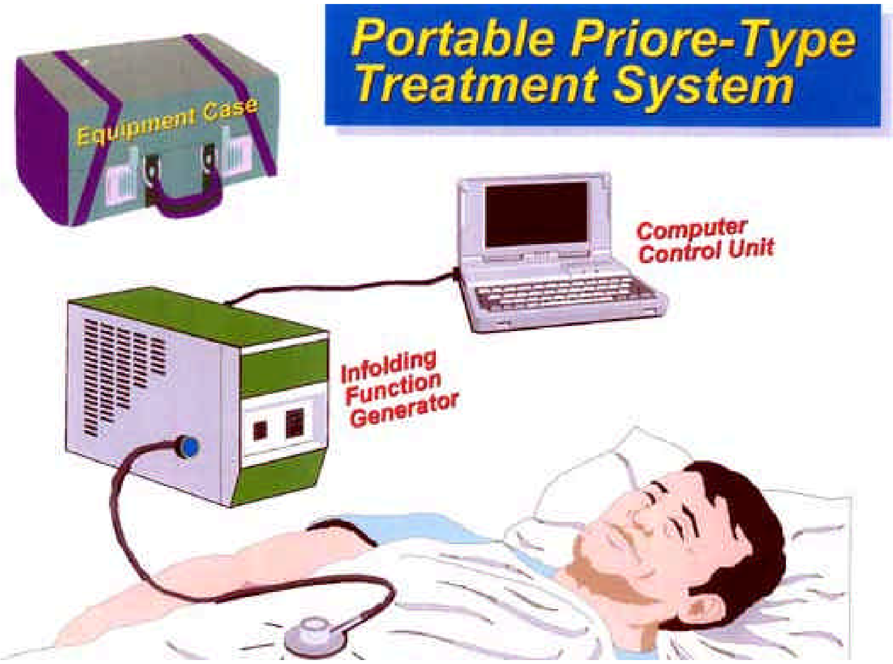
Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator

> emergency response teams, etc."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p04.htm

The Scalar Healing Blanket would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on befuddled government ears. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article "**Healing with Energy**": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:
Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the eventual deaths of three U.S. Ambassadors."

> Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

> "Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases

Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a conventional bio-attack would be greatly potentiated. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the "shadow" diseases that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these 'vacuum engines' or 'spacetime curvature engines' that acts on mass to generate the exact effects produced by anthrax. (Or any other disease one wishes) . . . Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .
>
> "And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the "shadow state" of a disease, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .
>
> "So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the necrotizing fasciitis disease engine. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. A few of those targeted bodies will now actually develop necrotizing fasciitis. So far, just like the U.S. Embassy targeting, except a different "carrier".
>
> "But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. The few cases of necrotizing fasciitis that break out will be most puzzling. They will be randomly scattered across the

> entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."
>
> "And that one happened, right here in the U.S., not so long ago. Check it out.
>
> "So carefully adjust the patterns for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has 12 shadow state disease engines in it.
>
> "Place that one on a populace. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "spreads" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.
>
> "Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. In the "thinned" set, one can at least double or triple, and even up to five times, that estimate. So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "That one is being set up, or already set up, right now. There are other such "augmentation" setups also under way.
>
> "So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "flesh eating disease." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This quantum potential disease induction weapon -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., one test used necrotizing fasciitis (flesh-rotting disease) engines.

> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. The major signature of the test was that the cases of the disease broke out totally statistically and randomly, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, but was a totally random set of occurrences. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

"Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.

> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that they would never be able to resist doing it."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the Gulf War Syndrome involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> "I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced... It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire electrical grids of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have no idea yet of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and engineer the human psyche. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no thing you can point to that is the mind. As Bearden says**, "The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the **interference zone** (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "Mind Wars."

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: '**One if by land, Two if by mind**."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden http://www.cheniere.org/misc/time.htm

In a slide Bearden illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself. For if my thoughts might no longer be "mine," then who and what am I? Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is “Mind Control and EM Wave Polarization Transductions” This is such serious stuff that Bearden includes a strong warning about misuse of this knowledge. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization's can be transformed one-into-the-other are presented."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p01.htm

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines

> (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired... Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gail’s collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p09.htm

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control1/p08.htm

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p01.htm
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p02.htm

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."

> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour. "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p06.htm

The paper "Mind Control and EM Wave Polarization Transductions" ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials. "The KGB scientists also know that something very like Gaia — a collective unconscious mind for all species on earth — also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials.""Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can

> then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (all systems tend to feudalism or some form thereof). "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.

> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> Bearden http://www.cheniere.org/explore%20articles/mind%20control3/p08.htm

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies." Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

"Anyway, it's presently a "Mexican stand-off" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."

"You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters can be made fairly readily and cheaply from nearly just Radio Shack parts or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take."

Bearden http://www.cheniere.org/correspondence/021702a.htm

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: this time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the wisdom of the wise could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, there is no limit to what humanity can achieve. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation. "So why not release these technologies? Because the social, economic and geo-political order of the world would be greatly altered. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the covert arsenal — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may

> endanger our future. Only a full, honest disclosure will correct this situation. It is not possible for me to convey in words the urgency of this."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has already leased some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo. They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. So the rogue Japanese acquired longitudinal EM weapons and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> Bearden http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. This is a world emergency.

And in Bearden's analysis, it all comes back to the issue of energy. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbert Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national Manhattan Project right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call. It was certainly bad enough, and one grieves for those lost American lives. But I fear the worst is yet to come, and it is coming at us like a runaway train hurtling right down the track at us."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 are in terrorist hands also, including the Yakuza and Aum Shinrikyo. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. So a great deal of the state of the world is not covered in the news at all, and will not be . . .
>
> "We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm
>
> "In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.
>
> "It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (it's going to go anyway, and catastrophically, the terrorists will see to that). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.
>
> "As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The

> scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.
>
> "Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).
>
> "It isn't. It has only just begun."
>
> Bearden http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The quantum potential weapons are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...
>
> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because it uses multiply connected spacetime, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> Bearden http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "Rule by Secrecy," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterward

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potential weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some agenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from Jean-Louis Naudin, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----

From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The Motionless Electromagnetic Generator from Tom Bearden is now PATENTED US 6362718 granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the magnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil extends around each of these posts. Input coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # 6,362,718

Patent Story on Rense.com:

Rense.com

MEG Scalar Energy Device
Patented - Production Starts Next Year

From Bill Morgan
wmorgan@nycap.rr.com
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan
wmorgan@nycap.rr.com

Do the magnets run down after a while?

Further developments: Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated experiments (such as Bohren's experiment,

which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism

Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

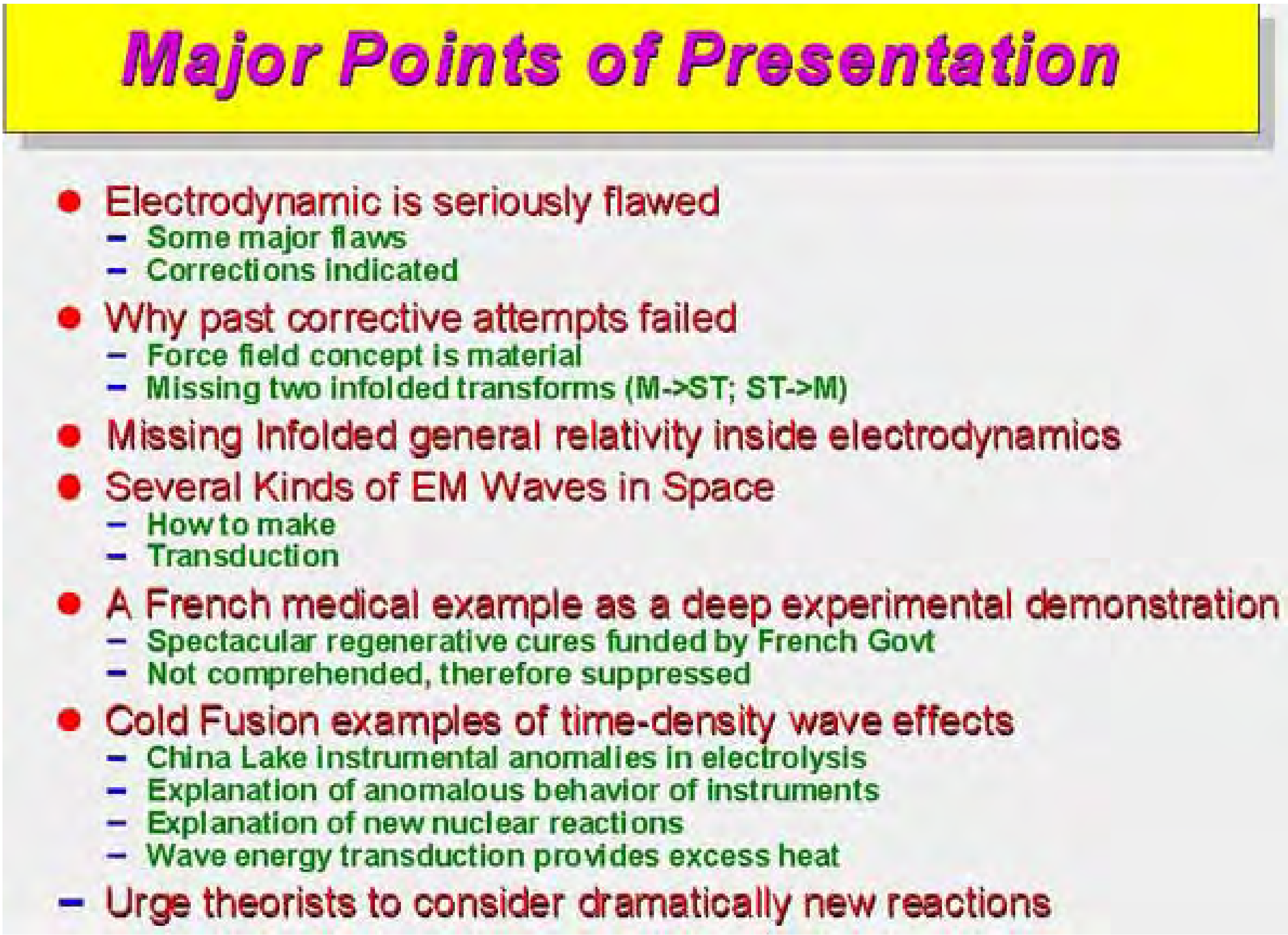

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

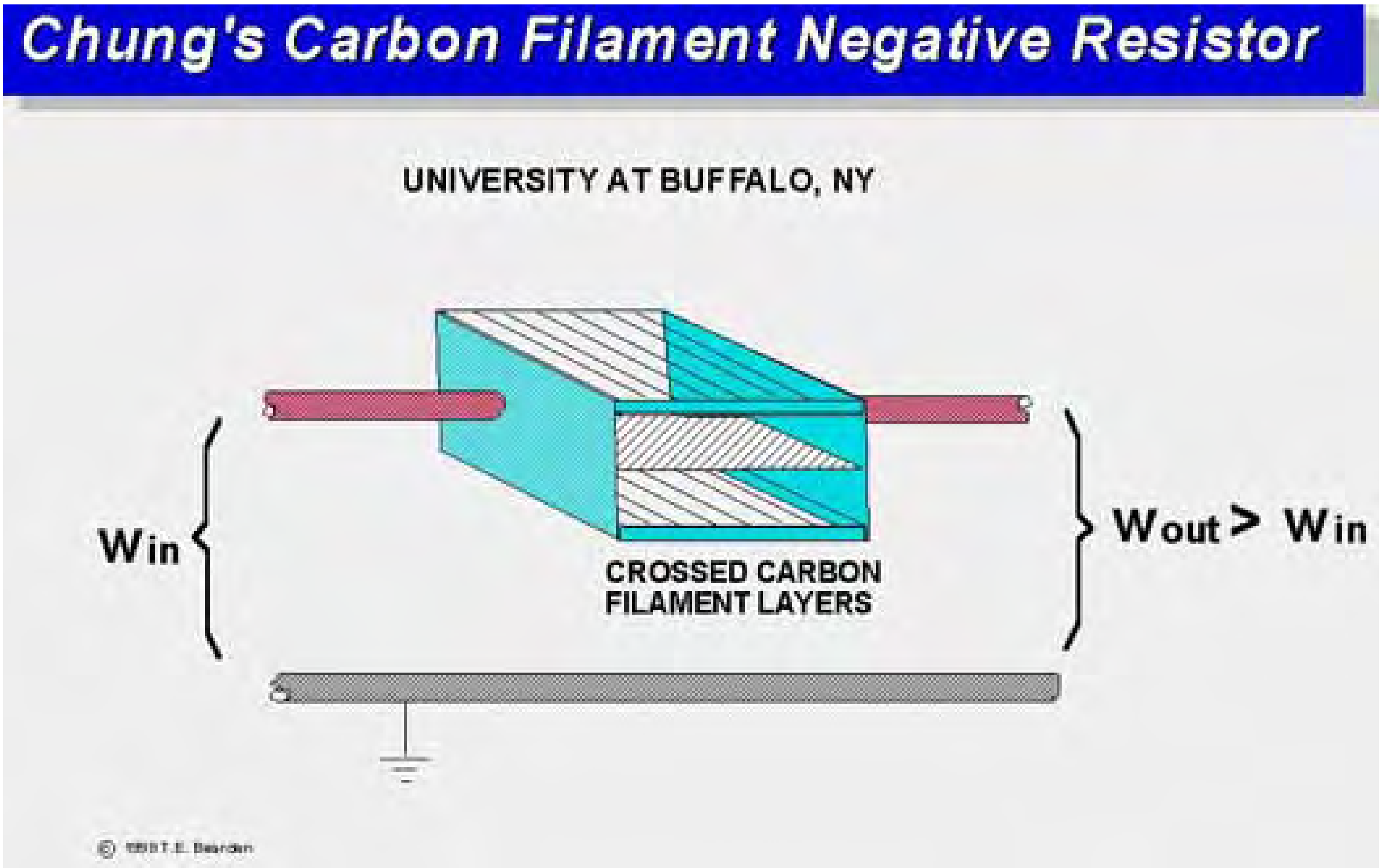

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

© [illegible] T.E. Bearden

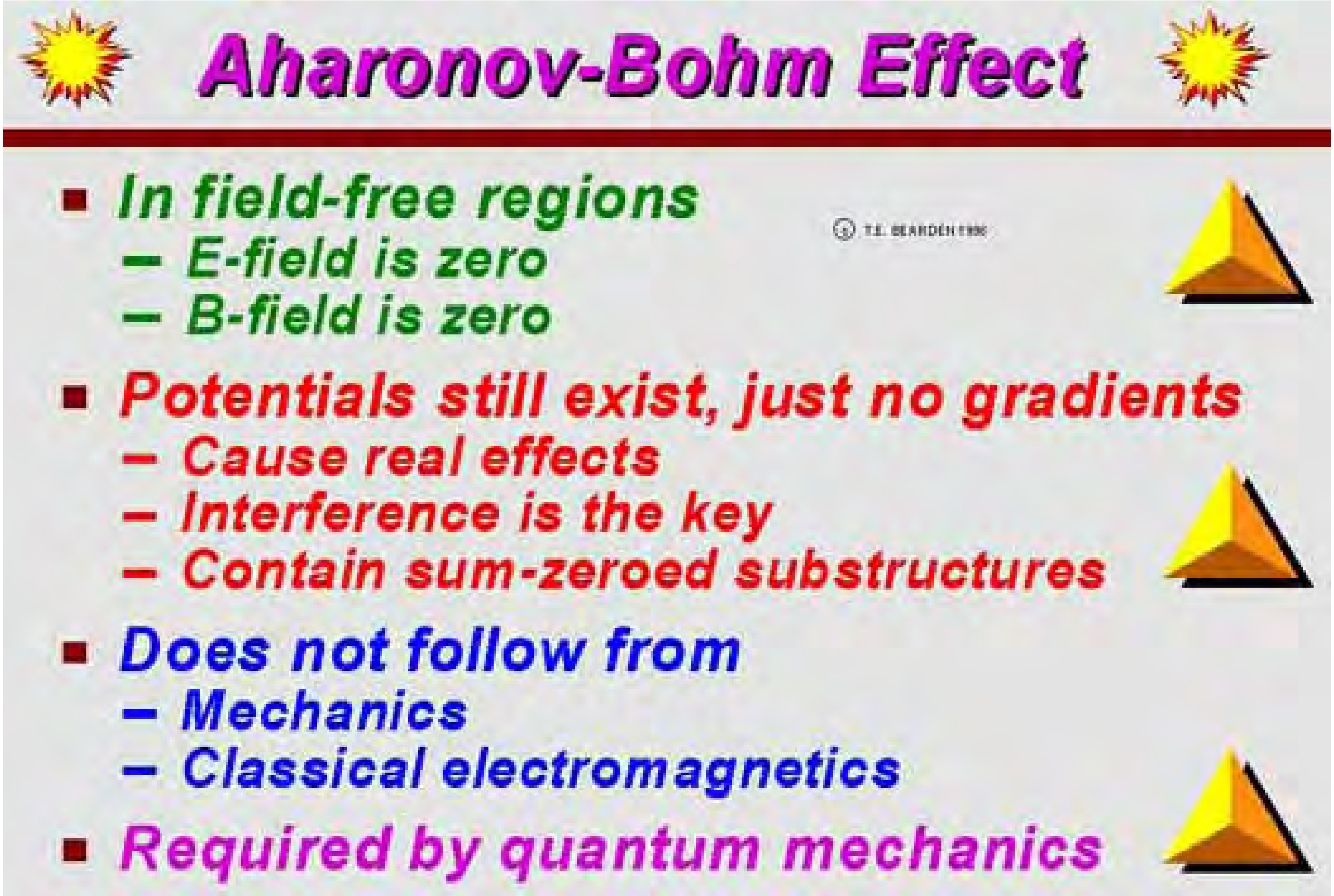

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

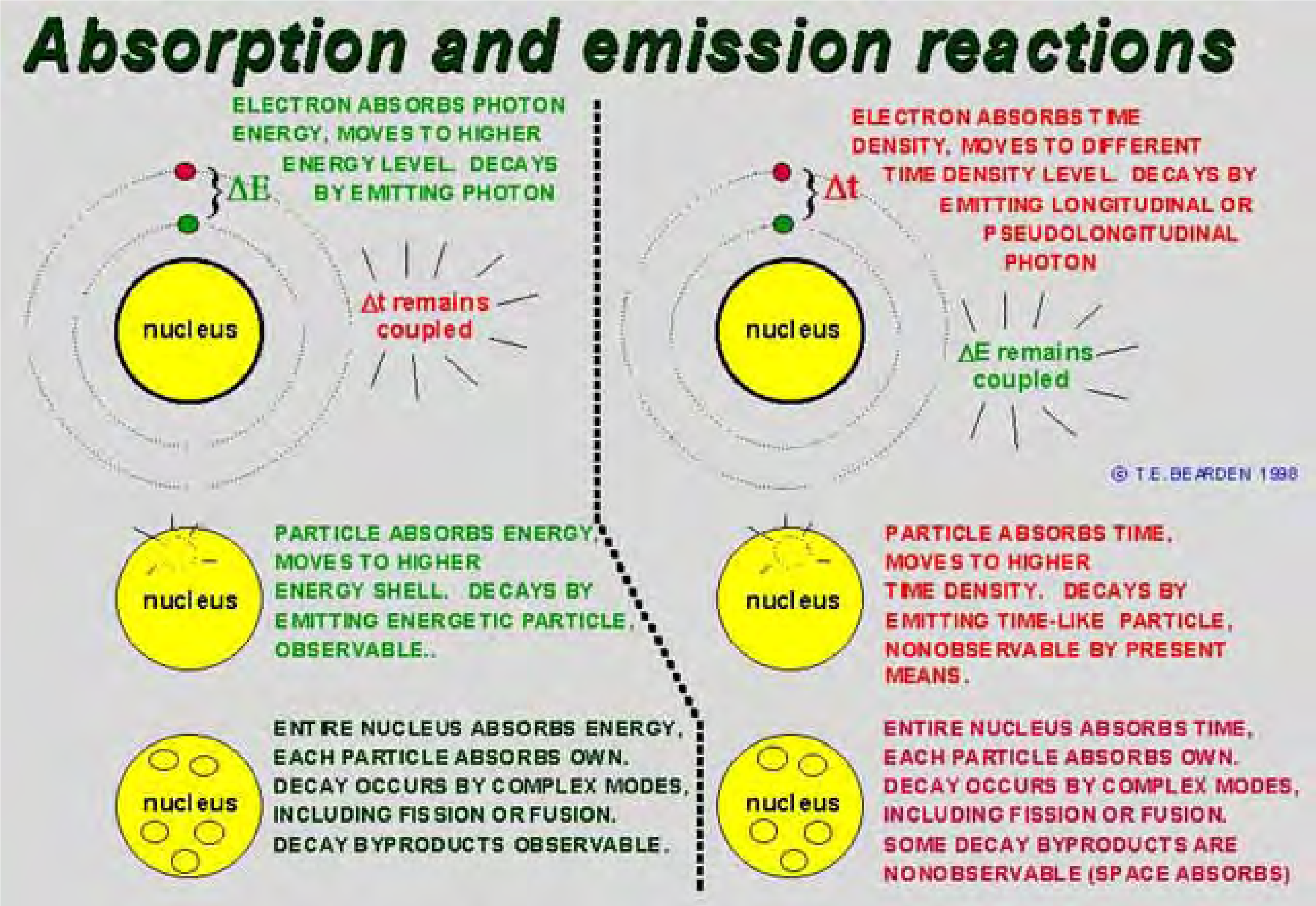

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

© T.E. BEARDEN 1998

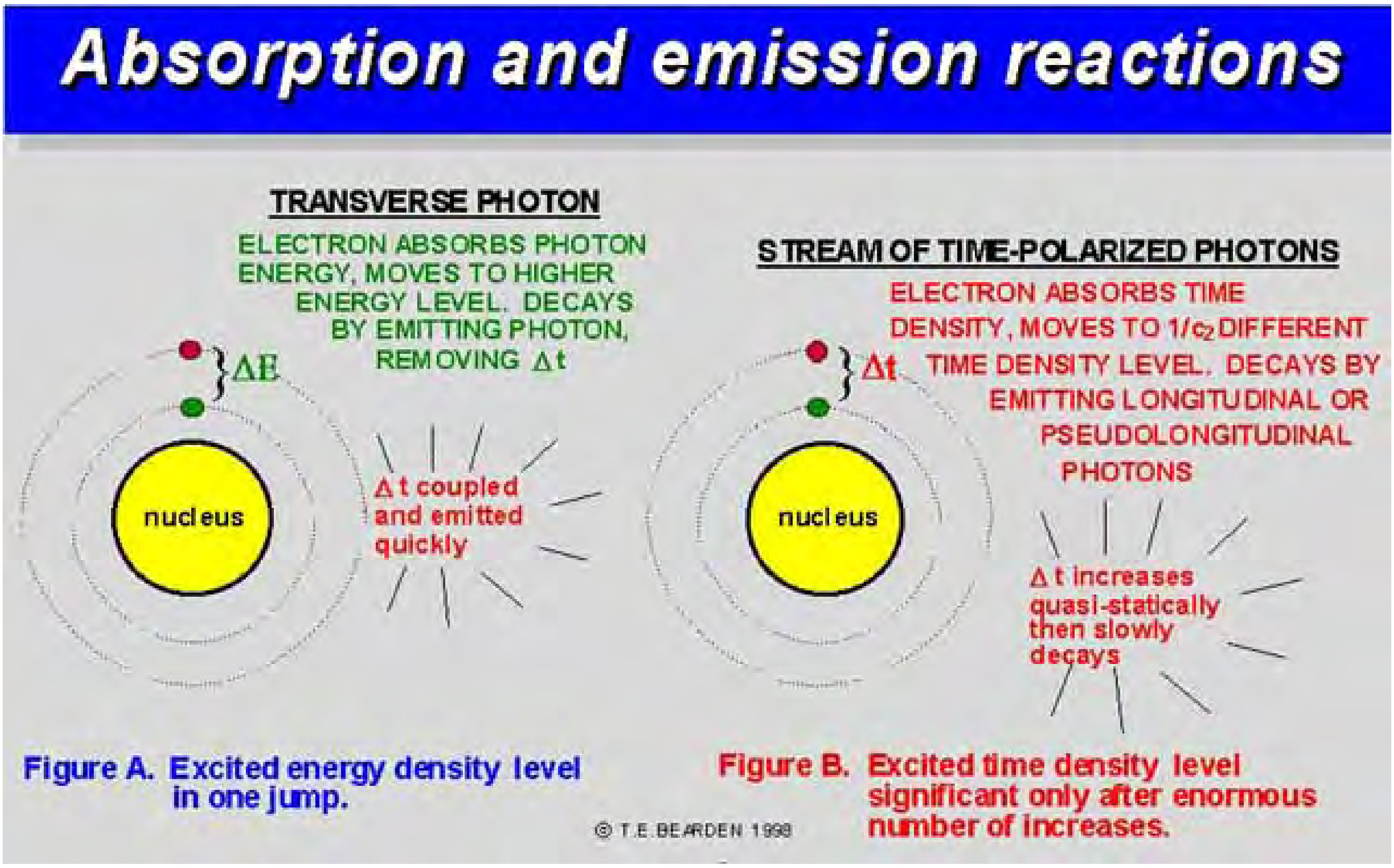

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

ΔE
nucleus
Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt
nucleus
Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

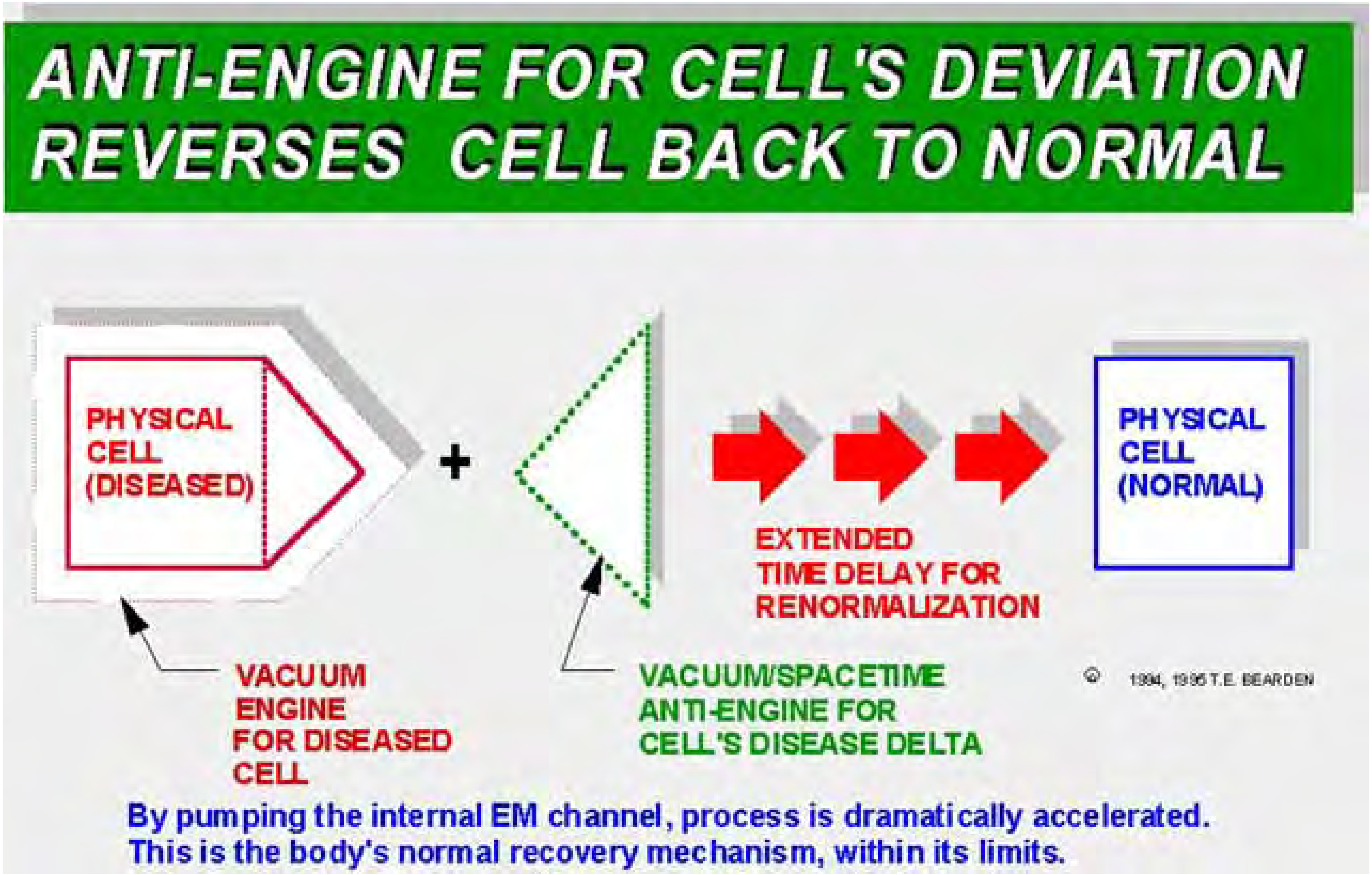

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

EXTENDED TIME DELAY FOR RENORMALIZATION

VACUUM ENGINE FOR DISEASED CELL

VACUUM/SPACETIME ANTI-ENGINE FOR CELL'S DISEASE DELTA

© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated. This is the body's normal recovery mechanism, within its limits.

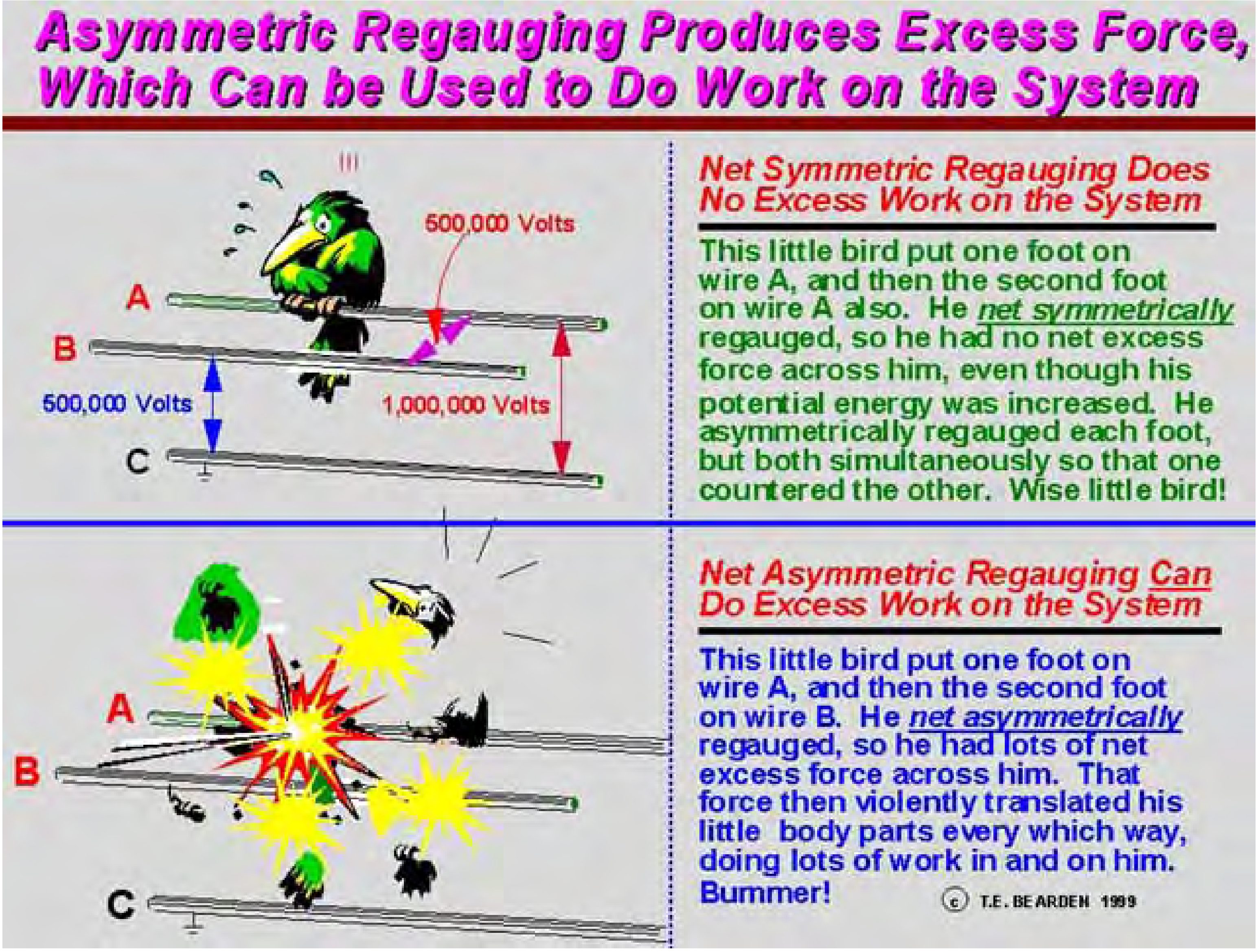

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

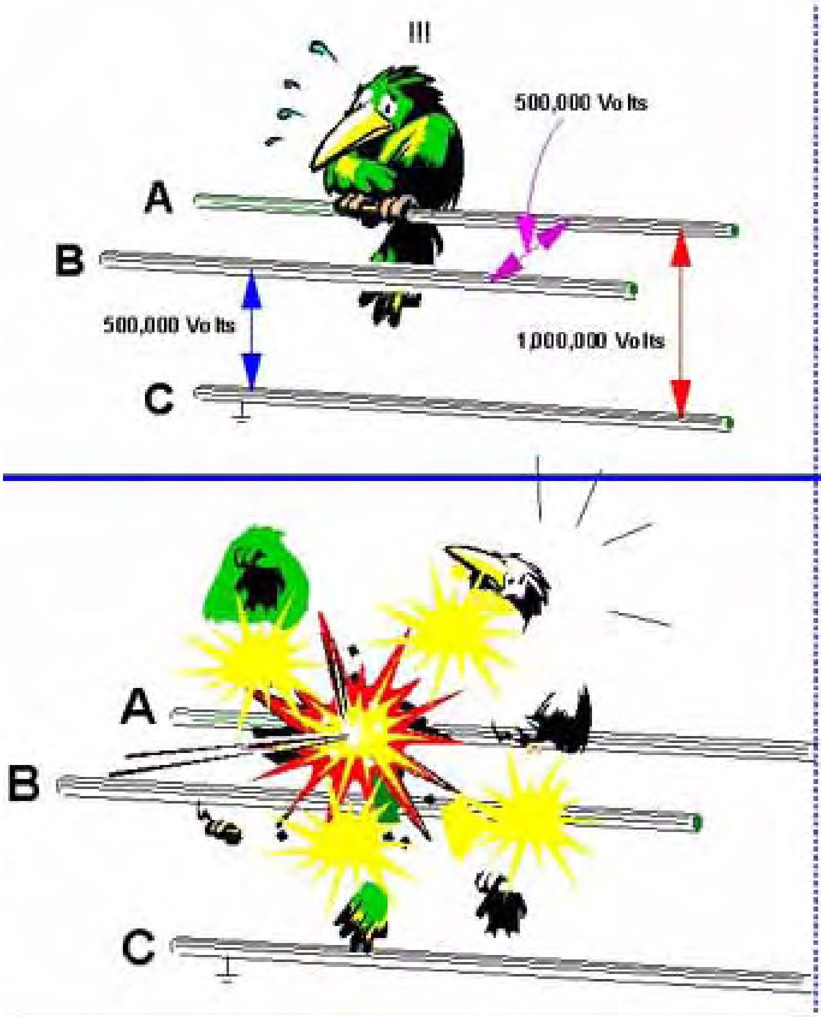

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1996

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

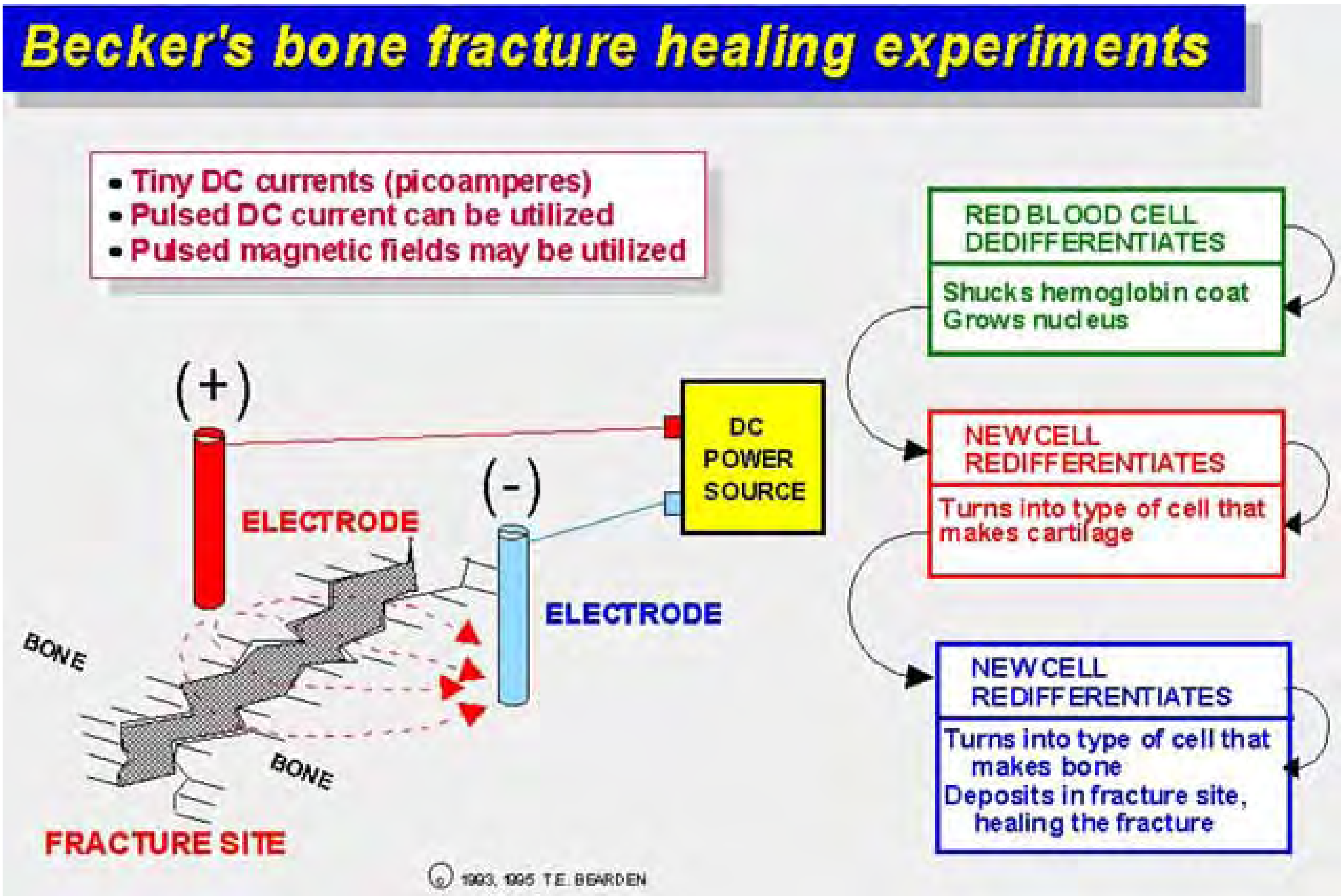

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

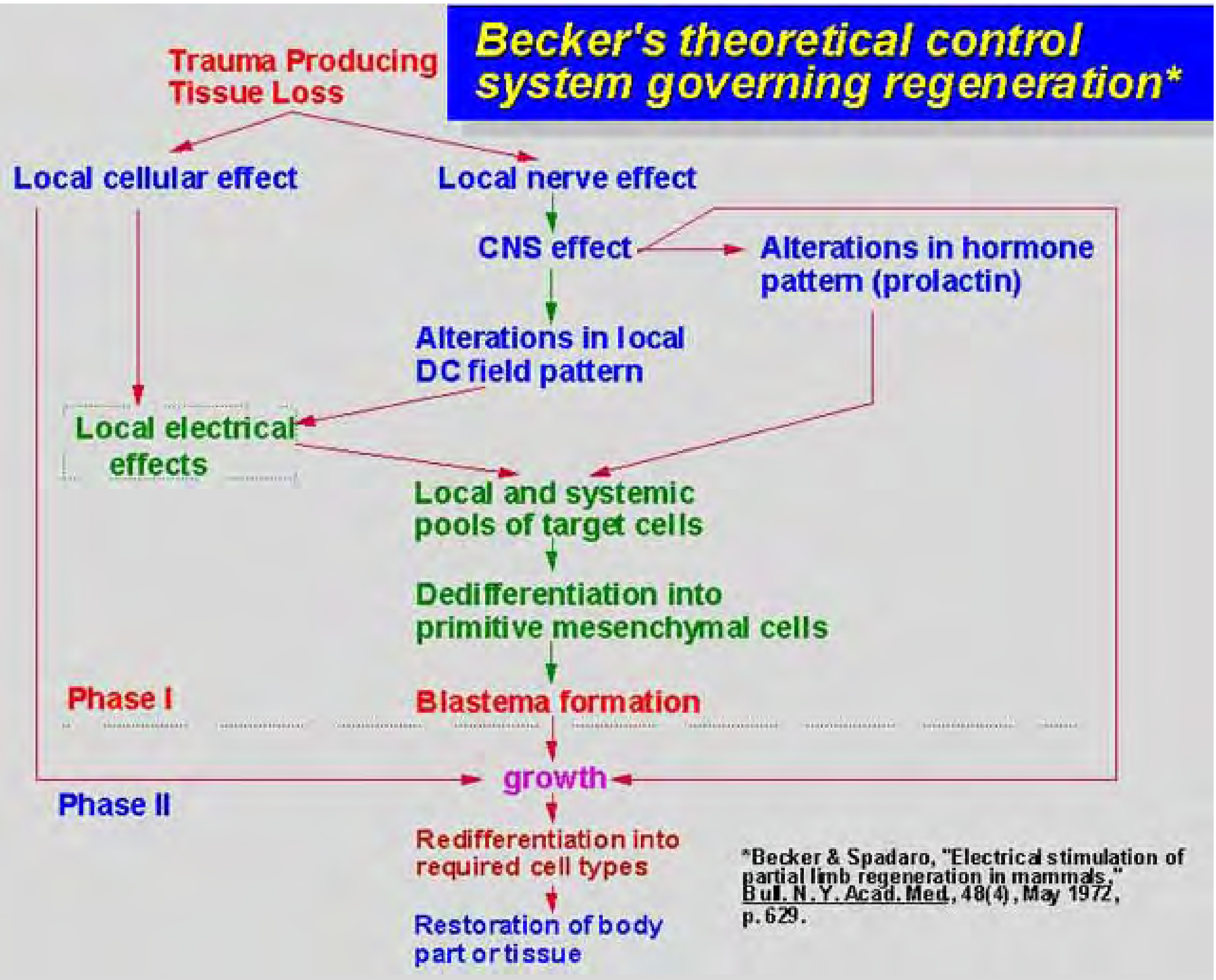

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bul. N.Y. Acad. Med., 48(4), May 1972, p. 629.

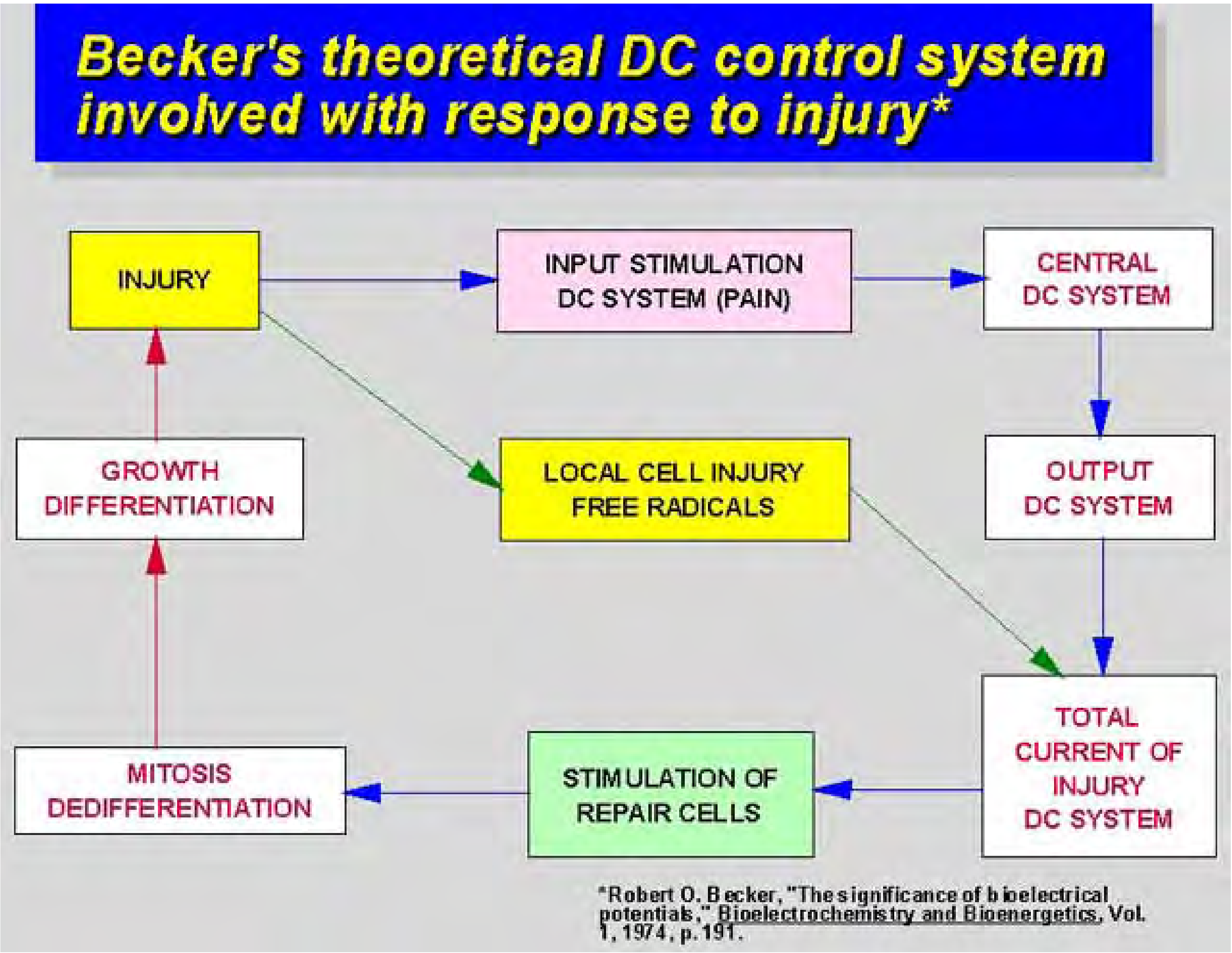

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

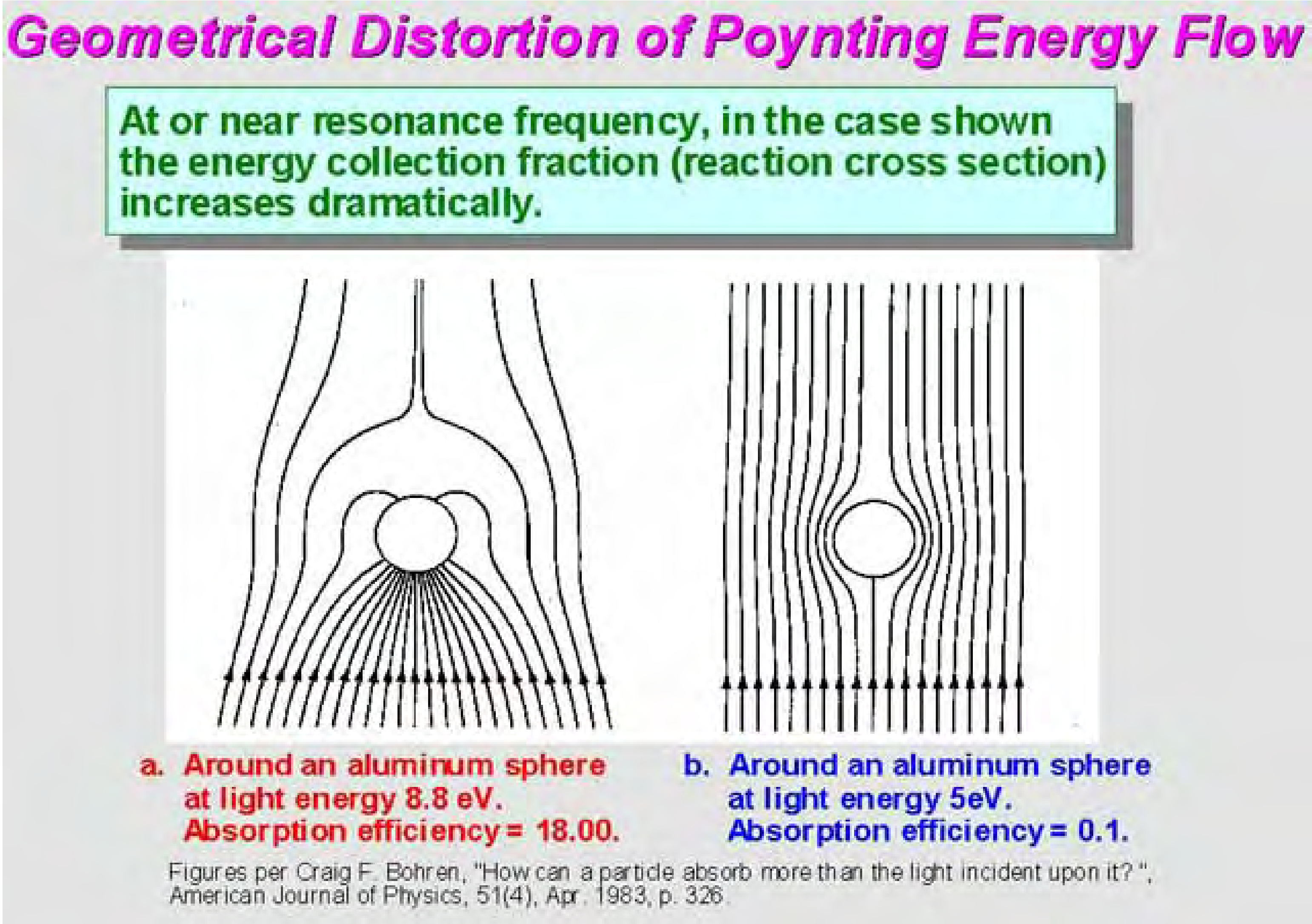

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it?", American Journal of Physics, 51(4), Apr. 1983, p. 326.

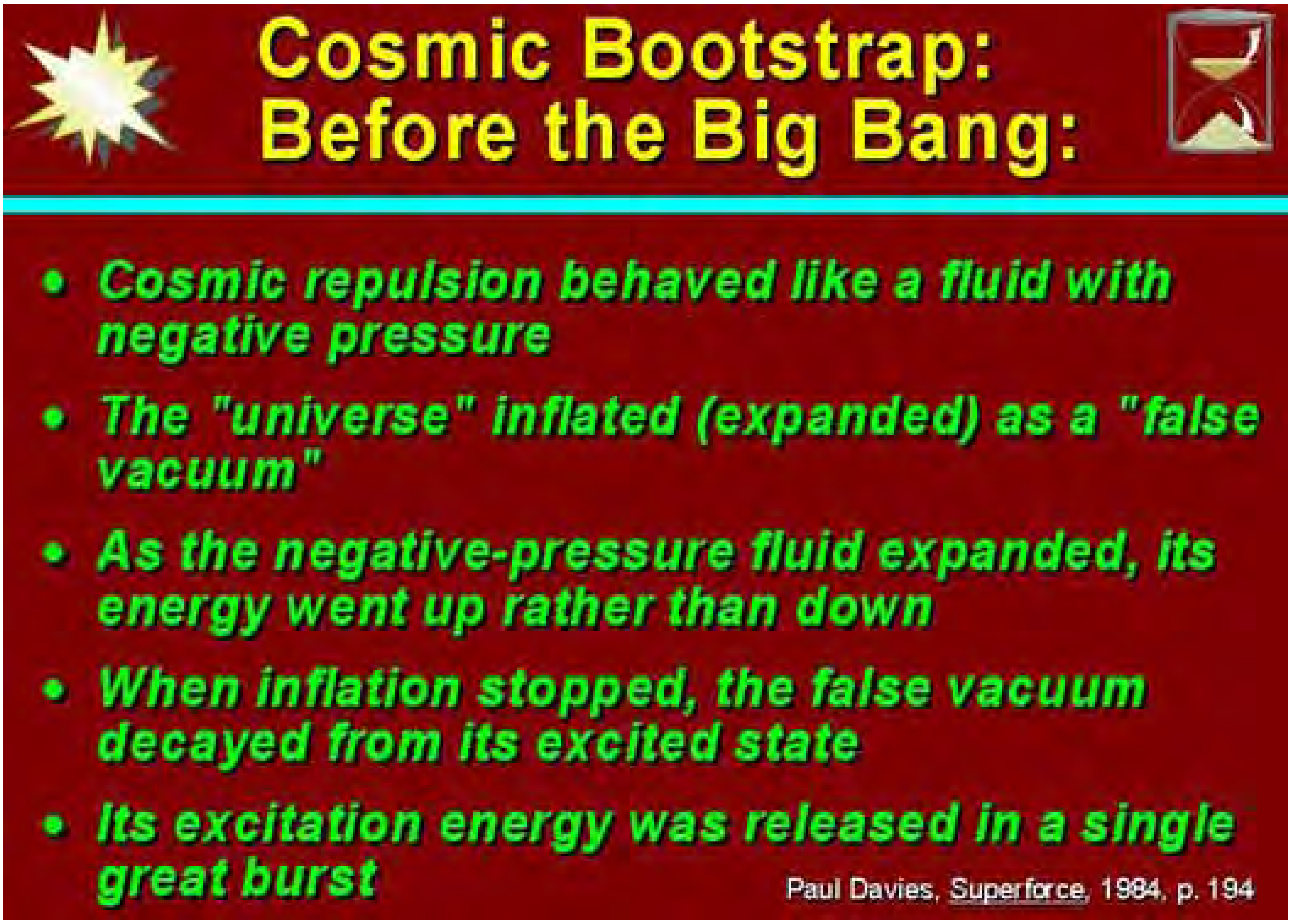

Cosmic Bootstrap: Before the Big Bang:

- *Cosmic repulsion behaved like a fluid with negative pressure*
- *The "universe" inflated (expanded) as a "false vacuum"*
- *As the negative-pressure fluid expanded, its energy went up rather than down*
- *When inflation stopped, the false vacuum decayed from its excited state*
- *Its excitation energy was released in a single great burst*

Paul Davies, Superforce, 1984, p. 194

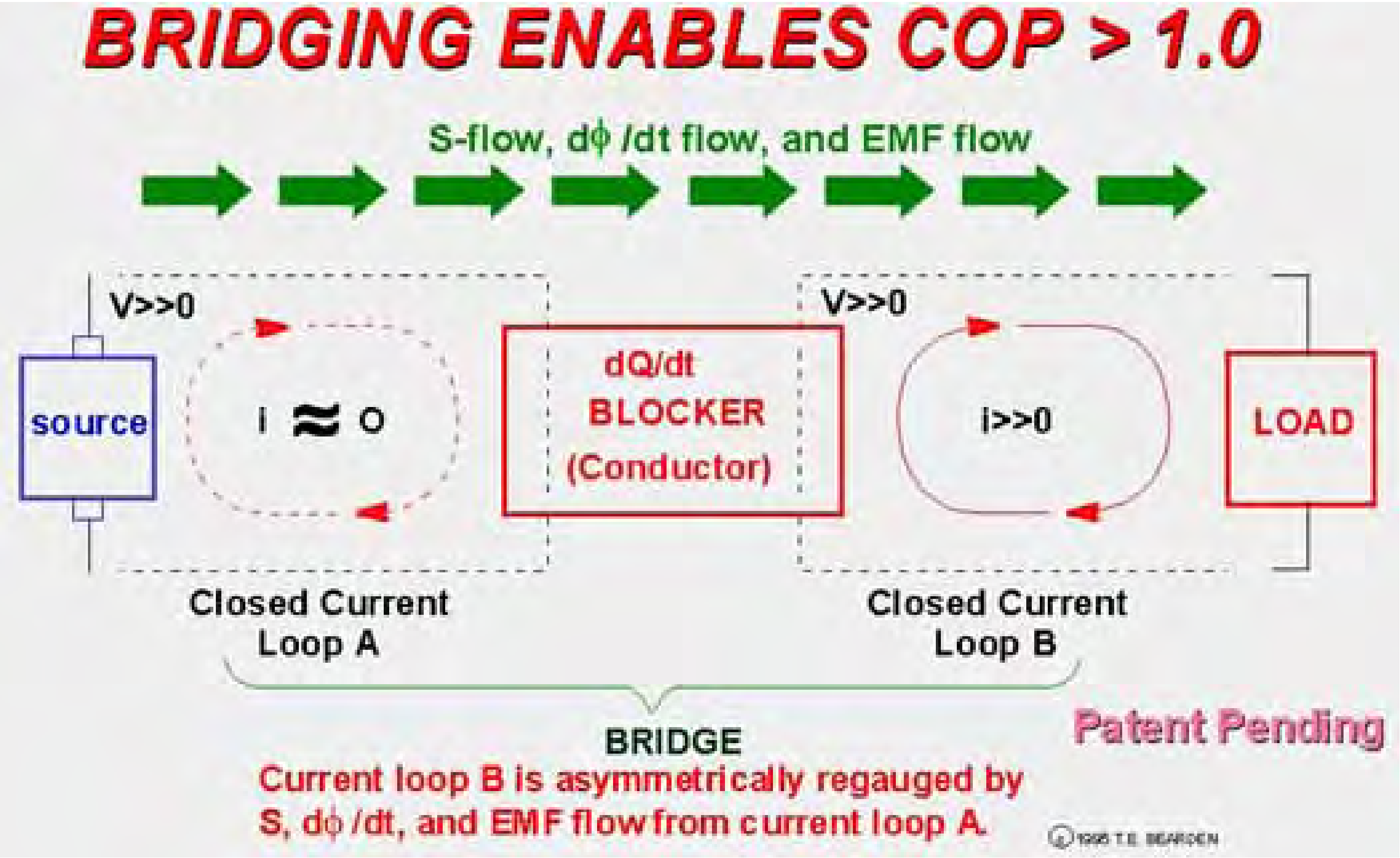

BRIDGING ENABLES COP > 1.0

S-flow, dϕ /dt flow, and EMF flow

V>>0

source

i ≈ 0

dQ/dt
BLOCKER
(Conductor)

V>>0

i>>0

LOAD

Closed Current
Loop A

Closed Current
Loop B

BRIDGE

Current loop B is asymmetrically regauged by
S, dϕ /dt, and EMF flow from current loop A.

Patent Pending

© 1996 T.E. BEARDEN

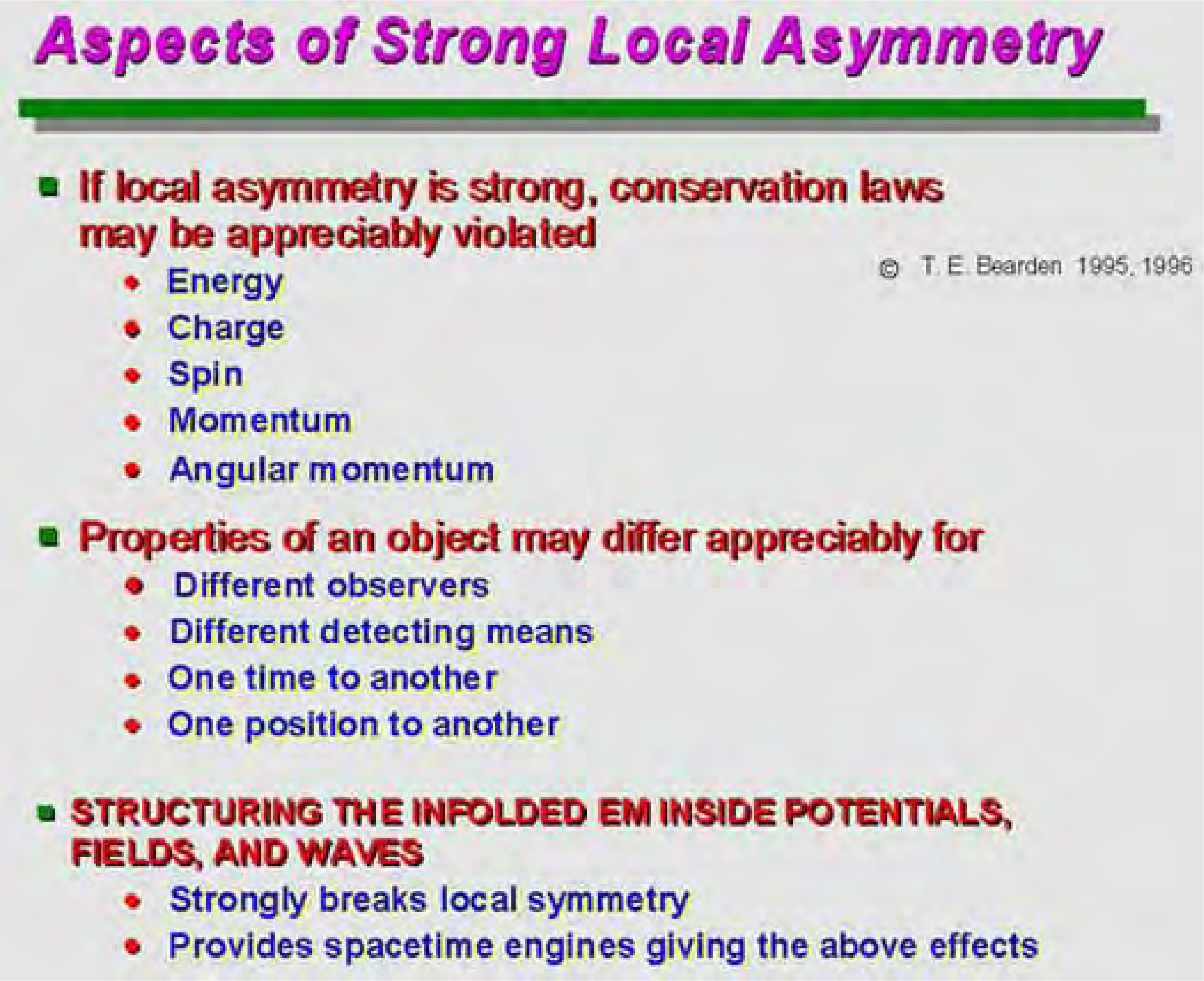

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

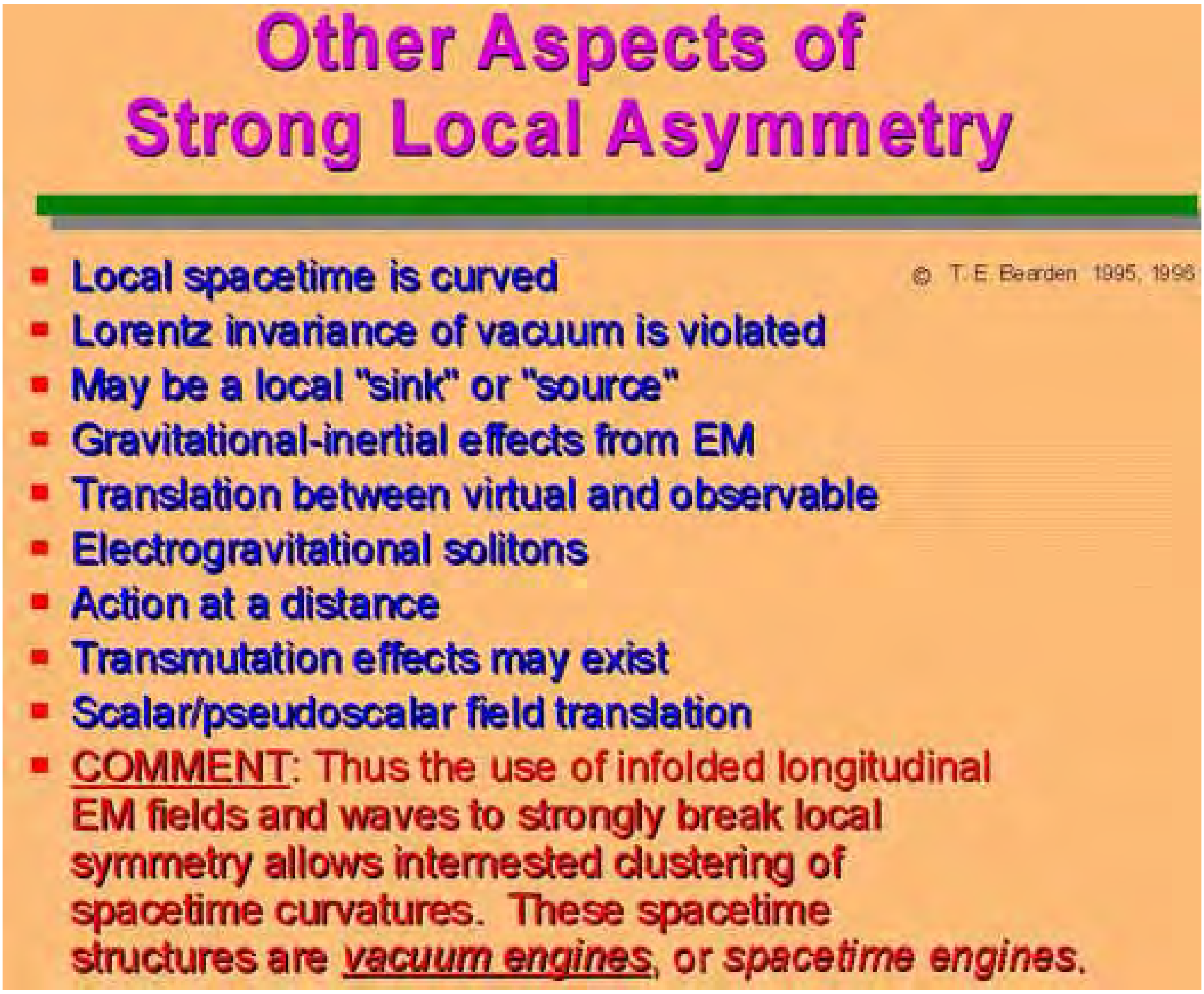

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

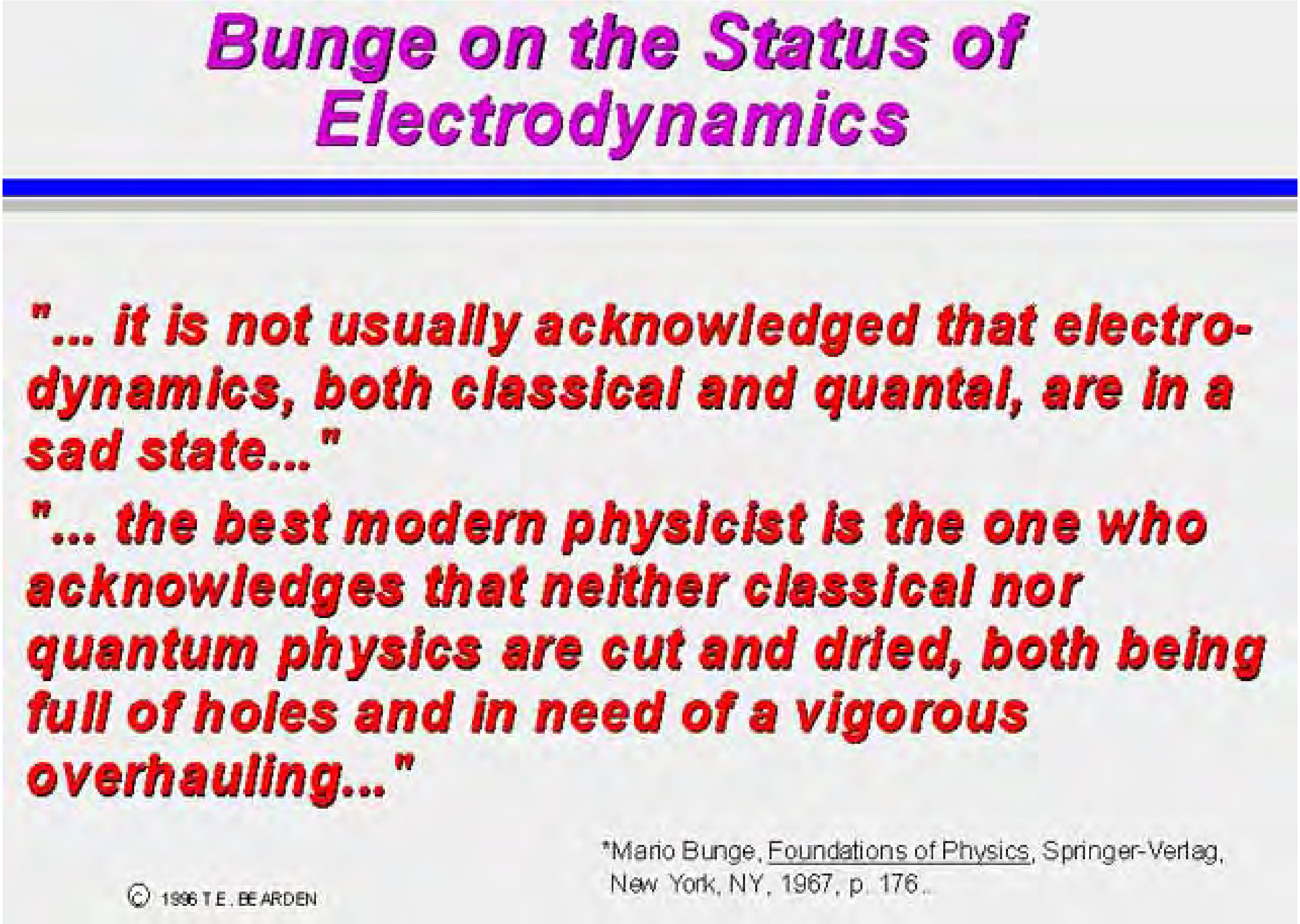

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

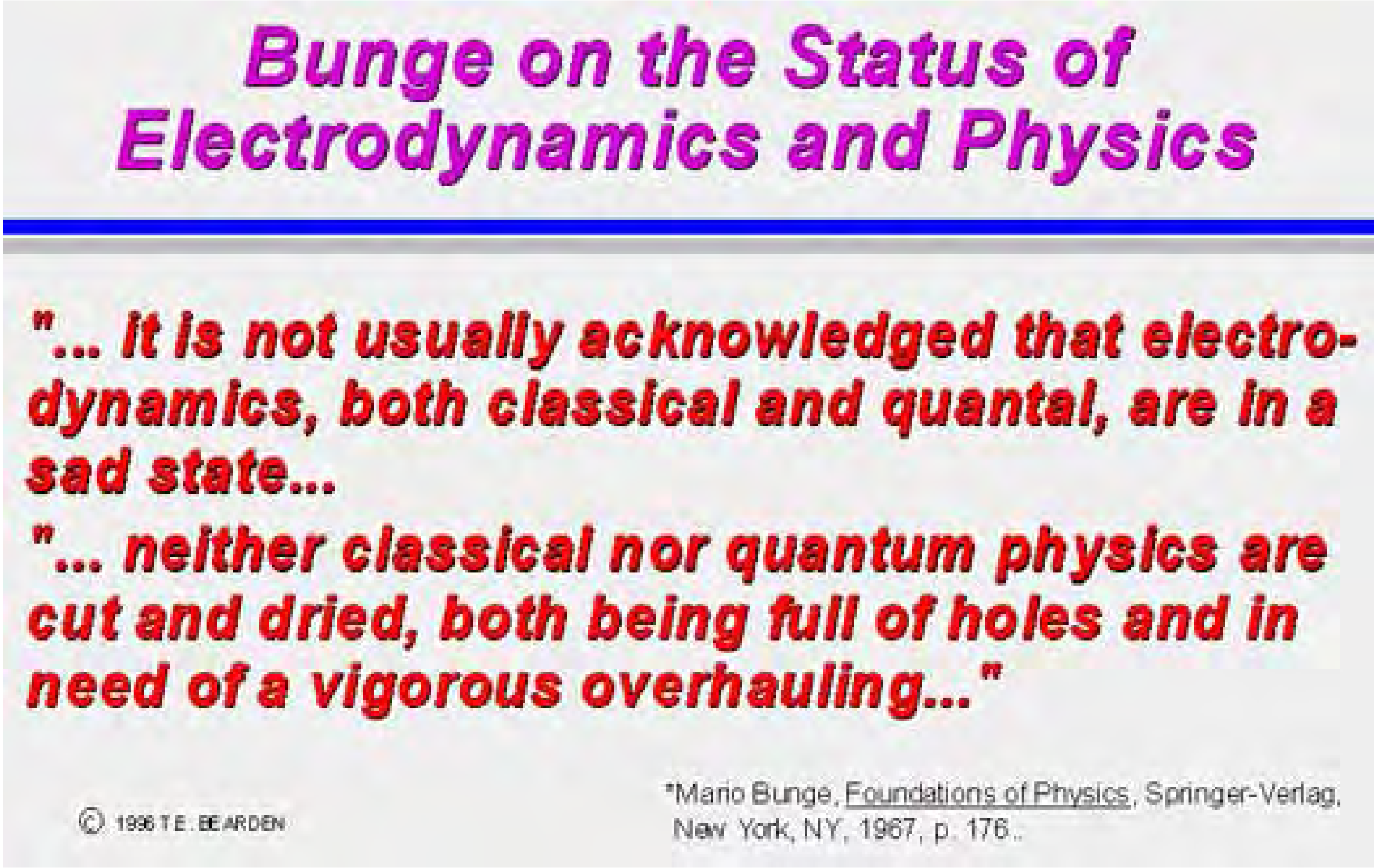

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© 1996 T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

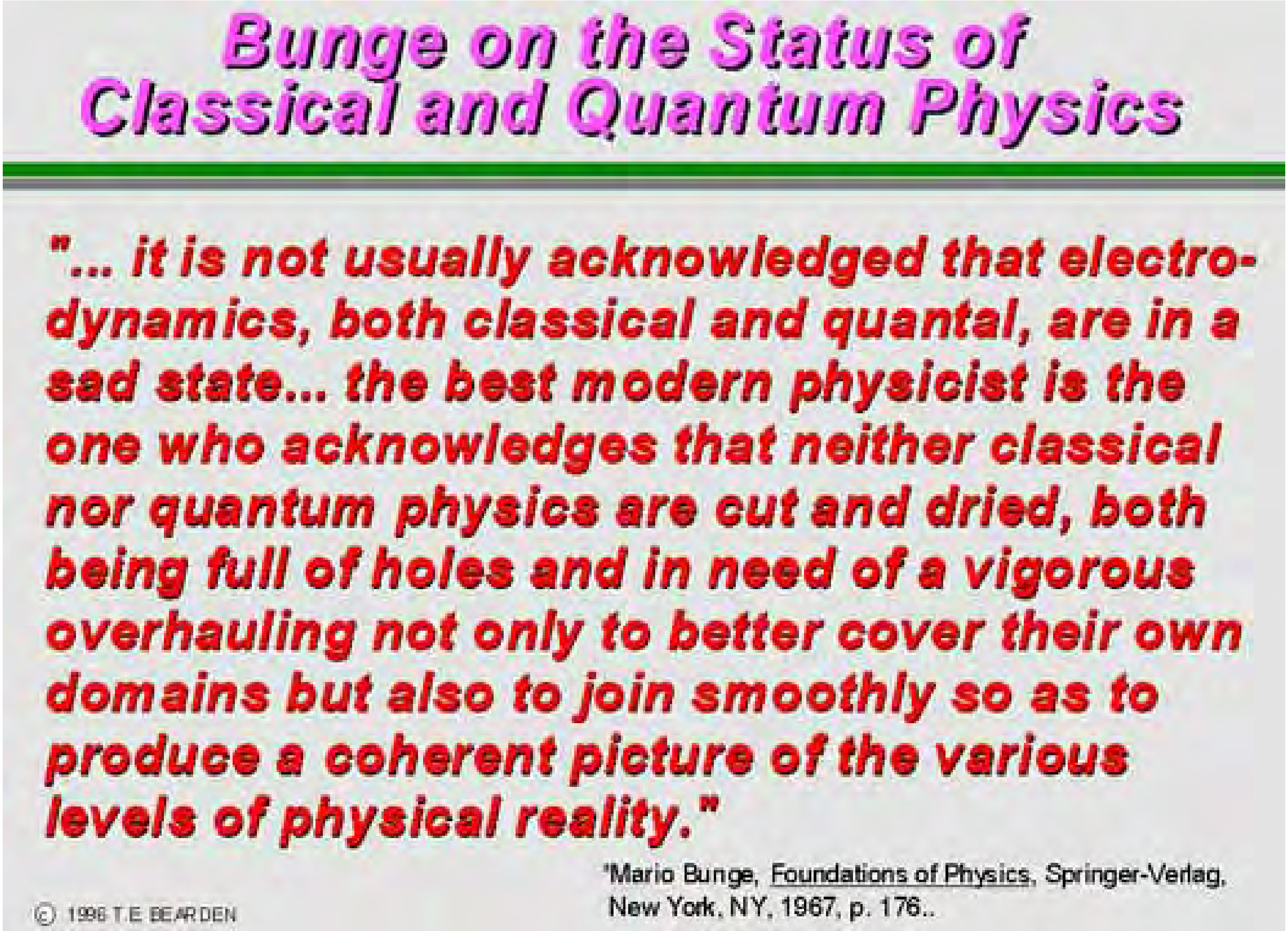

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

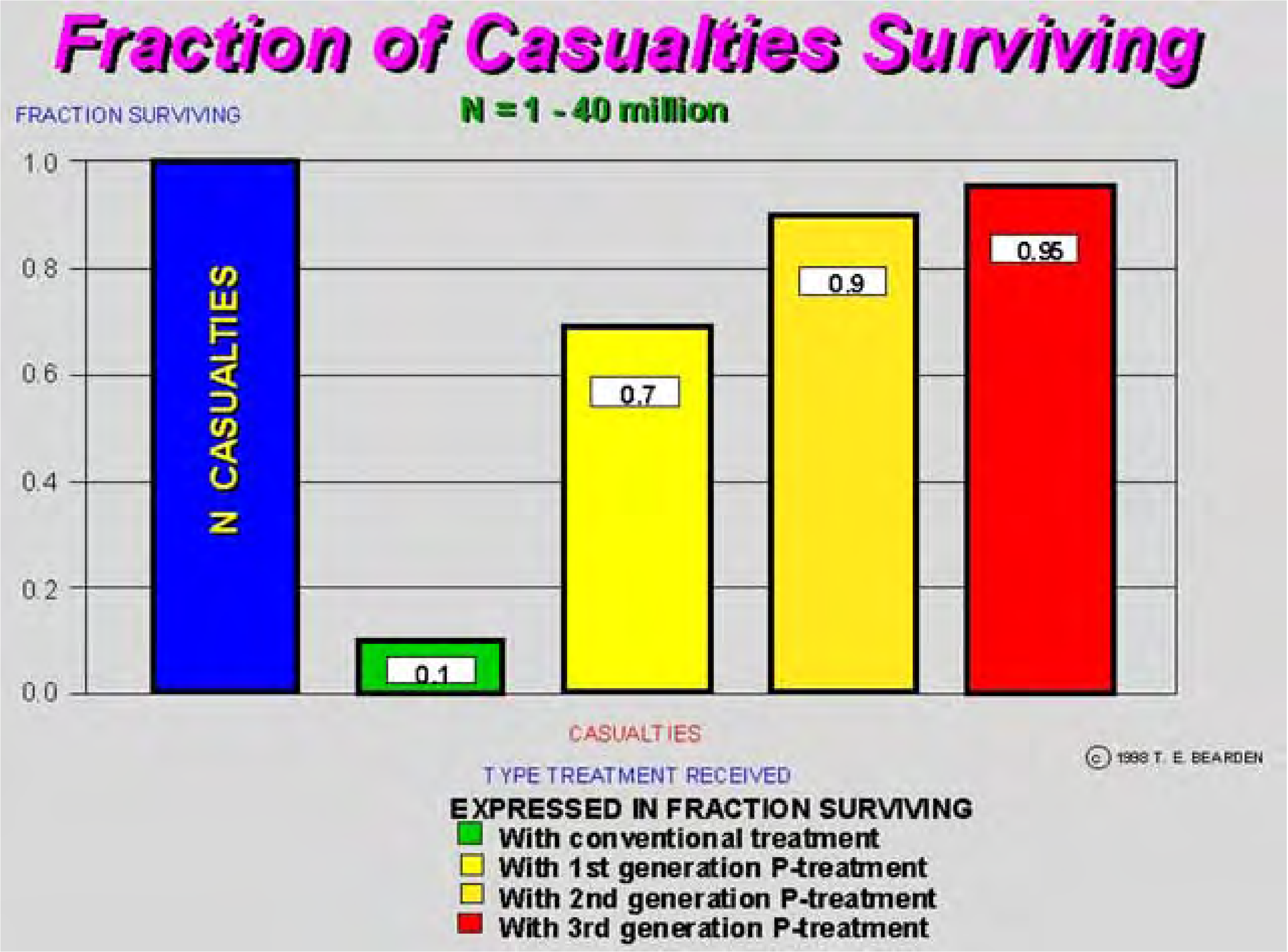
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
With conventional treatment
With 1st generation P-treatment
With 2nd generation P-treatment
With 3rd generation P-treatment

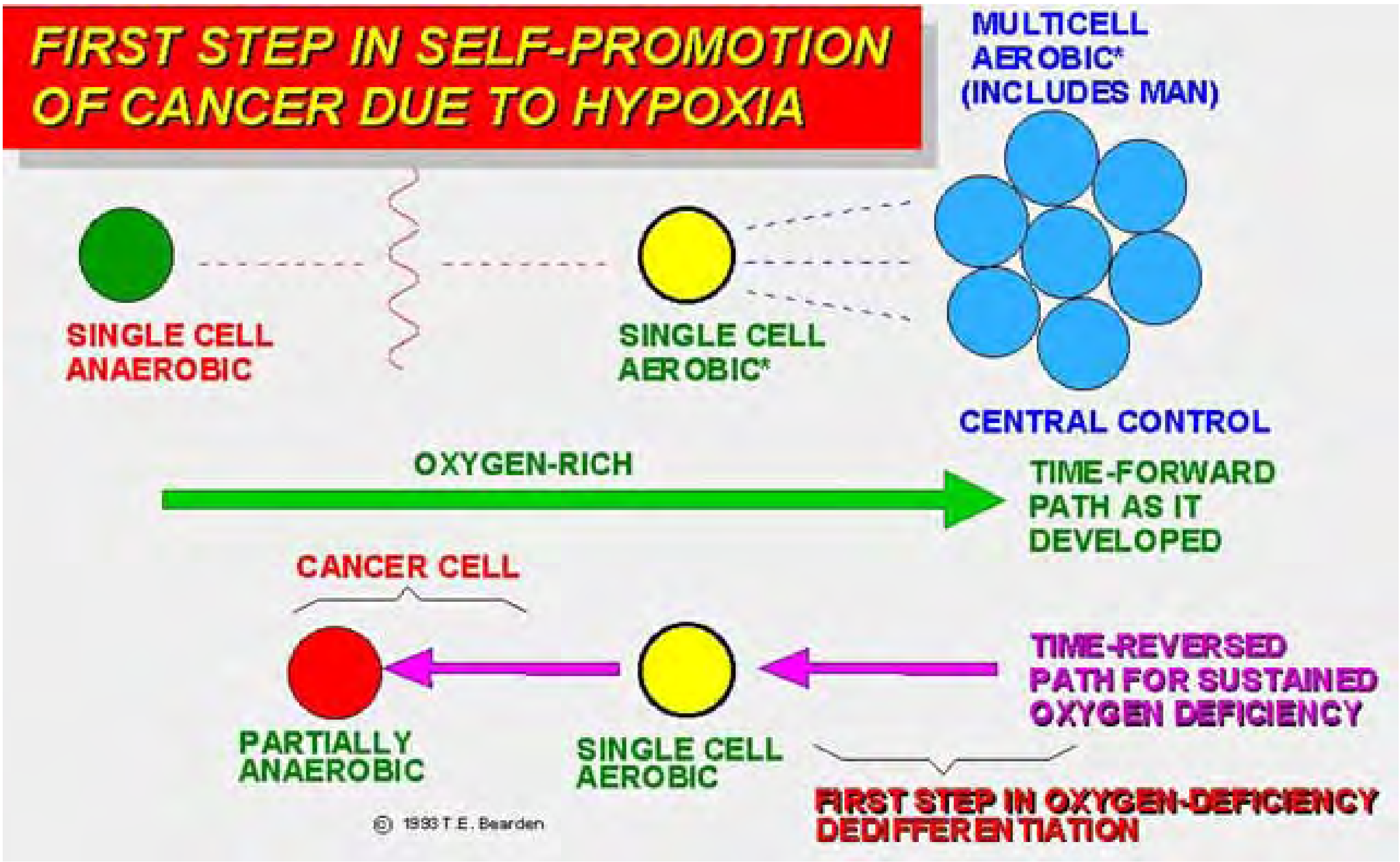

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)

SINGLE CELL
ANAEROBIC

SINGLE CELL
AEROBIC*

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD
PATH AS IT
DEVELOPED

CANCER CELL

TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY

PARTIALLY
ANAEROBIC

SINGLE CELL
AEROBIC

FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

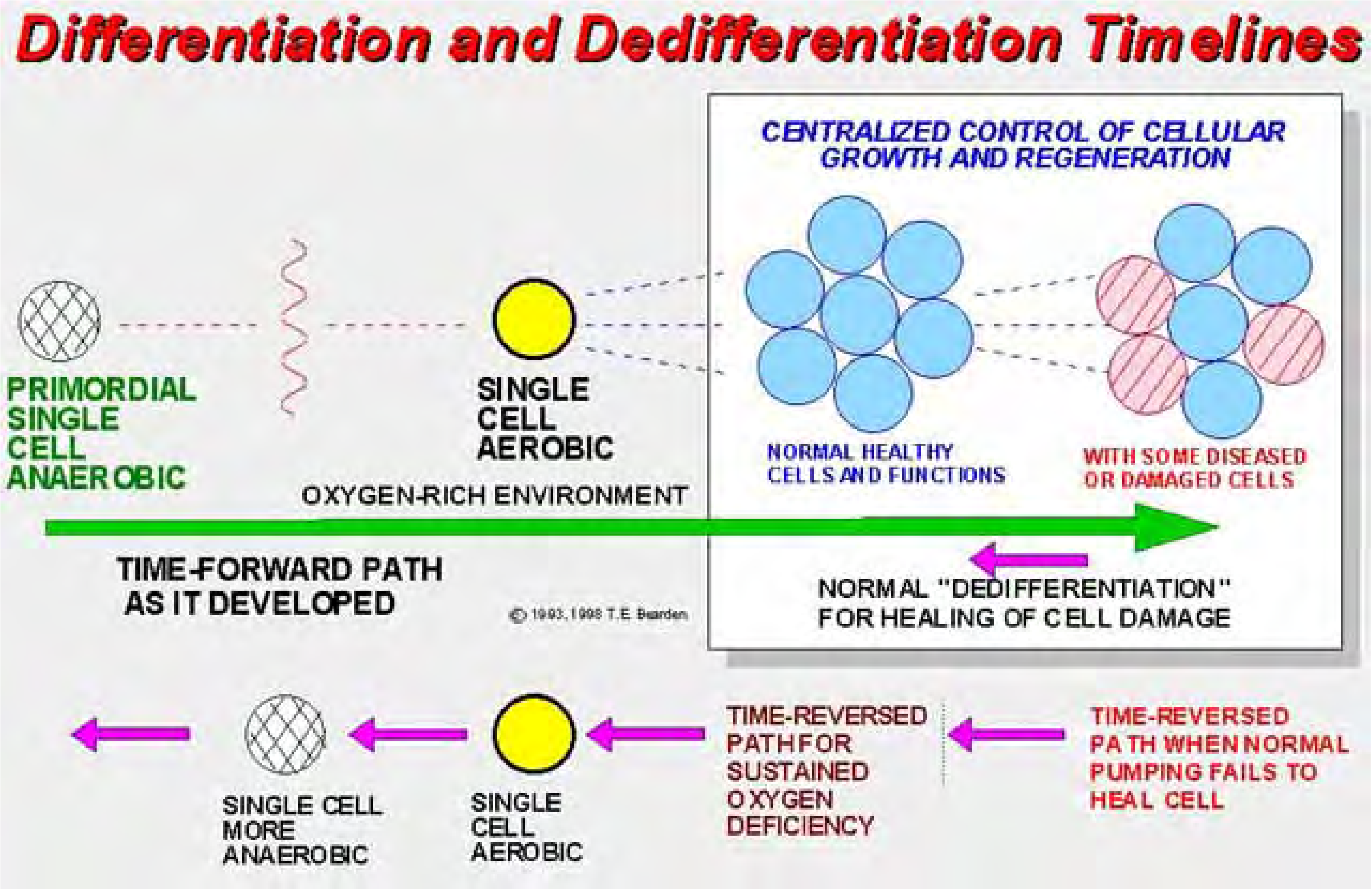
Differentiation and Dedifferentiation Timelines
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
OXYGEN-RICH ENVIRONMENT
TIME-FORWARD PATH AS IT DEVELOPED
© 1993, 1998 T.E. Bearden
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

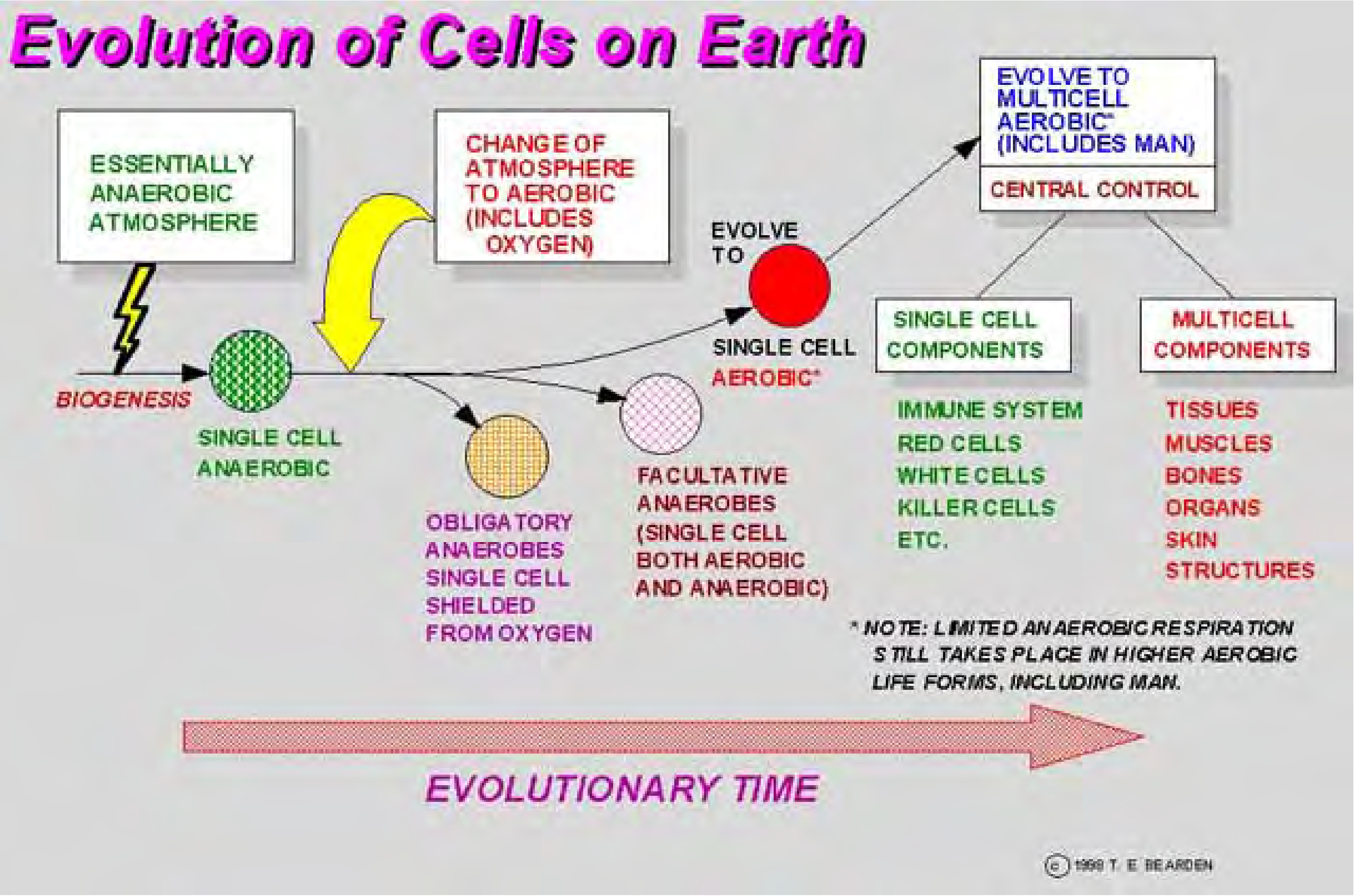

Evolution of Cells on Earth

ESSENTIALLY ANAEROBIC ATMOSPHERE

BIOGENESIS

SINGLE CELL ANAEROBIC

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

(c) 1999 T. E. BEARDEN

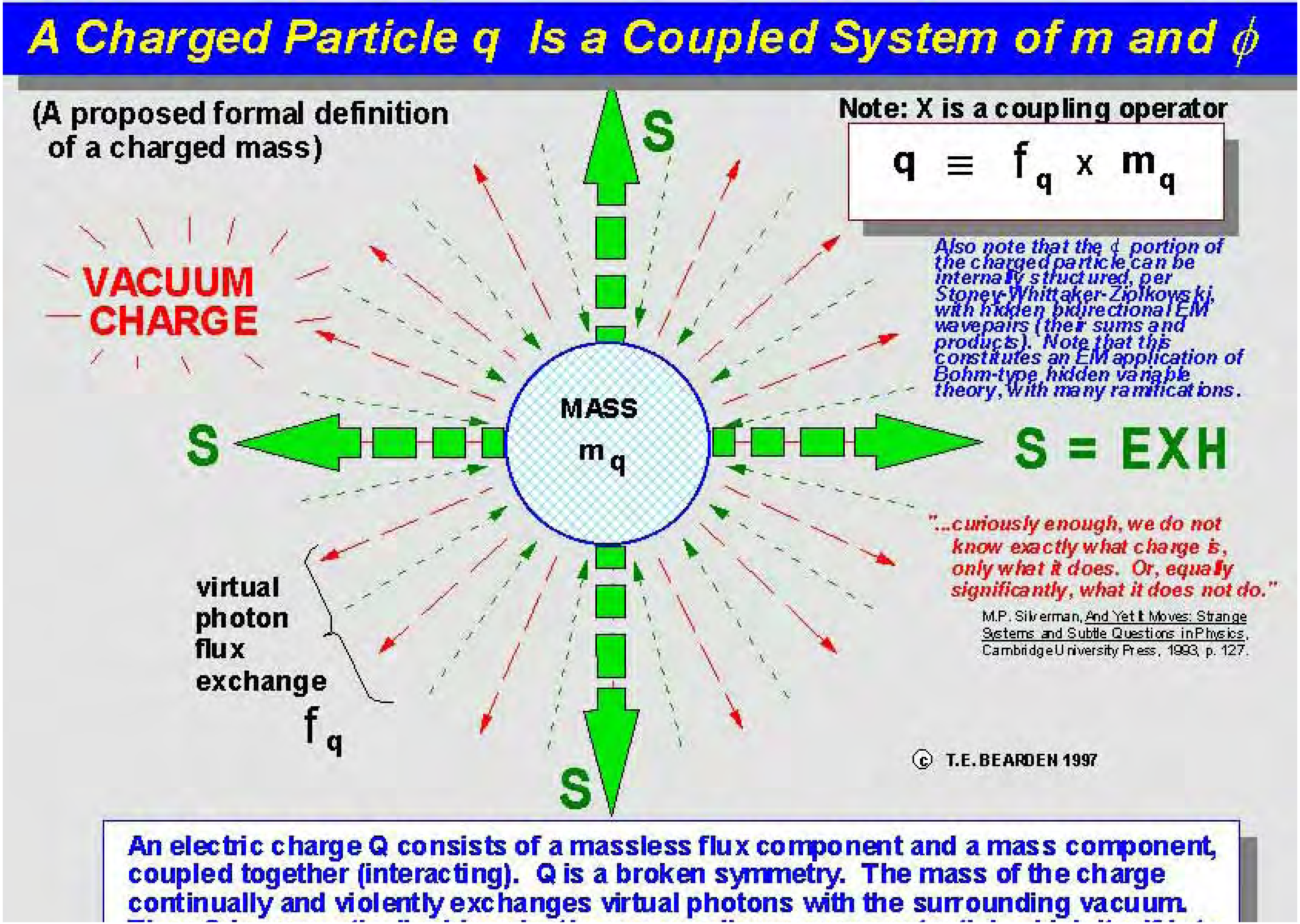

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

S = EXH

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

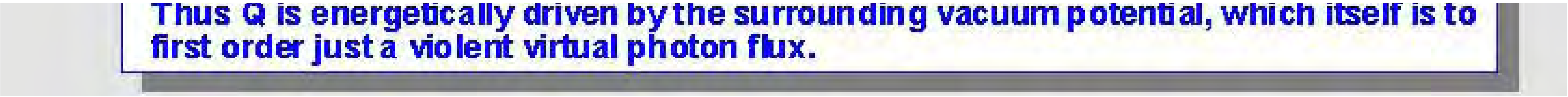

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

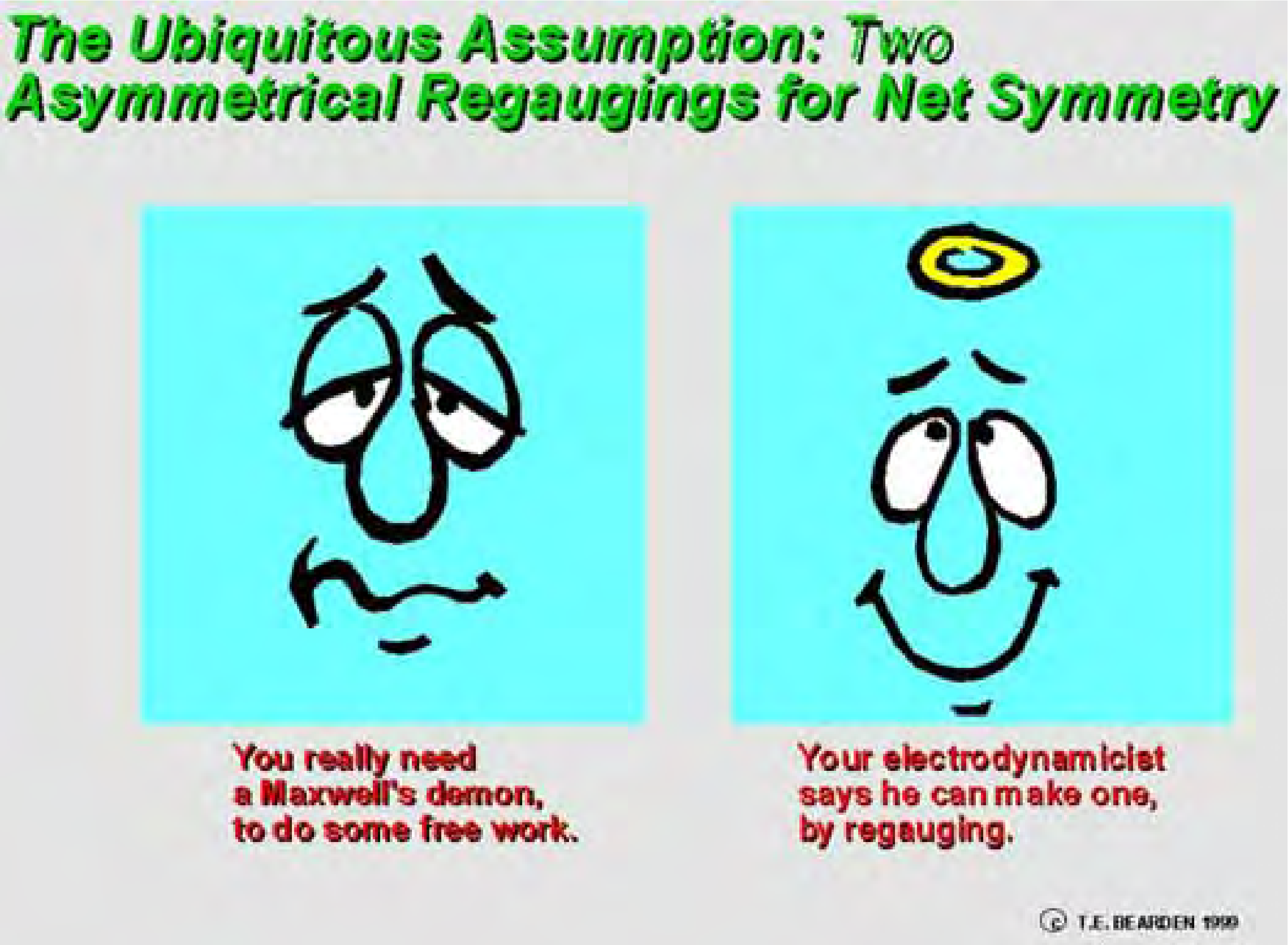

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

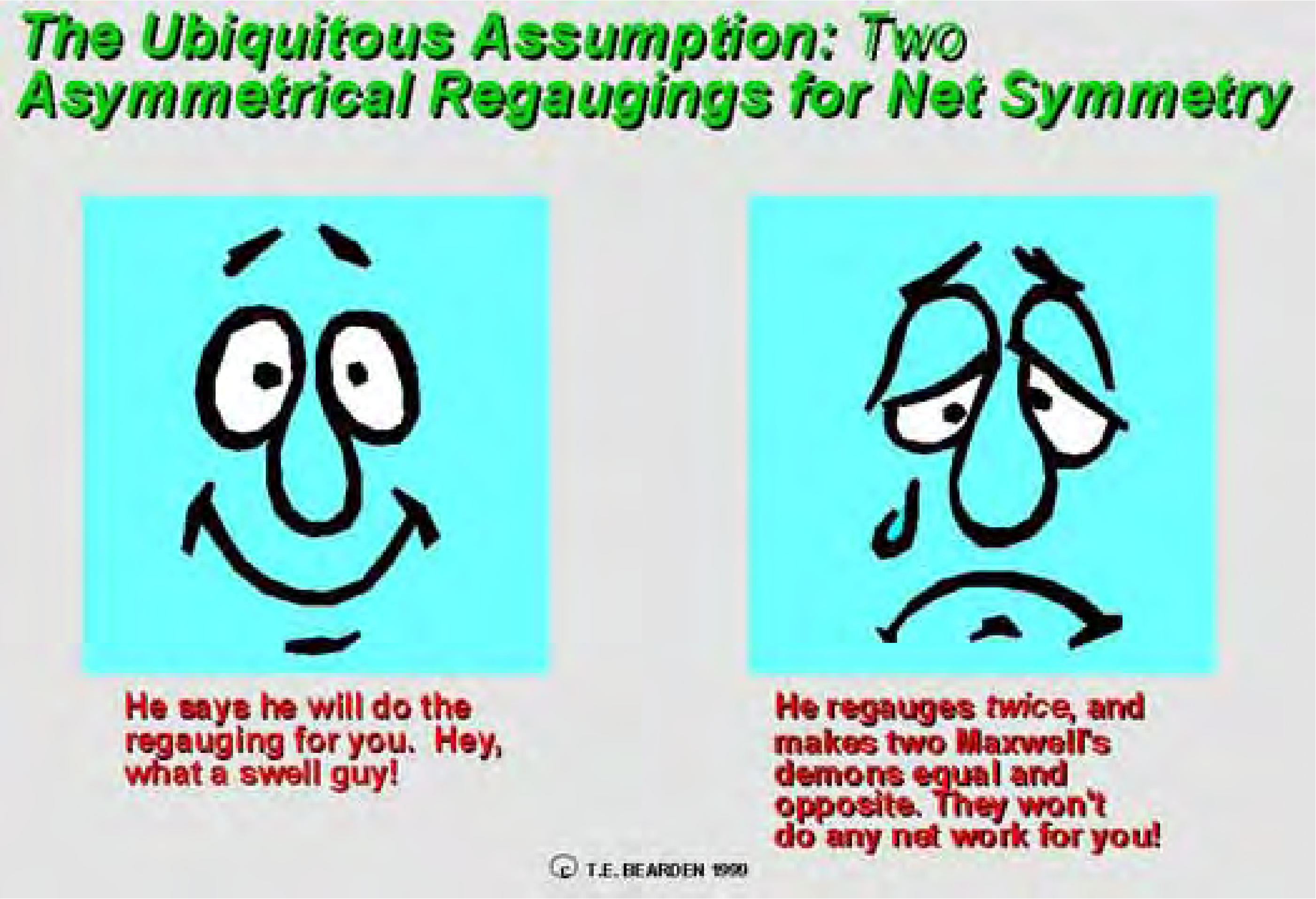

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

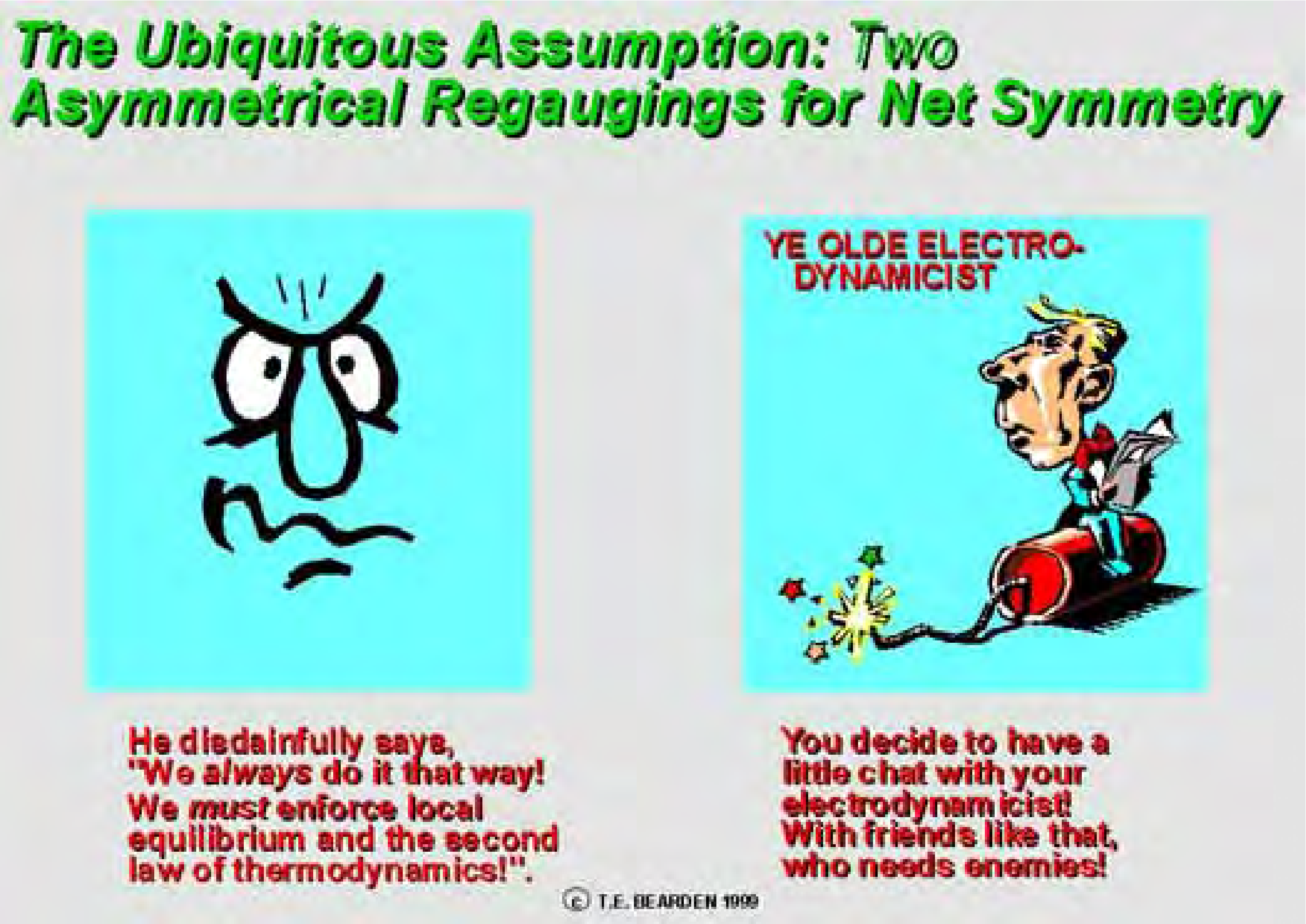

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

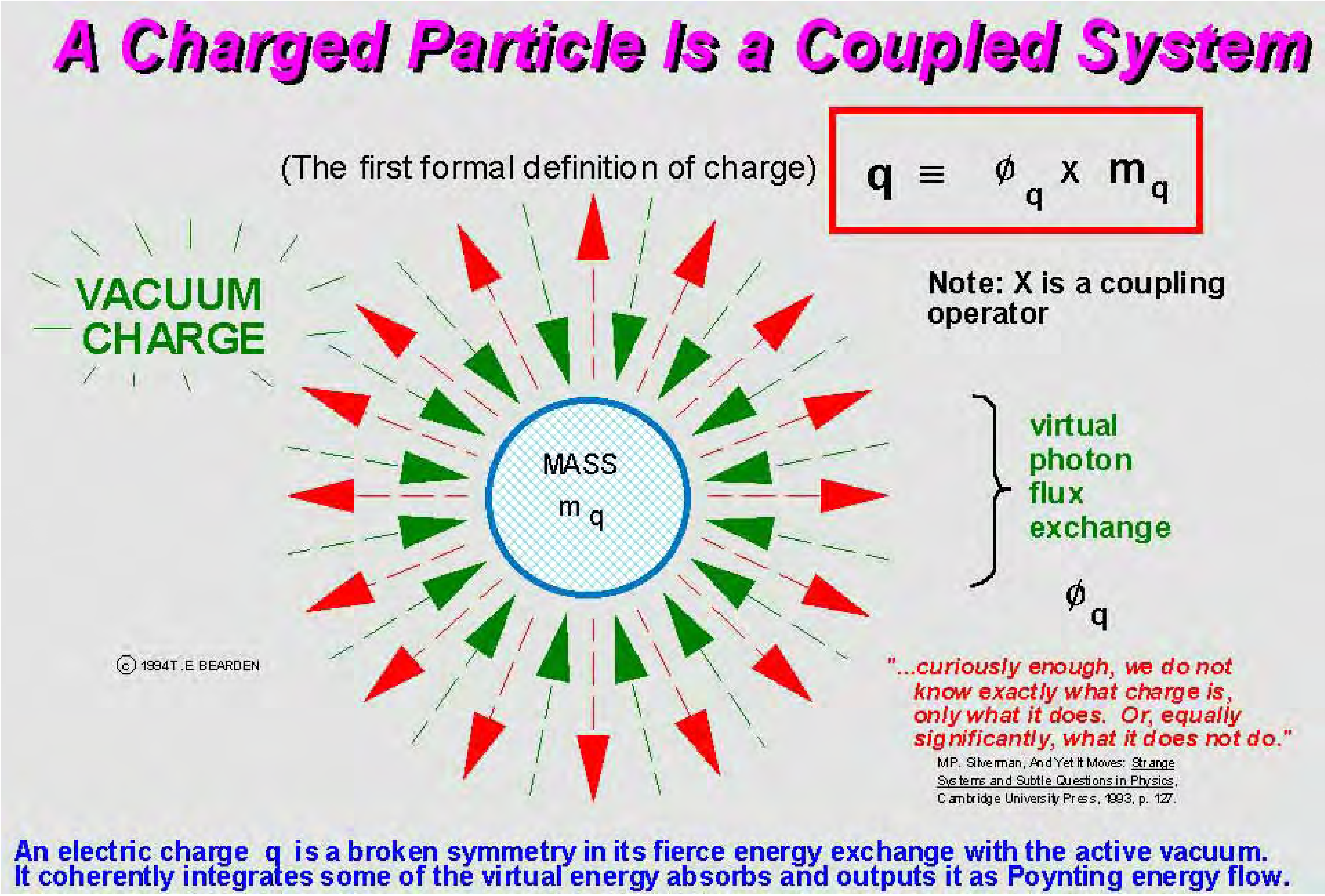

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$q \equiv \phi_q \; \mathrm{x} \; m_q$$

Note: X is a coupling operator

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

Chung's Carbon Filament Negative Resistor

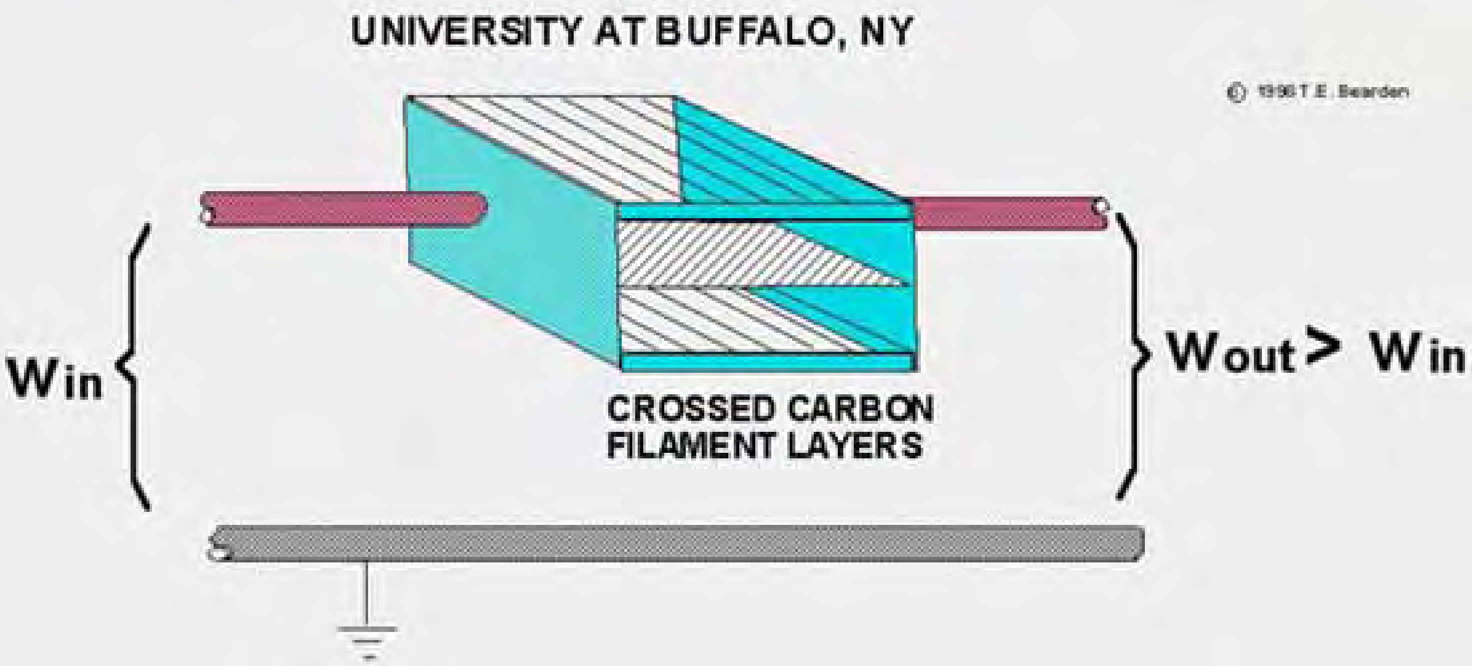
UNIVERSITY AT BUFFALO, NY
Win
CROSSED CARBON FILAMENT LAYERS
Wout > Win

© 1996 T.E. Bearden

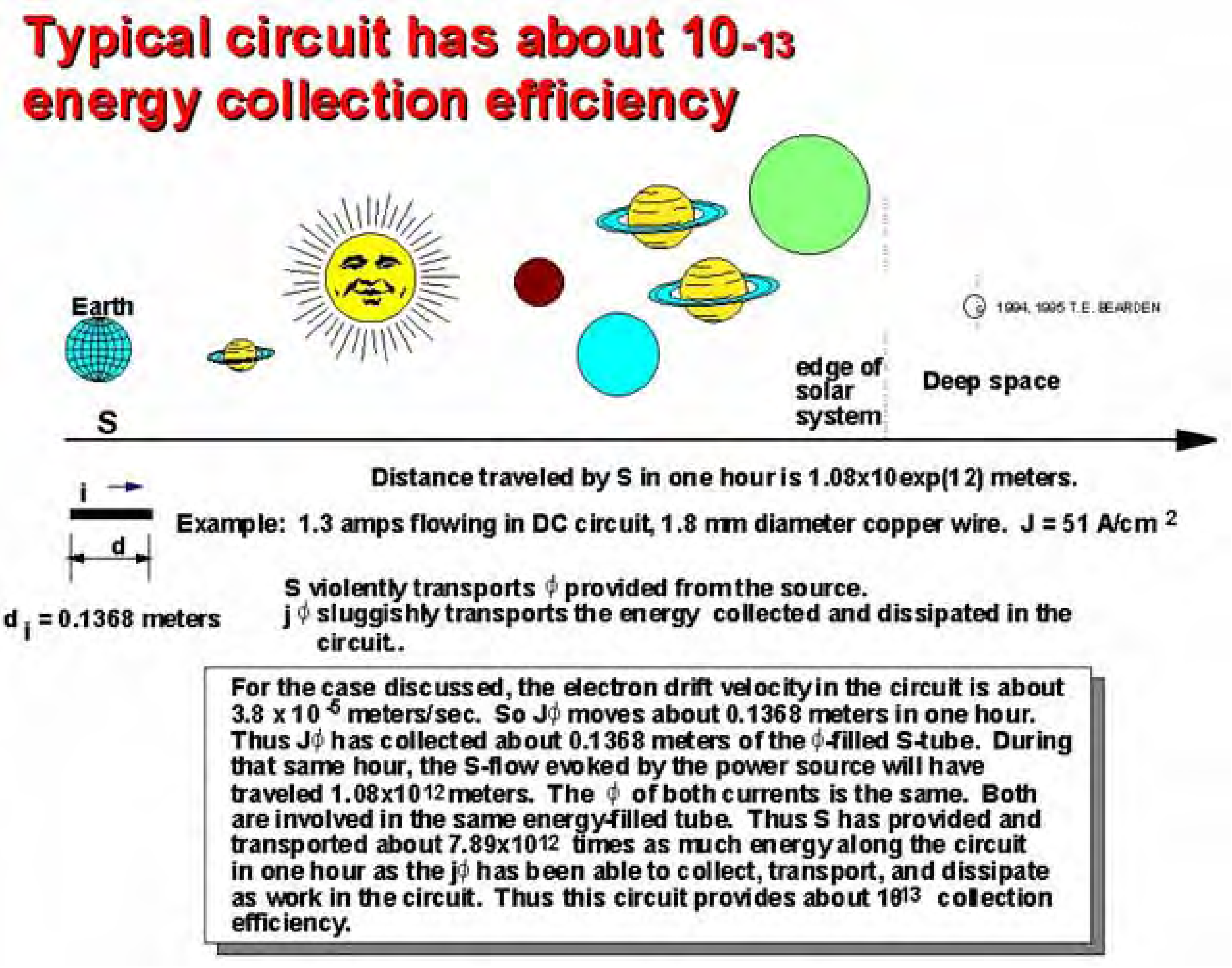

Typical circuit has about 10_{-13} energy collection efficiency

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. J = 51 A/cm ^{2}

d_i = 0.1368 meters

S violently transports ϕ provided from the source.
j ϕ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8 x 10 $^{-5}$ meters/sec. So Jϕ moves about 0.1368 meters in one hour. Thus Jϕ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled 1.08x10^{12} meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about 7.89x10^{12} times as much energy along the circuit in one hour as the jϕ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

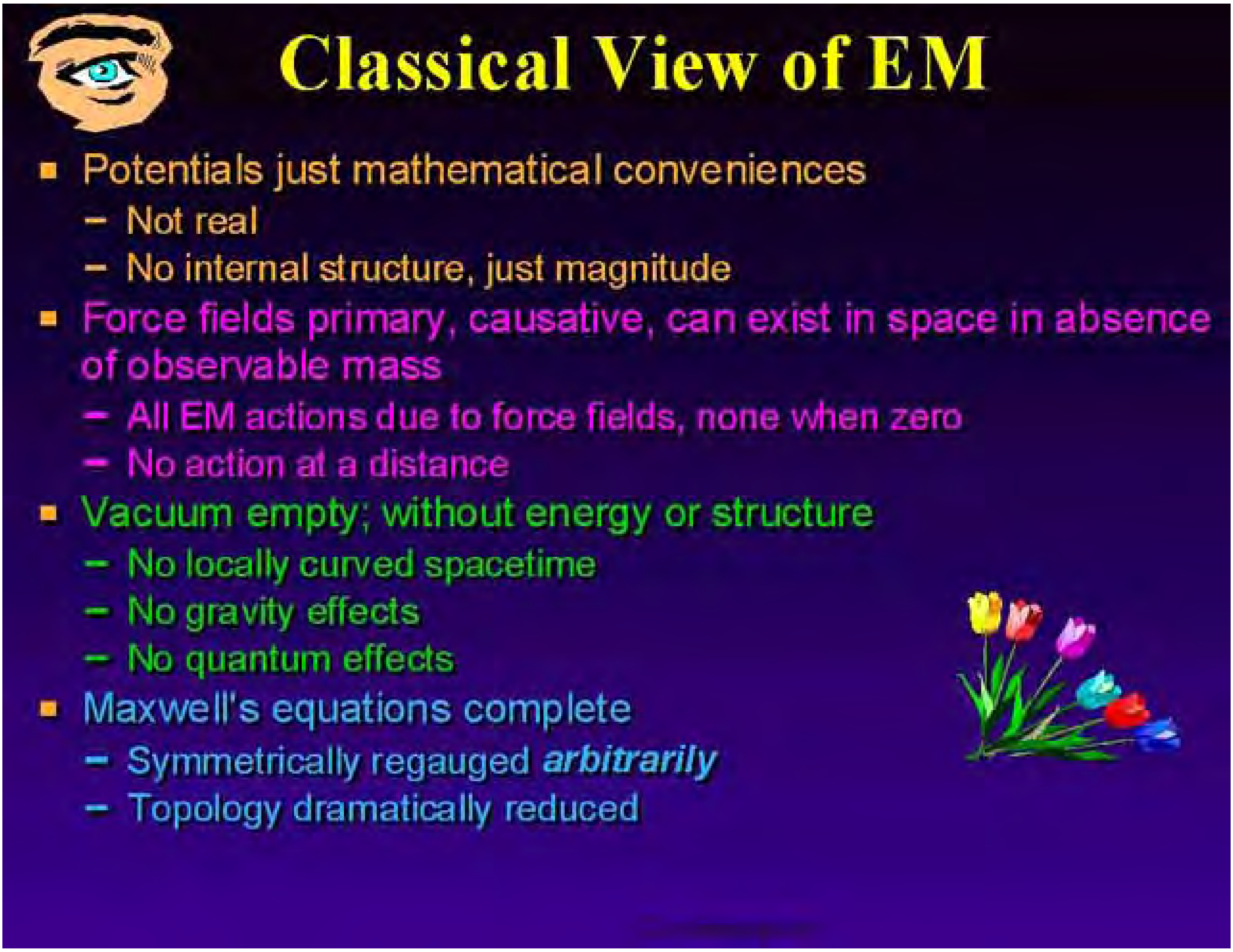

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

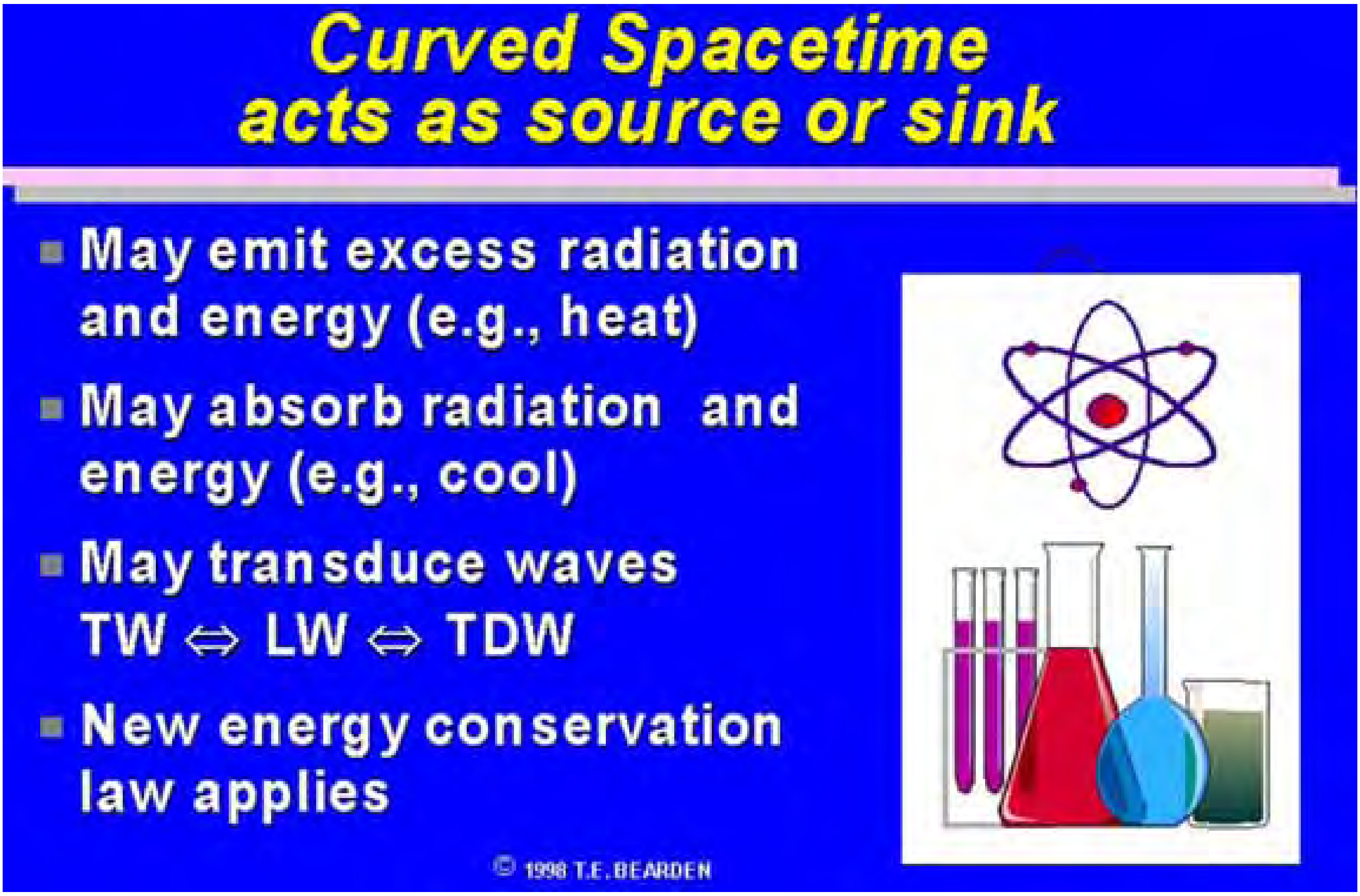

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves
  TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E.BEARDEN

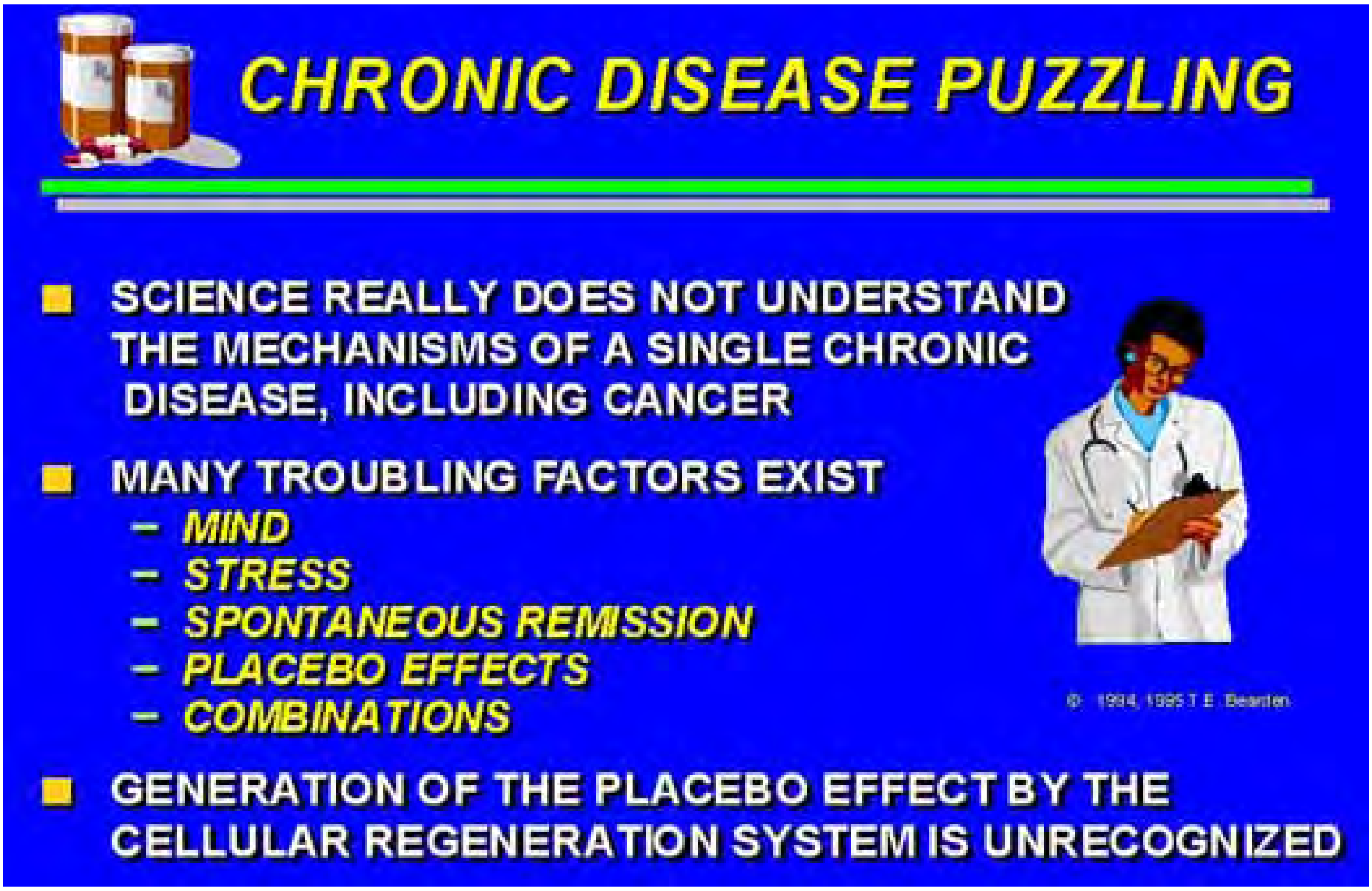

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

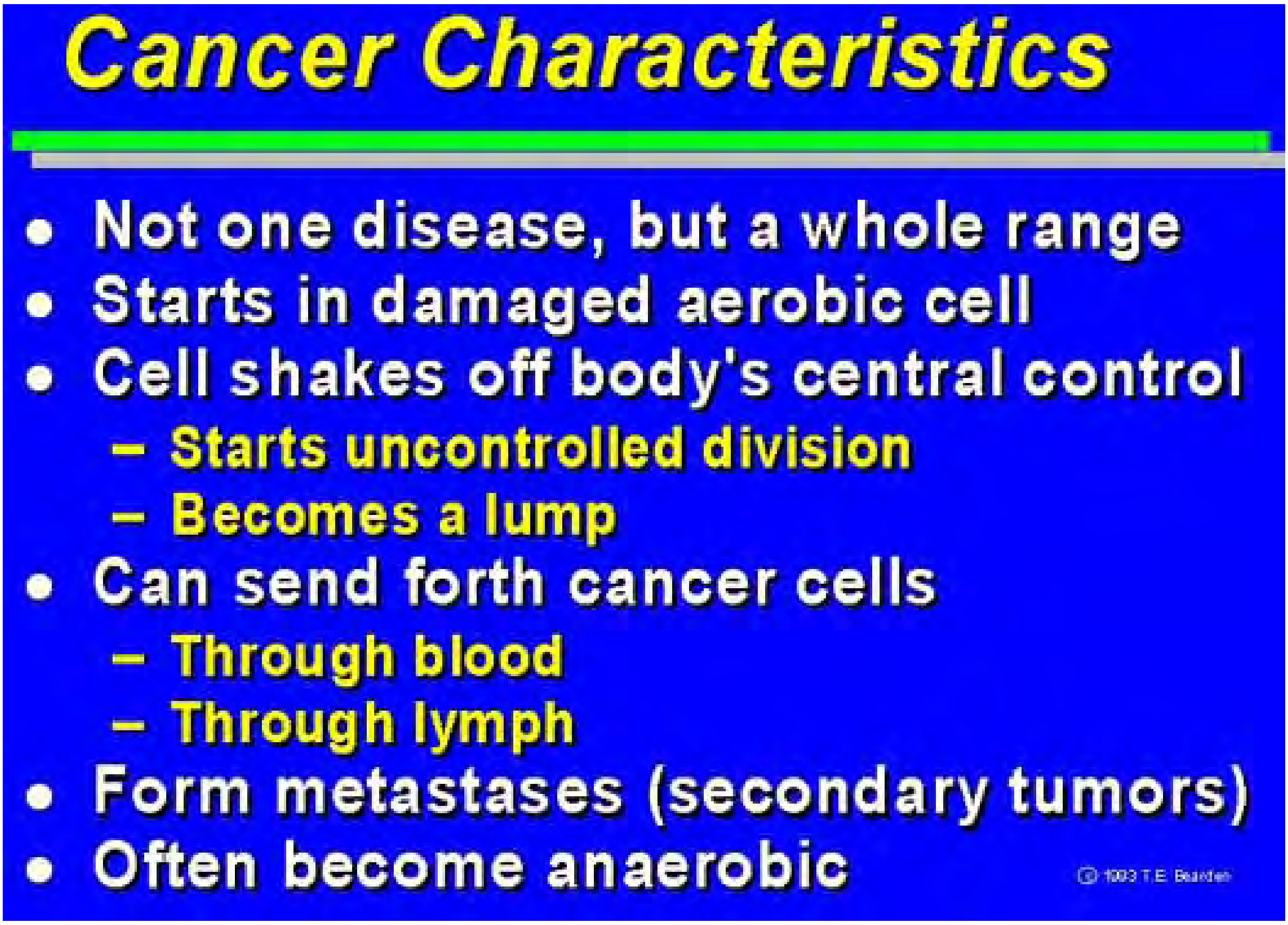

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

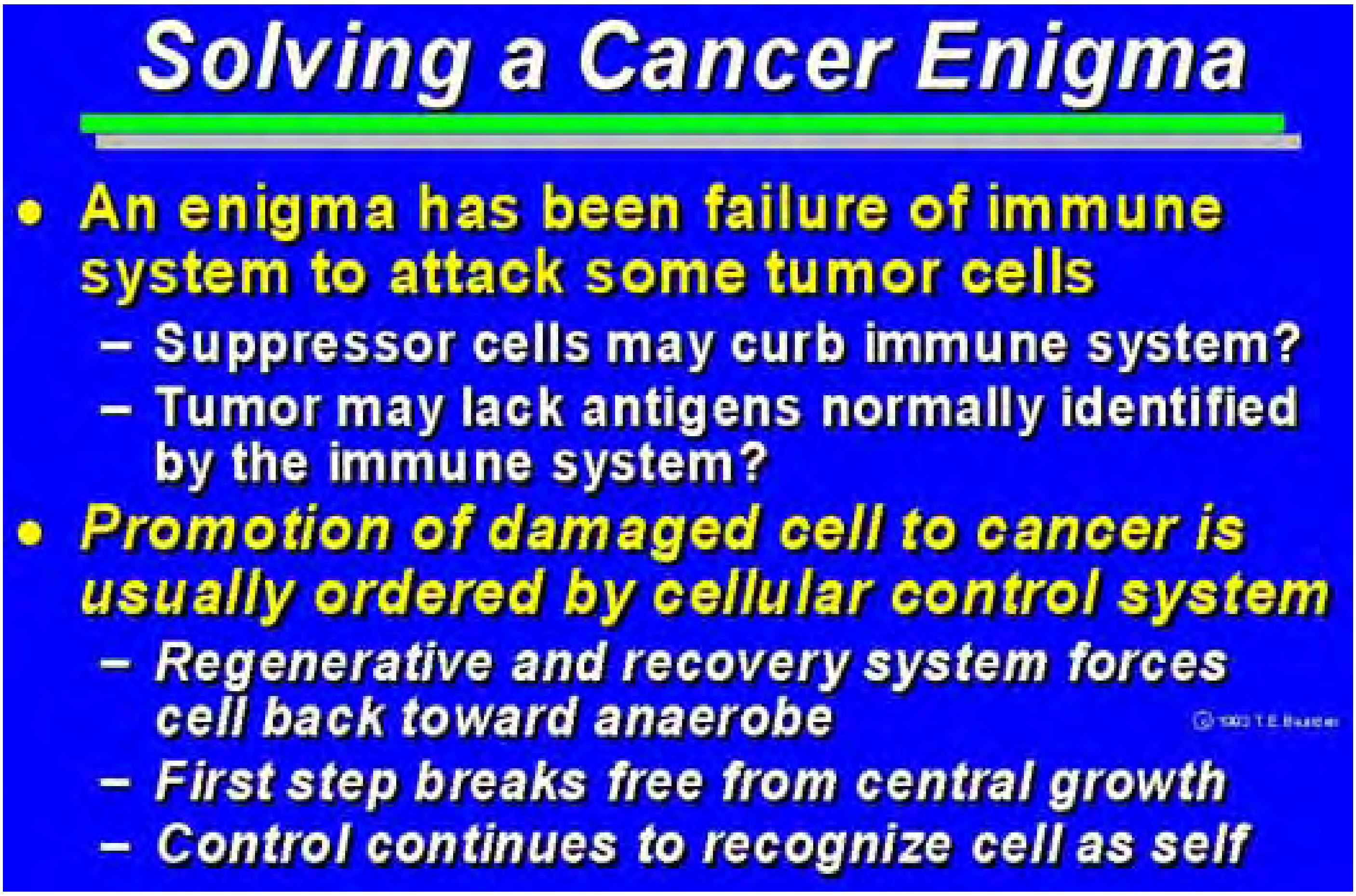

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

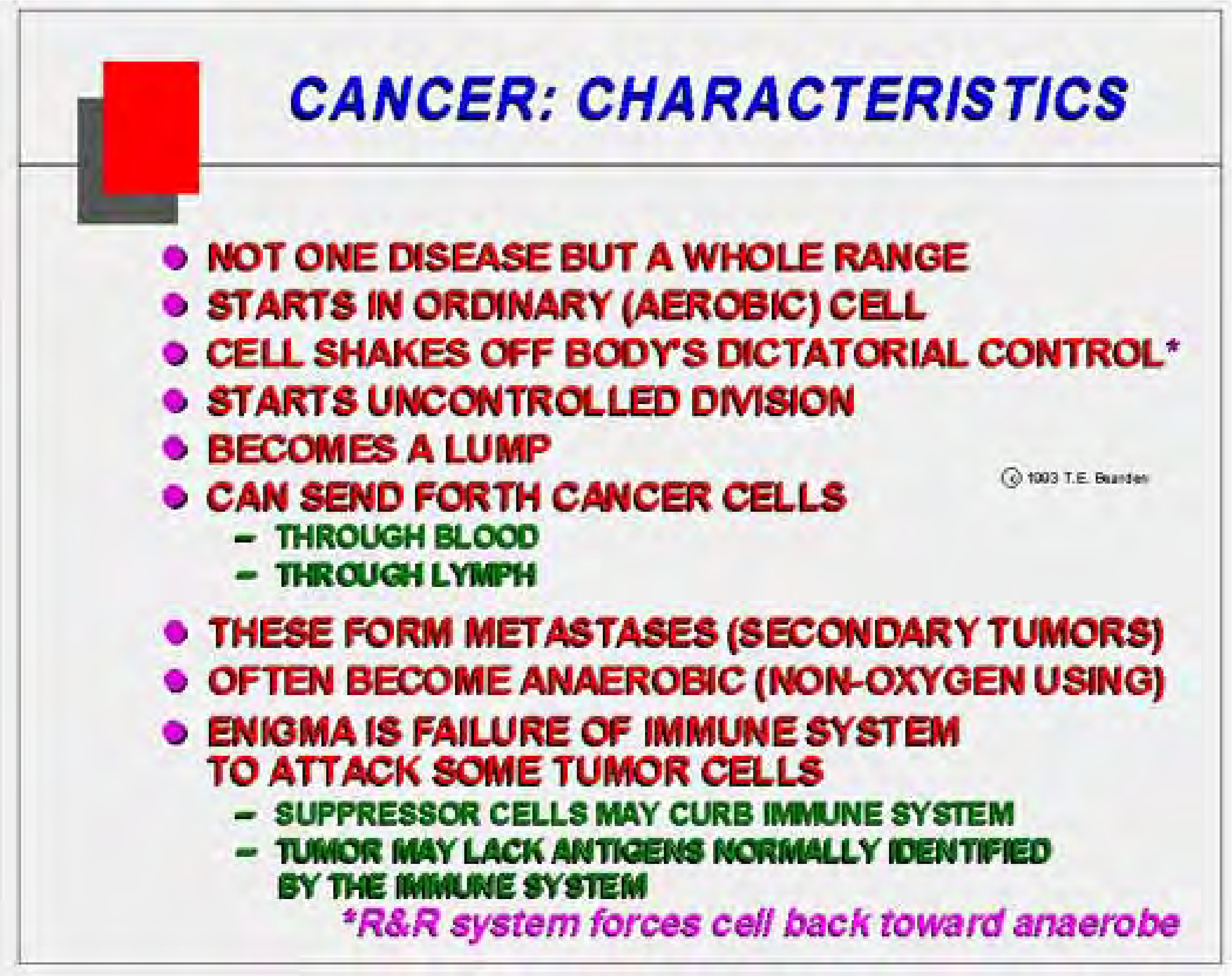

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

*R&R system forces cell back toward anaerobe

© 1993 T.E. Bearden

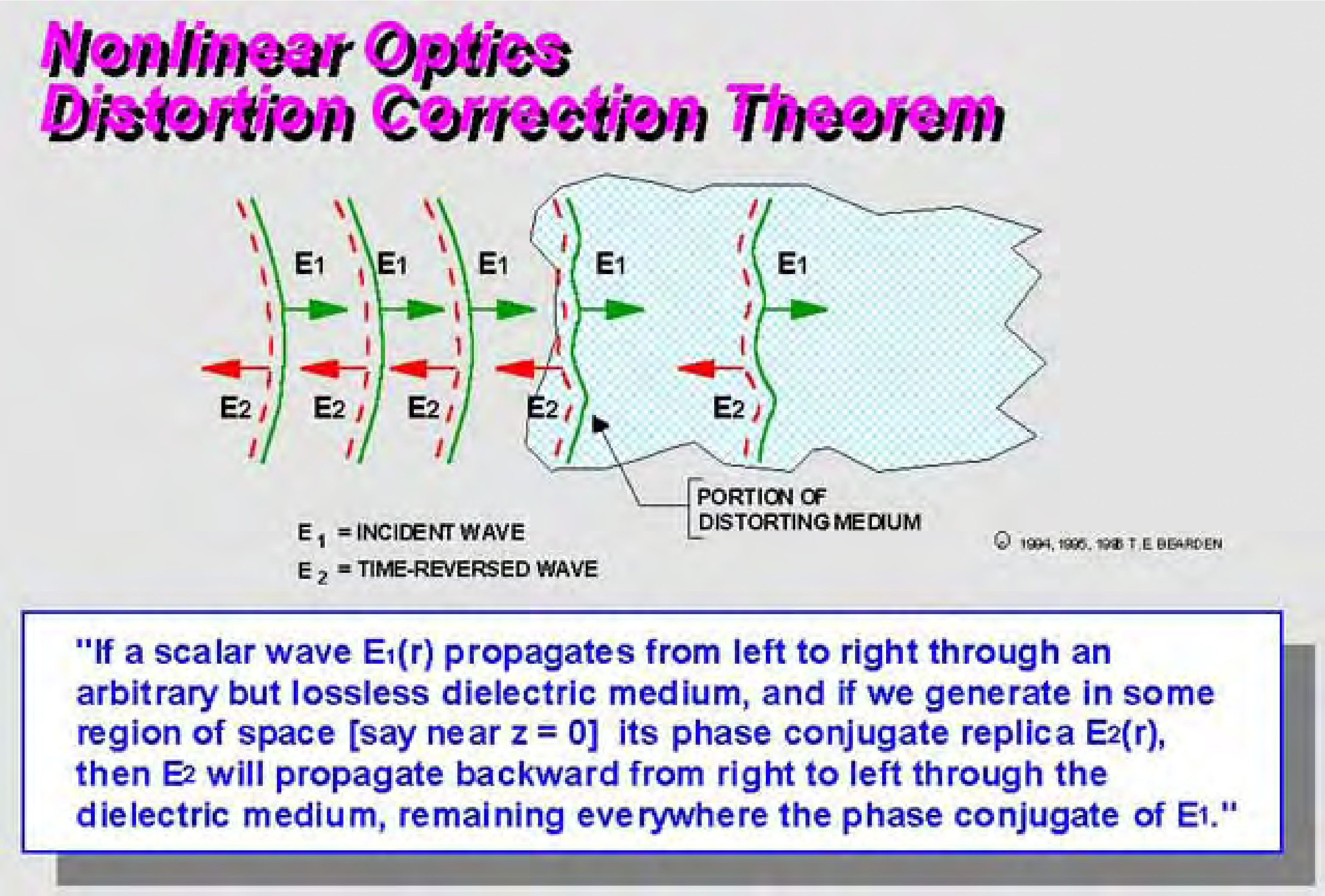

Nonlinear Optics
Distortion Correction Theorem

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T. E BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

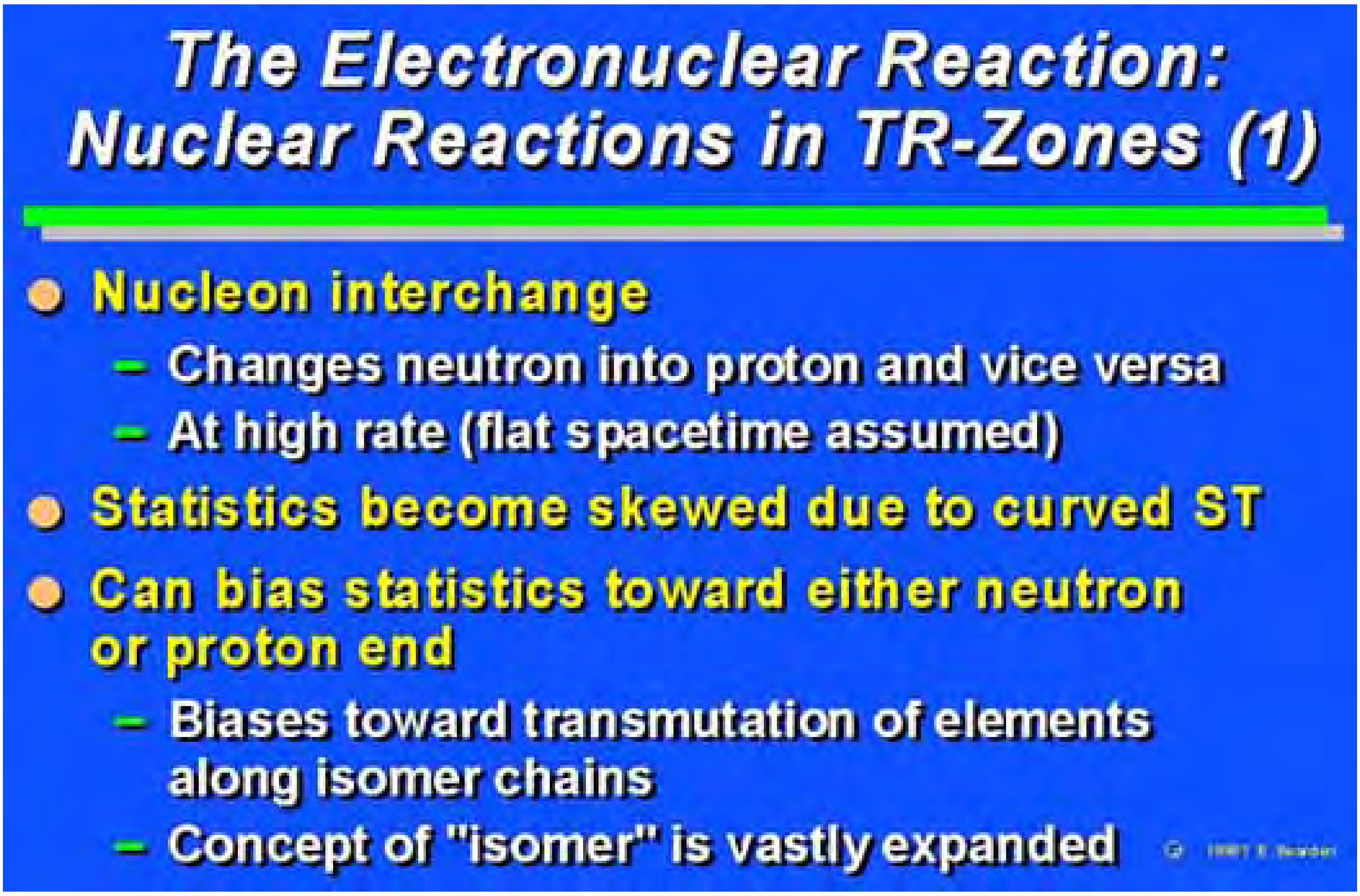

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- **Nucleon interchange**
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- **Statistics become skewed due to curved ST**
- **Can bias statistics toward either neutron or proton end**
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

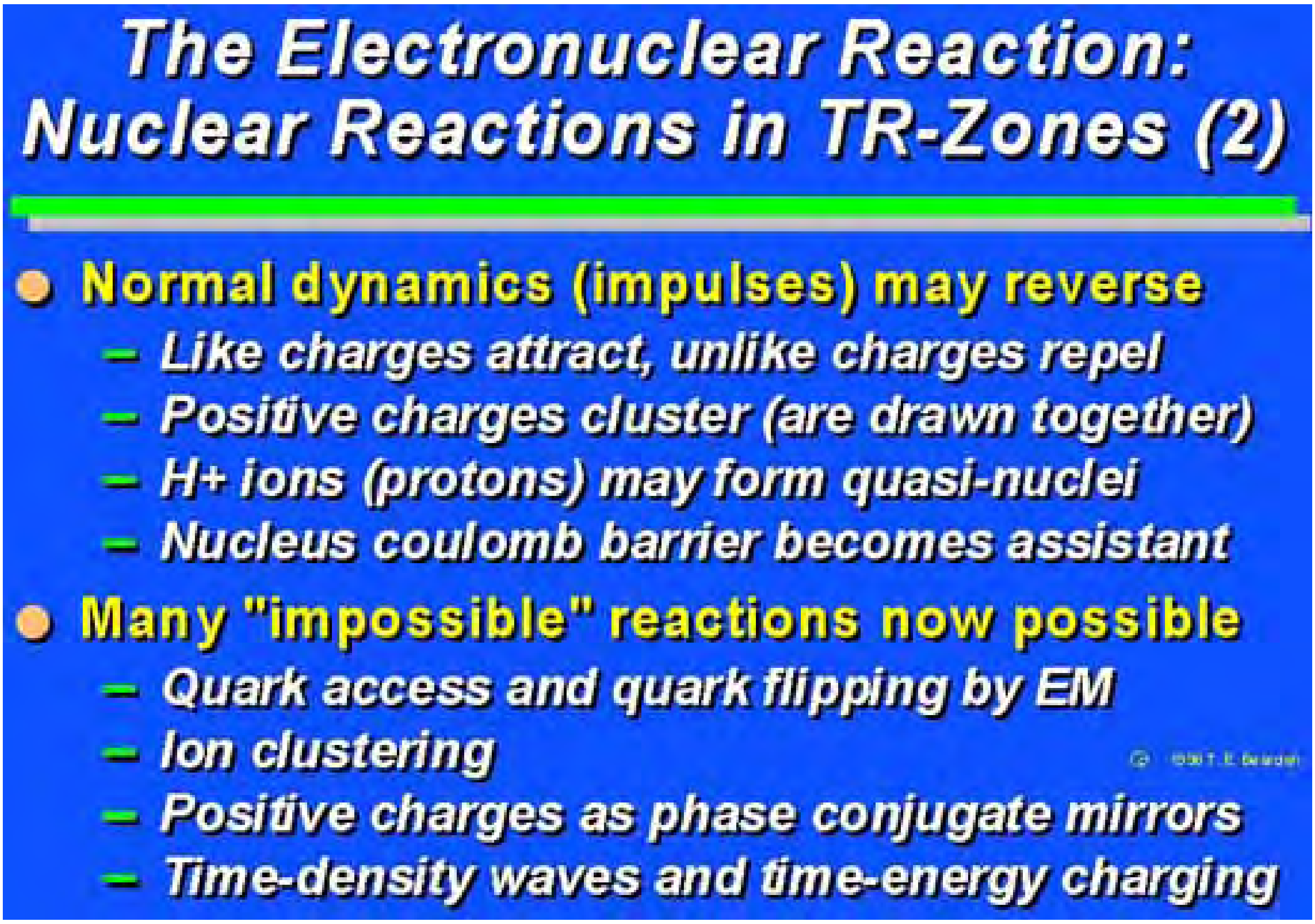

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- **Normal dynamics (impulses) may reverse**
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- **Many "impossible" reactions now possible**
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

© 1996 T. E. Bearden

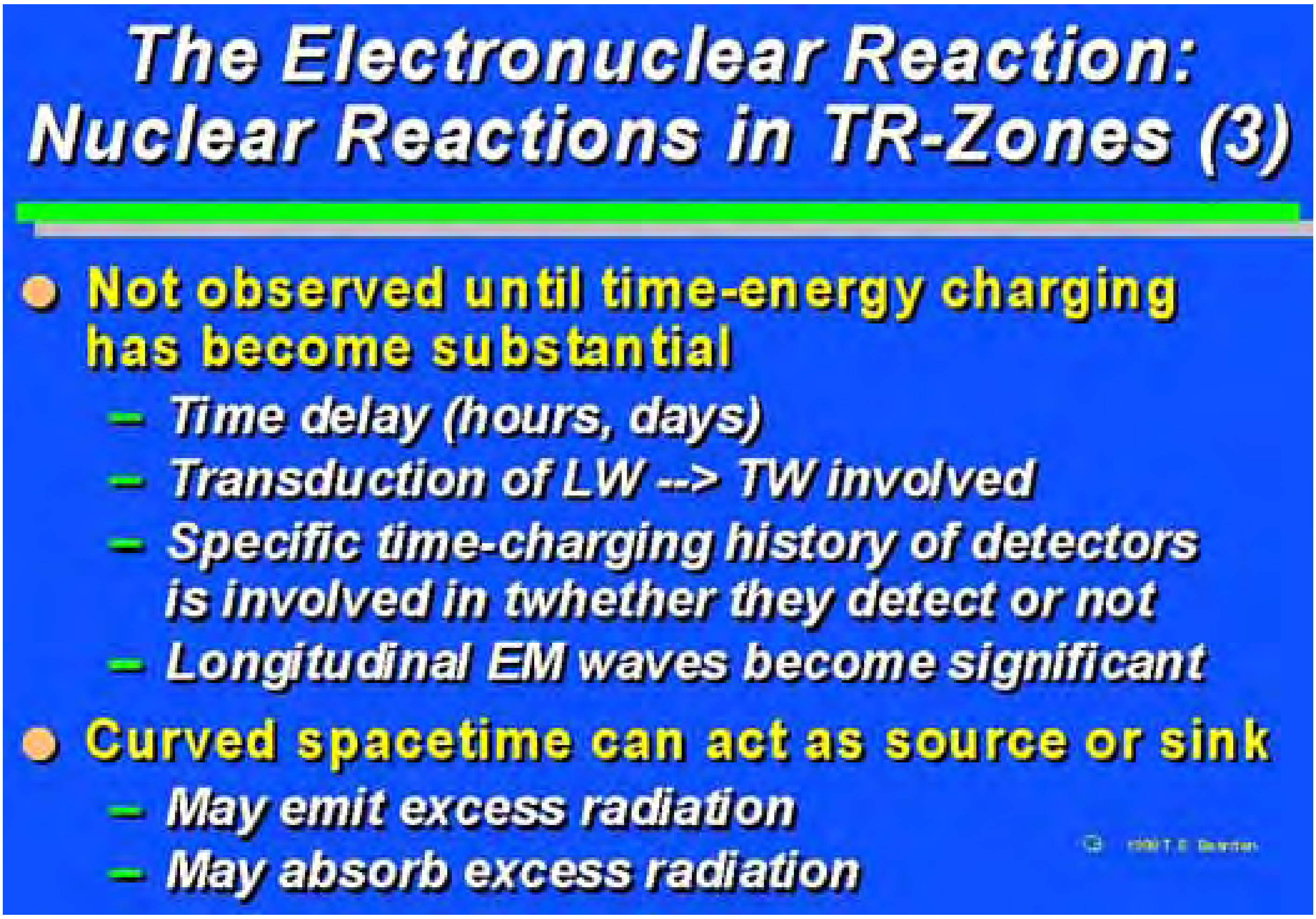

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- Not observed until time-energy charging has become substantial
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- Curved spacetime can act as source or sink
  - *May emit excess radiation*
  - *May absorb excess radiation*

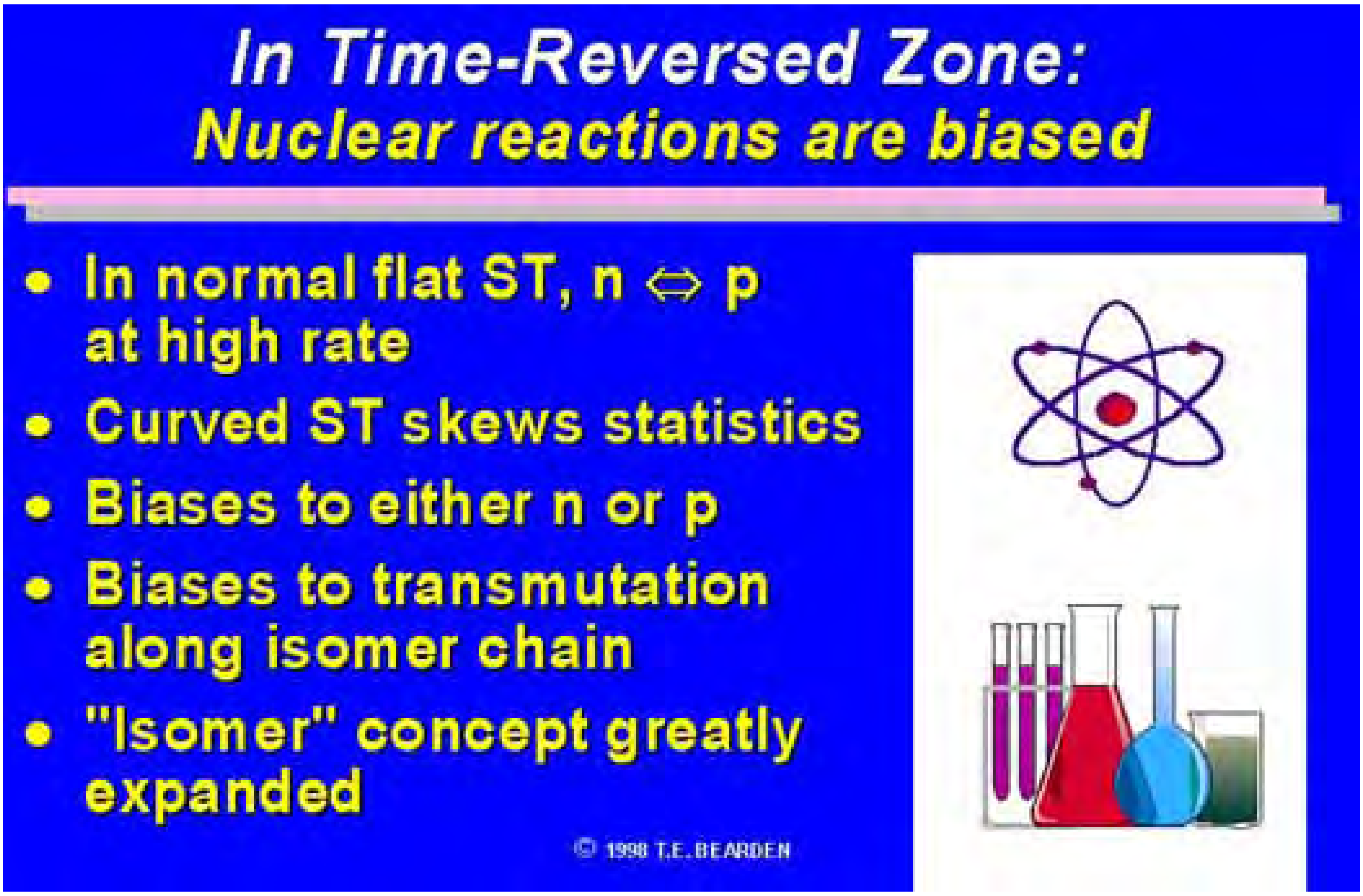

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E.BEARDEN

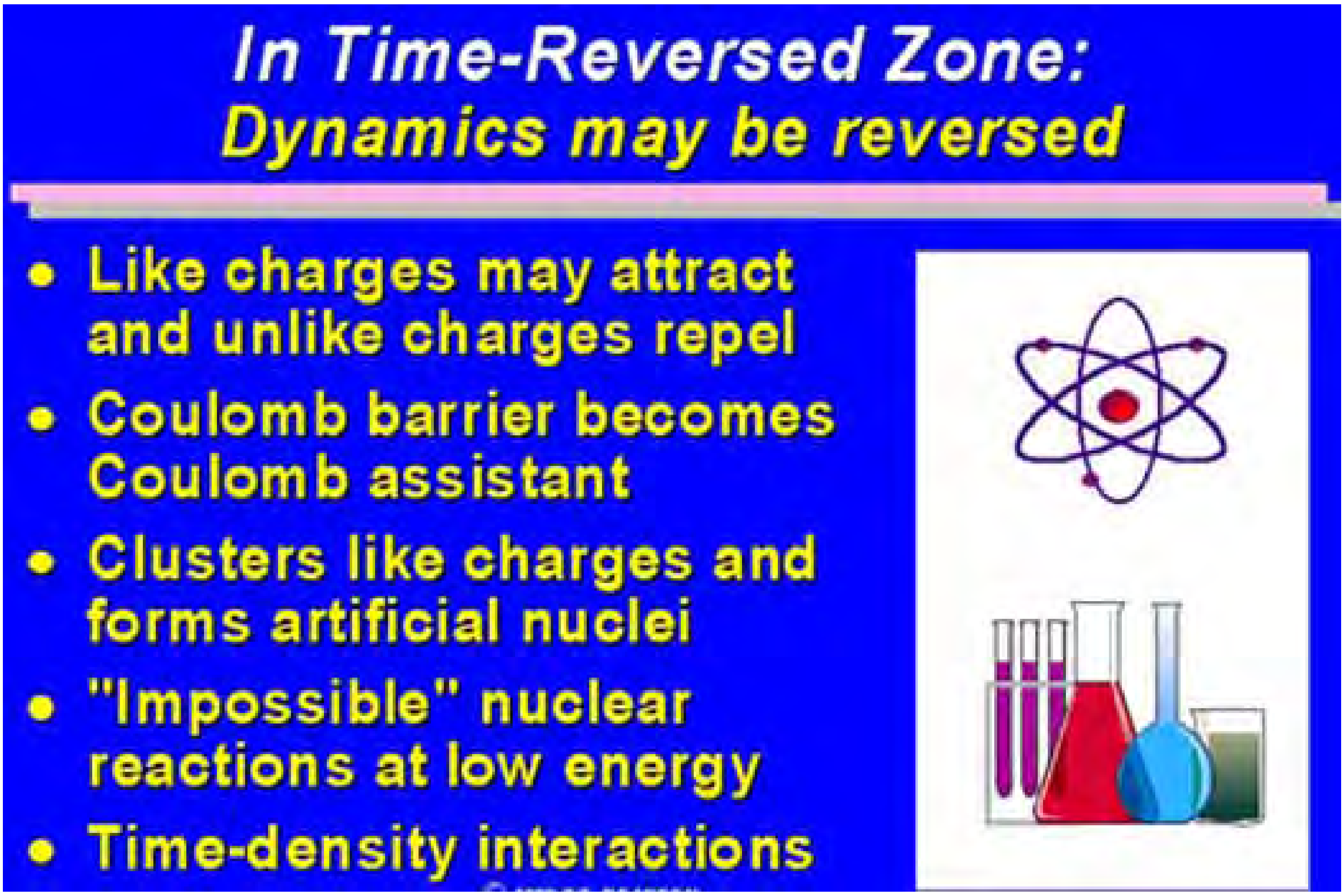

In Time-Reversed Zone: Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

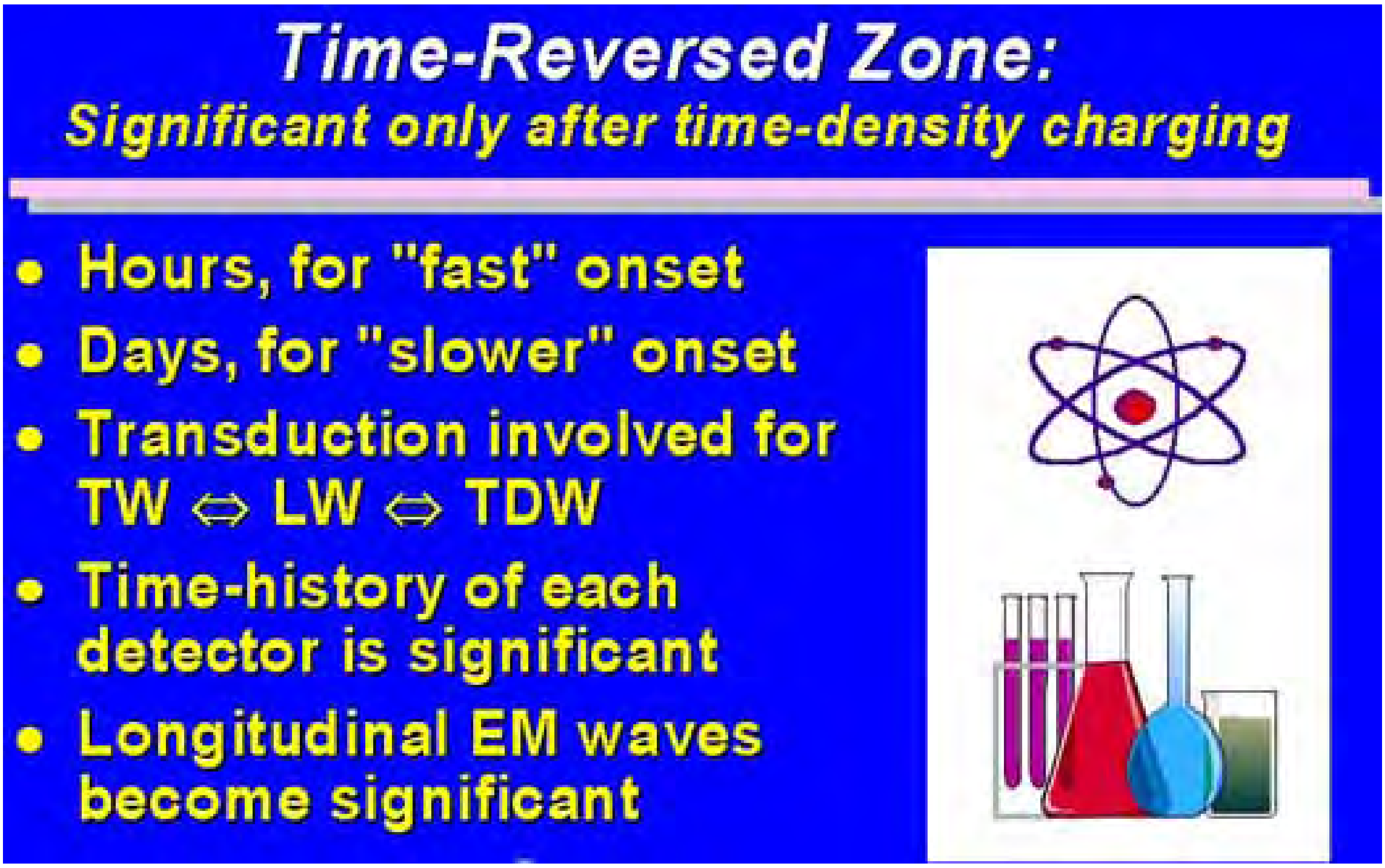

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

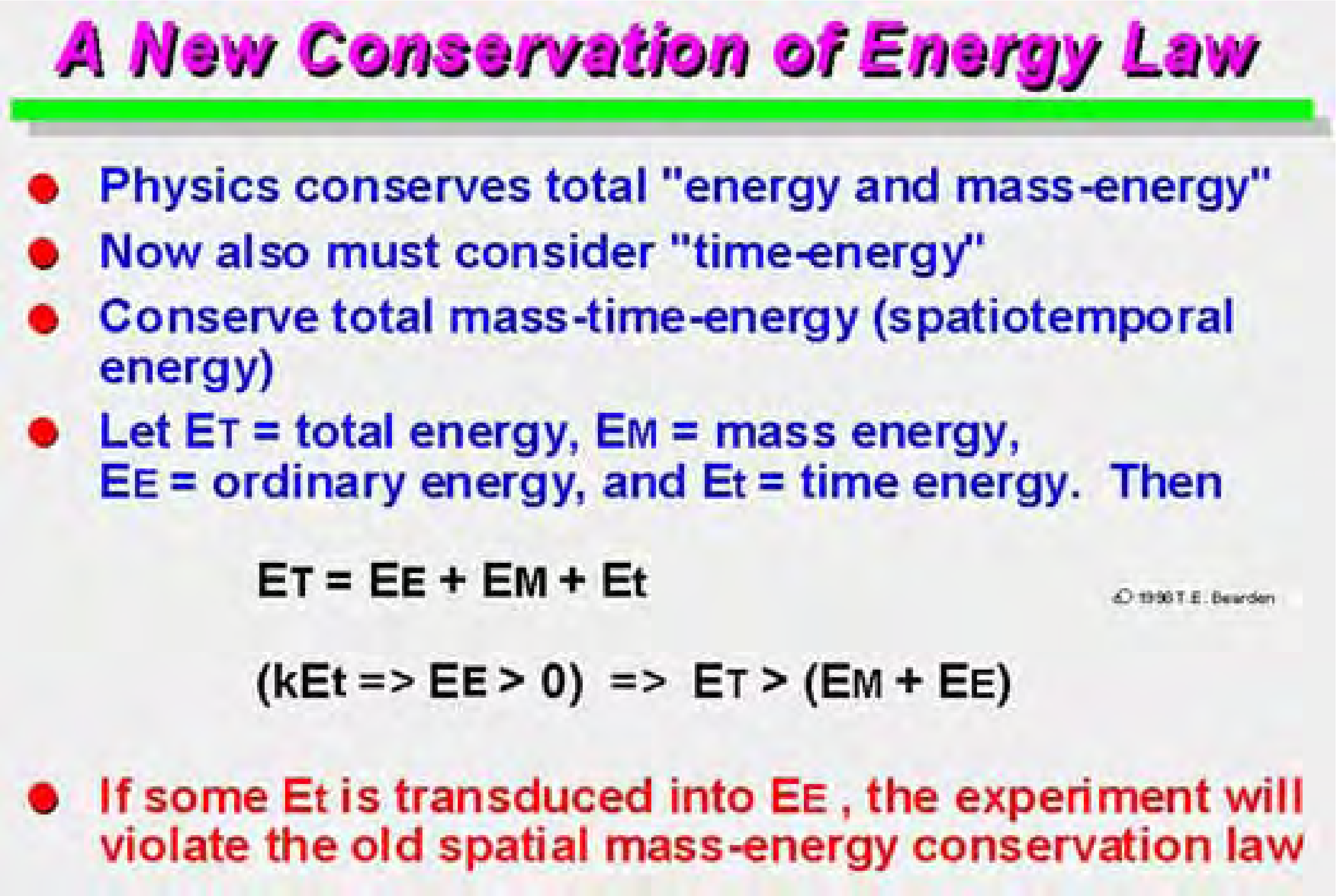

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T.E. Bearden

$$(kE_t \Rightarrow E_E > 0) \Rightarrow E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

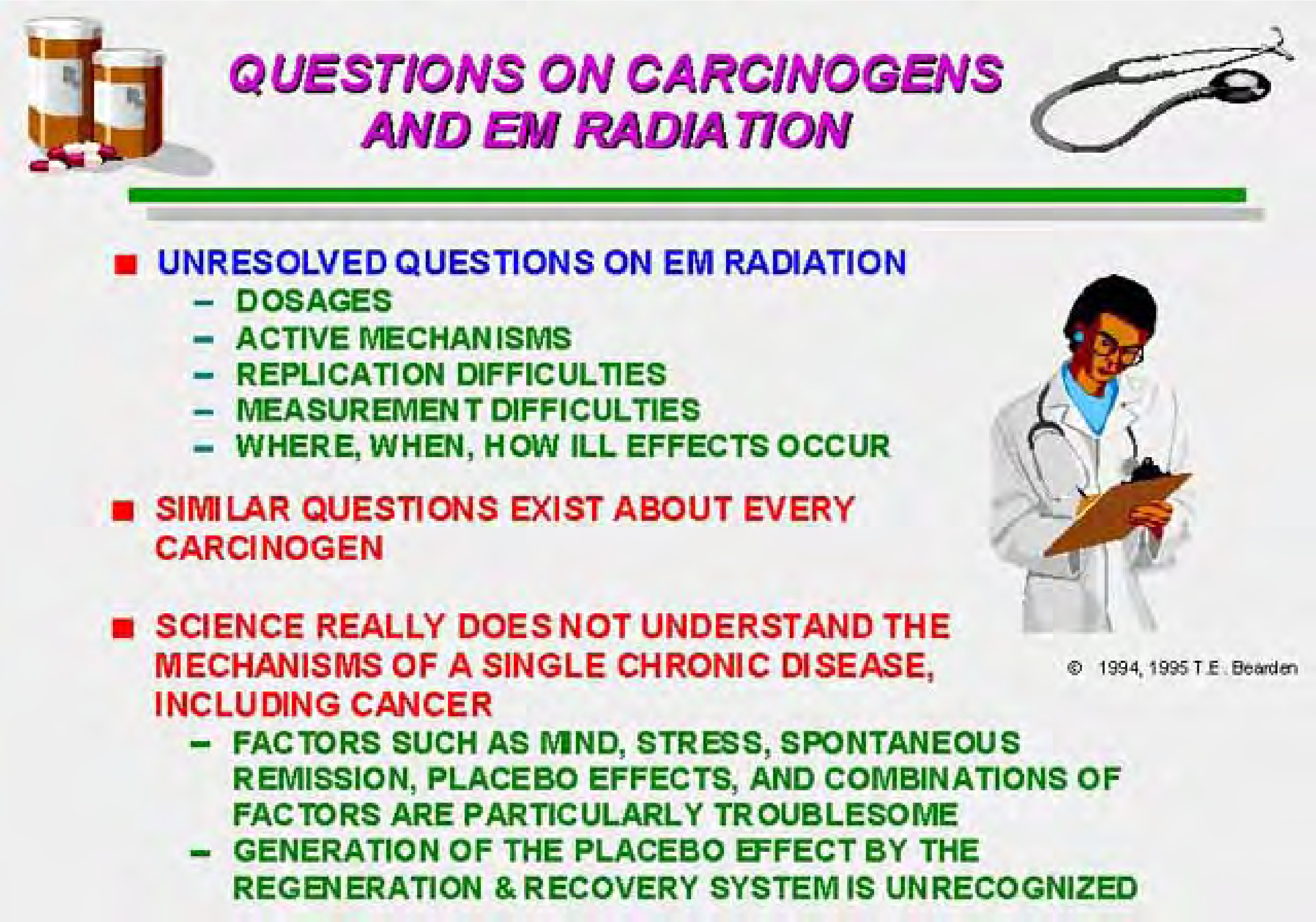

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

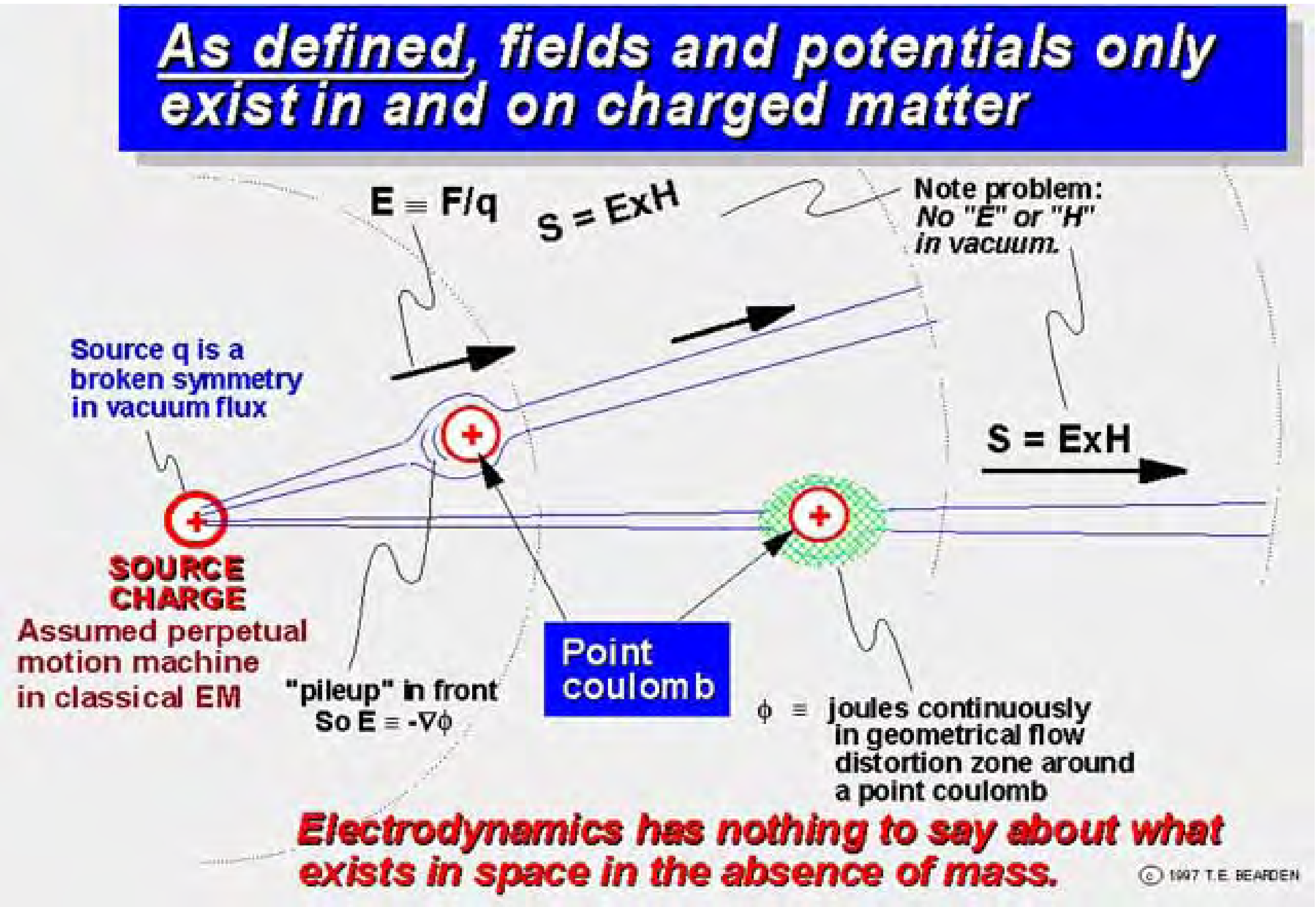

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a
broken symmetry
in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual
motion machine
in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point
coulomb

$\phi \equiv$ joules continuously
in geometrical flow
distortion zone around
a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

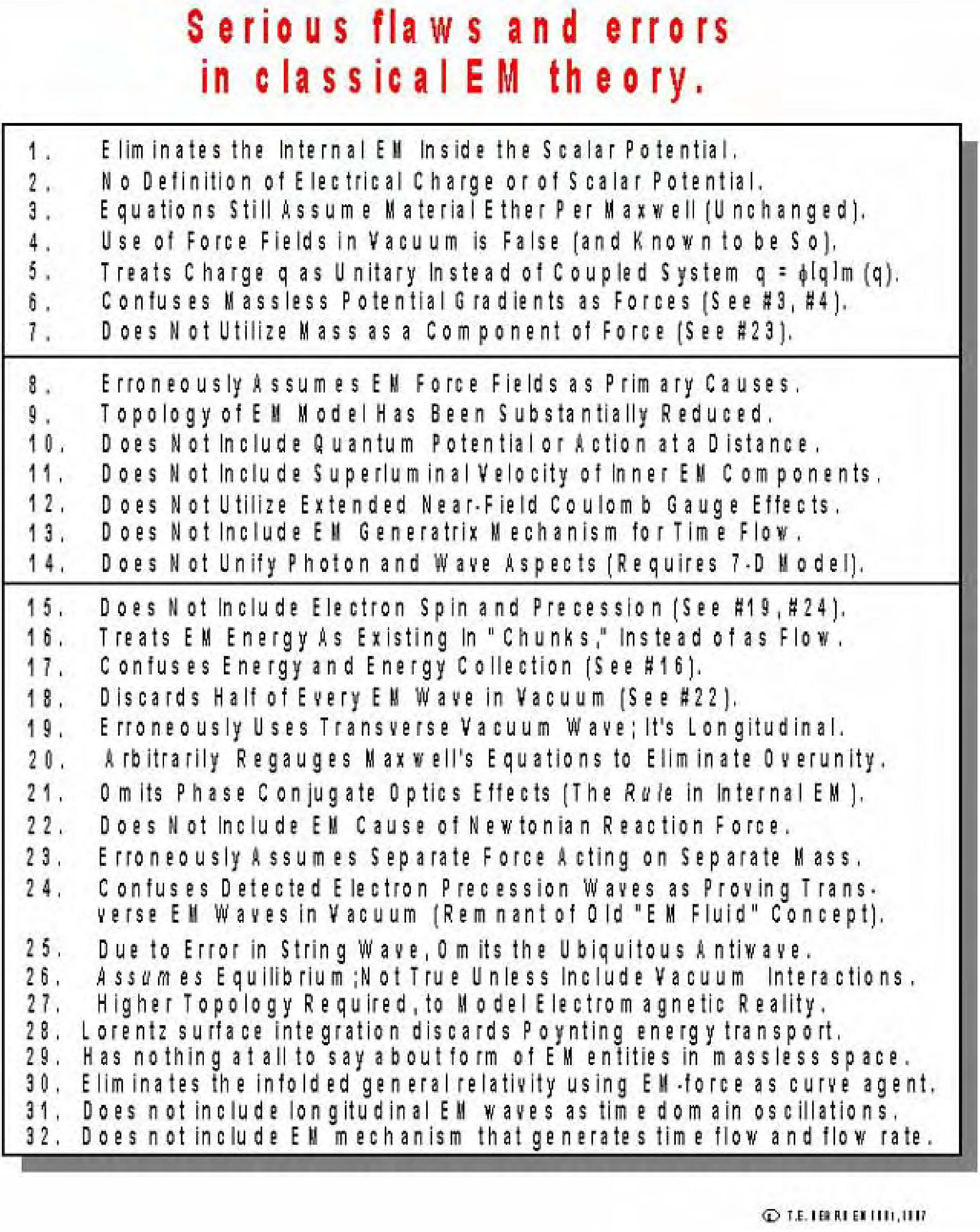

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assumes* Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991,1997

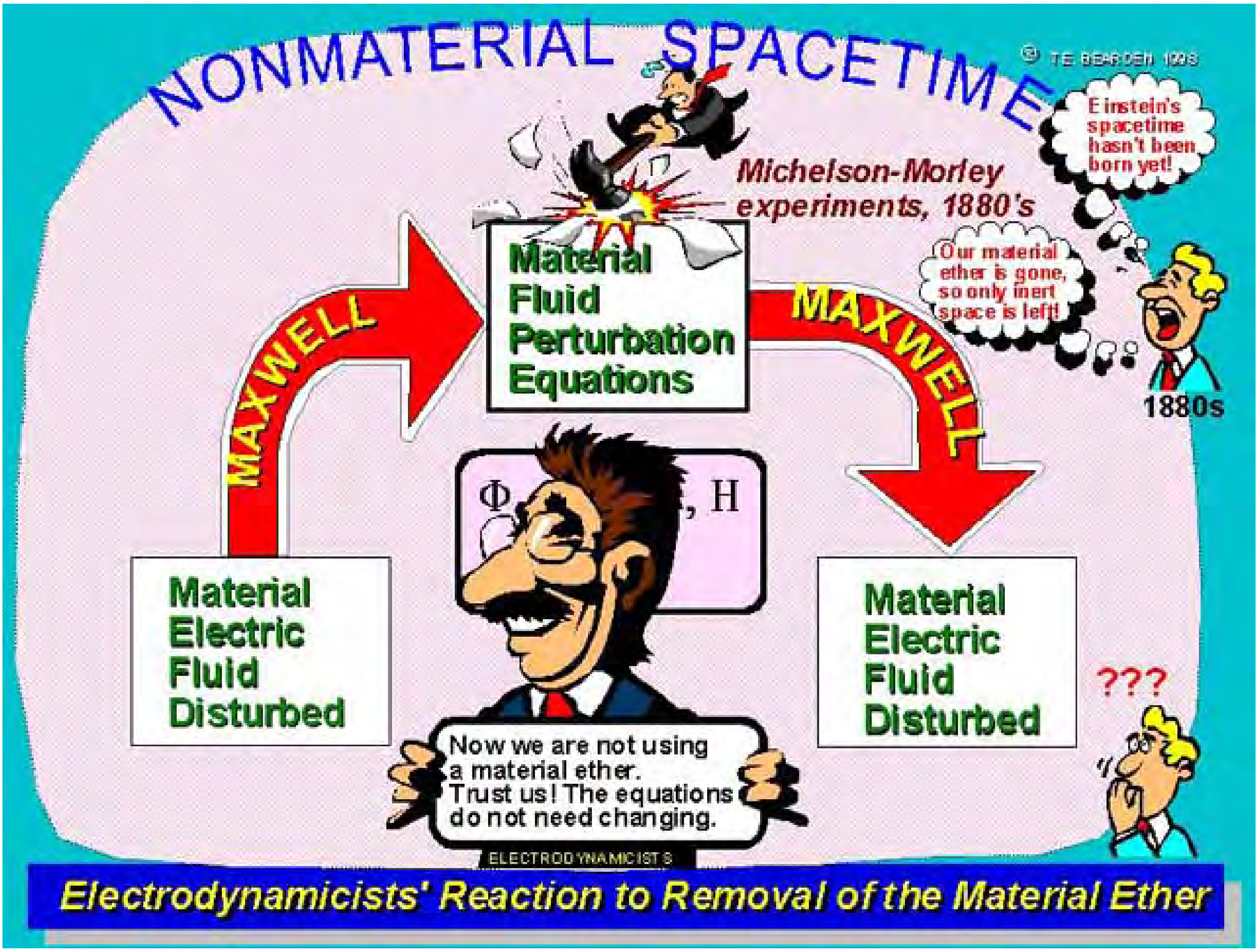
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ
, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

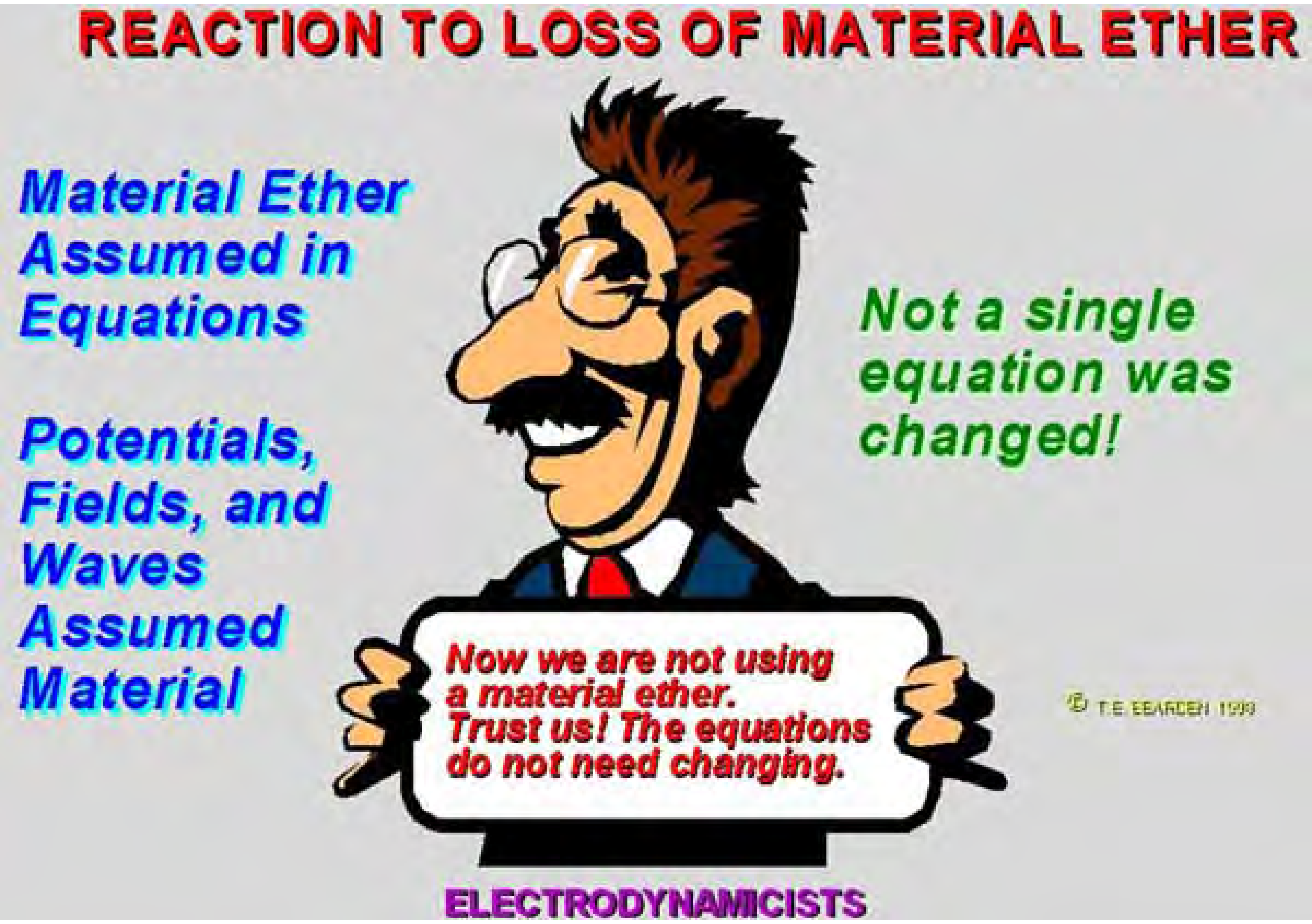

REACTION TO LOSS OF MATERIAL ETHER

Material Ether Assumed in Equations

Potentials, Fields, and Waves Assumed Material

Not a single equation was changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.
ELECTRODYNAMICISTS

© T.E. BEARDEN 1999

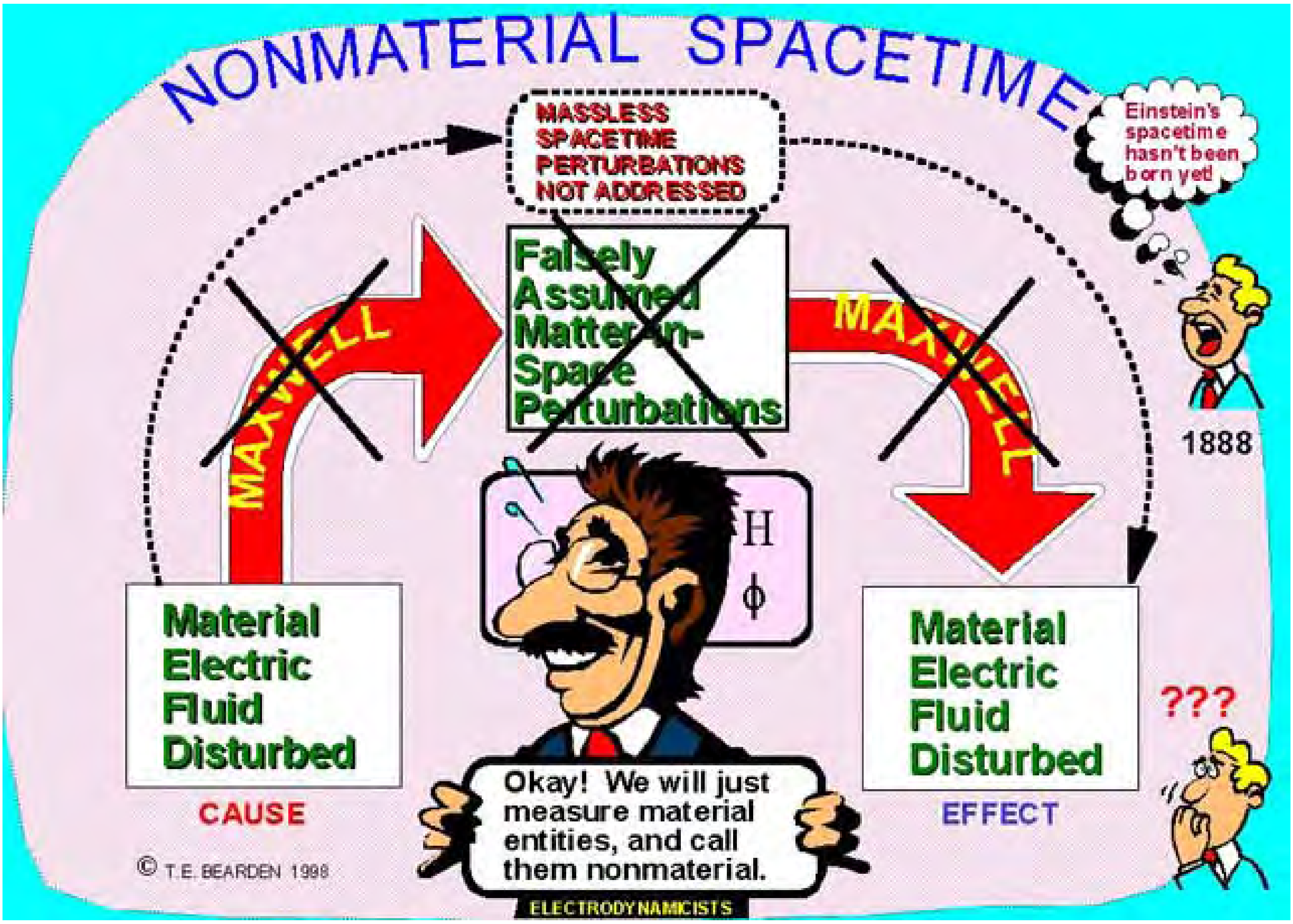
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
~~MAXWELL~~
~~Falsely Assumed Matter-in-Space Perturbations~~
~~MAXWELL~~
1888
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

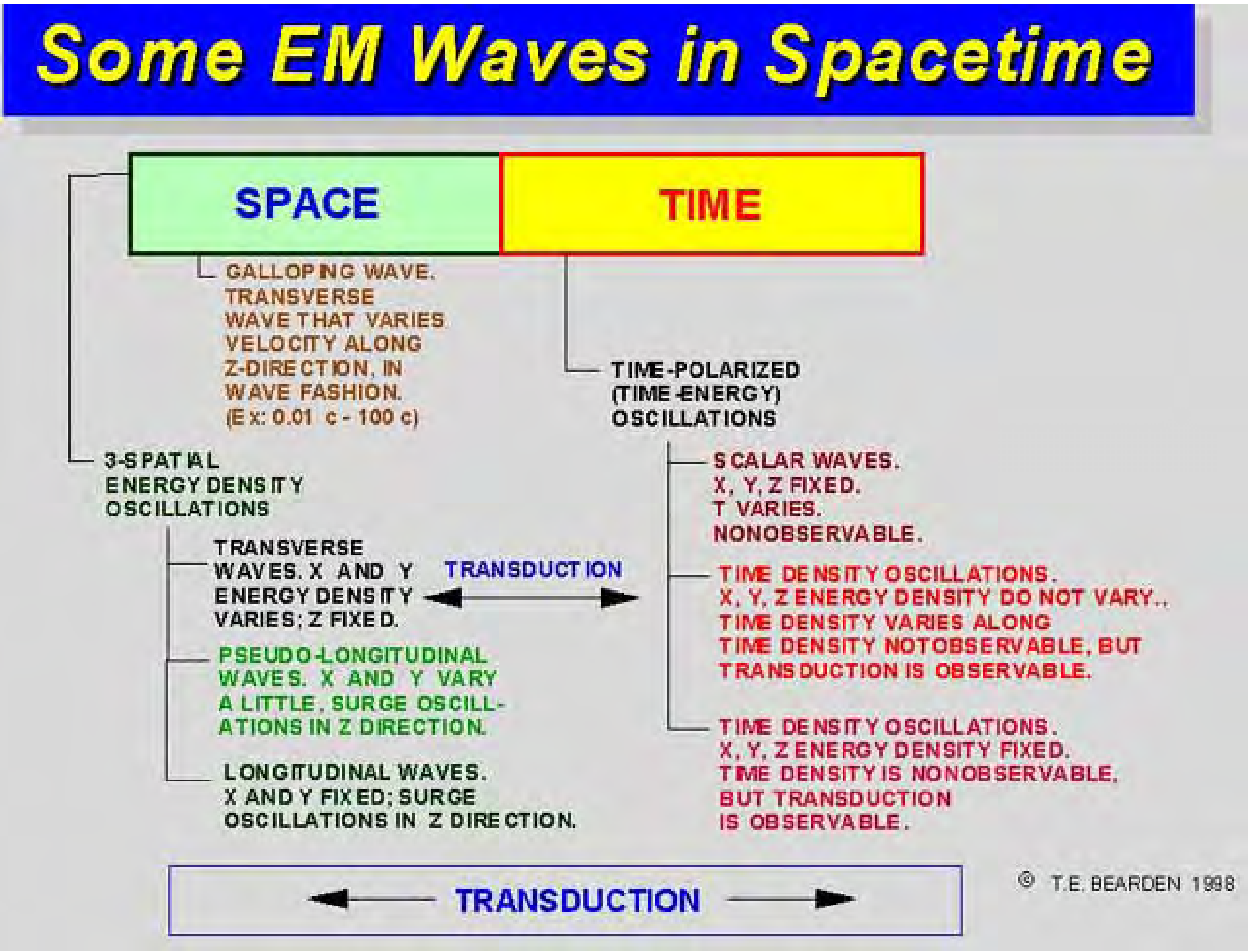

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE. TRANSVERSE WAVE THAT VARIES VELOCITY ALONG Z-DIRECTION, IN WAVE FASHION. (Ex: 0.01 c - 100 c)

TIME-POLARIZED (TIME-ENERGY) OSCILLATIONS

3-SPATIAL ENERGY DENSITY OSCILLATIONS

SCALAR WAVES. X, Y, Z FIXED. T VARIES. NONOBSERVABLE.

TRANSVERSE WAVES. X AND Y ENERGY DENSITY VARIES; Z FIXED.

TRANSDUCTION

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY DO NOT VARY.. TIME DENSITY VARIES ALONG TIME DENSITY NOTOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

PSEUDO-LONGITUDINAL WAVES. X AND Y VARY A LITTLE, SURGE OSCILLATIONS IN Z DIRECTION.

TIME DENSITY OSCILLATIONS. X, Y, Z ENERGY DENSITY FIXED. TIME DENSITY IS NONOBSERVABLE, BUT TRANSDUCTION IS OBSERVABLE.

LONGITUDINAL WAVES. X AND Y FIXED; SURGE OSCILLATIONS IN Z DIRECTION.

TRANSDUCTION

© T.E. BEARDEN 1998

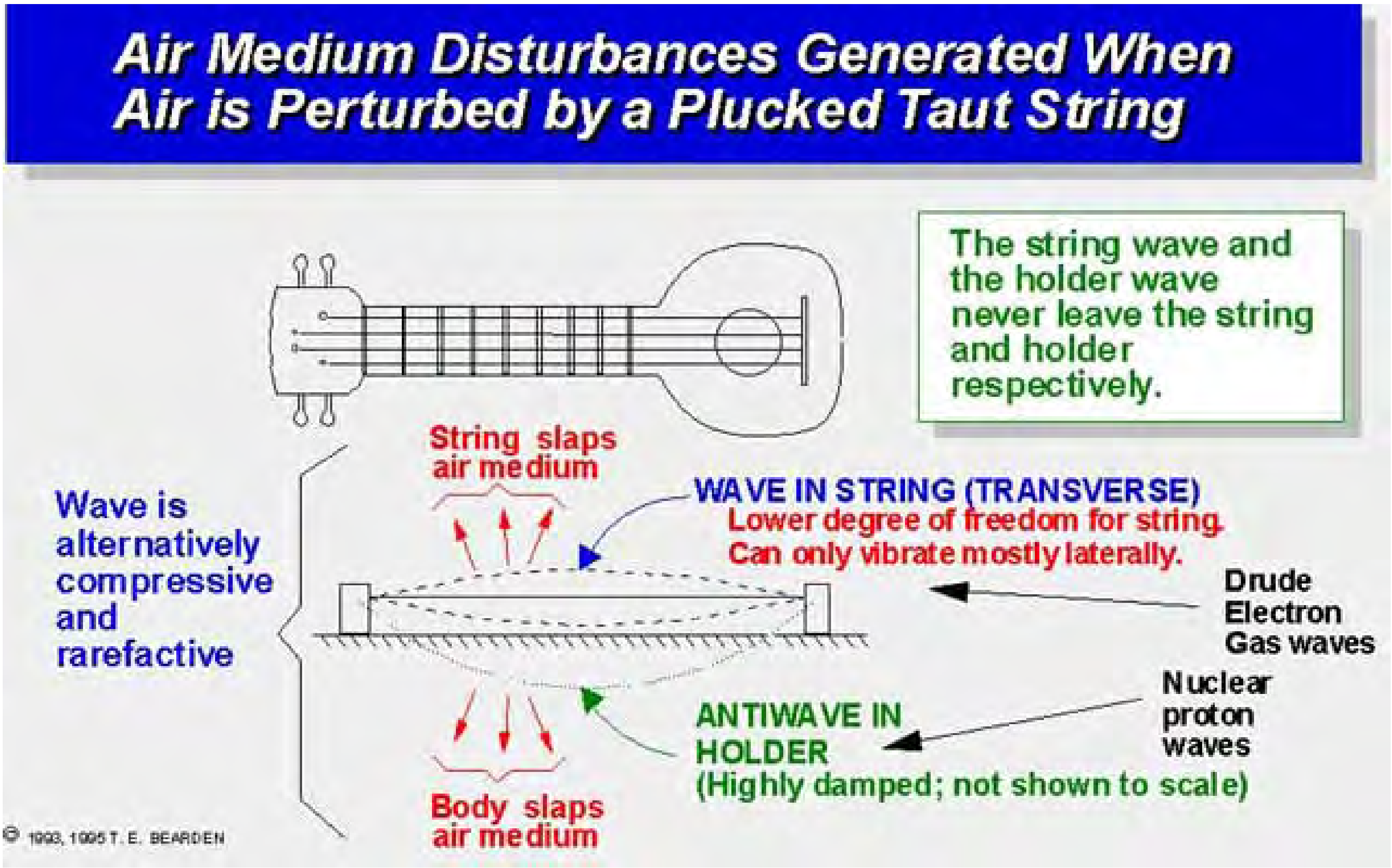

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string.
Can only vibrate mostly laterally.

Drude Electron Gas waves

Wave is alternatively compressive and rarefactive

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

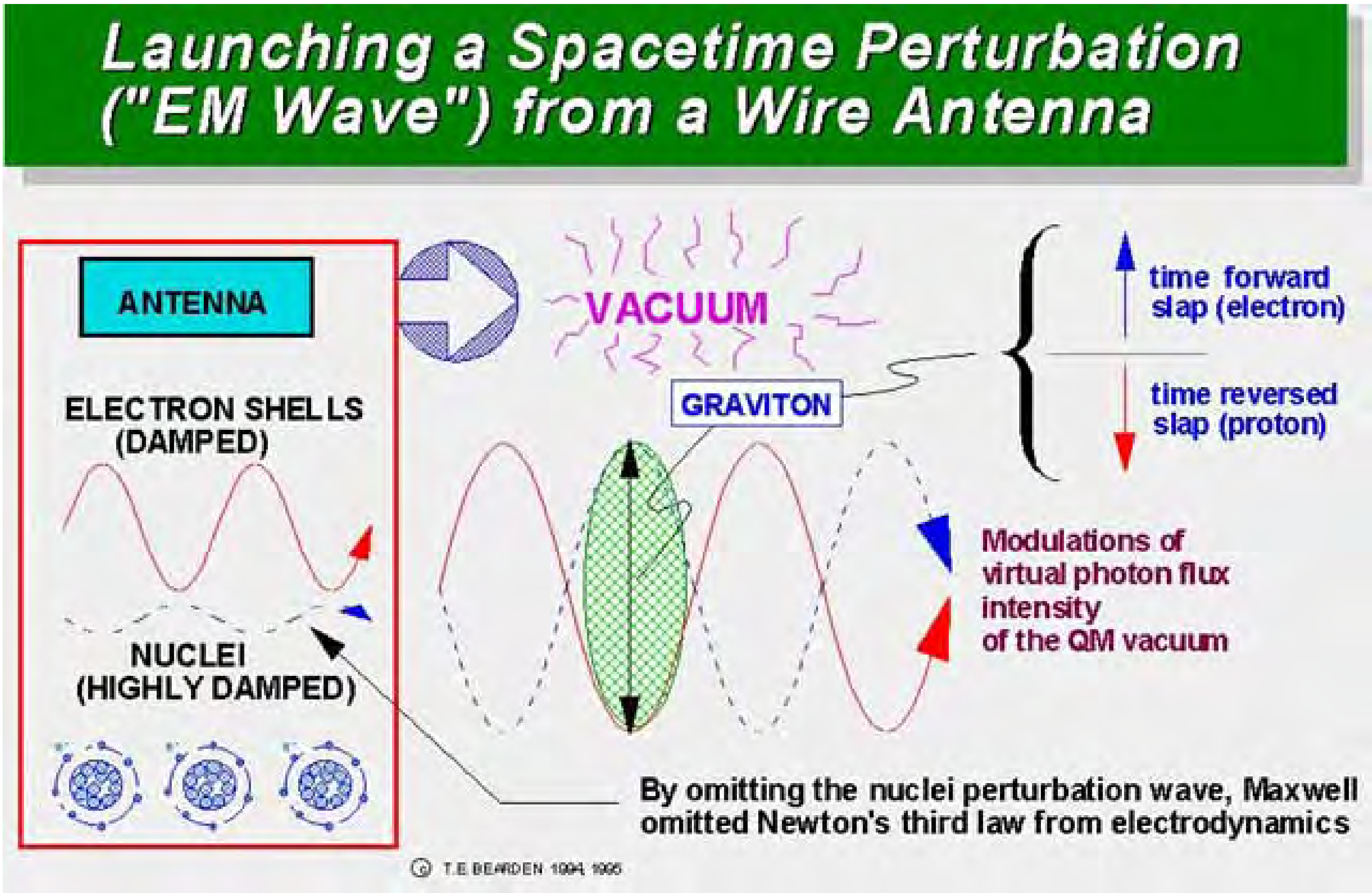

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA

ELECTRON SHELLS (DAMPED)

NUCLEI (HIGHLY DAMPED)

VACUUM

GRAVITON

time forward slap (electron)

time reversed slap (proton)

Modulations of virtual photon flux intensity of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

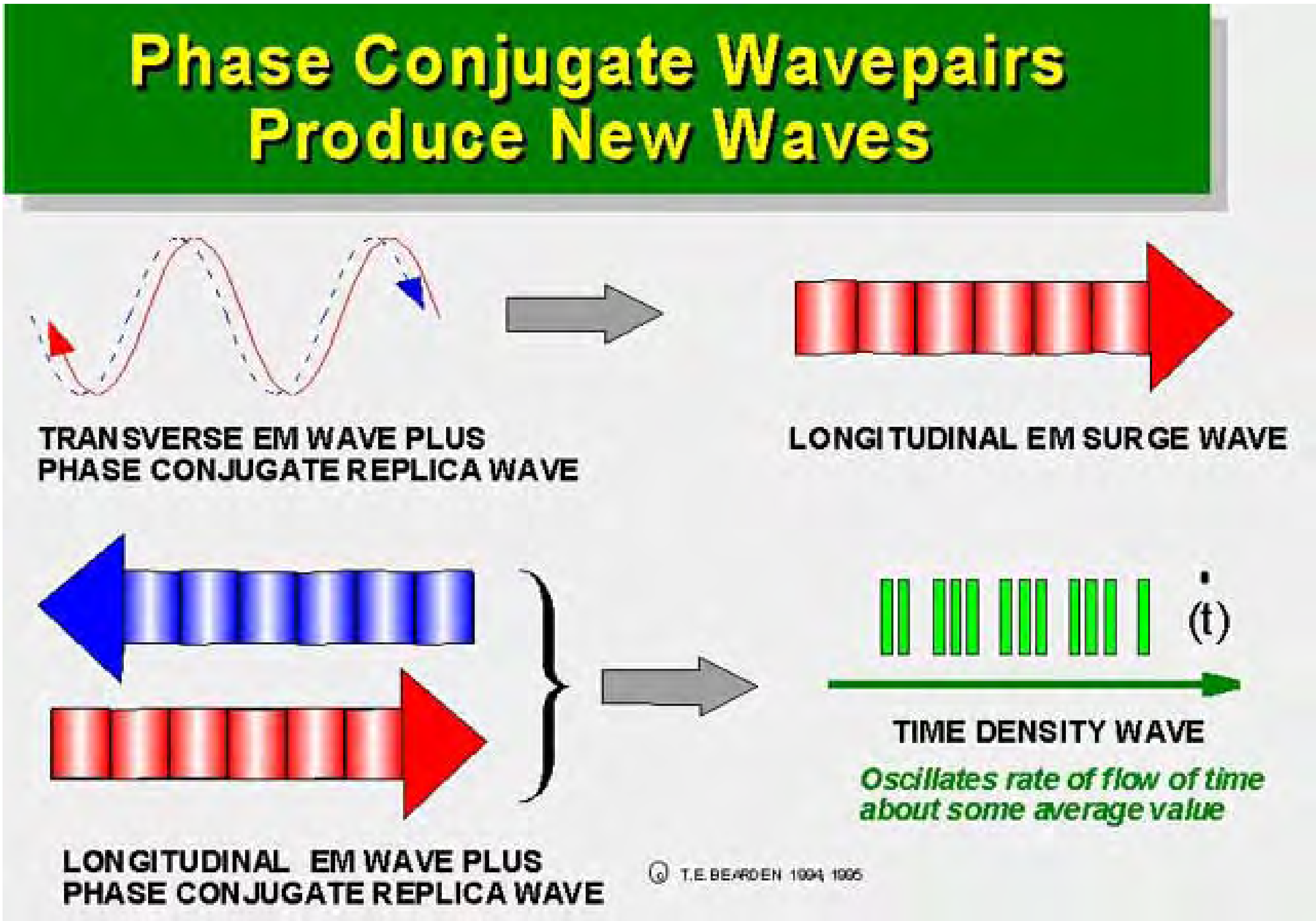

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

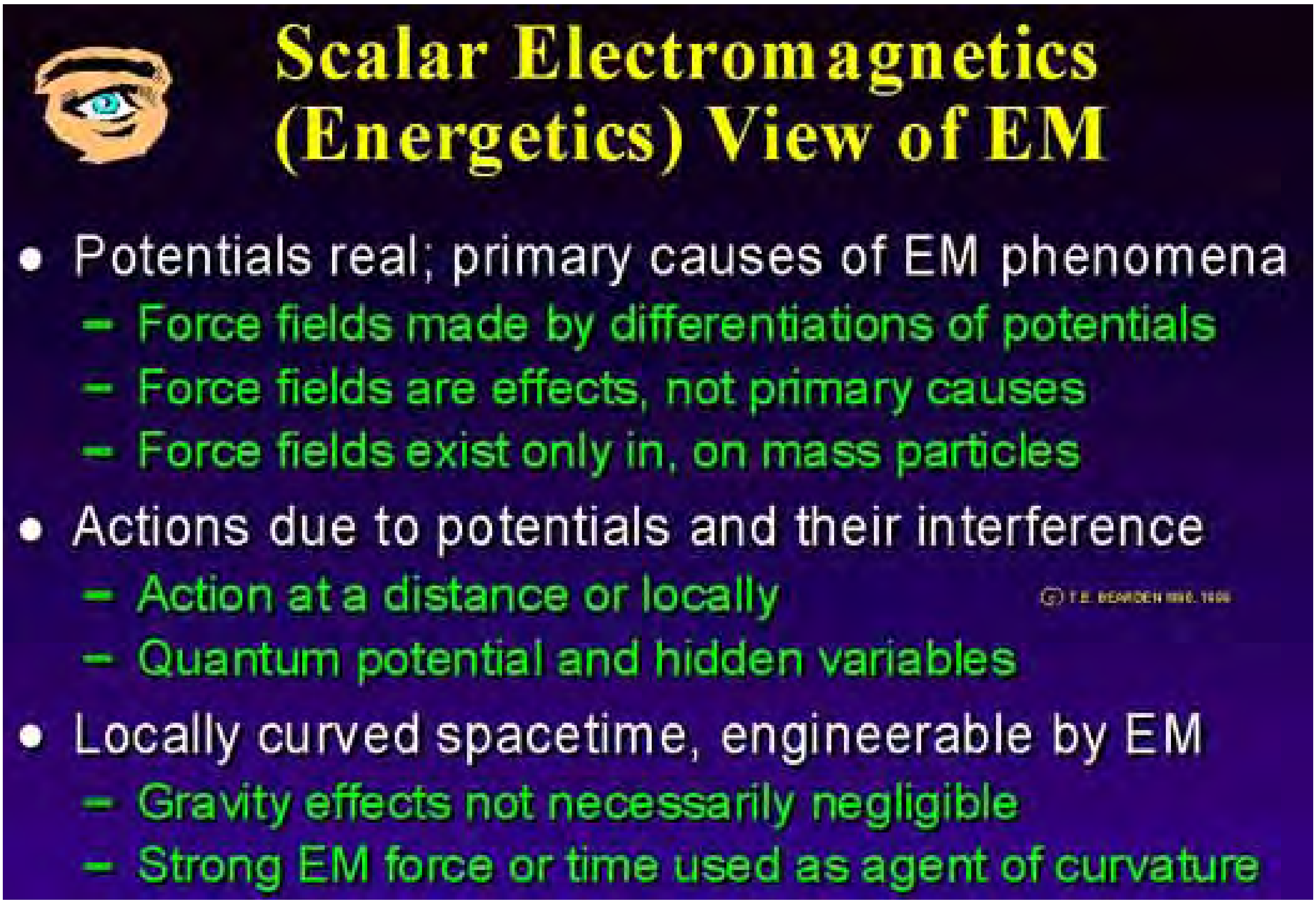

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN 1990, 1996

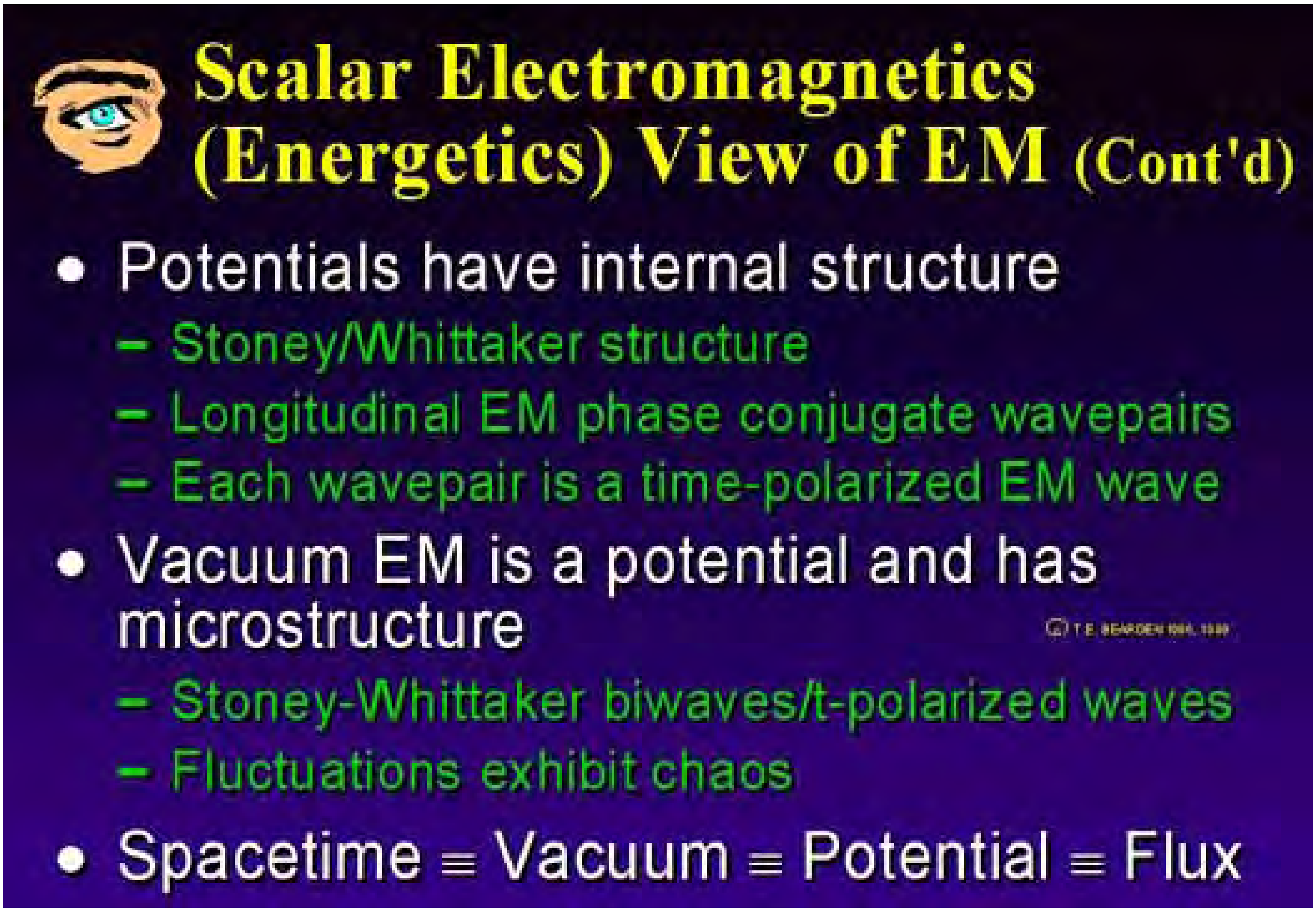

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

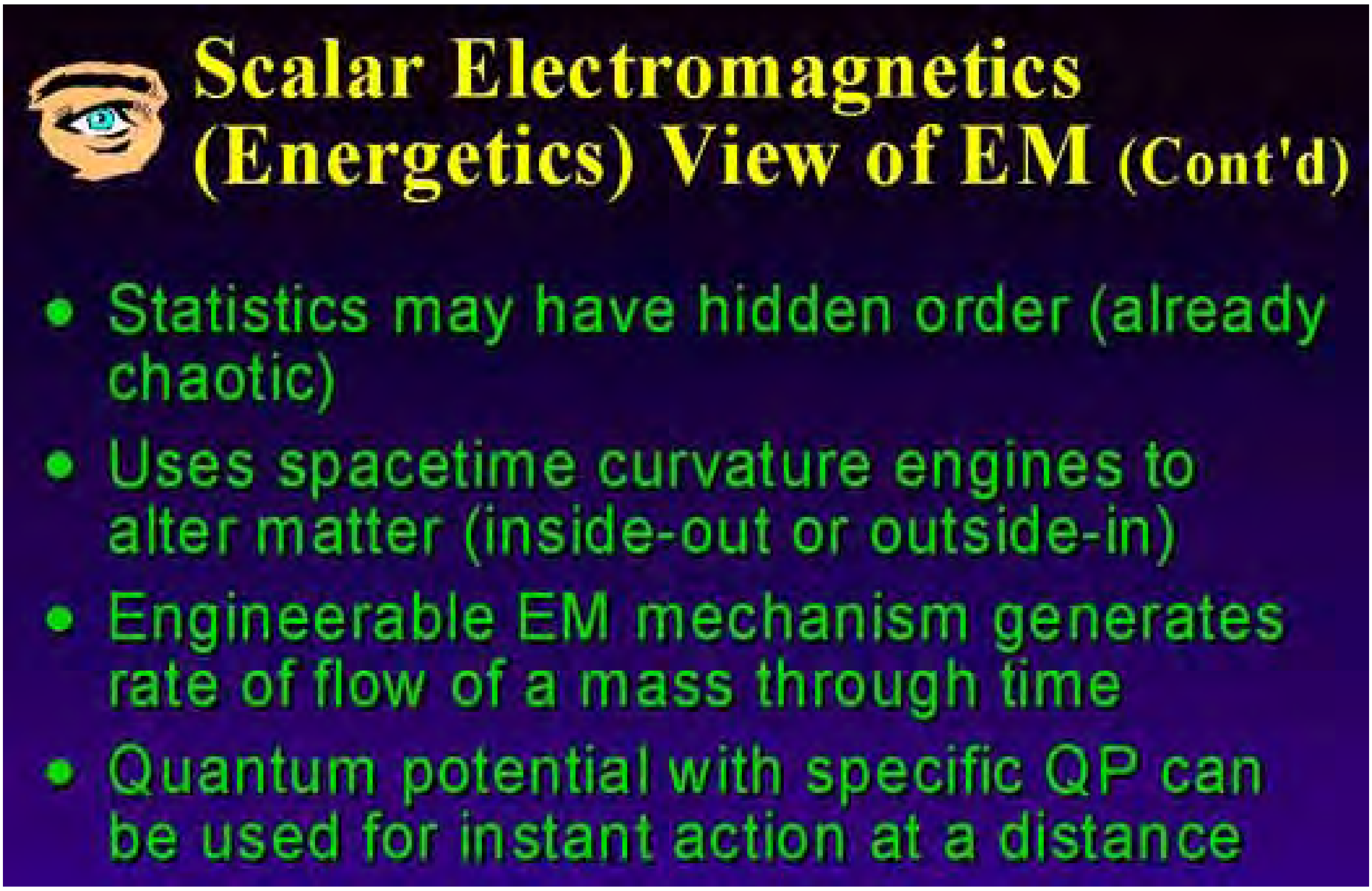

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

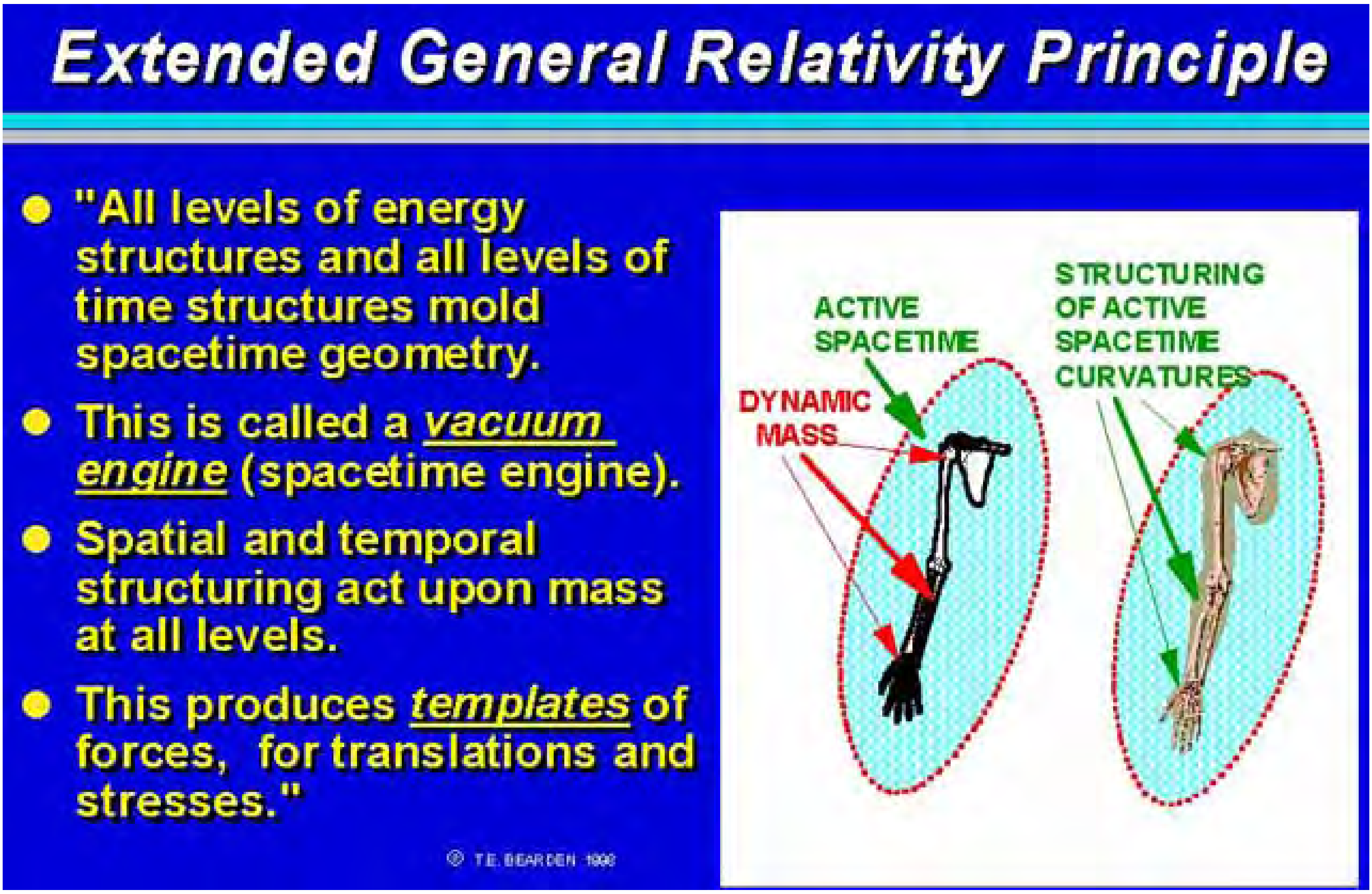

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a ***vacuum engine*** (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces ***templates*** of forces, for translations and stresses."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1990

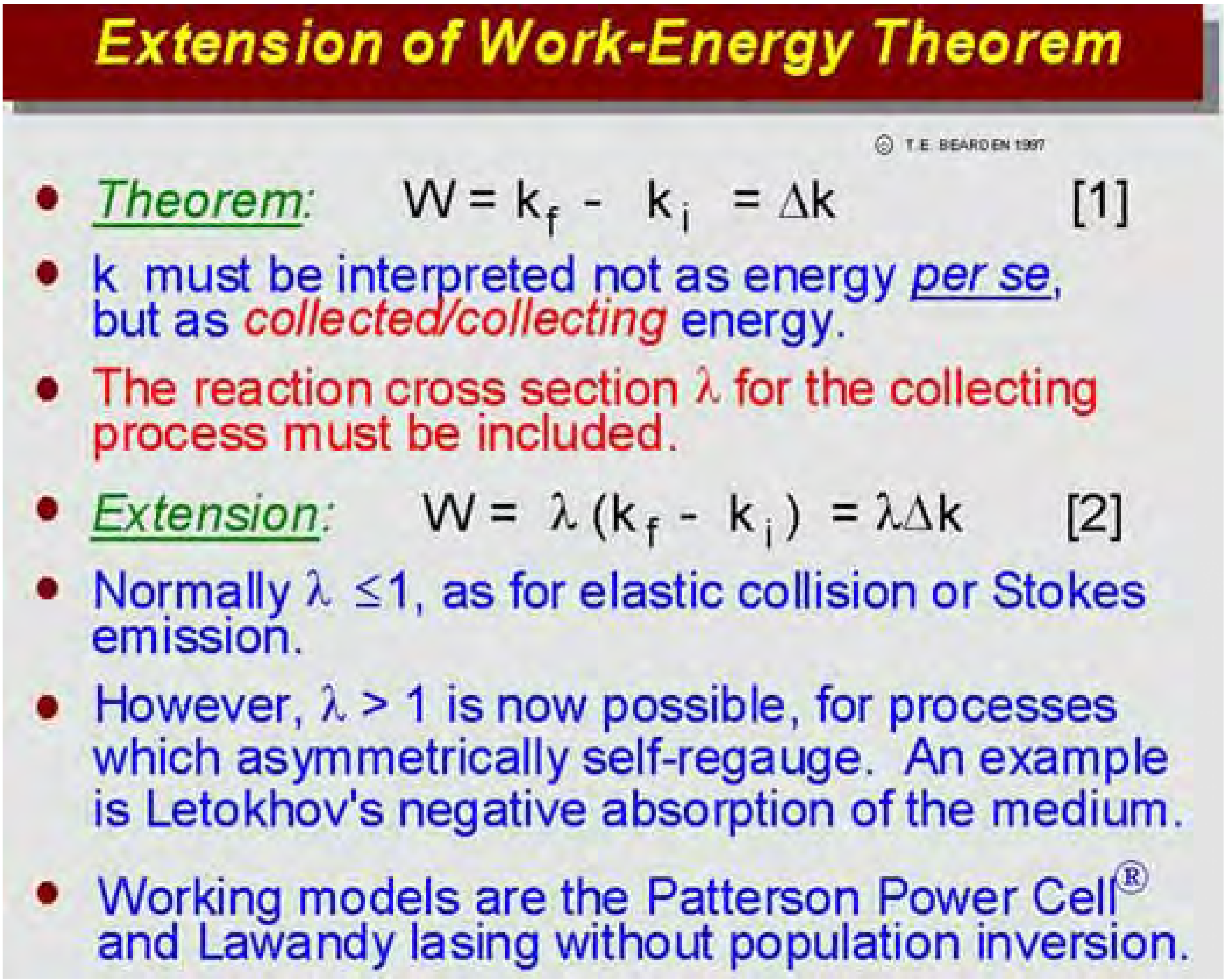

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem*: $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension*: $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

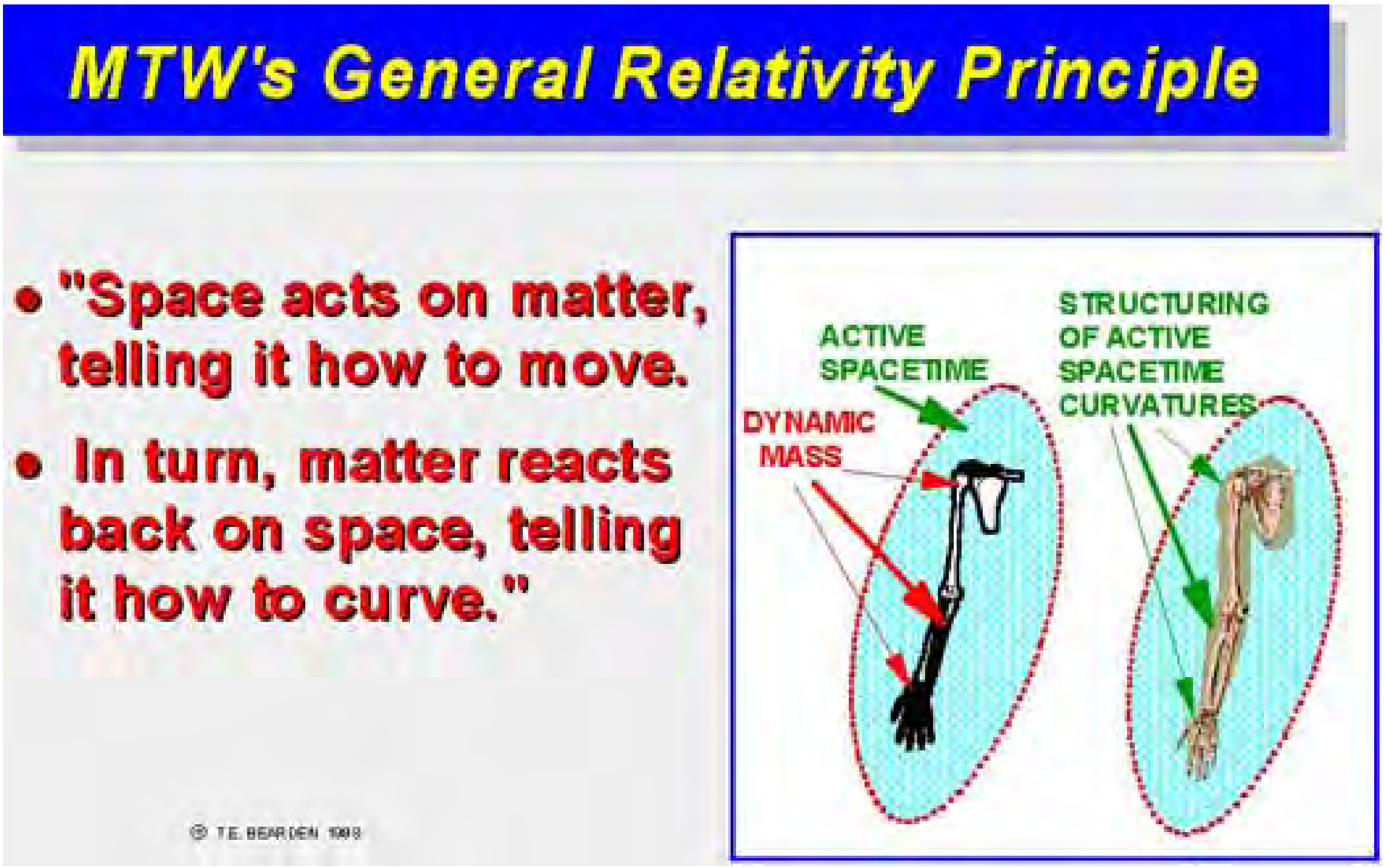

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
DYNAMIC MASS
STRUCTURING OF ACTIVE SPACETIME CURVATURES

© T.E. BEARDEN 1998

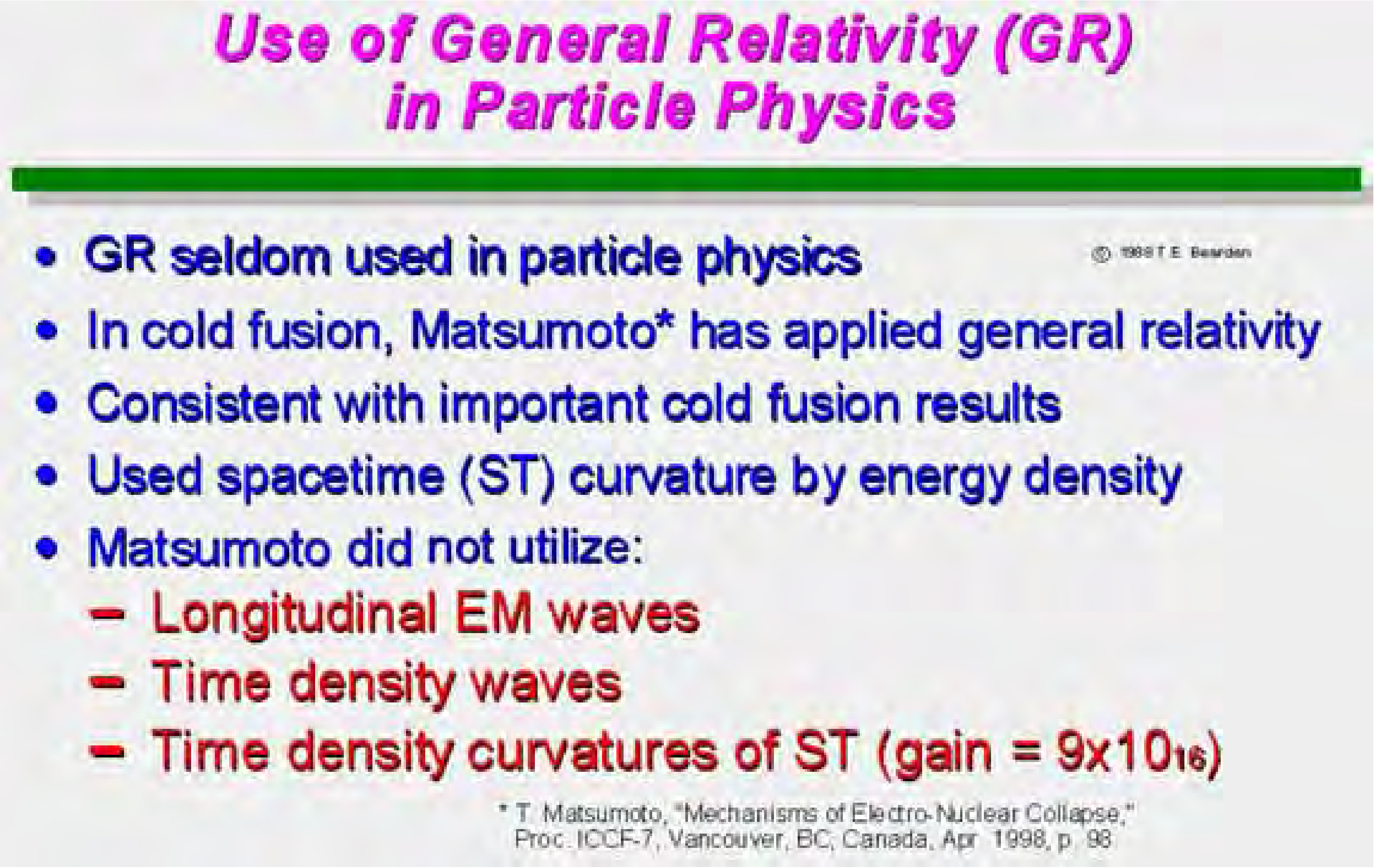

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = $9x10^{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98.

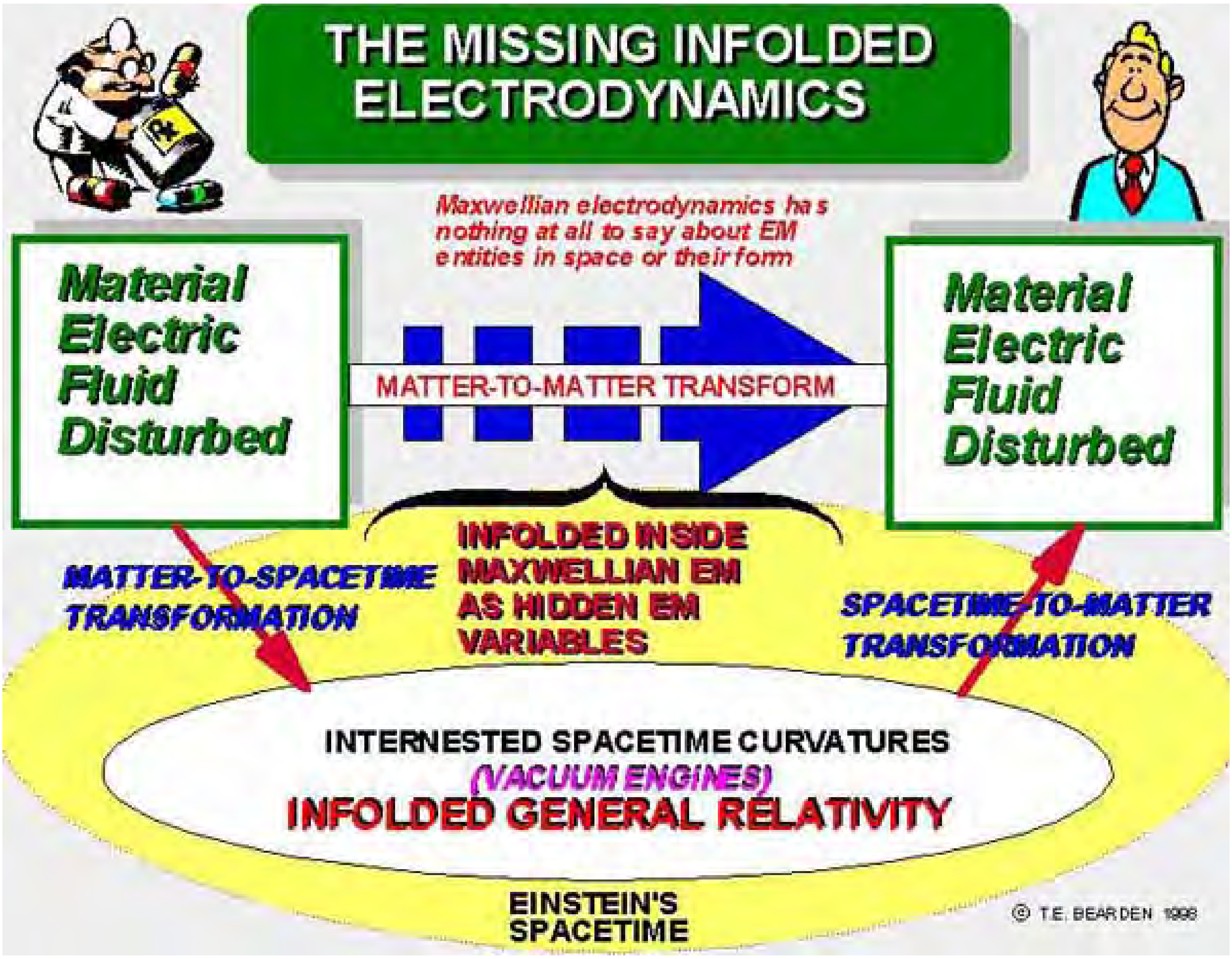

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

MATTER-TO-SPACETIME TRANSFORMATION

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

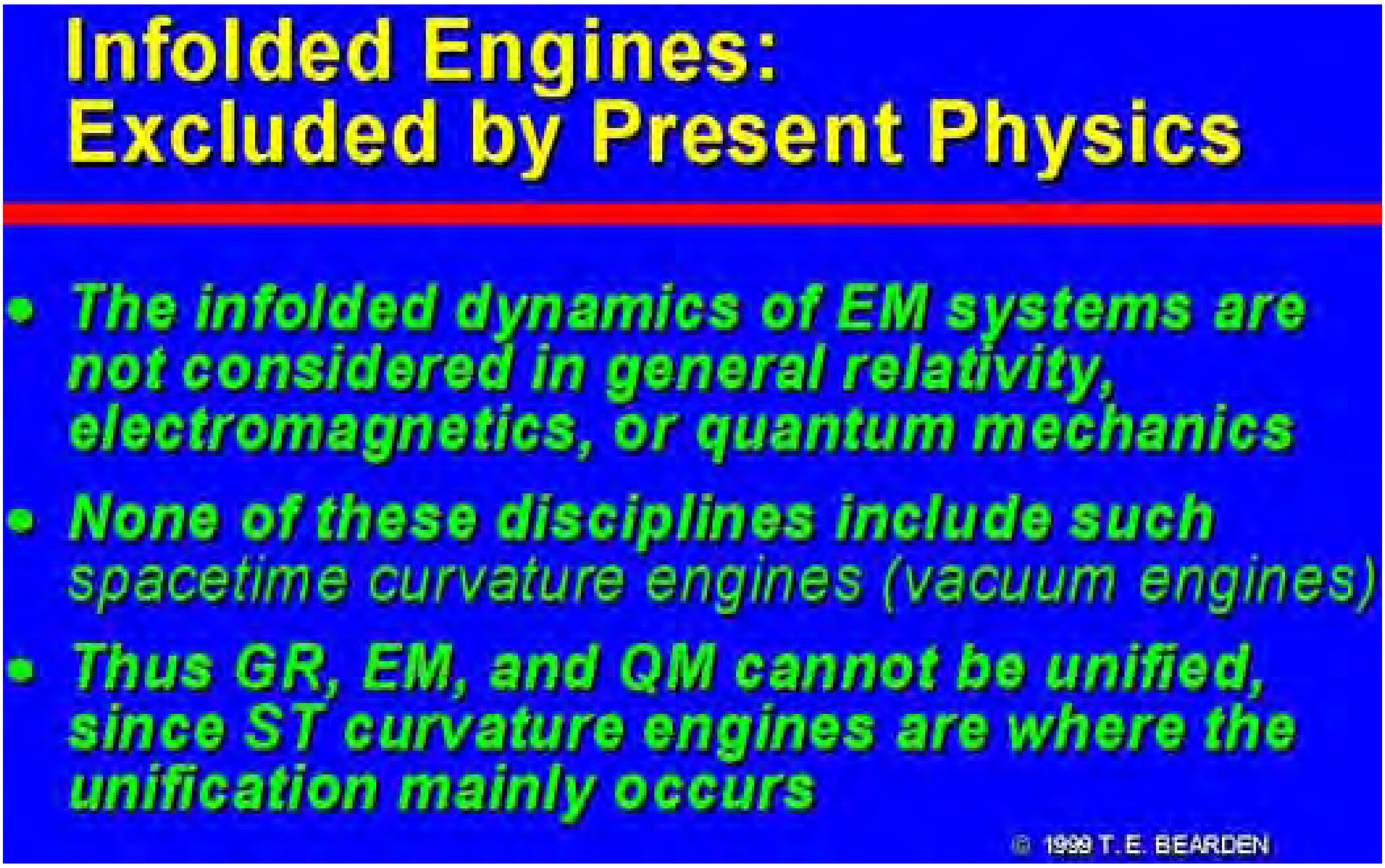

Infolded Engines: Excluded by Present Physics

- *The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics*
- *None of these disciplines include such spacetime curvature engines (vacuum engines)*
- *Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs*

© 1999 T. E. BEARDEN

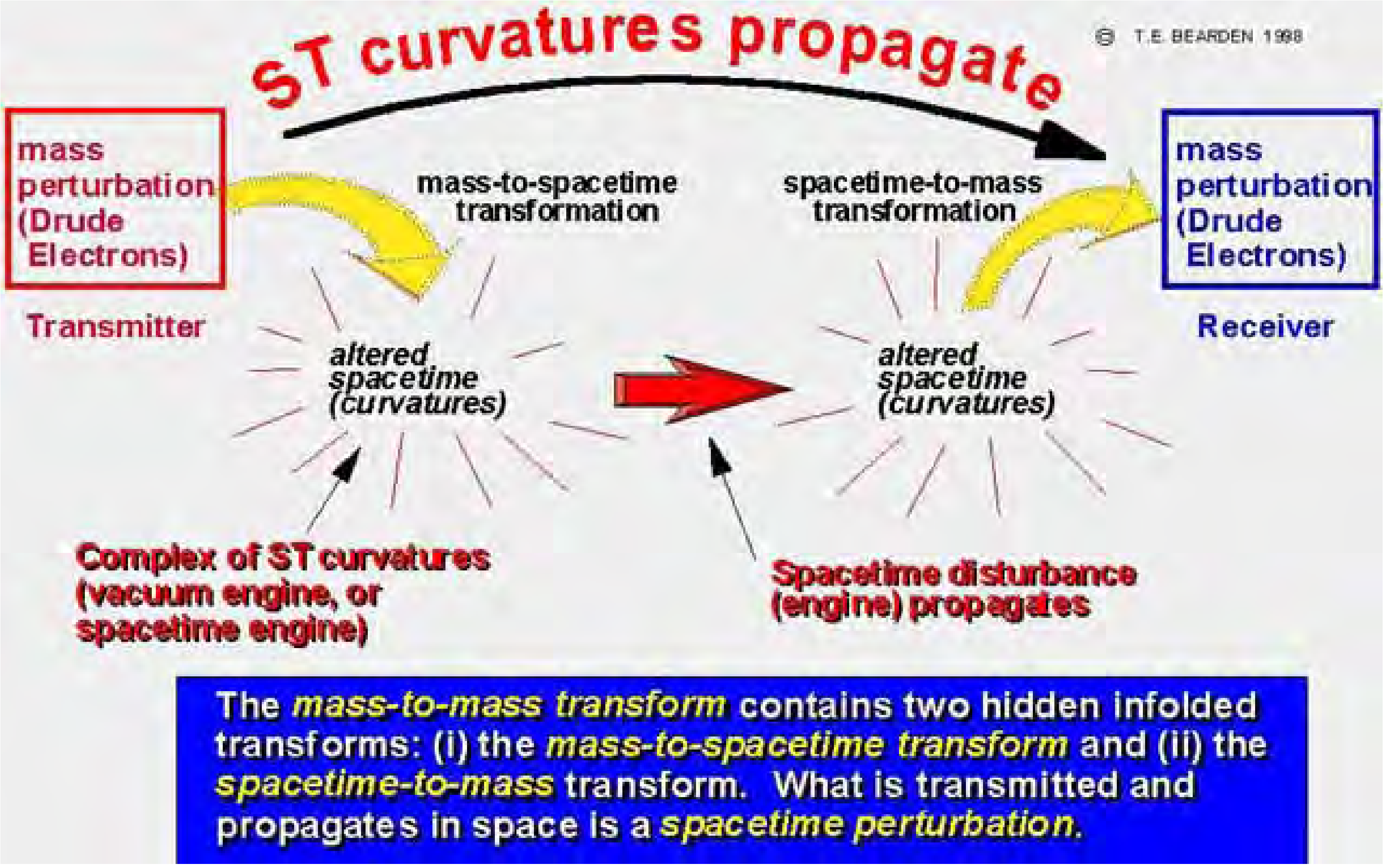

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

Complex of ST curvatures (vacuum engine, or spacetime engine)

spacetime-to-mass transformation

altered spacetime (curvatures)

Spacetime disturbance (engine) propagates

mass perturbation (Drude Electrons)

Receiver

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation*.

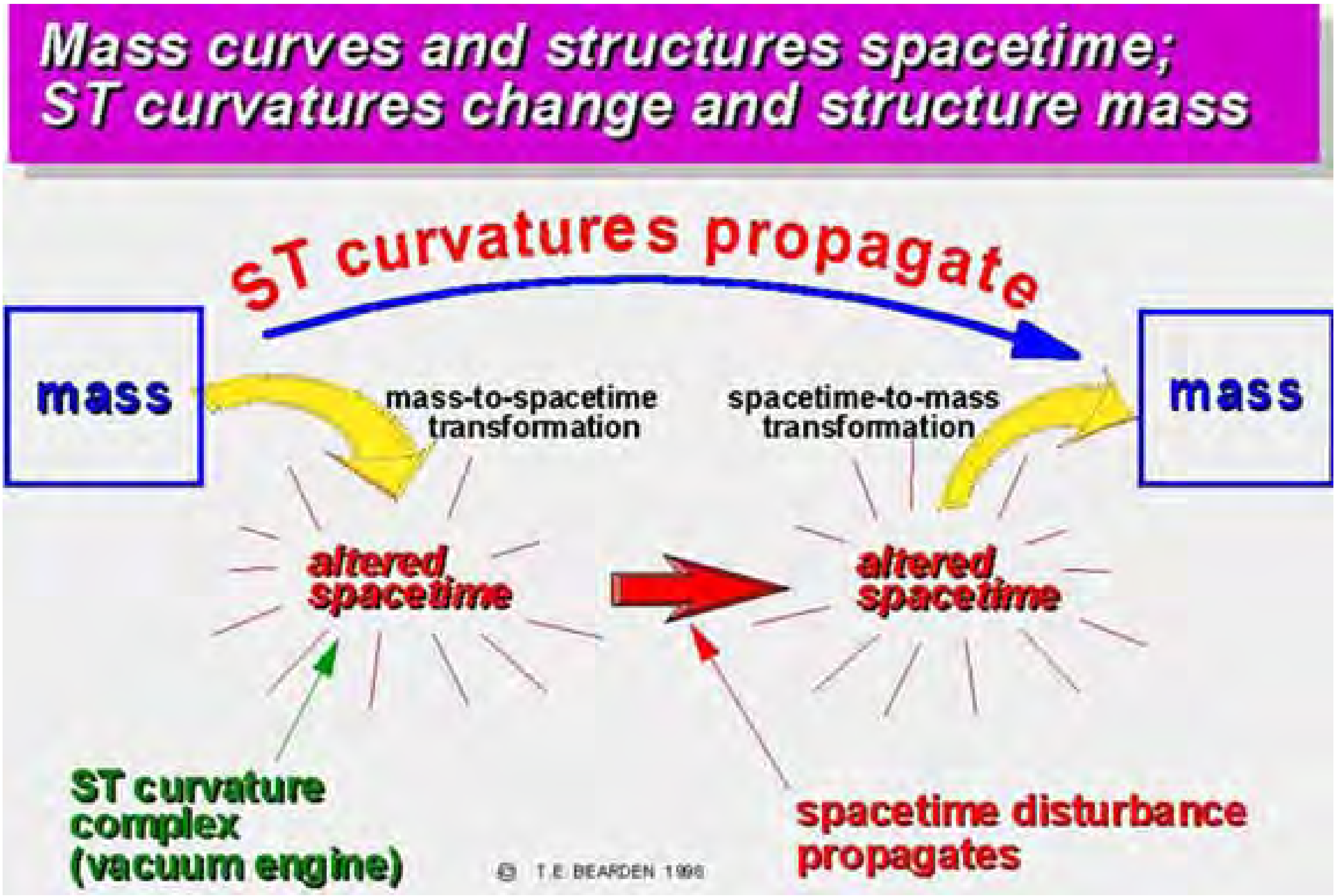

Mass curves and structures spacetime; ST curvatures change and structure mass

ST curvatures propagate

mass

mass-to-spacetime transformation

altered spacetime

spacetime-to-mass transformation

altered spacetime

mass

ST curvature complex (vacuum engine)

spacetime disturbance propagates

© T.E. BEARDEN 1996

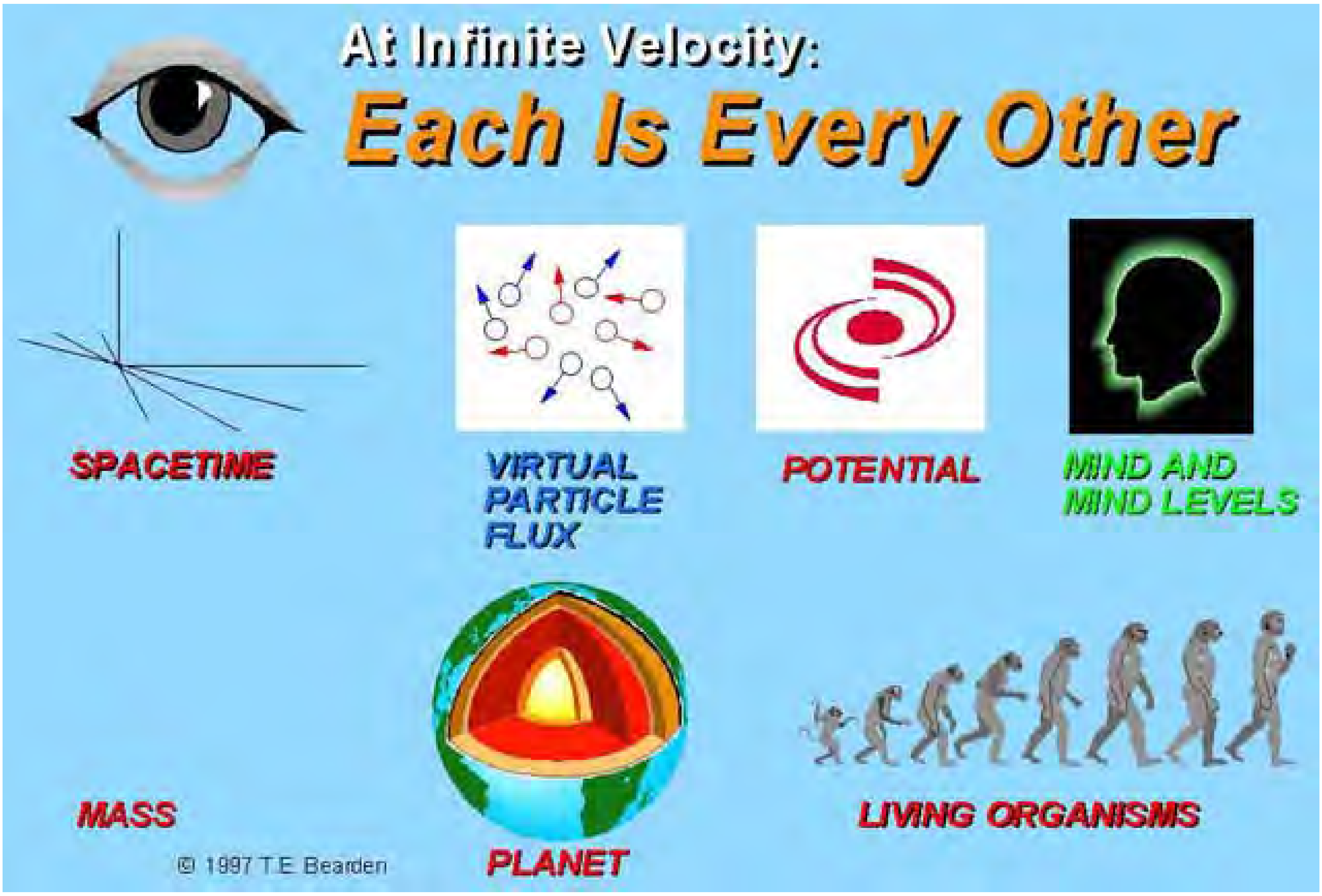

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

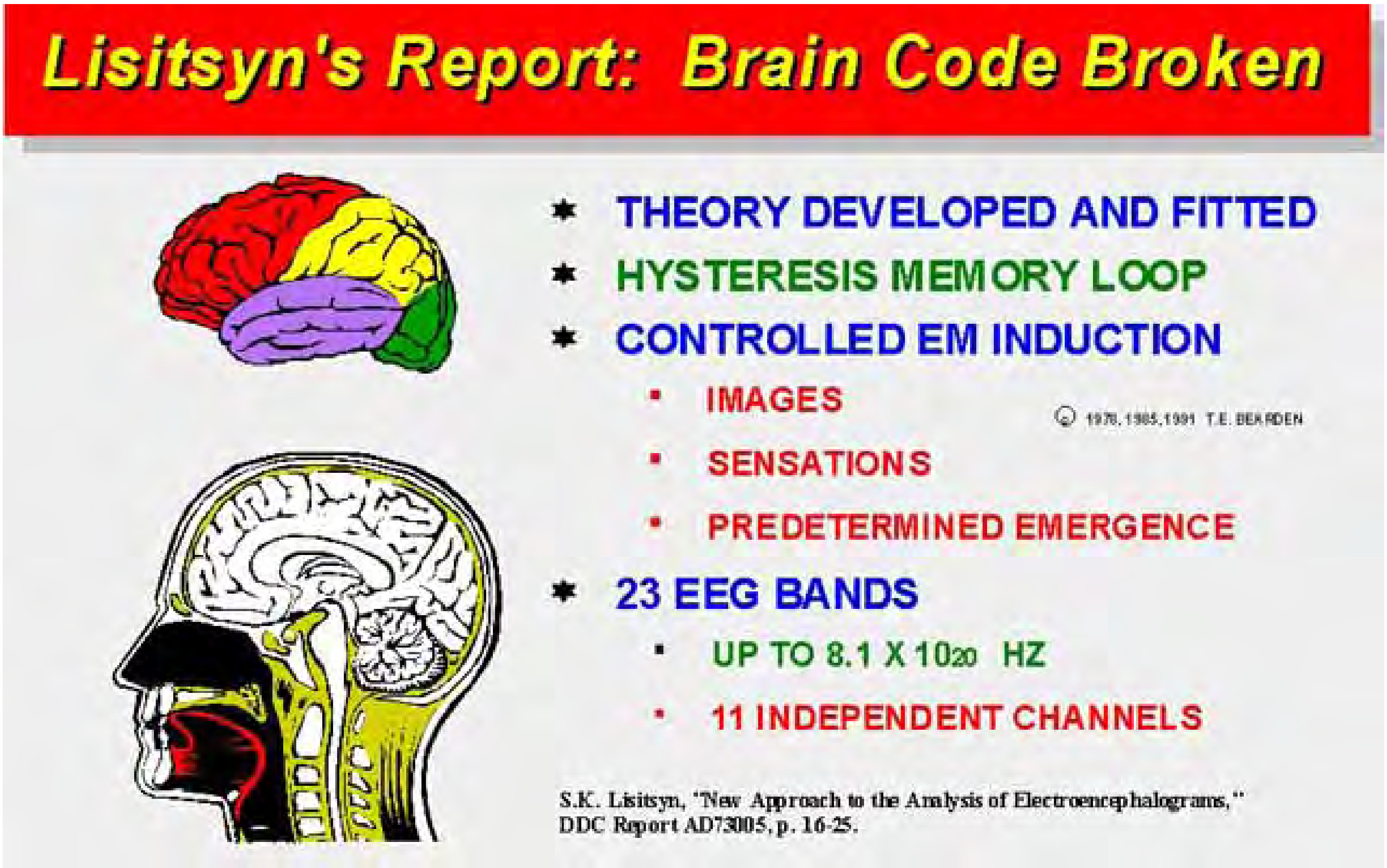

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

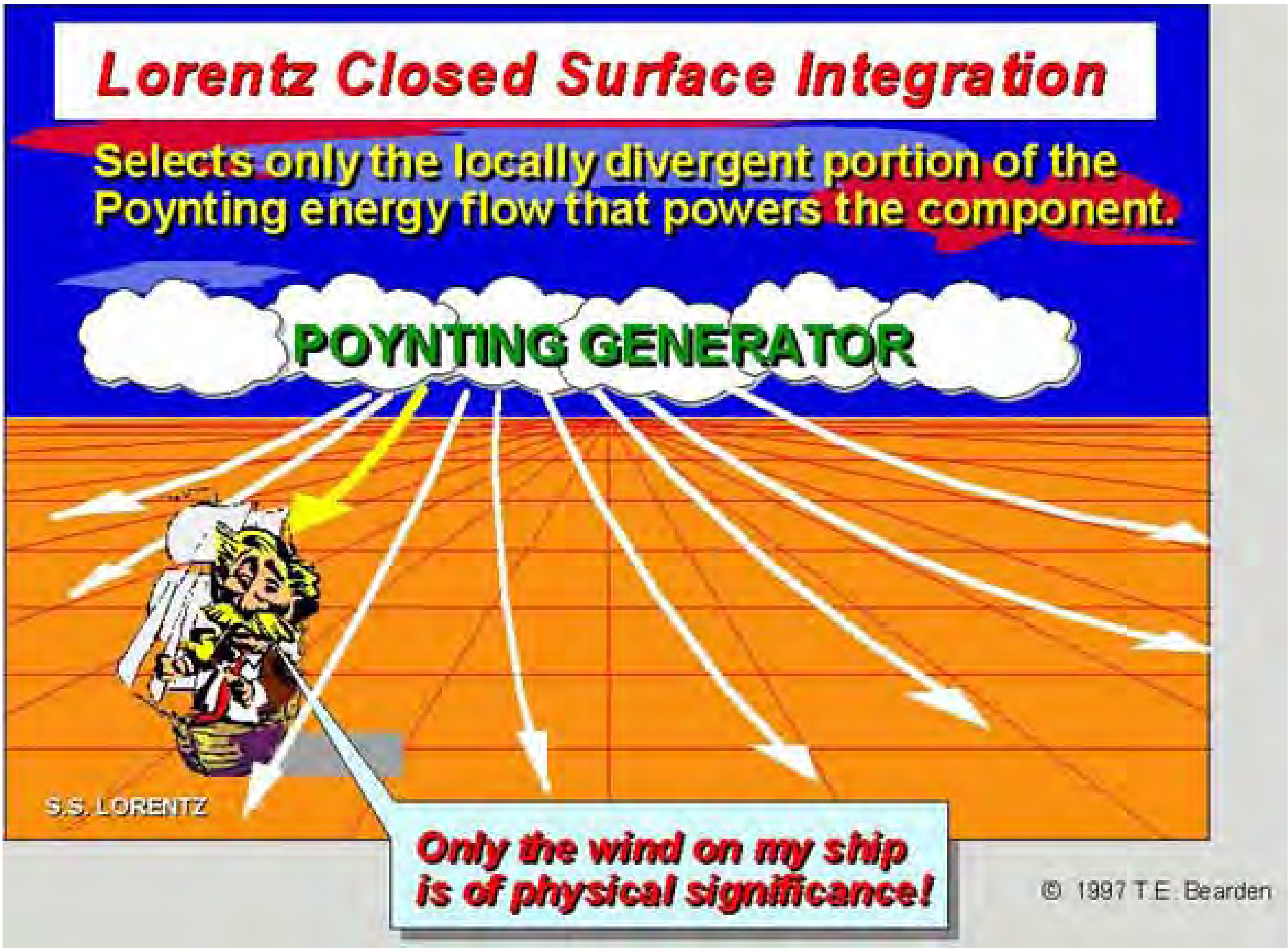

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

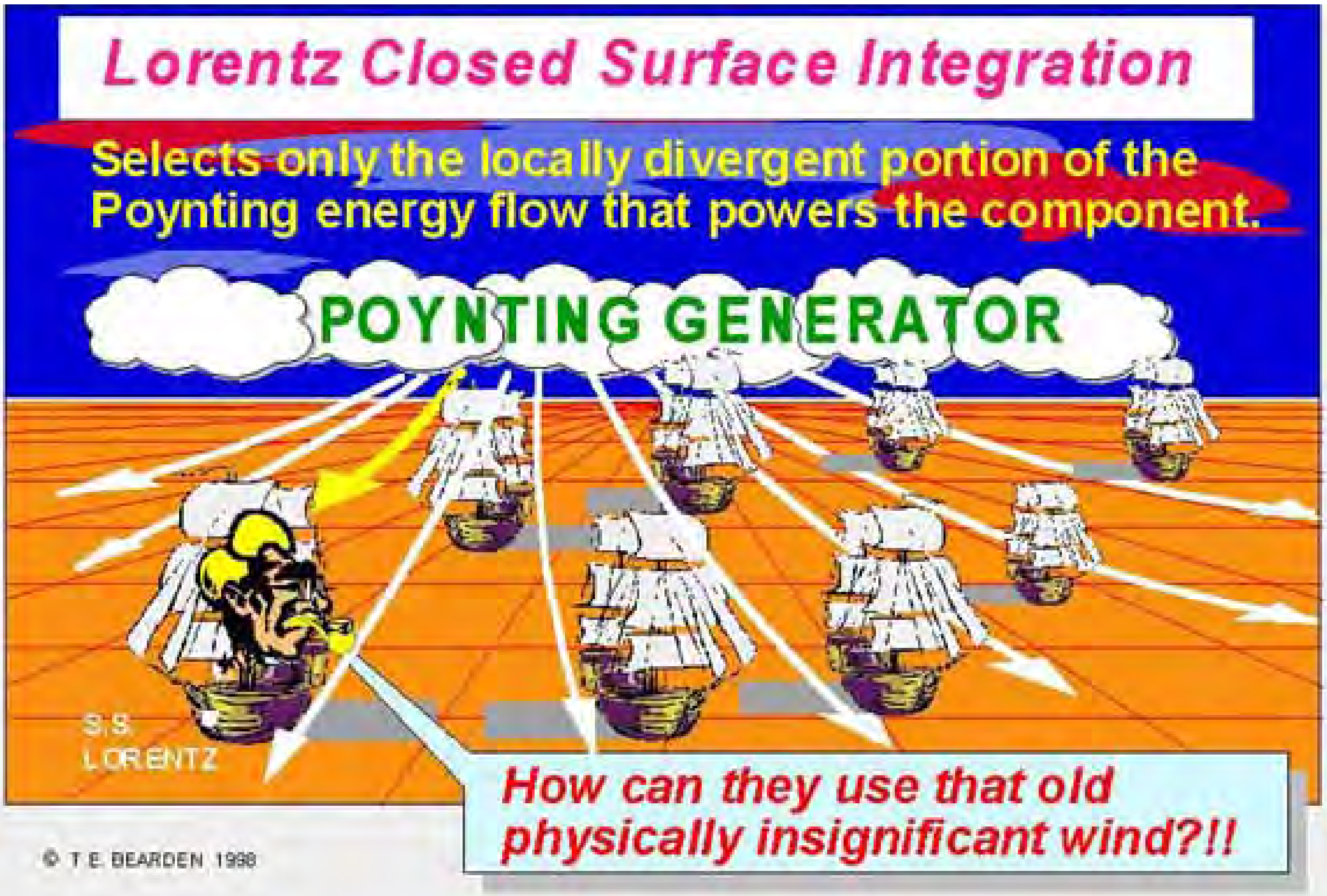

How can they use that old physically insignificant wind?!!

© T.E. BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

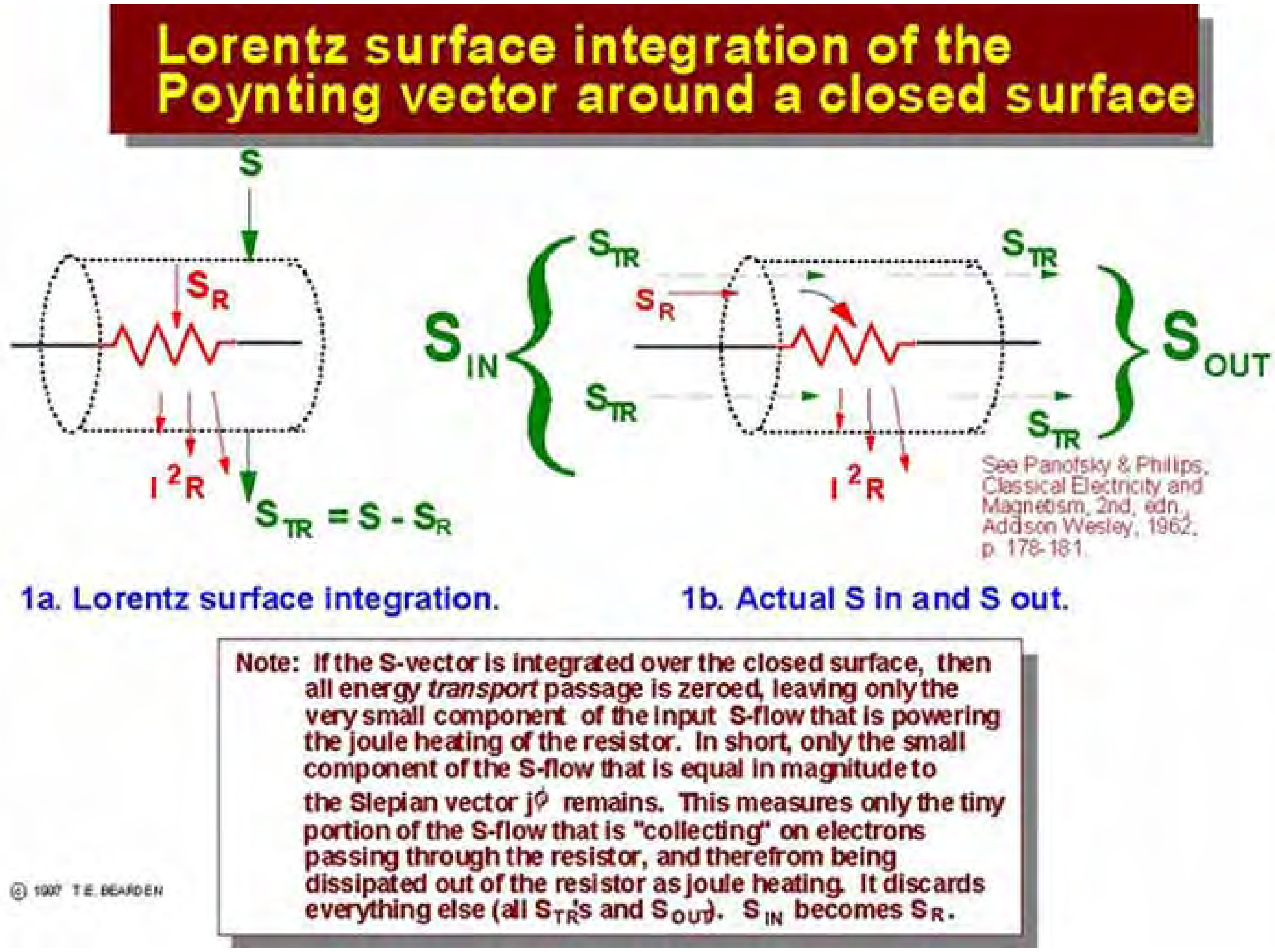

Lorentz surface integration of the Poynting vector around a closed surface

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1a. Lorentz surface integration.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

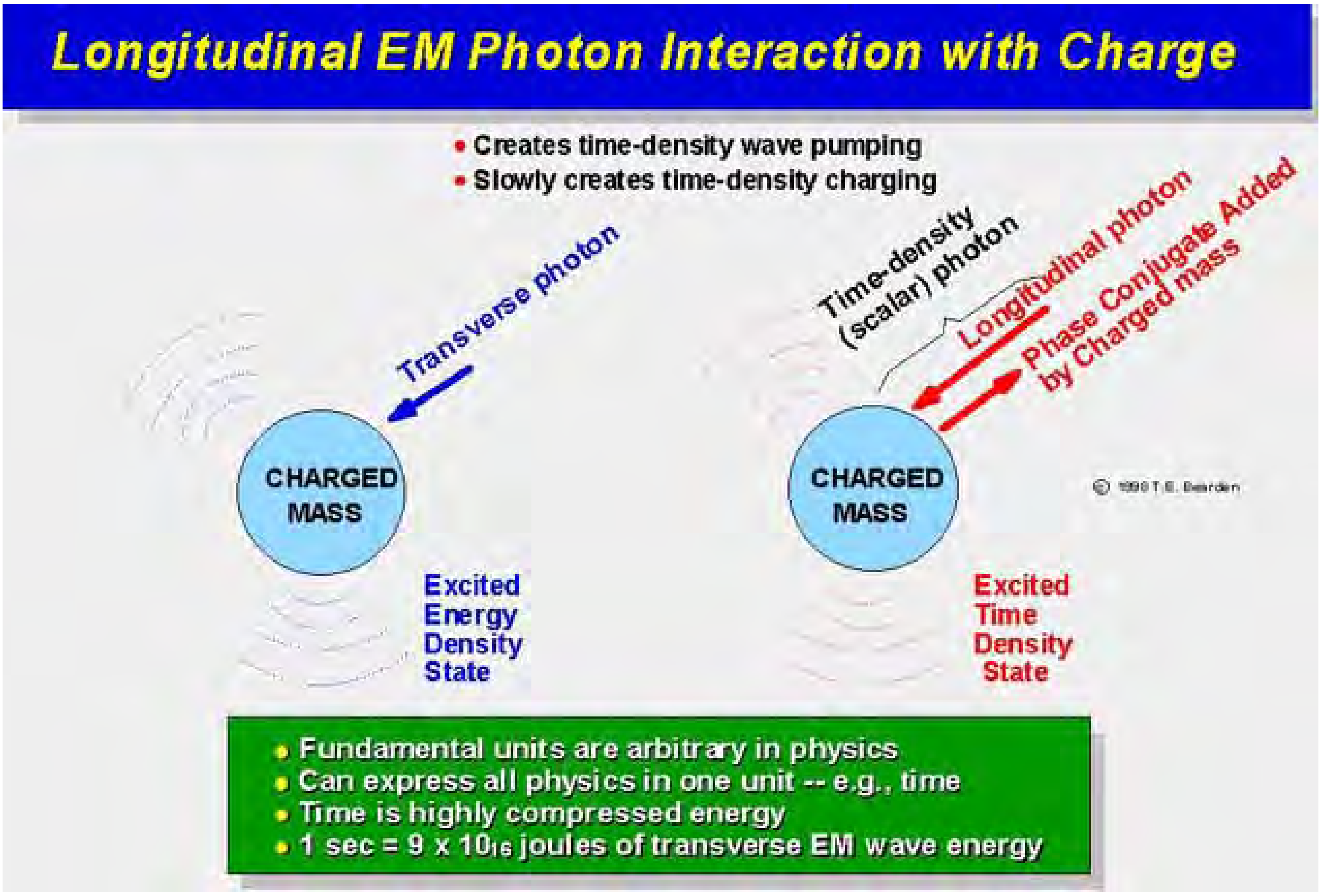

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon
CHARGED MASS
Excited Energy Density State
Time-density (scalar) photon
Longitudinal photon
Phase Conjugate Added by Charged mass
CHARGED MASS
Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9×10^{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

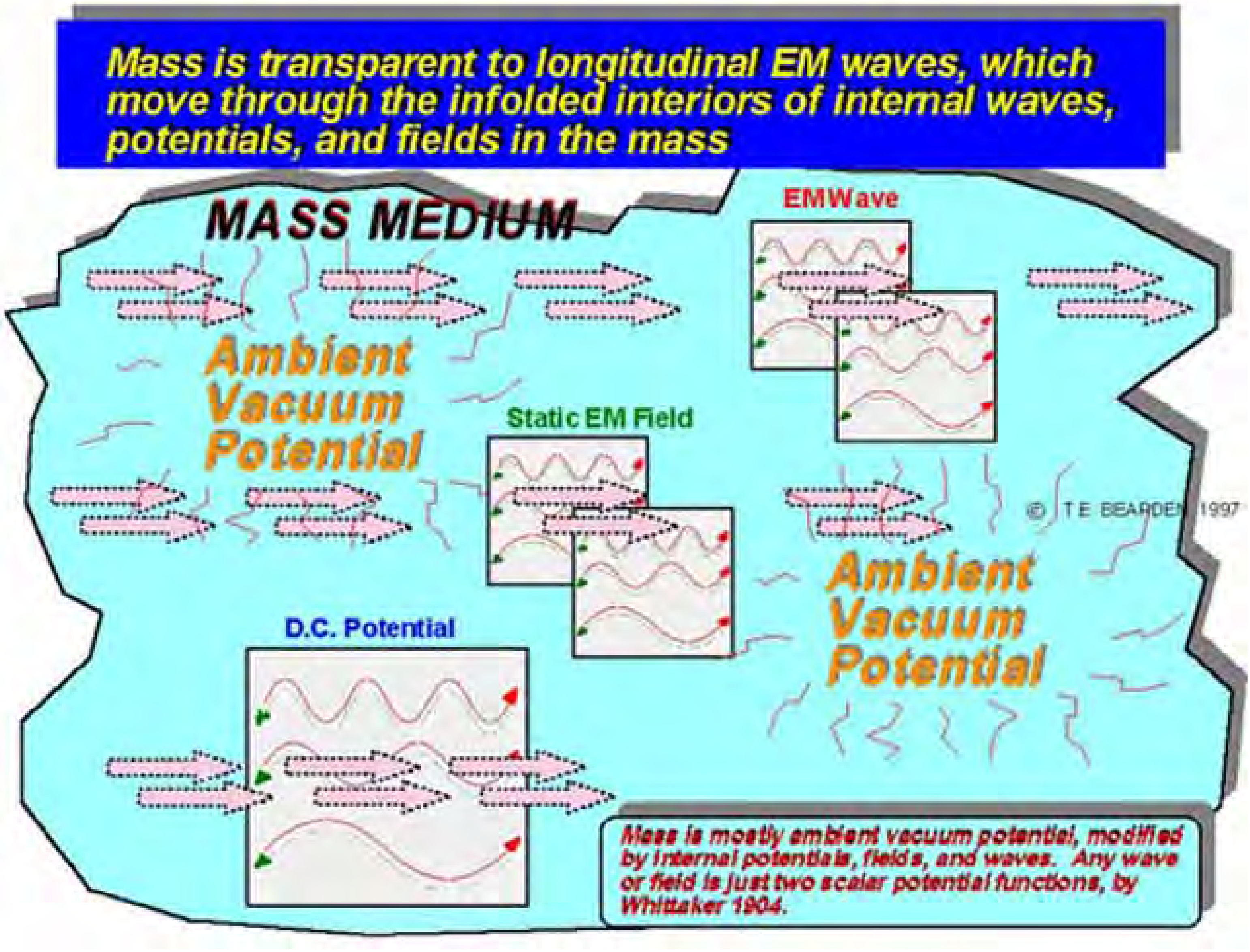

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
EMWave
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

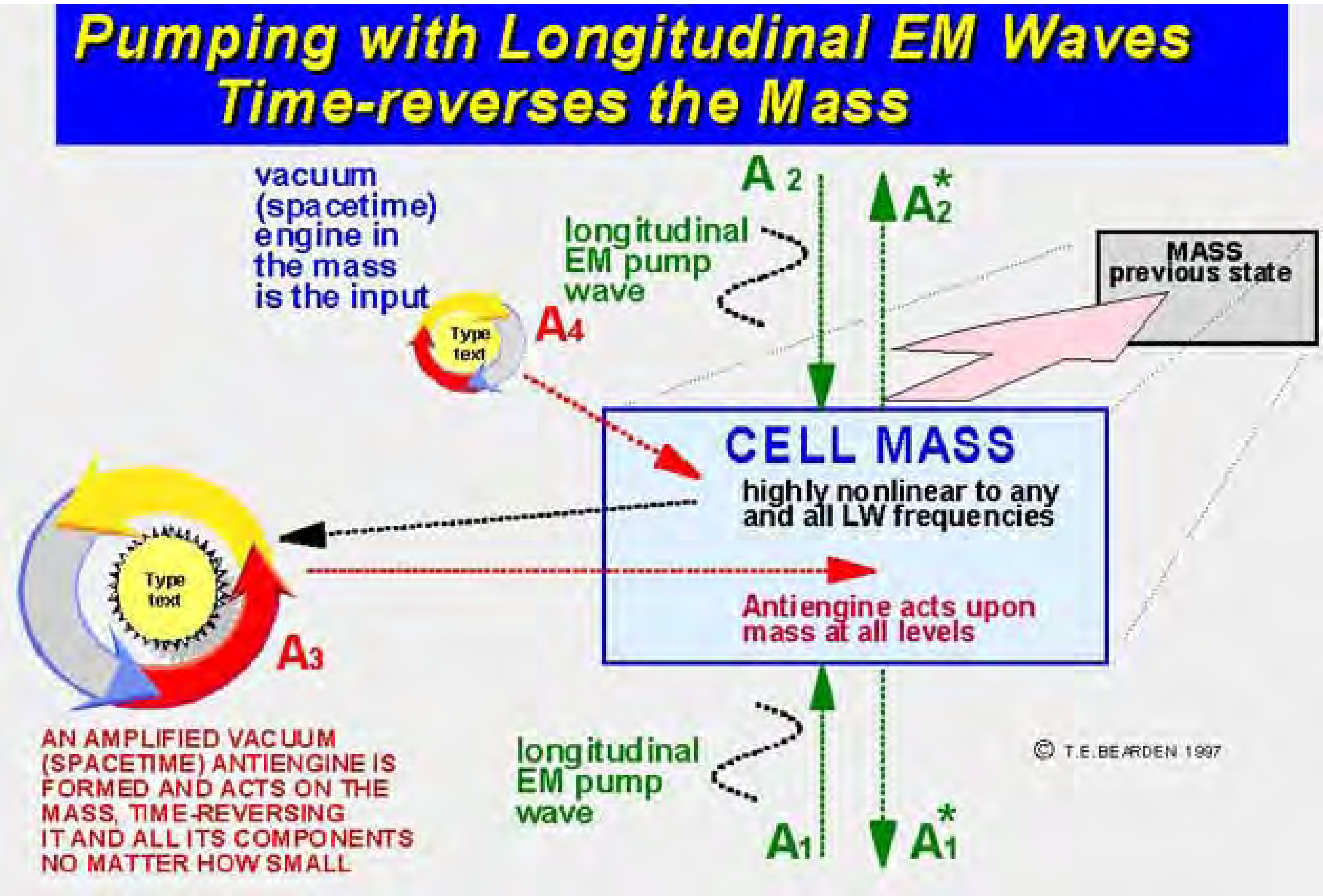

Pumping with Longitudinal EM Waves Time-reverses the Mass

vacuum (spacetime) engine in the mass is the input

longitudinal EM pump wave

MASS previous state

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

longitudinal EM pump wave

© T.E.BEARDEN 1997

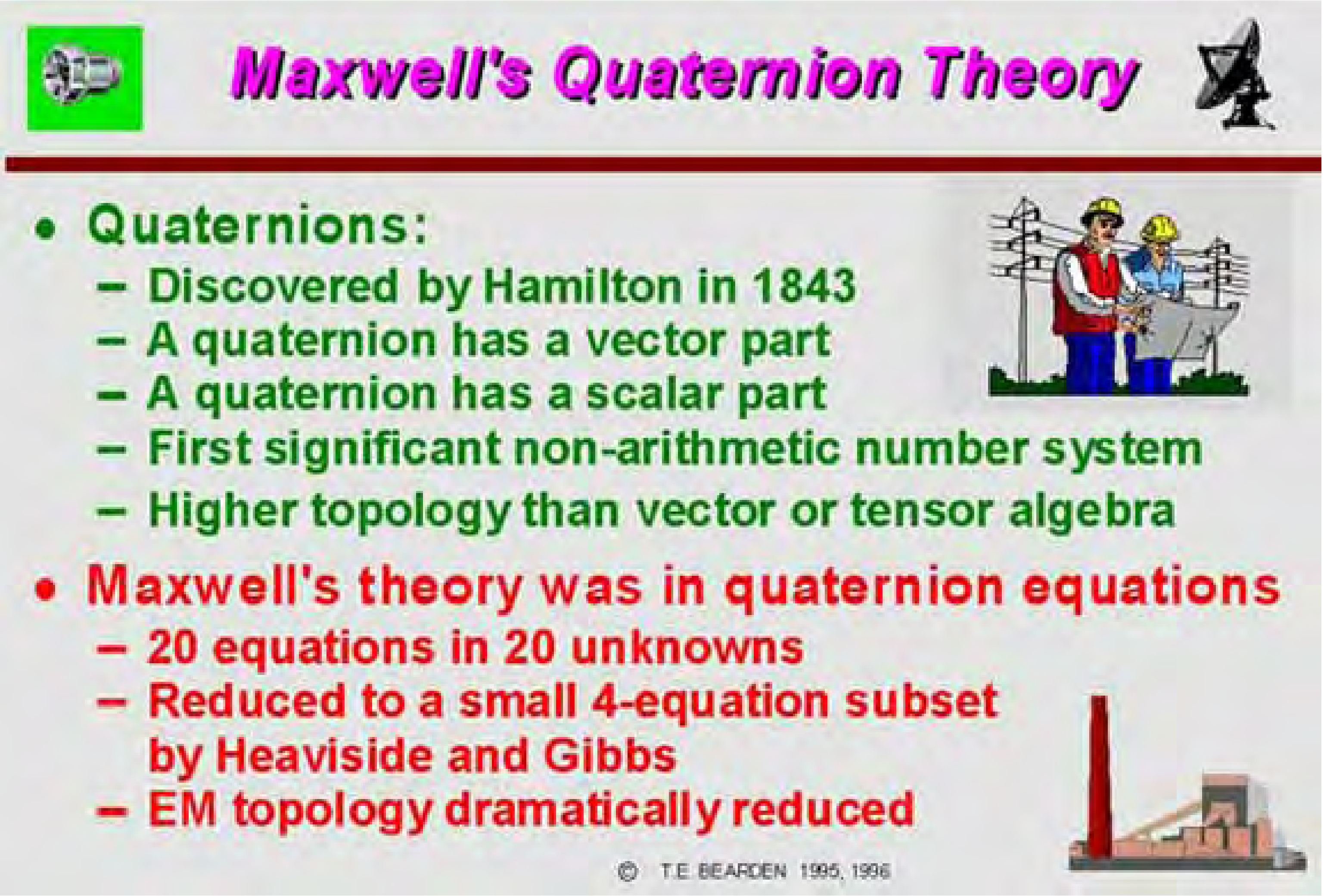

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

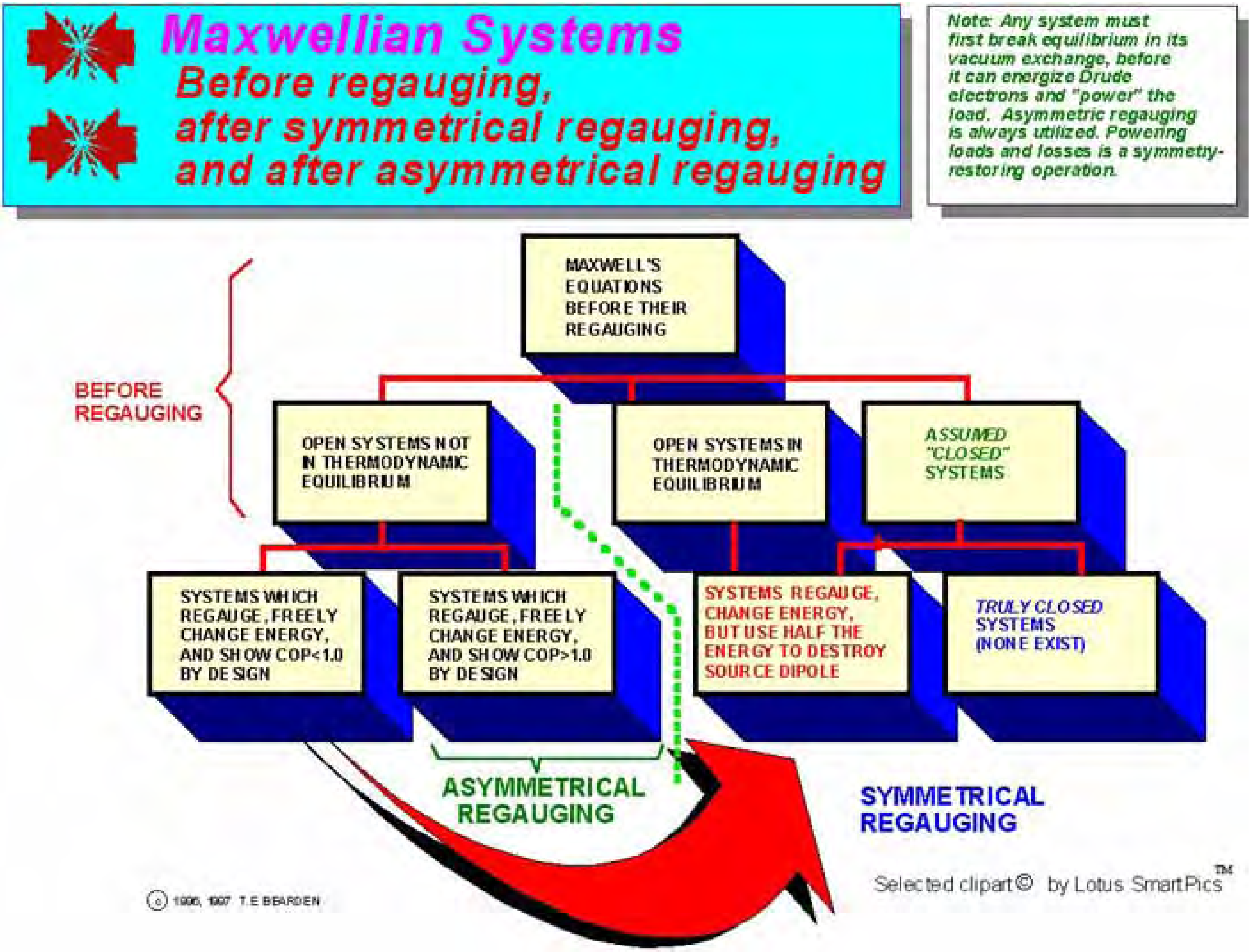

Maxwellian Systems
Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

BEFORE REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

(c) 1996, 1997 T.E BEARDEN

Selected clipart© by Lotus SmartPics™

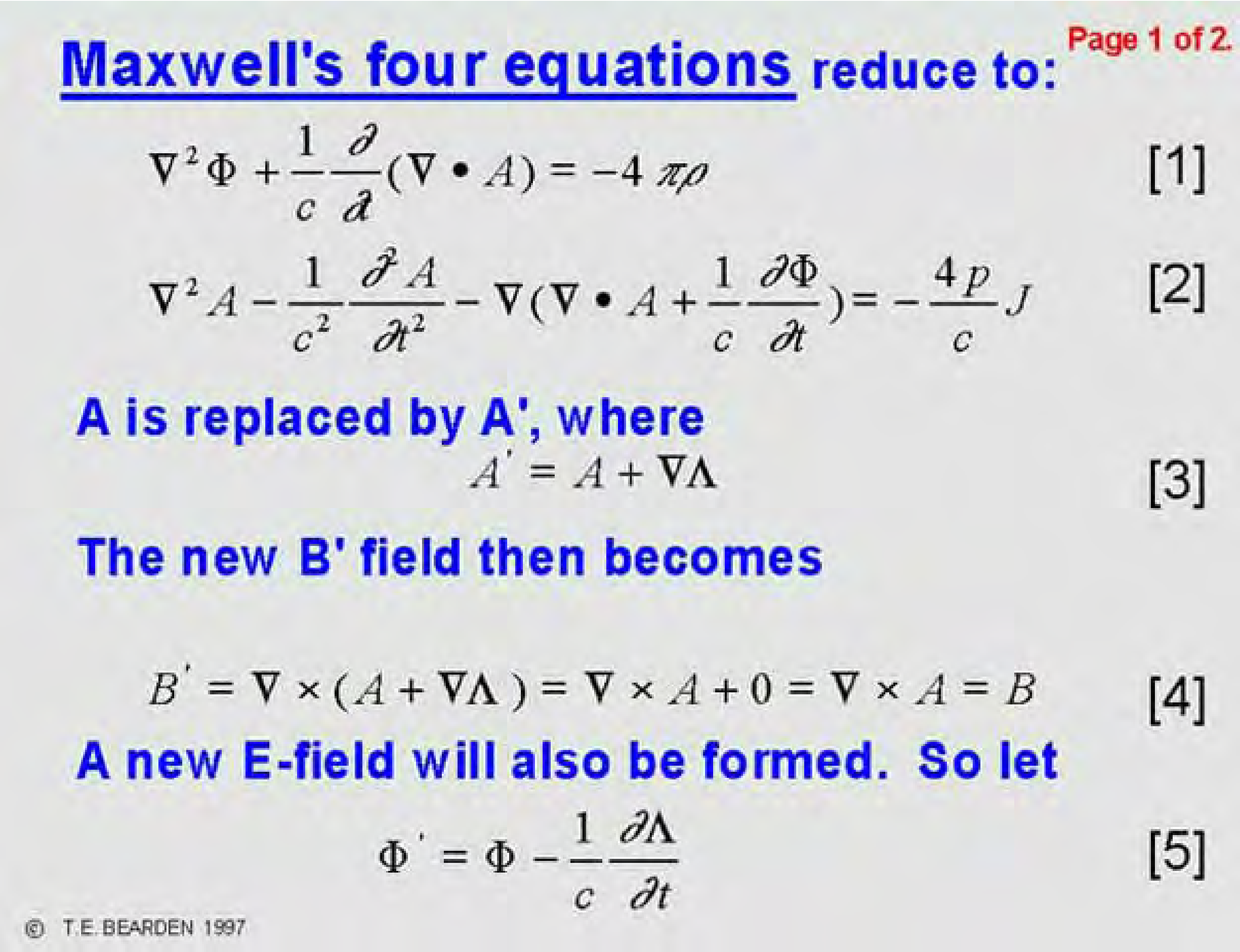

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2 \Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t}\right) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

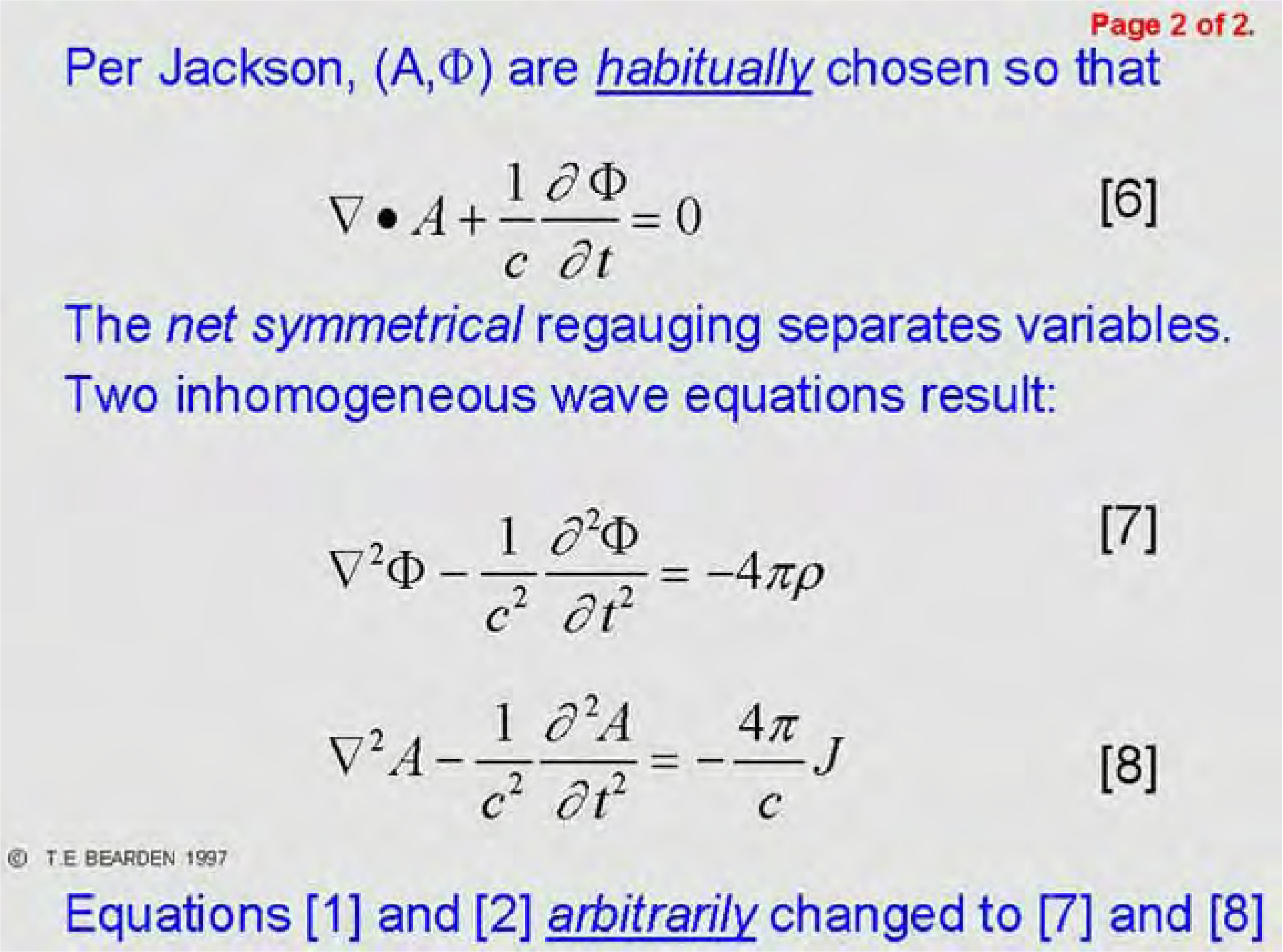

Page 2 of 2.

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables.

Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c}J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Energy, Observations and Comments

---

1. A Conservative Field Requires a Single-Valued Potential
2. Earth-Electrosphere and Earth-Ionosphere Capacitors
3. Earth's Poynting Energy Flow
4. The Hidden Energy Flow around us
5. Electrostatic Fields of the Earth's Electrosphere Capacitor in Fair Weather and Thunderstorms
6. Flow of Current and Energy Into, Through, and Along a Conductor
7. Energy "collected" on a charge
8. Fiber Fuse Effect
9. Impact of Vacuum Engineering
10. Lawandy's Experiment
11. Field may be nonconservative for a multivalued potential
12. The Ostrich Problem - Part 1
13. The Ostrich Problem - Part 2
14. Russia had Overunity Devices in 1930s
15. Energy output of human body
16. A "static" Poynting S-flow free-energy generator
17. Topology and Energy Conservation
18. Block Diagram: Sweet's Vacuum Triode

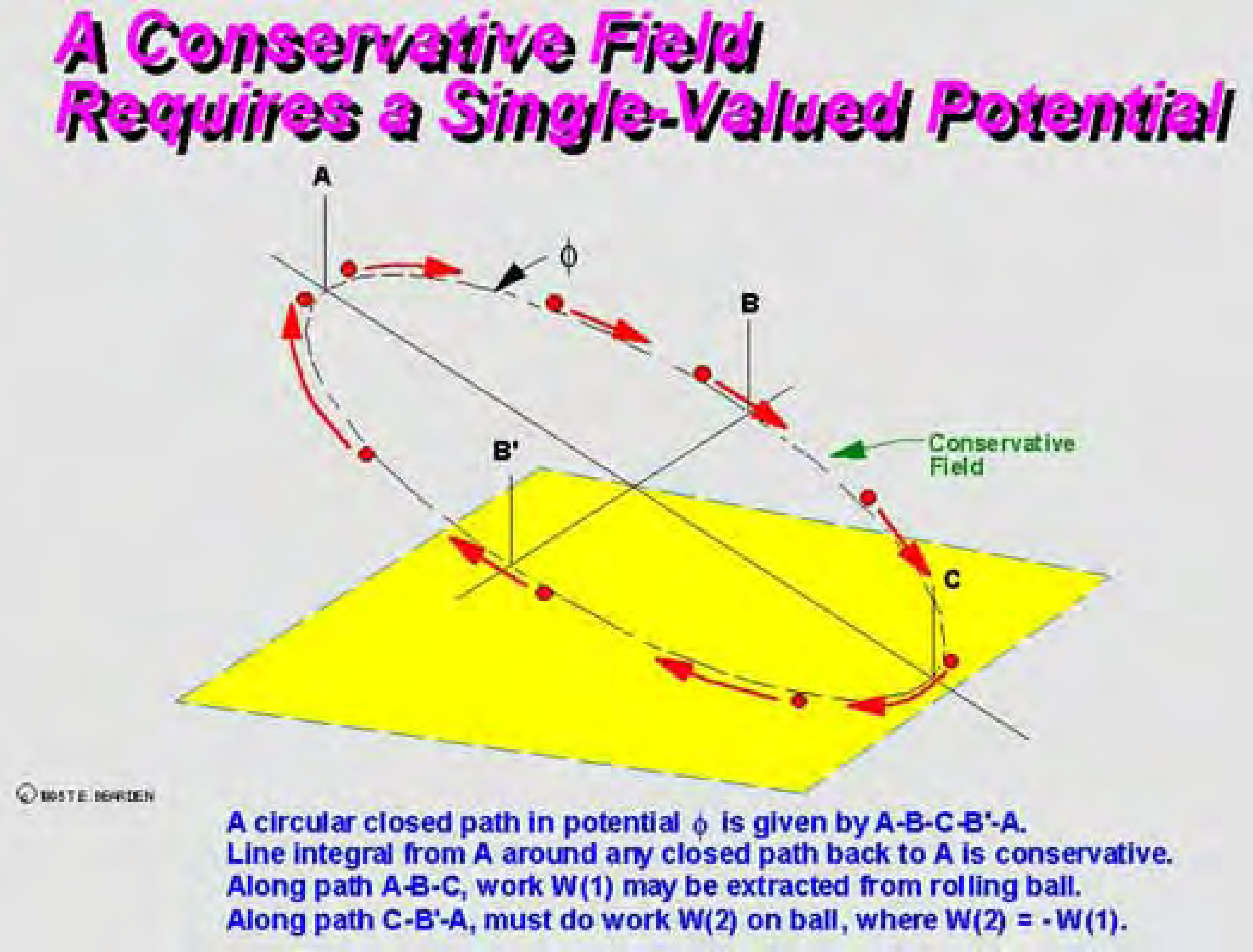

A Conservative Field Requires a Single-Valued Potential

© 1995 T.E. BEARDEN

A circular closed path in potential ϕ is given by A-B-C-B'-A.
Line integral from A around any closed path back to A is conservative.
Along path A-B-C, work W(1) may be extracted from rolling ball.
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1).

Return to Slide Index

Next Slide

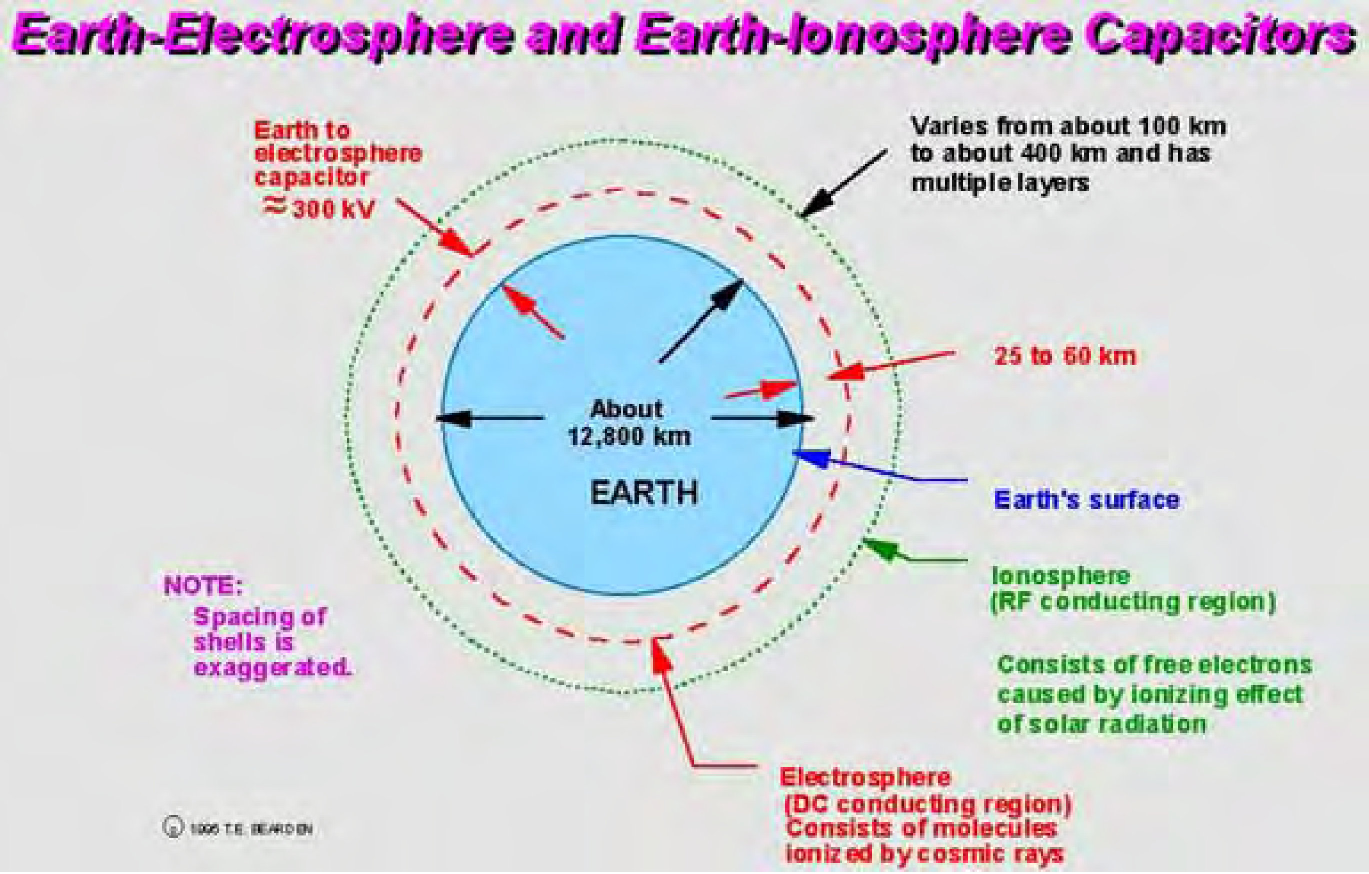

Earth-Electrosphere and Earth-Ionosphere Capacitors

Earth to electrosphere capacitor ≈ 300 kV

Varies from about 100 km to about 400 km and has multiple layers

About 12,800 km

EARTH

25 to 60 km

Earth's surface

Ionosphere (RF conducting region)

Consists of free electrons caused by ionizing effect of solar radiation

NOTE: Spacing of shells is exaggerated.

Electrosphere (DC conducting region) Consists of molecules ionized by cosmic rays

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

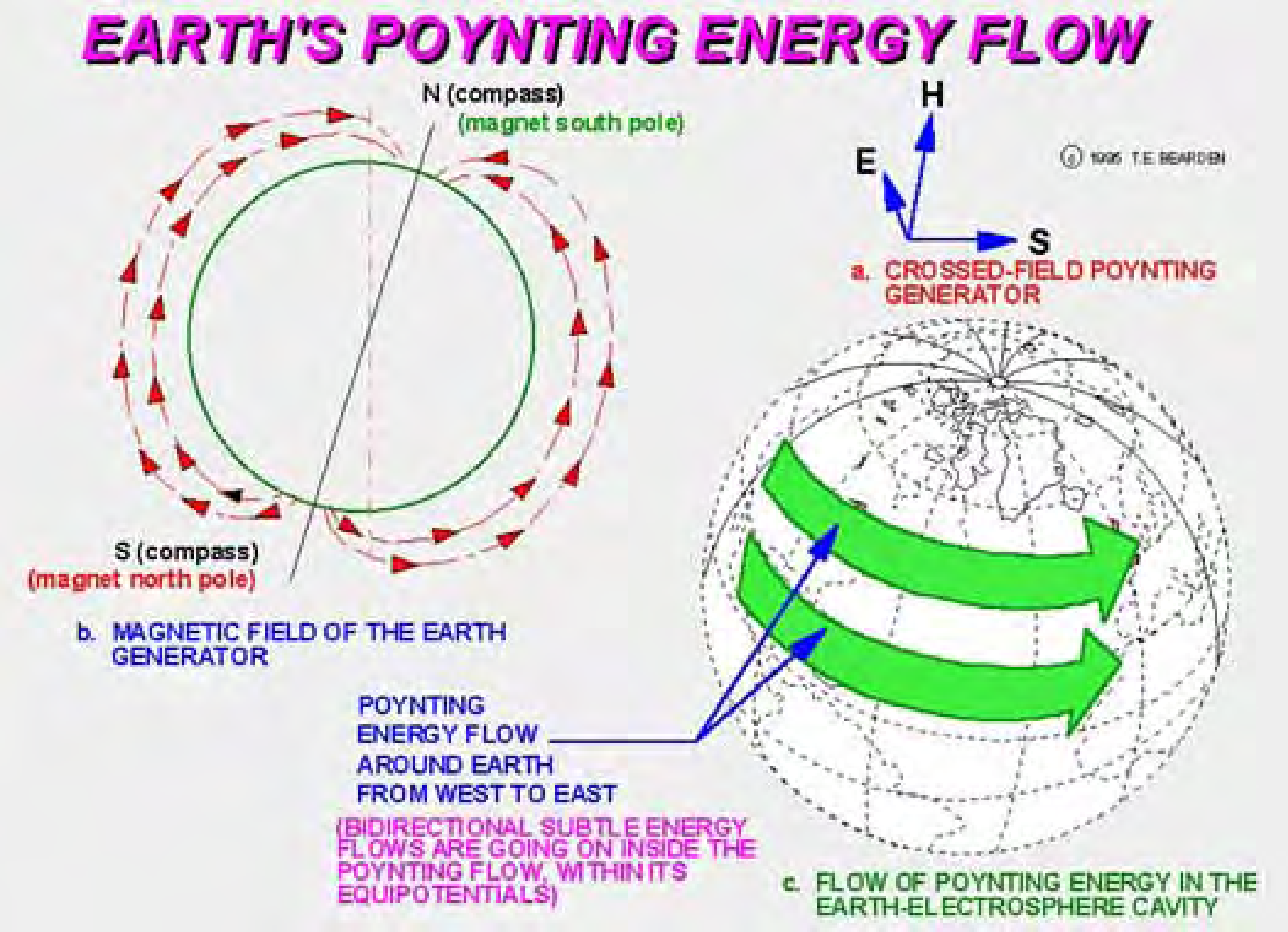
N (compass)
(magnet south pole)
S (compass)
(magnet north pole)
H
E
S
© 1995 T.E. BEARDEN

EARTH'S POYNTING ENERGY FLOW

a. CROSSED-FIELD POYNTING GENERATOR

b. MAGNETIC FIELD OF THE EARTH GENERATOR

POYNTING ENERGY FLOW AROUND EARTH FROM WEST TO EAST

(BIDIRECTIONAL SUBTLE ENERGY FLOWS ARE GOING ON INSIDE THE POYNTING FLOW, WITHIN ITS EQUIPOTENTIALS)

c. FLOW OF POYNTING ENERGY IN THE EARTH-ELECTROSPHERE CAVITY

Return to Slide Index

Next Slide

Previous Slide

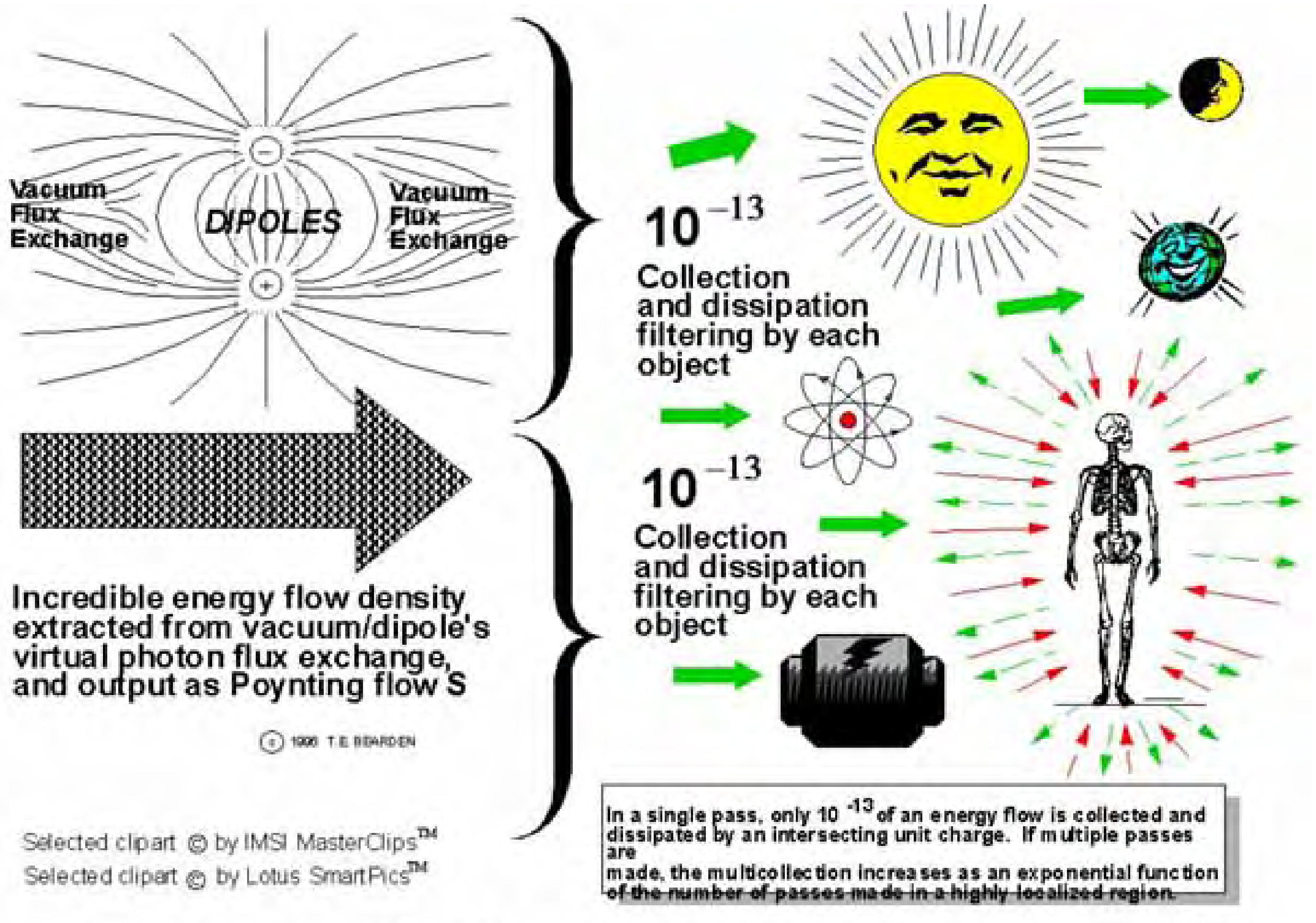
Vacuum Flux Exchange
DIPOLES
Vacuum Flux Exchange
Incredible energy flow density extracted from vacuum/dipole's virtual photon flux exchange, and output as Poynting flow S
© 1996 T.E. BEARDEN
10^{-13}
Collection and dissipation filtering by each object
10^{-13}
Collection and dissipation filtering by each object
Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

In a single pass, only 10^{-13} of an energy flow is collected and dissipated by an intersecting unit charge. If multiple passes are made, the multicollection increases as an exponential function of the number of passes made in a highly localized region.

Figure 19. The world of physical matter exists behind a dramatic 10^{-13} filter in a fiery cauldron of fierce and primeval EM energy flow.

Return to Slide Index

Next Slide

Previous Slide

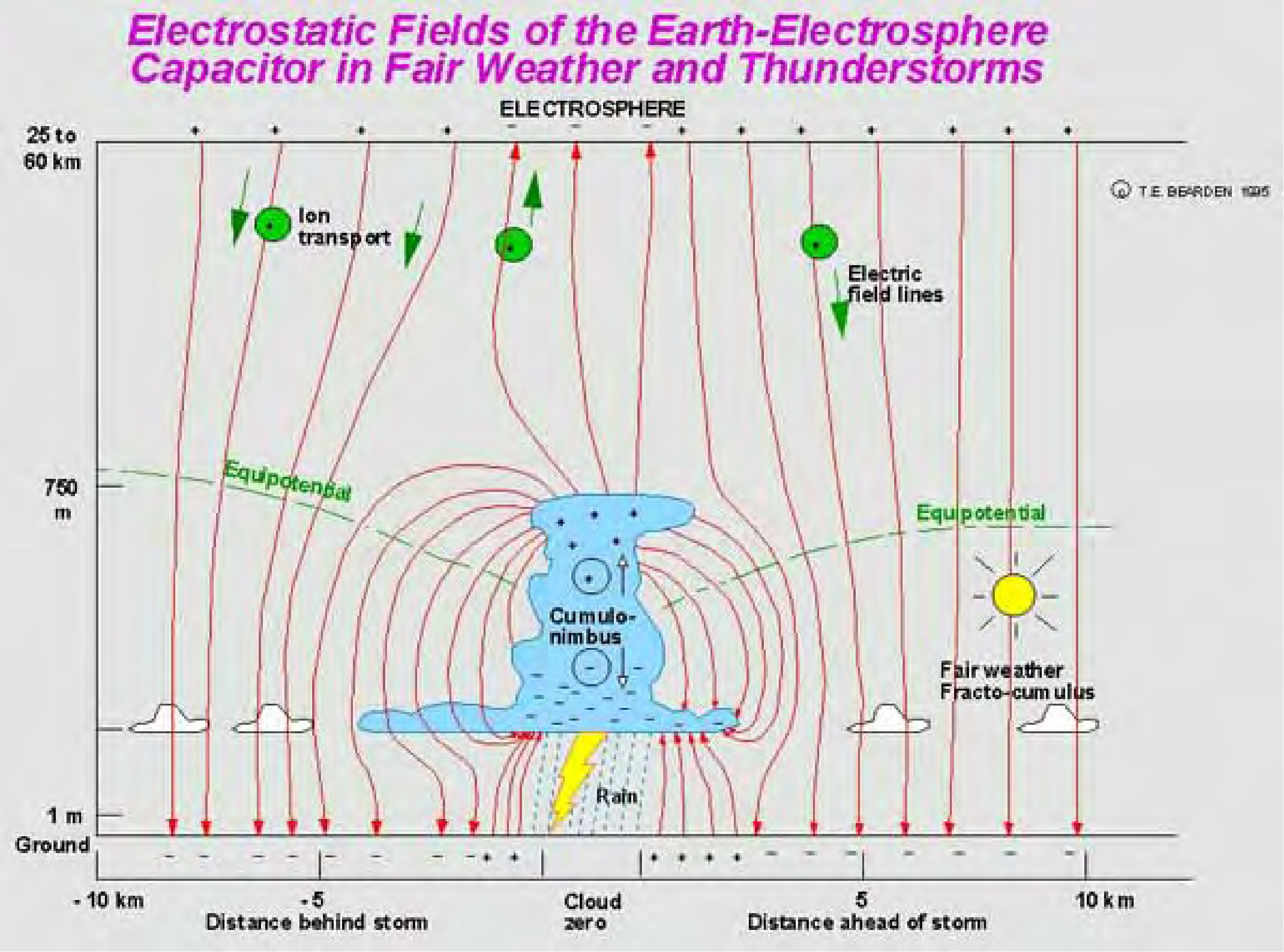
Electrostatic Fields of the Earth-Electrosphere Capacitor in Fair Weather and Thunderstorms

ELECTROSPHERE
25 to 60 km
Ion transport
T.E. BEARDEN 1995
Electric field lines
750 m
Equipotential
Equipotential
Cumulo-nimbus
Fair weather Fracto-cumulus
Rain
1 m
Ground
- 10 km
- 5
Distance behind storm
Cloud zero
5
Distance ahead of storm
10 km

[Return to Slide Index]

[Next Slide]

[Previous Slide]

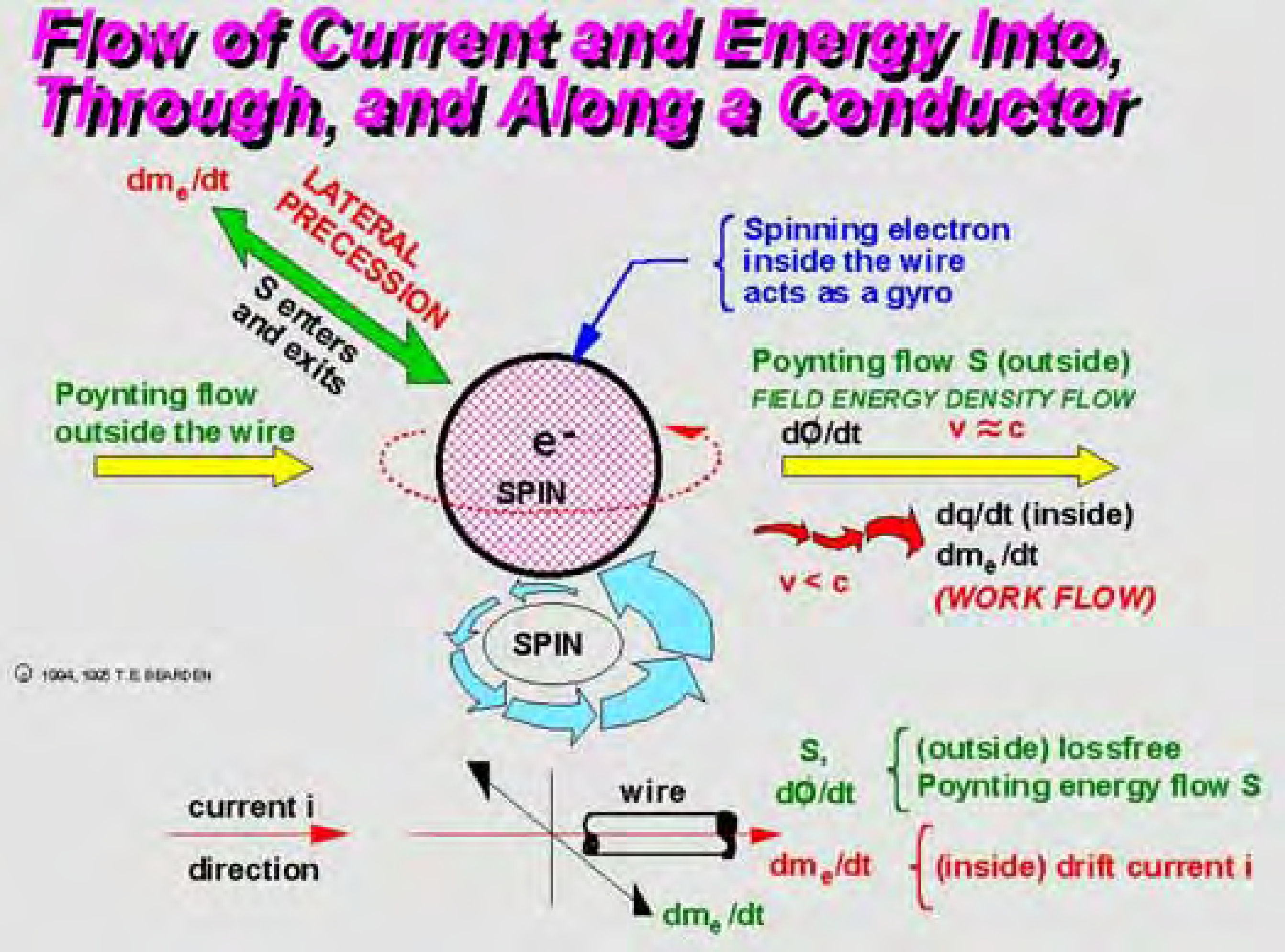

Flow of Current and Energy Into, Through, and Along a Conductor

dm_e/dt

LATERAL PRECESSION

S enters and exits

Spinning electron
inside the wire
acts as a gyro

Poynting flow
outside the wire

e^-

SPIN

Poynting flow S (outside)

FIELD ENERGY DENSITY FLOW

$d\phi/dt$ $v \approx c$

dq/dt (inside)

dm_e/dt

$v < c$

(WORK FLOW)

SPIN

© 1994, 1995 T.E. BEARDEN

current i

direction

wire

S,
$d\phi/dt$

(outside) lossfree
Poynting energy flow S

dm_e/dt

(inside) drift current i

dm_e/dt

[Return to Slide Index]

[Next Slide]

[Previous Slide]

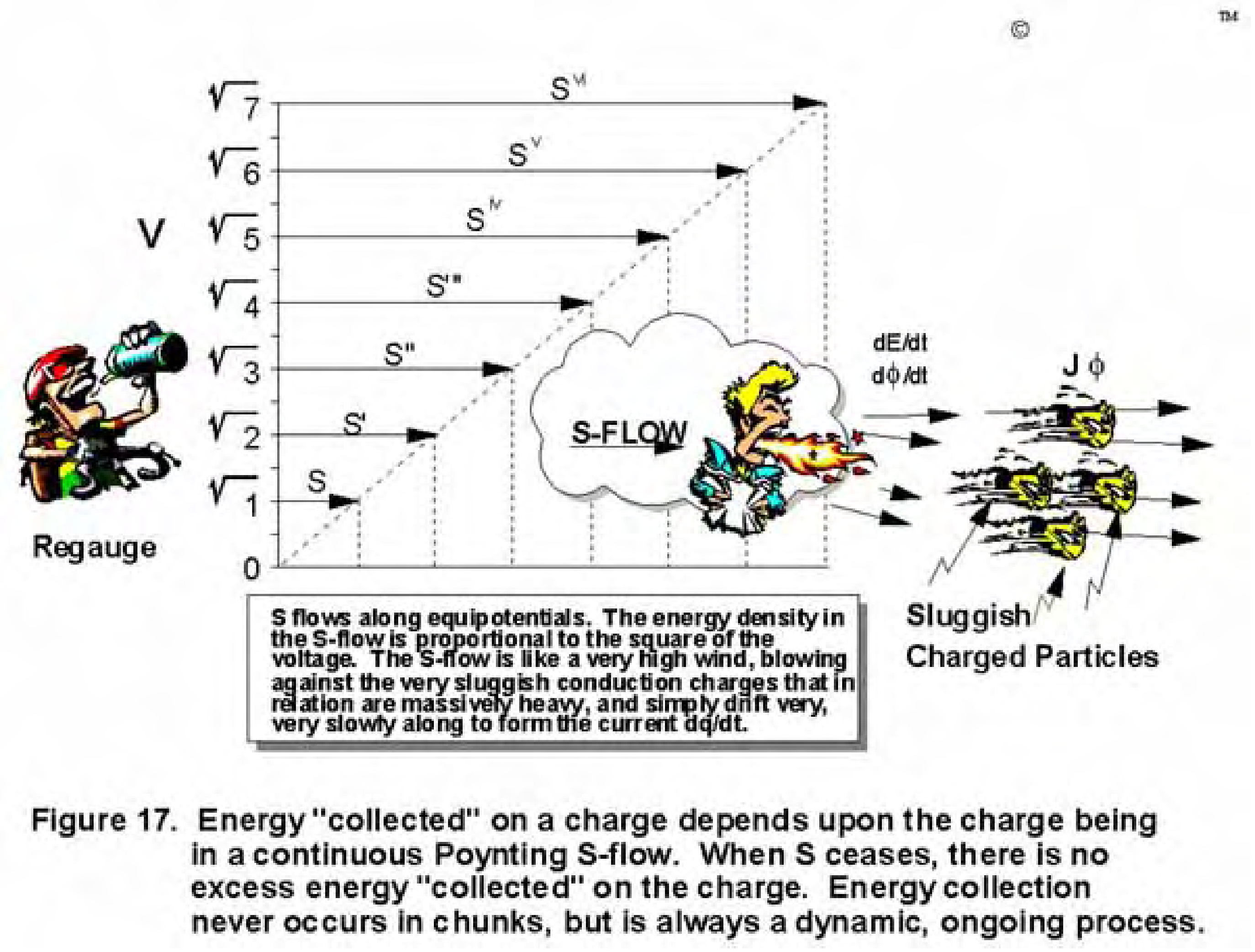
©
™
V
$\sqrt{7}$
$\sqrt{6}$
$\sqrt{5}$
$\sqrt{4}$
$\sqrt{3}$
$\sqrt{2}$
$\sqrt{1}$
0
S^{VI}
S^{V}
S^{IV}
S'''
S''
S'
S
Regauge
S-FLOW
dE/dt
dφ/dt
J φ
Sluggish
Charged Particles

S flows along equipotentials. The energy density in the S-flow is proportional to the square of the voltage. The S-flow is like a very high wind, blowing against the very sluggish conduction charges that in relation are massively heavy, and simply drift very, very slowly along to form the current dq/dt.

Figure 17. Energy "collected" on a charge depends upon the charge being in a continuous Poynting S-flow. When S ceases, there is no excess energy "collected" on the charge. Energy collection never occurs in chunks, but is always a dynamic, ongoing process.

Return to Slide Index

Next Slide

Previous Slide

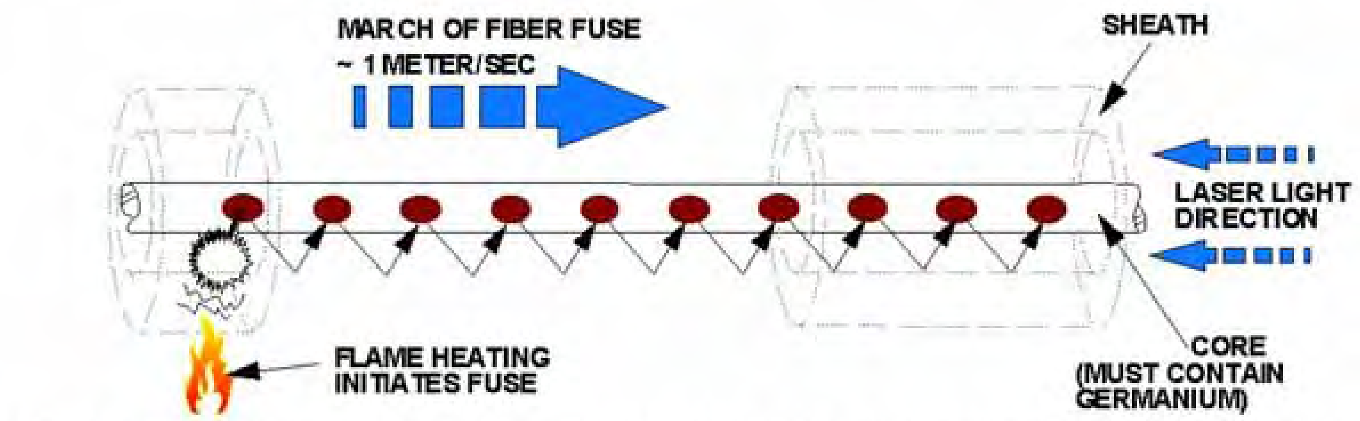
MARCH OF FIBER FUSE
~ 1 METER/SEC
SHEATH
LASER LIGHT DIRECTION
FLAME HEATING INITIATES FUSE
CORE (MUST CONTAIN GERMANIUM)

31A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

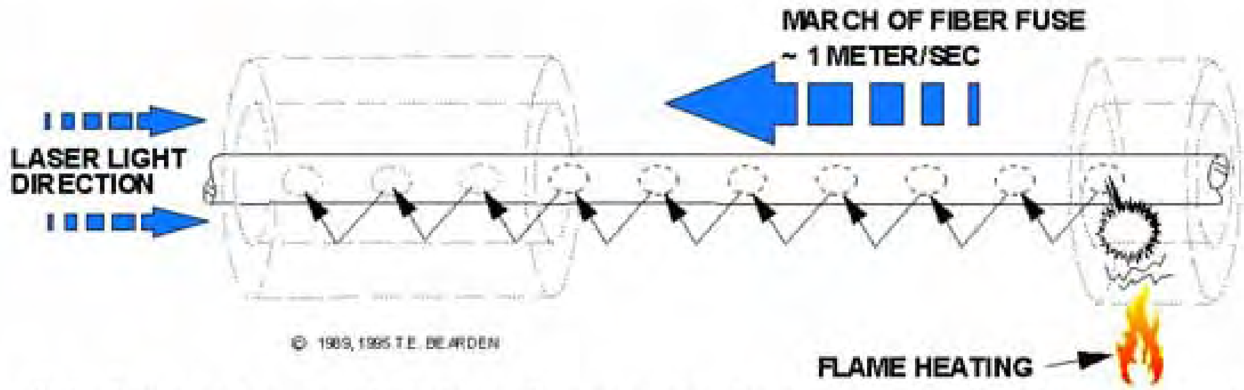
MARCH OF FIBER FUSE
~ 1 METER/SEC
LASER LIGHT DIRECTION
FLAME HEATING

© 1989, 1995 T.E. BEARDEN

31B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 31. Fiber fuse effect and its strange anomalies.

Return to Slide Index

Next Slide

Previous Slide

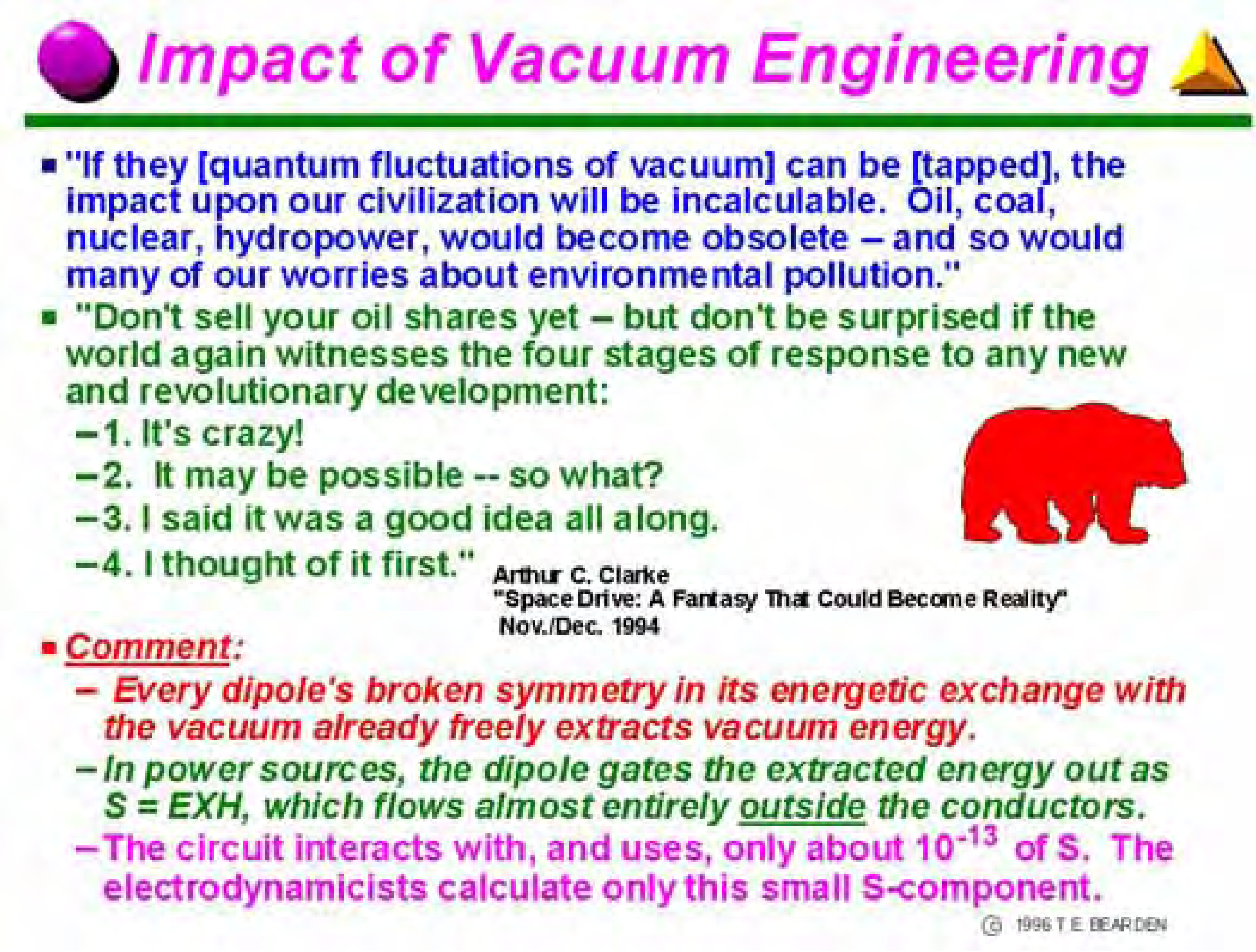

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete – and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet – but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

© 1996 T E BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

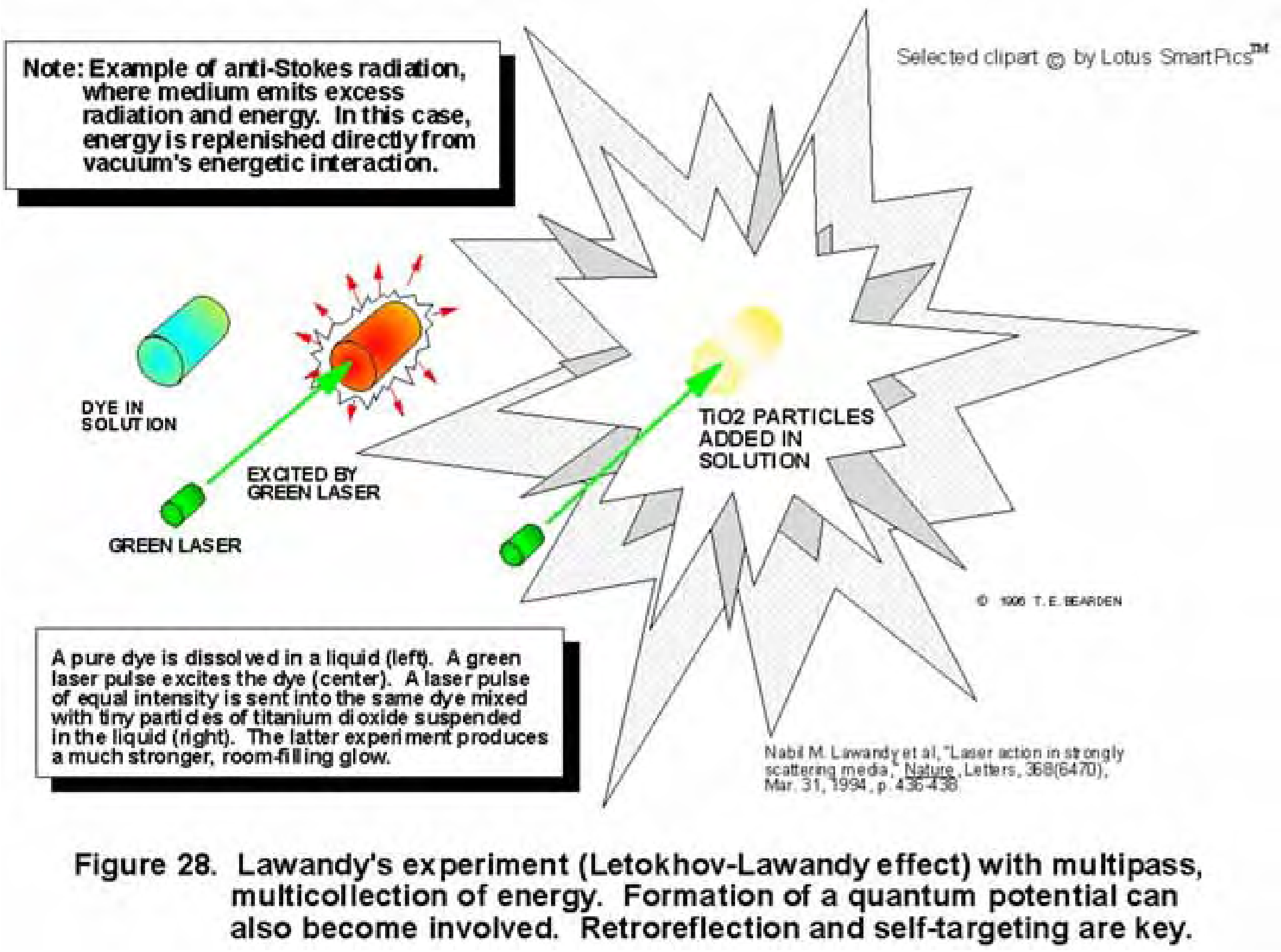
Selected clipart © by Lotus SmartPics™
DYE IN SOLUTION
EXCITED BY GREEN LASER
GREEN LASER
TiO2 PARTICLES ADDED IN SOLUTION
© 1996 T. E. BEARDEN

Note: Example of anti-Stokes radiation, where medium emits excess radiation and energy. In this case, energy is replenished directly from vacuum's energetic interaction.

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Nabil M. Lawandy et al, "Laser action in strongly scattering media," Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438.

Figure 28. Lawandy's experiment (Letokhov-Lawandy effect) with multipass, multicollection of energy. Formation of a quantum potential can also become involved. Retroreflection and self-targeting are key.

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

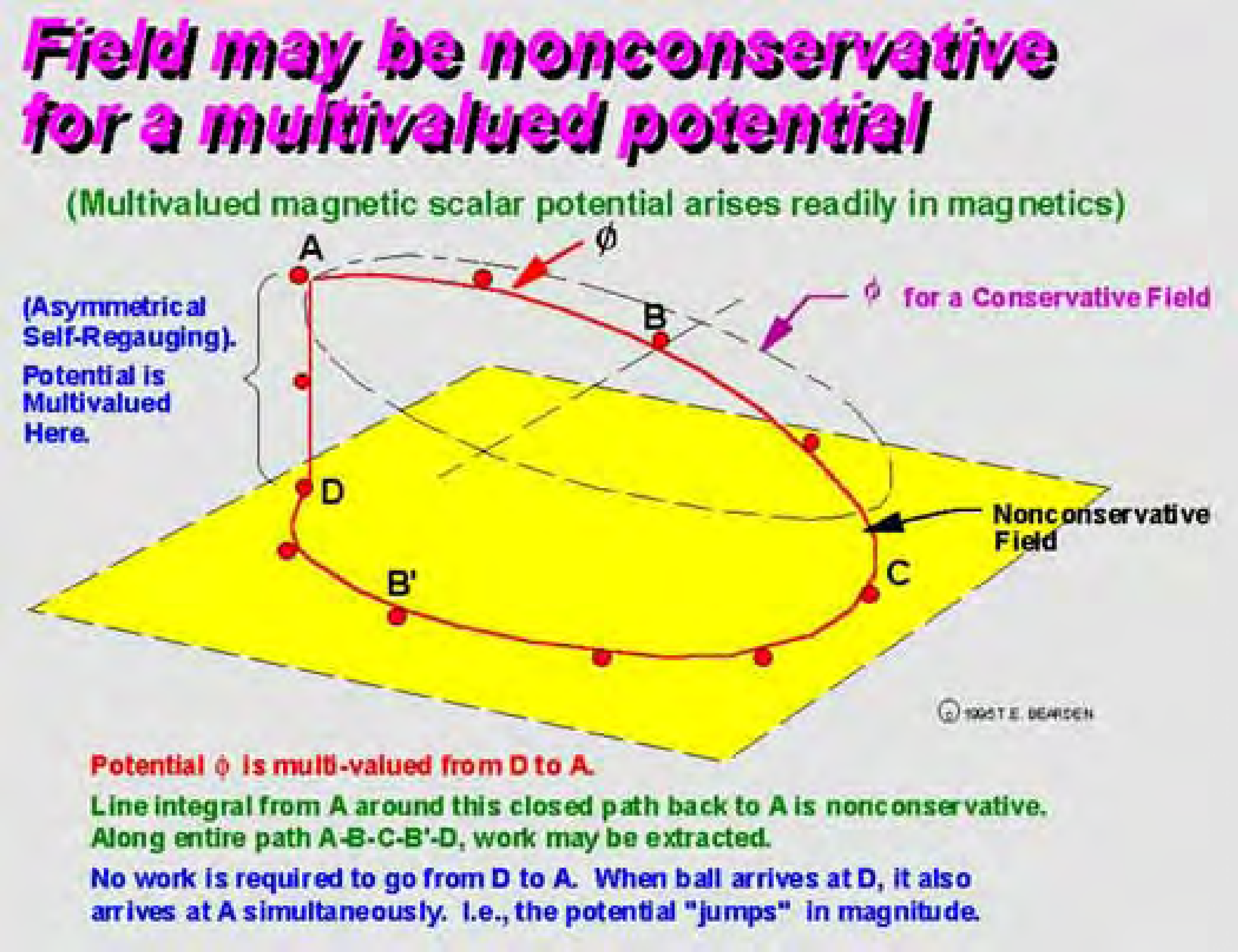

Field may be nonconservative for a multivalued potential

(Multivalued magnetic scalar potential arises readily in magnetics)

A
φ
B
φ for a Conservative Field
(Asymmetrical Self-Regauging).
Potential is Multivalued Here.
D
Nonconservative Field
C
B'
© 1995 T.E. BEARDEN

Potential φ is multi-valued from D to A.

Line integral from A around this closed path back to A is nonconservative.
Along entire path A-B-C-B'-D, work may be extracted.

No work is required to go from D to A. When ball arrives at D, it also arrives at A simultaneously. I.e., the potential "jumps" in magnitude.

Return to Slide Index

Next Slide

Previous Slide

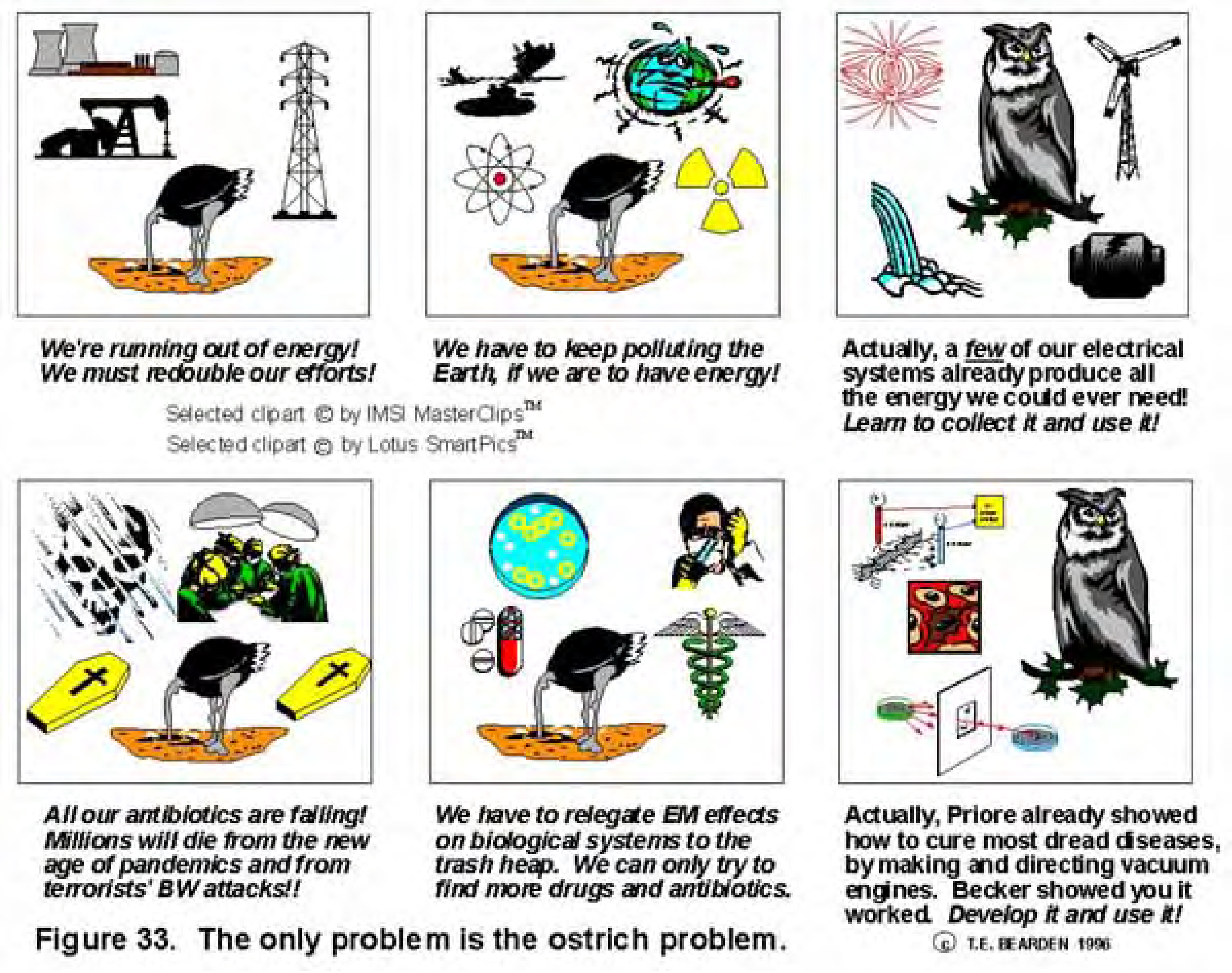

We're running out of energy! We must redouble our efforts!

We have to keep polluting the Earth, if we are to have energy!

Actually, a *few* of our electrical systems already produce all the energy we could ever need! *Learn to collect it and use it!*

Selected clipart © by IMSI MasterClips™
Selected clipart © by Lotus SmartPics™

All our antibiotics are failing! Millions will die from the new age of pandemics and from terrorists' BW attacks!!

We have to relegate EM effects on biological systems to the trash heap. We can only try to find more drugs and antibiotics.

Actually, Priore already showed how to cure most dread diseases, by making and directing vacuum engines. Becker showed you it worked. *Develop it and use it!*

Figure 33. The only problem is the ostrich problem.

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

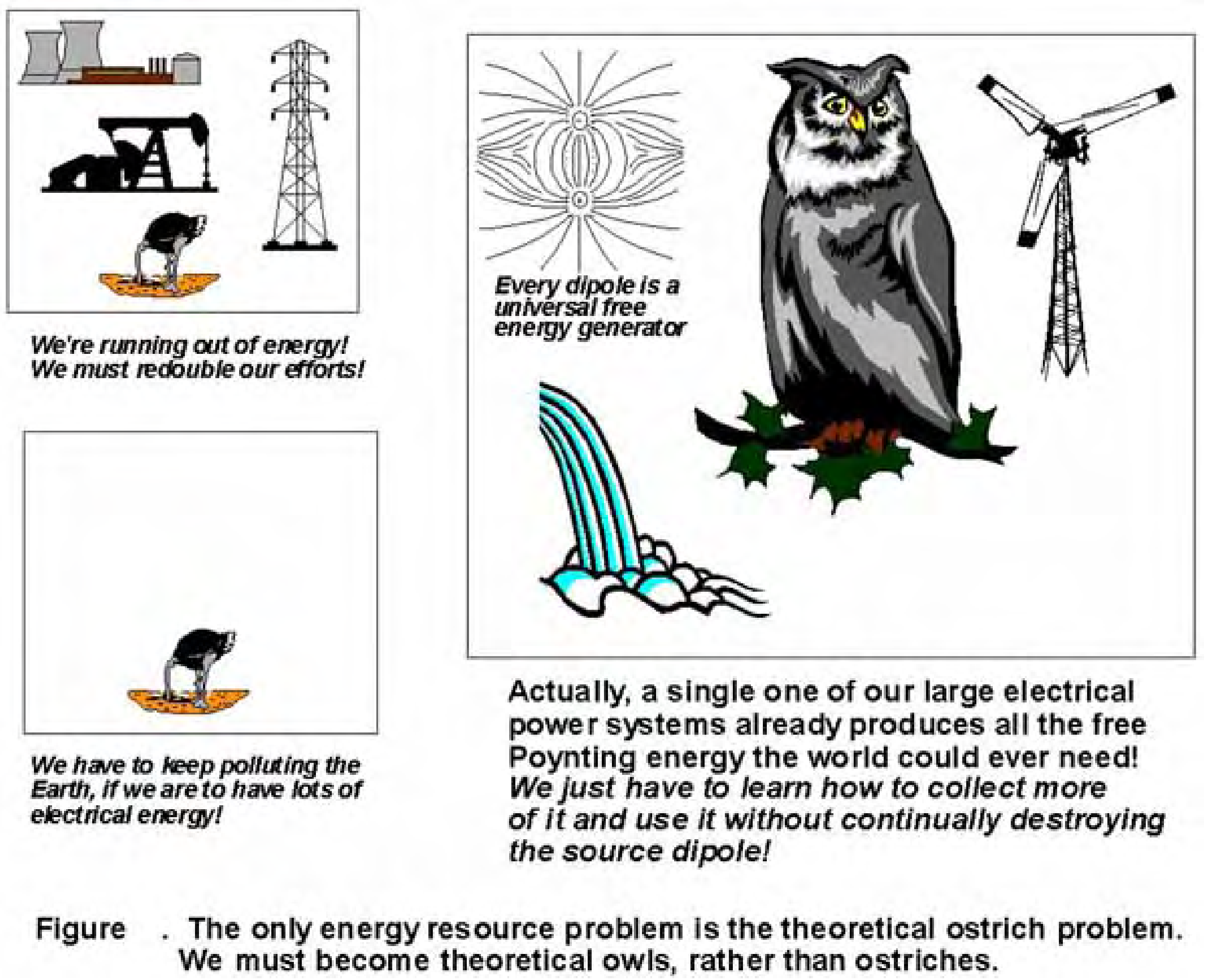

We're running out of energy!
We must redouble our efforts!

We have to keep polluting the Earth, if we are to have lots of electrical energy!

Every dipole is a universal free energy generator

Actually, a single one of our large electrical power systems already produces all the free Poynting energy the world could ever need! *We just have to learn how to collect more of it and use it without continually destroying the source dipole!*

Figure . The only energy resource problem is the theoretical ostrich problem. We must become theoretical owls, rather than ostriches.

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

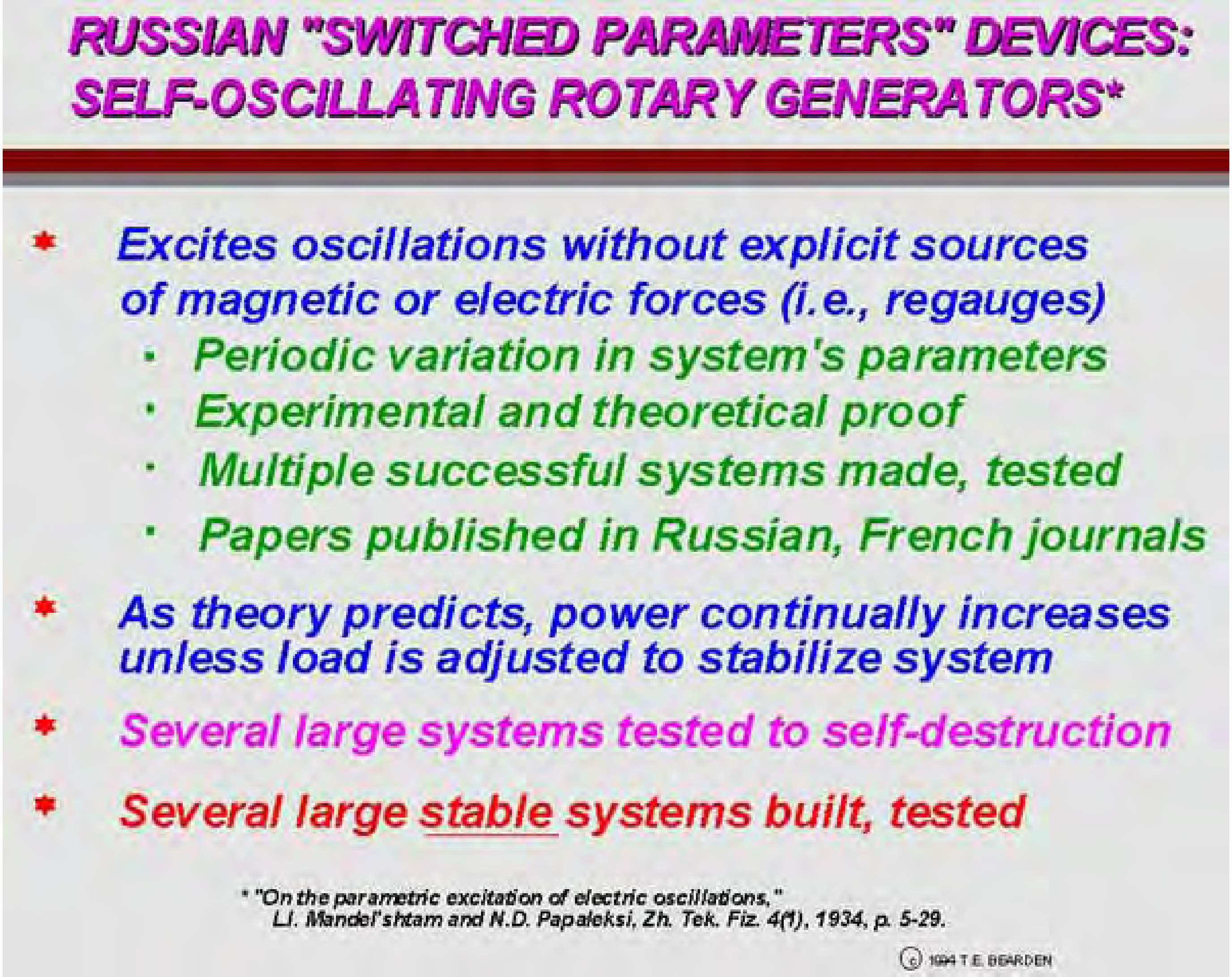

RUSSIAN "SWITCHED PARAMETERS" DEVICES: SELF-OSCILLATING ROTARY GENERATORS*

- Excites oscillations without explicit sources of magnetic or electric forces (i.e., regauges)
  - Periodic variation in system's parameters
  - Experimental and theoretical proof
  - Multiple successful systems made, tested
  - Papers published in Russian, French journals
- As theory predicts, power continually increases unless load is adjusted to stabilize system
- Several large systems tested to self-destruction
- Several large stable systems built, tested

* "On the parametric excitation of electric oscillations," L.I. Mandel'shtam and N.D. Papaleksi, Zh. Tek. Fiz. 4(1), 1934, p. 5-29.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

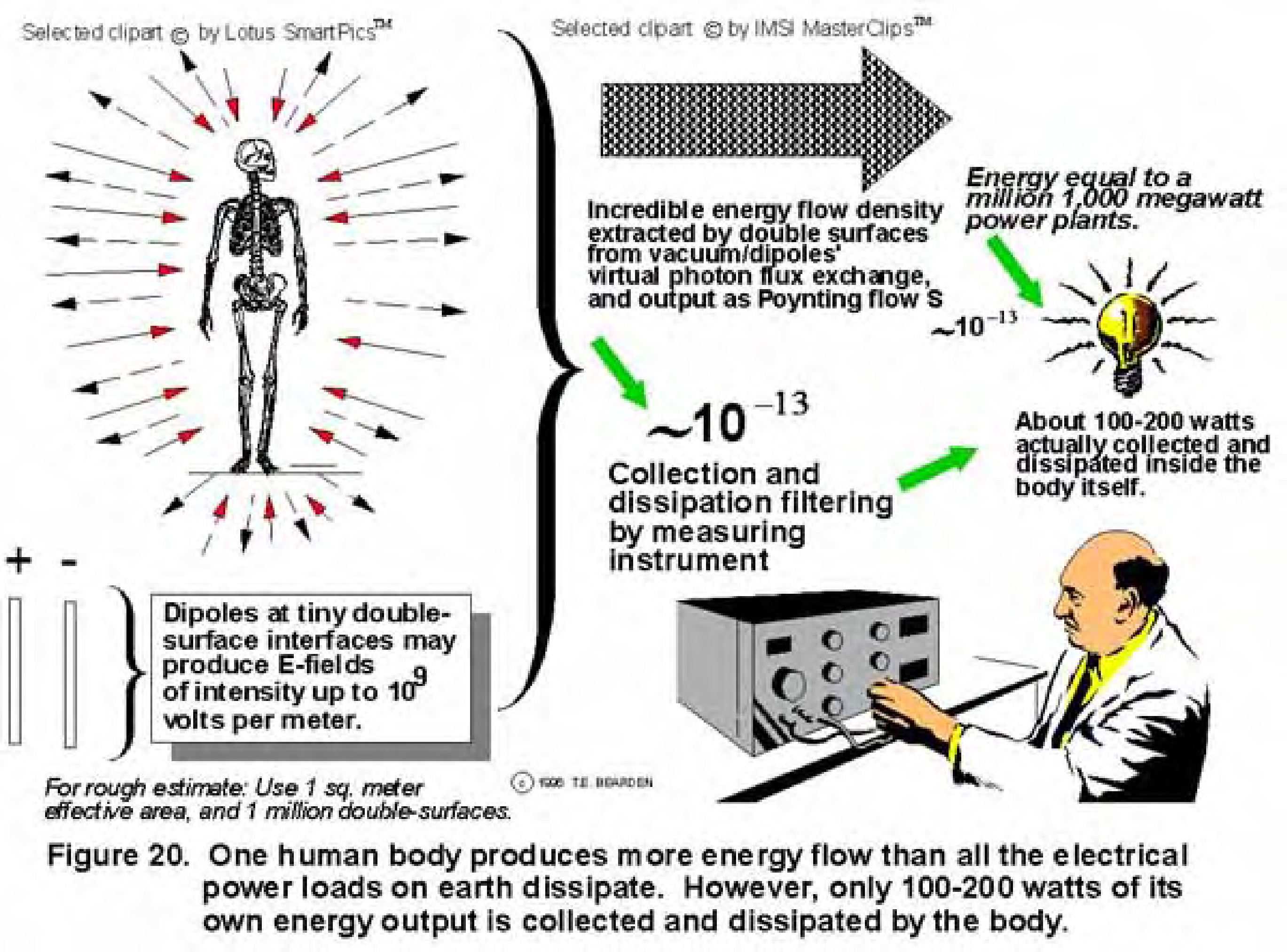
Selected clipart © by Lotus SmartPics™
Selected clipart © by IMSI MasterClips™
+
-
$\sim 10^{-13}$
$\sim 10^{-13}$
© [illegible] T.E. BEARDEN

Incredible energy flow density extracted by double surfaces from vacuum/dipoles' virtual photon flux exchange, and output as Poynting flow S

Energy equal to a million 1,000 megawatt power plants.

Collection and dissipation filtering by measuring instrument

About 100-200 watts actually collected and dissipated inside the body itself.

Dipoles at tiny double-surface interfaces may produce E-fields of intensity up to 10^9 volts per meter.

For rough estimate: Use 1 sq. meter effective area, and 1 million double-surfaces.

Figure 20. One human body produces more energy flow than all the electrical power loads on earth dissipate. However, only 100-200 watts of its own energy output is collected and dissipated by the body.

Return to Slide Index

Next Slide

Previous Slide

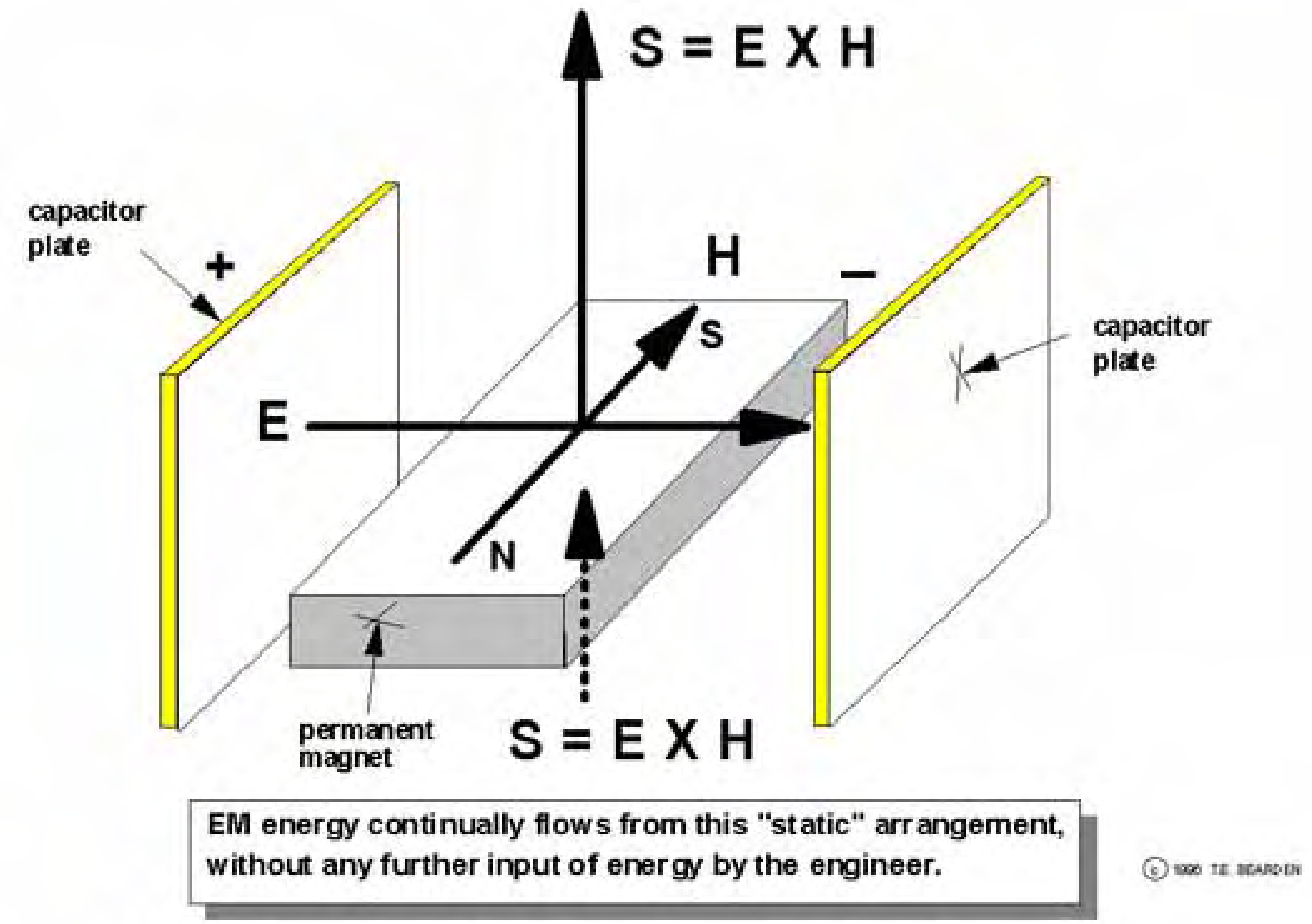
S = E X H
capacitor plate
+
H
–
S
capacitor plate
E
N
permanent magnet
S = E X H
© 1996 T.E. BEARDEN

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

Figure 18. A "static" Poynting S-flow free-energy generator.

[Return to Slide Index]

[Next Slide]

[Previous Slide]

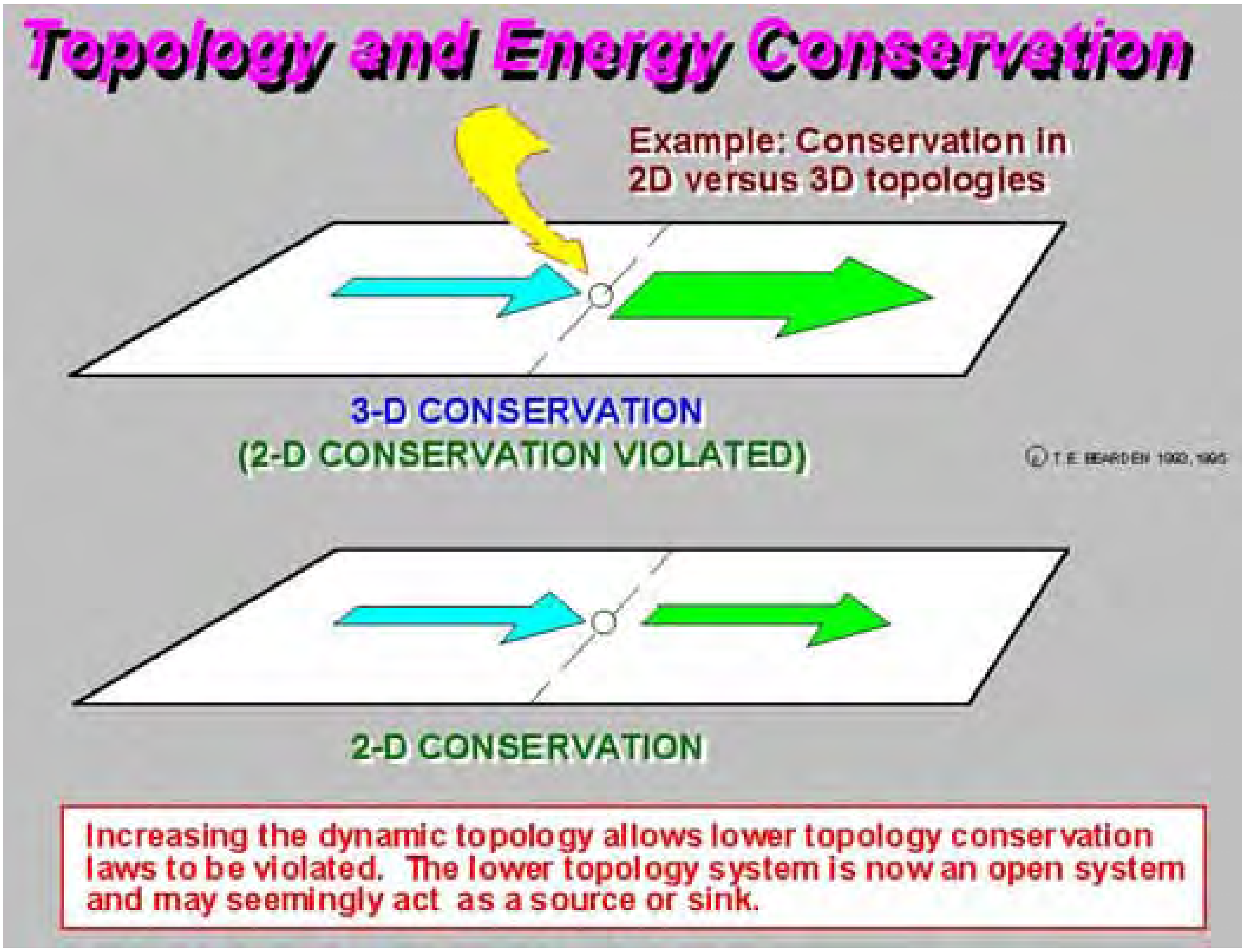

Topology and Energy Conservation

Example: Conservation in
2D versus 3D topologies

3-D CONSERVATION
(2-D CONSERVATION VIOLATED)

© T.E. BEARDEN 1992, 1995

2-D CONSERVATION

Increasing the dynamic topology allows lower topology conservation laws to be violated. The lower topology system is now an open system and may seemingly act as a source or sink.

Return to Slide Index

Next Slide

Previous Slide

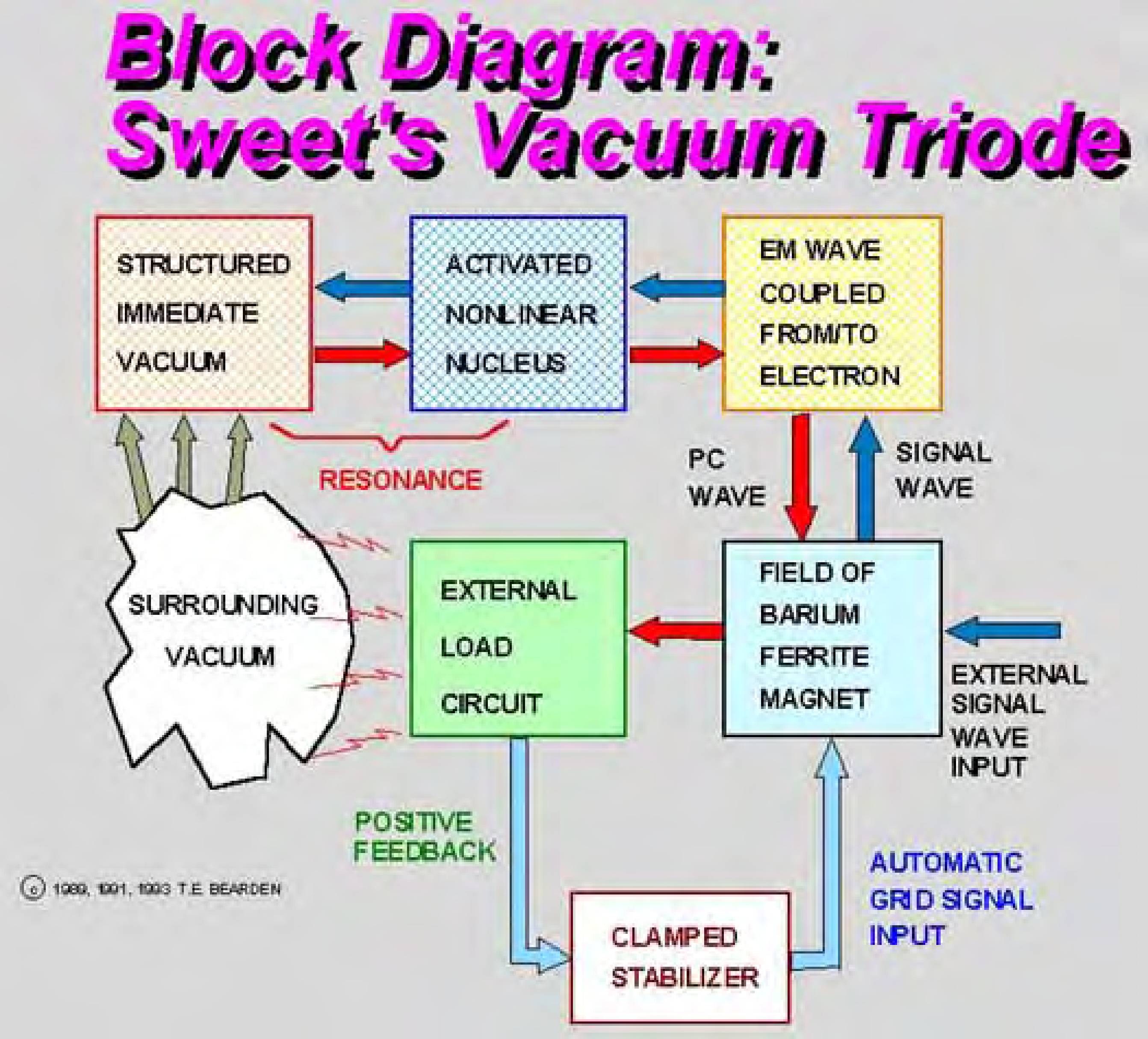

Block Diagram: Sweet's Vacuum Triode

STRUCTURED IMMEDIATE VACUUM

ACTIVATED NONLINEAR NUCLEUS

EM WAVE COUPLED FROM/TO ELECTRON

RESONANCE

PC WAVE

SIGNAL WAVE

SURROUNDING VACUUM

EXTERNAL LOAD CIRCUIT

FIELD OF BARIUM FERRITE MAGNET

EXTERNAL SIGNAL WAVE INPUT

POSITIVE FEEDBACK

AUTOMATIC GRID SIGNAL INPUT

CLAMPED STABILIZER

ⓒ 1989, 1991, 1993 T.E. BEARDEN

Return to Slide Index

Previous Slide

Defense Department Briefing on Priore Treatment to counter Biological Warfare

---

SLIDE LIST

Solutions to Mass Casualties Treatment Problem

Threat and Current Status

1. Title Page
2. Key Points in the New Technology
3. Title Page
4. Title Page
5. The Terrorist Threat
6. The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
7. Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.
8. Present Status of the WMD Threat to the United States
9. Some Present Problems in the WMD Defense Capabilities

Historical Solution: Priore Treatment Therapy

10. Title Page
11. Priore's laboratory device for treating small animals
12. Antoine Priore with the Mayor of Bordeaux
13. Raymond Pautrizel, eminent French parasitologist, who worked with Priore
14. 12-Foot plasma tube in Priore's large device
15. Priore's 3 1/2 stories high device for whole body treatment of humans
16. Mechanism of the Priore therapy
17. Priore's therapeutic methodology: using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain
18. Drastically reduced cumulative deterioration after Priore therapy

Extended Physics and Background

19. Title Page
20. Founders of Scalar Electromagnetics
21. Subdivisions of Soviet Energetics

22. Bunge on the Status of Classical and Quantum Physics
23. Einstein on Reviewing Foundations
24. Fields and Potentials have been "defined" only in and on matter
25. Maxwell's Equations are Matter-To-Matter Transforms
26. Electrodynamicists' Reaction to Removal of the Material Ether
27. The Missing Infolded Electrodynamics
28. Nonlinear Optics Distortion Correction Theorem
29. Infolded longitudinal EM biwave composition of a scalar potential
30. The Zero-Vector Axiom Destroys "Topologies Within Topologies"
31. Stoney-Whittaker Bidirectional Longitudinal EM Wave-Pair
32. Undistorted Progressive Waves (UPWs)
33. Velocity modulating a longitudinal EM wave oscillates rate of flow of time
34. The Graviton is a Coupled Photon/Antiphoton Pair
35. Mechanism for Physical Change and the Flow of Time
36. Any EM Field or Wave Pattern is Produced by interference of Two Scalar Potential Functions (E.T. Whittaker, 1904)
37. Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass
38. Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection
39. The Immune System
40. Popp's master cellular control system
41. Cancer: characteristics
42. The evolutionary path of multicellular organisms
43. Effect of fluid contamination on hemoglobin's oxygen transport
44. First step in Self-Promotion of Cancer due to Hypoxia
45. Becker's Cellular Growth Control
46. Becker's dc treatment of bone fractures
47. Becker's theoretical DC control system involved with responses to injury
48. Becker's theoretical control system governing regeneration
49. Link Between Standing DC Potentials and the Nervous System
50. Becker's Findings on Cellular Injury and Regeneration
51. Corrections For the Present EM Bioeffects Model
52. R&R System - Regeneration Example
53. Kaznacheyev's Cytopathogenic Mirror Effect
54. Microwave Radiation of U.S. Embassy in Moscow

55. Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
56. Quantum potential and hidden variable theory
57. Participants in a quantum potential share a common multiply-connected spacetime (MCST)
58. Gulf War Syndrome and Typical Comparative Disease Curves
59. Force Fields and Symmetry
60. Aspects of Strong Local Asymmetry
61. Other Aspects of Strong Local Asymmetry
62. Small Entities Emit Powerful S-Flows
63. What is the "magnitude" of a potential?
64. Table A1. The Vacuum Energetically Interacts With Every Particle, Continually
65. Extended Superpotential Theory (Adding Vacuum Engines)
66. Vacuum Engine: Jillions of Working ST Demons
67. Active Spacetime and a Specific Vacuum Engine (A General Relativistic Concept and Analogy)
68. Extension of Wheeler's general relativity principle
69. Signal versus Vacuum Engine
70. Electronuclear Interactions by Vacuum Engines
71. Vacuum Engineering
72. Impact of Vacuum Engineering
73. A Dramatic Extension to Nonlinear Optical Pumping
74. Time Conservation and Some Properties (Energy and time are canonical)
75. A Dramatic Extension to Nonlinear Optical Pumping
76. Key Points in the New Technology

Modern WMD Solution: Portable Treatment Systems

77. Title Page
78. Portable Priore-Type Treatment System
79. Portable System Block Diagram
80. Portable Priore Technology Unit
81. Saving the Stricken Civilians

Conclusions: Technology Practical, Proven

82. Title Page
83. Summary and Conclusions

Recommendation: Crash Development Program

84. Title Page
85. Recommend implementing:
86. Final Thoughts

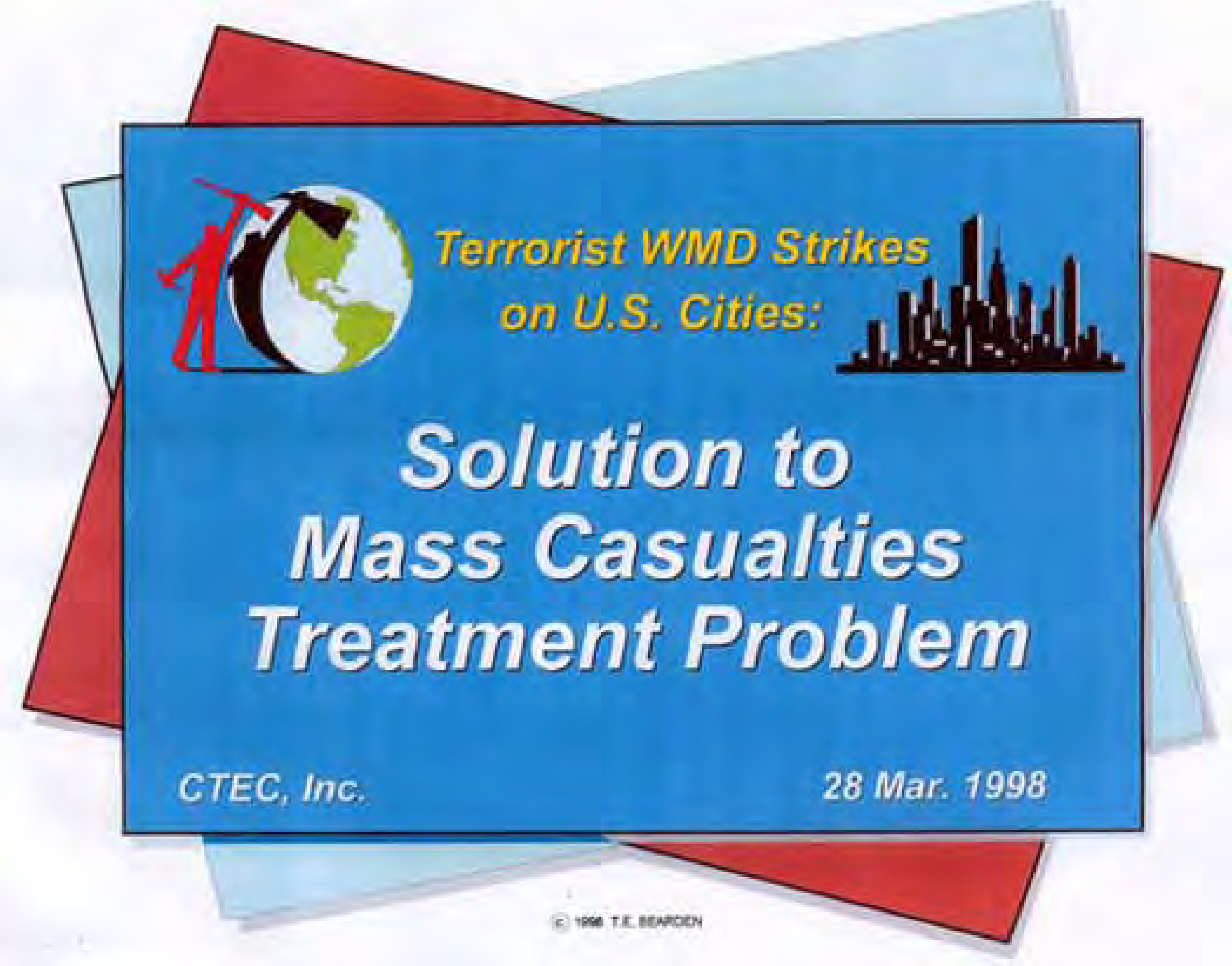

Terrorist WMD Strikes
on U.S. Cities:

Solution to
Mass Casualties
Treatment Problem

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

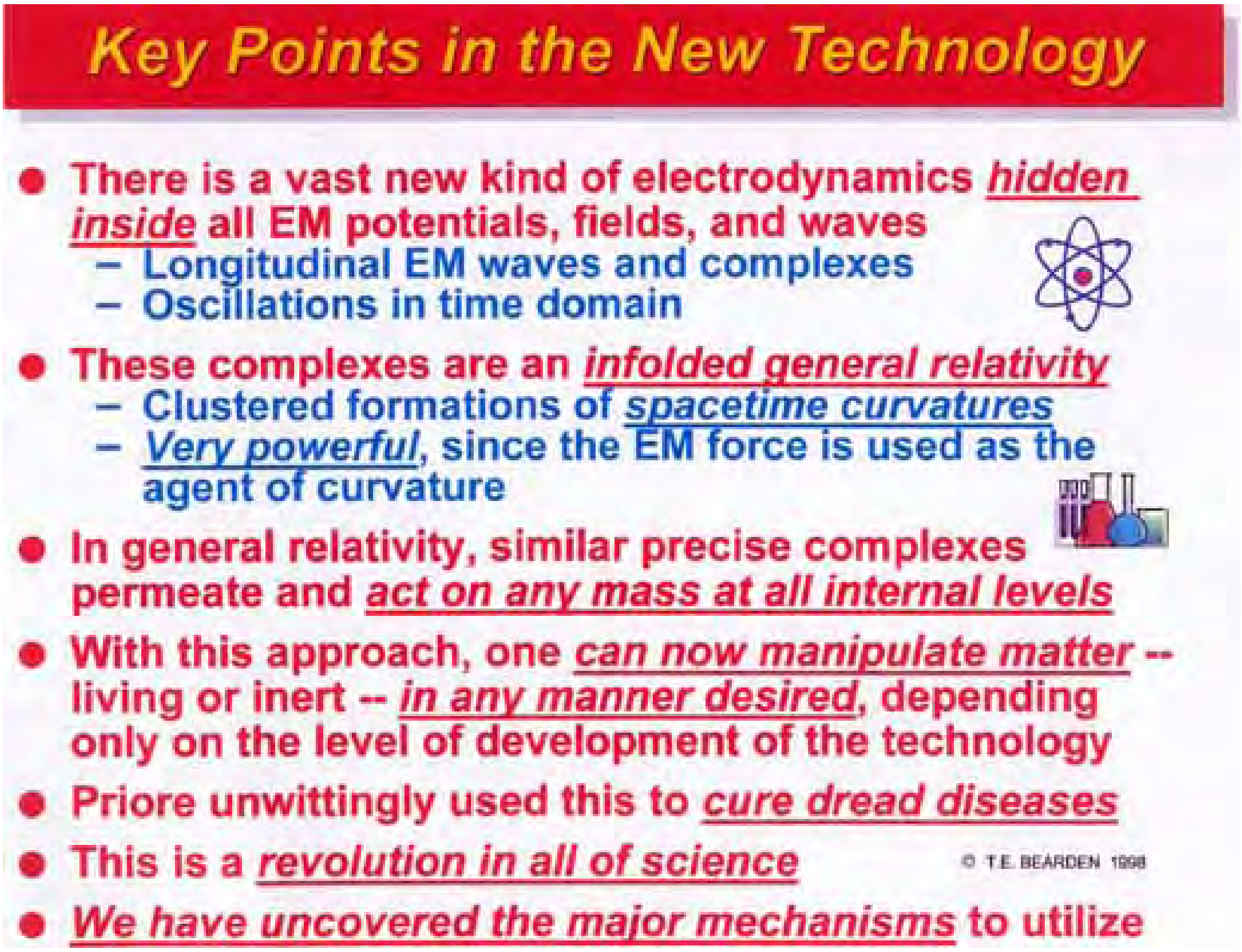

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*

© T.E. BEARDEN 1998

- *We have uncovered the major mechanisms* to utilize

Return to Slide Index

Next Slide

Previous Slide

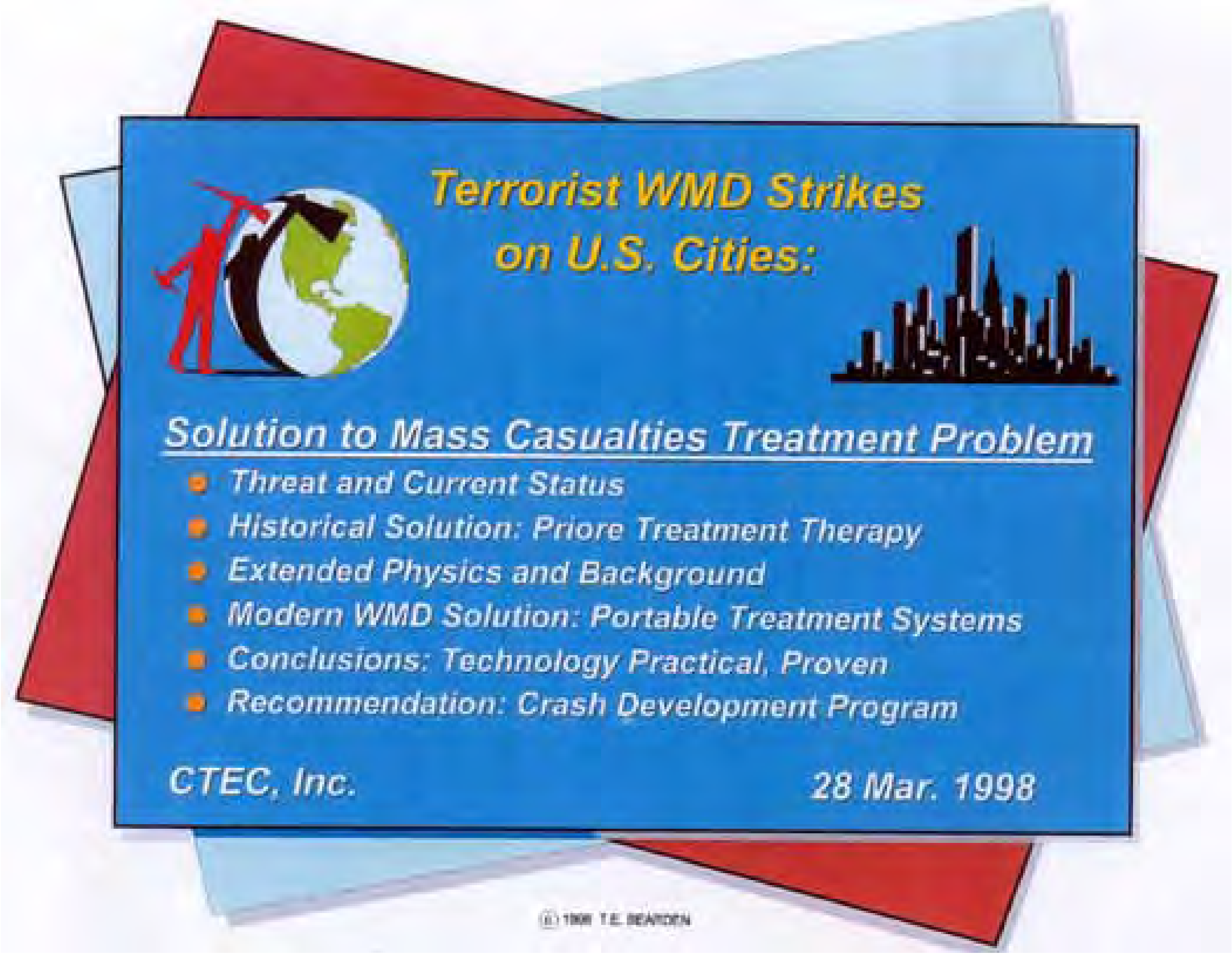

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

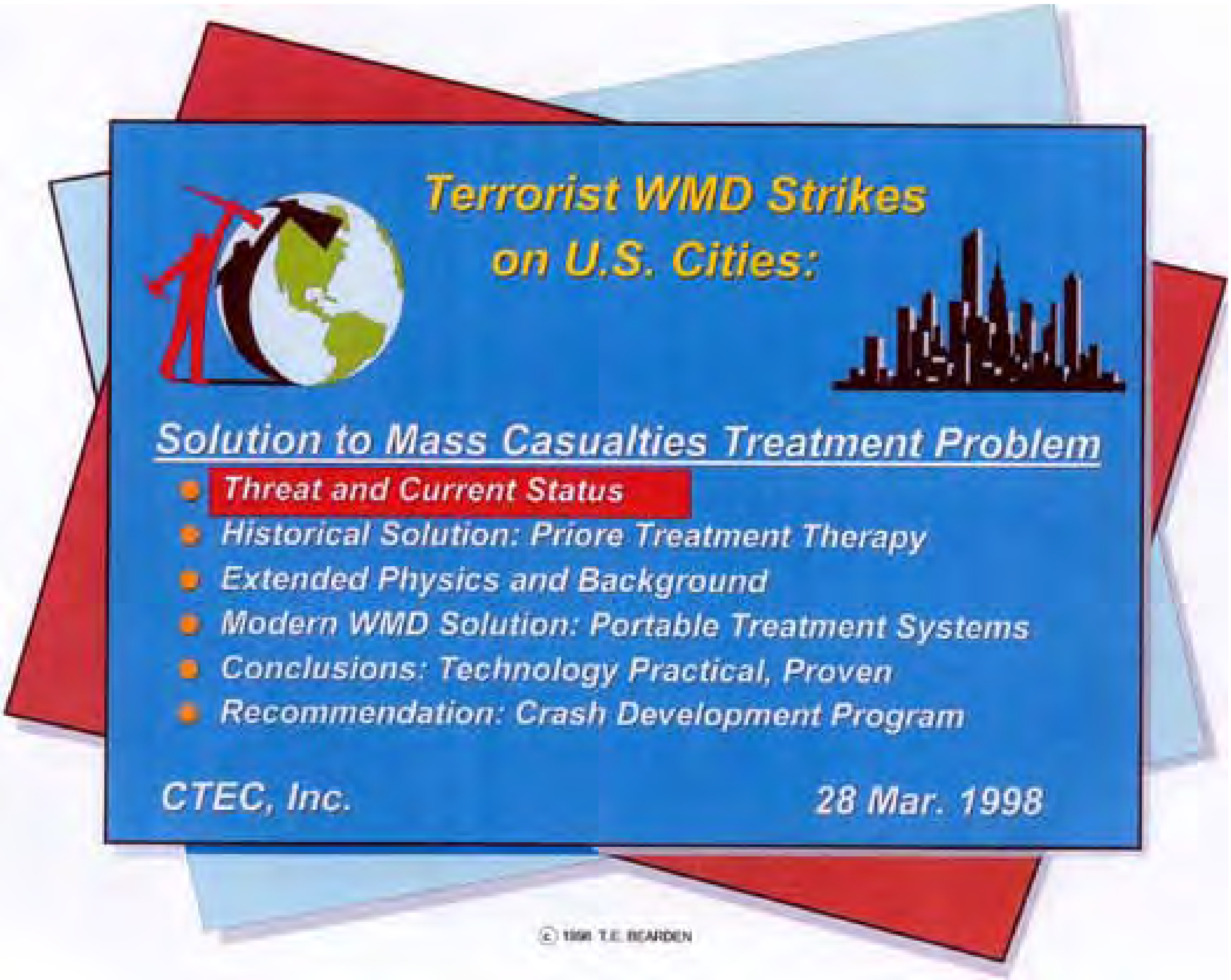

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

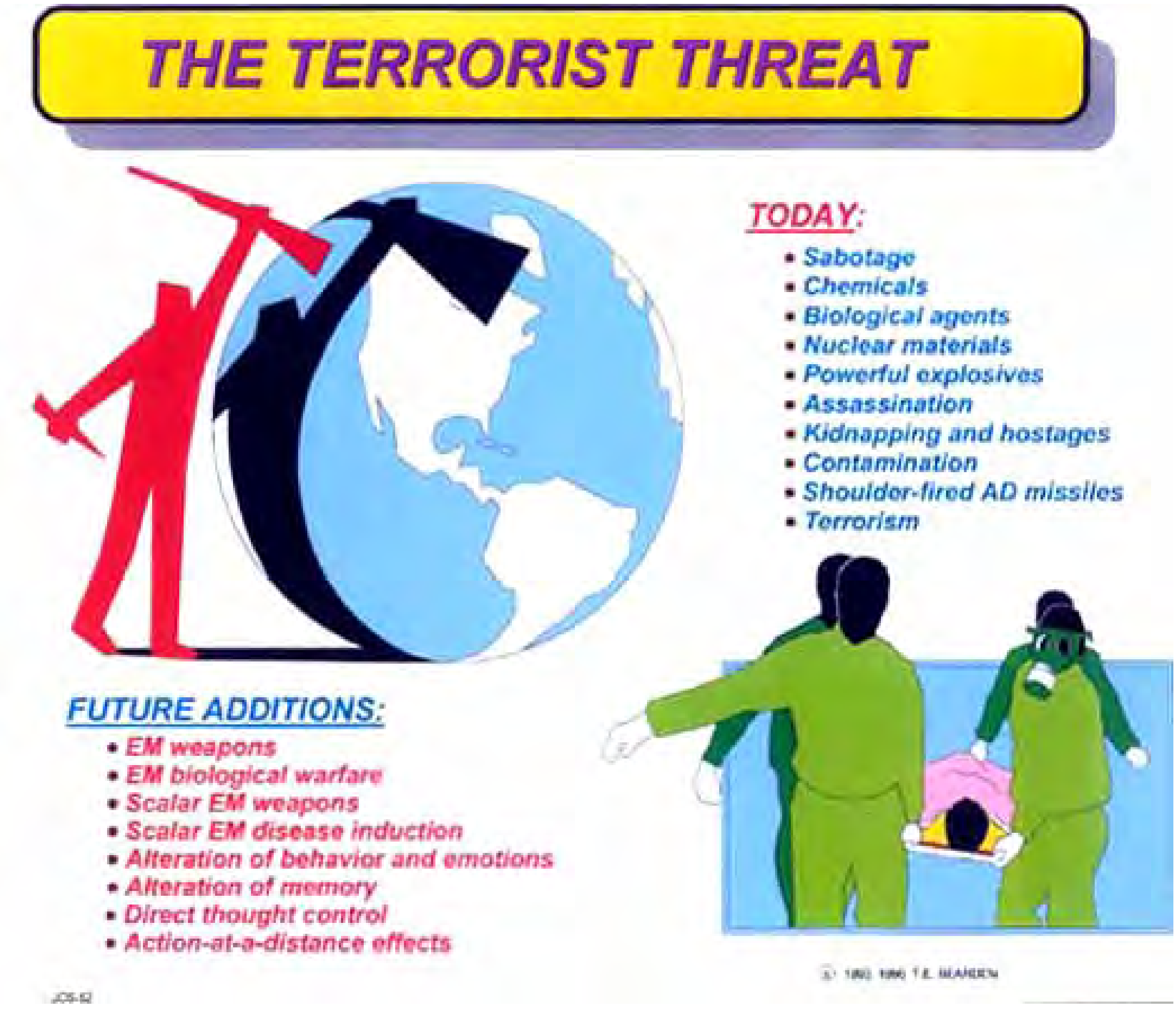

THE TERRORIST THREAT

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- EM weapons
- EM biological warfare
- Scalar EM weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct thought control
- Action-at-a-distance effects

© 1992, 1996 T.E. BEARDEN

JCS-62

Return to Slide Index

Next Slide

Previous Slide

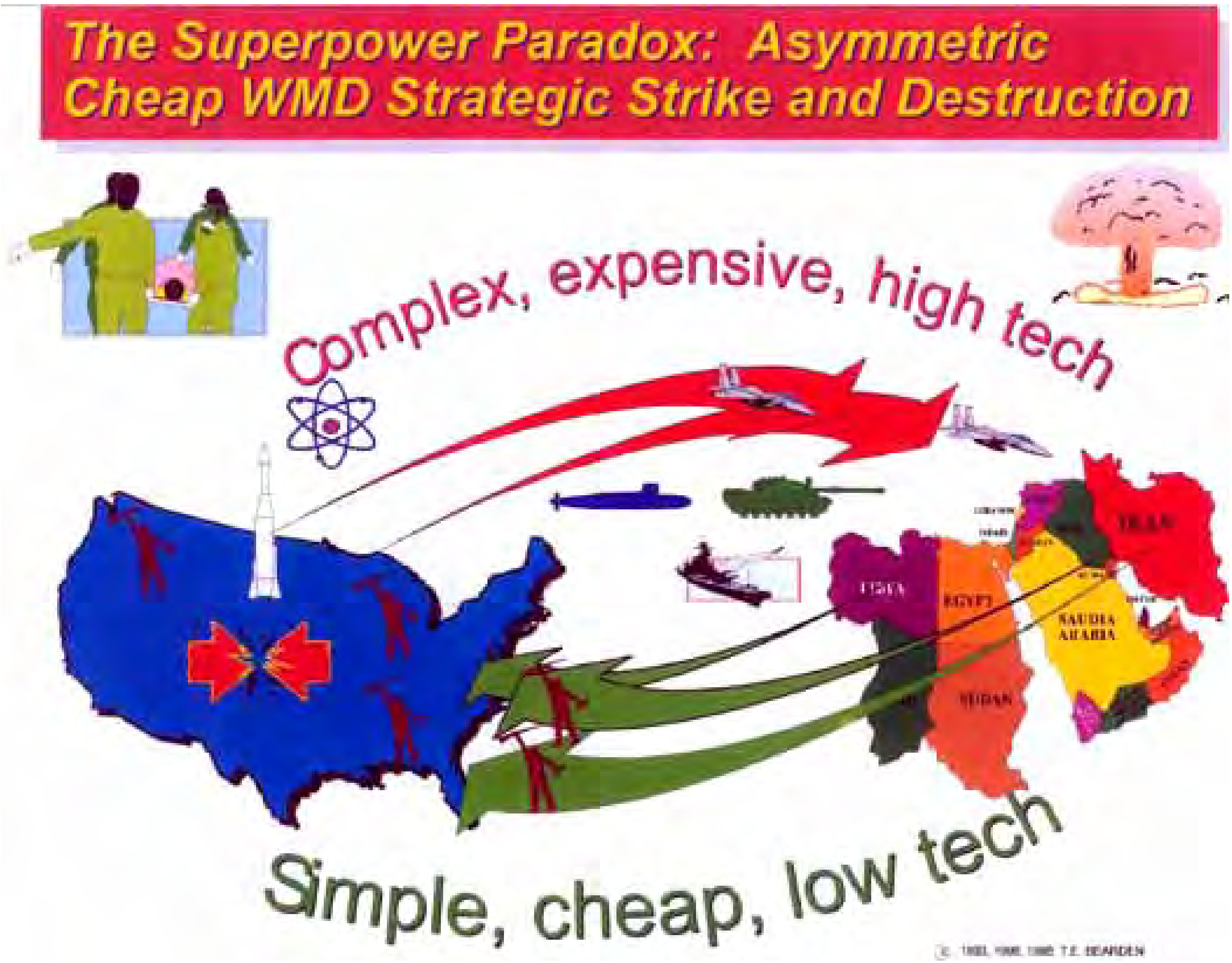

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

© 1993, 1996, 1998 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

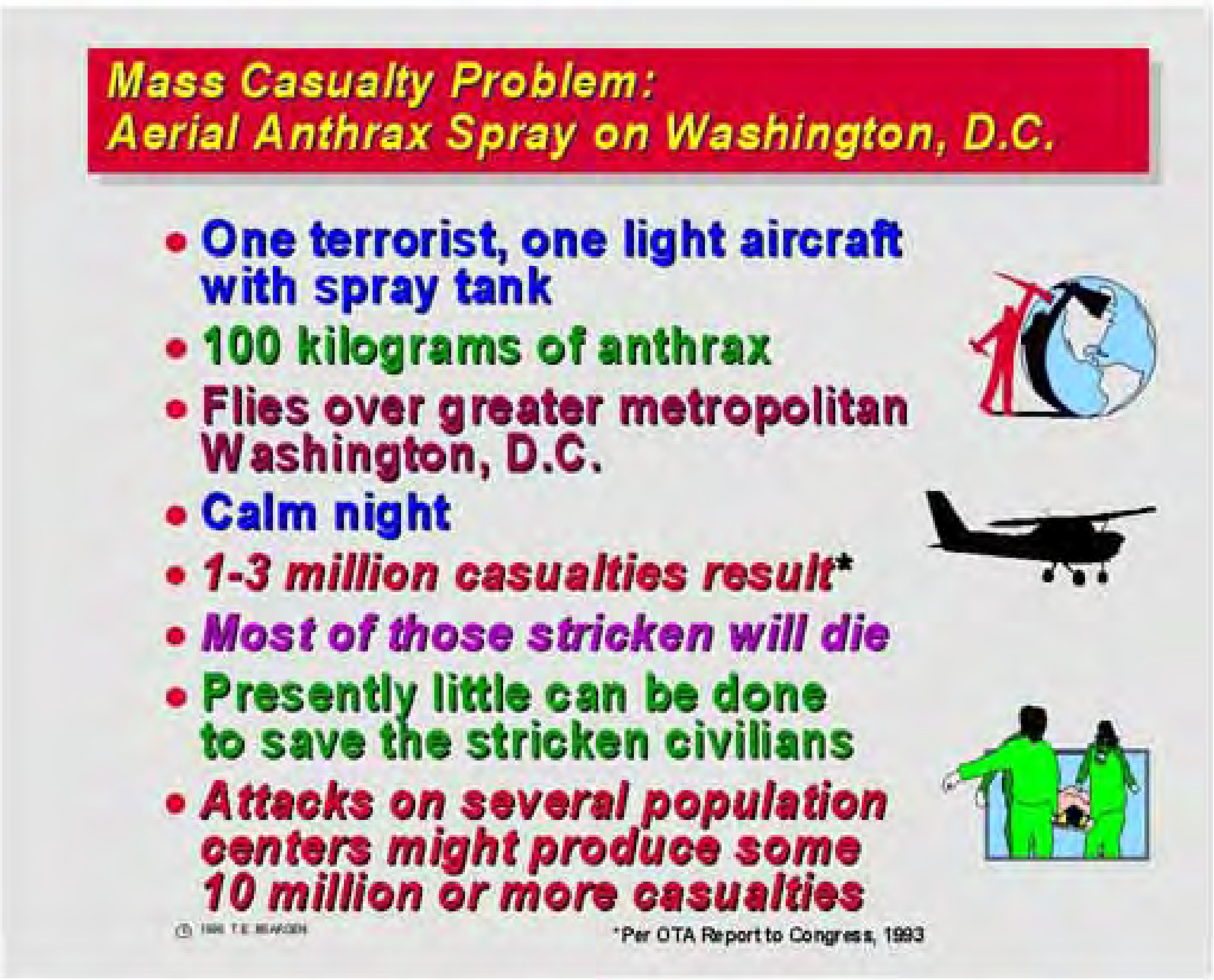

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- 100 kilograms of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© 1996 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

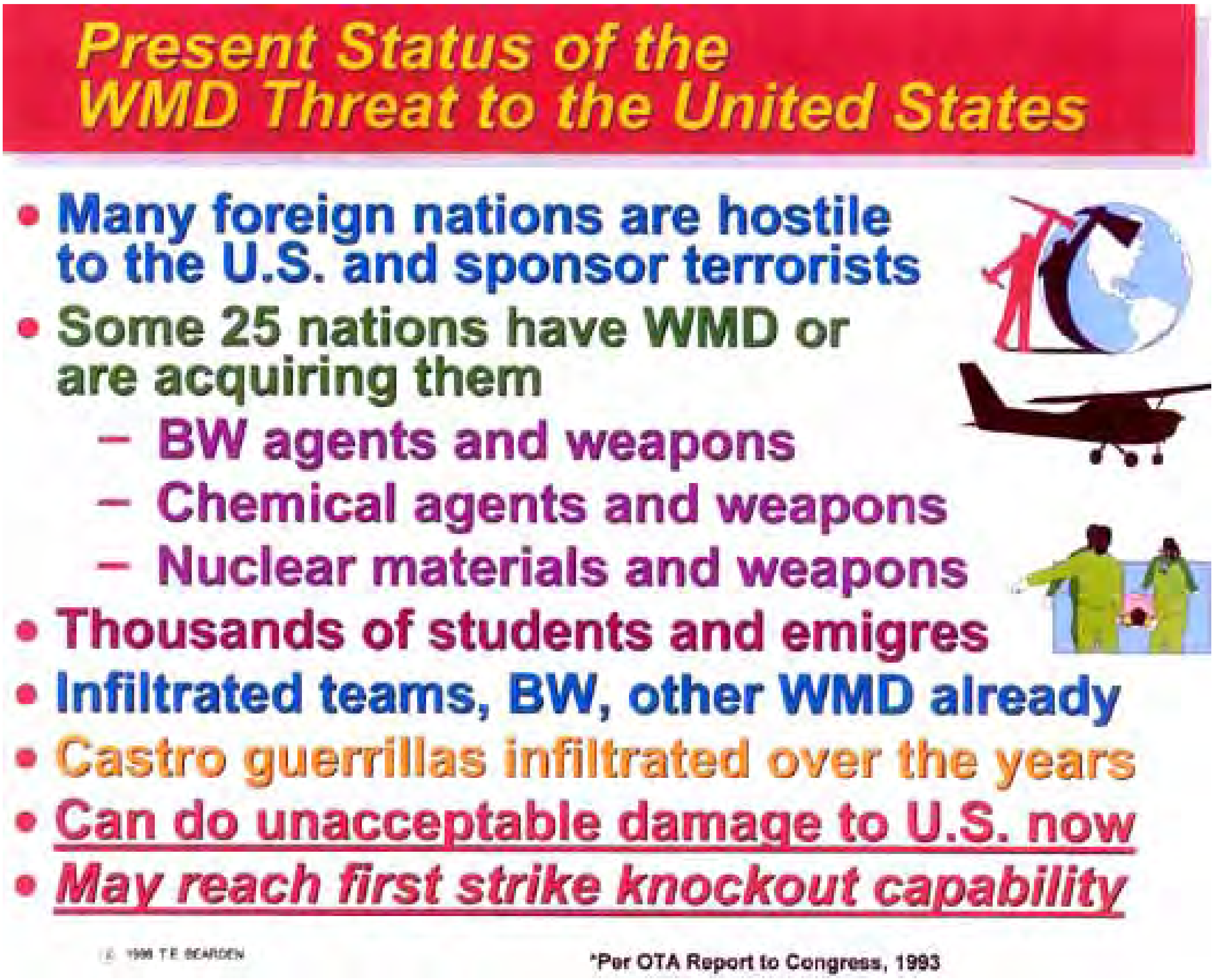

Present Status of the WMD Threat to the United States

- Many foreign nations are hostile to the U.S. and sponsor terrorists
- Some 25 nations have WMD or are acquiring them
  - BW agents and weapons
  - Chemical agents and weapons
  - Nuclear materials and weapons
- Thousands of students and emigres
- Infiltrated teams, BW, other WMD already
- Castro guerrillas infiltrated over the years
- Can do unacceptable damage to U.S. now
- *May reach first strike knockout capability*

© 1998 T.E. BEARDEN

*Per OTA Report to Congress, 1993

Return to Slide Index

Next Slide

Previous Slide

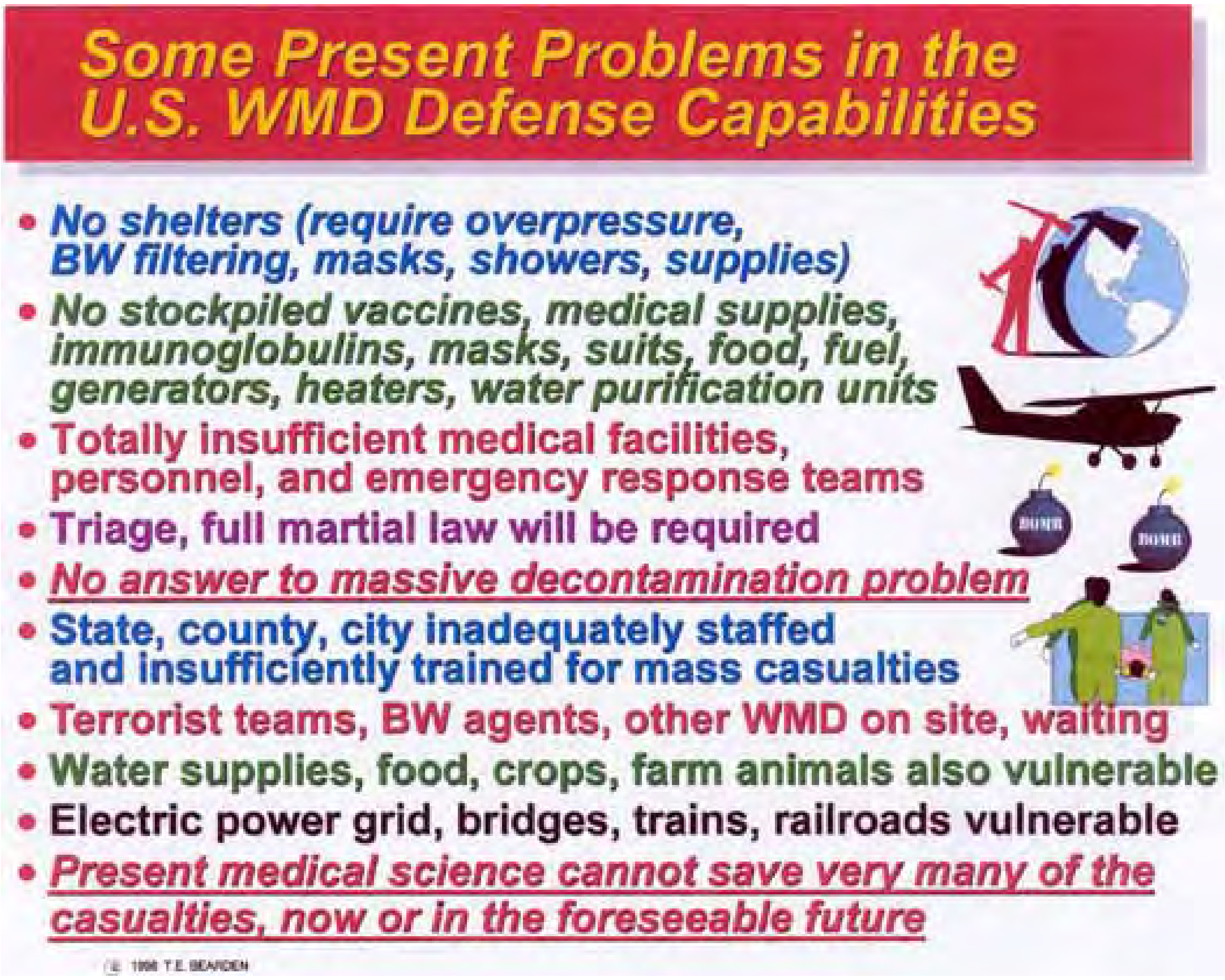

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

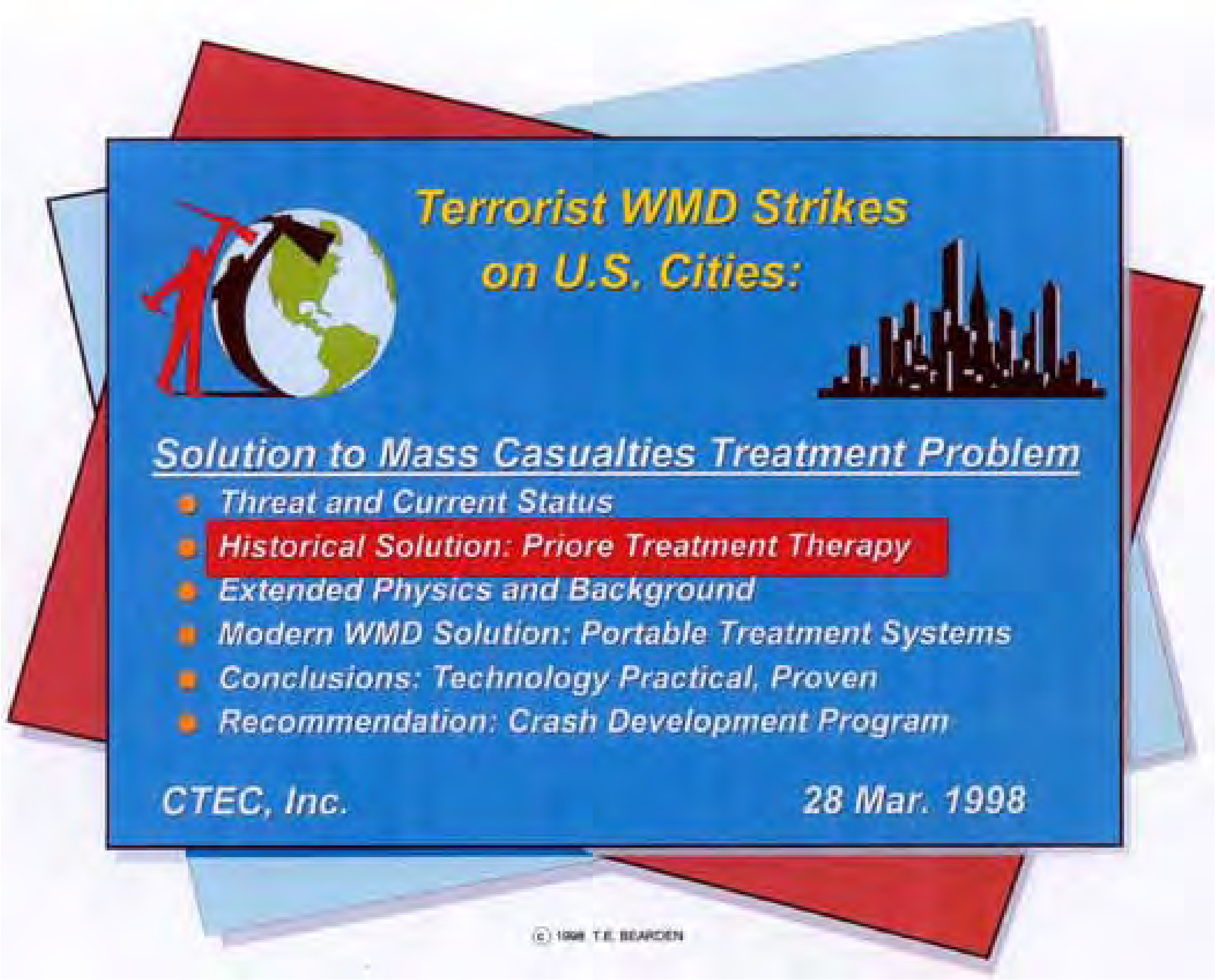

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

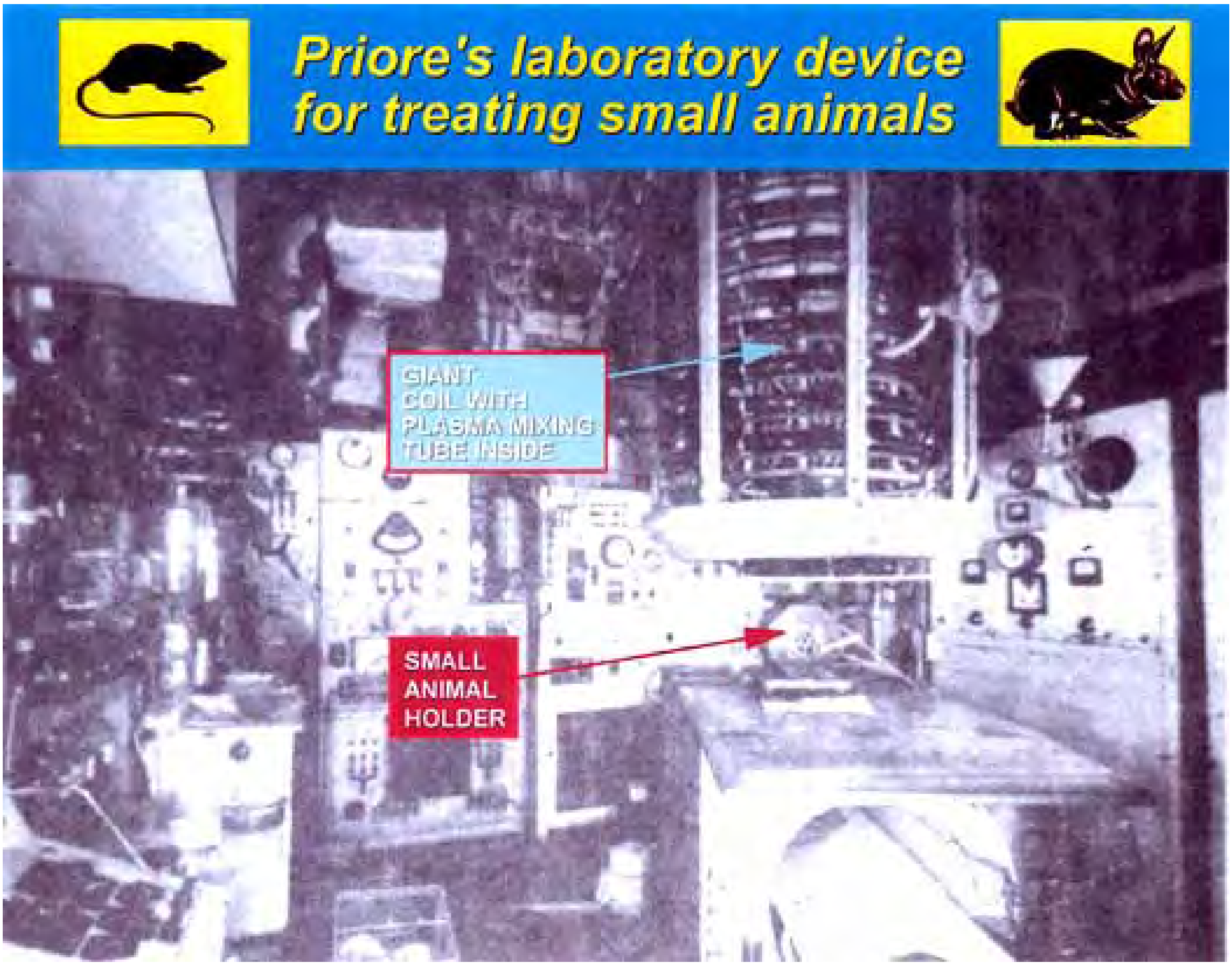

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE

SMALL ANIMAL HOLDER

Return to Slide Index

Next Slide

Previous Slide

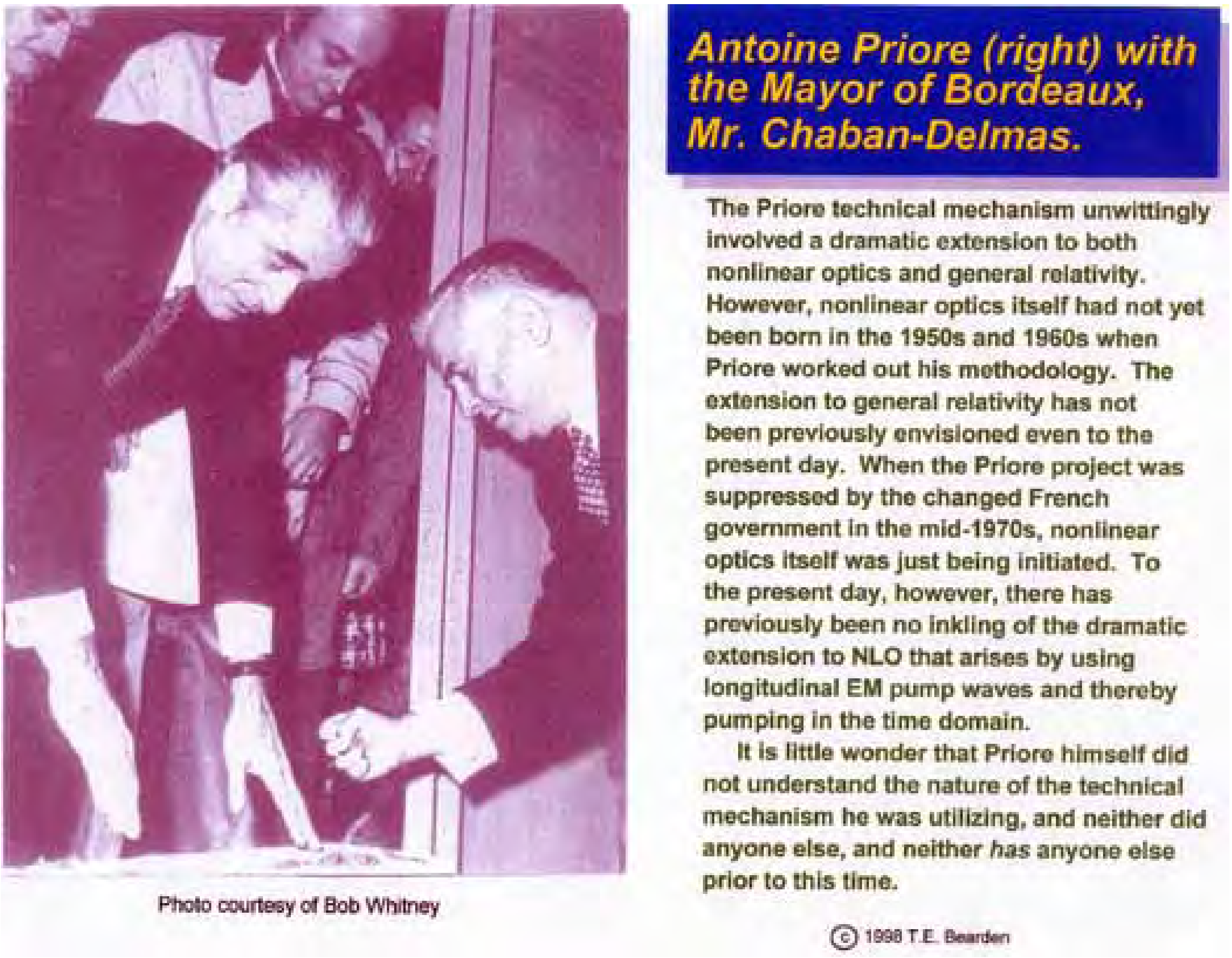

Antoine Priore (right) with the Mayor of Bordeaux, Mr. Chaban-Delmas.

Photo courtesy of Bob Whitney

The Priore technical mechanism unwittingly involved a dramatic extension to both nonlinear optics and general relativity. However, nonlinear optics itself had not yet been born in the 1950s and 1960s when Priore worked out his methodology. The extension to general relativity has not been previously envisioned even to the present day. When the Priore project was suppressed by the changed French government in the mid-1970s, nonlinear optics itself was just being initiated. To the present day, however, there has previously been no inkling of the dramatic extension to NLO that arises by using longitudinal EM pump waves and thereby pumping in the time domain.

It is little wonder that Priore himself did not understand the nature of the technical mechanism he was utilizing, and neither did anyone else, and neither *has* anyone else prior to this time.

© 1998 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

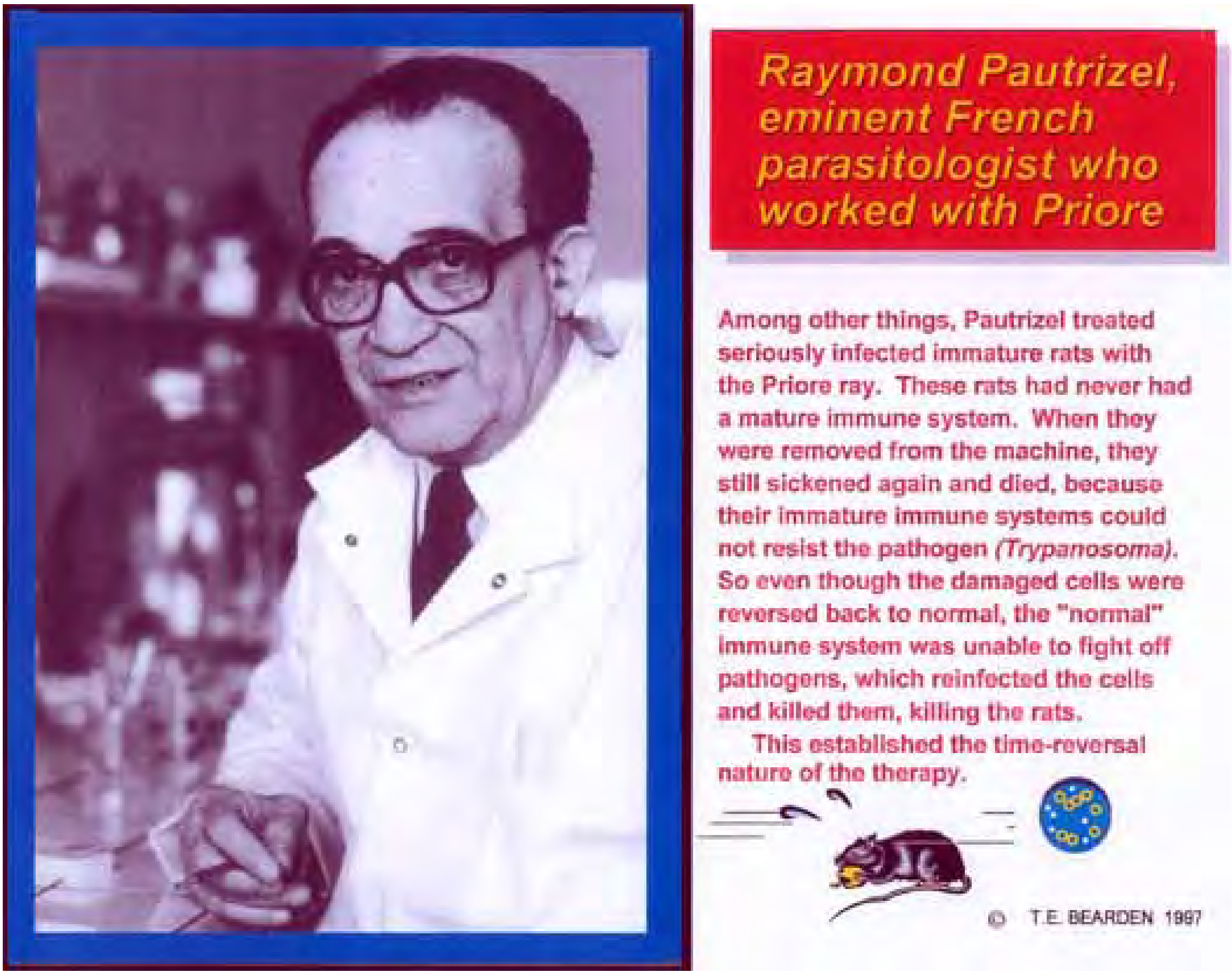

Raymond Pautrizel, eminent French parasitologist who worked with Priore

Among other things, Pautrizel treated seriously infected immature rats with the Priore ray. These rats had never had a mature immune system. When they were removed from the machine, they still sickened again and died, because their immature immune systems could not resist the pathogen *(Trypanosoma)*. So even though the damaged cells were reversed back to normal, the "normal" immune system was unable to fight off pathogens, which reinfected the cells and killed them, killing the rats.

This established the time-reversal nature of the therapy.

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

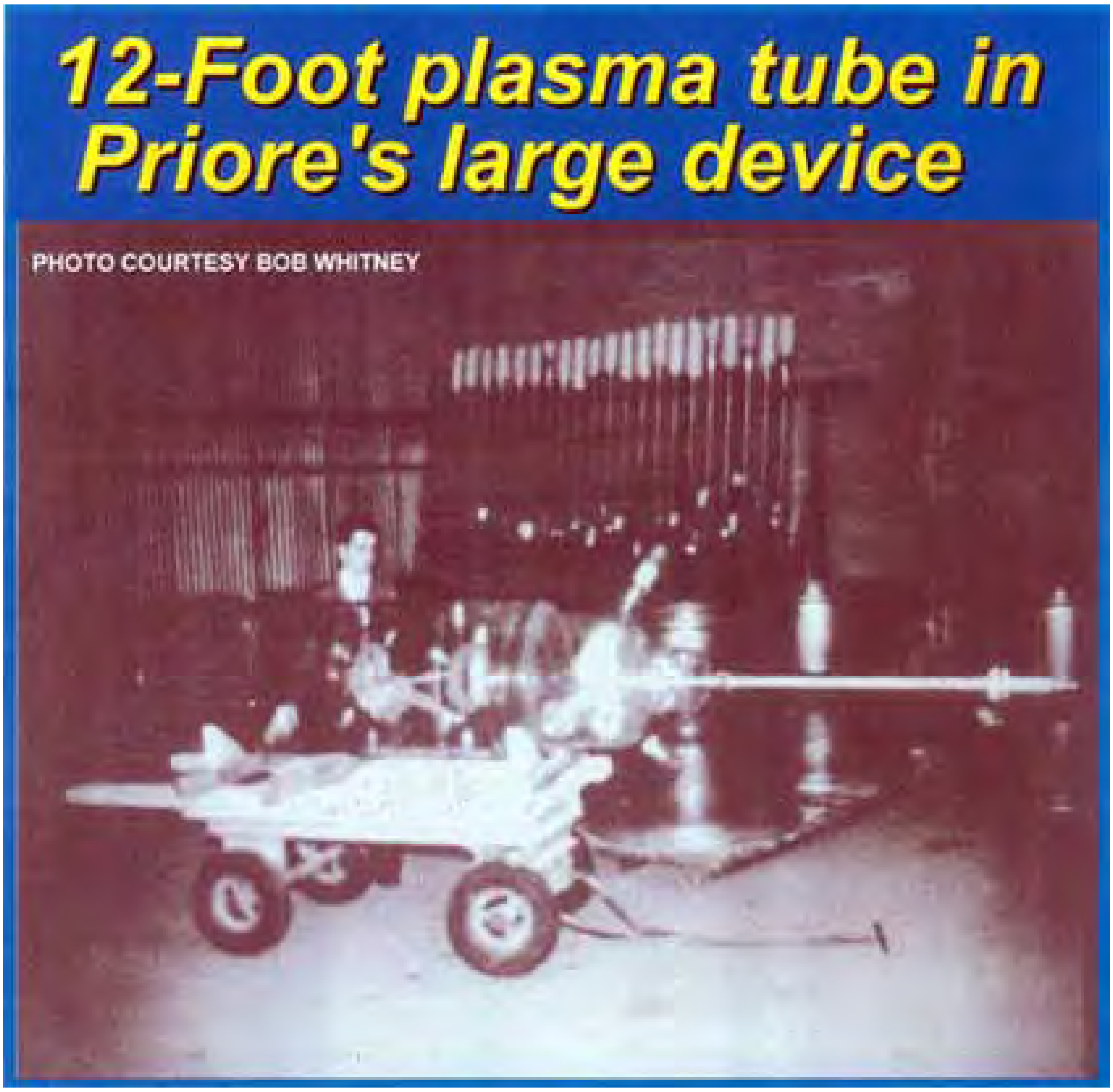

12-Foot plasma tube in Priore's large device

PHOTO COURTESY BOB WHITNEY

Return to Slide Index

Next Slide

Previous Slide

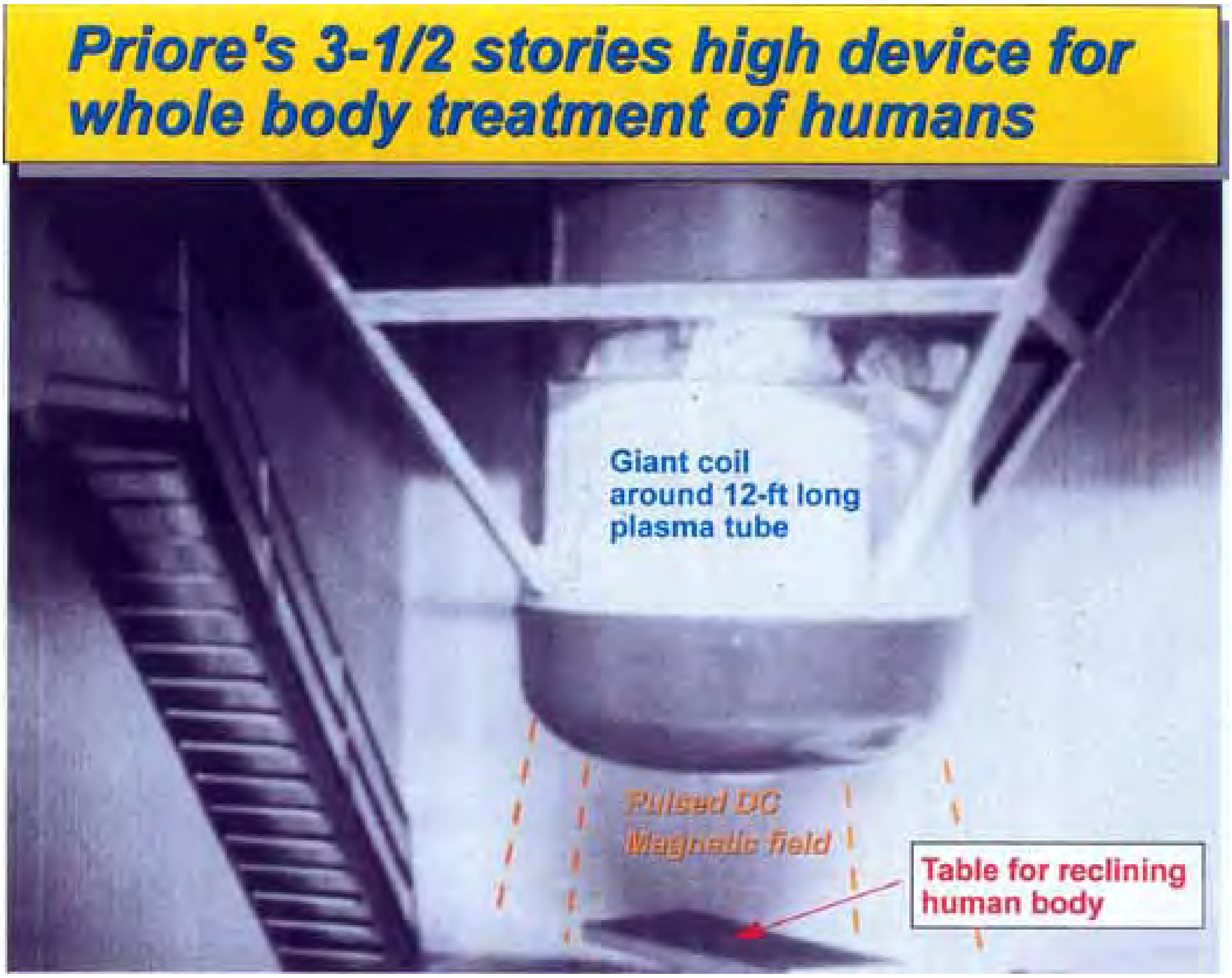

Priore's 3-1/2 stories high device for whole body treatment of humans

Giant coil
around 12-ft long
plasma tube
Pulsed DC
Magnetic field
Table for reclining
human body

Return to Slide Index

Next Slide

Previous Slide

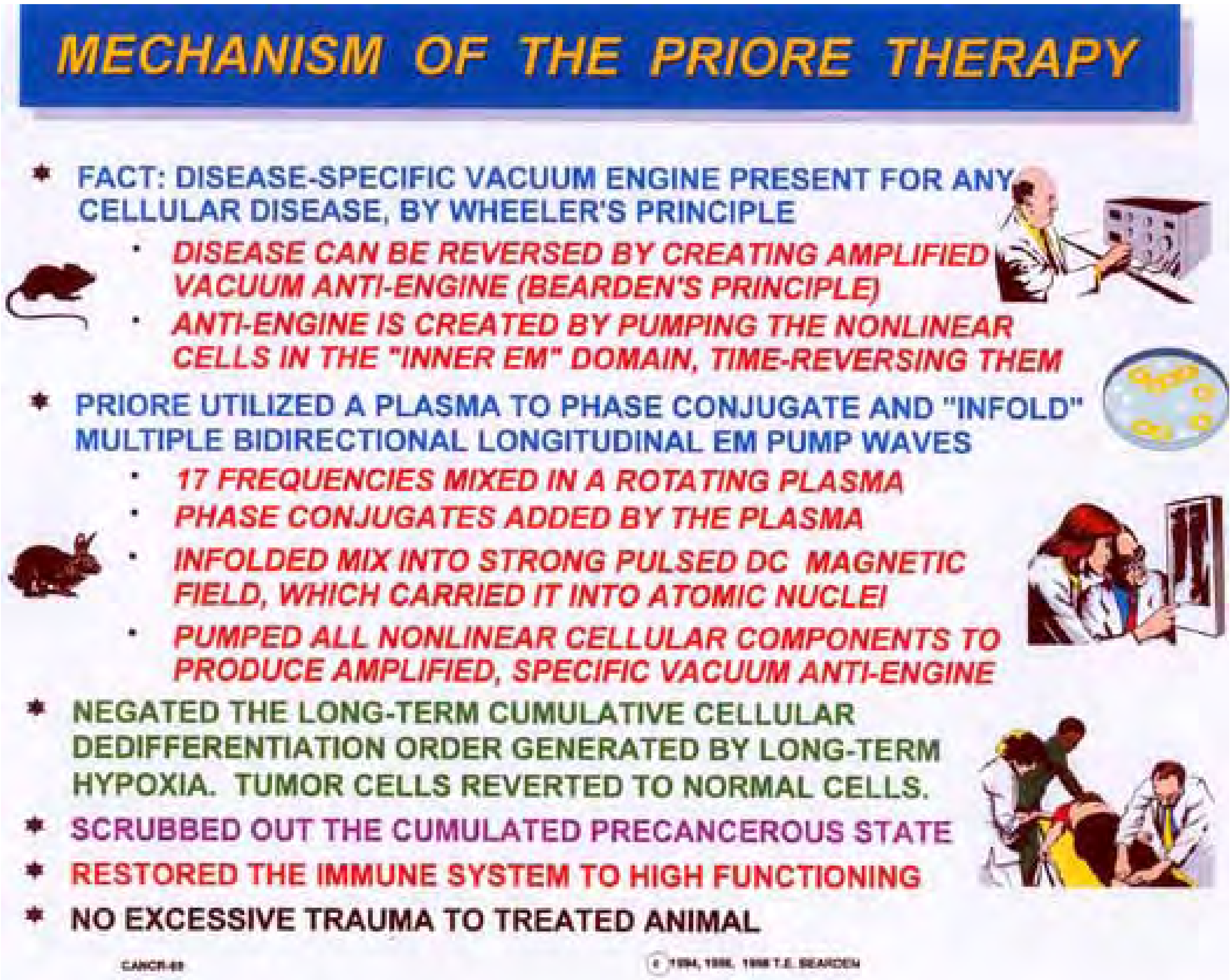

MECHANISM OF THE PRIORE THERAPY

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

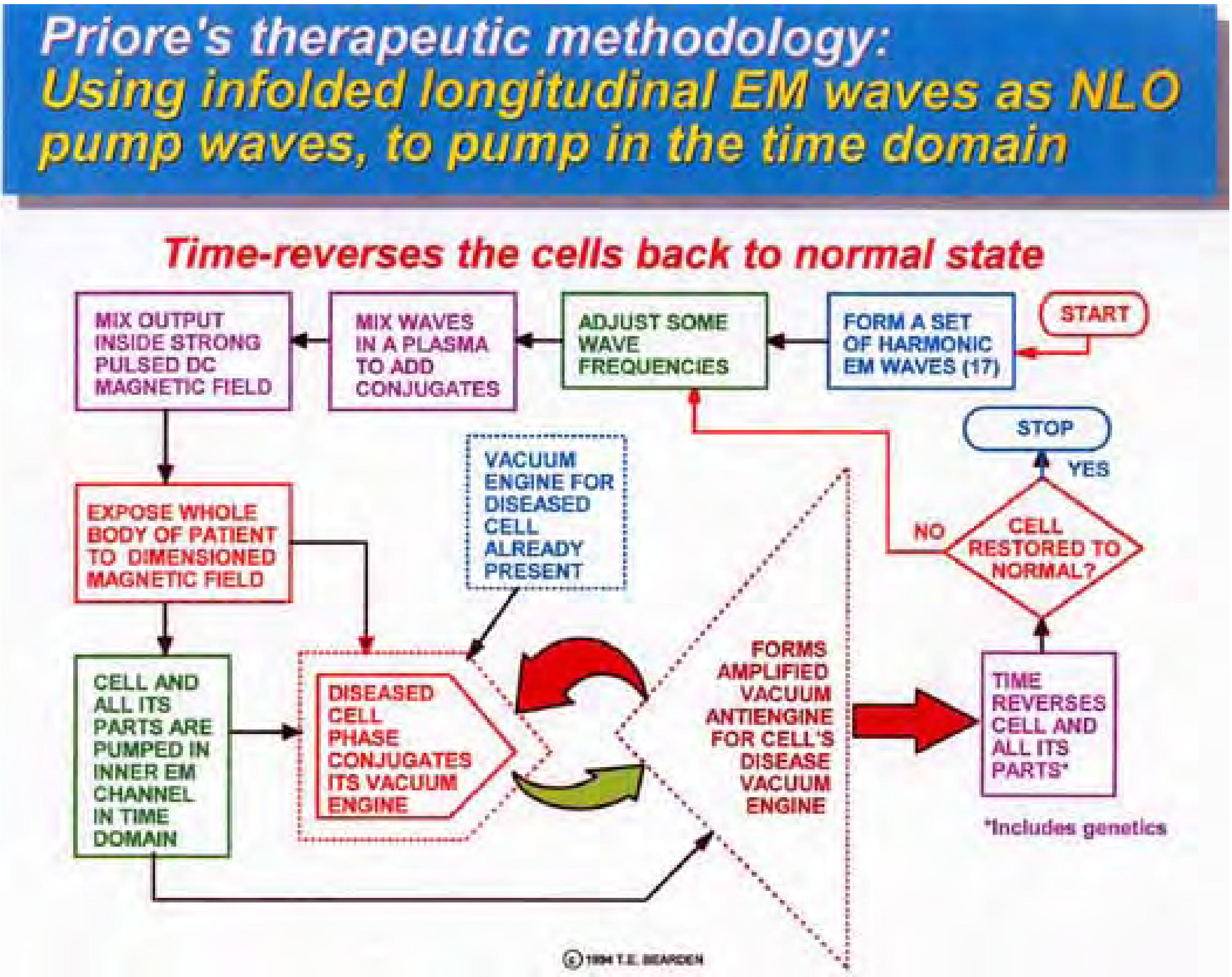

Priore's therapeutic methodology:
Using infolded longitudinal EM waves as NLO pump waves, to pump in the time domain

Time-reverses the cells back to normal state

START

FORM A SET OF HARMONIC EM WAVES (17)

ADJUST SOME WAVE FREQUENCIES

MIX WAVES IN A PLASMA TO ADD CONJUGATES

MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD

EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD

CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN

VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT

DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE

FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE

TIME REVERSES CELL AND ALL ITS PARTS*

*Includes genetics

CELL RESTORED TO NORMAL?

NO

YES

STOP

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

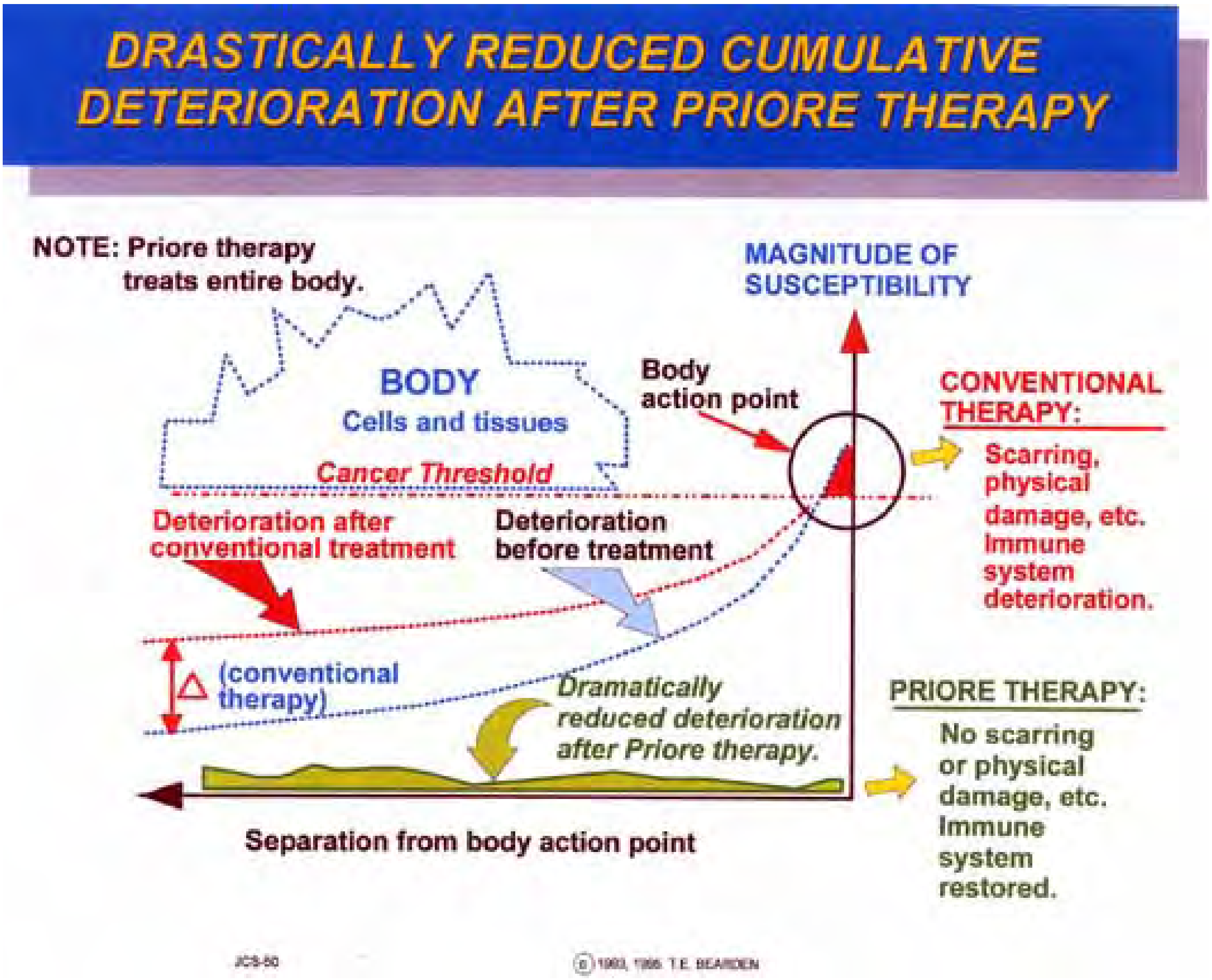

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy
treats entire body.

BODY
Cells and tissues

Cancer Threshold

MAGNITUDE OF
SUSCEPTIBILITY

Body
action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

Deterioration after
conventional treatment

Deterioration
before treatment

Δ (conventional therapy)

Dramatically reduced deterioration after Priore therapy.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

Separation from body action point

JCS-50

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

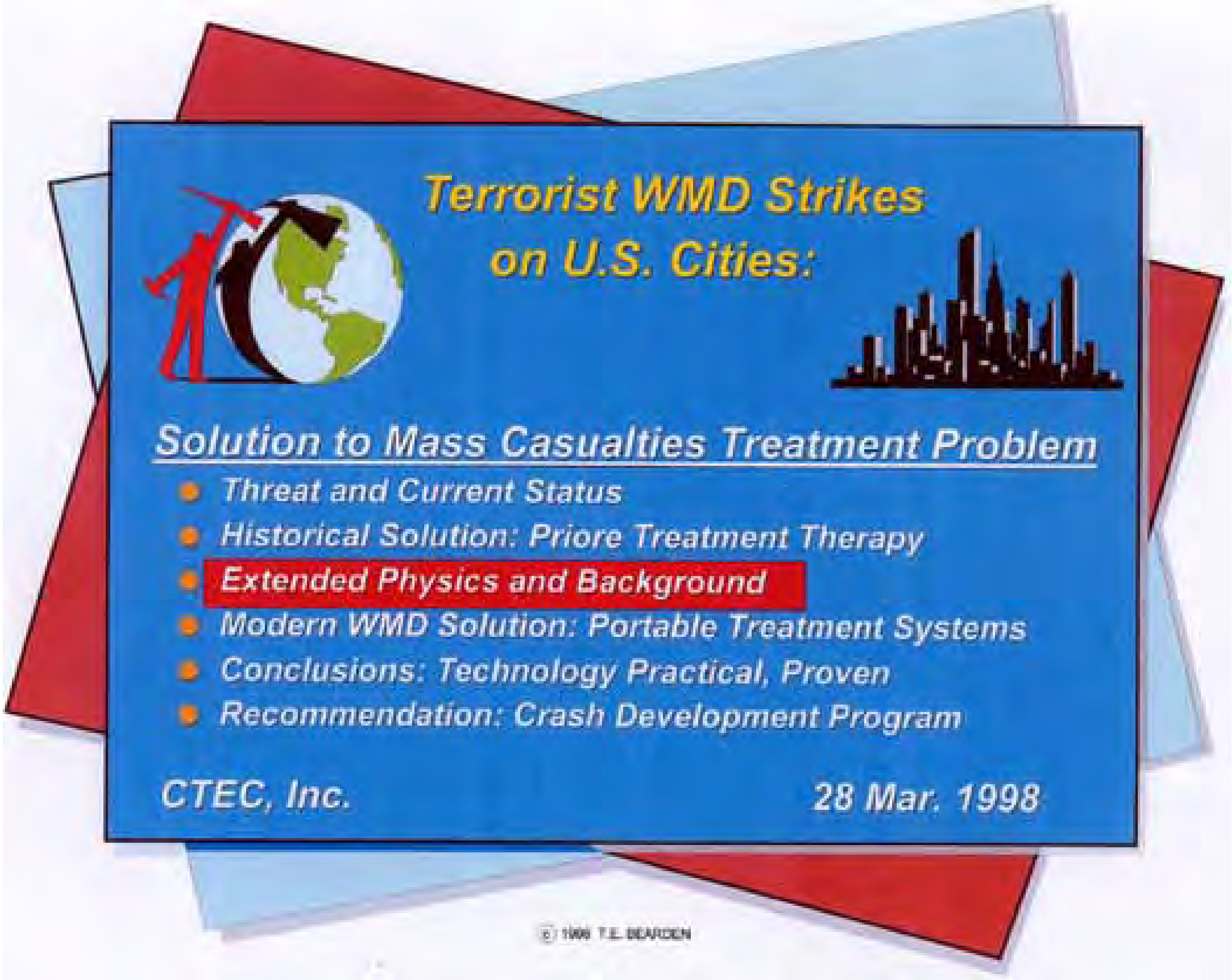

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

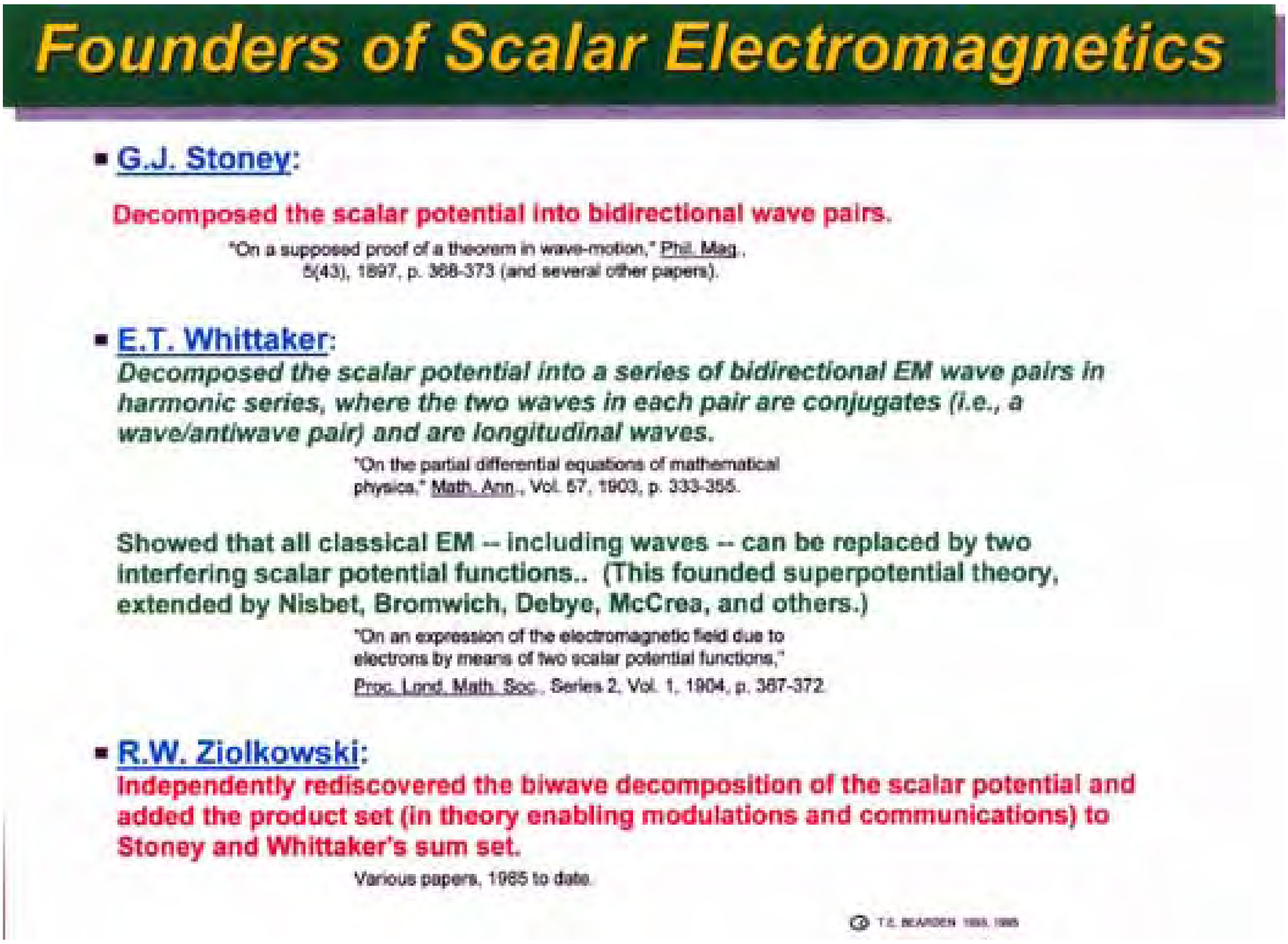

Founders of Scalar Electromagnetics

- **G.J. Stoney:**

Decomposed the scalar potential into bidirectional wave pairs.

"On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker:**

Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

"On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

"On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.

- **R.W. Ziolkowski:**

Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

Various papers, 1985 to date.

© T.E. BEARDEN [illegible]

Return to Slide Index

Next Slide

Previous Slide

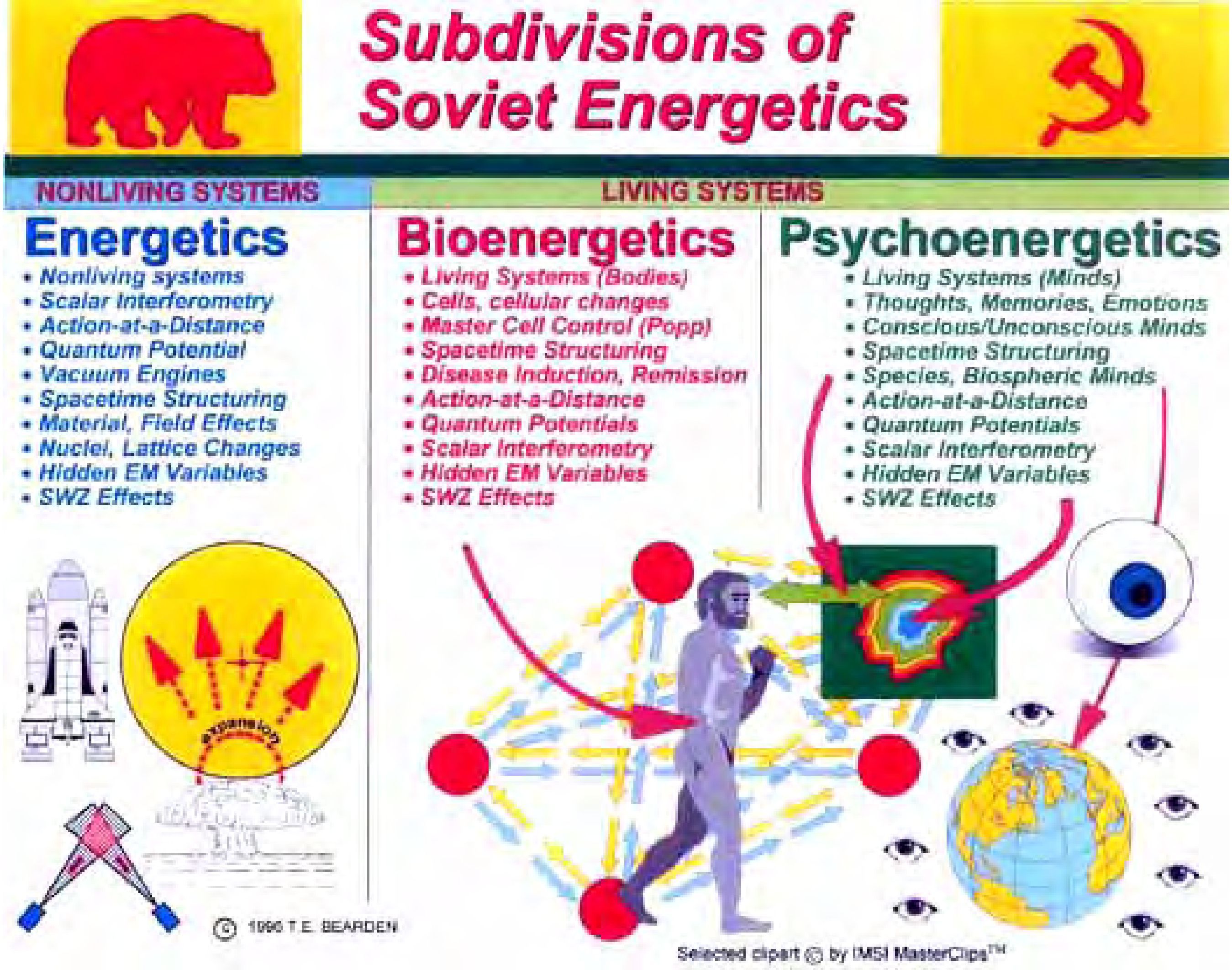

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

LIVING SYSTEMS

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1990 T.E. BEARDEN

Selected clipart © by IMSI MasterClips™

Return to Slide Index

Next Slide

Previous Slide

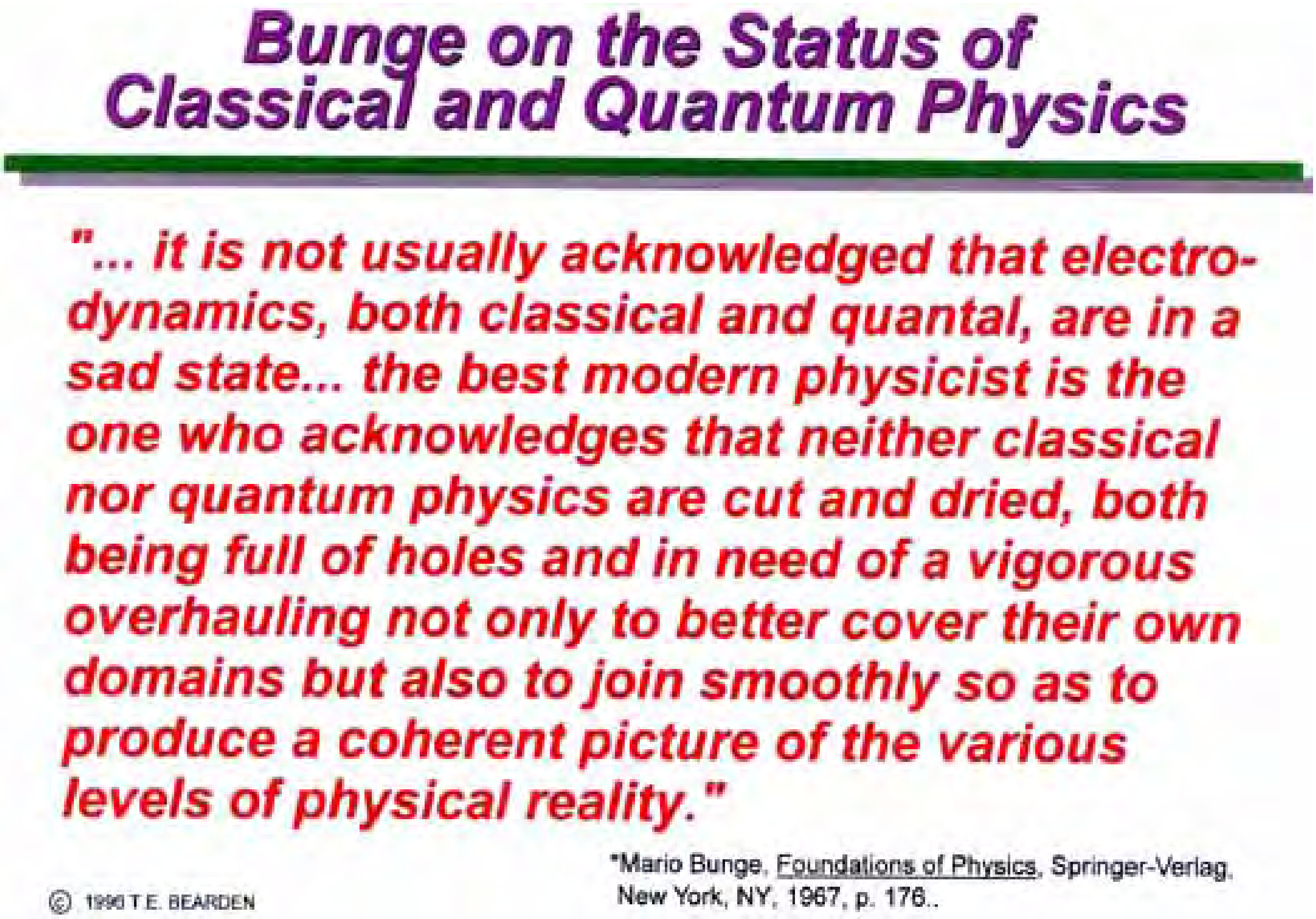

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

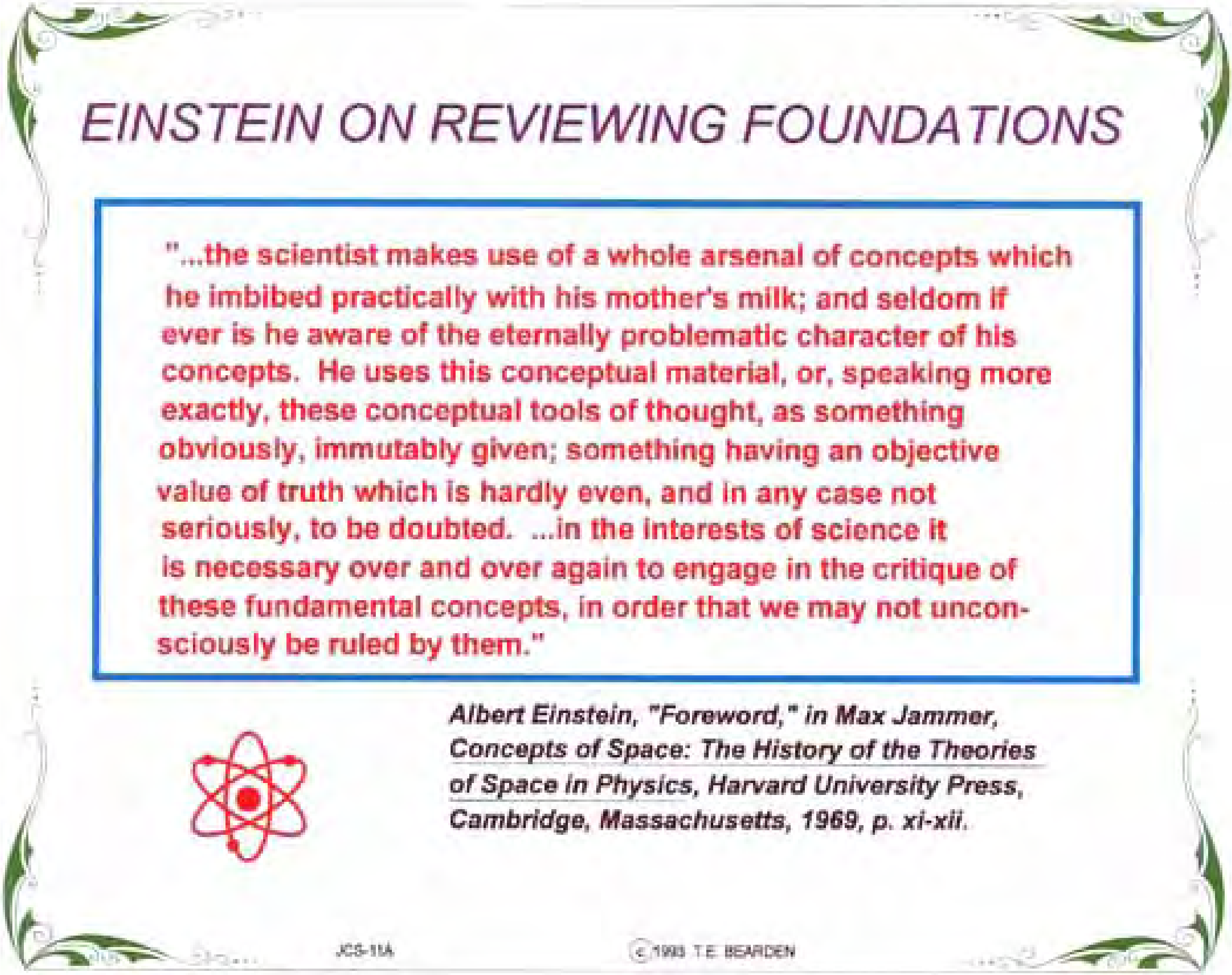

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

Fields and Potentials have been "defined" only in and on matter

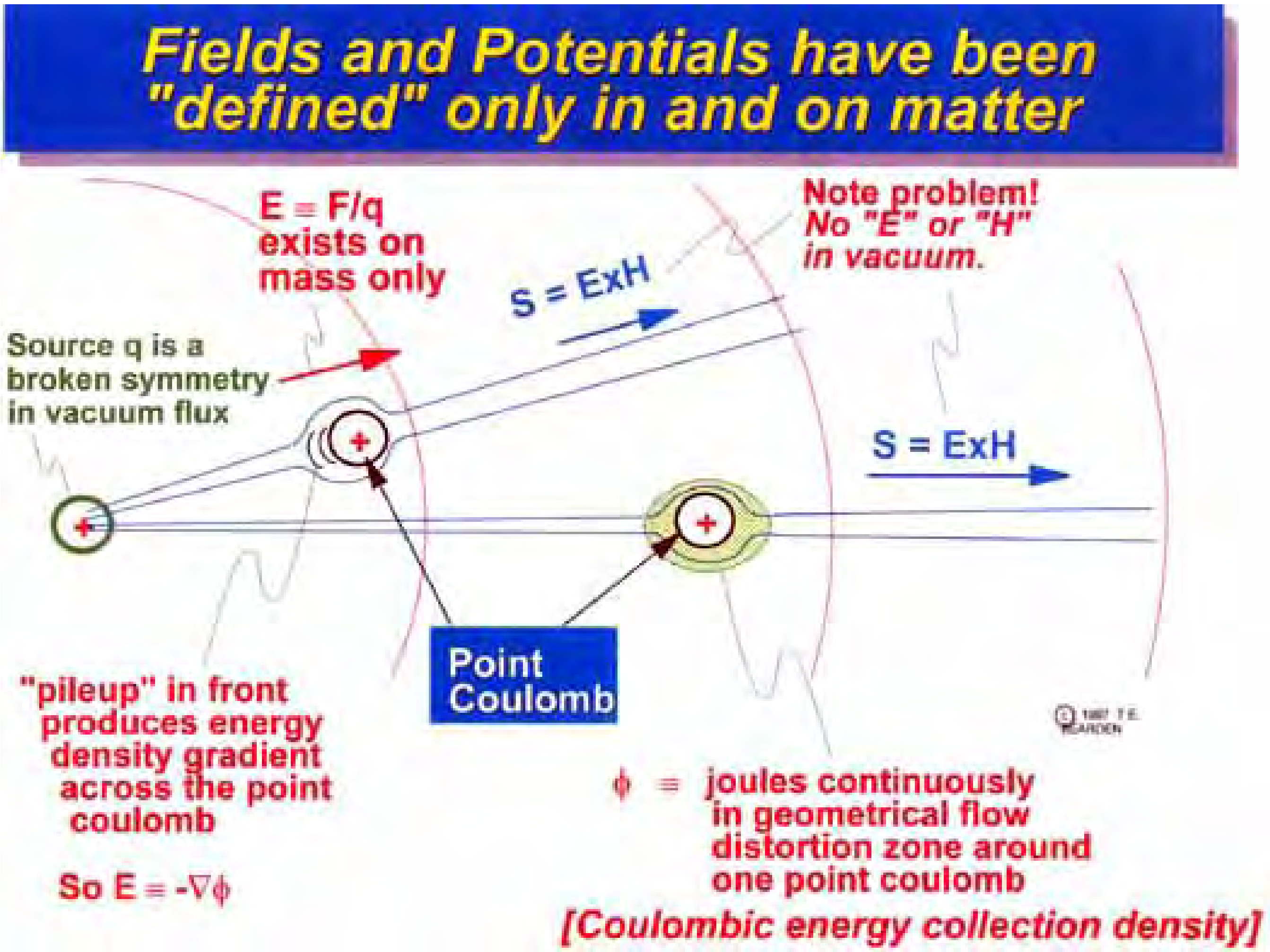

$E \equiv F/q$
exists on
mass only

Note problem!
No "E" or "H"
in vacuum.

$S = E \times H$

Source q is a
broken symmetry
in vacuum flux

$S = E \times H$

Point Coulomb

"pileup" in front
produces energy
density gradient
across the point
coulomb

So $E \equiv -\nabla\phi$

$\phi \equiv$ joules continuously
in geometrical flow
distortion zone around
one point coulomb

[Coulombic energy collection density]

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

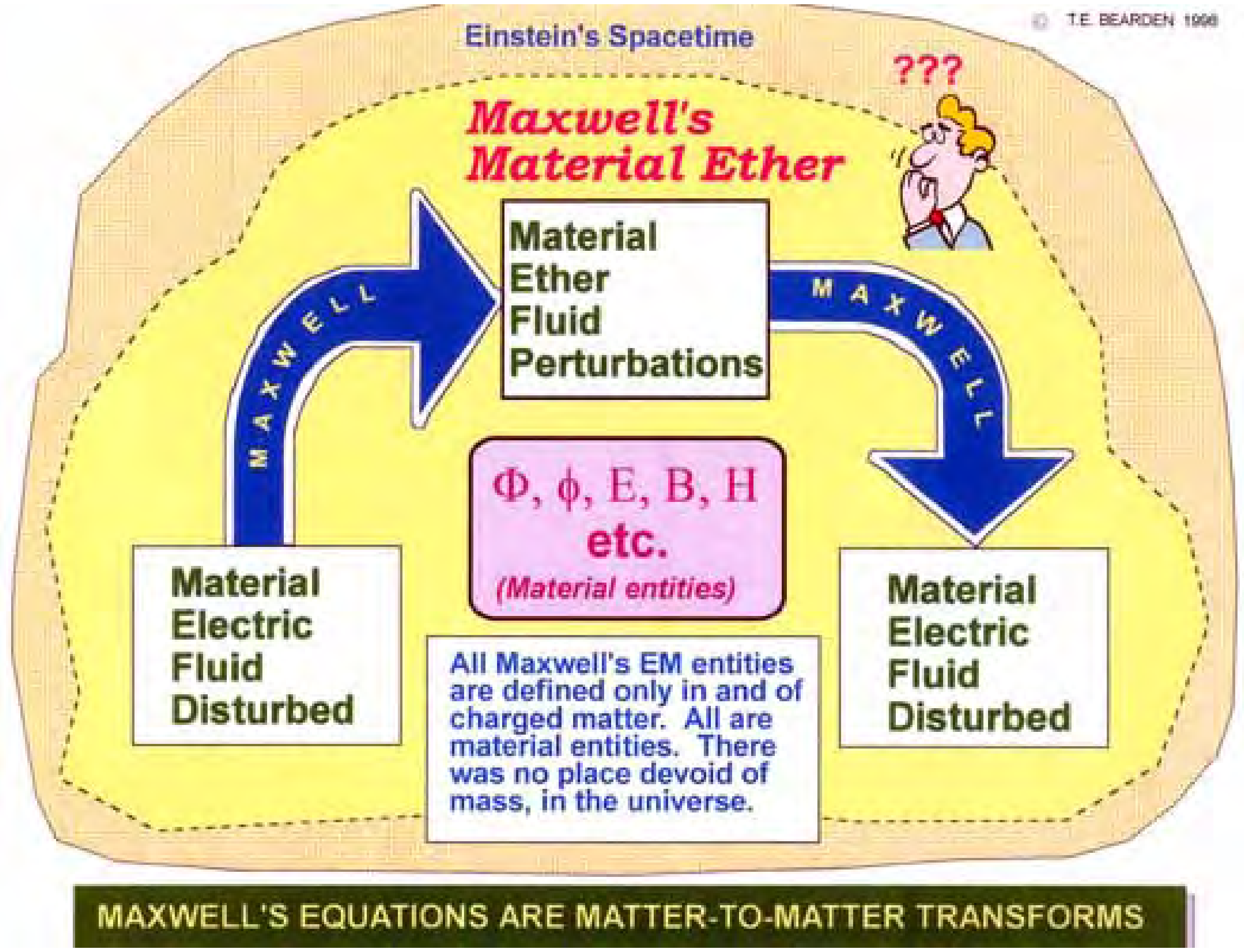

Einstein's Spacetime

© T.E. BEARDEN 1998

Maxwell's Material Ether

Material Ether Fluid Perturbations

Material Electric Fluid Disturbed

Φ, ϕ, E, B, H etc.

(Material entities)

Material Electric Fluid Disturbed

All Maxwell's EM entities are defined only in and of charged matter. All are material entities. There was no place devoid of mass, in the universe.

MAXWELL'S EQUATIONS ARE MATTER-TO-MATTER TRANSFORMS

Return to Slide Index

Next Slide

Previous Slide

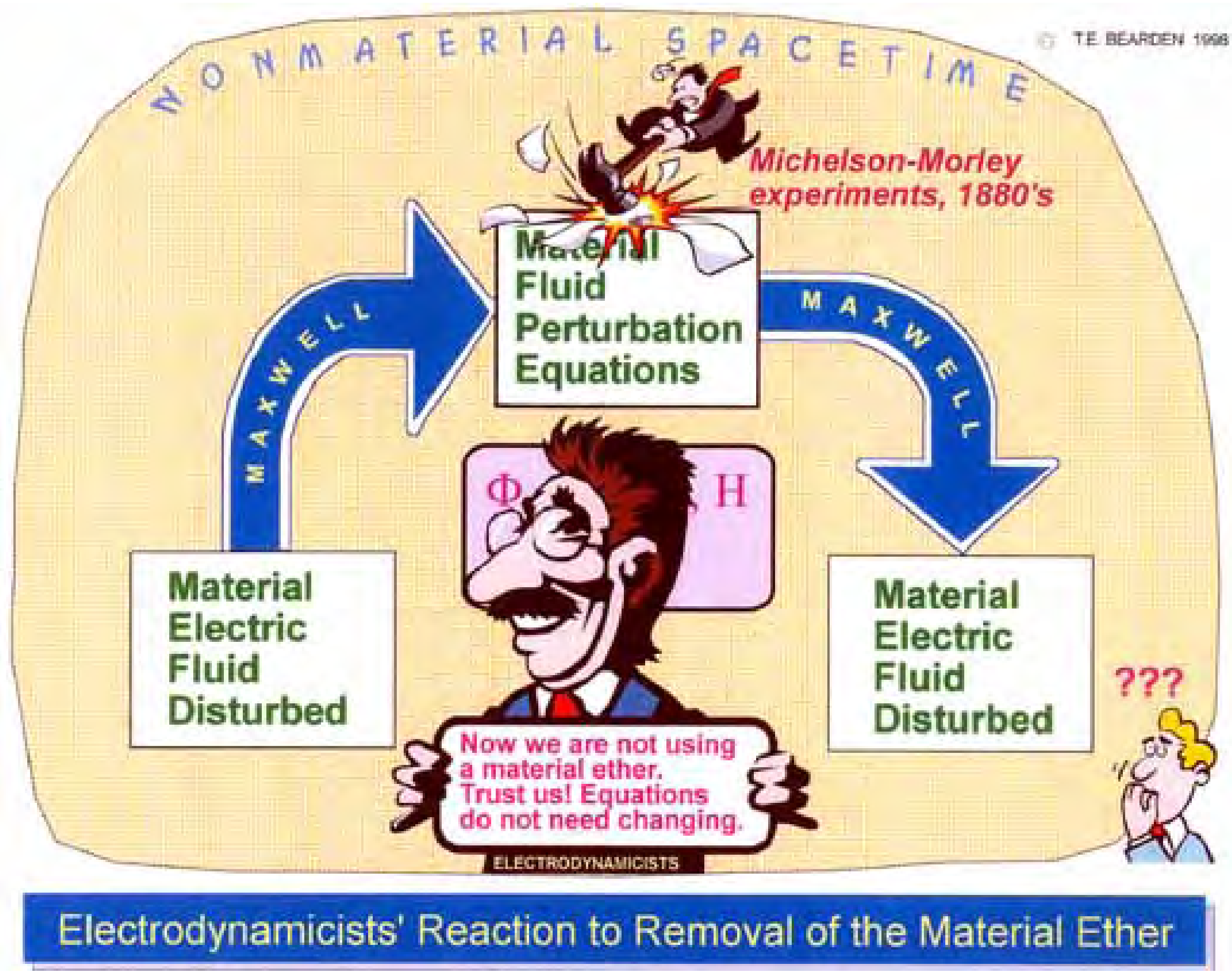

NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Michelson-Morley experiments, 1880's
MAXWELL
MAXWELL
Φ
H
???
ELECTRODYNAMICISTS

Material Fluid Perturbation Equations

Material Electric Fluid Disturbed

Material Electric Fluid Disturbed

Now we are not using a material ether. Trust us! Equations do not need changing.

Electrodynamicists' Reaction to Removal of the Material Ether

[Return to Slide Index]

[Next Slide]

[Previous Slide]

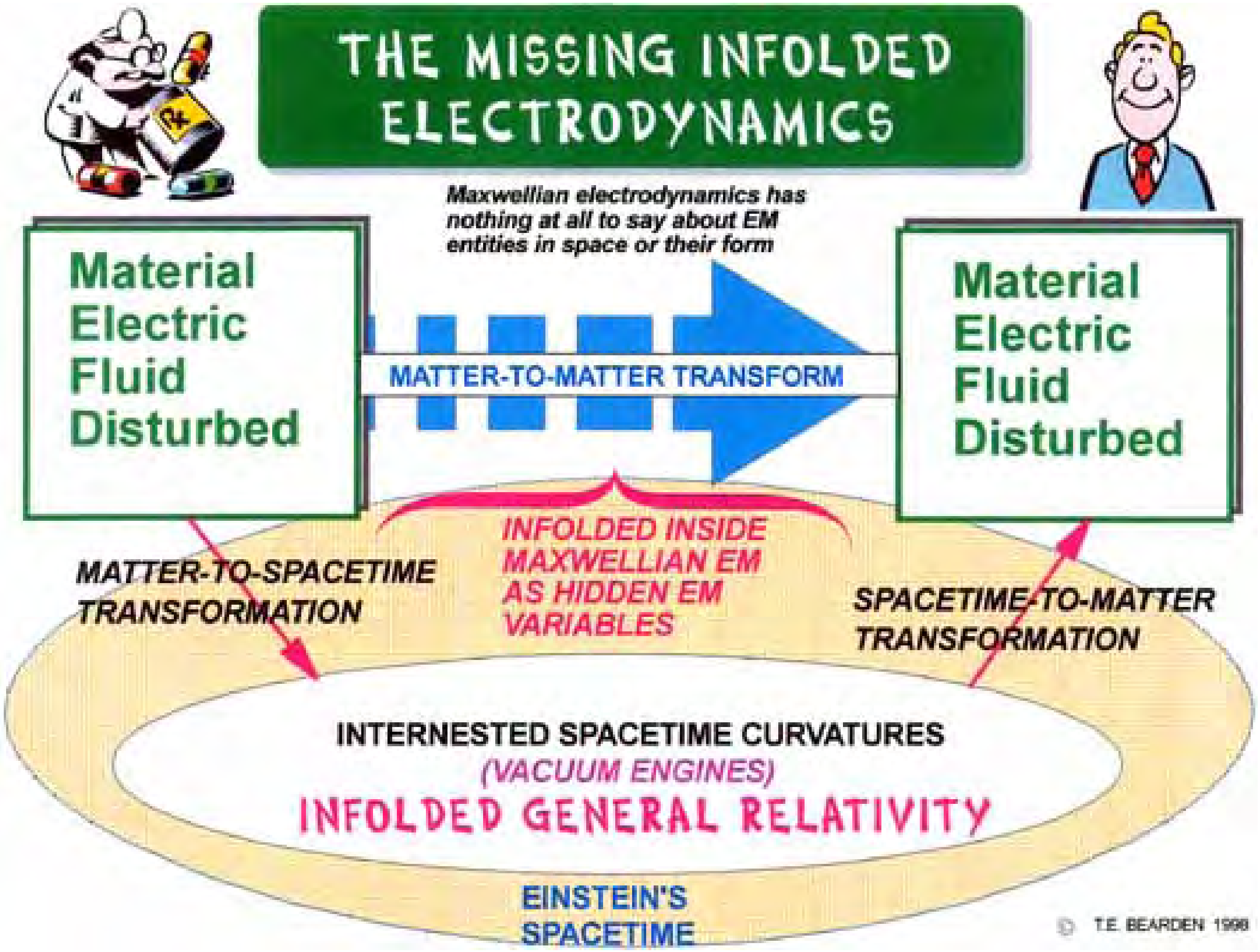

THE MISSING INFOLDED ELECTRODYNAMICS

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

Material Electric Fluid Disturbed

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

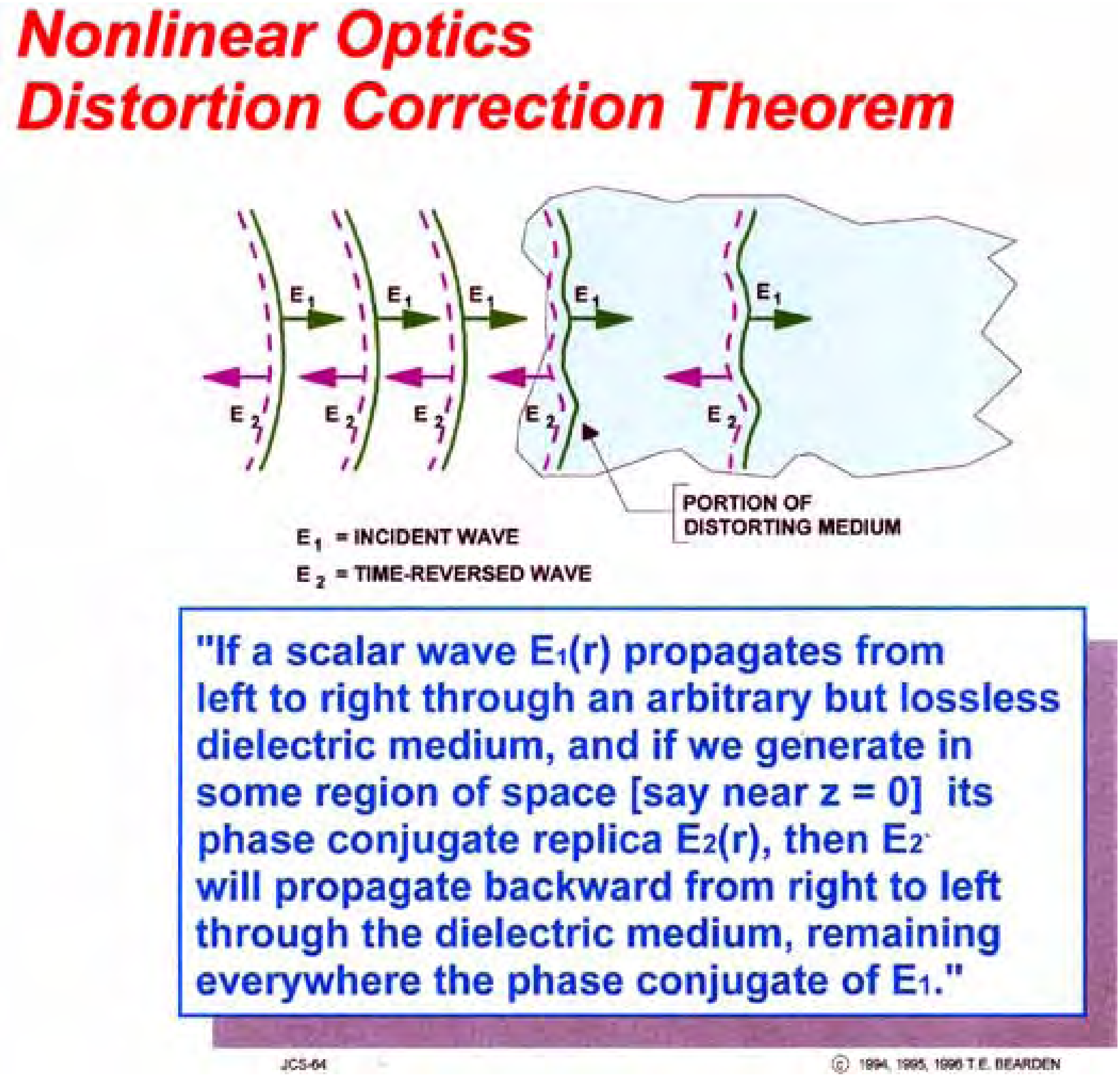

Nonlinear Optics Distortion Correction Theorem

E_1 = INCIDENT WAVE
E_2 = TIME-REVERSED WAVE

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z = 0] its phase conjugate replica $E_2(r)$, then E_2^{*} will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

JCS-64

© 1994, 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

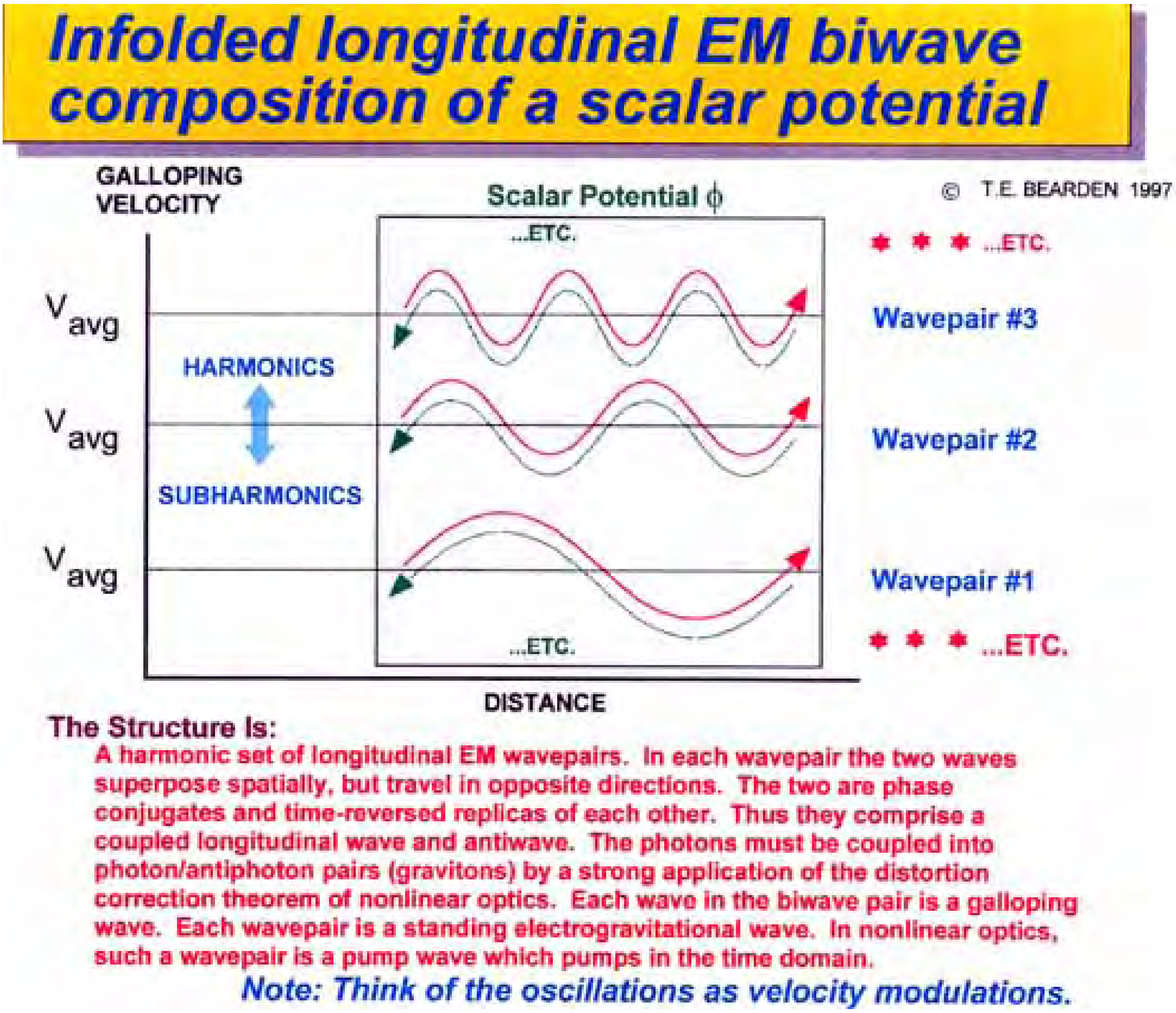

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
✱ ✱ ✱ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
SUBHARMONICS
V_{avg}
Wavepair #2
V_{avg}
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Return to Slide Index

Next Slide

Previous Slide

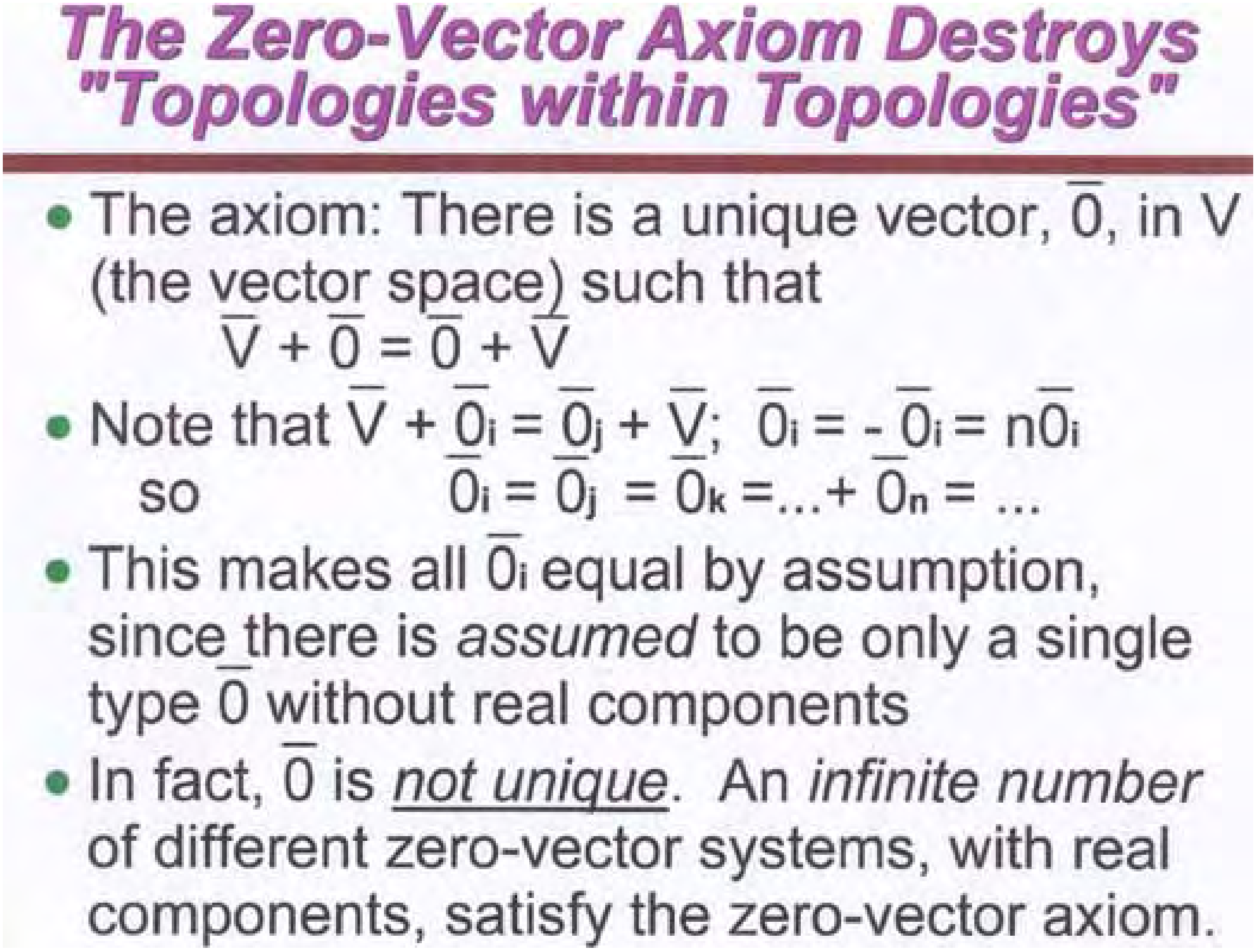

The Zero-Vector Axiom Destroys "Topologies within Topologies"

- The axiom: There is a unique vector, $\overline{0}$, in V (the vector space) such that
  $\overline{V} + \overline{0} = \overline{0} + \overline{V}$
- Note that $\overline{V} + \overline{0}_i = \overline{0}_j + \overline{V}$; $\overline{0}_i = -\overline{0}_i = n\overline{0}_i$
  so $\overline{0}_i = \overline{0}_j = \overline{0}_k = \ldots + \overline{0}_n = \ldots$
- This makes all $\overline{0}_i$ equal by assumption, since there is *assumed* to be only a single type $\overline{0}$ without real components
- In fact, $\overline{0}$ is *not unique*. An *infinite number* of different zero-vector systems, with real components, satisfy the zero-vector axiom.

Return to Slide Index

Next Slide

Previous Slide

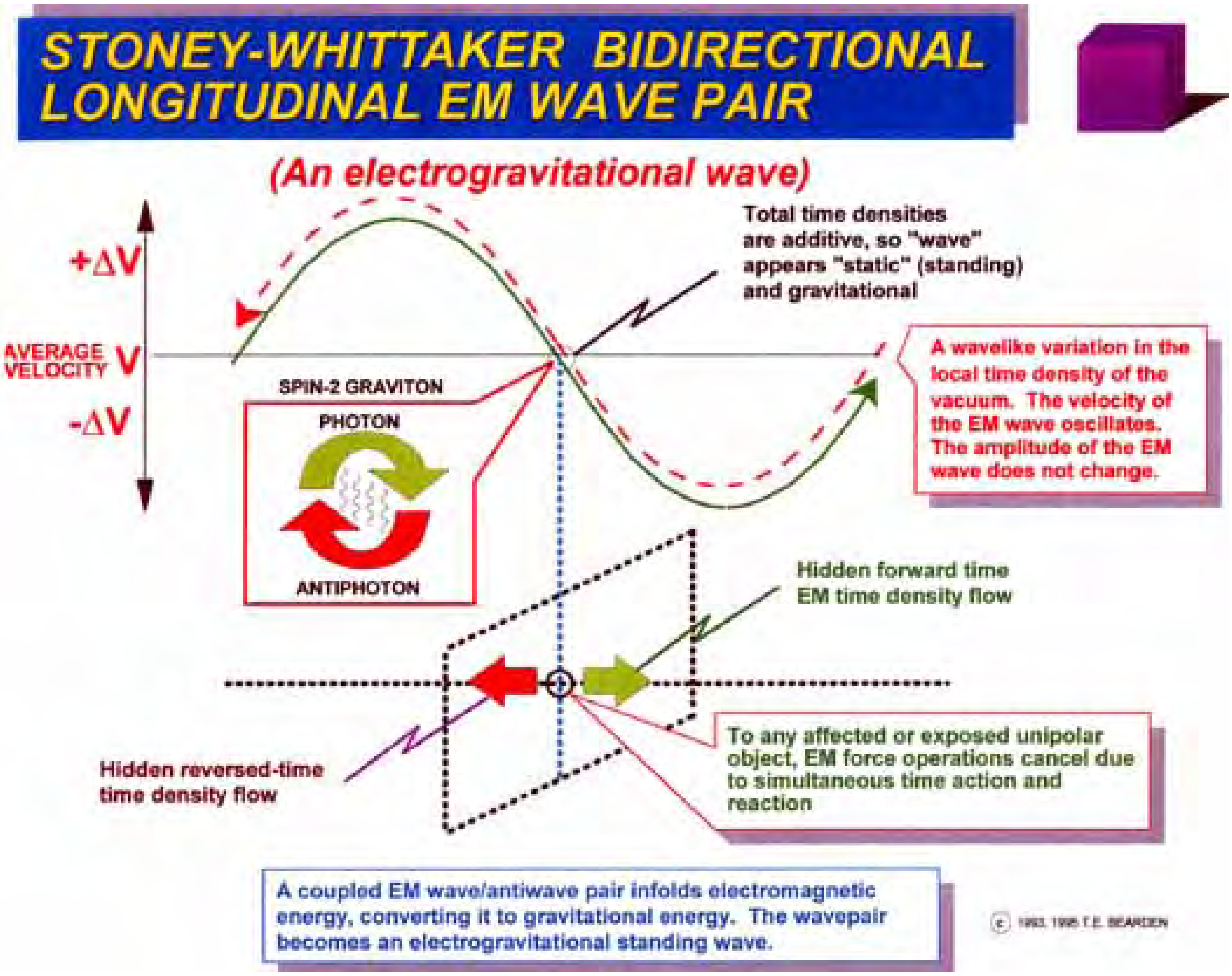

STONEY-WHITTAKER BIDIRECTIONAL LONGITUDINAL EM WAVE PAIR

(An electrogravitational wave)

Total time densities are additive, so "wave" appears "static" (standing) and gravitational

A wavelike variation in the local time density of the vacuum. The velocity of the EM wave oscillates. The amplitude of the EM wave does not change.

Hidden forward time EM time density flow

Hidden reversed-time time density flow

To any affected or exposed unipolar object, EM force operations cancel due to simultaneous time action and reaction

A coupled EM wave/antiwave pair infolds electromagnetic energy, converting it to gravitational energy. The wavepair becomes an electrogravitational standing wave.

© 1993, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

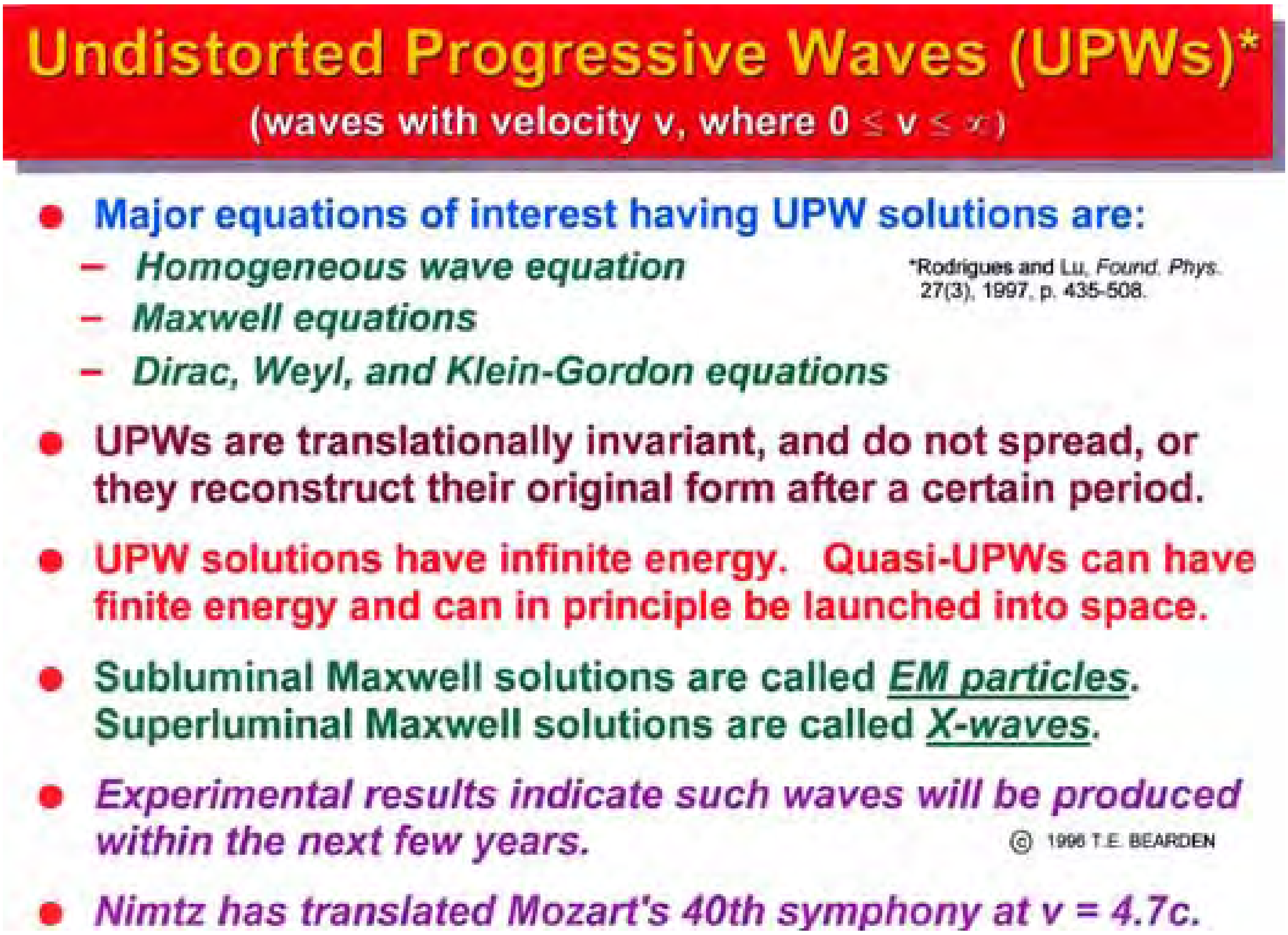

Undistorted Progressive Waves (UPWs)*

(waves with velocity v, where $0 \leq v \leq \infty$)

- Major equations of interest having UPW solutions are:
  - *Homogeneous wave equation*
  - *Maxwell equations*
  - *Dirac, Weyl, and Klein-Gordon equations*
- UPWs are translationally invariant, and do not spread, or they reconstruct their original form after a certain period.
- UPW solutions have infinite energy. Quasi-UPWs can have finite energy and can in principle be launched into space.
- Subluminal Maxwell solutions are called *EM particles*. Superluminal Maxwell solutions are called *X-waves*.
- *Experimental results indicate such waves will be produced within the next few years.*
- *Nimtz has translated Mozart's 40th symphony at v = 4.7c.*

*Rodrigues and Lu, *Found. Phys.* 27(3), 1997, p. 435-508.

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

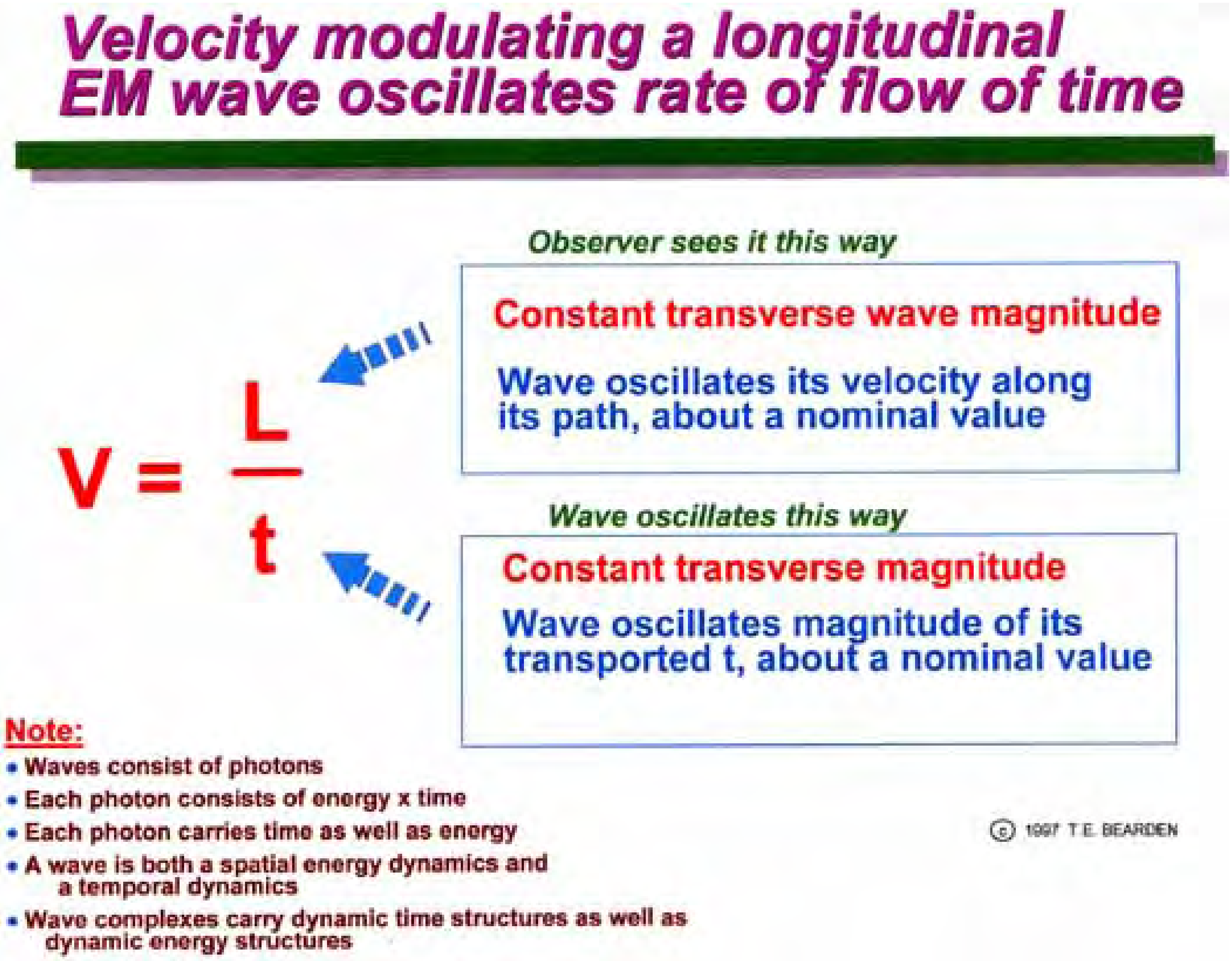

Velocity modulating a longitudinal EM wave oscillates rate of flow of time

$$V = \frac{L}{t}$$

Observer sees it this way

Constant transverse wave magnitude

Wave oscillates its velocity along its path, about a nominal value

Wave oscillates this way

Constant transverse magnitude

Wave oscillates magnitude of its transported t, about a nominal value

Note:

- Waves consist of photons
- Each photon consists of energy x time
- Each photon carries time as well as energy
- A wave is both a spatial energy dynamics and a temporal dynamics
- Wave complexes carry dynamic time structures as well as dynamic energy structures

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

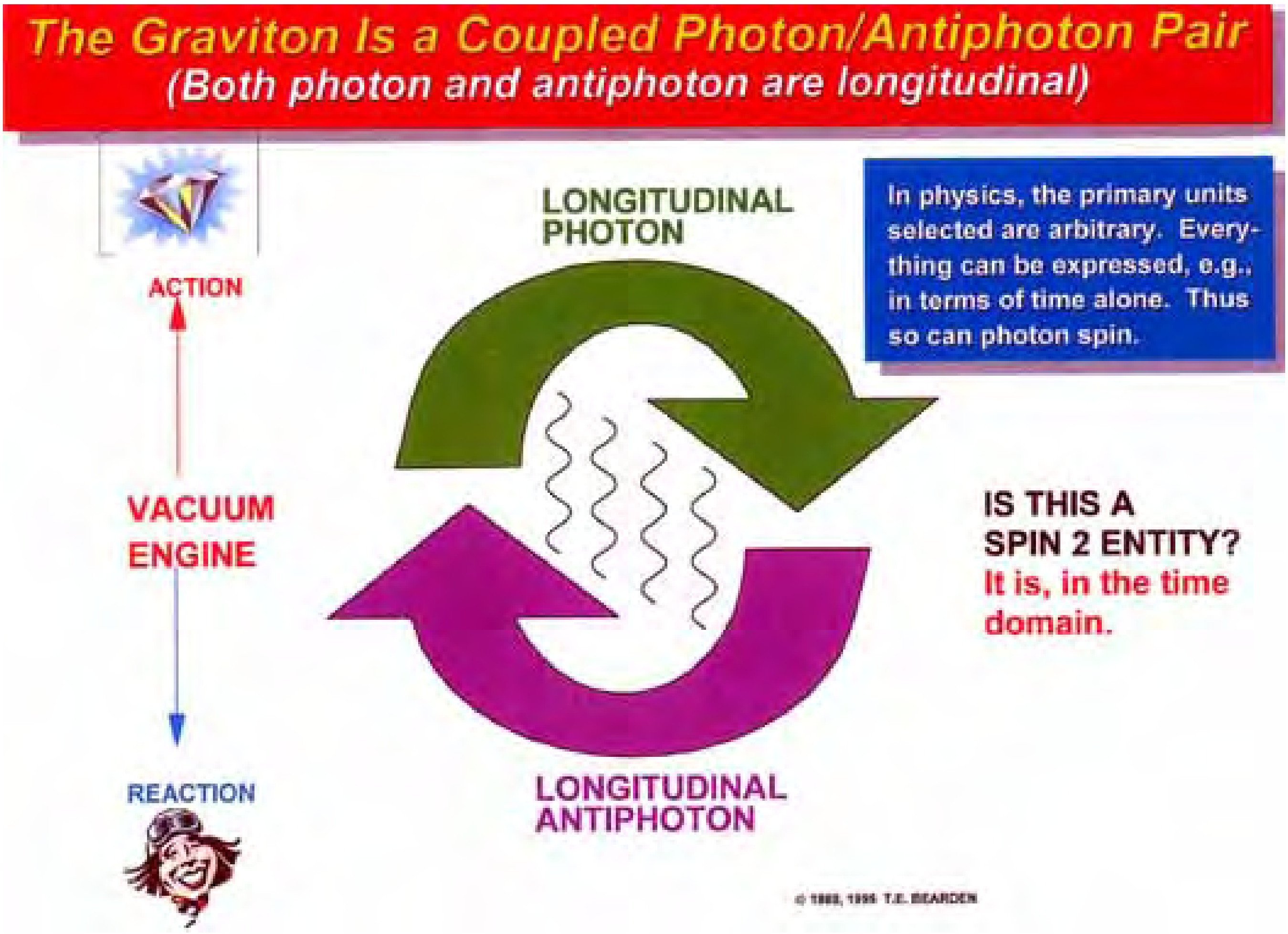

The Graviton Is a Coupled Photon/Antiphoton Pair

(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

ACTION

VACUUM ENGINE

REACTION

LONGITUDINAL PHOTON

LONGITUDINAL ANTIPHOTON

IS THIS A SPIN 2 ENTITY?

It is, in the time domain.

© 1988, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

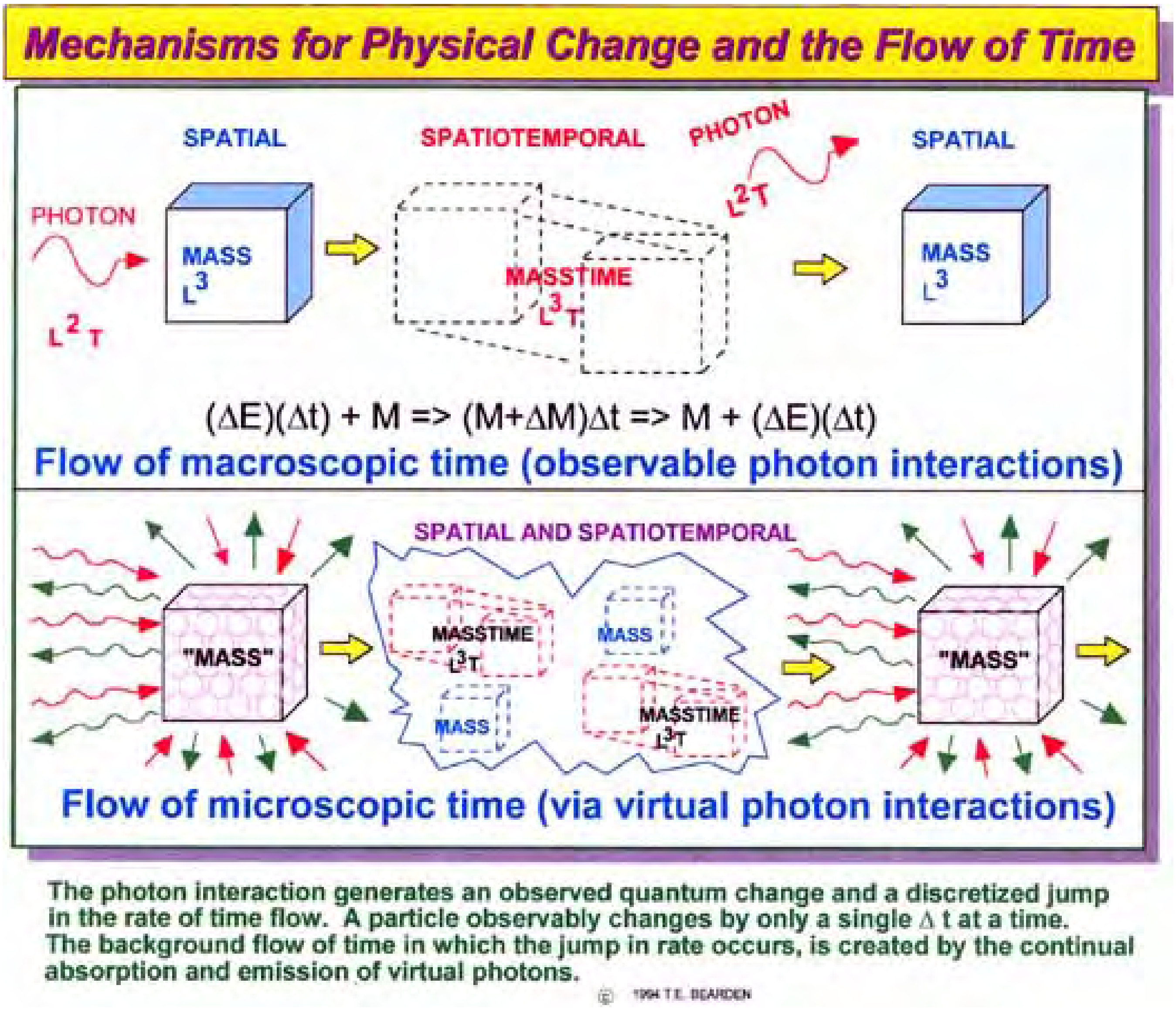

Mechanisms for Physical Change and the Flow of Time

SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L^3
L^2T
MASSTIME
L^3T
L^2T
MASS
L^3

$$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$$

Flow of macroscopic time (observable photon interactions)

SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L^3T
MASS
MASS
MASSTIME
L^3T
"MASS"

Flow of microscopic time (via virtual photon interactions)

The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.

© 1994 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

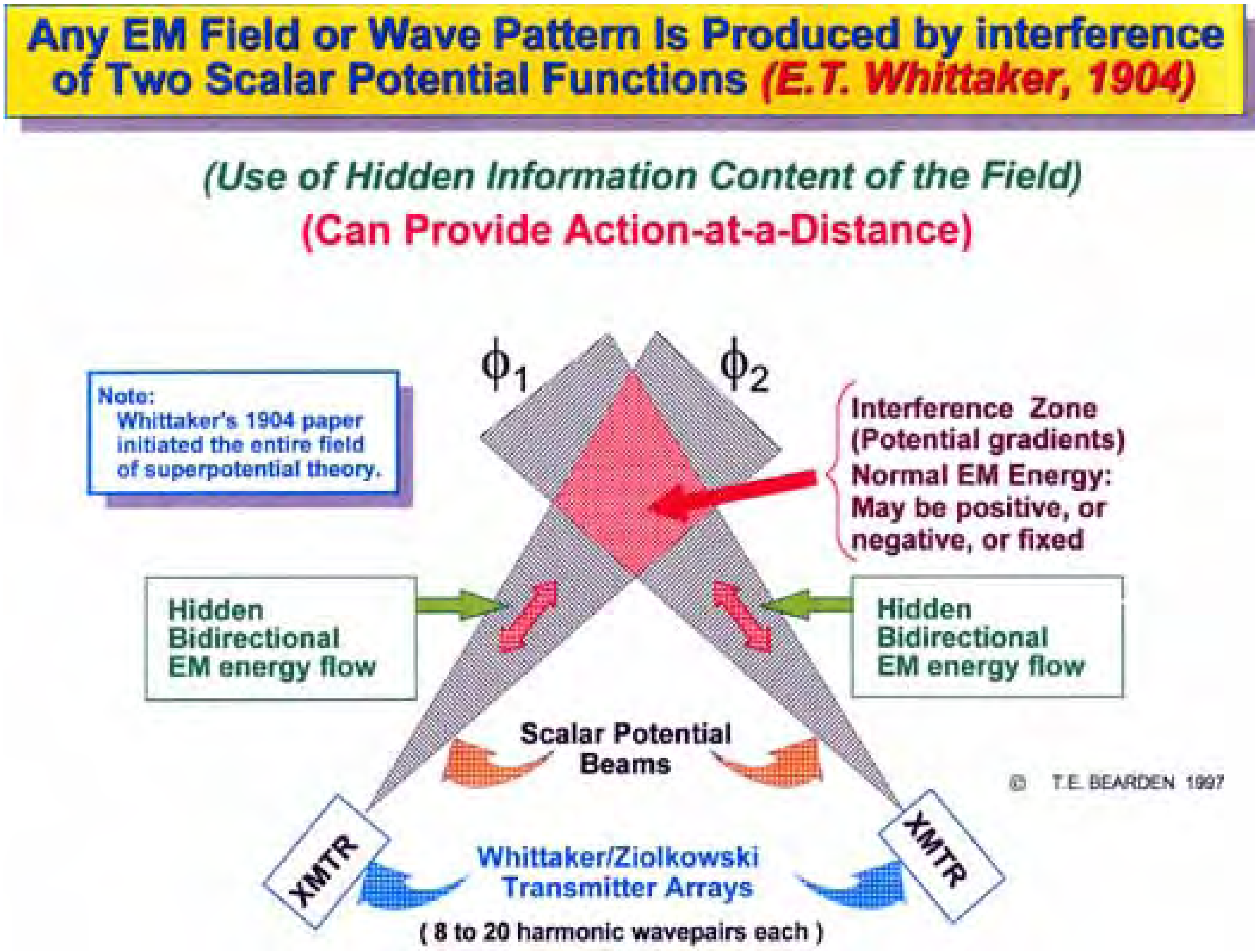

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

ϕ_1

ϕ_2

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or
negative, or fixed

Hidden
Bidirectional
EM energy flow

Hidden
Bidirectional
EM energy flow

Scalar Potential
Beams

© T.E. BEARDEN 1997

XMTR

XMTR

Whittaker/Ziolkowski
Transmitter Arrays

(8 to 20 harmonic wavepairs each)

Return to Slide Index

Next Slide

Previous Slide

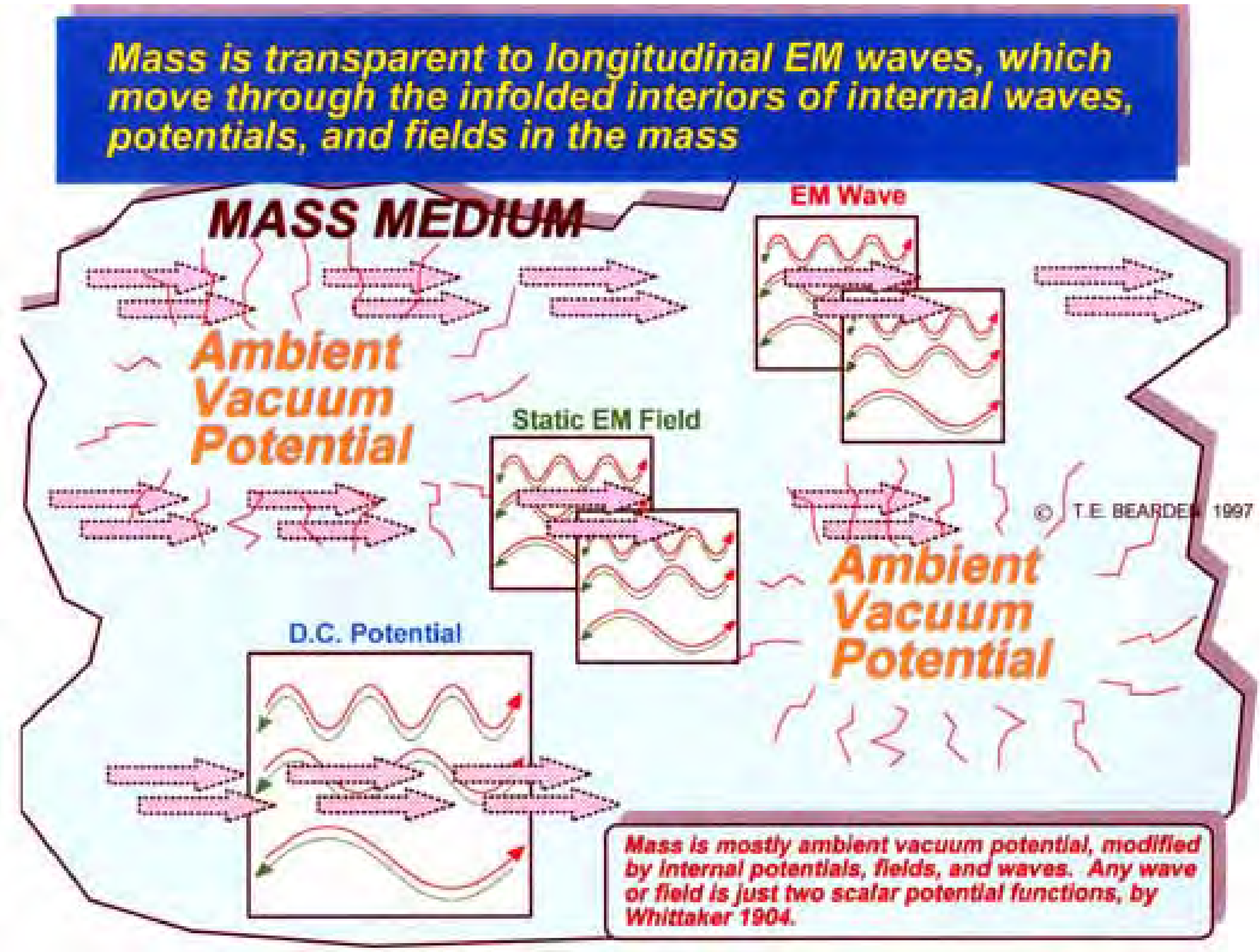

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

MASS MEDIUM
Ambient Vacuum Potential
Static EM Field
EM Wave
D.C. Potential
© T.E. BEARDEN 1997
Ambient Vacuum Potential

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

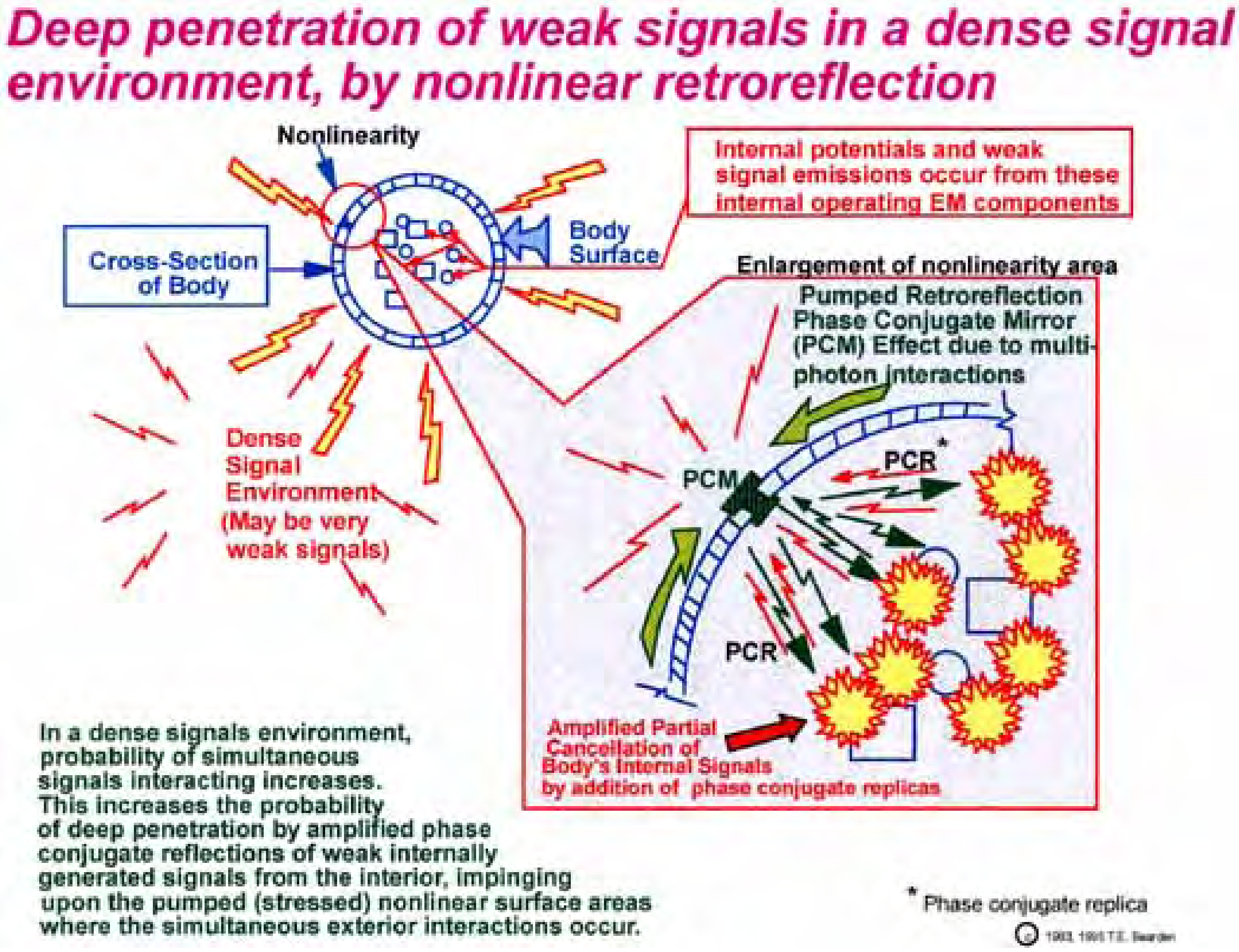

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity
Internal potentials and weak signal emissions occur from these internal operating EM components
Cross-Section of Body
Body Surface
Enlargement of nonlinearity area
Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions
Dense Signal Environment (May be very weak signals)
PCR*
PCM
PCR
Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

© 1993, 1995 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

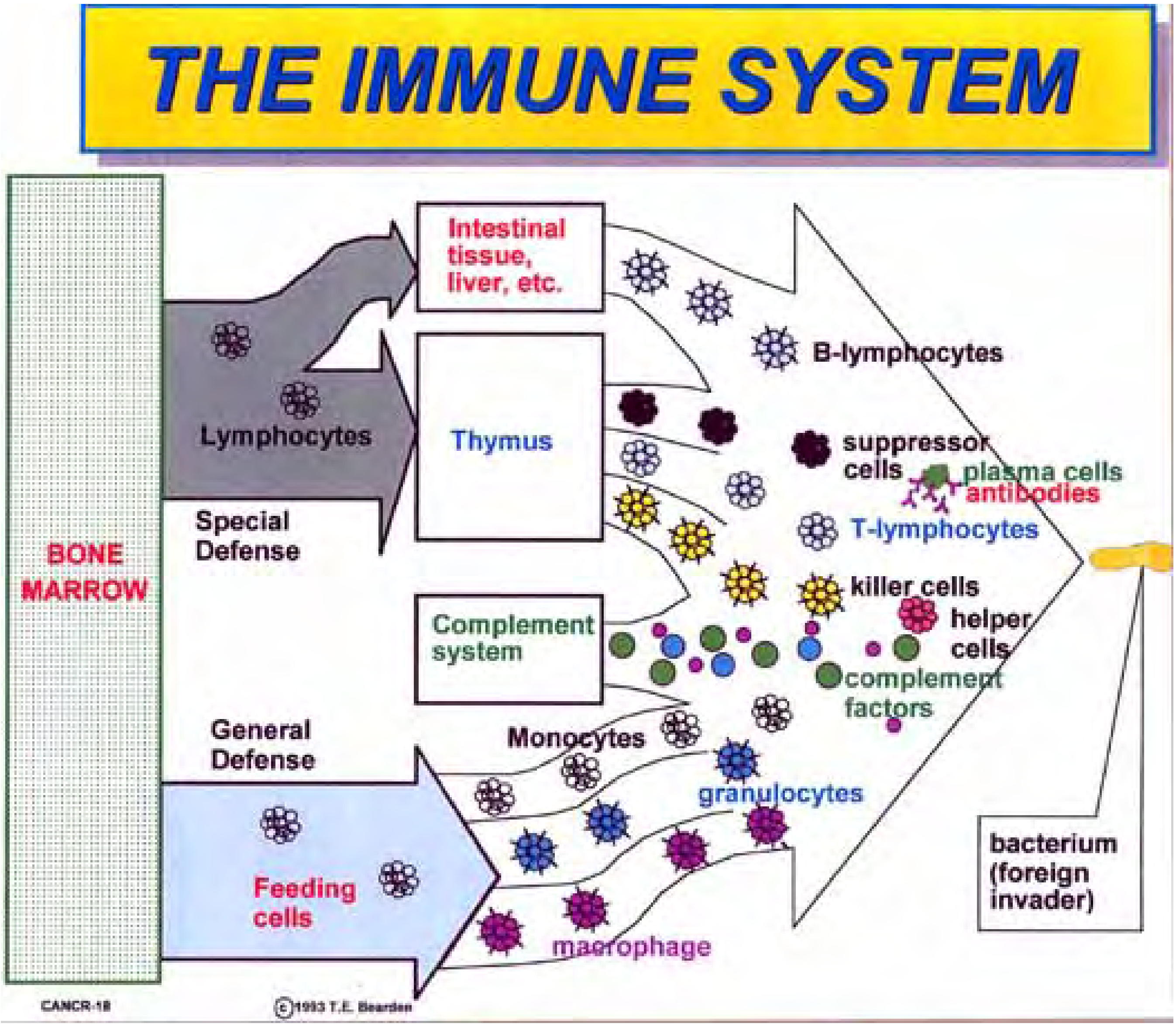

THE IMMUNE SYSTEM

BONE MARROW

Lymphocytes

Special Defense

General Defense

Feeding cells

Intestinal tissue, liver, etc.

Thymus

Complement system

B-lymphocytes

suppressor cells

plasma cells

antibodies

T-lymphocytes

killer cells

helper cells

complement factors

Monocytes

granulocytes

macrophage

bacterium (foreign invader)

CANCR-18

©1993 T.E. Bearden

[Return to Slide Index]

[Next Slide]

[Previous Slide]

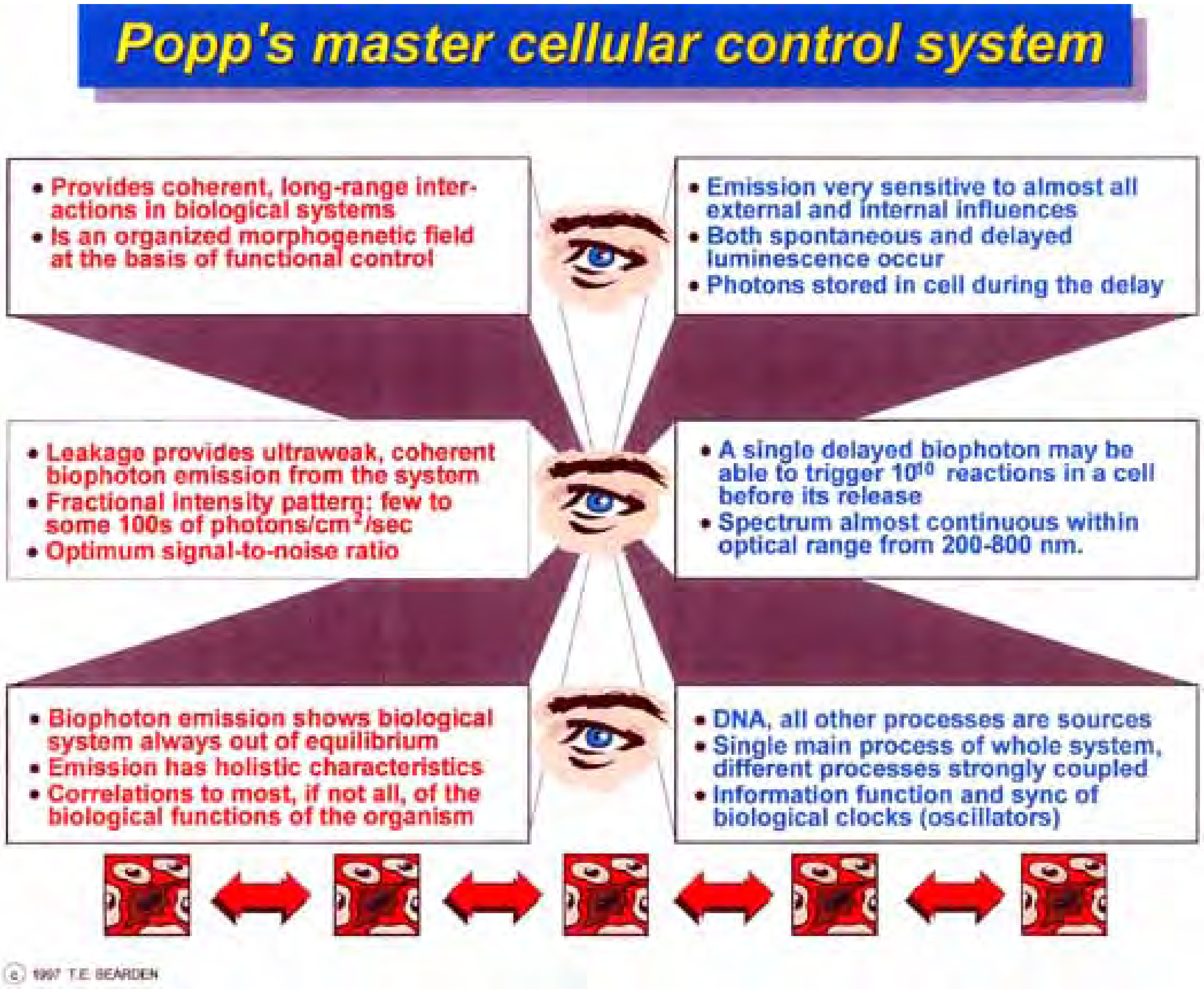

Popp's master cellular control system

- Provides coherent, long-range interactions in biological systems
- Is an organized morphogenetic field at the basis of functional control

- Emission very sensitive to almost all external and internal influences
- Both spontaneous and delayed luminescence occur
- Photons stored in cell during the delay

- Leakage provides ultraweak, coherent biophoton emission from the system
- Fractional intensity pattern: few to some 100s of photons/cm^2/sec
- Optimum signal-to-noise ratio

- A single delayed biophoton may be able to trigger 10^{10} reactions in a cell before its release
- Spectrum almost continuous within optical range from 200-800 nm.

- Biophoton emission shows biological system always out of equilibrium
- Emission has holistic characteristics
- Correlations to most, if not all, of the biological functions of the organism

- DNA, all other processes are sources
- Single main process of whole system, different processes strongly coupled
- Information function and sync of biological clocks (oscillators)

© 1997 T.E. BEARDEN

[Return to Slide Index]
[Next Slide]
[Previous Slide]

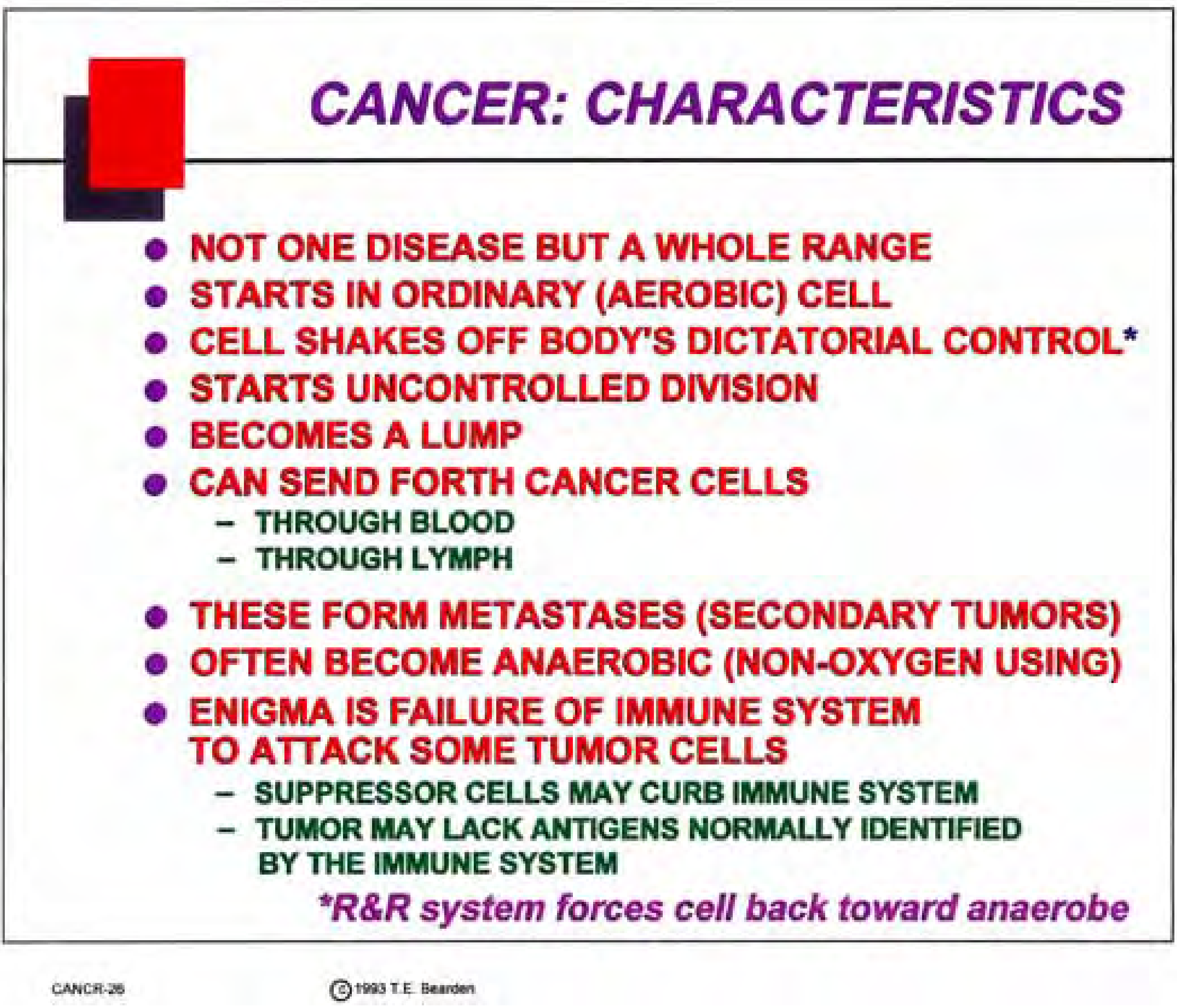

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

**R&R system forces cell back toward anaerobe*

CANCR-26

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

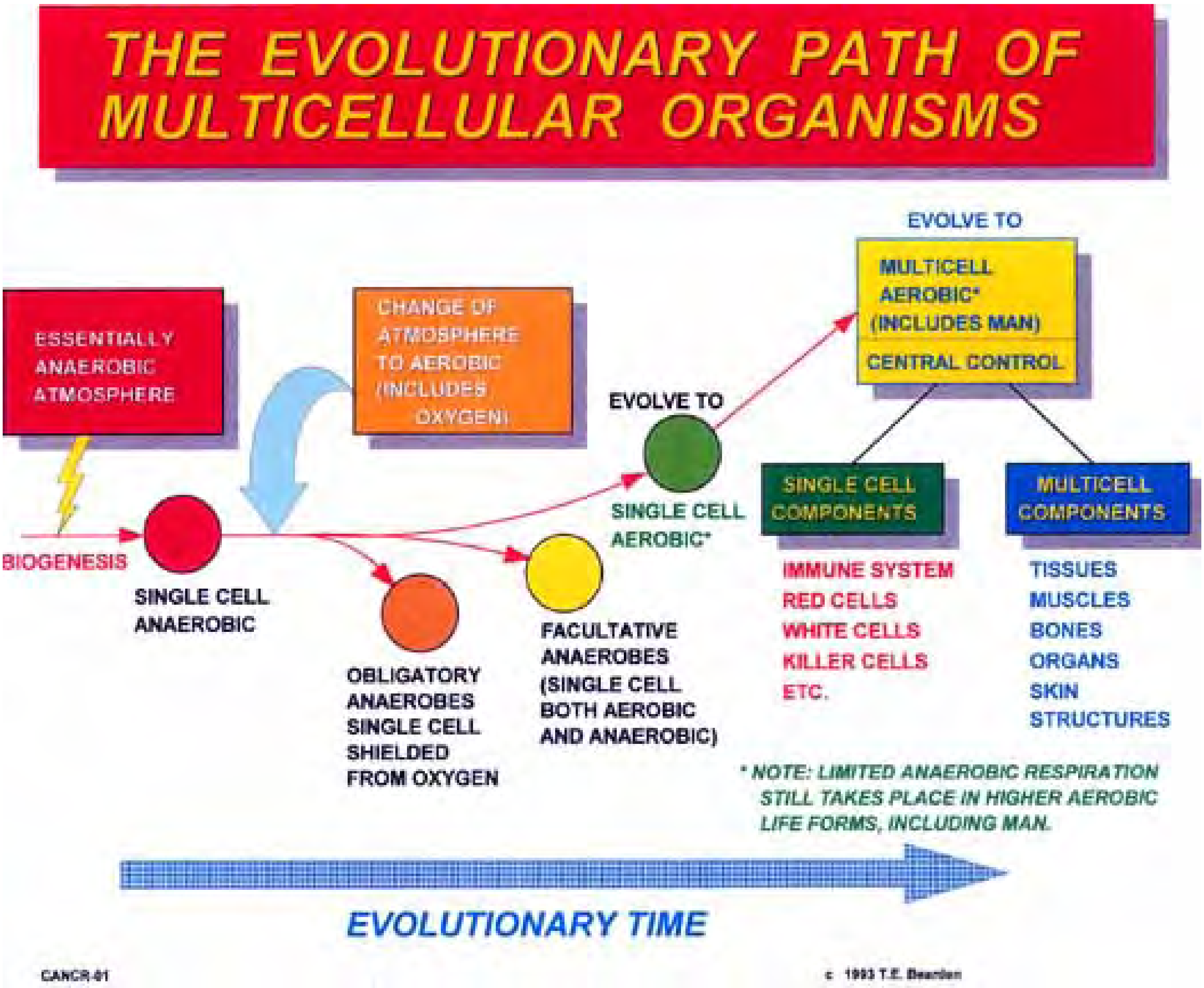

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

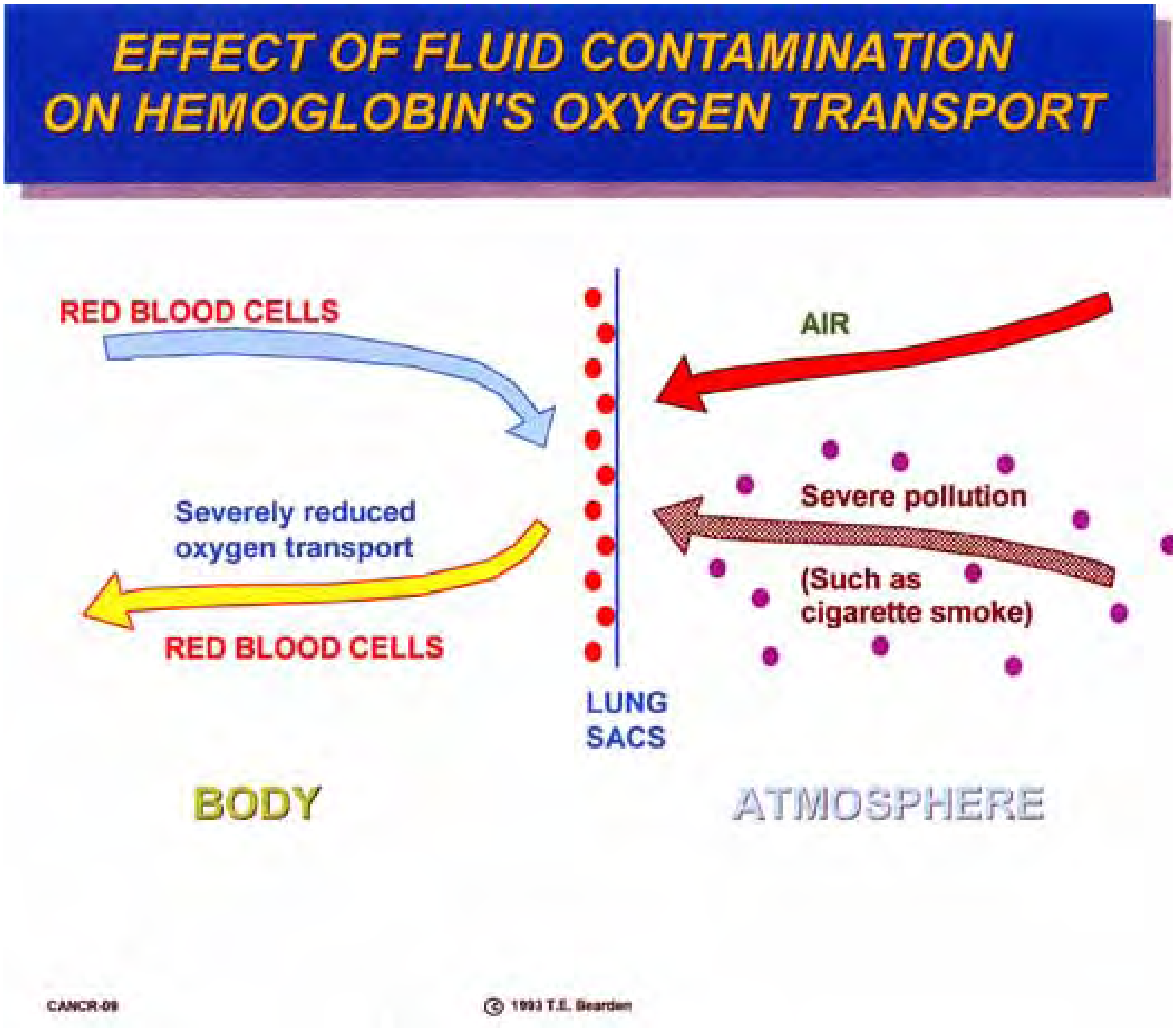

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severe pollution

(Such as cigarette smoke)

Severely reduced oxygen transport

RED BLOOD CELLS

LUNG SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

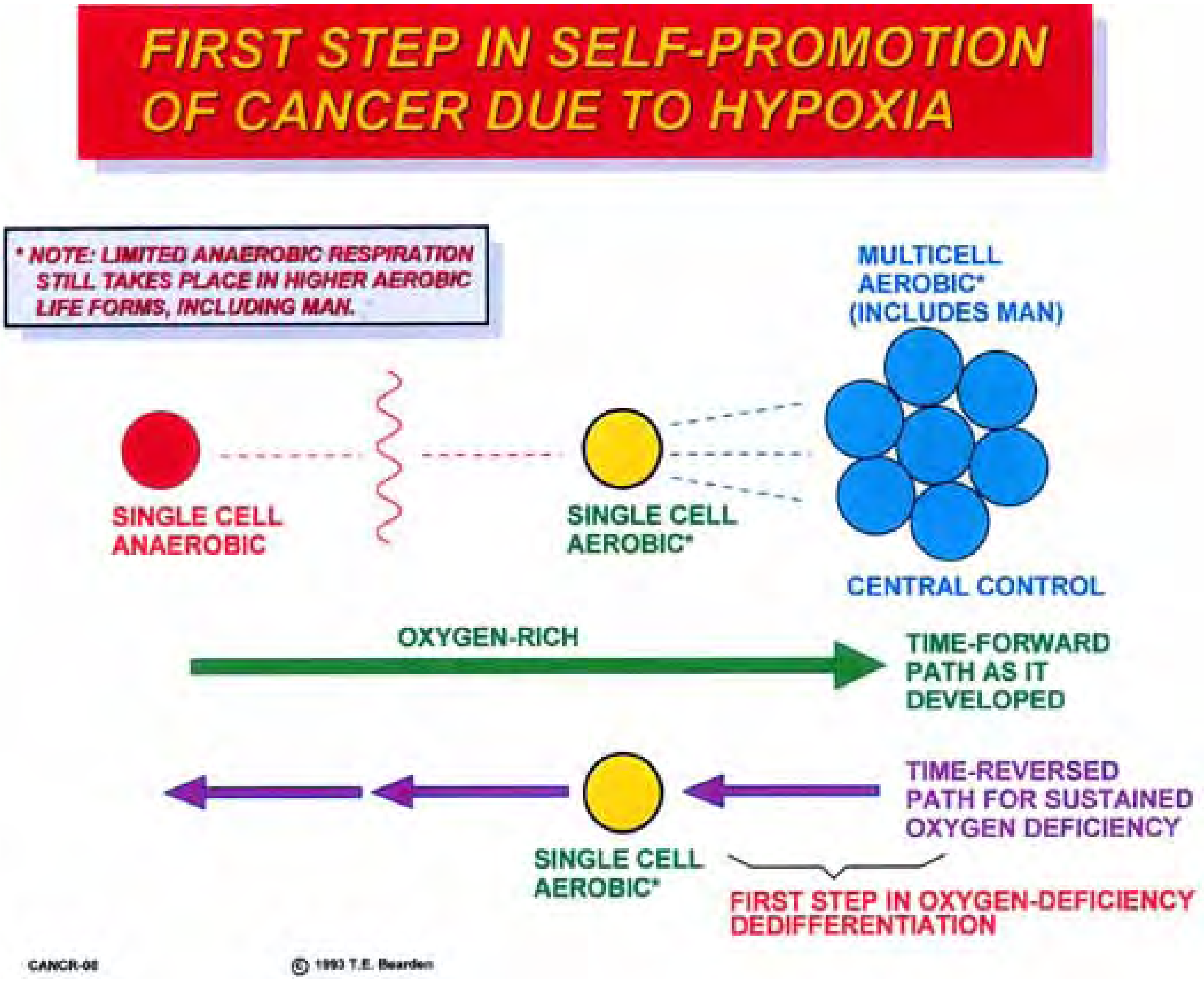

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

SINGLE CELL ANAEROBIC

SINGLE CELL AEROBIC*

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

OXYGEN-RICH

TIME-FORWARD PATH AS IT DEVELOPED

TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY

SINGLE CELL AEROBIC*

FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

[Return to Slide Index]

[Next Slide]

[Previous Slide]

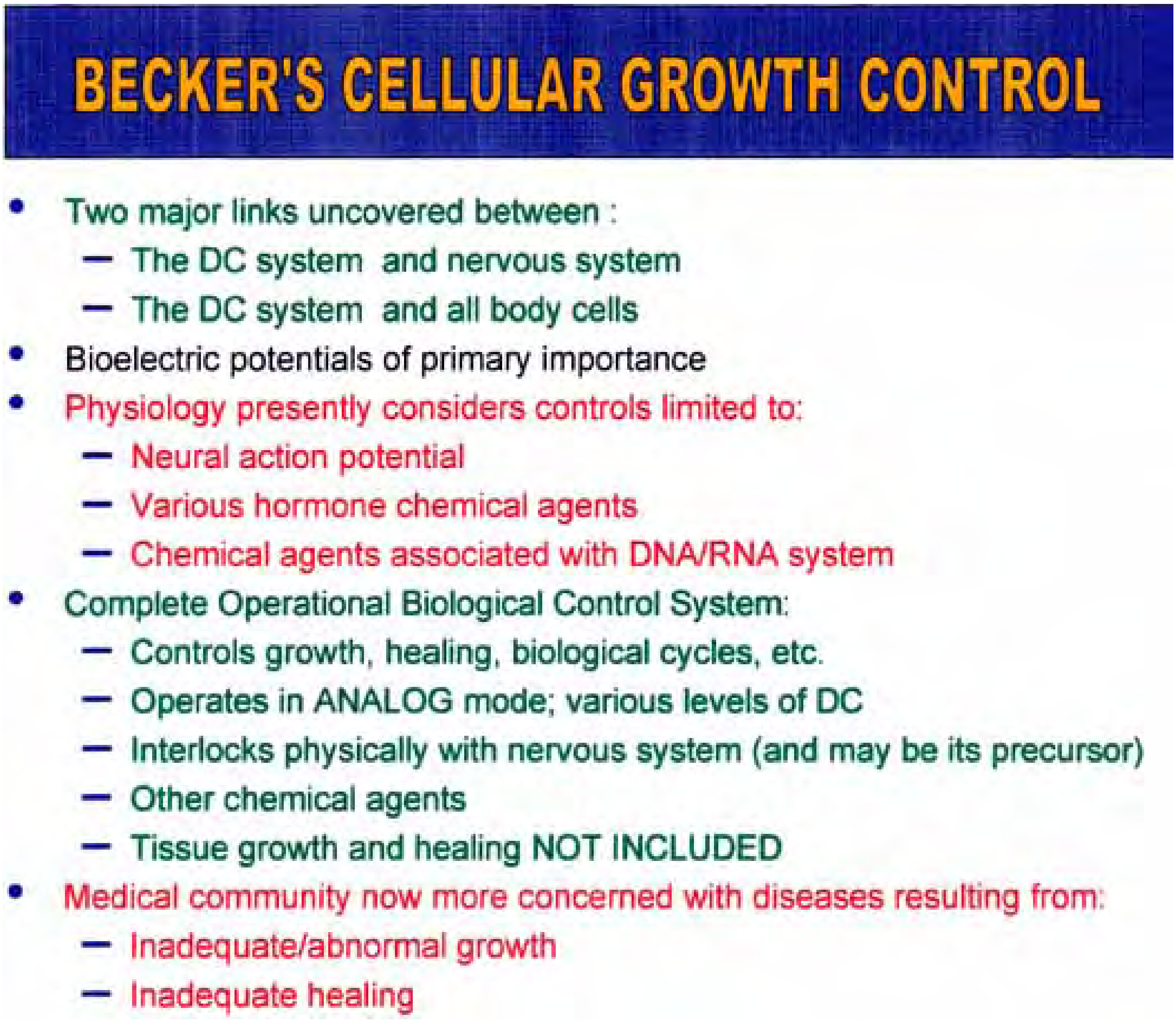

BECKER'S CELLULAR GROWTH CONTROL

- Two major links uncovered between :
  - The DC system and nervous system
  - The DC system and all body cells
- Bioelectric potentials of primary importance
- Physiology presently considers controls limited to:
  - Neural action potential
  - Various hormone chemical agents
  - Chemical agents associated with DNA/RNA system
- Complete Operational Biological Control System:
  - Controls growth, healing, biological cycles, etc.
  - Operates in ANALOG mode; various levels of DC
  - Interlocks physically with nervous system (and may be its precursor)
  - Other chemical agents
  - Tissue growth and healing NOT INCLUDED
- Medical community now more concerned with diseases resulting from:
  - Inadequate/abnormal growth
  - Inadequate healing

Return to Slide Index

Next Slide

Previous Slide

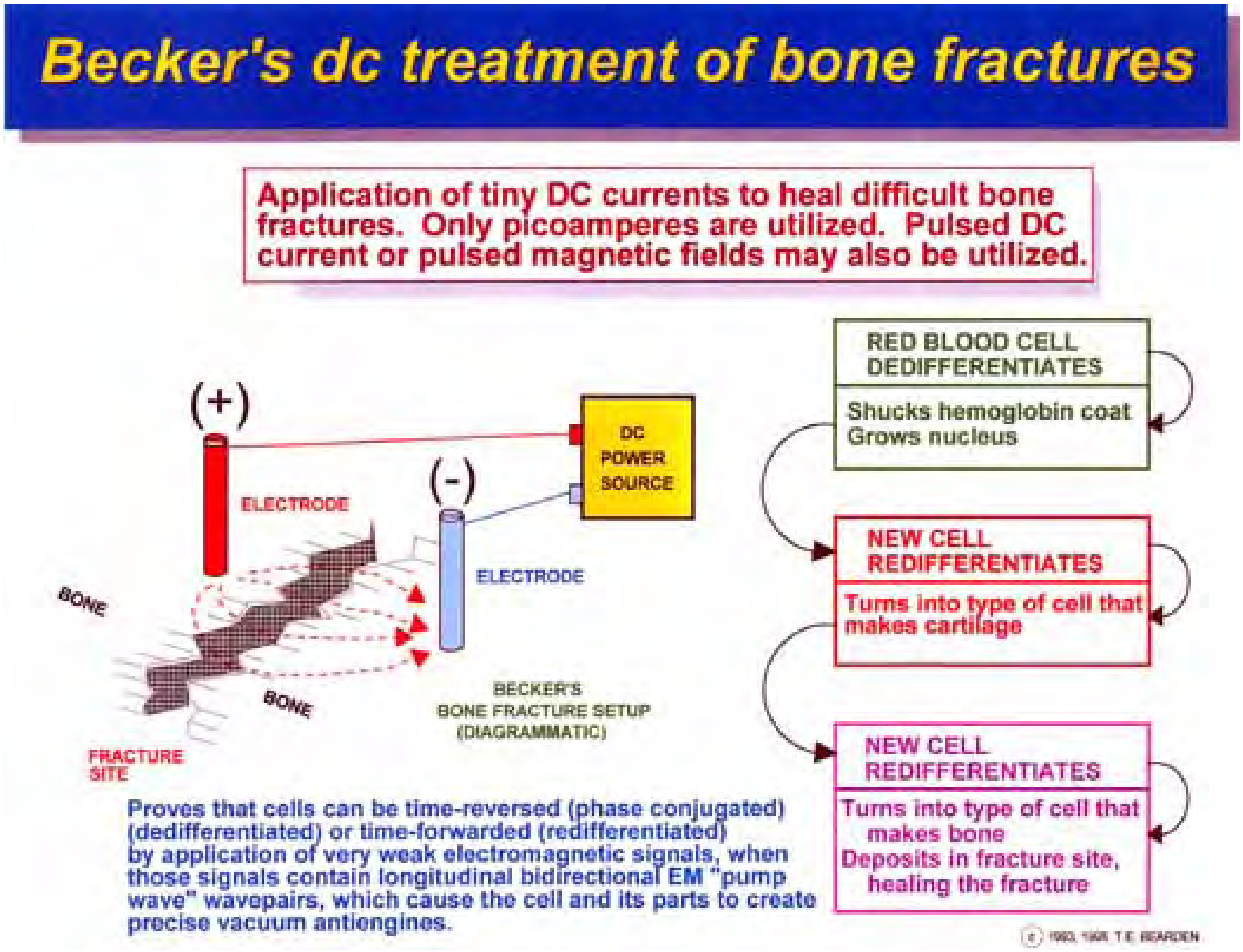

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

© 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

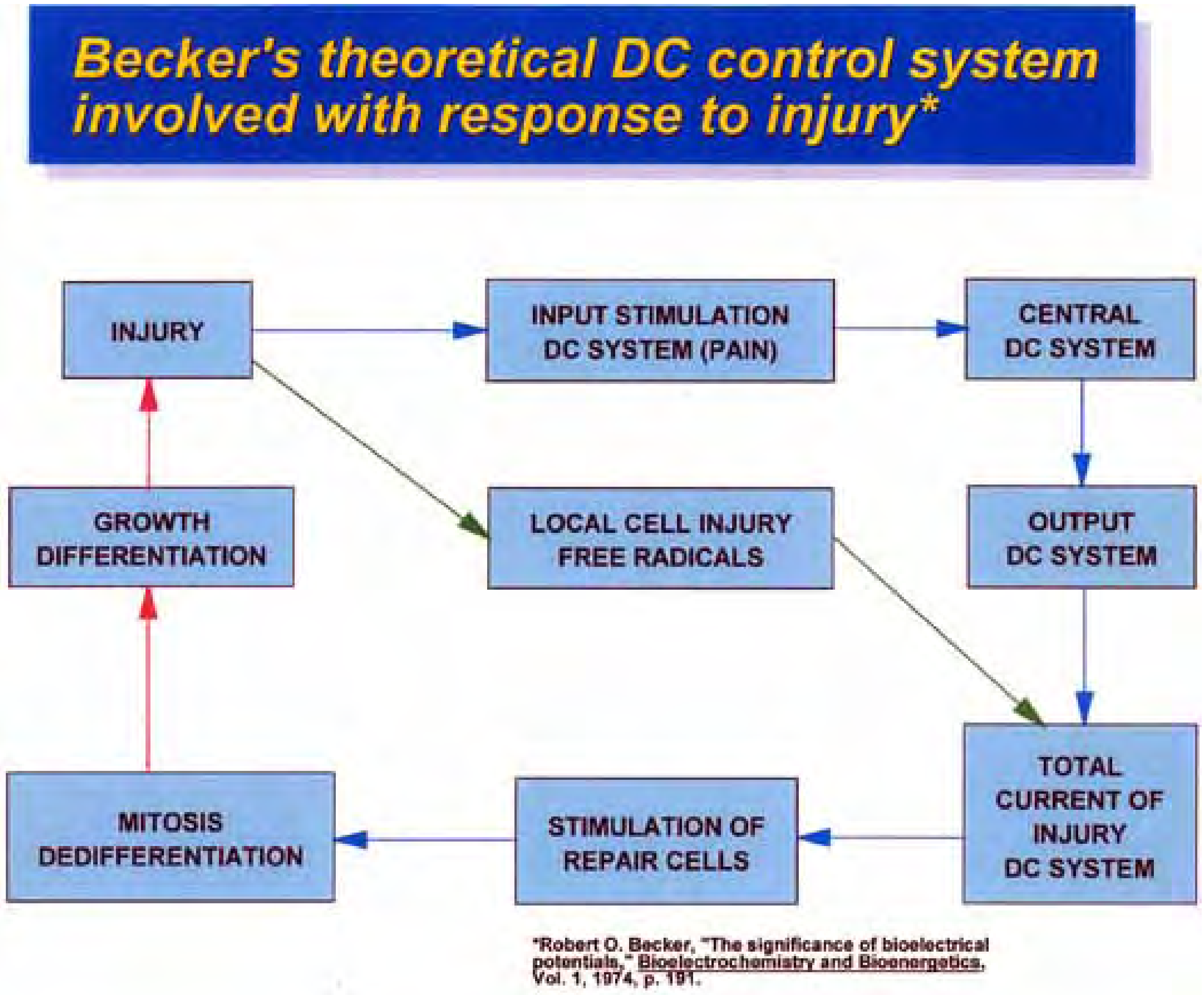

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

STIMULATION OF REPAIR CELLS

MITOSIS DEDIFFERENTIATION

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Return to Slide Index

Next Slide

Previous Slide

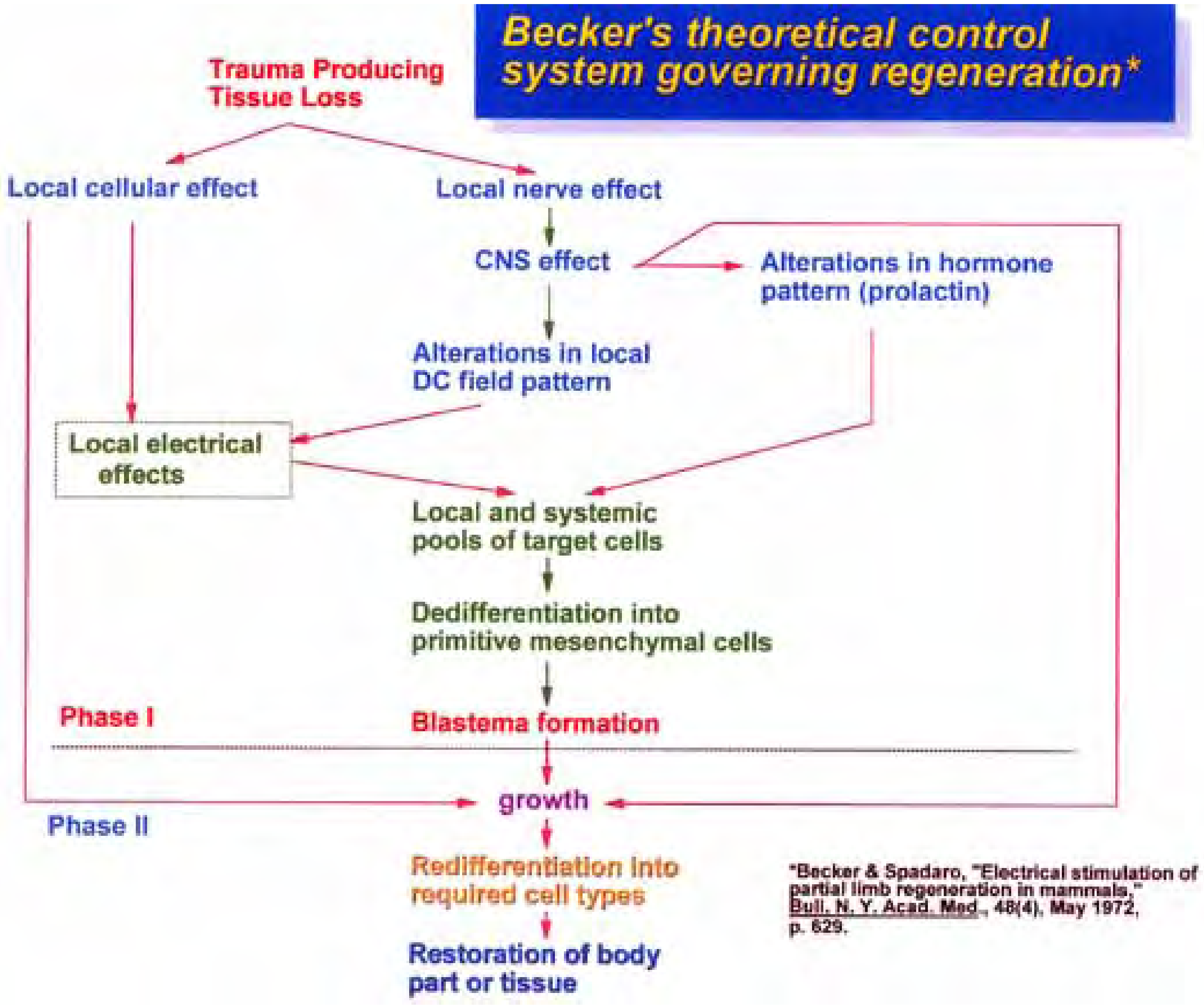

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Return to Slide Index

Next Slide

Previous Slide

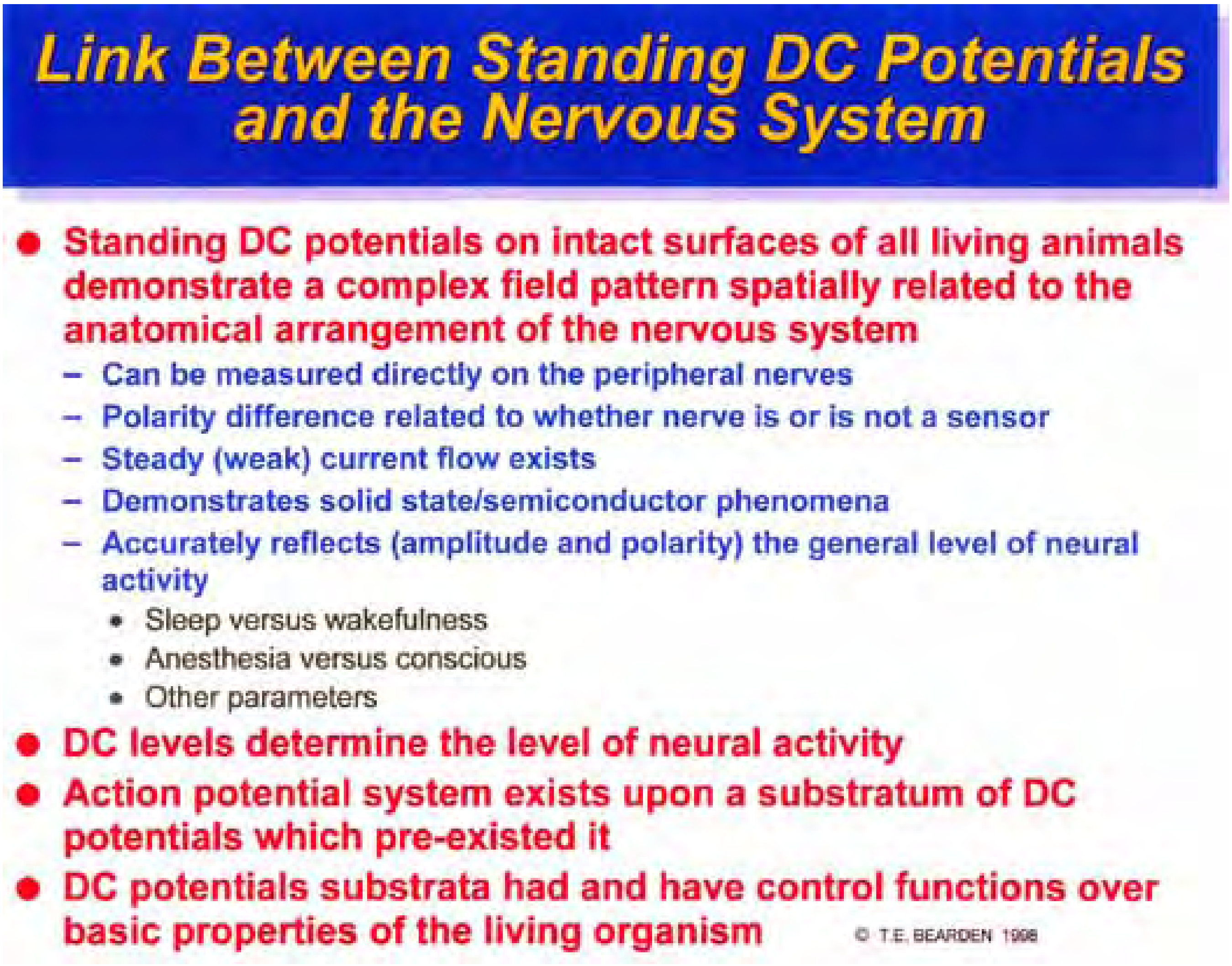

Link Between Standing DC Potentials and the Nervous System

- Standing DC potentials on intact surfaces of all living animals demonstrate a complex field pattern spatially related to the anatomical arrangement of the nervous system
  - Can be measured directly on the peripheral nerves
  - Polarity difference related to whether nerve is or is not a sensor
  - Steady (weak) current flow exists
  - Demonstrates solid state/semiconductor phenomena
  - Accurately reflects (amplitude and polarity) the general level of neural activity
    - Sleep versus wakefulness
    - Anesthesia versus conscious
    - Other parameters
- DC levels determine the level of neural activity
- Action potential system exists upon a substratum of DC potentials which pre-existed it
- DC potentials substrata had and have control functions over basic properties of the living organism

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

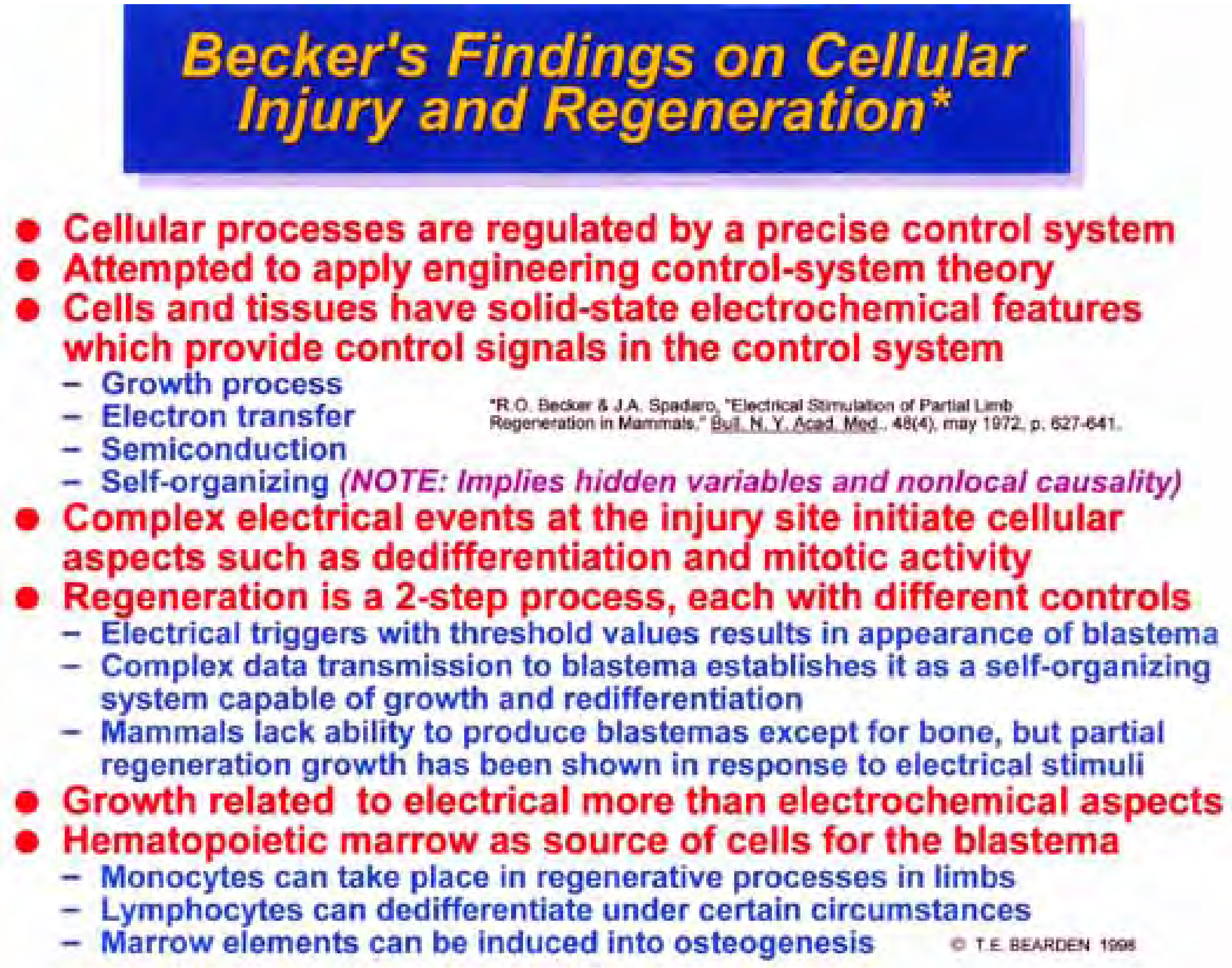

Becker's Findings on Cellular Injury and Regeneration*

- Cellular processes are regulated by a precise control system
- Attempted to apply engineering control-system theory
- Cells and tissues have solid-state electrochemical features which provide control signals in the control system
  - Growth process
  - Electron transfer
  - Semiconduction
  - Self-organizing *(NOTE: Implies hidden variables and nonlocal causality)*
- Complex electrical events at the injury site initiate cellular aspects such as dedifferentiation and mitotic activity
- Regeneration is a 2-step process, each with different controls
  - Electrical triggers with threshold values results in appearance of blastema
  - Complex data transmission to blastema establishes it as a self-organizing system capable of growth and redifferentiation
  - Mammals lack ability to produce blastemas except for bone, but partial regeneration growth has been shown in response to electrical stimuli
- Growth related to electrical more than electrochemical aspects
- Hematopoietic marrow as source of cells for the blastema
  - Monocytes can take place in regenerative processes in limbs
  - Lymphocytes can dedifferentiate under certain circumstances
  - Marrow elements can be induced into osteogenesis

*R.O. Becker & J.A. Spadaro, "Electrical Stimulation of Partial Limb Regeneration in Mammals," Bull. N. Y. Acad. Med., 48(4), may 1972, p. 627-641.

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

Corrections For the Present EM Bioeffects Model

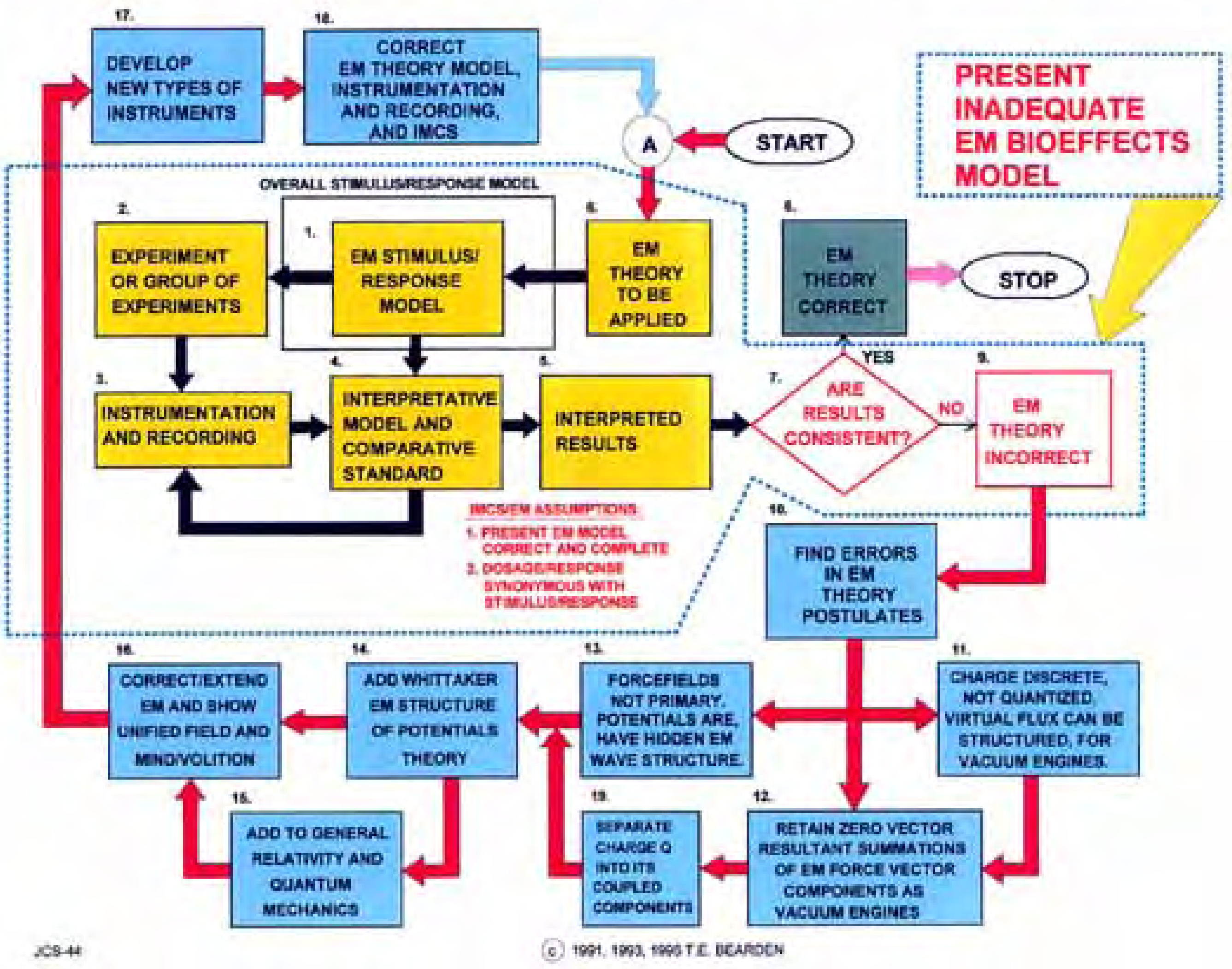

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

A

START

PRESENT INADEQUATE EM BIOEFFECTS MODEL

OVERALL STIMULUS/RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

1. EM STIMULUS/ RESPONSE MODEL

6. EM THEORY TO BE APPLIED

8. EM THEORY CORRECT

STOP

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

7. ARE RESULTS CONSISTENT?

YES

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

15. ADD TO GENERAL RELATIVITY AND QUANTUM MECHANICS

19. SEPARATE CHARGE Q INTO ITS COUPLED COMPONENTS

12. RETAIN ZERO VECTOR RESULTANT SUMMATIONS OF EM FORCE VECTOR COMPONENTS AS VACUUM ENGINES

JCS-44

© 1991, 1993, 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

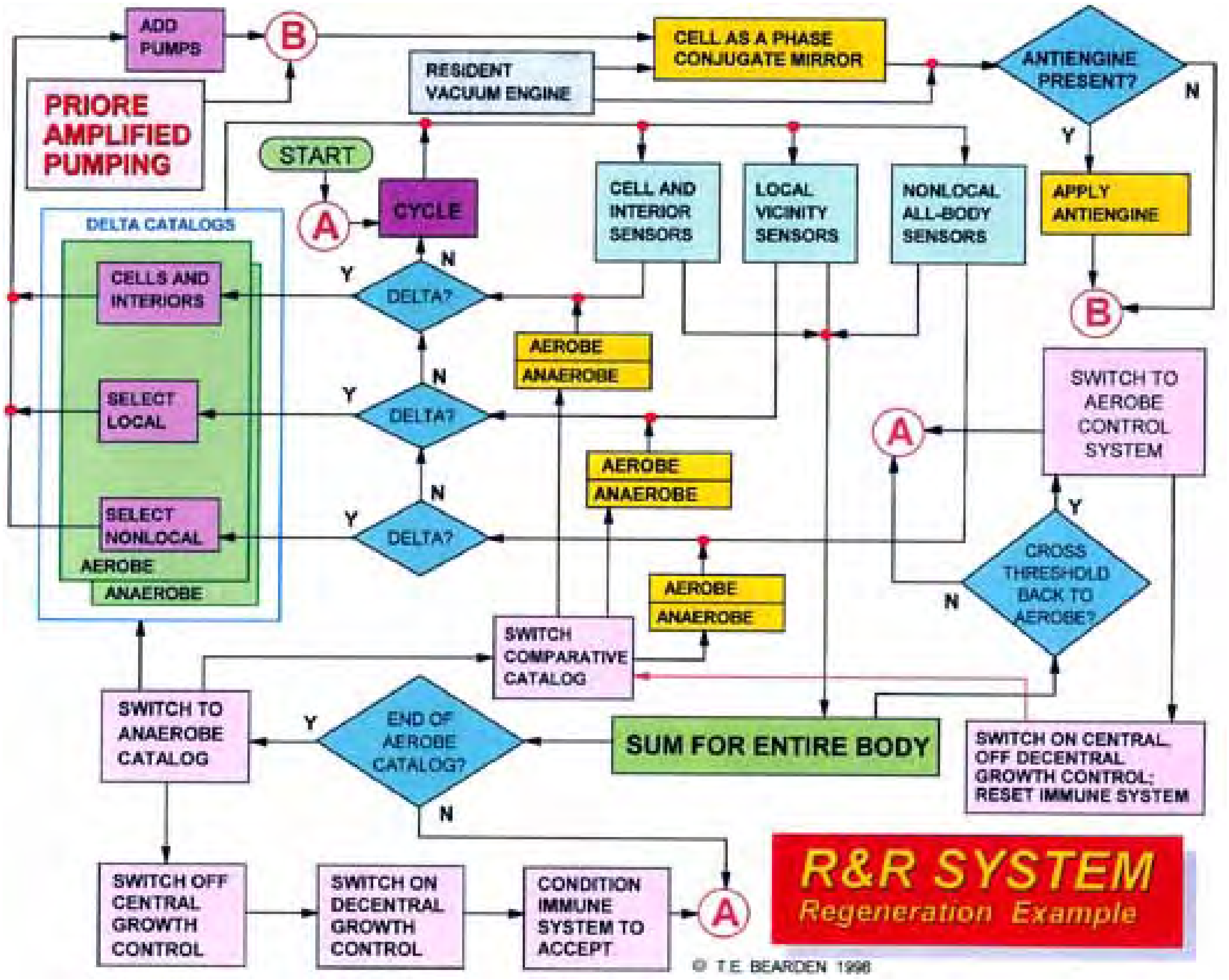

ADD PUMPS

B

CELL AS A PHASE CONJUGATE MIRROR

ANTIENGINE PRESENT?

N

RESIDENT VACUUM ENGINE

PRIORE AMPLIFIED PUMPING

START

Y

APPLY ANTIENGINE

A

CYCLE

CELL AND INTERIOR SENSORS

LOCAL VICINITY SENSORS

NONLOCAL ALL-BODY SENSORS

DELTA CATALOGS

CELLS AND INTERIORS

Y

N

DELTA?

B

AEROBE

ANAEROBE

SELECT LOCAL

Y

N

DELTA?

SWITCH TO AEROBE CONTROL SYSTEM

A

AEROBE

ANAEROBE

Y

N

SELECT NONLOCAL

Y

DELTA?

AEROBE

ANAEROBE

CROSS THRESHOLD BACK TO AEROBE?

N

AEROBE

ANAEROBE

SWITCH COMPARATIVE CATALOG

SWITCH TO ANAEROBE CATALOG

Y

END OF AEROBE CATALOG?

SUM FOR ENTIRE BODY

SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM

N

SWITCH OFF CENTRAL GROWTH CONTROL

SWITCH ON DECENTRAL GROWTH CONTROL

CONDITION IMMUNE SYSTEM TO ACCEPT

A

R&R SYSTEM

Regeneration Example

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

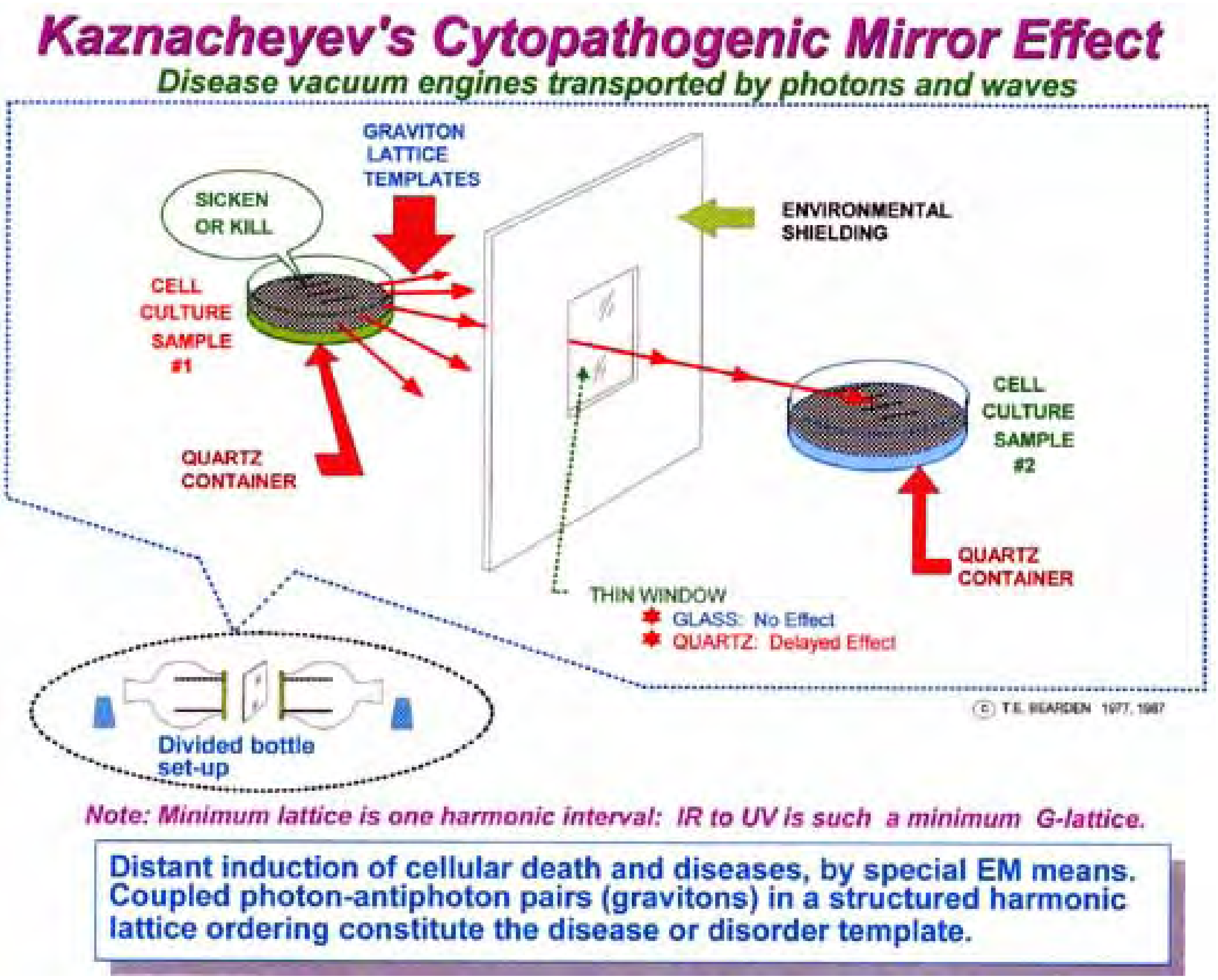

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Return to Slide Index

Next Slide

Previous Slide

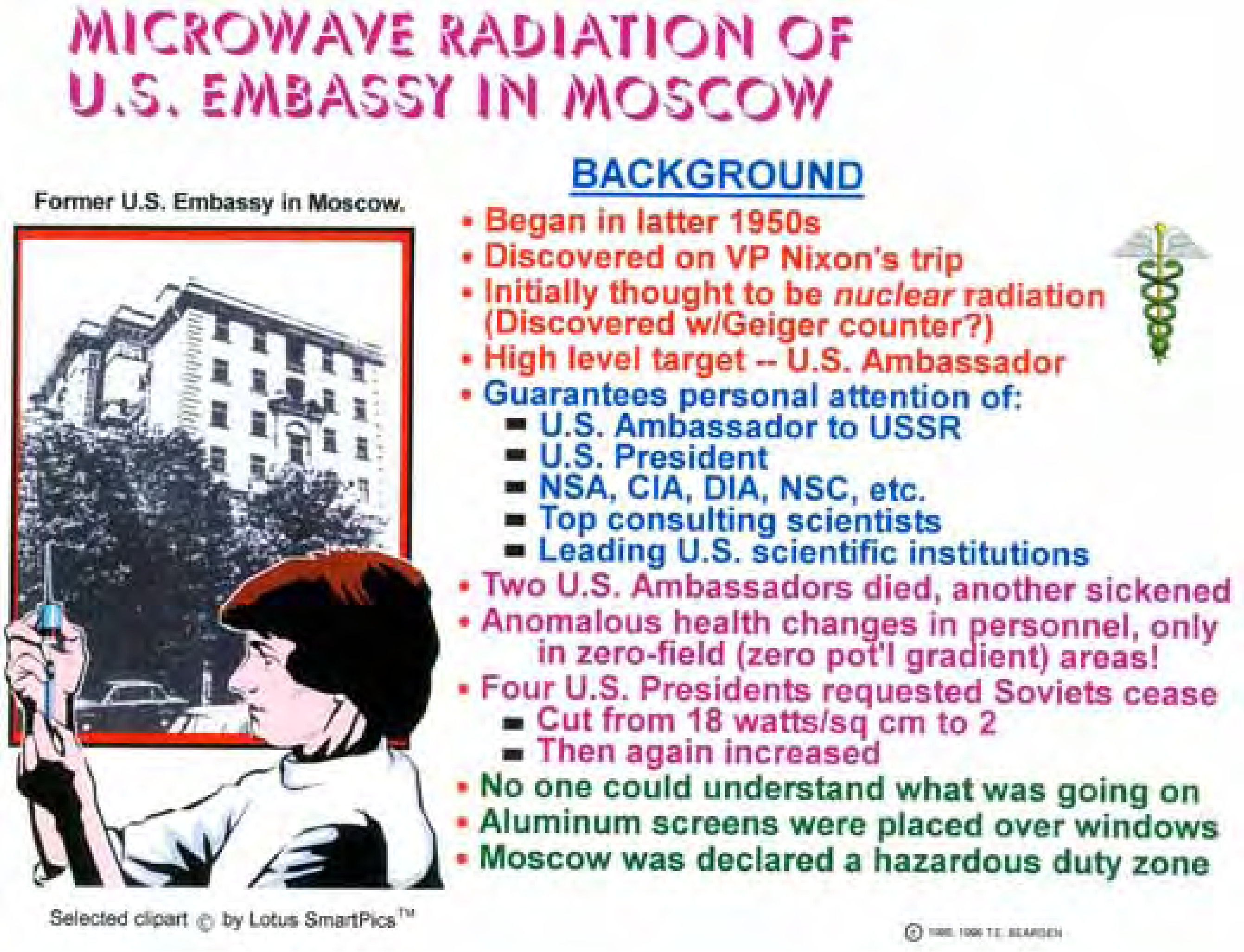

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© 1995, 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

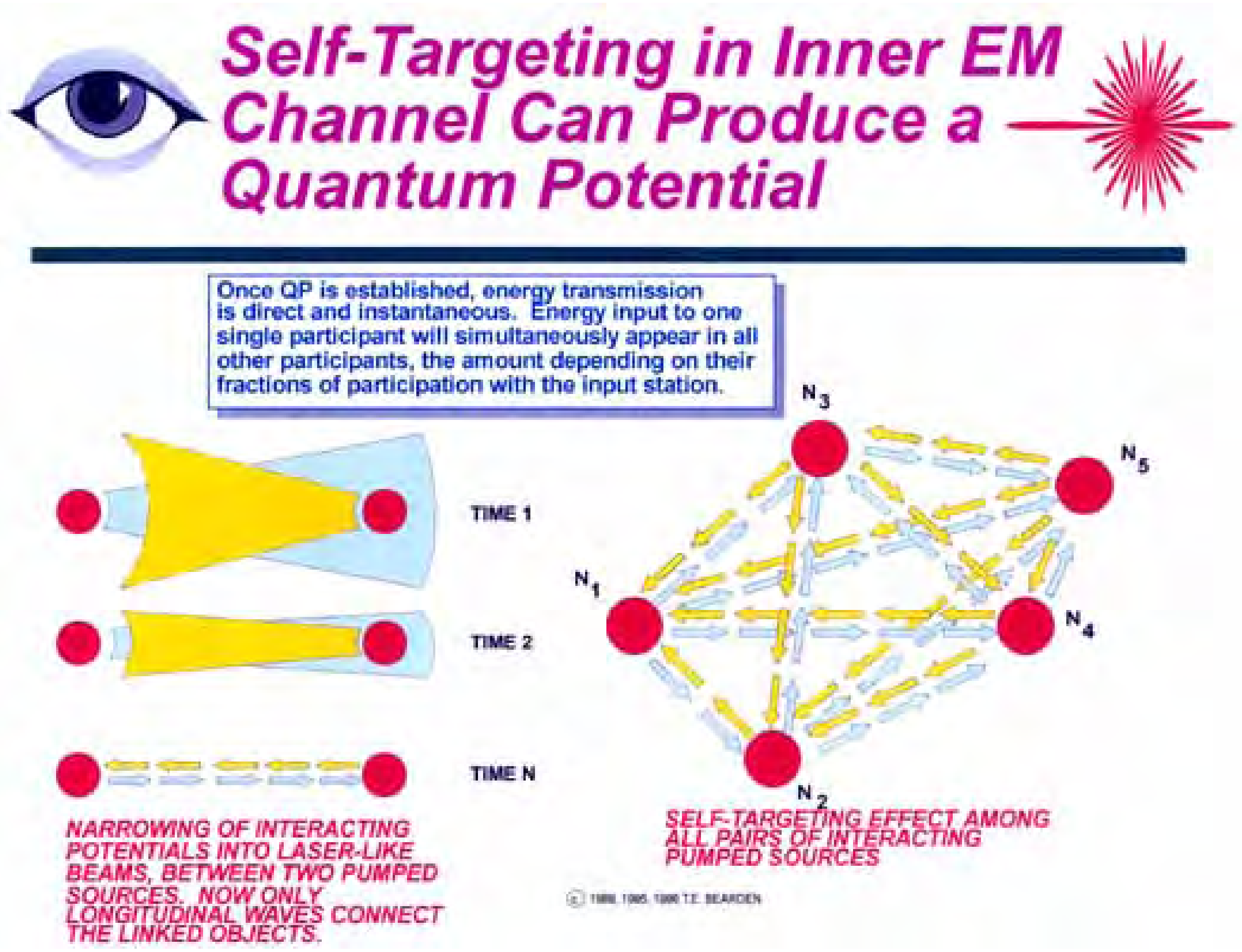

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1986, 1995, 1996 T.E. BEARDEN

[Return to Slide Index](#)

[Next Slide](#)

[Previous Slide](#)

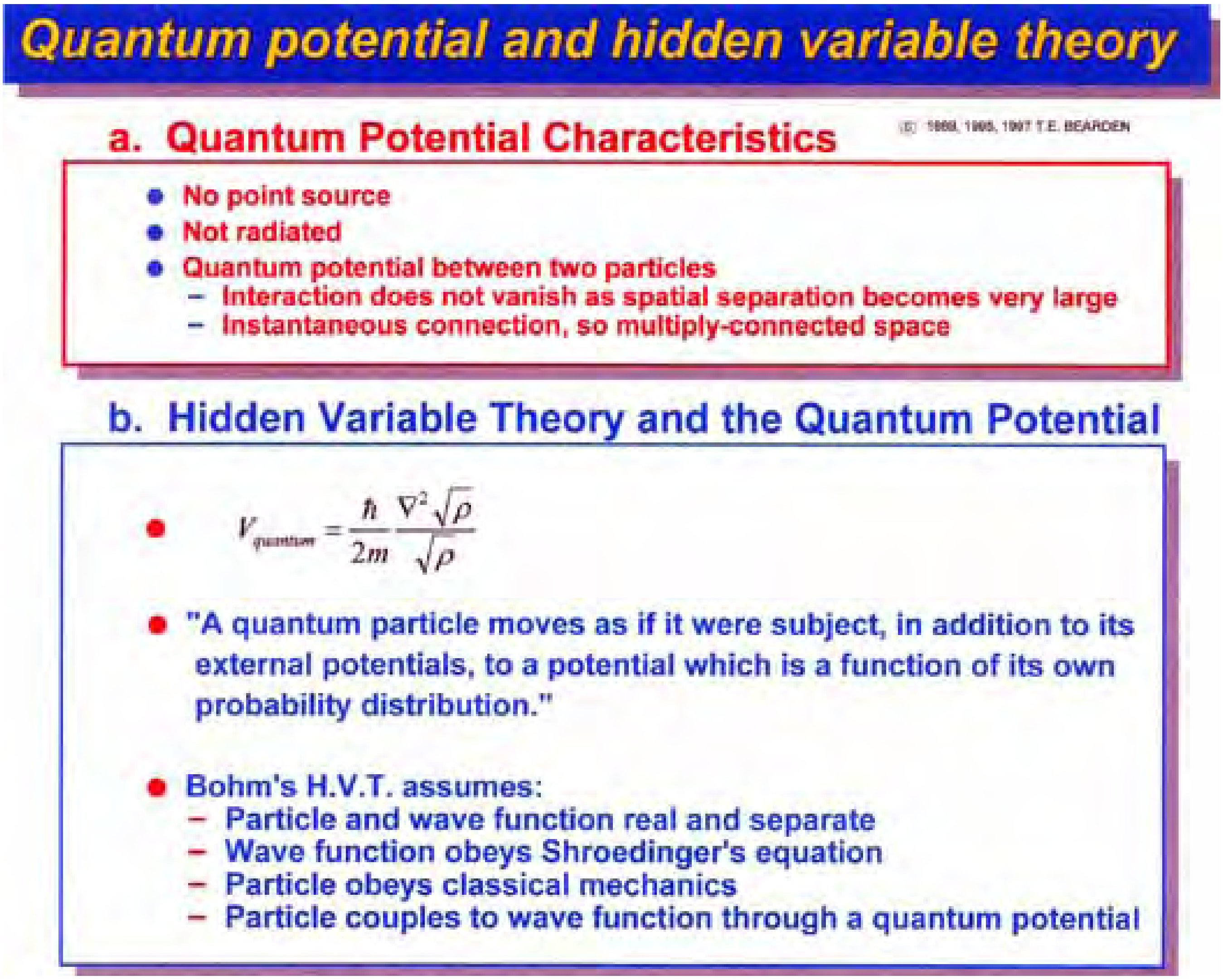

Quantum potential and hidden variable theory

a. Quantum Potential Characteristics

© 1988, 1995, 1997 T.E. BEARDEN

- No point source
- Not radiated
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection, so multiply-connected space

b. Hidden Variable Theory and the Quantum Potential

- $$V_{quantum} = \frac{\hbar}{2m} \frac{\nabla^2 \sqrt{\rho}}{\sqrt{\rho}}$$
- "A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution."
- Bohm's H.V.T. assumes:
  - Particle and wave function real and separate
  - Wave function obeys Shroedinger's equation
  - Particle obeys classical mechanics
  - Particle couples to wave function through a quantum potential

Return to Slide Index

Next Slide

Previous Slide

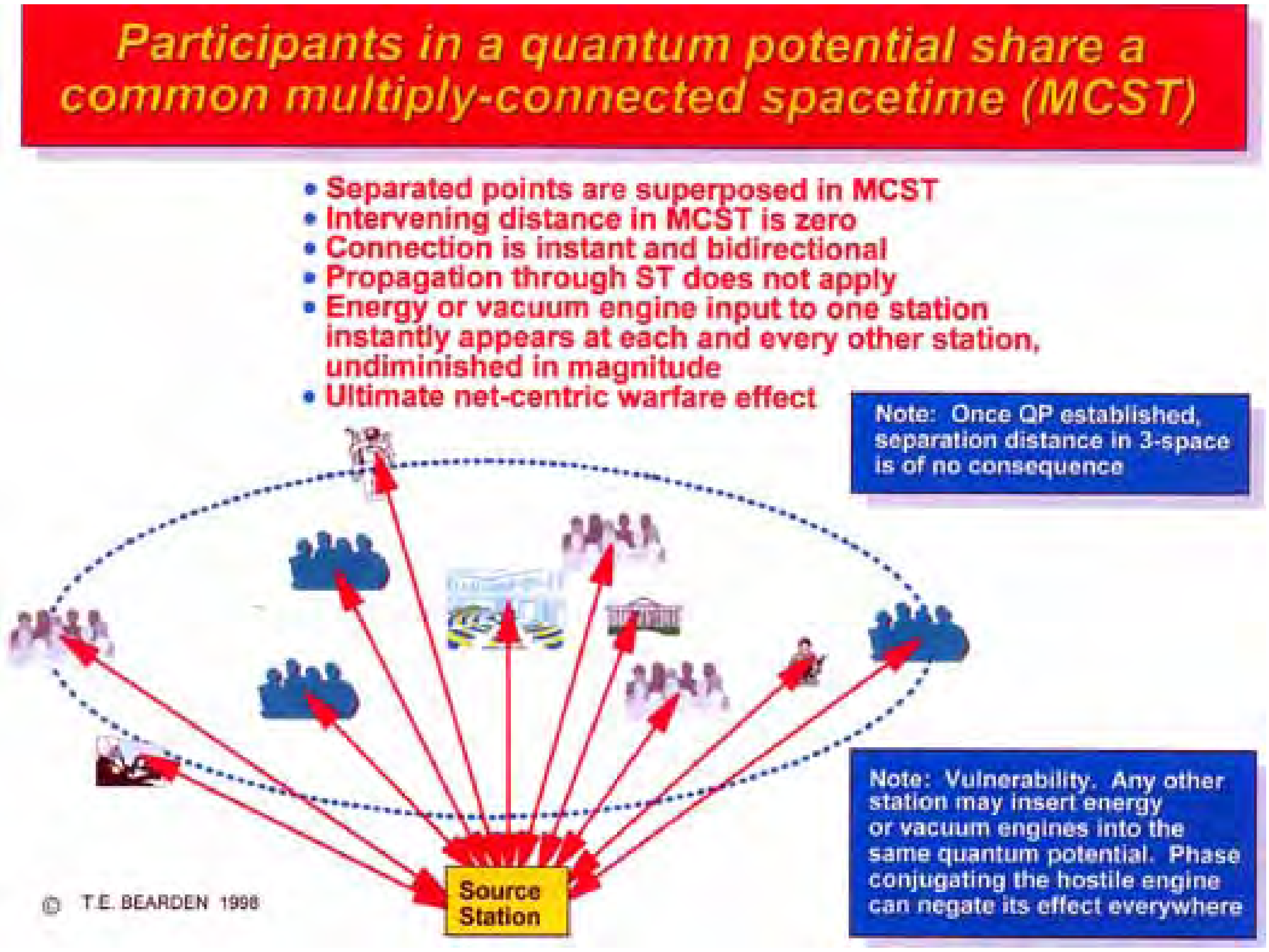

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Intervening distance in MCST is zero
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude
- Ultimate net-centric warfare effect

Note: Once QP established, separation distance in 3-space is of no consequence

Note: Vulnerability. Any other station may insert energy or vacuum engines into the same quantum potential. Phase conjugating the hostile engine can negate its effect everywhere

© T.E. BEARDEN 1998

[Return to Slide Index]
[Next Slide]
[Previous Slide]

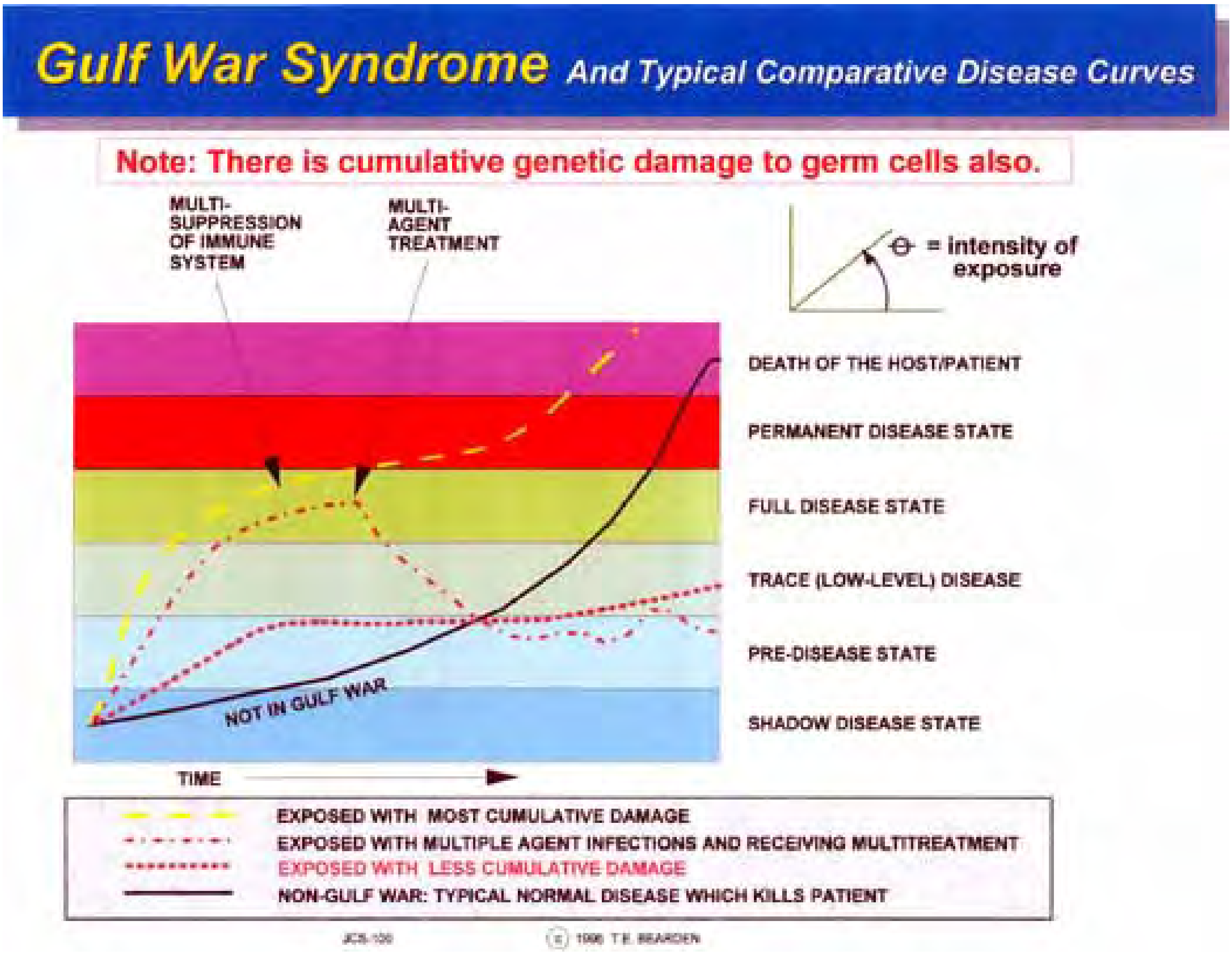

Gulf War Syndrome And Typical Comparative Disease Curves

Note: There is cumulative genetic damage to germ cells also.

- – – – EXPOSED WITH MOST CUMULATIVE DAMAGE
- -·-·-·- EXPOSED WITH MULTIPLE AGENT INFECTIONS AND RECEIVING MULTITREATMENT
- ·········· EXPOSED WITH LESS CUMULATIVE DAMAGE
- ——— NON-GULF WAR: TYPICAL NORMAL DISEASE WHICH KILLS PATIENT

JCS-100

(c) 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

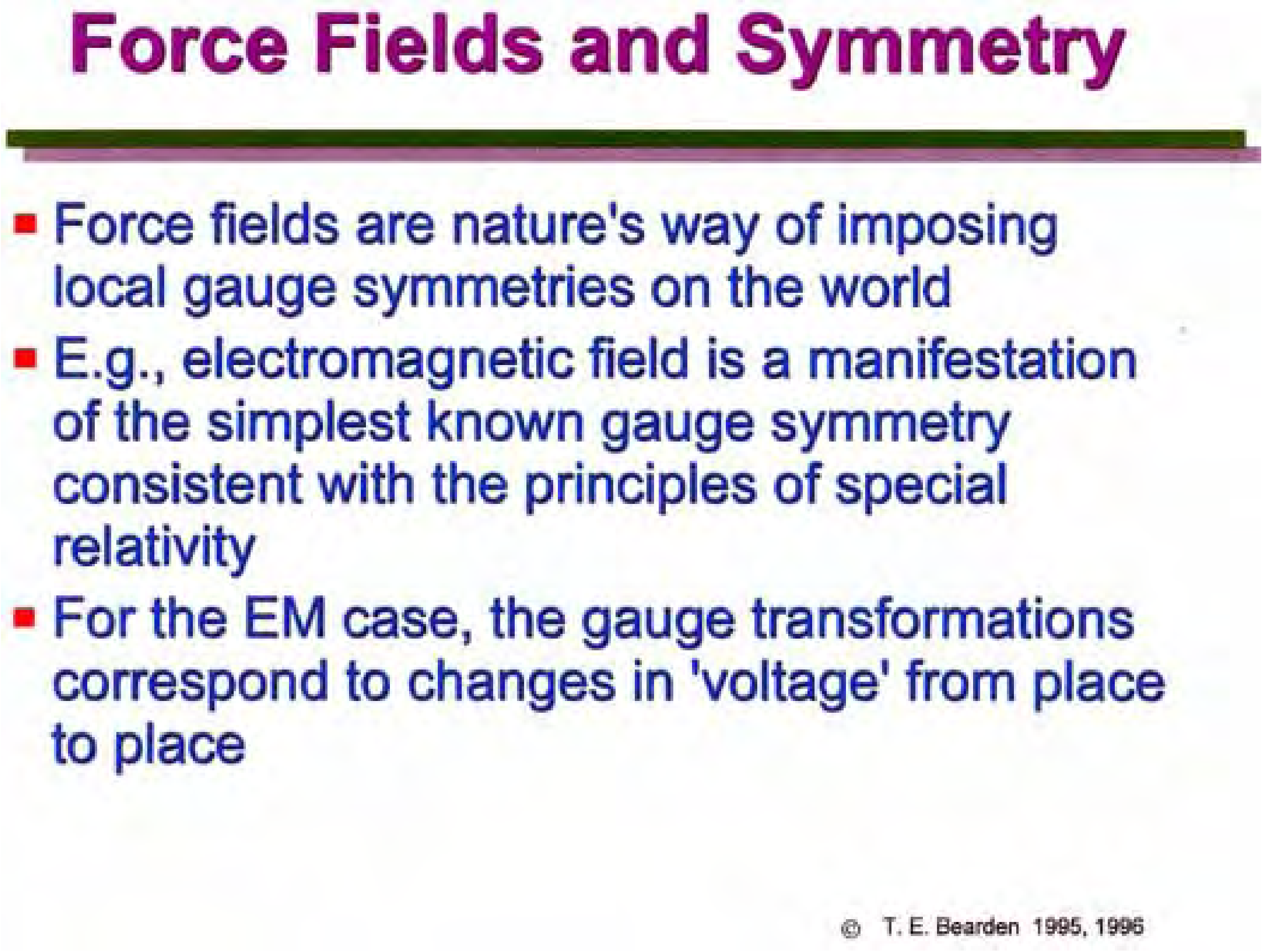

Force Fields and Symmetry

- Force fields are nature's way of imposing local gauge symmetries on the world
- E.g., electromagnetic field is a manifestation of the simplest known gauge symmetry consistent with the principles of special relativity
- For the EM case, the gauge transformations correspond to changes in 'voltage' from place to place

© T. E. Bearden 1995, 1996

Return to Slide Index

Next Slide

Previous Slide

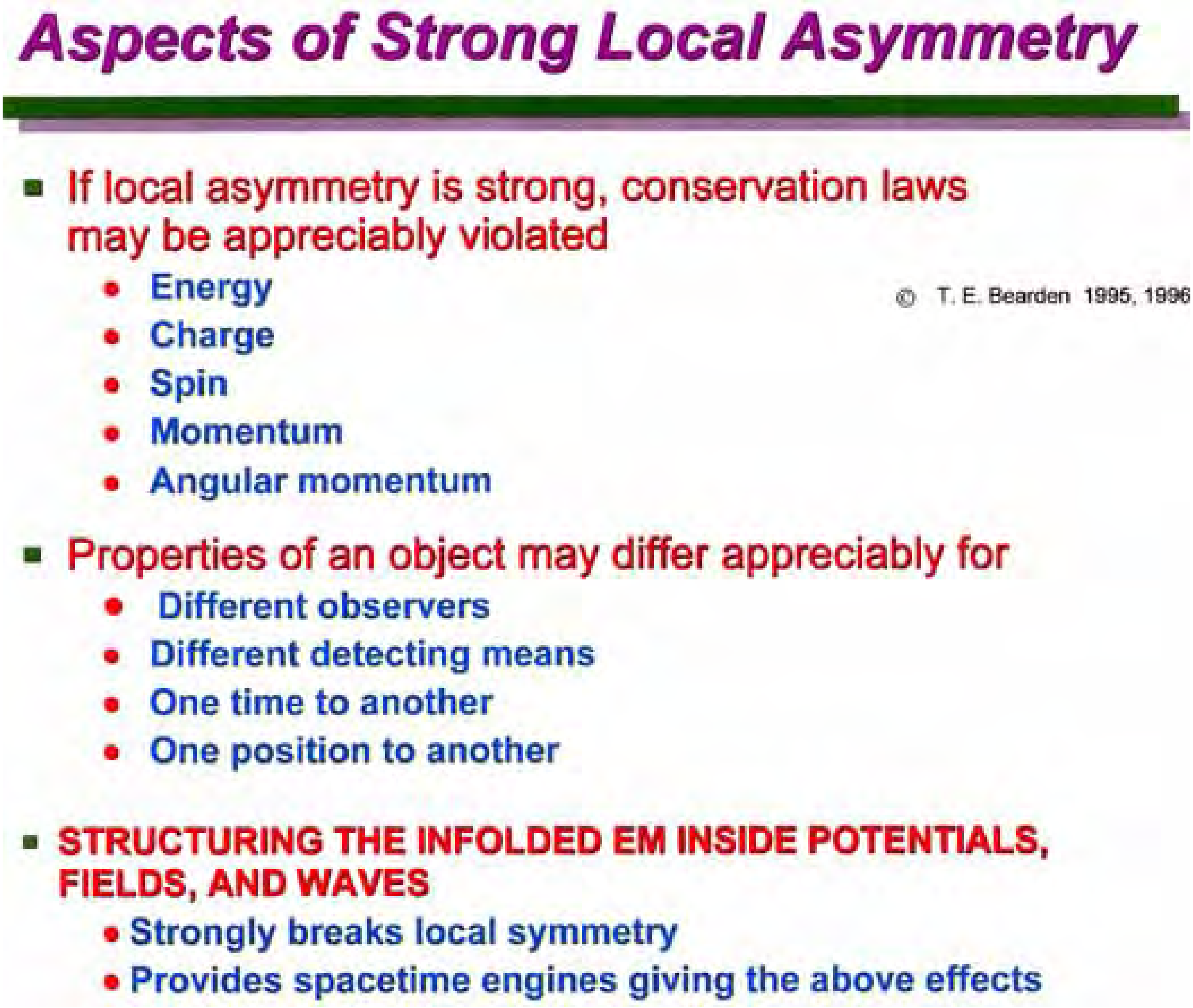

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

Return to Slide Index

Next Slide

Previous Slide

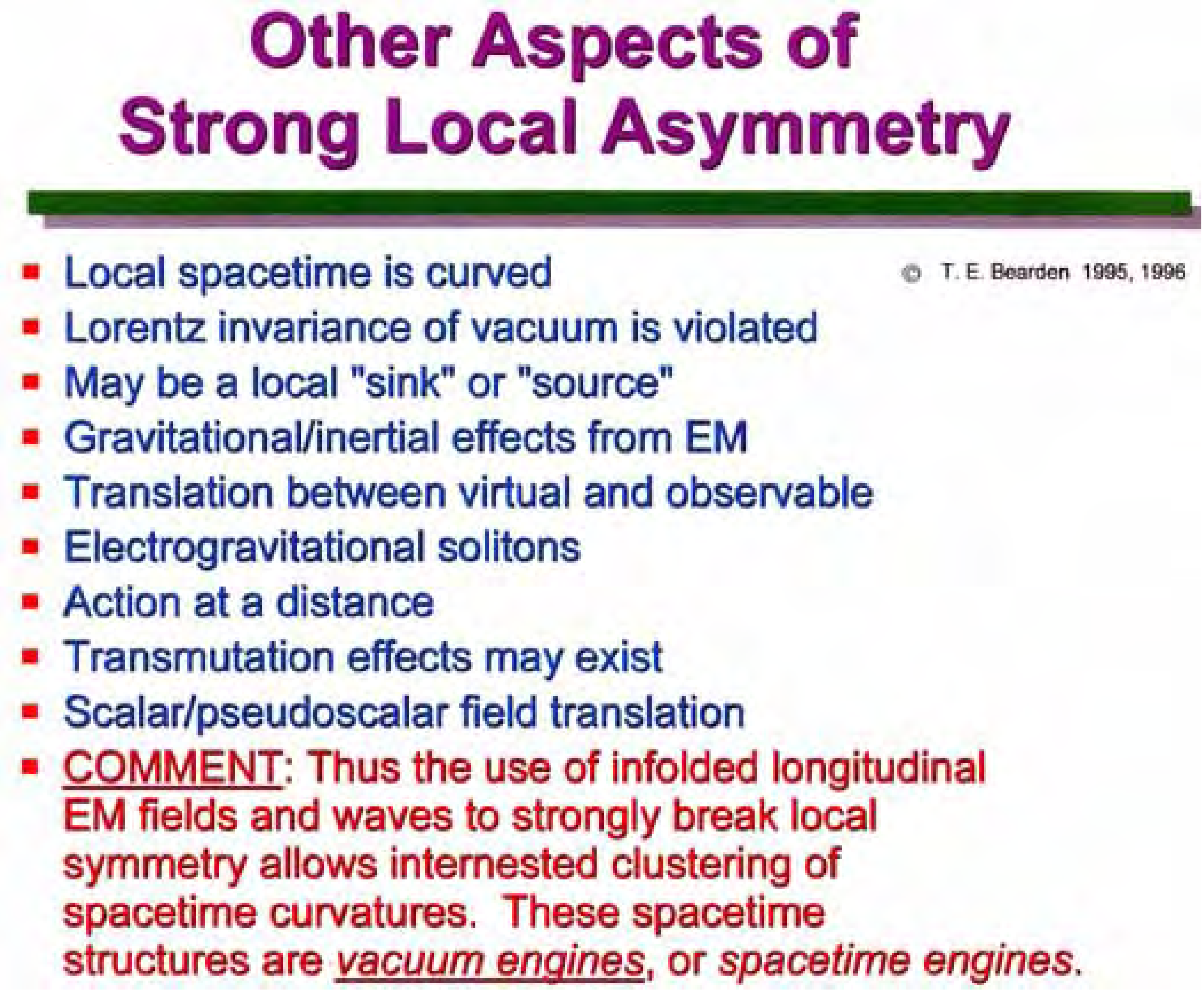

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational/inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

Return to Slide Index

Next Slide

Previous Slide

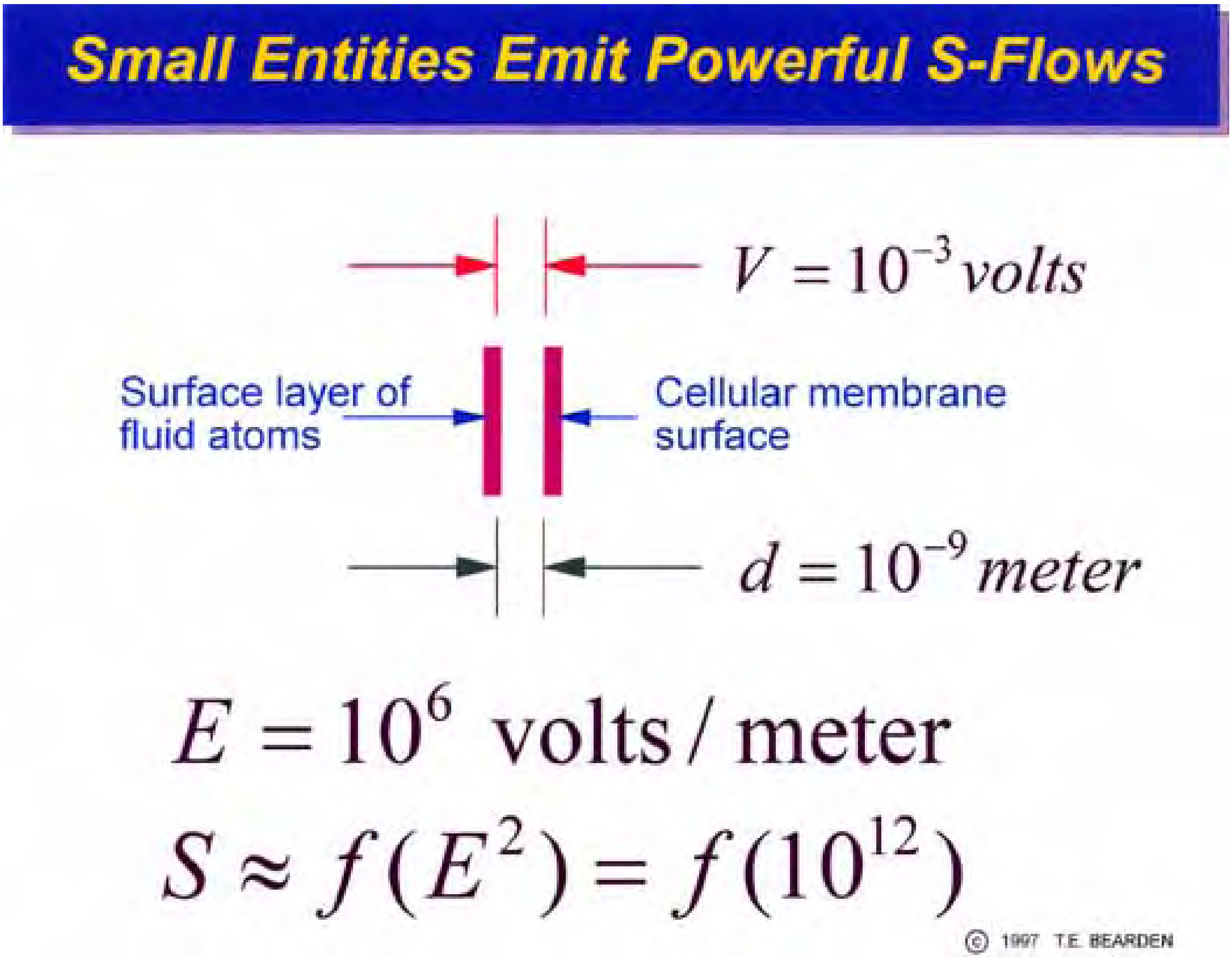

Small Entities Emit Powerful S-Flows

$V = 10^{-3}\,volts$

Surface layer of fluid atoms

Cellular membrane surface

$d = 10^{-9}\,meter$

$$E = 10^{6}\ \text{volts / meter}$$

$$S \approx f(E^{2}) = f(10^{12})$$

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

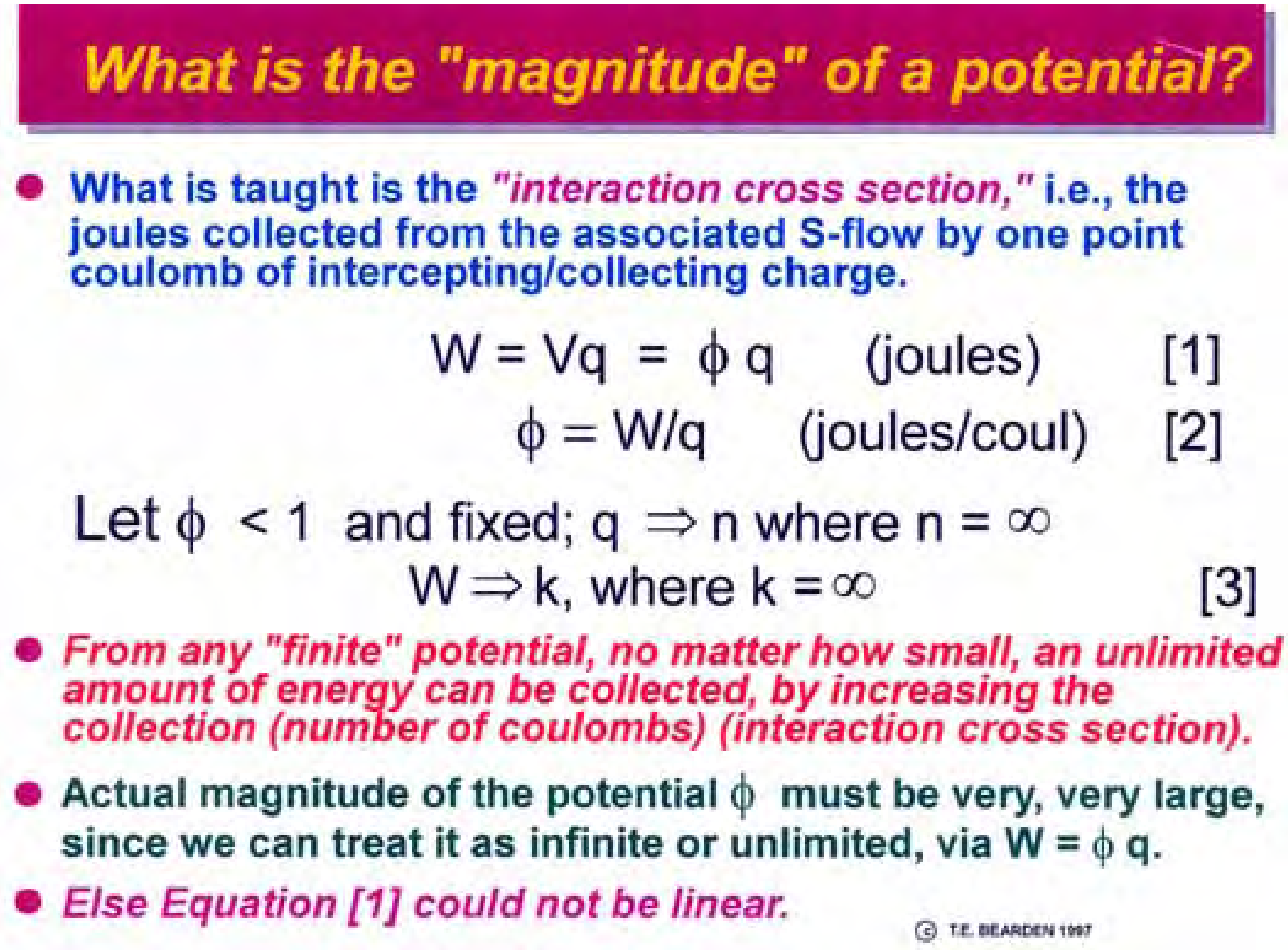

What is the "magnitude" of a potential?

- What is taught is the *"interaction cross section,"* i.e., the joules collected from the associated S-flow by one point coulomb of intercepting/collecting charge.

$$W = Vq = \phi q \quad \text{(joules)} \qquad [1]$$

$$\phi = W/q \quad \text{(joules/coul)} \qquad [2]$$

$$\text{Let } \phi < 1 \text{ and fixed; } q \Rightarrow n \text{ where } n = \infty$$

$$W \Rightarrow k, \text{ where } k = \infty \qquad [3]$$

- *From any "finite" potential, no matter how small, an unlimited amount of energy can be collected, by increasing the collection (number of coulombs) (interaction cross section).*
- **Actual magnitude of the potential ϕ must be very, very large, since we can treat it as infinite or unlimited, via $W = \phi q$.**
- *Else Equation [1] could not be linear.*

© T.E. BEARDEN 1997

Return to Slide Index

Next Slide

Previous Slide

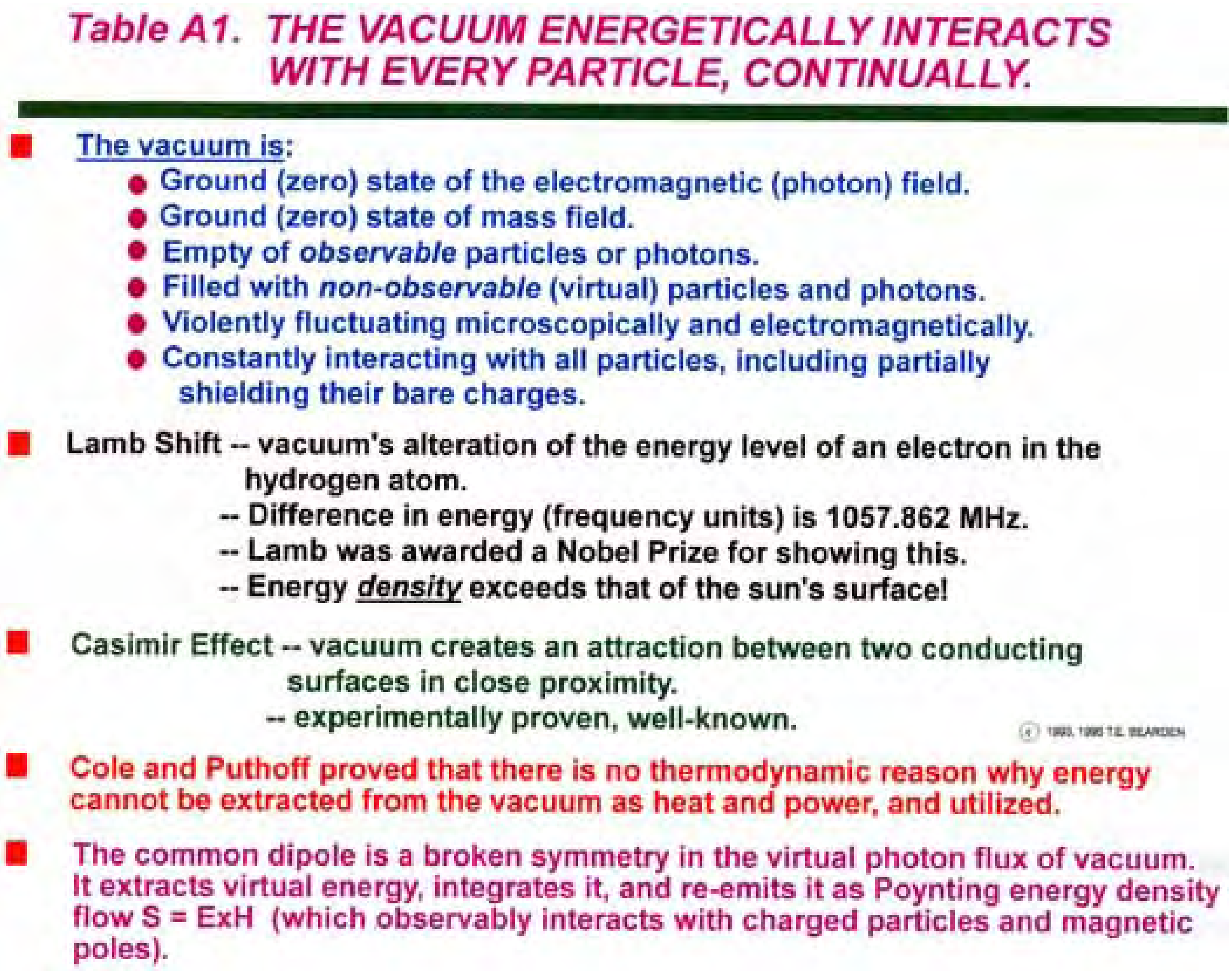

Table A1. THE VACUUM ENERGETICALLY INTERACTS WITH EVERY PARTICLE, CONTINUALLY.

- The vacuum is:
  - Ground (zero) state of the electromagnetic (photon) field.
  - Ground (zero) state of mass field.
  - Empty of *observable* particles or photons.
  - Filled with *non-observable* (virtual) particles and photons.
  - Violently fluctuating microscopically and electromagnetically.
  - Constantly interacting with all particles, including partially shielding their bare charges.
- Lamb Shift -- vacuum's alteration of the energy level of an electron in the hydrogen atom.
  - -- Difference in energy (frequency units) is 1057.862 MHz.
  - -- Lamb was awarded a Nobel Prize for showing this.
  - -- Energy ***density*** exceeds that of the sun's surface!
- Casimir Effect -- vacuum creates an attraction between two conducting surfaces in close proximity.
  - -- experimentally proven, well-known.

© 1993, 1995 T.E. BEARDEN

- Cole and Puthoff proved that there is no thermodynamic reason why energy cannot be extracted from the vacuum as heat and power, and utilized.
- The common dipole is a broken symmetry in the virtual photon flux of vacuum. It extracts virtual energy, integrates it, and re-emits it as Poynting energy density flow S = ExH (which observably interacts with charged particles and magnetic poles).

Return to Slide Index

Next Slide

Previous Slide

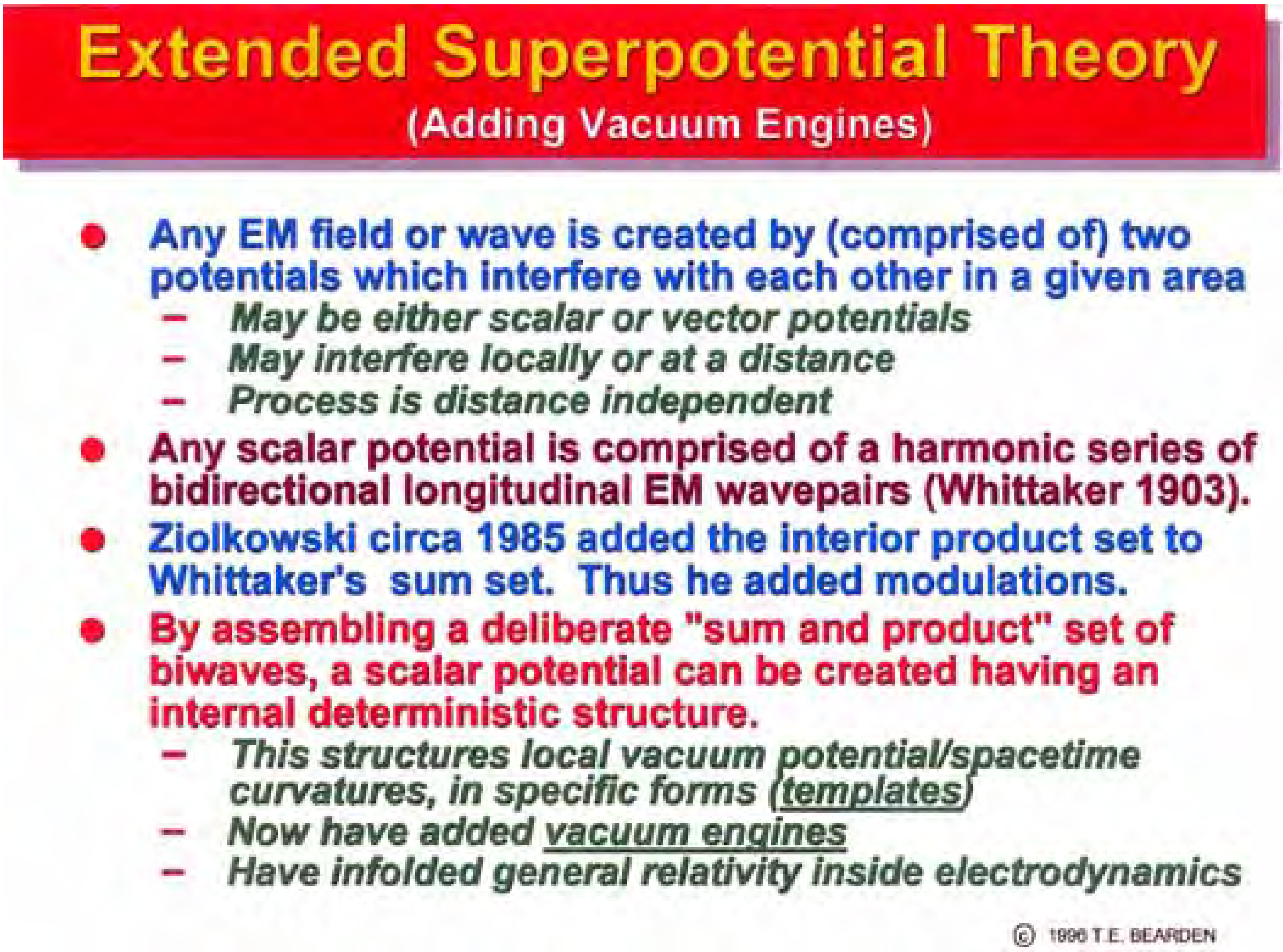

Extended Superpotential Theory

(Adding Vacuum Engines)

- Any EM field or wave is created by (comprised of) two potentials which interfere with each other in a given area
  - *May be either scalar or vector potentials*
  - *May interfere locally or at a distance*
  - *Process is distance independent*
- Any scalar potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs (Whittaker 1903).
- Ziolkowski circa 1985 added the interior product set to Whittaker's sum set. Thus he added modulations.
- By assembling a deliberate "sum and product" set of biwaves, a scalar potential can be created having an internal deterministic structure.
  - *This structures local vacuum potential/spacetime curvatures, in specific forms (templates)*
  - *Now have added vacuum engines*
  - *Have infolded general relativity inside electrodynamics*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

Vacuum Engine: Jillions of Working ST Demons

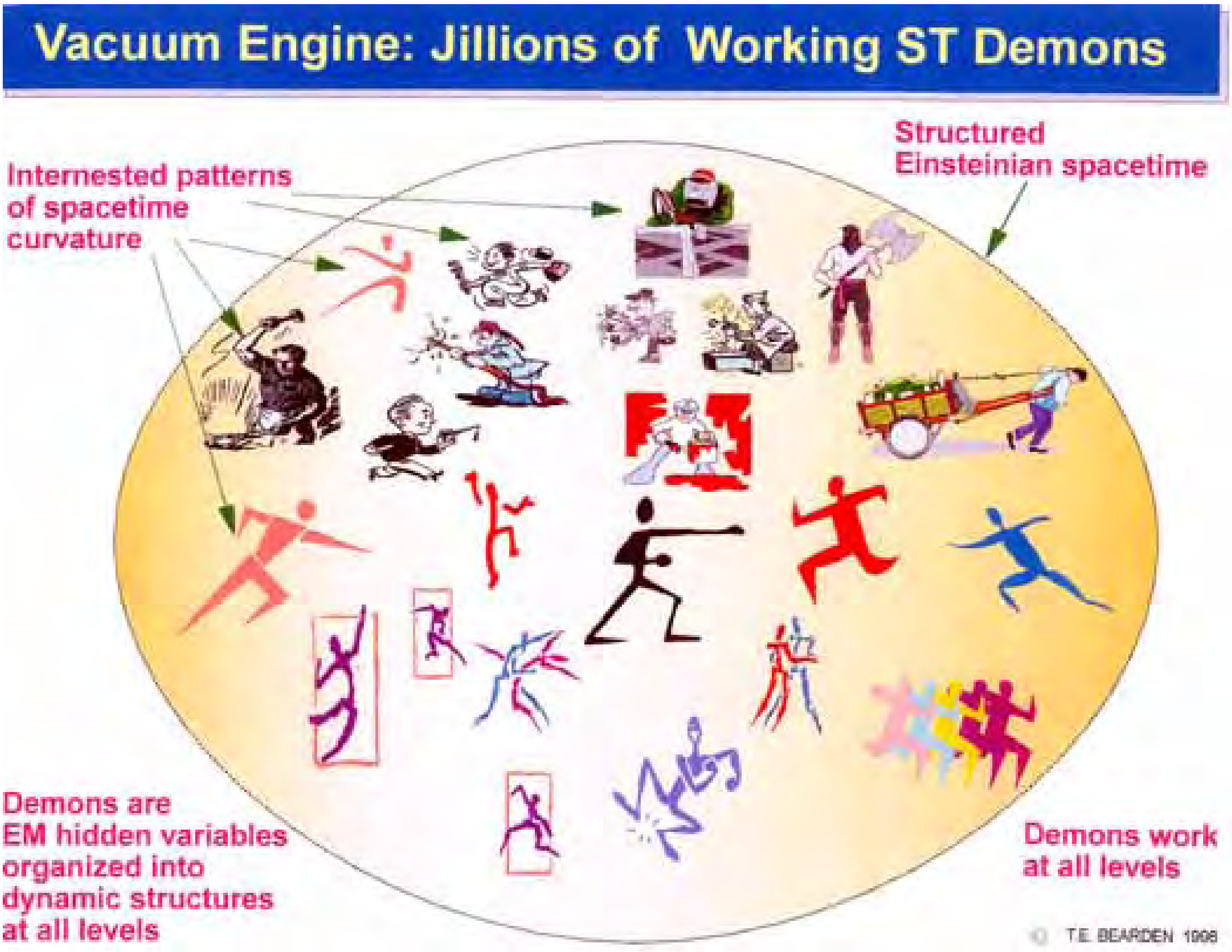

Structured Einsteinian spacetime

Internested patterns of spacetime curvature

Demons are EM hidden variables organized into dynamic structures at all levels

Demons work at all levels

© T.E. BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

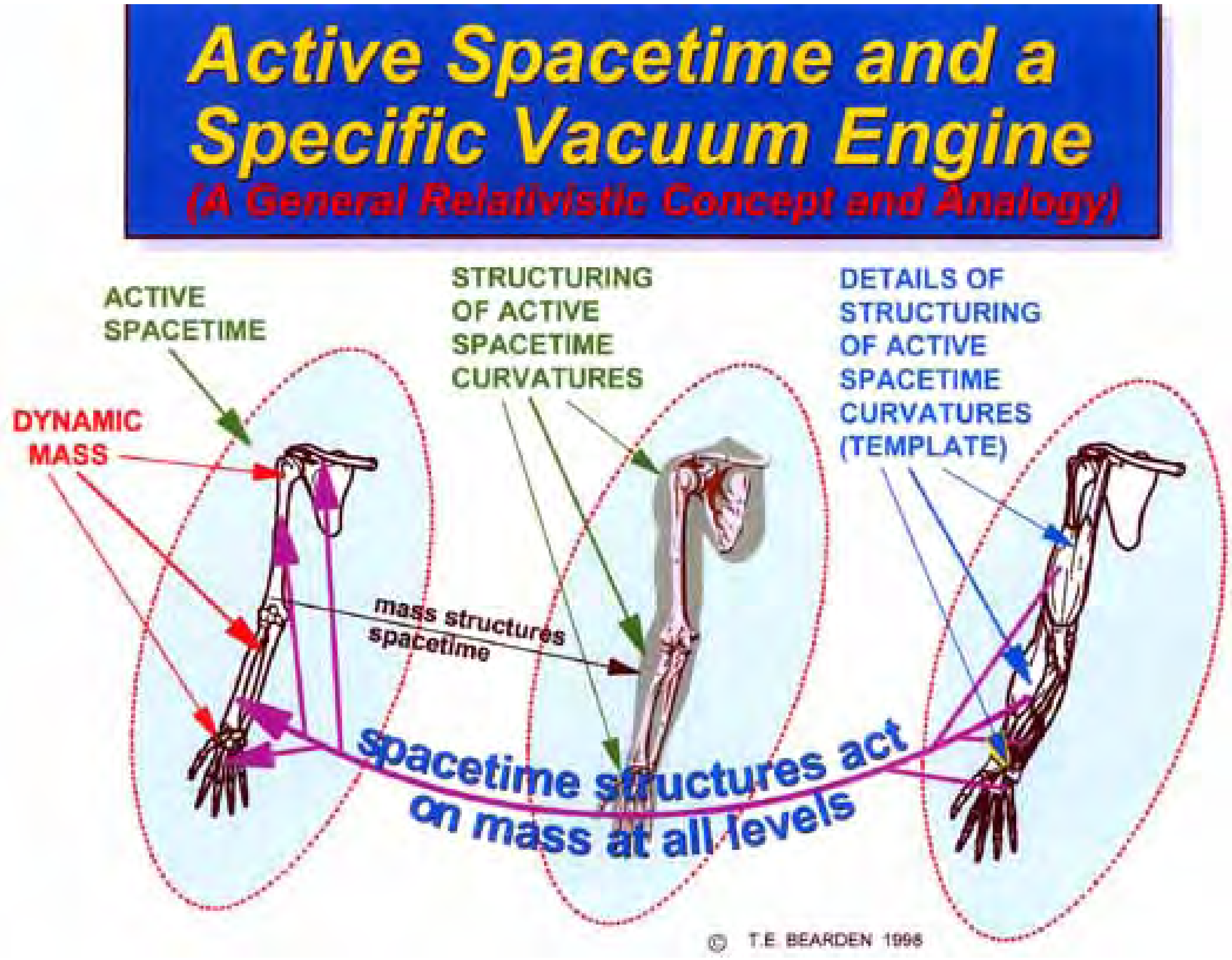

Active Spacetime and a Specific Vacuum Engine

(A General Relativistic Concept and Analogy)

ACTIVE SPACETIME

DYNAMIC MASS

STRUCTURING OF ACTIVE SPACETIME CURVATURES

DETAILS OF STRUCTURING OF ACTIVE SPACETIME CURVATURES (TEMPLATE)

mass structures spacetime

spacetime structures act on mass at all levels

© T.E BEARDEN 1998

Return to Slide Index

Next Slide

Previous Slide

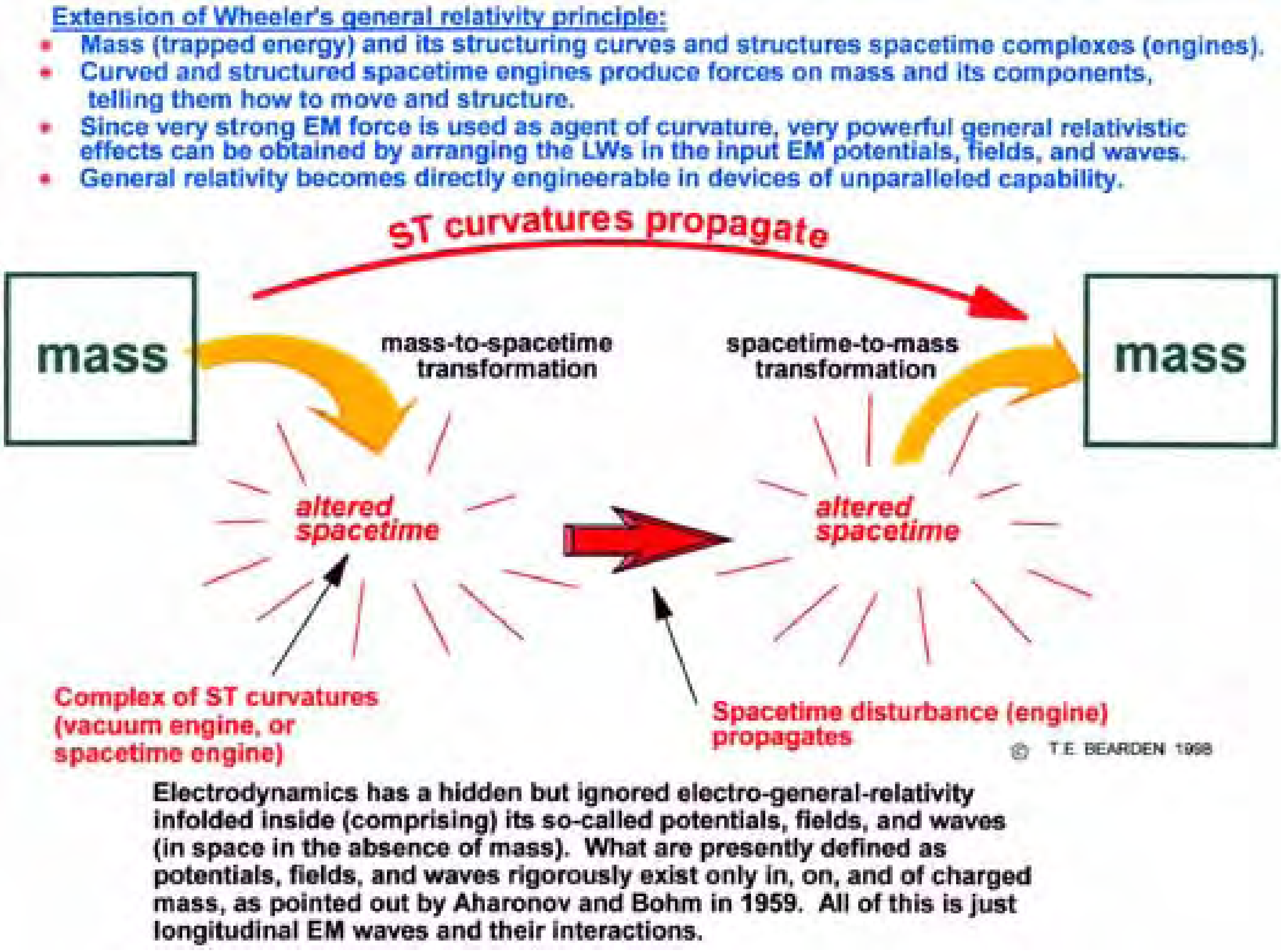

Extension of Wheeler's general relativity principle:

- Mass (trapped energy) and its structuring curves and structures spacetime complexes (engines).
- Curved and structured spacetime engines produce forces on mass and its components, telling them how to move and structure.
- Since very strong EM force is used as agent of curvature, very powerful general relativistic effects can be obtained by arranging the LWs in the input EM potentials, fields, and waves.
- General relativity becomes directly engineerable in devices of unparalleled capability.

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

© T.E. BEARDEN 1998

Electrodynamics has a hidden but ignored electro-general-relativity infolded inside (comprising) its so-called potentials, fields, and waves (in space in the absence of mass). What are presently defined as potentials, fields, and waves rigorously exist only in, on, and of charged mass, as pointed out by Aharonov and Bohm in 1959. All of this is just longitudinal EM waves and their interactions.

Return to Slide Index

Next Slide

Previous Slide

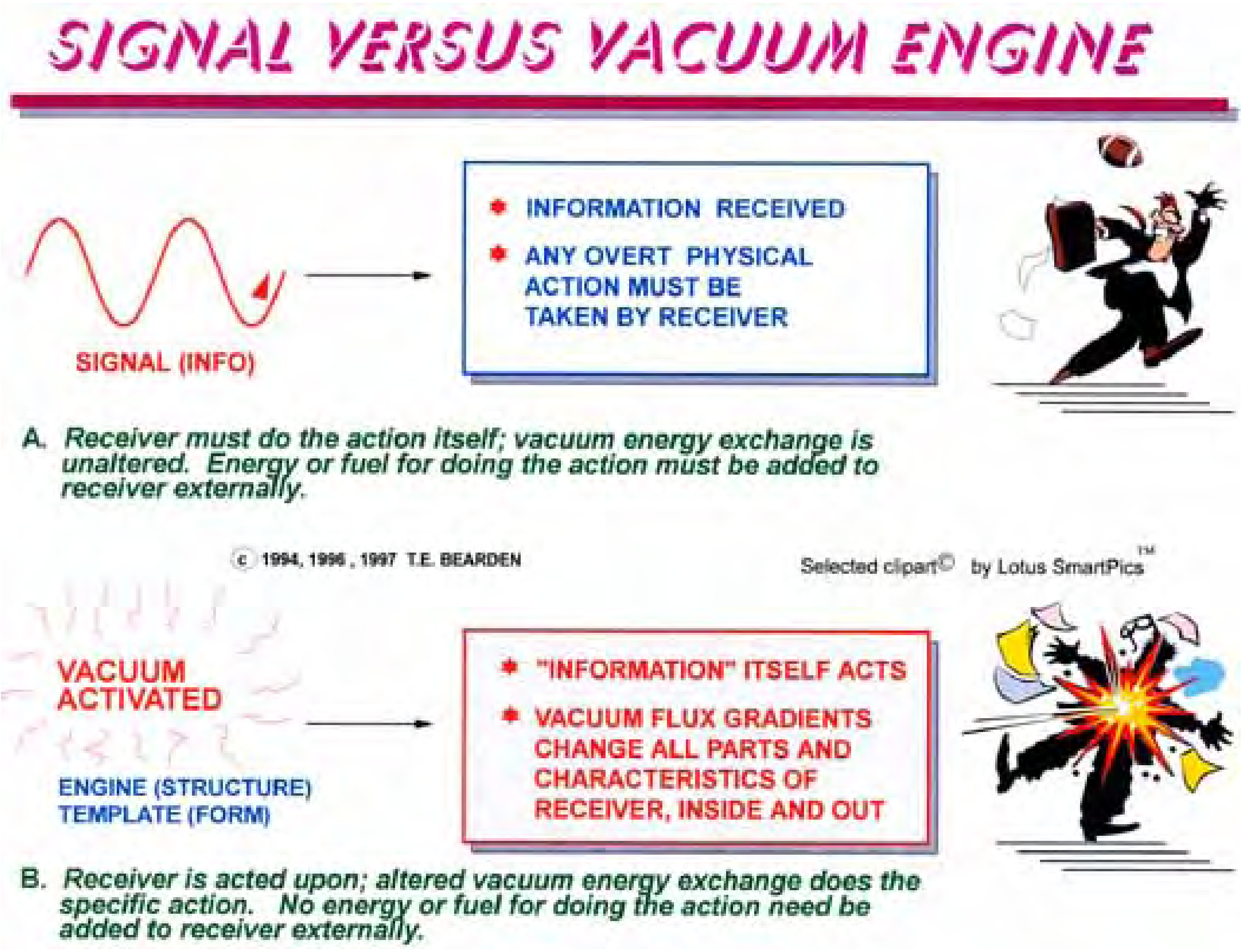

SIGNAL VERSUS VACUUM ENGINE

SIGNAL (INFO)

- INFORMATION RECEIVED
- ANY OVERT PHYSICAL ACTION MUST BE TAKEN BY RECEIVER

A. *Receiver must do the action itself; vacuum energy exchange is unaltered. Energy or fuel for doing the action must be added to receiver externally.*

© 1994, 1996, 1997 T.E. BEARDEN

Selected clipart© by Lotus SmartPics™

VACUUM ACTIVATED

ENGINE (STRUCTURE)
TEMPLATE (FORM)

- "INFORMATION" ITSELF ACTS
- VACUUM FLUX GRADIENTS CHANGE ALL PARTS AND CHARACTERISTICS OF RECEIVER, INSIDE AND OUT

B. *Receiver is acted upon; altered vacuum energy exchange does the specific action. No energy or fuel for doing the action need be added to receiver externally.*

Return to Slide Index

Next Slide

Previous Slide

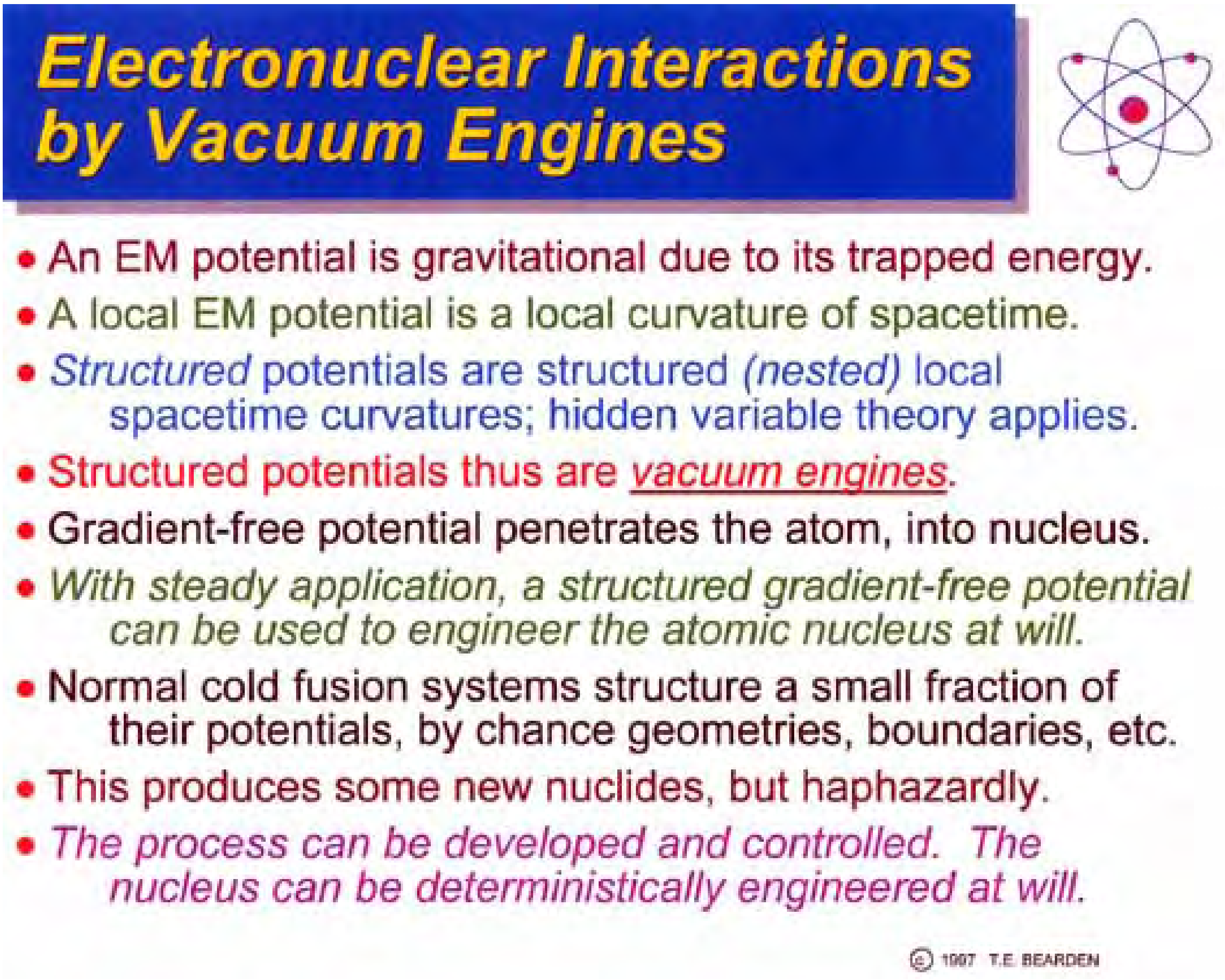

Electronuclear Interactions by Vacuum Engines

- An EM potential is gravitational due to its trapped energy.
- A local EM potential is a local curvature of spacetime.
- *Structured* potentials are structured *(nested)* local spacetime curvatures; hidden variable theory applies.
- Structured potentials thus are *vacuum engines*.
- Gradient-free potential penetrates the atom, into nucleus.
- *With steady application, a structured gradient-free potential can be used to engineer the atomic nucleus at will.*
- Normal cold fusion systems structure a small fraction of their potentials, by chance geometries, boundaries, etc.
- This produces some new nuclides, but haphazardly.
- *The process can be developed and controlled. The nucleus can be deterministically engineered at will.*

© 1997 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

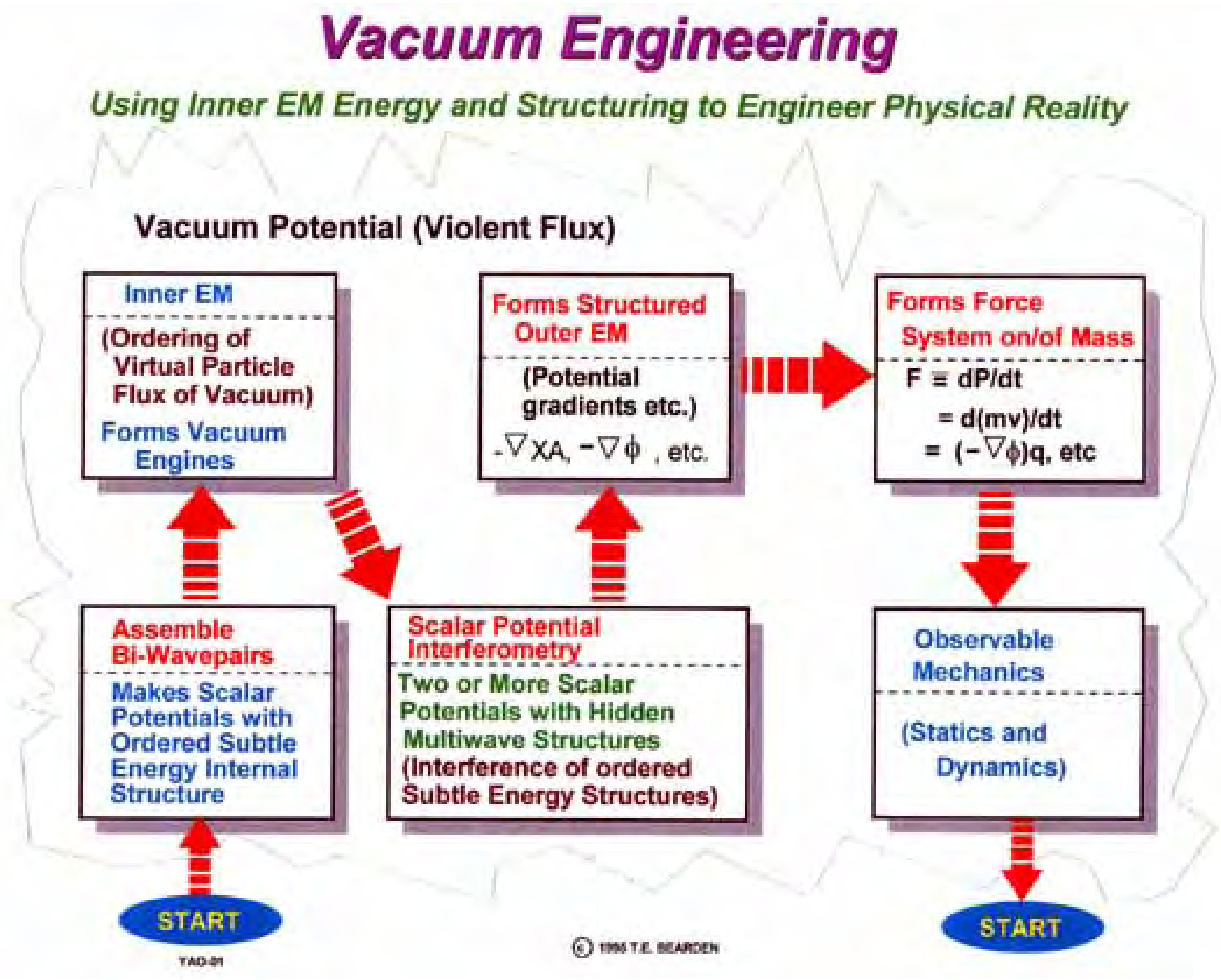

Vacuum Engineering

Using Inner EM Energy and Structuring to Engineer Physical Reality

Vacuum Potential (Violent Flux)

Inner EM

(Ordering of Virtual Particle Flux of Vacuum)

Forms Vacuum Engines

Forms Structured Outer EM

(Potential gradients etc.)

$-\nabla \times A$, $-\nabla \phi$, etc.

Forms Force System on/of Mass

$F \equiv dP/dt$

$= d(mv)/dt$

$= (-\nabla \phi)q$, etc

Assemble Bi-Wavepairs

Makes Scalar Potentials with Ordered Subtle Energy Internal Structure

Scalar Potential Interferometry

Two or More Scalar Potentials with Hidden Multiwave Structures (Interference of ordered Subtle Energy Structures)

Observable Mechanics

(Statics and Dynamics)

START

START

YAO-01

© 1995 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

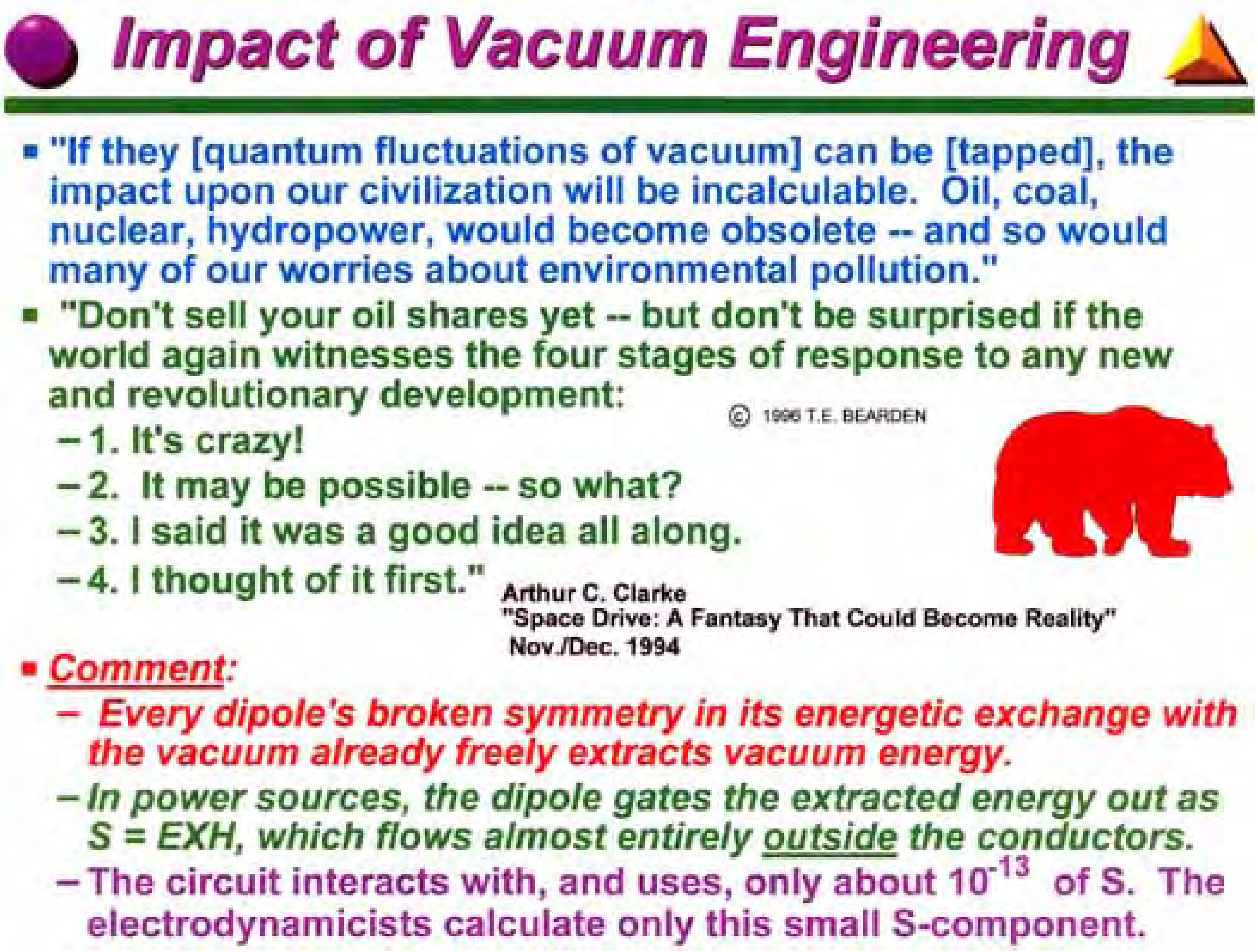

Impact of Vacuum Engineering

- "If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete -- and so would many of our worries about environmental pollution."
- "Don't sell your oil shares yet -- but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development:
  - 1. It's crazy!
  - 2. It may be possible -- so what?
  - 3. I said it was a good idea all along.
  - 4. I thought of it first."

© 1996 T.E. BEARDEN

Arthur C. Clarke
"Space Drive: A Fantasy That Could Become Reality"
Nov./Dec. 1994

- *Comment:*
  - *Every dipole's broken symmetry in its energetic exchange with the vacuum already freely extracts vacuum energy.*
  - *In power sources, the dipole gates the extracted energy out as S = EXH, which flows almost entirely outside the conductors.*
  - The circuit interacts with, and uses, only about 10^{-13} of S. The electrodynamicists calculate only this small S-component.

Return to Slide Index

Next Slide

Previous Slide

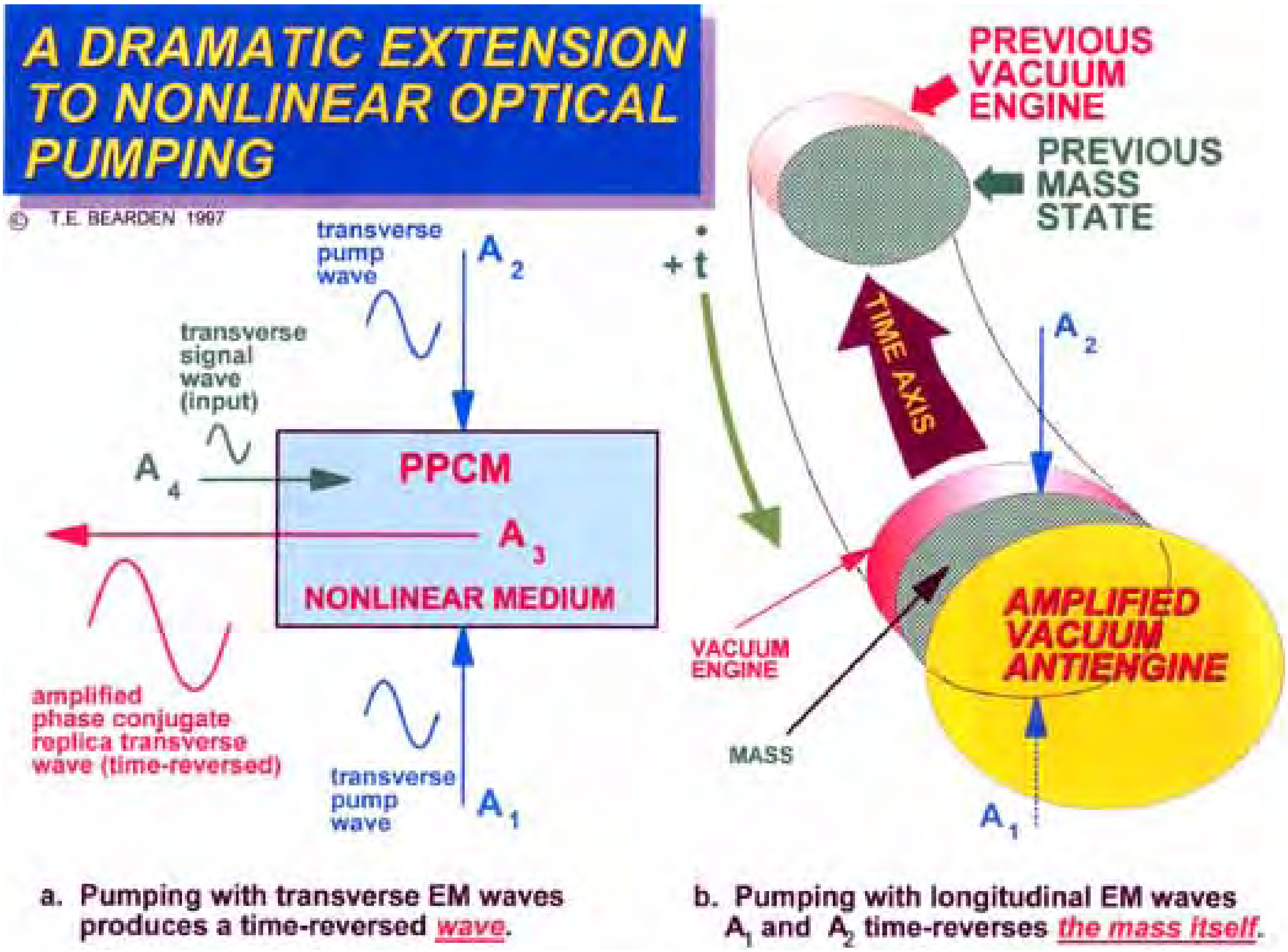

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

transverse pump wave

A_2

transverse signal wave (input)

A_4

PPCM

A_3

NONLINEAR MEDIUM

amplified phase conjugate replica transverse wave (time-reversed)

transverse pump wave

A_1

+ ṫ

PREVIOUS VACUUM ENGINE

PREVIOUS MASS STATE

TIME AXIS

A_2

VACUUM ENGINE

AMPLIFIED VACUUM ANTIENGINE

MASS

A_1

a. Pumping with transverse EM waves produces a time-reversed *wave*.

b. Pumping with longitudinal EM waves A_1 and A_2 time-reverses *the mass itself*.

Return to Slide Index

Next Slide

Previous Slide

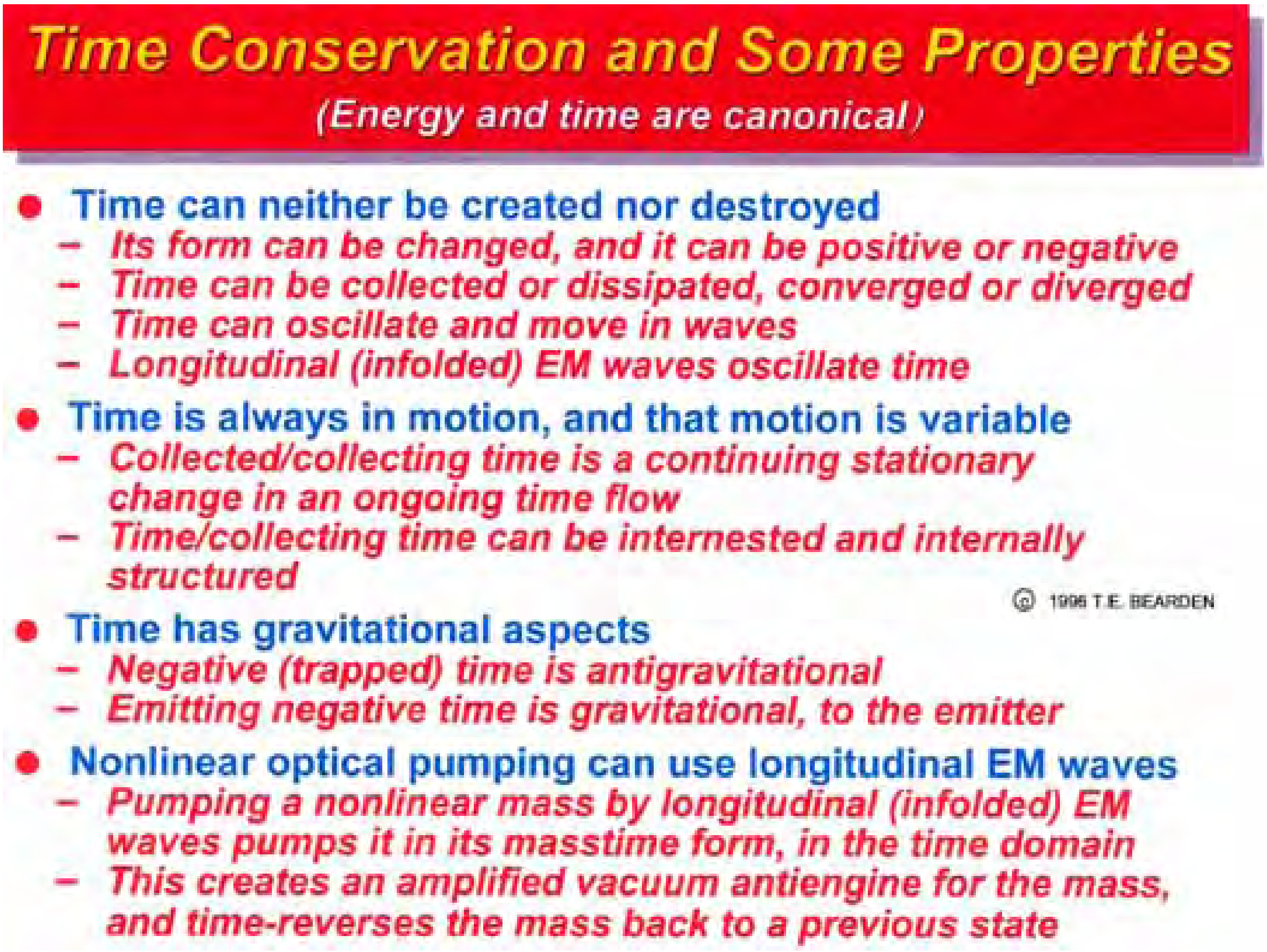

Time Conservation and Some Properties

(Energy and time are canonical)

- Time can neither be created nor destroyed
  - *Its form can be changed, and it can be positive or negative*
  - *Time can be collected or dissipated, converged or diverged*
  - *Time can oscillate and move in waves*
  - *Longitudinal (infolded) EM waves oscillate time*
- Time is always in motion, and that motion is variable
  - *Collected/collecting time is a continuing stationary change in an ongoing time flow*
  - *Time/collecting time can be internested and internally structured*
- Time has gravitational aspects
  - *Negative (trapped) time is antigravitational*
  - *Emitting negative time is gravitational, to the emitter*
- Nonlinear optical pumping can use longitudinal EM waves
  - *Pumping a nonlinear mass by longitudinal (infolded) EM waves pumps it in its masstime form, in the time domain*
  - *This creates an amplified vacuum antiengine for the mass, and time-reverses the mass back to a previous state*

© 1996 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

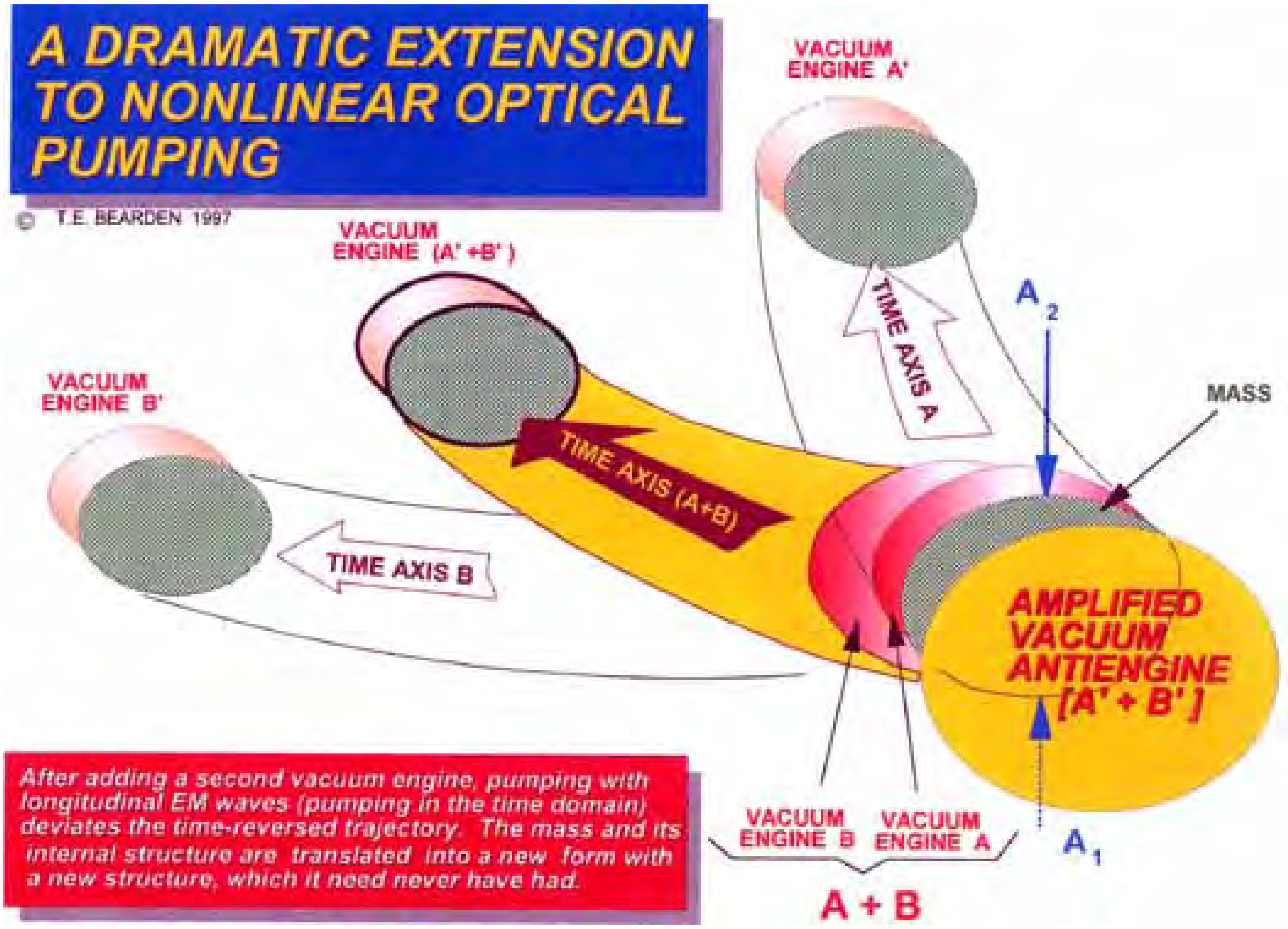

A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

VACUUM ENGINE A'
VACUUM ENGINE (A' +B')
VACUUM ENGINE B'
TIME AXIS A
TIME AXIS (A+B)
TIME AXIS B
A_2
MASS
AMPLIFIED VACUUM ANTIENGINE [A' + B']
VACUUM ENGINE B
VACUUM ENGINE A
A + B
A_1

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

Return to Slide Index

Next Slide

Previous Slide

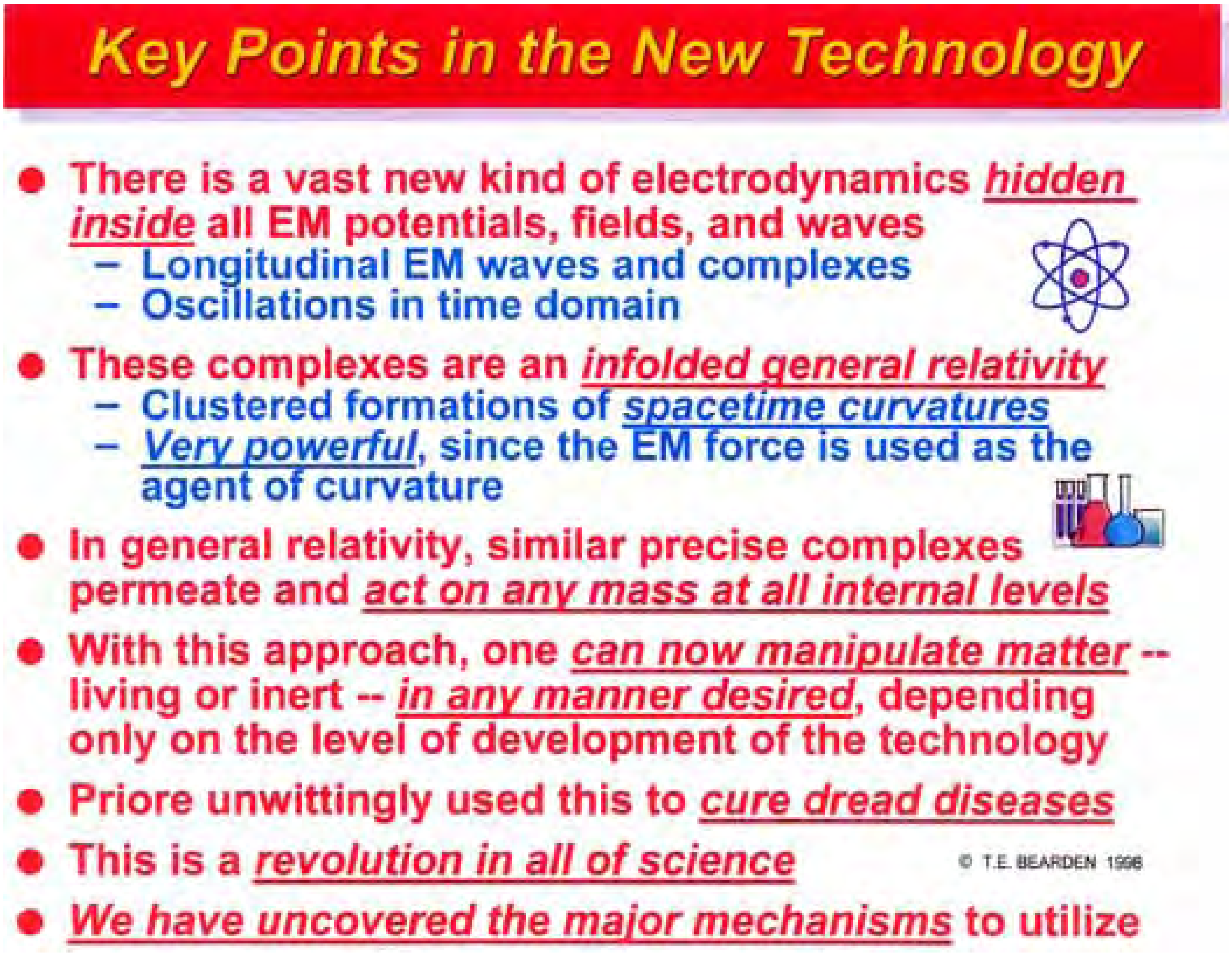

Key Points in the New Technology

- There is a vast new kind of electrodynamics *hidden inside* all EM potentials, fields, and waves
  - Longitudinal EM waves and complexes
  - Oscillations in time domain
- These complexes are an *infolded general relativity*
  - Clustered formations of *spacetime curvatures*
  - *Very powerful*, since the EM force is used as the agent of curvature
- In general relativity, similar precise complexes permeate and *act on any mass at all internal levels*
- With this approach, one *can now manipulate matter* -- living or inert -- *in any manner desired*, depending only on the level of development of the technology
- Priore unwittingly used this to *cure dread diseases*
- This is a *revolution in all of science*
- *We have uncovered the major mechanisms* to utilize

© T.E. BEARDEN 1996

Return to Slide Index

Next Slide

Previous Slide

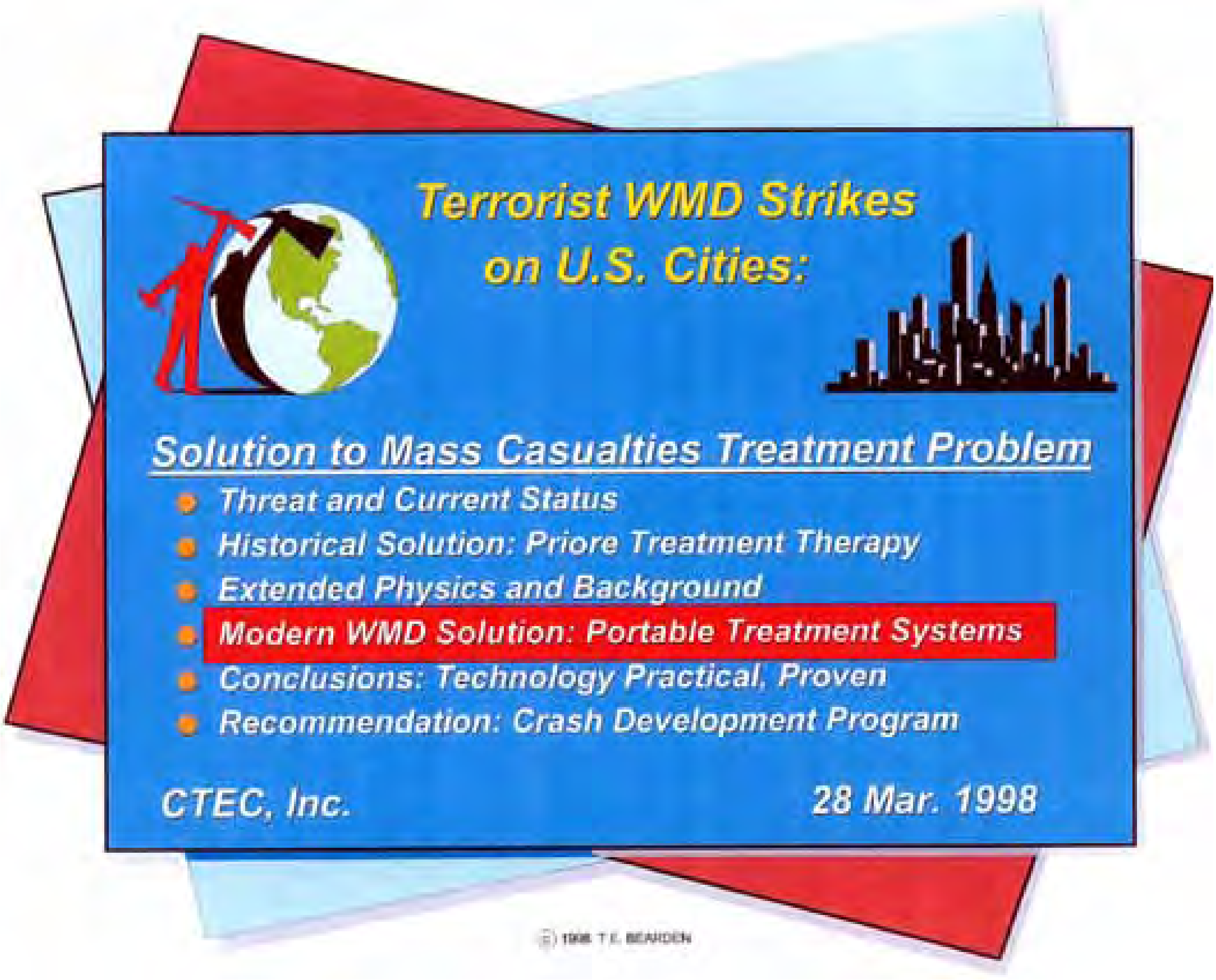

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

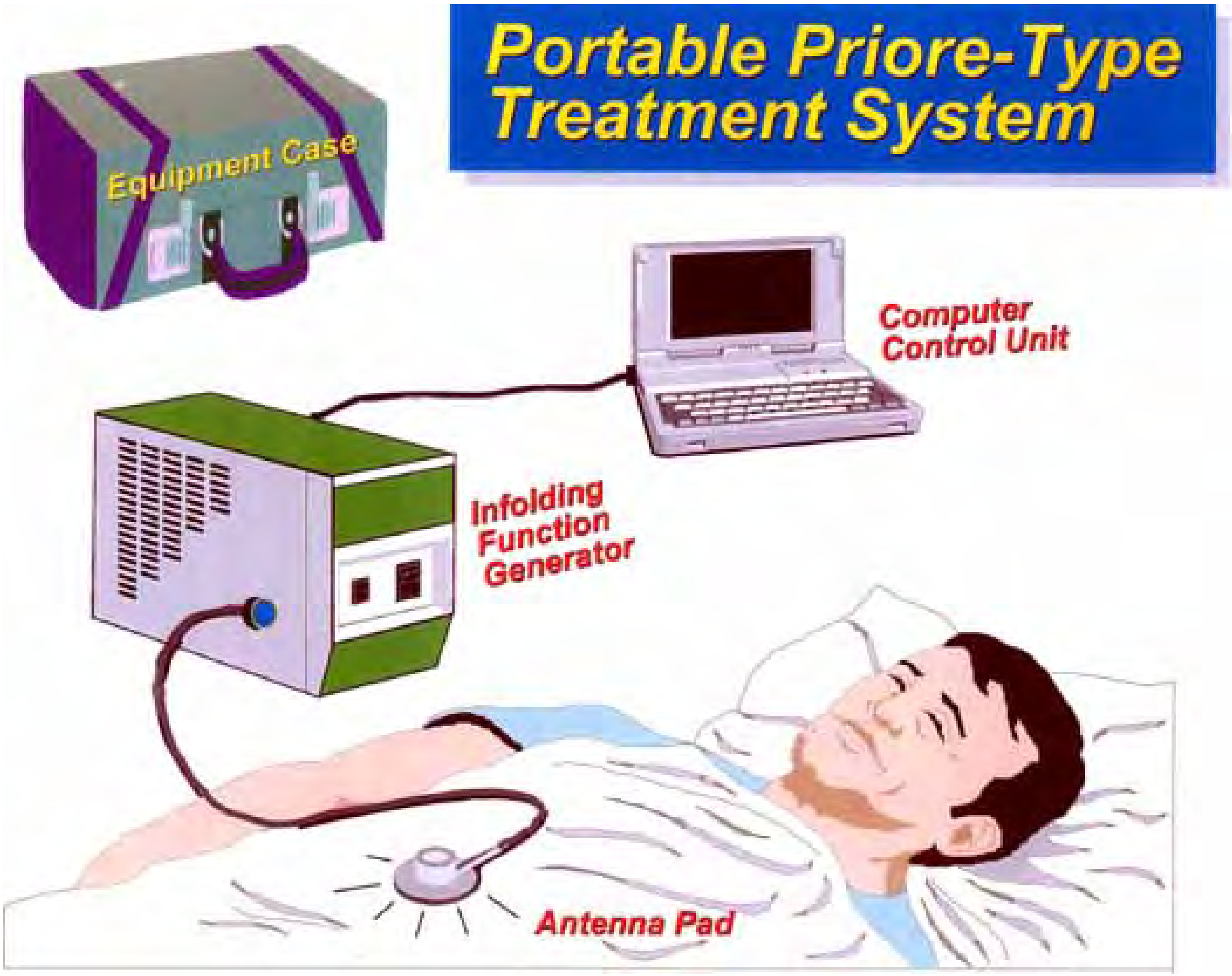

Portable Priore-Type Treatment System

Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Return to Slide Index

Next Slide

Previous Slide

Portable System Block Diagram

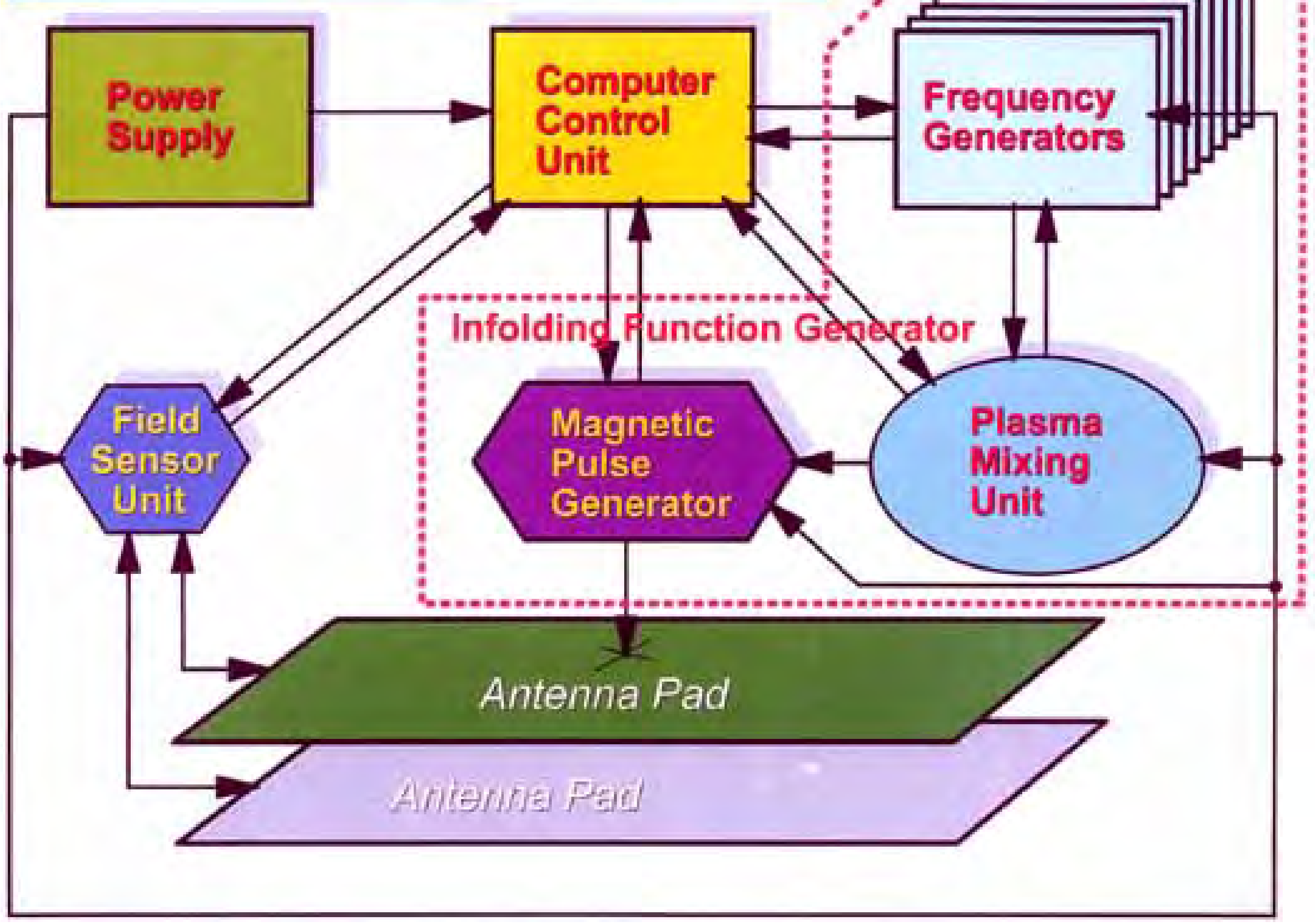

Power Supply

Computer Control Unit

Frequency Generators

Infolding Function Generator

Field Sensor Unit

Magnetic Pulse Generator

Plasma Mixing Unit

Antenna Pad

Antenna Pad

[Return to Slide Index]

[Next Slide]

[Previous Slide]

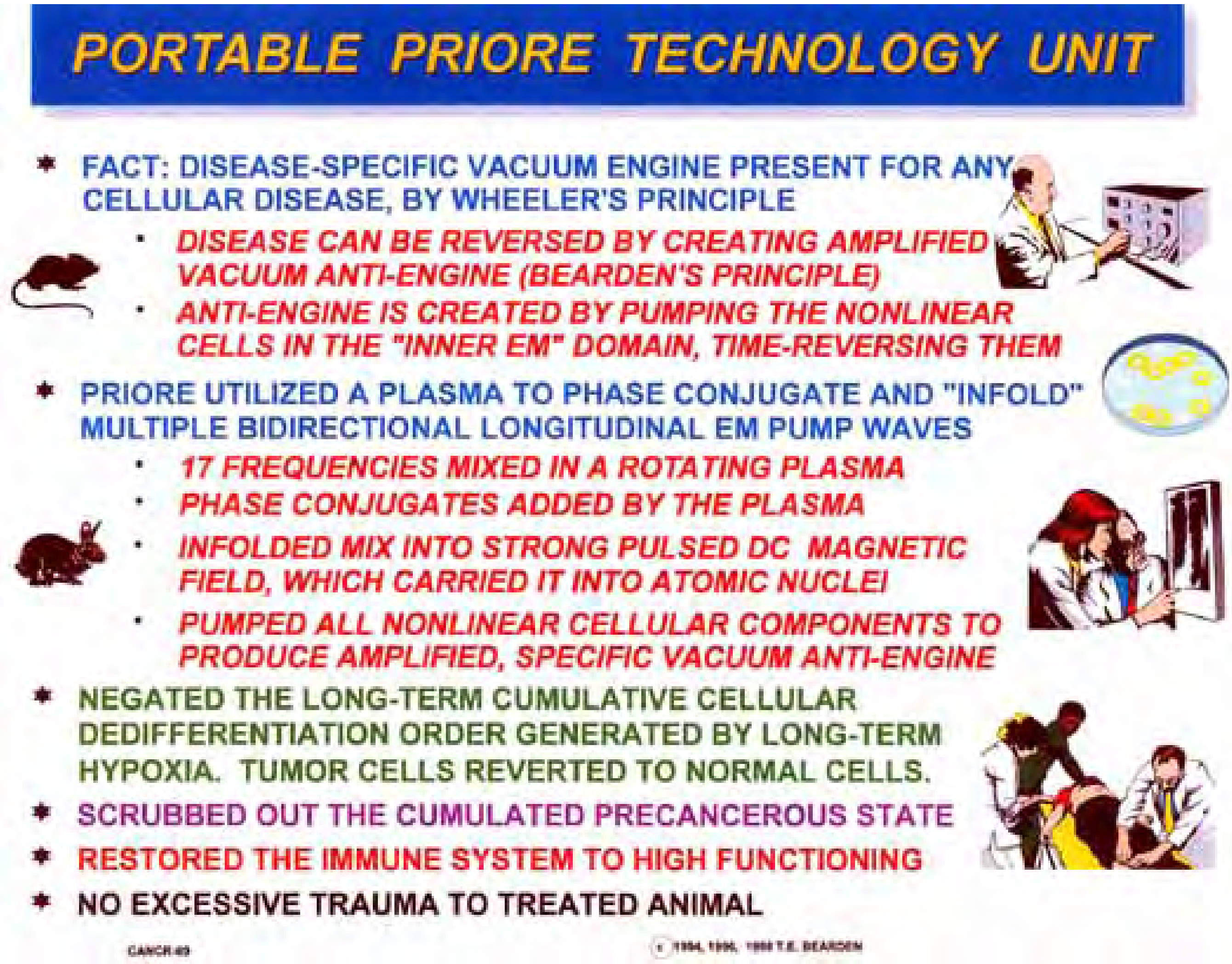

PORTABLE PRIORE TECHNOLOGY UNIT

- FACT: DISEASE-SPECIFIC VACUUM ENGINE PRESENT FOR ANY CELLULAR DISEASE, BY WHEELER'S PRINCIPLE
  - *DISEASE CAN BE REVERSED BY CREATING AMPLIFIED VACUUM ANTI-ENGINE (BEARDEN'S PRINCIPLE)*
  - *ANTI-ENGINE IS CREATED BY PUMPING THE NONLINEAR CELLS IN THE "INNER EM" DOMAIN, TIME-REVERSING THEM*
- PRIORE UTILIZED A PLASMA TO PHASE CONJUGATE AND "INFOLD" MULTIPLE BIDIRECTIONAL LONGITUDINAL EM PUMP WAVES
  - *17 FREQUENCIES MIXED IN A ROTATING PLASMA*
  - *PHASE CONJUGATES ADDED BY THE PLASMA*
  - *INFOLDED MIX INTO STRONG PULSED DC MAGNETIC FIELD, WHICH CARRIED IT INTO ATOMIC NUCLEI*
  - *PUMPED ALL NONLINEAR CELLULAR COMPONENTS TO PRODUCE AMPLIFIED, SPECIFIC VACUUM ANTI-ENGINE*
- NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CELLS.
- SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE
- RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING
- NO EXCESSIVE TRAUMA TO TREATED ANIMAL

CANCR-69

© 1994, 1996, 1999 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

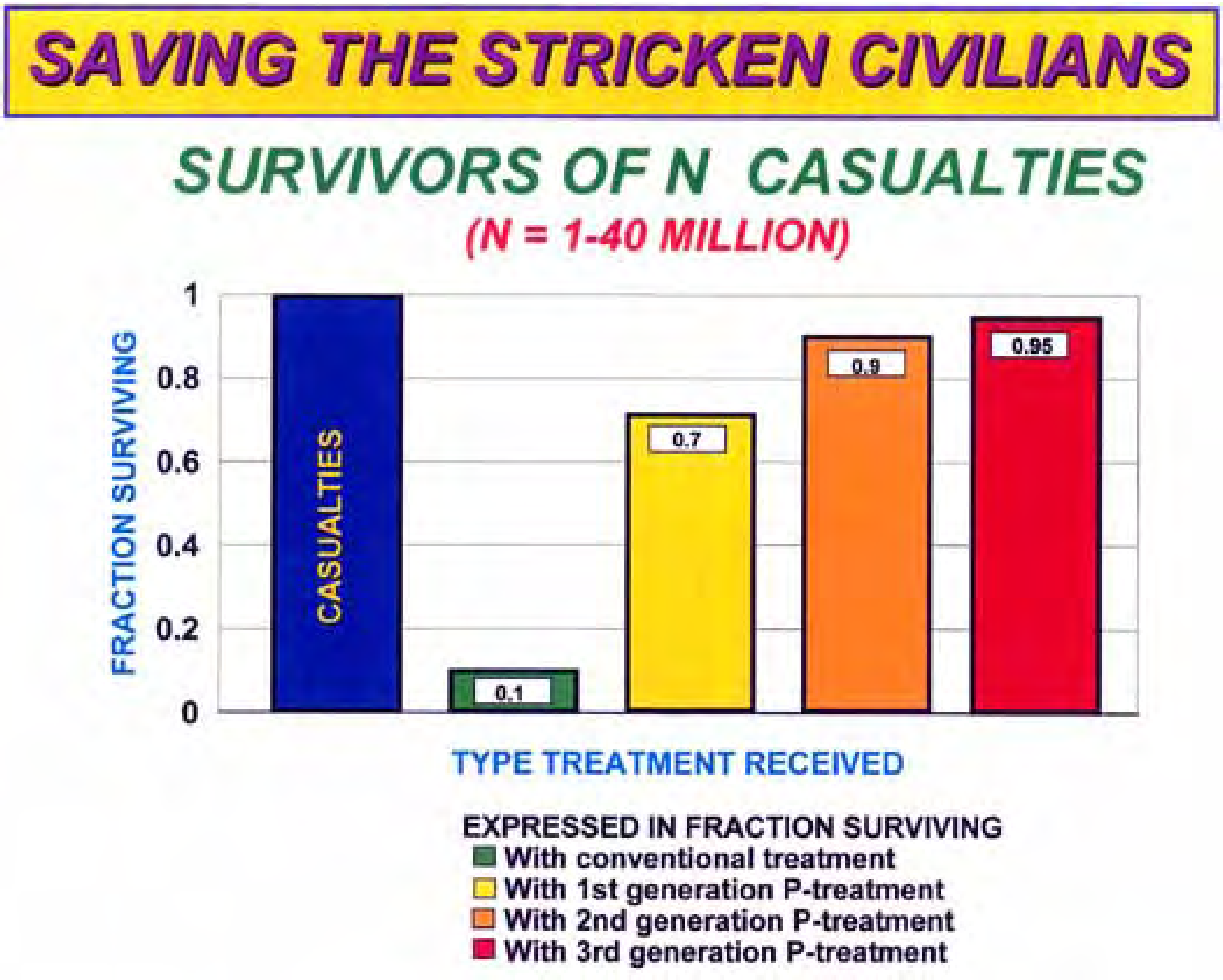

SAVING THE STRICKEN CIVILIANS

SURVIVORS OF N CASUALTIES

(N = 1-40 MILLION)

[Return to Slide Index]
[Next Slide]
[Previous Slide]

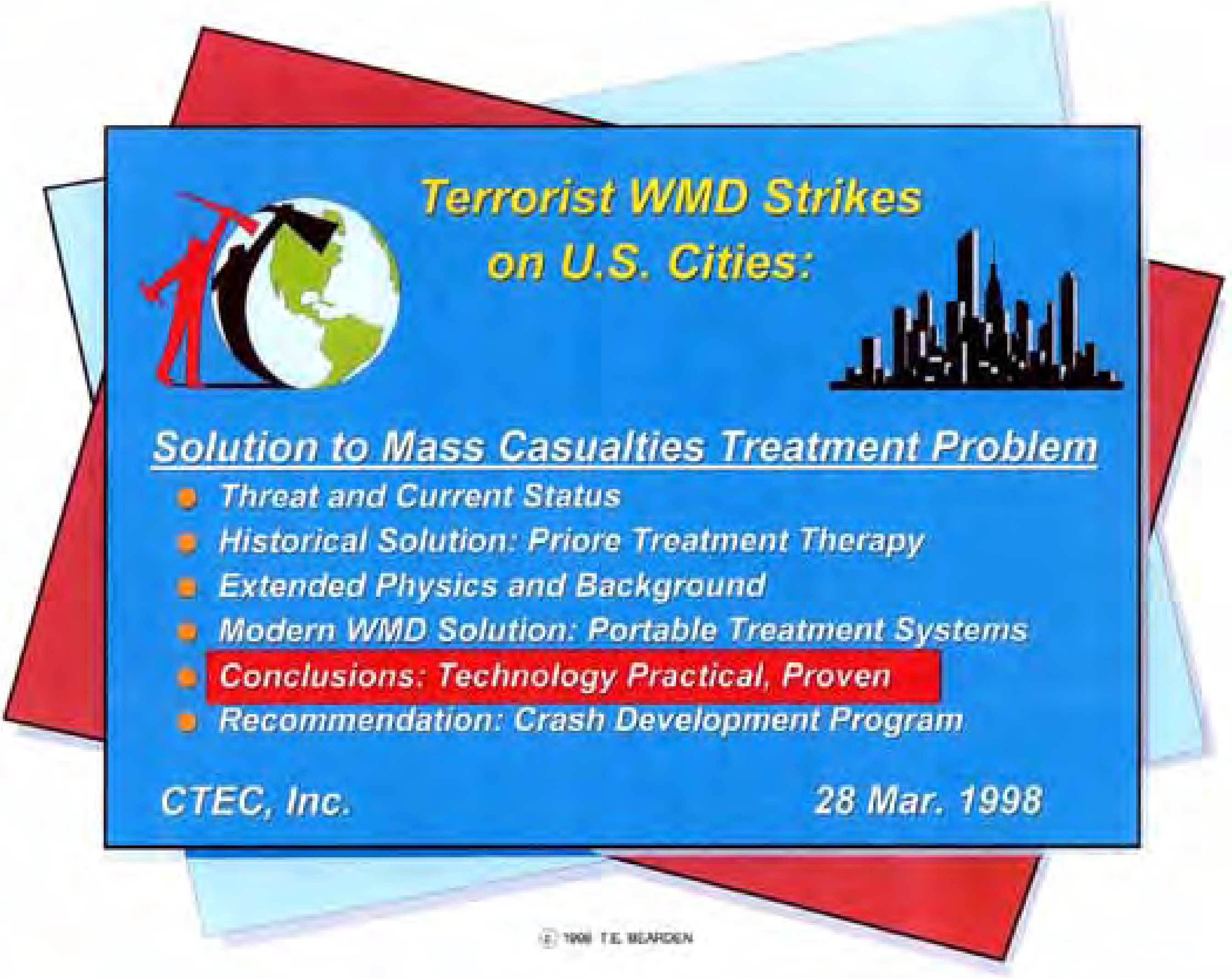

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

© 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

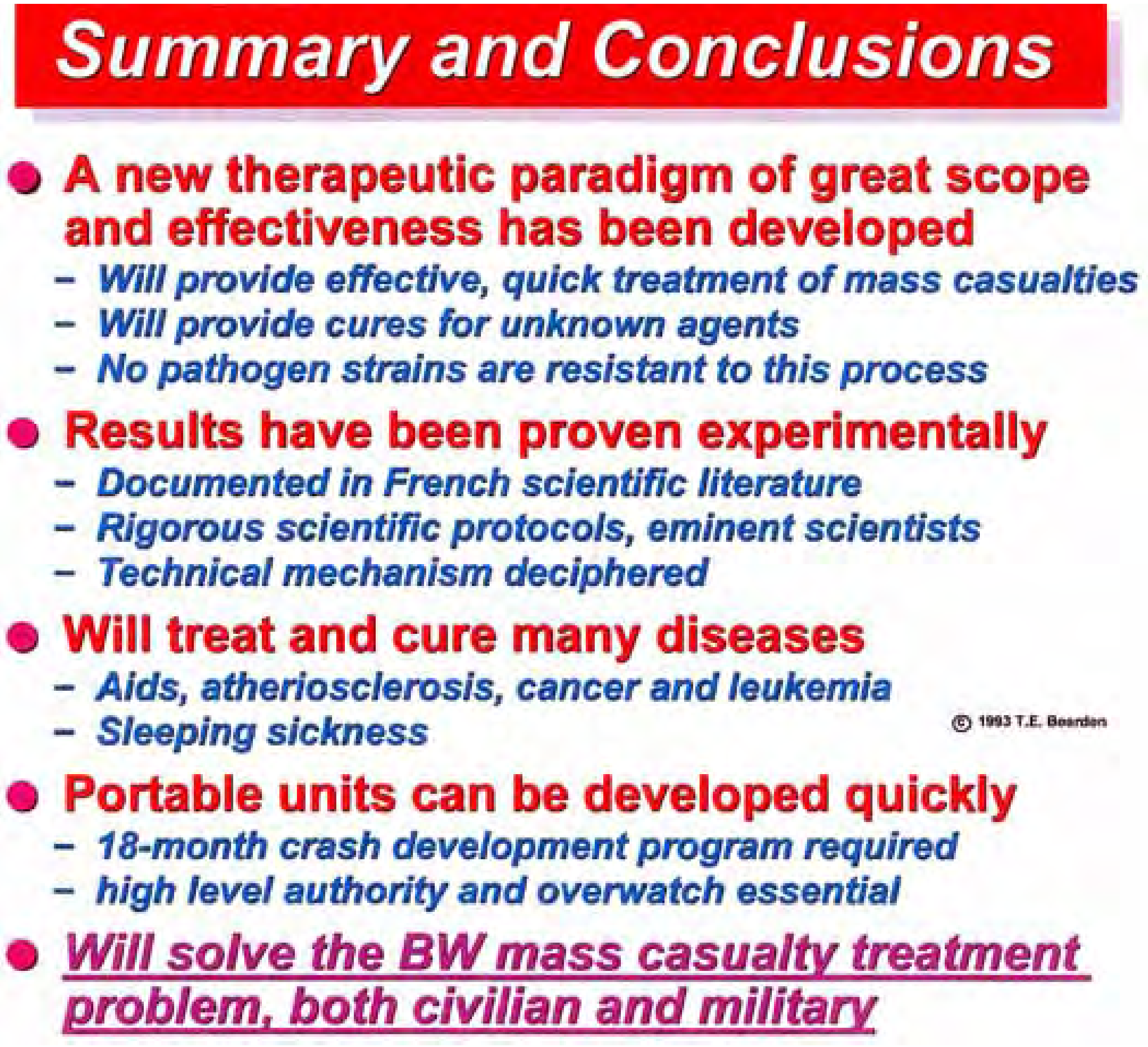

Summary and Conclusions

- A new therapeutic paradigm of great scope and effectiveness has been developed
  - Will provide effective, quick treatment of mass casualties
  - Will provide cures for unknown agents
  - No pathogen strains are resistant to this process
- Results have been proven experimentally
  - Documented in French scientific literature
  - Rigorous scientific protocols, eminent scientists
  - Technical mechanism deciphered
- Will treat and cure many diseases
  - Aids, atherioscIerosis, cancer and leukemia
  - Sleeping sickness
- Portable units can be developed quickly
  - 18-month crash development program required
  - high level authority and overwatch essential
- *Will solve the BW mass casualty treatment problem, both civilian and military*

© 1993 T.E. Bearden

Return to Slide Index

Next Slide

Previous Slide

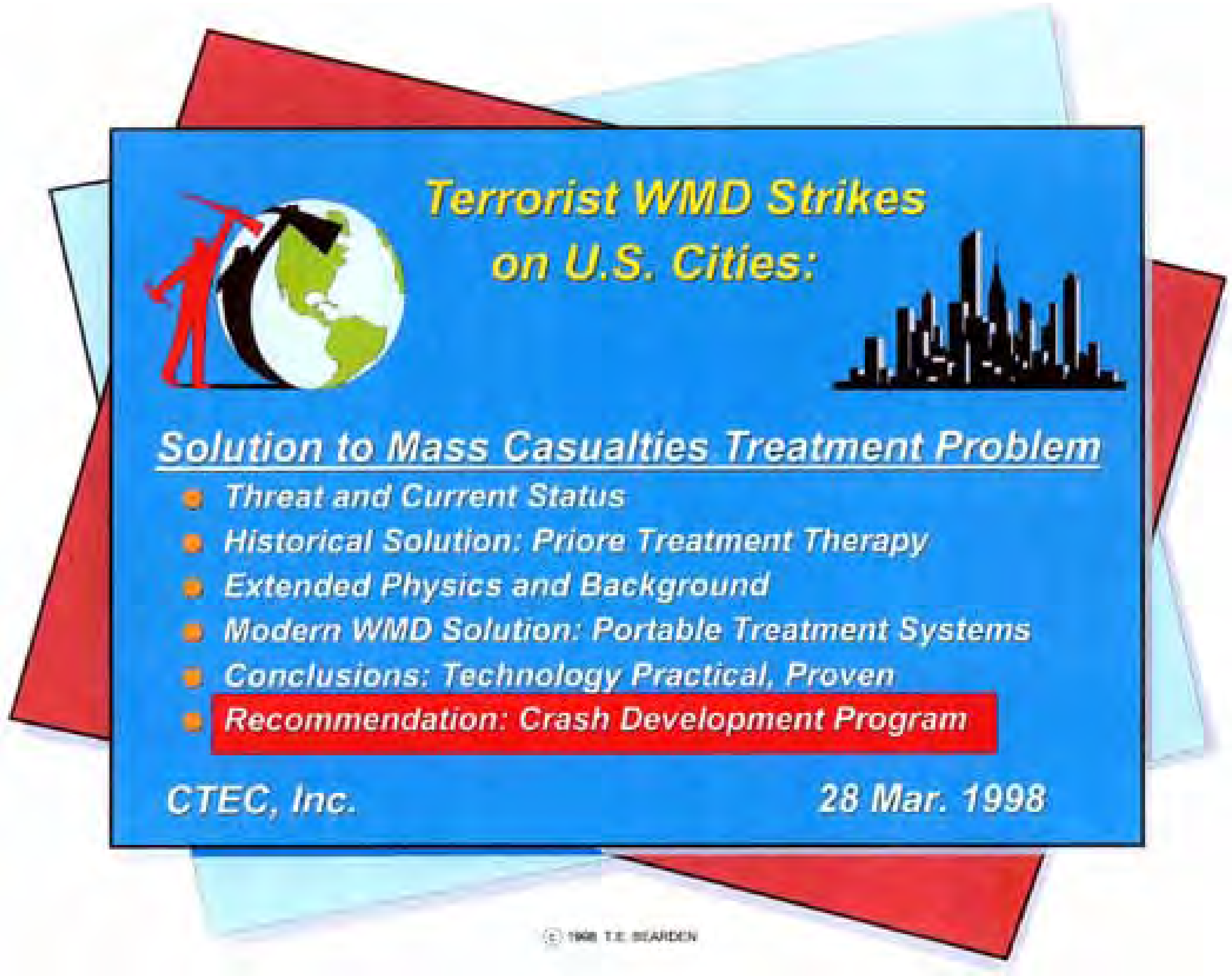

Terrorist WMD Strikes on U.S. Cities:

Solution to Mass Casualties Treatment Problem

- Threat and Current Status
- Historical Solution: Priore Treatment Therapy
- Extended Physics and Background
- Modern WMD Solution: Portable Treatment Systems
- Conclusions: Technology Practical, Proven
- Recommendation: Crash Development Program

CTEC, Inc.

28 Mar. 1998

(c) 1998 T.E. BEARDEN

Return to Slide Index

Next Slide

Previous Slide

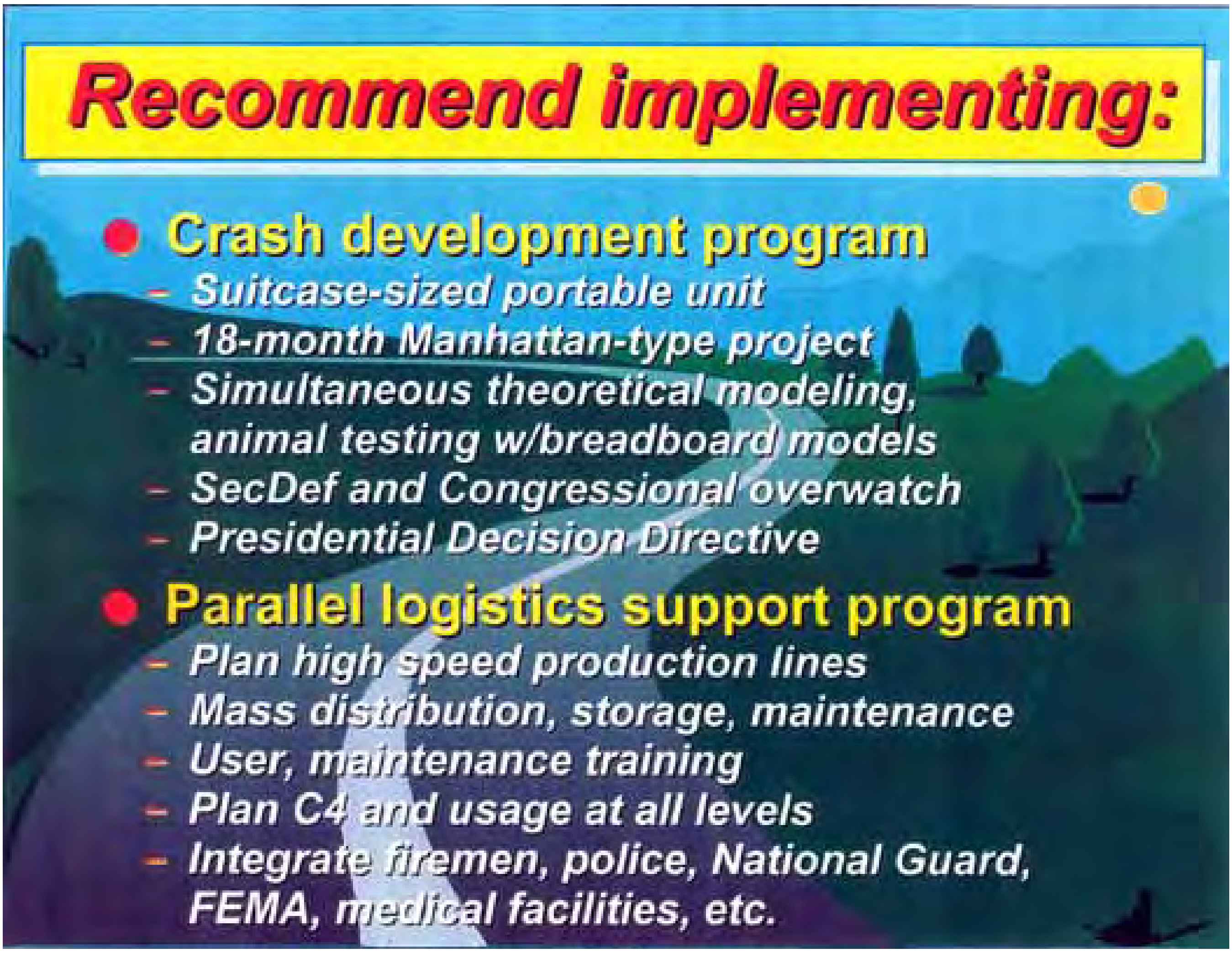

Recommend implementing:

- Crash development program
  - Suitcase-sized portable unit
  - 18-month Manhattan-type project
  - Simultaneous theoretical modeling, animal testing w/breadboard models
  - SecDef and Congressional overwatch
  - Presidential Decision Directive
- Parallel logistics support program
  - Plan high speed production lines
  - Mass distribution, storage, maintenance
  - User, maintenance training
  - Plan C4 and usage at all levels
  - Integrate firemen, police, National Guard, FEMA, medical facilities, etc.

Return to Slide Index

Next Slide

Previous Slide

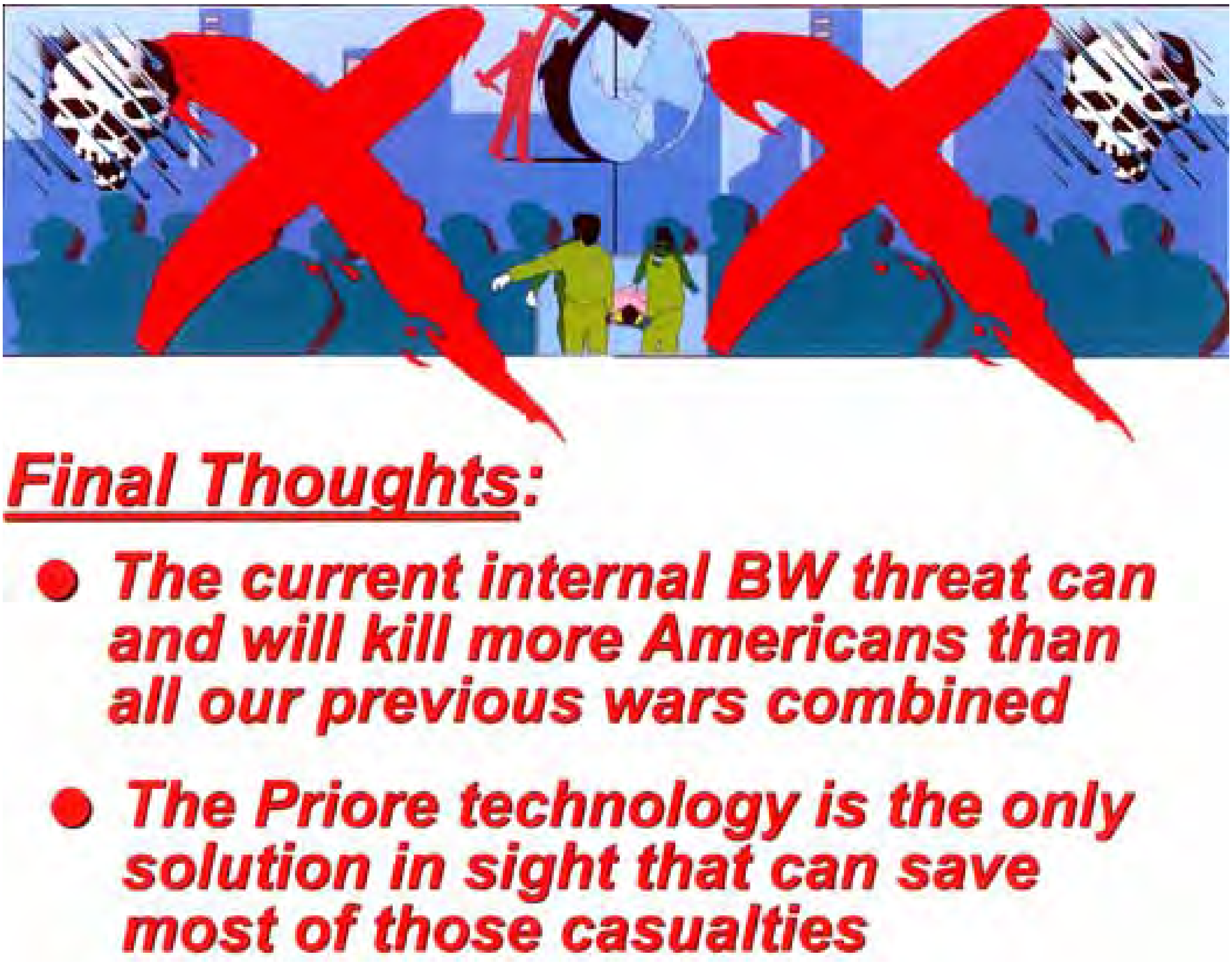

Final Thoughts:

- The current internal BW threat can and will kill more Americans than all our previous wars combined
- The Priore technology is the only solution in sight that can save most of those casualties

Return to Slide Index

Previous Slide

CTEC, Inc.
2311 Big Cove Road
Huntsville, AL 35801
(256) 533-3682 Ph/

CTEC, Inc.

June 24, 1998

Dr. Jane F. Kinsel, Director
Office of Policy Analysis
Department of Health & Human Services
Bathesda, MD 20892

Dear Dr. Kinsel:

Your letter of June 8 is appreciated. However, the entire action package had already gone to the Assistant SecDef and also to Major General Neary in the USAF.

The proof of a revolution in medical treatment was in the package. The basis for it is already demonstrated experimentally in the scientific literature, both in the U.S. and in France. We are speaking of something already proven, but missed by the entire U.S. medical community. Strong words, but true. Let me be specific.

Any cell in the body can be time-reversed (physics term) or dedifferentiated (biology term) back to an earlier state, by pumping (i.e., "squeezing" in simple terms) that cell with longitudinal EM waves as pump waves. There are considerable papers in the French scientific literature showing the experimental results.

Apply that to cancer. We are stating bluntly that all cancerous cells in the body can be "depromoted" back to normal cells easily. Further, the damage previously done to them (such as by sustained hypoxia due to contaminants reducing the oxygen-carrying ability normally given to it by some 60-to-80 surrounding water molecules per red cell and the consequent H-bonding interactions) will also be reversed. The cells can all be returned to healthy, normal cells again, without "killing" or "burning" or "cutting" anything.

With some development of the technology, you can quickly and easily cure every cancer patient in every hospital in the United States.

Now apply it to AIDS. Since you can time-reverse every cell in the body, you can time-reverse the HIV-infected cells – genetics and all – right back to normal cells with normal genetics. Those cells not infected and healthy, will just get a "little younger" again.

So with that small development, you can quickly and easily cure every AIDS patient in every hospital in the United States.

I would have thought that NIH would certainly be interested in such a revolutionary therapy which has already been demonstrated in laboratory experiments. World-renowned French scientists – such as the eminent parasitologist Pautrizel and research PhDs assigned directly by Robert Courrier, head of the Biology Section of the French Academy of Sciences, worked with Prioré in performing those startling experiments that proved what we are saying.

The method was rigorously shown to cure infectious diseases (such as trypanosomiasis), atherosclerosis (which now afflicts millions of Americans), and terminal tumors. It was shown to reverse suppressed immune systems back to normal vigorous functioning.

Please permit me to disagree with you that the examination of such is not the mission of the NIH. To the contrary, it is precisely the NIH which should have some of its very best scientists looking into this.

The action package to DoD was to propose a crash program to develop this methodology in the form of a small, portable unit for treatment of mass casualties resulting from a terrorist BW attack on our cities (now expected in the next few weeks, as you must be aware). The leading Arab terrorist of the world has already promised just this.

We are speaking of several million Americans who are going to die. They are going to die in spite of everything that NIH and its labs can do, even though I know you will make a heroic effort. With NIH present methods, you cannot do very much for those stricken Americans at all. Everything we have in the "normal" kit bag is almost useless against a knowledgeable spray attack of modified smallpox (as you know, the Russian secret BW labs have made and sold tons of that already, to guess who!). Or a spray of tularemia (you are well aware of the extreme lethality of that), or anthrax (easily obtained). It is also common knowledge that these terrorist teams with their BW agents are already on site in this country, waiting for the word to attack. This has now been officially raised to our primary Strategic Threat.

In the face of such looming strikes on the U.S., I simply cannot believe that the NIH does not feel it within their mission to rigorously check out a proposed method for saving perhaps 70% of those coming deaths (first generation equipment), and up to 90% with second generation equipment.

Perhaps you just did not read the package carefully. I urge you to do so.

Else NIH is going to look very, very bad when (1) the strikes do occur, (2) they can't handle them (and they cannot, as you well know), (3) they had within their hands in advance, a proposed methodology that could have been investigated intensely, to save millions of those stricken Americans.

As a staunch supporter of NIH and especially CDC, I have to believe that there is a different attitude there. If NIH is not interested in a previously experimentally demonstrated method of saving millions of Americans, then the only recourse left is the political channel.

I assure you that my critique of electrodynamics is well-founded, as supported by various leading physicists such as Nobelist Feynman (who stated bluntly that the field concept is wrong), Wheeler, Bunge, and so on. It is inexplicable why all our universities are still teaching a 130 year old abridged (Heaviside's) subset of Maxwell's theory, and one that still contains a material ether even though that was falsified 110 years ago!

You already have seen the misuse of this technology by the Russians, in the decades-long radiation of the U.S. Embassy in Moscow. Is it not NIH's mission to understand how they generated all those diseases and health changes? Have you ever looked into Kaznacheyev's experiments showing that ANY cellular disease or disorder can be induced in cells at a distance by purely EM means? Have you looked into what got eliminated from electrodynamics by ignoring the hidden longitudinal EM inside every potential, field, and wave? Do you realize what ignoring that has done to American chemistry, biochemistry, and medical science?"

I would hope that the NIH would at least be interested in its own mission area!

I challenge you bluntly to admit that no one presently understands the primary operation of the human cellular ***regenerative*** system. That's because its technical principle is not presently in our recognized physics, biochemistry, and medical science – even though it has been in the hard physics literature since 1903-4. Becker's epochal work came closest; since he only had access to the conventional EM model, it failed him. But he at least showed you rigorously that cells can be changed (both differentiated and redifferentiated) all over the map by laughably weak EM – picoamperes of current, which means persistent DC potentials. In the microwave radiation of the U.S. Embassy, all the health changes occurred in regions absent of fields – i.e., absent of bleed-off of the potentials. Hence in regions of stable DC potentials! With 100% correlation of the health changes to persistent, gradient free potentials, and 100% anticorrelation to changing potentials (i.e., fields), our own scientists concluded totally erroneously that it could not have been the EM radiation causing the changes, when their own results had just proven that it was.

Now I ask you: Can NIH explain how it was that the DC potentials caused those health changes? I think not. If it could, it would understand the action package I wrote, and its importance.

This is not a "political" letter. I'm just an ordinary American citizen who has done about 30 years very hard work on this particular area. The work is good enough now to stand on its own merits. The scientific community will eventually accept it, about 50 years from now. But we cannot wait 50 years! Those millions of Americans are going to die, because of the bureaucracy and inaction of our own scientific organizations. Whether they like it or not, the U.S. scientific community is now the forefront of the struggle for this nation to survive. And they are failing us dramatically.

So as a parent who wants his children to live, and his neighbors' children to live, and those millions of soon-to-be stricken Americans to live also, I urge you to at least have someone competent in physics (and knowledgeable that the foundations of physics itself are in woeful shape!) check the work. Has anybody there *read* any of the references to the Prioré work I cited? Does anyone there realize that *there are no EM forcefields in space?* Probably not.

Dr. Kinsel, there is a time for "business as usual" and there is a time to do some serious new work. Those stricken Americans are going to be lying there. Nothing NIH can do is going to stop that. That's a given, and it is not my own estimate. It's the official estimate. And you *know* you presently do not have the tools to do anything really effective about it.

Frankly, I interpret your letter to me as just the normal "spin control" for easy disposition. I would hope that NIH would seriously rethink the situation, get off the "spin control" and "business as usual" posturing, and do something dramatically different for a change.

What have you got to lose? A little time from perhaps two or three of your best scientists.

What have you got to gain? Saving untold millions of American lives. A new, *already partially demonstrated* cure for most dread diseases such as AIDS and cancer – diseases which you presently cannot do very much about in spite of decades of heroic struggle.

Sincerely,

cc: Senator Shelby, Congr. Cramer

T.E. Bearden
President and CEO

CTEC, Inc.
T.E. Bearden, President & CEO
2311 Big Cove Road
Huntsville, AL 35801-1351
(256)533-3682 ph/(256)536-0411 Fax

CTEC, Inc.

May 12, 1998

Dr. Harold Varmus, Director
National Institute of Health
Bathesda, MD 20892
(301) 496-1766

Dear Dr Varmus:

We have recently sent action documents to the Director of Nuclear and Counterproliferation, Office of the DCS, Air and Space Operations, HQ USAF and to the DoD (General Busbee) Assistant Secretary of Defense, as well as several Senate and House committees, which may be of interest to you. For one thing, we give the technical mechanism used to generate the Gulf War Syndrome. The same weaponry is now being used to rapidly induce new strains of certain pathogens worldwide. I urge you to closely read that section of the letter to MG Neary which explains the GWS, why the French did not get it, and why Southern Iraqis got an increase in cancer and leukemia but not GWS. Also explained is how the "microwave radiation" of the U.S. Embassy in Moscow induced diseases for decades, and how we failed to comprehend it because the standard EM model deceived us.

There is a way to treat and quickly cure most diseases, including the mass casualties resulting from a terrorist BW strike on our civilian population centers, our military bases, or our forces in the field. The method is little known, but was scientifically demonstrated in France in the 1960s and early 1970s before cancellation of the Prioré Project when the French Government changed. Eminent French scientists worked on the project, and the results are fully documented in leading French scientific journals. References are listed in a separate attachment.

Previously the Western scientific community has been unable to understand the technical mechanism responsible for such remarkable cures, because of serious foundations flaws in Western physics and electrodynamics. *Foundations* physicists have long been aware our science is flawed, and have pointed out many of these fundamental errors. However, the operational science establishment has not paid attention, even though physicists such as Feynman and Wheeler pointed out that the notion of "force fields in space" was totally wrong.

As a primary example, there is an "infolded" electrodynamics inside all potentials, fields, and waves that is far more primary than the coarse EM that is in our textbooks. It is actually an infolded general relativity, and this is what the Russians have weaponized. This infolded EM-GR has been ignored in the West (but not in Russia) since 1903.

As another example the potentials, fields, and waves of present electrodynamics are rigorously defined – and exist – only in the presence of mass. They do not exist in such form (force fields and oscillating force fields) at all in space, in the absence of mass. Rigorously, Maxwell's equations are mass-to-mass transforms, since Maxwell and everyone at the time assumed the ubiquitous presence of the material ether. In other words, to the founding electrodynamicists, there was no place in all the universe that was devoid of mass. Hence they defined all EM entities as entities containing mass. Even though Michelson-Morley experiments of more than a century ago destroyed that material ether, not a single Maxwellian equation has ever been changed! They still assume the material ether.

Specifically missing from our electrodynamics are two essential transforms: (1) the mass-to-spacetime transform, and (2) the spacetime-to-mass transform. In other words, these two transforms are "infolded" or hidden inside (comprise) the end point extremum case – the mass-to-mass transform actually written by Maxwell.

But these two missing transforms are just general relativity (they correspond precisely to Wheeler's principle of general relativity). So general relativity has always been infolded inside electrodynamics as we know it, but ignored. Further, this is a very powerful general relativity, since it uses the very strong EM force as the agent of spacetime curvature. Hence now one can engineer a powerful general relativity in the laboratory and in practical devices, and even action-at-a-distance, which one cannot do using merely the staid and weak G-force.

It turns out that longitudinal EM waves as shown by E.T. Whittaker in 1903 and 1904 are far more primary than our present EM theory and even our present GR theory. Both EM and GR are directly engineerable using longitudinal EM waves, including subluminal and superluminal EM waves in vacuum, and including action at a distance. One can engineer complexes of spacetime curvatures as desired, which in turn are *spacetime engines* (Wheeler's term). These engines can act upon mass – either living or inert – in any fashion desired, and at any distance desired.

The little-known and poorly understood Regeneration & Recovery (R&R) system – as contrasted to the immune system – utilizes this exact "spacetime engine" methodology to restore damaged or diseased cells, within its limitations – "dedifferentiating" or "time-reversing" them back to normal. The mechanism is fully explained in the attached package. The proposed program applies this fundamental method to greatly amplify the effect. Thus cancer cells can be reversed back to normal cells, HIV-infected cells (HIV-factories) can be reversed back to normal cells (genetics and all), etc. This represents a dramatic and unparalleled new medical therapeutic methodology of extreme power. The same techniques can simply be conjugated one additional time, and used to produce diseases and cellular changes at a distance, including in entire mass populations, as the KGB is doing right now. Specifically, by impressing an extremely weak "cocktail mix" of disease spacetime engines on the U.S. populace, the immune systems are reacting and being "spread thin" across several "shadow state" infectious pathogenic conditions. Thus when an actual terrorist BW attack occurs, the lethality of the strike will be remarkably enhanced because the immune systems will be "thinned" in their ability to fend off this new pathogen. By doing it this way, BW warfare can actually be conducted upon the U.S. without anyone being the wiser.

Our very survival as a nation is threatened by remarkable KGB weapons using this fully documented, extended electrodynamics. Bluntly, unless we take off our scientific blinders and understand Russian energetics weapon science, GWS and the Embassy health changes in Moscow were just a picnic before what is upon us. After two decades of work on the KGB energetics weapon problem, we have fully deciphered the technical mechanisms and what energetics is, as well as citing many of the Russian weapon tests and how they were done. Again, you can see that for yourself in the package.

Now that we understand the technical mechanisms, we are proposing rapid redevelopment of the Prioré process, in portable suitcase-sized units to be mass produced and filtered down into the entire emergency community for mass treatment of lethal infectious diseases resulting from BW strikes. And we are strongly urging the U.S. scientific community to quickly investigate and come to grips with this vast new *biological engineering* mechanism. Our very survival depends on it.

The inclosed documents provide an overview of the process. There is nothing else on the scientific horizon that can save the majority of all those stricken Americans, once the first foreign-sponsored terrorist teams strike our population centers with BW agents.

Sincerely,

T.E. Bearden
President & CEO

Incls: Full package with table of contents

Weapons Slides

The Tom Bearden Website

Help support the research

"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts." *

Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Slides
(covers period up to 1999)

- The Coming Strategic Attack on the United States
- Outline
- Loss of two separate A-10 Warthogs in 1997
- Typical ABM Defenses (Widely Deployed by mid-70s)

- Launch Phase Anti-Missile System
- Launch Phase Anti-Bomber System
- HC-130s Engines Quit, Aircraft Crashes
  Nov. 22, 1996
- 1996 Actions by this analyst
- Additional Information Exists: Not Analyzed, Not
  Integrated in Briefing
- Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985
- Satellite Photos of Plumes in Soviet Arctic Since 1974
  See "Fer de Lance" for additional information
- Aum Shinrikyo ("Supreme Truth" Cult), Shoko Asahara, Leader

  Aum Shinrikyo - Anatomy of a Dangerous Cult
- Ashkhabad Problem and Its Solution
- Strange Characteristics of Iran Quake, Sep. 1978
- Iran Earthquake, September 1978
- Situation Now (Assessment)
- Situation Now (Assessment 2)
- Aspects of Strong Local Asymmetry
- Mass Casualty Problem: Aerial Anthrax

Spray on Washington, D.C.

- Present Status of the WMD Threat to the United States
- Some Present Problems in the U.S. WMD Defense Capabilities
- Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future
- Attempt to Ban Energetics Weapons Worldwide
- Brezhnev's Schedule
- Large glowing Tesla globe witnessed in Red China by hundreds
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  Sequence of Events
- Cold Explosion -- 9 April 1984 Near Kurils, and formation and growth of spherical shell (globe)
  What They Were
- Cold Explosion -- 9 April 1984 Near Kurils (Plus additional weapons effects)
- Countering Insane MAD Systems:
  Find Loophole, Exploit It
- Next Generation Counter to Both Sane and Insane MAD systems
- Simultaneous Dudding of Nuclear Weapons Worldwide

- Simultaneous Dudding of Nuclear Weapons Worldwide
- Operation Divine Wind II
- Advanced EMBW Applications
- EM Biological Warfare (EMBW)
- Microwave Radiation of U.S. Embassy in Moscow
- Pentagon EMI Studies, 1989
- EM Missile Practice (6 slides)
- Subdivisions of Soviet Energetics Program
- Communism's Goal
- Manuilsky's Goal vs. Status Today (Dec. 1996)
- Professor Hellman, German Scientist Released from Work in Soviet Union
- Hemispheres and Globes of Light (4 slides)
- How Could Such Superweapons Ever Be Concealed?
- Soviets Easily Concealed These Superweapons for Four Decades
- Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes
- Longitudinal Wave Interferometry: Endothermic

- Longitudinal Wave Interferometry: Exothermic
- U.S. Air Attack on Libya, April 1986
- Lightning Wall
  circa 1959-1960
- Lisitsyn's Report: Brain Code Broken
- Large Longitudinal Wave Interferometer (LWI) Accident
- Large Longitudinal Wave Interferometer (LWI) Accident
- Use of Hidden Information Content of the Field Can Provide Action-At-A-Distance
- Two Types of Mutual Assured Destruction (MAD) System
- Psychoenergetics Weapons Teams Mentally Disable All Personnel
- Strikes Against Special Ship and Energetics Weapons Sites
- "Mindsnapper" Attack on Special Ships and Energetics Weapons Sites
- How Much Off Guard Are We?
  (7 Slides)
- Possible Targets: 1995-9 Escalation Phase
- Propagation of Weapon Effects
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15th, 1979

- Countering an Insane QP Weapon
- Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases
- Self-Targeting in Inner EM Channel Can Produce a Quantum Potential
- Mechanism for Producing a Quantum Potential
- Major Principles for Use of Quantum Potential Weapons
- Major Principles for Use of Quantum Potential Weapons (2)
- Instantaneous Communication by a Quantum Potential
- Participants in a quantum potential share a common multiply-connected spacetime (MCST)
- Quake in Tangshan, China
  28 July, 1976
- Exchanging Earthquakes
- Instant Communication by a Quantum Potential
- Westerners Do Not React to Slow Threats
- Characterizing a Given Risk or Risk Factor
- Technical Risk Management
- Aum Shinrikyo's Sarin Facility

- In Scalar Beam Interference Zone
- Physicist Schappeller Auralaz Munster, Austria
- Initial Soviet Plan for Cold War, After WWII
- Josef Stalin: Absolute Dictator
- Exploding any Resistance by Operational Scientists
- Significant Soviet Statements
- Captain Svoboda Dives Headlong to Her Death
- T-polarized Beam Projector
- Possible Tests (1)
- Possible Tests (2)
- Possible Tests (3)
- Possible Tests (4)
- Mushroom Cloud Rising From Sea
- Launches from Cape Canaveral and Vandenberg AFB Provided Practice Targets of Opportunity
- The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth
- Embodiment of an instrument array sufficient to discriminate when time-density charging has occurred, and that transduction

of time-charge decay has occurred.

- Time-Polarized EM Wave Interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone
- Transmutation of Radioactive Nuclei at a Distance
- Transmutation of Radioactive Nuclei
- Typical Tesla Shield, Northern Route of Japan
- A Curious Tide in Human Affairs
- Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958
- The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction
- The KGB/Yakuza-Aum War Plan
- Why Western Science is Ignorant of KGB Energetics
- Transmutation of Radioactive Nuclei at a Distance
- TheYakuza
- Yugoslav Earthquake, Followed by Cold Explosion
  April 15, 1979

---

* Note that the Secretary did not speak the words in square blocks; these were added by the writers later,

probably to "soften the tone" of the Secretary's information release. So the Secretary confirmed that some nations of the world -- not just some ragged terrorists, as the writers tried to imply for spin control -- do possess novel electromagnetic weapons and are using them to induce earthquakes, engineer the weather and climate, and trigger volcanic eruptions.

THE COMING STRATEGIC ATTACK ON THE UNITED STATES

By an AUM/Yakuza/KGB/Communist Coalition Using Energetics SuperWeapons

T.E. Bearden, LTC
U.S. Army (Retired)
M.S. Nuclear Eng.
Fellow Emeritus, AIAS

With Technical Background and History of Energetics Weapons

Partial support by the Association of Distinguished American Scientists is gratefully acknowledged.d

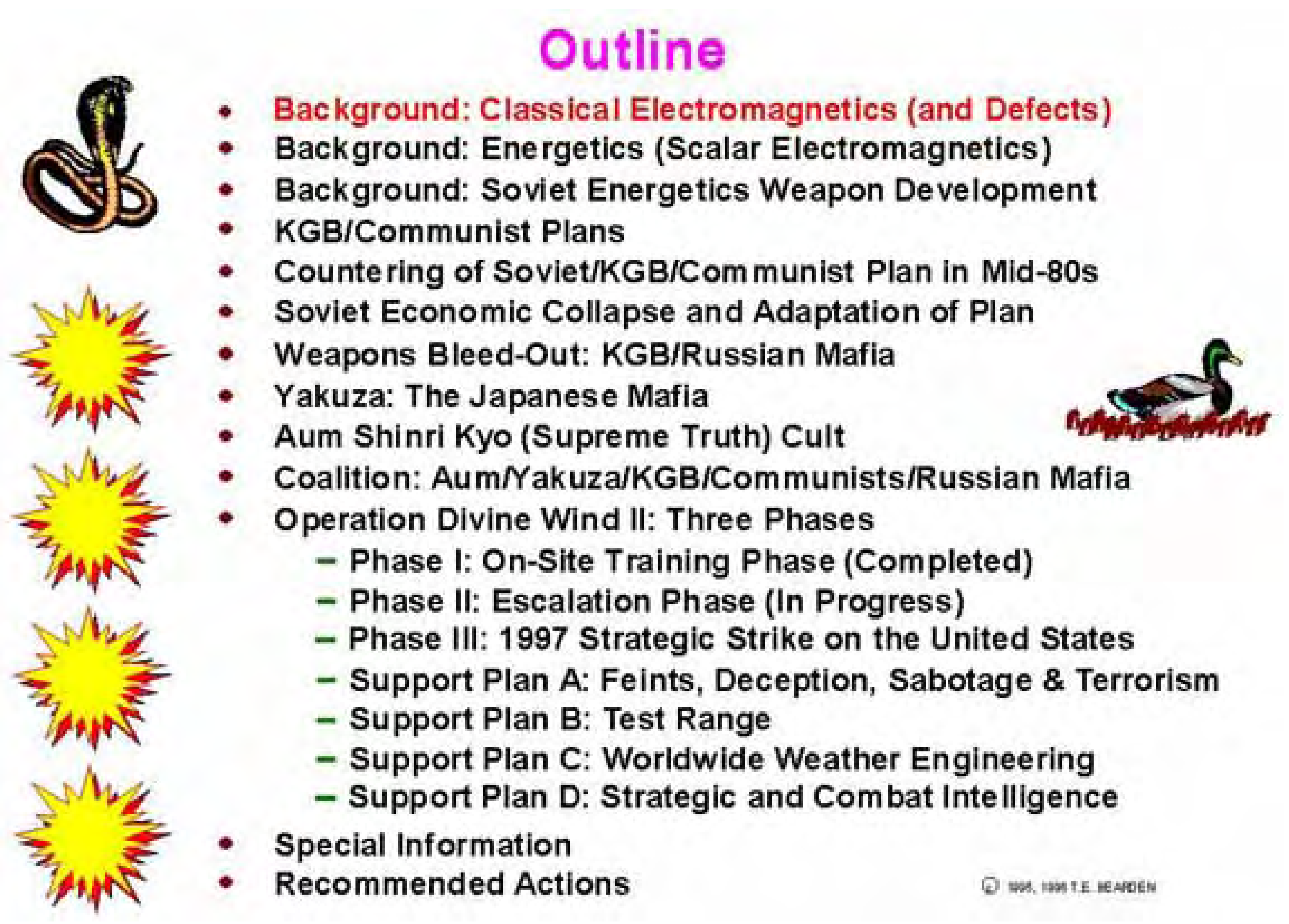

Outline

- Background: Classical Electromagnetics (and Defects)
- Background: Energetics (Scalar Electromagnetics)
- Background: Soviet Energetics Weapon Development
- KGB/Communist Plans
- Countering of Soviet/KGB/Communist Plan in Mid-80s
- Soviet Economic Collapse and Adaptation of Plan
- Weapons Bleed-Out: KGB/Russian Mafia
- Yakuza: The Japanese Mafia
- Aum Shinri Kyo (Supreme Truth) Cult
- Coalition: Aum/Yakuza/KGB/Communists/Russian Mafia
- Operation Divine Wind II: Three Phases
  - Phase I: On-Site Training Phase (Completed)
  - Phase II: Escalation Phase (In Progress)
  - Phase III: 1997 Strategic Strike on the United States
  - Support Plan A: Feints, Deception, Sabotage & Terrorism
  - Support Plan B: Test Range
  - Support Plan C: Worldwide Weather Engineering
  - Support Plan D: Strategic and Combat Intelligence
- Special Information
- Recommended Actions

© 1996, 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

Loss of two separate A-10 Warthogs in 1997

- A-10 Mysterious Flyaway and Crash - April 2, 1997
- Death of Second A-10 Pilot - May 27, 1997
- Captain Button's Mysterious Flyaway and Eventual Crash
- Death of Second A-10 Pilot - May 27, 1997

A-10 Mysterious Flyaway and Crash

April 2, 1997

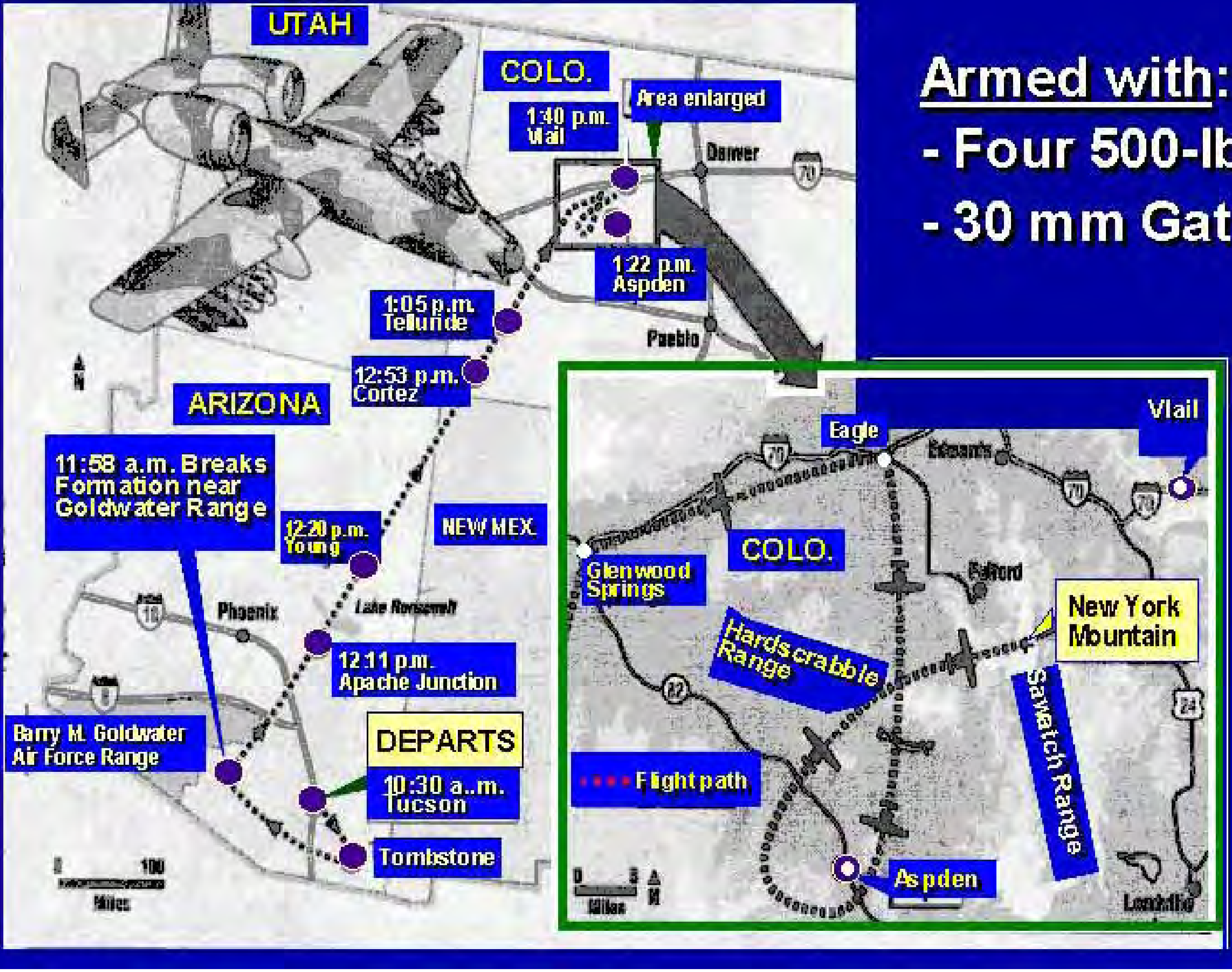
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
Denver
70
1:22 p.m.
Aspden
1:05 p.m.
Telluride
Pueblo
12:53 p.m.
Cortez
ARIZONA
11:58 a.m. Breaks
Formation near
Goldwater Range
12:20 p.m.
Young
NEW MEX.
Phoenix
10
Lake Roosevelt
12:11 p.m.
Apache Junction
8
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m.
Tucson
Tombstone
0 100
Miles
N
Eagle
Edwards
Vail
70
70
70
COLO.
Glenwood
Springs
82
New York
Mountain
Hardscrabble
Range
Sawatch Range
24
Flight path
Aspden
0 5
Miles
N
Leadville

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

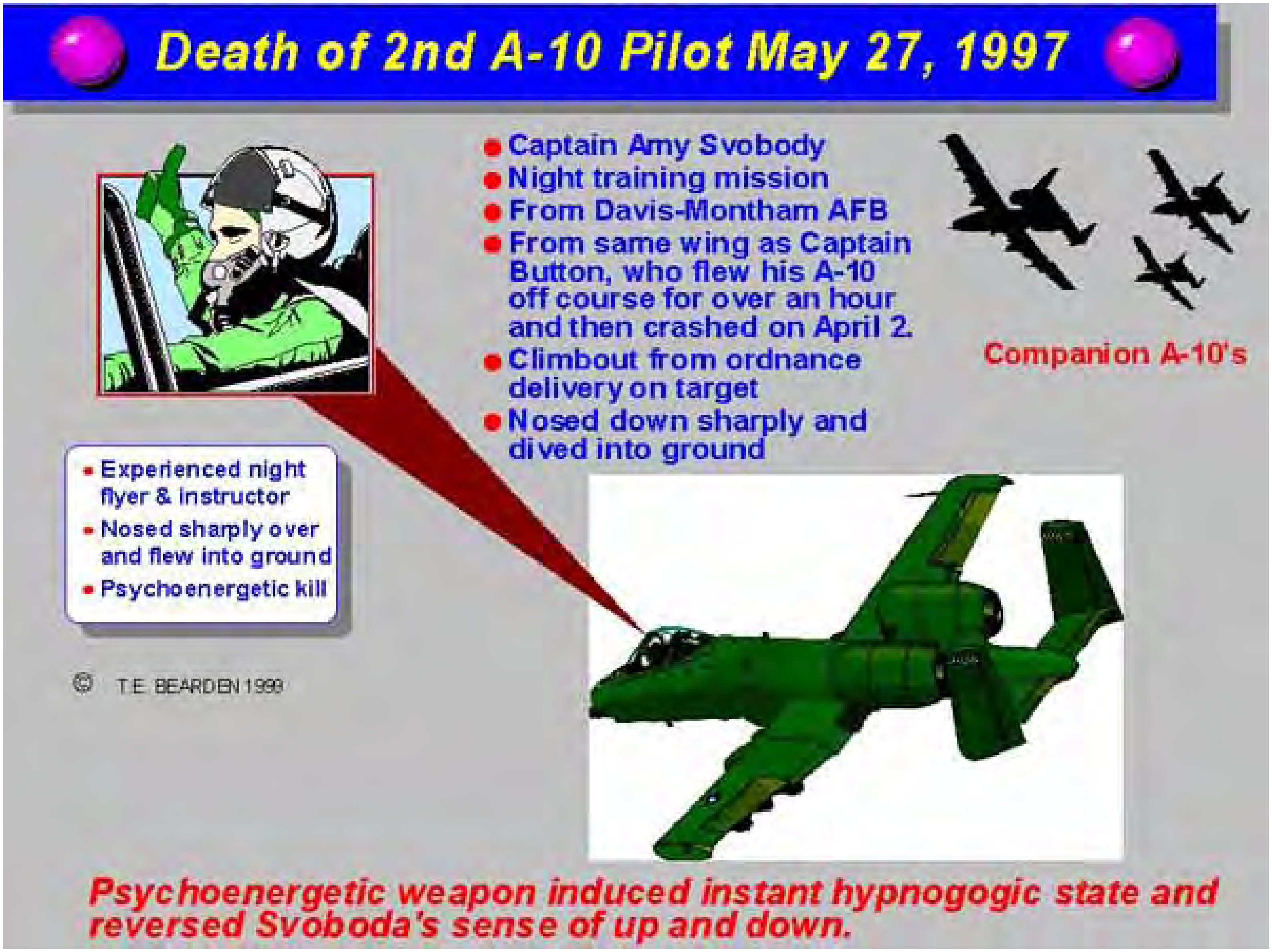

Death of 2nd A-10 Pilot May 27, 1997

- Captain Amy Svobody
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Climbout from ordnance delivery on target
- Nosed down sharply and dived into ground

Companion A-10's

- Experienced night flyer & instructor
- Nosed sharply over and flew into ground
- Psychoenergetic kill

© T.E. BEARDEN 1999

Psychoenergetic weapon induced instant hypnogogic state and reversed Svoboda's sense of up and down.

A-10 Mysterious Flyaway and Crash

April 2, 1997

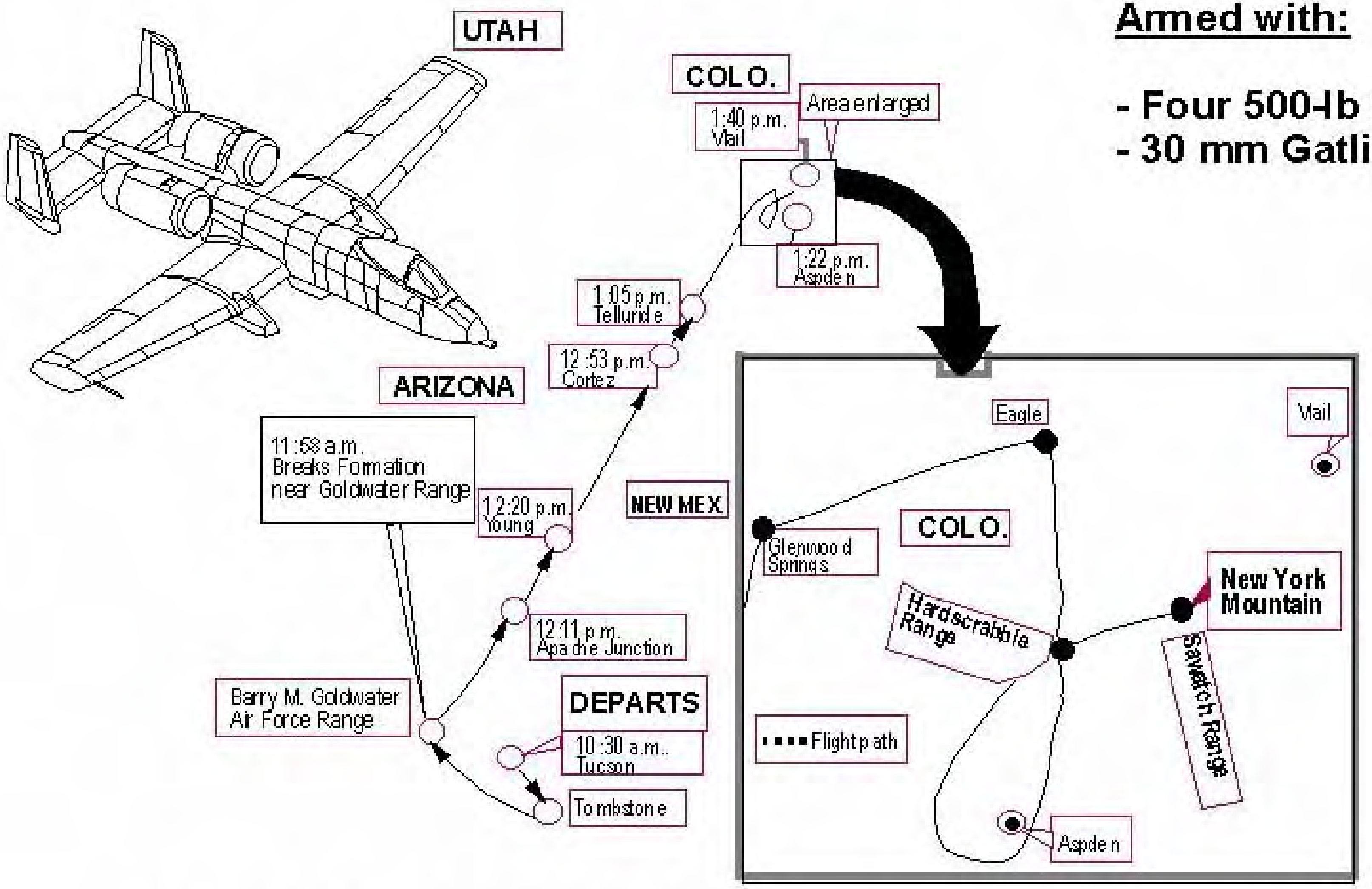
UTAH
COLO.
1:40 p.m. Vail
Area enlarged
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
Glenwood Springs
COLO.
Hardscrabble Range
New York Mountain
Sawatch Range
Flight path
Aspden

Armed with:

- Four 500-lb bombs
- 30 mm Gatling gun

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

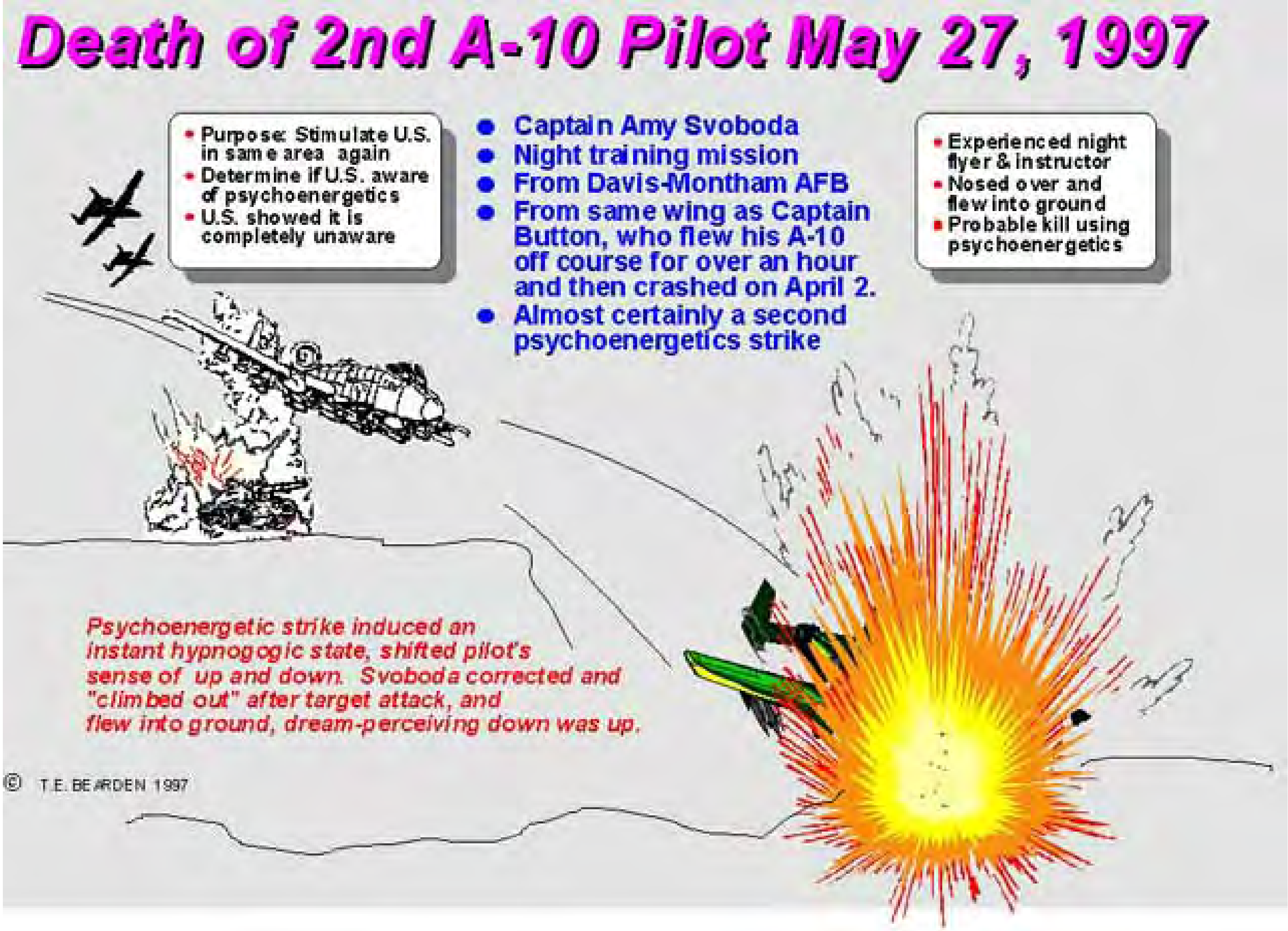

Death of 2nd A-10 Pilot May 27, 1997

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware

- Captain Amy Svoboda
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Almost certainly a second psychoenergetics strike

- Experienced night flyer & instructor
- Nosed over and flew into ground
- Probable kill using psychoenergetics

Psychoenergetic strike induced an instant hypnogogic state, shifted pilot's sense of up and down. Svoboda corrected and "climbed out" after target attack, and flew into ground, dream-perceiving down was up.

© T.E. BEARDEN 1997

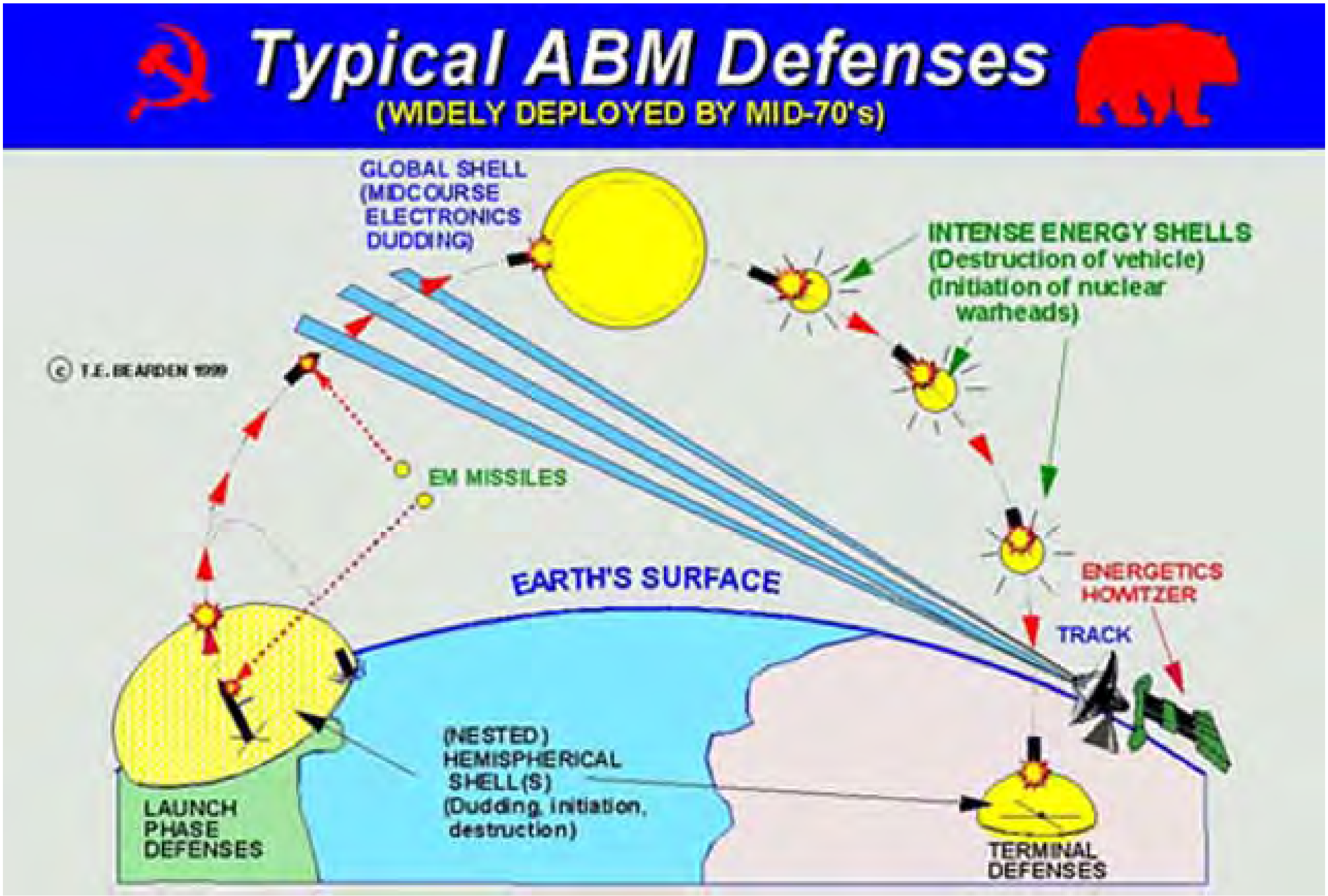
Typical ABM Defenses
(WIDELY DEPLOYED BY MID-70's)
GLOBAL SHELL (MIDCOURSE ELECTRONICS DUDDING)
INTENSE ENERGY SHELLS (Destruction of vehicle) (Initiation of nuclear warheads)
© T.E. BEARDEN 1999
EM MISSILES
EARTH'S SURFACE
ENERGETICS HOWITZER
TRACK
(NESTED) HEMISPHERICAL SHELL(S) (Dudding, initiation, destruction)
LAUNCH PHASE DEFENSES
TERMINAL DEFENSES

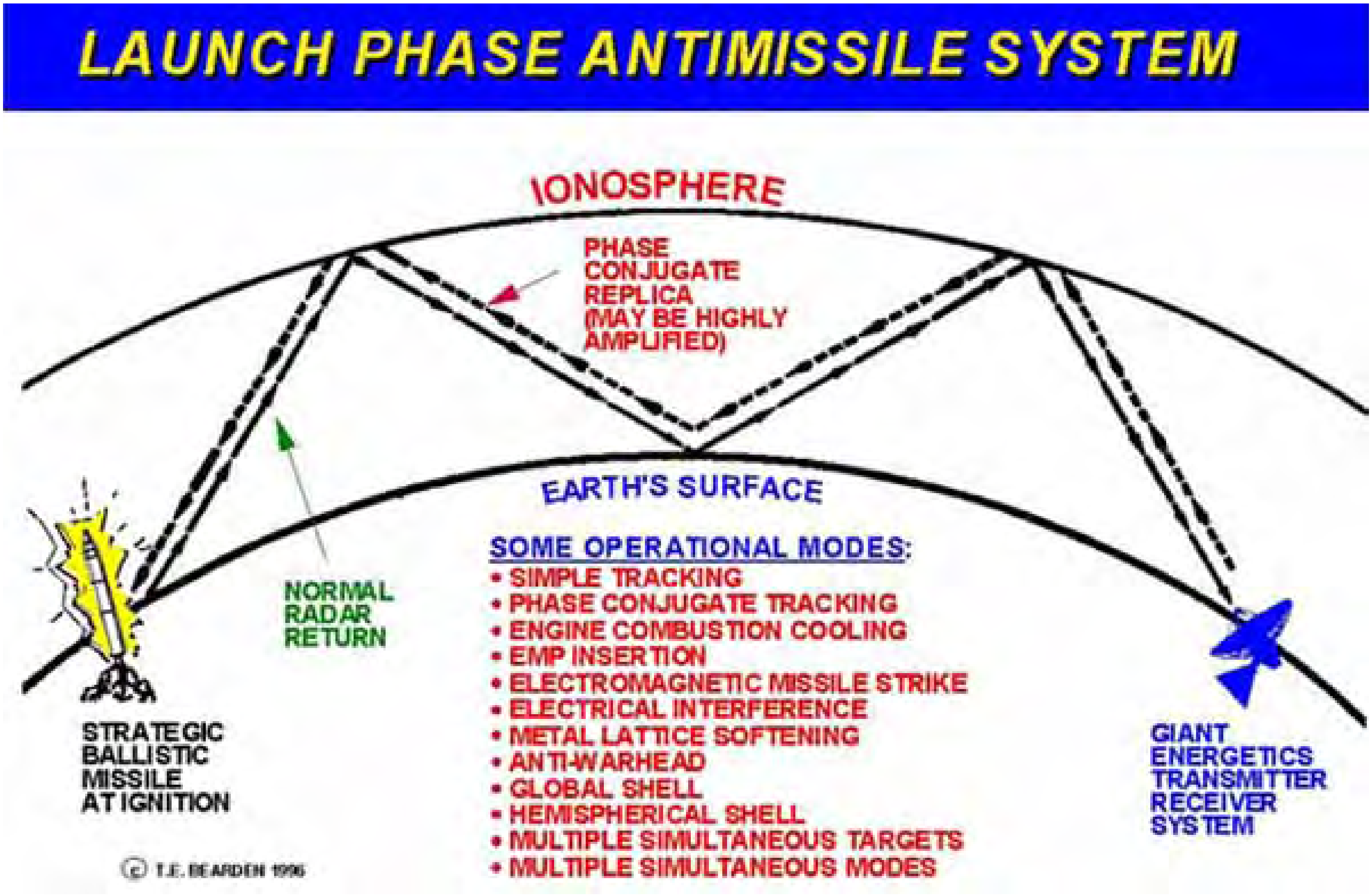

LAUNCH PHASE ANTIMISSILE SYSTEM

IONOSPHERE

PHASE
CONJUGATE
REPLICA
(MAY BE HIGHLY
AMPLIFIED)

EARTH'S SURFACE

NORMAL
RADAR
RETURN

STRATEGIC
BALLISTIC
MISSILE
AT IGNITION

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-WARHEAD
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

GIANT
ENERGETICS
TRANSMITTER
RECEIVER
SYSTEM

© T.E. BEARDEN 1996

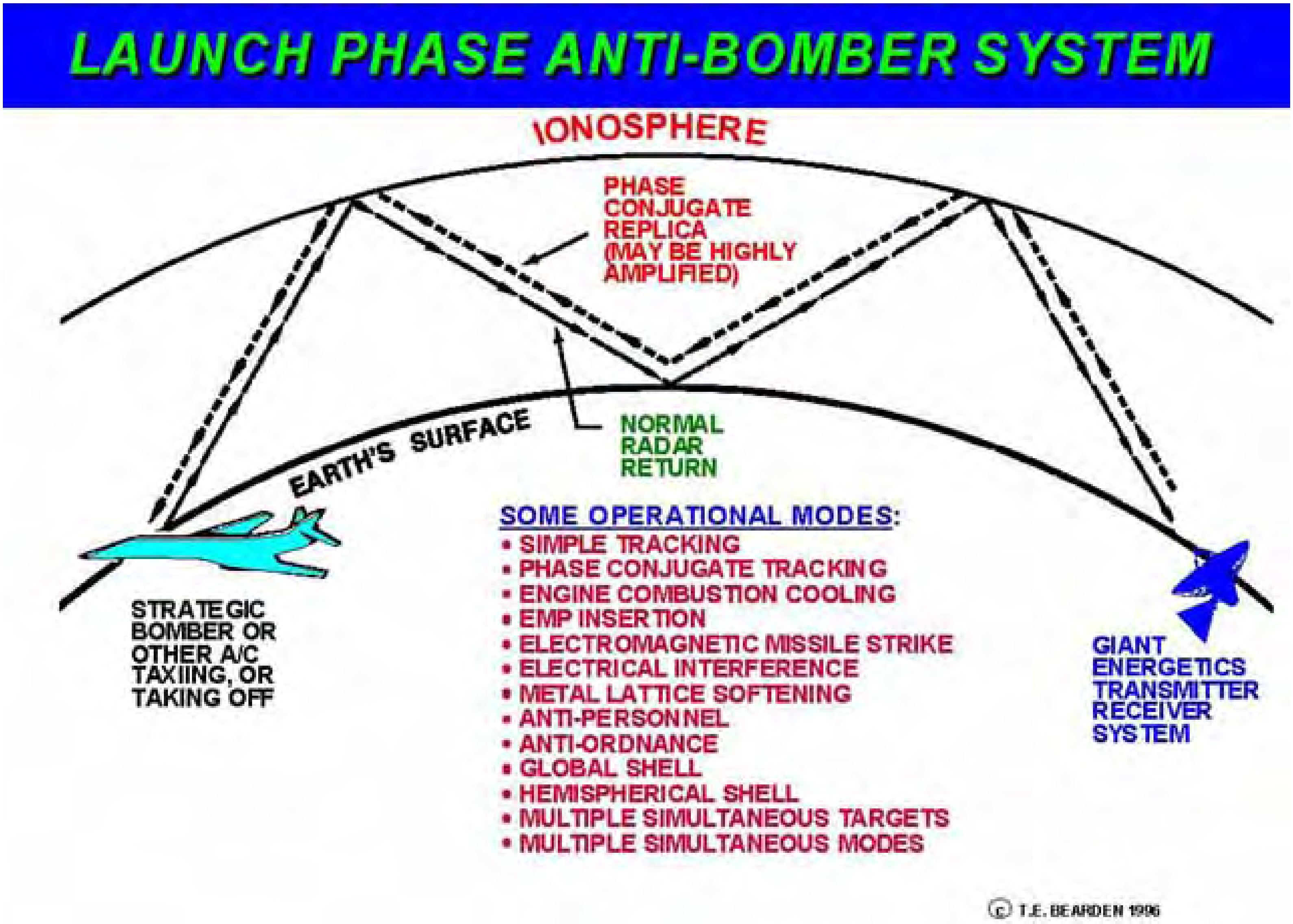

LAUNCH PHASE ANTI-BOMBER SYSTEM

IONOSPHERE
PHASE CONJUGATE REPLICA (MAY BE HIGHLY AMPLIFIED)
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER OR OTHER A/C TAXIING, OR TAKING OFF
GIANT ENERGETICS TRANSMITTER RECEIVER SYSTEM

SOME OPERATIONAL MODES:

- SIMPLE TRACKING
- PHASE CONJUGATE TRACKING
- ENGINE COMBUSTION COOLING
- EMP INSERTION
- ELECTROMAGNETIC MISSILE STRIKE
- ELECTRICAL INTERFERENCE
- METAL LATTICE SOFTENING
- ANTI-PERSONNEL
- ANTI-ORDNANCE
- GLOBAL SHELL
- HEMISPHERICAL SHELL
- MULTIPLE SIMULTANEOUS TARGETS
- MULTIPLE SIMULTANEOUS MODES

© T.E. BEARDEN 1996

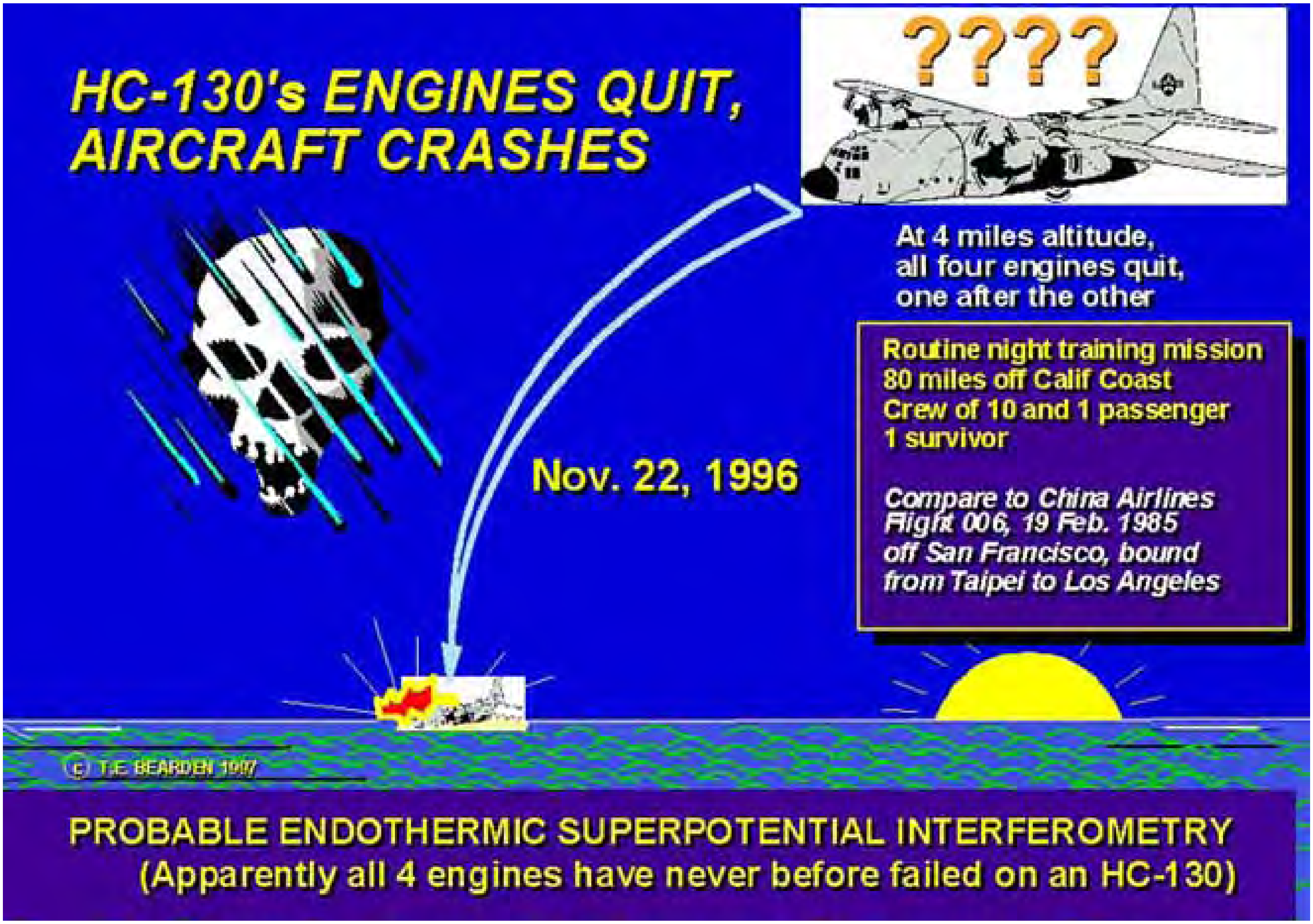

HC-130's ENGINES QUIT, AIRCRAFT CRASHES

At 4 miles altitude,
all four engines quit,
one after the other

Routine night training mission
80 miles off Calif Coast
Crew of 10 and 1 passenger
1 survivor

Compare to China Airlines Flight 006, 19 Feb. 1985 off San Francisco, bound from Taipei to Los Angeles

Nov. 22, 1996

© T.E. BEARDEN 1997

PROBABLE ENDOTHERMIC SUPERPOTENTIAL INTERFEROMETRY
(Apparently all 4 engines have never before failed on an HC-130)

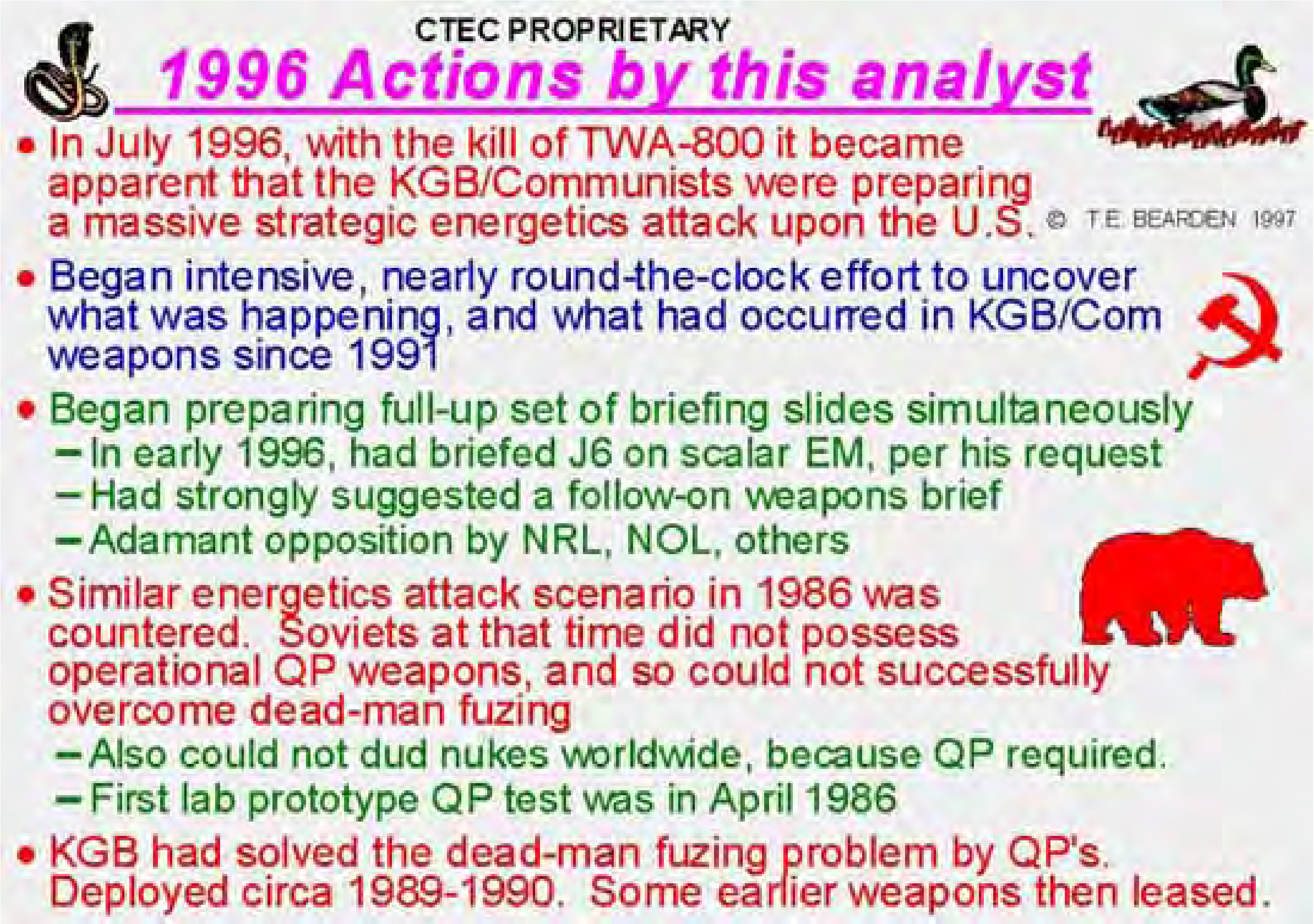

CTEC PROPRIETARY

1996 Actions by this analyst

- In July 1996, with the kill of TWA-800 it became apparent that the KGB/Communists were preparing a massive strategic energetics attack upon the U.S.

© T.E. BEARDEN 1997

- Began intensive, nearly round-the-clock effort to uncover what was happening, and what had occurred in KGB/Com weapons since 1991
- Began preparing full-up set of briefing slides simultaneously
  - In early 1996, had briefed J6 on scalar EM, per his request
  - Had strongly suggested a follow-on weapons brief
  - Adamant opposition by NRL, NOL, others
- Similar energetics attack scenario in 1986 was countered. Soviets at that time did not possess operational QP weapons, and so could not successfully overcome dead-man fuzing
  - Also could not dud nukes worldwide, because QP required.
  - First lab prototype QP test was in April 1986
- KGB had solved the dead-man fuzing problem by QP's. Deployed circa 1989-1990. Some earlier weapons then leased.

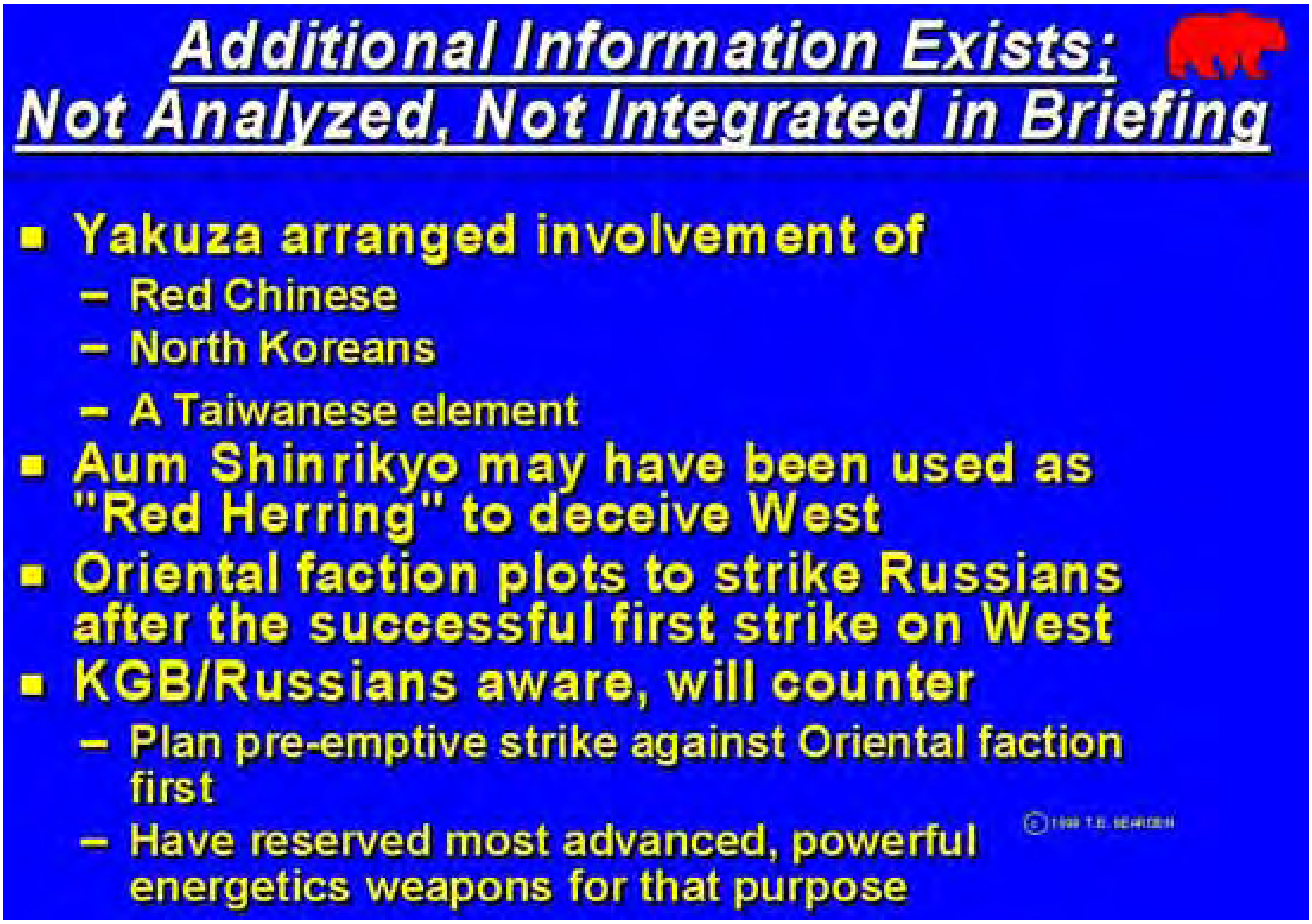

Additional Information Exists; Not Analyzed, Not Integrated in Briefing

- Yakuza arranged involvement of
  - Red Chinese
  - North Koreans
  - A Taiwanese element
- Aum Shinrikyo may have been used as "Red Herring" to deceive West
- Oriental faction plots to strike Russians after the successful first strike on West
- KGB/Russians aware, will counter
  - Plan pre-emptive strike against Oriental faction first
  - Have reserved most advanced, powerful energetics weapons for that purpose

The Tom Bearden Website

Help support the research

Kill of Arrow DC-8, Gander AFB, Dec. 12, 1985

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6

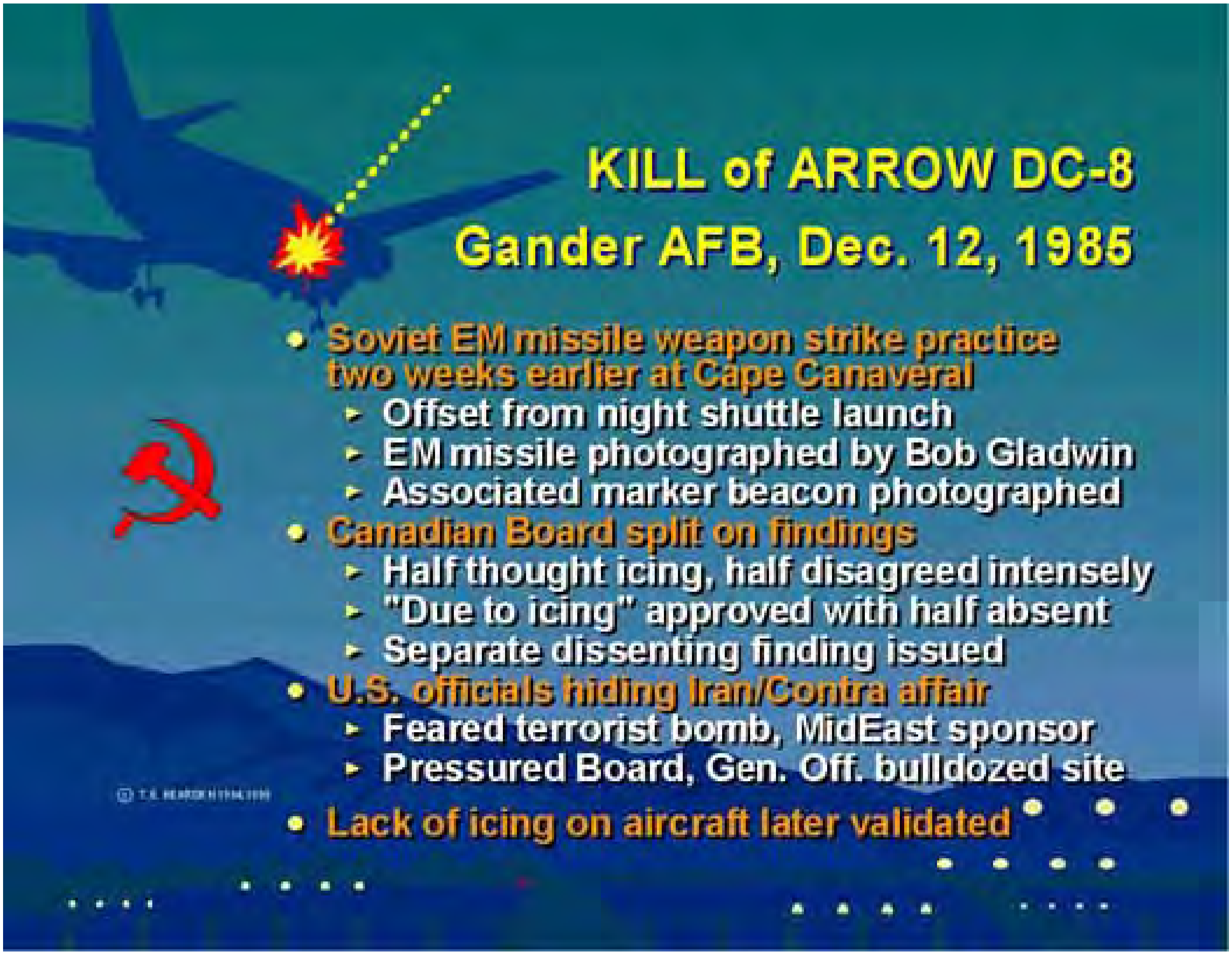

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. bulldozed site
- Lack of icing on aircraft later validated

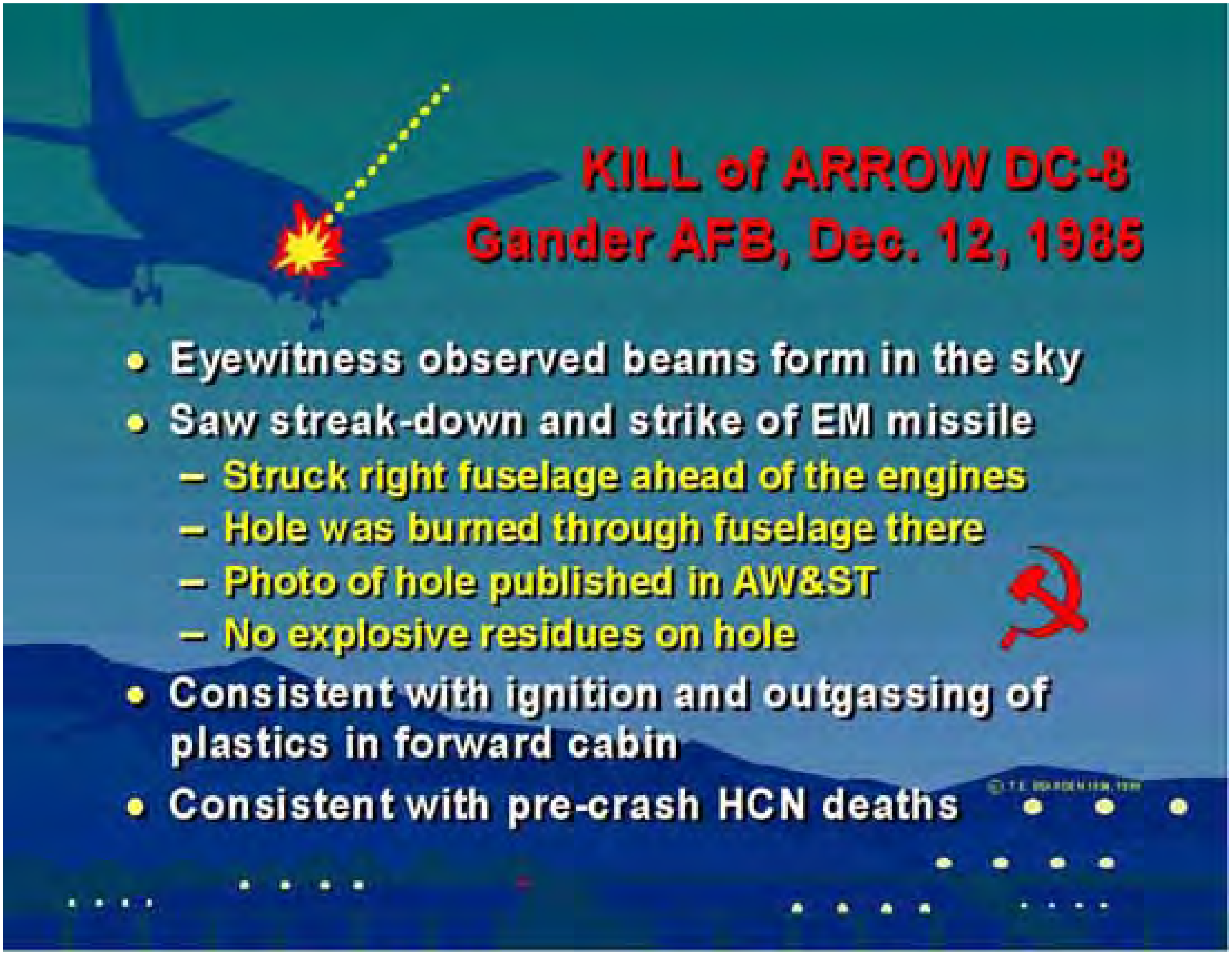

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Eyewitness observed beams form in the sky
- Saw streak-down and strike of EM missile
  - Struck right fuselage ahead of the engines
  - Hole was burned through fuselage there
  - Photo of hole published in AW&ST
  - No explosive residues on hole
- Consistent with ignition and outgassing of plastics in forward cabin
- Consistent with pre-crash HCN deaths

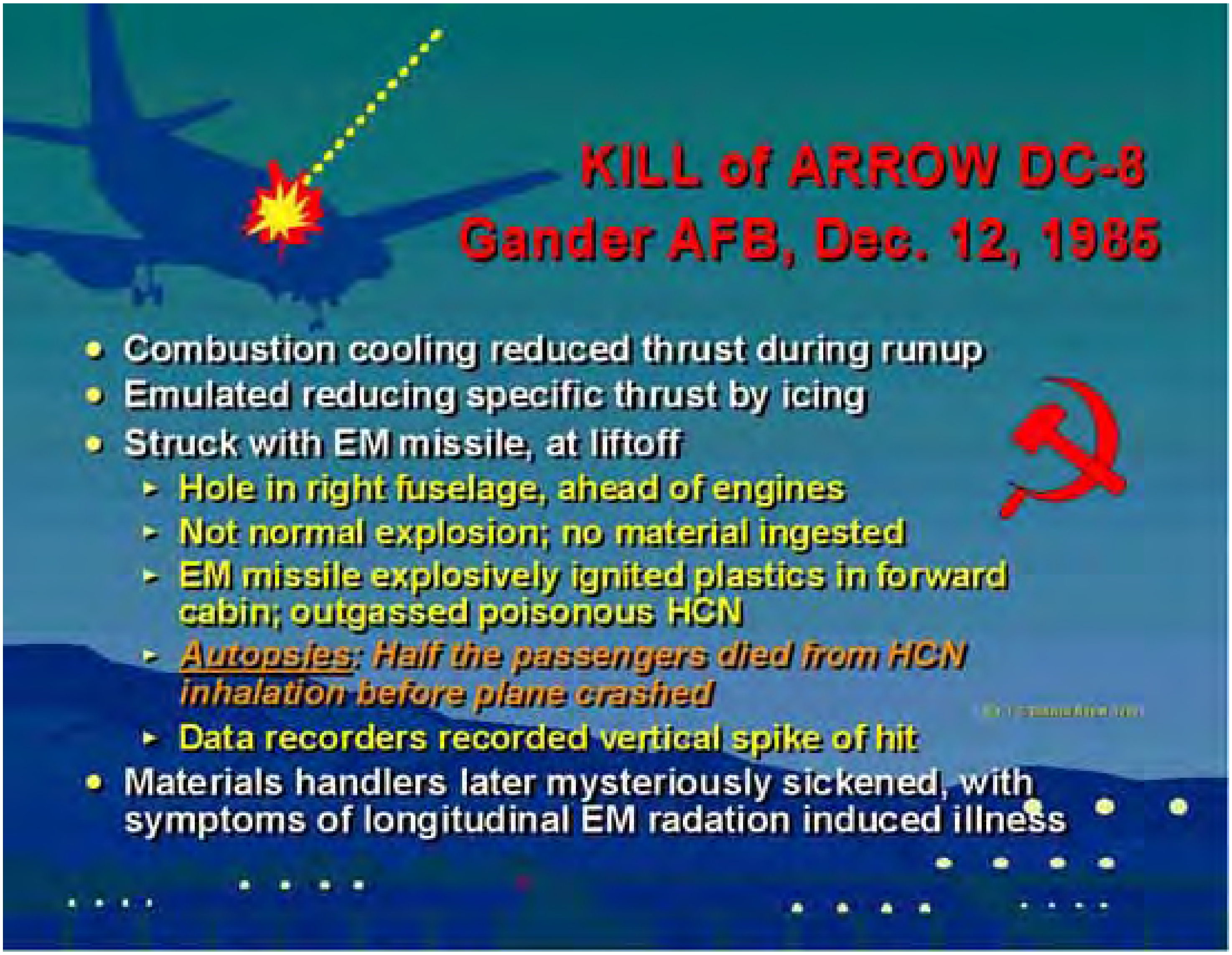

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of longitudinal EM radation induced illness

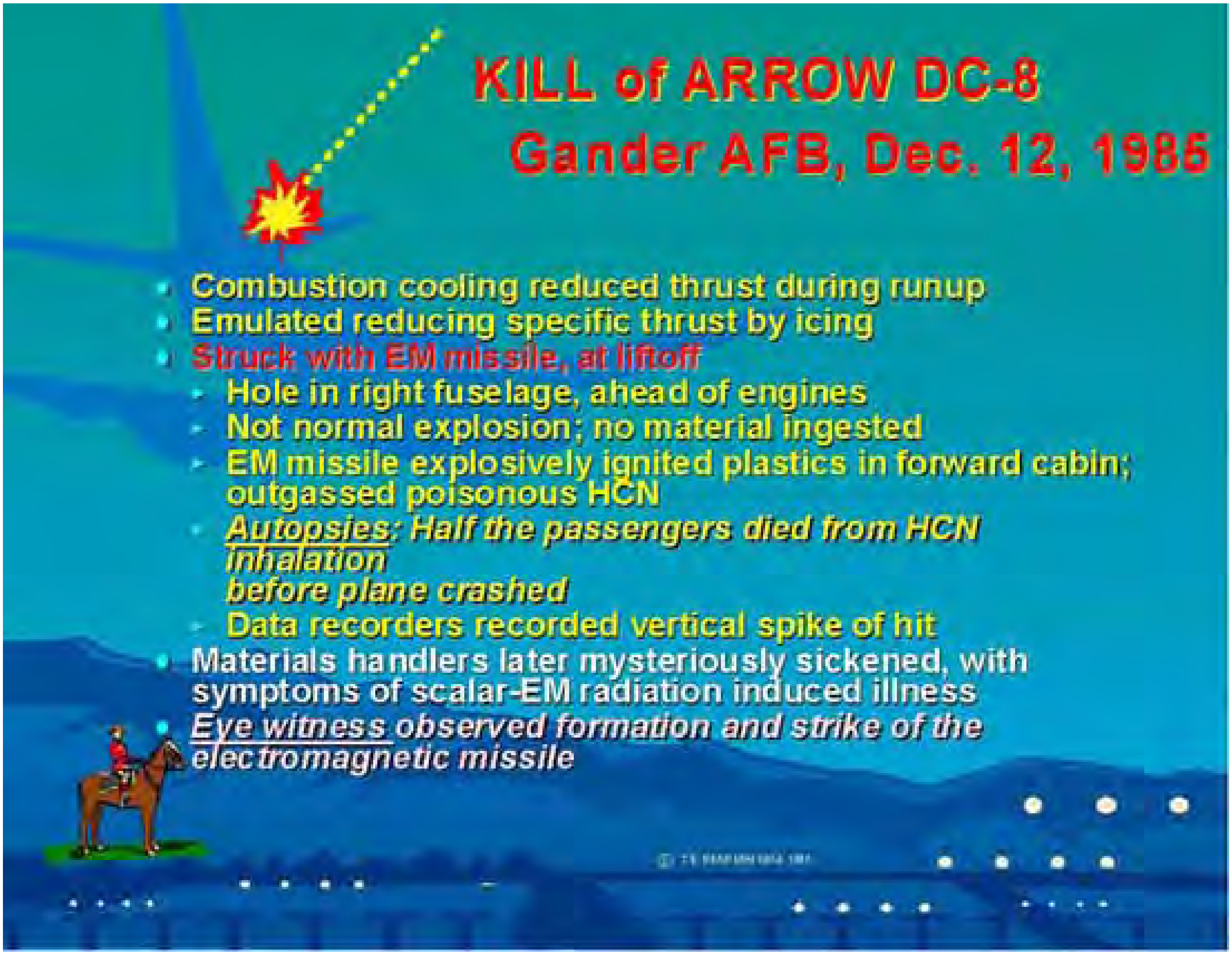

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Combustion cooling reduced thrust during runup
- Emulated reducing specific thrust by icing
- Struck with EM missile, at liftoff
  - Hole in right fuselage, ahead of engines
  - Not normal explosion; no material ingested
  - EM missile explosively ignited plastics in forward cabin; outgassed poisonous HCN
  - *Autopsies: Half the passengers died from HCN inhalation before plane crashed*
  - Data recorders recorded vertical spike of hit
- Materials handlers later mysteriously sickened, with symptoms of scalar-EM radiation induced illness
- *Eye witness observed formation and strike of the electromagnetic missile*

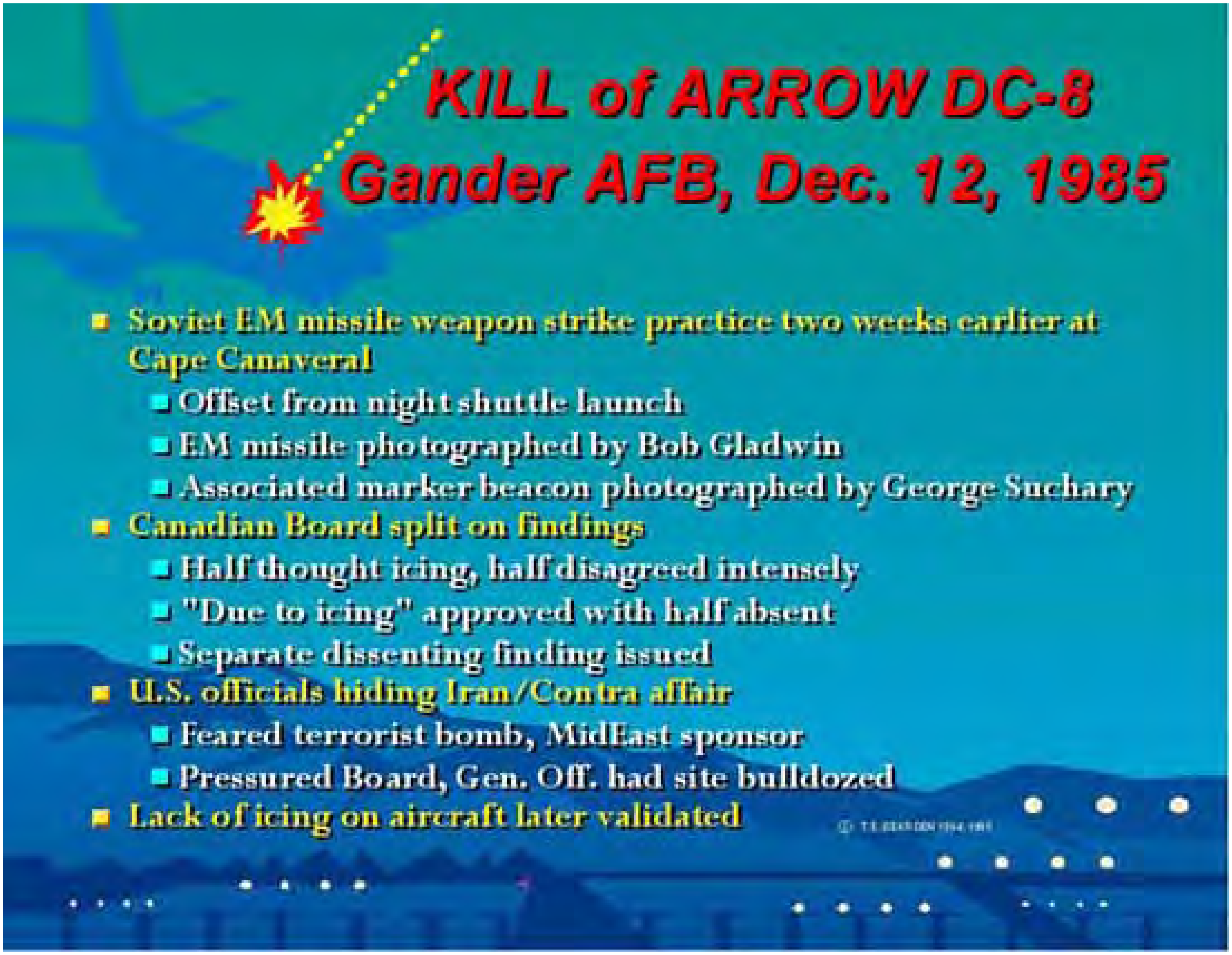

KILL of ARROW DC-8
Gander AFB, Dec. 12, 1985

- Soviet EM missile weapon strike practice two weeks earlier at Cape Canaveral
  - Offset from night shuttle launch
  - EM missile photographed by Bob Gladwin
  - Associated marker beacon photographed by George Suchary
- Canadian Board split on findings
  - Half thought icing, half disagreed intensely
  - "Due to icing" approved with half absent
  - Separate dissenting finding issued
- U.S. officials hiding Iran/Contra affair
  - Feared terrorist bomb, MidEast sponsor
  - Pressured Board, Gen. Off. had site bulldozed
- Lack of icing on aircraft later validated

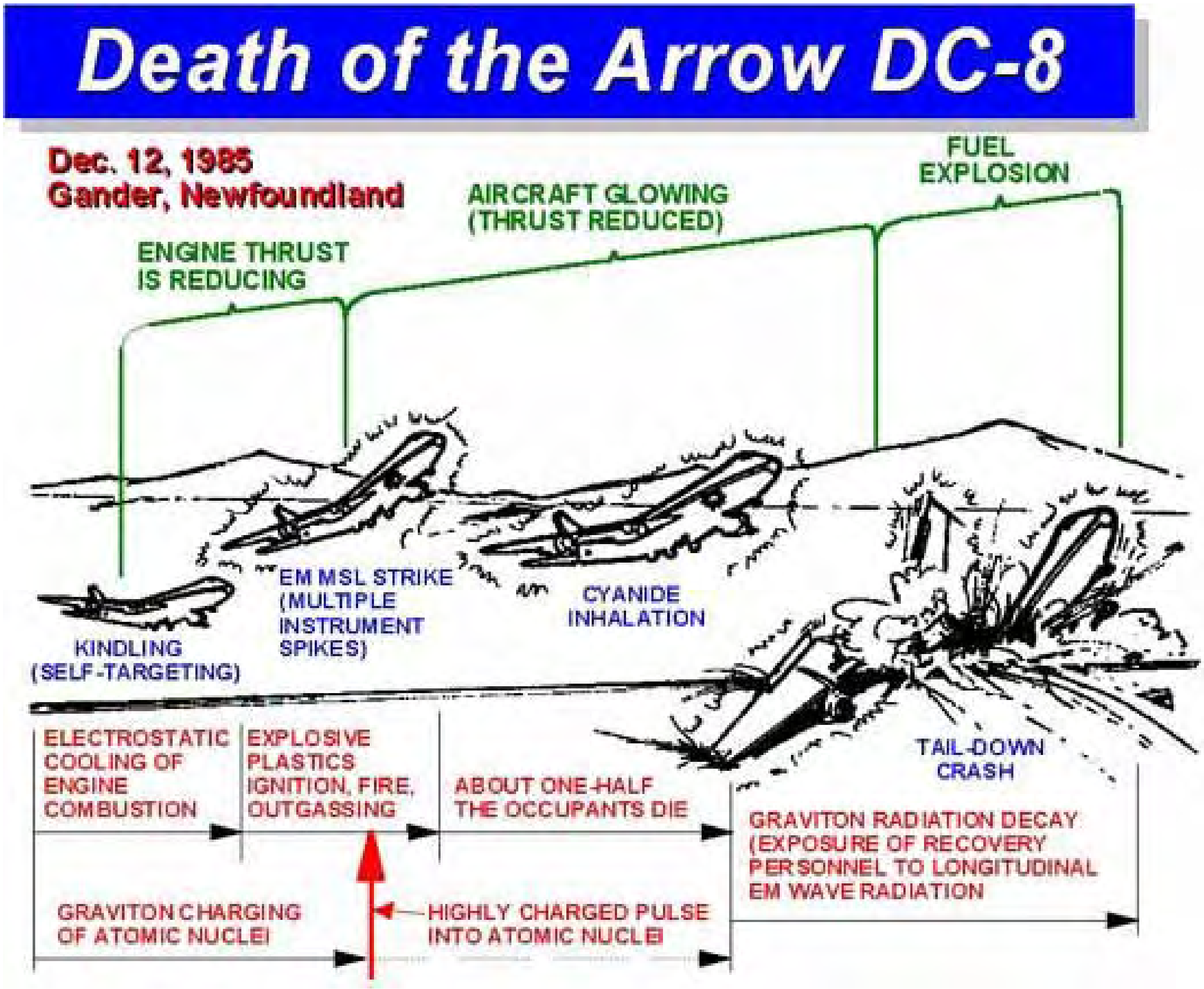
Death of the Arrow DC-8
Dec. 12, 1985
Gander, Newfoundland
ENGINE THRUST IS REDUCING
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
KINDLING (SELF-TARGETING)
EM MSL STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT ONE-HALF THE OCCUPANTS DIE
GRAVITON RADIATION DECAY (EXPOSURE OF RECOVERY PERSONNEL TO LONGITUDINAL EM WAVE RADIATION
GRAVITON CHARGING OF ATOMIC NUCLEI
HIGHLY CHARGED PULSE INTO ATOMIC NUCLEI

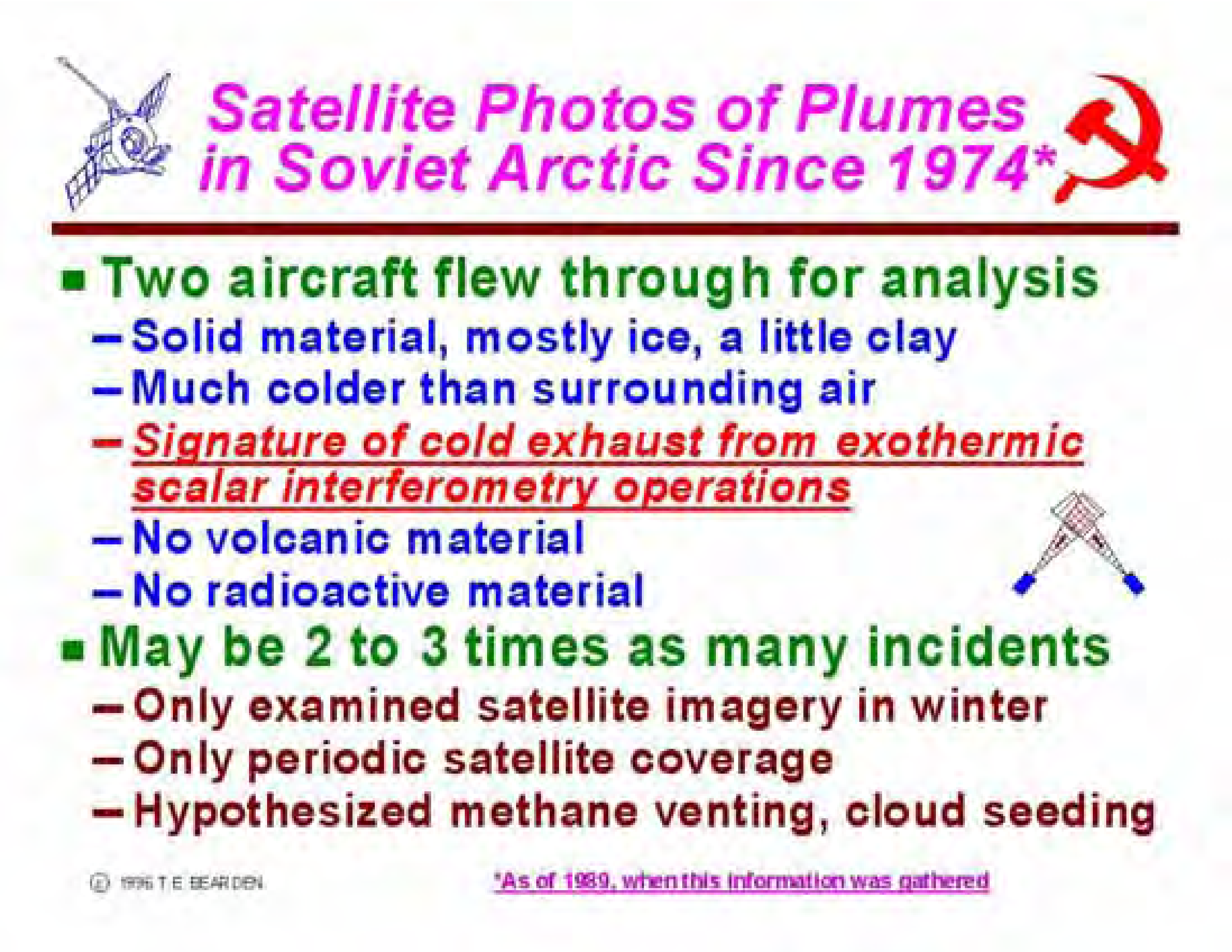

Satellite Photos of Plumes in Soviet Arctic Since 1974*

- Two aircraft flew through for analysis
  - Solid material, mostly ice, a little clay
  - Much colder than surrounding air
  - *Signature of cold exhaust from exothermic scalar interferometry operations*
  - No volcanic material
  - No radioactive material
- May be 2 to 3 times as many incidents
  - Only examined satellite imagery in winter
  - Only periodic satellite coverage
  - Hypothesized methane venting, cloud seeding

© 1996 T.E. BEARDEN

*As of 1989, when this information was gathered

AUM SHINRIKYO

("SUPREME TRUTH" CULT)

Shoko Asahara, Leader

© 1995, 1998 T.E. BEARDEN

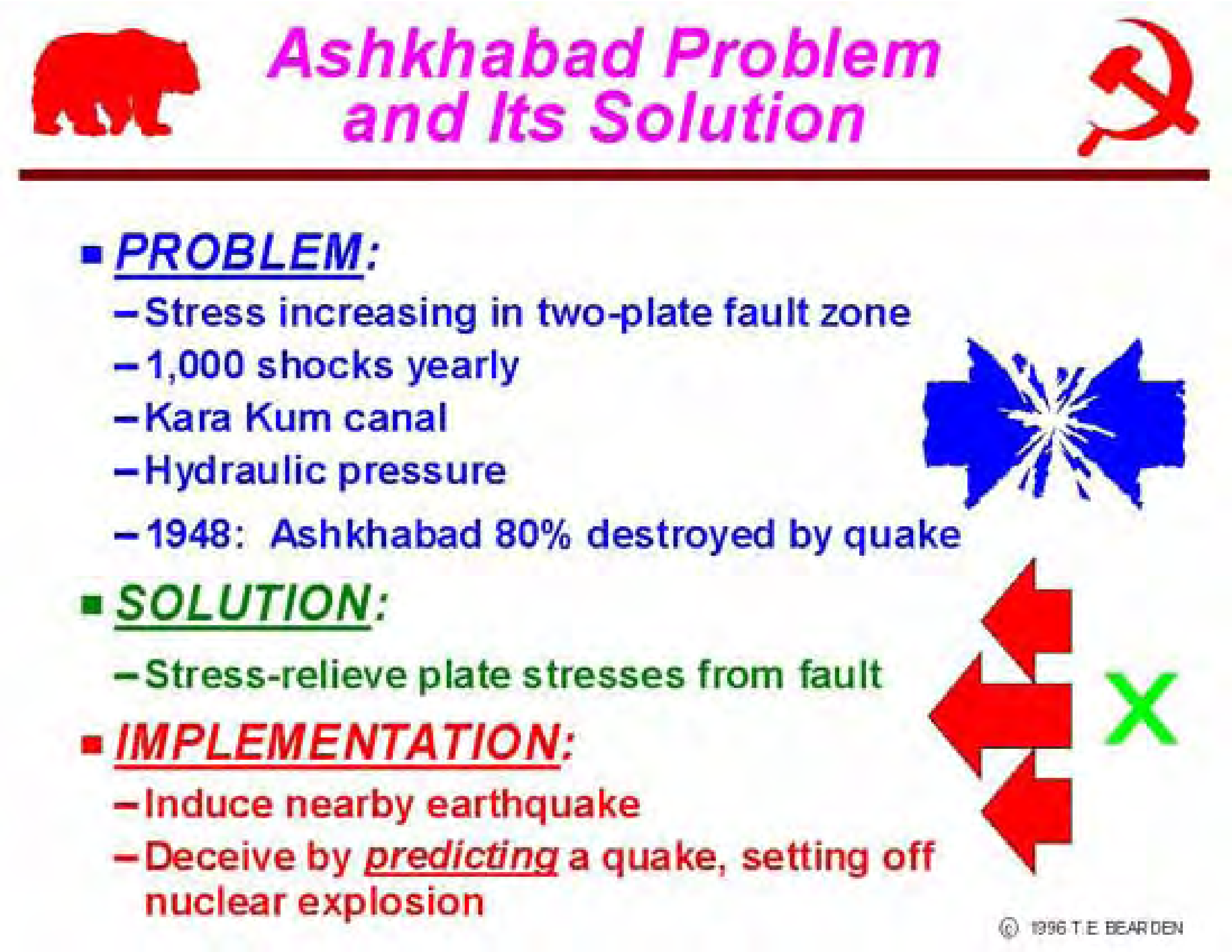

Ashkhabad Problem and Its Solution

- ***PROBLEM:***
  - Stress increasing in two-plate fault zone
  - 1,000 shocks yearly
  - Kara Kum canal
  - Hydraulic pressure
  - 1948: Ashkhabad 80% destroyed by quake
- ***SOLUTION:***
  - Stress-relieve plate stresses from fault
- ***IMPLEMENTATION:***
  - Induce nearby earthquake
  - Deceive by *predicting* a quake, setting off nuclear explosion

© 1996 T.E. BEARDEN

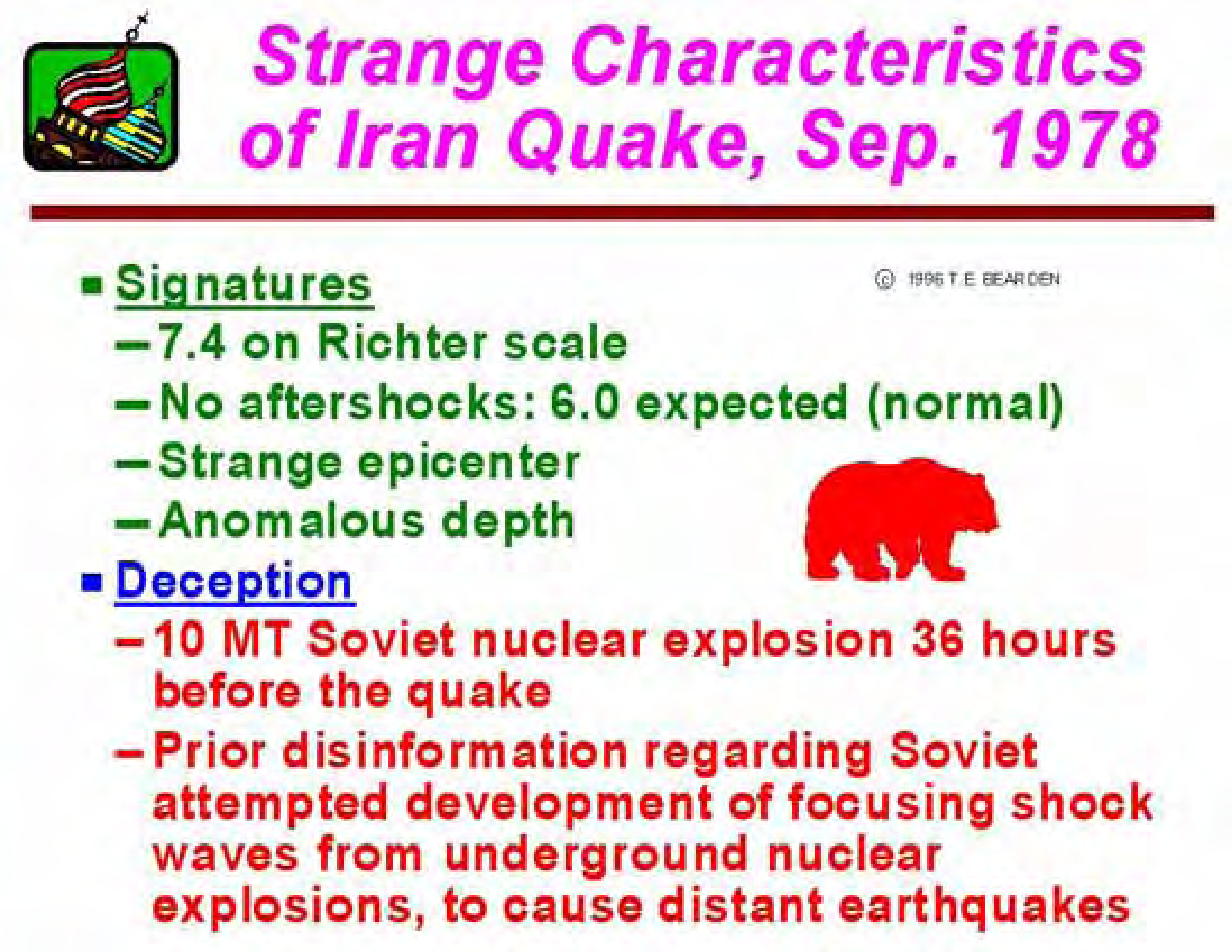

Strange Characteristics of Iran Quake, Sep. 1978

© 1996 T.E. BEARDEN

- Signatures
  - 7.4 on Richter scale
  - No aftershocks: 6.0 expected (normal)
  - Strange epicenter
  - Anomalous depth
- Deception
  - 10 MT Soviet nuclear explosion 36 hours before the quake
  - Prior disinformation regarding Soviet attempted development of focusing shock waves from underground nuclear explosions, to cause distant earthquakes

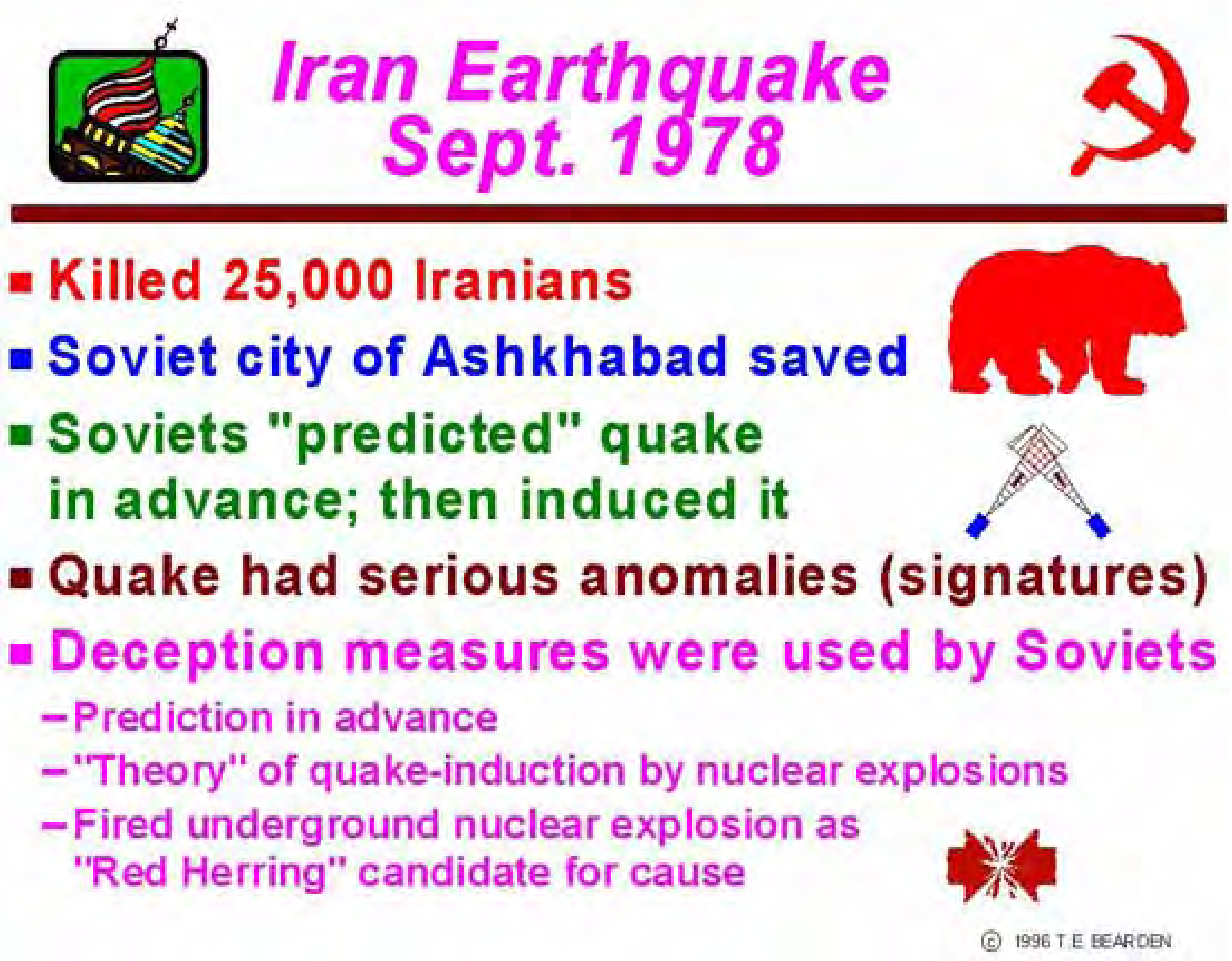

Iran Earthquake Sept. 1978

- Killed 25,000 Iranians
- Soviet city of Ashkhabad saved
- Soviets "predicted" quake in advance; then induced it
- Quake had serious anomalies (signatures)
- Deception measures were used by Soviets
  - Prediction in advance
  - "Theory" of quake-induction by nuclear explosions
  - Fired underground nuclear explosion as "Red Herring" candidate for cause

© 1996 T.E. BEARDEN

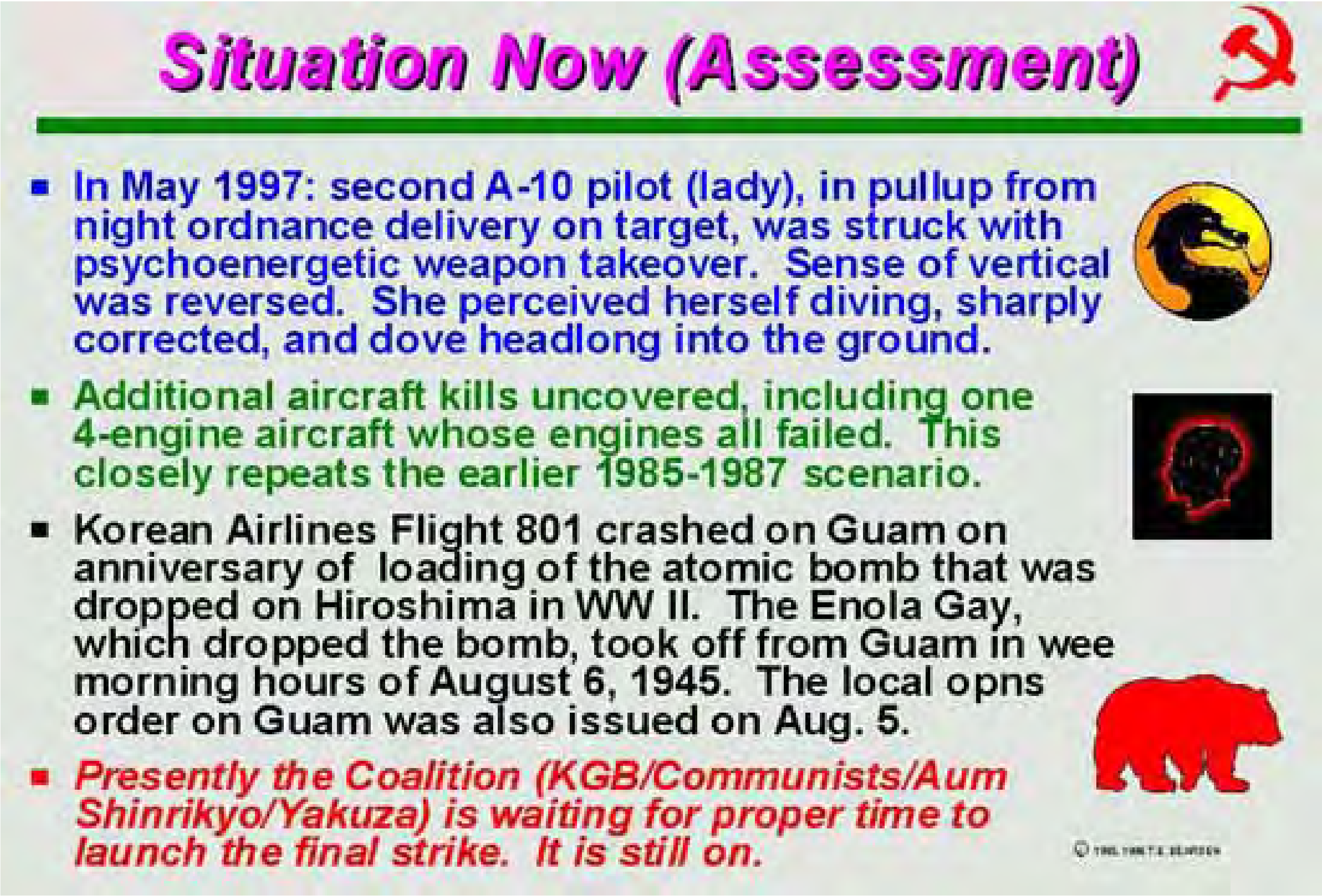

Situation Now (Assessment)

- In May 1997: second A-10 pilot (lady), in pullup from night ordnance delivery on target, was struck with psychoenergetic weapon takeover. Sense of vertical was reversed. She perceived herself diving, sharply corrected, and dove headlong into the ground.
- Additional aircraft kills uncovered, including one 4-engine aircraft whose engines all failed. This closely repeats the earlier 1985-1987 scenario.
- Korean Airlines Flight 801 crashed on Guam on anniversary of loading of the atomic bomb that was dropped on Hiroshima in WW II. The Enola Gay, which dropped the bomb, took off from Guam in wee morning hours of August 6, 1945. The local opns order on Guam was also issued on Aug. 5.
- *Presently the Coalition (KGB/Communists/Aum Shinrikyo/Yakuza) is waiting for proper time to launch the final strike. It is still on.*

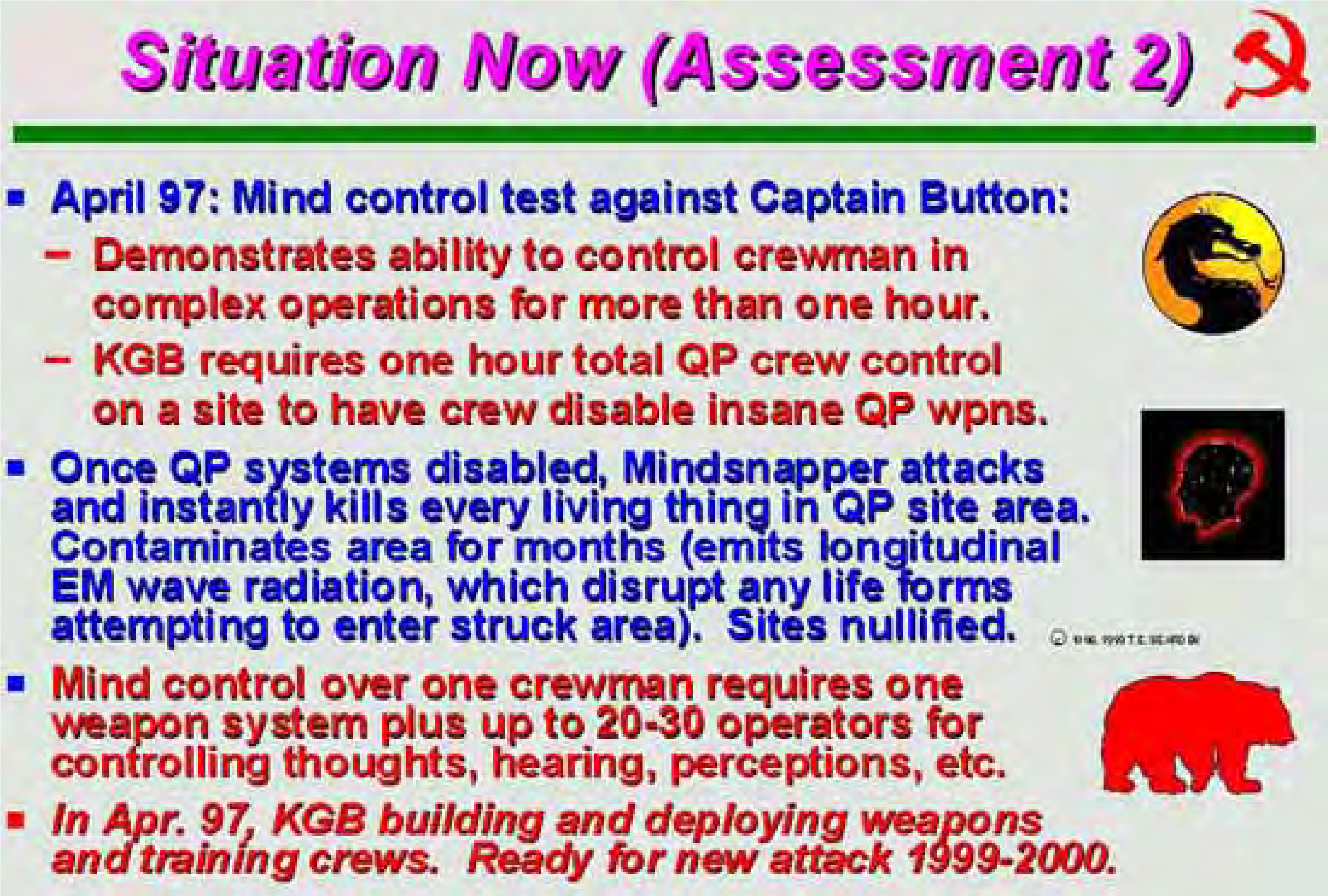

Situation Now (Assessment 2)

- April 97: Mind control test against Captain Button:
  - Demonstrates ability to control crewman in complex operations for more than one hour.
  - KGB requires one hour total QP crew control on a site to have crew disable insane QP wpns.
- Once QP systems disabled, Mindsnapper attacks and instantly kills every living thing in QP site area. Contaminates area for months (emits longitudinal EM wave radiation, which disrupt any life forms attempting to enter struck area). Sites nullified.
- Mind control over one crewman requires one weapon system plus up to 20-30 operators for controlling thoughts, hearing, perceptions, etc.
- *In Apr. 97, KGB building and deploying weapons and training crews. Ready for new attack 1999-2000.*

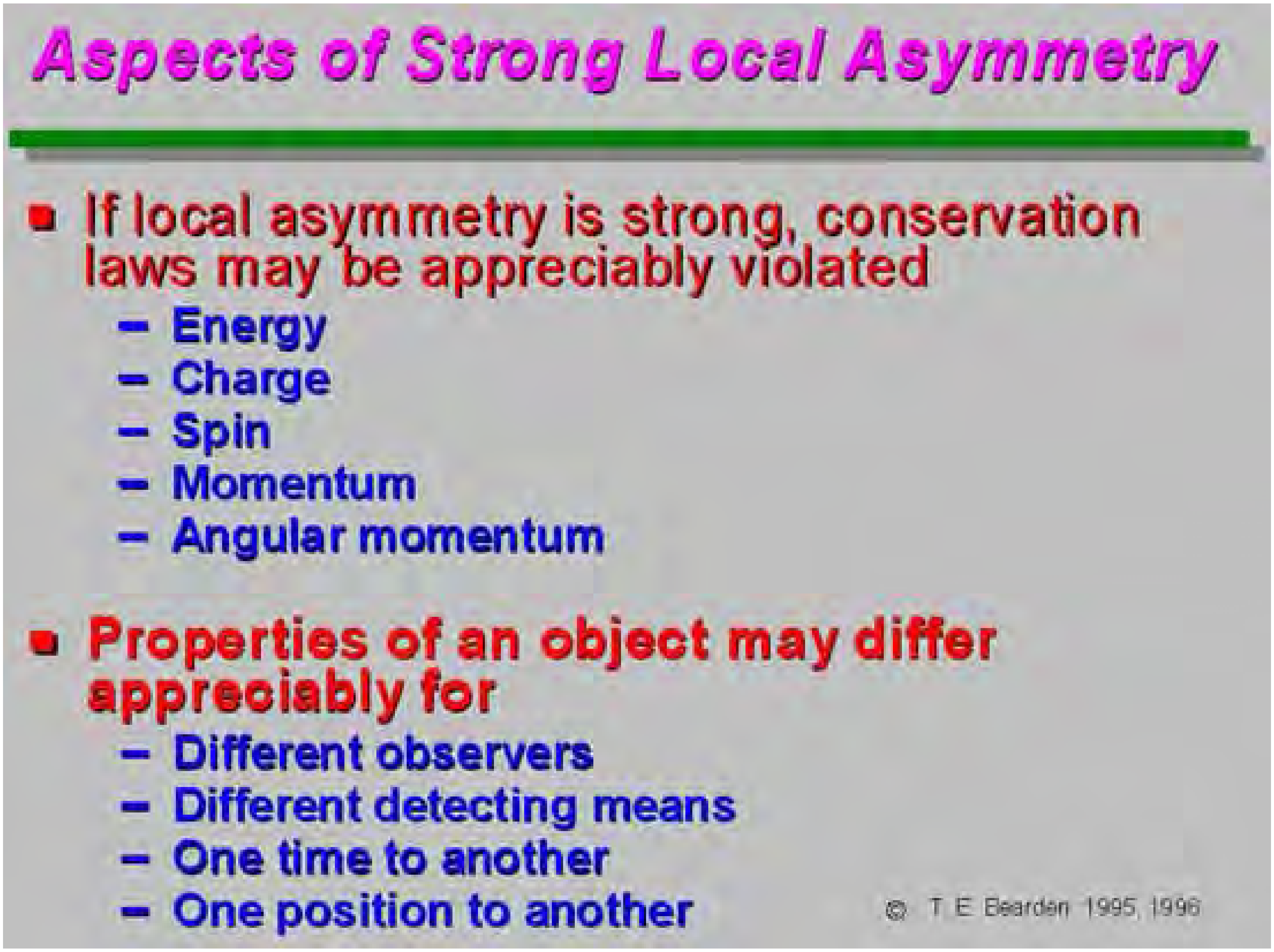

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum
- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another

© T. E. Bearden 1995, 1996

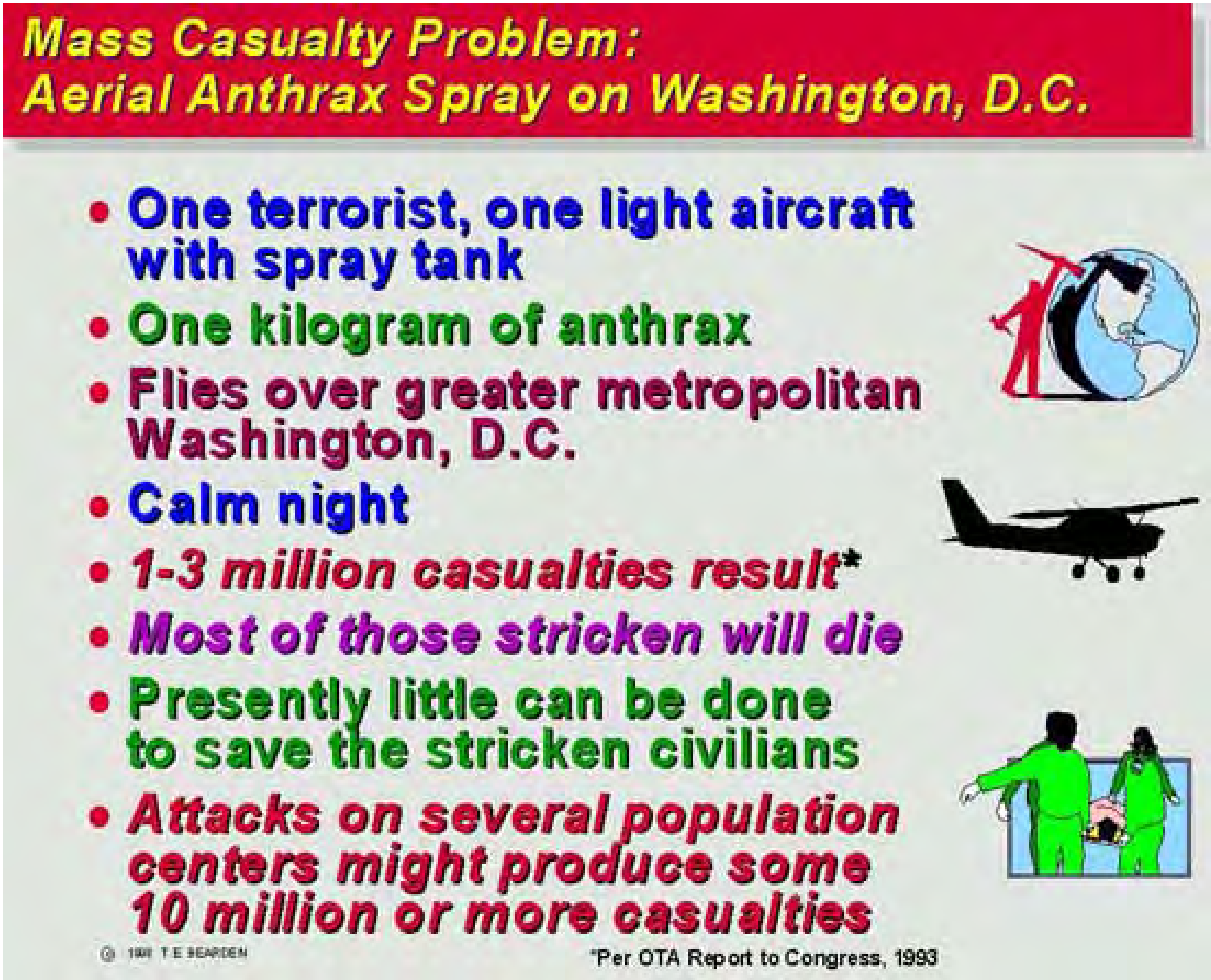

Mass Casualty Problem: Aerial Anthrax Spray on Washington, D.C.

- One terrorist, one light aircraft with spray tank
- One kilogram of anthrax
- Flies over greater metropolitan Washington, D.C.
- Calm night
- *1-3 million casualties result**
- *Most of those stricken will die*
- Presently little can be done to save the stricken civilians
- *Attacks on several population centers might produce some 10 million or more casualties*

© [illegible] T E BEARDEN

*Per OTA Report to Congress, 1993

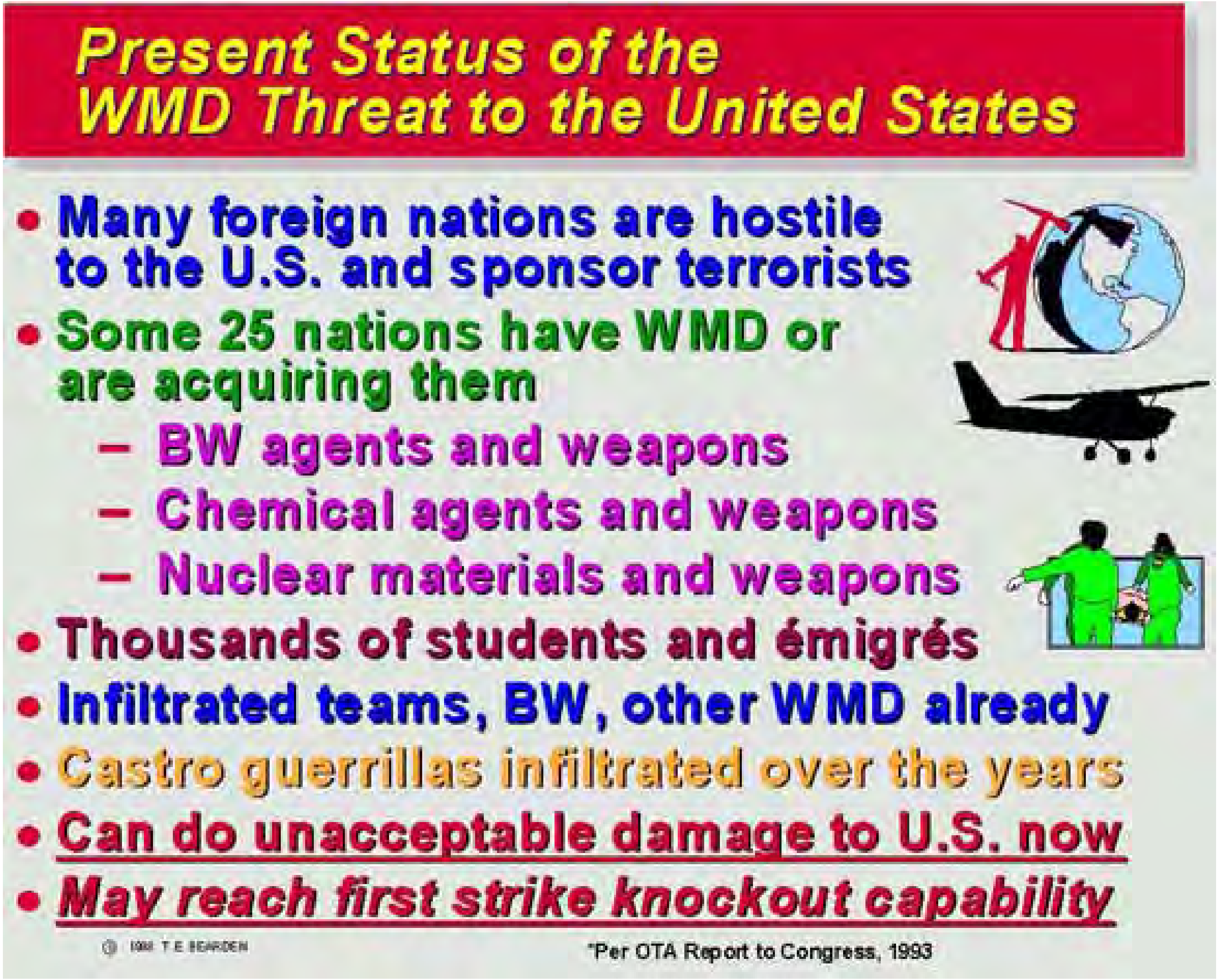

Present Status of the WMD Threat to the United States

- **Many foreign nations are hostile to the U.S. and sponsor terrorists**
- **Some 25 nations have WMD or are acquiring them**
  - **BW agents and weapons**
  - **Chemical agents and weapons**
  - **Nuclear materials and weapons**
- **Thousands of students and émigrés**
- **Infiltrated teams, BW, other WMD already**
- **Castro guerrillas infiltrated over the years**
- **Can do unacceptable damage to U.S. now**
- ***May reach first strike knockout capability***

© [illegible] T.E. BEARDEN

*Per OTA Report to Congress, 1993

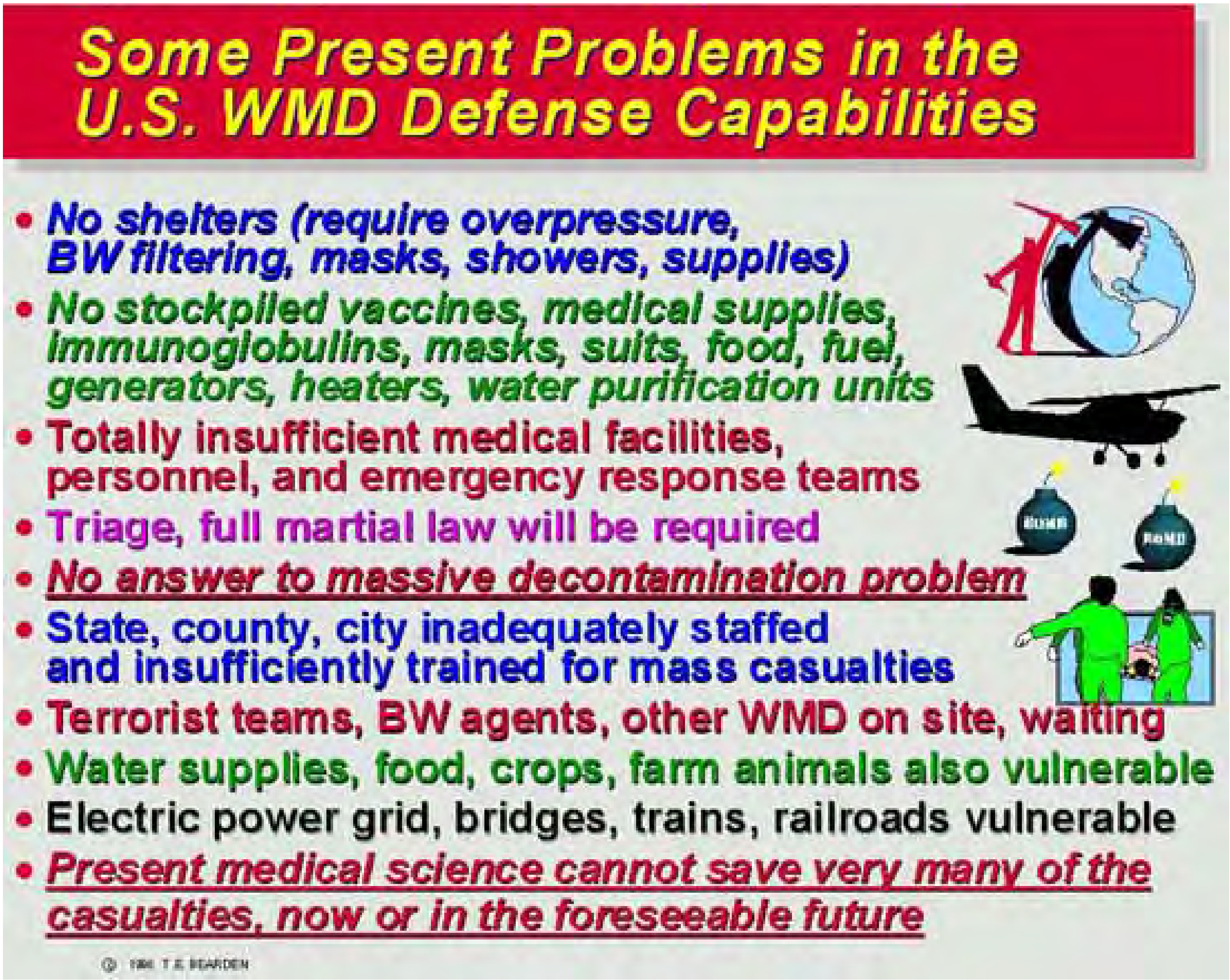

Some Present Problems in the U.S. WMD Defense Capabilities

- *No shelters (require overpressure, BW filtering, masks, showers, supplies)*
- *No stockpiled vaccines, medical supplies, immunoglobulins, masks, suits, food, fuel, generators, heaters, water purification units*
- Totally insufficient medical facilities, personnel, and emergency response teams
- Triage, full martial law will be required
- *No answer to massive decontamination problem*
- State, county, city inadequately staffed and insufficiently trained for mass casualties
- Terrorist teams, BW agents, other WMD on site, waiting
- Water supplies, food, crops, farm animals also vulnerable
- Electric power grid, bridges, trains, railroads vulnerable
- *Present medical science cannot save very many of the casualties, now or in the foreseeable future*

© [illegible] T.E. BEARDEN

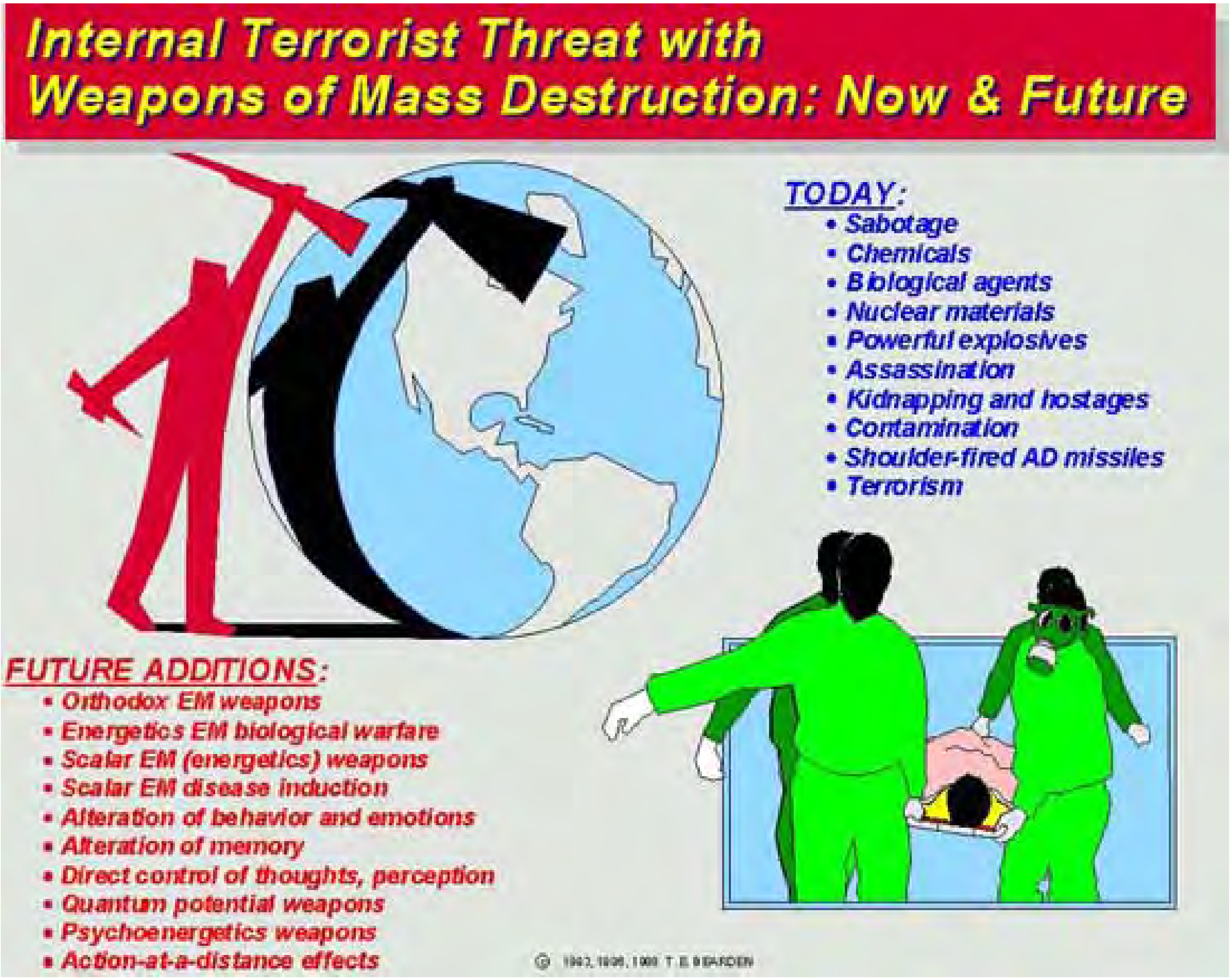

Internal Terrorist Threat with Weapons of Mass Destruction: Now & Future

TODAY:

- Sabotage
- Chemicals
- Biological agents
- Nuclear materials
- Powerful explosives
- Assassination
- Kidnapping and hostages
- Contamination
- Shoulder-fired AD missiles
- Terrorism

FUTURE ADDITIONS:

- Orthodox EM weapons
- Energetics EM biological warfare
- Scalar EM (energetics) weapons
- Scalar EM disease induction
- Alteration of behavior and emotions
- Alteration of memory
- Direct control of thoughts, perception
- Quantum potential weapons
- Psychoenergetics weapons
- Action-at-a-distance effects

© 1993, 1996, 1998 T. E. BEARDEN

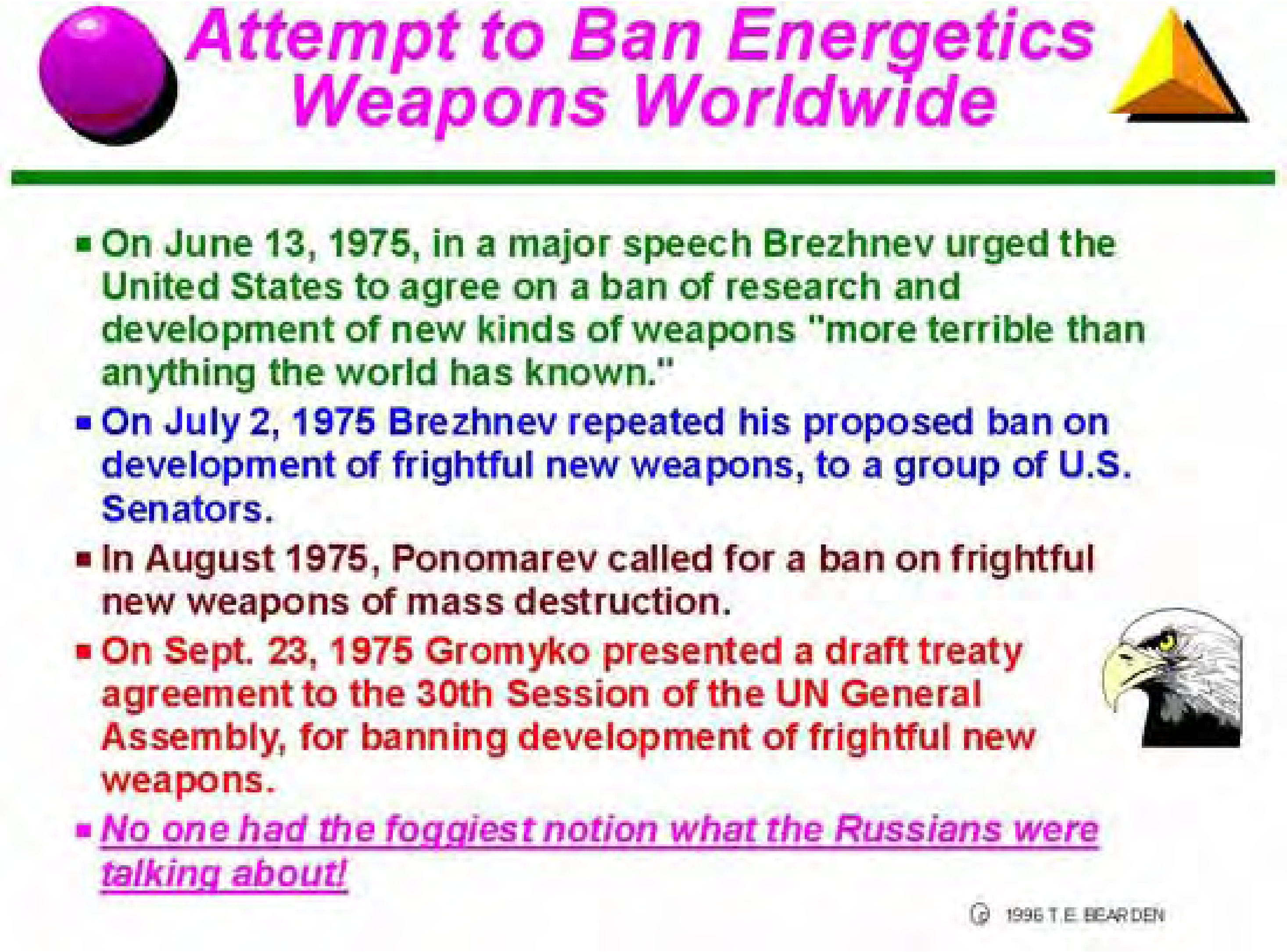

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

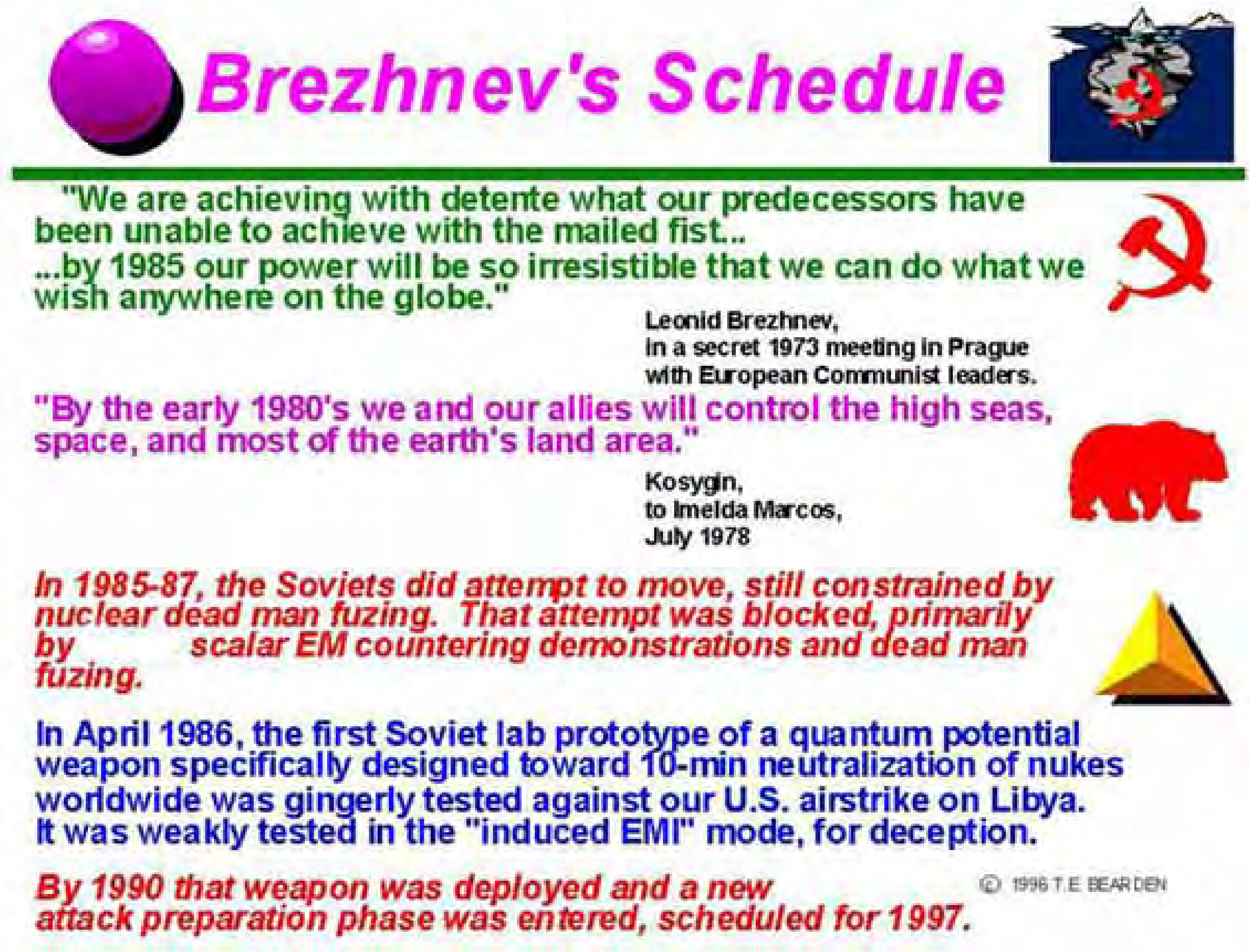

Brezhnev's Schedule

"We are achieving with detente what our predecessors have been unable to achieve with the mailed fist..
...by 1985 our power will be so irresistible that we can do what we wish anywhere on the globe."

Leonid Brezhnev,
in a secret 1973 meeting in Prague
with European Communist leaders.

"By the early 1980's we and our allies will control the high seas, space, and most of the earth's land area."

Kosygin,
to Imelda Marcos,
July 1978

In 1985-87, the Soviets did attempt to move, still constrained by nuclear dead man fuzing. That attempt was blocked, primarily by scalar EM countering demonstrations and dead man fuzing.

In April 1986, the first Soviet lab prototype of a quantum potential weapon specifically designed toward 10-min neutralization of nukes worldwide was gingerly tested against our U.S. airstrike on Libya. It was weakly tested in the "induced EMI" mode, for deception.

By 1990 that weapon was deployed and a new attack preparation phase was entered, scheduled for 1997.

© 1996 T.E. BEARDEN

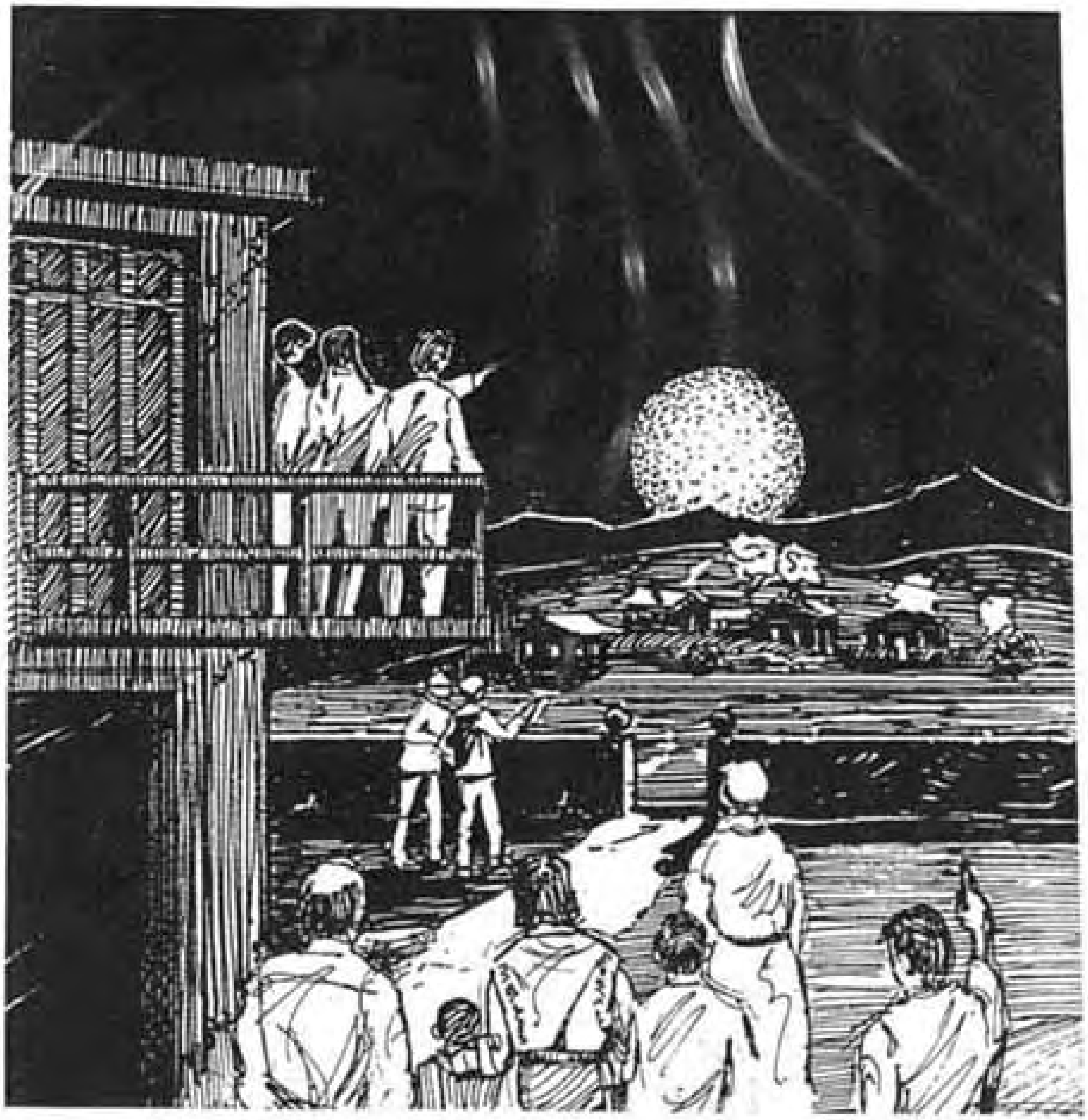

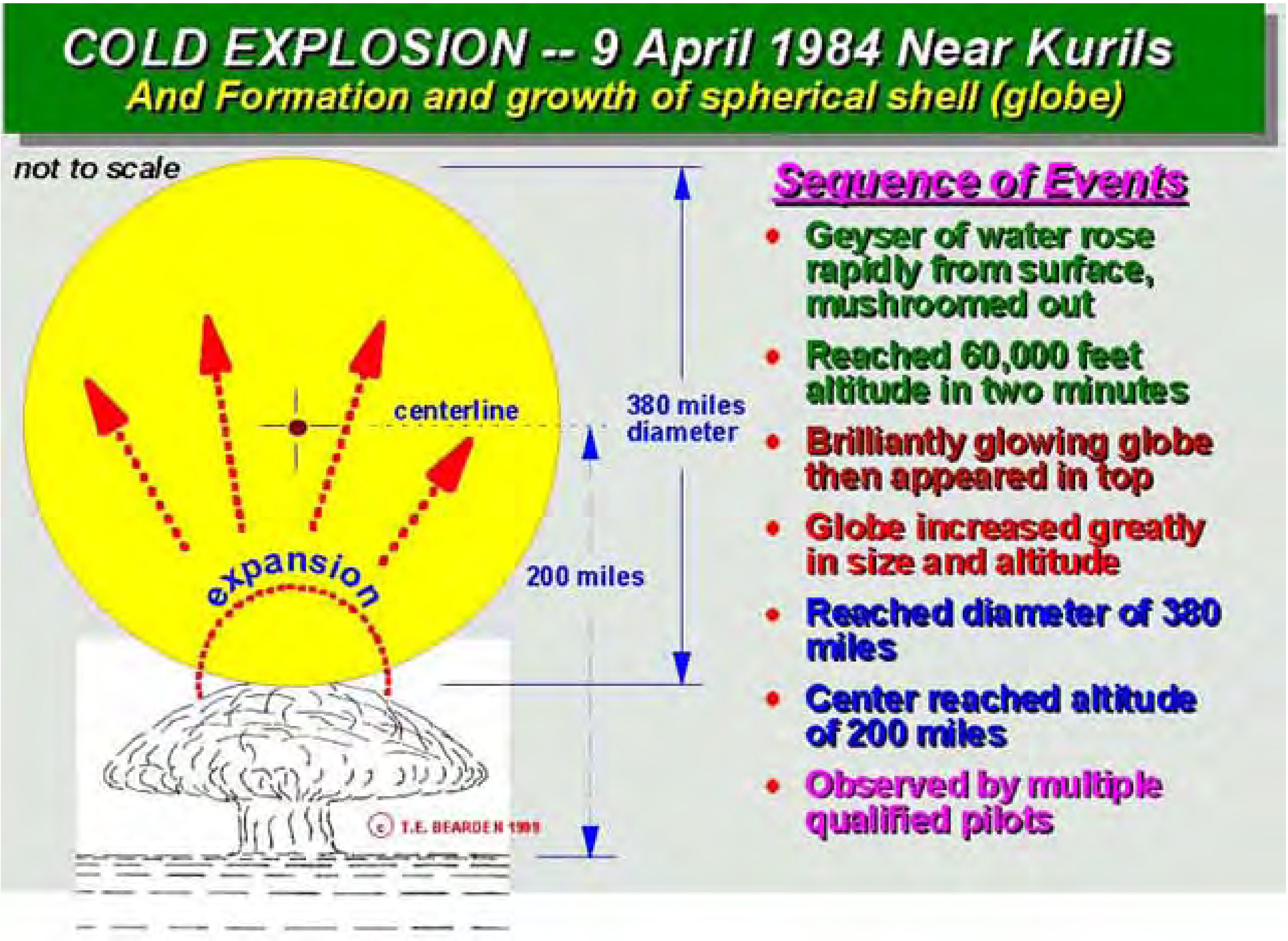

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles diameter
expansion
200 miles
© T.E. BEARDEN 1999

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in two minutes
- Brilliantly glowing globe then appeared in top
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles
- Observed by multiple qualified pilots

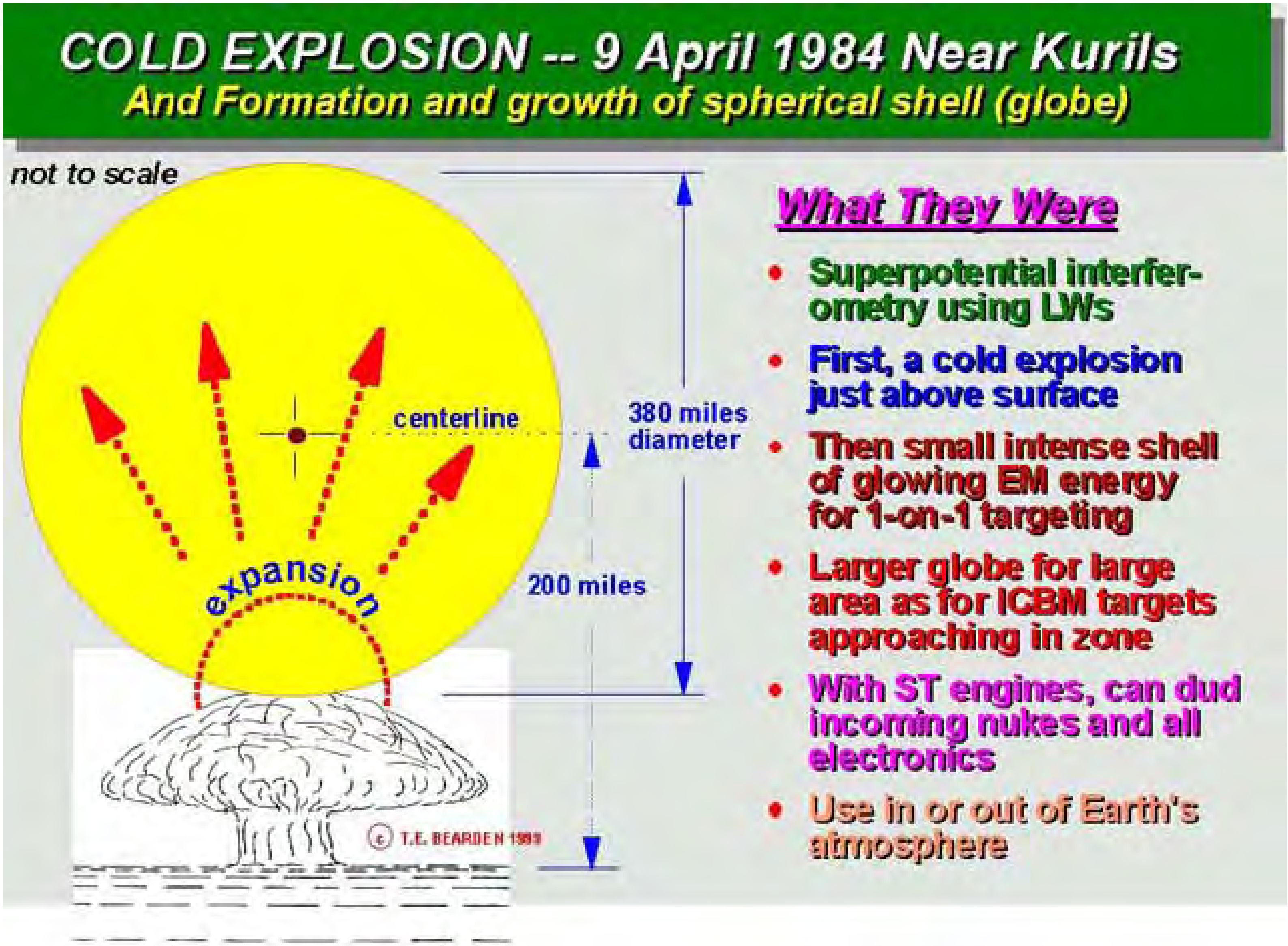

COLD EXPLOSION -- 9 April 1984 Near Kurils

And Formation and growth of spherical shell (globe)

not to scale
centerline
380 miles diameter
200 miles
expansion
© T.E. BEARDEN 1999

What They Were

- **Superpotential interferometry using LWs**
- **First, a cold explosion just above surface**
- **Then small intense shell of glowing EM energy for 1-on-1 targeting**
- **Larger globe for large area as for ICBM targets approaching in zone**
- **With ST engines, can dud incoming nukes and all electronics**
- **Use in or out of Earth's atmosphere**

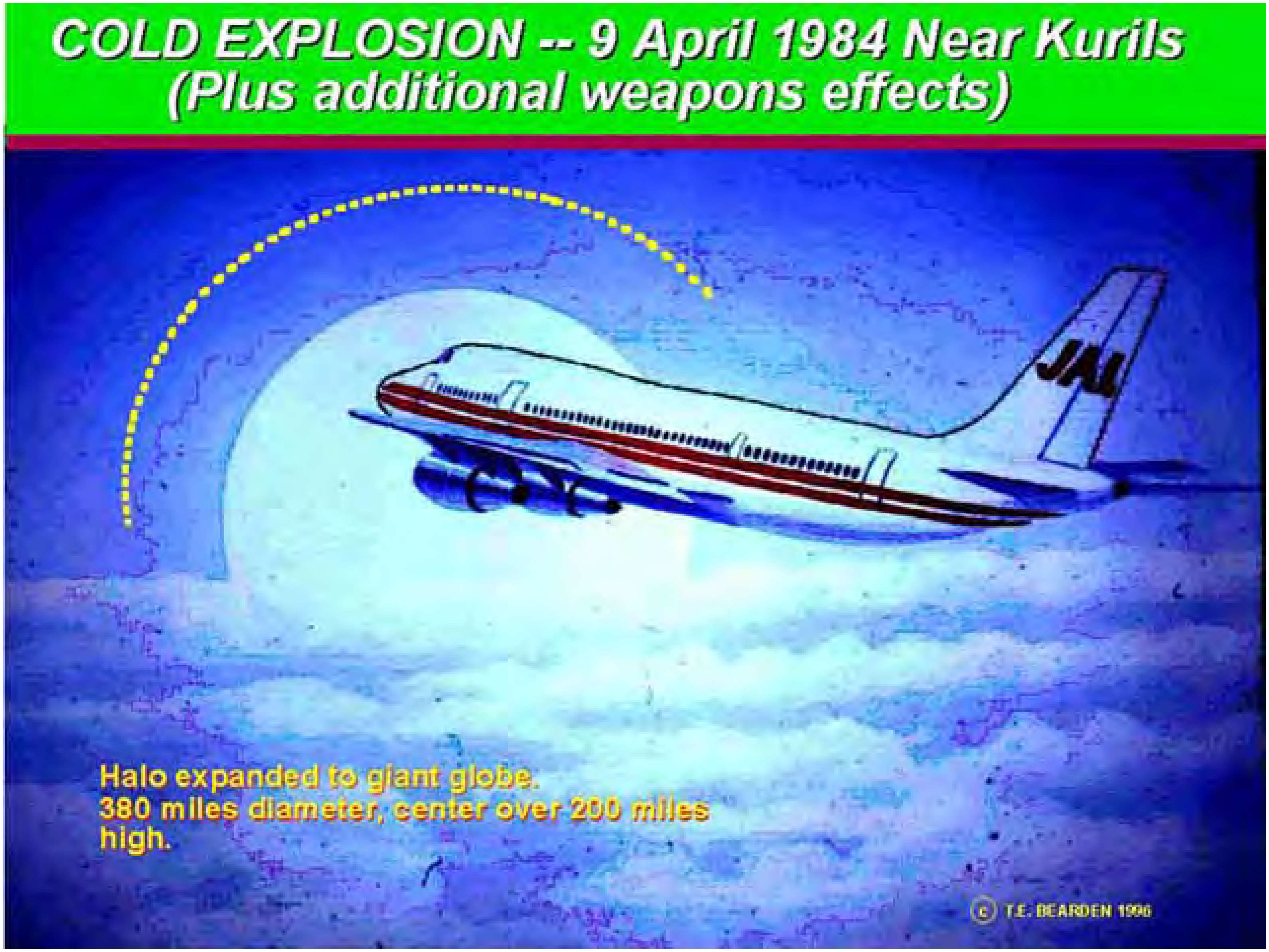

COLD EXPLOSION -- 9 April 1984 Near Kurils
(Plus additional weapons effects)

Halo expanded to giant globe.
380 miles diameter, center over 200 miles high.

© T.E. BEARDEN 1996

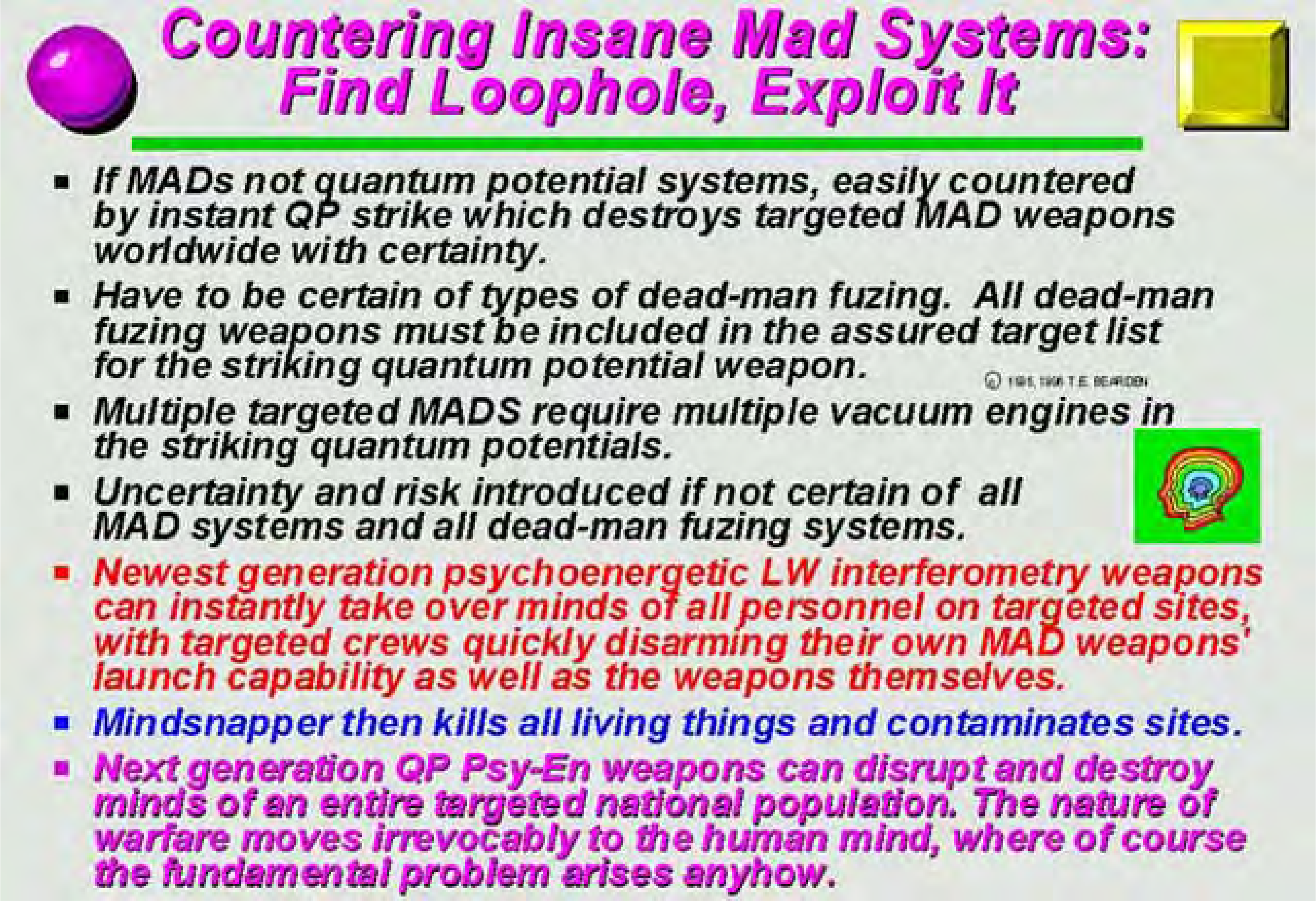

Countering Insane Mad Systems: Find Loophole, Exploit It

- *If MADs not quantum potential systems, easily countered by instant QP strike which destroys targeted MAD weapons worldwide with certainty.*
- *Have to be certain of types of dead-man fuzing. All dead-man fuzing weapons must be included in the assured target list for the striking quantum potential weapon.*

© 1995, 1996 T.E. BEARDEN

- *Multiple targeted MADS require multiple vacuum engines in the striking quantum potentials.*
- *Uncertainty and risk introduced if not certain of all MAD systems and all dead-man fuzing systems.*
- *Newest generation psychoenergetic LW interferometry weapons can instantly take over minds of all personnel on targeted sites, with targeted crews quickly disarming their own MAD weapons' launch capability as well as the weapons themselves.*
- *Mindsnapper then kills all living things and contaminates sites.*
- *Next generation QP Psy-En weapons can disrupt and destroy minds of an entire targeted national population. The nature of warfare moves irrevocably to the human mind, where of course the fundamental problem arises anyhow.*

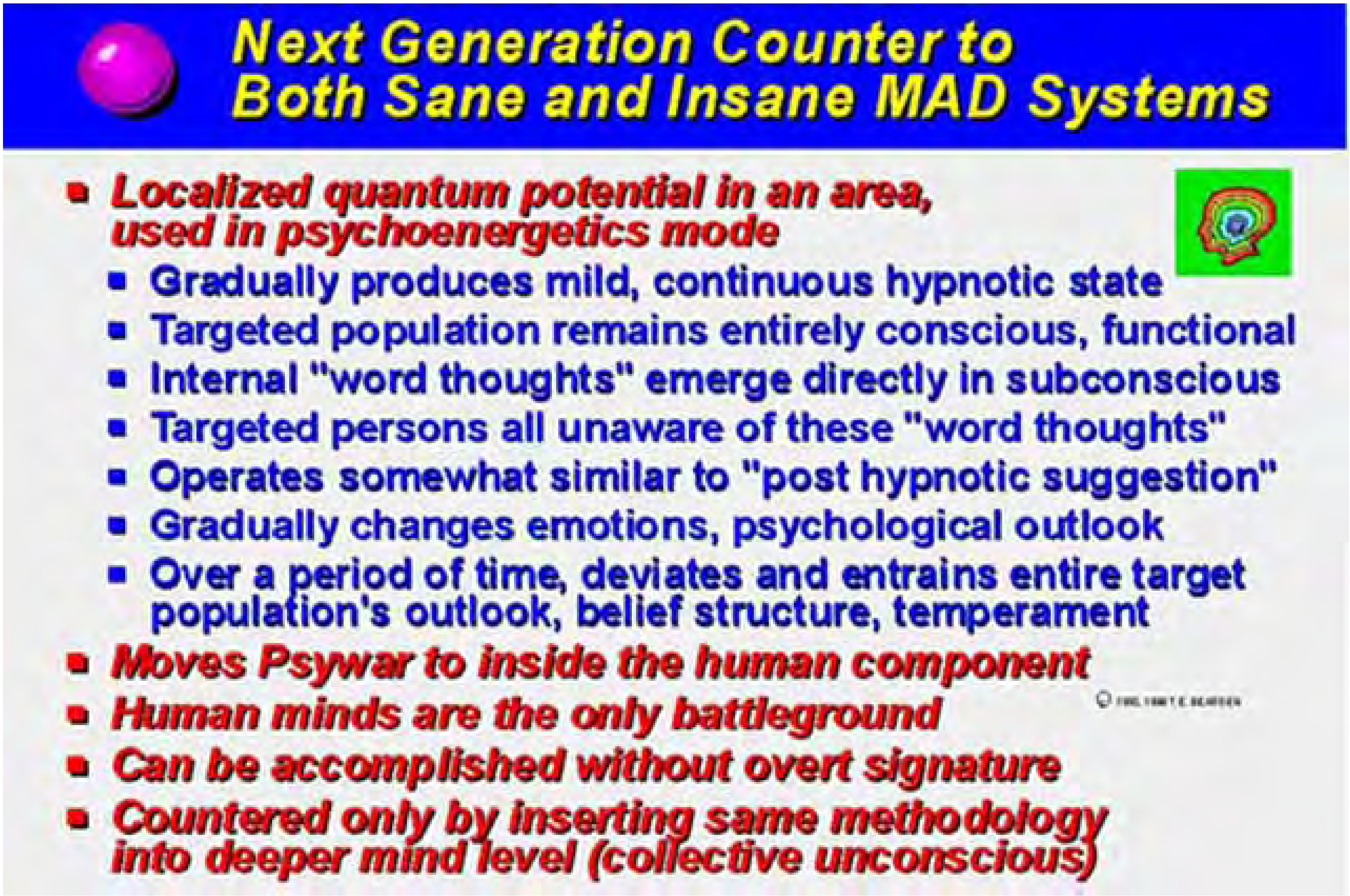

Next Generation Counter to Both Sane and Insane MAD Systems

- ***Localized quantum potential in an area, used in psychoenergetics mode***
  - **Gradually produces mild, continuous hypnotic state**
  - **Targeted population remains entirely conscious, functional**
  - **Internal "word thoughts" emerge directly in subconscious**
  - **Targeted persons all unaware of these "word thoughts"**
  - **Operates somewhat similar to "post hypnotic suggestion"**
  - **Gradually changes emotions, psychological outlook**
  - **Over a period of time, deviates and entrains entire target population's outlook, belief structure, temperament**
- ***Moves Psywar to inside the human component***
- ***Human minds are the only battleground***
- ***Can be accomplished without overt signature***
- ***Countered only by inserting same methodology into deeper mind level (collective unconscious)***

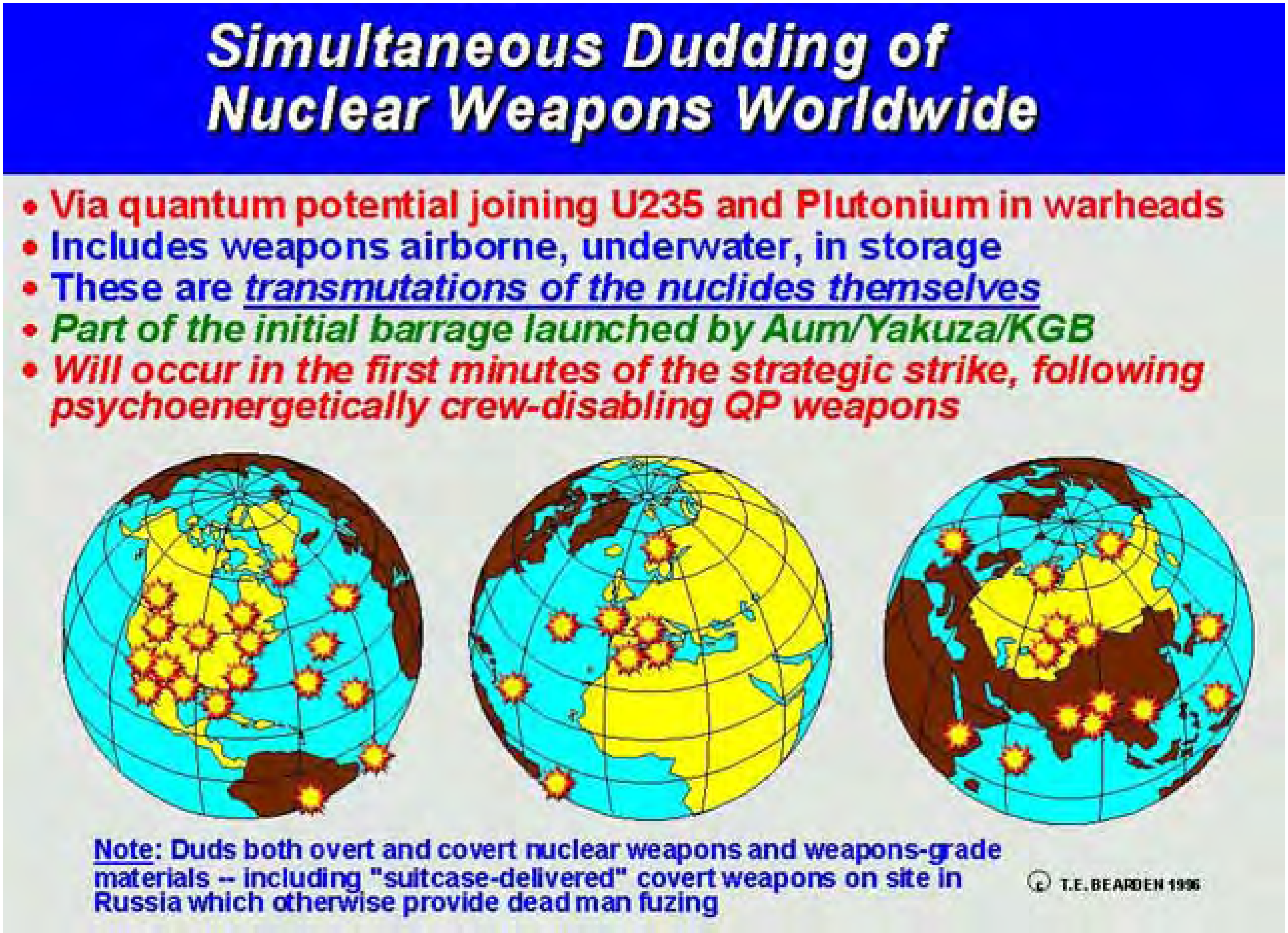

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- These are *transmutations of the nuclides themselves*
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike, following psychoenergetically crew-disabling QP weapons*

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

© T.E. BEARDEN 1996

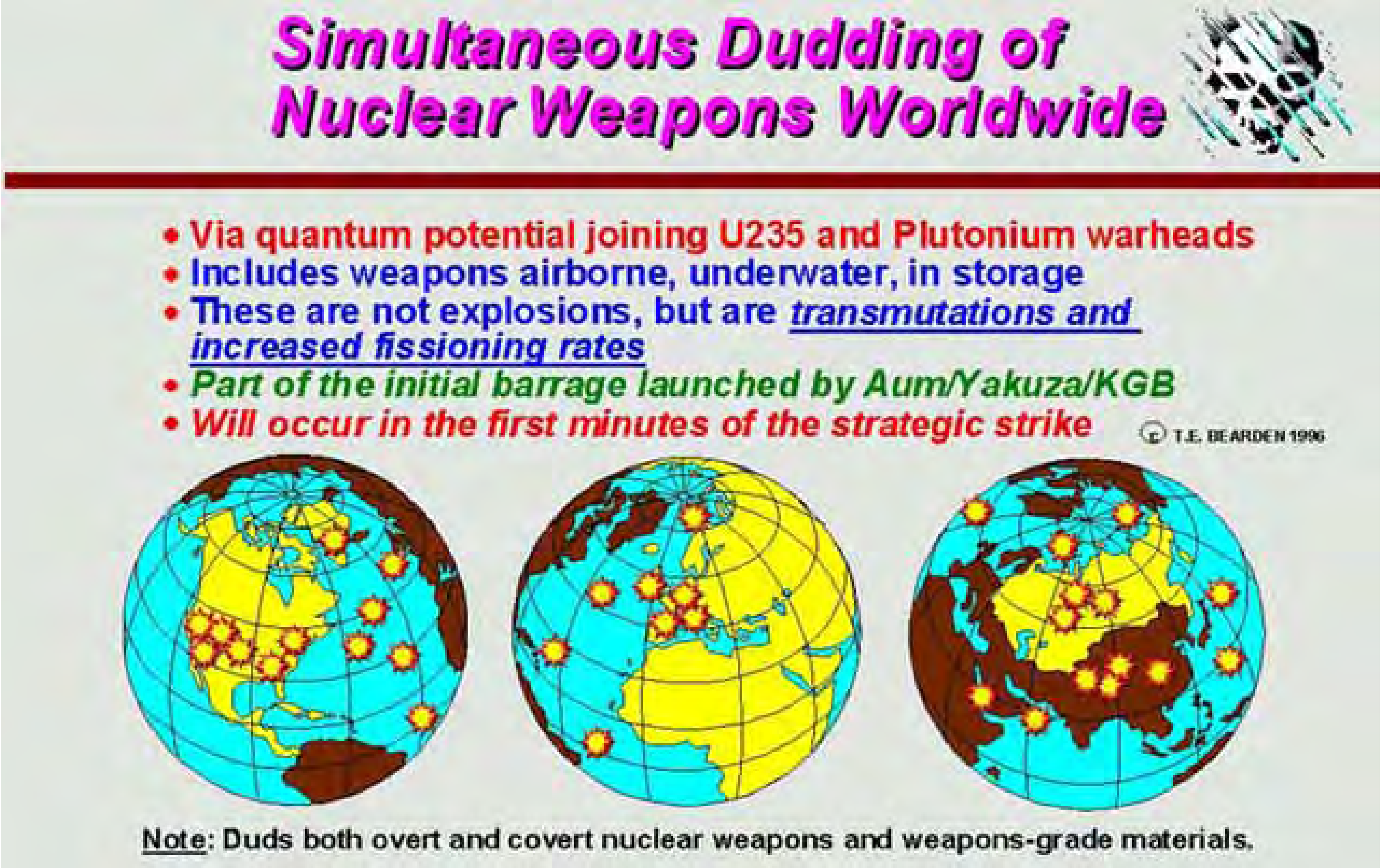

Simultaneous Dudding of Nuclear Weapons Worldwide

- Via quantum potential joining U235 and Plutonium warheads
- Includes weapons airborne, underwater, in storage
- These are not explosions, but are ***transmutations and increased fissioning rates***
- *Part of the initial barrage launched by Aum/Yakuza/KGB*
- *Will occur in the first minutes of the strategic strike*

© T.E. BEARDEN 1996

Note: Duds both overt and covert nuclear weapons and weapons-grade materials.

The Tom Bearden Website

Help support the research

Operation Divine Wind II

Slide 1

Slide 2

Slide 3

Slide 4

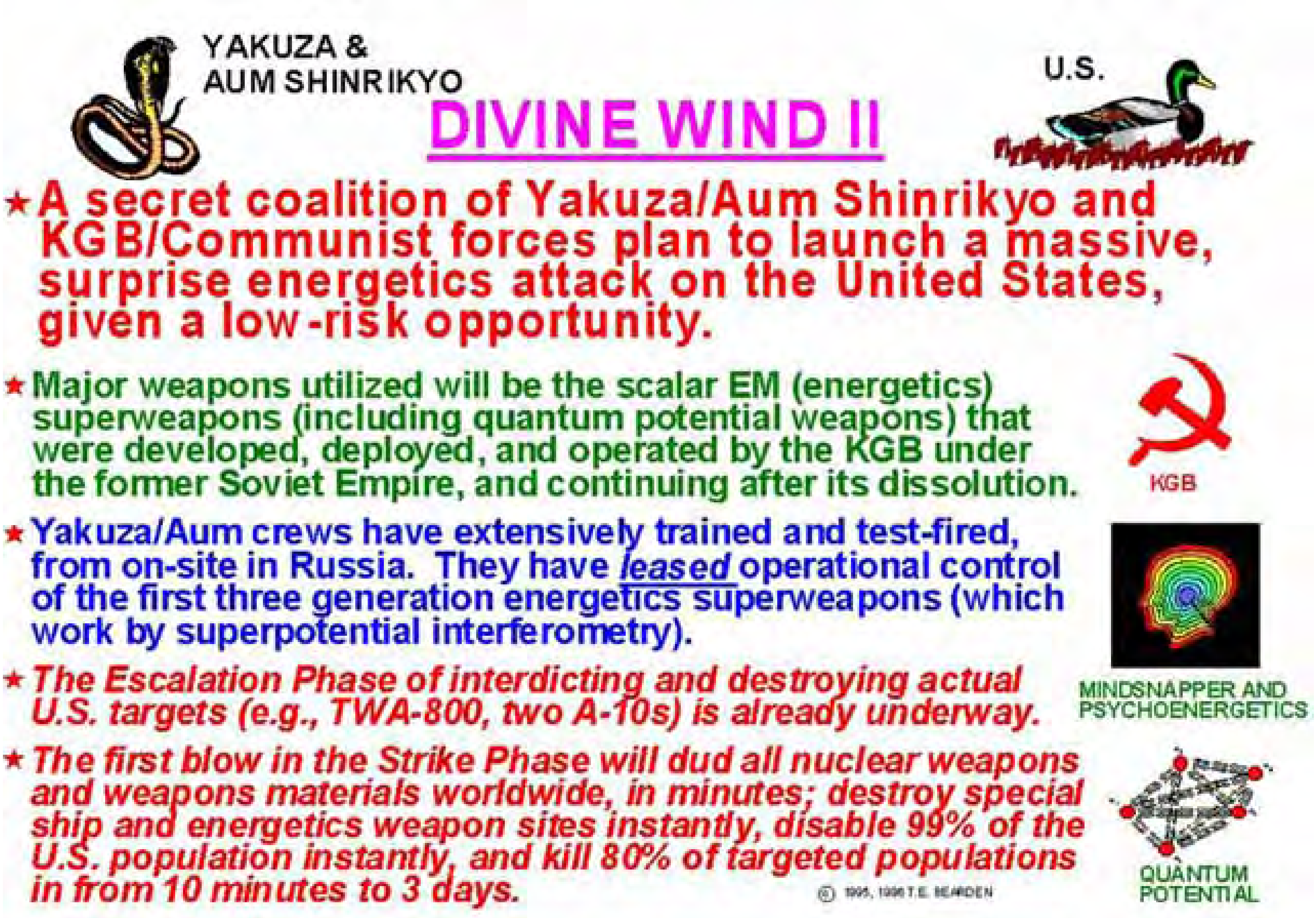

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plan to launch a massive, surprise energetics attack on the United States, given a low-risk opportunity.
- ★ Major weapons utilized will be the scalar EM (energetics) superweapons (including quantum potential weapons) that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
- ★ Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the first three generation energetics superweapons (which work by superpotential interferometry).
- ★ *The Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800, two A-10s) is already underway.*
- ★ *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, disable 99% of the U.S. population instantly, and kill 80% of targeted populations in from 10 minutes to 3 days.*

KGB

MINDSNAPPER AND
PSYCHOENERGETICS

QUANTUM
POTENTIAL

© 1995, 1996 T.E. BEARDEN

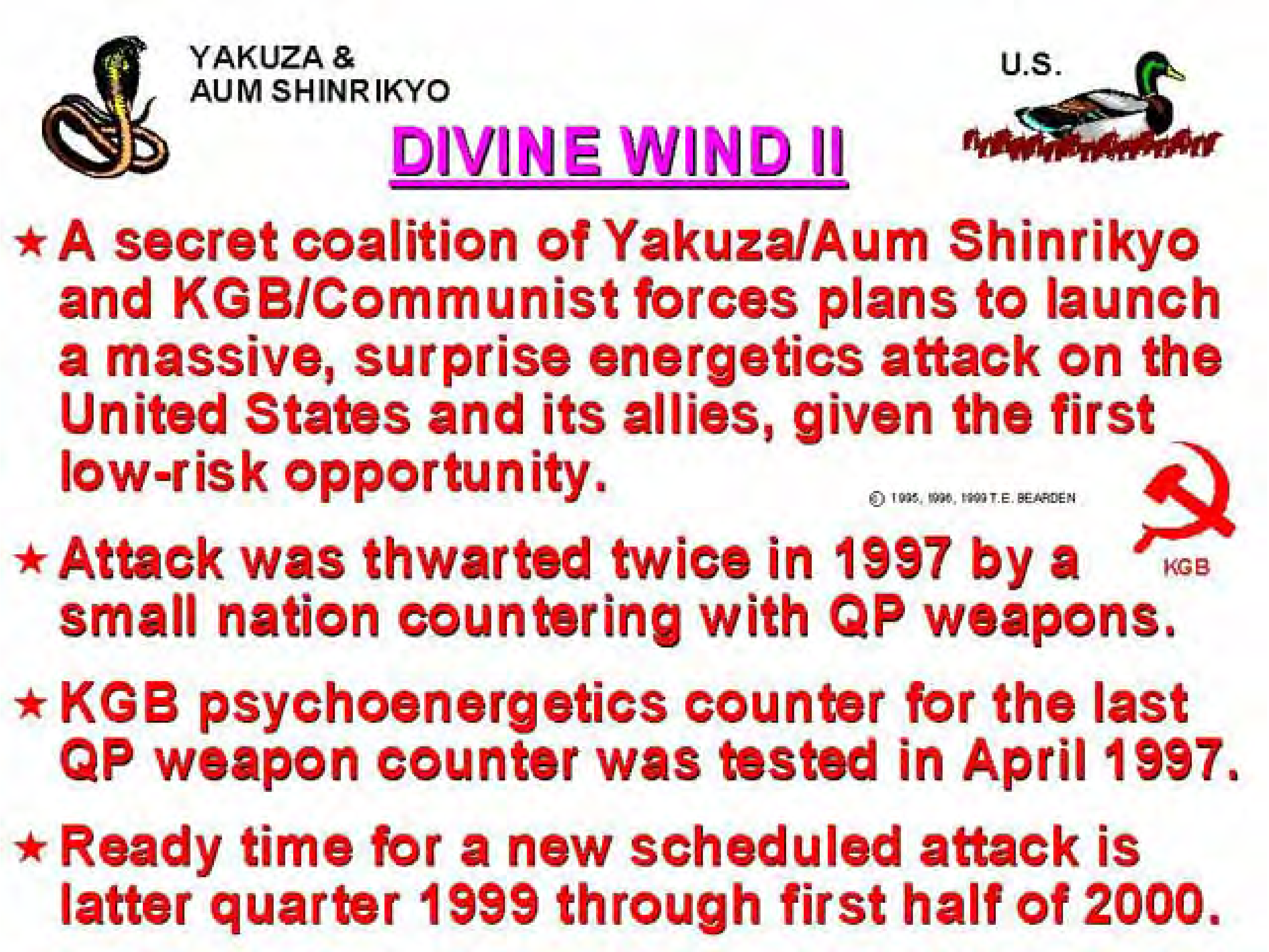

YAKUZA &
AUM SHINRIKYO

U.S.

DIVINE WIND II

- ★ A secret coalition of Yakuza/Aum Shinrikyo and KGB/Communist forces plans to launch a massive, surprise energetics attack on the United States and its allies, given the first low-risk opportunity.

© 1995, 1996, 1999 T.E. BEARDEN

KGB

- ★ Attack was thwarted twice in 1997 by a small nation countering with QP weapons.
- ★ KGB psychoenergetics counter for the last QP weapon counter was tested in April 1997.
- ★ Ready time for a new scheduled attack is latter quarter 1999 through first half of 2000.

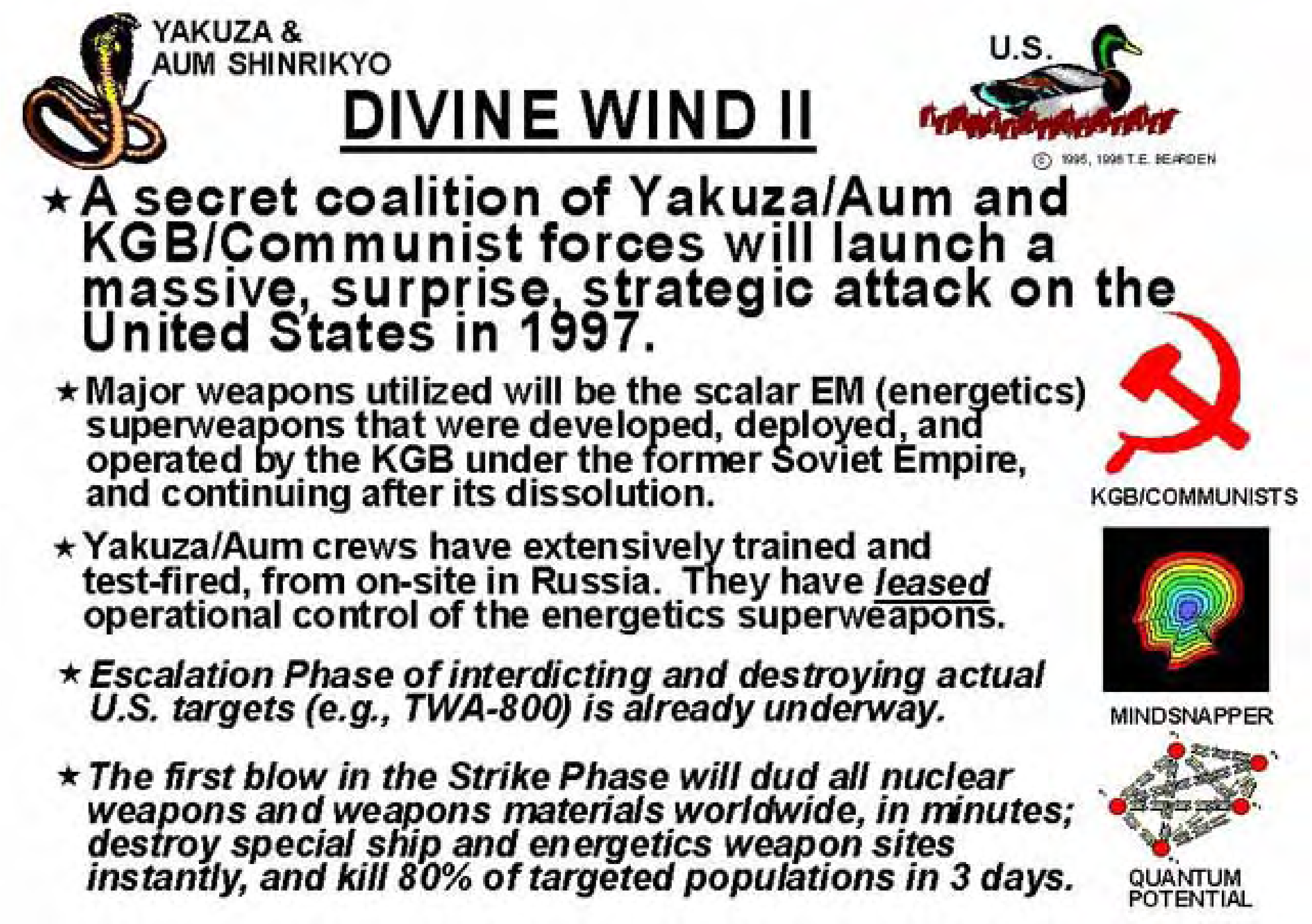
YAKUZA &
AUM SHINRIKYO
U.S.
KGB/COMMUNISTS
MINDSNAPPER
QUANTUM
POTENTIAL

DIVINE WIND II

© 1995, 1996 T.E. BEARDEN

- **A secret coalition of Yakuza/Aum and KGB/Communist forces will launch a massive, surprise, strategic attack on the United States in 1997.**
  - Major weapons utilized will be the scalar EM (energetics) superweapons that were developed, deployed, and operated by the KGB under the former Soviet Empire, and continuing after its dissolution.
  - Yakuza/Aum crews have extensively trained and test-fired, from on-site in Russia. They have *leased* operational control of the energetics superweapons.
  - *Escalation Phase of interdicting and destroying actual U.S. targets (e.g., TWA-800) is already underway.*
  - *The first blow in the Strike Phase will dud all nuclear weapons and weapons materials worldwide, in minutes; destroy special ship and energetics weapon sites instantly, and kill 80% of targeted populations in 3 days.*

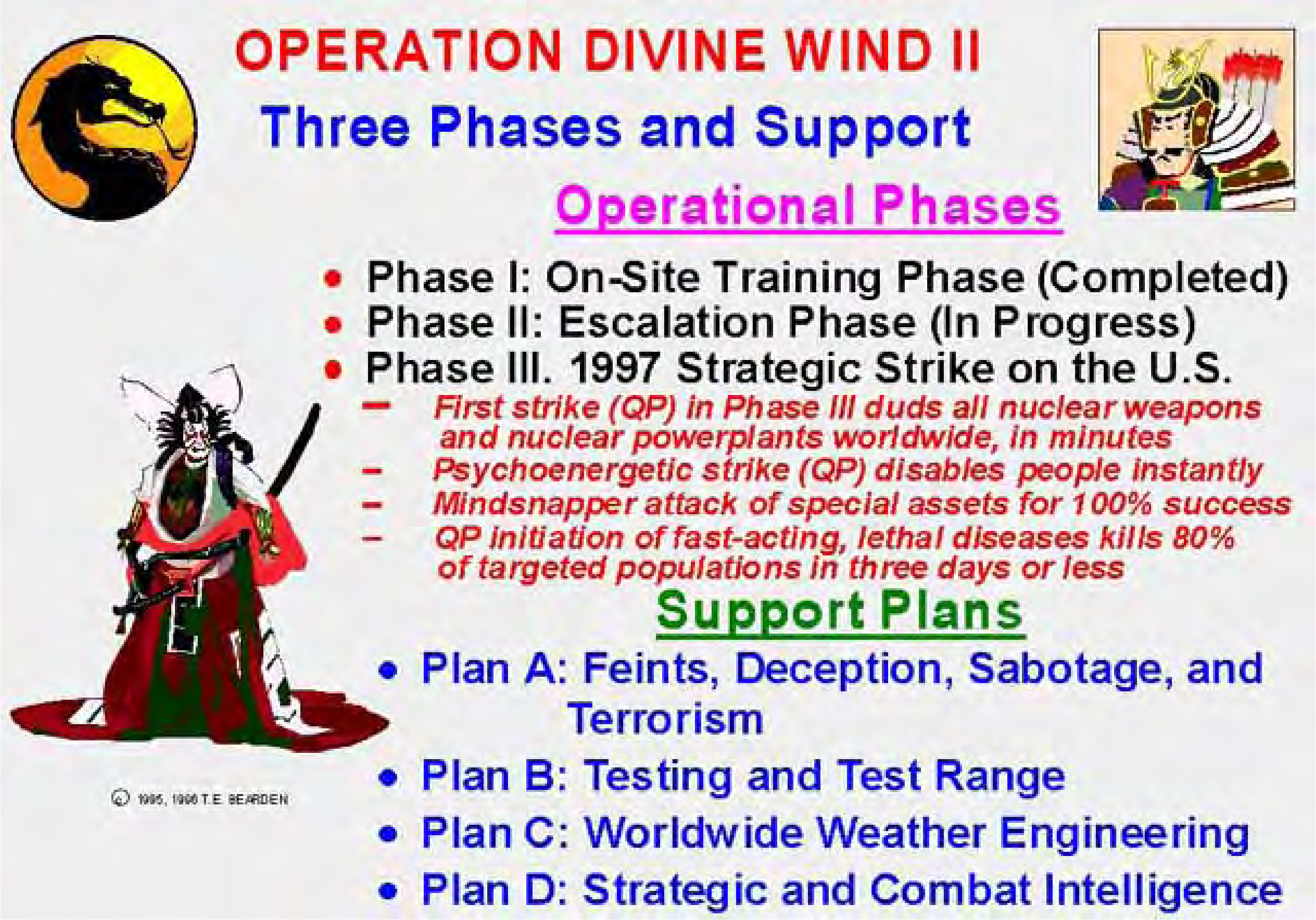

OPERATION DIVINE WIND II

Three Phases and Support

Operational Phases

- Phase I: On-Site Training Phase (Completed)
- Phase II: Escalation Phase (In Progress)
- Phase III. 1997 Strategic Strike on the U.S.
  - *First strike (QP) in Phase III duds all nuclear weapons and nuclear powerplants worldwide, in minutes*
  - *Psychoenergetic strike (QP) disables people instantly*
  - *Mindsnapper attack of special assets for 100% success*
  - *QP initiation of fast-acting, lethal diseases kills 80% of targeted populations in three days or less*

Support Plans

- Plan A: Feints, Deception, Sabotage, and Terrorism
- Plan B: Testing and Test Range
- Plan C: Worldwide Weather Engineering
- Plan D: Strategic and Combat Intelligence

© 1995, 1996 T.E. BEARDEN

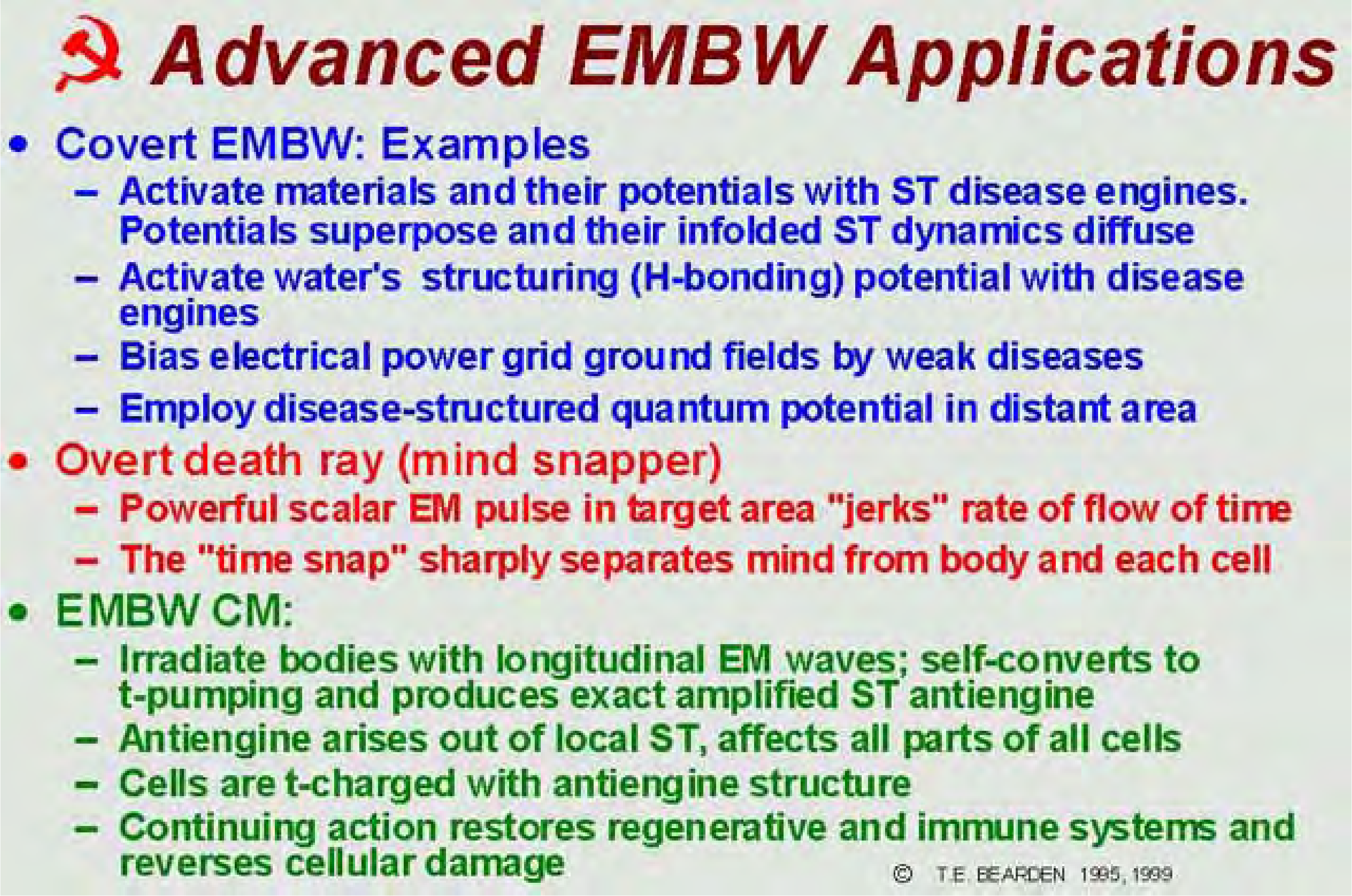

Advanced EMBW Applications

- Covert EMBW: Examples
  - Activate materials and their potentials with ST disease engines. Potentials superpose and their infolded ST dynamics diffuse
  - Activate water's structuring (H-bonding) potential with disease engines
  - Bias electrical power grid ground fields by weak diseases
  - Employ disease-structured quantum potential in distant area
- Overt death ray (mind snapper)
  - Powerful scalar EM pulse in target area "jerks" rate of flow of time
  - The "time snap" sharply separates mind from body and each cell
- EMBW CM:
  - Irradiate bodies with longitudinal EM waves; self-converts to t-pumping and produces exact amplified ST antiengine
  - Antiengine arises out of local ST, affects all parts of all cells
  - Cells are t-charged with antiengine structure
  - Continuing action restores regenerative and immune systems and reverses cellular damage

© T.E. BEARDEN 1995, 1999

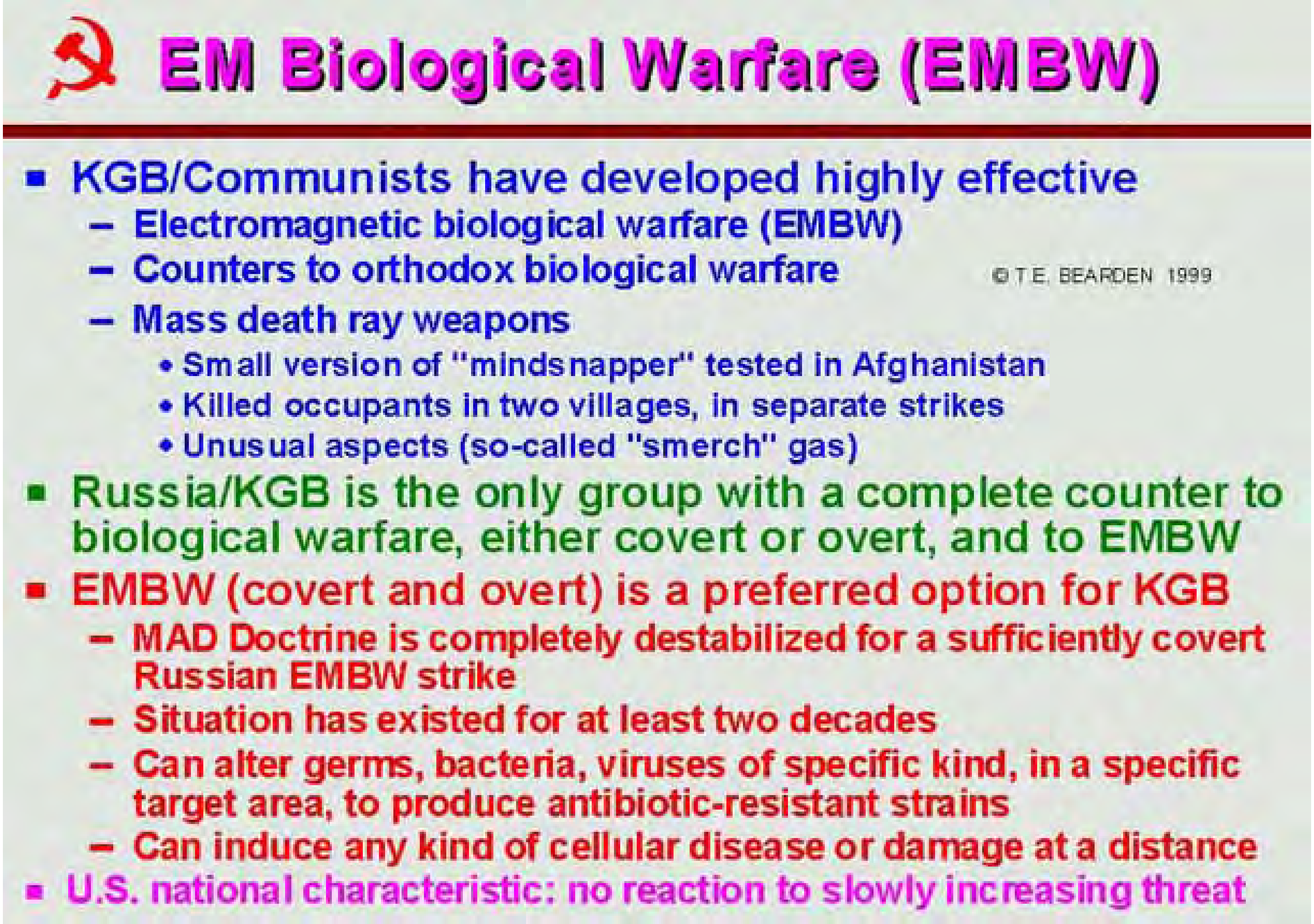

EM Biological Warfare (EMBW)

- KGB/Communists have developed highly effective
  - Electromagnetic biological warfare (EMBW)
  - Counters to orthodox biological warfare
  - Mass death ray weapons
    - Small version of "mindsnapper" tested in Afghanistan
    - Killed occupants in two villages, in separate strikes
    - Unusual aspects (so-called "smerch" gas)
- Russia/KGB is the only group with a complete counter to biological warfare, either covert or overt, and to EMBW
- EMBW (covert and overt) is a preferred option for KGB
  - MAD Doctrine is completely destabilized for a sufficiently covert Russian EMBW strike
  - Situation has existed for at least two decades
  - Can alter germs, bacteria, viruses of specific kind, in a specific target area, to produce antibiotic-resistant strains
  - Can induce any kind of cellular disease or damage at a distance
- U.S. national characteristic: no reaction to slowly increasing threat

© T.E. BEARDEN 1999

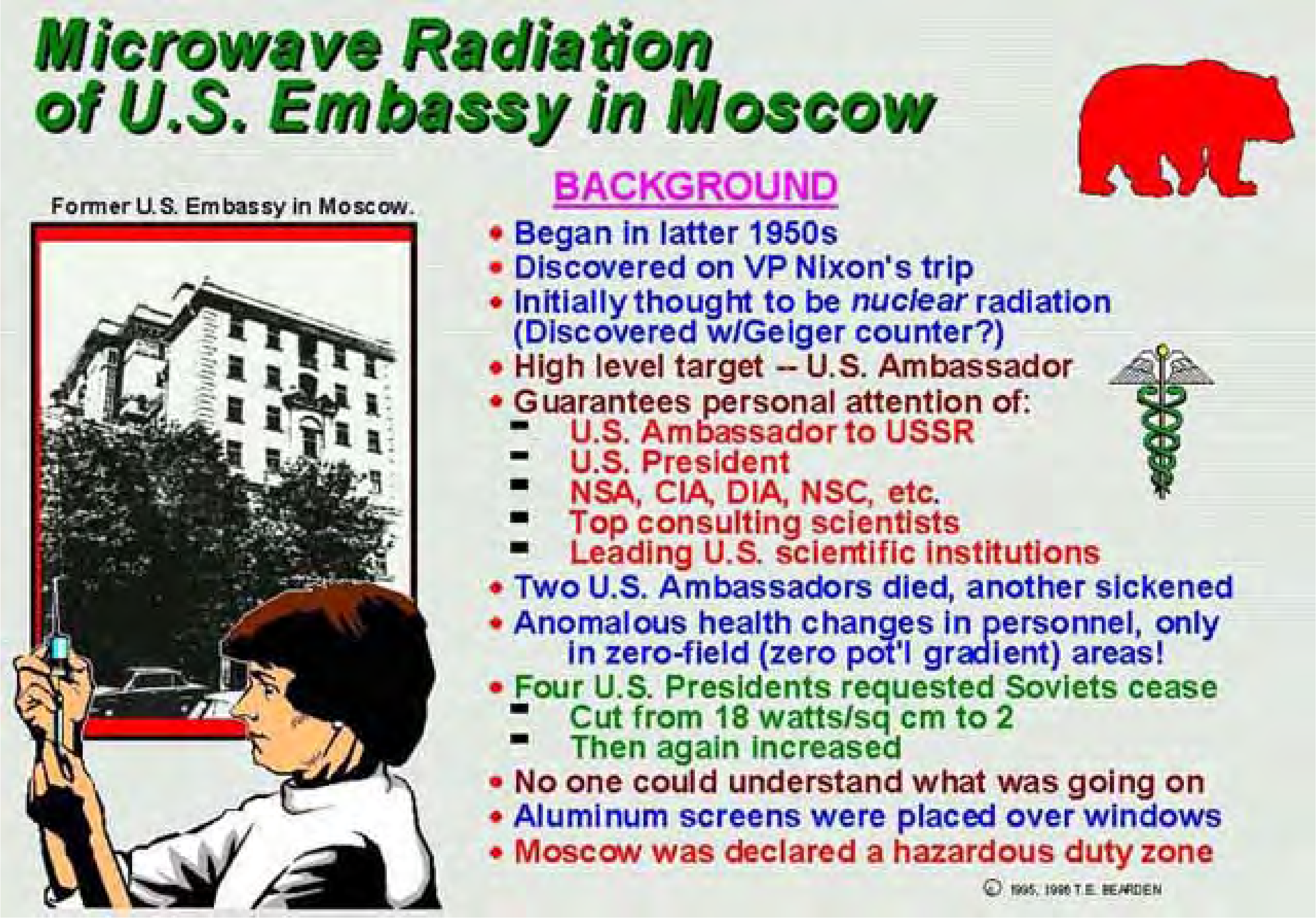

Microwave Radiation of U.S. Embassy in Moscow

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

© 1995, 1996 T.E. BEARDEN

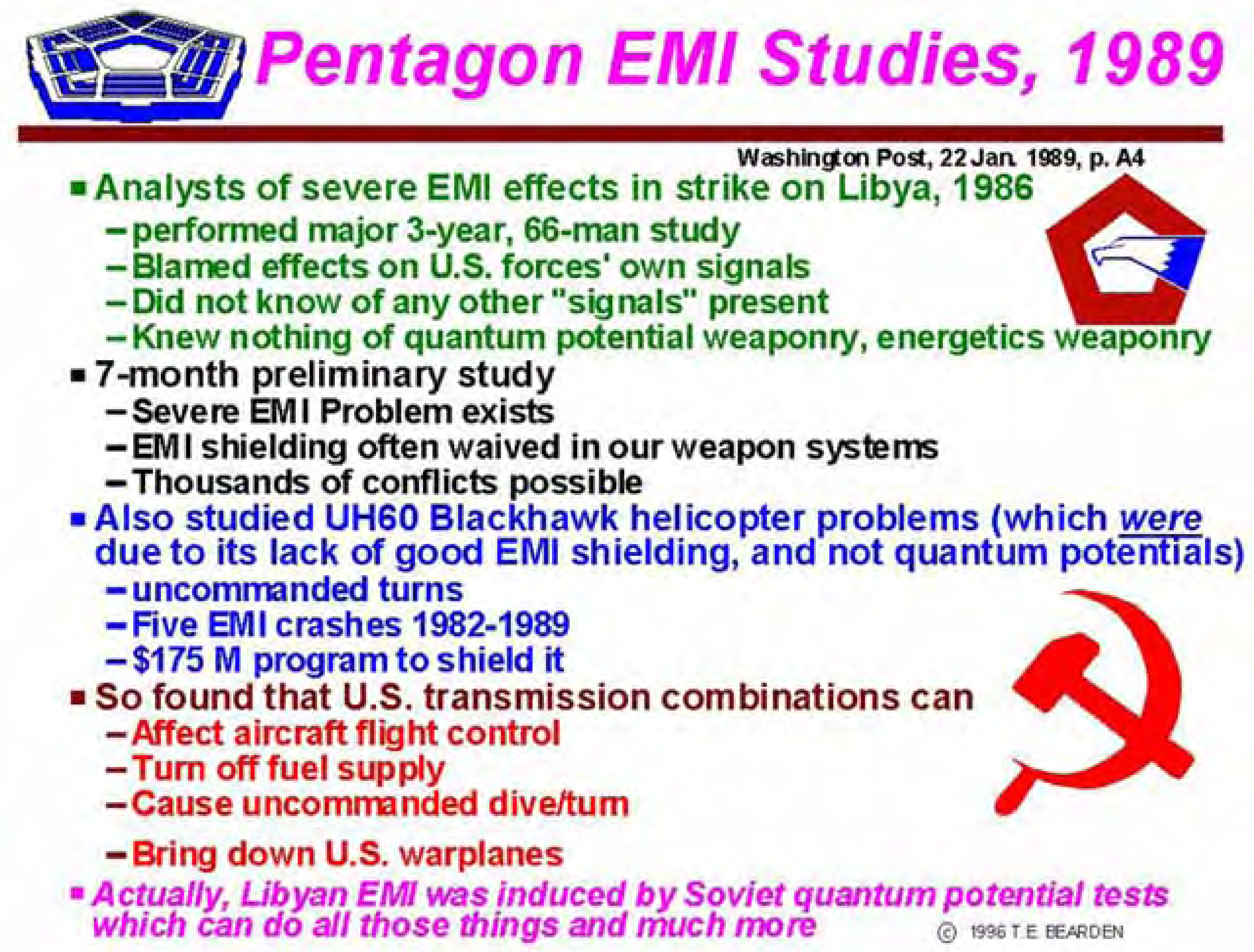

Pentagon EMI Studies, 1989

Washington Post, 22 Jan. 1989, p. A4

- Analysts of severe EMI effects in strike on Libya, 1986
  - performed major 3-year, 66-man study
  - Blamed effects on U.S. forces' own signals
  - Did not know of any other "signals" present
  - Knew nothing of quantum potential weaponry, energetics weaponry
- 7-month preliminary study
  - Severe EMI Problem exists
  - EMI shielding often waived in our weapon systems
  - Thousands of conflicts possible
- Also studied UH60 Blackhawk helicopter problems (which *were* due to its lack of good EMI shielding, and not quantum potentials)
  - uncommanded turns
  - Five EMI crashes 1982-1989
  - $175 M program to shield it
- So found that U.S. transmission combinations can
  - Affect aircraft flight control
  - Turn off fuel supply
  - Cause uncommanded dive/turn
  - Bring down U.S. warplanes
- *Actually, Libyan EMI was induced by Soviet quantum potential tests which can do all those things and much more*

© 1996 T.E. BEARDEN

The Tom Bearden Website

Help support the research

EM Missile Practice

- EM Missile Practice
  Latter Sept. - Early Nov. 1996
  Melbourne, Australia

- Deception: Using EM Missile to Reinforce Notion of SAM Attack
  Saturday evening, Nov. 16, 1996

- Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry
  Thursday, December 12th, 1996

- EM Missile Strike, Offset from Night Shuttle Launch, Cape Canaveral, Florida
  The Smoking Gun in November 1985

- Distant Operator Onsite in Russia Slew Away Registration Point
  Marker Beacon over Night Shuttle Launch, 26 November, 1985

- EM Missile Practice and Deception
  November 17, 1996

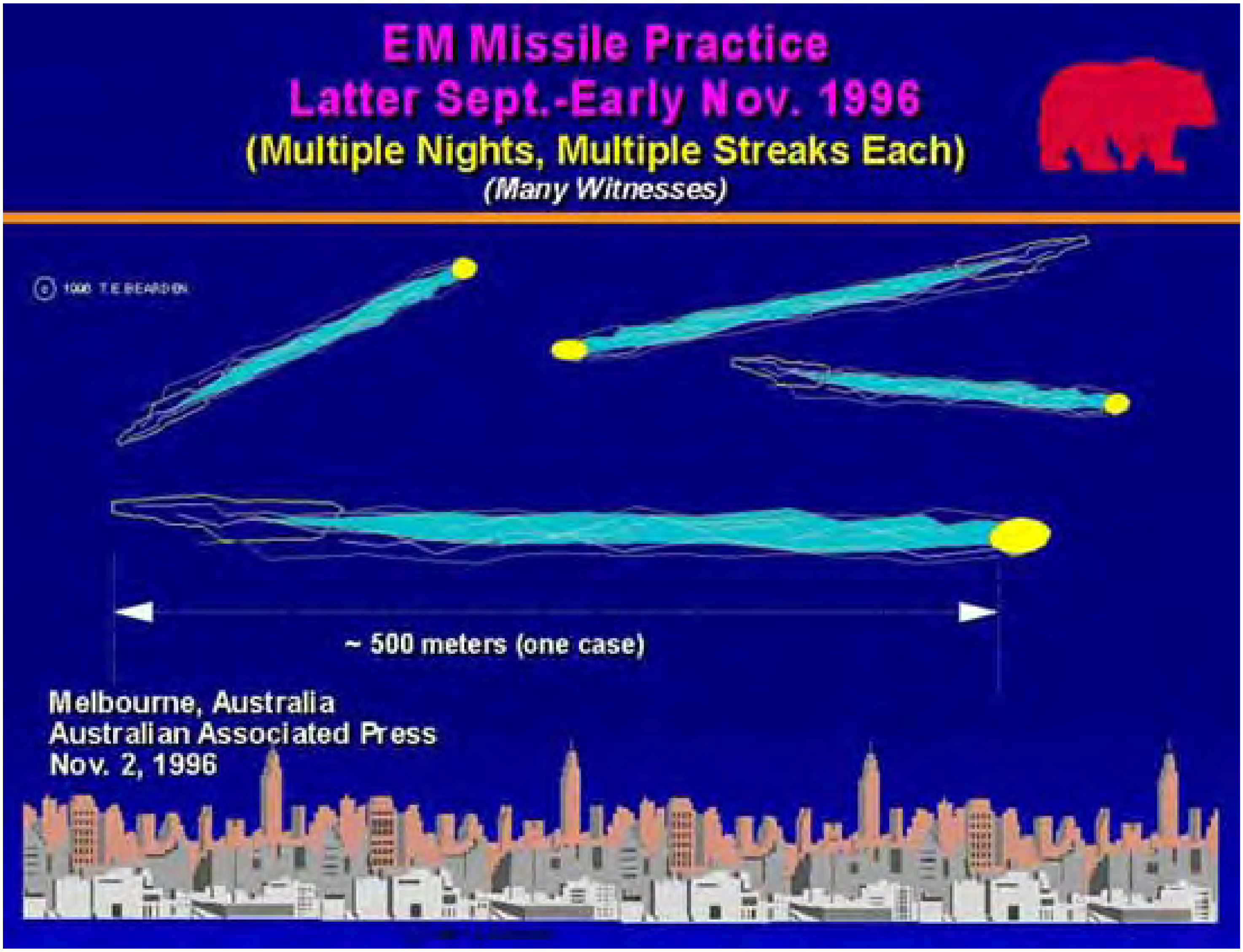

EM Missile Practice
Latter Sept.-Early Nov. 1996
(Multiple Nights, Multiple Streaks Each)
(Many Witnesses)

© 1996 T.E. BEARDEN

Melbourne, Australia
Australian Associated Press
Nov. 2, 1996

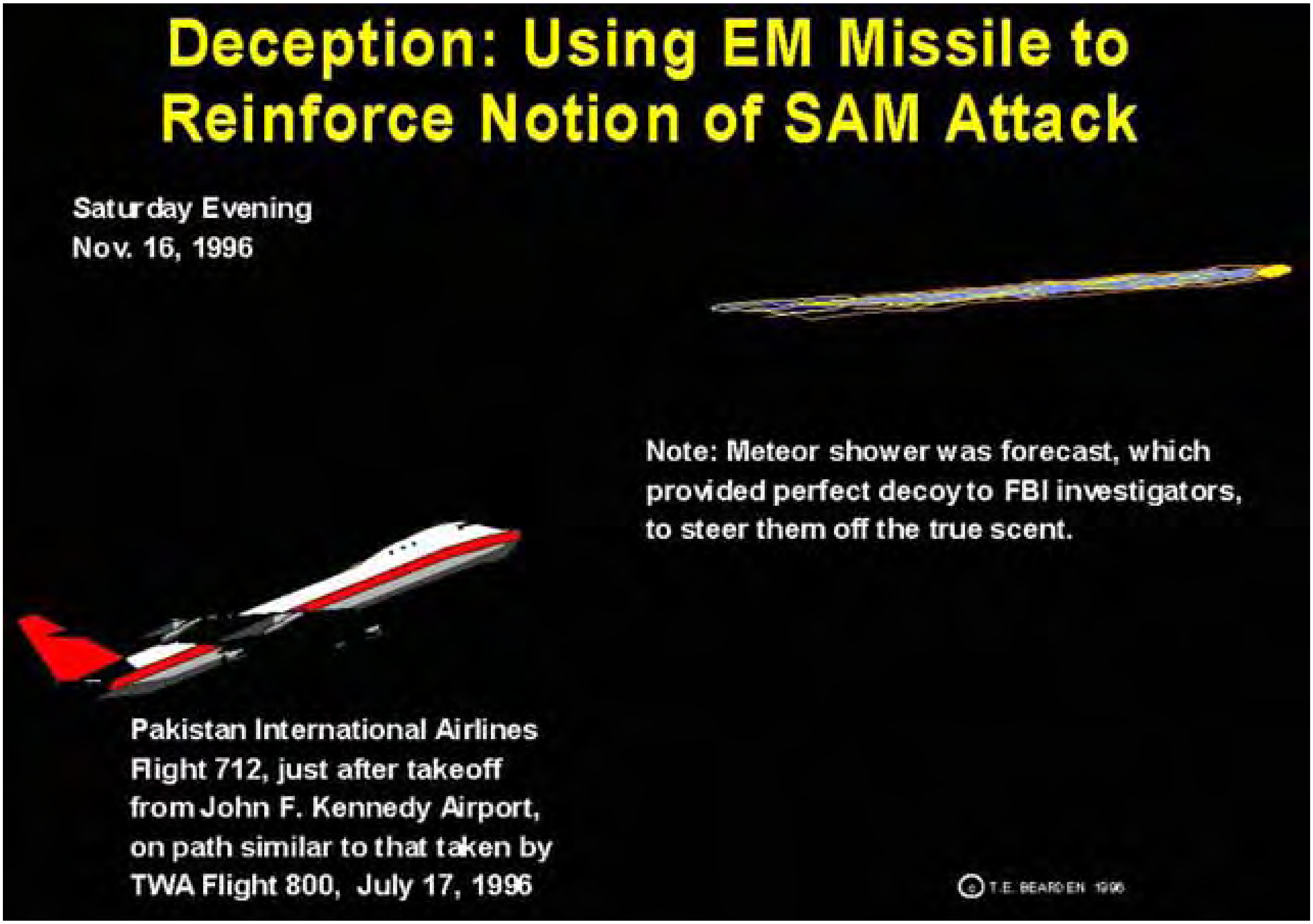

Deception: Using EM Missile to Reinforce Notion of SAM Attack

Saturday Evening
Nov. 16, 1996

Note: Meteor shower was forecast, which provided perfect decoy to FBI investigators, to steer them off the true scent.

Pakistan International Airlines Flight 712, just after takeoff from John F. Kennedy Airport, on path similar to that taken by TWA Flight 800, July 17, 1996

© T.E. BEARDEN 1996

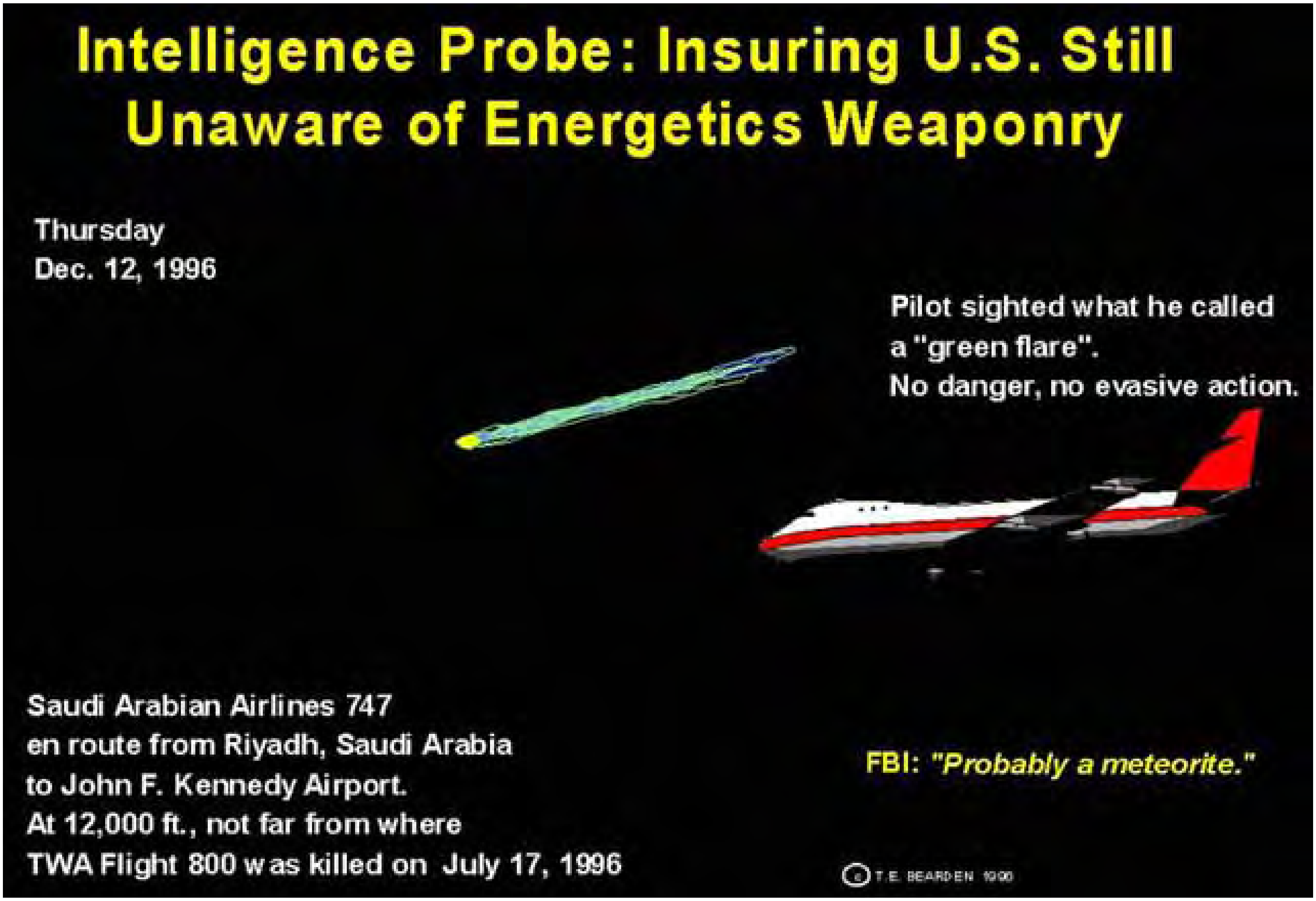

Intelligence Probe: Insuring U.S. Still Unaware of Energetics Weaponry

Thursday
Dec. 12, 1996

Pilot sighted what he called
a "green flare".
No danger, no evasive action.

Saudi Arabian Airlines 747
en route from Riyadh, Saudi Arabia
to John F. Kennedy Airport.
At 12,000 ft., not far from where
TWA Flight 800 was killed on July 17, 1996

FBI: *"Probably a meteorite."*

© T.E. BEARDEN 1996

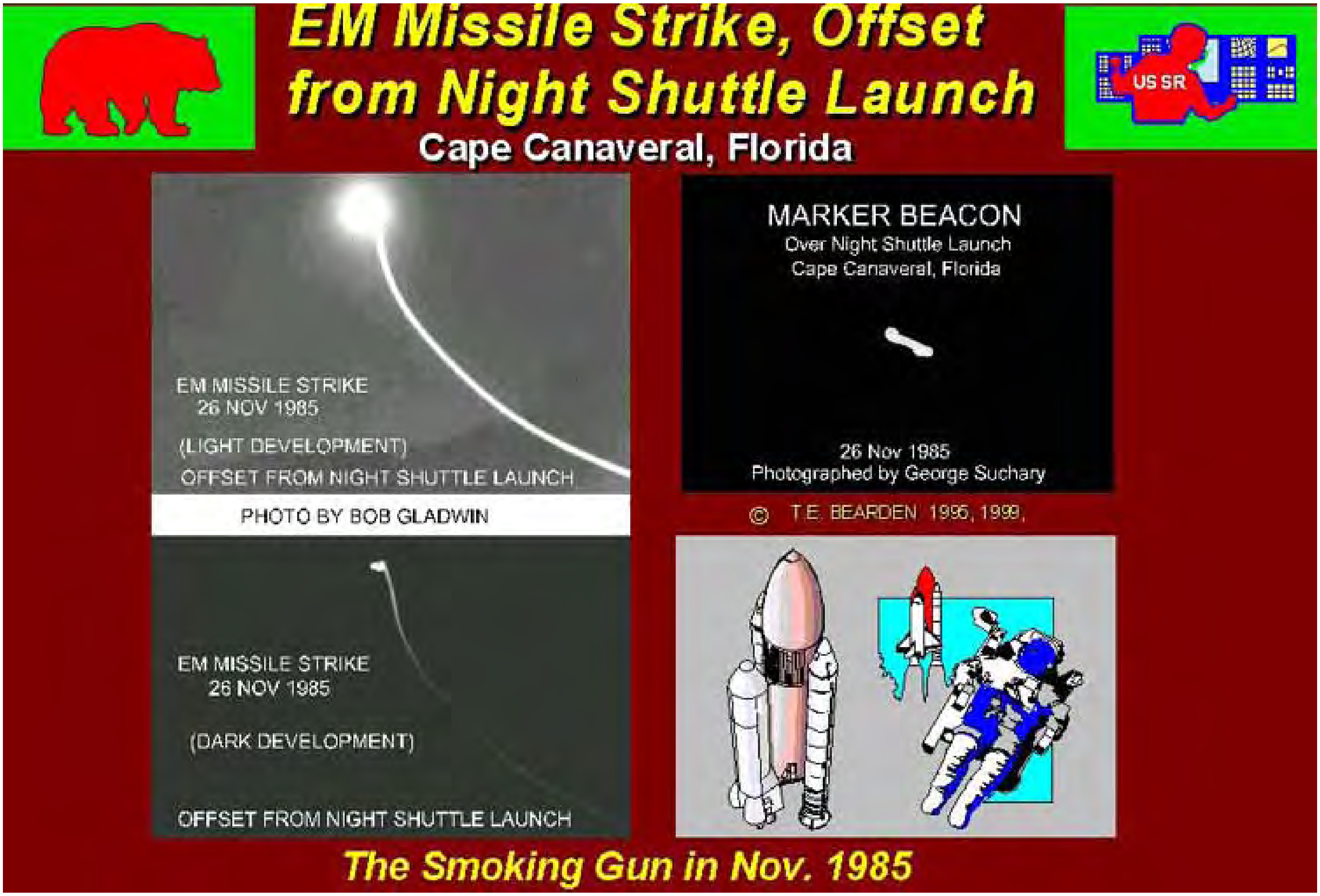

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)
OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1996, 1999,

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

The Smoking Gun in Nov. 1985

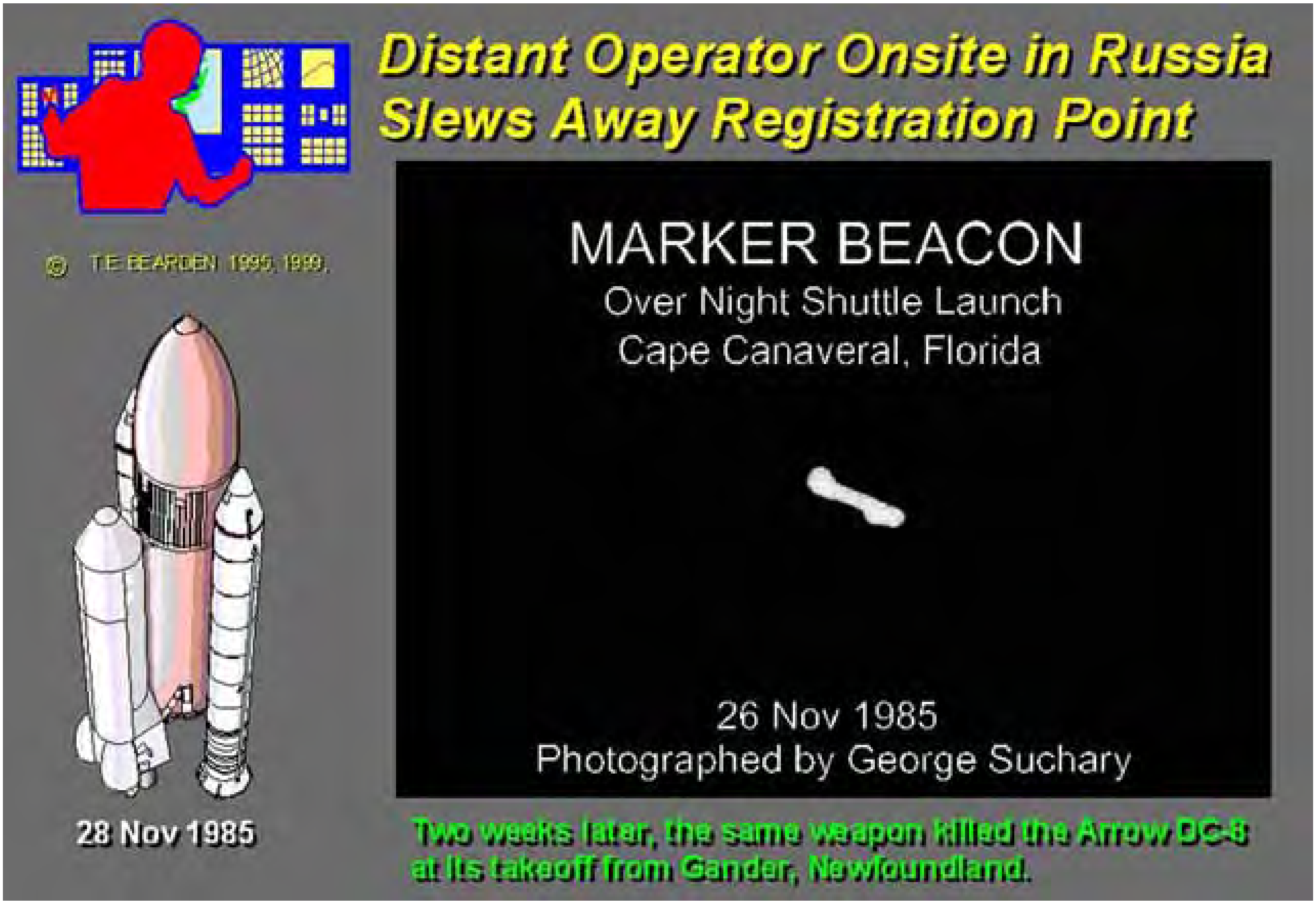

Distant Operator Onsite in Russia Slews Away Registration Point

© T.E. BEARDEN 1995, 1999.

MARKER BEACON

Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

28 Nov 1985

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

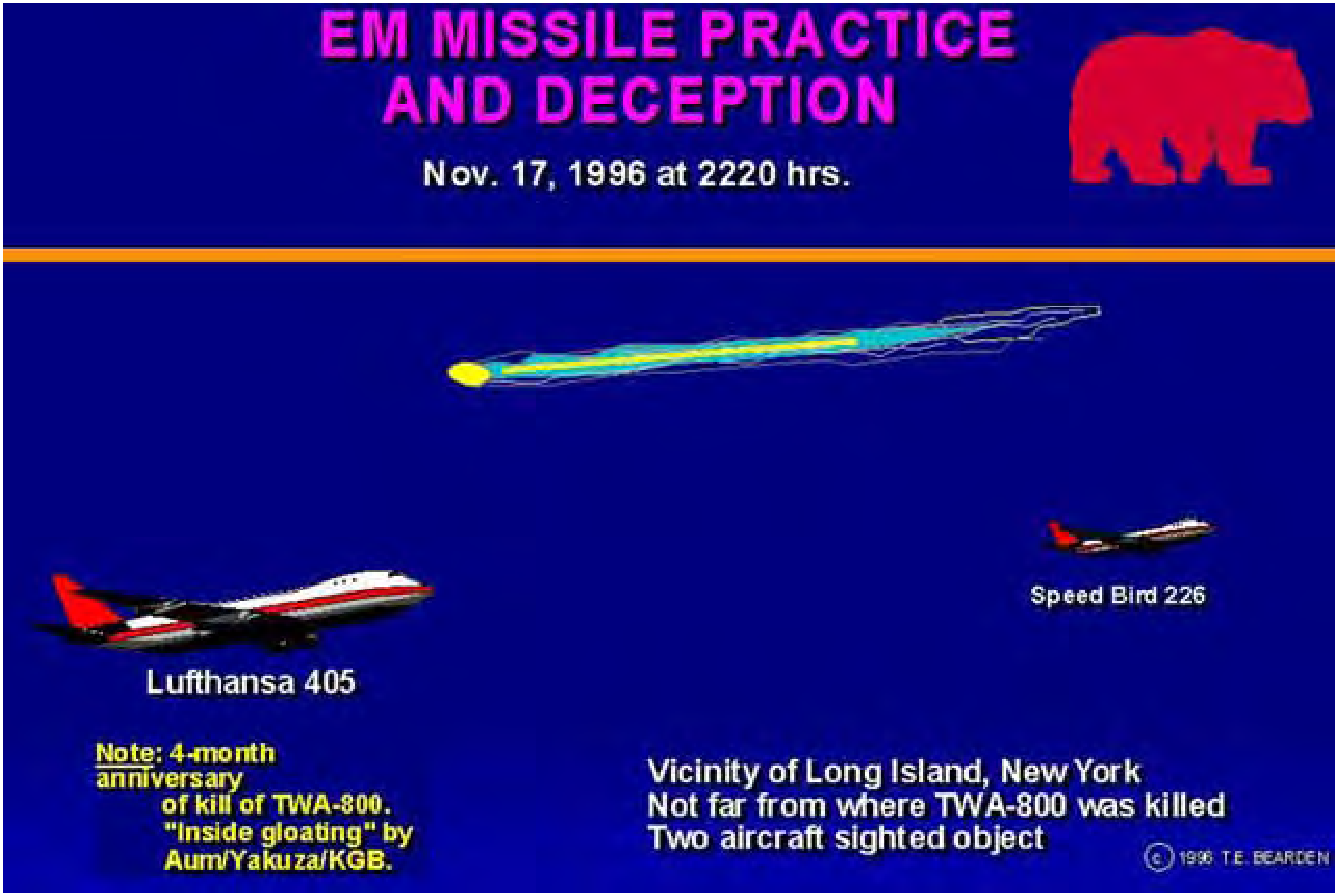

EM MISSILE PRACTICE AND DECEPTION

Nov. 17, 1996 at 2220 hrs.

Note: 4-month anniversary of kill of TWA-800. "Inside gloating" by Aum/Yakuza/KGB.

Vicinity of Long Island, New York
Not far from where TWA-800 was killed
Two aircraft sighted object

(c) 1996 T.E. BEARDEN

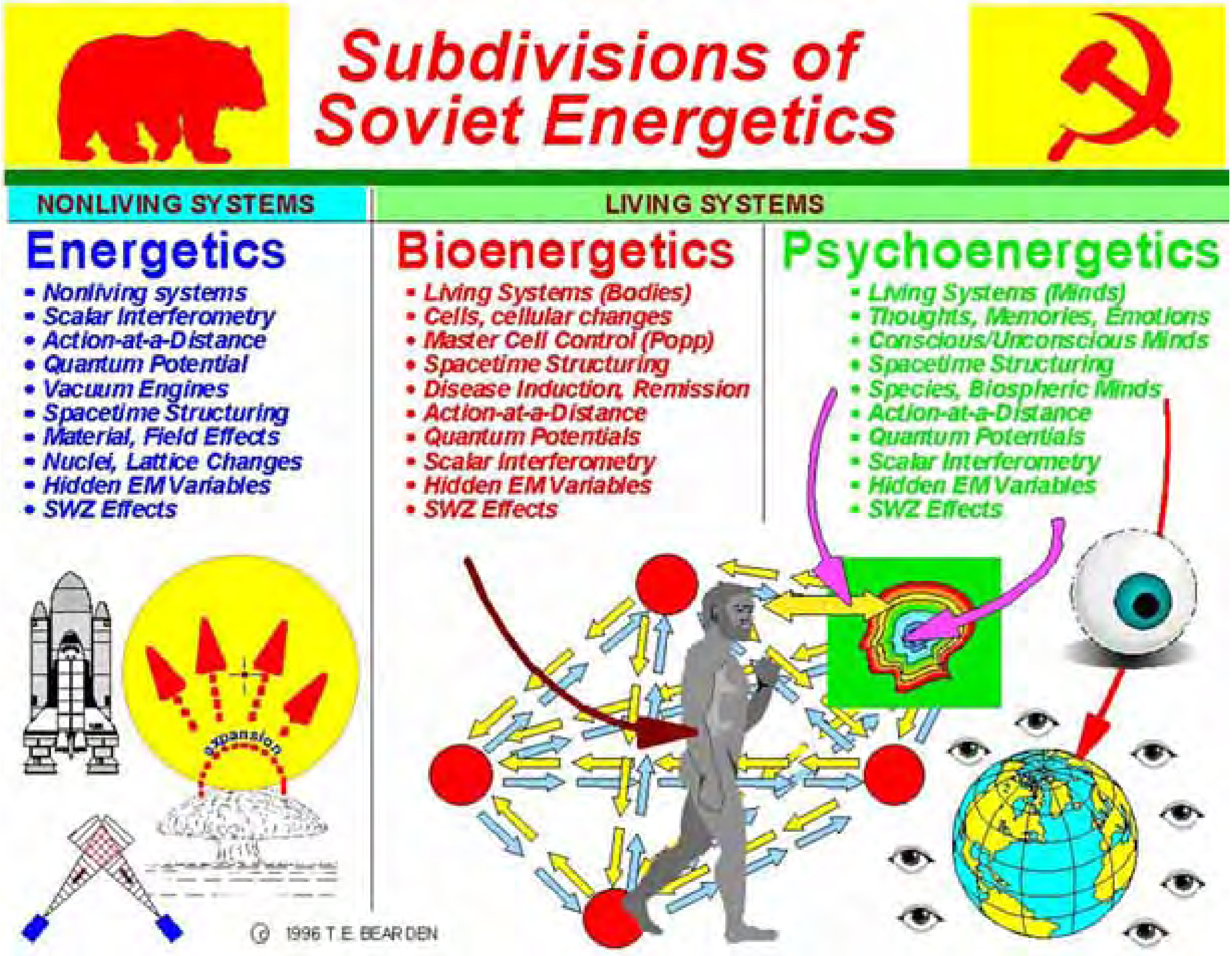

Subdivisions of Soviet Energetics

NONLIVING SYSTEMS

Energetics

- *Nonliving systems*
- *Scalar Interferometry*
- *Action-at-a-Distance*
- *Quantum Potential*
- *Vacuum Engines*
- *Spacetime Structuring*
- *Material, Field Effects*
- *Nuclei, Lattice Changes*
- *Hidden EM Variables*
- *SWZ Effects*

LIVING SYSTEMS

Bioenergetics

- *Living Systems (Bodies)*
- *Cells, cellular changes*
- *Master Cell Control (Popp)*
- *Spacetime Structuring*
- *Disease Induction, Remission*
- *Action-at-a-Distance*
- *Quantum Potentials*
- *Scalar Interferometry*
- *Hidden EM Variables*
- *SWZ Effects*

Psychoenergetics

- *Living Systems (Minds)*
- *Thoughts, Memories, Emotions*
- *Conscious/Unconscious Minds*
- *Spacetime Structuring*
- *Species, Biospheric Minds*
- *Action-at-a-Distance*
- *Quantum Potentials*
- *Scalar Interferometry*
- *Hidden EM Variables*
- *SWZ Effects*

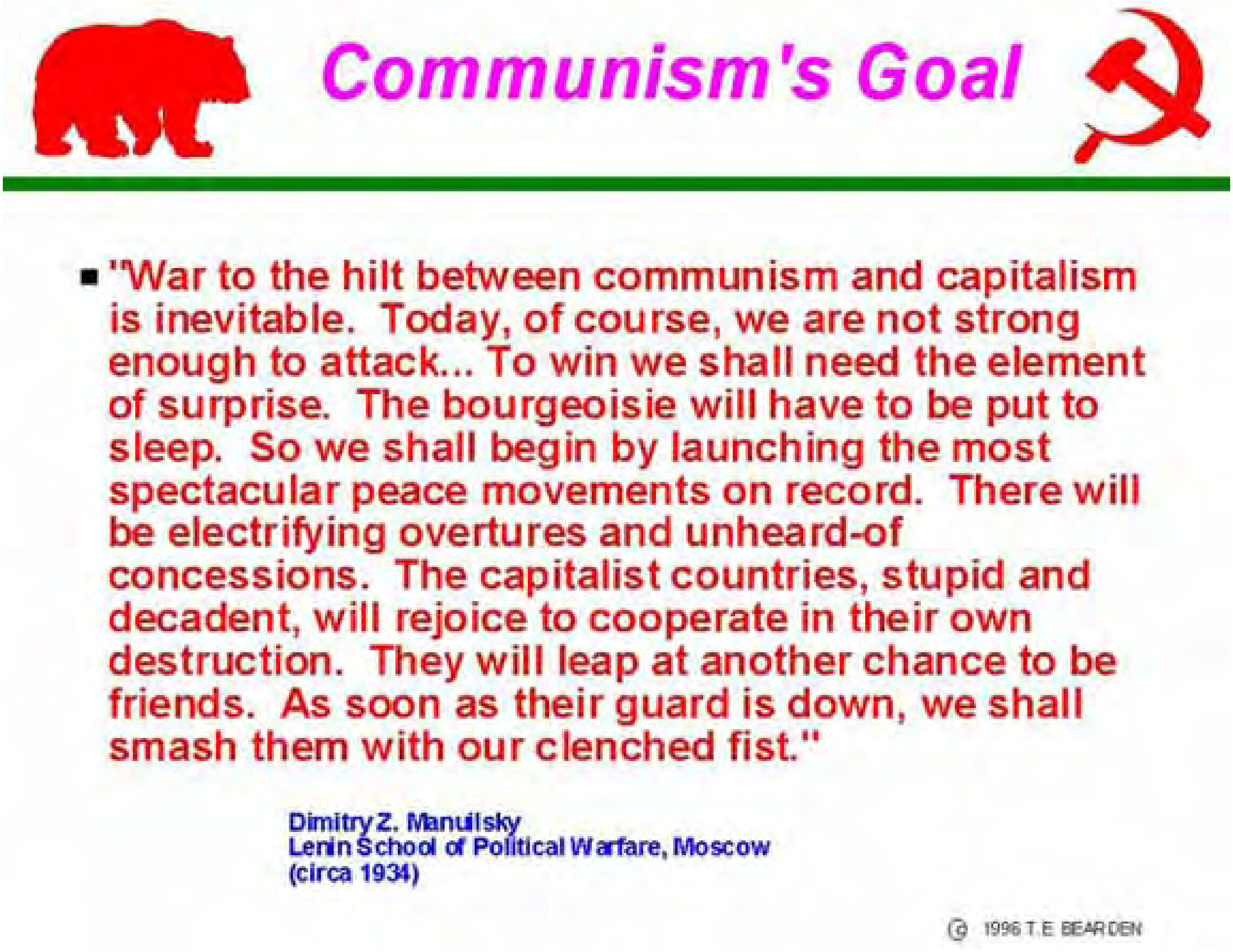

Communism's Goal

- "War to the hilt between communism and capitalism is inevitable. Today, of course, we are not strong enough to attack... To win we shall need the element of surprise. The bourgeoisie will have to be put to sleep. So we shall begin by launching the most spectacular peace movements on record. There will be electrifying overtures and unheard-of concessions. The capitalist countries, stupid and decadent, will rejoice to cooperate in their own destruction. They will leap at another chance to be friends. As soon as their guard is down, we shall smash them with our clenched fist."

Dimitry Z. Manuilsky
Lenin School of Political Warfare, Moscow
(circa 1934)

© 1996 T.E. BEARDEN

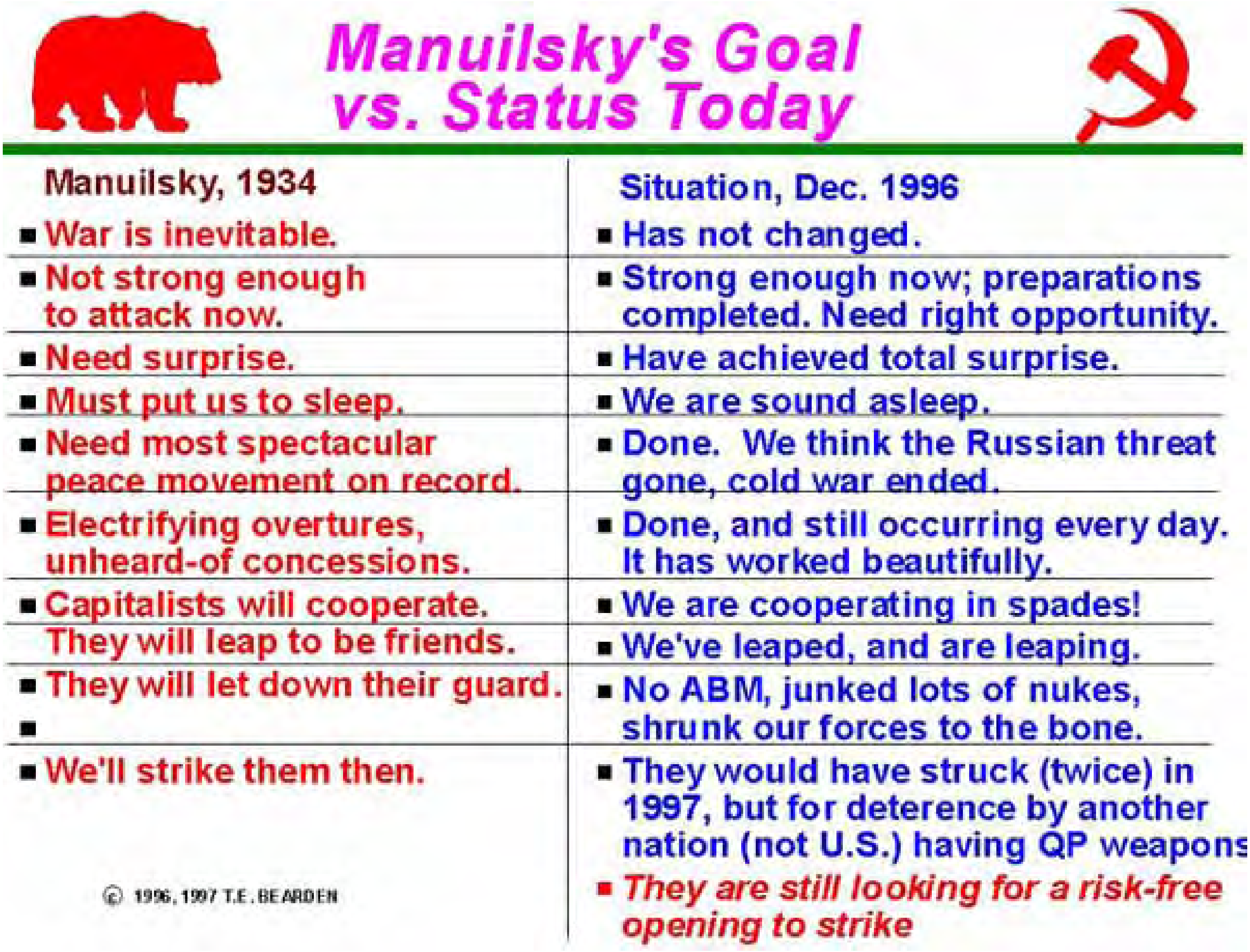

Manuilsky's Goal vs. Status Today

Manuilsky, 1934	Situation, Dec. 1996
■ War is inevitable.	■ Has not changed.
■ Not strong enough to attack now.	■ Strong enough now; preparations completed. Need right opportunity.
■ Need surprise.	■ Have achieved total surprise.
■ Must put us to sleep.	■ We are sound asleep.
■ Need most spectacular peace movement on record.	■ Done. We think the Russian threat gone, cold war ended.
■ Electrifying overtures, unheard-of concessions.	■ Done, and still occurring every day. It has worked beautifully.
■ Capitalists will cooperate.	■ We are cooperating in spades!
They will leap to be friends.	■ We've leaped, and are leaping.
■ They will let down their guard. ■	■ No ABM, junked lots of nukes, shrunk our forces to the bone.
■ We'll strike them then.	■ They would have struck (twice) in 1997, but for deterence by another nation (not U.S.) having QP weapons ■ *They are still looking for a risk-free opening to strike*

© 1996, 1997 T.E. BEARDEN

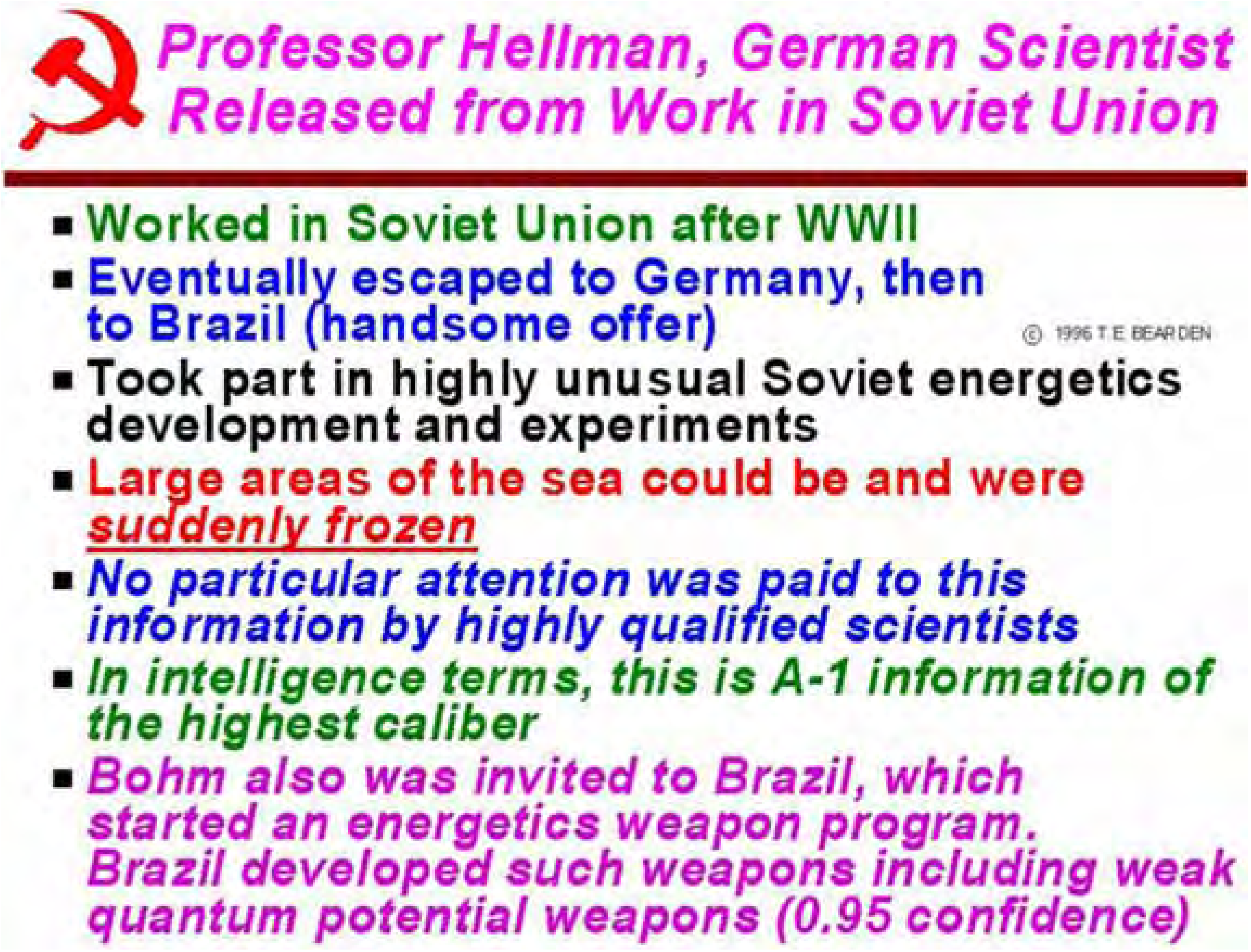

Professor Hellman, German Scientist Released from Work in Soviet Union

- Worked in Soviet Union after WWII
- Eventually escaped to Germany, then to Brazil (handsome offer)
- Took part in highly unusual Soviet energetics development and experiments
- Large areas of the sea could be and were *suddenly frozen*
- *No particular attention was paid to this information by highly qualified scientists*
- *In intelligence terms, this is A-1 information of the highest caliber*
- *Bohm also was invited to Brazil, which started an energetics weapon program. Brazil developed such weapons including weak quantum potential weapons (0.95 confidence)*

© 1996 T. E. BEARDEN

Hemispheres of Light

- March 20th, 1969
  Caribbean and Western North Atlantic
- Hemisphere and Globes, 1977
  24 March, 1977
- Virgin Islands Incident
  August 1969
- Typical Shield Test, Northern Route to Japan
  June, 1982

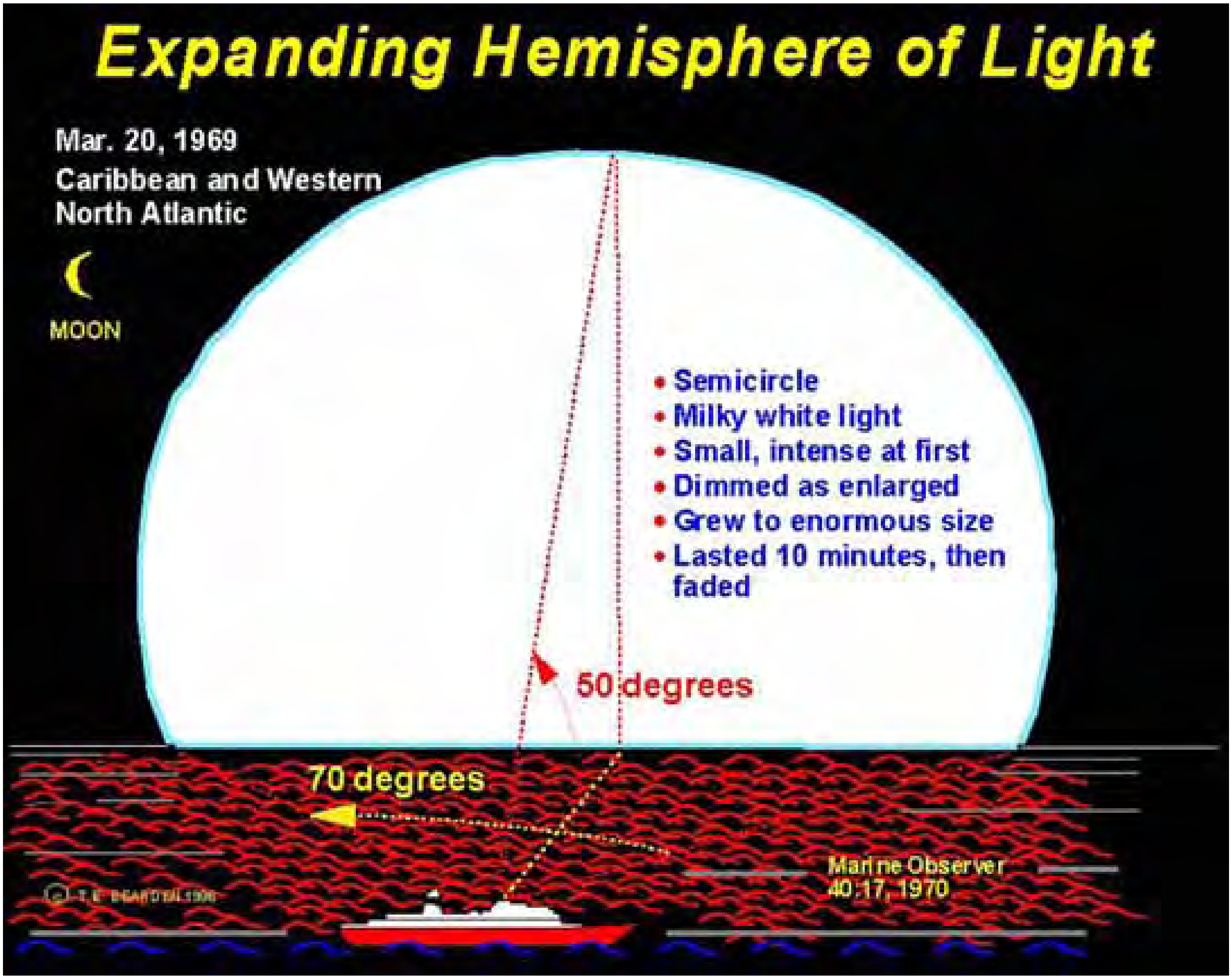

Expanding Hemisphere of Light

Mar. 20, 1969

Caribbean and Western
North Atlantic

- Semicircle
- Milky white light
- Small, intense at first
- Dimmed as enlarged
- Grew to enormous size
- Lasted 10 minutes, then faded

Marine Observer
40:17, 1970

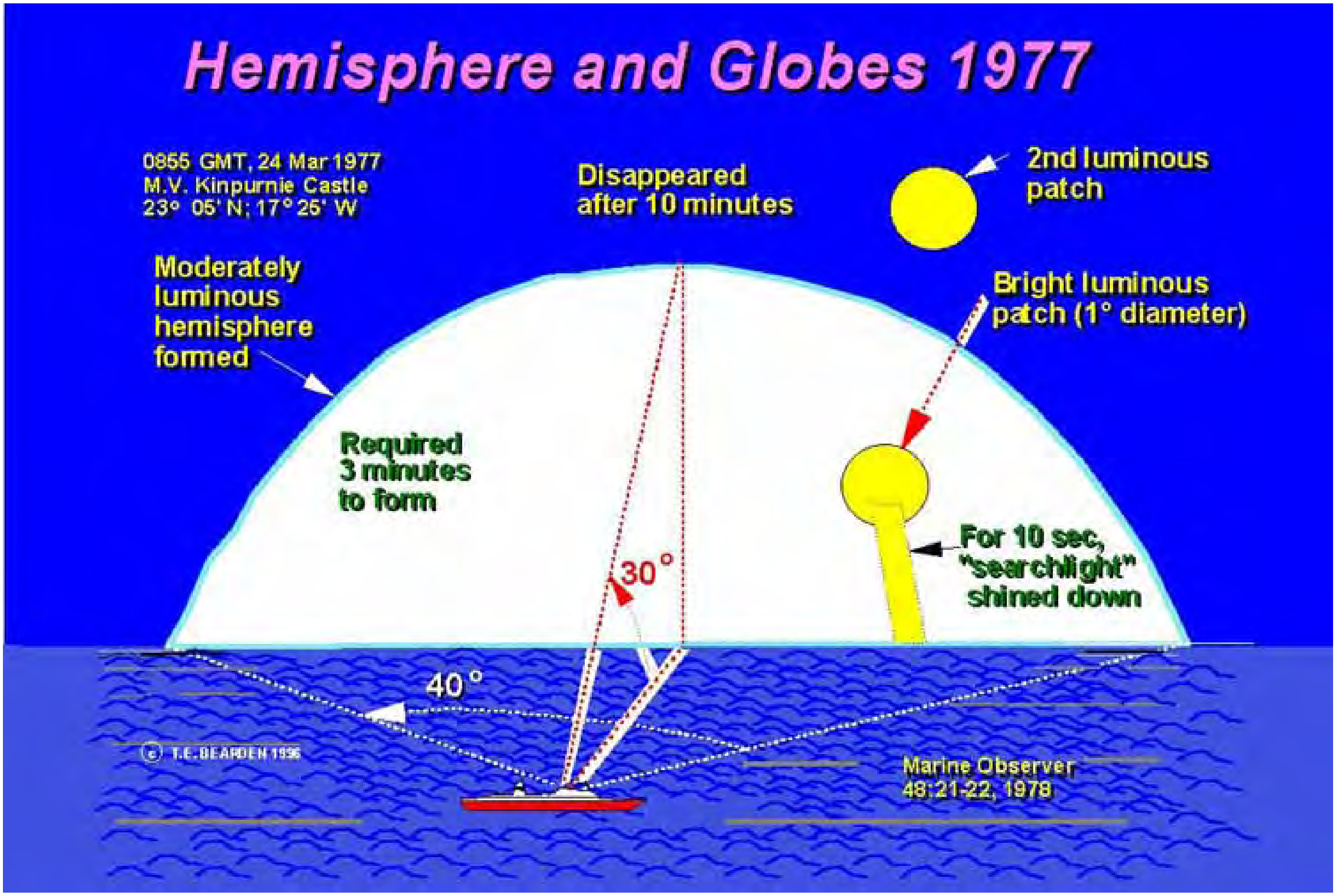
Hemisphere and Globes 1977
0855 GMT, 24 Mar 1977
M.V. Kinpurnie Castle
23° 05' N; 17° 25' W
Disappeared
after 10 minutes
2nd luminous
patch
Moderately
luminous
hemisphere
formed
Bright luminous
patch (1° diameter)
Required
3 minutes
to form
For 10 sec,
"searchlight"
shined down
30°
40°
© T.E. BEARDEN 1996
Marine Observer
48:21-22, 1978

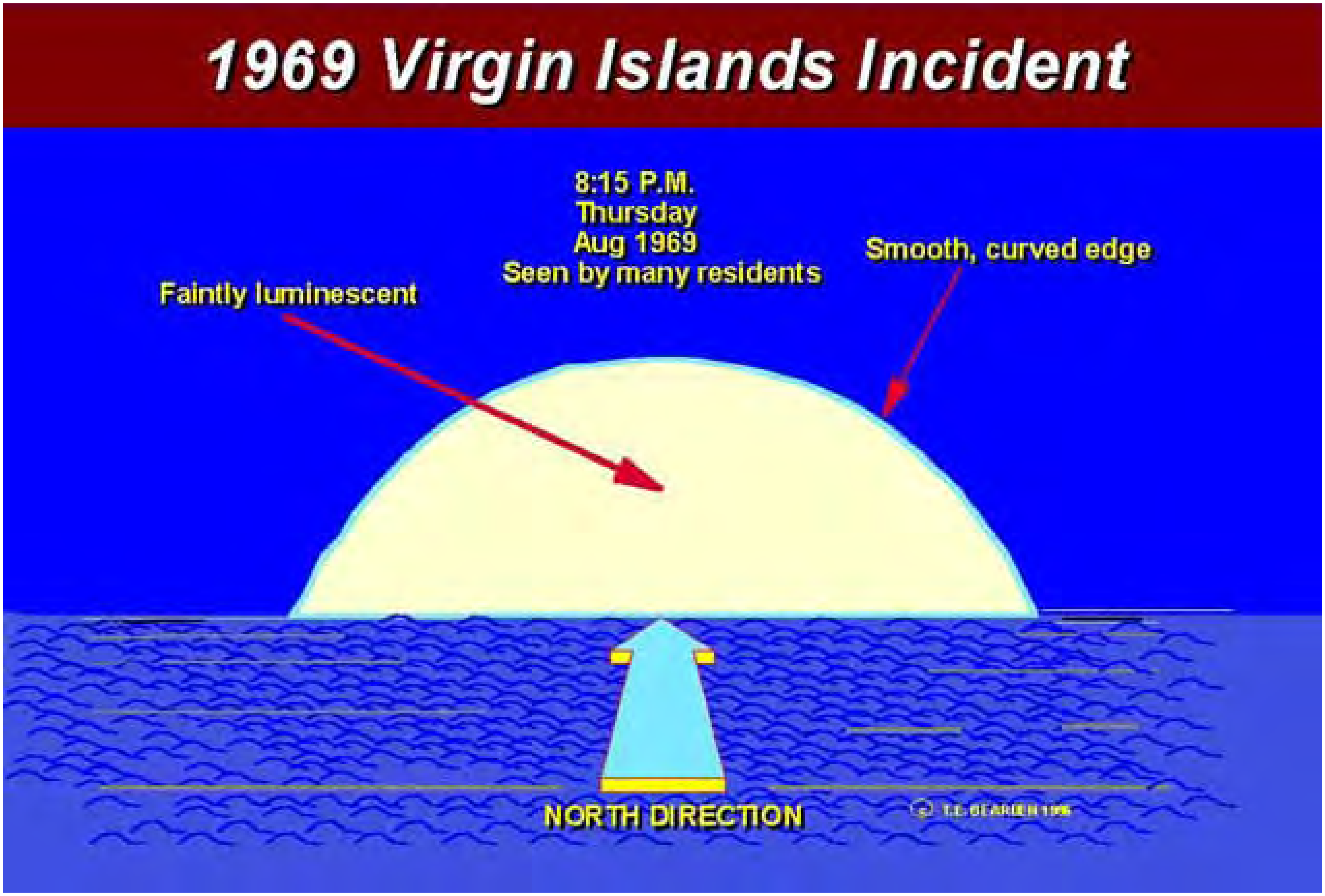

1969 Virgin Islands Incident

8:15 P.M.
Thursday
Aug 1969
Seen by many residents

Faintly luminescent

Smooth, curved edge

NORTH DIRECTION

© T.E. BEARDEN [illegible]

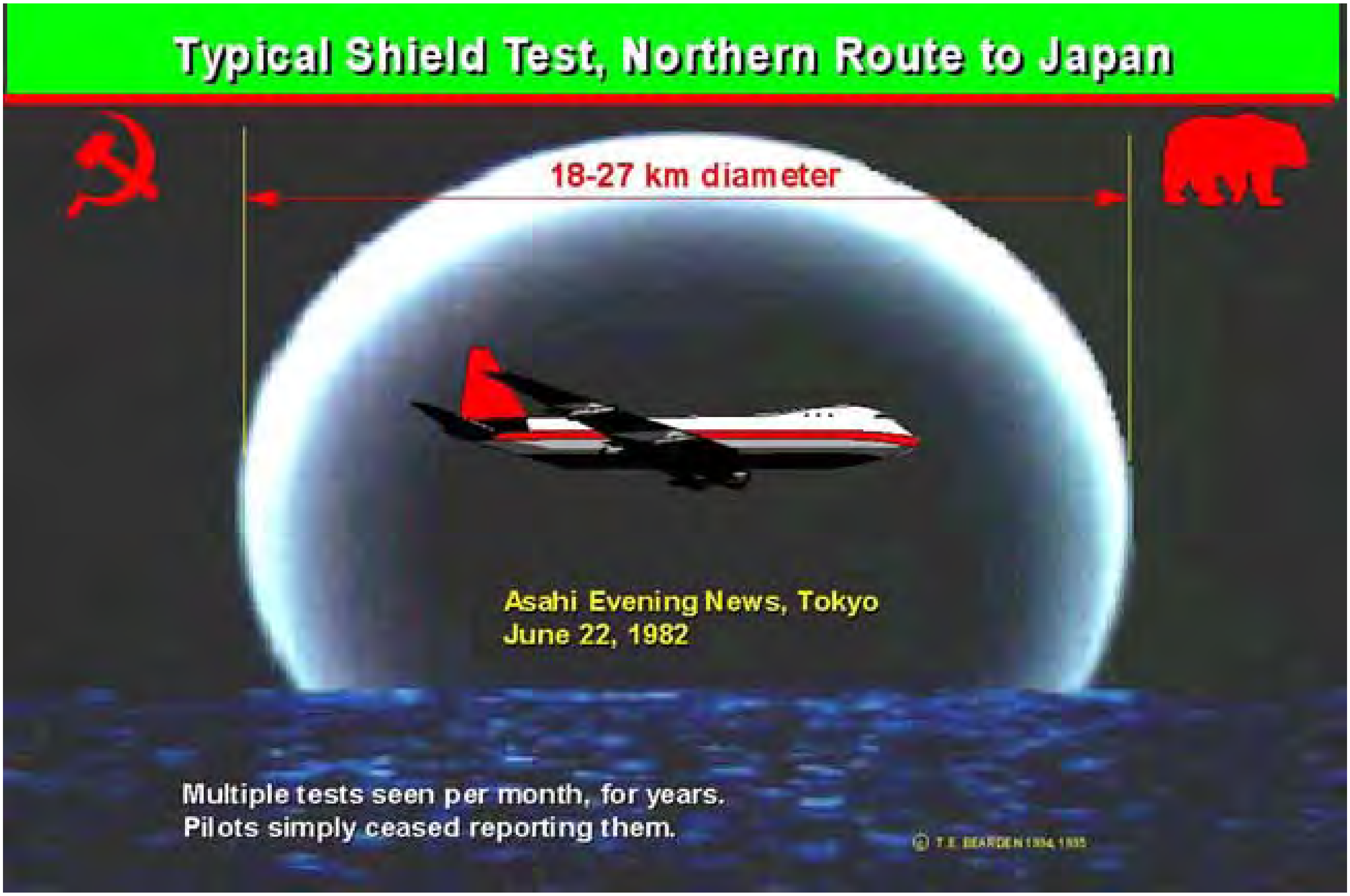

Typical Shield Test, Northern Route to Japan

18-27 km diameter

Asahi Evening News, Tokyo
June 22, 1982

Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

© T.E. BEARDEN 1994, 1995

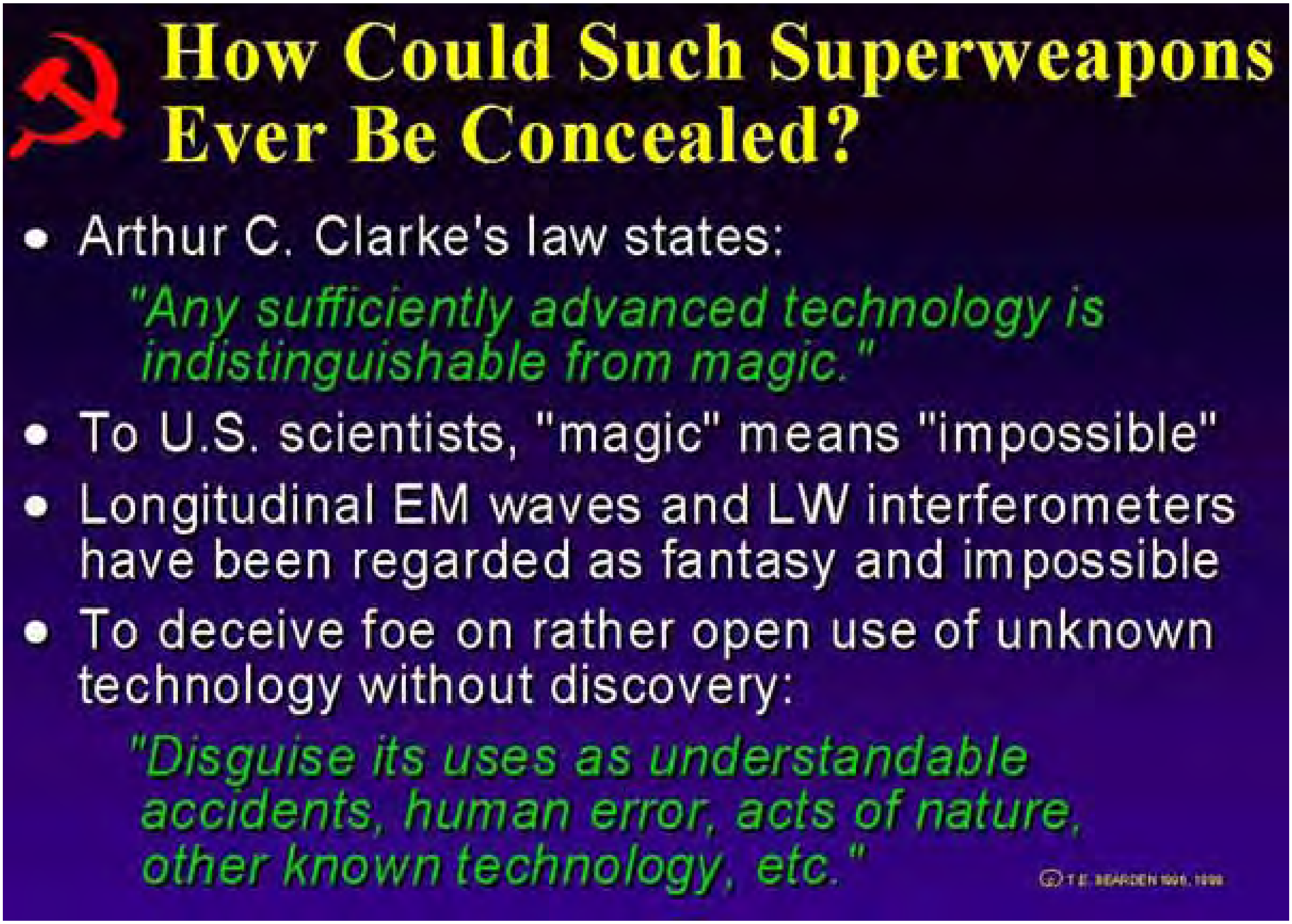

How Could Such Superweapons Ever Be Concealed?

- Arthur C. Clarke's law states:

  "Any sufficiently advanced technology is indistinguishable from magic."
- To U.S. scientists, "magic" means "impossible"
- Longitudinal EM waves and LW interferometers have been regarded as fantasy and impossible
- To deceive foe on rather open use of unknown technology without discovery:

  "Disguise its uses as understandable accidents, human error, acts of nature, other known technology, etc."

© T.E. BEARDEN 1996, 1998

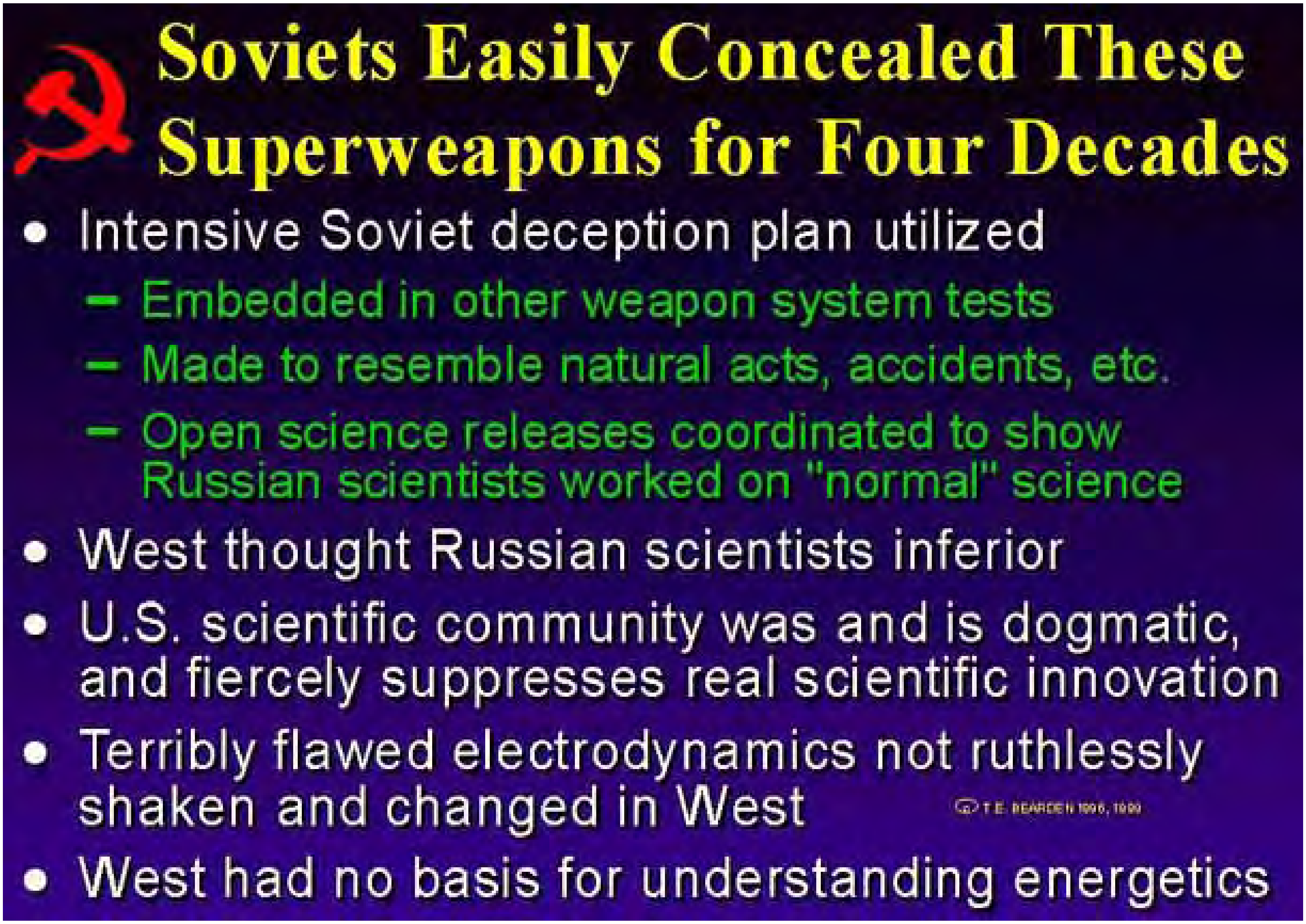

Soviets Easily Concealed These Superweapons for Four Decades

- Intensive Soviet deception plan utilized
  - Embedded in other weapon system tests
  - Made to resemble natural acts, accidents, etc.
  - Open science releases coordinated to show Russian scientists worked on "normal" science
- West thought Russian scientists inferior
- U.S. scientific community was and is dogmatic, and fiercely suppresses real scientific innovation
- Terribly flawed electrodynamics not ruthlessly shaken and changed in West
- West had no basis for understanding energetics

© T.E. BEARDEN 1996, 1999

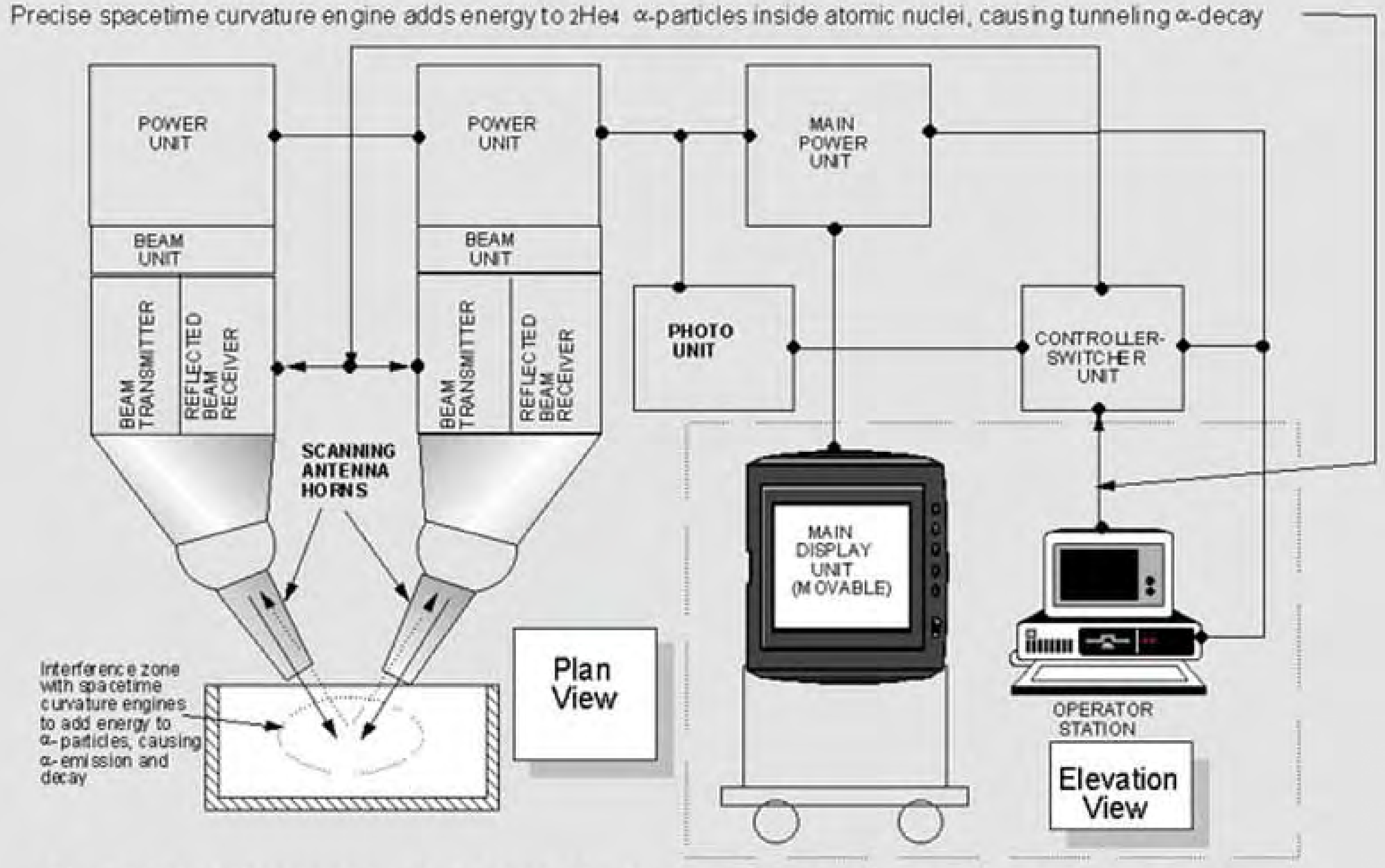
Precise spacetime curvature engine adds energy to 2He4 α-particles inside atomic nuclei, causing tunneling α-decay
POWER UNIT
POWER UNIT
MAIN POWER UNIT
BEAM UNIT
BEAM UNIT
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
BEAM TRANSMITTER
REFLECTED BEAM RECEIVER
PHOTO UNIT
CONTROLLER-SWITCHER UNIT
SCANNING ANTENNA HORNS
MAIN DISPLAY UNIT (MOVABLE)
OPERATOR STATION
Interference zone with spacetime curvature engines to add energy to α-particles, causing α-emission and decay
Plan View
Elevation View

Figure 26. Microscope-type, laser-like interferometer for inducing quick α-decay in samples of otherwise longer-lived isotopes.

© T.E. BEARDEN 1996

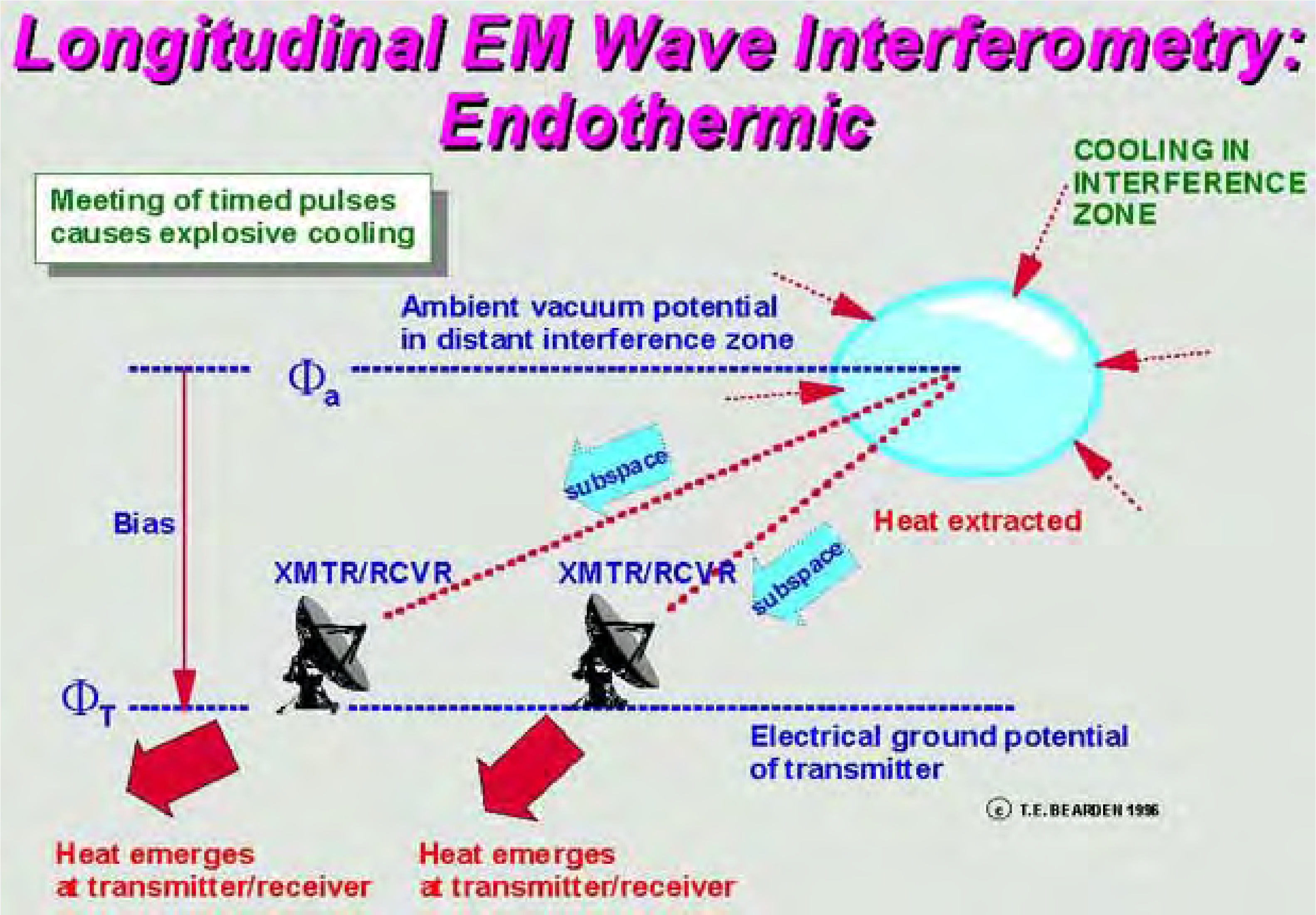

Longitudinal EM Wave Interferometry: Endothermic

Meeting of timed pulses causes explosive cooling

COOLING IN INTERFERENCE ZONE

Ambient vacuum potential in distant interference zone

Heat extracted

Electrical ground potential of transmitter

Heat emerges at transmitter/receiver

Heat emerges at transmitter/receiver

Φ_a
subspace
Bias
XMTR/RCVR
XMTR/RCVR
subspace
Φ_T

© T.E. BEARDEN 1996

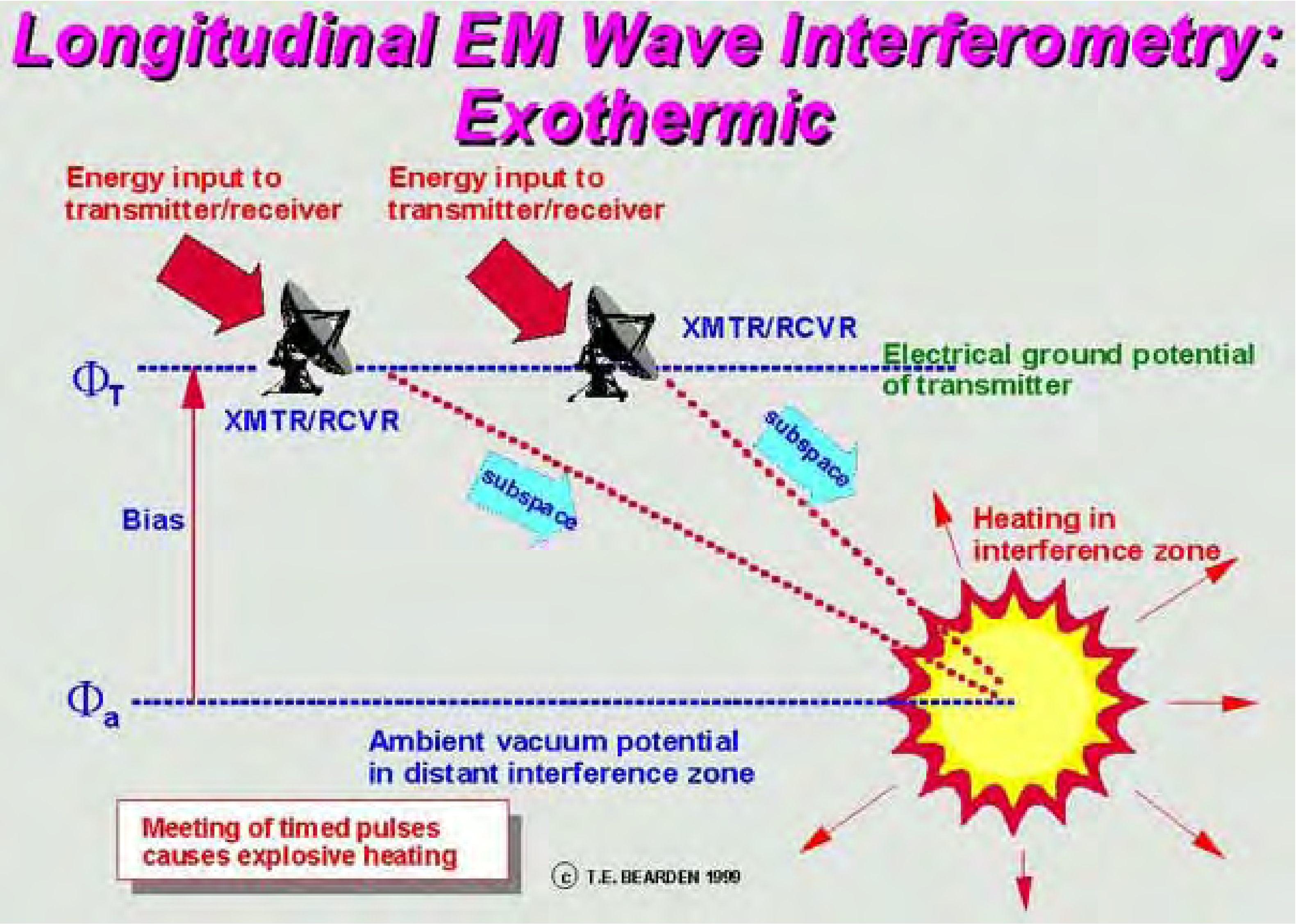

Longitudinal EM Wave Interferometry: Exothermic

Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
XMTR/RCVR
Φ_T
Electrical ground potential of transmitter
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone

Meeting of timed pulses causes explosive heating

© T.E. BEARDEN 1999

The Tom Bearden Website

Help support the research

U.S. Air Attack on Libya

April 1986

- Slide 1
- Slide 2
- "Washington Post" Article
  January 22, 1989

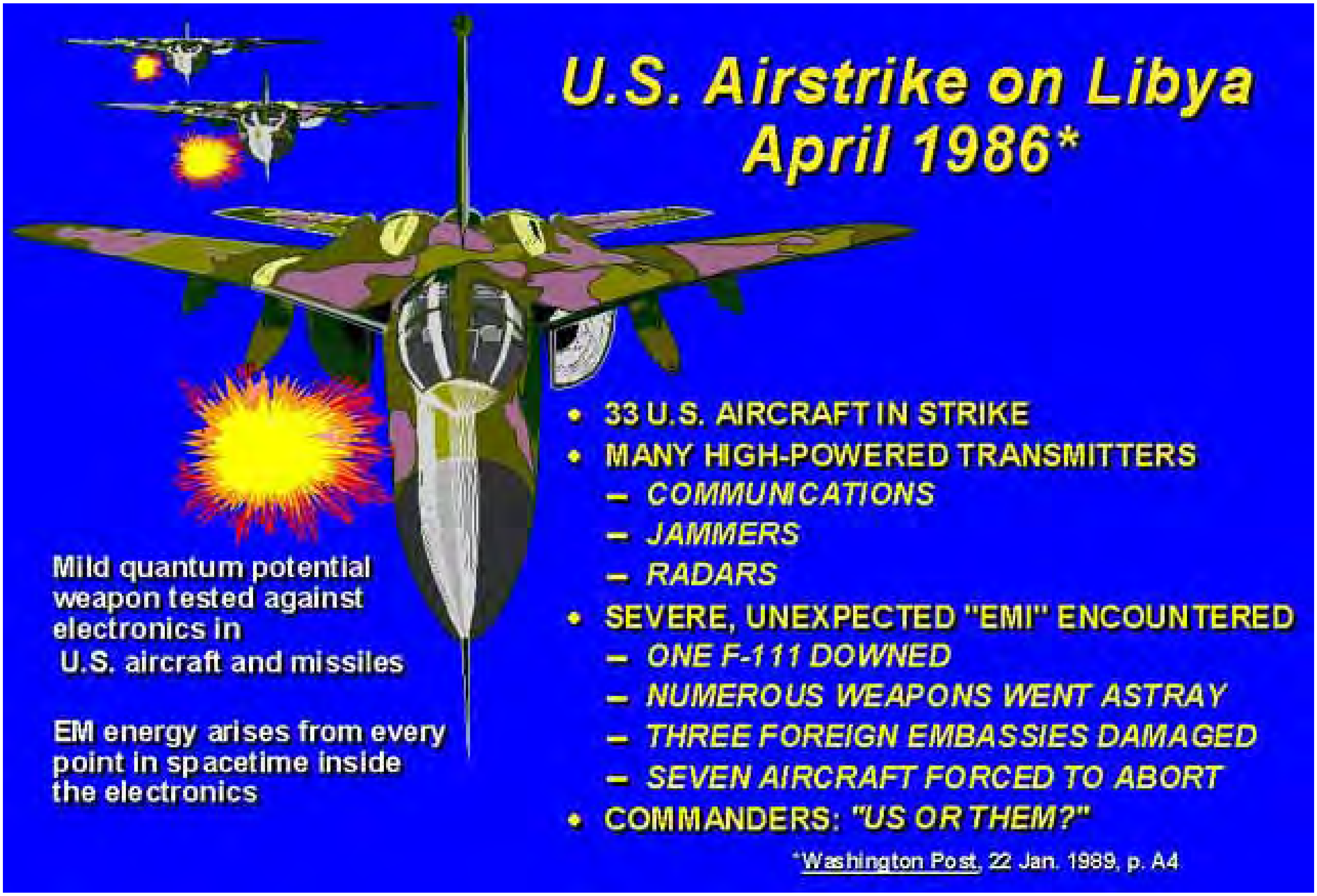

U.S. Airstrike on Libya April 1986*

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - *COMMUNICATIONS*
  - *JAMMERS*
  - *RADARS*
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - *ONE F-111 DOWNED*
  - *NUMEROUS WEAPONS WENT ASTRAY*
  - *THREE FOREIGN EMBASSIES DAMAGED*
  - *SEVEN AIRCRAFT FORCED TO ABORT*
- COMMANDERS: *"US OR THEM?"*

Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles

EM energy arises from every point in spacetime inside the electronics

*Washington Post, 22 Jan. 1989, p. A4

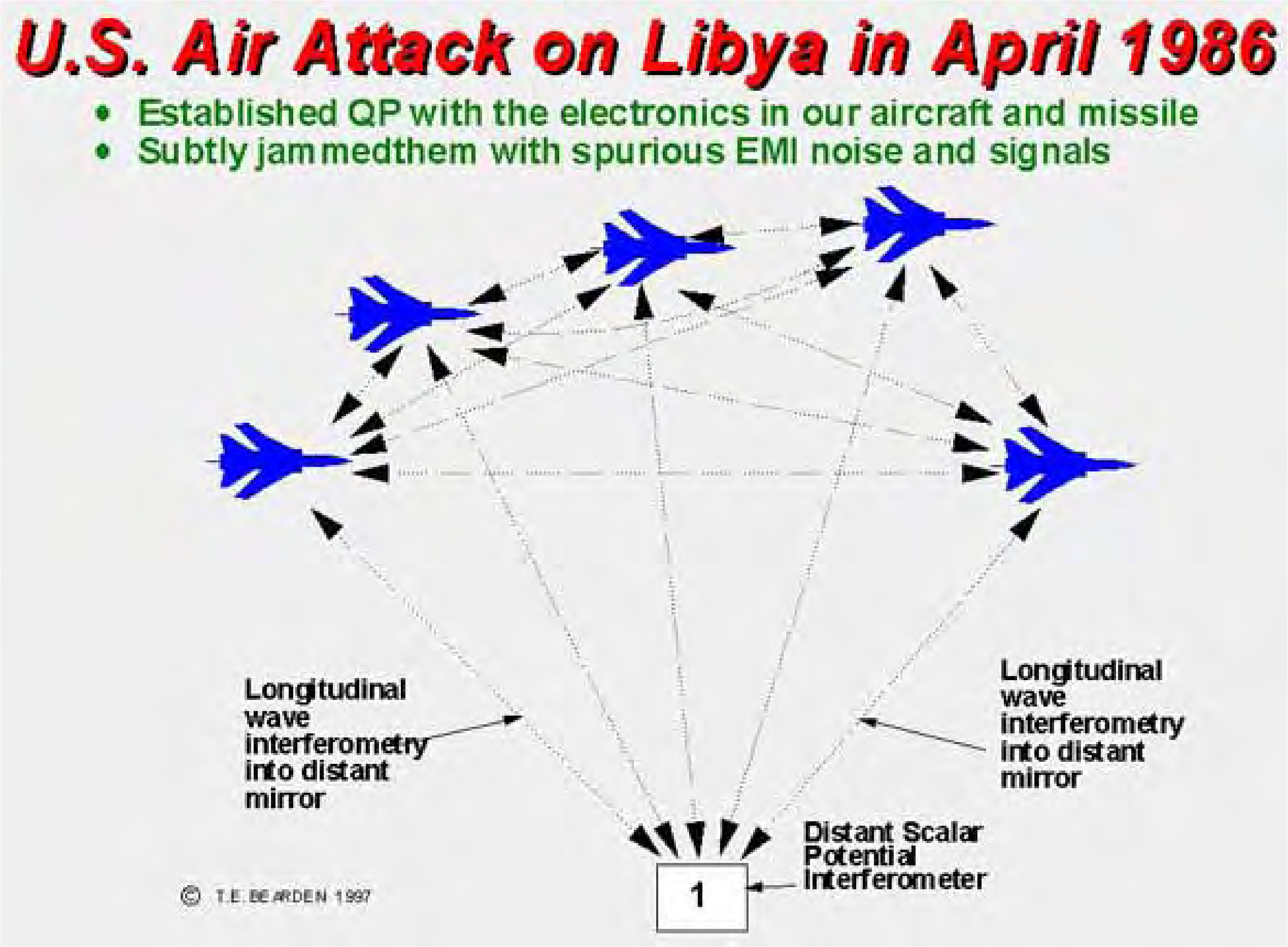

U.S. Air Attack on Libya in April 1986

- Established QP with the electronics in our aircraft and missile
- Subtly jammedthem with spurious EMI noise and signals

Longitudinal
wave
interferometry
into distant
mirror

Longitudinal
wave
interferometry
into distant
mirror

Distant Scalar
Potential
Interferometer

1

© T.E. BEARDEN 1997

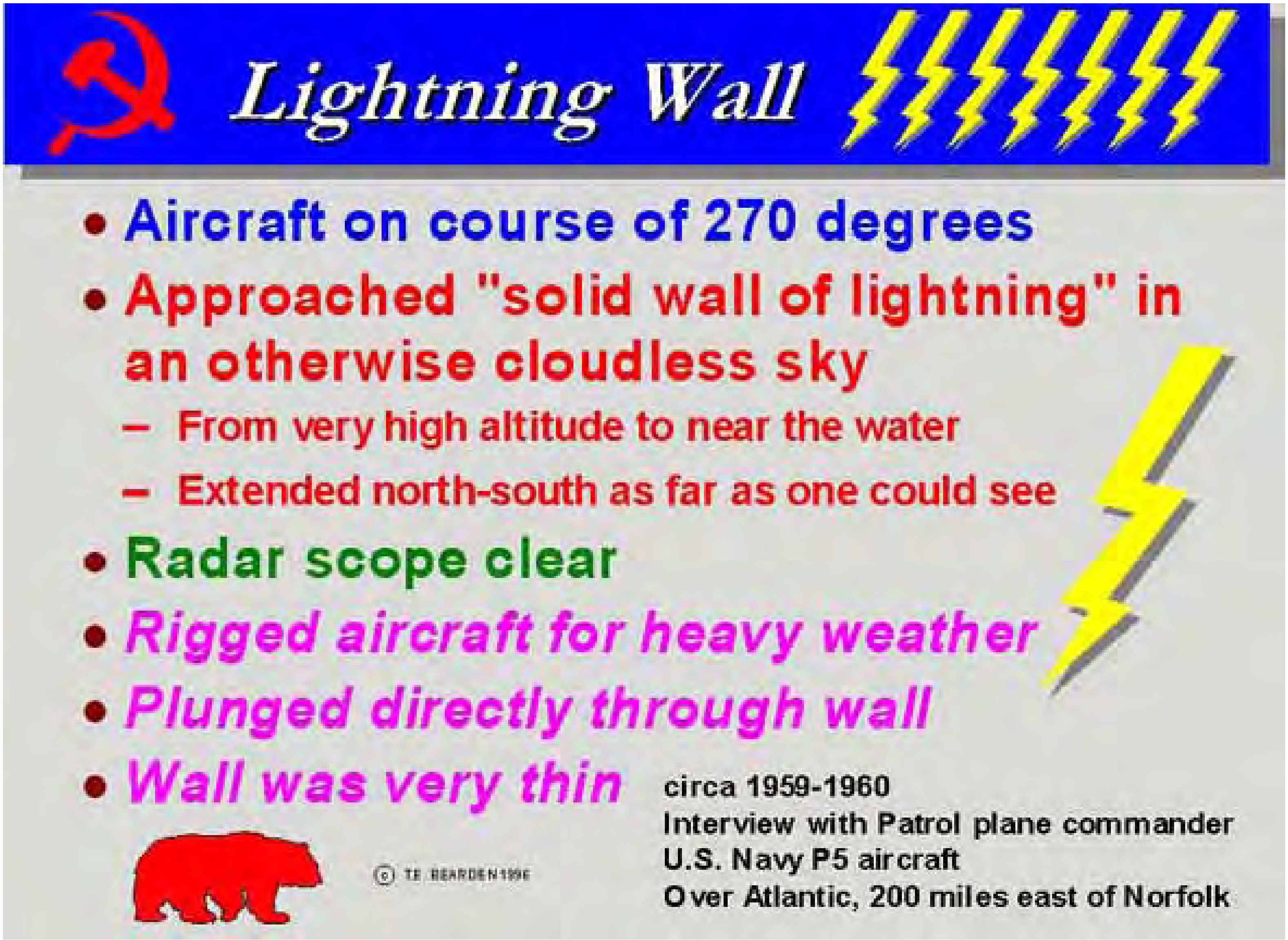

Lightning Wall

- Aircraft on course of 270 degrees
- Approached "solid wall of lightning" in an otherwise cloudless sky
  - From very high altitude to near the water
  - Extended north-south as far as one could see
- Radar scope clear
- *Rigged aircraft for heavy weather*
- *Plunged directly through wall*
- *Wall was very thin*

circa 1959-1960
Interview with Patrol plane commander
U.S. Navy P5 aircraft
Over Atlantic, 200 miles east of Norfolk

© T.E. BEARDEN 1996

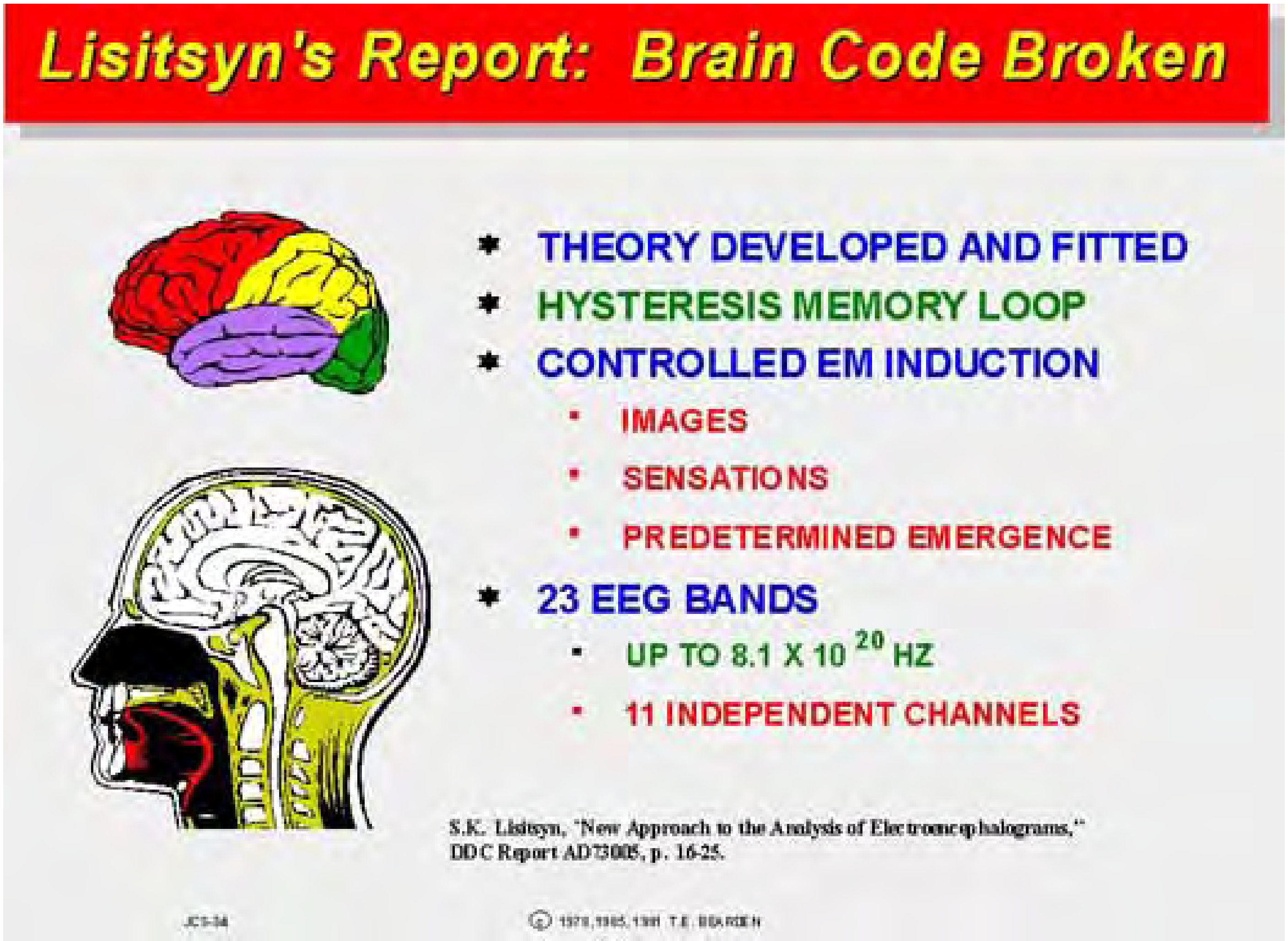

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

JCS-34

© 1978, 1985, 1991 T.E. BEARDEN

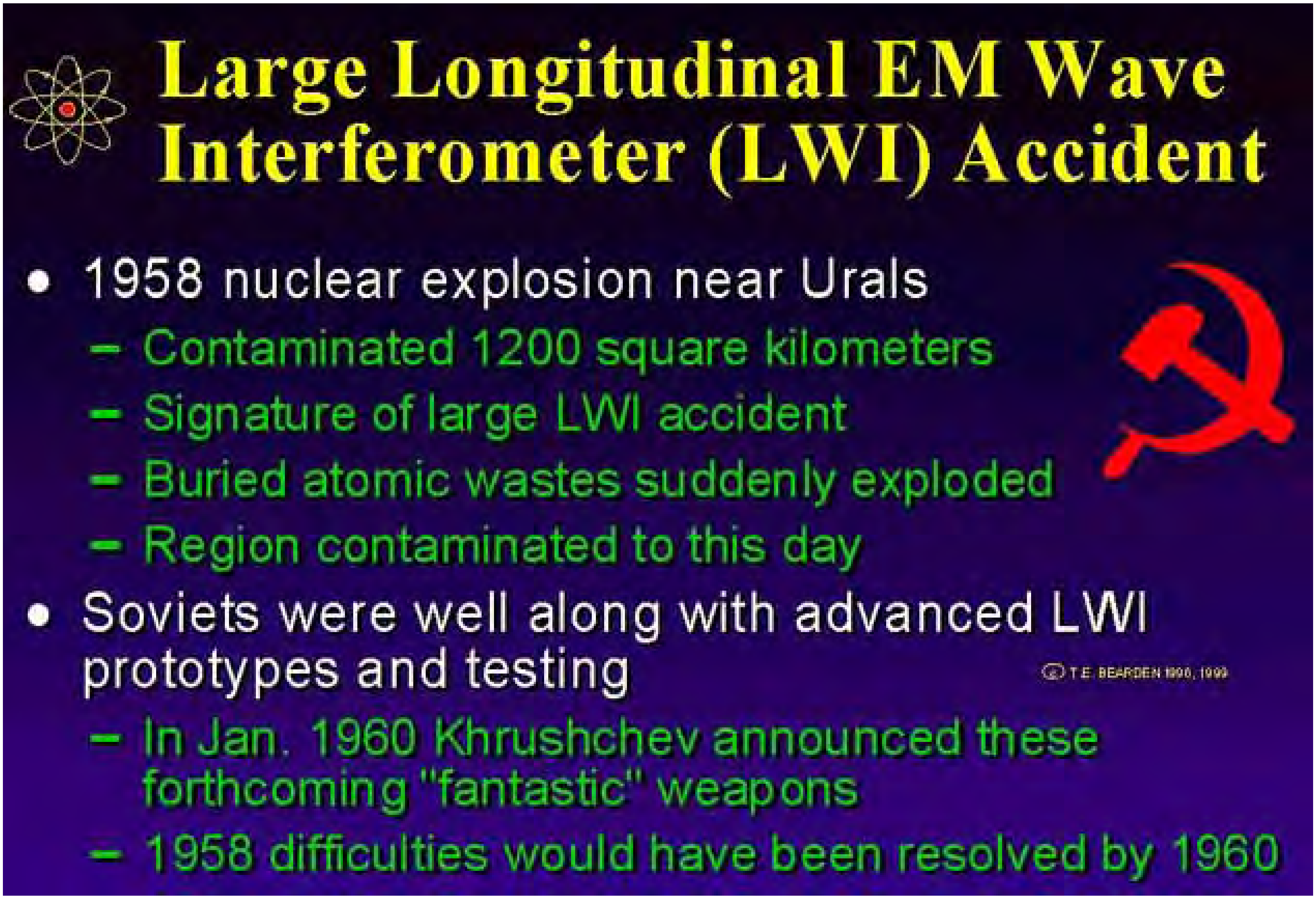

Large Longitudinal EM Wave Interferometer (LWI) Accident

- 1958 nuclear explosion near Urals
  - Contaminated 1200 square kilometers
  - Signature of large LWI accident
  - Buried atomic wastes suddenly exploded
  - Region contaminated to this day
- Soviets were well along with advanced LWI prototypes and testing
  - In Jan. 1960 Khrushchev announced these forthcoming "fantastic" weapons
  - 1958 difficulties would have been resolved by 1960

© T.E. BEARDEN 1996, 1999

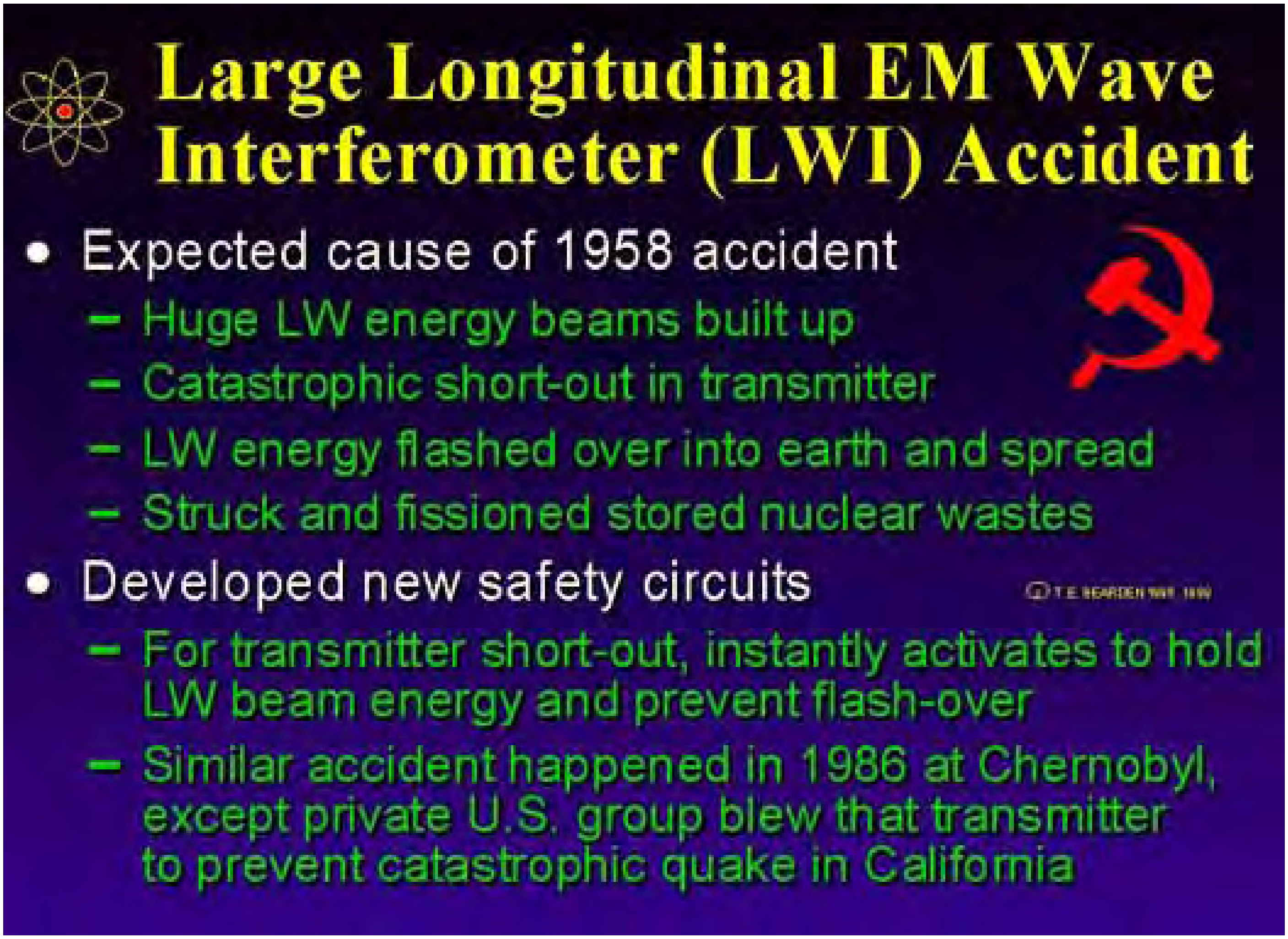

Large Longitudinal EM Wave Interferometer (LWI) Accident

- Expected cause of 1958 accident
  - Huge LW energy beams built up
  - Catastrophic short-out in transmitter
  - LW energy flashed over into earth and spread
  - Struck and fissioned stored nuclear wastes
- Developed new safety circuits
  - For transmitter short-out, instantly activates to hold LW beam energy and prevent flash-over
  - Similar accident happened in 1986 at Chernobyl, except private U.S. group blew that transmitter to prevent catastrophic quake in California

© T.E. BEARDEN [illegible]

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

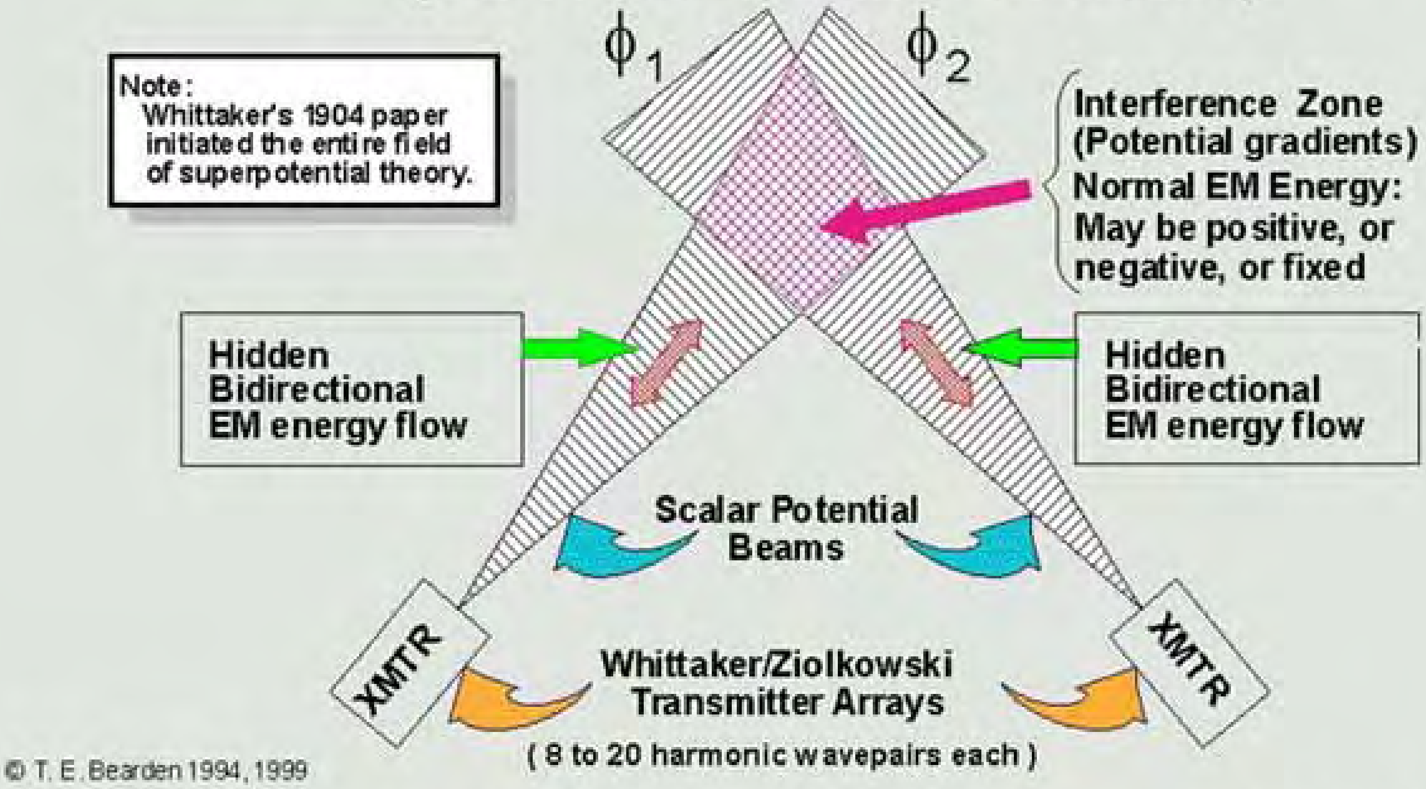
ϕ_1
ϕ_2
Note: Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

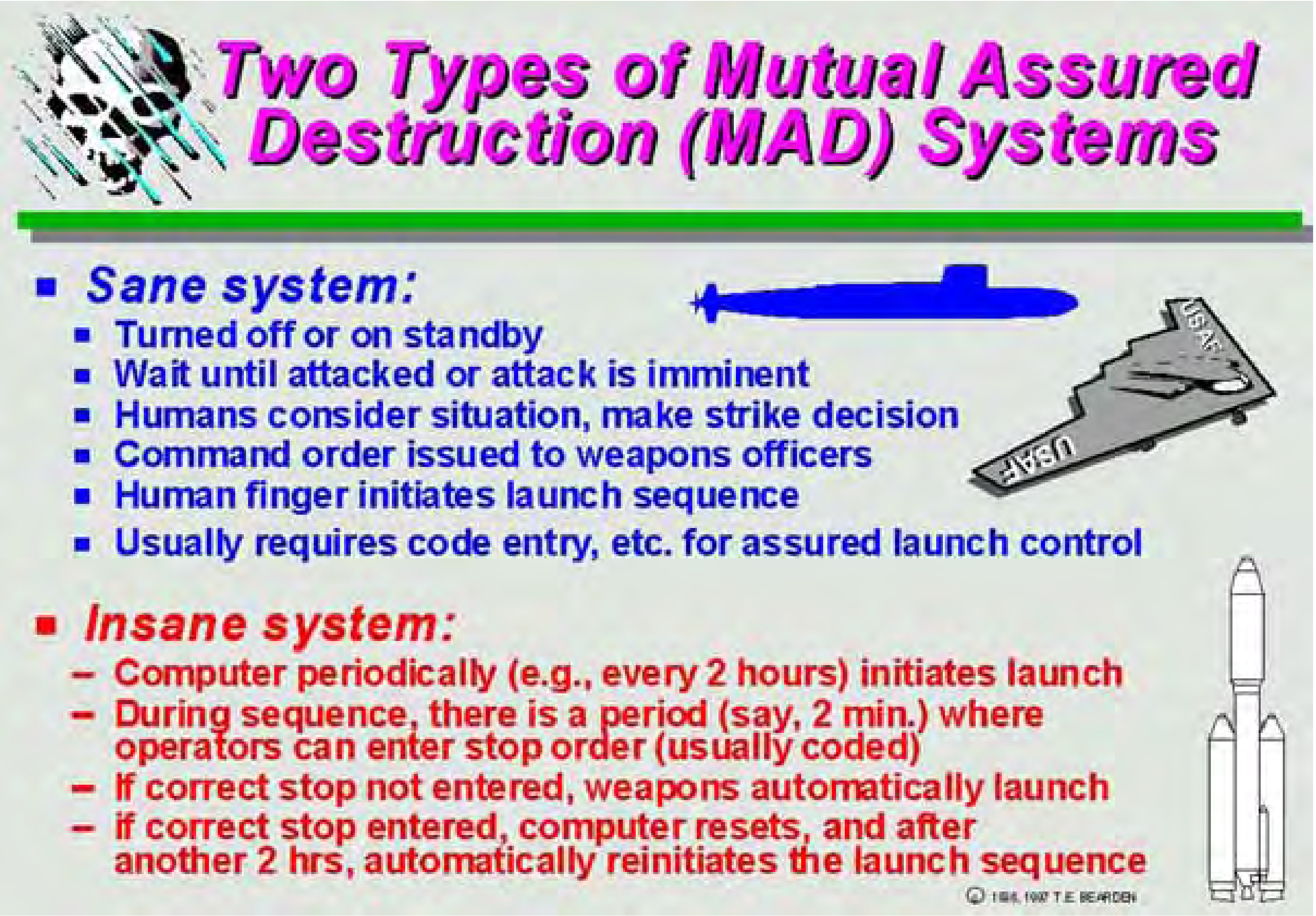

Two Types of Mutual Assured Destruction (MAD) Systems

- **Sane system:**
  - Turned off or on standby
  - Wait until attacked or attack is imminent
  - Humans consider situation, make strike decision
  - Command order issued to weapons officers
  - Human finger initiates launch sequence
  - Usually requires code entry, etc. for assured launch control
- **Insane system:**
  - Computer periodically (e.g., every 2 hours) initiates launch
  - During sequence, there is a period (say, 2 min.) where operators can enter stop order (usually coded)
  - If correct stop not entered, weapons automatically launch
  - If correct stop entered, computer resets, and after another 2 hrs, automatically reinitiates the launch sequence

© 1996, 1997 T.E. BEARDEN

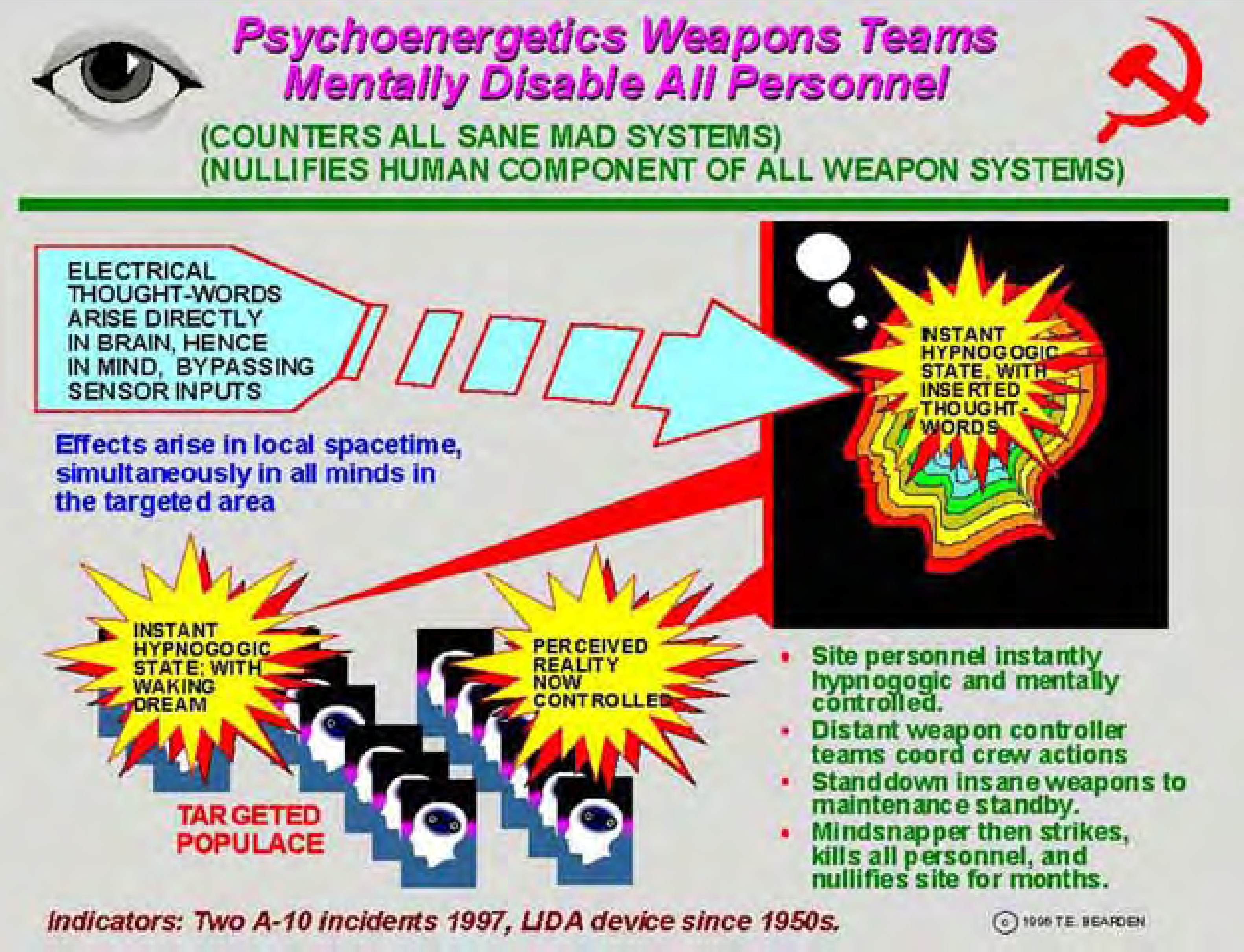

Psychoenergetics Weapons Teams Mentally Disable All Personnel

(COUNTERS ALL SANE MAD SYSTEMS)
(NULLIFIES HUMAN COMPONENT OF ALL WEAPON SYSTEMS)

ELECTRICAL THOUGHT-WORDS ARISE DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

INSTANT HYPNOGOGIC STATE, WITH INSERTED THOUGHT-WORDS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED POPULACE

- Site personnel instantly hypnogogic and mentally controlled.
- Distant weapon controller teams coord crew actions
- Standdown insane weapons to maintenance standby.
- Mindsnapper then strikes, kills all personnel, and nullifies site for months.

Indicators: Two A-10 incidents 1997, LIDA device since 1950s.

© 1996 T.E. BEARDEN

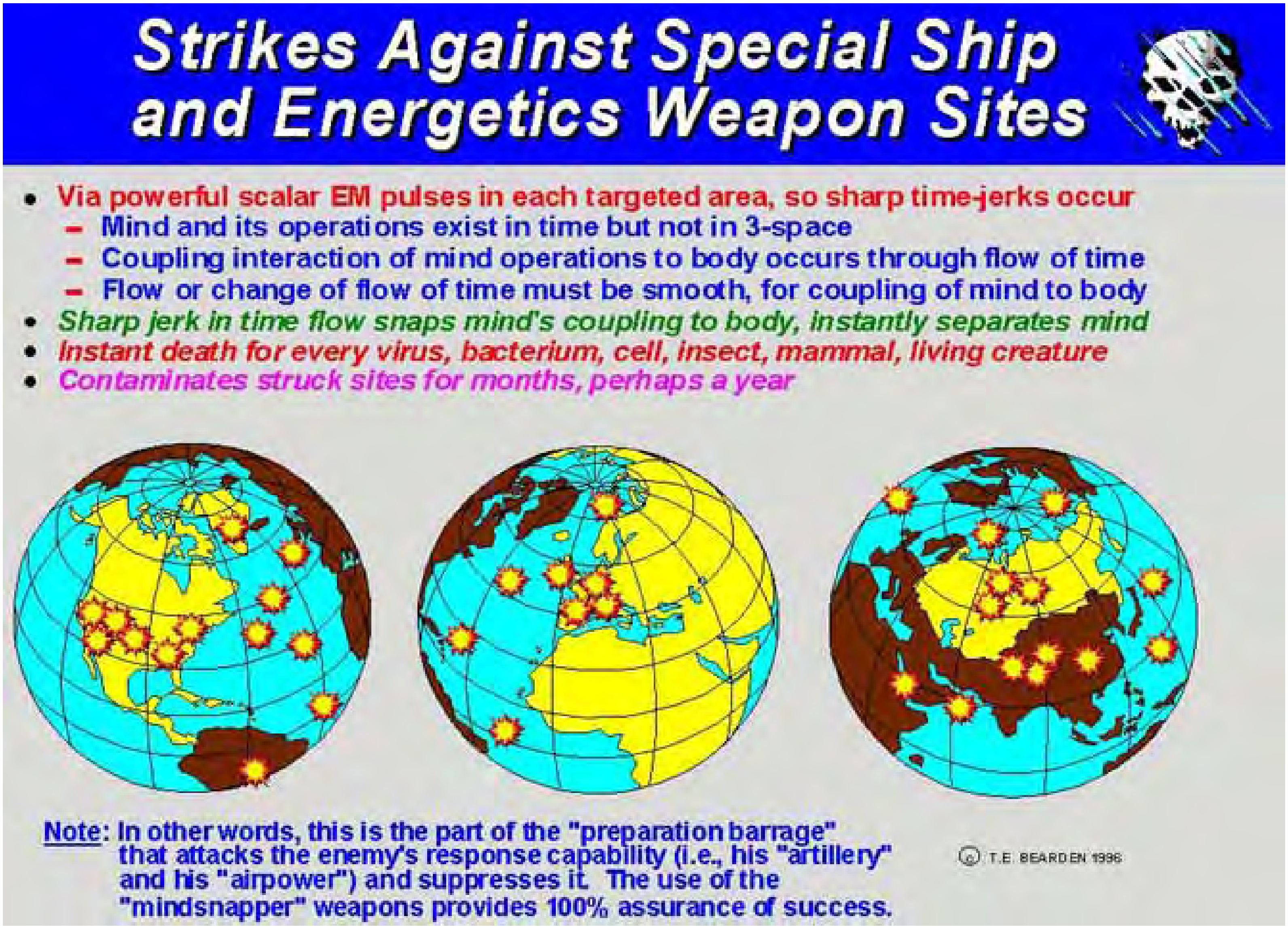

Strikes Against Special Ship and Energetics Weapon Sites

- Via powerful scalar EM pulses in each targeted area, so sharp time-jerks occur
  - Mind and its operations exist in time but not in 3-space
  - Coupling interaction of mind operations to body occurs through flow of time
  - Flow or change of flow of time must be smooth, for coupling of mind to body
- *Sharp jerk in time flow snaps mind's coupling to body, instantly separates mind*
- *Instant death for every virus, bacterium, cell, insect, mammal, living creature*
- *Contaminates struck sites for months, perhaps a year*

Note: In other words, this is the part of the "preparation barrage" that attacks the enemy's response capability (i.e., his "artillery" and his "airpower") and suppresses it. The use of the "mindsnapper" weapons provides 100% assurance of success.

© T.E. BEARDEN 1996

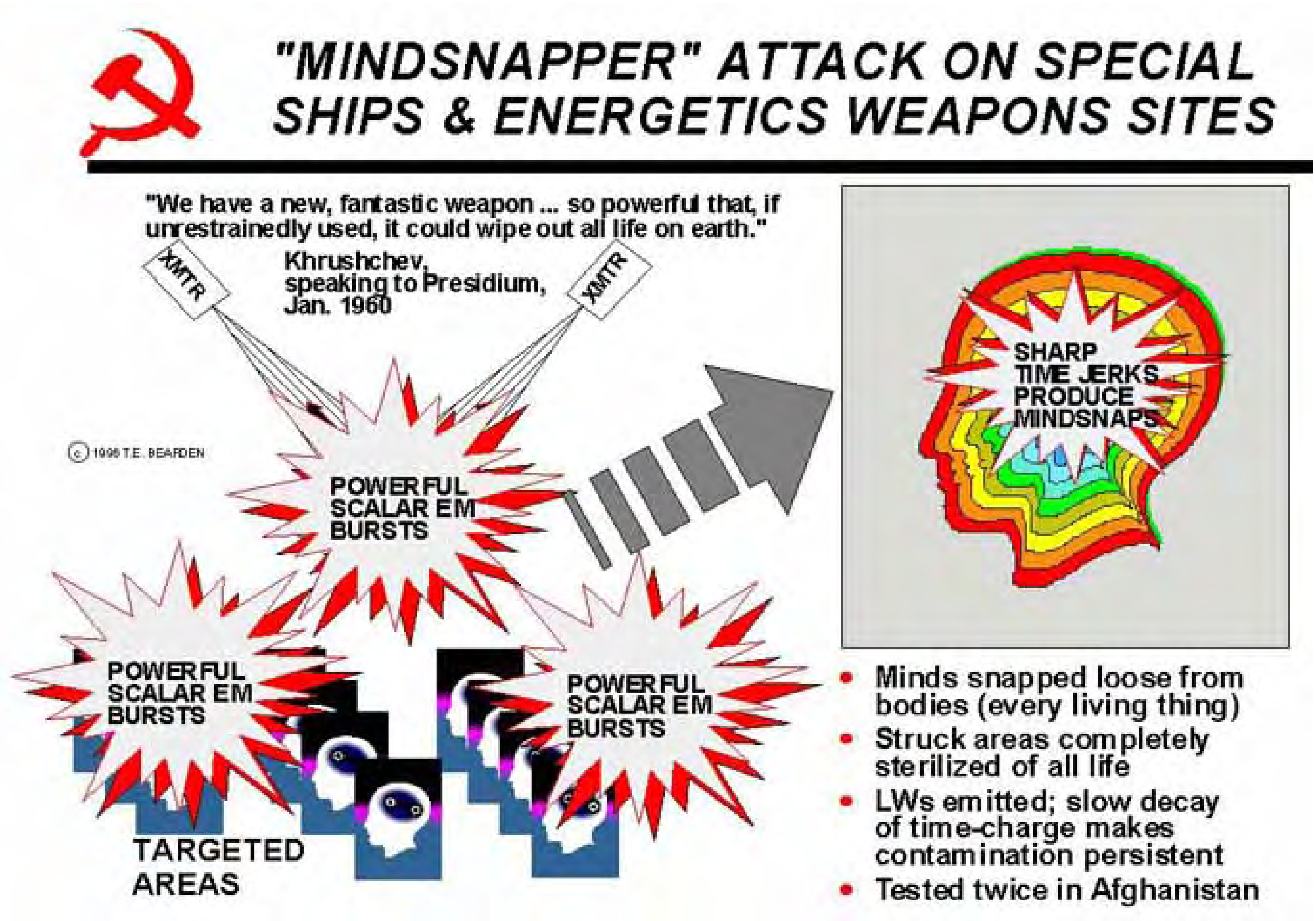

"MINDSNAPPER" ATTACK ON SPECIAL SHIPS & ENERGETICS WEAPONS SITES

"We have a new, fantastic weapon ... so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Khrushchev,
speaking to Presidium,
Jan. 1960

XMTR
XMTR
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
POWERFUL SCALAR EM BURSTS
TARGETED AREAS
SHARP TIME JERKS PRODUCE MINDSNAPS

© 1996 T.E. BEARDEN

- Minds snapped loose from bodies (every living thing)
- Struck areas completely sterilized of all life
- LWs emitted; slow decay of time-charge makes contamination persistent
- Tested twice in Afghanistan

The Tom Bearden Website

Help support the research

How Much Off Guard Are We?

- Slide 1
- Slide 2
- Slide 3
- Slide 4
- Slide 5
- Slide 6
- Slide 7

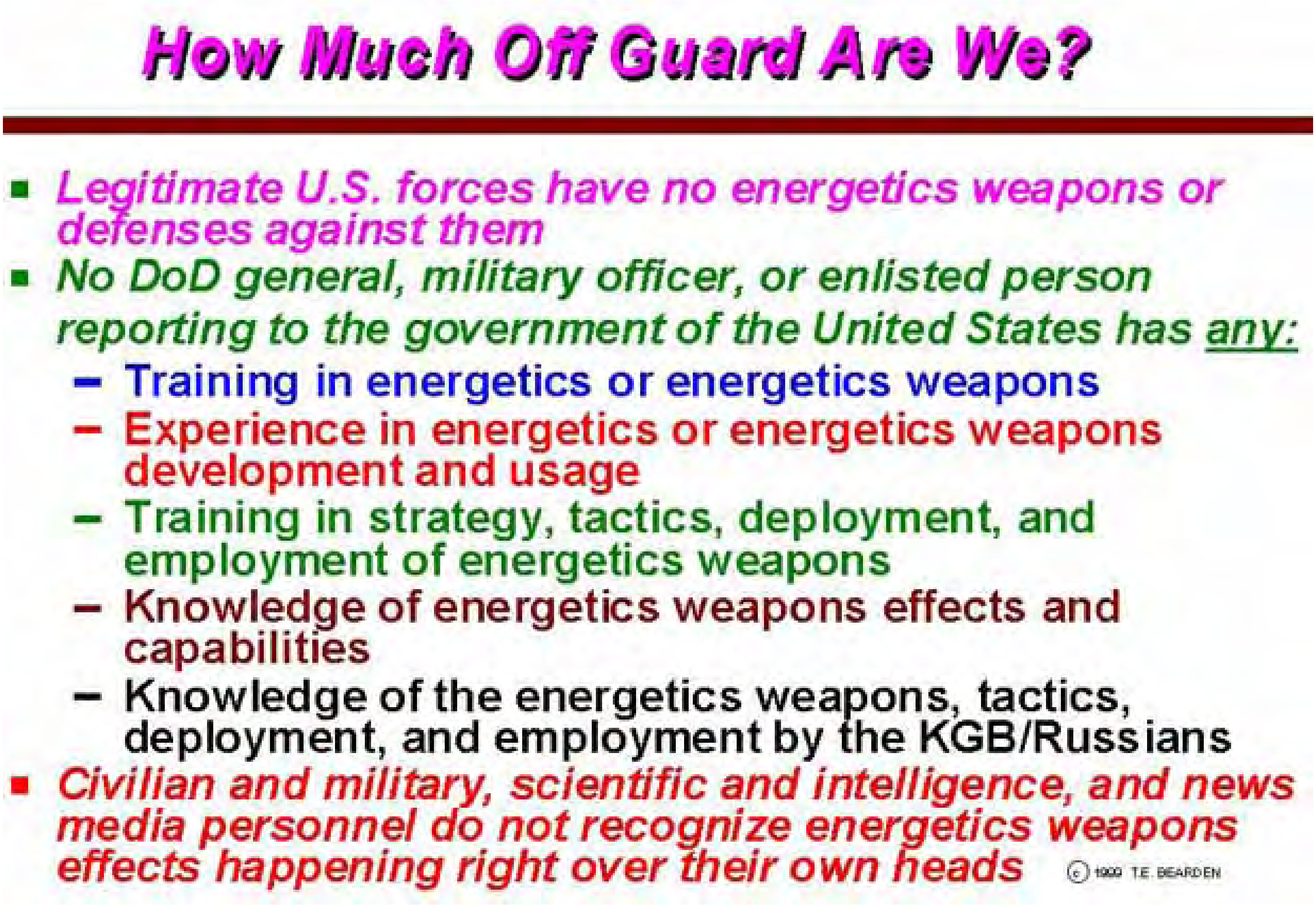

How Much Off Guard Are We?

- *Legitimate U.S. forces have no energetics weapons or defenses against them*
- *No DoD general, military officer, or enlisted person reporting to the government of the United States has any:*
  - Training in energetics or energetics weapons
  - Experience in energetics or energetics weapons development and usage
  - Training in strategy, tactics, deployment, and employment of energetics weapons
  - Knowledge of energetics weapons effects and capabilities
  - Knowledge of the energetics weapons, tactics, deployment, and employment by the KGB/Russians
- *Civilian and military, scientific and intelligence, and news media personnel do not recognize energetics weapons effects happening right over their own heads*

(c) 1999 T.E. BEARDEN

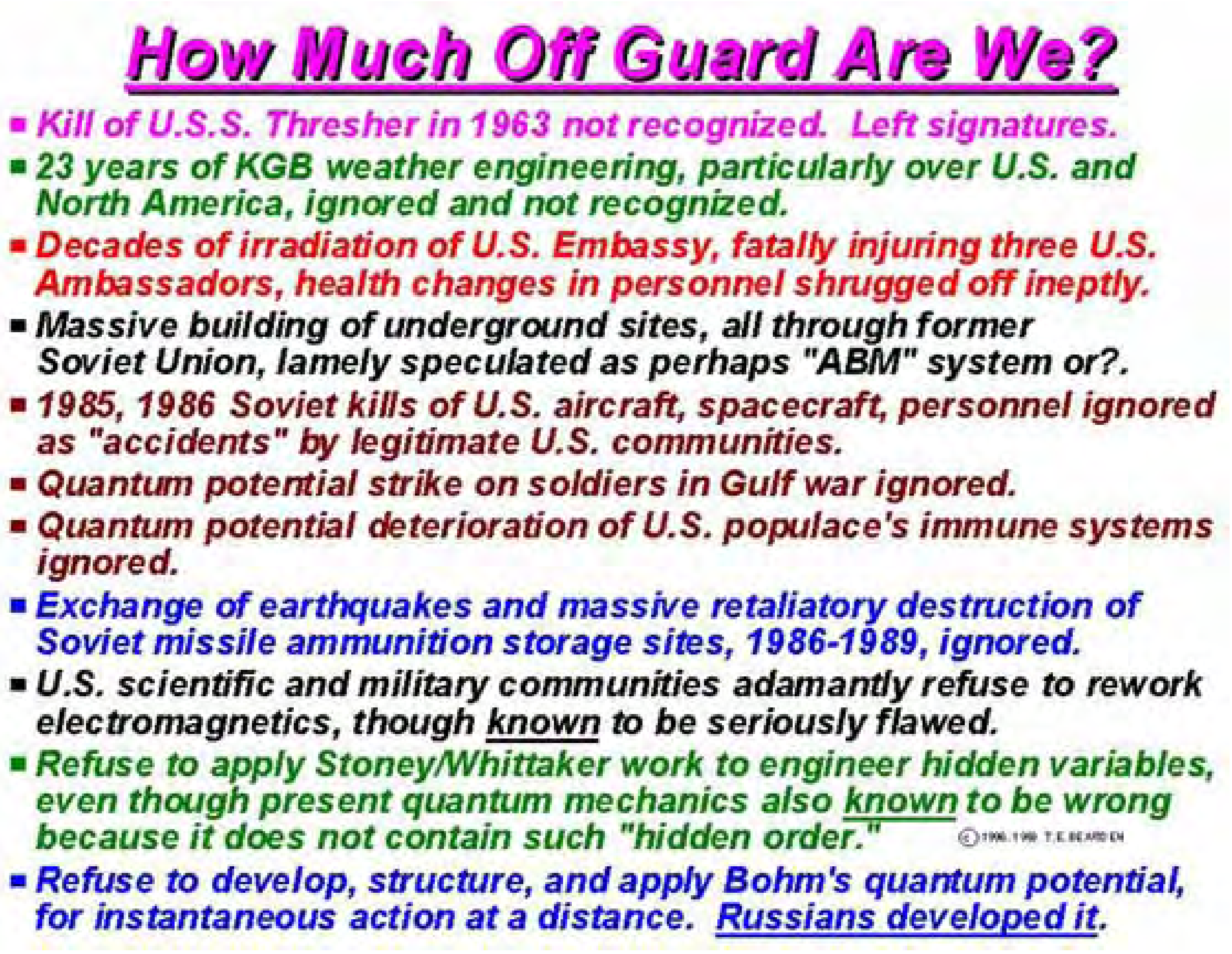

How Much Off Guard Are We?

- Kill of U.S.S. Thresher in 1963 not recognized. Left signatures.
- 23 years of KGB weather engineering, particularly over U.S. and North America, ignored and not recognized.
- Decades of irradiation of U.S. Embassy, fatally injuring three U.S. Ambassadors, health changes in personnel shrugged off ineptly.
- Massive building of underground sites, all through former Soviet Union, lamely speculated as perhaps "ABM" system or?.
- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.
- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, structure, and apply Bohm's quantum potential, for instantaneous action at a distance. Russians developed it.

© 1996, 1999 T.E. BEARDEN

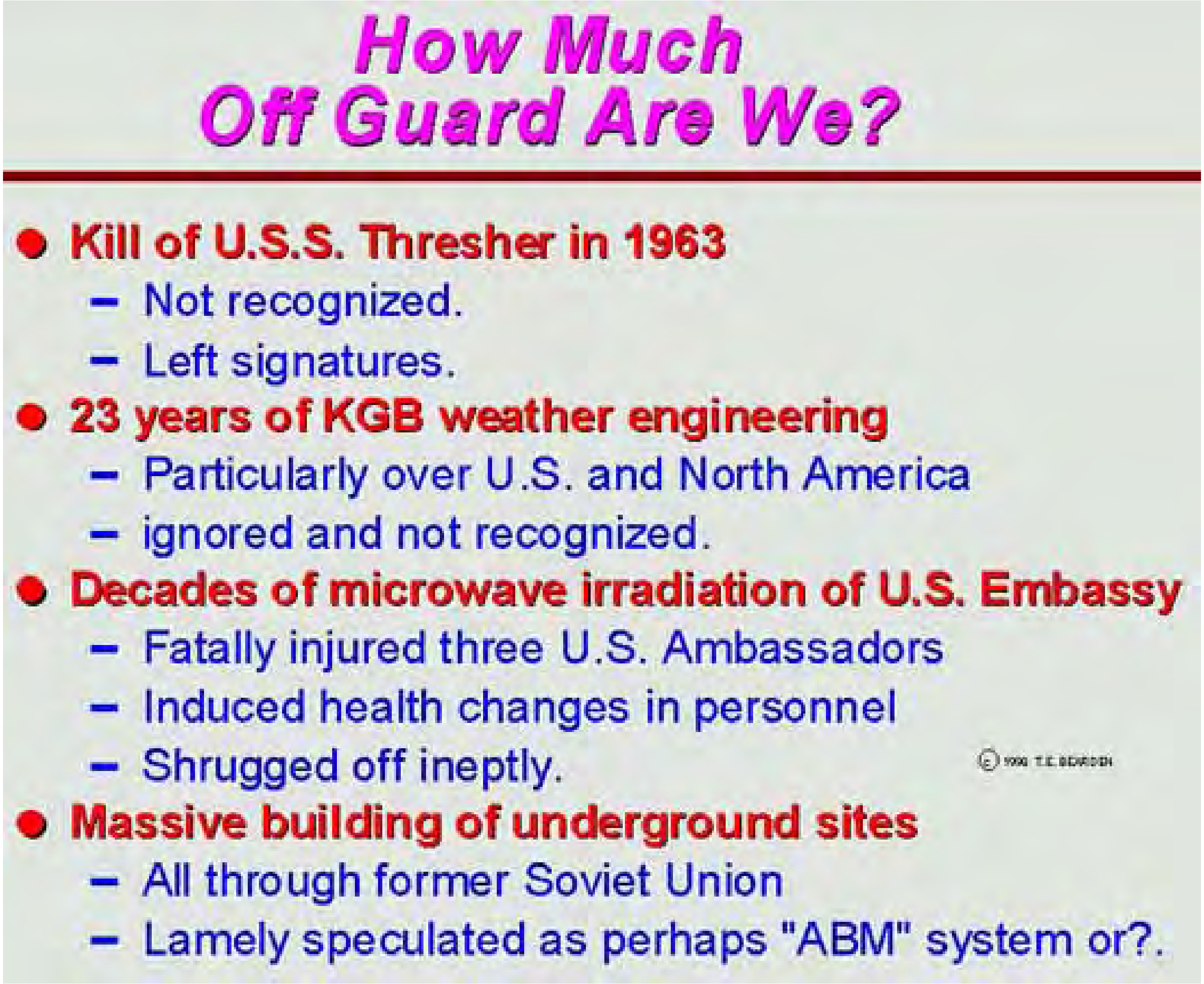

How Much Off Guard Are We?

- **Kill of U.S.S. Thresher in 1963**
  - Not recognized.
  - Left signatures.
- **23 years of KGB weather engineering**
  - Particularly over U.S. and North America
  - ignored and not recognized.
- **Decades of microwave irradiation of U.S. Embassy**
  - Fatally injured three U.S. Ambassadors
  - Induced health changes in personnel
  - Shrugged off ineptly.
- **Massive building of underground sites**
  - All through former Soviet Union
  - Lamely speculated as perhaps "ABM" system or?.

© [illegible] T.E. BEARDEN

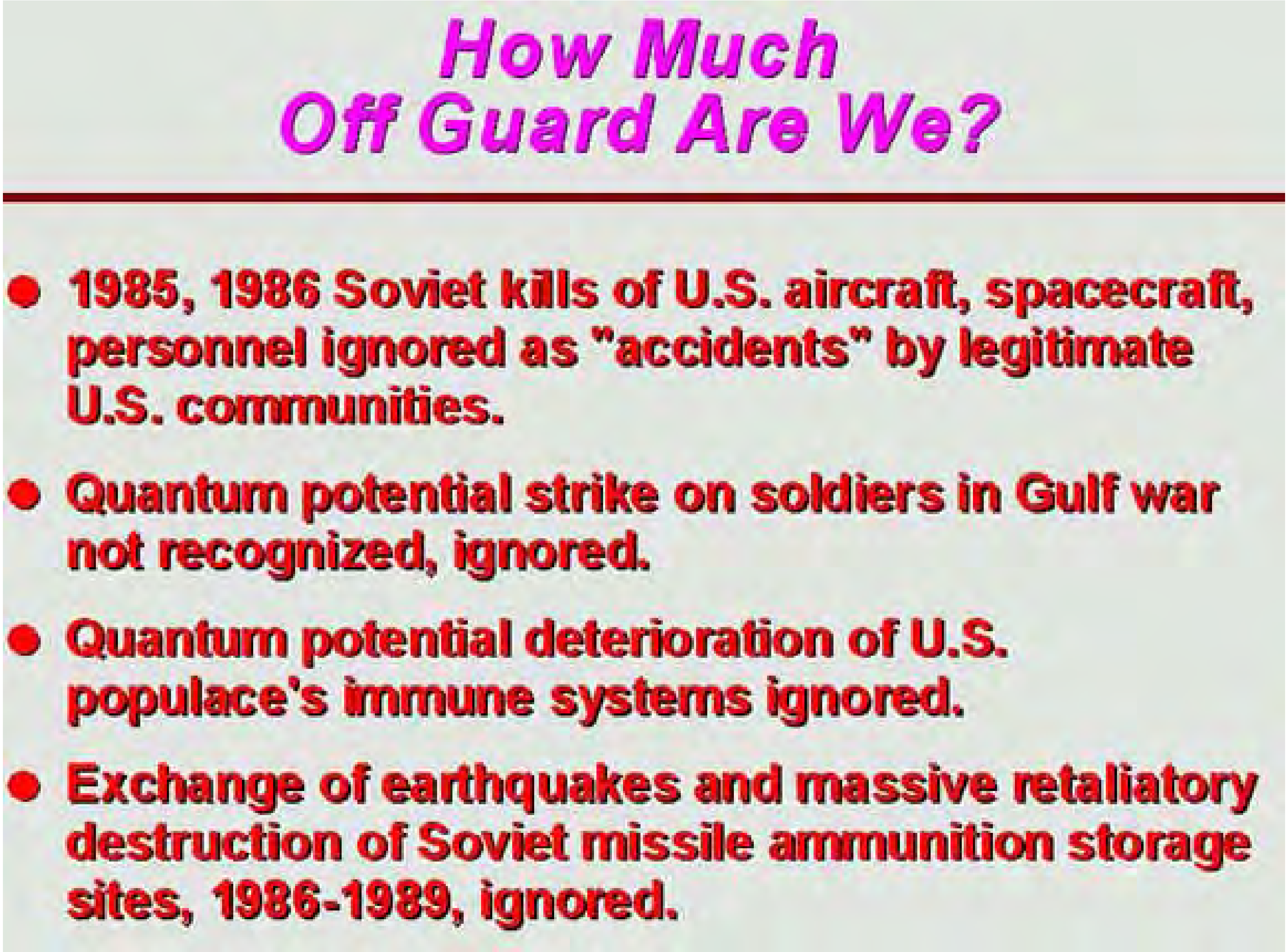

How Much
Off Guard Are We?

- 1985, 1986 Soviet kills of U.S. aircraft, spacecraft, personnel ignored as "accidents" by legitimate U.S. communities.
- Quantum potential strike on soldiers in Gulf war not recognized, ignored.
- Quantum potential deterioration of U.S. populace's immune systems ignored.
- Exchange of earthquakes and massive retaliatory destruction of Soviet missile ammunition storage sites, 1986-1989, ignored.

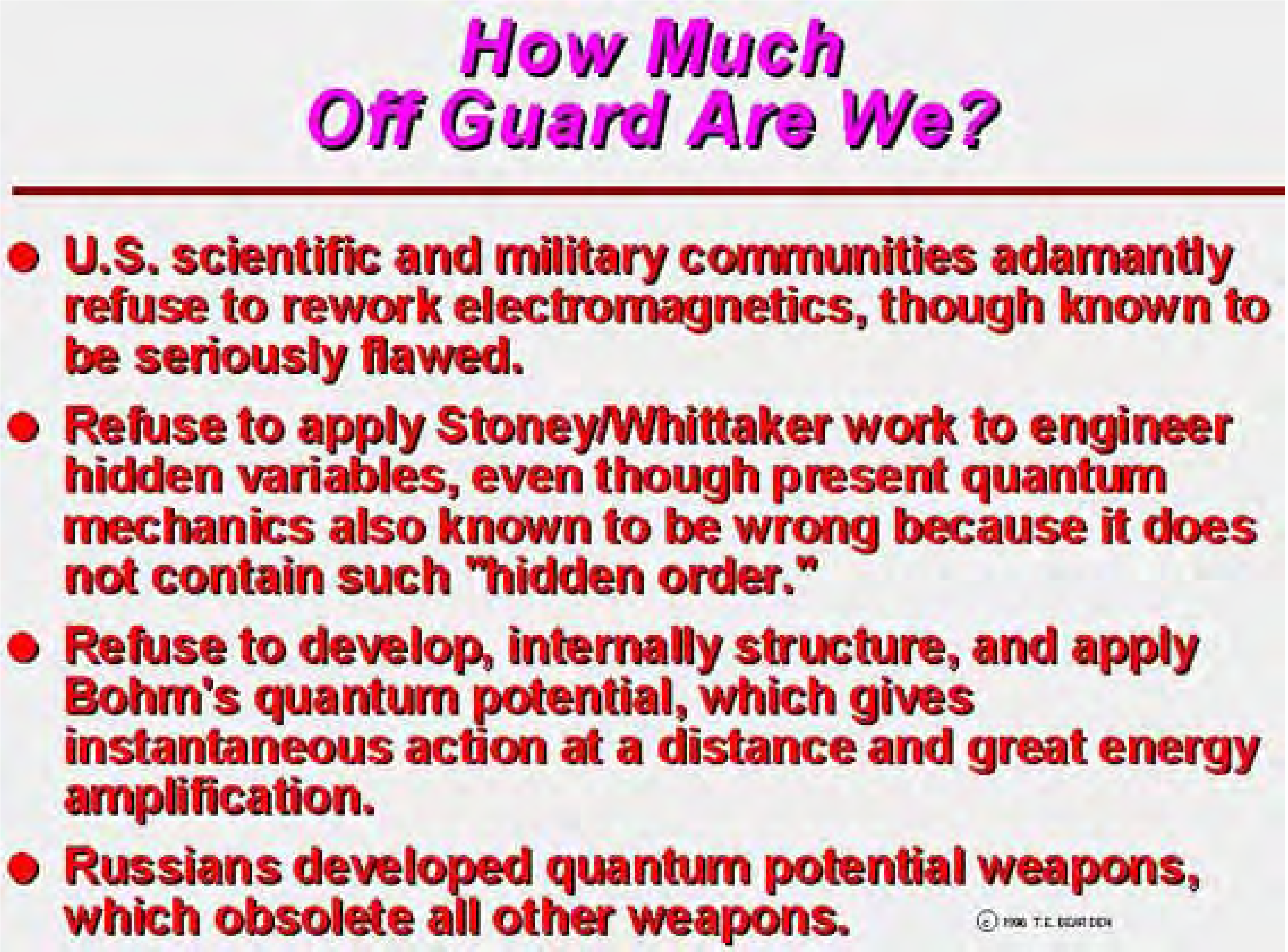

How Much Off Guard Are We?

- U.S. scientific and military communities adamantly refuse to rework electromagnetics, though known to be seriously flawed.
- Refuse to apply Stoney/Whittaker work to engineer hidden variables, even though present quantum mechanics also known to be wrong because it does not contain such "hidden order."
- Refuse to develop, internally structure, and apply Bohm's quantum potential, which gives instantaneous action at a distance and great energy amplification.
- Russians developed quantum potential weapons, which obsolete all other weapons.

© 1996 T.E. BEARDEN

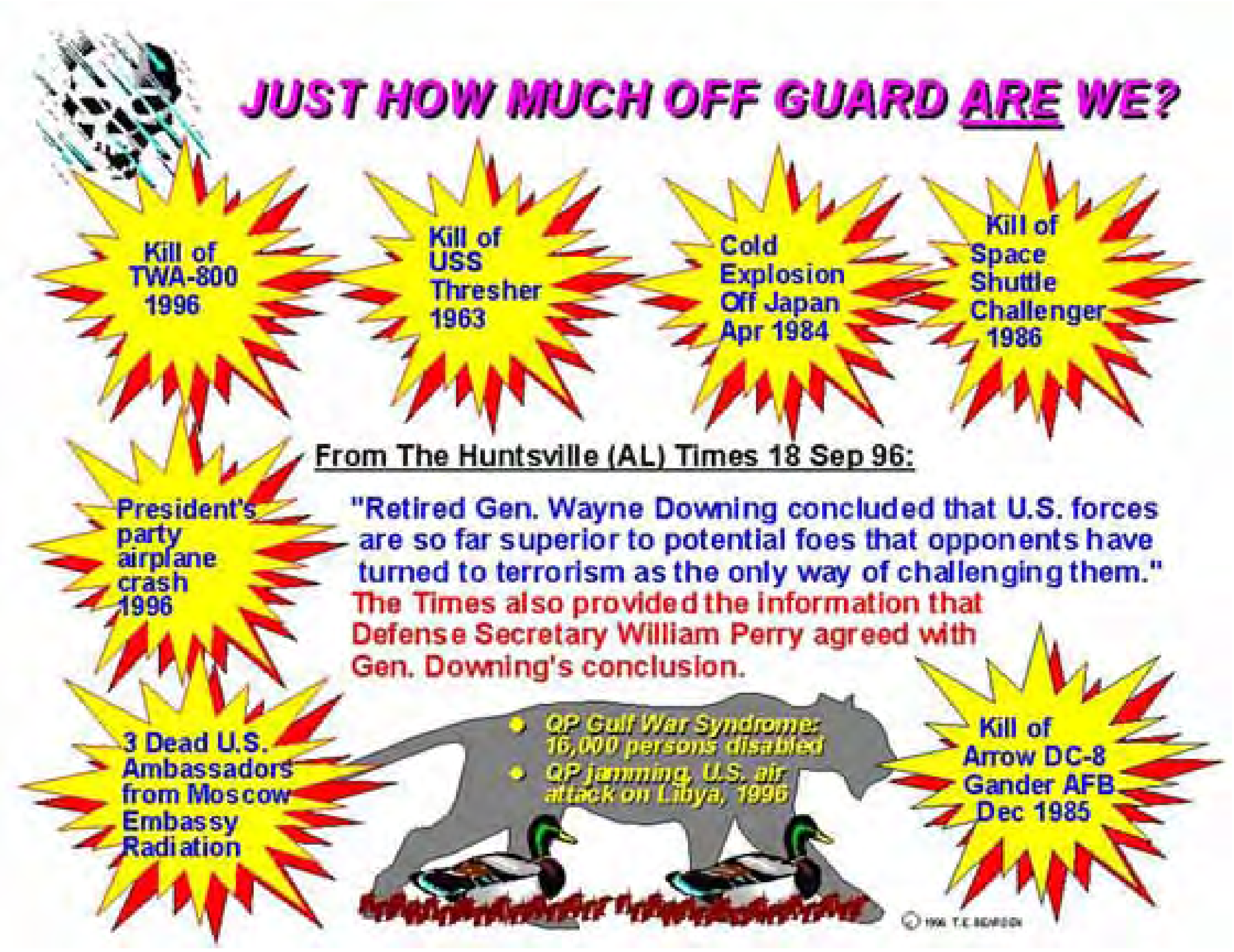

JUST HOW MUCH OFF GUARD ARE WE?

Kill of
TWA-800
1996

Kill of
USS
Thresher
1963

Cold
Explosion
Off Japan
Apr 1984

Kill of
Space
Shuttle
Challenger
1986

President's
party
airplane
crash
1996

From The Huntsville (AL) Times 18 Sep 96:

"Retired Gen. Wayne Downing concluded that U.S. forces are so far superior to potential foes that opponents have turned to terrorism as the only way of challenging them."

The Times also provided the information that Defense Secretary William Perry agreed with Gen. Downing's conclusion.

3 Dead U.S.
Ambassadors
from Moscow
Embassy
Radiation

- QP Gulf War Syndrome: 16,000 persons disabled
- QP jamming, U.S. air attack on Libya, 1996

Kill of
Arrow DC-8
Gander AFB
Dec 1985

© 1996 T.E. BEARDEN

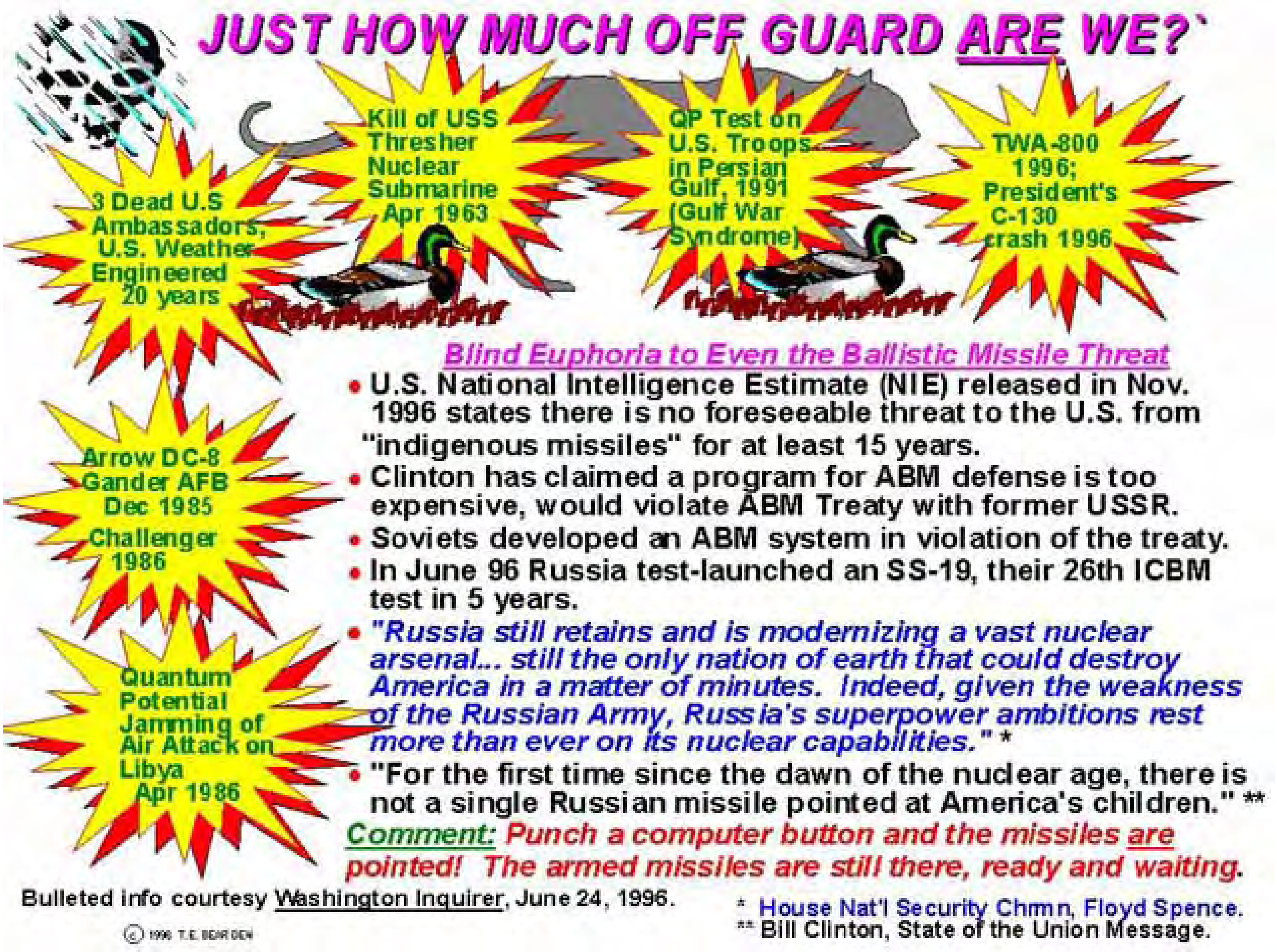

JUST HOW MUCH OFF GUARD ARE WE?

Kill of USS Thresher Nuclear Submarine Apr 1963

QP Test on U.S. Troops in Persian Gulf, 1991 (Gulf War Syndrome)

TWA-800 1996; President's C-130 crash 1996

3 Dead U.S Ambassadors, U.S. Weather Engineered 20 years

Arrow DC-8 Gander AFB Dec 1985

Challenger 1986

Quantum Potential Jamming of Air Attack on Libya Apr 1986

Blind Euphoria to Even the Ballistic Missile Threat

- U.S. National Intelligence Estimate (NIE) released in Nov. 1996 states there is no foreseeable threat to the U.S. from "indigenous missiles" for at least 15 years.
- Clinton has claimed a program for ABM defense is too expensive, would violate ABM Treaty with former USSR.
- Soviets developed an ABM system in violation of the treaty.
- In June 96 Russia test-launched an SS-19, their 26th ICBM test in 5 years.
- *"Russia still retains and is modernizing a vast nuclear arsenal... still the only nation of earth that could destroy America in a matter of minutes. Indeed, given the weakness of the Russian Army, Russia's superpower ambitions rest more than ever on its nuclear capabilities."* *
- "For the first time since the dawn of the nuclear age, there is not a single Russian missile pointed at America's children." **

Comment: Punch a computer button and the missiles are pointed! The armed missiles are still there, ready and waiting.

Bulleted info courtesy Washington Inquirer, June 24, 1996.

* House Nat'l Security Chmn, Floyd Spence.
** Bill Clinton, State of the Union Message.

© 1996 T.E. BEARDEN

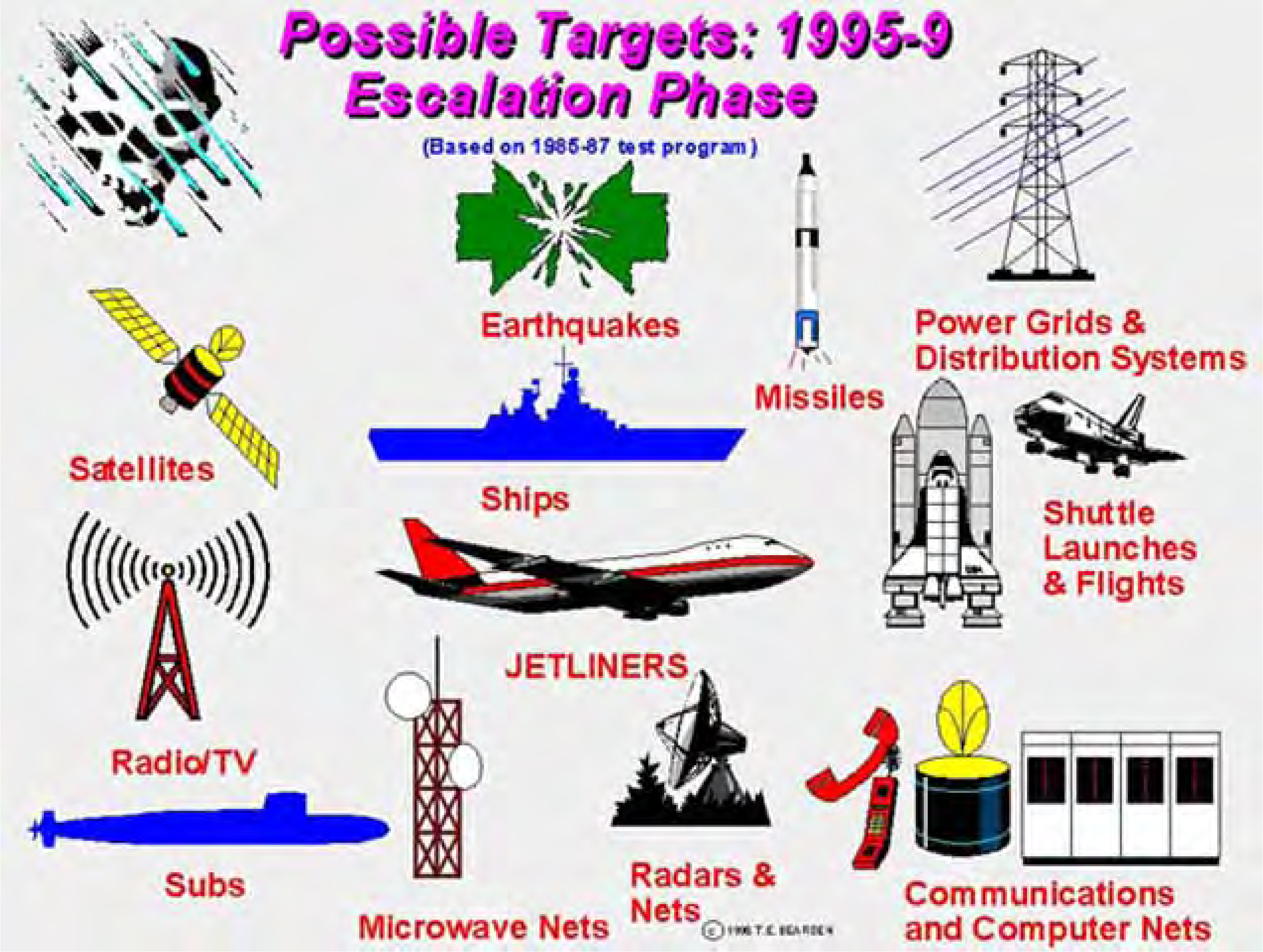
Possible Targets: 1995-9
Escalation Phase
(Based on 1985-87 test program)
Earthquakes
Missiles
Power Grids &
Distribution Systems
Satellites
Ships
Shuttle
Launches
& Flights
JETLINERS
Radio/TV
Subs
Microwave Nets
Radars &
Nets
Communications
and Computer Nets

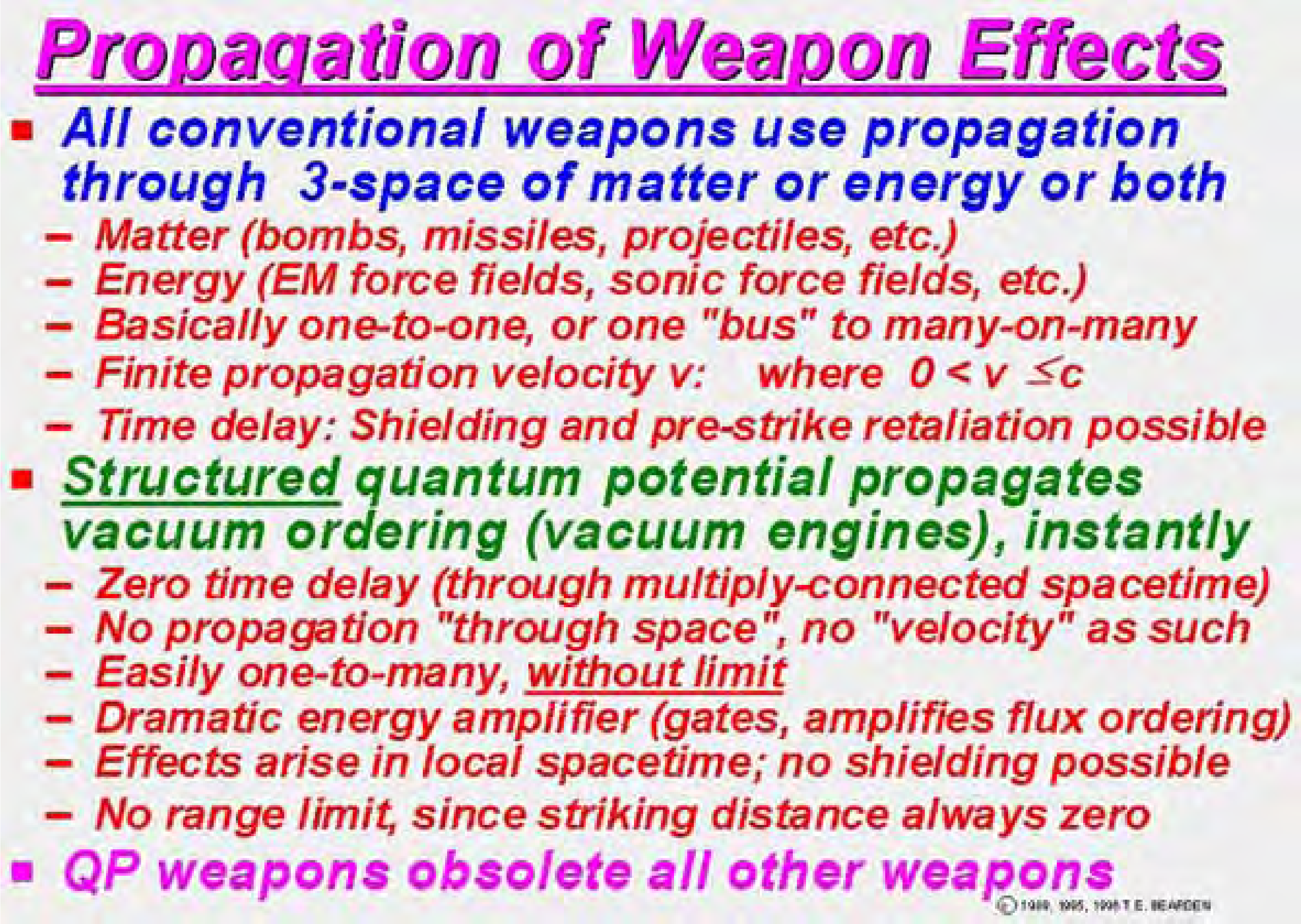

Propagation of Weapon Effects

- **All conventional weapons use propagation through 3-space of matter or energy or both**
  - *Matter (bombs, missiles, projectiles, etc.)*
  - *Energy (EM force fields, sonic force fields, etc.)*
  - *Basically one-to-one, or one "bus" to many-on-many*
  - *Finite propagation velocity v: where $0 < v \leq c$*
  - *Time delay: Shielding and pre-strike retaliation possible*
- **Structured quantum potential propagates vacuum ordering (vacuum engines), instantly**
  - *Zero time delay (through multiply-connected spacetime)*
  - *No propagation "through space", no "velocity" as such*
  - *Easily one-to-many, without limit*
  - *Dramatic energy amplifier (gates, amplifies flux ordering)*
  - *Effects arise in local spacetime; no shielding possible*
  - *No range limit, since striking distance always zero*
- **QP weapons obsolete all other weapons**

© 1989, 1995, 1998 T.E. BEARDEN

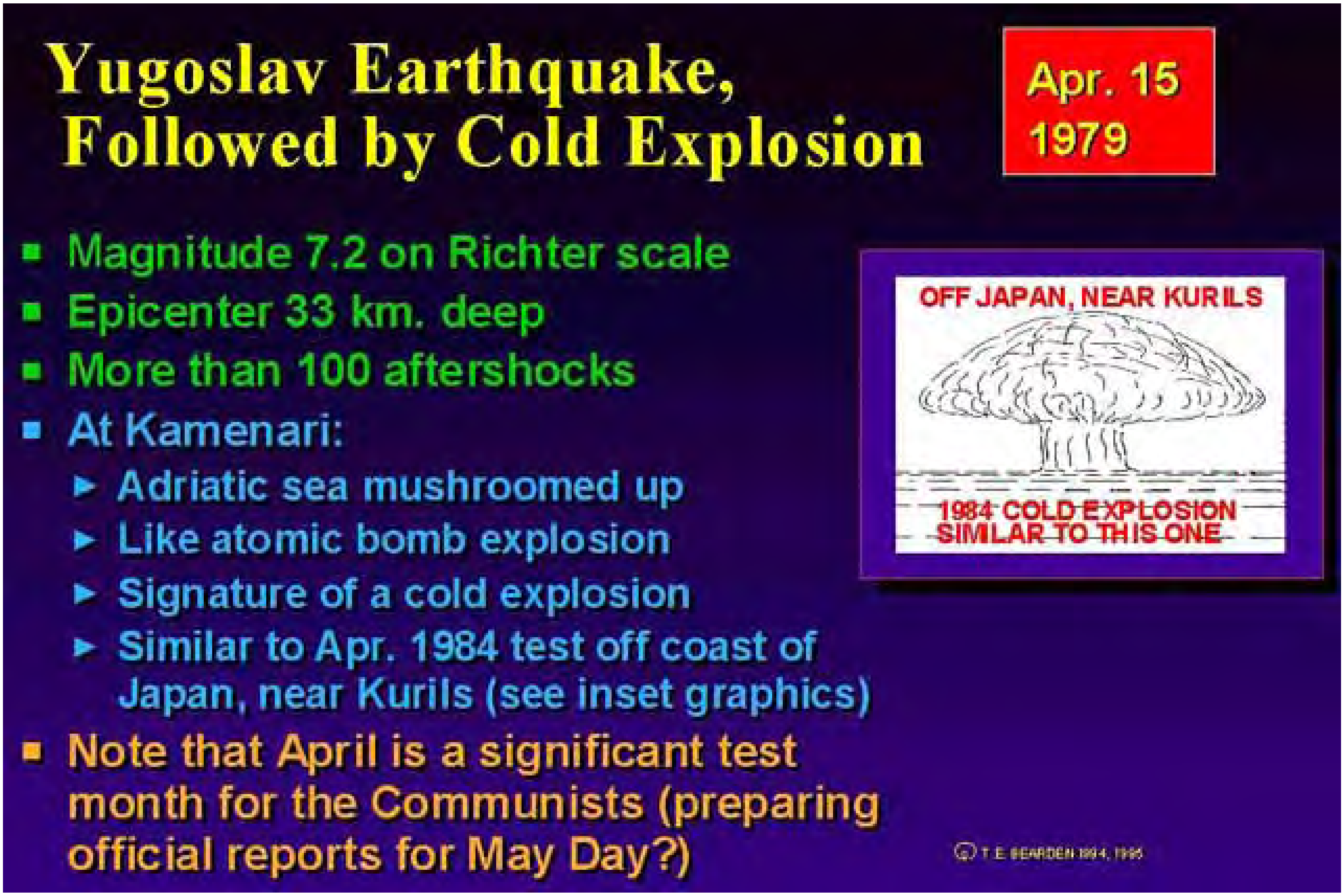

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- Note that April is a significant test month for the Communists (preparing official reports for May Day?)

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

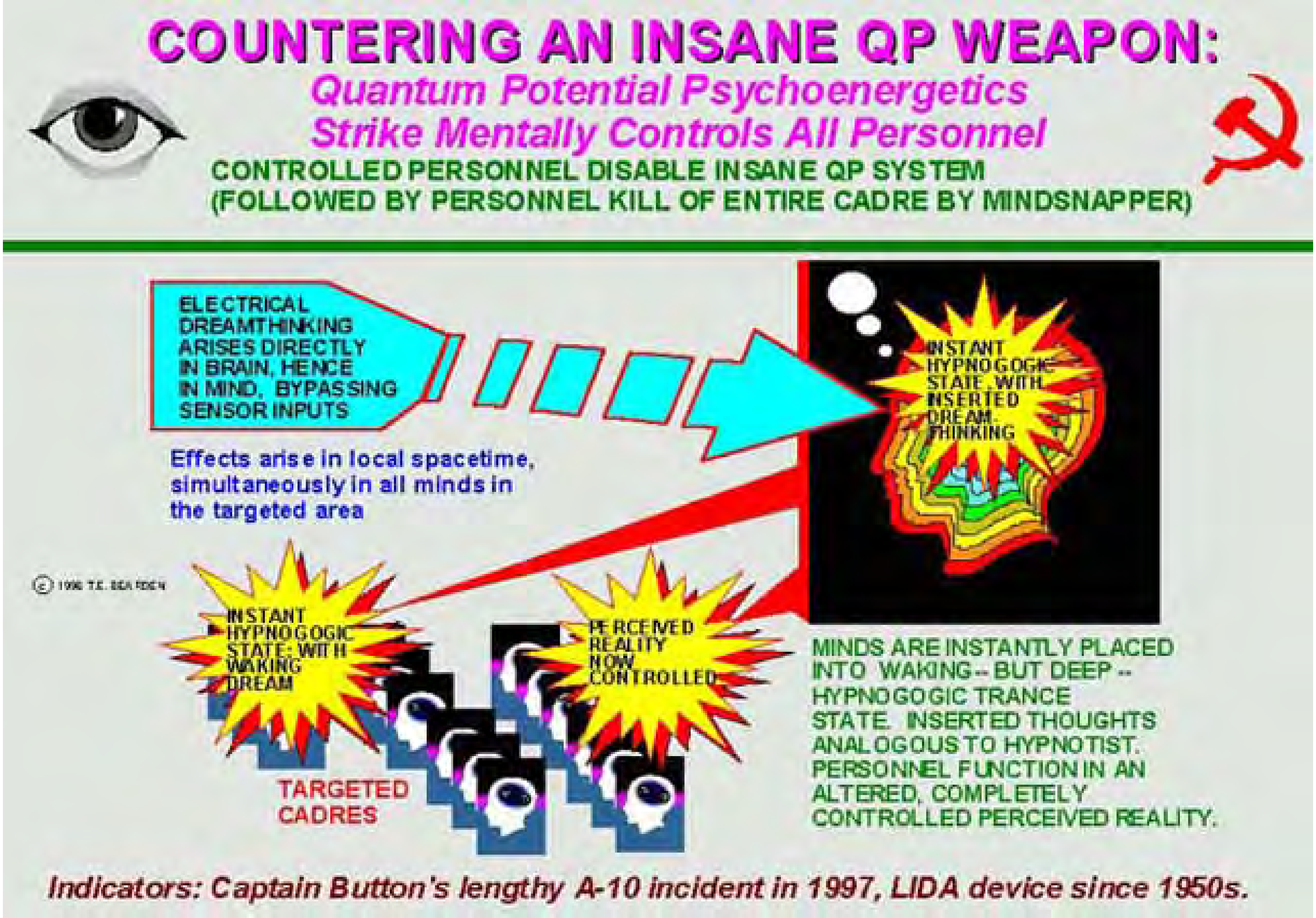

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM-THINKING

© 1998 T.E. BEARDEN

INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM

PERCEIVED REALITY NOW CONTROLLED

TARGETED CADRES

MINDS ARE INSTANTLY PLACED INTO WAKING-- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

Gulf War Disease: KGB QP Weapon Induced ST Curvature Engines for a Cocktail of Diseases

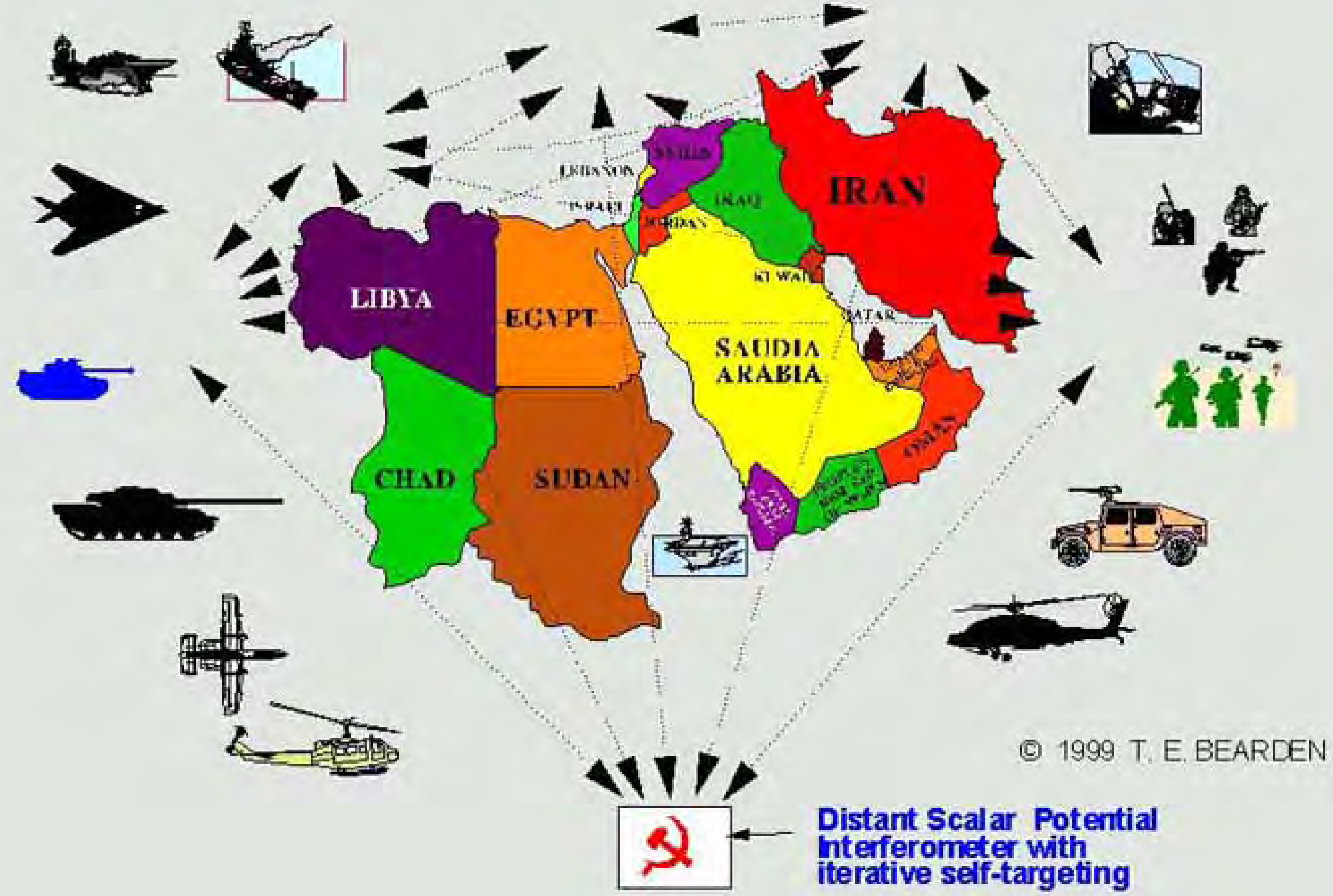
LEBANON
SYRIA
ISRAEL
JORDAN
IRAQ
IRAN
KUWAIT
QATAR
LIBYA
EGYPT
SAUDIA ARABIA
OMAN
CHAD
SUDAN
© 1999 T. E. BEARDEN

Distant Scalar Potential Interferometer with iterative self-targeting

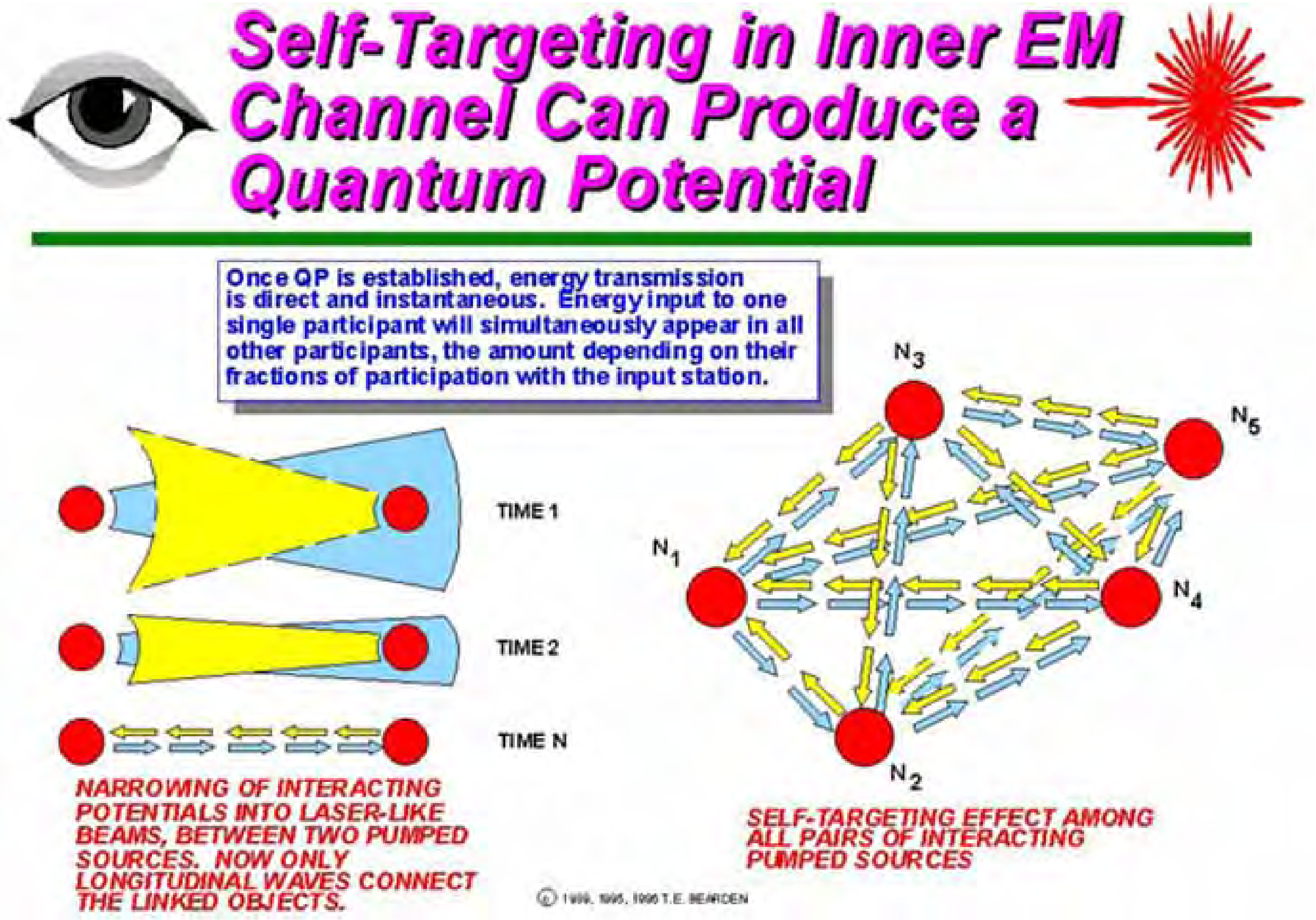

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

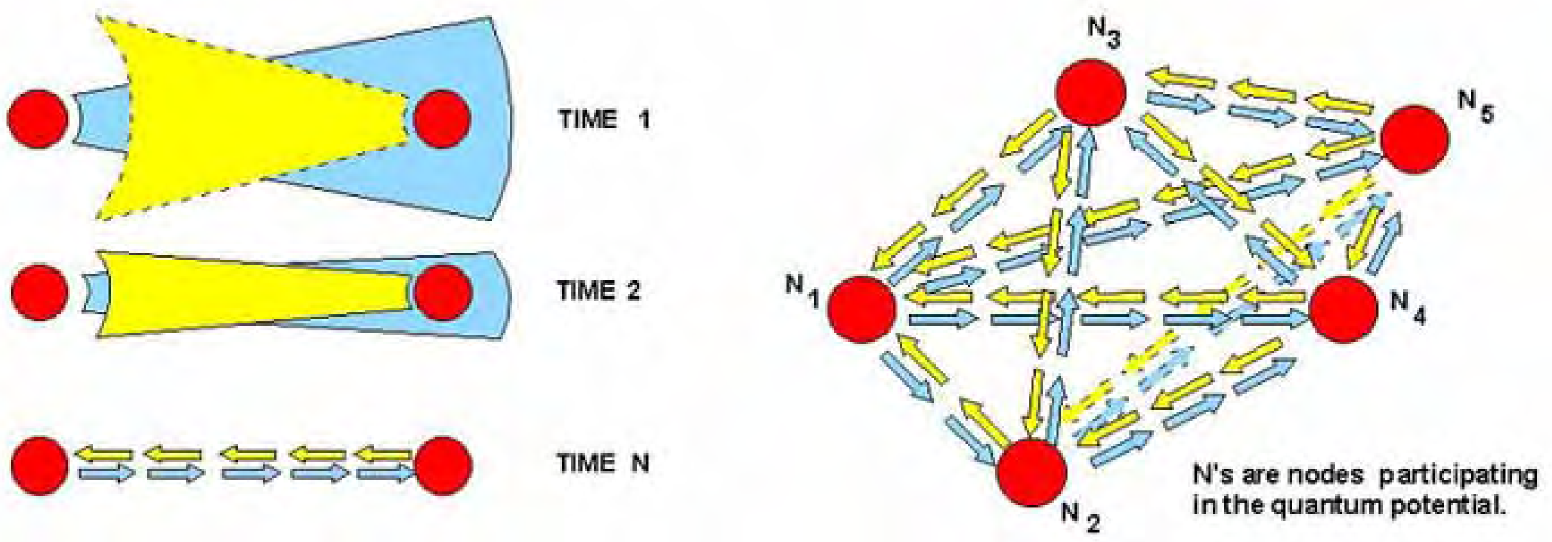

N's are nodes participating in the quantum potential.

a. Iterative retroreflection between any pair of nodes narrows the interconnecting potentials to a laser-like, bidirectional beam. Iterative phase conjugate reflection provides self-targeting.

b. The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

© 1989, 1995 T.E. BEARDEN

Consider the hidden Stoney/Whittaker bidirectional EM waves that comprise the potential from each source node. If these nodes have sufficient intersymmetry of dynamic material form, then self-targeting occurs between the two nodes of any nodal pair. By iterative phase conjugate reflection, the scattering is narrowed into interconnecting wavepairs, in laser-like beams, between each two nodes. Thus the interconnecting potentials narrow into laser-like, bidirectional beams. The Coulomb gauge may apply, and a quantum potential may be formed.

Figure 30. Mechanism for generating a quantum potential, by iterative self-targeting induced into a system of multiple scattering nodes. Once a QP is established, spatial separation has little or no effect.

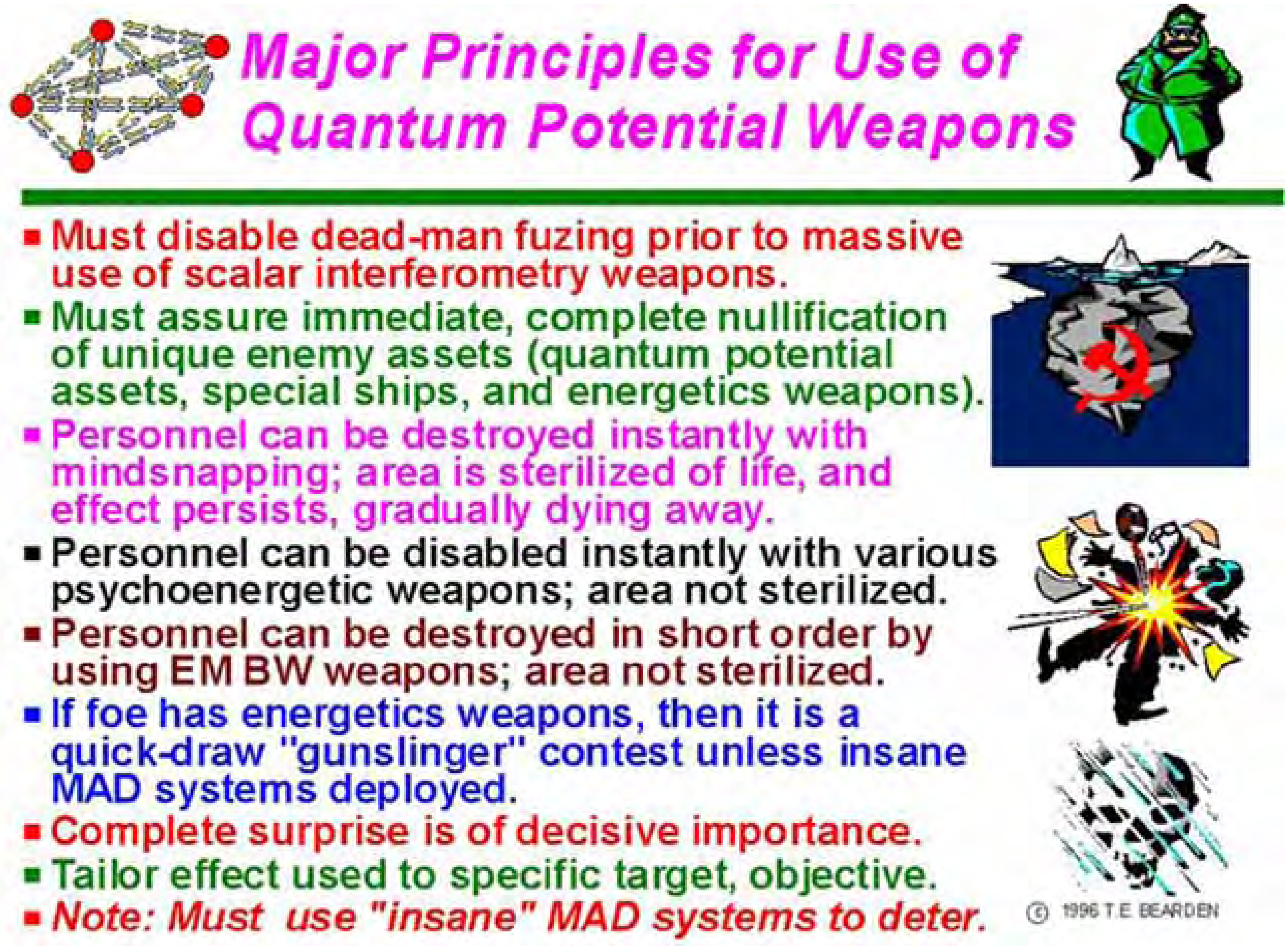

Major Principles for Use of Quantum Potential Weapons

- Must disable dead-man fuzing prior to massive use of scalar interferometry weapons.
- Must assure immediate, complete nullification of unique enemy assets (quantum potential assets, special ships, and energetics weapons).
- Personnel can be destroyed instantly with mindsnapping; area is sterilized of life, and effect persists, gradually dying away.
- Personnel can be disabled instantly with various psychoenergetic weapons; area not sterilized.
- Personnel can be destroyed in short order by using EM BW weapons; area not sterilized.
- If foe has energetics weapons, then it is a quick-draw "gunslinger" contest unless insane MAD systems deployed.
- Complete surprise is of decisive importance.
- Tailor effect used to specific target, objective.
- *Note: Must use "insane" MAD systems to deter.*

© 1996 T.E. BEARDEN

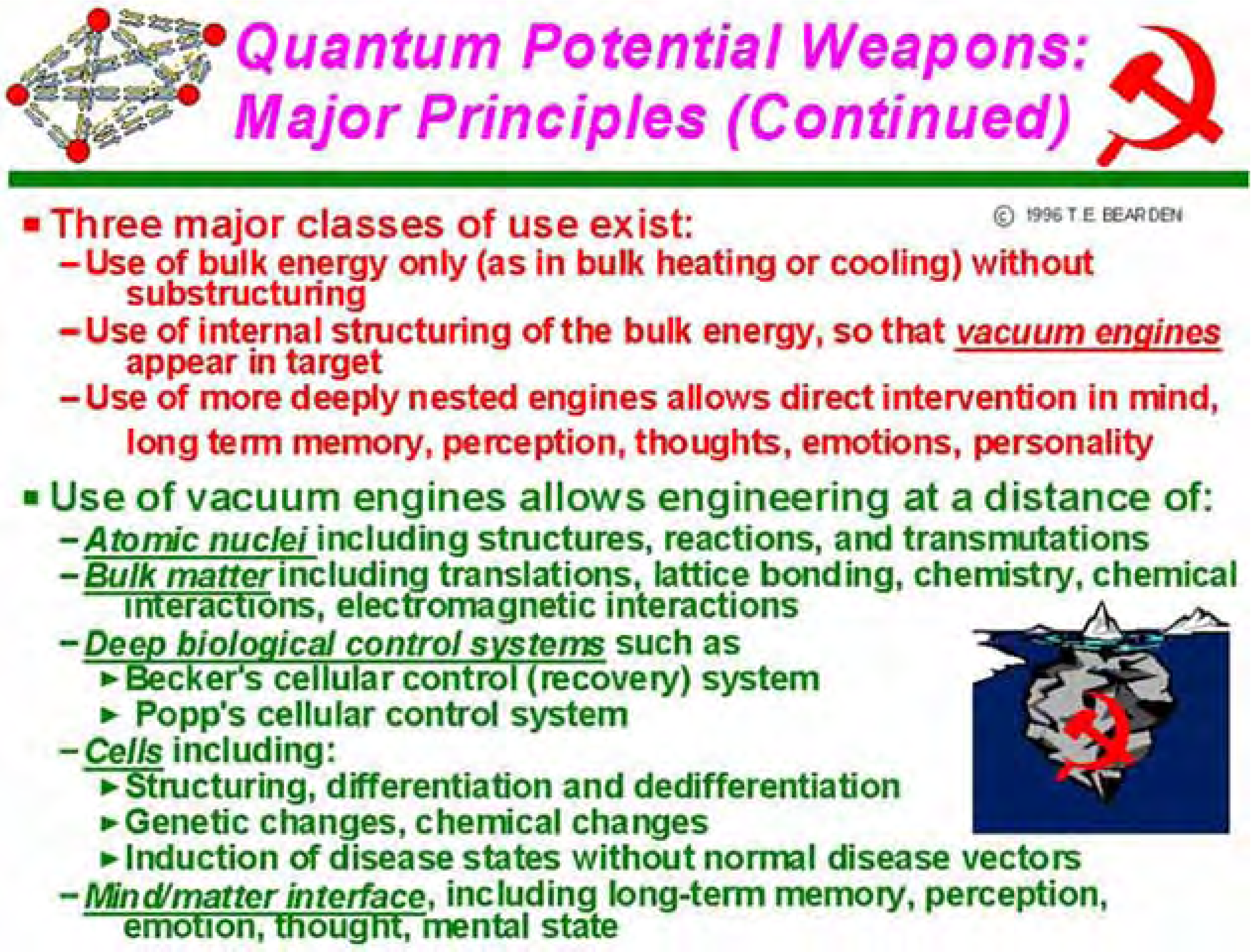

Quantum Potential Weapons: Major Principles (Continued)

© 1996 T. E. BEARDEN

- Three major classes of use exist:
  - Use of bulk energy only (as in bulk heating or cooling) without substructuring
  - Use of internal structuring of the bulk energy, so that *vacuum engines* appear in target
  - Use of more deeply nested engines allows direct intervention in mind, long term memory, perception, thoughts, emotions, personality
- Use of vacuum engines allows engineering at a distance of:
  - *Atomic nuclei* including structures, reactions, and transmutations
  - *Bulk matter* including translations, lattice bonding, chemistry, chemical interactions, electromagnetic interactions
  - *Deep biological control systems* such as
    - ► Becker's cellular control (recovery) system
    - ► Popp's cellular control system
  - *Cells* including:
    - ► Structuring, differentiation and dedifferentiation
    - ► Genetic changes, chemical changes
    - ► Induction of disease states without normal disease vectors
  - *Mind/matter interface*, including long-term memory, perception, emotion, thought, mental state

Instantaneous Communication by a Quantum Potential

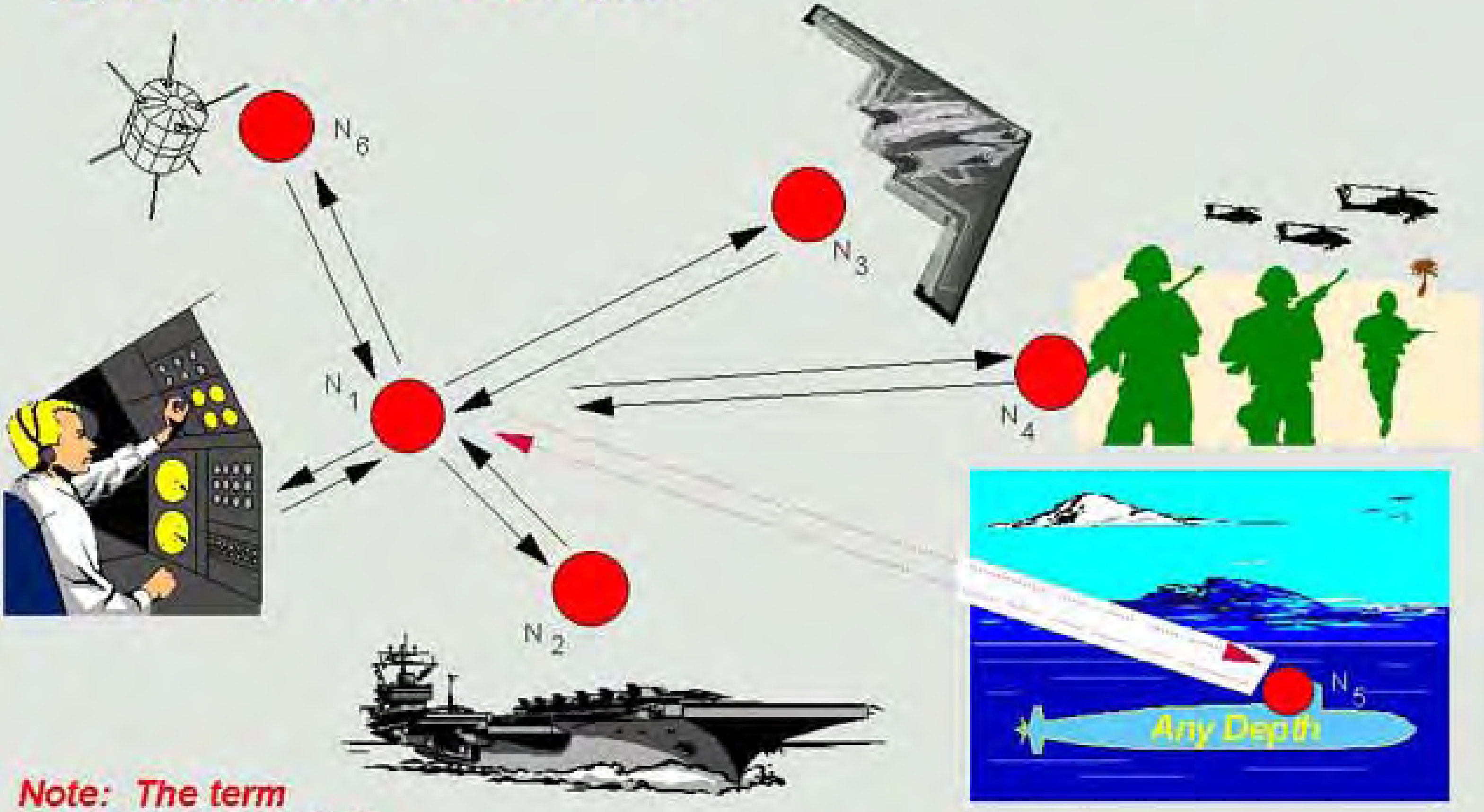

Note: The term "quantum potential" is used in an extended sense.

© 1989, 1995 T.E. BEARDEN

Selected clipart © by Lotus SmartPics ™

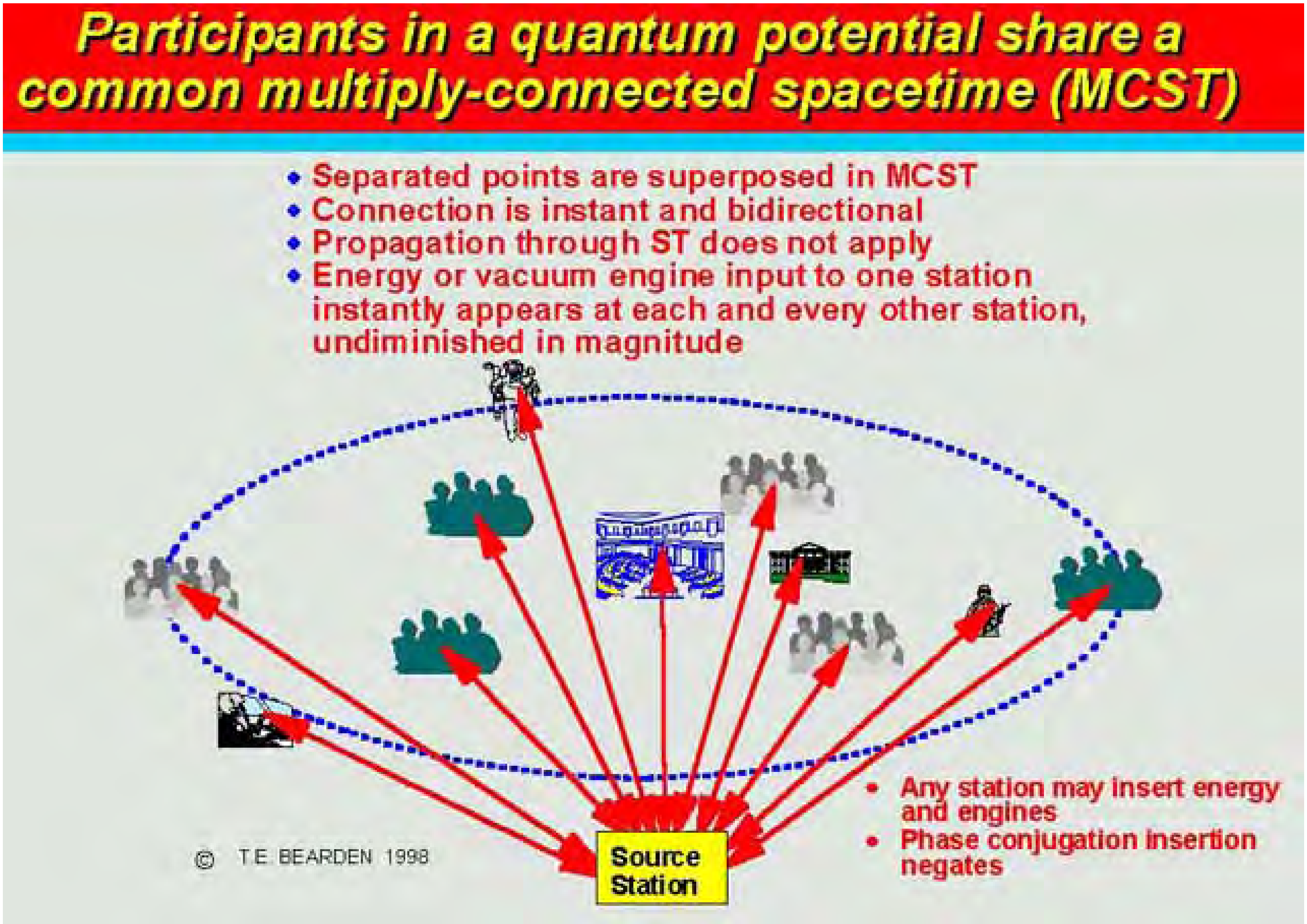

Participants in a quantum potential share a common multiply-connected spacetime (MCST)

- Separated points are superposed in MCST
- Connection is instant and bidirectional
- Propagation through ST does not apply
- Energy or vacuum engine input to one station instantly appears at each and every other station, undiminished in magnitude

© T.E. BEARDEN 1998

- Any station may insert energy and engines
- Phase conjugation insertion negates

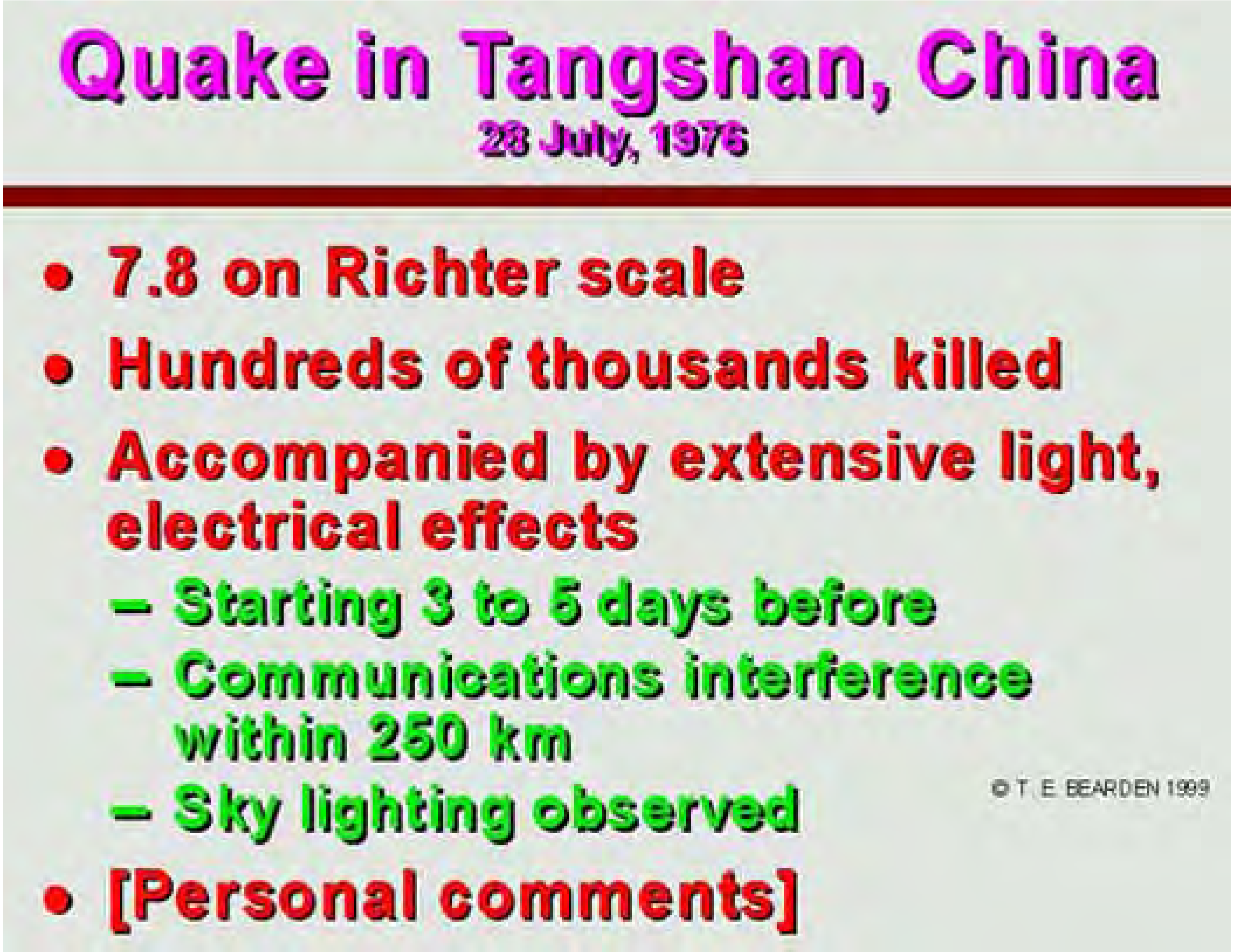

Quake in Tangshan, China

28 July, 1976

- 7.8 on Richter scale
- Hundreds of thousands killed
- Accompanied by extensive light, electrical effects
  - Starting 3 to 5 days before
  - Communications interference within 250 km
  - Sky lighting observed
- [Personal comments]

© T. E. BEARDEN 1999

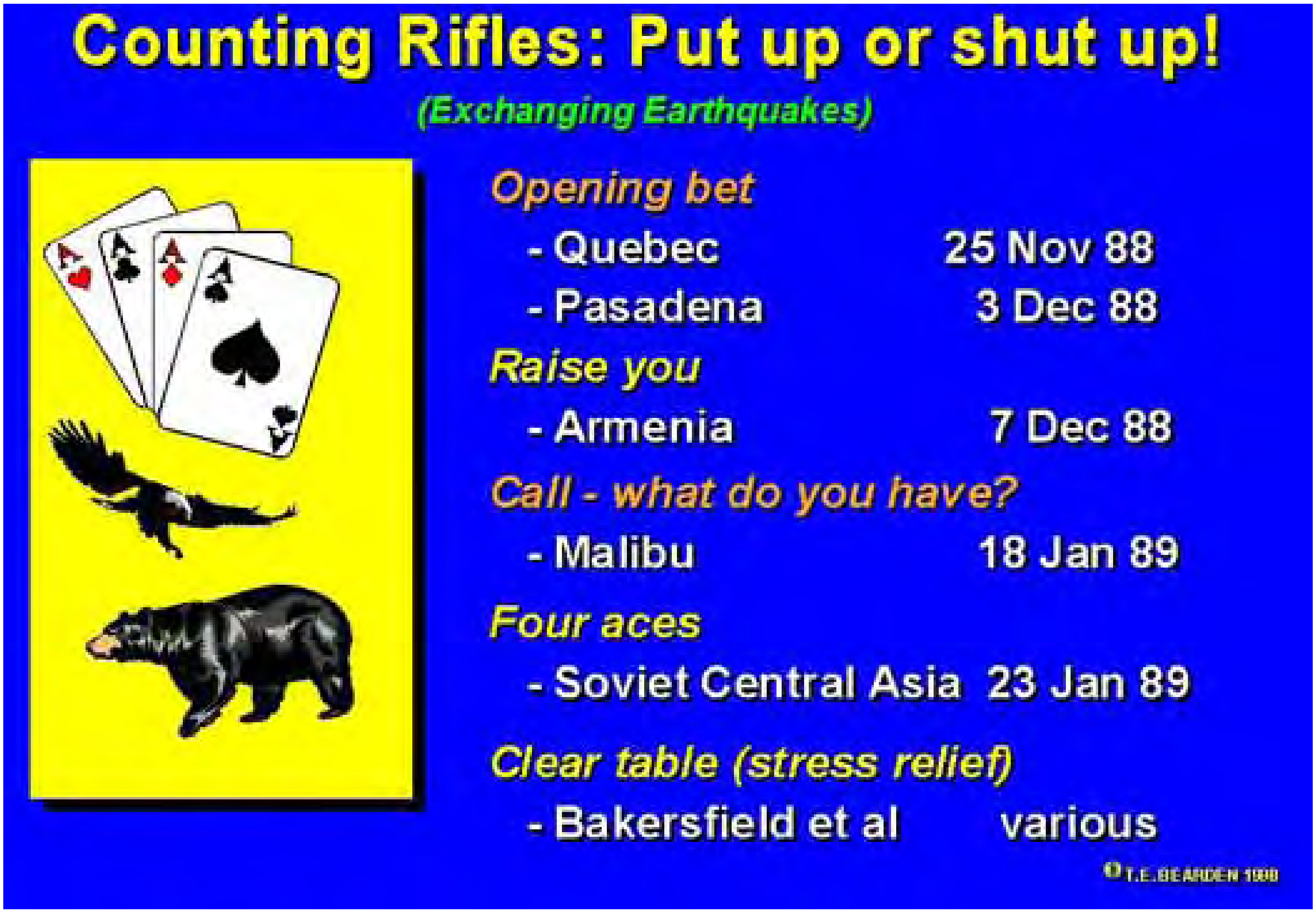

Counting Rifles: Put up or shut up!

(Exchanging Earthquakes)

Opening bet

- Quebec 25 Nov 88
- Pasadena 3 Dec 88

Raise you

- Armenia 7 Dec 88

Call - what do you have?

- Malibu 18 Jan 89

Four aces

- Soviet Central Asia 23 Jan 89

Clear table (stress relief)

- Bakersfield et al various

©T.E. BEARDEN 1990

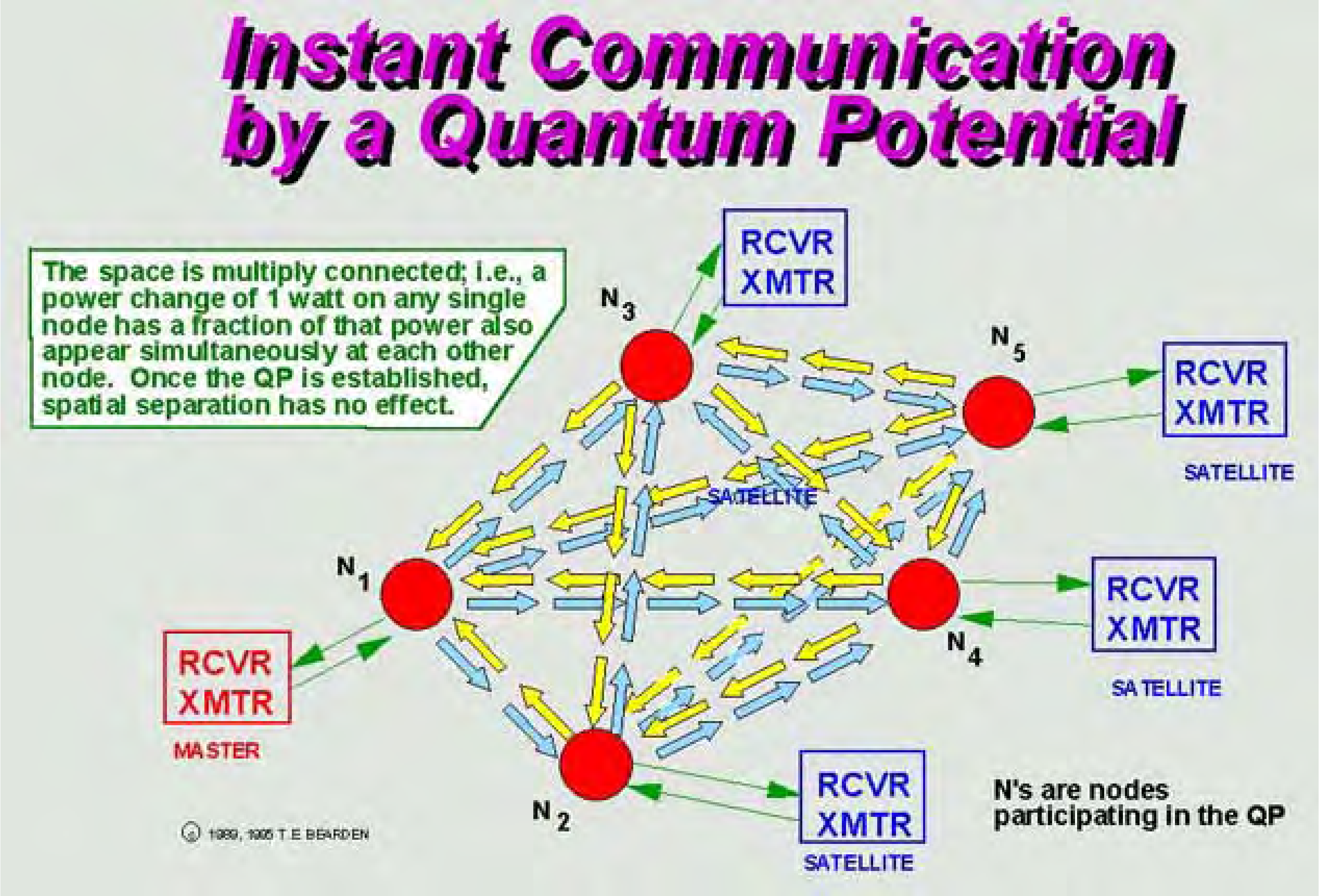

Instant Communication by a Quantum Potential

The space is multiply connected; i.e., a power change of 1 watt on any single node has a fraction of that power also appear simultaneously at each other node. Once the QP is established, spatial separation has no effect.

N's are nodes participating in the QP

N_3
RCVR
XMTR
SATELLITE
N_5
RCVR
XMTR
SATELLITE
N_1
RCVR
XMTR
MASTER
N_4
RCVR
XMTR
SATELLITE
N_2
RCVR
XMTR
SATELLITE

© 1989, 1995 T.E. BEARDEN

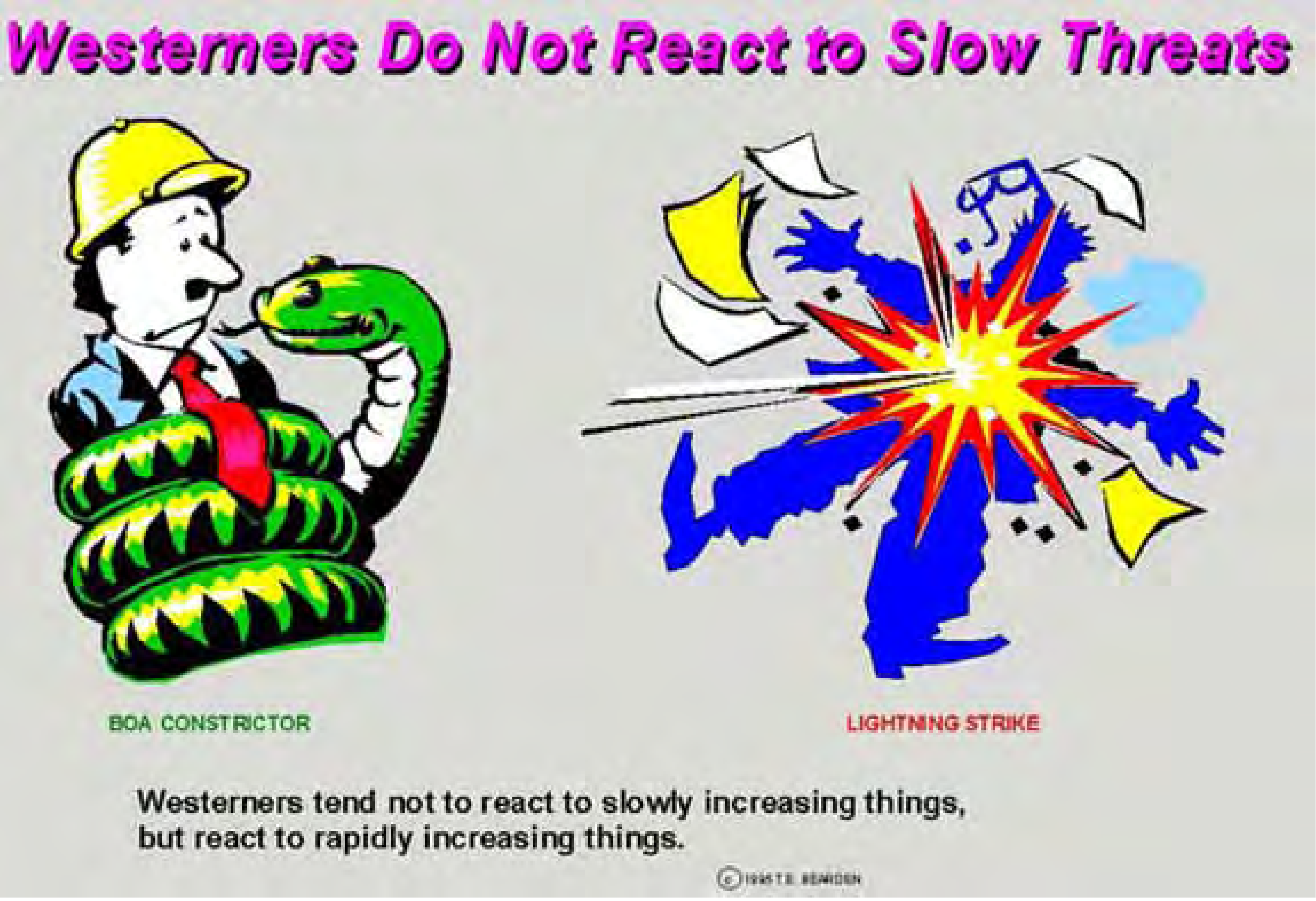

Westerners Do Not React to Slow Threats

BOA CONSTRICTOR
LIGHTNING STRIKE

Westerners tend not to react to slowly increasing things, but react to rapidly increasing things.

© 1995 T.E. BEARDEN

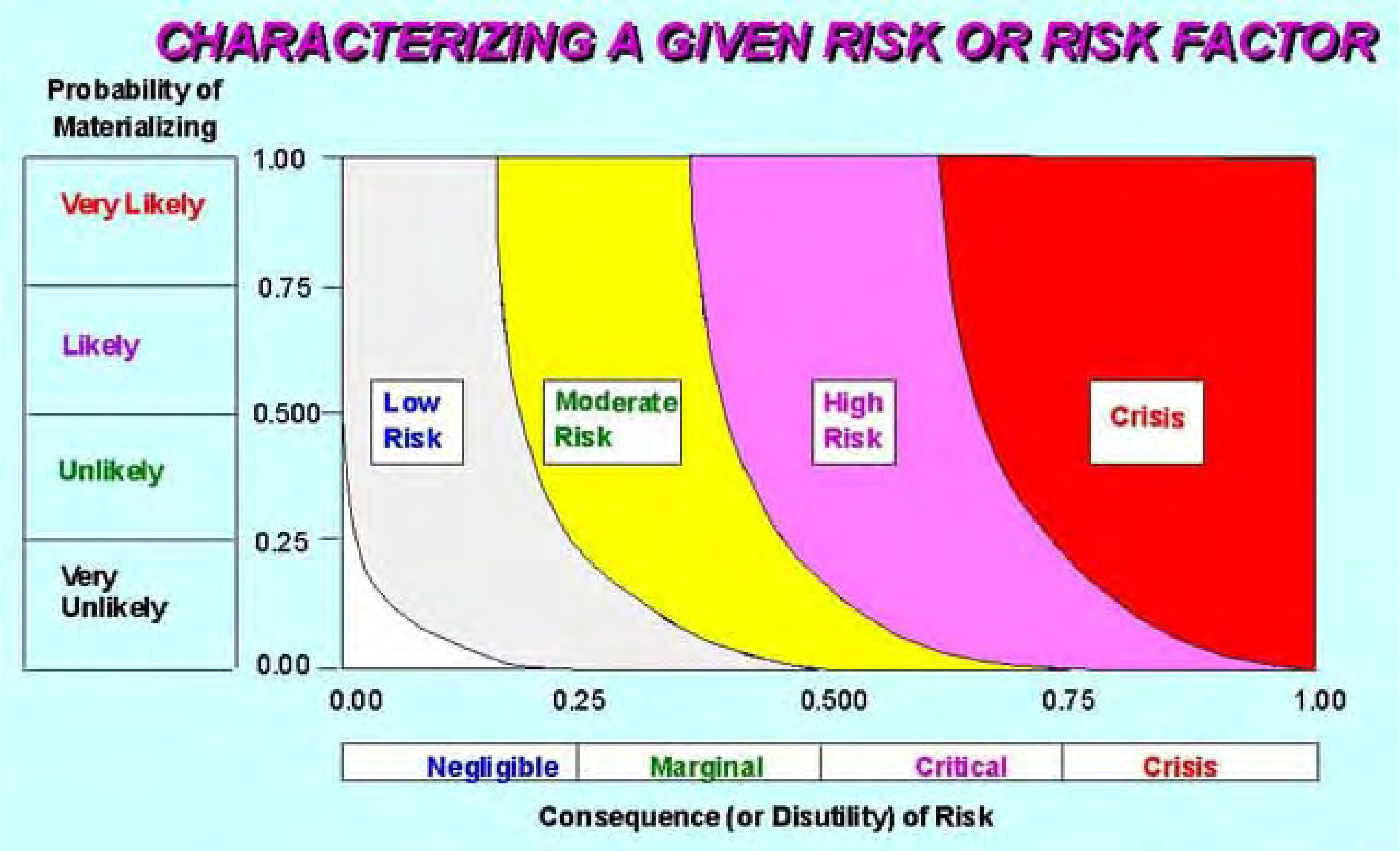
CHARACTERIZING A GIVEN RISK OR RISK FACTOR
Probability of Materializing
Very Likely
Likely
Unlikely
Very Unlikely
1.00
0.75
0.500
0.25
0.00
Low Risk
Moderate Risk
High Risk
Crisis
0.00
0.25
0.500
0.75
1.00
Negligible
Marginal
Critical
Crisis
Consequence (or Disutility) of Risk

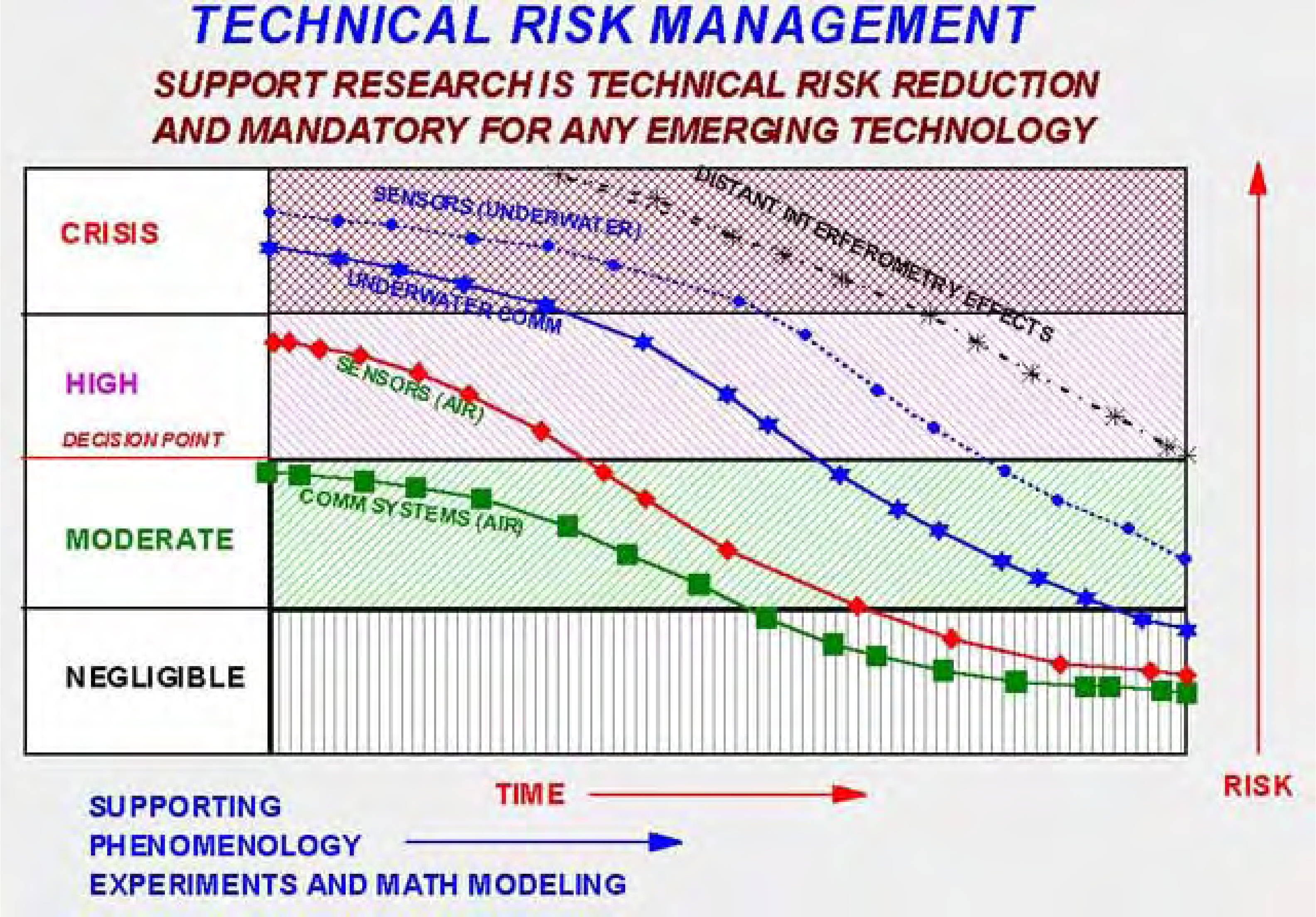

TECHNICAL RISK MANAGEMENT

SUPPORT RESEARCH IS TECHNICAL RISK REDUCTION AND MANDATORY FOR ANY EMERGING TECHNOLOGY

SUPPORTING
PHENOMENOLOGY
EXPERIMENTS AND MATH MODELING

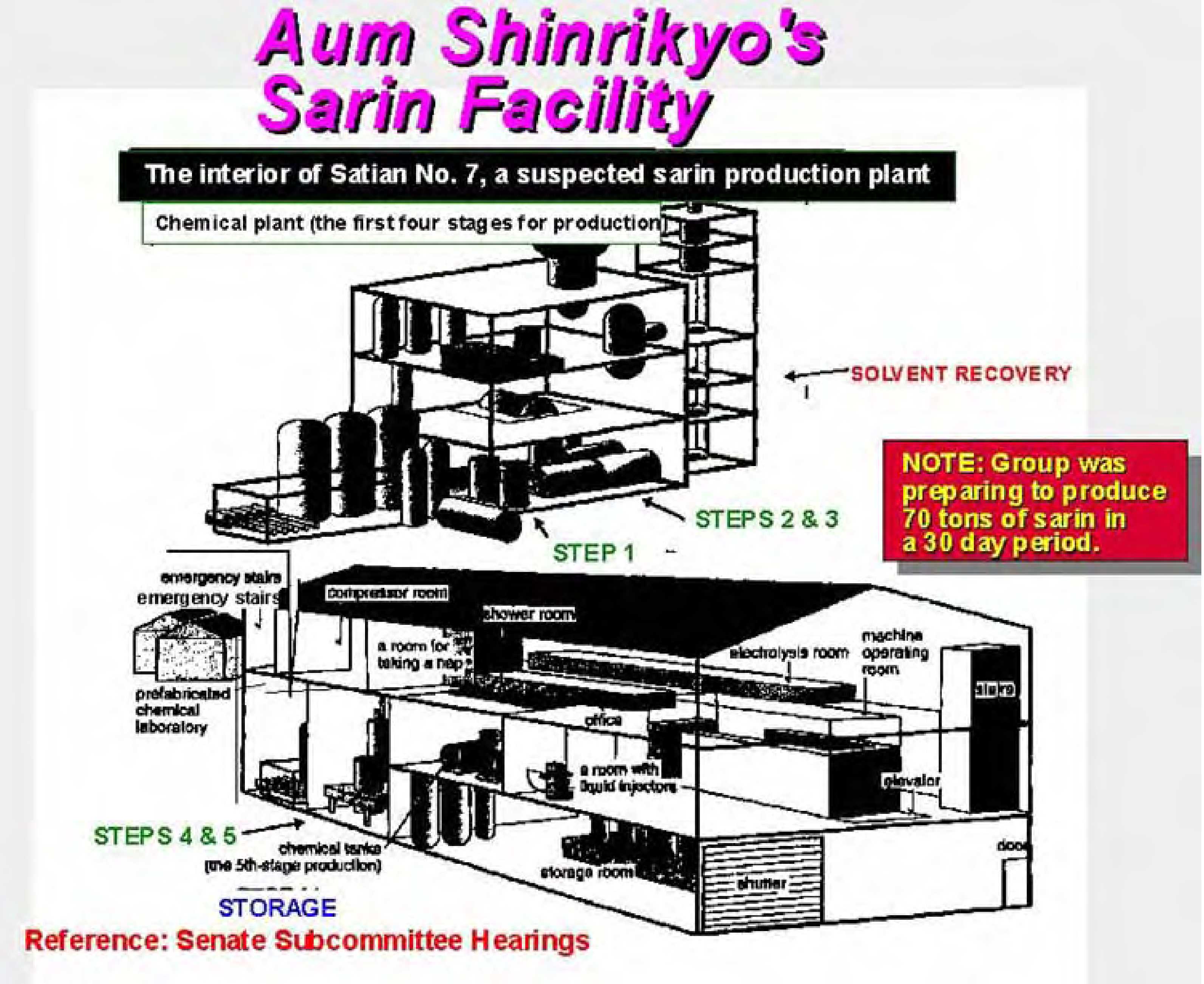

Aum Shinrikyo's Sarin Facility

The interior of Satian No. 7, a suspected sarin production plant

Chemical plant (the first four stages for production)

SOLVENT RECOVERY
STEPS 2 & 3
STEP 1
emergency stairs
emergency stairs
compressor room
shower room
a room for taking a nap
electrolysis room
machine operating room
stairs
prefabricated chemical laboratory
office
a room with liquid injectors
elevator
STEPS 4 & 5
chemical tanks (the 5th-stage production)
storage room
shutter
door
STORAGE

NOTE: Group was preparing to produce 70 tons of sarin in a 30 day period.

Reference: Senate Subcommittee Hearings

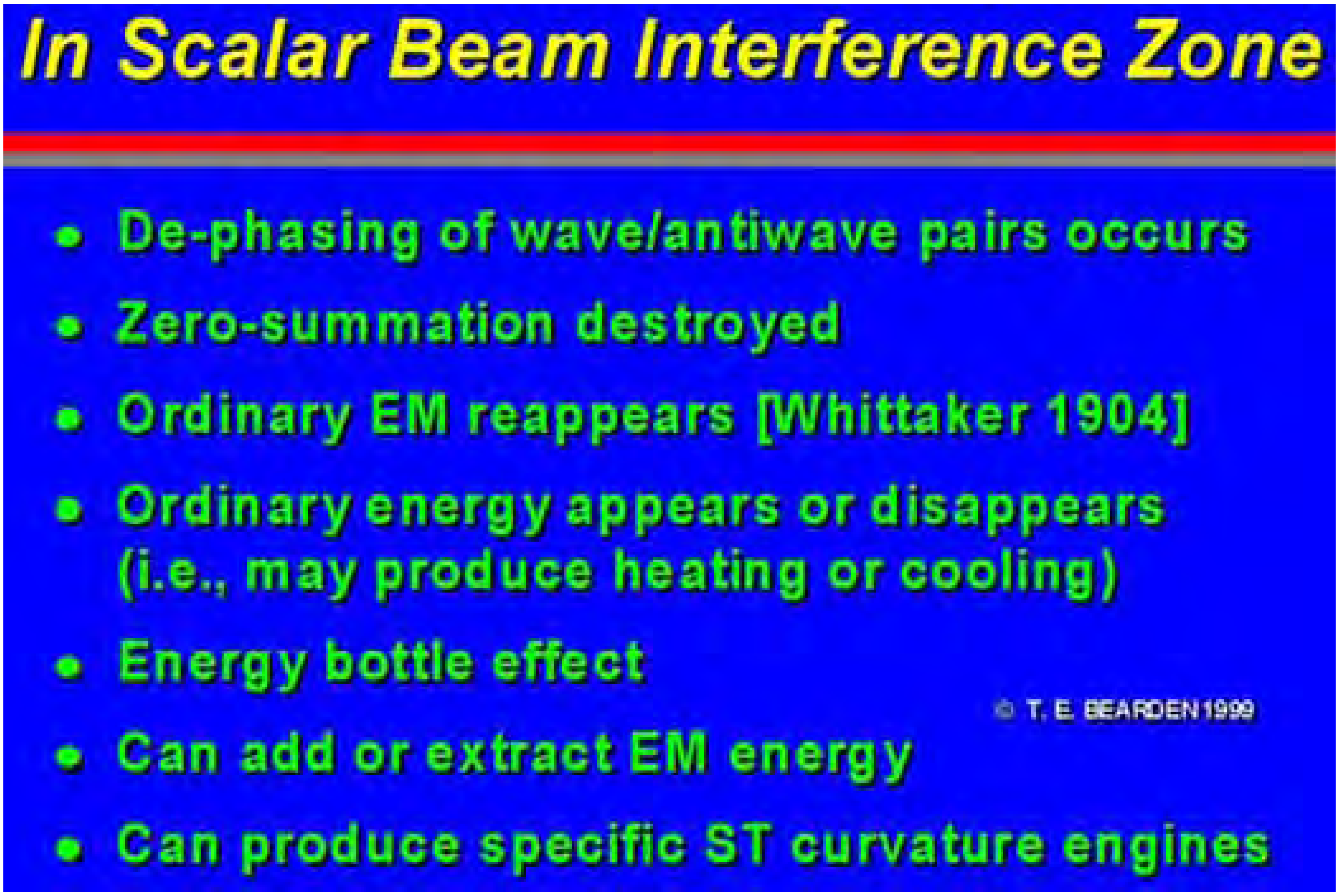

In Scalar Beam Interference Zone

- De-phasing of wave/antiwave pairs occurs
- Zero-summation destroyed
- Ordinary EM reappears [Whittaker 1904]
- Ordinary energy appears or disappears (i.e., may produce heating or cooling)
- Energy bottle effect
- Can add or extract EM energy
- Can produce specific ST curvature engines

© T. E. BEARDEN 1999

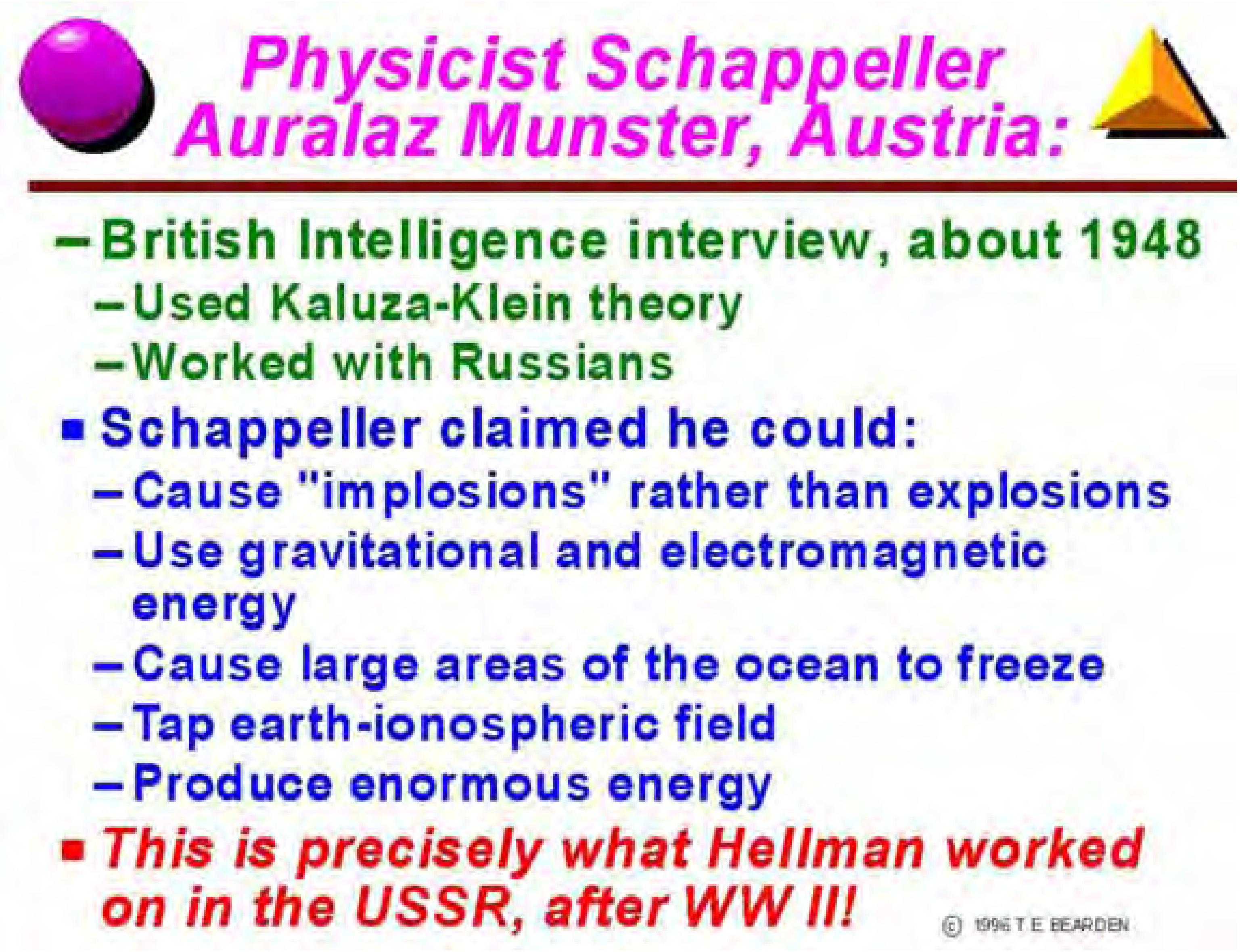

Physicist Schappeller Auralaz Munster, Austria:

- British Intelligence interview, about 1948
  - Used Kaluza-Klein theory
  - Worked with Russians
- Schappeller claimed he could:
  - Cause "implosions" rather than explosions
  - Use gravitational and electromagnetic energy
  - Cause large areas of the ocean to freeze
  - Tap earth-ionospheric field
  - Produce enormous energy
- *This is precisely what Hellman worked on in the USSR, after WW II!*

© 1996 T E BEARDEN

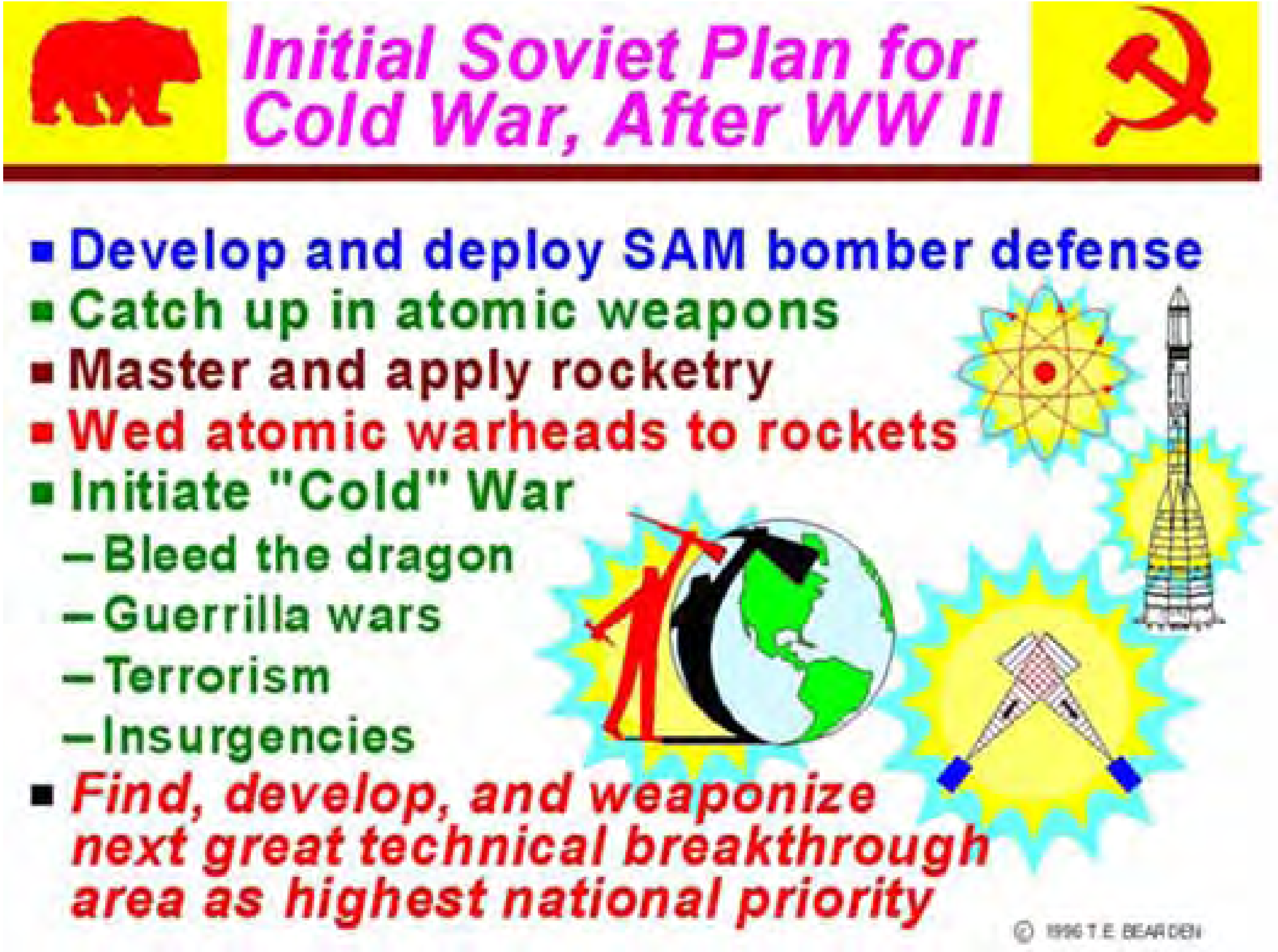

Initial Soviet Plan for Cold War, After WW II

- Develop and deploy SAM bomber defense
- Catch up in atomic weapons
- Master and apply rocketry
- Wed atomic warheads to rockets
- Initiate "Cold" War
  - Bleed the dragon
  - Guerrilla wars
  - Terrorism
  - Insurgencies
- *Find, develop, and weaponize next great technical breakthrough area as highest national priority*

© 1996 T.E. BEARDEN

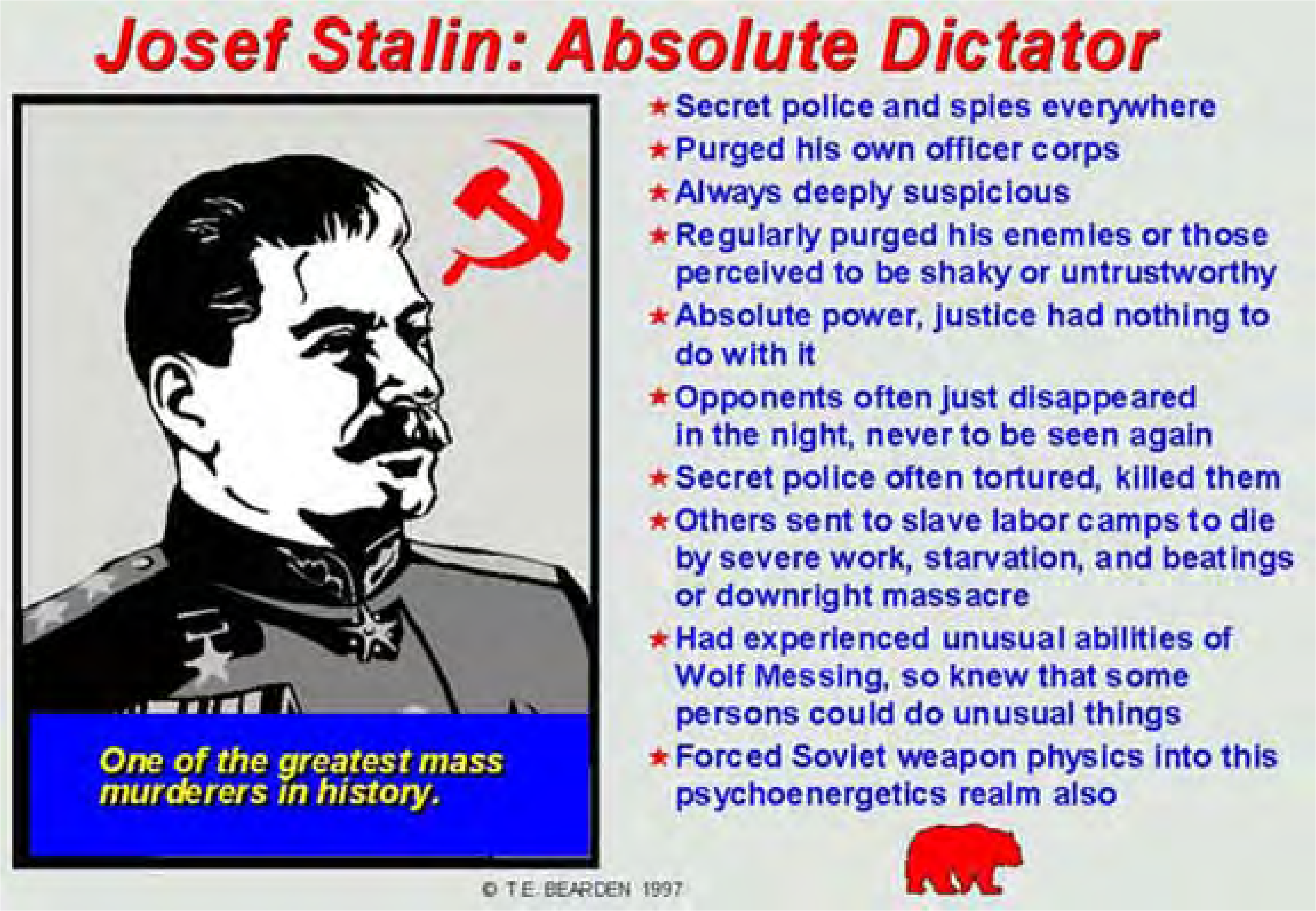

Josef Stalin: Absolute Dictator

One of the greatest mass murderers in history.

- ★ Secret police and spies everywhere
- ★ Purged his own officer corps
- ★ Always deeply suspicious
- ★ Regularly purged his enemies or those perceived to be shaky or untrustworthy
- ★ Absolute power, justice had nothing to do with it
- ★ Opponents often just disappeared in the night, never to be seen again
- ★ Secret police often tortured, killed them
- ★ Others sent to slave labor camps to die by severe work, starvation, and beatings or downright massacre
- ★ Had experienced unusual abilities of Wolf Messing, so knew that some persons could do unusual things
- ★ Forced Soviet weapon physics into this psychoenergetics realm also

© T.E. BEARDEN 1997

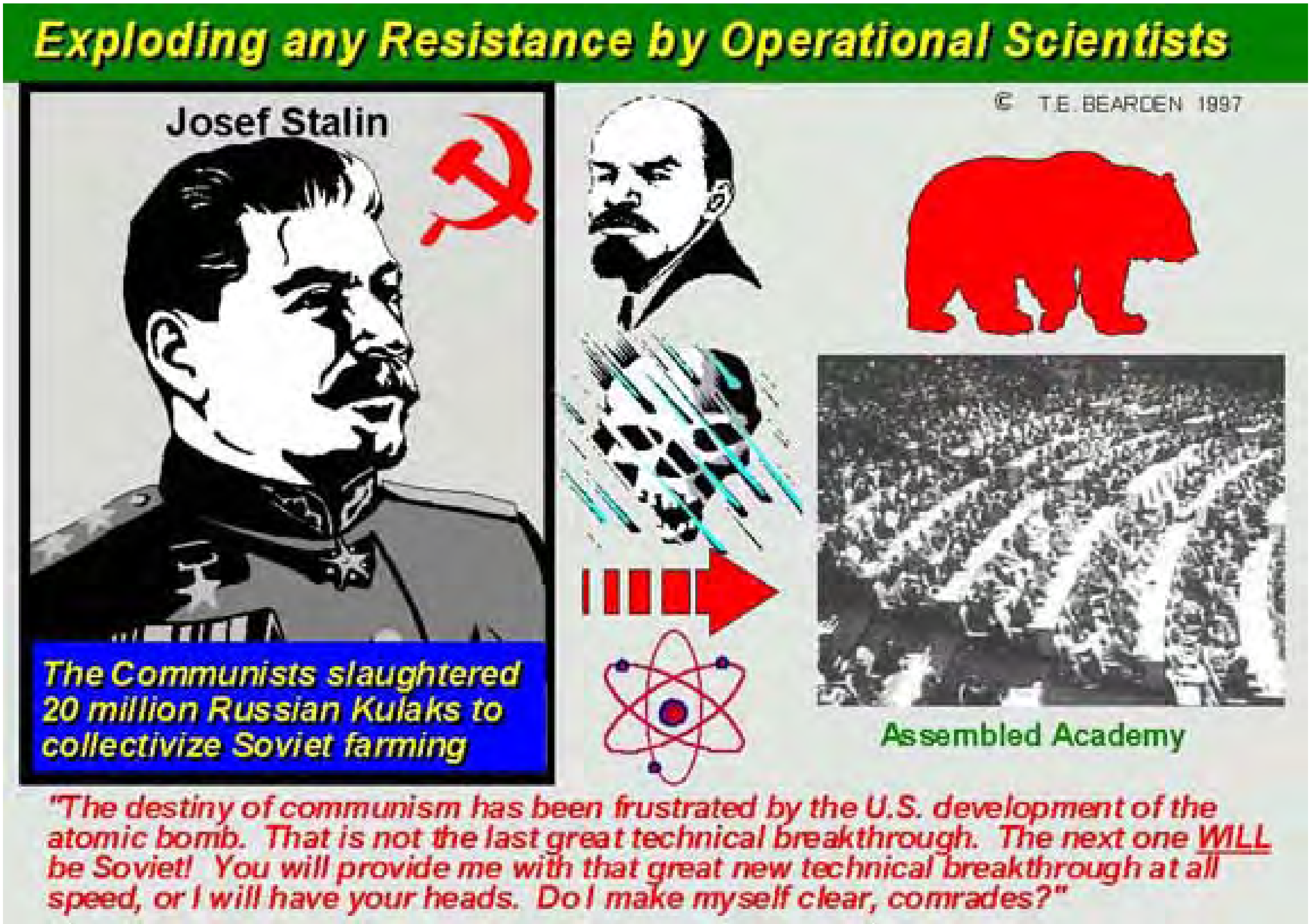

Exploding any Resistance by Operational Scientists

© T.E. BEARDEN 1997

Josef Stalin

The Communists slaughtered 20 million Russian Kulaks to collectivize Soviet farming

Assembled Academy

"The destiny of communism has been frustrated by the U.S. development of the atomic bomb. That is not the last great technical breakthrough. The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads. Do I make myself clear, comrades?"

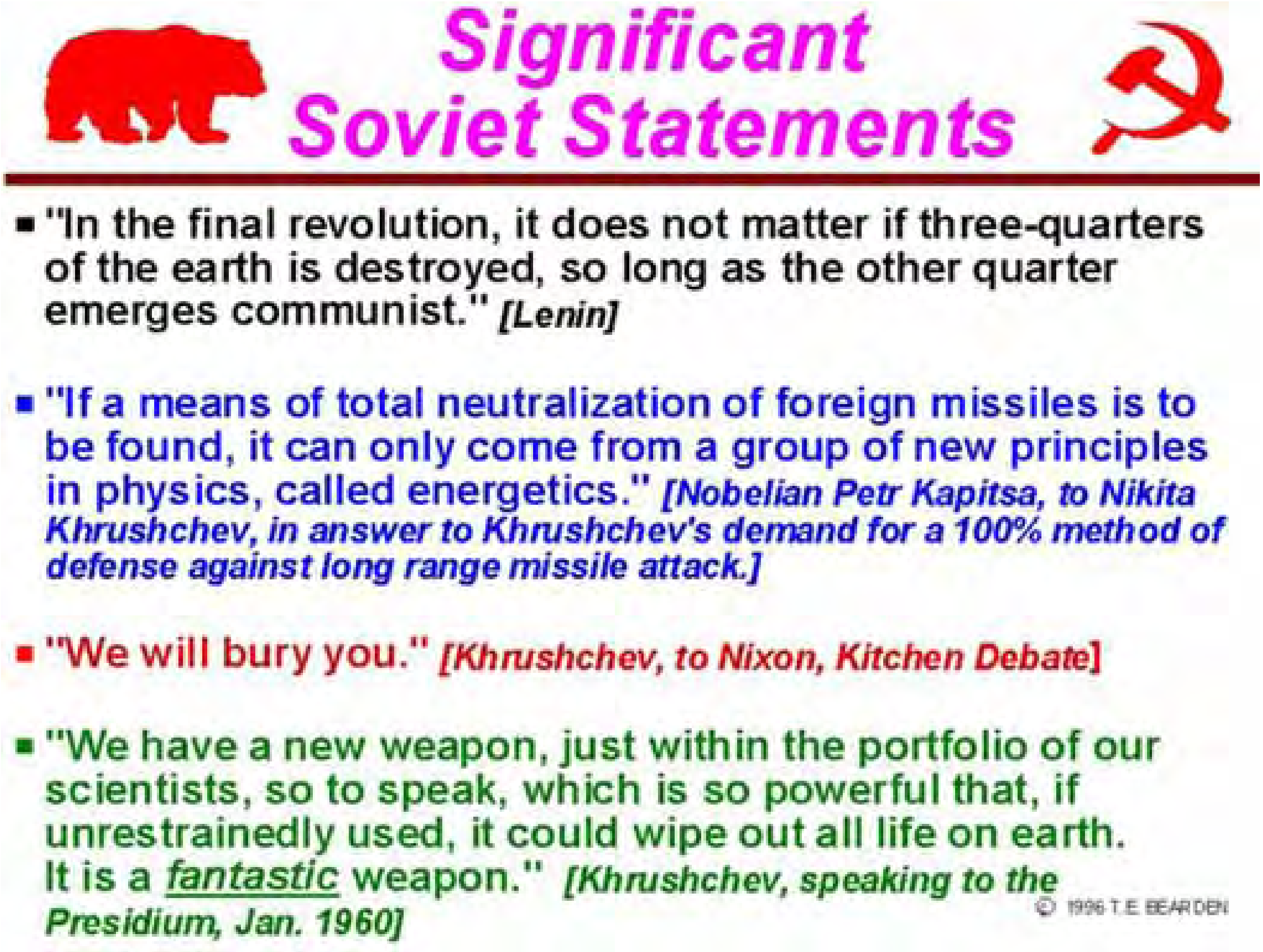

Significant Soviet Statements

- "In the final revolution, it does not matter if three-quarters of the earth is destroyed, so long as the other quarter emerges communist." *[Lenin]*
- "If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics." *[Nobelian Petr Kapitsa, to Nikita Khrushchev, in answer to Khrushchev's demand for a 100% method of defense against long range missile attack.]*
- "We will bury you." *[Khrushchev, to Nixon, Kitchen Debate*]
- "We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a *fantastic* weapon." *[Khrushchev, speaking to the Presidium, Jan. 1960]*

© 1996 T.E. BEARDEN

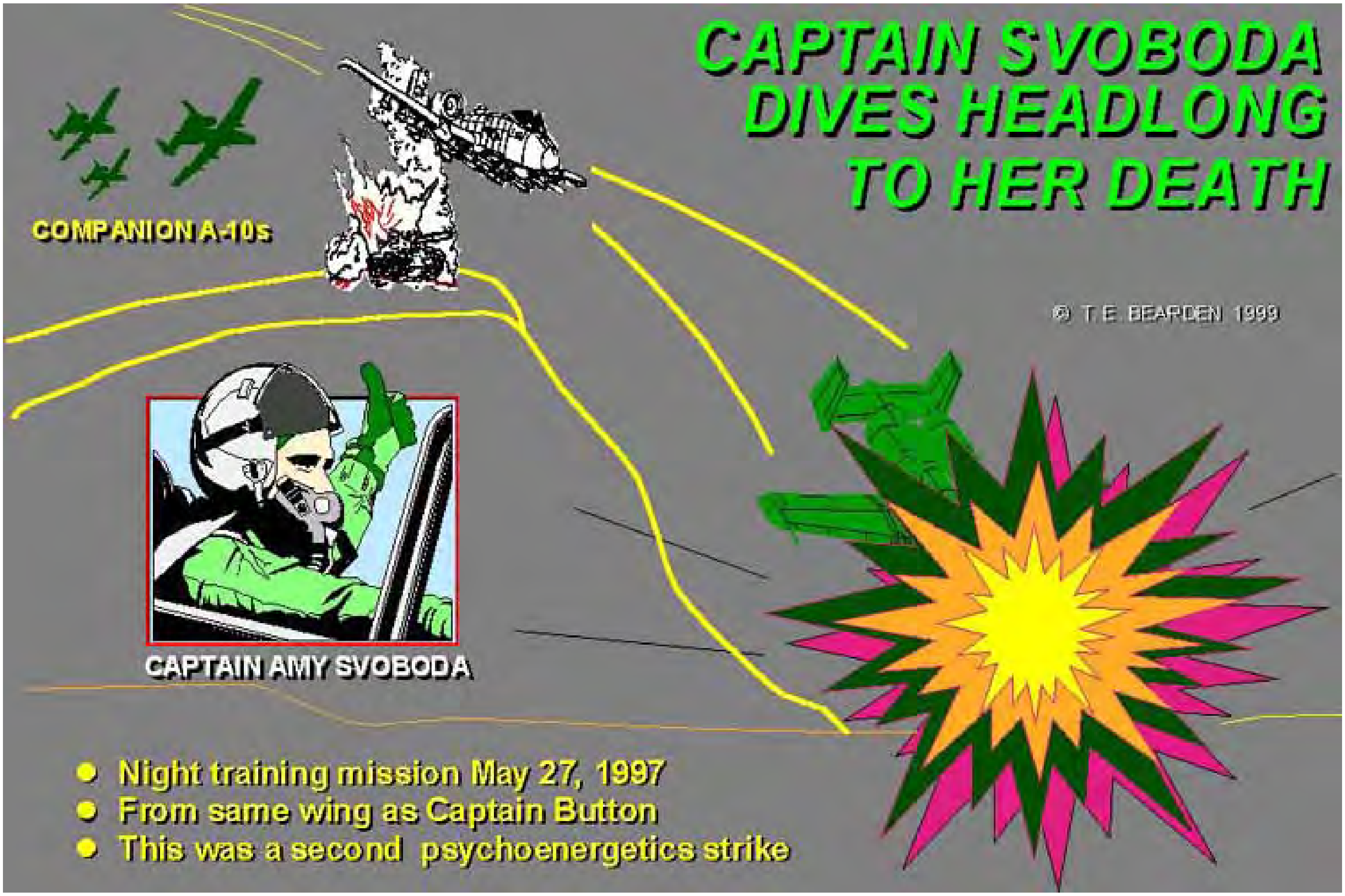

CAPTAIN SVOBODA DIVES HEADLONG TO HER DEATH

COMPANION A-10s
© T.E. BEARDEN 1999
CAPTAIN AMY SVOBODA

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

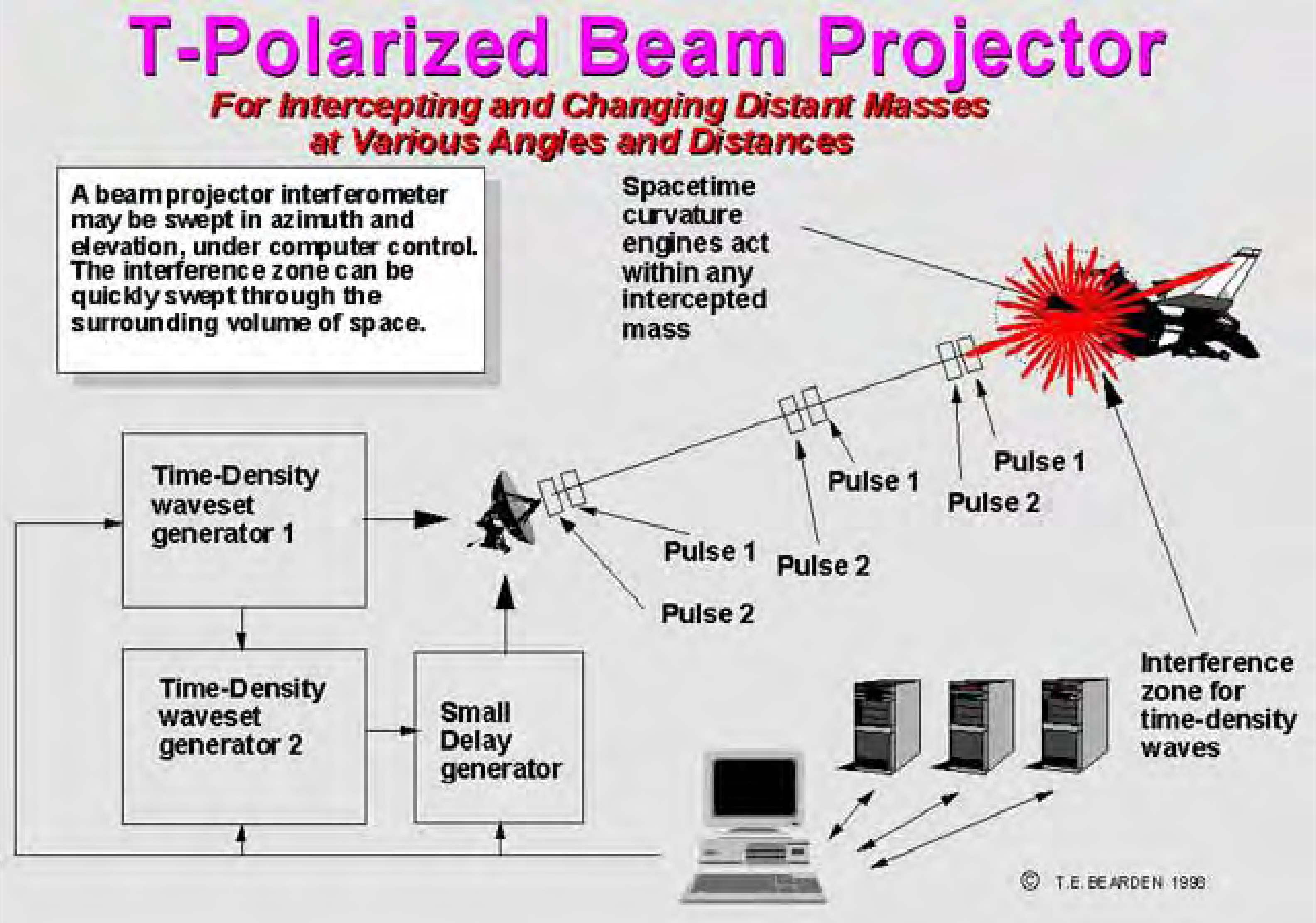

T-Polarized Beam Projector

For Intercepting and Changing Distant Masses at Various Angles and Distances

A beam projector interferometer may be swept in azimuth and elevation, under computer control. The interference zone can be quickly swept through the surrounding volume of space.

Spacetime curvature engines act within any intercepted mass

Time-Density waveset generator 1
Time-Density waveset generator 2
Small Delay generator
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Pulse 1
Pulse 2
Interference zone for time-density waves
© T.E. BEARDEN 1996

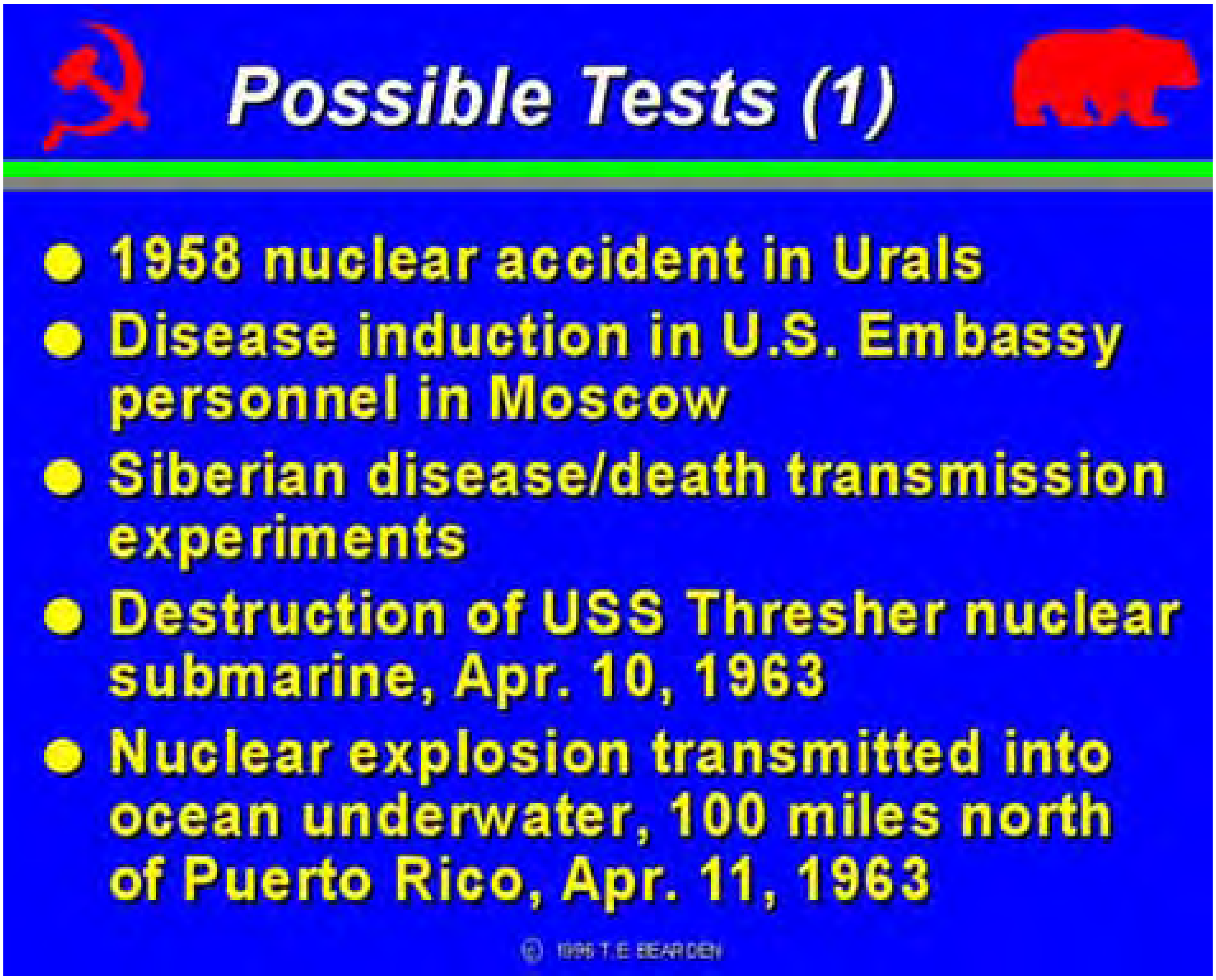

Possible Tests (1)

- 1958 nuclear accident in Urals
- Disease induction in U.S. Embassy personnel in Moscow
- Siberian disease/death transmission experiments
- Destruction of USS Thresher nuclear submarine, Apr. 10, 1963
- Nuclear explosion transmitted into ocean underwater, 100 miles north of Puerto Rico, Apr. 11, 1963

© 1996 T. E. BEARDEN

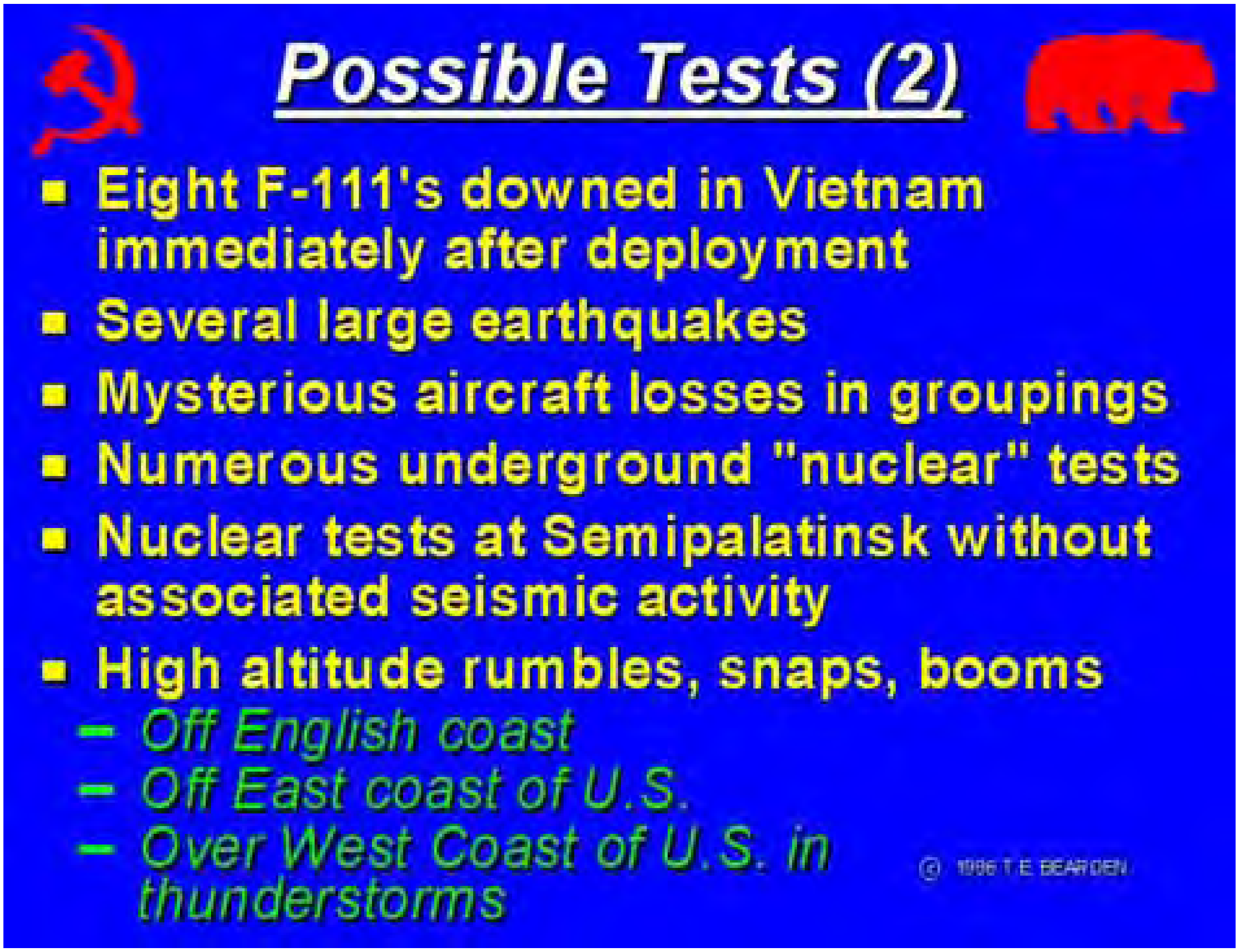

Possible Tests (2)

- Eight F-111's downed in Vietnam immediately after deployment
- Several large earthquakes
- Mysterious aircraft losses in groupings
- Numerous underground "nuclear" tests
- Nuclear tests at Semipalatinsk without associated seismic activity
- High altitude rumbles, snaps, booms
  - *Off English coast*
  - *Off East coast of U.S.*
  - *Over West Coast of U.S. in thunderstorms*

© 1996 T.E. BEARDEN

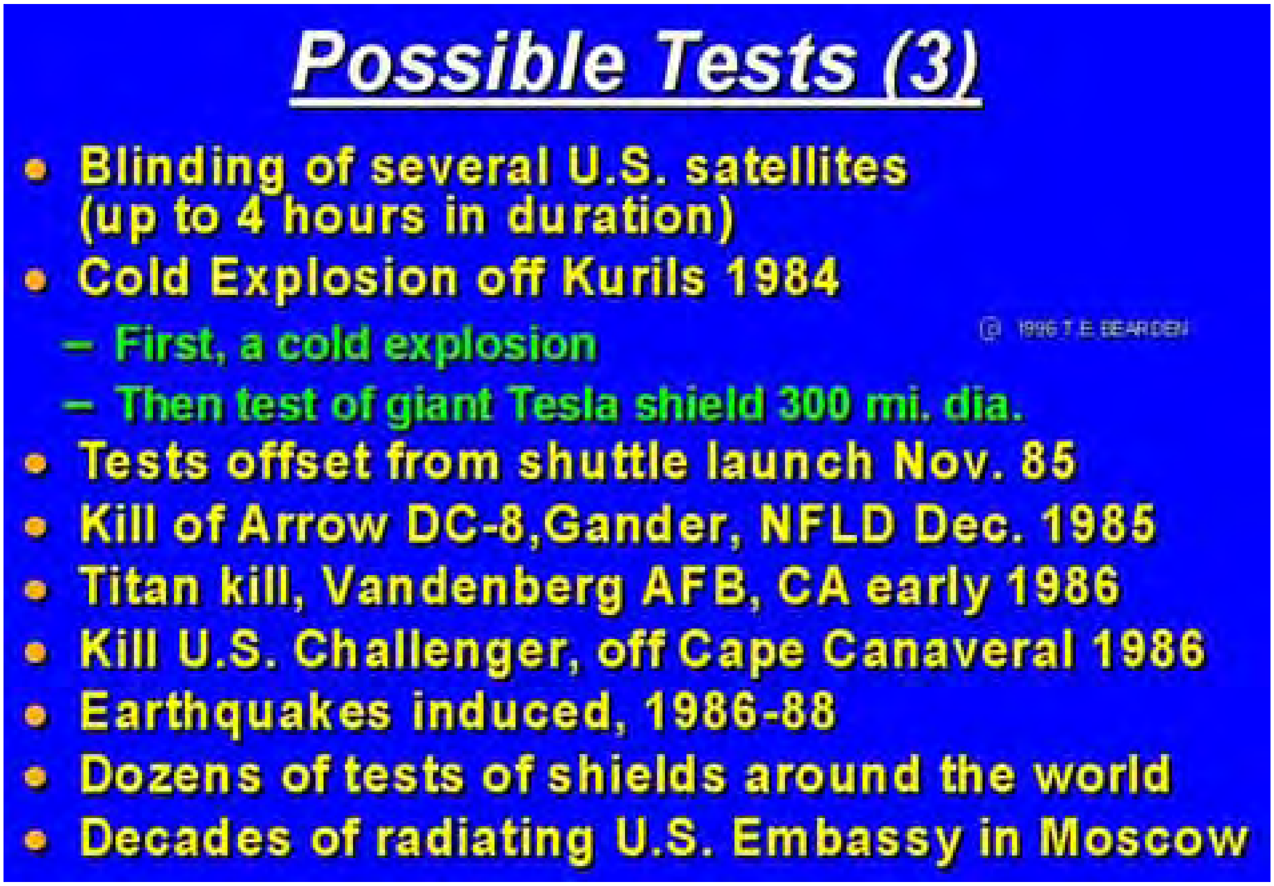

Possible Tests (3)

- Blinding of several U.S. satellites (up to 4 hours in duration)
- Cold Explosion off Kurils 1984
  - First, a cold explosion
  - Then test of giant Tesla shield 300 mi. dia.
- Tests offset from shuttle launch Nov. 85
- Kill of Arrow DC-8,Gander, NFLD Dec. 1985
- Titan kill, Vandenberg AFB, CA early 1986
- Kill U.S. Challenger, off Cape Canaveral 1986
- Earthquakes induced, 1986-88
- Dozens of tests of shields around the world
- Decades of radiating U.S. Embassy in Moscow

© 1996 T.E. BEARDEN

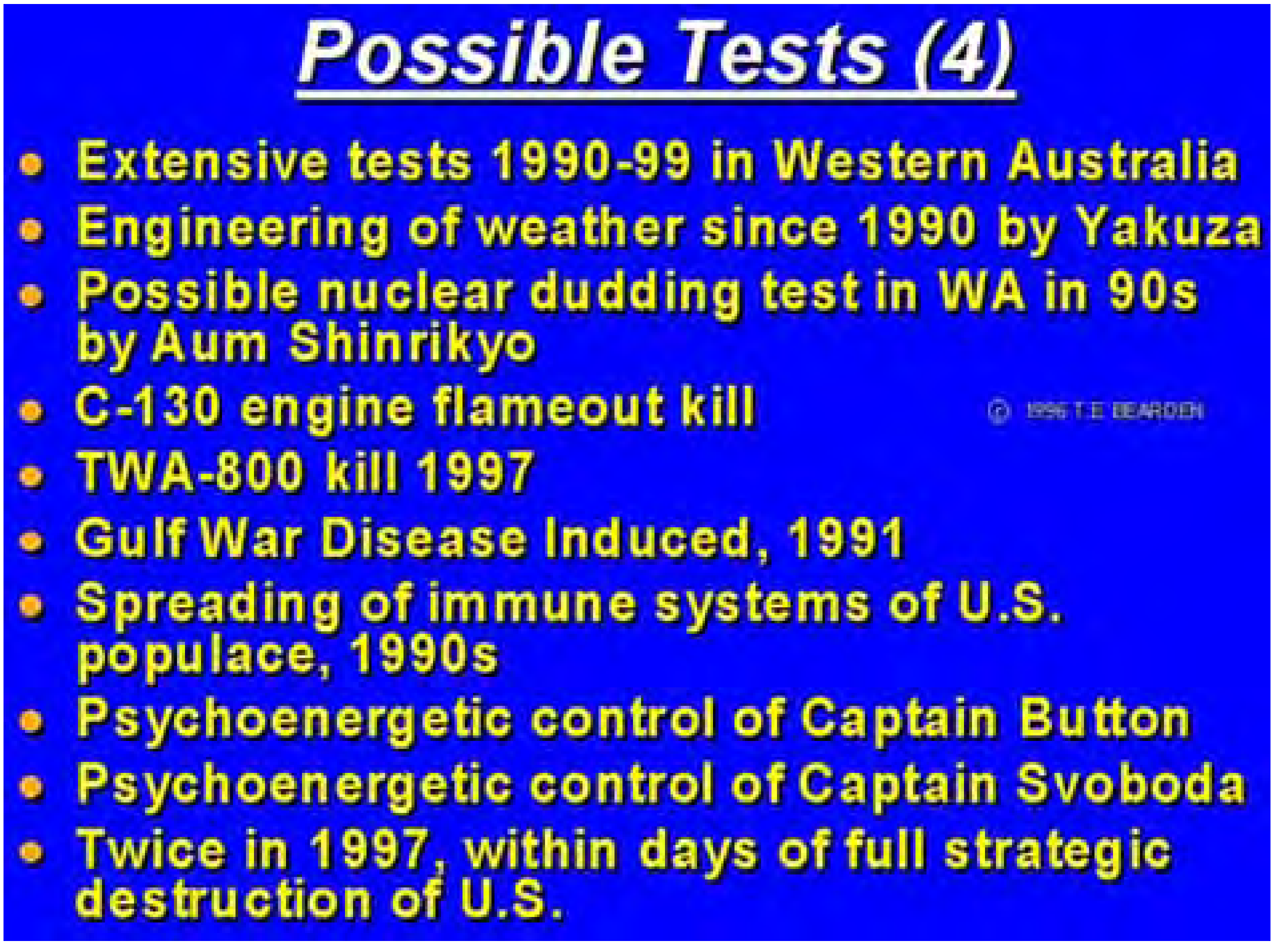

Possible Tests (4)

- Extensive tests 1990-99 in Western Australia
- Engineering of weather since 1990 by Yakuza
- Possible nuclear dudding test in WA in 90s by Aum Shinrikyo
- C-130 engine flameout kill
- TWA-800 kill 1997
- Gulf War Disease Induced, 1991
- Spreading of immune systems of U.S. populace, 1990s
- Psychoenergetic control of Captain Button
- Psychoenergetic control of Captain Svoboda
- Twice in 1997, within days of full strategic destruction of U.S.

© 1996 T.E. BEARDEN

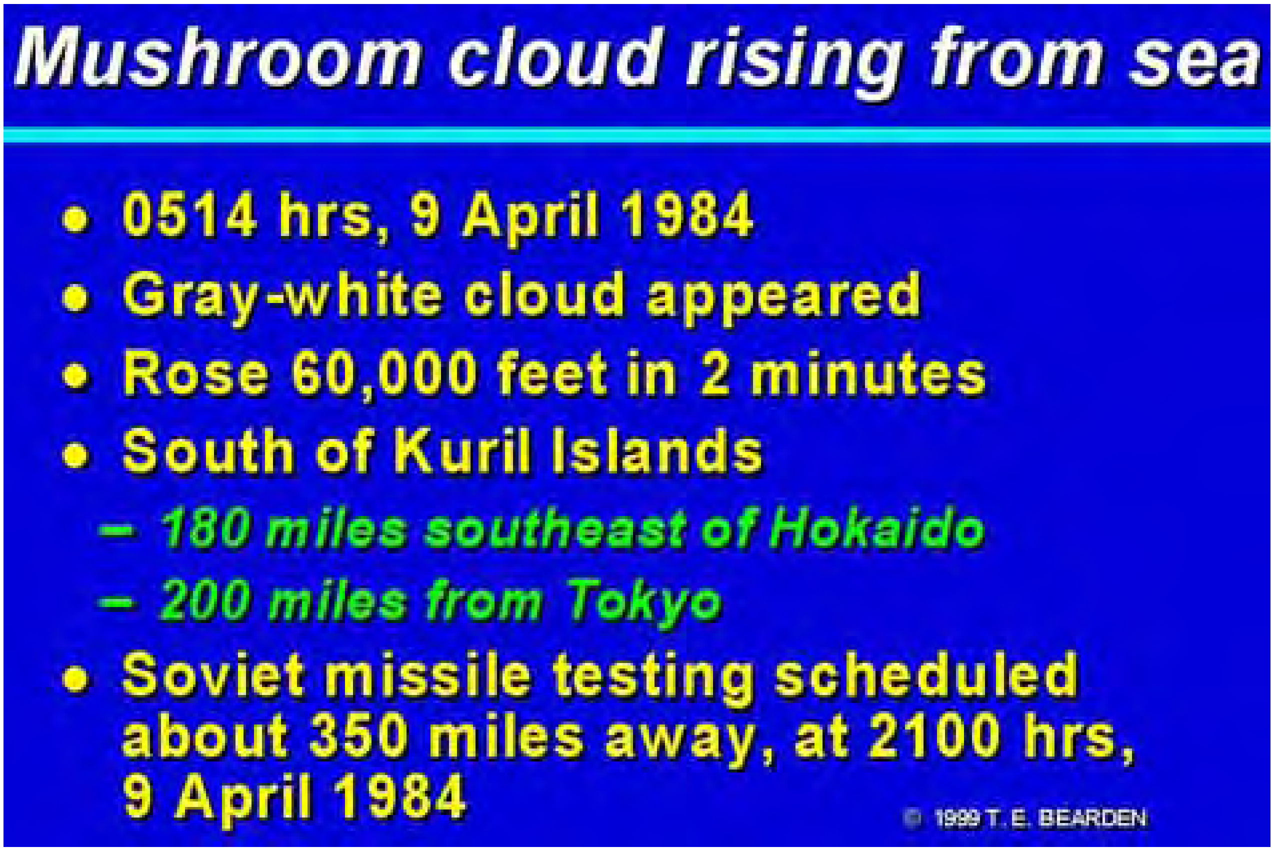

Mushroom cloud rising from sea

- 0514 hrs, 9 April 1984
- Gray-white cloud appeared
- Rose 60,000 feet in 2 minutes
- South of Kuril Islands
  - *180 miles southeast of Hokaido*
  - *200 miles from Tokyo*
- Soviet missile testing scheduled about 350 miles away, at 2100 hrs, 9 April 1984

© 1999 T. E. BEARDEN

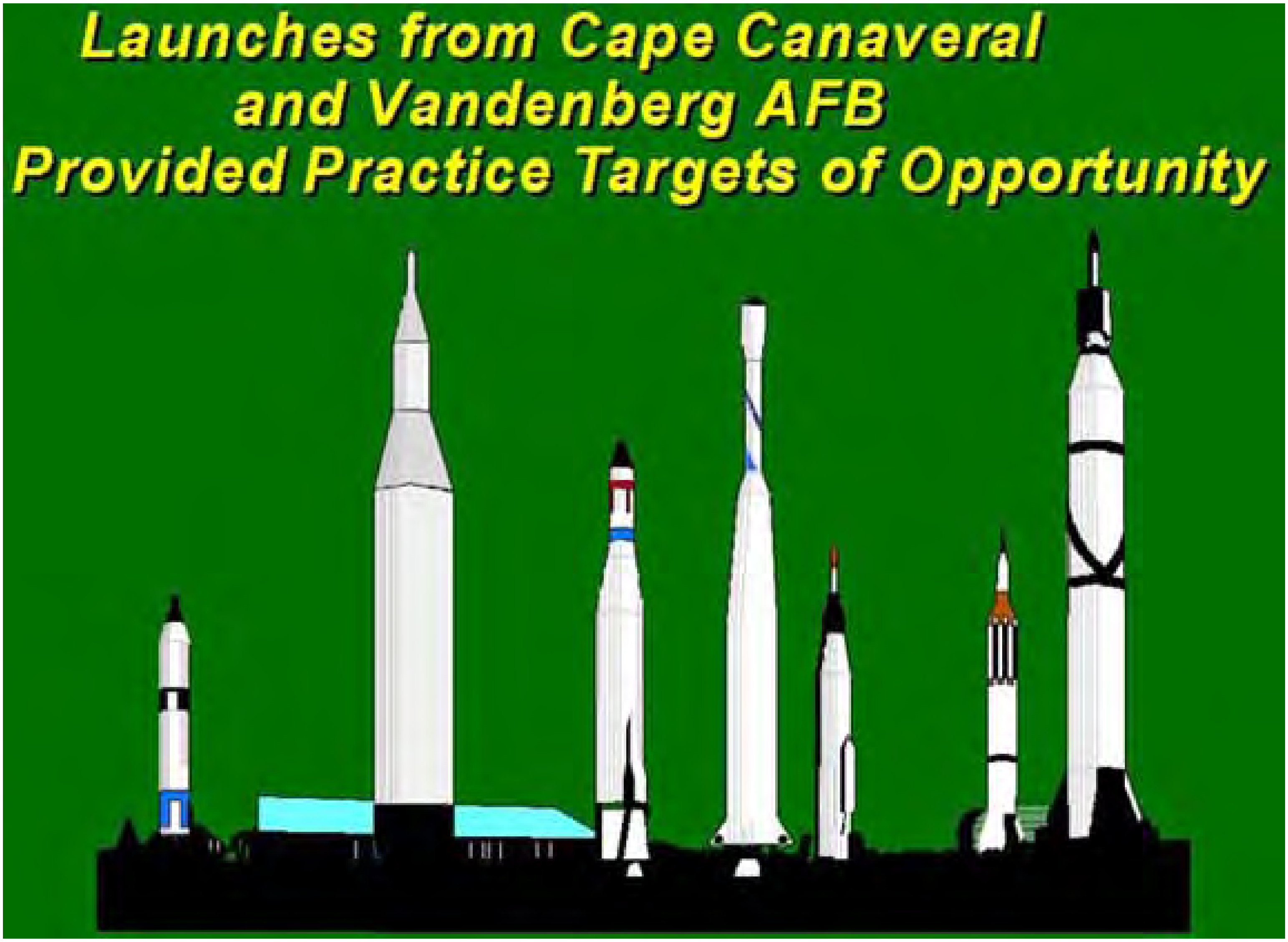
Launches from Cape Canaveral
and Vandenberg AFB
Provided Practice Targets of Opportunity

The sinking U.S.S. Thresher, with her controls jammed, implodes when she reaches crush depth

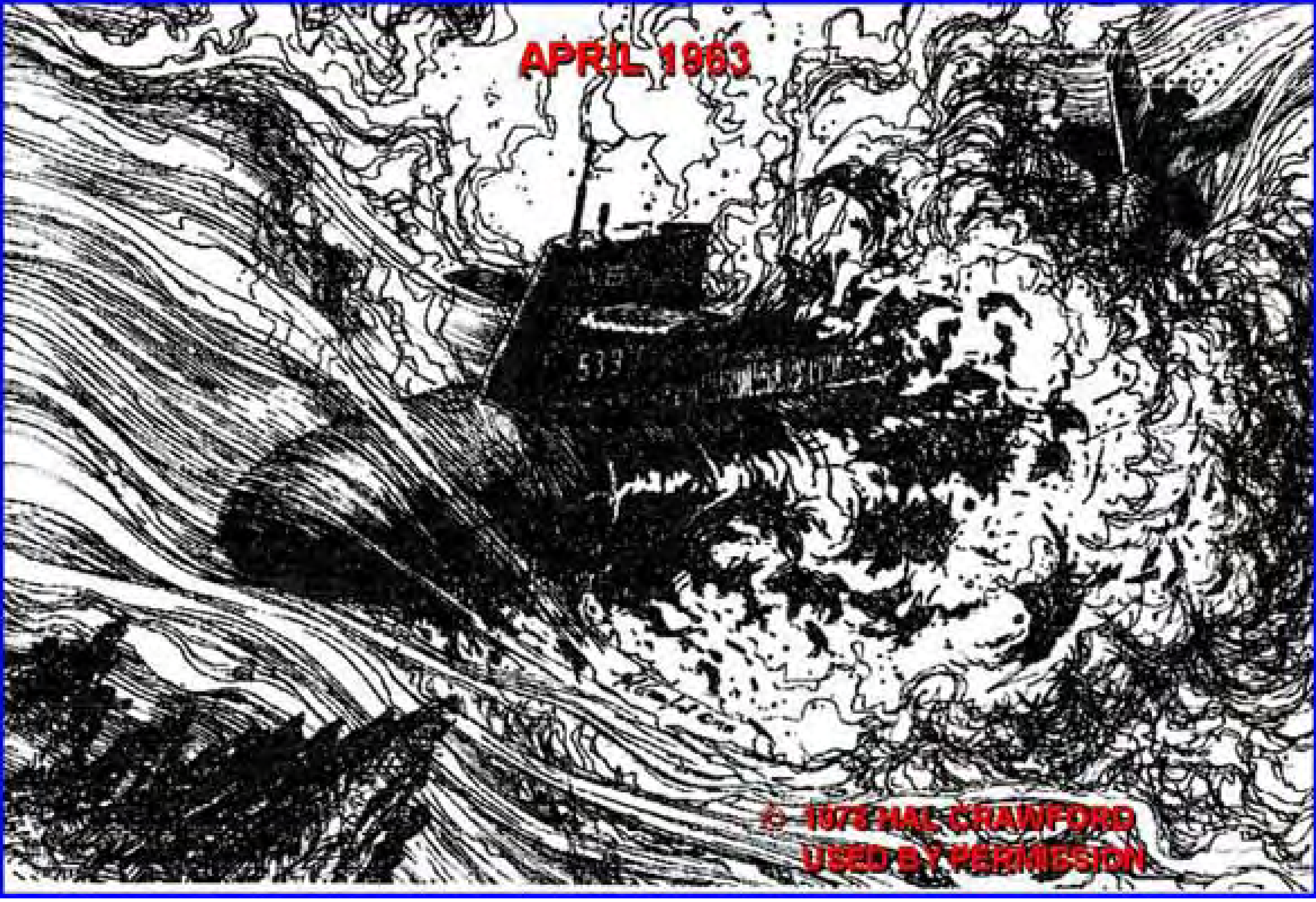
APRIL 1963
533
© 1976 HAL CRAWFORD
USED BY PERMISSION

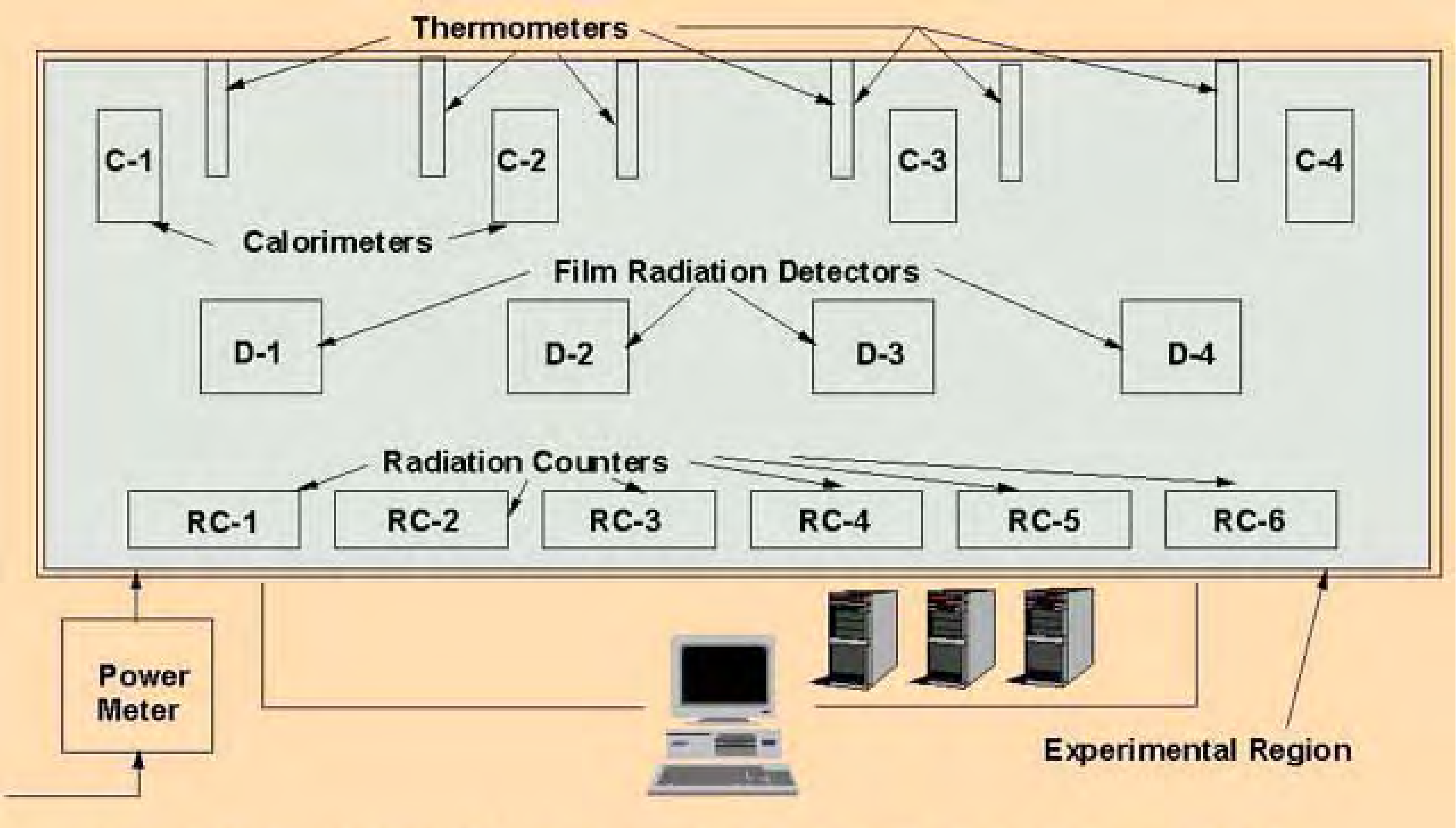
Thermometers
C-1
C-2
C-3
C-4
Calorimeters
Film Radiation Detectors
D-1
D-2
D-3
D-4
Radiation Counters
RC-1
RC-2
RC-3
RC-4
RC-5
RC-6
Power Meter
Experimental Region

Figure 32. Embodiment of an instrumental array sufficient to discriminate when time-density charging has occurred, and that transduction of time-charge decay is occurring.

© T.E. BEARDEN 1998

Time-Polarized EM Wave interferometer Creating Space-Time Curvature Engines in a Distant Interference Zone

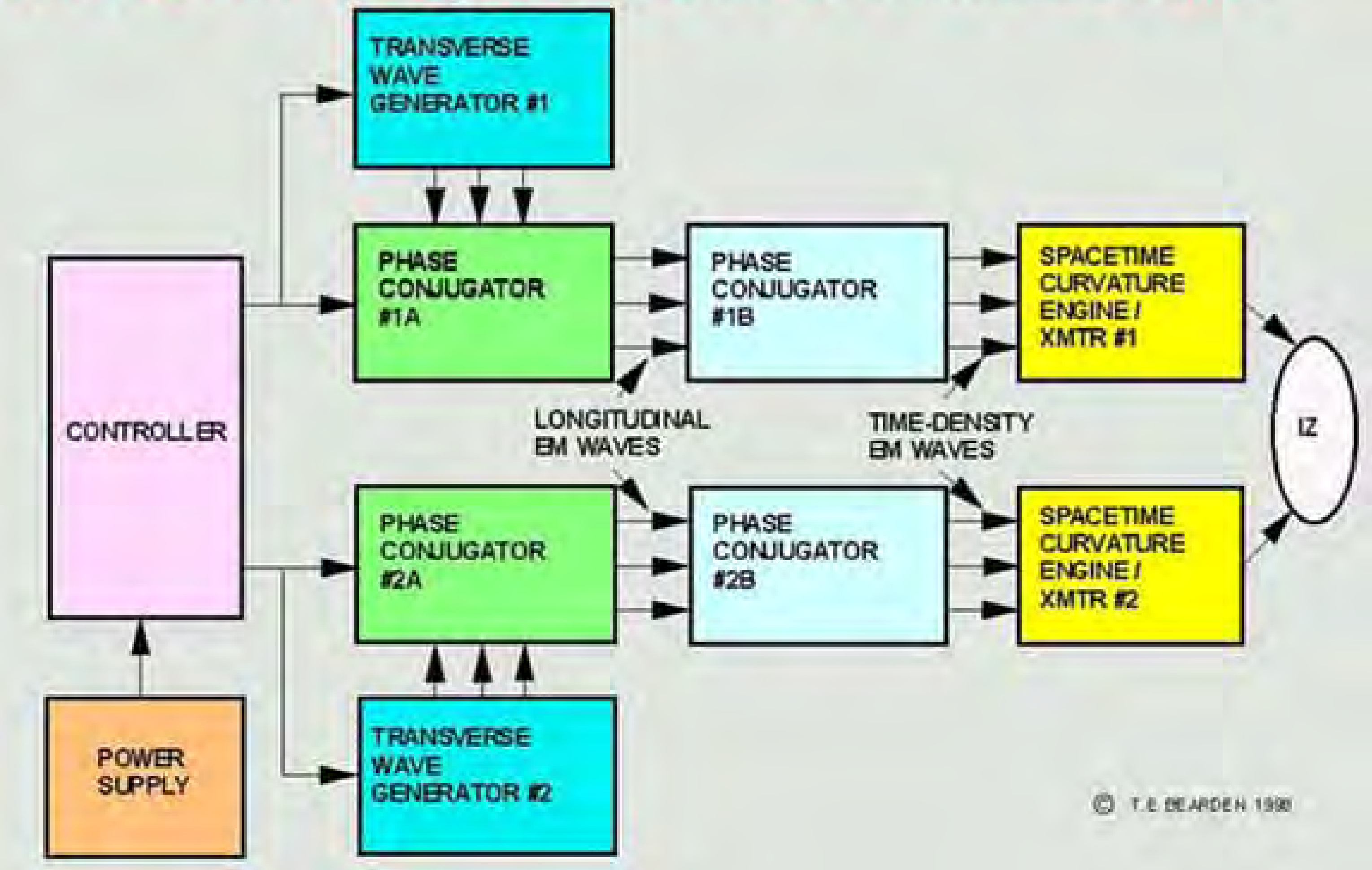
TRANSVERSE WAVE GENERATOR #1
CONTROLLER
PHASE CONJUGATOR #1A
PHASE CONJUGATOR #1B
SPACETIME CURVATURE ENGINE / XMTR #1
LONGITUDINAL EM WAVES
TIME-DENSITY EM WAVES
IZ
PHASE CONJUGATOR #2A
PHASE CONJUGATOR #2B
SPACETIME CURVATURE ENGINE / XMTR #2
POWER SUPPLY
TRANSVERSE WAVE GENERATOR #2
© T.E. BEARDEN 1990

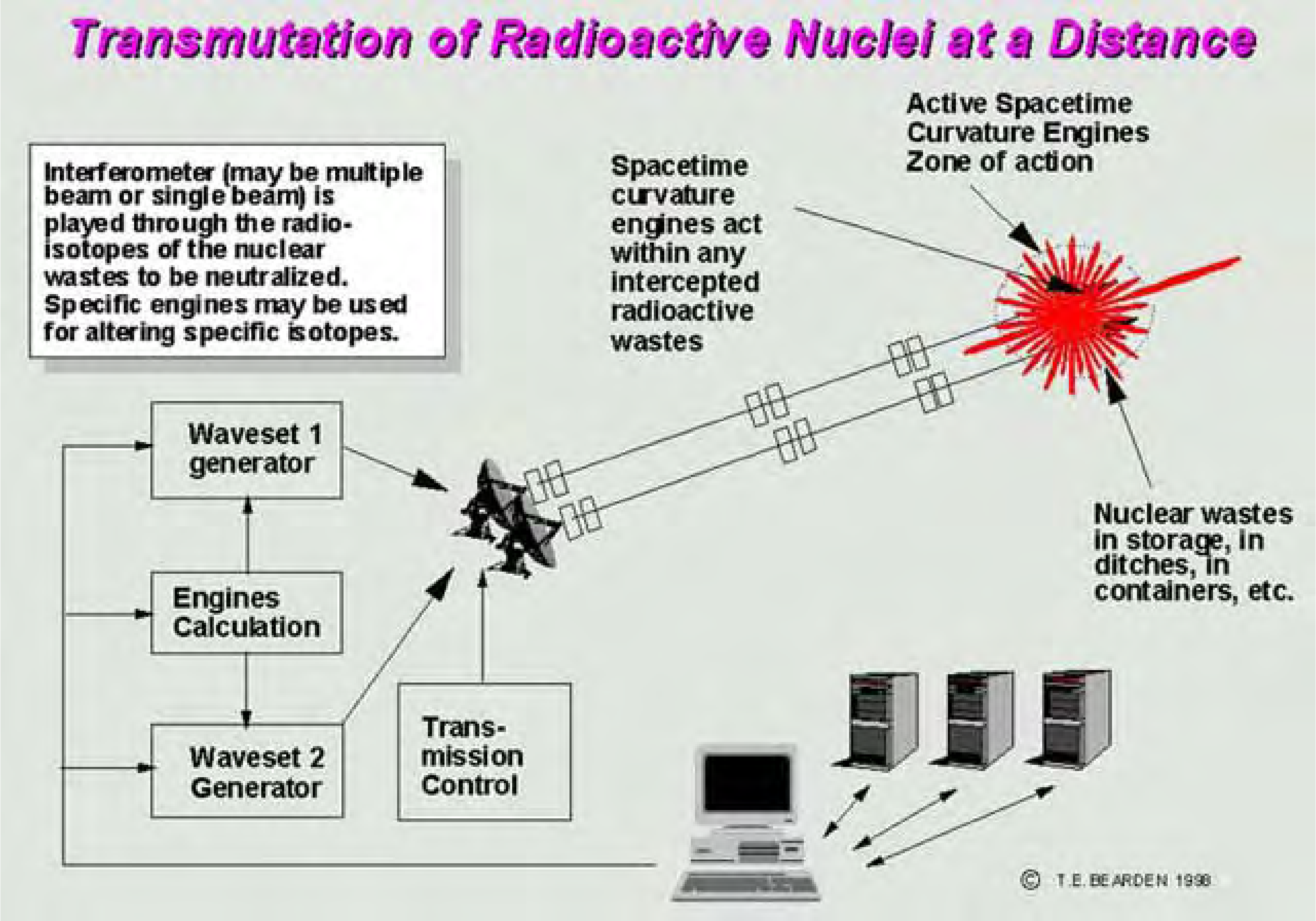

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Spacetime curvature engines act within any intercepted radioactive wastes

Active Spacetime Curvature Engines Zone of action
Nuclear wastes in storage, in ditches, in containers, etc.
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control

© T.E. BEARDEN 1998

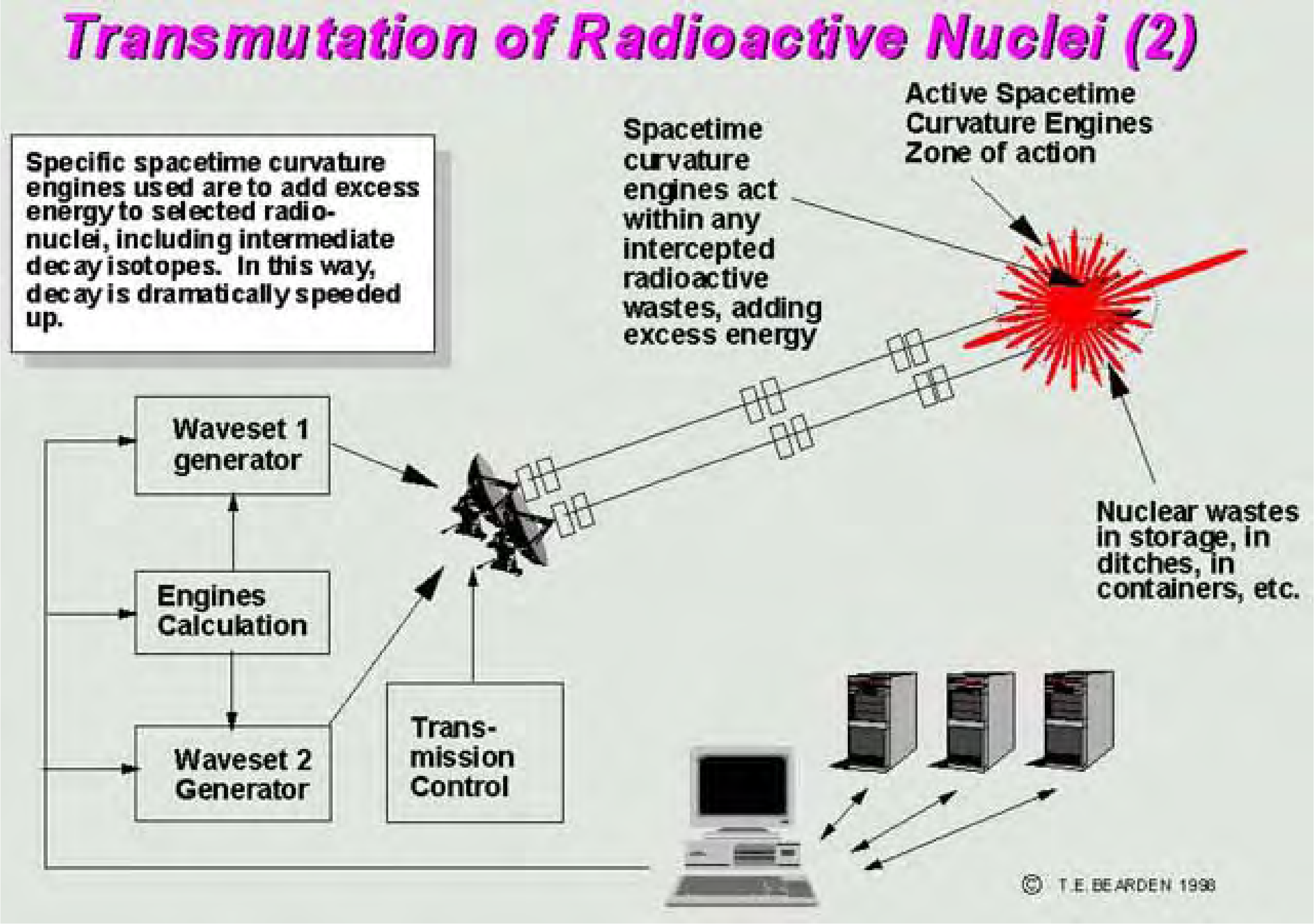

Transmutation of Radioactive Nuclei (2)

Specific spacetime curvature engines used are to add excess energy to selected radionuclei, including intermediate decay isotopes. In this way, decay is dramatically speeded up.

Spacetime curvature engines act within any intercepted radioactive wastes, adding excess energy

Active Spacetime Curvature Engines Zone of action

Nuclear wastes in storage, in ditches, in containers, etc.

Waveset 1 generator

Engines Calculation

Waveset 2 Generator

Transmission Control

© T.E. BEARDEN 1998

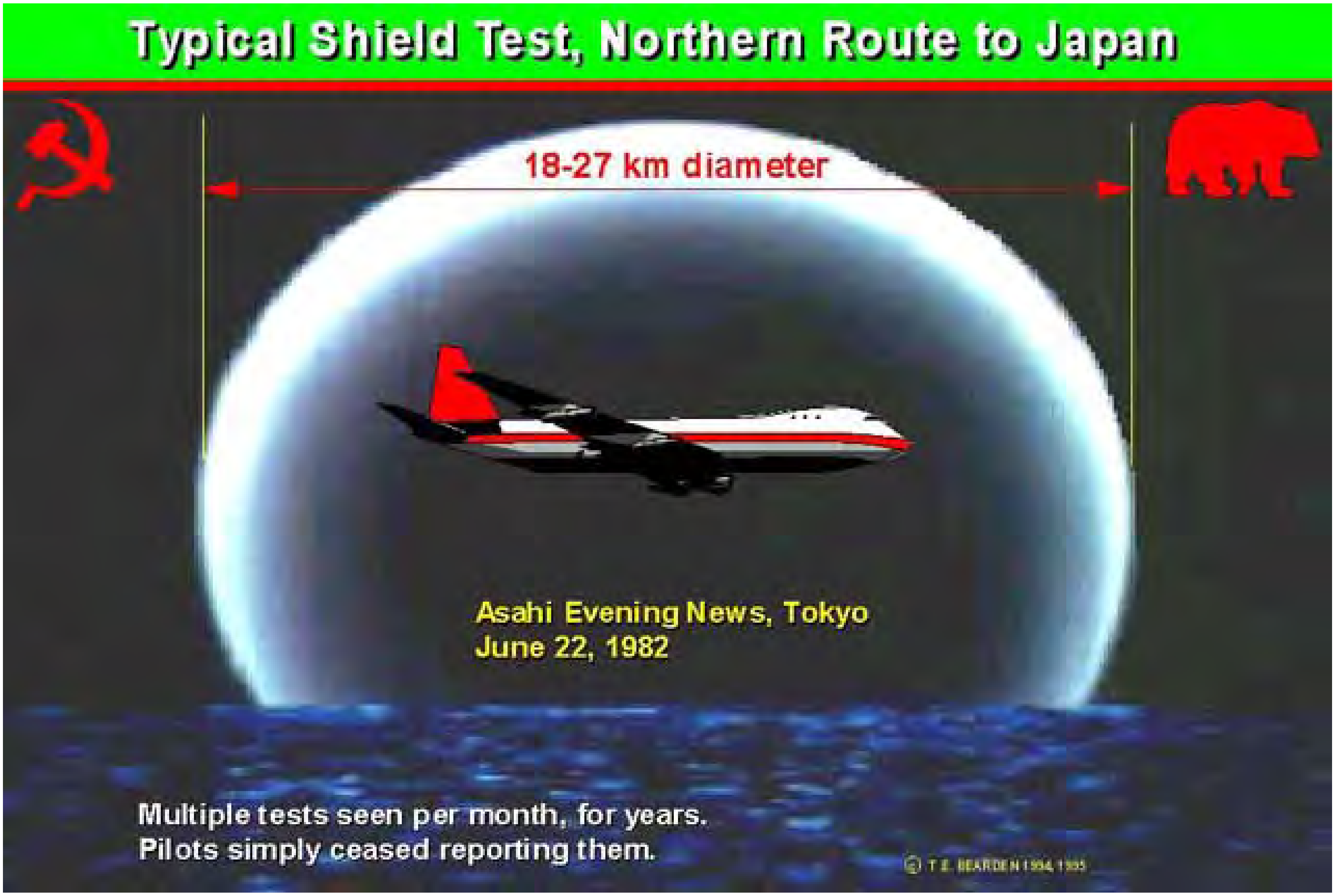
Typical Shield Test, Northern Route to Japan
18-27 km diameter
Asahi Evening News, Tokyo
June 22, 1982
Multiple tests seen per month, for years.
Pilots simply ceased reporting them.
© T. E. BEARDEN 1994, 1995

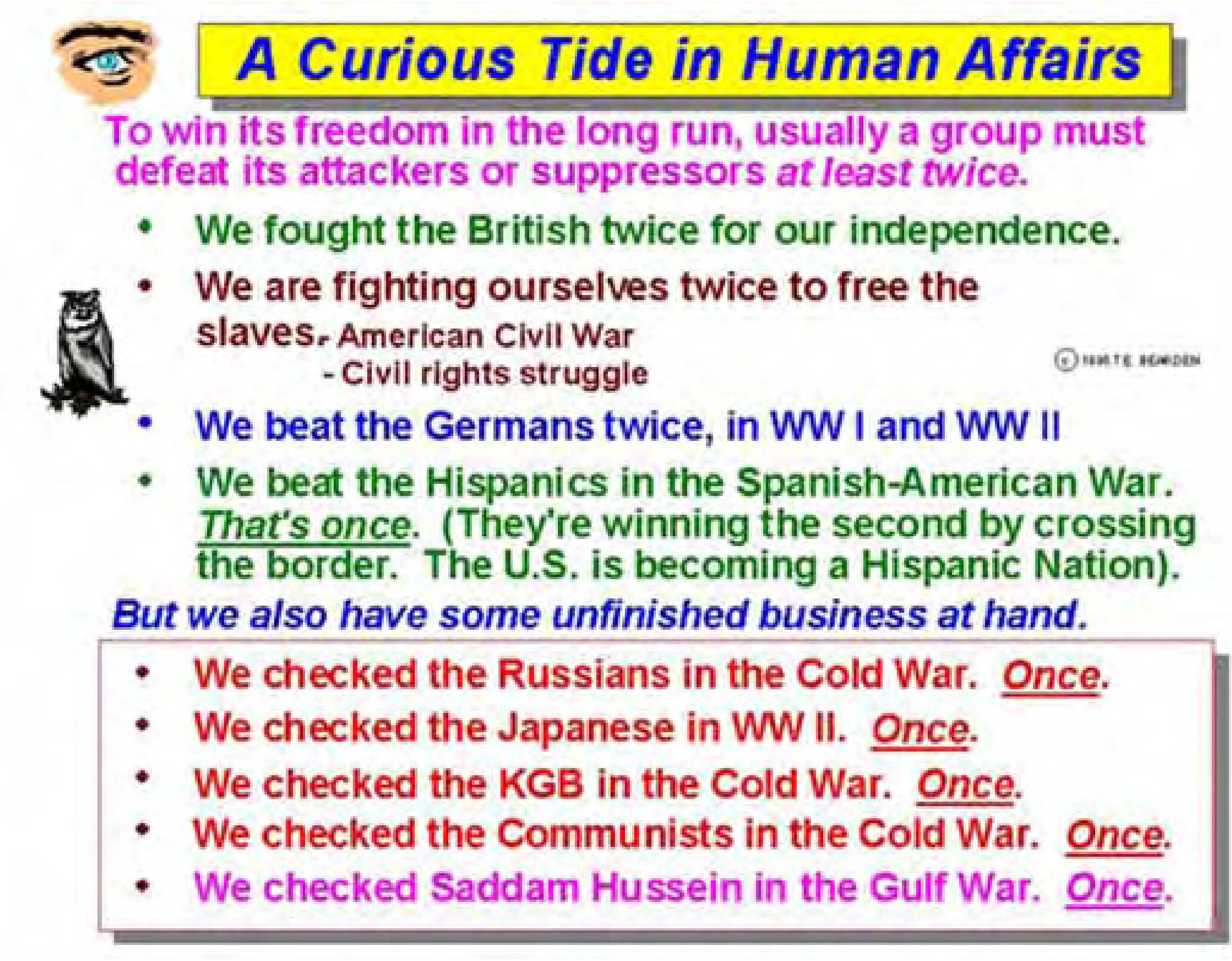

A Curious Tide in Human Affairs

To win its freedom in the long run, usually a group must defeat its attackers or suppressors *at least twice*.

- We fought the British twice for our independence.
- We are fighting ourselves twice to free the slaves. - American Civil War
  - Civil rights struggle
- We beat the Germans twice, in WW I and WW II
- We beat the Hispanics in the Spanish-American War. *That's once*. (They're winning the second by crossing the border. The U.S. is becoming a Hispanic Nation).

But we also have some unfinished business at hand.

- We checked the Russians in the Cold War. *Once*.
- We checked the Japanese in WW II. *Once*.
- We checked the KGB in the Cold War. *Once*.
- We checked the Communists in the Cold War. *Once*.
- We checked Saddam Hussein in the Gulf War. *Once*.

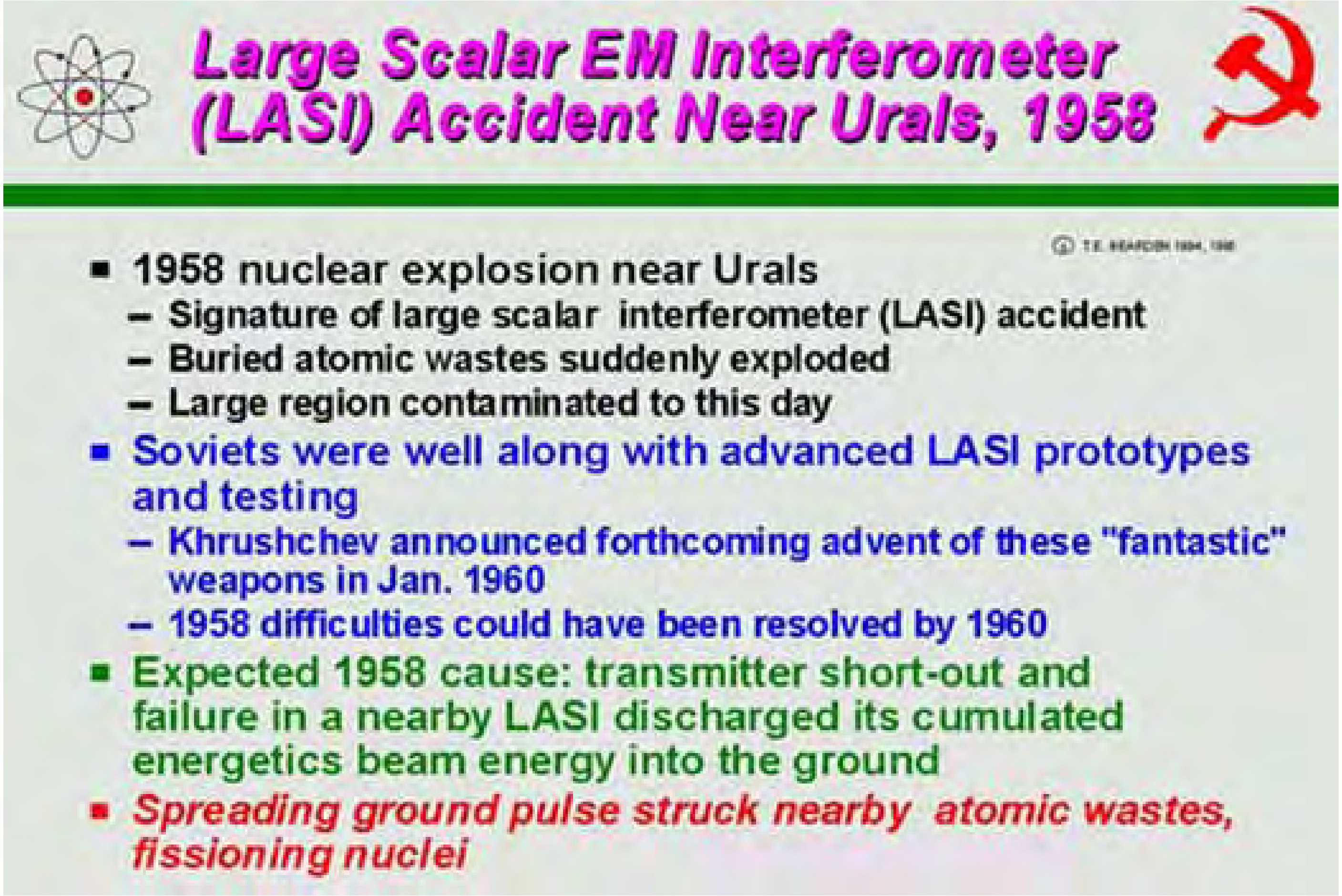

Large Scalar EM Interferometer (LASI) Accident Near Urals, 1958

- 1958 nuclear explosion near Urals
  - Signature of large scalar interferometer (LASI) accident
  - Buried atomic wastes suddenly exploded
  - Large region contaminated to this day
- Soviets were well along with advanced LASI prototypes and testing
  - Khrushchev announced forthcoming advent of these "fantastic" weapons in Jan. 1960
  - 1958 difficulties could have been resolved by 1960
- Expected 1958 cause: transmitter short-out and failure in a nearby LASI discharged its cumulated energetics beam energy into the ground
- *Spreading ground pulse struck nearby atomic wastes, fissioning nuclei*

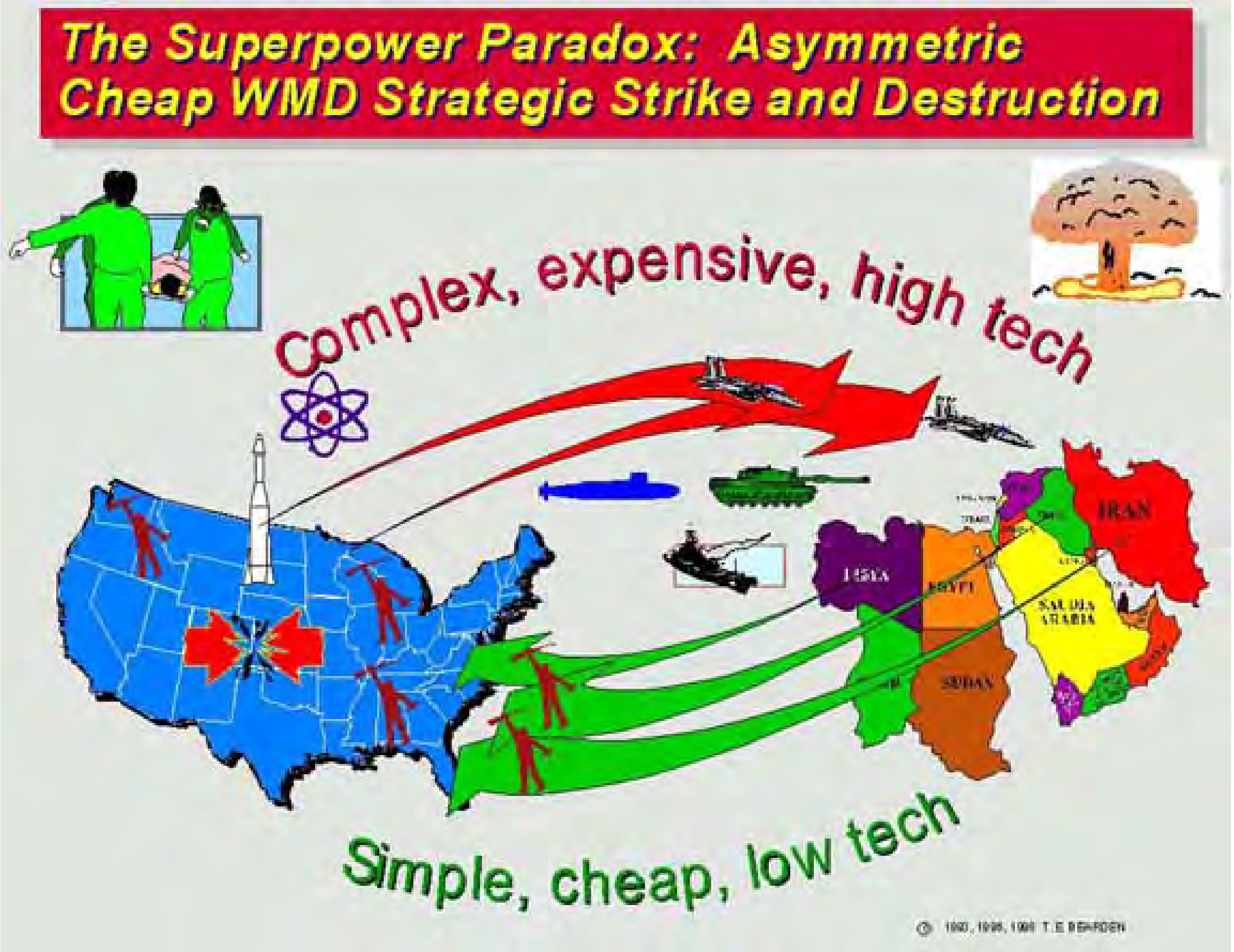

The Superpower Paradox: Asymmetric Cheap WMD Strategic Strike and Destruction

Complex, expensive, high tech

Simple, cheap, low tech

LIBYA
EGYPT
SUDAN
SAUDI ARABIA
IRAN

© 1990, 1998, 1999 T. E. BEARDEN

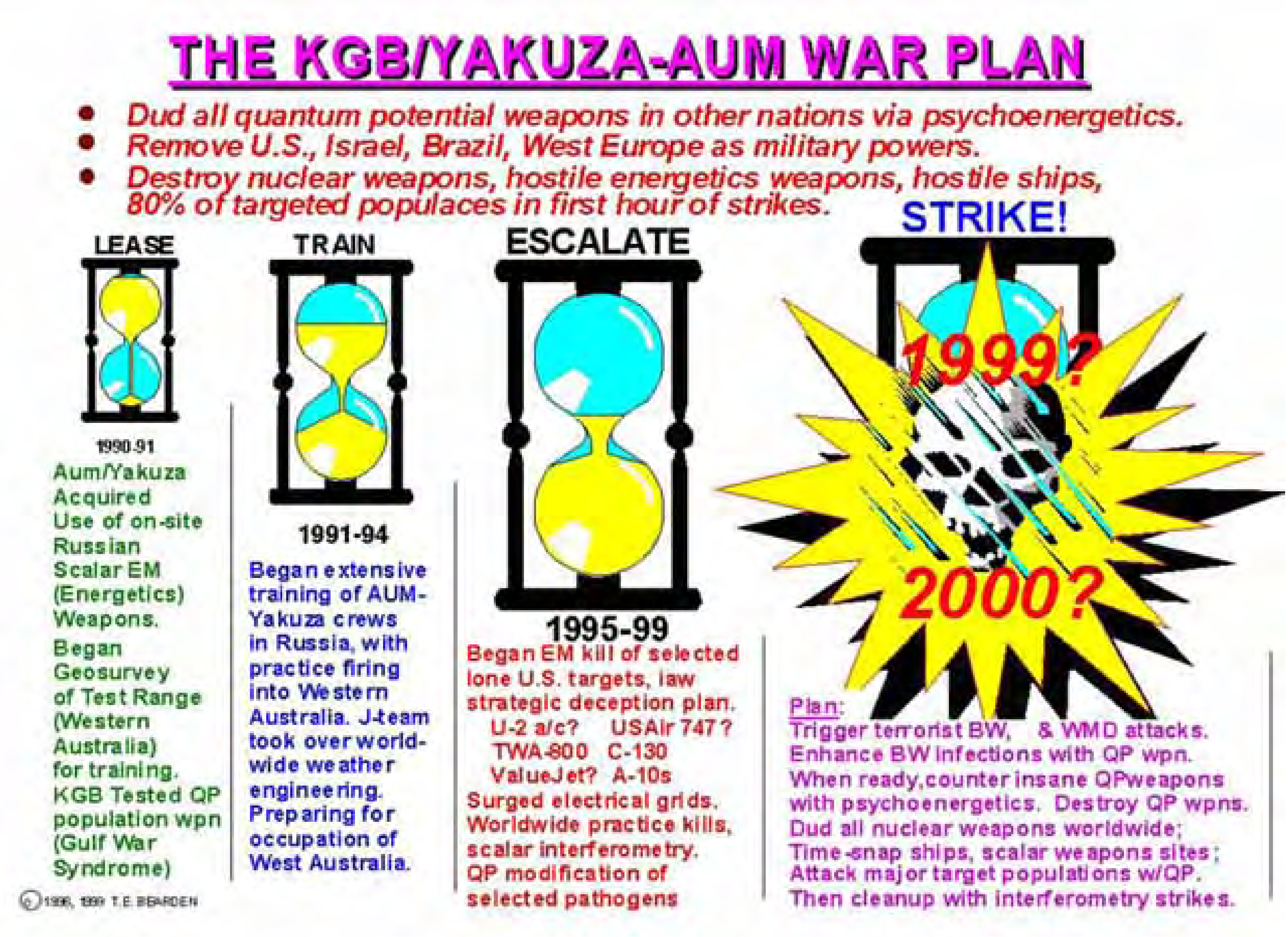

THE KGB/YAKUZA-AUM WAR PLAN

- *Dud all quantum potential weapons in other nations via psychoenergetics.*
- *Remove U.S., Israel, Brazil, West Europe as military powers.*
- *Destroy nuclear weapons, hostile energetics weapons, hostile ships, 80% of targeted populaces in first hour of strikes.*

LEASE

1990-91

Aum/Yakuza Acquired Use of on-site Russian Scalar EM (Energetics) Weapons.

Began Geosurvey of Test Range (Western Australia) for training.

KGB Tested QP population wpn (Gulf War Syndrome)

TRAIN

1991-94

Began extensive training of AUM-Yakuza crews in Russia, with practice firing into Western Australia. J-team took over world-wide weather engineering. Preparing for occupation of West Australia.

ESCALATE

1995-99

Began EM kill of selected lone U.S. targets, iaw strategic deception plan.

U-2 a/c? USAir 747?
TWA-800 C-130
ValueJet? A-10s

Surged electrical grids.
Worldwide practice kills, scalar interferometry.
QP modification of selected pathogens

STRIKE!

1999?

2000?

Plan:
Trigger terrorist BW, & WMD attacks.
Enhance BW infections with QP wpn.
When ready, counter insane QPweapons with psychoenergetics. Destroy QP wpns.
Dud all nuclear weapons worldwide;
Time-snap ships, scalar weapons sites;
Attack major target populations w/QP.
Then cleanup with interferometry strikes.

©1996, 1999 T.E. BEARDEN

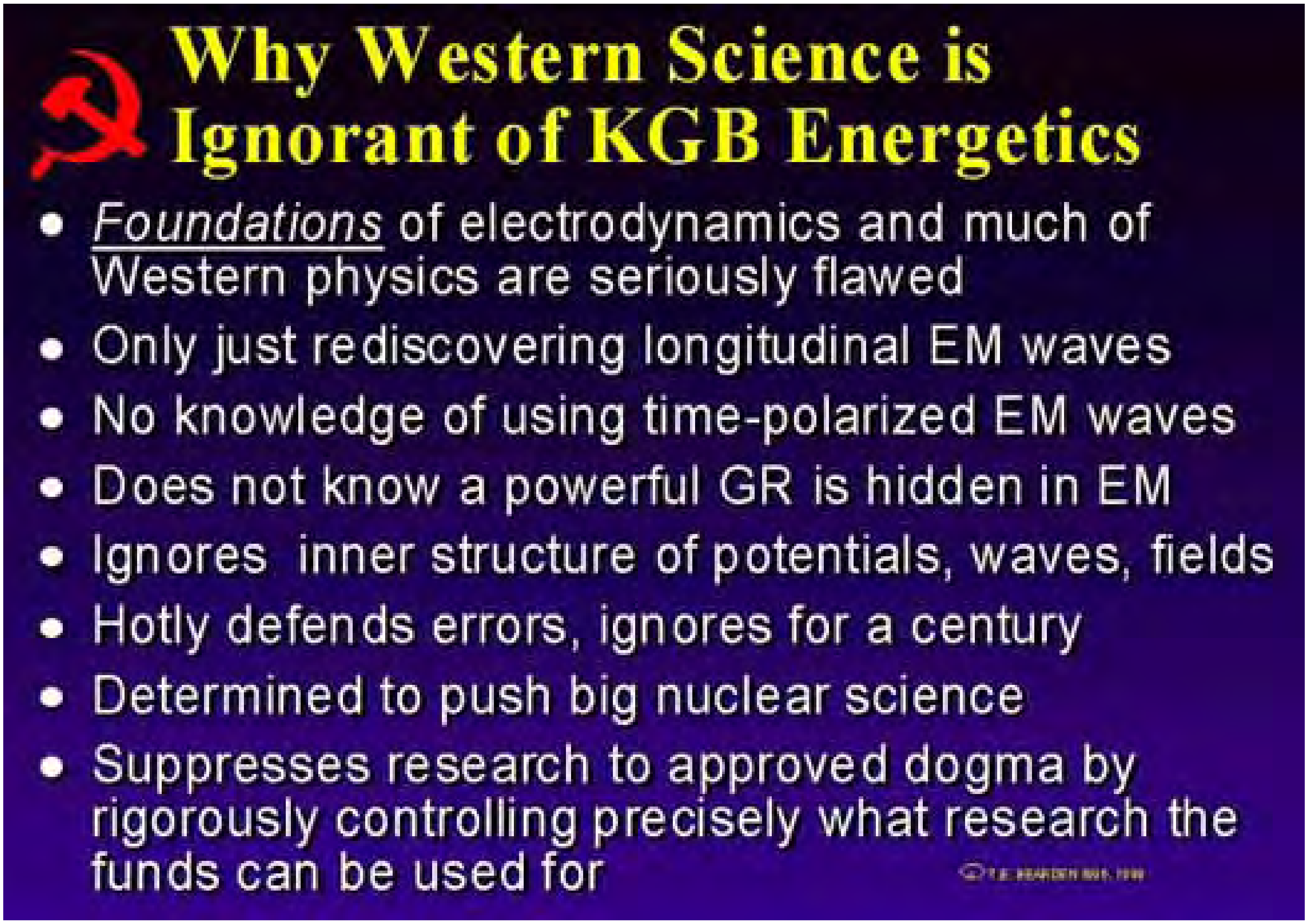

Why Western Science is Ignorant of KGB Energetics

- *Foundations* of electrodynamics and much of Western physics are seriously flawed
- Only just rediscovering longitudinal EM waves
- No knowledge of using time-polarized EM waves
- Does not know a powerful GR is hidden in EM
- Ignores inner structure of potentials, waves, fields
- Hotly defends errors, ignores for a century
- Determined to push big nuclear science
- Suppresses research to approved dogma by rigorously controlling precisely what research the funds can be used for

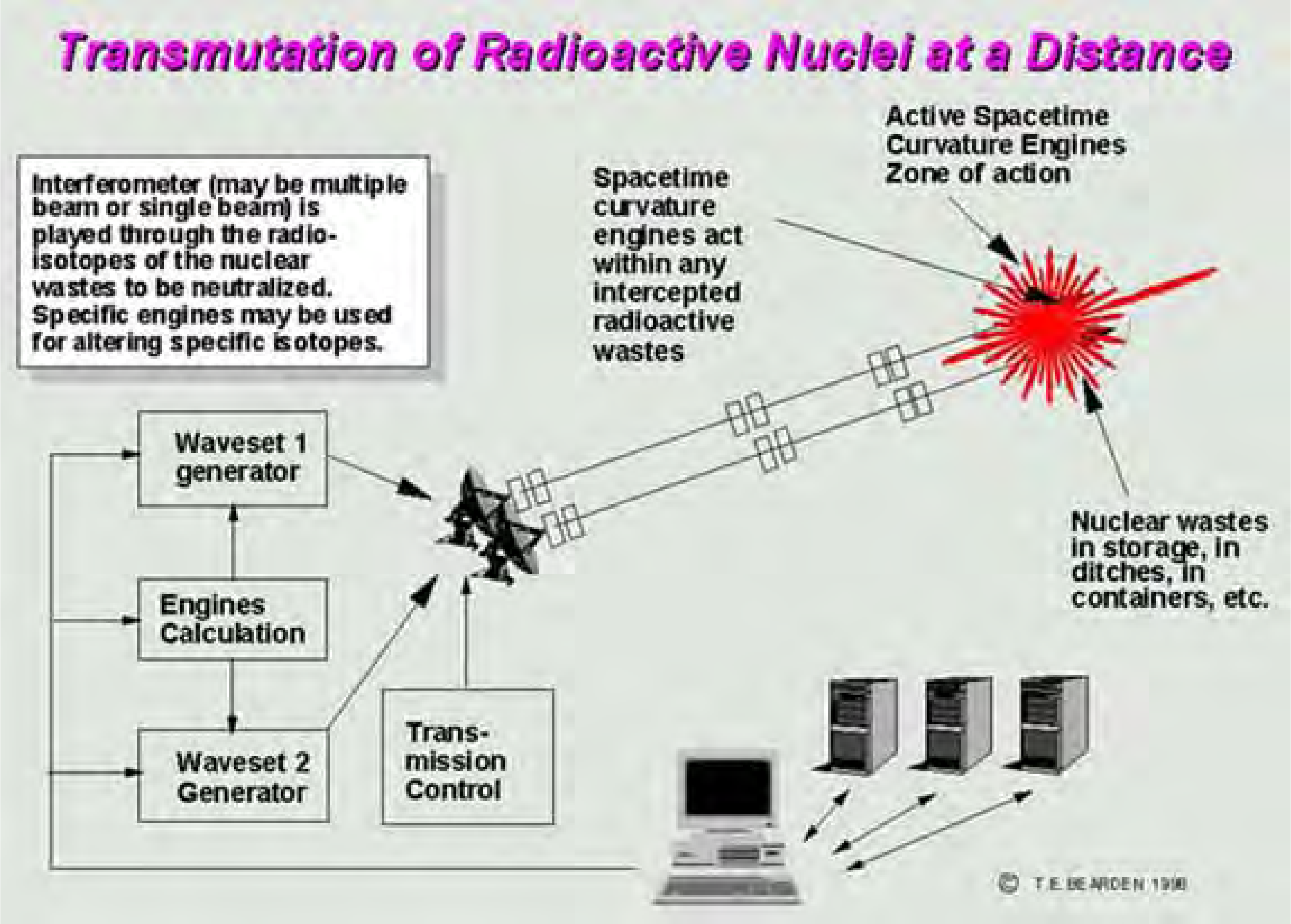

Transmutation of Radioactive Nuclei at a Distance

Interferometer (may be multiple beam or single beam) is played through the radioisotopes of the nuclear wastes to be neutralized. Specific engines may be used for altering specific isotopes.

Active Spacetime Curvature Engines Zone of action
Spacetime curvature engines act within any intercepted radioactive wastes
Waveset 1 generator
Engines Calculation
Waveset 2 Generator
Trans-mission Control
Nuclear wastes in storage, in ditches, in containers, etc.

© T.E. BEARDEN 1990

THE YAKUZA

(JAPANESE MAFIA)

To do business with Japan is to do business with the Yakuza

The Last Samurai, Defending Bushido

- **3 Major Yakuza Groups, In Coalition.**
- **Penetrate Japanese Government.**
- **Open Offices, Signs on Street.**
- **Public Romanticizes Yakuza as Last Japanese Samurai**
- **In Large High-Tech Corporations.**
- **In Control of Japanese Banks.**
- **Japanese Real-Estate Rip-off.**
- **$500 - 1,000 Billion per Year Income.**
- **Rapidly Spreading in U.S.**
- **Uniting Mafia Groups Worldwide.**
- **Coalition with Aum Shinrikyo.**
- **Leased Russian Energetics Weapons.**
- **Crews On-site in Russia, Trained.**
- **Practice Firings into Western Australia for Three Years. (Dozens of Shots per Month)**
- **Now in Escalation Phase, Killing U.S. Targets.**
- **All-out Strike on U.S. ready 1999-2000.**
- **Now held in check by KGB, dead-man fuzed by small nation friendly to U. S.**

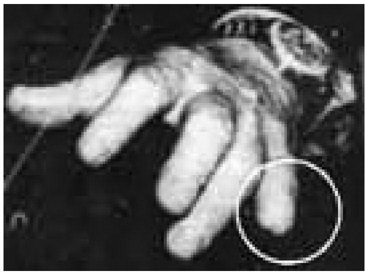

© 1995, 1996, 2001 T.E. BEARDEN

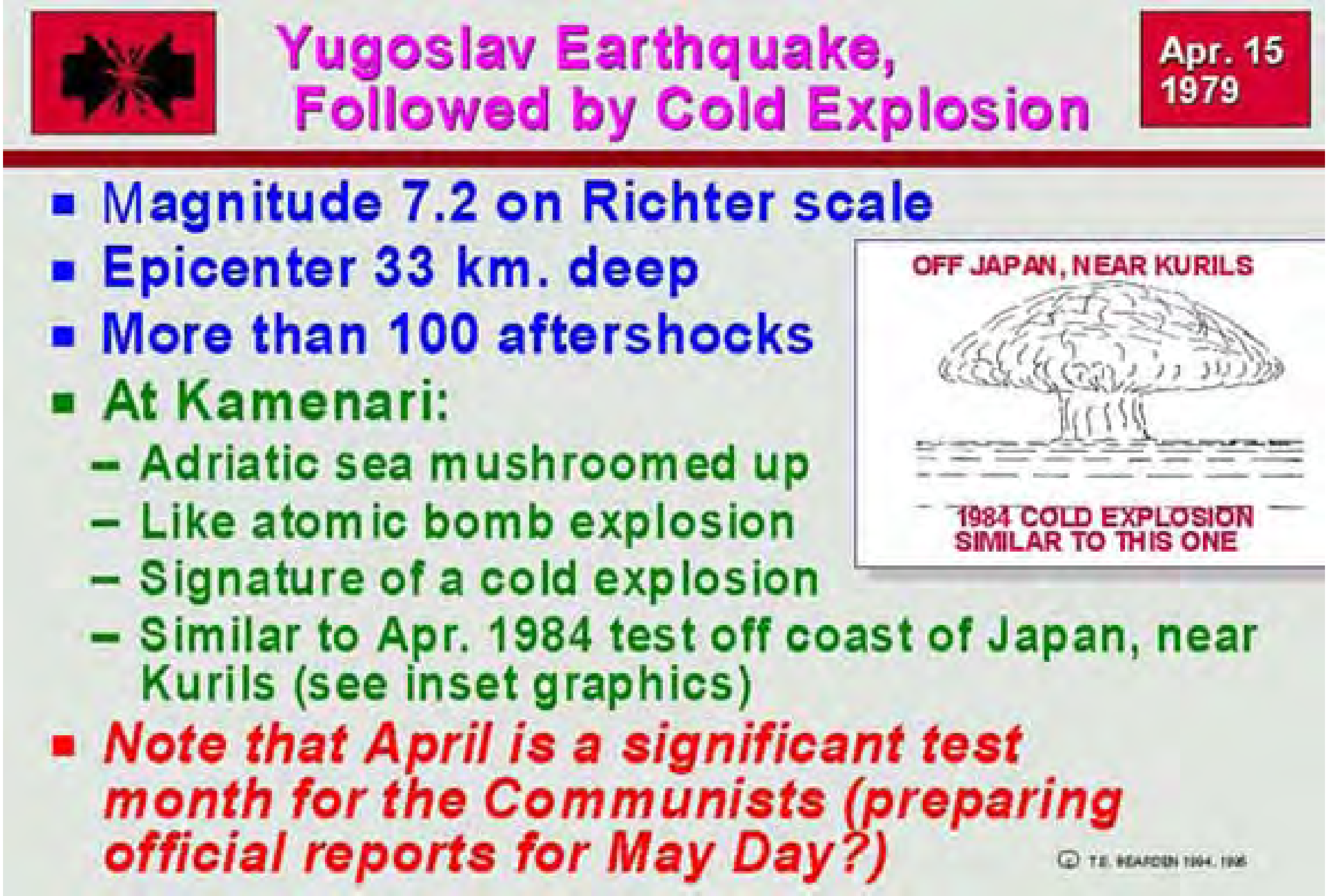

Yugoslav Earthquake, Followed by Cold Explosion

Apr. 15
1979

- Magnitude 7.2 on Richter scale
- Epicenter 33 km. deep
- More than 100 aftershocks
- At Kamenari:
  - Adriatic sea mushroomed up
  - Like atomic bomb explosion
  - Signature of a cold explosion
  - Similar to Apr. 1984 test off coast of Japan, near Kurils (see inset graphics)
- *Note that April is a significant test month for the Communists (preparing official reports for May Day?)*

OFF JAPAN, NEAR KURILS

1984 COLD EXPLOSION
SIMILAR TO THIS ONE

© T.E. BEARDEN 1994, 1996

Briefing to
Request for Information Symposium

14 February 2000

Air Force Research Laboratory

Directed Energy Directorate
AFRL/DE

Kirtland AFB, New Mexico

Colonel Doug Beason
Deputy Director

“VICTORY SMILES UPON THOSE WHO ANTICIPATE THE CHANGES IN THE CHARACTER OF WAR”

Guilio Douhet

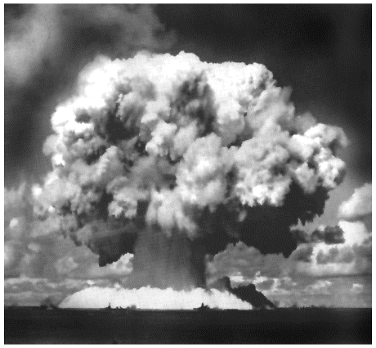

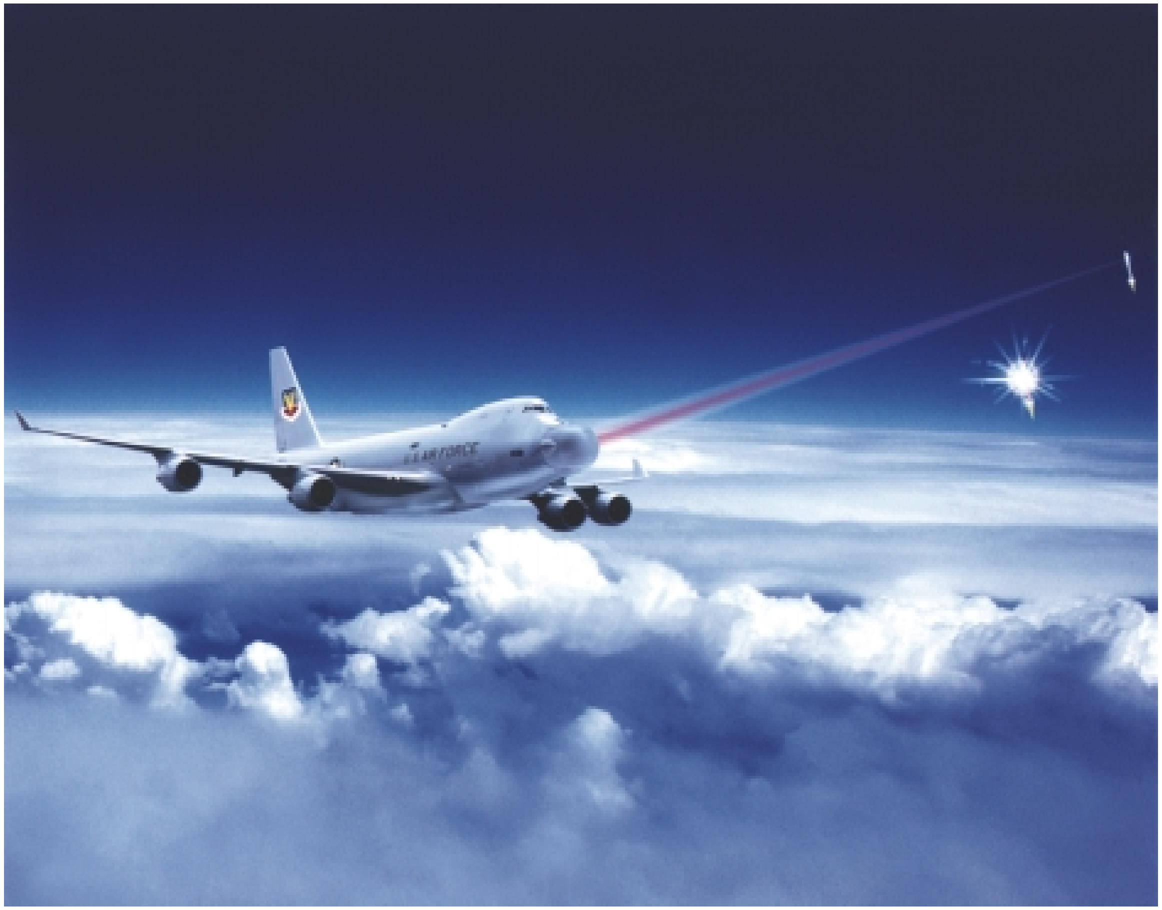
U.S. AIR FORCE

Directed Energy Directorate

A paradigm shift in weapon system technology for the 21st Century.

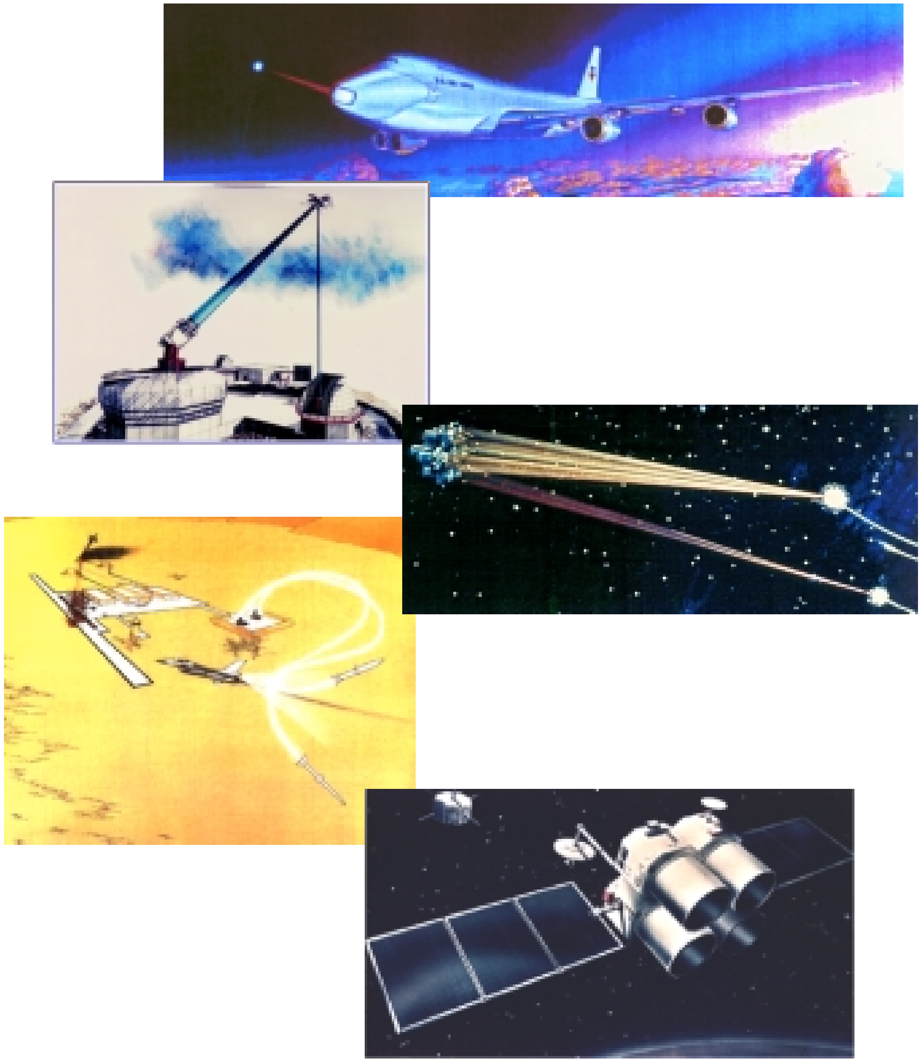

Our Mission

- ***To support*** user needs for DEW applications and to address mission area deficiencies
- ***To exploit*** the relevance of Directed Energy technology to Air Force and DoD needs
- ***To foster*** user awareness of the potential of Directed Energy for Air Force and DoD applications
- ***To avoid*** technological surprise
- ***To explore*** DE technology avenues that offer high payoff to directed energy capabilities and applications

To do what has never been done before!

Directed Energy Directorate

Our Vision of the Future:
Directed Energy will Dominate the Battlespace of the 21st Century

Strike

We will strike deep in the enemy's territory at the speed of light, with little or no collateral damage or loss of life, crippling his ability to wage aggression.

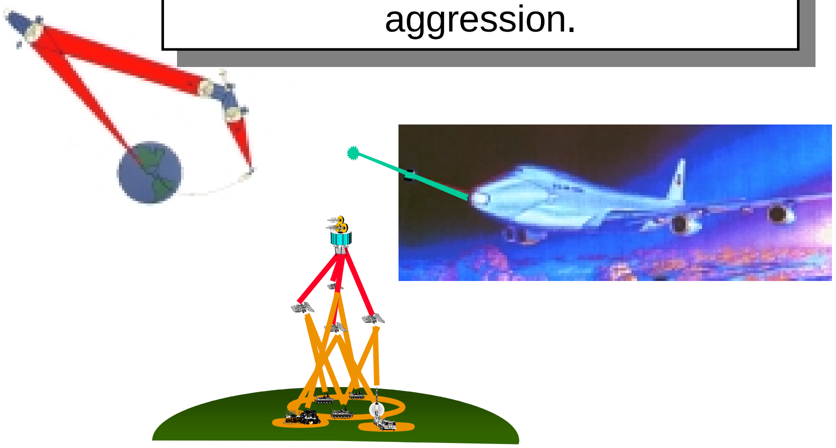

Protect

Our high value and limited military assets will be protected by invisible shields of directed energy.

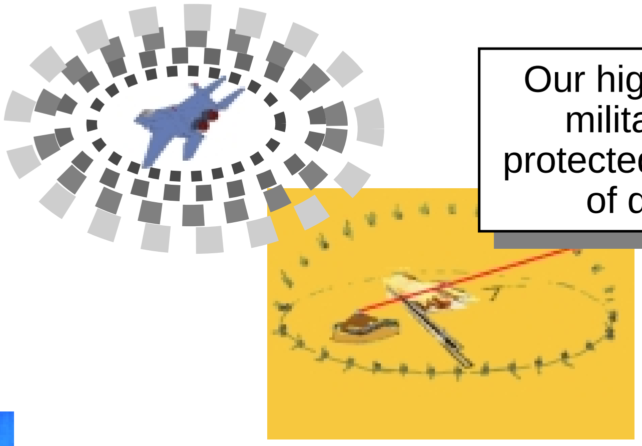

Graduated Deterrence

We will provide our warfighters with weapons that provide a wide range of graduated force for every military contingency.

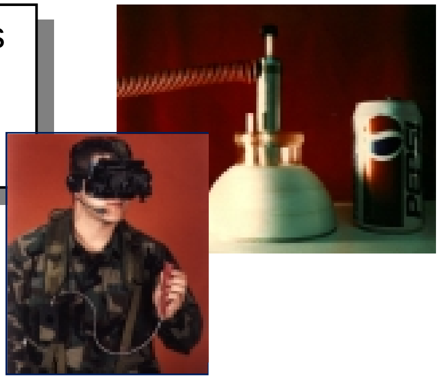

Awareness

Our battleflield awareness will be greatly enhanced with directed energy tools capable of detection and identification

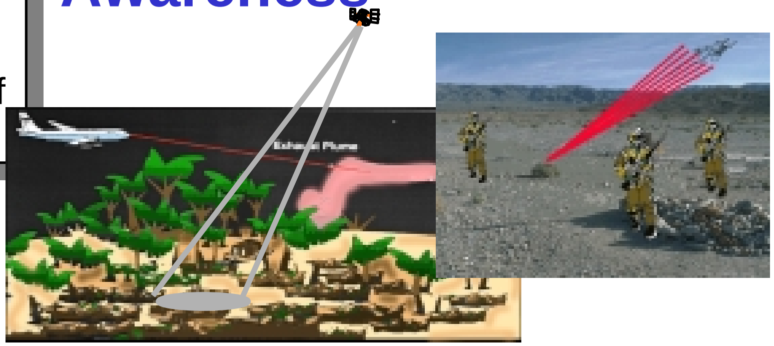

Security & safety in isolated canyons

Access to 50,000 acres for testing

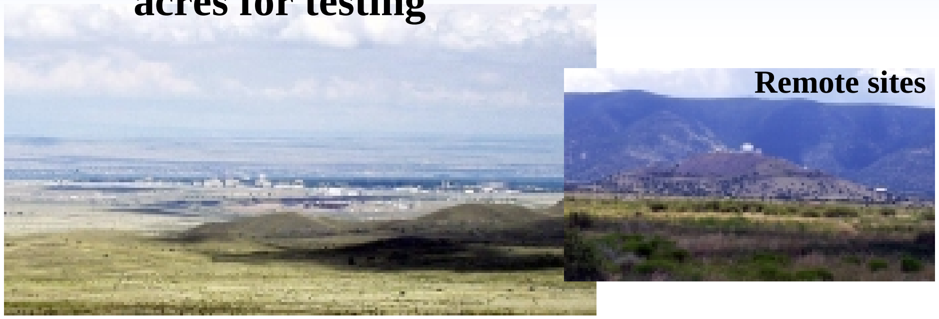
Remote sites

Kirtland AFB

Unique & ideally suited for the nation's directed energy technology development

Ideal weather
"thin" atmosphere

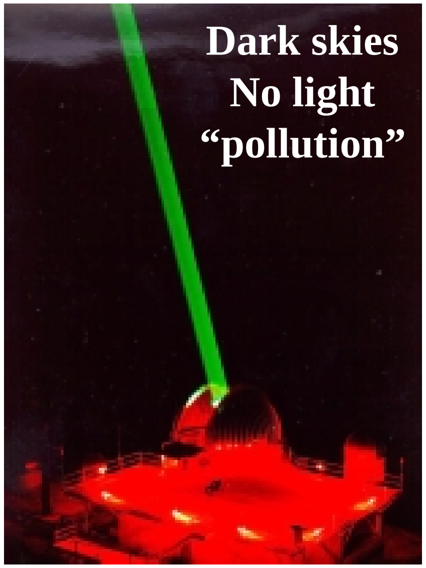
Dark skies
No light
"pollution"

Sandia National Laboratories

Many diverse, high tech tenants

International airport & metro area

Directed Energy Directorate
Remote Sites

- Maui, HI (leased land)
  - Relay Mirror Experiment Facility
  - Maui High Performance Computer Center
  - Maui Space Surveillance Complex

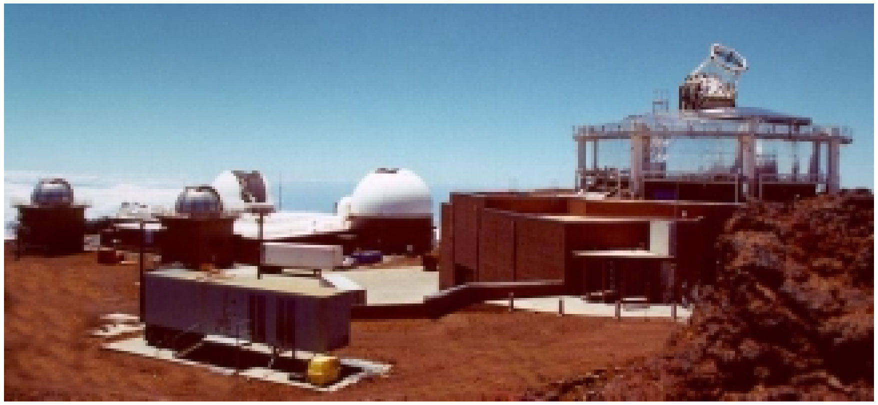

- North Oscura Peak, NM
  - Telescope site at White Sands Missile Range
  - Supports Space Imaging R&D (ABL Technology)

Lasers and Microwaves

Synergistic, Complementary and Capable

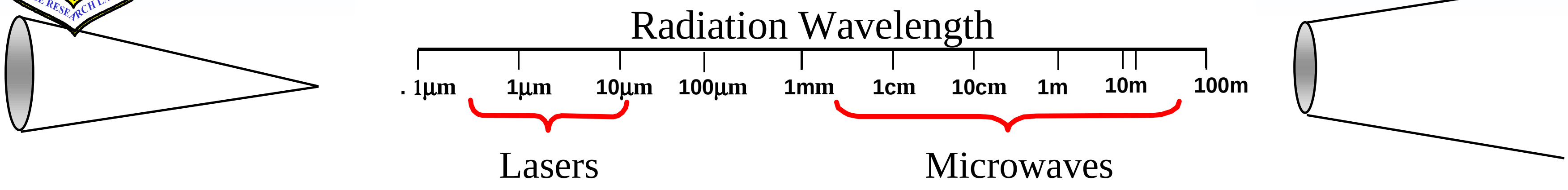
Radiation Wavelength
. 1µm
1µm
10µm
100µm
1mm
1cm
10cm
1m
10m
100m
Lasers
Microwaves

High Energy Lasers	High Power Microwave s
Small spot: surgical attack	**Flood target area: multiple simultaneous targets**
Affects targets from outside: structural destruction	**Affects components inside: electronic disruption**
Weather sensitive: clouds, dust, molecular absorption	**Weather insensitive: sees through clouds and dust.**
Modest apertures: glass mirrors, coatings	**Big apertures: metal antennas, ground planes**

But on the other hand,... they both

- **Travel at the speed of light to the target**
- **Capable of graduated effects from deny, disrupt, degrade, and destroy.**
- **Minimum collateral damage**

Directed Energy Directorate

Broad Technology Investments

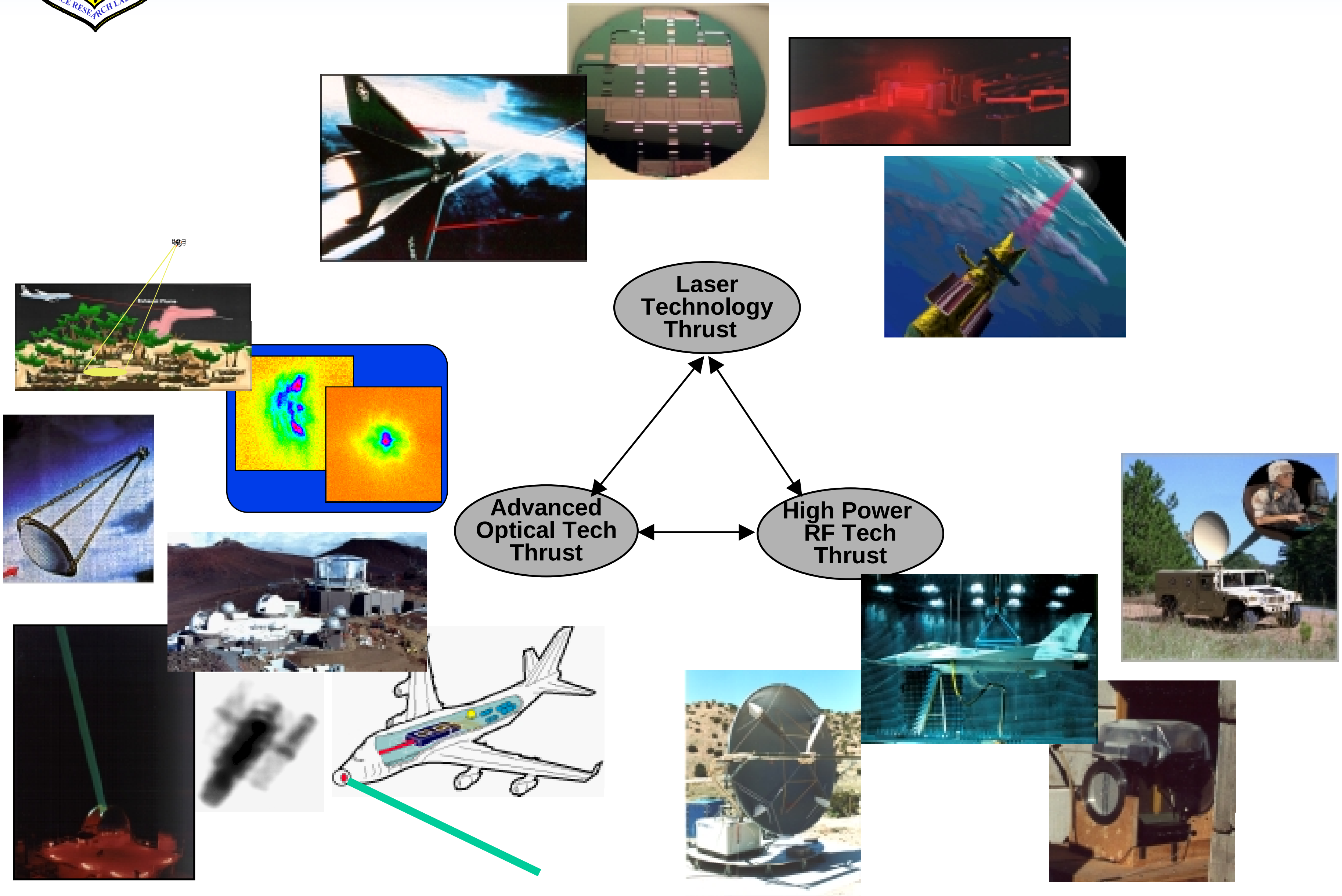

Laser Technology Thrust

Advanced Optical Tech Thrust

High Power RF Tech Thrust

Directed Energy Directorate

12 Major DE Technology Programs

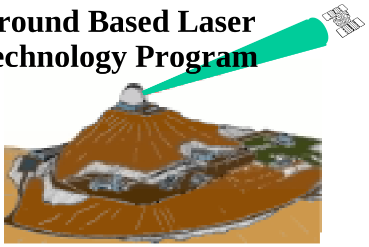

Ground Based Laser Technology Program

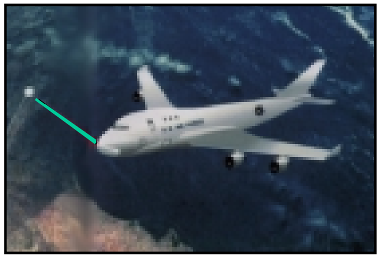

Airborne Laser Technology Program

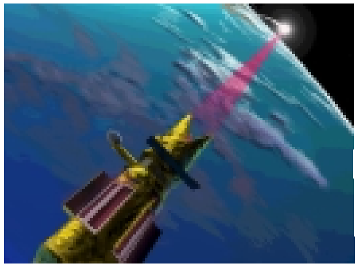

High Power Laser Program

Scalable Multi-Wavelength Laser Program

High Efficiency Electric Laser Program

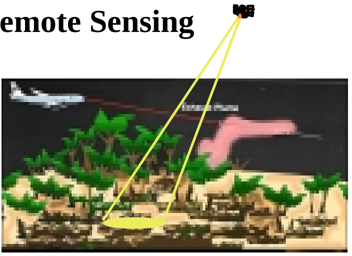

Remote Sensing

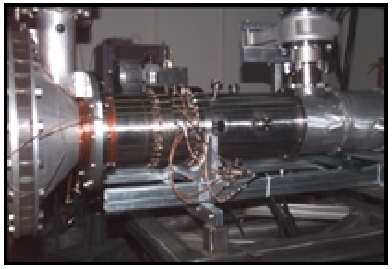

HPM Sources Program

HPM Pulsed Power Program

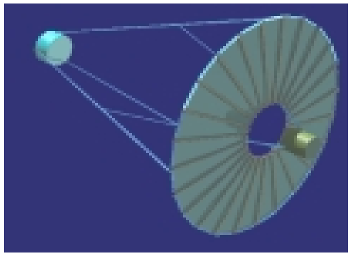

Large Optics

Space Awareness Program

HPM Applications Program

HPM Effects Program

AFRL
Directed Energy Directorate

Mission

Develop, integrate and transition science & technology for Directed Energy to include high power microwaves, lasers, adaptive optics, imaging and its effects to assure the preeminence of US in air & space

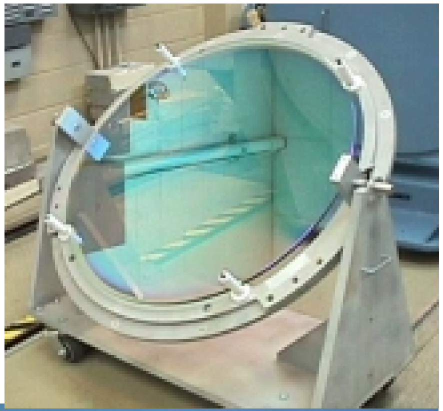

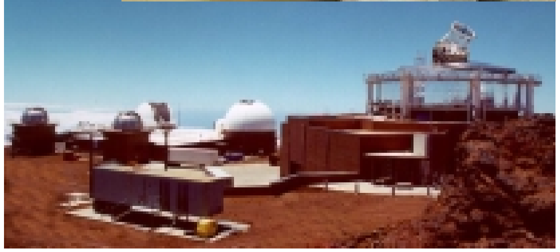

Major Technology Programs

Ground Based Laser
Airborne Laser
High Power Lasers
Multi-Wavelength Laser
Electric Lasers
Large Optics

Space Awareness
HPM Applications
HPM Effects
HPM Pulsed Power
HPM Sources
Remote Sensing

AFRL/DE - factoids

- **952 people (Civ, Mil, Contr)**
- **$109M FY00 budget**
- **4,325 acres**
- **757,000 sf of space**
- **Located throughout Kirtland**
- **Sites at Maui & WSMR**

Air Force Research Laboratory
Directed Energy Directorate

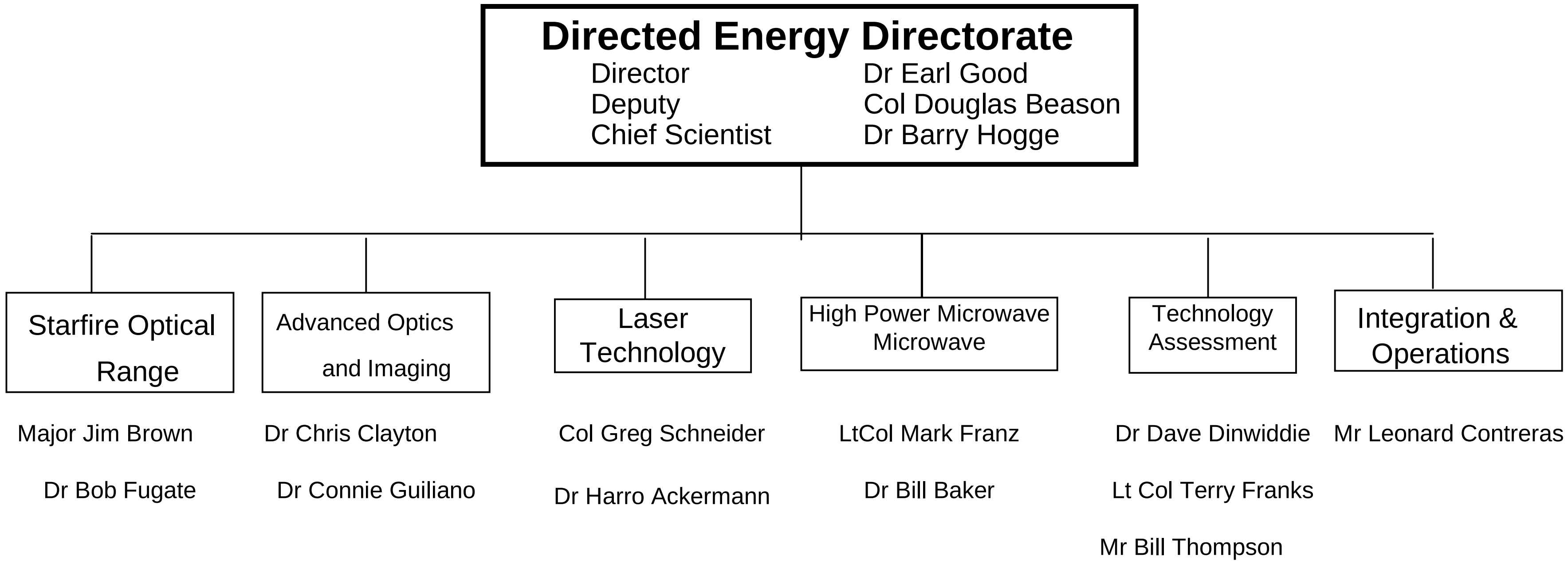

Directed Energy Directorate

Director — Dr Earl Good
Deputy — Col Douglas Beason
Chief Scientist — Dr Barry Hogge

Starfire Optical Range
Major Jim Brown
Dr Bob Fugate

Advanced Optics and Imaging
Dr Chris Clayton
Dr Connie Guiliano

Laser Technology
Col Greg Schneider
Dr Harro Ackermann

High Power Microwave Microwave
LtCol Mark Franz
Dr Bill Baker

Technology Assessment
Dr Dave Dinwiddie
Lt Col Terry Franks
Mr Bill Thompson

Integration & Operations
Mr Leonard Contreras

Directed Energy Directorate

Major Kirtland Facilities

Lasers, Optics, Imaging

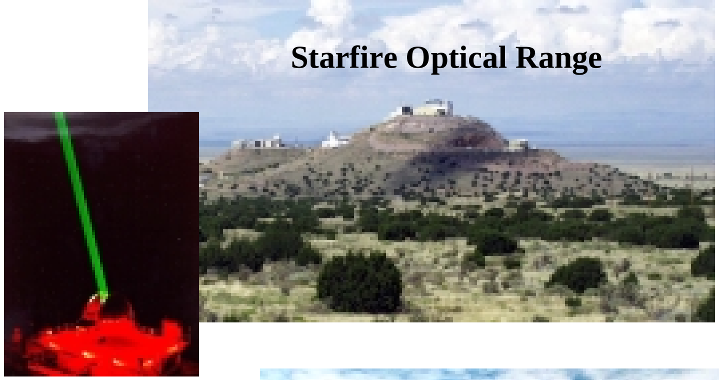
Starfire Optical Range

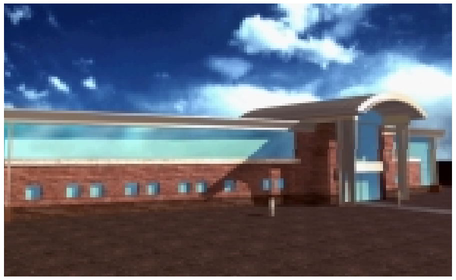
Advanced Laser Facility

Argus Facility

Optics Development
& Beam Control Laboratory

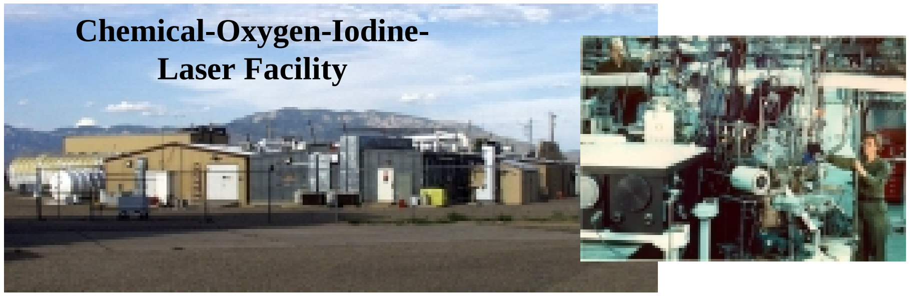
Chemical-Oxygen-Iodine-
Laser Facility

Optics Coating & Components Evaluation Laboratory

Laser Effects Test Facility

Directed Energy Directorate

Major Facilities - High Power Microwave

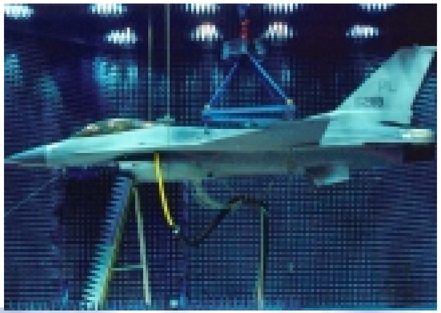

High Energy Microwave Laboratory

High Energy Research & Technology Facility

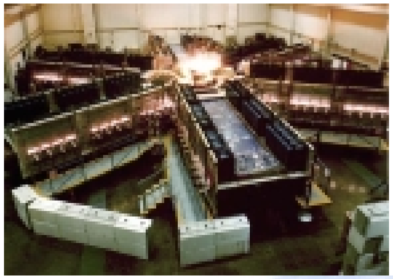

Plasma Research Laboratory

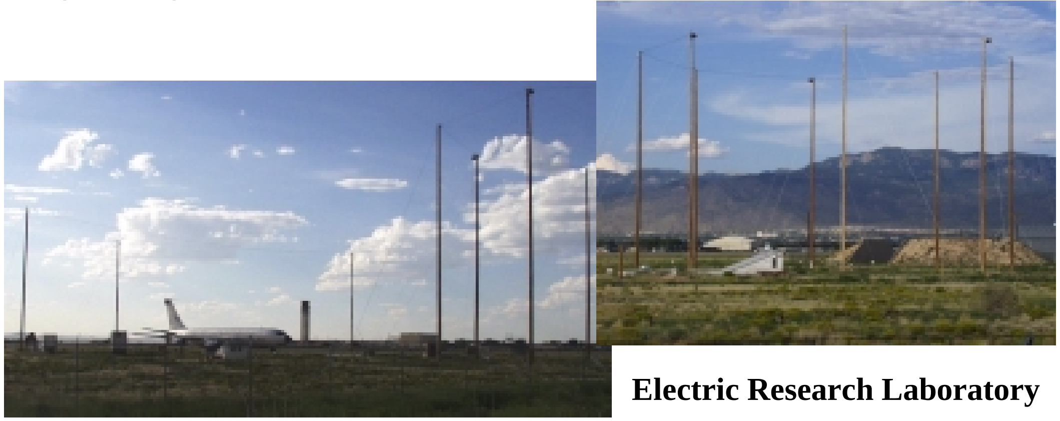

Electric Research Laboratory

Electromagnetic Energy/RF Effects Laboratory

Directed Energy Directorate

Manning

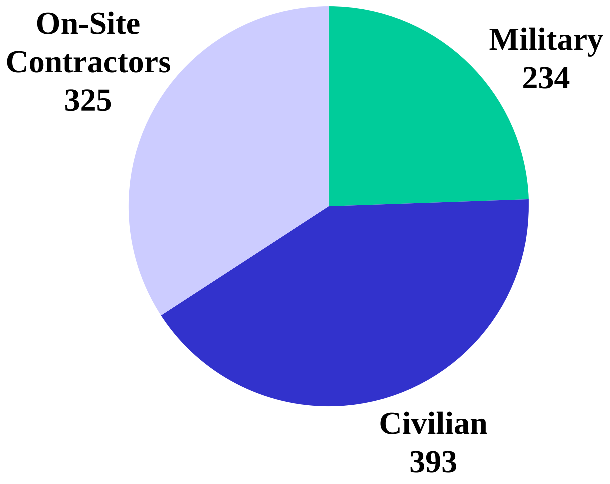

Assigned Personnel
952

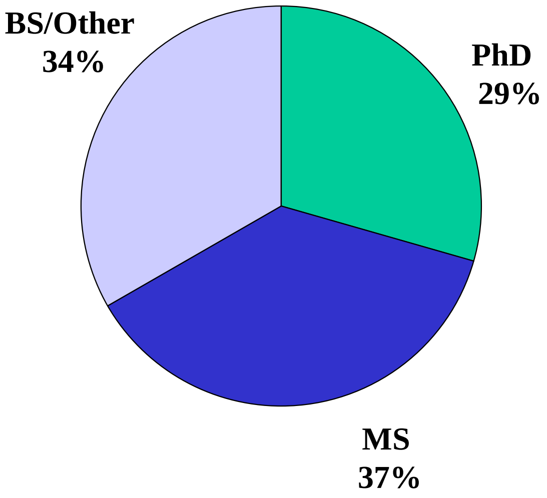

Government S&E Education Levels

Directed Energy Directorate

FY00 AFRL PB Resources

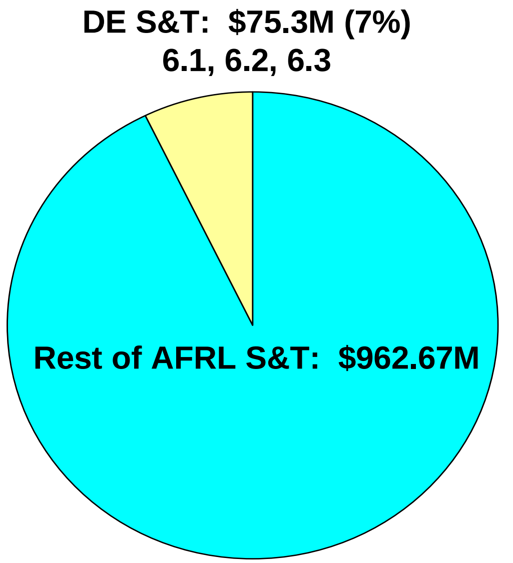

AFRL TOTAL S&T: $1,037.98M

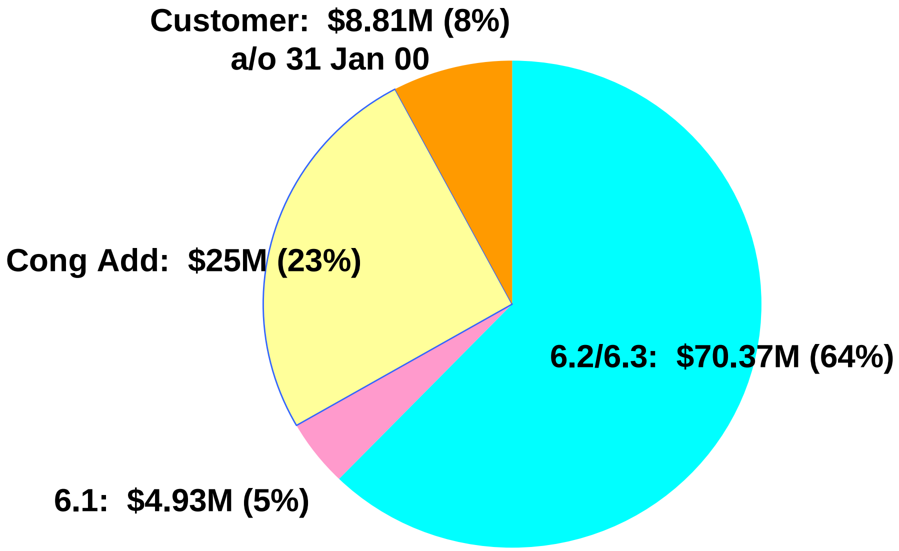

Directed Energy Total: $109.11M

Directed Energy Directorate

Potential Areas for Collaboration

- Electric Lasers
- Beam Control
- Large, Lightweight Optics
- Optical Coatings

- High Power RF Sources
- High Power RF Antennae
- Atmospheric Compensation
- Remote Sensing

For Information contact Mr Rich Garcia, AFRL/DEOB-PA (505)846-4583

AiMT Advances in Military Technology
Vol. 4, No. 2, December 2009

Directed Energy Weapons in Modern Battlefield

L. Palíšek[*]

Division VTÚPV Vyškov, VOP-026 Šternberk, s.p., Czech Republic

The manuscript was received on 25 June 2009 and was accepted after revision for publication on 21 September 2009.

Abstract:

The paper deals with electromagnetic threats caused due to Directed Energy Weapons (DEW). The area of interest is focused on HPM (High Power Microwaves) and UWB (Ultra Wide Bandwidth) as an important part of DEW area. Possible impact on electronic structures which can be part of military systems is considered in this paper. Some practical results achieved during experiments related to assessment of electronic equipment vulnerability is provided with overview of immunity of some equipment to HPM and UWB signals at the end of this study.

Keywords:

DEW, HPM, UWB, electromagnetic threats, NEC, vulnerability

1. Introduction

Directed Energy Weapons (DEW) with various principles seem to be more important in the future battlefield. There are several principles which are considered for use in the DEW area. Special optical technologies (lasers), sound technologies as well as microwave technologies are developed and can be used in special weapon systems to disturb military system behaviour or can be used for counter personnel weapons (for example Active Denial System working on frequency 94 GHz, see Fig. 1). DEW has been mainly considered as technologies suitable for less-lethal or even non-lethal technologies which have been subjects of interest within the past years.

Recent developed military systems consist of equipment which is more and more sophisticated and relying on electronics. Very high importance of electronics within the military systems is obvious from development of Network Enabled Capabilities (NEC) which is a key task of the Czech Armed Forces as well as within the NATO (NNEC – NATO Network Enabled Capability). The stage related to "Communication and information sharing" has been topical in recent years and characterised by development of communication networks. Common network infrastructure can be

---

[*] *Corresponding author: Division VTÚPV Vyškov, VOP-026 Šternberk, s.p., Olomoucká 1841/175, 785 01 Šternberk, Czech Republic, phone: +420 517 303 638, E-mail: l.palisek@vtupv.cz*

considered as a fundament of integrated NEC environment. Communication network and especially its reliability will have significant influence on mission success. Disturbance of these networks could be obviously critical.

Power electromagnetic fields can be generated by relevant power technologies and focused (directed) towards the target with electronics. Shortcuts like HPM (High Power Microwave) and UWB (Ultra Wide Bandwidth) are widely used in this area as well as HPEM (High Power Electromagnetics) and IEMI (Intentional Electromagnetic Interference). Examples of developed technologies which could be easily used like DEW against electronics are shown in Figs 2-6 (these technologies were offered or presented mainly during relevant conferences in past few years, radiated pulse power was hundreds of MW or higher in frequency range of microwaves).

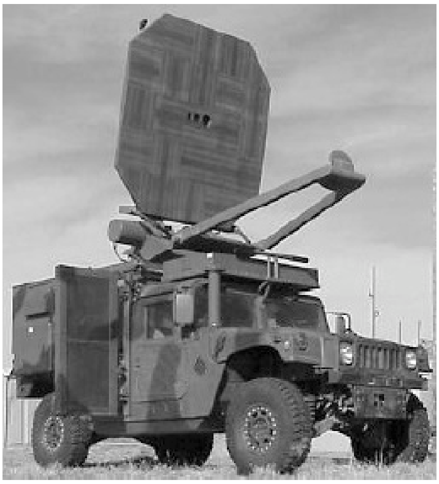

Fig. 1 Humvee with ADS (Active Denial System) mounted [1]

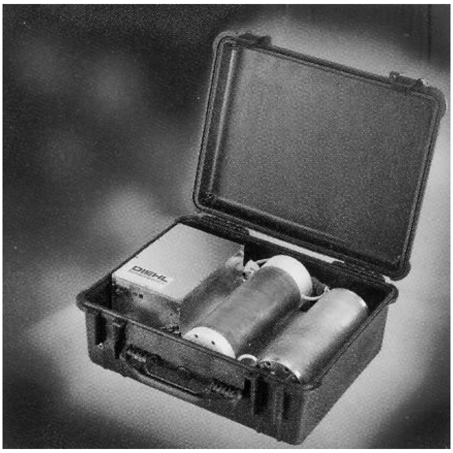

Fig. 2 HPM system DS-110 (built in suitcase, pulse power 250 MW [2]

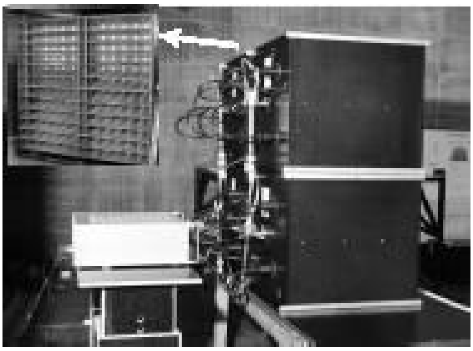

Fig. 3 HPM system with antenna array [2]

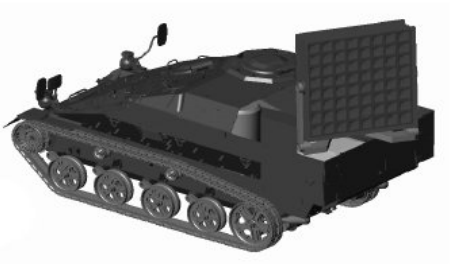

Fig. 4 HPM system with antenna array installed on a tank [2]

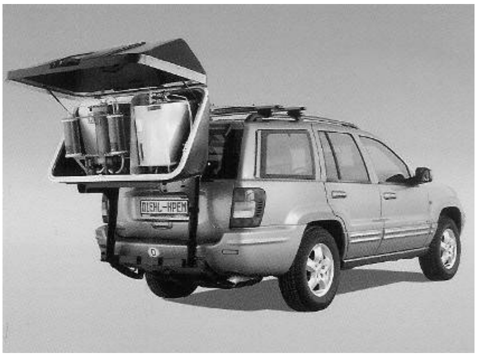

Fig. 5 Electromagnetic System for Selective Car Stopping [2]

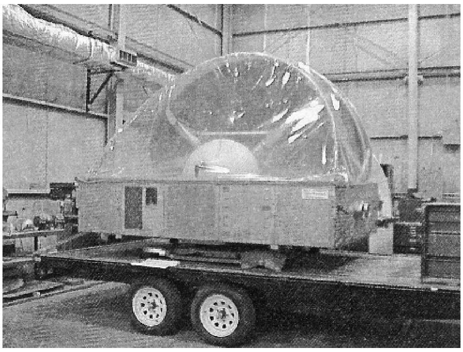

Fig. 6 JOLT hyperband system, UWB system [3]

2. Electromagnetic Threats

As it was mentioned in the introduction an important group of DEW technologies, non-lethal technologies, was represented by weapons with use of power electromagnetic field destructive effects.

Next electromagnetic environments are typically considered related to DEW (see Fig. 7):

- HPM: High Power Microwaves, narrowband signals in frequency range from hundreds of MHz up to several GHz, electromagnetic strength at the place of target typically up to tens of kV/m (strong dependence on range).
- UWB: Ultra Wide Bandwidth, pulses with rise time typically in hundreds of picoseconds and with duration of several nanoseconds, electromagnetic strength at the place of target typically up to tens of kV/m (strong dependence on range).

Because HPM and UWB environment is strongly dependent on range and moreover these environments may have a wide variety of wave shapes, their standardization is very difficult. Now there are only few standards which describe HPM and UWB environment generally (e.g. IEC 61000-2-13 [4]). A possible approach is to consider various types of HPM and UWB environments that have been produced and to consider their possible use in the future against sensitive targets. It is always necessary to make relevant analysis of electromagnetic threat where the shortest possible distance from HPM or UWB source is a very important parameter.

2.1. Electromagnetic Strength Assessment

As a typical possible scenario can be considered HPM (UWB) source with real pulse power $P = 1$ GW and as a worst case special antenna with extremely narrow radiation pattern ($G = 40$ dB, respectively $D = 8\,750$). Now the expected electromagnetic strength at the place of threatened equipment (system) can be calculated according to next Eq. (1)

$$E = \frac{\sqrt{30PD}}{r} \tag{1}$$

where P is radiated power, D is directivity of antenna, r is a distance of threatened equipment (system) from radiating antenna.
Note: Eq. (1) is valid only for direct propagation where reflections are neglected. This simplified equation is usually used for the microwave area and short distances.

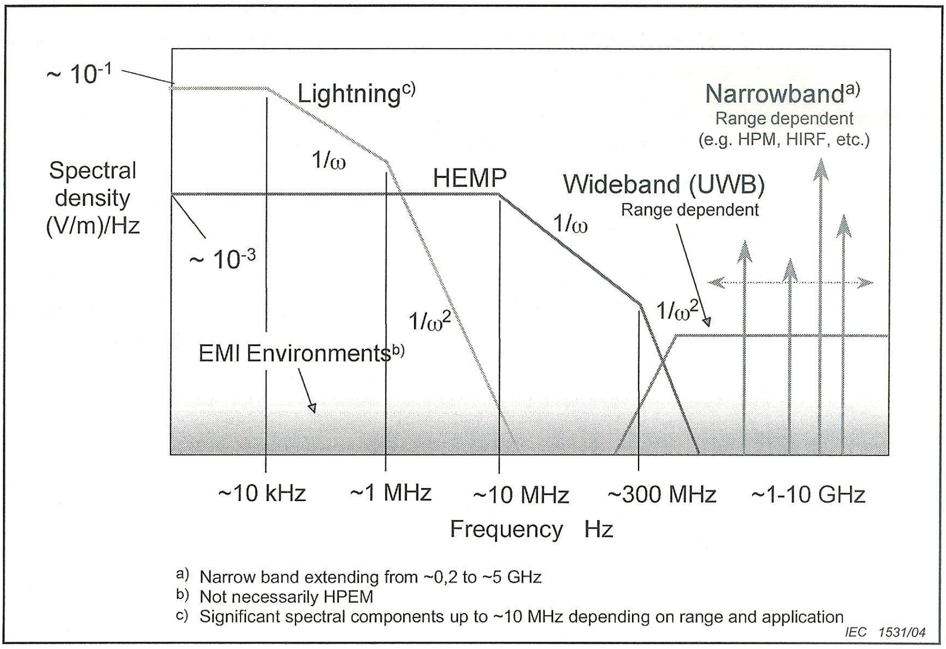

a) Narrow band extending from ~0,2 to ~5 GHz
b) Not necessarily HPEM
c) Significant spectral components up to ~10 MHz depending on range and application

Fig. 7 HPM and UWB comparison with other electromagnetic threats [4]

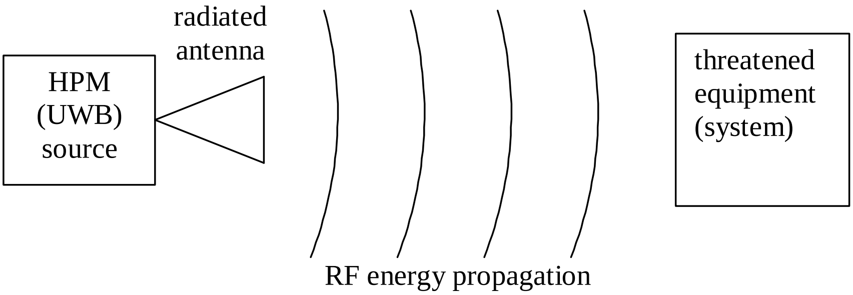

Fig. 8 RF energy propagation from source (HPM or UWB) to equipment (system)

It is obvious from Eq. (1) that for most cases (distance of directed energy weapon from equipment (system) is generally tens of meters), it is possible to consider electromagnetic strength up to tens of kV/m.

2.2. *Critical Frequencies Assessment*

It is very useful to consider transfer function (T) of investigated equipment (system) for critical frequencies assessment. Transfer function T_O (outer) is combined with T_I (internal) to form a composite transfer function from incident fields to response box inputs in Fig. 9 [5]. Note: T_O typically corresponds to relationship between incident electromagnetic field and surface current and charge densities on the outer surface. T_I corresponds to transfer function from the surface current and charge densities on the outer surface to some interior port of interest, producing voltage and current waveforms there.

From Fig. 9 it is obvious the most important part related to critical frequencies of considered system (equipment) is band 2 – resonance region. It is possible to use first of all analysis where sizes of equipment (system) as well as used cable lengths are considered and next it is important to make relevant measurement of transfer functions (shielding effectiveness of used boxes, induced currents and voltages etc.) for transfer function (T) assessment.

Considered frequency spectrum is divided as follows [5] (see Fig. 9):

Band 1: $f \le f_l \approx \dfrac{c}{4l}$ (aperture and small antenna coupling region),

Band 2: $f_1 \le f \le f_h \approx \dfrac{5c}{l}$ (resonance region, external and internal),

Band 3: $f_h \le f$ (integration region),

where l is characteristic dimension of the object, c is speed of the light, f is considered frequency.

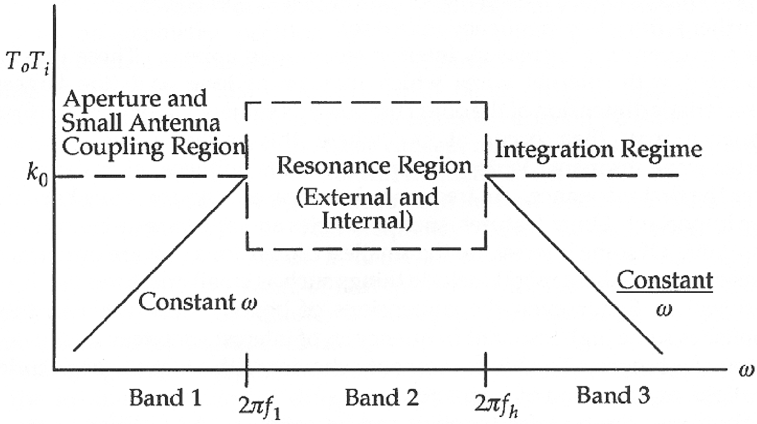

Fig. 9 System response as a function of frequency (logarithmic scales) [5]

3. HPM and UWB Impact on Electronics

Electronic systems can be affected by high intense electromagnetic fields causing the disturbance or even the permanent damage of the system. According to relevant EMC (Electromagnetic Compatibility) standards next performance criteria are usually used:

A - normal performance within the specification limits

B - temporary loss or degradation of function (self-recoverable)
C - temporary loss of degradation of function, which requires operator intervention or resetting
D - degradation or loss of function which is not recoverable due to damage of equipment (components) or software damage

On the one hand performance criteria are a very important parameter on the other hand consequence of IEMI (Intentional Electromagnetic Interference) attack has to be considered too. For consequence scaling it is possible to use for example such terms like "very limited", "limited", "severe", "very severe" and "catastrophic" [6].

4. Vulnerability Assessment

First of all it is necessary to make analysis where electromagnetic threat assessment is done (see chapter 2). Moreover it is necessary to consider criticality of relevant system (consequences – see chapter 3). Finally testing plan and testing can be prepared. The best way is to start with transfer function measurements (shielding effectiveness measurement, induced currents and voltages measurement) which is possible to carry out in wide frequency range up to few GHz or higher. After that it is necessary to choose possible critical frequencies according to transfer function measurements and according to analysis. Chosen frequencies can be used for final high power electromagnetic field measurement with HPM and UWB simulator use. Possibilities of these simulators are restricted to parameters which cannot be changed or changing these parameters takes too much time. Due to this fact it is necessary to choose suitable parameters according to analysis and transfer function measurement results for this testing. The data gained from analyses and testing has to be considered together with possible consequences which can occur. Such data can be used for vulnerability assessment.

5. Practical Results

A lot of experiments related to electronic equipment vulnerability have been carried out in previous projects within the last few years.

Transfer function measurements were carried out for better understanding of electromagnetic field coupling into relevant cables. Tested cables were irradiated with electromagnetic field with vertical polarisation and induced voltages were measured in frequency range 30 MHz to 1 GHz. It was low power measurement, electromagnetic strength of incident electromagnetic field was 1 V/m and measured results were recalculated for expected high power electromagnetic field 10 kV/m. Mainly cables which are widely installed in computer technologies were used for the purpose of these experimental measurements. Few examples are shown in Figs. 10-12. It is possible to expect induced voltage levels on regular widely used cables approximately up to 1 kV during electromagnetic field incidence with electromagnetic strength 10 kV/m on the basis of presented results. The highest induced voltages occur when the length of cables is comparable with wavelength of used signal (resonant region, see Figs. 10-12). For lower frequencies as well as for higher frequencies induced voltages are becoming lower.

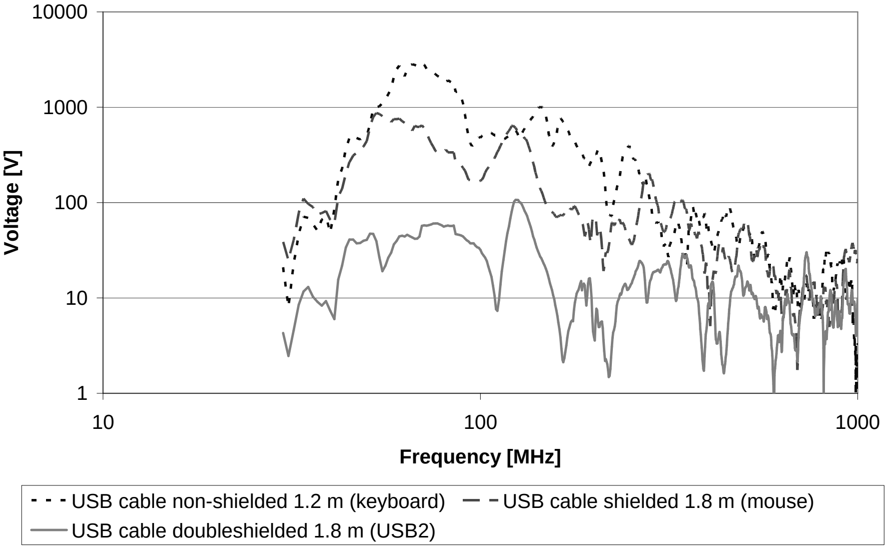

Fig. 10 Induced voltages on USB cables for electromagnetic strength E = 10 kV/m

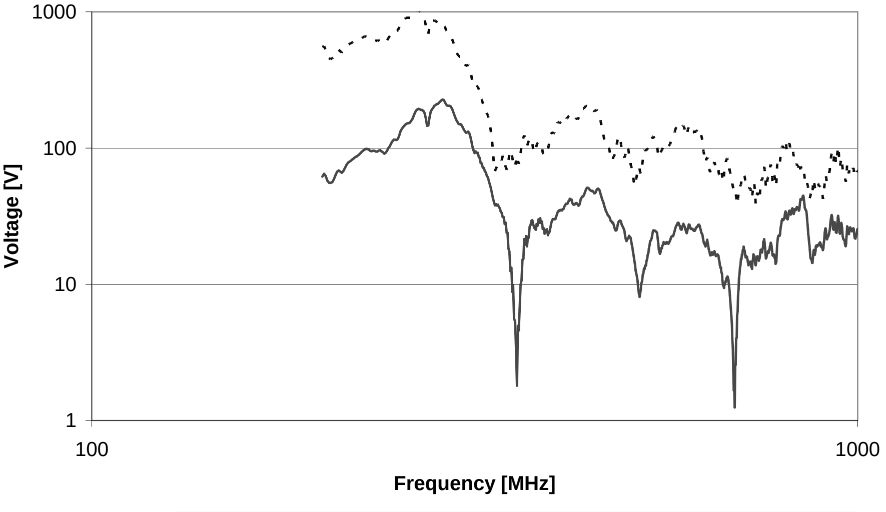

Fig. 11 Induced voltages on SATA cables for electromagnetic strength E = 10 kV/m

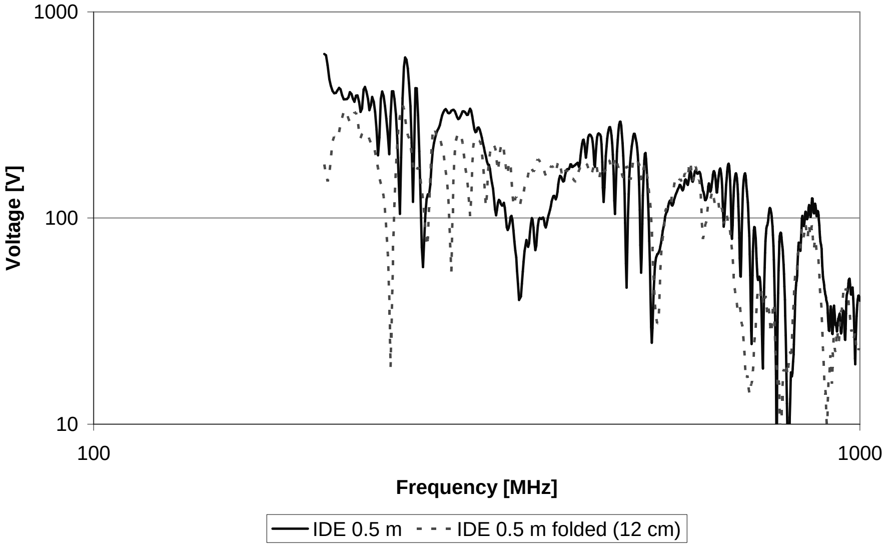

Fig. 12 Induced voltages on IDE cables for electromagnetic strength E = 10 kV/m

Other kinds of experiments in this area are testing with high power electromagnetic field irradiation of possible "targets" and evaluation of effects which can occur.

Some examples of tested setups are shown in Figs 13-15. It was verified that HPM signals can be used for stopping the car engine (see Fig. 13, HPM generator working on frequency 3 GHz with pulse power 500 kW was used for this experiment in shielded semi-anechoic chamber) as well as for effective disrupting of computer technologies and data transfers (see Figs. 14-16). Standard performance criteria (see chapter 3) are usually used for the evaluation of effects. From experiments which were carried out it was obvious that electromagnetic strength of few kV/m can be enough to achieve effective disturbance on electronics.

Repetition rate as a parameter of HPM and UWB signals can have an influence on some kind of temporary failures (see Fig. 16). Overview of immunity of some equipment to HPM and UWB signals is shown in Fig. 17. Typical electronic equipment was tested against effects of HPM 3 GHz, 6 GHz, and 9 GHz signals with pulse widths in hundreds of ns, repetition rates in hundreds of Hz and against effects of UWB signal with rise time of 0.5 ns, pulse duration of 3 ns in single shot regime. The most effective signal to disturb electronics from used ones was HPM 3 GHz. Very often the sensitivity level was very close to damage level (see Fig. 17). For UWB signal it was more typical to achieve temporary failures (sensitivity level) rather than some kind of damage. The results from testing showed the most vulnerable parts of recent systems are computer technologies including data transfers.

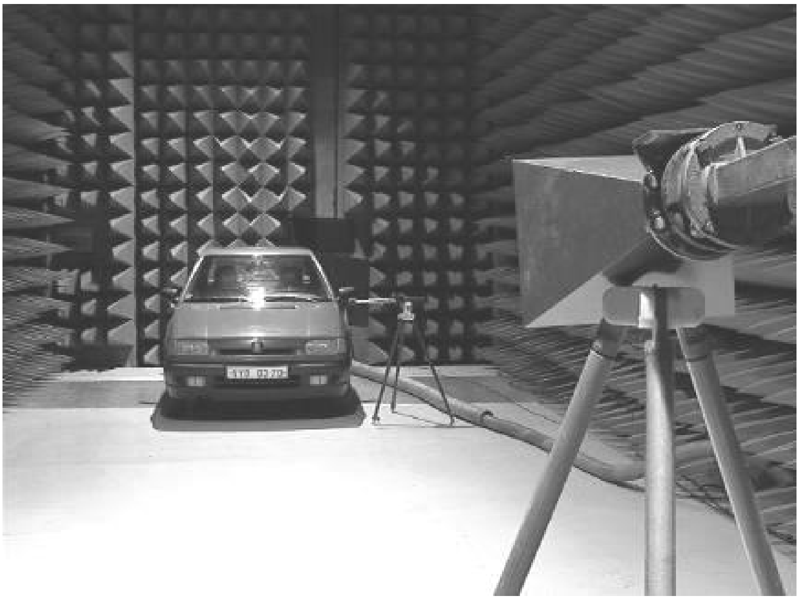

Fig. 13 Stopping the car engine with HPM irradiation use (HPM 3 GHz)

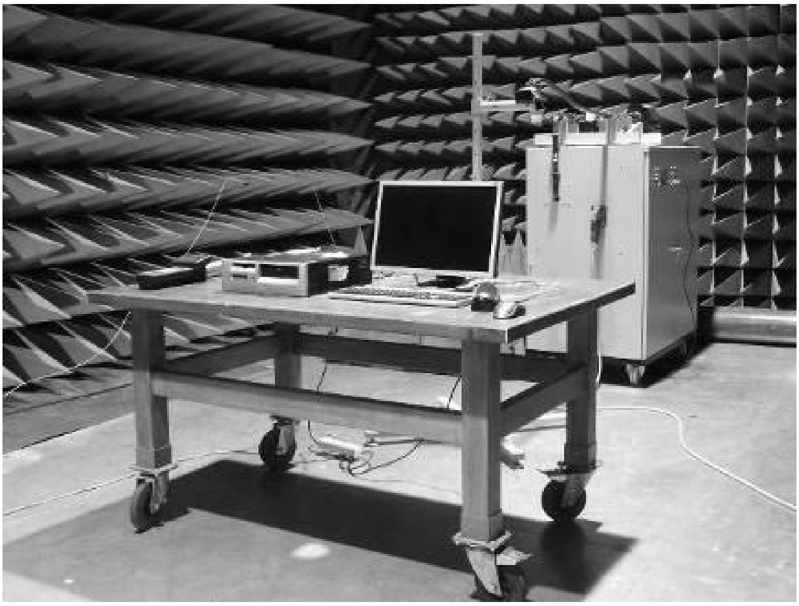

Fig. 14 PC setup irradiated with HPM 9 GHz

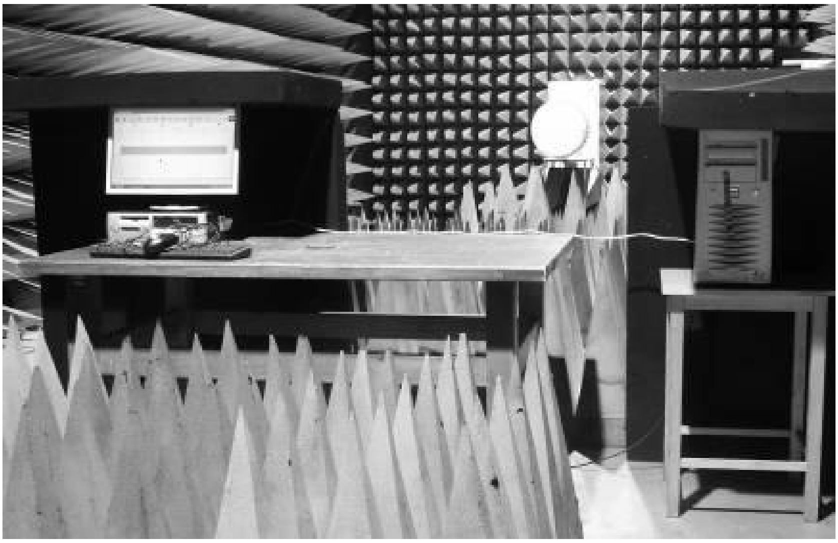

Fig. 15 Data transfer through UTP cable irradiated with UWB signal

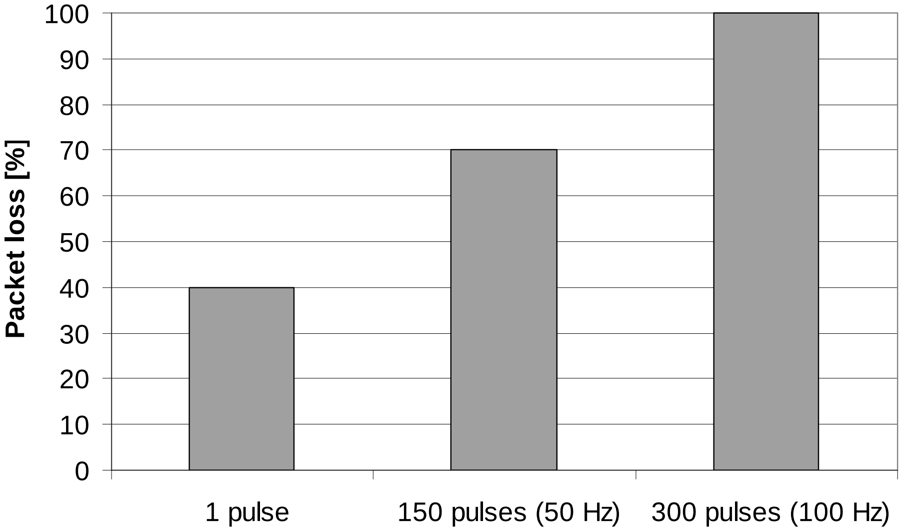

Fig. 16 UTP 2 m irradiated with UWB

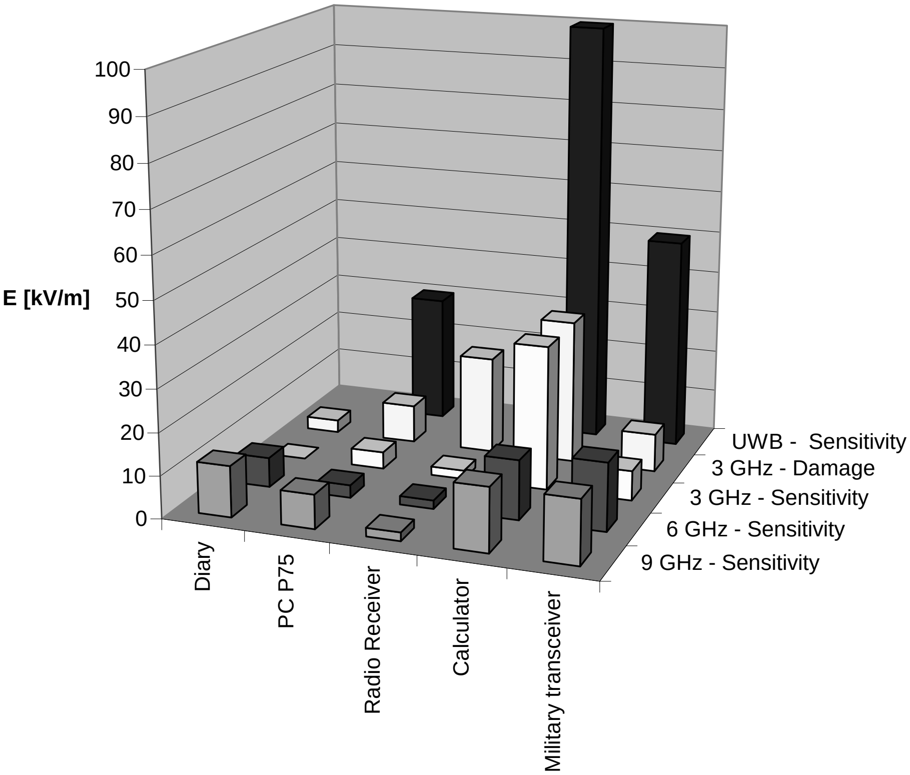

Fig. 17 Sensitivity and damage levels comparison for tested equipment against HPM and UWB effects

6. Conclusion

Electromagnetic threats HPM and UWB as a part of DEW technologies were briefly defined in this paper. Electromagnetic threats assessment as well as vulnerability assessment was introduced. Vulnerability of chosen tested equipment was presented on relevant examples. It was demonstrated that electronic equipment was vulnerable due to electromagnetic field irradiation. In case sensitive electronic equipment is used in crucial applications it is necessary to evaluate vulnerability of the system to such threats.

Gained results in the past have lead to starting a new project for the Ministry of Defence of the Czech Republic (OSPROZ-DEW1) focused on the assessment of vulnerability of weapon systems and parts of communication and information systems integrated in NEC which are installed to ordnance of the Army of the Czech Republic now. Assessment will be carried out on the basis of relevant analyses as well as relevant experimental measurements. The project started in 2009.

References

[1] *Wikipedia* [online]. Wikipedia, the Free Encyclopaedia. [cited 2009-09-15]. Available from: <http://en.wikipedia.org/wiki/Active_Denial_System>.

[2] *DIEHL Defence* [online]. DIEHL Defence. [cited 2009-09-15]. Available from: <http://www.diehl-bgt-defence.de>.

[3] GIRI, DV. *High – Power Electromagnetic Radiators: Nonlethal Weapons and Other Applications.* Cambridge (Massachusetts) and London : Harvard University Press, 2004. 194 p. ISBN 0-674-01569-X.

[4] International Standard IEC 61000-2-13 *Electromagnetic compatibility (EMC) – Part 2-13: Environment – High-power electromagnetic (HPEM) environments – Radiated and conducted*. 1st ed., 2005.

[5] TAYLOR, CD. and GIRI, DV. *High-Power Microwave Systems and Effects*. Washington : Taylor and Francis, 1994. 199 p. ISBN 1-56032-302-7.

[6] MANSSON, D., THOTTAPPILLIL, R. and BACKSTROM, M. Methodology for Classifying Facilities With Respect to Intentional EMI. *IEEE Trans. on Electromagnetic Compatibility*, 2009, vol. 51, p. 46-52. ISSN 0018-9375.

Acknowledgement

This paper is a part of study of the project OSPROZ-DEW1 "Defence capabilities against DEW – Vulnerability assessment of weapon systems and infrastructure C2" supported by the Ministry of Defence of the Czech Republic.

will be available to treat and possibly cure diabetes, hemophilia and other life-threatening and chronic illnesses.

Space manufacturing will develop the new metal alloys needed for high-temperature processing using advanced nuclear and fusion energy sources. Larger and more perfect crystals for the electronics and other industries will be produced.

By themselves, these new industries created by exploiting the unique environment of space, "pay" for the NASA programs. But the indirect impact to the economy has been even greater, and more important. These are the so-called "spin-offs" from space exploration which have been developed by industry using technological breakthroughs funded by NASA.

New materials, such as refractory ceramics and new alloys which withstand the temperature extremes of space flight, are used in everything from nuclear power plants to everyday kitchen implements. Developments in electronics, including miniaturization and sophisticated automation techniques, have streamlined the functioning of industry and shortened the time (thereby increasing the productivity) of almost everything we do.

The machines that monitor all of the vital bodily functions of the astronauts who walked on the Moon, now monitor premature infants in incubators, and have provided heart monitors and miniaturized pacemakers for thousands of people.

The same qualitative impact, magnified many times, will result from the development of beam weapons combined with the recommitment of the civilian space program to establish a permanently manned station in space toward the goal of colonizing the Moon and other planets.

However, studies done by Chase Econometrics and confirmed by independent studies using the Fusion Energy Foundation's econometric model show that there is a much greater impact that the development of a qualitatively new technology has on an economy. This impact is, strictly speaking, not measurable by adding up all the new products and new techniques that a new technology introduces; it is the *increase in productivity* throughout the economy as the result of the combination of higher manpower skill levels and new scientific knowledge entering industrial production. One study of this induced productivity effect estimated that U.S. productivity increased 0.1 percent for every billion dollars spent on the space program. This change in productivity alone represented an additional $3 billion to the GNP every year it was present. An interesting comparison is possible between the expenditure of a $1 billion aliquot of the federal budget on a high-technology R&D oriented program (like the Apollo program or the development of a beam weapon) and its expenditure on transfer payments, bureaucratic services, or the like. The Chase Econometrics study showed that the expenditure of this money on high-technology R&D actually *lowered* inflation, while the other expenditure had the opposite effect, raising inflation by 0.2 percent.

Four types of directed-energy weapons

by Mary McCourt

In his *EIR* multi-client report, *Beam Weapons: The Science to Prevent Nuclear War,* Dr. Steven Bardwell describes the types of beam weapons on line for development. Each type, laser beams, particle beams, microwave beams, and plasma beams, is, Bardwell states, "in principle capable of generating the required power and energy [to reach and disarm its target] in a form efficiently absorbed by the missile." A beam weapon effectively disarms a nuclear warhead. A hydrogen bomb can be detonated only by an initial powerful atomic-bomb explosion capable of setting off a chain reaction in the lithium-deuterium fuel. A beam weapon, by pumping energy into the very delicately balanced triggering mechanism, prevents the initial explosion and essentially turns the warhead into a "dud." The missile, like a satellite, might fall to the earth, but it can no longer be detonated.

Scientists agree that laser-defense battle stations, even with the lowest level of laser-beam technology, can be defended from other beam weapons themselves. But a missile cannot be effectively defended from the beam without such massive protection that it would lose both the necessary speed and distance.

Laser-beam weapons

Laser beams, particularly the chemical laser, will likely be the first deployable beam weapons developed. A laser is a beam of very intense, single wavelength electromagnetic waves, either of light or high energy X-rays. Such a weapon can be focused very precisely because either the light or X-ray wavelengths all have the same frequency and phase. The five different types of lasers, which can be applied to fusion energy as well as beam weapons, are all being researched at U.S. laboratories.

The **chemical laser,** which could be developed for military deployment within five years, uses a gaseous medium in which a chemical reaction is induced. The product of the reaction emits laser light. The Soviet Union used such a laser last year in tests that downed a ballistic missile.

In a **gas laser,** a burning gas such as a hydrogen and fluorine mixture is suddenly compressed, and the energy distribution that results from the compression is then stimulated to emit single-frequency light waves at very high energy. Both the United States and Japan are currently using huge gas lasers for nuclear fusion development.

An **electron discharge laser** uses replaceable energy

© 1982 EIR News Service Inc. All Rights Reserved. Reproduction in whole or in part without permission strictly prohibited.

from an electron beam to create the source of laser light. Such a laser would be very efficient for use in space because its energy source is electricity, not an exhaustible chemical fuel.

Two other types of lasers, **X-ray lasers** and **free-electron lasers,** yet to be perfected technologically, have greater advantages of energy density and flexibility than those listed above. The **X-ray laser** is widely recognized as the most promising long-range laser for ballistic missile defense, based in space. The X-ray laser, which is just a single pulse, is by far the most energy-dense, delivering thousands of times more energy per pulse than conventional lasers. In addition, the target absorbs the X-rays very efficiently, making this weapon capable of very efficient destruction of missiles.

Particle-beam weapons

Particle beams also deliver energy in a highly controlled pulse traveling at near the speed of light. But instead of a pulse of intense electro-magnetic radiation, the particle beam consists of **subatomic particles, (electrons** or **protons), neutral atoms** (usually hydrogen), or (usually magnetized) **macroscopic particles** accelerated to high speeds. A particle beam destroys its target, the triggering mechanism, by creating a very intense shock wave within the mechanism, like a very small, but extremely heavy and powerful hammer striking down on the target.

Electron beams can be generated in the range of millions of volts. Scientists researching the electron beam for military or civilian energy use have discovered that the electron beam becomes a complex structure of electrons and a magnetic field. Such structured beams are capable of carrying higher currents and more energy for much longer distances at much great levels of power output.

Proton beams, which have been researched intensively for the past 30 years, use an electron beam as a seed and then are accelerated in their own right. As protons are 2,000 times heavier than electrons, a proton beam of the same velocity has 2,000 times the energy of an electron beam.

Neutral particles eliminate many of the problems of charged-particle beams, which can degrade both the efficiency and controllability of the beam. By spring of 1983, U.S. researchers will have produced a beam of protons at an energy of 2.5 million electron volts, capable of traveling at 99 percent of the speed of light.

Control and targeting of a **macroscopic particle beam** would be more difficult, but the unparalleled power density of the beam—due to the large mass of the particles—gives it great potential as a beam weapon.

Microwave and **plasma beams** have only been discussed in the United States in the past two years, although it is estimated that the Soviet Union is two or three years ahead in the production of microwaves. Intense, directional microwaves are generated when electron beams are propagated at or near the speed of light through a plasma. Such beams, if focused, could destroy delicate electronic equipment in a target.

A plasma beam consists of the highest energy-dense form of matter, a gas so hot that the electrons and neutrons have separated. The plasma forms itself into a complex structure of particles and magnetic fields. The unique feature of a plasma beam is that it actually requires the atmosphere, which hinders the guidance and propagation of other beams, to hold in the plasma and sustain the structure.

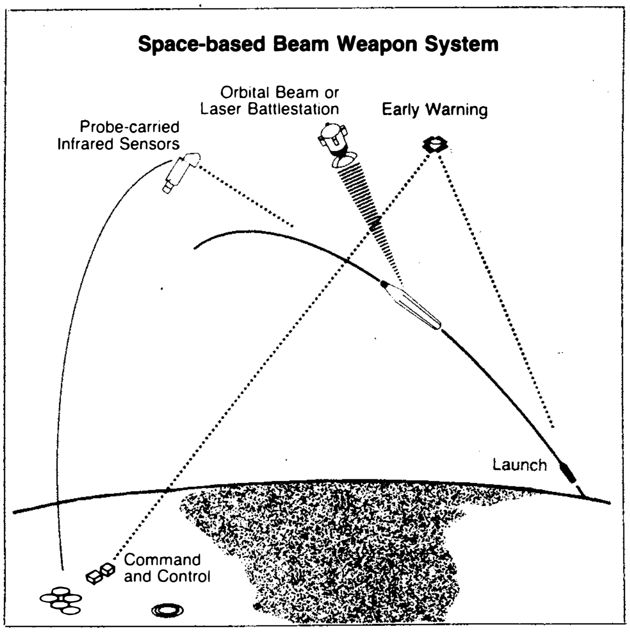
Space-based Beam Weapon System
Orbital Beam or Laser Battlestation
Early Warning
Probe-carried Infrared Sensors
Launch
Command and Control

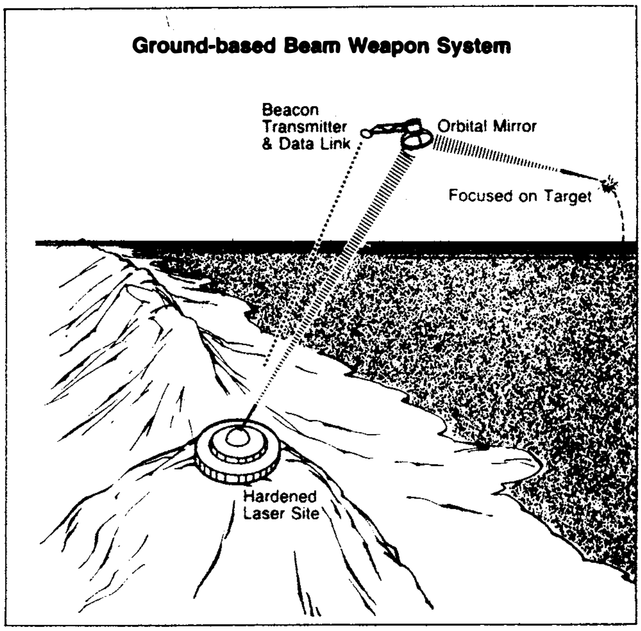
Ground-based Beam Weapon System
Beacon Transmitter & Data Link
Orbital Mirror
Focused on Target
Hardened Laser Site

Executive Intelligence Review Special Reports

Space-based beam weapons, (left) provide the only feasible means for area defense against nuclear-armed ballistic missiles. The deployment of a weapon capable of generating an intense beam of laser light, atomic particles, or plasma, in an orbit around the earth, would protect the entire United States from incoming ballistic missiles. **Ground-based beam weapons** (right) can provide both area and point defense. Shown here is a conceptual design of a laser beam-weapon system built on a mountaintop, which uses a relay mirror to provide aiming and tracking for the weapon.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/242532678

ELECTROMAGNETIC DIRECTED ENERGY WEAPONS FOR ELIMINATING ELECTRONIC SYSTEMS

Article

CITATIONS
2

READS
40

1 author:

Jan Valouch
Tomas Bata University in Zlín

28 PUBLICATIONS 20 CITATIONS

SEE PROFILE

All content following this page was uploaded by Jan Valouch on 06 March 2015.

The user has requested enhancement of the downloaded file.

ELECTROMAGNETIC DIRECTED ENERGY WEAPONS FOR ELIMINATING ELECTRONIC SYSTEMS

Capt. Dipl. Eng. Jan VALOUCH

Annotation:

The article presents information on the current situation in the development and application of electromagnetic weapons and indicates possible trends of their future employment. Under the term electromagnetic weapons we understand means and systems that use technology generating directed power electromagnetic pulses degrading the activities or destructing the electronic circuits of the enemy equipment. This will also enable to neutralize or completely destroy the operation of information, communication, control, firing and other electronic devices. Electromagnetic weapons are currently in use and greater number of countries realizes the perspectives of their future development.

* * *

1. CURRENT SITUATION OF DEVELOPMENT AND USE OF ELECTROMAGNETIC WEAPONS IN FOREIGN COUNTRIES

Principle of electromagnetic weapon activities

Typical representative of electromagnetic (EMC) weapons are so called **Directed Energy Weapons** (DEW) employing the technology of **High Power Microwave** (HPM). These weapons represent highly sophisticated technology of 21st century. Their destructive effect results from the electromagnetic field and they threaten the operation of all devices equipped with electronic circuits, especially by the effect of arcing, overload or by discharge of single electronic component parts. These weapons have strong physical and psychological impact.

In general, DEW consists of impulse source of energy, source of microwave radiation and antenna. The impulse source transforms accumulated energy to high power electric pulse with the length of duration in units of nanoseconds. Accumulated energy can be in kilo-Joule with the power of gigawatt. In the source of microwave radiation this impulse then acts upon diode and generates a beam of electrons with energy – approx, 400 kV, 10—60 kA. This energy is dissipated by the directional antenna.

Electromagnetic weapons can fundamentally differ in terms of its design and application e.g. according to the used frequency band. EMC wide band weapons emit in wide frequency range but with low density of energy. These equipments are suitable where it is not possible to exactly identify the characteristics of the target – especially its working frequencies. Contrary to the narrow-band EMC weapons emit pulses on individual frequencies with enormously high power. Their action upon the target is very effective since impulse resonate with the known frequency of the attacked device.

Development and application of electromagnetic weapons

Concerning a development and application of electromagnetic weapons we can say that after the EMC weapons emerged from "the unknown" world of classified projects, now they are tuned and EMC weapon capability is enhanced. And they are gradually mounted on suitable carriers – vehicles, aircraft, ships, bombs, missiles and even space ships. The interest is focused not only on single pieces of equipment but also on complete systems incl. sensors and equipment to control the combat functions.

Within the armed forces the employment of EMC weapons have the following advantages:

- very rapid effect against the enemy targets,
- usage irrespective of the weather conditions,
- coverage of a great amount of various targets with minimum need to be informed about their characteristics,
- threat to less available target – under ground,
- operational attack (neutralization, destruction or denial of activities of electronic assets) at selected levels of warfare,
- minimum of collateral destruction in politically sensitive environment and use of this environment after the end of conflict,
- reduction of minimum time for tracking and guidance to target.

USA, Russia, France, China and UK are the countries that achieved the greatest progress in the development and application of electromagnetic weapons. The interest of other countries (Germany, Belgium, the Netherlands, Denmark, Norway …) is for a long time focused on High Power Microwave (HPM) especially on protection against the effects of HPM. As to my opinion this is the beginning of the path for research in the Czech Republic.

Currently, discussions are held on their employment in Iraqi by the U.S. armed forces. They could facilitate quickly eliminate the command and control systems of Iraq and eliminate communication of their forces without any loss of lives and any collateral damage. There were some intentions to employ these weapons of electromagnetic warfare by the U.S. force in Kosovo and Afghanistan at least in trial operation. It is also interesting the assumption that Russian forces used DEW prior the action against terrorists in Moscow theatre in 2002. This weapon was designed to disable the electronic detonating primers. The fact is that no Chechen managed to initiate her primer.

Misuse of electromagnetic weapons

More frequent are the attacks of terrorists and criminal underworld using the electromagnetic means. Among the targets of terrorists can be financial institutions, medical facilities, aircraft, automobiles, computer network and other daily used civilian and military equipment. The EMC radiation generators can be for example in the briefcase and that is why it is quite easy to prepare such an attack. First known terrorist application of electromagnetic weapons is from 1995 when Chechen rebels used this technology against the security system of Russian facility.

Nobody knows about the use of DEW for criminal activities and terrorist attacks with exception of the offenders themselves and their victims. The German experts even officially recommended application of these means to the German police units since there were recorded several cases of use of DEW by the German underground. With DEW it is possible to commit perfect crime, as they leave no evidence. No doubt, in future a wide use of these weapons can be expected which can means an increased risk of their misuse.

Future of electromagnetic weapons

Future of DEW is often discussed in many articles and literature. A part of study called Air Force 2025 also deals with the future of DEW employment that was developed by the Air University Maxwell Air Force Base, Alabama. This study discusses a possibility to employ these means in the spaceships or satellites in order to destroy the hostile satellite information and communication channels.

Development of DEW in terms of their employment is mainly connected with mutual co-operation of the army, air force and navy forces. These will be the components that are the potential users of DEW. Protection of the aircraft and ships against the effects of enemy DEW is being solved. Further, e.g. active use of DEW against the air defense and employment of DEW in the aerospace. Within 3—5 years it is expected that DEW will be installed into the drones (USA). In respect of technologies for DEW, the research will be especially focused on increase of input power of DEW, reduction of size and weight and improvement of antenna systems. A very interesting is also a research of biological impact of DEW on human being and hazard for the personnel.

2. CURRENT SITUATION IN THE AREA OF ELECTROMAGNETIC WEAPONS IN THE CZECH REPUBLIC

The following part of this article briefly discusses NATO RTO – Research and Technology Organization (http://www.rta.nato.int), or the activities of selected panels with representation of the Czech Republic and which deal with the mentioned issue.

The panels dealing with this topic are:

- SCI–119 Tactical Implications of High Power Microwaves,
- SCI–132 High Power Microwave Threat to Infrastructure and Military Equipment.

Since most of activities of these technical panels is classified as NATO SECRET, only general character information not subject to classification are provided.

Panel SCI–119 Tactical Implications of High Power Microwaves started its activities in 1998 and terminated it in this year. The objective of the work of its members have been mainly solution of issues of use of HPM in the military sector both in terms of defense as well as of potential active employment in attack. The most important areas solved:

- design of resources generating HPM,
- vulnerability of information technology of military infrastructure,
- risk of back effect of electromagnetic weapons on friendly troops,
- testing of HPM effects on the off-the-shelf equipment equipped with electronic components (personal computers, cellular phones, vehicles, aircraft …),
- content of HPM national programs from the viewpoint of:
  - earmarking of specialized workplaces,
  - refinement of the content of solution in the given country,
  - allocation of funds,
  - planning of the number of students studying HPM at the universities,
  - gathering of results.

The issue of HPM is solved by NATO systematically since the beginning of 80-ies of the last century. In the short time, it is assumed that new advanced HPM weapons will be developed and within NATO this process must be systematically controlled. In this area – development of electromagnetic weapons – USA, Russian, France and the UK have the greatest success.

Panel SCI–132 High Power Microwave Threat to Infrastructure and Military Equipment commenced its activity by the introductory session in October 2002, held in Munster in Germany. Its orientation is linked to the preceding activity of the Panel SCI–119. The main activities of the Panel are planned for the period of 2003—2005 and single members will deal with the following issues:

- identification of potential threat of HPM on the military and civilian infrastructure and equipment,

- research of the penetration and dissemination of HPM in specific devices,
- research of DEW (Directed Energy Weapons),
- possibilities how to protect the military equipment against the effect of weapon assets on the basis of HPM.

Other areas considered important by the member of the Panel SCI–132 and that are to be in the area of interest are:

- resistance of military equipment against HPM,
- implication of HPM in the military field test and system of their evaluation,
- EMC terrorism.

The Panel consists of representatives of the following countries: Canada, Czech Republic, Denmark, Germany, France, UK, USA, the Netherlands, Norway and Italy.

CONCLUSION

Currently, the use of electromagnetic weapons plays the more important role. Though, the wider employment of these weapons is expected within 5—10 years, there already exist and are used the devices/equipment capable to reliably effect on the enemy electronic assets (control, information, communication…) and deny the enemy any further operations. The issue of electromagnetic weapons in the Czech Republic is in the stage of research and that is why the representation and involvement of the Czech representatives in the special Panels of NATO RTO or any other international organizations and boards is desirable.

REFERENCES

[1] MUNZERT, Reinhard: Targeting the Human with Directed Energy Weapons, 2002, Erlangen, Germany

[2] MERKLE, Laurence D.: Virtual Prototyping of RF Weapons, 2002, The Air Force Research Laboratory, Kirtland.

[3] COOP, Carlo: The Electro-magnetic Bomb: A Weapon of Electronic Mass Destruction, 2000, Melbourne Australia.

[4] VALOUCH, Jan: Activities of VTÚ PV Vyškov in the area of EMC and DEW, presentation of NATO RTO / SCI–132, 2002, Munster, Germany.

Selection of unclassified documents from the session of Panel NATO RTO/SCI:

[5] SCI–119 Workshop Tactical Implications of High Power Microwaves, Copenhagen, Denmark, 11—13.6.2002.

[6] Open Meeting of NATO RTO/SCI–132 on High Power Microwave to Infrastructure and Military Equipment, Munster, Germany, 22—24.10.2002.

View publication stats

James Madison University

From the SelectedWorks of George H Baker

August, 2011

High Power Electromagnetic Weapons: A Brief Tutorial

George H Baker, III, *James Madison University*

Available at: https://works.bepress.com/george_h_baker/33/

High Power Electromagnetic Weapons: A Brief Tutorial

George H. Baker

Background. High power electromagnetic weapons, also referred to as high power radiofrequency (HPRF) weapons, are a type of directed energy weapons. The system effects of high power electromagnetic environments are well recognized by world scientific and military communities. Former CIA Director John Deutch has said that, "the electron is the ultimate precision-guided weapon." In the course of the investigation ofnuclear EMP effects on electronics during the Cold War period, it became evident that garden variety, unprotected electronics would malfunction, in some cases burn out, in the presence of electromagnetic fields in the hundreds to thousands of volts per meter. The EMP experience has led to the development of non-nuclear high power electromagnetic sources to create fields that equal or exceed EMP levels, albeit over relatively small ranges. Achievable electronic effects could have serious consequences in terms of interruption or termination of critical system operation. The effects are of particular interest to the military in the context of information warfare and missile defense. Because most critical infrastructures are controlled by electronics, HPRF weapons are a concern for civilian systems as well. The weapons could be used to disrupt computer electronics controlling electric power grids, telecommunications networks, financial institution databases, security systems, and aircraft.

Military forces in many countries are pursuing the development of HPRF weapons. These programs are normally classified. HPRF weapons represent a revolutionary concept because they operate at the speed of light, can be used covertly, and harm equipment rather than humans (non-lethal). HPRF weapons may be enclosed in briefcases, packing containers, truck beds, or aircraft and maneuvered to expose critical military or infrastructure systems.

Electromagnetic Weapon Characteristics. There are three elements to be considered in discussing HPRF weapons systems: the weapon itself, the propagation of the weapon output to the target, and the target response (see Figure 1).

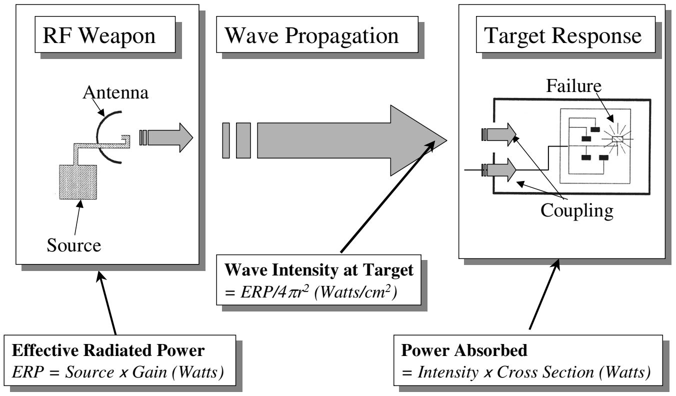

Figure 1. High Power RF Weapon Operation Elements

©2011 George H. Baker

Weapon Design. An HPRF weapon consists of a power source, or driver and a radiating antenna. The source is normally pulsed. Outputs may be sine waves (providing a narrow frequency band signal) or pulses (providing a wide frequency band signal). High power electromagnetic source technology includes electron beam devices (magnetrons, vircators, gyrotrons, backward wave oscillators), solid-state devices (bulk avalanche, optical switches, silicon carbide circuits) and explosive generators (magnetic flux compression using high explosives). Output power levels in the gigawatts are feasible over a wide range of frequencies. Design details will determine output characteristics. Power, frequency, bandwidth, repetition rate, and duty cycle are the important HPRF weapon output parameters.

The HPRF weapon source must be connected to an antenna with sufficient gain to "beam" the RF energy to a useful range. Portability requirements impose major limits on antenna area which in turn largely governs the maximum intensity (power per area) that can be delivered to a target system at a given range. Compact sources that may be moved covertly in briefcases, packing containers, truck beds, or aircraft and maneuvered for close-up exposures of critical military or infrastructure systems are of most interest. Antenna size limits largely determine the maximum intensity (power/area) that can be delivered to a target system at a given range. A table of approximate relative portable platform sizes is provided below.

Table 1. Portable Platform Size Comparisons (Approximate)

Capacity: Platform:	Volume	RF system Weight	Basis
Briefcase	0.02 m^3	5 kg	Typical hard side
Footlocker	0.15	40 kg	1 person portable
Pickup truck	4 m^3	1000 kg	1 ton capacity
Econoline van	8 m^3	2000 kg	Bed volume
Tractor trailer	80 m^3	20,000 kg	20 ton capacity

High Power Electromagnetic Wave Propagation. Electromagnetic waves propagate at the speed of light through the atmosphere. Under most conditions, the atmosphere will not attenuate HPRF waves as they travel from source to target. However, if the electromagnetic wave's peak field level exceeds the air ionization threshold a cascade process will occur in which an atmospheric plasma is created that will absorb most or all of the energy in the wave. This phenomenon is referred to as air breakdown.

A typical threshold for breakdown at sea level is approximately 1 megawatt/cm^2 . As an example, a 1 Gigawatt source radiating through a 1-foot radius circular antenna will produce an average intensity over the antenna of 0.35 megawatts/cm^2, a factor of 3 below air breakdown. Once away from the antenna, EM intensity falls off as the beam spreads (typically as $1/4\pi r^2$ where r is the distance to the target).

©2011 George H. Baker

The HPRF signal intensity on target depends on source transmission power (P_t), antenna gain, and range according to the following equation:

$$S_t = \frac{ERP}{4\pi r^2} = \frac{P_t \cdot G}{4\pi r^2} \frac{watts}{m^2}$$

The antenna gain may be approximated by $G \approx \frac{4\pi A_e}{\lambda^2}$ where A_e is the effective antenna area and λ is the wavelength of the radiated HPRF signal. This yields a simple formula for the intensity on target:

$$S_t = \frac{P_t A_e}{r^2 \lambda^2}$$

So, for example, a 1 Gigawatt source radiating at 1 GigaHz ($\lambda = 0.3$m) with a 1-m^2 antenna will produce a beam intensity of 11 kW/m^2 (or 1.1 watts/cm^2) on a target 1000 meters away. If the pulse duration were 1 microsecond, the energy fluence on target would be power x time or, 1.1 microjoules/cm^2.

Target Response. Once the electromagnetic wave signal reaches the target, its energy couples to the system in a very complex manner through various paths associated with the topology of the target system. The HPRF signal will induce currents on any external conductors (antennas, wires, etc) that penetrate to the system interior. The signal will also couple to any external metal shielding and then may reradiate to internal conductors. Signal waves will penetrate through any holes in external shielding to interior portions of the system. The main concern is the amount of energy that finds its way to critical electronic circuits, causing the system to malfunction.

To simplify coupling calculations, it is possible to determine an effective "coupling cross section" for critical internal circuits. In effect, this treats each internal circuit as a receiver. The power received by the circuit may be expressed as a simple function in incident wave power and an effective coupling cross section:

$$P_i = S_t \mathrm{A}_\sigma$$

P_i is the power induced in the internal circuit, S_t is the incident power at the system's exterior, and A_σ is the coupling cross section of the internal circuit. A_σ is a function of frequency and incorporates all the complexity of coupling including multiple paths, layers and mechanisms. There are two major coupling modes, front door and back door, as shown in figure 2.

©2011 George H. Baker

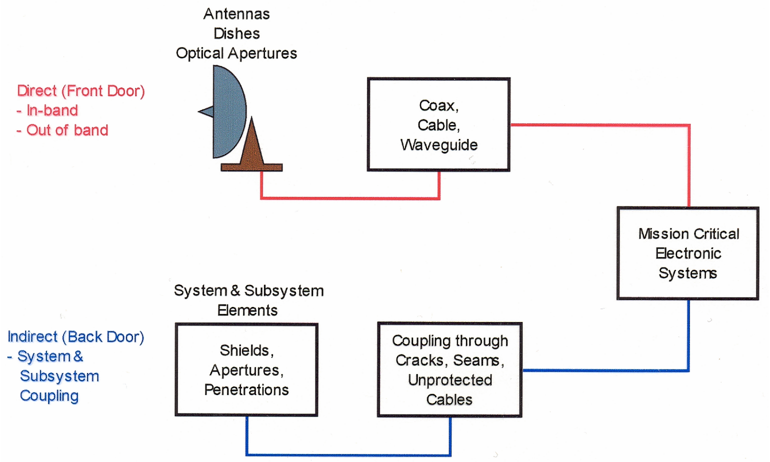
Antennas
Dishes
Optical Apertures
Direct (Front Door)
- In-band
- Out of band
Coax,
Cable,
Waveguide
Mission Critical
Electronic
Systems
System & Subsystem
Elements
Indirect (Back Door)
- System &
Subsystem
Coupling
Shields,
Apertures,
Penetrations
Coupling through
Cracks, Seams,
Unprotected
Cables

Figure 2. Front Door and Back Door Coupling Modes

The physics of front and back door coupling are the same. In each case, energy resident in the incident wave induces currents in the system that flow to a sensitive system circuit. Each can be characterized parametrically by equation $P_i = S_t A_\sigma$. The coupling cross-section, A_σ, is typically large for front door coupling and small for back door coupling. The coupling cross sections for front door coupling are of the order of the physical area of the antenna or aperture. Effective coupling cross section values for back door coupling of unhardened systems range from $10^{-4} – 10^{2}$ cm^2. Coupling cross-sections are hard to predict analytically and can be determined confidently only by direct measurement. This creates a problem for an attacker's confidence since HPRF effects are subject to much higher uncertainties than conventional weapon effects.

HPRF Weapon Effects on Systems. EMP and HPRF affect systems by disrupting the operation of electronic components either temporarily (upset or latchup) or permanently (component damage). Damage may be "direct" where energy inherent in the EMP/HPRF field is sufficient to cause malfunctions or "indirect" where EMP/HPRF energy triggers effects involving a system internal power supply. Even though the power delivered to a circuit by the HPRF field is small, the much larger energy available in the system's power supply (or fuel and ordnance) can be improperly diverted by an HPRF induced overvoltage arc or malfunction of system digital control circuit, causing major system damage.

Upset refers to an induced change of state in a digital circuit in which the system continues to operate, although possibly with erroneous data bit streams. Latchup refers to changes of state in digital circuits where the affected portion of the system ceases to operate until the system resets itself or (worse) a manual reset is required. In either case, system components are not directly damaged and recovery is often possible (depending on time criticality of affected function).

Exploitation of upset effects should not be discounted. In some cases they are tantamount to permanent damage. Upset may result in major system damage, e.g., a missile plunges into ocean due to a guidance system upset, or computer equipment is destroyed by an upset sprinkler system. Small, upset-level transient pulses can also act to trigger the release of energy from a system's own power supply causing components to burn out at fluences much lower than would normally cause permanent damage.

To give an indication of HPRF wave energies required to cause system effects, ITT industries (formerly Kaman Sciences) openly exposed circuit cards to 2.9 GHz microwave pulses with duration 1 microsecond. Onset of upset occurred at 1 watt/ cm^2 (equivalent to 1 microjoule/cm^2 for this pulse duration). All components were upset at wave intensities of 1000W/cm^2. Onset of damage occurred at 100W/cm^2 (for a 1 microsecond pulse, this is equivalent to 100 microjoules/cm^2). All components were damaged at wave intensities of 1000 W/ cm^2 (equivalent to 1000 microjoules/cm^2).

System Protection. System hardening involves a combination of operational and hardware techniques. Operational techniques may include the provision of spares for soft critical subsystems or boxes, disconnecting susceptible circuits upon warning, and/ or establishment of a physical keep-out perimeter (with barriers and/or security force) or zones around critical equipment to prevent positioning of HPRF weapons at close ranges. Operational controls may also be built into software to provide circumvention and reset, error-correcting codes, voting logic, and status detection. For some non time-sensitive systems, provisions for rapid system repair may be an option.

Conceptually, hardware approaches involve placing a conducting material between the incident HPRF wave and susceptible internal circuits. Hardening techniques have been successfully demonstrated and codified for EMP (ref. article on nuclear EMP). The EMP community has placed a heavy reliance on exterior shielding while limiting the number of penetrations that have to be individually protected. Such protection applied at the system exterior allows interior boxes to go untreated. This approach works well when designed in from the start. For retrofit protection, however, it is often prohibitively expensive.

Hardware approaches for HPRF protection, while conceptually similar, have some differences in emphasis from EMP. Because EMP is extremely broadband, typically only a small fraction of the energy comes through the front door in band. HPRF weapons can be tuned to the front door center frequency such that all the beam energy flows into the system. Also since HPRF weapons operate at higher frequencies, attention to smaller apertures (including cracks and seams) is required and dimensions of waveguide-beyond-cutoff penetration treatments will require changes (longer and more narrow waveguides are needed).

Front door in-band protection is one of the more challenging (but not insurmountable) HPRF protection problems. The high gains associated with most front door paths make these potentially the most susceptible portion of the system. However these well-characterized front door receive paths have received much attention in terms of protection engineering. Radar systems are often protected from their own or neighboring transmitters by a receiver protector or RP. Similar protection can be applied to communication receivers against in band HPRF environments. Table 2 summarizes HPRF hardening techniques.

©2011 George H. Baker

Table 2. HPRF Hardening Methods

Operational Techniques	Hardware Techniques	
	Front Door	**Back Door**
Disconnect on warning	Tuned metallic radomes	Shielding, topology control
Spares	Bandstop reflectors	Internal cable shields
Physical keep-out perimeters or zones	Antenna gain pattern control	Non corrosive mating surfaces
Software Techniques: Circumvention/reset Error correcting codes Voting logic Status detection/ alarm	Filters	Closely spaced fasteners
	TR devices/ limiters	RF gaskets and seals
	Terminal protection devices	Aperture screens and conductive films
		Selection of hard circuits and components
		Pin protection
		Fiber Optics
		Filtering, limiting at penetration points

Future Directions. HPRF generation techniques have matured to the point where practical devices have become technically feasible. Miniaturization of pulse-power source components, more efficient power supplies, and advances in electronic pulse forming, energy conversion, and antennas enable reduced size and increased efficiency of HPRF generation hardware. Simple weapons can also be built using inexpensive magnetrons from common microwave ovens.

Military trends to computerized battle management, weapons tracking, and landline/wireless network communications are a double-edged sword, introducing serious HPRF vulnerabilities. Military use of commercial off-the-shelf equipment and dependence on civilian infrastructure exacerbate the problem. Senior military officials have dropped hints about pursuing offense technology but there are no officially published details concerning weapon availability or capabilities.

Several other countries also have extensive background in the development of RF weaponry. The former Soviet Union pioneered the development of HPRF weapon technology and this technology is now being offered to other countries. According to a recent report from the Office of the US Secretary of Defense on the military power of China, “Captain Shen Zhongchang from the Chinese Navy Research Institute…envisions a weaker military defeating a superior one by attacking its space-based communications and surveillance systems…in future wars, Shen highlights radar, radio stations, communications facilities, and command ships as priority targets vulnerable to smart weapons, electronic attack, and electromagnetic pulse weapons.” It is expected that in future conflicts the United States will encounter adversaries using HPRF weapons as part of asymmetric tactics to disrupt information systems.

©2011 George H. Baker

See also: Nuclear Electromagnetic Pulse (EMP)

References and Further Reading

Abrams, M., "The Dawn of the E-bomb," IEEE Spectrum, October, 2003.

Backstrom, M. et al; *Preliminary Study Regarding the Resistance of Critical Societal Systems to High Intensity Electromagnetic Radiation*; Royal Swedish Defense Research Agency; Report FOA-R-97-00538-612-SE; August, 1997.

Barker, R., Schamiloglu, E., *High-Power Microwave Sources and Technologies,*Wiley-IEEE Press, 2001.

Benford, J., Segle, J., *High Power Microwaves*, 2nd Edition, Institute of Physics, 2004.

Guyatt, D. G., "Some Aspects of Antipersonnel Electromagnetic Weapons," ICRC Symposium on the Medical Profession and the Effects of Weapons, February, 1996.

Kopp, C., "The E-Bomb – A Weapon of Electrical Mass Destruction," INFOWAR Conference, Washington, D.C. September 1996.

Lucier, J. P., "Backyard Terrorism - E-Zapper Could Break the Bank,"
Insight Magazine, April 1, 2002.

McKenna, T., "Flash in the Plan," Journal of Electronic Defence, May 2003.

Price, D. et al, "Compact Pulse Power for Directed Energy Weapons," Journal of Directed Energy, Volume 1, Fall 2003.

Saxton, J., Chairman, *Congressional Hearing on Electromagnetic Weapons*, Joint Economic Committee, Intelligence and Security, Source: http://www.house.gov/jec/hearings/02-25-8h.htm, June 17, 1997.

National Research Council, Technology for the United States Navy and Marine Corps, 2000-2035, Volume 2: Technology, National Academy Press, Washington, DC, 1997.

About the author: George Baker is Professor of Integrated Science and Technology at James Madison University. He also serves as Technical Director of the University's Institute for Infrastructure and Information Assurance (IIIA). He is a consultant in the areas of critical infrastructure assurance, high power electromagnetics, nuclear and directed energy weapon effects, and risk assessment. He served as senior staff on the Congressional EMP Commission. Baker is former director of the Defense Threat Reduction Agency's Springfield Research Facility, a national center for critical system vulnerability assessment. Much of his career was spent at the Defense Nuclear Agency (DNA) as the Integrated Electromagnetics Team Leader directing the development of

©2011 George H. Baker

electromagnetic protection, underground testing, and standards including MIL-STD-188-125, MIL-STD-2169B, and MIL-HDBK-423. He also headed the Agency's Innovative Concepts Division overseeing the joint US-Russian space nuclear power technology, electro-thermal chemical (ETC) gun development, radiofrequency directed energy source concept development and testing, and DNA's university grants programs. Baker received the Agency Legacy Award for his leadership and innovation. He is a member of the NDIA Homeland Security Executive Board, the Institute of Electrical and Electronic Engineers, the Directed Energy Professional Society (Charter Member), and the Association of Old Crows. He is a Summa Foundation EMP Fellow and holds a Ph.D. from the U.S. Air Force Institute of Technology.

©2011 George H. Baker

archived as **http://www.stealthskater.com/Documents/Hamilton_04.pdf**

more of Dan Burisch is at http://www.stealthskater.com/Burisch.htm

note: because important websites are frequently "here today but gone tomorrow", the following was archived from http://www.skywatch-research.org/QuantumHolographic.htm on April 8, 2004 . This is NOT an attempt to divert readers from the aforementioned web-site. Indeed, the reader should only read this back-up copy if it cannot be found at the original author's site.

The Biophotonic Quantum Holographic Matrix

by **William F. Hamilton III**

ABSTRACT: *This paper is an attempt to integrate studies of the quantum potential, quantum holography, biophotonics and the enveloping matrix of biointegration, biocommunication, and bioinformation that composes the web of light and life in living organisms and their possible genesis in an electromagnetic infoton.*

Introduction:

Theoretical physicist David Bohm proposed a new hidden variables theory of Quantum Mechanics. He assumes the wave function does not represent just a set of probabilities but an actual field. This field exists and acts upon particles the same way a classical potential does. Thus it is a quantum potential which is associated to this field and is a function of the wave function. It is postulated here that the quantum potential is active -- as a field -- in living cells and organisms and acts upon them.

The equation Bohm used to represent this quantum potential is:

$$U = -\sum_k \left(\hbar^2 / 2m_k\right)\left(\partial_k^2 R / R\right) \qquad (1)$$

Research reveals that Bohmian quantum potential is still used, but that it may only represent a means to extend Quantum Theory into a realistic theory of particles and fields.

It resolves the dilemma of the appearance -- in one and the same phenomenon -- of both particle and wave properties in a rather straightforward manner. Bohmian mechanics is a theory of motion describing a particle (or particles) guided by a wave. This guiding wave is usually attributed to atomic particles. But it might be more insightfully seen as a traveling **soliton** or vortex.

Bohmian mechanics is manifestly nonlocal, and the velocity may depend upon the positions of other -- even distant -- particles whenever the wave function of the system is entangled.

The behavior of the particle is determined by the particle’s position and momentum, the wave field, and the sub-quantum fluctuations.

Bohm's quantum potential binds the entire Universe together into what he liked to call a seamless "unbroken wholeness". Every particle in the Universe is connected by the quantum potential to every other particle. He likened the cosmos to a **hologram** in which each point on the film carries information about the entire picture. Bohm's GWT -- far more sophisticated than de Broglie's crude version -- is a "holistic" vision in which all parts of the Universe are joined to every other part. "Interconnectedness" was one of Bohm's favorite words. He saw the Universe as resembling the unity of a living organism. A kind of pantheism not unlike Spinoza's -- a pantheism Einstein himself favored.” **[2]**

Bohm referred the holomovement as the basic reality. The pattern formed on a photographic plate by light waves that interfere and create a hologram. This pattern is the interaction or interference pattern of two parts of a laser light (coherent light). One beam reflects off an object, the other off a mirror. In addition, any portion of the holographic plate contains information on the whole object.

Biohologram:

Every cell in the human body contains nuclear DNA. Information on the whole body -- much like a hologram -- only the cells can go through a process of division and reproduction (mitosis).

In quantum holography, things are spookier still. While holograms are typically constructed with interfering beams of light, in quantum holography the researchers measure the simultaneous arrivals of an illuminating photon that is sent into a chamber and a companion photon in the other entangled beam.**[2]**

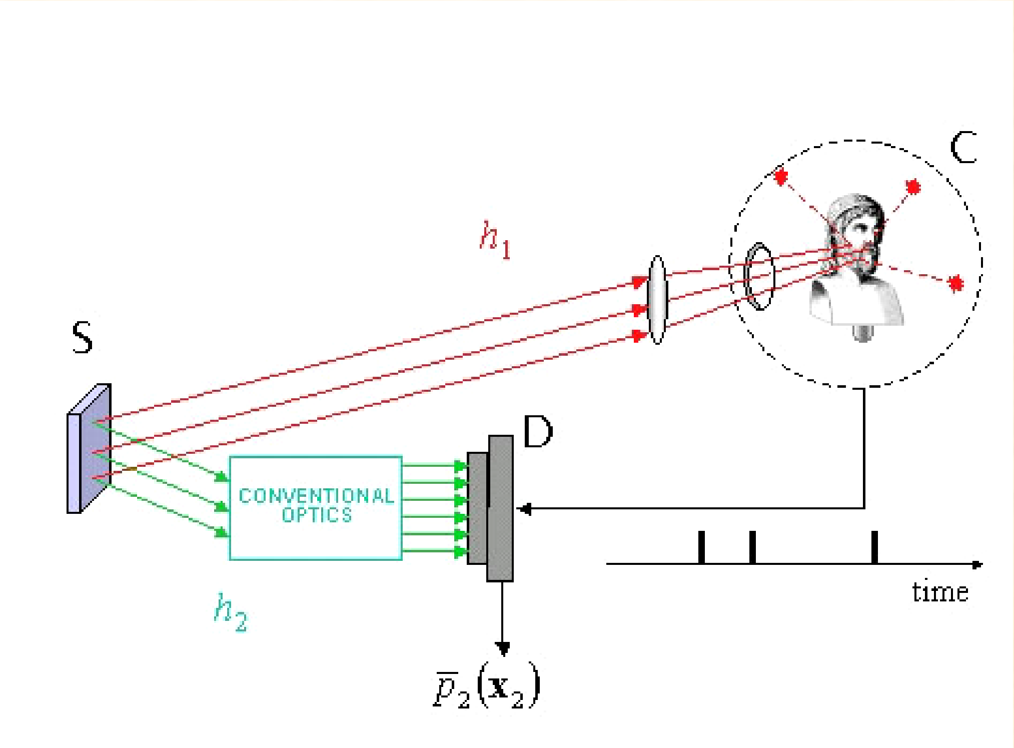

Quantum holography -- as depicted in this cartoon -- is a way that physicists could produce 3-dimensional images of objects that would be invisible to systems relying on classical physics. A source **S** produces 2 photon beams (**h_1** and **h_2**). The photons in **h_1** are quantum mechanically **entangled** with the photons in **h_2**. When a photon in **h_1** enters chamber **C**, it encodes 3-dimension information about the enclosed object (in this case a Grecian bust) in the interference of the 2 paths it could take -- a path that intercepts the object and subsequently is reflected to the chamber wall, and a path that misses the object and strikes the chamber wall directly. The second beam **h_2** passes through conventional optics and strikes a detector array **D**. A coincidence counter extracts the holographic information by monitoring the relative time between a photon in **h_1** striking the chamber wall and its entangled mate in **h_2** arriving at detector **D**.

Wave Genetics Inc. and the Moscow Institute of Control Sciences are exploring a new concept of the genetic code that asserts:

(1) that the evolution of biosystems has created genetic "texts" -- similar to natural context dependent texts in human languages -- shaping the text of these speech-like patterns.

2) that the chromosome apparatus acts simultaneously both as a source and receiver of these genetic texts, respectively decoding-and-encoding them.

3) that the chromosome continuum of multicellular organisms is analogous to a static-dynamical multiplex time-space holographic grating, which comprises the space-time of an organism in a convoluted form.

That is to say, the DNA action -- theory predicts and which experiment confirms -- is …

(i) that of a "gene-sign" laser and its solitonic electro-acoustic fields, such that the gene-biocomputer "reads and understands" these texts in a manner similar to human thinking but at its own genomic level of "reasoning". It asserts that natural human texts (irrespectively of the language used) and genetic "texts" have similar mathematical-linguistic and entropic-statistic characteristics, where these concern the fractality of the distribution of the character frequency density in the natural and genetic texts and where -- in case of genetic "texts" -- the characters are identified with the nucleotides, and

(ii) that DNA molecules -- conceived as a gene-sign continuum of any biosystem -- are able to form holographic pre-images of biostructures and of the organism as a whole as a registry of dynamical "wave copies" (or "**matrixes**”) succeeding each other. This continuum is the measuring, calibrating field for constructing its biosystem.

“Complex information can be encoded in EM fields, as we all know from coding-and-decoding of television and radio signals. Even more complex information can be encoded in holographic images. DNA acts as a holographic projector of acoustic and EM information which contains the informational quintessence of the biohologram. Only 3% of human DNA encodes the physical body. The remaining 97% of the 3 billion base pair genome contains over a million genetic structures -- called transposons -- that have the capacity to jump from one chromosomal location to another (Kelleher, 1999). We are 99.9% alike in our genetic legacy. Our individuality is expressed in 3 million small variations in our cells called single nucleotide polymorphisms.

"**Gene-expression** is the mechanism by which new patterns are called into being (Rossi, 2000). There is also a strong correlation between modulation of the brain’s EM field and consciousness (Persinger, 1987; McFadden, 2002). The Gariaev group has discovered a wave-based genome and DNA phantom effect which strongly supports the holographic concept of reality (Miller, Webb, Dickson, 1975). This main information channel of DNA is the same for both photons and radio waves. Superposed coherent waves of different types in the cells interact to form diffraction patterns, first in the acoustic domain and second in the electromagnetic domain -- a **quantum hologram** -- a translation process between acoustical and optical holograms.” **[4]**

The distribution of the character frequency in genetic texts is fractal, so the nucleotides of DNA molecules are able to form holographic pre-images of biostructures. This process of "reading and writing" the very matter of our being manifests from the genome's associative holographic memory in conjunction with its quantum nonlocality. Rapid transmission of genetic information and gene-expression unite the organism as holistic entity embedded in the larger Whole. The system works as a biocomputer -- a **wave biocomputer**.

The quantum nonlocality of the genetic information is fundamental. Experimental work of the Gariaev group shows how quantum nonlocality is directly related to laser radiation from chromosomes (coherent light), which jitterbugs its polarization plane to radiate or occlude photons. DNA and the genome have now been identified as active "laser-like" environments. Roughly speaking, DNA can be considered as a liquid crystal gel-like state that acts on the incoming light in the manner of a solitonic lattice. **[5]**

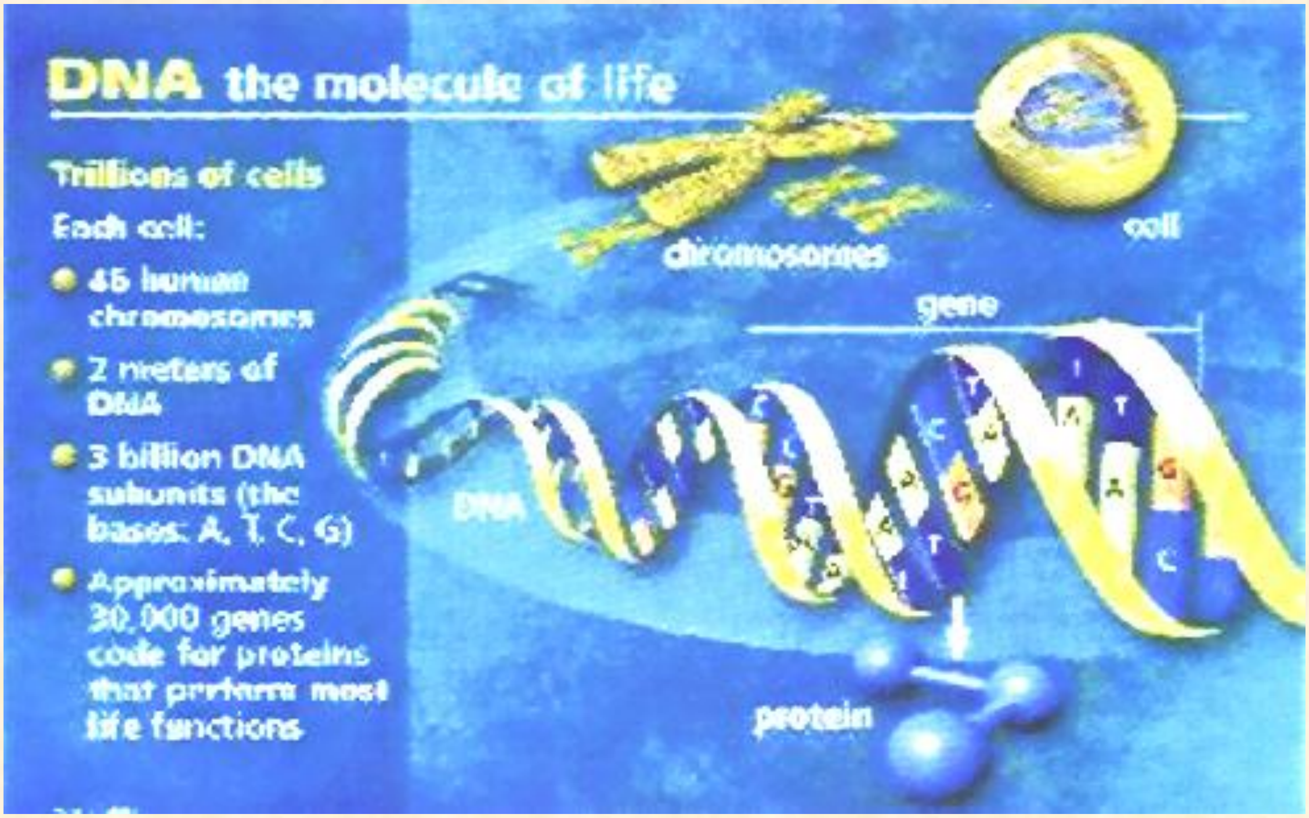
DNA the molecule of life
Trillions of cells
Each cell:
46 human chromosomes
2 meters of DNA
3 billion DNA subunits (the bases: A, T, C, G)
Approximately 30,000 genes code for proteins that perform most life functions
chromosomes
cell
gene
DNA
protein

Biophotonics

This entire section is extracted from the History of Biophotonics website:

Around 1923, Alexander Gurwitsch discovers an "ultra-weak" photon emission from living systems (onions, yeast, ...), since he suggested connections between photon emission and cell division rate. He calls this **photonemission "mitogenetic radiation**". His experiments indicate that the wavelength is in the range around 260 nm (Bibliography under Gurwisch and also Ruth (1977, 1979)).

Around 1950, Russian scientists rediscover "ultra-weak photon emission" from living organisms. Most results are published in "Biophysics" (Engl.) and originally in "Biofizika"). (Bibliography under Ruth, 1979).

Italian nuclear physicists discover by chance "bioluminescence" of seedlings. They do not think that this finding is significant, but they publish the results (Colli et al. 1954, 1955, Ruth 1979).

The Russian biophysicist and the American chemist enunciate the first theory of ultra-weak photonemission (UWPE) from biological systems -- the so called "**imperfection theory**". UWPE shall be an expression of the deviation from equilibrium -- some kind of distortion of metabolic processes (Zhuravlev 1972, Seliger 1975, Ruth 1979).

Independently from each other and by different motivations, scientific groups in Australia (Quickenden), Germany (Fritz-Albert Popp), Japan (Inaba), and Poland (Slawinski) show evidence of ultra-weak photon emission from biological systems by use of modern single-photon counting systems. Bibliography (Quickenden, Inaba, Popp &Ruth, Slawinski).

While Quickenden, Slawinski, and Inaba prefer the imperfection theory, Popp and his group enunciate just the opposite theory:

1. The radiation originates from an almost perfect coherent photon field.
2. Essential sources are the DNA and corresponding resonators in the cells.
3. The mechanism describes photon storage in cavities and information channels, tuned by Casimir forces.
4. There is a close connection to delayed luminescence which corresponds to excited states of the coherent photon field.
5. The radiation is not the product but essentially the initiator of chemical reactions in the cells. The radiation submits the information within and between cells.
6. The radiation is not limited to the optical range but follows an **f = const** rule (the occupation probability of the phase space is equal for all wavelengths) and extends to longer wavelengths including the so-called heat radiation of the body.
7. This radiation is the proper regulator and information carrier of life.

The Marburg group of Fritz-Albert Popp calls this phenomenon "**biophotons**" in order to stress the difference to "bioluminescence". Biophotons are single quanta which are permanently-and-continuously emitted by all living systems. They are subjects of quantum physics, and they display a universal phenomenon attributed to all living systems. Worldwide, all scientists who agree with these statements call the radiation "biophotons" and the scientific field "**biophotonics**".

From 1972 to 1980, the Marburg group of the leader -- the physicist and Dr. Habil Fritz-Albert Popp -- evaluated experimentally all the essential physical properties of biophotons.

1. The intensity ranges from a few up to some hundreds photons/(sec-cm^2).
2. The spectral distribution follows in the time average an **f = const** rule.
3. The modes are strongly coupled.
4. The delayed luminescence that approaches continuously the biophoton emission follows a hyperbolic rather than an exponential relaxation function.
5. The biophotons origin from an almost fully coherent field.
6. Cells are able to establish cavity resonators which contribute to biophoton regulation.

7. The essential source of non-equilibrium biophoton emission is the DNA.

This group introduces the first time photocount statistics (PCS) into biophotonics. They show evidence that biophotons are emitted according to a Poissonian PCS. Furthermore, they show evidence that (1) the delayed luminescence follows a hyperbolic relaxation function rather than an exponential one; (2) the modes are strongly coupled; and (3) there are hyperbolic oscillations around the continuous hyperbolic relaxation function. The group finds the first time intercellular communication by means of biophotons. Later this was confirmed by Albrecht-Bühler (Bacteria), Popp and Chang (dinoflagellates), Galle (daphnia), Shen (blood), and Vogel (bacteria).

(Bibliography under Popp and coworkers, i.e., Bahr, Böhm, Grass, Grolig, Herrmann, Kramer, Rattemeyer, Ruth, Schmidt, Wulle, Albrecht-Bühler, Chang, Galle, Shen, Vogel).

The papers of Popp and his group are examined mainly by the group of B. Chwirot (Kopernikus University, Torun). They confirm the essential results (Bibliography Chwirot et al.).

Herbert Klima (Atom Institute Vienna) performs his dissertation in Popp`s group at the University in Marburg. He transfers "Biophotonics" to the University in Vienna -- in particular investigations on laser excitation of living systems.

(Bibliography Klima or Atominstitut Wien).

J.Slawinski cooperates with the groups in Japan, USA and the Popp-group in Germany. He follows mainly the links between biophotons and biochemical reactions. There arises a branch biochemical biophotonics that becomes an essential part mainly in Japan and USA." **[6]**

Bioinfotons

This term "**bioinfoton**" is one that I am adopting to specify any particle (such as the **Ganesh Particle***), wave, or biomolecule that transfers information from an emitter to a receptor. It is possible that if a biophoton is emitted from a coherent source that it will carry holographic information as a form of modulation (possibly frequency-modulated) and produce action or results in biological systems.

It is possible that biophotons may even be carriers of ***psi*** information and that a coherent coupling can be established between 2 conscious lifeforms resulting in a transference of information from a higher potential field to a field at lower potential. Even though this is conjecture at this point, it suggests other means by which disparate lifeforms can communicate. We know there is a process by which trees communicate and even signal each other in the face of danger. We may be exchanging information with pets and other animals through biophotonic communication. This may be how some people who have a "green thumb" affect plants they care for.

Even communication with other intelligent life forms in the Cosmos may be effected by the amplification and vectoring of biophotons or -- via the quantum potential -- produce information at a distance.

A virtual photon flux in the quantum potential may be integrating the coordinated wholeness of the organism while the action of bioinfotons may keep every part of the organism in communication with every other part.

Conclusions

Life and living organisms are more than the sum of their parts and are composed of material and non-material parts. The spiral form of DNA evidently produces an electromagnetic wave field that informs the organism's growth and structure.

Scientist Richard Alan Miller is a pioneer in this field and has written:

"In a hologram, wave fields interfere with one another to lay the foundations for the reconstruction of the image of an object. But how are the wave fields produced?

The term "holography" comes from the Greek roots meaning "entire" and "to write". In holography, the image is projected by a coherent light source split into both the object wave and the reference wave background. This dichotomous nature is reflected in the particle/wave nature of the DNA molecule, which can be "read out" with biophotons from chromosomes to set up a holographically-produced wave field. This superposition of wave fields (object wave and reference wave) creates a wave guide for the formation of biological structure. The image is constructed according to the reference information contained in the genes. The reconstructed object wave is identical with the object wave field. The reconstructed wave fields reproduce exactly the recorded ones (the DNA with genetic code)." **[7]**

Mr. Miller also proposes this hypothesis:

"QUANTUM BIOHOLOGRAPHY

Hypothesis: The organization of any biological system is established by a complex electrodynamical field that is, in part, determined by its atomic physiochemical components. These, in part, determine the behavior and orientation of these components. This dynamic is mediated through wave-based genomes wherein DNA functions as the holographic projector of the psychophysical system -- a quantum biohologram."

I propose that the biohologram is embedded as an integral part of the biocosmic hologram we call the"Universe".

Dr. Dan Burisch has seen evidence that a **Ganesh Particle*** can transmute itself into a fully functional atypical eukaryotic cell in a matter of seconds -- not eons. Such evidence -- when it becomes available -- will change our entire perspective on the existence of a universal holonomic mind that informs its creation.

**Ganesh Particles are charged clusters that extracted from ancient bedrock and have a seemingly unknown origin, but carry information which is transferred to cells via a crossbridge. (discovered by Dr. Dan B.C. Burisch)*

References:

1. **http://plato.stanford.edu/entries/qm-bohm/#qp**
2. **http://www.ksharpe.com/Word/BM05.htm**

3. http://www.aip.org/enews/physnews/2001/split/566-1.html
4. http://www.rialian.com/rnboyd/dna-wave.doc
5. http://www.geocities.com/iona_m/Chaosophy4/bioholography/bioholography.htm
6. http://www.lifescientists.de/history.htm
7. http://www.nwbotanicals.org/oak/newphysics/Helix%20to%20Hologram.pdf

if on the Internet, Press <BACK> on your browser to return to the previous page (or go to www.stealthskater.com)

else if accessing these files from the CD in a MS-Word session, simply <CLOSE> this file's window-session; the previous window-session should still remain 'active'

The Tom Bearden Website

Help support the research

Foundations of Electro-Magnetism
Slide Series

- Major Points of Presentation
- Chung's Carbon Filament Negative Resistor
- Aharanov-Bohm Effect
- Absorption and emission reactions (1)
- Absorption and emission reactions (2)
- Anti-engine for cell's deviation reverses cell back to normal
- Asymmetric regauging produces excess force, which can be used to do work on the system
- Asymmetric regauging produces excess force, which can be used to do work on the regauging system
- Becker's bone fracture healing experiments
- Becker's theoretical control system governing regeneration
- Becker's theoretical DC control system involved with response to injury
- Geometrical Distortion of Poynting Energy Flow
- Cosmic Bootstrap: Before the Big Bang
- Bridging Enables COP > 1

- Aspects of Strong Local Symmetry
- Other Aspects of Strong Local Symmetry
- Bunge on the Status of Electrodynamics
- Bunge on the Status of Electrodynamics and Physics
- Bunge on the Status of Classical and Quantum Physics
- Fractions of BW Casualties Surviving
- First Step in Self-Promotion of Cancer Due to Hypoxia
- Differentiation and Dedifferentiation Timelines
- Evolution of Cells on Earth
- A Charged Particle q is a Coupled System of m and Ø
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (1)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (2)
- The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry (3)
- A Charged Particle is a Coupled System
- The Anomalies in Navy Electrolyte Experiments at China Lake
- Chung's Carbon Filament Negative Resistor

- Typical circuit has about 10^{-13} energy collection efficiency
- Classical View of EM
- Curved Spacetime acts as source or sink
- Chronic Disease Puzzling
- Cancer Characteristics
- Solving a Cancer Enigma
- Cancer: Characteristics
- Nonlinear Optics Distortion Correction Theorem
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)
- The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)
- In Time-Reversed Zone: Nuclear reactions are biased
- In Time-Reversed Zone: Dynamics may be reversed
- Time-Reversed Zone: Significant only after time-density charging
- A New Conservation of Energy Law
- Questions on Carcinogens and EM

Radiation

- AS DEFINED, fields and potentials only exist in and on charged matter
- Serious flaws and errors in classical EM theory
- Electrodynamicists' Reaction to Removal of the Material Ether
- Reaction to Loss of Material Ether (1)
- Reaction to Loss of Material Ether (2)
- Some EM waves in Spacetime
- Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String
- Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna
- Phase Conjugate Wavepairs Produce New Waves
- Scalar Electromagnetics (Energetics) View of EM (1)
- Scalar Electromagnetics (Energetics) View of EM (2)

- Scalar Electromagnetics (Energetics) View of EM (3)
- Extended General Relativity Principle
- Extension of Work-Energy Theorem
- 34 Flaws in Classical EM Theory
- MTW's General Relativity Principle
- Use of General Relativity (GR) in Particle Physics
- The Missing Infolded Electrodynamics
- Infolded Engines: Excluded by Present Physics
- Mass curves and structures spacetime; ST curvatures change and structure mass (1)
- Mass curves and structures spacetime; ST curvatures change and structure mass (2)
- At Infinite Velocity: Each Is Every Other
- Lisitsyn's Report: Brain Code Broken
- Lorentz Closed Surface Integration
- Lorentz's physically insignificant energy flow can be collected and utilized
- Lorentz surface integration of the Poynting vector around a closed surface
- Longitudinal EM Photon Interaction with Charge
- Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass.

- Pumping with Longitudinal EM waves Time-reverses the Mass
- Maxwell's Quaternion Theory
- Maxwellian Systems before and after Regauging
- Maxwell's four equations reduce to: Page 1
- Maxwell's four equations reduce to: Page 2
- More to come

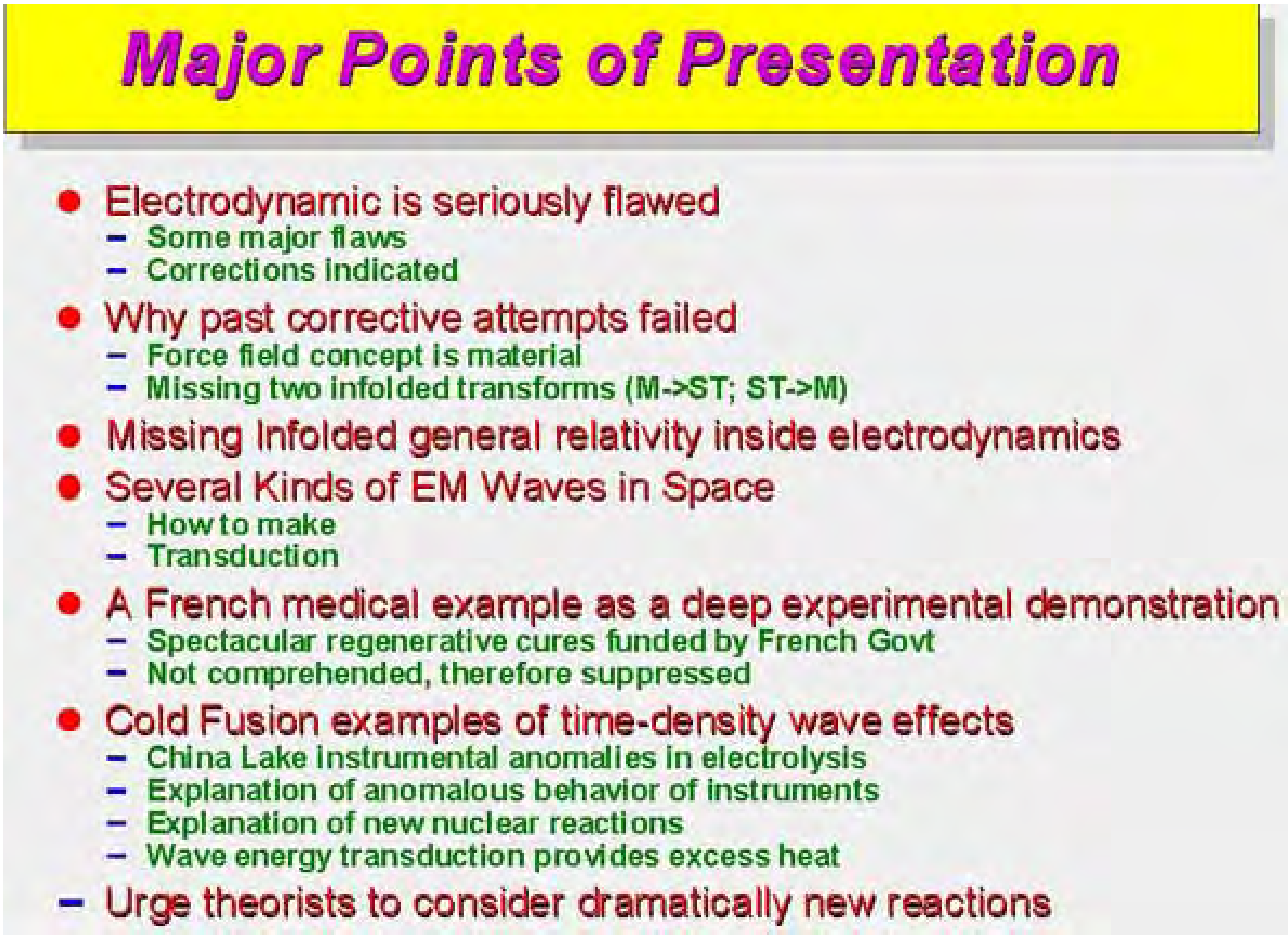

Major Points of Presentation

- Electrodynamic is seriously flawed
  - Some major flaws
  - Corrections indicated
- Why past corrective attempts failed
  - Force field concept is material
  - Missing two infolded transforms (M->ST; ST->M)
- Missing infolded general relativity inside electrodynamics
- Several Kinds of EM Waves in Space
  - How to make
  - Transduction
- A French medical example as a deep experimental demonstration
  - Spectacular regenerative cures funded by French Govt
  - Not comprehended, therefore suppressed
- Cold Fusion examples of time-density wave effects
  - China Lake instrumental anomalies in electrolysis
  - Explanation of anomalous behavior of instruments
  - Explanation of new nuclear reactions
  - Wave energy transduction provides excess heat
- Urge theorists to consider dramatically new reactions

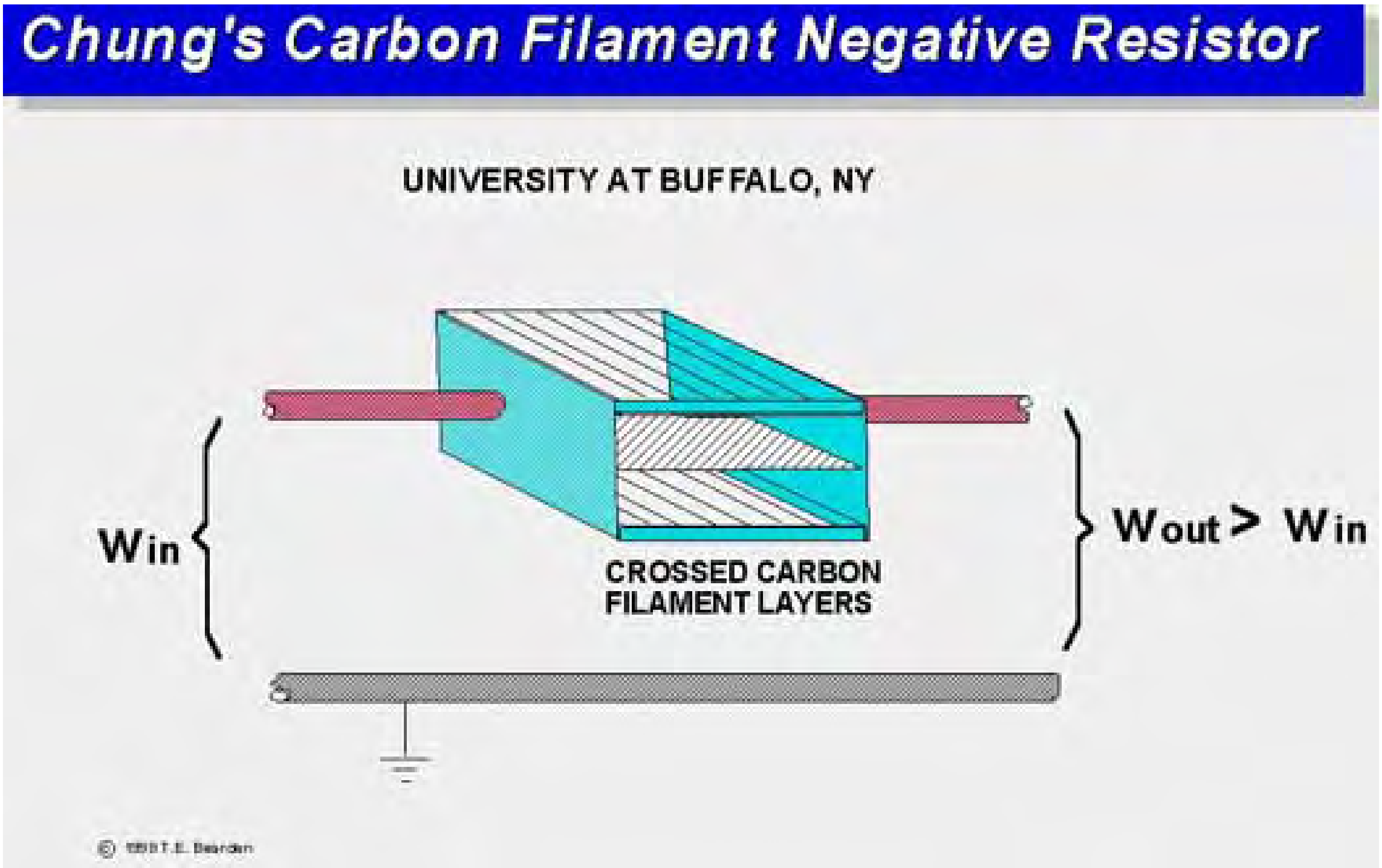

Chung's Carbon Filament Negative Resistor

UNIVERSITY AT BUFFALO, NY

Win
CROSSED CARBON
FILAMENT LAYERS
Wout > Win

© [illegible] T.E. Bearden

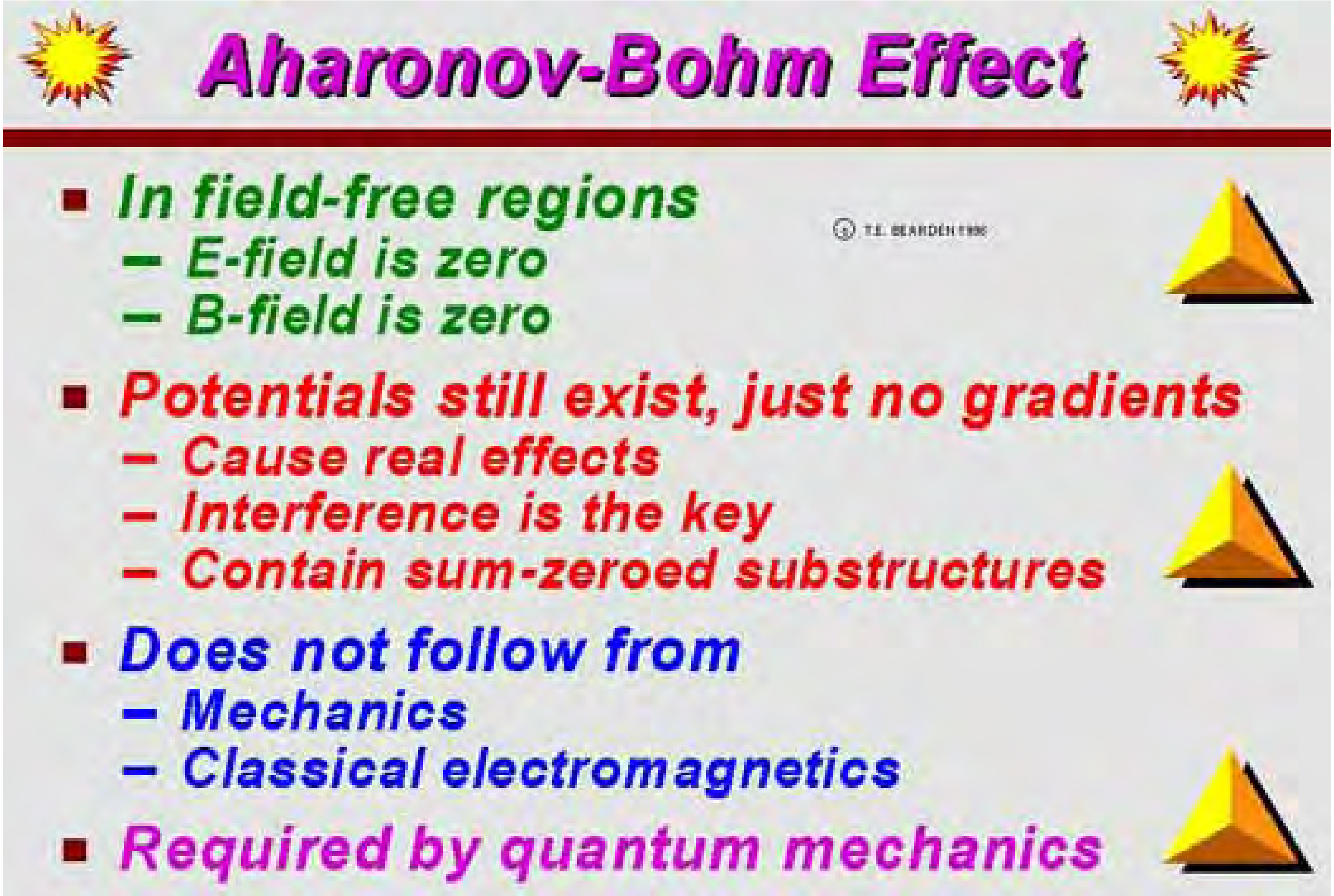

Aharonov-Bohm Effect

© T.E. BEARDEN 1996

- In field-free regions
  - E-field is zero
  - B-field is zero
- Potentials still exist, just no gradients
  - Cause real effects
  - Interference is the key
  - Contain sum-zeroed substructures
- Does not follow from
  - Mechanics
  - Classical electromagnetics
- Required by quantum mechanics

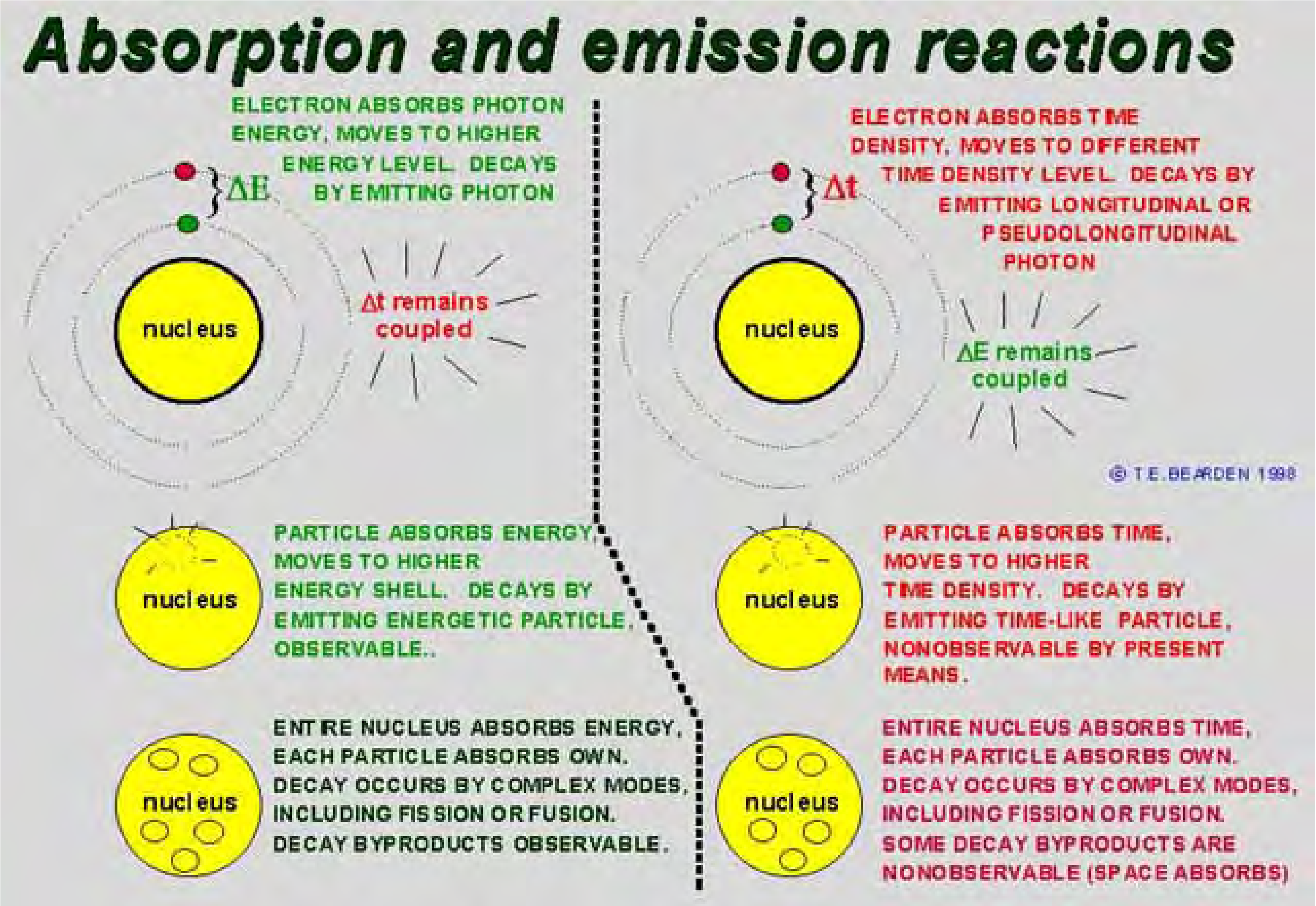

Absorption and emission reactions

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON

ELECTRON ABSORBS TIME DENSITY, MOVES TO DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTON

© T.E. BEARDEN 1998

PARTICLE ABSORBS ENERGY, MOVES TO HIGHER ENERGY SHELL. DECAYS BY EMITTING ENERGETIC PARTICLE. OBSERVABLE..

PARTICLE ABSORBS TIME, MOVES TO HIGHER TIME DENSITY. DECAYS BY EMITTING TIME-LIKE PARTICLE, NONOBSERVABLE BY PRESENT MEANS.

ENTIRE NUCLEUS ABSORBS ENERGY, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. DECAY BYPRODUCTS OBSERVABLE.

ENTIRE NUCLEUS ABSORBS TIME, EACH PARTICLE ABSORBS OWN. DECAY OCCURS BY COMPLEX MODES, INCLUDING FISSION OR FUSION. SOME DECAY BYPRODUCTS ARE NONOBSERVABLE (SPACE ABSORBS)

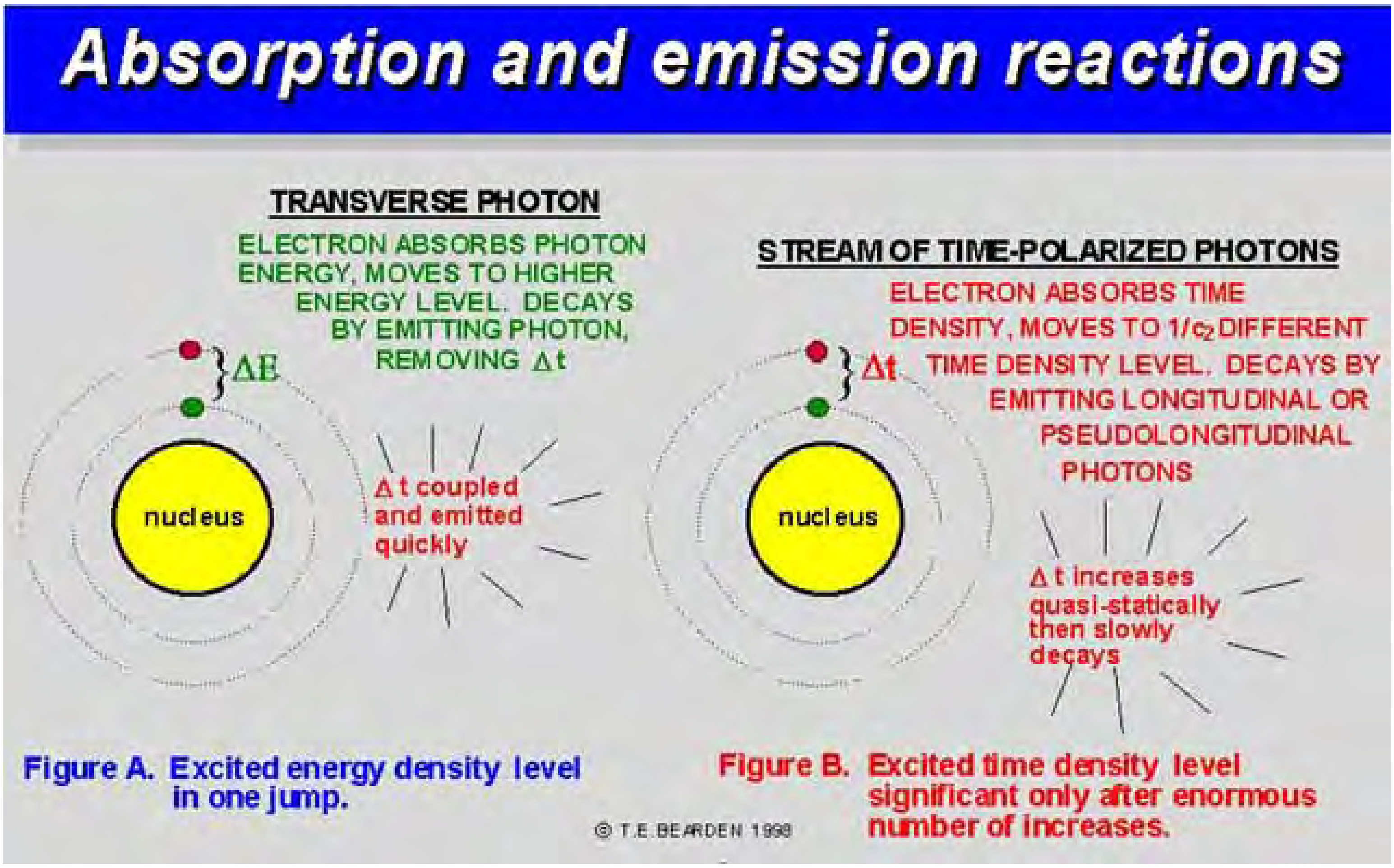

Absorption and emission reactions

TRANSVERSE PHOTON

ELECTRON ABSORBS PHOTON ENERGY, MOVES TO HIGHER ENERGY LEVEL. DECAYS BY EMITTING PHOTON, REMOVING Δt

ΔE
nucleus
Δt coupled and emitted quickly

Figure A. Excited energy density level in one jump.

STREAM OF TIME-POLARIZED PHOTONS

ELECTRON ABSORBS TIME DENSITY, MOVES TO $1/c_2$ DIFFERENT TIME DENSITY LEVEL. DECAYS BY EMITTING LONGITUDINAL OR PSEUDOLONGITUDINAL PHOTONS

Δt
nucleus
Δt increases quasi-statically then slowly decays

Figure B. Excited time density level significant only after enormous number of increases.

© T.E.BEARDEN 1998

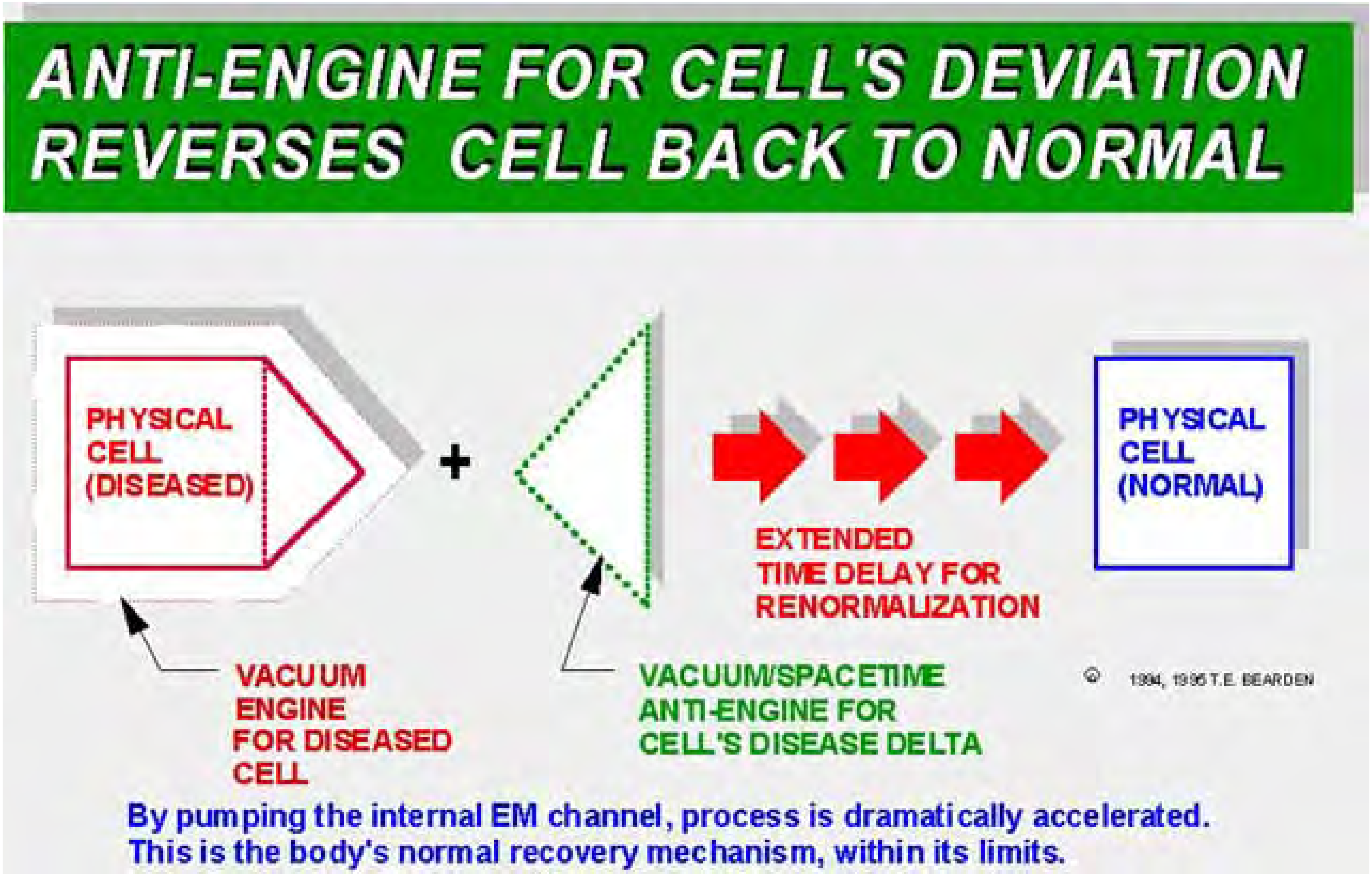

ANTI-ENGINE FOR CELL'S DEVIATION REVERSES CELL BACK TO NORMAL

PHYSICAL
CELL
(DISEASED)
+
EXTENDED
TIME DELAY FOR
RENORMALIZATION
PHYSICAL
CELL
(NORMAL)
VACUUM
ENGINE
FOR DISEASED
CELL
VACUUM/SPACETIME
ANTI-ENGINE FOR
CELL'S DISEASE DELTA
© 1994, 1995 T.E. BEARDEN

By pumping the internal EM channel, process is dramatically accelerated.
This is the body's normal recovery mechanism, within its limits.

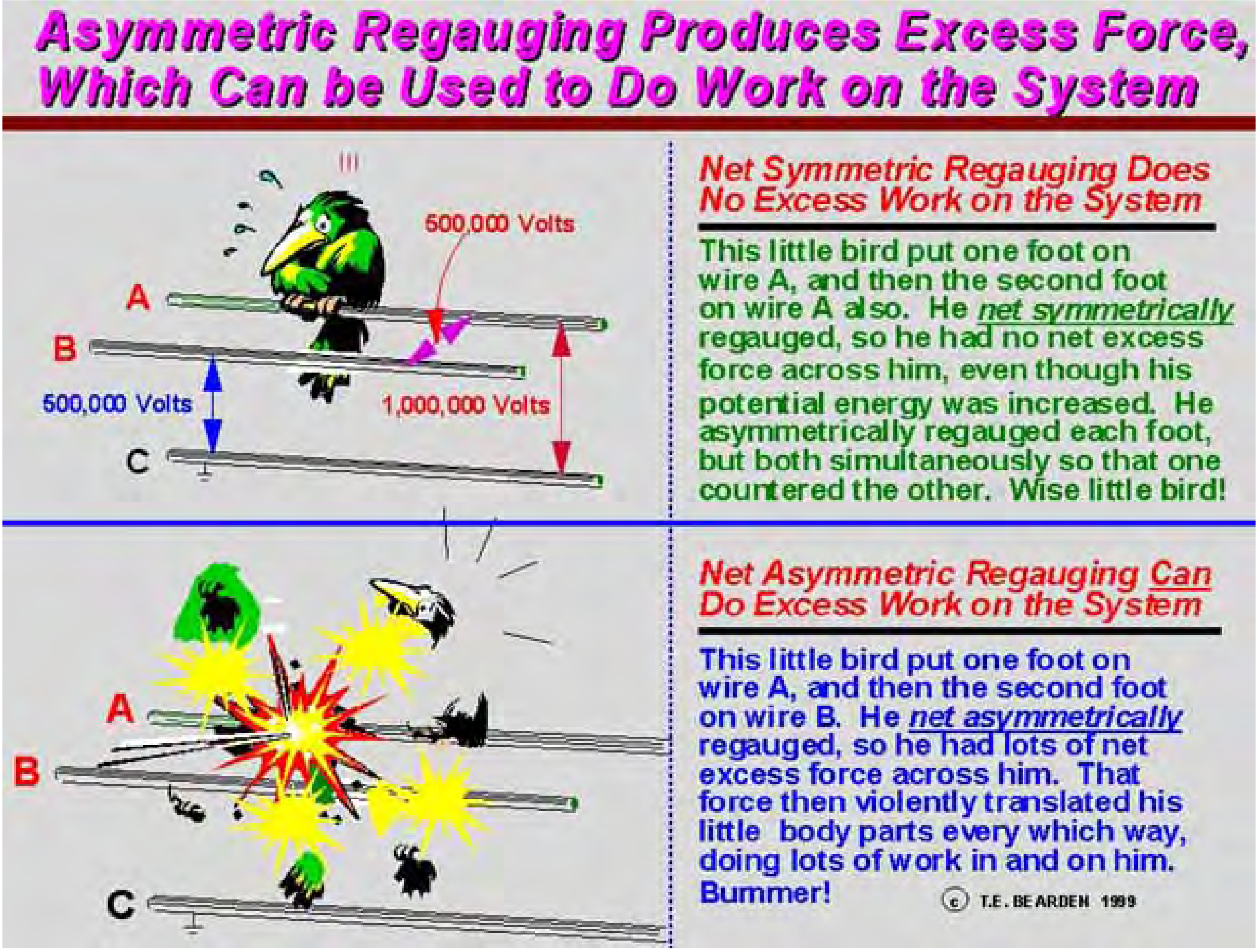

Asymmetric Regauging Produces Excess Force, Which Can be Used to Do Work on the System

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging *Can* Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

© T.E. BEARDEN 1999

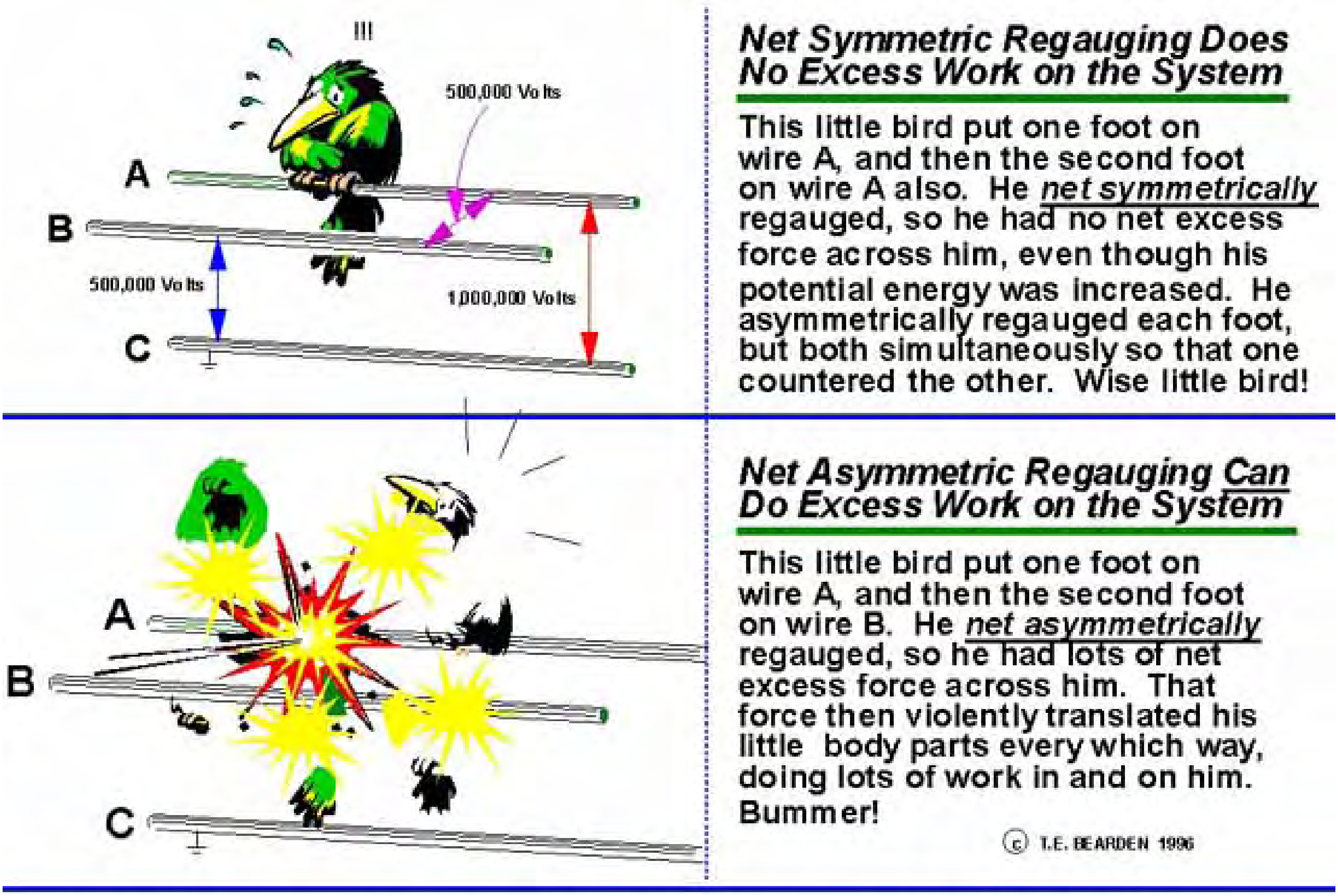

Net Symmetric Regauging Does No Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire A also. He *net symmetrically* regauged, so he had no net excess force across him, even though his potential energy was increased. He asymmetrically regauged each foot, but both simultaneously so that one countered the other. Wise little bird!

Net Asymmetric Regauging Can Do Excess Work on the System

This little bird put one foot on wire A, and then the second foot on wire B. He *net asymmetrically* regauged, so he had lots of net excess force across him. That force then violently translated his little body parts every which way, doing lots of work in and on him. Bummer!

Figure 32. Asymmetric regauging produces excess force, which can be used to do work on the regauging system.

Selected clipart © by Lotus SmartPics™

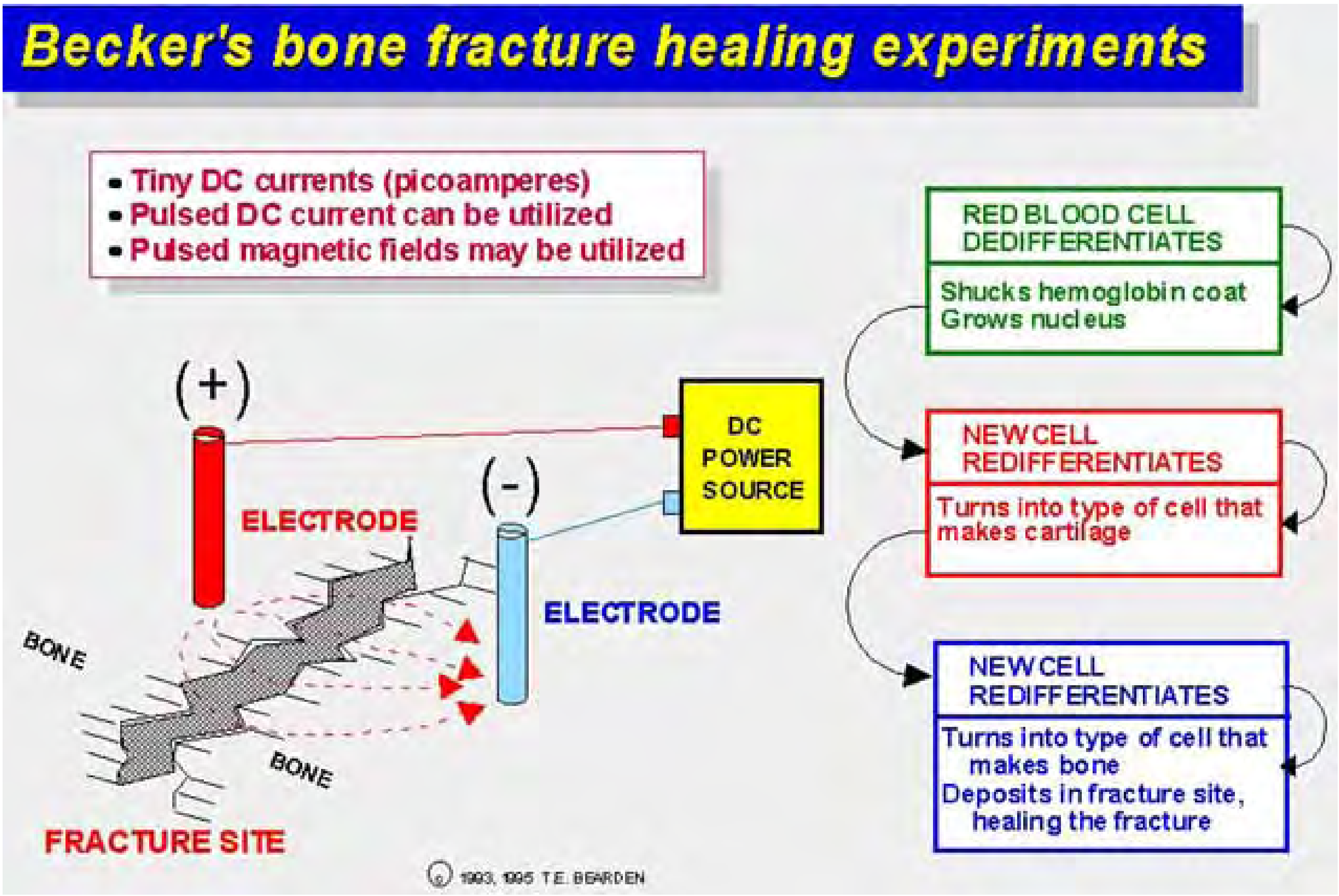

Becker's bone fracture healing experiments

- Tiny DC currents (picoamperes)
- Pulsed DC current can be utilized
- Pulsed magnetic fields may be utilized

(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1995 T.E. BEARDEN

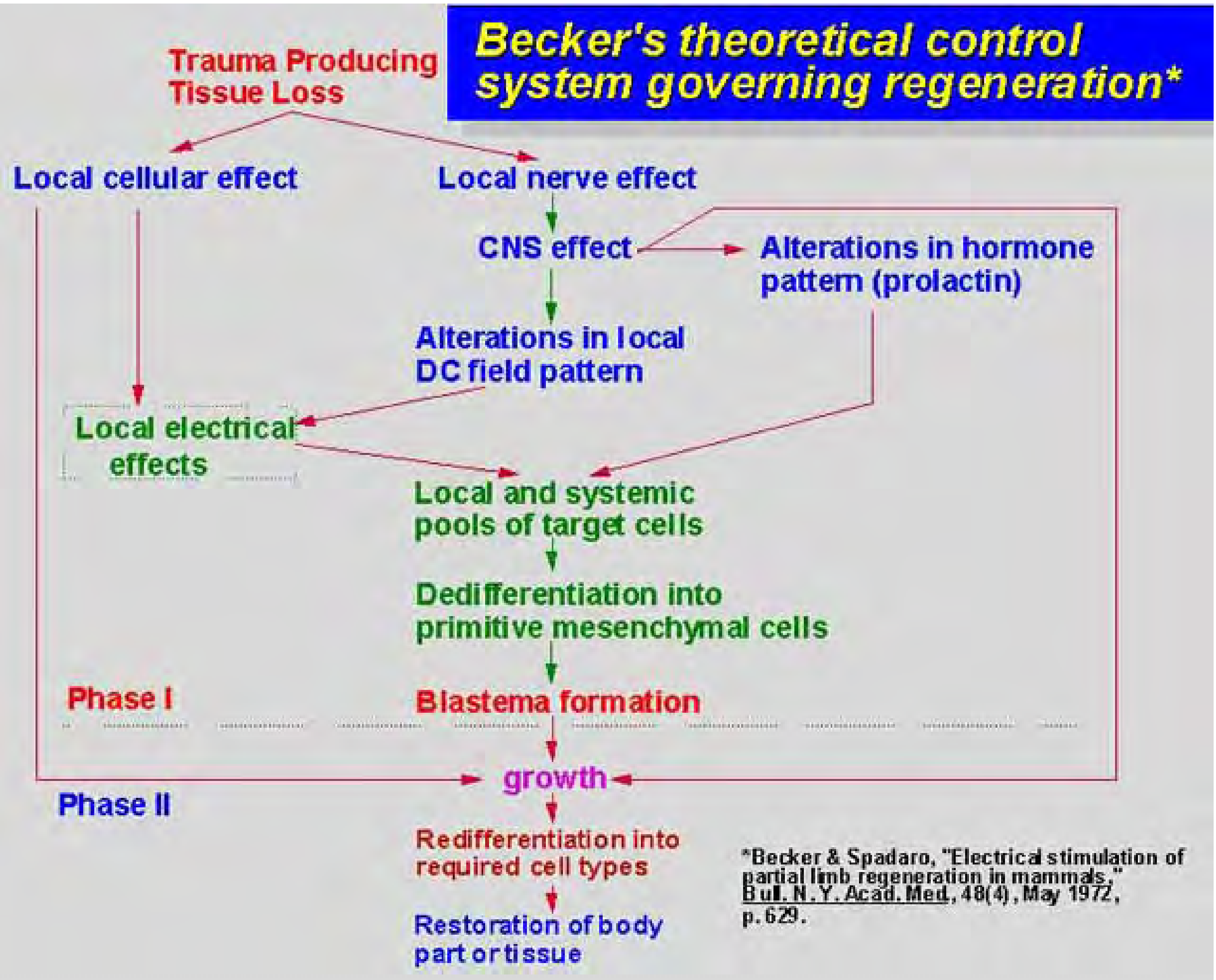

Becker's theoretical control system governing regeneration*

Trauma Producing Tissue Loss

Local cellular effect

Local nerve effect

CNS effect

Alterations in hormone pattern (prolactin)

Alterations in local DC field pattern

Local electrical effects

Local and systemic pools of target cells

Dedifferentiation into primitive mesenchymal cells

Phase I

Blastema formation

growth

Phase II

Redifferentiation into required cell types

Restoration of body part or tissue

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N.Y. Acad. Med., 48(4), May 1972, p. 629.

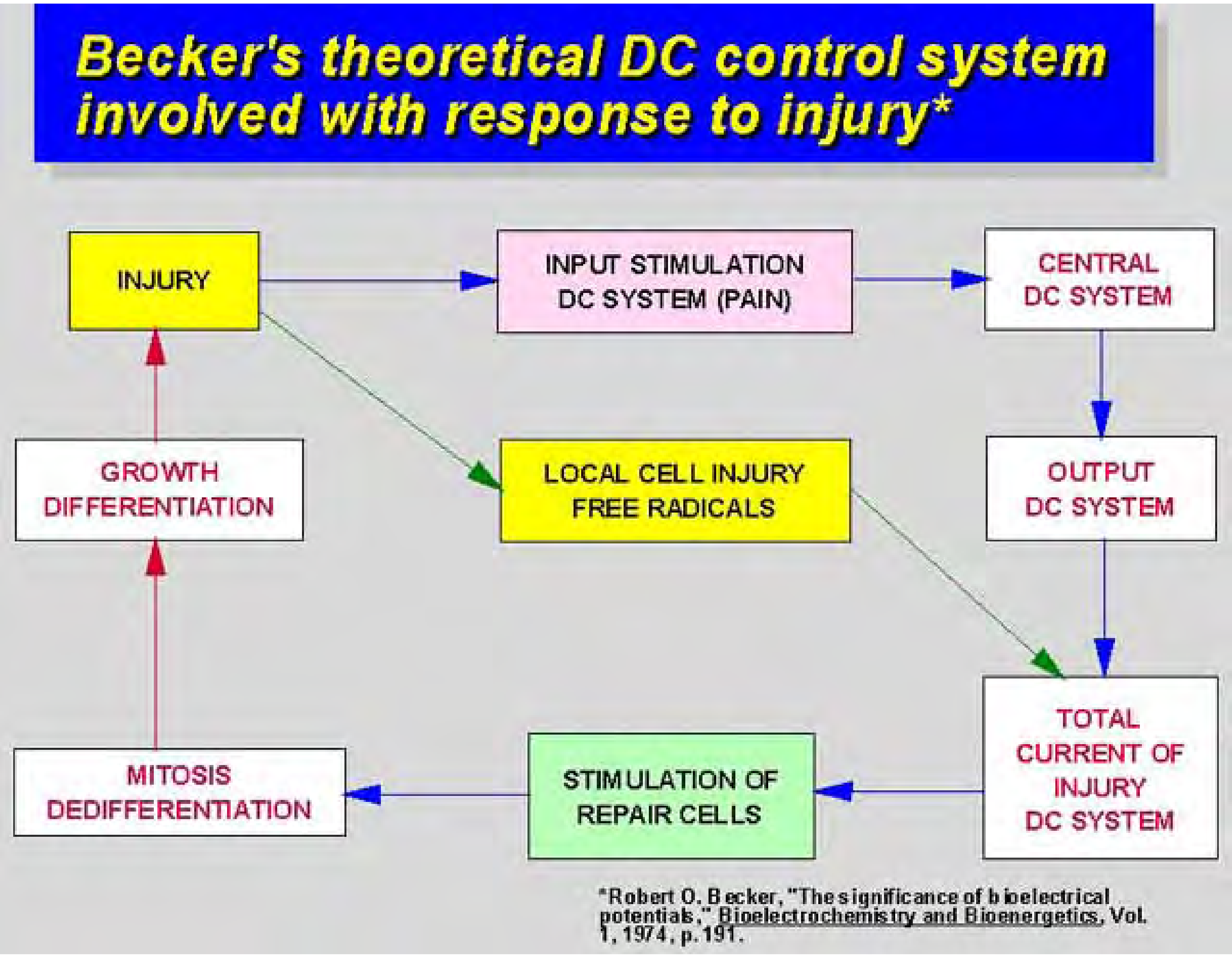

Becker's theoretical DC control system involved with response to injury*

INJURY

INPUT STIMULATION DC SYSTEM (PAIN)

CENTRAL DC SYSTEM

GROWTH DIFFERENTIATION

LOCAL CELL INJURY FREE RADICALS

OUTPUT DC SYSTEM

TOTAL CURRENT OF INJURY DC SYSTEM

MITOSIS DEDIFFERENTIATION

STIMULATION OF REPAIR CELLS

*Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

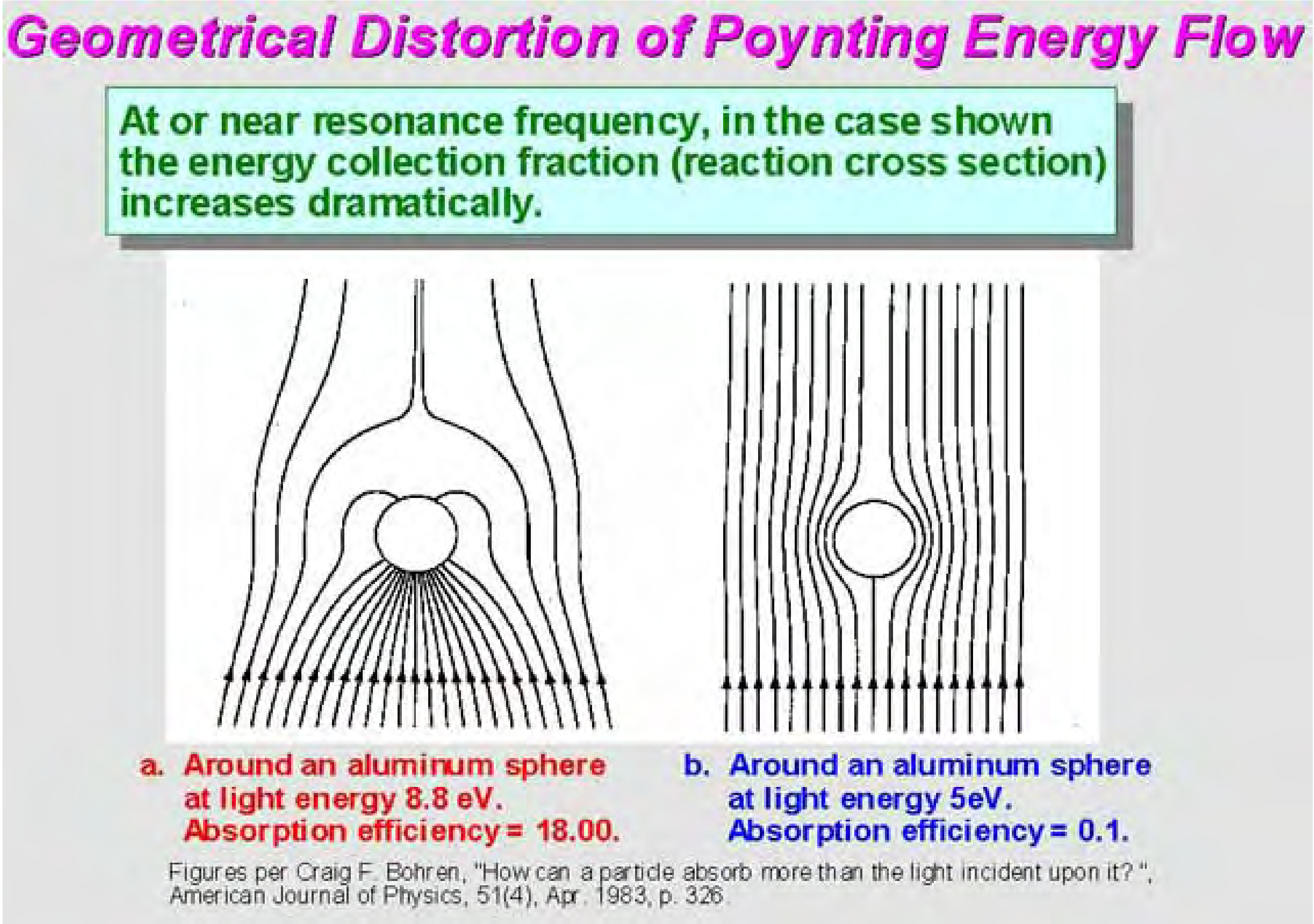

Geometrical Distortion of Poynting Energy Flow

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

a. Around an aluminum sphere at light energy 8.8 eV. Absorption efficiency = 18.00.

b. Around an aluminum sphere at light energy 5eV. Absorption efficiency = 0.1.

Figures per Craig F. Bohren, "How can a particle absorb more than the light incident upon it?", American Journal of Physics, 51(4), Apr. 1983, p. 326.

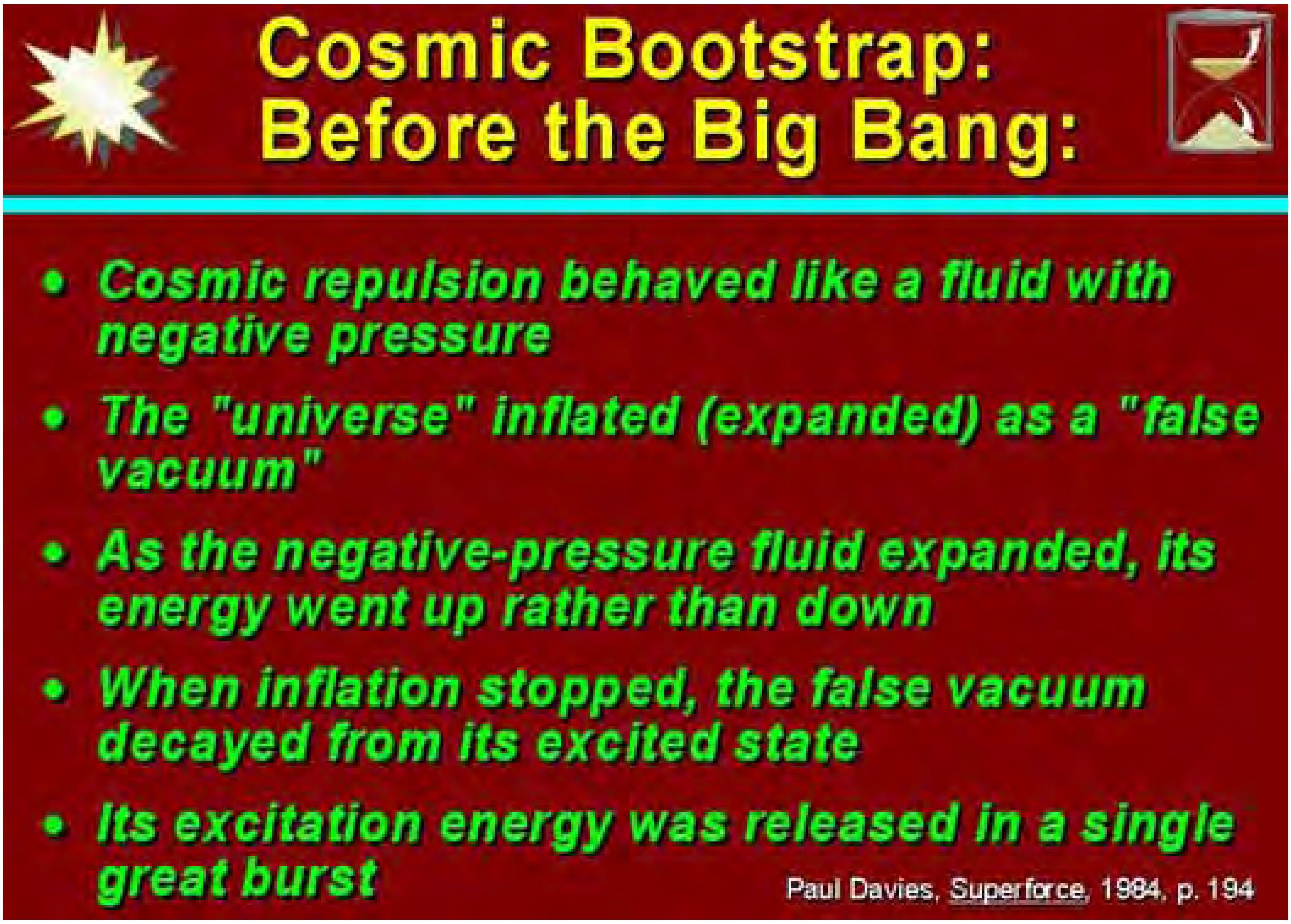

Cosmic Bootstrap: Before the Big Bang:

- Cosmic repulsion behaved like a fluid with negative pressure
- The "universe" inflated (expanded) as a "false vacuum"
- As the negative-pressure fluid expanded, its energy went up rather than down
- When inflation stopped, the false vacuum decayed from its excited state
- Its excitation energy was released in a single great burst

Paul Davies, Superforce, 1984, p. 194

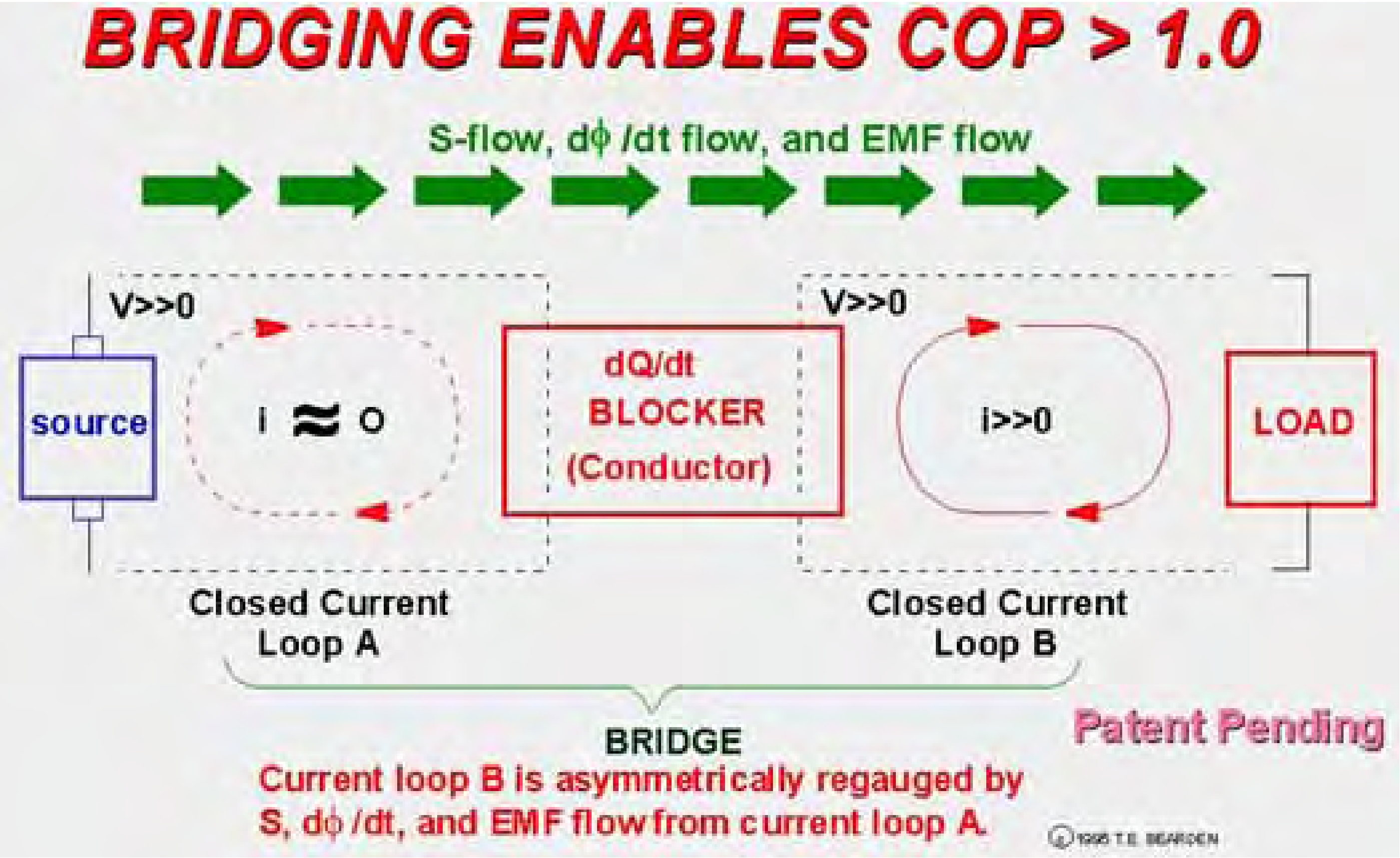

BRIDGING ENABLES COP > 1.0

S-flow, $d\phi$ /dt flow, and EMF flow

V>>0
source
$i \approx 0$
dQ/dt
BLOCKER
(Conductor)
V>>0
i>>0
LOAD

Closed Current Loop A

Closed Current Loop B

BRIDGE

Current loop B is asymmetrically regauged by S, $d\phi$ /dt, and EMF flow from current loop A.

Patent Pending

© 1998 T.E. BEARDEN

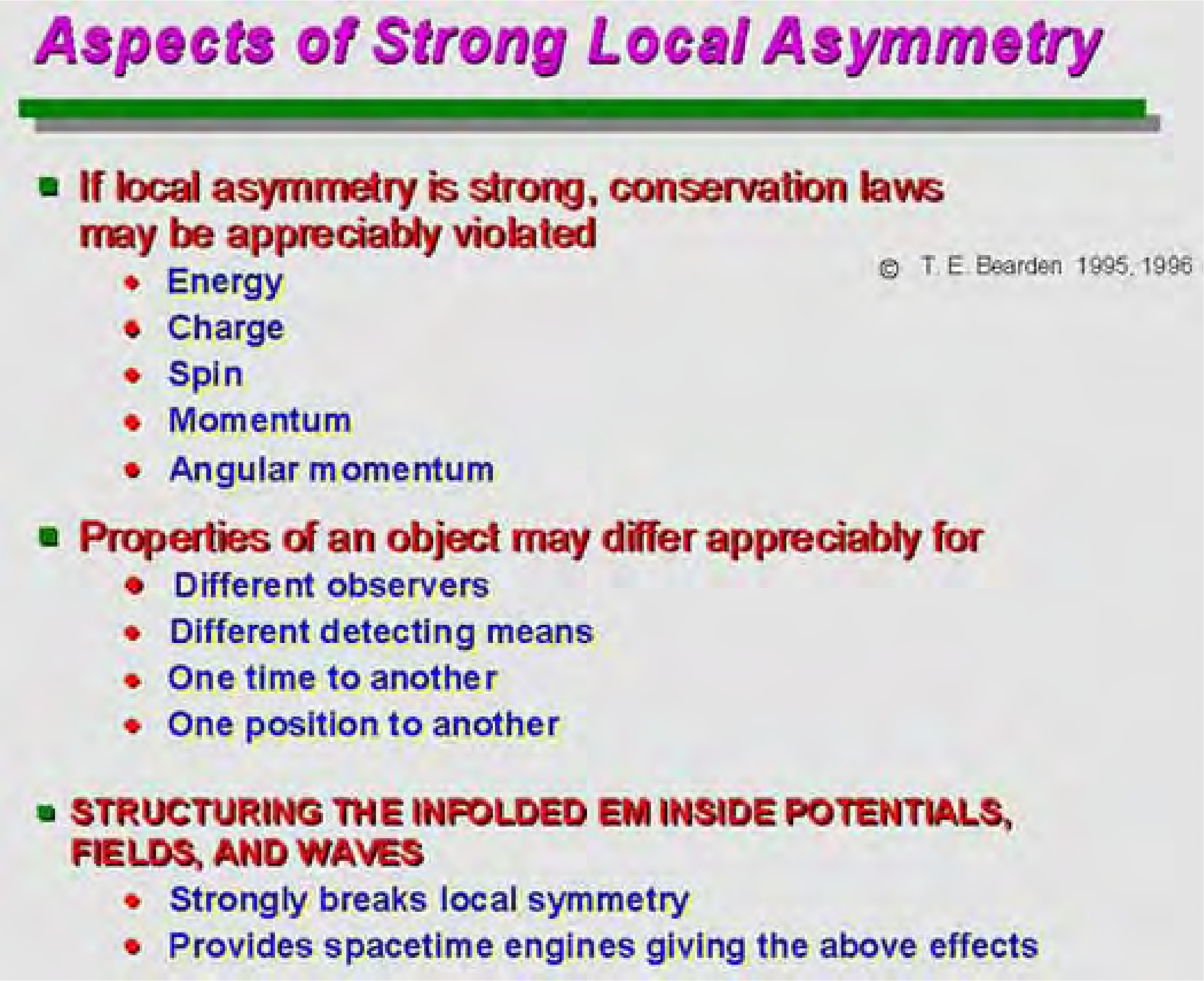

Aspects of Strong Local Asymmetry

- If local asymmetry is strong, conservation laws may be appreciably violated
  - Energy
  - Charge
  - Spin
  - Momentum
  - Angular momentum

© T. E. Bearden 1995, 1996

- Properties of an object may differ appreciably for
  - Different observers
  - Different detecting means
  - One time to another
  - One position to another
- STRUCTURING THE INFOLDED EM INSIDE POTENTIALS, FIELDS, AND WAVES
  - Strongly breaks local symmetry
  - Provides spacetime engines giving the above effects

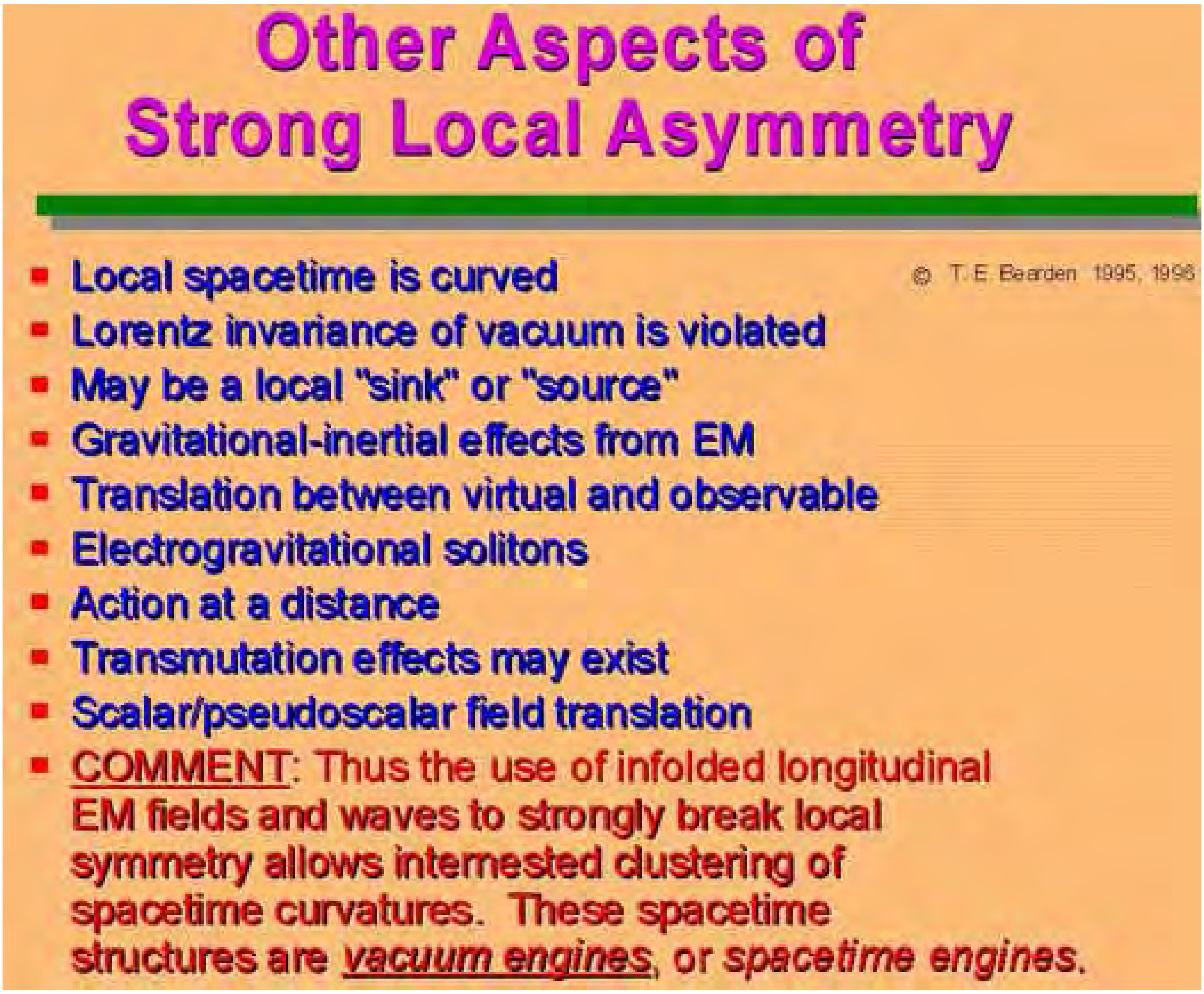

Other Aspects of Strong Local Asymmetry

© T. E. Bearden 1995, 1996

- Local spacetime is curved
- Lorentz invariance of vacuum is violated
- May be a local "sink" or "source"
- Gravitational-inertial effects from EM
- Translation between virtual and observable
- Electrogravitational solitons
- Action at a distance
- Transmutation effects may exist
- Scalar/pseudoscalar field translation
- COMMENT: Thus the use of infolded longitudinal EM fields and waves to strongly break local symmetry allows internested clustering of spacetime curvatures. These spacetime structures are *vacuum engines*, or *spacetime engines*.

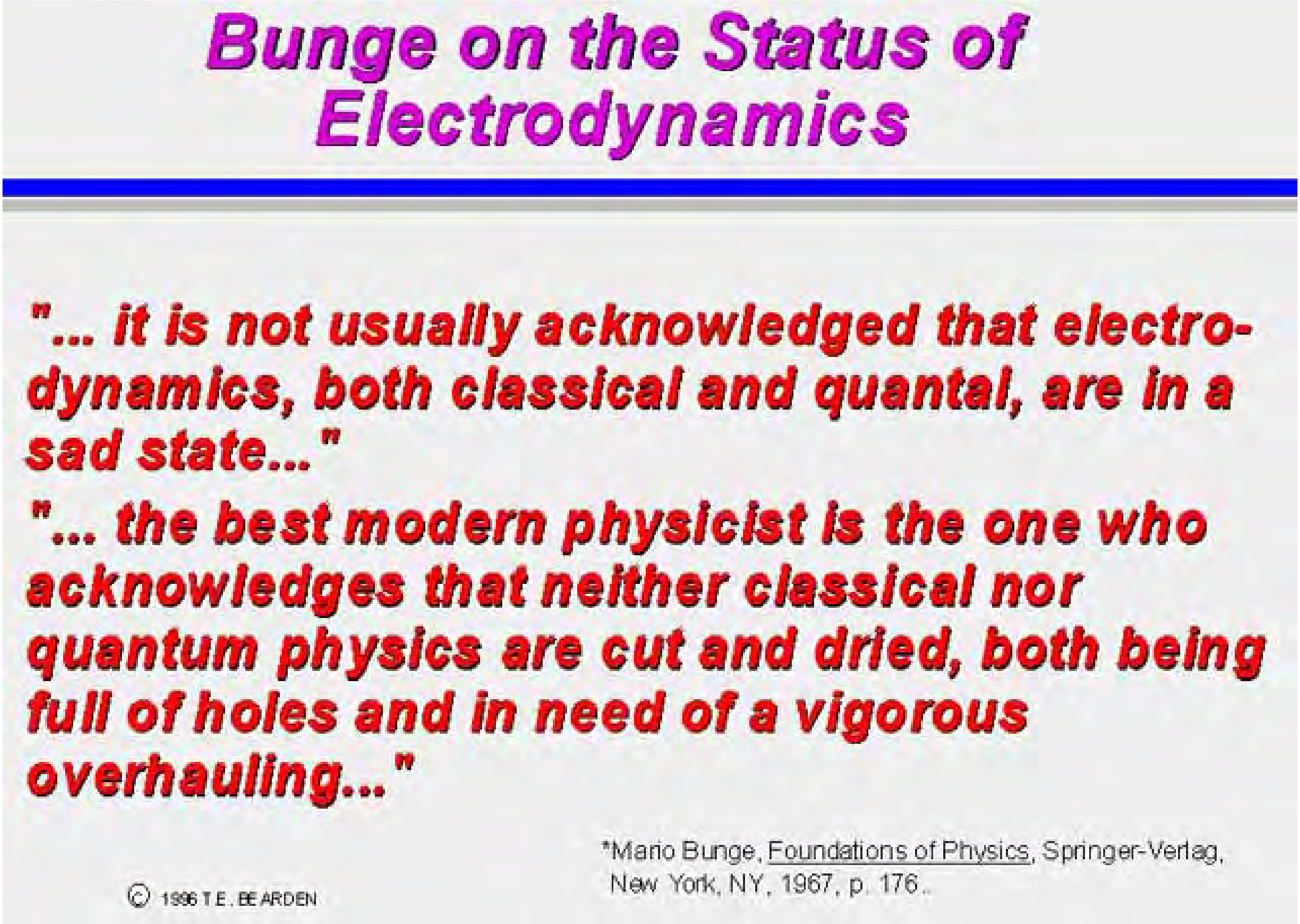

Bunge on the Status of Electrodynamics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state..."

"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

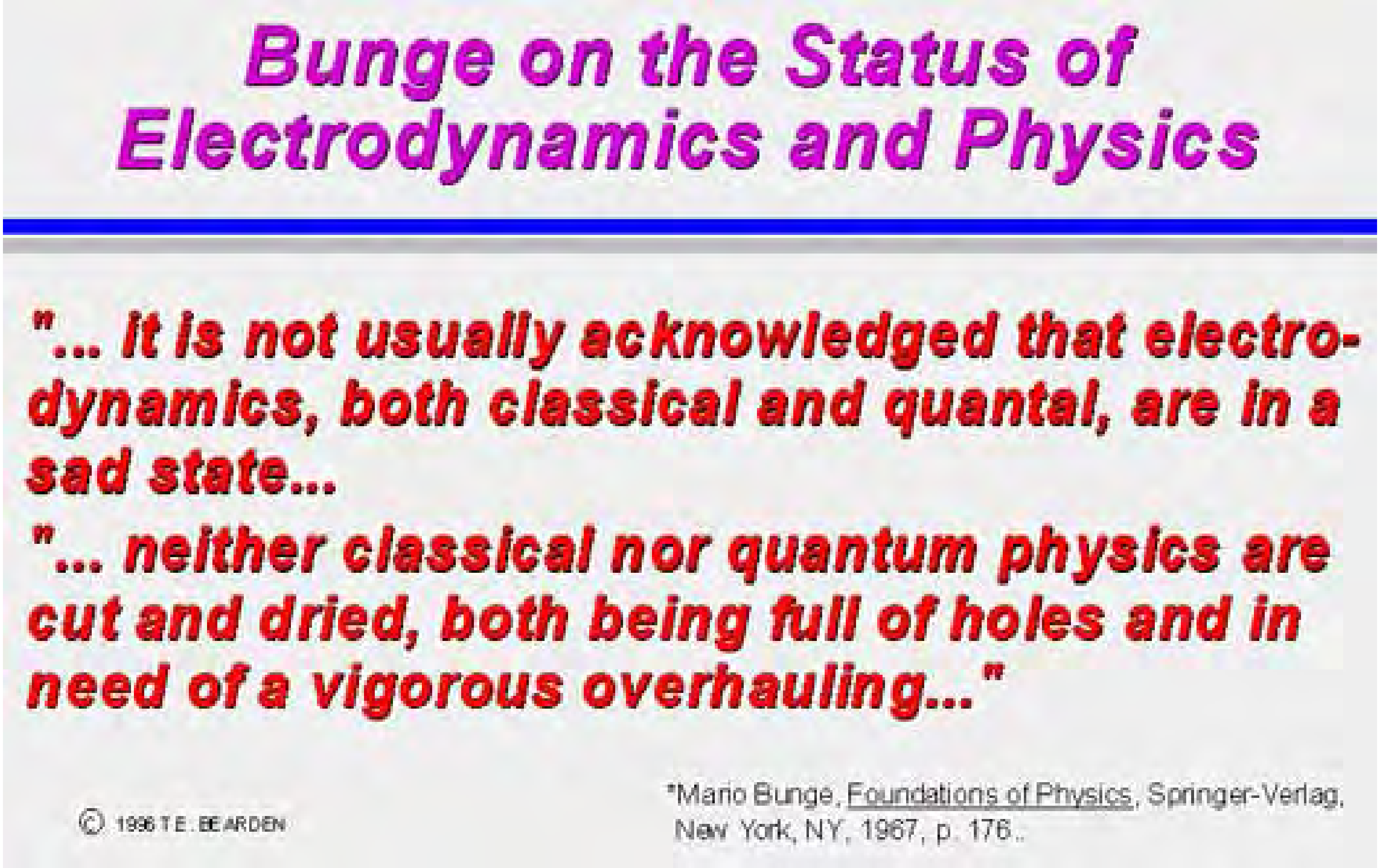

Bunge on the Status of Electrodynamics and Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state...

"... neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling..."

© 1996 T.E. BEARDEN

*Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

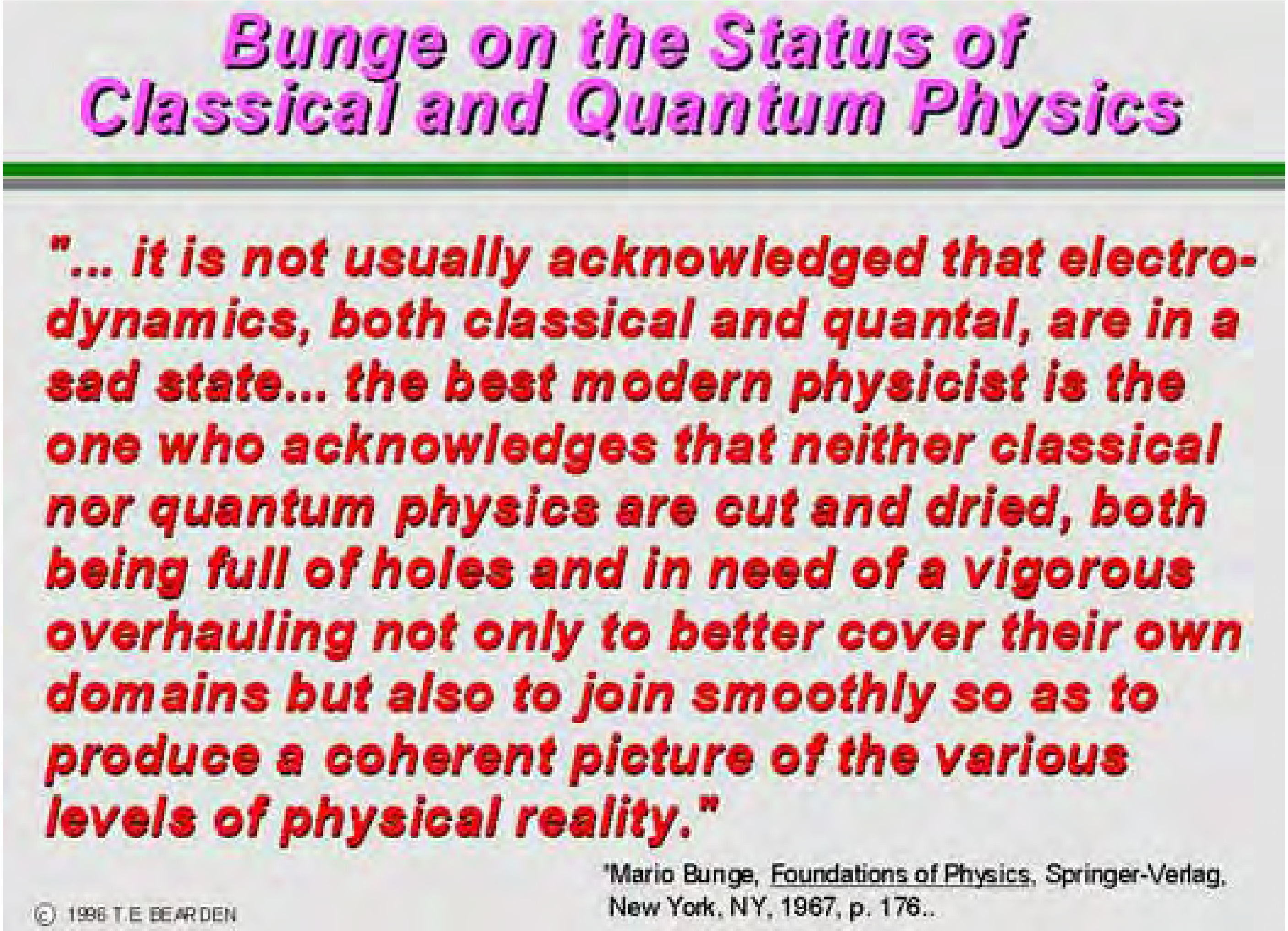

Bunge on the Status of Classical and Quantum Physics

"... it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."

'Mario Bunge, Foundations of Physics, Springer-Verlag, New York, NY, 1967, p. 176..

© 1996 T.E. BEARDEN

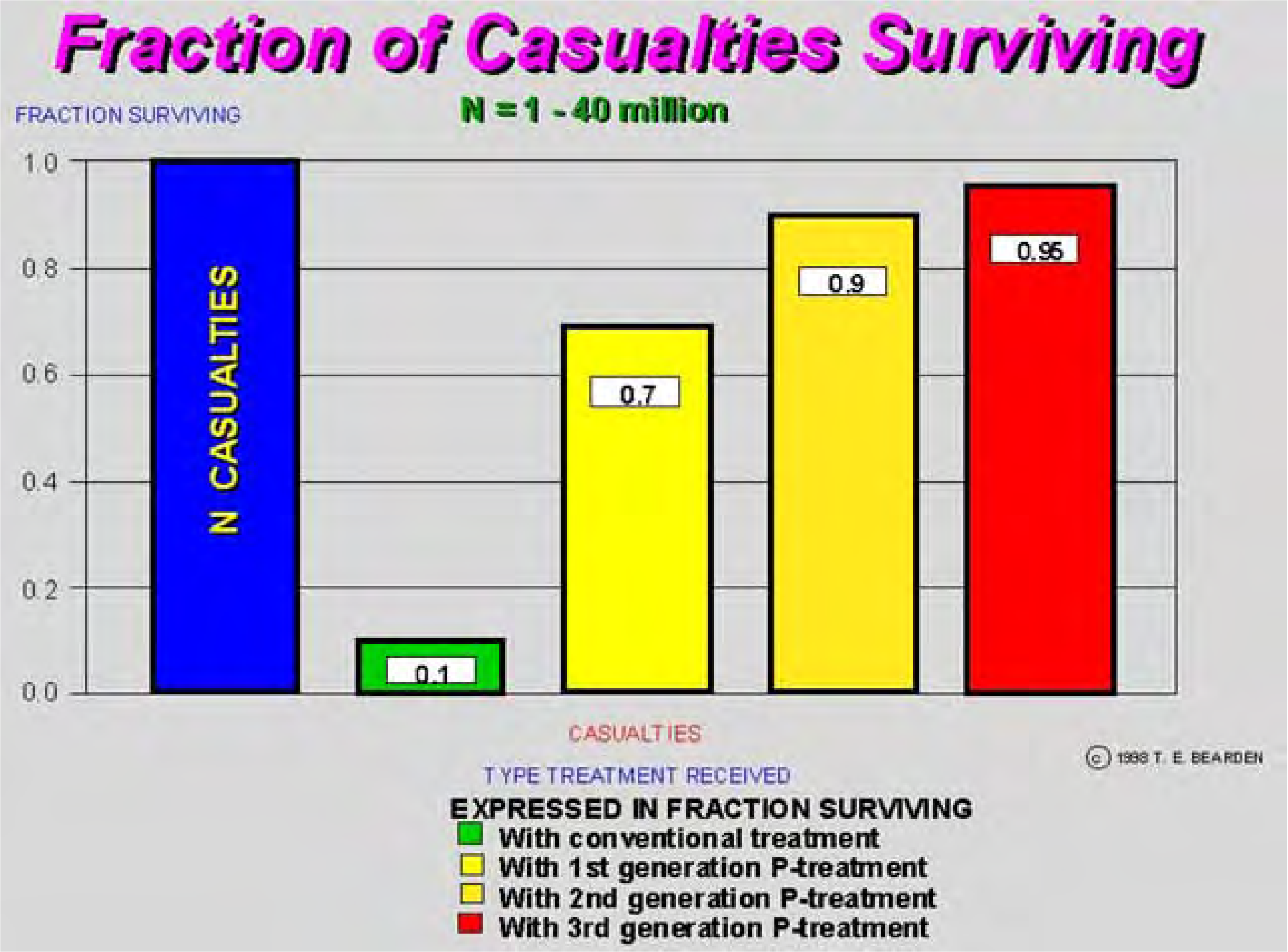
Fraction of Casualties Surviving
FRACTION SURVIVING
N = 1 - 40 million
1.0
0.8
0.6
0.4
0.2
0.0
N CASUALTIES
0.1
0.7
0.9
0.95
CASUALTIES
TYPE TREATMENT RECEIVED
© 1998 T. E. BEARDEN
EXPRESSED IN FRACTION SURVIVING
🟩 With conventional treatment
🟨 With 1st generation P-treatment
🟨 With 2nd generation P-treatment
🟥 With 3rd generation P-treatment

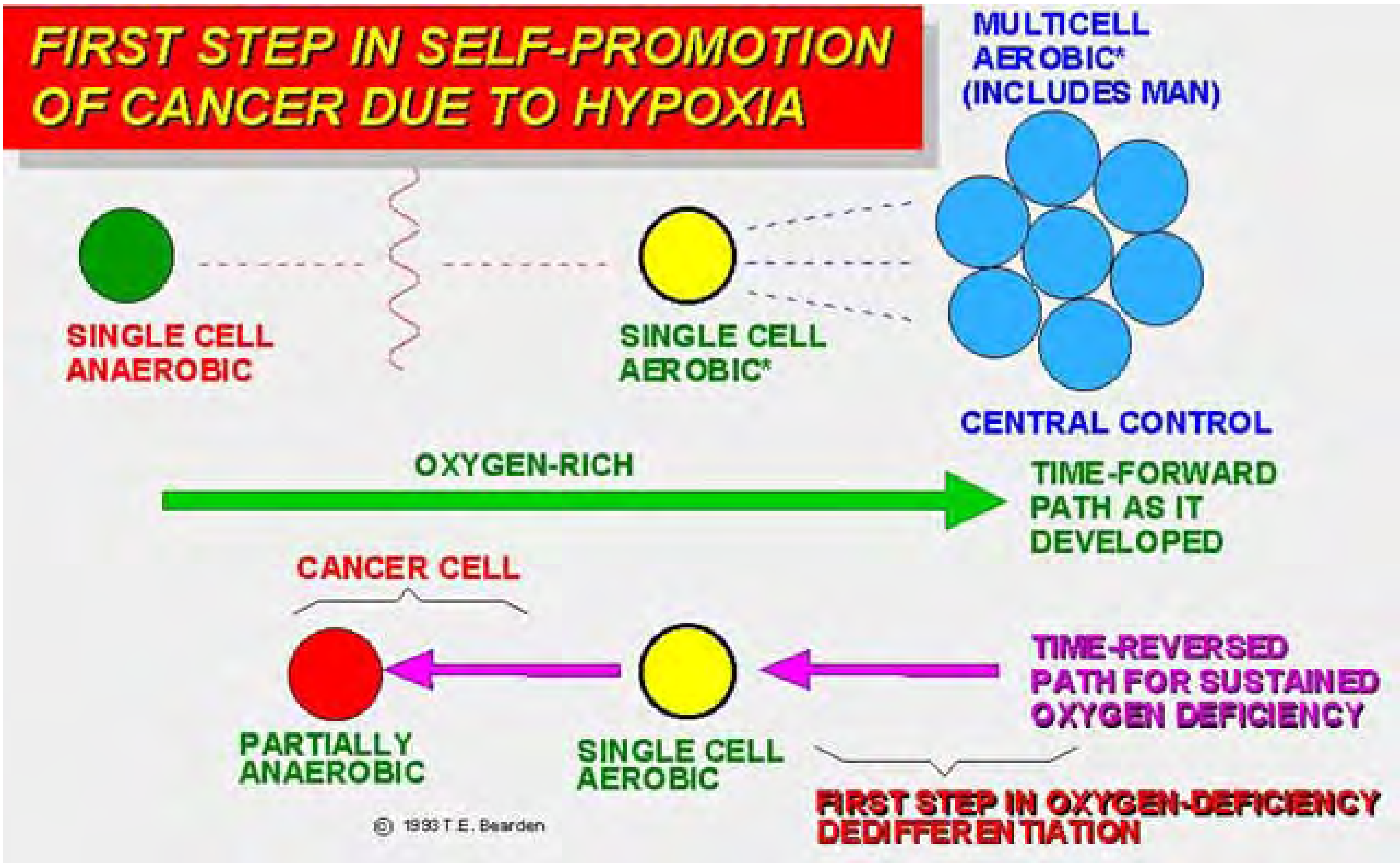

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

MULTICELL
AEROBIC*
(INCLUDES MAN)
SINGLE CELL
ANAEROBIC
SINGLE CELL
AEROBIC*
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD
PATH AS IT
DEVELOPED
CANCER CELL
TIME-REVERSED
PATH FOR SUSTAINED
OXYGEN DEFICIENCY
PARTIALLY
ANAEROBIC
SINGLE CELL
AEROBIC
FIRST STEP IN OXYGEN-DEFICIENCY
DEDIFFERENTIATION

© 1993 T.E. Bearden

Differentiation and Dedifferentiation Timelines

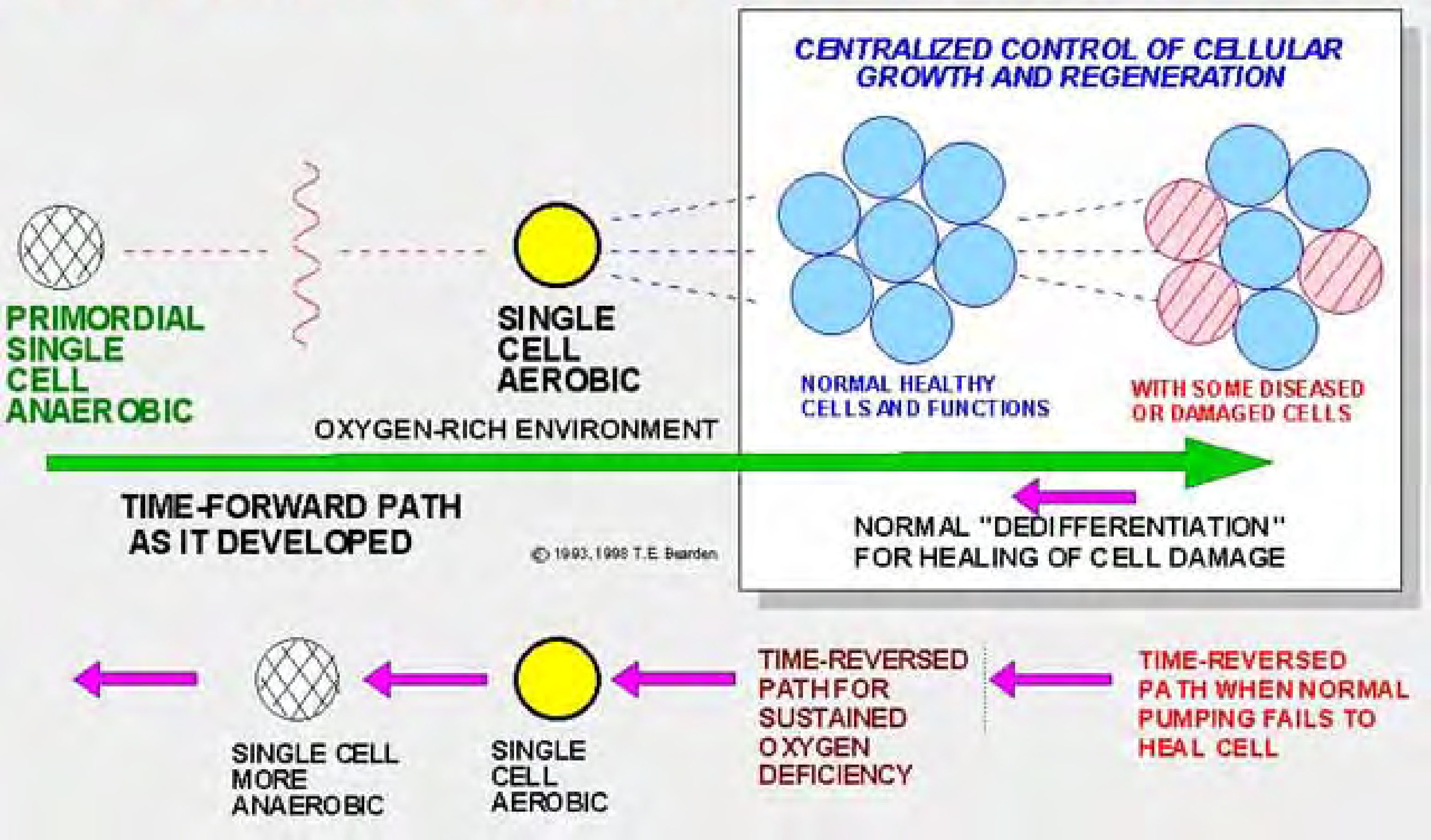
CENTRALIZED CONTROL OF CELLULAR GROWTH AND REGENERATION
PRIMORDIAL SINGLE CELL ANAEROBIC
SINGLE CELL AEROBIC
NORMAL HEALTHY CELLS AND FUNCTIONS
WITH SOME DISEASED OR DAMAGED CELLS
OXYGEN-RICH ENVIRONMENT
TIME-FORWARD PATH AS IT DEVELOPED
NORMAL "DEDIFFERENTIATION" FOR HEALING OF CELL DAMAGE
SINGLE CELL MORE ANAEROBIC
SINGLE CELL AEROBIC
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
TIME-REVERSED PATH WHEN NORMAL PUMPING FAILS TO HEAL CELL

© 1993, 1998 T.E. Bearden

Evolution of Cells on Earth

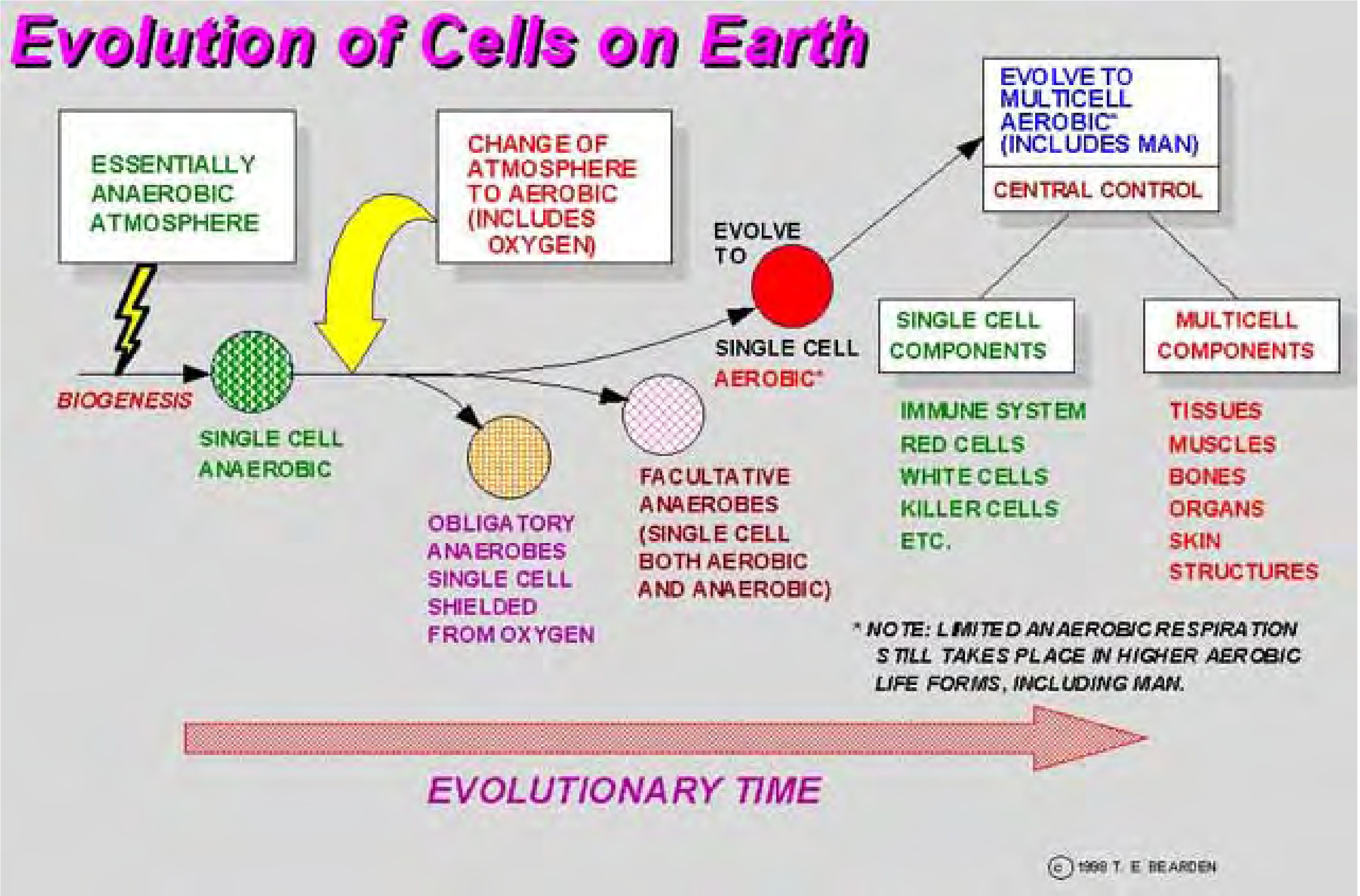

ESSENTIALLY
ANAEROBIC
ATMOSPHERE

CHANGE OF
ATMOSPHERE
TO AEROBIC
(INCLUDES
OXYGEN)

EVOLVE TO
MULTICELL
AEROBIC*
(INCLUDES MAN)

CENTRAL CONTROL

BIOGENESIS

SINGLE CELL
ANAEROBIC

EVOLVE
TO

SINGLE CELL
AEROBIC*

OBLIGATORY
ANAEROBES
SINGLE CELL
SHIELDED
FROM OXYGEN

FACULTATIVE
ANAEROBES
(SINGLE CELL
BOTH AEROBIC
AND ANAEROBIC)

SINGLE CELL
COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL
COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

** NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.*

EVOLUTIONARY TIME

(c) 1999 T. E. BEARDEN

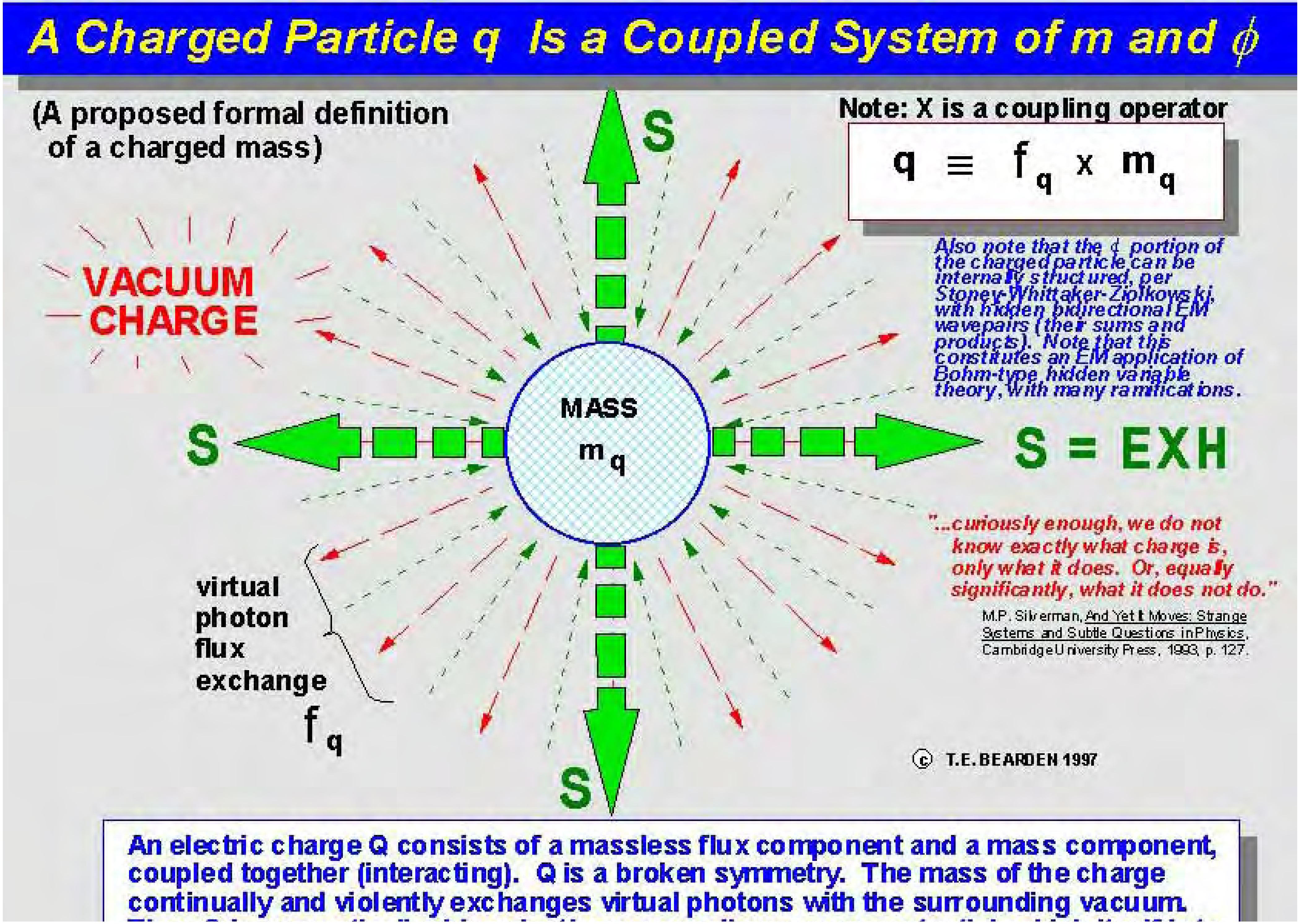

A Charged Particle q Is a Coupled System of m and ϕ

(A proposed formal definition of a charged mass)

Note: X is a coupling operator

$$q \equiv f_q \times m_q$$

Also note that the ϕ portion of the charged particle can be internally structured, per Stoney-Whittaker-Ziolkowski, with hidden bidirectional EM wavepairs (their sums and products). Note that this constitutes an EM application of Bohm-type hidden variable theory, with many ramifications.

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

© T.E. BEARDEN 1997

An electric charge Q consists of a massless flux component and a mass component, coupled together (interacting). Q is a broken symmetry. The mass of the charge continually and violently exchanges virtual photons with the surrounding vacuum.

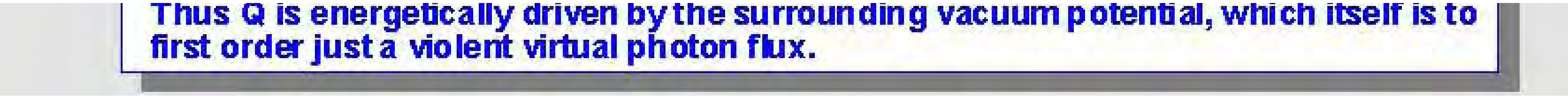

Thus Q is energetically driven by the surrounding vacuum potential, which itself is to first order just a violent virtual photon flux.

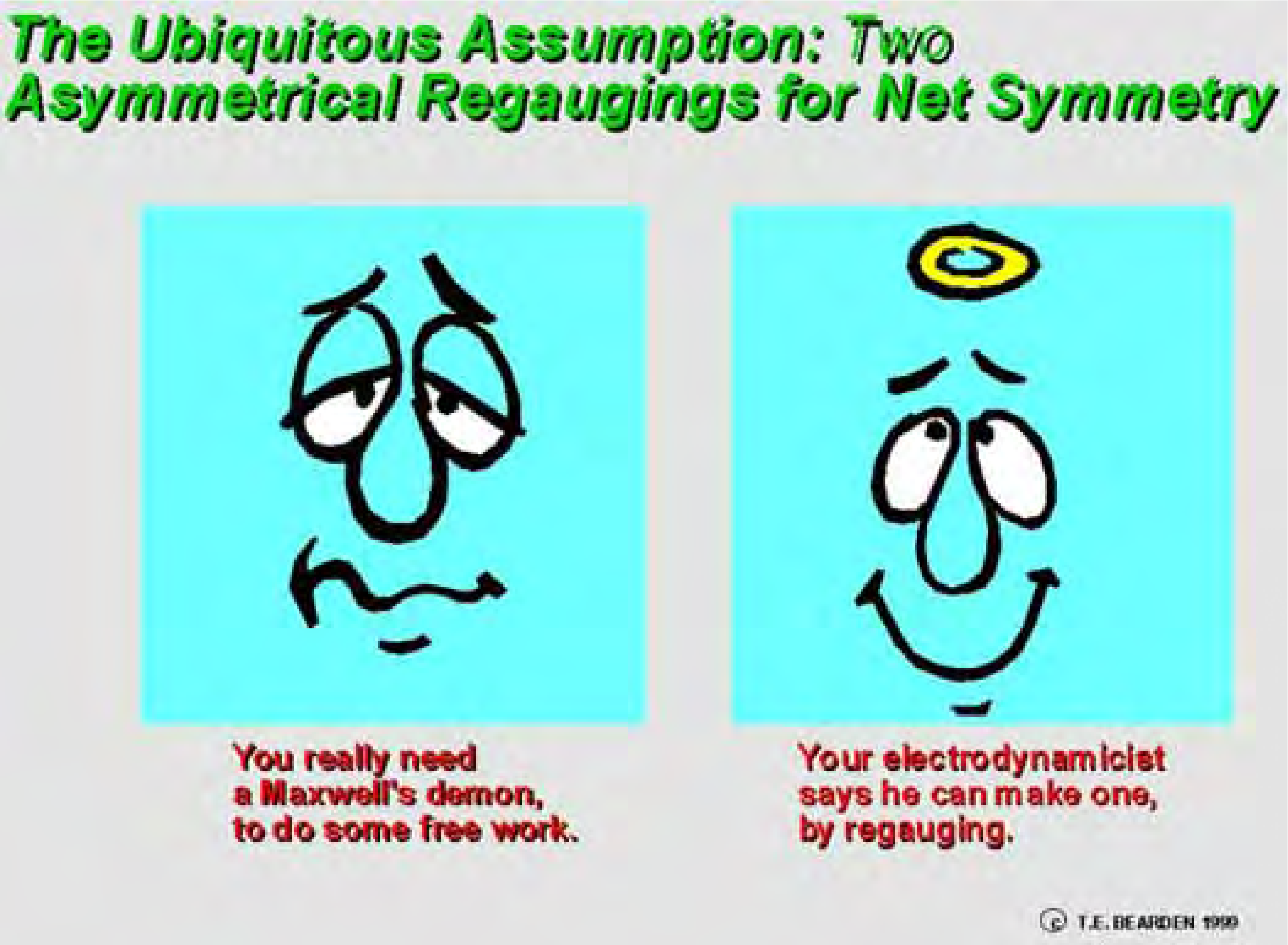

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

You really need
a Maxwell's demon,
to do some free work.

Your electrodynamicist
says he can make one,
by regauging.

© T.E. BEARDEN 1999

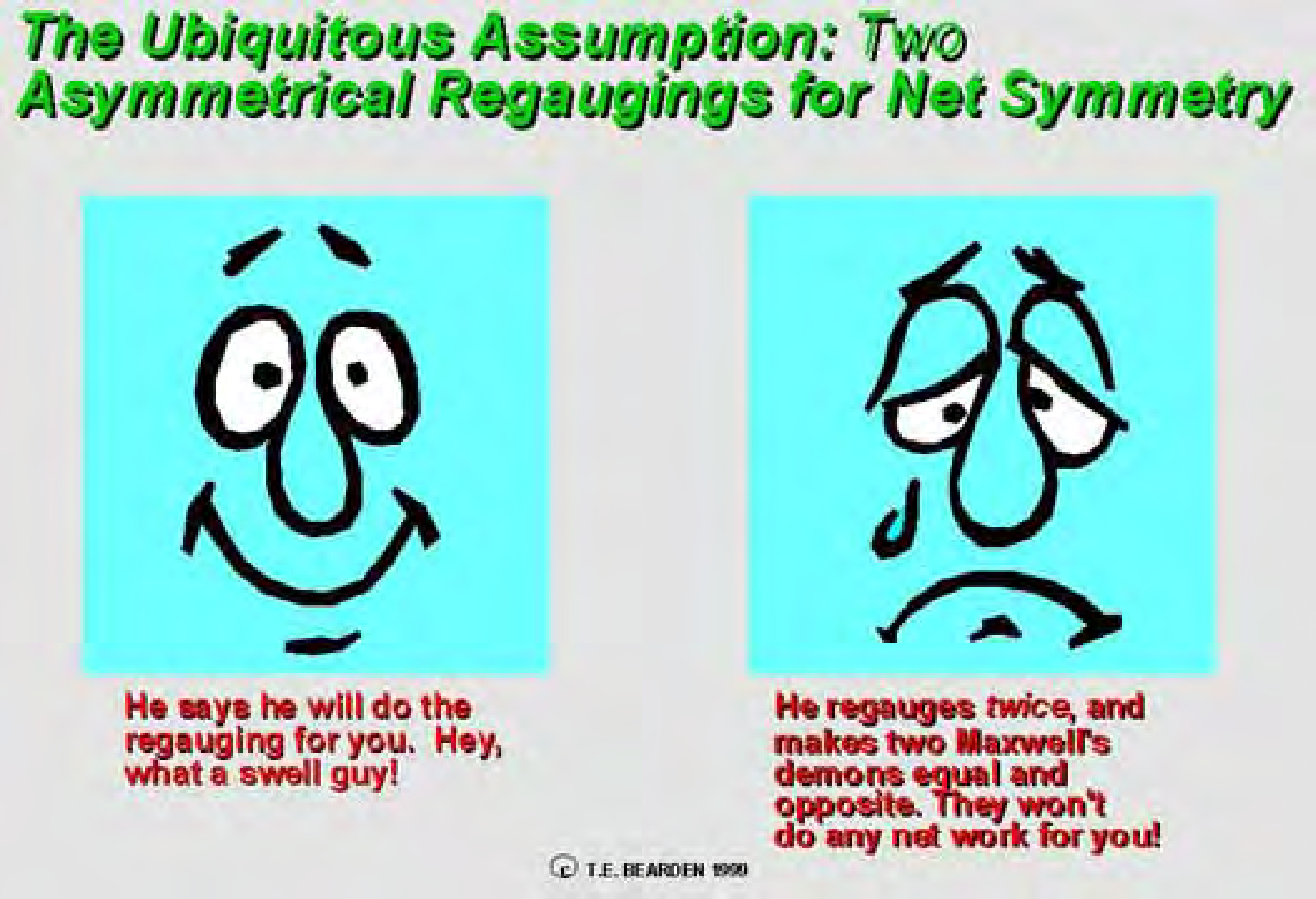

The Ubiquitous Assumption: *Two* Asymmetrical Regaugings for Net Symmetry

He says he will do the regauging for you. Hey, what a swell guy!

He regauges *twice*, and makes two Maxwell's demons equal and opposite. They won't do any net work for you!

© T.E. BEARDEN 1999

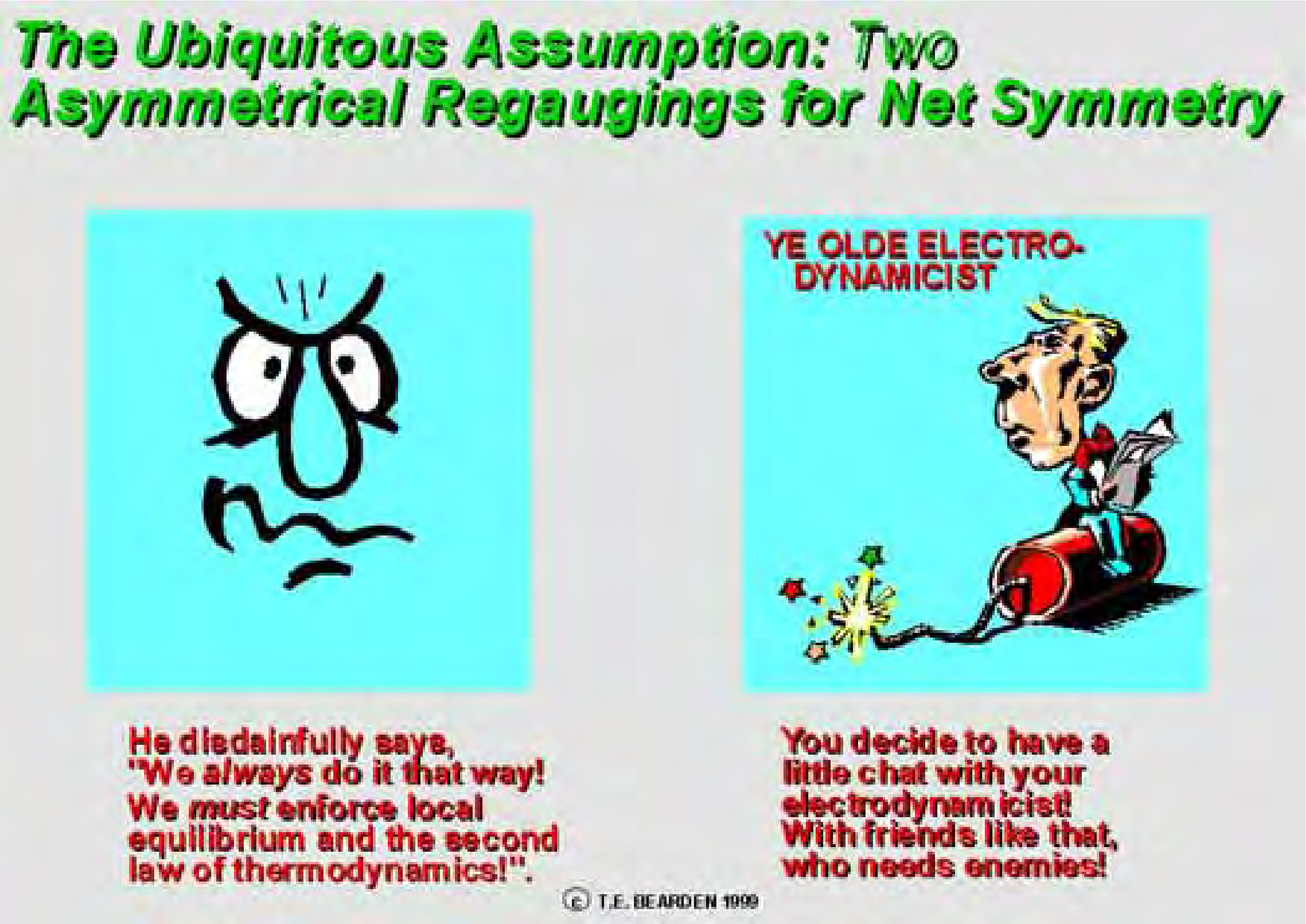

The Ubiquitous Assumption: Two Asymmetrical Regaugings for Net Symmetry

YE OLDE ELECTRO-
DYNAMICIST

He disdainfully says,
"We *always* do it that way!
We *must* enforce local
equilibrium and the second
law of thermodynamics!".

You decide to have a
little chat with your
electrodynamicist!
With friends like that,
who needs enemies!

© T.E. BEARDEN 1999

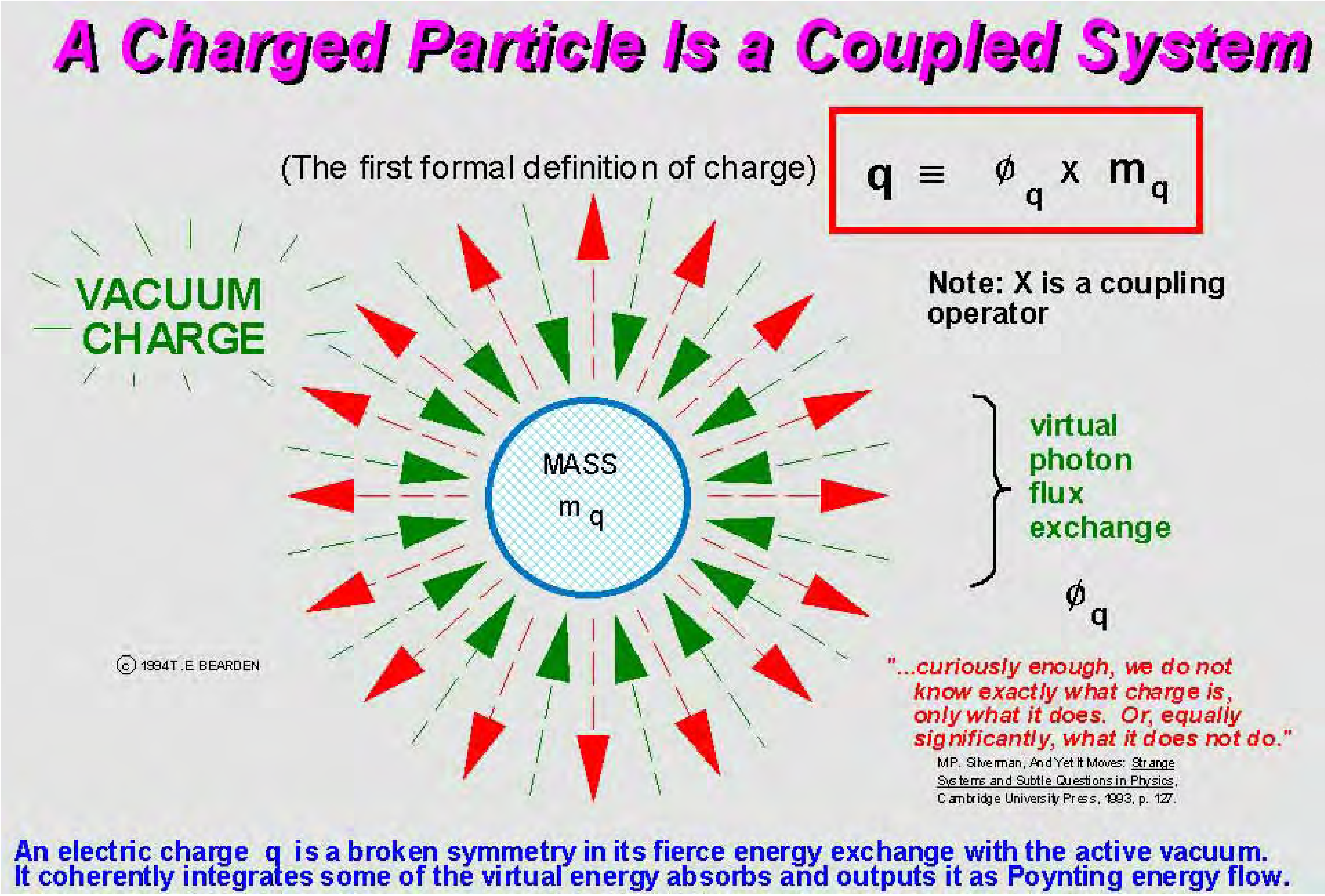

A Charged Particle Is a Coupled System

(The first formal definition of charge)

$$q \equiv \phi_q \times m_q$$

Note: X is a coupling operator

"...curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."

M.P. Silverman, And Yet It Moves: Strange Systems and Subtle Questions in Physics, Cambridge University Press, 1993, p. 127.

An electric charge q is a broken symmetry in its fierce energy exchange with the active vacuum. It coherently integrates some of the virtual energy absorbs and outputs it as Poynting energy flow.

Chung's Carbon Filament Negative Resistor

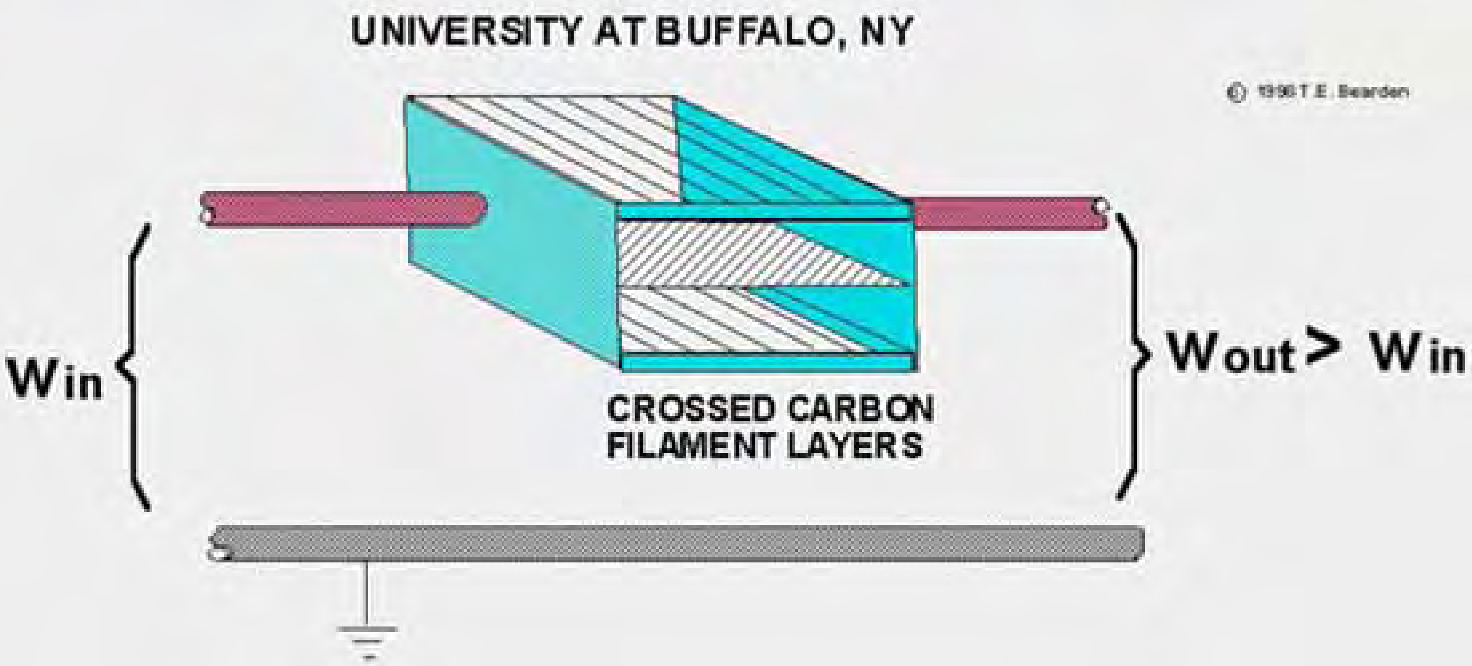
UNIVERSITY AT BUFFALO, NY
© 1996 T.E. Bearden
Win
Wout > Win
CROSSED CARBON
FILAMENT LAYERS

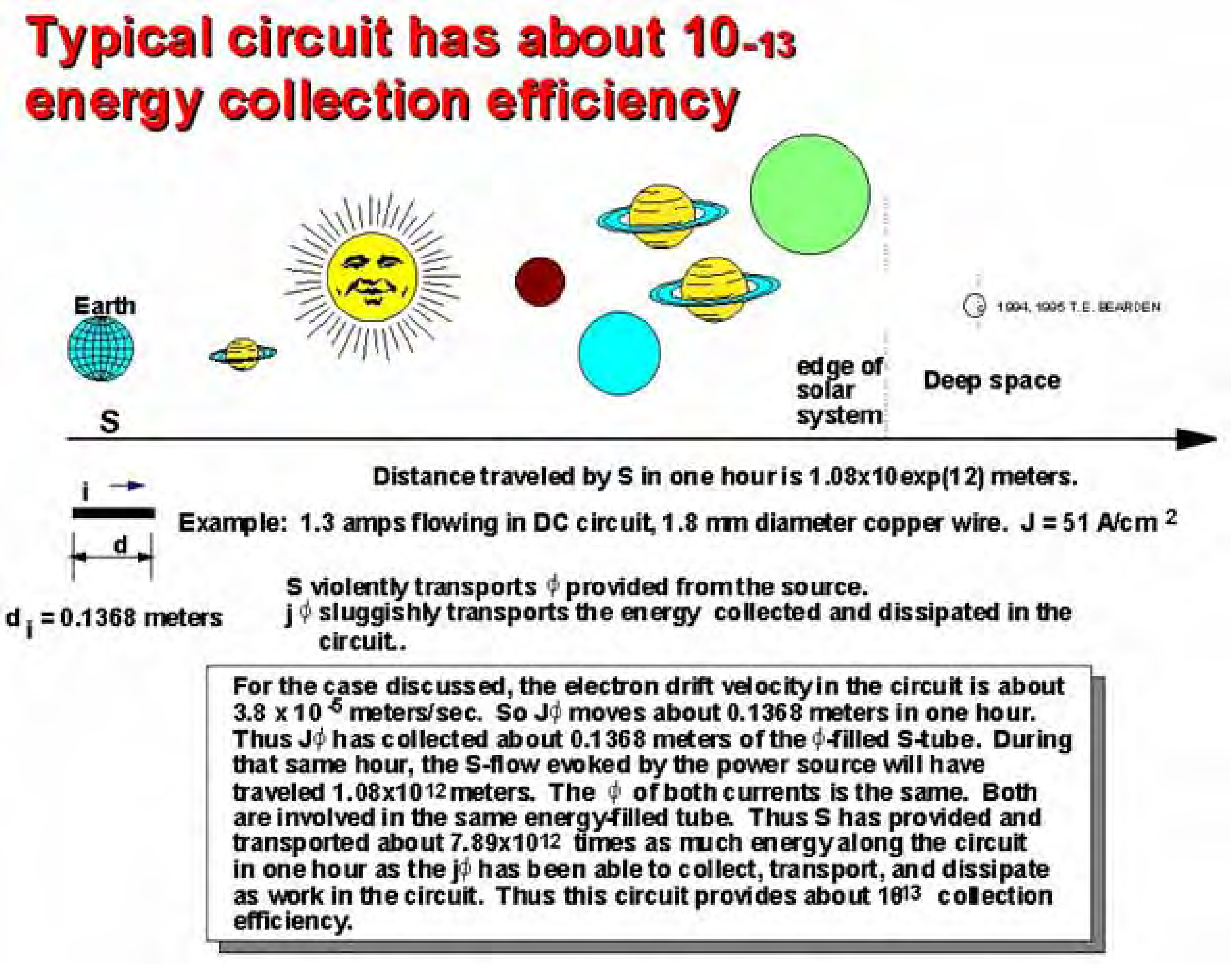

Typical circuit has about 10^{-13} energy collection efficiency

Distance traveled by S in one hour is 1.08x10exp(12) meters.

Example: 1.3 amps flowing in DC circuit, 1.8 mm diameter copper wire. $J = 51$ A/cm^2

$d_i = 0.1368$ meters

S violently transports ϕ provided from the source.
$j\phi$ sluggishly transports the energy collected and dissipated in the circuit..

For the case discussed, the electron drift velocity in the circuit is about 3.8×10^{-5} meters/sec. So $J\phi$ moves about 0.1368 meters in one hour. Thus $J\phi$ has collected about 0.1368 meters of the ϕ-filled S-tube. During that same hour, the S-flow evoked by the power source will have traveled $1.08\text{x}10^{12}$ meters. The ϕ of both currents is the same. Both are involved in the same energy-filled tube. Thus S has provided and transported about $7.89\text{x}10^{12}$ times as much energy along the circuit in one hour as the $j\phi$ has been able to collect, transport, and dissipate as work in the circuit. Thus this circuit provides about 10^{13} collection efficiency.

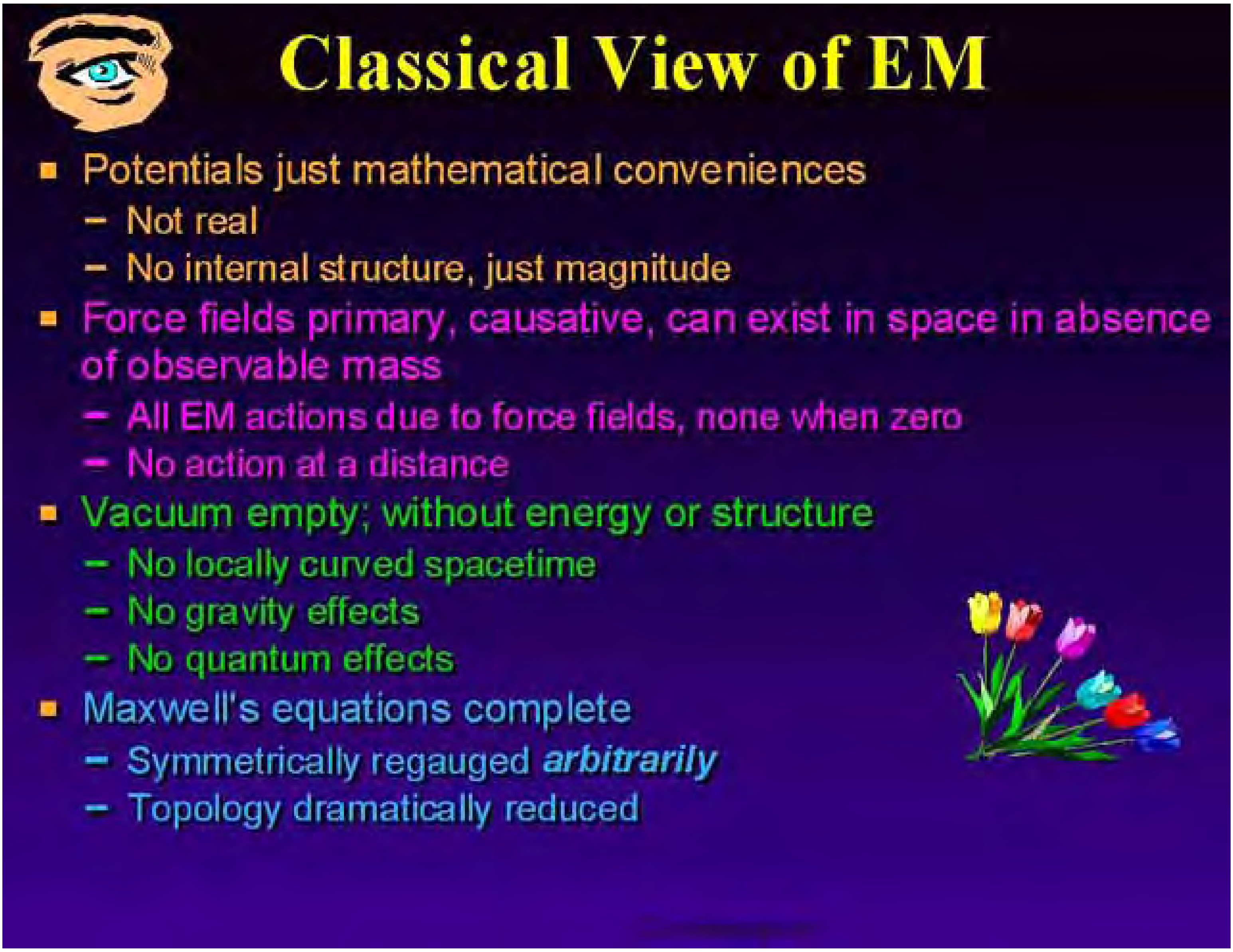

Classical View of EM

- Potentials just mathematical conveniences
  - Not real
  - No internal structure, just magnitude
- Force fields primary, causative, can exist in space in absence of observable mass
  - All EM actions due to force fields, none when zero
  - No action at a distance
- Vacuum empty; without energy or structure
  - No locally curved spacetime
  - No gravity effects
  - No quantum effects
- Maxwell's equations complete
  - Symmetrically regauged ***arbitrarily***
  - Topology dramatically reduced

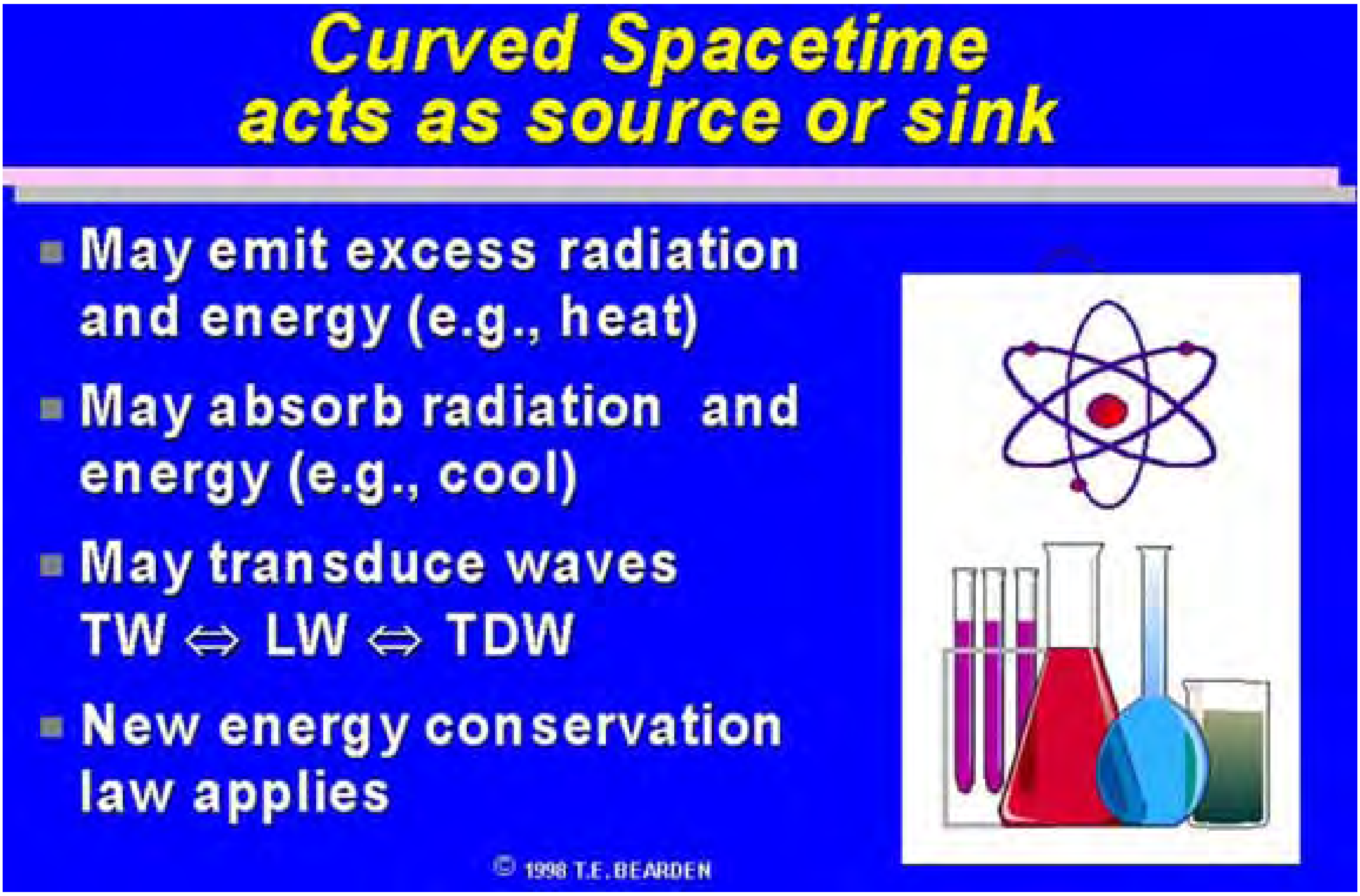

Curved Spacetime acts as source or sink

- May emit excess radiation and energy (e.g., heat)
- May absorb radiation and energy (e.g., cool)
- May transduce waves
  TW ⇔ LW ⇔ TDW
- New energy conservation law applies

© 1998 T.E.BEARDEN

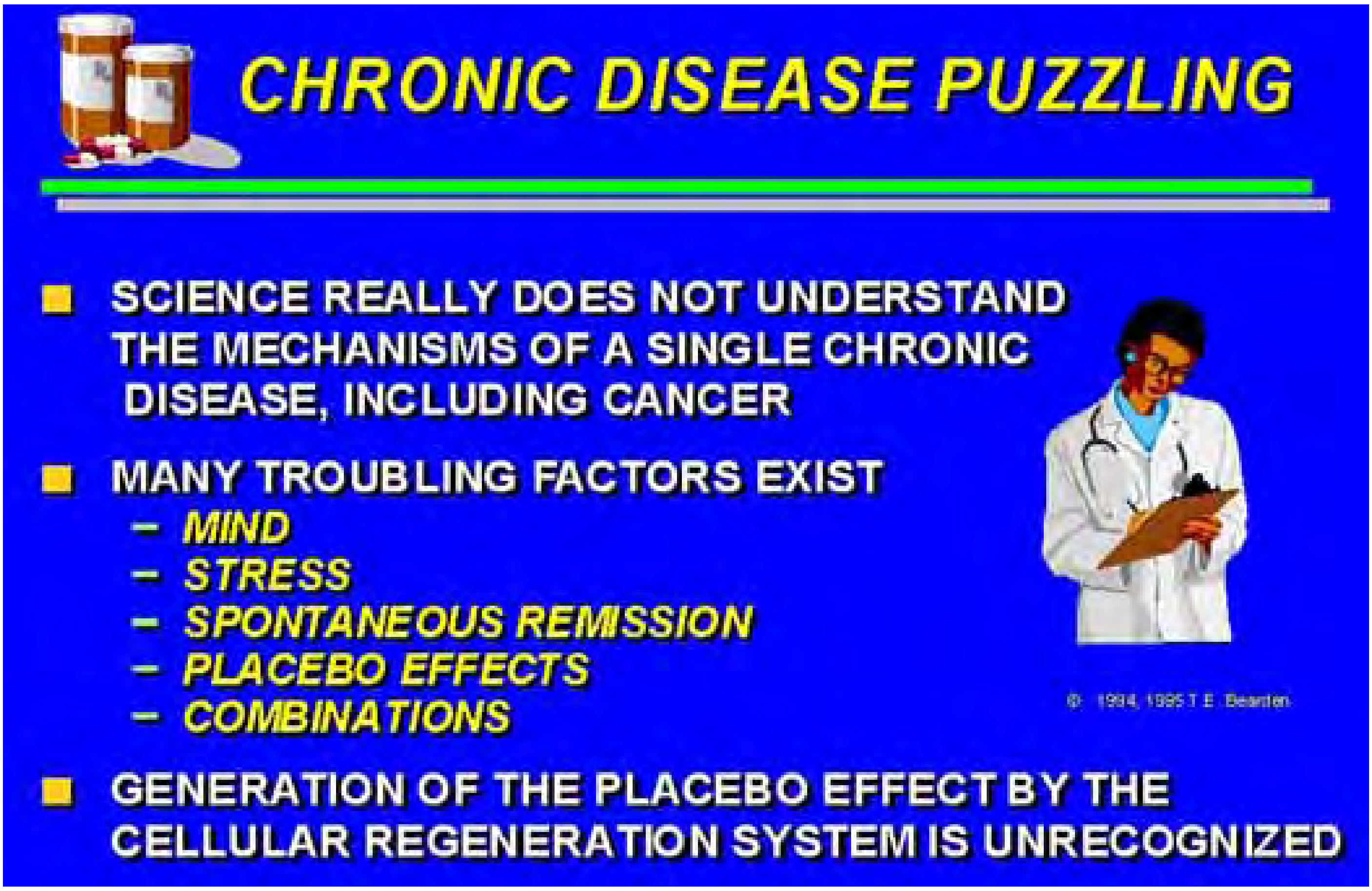

CHRONIC DISEASE PUZZLING

- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
- MANY TROUBLING FACTORS EXIST
  - *MIND*
  - *STRESS*
  - *SPONTANEOUS REMISSION*
  - *PLACEBO EFFECTS*
  - *COMBINATIONS*
- GENERATION OF THE PLACEBO EFFECT BY THE CELLULAR REGENERATION SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

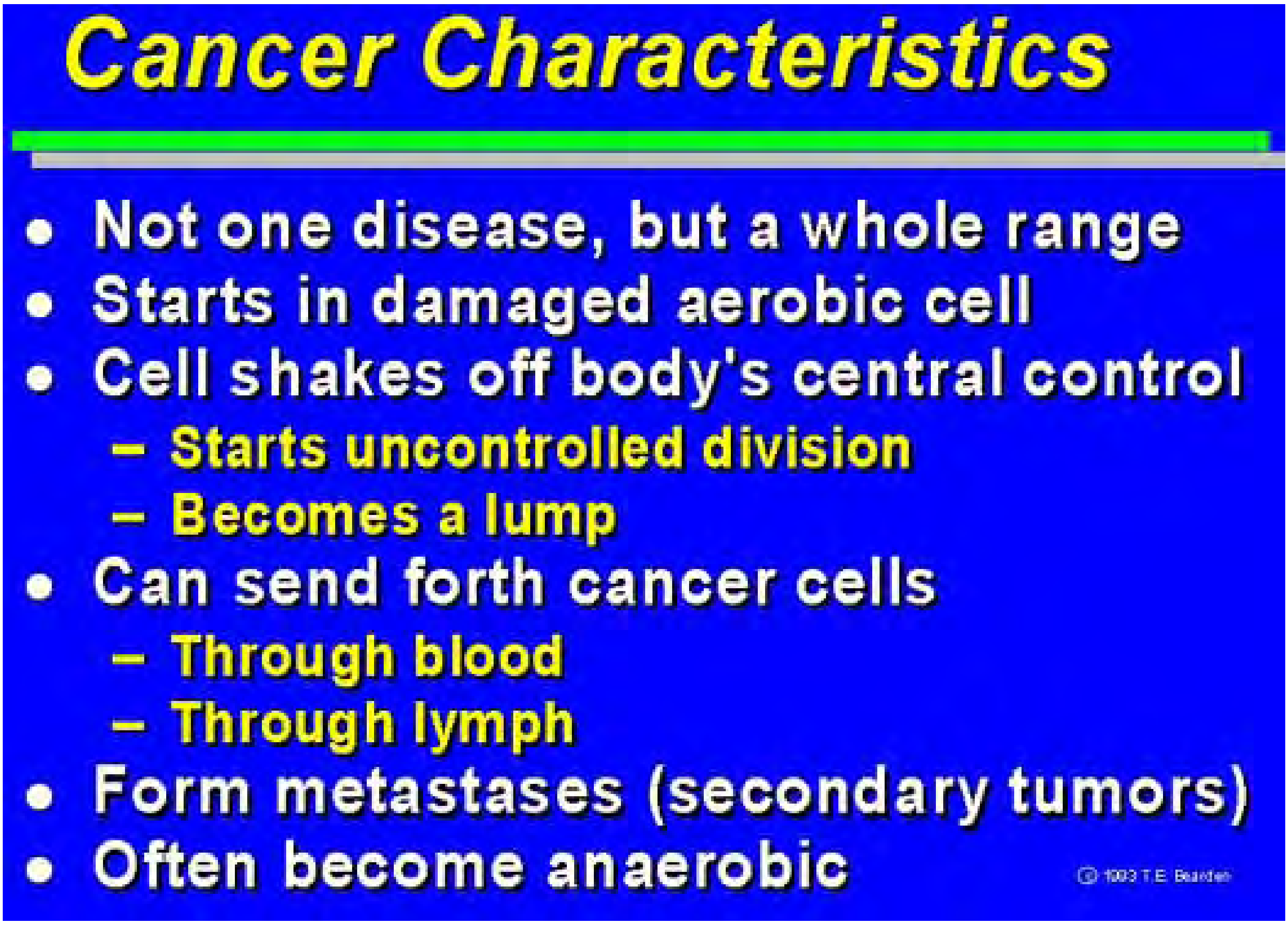

Cancer Characteristics

- Not one disease, but a whole range
- Starts in damaged aerobic cell
- Cell shakes off body's central control
  - Starts uncontrolled division
  - Becomes a lump
- Can send forth cancer cells
  - Through blood
  - Through lymph
- Form metastases (secondary tumors)
- Often become anaerobic

© 1993 T.E. Bearden

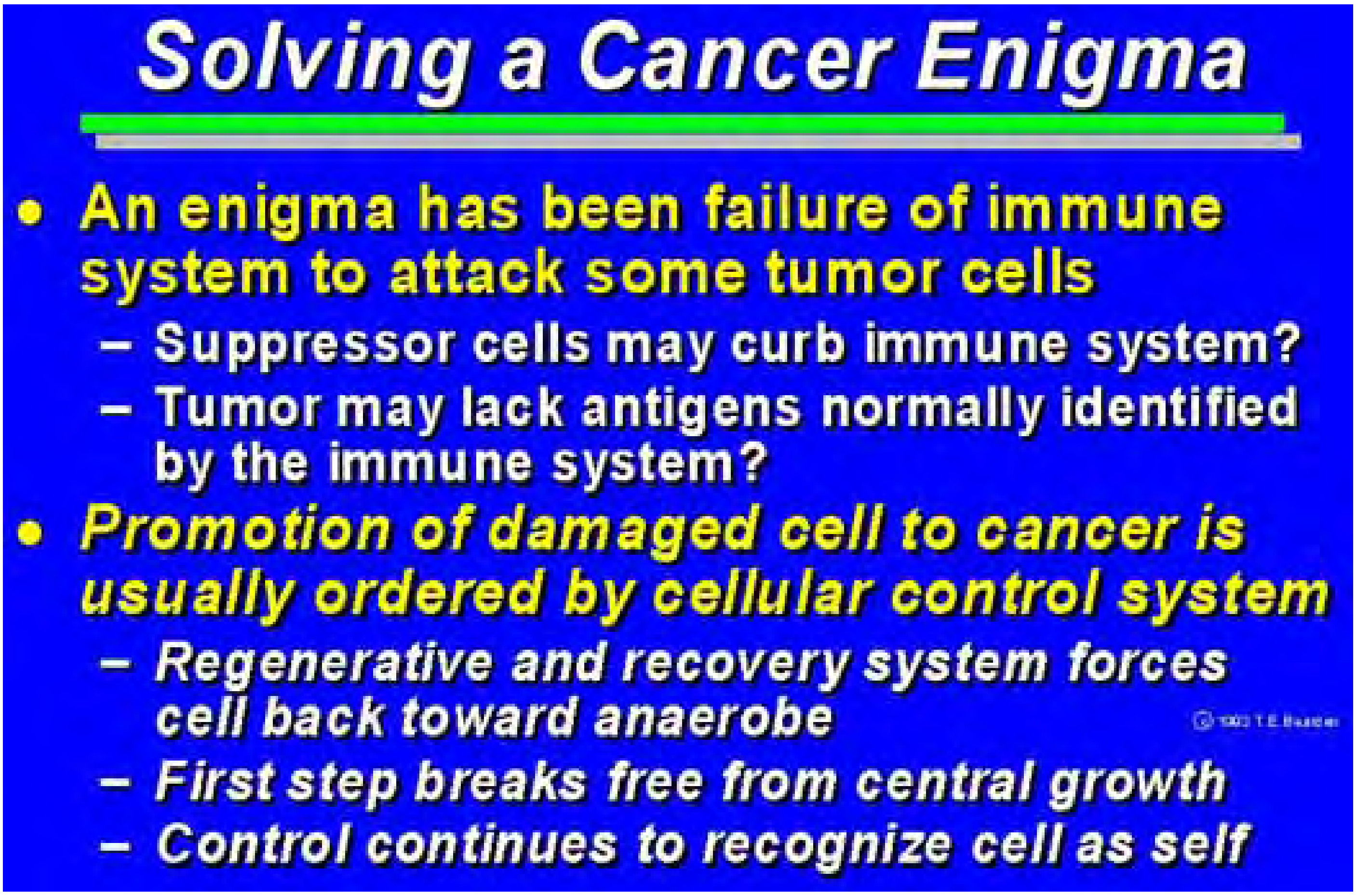

Solving a Cancer Enigma

- An enigma has been failure of immune system to attack some tumor cells
  - Suppressor cells may curb immune system?
  - Tumor may lack antigens normally identified by the immune system?
- *Promotion of damaged cell to cancer is usually ordered by cellular control system*
  - *Regenerative and recovery system forces cell back toward anaerobe*
  - *First step breaks free from central growth*
  - *Control continues to recognize cell as self*

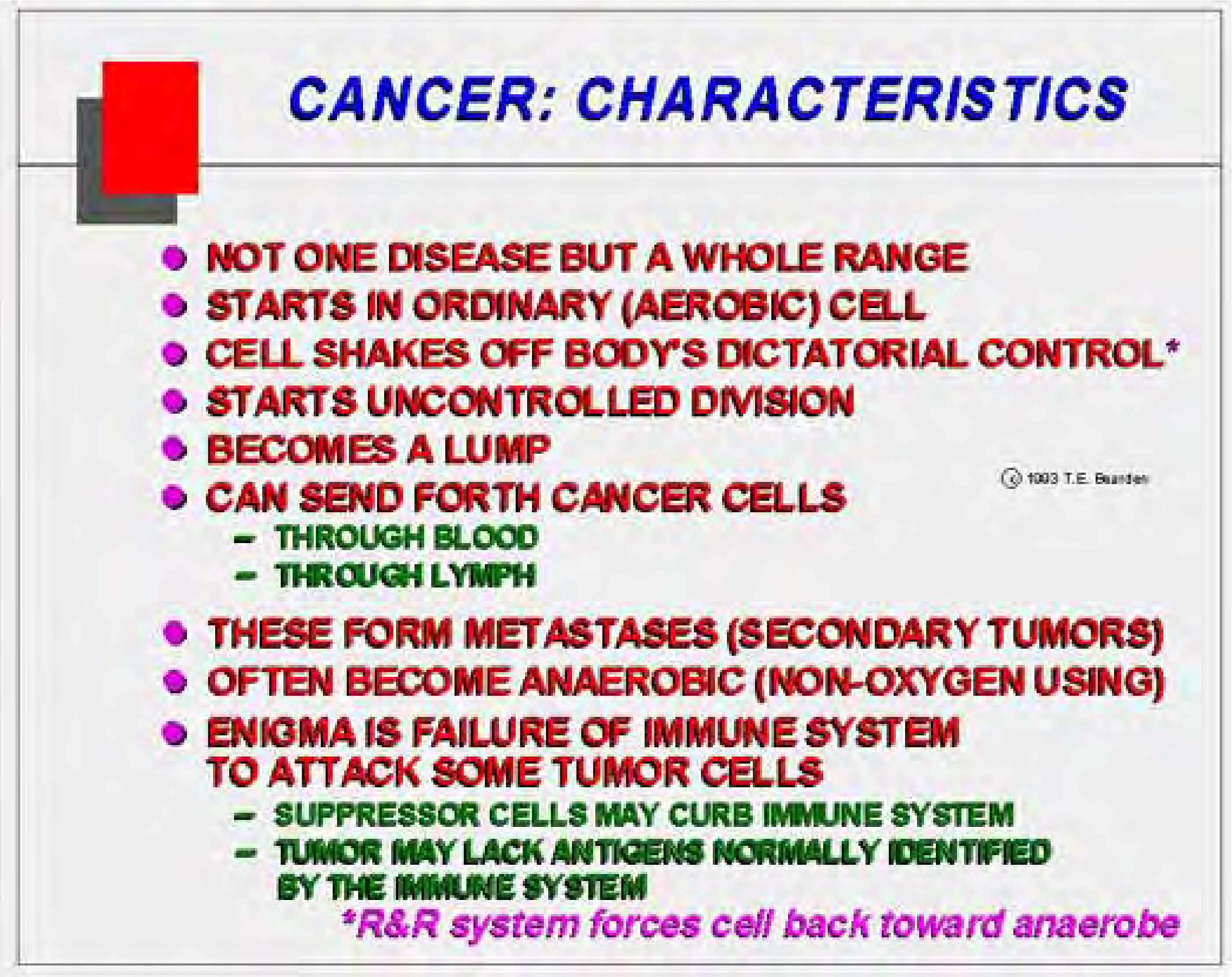

CANCER: CHARACTERISTICS

- NOT ONE DISEASE BUT A WHOLE RANGE
- STARTS IN ORDINARY (AEROBIC) CELL
- CELL SHAKES OFF BODY'S DICTATORIAL CONTROL*
- STARTS UNCONTROLLED DIVISION
- BECOMES A LUMP
- CAN SEND FORTH CANCER CELLS
  - THROUGH BLOOD
  - THROUGH LYMPH
- THESE FORM METASTASES (SECONDARY TUMORS)
- OFTEN BECOME ANAEROBIC (NON-OXYGEN USING)
- ENIGMA IS FAILURE OF IMMUNE SYSTEM TO ATTACK SOME TUMOR CELLS
  - SUPPRESSOR CELLS MAY CURB IMMUNE SYSTEM
  - TUMOR MAY LACK ANTIGENS NORMALLY IDENTIFIED BY THE IMMUNE SYSTEM

© 1993 T.E. Bearden

*R&R system forces cell back toward anaerobe

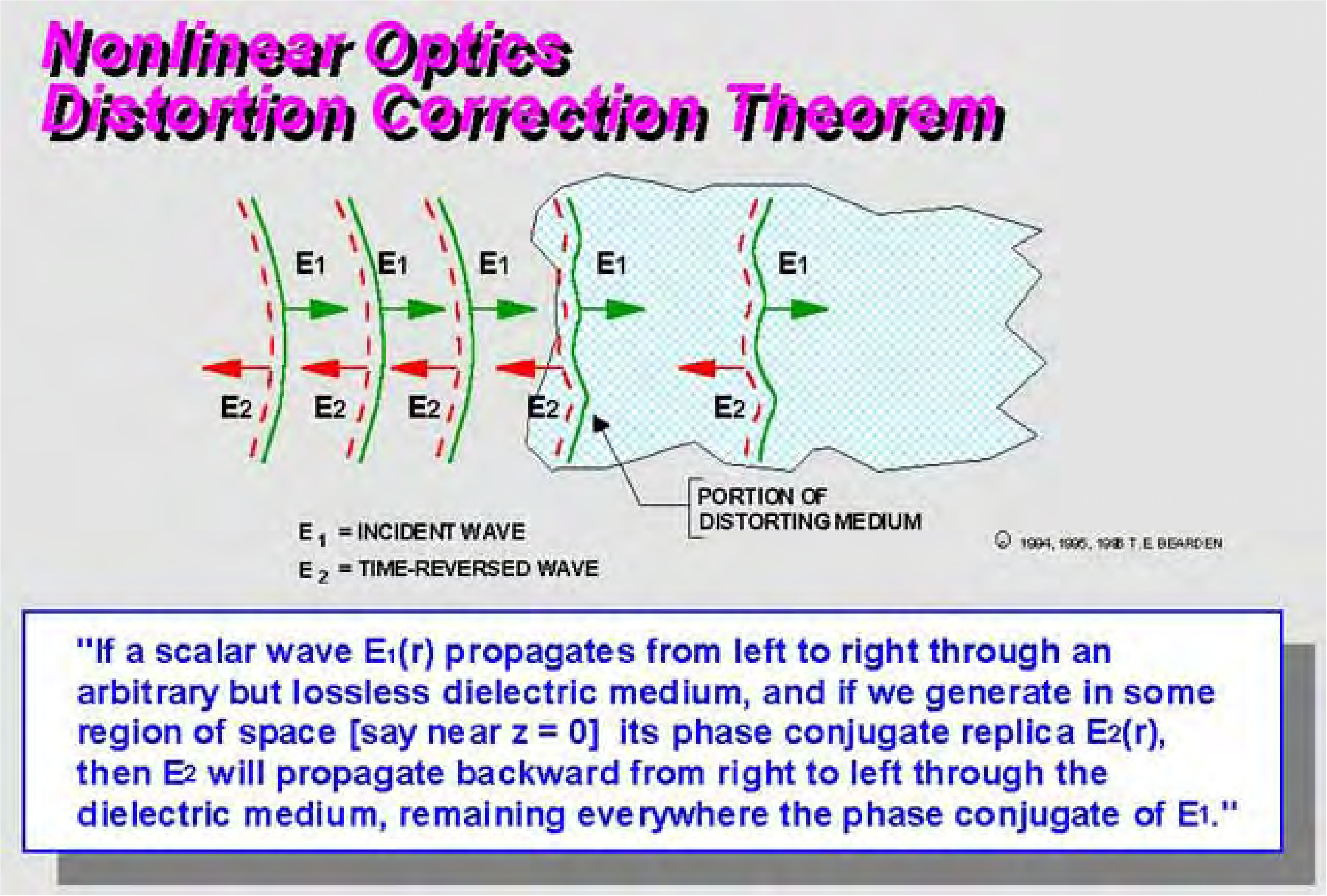

Nonlinear Optics
Distortion Correction Theorem

E1
E1
E1
E1
E1
E2
E2
E2
E2
E2
PORTION OF
DISTORTING MEDIUM

E_1 = INCIDENT WAVE

E_2 = TIME-REVERSED WAVE

© 1994, 1995, 1998 T. E. BEARDEN

"If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

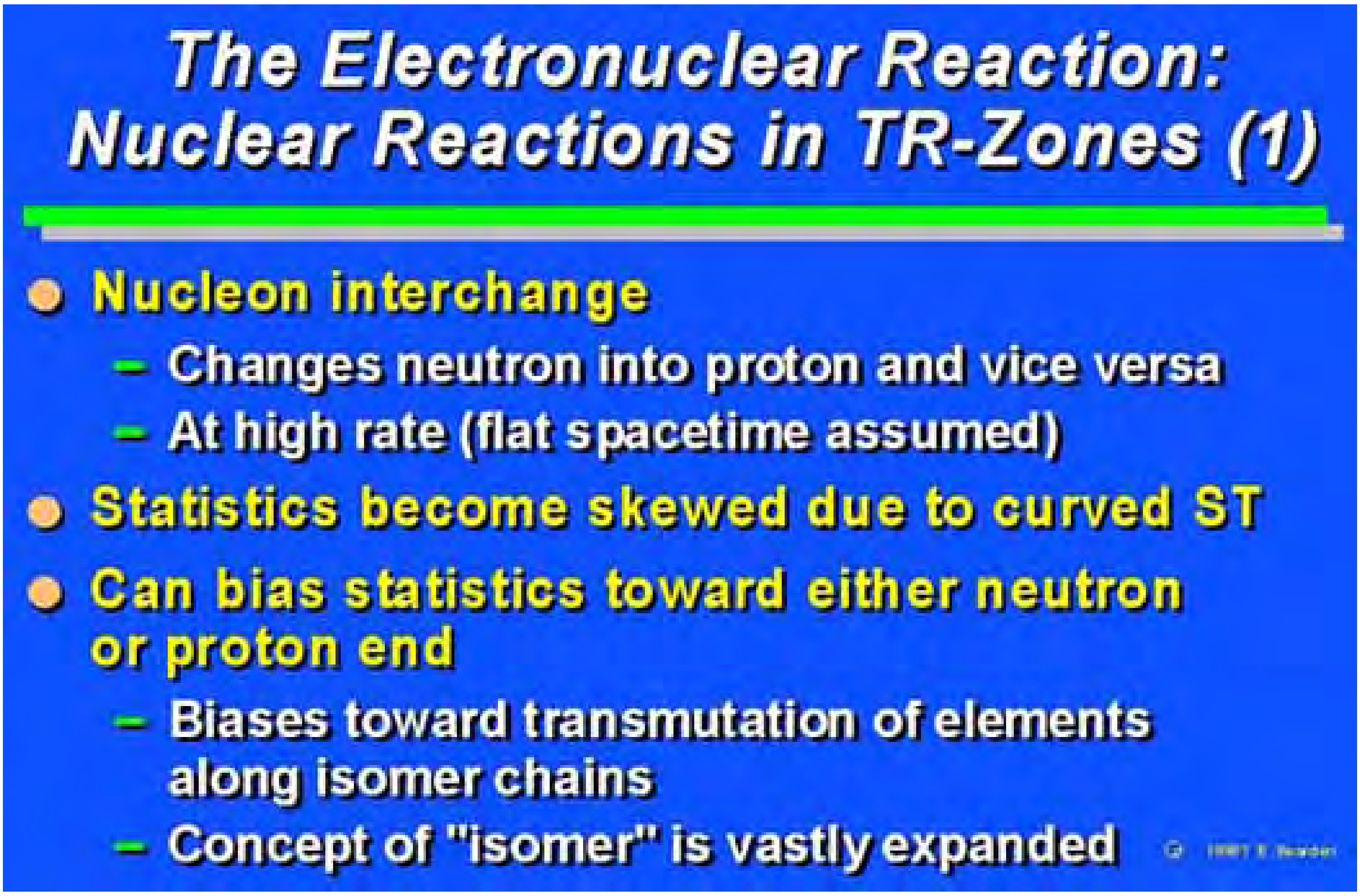

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (1)

- **Nucleon interchange**
  - Changes neutron into proton and vice versa
  - At high rate (flat spacetime assumed)
- **Statistics become skewed due to curved ST**
- **Can bias statistics toward either neutron or proton end**
  - Biases toward transmutation of elements along isomer chains
  - Concept of "isomer" is vastly expanded

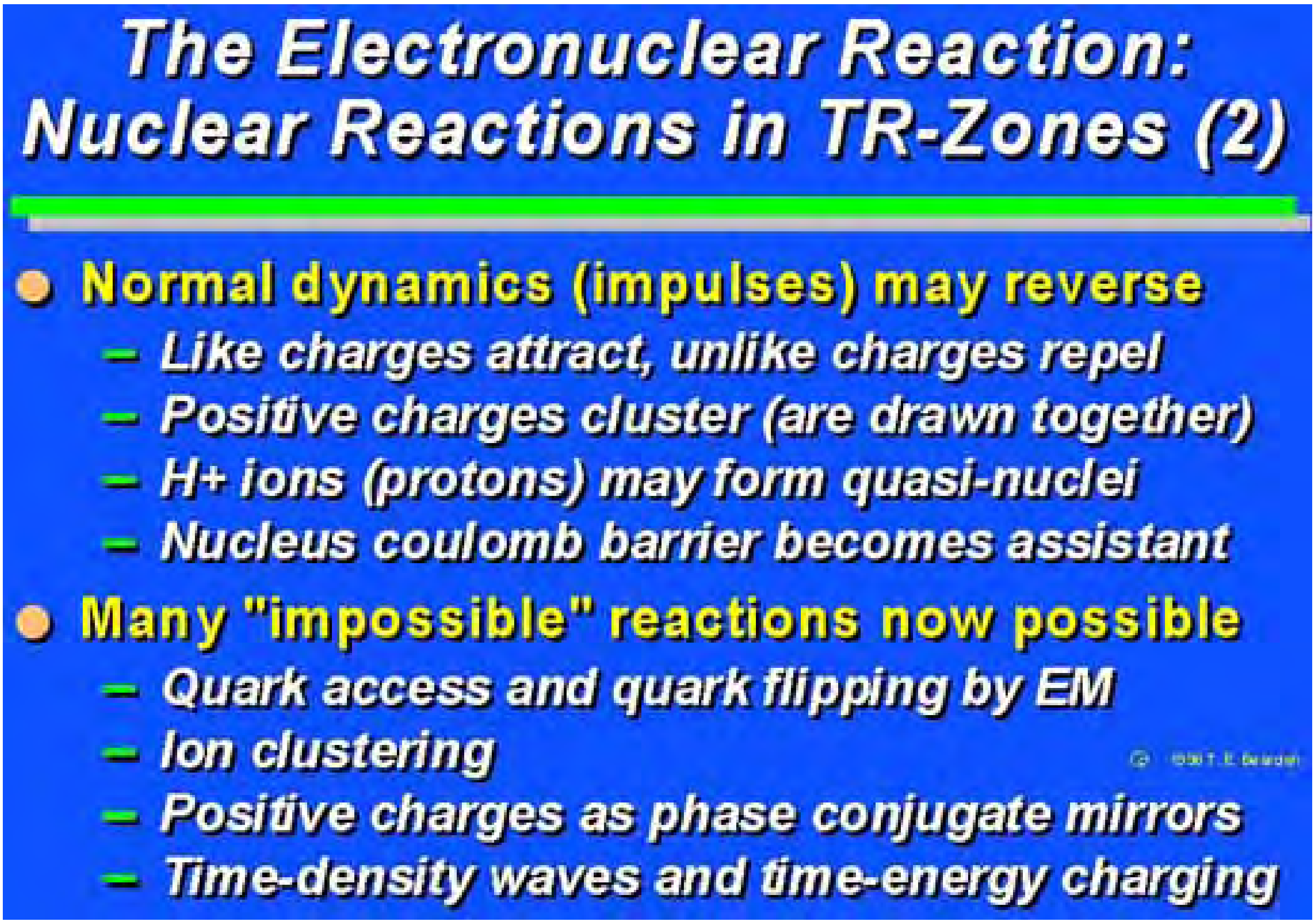

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (2)

- Normal dynamics (impulses) may reverse
  - *Like charges attract, unlike charges repel*
  - *Positive charges cluster (are drawn together)*
  - *H+ ions (protons) may form quasi-nuclei*
  - *Nucleus coulomb barrier becomes assistant*
- Many "impossible" reactions now possible
  - *Quark access and quark flipping by EM*
  - *Ion clustering*
  - *Positive charges as phase conjugate mirrors*
  - *Time-density waves and time-energy charging*

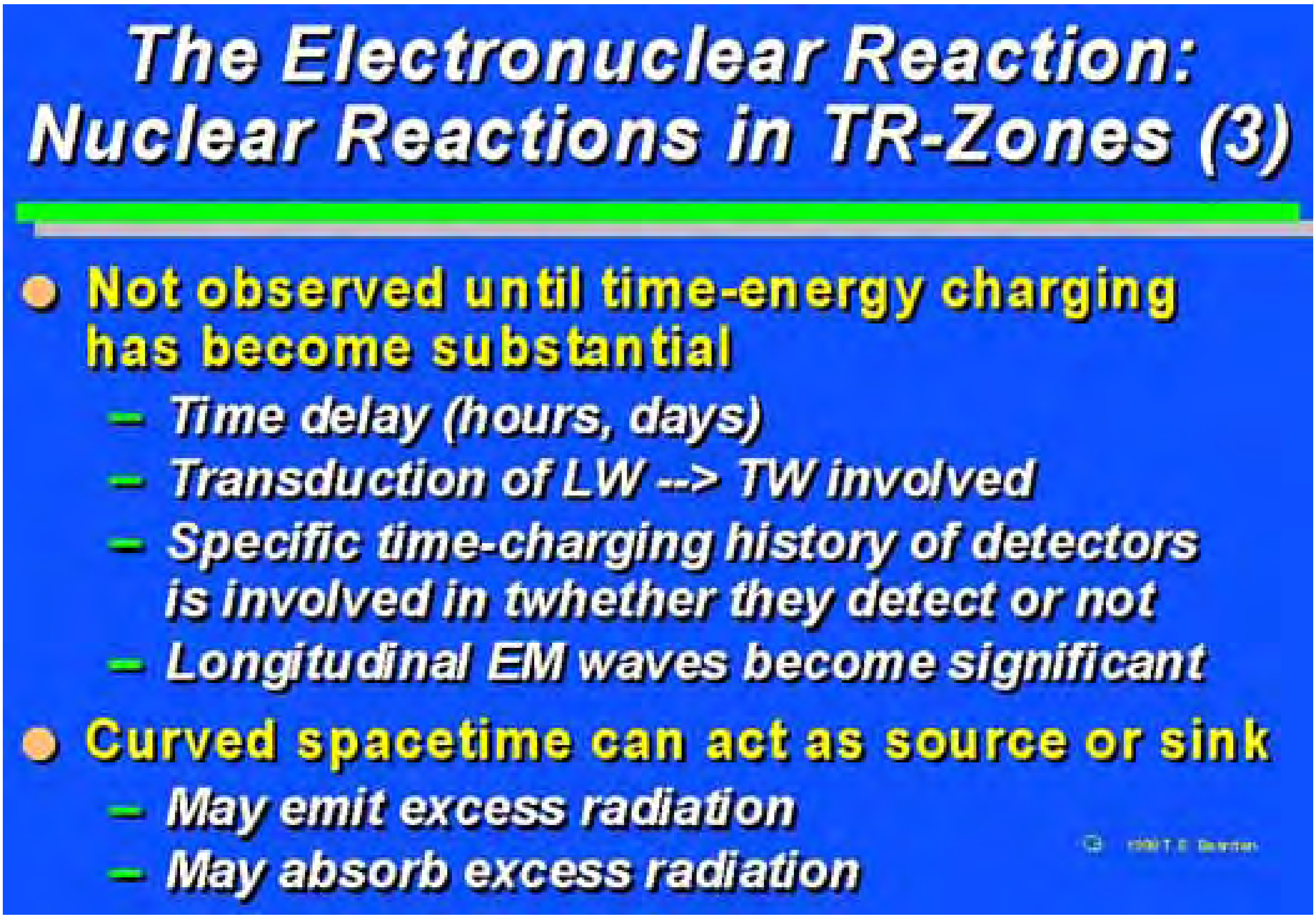

The Electronuclear Reaction: Nuclear Reactions in TR-Zones (3)

- **Not observed until time-energy charging has become substantial**
  - *Time delay (hours, days)*
  - *Transduction of LW --> TW involved*
  - *Specific time-charging history of detectors is involved in twhether they detect or not*
  - *Longitudinal EM waves become significant*
- **Curved spacetime can act as source or sink**
  - *May emit excess radiation*
  - *May absorb excess radiation*

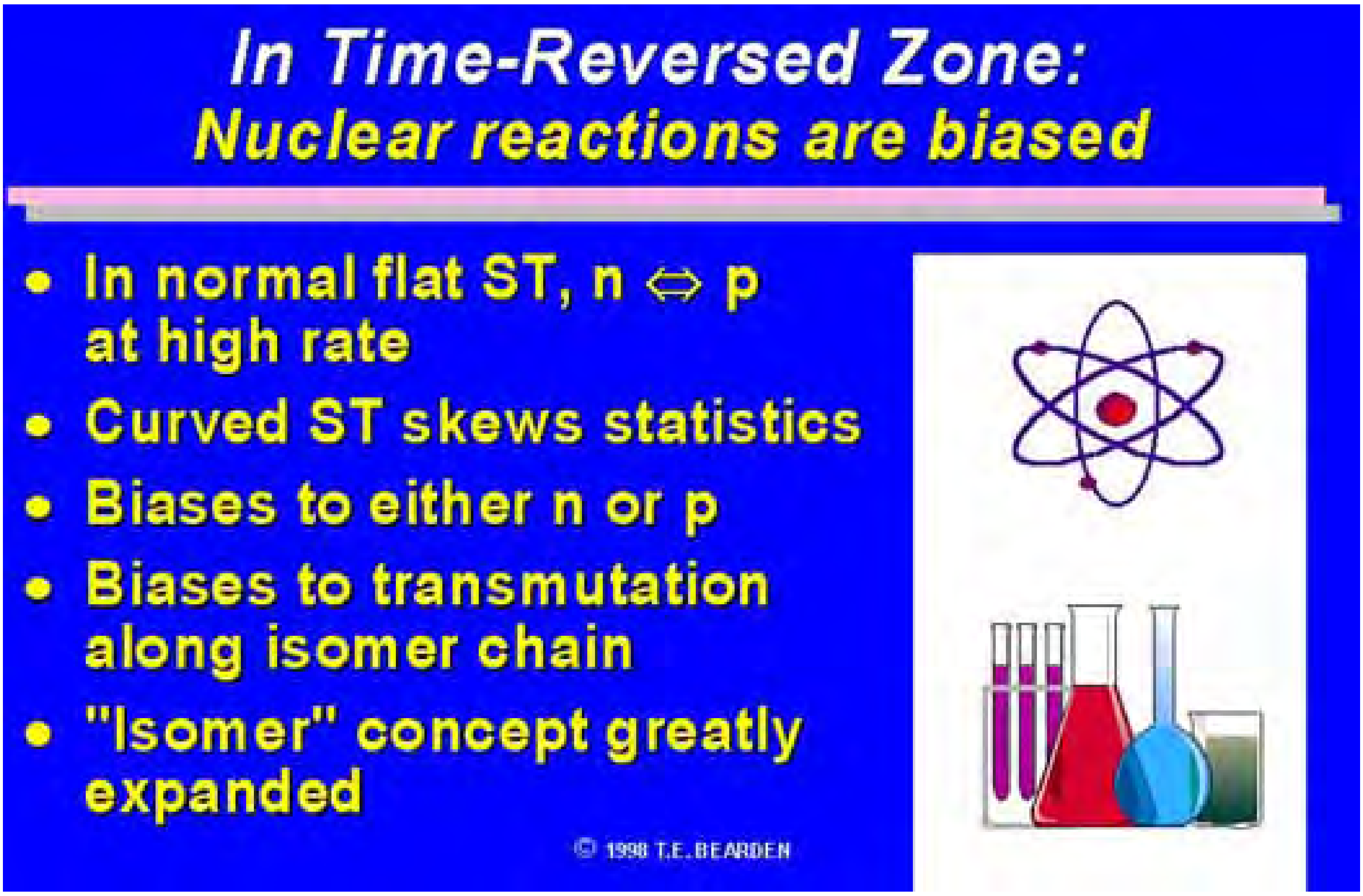

In Time-Reversed Zone:
Nuclear reactions are biased

- In normal flat ST, n ⇔ p at high rate
- Curved ST skews statistics
- Biases to either n or p
- Biases to transmutation along isomer chain
- "Isomer" concept greatly expanded

© 1998 T.E. BEARDEN

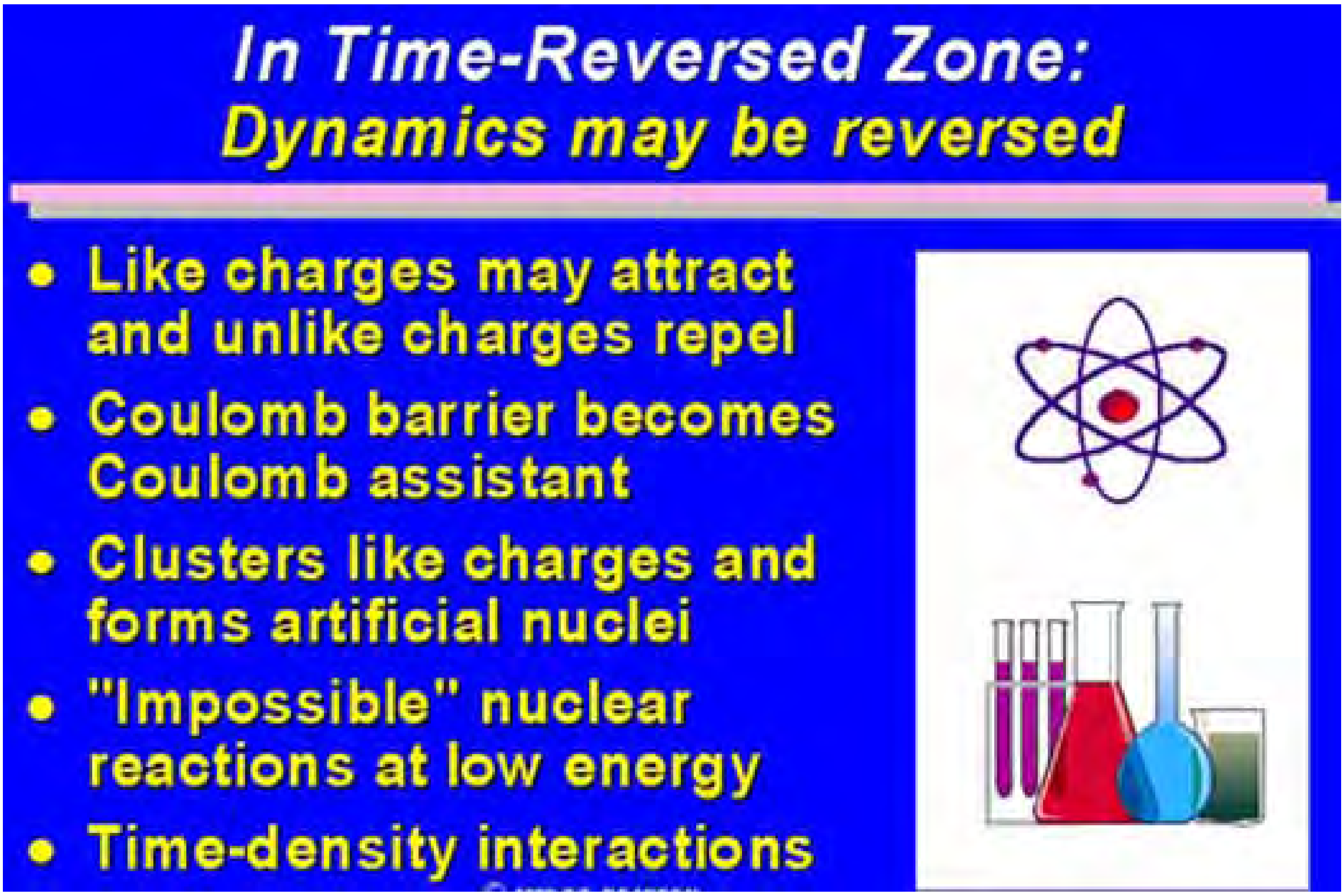

In Time-Reversed Zone:
Dynamics may be reversed

- Like charges may attract and unlike charges repel
- Coulomb barrier becomes Coulomb assistant
- Clusters like charges and forms artificial nuclei
- "Impossible" nuclear reactions at low energy
- Time-density interactions

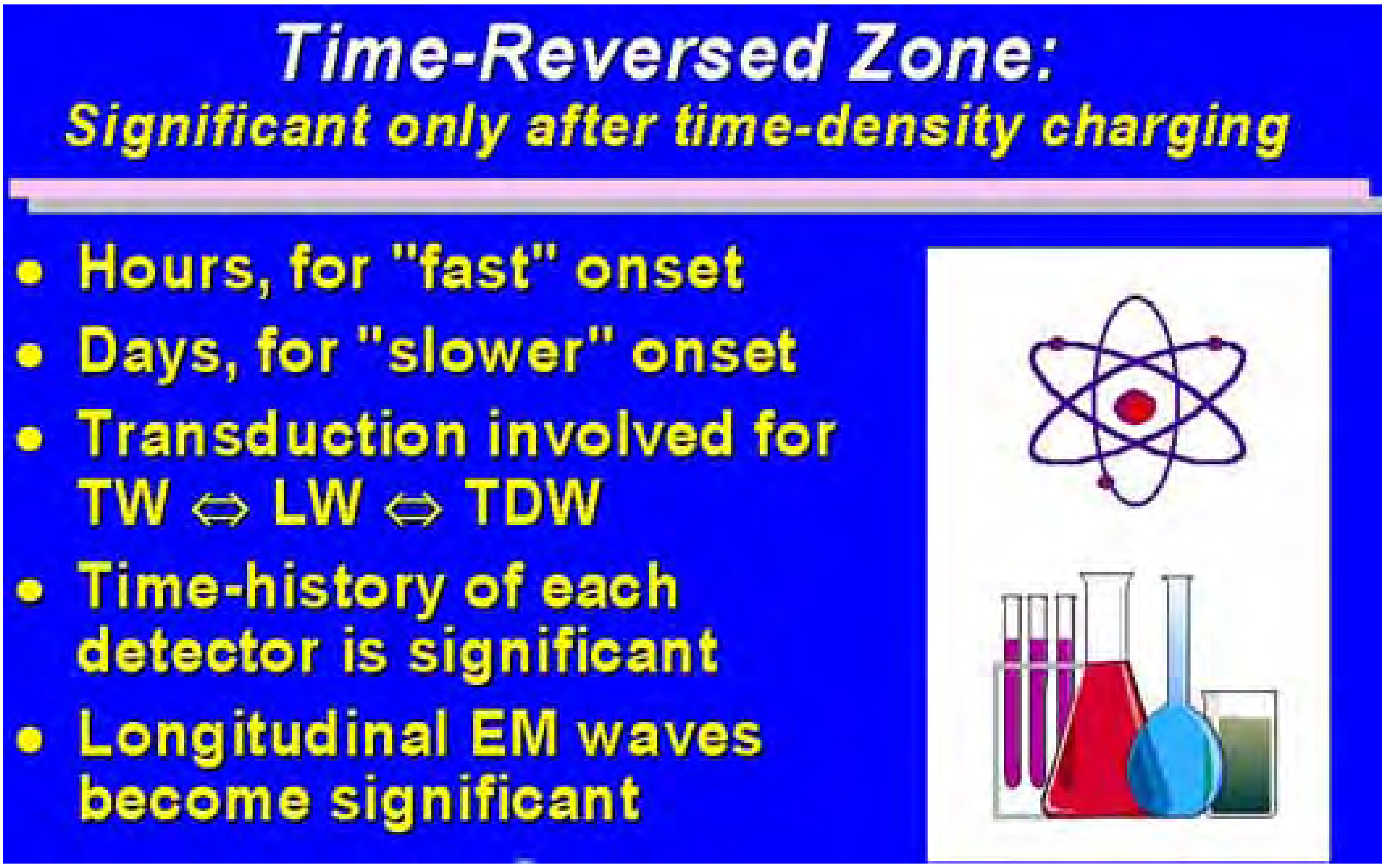

Time-Reversed Zone:

Significant only after time-density charging

- Hours, for "fast" onset
- Days, for "slower" onset
- Transduction involved for TW ⇔ LW ⇔ TDW
- Time-history of each detector is significant
- Longitudinal EM waves become significant

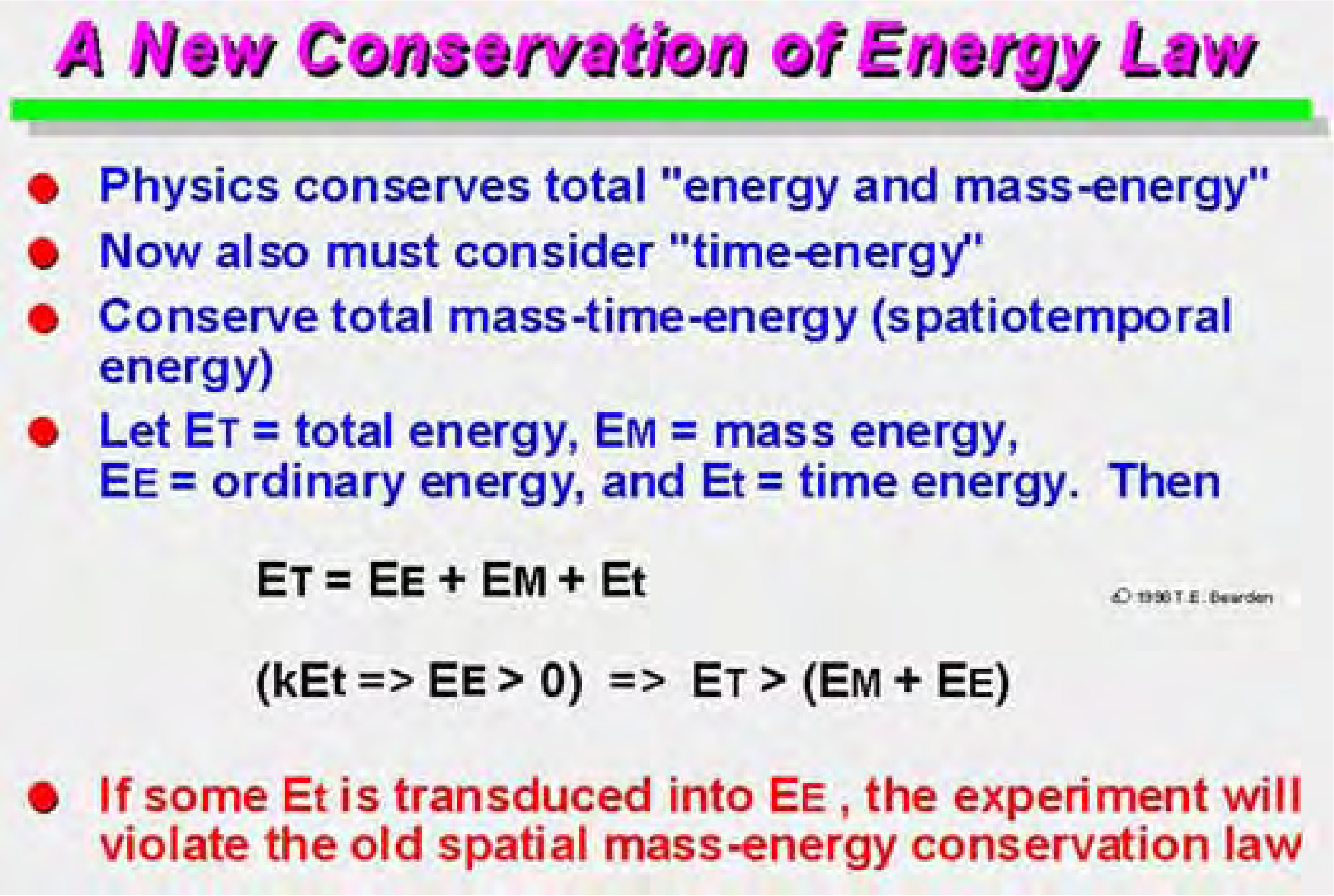

A New Conservation of Energy Law

- Physics conserves total "energy and mass-energy"
- Now also must consider "time-energy"
- Conserve total mass-time-energy (spatiotemporal energy)
- Let E_T = total energy, E_M = mass energy, E_E = ordinary energy, and E_t = time energy. Then

$$E_T = E_E + E_M + E_t$$

© 1998 T.E. Bearden

$$(kE_t \Rightarrow E_E > 0) \quad \Rightarrow \quad E_T > (E_M + E_E)$$

- If some E_t is transduced into E_E, the experiment will violate the old spatial mass-energy conservation law

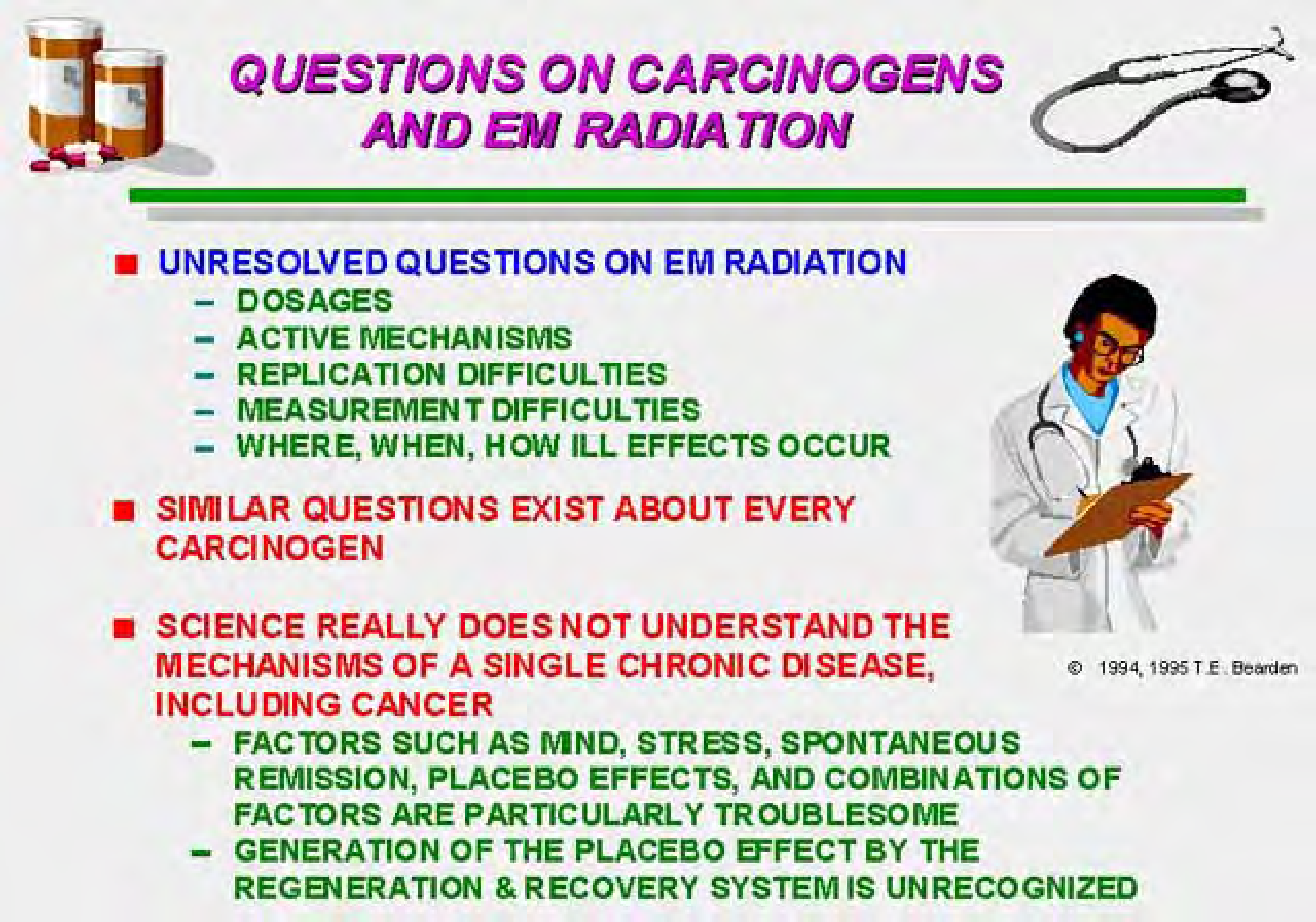

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

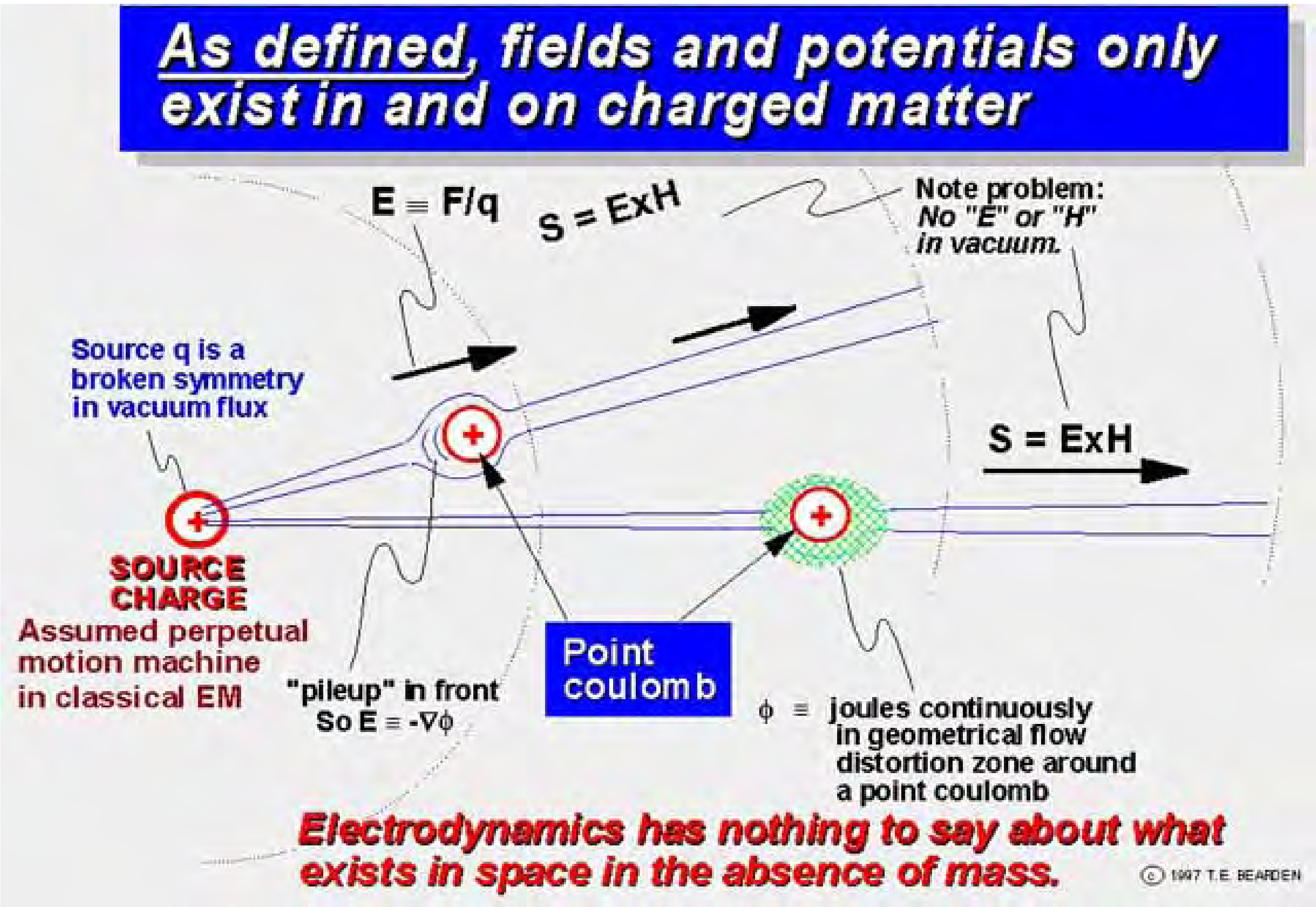

As defined, fields and potentials only exist in and on charged matter

$E \equiv F/q$

$S = E \times H$

Note problem:
No "E" or "H" in vacuum.

Source q is a broken symmetry in vacuum flux

$S = E \times H$

SOURCE CHARGE

Assumed perpetual motion machine in classical EM

"pileup" in front
So $E \equiv -\nabla\phi$

Point coulomb

$\phi \equiv$ joules continuously in geometrical flow distortion zone around a point coulomb

Electrodynamics has nothing to say about what exists in space in the absence of mass.

© 1997 T.E. BEARDEN

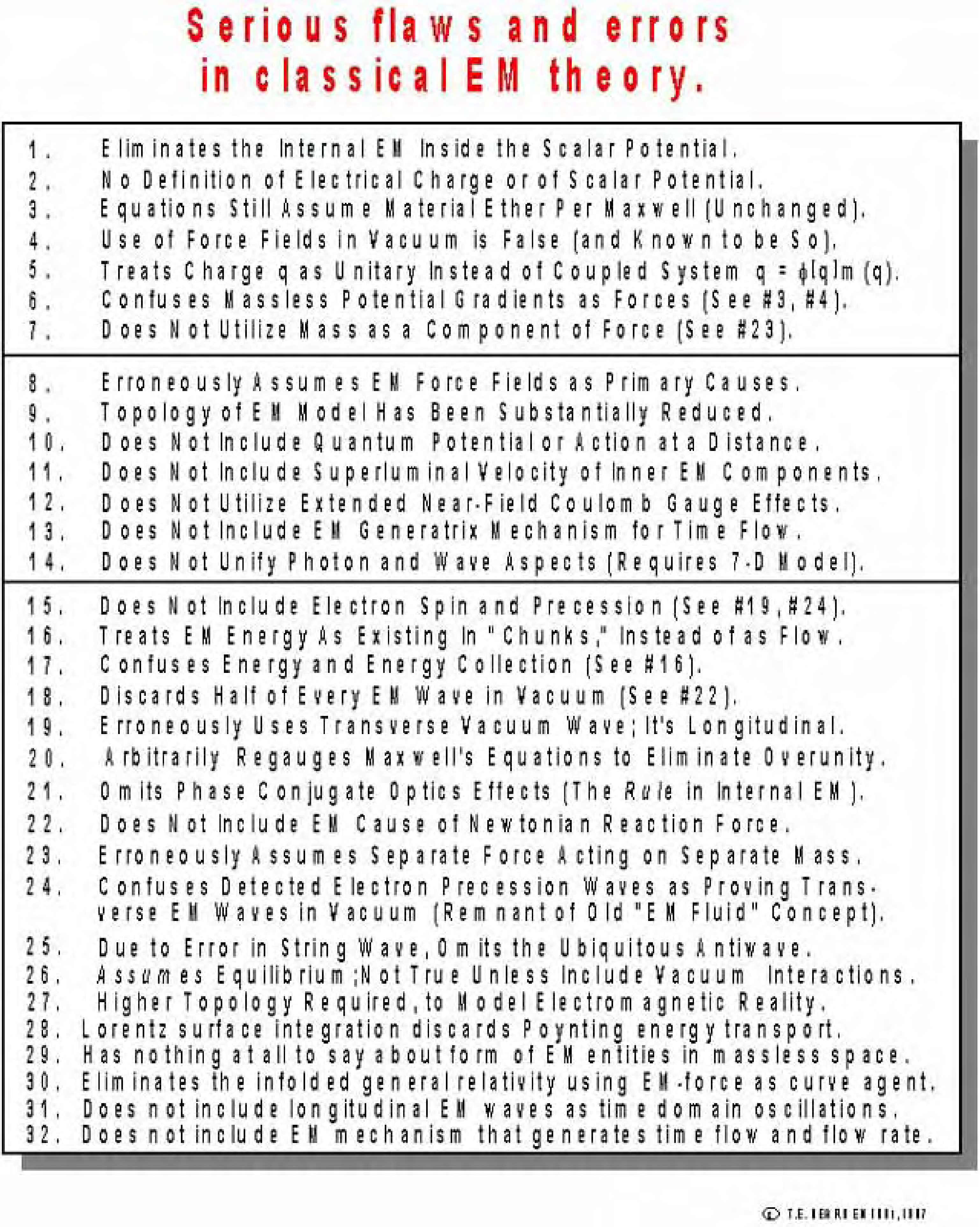

Serious flaws and errors in classical EM theory.

1. Eliminates the Internal EM Inside the Scalar Potential.
2. No Definition of Electrical Charge or of Scalar Potential.
3. Equations Still Assume Material Ether Per Maxwell (Unchanged).
4. Use of Force Fields in Vacuum is False (and Known to be So).
5. Treats Charge q as Unitary Instead of Coupled System $q = \phi[q]m(q)$.
6. Confuses Massless Potential Gradients as Forces (See #3, #4).
7. Does Not Utilize Mass as a Component of Force (See #23).

8. Erroneously Assumes EM Force Fields as Primary Causes.
9. Topology of EM Model Has Been Substantially Reduced.
10. Does Not Include Quantum Potential or Action at a Distance.
11. Does Not Include Superluminal Velocity of Inner EM Components.
12. Does Not Utilize Extended Near-Field Coulomb Gauge Effects.
13. Does Not Include EM Generatrix Mechanism for Time Flow.
14. Does Not Unify Photon and Wave Aspects (Requires 7-D Model).

15. Does Not Include Electron Spin and Precession (See #19, #24).
16. Treats EM Energy As Existing In "Chunks," Instead of as Flow.
17. Confuses Energy and Energy Collection (See #16).
18. Discards Half of Every EM Wave in Vacuum (See #22).
19. Erroneously Uses Transverse Vacuum Wave; It's Longitudinal.
20. Arbitrarily Regauges Maxwell's Equations to Eliminate Overunity.
21. Omits Phase Conjugate Optics Effects (The *Rule* in Internal EM).
22. Does Not Include EM Cause of Newtonian Reaction Force.
23. Erroneously Assumes Separate Force Acting on Separate Mass.
24. Confuses Detected Electron Precession Waves as Proving Transverse EM Waves in Vacuum (Remnant of Old "EM Fluid" Concept).
25. Due to Error in String Wave, Omits the Ubiquitous Antiwave.
26. *Assum*es Equilibrium; Not True Unless Include Vacuum Interactions.
27. Higher Topology Required, to Model Electromagnetic Reality.
28. Lorentz surface integration discards Poynting energy transport.
29. Has nothing at all to say about form of EM entities in massless space.
30. Eliminates the infolded general relativity using EM-force as curve agent.
31. Does not include longitudinal EM waves as time domain oscillations.
32. Does not include EM mechanism that generates time flow and flow rate.

© T.E. BEARDEN 1991, 1997

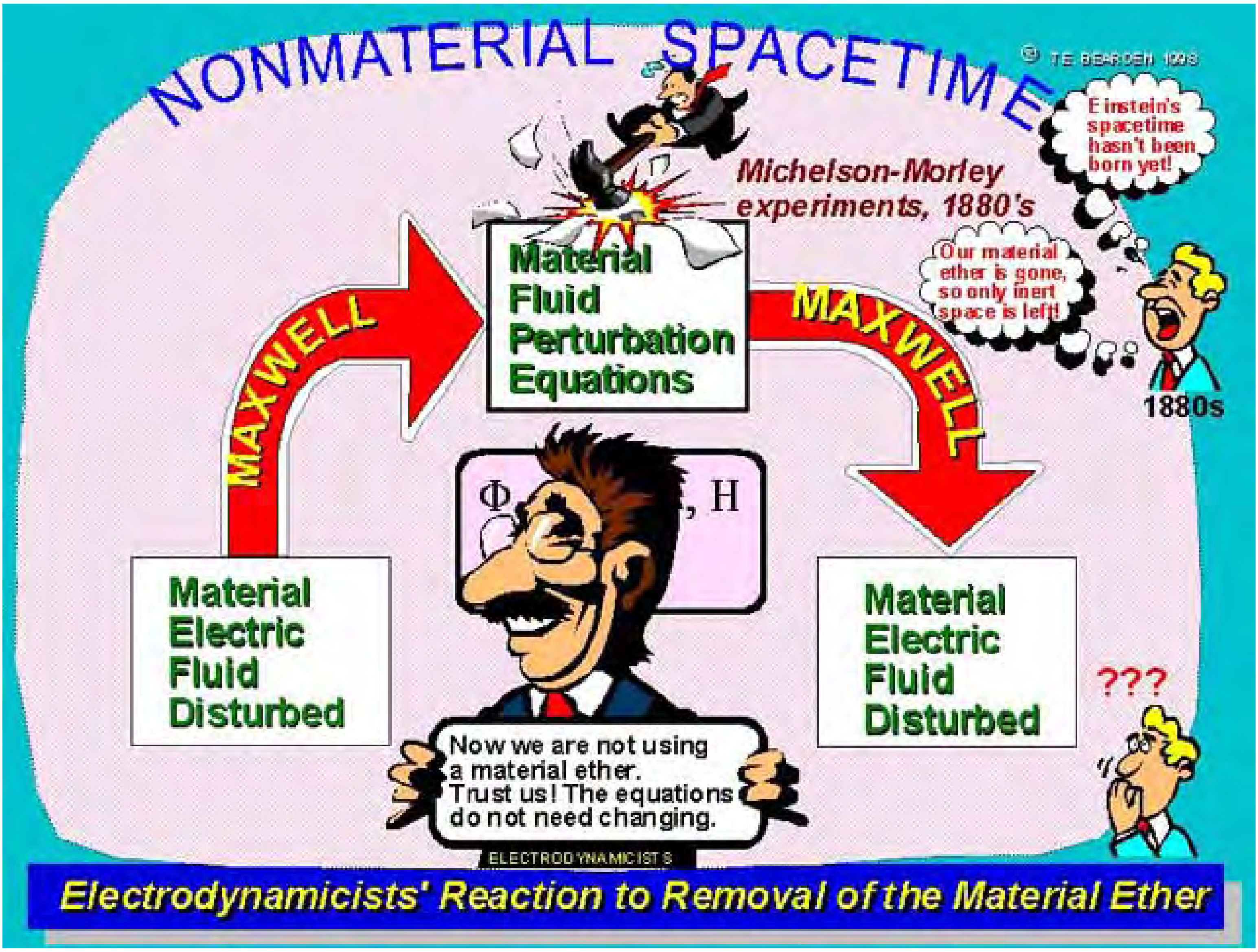
NONMATERIAL SPACETIME
© T.E. BEARDEN 1998
Einstein's spacetime hasn't been born yet!
Michelson-Morley experiments, 1880's
Our material ether is gone, so only inert space is left!
1880s
MAXWELL
Material Fluid Perturbation Equations
MAXWELL
Φ
, H
Material Electric Fluid Disturbed
Material Electric Fluid Disturbed
???
Now we are not using a material ether. Trust us! The equations do not need changing.
ELECTRODYNAMICISTS

Electrodynamicists' Reaction to Removal of the Material Ether

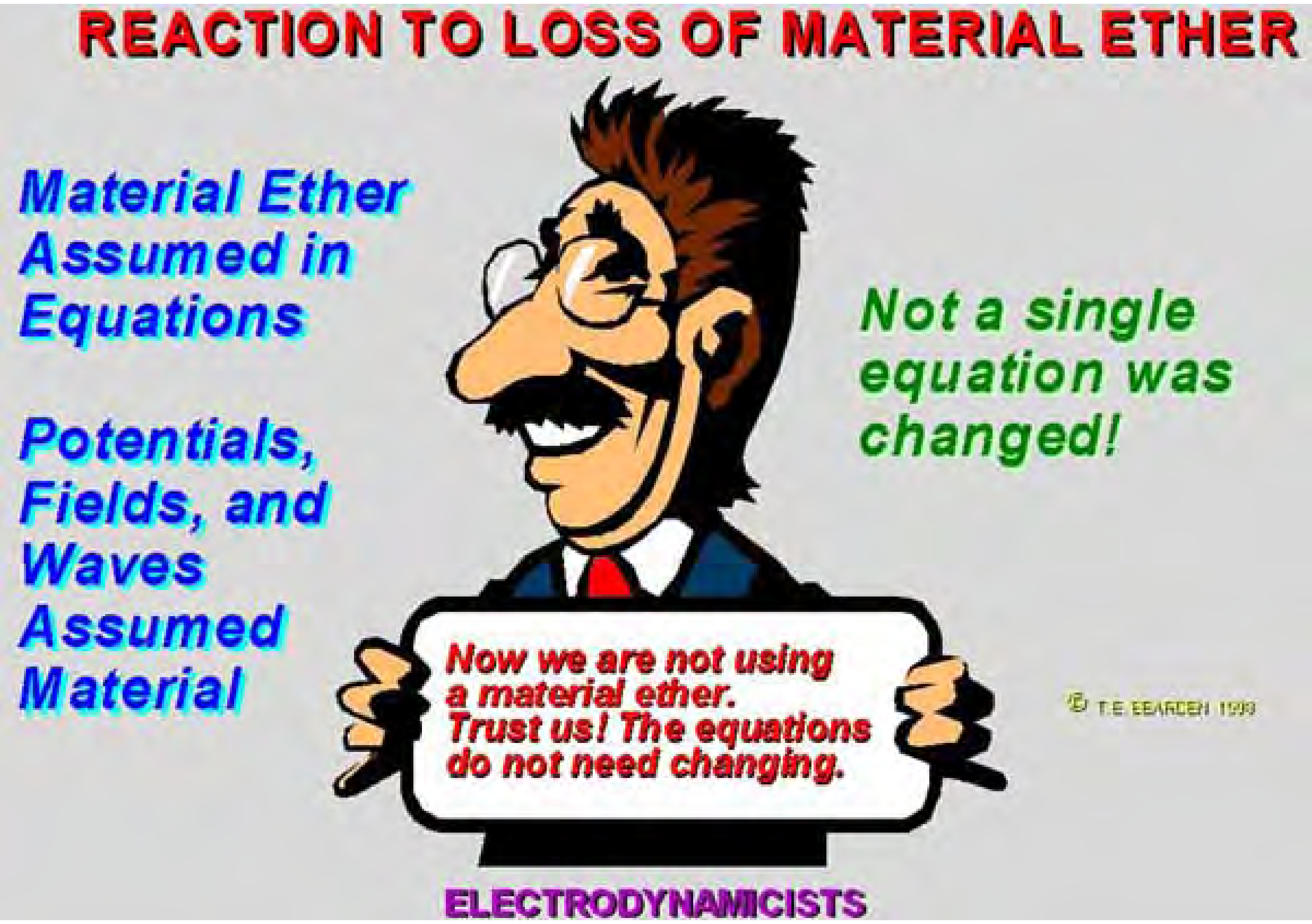

REACTION TO LOSS OF MATERIAL ETHER

Material Ether
Assumed in
Equations

Potentials,
Fields, and
Waves
Assumed
Material

Not a single
equation was
changed!

Now we are not using
a material ether.
Trust us! The equations
do not need changing.

© T.E. BEARDEN 1999

ELECTRODYNAMICISTS

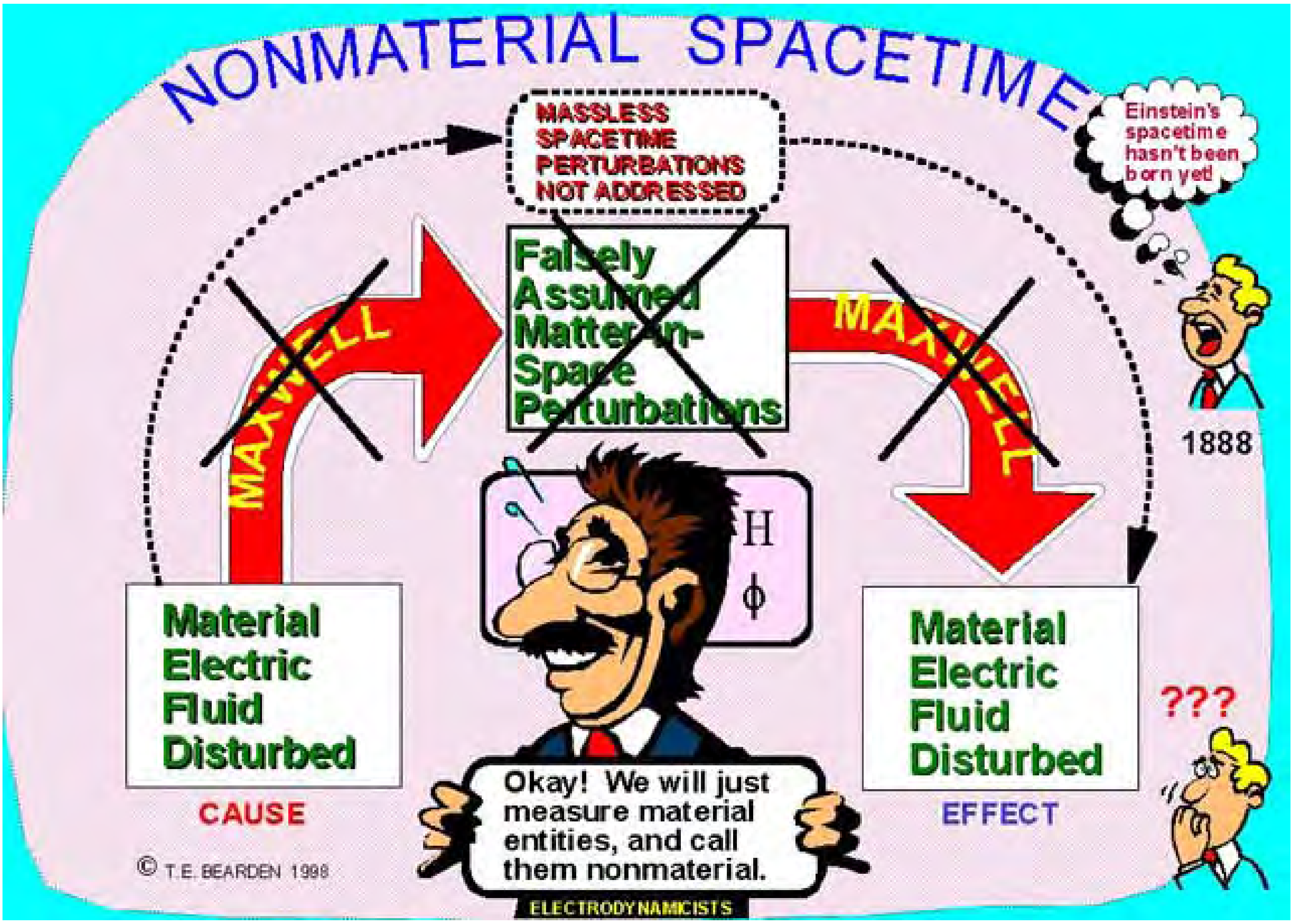
NONMATERIAL SPACETIME
MASSLESS SPACETIME PERTURBATIONS NOT ADDRESSED
Einstein's spacetime hasn't been born yet!
1888
MAXWELL
Falsely Assumed Matter-in-Space Perturbations
MAXWELL
H
ϕ
Material Electric Fluid Disturbed
CAUSE
Material Electric Fluid Disturbed
EFFECT
???
Okay! We will just measure material entities, and call them nonmaterial.
ELECTRODYNAMICISTS
© T.E. BEARDEN 1998

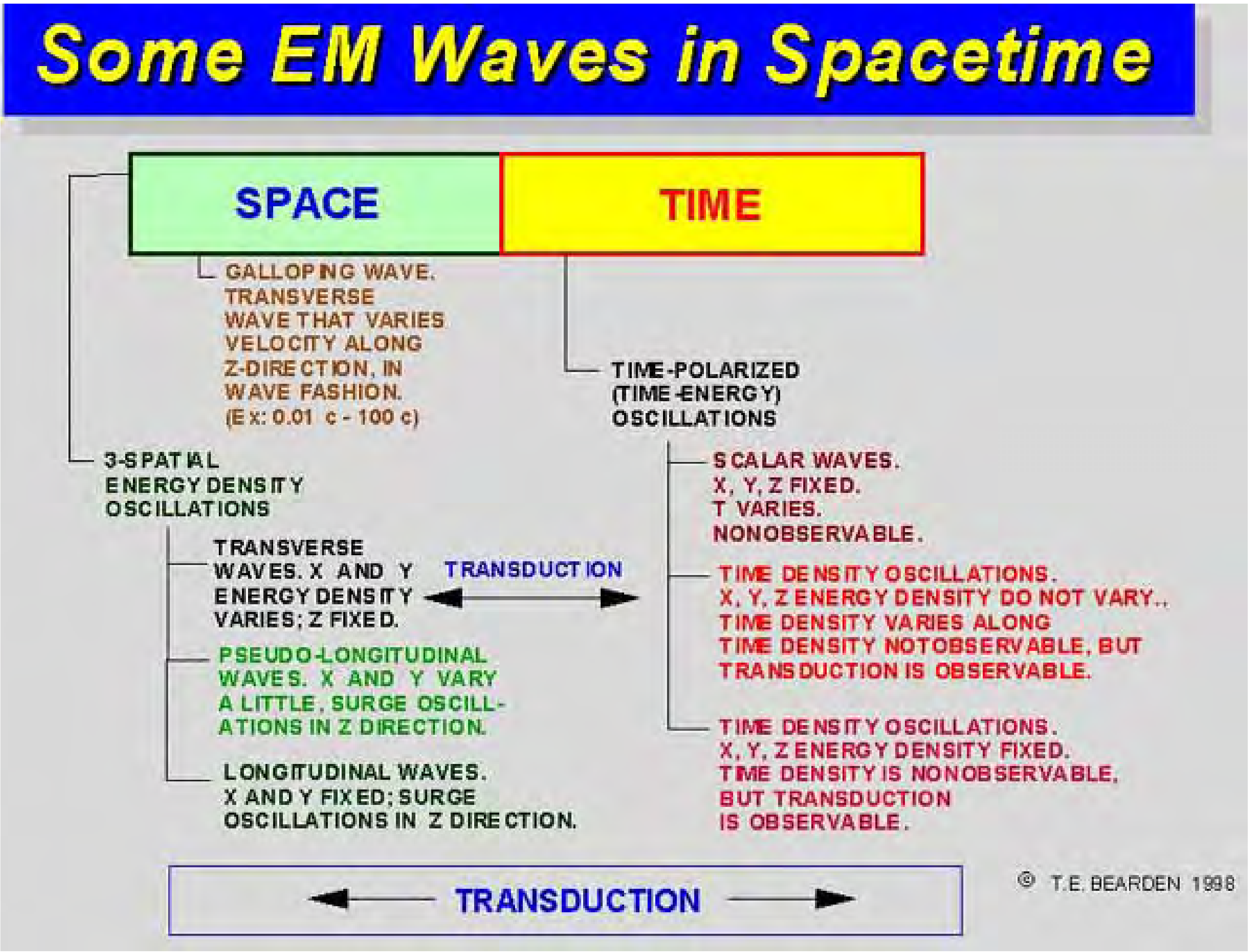

Some EM Waves in Spacetime

SPACE

TIME

GALLOPING WAVE.
TRANSVERSE
WAVE THAT VARIES
VELOCITY ALONG
Z-DIRECTION, IN
WAVE FASHION.
(Ex: 0.01 c - 100 c)

TIME-POLARIZED
(TIME-ENERGY)
OSCILLATIONS

3-SPATIAL
ENERGY DENSITY
OSCILLATIONS

TRANSVERSE
WAVES. X AND Y
ENERGY DENSITY
VARIES; Z FIXED.

TRANSDUCTION

PSEUDO-LONGITUDINAL
WAVES. X AND Y VARY
A LITTLE, SURGE OSCILL-
ATIONS IN Z DIRECTION.

LONGITUDINAL WAVES.
X AND Y FIXED; SURGE
OSCILLATIONS IN Z DIRECTION.

SCALAR WAVES.
X, Y, Z FIXED.
T VARIES.
NONOBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY DO NOT VARY..
TIME DENSITY VARIES ALONG
TIME DENSITY NOTOBSERVABLE, BUT
TRANSDUCTION IS OBSERVABLE.

TIME DENSITY OSCILLATIONS.
X, Y, Z ENERGY DENSITY FIXED.
TIME DENSITY IS NONOBSERVABLE,
BUT TRANSDUCTION
IS OBSERVABLE.

TRANSDUCTION

© T.E. BEARDEN 1998

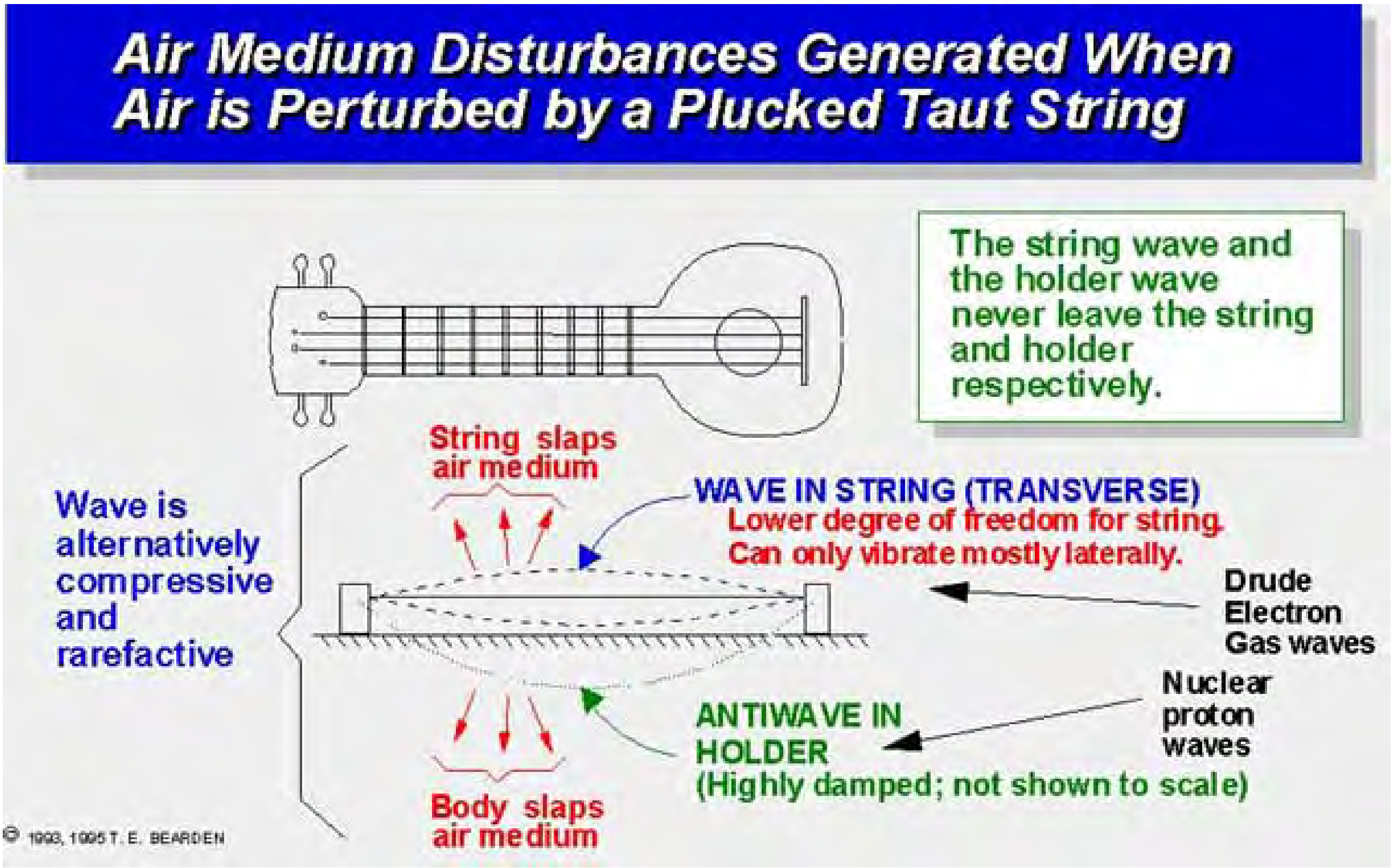

Air Medium Disturbances Generated When Air is Perturbed by a Plucked Taut String

The string wave and the holder wave never leave the string and holder respectively.

String slaps air medium

WAVE IN STRING (TRANSVERSE)

Lower degree of freedom for string.
Can only vibrate mostly laterally.

Wave is alternatively compressive and rarefactive

Drude Electron Gas waves

Nuclear proton waves

ANTIWAVE IN HOLDER
(Highly damped; not shown to scale)

Body slaps air medium

© 1993, 1995 T. E. BEARDEN

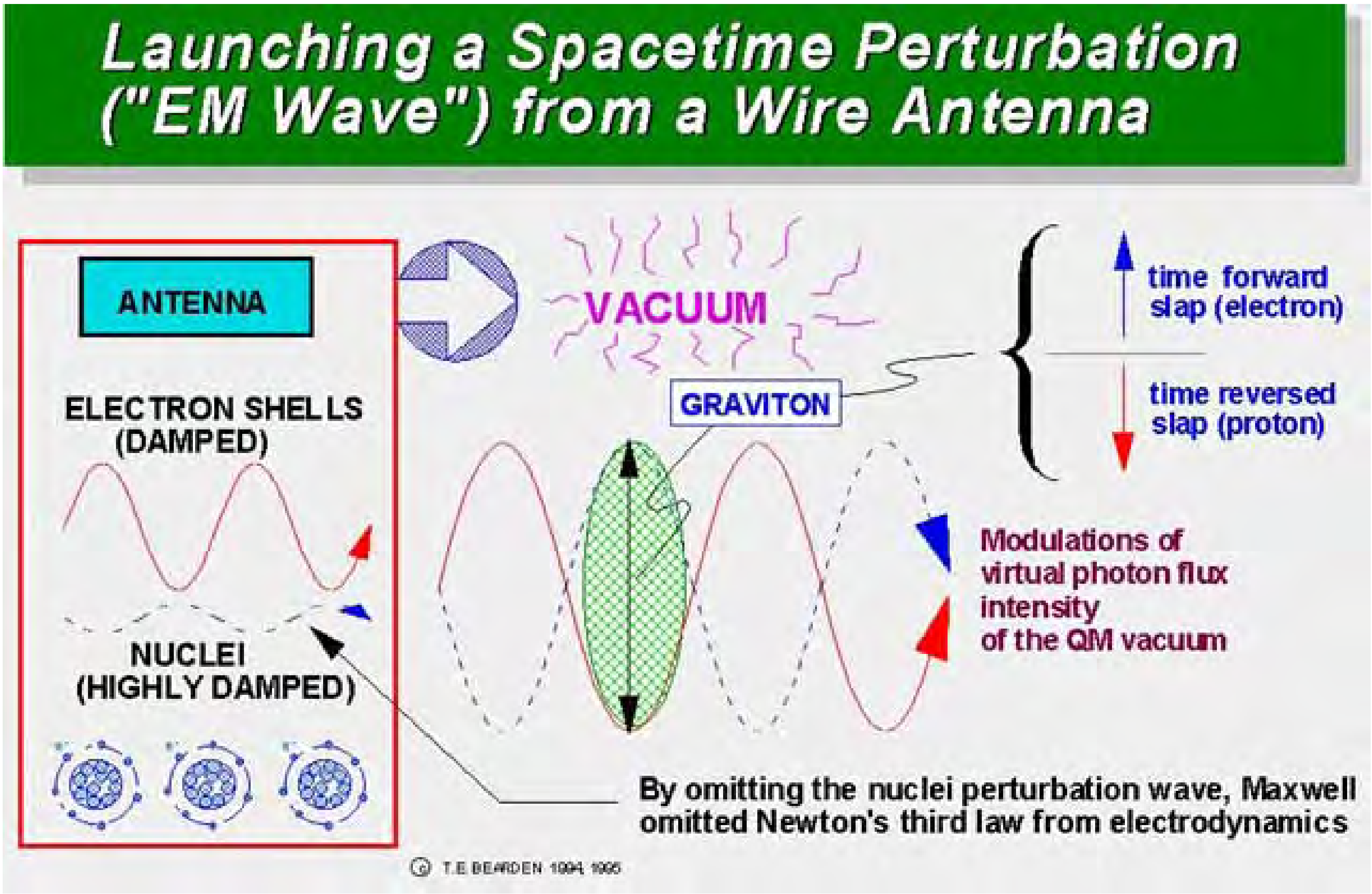

Launching a Spacetime Perturbation ("EM Wave") from a Wire Antenna

ANTENNA

ELECTRON SHELLS
(DAMPED)

NUCLEI
(HIGHLY DAMPED)

VACUUM

GRAVITON

time forward
slap (electron)

time reversed
slap (proton)

Modulations of
virtual photon flux
intensity
of the QM vacuum

By omitting the nuclei perturbation wave, Maxwell omitted Newton's third law from electrodynamics

© T.E. BEARDEN 1994, 1995

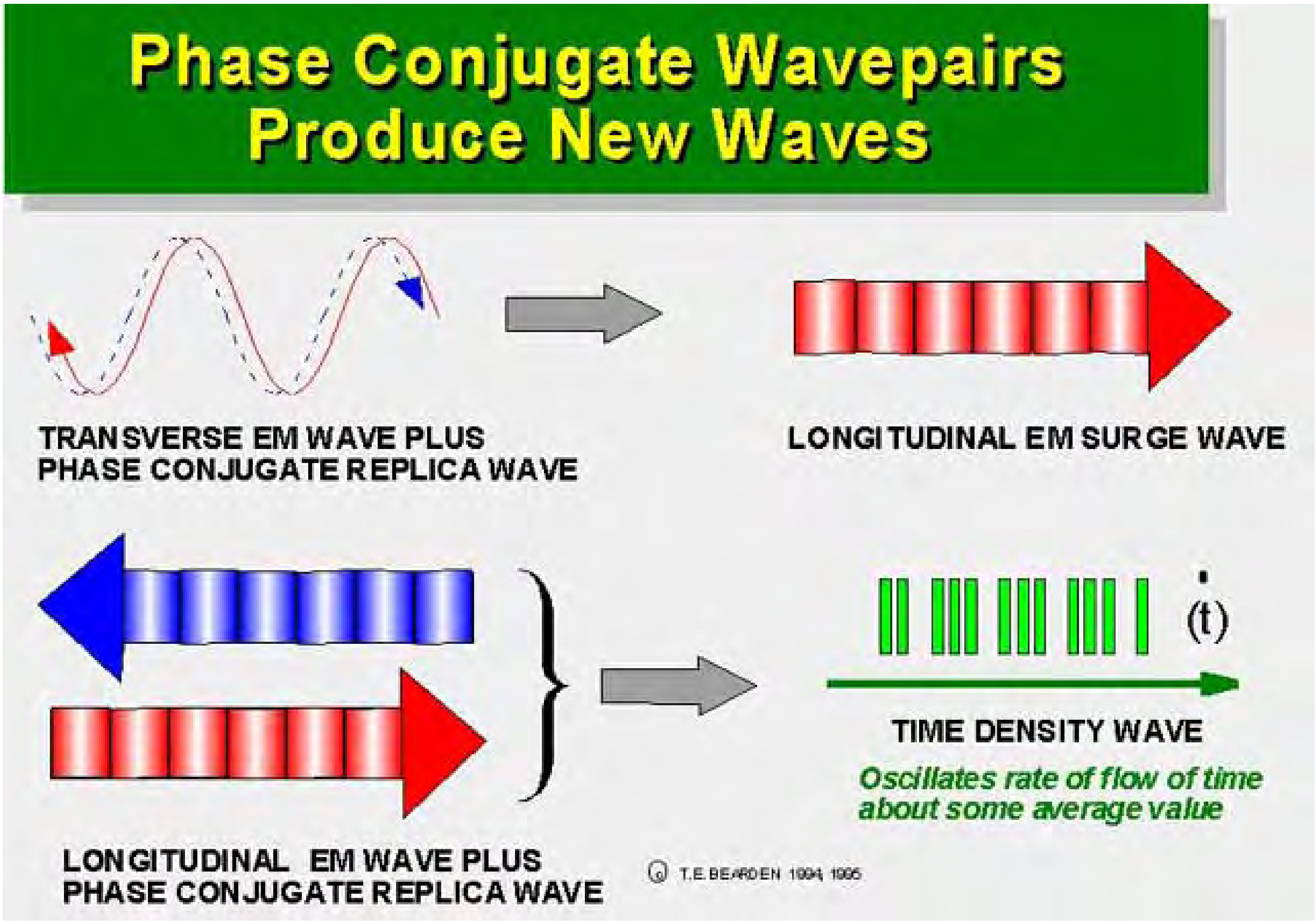

Phase Conjugate Wavepairs Produce New Waves

TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

LONGITUDINAL EM SURGE WAVE

LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE

TIME DENSITY WAVE

Oscillates rate of flow of time about some average value

© T.E. BEARDEN 1994, 1995

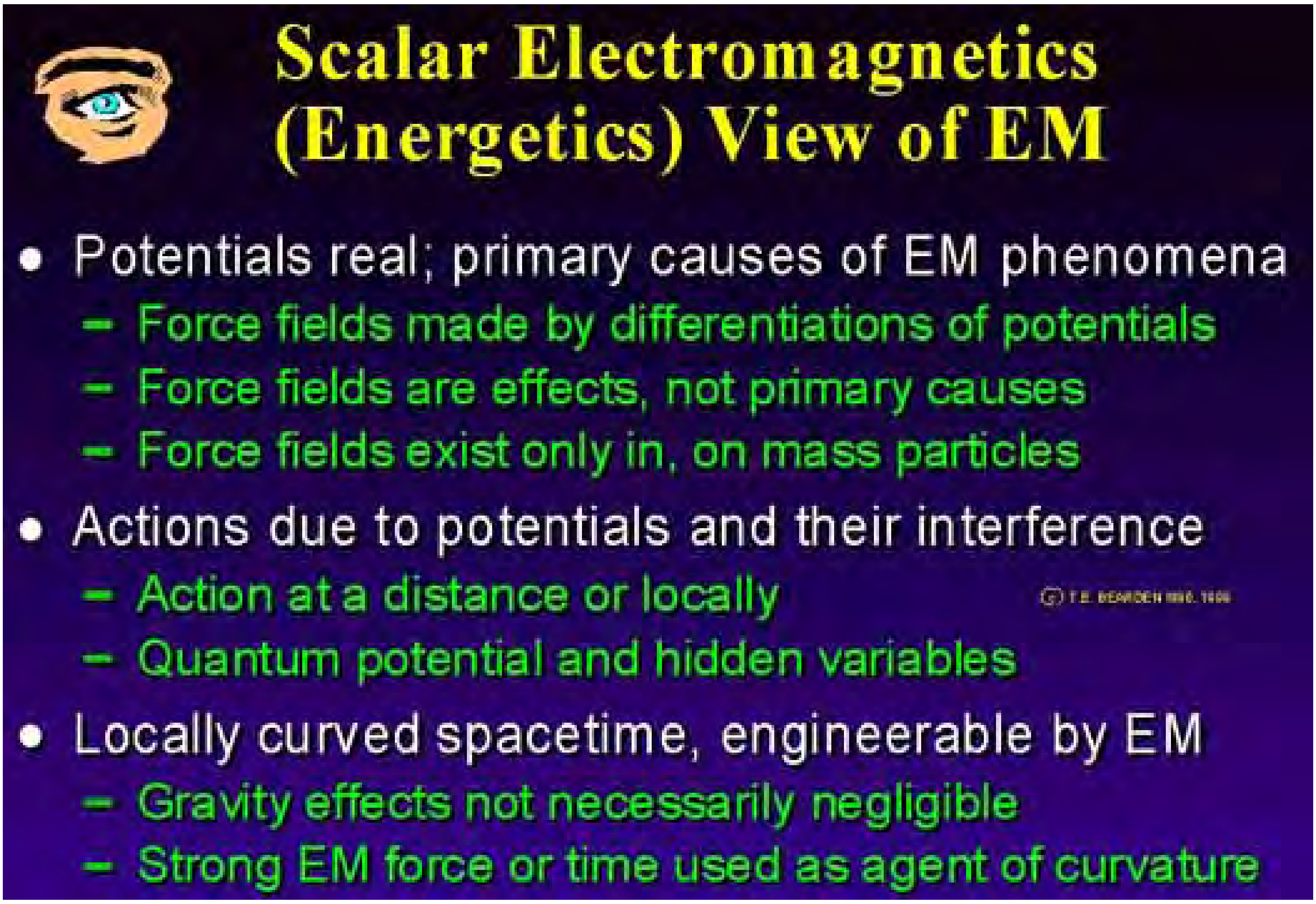

Scalar Electromagnetics (Energetics) View of EM

- Potentials real; primary causes of EM phenomena
  - Force fields made by differentiations of potentials
  - Force fields are effects, not primary causes
  - Force fields exist only in, on mass particles
- Actions due to potentials and their interference
  - Action at a distance or locally
  - Quantum potential and hidden variables
- Locally curved spacetime, engineerable by EM
  - Gravity effects not necessarily negligible
  - Strong EM force or time used as agent of curvature

© T.E. BEARDEN [illegible]

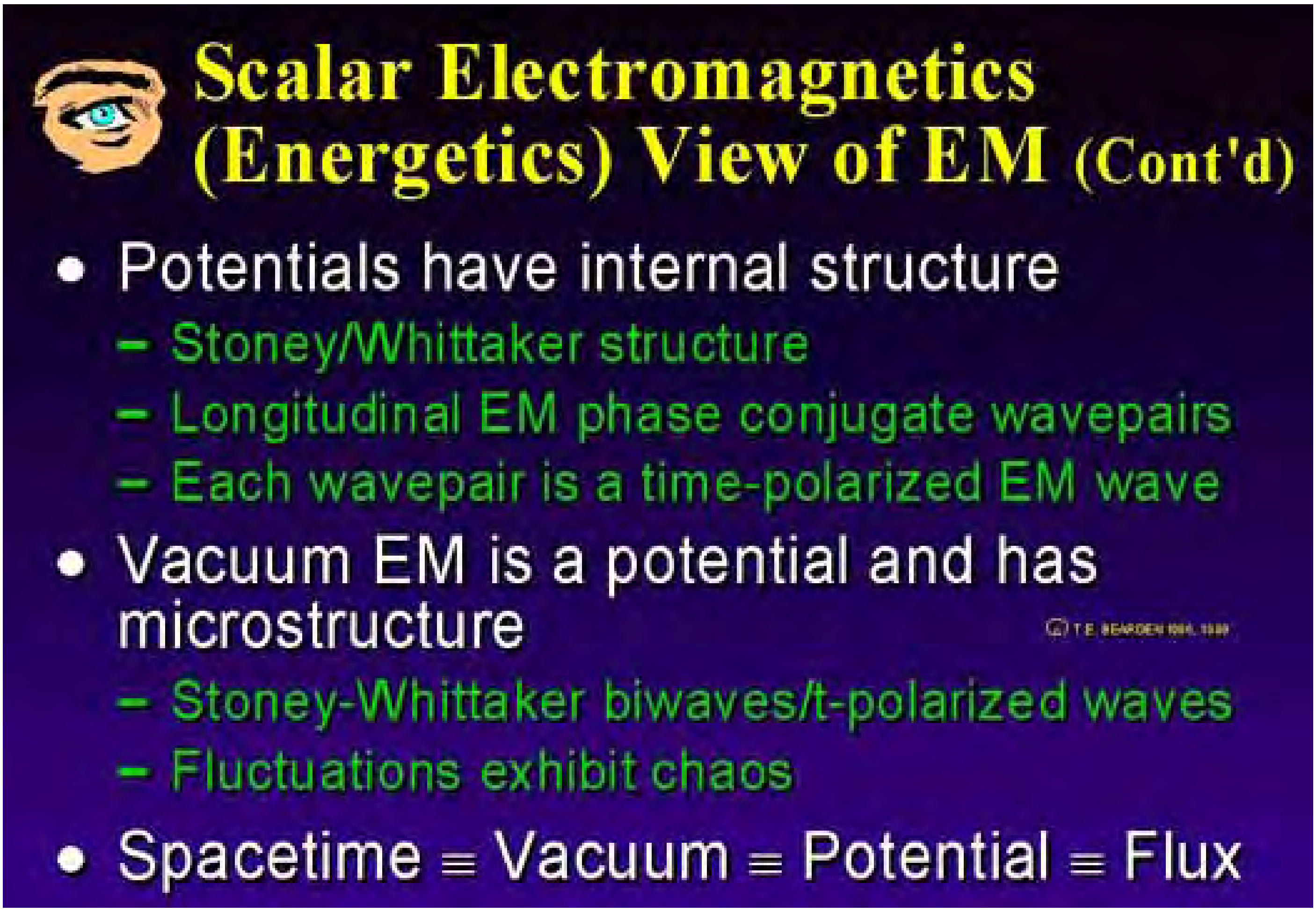

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Potentials have internal structure
  - Stoney/Whittaker structure
  - Longitudinal EM phase conjugate wavepairs
  - Each wavepair is a time-polarized EM wave
- Vacuum EM is a potential and has microstructure
  - Stoney-Whittaker biwaves/t-polarized waves
  - Fluctuations exhibit chaos
- Spacetime ≡ Vacuum ≡ Potential ≡ Flux

© T.E. BEARDEN [illegible]

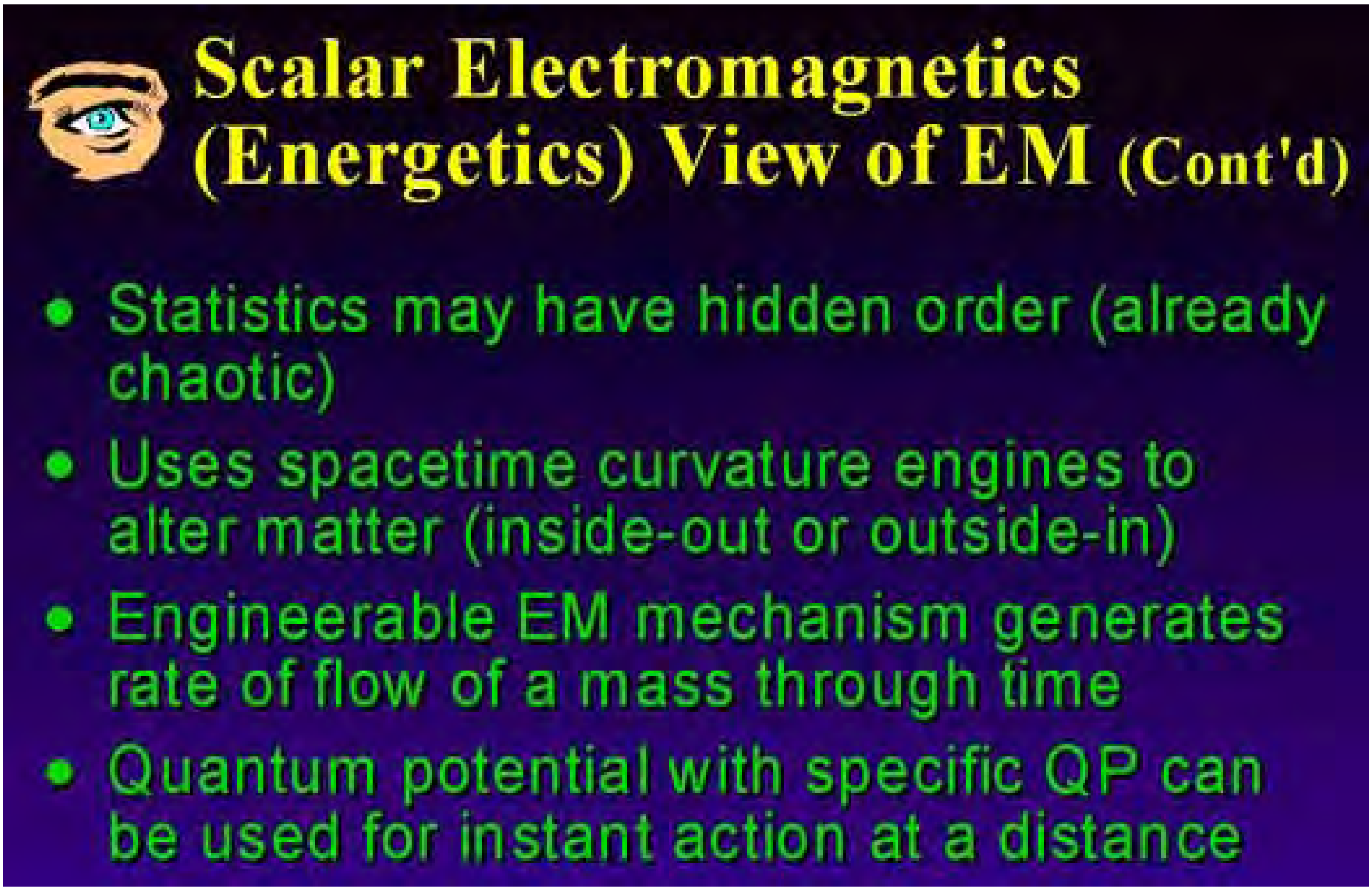

Scalar Electromagnetics (Energetics) View of EM (Cont'd)

- Statistics may have hidden order (already chaotic)
- Uses spacetime curvature engines to alter matter (inside-out or outside-in)
- Engineerable EM mechanism generates rate of flow of a mass through time
- Quantum potential with specific QP can be used for instant action at a distance

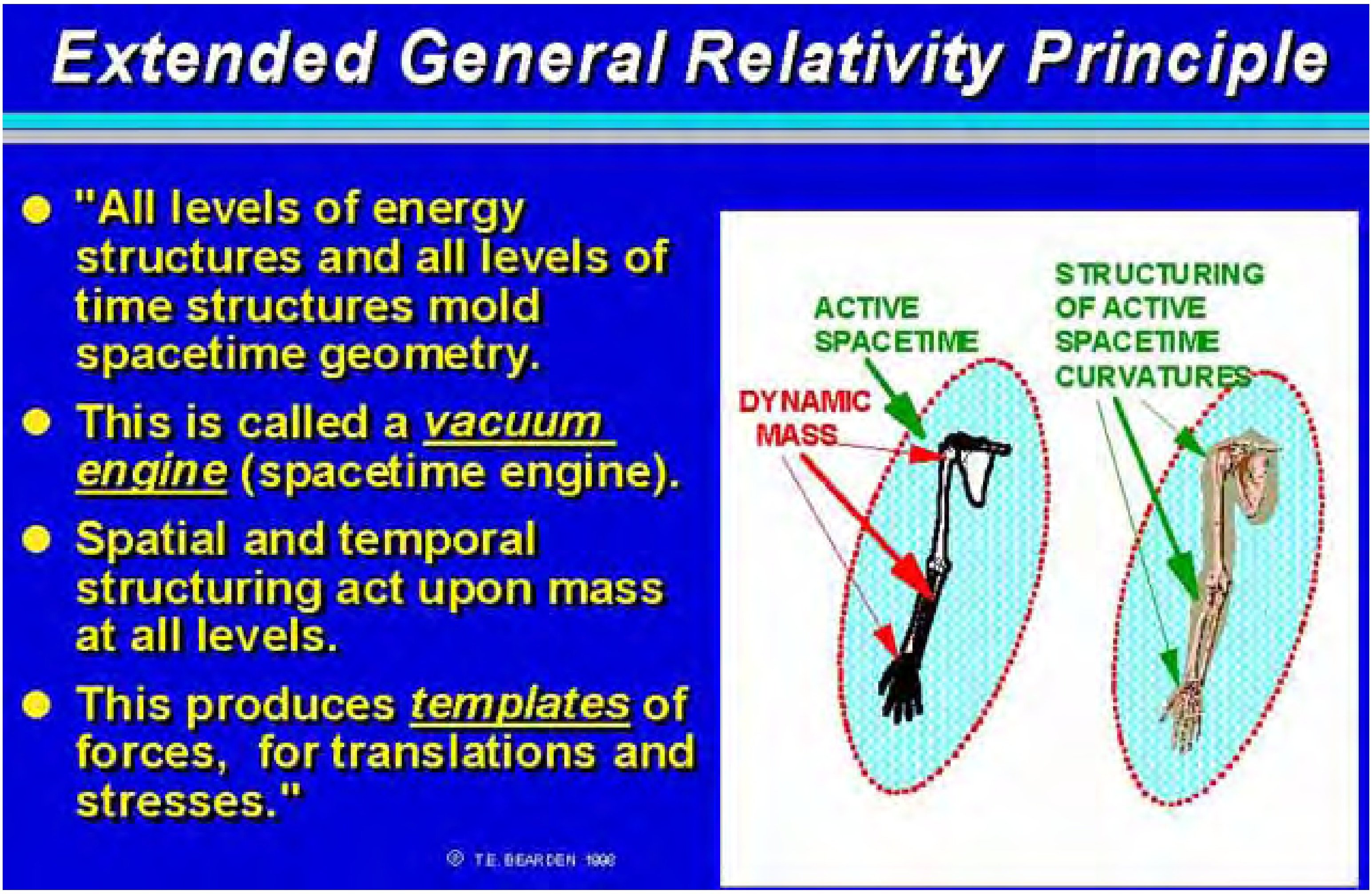

Extended General Relativity Principle

- "All levels of energy structures and all levels of time structures mold spacetime geometry.
- This is called a ***vacuum engine*** (spacetime engine).
- Spatial and temporal structuring act upon mass at all levels.
- This produces ***templates*** of forces, for translations and stresses."

© T.E. BEARDEN 1990

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

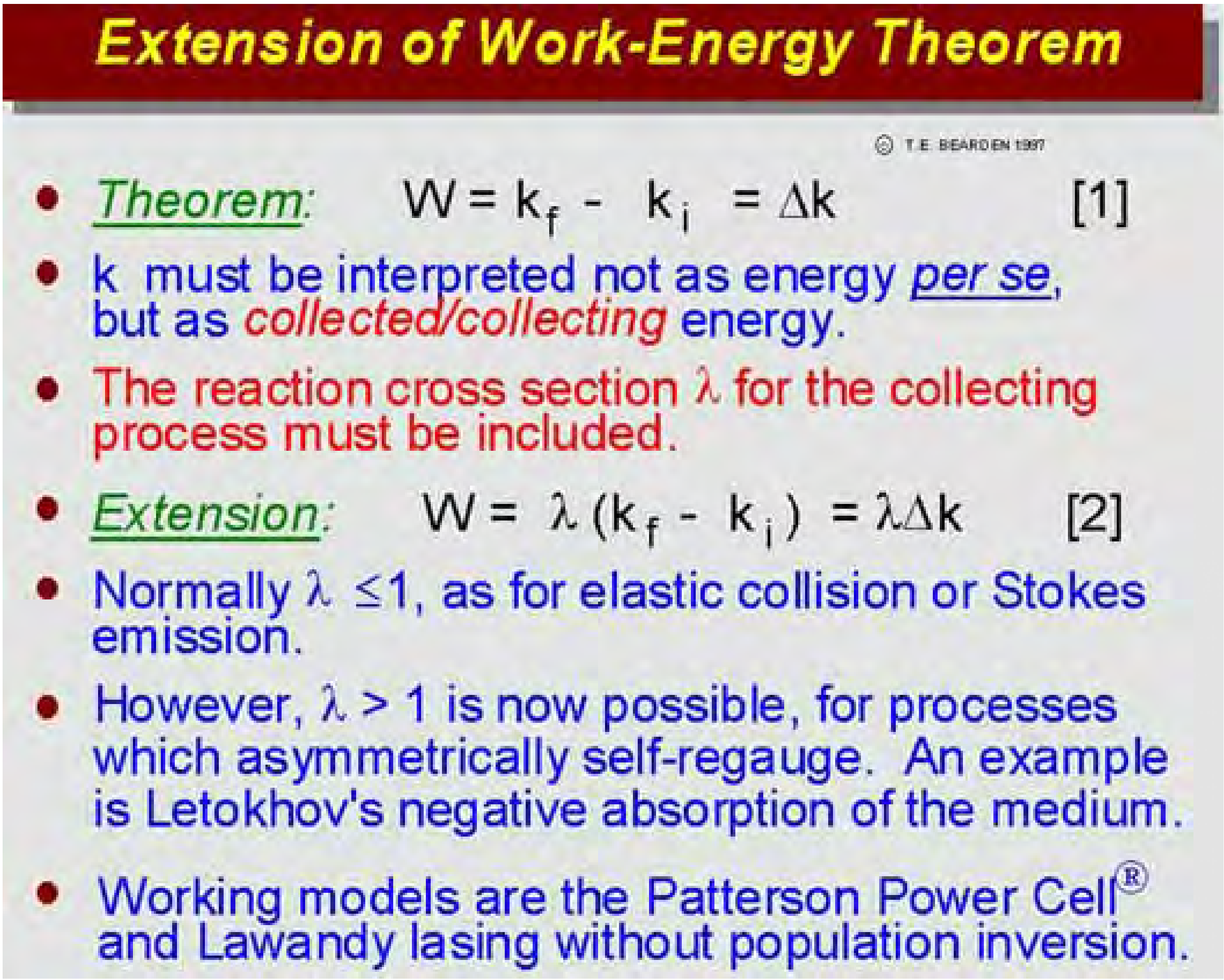

Extension of Work-Energy Theorem

© T.E. BEARDEN 1997

- *Theorem:* $W = k_f - k_i = \Delta k$ [1]
- k must be interpreted not as energy *per se*, but as *collected/collecting* energy.
- The reaction cross section λ for the collecting process must be included.
- *Extension:* $W = \lambda (k_f - k_i) = \lambda \Delta k$ [2]
- Normally $\lambda \leq 1$, as for elastic collision or Stokes emission.
- However, $\lambda > 1$ is now possible, for processes which asymmetrically self-regauge. An example is Letokhov's negative absorption of the medium.
- Working models are the Patterson Power Cell® and Lawandy lasing without population inversion.

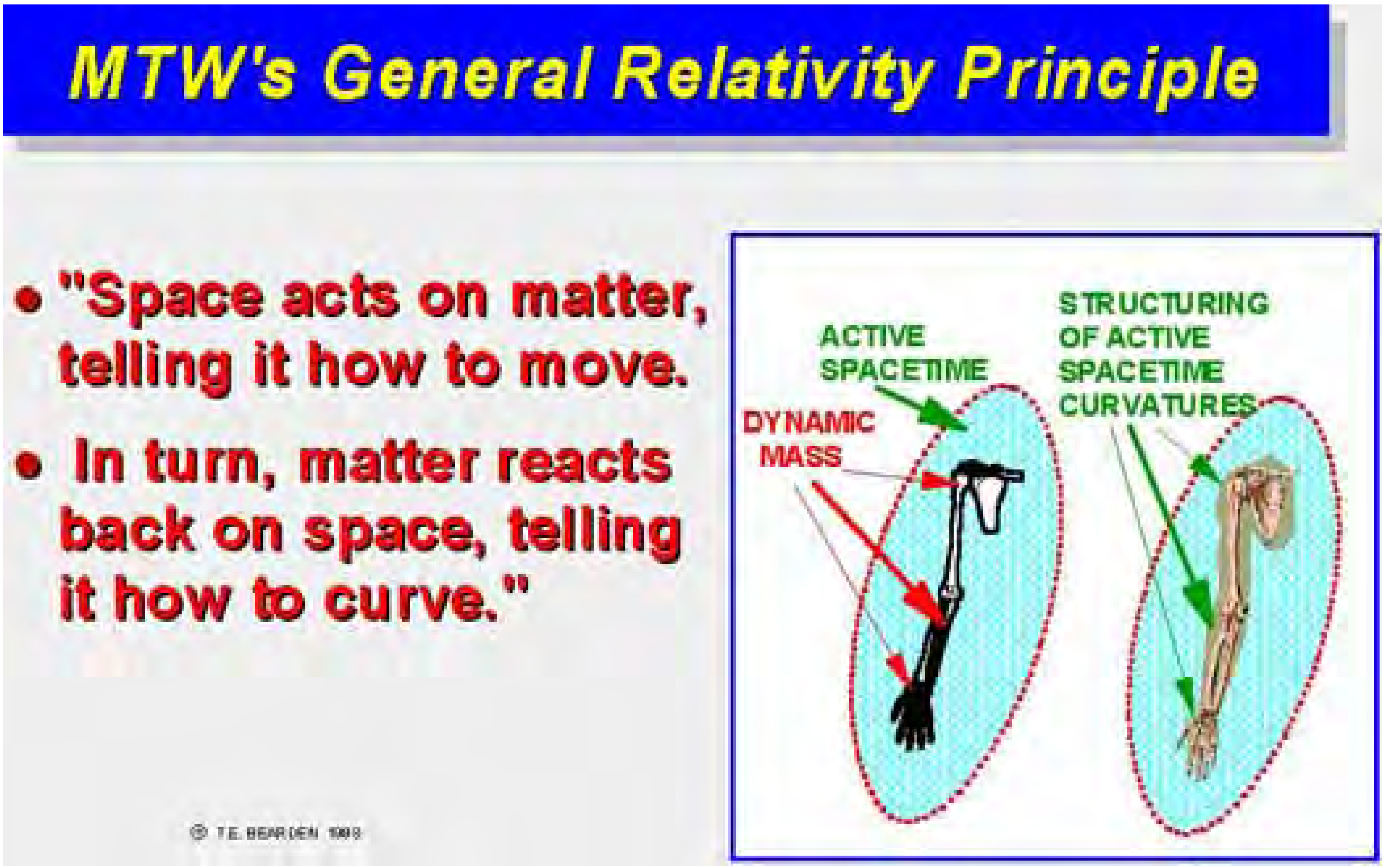

MTW's General Relativity Principle

- "Space acts on matter, telling it how to move.
- In turn, matter reacts back on space, telling it how to curve."

ACTIVE SPACETIME
STRUCTURING OF ACTIVE SPACETIME CURVATURES
DYNAMIC MASS

© T.E. BEARDEN 1998

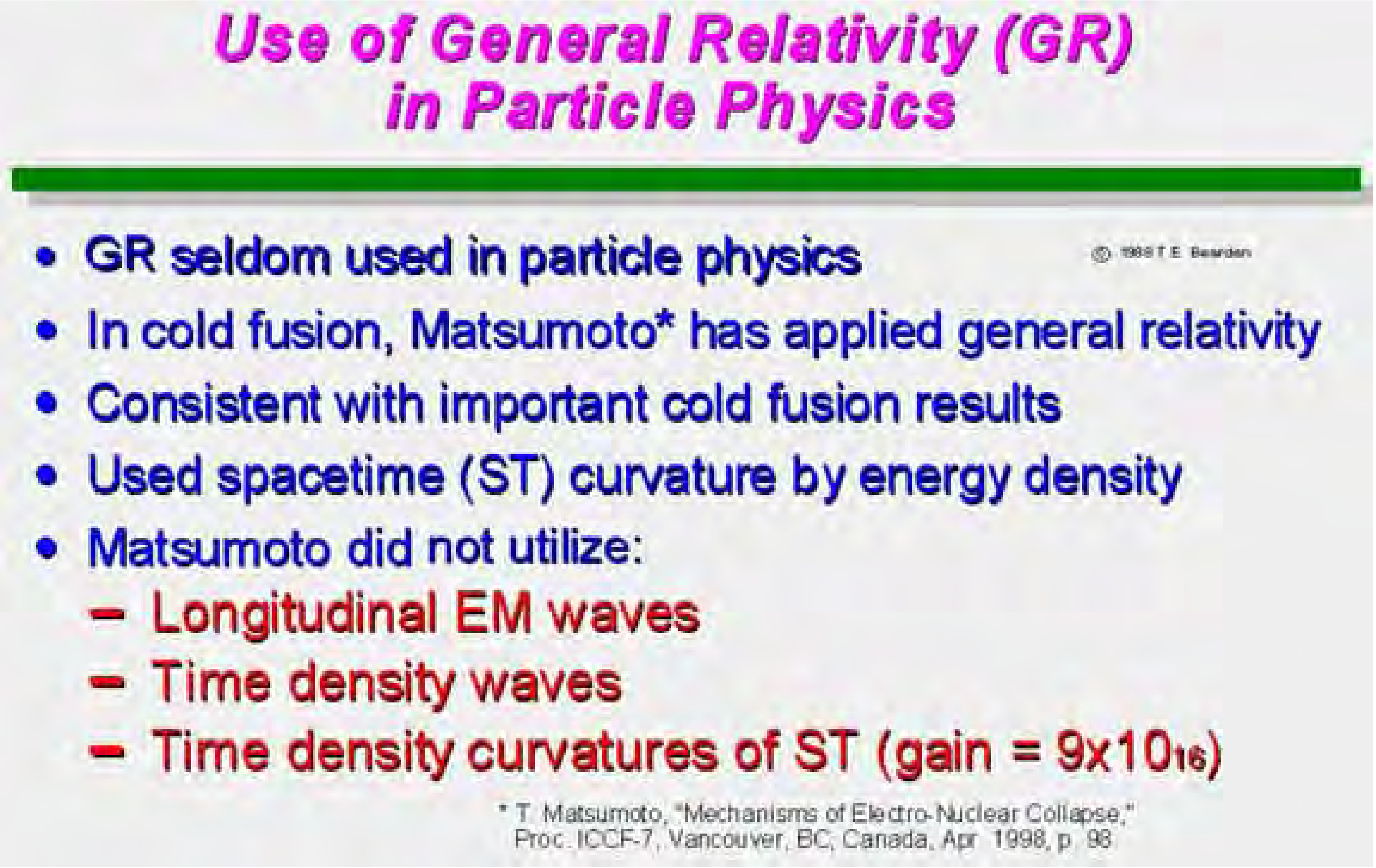

Use of General Relativity (GR) in Particle Physics

© 1998 T.E. Bearden

- GR seldom used in particle physics
- In cold fusion, Matsumoto* has applied general relativity
- Consistent with important cold fusion results
- Used spacetime (ST) curvature by energy density
- Matsumoto did not utilize:
  - Longitudinal EM waves
  - Time density waves
  - Time density curvatures of ST (gain = $9x10_{16}$)

* T. Matsumoto, "Mechanisms of Electro-Nuclear Collapse," Proc. ICCF-7, Vancouver, BC, Canada, Apr. 1998, p. 98.

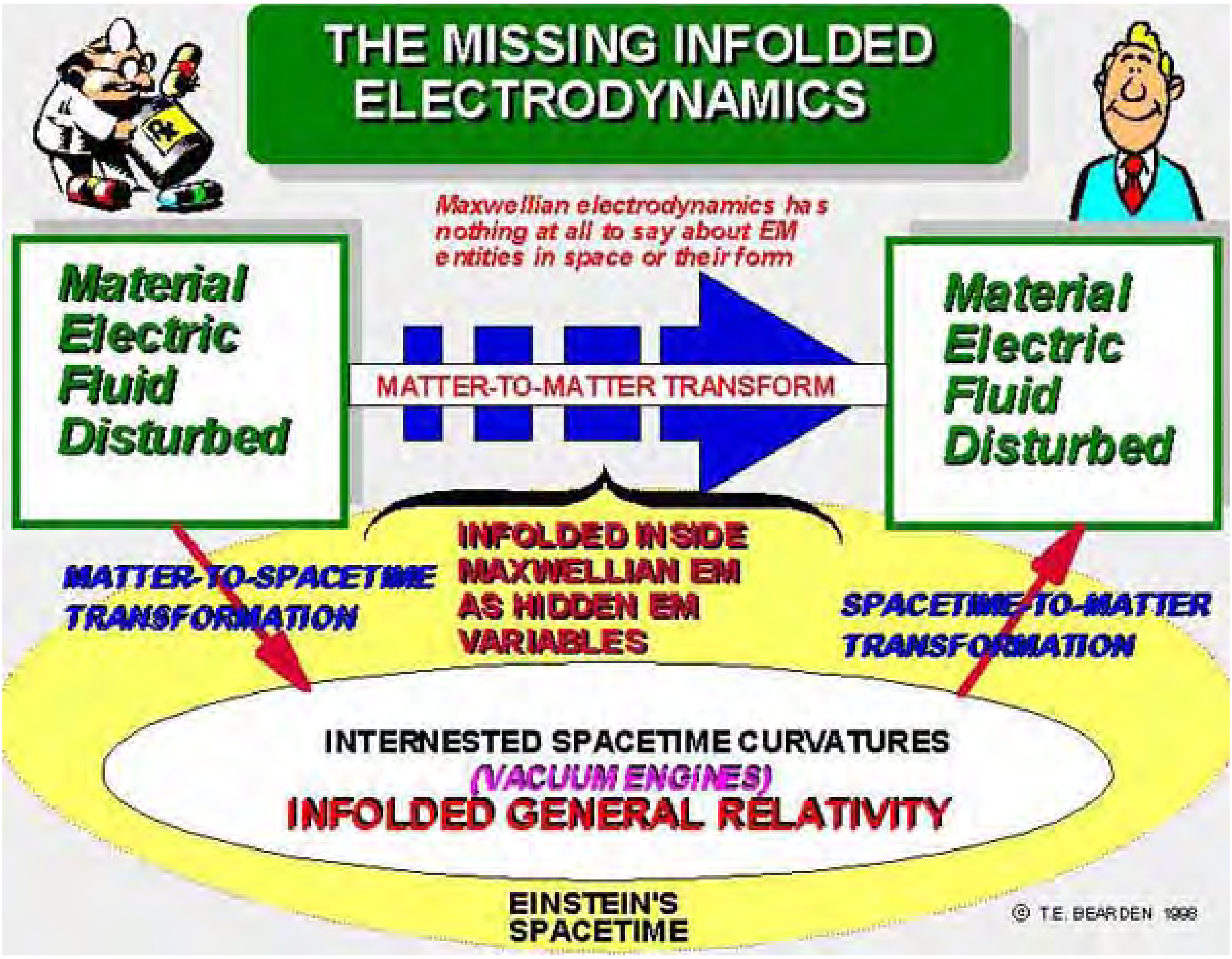

THE MISSING INFOLDED ELECTRODYNAMICS

Material Electric Fluid Disturbed

Maxwellian electrodynamics has nothing at all to say about EM entities in space or their form

MATTER-TO-MATTER TRANSFORM

Material Electric Fluid Disturbed

MATTER-TO-SPACETIME TRANSFORMATION

INFOLDED INSIDE MAXWELLIAN EM AS HIDDEN EM VARIABLES

SPACETIME-TO-MATTER TRANSFORMATION

INTERNESTED SPACETIME CURVATURES

(VACUUM ENGINES)

INFOLDED GENERAL RELATIVITY

EINSTEIN'S SPACETIME

© T.E. BEARDEN 1998

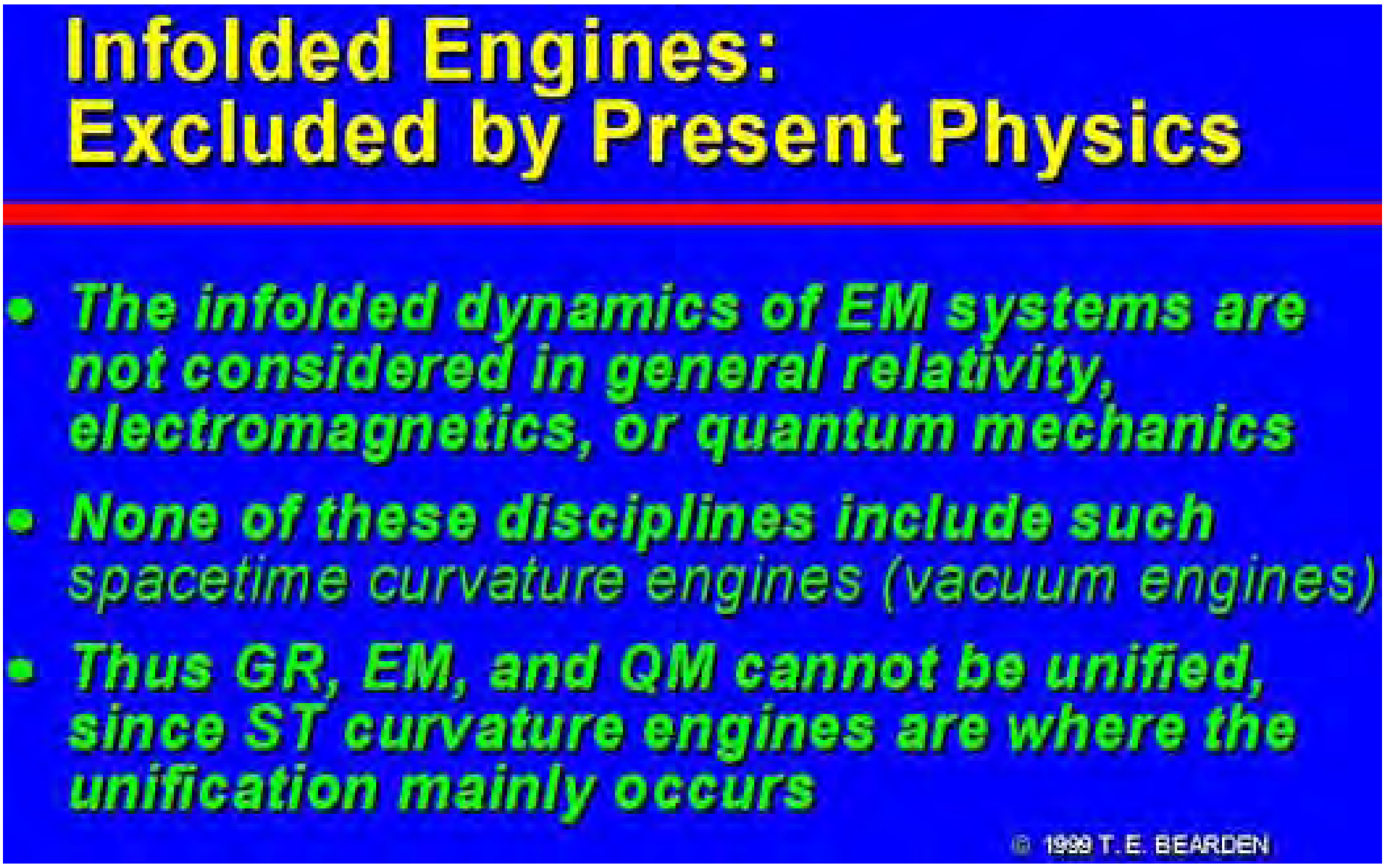

Infolded Engines: Excluded by Present Physics

- The infolded dynamics of EM systems are not considered in general relativity, electromagnetics, or quantum mechanics
- None of these disciplines include such spacetime curvature engines (vacuum engines)
- Thus GR, EM, and QM cannot be unified, since ST curvature engines are where the unification mainly occurs

© 1999 T. E. BEARDEN

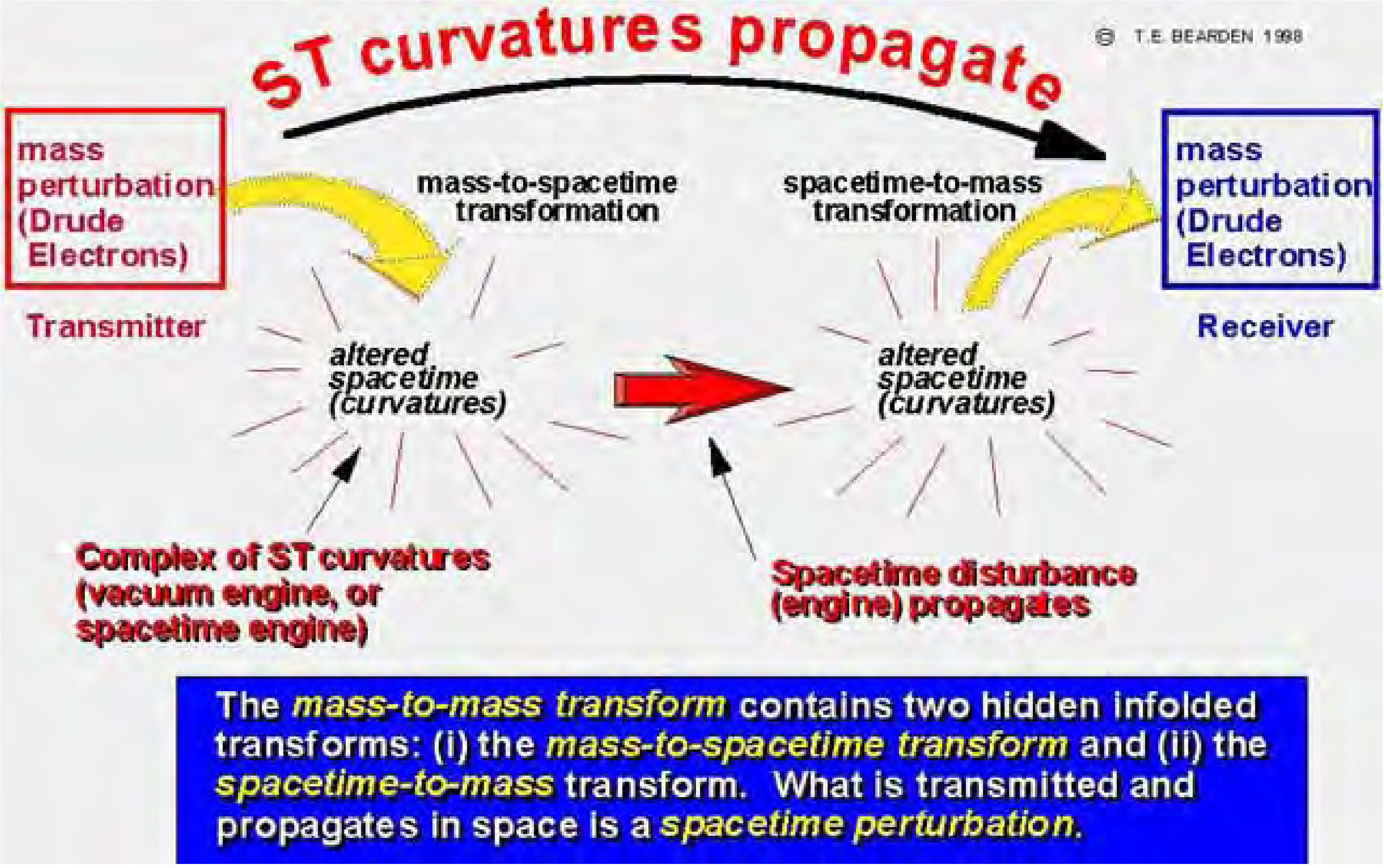

ST curvatures propagate

© T.E. BEARDEN 1998

mass perturbation (Drude Electrons)

Transmitter

mass-to-spacetime transformation

altered spacetime (curvatures)

spacetime-to-mass transformation

altered spacetime (curvatures)

mass perturbation (Drude Electrons)

Receiver

Complex of ST curvatures (vacuum engine, or spacetime engine)

Spacetime disturbance (engine) propagates

The *mass-to-mass transform* contains two hidden infolded transforms: (i) the *mass-to-spacetime transform* and (ii) the *spacetime-to-mass* transform. What is transmitted and propagates in space is a *spacetime perturbation.*

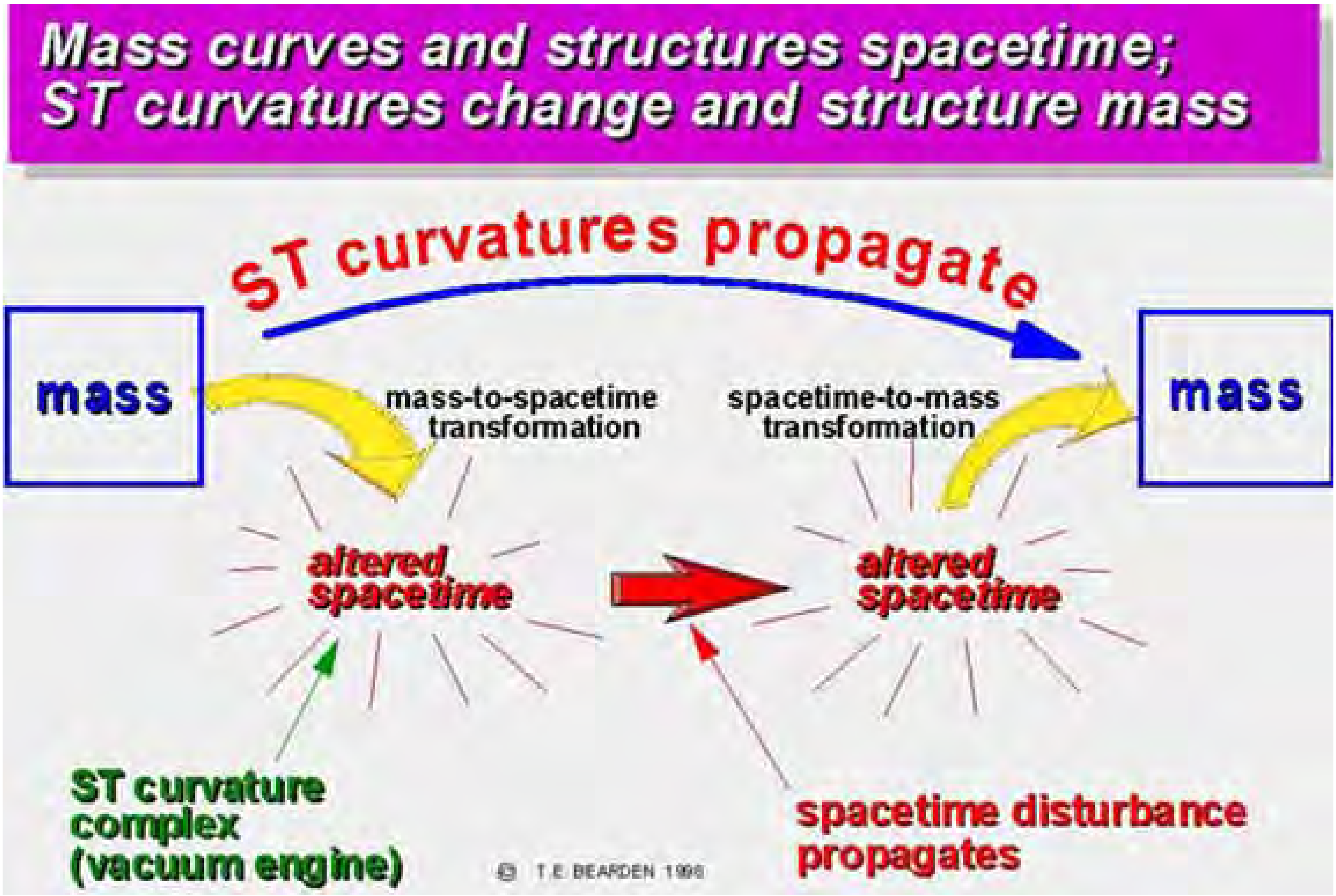

**Mass curves and structures spacetime;
ST curvatures change and structure mass**

ST curvatures propagate

mass

mass-to-spacetime transformation

spacetime-to-mass transformation

mass

altered spacetime

altered spacetime

ST curvature complex (vacuum engine)

© T.E. BEARDEN 1996

spacetime disturbance propagates

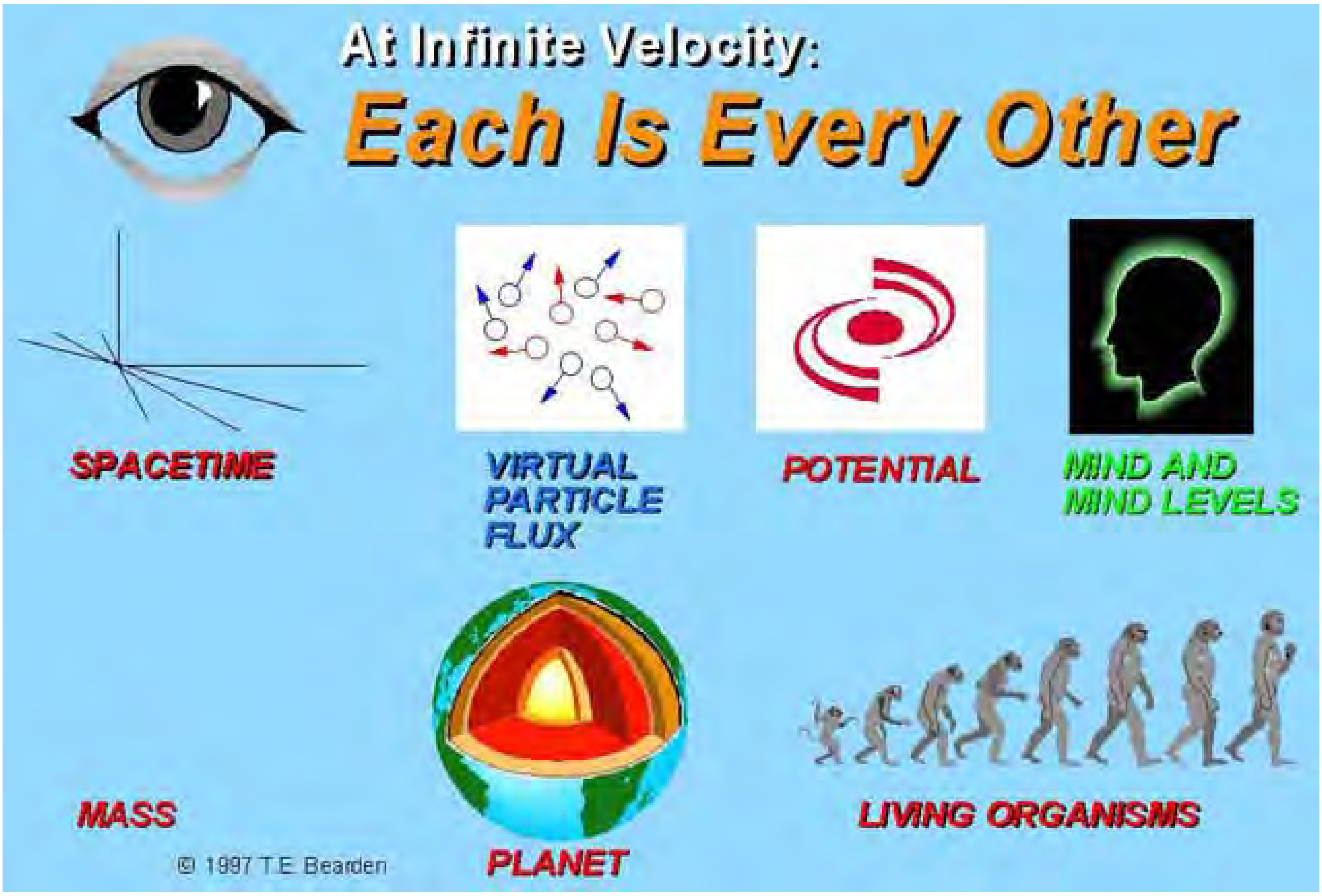

At Infinite Velocity:

Each Is Every Other

SPACETIME

VIRTUAL PARTICLE FLUX

POTENTIAL

MIND AND MIND LEVELS

MASS

PLANET

LIVING ORGANISMS

© 1997 T.E. Bearden

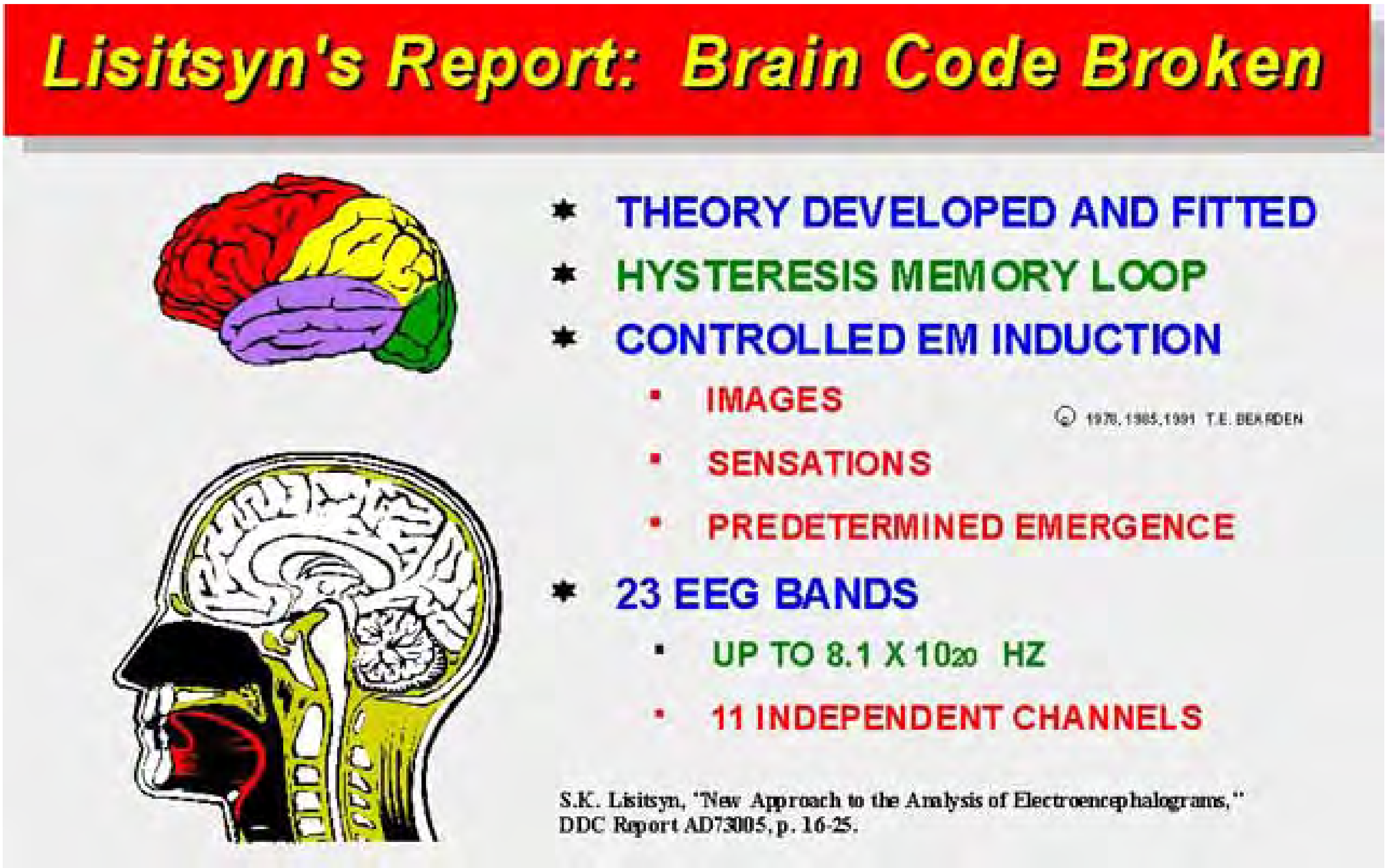

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10_{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

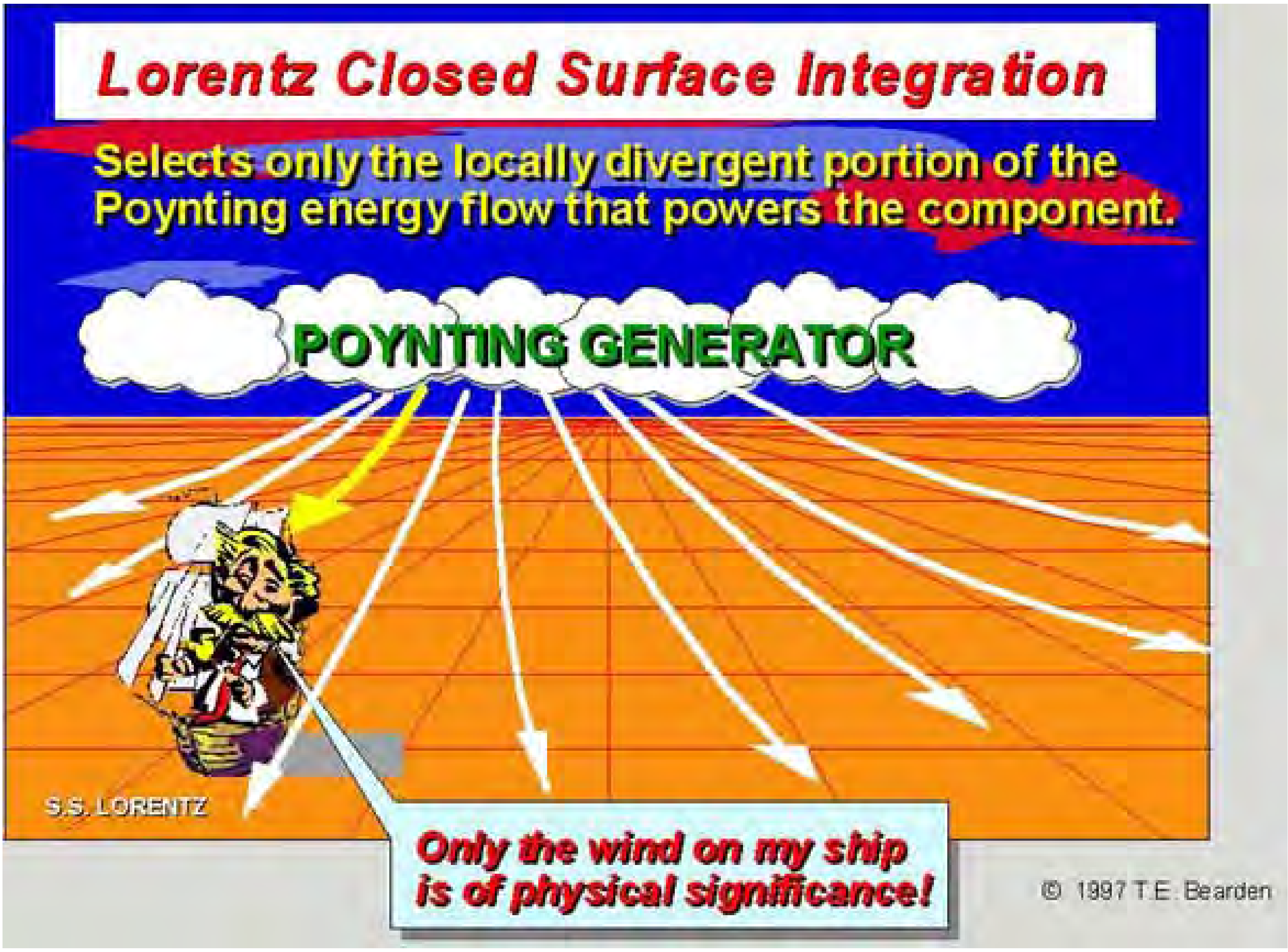

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

POYNTING GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

© 1997 T.E. Bearden

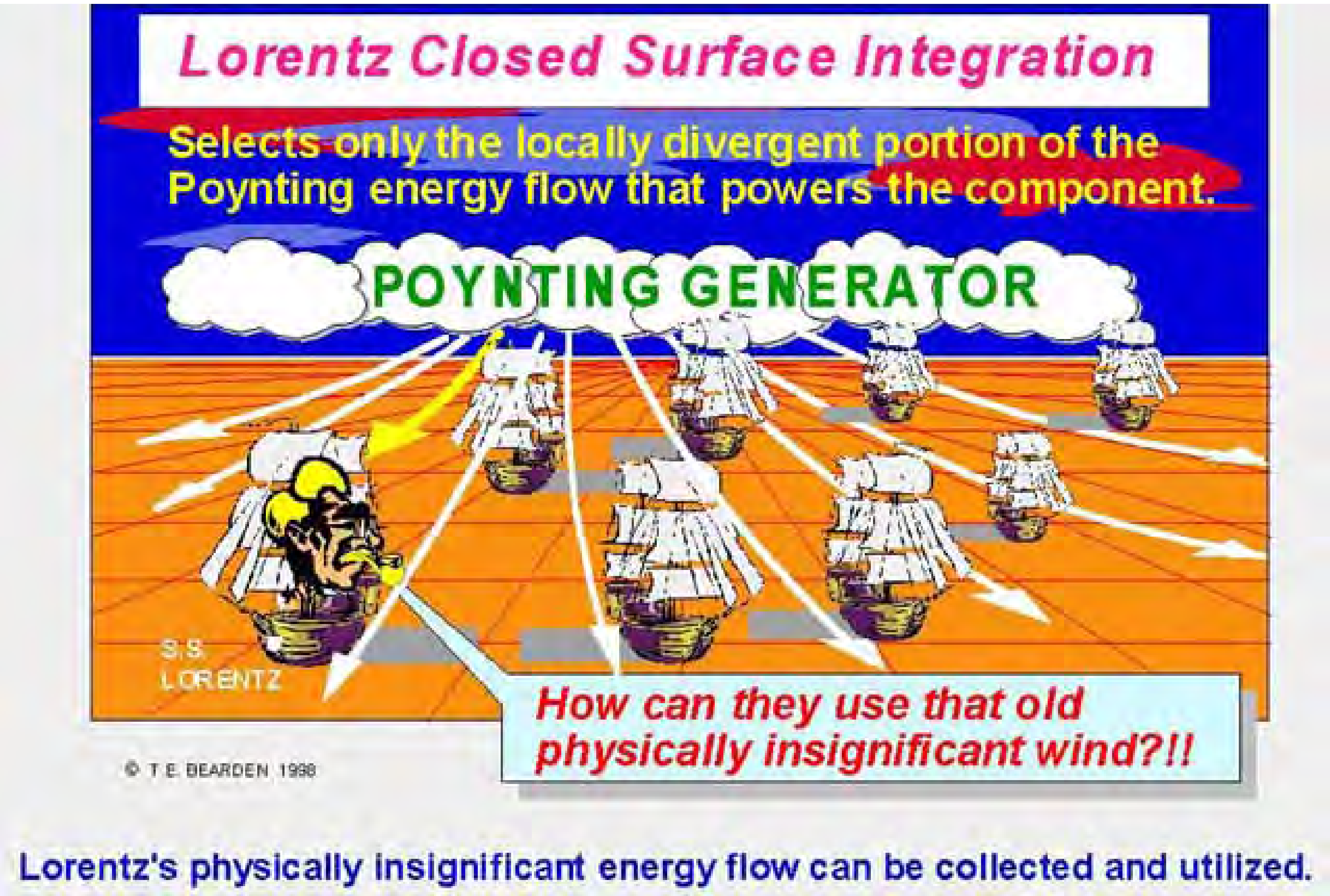

Lorentz Closed Surface Integration

Selects only the locally divergent portion of the Poynting energy flow that powers the component.

How can they use that old physically insignificant wind?!!

© T E BEARDEN 1998

Lorentz's physically insignificant energy flow can be collected and utilized.

Lorentz surface integration of the Poynting vector around a closed surface

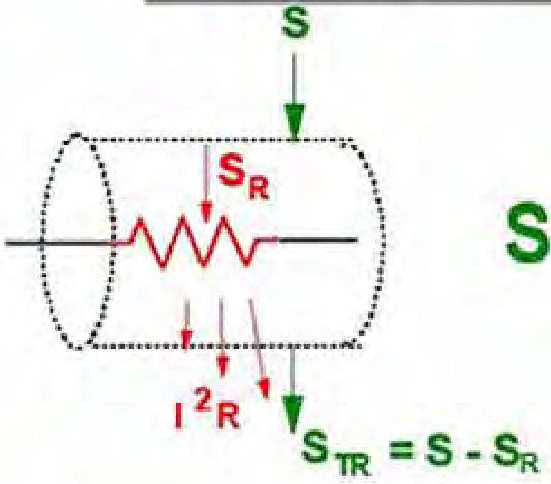

1a. Lorentz surface integration.

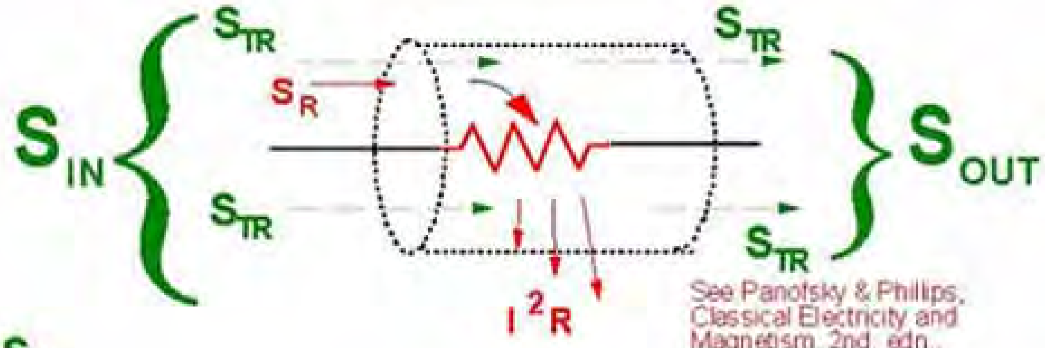

See Panofsky & Phillips, Classical Electricity and Magnetism, 2nd, edn., Addison Wesley, 1962, p. 178-181.

1b. Actual S in and S out.

Note: If the S-vector is integrated over the closed surface, then all energy *transport* passage is zeroed, leaving only the very small component of the input S-flow that is powering the joule heating of the resistor. In short, only the small component of the S-flow that is equal in magnitude to the Slepian vector $j\phi$ remains. This measures only the tiny portion of the S-flow that is "collecting" on electrons passing through the resistor, and therefrom being dissipated out of the resistor as joule heating. It discards everything else (all S_{TR}'s and S_{OUT}). S_{IN} becomes S_R.

© 1997 T.E. BEARDEN

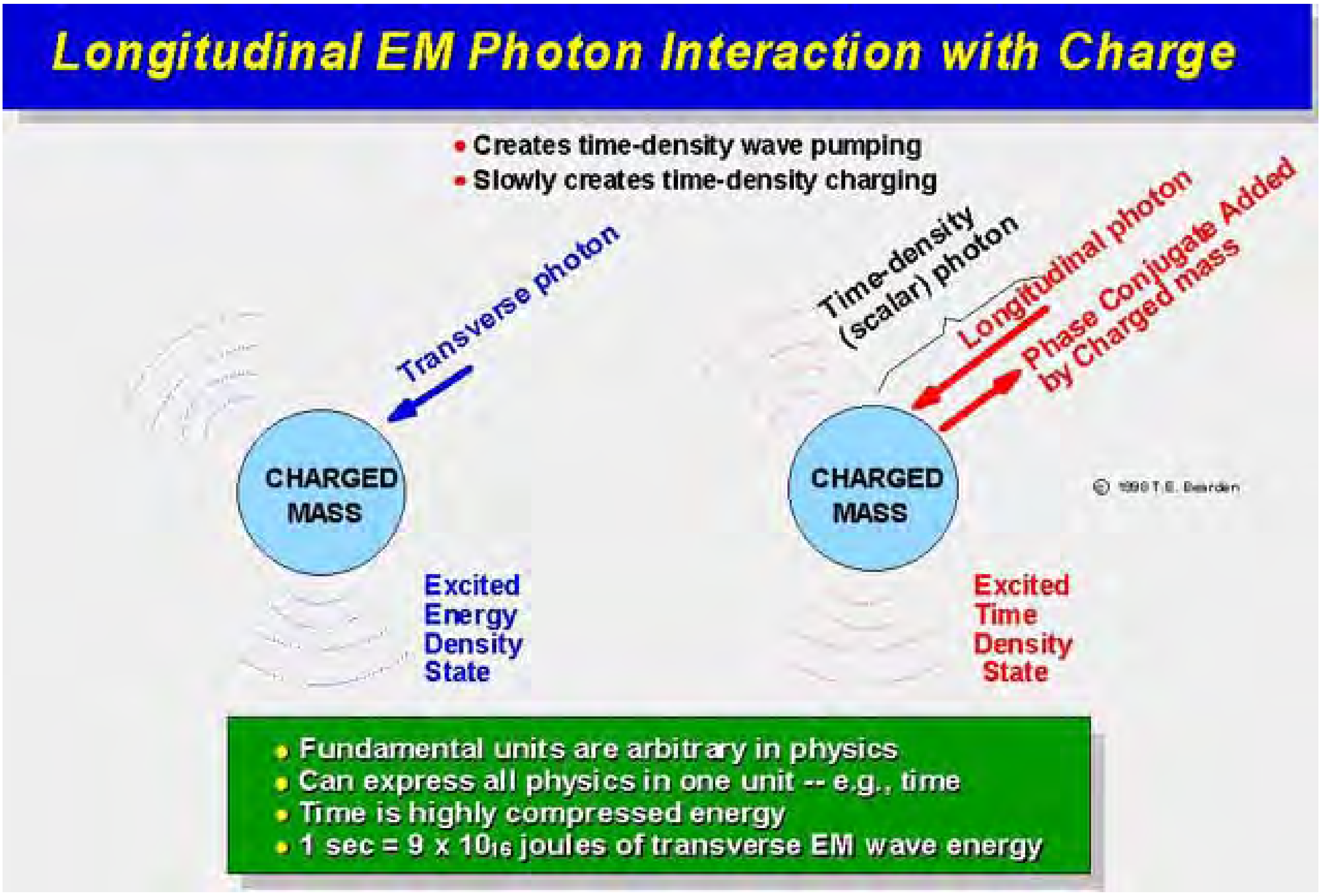

Longitudinal EM Photon Interaction with Charge

- Creates time-density wave pumping
- Slowly creates time-density charging

Transverse photon

CHARGED MASS

Excited Energy Density State

Time-density (scalar) photon

Longitudinal photon

Phase Conjugate Added by Charged mass

CHARGED MASS

Excited Time Density State

© 1990 T.E. Bearden

- Fundamental units are arbitrary in physics
- Can express all physics in one unit -- e.g., time
- Time is highly compressed energy
- 1 sec = 9 x 10_{16} joules of transverse EM wave energy

Figure 1-8. Infolded biwave composition of a scalar potential.

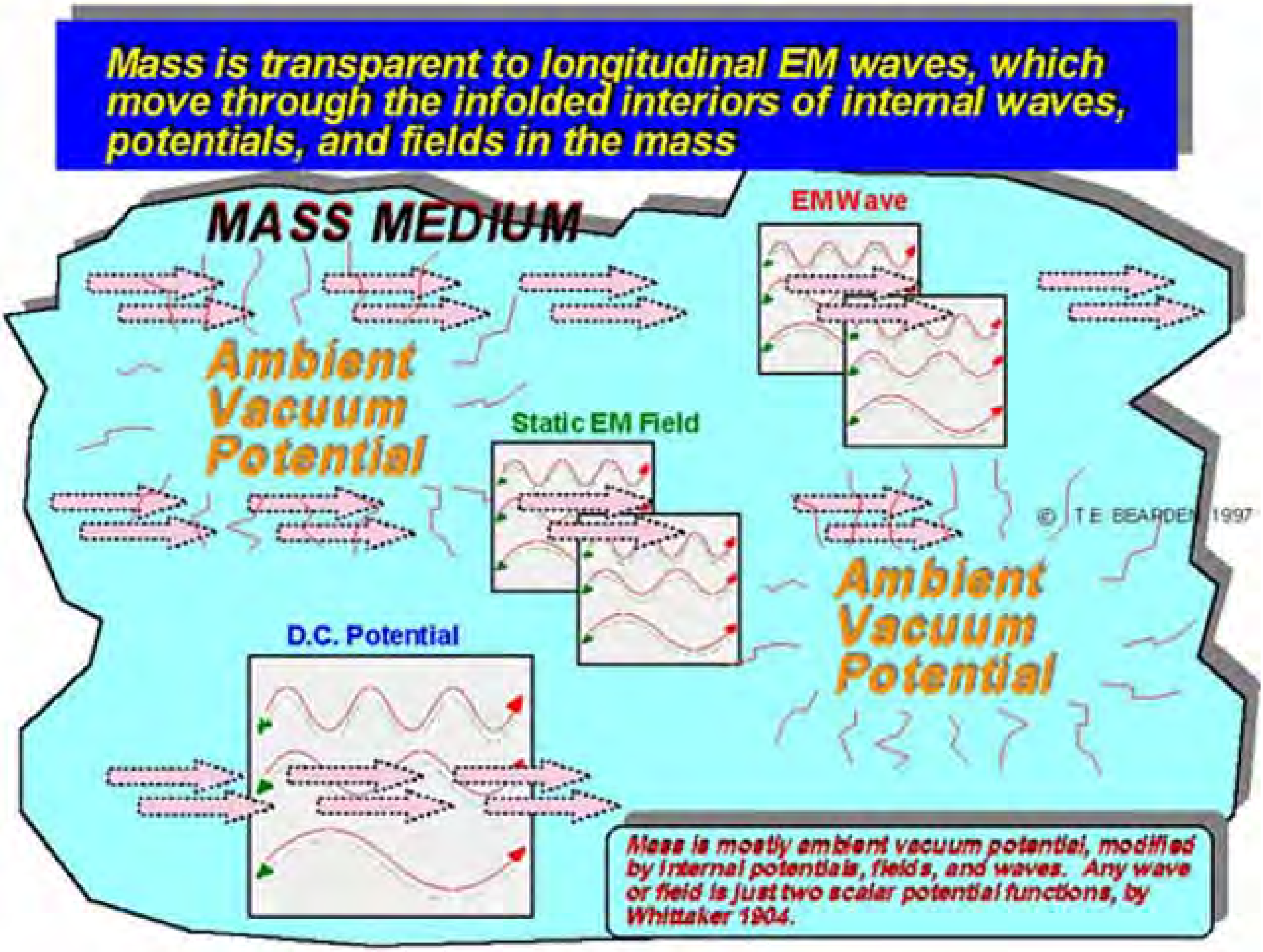

Mass is transparent to longitudinal EM waves, which move through the infolded interiors of internal waves, potentials, and fields in the mass

Mass is mostly ambient vacuum potential, modified by internal potentials, fields, and waves. Any wave or field is just two scalar potential functions, by Whittaker 1904.

MASS MEDIUM
EMWave
Ambient Vacuum Potential
Static EM Field
© T.E. BEARDEN 1997
Ambient Vacuum Potential
D.C. Potential

Pumping with Longitudinal EM Waves Time-reverses the Mass

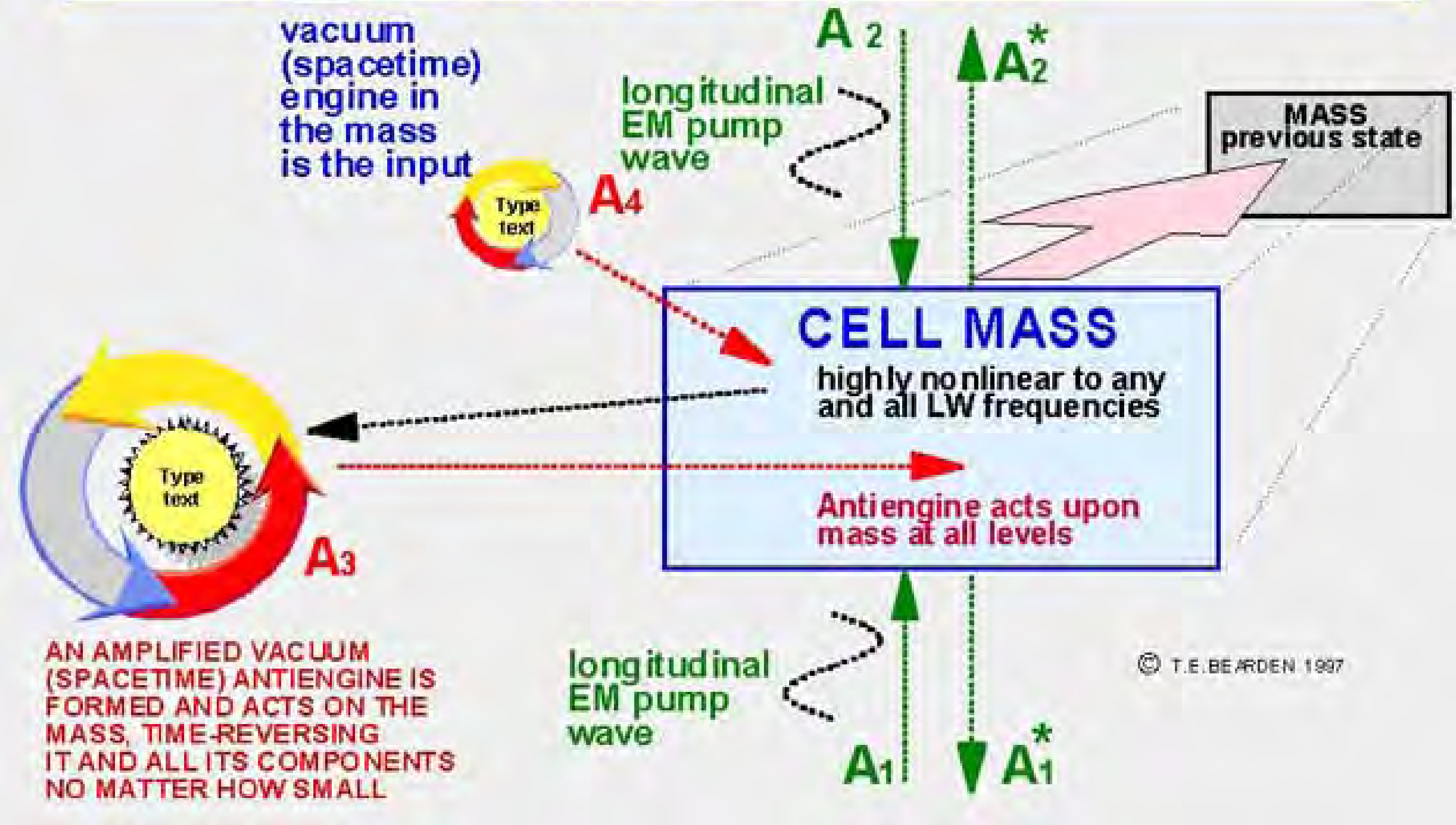

vacuum (spacetime) engine in the mass is the input

longitudinal EM pump wave

CELL MASS

highly nonlinear to any and all LW frequencies

Antiengine acts upon mass at all levels

AN AMPLIFIED VACUUM (SPACETIME) ANTIENGINE IS FORMED AND ACTS ON THE MASS, TIME-REVERSING IT AND ALL ITS COMPONENTS NO MATTER HOW SMALL

longitudinal EM pump wave

© T.E.BEARDEN 1997

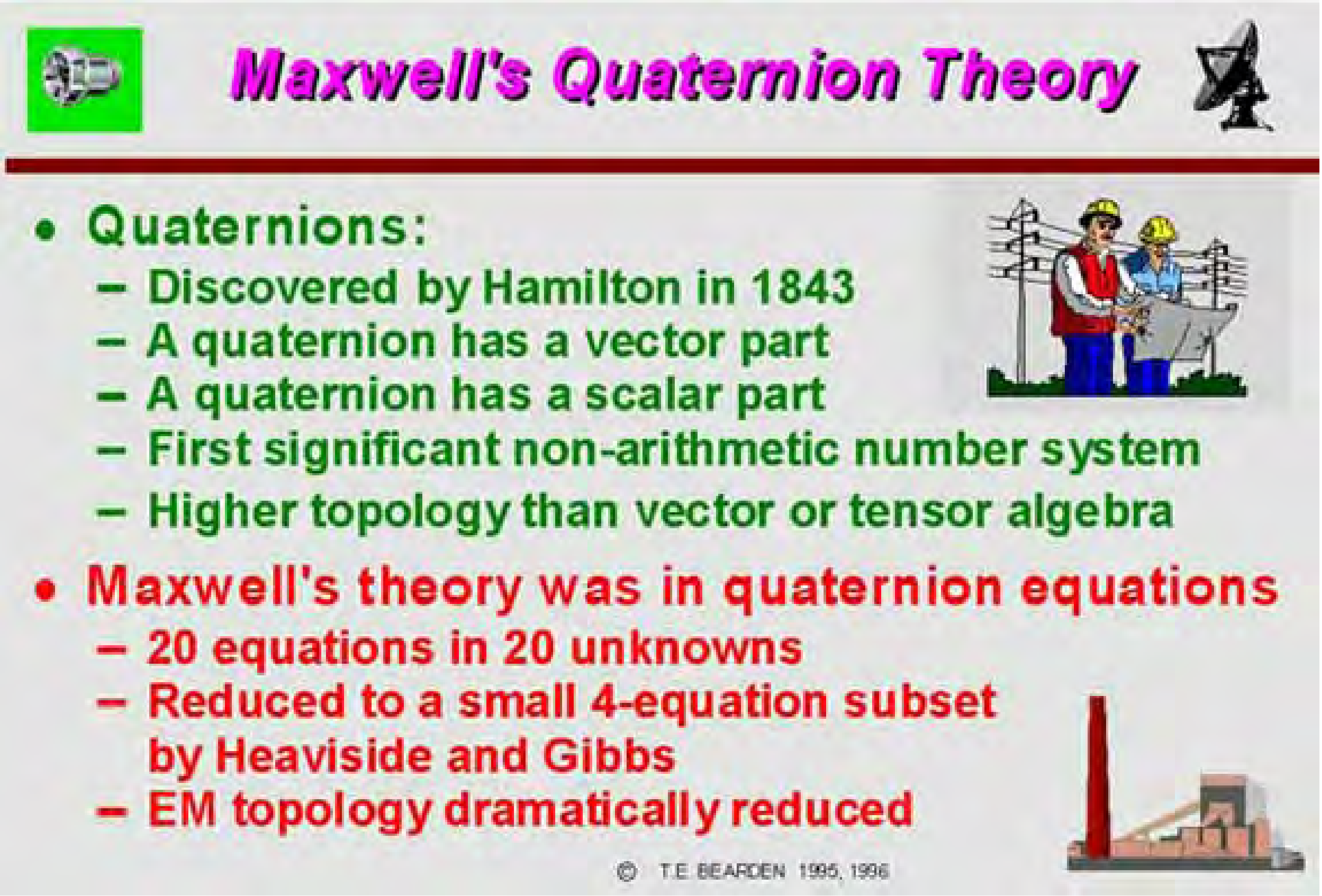

Maxwell's Quaternion Theory

- Quaternions:
  - Discovered by Hamilton in 1843
  - A quaternion has a vector part
  - A quaternion has a scalar part
  - First significant non-arithmetic number system
  - Higher topology than vector or tensor algebra
- Maxwell's theory was in quaternion equations
  - 20 equations in 20 unknowns
  - Reduced to a small 4-equation subset by Heaviside and Gibbs
  - EM topology dramatically reduced

© T.E. BEARDEN 1995, 1996

Maxwellian Systems

Before regauging, after symmetrical regauging, and after asymmetrical regauging

Note: Any system must first break equilibrium in its vacuum exchange, before it can energize Drude electrons and "power" the load. Asymmetric regauging is always utilized. Powering loads and losses is a symmetry-restoring operation.

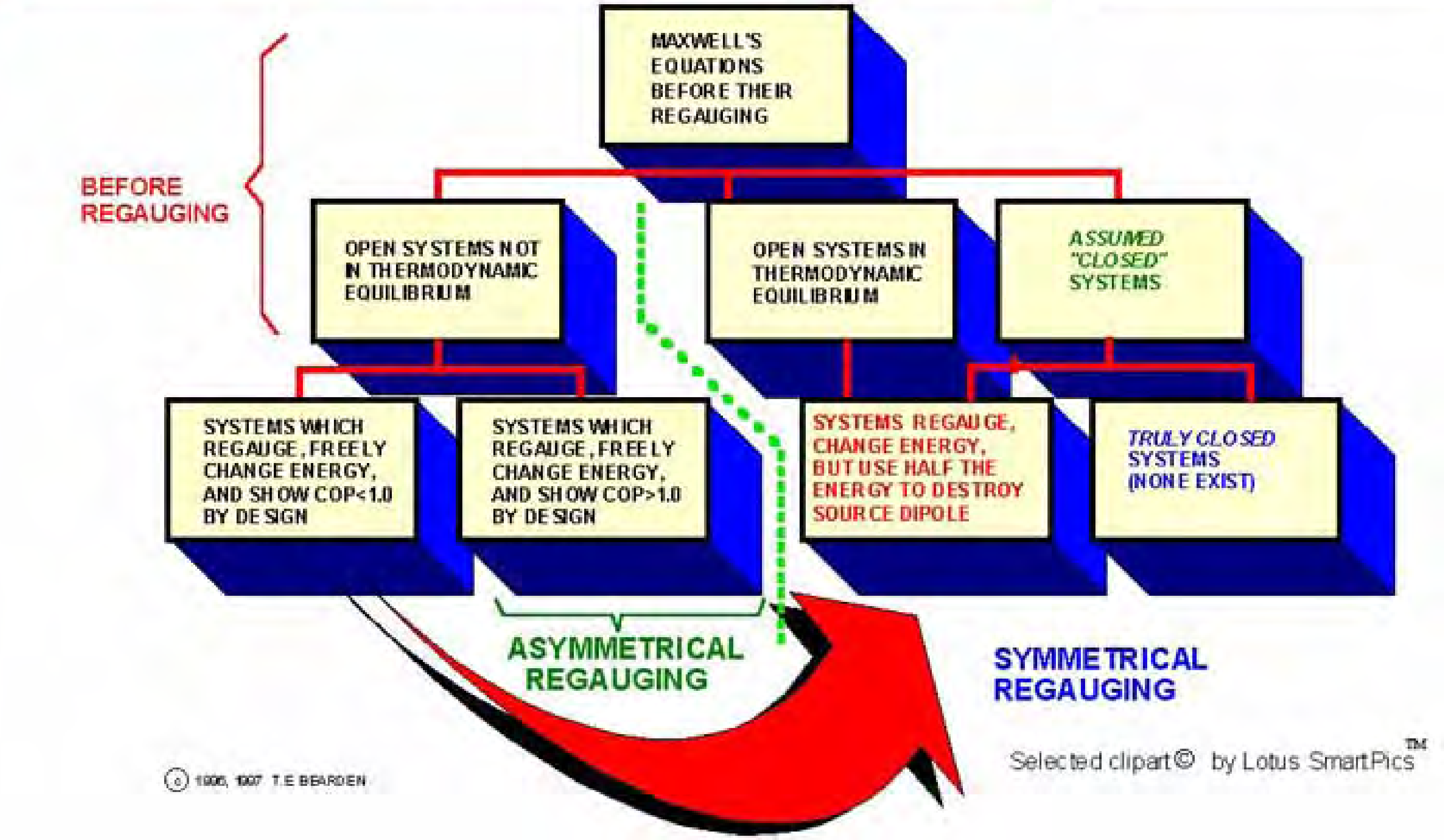

BEFORE REGAUGING

MAXWELL'S EQUATIONS BEFORE THEIR REGAUGING

OPEN SYSTEMS NOT IN THERMODYNAMIC EQUILIBRIUM

OPEN SYSTEMS IN THERMODYNAMIC EQUILIBRIUM

ASSUMED "CLOSED" SYSTEMS

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP<1.0 BY DESIGN

SYSTEMS WHICH REGAUGE, FREELY CHANGE ENERGY, AND SHOW COP>1.0 BY DESIGN

SYSTEMS REGAUGE, CHANGE ENERGY, BUT USE HALF THE ENERGY TO DESTROY SOURCE DIPOLE

TRULY CLOSED SYSTEMS (NONE EXIST)

ASYMMETRICAL REGAUGING

SYMMETRICAL REGAUGING

Selected clipart© by Lotus SmartPics™

© 1996, 1997 T.E BEARDEN

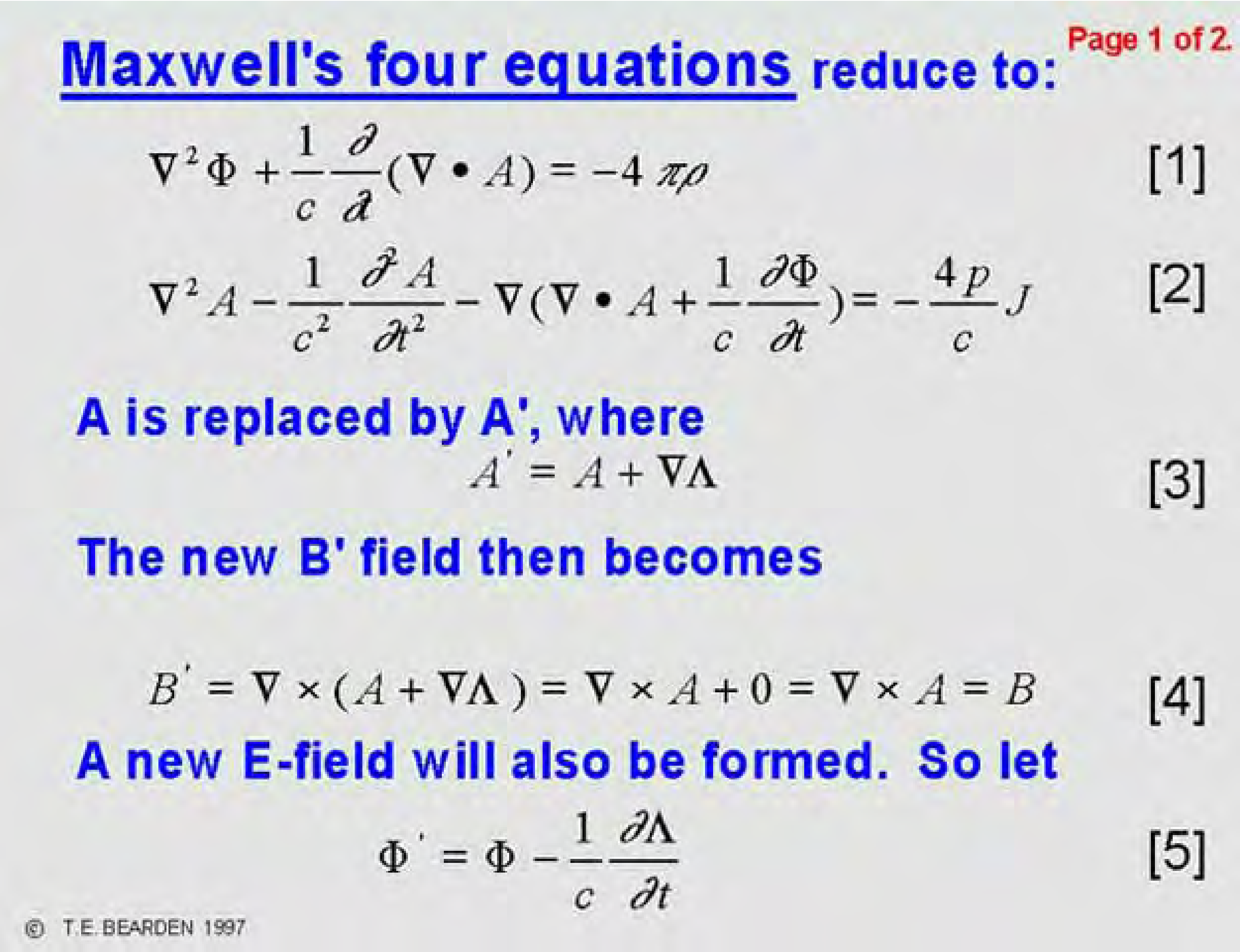

Page 1 of 2.

Maxwell's four equations reduce to:

$$\nabla^2 \Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet A) = -4\pi\rho \qquad [1]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} - \nabla\left(\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t}\right) = -\frac{4p}{c}J \qquad [2]$$

A is replaced by A', where

$$A' = A + \nabla\Lambda \qquad [3]$$

The new B' field then becomes

$$B' = \nabla \times (A + \nabla\Lambda) = \nabla \times A + 0 = \nabla \times A = B \qquad [4]$$

A new E-field will also be formed. So let

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial\Lambda}{\partial t} \qquad [5]$$

© T.E. BEARDEN 1997

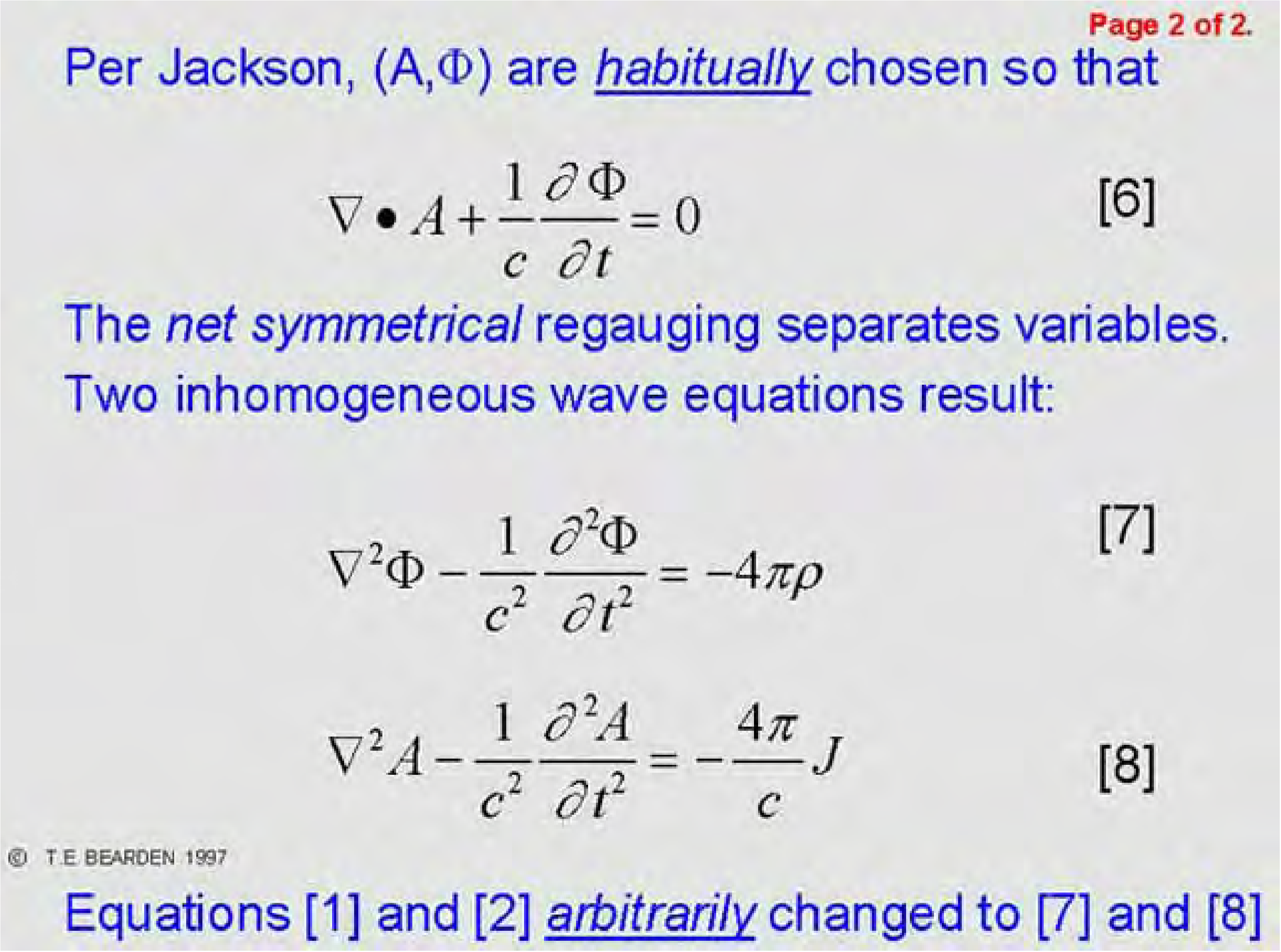

Page 2 of 2.

Per Jackson, (A,Φ) are *habitually* chosen so that

$$\nabla \bullet A + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [6]$$

The *net symmetrical* regauging separates variables. Two inhomogeneous wave equations result:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [7]$$

$$\nabla^2 A - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = -\frac{4\pi}{c} J \qquad [8]$$

© T.E. BEARDEN 1997

Equations [1] and [2] *arbitrarily* changed to [7] and [8]

The Tom Bearden Website

Mind Control and EM Wave Polarization Transductions - Part I, II & III

Reprinted from "Explore" magazine, 1999

Part I

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9
- Page 10

Part II

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8

Part III

- Page 1
- Page 2
- Page 3
- Page 4
- Page 5
- Page 6
- Page 7
- Page 8
- Page 9

- The U.S. approach to mind control

- "Voices in my head"
  Flanagan's neurophone - used to give instructions to diplomats during negotiations, etc.

The Tom Bearden Website

ORDER BOOKS & DVDs

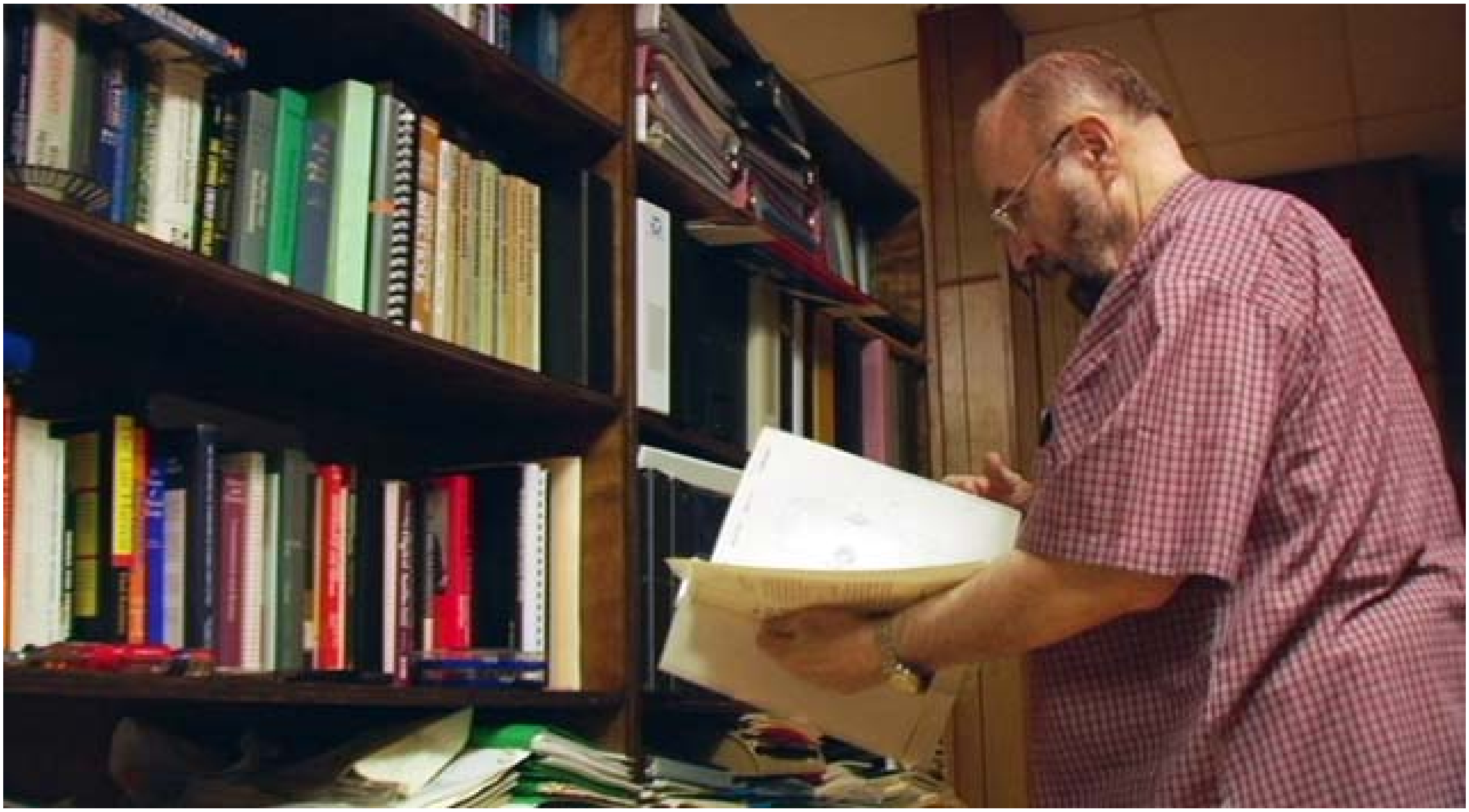

Articles

- **Job Opportunities in Energy**
  25 August, 2010

- **THE WORLD ENERGY CRISIS IN PERSPECTIVE:**
  The Breakthrough Technology that Solves it Quickly
  22 March, 2010

- Direct Engineering of Physical Reality Itself
  3 June, 2009
- Solving The Problem of Freely Using EM Energy from the Vacuum
  18 September, 2008
- How an EM circuit powers its load - and solving the world energy crisis
  16 September, 2008
- Rapid and Decisive Solution of the World Energy Crisis and Global Warming
  Updated
  9 September, 2008
- On Negative Energy EMP Weapons
  26 July, 2008
- In Memoriam - Howard Johnson
  31 March, 2008
- Suppression of the Self-Powering Battery?
  1st December, 2007
- Never Mind - Junk Science

and All That
11 September, 2007

- **The Deliberate Discard of Asymmetric Maxwellian Systems, Thus Preventing COP>1.0**
  and Self-Powering Energy-from-the-Vacuum Systems
  21 June, 2007

- **What We Have to Get Across to Our Electrical Engineering Profession:**
  To the EE Departments, Professors, Students, Curricula, and Textbooks
  May 24, 2007

- **Phases of Research and Development and Some Problems**
  18 February, 2007

- **Scalar Electromagnetic Weapons and their Terrorist Use: Immediate Strategic Aspects of the Asymmetric War on the U.S.**
  11 September, 2004

- **Reducing our Dependence on Fossil Fuels**
  28 April, 2003

- **Flash News: Second Law of Thermodynamics Curtailed**
  4th August, 2002

- **Maxwell's demons are alive and thriving**
  4 July, 2002

- **Clean Electrical Energy from the Active Vacuum**
  January, 2002

Mind Control and EM Wave Polarization Transductions

Part I
Vol. 9, No. 2 1999

Part II
Vol 9, No. 3 1999

Part III
Vol 9, No. 4,5 1999

Vacuum Engines and Priore's Methodology - The True Science of Energy-Medicine

Part I
Vol. 6 No. 1 1995

Part II
Vol. 6 No. 2 1995

Paranormal Pictures
Psychic Observer 39(3) ~ 224, 255-258

Many of Tom Bearden's older articles on a wide range of topics can be found in the back issues of Explore! magazine.

See http://www.explorepub.com

Mind Control and EM Wave Polarization Transductions, Part 1

©Copyright 1999 by T.E. Bearden, USA

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental pro-cedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Abstract

For some time we have been repeatedly queried about the technical mechanisms and unusual electrodynamics of advanced mind control research, both in the West and abroad. Calling full attention to the special note above, in this paper we present a high-level over-view of the novel electromagnetic nature of mind operations, mind and body coupling, and *intent*•the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented. We give two specific examples of lethal foreign military tests in 1997 of advanced mind control weapons against two military pilots over the central U.S. Some dangers and potential benefits of the emerging mind control technology are pointed out

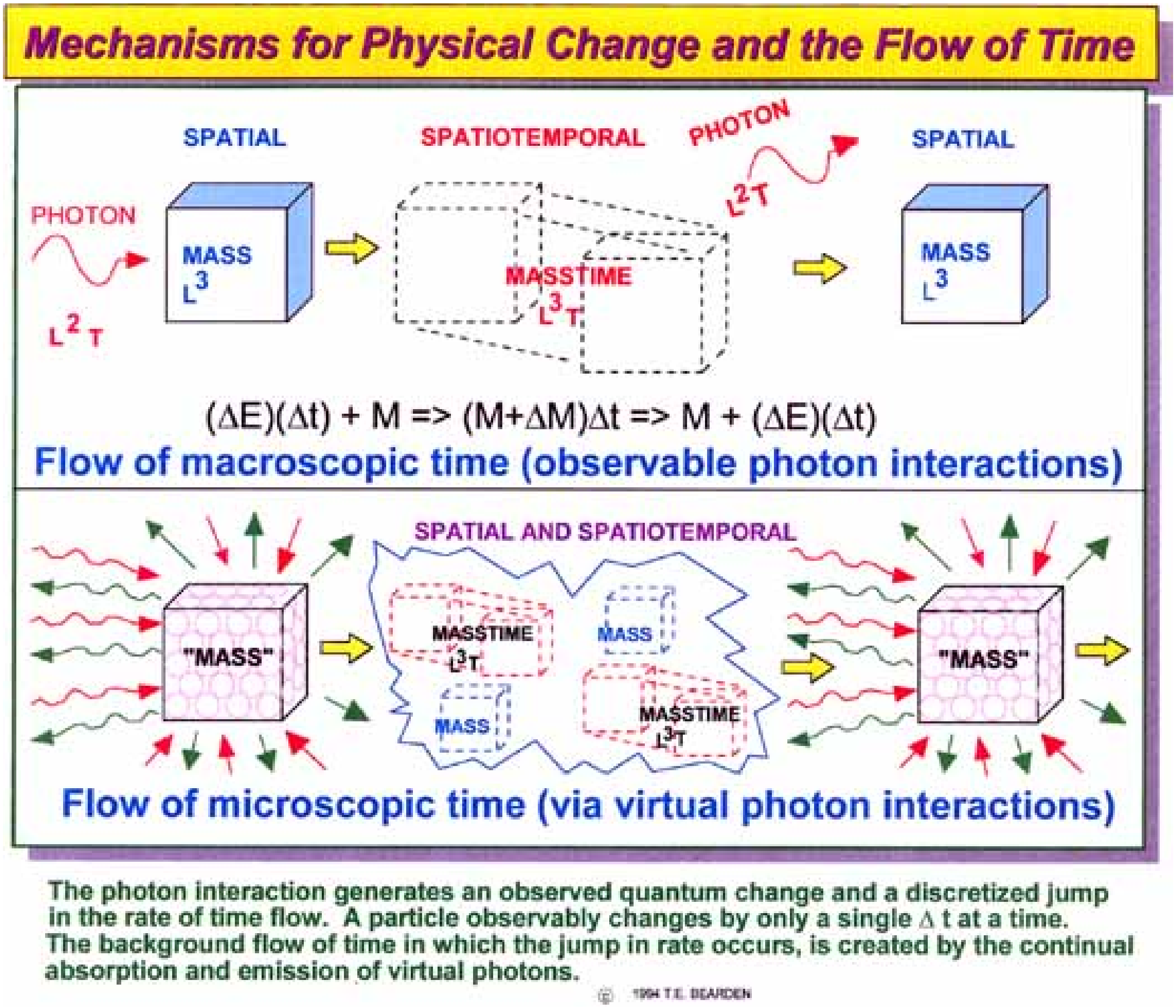
Mechanisms for Physical Change and the Flow of Time
SPATIAL
SPATIOTEMPORAL
PHOTON
SPATIAL
PHOTON
MASS
L3
L2T
MASSTIME
L3T
L2T
MASS
L3
(ΔE)(Δt) + M => (M+ΔM)Δt => M + (ΔE)(Δt)
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME
L3T
MASS
MASS
MASSTIME
L3T
"MASS"
Flow of microscopic time (via virtual photon interactions)
The photon interaction generates an observed quantum change and a discretized jump in the rate of time flow. A particle observably changes by only a single Δ t at a time. The background flow of time in which the jump in rate occurs, is created by the continual absorption and emission of virtual photons.
© 1994 T.E. BEARDEN

Figure 1. Mechanism that generates a mass's flow through time.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Introduction and Background

In quantum field theory, there are four polarizations of photons.[1] Using 4-space and the z-direction as the direction of propagation, we have x- and y- polarizations where the 3-spatial energy of the photon is oscillating laterally, in the x- or y- direction. These are *transverse polarized* photons, as is any combination of the two. The third polarization is along the z-direction, which is a longitudinal polarization. In other words, the 3-spatial energy of the longitudinal photon cannot oscillate in the x- or y- direction, and so it is oscillating to-and-fro along the line of motion, z-. The fourth polarization occurs when the energy is frozen in all three spatial dimensions x-, y-, and z- and it cannot oscillate in any of those directions. In that case, the photon oscillates its energy in the t- direction, providing the t-polarized photon. We will later discuss how time is actually highly condensed energy.

In Minkowski 4-space, body operations are space-like, and are so treated in conventional materialistic physics. Mind and mind-operations are time-like, not space-like, even though they are totally electromagnetic in nature.[2]

As is well-known, all observation in physics is considered 3-spatial..[3] Mass is a 3-spatial concept, and we detect changes to mass (as, e.g., in the shift of electrons in the circuits of electrical instruments). So one may model the common physical observation mechanism as a *time-differentiating process* which *a priori* discards time-change and retains 3-spatial energy change.[4] That is, Minkowski reality is modeled in the fundamental units of L^3t. Physical observation (via the transverse photon interaction) is the process given by applying the operator $\partial/\partial t$ to L^3t, yielding an L^3 output. Hence mind and mind operations are excluded by the usual physics instruments and observation, which simply exclude the time domain in their outputs and do not "measure" it.

For this reason, physicists have erroneously considered mind to be "metaphysical" and nonreal. Indeed, most physicists to one extent or the other are materialists, and consider the "mind" to be nothing but the operations ongoing in a "meat computer." Nonetheless, in the time-domain the time is absolutely real, and it is completely electrodynamic in nature. There is no metaphysics involved, and the temporal domain—along with mind and mind operations—is simply an erroneously neglected area of physics. In present physics, the notion of mind is comfortably disposed of by imposing the use of the "observer" concept, with out ever specifying that the observer has a consciousness and a mind. Indeed, "observation" is only about what that "observer" perceives. Obviously, one has a dramatically crippled physics if one eliminates that nonobservable called "time." Similarly, one also has a dramatically crippled physics when one eliminates the mechanisms and physics ongoing in those time-like and dynamic "things" such as mind, that occupy time and function in it.

From this viewpoint, Western physics adheres to its 3-spatial measurement foundation only by ignoring the transduction of time-polarized EM changes into detectable longitudinal and transverse EM wave changes. In this aspect, present Western physics is severely self-crippled.

Time is totally electromagnetic and energetic in nature. The flow of time is not a separate external river on which a mass floats along like a boat drifting down the current of a great flowing river. Instead, the flow of time is generated directly on every mass by its total set of photon interactions, both virtual and ob servable. We have previously presented the exact mechanism for the flow of time. *[Figure 1]*

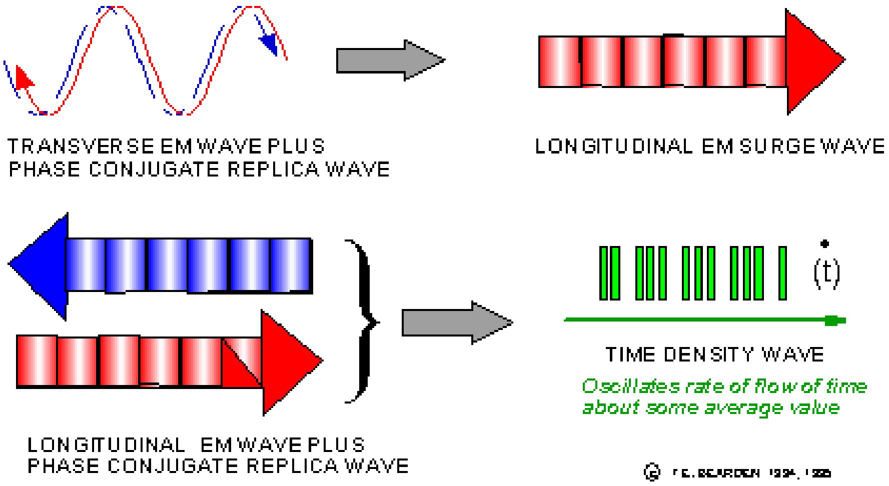
TRANSVERSE EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
LONGITUDINAL EM SURGE WAVE
LONGITUDINAL EM WAVE PLUS
PHASE CONJUGATE REPLICA WAVE
$\dot{(t)}$
TIME DENSITY WAVE
Oscillates rate of flow of time
about some average value
© T.E. BEARDEN 1994,

Figure 2. Transduction of EM wave type by successive phase conjugate pairing.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Further, a photon is comprised of angular momentum, therefore of (energy)x(time). It transports both energy and time, not just energy alone. When absorbed by a mass, not only does it "energy-excitation" charge the mass, but also it "time-excitation" charges it, converting the former "mass" to "mass time". Rigorously it is *masstime* that emits a photon, not *mass*. So a mass moves through time in little spurts, by the continual macroscopic addition and subtraction of little Dt components. Further, in being driven through time, mass is continually altered to masstime (a quite different critter from mass), to mass, to masstime, to mass, etc.

In any masstime state, a myriad of tiny virtual photon interactions made of very tiny (DE)(Dt)'s interact with that same mass during that particular larger Dt of the masstime state (m+Dm)(Dt). Hence mass in its alternate masstime state has that state internally structured by its surrounding environment's interactions with it. The masstime state is internally structured energetically in its (DE) component, and also internally structured temporally in its (Dt) component.

A standard charged fundamental particle such as an electron, e.g., is not necessarily identical with another, when the internal structuring of its masstime state is considered. Further, by two papers by Whittaker,[5] interferometry of masstime states with either energy reactions or temporal reactions can yield observable effects and changes due to these neglected "hidden variables" in the electron's masstime alternative states. To engineer the mind and its operations directly, one must perform electrodynamic engineering in the *time* domain, not in the 3-space EM energy density domain. The direct engineering of time-like mind and mind operations—in all levels and all aspects—requires the use of *time-polarized photons and time-polarized EM waves*. This is the rarest form of electrodynamics, almost untouched by Western physicists.

One can either painfully produce such time-polarized photons and EM waves and directly irradiate a target mass with specific assemblies of them, or one can force the mass itself to iteratively transduce ordinary transverse EM waves first into longitudinally-polarized EM waves and then into time-polarized EM waves. Both will be discussed briefly.

In the West, it appears that the present author's discovery of mechanisms for producing time-polarized (scalar) EM waves [see *Figure 2*] and for transducing between wave polarizations has no precedent. [See *Figure 3*]. While scalar (time-polarized) photons are known in the literature,[1] the creation and use of time-polarized EM waves does not seem to appear in the Western physics literature.[6]

Wave transduction or *polarization transductions* are terms used by the author for the process of transforming an EM wave's particular polarization into another polarization type. Such transduction to other than transverse polarization forms is little known in the West. It appears essentially haphazardly in some experiments, usually without any recognition by the experimenters themselves. Significant transduction in experiments will also produce "strange" and unexplained anomalies in the instruments being used for experimental measurements.[7]

So most probably Western mind control researchers have not recognized the methodology and mechanisms[8] for transducing one type of EM wave polarization directly into another.[9] They continue to seek the "mind" in 3-space and hence in the brain, rather than in the time domain. On the other hand, it appears

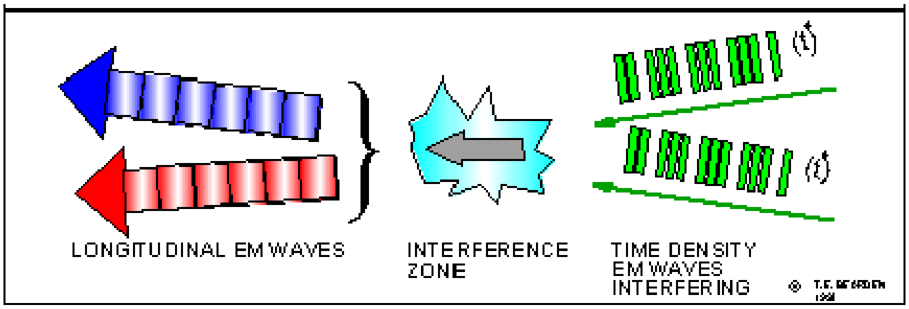

Figure 3. Interference processes for transducing higher EM wave polarizations into lower polarization states.

EXPLORE! VOLUME 9, NUMBER 2, 1999

that the Russian KGB *energetics*[10] weapons scientists—particularly those in *psychoenergetics*[11]—have known and used methods of transducing one wave polarization into another, for at least two decades. It follows that those same scientists have very probably developed *mind engineering* and *mind control* via novel time-polarized EM wave means and a dramatically extended electrodynamics of the time-like mind operations.

In the present paper we briefly develop the wave transductions and the basis for mind engineering, including the specialized use of ordinary transverse wave (TW) EM spectra to force internal EM wave transductions and time-domain operations inside irradiated bodies, cells, and tissues. Through the mind-body coupling mechanism, these transductions of transverse EM waves can operate upon the mind and its deepest operations as well as upon the body, every cell, and every part of every cell.

Russian Mind Control Uses Higher EM Wave Polarizations

It is apparent that the KGB psychoenergetics weapons scientists know and use the full extended EM wave polarization range. As a postulation, one would expect our own weapons scientists to know and utilize the orthodox transverse wave (TW) EM for similar research and experimentation. However, in the West scientists are *just beginning* to realize the importance of a very general (and weaker) type of longitudinal EM wave (LW) polarization, as evidenced by the appearance of dozens of papers on "undistorted progressive waves" (UPWs, which are essentially imperfect LW s with TW residues remaining). The Los Alamos National Laboratory web site has a number of such papers—particularly by Rodrigues[12] and Lu—available for free downloading. UPWs have very interesting characteristics. If they were perfect longitudinal EM waves, they would have infinite energy and infinite speed. Since only imperfect UPWs can be physically made, their speed can vary from slower than the speed of light to faster than the speed of light. Their energy can also vary over a great range.

Another characteristic of UPWs that are reasonably good longitudinal EM waves, is that they pass readily through a large depth of water and mass, including through the ocean and the earth with only small interactions and losses. Yet by interfering two such beams of "high quality" UPWs at a great distance, then in the interference zone ordinary EM energy will rise directly out of local spacetime potential, as essentially shown by Whittaker[13] nearly a century ago. Russian weapon research facilities have weaponized these effects[14] for nearly 50 years, under rigid KGB control and operation. Nonetheless, even if using only ordinary TW waves, Western mind control researchers may get some fairly good results, brute-force-like, by using gross correlates between just the input irradiating TWs and the exhibited behavioral responses of the individual. The net input-output correlations can be determined, even though not taking into account the actual wave polarization transduction mechanisms ongoing inside the irradiated dielectric (or brain, or mind, etc.). However, Western researchers appear to have no knowledge of the exact mechanisms by means of which coherent time-like mind operations of a biological organism couple to the organism's 3-spatial body to provide the mental control loop. They also appear unaware of how the coherent 3-spatial behavior responses of the body couple back to the time-like mind to provide it with a sensory feedback of the body's responses.

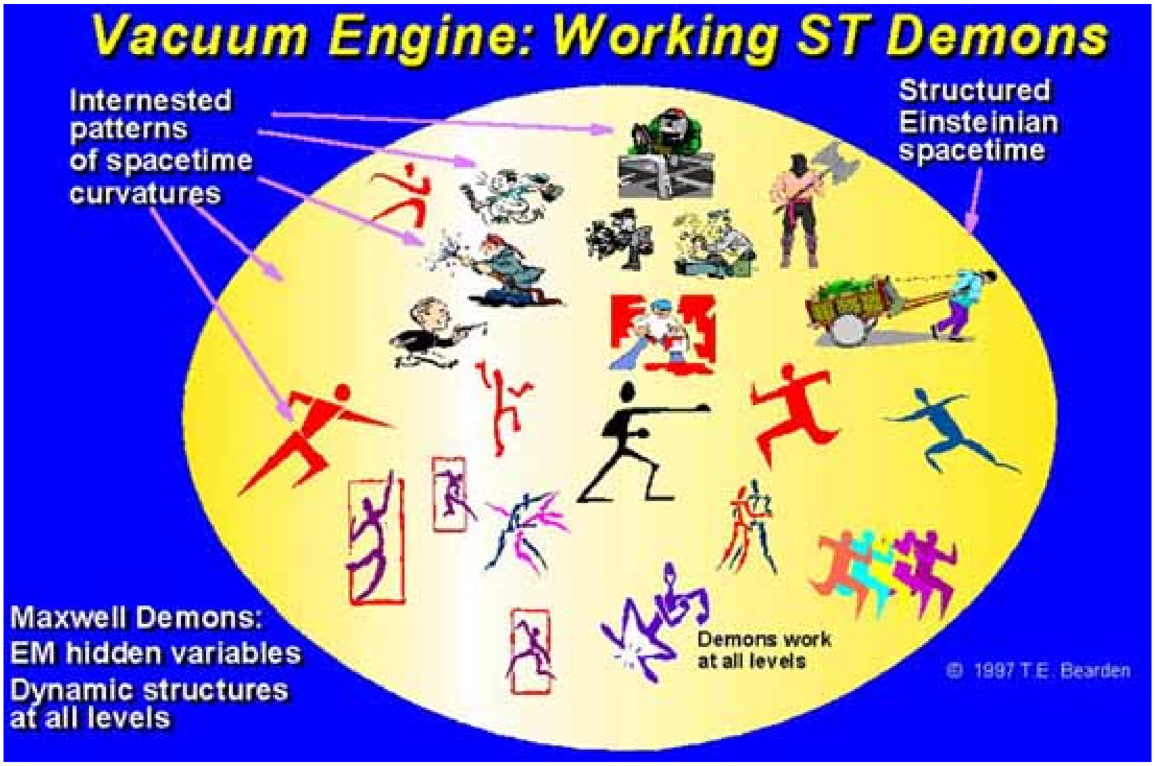

Vacuum Engine: Working ST Demons

Internested
patterns
of spacetime
curvatures

Structured
Einsteinian
spacetime

Maxwell Demons:
EM hidden variables
Dynamic structures
at all levels

Demons work
at all levels

© 1997 T.E. Bearden

Figure 4. A spacetime curvature engine has myriads of small ST curvature components which act upon mass at all levels.

EXPLORE! VOLUME 9, NUMBER 2, 1999

General Relativity Aspects

Both these coupling mechanisms can be taken directly from general relativity, if one puts one's mind to it. However, unless the wave polarization transformations are known, it is difficult or impossible to apply the ordinary GR *directly*, since relativists usually concentrate upon spacetime curvature by altering 3-space energy density rather than by altering time-polarized "time-energy" density.

In applied general relativity, spacetime itself is an active medium. One speaks of a change in spacetime as a "spacetime curvature". Any curvature of spacetime in a local region directly performs continuous work upon any mass embedded in that region. Vice versa, any mass (or other concentration of energy) in a local spacetime continuously acts upon that spacetime to "curve" it.

So if one wishes to perform actions upon mass in a distant region, one may generate local curvatures of spacetime in that region, and these curvature "engines" will directly act upon the mass at all levels. *[Figure 4]* This is very different from energy propagation through space from one point to another. Now the "ordinary EM energy and actions" arise from every spatiotemporal point within the mass, at every level, and move upwards (from inside to outside). So we speak of specific forms of "templates" of nested spacetime curvatures formed and utilized to engineer mass "from inside out" as *spacetime curvature engines or vacuum engines*.

Use of vacuum (spacetime curvature) engineering[15] is a far more powerful form of engineering than is provided by energy propagation through space. As an example, it is easy to alter the quarks in a nucleon, using spacetime curvature engines and time-charging decay. Indeed, we have developed the mechanisms for cold fusion and the electronuclear interaction (formation of new nuclides) at feeble energy. An Invention has been filed with the U.S. Patent Office, and formal patent applications are in preparation. Some information on these mechanisms and principles has been released[17].

Time as Dense EM Energy and a Strong Spacetime Curvature Agent

The advantage of using the time-polarized "time-energy" for spacetime curvature is that time is ordinary energy compressed by a factor of at least c^{2}—which, in the MKS system of units, is some 9¥10^{16}. Thus use of time-polarized EM photons and waves as ST curvature agents gives an amplification of 9 ¥10^{16} over the use of transverse-polarized EM waves for that purpose. In turn, the use of the strong EM force in ordinary TW waves as an agent of ST curvature is already a nominal 10^{40} times as strong as is the weak G force used as the agent of ST curvature.

The bottom line is this: For spacetime curvature effects, the use of the t-polarized domain provides amplification of some 9¥10^{56} greater than the weak G-force ST curvature agent usually considered in general relativity[18].

Western Science Remains Largely Materialistic

Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system.[8] This serious self-limitation exists because in the body we *measure* only weak TW EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to

Disclosure[16] on this work

measure "mind operations" directly.

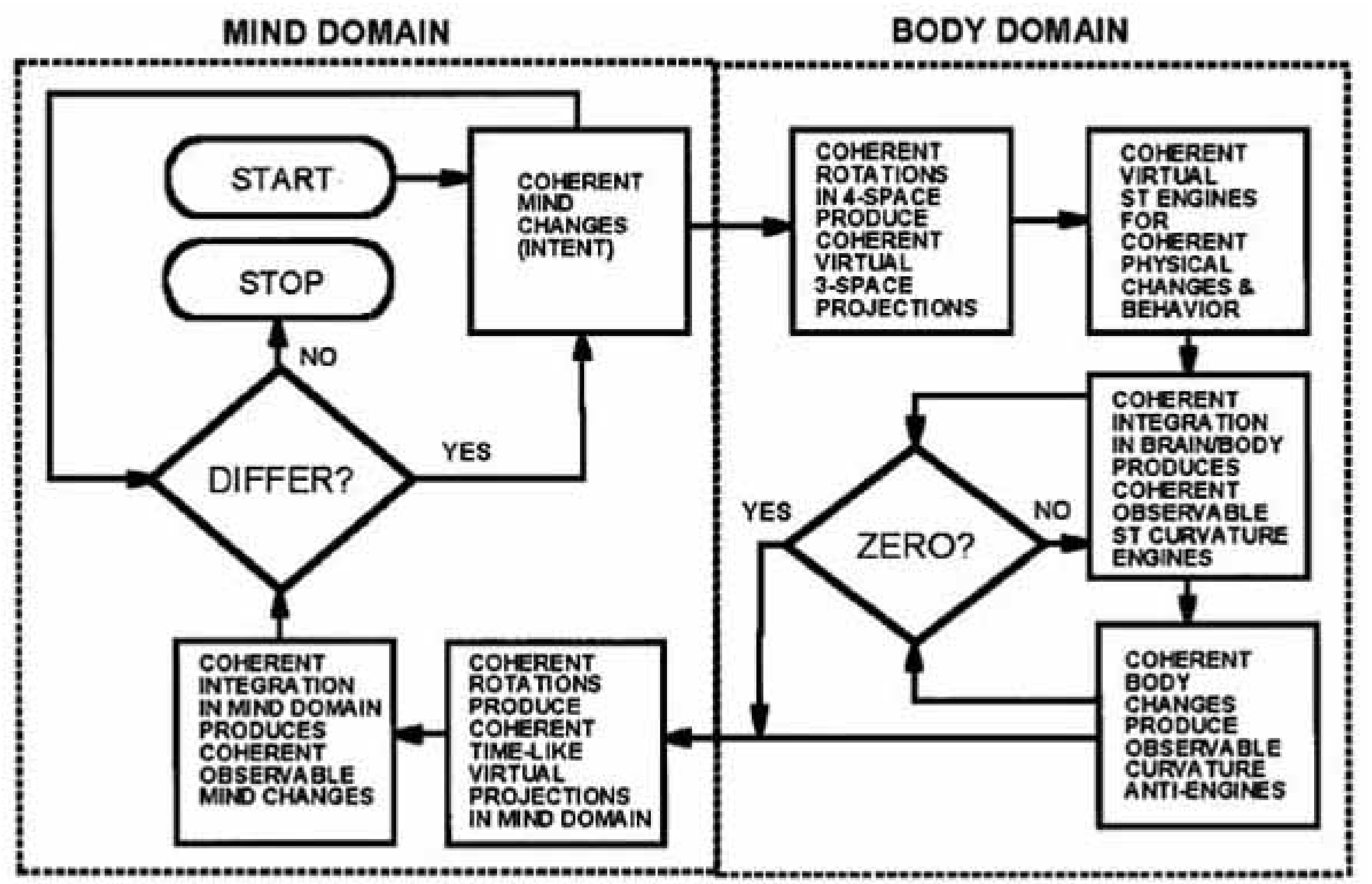
MIND DOMAIN
BODY DOMAIN
START
STOP
COHERENT MIND CHANGES (INTENT)
NO
DIFFER?
YES
COHERENT INTEGRATION IN MIND DOMAIN PRODUCES COHERENT OBSERVABLE MIND CHANGES
COHERENT ROTATIONS PRODUCE COHERENT TIME-LIKE VIRTUAL PROJECTIONS IN MIND DOMAIN
COHERENT ROTATIONS IN 4-SPACE PRODUCE COHERENT VIRTUAL 3-SPACE PROJECTIONS
COHERENT VIRTUAL ST ENGINES FOR COHERENT PHYSICAL CHANGES & BEHAVIOR
COHERENT INTEGRATION IN BRAIN/BODY PRODUCES COHERENT OBSERVABLE ST CURVATURE ENGINES
YES
ZERO?
NO
COHERENT BODY CHANGES PRODUCE OBSERVABLE CURVATURE ANTI-ENGINES

Figure 5. Mind-to body coupling and body-to-mind coupling.

EXPLORE! VOLUME , NUMBER 2, 1999

With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW functions exist in the body and in all of nature in general. Presently only highly theoretical quantum field theorists seem to even be aware of the existence of scalar (t-polarized) photons, and even they are unaware of t-polarized EM waves. Certainly our scientists do not measure t-polarized EM waves in and around the body and in nature, nor do they make them in the laboratory. Apparently they have not studied such waves and their interactions with matter—living and inert—at all.

But it's even worse. In ignoring the time component transported by photons and EM waves, science has also erroneously omitted half of the excitation charging and excitation decay processes whenever a mass interacts with photons and EM waves. More on that later.

Brute Forcing Time Functions versus Fine Control Methods

Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

While KGB scientists also use TW EM "brute force" TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired. They also do not hesitate to use LWs and t-polarized waves overtly. So in blunt terms, Western compared to Russian mind control research is probably like comparing an automobile body shop with a fine surgical ward. Having long ago worked out all those "exact correlates" to the internal "information content of the field" is where the KGB scientists are at least 20 years ahead of the West.

The reason one can get mind and behavior results with the TW electromagnetics, while bypassing the real mechanism which uses t-polarized waves and photons, is due to (1) the very peculiar nature of the EM emission from a dielectric, and (2) what can be done by re-radiating that dielectric with its emitted spectrum, deliberately and very carefully altered in selected parts. We will return to that important feature later.

Polarization and Observability

As we stated, there are four photon polarizations and therefore there must be correspondingly four EM wave polarizations[1] The first three polarizations are the x-, y-, and z- spatial polarizations. The x- and y- polarizations are transverse polarizations and the z-polarization is a longitudinal polarization along the direction of propagation (along the z-axis, by standard notation). Simply put, we may visualize the transverse polarizations as rather like the wiggling of a fish's tail from side to side as the fish moves forward, or a whale's flukes up and down as the whale moves forward, or some combination thereof. We may visualize the longitudinal polarization as a sort of "repetitive accordion effect" along the line of motion of the wave. Usually the z- polarization is neglected in EM wave theory, although in recent years physicists have "rediscovered" longitudinal EM waves and are now intensely researching the use and characteristics of such waves.[12]

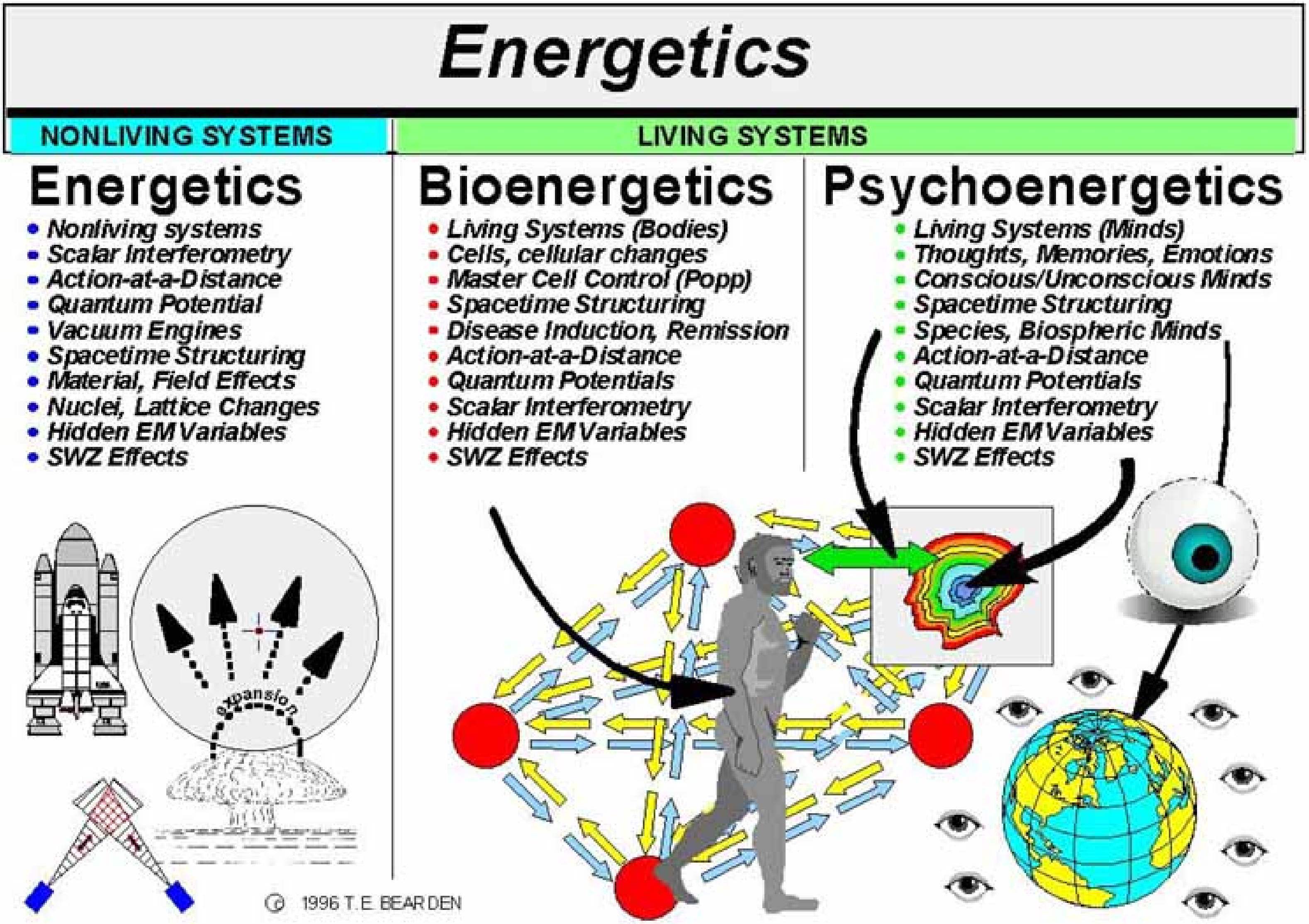

Energetics

NONLIVING SYSTEMS | **LIVING SYSTEMS**

Energetics

- Nonliving systems
- Scalar Interferometry
- Action-at-a-Distance
- Quantum Potential
- Vacuum Engines
- Spacetime Structuring
- Material, Field Effects
- Nuclei, Lattice Changes
- Hidden EM Variables
- SWZ Effects

Bioenergetics

- Living Systems (Bodies)
- Cells, cellular changes
- Master Cell Control (Popp)
- Spacetime Structuring
- Disease Induction, Remission
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

Psychoenergetics

- Living Systems (Minds)
- Thoughts, Memories, Emotions
- Conscious/Unconscious Minds
- Spacetime Structuring
- Species, Biospheric Minds
- Action-at-a-Distance
- Quantum Potentials
- Scalar Interferometry
- Hidden EM Variables
- SWZ Effects

© 1996 T.E. BEARDEN

Figure 6. The three divisions of Soviet energetics and their characteristics.

EXPLORE! VOLUME 9, NUMBER 2, 1999

Unfortunately, in quantum field theory there has been a tendency to regard the t-polarized or "scalar" photon—where the local rate of time is oscillated—as unobservable. Oscillation of the local rate of time produces powerful oscillations of local space-time curvatures, due to the extreme energy density of time. Individually, the t-polarized photon tends to be unobservable. However, a *coherent sequential group* of such t-polarized photons, each individually in the virtual state with respect to the external observer, may simply integrate in its interaction with mass over a short time period into observable change because of the coherent integration of virtual spacetime curvatures into a larger, observable spacetime curvature.

Rotations and Projections in 4-Space

Relativistically, any velocity change in 4-space is a rotation. Any such rotation in the time-domain also creates a tiny projection component into 3-space. Any such rotation in the 3-spatial domain also creates a tiny projection component into the time domain.

Hence a sufficient series of coherent time-domain (mental) changes produces a coherent series of virtual changes in 3-space (the body), thereby coherently integrating into an observable energy change in the body. A sufficient series of coherent 3-spatial energy changes produces a coherent series of virtual changes in the time domain which coherent integrate into an observable time-domain change.

Solution to the Age-Old Philosophical Problem of Intent

This is in fact the solution to the age-old problem of *intent*, or how the nonmaterial (i.e., non-3-spatial and non-observable) mind induces a 3-spatial, observable energy change upon the 3-spatial body. It is also the solution to the problem of *awareness*; i.e., how the mind is aware of itself[19] and of the responses of the body. Together the two form a closed loop coupling of the mind and body. The time delay in the loop together with memory recall for comparison, creates the sense of "persistence" of self in time. This also creates the sense of being a "separate, closed being" (i.e., of separate persistence in time—in the living entity). The sense of "separation of self from an external world" is created by comparing those body-to-mind sensory feedbacks which are not correlated to the mind's previous feed-forward intent.

Time As Energy and Why It Is Very Dense Energy

In addition to the three spatial polarizations of photons and EM waves, there is a very, very useful t-polarization along the time axis. In this polarization, the 3-spatial energy is not oscillating at all. Instead, the time or time-energy is oscillating. Time can be taken to be energy compressed by at least c^2, so it has at least the same energy density as mass. In other words, one second is $9x10^{16}$ joules of time-energy (energy compressed into time). The t-polarized photon or EM wave is called the scalar photon or scalar EM wave, respectively.

To demonstrate why time can be regarded as energy, we need only point out that the choice of fundamental units in any physics model is totally arbitrary. We usually choose these units to ease the mathematical manipulation, ease our understanding, and simplify the ability to visualize or "grasp" the physics. However, if one wishes, one can build all of physics from a single fundamental unit—e.g., energy. In that case, one readily sees that time is a function of energy and only energy. Hence it is perfectly proper to regard time as energy, and to seek out what form the function takes in our normal system of units (say, the MKS system).

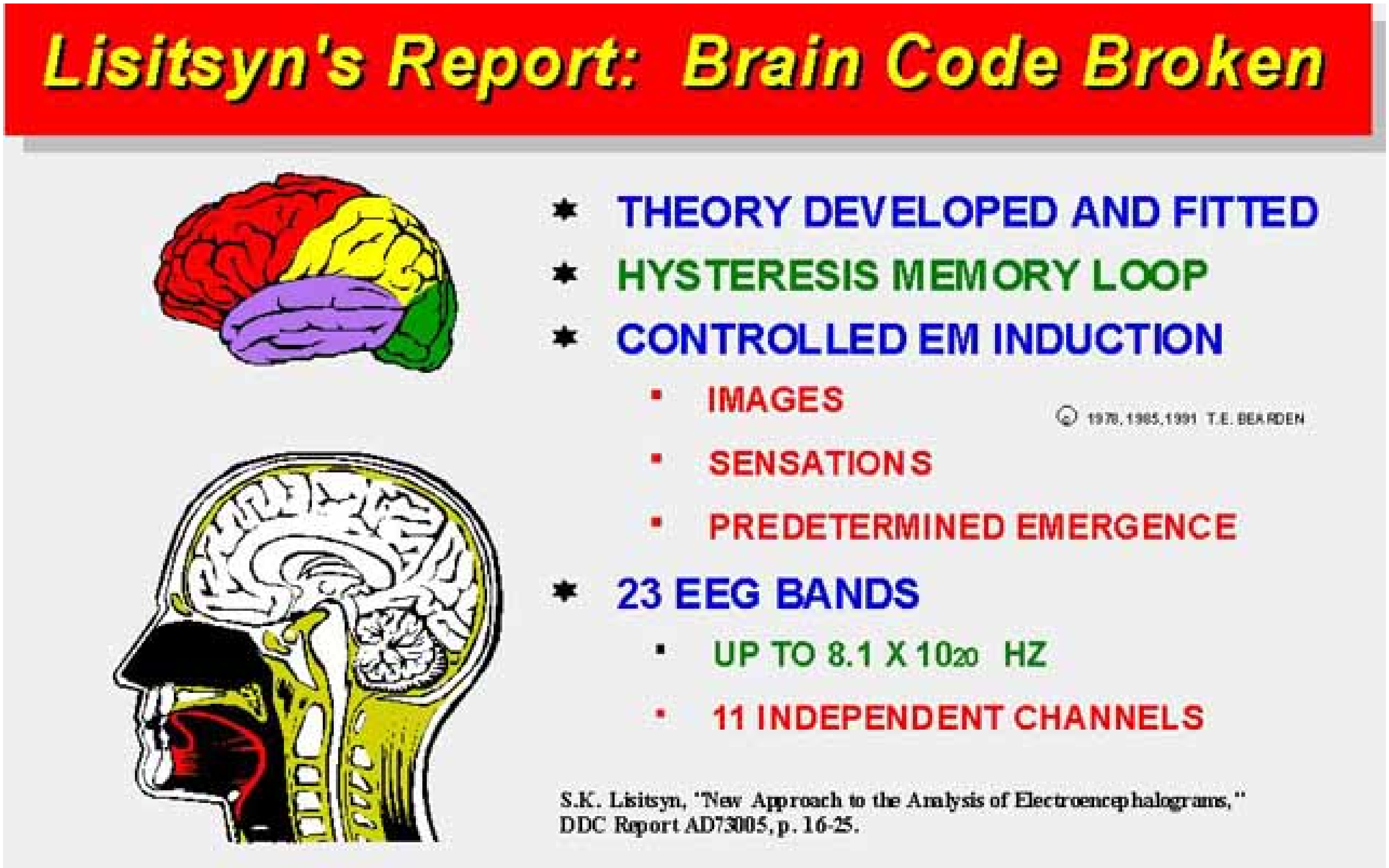

Lisitsyn's Report: Brain Code Broken

- THEORY DEVELOPED AND FITTED
- HYSTERESIS MEMORY LOOP
- CONTROLLED EM INDUCTION
  - IMAGES
  - SENSATIONS
  - PREDETERMINED EMERGENCE
- 23 EEG BANDS
  - UP TO 8.1 X 10^{20} HZ
  - 11 INDEPENDENT CHANNELS

© 1978, 1985, 1991 T.E. BEARDEN

S.K. Lisitsyn, "New Approach to the Analysis of Electroencephalograms," DDC Report AD73005, p. 16-25.

Figure 7. In the 1960s, Lisitsyn revealed Russian scientists had deciphered the human brain code.

EXPLORE! VOLUME 9, NUMBER 2, 1999

In that system, it appears that any change in time Δt can be taken as always a function of a corresponding change in energy ΔE, where $\Delta t \leq (\Delta E) \div c^2$ [1] from which it follows that $\Delta E \geq (\Delta t) \times c^2$ [2] For convenience, we take the special case where $\Delta t = (\Delta E) \div c^2$ [3] $\Delta E = (\Delta t) \times c^2$ [4].

Observation as Used in Physics Is Spatial

We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain, and thus "lost" by the stripping away of time in the ordinary observation process. Physical observation is essentially a time-differentiating mechanism applied to a 4-spatial change, or in terms of MKS fundamental units L and t, observation θ is $\theta = \partial/\partial t(L^3 t) = L^3$ [5].

Thus, as is well-known in quantum mechanics, physical observation is 3-spatial, and time is not a physical observable, even in theory. Since mind is time-like, it follows that mind is not a physical observable either, since discarding the time dimension also discards the mind. In short, one may also take physical operation as the mechanism that separates mind and body. To observe is to separate. We point out, however, that merely "separating and discarding" the time domain (as in physical observation of a single change) does not eliminate it, nor does it eliminate the single time-like mind change that may be involved in an intentional volition. The time domain certainly remains, even though only the 3-spatial intersection of the 4-spatial mechanism and the resulting "outputs". So a series of coherently integrated "mental intent" changes introduced into the human body's overall servo-mechanism provides the continuing input. From there, ordinary physics will generate the resulting actions induced in the body by that servomechanism and its amplifying mechanisms.

Two Coupling Mechanisms Make a Closed Loop

This is the "mind-to-body" coupling mechanism. [See *Figure 5*] It is the mechanism whereby the mind is coupled to the body, and whereby mental intent is able to induce a series of physical inputs into the body servomechanism.

So the body's servomechanism then generates the responses of the body (including everything from chemistry to electrical changes to muscular movements, etc.). These responses are changes in 3-space.

The conscious mind is a *serial processor*, though extremely rapid. It produces the series of coherent intent inputs for volitional behavior.

The unconscious mind is a *massively parallel processor*.[20] It continually produces the vast series of coherent "unconscious intent inputs" for control of all the deeper processes in the body, beyond usual conscious awareness.

However, as the body moves or changes in 3-space, each resulting quantum change in body 3-space is also a slight rotation out of 3-space and toward the time axis. Hence it induces—in the "virtual state" in the time domain—a precisely correlated projection. The body's responses are in general

Minkowski change is given by physical observation. So we may say *that the time-change remains in the virtual state, with respect to attempted physical observation of a single Minkowski 4-space change.*

Rotation, Coherent Integration, and Intent

Previously we pointed out that any change in a 4-space entity may be regarded as a rotation away from the "trajectory" of the entity. Hence each and every t-polarized change creates a very small projection into 3-space by its rotation slightly away from the time axis. All mind changes in the time-domain actually produce virtual 3-spatial projections in 3-space (in the physical domain). We define "intent" as the continued production of successively coherent mental changes, producing coherent virtual changes in the 3-space body system, with coherent integration of those 3-space virtual changes into observable 3-space energy inputs into the 3-space body system. Successively coherent mind-changes will produce successively coherent 3-space virtual changes. In short, coherent mind-changes will produce coherent integration of those 3-space virtual projections into an observable 3-space change. This is the creation of an ordinary 3-spatial energy change induced upon the 3-space body. In short, this is the mechanism whereby time-like mental intent is able to produce a series of coherent observable quantum changes in the physical body (as in the brain and nervous system).

The Body as a Servomechanism

From the standpoint of control theory, we may consider the body system to be a coherent, so a coherent series of virtual state changes in the time-domain (in the mind realm) are created successively. In short, again we have coherent integration, this time in the mind or time domain. This produces "observable" changes in the time-mind domain, which are coherent with the body's 3-spatial changes actually performed. Thus the mind receives feedback directly from the physical movements and changes of the body. This is the manner in which the body is coupled back to the mind.[21]

By comparing the "intent" behavioral move that was "fed forward" into the body, with the return "response" move analog that was "fed back" from the body to the mind, the mind is able to determine errors and differences, and originate additional correctional commands.

Thus the entire mind-body loop is a closed-circuit system of feedforward and feedback, together with corrections. It also has multiple levels of such, infolded in the larger volitional levels. See again *Figure 4*.

This solves the age-old problem of the mechanisms for the mind-body coupling, intent, volition, conscious and unconscious functioning, sense of the external world, sense of the internal world, sense of "being in" the external world, etc.[22]

All This Is Included in Russian Psychoenergetics

This is the highly summarized basis for *psychoenergetics*, the KGB's division of energetics that deals with the mind and body coupling and functions, and direct engineering of (1) the mind-body coupling and (2) the mind operations directly. [See *Figure 6*] The KGB intent, of course, has always been to exploit this science for the complex servomechanism system comprised of many subordinate servo systems with feedback and feedforward looping.

Given the input from intent, ordinary servo theory will take it from there. Once a servo has an input, servomechanism theory describes the response of the

degradation, killing, and control of human beings, including all humanity.

The Russians know full well that, if you produce and utilize scalar EM photons and t-polarized EM waves, you can directly affect and engineer mind and mind operations at any and all levels. Western researchers, who know nothing of how to make

EXPLORE! VOLUME 9, NUMBER 2, 1999

time-polarized photons and time-polarized EM waves, do not yet know that. Consequently the Russians have developed a highly secret science of *directly* engineering the mind and its operations, including thought, images, perceptions, feelings, emotions, memory, and mind processing. Indeed, Lisitsyn[23] wrote quite specifically of this capability in the 1960s. [See *Figure 7*]

Western clandestine mind control researchers are apparently still slowly and painfully fitting TW EM irradiation correlates to induced or resulting mental and physical behavior responses. They seem unaware of the actual wave transductions occurring inside the body and mind, but are unwittingly inducing those transductions in hidden fashion anyway and in "brute-force input-response fitting" models.

These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design—for example—of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.

This fine research article will be continued in Part II with Russian Methodology, Waves and Wave

spacetime that is highly active. The mind is rooted in the time-domain and projects from that domain into 3-space. The body is rooted in the 3-space domain and projects into the time-domain. To "change" or "function" in one domain is automatically to function in the other. EM and gravitational phenomena are still modeled separately by human scientists, but not by nature.

3. In quantum mechanics, time is not an observable, but merely a parameter.

4. It is convenient to consider mass as a 3-spatial form of condensed energy.

5. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics, " Mathematische Annalen, Vol. 57, 1903, p. 333-355;— "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372.

6. A NERAC document search yielded some 16 good references on t-polarized photons, but not a single reference on t-polarized EM waves.

7. E.g., such previously inexplicable instrument anomalies have accompanied some excellent and rigorous electrolyte experiments at U. S. Navy research facilities at China Lake. For a description, see Melvin H Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", Proc. ICCF - 7 (International Conference on Cold Fusion—7, Vancouver, BC, Canada, Apr. 1998, p. 101. For a brief explanation of the anomalies, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, " Explore!, 8(6), 1998, p. 7-16;— "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

8. Many U. S. researchers and journalists—and even many scientists—have wrestled with the problems of mind, intent, and mind-body

Transduction, The Cellular Control System, and other matters of interest. ♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations

coupling. Enumeration of those efforts would itself require an entire book. For an introduction into that domain, the reader is referred to the very determined exposé by Cheryl Welsh in her very timely book, The 1950's Discovery of the Code of the Brain, May 1988, published on the Internet. Simply employ any Net search engine and search on the name Cheryl Welsh. A connection to the site and to Welsh's book will immediately be found Welsh's important compendium is highly recommended, giving the reader a "birds eye view", so to speak, of how Western scientists think and proceed regarding the subject of mind control by electromagnetic means. Unfortunately all journalists so far researching the area have not been aware that Western electrodynamics itself is seriously flawed, and that the great Russian advances in mind engineering and mind control are a result of their secret but complete revision and correction of Western electrodynamics to provide the basis for energetics. Until Western scientists revise their own decrepit old electrodynamics fouling the textbooks, they will never catch up to the Russian mind control developments springing from intense development programs that did that revision in the late 1940s and early 1950s.

9. An Invention Disclosure on these processes and enhancement embodiments has been filed with the U. S. Patent Office. Formal patent applications are also in preparation.

10. Energetics is a unified science of an extended electrodynamics possessing a hidden infolded general relativity inside the potentials, fields, and waves. It deliberately employs higher polarizations (longitudinal and tempic) of photons and EM waves to engineer action at a distance, use of specific patterns of spacetime curvature created at a distance, and direct engineering of either inert or living bodies. It also encompasses the direct engineering of time-like mind and mind functions. It is divided into three divisions, depending upon the nature of the target. Targeted against inert materials, structures, fields, and waves it is called "energetics. "Against living bodies, their fields,

and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

---

1. E.g., see Lewis H Ryder, Quantum Field Theory, Cambridge University Press, 2nd Edition, 1996, p. 147+. For an even deeper discussion of photon polarization, see F. Mandl and G. Shaw, Quantum Field Theory, Wiley, 1984, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e. electrostatic) potential.
2. Nature is unified and not separated into the "separate" disciplines our science has prescribed. Thus the mind is also totally electrogravitational. The mind works not just with EM
forms in an inert space, but also with spacetime curvatures in a

and biochemistry, it is called "bioenergetics." Against the mind and mind functions, it is called "psychoenergetics".
11. Psychoenergetics is that branch of energetics used to directly affect and engineer the mind, memory, perception, thought, emotion, sensation, personality, and other mind functions of a living biological organism. It is usable against a single personal conscious (serial processor) mind and a single personal unconscious (massively parallel) processor mind In theory it is also usable against lungs collective unconscious mind and its functions, as well as the deeper unconscious for all species on Earth (i.e., against Gaia). We are speaking of physics, not metaphysics.
12. W A. Rodrigues Jr.. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 \leq v < \infty$ in nature," Foundations of Physics, 27(3), 1997, p. 435-5•08 A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory website.
13. E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions, "Proc. Lond. Math. Soc., Series 2, Vol. I, 1904, p. 367-372. The paper was published in 1904 and orally delivered in 1903.
14. E.g., see T. E. Bearden, Fer-de-Lance, Tesla Book Co., 1986;—AIDS: Biological Warfare, Tesla Book Co., 1988;—Soviet Weather Engineering Over North America, 1-hour videotape, 1985,—Gravitobiology; A New Biophysics, Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991,—Energetics of Free Energy Systems and Vacuum Engine

EXPLORE! VOLUME 9, NUMBER 2, 1999

Therapies, Tara Publishing, Internet node www. tarapublishing.com/books, July 1997; — Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ pages furnished to selected government agencies and key Congressional committee leaders.

15. E.g., see T.D. Lee, Particle Physics and Introduction to Field Theory, Harwood; New York, 1981. On P.383, Nobelist Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) is not utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25; Outlook; Possibility of Vacuum Engineering, " p. 824-828. With wave transductions and use of higher polarization EM waves, there are far better and more powerful ways to do vacuum (spacetime curvature) engineering.

16 T.E. Bearden, "Formation and Use of Time-Reversal Zones, EM Wave Transduction, Time-Density (Scalar) EM Excitation and Decay, and Spacetime curvature Engines to Alter Matter and Convert Time Into Energy," Invention Disclosure Document #446522, Oct. 26, 1998.

17. T.E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves, "Explore!, 8(6), 1998, p. 7-16; — "Toward a Practical Unified Field Theory and a Deep Experimental Example," Proc. INE Symposium, Univ. Utah, Aug. 14-15, 1998.

18. We have used the enormous energy density of time in our approach to a totally new kind of electronuclear interaction at feeble spatial energy. By use of a time-reversal zone (TRZ) in a fluid; the free H+ ions—which are free protons—can be induced to attract each other so tightly that the edge of the nuclear force region is entered. In that case the protons form a quasi-nucleus. When the TRZ decays back to a time-forward zone (TRZ), this new quasi-nucleus decays by changing quarks in one or more

means "sense of existing in time".

20. The "unconscious" is totally conscious, just multiply so. The conscious mind can see only "a single slide in the slide projector at a time, " so to speak. When it "looks" at the ongoings and outputs of the unconscious, it "sees" millions of slides in the slide projector simultaneously. Hence it cannot distinguish "one and one only". So it just sees nothing, where by "nothing" we mean "no single distinguishable thing." It therefore is unable to consciously perceive the actions of the unconscious. When the unconscious wishes to communicate with the conscious (because of unresolved conflicts, etc.), it must produce large groupings integrated into "symbolic scenarios". Hence the symbolism of the unconscious, and the basis for therapeutic psychology and psychiatry. A "symbol", after all, is just a thing which can have many different meanings, all folded into one.

21. The mind continually has activities of its own, each of which produces coherent changes in the body to a much smaller degree. The interweaving of these "smaller self-changes" infolded inside the 3-space major intent changes also produces slight body changes, which in turn "project back" into the mind. This huge group of "infolded self-loops" produces the sense of "my body" being "volitionally directed or occupied" simultaneously with the "greater intent" loops which represent "my intentional actions with my body". Further; the body servo continually receives a myriad of inputs from its sensory systems, both from changes inside the body and from incoming changes from the external universe. This interweaving produces the awareness of "being in the external world" by having (i) changes whose feedback do not match mind-intent changes, and (ii) changes whose feedback does match mind-intent changes. This of course applies to both the conscious and unconscious levels. The purpose here is to demonstrate that there exists an actual physics of the living being, the mind as well as

nucleons, thereby producing new nuclides. We point out that this is actually a very high energy reaction; however; since the time-energy is used and is very, very dense, the amount of 3-spatial energy that must be used is very small. If this new mechanism holds, it represents a revolution in particle physics.

19. "Self-awareness" refers to be continually "refreshed" by time-delayed coherent feed-back inputs from the body to the mind; where these inputs compare to the recalled memory of the minds inputs to the body. In short, the time-delay coherence creates the sense of "existing in time". Awareness usually

the body, and the coupling and interfiinctioning of mind and body.

22. The methodology also easily extends to include deeper interlays for the collective human unconscious mind and for Gaia—the collective unconscious mind level for all the species on earth.

23. S. K Lisitsyn, "New Approach to the Analysis of Electroencephalograms, " in Problems of Bionics— (Selected Articles), Defense Documentation Center Report AD 73005, date unknown, p. 16-25. Available through the National Technical Information System (NTIS).

Mind Control and EM Wave Polarization Transductions, Part 2

©Copyright 1999 by T.E. Bearden, USA

This article continues on with T.E. Bearden's article, Part 1 of which was published in Volume 9, #2 of Explore! for the Professional. Because we feel that Col. Bearden's caution below is so very important we are repeating its publication in all parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

A Strategic Threat Completely Unrecognized by the West

The Russian methodology leads directly to rather mind-bending, extended capabilities. KGB psychoenergetics scientists seem fixed on an eventual goal of directly engineering Jung's collective human unconscious, thereby converting the entire human species into a sort of "ant" species. In short, they seek to engineer what they regard as "perfect communism and perfect order" directly in the entire species. *Figure 8* indicates the direct connection of every human's conscious and unconscious mind to the collective species unconscious and to Gaia, the collective unconscious for all species on Earth. Since time is internally structured and layered, it follows that time-like mind levels are also internested and layered.

This appears to be a significant strategic threat, *well along in its development*, that our scientific and intelligence communities have completely failed to recognize.

The direct engineering of the entire human collective species is a doable.[1] It is going to be done, whether we like it or not. In that race (which may eventually determine whether or not we retain any personal freedom at all), the "democratic" nations are woefully behind. In fact, the West has not even realized there is such a new kind of "arms race" for dictatorship and control of the "collective species mind" and we have already almost lost it. The West has not yet even developed the overall theory of psychoenergetics.

Waves and Wave Transductions

A pure longitudinal EM wave (LW) would have infinite energy and infinite speed. In practice, one can only approach this perfect LW wave, and one always has transverse residues remaining. Hence the "imperfect" longitudinal EM waves we can make in suitable equipment and techniques, are known as "undistorted progressive waves" (UPWs). The UPWs may have velocity far greater than the speed of light, or less than the speed of light. So the speed v of UPWs can vary as

$$0 < v < \infty \qquad [6]$$

The Russians also know that, when a nearly-pure longitudinal EM wave interacts with matter (it usually doesn't interact very much, and has a low reaction cross section), the mass immediately adds the phase conjugate replica to the interacting LW portion, regardless of frequency, thereby transducing the input LW interactions into scalar (time-polarized) EM interactions.

One transduces from TW to LW to Scalar EM via iterative phase conjugation. One transduces from scalar EM to LW to TW via iterative interferometry. It took the present author 20 years to discover this simple thing; we now have filed an invention disclosure on the process and embodiments for performing it, and will be following with formal patent applications. Let TW = transverse EM wave, LW = longitudinal EM wave, and TDW = time-density EM wave (another name I use for t-polarized or scalar EM waves, since the density of time-energy is what is varying). Let (I) = interfering with, PCR = phase conjugate replica wave with a coupled set of parentheses enclosing the type of wave. The rules for wave type transduction are:

$$\text{TW} + \text{PCR(TW)} \Rightarrow \text{LW} \qquad [7]$$

$$\text{LW} + \text{PCR(LW)} \Rightarrow \text{TDW} \qquad [8]$$

Going the other way, the rules are:

$$\text{TDW (I) TDW} \Rightarrow \text{LW} \qquad [9]$$

$$\text{LW (I) LW} \Rightarrow \text{TW} \qquad [10]$$

In EM smog (dense EM signals, even though very weak), there is increased nonlinear multiwave interactions, both interferometric and phase conjugative. So in such EM smog there exists an increased, non-negligible component of LW and TDW irradiation, formed by the above transduction interactions in the irradiated living cells and bodies. This latter part accounts for many of the long term effects which show up in "correlative" studies, but which do not show up in the normal limited-frequencies TW experiments produced by so-called "physical experiments."

The standard EM bioeffects manner of experimenting uses a theoretical model of simple, direct energy deposition by means of absorption in cells and tissues. Now we must utilize an "open system" throughout the body dielectric in which higher polarized EM energy (either LW or t-polarized) travels readily throughout all areas of the body.

Explore! Volume 9, Number 3, 1999

EM Bioeffects and Wave Transductions

In the EM bioeffects experiments to date, there has been absolutely zero control of the internal transductions, and there has been no consideration of them. Almost all transductions of TWs into LWs will immediately result into transduction into TDW and time-excitation charging of the cells and all their internal components. These time-excitation charges (excited states of time-charging) will then slowly decay, emitting longitudinal EM waves in the process which travel throughout the body. These emitted LWs interfere with each other and the LW processes in the body, producing low level TW "noise jamming" of the cellular processes and—most especially—of the deep cellular control system of the body. The latter system functions by means of LWs and TDWs. Long-term effects of such "noise jamming" of biochemistry reactions and the control systems of the cells and body, can lead to degenerative diseases such as arthritis, atherosclerosis, Alzheimer's disease, heart disease, cataracts, tumors, leukemia, somewhat accelerated aging, immune system weakening and depressing, etc. As the body's natural defense and regeneration ability weakens, opportunistic infections further add to the degeneration.

It was shown by Whittaker in 1903 that any scalar EM potential is comprised of a harmonic series of bidirectional longitudinal EM wavepairs. [See *Figure 9*] Each longitudinal EM wave pair in the harmonic series is a phase conjugate pair. Unknown to Whittaker, that "LW phase conjugate pair" is actually a time-polarized EM "photon" structure, containing a phase conjugate LW photon substructure. So staid old "voltage" (potential) is comprised of a harmonic series of time-polarized EM waves!

The corresponding "t-polarized photons" are actually spin-4 *supergravitons* comprised of coupled antipairs of LW "photons". Each LW photon is a spin-2 *graviton* comprised of a coupled TW photon-antiphoton pair. Each TW photon and antiphoton in the coupled pair has spin-1. The couplet, being the longitudinal "photon" or graviton, thus has spin-2. And so on.

In 1904, Whittaker showed that any EM field or wave pattern can be decomposed into two scalar potential functions. By applying Whittaker 1903 decomposition to each scalar potential function, any EM field or wave has an enormously rich internal structure, comprised of scalar potential functions in interference. Each scalar potential function is comprised of time-polarized EM waves (made of spin-4 supergravitons). Each time-polarized EM wave is further comprised of bidirectional longitudinal EM wavepairs. Each longitudinal EM wave is further comprised of phase-conjugate pairs of transverse EM waves. Each transverse EM wave is further comprised of two scalar potential functions, each of which is in turn comprised of an infinite harmonic series of time-polarized EM waves, and so on *ad infinitum* and *ad nauseum*. Obviously all EM potentials, fields, and waves have an enormous "internal structure" or internal "information content of the field".

The Cellular Control System Uses t-Polarized EM Photons and Waves

The master cellular control system (MCCS) in the body utilizes these photon-gravitons, their corresponding EM wave polarizations, and transductions between them. By the use of very powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime to not merely send a "signal" and hope the cell itself acts upon it. Instead, it sends an "engine" which alters the local spacetime in which the cell and all its components are embedded. This "local ST curvature engine" then physically works directly upon the cell and all its parts, to alter the cell according to the specific "template" of the engine.

This is the direct application of the Wheeler's simple statement of general relativity: *Mass (trapped energy) acts upon spacetime to curve it, and curved spacetime acts upon mass to change it.*

Popp's "ordinary photon" study of this cellular control system[2] therefore will detect the coherent ordinary TW photon residues which result from the photon polarization transductions occurring from the actions of the ST curvature engines upon the cells.

Cellular division, biochemical processes, growth and growth control, cell differentiation and dedifferentiation, etc. are all controlled by the MCCS or one of its subsystems, using spacetime curvature engines. The ordinary TW electronic activity our biophysicists detect and correlate are simply the intermediate components of the actual ongoing polarization transductions.

The Cellular Control System Uses EM Time-Energy

One important subsystem of the MCCS controls the immune system's functions, including the deepest and most subtle functioning performed by all elements of the immune system. This system has been rather intensively studied, but since our biophysicists have not possessed the full electrodynamics utilized, only the correlated residue of normal TW EM waves and photons have been recognized and studied. Hence Western biophysics is in its very crude infancy.

Incorporation of the total wave and photon polarizations, together with their transductions, will eventually present a revolution in both biomedicine and biophysics. Unfortunately, presently our entire medical community is very firmly entrenched in only ordinary TW EM polarizations and ordinary biochemistry. Our scientists thus are studying less than 1% of the functioning of the living organism.

As an example, since Western scientists have no knowledge or research into the time-polarized mind domain and mind-body coupling, the mystery of the *placebo effect*—in which the mind has been verified to produce actual physical effects – will indefinitely remain a mystery.[3] As one result of that "crippling" of our medical theory, medical experimenters have to deliberately drive experiments on drug effects, etc. beyond the "mind's own threshold" of physical effect. In short, they have to drive the experimental results over and above the results being accomplished by the placebo effect. They must perform valid placebo control experiments, and subtract the placebo component results from their own experimental results, to determine the "delta" induced purely by the drugs.

The Regenerative Subsystem

Another important subsystem of the MCCS controls the *regenerative system* of the body. The regenerative system has various mechanisms, including growth control, differentiation and dedifferentiation of cells, etc. Becker's portrayal of the regenerative system is shown in *Figure 10*. One of its most important functions is healing and restoring damaged or diseased cells. It does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] powerfully condensed gravitons and supergravitons, it is able to utilize small but highly significant direct curvatures of spacetime in precise patterns. In short, it utilizes general relativistic spacetime curvature engines, rather than "EM photon signals", to accomplish its control of the cells and their processes. It does

does that by a most remarkable process.

This system has been very poorly studied, mostly by Becker[4] *et al.*, although there is still sporadic, ill-funded, and ill-planned research on regenerative effects from time to time.[5] Becker and his colleagues came closest, and—considering the severe limitations of the orthodox U(1) EM model available to them—Becker's work is particularly revolutionary. He demonstrated beyond question that simple DC potentials could generate cellular differentiation and redifferentiation, and produce results at global sites in the body and not just at the local injury sites where the DC potentials were applied.

Becker was nominated for a Nobel Prize for his Herculean efforts and remarkable results. For his trouble he was hounded, his funds were withdrawn, and he was forced to retire at an early age in his 50s. His biggest "heresy" was to bravely point out that power line radiations had deleterious effects upon biological systems, and to testify to it in a series of hearings and court cases where a very powerful electrical power industry essentially ran rough shod over science and the public.

Organized Science Is Highly Politicized and Imperfect

It is not only politicians that are "bought and paid for" by powerful vested financial interests. In organized science, precisely the same arrangement holds true. Scientists are no better or no worse than any other segment of humanity. A surprising percentage will produce the results desired by their benefactors, if provided with cushy jobs, good income, prestige, and secure positions.

It has been said that money is power. It has also been said that power corrupts, and absolute power corrupts absolutely. In the scientific community, as well as in the political community, this "great corruption" by great financial power has been elevated to an art form. Simply examine the funds made available to scientists. There is not enough for all the scientists, so it is fiercely competitive. If the scientist does not get grants (e.g., a would-be professor at a university), he does not bring in extra money to the university. He cannot get his graduate students funded, so his own innovative (read: heretical) research is thwarted. His papers are rejected by the journals, and before long he has no funded position in science. If lucky, he will be employed in the grocery store or the butcher shop.

Every bit of funding available for the scientists to compete for (and apply for grants) already has firmly fixed controls of exactly what research can be done with that money. The "science" that will be applied is already formulated. That and only that can be done. It is a tribute to the dedication and resourcefulness of our working scientists themselves, that anything innovative at all gets done.

One can have a very fine scientific career, tenure as a professor, get one's graduate students funded, get one's papers published in prestigious journals, win awards and achieve stature and prestige, if one just plays the desired game prescribed by the "old boys schools" in control of almost all the scientific funding. Yes, one can permissibly move the decimal point in the present models a bit, and be rewarded handsomely for doing so. But one is damned if one attempts any serious change or overthrow of obsolete scientific models and practices favored by the "system".

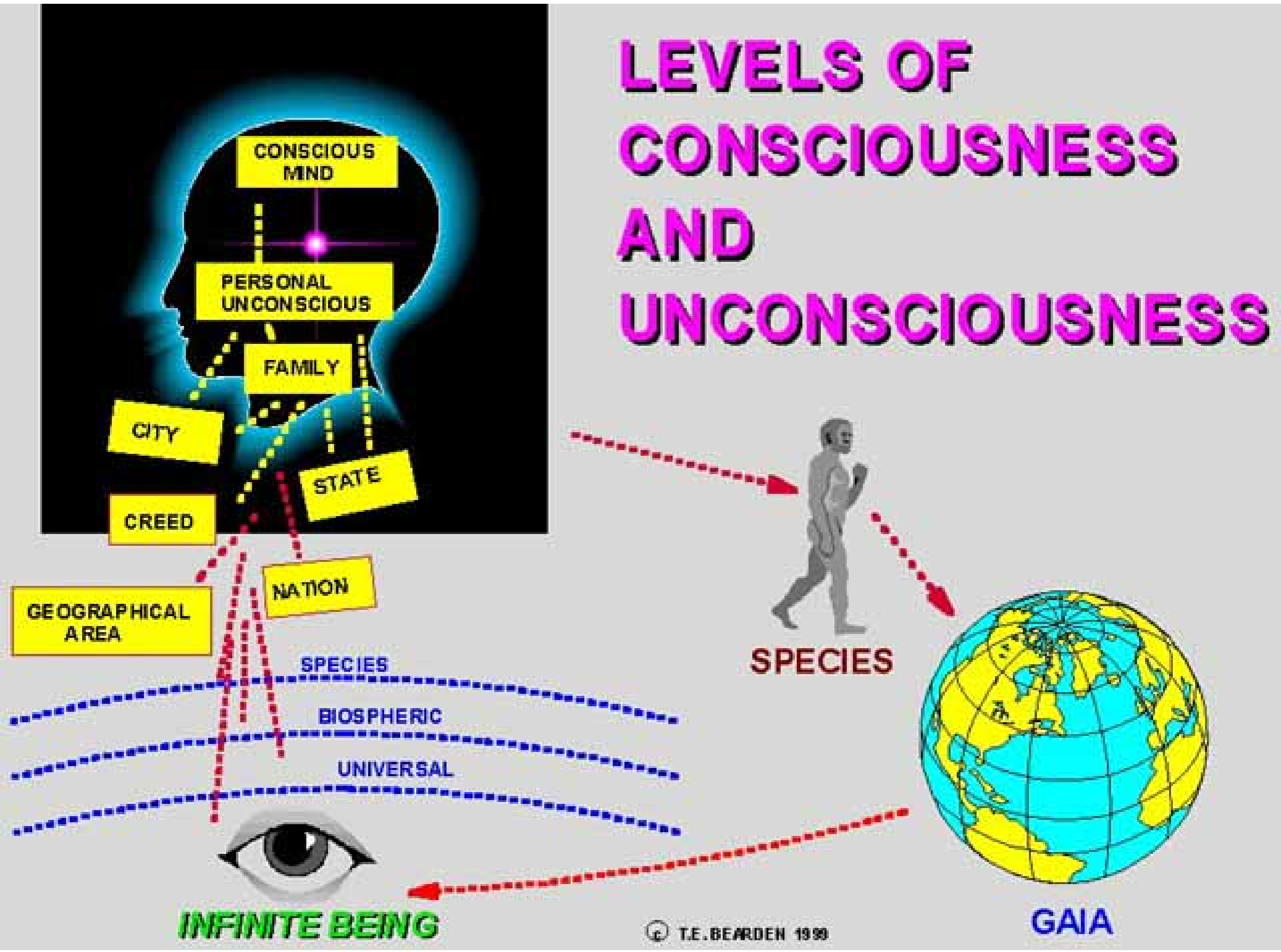

Figure 8. Nested levels of consciousness and unconsciousness.
The mind levels contain the "life" and "mind" dynamics.

EXPLORE! VOLUME 9, NUMBER 3, 1999

That is precisely why all our universities continue to teach a terribly flawed electrodynamics that is a caricature with at least 32 major foundations flaws in it.[6]

And that's why the organized Western scientific community has betrayed the taxpayers who largely fund its research. In short, the scientific community fiercely enforces dogma, suppresses innovation, and has denied the U.S. the very defense we need to survive against our pressing enemies who have not been so scientifically dogmatic.

Microwave Radiation of the U.S. Embassy in Moscow

Because of the strongly enforced adherence to the insane foundations errors in classical electrodynamics, our scientists still do not comprehend the decades of so-called "microwave radiation" of personnel in the U.S. Embassy in Moscow.[7] All diseases and health changes produced in those personnel were in *field-free areas* (i.e., where the potentials were persistent and unchanging, with no gradients etc.)[8] No one even thought to look at the "inner longitudinal EM phase conjugate biwave composition" of those potentials, where they would have found the Whittaker internal bidirectional longitudinal EM wavepairs and the internal time-polarized EM waves. They would also have found the infolded, very powerful general relativity: the spacetime curvatures deterministically formed and being utilized to directly alter the irradiated cells and their constituents.

Instead, government-funded studies did a rather straightforward "force field" analysis, and concluded that, since there were no fields there, it could not have been the microwave radiation.

Elementary statistics would immediately indicate a different conclusion: The diseases and health changes were 100% correlated to the presence of potentials, and 100% anticorrelated to the presence of fields. Hence it was the field-free potentials that had to be inducing the diseases and health changes. If it were NOT the radiation at all, then some changes would have occurred in personnel in areas where fields were *present*, as well as in areas where fields were *absent*. Since this did not occur, such a premise is destroyed by self-contradiction. In short, the data actually proved (i) it was indeed the radiation that was causing the diseases and health changes, and (ii) further, it was the field-free potentials that were the culprit.

That this elementary statistical conclusion eluded our scientific community and our government community, is absolutely inexplicable.

Three U.S. Ambassadors eventually died of diseases induced by their exposure to the Embassy radiation. Ironically, Kaznacheyev[9] had openly released some of the results of thousands of Soviet military experiments showing that any kind of cellular disease or disorder could be induced in targeted cells by novel EM radiation emitted from diseased or damaged cells. Apparently no one in the West deciphered the mechanisms involved in Kaznacheyev's epochal work.

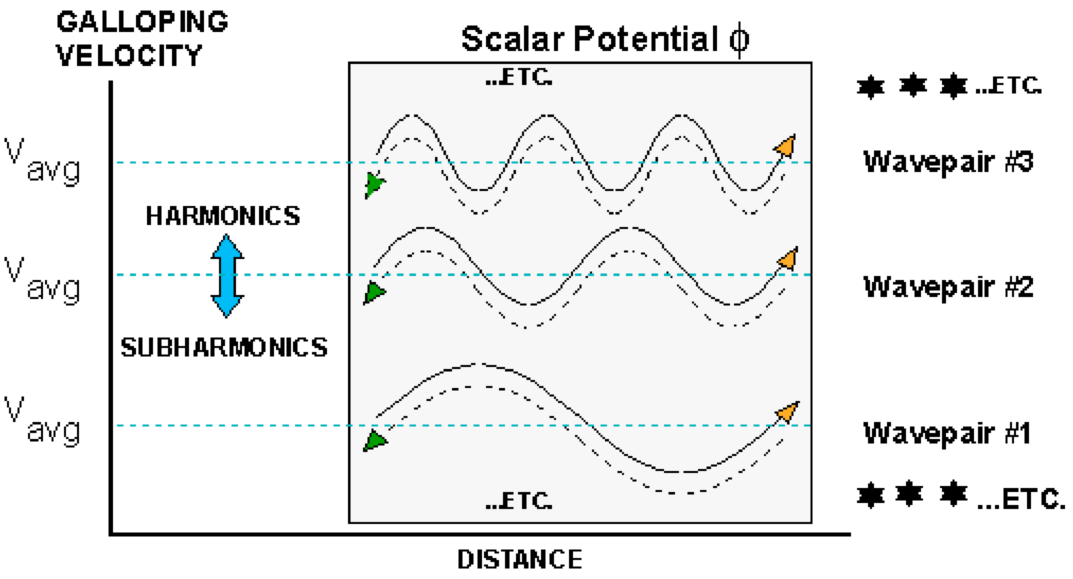

The Structure Is:

- **A harmonic set of longitudinal phase conjugate wavepairs.**
- **In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.**
- **The convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.**
- **The charge's spin is 720 degrees, 360 in the real plane and 360 in the imaginary plane.**
- **Hence the charge receives the complex convergent EM energy, transduces it into real EM energy, and emits enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").**
- **This produces the fields and potentials from the "source charge or dipole."**
- **Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field**

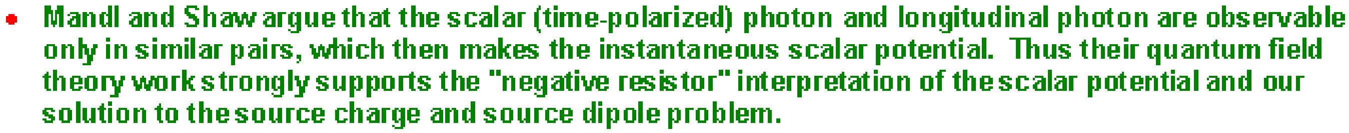

- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus their quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 9. Infolded longitudinal biwaves (time-polarized EM waves) composition of a scalar potential.

EXPLORE! VOLUME 9, NUMBER 3, 1999

Figure 10

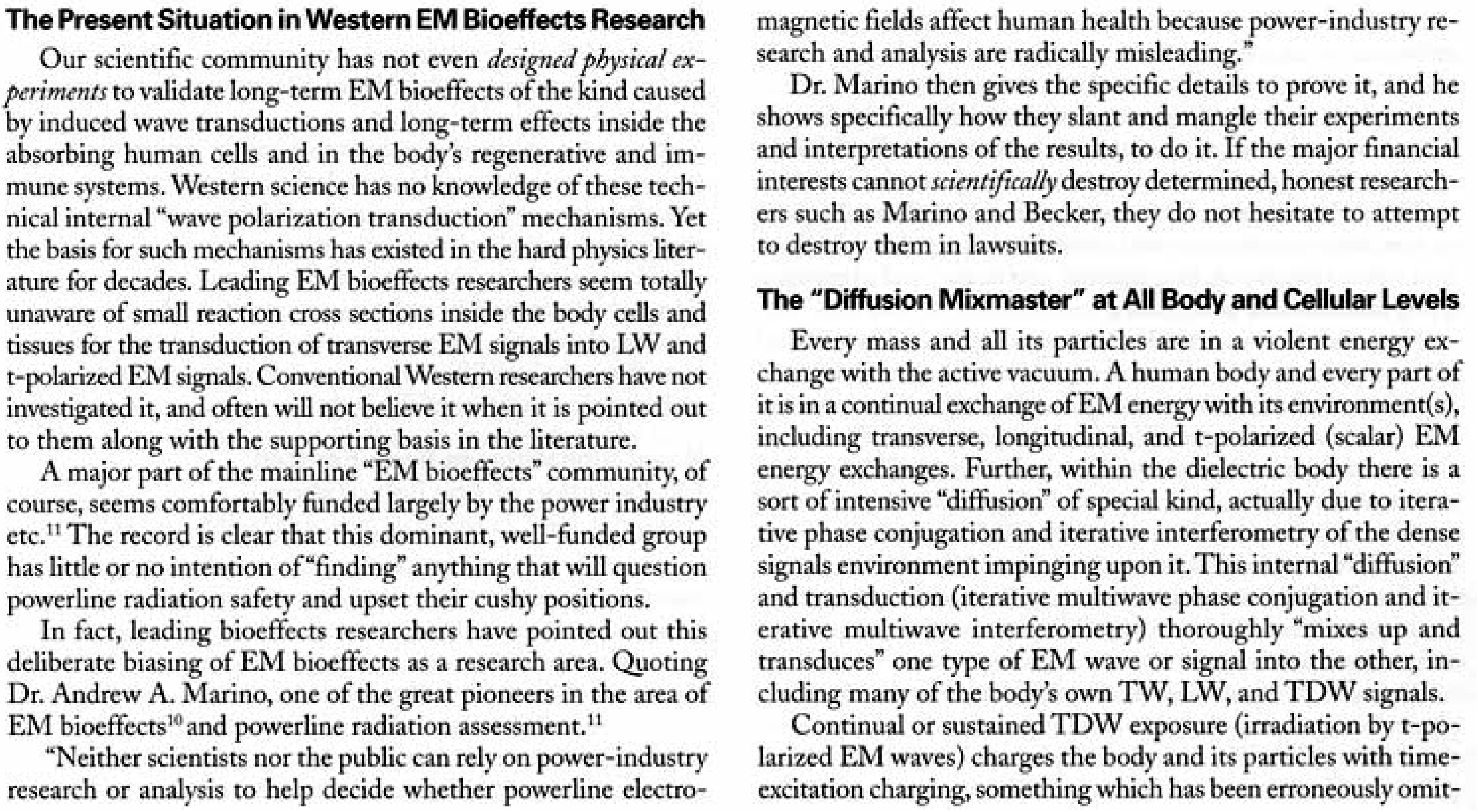

The Present Situation in Western EM Bioeffects Research

Our scientific community has not even *designed physical experiments* to validate long-term EM bioeffects of the kind caused by induced wave transductions and long-term effects inside the absorbing human cells and in the body's regenerative and immune systems. Western science has no knowledge of these technical internal "wave polarization transduction" mechanisms. Yet the basis for such mechanisms has existed in the hard physics literature for decades. Leading EM bioeffects researchers seem totally unaware of small reaction cross sections inside the body cells and tissues for the transduction of transverse EM signals into LW and t-polarized EM signals. Conventional Western researchers have not investigated it, and often will not believe it when it is pointed out to them along with the supporting basis in the literature.

A major part of the mainline "EM bioeffects" community, of course, seems comfortably funded largely by the power industry etc.[11] The record is clear that this dominant, well-funded group has little or no intention of "finding" anything that will question powerline radiation safety and upset their cushy positions.

In fact, leading bioeffects researchers have pointed out this deliberate biasing of EM bioeffects as a research area. Quoting Dr. Andrew A. Marino, one of the great pioneers in the area of EM bioeffects[10] and powerline radiation assessment.[11]

"Neither scientists nor the public can rely on power-industry research or analysis to help decide whether powerline electromagnetic fields affect human health because power-industry research and analysis are radically misleading."

Dr. Marino then gives the specific details to prove it, and he shows specifically how they slant and mangle their experiments and interpretations of the results, to do it. If the major financial interests cannot *scientifically* destroy determined, honest researchers such as Marino and Becker, they do not hesitate to attempt to destroy them in lawsuits.

The "Diffusion Mixmaster" at All Body and Cellular Levels

Every mass and all its particles are in a violent energy exchange with the active vacuum. A human body and every part of it is in a continual exchange of EM energy with its environment(s), including transverse, longitudinal, and t-polarized (scalar) EM energy exchanges. Further, within the dielectric body there is a sort of intensive "diffusion" of special kind, actually due to iterative phase conjugation and iterative interferometry of the dense signals environment impinging upon it. This internal "diffusion" and transduction (iterative multiwave phase conjugation and iterative multiwave interferometry) thoroughly "mixes up and transduces" one type of EM wave or signal into the other, including many of the body's own TW, LW, and TDW signals.

Continual or sustained TDW exposure (irradiation by t-polarized EM waves) charges the body and its particles with time-excitation charging, something which has been erroneously omit-

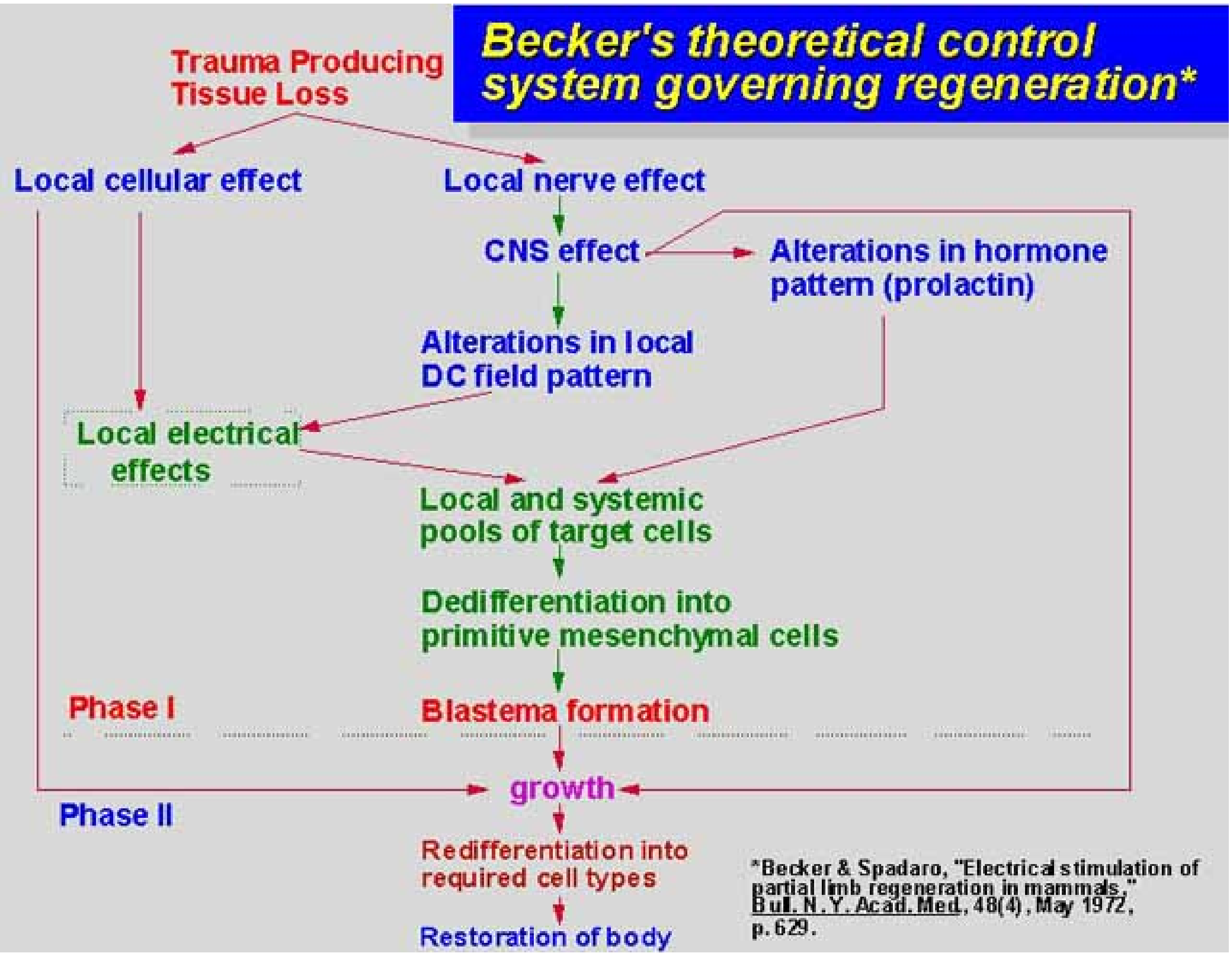
Becker's theoretical control system governing regeneration*
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
growth
Phase II
Redifferentiation into required cell types
Restoration of body

*Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

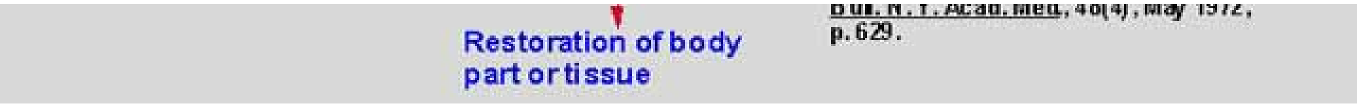
Restoration of body part or tissue
Bull. N. Y. Acad. Med, 48(4), May 1972, p.629.

Figure 10. Becker's model of the Regenerative system of the body.

EXPLORE! VOLUME 9, NUMBER 3, 1999

ted in physics, but is true nonetheless. Inexplicably, physicists just ignore what happens time-wise when a mass absorbs a photon. We have covered that elsewhere.

So inside the body dielectric, there is a "diffusion mix" correlate of any and every type of the waves, to any or all of the exchanges, plus the additional internal interactions of the body in its TW, LW, and scalar EM wave transductions. There is a one-to-one correspondence to any one of the three types of polarization waves (transverse, longitudinal, and scalar) and everything going on anywhere in the body.

Note that the type of LW we are using is actually made of phase conjugate pairs of transverse EM waves, so its "photons" are made of photon/antiphoton pairs. That is, our longitudinal EM wave has an internal TW structure, and the LW wave is comprised of spin-2 *gravitons*. Our longitudinal EM wave has an internal structure, and it dramatically differs from the LW used by the U.S. scientific community which ignores wave, field, and potential infolded substructures. Our LW (which seems to be what the Russians use) is also a gravitational wave because it is comprised of oscillations in its distributions of spin-2 photon/antiphoton pairs (i.e., *gravitons*).

The type of t-polarized (scalar) EM wave we are using is a phase conjugate coupled pair of the specialized electrogravitational (EG) LWs. The "t-polarized photons" comprising this t-polarized "EM" wave are comprised of spin-4 *supergravitons*. These are, I believe, what the KGB energetics weapons scientists use also. By iteration, we can go into as deep an internal multilevel structuring of EM and EG waves, of any type, via methods shown by Whittaker.[12]

Now let us examine a body as a dielectric, sitting there in its exchange, in relative equilibrium. It turns that every tiniest piece of a dielectric participates in any emission from it[13]—even the emission of a single photon, graviton, or supergraviton. Well, it *would* so participate, due to the special sort of correlated diffusion and interferometry and phase conjugation that we pointed out. Anything that escapes (is re-emitted) from the body, has been through this "diffusion wringer". It has escaped from each and every internal part of the body.

Novel Signatures in Body-Emitted EM Radiation of Any Polarization

A single photon or wave emitted from a dielectric carries in its internally structured time component a *template* of the exact internal time structuring of every part of the dielectric emitting that photon. It carries in its internally structured energy component a *template* of the internal energy structuring of every part of the dielectric emitting it. An emitted EM wave, being a collection of emitted photons and photon structures, thus carries the same internal structuring of both its transported energy and its transported time.[14] Western scientists do not use the internal

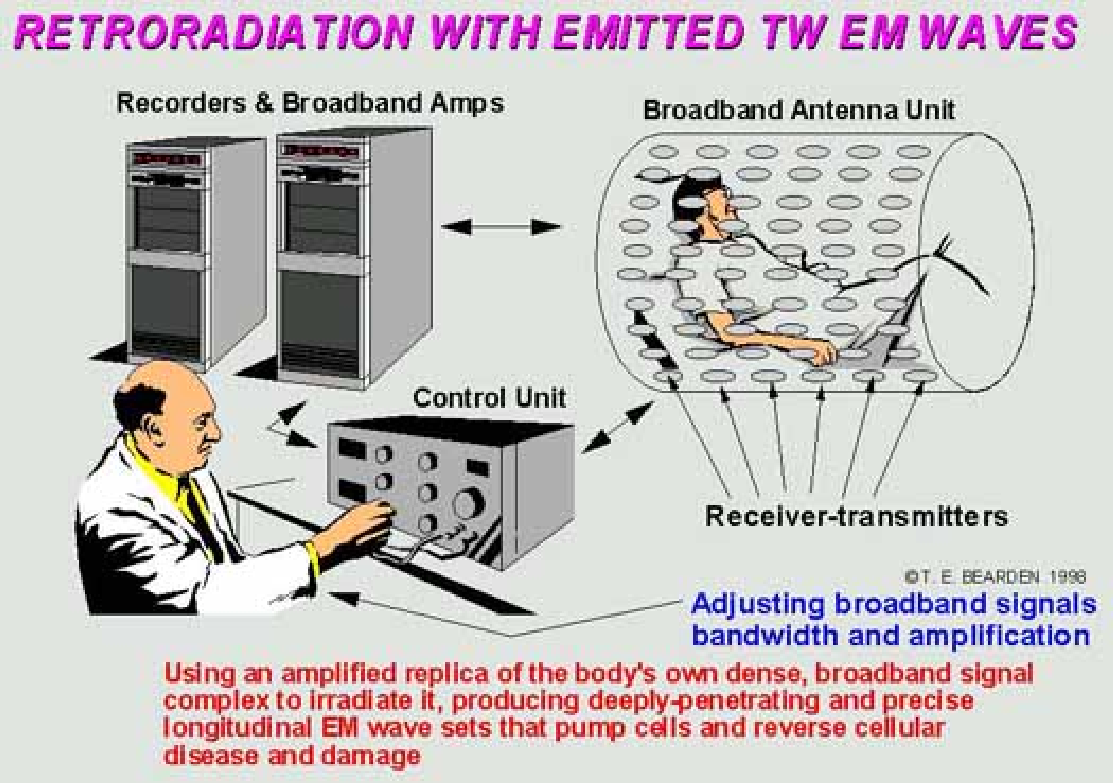
RETRORADIATION WITH EMITTED TW EM WAVES
Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification
Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

Figure 11.

EXPLORE! VOLUME 9, NUMBER 3, 1999

structuring of photons and EM sine waves, have no instruments for that, and so do not know it. The Russians know it,[15,16] use it, and have developed sophisticated instruments for it for the last 30 years or more.

Shortly we will postulate how conventional scientists experimenting with mind and behavior control electromagnetically might utilize an immediate consequence of the fact that the specific emission from a body at any time—whatever type of wave component we look at—is a total correlate to all that body's internal functions, its energy exchanges with its environment, its states, its mental states and conditions, etc.

And we accent that the coherently integrated intent-changes in that body from its coupled mind's operations are also in the substructure of photons and EM waves emitted from the living body into its environment.

How Western Mind Control Scientists Might Proceed

Conventional scientists would ignore scalar photons and waves, determine the TW spectrum, and then slowly and painfully correlate that to the various physical, mental, and emotional things. That's the way they would *have* to do it, since apparently they use only standard electrodynamics. They can in fact do this via brute force, because that emission TW spectrum is of waves escaping from the entire "diffusion machinery" mixmaster in the body dielectric. It does contain all those integrated *mind-intent* 3-space correlates as well as the ordinary body 3-space energy correlates.

Anyhow, in envisioning Western procedures, the test body would be sitting there, in its normal (let's say fairly normal and quiet) EM environment. [See *Figure 11*] Let's assume this body is not sick, but is healthy and normal. Its emissions are a part of its equilibrium condition. It follows that the totality of its TW emissions is correlated to the sum total of everything that interacted with the body, plus its own internal reactions, including from mind-body coupling. Indeed, everything that has ever happened to that entity—physically, mentally, emotionally, etc.—has internal correlates infolded inside the emitted transverse EM wave spectrum.[17]

So first the scientists would simply measure the full environmental TW spectrum down to very tiny levels in all frequency bands possible—everything that is coming in to the body from its environment. They would probably perform the experiments in a triply-shielded Faraday cage, so the external environment outside the cage contributes very little transverse EM wave noise (perhaps some Schumann resonance). Magnetic wave shielding and ELF shielding is something else again, and longitudinal EM wave shielding is in general not possible unless longitudinal EM waves are employed in the shielding. But a "quiet external TW environment" enhances the isolation and measurement of the specific TW signal transforms the Western experimenters would be seeking.

Then they would measure everything likewise that the body is *emitting* when in that environment. That is the "zero reference" or "normal reference" level. They would measure every aspect: frequency, amplitude, phase, wave shape, modulation, etc.

Probable Experimental and Measurement Techniques

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using

They would consider the body dielectric as a "ship with portholes". Light (EM transverse wave energy) coming out of the portholes has originated and adapted from everywhere inside the ship and its operations. When the experimenters reinsert energy through one or more portholes, they are interested in determining precisely which operations inside the ship are affected. So the body physiological functions are also heavily instrumented.

By feeding back in the exact EM pattern and magnitude escaping the body dielectric through a single porthole, one "nullifies" that part of the emission by producing a net equilibrium in that component. This will in turn produce changes elsewhere in the spectrum emission from other portholes, and this can be measured. It will also produce "stress changes" in specific portions of the body dielectric, and in the heavily instrumented body these physiological, chemical, electrical, etc. changes can also be detected. Monitoring of the skin conductance, e.g., indicates the level of stress, and EEG monitoring with sophisticated instrumentation can indicate something of the brain wave states.

In short, the necessary measurements from which to construct specific correlates can be performed, recorded, and later analyzed.

This article will be concluded in Volume 9, #4 with Improving the Methodology, Further Phenomenology Experiments, Rejuvenation and other matters. ◆

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel U.S. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, U.S. Army. Graduate of Guided Missile Staff Officer's Course, U.S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor. He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuum's virtual particle flux is organized and used to energetically shape and manipulate matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

1. *In fact, it simply applies in the laboratory that method that is already used by species for the so-called "directed evolution" component of evolution. The exact mechanisms for directed evolution also arise naturally in the new approach. Directed evolution is required; e.g., no reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by*

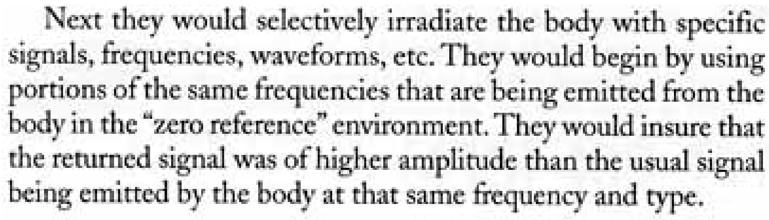

Next they would selectively irradiate the body with specific signals, frequencies, waveforms, etc. They would begin by using portions of the same frequencies that are being emitted from the body in the "zero reference" environment. They would insure that the returned signal was of higher amplitude than the usual signal being emitted by the body at that same frequency and type.

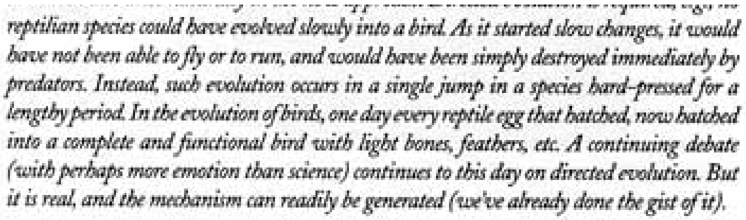

reptilian species could have evolved slowly into a bird. As it started slow changes, it would have not been able to fly or to run, and would have been simply destroyed immediately by predators. Instead, such evolution occurs in a single jump in a species hard-pressed for a lengthy period. In the evolution of birds, one day every reptile egg that hatched, now hatched into a complete and functional bird with light bones, feathers, etc. A continuing debate (with perhaps more emotion than science) continues to this day on directed evolution. But it is real, and the mechanism can readily be generated (we've already done the gist of it).

EXPLORE! VOLUME 9, NUMBER 3, 1999

2. *Fritz Albert Popp, "Photon Storage in Biological Systems," in Fritz Albert* Popp et al, *eds.*, Electromagnetic Bio-Information: Proceedings of the Symposium, Marburg, Federal Republic of Germany, Sep. 5, 1977; *Urban & Schwarzenberg, Baltimore, 1979, p. 123-149;—"On the coherence of ultraweak photoemission from living tissues," in* Disequilibrium and Self-Organization, *ed. by C. W. Kilmister, D. Reidel, Hingham, MA, 1986;*—et al., *"Physical aspects of biophotons,"* Experientia, *Vol. 44, 1988, p. 576-585;—"Biophoton emission: New evidence for coherence and DNA as source,"* Cell Biophys., *Vol. 6, 1984, p. 33-52. See particularly Fritz Albert Popp, Ke-hsueh Li, and Qiao Gu, Eds.*, Recent Advances in Biophoton Research and its Applications, *World Scientific Publishing Co., New York, NY, 1992.*
3. *With novel use of EM time-energy to curve spacetime, the physical actions of the "placebo effect" are due to the use of such "spacetime curvature engines" to produce actual physical effects—at least temporarily —in the cells and throughout the body's chemistry, physiology, etc. Experimental detection of these physical effects are well-established in the orthodox medical literature.*
4. *For example, see particularly R. O. Becker, "The direct current field: A primitive control and communication system related to growth processes,"* Proceedings of the. XVI International. Congress of Zoology, *Washington, D.C., Vol. 3, 1963, p. 179-183;— "The direct current control system: A link between environment and organism,"* New York State Journal of Medicine, *Vol. 62, 1962, p. 1169-1176;—and D. G. Murray, "A method for producing cellular dedifferentiation by means of very small electrical currents,"* Trans. N.Y. Acad. Sci., *Vol. 29, 1967, p. 606-615;—and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals,"* Bulletin of the New York Academy of Medicine, *Second Series, 48(4), May 1972, p. 627-64.*
5. *E.g., a popular discussion is given by John Travis, "Starting Over: Some Animals can regenerate limbs or even most of their bodies. How?"*, Science News, *152(18), Nov. 1, 1997, p. 280-282. For a typical technical paper, see Jeremy P. Brockes, "Amphibian limb regeneration: Rebuilding a complex structure,"* Science, *276(5309), Apr. 4, 1997, p. 81-87.*
6. *I have one viewgraph which lists those 32 major flaws in electrodynamics. Further, leading theorists (Evans, Crowell, and others) in the Alpha Foundation's Institute for Advanced Study (AIAS)—of which I am a Fellow Emeritus by AIAS's kind tolerance—have added about another very technical and very fundamental dozen or so errors to my own list. A leading foundations scientist with over 500 papers in the literature commented dryly to the AIAS that "I broadly agree. that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science. Faraday visualized that "EM disturbances" surrounding wires, magnets, etc. were simply the transverse plucking of taut strings. As he states in his book, Maxwell—who also firmly believed in the material ether—simply captured Faraday's thinking and assumed that transverse "plucked string" wave disturbance in the material ether. At the time all this was expressed, the electron had not been discovered, the atom had not been discovered, the molecule was a featureless blob, and no one had the slightest notion of an atomic nucleus filled with protons and neutrons. Charge was simply a "certain volume of electric fluid" and everything electromagnetic was expressed in fluid dynamics. As an example, that "material fluid" and "material ether" theory was destroyed in the late 1880s by the Michelson-Morley experiment. All the Maxwell equations assume that material ether to this day; not a single equation was ever changed to eliminate the material ether assumptions during the last 110 years.*
7. *For full explanations of the "Moscow Signals," see T. E. Bearden, .,* Gravitobiology: A New Biophysics, *Tesla Book Co., P.O. Box 121873, Chula Vista, CA 91912, 1991;—* AIDS: Biological Warfare, *Tesla Book Co., 1988. For an excellent Government study of the Embassy radiation by excellent but orthodox Johns Hopkins scientists, see A. M. Lilienfeld* et al., Foreign Service Health Status Study: Evaluation of Health Status of Foreign Service and Other Employees from Selected Eastern European Posts, *Department of Epidemiology, of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland, Final Report, July 31, 1978. On p. 245, it is noted that a marked excess of symptoms were noted in the Moscow employee group as compared to other East European employees in other East European nations.*
8. *Johns Hopkins researchers established the exact field patterns present in the Embassy from the microwave carrier radiation; see Robert C, Mallaliew,* A Model of the Microwave Intensity Distribution within the U.S. Embassy, *Moscow 1966 to 1977, Report FS-80-166, Applied Physics Laboratory, Johns Hopkins University, Aug. 1980. Nowhere in any of the Government-funded studies is there any consideration of field-free scalar potentials and their internal time-polarized EM wave structures (the bidirectional phase conjugate wavepairs shown by Whittaker to comprise the scalar potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.*

issue, see "A Comment" by William A. McGarey, p. 143; "A Comment" by Arthur C. Hastings, p. 143-144. See also Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," Psychoenergetic Systems, *1(1), Dec. 1974. The Soviet Union awarded Kaznacheyev a very high medal for his achievements.*

10. *For a very nice summary of EM bioeffects from the more conventional view, see Robert O. Becker and Andrew A. Marino,* Electromagnetism and Life, *State University of New York Press, Albany, 1982. The book is also available at http://www.ortho.lsumc.edu/Faculty/Marino/El/ElPub.html.*
11. *Andrew A. Marino,* Powerline Electromagnetic Fields and Human Health, *http://www.ortho.lsumc.edu/Faculty/Marino/Marino.html. This book is highly recommended. Particularly see "Chapter 5, Blue-Ribbon Committees and Powerline EMF Health Hazards," and "Chapter 6: Power-Industry Science and Powerline EMF Health Hazards." Biophysicist Marino is one of the leaders in the field and has been personally involved in many skirmishes with powerline-dominated studies and findings.*
12. *E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355; — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372.*
13. *Reali, G. C., "Reflection from dielectric materials,"* American Journal of Physics, *50(12), Dec. 1982, p. 1133-1136. The reflected field from a dielectric material is not generated just at its surface but comes from everywhere in the interior of it.*
14. *Present physics models the EM wave as transporting energy, but erroneously omits that it also transports time. A photon is made of angular momentum, or "energy x time". Since EM waves are made of photons, it follows that they transport both the energy components and the time components making up those photons. An EM wave is a wave of angular momentum, not just a wave of energy.*
15. *KGB energetics weapons scientists call this internal structuring the "information content of the field." However, the Russians use a good deception plan, which implies that the "information content of the field" is just the normal transverse EM spectral content of EM signals. Appreciable "pertinent" work using that phrase is produced by the open Russian scientific community, thus confusing the issue. Our electrodynamicists seem to buy it hook, line, and sinker. The deception plan also tricks Western scientists into thinking that psychoenergetics is nothing but a Russian version of Western parapsychology. Again, our scientists comfortably accept this with smug satisfaction.*
16. *Presently the Russians are in a great new version of this deception plan, with the goal of deceiving Western scientists into thinking that the "secret" Russian energetics is based on torsion fields. Consequently they are releasing substantial material on torsion wave theory, including replicable experimental results. The released version of torsion fields has one glaring omission: the time components and structuring. This deception plan is excellent, because the torsion field theory that is released, is indeed "real" and testable. It is also revolutionary, with respect to present Western science. It almost positively assures diverting the attention of much of the Western science community into this highly limited subset of the full Russian torsion theory—effectively diverting the West away from psychoenergetics and the direct use of t-polarized EM waves and fields, torsion or otherwise.*
17. *We point out that E. T. Whittaker, in "On the Partial Differential Equations of Mathematical Physics,"* Mathematische Annalen, *Vol. 57, 1903, p. 333-355, showed that every scalar potential is comprised of an internal structure of bidirectional longitudinal EM phase conjugate wavepairs in an infinite set of harmonics. Each of those phase conjugate longitudinal EM wavepairs comprises a time-polarized EM wave. So the common potential contains an internal longitudinal EM wave structure that is further organized into an internal time-polarized EM wave structure. The mind-intent correlates and the total mind and emotional experience correlates from fertilization of the egg that led to development of that person, is in the internal substructure of every potential in the living body. In his "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions,"* Proc. Lond. Math. Soc., *Series 2, Vol. 1, 1904, p. 367-372, Whittaker also showed that any EM field or wave is comprised of two scalar potential functions. It follows from Whittaker and Reali that the EM emissions from the living body carry the internal correlates of everything that has been experienced by that person from conception in the womb. Sometimes this deep embedded "hidden information content" can even be accessed in deep hypnosis. I once observed the deep hypnotic regression of a subject back to her prenatal experiences prior to birth, where the subject relived hearing words of rejection by the father, before the subject was born. The occurrence of the incident and the father's words were directly confirmed on the spot by the mother, who was closely observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological*

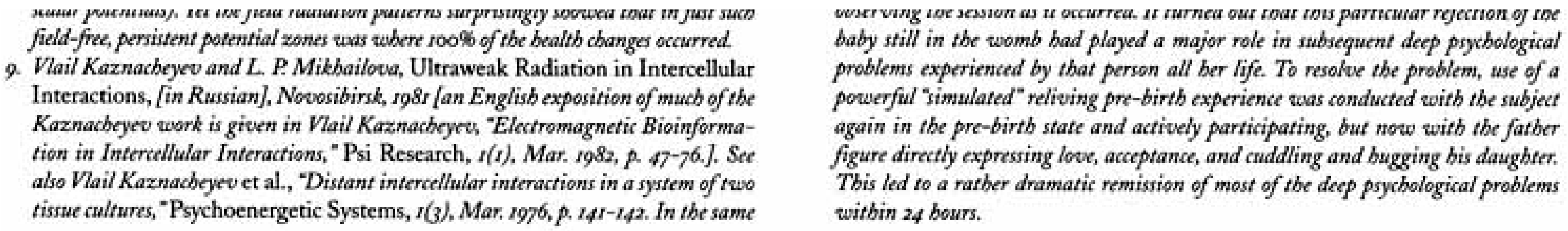

ular potentials). Yet the field radiation patterns surprisingly showed that in just such field-free, persistent potential zones was where 100% of the health changes occurred.

9. *Vlail Kaznacheyev and L. P. Mikhailova,* Ultraweak Radiation in Intercellular Interactions, *[in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions,"* Psi Research, *1(1), Mar. 1982, p. 47-76.]. See also Vlail Kaznacheyev* et al., *"Distant intercellular interactions in a system of two tissue cultures,"* Psychoenergetic Systems, *1(3), Mar. 1976, p. 141-142. In the same*

observing the session as it occurred. It turned out that this particular rejection of the baby still in the womb had played a major role in subsequent deep psychological problems experienced by that person all her life. To resolve the problem, use of a powerful "simulated" reliving pre-birth experience was conducted with the subject again in the pre-birth state and actively participating, but now with the father figure directly expressing love, acceptance, and cuddling and hugging his daughter. This led to a rather dramatic remission of most of the deep psychological problems within 24 hours.

EXPLORE! VOLUME 9, NUMBER 3, 1999

Mind Control and EM Wave Polarization Transductions, Part 3 of 3

©Copyright 1999 by T.E. Bearden, USA

This outstanding article is the last, Part 3, Part I and 2 of which were published in Volume 9, #2 and #3 of Explore! for the Professional. Because we feel that Col. Bearden's Special Note below is so very important we are repeating its publication in all 3 parts of this article.

Special Note

This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation

achieving many more effects in the targeted bio-organism or groups of them. With Hunt's method, the system as an *open dissipative system far from thermodynamic equilibrium in its environment* is being measured as such. This more directly allows for determination of the environmental inputs on a much broader range of subsystems, both space-like and time-like. Particularly for mental and emotional correlates to environmental stimuli, such measurements and correlations are essential.

More Comprehensive Phenomenology Experiments

To return: Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.

In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion.

of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral.

Two Ways to Improve the Methodology and Experimental Correlates

We diverge momentarily: If they wish to improve this methodology, they will consider the body dielectric as a nonlinear isotropic medium. Then they will pretend that the spectrum of *difference frequencies* between each two adjacent waves is the spectrum of "emissions" from the body. They would also use those, as if they were actual emitted spectral signals, instead of the actual signals emitted. That procedure will give superior correlations, but it is more complicated. Part of that is known for underwater sonar,[1] and so our own scientists may or may not have realized it for use in mind and behavior correlate determination with respect to the body dielectric.

Another dramatic improvement in the depth to which the correlates can be constructed, is provided by detecting each porthole output in a two-channel device, where one channel has an adjustable time-delay as utilized by Dr. Valerie Hunt.[2] The two channels—one real-time and one slightly delayed—are then mixed, as by Hunt's method, producing an instrumental measurement of the chaotic part of the functions. Again, this can be measured, recorded, and later analyzed. It adds (i) a completely different kind of thermodynamics in the system measurements, (ii) a dramatically extended set of correlates, (iii) more direct measurement of the time-polarization causative signals, and (iv) an applications Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions.

A particularly vulnerable aspect of every mammal is the pleasure center in the brain. Experiments have shown that, when this center is stimulated, it is the most addictive experience possible. Rats will forego food, endure electrical shocks, starve, and even die to obtain stimulation of this center. Obviously a device capable of generating signals that evoke direct stimulation of the pleasure center in humans would be a powerful weapon.[3]

Extending the Data Base

See Figure 12. After the researchers have made an "individual porthole signal insertion" correlate database, then they would make a "two, three, four," porthole signal insertion correlate database by inducing selected multiple changes at once. Again they would analyze everything; the program would need one or more supercomputers. It would also require quite a highly qualified multidisciplinary team, extensive facilities, substantial funding, and would be many years in duration. We are describing a decades-long program. But it's a doable.

With sufficient development and performance of such a program, the researchers would have produced (with sufficient testing and analysis, and sufficient retest validation and verification) a database of "specific correlates for a given overall technology eventually having many additional degrees of freedom and therefore capable of

desired physical, mental, or physical and

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

mental delta."

A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment.

The Dielectric Body as a Ship With Portholes

Again let us review what the EM transverse wave irradiation from the body dielectric represents. Think of the body dielectric as a sort of "ship", with lots of portholes. When it gets irradiated, certain operations happen internally to that radiation, i.e., certain "processes" and "interactions" occur, which change both the mental and physical states. The output emissions back from the body dielectric, out through the filtering "portholes", will directly correlate to what has happened inside the ship.

It is terribly important to realize that one can use the same frequencies and spectra coming back out of those "portholes" to insert EM signals and waves etc. back in there, and into the deepest processes going on in the body and in the mind (at all levels, times. So it will reach directly into and affect any and every part of the dielectric, as you wish, and in whatever manner you have correlates for.

We filed an invention disclosure[4] on the above correlate and "back-porthole" process for use in directly inducing full and efficient conversion directly to time-density (scalar, or time-polarized) EM waves inside the body/ cells, *therefore immediately and directly time-excitation charging the body cells and all their internal components.*

Subtracting the Normal Correlates to Obtain the Deltas

A specifically diseased body, e.g., has a specifically tailored change in its "normal" emission spectra, because of the disease condition alteration of its resident spacetime curvature engine. By having a database for the normal emission spectra of the healthy body, one can subtract it from the actual emitted spectra, to obtain the "delta spectrum".

If the correlates have been previously determined by the above mentioned research program, then one may simply introduce an amplified "delta spectrum" into the body, i.e., through the portholes. This will immediately time-excite the cells and their components throughout the body, particularly where that disease does exist. It turns out that the time-excitation charging by that method will "pump" those diseased including the deep unconscious).

Remember, whatever you put back in there, is going back through the "mixmaster" hopper. It will be iteratively phase conjugated and interfered with by everything, countless cells in the time domain. The cells themselves (and all their parts) are highly nonlinear—to time-excitation charging, they are *totally* nonlinear. So they will act as rather perfect pumped phase conjugate mirrors, but pumped in the time-energy domain rather than in the spatial

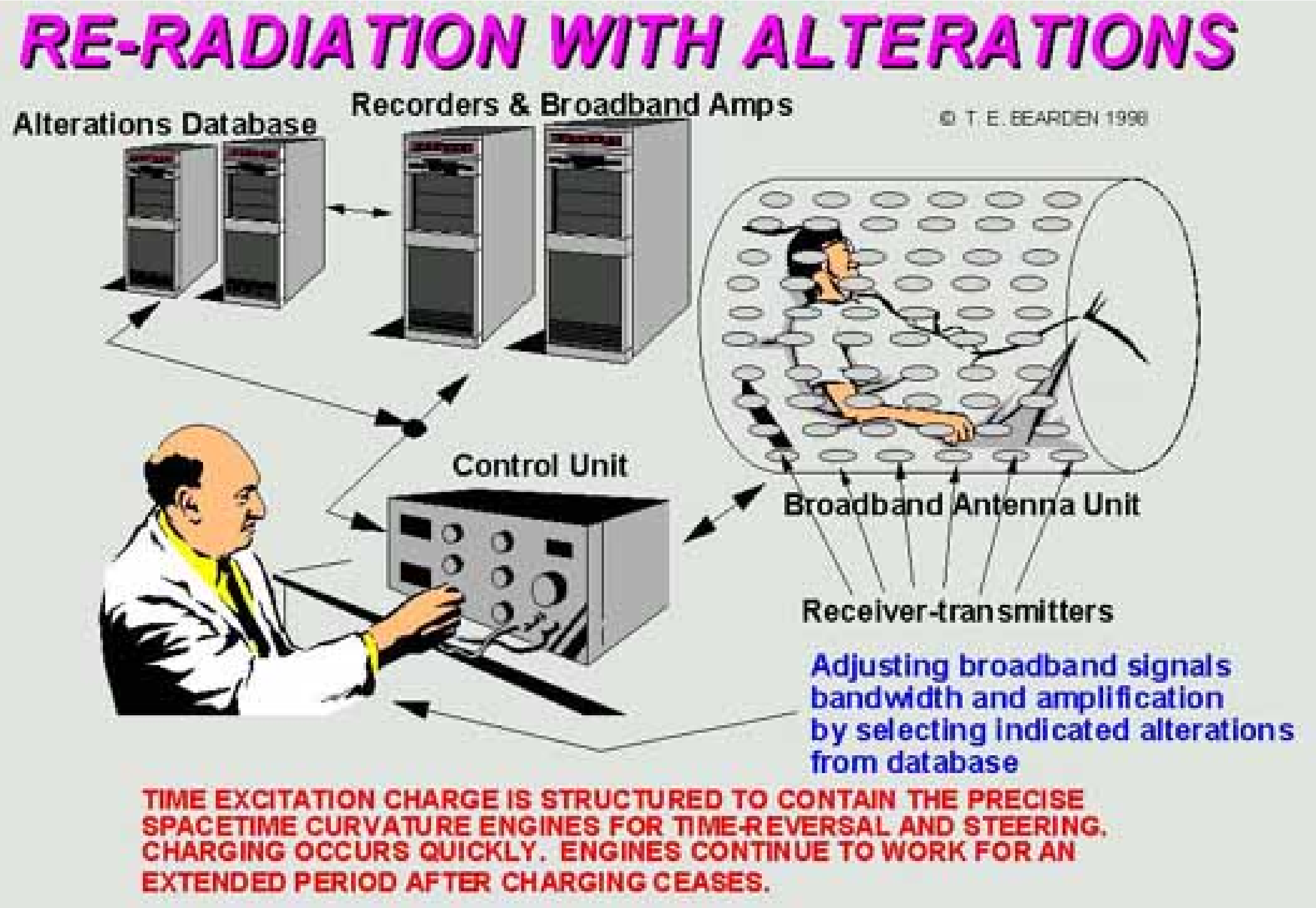
RE-RADIATION WITH ALTERATIONS
Alterations Database
Recorders & Broadband Amps
© T. E. BEARDEN 1998
Control Unit
Broadband Antenna Unit
Receiver-transmitters
Adjusting broadband signals bandwidth and amplification by selecting indicated alterations from database
TIME EXCITATION CHARGE IS STRUCTURED TO CONTAIN THE PRECISE SPACETIME CURVATURE ENGINES FOR TIME-REVERSAL AND STEERING. CHARGING OCCURS QUICKLY. ENGINES CONTINUE TO WORK FOR AN EXTENDED PERIOD AFTER CHARGING CEASES.

Figure 12. Using an amplified replica of the body's own dense, broadband signal complex, containing added alterations, to irradiate the body and produce deeply-penetrating precise time-density charging for cellular time-reversal while steering the time-reversal trajectory to also correct a congenital condition.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

energy domain. This is a dramatic extension to nonlinear phase conjugate optics theory.[5]

General relativity tells us that each specific detail of the mass and energy structure of each cell and all its parts—down to the finest details—is accompanied by a precise set of spacetime curvatures. By reintroducing the same energy as the disease delta TW waves, the affected cells add the phase conjugates, converting the inserted TW signals into longitudinal EM wave signals. The longitudinally-pumped cells again phase conjugate, converting these longitudinal EM waves to time-polarized EM waves, and "charging" the cells and all their parts in the time domain.

Suddenly, at that point the *resident spacetime curvature engine*—rather than TW or LW EM waves-serves as the "input signal" to the cell-and-its-parts as pumped phase conjugate mirrors (PCMs), pumped in the time domain. An extension to the distortion correction theorem of nonlinear optics results. An exact, amplified antiengine for the diseased cells—i.e., a phase conjugate replica of the resident spacetime curvature engine—is formed. This amplified antiengine overpowers the resident disease engine, powerfully time-reversing the cells and their components back to a previous healthy state. The time-reversal actually occurs over a period of time after the "time excitation charge-up" ceases, say over the next week or two. Usually two to three quick irradiations are required, one week apart, to give a time-reversing period of three to four weeks.

force upon the cells or any part. E.g., as Pepper[6] states,

"On a more fundamental level, the ideal lossless PCM reverses all the quantum numbers of the incident photon (i. e., linear momentum, angular momentum, etc.). It can be shown that the PCM therefore experiences no linear or angular momentum transfer from the incident photons; hence, the PCM is free from photon radiation pressure and torques."

Note that, in ordinary EM theory, it is precisely translation forces that are engendered upon charged mass by EM fields. This process is quite different. In biological terms, this process dedifferentiates (in physics: time-reverses) the diseased cells and all their components—including the genes—back to an earlier healthy state.

By taking a diseased body and feeding back into it an exact amplified replica of its emitted TW spectrum, one can force the body to "time-excitation charge" all cells and their components immediately, with high efficiency. The human body and every cell and every part of every cell does the necessary "calculation" for the exact antiengine for each part. Since time is extremely condensed spatial EM energy (by a factor of almost 10^{17}), a startling contraction occurs in the radiation time required to time-charge the cells and their components and charged particles. This is a method for instant initiation of time-charging, hence it is extremely rapid. What took Priore several hours of LW irradiation to accomplish in the body,[7] can now be done

Even though appreciable forces are operating internally upon the cells and all their parts, there is no net translation

in seconds with this "thwarting" and spectrum transduction method. In fact, it is an intense stress on the body, and—say—should only

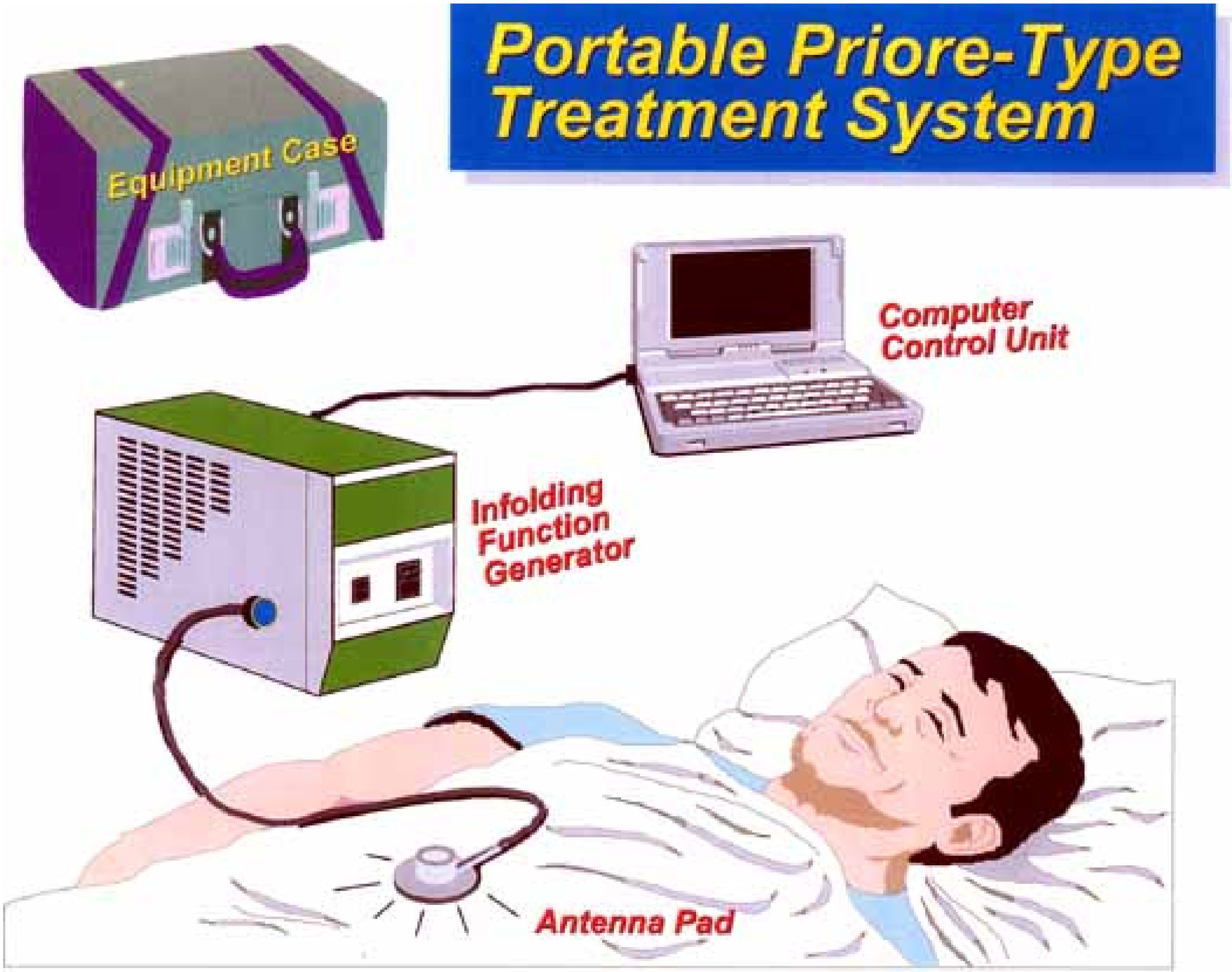
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

Figure 13. Portable first-method treatment unit proposed to the Department of Defense. For treating and curing mass casualties from terrorist BW attacks on U.S. population centers.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

be continued for 30 seconds or so. We highly stress that overcharging in this time domain can be detrimental or even lethal if prolonged.

At the end of the very short time-excitation charging time, the body is "time-excitation charged in all its minutest particles". So everything is therefore pumped in the time domain for the next several days after the irradiation is removed, since the time-charge decays very gradually over a period of time. During that decay period, every part of the cell and body acts as a pumped phase conjugate mirror material, pumped in the time domain rather than in the 3-space energy domain. When time-pumped, the mass itself reverses to a previous physical state. This can be used for both nonliving and living masses. Because there is no forcible translation manipulation of the cells and their parts, it can be utilized on living cells in vivo. In the case of diseased cells—e.g., AIDS-it appears that three precisely determined 30-second "AIDS correlate" irradiations a week apart, should be sufficient to completely cure AIDS, including removal of all HIV-genetics changes from every previously infected cell of the body. Also, it would accomplish considerable rejuvenation of the body (making the body younger).

Rejuvenation: Restore the Telomeres and Produce Immortal Cells

Saving Most of the Casualties from Mass Terrorist Attacks

To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S. Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines could be developed cheaply and quickly, and present government facilities could be used to massively attack the correlate database production problem. The portable treatment units could be cranked out by the hundreds of thousands and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc.

Each device could be used in "assembly line" fashion upon hundreds of sickened patients, in case of a WMD (weapons of mass destruction) attack upon one or more of our cities. The first generation equipment would save probably 70% of the stricken (as from anthrax, etc.). With second generation equipment, one would expect 90% of the stricken casualties could be saved.

Just now, there is a very dim prognosis for all those casualties, using present available treatment capabilities. Time and lack of vaccines defeats just about everything that can be done. Brutally speaking, most of those unfortunate enough to breath in

In fact, it is now known that the gradual loss of the telomeres—the natural ends of chromosomes—in a cell's chromosomes is what causes the cell to age. The gradual loss of the telomeres and the resulting shortening of the chromosomes represents a cellular "disease state" or "disorder" state" accompanied by a precise delta in the spacetime curvature engine accompanying every affected (aging) cell.

By determining the correlate for that precise telomere loss in the aging body's emitted radiation spectrum, an effective telomere regeneration process immediately emerges. One simply amplifies that aging "precise telomere reduction" correlate and reradiates the body with it, passing an amplified correlate back into every cell of the body. The cells and all their parts—including the chromosomes—add *the phase conjugate replicas*, converting the incoming TW EM correlates to longitudinal EM wave correlates. The cells again phase conjugate the LWs and add *the phase conjugate replicas*, converting the LW correlates to time-polarized EM wave pumping correlates. This latter process *time-excitation-charges* the aging cells and all their parts, including the genes and chromosomes. The cellular matter itself produces the amplified "aging antiengine" for each and every cell and telomere ending. The result is to again add telomeres, converting the cell back to a younger, more vigorous cell, and reversing the process of aging.

We point out without further elaboration that these methods also introduce direct engineering signals into the body's master cellular control system, including the regenerative subsystem. Hence the body's regenerative subsystem is involved in helping produce the necessary engines and anthrax spores from an attack, are going to die. At least 80% of them will expire, in spite of everything the present system can do for them.

Note each device could rapidly treat a whole series patients, one right after the other, merely laying them between "blankets" containing wire antennas, irradiating with the proper correlate spectrum quickly, moving on to the next patient, irradiating quickly (seconds), etc. The entire apparatus could be made to fit in a large suitcase-sized container, and it would be highly portable and controlled by a laptop-sized computer.

Unfortunately, so far the DOD doesn't even understand what the heck we're talking about! The National Institute of Health (NIH) and the National Science Foundation (NSF) are not interested at all. After all, it isn't drugs, vaccines, antibiotics, etc. —which are simply not going to do the job. However, those are what the government and scientific communities are going to finance to the tune of hundreds of millions of dollars. They are then going to lose almost all those mass casualties when the terrorist attacks on our population centers inevitably occur.

Those are Americans. They are going to die. Most of them could be saved. We ought to save them.

For any change in state (physical, chemical, emotional, mental, perceptual), the body dielectric "portholes" change and the output spectrum changes. The delta from zero reference now represents that exact change.

Where Western Mind Control Research Has Probably Arrived

Western mind control researchers

operate upon the short telomere chains to add back telomeres. More specifically, the regenerative subsystem is changed into perfect consonance with the external introduction of the amplified antiengines.

That exact process can be used to rejuvenate the aged population and remedy most of the debilities of growing old. That is how we eventually intend to do it, if we can convince the U.S. government and the U.S. scientific community to perform the necessary research and validation.

probably first established a correlation database of porthole emission correlates to changes in internal states, both physical and mental. They almost certainly recognize the peculiar participation of every part of the body dielectric in the TW emission spectrum from the body. They will almost certainly have opted for this "porthole" notion of some similar expression of it, and the "thwarting" of these "relaxation and release" emission processes to engender internal physical changes in the body and in the brain-meat-computer, as they see it.

EXPLORE! VOLUME 9, NUMBERS 4 & 5, 1999

They probably experimented with different TW EM radiation mixes, to see what the body emission change to, and what states are induced. Undoubtedly they found that the induced states began to correlate with the former emission delta data base. That is, they found out which delta emissions spectra, reversed and amplified and fed back thorough the portholes, produce what internal changes.

Use for healing is the simplest of all. Use for killing or debilitating is more complicated. Of course weapons researchers wish to look only at the killing aspects. That is something they understand. They do not understand healing, for they do not even have the "regenerative" or "healing" mechanism in any of the scientific papers or texts. Becker[8] came closest in this country, and Priore did it in France. Both of those scientists were suppressed and paid a bitter price for their revolutionary and innovative research.

But look at what has been done, by "input-output" correlation research. The scientists will have a database now, that tells what state emerges in the body and in the mind, for what irradiation spectra (frequency, relationships, phases, amplitudes, wave shapes, etc.) one uses. And they did not work out all the wave-to-wave transductions ongoing in the mind and body connection and in the body and in the mind, individually, between TWÛLWÛTDW.

We suspect that this is the course Western mind control researchers will have taken, without even realizing the internal wave transductions occurring. And they would not be concerned with that. They would have what they want, for weapons, mind control, and behavior control. And they of the internal "mix" of the dielectric and all its tiniest portions. So Western mind control researchers and rogue groups probably do get good results.

We are not just picking on the U.S. government here! Internal divisive groups exist in all large and powerful groups, and joust for power. In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old "Power corrupts, and absolute power corrupts absolutely" routine.

The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.

In a nutshell, that's what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, *it could account for what seems to be actually occurring*. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the would of course ignore the healing aspects. And they would get results of course, revolutionary results. So they would undoubtedly classify the program very highly, thinking that they had the best of the best.

Rogue Groups May Also Have Arisen to Divert the Purpose of the Projects

And that is where real problems may arise. Every human group has within it competing subgroups, jousting for power. The more highly classified the group and its activities, and the less that is known about it outside its confines, the more strongly and secretly the subgroups can play. So they can emerge as **rogue groups** pursuing their own agendas, not that of their duly elected governments. In the extreme, such a deep black program can even become a "captured" program, which is totally in the hands of rogues and no longer reports to or is bound by the dictates of the parent government.

It has long been suspected in some quarters that even the US's highly classified research community may be riddled by such rogue groups, and so may be certain highly sensitive parts of the Research and Development community.

Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A world." They may actually believe they are ahead of the Russians.

Where the Russians Probably Have Arrived

On the other hand, the Russians will have developed instruments for detecting and isolating the LWs and the scalar waves, in addition to the ordinary mundane instruments for measuring the TWs. So they will have not only the TW correlates, but also the much finer TWÛLWÛTDW (scalar EM wave) correlates. They will understand and employ the use of vacuum engines (spacetime curvature engines) which Western researchers will often invoke but not understand or even be aware that this is what is being invoked. In other words, Western researchers probably just use the old method of "input fi set of functions fi output". That is the standard basis for almost all Western system analysis. With it, one doesn't have to know the exact functions and processes going on in the middle. One just determines a transfer function from a given input to the output that results. Then one catalogs that transfer function. One does that for lots and lots of transfer functions for particularly desirable results, and one has the system analysis. One also has the immediate basis for a rapidly-developed applications technology.

Western scientists are good system analysts. But they have little or no comprehension of what is really going on in that "inside the ship functions" portion, even though they have determined the transform function which acts upon the input to give the "from the portholes" output.

The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced certain percentage of a highly secret rogue group will wind up using all these things and more. It's a **human** characteristic, the old primate dominance game. Only now disguised and hidden under deep classification.

Nonetheless, the "simplified" TW EM "porthole correlate" approach will yield very positive results, due to the correlation existing with the LW s and scalar portions than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW interferometers, through intervening mass

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

including the earth and ocean. Western mind-control scientists and their rogue groups—assuming they exist as postulated—largely will not be able to do that, and probably won't even recognize Russian testing of psychoenergetics weapons directly in or over Western nations.

Russian Lethal Psychoenergetics Mind Control Tests Over the U. S.

In fact, the KGB tested just such highly advanced mind control weapons over the middle of the United States in 1997 on two occasions. The tests were conducted against Captain Button in his A-10 "Warthog" aircraft on April 2, 1997 [See Figure 14], and upon Captain Svoboda in her A-10 "Warthog" on May 27, 1997. [See Figure 15].

Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona. At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—dream-thinking until he died that everything was normal. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.

Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks. A secondary purpose was to stimulate the U.S. government, watch its investigation and findings, and ascertain if the U.S. knew of the extent to which Russian KGB mind control weaponry had advanced. Obligingly, we proved to me KGB that we did not understand what had actually happened to Captain Button.

Captain Svoboda was from the same Wing as Captain Button and was its most experienced night fighter A-10 pilot. In fact, she taught other pilots the fine points of night flying combat missions in the A-10. About two months after Captain Button's bizarre flight and death, Svoboda was engaged in night target bombing on the range. Rising from her ordnance drop on me target, Svoboda was suddenly struck and rendered hypnogogic and in a waking-dream state. Her sense of the vertical was instantly reversed. Superb pilot that she was, she immediately dream-perceived she was diving, not climbing out of the delivery. She instantly corrected to "climb out". So in the real world she dived sharply into the ground and her aircraft exploded.

This test demonstrated the instant alteration of a skilled pilot's perception, causing that pilot to inadvertently react promptly in an unintentionally suicidal manner. A secondary purpose of the test was to stimulate the U.S. government system again, to see if the system recognized what had

A-10 Mysterious Flyaway and Crash

April 2, 1997

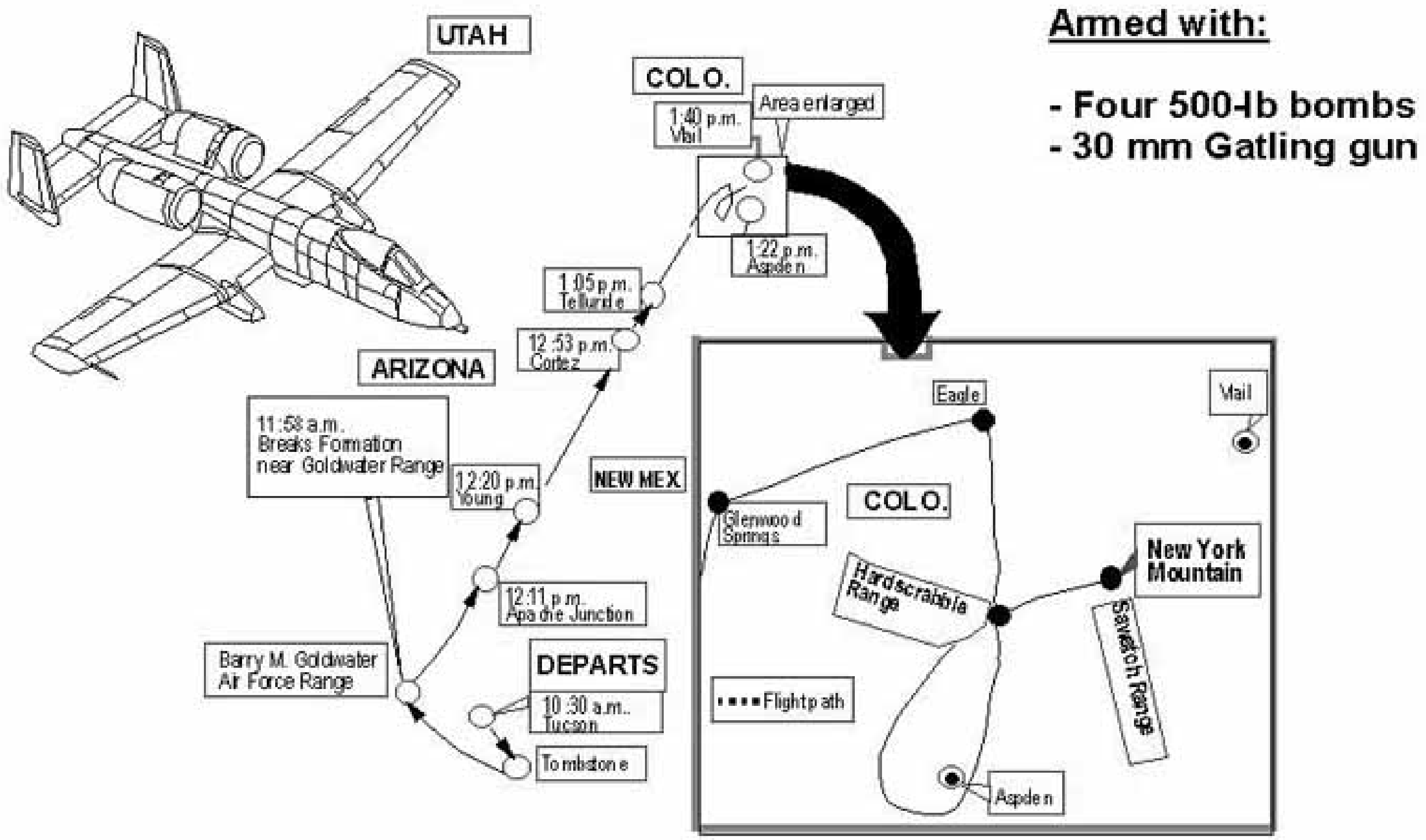
UTAH
COLO.
1:40 p.m.
Vail
Area enlarged
1:22 p.m.
Aspden
1:05 p.m.
Telluride
12:53 p.m.
Cortez
ARIZONA
11:58 a.m.
Breaks Formation
near Goldwater Range
12:20 p.m.
Young
NEW MEX.
12:11 p.m.
Apache Junction
Barry M. Goldwater
Air Force Range
DEPARTS
10:30 a.m.
Tucson
Tombstone
Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun
Eagle
Vail
Glenwood
Springs
COLO.
Hardscrabbla
Range
New York
Mountain
Savatch Range
Flightpath
Aspden

Figure 14a. Captain Button's mysterious flyaway and eventual crash.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

occurred. Sad to say, it did not and it has not. It was lamely speculated that Svoboda probably looked over her shoulder like a rank greenhorn, and inadvertently dived her aircraft into the ground while looking back at the target. If there was one pilot in the USAF who would have not done that, it was Captain Svoboda. At any rate, the KGB determined that the U.S. scientific, intelligence, and governmental communities did not have the foggiest notion of what had occurred.

In private proprietary communications[9] we have explained further, to proper parties, exactly the reasons for these two tests, and what exactly was being tested, and what is its planned strategic use.

The Final Race for Direct Mind Control of the Entire Human Species

The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.

The KGB scientists also know that something very like Gaia— a collective unconscious mind for all species on earth—also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials. If one reads some later books by Kaznacheyev,[10] e.g , this human species unconscious mind aspect and the biosphere unconscious mind aspect emerges loud and clear, if one interpolates between the lines, so to speak. Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society"—communist style—is still a variation of feudalism (all systems tend to feudalism or some form thereof).

The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now. It is simply a matter of developing sufficiently complex equipment to be able to directly engineer the time-domain substructuring at sufficiently deep levels. It is a doable, although it is very, very complex and very, very difficult. On the other hand, they have been at it for decades. Even in the 50s the Russians already had the old LIDA device,[ll] capable of inducing a catatonic trance-like state in a mammal—either a human or a cat,[12] for example. Some versions of Russian mind control devices were used on Russian soldiers in the

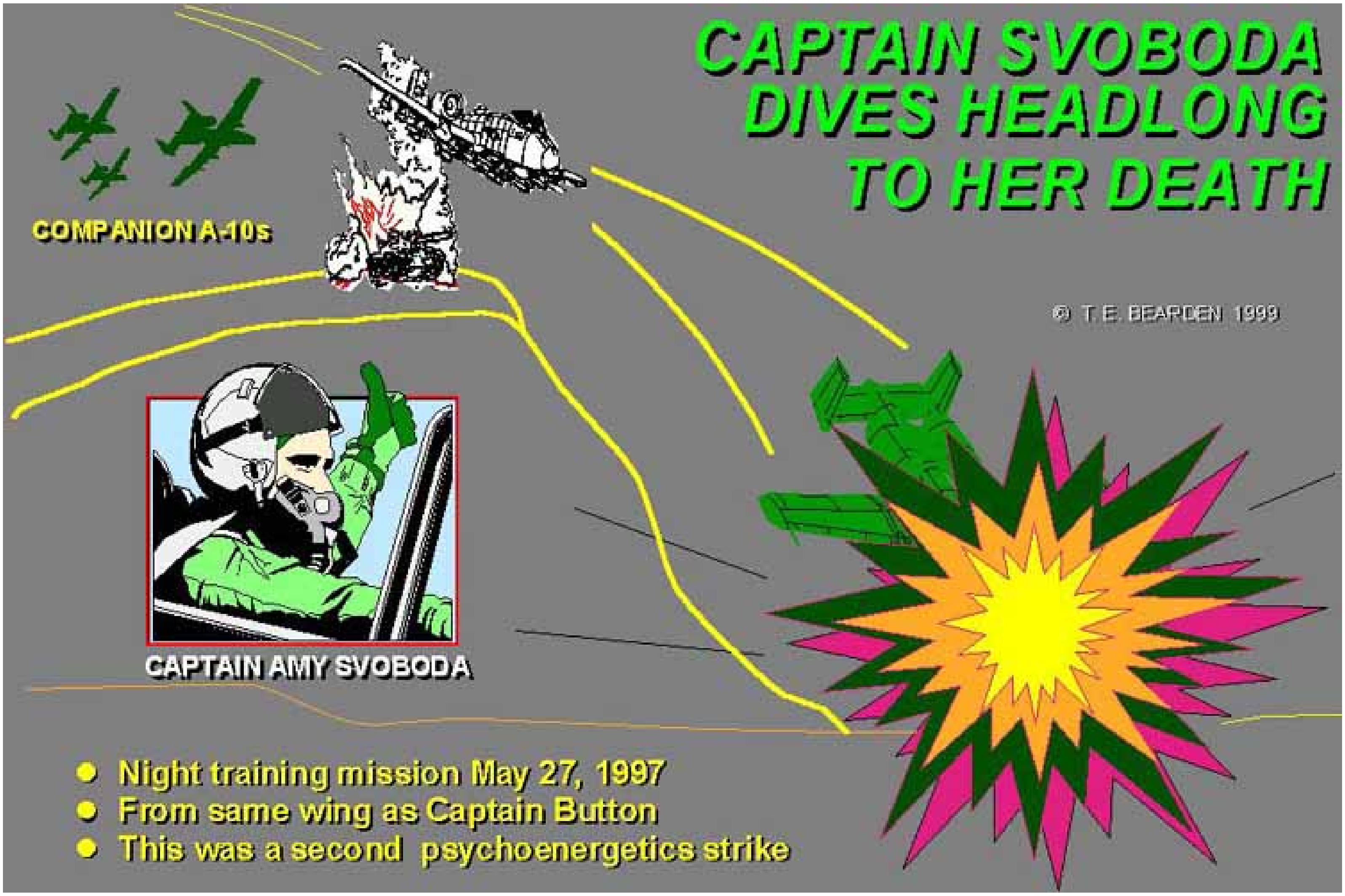
CAPTAIN SVOBODA
DIVES HEADLONG
TO HER DEATH
COMPANION A-10s
© T. E. BEARDEN 1999
CAPTAIN AMY SVOBODA

- Night training mission May 27, 1997
- From same wing as Captain Button
- This was a second psychoenergetics strike

Figure 15. Captain Svoboda dives headlong to her death.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

Afghan War to condition them for performance of merciless acts.[13]

Potential U.S. Rogue Groups and Two Recent Examples

Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.

Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed.

We Certainly Have Proof of Such Rogue Activity. For Example, Here are Two Prominent Cases:

as a result of those ghoulish experiments. So what happened when this gruesome thing was revealed? Again President Clinton publicly apologized! The head of the Executive Branch did not turn the macabre matter over to his Attorney General to prosecute. How many of these scientists were indicted? Not one. How many are going to be indicted? Not one. Suppose again this had been done by a private doctor in his own private clinic. You get the point.

There is Little or No Punishment of Rogue Scientific Groups

Shockingly, the U.S. government at the highest level has directly shown (and these are not all such cases by any means!) that mass crimes against U.S. civilians, *perpetrated by portions of the U.S. scientific community in direct conspiracy and in secret*, will likely be condoned. The perpetrators will not be indicted, tried, or convicted.

Note the connection of the "rogue groups" thesis to the above incidents. Note how rogue **scientific** groups got away with it in both these cases. There may be other rogue

1. For three decades, scientists from the U.S. government, universities, and civilian contractors secretly conspired to treat unsuspecting human patients with whole body nuclear radiation, including some retarded children and cancer patients. Some of those patients—including some of the retarded children—died as a result. These illegal experiments were conducted in great secrecy, and the results were highly classified. Eventually these actions were revealed, and a Presidential investigation committee investigated. These experiments would seem to be little different from the WWII Nazi and Japanese experiments on human prisoners. Are those U.S. scientists who were responsible for those retarded children's' deaths any different from the Nazi scientists and criminals we executed at Nuremberg? So what happened when this terrible thing finally came out into the open? President Clinton publicly apologized! How many of the responsible scientists were indicted for murder? Not one. None is ever going to be.
   Suppose a doctor here in South Alabama, with his own private clinic, had secretly irradiated those human patients and retarded children with whole body nuclear radiation, over a period of 30 years, resulting in some agonizing deaths. How many state and federal agencies would have come after him with arrest warrants? Probably about 50 or so. Would he have been indicted, tried, convicted, and executed or sent to prison for scientific groups getting away with such things today. They should not be able to get away with it. But they can.

Perhaps a most startling additional part of those two incidents is that there was no great public outcry from the scientific community, deploring these murders and demanding that the criminal scientists be indicted and tried.

As the old saying goes, "By their silence they have convicted themselves." The organized scientific community, **as a community**, has shown that it has little or no ethics, and—while deploring any scientific murdering that "gets revealed,"—is not really interested in justice. In short, much of the U.S. scientific community may now have very little ethics left.

We are **not** talking about normal individual scientists, but the Big Science community. There is a whale of a lot of difference between the two. In the Big Science community, there can be and there are rogue groups. Lots of them. There is deep cover, deep classification. And there is very probably advanced mind control research and testing, be it legal or illegal. Hopefully most of it is legal and constrained. However, some of it is almost certain to be illegal and ill constrained.

Human beings are still human beings. All the good and evil is still there, regardless of the group. The stage settings change, but the cast of characters and the play never change.

It's sad, but 'twas ever thus. Hidden parts of our own governments—and other governments throughout the world—are no different from the old medieval groups, where nobles etc. were always plotting against the king, or using the king's power for their own nefarious end. The rogue groups today are no different from all the scheming and conniving groups that life? Absolutely! Would the President of the United States have apologized? Not on your life; he would have turned it over to the U.S. Attorney General with a strong directive to investigate and prosecute on criminal charges.

2. Again for three decades, scientists from (i) the U.S. Government, (ii) universities, and (iii) civilian contractors conspired to give poor syphilitic blacks in Tuskegee, Alabama a placebo while ostensibly treating them for syphilis. The purpose of the program was to deliberately observe and record the ravages of the disease to its fruition in their wracked bodies. So the scientists deliberately sat there and watched their brains rot and their bodies rot. They kept meticulous notes, of course. It was very scientific. And it was macabre. They "treated" some 400 blacks in the "program." Over 100 blacks died destroyed the Roman Empire. *Great empires fall from within, not usually from without!*

Conclusion

This concludes our brief portrayal of the probable directions that mind control research has taken, in Russia and in the West. It is a novel and powerful area, and it poses both severe dangers and great benefits to all the peoples of the Earth.

Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction of helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth. Yet it can revolutionize medical science, education, communication, and psychology. Eventually it can engineer the mind and memory directly, beneficially. We foresee the day—perhaps 30 years hence—when education will be

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

accomplished *by directly loading the software into the mind.* Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks one will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M. D. And so on. When that happens, then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.

We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on Earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to

446955, Oct. 23, 1998, forwarded to US. Patent Office on Nov. 4, 1998.

5. A very good introduction to nonlinear phase conjugate optics is given by David M. Pepper, "Nonlinear Optical Phase Conjugation," Optical Engineering, 21(2), March/April 1982, p. 156-183. Also see David M. Pepper, "Applications of Optical Phase Conjugation," Scientific American, 254(1), Jan. 1986, p. 74-83.

6. Pepper, Optical Engineering, ibid., p. 158.

7. See T.E. Bearden, Energetics of Free Energy Systems and Vacuum Engine Therapies, Tara Publishing, Internet node www.tarapublishing.com/books,July 1997. See also T.E. Bearden, "Energetics Update and Summary, " Part I, Explore, 7(6), 1997, p. 60-67, Part II, Explore!, 7(7), 1997, p. 53-56; Part III Explore!, 8(1), I997, p. 53-56; Part IV; Explore!, 8(3), 1997, p. 56-63. See also T.E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy-Medicine: Parts I and II " Explore!, 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

8. E.g., see R.O. Becker; Charles H. Bachman, and Howard Friedman, The direct current system: A link between the environment and the organism, " New York State Journal of Medicine, Vol. 62,

be the Sword of Damocles and turn it into the golden Millennium.

We are reminded of a poignant quote from Teilhard de Chardin:

"Someday, after we have mastered the winds, the waves, the tides and gravity, we shall harness for God the energies of love. Then for the second time in the history of the world man will have discovered fire."

May the nations of the world utilize the new mind engineering science to truly free all minds into a common love of, and respect for, their fellow humans of every color, creed, and circumstance. As de Chardin envisioned, may we use the new science to discover a new fire, burning clearer and brighter than anything that has gone before.

♦

ABOUT THE AUTHOR

President and Chief Executive Officer, CTEC, Inc. Lieutenant Colonel US. Army (Retired). MS Nuclear Engineering, Georgia Institute of Technology. BS mathematics, Northeast Louisiana University. Graduate of Command & General Staff College, US. Army. Graduate of Guided Missile Staff Officers Course, U. S. Army (equivalent to MS in Aerospace Engineering). Numerous electronic warfare and counter-countermeasures courses.

Tom is the leading advocate of scalar potential electromagnetics and the Fogal charge-blocking semiconductor: He is involved in alternate energy devices and scalar electromagnetic system prototypes. He has defined charge q as a coupled system of two components, advanced a mechanism for a vacuum engine, whereby the vacuums virtual particle flux is organized and used to energetically shape and manipulate

April 15, 1962, p. 1169-1176. See also R. 0. Becker; "The direct current field: A primitive control and communication system related to growth processes, "Proceedings of the XVI International Congress of Zoology, Washington, D.C., VOL.3, 1963, p. 179-183; —"The neutral semiconduction control system and its interaction with applied electrical current and magnetic fields," Proc. 1lth International. Cong. Radiol., Amsterdam, Excerpta Medica Foundation, 1966, Series 105, p. 1753-1759. See particularly Robert O. Becker and David G. Murray, "The electrical control system regulating fracture healing in amphibians," Clinical Orthopaedics and Related Research, Vol. TBD, No. 73, Nov.-Dec. 1970, p. 169-198.

9. Bearden, T.E., Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications, CTEC Proprietary, May 1, 1998, 200+ page enclosure to CTEC Letter, "Saving the Lives of mass BW Casualties from Terrorist BW Strikes on US. Population Centers," distributed to selected high U.S. Government agencies and Legislative subcommittees, May. 4, 1998.

10. E. g. , Vlail Petrovich Kaznacheyev, V. I. Vernadsky's Study on Biosphere and NooSphere, 1989 [in Russian]. Vernadskys concepts included living matter, biosphere, and human-transformed biosphere (noosphere). Kaznacheyev writes of evolution and

matter and energy, and advanced a mechanism for a Bohm-type quantum potential, which theoretically allows the engineering of action at a distance in physical systems.

He is president of the Association of Distinguished American Scientists (ADAS), a life member of the Alabama Academy of Science, Emeritus member of the American Association of Physics Teachers, and has served on the Board of Directors of several associations and corporations. Tom is a guitarist, a retired aikidoist (Sandan, Yoseikan style), is happily married and has two sons and two daughters. He and his wife Doris live in Huntsville, Alabama where Tom is retired from aerospace, continues private research, and serves as a consultant to industry on Scalar electromagnetics processes.

REFERENCES

1. Owen Flynn, "Parametric arrays: A new concept for sonar, "Electronic Warfare Magazine, June 1977, p. 107-112.
2. Valerie Hunt, presentation to the U.S. Psychotronics Associations Annual Symposium, Bergamo Center, Ohio, 1988. Hunt presented epochal work showing chaos—and thus hidden order—in recorded ELF electromagnetic radiation from human bodies. Hunt split the received EM signals into two identical channels, time-delayed one channel by a very slight amount, then recombined (mixed) the nondelayed and delayed signals. The result was that attractors precisely indicative of chaos appeared. This positively proves that human systems produce ordered time-

organization of the living matter as the biosphere of the earth. Without specifying the exact technical details, Kaznacheyev discusses the gist of much of what is in this present article. The present author has an autographed copy of Kaznacheyevs book referenced above.

11. Rabichev, Lev Yakovlevich; Vladislav Fedorovich Vasiliev; Alexandr Sergeevich Putilin; Tatyana Grigorievna Ilina; Petr Vasilievich Raku; and Leopold Pavlovich Kernitsky, "apparatus for the treatment of neuropsychic and somatic diseases with heat, light, sound; and VHF electromagnetic radiation," U.S. Patent. No. 3,773,049, Nov. 20, 1973. This is the U. S. patent for the LIDA machine, developed by the Russians prior to the Korean War. Additional patents or applications are as follows: Italy, #619,129, March 1961. Germany, #1,183,607, Dec. 1964. Great Britain, #826,766, Jan. 1960.

12. Stefan T. Possony, "Psy-War: Soviet Device Experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet device called LIDA which is used to bombard human brains with radio waves in the 40-MHz region. Also reports on work by A.S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected. See also "Soviet device uses radio

structuring of EM.

3. Years ago, I read an obscure reference to Russian research directly along that line, with successful development of just such a portable device. Unfortunately I lost the reference in a disastrous flood in my basement and have been unable to recover it.
4. T. E. Bearden, "Alteration of Materials by Wideband Multiwave Irradiation Combined with Selective Emission Manipulation," Invention Disclosure Document Number

waves as tranquilizer, " Associated Press release, Huntsville [Alabama] Times, May 19, 1983. Dr. Ross Adey tested the Soviet LIDA machine on a cat, and in two or three minutes the cat was sitting quietly, almost as though transfixed, and uninterested in its surroundings. The animal remained uninterested in its surroundings for some 20 to 30 minutes after the machine was turned off.

13. "KGB Used PIS to Program Soldiers, "Resonance, No. 30, Mar. 1996, p.29. Quoting: "During the 1970s, the Soviet KGB developed a Psychotronic Influence System (PIS) that was used to turn people into programmable "human weapons." Soldiers fighting in the Afghan conflict were trained using the system which employs a combination of high frequency radio waves and hypnosis. Commands are activated by code words or cyphers. The purpose of PIS was to render regular soldiers and members of the special forces incapable of feeling sympathy.

EXPLORE! VOLUME 9, NUMBER 4 & 5, 1999

Latest cloning research findings by Bedini/Bearden

The huge dangers of current cloning technology

Mystery of the death of Australia's cloned sheep explained

"Spark of Life"

From "Stranger than Science"

by Frank Edwards © 1959, 1987

Did an obscure amateur scientist discover the spark of life? Did Andrew Crosse accidentally stumble upon a mixture of chemicals and electricity which produced living things from a lifeless compound? If he did not, then just what did he accomplish? The record of his work is clear, only the enigma of the results remains to plague science.

The neighbors of Andrew Crosse regarded him as more devil than man. They did not understand the bright flashes that lighted his laboratory windows at night when he was tinkering with his crude electrical devices. Not only was he dreaded and shunned as "the thunder and lightning man," but he was denounced as an atheist, a blasphemer, and a Frankenstein who had best be put in chains for the common safety.

Andrew Crosse minded his own business, it is true, but his was a very strange business for the early 1800's. And one of his experiments remains a very strange business to this day.

Andrew was an enthusiastic amateur experimenter in the new field of electricity. Out of touch with others in the same work, he labored under the double handicap of not knowing what had already been done - and of not understanding what he was doing himself. Yet it may have been this very lack of knowledge which led him to undertake the experiments which were to inscribe his name, however faintly, in the annals of science.

He decided to induce the development of artificial crystals by subjecting chemicals to prolonged exposure to weak electrical currents. Andrew mixed up some silicate of potash and hydrochloric acid and into this he dropped a fistsized chunk of oxide of iron. By inducing the current from a small battery to trickle through the solution to the oxide of iron, he hoped to bring about the growth of artificial crystals of silica.

Having arranged this combination of chemicals and current, Andrew set it aside and went back to his favorite pastime of studying the spark-gap - the flashes which alarmed his neighbors.

Did he accidentally stumble upon an arrangement which created life from inorganic matter?

In a paper which he wrote for the London Electrical Society in that same year of 1837, Andrew set down this account of his experience.

He wrote "On the fourteenth day after the commencement of this experiment, I observed through a small magnifying lens a few small whitish specks clustered around the middle of the electrified stone. Four days later these specks had doubled in size and had struck out six or eight fine filaments around each speck . . . the filaments longer than the hemisphere from which they projected.

"On the 26th day of the experiment, the objects assumed the form of perfect insects, standing erect on the bristles which they were growing. Although I regarded this as most unusual I attached no singular significance to it until two days later, the 28th day of the experiment, when the magnifying lens showed that these things were moving their legs. I must say now that I was quite astonished. After a few more days they detached themselves from the stone and moved about through the caustic acid solution.

"In the course of a few weeks more than a hundred of them made their appearance on the oxide of iron. Under a microscope I examined them and found that the smaller ones had six legs, the larger ones had eight. Others who have examined them pronounced them to be of the genus acari, but some say they are an entirely new species.

"I have never ventured an opinion on the cause of their birth for the reason that I have never been able to form one. I thought they might have been airborne creatures that had drifted into the liquid and prospered, but later experiments with closed vessels, in which the ingredients had been purified by baking in the oven, produced identical creatures; therefore, I suggest that they must originate in the electrified liquid by some process unknown to me."

Andrew Crosse realized that he was walking a tightrope before the top scientists of his day. He was describing an experience foreign to their accepted understanding and therefore he was inviting ridicule. He did not have long to wait. Cries of fraud and hoax engulfed him. He and his alleged insects were denounced as nothing more than humbugs.

Amid all the furor that his announcement had created, Crosse stood alone and helpless. Even other scientists who had duplicated his tests with similar results kept their silence.

All but one. That voice was raised in his defense and it was such a powerful voice that none dared challenge it; for it belonged to the great Michael Faraday. At last Crosse had found his champion.

Faraday reported to the Royal Institution that he too had experienced development of these little creatures in the course of his experiments. But he added that he could not decide whether they had been created in the sterile solutions or brought back to life by the electricity! Either development would have constituted a milestone in scientific advance, as Faraday realized, but he left it to his fellow scientists to make the decision, if any.

Crosse never claimed to have discovered anything. He was merely reporting what had happened. After the attacks upon him subsided, he retired to his home in the Quontock hills for many more happy years among his test tubes and batteries.

Although he spent his long life as a humble searcher after scientific truths, the contribution for which he is remembered is the controversy over his acari - unwanted then, and unexplained to this day.

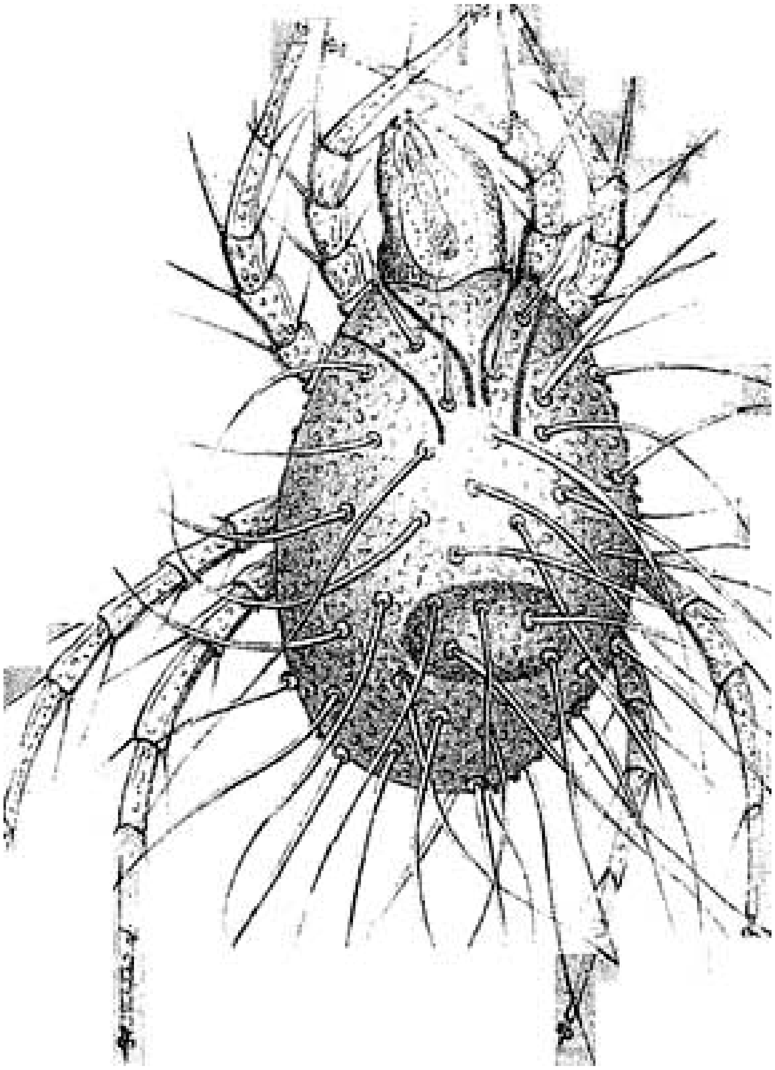

Genus Acari

Tom Bearden comments (finalized and slightly updated in March 2001).

It is difficult to make an understandable comment on the business of life, because presently scientists have not the foggiest notion what mind, being, life, death, time, etc. are. For that matter, they don't really know what matter and energy and charge are either. In the scheme of things, our modern "science" is only about three hundred years old, and that's a snap of the finger in even just the existence of this world. Much of modern science still considers the mind as a "meat computer", or just the physical electrodynamic activity of the brain, etc. Materialism has not died, even though quantum mechanics essentially wiped it out decades ago, and the old concept of "mass" as something hard, permanent, and "material" has also long since vanished. One is referred to the works of Max Jammer to see what the present view of "mass" is. Also, without looking up the references again, recent experiments with quarks and gluons wound up with experiments in which it mattered not a whit whether one considered the mass as zero or nonzero. Nature, it seems, does not have to play by our usual "either-or" rules. In physics, of course, this is the duality problem, which to this day has not been resolved because attempts to resolve it have applied Aristotelian logic -- and there is no solution to the problem of "opposites being identical" in that logic, which forbids that very thing by authoritarianism.

The modern biogenesis experiments, e.g., never took place in the absence of life. There was no "statistical generation of living or quasi-living forms" in the absence of engines for such forms. The great rush of Poynting energy flow around the earth, generally parallel to the equator, alone carries the "engines" (internal sets of spacetime curvatures) for every form of life that ever existed upon the earth. So the experiments were bathed in "living forms and dynamics", if one accounts for the internal Whittaker-structuring of electrodynamics -- noted but ignored by our scientists for nearly 100 years.

Quite some time ago, we worked out the nature of mind, and the mind-matter coupling mechanism (which is engineerable). We did that in pursuit of the mechanisms in the KGB/Russian psychoenergetics weapon work. I had the last of a 3-part article published in "Explore", dealing briefly with mind control technology both the way it is probably done clandestinely in this country and how it is done by the KGB/Russians. I'm still trying to get the proper parties to believe it; not much progress. However, some very neat information has been released on the Internet that very probably reveals why officially there is no interest. It's because three decades ago some of our "spooky folks" found (in deep hypnosis experiments) that humans are inordinately sensitive to extremely low-level EM signals and fields. Even though no detection or sensation may rise to the conscious level, the unconscious (which is totally conscious, just massively parallel rather than serial) does sense these signals. Voila! It would appear that "very deep black" parts of the community probably developed quite good mind influence and mind and behavior control or interference using techniques adapted from that work. The reader must always remember that just because the *overt* government does not know something or have something, does not at all mean that the *covert* government doesn't have it. And even in the covert government, there are layers within layers. So there, even when the upper levels do not have something, it still does not mean that the deeper levels do not have it. The problem with the best knowledge and technology, when unknown to the upper covert levels and the overt level, is that it is then quite likely to be subverted into the deepest levels, which may have their own agendas and may not report to the duly elected U.S. government at all. Indeed, such levels often assume that the President himself does not have a "need to know".

Present physics has erroneously omitted the time increment carried by the photon. The overemphasis on the so-called transverse wave (which does not exist as such in vacuum, fairly readily shown) has led to our missing the greatest and most important part of electrodynamics: the longitudinally polarized EM wave and the time-polarized EM wave. For example, mind and mind operations are totally electromagnetic, but use time-polarized EM waves rather than the standard stuff. Overtly we do not even have detectors for LPWs and TPWs yet; the Russians do. However, every major weapons lab on earth now is aware of longitudinal EM waves and their powerful characteristics. Simply read some of the summary material on Undistorted Progressive Waves (UPWs) on the Los Alamos National Laboratory website, e.g. by Rodrigues et al. A UPW is an "imperfect" longitudinal EM wave with some remaining transverse wave residue (that is the way it is modeled).

Let me give you one example of how fouled we are. We still use Aristotelian logic primarily. Yet it is an incomplete logic. Let me show you instantly. Take the simple Venn diagram of a rectangle lying on its side, divided in half by a vertical line in the middle. Label the right rectangular area A. Label the left rectangular area "not-A". What is the line in the middle and the border? Both A and not-A, so that A is identical to not-A. Yet the third law (law of the excluded middle) forbids A identical to not-A. Apply the third law to the Venn diagram. The border disappears, and the middle dividing line disappears. The logic self-destructs, because now there is no separate A and no separate not-A. Yet so far as I can determine, no one seems to have seen that very simple thing and questioned it. Aristotelian logic cannot be applied to the very Venn diagrams used to establish and "prove by demonstration" its propositions. We corrected that nearly 20 years ago, and got called every kind of lunatic imaginable -- the usual reaction when something fundamentally wrong is pointed out in a comfortable part of the prevailing paradigm. These days I simply point to Morris Kline's "Mathematics: The Loss of Certainty" to the Aristotelians, and tell skeptics to go and read it and then we will discuss the foundations of logic. With a 5-law logic (or 4-law with an additional applications rule), one solves the "dividing line" problem nicely. Then suddenly things like spacetime and being and mind and nothing etc. can be solved -- in the 5-law logic, but never in the 3-law logic. The problem of life itself -- whether terrestrial or extraterrestrial -- has no solution in 3-law logic.

The entire method of Zen is to try to break up the iron Aristotelian bent of the conscious (serial processor) mind, and get into the far broader unconscious (totally conscious, massively parallel processor) mind. Few have done that. But one cannot present a 5-law solution to a 3-law processor; it just "does not compute". It's like putting a thousand slides at once on the normal "single slide at a time" projector. The "conscious mind" can see only a single slide at a time. So it just sees "black" or "nothing at all". That's why we "see" the "massively parallel conscious" mind as "unconscious"! It isn't. But it is a quite different *form* of consciousness which we all have.

Time is never observed, and no observable is a *cause*, a priori, because an observable is only a frozen, nonchanging 3-spatial "instant snapshot" and therefore the *output* of the observation process, while the *causal input* is what exists in 4-space prior to its interaction with the observation process. When one adds time and time dynamics to one's science, one moves far in front of much of present Western physics. E.g., physics has nearly hopelessly confused the *causal* side of the observation mechanism/process with the *effect* side, everywhere, throughout physics -- simply by assuming that observables persist in time -- a non sequitur since an observation is a frozen 3-space snapshot at one single instant.

And sadly, the physicists do not seem to have even realized it!. Not a single paper or text in the West, e.g., illustrates an EM wave in spacetime prior to its interaction with a unit point charge. E.g., simply check a fine editorial by Romer. Romer is the editor of American Journal of Physics, and we are referring to Robert H. Romer, "Heat is not a noun," American Journal of Physics, 69(2), Feb. 2001, p. 107-109; and specifically to his end note #24, p. 109. Romer takes to task *"…that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. …it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "…perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

But in general physicists continue to charge on, confusing effect as cause. Electrodynamics, e.g., is particularly guilty of greatly confusing the two, which is really what is preventing a successful and engineerable unified field theory. Further, contemporary classical electrodynamicists do not calculate either the field or the potential; instead they calculate the *reaction cross section* of each, interacting with a unit point charge assumed out of nowhere. In short, they calculate how much is diverged from the actual field or potential, and then call that "little extracted part" the field or potential! Again, they mistake the causal entity existing before the interaction with the effect entity existing after the interaction. Well, after the interaction has occurred, that result is the *effect*, not the *cause prior to interaction*. Physics has this so hopelessly snarled that it will require 50 to 100 years before the journals and physics societies are likely to even realize it and admit it, and then it will require them another 50 years to straighten it out and purge this huge non sequitur from physics theory and modeling.

Anyway, we did model how mind and mind operations operate, including their precise nature, how they fit into all this, etc. We thus were able to resolve many previously unsolved riddles such as the nature of intent, how a space-less mind operation can generate a spatial input to a spatial body and vice versa, etc. We also worked out how to use "effects" systems to generate "causal" systems, completely independent of mass -- because the KGB has recently completed a completely different kind of weaponry based on this, and different from anything that has ever existed in the history of mankind (at least on this planet!). The Russians of course were already there; I got there much more crudely by intensely working on how to explain what they were doing.

To see that the effect can generate the cause, just consider the normal "forward time" observation process having to operate in a "time-reversed" process. You see, not only must we shift observers from frame to frame, but we must also be able to shift them from forward-time (of some conveniently chosen outside reference observer) to reversed time (of that same reference observer).

At any rate, it turns out that the entire universe is already "alive" in a very special sense. There is no death; there is mind (I use the term "mind" in the most general sense, not just the conscious mind which is merely a "periscope" put up and taken down from the unconscious). And mind is not a part of the brain, and it is indeed separated from body but intimately linked to it by the mind-body coupling dynamics. Further, any life that has "ever existed" still dynamically exists in a special kind of more fundamental electrodynamics infolded inside all normal EM fields, waves, and potentials. The Tibetans knew much of this very well, but of course in terms that our materialistic science rejects or completely fails to understand (as we stated, most of our scientists believe the human is a robot and the mind is simply the electrochemical reactions of a "meat computer"!).

Since ambient spacetime may be modeled as a special kind of "scalar" potential with intense energy density, it then decomposes into a series of bidirectional EM longitudinal wavepairs, a la Whittaker 1903, except one must use impulse waves so as to include the time dynamics part of spacetime, rather than forcefield 3-space type waves. The entire thing -- mind, life, matter, fields, waves, spacetime, etc. -- then becomes structurable and directly engineerable by novel EM means (such as the O(3) higher symmetry electrodynamics pioneered by Evans and Vigier). Longitudinal EM wave technology is the key to developing such mind-body coupling mechanisms and engineering of them, at all. One only gets the time-polarized EM wave technology *after* the LPW technology has "done its thing" -- the time-polarized technology is far and away the most powerful and primary.

Also, low frequency photons have far more energy than high frequency photons, if one includes the time component, since time is spatial energy compressed by c-squared and has the same energy density as mass. Time-energy (highly compressed spatial energy, compressed by c-squared but in the time domain rather than in 3-space) and spatial energy in the photon are canonical. As the frequency of the photon decreases, its spatial energy decreases also with the decreasing frequency. However, the time-energy component increases, and its compacted energy increases as the square of the time-component expressed in seconds. Hence at lower frequencies the total energy (sum of spatial energy and time energy) in the photon increases. Indeed, if one wishes to engineer gravitational effects, it is the low frequency photons (low ELF) that are important, because they have far and away more "spacetime curving energy" than high frequency photons such as gamma rays.

At least eight nations on earth now have longitudinal EM wave technology weapons to one extent or another. Several of them -- hostile to one another -- are using Western Australia as a convenient nearly uninhabited target range.

Anyway, when you come at the mind-body-life problem with that background (which unfortunately required more than 20 years of very hard work!) there appears to be no great difficulty in explaining the Faraday, Crosse, Fox, Miller, and Urey biogenesis experiments. And yes, there appears to be no difficulty in explaining the "appearance" of "life on earth" in the prebiotic soup. It did not really appear out of nowhere spontaneously at all, but simply was a sort of sympathetic resonant coupling of physical forms to mind-forms already there. None of the experimenters realized that the great Poynting energy flow (and one must add back in the missing Heaviside component which is some 10exp13 times larger than the accounted Poynting flow) around the earth parallel to the equator and from West to East, is already internally structured with every possible form of life ever in the universe anywhere. If we further multiply that Heaviside component by c-squared, so that we can account for the time-energy portion of that great energy flow, we can see immediately that life and its environment are both living and intercommuting all the while. The creator in fact created a living universe, not an inert one (as best we can speak it in the inadequate 3-law logic). If one believes in big bangs, then life structuring emerged right along in all the fields and waves and energy, right in and out of the big bang. There is not a single point in the universe anywhere that is absent of energy and therefore absent of life, in the view of the model I am using.

Understand, one does not have to agree with this model! That is one's own decision. We are just pointing out that, if we correct many of the errors known to be in the presently used models, this is the kind of reality that readily emerges. We get a physics that is not just an "inert" physics, but a living physics where literally the entire universe is "alive" in very special manner.

In deciphering (after 14 years!) the Priore mechanism, I found that incredibly the cellular regenerative system does retain the form and "memory" of all the precessor cellular forms. Cancer, in fact, is very often due to a direct "order" by that system to a cell (aerobic) to start back down the "chain" to an anaerobic cell, because of severe hypoxia and the sustained inability of the regenerative system to maintain the cell in aerobic form. It doesn't just send an "order", but sends a "precise set of spacetime curvature engines" that go to work on that cell and all its parts. The first major step back is to break it loose from centralized control of its growth -- whereupon we recognize it as a tumor cell. There are two requirements: (1) cellular damage, and (2) "promotion". Biologists do not really understand "promotion", because they do not understand the cellular regenerative system—which used time-polarized EM waves, and nonlinear optical pumping of the cell and all its parts in the TIME-energy domain rather than the SPATIAL energy domain. That sort of pumping creates an "antiengine" to its forward time engine resident in the sick cell, thereby time-reversing the cell and all its parts back to a previous state on the trail from the early anaerobic single cells to the present aerobic cells.

By the way, recent experiments in cloning are beginning to bear out the fallacy that exact electrochemical processes produce the same end physical results, particularly when the persistence of the created forms is examined. E.g., cloned rats may live half their lives in apparently normalcy—then suddenly grow greatly obese and die. Many other anomalies are being uncovered in this fashion. The reason is quite simple: We have some diagrams showing that one may introduce an additional "engine" to the pumping of a diseased or damaged cell (or a healthy one for that matter), and in the time-reversal process then induced, the "time-reversal" process will be diverted from the original trajectory. In short, the original full state will not be recovered.

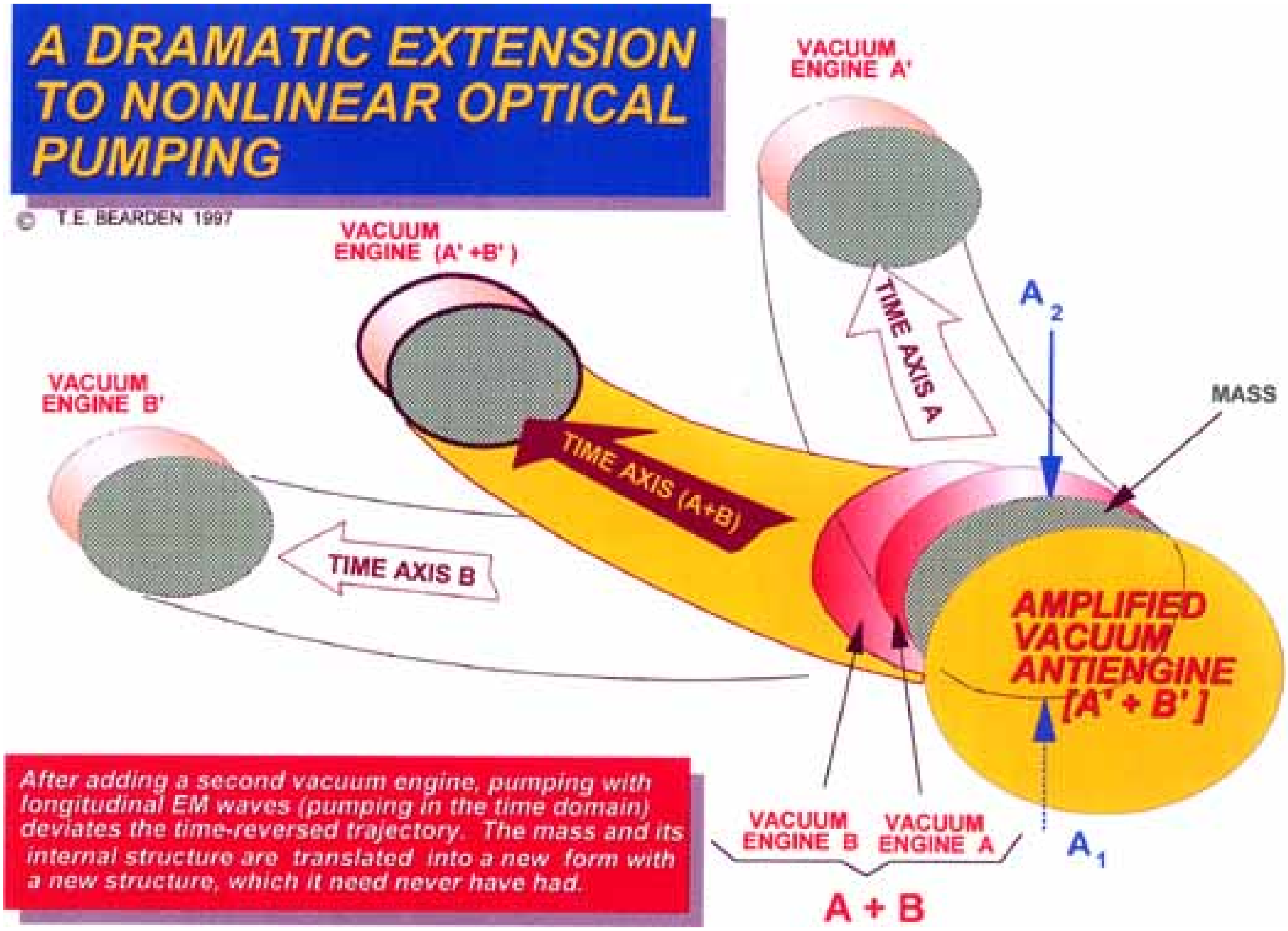
A DRAMATIC EXTENSION TO NONLINEAR OPTICAL PUMPING

© T.E. BEARDEN 1997

After adding a second vacuum engine, pumping with longitudinal EM waves (pumping in the time domain) deviates the time-reversed trajectory. The mass and its internal structure are translated into a new form with a new structure, which it need never have had.

This, we pointed out, is a second generation technology to be developed, where a disease or disorder must be treated and where also there was never a normal past state anyway. A case in point is birth defects, etc. Eventually, if the technology ever is allowed to be funded and developed, even those kinds of medical conditions will be treatable and curable. The point is that in cloning, one does not use the normal sperm cell and egg, without abnormal disruption. There is indeed an abnormal disruption -- and every such physical action or change and dynamics certainly involves an additional set of engines. In short, with present methods the cloned entity also has introduced into the equation an additional "engine". In the fertilization process, there is quite a jump back to a "time-reversed" situation (fertilization is the beginning of a physical embodiment, hence from two parts of an embodiment already in being, there is a rather dramatic "jump backwards"). The "normal" backwards jump of that kind is a normal fertilization of the egg by the sperm. Any change to that procedure, represents the insertion of an abnormal engine to be processed in that same "jump back". For the methods being used, it appears that the abnormal engines are sufficiently weak that, in the cloned animal, some time must expire in its lifetime before the impact of that little constant deviation from normal grows to become suddenly significant.

Indeed, all "living forms" are materialized from energy, but from both matter-energy and time-energy and the dynamics of the two. Natural processes include fertilization, birth, etc., but there is no technical reason whatsoever that the process cannot be direct and even directly engineerable. In this "expanded energy" sense, mind energy and dynamics can be engineered as well as body matter energy and dynamics.

To cinch the matter of "engineering living forms", I once participated in a accidental and unplanned experiment where Golden materialized living things. Not little bitty things, but big things. In my view we actually materialized thought forms, since much later we were able to piece together exactly what apparently engendered each part of the phenomena. There were multiple (scared) witnesses who prefer never to discuss it. The experiment also altered all the clocks in the area for four days before the "time-energy charge" we had inadvertently created dissipated. Let me just say that it really got my attention! We never repeated that experiment again. But the changes were substantial and the phenomena were substantial, and to this day I prefer not to discuss that experiment. It simply cannot be understood without the outline sketch we are speaking of, where one accounts for time-energy and dynamics as well as material energy and dynamics, and also for the coupling dynamics between the two.

Not only can one have a "material entity" or thing communicating with another "material entity" or thing, but one can also have a "mental thing" (time-domain thing with no spatial existence) communicating with a "material entity or thing" with only spatial existence, and vice versa. There is indeed a coupling two-way mechanism

Look at the so-called "physical form" -- which almost always one is speaking of as purely 3-space. Most scientists have not yet understood that a priori a piece of 3-dimensional mass *as observed* cannot even persist in time! Masstime exists in time, but "mass" is a frozen 3-spatial intersection of the spacetime entity. It is exactly comparable to a frozen frame in a movie film. The individual frame never has any existence except at a frozen moment in the slide projector. "Motion" and "movement" etc. involve the time domain a priori -- hence something without any time-component at all, cannot persist or even exist except at one frozen moment in time. Hence any observable existed in that single "observation slice" and at that "point in time" only. After all, a "point in time" is where time symmetry is broken or ceases. In short, it is a time-discontinuity. All observations are such time-discontinuities.

The SETI problem will not be solved, in my personal opinion (understand, I do not object at all if others do not see it this way) until a far greater and deeper understanding is reached by our modern physics. Forget communicating with "transverse" waves; longitudinal EM waves can travel at any velocity, limited only by how much "residue" of the normal EM wave junk is still hanging in there (how much "transverse EM wave noise" is still present. A pure LW moves at infinite velocity. As we suggested, simply download the Rodrigues and Lu papers carried on the Los Alamos National Laboratory Web site. It appears that now most major weapon labs have "discovered" the fact that there can be longitudinal EM waves, and some eight or so nations have developed weapons technology from such. If one has such technology, all the rest is obsolete.

For the skeptic, here is a statement made in 1997 by Dr. William Cohen, then Secretary of Defense:

> *"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves… So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations…It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."*
>
> [Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.].

Any technological civilization that has not solved the fundamental "self-killer" problem of the technological species, must and will self-destruct. This follows from simple servo-mechanism theory, as we have previously pointed out. The KGB is working on a "solution" that is far along: engineer the entire human species quantum potential so that the human species can be turned into an "antlike" species, which of course would be the "ideal" communist state, in the minds of many Communist party theoreticians. Our guys haven't even got started on engineering the quantum potential yet. They have not yet even discovered the big ball park, let alone the big ball game.

Anyway, that is just some comments for your personal information. What really is happening in the "hidden weapons" area and "hidden weapons science area" is already so advanced in some quarters that science fiction pales by comparison. The real question is whether we shall survive very much longer, and whether we shall mature rapidly enough to begin to use these technologies for the benefit of humanity, rather than for its destruction.

Sincerely,
Tom Bearden

Additional Comments from Tom dated 9 July, 2001

CHENIERE PRESS

The Tom Bearden Website

Return to Order *Excalibur Briefing*

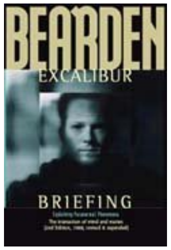

Help support the research

CHECK ONLINE STORE FOR UPDATES

EXCALIBUR BRIEFING

THE NEUROPHONE

Another device that uses the new hyperspatial, virtual state, nested modulation technology (and has done so for seventeen or eighteen years) is Dr. Pat Flanagan's neurophone. With brilliant insight and intuition far beyond that of science at the time, Pat invented and patented the instrument by the time he was seventeen years old. The neurophone is a device that, contrary to all present theory and knowledge, will directly “pump the brain” and reproduce sound and information directly in the brain and mind system, without going through the auditory system at all. A simplified diagram of Pat's improved neurophone is shown in Figure 11.

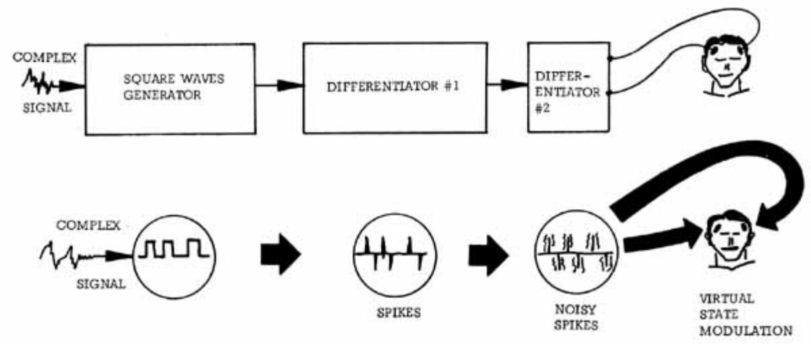

Figure 11. A simplified diagram of the improved Flanagan neurophone

Briefly, the device takes a complex signal, such as the sound of an orchestra playing a musical interlude, and electrically processes it as shown in the figure. First the signal is passed into a section that clips everything into a series of square waves, remarkably analogous to the sort of clipped waves Lisitsyn confirms are the information carriers of the human brain. Next the square waves are differentiated, yielding a series of sharp spikes (note that these spikes retain the pulse-time content of the clipped signal). These spikes are again differentiated, and since these are finite spikes with real nonzero rise times and decay times rather than theoretical constructs, a series of noisy spikes results from the second differentiator section. From here, the noisy spikes are introduced to special contact electrodes, one of which is normally placed on the forehead, while the other may be placed almost anywhere, including on the foot. Nowhere are any sound waves introduced to the head.

The square-wave clipper section reduces the complex signals, their overtones, and their complex modulations to square waves, retaining the temporal content of the wave mix but not the waves themselves. The first and second differentiators heighten or filter through the temporal content of the higher-order differentiations, that is, they serve as a band pass filter unit to accent the time keying of the neutrinic and mindfield portions or aspects of the signal. When these time spikes are then introduced across the body as pulsed voltages, they are modulated directly on the dendrite firings of the brain and nervous system, providing direct and pulsed modulation of the neutrinic and mindfield component channels of the mind-brain-consciousness-life loop itself. Thus the neurophone directly inputs information into the brain and nervous system, bypassing all the normal sensory systems that lie between the mind-brain loop and the outside environment.

Indeed, a similar development has been reported by two University of Missouri researchers. Dr. Donald York, a neurophysiologist, and Dr. Thomas Jensen, a speech pathologist, have recently reported identifying and decoding twenty-seven words and syllables in specific brain wave patterns and correlating these electroencephalographic patterns with both the spoken word and the silently thought word in about forty subjects. At present, Dr. York and Dr. Jensen are programming a computer with a brain wave vocabulary, to monitor and read the EEG of a stroke victim's brain and help stroke victims who have lost their powers of speech to communicate.

While all such developments can obviously be misused, their potential for assistance to mankind is enormous. Indeed, my own premise that brain linkage is possible and can be used to advance all mankind to the sixth stage of species evolution is slowly being proven. It is within our reach now to develop a multi-channel communications system that will directly link brains, minds, and consciousness into a single functional being; and I have already pointed out that the psychokinetic power of the emergent entity increases exponentially with the number of linked stages. The normal corpus callosum linkage of the two cerebral halves in each human head proves that brains, minds, and beings can be linked and integrated into one.

The Excalibur Briefing details the basic concepts of the theory involved in hyperspace-virtual state engineering, and thus in the engineering of mind-life-consciousness links to multicellular organisms here on earth. Pat Flanagan's neurophone proves that complex information can be implanted directly in the brain and mind, bypassing the normal isolating barriers. Drs. York and Jensen have shown that the brain wave does indeed contain recognizable, decodable information analogues. Lisitsyn has already reported measuring the number of independent channels involved and has documented the importance of the clipped brain waves as opposed to the sine-wave content. If we have the vision, we can still develop a brainlink and rapidly apply it to link sixty to one hundred persons, thereby opening a conscious, multidimensional “eye” for the collective human unconsciousness, gently rousing ZARG into wakefulness, and fulfilling human destiny without undergoing the psychotronic Armageddon planned by the Soviets. The time is short, but it can still be done if the funding, the vision, and the team can be established on a crash priority basis.

SOVIET “WOODPECKER” SIGNALS

But, as I have said, all this can be used for the detriment of mankind as well as for its benefit. Unfortunately, it appears that the Soviet Union has chosen to weaponize the effects on a global scale. A prime example is the “woodpecker” signals emanating from the USSR in the 5-30 megahertz region and interfering with communications around the earth.

These complex woodpecker signals appear to originate from two or three dozen powerful Soviet transmitters, each with a power estimated as high as 40 megawatts (Figure 12).

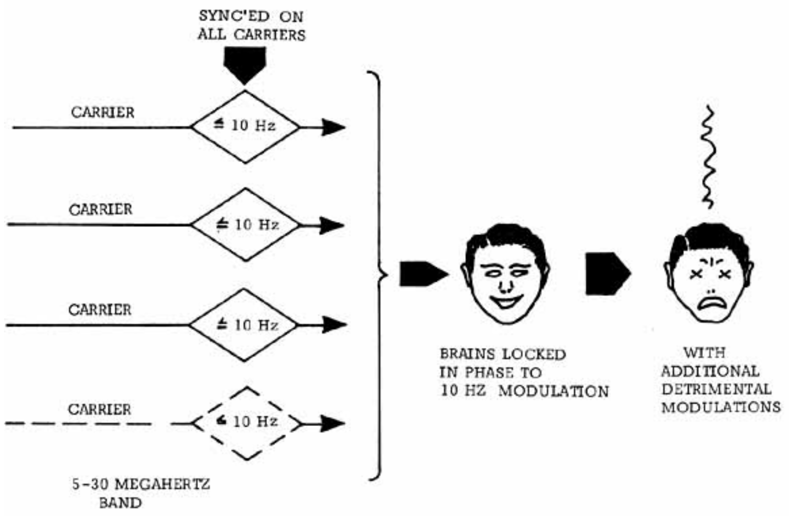

Figure 12. The Soviet “woodpecker” signals

The pioneering experimental measurements of these signals by Dr. Robert Beck and William Bise have shown just how deadly a potential may be possessed by the signals. These measurements have been performed in Eugene and Portland, Oregon; Los Angeles, California; Huntsville, Alabama; and several other locations. They have been particularly significant on the West Coast in and around Eugene, Oregon because of the presence of a direct current transmission line several hundred miles long, which has acted as a long wire antenna, picking up the signals and rebroadcasting them with appreciable gain in the vicinity.

Typically the signals may be found on, say, sixteen different carriers between 10 and 20 megahertz. Twelve of the carriers may appear normal, with normal side-bands, and the other four may have the carrier and both sidebands suppressed but still show the biologically significant modulation (for example, 10 hertz). On all sixteen channels a strong 10 hertz modulation may appear, all perfectly time synchronized and in phase channel-to-channel. The received signal from one of these carriers may be twenty-five to thirty times as strong as the earth's back-ground magnetic field, which is oscillating at approximately 7.0 to 7.5 hertz. Other complex modulation frequencies, many of them changing, are present on the various channels.

Normally, the brainwaves of mammals in an area are gently entrained by the normal oscillations of the earth's magnetic field and are oscillating along at, say, 7.5 hertz. Under continuous radiation from the Soviet woodpecker signals, a percentage (say 30 percent) of the mammalian brains may be captured by the 10 hertz modulation on the overpowering Soviet signals. These captured brains are now in forced oscillation and are riding along in phase with the 10 hertz Soviet modulation. In other words, a certain percentage of the human brains in an area will phase lock to the 10 hertz modulation, and that will effectively lock into those brains the sixteen carrier frequencies with their concomitant frequency mixes and complex signal modulations. If the signals are made much more powerful (for example, by focusing them onto one particular area or increasing the power of the transmitters), then a much higher percentage of brainwave entrainment can be accomplished in the targeted area. By simply adding on the desired modulation patterns, the Soviets can now pump material directly into the mind-brain-consciousness-life loops of the entrained brains. The weapon implications are enormous: raw emotion such as sheer terror or panic can be transmitted; death or disease patterns of all kinds can be transmitted; informational content (thoughts and ideas) can be impressed directly into the captured brains and minds and processed as if originating inside each brain itself. Indeed, as long as the Soviets keep the effects rather gentle, an entire population can be subtly influenced without governmental notice. Certain specific ELF frequencies can rapidly disable or even kill, and for wartime use those frequencies can be directly implanted in the captured brains themselves by the woodpecker signals. It appears that 1984 came a little early, and one can perhaps now understand Brezhnev's strange 1975 proposal to the SALT negotiations that we should also consider outlawing the development of new weapons more frightful than the mind of man has ever conceived.

The signals also have direct application against strategic targets other than personnel. For example, by incorporating the correct nested modulation pattern, the positron kindling (electron squelching) pattern can be impressed. By beaming through the earth (with a circularly polarized signal or with other complex modulation forms), a beam with a strong neutrinic component, which will squelch electrons and which will interact with great preference for strong voltages or strong magnetic fields, can be directed at a target complex. In the targeted area aircraft and communications systems can be disabled; automobile, truck, and other ignition systems will fail; power transformers and large motors and generators will fail; power transmission systems with their associated high voltages will fail; and so on. Even independent, self-contained emergency power equipment, such as at military installations and strategic missile sites, will fail to operate. Further, to some of these signals, the earth and the sea are as transparent as glass, and so underwater nuclear submarines can be electrically disabled on station before their missiles can be fired, while strategic missile complexes can be electrically disabled before their deadly nuclear missiles can be hurled toward their distant targets. If some submarines and land complexes succeed in firing missiles before being disabled, these missiles can be electrically dudded and scrubbed from the sky over Soviet territory while still far from their intended targets.

So, as stated at the end of *Excalibur*, the players are on stage and in place. The final act begins. One thing is sure—mankind will never be the same again after seeing the play!

Return to top of page

LASER WEAPON SYSTEMS

TURNING UP THE HEAT WITH SPEED-OF-LIGHT RESPONSE

LOCKHEED MARTIN
We never forget who we're working for®

LASER WEAPON SYSTEMS

Lockheed Martin has specialized in laser weapon system development for 40 years, with advancements in areas such as precision pointing and control, line-of-sight stabilization and adaptive optics – essential functions in harnessing and directing the power of a laser beam – and in compact, robust, spectrally beam-combined fiber laser devices that provide unmatched performance.

Lockheed Martin's directed energy (DE) laser program draws upon our proven expertise across a broad range of capabilities, as well as our spirit of innovation. Our experience includes:

- The beam-control/fire-control system for a megawatt-class laser that destroyed a ballistic missile in flight
- High-power spectrally beam-combined fiber lasers that provide the most efficient conversion of platform prime power into lethal power on target
- Unique SWIR component capabilities that significantly extend the ISR (intelligence, surveillance, and reconnaissance) range of the DE laser system
- Widespread weapons-system and platform integration experience across all DoD services

REVOLUTIONARY TECHNOLOGY

Laser weapons are a revolutionary technology because of the advantages of speed, flexibility, precision and low cost per engagement that are only possible with lasers.

These advantages apply to stand-alone DE laser systems as well as to weapon systems that combine DE and kinetic energy capabilities. In these cases, DE operates as a force multiplier, enabling the warfighter to counter a growing range of emerging threats.

DE weapons are characterized by:

- Very deep magazine
- Extremely low cost per engagement
- Speed of light delivery

These strengths mean that they are well-suited to countering large numbers of inexpensive, highly maneuverable threats that might otherwise exhaust the magazines of our current defensive kinetic energy (KE) weapons. By using DE and KE weapons together, the warfighter will be able to neutralize emerging swarming threats while reserving our most capable kinetic weapons to defend against our adversaries' largest and most hardened threats. DE weapon systems support multi-mission scenarios, and can be readily augmented by communications functions, as well as unprecedented intelligence, surveillance and reconnaissance range, and precision.

In addition, Lockheed Martin's directed energy weapons bring some important and unique advantages to the warfighter:

- A robust and highly-maintainable laser architecture that is designed to minimize life-cycle cost and to maximize up-time
- Output laser power that can be rapidly adjusted between low and maximum power to support disrupt, disable, and destroy capabilities
- Highly parallel laser architecture that supports graceful degradation by eliminating almost all single points of failure

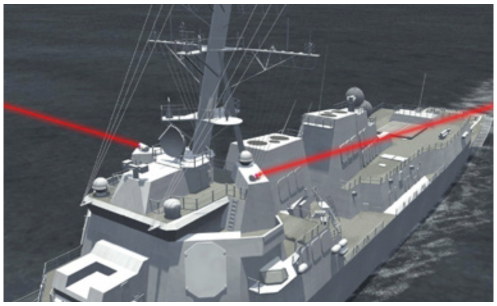

- Highest system efficiency demonstrated in any DE laser system; minimizing size weight and power requirements for DE platforms such as Army and Marine Corps ground vehicles, Navy DDG and LCS ships, and SOCOM AC-130
- Highest lethality laser to minimize engagement time

LOCKHEED MARTIN IS ADVANCING AND DEMONSTRATING A RANGE OF LASER WEAPON SYSTEM TECHNOLOGIES:

ADAM (Area Defense Anti-Munitions)

- Autonomous operations for rocket threats; accepts external sensor cue when required; capable of continuous operation
- Successfully engaged constrained and free-flying rockets, an unmanned aircraft systems (UAS) target in flight, and a small boat
- Affordable commercial-off-the-shelf (COTS) based system with very low cost-per-kill; deep magazine; scalable precision effects
- Capable of close-in defense (1-4 km)

ALADIN (Accelerated Laser Demonstration Initiative)

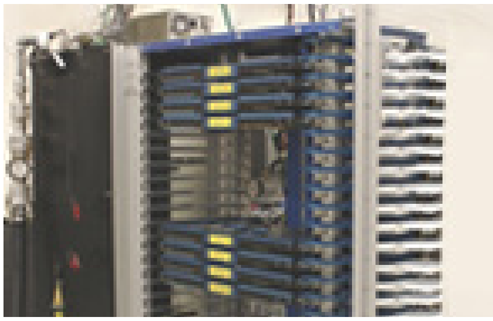

- 30-kilowatt laser made by combining many fiber lasers into a single, near-perfect-quality beam of light
- Uses approximately 50 percent less electricity than alternative solid-state laser technologies
- Spectral Beam Combining sends beams from multiple fiber laser modules, each with a unique wavelength, into a combiner that forms a single, powerful, high-quality beam

ATHENA (Advanced Test High Energy Asset)

- Represents highest power level documented by a laser weapon system of this type, while retaining excellent beam quality and electrical efficiency
- First field testing of an integrated 30-kilowatt single-mode fiber laser weapon system prototype
- Uses the proven high-energy laser weapon system architecture from our ADAM system, and incorporates the 30-kilowatt ALADIN laser

ABC Turret (Aero-adaptive Aero-optic Beam Control)

- Prototype turret with the ability to fire in any direction mounted on tactical aircraft
- First turret to demonstrate a 360-degree field of regard for laser weapon systems on an aircraft flying near the speed of sound
- Validated performance with nearly 60 flight tests conducted in 2014 and 2015 using a business jet as a low-cost flying test bed

RELI (Robust Electric Laser Initiative) for Army HELMTT (High Energy Laser Mobile Test Truck) Demonstrator

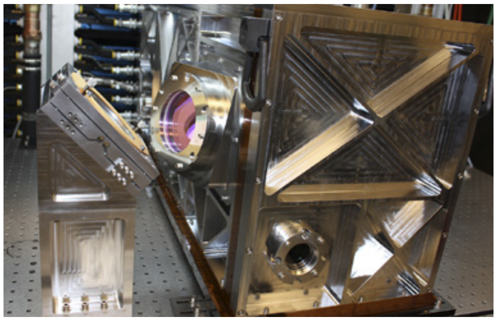

- Modular 60-kW laser in development for the U.S. Army
- Proven affordable weapon architecture that supports size, weight and power constraints for air, sea and land platforms
- Scalable design combines multiple kilowatt lasers to attain various weapon power levels
- Modularity results in a reliable and low maintenance laser system that minimizes single points of failure

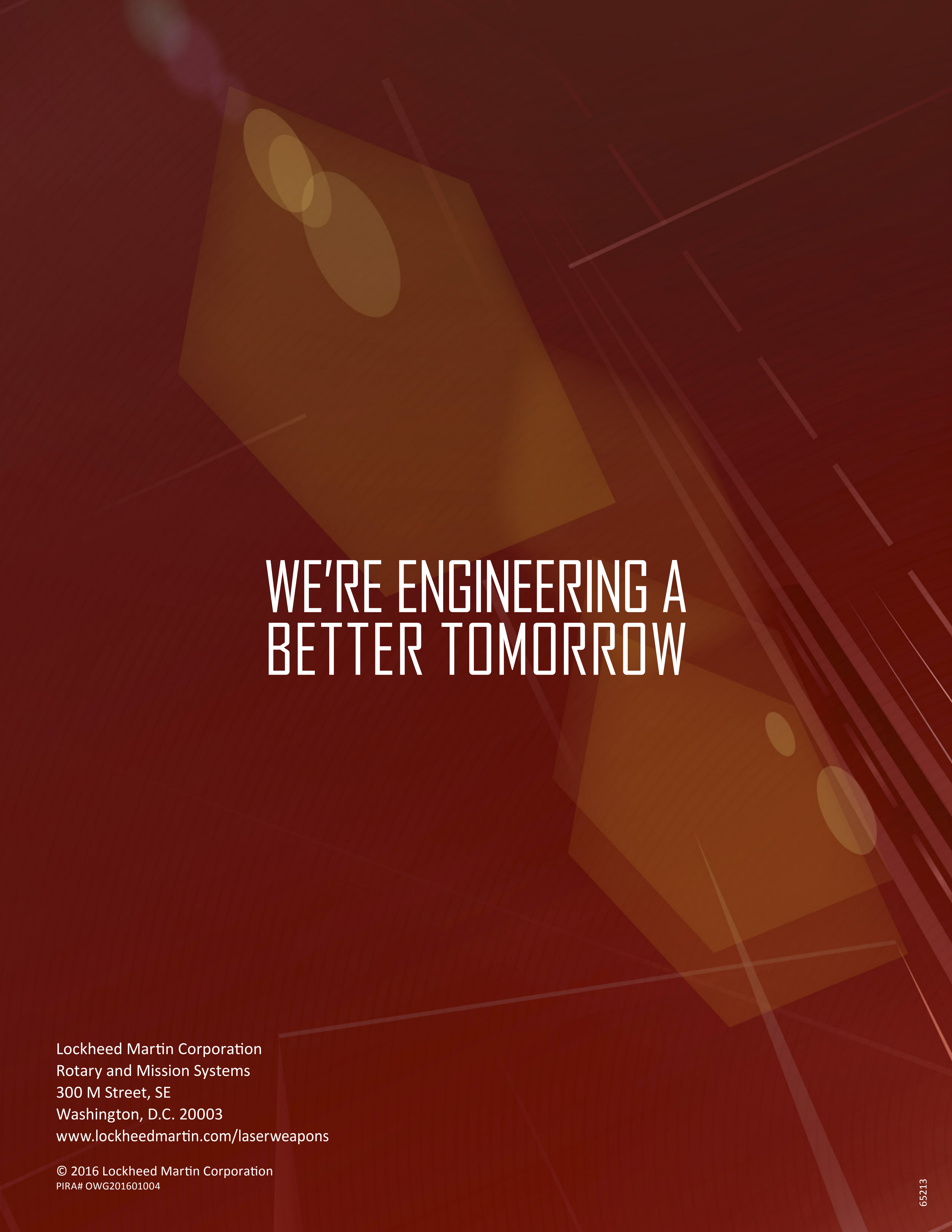

WE'RE ENGINEERING A BETTER TOMORROW

Lockheed Martin Corporation
Rotary and Mission Systems
300 M Street, SE
Washington, D.C. 20003
www.lockheedmartin.com/laserweapons

© 2016 Lockheed Martin Corporation
PIRA# OWG201601004

Axial Vircator for Electronic Warfare Applications

Libor DRAŽAN, Roman VRÁNA

Dept. of Radar, University of Defence, Kounicova 65, 662 10 Brno, Czech Republic

libor.drazan@unob.cz, roman.vrana@unob.cz

Abstract. *This paper deals with a high power microwave generator with virtual cathode – vircator in axial release for electronic warfare applications. The classification of directed energy weapons microwave (DEWM) is introduced together with basic block diagrams of a particular class of DEWM. In the paper, methods for designing vircator pulsed power supply, axial vircator structure, measurement methods and experimental results are presented. The vircator in electromagnetic ammunition is powered by magneto-cumulative generator and in weapons for defense of objects (WDO), it is powered by Marx generator. The possible applications of a vircator in the DEWM area are discussed.*

Keywords

Electronic warfare, directed energy weapons microwave, high power microwave, magneto-cumulative generator, Marx generator, pulse forming line, vircator.

1. Introduction

Electronic warfare (EW) is defined as the art and science of preserving the use of the electromagnetic spectrum for friendly use while denying its use to the enemy. Electromagnetic spectrum is, of course, reaching from DC to light (and beyond). Thus, electronic warfare covers the full radio frequency spectrum, the infrared spectrum, the optical spectrum, and the ultraviolet spectrum.

Electronic warfare has classically been divided into:

- Electronic warfare support (ES), which includes the receiving part of EW.
- Electronic attack (EA), which includes jamming, chaff, flares used to interfere with the operation of radars, military communication, and head-seeking weapons. Over the last years, EA also includes antiradiation weapons (ARW) and directed energy weapons (DEW).
- Electronic protection (EP), which is directed into design or operation of radars or communication systems to counter the effect EA.

A detailed classification of EW can be found in [1].

Today the directed energy weapons (DEW) are significant in terms of the electronic attack (EA). Considerable financial resources are directed especially to the research of DEW working in microwave band (directed energy weapons microwave − DEWM), which derive their benefits from incidence of electromagnetic pulses (EMP) on electronic devices. The EMP effect was first observed during the early testing of high altitude airburst nuclear weapons [2]. The effect is characterized by production of a very short (hundreds of nanoseconds) but intense electromagnetic pulse, which propagates away from its source with ever diminishing intensity, governed by the theory of electromagnetism. The EMP is in effect an electromagnetic shock wave. This pulse of energy produces a powerful electromagnetic field, particularly within the vicinity of the weapon burst. The field can be sufficiently strong to produce short lived transient voltages of hundreds of volts to kilovolts on exposed electrical conductors, such as wires or conductive tracks on printed circuit boards.

At present, a high intensity EMP can be successfully generate without using nuclear weapons. This fact is very significant especially for applications in the EW area. It is this aspect of the EMP effect which is of military significance, as it can result in irreversible damage to a wide range of electrical and electronic equipment, particularly computers and radio or radar receivers. Depending on the electromagnetic hardness of the electronics (a measure of the equipment's resilience to this effect) and the intensity of the field produced by the weapon, the equipment can be irreversibly damaged or even electrically destroyed. The damage inflicted is not unlike the damage experienced by exposure to close proximity lightning strikes, and may require complete replacement of the equipment, or at least substantial portions of it [3], [4].

2. Directed Energy Weapons Microwave

Directed energy weapons microwave (DEWM) generate very high-powered electromagnetic impulse (EMP) in the microwave band. From the generated signal bandwidth point of view, there are two classes of generators for the design and construction of DEWM in order to produce electromagnetic field with sufficient power:

- Narrowband – very similar to the radar transmitters; usually generate a modulated harmonic signal. This class is called the high power microwave (HPM).
- Ultra-wideband (UWB) – generates high power videosignal illuminated by a broadband antenna.

The application of directed energy weapons microwave (DEWM) determines their design. There are two groups of directed energy weapons microwave:

- Single shot DEWM,
- Repetitive pulsed power systems.

The electromagnetic weapon modes closely relate to the mission of these weapons. Electromagnetic ammunition usually uses single shot DEWM while weapons for defense of objects (WDO) usually use repetitive pulsed power systems. The typical pulse repetition frequency of DEWM is nearly one thousand pulses per seconds. Recently, there has been a trend of increase in the pulse repetition frequency beyond tens of thousands pulses per seconds. The actual DEWM construction depends on the concrete supposed combat use, on the required measure, range and so on. The following text will be dedicated to experimental verification of some technology usable in narrowband DEWM or HPM. Questions regarding classification, construction and possible use of generators for DEWM are answered in more detail in [4], [5] and [6].

2.1 Basic Block Diagram of DEWM

HPM design depends on the specific assumed application of DEWM. There are difference approaches to the design and realization of electromagnetic ammunition and to the design and realization of weapons for defense of objects (WDO). These differences mostly affect the choice of the power sources and the construction of modulators. In electromagnetic ammunition, a highly limiting factor is the demand for low weight and small proportions. This fact also sharply limits the choice of the power microwave tube.

Electromagnetic ammunition usually uses single shot DEWM. The single shot high-energy high-voltage pulsed source has to fulfill hard requirements on the high amount of the delivered energy with respect to the source volume [MJ/m^3]. The best solution of this requirement is a magneto-cumulative generator that reaches up to 8000 MJ/m^3. A block diagram of a HPM generator in electromagnetic ammunition is shown in Fig. 1.

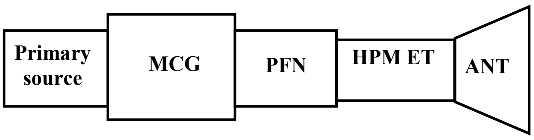

Fig. 1. Block diagram of a HPM generator in electromagnetic ammunition.

In the block diagram, the primary source feeds the stator coil of a magneto-cumulative generator (MCG), which creates a powerful magnetic field in the surrounding fittings filled with an explosive. At the moment of achieving the primary current peak value, the explosive is shooting. The explosion causes expansion of fitting that performs compression of the magnetic field inside the stator coil of MCG. Compression of magnetic field causes dramatic growth of the current flowing through MCG (hundreds of kiloamperes). The MCG output is connected to a pulse forming network that provides transformation of heavy current to an impulse high voltage (hundreds of kilovolts), which feeds a high power microwave electron tube (HPM ET). The generated microwave is emitted through antenna (ANT) to the space.

A block diagram of a HPM generator in weapons for defense of objects (WDO) is shown in Fig. 2.

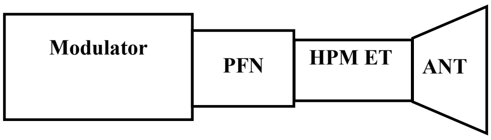

Fig. 2. The block diagram of HPM generator in Weapons for Defense of Objects (WDO).

In weapons for defense of objects (WDO), the HPM generator works similarly as the HPM generator in electromagnetic ammunition. Since WDOs usually generate power microwave pulses repetitively, a modulator with this capability has to be used instead of MCG. Modulators with Marx capacitor banks, Tesla transformers or transductors with fast opening switches are used most often. In this class of DEWM, other types of high power microwave electron tubes (HPM ET) are also used. They may generate microwave pulses with high pulse repetition frequency.

3. Energy Sources for Directed Energy Weapons Microwave

The electromagnetic pulse sources for DEWM are entirely based on energy acquired from the chemical energy of explosion that compresses the magnetic flux made by coil (MCG) or by discharging a capacitor bank. These two possibilities are the best solutions for generation of pulses with sufficient amount of energy in a short time. Other methods were used to generate pulses in the beginning of DEWM research or in laboratory.

3.1 Helical Magneto-Cumulative Generator

MCGs are entirely based on the energy acquired from chemical energy of explosion that compresses magnetic flux made by coil. MCG physical principles, classification and applications are described in more detail in [7]. The main part of research was focused on the possibility of using a magneto-cumulative generator for electromagnetic ammunition power supply. For that reason, 4 versions of low power MCGs were designed, realized and tested in order to verify their characteristics, construction principle, explosive initialization method and verifying a measurement method.

The tested magneto-cumulative generator in Fig. 3 consists of an aluminous cylindrical stator, inside is a 16 screw helical coil from 2 mm copper wire, a copper armature tube with 38 mm outer diameter and 4 mm wall thickness filed by octogen (explosive). The explosive is electrically initiated. The load coil represented by a copper strip in shape of a loop connects the stator with the armature. In the right side of stator, there is a crowbar switch made of a brass plate to perform disconnecting of the primary current source. An insulator circle is responsible for separation of the stator and the armature.

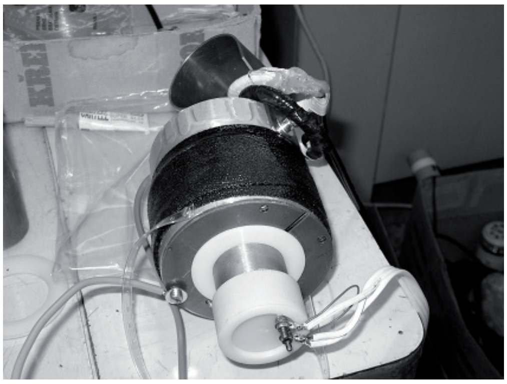

Fig. 3. Low power MCG.

Low power magneto-cumulative generator characteristics:

C_1 = 190 μF − primary capacitor capacity

U_0 = 800 V − voltage of capacitor

L_G = 13.1 μH − stator coil inductance

R = 0.0175 Ω − stator coil resistance

T_0 = 3.13 x 10^{-4} s − primary current period

I_m = 2750 A − maximum primary current

T_d = 73 μs − explosive initiation delay

Fig. 4 presents the low power magneto-cumulative generator output current time history calculated from the voltage measured by Rogowski coil. The primary current I_m = 2750 A gives the generator peak output current I_{out} = 27 kA. The detailed test description and measured results can be found in [8], [9], [10] and [11]. The purpose of the mentioned tests was:

- Verify the method of generating a high energy pulse by magnetic flux compression using explosive;
- Solve the primary current magneto-cumulative generator power supply;
- Find the peak primary current and an explosive initiation synchronization method;
- Find a high current measurement method.

All the requirements were met.

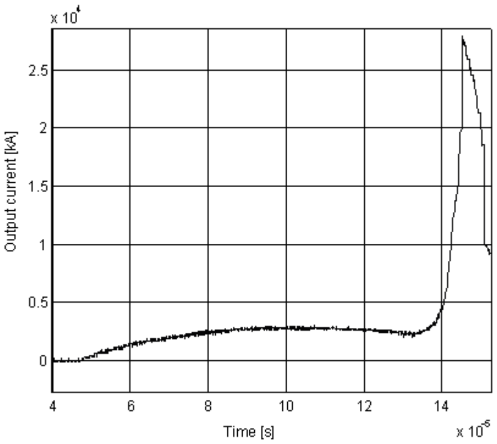

Fig. 4. Low power magneto-cumulative generator output current time history.

To use a magneto-cumulative generator for single shot high power high voltage pulse generator, it is necessary to obtain a high-level output current. Because of that reason, a middle power MCG was designed and realized. It is a two-stage helical magneto-cumulative generator using a dynamic transformer to increase the output voltage. In Fig. 5, there is a schematic plot of the generator's first stage. The magneto-cumulative generator power supply primary current, the initiation of the explosive and the output high current measure was analogical to the low power magneto-cumulative generator described above.

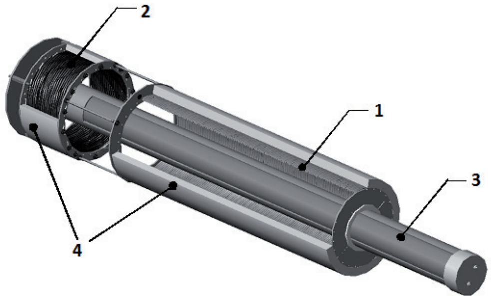

1 – First stage coil, 2 – Load coil, 3 – Armature with explosive, 4 – Insulator

Fig. 5. Schematic drawing of first stage of middle power MCG.

Fig. 6 shows a photo of the middle power magneto-cumulative generator's first stage prepared to the test.

Magneto-cumulative generator characteristics:

C_1 = 125 μF − primary capacitor capacity

U_0 = 5000 V − voltage of capacitor

L_G = 323 μH − stator coil inductance

T_0 = 3x10^{-3} s − primary current period

I_m = 2650 A − maximum primary current value

L_L = 300 nH − load inductance

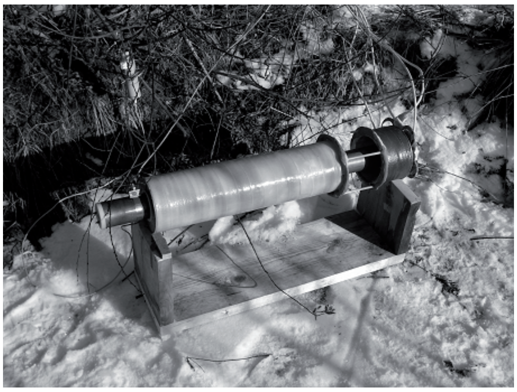

Fig. 6. The picture of middle power MCG prepared to testing.

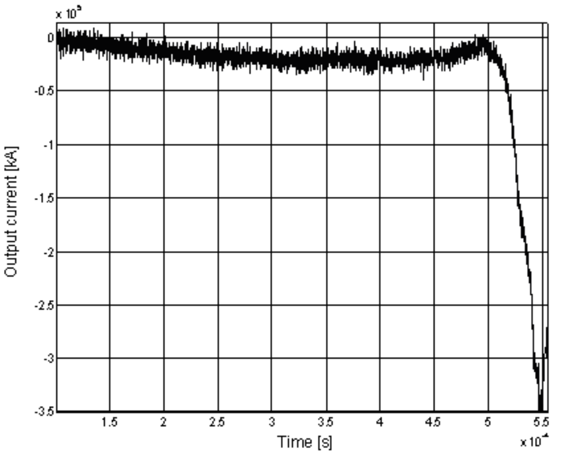

Fig. 7. Middle power magneto-cumulative generator output current time history.

In Fig. 7, the generator output current time history measured during test is shown. The primary current $I_m = 2650$ A gave the generator peak output current $I_{out} = 350$ kA.

The next step in the magneto-cumulative generator research is the implementation and verification of a dynamic transformer. The dynamic transformer's role is an increase of the output voltage to about 70 kV. Magneto-cumulative generator's second stage supplies a load coil through an electric breaker. The beaker interrupts the current in the load circuit in a very short time to enable transfer of the energy from the magnetic field to the high voltage pulse. These activities will be done in the next step of the research and development. The further trends are described in [12].

3.2 Pulsed Power Supply With Capacitive Storage of Energy

For some pulse applications in weapons for defense of objects (WDO), it is desirable to couple the Marx generator directly to the vacuum diode; however, the pulse rise time is then limited by the Marx inductance and capacitance, and the impedance of the generator is greater than typically several tens of ohms. In order to produce short, fast-rising, low-impedance beam outputs, it is customary to use the Marx to charge a pulse-forming line (PFL). Although the PFLs may be constructed in a variety of shapes (strip, coaxial, radial, etc.), they are typically used in only two types of circuits— the simple transmission line, and the double, or Blumlein line. In contrast to the simple transmission line, an alternate circuit invented by A. D. Blumlein is capable of producing an output pulse into a matched load that equals the charge voltage. A cylindrical version of the Blumlein circuit fabricated by our team is represented in Fig. 8. It consists of three coaxial cylinders with the intermediate cylinder being charged by the Marx generator. The Marx generator is shown in Fig. 9. The center cylinder is connected to the outer grounded cylinder by an inductor. Ideally the inductor acts as a short during the charge cycle, and then as an open for the short duration of the output pulse [13]. Pulsed power supply parameters are shown in Tab. 1.

Marx generator		Pulse forming line	
Number of stage	n=18	Impedance	9,1 Ω
Capacity of 1stage	100 nF	Length	1 m
Input voltage	U_{in}=25 kV	Pulse duration	60 ns
Output voltage	U_{out}=450 kV	Working medium	water

Tab. 1. Pulsed power supply parameters.

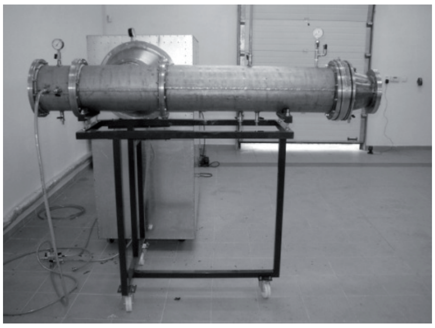

Fig. 8. Cylindrical version of the Blumlein circuit.

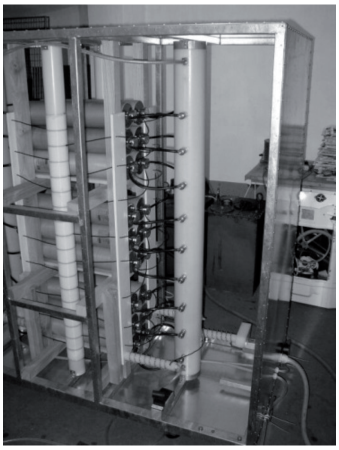

Fig. 9. Marx generator.

4. High Power Microwave Generator

A wide range of HPM devices exists. Relativistic klystrons, magnetrons, slow wave devices, reflex triodes and vircators are all examples of the available technology base [5]. From the perspective of an electromagnetic weapon or warhead designer, the device of choice will be the vircator at this time. The vircator is mainly a one shot device capable of producing a very powerful single pulse of radiation, yet it is mechanically simple, small and robust, and can operate over a relatively broad band of microwave frequencies. The physics of the vircator tube are substantially more complex than those of the preceding devices. The fundamental idea behind the vircator is that of accelerating a high current electron beam against a foil or a grid anode. Many electrons will pass through the anode, forming a bubble of space charge behind the anode. Under the proper conditions, this space charge region will oscillate at microwave frequencies. If the space charge region is placed in a resonant cavity which is appropriately tuned, very high peak powers may be achieved. Conventional microwave engineering techniques may then be used to extract the microwave power from the resonant cavity. Because the frequency of oscillation is dependent upon the electron beam parameters, vircators may be tuned or chirped in frequency, where the microwave cavity will support appropriate modes. Power levels achieved in vircator experiments range from 170 kW to 40 GW over frequencies spanning the decimeter and centimeter bands [5].

4.1 Vircator Operation Fundamentals

The basic idea of the vircator is to accelerate a dense flush of an electron beam against a grid or a foil anode. Plenty of electrons pass through the anode and form a region of a space charge behind the anode called ”virtual cathode“. This region of a space charge at corresponding conditions can oscillate in a region of microwave frequencies. It is possible to tune the vircator in a broad band of frequencies using only a change of a space charge density. There is no necessity to have an external magnetic field for a correct vircator function. In Fig. 10, an axial vircator is shown. Electron beam passes through the foil or the grid anode. Microwave power is brought out axially, too. Frequency changes appear in vircator with standard geometry when the distance between the anode and the cathode gets smaller due to filling the working space by plasma. Efficiency of a standard geometry vircator is ordinarily about ones percent.

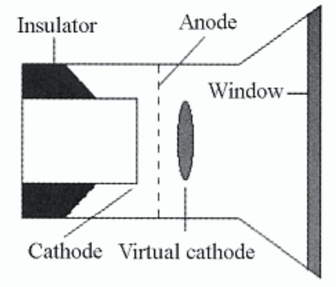

Fig. 10. Schematic drawing of axial vircator.

Despite its low efficiency, vircator is very attractive for army applications because it is very simple to be made, it is a compact device and there it is no need to have an external magnetic field. For more information about construction and properties of various types of vircators you can see e. g. [5], [13] and [14].

4.2 Mathematical Description of Vircator Operation

For microwave frequency generation, it is necessary to meet a number of conditions relevant to the power supply and geometric proportions of a vircator electrode. Determination of these conditions results from [15] and we use simplified geometry displayed in Fig. 11 to derive the mathematical characterization.

The most common case is that a pulse duration τ is much longer than a beam transit time across cavity $\tau >> L/C$.

We assume that an external axial magnetic field obstructs transverse electron motion. The space charge of the beam makes a negative potential energy $e\Phi$ in drift space which breaks the electrons. If the space charge potential reaches the value of accelerating voltage (electrons are stopped in beam), the beam with bigger current cannot expand. There is a value of current which causes stopping of electrons. This value is called the vacuum critical current. There has to be the above mentioned critical current to make the vircator generate microwave oscillations.

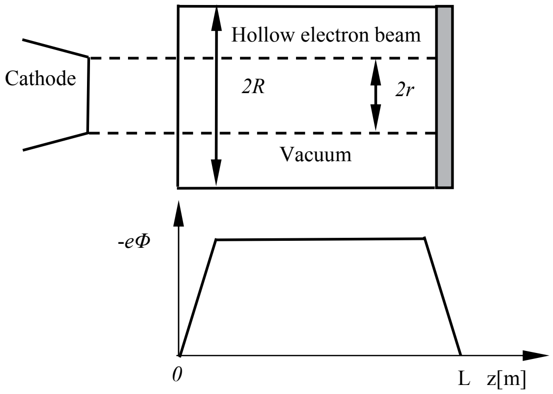

Fig. 11. Simplified geometry of axial vircator.

We suppose a hollow beam; whose charge is concentrated in a thin layer with radius r. Behind the transition space (which is comparable to the radius of chamber – backward conductor), there is nearly homogenous potential. It is possible to determine the potential quantity with consideration that it is a coaxial capacitor and its inner cylinder is saturated by beam charge.

Capacity of the coaxial capacitor is given by

$$C = \frac{2 \cdot \pi \cdot \varepsilon_0}{\lg\left(\frac{R}{r}\right)} \tag{1}$$

where ε_0 is the vacuum electric permittivity, R is the chamber radius, r is the beam radius.

The potential is given by the equation

$$\Phi = \frac{Q}{C} \tag{2}$$

where Q is the charge linear density, C is the capacity per length unit. We can define the charge linear density as

$$Q = \frac{I}{v} \tag{3}$$

where v is an electron module velocity and the current is defined by the equation

$$I = e \cdot n \cdot v \cdot S \tag{4}$$

where e is the charge of an electron, n is the linear electron density, S is the profile of the electron beam.

When (1) and (3) are substituted into (2), we can obtain expression for the potential in the form

$$\Phi = \frac{Q}{C} = \frac{I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot v}. \tag{5}$$

The electron velocity v is associated with its initial energy and potential by the energy preservation law

$$m_e \cdot \gamma_0 \cdot c^2 + e\Phi = m_e \cdot \gamma \cdot c^2 \tag{6}$$

where γ_0 is the initial relativistic factor of the beam, γ is the relativistic factor inside a system reduced due to potential.

The relativistic factor inside a system can be expressed by the equation

$$\gamma = \gamma_0 + \frac{e\Phi}{m_e \cdot c^2} \tag{7}$$

where m_e is the electron mass and c is the speed of light. The relativistic factor can also be expressed by

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \ . \tag{8}$$

The electron velocity in drift space v can be specified from following equations:

$$\gamma^2 = \left(\gamma_0 + \frac{e\Phi}{m_e \cdot c^2}\right)^2 , \tag{9}$$

$$v = c \cdot \sqrt{1 - \frac{1}{\left(\gamma_0 - \frac{e\Phi}{m_e \cdot c^2}\right)^2}} \ . \tag{10}$$

Substituting (10) into (3) and multiplying by $(e\cdot v) / (m\cdot c^3)$, we can get

$$F(\Phi) = \frac{e \cdot \Phi}{m \cdot c^2} \cdot \sqrt{1 - \frac{1}{\left(\gamma_0 - \frac{e\Phi}{m_e \cdot c^2}\right)^2}} = \frac{e \cdot I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3}. \tag{11}$$

The function $F(\Phi)$ has its peak value under the conditions describes by

$$\frac{e \cdot \Phi}{m \cdot c^2} = \gamma_0 - \gamma^{\frac{1}{3}} \ . \tag{12}$$

The evaluation is written in the form

$$F(\Phi) = \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} = \frac{e \cdot I \cdot \lg\left(\frac{R}{r}\right)}{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3} \ . \tag{13}$$

There is no possibility to find the solution of (11) for bigger values of function $F(\Phi)$. According to (13), the critical vacuum current is

$$I_{cr} = \frac{2 \cdot \pi \cdot \varepsilon_0 \cdot m \cdot c^3}{e \cdot I \cdot \lg\left(\frac{R}{r}\right)} \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} \ . \tag{14}$$

When we substitute the known constants (like π = 3.1415), we obtain the equation

$$I_{cr} = \frac{8.5}{\lg\left(\frac{R}{r}\right)} \sqrt{\left(\gamma^{\frac{2}{3}} - 1\right)^3} \tag{15}$$

where I_{cr} is the critical vacuum current [kA].

For vircator implementation, energy of electrons of about 500 keV is considered, which corresponds to the relativistic factor of $\gamma_0 = 2$. Thus planar diode current limited by space charge follows the three-half Child-Langmuir law

$$I \approx \frac{U^{\frac{3}{2}} \cdot S_k}{d_{ka}^{\ 2}} \tag{16}$$

where U is the voltage between anode and cathode, S_k is the surface of cathode, d_{ka} is the distance between anode and cathode.

An important characteristic of a vacuum planar diode is its impedance, which is given by

$$Z = 1.36 \cdot 10^5 \cdot U^{-\frac{1}{2}} \cdot \frac{d_{ka}^2}{r_k^2} \tag{17}$$

where r_k is the radius of the cathode.

The vacuum critical current is the maximal current that can distribute itself through vacuum. It means it is possible to measure the same current until the distance is bigger than the diameter of beam. It is just the same value when the potential of space charge of electron beam is equal to the accelerating voltage. That is the limit where the virtual cathode is generated and the vircator starts working. The generated frequency depends on the plasma frequency of the electron beam. The plasma frequency is determined by the electron beam current density. The plasma frequency is the frequency of electrons in space charge field oscillations. The electrons can oscillate due to the influence of repulsive force which takes effect between particles with the same charge. The plasma frequency depends on the concentration of electrons in the beam. The current density is given by the number of charges which are able to pass through the surface of 1 m^2 per one second. Thus we can calculate the current density i [Am^{-2}]

$$i = e \cdot n_V \cdot c_e \qquad (18)$$

where e is the electron charge, n_V is the electron volume density and c_e is the electron velocity.

Plasma frequency of the electron beam is approximately given by

$$f_{pe} = 9 \cdot 10^7 \cdot \sqrt{i} \ . \qquad (19)$$

4.3 Axial Vircator Design

Following the theoretical assumptions given above, the axial vircator with grid anode was fabricated. The structure of the grid anode enables repeated operation in contradistinction to vircator with a foil anode. The cathode diameter $r_k = 17.5$ mm and the working chamber radius $R = 40$ mm were chosen in order to decrease the value of the vacuum current. The critical vacuum current is $I_{cr} = 4.63$ kA in this configuration. The vircator impedance matching to the forming line is a very important requirement for vircator's correct operation. Because of that, a cathode with a possibility to change the distance d_{ka} between cathode and anode is fabricated. The cathode is made as a cylindrical carbon block. The anode grid is made from pyrolitic carbon.

In Fig. 12, there is section picture of the built-up axial vircator with the possibility to the change distance between anode and cathode. In Fig. 13, there is a view of the fabricated vircator. The vircator is fed by a pulse generator formed by Marx generator and pulse forming line described in Section 3.2. The connection between the axial vircator and the output of pulse forming line is shown in Fig 14. Vircator voltage is measured by a cylindrical capacitive probe, which is a part of pulse forming line output circuit. Vircator current is measured by Rogowski coil placed in the outer cylinder of the pulse forming line. Vircator voltage time history is displayed in Fig. 15 and vircator current time history is displayed in Fig. 16.

Compared to other types of microwave tubes, the microwaves generated by an axial vircator cover a relatively wide frequency band. In Fig. 17, the frequency spectrum of the signals generated by the realized axial vircator is shown. The dominant frequencies of the generated microwave signals are placed in the band of 0.5 to 1.5 GHz. The frequency spectrum was acquired on the basis of information measured by a digital oscilloscope Tektronix TDS 7704B connected to broadband antenna EMCO 3115. For the measurement the microwave output power of the vircator, a measuring device was designed and made, using calorimetric method for the power measurement. A detailed power sensor description can be found in [18]. A more detailed description the of axial vircator research results is given in [6], [14], [16] and [17].

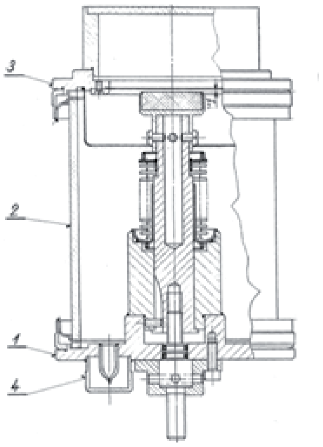

1 cathode contact ring, *2* ceramic insulator, *3* anode contact ring, *4* pipe for vacuum pumping shelter

Fig. 12. Section picture of built-up axial vircator with possibility to change distance between anode and cathode.

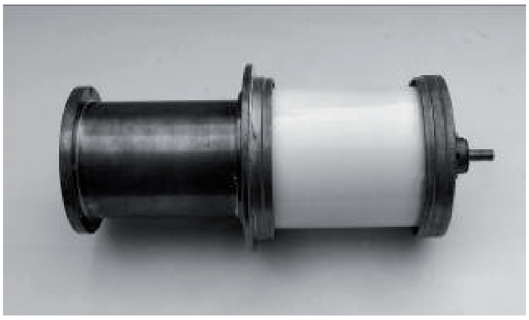

Fig. 13. View of fabricated axial vircator with adjustable distance between anode and cathode.

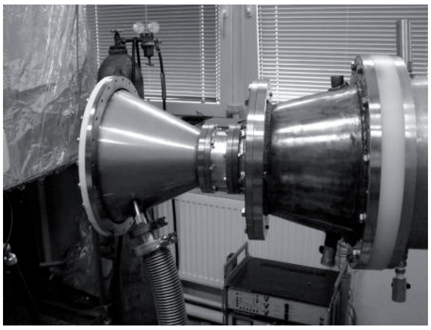

Fig. 14. View of fabricated axial vircator connected to the pulse forming line.

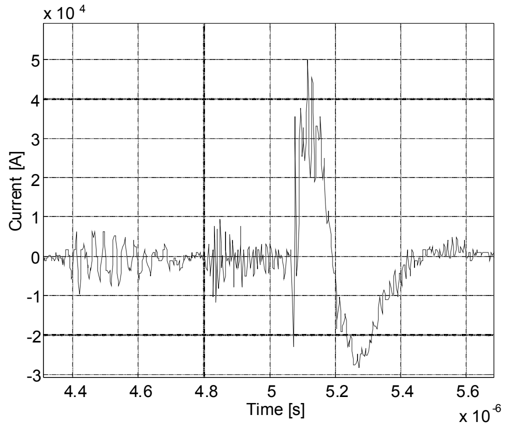

Fig. 15. Vircator current time history.

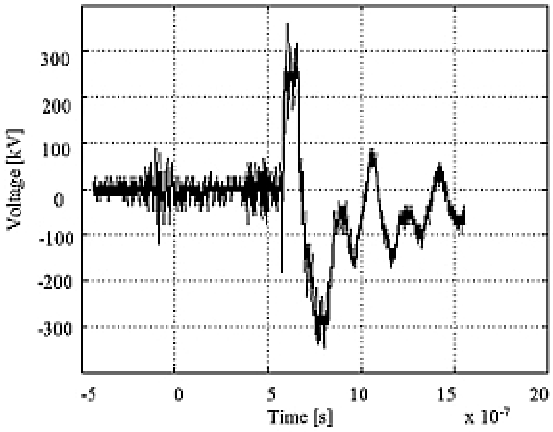

Fig. 16. Vircator voltage time history.

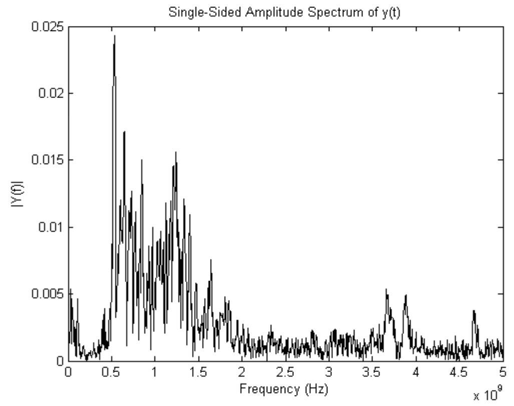

Fig. 17. Frequency spectrum of signals generated by realized axial vircator.

5. Conclusion

The research of directed energy weapons microwave validated substantial technology applicable in electromagnetic ammunition and in weapons for defense of objects (WDO). The axial vircator is the most important component part applicable in both categories of DEWM. Dominant frequencies of the generated microwave signal are placed in band of 0.5 to 1.5 GHz. To fulfill the vircator minimal proportion requirements to be able to use the vircator as a DEWM, it is necessary to shift the vircator dominant frequencies to the 3 GHz band. This requirement can be reached by arranging the cathode size efficiently and consequentially optimizing the distance between anode and cathode. Great attention also has to be paid to increase the specific dielectric strength of the output window of the axial vircator.

Acknowledgements

Research of the directed energy weapons microwave substantial technology was realized within the scope of ŽELETON, APPM and GENVLN research projects supported by the Ministry of Defense of Czech Republic.

References

[1] ADAMY, D. *EW101: A First Course in Electronic Warfare*. Norwood: Artech House, 2001.

[2] GLASSTONE, S., DOLAN, P. *The Effect of Nuclear Weapons.* 3rd ed. USA: United States Department of Defence, 1977.

[3] COOP, C. *An Introduction to the Technical and Operational Aspects of the Electromagnetic Bomb.* Canberra: RAAF Air Power Studies Centre Technical Reports, 1996.

[4] TAYLOR, C. D., GIRI, D. V. *High-Power Microwave Systems and Effects*. Washington: Taylor and Francis, 1994.

[5] BARKER, R. J., SCHAMILOGLU, E. *High-Power Microwave Sources and Technologies.* IEEE Press, 2002.

[6] DRAŽAN, L. *The Chosen Problems of Directed Energy Weapons Radio Frequency and Microwave Design.* Inaugural dissertation. Brno: University of Defence, 2008. 144 p. [in Czech].

[7] ALGILBERS, L. L., GRISHNAEV, I., TKACH, Y. *Magnetocumulative Generators.* New York: Springer-Verlag, 2000.

[8] DRAŽAN, L., ZAJONC, J. *Magneto-cumulative generator test results - version G3 and G4 (22.10.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[9] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results, ver. G3 and G4 (26.11.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[10] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results – ver. G3 (9.12.2002).* APPM project technical report. Brno: Military Academy, 2002. [in Czech].

[11] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Test Results – ver. G3 (7.1.2003).* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[12] DRAŽAN, L., ZAJONC, J. *Magneto-Cumulative Generator Optimisation Possibility.* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[13] MILLER, R., B. *An Introduction to the Physics Intense Particle Beams.* New York: Plenum Press, 1982.

[14] BEZDĚK, M., TECL, S., DRAŽAN, L. *Report Concerning the Vircator Research Results.* Research report. Vyškov: VOP-026 Šternberk, 2005. [in Czech].

[15] ŠUNKA, P. *High Power Pulse Generation*. Lectures collection. Prague: Academy of Sciences of the Czech Republic, Institute of Plasma Physics, 2004. [in Czech].

[16] DRAŽAN, L.,ZAJONC, J. *Test of Pulse Forming Line and Design of Microwave Generator with Virtual Cathode.* APPM project technical report. Brno: Military Academy, 2003. [in Czech].

[17] DOSTÁL, F., PLCH, Z. *GENVLN - Measurement.* Technical report. Vyškov: VOP 026 Šternberk, 2008. [in Czech].

[18] DREXLER, P., FIALA, P. Methods for high power EM pulse measurement. *Radioengineering*, 2006, vol. 15, no. 4, p. 43-49.

About Authors...

Libor DRAŽAN was born in Strakonice, Czech Republic, in 1960. He received his Ing. (M.Sc.) and CSc. (Ph.D) degrees from Brno Military Academy in 1984 and 1998, respectively. Since 1990, he has been with the Dept. of Radar at Military Academy and University of Defence in Brno, as an assistant professor (1990 to 2002), head of Dept. (since 2003) and associate professor (since 2008). His research interests include active and passive radar theory, electronic warfare and directed energy weapons microwave.

Roman VRÁNA was born in Havířov, Czech Republic, in 1972. He received his Ing. (M.Sc.) degree from Brno Military Academy in 1995. Since 1999, he has been with the Dept. of Radar, Military Academy and University of Defence in Brno, as an assistant professor (since 2008). His research interests include active radar theory and electronic warfare.

DARPA NESD

INDUSTRY GROUP

The *Biological Technologies Office (BTO)* of the *Defense Advanced Research Projects Agency (DARPA)* is sponsoring an Industry Group to support the *Neural Engineering System Design (NESD) Program*.

NESD is a highly interdisciplinary program requiring extensive integration of new research and technology. To facilitate the development and adoption of NESD technologies, DARPA has organizedan industry group for potential proposers. Members of this group have agreed to provide access to rapid prototyping and manufacturing of advanced neuro-engineering components, including state-of-the-art electronics, photonics, computing, assembly and packaging.

This document contains all responses to Special Notice 16-17 as of March 10, 2016. Any updates will be posted at **http://www.darpa.mil/work-with-us/opportunities**.

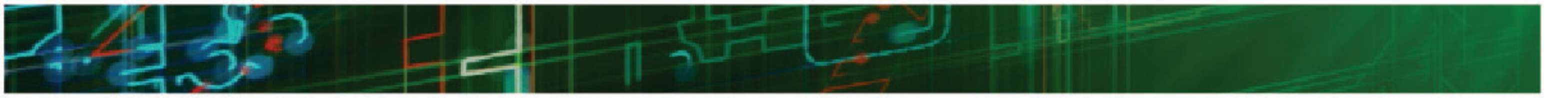

ALLEN INSTITUTE for
BRAIN SCIENCE

Allen Institute for Brain Science

Primary Contact

Christof Koch, Ph.D. / christofk@alleninstitute.org

About

The Allen Institute for Brain Science is an independent, non-profit medical research organization that has embarked on a concerted, ten-year effort, involving more than 300 scientists, engineers, technologists and other specialists, to understand the cerebral cortex in the mouse and in the human at the cellular level. We are built upon the notion of Big Science, Team Science and Open Science in the service of Basic Science with an explicit mission to accelerate understanding of the human brain in health and disease. Our data are freely and publicly available at www.brain-map.org.

Capabilities

- We seek to experimentally and computationally characterize the components (cell types) of the rodent visual thalamo-cortical system at the molecular, transcriptional, electron-microscopic, electrophysiological, morphological and functional levels to study how these components integrate information to generate high-level cognitive functioning such as object recognition.
- To study such high-level functioning, we use various in vivo perturbation and monitoring techniques (e.g. optical stimulation, extracellular depth recordings, wide and narrow-field Ca-imaging, etc.), very high density silicon probes (364 probes per shaft) as well as state-of-the art computer models that capture cortical information processing at varying levels of granularity.
- To understand the human brain, the Institute is also engaged in cell typing from human brain tissue harvested from patients undergoing brain surgery.
- Thus, we are routinely recording intracellularly from single and multiple cells from ex vivo human lateral temporal cortex and as well as performing morphological characterization on these same cells).

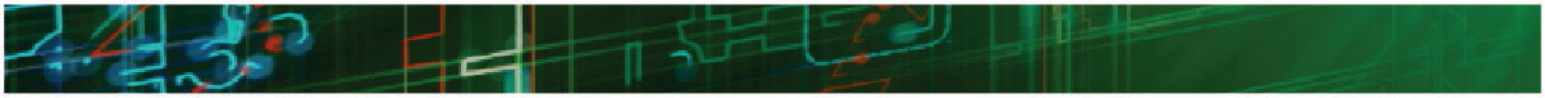

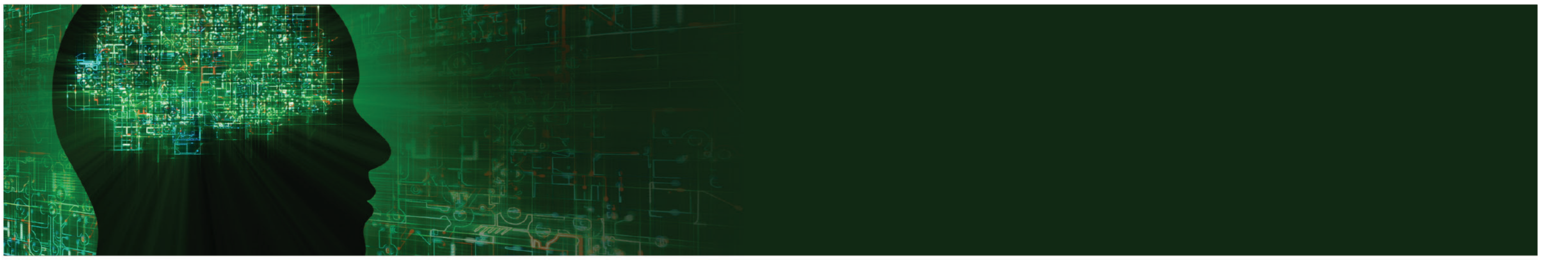

Capabilities

- We also have extensive experience with the genomic and transcriptional architecture of the human brain.
- We are interested in applying this knowledge for medical and translational purposes and to help facilitate advancement of CNS disorder therapies.

Facilities II

The Allen Institute headquarters, a 270,000 square foot life sciences building, is located in Seattle's South Lake Union neighborhood, the city's biotechnology hub. We are a high-throughput facility, home to custom-designed robotic systems, automated technologies, premier bench laboratory space and all the high-powered computing hardware and software necessary for processing petabytes of data.

Additional Contacts

Erica Sessle, MPH, MPhil / ericas@alleninstitute.org
Jennifer Pawlosky / jenniferp@alleninstitute.org

Bionics Institute

Primary Contact
Rob Shepherd / rshepherd@bionicsinstitute.org

Capabilities

Medical device safety and efficacy testing (in vitro and in vivo)

- Design, rapid prototyping & validation testing
- Accelerated aging & flexion testing
- Electrochemical impedance spectroscopy
- Acute feasibility studies
- Chronic safety and biocompatibility studies
- Drug delivery techniques
- Histology & data analysis
- Scanning electron microscopy of implanted materials
- Preclinical animal model development

Custom device prototyping & fabrication

- Design and fabrication of custom electrode arrays and surgical tools
- Use of FDA approved materials

'First in human' clinical trials

Clinical and Scientific Track Record

- 35 years of internationally recognised experience & scientific excellence
- Extensive clinical, academic & commercial partnerships
- Proven research & clinical development pathways

Staff and Facilities

- Experienced and qualified scientists, engineers and clinicians
- Co-located in a major Australian hospital
- Purpose built equipment and facilities

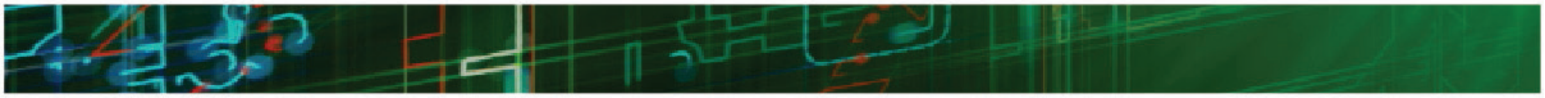

BOSTON
MICROMACHINES
CORPORATION

Boston Micromachines Corp

Primary Contact

Paul Bierden / pab@bostonmicromachines.com

About

Boston Micromachines Corporation (BMC) is the leading provider of microelectromechanical systems (MEMS) deformable mirror products and has expertise in the design of adaptive optics systems. The company's suite of award-winning compact DM products is the most cost-effective, highest performance mirrors in the market today. By applying wavefront correction, BMC devices can be used in a variety of applications which include laser beam shaping, microscopy, astronomy, and vision science. Located in Cambridge, MA, BMC is privately held and offers custom-designed manufacturing services in addition to its portfolio of standard DM products and adaptive optics systems.

Capabilities

The specific capabilities/services the company provides in relation to the NESD falls into two categories:

Deformable mirrors systems for use in neurological imaging instruments

- MEMS deformable mirrors ranging in size from 140 – 1000 actuators
- Mirror surface geometries of continuous, segmented, and hexagonal tip-tilt
- High speed drive electronics

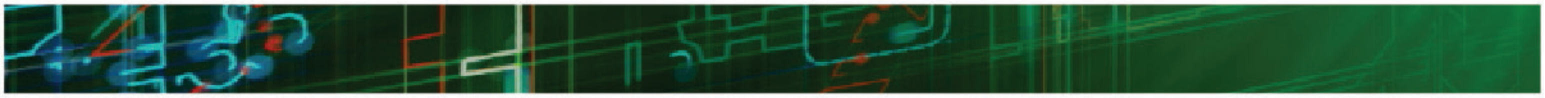

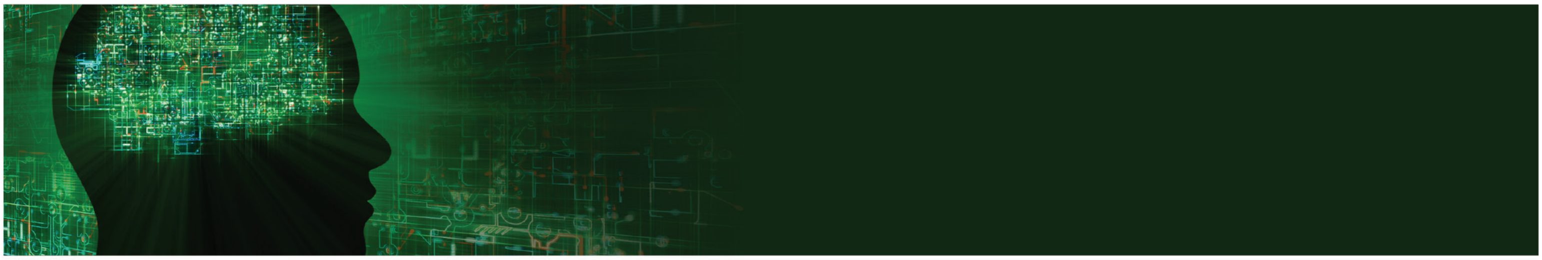

These devices are currently being used in in vivo neural imaging systems at various premier research institutions around the country. Boston Micromachines also provides research and development services for advancing the deformable mirror and drive electronics technology and design to better fit specific applications.

- Advanced imaging systems utilizing adaptive optics to improve resolution and contrast in in vivo subsurface imaging application (i.e. deep penetration brain imaging)

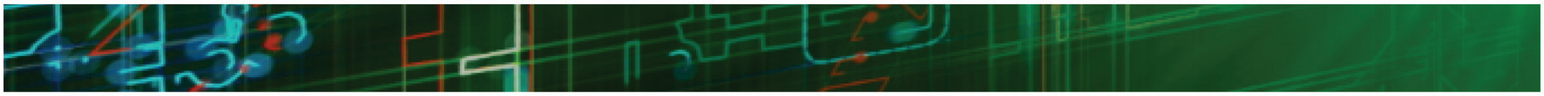

Boston Scientific
Advancing science for life™

Boston Scientific

Primary Contact

Rafael Carbunaru / Rafael.Carbunaru@bsci.com

Capabilities

- Lead prototyping
- Firmware modifications to the implantable pulse generators (IPG)
- Software modifications to the computer programmers (CP)
- Use of advanced computational models and predictive algorithms for parameters and programming guidance for neurostimulation

Additional Contacts

- Stephen Carcieri / Stephen.Carcieri@bsci.com
- Hemant Bokil / Hemant.Bokil@bsci.com
- Michael Moffitt / Michael.Moffitt@bsci.com

Chronocam

Primary Contact

Luca Verre / lverre@chronocam.com

Capabilities

- Extensive know-how in designing, developing and industrialising asynchronous, event-based, bio-inspired vision sensors
- Expertise in developing and customising event-based, bio-inspired computer vision algorithms for a large number of applications spanning from autonomous navigation, security and surveillance, aerospace and defence to biomedical devices and smart objects
- System integration capabilities: development of camera systems relying on Chronocam's proprietary vision sensors and development of embedded software porting Chronocam’s proprietary computer vision algorithms on a large number of processing platforms (such as: CPUs, GPUs, FPGA)

Additional Contacts

Christoph Posch / cposch@chronocam.com
Xavier LaGorce / xlagorce@chronocam.com
Thomas Finateau / tfinateu@chronocam.com

Cirtec

Primary Contact

Heather Dunn / Heather.Dunn@cirtecmed.com

About Cirtec

Cirtec Medical has been providing contract design, development, and manufacturing services to the medical device industry since our founding in Massachusetts in 1986. Today, Cirtec has over 150 employees in our two facilities on the east and west coasts.

Cirtec has supported active implantable developers throughout our history. Early projects focused on hermetic welding and testing services for implantable neurostimulation devices, with service offerings expanding to provide full design and development. In all, Cirtec has worked with over 60 different companies and institutions developing active implantable devices, with over 30 in the neuromodulation space. We work with companies and institutions of all sizes, from early stage starts-ups to large, well-established medical device manufacturers. We also work frequently with research laboratories and academic groups; our current projects include development and manufacturing of an implantable neurostimulator used for research and recently approved for clinical study under a feasibility IDE.

Capabilities

Cirtec is ISO 13485 certified and FDA registered.
Cirtec supports the full product development cycle from conceptual development and requirements definition through clinical and commercial manufacturing.

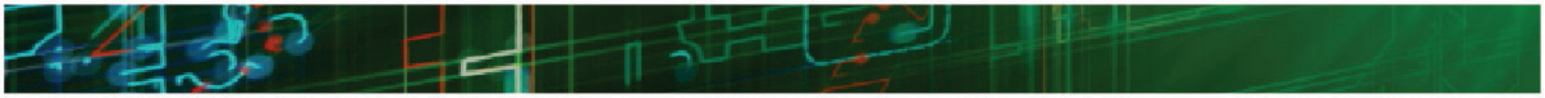

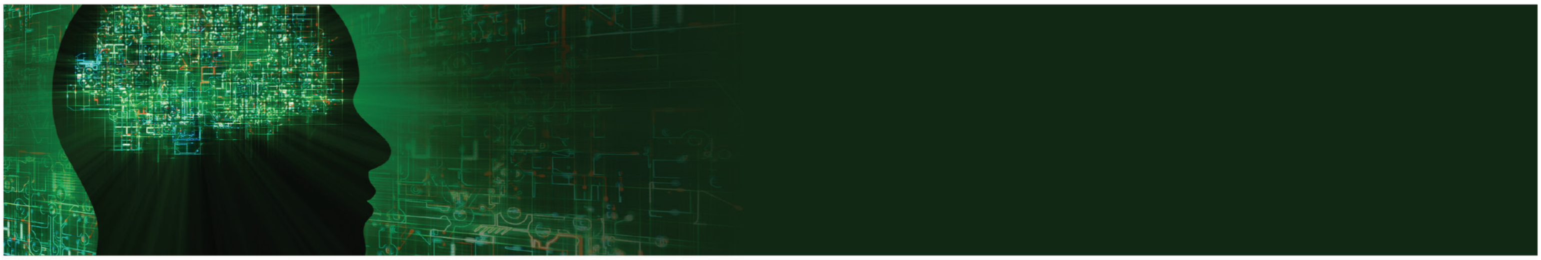

Our engineering teams including mechanical, electrical, and software design resources as well as process development, manufacturing, and quality engineering. Cirtec has a robust and flexible quality system; start-up companies and research groups often choose to work under our quality system rather than developing their own. We provide design history file management and document control services. We can also perform or contribute to requirements documentation, risk management activities, and design verification activities. Our manufacturing capabilities relevant to active implantable devices include:

- Laser welding
- Resistance welding
- Electrical assembly
- Laser marking
- Laser ablation
- Surface treatment
- Ultrasonic cleaning
- Epoxy molding and adhesion
- Silicone molding and adhesion
- Packaging and labelin

Manufacturing capabilities are supported by in-house test capabilities. These include mechanical testing, hermetic leak testing, bench level and automated electrical testing, environmental conditioning, and custom test setups such as soak or pressure testing.

Our manufacturing infrastructure includes multiple ISO class 7 cleanrooms and ESD safe handling procedures for all devices. Cirtec provides complete supply chain management services, leveraging our extensive Approved Supplier List and robust inventory management and material traceability procedures.

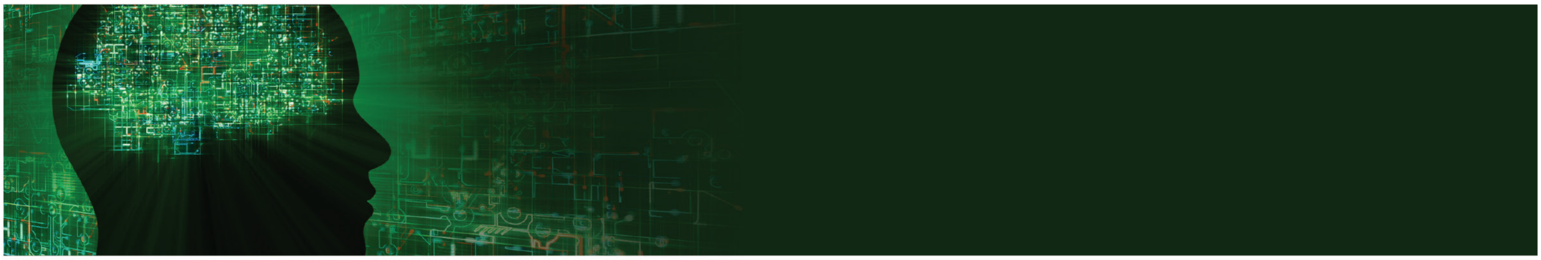

VALUE TO PROPOSING TEAMS: There are six primary areas where Cirtec capabilities provide value to proposing teams.

Hermetic barrier development	Hermetic package and barrier design
	Leak specification development
	Internal moisture management: Vacuum bake, desiccant handling, residual gas analysis testing
	Hermetic sealing (welding, brazing) and testing
Electrical assembly	Soldering, wire bonding, and die bonding
	Coil winding
	ESD-safe workspaces
	Lithium battery safe handling procedures
	Bench and automated testing
Interconnect development	Feedthrough design and sourcing
	Hardwired and disconnectable design
	Fabrication and testing
Encapsulation	Epoxy, silicone, and parylene encapsulation
	Layered encapsulation design to manage fluid ingress
	In-house design, tooling, molding, testing
Sterile packaging	Package design
	Pouch and tray packaging and labeling
	Off-the-shelf double sterile tray system
	Sterilization service provider management
Quality system support	Design control support and design history file maintenance
	GMP manufacturing
	Supply chain management
	Procedures for and support of requirements definition, risk management, design verification, and process validation

Dragon ID

Primary Contact

Eugene Malinskiy / eugene@dragonid.com

Capabilities

- Our staff is very experienced working within the complexities of the medical system from observing physicians in the operating room to bringing devices through regulatory approval.
- We have engineers with process development and industrial experience, including interfacing directly with manufacturers.
- We have software engineers experienced in both hardware and software engineering and experience in microcode and embedded systems, end user applications, and high performance computing.
- In addition to engineering talent, we staff researchers, designers, and SMEs to bring a truly interdisciplinary approach to innovation and design for a project.
- We have experience recruiting, establishing, and coordinating relationships between global stakeholders for complex projects to maximize accountability and efficiency for a product while minimizing bureaucracy required.
- We work with our world--class partners in industries such as industrial design, rapid multi--material prototyping, materials selection, and manufacturing in order to round out our capabilities for a product.

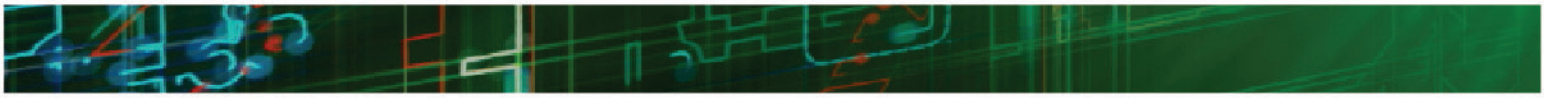

Freedom Photonics LLC

Primary Contact

Leif Johansson / leif@freedomphotonics.com

About

Founded in 2005 we are privately held Limited Liability Company based in Santa Barbara, California with ISO Qualification in process, certification in Q2 2016. We develop, manufacture and market complete photonic integrated circuit based products (Indium Phosphide, PLC, Si Photonics).

Our main markets in components, modules and subsystems are for fiber-optic communication, free-Space-optical (FSO) communication and infrared sensors and illuminators.

Capabilities

We offer electronics design, assembly and testing with expertise in photonic design, packaging, control electronics, manufacturing and qualification.

We offer services across development and have a premier R&D compound semiconductor facility with lithography (steppers, DUV stepper, EBL), wet and dry etching, deposition and characterization capabilities.

Our Business Model

Conceptualization, prototyping and development

- Fast R&D turn-around and low cost device fabrication in Nanotech cleanroom facility (Freedom Photonics employees)
- Internal multi-project wafer runs

Volume device production with commercial foundries worldwide

- Fabrication processes designed for manufacturability
- Access to manufacturability and production capacity, qualified processes
- Experience in outsourcing volume device manufacturing

Significant internal capability for assembly, packaging, test

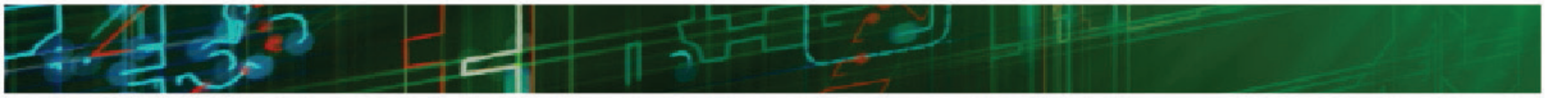

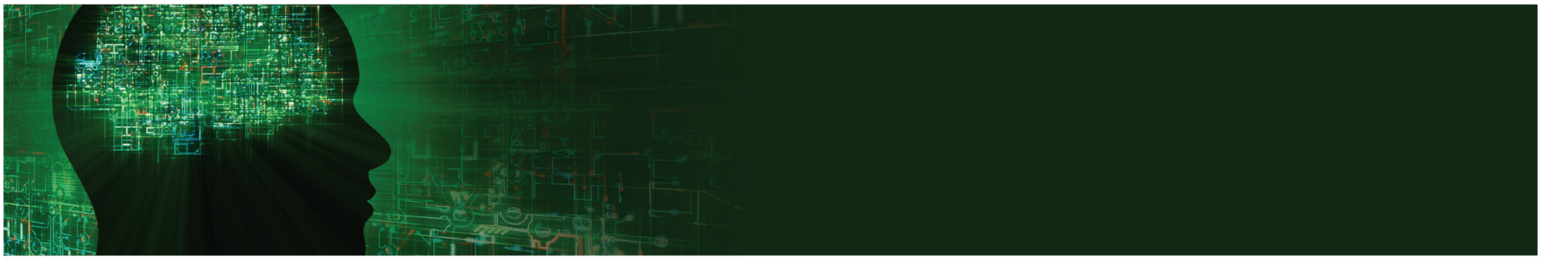

Capabilities and Facilities I – Photonics Design and Fabrication

Design

- Commercial and proprietary tools for passive and active PIC modeling, (Lumerical, native FDTD/BPM code, mode solvers, mask layout, SOA/active region models, Simwindows, SRIM)

Mask layout

- Proprietary scripting tool, with design rule checker and automated mask plate layout

Fabrication

- Full wafer process, die pick-and-place, wedge bonder, ball- bonder, auto-cleaver, pull-tester, etc.

Capabilities and Facilities II – Photonics Production Test

- Photonic device bar level probe
- Chip-on-carrier level (probe card)
- Package; DC, spectral, and data testing
- Burn in – monitored, environmental chambers

Capabilities & Facilities III – Packaging

Several in-house systems with laser welding and UV curing capability.

- Custom tooling, TOSA packages, 14 pin butterflies etc.
- ZEMAX commercial optical train design tool

Capabilities and Facilities IV – Electronics

- Digital and Analog Electronics
- FPGA design and verification tools in house
- Commercial tools for board layout
- Outside foundry for PCB fabrication and stuffing
- Electronics workstations for assembly and test
- Oscilloscopes, sources, multi-meters

We perform contract fabrication and engage with a number of domestic and international fabrication providers, both full process or selected fabrication steps are available.

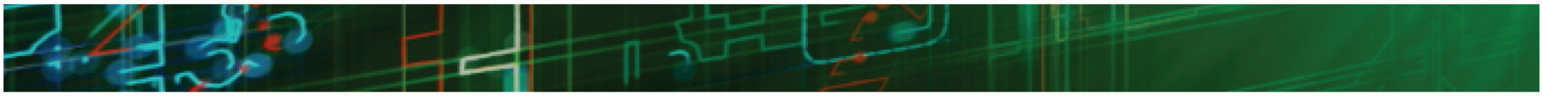

IMEC

Primary Contact

Barun Dutta / Barundeb.Dutta@imec.be

Capabilities

IMEC is interested in making its comprehensive capabilities available for joint-research programs for solutions in Life Sciences Technologies, Healthcare and Bio-medical engineering, including:

- Semiconductor/nanotechnology
- Process technology
- Design technology and systems/software technologies

Additional Contacts

Maarten Willems / Maarten.Willems@imec.be

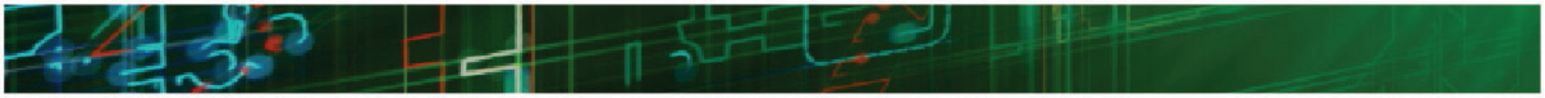

Infinite Arthroscopy

Infinite Arthroscopy

Primary Contact

Daniel Dudley / d.dudley@infinitearthro.com

About

We are exclusively focused on developing new medical technologies with an emphasis on translating innovations from other industries to the operating theatre. We have experience in the full life cycle of medical device development from initial conception to regulatory submission. We have large number of rapid prototyping and fabrication techniques available in-house including but not limited to software-based design, stereolithography, fused deposition modelling, PCB milling, and machining. Our staff includes designers, engineers, health care professionals, and subject matter experts from a diversity of industries allowing us to leverage a far wider knowledge base than the standard medical device company. We maintain a nationwide network of leading physicians and institutional partners to develop, refine, and validate our technologies according to contemporary healthcare practice and needs.

Capabilities

Infinite Arthroscopy also maintains a portfolio of intellectual property derived from our medical device technologies that would be available to the NESD Industry Group. Two relevant examples are listed below.

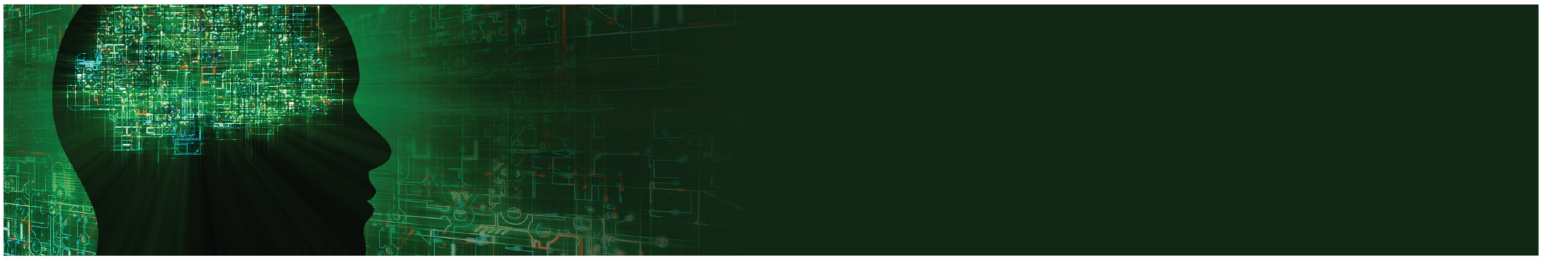

Infinite Arthroscopy, Inc. Proprietary Technologies:

Novel Light Source

- Miniaturized / Ruggedized
- Low power (sub 1 watt)
- Low heat (to LEDs of comparable lumen-output)
- High lumen output
- Long Lifetime (50,000+ hours)
- Emits only target wavelength(s) (Capacity to customize to project needs)

Command/Control Architecture for Medical Devices

- HIPAA/FTC Approved for use in Operating Rooms
- Uninterrupted High-bandwidth capacity
- Transmission below standard RF noise floor
- Encrypted protocol

Additional Contacts

Howard Fein / h.fein@infinitearthro.com

Inscopix

Primary Contact

Pushkar Joshi / pushkar@inscopix.com

Capabilities

- Neuroscience application development and validation (surgical procedures for rodents and non-human primates, application of biological reagents within our own life science lab)
- Optical system development (lens design and simulation, LED light source development, prototyping)
- Hardware system (FPGA-based and/or microcontroller based mixed-signal systems especially involving imaging, mechanical packaging) development, testing, and prototyping
- Software (application, image processing, computer vision, analytical) development and testing
- Mechanical accessories development (design, prototyping via 3D printing)
- With these capabilities, we can customize Inscopix's existing and emerging systems and applications, or develop new systems and applications for or with other proposers in support of this program.

Additional Contacts

Shung Chieh / schieh@inscopix.com

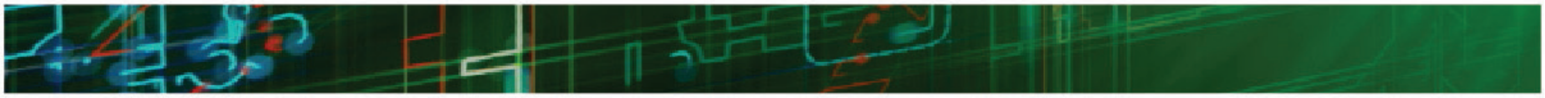

Nanomedical Diagnostics

Primary Contact

Brett Goldsmith / brgoldsmith@nanomedicaldiagnostics.com

Capabilities

- Graphene FET sensor array chips appropriate for integration with nearly any biochemical or electrochemical interaction
- Current sensor hardware capable of measuring biochemical and bioelectric interactions over time with 10 millisecond time resolution, and spatial resolution less than 10 microns
- Capability to produce sensor hardware with time resolution approaching one microsecond with a spatial resolution less than 2 microns
- Expertise in combining electronic materials, packaging and assembly with active biological materials
- Projected FDA clearance for an in vitro clinical product, and associated manufacturing chain, using graphene sensors by mid 2017
- Graphene appropriate for use in medical sensors integrated into an electronic or implantable platform of your choosing
- Chemical and biochemical expertise in biochemical attachment and blocking techniques specialized for boosting specificity of electronic readout of biological interactions
- A complete "turnkey" R&D system for using graphene sensor chips
- Hardware and software appropriate for reading graphene biosensors

Additional Contacts

Francie Barron / fbarron@nanomedicaldiagnostics.com

Naval Research Laboratory

Primary Contact

Dr. Jas. S. Sanghera / sanghera@nrl.navy.mil

About

NRL is the corporate research laboratory for the Navy and Marine Corps and conducts a broad program of scientific research, technology and advanced development.

The Optical Materials and Devices Branch has five primary areas: high purity chemicals, specialty optical materials, silica fiber technology, optical devices and advanced concepts.

Capabilities

We have expertise in the research, development and fabrication of glasses, glass-ceramics, ceramics, crystals, bulk optics, films, waveguides and optical fibers.

We offer a full suite of capabilities to make transparent ceramics to solve the ceramic packaging needs of NESD, some key properties:

- UV to RF transparency tailoring
- Tougher, stronger, harder than glass
- Electrically inert
- Demonstrated conformal shapes
- Demonstrated bonding technologies
- Vertical integration

Our facilities include class 100 clean rooms, powder processing equipment, sintering systems, and characterization capabilities (XRD, SEM, TEM, optics labs, mechanical testing).

Some of our prior applications include: Conformal Optics, IR simulators, photonic mast design, flexible PV's, IRCM, and Directed Energy Weapons.

Additional Contacts

Shyam Bayya / shyam.bayya@nrl.navy.mil

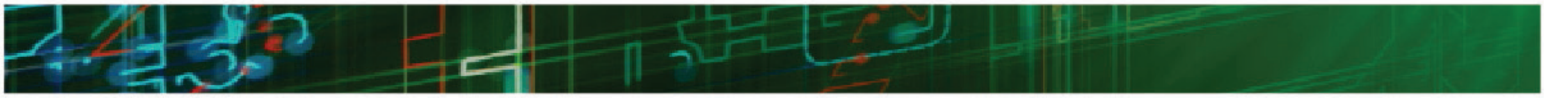

Star Lab

Primary Contact

Adam Fraser / adam.fraser@starlab.io

About

Star Lab is a high-tech small business focused on complex systems security research and development, with significant expertise and a pedigree of success developing security technologies for safety-critical and mission-critical systems. The company specializes in the areas of high-assurance operating systems, technology protection, applied cryptography, and anti-tamper, and is also a leader in kernel/hypervisor/embedded software development. Star Lab has offices in Washington D.C. and Huntsville, AL.

Additional Contacts

Irby Thompson / irby@starlab.io

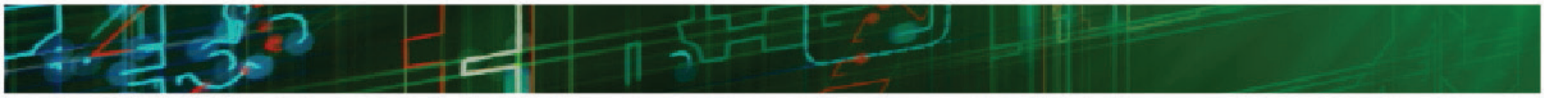

Qualcomm

Primary Contact

Karim Arabi / karabi@qti.qualcom.com

Capabilities

- Experience in building an end-to-end platform as a reference design for interface with brain
- Experience in driving an industry wide standard effort to standardize a communication medium and protocol for brain interface
- Advanced low power wireless solutions, RF interface, MODEM system solutions and related electronics for brain interface
- Advanced wireless charging and powering solutions for brain implants
- Ultra low power electronic design solutions at advanced technology nodes suitable for brain implants to enable always on monitoring and processing
- Ultra low power DSP solutions, stream data processing and data compression and decompression solutions

Additional Contacts

Rudy Beraha / rberaha@qti.qualcomm.com

New World War

Revolutionary Methods for Political Control

Mind-Reading

History

Probably no more “intrusive and persistent” method of obtaining information about a person exists than reading their mind.[1] Research on mind-reading has been vigorously pursued by US government agencies and various academic centers since the 1970s, and continues to this day.

Since 1973 DARPA has been studying mind-reading with EEG hooked to computers, using scientists at the University of Illinois, UCLA, Stanford Research Institute, Massachusetts Institute of Technology, and the University of Rochester.

They developed a system that could determine how a person perceived colors or shapes and were working on methods to detect daydreaming, fatigue, and other brain states. Although the device had to be calibrated for each person’s brain by having them think a series of specific thoughts, the calibration was quick.

In 1974 another very basic mind-reading machine was created by researchers at Stanford Research Institute. It used an EEG hooked to a computer which allowed a dot to be moved across a computer screen using thought alone. When interpreting people’s brainwaves, it was right about 60% of the time. During these tests scientists discovered that brain patterns are like fingerprints, each person has their own. So, each computer would have to be calibrated for a specific person.

Another method to address this issue was to store a large amount of generic patterns on the computer, so when it encountered a brain pattern it didn’t recognized, it used one that most resembled it. Since then, DARPA has sponsored Brain-Computer Interface (BCI) and mind-reading programs at Duke University, MIT, University of Florida, and New York State University, Brooklyn.

The Human Computer Interaction group at Tufts University has studied mind-reading funded by grants from a government research and education agency known as the National Science Foundation (NSF). Carnegie Mellon University, Stanford University, and the MIT Sloan School of Management have studied mind-reading. The Computer Laboratory at the University of Cambridge in England has developed mind-reading machines based on facial expressions.

Other academic institutions that have participated in mind-reading projects include the University of California, Berkeley, University of Maryland, and Princeton University in New Jersey. Microsoft has studied mind-reading using EEG to better accommodate its users. Emotiv Systems built a mind-reading gaming device which uses EEG to infer the mental states of video game

Main Menu

Home

Read *New World War* Online

Articles

Solutions

Biography

Download my books for free

Updates

Update: January 20, 2018
A freemasonic death threat....more

Important Updates

All Previous Updates

Donations

Your Donations Are Appreciated

Available Books

PDF, Kindle, & ePub editions of my books can be downloaded for free, or purchased on Amazon & LuLu.

players. Honda Motors and Advanced Telecommunications Research Institute International (ATR) have studied mind-reading.

Neuroimaging Devices

Scientists discovered that the neural code of the human brain is similar to the digital code of a computer. To some extent, they have deciphered this code. Prior to this, they assumed that it was necessary to identify the neurons associated with specific acts, which would have made mind-reading much more difficult.[2]

They now understand that it's not necessary to monitor billions of neurons to determine which are connected to a particular thought or act. Only a small number of them need to be monitored to accomplish this. To monitor these neurons researchers use neuroimaging devices. They include event-related optical signal (EROS), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), functional near-infrared imaging (fNIR), magnetoencephalography (MEG), and positron emission tomography (PET).[3] These devices may be combined for a more accurate reading.

There are basically two types of measurements, *direct* methods and *indirect* methods. Direct methods measure changes in electromagnetic fields and currents around the brain which are emitted from the surface of the scalp, or they monitor the neurons themselves. Indirect methods measure hemodynamic (blood movement) changes of hemoglobin in specific tissue compartments.

Both of these methods are almost simultaneous with neuronal activity. Regarding sensors, there are invasive ones which must be implanted, and non-invasive ones which can be worn on the scalp, in the form of a headband.

Electroencephalography (EEG) provides a direct method for determining brain states and processes by measuring the electrical activity on the scalp produced by the firing of neurons in the brain. EEG has been around for over 100 years. EEG is commonly used in neuroscience, cognitive science, and cognitive psychology. It is inexpensive, silent, non-invasive, portable, and tolerates movement.

Wireless EEG which uses non-invasive sensors that have physical contact with the scalp can transmit the signals to a remote machine for deciphering. Although, in 1976 the *Los Angeles Times* reported that DARPA was working on an EEG to detect brain activity several feet from a person's head, which was to be completed in the 1980s. EEG normally produces only a general indicator of brain activity.

However, in 2008 *Discovery News* reported that a company called Emotiv Systems developed an algorithm that decodes the cortex, providing a more accurate measurement. "We can calibrate the algorithm across a wide range of technologies with the same resolution you would get from placing an invasive chip inside the head," said Tan Le, president of Emotiv Systems.

Functional magnetic resonance imaging (fMRI) measures the blood flow in the brain in response to neural activity. Active neurons use oxygen, which is brought to them by blood. The more active a region of the brain is the more

GEOENGINEERING
PLANETARY
SABOTAGE
&
HUMAN
EXTERMINATION
Mark M. Rich

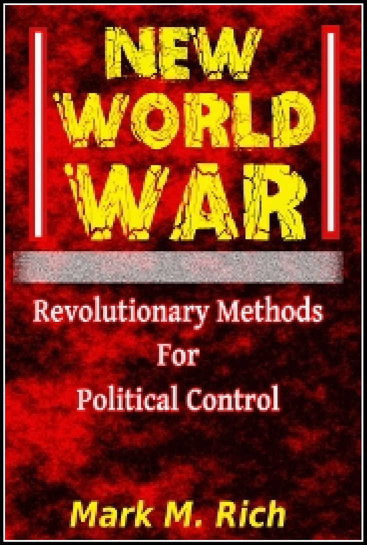

blood flows in the area. This movement of blood is referred to as *hemodynamic activity*. FMRI can detect which areas are receiving blood, which indicates that they're processing information.

The fMRI provides an indirect measurement of brain processes. It is the most common method of neuroimaging, and can produce 2 and 3-dimensional images. It is non-invasive, and can record signals from all brain regions, unlike EEG which focuses on the surface only.

Functional near-infrared imaging (fNIR) provides an indirect measurement of brain activity by detecting hemodynamic changes in the cortex. Although it is based on different principles, in that it uses light, it functions in the same manner as fMRI. FNIR can provide an almost continuous display of these changes in the cortex. It is inexpensive, non-invasive, and portable. A wireless headband with sensors exists for this device.

Event-related optical signal (EROS) is a brain-scanning device that focuses near-infrared light into the cerebral cortex to detect the density of neurons indicated by the transparency of brain tissue. Because it can only detect these changes a few centimeters deep, it can only image the cerebral cortex. Unlike fNIR, which is an optical method for measuring blood flow, EROS detects the intensity of neurons themselves and provides a direct measurement of brain activity. It is very accurate, portable, inexpensive, and non-invasive. A wireless headband with sensors exists for this device.

Capabilities

Mind-reading can be accomplished by first having a computer learn which brain patterns are associated with specific thoughts, then store the decoded information in a database. This machine learning is accomplished using a type of *artificial intelligence* (AI) called an *algorithm*. A very basic algorithm is a spell checker, which uses a database of common mistakes associated with a particular sequence of letters to present suggestions to a user.

"The new realization is that every thought is associated with a pattern of brain activity," proclaimed neuroscientist John Dylan Haynes, in *Newsweek International* on February 4, 2008. "And," says Haynes, "you can train a computer to recognize the pattern associated with a particular thought."

In a January 2000 issue of *US News and World Report*, Lockheed Martin neuroengineer Dr. John Norseen announced, "Just like you can find one person in a million through fingerprints ... you can find one thought in a million." This can be accomplished using AI and HCI, or what Dr. Norseen calls biofusion.

The decoded brain signals can be stored in a database. Then when someone is scanned, the computer detects the pattern and matches the signals to the database of known meanings. But it's not necessary to scan a brain to decode its signals for every single thought, such as a picture.

Instead, after the machine has learned how to decipher patterns associated with specific thoughts such as images, more images can be added to the program and the computer can use the process it used for the other images as a model to somewhat accurately detect additional thoughts.

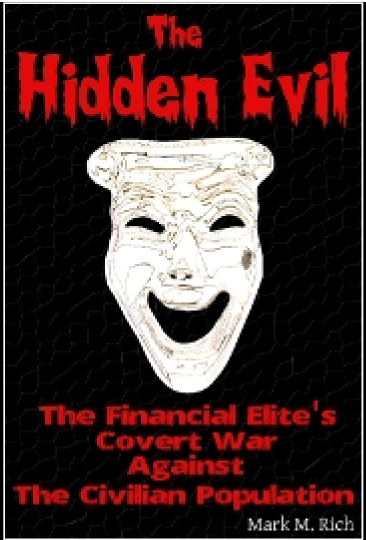
The Hidden Evil
The Financial Elite's Covert War Against The Civilian Population
Mark M. Rich

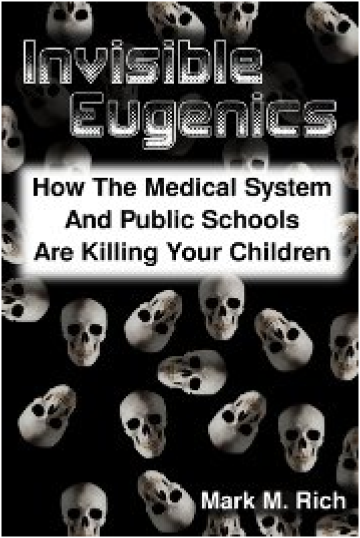
Invisible Eugenics
How The Medical System And Public Schools Are Killing Your Children
Mark M. Rich

New World War: Revolutionary Methods for Political Control

Dedication & Thanks

Peer Review & Thanks
Dedication

Volume I: Current Political Situation

Overview
Introduction
Revolution in Warfare
The Other World
Dictatorship Creation
The Groups Facilitating the Revolution
Their Goal is Neo-Feudalism
Problem-Reaction-Solution
Volume I Commentary

Both words and images can be detected using mind-reading devices with varying degrees of accuracy. This can occur for words and images being viewed by a person on an external display, such as a book, or words and images just being thought of with no external stimuli.

“It is possible to read someone’s mind by remotely measuring their brain activity,” announced *New Scientist* in their *Mind-Reading Machine Knows What You See* article of April of 2005. The Computational Neuroscience Laboratories at the Advanced Telecommunications Research Institute International (ATR) in Kyoto Japan, and Princeton University in New Jersey, proved that by monitoring the visual cortex with fMRI they could determine which basic objects (sets of lines) a person was looking at.

When the objects were combined, they could even determine which one was being focused on. According to the scientists, it may be possible not only to view but also to record and replay these images. They announced that the technology could be used to figure out dreams and other secrets in people’s minds.

Vanderbilt University in Nashville has conducted simple mind-reading tests using an fMRI/Computer, which learned what basic images a group of test subjects was looking at. They were able to predict with 50% accuracy which objects the test subjects were thinking of when they were asked only to remember what they had seen, without being shown the images.

On March 6, 2008 *ABC News* reported that neuroscientists at the University of California at Berkeley accomplished mind-reading by monitoring the visual cortex with an fMRI connected to a self-learning (artificial intelligence) computer program.

First, they used 1750 pictures to build a computational database for the computer to learn with by flashing the pictures in front of test subjects connected to an fMRI. This allowed the algorithm to decipher the brain patterns and associate them with the images.

In addition to deciphering these brain patterns, the computer recorded the process that it used to accomplish this, and built a model based upon it. Then, without scanning the test subjects, they added 120 new pictures to the program and allowed it to create its interpretation of what the new brain signals would be, based on the previous model.

Then they had the test subjects look at these pictures which they had never seen while being scanned. The computer predicted what they were looking at 72% of the time. The scientists announced that the model could be used as a basis to predict the brain activity associated with any image.

What this means is that it’s not necessary to scan a brain to obtain the meaning of each signal. Once the model had been developed, they could simply add new pictures to the database/dictionary. The scientists suggested that out of 1 billion pictures, the computer would be accurate about 20% of the time.

Images which are not consciously seen by a person can even be detected by mind-reading machines. Researchers at University College London flashed

Volume II: The New War

The New War
The New Enemy
Initiatives to Remove Civil Liberties
The Investigation
Surveillance Technology and Methods
Mind-Reading
Volume II Commentary

Volume III: Weapons of The New War

Introduction to Nonlethal Weapons
Psychological Operations
Introduction to Directed-Energy Weapons
High-Powered Microwaves
High-Powered Lasers
Sonic Weapons
Computer Network Operations
Microwave Hearing
Silent Subliminals
Use of Citizen Informants
Chemical and Biological
Weather Warfare
Miscellaneous Weapons and Tactics
Volume III Commentary

Volume IV: The Coverup

Volume IV Introduction
Schizophrenia Spectrum Disorders
Control of the Medical Industry
Another Look at Schizophrenia
Political Considerations
Punitive Psychiatry in Communist Russia
Coverup Initiatives
Volume IV Commentary
Conclusion

Appendix

A Brief History of PsyOp
Small-Scale Wars
Nongovernmental Organizations
Human-Computer Intelligence Network

pictures in quick succession to test subjects connected to an fMRI. Although some of these pictures were invisible to the subjects, they were accurately recorded 80% of the time by the computer.

Like a fingerprint, each person has their own *brainprint*. Therefore, calibration for each brain is necessary. This is accomplished by having the person think a series of specific thoughts. In the case of EEG, this calibration can take less than a minute. However, because the signals which represent thoughts are similar from one person to the next, a universal mind-reading database has been suggested.

Using fMRI, scientists at Carnegie Mellon University (CMU) discovered that the brain patterns associated with specific thoughts are quite similar among multiple people. This, they stated, would provide the opportunity to create a universal mind-reading dictionary.

Scientists at the University of California at Berkeley mentioned that a “general visual decoder” would have great scientific use. Likewise, the brain patterns associated with specific words that occur when people are reading are also basically the same. This similarity of brain functions associated with words seems to have been an evolutionary development which allowed for an advantage in communication.

A mind-reading machine capable of determining the brain pattern associated with a specific word was developed by scientists at CMU. Brain scans using fMRI were taken of test subjects who were given a variety of words to think of in order to train the computer. An important consideration here is that they were not viewing these words on an external display, only thinking about them. After the computer identified the brain patterns associated with those words, the subjects were given two new words to think about, which the computer accurately determined.

Although, in this particular study only a couple of words were tested, it proves that after a model of how to decipher brain signals was created, AI could accurately determine new words that subjects were thinking about. “These building blocks could be used to predict patterns for any concrete noun,” proclaimed Tom Mitchell of the Machine Learning Department.

In February of 2004 *Popular Science* announced that a mind-reading computer could, in theory, translate a person’s working verbal memory onto a computer screen. “You could imagine thinking about talking and having it projected into a room 2,000 miles away,” says Professor Craig Henriquez at Duke University’s Center for Neuroengineering, who has studied mind-reading for DARPA. He added, “It’s very, very possible.”

FMRI can be used to determine if someone is reading or writing. Neuroscientists can determine when a person is reading by monitoring their brainwaves. They can almost determine exactly what they’re reading. And because these patterns are similar from one person to the next, a universal device for determining what people are reading is possible.

In March of 2008 both Technology Review and *ABC News* revealed that an fMRI could in theory be used to display a person’s dreams. Then in December of 2008, scientists at ATR in Kyoto Japan announced that they developed a

Electronic Tyranny
Other Devices Connected to the GIG
My Experience

Sources

Other Sites

www.TheHiddenEvil.com
www.RichEssence.com

technology that would eventually allow them to record and replay a person's dreams.

Emotions from love to hate can be recognized by neuralimaging. The level of stress a person is experiencing can also be measured. Brain states such as honesty, deception, and even self deception, can be measured.

Patterns associated with decisions can also be read. Scientists from CMU, Stanford University, and the MIT Sloan School of Management were able to accurately predict the purchasing decisions of test subjects in a virtual shopping center. They monitored the subject's level of interest in a product as well as their decision to purchase it.

Neuroimaging can also detect decisions about how someone will later do a high-level mental activity. Neuroimaging can be used to determine if someone is speaking or reading. It can be used to detect areas of the brain that are active when someone is hearing a sound, or touching an object.

Brain patterns associated with specific physical movements, such as a finger, can be deciphered with neural imaging. The mere intention to make a physical movement can be detected before the actual movement is made.

Cameras

A type of mind-reading is possible with cameras connected to computers. One such device, called the Emotional Social Intelligence Prosthetic (ESP), was developed at the MIT Media Laboratory in 2006. It consists of a tiny camera that can be worn on a hat, an earphone and a small computer that is worn on a belt. It infers a person's emotional state by analyzing combinations of subtle facial movements and gestures.

When an emotional state is detected, the wearer is signaled through the earphone to adjust their behavior in order to gain the attention of the target. The computer can detect 6 emotional states. It can also be adjusted for cultural differences and configured specifically for the wearer.

Around this time, the Computer Laboratory at the University of Cambridge, UK, developed a similar camera-based mind-reading machine. It uses a computer to monitor, in real-time, combinations of head movement, shape, color, smiles, and eyebrow activity to infer a person's emotional state.

It detects basic emotional states such as happiness, sadness, anger, fear, surprise, and disgust, as well as more complex states. It's accurate between 65 and 90 percent of the time. "The mind-reading computer system presents information about your mental state as easily as a keyboard and mouse present text and commands," they announced.

Used for Surveillance

Mind reading exists. The DOD and various institutions have vigorously researched this subject since at least the mid 1970s. "Mapping human brain functions is now routine," declared *US News and World Report*, in an article entitled *Reading Your Mind—And Injecting Smart Thoughts*, of January of 2000.

Both words and images, being viewed or thought, can be mind-read. Various emotional states as well as mental processes such decisionmaking, reading, writing, movement and the intention to make a movement, can be detected with mind-reading devices. Perceptions such as touch, sound, and light can also be detected.

The proposed uses for mind-reading technology are positive. Some include determining if people in comas can communicate, helping stroke patients and those who suffered brain injuries, aiding those with learning disorders, assisting with online shopping, and improving people's communications skills.

However, other uses that have been suggested include the monitoring of unconscious mental processes, and interrogation of criminal suspects and potential terrorists. Dr. John Alexander mentioned that the recent developments in mind-reading technology would take surveillance to new levels by allowing investigators to "peer into the inner sanctum of the mind," in order to determine if a suspect has caused, or will likely cause a crime.

Dr. Norseen has sent R&D plans to the pentagon to have tiny mind-reading devices installed at airports to profile potential terrorists. He suggested that these devices could be functional by 2005. In August of 2008, *CNN* stated that the US military's knowledge obtained from mind-reading research could be used to interrogate the enemy.

Law Enforcement Technology announced in September of 2005 the existence of a new forensic technology known as *Brain Fingerprinting*, which has already been used in hundreds of investigations as a lie detector by the CIA, FBI and law enforcement agencies in the United States.

Brain Fingerprinting is admissible in court, because unlike a polygraph, which relies on emotional responses, it uses EEG to see how the brain reacts to words and pictures related to a crime scene. Dr. Larry Farwell, its inventor, says it is completely accurate. According to the report it will be used to help speed-up investigations.

Sources Next

Endnotes

[1] Another possible method to obtain information is Remote Viewing. RV is the ability to produce correct information about people, events, objects, or concepts that are somewhere else in space and time, and are completely blind to the viewer collecting the information. It can be used to describe people or events, produce leads, reconstruct events, make decisions, and make predictions about the future. See *Remote Viewing Secrets* by Joseph McMoneagle. RV tests were conducted by the US government over a 20-year period during Project Stargate, a classified initiative by the CIA which began in 1972 and lasted until about 1994. Most of the 154 tests and 26,000 trials took place at the Cognitive Sciences Laboratory at Forte Meade, Maryland. A majority of the results of the project are still classified. See the *Journal of Parapsychology* articles, *Remote Viewing by Committee*, September 22, 2003, by Lance Storm and *Experiment One of the SAIC Remote Viewing Program*, of December 1, 1998, by Richard Wiseman and Julie Milton. The success of the project varies depending on the source. Allegedly the original tests were conducted under rigid scientific conditions, which had impressive results. However, the same sources describe RV

in general as ineffective. See *Discover Magazine's* article, *CIA ESP*, on April 1, 1996, by Jeffrey Kluger, and the *Washington Post's*report, *Many Find Remote Viewing a Far Fetch from Science*, on December 2, 1995, by Curt Suplee. According to author McMoneagle, an original viewer during Project Stargate, it is accurate about 50 or 60 percent of the time. Nevertheless, RV will be used to obtain intelligence, according to John B. Alexander. See *The New Mental Battlefield*, which appeared in the December 1980 issue of *Military Review*. Also see the June 1998 *Research Report Number 2* of the University of Bradford's Non-Lethal Weapons Research Project (BNLWRP), for how RV has been added to the NLW arsenal. According to multiple sources, US government agencies are now using the consulting services of RV professionals. This was reported on January 9, 2002 in the *University Wire's (Colorado Daily)* article, *Clairvoyant Discusses Reveals Details of Remote Viewing*, by Wendy Kale, and in the *Bulletin of the Atomic Scientists* on September 1, 1994, in its report, *The Soft Kill Fallacy* by Steven Aftergood. In his book *Winning The War: Advanced Weapons, Strategies, and Concepts for the Post-911 World*, Alexander had this to say regarding RV: "Since the beginning of history, humans have made anecdotal references to innate abilities to foretell the future, to know what was occurring at distant locations or the status of people separated from them, and to find resources they need without any traditional means of accessing that information." He continued: "Studies have demonstrated beyond any doubt that these nontraditional capabilities exist. ... [RV can] radically change our means of gathering intelligence. It holds the promise of providing information about inaccessible redoubts and advances in technology. More importantly, once these skills are understood, those possessing them will be able to determine an adversary's intent and be predictive about the events."

[2] Because neuroimaging technology decodes brain patterns to thoughts, some argue that it technically doesn't read a person's mind. However, because specific thoughts and brain states can be deciphered, here it is referred to as mind-reading. Additionally, most mainstream documents refer to this as mind-reading, despite the fact that it is actually brainwave-reading.

[3] Magnetoencephalography (MEG) and positron emission tomography (PET) can also be used to infer a person's neurophysiological state. But because they are impractical for field use due to their large size and harmful radiation, MEG and PET won't be considered here. However, DARPA is in the process of developing a small helmet-sized MEG device which would be connected to a portable computer. See the article *Mind over Machine* in the February 1, 2004 issue of *Popular Science*, by Carl Zimmer.

Copyright © 2011 Mark Rich. All rights reserved. Mark (at) NewWorldWar.org

Congressional Research Service
Informing the legislative debate since 1914

Navy Lasers, Railgun, and Hypervelocity Projectile: Background and Issues for Congress

Ronald O'Rourke
Specialist in Naval Affairs

May 25, 2016

Congressional Research Service
7-5700
www.crs.gov
R44175

CRS REPORT
Prepared for Members and
Committees of Congress

Summary

The Navy is currently developing three potential new weapons that could improve the ability of its surface ships to defend themselves against enemy missiles—solid state lasers (SSLs), the electromagnetic railgun (EMRG), and the hypervelocity projectile (HVP).

Any one of these new weapon technologies, if successfully developed and deployed, might be regarded as a "game changer" for defending Navy surface ships against enemy missiles. If two or three of them are successfully developed and deployed, the result might be considered not just a game changer, but a revolution. Rarely has the Navy had so many potential new types of surface-ship missile-defense weapons simultaneously available for development and potential deployment.

Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain. Overcoming these challenges will likely require years of additional development work, and ultimate success in overcoming them is not guaranteed.

The issue for Congress is whether to approve, reject, or modify the Navy's funding requests and proposed acquisition strategies for these three potential new weapons. Potential oversight questions for Congress include the following:

- Using currently available approaches for countering anti-ship cruise missiles (ASCMs) and anti-ship ballistic missiles (ASBMs), how well could Navy surface ships defend themselves in a combat scenario against an adversary such as China that has large numbers of ASCMs (including advanced models) and ASBMs? How would this change if Navy surface ships in coming years were equipped with SSLs, EMRG, HVP, or some combination of these systems?
- How significant are the remaining development challenges for SSLs, EMRG, and HVP?
- Are current schedules for developing SSLs, EMRG, and HVP appropriate in relation to remaining development challenges and projected improvements in enemy ASCMs and ASBMs? To what degree are current schedules for developing SSLs, EMRG, or HVP sensitive to annual funding levels?
- When does the Navy anticipate issuing roadmaps detailing its plans for procuring and installing production versions of SSLs, EMRGs, and HVP on specific Navy ships by specific dates?
- Will the kinds of surface ships that the Navy plans to procure in coming years have sufficient space, weight, electrical power, and cooling capability to take full advantage of SSLs (particularly those with beam powers above 200 kW) and EMRG? What changes, if any, would need to be made in Navy plans for procuring large surface combatants (i.e., destroyers and cruisers) or other Navy ships to take full advantage of SSLs and EMRG?
- Are the funding sources for SSLs, EMRG, and HVP in Navy and Defense-Wide research and development accounts sufficiently visible for supporting congressional oversight?

Contents

Introduction 1
  Issue for Congress 1
  Scope of Report 1
Background 2
  Strategic and Budgetary Context 2
    Concern about Survivability of Navy Surface Ships 2
    Depth of Magazine and Cost Exchange Ratio 3
  SSLs, EMRG, and HVP in Brief 4
    SSLs 4
    EMRG 13
    HVP 16
  Indirectly Improving Ability to Counter ASCMs and ASBMs 22
  Remaining Development Challenges 22
    SSLs 23
    EMRG and HVP 24
Issues for Congress 26
  Potential Oversight Questions 26
Legislative Activity for FY2017 27
  Summary of Congressional Action on FY2017 Funding 27
  FY2017 National Defense Authorization Act (H.R. 4909/S. 2943) 27
    House 27
    Senate 30
  FY2017 DOD Appropriations Act (H.R. 5293) 33
    House 33
  Directed Energy Weapon Systems Acquisition Act of 2016 (H.R. 4964/S. 2778) 34
    House 34
    Senate 35

Figures

Figure 1. Laser Weapon System (LaWS) on USS Ponce 7
Figure 2. Laser Weapon System (LaWS) on USS Ponce 7
Figure 3. Laser Weapon System (LaWS) on USS Ponce 8
Figure 4. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship 10
Figure 5. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship 11
Figure 6. ONR Graphic of LWSD Components 12
Figure 7. Industry-Built EMRG Prototype Demonstrator 14
Figure 8. Industry-Built EMRG Prototype Demonstrator 15
Figure 9. EMRG Prototype Demonstrator Installed on a JHSV 16
Figure 10. Photograph Showing HVP 17
Figure 11. HVP 18
Figure 12. HVP Launch Packages 19
Figure 13. HVP Application to Various Launchers 20

Figure 14. Navy Slide Depicting Operations Against Various Target Types 21
Figure 15. Development Challenges for SSLs 24
Figure 16. Development Challenges for EMRG 25

Tables

Table 1. Summary of Congressional Action on FY17 Funding.. 27

Appendixes

Appendix. Potential Advantages and Limitations of Shipboard Lasers 38

Contacts

Author Contact Information 40

Introduction

Issue for Congress

This report provides background information and issues for Congress on three potential new weapons that could improve the ability of Navy surface ships to defend themselves against enemy missiles—solid state lasers (SSLs), the electromagnetic railgun (EMRG), and the hypervelocity projectile (HVP).[1]

Any one of these new weapon technologies, if successfully developed and deployed, might be regarded as a "game changer" for defending Navy surface ships against enemy missiles. If two or three of them are successfully developed and deployed, the result might be considered not just a game changer, but a revolution. Rarely has the Navy had so many potential new types of surface-ship missile-defense weapons simultaneously available for development and potential deployment. Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain.

The issue for Congress is whether to approve, reject, or modify the Navy's funding requests and proposed acquisition strategies for these three potential new weapons. Congress' decisions on this issue could affect future Navy capabilities and funding requirements and the defense industrial base.

Scope of Report

SSLs are being developed by multiple parts of the Department of Defense (DOD), not just the Navy. SSLs, EMRG, and HVP, moreover, have potential application to military aircraft and ground forces equipment, not just surface ships. And SSLs, EMRG, and HVP can be used for missions other than defending against ASCMs and ASBMs.[2] This report focuses on Navy efforts to develop SSLs, EMRG, and HVP for potential use in defending Navy surface ships against ASCMs and ASBMs. It supersedes an earlier CRS report that provided an introduction to potential Navy shipboard lasers.[3]

Note that while fictional depictions of laser weapons in popular media often show them being used to attack targets at long ranges, the SSLs currently being developed by the Navy for potential shipboard use would be used to counter targets at short ranges of about a mile to perhaps a few miles.

---

[1] Railgun is also spelled as rail gun; EMRG is also abbreviated as EM railgun; hypervelocity is also spelled as hyper-velocity or hyper velocity.

[2] As discussed later in the report, the Navy is exploring the potential for using shipboard lasers to counter small boats and unmanned aerial vehicles (UAVs), and EMRG can be used to attack land targets.

[3] CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke. This earlier CRS report has been archived and remains available as a supplementary reference source on potential Navy shipboard lasers.

Background

Strategic and Budgetary Context

Concern about Survivability of Navy Surface Ships

Although Navy surface ships have a number of means for defending themselves against anti-ship cruise missiles (ASCMs) and anti-ship ballistic missiles (ASBMs),[4] some observers are concerned about the survivability of Navy surface ships in potential combat situations against adversaries, such as China, that are armed with advanced ASCMs and with ASBMs.[5] Concern about this issue has led some observers to conclude that the Navy's surface fleet in coming years might need to avoid operating in waters that are within range of these weapons, or that the Navy might need to move toward a different fleet architecture that relies less on larger surface ships and more on smaller surface ships and submarines.[6] Such changes in Navy operating areas and fleet architecture could substantially affect U.S. military strategy and the composition of the Navy's shipbuilding expenditures.

Navy surface fleet leaders in early 2015 announced a new organizing concept for the Navy's surface fleet called distributed lethality. Under distributed lethality, offensive weapons such as ASCMs are to be distributed more widely across all types of Navy surface ships, and new operational concepts for Navy surface ship formations are to be implemented. The aim of distributed lethality is to boost the surface fleet's capability for attacking enemy ships and make it less possible for an enemy to cripple the U.S. fleet by concentrating its attacks on a few very-high-value Navy surface ships (particularly the Navy's aircraft carriers).[7] Perspectives on whether it would be cost effective to spend money spreading offensive weapons across a wider array of Navy surface ships might be influenced by views on whether those surface ships can adequately defend themselves against enemy missiles.

---

[4] These include the following: operating ships in ways that make it hard for others to detect and accurately track Navy ships; jamming or destroying enemy targeting sensors; interfering with the transmission of targeting data from sensors to weapon launchers; attacking weapon launchers (which can land-based launchers or launchers on surface ships, submarines, or aircraft); and countering ASCMs and ASBMs headed toward Navy ships. Navy measures for countering ASCMs and ASBMs headed toward Navy ships include the following: jamming a missile's guidance system; using decoys of various kinds to lure enemy missiles away from Navy ships; and shooting down enemy missiles with surface-to-air missiles and the Phalanx Close-In Weapon System (CIWS), which is essentially a radar-controlled Gatling gun. Employing all these measures reflects a long-standing Navy approach of creating a multi-layered defense against enemy missiles, and of attacking the enemy's "kill chain" at multiple points so as to increase the chances of breaking the chain. (The kill chain is the sequence of steps that an enemy must complete to conduct a successful missile attack on a Navy ship. This sequence includes, at a basic level of description, detecting and tracking the Navy ship, passing that information from sensors to the weapon launcher, launching the weapon, and guiding the weapon all the way to the Navy ship. Interfering with any one of these actions can break the kill chain and thereby prevent or defeat the attack.)

[5] See, for example, Andrew F. Krepinevich, Maritime Warfare in a Mature Precision-Strike Regime, Washington, Center for Strategic and Budgetary Assessments, 2014, 128 pp. For more on China's ASCMs and ASBMs, see CRS Report RL33153, *China Naval Modernization: Implications for U.S. Navy Capabilities—Background and Issues for Congress*, by Ronald O'Rourke.

ASCMs and ASBMs are not the only reasons that some observers are concerned about the future survivability of U.S. Navy surface ships in combat situations; observers are also concerned about threats to U.S. Navy surface ships posed by small boats, mines, and torpedoes.

[6] See, for example, Phillip E. Pournelle, "The Deadly Future of Sea Control," *U.S. Naval Institute Proceedings*, July 2015: 26-31.

[7] See, for example, Thomas Rowden, Peter Gumataotao, and Peter Fanta, "Distributed Lethality," *U.S. Naval Institute Proceedings*, January 2015: 18-23; Sam LaGrone, "SNA: Navy Surface Leaders Pitch More Lethal Ships, Surface Action Groups," *USNI News*, January 14, 2015; Kris Osborn, "Navy Unveils New Surface Warfare Strategy," *Military.com*, January 14, 2015; Sydney J. Freedberg Jr., "'If It Floats, It Fights,': Navy Seeks 'Distributed Lethality,'" *Breaking Defense*, January 14, 2015; Mike McCarthy and Megan Eckstein, "Navy Eyeing A 'Hunter Killer' Surface (continued...)

Depth of Magazine and Cost Exchange Ratio

Two key limitations that Navy surface ships currently have in defending themselves against ASCMs and ASBMs are limited depth of magazine and unfavorable cost exchange ratios. Limited depth of magazine refers to the fact that Navy surface ships can use surface-to-air missiles (SAMs) and their Close-in Weapon System (CIWS) Gatling guns to shoot down only a certain number of enemy unmanned aerial vehicles (UAVs) and anti-ship missiles before running out of SAMs and CIWS ammunition[8]—a situation (sometimes called "going Winchester"), that can require a ship to withdraw from battle, spend time travelling to a safe reloading location (which can be hundreds of miles away),[9] and then spend more time traveling back to the battle area.

Unfavorable cost exchange ratios refer to the fact that a SAM used to shoot down a UAV or anti-ship missile can cost the Navy more (perhaps much more) to procure than it cost the adversary to build or acquire the UAV or anti-ship missile. In the FY2016 defense budget, procurement costs for Navy SAMs range from about $900,000 per missile to several million dollars per missile, depending on the type.[10]

In combat scenarios against an adversary with a limited number of UAVs and anti-ship missiles, an unfavorable cost exchange ratio can be acceptable because it saves the lives of Navy sailors and prevents very expensive damage to Navy ships. But in combat scenarios (or an ongoing military capabilities competition) against a country such as China that has many UAVs and anti-ship missiles and a capacity for building or acquiring many more, an unfavorable cost exchange ratio can become a very expensive—and potentially unaffordable—approach to defending Navy surface ships against UAVs and anti-ship missiles, particularly in a context of constraints on U.S. defense spending and competing demands for finite U.S. defense funds.

(...continued)

Fleet, Would Require Upgunning Existing Ship Fleets," *Defense Daily*, January 15, 2015: 1-3; Richard Scott, "Offensive Language: USN Sets Out Surface Firepower Strategy," *Jane's International Defence Review*, May 2015: 42-47; Megan Eckstein, "Navy Studying Implications of Distributed Lethality in Wargames Series," *USNI News*, July 9, 2015; Lara Seligman, "Navy Establishes Task Force To Study Impact of Distributed lethality," *Inside the Navy*, July 10, 2015.

[8] Navy cruisers have 122 missile cells; Navy destroyers have 90 or 96 missile cells. Some of these cells are used for storing and launching Tomahawk land attack cruise missiles or anti-submarine rockets. The remainder are available for storing and launching SAMs. A Navy cruiser or destroyer might thus be armed with a few dozen or several dozen SAMs for countering ASCMs and ASBMs. Countering ASCMs or ASBMs with SAMs might sometimes require shooting two SAMs at each ASCM or ASBM.

[9] The missile cells on a Navy cruiser or destroyers are clustered together in an installation called a Vertical Launch System (VLS). VLS cells cannot be reloaded while the ship is underway; a ship needs to return to a port or a calm anchorage to reload its VLS.

[10] Unit procurement costs for ship-launched SAMs in the FY2016 are as follows: about $900,000 for the Rolling Airframe Missile (RAM), about $1.1 million to about $1.5 million for the Evolved Sea Sparrow Missile (ESSM), about $3.9 million for the SM-6 Block 1 missile, about $14 million for the SM-3 Block 1B missile, and more than $20 million for theSM-3 Block IIA missiles. RAM and ESSM are short-range missiles for defense against aircraft and ASCMs. The SM-6 Block 1 is a medium-range missile used for both defense against aircraft and ASCMs, and terminal (i.e., endo-atmospheric) defense against theater-range ballistic missiles. The SM-3 Block 1B and SM-3 Block IIA are used for mid-course (i.e., exo-atmospheric) defense against theater-range ballistic missiles.

SSLs, EMRG, and HVP offer a potential for dramatically improving depth of magazine and the cost exchange ratio:

- **Depth of magazine.** SSLs are electrically powered, drawing their power from the ship's overall electrical supply, and can be fired over and over, indefinitely, as long as the SSL continues to work and the ship has fuel to generate electricity. The EMRG's projectile and the HVP (which are one and the same—see next section) can be stored by the hundreds in a Navy surface ship's weapon magazine.[11]
- **Cost exchange ratio.** An SSL can be fired for a marginal cost of less than one dollar per shot (which is the cost of the fuel needed to generate the electricity used in the shot), while the EMRG's projectile/HVP has an estimated unit procurement cost of about $25,000.[12]

For additional discussion of the strategic and budgetary context in which the programs discussed in this report and other Navy programs may be considered, see CRS Report RL32665, *Navy Force Structure and Shipbuilding Plans: Background and Issues for Congress*, by Ronald O'Rourke.

SSLs, EMRG, and HVP in Brief

SSLs

Overview

The Navy in recent years has leveraged both significant advancements in industrial SSLs and decades of research and development work on military lasers done by other parts of DOD to make substantial progress toward deploying high-energy SSLs[13] on Navy surface ships. Navy surface ships would use high-energy SSLs initially for countering small boats UAVs, and potentially in the future for countering ASCMs and ASBMs as well.[14] High-energy SSLs on Navy ships would be short-range defensive weapons—they would counter targets at ranges of about one mile to perhaps eventually a few miles.[15]

---

[11] In July 2015, the Navy issued a request for information (RFI) to industry for the fabrication of a prototype EMRG mount that would store a minimum of 650 rounds. (RFI for Fabrication of Prototype Mount for Naval Railgun, Solicitation Number: N00024-15-R-4132, FedBizOpps.gov, July 29, 2015. See also Justin Doubleday, "Navy Developing Integrated Mount For Electromagnetic Railgun," *Inside the Navy*, July 31, 2015.)

[12] Sources for cost of HVP: David Martin, "Navy's Newest Weapon Kills at Seven Times the Speed of Sound," CBS News (cbssnews.com), April 7, 2014; Kris Osborn, "Navy Will Test its Electromagnetic Rail Gun aboard DDG 1000," *DefenseTech*, April 15, 2015.

[13] In discussions of potential Navy shipboard lasers, a high-energy laser is generally considered to be a laser with a beam power of at least 10 kilowatts (kW).

[14] In general, lasers would counter small boats and missiles by heating and burning holes in their skins, and causing thermal damage to their interiors. Lasers can also be used to "dazzle" (i.e., interfere with) electro-optical sensors on a boat or missile.

[15] The Navy has also performed research and development work on a different kind of laser, called the free electron laser (FEL). In recent years, Navy research and development work on potential shipboard lasers has shifted more to SSLs. For background information on the FEL, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

In addition to a low marginal cost per shot and deep magazine, potential advantages of shipboard lasers include fast engagement times, an ability to counter radically maneuvering missiles, an ability to conduct precision engagements, and an ability to use lasers for graduated responses ranging from detecting and monitoring targets to causing disabling damage. Potential limitations of shipboard lasers relate to line of sight; atmospheric absorption, scattering, and turbulence (which prevent shipboard lasers from being all-weather weapons); an effect known as thermal blooming that can reduce laser effectiveness; countering saturation attacks; possible adversary use of hardened targets and countermeasures; and risk of collateral damage, including damage to aircraft and satellites and permanent damage to human eyesight, including blinding. These potential advantages and limitations are discussed in greater detail in the **Appendix**.

Selected Key Developments

Key developments in the Navy's high-energy SSL development effort include the following:

- Between 2009 and 2012, the Navy successfully tested a prototype SSL called the Laser Weapon System (LaWS) against UAVs in a series of engagements that took place initially on land and subsequently on a Navy ship at sea.
- Between 2010 and 2011, the Navy tested another prototype SSL called the Maritime Laser Demonstration (MLD) in a series of tests that culminated with an MLD installed on a Navy ship successfully engaging a small boat.
- In April 2013, the Navy announced that it planned to install LaWS on the USS *Ponce* (pronounced pon-SAY)—a converted amphibious ship that is operating in the Persian Gulf as an interim Afloat Forward Staging Base (AFSB[I]) [16]—to conduct evaluation of shipboard lasers in an operational setting against swarming boats and swarming UAVs.[17] The system was installed in August 2014 (see **Figure 1**, **Figure 2**, and **Figure 3**).
- In March 2014, it was reported that the Navy anticipated moving to a shipboard laser program of record in "the FY2018 time frame" and achieving an initial operational capability (IOC) with a shipboard laser in FY2020 or FY2021.[18]

---

[16] An AFSB operates as a "mother ship" for Navy helicopter and small boat operations. The Ponce is serving as an interim AFSB pending the arrival of a new AFSB that is currently being built.

[17] "Navy Leaders Announce Plans for Deploying Cost-Saving Laser Technology," *Navy News Service*, April 8, 2013; Thom Shanker, "Navy Deploying Laser Weapon Prototype Near Iran," *New York Times*, April 9, 2013: 4; Mike McCarthy, "Navy Deploying Laser For Taking Out Drones," *Defense Daily*, April 9, 2013; Graham Warwick, "U.S. Navy Planning Gulf Deployment For Laser Weapon," *Aerospace Daily & Defense Report*, April 9, 2013: 6; Megan Eckstein, "Navy-Built Laser Weapon System Will Begin Demo On Ponce In Early 2014," *Inside the Navy*, April 15, 2013. See also Lara Seligman, "Navy-built LaWS To Begin Demo This Summer, IOC Slated For FY-20-21," *Inside the Navy*, March 24, 2014; Office of Naval Research, "All Systems Go: Navy's Laser Weapon Ready for Summer Deployment," *Navy News Service*, April 7, 2014.

Swarming refers to the use of boats and UAVs in large numbers, or swarms, in an attempt to confuse and overwhelm a target ship's defensive systems.

[18] Lara Seligman, "Navy-built LaWS To Begin Demo This Summer, IOC Slated For FY-20-21," *Inside the Navy*, March 24, 2014. A program of record, or POR, is a term sometimes used by DOD officials that means, in general, a program in the Future Years Defense Plan (FYDP) that is intended to provide a new, improved, or continuing materiel, weapon, or information system or service capability in response to an approved need. The term is sometimes used to refer to a program in a service's budget for procuring and deploying an operational weapon system, as opposed to a research and development effort that might or might not eventually lead to procurement and deployment of an operational weapon system.

- In December 2014, the Navy declared LaWS on the Ponce to be an "operational" system.[19]
- In January 2016, the Navy stated that it anticipated releasing a directed energy weapon roadmap in February 2016.[20]

---

[19] A December, 11, 2014, press report stated

> The Navy's first-of-a-kind laser deployed on a vessel sailing in the Persian Gulf has been declared operational and can be used by the crew to defend itself against potential threats, the service's head of the Office of Naval Research said on Wednesday [December 10, 2014].
>
> Rear Adm. Matthew Klunder told reporters on a conference call that Central Command has been green lighted to use the laser in the event of a threat, approval that has been passed along to the ship's commanding officer. The 30-kilowat laser, known as the Laser Weapon System, or LaWS, was installed on the USS Ponce in August [2014].
>
> The ship later departed for the Persian Gulf and the LaWS successfully carried out operational testing recently by striking a fast attack boat and drone, Klunder said, adding that this marks the "historic" first ever operational deployment of a directed energy weapon.
>
> (Mike McCarthy, "Navy Authorized To Use Ship-Based Laser In Battle," *Defense Daily*, December 11, 2014: 3. See also Sam LaGrone, "U.S. Navy Allowed to Use Persian Gulf Laser for Defense," *USNI News*, December 10, 2014; Philip Ewing, "Navy Declares Laser Weapon 'Operational,'" *Politico Pro (Pro Defense Report)*, December 10, 2014.)

The Navy testified on February 24, 2016, that

> the Solid State Laser Quick Reaction Capability (SSL-QRC) was fielded as a science and technology demonstration aboard the USS PONCE. It was successfully demonstrated as an effective weapon system and was subsequently transitioned to the fleet in the Central Command area of responsibility and is now an operational system.
>
> (Statement of Rear Admiral Mathias W. Winter, United States Navy, Chief of Naval Research, Before the Emerging Threats and Capabilities Subcommittee of the House Armed Services Committee on The Fiscal Year 2017 Budget Request, February 24, 2016, p. 15.)

[20] Justin Doubleday, "Winter: Navy Directed-Energy Strategy To Be Released This Month," *Inside the Navy*, February 1, 2016.

Figure 1. Laser Weapon System (LaWS) on USS Ponce

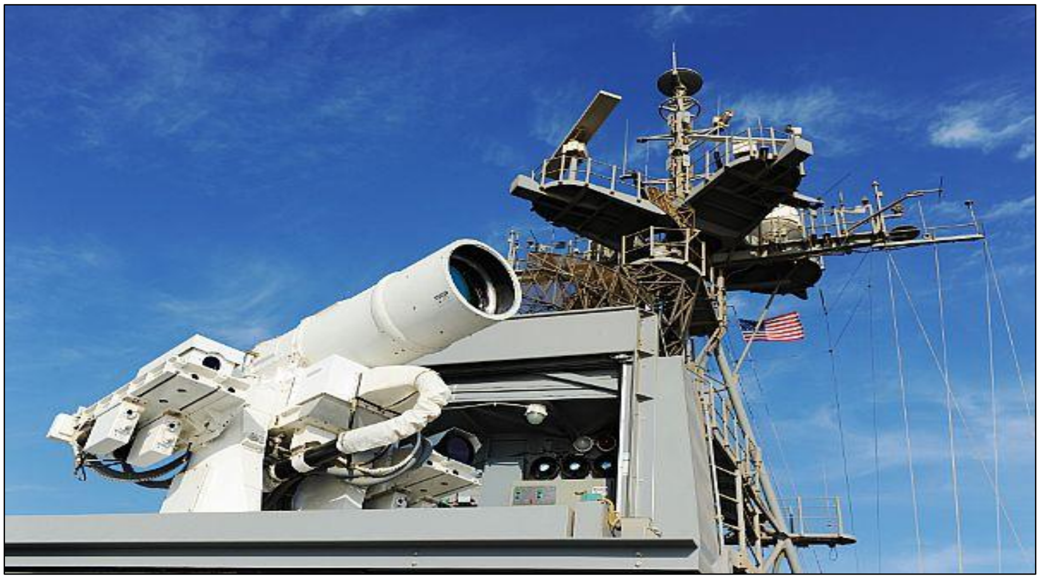

Source: Navy photograph dated November 16, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

Figure 2. Laser Weapon System (LaWS) on USS Ponce

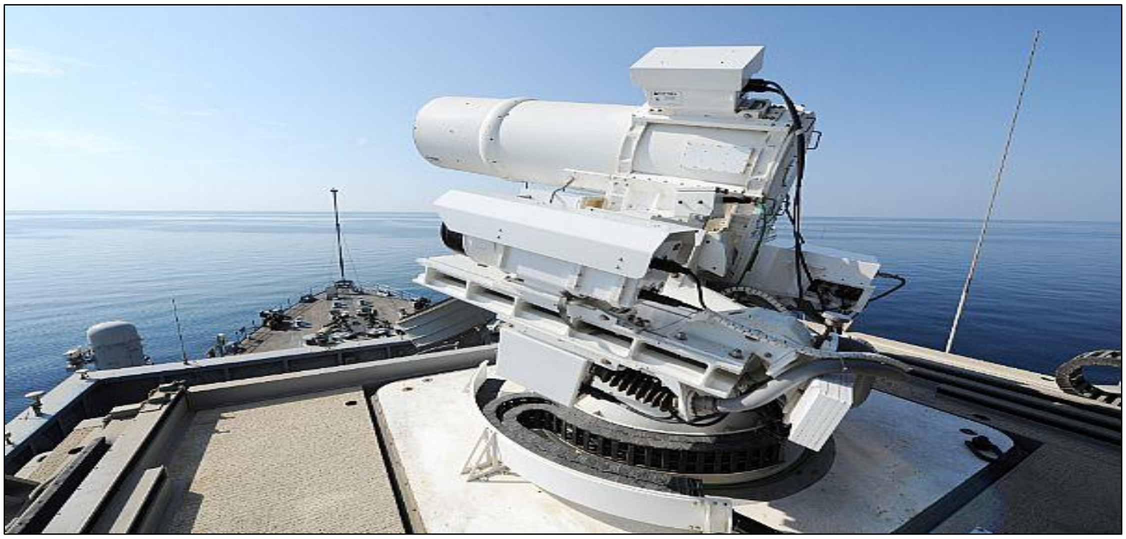

Source: Navy photograph dated November 17, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

Figure 3. Laser Weapon System (LaWS) on USS Ponce

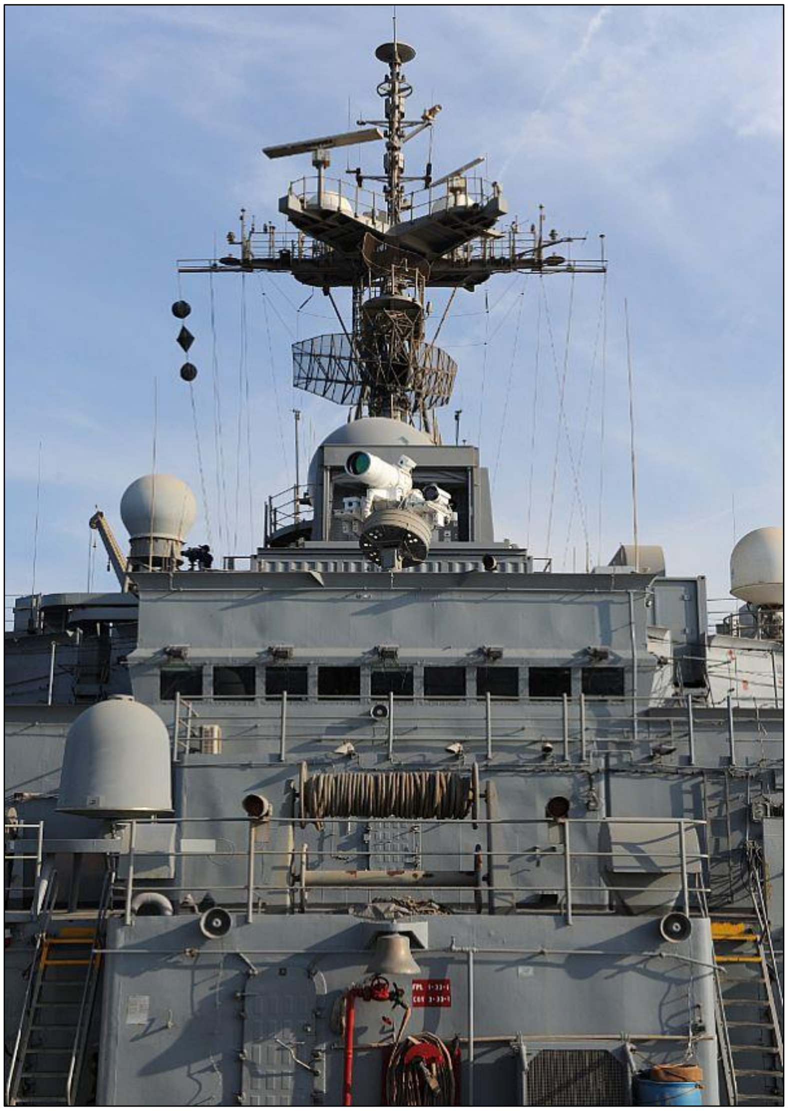

Source: Navy photograph dated November 16, 2014, accompanying David Smalley, "Historic Leap: Navy Shipboard Laser Operates in Arabian Gulf," *Navy News Service*, December 10, 2014, accessed August 12, 2015, at http://www.navy.mil/list_all.asp?id=84805.

SSL Technology Maturation (SSL-TM) Program

LaWS has a reported beam power of 30 kilowatts (kW),[21] which is strong enough to counter small boats and UAVs. As a follow-on effort to LaWS and MLD, the Navy initiated the SSL Technology Maturation (SSL-TM) program, in which industry teams led by BAE Systems, Northrop Grumman, and Raytheon, among others, competed to develop a shipboard laser with a beam power of 100 kW to 150 kW, which would provide increased effectiveness against small boats and UAVs.[22] Boosting beam power further—to something like 200 kW or 300 kW—could permit a laser to counter at least some ASCMs. Even stronger beam powers—on the order of at least several hundred kW, if not one megawatt (MW) or more—could improve a laser's effectiveness against ASCMs and enable it to counter ASBMs.[23]

---

[21] See, for example, Mike McCarthy, "Navy Authorized To Use Ship-Based Laser In Battle," *Defense Daily*, December 11, 2014: 3.

On October 22, 2015, DOD announced that it had selected Northrop Grumman as the winner of the SSL-TM competition. DOD's contract-award announcement stated:

> Northrop Grumman Space and Mission Systems Corp., Redondo Beach, California, is being awarded a $53,151,809 cost-plus-fixed-fee contract for the Solid State High Power Laser Weapon System Demonstrator (LWSD) program.... The Office of Naval Research seeks to continue the advancement of SSL weapon system designs, architectures, and component technologies. The government believes that improvements in lethality may be achieved through maturation and optimization of a variety of system characteristics, including laser power, beam quality, beam director architecture, and other physical and optical aspects of the laser, beam director, and system design. Leveraging our experience and internal investments, the Northrop Grumman team is ready to fully support the three phases of the LWSD program. This contract contains options, which if exercised, will bring the contract value to $91,057,597. Work will be performed in Redondo Beach, California, and is expected to be completed Oct. 21, 2016. If options are exercised, work will continue through July 7, 2018.... This contract was competitively procured under the Office of Naval Research broad agency announcement 15-0005 entitled "Solid State, High Power Laser Weapon System Demonstrator (LWSD) Design, Development and Demonstration for Surface Navy, USN." Six proposals were received in response to this solicitation.[24]

A December 22, 2016, Northrop Grumman news release about the October 22, 2016, contract award stated:

> During Phase 1 of the LWSD contract, Northrop Grumman will develop a detailed design for the new system. Phase 2 will include assembly and ground test of the system, while Phase 3 will comprise at-sea testing of the system aboard the Navy's Self Defense Test Ship (SDTS). The Navy will lead this testing with Northrop Grumman providing technical support. The SDTS is the former USS Paul F. Foster (DD-964).
>
> According to Renard, Northrop Grumman's LWSD is well suited to support the Navy's planned initial testing on the SDTS. The company has designed its system to be installed, however, with minimal modification or additional costs, for demonstration on the Navy's DDG-51 FLT II class destroyers.[25]

---

[22] For more on the SSL-TM program, see Office of Naval Research, "Solid-State Laser Technology Maturation Program," accessed August 11, 2015, at http://www.onr.navy.mil/Media-Center/Fact-Sheets/Solid-State-Laser-Technology-Maturation-Program.aspx; Office of Naval Research, "Solid State Laser Technology Maturation Program," September 2012, accessed August 11, 2015, at http://www.onr.navy.mil/~/media/Files/Fact-Sheets/35/Solid-State-Laser-Technology-Maturation-Program-2012-a.ashx; Office of Naval Research, "Research and Development/Technology Maturation of Solid State High Power Laser Weapon Systems, Subsystems, and/or Components for Surface Navy, USN, Broad Agency Announcement (BAA)," ONR BAA # 12-019, 2012, accessed August 11, 2015, at http://www.onr.navy.mil/~/media/files/funding-announcements/baa/2012/12-019.ashx; Future Force, "Developing a High-Energy Laser for the Navy," January 23, 2015, accessed August 11, 2015, at http://futureforce.navylive.dodlive.mil/2015/01/high-energy-laser/.

[23] For additional discussion, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke, particularly the section entitled "Required Laser Power Levels for Countering Targets" and Appendix A on "Laser Power Levels Required to Counter Targets."

[24] DOD contract award announcements for October 22, 2015, accessed December 18, 2015, at: http://www.defense.gov/News/Contracts/Contract-View/Article/625630.

[25] "US Navy Selects Northrop Grumman to Design and Produce Shipboard Laser Weapon System Demonstrator," (continued...)

Figure 4 shows an artist's rendering of LWSD installed on the Navy's Self Defense Test Ship (the USS Paul F. Foster [DD-964], an old Spruance [DD-963] class destroyer).

Figure 4. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship

Artist's rendering

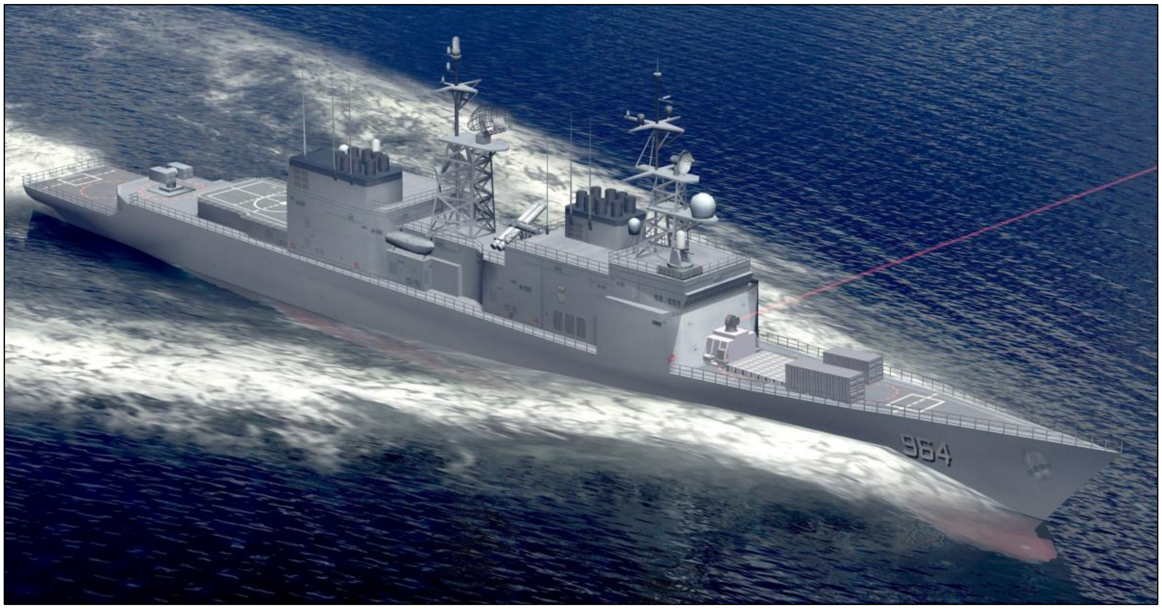

Source: Cropped version of image accessed on March 18, 2016, at http://media.globenewswire.com/cache/189/hires/39412.jpg.

Figure 5 is a detail from the above photo.

(...continued)

December 22, 2016, accessed March 18, 2016, at: http://www.globenewswire.com/newsarchive/noc/press/pages/news_releases.html?d=10158731 . See also Richard Scott, "Northrop Grumman To Build on MLD for SSL Demonstrator," *IHS Jane's International Defence Review*, February 2016: 5.

Figure 5. Laser Weapon System Demonstrator (LWSD) on Self Defense Test Ship

Artist's rendering

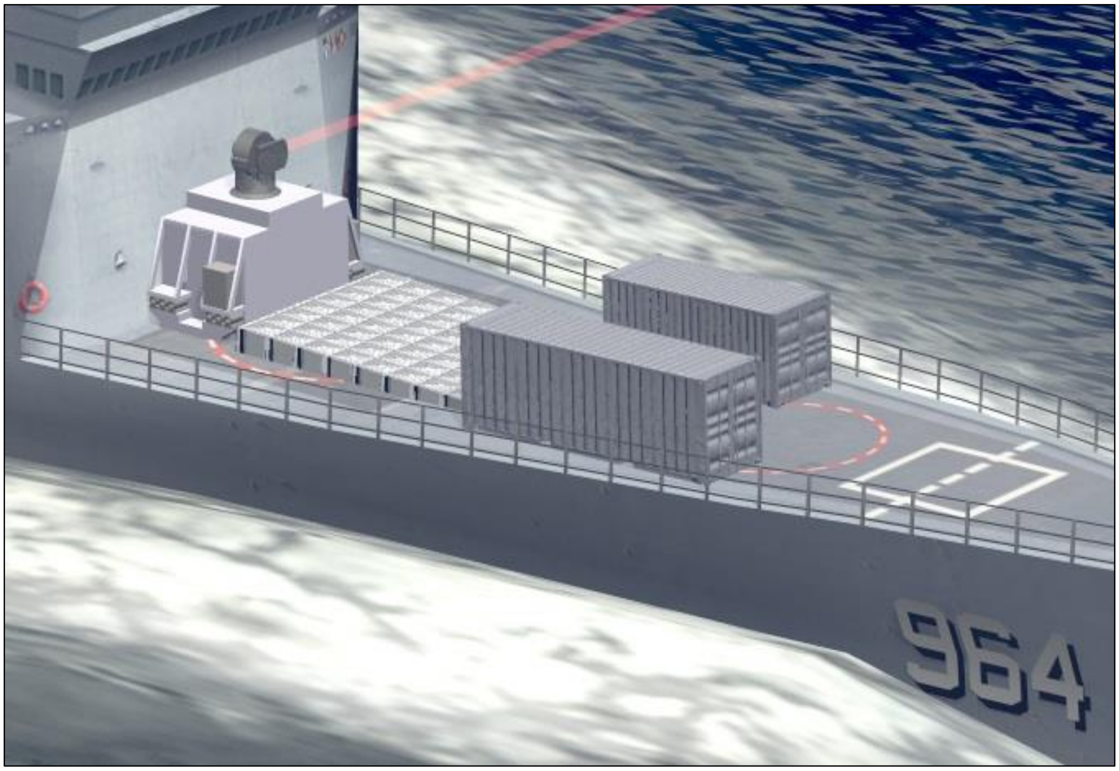

Source: Cropped version of image accessed on March 18, 2016, at http://media.globenewswire.com/cache/189/hires/39412.jpg.

Figure 6 Is an Office of Naval Research (ONR) graphic illustration of the LWSD's major components.

Figure 6. ONR Graphic of LWSD Components

Artist's rendering

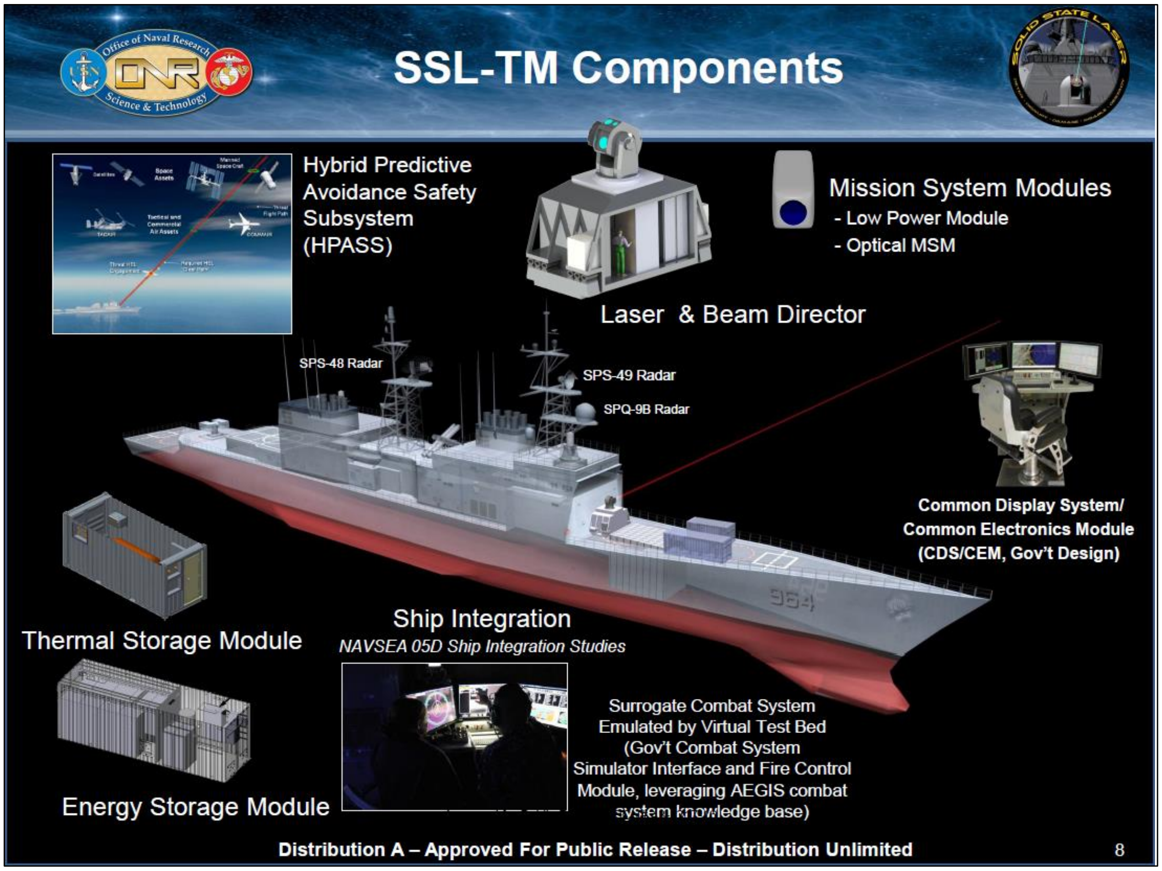

SSL-TM Components

Hybrid Predictive Avoidance Safety Subsystem (HPASS)

Mission System Modules
- Low Power Module
- Optical MSM

Laser & Beam Director

SPS-48 Radar

SPS-49 Radar

SPQ-9B Radar

Common Display System/ Common Electronics Module (CDS/CEM, Gov't Design)

Thermal Storage Module

Ship Integration

NAVSEA 05D Ship Integration Studies

Energy Storage Module

Surrogate Combat System Emulated by Virtual Test Bed (Gov't Combat System Simulator Interface and Fire Control Module, leveraging AEGIS combat system knowledge base)

Distribution A – Approved For Public Release – Distribution Unlimited

8

Source: Slide from February 2016 ONR briefing to CRS on SSL-TM program, received from Navy Office of Legislative Affairs February 26, 2016.

Directed Energy Roadmap

A July 28, 2015, press report stated:

> [Secretary of the Navy Ray] Mabus said he would release a DE [directed energy][26] roadmap this fall that "charts our course for research, development, and fielding of high power radio frequency weapons, lasers, and directed energy countermeasures. And I will follow it up with my guidance to the Program Objective Memorandum for [Fiscal Year 2018],[27] which, importantly, establishes a resource sponsor and a program of record."...
>
> Also meant to help quicken the pace of progress, the Office of Naval Research will take lessons learned from the [USS] Ponce to inform the Solid State Laser Technology Maturation program that aims to produce a 100-150 kilowatt laser prototype for at-sea testing in 2018, or sooner if possible. Rear Adm. Bryant Fuller, Naval Sea Systems Command (NAVSEA) chief engineer, said... that everything the Navy learned about rules

---

[26] Lasers and another class of weapons called high-power microwave (HPM) weapons are referred to collectively as directed-energy weapons because they achieve their effects by directing electromagnetic energy at their targets.

[27] The Program Objective Memorandum (POM) is an internal DOD document that guides the preparation of a budget for a particular fiscal year.

> of engagement and how to use LaWS in an operational environment would apply to larger laser weapons as well. Leveraging the operational knowledge Ponce gained will help the Navy field whatever comes out of the SSL-TM effort much more rapidly.
>
> In the meantime, Mabus said the Laser Weapon System (LaWS) will continue its work in the Middle East after early success led officials to extend its deployment.[28]

EMRG

In addition to SSLs, the Navy since 2005 has been developing EMRG, a cannon that uses electricity rather than chemical propellants (i.e., gunpowder charges) to fire a projectile.[29] In EMRG, "magnetic fields created by high electrical currents accelerate a sliding metal conductor, or armature, between two rails to launch projectiles at [speeds of] 4,500 mph to 5,600 mph,"[30] or roughly Mach 5.9 to Mach 7.4 at sea level.[31] Like SSLs, EMRG draws its power from the ship's overall electrical supply.[32] The Navy originally began developing EMRG as a naval surface fire support (NSFS) weapon for supporting U.S. Marines operating ashore, but subsequently determined that the weapon also has potential for defending against ASCMs and ASBMs.[33] In response to Section 243 of the FY2012 National Defense Authorization Act (H.R. 1540/P.L. 112-81 of December 31, 2011), the Navy in September 2012 submitted to the congressional defense committees a report on the EMRG development effort.[34]

Following tests with early Navy-built EMRG prototypes, the Navy funded the development of two industry-built EMRG prototype demonstrators, one by BAE Systems and the other by General Atomics (see **Figure 7** and **Figure 8**).

---

[28] Megan Eckstein, "Mabus: Adversaries Showing Interest in Directed Energy; Navy Needs to Move Faster," *USNI News*, July 28, 2015.

[29] Because it uses electricity rather than a powder charge to accelerate the projectile, Navy officials sometimes refer to EMRG as a launcher rather than a gun or cannon.

[30] Grace Jean, "With a Bang, Navy Begins Tests on EM Railgun Prototype Launcher," Navy News Service, February 28, 2012, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=65577.

[31] The speed of sound in air (i.e., Mach 1), varies with altitude; at sea level, it is approximately 761 miles an hour. (See for example, the table entitled "Speed of Sound at Different Altitudes," accessed August 12, 2015, at http://www.fighter-planes.com/jetmach1.htm.

[32] Unlike SSLs, however, EMRG is not a directed energy weapon, because it achieves its effects by firing a physical projectile at the target, not by directing electromagnetic energy at the target. See also footnote 26.

[33] For a recent article discussing the use of EMRG in countering ASCMs and ASBMs, see Sam LaGrone, "Navy Wants Rail Guns to Fight Ballistic and Supersonic Missiles Says RFI," *USNI News*, January 5, 2015.

[34] U.S. Navy, *Electromagnetic Railgun System: Final Report to the Congressional Defense Committees*, August 2012, with cover letters dated September 18, 2012. For a press report discussing the Navy's report to Congress, see Dan Taylor, "Stackley: Navy Identifies Four Technical Hurdles To Railgun Development," *Inside the Navy*, November 19, 2012.

Figure 7. Industry-Built EMRG Prototype Demonstrator

BAE prototype

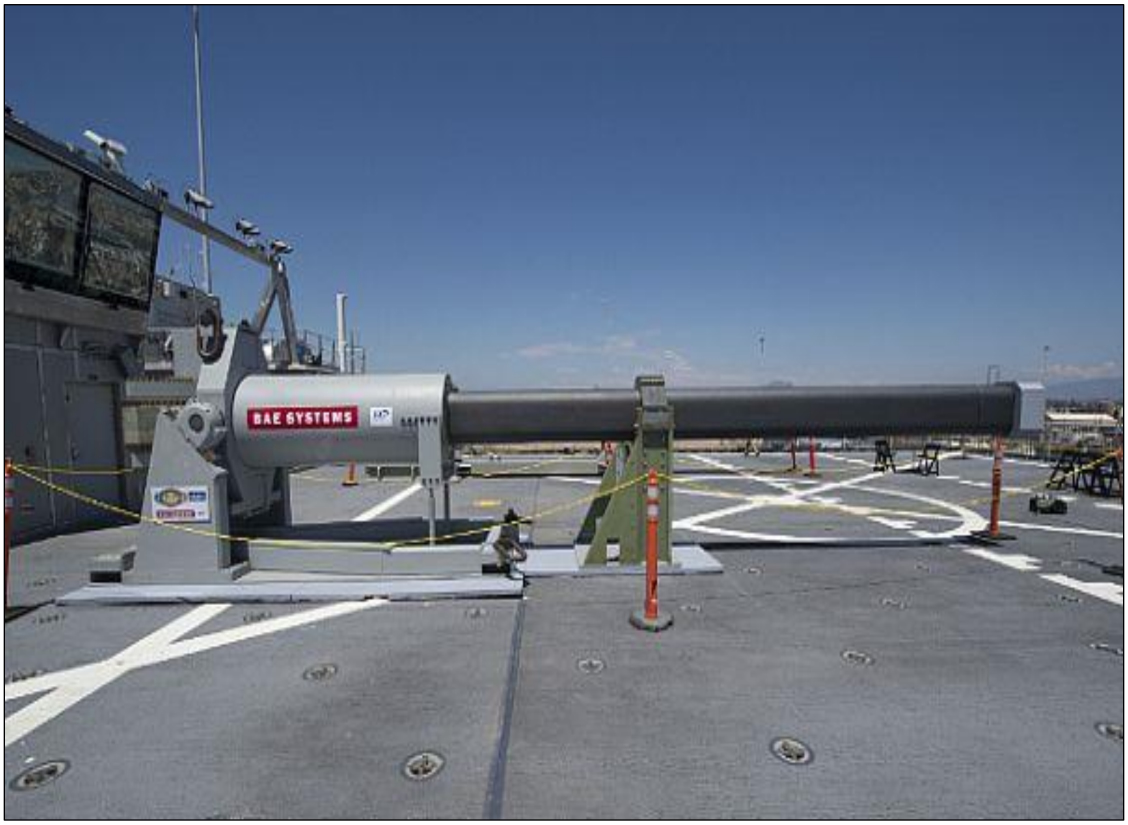

Source: Navy photograph dated July 8, 2014, associated with Office of Naval Research Public Affairs, "From Research to Railgun: Revolutionary Weapon at Future Force EXPO," *Navy News Service*, January 13, 2015, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=85166.

Figure 8. Industry-Built EMRG Prototype Demonstrator

General Atomics prototype

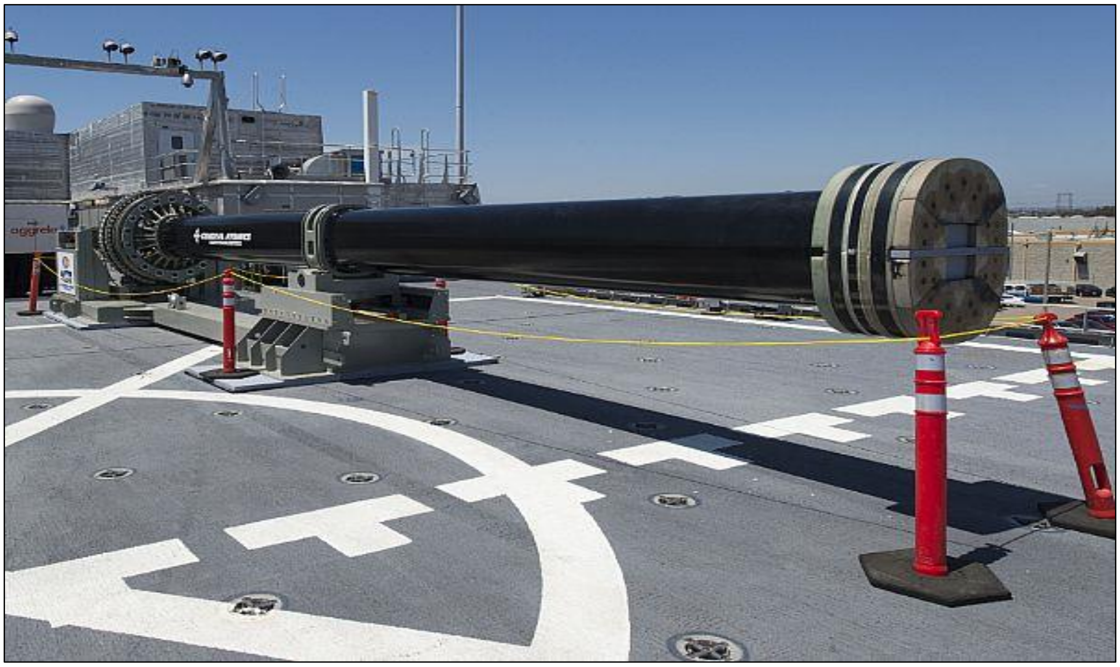

Source: navy photograph dated July 8, 2014, accessed August 12, 2015, at http://www.navy.mil/view_image.asp?id=180994.

The two industry-built prototypes are designed to fire projectiles at energy levels of 20 to 32 megajoules,[35] which is enough to propel a projectile 50 to 100 nautical miles.[36] (Such ranges might refer to using the EMRG for NSFS missions. Intercepts of ASCMs and ASBMs might take place at much shorter ranges.) The Navy began evaluating the two industry-built prototypes in 2012.

In April 2014, the Navy announced that it plans to temporarily install a prototype EMRG aboard a Navy Joint High Speed Vessel (JHSV) in FY2016, for use in at-sea tests.[37] **Figure 9** is an artist's rendering of that installation.

In January 2015, it was reported that the Navy is projecting that EMRG could become operational on a Navy ship between 2020 and 2025.[38] In April 2015, it was reported that the Navy is considering installing an EMRG on a Zumwalt (DDG-1000) class destroyer by the mid-2020s.[39]

---

[35] The Navy states that "a megajoule is a measurement of energy associated with a mass traveling at a certain velocity. In simple terms, a one-ton vehicle moving at 100 mph equals a magajoule of energy." (Office of Naval Research Public Affairs, "Navy Sets New World Record with Electromagnetic Railgun Demonstration," *Navy News Service*, December 10, 2010, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=57690.)

[36] Grace Jean, "With a Bang, Navy Begins Tests on EM Railgun Prototype Launcher," Navy News Service, February 28, 2012, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=65577.

[37] Naval Sea Systems Command Office of Corporate Communication, "Navy to Deploy Electromagnetic Railgun Aboard JHSV," *Navy News Service*, April 7, 2014, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=80055.

[38] Sam LaGrone, "Navy Wants Rail Guns to Fight Ballistic and Supersonic Missiles Says RFI," *USNI News*, January 5, 2015.

[39] Sam LaGrone, "Navy Considering Railgun for Third Zumwalt Destroyer," *USNI News*, February 5, 2015 (updated (continued...)

Figure 9. EMRG Prototype Demonstrator Installed on a JHSV

Artist's rendering

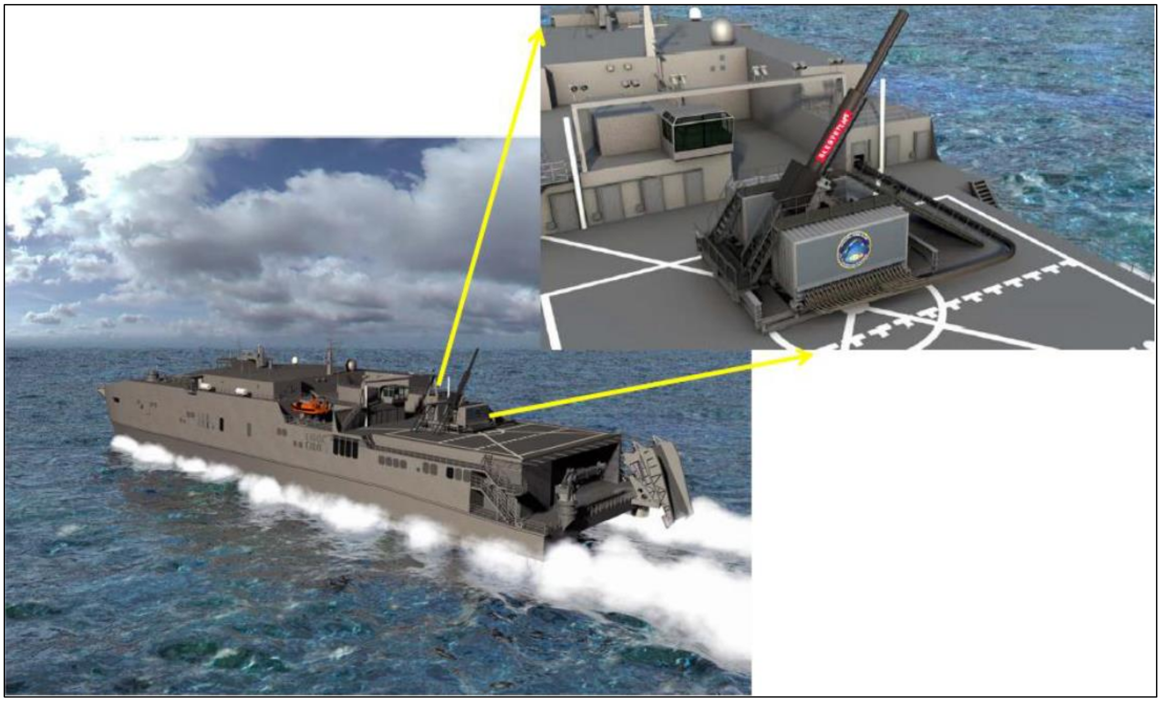

Source: Briefing slide entitled "FY16 At-Sea Test of Railgun" in Navy briefing entitled "Railgun Program Overview," undated but posted at InsideDefense.com on April 14, 2015. (InsideDefense.com states that the briefing was presented at a public conference on April 14, 2015.)

HVP

As the Navy was developing EMRG, it realized that the guided projectile being developed for EMRG could also be fired from 5-inch and 155mm powder guns.[40] Navy cruisers each have two

---

(...continued)

February 11, 2015); Mike McCarthy, "Navy Aiming To Put Railgun On Third Zumwalt Destroyer," *Defense Daily*, February 6, 2015; Kris Osborn, "Navy Will Test its Electromagnetic Rail Gun aboard DDG 1000," *DefenseTech*, April 15, 2015. For more on Zumwalt-class destroyers, see CRS Report RL32109, *Navy DDG-51 and DDG-1000 Destroyer Programs: Background and Issues for Congress*, by Ronald O'Rourke.

[40] The Navy describes the HVP as "a next generation, common, low drag, guided projectile capable of completing multiple missions for gun systems such as the Navy 5-Inch, 155-mm, and future railguns.... HVP's low drag aerodynamic design enables high velocity, maneuverability, and decreased time-to-target. These attributes coupled with accurate guidance electronics provide low cost mission effectiveness against current threats and the ability to adapt to air and surface threats of the future." (Office of Naval Research, Hypervelocity Projectile," September 2012, accessed August 14, 2015, at http://www.onr.navy.mil/~/media/Files/Fact-Sheets/35/Hypervelocity-Projectile-2012B.ashx.) The Navy states that HVP weighs 23 pounds. (Source: David Martin, "Navy's Newest Weapon Kills at Seven Times the Speed of Sound," *CBS News (cbssnews.com)*, April 7, 2014.)

BAE Systems states that HVP is 24 inches long and weighs 28 pounds, including a 15-pound payload. The total length and weight of an HVP launch package, BAE Systems states, is 26 inches and 40 pounds. BAE states that the maximum rate of fire for HVP is 20 rounds per minute from a Mk 45 5-inch gun, 10 rounds per minute from the 155mm gun on DDG-1000 class destroyers (called the Advanced Gun System, or AGS), and 6 rounds per minute from EMRG. HVP's firing range, BAE Systems states, is more than 40 nautical miles (when fired from a Mk 45 Mod 2 5-inch gun), more than 50 nautical miles (Mk 45 Mod 4 5-inch gun), more than 70 nautical miles (155mm gun on DDG-1000 class destroyers), and more than 100 nautical miles (EMRG). (BAE Systems, "Hypervelocity Projectile (HVP)," 2014, (continued...)

5-inch guns, and most Navy destroyers each have one 5-inch gun. The Navy's three new Zumwalt class (DDG-1000) destroyers, which are under construction, each have two 155mm guns.

The projectile is a hypervelocity projectile when fired from either EMRG or a powder gun, but the term HVP tends to be used more frequently in connection with the concept of firing it from a powder gun. **Figure 10** and **Figure 11** show the HVP.

Figure 10. Photograph Showing HVP

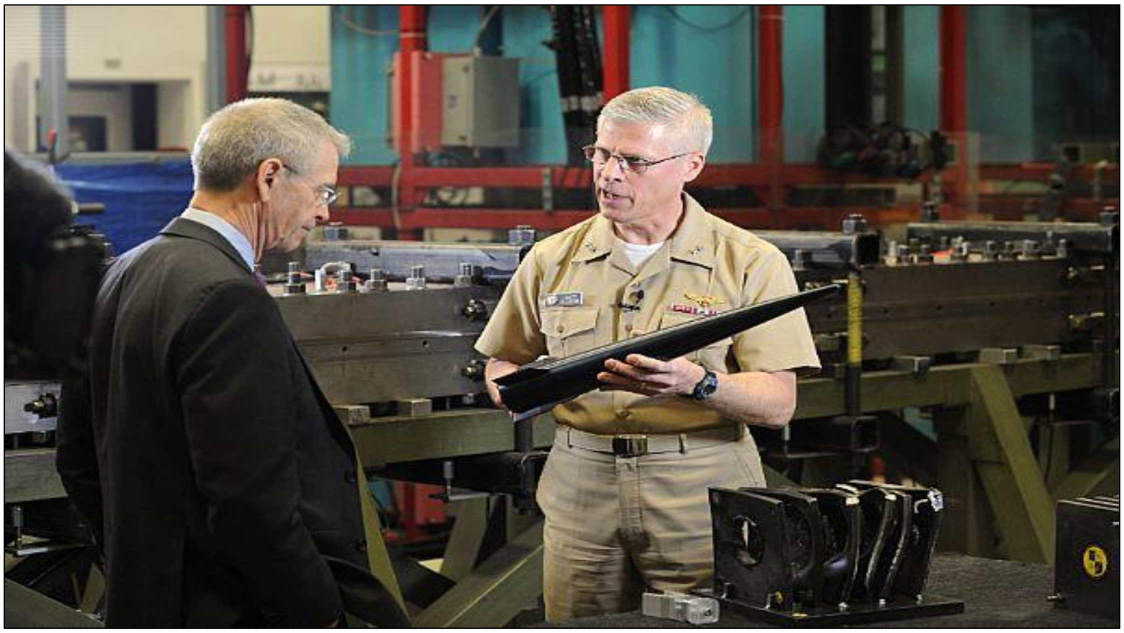

Source: Navy photograph dated April 4, 2014, with a caption that reads: "Rear Adm. Matthew Klunder, chief of naval research, shows off a Hypervelocity Projectile (HVP) to CBS News reporter David Martin during an interview held at the Naval Research Laboratory's materials testing facility. The HVP is a next-generation, common, low drag, guided projectile capable of completing multiple missions for gun systems such as the Navy 5-inch, 155-mm, and future railguns," accessed August 12, 2015, at http://www.navy.mil/view_image.asp?id=174517.

(...continued)

accessed August 14, 2015, at http://www.baesystems.com/download/BAES_178505/hyper-velocity-projectile-hvp-datasheet.)

In July 2015, the Navy issued a request for information (RFI) to industry for the fabrication of a prototype EMRG mount capable of handling an integrated launch weight package of 22 kg, or about 48.5 pounds. (RFI for Fabrication of Prototype Mount for Naval Railgun, Solicitation Number: N00024-15-R-4132, FedBizOpps.gov, July 29, 2015. See also Justin Doubleday, "Navy Developing Integrated Mount For Electromagnetic Railgun," *Inside the Navy*, July 31, 2015.)

Figure 11. HVP

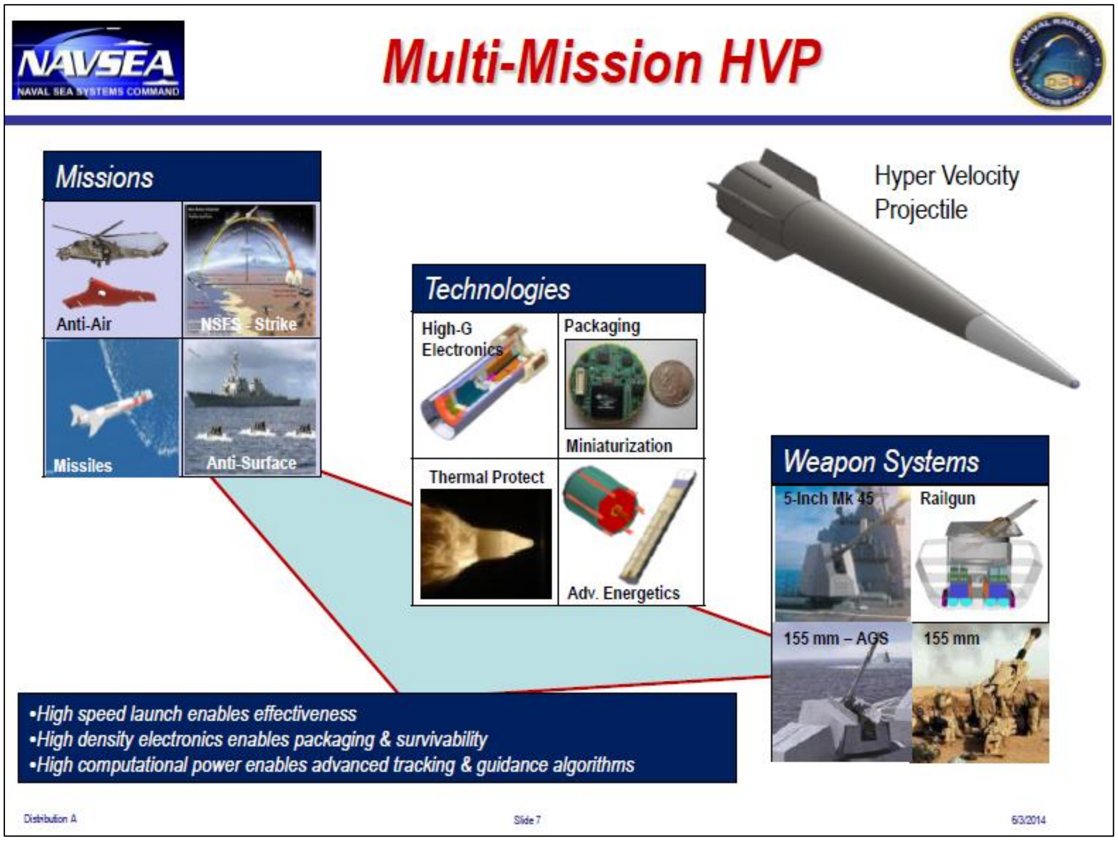

NAVSEA
NAVAL SEA SYSTEMS COMMAND

Multi-Mission HVP

Missions
- Anti-Air
- NSFS - Strike
- Missiles
- Anti-Surface

Technologies
- High-G Electronics
- Packaging
- Miniaturization
- Thermal Protect
- Adv. Energetics

Hyper Velocity Projectile

Weapon Systems
- 5-Inch Mk 45
- Railgun
- 155 mm – AGS
- 155 mm

- •High speed launch enables effectiveness
- •High density electronics enables packaging & survivability
- •High computational power enables advanced tracking & guidance algorithms

Distribution A

Slide 7

6/3/2014

Source: Slide 7 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

When fired from 5-inch powder guns, the projectile achieves a speed of roughly Mach 3, which is roughly half the speed it achieves when fired from EMRG, but more than twice the speed of a conventional 5-inch shell fired from a 5-inch gun.[41] This is apparently fast enough for countering at least some ASCMs. The Navy states that "The HVP—combined with the MK 45 [5-inch gun][42]—will support various mission areas including naval surface fire support, and has the capacity to expand to a variety of anti-air threats, [and] anti-surface [missions], and could expand the Navy's engagement options against current and emerging threats."[43]

One advantage of the HVP/5-inch gun concept is that the 5-inch guns are already installed on Navy cruisers and destroyers, creating a potential for rapidly proliferating HVP through the cruiser-destroyer force, once development of HVP is complete and the weapon has been integrated into cruiser and destroyer combat systems. **Figure 12** shows HVP launch packages configured for 5-inch guns, 155mm guns, and EMRG.

---

[41] Source: Sam LaGrone, "Updated: Navy Researching Firing Mach 3 Guided Round from Standard Deck Guns," *USNI News*, June 1, 2015 (updated June 2, 2015).

[42] The type of 5-inch gun on Navy cruisers and destroyers is called the Mark 45.

[43] Naval Surface Warfare Center Dahlgren Division Corporate Communications, "DEPSECDEF Loads HVP on Test Range, Observes Repetitive Rate Electromagnetic Railgun's Commissioning Series," *Navy News Service*, May 8, 2015, accessed August 12, 2015, at http://www.navy.mil/submit/display.asp?story_id=86987.

Figure 12. HVP Launch Packages

Launch packages for 5-inch gun, 155mm gun, and EMRG

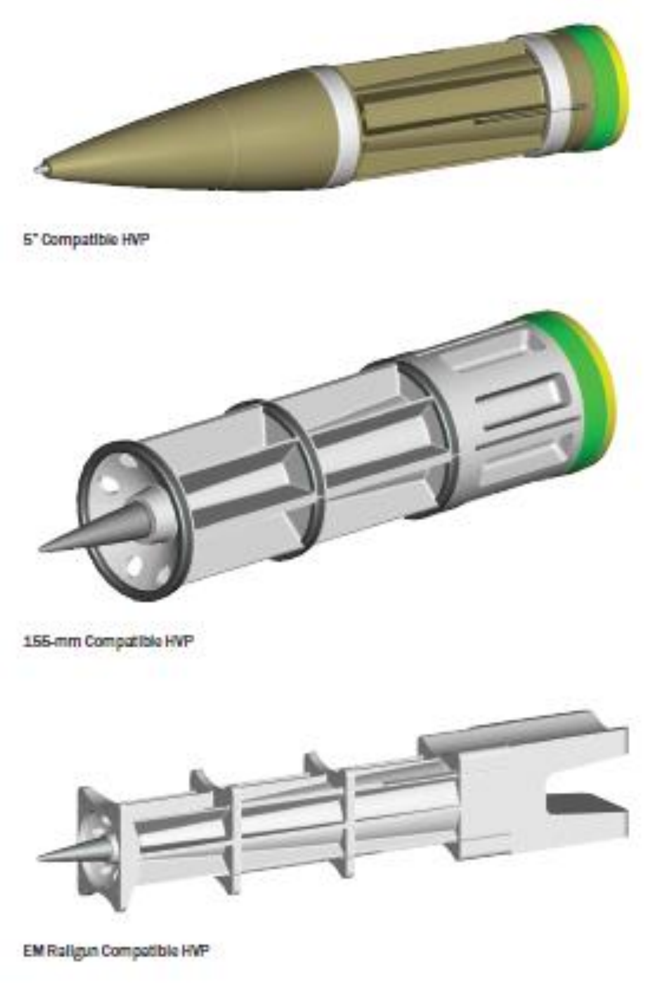

Source: BAE Systems, "Hypervelocity Projectile (HVP)," 2014, accessed August 14, 2015, at http://www.baesystems.com/download/BAES_178505/hyper-velocity-projectile--datasheet.

Figure 13 is a slide showing the potential application of HVP to 5-inch power guns, 155mm powder guns, and EMRG. The first line of the slide, for example, discusses HVP's use with 5-inch powder guns, stating that it uses a high-explosive (HE) warhead for the NSFS mission;[44] that a total of 113 5-inch gun barrels are available in the fleet (which could be a reference to 22 cruisers with two guns each, and 69 destroyers with one gun each); and that as a game-changing capability, it is guided and can be used at ranges of up to 26 nautical miles to 41 nautical miles for NSFS operations, for countering ASCMs, and for anti-surface warfare (ASuW) operations (i.e., attacking surface ships and craft).

---

[44] The "KE" in the next line down means that when fired from EMRG, the projectile can alternatively attack targets using its own kinetic energy (i.e., by simply impacting the target at hypersonic speed).

Figure 13. HVP Application to Various Launchers

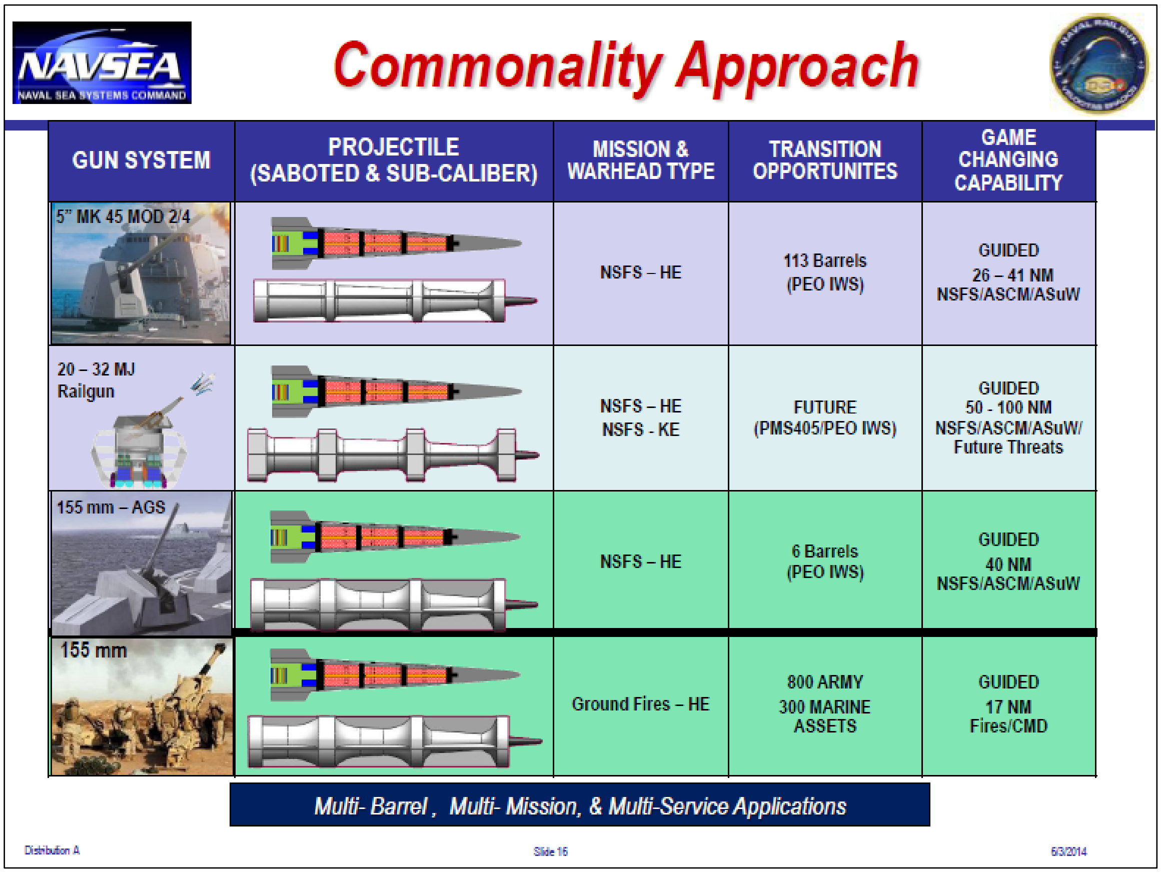

Commonality Approach

GUN SYSTEM	PROJECTILE (SABOTED & SUB-CALIBER)	MISSION & WARHEAD TYPE	TRANSITION OPPORTUNITES	GAME CHANGING CAPABILITY
5" MK 45 MOD 2/4		NSFS – HE	113 Barrels (PEO IWS)	GUIDED 26 – 41 NM NSFS/ASCM/ASuW
20 – 32 MJ Railgun		NSFS – HE NSFS - KE	FUTURE (PMS405/PEO IWS)	GUIDED 50 - 100 NM NSFS/ASCM/ASuW/ Future Threats
155 mm – AGS		NSFS – HE	6 Barrels (PEO IWS)	GUIDED 40 NM NSFS/ASCM/ASuW
155 mm		Ground Fires – HE	800 ARMY 300 MARINE ASSETS	GUIDED 17 NM Fires/CMD

Multi- Barrel , Multi- Mission, & Multi-Service Applications

Distribution A

Slide 16

6/3/2014

Source: Slide 16 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

Figure 14 is a not-to-scale illustration of how HVPs fired from EMRGs and 5-inch guns can be used to counter various targets, including ASCMs and ASBMs.

Figure 14. Navy Slide Depicting Operations Against Various Target Types

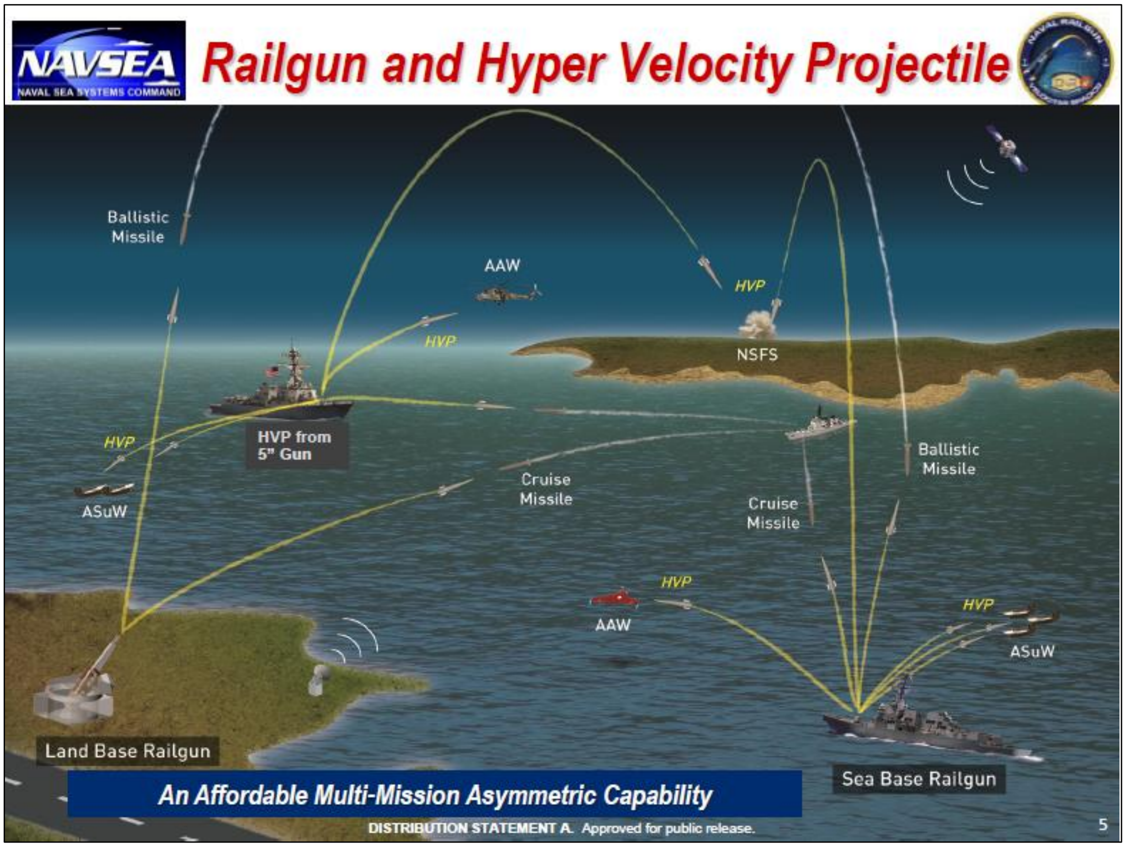

NAVSEA NAVAL SEA SYSTEMS COMMAND

Railgun and Hyper Velocity Projectile

An Affordable Multi-Mission Asymmetric Capability

DISTRIBUTION STATEMENT A. Approved for public release.

Source: Slide 5 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

An April 11, 2016, press report states:

> The Pentagon wants to take a weapon originally designed for offense, flip its punch for defense and demonstrate by 2018 the potential for the Army and Navy to conduct missile defense of bases, ports and ships using traditional field guns to fire a new hypervelocity round guided by a mobile, ground variant of an Air Force fighter aircraft radar.
>
> The Strategic Capabilities Office is working with the Army, Navy and Air Force to craft a Hypervelocity Gun Weapon System that aims, in part, to provide China and Russia an example of a secret collection of new U.S. military capabilities the Defense Department is bringing online in an effort to strengthen conventional deterrence.
>
> "It is a fantastic program," Will Roper, Strategic Capabilities Office director, said in a March 28 interview with reporters, who said the project aims "to completely lower the cost of doing missile defense" by defeating missile raids at a lower cost per round and, as a consequence, imposing higher costs on attackers.[45]

A May 5, 2016, press report states:

---

[45] Jason Sherman, "SCO Aims To Flip The Script on Missile Defense With Hypervelocity Gun," *Inside the Navy*, April 11, 2016.

> Come January [2017], the Pentagon will almost assuredly have new leadership, complete with a new vision for how the Department of Defense should operate, organize and plan for the future.
>
> It’s a reality facing down Defense Secretary Ash Carter and Deputy Secretary Bob Work as they try to complete a transformation at the Pentagon, one which both men have said is vital to making sure the US is able to maintain its technological edge against great powers like Russia and China in the future.
>
> While Carter has made his reach-out to nontraditional communities the centerpiece of his initiative, Work’s focus has been the “Third Offset,” a series of technological bets, largely focused on man-unmanned teaming, that he believes will pay off in the future.
>
> So how do you make sure that a long-term project like the Third Offset keeps going under a new administration?
>
> “We have three really complimentary ways to go about this,” Work said Monday at an event hosted by the Atlantic Council.
>
> The first, he explained, is giving many options to the next administration.
>
> “First of all, I’m going to be central to the transition, so I am going to be able to personally talk with the transition team and explain to them what we have pursued and why we have pursued it, and let them make their own decisions,” Work said.
>
> “One of the things we have done in our program is build in a lot of different options that they can pull levers on,” Work explained.
>
> As an example, he pointed to the idea of an electromagnetic railgun. Initially, Work and his team thought that was an area that would be a major focus of development, but as they experimented they realized that a powder gun with a hypervelocity round could have almost the same impact — but at a fraction of the cost, because it did not require the development, testing and adaptation of a new gun.
>
> “We’re going to say ‘look, this is the place where [we think] you want to put your money,’ but we’re going to have enough money in both the electromagnetic railgun and the powder gun that if the new administration says ‘I really want the electromagnetic railgun, this is the way I want to go,’ knock yourself out,” Work said. “We’ve set you up for success.”[46]

Indirectly Improving Ability to Counter ASCMs and ASBMs

As discussed earlier, SSLs currently under development have enough beam power to counter small boats and UAVs, but not enough to ASCMs or ASBMs. Even so, such SSLs could indirectly improve a ship’s ability to counter ASCMs and ASBMs by permitting the ship to use fewer of its SAMs for countering UAVs, and more of them for countering ASCMs and ASBMs. Similarly, even though HVPs fired from 5-inch powder guns would not be able to counter ASBMs, they could indirectly improve a ship’s ability to counter ASBMs by permitting the ship to use fewer of its SAMs for countering ASCMs and more of its SAMs for countering ASBMs.

Remaining Development Challenges

Although the Navy in recent years has made considerable progress in developing SSLs, EMRG, and HVP, a number of significant development challenges remain. Overcoming these challenges will likely require years of additional development work, and ultimate success in overcoming them is not guaranteed.[47]

---

[46] Aaron Mehta, “Pentagon No. 2: How to Keep Third Offset Going in the Next Administration,” *Defense News*, May 5, 2016.

SSLs

As shown in **Figure 15**, remaining development challenges for SSLs include, among other things, making the system rugged enough for extended shipboard use, making the beam director (the telescope-like part of the laser that sends the beam toward the target) suitable for use in a marine environment (where moisture and salt in the air can be harsh on equipment), and integrating the system into the ship's electrical power system and combat system.

A January 23, 2015, blog post co-authored by the Office of Naval Research's program officer for the Navy's SSL program states:

> In the near term, many challenges remain to develop and operate high-energy laser systems in the maritime environment that are unique to the Navy and Marine Corps. Among these challenges is dealing with the heat generated as power levels increase. A second issue is packing sufficient power on the platform, which will require advanced battery, generator, power conditioning, and hybrid energy technologies. Current laser technologies are approximately 30 percent electrically efficient. Corrosion and contamination of optical windows by shipboard salt spray, dirt, and grime also are technical challenges. In addition, atmospheric turbulence resulting from shifting weather conditions, moisture, and dust is problematic. Turbulence can cause the air over long distances to act like a lens, resulting in the laser beam's diffusing and distorting, which degrades its performance.
>
> Much progress has been made in demonstrating high-energy laser weapon systems in the maritime environment, but there is still much to be done. Additional advances will be required to scale power levels to the hundreds of kilowatts that will make high[-]energy lasers systems robust, reliable, and affordable. Higher power levels are important for the ability to engage more challenging threats and improve the rate and range at which targets can be engaged.
>
> The programs managed by ONR are addressing these remaining issues while positioning this important warfighting capability toward an acquisition program and eventual deployment with the fleet and force.[48]

---

[47] Laser skeptics sometimes note that laser proponents over the years have made numerous predictions about when lasers might enter service with DOD, and that these predictions repeatedly have not come to pass. Viewing this record of unfulfilled predictions, skeptics might argue that "lasers are X years in the future—and always will be." Laser proponents acknowledge the record of past unfulfilled predictions, but argue that the situation has now changed because of rapid advancements in SSL technology and a shift from earlier ambitious goals (such as developing megawatt-power lasers for countering targets at tens or hundreds of miles) to more realistic goals (such as developing kilowatt-power lasers for countering targets at no more than a few miles). Laser proponents might argue that laser skeptics are vulnerable to what might be called cold plate syndrome (i.e., a cat that sits on a hot plate will not sit on a hot plate again—but it will not sit on a cold plate, either).

[48] Peter Morrison and Dennis Sorenson, "Developing a High-Energy Laser for the Navy," Future Force, January 23, 2015, accessed August 13, 2015, at http://futureforce.navylive.dodlive.mil/2015/01/high-energy-laser/. The authors are identified at the end of the post as follows: "Peter Morrison is the Office of Naval Research's program officer for the Navy's Solid-State Laser program. Dennis Sorenson is a contractor with the Office of Naval Research."

Figure 15. Development Challenges for SSLs

As of February 2013

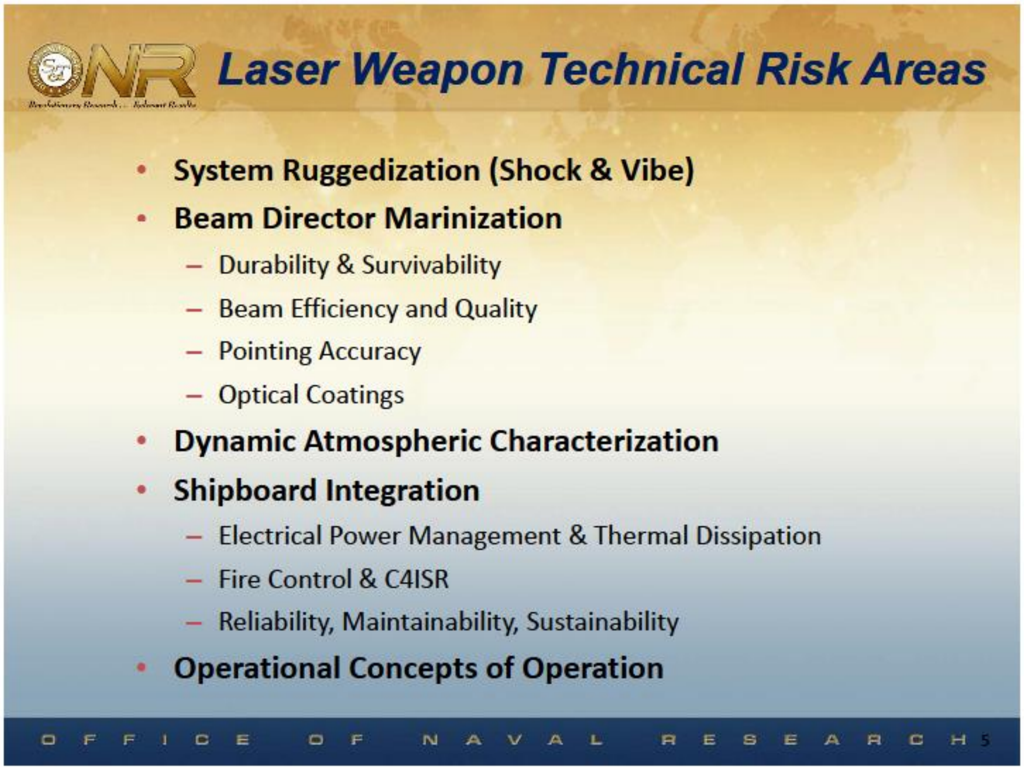

Laser Weapon Technical Risk Areas

- **System Ruggedization (Shock & Vibe)**
- **Beam Director Marinization**
  - Durability & Survivability
  - Beam Efficiency and Quality
  - Pointing Accuracy
  - Optical Coatings
- **Dynamic Atmospheric Characterization**
- **Shipboard Integration**
  - Electrical Power Management & Thermal Dissipation
  - Fire Control & C4ISR
  - Reliability, Maintainability, Sustainability
- **Operational Concepts of Operation**

OFFICE OF NAVAL RESEARCH

Source: Slide from Navy briefing entitled "Navy Solid State Laser Program Overview," ASNE Day 2013, Mr. Peter "Rollie" Morrison, ONR 35 S&T Program Office, February 22, 2013, accessed August 13, 2015, at https://www.navalengineers.org/ProceedingsDocs/ASNEDay2013/Morrison_Pres.pdf.

EMRG and HVP

As shown in **Figure 16**, remaining development challenges for EMRG involve items relating to the gun itself (including increasing barrel life to desired levels), the projectile, the weapon's electrical power system, and the weapon's integration with the ship. Fielding HVP on cruisers and destroyers ships equipped with 5-inch and 155mm powder guns would additionally require HVP to be integrated with the combat systems of those ships.

Figure 16. Development Challenges for EMRG

As of May 2014

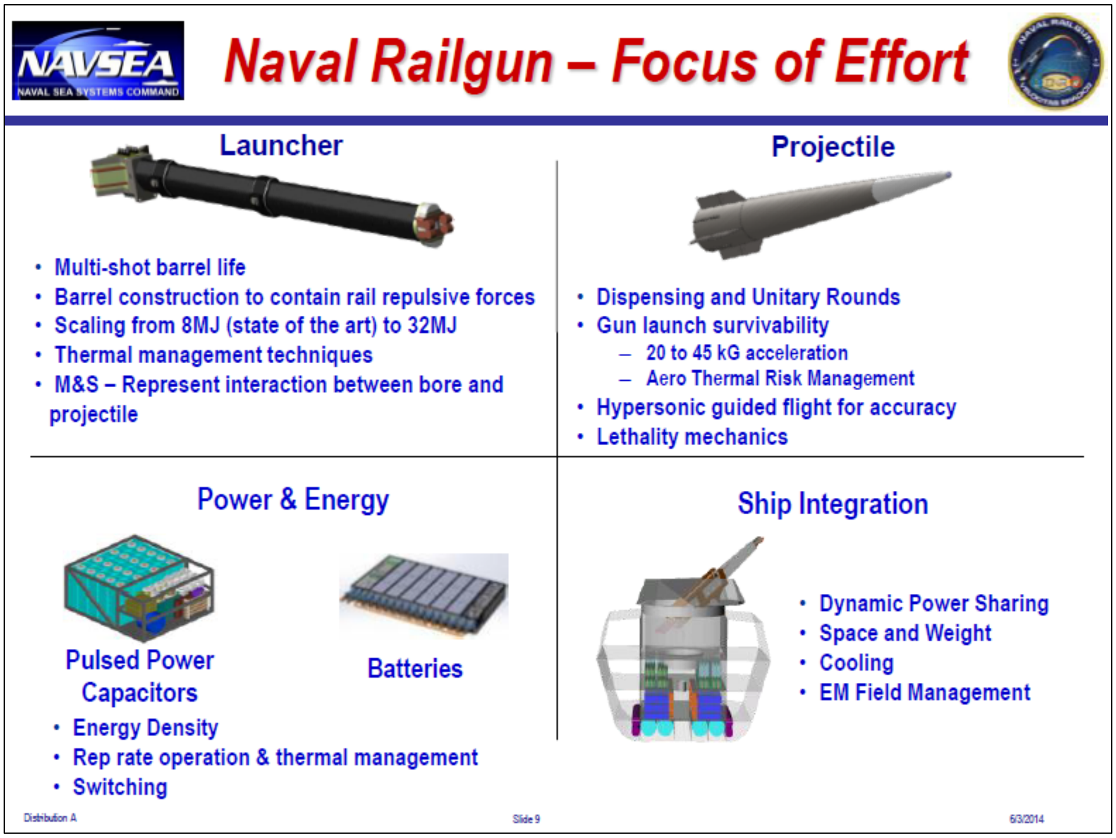

NAVSEA NAVAL SEA SYSTEMS COMMAND

Naval Railgun – Focus of Effort

Launcher

- Multi-shot barrel life
- Barrel construction to contain rail repulsive forces
- Scaling from 8MJ (state of the art) to 32MJ
- Thermal management techniques
- M&S – Represent interaction between bore and projectile

Projectile

- Dispensing and Unitary Rounds
- Gun launch survivability
  - 20 to 45 kG acceleration
  - Aero Thermal Risk Management
- Hypersonic guided flight for accuracy
- Lethality mechanics

Power & Energy

Pulsed Power Capacitors

Batteries

- Energy Density
- Rep rate operation & thermal management
- Switching

Ship Integration

- Dynamic Power Sharing
- Space and Weight
- Cooling
- EM Field Management

Distribution A

Slide 9

6/3/2014

Source: Slide 9 from Navy briefing entitled "Electromagnetic Railgun," NDIA Joint Armaments Forum, Exhibition & Technology Demonstration, May 14, 2014, LCDR Jason Fox, USN, Assistant PM [Program Manager], Railgun Ship Integration, Distribution A, Approved for Public Release, accessed August 13, 2015, at http://www.dtic.mil/ndia/2014armaments/WedFox.pdf.

The Navy states:

> The EMRG effort began in FY 2005 with a focus on the barrel, power storage, and rail technology. In 2015, the Navy is testing full-scale industry advanced composite launchers for structure strength and manufacturability, and has advanced the pulsed-power system design from single-shot to actively cooled repeated rate operations. Building on the success of the first phase, the second phase started in 2012 with a focus on developing equipment and techniques to fire ten rounds per minute. Thermal-management techniques required for sustained firing rates are in development for both the launcher system and the pulsed-power system. The Office of Naval Research will develop a tactical prototype EMRG launcher and pulsed-power architecture suitable for advanced testing both afloat and ashore. Railgun demonstration has been funded to occur in FY 2016.[49]

A June 2015 press report states:

> As the Navy prepares to test its electromagnetic railgun at sea for the first time in 2016, service leaders said one of the biggest challenges will be integrating the new technology onto existing platforms.....

---

[49] U.S. Navy, *U.S. Navy Program Guide 2015*, p. 169.

> [Vice Adm. William Hilarides, commander of Naval Sea Systems Command] said he is positive the Navy will successfully demonstrate the weapon's ability to fire from the Trenton, but one of the biggest challenges will be configuring the railgun so that it fits within the power structure of other existing platforms.
>
> "Those are not 600-ton margin ships," he said [meaning ships with 600 tons of growth margin available to accommodate EMRG]. "If they have 60 tons, if they have 16 tons, then we'll be talking about what do we take off our existing destroyers, cruisers and other ships in order to get this incredible capability [on them]."
>
> These types of discussions are influencing ship designs as program managers look at what systems are indispensable and what can be exchanged, Hilarides said.
>
> Integrating the railgun into the fleet won't be a swift process.
>
> It will be at least 10 years until the railgun is fielded on new ships and potentially 30 years past that before the Navy considers removing powder guns from the fleet entirely and transitioning to energy weapons alone, according to Hilarides.[50]

Issues for Congress

Potential Oversight Questions

Potential oversight questions for Congress regarding Navy programs for SSLs, EMRG, and HVP include the following:

- Using currently available approaches for countering ASCMs and ASBMs, how well could Navy surface ships defend themselves in a combat scenario against an adversary such as China that has large numbers of ASCMs (including advanced models) and ASBMs? How would this change if Navy surface ships in coming years were equipped with SSLs, EMRG, HVP, or some combination of these systems?
- How significant are the remaining development challenges for SSLs, EMRG, and HVP?
- Are current schedules for developing SSLs, EMRG, and HVP appropriate in relation to remaining development challenges and projected improvements in enemy ASCMs and ASBMs? To what degree are current schedules for developing SSLs, EMRG, or HVP sensitive to annual funding levels?
- When does the Navy anticipate issuing roadmaps detailing its plans for procuring and installing production versions of SSLs, EMRGs, and HVP on specific Navy ships by specific dates?
- Will the kinds of surface ships that the Navy plans to procure in coming years have sufficient space, weight, electrical power, and cooling capability to take full advantage of SSLs (particularly those with beam powers above 200 kW) and EMRG? What changes, if any, would need to be made in Navy plans for procuring large surface combatants (i.e., destroyers and cruisers) or other Navy ships to take full advantage of SSLs and EMRG?

---

[50] Allyson Versprille, "Integration Biggest Challenge for Railgun," *National Defense*, June 2015. See also Lance M. Bacon, "3-Star: 'Lot of Work' Before Railgun Arrives in Fleet," *Navy Times*, February 5, 2015.

- Are the funding sources for SSLs, EMRG, and HVP in Navy and Defense-Wide research and development accounts (see "Summary of Congressional Action on FY2017 Funding" below) sufficiently visible for supporting congressional oversight?

Legislative Activity for FY2017

Summary of Congressional Action on FY2017 Funding

Funding in the defense budget for research and development work on Navy SSLs, EMRG, and HVP is spread across several research and development account line items (which are known as program elements, or PEs). The PEs shown in **Table 1** capture much but not necessarily all of the funding for developing Navy SSLs, EMRG, and HVP. The PEs shown in the table, moreover, include funding for efforts other than Navy SSLs, EMRG, and HVP, so congressional changes from requested amounts might or might not relate to SSLs, EMRG, or HVP.

Table 1. Summary of Congressional Action on FY17 Funding

In millions of dollars, rounded to nearest tenth

Program Element (PE) number, PE name, FY16 budget line number	Req.	Authorization			Appropriation		
		HASC	SASC	Conf.	HAC	SAC	Conf.
0602114N, Power Projection Applied Research, line 4	41.4	41.4	41.4		41.4		
0602750N, Future Naval Capabilities Applied Research, line 13	165.1	165.1	165.1		157.1		
0603114N, Power Projection Advanced Technology, line 16	96.4	106.4	81.4		76.6		
0603673N, Future Naval Capabilities Advanced Technology Development, line 21	249.1	249.1	239.1		252.1		
0603925N, Directed Energy and Electric Weapon System, line 75	32.7	32.7	32.7		32.7		
0604250D8Z, Advanced Innovative Technology, line 95	844.9	804.9	844.9		844.9		

Source: Table prepared by CRS based on Navy FY17 budget submission and committee and conference reports.

Notes: HASC is House Armed Services Committee; **SASC** is Senate Armed Services Committee; **HAC** is House Appropriations Committee; **SAC** is Senate Appropriations Committee; **Conf.** is conference agreement. The PEs shown in the table below capture much but not necessarily all of the funding for work on Navy SSLs, EMRG, and HVP. The PEs shown in the table, moreover, include funding for efforts other than Navy SSLs, EMRG, and HVP.

FY2017 National Defense Authorization Act (H.R. 4909/S. 2943)

House

The House Armed Services Committee, in its report (H.Rept. 114-537 of May 4, 2016) on H.R. 4909, recommended the funding levels shown in the HASC column of **Table 1**. The recommended increase of $10 million for PE 0603114N, Power Projection Advanced Technology (line 16) is for "Program increase for common mount." (Page 498) The recommended reduction of $40 million for PE 0604250D8Z, Advanced Innovative Technology (line 95) is for "SCO" (the Strategic Capabilities Office). (Page 518) H.Rept. 114-537 states:

> *Common mount for electromagnetic railgun*
>
> The budget request contained $96.4 million in PE 63114N for power projection advanced technology. Of this amount, $15.4 million was included for the Navy's electromagnetic railgun prototype.
>
> The committee remains supportive of the Navy's program for developing and deploying an electromagnetic railgun. The committee recognizes the growing imperative for the Navy to field this type of weapon, not only to increase capabilities for naval surface fire support and ballistic missile defense, but to also decrease the cost exchange model when comparing the railgun to conventional missiles or guns. However, the committee is increasingly concerned that the shift in emphasis to the hypervelocity projectile by the Strategic Capabilities Office has left the Navy with a funding gap in developing the requirements and design for a common mount, which is a necessary prerequisite to getting this capability into operational use. Therefore, the committee directs the Secretary of the Navy to provide a briefing to the House Committee on Armed Services by February 15, 2017, on the plan and milestone schedule for demonstrating and deploying a common railgun mount for sea- and land-based applications.
>
> The committee recommends $106.4 million, an increase of $10.0 million, in PE 63114N to support the development of a common mount for the sea-based and land-based electromagnetic railgun. (Page 61)

H.Rept. 114-537 also states:

> *Strategic Capabilities Office*
>
> The budget request contained $844.9 million in PE 64250D8Z for development activities of the Strategic Capabilities Office (SCO).
>
> Created in 2012 by the Deputy Secretary of Defense, SCO has the mission to identify, analyze, demonstrate, and transition game-changing applications of existing and near-term technology to shape and counter emerging threats. SCO is comprised of a relatively small number of personnel and relies on other program office personnel and resources to execute its mission. The committee appreciates the nature of SCO's mission and sustained leanness of the organization; however, the committee notes the budget for SCO has grown exponentially each fiscal year. For example, the fiscal year 2017 budget request is nearly double the request for fiscal year 2016.
>
> The committee is concerned that such rapid budget growth may bring with it some risks, including the demands on SCO's small staff, demands on other Department of Defense personnel, and impact of SCO decisions on existing programs. For example, the committee is aware of SCO's inclusion on the electromagnetic railgun development, and subsequent reprioritizing of its planned investment in that program for fiscal year 2017, resulting in a funding gap that could not be covered by the program office.
>
> Additionally, the committee remains concerned that the transition of technologies from SCO has not been adequately captured and conveyed to the oversight committees. The report required by the committee report (H. Rept. 114–102) accompanying the National Defense Authorization Act for Fiscal Year 2016 has not been delivered and is now almost 6 months late. In order to support prudent use of taxpayer resources, and to ensure proper oversight of these activities, the committee believes this report should be provided and concerns addressed before supporting full funding of planned activities.
>
> Therefore, the committee recommends $804.9 million, a decrease of $40.0 million, in PE 64250D8Z for development activities of the Strategic Capabilities Office. (Page 92)

Section 220 of H.R. 4909 as reported states:

> SEC. 220. Designation of Department of Defense senior official with principal responsibility for directed energy weapons.
>
> Not later than 180 days after the date of the enactment of this Act, the Secretary of Defense shall—
>
> (1) designate a senior official already serving within the Department of Defense as the official with principal responsibility for the development and demonstration of directed energy weapons for the Department; and
>
> (2) set forth the responsibilities of that senior official with respect to such programs.

H.Rept. 114-537 also states:

> *Five-inch precision guided projectile development for naval surface fire support*
>
> In the committee report (H. Rept. 114–102) accompanying the National Defense Authorization Act for Fiscal Year 2016, the committee noted "that current surface Navy gunnery requirements are outdated and that new technologies such as railgun and directed energy weapons are nearing readiness for technology transition." The committee referenced the Advanced Naval Surface Fires (ANSF) initiative and noted the ANSF was assessing options for providing a near-term 5-inch guided munition capability. The committee understands this capability would provide for improved and extended-range naval surface fire support. The committee continues to support the need for this precision guided capability and is also aware of the Hypervelocity Gun Weapon System (HGWS) program that is currently under consideration by the Strategic Capabilities Office (SCO). The committee notes the HGWS program would "flip the cost equation using conventional guns to defend forward bases against raids of advanced cruise and ballistic missiles" and believes there could be applications for use in 5-inch gun systems for naval surface fires support. The committee is encouraged by the development of both of these initiatives and expects the Navy and SCO to coordinate on these capabilities. The committee also expects the Navy to proceed forward with an accelerated development and acquisition strategy for this needed capability that is consistent with acquisition reform principles. (Pages 62-63)

H.Rept. 114-537 also states:

> *Technology enablers for directed energy weapon systems*
>
> The committee is aware that the Department of Defense has made significant advances in the development and operational demonstration of directed energy weapons systems. Each military department has demonstrated a marquee program in this area, such as the Navy's Laser Weapon System deployed on the USS Ponce, the Army High Energy Laser Mobile Demonstrator, and the Marine Corps' Ground Based Air Defense System. Along with technology demonstration activities like the Robust Electric Laser Initiative and the High Energy Liquid Laser Area Defense System, each of these programs demonstrated the increased power output and power on target necessary to develop a militarily useful directed energy weapon.
>
> However, as the Department has made progress in raising the power levels of these systems, it has also demonstrated the need for emphasis on development in other technology areas necessary to realize the full potential of laser weapons. For example, higher power output requires improved beam control to engage targets at greater distances, as well as better thermal management systems to dissipate the increased heat load. As the Department has been overcoming foundational technical challenges, new challenges have emerged that will impact the operational uses for directed energy weapons.

> Therefore, the committee directs the Assistant Secretary of Defense for Research and Engineering, in coordination with the research components of the military departments and the High Energy Laser Joint Technology Office, to provide a briefing to the House Committee on Armed Services by January 20, 2017. This briefing should provide a roadmap for enabling technologies, including:
>
> (1) Beam directors and adaptive optics, including deformable mirrors;
>
> (2) Thermal management needs and capabilities;
>
> (3) Integration challenges with fire control systems, including potential
>
> future needs for fire control for laser systems;
>
> (4) Power architectures and power electronics needs;
>
> (5) Facilities and test range capabilities; and
>
> (6) Other areas as deemed by the Secretary. (Page 93)

Senate

The Senate Armed Services Committee, in its report (S.Rept. 114-255 of May 18, 2016) on S. 2943, recommended the funding levels shown in the SASC column of **Table 1**. The recommended reduction of $15 million to PE 0603114N, Power Projection Advanced Technology (line 16) is for "General decrease." (Page 481) The recommended reduction of $10 million to PE 0603673N, Future Naval Capabilities Advanced Technology Development (line 21) is for "Capable manpower, and power and energy." (Page 481).[51] S.Rept. 114-255 states:

> **Power projection advanced technology**
>
> The budget request included $96.4 million in PE 63114N for power projection advanced technology. The committee notes that the Navy, Air Force, Defense Advanced Research Projects Agency, Strategic Capabilities Office, and other elements within the Department of Defense are all pursuing advanced power projection technologies and systems. The committee is concerned that these efforts are not well-coordinated and have uncertain pathways for transition to programs of record. In addition, the committee notes that the budget request represents an almost 200 percent increase over the amount enacted for fiscal year 2016. The committee believes that such a large increase in budget is not warranted and is concerned about the ability of the programs to absorb the additional funding. Consequently, the committee recommends a decrease of $15.0 million in PE 63114N, but directs that this reduction not be assessed against solid state laser maturation efforts. (Page 52)

Section 216 of S. 2943 as reported states:

---

[51] This recommended reduction of $10 million does not appear to relate (at least not directly) to lasers, the electromagnetic railgun, or the HVP. S.Rept. 114-255 states:

> **Capable manpower and power and energy**
>
> The budget request included $249.1 million in PE 63673N for future naval capabilities advanced technology developments. The activities listed under this program element include capable manpower and power and energy. The committee believes that the work plans for fiscal year 2017 on these activities does not warrant the level of funding included in the budget request. For example, the committee notes that the research included in these two projects include development of new personnel and management methodologies, and capabilities in energy security. Both of these efforts could be better coordinated with other organizations performing similar research. Consequently, the committee recommends a decrease of $10.0 million in PE 63673N to be distributed appropriately from capable manpower and power and energy. (Page 53)

SEC. 216. Directed energy weapon system programs.

(a) Inclusion of Directed Energy Weapon System programs in the rapid acquisition authority program.—

(1) IN GENERAL.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:

“(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.

“(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:

“(I) The rapid acquisition authority provided under this section.

“(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.

“(III) The acquisition of commercial items using simplified acquisition procedures.

“(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.

“(iii) In this subparagraph, the term ‘directed energy weapon systems’ means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.”.

(2) CONFORMING AMENDMENTS.—Section 2373 of title 10, United States Code, is amended—

(A) in subsection (a), by striking “and aeronautical supplies” and inserting “, aeronautical supplies, and directed energy weapon systems”; and

(B) by adding at the end of the following new subsection:

“(c) Directed energy weapon systems defined.—In this section, the term ‘directed energy weapon systems’ means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.”.

(b) Joint Directed Energy Program Office.—

(1) REDESIGNATION.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the “Joint Directed Energy Program Office” (in this subsection referred to as the “Office”).

(2) STRATEGIC PLAN FOR DEVELOPMENT AND FIELDING OF DIRECTED ENERGY WEAPONS CAPABILITIES.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and fielding of directed energy weapons capabilities for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.

(3) ACCELERATION OF DEVELOPMENT AND FIELDING OF DIRECTED ENERGY WEAPONS CAPABILITIES.—

(A) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this section and new acquisition and management

> practices established pursuant to this section to accelerate the development and fielding of directed energy capabilities.
>
> (B) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in subparagraph (A) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Regarding Section 216, S.Rept. 114-255 states:

> **Directed energy weapon system programs (sec. 216)**
>
> The committee remains concerned about the Department of Defense's inability to field an operational directed energy system. The committee is aware that the military services and industry partners have developed sufficient directed energy weapon capabilities for specific scenarios—like the High Energy Laser Mobile Demonstrator (HEL–MD) to counter rocket, artillery and mortar for base protection purposes and the Counter Electronics High Powered Microwave Advanced Missile Project (CHAMP) for disabling an adversary's electronics while avoiding collateral damage. These programs, as well as other high energy laser weapon systems, have been tested and demonstrated, but have failed to transition to acquisition programs of record.
>
> The committee notes that directed energy capabilities have the potential to support many operational missions in cost effective and efficient manners. In response to these factors, the committee recommends a provision that would amend section 806 of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314) to grant rapid acquisition authorities for directed energy weapon systems to accelerate the development and fielding of this technology and to help offset the gains of potential adversaries.
>
> The committee notes that since 1960, the Department of Defense has invested more than $6.0 billion in directed energy science and technology initiatives. However, the committee remains concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs. The committee notes with concern that years of investment have not to date resulted in any operational systems with high energy laser capability.
>
> The committee highlights that the Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications found that "directed energy offers promise as a transformational 'game changer' in military operations, able to augment and improve operational capabilities in many areas." The task force further concluded that the range of potential applications is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons. Consistent with the findings of the task force, the committee believes that directed energy weapons systems offer significant benefits in terms of cost effectiveness, sustainability, magazine capabilities, and precision targeting. (Pages 46-47)

S.Rept. 114-255 also states:

> **High energy laser joint technology office**
>
> The budget request included $42.3 million in PE 62890F for high energy laser research. The committee notes that this program element funds defense high energy laser applied research through the High Energy Laser Joint Technology Office. However, the committee is concerned that the Joint Technology Office has not received sufficient funding in recent years to drive the maturation of high energy laser technology. As an example, the committee notes with concern that no laser technologies have yet been fielded or deployed, despite promising development and field tests. Given the importance of directed energy weapons systems in general as noted elsewhere in this Act, and of high energy laser systems in particular, the committee is concerned that budget request for this

> program element will be insufficient for supporting the joint technology office. Accordingly, the committee recommends an increase of $5.0 million in PE 62890F for the high energy laser joint technology office. (Pages 54-55)

S.Rept. 114-255 also states:

> **Directed energy systems prototyping**
>
> The budget request included no money in PE 64342D8Z for defense technology offsets. The committee notes with disappointment that the administration did not view it as a priority to request funds through this program element. Particularly with the high-profile emphasis placed on the Department of Defense's Third Offset Strategy, the committee is disappointed to see this program be unfunded. In addition, as noted elsewhere in this report, the committee is deeply disappointed with how the technology offset funding enacted in fiscal year 2016 was allocated. As noted, none of the money was put towards directed energy, in contradiction to the clear intent of Congress that half of the money be used to bolster directed energy technologies. While the committee does not recommend additional unrestricted funds for the technology offsets program, the committee underscores that directed energy systems are still critical areas of work in need of greater support and attention. The committee believes that the Department needs to focus in particular on the transition from lab development to deployment and fielding. Consequently, the committee recommends a general increase of $25.0 million in PE 64342D8Z to be used only for the purposes of directed energy systems prototyping. (Pages 59-60)

S.Rept. 114-255 also states:

> **Laser weapon system demonstrator**
>
> The Committee commends the Navy for initiating and funding the Laser Weapon Systems Demonstrator (LWSD) and believes that this is an important step toward maturing technologies that could ultimately enable the deployment of a shipboard maritime laser weapons system. While the Committee understands that the Navy envisions transitioning laser weapons to a formal Program of Record in the 2020s, it appears that the Navy has not programmed funding beyond the LWSD sea-based tests to support the installation of LWSD on a DDG or for the design and procurement of a formal maritime laser program.
>
> The committee expects that the Secretary of the Navy will keep the congressional defense committees updated on its plan to seamlessly transition the LWSD to a shipboard weapons system following sea-based testing and to a formal maritime laser Program of Record, technical progress toward developing the capability, and programmatic steps being taken to move to demonstration and deployment of advanced laser systems. (Page 70)

FY2017 DOD Appropriations Act (H.R. 5293)

House

The House Appropriations Committee, in its report (H.Rept. 114-577 of May 19, 2016) on H.R. 5293, recommended the funding levels shown in the HAC column of **Table 1**. The recommended reduction of $8 million to PE 0602750N, Future Naval Capabilities Applied Research (line 13) is for "FORCENET excessive growth" ($5 million), "Power and energy previously funded efforts" ($2 million), and "Sea shield previously funded efforts" ($1 million). (Page 228) The recommended reduction of $19.8 million to PE 0603114N, Power Projection Advanced Technology (line 16) is for "Precision strike technology excess growth." (Page 228) The recommended net reduction of $3 million to PE 0603673N, Future Naval Capabilities Advanced

Technology Development (line 21) consists of a reduction of $2 million for “Power and energy previously funded efforts,” a reduction of $2 million for “Sea shield previously funded efforts,” and an increase of $7 million for “Program increase—ASW [antisubmarine warfare] research.” (Page 228)

H.Rept. 114-577 states:

> HIGH ENERGY LASERS
>
> The Committee is aware of efforts within the High Energy Laser Joint Technology Office to develop advanced, directed-energy, high energy laser weapons that have the potential to perform a wide variety of military missions. The Committee encourages the Secretary of Defense to explore further development and evaluation of this important technology. (Page 264)

Directed Energy Weapon Systems Acquisition Act of 2016 (H.R. 4964/S. 2778)

House

H.R. 4964 was introduced in the House on May 5, 2016. The text of H.R. 4964 as introduced states:

> SECTION 1. Short title.
>
> This Act may be cited as the “Directed Energy Weapon Systems Acquisition Act of 2016”.
>
> SEC. 2. Findings.
>
> Congress makes the following findings:
>
> (1) The Committee on Armed Services of the Senate noted in the report accompanying S. 1356 (S. Rept. 114–49; 114th Congress) that since 1960, the Department of Defense has invested more than $6,000,000,000 in directed energy science and technology initiatives, and that the Committee is concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs.
>
> (2) The Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications (the “Task Force”) found that “directed energy offers promise as a transformational ‘game changer’ in military operations, able to augment and improve operational capabilities in many areas”.
>
> (3) Despite this potential, years of investment have not resulted in any operational systems with high energy laser capability.
>
> (4) The Task Force believes that the range of potential application is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons.
>
> SEC. 3. Inclusion of directed energy weapon system programs in the rapid acquisition authority program.
>
> (a) In general.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:
>
> “(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate

> a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.
>
> “(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:
>
> “(I) The rapid acquisition authority provided under this section.
>
> “(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.
>
> “(III) The acquisition of commercial items using simplified acquisition procedures.
>
> “(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.
>
> “(iii) In this subparagraph, the term ‘directed energy weapon system’ means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.”.
>
> (b) Conforming amendments.—Section 2373 of title 10, United States Code, is amended—
>
> (1) in subsection (a), by striking “and aeronautical supplies” and inserting “, aeronautical supplies, and directed energy weapon systems”; and
>
> (2) by adding at the end of the following new subsection:
>
> “(c) Directed energy weapon system defined.—In this section, the term ‘directed energy weapon system’ means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.”.
>
> SEC. 4. Joint Directed Energy Program Office.
>
> (a) Redesignation.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the “Joint Directed Energy Program Office” (in this section referred to as the “Office”).
>
> (b) Strategic plan for development and transition of directed energy weapons capabilities toward fielding.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and transition of directed energy weapons capabilities toward fielding for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.
>
> (c) Acceleration of development and transition of directed energy weapons capabilities toward fielding.—
>
> (1) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this Act and new acquisition and management practices established pursuant to this Act to accelerate the development and transition of directed energy capabilities toward fielding.
>
> (2) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in paragraph (1) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Senate

S. 2778 was introduced in the Senate on April 12, 2016. The text of S. 2778 as introduced states:

SECTION 1. Short title.

This Act may be cited as the "Directed Energy Weapon Systems Acquisition Act of 2016".

SEC. 2. Findings.

Congress makes the following findings:

(1) The Committee on Armed Services of the Senate noted in the report accompanying S. 1356 (S. Rept. 114–49; 114th Congress) that since 1960, the Department of Defense has invested more than $6,000,000,000 in directed energy science and technology initiatives, and that the Committee is concerned that, despite this significant investment, the Department's directed energy initiatives are not resourced at levels necessary to transition them to full-scale acquisition programs.

(2) The Defense Science Board Task Force on Directed Energy Weapon Systems and Technology Applications (the "Task Force") found that "directed energy offers promise as a transformational 'game changer' in military operations, able to augment and improve operational capabilities in many areas".

(3) Despite this potential, years of investment have not resulted in any operational systems with high energy laser capability.

(4) The Task Force believes that the range of potential application is sufficient to warrant significantly increased attention to the scope and direction of efforts to assess, develop, and field appropriate laser, microwave, and millimeter wave weapons.

SEC. 3. Inclusion of directed energy weapon system programs in the rapid acquisition authority program.

(a) In general.—Section 806(c)(1) of the Bob Stump National Defense Authorization Act for Fiscal Year 2003 (Public Law 107–314; 10 U.S.C. 2302 note) is amended by adding at the end the following new subparagraph:

"(D) (i) In the case of any supplies and associated support services that, as determined in writing by the Secretary of Defense without delegation, are urgently needed to eliminate a deficiency in directed energy weapon systems, the Secretary may use the procedures developed under this section in order to accomplish the rapid acquisition and deployment of needed offensive or defensive directed energy weapon systems capabilities, supplies, and associated support services.

"(ii) For the purposes of directed energy weapon systems acquisition, the Secretary of Defense shall consider use of the following procedures:

"(I) The rapid acquisition authority provided under this section.

"(II) Use of other transactions authority provided under section 2371 of title 10, United States Code.

"(III) The acquisition of commercial items using simplified acquisition procedures.

"(IV) The authority for procurement for experimental purposes provided under section 2373 of title 10, United States Code.

"(iii) In this subparagraph, the term 'directed energy weapon system' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".

(b) Conforming amendments.—Section 2373 of title 10, United States Code, is amended—

(1) in subsection (a), by striking "and aeronautical supplies" and inserting ", aeronautical supplies, and directed energy weapon systems"; and

(2) by adding at the end the following new subsection:

"(c) Directed energy weapon system defined.—In this section, the term 'directed energy weapon system' means military action involving the use of directed energy to incapacitate, damage, or destroy enemy equipment, facilities, or personnel.".

SEC. 4. Joint Directed Energy Program Office.

(a) Redesignation.—The High Energy Laser Joint Technology Office of the Department of Defense is hereby redesignated as the "Joint Directed Energy Program Office" (in this section referred to as the "Office").

(b) Strategic plan for development and transition of directed energy weapons capabilities toward fielding.—In addition to the functions and duties of the Office in effect on the day before the date of the enactment of this Act, the Office shall develop a strategic plan for development and transition of directed energy weapons capabilities toward fielding for the Department, in which the Office may define requirements for directed energy capabilities that address the highest priority warfighting capability gaps of the Department.

(c) Acceleration of development and transition of directed energy weapons capabilities toward fielding.—

(1) IN GENERAL.—To the degree practicable, the Office shall use the policies of the Department that are revised pursuant to this Act and new acquisition and management practices established pursuant to this Act to accelerate the development and transition of directed energy capabilities toward fielding.

(2) ENGAGEMENT.—The Secretary shall ensure that use of policies and practices described in paragraph (1) include engagement with defense and private industries, research universities, and unaffiliated, nonprofit research institutions.

Appendix. Potential Advantages and Limitations of Shipboard Lasers

This appendix presents additional information on potential advantages and limitations of shipboard lasers.

Potential Advantages

In addition to a low marginal cost per shot and deep magazine, potential advantages of shipboard lasers include the following:

- **Fast engagement times.** Light from a laser beam can reach a target almost instantly (eliminating the need to calculate an intercept course, as there is with interceptor missiles) and, by remaining focused on a particular spot on the target, cause disabling damage to the target within seconds. After disabling one target, a laser can be redirected in several seconds to another target.
- **Ability to counter radically maneuvering missiles.** Lasers can follow and maintain their beam on radically maneuvering missiles that might stress the maneuvering capabilities of Navy SAMs.
- **Precision engagements.** Lasers are precision-engagement weapons—the light spot from a laser, which might be several inches in diameter, affects what it hits, while generally not affecting (at least not directly) separate nearby objects.
- **Graduated responses.** Lasers can perform functions other than destroying targets, including detecting and monitoring targets and producing nonlethal effects, including reversible jamming of electro-optic (EO) sensors. Lasers offer the potential for graduated responses that range from warning targets to reversibly jamming their systems, to causing limited but not disabling damage (as a further warning), and then finally causing disabling damage.

Potential Limitations

Potential limitations of shipboard lasers include the following:

- **Line of sight.** Since laser light tends to fly through the atmosphere on an essentially straight path, shipboard lasers would be limited to line-of-sight engagements, and consequently could not counter over-the-horizon targets or targets that are obscured by intervening objects. This limits in particular potential engagement ranges against small boats, which can be obscured by higher waves, or low-flying targets. Even so, lasers can rapidly reacquire boats obscured by periodic swells.
- **Atmospheric absorption, scattering, and turbulence.** Substances in the atmosphere—particularly water vapor, but also things such as sand, dust, salt particles, smoke, and other air pollution—absorb and scatter light from a shipboard laser, and atmospheric turbulence can defocus a laser beam. These effects can reduce the effective range of a laser. Absorption by water vapor is a particular consideration for shipboard lasers because marine environments feature substantial amounts of water vapor in the air. There are certain wavelengths of light (i.e., “sweet spots” in the electromagnetic spectrum) where atmospheric absorption by water vapor is markedly reduced. Lasers can be

designed to emit light at or near those sweet spots, so as to maximize their potential effectiveness. Absorption generally grows with distance to target, making it in general less of a potential problem for short-range operations than for longer-range operations. Adaptive optics, which make rapid, fine adjustments to a laser beam on a continuous basis in response to observed turbulence, can counteract the effects of atmospheric turbulence. Even so, lasers might not work well, or at all, in rain or fog, preventing lasers from being an all-weather solution.

- **Thermal blooming.** A laser that continues firing in the same exact direction for a certain amount of time can heat up the air it is passing through, which in turn can defocus the laser beam, reducing its ability to disable the intended target. This effect, called thermal blooming, can make lasers less effective for countering targets that are coming straight at the ship, on a constant bearing (i.e., "down-the-throat" shots). Other ship self-defense systems, such as interceptor missiles or a CIWS, might be more suitable for countering such targets. Most tests of laser systems have been against crossing targets rather than "down-the-throat" shots. In general, thermal blooming becomes more of a concern as the power of the laser beam increases.
- **Saturation attacks.** Since a laser can attack only one target at a time, requires several seconds to disable it, and several more seconds to be redirected to the next target, a laser can disable only so many targets within a given period of time. This places an upper limit on the ability of an individual laser to deal with saturation attacks—attacks by multiple weapons that approach the ship simultaneously or within a few seconds of one another. This limitation can be mitigated by installing more than one laser on the ship, similar to how the Navy installs multiple CIWS systems on certain ships.
- **Hardened targets and countermeasures.** Less-powerful lasers—that is, lasers with beam powers measured in kilowatts (kW) rather than megawatts (MW)—can have less effectiveness against targets that incorporate shielding, ablative material, or highly reflective surfaces, or that rotate rapidly (so that the laser spot does not remain continuously on a single location on the target's surface) or tumble. Small boats (or other units) could employ smoke or other obscurants to reduce their susceptibility to laser attack.[52] Measures such as these, however, can increase the cost and/or weight of a weapon, and obscurants could make it more difficult for small boat operators to see what is around them, reducing their ability to use their boats effectively.
- **Risk of collateral damage to aircraft, satellites, and human eyesight.** Since light from an upward-pointing laser that does not hit the target would continue flying upward in a straight line, it could pose a risk of causing unwanted collateral damage to aircraft and satellites. The light emitted by SSLs being developed by the Navy is of a frequency that can cause permanent damage to human eyesight, including blinding. Blinding can occur at ranges much greater than ranges for damaging targeted objects. Scattering of laser light off the target or off fog or particulates in the air can pose a risk to exposed eyes.[53]

---

[52] See, for example, "Kelsey D. Atherton, "China Plans To Defeat American Lasers With Smoke," *Popular Science*, May 3, 2016.

[53] The United States in 1995 ratified the 1980 Convention on Prohibitions or Restriction on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects. An (continued...)

For additional background information on potential Navy shipboard SSLs, see CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

Author Contact Information

Ronald O'Rourke
Specialist in Naval Affairs
rorourke@crs.loc.gov, 7-7610

(...continued)

international review of the convention began in 1994 and concluded in May 1996 with the adoption of, among other things, a new Protocol IV on blinding laser weapons. The protocol prohibits the employment of lasers that are specifically designed to cause permanent blindness to the naked eye or to the eye with corrective eyesight devices. The United States ratified Protocol IV on December 23, 2008, and it entered into force for the United States on July 21, 2009. DOD views the protocol as fully consistent with DOD policy. DOD believes the lasers discussed in this report are consistent with DOD policy of prohibiting the use of lasers specifically designed to cause permanent blindness to the naked eye or to the eye with corrective eyesight devices. For further discussion, see Appendix I ("Protocol on Blinding Lasers") in CRS Report R41526, *Navy Shipboard Lasers for Surface, Air, and Missile Defense: Background and Issues for Congress*, by Ronald O'Rourke.

Torpedo & Directed Energy Weapons Indian Navy Perspective

RAdm Om Prakash Singh Rana, VSM,
Director General of Naval Armament Inspection

IN supporting Indigenisation

IN Indigenisation Plan 2015-30

Maritime Cap Perspective Plan

Self-Reliance

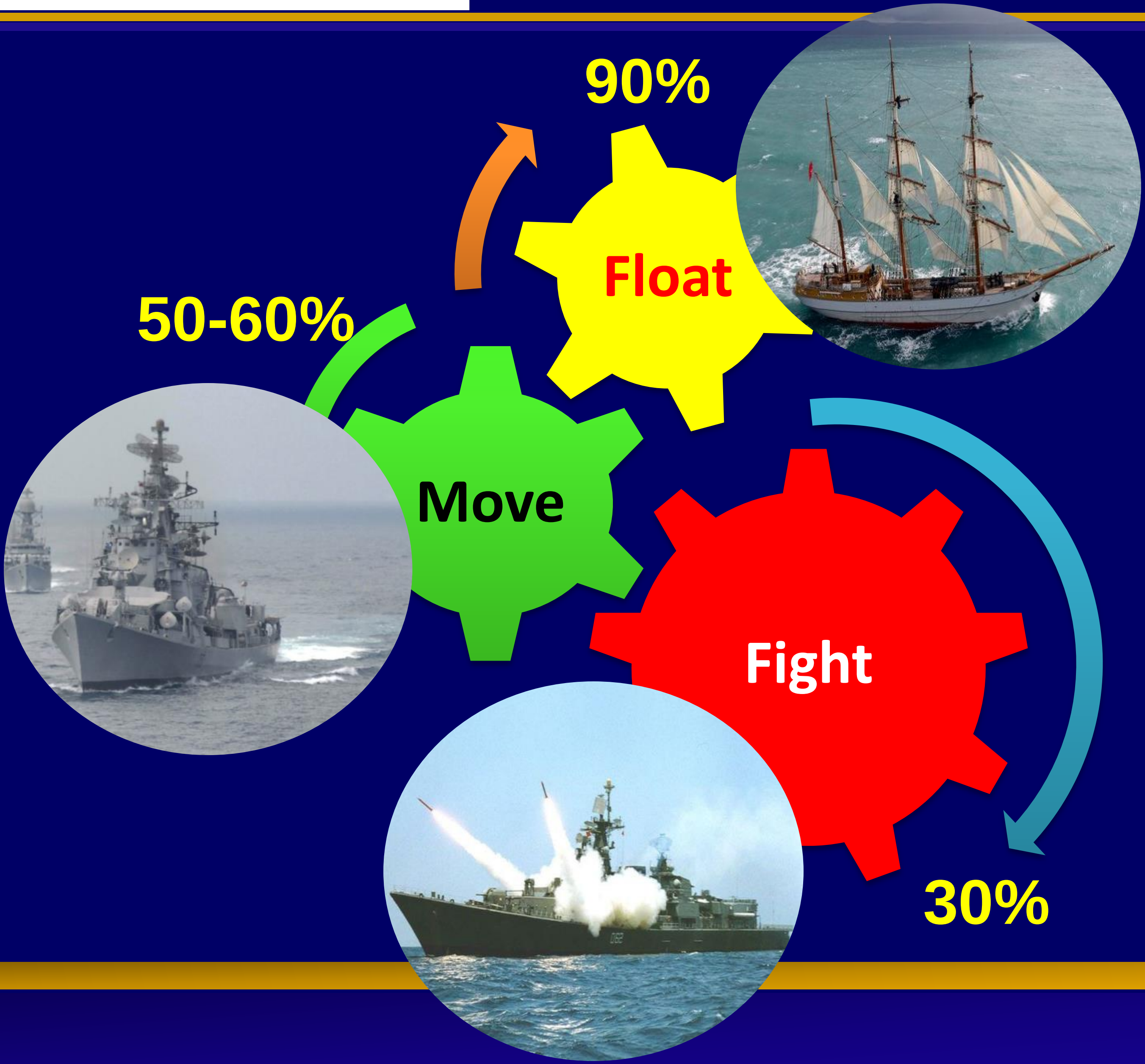

90%

Float

50-60%

Move

Fight

30%

Scope

MAD
RF
Sonobuoy
Lightweight Torpedo
Decoy
HMS
Lightweight Torpedo
TAS
Decoy
Dipping Sonar
Sonobuoy
Lightweight Torpedo

- Torpedo & Directed Energy Weapons
  - Technological advancement
  - Availability
- QRs for Torpedo & DEWs
- Way Ahead

Advancement
in
Torpedo Technology

Torpedo…….

Advancement in Torpedo Technology Contd...

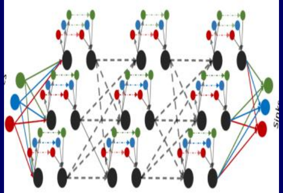

- Broadband sonar, multi-beam & INS
- Multi target tracking capability
- S/W based & high speed sig processing
- Re-programmable S/W & search pattern
- ACCM features
- Digital servo and retractable controls

Torpedo…….

Advancement in Torpedo Technology Contd...

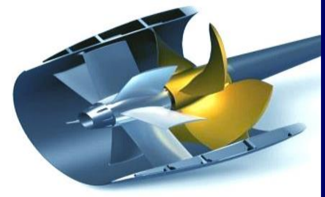

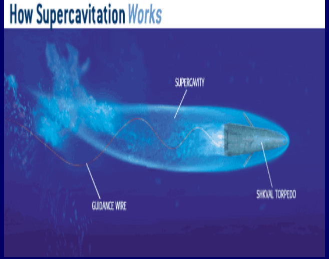

How Supercavitation Works
SUPERCAVITY
GUIDANCE WIRE
SHKVAL TORPEDO

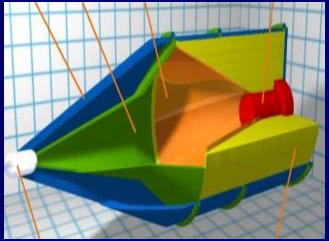

- Al-Ag_2O, Li-ion & Cu-Ni batteries
- Brushless motor & pump jet propulsion
- Semi open/close loop thermal propulsion
- Super-cavitation propulsion
- Tandem shaped charge insensitive W/H

Torpedo…….

Advance Torpedoes

Torpedo	Origin	Speed (Kts)	Range (Kms)
Black Shark Adv	Italy	18- 50	>50
DM2 A4 Sea Hake	Ger	50	> 50
A244S MOD-3	Italy	36	13.5
Flash Black	Italy	50	20
F21 HWT	France	25-50	> 50
Spearfish HWT	UK	40	48
Torpedo (2000)	Swed	40	> 40
MU-90	Eutorp	50	23
MK 48 ADCAP	USA	55	38
MK 54	USA	28 -43	12
Shkval E	Russia	> 200	7 - 10
UGST	Russia	50	40

Torpedoes in Indian Navy

Torpedo		Origin/ Country
A244S		Italy
AT1ME		Russian
MK 54		USA
TAL MK-I		India
53-65 KE		Russian
CET 65 E		Russian
TEST 71ME		Russian
SUT		Germany
Varunastra	VARUNASTRA NSTL, DRDO	India

Directed Energy Weapons

Directed Energy Weapons

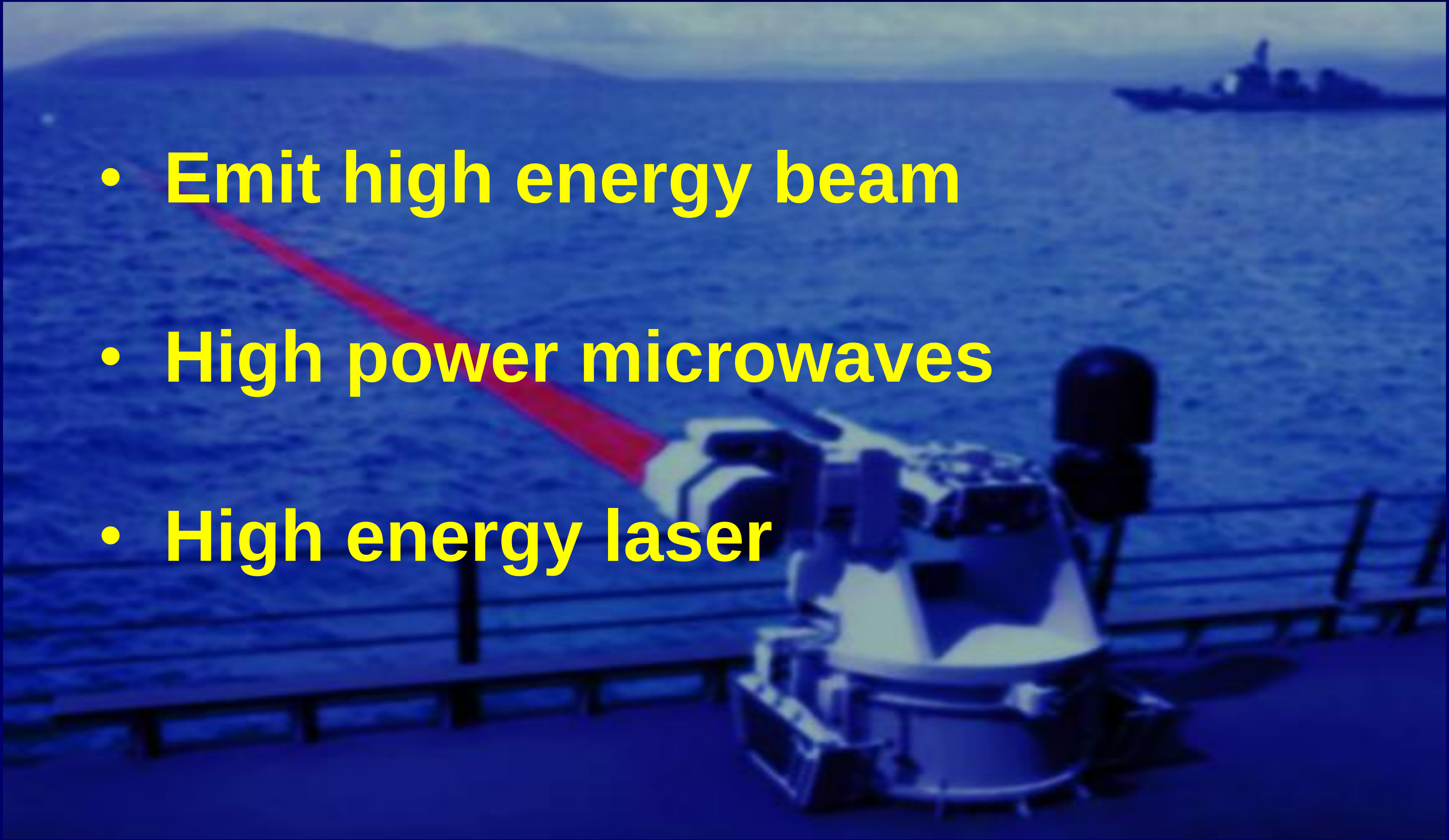

- Emit high energy beam
- High power microwaves
- High energy laser

DEW Advantages

- High Speed of Engagement
- Effective for Defensive/Offensive Role
- Immunity from Gravitation Constraints
- Space Requirement – Limited
- Effective against Moving Air Targets
- Minimal Collateral Damage
- Low Marginal Cost per Shot

High Energy Laser (HEL)
on Rocket from 1.6km away

DEWs from Sea, Air & Land Platforms

HE Laser Mobile Demonstrator **(USA)**

Electromag Rail Gun

(USA)

Active Denial System **(USA/ Israel)**

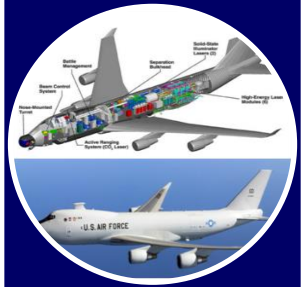

Airborne Ballistic Laser **(USA)**

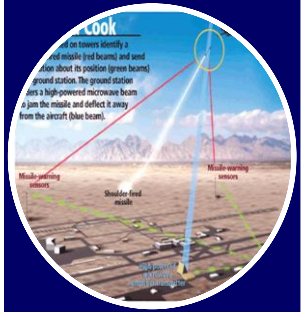

Vigilant Eagle **(USA)**

Laser Weapon System

(USA)

DEWs in Indian Navy

DEWs not as potential destroyer

10 KW HEL demonstrated by DRDO

Work in progress for 25 KW HEL

Develop 100-500 KW HEL

Qualitative Requirements (QRs)

QRs....

QRs for Torpedoes

Require-ment	**Remark**
Speed	Upto 55 Knot with variable speed
Range	Upto 40 Kms
Battery	$Mg\text{-}Cu_2Cl_2$, $Al\text{-}Ag_2O$ & Li-ion batteries
Propulsion	Gas turbine, brushless motor & super-cavitation
Warhead	Tandem insensitive munition W/H & prox EBW fuze
Guidance	CBASS & improved signal processing algorithm
Launch I/ P	Tactical S/W algorithm with target data & upgrades
Control	Intelligent digital controllers & actuators
Camouflage	ACCM features

QRs for DEWs

- 10-25 KW for non-lethal engagement
- 100-500 KW against:-
  - Missile
  - Rocket
  - Aircraft
  - UAVs

Way Ahead

Way Ahead….

Joint Venture

Requirement	Scope for Designers & Industry
Torpedo & Sub-systems	
Mech Parts, W/H & Prop	Al-Mg alloy, FRP & nano materials, **IM**, Mg-Cu, Al-Ag_2O & Li-ion batteries, brushless motor, GT, pump jet & supercavitation prop
Guidance & Control	Sonar imaging & det, CABSS, PCBs, tgt wake seeking, mag anomaly det & super quantum induction device homing, adv accl & sensors.
Process Control	Adv sig processing, tact S/W algorithm for tgt data, friendly forces positions, sound profile & mid-course upgrade, dig ctrl & actuator
Miscellaneous	Fibre optic cable, gene, invertor, shaft, bearing, gear box, bottle, gyro, pump, valve, fastener, cable, pipe, dynamic seal, loc & rec aids
Directed Energy Weapons	
DEW Systems	Ship borne 25-500 KW using High Power Microwaves or Lasers for non-lethal & lethal engagement

Way Ahead….

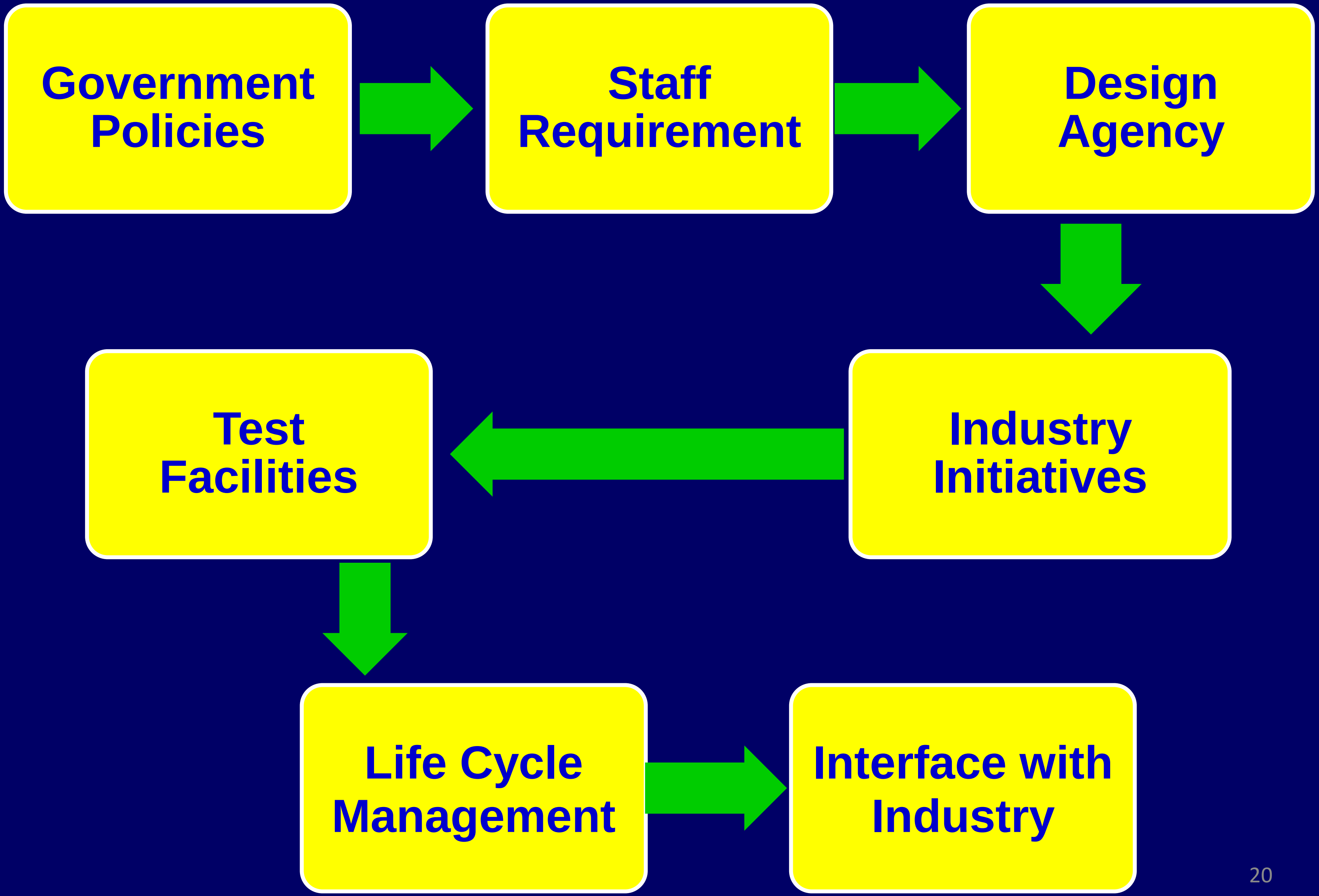

Government Policies

Staff Requirement

Design Agency

Industry Initiatives

Test Facilities

Life Cycle Management

Interface with Industry

Conclusion

- Govt promulgated various major policy guidelines
- ***IN*** "Indigenisation & MCPP"
- Harnessing knowledge base, infrastructures for "**Make in India**" prg under "**FIGHT**" category
- DEWs identified as core mil technology requirement
- Opportunities for strategic partnership

Thank You

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply *using the scalar interferometers to manipulate and engineer the human psyche*. To me it is the most mind-boggling and frightening aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is nothing you can point to that is the mind. As Bearden says, "The mind is time-like."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, a thought or feeling may be made to arise in a person or persons who are in the interference zone (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of *psychoenergetics and the coming "Mind Wars."*

Scalar Psychoenergetics in its most primitive form simply "entrains" all minds in the target area into a deep hypnogogic trance. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of Paul Revere, who heroically conveyed important warnings for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: "One if by land, Two if by mind."

Indeed one strategy in a Mind War would be to simply take over the minds of the "leaders" of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure.

The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could take over control of a person's mind, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)."
> **Bearden: http://www.cheniere.org/misc/time.htm**

In a slide (below image) **Bearden** illustrates how a Scalar War would involve a psychoenergetics attack on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and getting them to shut down their systems.

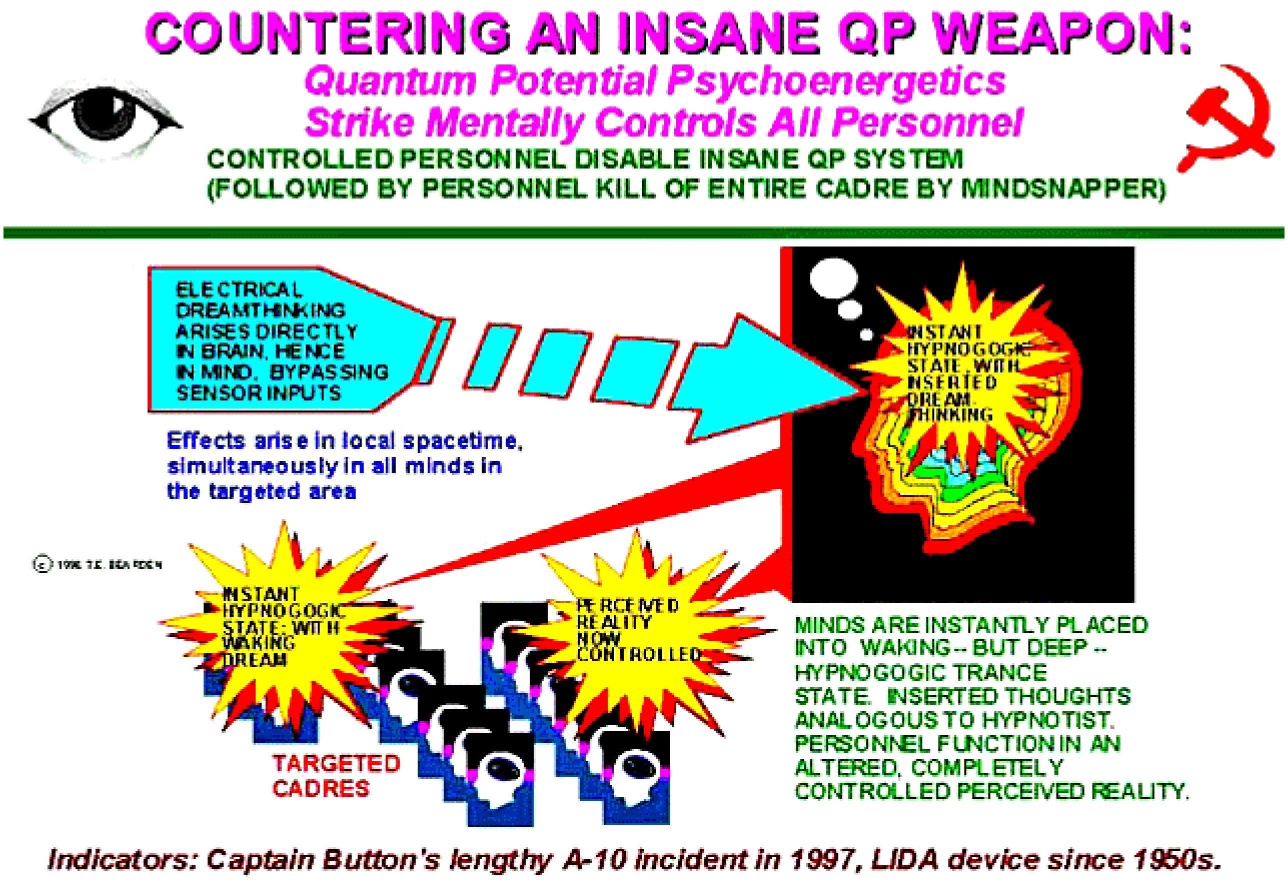

COUNTERING AN INSANE QP WEAPON:

Quantum Potential Psychoenergetics
Strike Mentally Controls All Personnel

CONTROLLED PERSONNEL DISABLE INSANE QP SYSTEM
(FOLLOWED BY PERSONNEL KILL OF ENTIRE CADRE BY MINDSNAPPER)

ELECTRICAL DREAMTHINKING ARISES DIRECTLY IN BRAIN, HENCE IN MIND, BYPASSING SENSOR INPUTS

Effects arise in local spacetime, simultaneously in all minds in the targeted area

MINDS ARE INSTANTLY PLACED INTO WAKING -- BUT DEEP -- HYPNOGOGIC TRANCE STATE. INSERTED THOUGHTS ANALOGOUS TO HYPNOTIST. PERSONNEL FUNCTION IN AN ALTERED, COMPLETELY CONTROLLED PERCEIVED REALITY.

INSTANT HYPNOGOGIC STATE, WITH INSERTED DREAM THINKING
© 1998 T.E. BEARDEN
INSTANT HYPNOGOGIC STATE: WITH WAKING DREAM
PERCEIVED REALITY NOW CONTROLLED
TARGETED CADRES

Indicators: Captain Button's lengthy A-10 incident in 1997, LIDA device since 1950s.

And ultimately psychoenergetic warfare goes to the very heart of human identity itself.

For if my thoughts might no longer be "mine,"

- Then who and what am I?
- Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme?
- Will the "secret government" eventually dictate directly into your mind how you feel about yourself?
- Or what you think you are?
- Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of **scalar psychoenergetics** as expounded by **Col. Bearden** seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is **Mind Control** and **EM Wave Polarization Transductions.** This is such serious stuff that **Bearden** includes a strong warning about misuse of this knowledge.

Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories.
>
> We do not propose or condone any use of these procedures for non-approved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results.
>
> We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral."
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p01.htm**

After the serious warning comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations.
>
> We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transaction mechanisms whereby differing EM wave polarization s can be transformed one-into-the-other are presented."
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p01.htm**

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman.

Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" *(longitudinal wave)* nature of the human mind.

Western Science Remains Largely Materialistic

> "Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.
>
> "With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.
>
> "Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general."
> Bearden: http://www.cheniere.org/explore articles/mind control1/p05.htm

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum.

If only transverse waves are used one pretty well has to hit the target mind with a sledgehammer of waves.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

> "Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.
>
> "While KGB scientists also use TW EM "brute force", TW waves when necessary, they do "imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures required to directly perform the mind engineering desired...
> Bearden
>
> "These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for engineering the entire human collective unconscious simultaneously, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely engineering the memory and knowledge base of an individual.
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p09.htm**

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that all mind operations are time-like, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..."
> **Bearden: http://www.cheniere.org/explore articles/mind control1/p08.htm**

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states.

A kind of "gnome" of the psyche. Bearden describes the process of the decoding of this "gnome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to provoke any body or mind change they wish. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc. changes and interactions."
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p01.htm**
>
> "A second stage in the research would be to test the correlates and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment."
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p02.htm**

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is.

The intense secrecy with which this knowledge has been kept from the public has led to the current farce of the "government" pretending it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time.
>
> Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, may not operate under very much legitimate government control at all, and may operate specifically for the purposes of the rogue group or groups that have gained control.
>
> "In a nutshell, that’s what may be going on in the clandestine mind control projects in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well outside all laws and ethics, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, it could account for what seems to be actually occurring.
>
> As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is power and secrecy. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are to rogue groups often just idle words. They have their own agendas.
>
> And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It’s a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification."
>
> "The Russians, with their additional knowledge of the actual mechanisms in the transforms, will be much more advanced than the West, because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U (l) electrodynamics.
>
> Further, the Russians have decades of use of longitudinal interferometry beams to reach right through the earth and ocean and produce stringent EM effects at a distance.
>
> So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass."
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p05.htm**

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers.

One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion dream-thinking until he died that everything was normal.
>
> All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour.
>
>> "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."
>
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p06.htm**

The paper Mind Control and EM Wave Polarization Transductions ends with the most chilling and frightening possibility of all: the psychoenergetic engineering of the human species as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

> "The KGB psychoenergetics weapons scientists because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials.
>
> "The KGB scientists also know that something very like Gaia a collective unconscious mind for all species on earth also exists, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials."
>
> "Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.
>
> "I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then convert the human species easily into a sort of "ant" society, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" communist style is still a variation of feudalism (all systems tend to feudalism or some form thereof).
>
> "The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...
>
> "Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to condition them for performance of merciless acts." 13
> **Bearden:** http://www.cheniere.org/explore_articles/mind_control3/p07.htm
> More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed largely criminal elements to take over this technology while leaving the civilized elements in the dark.

Openness and the spreading of this knowledge among good and decent people is essential, and for that we will have to overcome our frightened denial.

> "Meanwhile, rogue groups amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, unless great care is exercised by the oversight committees in the House and the Senate.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed."
> **Bearden:** http://www.cheniere.org/explore_articles/mind_control3/p08.htm

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes a plea to the nations of the world regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. The excesses in its bad use are a potent threat to all nations on earth.
>
> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day - perhaps 30 years hence - when education will be accomplished by directly loading the software into the mind. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering.
>
> Another three weeks, an MD And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized.
>
> There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium."
>
> **Bearden: http://www.cheniere.org/explore articles/mind control3/p08.htm**

(Perhaps after so many frightening topics a slight break is in order here to consider the lighter side of psychoenergetics. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley’s fanciful idea in his novel "Brave New World," an entertainment format which he called "The Feelies."

Those were something you went to, like going to the movies, but you would experience all the actual feelings themselves, as if the depicted events were actually happening to you.

Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that The Feelies are now possible, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more new tracks now recorded on the film next to the sound track.

These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater’s interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but can he really feel it? Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person’s feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be deep sobbing and a flood of tears, if the director so desired it. You would feel it at your core. Or, take a big chase scene, you would be terrified. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of Horror-Feelie would be too much for many people to take. Do I really want to feel that an actual vampire is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the definite feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up that despair for scanning?

Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience’s minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

Go Back to Scalar War

Go Back to Psychoenergetics

<<< Scalar Wars

The Brave New World of Scalar Electromagnetics

by Bill Morgan

(See also **NEW 2007**: Book Review of Bearden's "Oblivion - America at the Brink")

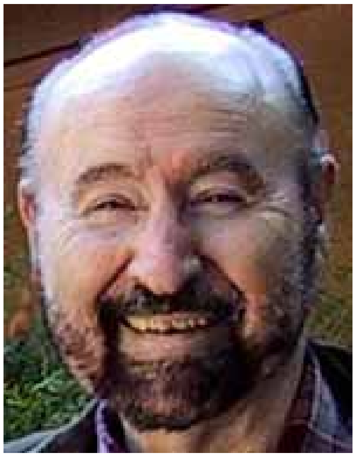

Tom Bearden (2001)
Lieutenant Colonel, U.S. Army (Retired)

For the past six months I have been undergoing the greatest **paradigm shift** I have ever had to go through. It has rattled my nerves and shaken my bones. This intense adjustment of my "world"has come about by studying the information given by Col. Tom Bearden at his website **Cheniere**. The new knowledge there has necessitated a total revision of my ideas about physical reality, the world we live in, and the future of humanity.

This paradigm shifting even actually made me dizzy on certain days as I tried to absorb and digest Bearden's vast amount of information. I am not a scientist at all, just a layman, and I have little comprehension of the math and high physics of this new science called "Scalar Electromagnetics."But there is a great deal of information at **Cheniere** which **needs to become common knowledge** as fast as possible, for the sake of the survival of life on earth.

To that end I have put together this small primer of Bearden's ideas, as a kind of "beginner's guide" to his website. (I have tried to underline certain words and phrases which are part of the new "jargon" of talking about these "new" phenomena. Also I have taken the liberty of **bolding** certain words and phrases to help them stick in the mind).

(Note: Throughout "**EM**" means "**electromagnetic,**" and "**LW**" means "**longitudinal wave**")
You can download this article in Word format

This article has six sections:

1. New Waves 2. Tapping the Waves 3. Weaponization 4. Healing 5. Psychoenergetics 6. As It Stands

2006 NOTE: Click here for a current status history of the MEG from Tom Bearden

1. New waves discovered

Longitudinal EM energy fills vacuum of space, the time domain of spacetime, time as compressed energy, E=tc2, waves of time, phase conjugate wave pairs.

> **"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels."** Nikola Tesla
>
> **"At any point and at any time, one can freely and inexpensively extract enormous EM energy flows directly from the active vacuum itself."** Tom Bearden

I guess the first thing to try to comprehend is that a **"new" kind of electromagnetic (EM) wave has been discovered** in the empty vacuum of space which, when engineered, can be an inexhaustible supply of energy in great magnitude at any place in the universe. The word "new" is in quotes because the

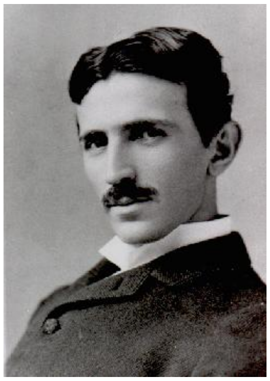

discovery really goes back to Nikola Tesla and his discovery of what he called "radiant energy." It is also not "new" because the Russians (KGB) have been working on this technology **for over 30 years** and have **weaponized** these "new" **longitudinal scalar waves** to a great degree.

These are the weapons Nikita Khrushchev spoke of in January, 1960.

> "By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics - weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon so powerful that it could **wipe out all life on earth** if unrestrainedly employed.'" Tom Bearden

After over 30 years of development, and extensive testing around the globe, these new scalar electromagnetic weapons are up and running and ready to go. Tom Bearden, at his website Cheniere, dicusses the history of these new scalar electromagnetic weapons in his paper "Historical Background of Scalar EM Weapons."

Some Immediate Implications

> The implications of successful engineering of the longitudinal waves are enormous, and will change the world as we know it, one way or another. Among other things, these discoveries mean that:
>
> **1. The solutions to the energy crisis and the "oil problem" are in hand. These oil wars are unnecessary. There is endless energy available freely from the domain of time.**
>
> **2. Unbelievably powerful weapons are not only possible, but are already operating in several nations. The many powers of these weapons are unprecedented and mind-boggling.**
>
> **3. The cure of diseases such as cancer and AIDS, in fact nearly any disease, has become possible within a few years of sufficient funding. Everyone can be made healthy and stay healthy.**
>
> **4. Mind control on a mass scale has now become possible, and the machines to do it are already in place in certain nations. It has become possible to mentally enslave whole populations with the twist of a few dials.**

So the layman will need to understand that **there is a new kind of electromagnetic energy** that is altogether different from what he knows, e.g. radio, TV, cell phones, etc. The ordinary EM waves that we have known about are called **transverse** EM waves, to distinguish them from the new **longitudinal** EM waves. These scalar waves do not actually exist in our "material" world, but exist only in the vacuum of empty space, or the time domain. And we must keep in mind that this vacuum of space we speak of exists all through everything. Even our bodies are mostly empty space between atoms and molecules. So the gateway to this seething ocean of energy can be there at every point in the universe. This seething ocean of energy is all around us and all through us.

Emptiness is Full

This amazing discovery announces that **the "emptiness" of empty space is in fact not empty**, but a **great ocean** of seething energy!

Col. Bearden refers to this ocean of energy as being of the "time domain." Energy out of time? It seems like something from Star Trek but this is the point which the new science of scalar electromagnetics has reached. And where it is going may be beyond anything Star Trek could have dreamed of.

We live in a 3-dimensional world, which physics calls "3-space." But there is also **spacetime**, or 4-space, or the "4th dimension." Then suddenly comes this amazing new knowledge that **time itself is actually compressed energy**. And it is energy which is compressed by exactly the same factor by which **matter** is considered compressed energy: the speed-of-light-squared!

So we have a new companion to the famous **E=mc2**. It is now paired with **E=tc2** (where t is actually "delta-t," or change in time). Or (amaze your family and friends and) say:

"E equals em-cee-squared"

and

"E equals delta-tee-cee-squared"

It has a nice ring to it, like some freedom bell announcing a new era.

And as the atomic bomb released the compressed energy in matter, so can we now unleash **the tremendous energy that is compressed into time itself.** It gives a completely new meaning to the term "time bomb."

Some New Terminology of Scalar Electromagnetics

A new science, **several** new names

"Those [Russian] weapon scientists resurrected an old term from the history of electrodynamics, called "**energetics**". That is their approach to a **unified field theory**, where everything is based on "energetics". This model as its foundations uses a very similar approach to that "single fundamental unit" model, where energy is the unit. If one makes the energy EM in nature, then one has the Russian energetics approach. This unified approach gathers everything in, including all energy actions and relations in inert matter (the first branch of

> energetics, called by the same name), all field and matter interactions in living matter (the second branch of energetics, called "**bioenergetics**"), and all mind operations and mind-matter interactions (the third branch of energetics, called "**psychoenergetics**"). [ed: sometimes also called "psychotronics" by the Russians]
>
> "Note that the weaponeers sliced energetics into three branches, with each branch depending on the nature of the "targeting".
>
> "So I took a hint from them, since the stuff and the approach worked. They had already weaponized it highly."
>
> Bearden : http://www.cheniere.org/correspondence/021502a.htm
> See also "Subdivisions of Soviet Energetics"

Rick Andersen in an article called "What is Scalar Electromagnetics?" describes the **new science** this way:

"Scalar EM is the brainchild of Lt. Col. (retired) Thomas E. Bearden, a systems analyst and wargames specialist who has been advocating a view of electromagnetics which is based on the notion of a vast, unseen background of scalar energies (as opposed to vector energies) **which underlie all physical reality.**

"If Bearden is correct in his Scalar EM theory, then we can build devices which would enable us to **alter gravity**, time, inertia, and the apparent mass of an object. This of course has ENORMOUS implications for military applications, **space-vehicle drives, time-travel, teleportation, paranormal phenomena**, and just about every other area one can think of."
http://twm.co.nz/Beard_scalem.html

> **"Time-energy, time-currents, and time-structuring play the dominant role in electromagnetics. *Time-as-energy* eventually becomes engineerable, as easily as is spatial energy now. We are always dealing with *spacetime* and with *spacetime curvature*."**
>
> **Bearden:http://www.cheniere.org/techpapers/Vision%202000%20paper.doc**

The scalar energy discovered in the vacuum of space is sometimes called "**zero-point**" energy. Thomas Valone explains how this term came about.

What does "Zero Point" mean?

"Boyer traces the historical “creation of the vacuum” as proceeding in stages in parallel with the historical development of ideas about the vacuum. To paraphrase, he says that in the 17th century, it was thought that a totally empty volume of space could be created by simply removing all matter and, in particular, all gases. That was our first concept of the vacuum. Just get rid of all the gas.

Late in the 19th century, it became apparent that the region still contained thermal radiation. But it seemed that the radiation might be eliminated by cooling. So the second concept of getting a real vacuum is to cool it down to zero temperature. Just go all the way to absolute zero. Then we've got a real vacuum. Right? Well, since then, both theory and experiment have shown that **there is a non-thermal radiation in the vacuum** and that **it persists** even if the temperature could be **lowered to absolute zero**. Therefore, it was simply called the "**zero point**" radiation."

Thomas Valone http://www.seaspower.com/InsideZeroPoint.htm

"For our very survival, it is absolutely imperative that informed citizens be aware of this dramatic change, which is just now starting. The powerful new science and engineering must be controlled and used for humanity's benefit, not its detriment. Else it will eventually be let loose unrestrainedly, to destroy all life on earth - a possibility indicated by Nikita Khrushchev in 1960." - Bearden

http://www.cheniere.org/explore%20articles/priore1/p4.jpg

2. Tapping the waves

Endless free energy, Tesla's radiant energy, the dipole as the gateway to the energy of the void, MEG in one year, Bedini, Yakuza, etc. End of Oil, geopolitical chaos, rogue groups

Move Over G.E. & Big Oil, Energy is Everywhere!

Third-world nations set to rise

> "Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to **free energy** and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed ... Today it is still being suppressed by the orthodox establishment."
> http://www.cheniere.org/books/ferdelance/s25.htm

The vast seething ocean of energy of the vacuum, the longitudinal scalar EM waves of the time domain, can now be tapped and "transduced" into ordinary (transverse) EM energy in our 3-Space world. The process can be compared to putting a paddlewheel into a river. The energy acquired is free, since the river is there flowing whether we tap it or not. And it is a mighty river, and is not diminished by our paddlewheel.

This means that **a permanent solution to the "energy crisis" is at hand**, if only those who hold it **so secret** would give it up to humankind.

It means our species' **insane use of oil** is no longer necessary. We can make automobiles that need no fuel at all. We can build motors that power themselves. It means that unlimited electrical power can be available for free anywhere in the world. Every house and even shack can have its own power supply.

Of course there is the cost of the equipment at first, but the energy flow is free for the presumably very long life (no moving parts) of the generator. It will probably take decades to shift over to scalar power.

But the ability to tap the scalar (longitudinal) EM waves of the vacuum (which exists everywhere) also means that **unbelievably powerful weapons** can be, and already have been developed. In fact these weapons are so dangerous that the **fear engendered** becomes a psychological obstacle to the grasping of the facts.

I struggled with this myself for months, and even though Tom Bearden had completely convinced me already, it is hard to accept the fact that for everyone on earth the "worst that might happen" has suddenly gotten much worse. It takes time to deal with that, to realize it and get over it.

Indeed, the possibilities of these new weapons, which are tapping the **huge energies of time**, are so terrifying it is natural for the mind to simply say: "I don't want to hear about that."

And the government, likewise, pretends these weapons do not exist and tries to keep it secret from the people. Because **most people would indeed be truly alarmed** to know the actual situation. Col. Tom Bearden has decided that the people need to know the truth anyway. As a People on a Planet we must face this now, and **ban the use of scalar weapons in war or in mass mind control**. The current deployment of these weapons is a **world emergency** which we must all face.

The physics and mathematics of the longitudinal waves and the time domain are clearly beyond the layman, but I recommend scanning through them anyway just to get a sense of the wonder and the elegance of it all and some of the exotic terms which are used. Two of the most important papers by Col. Bearden are:

Giant Negentropy from the Common Dipole
The Final Secret of Free Energy.

Despite the complexity of the science involved, Bearden sums up the **essential secret** of it all in a few words.

> "There is no problem at all in extracting all the energy one wishes from the active vacuum, anywhere in the universe, at any time. **Just make a dipole.**
>
> The problem is in (1) catching some of that freely gushing EM energy in a circuit containing a load, and (2) dissipating the caught and collected EM energy in that load to power it, without using half the caught energy to destroy the source dipole's).
>
> That is the ONLY real energy problem on the planet, and always has been."
>
> http://www.cheniere.org/correspondence/022502.htm

A dipole could be a battery, which is polarized into negative and positive. Or the terminals of a generator. Or any magnet with its two poles. Or an atom with its positive and negative charges. Or the earth's magnetic field, the sun, and all heavenly bodies. Even a galaxy must be some kind of mega-dipole. Wherever there is a dipole there is already immense scalar EM energy ushering out of and back into the vacuum. And there have already been demonstrated a number of diverse methods for tapping that energy as it flows. (John Bedini, Edwin Gray, Bruce de Palma, and others). Bearden provides what he calls a "partial list" on his website.

> "In other words, **nature then gladly gives us** as much EM energy flow as we need, indefinitely — just for paying a tiny little bit initially to "make the little dipole." After that, we never have to pay anything again, and nature will happily keep on pouring out that 3-flow of EM energy for us. This is the **giant negentropy** mechanism I uncovered, performed in the simplest way imaginable: just make an ordinary little dipole." Bearden
> http://www.cheniere.org/techpapers/Unnecessary%20Energy%20Crisis.doc

Important note 2006: Since the enthusiastic early report below, the hopeful timetable envision has gone quite off track. Bearden had said that the MEG could appear as early as 1-2 years ***after getting funding*** for the final research and scale-up. They have **still** NOT gotten the funding as of 2006.

Click here for a current status history of the MEG from Tom Bearden

The original text of article continues here...

How to Buy a Free Energy Machine

Answer: Just wait one more year! 2003. Then get off the Grid!
[note 2006: The above line was unfortunately naive and overly optimistic! - author]

Perhaps the most eagerly awaited event now is the marketing of Bearden's **M**otionless **E**lectromagnetic **G**enerator (MEG) scheduled to begin in about a year's time. This is when the proof will hit the fan. Many people have said to me, when talking about free energy, "I'll believe it when I can go to the store and buy one." Hopefully in a year's time they will be able to order their own MEG. And then they will believe, because it works and is working.

Says Bearden: "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat."
Bearden : http://www.cheniere.org/correspondence/011202.htm

Soon the cat will be fully out of the bag. The first MEG units will supply 2.5 kilowatts of free electricity. Forever. Units can be connected together to double or triple the wattage. After that gets going, 10 kilowatt generators are planned. Energy will flow freely from the time domain indefinitely, and there are no moving parts to wear out. This manufacturing project is a heroic effort and will change the world as we know it. For once people are in possession of such devices there can be no further denial from anyone, including the government and/or "secret-government."

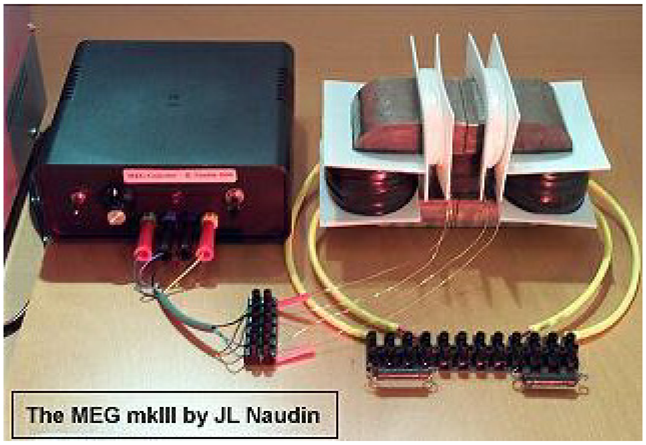

The Motionless Electromagnetic Generator
(Successfully replicated here by J. L. Naudin)

The physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum."

The device was successfully replicated by J.L. Naudin in France, and others, like Steve Utne, are also working on their own models. But **Magnetic Energy Limited's MEG** is actually scheduled for production soon and looks to be the first commercially available free-energy generator in history!

I imagine it will not take long for tinkerers to put them into their electric cars and make the first fuelless automobiles since Tesla's unheralded (and unheeded) experiments. Or for builders to begin building houses powered by MEG's. Houses which will never see electric bills. How can they ever make enough of these to meet the demand that will suddenly be there? Bearden is right, it needs to be an all-out crash program of the government **on the order of the Manhattan Project**.

I must include here that Bearden is quick to point out that he is not himself the inventor of the MEG, but is a partner and advisor in bringing it to the public market through **Magnetic Energy Limited, Inc**. I must also admit that I am praying for the success of this venture, and the ending of the absurd and criminal withholding of this technology from the people by the black-ops military-industrial people. This secrecy has cost the planet over 30 years of unnecessary pollution.

The first units off the assembly line, which is being set up in a "friendly" unnamed nation, will usher the public at last into the new era of scalar electromagnetics.

As of December, 2001, the situation with the MEG was described by Bearden in a letter.

> "What we presently have with the MEG is a successful laboratory experiment. At least a year's very hard research will have to be done before we will be ready to put a commercial power supply into production.
>
> "Consequently, we have made an agreement with a foreign partner (the National Materials Science Lab of the National Academy of Sciences of a friendly foreign nation) to do that year's research. At the same time, we are trying to make an agreement with one or more large financial partners here in the U.S." http://www.cheniere.org/correspondence/121901.htm

> "So we are hopeful that we will be able to start introducing units on the market about a year from now. Quite simply, we will either succeed or we will fail. But we will give it our very best effort."
> http://www.cheniere.org/correspondence/012202a.htm

Bearden has given another technical explanation of how the MEG works here in a letter.

Update : Motionless Electromagnetic Generator patented, 3-26-2002

Click here to read the patent

"In Magnetic Energy Limited, all business discussions and arrangements are taken care of by Dr. Lee Kenny, Managing Partner. There are five of us who are co-inventors of the MEG, and two of the others are really the principal inventors." Bearden
http://www.cheniere.org/megstatus.htm

Lee Kenny can be reached by E-mail at: jlkenny@ingr.com

Free Energy has been Suppressed

Secret Government Aware of Free Energy

"In this world of contemporary times, all the agencies, CIA, FBI, KGB, NSA, Electric Power Research Institute, DARPA, Brookings Institution, Henry Kissinger, Edward Teller, et al., all are aware of [free-energy researcher Bruce] DePalma and his ideas.

"Because these individuals and institutions are employed by the ruling elite to forecast the future and satisfy present needs and demands, it is clear that **free energy is a threat** to the world order constructed by **business** and the **mindset of those who want to own the world**.

"Free energy represents Man's aspirations and dreams of freedom and equality, uniform division of resources and the ability to choose one's own future.

"The fact that free energy is suppressed speaks to the greed and self-interest of a ruling elite which, even in the face of an emergency of global starvation, resource depletion and environmental pollution, will not give one inch if it means loss of control. This is an attitude of **paranoid delusion** and fantasy which can only arise from the alienation of a **class of elitists** who, through their money, are completely **insulated** from the reality of the day to day fight for existence of the common man."

http://depalma.pair.com/Absurdity/Absurdity07/ProblemOfFreeEnergy.html

"Before the turn of the century, **Nikola Tesla** had discovered and was utilizing a **new type of electric wave**. Tesla repeatedly stated his

> waves were non-Hertzian, and his wireless transmissions did not fall off as the square of the distance. His discovery was apparently so fundamental (and his intent to provide free energy to all humankind was so clear) that it was responsible for the withdrawal of his financial backing, his deliberate isolation, and the gradual removal of his name from the history books."
>
> Bearden http://www.cheniere.org/books/part1/starting%20pages.htm

Bearden describes one clear case of the suppression of free energy when the Yakuza killed the **Kawai engine** in his very presence in Huntsville, Alabama.

> "That night a jet arrived from Los Angeles, with a Yakuza on board. The next morning Kawai no longer controlled his company, his invention, or his own fate. The Kawai party was in fear and trembling. The Yakuza coldly dissolved the agreement, they packed up the two Kawai engines we had, and left. And that was that."

So it is not only governments who are keeping this revolutionary technology from seeing the light of day. The Yakuza? KGB? The Aum Shinrikyo cult? Our "leaders" need to be honest about what is going on, otherwise they look foolish, acting out a kind of "Alice In Wonderland" script before the public, a script that has become completely unbelievable.

Concerning the criminal withholding of this scalar technology by the military black-budget people, and the cover-up of the actual situation regarding scalar weapons throughout the world, Tom has this to say:

> "The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to **defend** the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.
>
> **"They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up.**"
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Energy Crisis Solved

One of the "must read" documents at Cheniere is The Unnecessary Energy Crisis: How to Solve It Quickly, especially since the "**oil wars**," which Bearden has foreseen for decades, have now already begun.

Here are some clips from that Document:

> "To avoid the impending collapse of the world economy and/or the destruction of civilization and the biosphere, we must quickly replace much of the "electrical energy from oil" heart of the crisis at great

> speed, and simultaneously replace a significant part of the "transportation using oil products" factor also.
>
> "The technical basis for that solution and part of the prototype technology required, are now at hand. We discuss that solution in this paper." Bearden

And although the solution to the energy problem and the oil wars may be at hand, Bearden suggests there is still an urgent rush that needs to be made to save human civilization from collapse.

> "Once the technology hardware solutions are ready for mass production, even with a massive worldwide deployment effort some five years are required to deploy the new systems sufficiently to contain the problem of world economic collapse. This means that, by the end of 2003, those hardware technology solutions must have been completed, and the production replacement power systems must be ready to roll off the assembly lines en masse.
>
> "The 2003 date appears to be the critical "point of no return" for the survival of civilization as we have known it.
>
> "Reaching that point, say, in 2005 will not solve the crisis in time, and the collapse of the world economy as well as the destruction of civilization and the biosphere will still almost certainly occur, even with the solutions in hand."
>
> "The **good news** is that we now know how to easily initiate continuous and powerful "electromagnetic winds" from the vacuum at will. Once initiated, each free EM energy wind flows continuously so long as the simple initiator is not deliberately destroyed."

On page 9 he explains the concept of the **dipole**, and how there is already huge scalar energy flowing in the time domain around any dipole. Hold a magnet or a battery in your hand and you hold a stress in time and a beautiful flowing loop of longitudinal waves!

Bearden decries the scientific community in the West for being unable to let go of their dogma long enough to see that the **greatest discovery in human history** lies at their feet, but for their stubborn ignorance.

> "The present energy crisis then is due totally to that "**conspiracy of ignorance**" we referred to, that is maintained by the scientific community, and that has been maintained by it for more than 100 years.
>
> "This is the real situation that the environmentalists must become aware of, if they are to see the correct path into which their energies and efforts should be directed — to solve both the energy crisis and the problem of gigantic pollution of the biosphere."

As far as anyone in the legislative branch of government ever understanding the profundity of the discovery and engineering of scalar waves, Tom Bearden says they are at a great disadvantage in doing so.

> "The nontechnical person — e.g., a Senator or a Congressperson — is operating under a distinct disadvantage. If he or she takes the stance that much better electrical power systems can readily be achieved, **he or she is in fact opposing almost the entire set of University, Government Technical, Power Company, Battery Company, and Organized Science communities**. Further, in most cases his technical advisors are themselves from one or the other of those communities, and likely to go back into that community or those communities when the Senator or Congressperson leaves office, or even before. So the Congress and the non-technical government community at large operate at a great disadvantage...Bearden

National Emergency Declaration is Needed

"The Government *Non-Technical* community (the Senate and the Congress, in particular) are in far better shape than the Government *Technical* community, to appreciate the **world implications of the pending economic disaster**. I am hopeful that both the environmentalists and the Government Non-Technical community will rapidly unite in a common goal to **get this vacuum energy program launched**, under a **National Emergency** declaration. If so, then they can solve the energy crisis and the pending economic crisis, in fairly short order, and permanently."

unnecessary_energy_crisis.htm#congress_disadvantage

In the conclusion of his paper The Unnecessary Energy Crisis: How to Solve It Quickly Bearden again states the urgency of replacing our centralized power grids with highly dispersed and local taps on the freely flowing energy of the active vacuum. Given the abilities of the Tesla howitzers, the power grids of all nations are sitting ducks.

"Personally, the present author regards the increasing **energy crisis as the greatest strategic threat** to the United States in its entire history. I will do anything within my power to help prevent what I perceive to be the looming economic collapse of the Western world, preceded or accompanied by a sudden, explosive, all-out and continuing exchange of the WMD arsenals of most of the world.

...*In the name of all humanity, let us begin!* Else by the time this first decade of the new millennium ends, much of humanity may not remain to see the second decade."

3. Weaponization

Scalar interferometry, Tesla howitzer, Tesla dome, heat/cold, dudding, 10 nations have them, dead-man fusing & Insane Systems, scalar strategy, earthquakes and weather wars, solar storms, mind attacks, leased to Yakuza, many examples of testing, quantum potential weapons.

> **"Indeed, Soviet energetics weapons are now capable of destroying both our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap."** Bearden
> http://www.cheniere.org/books/ferdelance/intro.htm
>
> "**We have a new weapon, just within the portfolio of our scientists, so to speak, which is so powerful that, if unrestrainedly used, it could wipe out all life on earth. It is a fantastic weapon."** Khrushchev, to the Presidium, Jan. 1960
>
> "We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.
>
> "Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.
>
> "History does repeat itself. But this time it's not we Americans who got the great new superweapon first."
>
> Bearden http://www.cheniere.org/books/ferdelance/s75.htm

Warfare has been changed forever by the development of these scalar energy longitudinal wave howitzers. To get a basic understanding of scalar waves is to have the imagination suddenly run wild as all the implications and possibilities regarding warfare fall into place. One realizes with a certain horror that the world has totally changed, and that there are some very fearsome possibilities. Remember, the power for these weapons comes from the time domain, longitudinal EM waves in the vacuum of empty space, and the power is tremendous and mind-boggling. Being able to blast away at any target from a distant control booth is something that has never happened before. This is incredible power to be in control of and it divides the history of weaponry into "before" and "after." And the destructive power of these weapons is delivered instantaneously to the target **from the local vacuum at the place of the target.**

From an interview:

Questioner: "Are you saying that real energy can be transmitted faster than the speed of light?"

Bearden: "Yes. But hyperspatially. Not through 3-space, but "around" it. Again, this means that it's possible to produce energetic changes in a distant system at a distant place, without transmitting energy "through space" in the normal sense. You transmit "around" space, so to speak, and directly in time or through a higher dimension, depending on the model you're using to understand this. Further, the speed of the internal EM energy transmission is not limited to the speed of light."

http://twm.co.nz/beard_interview.htm

Here are some initial considerations about scalar weapons and scalar wars

1. Tesla Howitzer 2. Tesla Domes 3. Rogue groups 4. WW III 5. Russian "UFOs?"

1. The Tesla Howitzer and its modes of action

Source: Tom Bearden's weapons slides and Fer-de-Lance Briefing

For one thing there is no delivery of any thing to any place. Bombs are obsolete. So are the planes to get them there. No helicopters necessary. All current nuclear devices are rendered nearly useless by electronic "dudding." All distant-destruction can be done from a control room, and to any point on earth. The destructive power of "men" has just increased by orders of magnitude. The new weapons could even cause storms on the sun! Thus there is great urgency to make these **new facts public** and known, with the hope that all nations would come together to ban the use of scalar weapons.

Two scalar antennae together along with the computers to control them, make up a scalar interferometer, and according to Col. Bearden the Russians have hundreds of such installations already. These interferometers are called "**Tesla Howitzers**." They can deliver a giant blast of energy to the distant target site, true "action-at-a-distance".

The **first howitzer mode** is called the "**exothermic**" mode of operation because immense EM energy blasts outward at the target site. The blast of a scalar howitzer can be of near nuclear level in destructiveness, and can be repeated easily, at that place, or nearby, or anywhere.

The howitzer can use a lesser exothermic power setting and simply **destroy all electronics** in the target area. Thus they can render our nuclear missiles inoperable as they sit in their silos by "frying" the electronic circuits that guide them. They can bring down any airplane, anywhere in the world, at any time. Any person anywhere, if their exact position is known, can be assassinated without a shot being fired.

The howitzers can also bring down power grids anywhere in the world, at any time. Thus they could bring Wall Street to an instant halt by disabling all its electronics. Even this one little easy tweak of the howitzer could cripple the U.S. economy in a single day. No wonder the "**leaders**" don't want you to know! This is a new era we are in. The discovery and development of scalar interferometry has put us in a **new world**, whether we like it or not. Anyone could be expected to go into denial, the **befuddled** senators and congressmen included. But we have **to wake up**.

In a **second howitzer mode** called the "**endothermic**" mode, the howitzer sucks energy out of the target area, essentially creating a blast of cold at the distant target. It is even capable of freezing parts of the ocean. Bearden gives a number of cases where these "**cold explosions**" have been witnessed, mostly by airline pilots. April 9, 1984: 1 2 3) A huge mushroom cloud of mostly water is seen rising miles into the sky from out of the ocean, an awesome sight.

In the endothermic mode the sucked-out energy must go somewhere, so it is vented out at some other chosen spot on the earth. These endothermic plumes have been photographed by satellite.

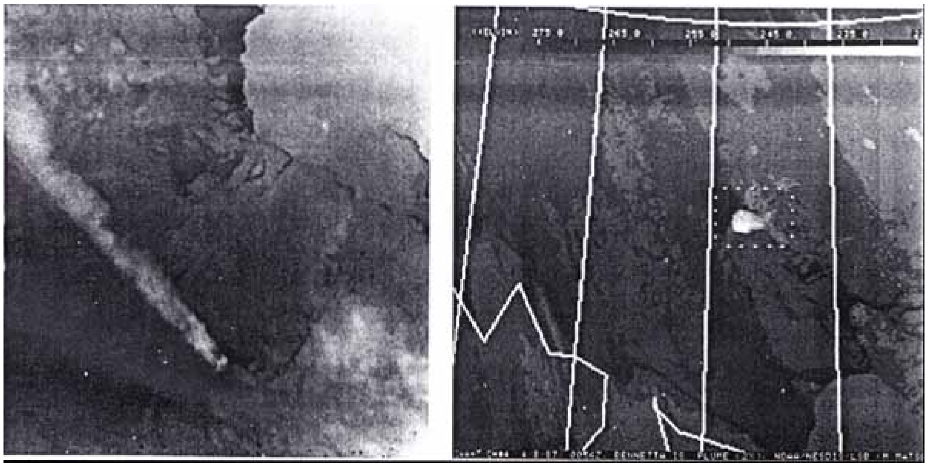

"The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)" Bearden More Plumes

By using both of these **heat** (exothermic) and **cold** (endothermic) modes together the weather can be altered anywhere. Warm the air over here, cool it down over there, put a curl in the jet stream, dissipate clouds, create clouds, whip up a tornado. Bearden speaks about using the **endothermic howitzer mode** as a weapon of war:

> "Cold explosions can be used to freeze tanks, personnel, and equipment. The equipment and tanks thaw out. The personnel thaw out too, but they are dead when they do." Bearden, Fer-de-Lance

(This endothermic mode of the howitzers makes me think that there might one day be a cure for the problem of global warming after all. Could not a big device, situated in space or on the moon, be able to vent a great deal of excess heat outside the atmosphere into space, thus functioning as a kind of planetary thermostat?)

While the United States has experimented with electomagnetic weapons using ordinary **transverse** EM waves, the **longitudinal** [**LW**] weaponized waves are truly of another order altogether.

For one thing they do not have to travel through space, for they come from the time domain, which is everywhere. The energy of the blast, the heat or light or whatever, comes from the vacuum of space **at the location of the target itself**. This is astonishing. The weapons actually just trigger the release of immense energy from the vacuum at the target location.

> "... it is possible to focus the potential for the effects of a weapon **through spacetime itself**, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There **interference of these ripple patterns** creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location." Bearden Fer-de-Lance

The big Tesla howitzers are aimed at their targets by using a worldwide electromagnetic pattern called the "Woodpecker Grid," begun by the Russians in 1976.

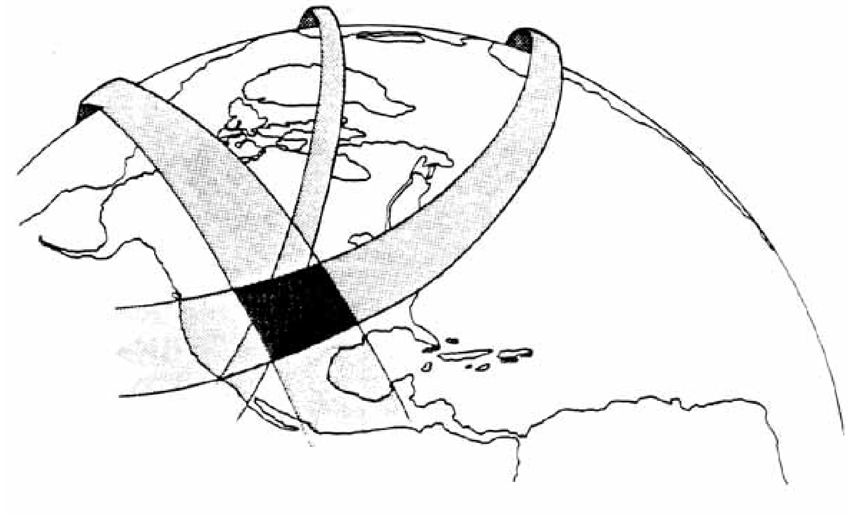

You can hear the Woodpecker signal here:
http://www.cheniere.org/misc/woodpeckersound.htm

The woodpecker grid is just that, a grid of EM waves which provide a channel for the **LW** waves to any point on earth. Computers use the woodpecker grid information to hone down the aiming of the howitzers **to a pinpoint location on earth**. (For that matter, it could be under the earth, in space, or under the sea).

A third howitzer mode is what Col. Bearden calls "**Mindsnapper**" mode. This is a truly frightening mode which affects the electromagnetic mind-body connection. Being mind-snapped at a low level would cause you to lose consciousness. Being mind-snapped at a slightly higher power would "entrain" all minds in the target area into a kind of hypnogogic trance, a state in which they would all be highly susceptible to suggestions and orders.

Being mind-snapped at high level would so disrupt the mind-body connection that everyone in the target area would fall down, instantly dead. Not only that, everything around you and inside you would be killed, so that your sterilized body would **not rot** for quite a long time.

The **Mindsnapper** can be used for a small area, or to **kill an entire population**. In fact, using the scalar waves to affect the human mind has become a new field in its own right, and has been dubbed "psychoenergetics." (This is discussed below in section 5).

Bearden describes the use of the howitzer in **mind-snapping mode** against people, to cause unconsciousness or death:

> "From a military viewpoint, these scalar EM weapons are really quite all-around weapons.
>
> "For example, they are very lethal against personnel.
>
> "Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for

> example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.
>
> "Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious. Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise." Bearden http://www.cheniere.org/books/ferdelance/s80.htm
>
> "Those hit by the scalar EM weapon, however, have a most peculiar death mode.
>
> "Death comes-instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.
>
> "A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in."
>
> Bearden http://www.cheniere.org/books/ferdelance/s81.htm

He goes on to say that the same "bazooka-sized scalar EM pulse weapon" can also be used to knock out a tank with one shot (or pulse). So tank warfare is obsolete. Helicopters? Bearden points out: "... if he's attacked by a helicopter or a fixed-wing aircraft, he can bring the same weapon to bear on it. He can knock down a chopper or a jet. With one shot."

Obituary for Tank Warfare

Tanks useless against portable LW weapons

"For example, suppose one attacks a tank. [with an LW weapon]
With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.
All electronic systems of the tank are dudded. Total systems kill is achieved.
The ammunition in the tank explodes. Total kill is achieved.
The fuel explodes. This is another total kill mechanism.
That tank has been killed totally, by a variety of mechanisms, all simultaneously.
That's a k-kill in any analyst's book.
And don't worry about retrieval and repair. That tank is finished permanently.
Even the most modern tank is just as vulnerable as the most obsolete."

Bearden http://www.cheniere.org/books/ferdelance/s88.htm

One begins to get the idea how profoundly the conventional means of warfare have been obsoleted. All the usual weapons of war have been rendered nearly useless! And **Star Trek's "Phaser**," with its "stun" and "kill" modes is here at last.

Although it is the weapons aspect of scalar electromagnetic which has received the most developmental effort, Bearden points out that a great many other amazing things that may be accomplished in this new field of science. Beam me up Scotty!

> "Superluminal communications systems, hyperspace drive, and **materialization** and **dematerialization** are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science- fiction."
>
> Bearden http://www.cheniere.org/books/ferdelance/s15.htm

Bearden describes a huge test of the Russian scalar weapons systems which occurred in 1985, a test that went completely unnoticed by western scientists:

> "Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.
>
> "For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.
>
> "Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened."
>
> Bearden http://www.cheniere.org/books/ferdelance/s38.htm
>
> "The **Woodpecker grid/howitzer** weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force."
>
> (Bearden http://www.cheniere.org/books/ferdelance/s65.htm)

The whole series of the Fer-de-Lance briefing slides and comments is a "must read" for anyone wanting to be aware of the actual state of things in this world right now. It is a new world and it takes some getting used to. Fer-de-Lance speaks clearly, for the layman, of the way things stand.

What is it like to operate these Tesla Howitzers? Col. Bearden gives us an idea in his "Fer-de-Lance" briefing, Slides 66-71.

Who needs armies anymore? A entire war can be fought and won by a handful of men sitting in a comfortable control booth! The scalar electromagnetics revolution has made all previous forms of warfare nearly obsolete.

Why haven't the Russians used the scalar weapons yet?

Bearden gives his thoughts on this in "Fer-de-Lance." The power of these weapons is so great that an accident or war could wipe out all life on earth, and even **affect the sun**.

> "As can be appreciated, this entire concept [MAD] is now totally obsoleted by the Soviet possession of scalar EM weapons.
>
> 'The "mutual assured destruction" capability -- or MAD concept as it was so aptly named! -- is **no longer mutual.** And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.
>
> 'The restraint on the Soviets is not because of our Triad forces. The restraint is twofold:
>
>> (1) a second nation has scalar EM weapons, and would undoubtedly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?);
>>
>> (2) the closed-loop **scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked"** in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply **wipe out all life on Earth in a fiery shower**.
>
> 'The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.
>
> 'The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.
>
> 'Put another way, the pucker factor -- even on the Russians who may contemplate unleashing Armageddon -- is higher than any Western strategic analyst has ever imagined, even **in his wildest nightmares**.
>
> 'Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.
>
> 'The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything."
>
> Bearden http://www.cheniere.org/books/ferdelance/s74.htm

But the Russians (KGB) are not the worst of it. Bearden claims that some of the scalar weapons have been leased to the Yakuza for about a billion a year, and that only the KGB is holding the Yakuza in check from making a **full-scale all out scalar attack on the west.**

U.S. Defense Secretary William Cohen

Warns about eco-terrorism using scalar electromagnetic weapons

> "Others [terrorists] are engaging even in an eco-type of terrorism whereby they can **alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...** So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...**It's real**, and that's the reason why we have to intensifour [counter terrorism] efforts." - Defense Secretary William Cohen, 1997

In a letter to a writer named "Russell" Bearden says,

"In short, Russell, the Secretary of Defense of the United States confirmed that there are indeed novel kinds of EM weapons, right now and have been for some time, which have been and are being used to (1) initiate earthquakes, (2) engineer the weather and climate, and (3) initiate the eruption of volcanoes. We wrote about those exact uses of the weaponry decades ago. Several nations now have such weapons. Three of them (two on one side and the other on a hostile side) are even firing practice shots into Western Australia, as a convenient test range."

http://www.earthchangestv.com/ufo/0209gandor.htm

See Also:
Examples of scalar weapons testing around the globe

See Also: Powers of the Longitudinal Interferometers

Tesla Howitzers in Weather mode

How to Make Weather

"Engineering the weather is **duck soup**; they tested that over the U.S. in 1967, and entered upon continuing operations over North America on July 4, 1976 as a grim kind of KGB "Bicentennial Gift" to the United States. Here's how it is done with several interferometers.

"First, the interferometers can deliberately make "high pressure areas" (cool the air so it shrinks and its "footprint" pressure on the ground increases because its density increases) and "low pressure areas" (heat the air so that expands and its "footprint" pressure on the ground decreases because its density decreases). Well, if one makes the highs and lows where one wishes them, and judiciously and somewhat slowly moves them along a given path, these highs and lows will entrain the jet streams and thus "steer" the weather.

"So if you wish a very cold snap or attack, go up into Canada and start these actions to steer down some streams that bring some very frigid air. Establish other highs and lows judiciously to "block" or "slow" other jet streams and flows as desired. In this way, one can pull the "large cold air masses coming down from Canada" deep into the southern U.S. at will. They do it regularly. If you wish an ice storm, add an additional current of warm moist air you bring up (steer up by using artificial highs and lows) from the ocean. Where they meet, you will get freezing rain, then sleet, then severe icing accumulating. This type of attack does rather large damage to the struck area."

http://www.earthchangestv.com/ufo/0209gandor.htm

Whole websites have now sprung up showing weather **radar anomalies** which may well be indications of scalar weather engineering. One list of over 500 images is here. Cheniere has its own collection here.

Here is an animated anomaly with cloud formation.

In another section at Cheniere are many photos of **cloud anomalies** which reveal the underlying grid along which the clouds are forming or un-forming. Bearden relates a startling incident in these **ongoing "weather wars**."

> "During the spring of 1986, abnormally strong Soviet weather engineering occurred over the U. S., causing a drastic drought in the southeastern U.S. **This drought was broken by a colleague who used an extremely powerful scalar EM device** to redirect jetstreams. A most unusual and unique signature of the "blocking" against the Soviet scalar EM actions resulted: Two huge circulations developed in the atmosphere, clearly showed as two adjacent giant "holes" in the swirling cloud cover over the middle and eastern U.S.
>
> "Between these two giant holes, the cloud circulations formed a stream of clouds, moving to the south, looking very: much like a giant vertical "bar" of a huge "Y-shaped" cloud flow.* Several national weathermen commented on this unusual pattern, which had not been observed before. The pattern continued, day after day, as the blocking continued."
>
> * On the weather maps.
>
> http://www.cheniere.org/books/analysis/history.htm

Tesla Howitzers in Earthquake mode

How to Make an Earthquake

"Here's how you initiate a very large earthquake with such weapons. Take a convenient fault zone of set of them. Focus the interferometry on the fault zone, in the "diverging" mode, and deposit EM energy there in the rocks on both sides, increasing (slowly) the stress in the rocks by the reverse piezoelectric effect (deposit excess energy, get crystal mechanical movements).

"Do it slowly, and the stress will build up to large pressures well-above a plate slip minimum energy required. At some point, the rocks yield and one or both sides "slip" and move rather sharply, giving a very large earthquake in that zone.

"Do the same thing down in the earth (remember, LWs easily penetrate right through the earth and ocean at will, and so the "interference zone" focus can be inside the earth or beneath the ocean, at will.

"Anyway, focus this thing down to where the active part of the volcano is still slumbering, down where the hole in the plate has been made. Keep increasing the deposition of energy in the magma itself, and eventually the increasing pressure from deep within that volcano, underground, will cause an eruption. Build the energy slow, and the eruption will likely be much larger."

http://www.earthchangestv.com/ufo/0209gandor.htm

Yugoslav Earthquake, Followed by Cold Explosion
Strange Characteristics of Iran Quake, Sep. 1978

Tesla Howitzers in Tesla's "Big Eye" mode

How to make a Big Eye in the Sky

By using longitudinal wave interferometry in a weakly endothermic mode a new kind of "radar" is now possible, one which can look right through the earth and oceans. Because the endothermic mode extracts energy from the distant target, a pattern of the target area can be scanned for in the returned energy.

"Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and -- with modern techniques -- the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.

"Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.

"The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a system. One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had little difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible."

Bearden http://www.cheniere.org/books/ferdelance/s42.htm

Tesla Howitzers can destroy all life on earth

How to Make a Storm on the Sun!!

(Just make a big mistake with your Tesla howitzers)

The **deepest dangers** of the use of scalar weapons are frighteningly pointed out by Bearden here where he considers that the sun, earth and moon (actually the whole solar system) is a delicately balanced arrangement of scalar (longitudinal) EM radiation. There is a scalar connection between the earth and the sun, which are both giant dipoles gushing energy in the time domain. Big scalar events on the earth could alter this balance causing **true catastrophe**.

"Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects."

Unrestrained scalar warfare could cause huge solar storms? To realize that **human activity** could actually **affect the sun itself** is truly mind boggling.

Bearden goes on to say:

> "Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the **destruction of the earth itself**. It is not accidental that in 1960 Khrushchev stated that his new fantastic weapon could -- if unrestrainedly used -- destroy all life on earth.
>
> "This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western

> world did not even know what the Russians were talking about."
>
> http://www.cheniere.org/books/ferdelance/s24.htm

2. Tesla domes provide near-perfect defense shield.

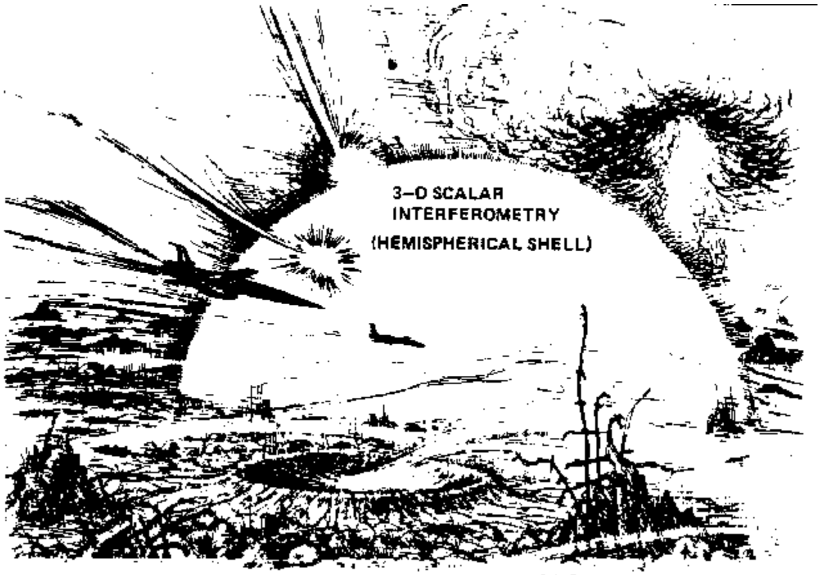

(Slide from Cheniere.org)

The Tesla howitzers can be used in a nearly **impregnable defensive mode** whereby they throw up a dome (**Tesla Dome**) or a sphere (**Tesla Globe**) of highly powerful electromagnetic energy, enough to "dud" or destroy missiles which try to penetrate them. Once again, this energy is not going through space from the howitzer, but being made to emerge from the local vacuum **at the location of the shell**. KGB tests of these domes have been witnessed by airline pilots around the world. Bearden gives many examples of the Tesla domes being sighted around the world in his briefing paper "Fer-de-Lance." (Russian dome test, another globe incident, yet another globe test). These giant electromagnetic domes can be hundreds of miles across, or narrowed down more to total impermeability.

> "Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the **molten core of the earth itself**, as previously explained. In late April/early May of 1985, 27 such "**power taps**" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the **Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons** capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly after launch, detecting and

> destroying long range strategic bombers as soon as they are airborne, etc."
> http://www.cheniere.org/books/ferdelance/s52.htm

By using nested domes one can protect the domed area **even against nuclear radiation itself**. Tesla globes can be used to hit airborne targets by simply placing a globe of any chosen size in the flight path of the incoming missile. It does double duty because the missile hits it going in, and then the rubble hits it again going out the other side.

While the U.S. persists in its useless boondoggle "missile defense system" by shooting a missile with a missile (missiles which would already have been "dudded" by the Tesla howitzers), between the domes and the globes the Russians (KGB) have in hand (**operational**) a nearly **perfect missile defense system**. Someone should ask the Secretary of Defense if the U.S. Tesla domes are in place. And over which cities? How powerful are our American Tesla domes?

Bearden cites a number of examples of these domes being tested around the world: March 20, 1969 | March 24, 1977 | August 17, 1980 | June 17, 1966 | June 18, 1982 June 22, 1976

In a slide from the Fer-de-Lance briefing it is shown how the "woodpecker grid" itself also acts as a sensor, detecting any missile launch anywhere, and how the howitzers can respond instantly by delivering a blast to exactly the right coordinates, right along the channels of that same grid which is feeding the location information about the missile launch. This certainly beats "trying to hit a bullet with a bullet." Especially since YOUR bullet would probably be instantly inoperable in a real scalar war situation.

3. Rogue Groups in the government

Bearden fears that these scalar electromagnetic secrets may have fallen into the hands of what he called secret "**rogue groups**" within the labyrinthine depths of the black-ops special-access domain of the secret government. "In the extreme, such a deep black program can even become a "captured" program," he says, "which is totally in the hands of rogues and **no longer reports to or is bound by the dictates of the parent government**."

> "They have their own agendas. And being rogue groups, they may well bring in **unethical, immoral tricks**: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. **It's a human characteristic, the old primate dominance game. Only now disguised and hidden under deep classification. ...**
>
> "In highly classified groups, these groups are greatly enabled to joust more widely, unethically, and immorally because it is so deeply hidden. So very deep classification evokes the growth and intensity of rogue groups. It's the old 'Power corrupts, and absolute power corrupts absolutely' routine. ... Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, **may not operate under very much legitimate government control at all**, and may operate specifically for the purposes of the rogue group or groups that have gained control."

> "The involvement of at least some rogue groups, some being "**cowboys**" who operate well **outside all laws and ethics**, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, ***it could account for what seems to be actually occurring***. "
>
> http://www.cheniere.org/explore%20articles/mind%20control3/p05.htm

4. World War III has already begun

Because it is now possible to cripple the power grids and electronics of any nation, that nation could be reduced to economic ruin without ever firing an actual "shot." Without electricity Wall Street cannot operate. Without electricity you quickly need martial law. One no longer needs to wage a conventional war anymore at all. There are now **lots of new ways to bring any nation to its knees**.

> "We are at war, whether we like it or not. And we are going to have strategic strikes on an unparalleled scale for us. Pearl Harbor was nothing compared to what is coming down, once we consider the "augmentation" and such.
>
> "When that factor is considered, then it's a totally new ball game on replacing the power system. There isn't going to be much choice; the old one is going to be destroyed or so badly crippled that **it will collapse the U.S. economy** in a way we have not seen since **1929**. Of course, in such a case hitting us, we will be under martial law, the works, everywhere. For one thing, it's the only way you can keep the cities from imploding from within, by the criminal element, looting, robbing, killing, etc. Simply wait and see; it's on the way."

Although Bearden has for years decried the "scalar-gap," and chided the military for falling so far behind the Russian KGB in the field of scalar electromagnetics, recent statements indicate that things may at last have changed somewhat.

> "In the new war, the first phase of WW III has already been completed. Hence President Bush's recognition that this is a total war, and has to be taken to the enemy no matter where he may be hiding or in what sanctuary or in what foreign nation. This is an entirely different war than anything the U.S. has ever fought in its history. **And before it is over, it is going to kill more Americans than have died in all the previous wars in our history. The news media and the public have not yet recognized that awful situation and predicament.** Most are already thinking that, well, we won in Afghanistan so it's just about over. Instead, it has only just begun.
>
> "Anyway, that's a brief uptake on some of the weapon situation today. Some 10 nations or more now have LWIs, and five have the QPs. I'm still working on how many have the appalling negative EMP weapons, but would estimate 6 or 7 nations have them in one stage or the other, either deployed or in at least advanced development. And in my estimation, the Yakuza will have them within three to five years if they have to build them themselves, and quicker if they can buy them from the Russians, Chinese, or???" Bearden

When he was asked the question "**Does the government listen to Tom Bearden?**" Bearden said:

> "Let me speak carefully. I am very satisfied with how I've been listened to in the quarters where the rubber meets the road. **The West is no longer defenseless, not by any means.** Cannot amplify it further than that."
> http://www.cheniere.org/correspondence/021102a.htm

Bearden has briefed the government a number of times and the Cheniere website contains a great number of his graphic military briefing slides.

5. Are the Russians making "UFOs?"

Scalar electromagnetics engineering opens the possibility of antigravity vehicles and "hyperspace travel." In other words we can now make our own UFO's! Of such **scalar-powered antigravity vehicles**, which have now become possible, Bearden has this to say:

> "We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.
>
> "They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. **Actually they would not be moving "through" the atmosphere at all, but through a higher space outside each particle of atmospher**e.
>
> "They could seem to **materialize** and **dematerialize**.
>
> "They could seem to plunge into the ocean or rise out of it.
>
> "They could even seem to operate under the ocean or inside the earth itself.
>
> "**Such anomalous vehicle performances have been seen all over the world**, particularly since a few years after WWII.
>
> "It appears that one or more nations of Earth are operating such vehicles now **in great secrecy**. Our own "government" may even be one of these.
>
> "If so, it **isn't** our true "government/government." Instead, it's the "control group/government." It's government at the operational level, but at higher level **it belongs to certain control groups who have penetrated our government's vitals and taken over all such projects**."
>
> Bearden http://www.cheniere.org/books/ferdelance/slide32.htm

In an excellent article on her Earthfiles website, Linda Moulton Howe reports on the utter bafflement of Norwegian scientists over mysterious lights appearing in the valley of Hessdalen, Norway.

She writes of the research there: "The results can be broken down into two groups: 95% are thermal plasmas and 5% are unidentified solid objects. The plasmas emit long wave radio frequencies and strangely, their temperatures do not vary with change in size or brightness." She quotes the scientists' research summary:

> "1) most of the luminous phenomenon is a thermal plasma;
> 2) the light-balls are not single objects but are constituted of many small components which are vibrating around a common barycenter;
> 3) the light-balls are able to eject smaller light-balls;
> 4) the light-balls change shape all the time;
> 5) the luminosity increase of the light balls is due to the increase of the radiating area. **But the cause, and the physical mechanism with which radiation is emitted, is currently unknown**."

This would almost certainly seem to be another example of the testing of the scalar howitzers. It is baffling because, as Bearden points out repeatedly, Western scientists have limited knowledge of the scalar electromagnetic principles by which the howitzers operate. The Hessdalen scientists were completely **baffled** as to where the energy of these plasma balls **was coming from**.

Ms. Howe interviews Massimo Teodorani, Ph.D., Astrophysicist, who released the report on the 2001 Hessdalen sightings. The scientist speaks of the baffling phenomon observed.

> "I don't know how it is possible that Nature is spontaneously able to do that. Anyway, we deduce that the plasma is trapped inside a sort of magnetic cage and the magnetic cage closes around the plasma and keeps it fixed in some way, prevents it from expanding. **But where does it come from? We don't know**."

By what we have been considering here, we can recognize the signatures of scalar electromagnetic engineering. If it is a Tesla howitzer, then the energy for the plasma balls **is coming from the vacuum of spacetime** at the very **location of the balls themselves**, triggered by scalar interferometers aimed through the woodpecker grid.

These kind of balls can be used as **marker beacons** giving feedback for precision aiming of the howitzers. The energy of the marker beacon can be read back into the computers giving precise location information for pinpoint aiming. The target area can be very small, or widened out.

There is a little movie of one of the Hessdalen plasma balls here. It would seem to be a clear documentation of a scalar howitzer marker-beacon. How many other current "UFO" sightings are actually the witnessing of Tesla howitzer tests?

In fact there are a great many **anomalous events** of the past few decades which might very well be explained by the Russian testing of their longitudinal electromagnetic weapons. It would seem that the mysterious appearance of crop circles around the world could be accomplished by feeding a precise mathematical graphic pattern into the computerized aiming software and change the very molecular structure of the crops themselves along the lines of that pattern. The stalks fall over from the localized effect of the longitudinal waves. A plasma ball tracing patterns? Are they slyly sending a message to see if we are "getting it" yet?

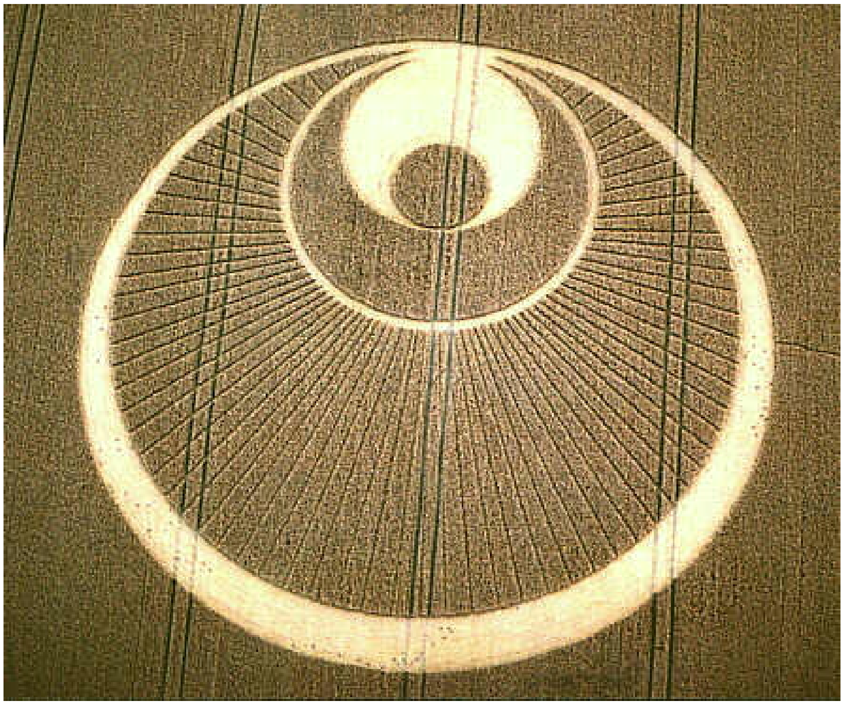

The "Cambridge Angel" - Made by Russian scalar plasma balls?

Also see: Interview with Tom Bearden

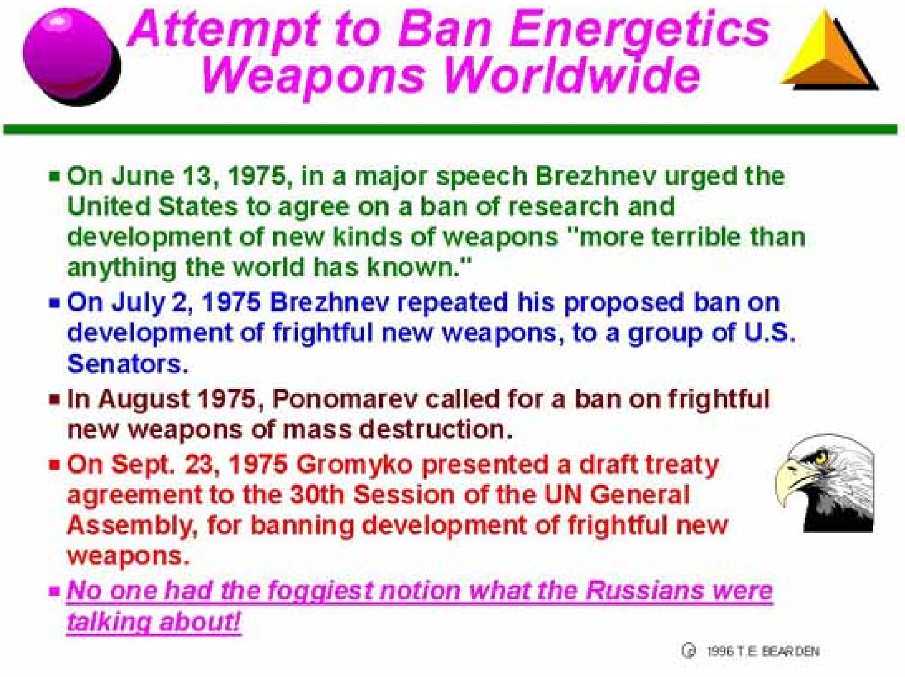

Attempt to Ban Energetics Weapons Worldwide

- On June 13, 1975, in a major speech Brezhnev urged the United States to agree on a ban of research and development of new kinds of weapons "more terrible than anything the world has known."
- On July 2, 1975 Brezhnev repeated his proposed ban on development of frightful new weapons, to a group of U.S. Senators.
- In August 1975, Ponomarev called for a ban on frightful new weapons of mass destruction.
- On Sept. 23, 1975 Gromyko presented a draft treaty agreement to the 30th Session of the UN General Assembly, for banning development of frightful new weapons.
- *No one had the foggiest notion what the Russians were talking about!*

© 1996 T.E. BEARDEN

(Slide from Cheniere.org)

---

4. Healing powers of longitudinal waves

Healing via longitudinal waves, time reversal of cells (de-differentiation), Priore's machine, Bearden's proposed "blanket" version, using the electric power grid for healing waves or for broadcasting disease patterns. Scalar Biowar.

The good news!

"**Any sort of disease whatsoever - physical or mental** - will be rapidly curable, simply and cheaply. It will be possible to reverse aging and rejuvenate the person. It will be possible to regrow limbs and straighten misshapen spines - and directly remove the causes of mental diseases and cure them as well. Lifespan will be staggeringly increased, without 'old age's debilitation.' AIDS, cancer, leukemia, and genetic diseases will be completely conquered. The potential for a true golden era is upon us, for all humanity."

Bearden- http://twm.co.nz/beard_interview.htm

Proof of the healing powers of longitudinal waves from the vacuum is seen in the work of Antoine Priore who produced longitudinal scalar waves using a **gigantic glass tube** filled with plasma. He himself did not understand why it worked or how it cured cancer in animals.

> "The Priore team cured terminal tumors, trypanosomias, and other dread diseases in laboratory animals in France, in the 1960s and 1970s, under rigorous scientific protocols by eminent scientists. The Priore

> project was funded by the French Government to the tune of several millions of dollars (equivalent). http://www.explorepub.com/articles/energetics.html
>
> ". . . Prioré's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960." http://www.cheniere.org/priore/index.html

Bearden has stated that the longitudinal EM waves can have a healing effect because they always come in pairs called "**phase conjugate pairs**" and one of the pair is **time-reversed**. This enables engineering of the waves to "pump" the cells of the body (with waves from the time domain) and actually **time-reverse the cells** back to a previous healthy condition. The scalar cure of cancer does not involve killing the cancer cells but rather pumping them with time energy into a previous condition. I know, this seems miraculous, and is it not?

Priore and his giant plasma tube

This healing power in the longitudinal waves of the time domain is a mind-boggling thing to face. It took me months to get through it, and to **suffer the fact** that the **cure** for cancer and AIDS was right there, **in this new scalar electromagnetics technology**, and that it has been kept back from the people for so many decades now.

Bearden has a great deal to say about the work of Antoine Prioré, who himself did not understand why his machine worked. Priore did not know about the longitudinal scalar waves of the time domain. In fact, Tom Bearden has supplied, at last, the missing explanation as to how that **big giant two-story glass tube** consistently cured cancer in animals.

> "**The medical implications are enormous**. By just reversing the damaged cells back to a previous physical state, this gives a physics mechanism for *controlled cellular dedifferentiation*, in biology terms. Physicians will **time-reverse diseased, damaged or aged cells** back to an earlier healthy condition, including all the cellular genetics. Normal cells are just reversed to a slightly younger condition."
>
> "We believe this mechanism explains the revolutionary but baffling electromagnetic cures achieved by a team of outstanding French scientists in the 1960s and early 70s, working with inventor **Antoine Prioré**. In lab animals, cures were achieved for terminal tumors, infectious diseases such as trypanosomiasis, and atheriosclerosis. In addition, deliberately suppressed immune systems were dramatically restored.

> "We hypothesize that this is the long-sought mechanism of how the living body heals itself. In short, the body's cellular control system induces **cellular reversal by weak time-domain pumping** of the damaged cells and the affected area. The cellular regeneration system performs all healing of cellular damage in the body; the immune system heals nothing, not even its own damaged cells) .
>
> "For the first time medical science can develop as a science of *healing* rather than a science of *intervention*. And it will develop as a science of *unlimited* healing, since **no pathogen can resist action by curved spacetime engines**." Bearden

http://www.cheniere.org/techpapers/Vision%202000%20paper.doc

See Priore's machine See Slides of how it works.

Bearden has extended the Priore work, with the previously missing knowledge now in place, and designed a revised system (for irradiating the body with longitudinal waves) which gets rid of Priore's giant glass plasma tube and uses instead an **antenna unit which surrounds the body**. He has also provided the missing explanation of how it works. There is an article from "Explore" magazine, Part I and Part II.

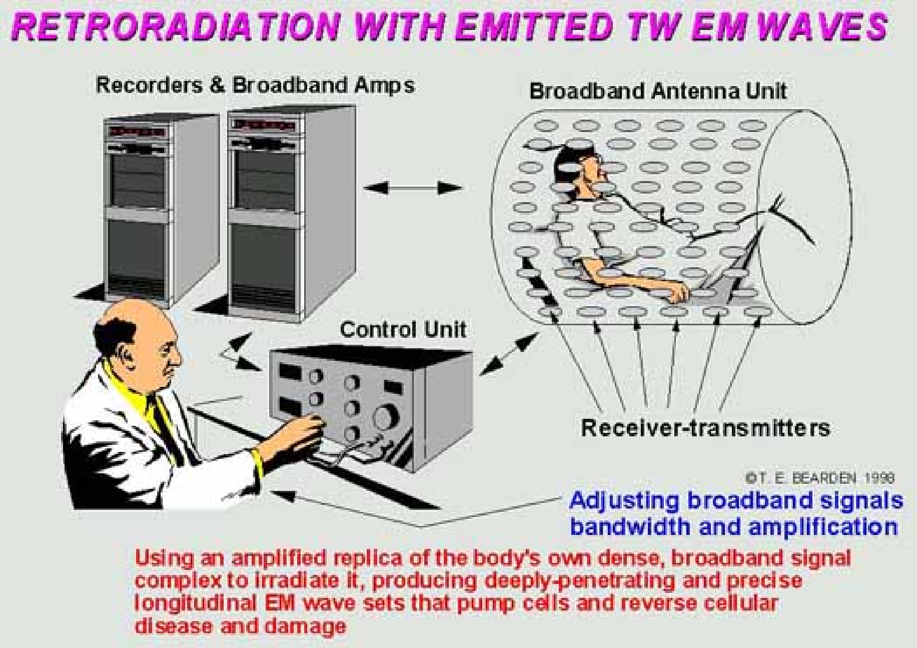

RETRORADIATION WITH EMITTED TW EM WAVES

Recorders & Broadband Amps
Broadband Antenna Unit
Control Unit
Receiver-transmitters
©T. E. BEARDEN 1998
Adjusting broadband signals bandwidth and amplification

Using an amplified replica of the body's own dense, broadband signal complex to irradiate it, producing deeply-penetrating and precise longitudinal EM wave sets that pump cells and reverse cellular disease and damage

(Slide from Cheniere.org)

In fact, Bearden has presented to the U.S. government a plan to use a **portable version** of a **Scalar Healing Blanket** where wire channels in the blanket act as a scalar antenna, which can both send and receive scalar electromagnetic waves. A **two minute treatment** of the waves is sufficient to restore the body to its earlier healthy condition. Thus even **thousands could be treated easily and quickly in a biowar mass attack** situation.

> "To defend against things like anthrax attacks on our civilian population centers, we also have recommended to the U.S.

> Government the crash development of small, portable treatment machines—using similar "antiengine-forming" and cellular reversing "porthole technology". [See Figure 13 on previous page]. Such portable machines **could be developed cheaply and quickly**, and present government facilities could be used to massively attack the correlate database production problem.
>
> "The portable treatment **units could be cranked out by the hundreds of thousands** and flooded down through emergency response agencies such as the police forces, the National Guard, emergency hospitals, emergency response teams, etc." - Bearden

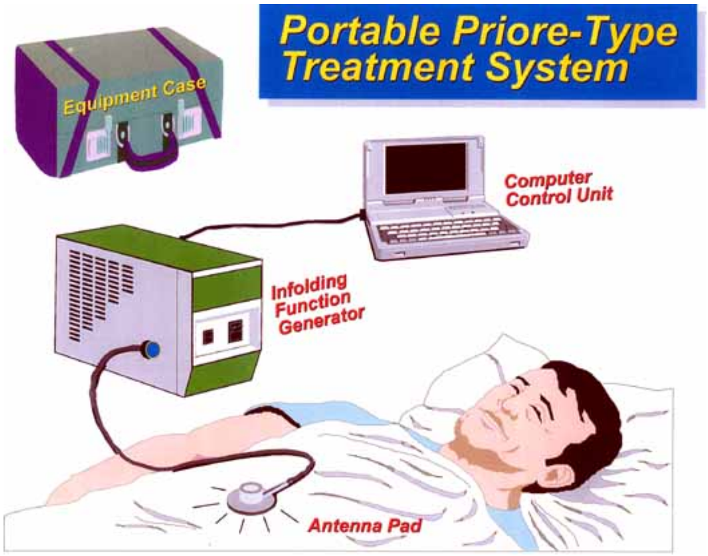
Portable Priore-Type Treatment System
Equipment Case
Computer Control Unit
Infolding Function Generator
Antenna Pad

(Slide from Cheniere.org)

The **Scalar Healing Blanket** would fit into large suitcase. It has three basic parts, the longitudinal wave generator, a laptop computer, and the antenna-blanket. The plan was a stroke of Bearden's genius, but unfortunately it fell on **befuddled government ears**. "Unfortunately, so far the DOD [Dept. of Defense] doesn't even understand what the heck we're talking about!"

Article "Healing with Energy": http://www.explorepub.com/articles/energetics.html

> "We tried very hard in 1998 to call attention to the extreme national need for the most urgent possible development of a portable computer-controlled treatment device based on the experimentally proven Prioré methodology, but extended to a much simpler and far more rapid method of treatment.
>
> "Quick development of this technology could lead to **a suitcase size device, capable of treating a patient in less than one minute, and stopping the symptoms and disease progress cold, then reversing it**. Three "less than a minute" treatments one week apart would be required for complete cure. As an added advantage, partial reversal of aging in older patients would also be accomplished as a bonus, in addition to time-reversing the damaged and diseased cells...
>
> "The method proposed to the DoD in 1998 used normal EM in a peculiar way to force the body itself to make its own time-polarized

> EM waves, and pump every part of its cells -- including the genetics -- in the time domain.
>
> "Quite simply, we are going to be struck this way (QP spreading the immune systems of the populace plus professional smallpox, anthrax, etc. BW attacks on our population centers) and by other means we've spoken of in other papers and briefings."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

The Dark Side of the Force:

Creating and broadcasting disease

Speaking of the scalar wave induction of disease by longitudinal wave patterns Bearden refers to work by French scientist Kervran and Russian scientist Kaznacheyev, who proved that any disease could be transmitted electromagnetically.

> "It works this way: the Kaznacheyev experiments (and others) demonstrated that any cellular disease or disorder can be initiated into cells at a distance, by electromagnetic means (albeit unusual EM in nature). The decades of microwave radiation of the U.S. Embassy in Moscow, was an ongoing set of "stimuli" to ping the U.S. system and, by its response, ascertain how far along (or if it was "along") in its knowledge of such weapons. The radiation was responsible for many health changes in personnel and for the **eventual deaths of three U.S. Ambassadors**." Bearden http://www.cheniere.org/correspondence/021702a.htm

In a letter he writes,

"Yes, the Kaznacheyev experiments were quite real; 15,000 or so in military research institutes in Siberia. As with so many other novel areas investigated and developed by the Soviets, most U.S. investigators just assumed normal EM, which is not what the Kaznacheyev experiments involved, and not what was carrying disease induction patterns from one cell culture to another in rigorous lab tests."

How to "Broadcast" Diseases

Longitudinal EM Biowars

Bearden explains how the new quantum potential weapons can be used to induce disease-at-a-distance in a population, or to "spread" the immune system so thin that a **conventional bio-attack would be greatly potentiated**. By "broadcasting" virtual disease patterns over a population, the immune system is so overloaded by the presence of the **"shadow" diseases** that its effectiveness is thinned out.

> "In short, alter the internal wavestructure, and one creates a curved spacetime 'engine' that acts on mass in any fashion one chooses to design -- including initiate diseases . . . Now visualize one of these '**vacuum engines**' or 'spacetime curvature engines' that **acts on mass** to **generate the exact effects produced by anthrax**. (Or any other disease one wishes) . . .

Call the virtual state of a disease engine the disease pattern in the "shadow" state, just prior to becoming observable. . . .

"And the master cellular control system responds to "coming events that cast their shadows before". In short, it responds to the upper level of the **"shadow state" of a disease**, still in the virtual state! (Hey, virtual particles are real, virtual photons are real, virtual ST curvatures are real --- just very fleeting, only for another to immediately arise). The exchange of virtual particles is known to generate all forces in physics anyway! . . .

"So the cellular control system responds to "shadow state" disease patterns. Enter a QP weapon application. Place a quantum potential weapon "area" on the United States, so that the entire populace is in it. The slowly introduce and bring up in the shadow state, the **necrotizing fasciitis disease engine**. At some point, the cellular control systems will react, and order the immune system into action. In other words, immune system resources will be committed against this "immediately coming" enemy. Bring it on up a little, but still just in the shadow state. Voila! Humans form a bell-shaped distribution curve. And even if the structure of the QP fluctuates a little, fluctuating a bit stronger and a bit weaker, etc. **A few of those targeted bodies will now actually develop necrotizing fasciitis**. So far, just like the U.S. Embassy targeting, except a different "carrier".

"But now there is an exact signature that this is what is doing it, and not normal vector carriers of disease and contact or exposure infection. **The few cases of necrotizing fasciitis that break out will be most puzzling.** They will be randomly scattered across the entire nation, and so few that the vector carrier and contact or exposure explanation completely fails."

"And that one happened, right here in the U.S., not so long ago. Check it out.

"So carefully adjust the **patterns** for, say, 12 major terrible diseases, to that "adjusted and desired shadow state level". Now add all 12 patterns into the internal structure of the quantum potential. Now it has **12 shadow state disease engines** in it.

"**Place that one on a populace**. The cellular control systems order the immune systems into action, and they now have to commit their finite resources to 12 different areas. This greatly thins the response an immune system can make to any one of those 12 disease patterns! In short, it "**spreads**" the capabilities of the immune system, which can be galvanized to great effort, but only can perform a small bit against each area.

"Suppose one of those "galvanized" areas in the targeted populace's immune systems is anthrax. And suppose the bad guys now hit one or more of those cities with a professional anthrax attack. A former OTA study showed that a professional anthrax aerial spray attack against Washington D.C., dispensing

> 100 kilograms of anthrax spores in the spray, will account for 1 to 3 million casualties. And that's in an "unthinned" set of immune systems. **In the "thinned" set, one can at least double or triple, and even up to five times, that estimate.** So conservatively, now there would be from 2 to 6 million casualties. And even mild anthrax spray attacks in other "thinned" cities would also be amplified.
>
> "**That one is being set up, or already set up, right now.** There are other such "augmentation" setups also under way.
>
> **"So the weapons alluded to by the SecDef in 1997 are already being employed. WW III has already begun, and our populace (and our scientific community) is blissfully unaware of it."**
>
> http://www.cheniere.org/correspondence/021702a.htm

Most of us can remember the strange outbreaks of the so-called "**flesh eating disease**." It seemed to break out here and there randomly with no known cause. Bearden believes this was a test of the newer quantum potential weapons which go even beyond the capabilities of the Tesla howitzers.

> "The Kaznacheyev "disease induction by novel electrodynamics" work is also the basis for the spacetime curvature engines (not EM signals!) carried by quantum potential EM biological weapons for distant induction of diseases etc. into the population of an entire nation or area, such as the U.S.
>
> "This **quantum potential disease induction weapon** -- capable of attacking an entire population of a nation or area -- has been tested several times in the U.S. at very low levels, in the shadow state only (we explain shortly) and it has a most peculiar signature. E.g., **one test used necrotizing fasciitis (flesh-rotting disease) engines**.
>
> "The disease 'engine' was deliberately kept down in the 'shadow state' (virtual state, just below the observable state) so that only a tiny bit of the populace with depressed thresholds would "breach the observable threshold' due to the population forming a bell-shaped curve as to the threshold level. **The major signature of the test was that the cases of the disease broke out totally statistically and randoml**y, spread all over, without any 'disease vector' in between.
>
> "In short, it was not spread in any manner required by normal deterministic disease vectoring, **but was a totally random set of occurrences**. Several other diseases were also tested this way in the American populace, with precisely the same signature."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm
>
> "Apparently one planned use of such an augmented clandestine BW weapon will be smallpox, e.g. As you know, the Russians

> did develop smallpox for biological warfare, and even developed a new strain against which previous vaccination is ineffective.
>
> "The vast numbers of expected casualties, coupled with immune system spreading, are absolutely irresistible to the Russian mind. You can count on it that they have seen that the smallpox is already on site here, in the U.S. and in the hands of professional terrorist teams. Of course, the Russians/KGB will protest they have nothing to do with that! Yet likely the KGB will have flatly arranged it or at least assisted it. The capability for mass destruction of the nation is so lucrative and easy to achieve in advance, that **they would never be able to resist doing it**."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

Bearden believes the **Gulf War Syndrome** involved the use of scalar technology to induce a disease state, as he stated in a 1997 interview:

> **"I'll say this flatly, and I won't answer any further questions about it; the Gulf War Syndrome was induced...** It was induced with this stuff. All the rest... was contributing factors that everybody's talking about... There are some deeper signatures, if you look into it very deeply, that show you exactly how it was done and the fact that it was induced. It was a test. It was a test of a very special kind of weapon I have not talked about yet. Anyway you can use it to create diseases."
>
> Bearden http://twm.co.nz/beard_weapon.htm

Besides using the woodpecker grid to target a population, the entire **electrical grids** of nations can be used as "channels" for disease patterns to be broadcast to all people near that grid. Your own house wiring can become an antenna, broadcasting these "shadow" patterns of disease, weakening your immune system.

And here one faces the frustrating fact that the same power grids could be set to channel patterns which would keep the entire population healthy, rejuvenated, continuously bathing in the time-reversing old-age-defying energy of the spacetime vacuum. We have **no idea yet** of how long the human life span could be increased with this new technology. People might be able to live two centuries given full humane development of this technology.

5. Psychoenergetics

Mental entrainment, inducing hypnogogic trance, mind control, broadcasting thoughts and feelings, the Ant World, education via EM [LW] mental download, the "Feelies," revolution in spirituality, "enlightenment machine?" Scalar churches?

Psychoenergetics is simply using the scalar interferometers to manipulate and **engineer the human psyche**. To me it is the most **mind-boggling and frightening** aspect of longitudinal wave engineering.

The mind is electromagnetic in nature, and itself of the time-domain. It is not observable in 3-space. There is no **thing** you can point to that is the mind. As Bearden says, "**The mind is time-like**."

The longitudinal scalar EM waves are also of the time-domain. By overlaying certain infolded patterns (oscillations in time) on the longitudinal waves, **a thought or feeling may be made to arise in a person** or persons who are in the interference zone (where the wave beams cross).

The person will not notice anything, feeling that this thought or idea is his or her own idea. The longitudinal wave pattern might be a wave of panic or fear that spontaneously sweeps over you, perhaps inexplicably. It could even be in the form of a sudden intense patriotism. It could be explicit words and phrases, which everyone would think they had thought of by themselves. Or it could be rage and hatred, angry and violent. On the other hand It could also be a continuous sense of docility and placidity. This is the brave new world of psychoenergetics and the coming "**Mind Wars**."

Scalar Psychoenergetics in its most **primitive form** simply "entrains" all minds in the target area into a **deep hypnogogic trance**. In this state of mind people would be suddenly extremely suggestible, and would likely believe anything they are told, and would obey any orders given. Talk about winning the hearts and minds of the American People!

Tom Bearden makes me think of **Paul Revere**, who heroically conveyed important **warnings** for the sake of the new nation. But the words of the famous phrase have now strangely and curiously mutated into the once-inconceivable: "**One if by land, Two if by mind.**"

Indeed one strategy in a **Mind War** would be to simply **take over the minds of the "leaders"** of the enemy nation. The targeted leader would not be aware that anything was amiss, although he might begin to make unexpected changes in policy.

And are, in fact, any of our "leaders" already having secret thoughts that are not their own? Who knows? Frighteningly, we can no longer be sure. The new sciences of scalar electromagnetics and psychoenergetics are even now plunging humanity into a sudden science-fiction-like world beyond anything in humanity's previous imaginings.

> "The Russians reached the point in the mid 1990s that they could **take over control of a person's mind**, with modified longitudinal EM waves including some time-polarized EM waves, and with a team of specialists (estimate 25-30 per transmitter, and one transmitter and team per controlled person)." Bearden

In a slide Bearden illustrates how a Scalar War would involve a **psychoenergetics attack** on the operators of the enemy scalar installation, entraining their minds into hypnogogic trance and **getting them to shut down their systems**.

And ultimately psychoenergetic warfare goes to the very heart of **human identity itself**. For if my thoughts might no longer be "mine," then **who and what am I?** Can my very sense of being "me" be hijacked by some nefarious psychoenergetic scheme? Will the "secret government" eventually dictate directly into your mind how you feel about yourself? Or what you think you are? Or what you should do?

How Does Longitudinal EM Mind Control Work?

The physics of scalar psychoenergetics as expounded by Col. Bearden seems even more arcane and over-my-head than that of extracting energy from the vacuum. I can't understand it, but I look through it anyway. Each time another tiny understanding might occur, or a term might fall into place. I know the mathematics is forever beyond me.

One paper to begin with is Mind Control and EM Wave Polarization Transductions This is such serious stuff that Bearden includes a **strong warning about misuse of this knowledge**. Psychoenergetics weapons can mentally maim and physically kill.

Warning!

(From "Mind Control and EM Wave Polarization Transductions")

> "This article refers to experimental research techniques which can be detrimental or lethal in the hands of any but highly skilled, qualified experimental scientists proceeding under proper laboratory safety procedures. The purpose of this article is strictly for information to properly qualified and authorized scientists in certified laboratories. We do not propose or condone any use of these procedures for nonapproved practice of medicine without a license. Neither the publisher nor the author are responsible for accidents or outcomes in the use of these experimental procedures and techniques. Any researcher who performs these procedures and experiments is acting on his or her own volition, and is solely responsible for insuring safety, qualifications, and legality of the acts and their results. **We neither suggest nor condone unauthorized experimentation on human subjects. Such is a criminal violation of the constitutional rights of the subject under Federal and State laws, and is both illegal and immoral**." Bearden

After the **serious warning** comes the description of the paper. This is heady stuff for the layman.

> "Calling full attention to the special note above, in this paper we present a high-level overview of the novel electromagnetic nature of mind operations, mind and body coupling, and intent - the induction of physical 3-space EM energy changes into the brain and nervous system, and into every cell of the body, from the mind's time-like coherent operations. We summarize the time-polarized electrodynamics used to engineer and affect mind operations and the mind-body coupling loop. Transduction mechanisms whereby differing EM wave polarizations can be transformed one-into-the-other are presented." Bearden

I will leave it to the technically astute to try to understand the mechanisms of operation which are described in scientific language. But there are tidbits for the layman. Here he speaks of the difficulty Western scientists have in opening up to the "immaterial" (longitudinal wave) nature of the human mind.

Western Science Remains Largely Materialistic

"Ironically, most Western scientists are materialists and consider "mind" as a mystical and nonscientific concept. They tend to consider mind operations and functions either to be simply "meat computer" operations and functions, or at best to be very weak ordinary transverse-wave EM operations and functions in the brain and nervous system. This serious self-limitation exists because in the body we measure only weak TW [transverse wave] EM operations and functions correlated to biological behavior and brain operations. We simply do not know how to measure "mind operations" directly.

"With no mind measurements possible and no instruments, it is understandable that Western science considers only the physical side of the mind-matter interface.

"Presently our scientists do not measure the longitudinally-polarized EM wave operations and functions in the body and around it in nature. Few of them are aware that a maelstrom of such LW [longitudinal wave] functions exist in the body and in all of nature in general." Bearden

Western methods of influencing the mind with EM waves have only used transverse waves, not the longitudinal waves of the vacuum. If only transverse waves are used one pretty well has to hit the target mind with a **sledgehammer of waves**.

Western Science knows only transverse EM waves

Brute Forcing Time Functions versus Fine Control Methods

"Ironically, Western mind control researchers using transverse EM waves for mind control research, are using a brute force method of evoking and using vacuum engines (spacetime curvature engines) and a special form of general relativity, although they do not appear to realize it.

"While KGB scientists also use TW EM "brute force" TW waves when necessary, they do **"imprint" or "activate" those waves with the desired internal LW and time-polarized EM wave and photon structures** required to directly perform the mind engineering desired... Bearden

"These "fitted brute-force models" certainly can be very powerful, and certainly can produce the exact results shown in the experimental verifications of the fittings. However, they do not of themselves allow sophisticated design for example of the necessary time-polarized wave assemblies for **engineering the entire human collective unconscious simultaneously**, or for engineering the entire collective unconscious of all species on Earth (i.e., Gaia's collective unconscious), or even for precisely **engineering the memory and knowledge base of an individual**. Bearden

The seemingly magical operations of psychoenergetics, sometimes called "psychotronics," cannot be achieved with ordinary transverse EM wave energy. But they are possible with longitudinal scalar waves because the mind itself is scalar in nature.

> "We note that **all mind operations are time-like**, i.e., they are comprised as scalar EM photon functions and scalar EM wave functions. Thus the mind is a very special kind of electromagnetic system, existing in the time domain..." Bearden

The engineering of the mind can only be accomplished with the waves of time.

The course of psychoenergetics development necessitates the building of a vast database of settings and patterns which bring about given mental and/or emotional states. A kind of "**genome**" of the psyche. Bearden describes the process of the decoding of this "genome."

> "Now the scientists would perform many phenomenology experiments, making one little change at a time and profusely recording the data. Each time, they would establish the physical change(s) that occur in the body and/or the mental and emotional changes that occur in the mind for each spectral reinsertion back through the "ship's portholes". They would simply but painstakingly (over some years) build up an extensive database of those individual correlates.
>
> "In these experiments, the experimenters will eventually be able to **provoke any body or mind change they wish**. Strong emotion. Intense pain. Intense pleasure. Painful thoughts. Images. Memories. Perceptions. Dreams. Visions. Memory losses. Memory changes. Personality changes. Etc. The "delta" in the emission spectrum (the changes from zero reference spectrum) represent the precise totality of all mental, physical, organic, chemical, etc.changes and interactions." Bearden
>
> "A second stage in the research would be to **test the correlates** and porthole insertions upon a statistically significant cross section of ordinary people, and/or specialized populations (such as toughened soldiers). The program would evolve a highly complex, very effective, ever-improving science and technology of mind and behavior control and engineering mechanisms, complete with finished database and developed applications equipment." Bearden

Summing up his paper Bearden speculates on what the actual situation the science of psychoenergetics probably is. The intense secrecy with which this knowledge has been kept from the public has led to the **current farce of the "government" pretending** it all does not exist, and this secrecy may have now put us at great risk.

> "The implication is that in the West one or more highly classified, sustained, heavily funded developments in advanced mind control programs, probably exists and probably has existed for some time. Due to loose formation of rogue groups inside such programs, they may have dual or triple purposes, **may not operate under very much legitimate government control at all**, and may operate specifically for the purposes of the rogue group or groups that have gained control.

> "In a nutshell, that's what may be going on in the **clandestine mind control projects** in several Western governments. The involvement of at least some rogue groups, some being "cowboys" who operate well **outside all laws and ethics**, could also result in such things as assassinations, clandestine testing on individuals without their consent, etc. In short, ***it could account for what seems to be actually occurring***. As also is "usual" in such a mess, one or more of the rogue groups eventually may become very powerful because their secret weapons are very powerful. They may become confident, thinking they have the "best in the world." They may actually believe they are ahead of the Russians..."
>
> "Remember that, in every large and powerful human organization, the basis for rogue groups is **power** and **secrecy**. They are always seeking to increase their power, control, influence, prestige, etc. Nothing else. Patriotism and mission are—to rogue groups—often just idle words. They have their own agendas. And being rogue groups, they may well bring in unethical, immoral tricks: assassination, bribery, entrapment, disinformation, plausible deniability, etc. A certain percentage of a highly secret rogue group will wind up using all these things and more. It's a **human** characteristic, the **old primate dominance game**. Only now disguised and hidden under deep classification."
>
> "The **Russians**, with their additional knowledge of the actual mechanisms in the transforms, will be **much more advanced than the West,** because their fundamental psychoenergetics science is far more advanced, so long as we continue to use the old U(l) electrodynamics. Further, the Russians have decades of use of **longitudinal interferometry beams to reach right through the earth and ocean** and produce stringent EM effects at a distance. So they will also be able to do the same things here in "mind control" with LW [longitudinal wave] interferometers, through intervening mass." Bearden

Bearden cites two examples where he believes the Russian KGB tested the psychoenergetics capabilities of their scalar interferometers. One was the case of Captain Button in his A-10 "Warthog" aircraft on April 2, 1997.

> "Over Arizona, Captain Button was thrown into a hypnogogic state, and his perceptions instantly altered and controlled. In his instant "dream-waking" state, everything seemed perfectly normal. His sense of direction was altered a bit more than 90 degrees, so he simply corrected and turned and "flew toward the range", actually flying off course by more than 90 degrees and ignoring radio contacts. He flew right on out of Arizona."
>
> "At one point he circled, probably thinking he was over the range, and he probably dropped his ordnance there. Then he "flew on back toward home," as he thought in his waking dream state, until his fuel ran out and he crashed and died in the explosion—**dream-thinking until he died that everything was normal**. All the while, his sense of the passage of time was altered. To him, in his dream-thinking, dream-acting state, everything was normal and nothing untoward had happened. **So the distant KGB transmitters and associated psychoenergetics team controlled him for over an hour**.

> "Indeed, that was the exact purpose of the test: demonstrate control (at a great distance) of a skilled person performing highly skilled tasks."

The paper Mind Control and EM Wave Polarization Transductions ends with the most chilling and frightening possibility of all: the **psychoenergetic engineering of the human species** as a whole.

The Ant World

The Final Race for Direct Mind Control of the Entire Human Species

"The KGB psychoenergetics weapons scientists—because of their direct measurements and detections (and use) of t-polarized waves and LWs [longitudinal waves] for decades—also understand that Jung's collective unconscious mind (of the entire human species) also has its own time-like operations and correlates, which one measures also when using t-polarization measurements and sorting it all out. **The collective unconscious mind operations are buried several levels deeper inside the recursive Whittaker structuring inside the EM fields, waves, and potentials**.

"The KGB scientists also know that something very like **Gaia** — a collective unconscious mind for all species on earth — **also exists**, and they are striving to be able to sort out and measure that one as well. The Gaia operations are buried even deeper in the recursive structuring inside the EM fields, waves, and potentials."

"Actually, from day one, the Russian mind control scientists have had their eventual goal set upon this "deeper area of mind and possible mind control" of the entire human species.

"I am convinced that the KGB psychoenergetics scientists understand this deeper area. If they can learn to directly engineer the collective human species unconscious, they can then **convert the human species easily into a sort of "ant" society**, modeled along ideal Communism lines, except of course with an excluded "hierarchy at the top" running everything. The "ant society" — communist style — is still a variation of feudalism (**all systems tend to feudalism or some form thereof**).

"The KGB psychoenergetics weapon scientists are seriously pressing on toward that very goal. In my opinion they are not very far from it right now...

"Some versions of Russian mind control devices were used on Russian soldiers in the Afghan War to **condition them for performance of merciless acts**." [13]

http://www.cheniere.org/explore%20articles/mind%20control3/p07.htm

More about "The Ant World"

Whether it is the Russians planning an "ant world" society or rogue U.S. black-ops groups withholding a great boon from human race, the secrecy on all sides around has allowed **largely criminal elements** to take over this technology while leaving the civilized elements in the dark. Openness and the spreading of this knowledge among good and decent people is essential, and for that **we will have to overcome our frightened denial**.

> "Meanwhile, **rogue groups** amongst Western clandestine mind control researchers will probably arise if they have not already done so. They will likely seek to increase their personal control and further isolate the programs from orthodox government review and from government and legislative control. They may even divert the research into highly illegal and unethical means, because it furthers their own rogue agendas. That is how clandestine U.S. government research can sometimes go sour, **unless great care is exercised by the oversight committees in the House and the Senate**.
>
> "Sometimes when rogue groups do gain control and total secrecy of a given new technological area, then what appears to be "U.S. government operations" do start to encompass a criminal and unethical operations, hidden usually beneath the deep veil of high classification. Also, if it's "scientific," no one is ever brought to justice, even if the "evil science actions" are uncovered and publicly revealed." Bearden

Like the huge energy of the Tesla howitzers, the ability to engineer the mind with longitudinal waveforms is a seeming "magical" thing which can be used for good or ill. Bearden envisions that further developments will allow something like the "downloading" of knowledge via scalar interferometry. Everyone could be highly educated, and quickly and easily.

In the conclusion of the paper "Mind Control and EM Wave Polarization Transductions" Bearden makes **a plea to the nations of the world** regarding the use of psychoenergetic engineering.

> "Let us hope this great new area, already off to a bad start, can be bridled and steered in the direction helping and healing people, rather than killing or abusing them. **The excesses in its bad use are a potent threat to all nations on earth**.
>
> "Yet it can revolutionize medical science, education, communication, and psychology. We foresee the day -- perhaps 30 years hence -- when **education will be accomplished by directly loading the software into the mind**. Then in three weeks one will "load" a doctorate, say, in physics. In three more weeks on will also load a doctorate in chemistry. Another three weeks, in electrical engineering. Another three weeks, an M.D. And so on. When that happens then truly everyone on earth can be educated. Freedom from ignorance may well be another great freedom that is legally recognized. There will be no impoverished large groups lacking the education to find decent, productive jobs.
>
> "We urge all nations to use the principles involved: not for human abuse, but for healing, educating, uplifting, and life-expanding of every person on earth. If we do, we shall all have a far brighter future. Then we shall check what has started out to be the Sword of Damocles and turn it into the golden Millennium." Bearden

(Perhaps after so many frightening topics a slight break is in order here to consider the **lighter side of psychoenergetics**. For example, what will be the ultimate impact on the entertainment industry?)

The Feelies at Last!

Psychoenergetics will be able to bring about at last, for better or for worse, Aldous Huxley's fanciful idea in his novel "**Brave New World**," an entertainment format which he called "**The Feelies**." Those were something you went to, like going to the movies, but you would **experience** all the actual feelings themselves, as if the depicted events were actually happening to you. Although our "leaders" seem too dim to grasp what is going on in the world of scalar electromagnetics, it should not take some young Hollywood genius too long to understand that **The Feelies are now possible**, and that the immense megabucks of potential profit more than justify heavy investment in bringing it to market.

One can imagine that first "Feelie" theater, "The Huxley" perhaps, a large comfortable area surrounded by nine giant screens (not just one screen!), which become 3D if you put on your headset. And The Huxley is fitted with the finest sound, and oh, by the way, a modest sized scalar interferometer with computer. And there are one or more **new tracks** now recorded on the film next to the sound track. These are the digital patterns of thoughts, feelings, and ideas to be fed into the theater's interferometer. And these thoughts and feelings and ideas will be felt by the audience as if they were their own thoughts and ideas. It is an entirely new art form. Psychevision.

As important (and no doubt famous) as the actors would be the "feelers," those from whose minds the feelings were lifted in the original programming of the Feelie tracks. Because, for example, Brad Pitt might be a good and handsome actor, but **can he really feel it?** Like the dubbing of foreign language films, it might be necessary to dub the "emotional track" using a deeper and more profound person's feelings, thoughts and ideas.

It might even be engineered so that if you sat on one side of the theater you would feel like the hero, and on the other side you would feel exactly like the bad guy! What would that lead to?

We have all cried at a movie some time, most people have, though they might not want to admit it. But this would be **deep sobbing** and a **flood of tears**, if the director so desired it. You would feel it at your core. Or, take a big chase scene, **you would be terrified**. They are breathing down your neck! You would be gripped by panic.

For that matter, the whole genre of **Horror-Feelie** would be too much for many people to take. Do I **really** want to feel that an **actual vampire** is tearing at my throat with his fangs?

A Religious Feelie would make you feel the exaltation of a saint at his prayers, or give you the **definite** feeling that you are Christ suffering on the cross. All the pain could be there too. The stickiness of the blood. The very feeling itself: "Why hast thou forsaken me?" Who among us could come up **that** despair for scanning? Had this been developed a bit earlier we might watch a documentary of, say, Mother Theresa, containing her own actual deep sense of compassion, which we would feel arising in us as our own.

Then again, a Psychedelic Feelie could induce an LSD-like state and simply blow the audience's minds.

Even in its simplest most primitive mode, the hypnogogic trance induction with its increased suggestibility, would completely alter the experience of even ordinary current movie going.

6. As it Stands - 2002

[Note 2006: Click here for a current status history of the MEG from Tom Bearden]

> "**Anyway, it's presently a "Mexican stand-off**" with the various parties maneuvering with check and countercheck. That is the real "balance of terror" that is held by such a fragile balance."
>
> "You can see the implications if such weapons fall into the hands of the radical terrorists. Ordinary positive energy EMP shooters **can be made fairly readily and cheaply** from nearly just **Radio Shack parts** or surplus parts. It just takes some know-how... When and if the Yakuza [Japanese mafia] gets those (and they will, it's just a question of time), you can begin to appreciate some of the turns this present terrorism and our war on it will take." Bearden

The very thought of the building of one of these howitzers by "some lone-nut gunman" from Radio Shack parts makes one realize how serious it is for humanity to face the crisis it has come to.

Once again, as at the dawn of the nuclear era, humanity finds itself at the crossroads of Heaven and Hell. As a species, we must stop and consider. And there is at least one real difference between this visit to the crossroads and the last: *this* time there is a very real possibility of a true "earthly heaven" on the Heaven side.

The poor nations could come to a good life. Disease could be largely eliminated. Fossil fuel pollution could end. Global warming could be solved, and the weather could be managed in a humane way, bringing water where there is drought, and sunny skies where there is flood. Scalar interferometry would be there to avert any possible hits by asteroids or comets. Mars could be colonized and terraformed. Everyone would be highly educated through psychoenergetic downloading of knowledge. And the **wisdom of the wise** could be realized directly through scalar connection to their minds and hearts.

These stunning sudden possibilities give me pause to think: if our species has reached this crossroads in such a short span of centuries, then surely other species in this vast universe have passed this way before and long ago. And of those who chose the path to "planetary heaven," what is their life like now? It might be good if some of them were with us now, to guide us through this last struggle over this pesky old good-and-evil problem. **It is our last chance to choose the path which benefits the whole species, and not just some elite group or faction**.

The Disclosure Project

One person working tirelessly to bring to bring these technologies into the open is Dr. Stephen Greer of the Disclosure Project, which is calling for congressional hearings into the issues of UFOs and the free energy technology that was, perhaps in part, reverse-engineered from them.

> "Once abundant and nearly free energy is available in impoverished areas for agriculture, transportation, construction, manufacturing and electrification, **there is no limit to what humanity can achieve**. It is ridiculous — obscene even — that mind-boggling poverty and famine exists in the world while we sit on classified technologies that could completely reverse this situation.
>
> "So why not release these technologies? **Because the social, economic and geo-political order of the world would be greatly altered**. Every deep insider with whom I have met has emphasized that this would be the greatest change in known human history. The matter is so highly classified not because it is so silly, but because its implications are so profound and far reaching. By nature, those who control such projects do not like change. And here we are talking about the biggest economic, technological, social and geo-political change in known human history. Hence, the status quo is maintained, even as our civilization hurtles towards oblivion...
>
> "With the types of weapons currently in the **covert arsenal** — weapons more fearsome even than thermonuclear devices — there is no possibility of a survivable conflict. Yet in the darkness of secrecy, actions have been taken on behalf of every human that may endanger our future. Only a full, honest disclosure will correct this situation. **It is not possible for me to convey in words the urgency of this**."
>
> Dr. Stephen Greer http://www.disclosureproject.org/ES-DisclosureImplications-2.htm

Greer and his team have assembled hundreds of witnesses, many of whom are military or ex-military, who are ready to testify to congressional hearings what they know about the covert black projects which have kept free-energy and antigravity propulsion technology secret from the people.

Greer points out the need to regulate these technologies so they are limited to peaceful uses, but this is of course difficult when the government simply chooses to pretend they do not exist. In the meantime, as the President and the Defense Department continue the charade of war-as-usual, those scalar (longitudinal wave) installations which DO exist continue to fall into ever more sinister hands. And there is no reporter in the press brave enough to ask, "**Mr. Secretary of Defense, what steps are you taking to defend our MINDS from attack by the big Russian longitudinal interferometers?**"

Longitudinal Interferometers are Proliferating

In a later statement Bearden reveals that in fact the Yakuza [Japanese mafia] has **already leased** some of the interferometers with a nearly $1 billion "down payment:"

> "After the collapse of the Soviet Union's economy, lots of things got available for lease or purchase that would never have been under the old system. That happened in the resulting economic chaos. The Russian Mafia rose to prominence, as did some wheeling and dealing entrepreneurs who played both ends against the middle.
>
> "**In that atmosphere, and in the need for money, the KGB allowed leasing of the earlier longitudinal EM wave interferometers to a consortium of the Yakuza and Aum Shinrikyo.** They did this only after they had tested and deployed their great new quantum potential weapons. That class of weapon is the dominant weapon on earth. Five nations now have it: Brazil, Russia (KGB), China, and two nations friendly to the U.S. The deal to lease many of the earlier LWIs [longitudinal wave interferometers] to the Japanese Yakuza and Aum Shinrikyo was consummated at the end of 1989, with a down payment of $900 million in gold bullion as "up front" money. I don't know what the lease per year was and is, but probably something like a billion dollars a year. **So the rogue Japanese acquired longitudinal EM weapons** and that technical know-how. In fact, the Yakuza is producing certain kinds of those in its own facilities in Japan now."
>
> http://www.cheniere.org/correspondence/030202a.htm

Personally, I am a little uncomfortable with the knowledge that the Japanese mafia may now be able to insert things into my mind. I'd like to sleep at night knowing that my dreams are my own, such as they may be, and not some weird fantastic visions cooked up by a criminal organization. **This is a world emergency**.

And in Bearden's analysis, it all comes back to the issue of **energy**. It is the energy crisis which is now provoking the current "oil wars," since oil production has hit the "Hubbard Peak" and will decline. And the electrical power grids must be replaced with scalar power as fast as possible, for these grids are now totally vulnerable to any possessor of longitudinal interferometers. They can all be wiped out in minutes. By anyone who possesses or controls a Tesla howitzer.

> "Anyway, in my personal view, we should have a national **Manhattan Project** right now, to develop decentralized electrical power systems taking their energy from the vacuum . . . The MEG is real and so are several other systems invented by private inventors and colleagues. All these efforts should be massively funded and massively and quickly developed. The survival of this nation is going to depend upon it. . . . There are many other facets, but the major point is that **our scientific apparatus and the public are still unaware of what a desperate struggle we have entered, since it was thrust upon us. 9-11 was a wake-up call**. It was certainly bad enough, and one grieves for those lost American lives. But I fear the **worst is yet to come**, and it is coming at us like a runaway train hurtling right down the track at us."
>
> http://www.cheniere.org/correspondence/021702a.htm
>
> "The electromagnetic weapons mentioned by the Secretary of Defense in April 1997 **are in terrorist hands also**, including the **Yakuza and Aum Shinrikyo**. That rogue group leased those earlier weapons on site in Russia, from the KGB, at the end of 1989. The SecDef stated that such weapons were being used to initiate earthquakes, initiate volcanoes into eruption, and control and engineer the weather. Quite true. And presently

there is a truly massive set of weather engineering operations going on over North America, from that rogue Japanese group manning the weapons in Russia. Some 10 nations of the world now have the type of longitudinal EM wave interferometer weapons (which are what the SecDef was most probably referring to). There are other even more fearsome weapons, possessed by five nations. **So a great deal of the state of the world is not covered in the news at all, and will not be . . .**

"We have been in an undeclared war of an eerie kind for some decades. That war seems now to have started to heat up also. The destructive capability of some of these weapons is awesome, and far greater than nuclear."

http://www.cheniere.org/correspondence/021702a.htm

"In that sense, the first phase of WW III is already completed. Now we are only waiting for Phase II.

"It is against that backdrop that I think one must "analyze" such things as how fast we will be replacing the power grid (**it's going to go anyway, and catastrophically, the terrorists will see to that**). And everything we have is critically dependent upon energy, from the economy to jobs to mobility, to the millions of trucks that transport our goods every day, etc.

"As far as I am concerned, we are already late in declaring a great national emergency in energy, because of the coming destruction of the present energy infrastructure. The scientific community should lead, follow, or get out of the way. The emergency is not in seeking "alternative but conventional" systems. It is in developing "energy from the vacuum" on an emergency, crash basis.

"Unfortunately, everywhere I look I just see business as usual. Most Americans (and news media) seem to think, hey, Afghanistan is finished except for a bit of cleanup, and it's just about over).

"It isn't. It has only just begun."

http://www.cheniere.org/correspondence/021702a.htm

Beyond the Howitzers: Quantum Potential Weapons

Do not use EM propagation through space

As more and more nations acquire the scalar interferometers, Bearden speculates on the next generation of longitudinal weapons called "quantum potential" [QP] weapons.

> "The **quantum potential weapons** are presently only held by three nations: Russia (in the hands of the KGB, which is now know by its new name), Brazil, and the "friendly little nation." Red China may be working on them, but is just at the beginning...

> "The QP weapons have been highly reserved by the Russians. However, the complete extent of their capabilities is still being figured out...
>
> "The quantum potential weaponry (possessed by only three nations) is supreme, because **it uses multiply connected spacetime**, and DOES NOT involve propagation of EM energy through space as does the longitudinal EM wave interferometers.
>
> http://www.cheniere.org/misc/qp.htm
>
> "The operational deployment of full-scale strategic QP weapons on site in Russia occurred at the end of 1989, which is when the formal weapons first went operational. These weapons are not in the inventory of the regular Russian armed forces, but all research and development, manufacturing, deployment, an manning and employment is under the ruthless control of the KGB. Shortly thereafter (within weeks), the KGB leased many of their earlier longitudinal EM wave interferometer (LWI) weapons to a rogue Japanese group comprised of the Aum Shinrikyo and Yakuza."
>
> Bearden http://www.cheniere.org/correspondence/021401.htm

In another paper, "Time Polarization - Significance and Weaponization," Bearden tells the story of how a KGB longitudinal wave attack was averted by the longitudinal weapons of a "friendly small nation."

So as it stands, it is precarious. Not until the people demand an end of what author Jim Marrs has dubbed the "**Rule by Secrecy**," can the people's representatives tackle the big problems of a worldwide ban on Scalar Wars and the shifting over to an essentially oil-free economy.

Latest Update from Tom Bearden - March 2002

The whole situation is changing so fast that when I asked Tom to look over this article he mentioned the seeming contradiction when some of the above statements are taken together. At different times the information he has is updated so one has to keep a kind of running track. He was kind enough to clear up the confusion here with the situation as it stands as of March, 2002.

> "We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy." Tom Bearden

Emptiness is Full of Energy

"The Tao is an empty vessel;
it is used, but never filled.
Oh, unfathomable source
of ten thousand things!

"Oh, hidden deep but ever present!
I do not know from whence it comes.
It is the forefather of the gods."

Lao Tzu -Tao te Ching

Afterword

Because of the extreme seriousness of the topics brought forward in this article, I asked Tom Bearden to look it over to make sure I had not gotten anything too terribly wrong or mixed up. I did not want to pass on any misunderstanding of my own. I felt a little nervous writing to an eminent scientist, and hoped I had not presumed upon his time.

But Tom was so kind as to answer immediately, and to my relief I had not gotten anything too terribly wrong. Here is his letter, in which he clears up a misunderstanding, and reports on his current health treatment, and gives a call to the young grad students to turn themselves loose in the whole new field of scalar electromagnetics and free energy.

Subject: RE: Article on Tom Bearden and "Scalar Wars"
Date: Fri, 22 Mar 2002 22:54:01 -0600

Bill,

Well, I must say I'm astounded at your persistence and hard work in congealing all that together. My compliments on a hard job quite well done.

Couple tiny things: In the area mentioning Kervran and Kaznacheyev: Kervran was French; Kaznacheyev was and is Russian. Little typo of "spredding" the immune system. Should be "spreading".

We knew of three nations having the QP weapons for some time. Then we found out last year that five nations had it. So that explains my using the figure "three" first (which represented what I knew at the time). Later I found out about the other two nations. China, e.g., only deployed its QP weapon this past year. So that explains the apparent contradiction between me using "five" in later papers and "three" in earlier ones. It was the best information I had at that time. Might mention that so the reader will understand the apparent discrepancy.

Otherwise, it's pretty accurate.

Just now I'm letting the weapons take care of themselves; have to, as I will be recovering now for nearly a year from now after the heart attack last year and discovering that I had contracted mycoplasia (the BW modified kind) in Canada in 1968. Finally got a test which confirmed the chronic mycoplasia (which I had for 33 years!). The stuff burrows inside one's red corpuscles, then draws nutrients from the hemoglobin, hardening it and reducing its ability to take on oxygen (by as much as 50% or more). The resulting symptom is loss of endurance (which I had for that 33 years) -- the so-called "chronic fatigue syndrome".

Many of the Gulf War Veterans who are sick also have it, though they apparently have other complicating factors as well. A very high percentage of persons with chronic fatigue syndrome also have mycoplasia. The treatment for the long-standing mycoplasma infection of that nature is to stay on antibiotics for a year. The only time the mycoplasma is vulnerable is when one wears out a red cell in normal usage, and the body makes a new red cell to replace it. The mycoplasma comes out of the dying cell to infect the new replacement cell, so it is exposed and vulnerable at that specific time and that time only. The antibiotics kills it at that time. Hence the necessity to stay on antibiotics for such a long time so it is continuously in your bloodstream (I've already been on antibiotics for over two months, and have at least 10 more months to go). We also take a little medical oxygen every day.

Another symptom of the mycoplasma infection is the association of runaway fibrillation of the heart. To stop that, there are pills one takes, and they stop it. But they also further reduce the volume of blood pumped by the heart, so that this projects one right back into the hypoxia, dizzy, loss of balance, etc. Catch 22 situation; the fibrillation will build and cause heart attack, stroke, etc. if not stopped, and if stopped it puts you back in the hypoxia which can cause many things! With the conventional treatment, you're damned if you do and damned if you don't. Hence the need for a little extra medical oxygen, so you can take the "heart clamp" pill and stop the fibrillation, but also augment the oxygen enough to continue to survive without a stroke or some such. Interestingly, it's very difficult to even obtain a mycoplasma test, and Medicare fights you tooth and nail over paying for the oxygen. Either they will pay for it or I'll pay for it myself; it is absolutely necessary.

After a heart attack, one is treated in the heart clinic by specialists who are very caring and excellent practitioners, but know nothing of mycoplasma infection, so not consider it even a part of "heart disease", and who are rather puzzled when one has no clogging of the arteries, has the right count of red corpuscles (largely nonfunctional, of course, but there), pumps the right amount of blood, etc. So they identify and treat only the runaway fibrillation, do not prescribe the oxygen -- and essentially place you on a regime as best they know, but one that is guaranteed to kill or maim you or turn you into a totally disabled person or human vegetable via stroke, paralysis, etc. Fortunately my family doctor will prescribe the oxygen anyway, though I may have to pay for it myself. Whatever works!

However, I'm slowly improving little by little, working as much as I'm able to and trying to finish my book: Energy from the Vacuum: Concepts and Principles, and get it to the publisher for publication later this year. Hopefully the book will turn all the interested young grad students and post-docs loose in the free-energy area. They will not have to spend 30 agonizing years to get to where I am, and where a few others are, but can simply start here and go forward. Since speed to develop energy from the vacuum is of the essence because of the world situation, I'm moving on it as fast as I am physically able to.

Meanwhile, working with the AIAS (Alpha Foundation's Institute for Advanced Study), the AIAS (15 authors) has now gotten about 20 scientific papers published in leading journals (Foundations of Physics, Foundations of Physics Letters, Physical Scripta, Optik, etc.) dealing with electromagnetic energy from the vacuum. Several more are either approved for publication or in the review process also. So slowly we are getting it into the mainstream scientific literature. Hopefully it will be in time, but it's cutting it close.

Very best wishes,

Tom Bearden

Late update

Massive display of longitudinal interferometers?

UFO Fleet Taped
Over Bedhampton, England

Is this sighting a case of the testing of all the Russian (KGB) longitudinal interferometers at once? And coordinating their targeting through the woodpecker grid to a common place, as an exercise in making a massive scalar attack upon a given region? Was this a confluence of Russian Tesla howitzer marker-beacons?

Bedhampton, England - March 22, 2002 - Video Frames by: Anthony Woods

Certainly for a massive all-out scalar attack you would want coordinated movements of the target areas of many howitzers at once, for you could then just cut a bloody swath across a wide area, all the howitzer blasts marching across the land in a kind of scalar version of the "scorched-earth policy." All in a line, blasting, and heat, and mind waves. Or you might want to put the blasts all in the same place for a truly horrendous destruction. And on top of that biowar on citizens whose immune systems have been weakened with the quantum potenial weapons. This is the latest version of Armageddon.

And should any madmen gain control of these weapons, any rogue group at all, and should they have some aggenda to radically depopulate the world, to simply kill by the billion, the means will be in their hands to do so.

Update: MEG patent granted

Motionless Electromagnetic Generator Patent Granted

This message is from **Jean-Louis Naudin**, the French scientist who successfully replicated the Motionless Electromagnetic Generator (MEG) in France. He announces the granting of the patent to the MEG Builder's group at Yahoo.

----- Original Message -----
From: jnaudin509@aol.com
To: jlnlabs@yahoogroups.com
Sent: Tuesday, March 26, 2002 2:20 PM
Subject: [jlnlabs] (Info) GOOD NEWS : The Tom Bearden's MEG IS PATENTED

Dear ALL,

GOOD NEWS !!!!

The **Motionless Electromagnetic Generator**
from Tom Bearden is now PATENTED
US **6362718** granted on March 26, 2002

US Patent 6,362,718 :
Motionless Electromagnetic Generator (MEG)

Abstract: An electromagnetic generator without moving parts includes a permanent magnet and a magnetic core including first and second magnetic paths. A first input coil and a first output coil extend around portions of the first magnetic path, while a second input coil and a second output coil extend around portions of the second magnetic path. The input coils are alternatively pulsed to provide induced current pulses in the output coils. Driving electrical current through each of the input coils reduces a level of flux from the permanent magnet within the magnet path around which the input coil extends. In an alternative embodiment of an electromagnetic generator, the agnetic core includes annular spaced-apart plates, with posts and permanent magnets extending in an alternating fashion between the plates. An output coil xtends around each of these posts. nput coils extending around portions of the plates are pulsed to cause the induction of current within the output coils.

Inventors: Patrick; Stephen L. (2511 Woodview Dr. SE., Huntsville, AL 35801); Bearden; Thomas E. (2211 Cove Rd., Huntsville, AL 35801); Hayes; James C. (16026 Deaton Dr. SE., Huntsville, AL 35803); Moore; Kenneth D. (1704 Montdale Rd., Huntsville, FL 35801); Kenny; James L. (925 Tascosa Dr., Huntsville, AL 35802)

Appl. No.: 656313
Filed: September 6, 2000

For some technical info see my web site at : http://jnaudin.free.fr/html/meg.htm

Best Regards
Jean-Louis Naudin
Email: JNaudin509@aol.com
Main Web site : http://jlnlabs.org
Site France : http://jlnlabs.multimania.com

Click here to view patent # **6,362,718**

Patent Story on Rense.com:

Important Note 2006! MEG delayed: **Click here for a current status history of the MEG from Tom Bearden**

Rense.com

MEG Scalar Energy Device Patented - Production Starts Next Year

From Bill Morgan
3-28-2

Patent was granted on March 26, 2002 for "The Motionless Magnetic Generator," (MEG) US Patent 6,362,718, which is likely to become the first commercially available free energy device in history in about one year from now. The machine will provide free electricity from the vacuum, for the life of the device, which should be a very long life since it has no moving parts. You can see a picture of scientist Jean-

Louis Naudin's MEG replication model at: http://jnaudin.free.fr/html/meg.htm.

It has strong magnets, coils, and a controller unit with the electronics. Naudin made the announcement to the MEG-builder's Yahoo group. The announcement has significance since the patent office has always been skeptical of devices which seem to "get-something-for-nothing." But according to the new science of scalar electromagnetics, the MEG does not break the law of conservation of energy. It's just that the energy is conserved in the fourth dimension, time, and not our 3-space world.

The MEG provides electrical energy by tapping the longitudinal electromagnetic (EM) waves which exist in almost infinite abundance in the vacuum of space. This ocean of energy which permeates everything is sometimes called the "zero point" energy, since it remains there even at absolute zero temperature.

Four inventors are listed: Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore.

Tom Bearden has explained the operation of the MEG on his website Cheniere.org, and also speaks about the new fearful weapons that can and have been made using the same "longitudinal waves" of the vacuum.

The complicated physics of how the MEG works is explained in the paper "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy Replenishment from the Active Vacuum," which can be found at Tom Bearden's website:

cheniere.org.(http://www.help4all.de/energy/MEGpaper.pdf)

The first MEG units to be produced for sale will output 2.5 kilowatts of free electricity. Forever. They should be in production about a year from now. Facilities for manufacturing the device are being set up in an unnamed "friendly nation."

This free electricity will flow indefinitely, without much, or any maintenance. The units may be hooked together to provide more wattage, so four of them would provide 10 kilowatts. After some production experience units will be made which output 10 kilowatts each. With a couple of those units a house could get off the electrical grid.

Tom Bearden, one of the inventors has said "I will admit that the chief scientist of an important experimental group in a large company was rather stunned at the type of output we were able to obtain. The MEG may look like just a transformer, but it is not. It is a completely different breed of cat." This cat, it would seem, is out of the bag now.

The ordinary EM waves we are familiar with are called "transverse" waves, to distinguish them from the new "longitudinal" EM waves of the vacuum. Bearden has explained in depth on his website cheniere.org that wherever there is a dipole (battery, generator, magnet) there is an unseen flow of longitudinal EM waves in that local vacuum, the only problem is in tapping that energy and "transducing" it to electricity. Bearden says that the problem with all the electrical circuits we have is that they are two-wire circuits, a loop by which half the energy goes back to destroy the dipole. In the MEG that closed loop is never made. So the dipole is not being destroyed.

Patents are not granted on devices which do not work, so in a sense this announcement proclaims a new era. The MEG device itself proclaims and proves that energy is abundantly available everywhere, for free. We only have to build the devices to tap it. MEG type devices could be put into a car with an electric engine to make a truly fuel-less automobile. Electricity can be available in very remote places. People will be able to get off our very terrorist-vulnerable power grids. In an era of terrorism a highly dispersed power system would be most desirable rather than our centralized systems.

What is somewhat astonishing to me is the discovery by Bearden et. al. that time itself is actually compressed energy, and that this free energy is actually coming from the time domain, the ocean of longitudinal EM waves which fill the empty vacuum of spacetime. In fact, time is energy compressed by the same factor that matter is compressed energy: the speed-of-light squared. Thus there is a new companion to Einstein's E=mc2. Can you say "E equals delta-tee-cee-squared?" The "tee" is time and delta-tee is change in time.

A company has been set up to manufacture the MEG called Magnetic Energy Limited.

The ramifications of free energy are enormous. The oil wars are not necessary. If we threw as much money at this technology as we are spending on the oil wars we would be free of the need for oil in less than a decade. With fuel-less cars air pollution will be greatly lessened. Third world nations can raise their standard of living eventually. And the energy is free. And it never runs out.

I have put together a kind of "Bearden for Beginners" article which explains some of the basic concepts of the current state of scalar technology.

http://www.prahlad.org/pub/bearden/scalar_wars.htm

The granting of the MEG patent is the herald of the new era of scalar electromagnetics, and the free energy which flows forever and never runs out.

Bill Morgan

Status History of the MEG

updated Oct. 26, 2006

Important notes on the status of the MEG (Motionless Electromagnetic Generator)

Since the time this "**Scalar Wars**" article was written the timetable of the **MEG** has been thrown off considerably. Bearden had originally said the MEG could be marketed within 1-2 years ***of receiving funding for final research and scale-up costs***. I thought certainly it would not take very long, but over 5 years later they still have not gotten the $10 million or so necessary. There are various reasons for this, some of which Tom Bearden has written about.

The place for updates on the MEG status is Tom Bearden's correspondence section at his website Cheniere.org:

http://www.cheniere.org/correspondence/index.html

Here are a few of his reports which detail the history of the MEG status :

Mar. 6, 2006
http://www.cheniere.org/correspondence/030206.htm

May 16, 2005
http://www.cheniere.org/correspondence/051605.htm

Jan. 19, 2005
http://www.cheniere.org/correspondence/011905.htm

Dec. 5, 2004
http://www.cheniere.org/correspondence/120504.htm

Do the magnets run down after a while?

Further developments

Tom Bearden clears up a question

When the MEG patent was released many questioned the patent language which seemed to imply that permanent magnets of the MEG would "wear out" after awhile. I myself wondered about the same thing so I put the question to Tom, and here is his reply.

> Dear Bill,
>
> Thanks for the kind words; much appreciated.
>
> I'm not up to going on radio or TV, and won't be for quite some time. Still suffering from hypoxia, though very slowly improving, and will still be on antibiotics for 10 more months at least (maybe longer).
>
> Obviously an extensive dialog developed between the Patent office and our patent attorneys. As best I can understand it, the Patent office accepts the fact that a permanent magnet produces and emits energy, or has a history of accepting it. They do not appear to be at all cognizant of the broken symmetry of opposite charges, and hence of a dipole. So in their opinion (which seems to be vague), the magnet has to use itself up very gradually, or some such. So some rewrites were essentially required by the Patent office.
>
> All I can say is that a magnet is a very durable thing, so long as you don't heat it up too much or shock it to much. Certainly one will make a good 20 years, and of course there are plenty of magnets that have been around awhile longer.
>
> That said, in our earliest experiments we did "measure" some such effect, or so we thought, until we found that a particular instrument was bad and giving erroneous readings. So with a new instrument, we have not observed since then any such readings.
>
> Out of all that fuzziness came the inclusion of "using up the magnets gradually". So far as we are aware (after correcting those early readings on a defective instrument), we have seen no such measurements or effects since then at all. Nonetheless, just in case there is some very small effect in that respect, at least it's in there.
>
> Actually, as you are aware, capacitors and solid state components are the major concerns for failure, and any system does require maintenance, both periodic and whenever malfunction or failure occurs. The MEG is no different from any other device in that respect, so of course it will have a component failure rate and maintenance required from time to time, just as any other electromagnetic device. But without moving parts, it can be ruggedized and made very durable indeed.
>
> Please give Jeff Rense my regrets, and I do admire his work and his show. He's laying it in there.
>
> The depth of the scientific mindset against COP>1.0 is inexplicable, particularly when physics itself already has COP>1.0 validated

experiments (such as Bohren's experiment, which outputs 18 times as much as one inputs by Poynting calculations), and it works every time. Any competent university optics lab can perform the experiment. Also, even the conventional texts admit that a charged capacitor or electret laid on a permanent magnet, so that the E of the cap or electret and the H of the magnet are at right angles, is a "free energy machine". That silly thing will just sit there and pour out Poynting energy S = EXH indefinitely. Years and years and years. If you wait just a single year, that silly thing has changed the energy density of a volume of space a light year in radius, reaching out well beyond the solar system. Now that's a WHALE of a lot of EM energy that beast poured out in that year, and it will still be going strong and pouring it out at a steady rate. The charges and dipoles in original matter in the universe have been pouring out EM energy in that fashion for some 14 billion years, give or take a billion or so. It's also easy to see that, prior to Lorentz's ARBITRARY symmetrical regauging of the Maxwell-Heaviside equations, those equations prescribe both EM systems in equilibrium with their active environment, and EM systems not in equilibrium with it. The latter systems are permitted by the thermodynamics of open disequilibrium systems to perform five magic functions: (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator inputs (the active environment inputs the rest of it), (4) power itself and its load (the active environment inputs all the energy, like a windmill), and (5) exhibit negentropy. That is a well-established thermodynamics. Lorentz arbitrarily discarded all such open disequilibrium EM systems in order to get simpler equations having analytical solutions and not requiring numerical methods. So he simply scrapped all the really interesting and challenging EM systems in Maxwell's theory. ARBITRARILY!

That our scientific establishment continues to teach only that "half" of the theory, and not point out the rest to the young students, and not fund research into developing such systems that take energy from the vacuum, is simply inexplicable. It's a matter of total dogma, not science at all.

But hopefully it is changing a little bit. At least now we cite chapter and verse when these things were done, and what they are, and the young grad students and post docs can go check the papers and make up their own minds.

That is the real hope of the future: that we get a generation of young fellows, with excellent skills, who now know where all the skeletons got hidden in the closet, and understand that the electrical engineering model arbitrarily assumes an inert vacuum environment (falsified for more than a half century in particle physics) and a local flat spacetime (falsified for nearly a century by general relativity).

Once they know that past history, and go read into what broken symmetry of opposite charges means for a dipole and for all dipolar EM circuits, they will set to and change that more than a century of error. In that case, we shall have numerous solutions to the energy crisis rather quickly, popping up all over.

Good thing. It will certainly displace some great energy barons, but it will also free the populace. One can extract all the EM energy from the vacuum one wishes, anywhere, anytime, with ridiculous ease. One can produce that "electrical wind of energy" at will, by making a simple dipole and then letting it alone. The only energy problem is in how to then intercept some of that energy flow and "catch" it in a circuit, discharge it in a load to power it, and not use half the "caught" energy to kill that dipole that is gushing out the extracted EM energy from the vacuum.

Very best wishes,

Tom Bearden

END

Revised: May 16, 2007

Download this article in Ms Word format
(ScalarWars.doc)
Send a comment to Bill Morgan

See also:

Scalar Electromagnetics

Scalar Wars:

The Brave New World of Scalar Electromagnetics
(Read this one first!)

Horrifying US Secret
Weapon Unleashed In Baghdad

Scalar Heaven or Hell:
Psychoenergetics

Engineering the Mind

Sacred_Scientist thread at Rumor Mill News

Are any of these Earthquakes Being caused by Scalar Weapons?

Safe Places, Earth Changes, and Gordon Scallion

Examples of probable Scalar Weapons Tests

Iraq: World's First Scalar War?

Space Energy Access Inc. Announces
New Free Energy Device

SINGLE MOST IMPORTANT BREAKTHROUGH IN HISTORY

Comparison: Ebomb and Scalar Waves

IT COULD COME AS SIMPLY AS THIS
Scalar destruction of U.S. power grid

WHAT A SCALAR "SMOKING GUN" WOULD BE

BEARDEN REPONDS TO SCALAR ATTACK POSSIBILITY (Columbia)

THE SCALAR ATTACK POSSIBILITY
Columbia Disaster

TOM BEARDEN ON THE CHALLENGER DISASTER

ANOTHER BEARDEN ACCOUNT OF '86 CHALLENGER SCALAR ATTACK

Broadcasting Thoughts & Disease

Thoughts on Diane Harvey's Appeal to the U.N.

Scalar War Has Already Begun

Musings on Scalar Electromagnetic Phenomena

The New Scalar Paradigm

ELECTROMAGNETIC WEAPONS TO BE USED

SCALAR: BEARDEN SAYS U.S. NOW DEFENDED

American Karma & The River of Tears

Open Letter to Donald Rumsfeld

My Rumor Mill Posts on All Topics

Your comments are welcome: prema@prahlad.org

Journal of Directed Energy, **1**, Fall 2003, 48–72

Compact Pulsed Power for Directed Energy Weapons

**David Price,* Carl Bloemker, Edward Goldman, David Nett,
Sidney Putnam, Doug Weidenheimer, Roger White,
and Tony Wynn**

Titan Corporation, 2700 Merced Street, San Leandro, California 94577

The Titan Corporation has been developing directed energy technologies since the early 1980s. Over this time Titan has provided advanced power supplies for electromagnetic gun and laser research while making broader contributions in the field of high-power microwaves (HPM). Titan's HPM work expanded from initial efforts fielding and operating the first gigawatt-level oscillators for susceptibility test applications, to research and development maximizing the peak and average output powers and overall efficiency realized from such systems. As interest in HPM technology has spread globally, Titan has leveraged this core competency and provided the HPM equipment for nearly all of the major European effects test facilities. Titan's current interests focus on compact, efficient and reliable directed energy weapon systems and the advanced subsystems and components that enable the same. Specific subsystems of interest include pulse-forming networks and intermediate energy storage and power conditioning elements (electronically reconfigurable batteries and power ride-thru subsystems). These subsystems are designed for reduced size and weight while still meeting severe service, platform integration, lifetime, and thermal management constraints. Specific components under development at Titan include laser-gated solid-state switches and both high peak and high average power, frequency-agile, HPM oscillators. Status and recent results from this research are presented.

KEYWORDS: Electronically reconfigurable batteries, High-power microwaves, Laser-gated solid-state switches, Magnetron, Power-ride thru, Pulsed power, Reltron

1. Introduction and Previous Work

Directed energy weapons (DEWs) are emerging as a key defense technology of the early 21st century. The first laser antimissile and antiartillery systems are within perhaps five years of deployment, and newly envisioned information warfare and nonlethal point defense missions are hastening the deployment of high-power microwave (HPM) systems. Department of Defense (DoD) programs in "more electric" and "all-electric" platforms and electrically driven weapon and self-defense systems are promoting the development of laser, high-power radio frequency (RF) and both electromagnetic (EM) and electrothermal/chemical (ETC) gun technologies. In fact, the DoD's current fundamental "transformation" efforts are based on stressing the development of capabilities to deal with threats emerging in the

---

Received February 13, 2003; revision received May 28, 2003.

*Corresponding author; e-mail: dprice@titan.com.

© 2003 Directed Energy Professional Society

Table 1. Triservice directed energy applications

U.S. Army [Future Combat System (FCS)—hybrid electric drive vehicles]
- Advanced armaments—electrothermal/chemical guns
- Enhanced area air defense—solid-state heat capacity laser system
- Counter-HARM
- Counter-ATGM
- Countermine (electronically fuzed and wide-area threats)

U.S. Navy [All-Electric Ship Programs (Integrated Propulsion System and DDX)]
- Advanced stand-off shore bombardment—electromagnetic guns
- Advanced anti-ship missile defense (free-electron laser and HPM)
- Nonlethal area denial

U.S. Air Force (more electric aircraft program and advanced UAV programs)
- Enhanced self-defense—solid-state heat capacity laser system
- Space-based laser program
- Airborne laser
- Suppression of enemy air defense
- Aircraft self-defense
- Close air support
- Defensive and offensive counter air
- Strategic attack/strike warfare, attack operation/air interdiction
- Combat search and rescue
- Area denial
- Cruise missile defense

Table 2. Countries now developing high-power RF technology

Programs	Countries
Major	United States, China
Medium-scale	United Kingdom, France, Germany, Russia
Emerging	Sweden, Japan, India, Taiwan, Australia, Israel, South Korea

Information Age instead of emphasizing countermeasures to specific threats. DEWs hold the promise of engaging multiple threats and multiple threat types, thereby fitting very well into this conceptual paradigm. Some specific examples of defense applications of DEW technologies are shown in Table 1.

The initiatory HPM programs in the USSR and United States have evolved and influenced foreign allies and threats to take an interest (Table 2). The collapse of the USSR led to the dispersal of Russian and Ukrainian HPM workers and the proliferation of the Soviet technology to the Third World. This global emergence is driven by at least two factors: 1) The dependence of both military systems and civilian support infrastructure on sensitive electronics is increasing, and 2) after a more-than-20-year effort, an understanding of the susceptibilities of both military and commercial systems to HPM threats is maturing.

For these reasons the Titan Corporation has taken a keen interest in the development of systems, subsystems, and advanced components across several different technologies to support HPM and laser directed energy programs.

In 1997, Maxwell Technologies purchased Physics International (PI) from Primex Technologies and merged two of the three most prominent, commercial, pulsed power entities in the United States. PI and Maxwell pulsed power capabilities were combined into a new entity, Maxwell Physics International, which became a component of Maxwell's Systems Division. Maxwell subsequently decided to leave the DoD and Department of Energy (DOE) pulsed power business altogether and sold the combined Maxwell/PI pulsed power divisions to the Titan Corporation, the parent company of Pulse Sciences Incorporated (PSI). Now, the three most prominent names in commercial pulsed power are combined as the Pulse Sciences Division (Titan-PSD) of the Titan Corporation.

In March 2002, Titan completed the acquisition of Jaycor and its subsidiary California Tube Laboratory (CTL). This acquisition established Titan as one of the world's primary industrial authorities on electromagnetic effects and suppliers of RF and narrow-band microwave DEW technologies and associated electrical power systems. High average power, high efficiency, magnetrons; high peak power, pulsed, relativistic magnetrons and superreltrons; short pulsed, ultra-broad-bandwidth systems and megawatt-class power supplies and power conditioning subsystems are all designed and manufactured by Titan. This paper will chronicle some of the past and report on some of the current contributions that Titan has made to the development of pulsed power technology for DEW applications.

1.1. Previous Titan work in support of HPM programs

Titan-PSD has been developing directed energy technologies since the early 1980s. In this time, Titan has delivered capacitor banks for electromagnetic gun research with a total stored energy exceeding 200 MJ and advanced pulsed power systems for several large laser systems (Fig. 1). Over 40 years, Titan-PSD has delivered more than 200 pulsed power systems for electromagnetic pulse (EMP), x-ray, and lightning simulation; flash radiography; and other applications worldwide. Although somewhat peripheral to DEW research proper, these projects have nonetheless formed much of the experience base supporting Titan's directed energy technologies contributions.

Titan's primary contributions have been in the field of HPM (Fig. 2). Work started in 1984 when Titan set up and began operating a susceptibility effects test facility for the Defense Threat Reduction Agency (DTRA). A 30 × 20 × 16 ft anechoic chamber to support this work was originally designed to operate in X-band and above. At the outset of the U.S. HPM directed energy program the community felt that the higher gains possible from fixed antenna sizes at higher frequencies would drive technology development into these bands. Over the next 10-plus years, the 25 effects test programs conducted on strategic and tactical systems in the Titan facility and tests conducted at other sites both confirmed that the lowest susceptibility thresholds are observed in the S-band and below. HPM source development history has mirrored these findings.

1.1.1. HPM source developments at Titan. In 1983 Titan initiated HPM source research in its test facility by developing first C-X band overmoded vircators[17] and later an L-band vircator.[13] These systems produced gigawatt outputs through the oscillating virtual cathode mechanism. Their main shortcoming was their broad bandwidth due to chirping induced by diode gap closure. To address this shortcoming Titan developed the first cavity vircator[3] that reduced the output bandwidth from several hundred megahertz to a few tens of megahertz. Versions of these devices were soon after sold to Nucletudes in France

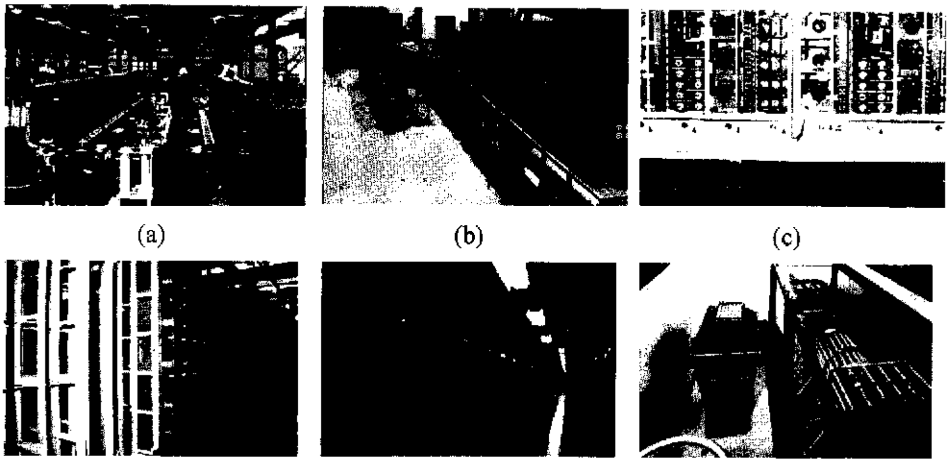

(a) (b) (c)

(d) (e) (f)

Fig. 1. Titan-PSD has designed, fabricated, and delivered the pulse power subsystems for most of the major electric gun facilities worldwide: (a) the decommissioned 32-MJ Thunderbolt System built for a Strategic Defence Initiative Office (SDIO) project; (b) the existing 32-MJ Kirkcudbright system currently being operated by the United Kingdom; (c) the first transportable pulse power subsystem designed for an electric weapon system. [This 8.5-MJ stored energy subsystem, consisting of four separate PFNs, was designed by Titan-PSD under the Army/Army Research Development and Engineering Center (ARDEC) Balanced Technology Initiative and was capable of delivering up to 5 shots at 3 rounds per minute.]; (d) the 52-MJ bank delivered to the ARDEC for electromagnetic and coil gun research; (e) the pulsed power driver for the 40-MJ OMEGA laser at the University of Rochester; and (f) a portion of the 100-Hz, 750-kV, 1.25-μs modulator for NRL's EMRLD laser.

and Culham Laboratory in the United Kingdom to help kick-start HPM effects testing and technology development programs at each location.

Titan initiated work on magnetrons by acquiring S- and X-band, A-6 versions from Bekifi at the Massachusetts Institute of Technology.[2] These devices were integrated into the Titan test facility, and within two years an L-band version of this same design was developed.[16] An S-band magnetron was sold commercially to Thales in France in 1987 to help kick-start HPM effects testing and technology development programs there.

This HPM source development work soon expanded from the initial efforts to field gigawatt-level vircators and magnetrons for the effects testing applications to the development of other high-power sources. These include narrow-band klystrons (both high- and low-perveance variants) and reltrons, as well as ultra-wide-band systems. In parallel, the peak and average power and spectral characteristics of these HPM sources were being explored in research involving phase locking, repetitive pulsing, and frequency agility. Other related Titan work included the development of pulsed power systems to drive these HPM sources involving several different architectures [pulse-forming networks (PFNs) with capacitive energy storage and voltage adders with magnetic compression power conditioning]. These advancements are described briefly in the following.

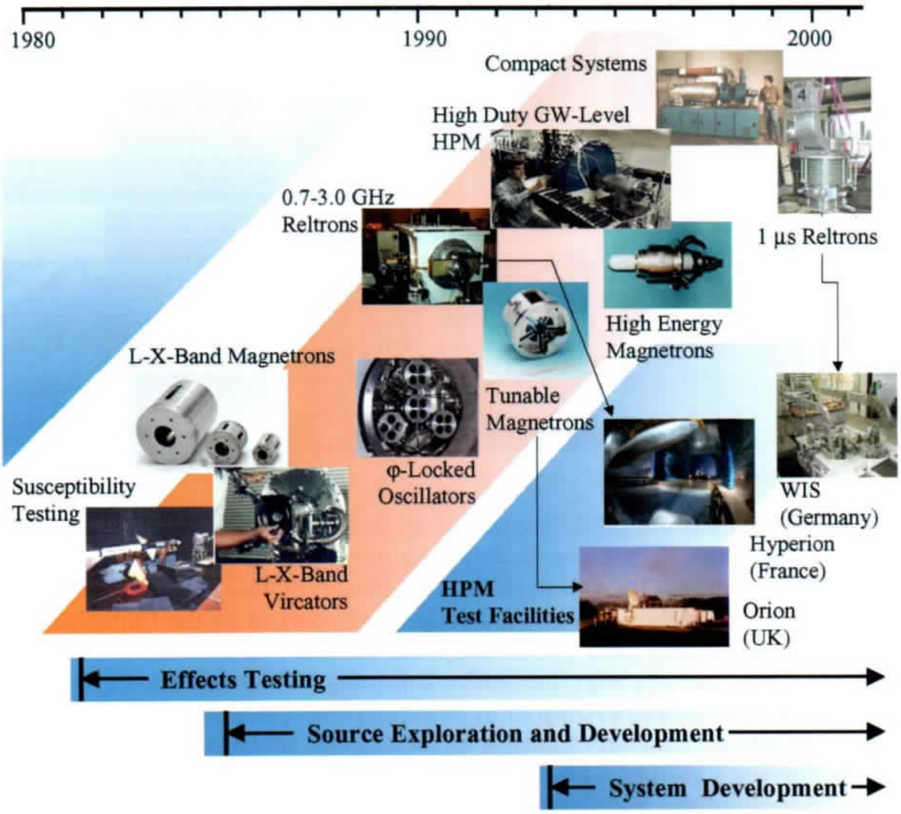
1980
1990
2000
Compact Systems
High Duty GW-Level HPM
0.7-3.0 GHz Reltrons
1 μs Reltrons
High Energy Magnetrons
L-X-Band Magnetrons
Tunable Magnetrons
φ-Locked Oscillators
Susceptibility Testing
L-X-Band Vircators
WIS (Germany)
Hyperion (France)
HPM Test Facilities
Orion (UK)
Effects Testing
Source Exploration and Development
System Development

Fig. 2. Titan contributed to the field of HPMs in three distinct eras: 1) susceptibility effects testing, 2) HPM source development, and (currently) 3) DEW system development. The effects testing work started in 1984. Then, Titan operated the DTRA susceptibility effects test facility supporting tests for external DoD users. Soon thereafter Titan began using the facility for its own end-to-end DEW research programs. In parallel, Titan HPM source development efforts initially focused narrowly on providing the facility with minimum essential capability but soon expanded to meet specific weaponization and testing requirements. During this period, Titan performed some of the seminal narrowband HPM work associated with phase-locking, frequency agility, and repetitive operation. Source development work continues today, focusing on improving energy output and tunable bandwidth. These efforts are combined with the development of other advanced pulsed power subsystems and components needed to realize compact, efficient, and reliable DEW systems.

In the late 1980s, Titan S-band magnetrons were used in a three-year effort to understand how to achieve output powers exceeding 10 GW from a phase-locked array.[9] Many mission scenarios for DEW systems that could produce burnout or upset in military targets using only a single HPM pulse drove output requirements to this 10-GW level. The combination of gigawatt-level sources locked in phase appeared an effective way to realize this output level without increasing the RF fields and beam energies within individual diodes (and dealing with the RF breakdown and thermal issues that would certainly ensue). In this effort both master/slave and peer configurations were investigated to understand the optimal architecture for such an array. In the end, a module of seven "peer-coupled," 2.8-GHz, magnetrons was built that produced an output power of 2.9 GW. Extrapolation to even higher output levels appeared feasible.

At this time, the HPM community was beginning to recognize that there must exist a middle ground between HPM and electronic warfare (EW) DEW concepts. The HPM approaches were characterized by superpower, single-shot, single-frequency, pulse formats, while the EW approaches made use of very low powers but incorporated modulations in the output pulse formats to effectively exploit susceptibilities within their intended targets. To access this middle ground the HPM sources had to first be shown capable of operation in repetitive or burst modes. This burst capability would be surely required of any DEW system, if only to engage multiple threats deployed in sorties or clusters. In response to these weaponization considerations Titan-PSD developed relativistic magnetrons capable of operation in 1-kHz bursts.[4] A 1.1-GHz, L-band, magnetron was driven by a magnetic pulse compression modulator (see discussion below) and achieved 250-pps repetitive operation at 1.2-GW peak power, in a burst of 100 shots. The average power of 12.6 kW during the burst remains today the state of the art for gigawatt peak power sources. To demonstrate that the magnetron's diode could recover in a time short enough to support higher repetition rates, the device was run for ~5 pulses (limited by the modulator) with 1 ms between individual pulses.

In the early 1990s both Jaycor and PI were actively extolling the use of frequency-agile waveforms to enhance lethality in backdoor attacks. Frequency agility exploits the characteristics of typical coupling cross sections that show pronounced dependence on microwave frequency. An HPM source that can sweep or hop frequency across a band has a higher probability of matching to a coupling resonance than a source fixed at a single frequency. Because of this strong dependence of backdoor coupling and susceptibility levels on microwave frequency, it was apparent that test facilities needed the capability to vary frequency continuously in order to obtain a comprehensive and accurate assessment of any given test asset's susceptibilities. In response to these testing and weaponization considerations Titan-PSD developed frequency-tunable magnetrons. These oscillators can be tuned ±17% about a central frequency,[10] which today still represents the state of the art. In the mid-1990s these magnetrons were integrated into advanced concept demonstrations for the Army and were also exported to the United Kingdom as part of an HPM test facility.

In the late 1980s and early 1990s Titan-PSD developed several ultra-broad-bandwidth (UWB) systems for Army and Air Force DEW research programs. The first system was based on a spark gap-switched Marx that drove a variety of different broadband antennas. It could launch 200-MW total RF power in 3-ns pulses with rise times of ~300 ps at a repetition rate of a few pulses per second. This compact, self-contained system was used in both indoor susceptibility tests and outdoor technology demonstrations. A second project produced a system that could launch impulse waveforms at 200 pps in burst mode.[15] A 3.66-m-diameter paraboloidal reflector was driven at its feed by a hydrogen switch that was integrally coupled to a novel electromagnetic lens. The radiated waveform had a rise time of ~100 ps and a 10–90 rise-to-fall pulse width of 45 ns. The peak electric field on bore-sight measured at 305 m was 4.2 kV/m.

Titan invented the super-reltron in 1992. A suite of tubes that operate between 0.7 and 4.5 GHz has been developed since that time.[12] These novel tubes represent a marriage between pulse power and conventional high-power klystron techniques. An energetic electron beam is modulated by periodic virtual cathode formation, and the bunched electron beam is postaccelerated to energies approaching 1 MeV. Microwave power is extracted with ~50%

Table 3. Operating parameters of Titan-built modulators for DEW applications

Modulator	Voltage, kV	Current	Pulse width	Pulse repetition frequency	Pulse energy	Average power	Switch type
LS-15	50	10 kA	100 ns	1.9 kHz	50 J	95 kW	Spark gap
NED laser driver	±40	260 A	2 μs	10 Hz	3 kJ	30 kW	Thyratron
EMRLD PFN	750	38 kA	1.3 μs	125 Hz	35 kJ	4.4 MW	Spark gap
EMRLD (trigger)	600	260 A	400 ns	125 Hz	100 J	12.5 kW	Spark gap
OMEGA[a] laser driver	15	128 kA	200 μs	Single shot	40 MJ	—	Ignitrons
CLIA HPM driver	750	10 kA	100 ns	250 Hz	750 J	188 kW	Magnetic
ORION HPM driver	500	10 kA	50–500 ns	100 Hz	250 J	25 kW	Thyratron
Compact HPM driver	450	9 kA	450 ns	10 Hz	1.8 kJ	18 kW	Spark gap

[a] 120 modules, each storing 0.33 MJ.

conversion efficiency downstream in a dual-cavity output section that is tuned to the bunch frequency. The reltrons can produce microwave pulse widths approaching 1 μs and pulse energies of a few 100 J, outstripping the relativistic magnetrons that to date are limited to pulse widths of about 100 cycles and pulse energies below 100 J.

In response to the HPM effects testing considerations discussed above, Titan developed frequency-tunable reltrons. These oscillators can be tuned ±10% about a central frequency.[11] These devices are still in development today (see Sec. 2) and have been provided to both France and Germany as HPM threat simulators.

1.1.2. Development of pulse power modulators for DEW applications at Titan. Titan has developed a large number of modulators for a variety of laser and HPM DEW applications, with voltage ratings from tens of kilovolts to several megavolts and average power levels from tens of kilowatts to multimegawatts (Table 3). They also range from relatively simple units to complex, state-of-the-art devices with various different architectures and subsystems. These include PFNs for generating the required pulse shapes; thyratron switching for precise timing of the pulse switching into the load; low inductance layouts for fast output pulse rise times and high-turns-ratio pulse transformers for generating the high voltage outputs. The switching technologies include thyratrons, spark gaps, magnetic switches, and solid-state devices, such as metaloxide semiconductor field effect transistors (MOSFETs), insulated gate bipolar transistors (IGBTs), and thyristors (Fig. 3).

Several of these modulators have been designed for long-lifetime, reliable operations. Such features are essential if these systems are to find their way into military applications. As an example, the Defense Advanced Research Projects Agency (DARPA) excimer Raman-shifted laser device (EMRLD) modulator (a 100-Hz, 750-kV, 1.25-μs, lumped-element PFN to pump an excimer laser) was specifically designed for and demonstrated a lifetime of 10^{10} shots.

The ORION pulser (Fig. 3b) design has command resonant charge and intermediate energy storage sections each switched by thyratrons that drive a step-up transformer and

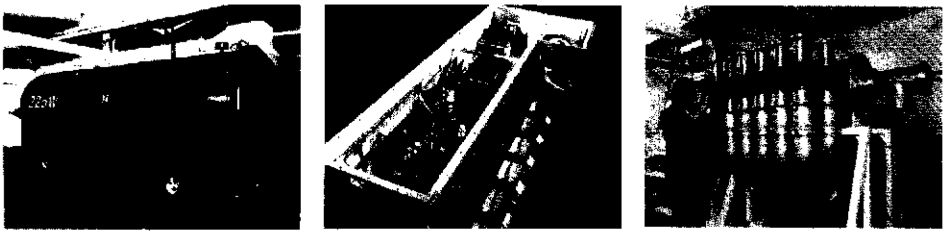

(a) (b) (c)

Fig. 3. Titan-PSD pulsed power systems based on (a) spark gap, (b) thyratron, and (c) magnetic switching.

Table 4. Key operating parameters in European HPM test facilities

HPM facility	Power density, W/cm^2	Range, m	Frequency, MHz	Maximum pulse width, ns	Maximum repetition rate, Hz	Shots per burst	Source type
UK—Orion	>120	100	1,070–3,300	100	100	100	Magnetrons
France—Hyperion	>1,000	~100	700–3,300	300	1	—	Reltrons and magnetrons
Germany	>170	15	700–1,450	300	10	100	Reltrons
Sweden[a]	~120	15–25	L-Ku	500–5,000	1,000	Continuous	Reltrons

[a] This facility uses conventional high power klystrons in L-, S-, C-, X-, and Ku-bands, with maximum power 25–0.25 MW.

PFN. The modulator can fire 1,000 pulses in a burst at repetition rates up to 100 pps and produces 500 kV into a 50Ω load. It has been used to drive relativistic magnetrons[6] for HPM effects measurements.

The Compact Linear Induction Accelerator (CLIA) (Fig. 3c) was developed for moderate-repetition-rate HPM source research and development.[1] CLIA operates at 250 pps with 750-kV, 10-kA, 100-ns output pulses. The system uses thyratron and magnetic switching throughout and produces 188-kW average power into the load. CLIA is not actively cooled and so is restricted to burst operation with 5,000 pulses in a burst. CLIA has been used to drive both magnetron and klystron loads. This technology is scalable to kilohertz repetitive operation.

1.1.3. Titan pulse power systems for HPM simulation. In addition to work done supporting domestic RF and microwave DEW research and development, Titan has provided HPM simulation systems for all of the world's major HPM test facilities in Sweden, France, the United Kingdom,[14] and, most recently, Germany. The microwave test parameters and operational capabilities achievable in these international facilities are summarized in Table 4. The types of Titan HPM sources that produce the intense microwave environments are also given. Work continues today upgrading those systems and providing similar hardware for several non-European countries.

2. Most Recent Contributions

Titan-PSD has taken an interest in the full spectrum of developments of advanced subsystems and components that the three U.S. services will require to realize their DEW and electric platform goals. These developments enable many different military missions and cut across many different technologies: advanced capacitors, switches, transmission systems, RF, and optical systems. Some of the pulse power and power electronic subsystem and component developments that Titan is pursuing are summarized in Table 5.

The following subsections describe five electrical subsystems and three components that Titan-PSD is currently developing for DEW applications.

2.1. Titan-PSD development of subsystems for DEW applications

2.1.1. HPM modulator. Titan-PSD has a conceptual design of a high-voltage-pulse power subsystem for a short-pulse HPM system. The given RF pulse output parameters led us to choose a specific tube that is compatible with the constraints of an airborne system, namely the magnetron with a 20Ω impedance and 33% efficiency in a short-pulse mode. Our pulse power design concept is driven by the tube input requirements (20Ω load, 40-ns pulse width, 10 pps) and comprises a low-impedance, water-filled, stepped-pulse forming line (PFL) charged by a fast Marx and discharged through a self-breaking, spark-gap output switch. An advantage to using a liquid dielectric PFL is that its impedance can be easily varied along its length to accommodate variations in load impedance. Input power to the Marx is from two 400-V series strings of lithium ion batteries with solid-state switched intermediate voltage multipliers providing ±40 kV. Batteries are recharged from aircraft power through a small voltage converter.

Figure 4 is a simplified circuit diagram of the pulse power modulator coupled to a 20Ω constant impedance load though a transit-time isolator. The transit-time isolation is integral to the two parallel cables that couple the output of the pulsed power modulator to the microwave source. The Marx stores 1,140 J at a voltage of ±40 kV. It erects to a 1,280-kV open circuit voltage after a trigger pulse is applied to several of its switches from the trigger generator. Near peak PFL voltage, the self-firing spark gap closes, and energy is transferred to the tube at a peak voltage of 640 kV.

Figure 5 shows the voltage and current in the load computed from a more detailed version of the circuit depicted in Fig. 4. Note that the current is multiplied by a factor of 10. Voltage rises in about 11 ns (10–90%), and the width at the peak is approximately 40 ns. We recognize that the impedance of the magnetron can vary during RF generation, but the constant impedance model is sufficient here. An advantage to using a liquid dielectric PFL

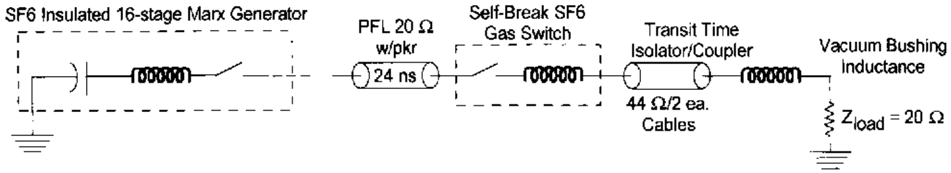
SF6 Insulated 16-stage Marx Generator
PFL 20 Ω w/pkr
24 ns
Self-Break SF6 Gas Switch
Transit Time Isolator/Coupler
44 Ω/2 ea. Cables
Vacuum Bushing Inductance
Z_{load} = 20 Ω

Fig. 4. Simplified circuit diagram of the pulse power modulator coupled to a 20Ω constant impedance load though a transit time isolator. This impedance represents a magnetron electron tube load.

Table 5. Titan advanced DEW power technology development rationales

What need does the development of pulsed power technology fulfill?		What need does the development of high power/high current technology fulfill?	
It is an enabler for advanced electric weapons and self-defense systems.		It is an enabler for advanced electric weapons and self-defense systems.	
System	User	System	User
Dynamic protection	Army (FCS), Navy (carrier)	Solid-state lasers	Army (Anti-Aircraft and Anti-Missile), Navy (ASMD)
ETC	Army (FCS), Navy (shore bombardment)	HPM	Army (FCS), Navy (non-lethal area denial)
HPM/electromagnetic pulse	Army (FCS), Navy (non-lethal area denial and ASMD), Air Force (SEAD)	EMALS	Navy (carriers)
EM gun	Navy (shore bombardment)	Electric arresting gear	Navy
UWB	All services (counter C4I)	Ride-thru	All services
		It is an enabler for electrically driven platforms: • FCS platforms for the Army • Combatants such as DD21 for the Navy • More electric aircraft for the Air Force	
It is an enabler for commercial spin-offs: • Environmental remediation systems • Material surface treatments • Pulsed thrusters for satellite station keeping		It is an enabler for commercial spin-offs: • Electric vehicles • Commercial power distribution and control • Advanced semiconductor lithography	
What must we accomplish at the subsystem level? • Develop compact, reliable, affordable PFNs • Develop compact, reliable, affordable modulators • Develop compact, reliable, affordable intermediate energy stores		What must we accomplish at the subsystem level? • Develop compact, reliable, affordable prime power, power conversion and controls • Develop compact, reliable, affordable intermediate energy stores • Develop compact, reliable, affordable ride-thru systems	
What must we accomplish at the component level? • Develop high-energy-density, fast-discharge capacitors • Develop long-life, reliable, high-current switches • Develop compact inductors for pulse control • Develop affordable ultracapacitors • Develop compact, reliable, efficient, high-average-power HPM sources		What must we accomplish at the component level? • Develop high-current-density, high-voltage stand-off solid-state switches • Develop affordable ultracapacitors • Develop compact, reliable, efficient, high-average-power HPM sources	

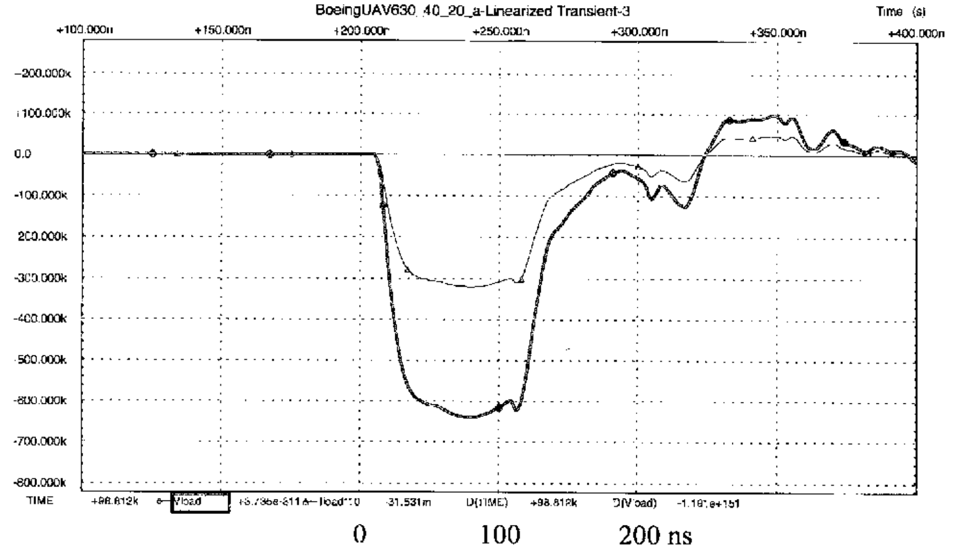

Fig. 5. Voltage and current (multiplied 10×) in the magnetron load computed using detailed version of the simplified circuit diagram in Fig. 4.

is that its characteristic impedance can easily be varied along its length to accommodate variations in load impedance.

Our compact pulse power system conceptual design mounts all of the subsystems and components in a frame that is 24 in. wide × 24 in. high × 144 in. long. The total weight of the system, including the frame, sealed enclosures, and internal mountings, is 1,420 lb. Figure 6 is a dimensioned diagram of the subsystem that is designed to fit into the port bay of an unmanned aerial vehicle (UAV) with the source in the starboard bay. Frames and supports have been omitted for clarity.

2.1.2. Compact PFN. Titan-PSD has designed, fabricated, and delivered the pulse power subsystems for most of the major electric gun facilities worldwide. These facilities were dedicated to electric gun research and for this reason were designed with a premium on reliability and ease of use of the pulse power subsystem. Compactness was a secondary consideration given the large amount of real estate available for the subsystem.

The 8.5-MJ stored energy subsystem, consisting of four separate PFNs (Fig. 1c), was the first design to use the high-energy-density, 2.5-J/cc (2.5-MJ/m^3) polyvinylidene fluoride (PVDF) Aerovox capacitors; the overall energy density of the PFNs is less than 0.5 MJ/m^3. This low energy density is partially due to the use of "jelly-roll" inductors whose large external fields forced the separation of components and resulted in a low component packing fraction.

Figure 7 is a photograph of a 250-kJ PFN that was constructed under a DTRA/Navy-sponsored ETC gun development project and that was incorporated for a time into the Army/Tank-Automotive and Armaments Command (TACOM) Combat Hybrid Power System (CHPS) system integration laboratory.[5] Titan-PSD designed this PFN module for compactness by developing innovative bus work and a closed-field inductor for use with the 2.5-J/cc PVDF capacitors. These advances in component design allowed us to achieve a

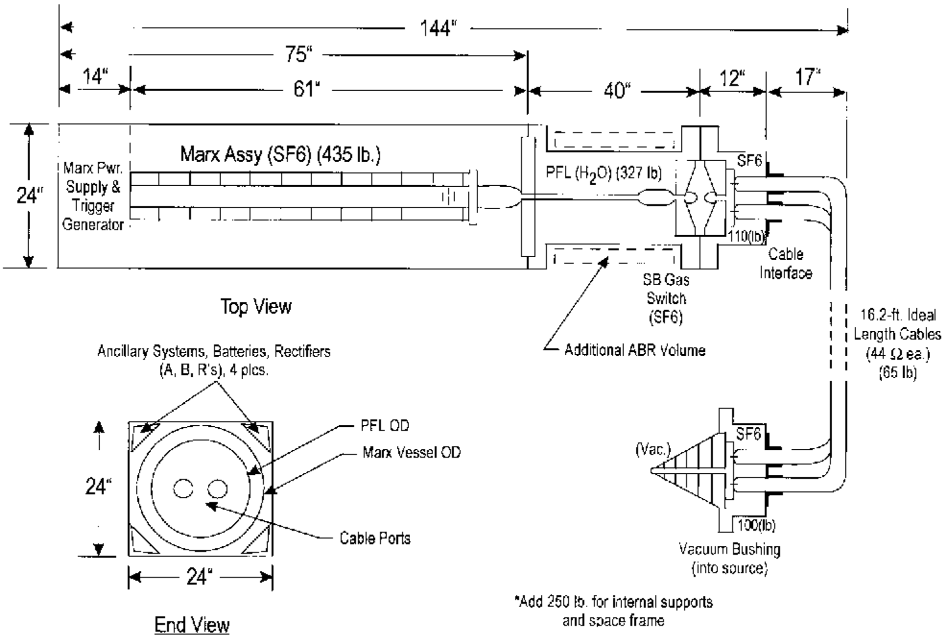
144"
75"
14"
61"
40"
12"
17"
24"
Marx Pwr. Supply & Trigger Generator
Marx Assy (SF6) (435 lb.)
PFL (H_2O) (327 lb)
SF6
110(lb)
Cable Interface
SB Gas Switch (SF6)
Additional ABR Volume
Top View
16.2-ft. Ideal Length Cables (44 Ω ea.) (65 lb)
Ancillary Systems, Batteries, Rectifiers (A, B, R's), 4 plcs.
PFL OD
Marx Vessel OD
Cable Ports
24"
24"
End View
(Vac.)
SF6
100(lb)
Vacuum Bushing (into source)
*Add 250 lb. for internal supports and space frame

Fig. 6. Top and end views of the pulse power subsystem concept design. The upper portion will be mounted in a frame and placed in one bay of a UAV. The vacuum bushing and the source will be placed in a second bay.

very high packing factor and an overall energy density of 1.25 MJ/m^3. This module designed in 1994 still represents the state of the art in PFN design. Minor variants to the design are under consideration today for EM gun applications.

Titan-PSD has estimated the increases in the overall energy density of small (250–500-kJ) PFN modules that could be realized by achieving specific advances in component technologies.[†] This includes the use of an existing, more compact vacuum output switch (developed by Titan-PSD) and existing, more compact diodes for the crowbar switches. We would design a higher-energy-density, closed-field inductor and eliminate the dump resistors and associated hardware. Rather than having to dump energy into a resistor, advanced pulsed power systems would be charged by four-quadrant converters capable of returning energy from the PFN to an intermediate energy store. When 5-J/cc film or ceramic, high-energy, reversal tolerant capacitors become available, the utilization of advanced components would lead to a PFN design with an energy density of 3.25 MJ/m^3.

2.1.3. Power supplies for laser applications. Titan-PSD's San Diego operations are actively engaged in building innovative high-average-power systems. Titan is currently delivering power conditioning and control elements for the high-power RF klystrons that will power the DOE Spallation Neutron Source that is now under construction at Oak Ridge, Tennessee, and a series of multimegawatt rectifiers for the U.S. Navy's Electro-Magnetic

---

[†] Higher energy densities could be achieved in the design of larger modules due to economics of scale.

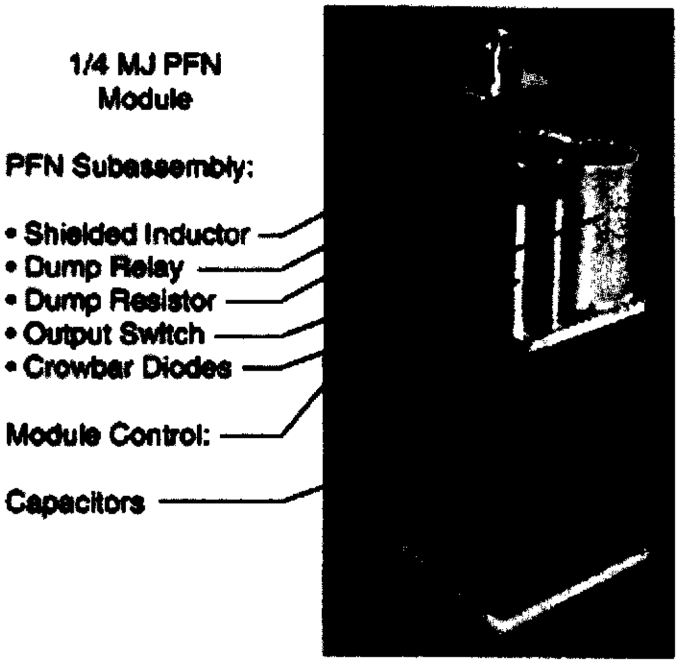

Fig. 7. Photograph of the 250-kJ PFN with the state-of-the-art, 1.25-MJ/m^3 overall energy density. Titan-PSD has operated this PFN into resistive loads at a rate of 1/3 Hz for extended periods. Operational limits were imposed by the heating of the load, not by the performance of any component of the PFN.

Aircraft Launch System (EMALS). Most recently, Titan has developed conceptual compact power supply designs for deployable pulse power systems for the Strategic Illumination Laser (SIL). The SIL is a diode-pumped, solid-state laser for use as a designator as a part of the U.S. Air Force's Missile Defense Agency's Airborne Laser (ABL). These power supplies will drive compact arrays of light-emitting diodes (LEDs), which will in turn pump the solid-state lasing medium.

Titan's proposed SIL power supply system will consist of a group of innovative ballast-type isolated power converter modules, each fed by its own active-power-factor-corrected rectifier module. By this method the input current harmonic distortion will be minimized, resulting in an overall system power factor of 0.995, which will be acceptable for the Boeing-747 type aircraft (ABL platform) power system. The whole power system will be protected against faults by appropriate fuses and circuit breakers. Initial size and weight estimates for the total power system is 65 liters of total volume and 140 kg of mass. The driving factor in the weight and volume is the need to meet the power factor requirements. Given a dedicated power source, the size and weight could be halved.

Two basic topologies have been examined for power factor correction. They are the "Vienna rectifier,"[8] first proposed by researchers in the Vienna Technical University, and the hex-bridge bidirectional inverter. Both topologies do not contain any magnetic components, operating at line frequency, and both can provide a power factor over 0.99. All merits and drawbacks of both topologies have been initially analyzed, and the Vienna approach is favored for this application.

For the ballast inverter Titan-PSD selected the variable-frequency, zero-power switching full bridge quasi-resonant inverter topology. It is new and until recently not well known but is very promising for multikilowatt power supplies. It shows excellent average to peak ratio for the semiconductor switches, close to the theoretical limit of 50%, and a good power transformer copper utilization factor, typically 90%. It is intrinsically output current limiting and therefore operates well in a current regulated mode and is tolerant to output short circuits. Using this converter topology allows the design to reach unprecedented power conversion efficiency (real value 95%) with excellent reliability.

2.1.4. Electronically reconfigurable battery. We consider a DEW power system in which a 10-kV capacitive store is required to operate repetitively. If the system operates at 3 pps, even for a short period of time, the recharge time would be about 300 ms and the recharge power would be about 500 $\mathrm{kW_{avg}}$ (1 $\mathrm{MW_{pk}}$). An important system issue is the size and weight of the 500-kW power conditioning unit.

Titan-PSD has developed a means to eliminate the need for intermediate power conditioning for many DEW systems mounted on hybrid electric vehicles by using the on-board batteries to charge the capacitive store directly. Our approach uses electronic switching to convert the battery modules in the mobility store from a parallel configuration providing vehicle load leveling and/or silent mobility capability, to a series configuration that delivers 500 kW at 10 kV to recharge the capacitor bank and back again. We refer to this as an "electronically reconfigurable battery" (ERB). Note that all of the batteries are recharged in parallel. A similar approach is used in capacitive Marx generators to achieve very high voltage pulsed output.

Figure 8 is a schematic drawing of the circuit of an ERB in a hybrid electric drive vehicle. In this schematic circuit, some battery modules are always in parallel and connected to the vehicle bus. Others, along the left-hand side of the drawing, can be switched between a parallel configuration, which supports the vehicle bus, and series operation for charging the DEW store. Electronic reconfiguration requires three switches per battery module. All switches, with the exception of the output switch, need only block the voltage of a single battery module and open at near zero current (characteristic of capacitor charging) in normal operation. Passive snubbing controls the transient conditions during erection and de-erection. Reconfiguration can be accomplished in less than 1 ms using off-the-shelf, solid-state switches such as IGBTs or MOSFETs.

Figure 8 shows an erectable battery module charging a Guillemin E-type network. Switches 2 and 3 are battery isolation switches, and switch 4 is a series switch. These switches are rated only for the module voltage (1 kV in this case). Mechanical contactors parallel the isolation switches for better efficiency during periods when the pulse power system is not in use. Switch 5 is the output switch, which is rated for the full output of the ERB (10 kV and 50 A) and is most likely a series stack of the same switches used for 2–4. For fault protection and charge interrupt at current, a vacuum contactor and fuse are placed in series with the output switch. All switches are optically isolated with gate power drawn from their adjacent modules. Switch 6 is a high-voltage and current-closing switch, either vacuum or solid state.

Assuming a 20-ton-class vehicle and extrapolating from CHPS requirements, we have developed a conceptual design of an ERB for a hybrid electric vehicle with a system meeting the following requirements: 1) Deliver up to 400 kW to the vehicle bus at 1 kV in parallel operation; 2) recharge a 150-kJ capacitive store to 10 kV in 300 ms; and 3) support 30–45-min silent operation at 80 kW. These requirements mandate the use of very high energy

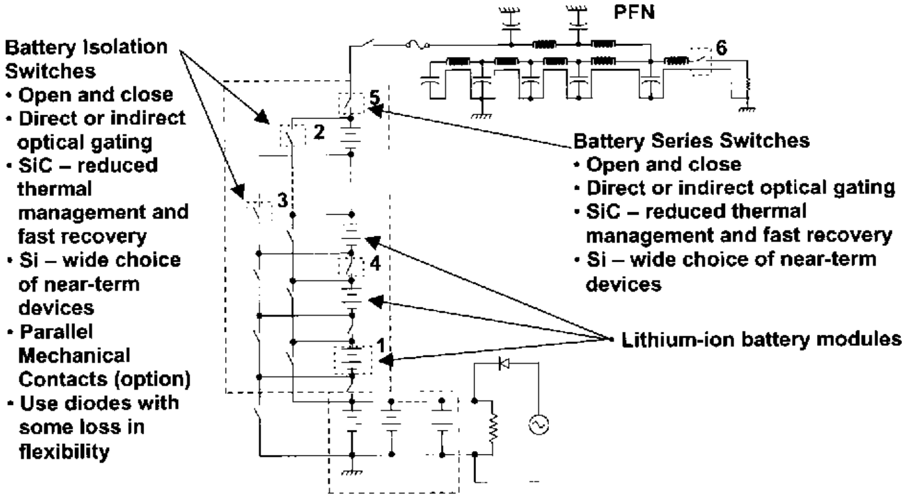
PFN
6
Battery Isolation Switches
• Open and close
• Direct or indirect optical gating
• SiC – reduced thermal management and fast recovery
• Si – wide choice of near-term devices
• Parallel Mechanical Contacts (option)
• Use diodes with some loss in flexibility
5
2
3
4
1
Battery Series Switches
• Open and close
• Direct or indirect optical gating
• SiC – reduced thermal management and fast recovery
• Si – wide choice of near-term devices
Lithium-ion battery modules

Fig. 8. Schematic circuit of the ERB. Shown is a generic circuit topology for charging a Guillemin E-type network. For the present application, the 10 modules in the dynamic store are all erected in series to produce a 10-kV output. A feature of the ERB is that the dynamic portion of the store may be erected in combinations of series/parallel modules to provide any output voltages that are an integer multiple of the DC bus voltage. This voltage agility, in theory, would allow a single dynamic store to power several loads with different input voltage requirements on the same platform such as ETC guns and HPM systems.

and power density batteries such as the HP series of lithium ion batteries being developed by SAFT and the lithium polymer batteries produced by Ultralife Batteries, Inc., for use in cell phones. The SAFT batteries have a slight advantage in usable power density and packaging for military use, whereas the Ultralife batteries have an advantage in cost due to volume production and a potential for more compact packaging (thin, rectangular) in ERB service. A design and trade study will select the most appropriate cell for the application. We chose the Ultralife cells for the conceptual design used for purposes of discussion in this paper because of our experience with them in our laboratory tabletop ERB and because of the availability of these cells over a range of capacities from 120 mA-h to 3.5 A-h.

The ERB system in this case constitutes only 1/3rd of the total battery for erected (dynamic) operation. The remaining 2/3rds of the store (static) is dedicated to load leveling and silent mobility. The total capacity of ~290 MJ (80 kW-h) accommodates silent mobility requirements. With only 1/3rd (90 MJ, 25 kW-h) of the total capacity configured for on-command electronic erection and de-erection, the vehicle energy storage system maintains its load leveling and silent operation capability even when the pulse power system is active.

For our conceptual design, we assemble 53 Ultralife polymer batteries into individual 200-V stacks. Five of these stacks are placed in series to obtain an output voltage equal to that of the vehicle bus. In the static portion of the store, 20 such series stacks of the UBC34106102 cells (5,300 total cells) are needed to provide the 200-MJ nominal capacity. In the dynamic portion of the store, the UBC383562 cell is used because of its higher current rating. Due to the smaller size of these cells, there can be as many as 4 in parallel by 53 in series per 200-V unit, and five such units will form an erectable module. The module will incorporate all necessary switches and isolation and thermal management hardware.

Ten modules will constitute the entire dynamic store (~90 MJ, 10,600 total cells). The total battery volume in the static and dynamic stores is ~0.282 m^3, and the accessories are expected to add another 0.4–0.6 m^3 for a total volume of less than 1 m^3. The weight for just the batteries would be ~572 kg, with accessories of much lower density adding another 100–200 kg, for a total weight of less than 700 kg. Charging of the ERB is always performed at bus voltage with the dynamic store in parallel configuration.

In load leveling or silent mobility operation, the charge state of the ERB is continuously monitored and controlled by the vehicle's systems. When the pulse power mode is activated, the mechanical contactors bypassing the isolation switches are opened. Next, all isolation switches are opened, followed by the closing of all series switches, and the ERB begins charging the capacitive store. Initially, the load looks like a short circuit, and the charging current is limited by the internal impedance of the ERB cells. As the charge on the bank increases, the charging current diminishes, and the cell voltage begins to rise. Near full charge, the current approaches zero and the ERB output voltage rises to its maximum of ~10 kV. Voltage on the store is monitored and is regulated by chopping the ERB output (i.e., opening and closing the series and output switches as necessary to maintain charge against leakage; this has been demonstrated in the laboratory at rates of up to 1 kHz). When the store discharges, the output switch opens immediately and remains open long enough for the closing switch to recover (1–10 ms). The ERB can remain erected for as long as the pulse power system is active. Recharging of the store is then initiated by reclosing the ERB output switch. When the pulse power system is deactivated, the dynamic store is returned to its parallel configuration by opening the series switches (de-erecting). The dynamic store can then be recharged from the vehicle power bus.

We have created an ERB laboratory demonstration consisting of five erectable modules of 40 V. The Ultralife UBC383562 cells were used and arranged on PC boards that were in turn stacked via standoffs. Isolation and series switching was via optically coupled IGBTs. Some figures of merit (FOM) from the ERB testing include a demonstrated current rise time to 80 A of 4 μs at 200 V and an erection/de-erection frequency of 1 kHz for use in voltage regulation. A FOM from the four series cell tests is a demonstrated current rise time to 120 A of 1 μs, in a maximum power transfer configuration.

2.1.5. Intermediate energy storage. In the preceding section Titan-PSD identified a need for and established the technical feasibility of an advanced technology power conditioning element to charge capacitive stores in DEW systems. The ERB is the solution with the highest efficiency and energy and power density possible. This is critical in order to realize a manageably sized DEW system. A second similar, critical issue in the design and use of pulsed electrical weapon systems involves the level of available platform power. If this power is insufficient to support the required operational burst length and repetitive fire rate of the system, then it may be necessary to incorporate intermediate energy storage into the system design. Intermediate storage could be practically achieved through the use of secondary batteries, flywheels, and ultracapacitors (also referred to as electric double-layer capacitors or EDLCs). Figure 9 is a schematic diagram of intermediate energy stores in two DEW systems.

Figure 10 shows a present application of an EDLC energy storage system in providing ride-through of voltage sags and momentary outages for an induction motor. The output of the store is connected across the direct current (DC) link of the asynchronous drive and supports its voltage. Titan-PSD has manufactured a ride-through system that has been demonstrated to respond to sags and outages within 10 ms. It can provide

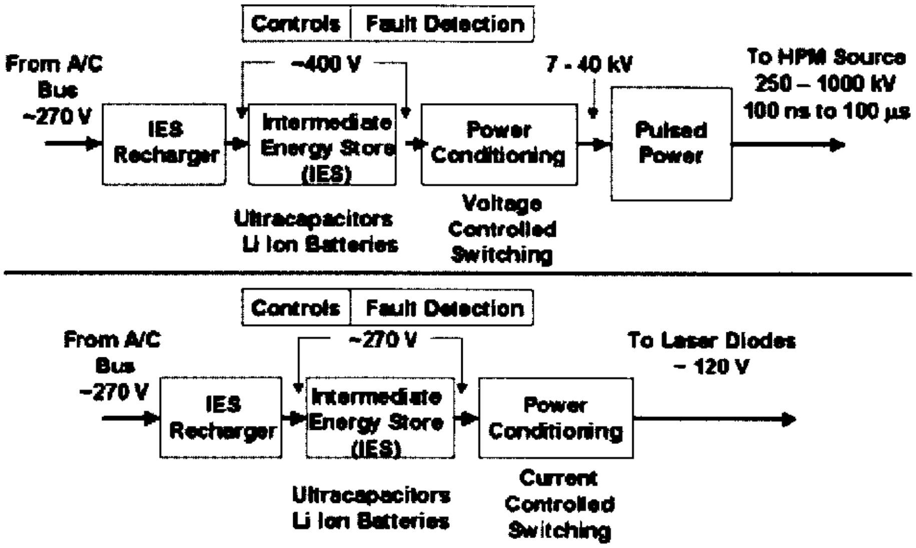
Controls
Fault Detection
From A/C Bus ~270 V
IES Recharger
~400 V
Intermediate Energy Store (IES)
Power Conditioning
7 - 40 kV
Pulsed Power
To HPM Source 250 – 1000 kV 100 ns to 100 μs
Ultracapacitors Li Ion Batteries
Voltage Controlled Switching
Controls
Fault Detection
From A/C Bus ~270 V
IES Recharger
~270 V
Intermediate Energy Store (IES)
Power Conditioning
To Laser Diodes ~ 120 V
Ultracapacitors Li Ion Batteries
Current Controlled Switching

Fig. 9. Two DEW block system schematics with intermediate energy storage.

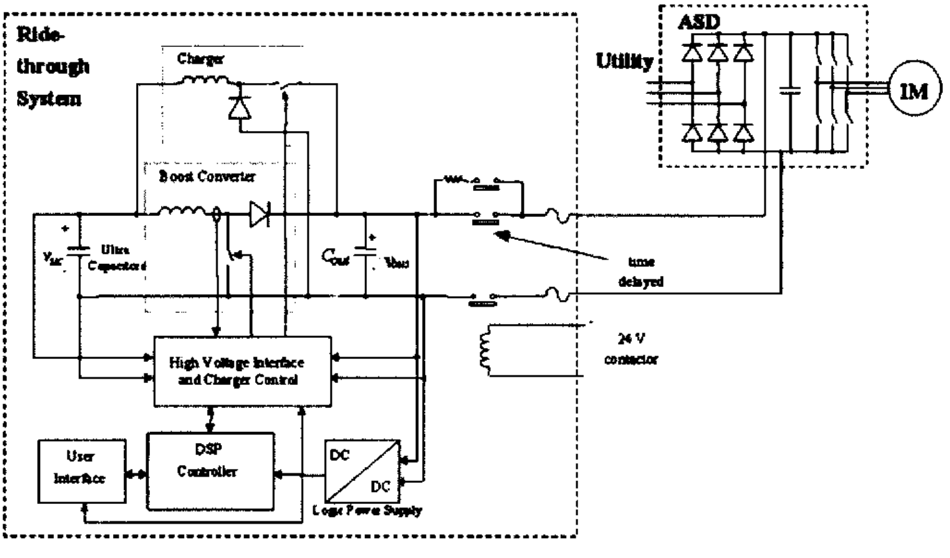
Ride-through System
Charger
Boost Converter
Ultra Capacitors
V_{uc}
C_{out}
V_{out}
High Voltage Interface and Charger Control
User Interface
DSP Controller
DC
DC
Logic Power Supply
Utility
ASD
IM
time delayed
24 V contactor

Fig. 10. Circuit diagram of an EDLC ride-through system supporting the DC link of an asynchronous drive (ASD).

100 kW for up to 5 s, which is sufficient to allow for orderly shutdown in a prolonged outage.

Titan-PSD has also developed several concept designs of EDLC energy storage systems for military applications. For example, one EDLC system design would be capable of supporting the DC link in a Navy shipboard system at 480 V for 12.5 s at 250 kW. This system is estimated to be 50% smaller and lighter than a corresponding flywheel system (i.e., the Caterpillar UPS-250 system). With minor modifications, the same system (or several systems in parallel) could be used as an intermediate energy store for DEW systems.

2.2. Titan advanced component developments to enable DEWs

2.2.1. Solid-state switches for severe service. The key enabling technology for nearly all pulse power systems for DEW applications is the switch that initiates and delivers the pulsed electrical output from the electrical energy store to the load. This switch must hold off high voltage (typically ≥ 10 kV), close rapidly (<1 μs), carry high peak current (100–500 kA) with high initial rate of rise (50–100 kA/μs), sustain a high total action ($\geq 10^7$ A^2-s), and recover rapidly (<10 ms). In addition, the switch must support high current density, operate reliably under mechanically stressing conditions, have minimum weight and volume claims, and not require ancillaries such as vacuum pumps or gas flow subsystems. At present, the hydrogen thyratron is most frequently selected to meet these simultaneous requirements, but thyratrons are large and heavy devices, with auxiliary equipment that consumes several kilowatts and often require immersion in oil. The Army Research Laboratory has been actively seeking solid-state devices for DEW utility[7] and Titan-PSD is weighing in by developing a high fluence, laser gated and pumped (silicon) thyristor (LGPT) that will satisfy these requirements.

High fluence optical gating of thyristor switches has been chosen for numerous pulsed power switching applications. Light activation with high fluence laser sources enables the use of solid-state switching devices in performance parameter spaces previously serviced only by vacuum and gas spark gaps. If the output from the high fluence source is properly distributed within the switching medium, the di/dt capability of the switch does not depend on the rate of plasma spreading. Rather, di/dt becomes a function of the rate of rise of the optical fluence. In addition, optical gating eliminates the extensive and intrusive physical gate structure that limits the conduction area of the highest di/dt conventional solid-state devices. Properly designed laser-gated devices are typically capable of higher di/dt, peak current, and charge transfer for a given silicon area than conventional solid-state devices.

The band edge for intrinsic silicon is ~1.12 eV, corresponding to ~1,111 nm. However, silicon is an indirect band-gap material so that the optical phonon energy (0.063 eV) adds to the photon energy to create e–h pairs out to 1,170 nm. In most previous efforts involving direct laser-gating of silicon devices, a solid-state laser was used; typically either Nd:YAG or Nd:YVO at 1,064 nm. However, over the past 10 years, the development of compact diode-pumped solid-state lasers has significantly reduced the size of the requisite optical sources and improved their lifetime and reliability. More important, the InGaAs laser diode bars that were developed for pumping these solid-state lasers have been pushed to longer wavelengths, approaching and exceeding that of the silicon band edge. In principle, the center wavelength of InGaAs laser diode bars is selectable by design over a continuum from 850 nm through at least 1,150 nm. This allows direct optical gating and pumping of silicon by the diodes without any of the intermediary options of the preceding paragraph. Figure 11 is an illustration of an on-board laser diode gated and pumped silicon thyristor, which is one of the candidate approaches in development now.

Titan-PSD is currently performing on two programs and has finished a third that provide technical leverage for further switch development. Our Electra Advanced Pulsed Power Program is a five-year, Naval Research Laboratory (NRL)/DOE-sponsored effort to develop advanced pulsed power components and systems for KrF laser IFE (krypton fluoride laser inertial fusion energy). The most critical component is the primary switch, and we are

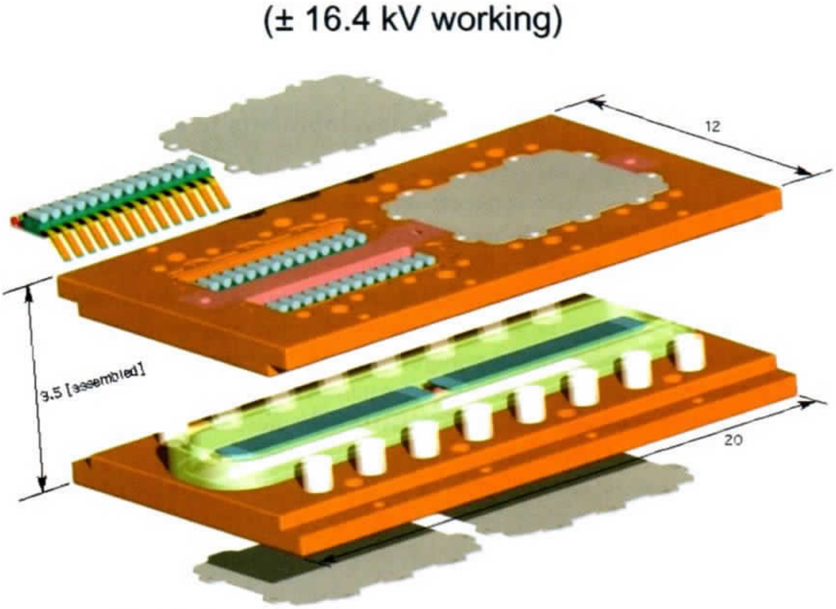

Fig. 11. Illustration of a laser gated and pumped thyristor.

developing an on-board LGPT for this role. The switching requirements are repetitive, high peak current and di/dt, as well as high efficiency (the switch dissipates $<1.5\%$ of stored energy) and long lifetime (10^9 shots). The LGPT is to be used in a Marx generator, and the single device operating parameters are 16.4-kV working voltage, current density 2.25 kA/cm^2, peak current 225 kA, max di/dt 900 kA/μs, pulse width 800 ns, and repetition rate 5 pps continuous. The lifetime requirement limits the thermal excursion per pulse to a few degrees, and therefore the action per square centimeter is low ($\sim$2 A^2-s).

Another current Titan-PSD program is to develop compact pulsed power components for the Air Force Research Laboratory (AFRL) at Kirtland Air Force Base. Components under development enable a compact, long-lifetime Marx-PFN driver for a narrowband HPM source. The di/dt requirement is a few hundred kA/μs/cm^2(Si) or 2 MA/μs total, with a peak current of 30 kA. To satisfy the switching requirements for this application, we are considering the LGPT, a solid-state laser activated switch, and a hybrid combination of the two.

Although the action and charge transfer in these Electra and AFRL applications are substantially less than that found in some pulse power applications, they serve to illustrate our confidence in the high power laser diode bars for high di/dt service. In both cases, the pulse width of the energy transfer is short enough that thyristor action has barely begun by the time the pulse is over. Because of this, a large fraction of the total charge conducted during the pulse must be supplied optically. The diode laser bars and their drive circuitry are capable of supplying an illumination fluence of up to 6 kW/cm^2(Si) within 20 ns and sustaining that level for at least several microseconds. Fluence of several kilowatts per square centimeter is desirable for initial gating in many DEW applications. The required fluence drops to 250–500 W/cm^2 for continuous pumping.

Another effort was recently undertaken in a conceptual design study for an optically gated and pumped switch for rectification of compulsator (pulsed alternator) output for driving an electromagnetic gun. This completed study was performed by Titan-PSD and the Optiswitch Technology Corporation for the Institute for Advanced Technology (IAT). The on-board LGPT design was chosen because of the customer's desire to reduce the overall device count and silicon area compared with conventional devices. Specifically, we designed a 16.7 kV, 100–115-cm^2 asymmetric (blocks voltage forward direction only) thyristor and series diode. Depending on the allowable temperature excursion per shot, our approach replaces ~80 conventional devices with 3–5 series pairs (6–10 devices total; 340–500 cm^2 active) for the same forward losses. This is possible because continuous pumping of thick (2.5-mm), high-voltage (15–20-kV) devices with photo-carriers mitigates the carrier population deficit in the n-base. Electrically gated devices of the same voltage capability typically exhibit huge commutation and conduction losses and are therefore impractical for this type of service. Reverse blocking was handled by a series optically pumped diode. Separating the forward and reverse blocking functions allowed more freedom in designing for reverse recovery, a feature deemed necessary due to repeated failures of symmetric (blocks voltage both directions) devices in this service. Laser diode pumping served to reduce the forward losses in the series diode element as in the thyristor.

The switch was designed to handle a peak current of 1.6 MA (3.3–4.7 kA/cm^2) for 600 μs when installed in a half-wave, four-phase rectifier. Nine total cycles are rectified per shot of the electromagnetic gun. The initial design (500 cm^2 active) limited the temperature excursion per shot to ~100°C, commensurate with a service life of ~10^4 shots. In the extreme, the current density could be pushed to 4.7 kA/cm^2 and optical pump power doubled, resulting in a ΔT approaching 180°C and ~200-shot lifetime. The laser diode bars in the initial case provided a constant 500 W/cm^2 of silicon for each 600-μs conduction phase. This goes to 1 kW/cm^2 in the extreme case.

2.2.2. Next generation reltron microwave generator development. Reltron HPM sources were invented and developed by R. Bruce Miller of Titan Advanced Innovative Technologies from the mid-1980s through 1998. The PSD of Titan assumed responsibility for reltron contracts in 1998. Projects included one complete HPM simulator for a government research laboratory in Germany and eight reltron tubes of various specifications. In developing these eight tubes, enough mechanical and electrical modifications have been included to justify their designation as "second-generation" designs.

Reltrons are both compact and efficient. The electron beam is highly modulated over ~10 cm and postaccelerated prior to drifting to the output cavities and beam collector. This beam transport requires only ~75-cm overall length for L-band tubes (Fig. 12). The output cavities are incorporated into a standard rectangular waveguide. Microwave power is efficiently extracted (~80%) directly in TE$_{10}$ mode. The overall efficiency (peak electrical to microwave power) is 30–40%.

In general these new tubes move the reltron technology base in the directions of longer pulsewidths (~1–2 μs), increased frequency tunability (~±13%), lower driver voltages (500–600 kV), and higher repetition rates (~10 Hz for demountable versions, ~300 Hz for sealed tubes). Second-generation improvements include a new high voltage insulator design, new movable current contacts for the modulation and output cavity tuners, reduced outgassing, and monolithic grids. The frequency range for these eight tubes covers 700–1,450 MHz. Programs are in place now to extend this range.

In addition to the objective to extend the frequency coverage, efforts are underway to improve beam optics to reduce current losses and thereby improve efficiency and microwave output power. A new beam current diagnostic package has been recently implemented to

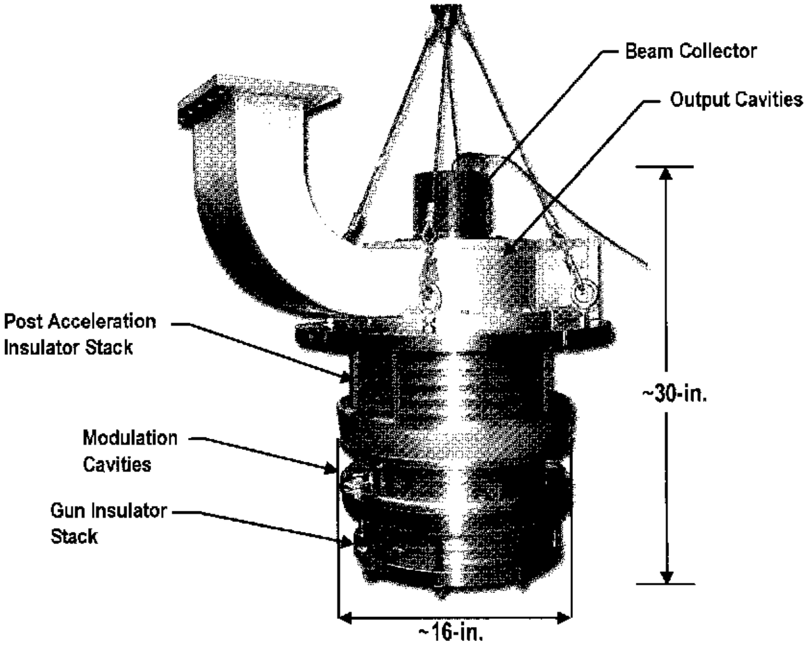

Fig. 12. L-band reltron capable of generating flat 200-MW, 900-ns pulses with frequency stability $\delta f/f < 0.3$. This generation of reltrons is capable of RF pulse widths of ~300 RF cycles, repetitive pulse operation from single-shot to 10 Hz, and ±10% continuous-frequency tunability.

better infer basic performance parameters and to assist in comparing benchmarking simulations. These calculations simulate performance in the actual geometry and are generated with a three dimensional particle-in-cell code (the LSP code from Mission Research Corporation). Measured waveforms generated by a recently delivered L-band tube are shown in Fig. 13. This tube produces ~185 MW in flat, 900-ns pulses with excellent frequency stability. The LSP code was essential during this tube's development. A new era in microwave tube design has emerged with the development of accurate three-dimensional models.

2.2.3. HEM magnetron. Backdoor microwave effects on electronics systems in modern weapons fall into three broad categories. Two categories are characterized by high peak power density and microwave formats that overwhelm the target system with an electric field that is high enough to cause circuit upset or burnout. In the third category, effects are caused by coupling spurious signals into the target system, akin to classical smart jamming, interference or spoofing[‡] techniques used by the EW countermeasures community. The DEW microwave pulse formats appropriate to the latter class of effects are generally distinguished from upset and damage pulse formats by lower peak power density thresholds but more specific requirements on frequency and amplitude modulations and

---

[‡] Low incident microwave levels can cause "interference" with the operation of threat weapon electronic systems, producing high levels of noise in sensitive circuitry to "mask" true signals. The incident microwave signal is said to "spoof" if it is tailored to "mimic" the true signal and to thereby produce false commands.

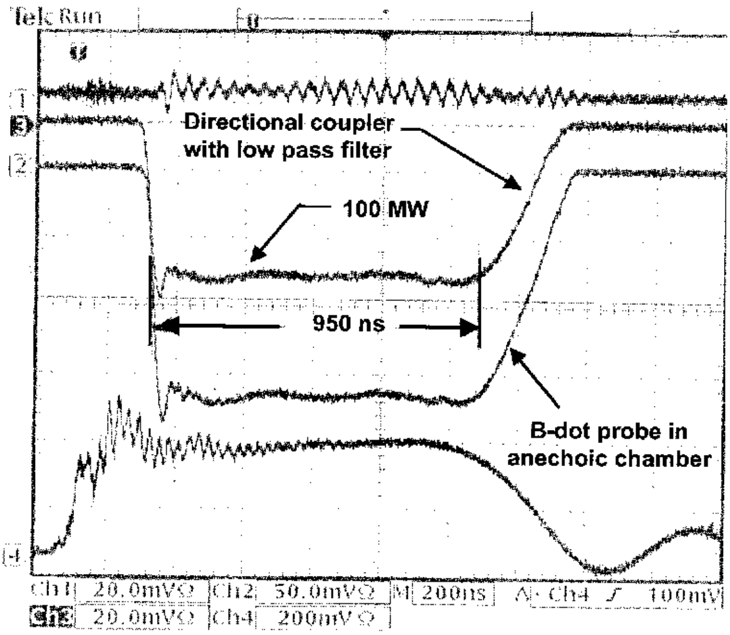

Fig. 13. Waveforms from the 185-MW L-band reltron: the top trace is the heterodyne (l.o. = 1.3 GHz) showing excellent frequency stability, and the second trace is the steady, flattop output power measured with a calibrated directional coupler. (The data are conditioned with a low-pass filter; higher harmonic content is know to be more than 13 dB below the primary L-band signal.) The third trace is the radiated power monitored with an uncalibrated, free-field B-dot, and the last trace is an uncalibrated profile of the Marx current.

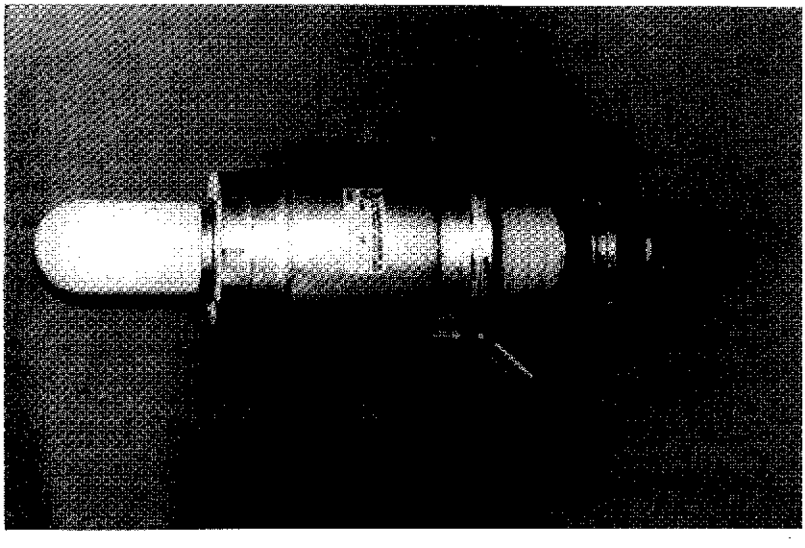

Fig. 14. Titan's L-band HEM magnetron. This is the highest-average-power magnetron that is commercially available.

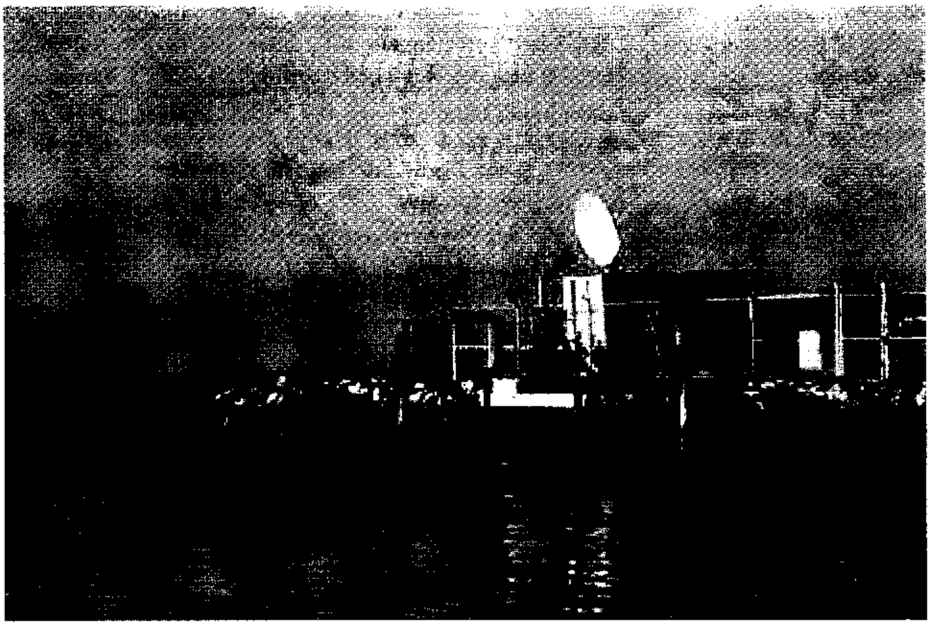

Fig. 15. The principal system elements of the ATD system are a power supply, a modulator, an HEM source, and an antenna. These components are supported by a control system and a cooling system. For the DTRA ATD ground-based system the power supply and operators' console(s) are housed in a shelter mounted on one flatbed trailer while the remaining components are mounted on another flatbed trailer (out of the field of view to the right). The system was designed to broadcast 2-MW peak power pulses with an average power of 300 kW for 30-s missions. This yields a peak fluence of 1 mW/cm^2 at a range of 2,400 m.

repetition rate. All three categories of effects can be induced by DEW systems based on high-power RF, narrow-band, pulsed HPM or long-pulse, high-energy microwaves (HEMs).

Titan has developed an HEM laboratory source and an HEM advanced technology demonstration (ATD) source for DEW purposes at its California Tube Laboratory. These HEM sources have efficiencies of approximately 85%. This high efficiency minimizes prime power and cooling requirements, thus significantly minimizing overall DEW system size and weight. The HEM magnetron, shown in Fig. 14, has produced 900-kW peak and 150-kW average powers and can easily be upgraded to 300-kW average power. The unique modulation characteristics of this magnetron can produce modulation effects of 30 MHz or greater in the target systems. More operational details are given in Table 6.

The ATD system shown in Fig. 15 combines the outputs of two 150-kW-average-output power magnetrons. The maximum peak output power obtained to date is 1.2 MW. The magnetrons with their respective magnet yokes each weigh about 200 lb, and each takes up about 4 ft^3. When combined their efficiency is over 76%, therefore together they absorb only 95 kW and are actively cooled. Variants of this RF system are suitable for Army, Navy, and Air Force missions.

Table 6. California tube laboratory HEM magnetron operational specifications

Parameter	Value
Frequency, fixed	915 MHz (890–920 MHz possible as well)
Power output	
Continuous wave	300-kW minimum
Pulse	600-kW peak (900 actually achieved)
	300-kW avg (nominal 50% duty)
Pulse width	10s of microseconds to 5 ms
Peak anode voltage	42 kV
Peak anode current	17 A
Beam efficiency	88% typical
Water-cooled anode	20 gpm typical
Filament	12 VAC, 200-A typical standby

3. Conclusions and Significance

The first laser and HPM DEWs are just now being deployed, and new threats emerging in the information age will sustain the need for continued DEW development and acceptance. The deployment of even more advanced DEW systems and in fact the DoD's intent to field more (or all) electric fighting platforms depend critically on the availability of compact, lightweight, efficient pulsed power subsystems and long-lived, reliable advanced power conditioning and front-end components.

Titan has been supporting U.S. directed energy programs by providing services and analysis and developing advanced technologies since the early 1980s. Today, Titan is the only U.S. industrial supplier of commercial HPM generators. We have advanced pulsed power and power electronics technology development programs involving intermediate and primary electrical energy storage, solid-state switching, power conditioning, electrical-to-microwave conversion, and novel microwave antennas underway now. Titan is focused on providing pulsed power and C4ISR (command, control, communications, computers, intelligence, surveillance, and reconaissance) solutions for national defense and is intent on maintaining its industrial leadership in military information technologies.

Acknowledgments

The authors recognize and acknowledge contributions made by Jim Benford, Jerry Levine, R. Bruce Miller, and David Blank to the development of the primary narrow-band microwave sources that were the precursors of much of the HPM and HEM technology that Titan offers today.

References

[1] Ashby, S., D. Drury, G. James, P. Sincerny, and L. Thompson, "CLIA—A Compact Linear Induction Accelerator System," *Proceedings 8th International IEEE Pulsed Power Conference*, San Diego, CA, p. 940 (1991).

[2] Bekifi, G., and T.J. Orzechowski, Phys. Rev. Lett. **37**(6) 379 (1976).

[3] Benford, J., D. Price, H. Sze, and D. Bromley, J. Appl. Phys. **61**(5), 2098 (1987).

[4] Benford, J.N., N.J. Cooksey, J.S. Levine, and R.R. Smith, IEEE Trans. Plasma Sci. **21**(4), 388 (1993).

[5] Bhasavanich, D., C.M. Gilman, H.G. Hammon, and K. Nielsen, "80 kW, 1/3 Hz Repetitive-Fire PFN for Electrothermal Launchers," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 155 (1995).

[6] Hammon, J., S.K. Lam, and S. Pomeroy, "A Transportable 500 kV, High Average Power Modulator with Pulse Length Adjustable from 100 ns to 500 ns," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 429 (1995).

[7] Kingsley, L.E., R. Pastore, and H. Singh, "Solid-State Power Switches for HPM Modulators," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 65 (1995).

[8] Kolar, J.W., and F.C. Zach, "A Novel Three-Phase Interface Minimizing Line Current Harmonics of High-Power Telecommunications Rectifier Modules," *Record of the 16th IEEE International Telecommunications Energy Conference*, Vancouver, Canada, p. 367 (1994).

[9] Levine, J.S., N. Aiello, J. Benford, and B. Harteneck, J. Appl. Phys. **70**(5), 2838 (1991).

[10] Levine, J.S., B. Harteneck, and H.D. Price, *Intense Microwave Pulses III*, SPIE Vol. 2557, p. 74 (1995).

[11] Miller, R.B., K.W. Habiger, W.R. Beggs Jr., and J.R. Clifford, "Advances in Super-Reltron Source Development," *Intense Microwave Pulses III*, SPIE Vol. 2557, p. 2 (1995).

[12] Miller, R.B., C.A. Muehlenweg, K.W. Habiger, J.R. Smith, and D.A. Shiffler, *Intense Microwave Pulses II*, SPIE Vol. 2154, p. 99 (1994).

[13] Price, D., D. Fittinghoff, J. Benford, H. Sze, and W. Woo, IEEE Trans. Plasma Sci. **16**(2), 177 (1988).

[14] Price, D., J.S. Levine, and J. Benford, "ORION-A Frequency-Agile HPM Field Test System," *Seventh National Conference on High Power Microwave Technology*, Laurel, MD (1997).

[15] Smith, I.D., D.W. Morton, D.V. Giri, H. Lackner, C.E. Baum, and J.R. Marek, "Design, Fabrication and Testing of a Paraboloidal Reflector Antenna and Pulser System for Impulse-Like Waveforms," *Proceedings 10th International IEEE Pulsed Power Conference*, Albuquerque, NM, p. 56 (1995).

[16] Smith, R.R., J. Benford, D. Harteneck, and H.M. Sze, IEEE Trans. Plasma Sci. **19**(4), 628 (1991).

[17] Sze, H., J. Benford, W. Woo, and B. Harteneck, Phys. Fluids **29**(11), 3873 (1986).

The Tom Bearden Website

TOWARDS A NEW ELECTROMAGNETICS
PART I: SOLUTION TO TESLA'S SECRETS AND THE SOVIET TESLA WEAPONS

Table of Contents

Foreword **NEW**

Starting Pages **NEW**

Soviet Tesla Weapons

References

The Tom Bearden Website

TABLE OF CONTENTS

Acknowledgements

Foreword

Part I

Solutions to Tesla's Secrets and the Soviet Tesla Weapons

References

Part II

Reference Articles for Solutions to Tesla's Secrets

Bottone, A., "Nikola Tesla's New Wireless", *Electrical Engineer*, London, Dec. 24, 1909.

"Tesla's New Device Like Bolts of Thor", *N. Y. Times*, Dec. 8, 1915.

Secor, H. Winfield, "Tesla's views on Electricity and the War", *Electrical Experimenter*, Aug., 1917.

Tesla, Nikola, "The Effect of Statics on Wireless Transmission", *Electrical Experimenter*, Jan., 1919.

Gernsback, H., "Underground Wireless", *Electrical Experimenter*, March, 1919

Secor, H. Winfield, "America's Greatest War Invention, The Rogers Underground Wireless", *Electrical Experimenter*, March, 1919.

Tesla, Nikola, "Tesla on High Frequency Generators", *Electrical Experimenter*, April, 1919.

Tesla, Nikola, "The True Wireless", *Electrical E:x:perimenter*, May, 1919.

Tesla, Nikola, "My Inventions, v. The Magnifying Transmitter", *Electrical Experimenter*, June, 1919.

"The Rogers Underground Aerial For Amateurs", *Electrical Experimenter,*
June, 1919.

"A Giant Eye to See Round the World", *Albany Telegram,* Feb. 25, 1923.

Tesla, Nikola, "'World System' of Wireless Transmission of Energy",
Telegraph and Telephone Age, Oct. 16, 1927.

Secor, H. Winfield, "Tesla Maps Out Our Electrical Future", *Science and Invention,* April, 1930.

Tesla, Nikola, "Man's Greatest Achievement, *N.* Y. *American,* July 6, 1930.
"Tesla at 75", *Time,* July 20, 1931.

Gernsback, Hugo, "Faster Than Light:", *Everyday Science and Mechanics,*
Nov., 1931.

"No High Speed Limit, Says Tesla", *The Literary Digest,* Nov. 7, 1931.

O'Neill, John J., "Tesla Cosmic Ray Motor May Transmit Power 'Round Earth",
Brooklyn Eagle, July 10, 1932.

"Tesla, 76, Reports His Talents At Peak", *N. Y. Times,* July 10, 1932.

Tesla, Nikola, "Pioneer Radio Engineer Gives Views on Power", *N. Y. Herald*
Tribune, Sept., 11, 1932.

Bird, Carol, "Tremendous New Power Soon to be Unleashed", *Kansas City*
Journal-Post, Sept. 10, 1933

"Tesla 'Harnesses' Cosmic Energy", *Philadelphia Public Ledger,* Nov. 2, 1933.

"Tesla Sees Evidence Radio and Light are Sound", *N.* Y. *Times,* April 8, 1934.

"Invents Peace Ray", *N. Y. Sun,* July 10, 1934.

"Death-Ray Machine Described", *N. Y. Sun,* July 11, 1934.

"Tesla, at 78, Bares New 'Death-Beam'", *N. Y. Times,* July 11, 1934.

Alsop, Joseph W., Jr., "Beam to Kill Army at 200 Miles, Tesla's Claim",
N. Y. Herald Tribune, July 11, 1934

Tesla, Nikola, "Tesla on Power Development and Future Marvels", *N. Y. World*
Telegram, July 24, 1934.

Wilshimer, Helen, "Dr. Tesla Visions the End of Aircraft in War", *Every*
Week Magazine, Oct. 21, 1934.

"A Machine to End War", by Nikola Tesla as told to George Sylvester Viereck,
Liberty, Feb., 1935.

"Tesla Predicts Ships Powered by Shore Beam", *N. Y. Herald Tribune,*
June 5, 1935.

"3 Tesla Inventions", *N. Y. Sun,* July 9, 1935.

Sparling, Earl, "Nikola Tesla, at 79, Uses Earth to Transmit Signals;
Expects to Have $100,000,000 Within Two Years", *N. Y. World Telegram,*
July 11, 1935.

"Tesla's Controlled Earthquake", *N. Y. American*, July 11, 1935.

"Tesla, 79, Promises to Transmit Force", *N. Y. Times,* July 11, 1935.

Tesla, Nikola, "Expanding Sun will Explode Someday Tesla Predicts",
N. Y. Herald Tribune, Aug. 18, 1935.

"Tesla, 80, Reveals New Power Device", *N. Y. Times*, July 11, 1936.

"Sending of Messages to Planets Predicted by Dr. Tesla on Birthday",
N. Y. Times, July 11, 1937.

"Tesla Has Plan to Signal Mars", *N. Y. Sun*, July 12, 1937.

"In the Realm of Science: Tesla, Who Predicted Radio, Now Looks Forward
to Sending Waves to the Moon", *N.* Y. *Herald Tribune,* Aug. 22, 1937.

"Aerial Defense 'Death-Beam' Offered to U. S. By Tesla", *Baltimore Sun,* July 12, 1940.

"'Death Ray' for Planes", *N. Y. Times,* Sept. 22, 1940.

"Proposing the 'Death Ray' for Defense", *Philadelphia Inquirer,* Oct. 20, 1940.

Swezey, Kenneth M., "Nikola Tesla", *Science,* May 16, 1958.

"Cableless Power", *Electronics* (Digest), July 22, 1960.

"Fireballs for Defense?", *Christian Science Monitor,* Feb. 7, 1962.

Lagus, George W., "The Use of Ground in High-Frequency Circuits", *Radio-TV and Science and Electronics,* June-July, 1969.

Jamison, S. L., "Life Energy", *Probe, The Unknown,* June, 1973.

Jueneman, Frederic B., "The First Law of Thermodynamics", *Industrial Research,* Feb., 1974.

Wade, Nicholas, "Charged Debate Erupts over Russian Beam Weapons", *Science,* May 27, 1977.

Benford, Gregory, "Zeus in Orbit", *Omni,* Sept., 1981.

JAMES HARRIS ROGERS' PATENTS

(In order of patent application date)

958,829, Method and Apparatus for Producing High Frequency Oscillating Currents. James H. Rogers. Filed Jan. 20, 1910.

1,220,005, Wireless Signaling System. James Harris Rogers and Henry H. Lyon. Filed Nov. 10, 1916.

1,322,622, wireless Signaling System, James H. Rogers and Henry H. Lyon, Filed Nov. 10, 1916.

1,349,103, Radiosignaling System. James Harris Rogers. Filed May 2, 1917.

1,315,862, Radiosignaling System. James Harris Rogers. Filed Jan. 17,

1918.

1,349,104, Radiosignaling System. James Harris Rogers. Filed Jan. 17, 1918.

1,303,729, Wireless Signaling System. James Harris Rogers. Filed Jan. 10,
1919.

1,303,730, Radiosignaling System. James Harris Rogers. Filed Jan. 11, 1919.

1,316,188, Radiosignaling System. James H. Rogers. Filed Mar. 28, 1919.

The Tom Bearden
Website

ACKNOWLEDGEMENTS

PART I

I am much indebted to Mr. Carey Briggs for furnishing references which strongly support the existence of scalar potential waves and scalar interferometry. Among those papers are Whittaker (1903), Nisbet (1955) , and other papers dealing with Hertzian potential theory. As ever, my profound gratitude is due Hal Crawford for his splendid illustrations which accompany this paper. I am also indebted to John Ratzlaff and the Tesla Book Company for graciously furnishing important Tesla references for me to analyze. My sincere thanks go to Messrs. Jim Beal and Williard Van de Bogart for furnishing important references on exoelectron emission and tectonic fault zones, and to Dr. Robert Beck for furnishing important material on the Rogers underground and undersea transmission system. Additional assistance was also received from Mr. Peter Kelly, Joe Siegeldorf, Joe Gambill, and Bob Beutlich. I am also very appreciative of the information furnished by Bob Jones and Vestigia on the beautiful Vestigia mystery light experiments.

PART II

A. L. Lez, San Anselmo, Calif.
Thorn L. Mayes, Saratoga, Calif.
Bob Nelson, Berkeley, Calif.
Mildred Ratzlaff, Millbrae, Calif.

The Tom Bearden Website

FOREWORD

The long-searched-for Tesla secrets have been solved by Tom Bearden. Part I contains the specific description of the solutions to Tesla's secrets and the Soviet Tesla weapons.

The present electromagnetic theory has several flaws, and it is these errors that have hidden the long-sought unified field theory. The explanations presented by the author provide clarity and understanding to related Tesla writings and patents.

Before the turn of the century, Nikola Tesla had discovered and was utilizing a new type of electric wave. Tesla repeatedly stated that his waves were non-Hertzian, and that his wireless transmissions did not fall off as the square of the distance. The author has found that the wave which Tesla was using is a longitudinal scalar wave, capable of many times the speed of light. When coupled with another scalar wave (Tesla wave),
the resulting vector wave can be the source of electromagnetic energy for electrical power, a protective force field, or a destructive weapon.

The booklet is well illustrated with special drawings by Hal Crawford. The bibliography includes corroborating information on all major points. Part II presents reprints or the entire subject matter of selected articles from the bibliography, plus additional references of importance. Articles are arranged in chronological order. Also included are nine patents by James Harris Rogers which were the basis of successful wireless communication during World War I. These principles of underground and sub-sea wireless were based on Tesla's discoveries.

Bearden's solutions are truly a major breakthrough, and an absolute necessity to the understanding of existing Tesla writings. It is also important that more should be known of this extraordinary individual, Thomas E. Bearden.

BIOSKETCH:

A nuclear engineer, wargames analyst, and military tactician, Lieutenant Colonel (Retired) Tom Bearden has over 26 years experience in air defense systems, tactics and operations; technical intelligence, anti-radiation missile countermeasures; nuclear weapons employment, computerized wargames; and military systems requirements.

He is currently with the Alabama division of a large aerospace company where he is involved in determining the future requirements for laser weapons.

Lieutenant Colonel Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in mathematics from Northeast Louisiana University. He is also a graduate of the U.S. Army Command and General Staff College and several U.S. Army artillery and missile schools.

He has had direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot missile systems and the production of technical intelligence on Soviet surface-to-air missile systems for the U.S. Army/Department of Defense.

Tom holds a Sandan (third degree black belt) in Yoseikan aikido; is a singer, guitarist and songwriter; member of Mensa, Society for the Investigation of the Unexplained, Vestigia, the American Association of Meta-Science: and is on the board of governors of the U.S. Psychotronics Association. In addition, he is Alabama Director of a small nonprofit foundation, the Association of Distinguished American Scientists and is a MUFON adviser on nuclear engineering.

He also edits *Specula,* Journal of the American Association of Meta-Science.

Tom is primarily active as a theorist on UFOs, parapsychological and paranormal phenomena, and psychotronics. He is the author of a book, *The Excalibur Briefing,* Strawberry Hill Press, 2594 15th Avenue, San Francisco, California 94127 and has numerous technical papers dealing with new paradigm physics, paranormal phenomena, and Soviet psychotronic weapons.

Finally, Bearden's *Solutions to Tesla's Secrets and the Soviet Tesla Weapons* promises to be recognized in the future as a major contribution to our present technology.

JTR

The Tom Bearden Website

TESLA'S SECRET AND THE SOVIET TESLA WEAPONS

© T.E. Bearden 1981

With Special Drawings by Hal Crawford

Before the turn of the century, Nikola Tesla had discovered and was utilizing a new type of electric wave. Tesla repeatedly stated his waves were non-Hertzian, and his wireless transmissions did not fall off as the square of the distance. His discovery was apparently so fundamental (and his intent to provide free energy to all humankind was so clear) that it was responsible for the withdrawal of his financial backing, his deliberate isolation, and the gradual removal of his name from the history books.

By 1914 or so, Tesla had been successfully isolated and was already nearly a "nonperson." Thereafter Tesla lived in nearly total seclusion, occasionally surfacing (at his annual birthday party for members of the press) to announce the discovery of an enormous new source of free energy, the perfection of wireless transmission of energy without losses, fireball weapons to destroy whole armies and thousands of airplanes at hundreds of miles distance, and a weapon (the "Tesla Shield," I've dubbed it) that could provide an impenetrable defense and thus render war obsolete.

In my pursuit of Tesla's secret, it gradually became apparent to me that present orthodox electromagnetic theory is seriously flawed in some fundamental respects. One of these is in the definition and use of ϕ, the scalar electrostatic potential. It is this error which has hidden the long-sought unified field theory from the theorists.

In the theory of the scalar electrostatic potential (SEP), the idea is introduced of work accomplished on a charge brought in from a distance against the scalar field. The SEP is not a vector field, but is a scalar field. Indeed, scalar potential cannot of itself perform work on a charged mass; if it could do so, then tremendous force would exist on every mass due to the extremely high SEP of the vacuum itself. Only a differential of SEP between two spatial points can produce force or accomplish work. (Rigorously, a differential of scalar potential between two spatial points constitutes a vector. Only a vector can produce force and do work.)

Also, work can only be done on a mass. Further, it takes time* to move an electron or other charged mass between two spatial points, and so the work performed by a spatial differential of the ϕ-field requires time. Rigorously, the delta SEP is voltage, not SEP per se, and is directly related to

E field. The entire voltage concept depends on the work performed in moving a mass, after that mass has moved. The idea of "voltage" always implies the existence of a steady differential of ϕ between two spatial points for a finite length of time, and it also involves the assumption of a flow of actual mass having occurred. SEP, on the one hand, is always a single-point function; on the other hand, difference in potential (i.e., V) is always a two point function, as is any vector.

Yet many graduate level physics and electromagnetics papers and texts erroneously confuse ϕ and V in the static case! Such an interpretation is of course quite incorrect.

---

* *Two spatial points involve at least $\Delta t = \Delta L/c$ in time. All vectors and gradients involve 2 separated spatial points, and thus present timelines in 4-space. ϕ_4 is a point, not a line, in 4-space.*

Another common assumption in present EM theory -- that the electrostatic potential (ϕ_0) of the normal vacuum is zero -- has no legitimate basis. In fact, we know ϕ_0 is nonzero because the vacuum is filled with enormous amounts of fluctuating virtual state activity, including incredible charge fluctuations. And by virtue of its point definition, ϕ_0 must be the "instantaneous intensity" of these fluctuations -- but both in space and time. The scalar electrostatic potential is therefore the "instantaneous stress" on spacetime itself, and a measure of the intensity of the virtual state flux through a 4-dimensional spacetime point.

Potential theory was largely developed in the 1800's, before the theory of relativity. Time flowrate was then regarded as immutable. Accordingly, electrostatic "intensity" was chosen as "spatial intensity," with the connotation of "spatial flux density." This assumes a constant, immutable rate of flow of time, which need not be true at all if we believe relativity. Such a spatial "point" intensity is actually a "line" in 4-space, and not a 4-dimensional "point" at all. Thus the spatial potential -- ϕ_3 -- is a very special case of the real spacetime potential -- ϕ_4, or charge and electromagnetic theory today is accordingly a special case of the real 4-space electromagnetism that actually exists! Note also that charge is a 4-dimensional concept.

Now mass is a spatial, 3-dimensional concept. Rigorously, mass does not exist in time -- masstime exists in time. Mass and charge are thus of differing dimensionalities!

Also, according to quantum mechanics, the charge of a particle -- e.g., of an electron -- is due to the continual flux of virtual particles given off and absorbed by the observable particle of mass. Thus charge also is conceptually a measure of the virtual flux density, and directly related to ϕ. Further, since the charge exists in time, it is the charge of a particle of spatial mass that gives it the property of masstime, or existing in time.

Here a great confusion and fundamental error has been thrown into the present EM theory by the equating of "charge" and "charged mass." As we have seen, the two things are really very different indeed.

To speak of a spatial "amount" of charge erroneously limits the basic EM theory to a fixed time flowrate condition (which of course it was considered to be, prior to Einstein's development of relativity). Thus when the limited present theory encounters a "relativistic" case (where the time flowrate changes), all sorts of extraordinary corrections must be introduced. The real problem, of course, is with the fundamental definitions of electrostatic potential and charge. The spatial "amount" of charge (i.e. the coulomb) as we presently erroneously use the term, is actually the spatial amount of observable "charged mass." To correct the theory, one must introduce the true 4-space SEP and separate the definitions of charge and charged mass.

Only when a mass is moved does one have work -- and voltage or vector fields. (The reason one has voltage and E field connected to a normal electrostatically charged object in the laboratory is because an excess of charged-particle masses are assembled on the object, and these masses are in violent motion! A true static charge would have no E field at all.)

Table 1. Some Present Theoretical Facts

- $Ø_3$-FIELD CONCEIVED BEFORE RELATIVITY
    - POINT FUNCTION IN SPACE
    - LINE FUNCTION IN SPACETIME
    - TIME FLOWRATE CONSIDERED IMMUTABLE
    - SPECIAL CASE OF $Ø_4, \ldots\ldots, Ø_N$
- GRADIENT BETWEEN TWO POINTS IN $Ø_3$ FIELD
    - PRODUCES VOLTAGE
    - PRODUCES E FIELD
    - MOVES ELECTRON
    - IS AN AREA IN 4-SPACE
- GRADIENT MAY BE BETWEEN SEPARATED POINTS IN TWO SUPERPOSED Ø-FIELDS
- Ø-FIELD CAN "MOVE" AT INFINITE VELOCITY*

* JACKSON, CLASSICAL ELECTRODYNAMICS, 2ND EDITION, P. 223

Table 2. Some Proposed $Ø_4$ Characteristics

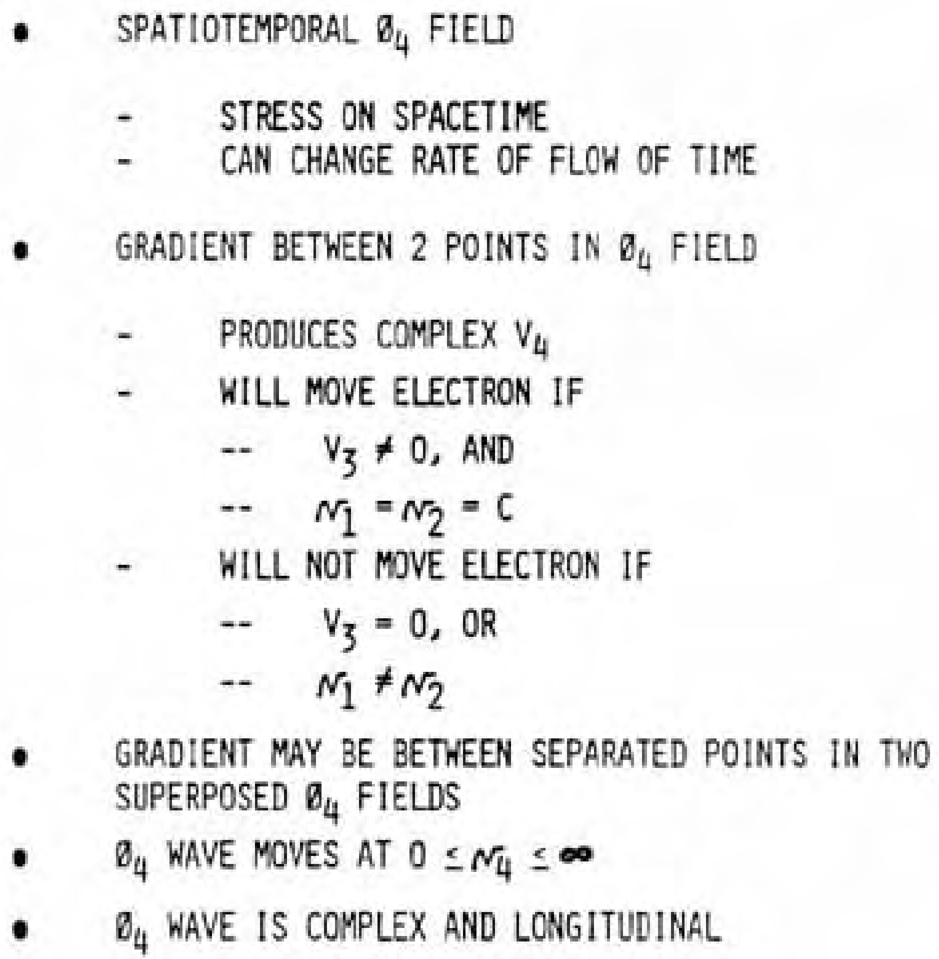

- SPATIOTEMPORAL $Ø_4$ FIELD
  - STRESS ON SPACETIME
  - CAN CHANGE RATE OF FLOW OF TIME
- GRADIENT BETWEEN 2 POINTS IN $Ø_4$ FIELD
  - PRODUCES COMPLEX V_4
  - WILL MOVE ELECTRON IF
    - -- $V_3 \neq 0$, AND
    - -- $\nu_1 = \nu_2 = C$
  - WILL NOT MOVE ELECTRON IF
    - -- $V_3 = 0$, OR
    - -- $\nu_1 \neq \nu_2$
- GRADIENT MAY BE BETWEEN SEPARATED POINTS IN TWO SUPERPOSED $Ø_4$ FIELDS
- $Ø_4$ WAVE MOVES AT $0 \leq \nu_4 \leq \infty$
- $Ø_4$ WAVE IS COMPLEX AND LONGITUDINAL

ϕ-field need not involve observable mass accumulation, but only charge (virtual flowrate intensity) accumulation. Accumulated masses are like so many gallons of water; accumulated charge is like so much pressure on both the water (space) and the time in which the water is existing.

Now, if one varies the SEP solely as a point function, one would have a purely scalar complex longitudinal wave, and not a vector wave at all. This is the fundamentally new electrical wave that Tesla discovered in 1899.

Rigorously, all vector fields are two-point functions and thus decomposable into two scalar fields, as Whittaker showed in 1903. It follows that any vector wave can be decomposed into two scalar waves. By implication, therefore, a normal transverse EM vector wave, e.g., must simply be two coupled scalar (Tesla) waves -- and these scalars independently would be longitudinal if uncoupled. An ordinary transverse EM vector wave is thus two pair-coupled Tesla scalar longitudinal waves, and only a single special case of the much more fundamental electromagnetics discovered by Nikola Tesla.

A Tesla (scalar potential) wave -- i.e., a massless wave in pure ϕ_0, the stress of the spacetime medium would have very strange characteristics indeed. For one thing, since it moves in a complex 4-space, it has many more modes of movement than does a simple wave in 3-space. And for another thing, it need not be bound at all by the speed of (vector) light. In current theory, one ϕ_3-field does not directly interact or couple with other existing ϕ_3-fields except by simple superposition. Therefore presently the ϕ-field is considered to have no drag limitation at all, hence infinite velocity. (E.g., as stated in Jackson, Classical Electrodynamics, 2nd edition, page 223.)

Actually, a ϕ_4-wave can and will interact with some of the other existing ϕ_4-waves in the medium transversed, and this interaction can involve pair-coupling into EM vector fields and waves, an interaction not presently in the electrodynamics theory. The result of scalar pair-coupling creates a finite amount of vector "drag" on the ϕ_4-wave, so it then has less than infinite velocity. However, if this drag is small due to limited pair-coupling, the scalar wave's velocity through the slightly dragging medium still may be far greater than the speed of vector EM waves (light) in vacuum. On the other hand, if the pair-coupling is made severe, the ϕ-wave may move at a speed considerably below the speed of vector light waves in vacuum. The velocity of the ϕ_4-wave is thus both variable and controllable or adjustable (e.g., simply by varying its initial amplitude, which through a given medium changes the percentage of pair-coupling and hence the degree of drag on the scalar wave). The Tesla scalar wave thus can have either subluminal or superluminal velocity, in contradiction to present theory.

Note that the scalar wave also violates one of Einstein's fundamental postulates -- for the speed of our "new kind of light" wave is not limited to c, and need not be the same to every observer. Thus Tesla scalar waves lead to a new "superrelativity" of which the present Einstein relativity is only a highly special case!

Table 3. Tesla Waves Can:

- **ESTABLISH STANDING WAVES**
  - **in the earth**
  - **in the ionosphere**
- **TAP ENERGY FROM THE EARTH'S CORE**
- **TRAVEL FASTER OR SLOWER THAN LIGHT**
- **CHANGE RATE OF TIME FLOW**
- **AFFECT ALL FIELDS, INCLUDING GRAVITY**
- **COMMUTE BETWEEN VIRTUAL AND OBSERVABLE**

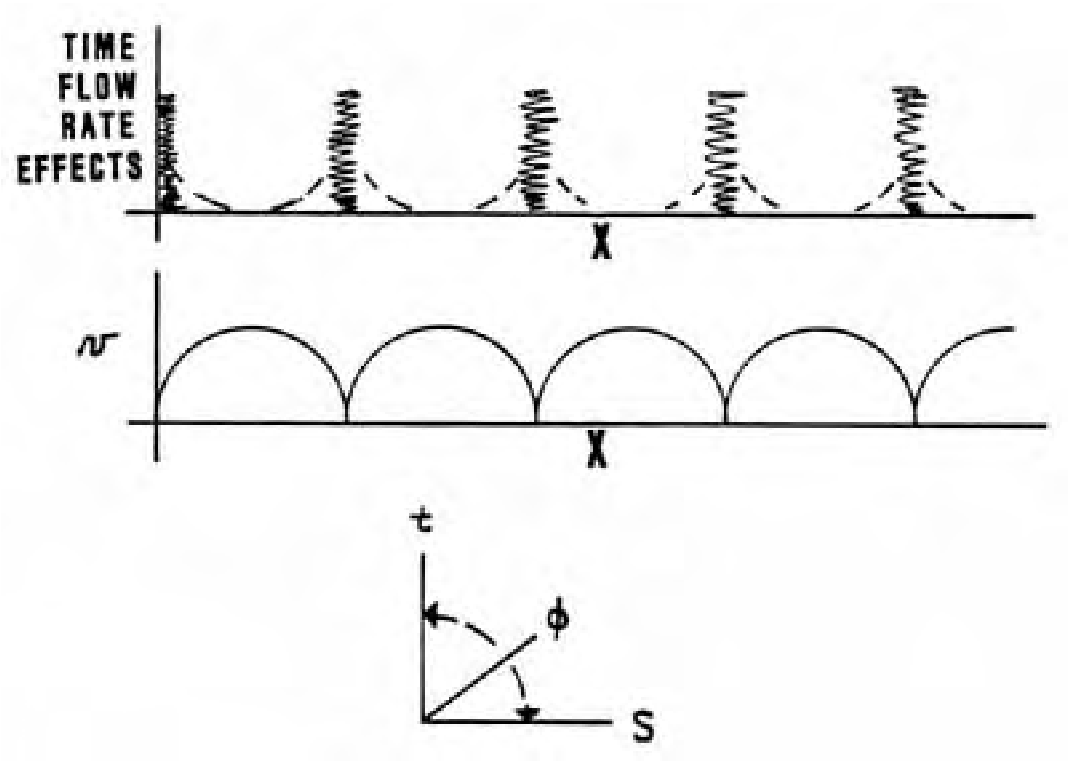

Figure 1. Tesla's Standing Columnar Wave

But let us now look for some subtle but real examples of scalar waves and scalar pair-coupling in nature. As is well known, a tectonic fault zone can provide anomalous lights, sounds, etc. from stresses, piezoelectrical activity, and telluric currents in the earth and through the fault zone. In examining the fault zone phenomena, I finally realized that a fault zone was literally a scalar interferometer i.e., if one can have scalar ϕ_4-waves, they can interfere, either constructively or destructively. Their interference, however, produces scalar pair-coupling into vector EM waves. This coupling may be at a distance from the interferometer itself, and thus the interferometer can produce energy directly at a distance, without vector transmission through the intervening space. Coupling of ϕ waves with the paired scalars comprising ordinary EM vector waves can also occur. If this triplex coupling forms additional EM vector waves 180 degrees out of phase, the ordinary EM wave is diminished or extinguished. If the scalar triplex coupling occurs so as to create vector EM waves in phase with the interacting vector EM wave, the amplitude of the ordinary vector wave is increased.

Scalar potential waves can thus augment or diminish, or create or destroy, ordinary EM waves at a distance by pair-coupling interference under appropriate conditions, and this is in consonance with the implications of Whittaker's fundamental 1903 work.

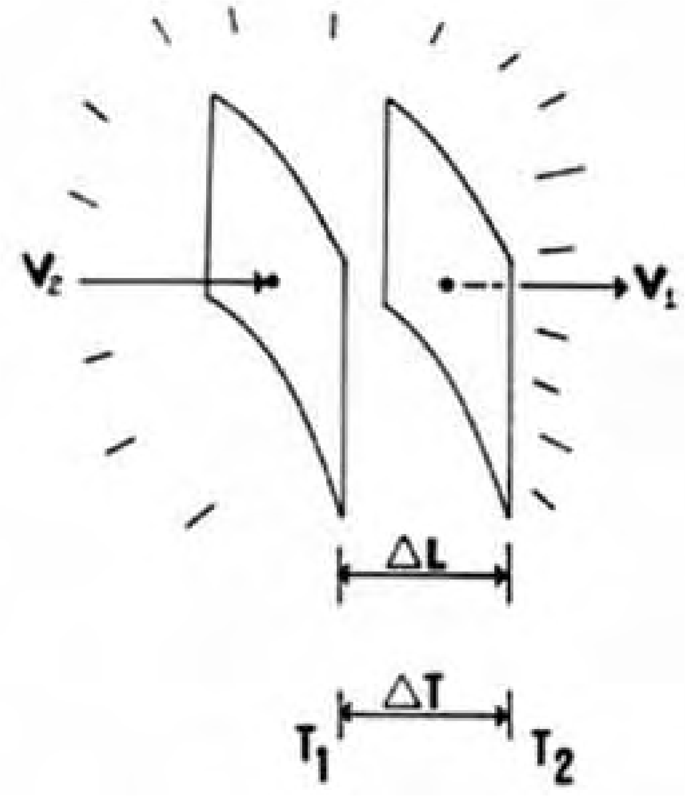
V_2
V_1
ΔL
ΔT
T_1
T_2

ALL VECTORS AND VECTOR FIELDS ARE CREATED BY SCALAR INTERFEROMETRY AND PAIR COUPLING

EM COUPLING CONDITION

$$\Delta L/c = \Delta T$$

Figure 2. Pair Coupling: Producing EM at a Distance

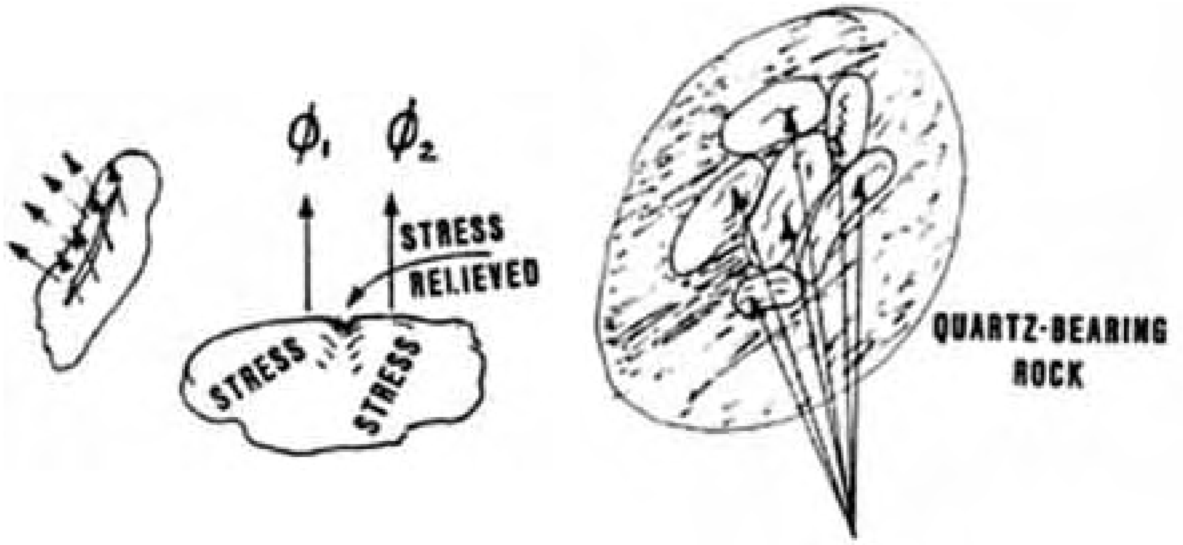
ϕ_1
ϕ_2
STRESS RELIEVED
STRESS
STRESS
QUARTZ-BEARING ROCK

LOCKED-IN STRESSED CRACKS

EACH TINY STRESSED CRACK ACTS AS A SCALAR INTERFEROMETER

BECOMES A PRIGOGINE CRYSTAL

Figure 3. Natural Scalar Interferometers

An earthquake fault zone is such a scalar interferometer. Stresses and charge pileups exist in the plates on each side adjacent to the fault, with stress relief existing in the middle in the fault fracture itself. Since the rock is locally nonlinear, the mechanical stresses and electrical currents in it are also locally nonlinear. This results in the generation of multiple frequencies of ϕ_4-waves from each side of the fault interferometer, yielding two complex Fourier expansion patterns of scalar potential waves. On occasion these two Fourier-transformed scalar wave patterns couple at a distance to produce stable ordinary electromagnetic field in a 3-dimensional spatial pattern -- e.g., a stress light such as the Vestigia light covered in Part I of The Excalibur Briefing. Driven by the erratic two scalar Fourier expansion patterns of the scalar interferometer (whose input stresses normally slowly change), an erratic, darting, hovering "spooklight" of the variety studied by Vestigia is produced.

As the stresses change in each side of the interferometer, the distant scalar coupling zone is affected. Thus the stresslight moves and its form changes, but it may be relatively stable in form for seconds or minutes. Since the stresses in the rock may be intense, the stress light may involve an intense pair of ϕ-patterns coupling into the sphere or ball of vector EM energy. The atoms and molecules of the air in the region of the coupled stresslight ball thus become highly excited, giving off radiant energy as the excited states decay.

Since much of the piezoelectric material in the stressed rocks is quartz grains, the features of quartz are of particular interest. Each little quartz grain is itself highly stressed, and has stress cracks. It is therefore a little scalar interferometer. Further, quartz is transparent to infrared and ultraviolet; and the random orientation of all the quartz scalar interferometers may also form a Prigogine system far from thermodynamic equilibrium. If so, this system can tap into highly energetic microscopic electromagnetic fluctuations to produce large-scale, ordered, relatively stable patterns of electromagnetic energy at a distance.

In short, all of this lends support to the formation of relatively stable but somewhat erratic patterns of electromagnetic energy at a distance from the fault itself. In the atmosphere, such scalar interferometers could form in clouds or even in the air or between clouds and earth. If so, such rare but occasional "weather" scalar interferometers could account for the rare phenomenon of ball lightning. The intense energy of the ball of lightning, as compared to the lesser energy of an earthstress light, could well be due to the enormous electrical charges between clouds or between cloud and earth, available to fuel the scalar interferometer. Very probably it is this phenomenon which gave Tesla the clue to scalar wave interferometry.

Thus such phenomena as earthstress lights, ball lightning, and the Tesla system of wireless transmission of energy at a distance with negligible losses and at speeds exceeding the speed of light may be explained. They are complex, however, and involve fundamental changes to present electromagnetic theory. These changes include utilizing 4-space scalar electrostatic potentials, scalar waves, pair coupling, ordinary 3-dimensional Fourier expansion, the Prigogine effect, and the properties of piezoelectric materials in rocks.

Since the scalar potential also stresses time, it can change the rate of flow of time itself. Thus it affects anything which exists in time -- including the mind, both of the individual and at various levels of unconsciousness. Therefore the same functions that result in earthstress lights also affect mind and thought, and are in turn affected by mind and thought. This is the missing ingredient in Persinger's theory that UFO's are correlated with, and a result of, fault zones and earth stresses. While Persinger seems to feel this is a "normal physics" explanation, it indeed involves a paranormal explanation.

The time-stressing ability of the true ϕ scalar wave also explains the interaction of such earthstress lights with humans and human intent, as noted by other researchers. (E.g., the lights that repeatedly seemed to react to the observers, as detailed by Dr. Harley Rutledge in his epoch-making Project Identification, Prentice-Hall, 1981.)

These ideas in condensed form comprise the concepts required to violate the speed of light and produce an ordinary electromagnetic field at a distance, using scalar interferometry, without losses -- as Tesla had done in his wireless transmission system which he had tested prior to 1900 and had perfected by the 1930's. Scalar interferometry can give stable regions of EM or "light energy" at a distance without losses, particularly as detailed in the beautiful Vestigia experiments, and it is within our grasp to utilize the new effects. Indeed, any stress crack in a material can result in the scalar potential interferometer effect. Exophoton and exoelectron emission -- poorly understood but already known in fatiguing of materials -- must be at least partly due to the scalar interferometer effect.

However, one additional caution should be advanced. Normal movement of electrons allows so much "sideplay" movement of the electrons -- and there is so much such sideplay electron motion in the surrounding vicinity -- that pair coupling is almost instantaneous for small waves. Thus orbital electrons in atoms seem to absorb and emit vector EM photons. Actually they also emit some percentage of scalar waves as well. Since a scalar wave is comprised exclusively of disturbance in the virtual state, it need not obey the conservation of energy law. Further, a scalar wave of itself does not "push electrons" or other charges; hence it is nearly indetectable by present detectors. Ionization detectors such as a Geiger counter tube, e.g., are exceptions if the scalar wave encountered is fairly strong. In that case sufficient triplex coupling with the ionized gas occurs to produce additional ionization or charge, breaching the tube's cutoff threshold and producing a cascade discharge of electrons and voltage which is detected. But weak scalar waves are presently indetectable by ordinary instruments. However, these small scalar waves are detectable by sensitive interferometry techniques e.g., such as an electron interferometer. Since the use of such instruments is quite rare, then indeed we have been living immersed in a sea of scalar waves without knowing it.

Finally, the percentage of scalar waves produced by changes in charged mass pileups can be increased by utilizing charged mass streaming. Essentially the charged masses must be moved suddenly, as quickly as possible, at or near the complete breakdown of the medium. For this reason, Tesla utilized sparkgaps in his early transmission systems, but also found that he could induce ionized media to "breakdown" in such fashion by a slow growth process. One of his early patented atmospheric wireless transmission systems is based on this fact. However, it was necessary to use a very high voltage, insuring extreme stress on the medium and hence some spillover stress onto time itself. In other words, ϕ_3 is always an approximation; at sufficiently high spatial stress, sufficient spillover ϕ_4 exists to give Tesla scalar waves. For this reason, Tesla used very high voltages and extremely sharp discharges to give "streaming" of the charged masses and thus high percentages of ϕ_4 waves. This suggests that the breakdown of dielectrics is a much richer phenomenon than is presently allowed for in the conventional theory.

To summarize, electrostatic potential -- ϕ-field is stress on the spacetime medium at a four-dimensional point. I. e. , it is a sort of pressure on the medium, but pressure on all four dimensions, not just on the three spatial dimensions. Thus in the new standard theory, ϕ_4 may have complex values. In addition, a ϕ-wave is to be interpreted as a scalar longitudinal wave in complex spacetime -- directly in ϕ_0, the normal average 4-space stress itself. And charge and charged mass must be recognized as two separate concepts. This is the gist of what I finally recognized about Nikola Tesla's work and fundamental discovery.

This is exciting, for it means that Tesla stress waves can affect either space or time individually, or both space and time simultaneously, or even oscillate back and forth between primarily affecting time and primarily affecting space, Tesla's waves were actually these ϕ-field scalar waves. As such, they were fundamentally different from ordinary electromagnetic waves, and had entirely different characteristics, just as Tesla often stated. E. g., a Tesla wave can either move spatially, with time flowing linearly; move temporally only (sitting at a point and waxing and waning in magnitude -- but changing the rate of flow of time itself in doing so, and affecting gravitational field, fundamental constants of nature, etc,), or move in a combination of the two modes. In the latter case, the Tesla wave moves in space with a very strange motion -- it oscillates between (1) spatially standing still and flexing time, and (2) moving smoothly in space while time flows smoothly and evenly. I.e., It stands at one point (or at one columnar region), flexing for a moment; then slowly picks up spatial velocity until it is moving smoothly through space; then slows down again to a "standing column," etc. This is Tesla's fabulous "standing columnar wave."

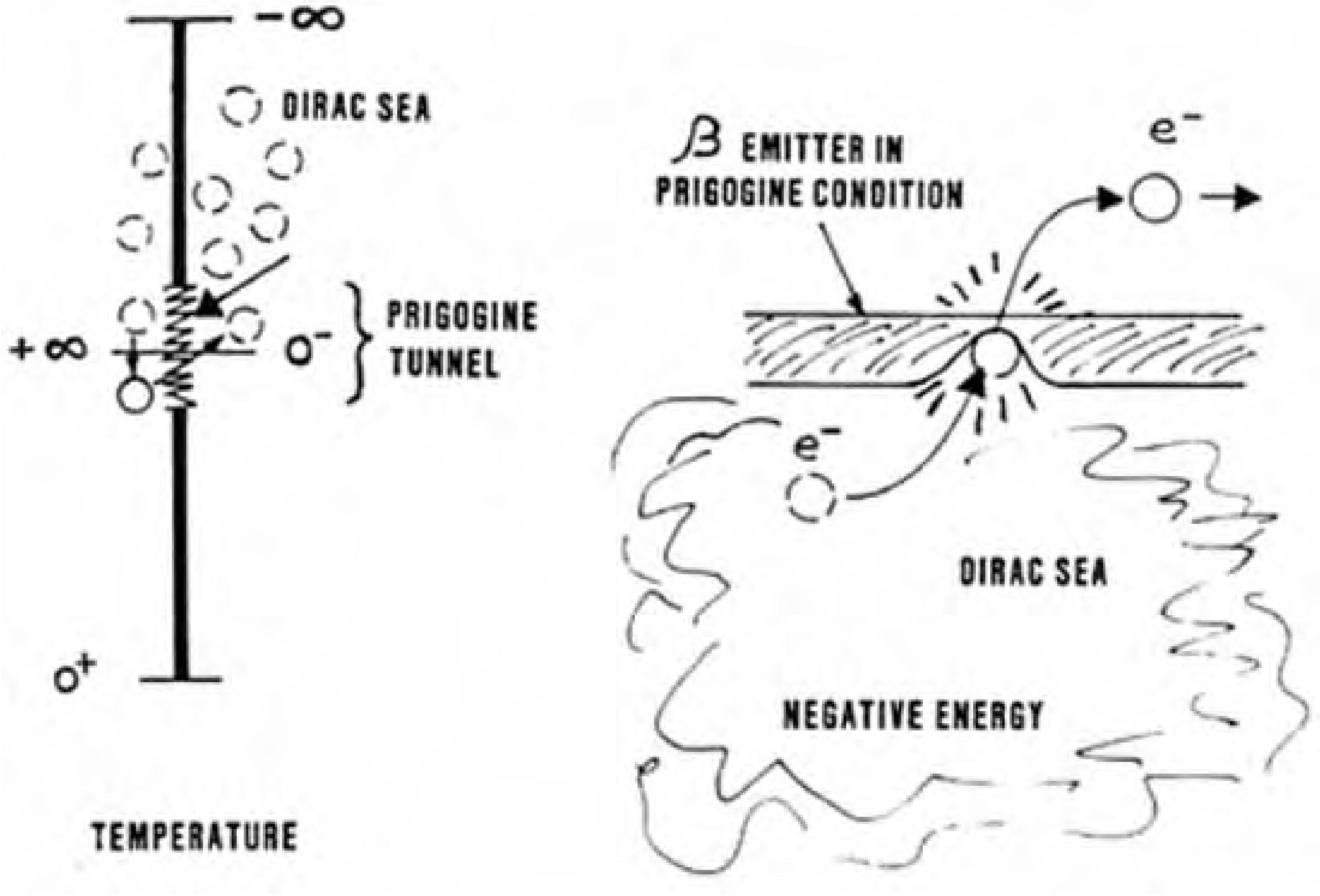
−∞
DIRAC SEA
+∞
0⁻
PRIGOGINE TUNNEL
0⁺
TEMPERATURE
β EMITTER IN PRIGOGINE CONDITION
e⁻
e⁻
DIRAC SEA
NEGATIVE ENERGY

Figure 4. Exoelectric Emission

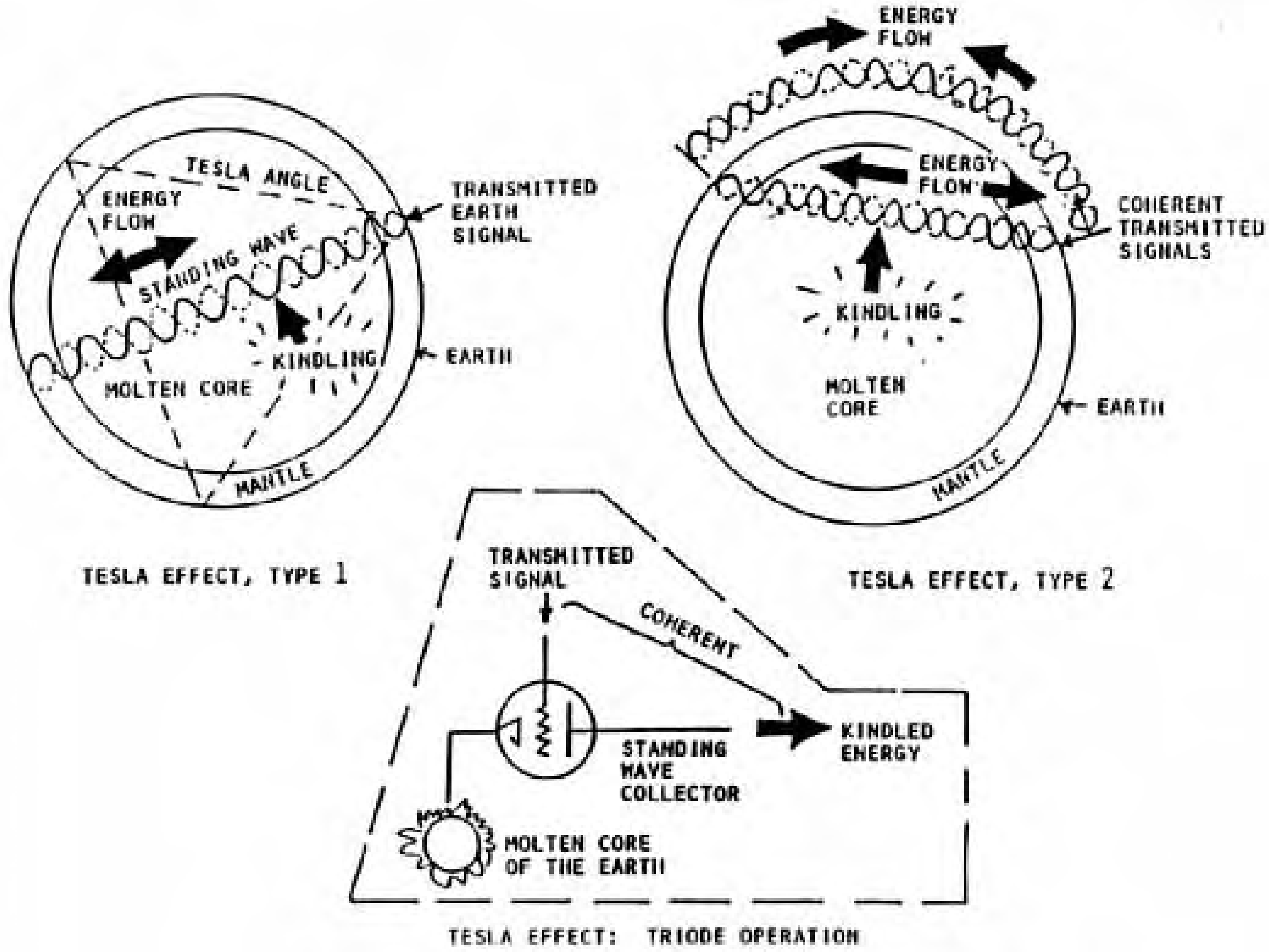
TESLA ANGLE
ENERGY FLOW
STANDING WAVE
TRANSMITTED EARTH SIGNAL
KINDLING
EARTH
MOLTEN CORE
MANTLE
TESLA EFFECT, TYPE 1
ENERGY FLOW
ENERGY FLOW
COHERENT TRANSMITTED SIGNALS
KINDLING
MOLTEN CORE
EARTH
MANTLE
TESLA EFFECT, TYPE 2
TRANSMITTED SIGNAL
COHERENT
KINDLED ENERGY
STANDING WAVE COLLECTOR
MOLTEN CORE OF THE EARTH
TESLA EFFECT: TRIODE OPERATION

Figure 5. The Tesla Effect

Another wild characteristic of the Tesla wave is that it can affect the rate of flow of time itself; hence it can affect or change every other field including the gravitational field -- that exists in time flow. It can also affect all universal constants, the mass of an object, the inertia of a body, and the mind and thoughts as well! All of these exist in the flow of time, and they are affected if the time stream in which they exist is affected. This was the awful secret that Tesla partially discovered by 1900, and which he came more and more to fully realize as he pursued its nature and its ramifications into the 1920's and 1930's.

Tesla also found he could set up standing ϕ-field waves through the earth. He in fact intended to do so, for he had also discovered that all charges in the highly stressed earth regions in which such a standing wave existed produced ϕ-fields which would feed (kindle) energy into the standing ϕ-field wave by pair coupling. I.e., normal vector field energy would "assemble" onto the scalar matrix wave by means of pair-coupling. Thus by transmitting a scalar standing wave into the earth, he could easily tap the fiery scalar fields produced in the molten core of the planet itself, turning them into ordinary electromagnetic energy. In such case, a single generator would enable anyone to put up a simple antenna and extract all the free energy desired.

When Tesla's alarmed financial backers discovered this was his real intent, they considered him a dangerous madman and found it necessary to ruthlessly stop him at all costs. And so his financial support was withdrawn, he was harassed in his more subtle patent efforts (and the patents themselves were adulterated), and his name gradually was removed from all the electrical textbooks. By 1914 Tesla, who had been the greatest inventor and scientist in the world, had become essentially a nonperson.

A few other persons in the early 1900's also were aware that potential and voltage are different. And some of them even learned to utilize Tesla's ϕ-field, even though they only vaguely understood they were utilizing a fundamentally different kind of electromagnetic wave. For example, James Harris Rogers patented an undersea and underground communications system which Tesla later confirmed utilized Tesla waves. The U.S. secretly used the Rogers communications system in World War I to communicate with U.S. submarines underwater, and to communicate through the earth to the American Expeditionary Force Headquarters in Europe. The Rogers system was declassified after the War and very shortly after that, it had mysteriously been scrubbed off the face of the earth. Again, potential stress waves -- Tesla waves -- were eliminated and "buried."

Probably the most brilliant inventor and researcher into Tesla's electromagnetics was T. Henry Moray of Salt Lake City, Utah. Dr. Moray actually succeeded in tapping the limitless zero-point energy of vacuum (spacetime) itself. By 1939, Dr. Moray's amplifier contained 29 stages and its output stage produced 50 kilowatts of power from vacuum. Interestingly, another 50 kilowatts could be tapped off any other stage in the device -- which consequently could have produced almost 1.5 megawatts of electrical power! Dr. Moray's epoch-making work was suppressed also. His device -- which represented over 20 years of heartbreaking accumulation of 29 working tubes from thousands made -- was destroyed by a Soviet agent in 1939, but not before the agent had obtained the drawings for building the tubes and the device itself. Today the Moray amplifier is a standard component of many of the Soviet secret superweapons and Tesla weapons.

In the 20's and 30's, Tesla announced the final perfection of his wireless transmission of energy without losses even to interplanetary distances. In several articles (e.g., H. Winfield Secor, "Tesla Maps Our Electrical Future, " Science and Invention, Vol. XVII, No.12, pp. 1077, 1124-1126), Tesla even revealed he used longitudinal stress waves in his wireless power transmission. Quoting from the article, "Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound." "He says that a Hertz wave would only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves are absolutely impossible." The wily Tesla did not reveal, of course, that such scalar waves nearly always immediately pair-coupled into vector waves when produced by normal means. Tesla himself was working with longitudinal scalar waves.

Table 4. Rogers' Undersea and Underground Communication System

- **SECRETLY USED BY U.S. IN WWI**
- **COMMUNICATED TO SUBS UNDERWATER**
    - **- ANY DEPTH, UP TO 30 KHZ**
    - **- REDUCED DEPTH, >30 KHZ**
- **COMMUNICATED TO EXPEDITIONARY HQ OVERSEAS**
    - **- 5,000 TIMES STRONGER THAN AERIAL**
    - **- STATIC-FREE**
- **DECLASSIFIED IN 1919**
    - **- TO REVOLUTIONIZE COMMUNICATIONS**
    - **- TO END ALL SURFACE ANTENNAS**
- **MYSTERIOUSLY "LOST"**
- **REDISCOVERED AND "LOST" THREE TIMES AFTER WWII**

continued

BEARDEN
FER DE LANCE
12.14 15:24
BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

ORDER THE ALL NEW 2ND EDITION OF "FER DE LANCE"

Updated to the year 2002

Stunning new revelations about the secret scalar EM wars the media ignores

THIS BOOK IS THE ONLY SOURCE FOR THIS INFORMATION

The Tom Bearden Website

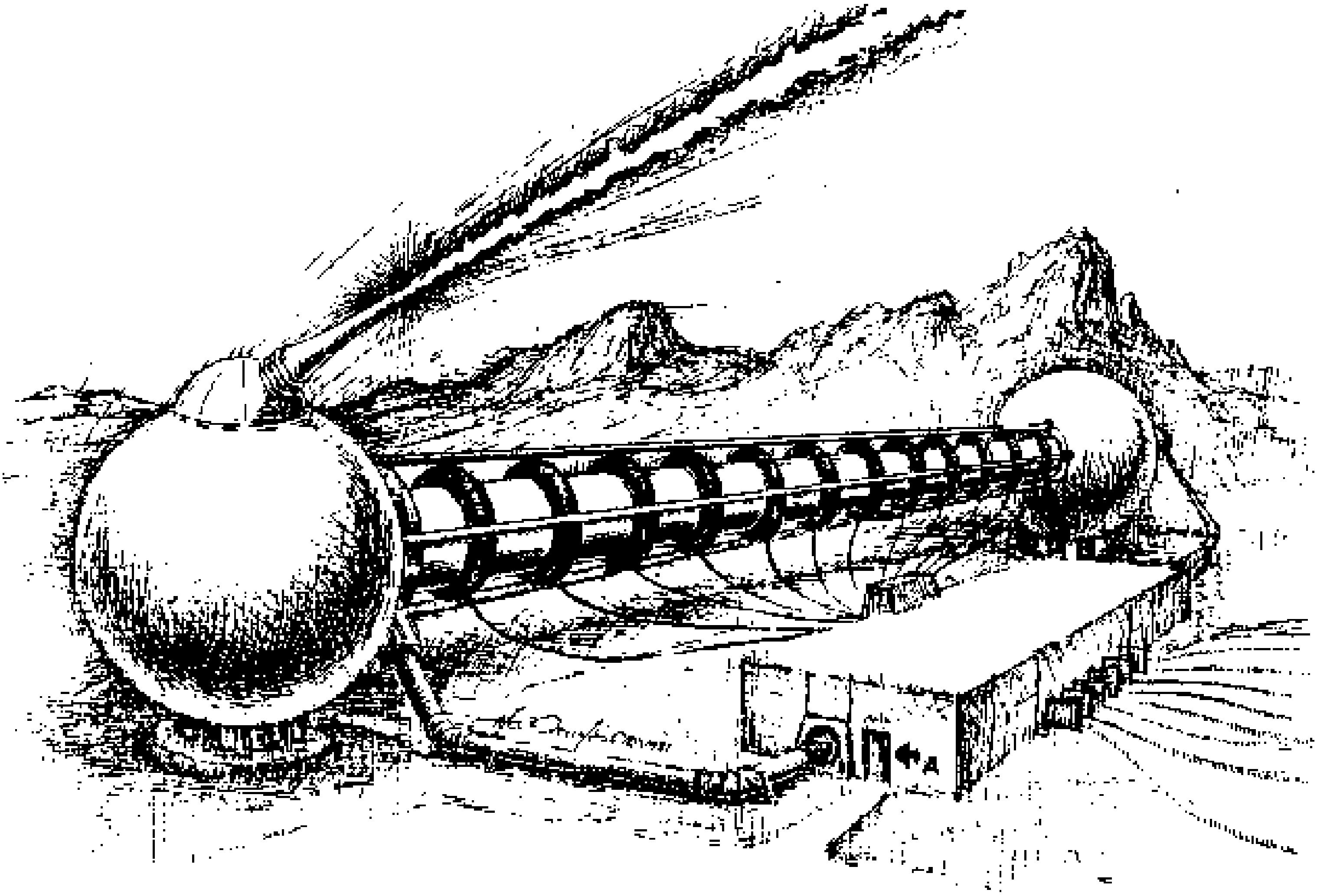

Scalar Potential Interferometer

Figure 6. Multimode Tesla Weapon

In the 1930's Tesla announced other bizarre and terrible weapons: a death ray, a weapon to destroy hundreds or even thousands of aircraft at hundreds of miles range, and his ultimate weapon to end all war -- the Tesla shield, which nothing could penetrate. However, by this time no one any longer paid any real attention to the forgotten great genius. Tesla died in 1943 without ever revealing the secret of these great weapons and inventions.

Unfortunately, today in 1981 the Soviet Union has long since discovered and weaponized the Tesla scalar wave effects. Here we only have time to detail the most powerful of these frightening Tesla weapons -- which Brezhnev undoubtedly was referring to in 1975 when the Soviet side at the SALT talks suddenly suggested limiting the development of new weapons "more frightening than the mind of man had imagined." One of these weapons is the Tesla howitzer recently completed at the Saryshagan missile range and presently considered to be either a high-energy laser or a particle beam weapon, (See Aviation Week & Space Technology, July 28, 1980, p. 48 for an artist's conception.)

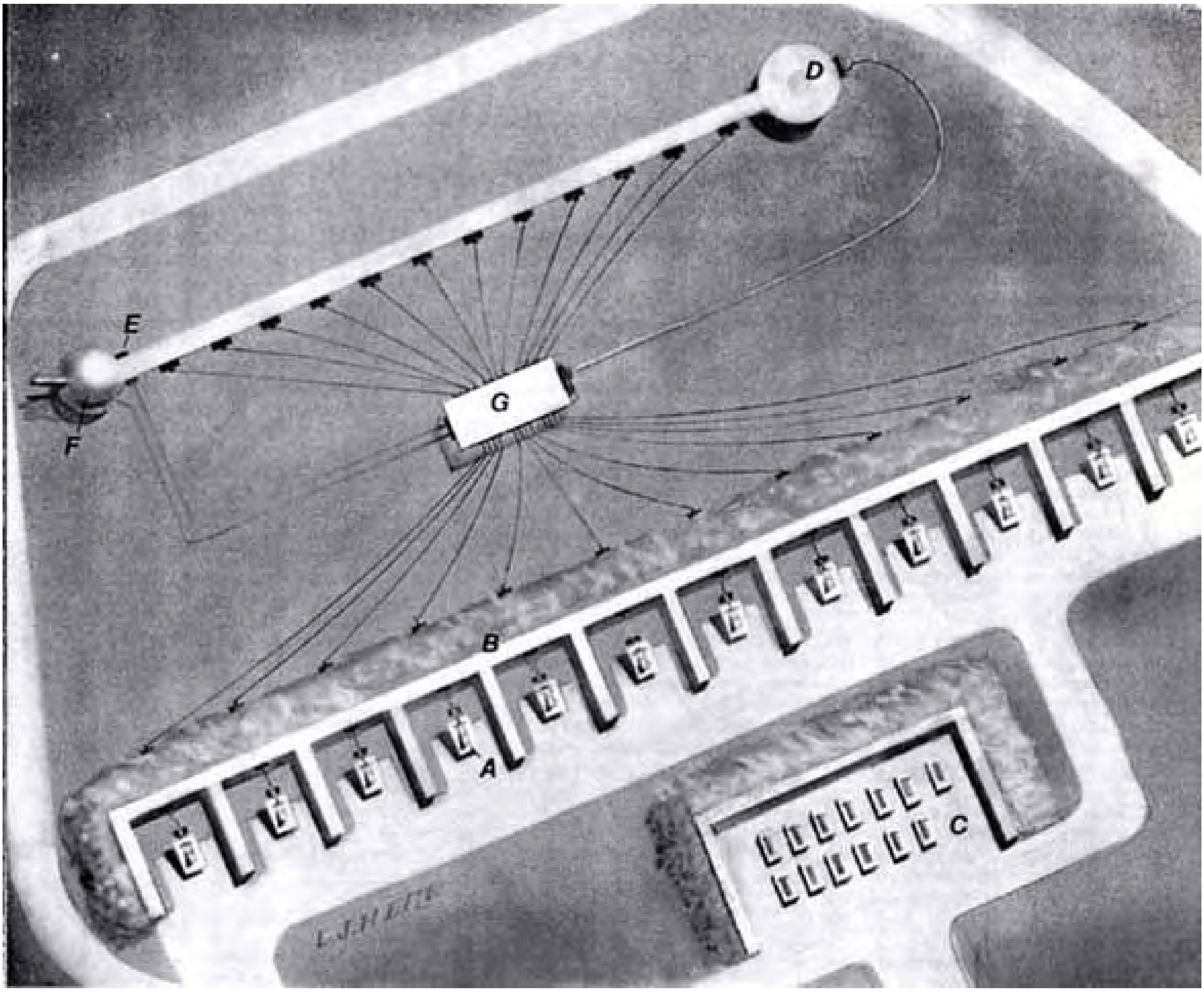

Aviation Week & Space Technology July 28, 1980

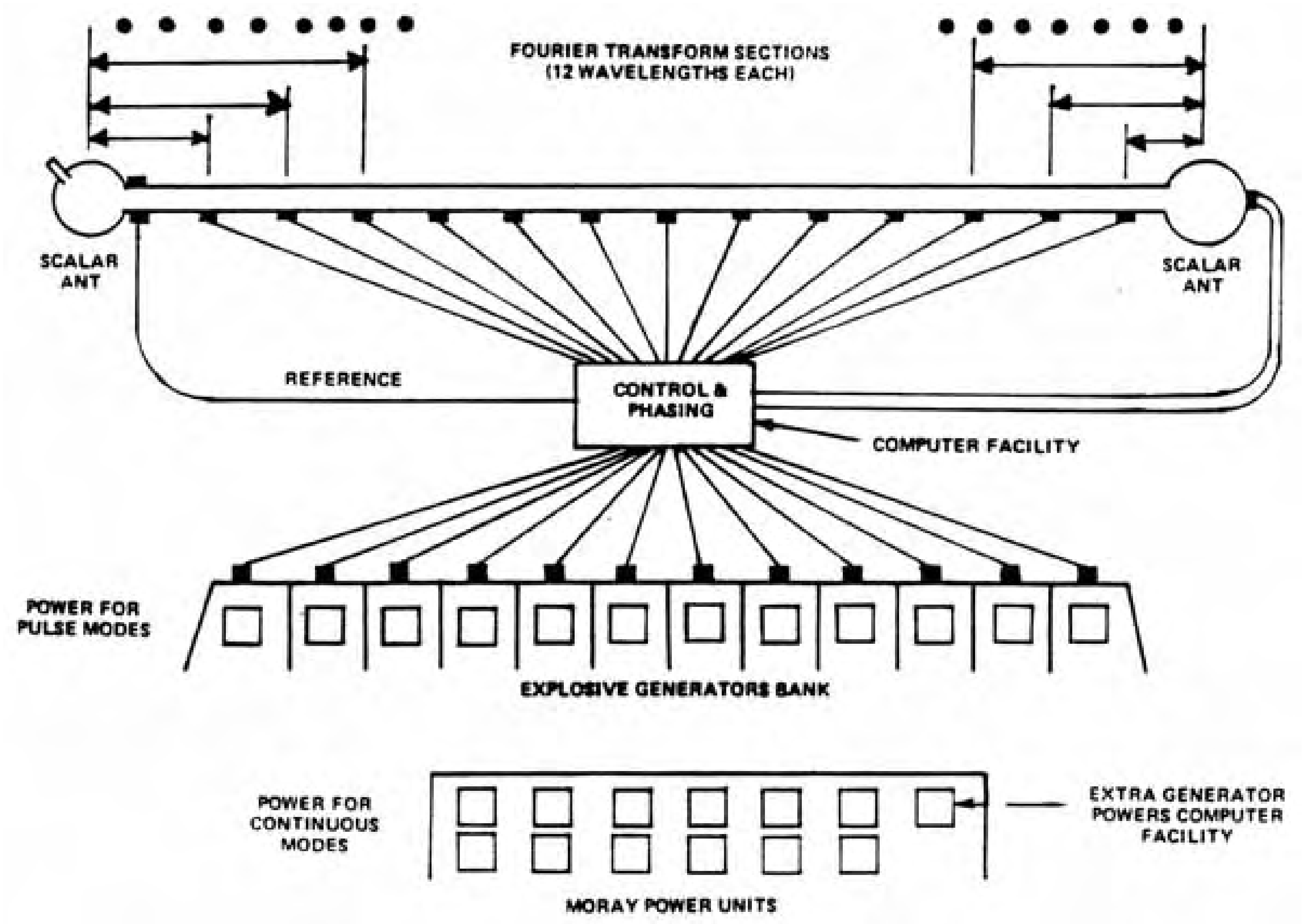
FOURIER TRANSFORM SECTIONS
(12 WAVELENGTHS EACH)
SCALAR ANT
SCALAR ANT
REFERENCE
CONTROL & PHASING
COMPUTER FACILITY
POWER FOR PULSE MODES
EXPLOSIVE GENERATORS BANK
POWER FOR CONTINUOUS MODES
EXTRA GENERATOR POWERS COMPUTER FACILITY
MORAY POWER UNITS

Figure 7. Tesla Weapons at Saryshagan

The Saryshagan howitzer actually is a huge Tesla scalar interferometer with four modes of operation. One continuous mode is the Tesla shield, which places a thin, impenetrable hemispherical shell of energy over a large defended area. The 3-dimensional shell is created by interfering two Fourier-expansion, 3-dimensional scalar hemispherical patterns in space so they pair-couple into a dome-like shell of intense, ordinary electromagnetic energy. The air molecules and atoms in the shell are totally ionized and thus highly excited, giving off intense, glowing light. Anything physical which hits the shell receives an enormous discharge of electrical energy and is instantly vaporized -- it goes pfft! like a bug hitting one of the electrical bug killers now so much in vogue.

If several of these hemispherical shells are concentrically stacked, even the gamma radiation and EMP from a high altitude nuclear explosion above the stack cannot penetrate all the shells due to repetitive absorption and reradiation, and scattering in the layered plasmas.

In the continuous shield mode, the Tesla interferometer is fed by a bank of Moray free energy generators, so that enormous energy is available in the shield. A diagram of the Saryshagan-type Tesla howitzer is shown in figure 7. Hal Crawford's fine drawing of the interferometer end of the Tesla howitzer is shown in figure 6. Hal's exceptional rendition of the Tesla shield produced by the howitzer is shown in figure 8.

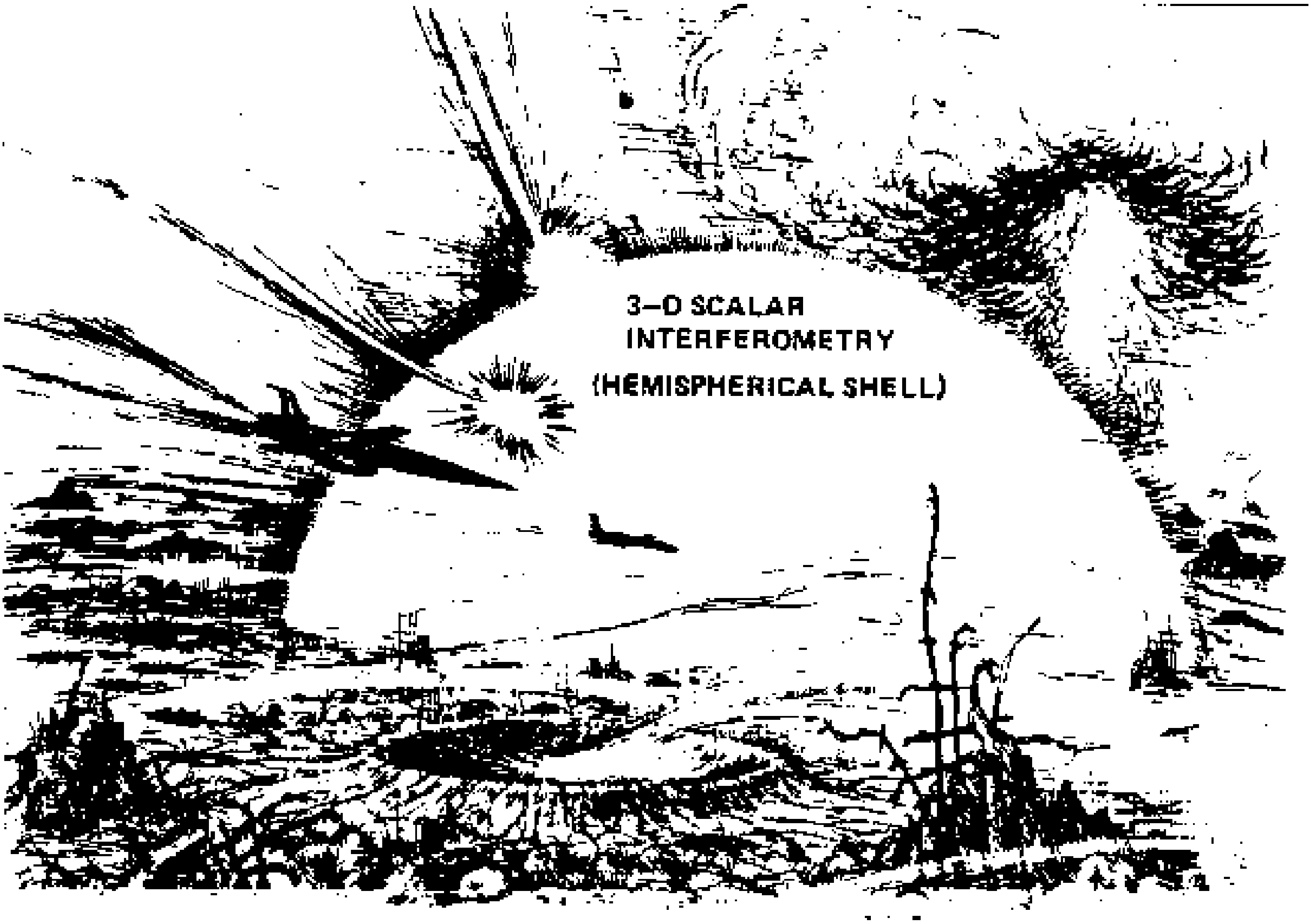
3–D SCALAR
INTERFEROMETRY
(HEMISPHERICAL SHELL)

Figure 8. The Tesla Shield

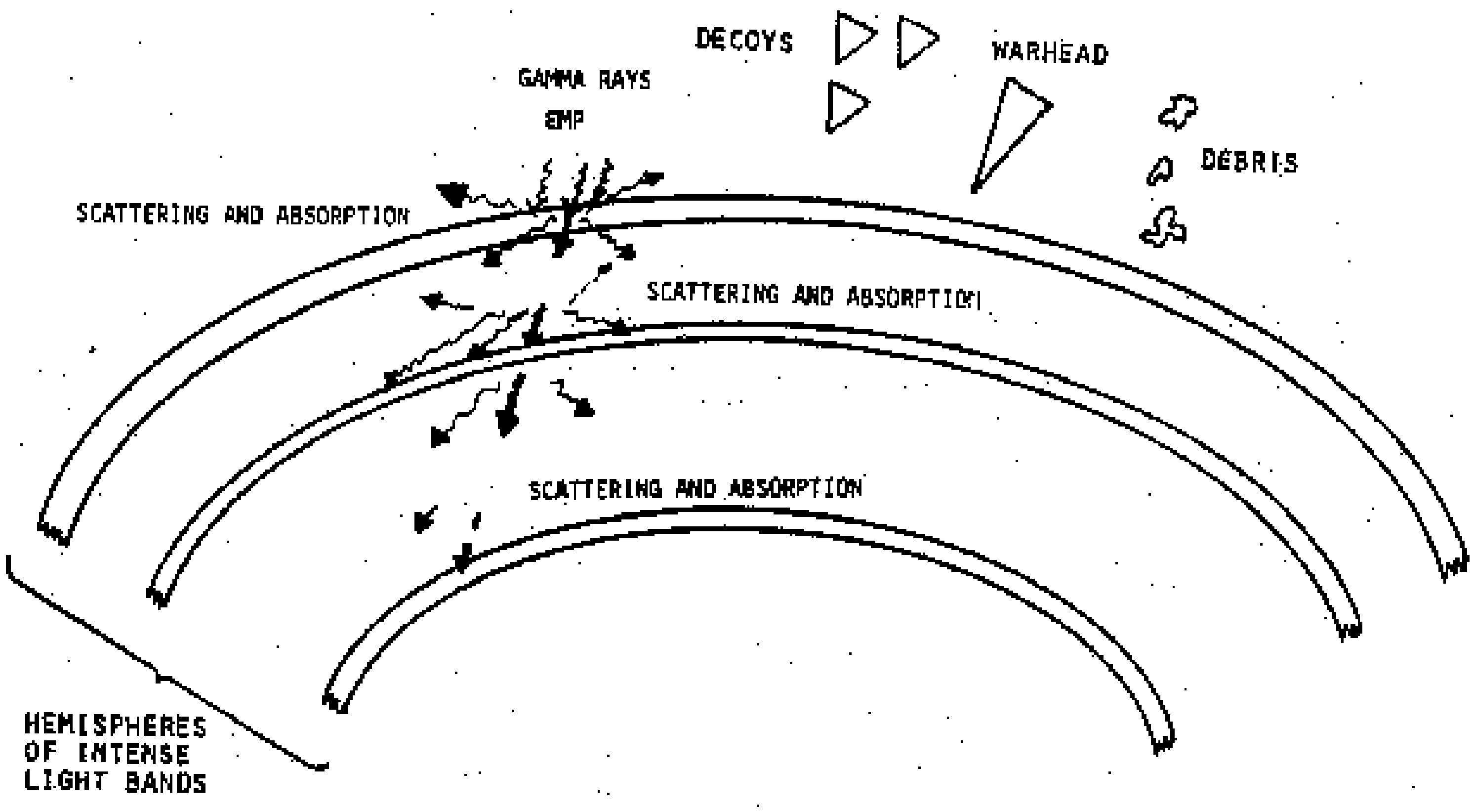
DECOYS
GAMMA RAYS
EMP
WARHEAD
DEBRIS
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
SCATTERING AND ABSORPTION
HEMISPHERES
OF INTENSE
LIGHT BANDS

Figure 9. Tesla Terminal Area Defense System

In the pulse mode, a single intense 3-dimensional scalar phi-field pulse form is fired, using two truncated Fourier transforms, each involving several frequencies, to provide the proper 3-dimensional shape (Figure 10). This is why two scalar antennas separated by a baseline are required. After a time delay calculated for the particular target, a second and faster pulse form of the same shape is fired from the interferometer antennas. The second pulse overtakes the first, catching it over the target zone and pair-coupling with it to instantly form a violent EMP of ordinary vector (Hertzian) electromagnetic energy. There is thus no vector transmission loss between the howitzer and the burst. Further, the coupling time is extremely short, and the energy will appear sharply in an "electromagnetic pulse (EMP)" strikingly similar to the 2-pulsed EMP of a nuclear weapon.

This type weapon is what actually caused the mysterious flashes off the southwest coast of Africa, picked up in 1979 and 1980 by Vela satellites. The second flash, e.g., was in the infrared only, with no visible spectrum. Nuclear flashes do not do that, and neither does superlightning, meteorite strikes, meteors, etc. In addition, one of the scientists at the Arecibo Ionospheric Observatory observed a gravitational wave disturbance -- signature of the truncated Fourier pattern and the time-squeezing effect of the Tesla potential wave -- traveling toward the vicinity of the explosion.

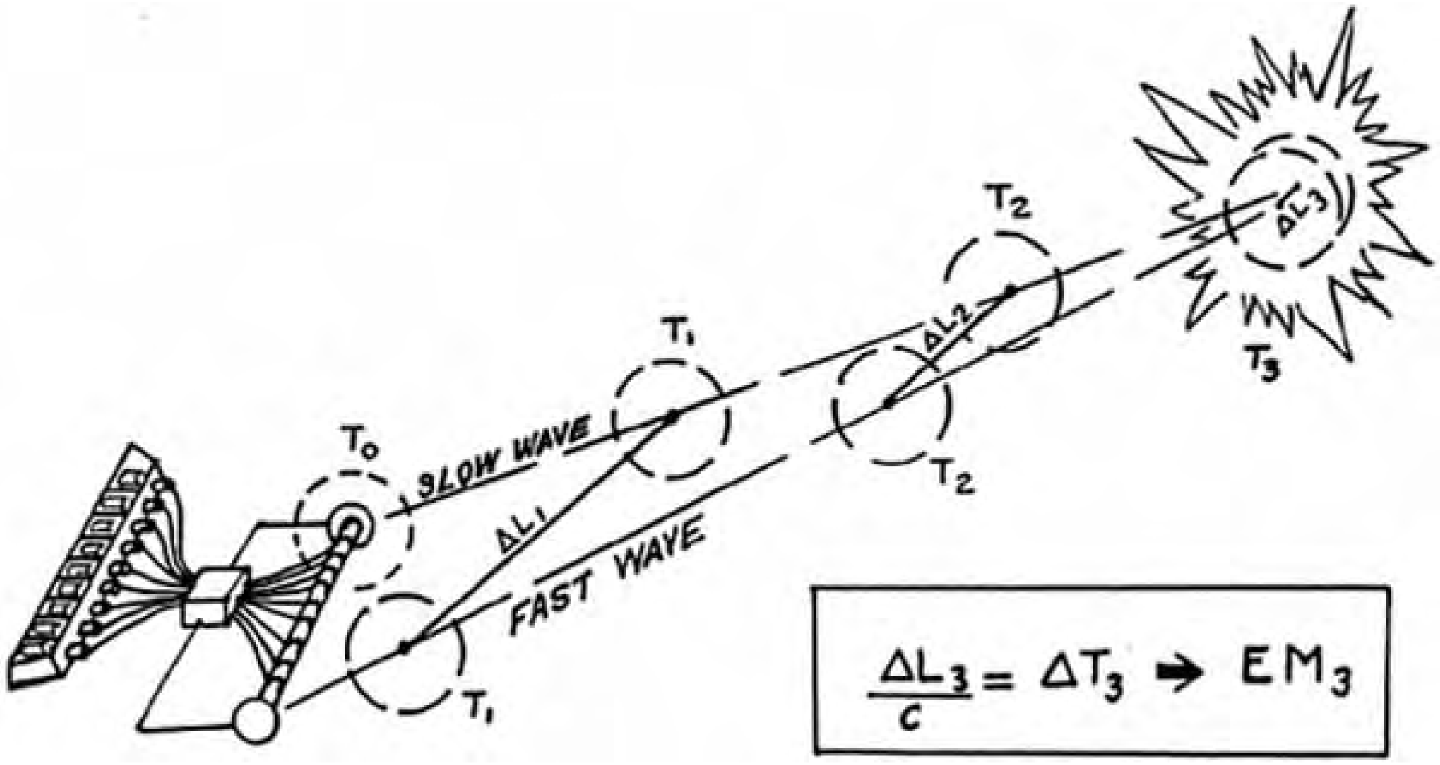
T₀
SLOW WAVE
FAST WAVE
ΔL1
T₁
T₁
T₂
ΔL2
T₂
ΔL3
T₃
ΔL₃/c = ΔT₃ → EM₃

TESLA HOWITZER
(SCALAR INTERFEROMETER)

Figure 10. "Nuclear" Flashes off the Coast of Africa

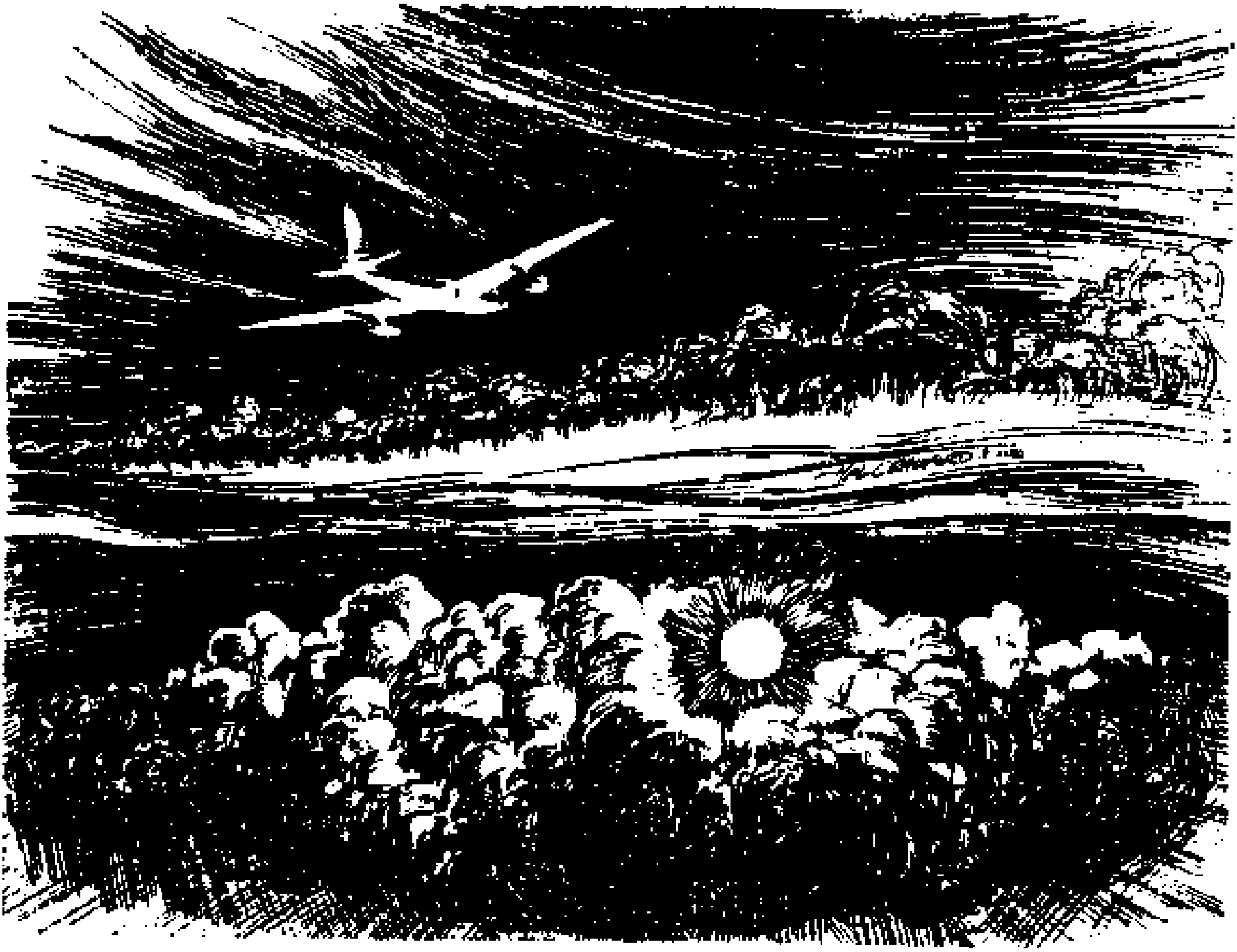

Lithuania - 10 Sep 1976 - British European Airways Flight #831 between Moscow and London

CIA Report Released under FOIA

Figure 11. Continuous Tesla Fireball

The pulse mode may be fed from either Moray generators or -- if the Moray generators have suffered their anomalous "all fail" malfunction -- ordinary explosive generators. Thus the Tesla howitzer can always function in the pulse mode, but it will be limited in power if the Moray generators fail.

In the continuous mode, two continuous scalar waves are emitted -- one faster than the other -- and they pair-couple into vector energy at the region where they approach an in-phase condition. In this mode, the energy in the distant "ball" or geometric region would appear continuously and be sustained -- and this is Tesla's secret of wireless transmission of energy at a distance without any losses. It is also the secret of a "continuous fireball" weapon capable of destroying hundreds of aircraft or missiles at a distance. An example of a Soviet test of this mode of operation is shown in figure 11.

Witness to a super weapon?

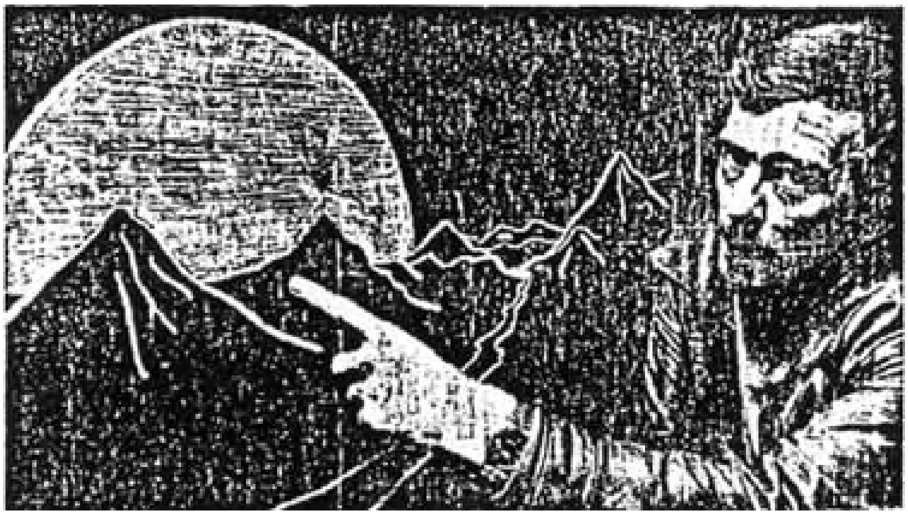

Nick Downie describes the strange lurid glow that flared silently over the Hindu Kush

THE SUNDAY TIMES, 17 AUGUST 1980

(Multiple incidents in Sept., 1979)

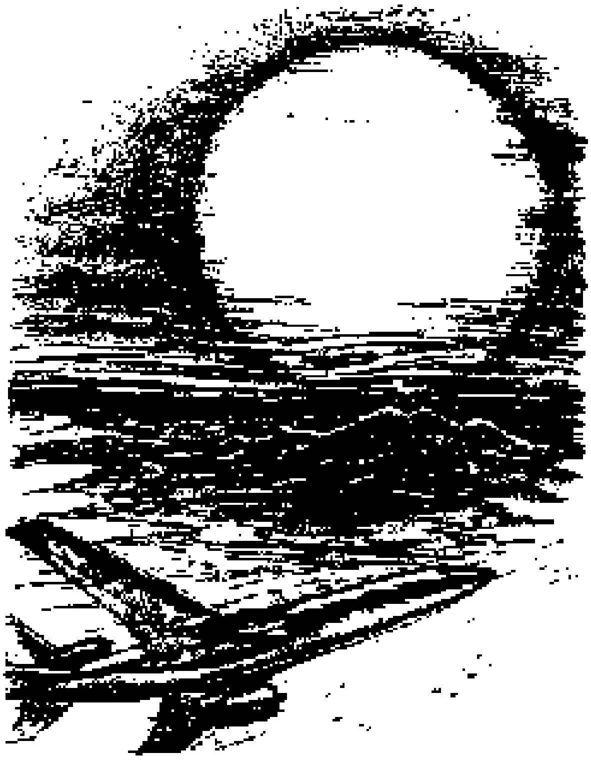

TEHERAN, IRAN 17 JUNE 1966

SEEN FROM NEAR MEHRABAD AIRPORT

OBSERVED 4-5 MINUTES

SEEN BY 2 AIRCRAFT

CIA REPORT RELEASED UNDER FOIA

Figure 12. Tesla EMP Globe

The volume of the Tesla fireball can be vastly expanded to yield a globe which will not vaporize physical vehicles but will deliver an EMP to them to dud their electronics. A test of this mode is shown in figure 12. (See also Gwynne Roberts, "Witness to a Super Weapon?", the London Sunday Times, 17 August 1980 for several other tests of this mode at Saryshagan, seen from Afghanistan by British TV cameraman and former War Correspondent Nick Downie.)

If the Moray generators fail anomalously, then a continuous mode limited in power and range could conceivably be sustained by powering the interferometer from more conventional power-sources such as advanced magnetohydrodynamic generators.

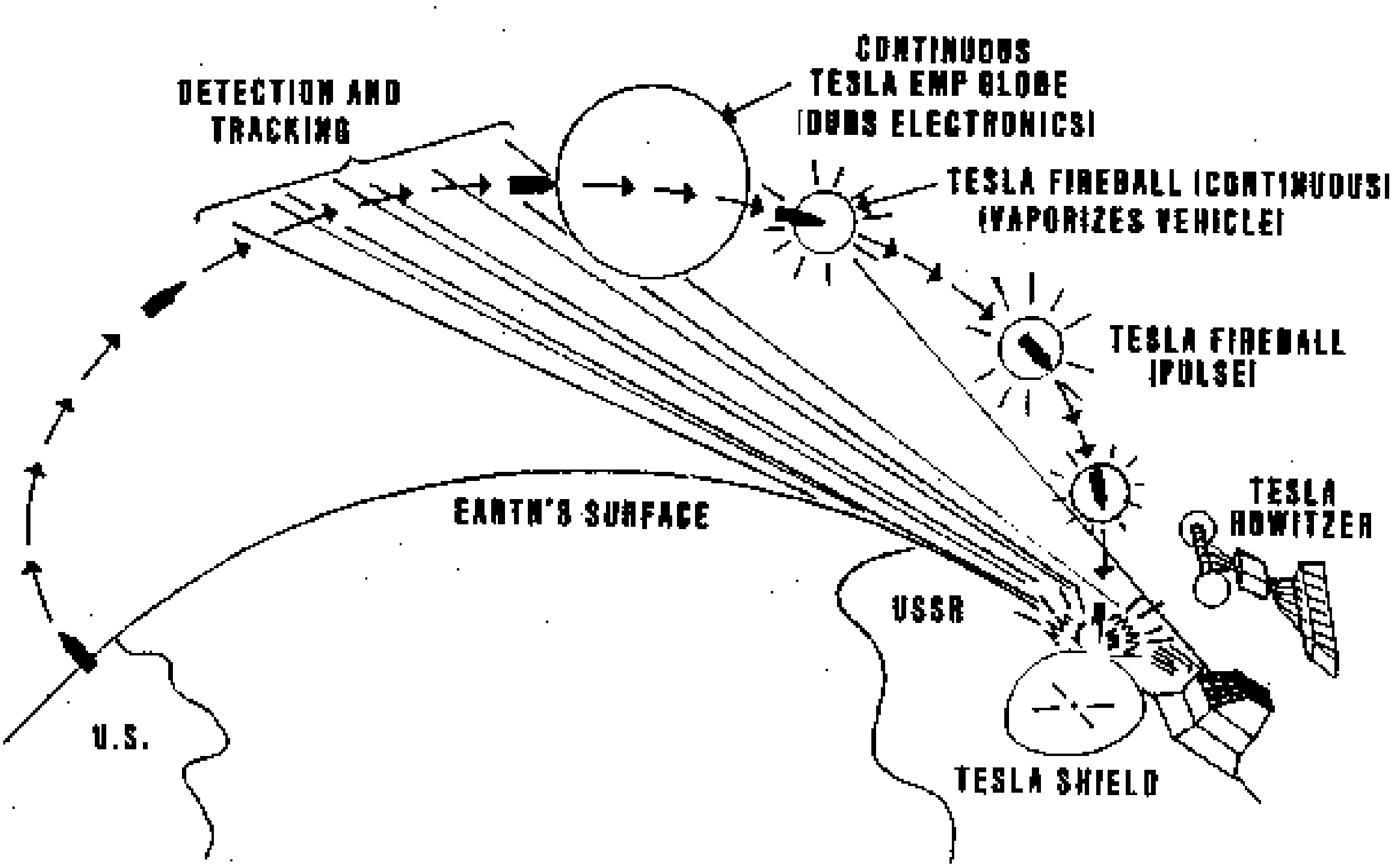
DETECTION AND TRACKING
CONTINUOUS TESLA EMP GLOBE (DUDS ELECTRONICS)
TESLA FIREBALL (CONTINUOUS) (VAPORIZES VEHICLE)
TESLA FIREBALL (PULSE)
TESLA HOWITZER
EARTH'S SURFACE
USSR
U.S.
TESLA SHIELD

Figure 13. Tesla ABM Defenses

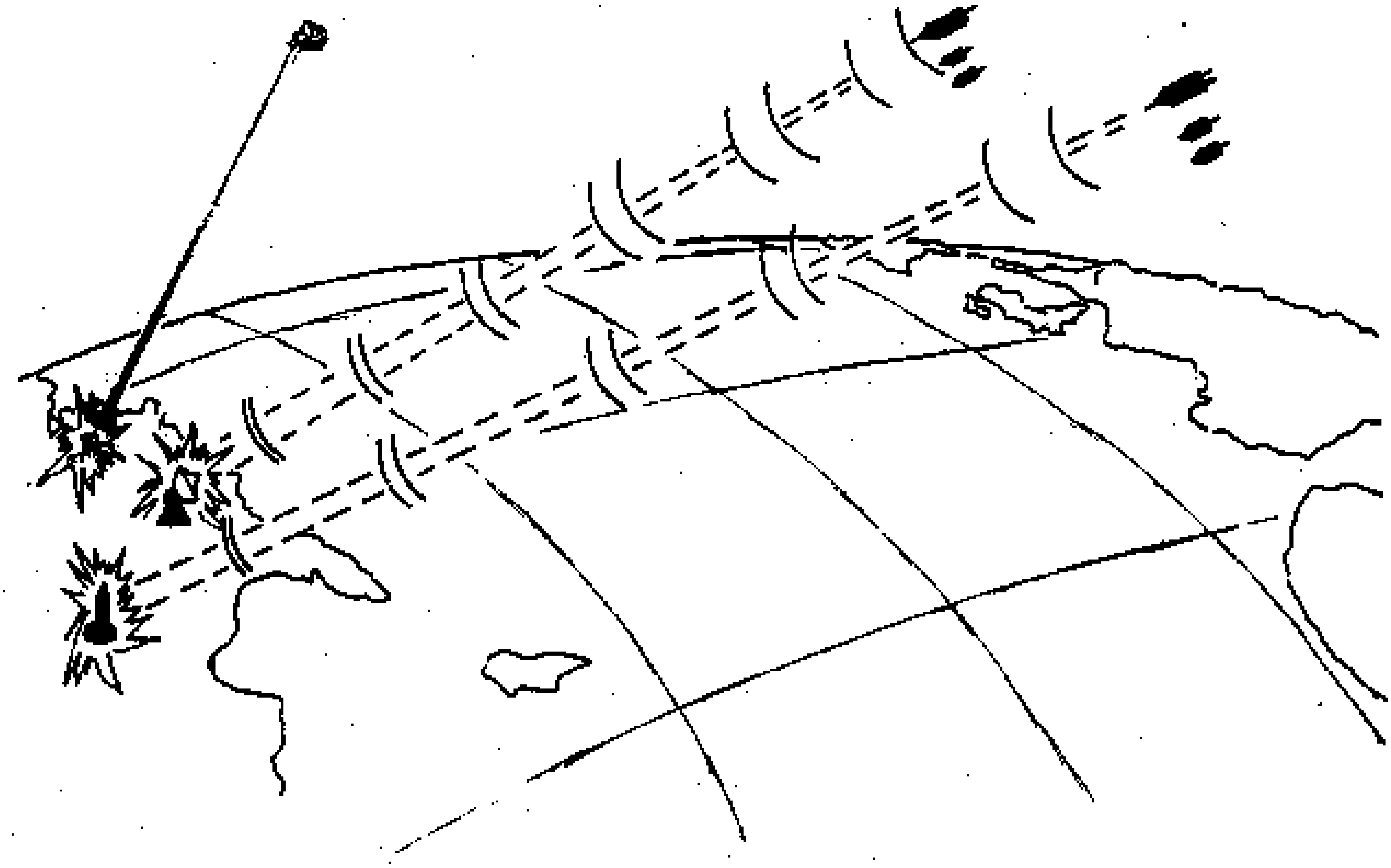

Figure 14. Moray/Tesla Technology: Star Wars Now

Typical strategic ABM uses of Tesla weapons are shown in figure 13. In addition, of course, smaller Tesla howitzer systems for anti-tactical ballistic missile defense of tactical troops and installations could be constituted of more conventional field missile systems using paired or triplet radars, of conventional external appearance, in a scalar interferometer mode.

With Moray generators as power sources and multiply deployed reentry vehicles with scalar antennas and transmitters, ICBM reentry systems now can become long range "blasters" of the target areas, from thousands of kilometers distance (figure 14). Literally, "Star Wars" is liberated by the Tesla technology. And in air attack, jammers and ECM aircraft now become "Tesla blasters." With the Tesla technology, emitters become primary fighting components of stunning power.

The potential peaceful implications of Tesla waves are also enormous. By utilizing the "time squeeze" effect, one can get antigravity, materialization and dematerialization, transmutation, and mindboggling medical benefits. One can also get subluminal and superluminal communication, see through the earth and through the ocean, etc. The new view of phi-field also provides a unified field theory, higher orders of reality, and a new super-relativity, but detailing these possibilities must wait for another book.

With two cerebral brain halves, the human being also has a Tesla scalar interferometer between his ears. And since the brain and nervous system processes avalanche discharges, it can produce (and detect) scalar Tesla waves to at least a limited degree. Thus a human can sometimes produce anomalous spatiotemporal effects at a distance and through time. This provides an exact mechanism for psychokinesis, levitation, psychic healing, telepathy, precognition, postcognition, remote viewing, etc. It also provides a reason why an individual can detect a "stick" on a radionics or Hieronymus machine (which processes scalar waves), when ordinary detectors detect nothing. Unfortunately there is not room to develop the implications of this human Tesla interferometry in detail, for that must wait for yet another book, presently in its initial stages, that Hal Crawford and I are writing.

Table 5. Orders of Reality

- **PHOTONS ARE**
  - **> PAIR-COUPLED SCALARS**
  - **> VELOCITY-LIMITED TO C**
  - **> CARRIERS OF T**

- **PHOTON INTERACTION**
  - **> IS UBIQUITOUS**
  - **> PRODUCES $\dot{T}$ = C**
  - **> YIELDS 1st ORDER REALITY**

- **SCALAR O WAVES**
  - **> NOT VELOCITY-LIMITED**
  - **> YIELD HIGHER ORDER REALITIES**

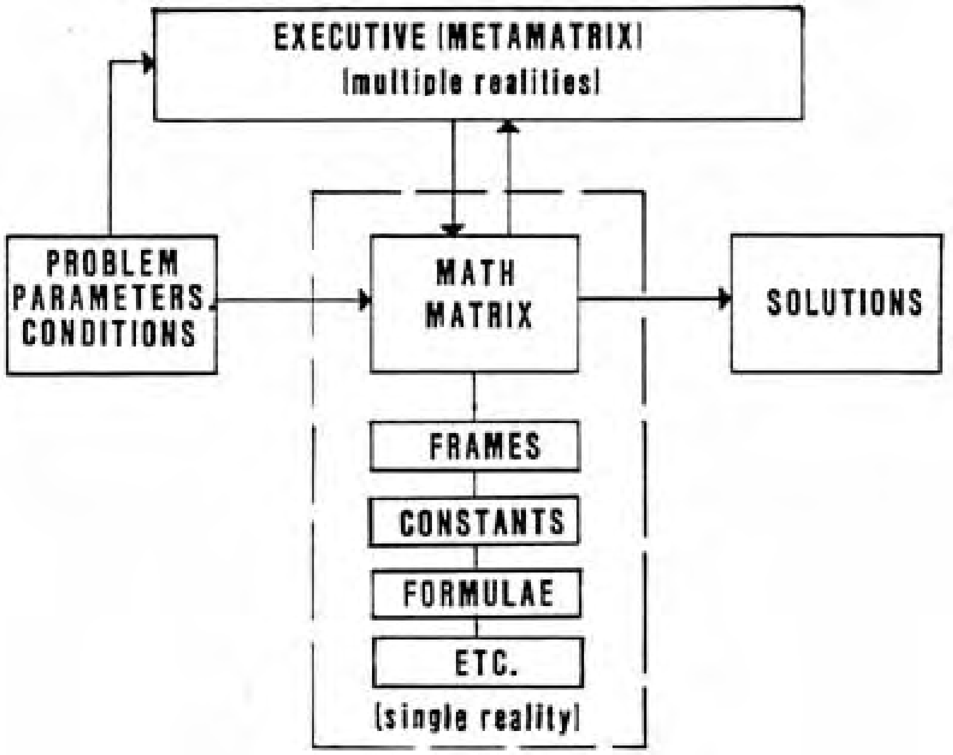
EXECUTIVE (METAMATRIX)
(multiple realities)
PROBLEM PARAMETERS, CONDITIONS
MATH MATRIX
SOLUTIONS
FRAMES
CONSTANTS
FORMULAE
ETC.
(single reality)

Figure 15. Implications of Tesla Potential

At the July 1981 U.S. Psychotronics Association's Annual Conference in Dayton, Ohio, I presented the first rough paper on the Tesla secret and scalar interferometry. A videotape of the presentation was made and will shortly be available. I am also scheduled to make a special presentation at the Alternate Energy Conference in Toronto, Canada in latter October, 1981. A professional, videotaped two-hour presentation on this subject is also being prepared. Wide distribution of the material through the international underground physics and technology network has already been made. This time, God willing, Tesla's secret will not be suppressed for another 80 years!

And perhaps it is not yet too late. The material has cost me (now) some 16 years of agonizing labor and nearly $100,000 of my own personal funds. No orthodox university, scientific group, foundation, or governmental agency would support such an effort, either financially or otherwise. Indeed, most ordinary journals will not even accept material on such matters. Nonetheless, the area is of overwhelming importance and I truly believe Tesla's lost secret will shortly affect the lives of every human being on earth.

Perhaps with the free and open release of Tesla's secret, the scientific and governmental bureaucracies will be. shocked awake from their slumber, and we can develop defenses before Armageddon occurs. Perhaps there is hope after all -- for even Brezhnev, in his strange July, 1975 proposal to the SALT talks, seemed to reveal a perception that a turning point in war and weaponry may have been reached, and that human imagination is incapable of dealing with the ability to totally engineer reality itself. Having tested the weapons, the Soviets must be aware that the ill-provoked oscillation of timeflow affects the minds and thoughts -- and the very lifestreams and even the collective species unconsciousnesses -- of all lifeforms on earth. They must know that these weapons are two-edged swords, and that the backlash from their use can be far more terrible to the user than was the original effect to his victim.

If we can avoid the Apocalypse, the fantastic secret of Nikola Tesla can be employed to cure and elevate man, not kill him. Tesla's discovery can eventually remove every conceivable external human limitation. If we humans ourselves can elevate our consciousness to properly utilize the Tesla electromagnetics, then Nikola Tesla -- who gave us the electrical twentieth century in the first place -- may yet give us a fantastic new future more shining and glorious than all the great scientists and sages have imagined.

References

The Tom Bearden Website

REFERENCES

1. Whittaker, E. T., Proc. Lond. Math. Soc. 1, 367 (1903).

2. Debye, P., Ann. Phvs. (Leipz.) 30, 57 (1909).

3. Bromwich, T. J. I'A. Phil. Trans. A, 220, 175 (1920); also Phil. Mag., 38, 143 (1919).

4. Laporte, O. & Uhlenbeck, G. E., Phys. Rev. 37, 1380 (1931).

5. Nisbet, A., Proc. R. Soc. London A 231, 250 (1955).

6. Essex, E. A., Am. Jour. Phys. 45, 1099 (1977).

7. Braunlich, P., Ed., Thermally Stimulated Relaxation in Solids, Springer-Verlag, New York, 1979.

8. Ratzlaff, John. T., Dr. Nikola Tesla: Selected Patent Wrappers, Volumes I, II, III, & IV, Tesla Book Company, 1580 Magnolia Avenue, Millbrae, CA 94030, 1980.

9. Ratzlaff, John T. & Anderson, Leland I., Dr. Nikola Tesla Bibliography, Ragusan Press, 936 Industrial Ave., Palo Alto, CA 94303, 1979.

10. "Tesla- 85th Birthday," N. Y. Sun, July 11, 1941. (Tesla says he could build, within three months, a plant at a cost of $2 million that would melt the engines of an approaching aircraft at a distance of thousands of miles.)

11. "Tesla Promises to Light Dark Spot on Moon. It's Part of a Scheme of His for Interplanetary Radio; Distance Means Nothing. Has 4 New Inventions. Tells of Them on 81st Birthday; 2 Nations Honor Him," N. Y. Herald Tribune, July 11, 1937.

12. O'Neill, John J., "In the Realm of Science: Tesla, who predicted radio, now looks forward to sending waves to the Moon," N. Y. Herald Tribune, Aug. 22, 1937. (Inventor hopes to use energy-transmitting device to make spot glow on the lunar surface. Theory is traced to 1897 experiments. His mechanism is to use vast natural forces, possibly cosmic rays.)

13. "Tesla, 80, Reveals New Power Device," N. Y. Times, July 11, 1936, p. 13, col. 2. (Says his wireless system of power transmission will supply the earth with energy for industry.)

14. Sparling, Earl. "Nikola Tesla, at 79, Uses Earth to Transmit Signals; Expects to have $100,000,000 Within Two Years," N. Y. World-Telegram, July 11, 1935. (Inventor tells of 'quake' in his laboratory that brought police and ambulances during experiments with mechanical oscillator.)

15. "Tesla's Controlled Earth Quakes Power Through the Earth, A Startling Discovery," N. Y. American, July 11, 1935, Section 2. (Announces the successful passage of an induction current with a varying flux through a circuit without the use of a commutator. Cosmic ray studies indicate many tenets of theory of relativity to be fallacious. Possible to convey mechanical effects to any distance.)

16. "Tesla, 79, Promises to Transmit Force -- Transmission of Energy Over World," N. Y. Times, July 11, 1935, p. 23, col. 8. (Tesla has plans to send energy over the entire world. Measurement of cosmic rays said to be 50 times greater than the speed of light, demolishing the theory of relativity.)

17. Tesla, Nikola, "Expanding Sun Will Explode Some Day, Tesla Predicts," N. Y. Herald-Tribune, Aug. 18, 1935. (Present literature on cosmic rays is erroneous. Some cosmic rays reach speeds 50 times that of light. Sun will increase in mass and energy and will ultimately explode. Tesla's view is that the condensation of primary substance is going on continuously. Finds secret of cosmic rays in the positive electrical charge of the sun. Discusses radioactive emanations.)

18. "3 Tesla Inventions -- Famous Scientist Will Tell Them. Tomorrow," N. Y. Sun, July 9, 1935. (One of the discoveries is a new way of transmitting energy, an entirely new principle, nothing like wireless. Also method of harnessing cosmic rays.)

19. Welshimer, Helen, "Dr. Tesla Visions the End of Aircraft in War," Every Week Magazine, Oct. 21, 1934, p. 3. (Claims to have created a new agent, which kills without a trace and yet pierces the thickest armor. Can destroy armies or aircraft.)

20. Tesla, Nikola. "Tesla on Power Development and Future Marvels," N. Y. World Telegram, July 24, 1934. (Source of reference is Prodigal Genius by J. J. O'Neill, p. 241. Reply to articles of June 29, July 12, and July 13, 1934. Praises Westinghouse and Insul to create the power system he had barely suggested in 1893. With regard to death ray effect, Tesla employs an agent

in which intensity does not diminish with the square of the distance.)

21. Dunlap, Orrin E., Jr., "Tesla Sees Evidence That Radio and Light Are Sound, " N. Y. Times, Apr. 8, 1934, X, p. 9, Col. 1. (Tesla points to errors of the past, explains radio as he sees it at age 77. He expects television.). (Note: Sound is a longitudinal wave -- as is the Tesla wave. Hertzian waves are transverse waves, not longitudinal.)

22. Bird, Carol, "Tremendous New Power Soon to be Unleashed," Philadelphia Public Ledger, Sep. 10, 1933, Magazine Section, p. 6. (Revolutionary power project by Tesla, who is also completing process for thought photography.) (Note: Here we have a possible clue that the same principle may be used both in Tesla's energy device and in Tesla's approach to thought photography.)

23. "Tesla 'Harnesses' Cosmic Energy," Phila. Public Ledger, Nov. 2, 1933. (A principle has been discovered to derive cosmic energy which operates the universe. Power is "everywhere present in unlimited quantities." Will eliminate the need for coal, oil, gas, or any of the common fuels.) (Note: Was Tesla referring to what we today call "zero-point energy of the vacuum"? His principle will tap it.)

24. Blakeslee, Howard W., "Discovery of Force to Surround Nations & Smash Attacker Claims of Aged Inventor - Nikola Tesla Makes Announcement on 75th Birthday -- Will Turn Plans Over to Geneva," Minneapolis Tribune, July 11, 1934. (Note: Here the inventor is referring to the "Tesla Shield.")

25. "Tesla, at 78, Bares New 'Death Beam'", N. Y. Times; July 11, 1934, p. 18, col. 1. (Invention powerful enough to destroy 10,000 airplanes at 250 miles away. Defense weapon only.)

26. Alsop, Joseph W., Jr., "Beam to Kill Army at 200 Miles Tesla's Claim on 78th Birthday," N. Y. Herald Tribune. July 11, 1934, pp. 1, 15. (Beam of force similar to death ray, involves four electrical devices. Can also be used in peacetime to transmit power over distances limited only by the curvature of the earth.) (Note: In this reference, we note that the Tesla death ray and the Tesla wireless transmission devices apparently use the same effect or basic principle. There appear to be four devices involved in an operational system.)

27. "A Giant Eye to See 'Round the World' ", Albany Telegram, Feb. 25, 1933. (Based on the mechanism of the human eye. The first two parts of Tesla's invention have been completed. Will allow man to see any part of the earth.) (Note: Similarity of binocular vision to interferometry.)

28. Tesla, Nikola. "Pioneer Radio Engineer Gives Views on Power," N. Y. Herald Tribune, Sept. 11, 1932. (Tesla says wireless waves are not electromagnetic, but sound in nature. Holds space is not curved.) (Note again that sound waves are longitudinal, as are Tesla's waves. Electromagnetic waves are transverse.)

29. "No High-Speed Limit, Says Tesla, " Literary Digest, Nov. 7, 1931, p. 28. (Speeds greater than light, deemed impossible by the Einstein theory, have been accomplished. As early as 1900 Tesla showed that power from his transmitter passed over the earth at a speed of 292,830 miles per second.) (Note: As is well known, velocity represents the rotation of an object out of the normal 3-dimensional space toward a 4th dimensional direction. The speed of light, c, represents a full orthogonal turn. A normal electromagnetic wave, being transverse oscillatory, is a priori limited to a single orthogonal turn, for that is all it possesses. It thus moves at the speed c. A longitudinal scalar wave, on the other hand, need not be so limited at all.)

30. O' Neill , J. J. "Sun Emits Super Ray, Nikola Tesla Asserts, " Brooklyn Eagle, Feb. 8, 1932, p. 4, col. 1. (Note: Logically, the sun should also emit scalar waves as well as transverse vector waves.)

31. O'Neill, J. J., "Tesla Cosmic Ray Motor May Transmit Power 'Round Earth'", Brooklyn Eagle, July 10, 1932, A, 1:4, pp. 1, 17. (Efforts by Tesla to harness cosmic rays started 25 years ago, and he now announces success in operating a motive device by means of these rays. Hopes to build a motor on a large scale.)(Note: possibly implies that all these Tesla devices for the last 25 years have been dealing with the same basic principle.)

32. "Tesla, 76, Reports His Talents at Peak," N. Y. Times, July 10, 1932, p. 19, col. 1. (New invention in tapping tremendous and unused source of energy. One invention to permit generation of all kinds of rays of almost unlimited intensity.) (Note: By scalar interferometry, electromagnetic waves of any frequency -- and hence of any "type" can be assembled by coupling the appropriate scalars together. A Hertzian wave is just two coupled Tesla scalar waves.)

33. Tesla, Nikola, "Man's Greatest Achievement," N. Y. American, July 6, 1930, p. 10. (Editorial Section). ("To create and annihilate material substance, cause it to aggregate in forms according to his desire...would place him beside his Creator and fulfill his ultimate destiny.") (Note: Here Tesla seems to be referring to controlled materialization and dematerialization of matter.)

34. Tesla, Nikola. "World System of Wireless Transmission of Energy," Telegraph & Telephone Age -- N. Y., Oct. 16, 1927, pp. 457-460.

Transmission of power without wires is not a theory, but "a fact demonstrated by Tesla." Mode of propagation of currents from transmitter has mean speed 57% greater than Hertz waves.) (Note: a faster than light, non-Hertzian type of wave is involved. The Tesla wave is not a normal Hertzian wave.)

35. Secor, H. Winfield, "The Rogers Underground Wireless, Electrical Experimenter, Mar., 1919, pp. 787-789, 832-835, 839. (Rogers system does away with aerial wires.)

36. Gernsback, Hugo, "Underground Wireless," Electrical Experimenter, Mar. 1919, p. 762. (Development of James H. Rogers. Receives messages from Europe during thunderstorm. Tesla agrees that messages are not Hertzian waves.)

37. Tesla, Nikola, "The True Wireless," Electrical Experimenter, May 1919, pp. 28-30, 61-63, 87. (Tesla says Hertz wave theory is a delusion. Signals must be from earth currents. Note again that Tesla does not like Hertzian waves, but himself has something more fundamental.)

38. Tesla, Nikola. "The Effects of Statics on Electrical Transmission," Electrical Experimenter, Jan., 1919, pp. 627, 658. (Tells of defects in Hertz waves. System devised by Tesla releases energy at infinite velocity.) (Note: electrostatic potential is already known to be able to travel at infinite velocity. See Jackson, Classical Electrodynamics, 2nd Edition, 1975, p.223. This is strong evidence that Tesla used scalar potential waves or Ø-fields.)

39. Bot tone, A., Nikola Tesla's New Wireless," Electrical Engineer, London, Dec. 24, 1909, p. 893. (Experiments conducted on Long Island have resulted in principles of transmission which are direct opposite of Hertzian wave transmission.) (Note: In other words, using uncoupled scalar waves -- Tesla waves instead of pair-coupled scalar waves -- Hertzian waves?)

40. Tesla, Nikola. "Electrical Control of the Weather Will Soon Be an Accomplished Fact," St. Louis Republic, Nov. 15, 1908, V, p. 3. (Tesla concurs in concept of electrical control of weather.)

41. Swezy, Kenneth M. , "Nikola Tesla, " Science, May 16, 1958 , pp. 1147-1158. (Electricity today is generated, transmitted, and converted to mechanical power by means of his inventions.)

42. Ramsay, Jack, "Cableless Power, " Electronics (Digest) , July 22, 1960, pp. 6, 8. (Soviet attempt at development of wireless power transmission, but a failure according to Swedish observers, and the idea having been bandied about in the U. S. since Tesla.) (Note: Establishes that the Soviets were at least trying to work with the Tesla techniques.)

43. Anderson, Leland I., "Correspondence: Sub-Surface Communications Systems," Proceedings of the I.R.E., Mar., 1961, p. 645. (Reference to subsurface or "earth current" communication systems. Pioneering work by Nikola Tesla and James H. Rogers.)

44. Rebert, Edwina, "Fireballs For Defense?", Christian Science Monitor, Feb. 7, 1962, p. 9c. (Feature article on possibility of using ball lightning as defense weapon -- may be inspired by early experiments of Tesla at Colorado Springs.)

45. Matsch, Lee and Rice, Warren, "Potential Flow Between Two Parallel Circular Disks with Partial Transmission," Journal of Applied Mechanics, Trans. ASME, Mar., 1967, Vol. 34, Series E., No.1, pp. 239-240. (Reference article -- bladeless turbines.)

46. Lagus, George W., "The Use of Ground in High-Frequency Circuits," Radio-TV Experimenter and Science and Electronics, June-July, 1969, PP. 71, 72. (Tesla gave several demonstrations of feasibility of electrifying large areas of land by ground propagation of high frequency currents.)

47. Jamison, S. L. , "Life Energy, " Probe. The Unknown, June , 1973, pp. 3,4. (Letter to editor -- report that vitamins and minerals could be impressed on the body using a Tesla coil.)

48. Jueneman, F. B. J. , "The First Law of Thermodynamics, " Industrial Research, Feb., 1974, pp. 17-18. (Nikola Tesla, at Colorado Springs in 1899, found that the earth was a large electron sink, and electrical storms generated standing wave potentials.)

49. Puharich, Andrija, "The Work of Nikola Tesla Ca. 1900 and its Relationship to Physics, Bioenergy and Healing," paper read at the International Inter-disciplinary Conference on Consciousness and Healing, Oct. 13, 1976, University of Toronto.) ,

50. Golka, Robert K. and Bass, Robert W., "Tesla's Ball Lightning Theory, a BKG-Wave, The Ransworth Effect, and the 'Hydrotron' Electrostatic-Inertial Self-Confined Plasmoid Concept." (Paper presented at the Annual Controlled Fusion Theory Conference, San Diego, California, May 4-6, 1977.)

51. Popovic, Prof. Vojin, "Research of Nikola Tesla in the Light of His Diary From Colorado Springs," 10 pp. (Diary contains details of investigations for determining nodal points in the earth as a conductor of waves. Comments on "fire globes.")

52. Curtis, George D., PhD, "An Electromagnetic Radiation Pattern Over the Ocean," Undersea Technology, Vol. 5, no.8, August 1964. Curtis reports the presence of a previously unknown weak electromagnetic radiation pattern over the ocean. When all known effects are accounted for, the anomalous pattern still remains.

53. Santilli, R. M., "Partons and Gravitation: Some Puzzling Questions," Annals of Physics, Vol. 83, No.1, March 1974, pp. 108-157. In this paper Santilli proved that one of the cornerstone assumptions of physics -- that electric field and gravitation are separate things -- is false. One is therefore left with only two alternatives: either they are totally the same thing, or they are partially the same thing. (Comment: Note that the ~4 Tesla potential provides the unifying connection.)

54. Science News, Vol. 113, No.1, January 7, 1978, p. 3. A photograph by T. Kuribayashi of the earthquake lights involved in the Matsushiro earthquake swarm (1965-1967). This is the only known photo of earthquake lights.

55. Wiedemann, C. Louis, "Results of the N.J. 'Spook Light' Study," Vestigia Newsletter, Vestigia, RD 2, Brookwood Rd., Stanhope, NJ 07874, May 1977, pp. 1-3. This article describes an ongoing Vestigia experiment in which a nocturnal light or "mystery light" was scientifically photographed in conjunction with visual sightings and instrumental recordings of data. Various later issues of the Vestigia Newsletter contain other update articles on the phenomenon.

56. Moray, T. Henry, The Sea of Energy, fifth edition, History and Biography by John E. Moray, Foreword by Tom Bearden; Cosray Research Institute, 2505 South 4th East, Salt Lake City, Utah 84115, 1978. In addition to his free-energy device, Moray also built a special radio which he used to clearly listen to Admiral Byrd at the South Pole. No static was experienced on Moray's radio, showing he was' not using ordinary electromagnetics. In addition, he built a device which he could tune to listen to persons several miles distant, so that their voices sounded as if they were immediately nearby. Both these devices were demonstrated to a visiting Russian Colonel/ PhD.

57. I. Procaccia, J. Ross, Science, 198, 716 (18 Nov 1977). Describes Prigogine's Nobel Prize-winning work on thermodynamics of nonlinear systems far from thermodynamic equilibrium. Prigogine's work shows that such a system can indeed demonstrate negentropy. Out of unrestrained disorder can and does arise order, contrary to the old thermodynamics.

58. "Persinger's 'Earthquake Lights' ...Ho-Hum," Frontiers of Science, Vol. III, No.3, March-April 1981, p. 15-16. Gives cogent and "fatal" objections to

Persinger's "earthquake light" scheme for explaining UFO's. Mentions Dr. Brian Brady's experiments obtaining fast, short electrical lights or sparks from crushing quartz-bearing granite cylinders. (Comment: Brady's experiments are certainly repeatable, so it is true that "earthstress lights" can be produced at a distance, although the piezoelectric effect alone in no way explains the production of the phenomena outside the piezoelectric crystal. The fact that UFO's occurring near a fault zone are not necessarily repeatable simply means that at least one other major factor must be involved in the production of UFOs than is involved in the production of earthstress lights. UFOs may sometimes involve scalar interferometry from a fault zone, but other causative factors must also be involved. Earthstress lights, on the other hand, need have no additional causative factors than what lies in the earth itself.)

59. Jackson, John David, Classical Electrodynamics, Second Edition, John Wiley & Sons, New York, 1975, p. 223.

60. Roberts, Gwynne, "Witness to a Super Weapon ?", The Sunday Times, London, 17 August, 1980. Article deals with a strange semicircular phenomenon seen inside the Soviet Union from Afghanistan in September, 1979 by Nick Downie, a former member of the SAS who has established a considerable reputation as a war cameraman. Downie saw the phenomenon twice, and later met an Afghan who had seen it several months before. A possibly related phenomenon is reported in Brackenbridge, M., "Unidentified Phenomenon," Marine Observer, 48: 21-22, 1978. Note the first "nuclear flash" was detected by Vela satellite on September 22, 1979 - close in time to Downie's sightings.

61. Bearden, Tom and Crawford, Hal, "Possible Soviet Test of a Tesla Weapon," Specula, Journal of the A.A.M.S., Vol. 3, No.2, April-June 1980, p. 29. See also Specula 3, 2, 30-32 by Bearden and Crawford. These reports are from CIA reports released under the Freedom of Information Act.

62. Bearden, Thomas E., The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980.

63. Muldrew, D. B., "Generation of Long-Delay Echoes," Journal of Geophysical Research, 84: 5199-5215, 1979. For anomalous long delays of electromagnetic signals in the atmosphere, Muldrew favors a rather complex interaction between signals from separate transmitters that (theoretically at least) can create a long-lived electrostatic wave that travels in the ionosphere -- a sort of natural memory device. The coded signals could then be read out much later when the proper natural conditions developed. Delays of up to 40 seconds might be possible with this "ionospheric memory."

64. Zhugzhda, Yu. D., Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation, "Magnetogravitational Waves in a Conducting Isothermic Atmosphere," Moscow ASTRONOMICHESKIY ZHUGZHDA, Vol. 56, No.1, 1979, pp. 74-83. Among other things, the paper indicates the possibility of transforming longitudinal waves into transverse waves in the region of a strong magnetic field.

65. Ranada, Antonio and Vazquez, Luis, "Kinks and the Heisenberg Uncertainty Principle," Physical Review D, Vol. 19, No.2, 15 January 1979, pp. 493-495. Shows that the velocity and position of the center of a kink, soliton, or solitary wave can be known with arbitrary precision, in violation of the Heisenberg uncertainty principle. (Comment: A Fourier-transformed scalar Tesla wave pattern is either a soliton or closely related to it.)

66. Don Moser, Photographs by Blair Pittman, "Big Thicket of Texas, National Geographic, Vol. 146, No.4, October 1974, pp. 504-529. The author photographed a "ghost light" or nocturnal spooklite -- i.e., an earths tress light. The photo is shown in the article.

67. Gehring, Gillian, "Actinide magnetism: an extraordinary tale," Nature, Vol. 279, 3 May 1979, pp. 16-18. Anomalous magnetic properties -- including spin orbit coupling as a large effect -- are exhibited by actinide intermetallics containing uranium. Present theory cannot explain these effects.

68. Gintsburg, M. A., Institute of Terrestrial Magnetism, the Ionosphere, and Radio-Wave Propagation, Academy of Sciences of the USSR, Astron. Zh. 51: 218-221, Jan-Feb 1974, English translation in Sov. Astron., Vol. 18, No.1, July-August 1974, pp. 128-130. Among other things, the article points out that thus far a Fourier analysis in mass electrodynamics has not been made. One property characteristic of mass electrodynamics is the existence of longitudinal electromagnetic waves in vacuo, i.e., longitudinal photons. They have not yet been discovered but theoretically they are entirely possible.

69. Burman, R., "A Photon rest mass and the propagation of longitudinal electric waves in interstellar and intergalactic space," J. Phvs. A: Math. Nucl. Gen., Vol. 6, March 1973, pp. 434-444. This paper on longitudinal electric waves in plasma may be very important. Weber may be detecting longitudinal electric waves. Proca waves can go faster or slower than c.

70. Klass, Philip J., "Anti-Satellite Laser Use Suspected," Aviation Week & Space Technology, December 8, 1975, pp. 12-13. Some details on the "laser blinding" of U. S. satellites by the Soviets. On one occasion, the satellite was blinded for 4 hours. (Comment: difficult to explain with current laser

technology, but simple to explain with Tesla scalar interferometry.)

71. "Soviets Build Directed-Energy Weapon," Aviation Week & Space Technology, July 28, 1980, pp. 57-50. Presents a background summary in Soviet particle beam weapon work, particularly the device being constructed at Saryshagan.

72. "Scientists Fail to Solve Vela Mystery," Science, Vol. 207, 1 February 1980, pp. 504-506. See also article in Science, 30 November, 1979.

73. "Satellite Evidence Shows 'Possibility' of Nuclear Test, DOD Says," Aerospace Daily, October 29, 1979, p. 286.

74. "A Flash of Light," Newsweek, November 5, 1979, pp. 64-65.

75. "Was It a Nuclear Device?", Newsweek, July 21, 1980, p. 19.

76. "A Nuclear Blast -- or 'Zoo Animals'?", Newsweek, April 7, 1980, p. 21.

77. "Diverging Views," Washington Roundup, Aviation Week & Space Technology, July 21, 1980, p. 15.

78. Klass, Philip J. , "Clandestine Nuclear Test Doubted," Aviation Week & Space Technology, August 11, 1980, pp. 67, 69, 71-72.

79. "Debate Continues on the Bomb That Wasn't," Science, Vol. 209, 1 August 1980, pp. 572-573.

80. "Navy Lab Concludes the Vela Saw a Bomb," Science, Vol. 209, 29 August, 1980, pp. 996-997.

81. Bloch, Ingram and Crater, Horace; "Lorentz-invariant potentials and the nonrelativistic limit," Am. J. Phys. 49(1), Jan. 1981, pp. 67-75. Among other things, this interesting paper points out some decidedly unusual and unfamiliar influences of the size of the scalar potential upon what is considered nonrelativistic behavior. For scalar potential energyof appreciable size relative to a particle's rest energy, ordinary Newtonian mechanics and the Schrodinger equation may be inadequate, even if v/c is small. This reference by itself is sufficient to establish that Ø-field alone can alter the flow rate of time.

82. Semon, Mark D. and Schmieg, Glenn M., "Note on the analogy between inertial and electromagnetic forces," Am. J. Phys. 49(7), July 1981, pp. 689-690. Compares physical forces in a rotating frame with electromagnetic forces in a Newtonian frame. The transverse mechanical force in a rotating

frame is associated with an induced EMF in a Newtonian frame.

83. Harlacher, Von Wolfgang M., "Bomben aus dem Hyperraun," Esotera, 4 April 1979, pp. 359-365.

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

TABLE OF CONTENTS

Cover

List of Figures

Implications

Electrical Physics Presently Has a Mindset

It Started With Geometry and Grew

Points and Motion

Four Types of Vectors Actually Emerged

Quantum Mechanics Compounds the Problem

To Summarize Briefly

Uncharged Spatial (Massless) System Vector

Chargeless Mass System Vector

What Force Is

Force and Hertzian Waves Cannot Exist in Vacuum

Charged-Mass-System Vector

Charged Spatial (Massless) Vector

The Shadow Vector

A Scalar is a Zero Vector

Virtual and Observable Aspects

Scalars and Vectors Have Substructures

Substructures, Virtual Levels, and Hyperspaces

Notes and References

Additional Notes and References

TOWARD A NEW ELECTROMAGNETICS

PART III: CLARIFYING THE VECTOR CONCEPT

by

T. E. Bearden

© Copyright 1983

by T. E. Bearden

All rights reserved

LIST OF FIGURES

1. Uncharged spatial (massless) system vector
2. Chargeless mass system vector
3. Detection of "transverse" and longitudinal waves
4. Charged mass system vector
5. The "charge" on an electron mass consists of a flux of virtual particles on and off the mass
6. A test charge (charged mass) brought near a fixed charge (charged mass) experiences an acceleration
7. A charged mass system vector
8. Repeating the "test charge" experiment
9. Charged spatial (massless) vector
10. Removing the bare particle (mass) from a charged particle leaves the charge. The vacuum is DEFINED AS the charge
11. Assigning a spatial vector to the charged vacuum
12. A "shadow vector" ε_s
13. Substructures of a "zero resultant" spatial vector. (The substructures of two zero vectors may be vastly different)

TOWARD A NEW ELECTROMAGNETICS

Part III: Clarifying the Vector Concept

© 1983 T.E. Bearden

-- IMPLICATIONS --

Some of the fundamental concepts of the new Tesla electromagnetics are presented. The new concepts have startling implications:

(1) No force or force field exists as such in vacuum.

(2) Hertzian (transverse) electromagnetic waves do not travel through the vacuum, just as Tesla stated.

(3) Forceless, massless Tesla (scalar) longitudinal waves actually transit the vacuum. Tesla called them "electrical sound waves."

(4) At present there are actually four different FUNDAMENTAL TYPES of vectorial entities in physics, erroneously confused as one and the same.

(5) Tesla longitudinal scalar waves are also "time" waves and can affect anything and everything that exists in time.

(6) The fundamental constants of nature (which exist in time) can be altered by Tesla scalar waves, which oscillate the values of the constants.

(7) Every vector and scalar has an internal substructure, which can be independently affected and changed. This allows the direct engineering of the virtual state and the vacuum itself.

(8) All observable forces (electrical, mechanical, gravitational, etc.) arise in, on, and OF the actual substructure of the "accelerating mass particle" itself, not as an "external" massless force or force field applied "to" a mass .

(9) Physical reality itself -- and the "physical laws of nature"

can be deliberately changed and engineered.

(10) All "physical reality" is totally internal to the physical changes of the mass particles of the detector system of the observer.

(11) Relativity's speed of light limitation applies only to the changes of the basic mass particles of the detecting instrument.

(12) Detection of superluminal effects cannot be accomplished by a "single stage" or "single shift " (single interaction) detector .

(13) Detection of superluminal effects is permitted by "multiple stage" or "multiple shift" interactions where the last interaction is a conventional interaction of photon vs. detector particle. (The two-slit apparatus for detection of electron diffraction is an example. First, the superluminal DeBroglie waves interact with the slits, which are "tuned" toward the electron's DeBroglie wavelength. The interaction with TWO slits produces subluminal interference effects, which then interact back upon the physical electron. The apparatus is thus an electron interferometer capable of detecting superluminal waves by a two-stage interaction).

(14) Interference is the most common first-stage superluminal interaction to accomplish "downshifting" superluminal entities to luminal or subluminal velocities. Superposition of superluminal "phase" waves (such as deBroglie waves, which individually always move faster than the speed of light) interferes the waves to create a subluminal group velocity, which may then interact with an ordinary mass particle in the detection system.

(15) Any otherwise physical vector must exist as an unzipped (segmented) or "shadow" vector in vacuum. "Radiation" of a vector EM wave from the electron gas in an antenna into vacuum results in the "choking off" of the mass of the transversely oscillating electrons in the antenna. Since the spinning electron mass is the "zipper" that makes or comprises the physical vector in the first place, this throttling of the mass flow unzips the E and B vectors, leaving whirling (massless magnetic scalar potential) segments of massless charge flux (massless electrostatic scalar potential). This unzipped whirling pattern of charge flux (scalar massless A/Ø) is what radiates into vacuum and propagates through it. This is a special kind of

scalar wave pattern, not a physical or vector wave.

(16) The spin of a charged particle is the mechanism for integrating or "zipping together" the individual virtual fragments of a shadow vector into a real (observable) vector. For "uncharged particles" such as neutrons, it is the spins of its virtual charged components. that accomplish the integration or zipping.

(17) All fundamental charged particles are constantly accelerated. There is no such thing as an "unaccelerated" particle, except as a gross average over time or length. Further, all of them are spinning.

(18) All changes to and from a physical vector or scalar system must arise in and come from its own internal substructure, which is zipped to its spinning particle of mass.

(19) All fundamental particles are charged internally. That is, they are dynamic assemblages of smaller charged particles. If the average sum of the total internal charge is essentially zero over some finite, small increment of time, the particle is externally uncharged. If the sum is not essentially zero, the particle is also externally charged.

(20) There are no static physical things in existence. In physical reality, something appears "static" only at a particular level. Upon sufficiently fine examination, it is composed of accelerating parts, and thus comprised of "fluctuations."

(21) Since (a) the basic physical (mass) vector consists of a "smeared particle," where particle and smear are inseparable, (b) the conceptual particle also is accelerated, and (c) the "smearing" is for a small increment of time and a small increment of length; then the basic constituency of "physical reality" is inseparable "force x time x length," or action. The basic "quantum" of physical change is thus comprised of action.

(22) Since to "detect" we must "stop" the action, separate or split the quantum into two pieces ("canonical" pieces) , and compare (measure) one piece by throwing away the other, then each physical observable must have a differential operator (the "separating agent") corresponding to it. This accounts for the fundamental postulate of quantum mechanics whereby every observable has a corresponding operator. Further, since what remains is totally relative to what was split out and thrown away, physical change

is totally "relative." This accounts for the fact that observed reality is relative, each part to each other.

(23) As a special case, we may assume that we can evaluate a physical change at a point (without length). If so, when we discard length, the remaining basic vector is momentum. This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of length discarded -- i.e., it holds for the linear case. Conservation of momentum, then, is violated when sufficient nonlinearity in length is present.

(24) As a second special case, we may assume that we can evaluate a physical change in a spatial manner (without time). If so, when we discard time, the remaining basic vector is energy (has the units of energy or work). This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of time discarded -- i.e., it holds for the linear case. Conservation of energy, then, is violated when sufficient nonlinearity in time is present. Since a "virtual change" a priori is defined as a total nonlinearity in the observer's quantal time increment but not outside it, then virtual interactions can and do violate conservation of energy within that time increment, but not out of it -- so long as the time interval itself is considered linear. If the time interval is sufficiently nonlinear, then the virtual change may result in violation of the conservation of energy externally to the time increment. In that case, an "observable change" results .

(25) As a third special case, we may assume that we can evaluate the "instantaneous value" of a physical change at a static point in space. To do so, we must discard both time (to be instantaneous) and length (to be at a spatial point), and the remaining basic vector is force. This approximation holds only so long as the system to which it is applied essentially does not change over the quantal fragment of time or the quantal fragment of length discarded. Conservation of force, then, is violated when sufficient nonlinearity in time or length is present.

(26) A new conservation of energy law is required, one which unites the present conservation of energy law with an altered form of the conservation of charge law. Briefly, the total equivalent of mass,

observable energy, and massless charge (anenergy) is conserved.

(27) All AL and At fragments are produced and destroyed one at a time, in the action fissioning of a single quantum of action (detection process). Each ∇L and ∇t is discretized but not quantized. Since quanta do not superpose, the "external universe" is continually created and destroyed in the detector's mass system, one quantum at a time, at a very high rate. This interpretation gives physical meaning to the creation and annihilation operators of quantum mechanics.

(28) Since the detecting mass system is itself continually created and destroyed one quantum at a time, ultimately all is mind changes, and only mind changes. The observer's life, mind, and being transcend all materialistic interpretations of reality -- as indeed does the very fact of the "existence" of a perceived external universe.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Electrical Physics Presently Has a Mindset --

In examining the foundations of geometry, mechanics, and electromagnetics, it becomes strikingly clear that substantial -- even grave theoretical errors were made early on and perpetuated into the existing theory. These errors are now so firmly entrenched that they form a part of the "mindset" of almost all physicists, engineers, and scientists.

So ingrained are these errors and inconsistencies that the orthodox scientist/theoretician finds it almost impossible to break out of them.

The present mindset is analogous to the Newtonian mindset which so fiercely resisted the new ideas of relativity, shortly after the turn of the century. However, after a few scientists formulated the rules and theory of the "relativity mathematics game," a new generation of students, not yet so firmly engrained in the Newtonian mindset, could grasp the new relativity when their teachers expounded it.

In this short series of papers[1],[2] I will roughly outline where the founders of mechanics and electromagnetics went wrong, and indicate the way to correct the fundamental errors. In addition, I will briefly point out some of the implications, and speak of some direct experimental proof.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

NOTES AND REFERENCES

1. Bearden, T. E., "Comments on the New Tesla Electromagnetics: Part I: Discrepancies in the Present EM Theory;" "Part II: The Secret of Electrical Free Energy," Tesla Book Company, 1580 Magnolia Ave., Millbrae, CA 94030, 1982.

2. Bearden, T. E., "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Co., 1981. Also Ratzlaff , John T., "Reference Articles for Solutions to Tesla's Secrets," Tesla Book Co., 1982.

3. Note we are applying the rule, "A thing is that which it does, and it does that which it is." Actually this is one statement of a fourth fundamental law of logic not incorporated by Aristotle. See Bearden, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," DTIS report, available through the National Technical Information System, Port Royal Road, Springfield, VA 22161.

4. Specifically, the resulting theory becomes a curtailed, special case of the much more fundamental electrodynamics and electromagnetics that actually exist .

5. We point out here that measuring a field of force existing in the electron gas in a probe of the measuring instrument is not at all the same thing as measuring a force in the vacuum, nor does it establish that a force exists in vacuum. Indeed, it is already well known that the FIELD concept itself will not withstand rigorous logical examination. For a discussion rather clearly showing the present difficulty in defining a field, see Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, pp. 283-287. Note particularly on p. 283 that a "field of force" at any point is actually DEFINED only for the case when a unit mass is present at that point. It is then illogically ASSUMED that the force

continues to exist at the point in the ABSENCE of the mass, which of course need not follow at all. On p. 284, note the similar logical paradox connected with the idea of a scalar gravitational potential field. The potential (field) is only defined at a point when mass is present at that point, .and it is specifically defined as the potential energy per unit mass for a particle present at that point. IF THERE IS NO MASS PRESENT, NEITHER A FORCE VECTOR FIELD NOR A SCALAR POTENTIAL FIELD IS DEFINED THERE. ASSIGNMENT OF THESE FIELDS TO THE POINT IN THE ABSENCE OF THE MASS IS AN ASSUMPTION, NOT AT ALL A DEFINITION. SINCE A TRUE DEFINITION IS AN IDENTITY, THEN THE ENTITY IDENTIFIED (DEFINED) TO INCLUDE THE PRESENCE OF MASS IS NOT IDENTICAL TO THE ENTITY RESULTING WHEN THAT MASS IS ABSENTED.

To see just how arbitrary and postulational are present. "definitions" of mass and force, see Lindsay and Margenau, op. cit., pp. 84-101. Also see Richard P. Feynman, Robert B. Leighton, and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol.1,. 1963, Fourth Printing July 1966, p.2-4 for a definition of the electric field in the context of its POTENTIALITY for producing a force. Again, the force only exists when a particle of mass is present. From these examples, one can see the implication that A PHYSICAL FIELD IS SOMETHING SUCH THAT, WHEN A MASS IS INTRODUCED INTO IT, THE MASS EXHIBITS AN EFFECT. For a "force field," this is tantamount to stating that there exists some mechanism connected with a field which, in the presence of a mass causes a force to be exhibited. In that case the force is an EFFECT, not a cause, and there is a more fundamental mechanism that GENERATES FORCE ITSELF. See also field discussions in Feynman, Richard, The Character of Physical Law, M.I.T. Press, Cambridge, MA, March 1967, 2nd printing September 1967, passim.

6. While in Europe prior to 1881, Albert Abraham Michelson performed his first interferometer experiments to determine the velocity of the earth through the ether, obtaining essentially null results. At the Case School of Applied Science in Cleveland,

Ohio, he perfected his interferometer experiment from 1883 to 1887, assisted by a colleague, chemist Edward Williams Morley. By 1887 the results were ready and announced. Michelson himself thought his experiment had proven Stokes' theory of an ether dragged along by the earth in motion, and thus motionless with respect to the earth. This was at odds, however, with certain other experiments indicating a moving ether. The Michelson-Morley experiment was finally reconciled with these other experiments by Fitzgerald's suggestion in 1892 that the physical dimensions of material bodies are altered when they are in motion. In 1907 Michelson was awarded the Nobel prize, the first American to receive it in the sciences.

7. See Lindsay and Margenau, Foundations of Physics, 1963, pp. 324-326; D. C. Miller, "The Ether Drift Experiment and the Determination of the Absolute Motion of the Earth," Reviews of Modern Physics, Vol. 5, p. 203, 1933. Actually the experiments did not yield a conclusively null result, but rather showed large systematic trends. For a typical elimination of the systematic trends, see Handschy, M. A., "Re-examination of the 1887 Michelson-Morley experiment," American Journal of Physics, Vol. 50, No. 11, Nov. 1982, pp. 987-990. See Rho Sigma, Ether-Technology, CSA Printing & Bindery, Lakemont Georgia, 30552, 1977 for several enlightening points on the vacuum ether: See A. K. Lapkovskii, "Relativistic Kinematic Equations and the Theory of Continuous Media," Soviet Physics Journal, Vol. 21., No. 6, June 1978 for an abstract describing Soviet utilization of the concept of a small particle (called by Bearden a quiton, in Quiton/Perceptron Physics, DTIS, 1973) of the medium. See Belyaev, B. N. , "On Random Fluctuations of the Velocity of Light in Vacuum," Azvestiya Vysshikh Uchebnykh Zavedenii, Fizika, No. 11, Nov. 1980, pp. 37-42, translation by Plenum, for discussion of the proven variation of the speed of light in vacuum; the velocity of light in a vacuum on earth is measured to be higher than the velocity of light in the vacuum of deep space. See Graham, G. M. & Lahoz, D. G., "Observation of static electromagnetic angular momentum in vacuo," Nature, Vol. 285, 15 May 1980, pp. 154-155 for the first direct observation of free electromagnetic angular

momentum in vacuum. See Davies, Paul, "Something for nothing," New Scientist, 27 May 1982, pp. 580-582 for a discussion showing that modern theories of the vacuum reveal that even empty space is seething with activity; an ether of sorts emerges from vacuum fluctuations due to quantum mechanics considerations. See Hooper, William J., "All-Electric Motional Electric Field Generator," U. S. Patent No. 3,610,971, October 5, 1971 for a generator which produces a gravitational or inertial field. Einstein suggested that vacuum, complete with electromagnetic and gravitational fields, be called the ether. Dirac certainly did not abandon an ether, for in 1954 he stated "The aetherless basis of physical theory may have reached the end of its capabilities and we see in the aether a new hope for the future." James Clerk Maxwell derived his famous equations based on an ether theory. Sir Arthur Eddington also believed firmly in an ether. Sir Oliver Lodge actually pointed out the dilemma which yields the approach in this paper: writing of the ether in his book, The Ether of Space, Harper & Bros., New York, 1909, he stated: "We have no means of getting hold of the ether mechanically; we cannot grip it or move it in the ordinary way: we can only get it electrically. We are straining the ether when we charge a body with electricity; it tries to recover, it has the power of recoil.. . . "But when electrical theory was being founded, scientists thought of space as something rather fixed, and FILLED WITH a thin material ether. They did not realize that space itself does not exist except after an observation; before the observation, spacetime exists -- indefinite in both length and time. They did not know that electrostatic scalar potential in fact was spacetime, hence also the vacuum and the ether. In assuming that the charge of vacuum is zero and that charge and charged mass are identical, they hid the answer to the dilemma and placed the foundations of electromagnetics on its present unsound basis.

8. The field, of course, is indeed a highly useful concept and this author certainly does not recommend its abandonment. Instead, he recommends that it be placed on a sounder logical basis.

9. Specifically, they came to feel that the "electric

field" which was improperly defined -- was what was waving.

10. In fact Einstein once proposed that the vacuum, complete with its electromagnetic and gravitational fields, should be called the ether. His proposal was not adopted. (See Born, Max, Einstein's Theory of Relativity, Revised Edition, Dover Publications, New York, 1965, p. 224.)

11. Particularly from the work of Schrodinger, Born, Dirac and others.

12. For example, see Lindsay and Marge. nau, op. cit., pp. 287-288. A physical vector is thought to be defined by its magnitude, its direction in space, and its transformation characteristics. Actually that is a geometrical vector, not a physical vector. It does not tell us WHAT A VECTOR CONSISTS OF, but only tells us some of its important characteristics. Remember that a true definition must be an identity.

13. Here a reading of Lindsay and Margenau, op. cit., pp. 79-81 may prove enlightening. Also note that velocity, or L/T considered "at an instant" (stopped), represents an idea of "motionless motion" and is an application of the fourth law of logic. For a discussion of the fourth law of logic and its usage, see Bearden, Thomas E., "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," Specula, Journal of the A.A.M.S., P.O. Box 1182, Huntsville, AL 35807, combined Vol. 3, No. 4/Vol. 4, No. 3, Oct 1980 - Mar 1981, pp. 50-57 (also available from Defense Technical Information Service).

14. Note this is an identity of opposites, which explicitly violates the three Aristotlean laws of logic. See Bearden, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," loc. cit., 1981.

15. Again note the fourth law of logic: zero motion (the absence of motion) being recognized as a special case of the presence of motion. Also, physical reality consists of internested levels, and any physical object has an internal substructure of nested levels of finer structure, extending down into the virtual (nonobserved) state. For a vector to model (apply to) a

physical object, it itself must be modeled in such fashion as to reflect this kind of substructure. Thus the use of geometrical vectors as models of physical objects in motion is presently flawed in a fundamental fashion.

16. The reader is most strongly urged to read Morris Kline, Mathematics: The Loss of Certainty, Oxford University Press, New York, 1980 as a prelude to understanding what mathematics is and is not, and what it does and does not.

17. Refer to Lindsay & Margenau, op. cit., pp. 79-81 to see how the ideas of motion and vector are inextricably entangled with the idea of a particle.

18. Call it uncertainty or call it constituency; a quantum change is composed of two canonical entities inextricably welded together into a single entity.

19. Time is an unavoidable, nonexclusive constituency of the welded quantum.

20. Simply from the definition of force as CONSISTING OF a time- and length-smeared mass motion change.

21. The force is an effect, not a cause. It IS the smeared charged particle. It is CAUSED by a more fundamental mechanism. It is the result of the combination of (1) a nonzero del phi, and (2) the presence of a spinning charged particle. IN A DEL PHI, THE SPINNING CHARGED PARTICLE ACCELERATES ITSELF! This is the fundamental secret of free energy that was suppressed, to bury the fundamental work of Nikola Tesla, shortly after Tesla was forced to abort his Wardenclyffe attempt to provide the world with free energy.

22. To quote: "The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history. " Nikola Tesla, "The True Wireless, " Electrical Experimenter, May 1919, p. 87.

23. De Beauregard, O. Costa, "Running backwards the Mermin device: Causality in EPR correlations," American Journal of Physics, Vol. 51, No. 6, June 1983, p. 515.

24. Note the Soviet scientist Kozyrev's experiments with time waves. See Kozyrev, N. A., "Possibility of Experimental Study of the Properties of Time," September 1967, pp. 1-49, in JPRS 45238, May 1968. Kozyrev reports real physical effects from the oscillation of time. Also, note that scalar potential energy of appreciable size with respect to a particle's rest energy can force the situation to be relativistic, even though the velocity of the particle with respect to the velocity of light is small. That is, electrostatic scalar potential alone can cause variation in the rate of flow of time and hence vary physical characteristics. See Bloch & Crater, "Lorentz-invariant potentials and the non-relativistic limit," American Journal of Physics, Vol. 49, No.1, 1981, pp. 67-75. By inference, oscillating the electrostatic scalar potential can produce time waves and lead to direct physical effects.

25. It is already shown in the literature that the electrostatic scalar potential (ESP) can affect spacetime (ST) in the same manner as velocity. Cf Bloch & Crater, op. cit., 1981. Now note that, to any quantal or macroscopic observer, the existence of the 4-space volume of ST implicit in $(\Delta t)(\Delta v)$, where v is volume, cannot be separated from the existence of the subquantal entities that exist therein. We therefore DEFINE the magnitude of the ESP as the summation of the absolute values of all the internal virtual vectors in the $(\Delta t)(\Delta v)$ quantum of ST, divided by the absolute value (magnitude) of $(\Delta t)(\Delta v)$. We take the view that no such thing as "unstressed" ST physically exists, and that "spacetime" and "stressed spacetime" are identical. Hence ESP and ST are one and the same thing. Note that this implies that the virtual density of ST is variable, and is nothing but the magnitude of the ESP. In EM theory, the assumption that the ESP of vacuum ($Ø_O$) is equal to zero is in serious error. In fact, $Ø_O$ IS "spacetime of the laboratory observer," in the new view.

26. For example, the definition of the electrostatic

potential (ESP) is usually taken as "the work which must be done against electric forces to bring a unit charge from a reference point to the point in question; the reference point is located at an infinite distance, or, for practical purposes, at the surface of the earth or some other large conductor." (McGraw-Hill Dictionary of. Scientific and Technical Terms, ed. Daniel N. Lapedes, second edition, 1978, p. 518.) Note that this is NOT a definition at all, for it is not an identity. Instead, it is the statement that, if an ESP exists at a point and a unit charged mass (assumed to be at a point) is brought in from infinity toward the ESP location point, the amount of work it is necessary to expend upon the mass of the particle is equal to the magnitude of the ESP. The ESP exists whether or not any work at all is expended, and whether or not a charged unit mass is brought in. To adequately define ESP, we must define its identity, or what it consists of, in the absence of mass, since we have conceived the ESP to exist at a vacuum point. Further, the definition usually taken is completely a 3-space definition. Instead, in our new view the ESP is to be taken at a point in n-space, where n is equal to or greater than 4.

27. See Bearden, Quiton/Perceptron Physics, 1973, available through the DTIS. See also Bearden, The Excalibur Briefing, Strawberry Hill Press, San Francisco, CA, 1980. Ultimately all physical phenomena are mindchanges in the minds of all the observers.

28. And then assumes this summation value is zero.

29. See note 25 above.

30. Bearden, The Excalibur Briefing, Strawberry Hill Press, San
Francisco, CA, 1980.

31. Cf Rauscher, E. A., "Electromagnetic and Non-Linear Phenomena in Complex Minkowski Spaces," Tecnic Research Laboratories, 64 Santa Margarita, San Leandro, CA 94579. Presented at the 1983 March Meeting of The American Physical Society in Los Angeles, CA 21-25 March, 1983. This is a truly remarkable paper of great significance. Rauscher, a world-class physicist, has presented a new theoretical

model for some rather extraordinary possible extensions of present electromagnetics.

32. Cf Muses, Charles, Introduction to Jerome Rothstein's Communication, Organization, and Science, The Falcon's Wing Press, Indian Hills, Colorado, 1958. The entire foreword by Muses is a remarkable document, which analyzes the structure of time itself. See also his profound summary paper, "Hypernumbers II" in the January 1978 issue of Applied Mathematics and Computation, published by Elsevier.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

ADDITIONAL NOTES AND REFERENCES

Although quotes and direct utilization of material from these references were not incorporated in this paper, the following references were also consulted. In addition,' several notes are added for further clarity.

33. Rupert Sheldrake, A New Science of Life: The Hypothesis of Formative Causation, J. P. Tarcher, Inc., Los Angeles, CA, 1981. Ø, the electrostatic scalar potential field, in my opinion is actually the morphogenetic field that Sheldrake proposes.

34. Briefly, by a "particle" we mean an entity so constructed that, if any part of it changes all of it changes. From the viewpoint of the particle, this implies that to change is to detect, and to detect is to change. Also, internal and external become synonymous, in the "detected" sense. The idea of a "fundamental particle" in physics actually invokes the fourth law of logic implicitly.

35. Only if a thing dimensionally contains time, can it "occupy time." This point is so obvious that one wonders how so many of the scientists and mathematicians seem to have missed it. By this criterion, e.g., mass does not exist in time, a priori. To "observe" or detect, in fact, means to stop time, thus collapsing the wave function. However, it reduces the observable or detectable to a spatial quantity, not a spatiotemporal quantity. In other words, the ordinary scientific method destroys a part of reality in each detection or measurement, yielding only a partial truth. not fundamental truth.

36. Note that electrostatic scalar potential is actually infinite-dimensional and hyperspatial. The coverage of

this paper is still only a special case. By Tesla technology, it is possible to do direct engineering in hyperspace -- beyond our present space and time, with all that that statement implies.

37. Bob Sloan, "Nikola Tesla: The Greatest Inventor of all Time?". IEEE Antennas and Propagation Society Newsletter, June 1983, pp. 9-11. A very succinct summary of the importance Tesla played in ushering in the modern age.

38. Gerald E. Brown and Mannque Rho, "The structure of the nucleon," Physics Today, Vol. 36, No.2, February 1983, pp. 24-32. Recommended as a summary of the new thinking as to the structure of the nucleon: a bag containing three quarks, surrounded by a cloud of mesons which squeeze the bag.

39. John J. O'Neill. Prodigal Genius: The Life of Nikola Tesla. Angriff Press, P.O. Box 2726, Hollywood, CA 90028, new printing 1981.

40. Margaret Cheney, Tesla: Man Out of Time, Prentice-Hall, Englewood Cliffs, NJ, 1981.

41. John T. Ratzlaff and Leland I. Anderson. Dr. Nikola Tesla Bibliography. Palo Alto. CA, 1979. Indispensable.

42. Dr. Nikola Tesla: Selected Patent Wrappers, compiled by John T. Ratzlaff, multiple volumes. 1980. Available from The Tesla Book Company, 1580 Magnolia, Millbrae, CA 94030. Tesla's correspondence with the U.S. Patent Office, when patiently trying to obtain patents. He spent a great deal of time trying to convince the Patent Office that his inventions would indeed work. Some of them required 12 years to obtain, and then were "watered down" in the process.

43. Thomas Commerford Martin, The Inventions, Researches and Writings of Nikola Tesla, Originally published in 1894 by The Electrical Engineer, New York; republished in 1977 by Omni Publications, Hawthorne, CA 90250.

44. Ernest Nagel and James R. Newman, Godel's

Proof, New York University Press, 1958.

45. Yakov P. Terletskii, Paradoxes.in the Theory of Relativity, With a Foreword by Banesh Hoffman, translated from the Russian, Plenum Press, New York, 1968. Of particular interest is the discussion on particles with imaginary masses, moving faster than the speed of light, contained in pp. 104-107. Such particles can in principle be experimentally detected. In fact, it would appear that the well known exchange of virtual particles between two other particles, such that each turns into the other, is such a case. (Note that protons and neutrons in the nuclei of atoms do precisely this.)

46. Robert M. Besancon, Ed., The Encyclopedia of Physics, Second Edition, Van Nostran Reinhold, New York, 1974. Particularly see the discussion on the electron, pp. 272-274. Note this discussion predates Stanford University's experiments yielding fractional charge, though it does point out that several physicists had also reported measuring fractional charges on the electron. See also the discussions of ionization, Michelson-Morley experiment, the photon, and propagation of electromagnetic waves.

47. Robert Eisberg and Robert Resnick, Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles, John Wiley & Sons, New York, 1974.

48. R. K. Bullough and P. J. Caudrey, eds., Solitons, Springer-Verlag, New York, 1980.

49. James Dale Barry, Ball Lightning and Bead Lightning, Plenum Press, New York, 1980. Note particularly p.196, for a short discussion on flashless discharges. An extensive bibliography is also included.

50. Harley D. Rutledge, Project Identification, The First Scientific Field Study of UFO Phenomena, Prentice-Hall Inc., Englewood Cliffs, NJ, 1981.

51. John J. Reitz, Frederick J. Milford, and Robert W. Christy, Foundations of Electromagnetic Theory, Third Edition, Addison-Wesley, Reading, MA, 1979.

52. The entire series of handbooks by William Corliss, dealing with anomalies and unusual natural phenomena

of all kinds. Corliss is a national treasure, and his handbooks are absolutely indispensable.
See particularly his Handbook of Unusual Natural Phenomena, The Sourcebook Project, Box 107, Glen Arm, MD 21057, 1977 and his Lightning, Auroras, Nocturnal Lights, and Related Luminous Phenomena, 1982.

53. Bernard d'Espagnat, Conceptual Foundations of Quantum Mechanics, W. A. Benjamin, Menlo Park, CA, 1971.

54. D.W.G. Ballentyne and D.R. Lovett, A Dictionary of Named Effects and Laws in Chemistry, Physics and Mathematics, Fourth Edition, Chapman and Hall, New York, 1980. Check this neat little book to discover some very odd effects in materials.

55. David Bohm, The Special Theory of Relativity, W. A. Benjamin, New York, 1965.

56. Albert Einstein, Relativity: The Special and the General Theory, Crown Publishers, New York, 1961. See particularly the discussion of relativity and the problem of space, in Appendix V.

57. Edwin F. Taylor and John Archibald Wheeler, Spacetime Physics, W. H. Freeman and Co., San Francisco, 1966. Note particularly the discussion on observers and frames in the first two dozen pages. On p.19, note that the notion of a frame requires an infinite observer distributed through each and every "point" that was clock-synchronized in a frame. Since all observers are localized, a better idea is to realize that all the "external" modeling just represents the relative changes inside the physical detection system of the observer himself/herself. All detection/observation is totally inside the physical observer.

58. Paul Edwards, Ed. in Chief, The Encyclopedia of Philosophy, Vols. 1-8, Macmillan Publishing Co., New York, 1967.

59. Michael A. Persinger and Gyslaine F. Lafreniere, Space-Time Transients and Unusual Events, Nelson-Hall, Chicago, IL, 1977.

60. John David Jackson, Classical Electrodynamics, Second Edition, John Wiley & Sons, New York, 1975.

61. James Clerk Maxwell, A Treatise On Electricity & Magnetism, Vols. 1 & 2, Third Edition, Dover Publications, New York, 1954. An unabridged, slightly altered, republication of the third edition, originally published by the Clarendon Press in 1891.

62. Jack S. Greenberg and Walter Greiner, "Search for the sparking of the vacuum," Physics Today, August 1982, pp. 24-32. A beautiful summary article on the present concept of the vacuum, from the standpoint of quantum mechanics and quantum field theory. I specifically urge anyone interested in tapping the vacuum energy to read this article.

63. Max Jammer, Concepts of Mass, Harvard University Press, Cambridge, MA, 1961. Recipient of the monograph prize of the American Academy of Arts and Sciences for the year 1960 in the field of physical and biological sciences. This book, strange as it may seem, was the first monograph to subject the notion of mass to an integrated and coherent historical investigation, something which had never before been done. The reading of this book is an absolute must for anyone seriously concerned about whether or not science speaks fundamental truth, or to what degree it does so. Most scientists assume foundations concepts such as "field," "force," "mass," "time," etc. are well-defined and well-understood in physics, since they are ubiquitous and so familiar. In fact, none of the absolute fundamentals in the foundations of physics is unambiguously understood!

The last paragraph by Jammer is illuminating: "The modern physicist may rightfully be proud of his spectacular achievements in science and technology. However, he should always be aware that the foundations of his imposing edifice, the basic notions of his discipline, such as the concept of mass, are entangled with serious uncertainties and perplexing difficulties that have as yet not been resolved."

64. G. Burniston Brown, "Gravitational and inertial mass," American Journal of Physics, Vol. 28, p. 475, 1960: "...one of the most astonishing features of the

history of physics is the confusion which surrounds the definition of the key term in dynamics, mass."

65. Whittaker, Sir Edmund, A History of the Theories of Aether and Electricity, Vol. 1: The Classical Theories, and Vol. 2: The Modern Theories 1900-1926, Harper Torchbooks, Harper & Brothers, New York, 1960.

66. Weyl, Hermann, Space -- Time -- Matter, Fourth Edition, translated from the German by Henry L. Brose, Dover Publications, New York, 1922.

67. Jammer, Max, Concepts of Space: The History of Theories of space in Physics, Second Edition, Foreword by Albert Einstein, Harvard Univ. Press, Cambridge, MA, 1969.

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- It Started With Geometry and Grew --

At the very beginning of what we call the "scientific period," mathematics was both king and queen, and Euclidean geometry was its handmaiden. So we ask, "What precisely is geometry?" Here we are not interested in a "textbook" answer, but in an answer indicating what geometry really does.[3] In other words, with what does geometry concern itself, and what is the fundamental nature of those things with which it concerns itself?

Briefly, geometry -- at its foundation -- is totally spatial. It is fitted to, and expressed in terms of, the TOTAL ABSENCE OF MASS. Thus the geometer deals in abstract, massless entities called "points," "lines," "planes" etc. When the geometer speaks of "motion," he speaks of a time-smeared, length-smeared point. Geometry at heart is massless, and a "geometer's vector" is a highly specific type of "system." In fact, it represents the "time-smearing" and "length-smearing" of a point. A priori, the fundamental concept of the geometrical vector has taken a "spatial" entity and introduced a hidden involvement with "time."

Modern mathematics and physics have followed an intertwined development for several hundred years. And both sprang as offshoots of the original work of the geometers. Let us briefly sketch the overall path of interest taken by these two developing disciplines.

With the advent of Descartes's fundamental work, algebra was combined with geometry to yield analytic geometry, a new and powerful mathematical tool. With the invention of calculus by Leibniz and Newton, both mathematics and physics received a giant impetus. Differential geometry and vector mathematics arose in full splendor and, in physics, mechanics leaped to the forefront with Newton's profound work.

But the mechanics made a most fundamental error when they simply applied the geometer's vector to a mass, to produce -- so they thought -- a mass vector. That which rigorously applies only to the absence of mass cannot be so lightly applied to the presence of mass without the risk of serious limitations in the resulting theory. The precise difference between a geometer's massless vector and a mechanic's mass-vector is one of the issues to be developed in this thesis.

As rapid development continued in mechanics and mathematics, certain physicists were involved in intense experimental work on charged matter, becoming the first electricians. Both the preceding mathematical ideas and constructs as well as the preceding (partially erroneous) mechanics constructs and ideas were applied by the electricians, struggling with their pith balls, cat fur, and glass rods to understand, quantify, and model electrical forces and the phenomena of charged matter. In other words, the electricians strove to formulate the physics and dynamics of charged matter and its interactions by simply "adding to" the work of the geometers and mechanics. Here again, a fundamental logical error was made. That (geometry) which a priori applies only to the absence of mass, and that (mechanics) which a priori applies only to the absence of charge, cannot be lightly applied to the presence of charged mass (both mass and charge)[4] without risking the incorporation of grave limitations in the resulting theory.

After the profound work of Maxwell, the idea of FIELDS OF FORCE became more prominent, until the field concept ruled the day[5]. The electricians continued, pushing the idea of fields into space and vacuum itself, along the way inventing the idea of "charge effects" existing even in the massless vacuum, with concomitant fields. Meanwhile, they had thoroughly confused chargeless point-smeared, chargeless mass-smeared, length-smeared and time-smeared vectors.

After a set of fundamental experiments designed to detect motion of the material ether yielded essentially null results[6], Michelson and Morley were regarded as having completely disposed of the ether -- even though the experiments only disposed of material ethers, and not Lorentz-invariant non-material ethers[7]. Maxwell's equations and the field concept were elevated to profound importance.[8] Then, after Einstein's fundamental relativity work shortly after the turn of the century, the ether concept faded away and the field concept reigned supreme. Indeed, in their enthusiasm the interpreters of relativity went so far as to affirm that one can have a wave without any medium; that is. that something can be moving (waving) without anything there to move![9] And with great glee they pronounced the final end to the idea of "ether" as a medium, even though Einstein himself never did any such thing.[10] With the advent of Einstein's General Theory of Relativity, even matter came to be regarded as just a special "kink" or curvature in spacetime or "vacuum nothing."

Quantum mechanics arose and even certainty and determination fell. Chaos, probability, and randomness now assumed the ruling position. Probability waves (and probability fields) arose,[11] as did quantum fields of various kinds. The intermingling of these concepts with the concepts of electrodynamics pushed the idea of the field even farther into esoteric realms.

The point is, each of these developing disciplines incorporated and built on the foregoing disciplines. From the beginning of geometry, there was no rigorous definition of a vector, and there is none today.[12] From the beginning of mechanics, in their foundations the theorists made grave logical errors by incorporating the geometer's vector; errors so great that today mechanics and electromagnetics are severely flawed, as is everything that came after them and built upon their illogical foundations.

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Points and Motion --

It is my purpose in this paper to expose in a very simple fashion the most basic errors that were made. One basic error involves the idea of motion itself.[13]

In formulating concepts of motion, the geometers used a "point in motion" to determine or specify, for example, velocity. Now a "point" is a static concept a priori. To determine (or even to think and perceive) motion, one must determine that it occupies two different points (positions or locations) at two different times, yet consider both points at the same time. Indeed, that is precisely what the arrow means that is used to represent a vector. A "point in motion" therefore represents a contradiction of opposites. That is, it represents the idea that "that which is motionless has motion.".[14] Even with this, there is a difference in a spatial point and a spatiotemporal (spacetime) point. To exist at all, a spatial point must be moving in time; in other words, it is a spatiotemporal line, even if it is a static spatial point.

Vector analysis was constructed in the abstract -- again, a massless point in motion possessed or constituted a velocity vector, etc. In massless (and timeless) space, FIELDS were defined: "scalar" fields constituted the assignment of a simple motionless number (magnitude) to each spatial point, while "vector" fields constituted the assignment of a "simple vector" (magnitude and velocity) to each spatial point. But the MATHEMATICAL vector system consisted of massless (point) motional relationships, recognizing zero motion as a special case of motion.[15]

Of course mathematics development was also always intertwined with practical problems. With the sustained application of mathematics to gross physical material problems, mechanics slowly arose.

These developments required decades and even centuries to occur completely. All along the way, innovations and changes -- and additions to the mathematical formulism were being derived and taught to students as the "natural" system of reality. A permanent mindset was being forged.

Indeed, mathematics was regarded as THE single human expression of fundamental truth. Not until Godel's work in the twentieth century did it become evident that MATHEMATICS IS SIMPLY A GAME PLAYED ACCORDING TO ASSIGNED RULES, AND THERE IS NO ULTIMATE TRUTH IN MATHEMATICS ALONE.[16] It is a most useful game, of course, since it is the game fitted to perception. Thus it applies, essentially, to whatever can be perceived. But to be applied to physical systems, it must be changed, altered, updated, and fitted as the perceiving/detecting instruments become ever more subtle.

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Four Types of Vectors Actually Emerged --

As the physical sciences slowly developed and incorporated abstract geometry and mathematics, in actuality four major types of vectors and two major states of observation evolved, although this fact did not become apparent to the scientists. Specifically, the mathematicians and scientists failed to recognize the differences in the four types of vectors, hopelessly intermingling them and confusing them as a single class of vector. Further, they did not appreciate that a fundamental vector conceptually is a UNITARY SYSTEM, and the system represented by one of these four types of vectors utilizes and is comprised of different components, "welded together with no seam in the middle."

Conceptually (and from a systems viewpoint), the four types of vectors are (1) the chargeless, massless spatial system vector (geometer's vector), (2) the uncharged mass system vector (mechanic's vector), (3) the charged mass system vector (electrician's vector), and (4) the charged space system vector (advanced electrician's vector). These four fundamentally different vectors are shown in Figures 1, 2, 4, and 9 in a simplified manner.

The major problem was that, beginning with the geometer's vector, these four major types of vectors were not treated as systems. Instead, their "vector" aspects were hopelessly confused and intermeshed, and no distinctions were made between them. And in the foundations of the mathematical constructs, time-smearing was not recognized at all.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- Quantum Mechanics Compounds the Problem --

In addition, the two presently recognized observation states -- observable and nonobservable (virtual) -- were of course unknown to the early geometers and electricians, and these ideas were not incorporated directly into the theoretical foundation.

From particle physics and quantum mechanics, we now understand that physical reality is structured of an observable state, underlaid with an infinite number of ever finer, successive levels of virtual (unobservable) states. At least reality is most accurately modeled in that fashion, according to particle physics today.

It is also well known, for example, that at the most fundamental level, one cannot actually separate nonmotion from motion (which implies, for example, that one cannot separate mass and velocity). In other words, a "mass in motion" idea is actually incorrect, at the most basic level. What actually exists is a sort of "smeared mass". That is, "mass-motion" is fundamentally what exists, not mass IN motion .

Actually, all that the Heisenberg Uncertainty Principle implies is this fact: If one examines the concept of "static (non-smeared) thing in non-static (smeared) motion", in ever finer detail, one reaches a degree of fineness where the "smearing" is paramount and one cannot have an un-smeared or "separate static thing" to be in motion. Instead, one only has the smeared, 4-dimensional spacetime entity, without 3-dimensional spatial separations.

This means, for example, that at the most basic level, it is actually incorrect to represent a momentum with a little static particle of mass connected to a spatial velocity vector. It is incorrect to think of the system as comprised of TWO SEPARATE ENTITIES, (1) a mass, and (2) a massless spatial system velocity vector (a geometer's vector).

We mention in passing that, presently, we understand that every particle is continually accelerating. First, the particle has spin, which involves rotation, which means that every "part" of the periphery of the particle is accelerated toward the center axis of spin. Second, every particle is continuously "fluctuating," and these fluctuations are accelerations. Further, we must consider any change such as an acceleration -- as existing in a small time increment, and occurring in a small length increment. Thus mass particles actually exist as (mass x acceleration x time x length). This of course has the dimensions of ACTION or angular momentum. The "real" world of physical matter, then, is composed of building blocks of action, called "quanta." Any other physical "quantity" must be obtained by fissioning (differentiating) the action quantum. For this reason, quantum mechanics presently must postulate that to every observable there corresponds an operator.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help Support the Work

-- To Summarize Briefly --

Let us now summarize these concepts and further examine their impact.

In physics, there must actually exist four major KINDS of vectors, rather than just one as prescribed by present interpretation. These vectors are "built" starting with four major kinds of particles.[17] These are: (1) the spatial point, (2) an uncharged mass particle, (3) a charged mass particle, and (4) a charged spatial point. Further, each of the four vectors at its most fundamental level (that is, at the quantum level)[18] is actually an inseparable, unitary SYSTEM welded into a single undivided entity containing time and existing nonexclusively in time.[19] When we look at or represent the so-called "parts" of the system, we are looking at them before they are welded together into the physical vector. That is, whenever we speak of "parts, " we imply that a "cutting" or "differentiating" action has been implied to separate the system into such "parts."

From such considerations, four different kinds of system vectors result.

Next Page

Help Support the Work

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

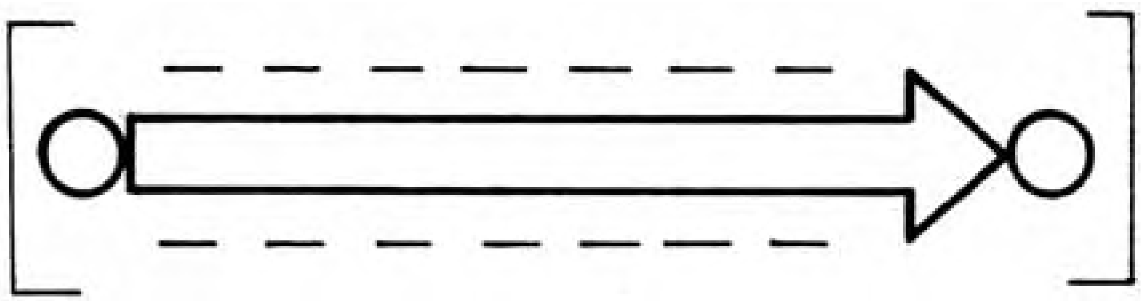

(GEOMETER'S VECTOR)

Figure 1. Uncharged spatial (massless) system vector.

UNCHARGED SPATIAL (MASSLESS) SYSTEM VECTOR
-- (Figure 1) --

This is the geometer's abstract vector, consisting of a "point in motion." (Actually, it is a "smeared point," for example.) However, so ingrained is the concept of a "point in motion" velocity vector that we now consider it to be "natural" because of its total familiarity. But simply ask, "WHAT is in motion?"[17] and you immediately see the difficulty. To have a WHAT, one must "stop the action" (detect or measure), separating "static" from "non-static." Acceleration and other vectors, etc. have also been derived by the geometer and utilized in similar fashion. All are massless.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

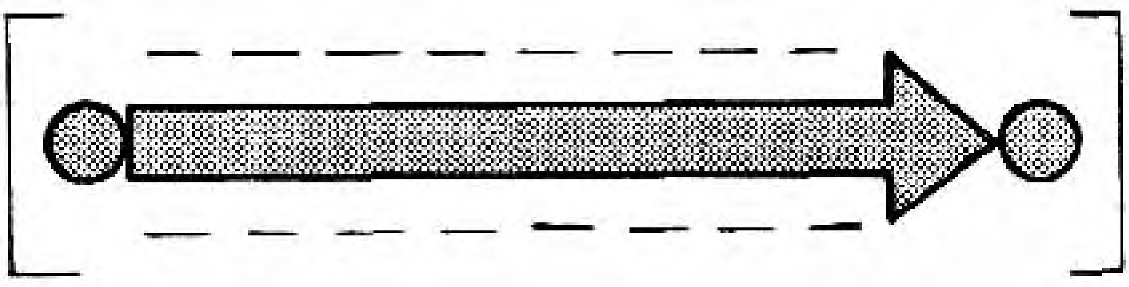

(MECHANIC'S VECTOR)

Figure 2. Chargeless mass system vector.

CHARGELESS MASS SYSTEM VECTOR
-- (Figure 2) --

This type of vector actually is the essential vector of mechanics, involving mass-motion (momentum), force (mass-motion change), etc. The fundamental difference between this type of vector and the geometer's vector is the presence of smeared mass existing in time (i.e., mass-time), welded together with a geometer's vector, but with no seam in the middle, into a new kind of "vector" AS A SYSTEM. The vector is the SYSTEM EXISTING IN TIME .

To illustrate:

"Momentum" is more properly referred to as "mass-motion" rather than "mass in motion." That is, at the fundamental level, the mass is NOT "separate" from its motion. It is NOT separated from the time in which the smearing occurs. The idea of momentum, however, is really to express the time-density of the mass-smeared-through-length. It is, in other words, the time rate of length-smearing of mass. Now in our minds we have conceived that

$$\bar{p} = m\odot\bar{v} \qquad (1)$$

where, by the symbol $\odot$ we mean "coupled to"; but we actually have

$$\bar{p} \equiv [\overline{m\bar{v}}] \qquad (2)$$

where we are not allowed to separate (even in thought) m and $\bar{V}$. Quantum mechanics agrees with this essentially, because $\bar{P}$ is a canonical variable linked to length, in any observable physical change. (The REASON p is canonically linked to ΔL is because $\bar{P}$ is the time-rate of length-smearing of m. If there is no length, there is no length-smearing of m to have a time rate OF in the first place.)

Note that, not only is

$$\bar{p} = [\overline{m\bar{v}}] \qquad (3)$$

but also

$$\bar{p} \equiv [\overline{m\bar{v}}] \qquad (4)$$

which is a much stronger and quite different statement. That is, $\bar{P}$ is IDENTICALLY $(\overline{mv})$, not just calculably EQUAL TO (m)($\bar{V}$). This means that $\bar{P}$ is a SYSTEM that is COMPRISED of mass-motion.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

-- What Force Is --

We now note that force, for example, is -- and may be defined as -- the time-rate of change of momentum, or

$$\bar{F} \equiv d/dt[\bar{p}] \qquad (5)$$

and this identity states that a force -- any force is COMPRISED OF time-changing "mass-motion." As such, the force vector is a mass-system vector, not just a massless spatial vector. Fundamentally, this mass-system vector is a totally different creature from a massless spatial vector. Our present manner of considering force as a geometer's vector "separately applied to" a mass particle is completely erroneous at the quantum level. Instead, fundamentally force is always a mechanic's vector. Force is an EFFECT, not a CAUSE.

And here mechanics made a most fundamental error, in not recognizing the difference between its kind of vector and that of the geometer.

Next Page

TOWARD A NEW ELECTROMAGNETICS
PART III: CLARIFYING THE VECTOR CONCEPT

--Force and Hertzian Waves Cannot Exist in Vacuum --

Note that one cannot have an observable "force vector" existing' in vacuum a priori.

For example, we have the definition of force as

$$\bar{F} \equiv \frac{d\bar{p}}{dt} \qquad (6)$$

or

$$\bar{F} \equiv \overline{[m\bar{a} + \dot{m}\bar{v}]} \qquad (7)$$

and we see that, rigorously, a force vector CONSISTS OF (not, "is equal to") a time-changing mass-motion vector system. IF THERE IS NO OBSERVABLE ACCELERATING MASS PRESENT, THEN THERE CAN BE NO OBSERVABLE FORCE PRESENT. The mass can accelerate in time (increase or decrease of mass) or space (increase or decrease of velocity) or both.

Observable force CANNOT exist in vacuum (in the absence of mass), a priori.[20]

However, assume for a moment that one could have a massless force vector, as assumed in present electrical theory. Let this force vector appear at a point in the vacuum. Since the vacuum has zero observable mass, it would have zero inertial resistance to this hypothetical observable force -- hence the observable force would instantly produce an "infinite" acceleration of its point of application, vanishing with it into the infinite distance. Therefore our fictitious force would disappear the instant it appeared! In any case, it could not be retained at a point in the vacuum for any finite length of time, however small.

The direct implications are that (1) something other than an observable electrical force field exists in the vacuum, and (2) there must exist a more fundamental mechanism by which this "something else" generates or CREATES a force on/of a moving electrically charged mass. (Note again that, at the most basic level, any particle of mass is always quivering, in motion, and accelerating. from quantal considerations alone.)

Thus immediately we have discovered something unique about so-called "force-fields" in vacuum: for example, about gravitational field, electrical field, and magnetic field (and the strong force and the weak force as well). These fields do not exist at all as ordinary force vectors -- and real force fields -- in vacuum! $\bar{E}$ and $\bar{B}$ fields, e.g., are defined in terms of force per unit electrical charged mass and magnetically charged mass, respectively. In the absence of mass, they cannot exist.

And this in turn means that transverse $\bar{E}-\bar{H}$ field waves (Hertzian waves) cannot exist in a vacuum. Indeed, they appear on, and ARE CONSTITUTED of, the charged-mass-motion that changes, and they appear where such change occurs, as a result of an introduced mass.[21] But in the absence of the spinning charged particle of mass, they do not exist as force fields at all.

Hertzian waves exist in a transmitting whip antenna, for example, in the oscillating electron gas along its length. Something else entirely different exists in vacuum between the transmitting antenna and the receiving antenna. Then in the receiving antenna, Hertzian waves again exist in the oscillating electron gas along its length. (See Figure 3.)

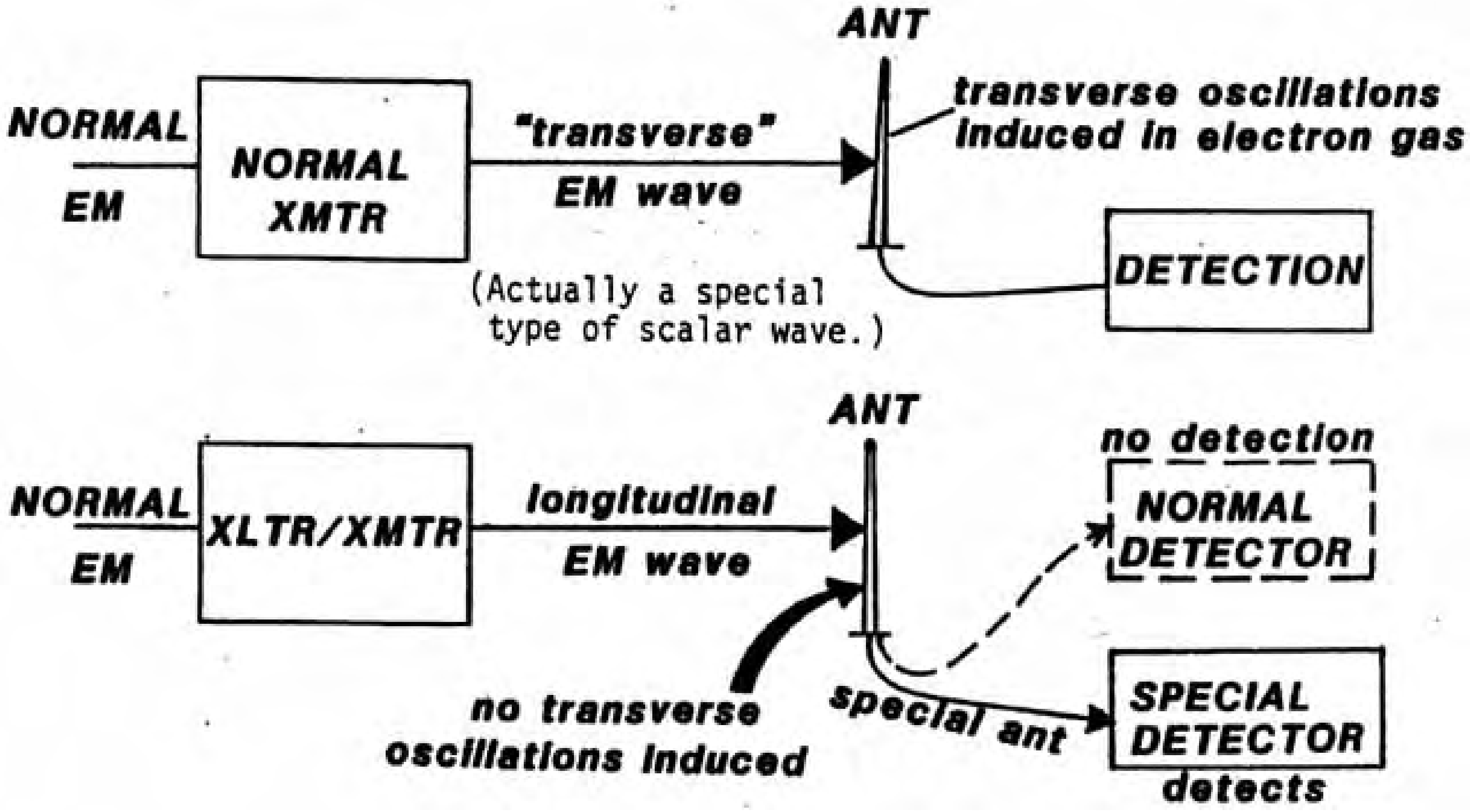
ANT
NORMAL
EM
NORMAL XMTR
"transverse"
EM wave
transverse oscillations induced in electron gas
DETECTION
(Actually a special type of scalar wave.)
ANT
NORMAL
EM
XLTR/XMTR
longitudinal
EM wave
no detection
NORMAL DETECTOR
no transverse oscillations induced
special ant
SPECIAL DETECTOR
detects

Figure 3. Detection of "transverse" and longitudinal waves

This is interesting, for Nikola Tesla stated several times that HERTZIAN WAVES CANNOT BE PRODUCED IN A VACUUM, NOR CAN THEY TRAVEL IN A VACUUM.[22]

Tesla was correct, as we are beginning to see.

We shall later return to show in what form so-called "force-fields" actually exist in vacuum.

For now, I point out that I am stating a fundamental change to all of physics, including both mechanics and electromagnetics.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support this work

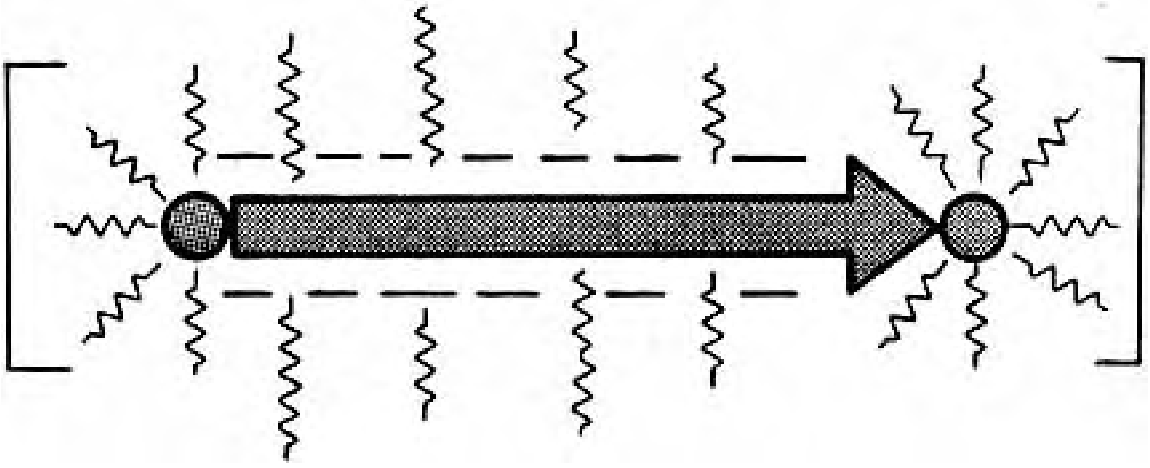

(ELECTRICIAN'S VECTOR)

Figure 4. Charged mass system vector.

-- CHARGED-MASS-SYSTEM VECTOR --
-- (Figure 4) --

The third type of vector we meet is the vector mass system where the mass is charged. First, we point out a serious error in present electromagnetic (EM) theory. That is, in present theory it is implicitly assumed that

$$\mathbf{q} \equiv \mathbf{q_m} \qquad (8)$$

In other words, "charge" and "charged mass" are erroneously assumed to be identically the same thing.

In the days when electricians were playing with pith balls and striving to uncover the secrets of electricity, they knew nothing at all about the virtual state, and consequently nothing about a "virtual particle flux" on a particle of mass causing (and comprising) the "charge" of that mass.

Today, of course, we know from particle physics and quantum mechanics that the "charge" on an observable particle of mass IS due to a flux of virtual (nonobservable) particles on and off the mass of the observable particle (see figure 5). A charged mass is thus presently known to be a SYSTEM: a massless charge flux, coupled to a bare particle (chargeless mass) constitutes a "charged particle."

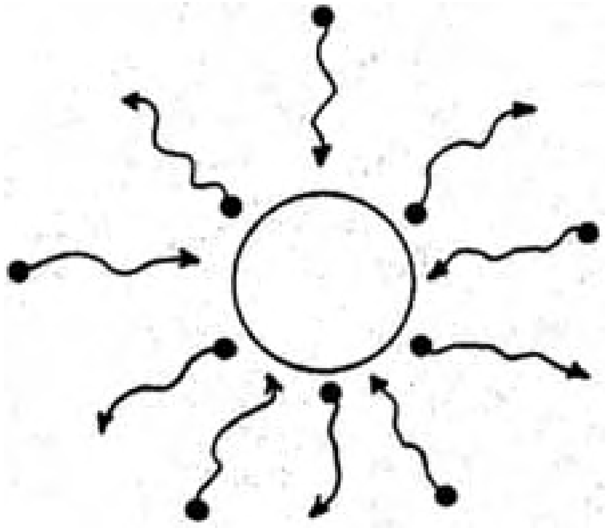

Figure 5. The "charge" on an electron mass consists of a flux of virtual particles on and off the mass.

Thus, actually the "charge" is the virtual (unobservable, or SPATIO-TEMPORAL) flux to and from the observable SPATIAL particle of mass. So, rigorously,

$$\mathbf{q \not\equiv q_m} \qquad (9)$$

But instead,

$$\mathbf{q \equiv [d/dm(q_m)]} \qquad (10)$$

and this is a definition and therefore an identity. This definition alone affects all present electromagnetics theory.

To illustrate: In founding electrical theory, early scientists dealt with forces generated by charged masses (for example, charged pith balls). They later extrapolated the experimental results they obtained (or thought they obtained) with the smallest charged mass, a charged particle. In Figure 6, I show the classic situation for derivation of the idea of E-field (except we have used an electron for our test charge, rather than a pith ball).

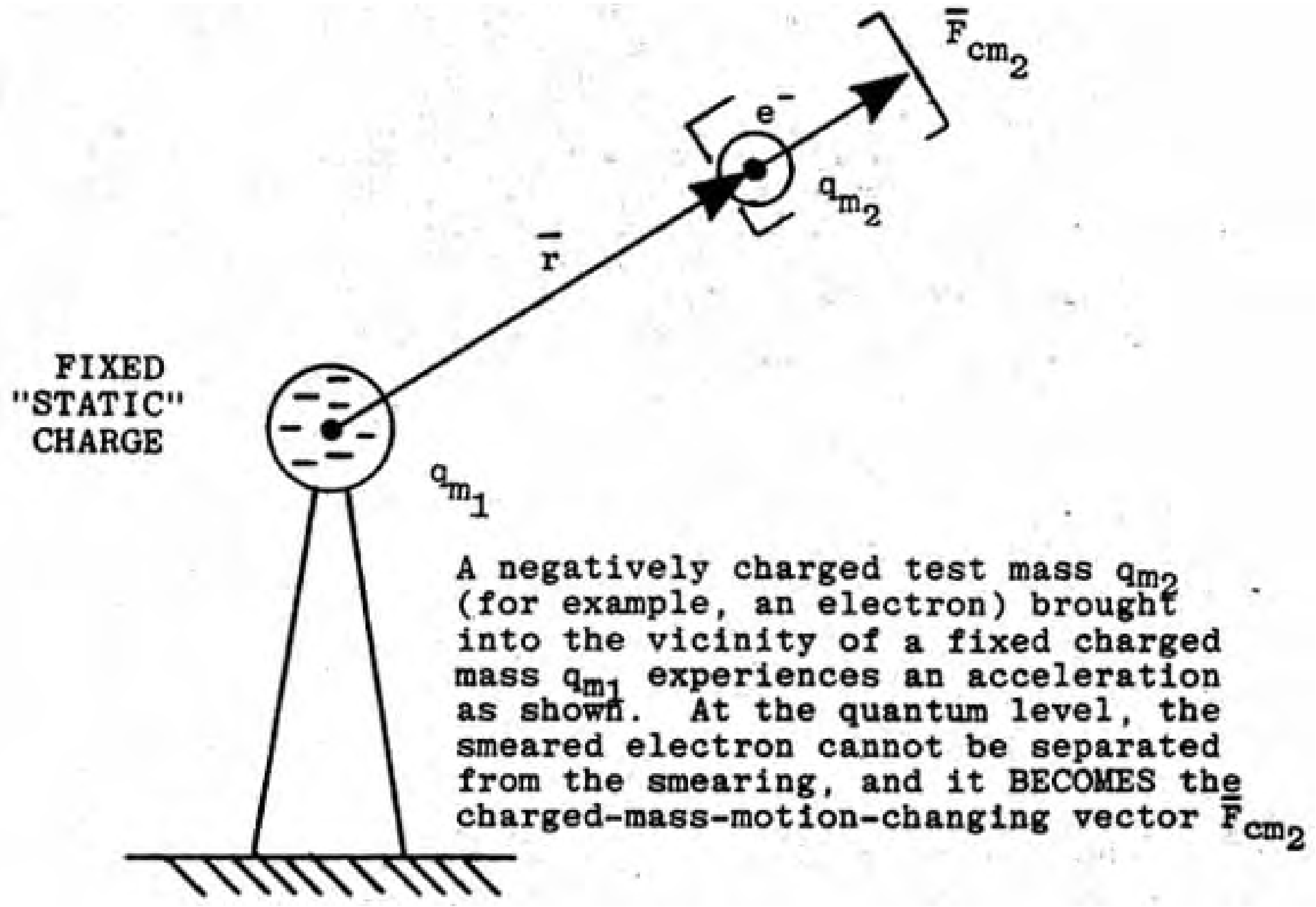

A negatively charged test mass q_{m_2} (for example, an electron) brought into the vicinity of a fixed charged mass q_{m_1} experiences an acceleration as shown. At the quantum level, the smeared electron cannot be separated from the smearing, and it BECOMES the charged-mass-motion-changing vector $\bar{F}_{cm_2}$

Figure 6. A test charge (charged mass) brought near a fixed charge (charged mass) experiences an acceleration.

Now note that what actually happens is that the unrestrained test charge becomes a CHARGED MASS SYSTEM VECTOR (a "smeared charged mass-motion changing"). The "test charge" BECOMES a charged mass force vector; it does not have a separate geometer's vector "appear on it." What actually happens is shown in Figure 7.

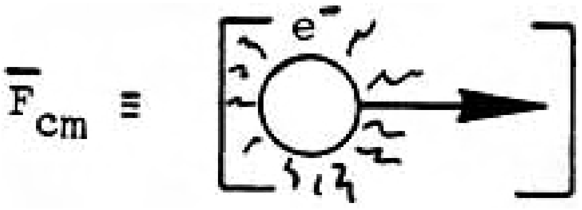

Figure 7. A charged-mass-system vector.

That is, in the simplest (nonrelativistic) case, for an electron what happens is

$$\bar{F}_{cm} \equiv [\overline{q_e \odot m_e \odot \bar{a}_s}] \qquad (11)$$

and this is a DEFINITION. That is, considered instantly, the electron exists as a charged-mass electrical force CONSISTING OF/COMPRISED OF a charge flux q_e canonically coupled to a mass, with that subsystem then canonically coupled to a spatial acceleration vector, ALL AS A SINGLE ENTITY, WITHOUT ANY "SEAMS" BETWEEN ITS "PARTS." The $\bar{F}_{cm}$ IS THE ELECTRON SYSTEM ITSELF; it is NOT a "spatial vector." Rigorously, it does not exist in the absence of the smeared electron mass, a priori.

Again, in assuming this force exists in the absence of the smeared mass of the moving particle, electromagnetics theory is in serious logical error.

Referring back to Figure 6, we see that, if we repeat the experiment many times and with the test charge in many locations, we have the situation shown in Figure 8.

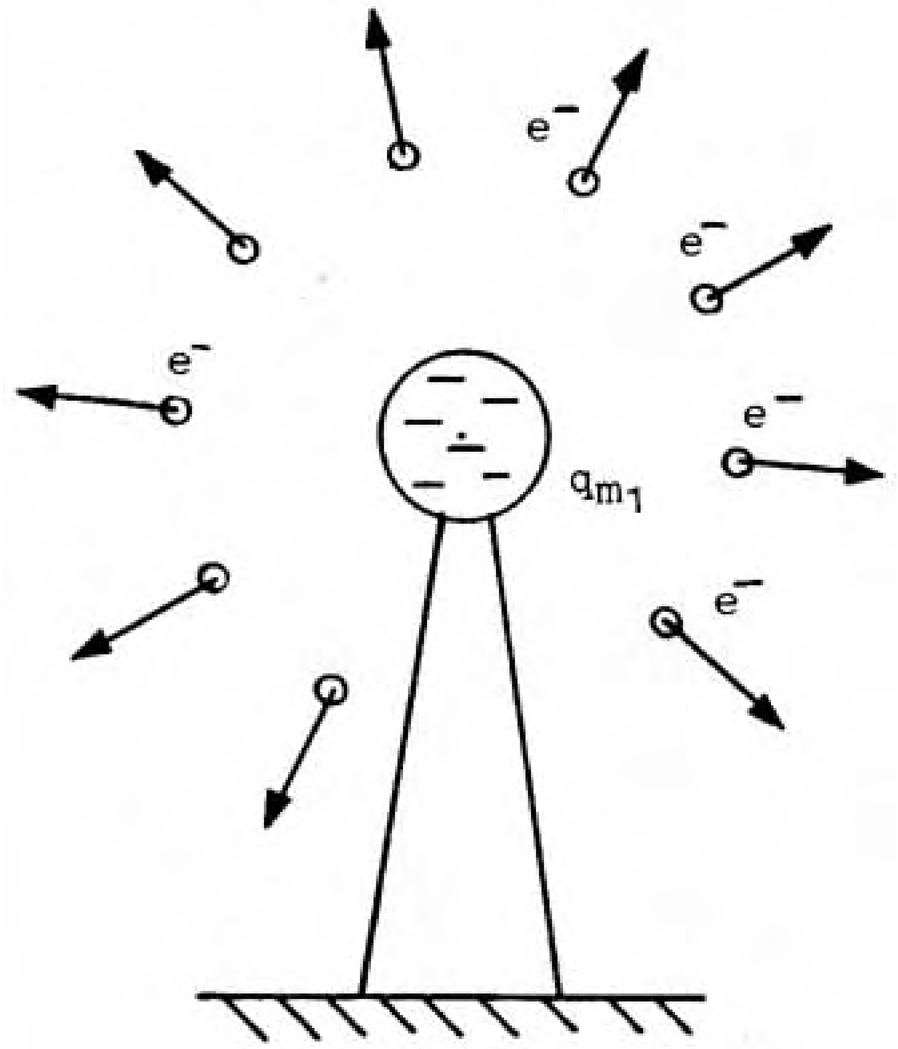

Figure 8. Repeating the "test charge" experiment.

It is found that, rigorously,

$$|\bar{F}_{cm}| \propto \left|\frac{q_{m_1} q_{m_2}}{r^2}\right| \qquad (12)$$

where $\bar{F}_{cm}$ is a charged mass system vector. Erroneously, this has been stated one way or another as

$$|\bar{F}| \propto \left|\frac{q_1 q_2}{r^2}\right| \qquad (13)$$

where $\bar{F}$ is assumed to be a spatial system vector. Further, this confusion has been carried over into the definition of the $\bar{E}$-field as:

$$\bar{E}_{cm} \equiv \frac{\bar{F}_{cm}}{m_{cm}} \qquad (14)$$

In this definition, $\bar{E}$ -- which is a charged mass system vector -- has been confused as a charged spatial system vector, where $\bar{F}$ is regarded simply a spatial system vector! Actually, the definition of the $\bar{E}$-field should be

$$\bar{E}_{cm} \equiv \frac{\bar{F}_{cm}}{m_{cm}} \qquad (15)$$

where $\bar{F}_{cm}$ is a charged mass system vector. Failure to properly define the $\bar{E}$-field has caused the conception of the $\bar{E}$-field to be falsely perpetuated as existing in vacuum.

The $\bar{E}$-field is TREATED this way in present EM theory. Hence present theory falsely assumes that the observable $\bar{E}$-field can exist in vacuum.

What actually exists in space, $\bar{E}$-field-wise, is a special kind of ordered virtual state pattern in a series of spinning "scalar" fields. This virtual state pattern or "shadow vector" field will be explained later.

Note again that one cannot have a "force vector" existing in vacuum - a priori.

However, assume for a moment that one could have a massless force vector, as presently assumed. Let this force vector appear at a point in the vacuum. Since the vacuum has zero observable mass, it would have zero inertial resistance to this hypothetical observable force hence the observable force would instantly produce an "infinite" acceleration of its point of application, vanishing with it into the distance. Therefore our fictitious force would disappear the instant it appeared! In any case, it could not be retained at a point in the vacuum for any finite length of time, however small.

The direct implications are that (1) something other than an observable electrical force field exists in the vacuum, and (2) there must exist a more fundamental mechanism by which this "something else" generates or CREATES a change on/of an accelerating electrically charged mass particle. (Note again that at the basic level, any particle of mass is ALWAYS quivering and accelerating, from quantal fluctuation considerations alone.) Causality has no arrow microscopically.[23]

Next Page

Help support this work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

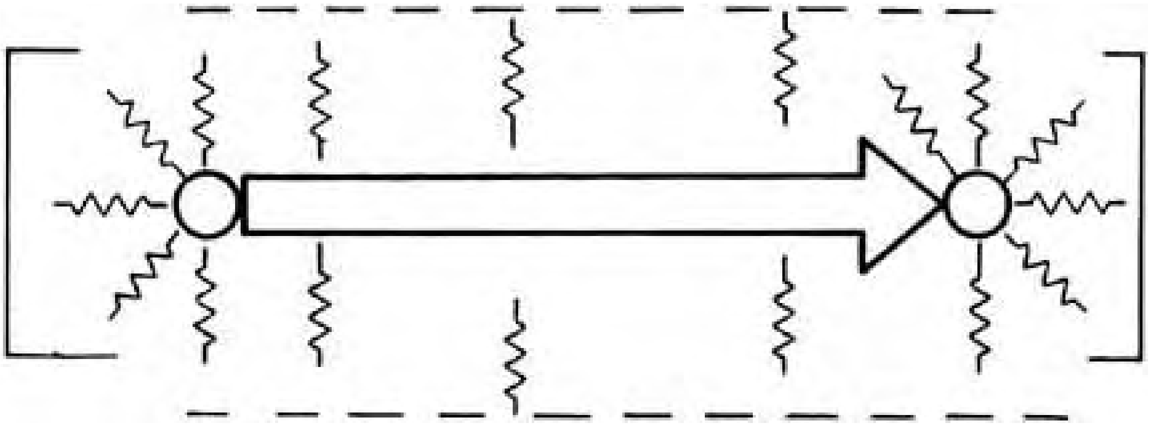

(ADVANCED ELECTRICIAN'S VECTOR)

Figure 9. Charged spatial (massless) vector.

-- CHARGED SPATIAL (MASSLESS) VECTOR --

We recognize now that

$$q_c \neq q_m \qquad (16)$$

and that q_c is simply the virtual-particle flux that constitutes charge -- and indeed constitutes vacuum itself!

We DEFINE vacuum, based on Figure 5, as shown in Figure 10.

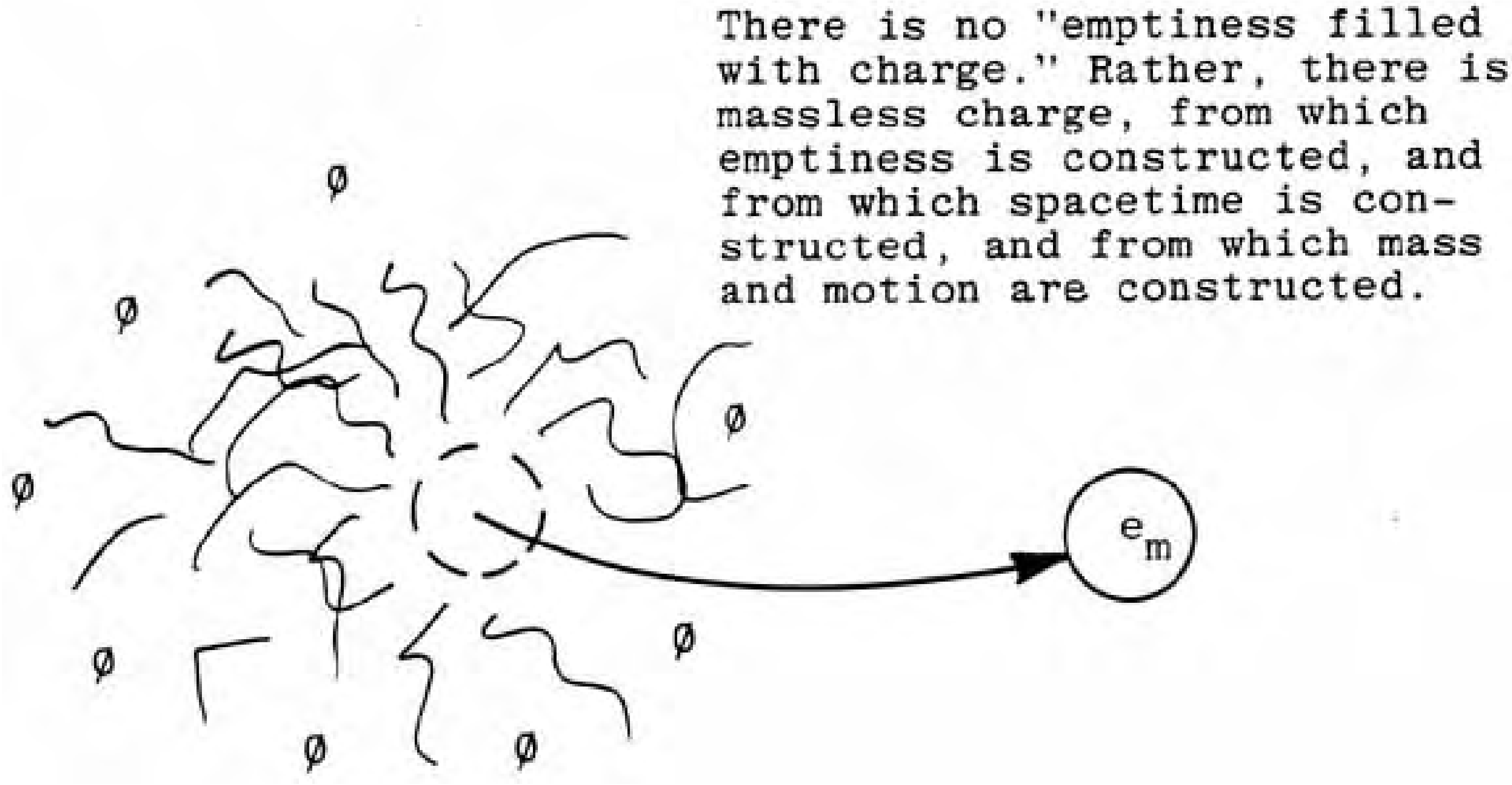

There is no "emptiness filled with charge." Rather, there is massless charge, from which emptiness is constructed, and from which spacetime is constructed, and from which mass and motion are constructed.

Figure 10. Removing the bare particle (mass) from
a charged particle leaves the charge.
The vacuum is DEFINED AS the charge.

That is, vacuum may be defined as pure massless charge flux. This flux IS identically "spacetime" as well. Vacuum is pure Ø-field (electrostatic scalar potential). Here again, in present theory it is assumed that

$$Ø_O \equiv 0 \qquad (17)$$

which, by our new definition of vacuum, is quite false.

We now note that, if we insist on assigning a spatial vector to the vacuum, we have the situation shown in Figure 11.

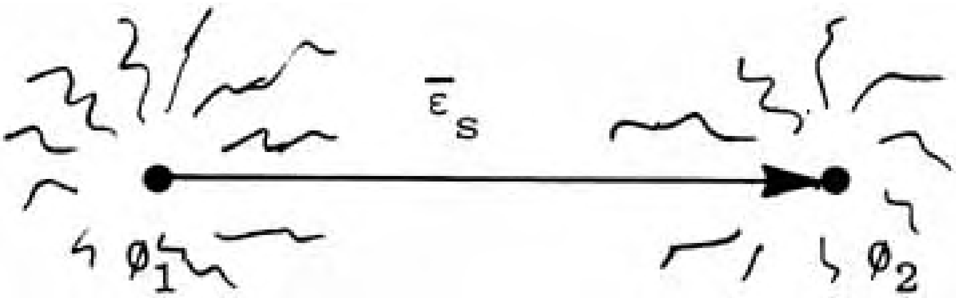

Figure 11. Assigning a spatial vector to the
charged vacuum

Next Page

Help support the work

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- THE SHADOW VECTOR --

Note that this spatial vector $\bar{\epsilon}_s$ represents $-\overline{\nabla\phi_s}$, that is,

$$\bar{\epsilon}_s = -\overline{\nabla\phi}_s \qquad (18)$$

but $\bar{\epsilon}_s$ cannot be a force (mass system) vector. It can only exist as an ordered pattern in the virtual flux between two separated points of the vacuum; that is, as an ordered pattern in the virtual state. Literally, $\bar{\epsilon}_s$ exists only as a tiny bit of order existing in great disorder.

In other words, the present EM theory is incorrect in stating that

$$\bar{E}_m = -\overline{\nabla\phi}_s \qquad (19)$$

in vacuum in the absence of an observable spinning charged particle, since

$$\bar{\epsilon}_s \neq \bar{E}_m \qquad (20)$$

The actual existence of $\bar{\epsilon}_s$ may be visualized in terms of successive differentials of $\bar{E}_m$, broken into differentials $\bar{e}_m$ so small that, observably, each little differential's mass component m has become virtual, so that

$$\bar{e}_m \equiv \bar{e}_{vm} \qquad (21)$$

where subscript m stands for mass, subscript v for virtual, and observably

$$\bar{e}_{vm} \equiv \overline{dE}_m \qquad (22)$$

but, in the absence of a spinning charged particle,

$$\bar{E}_m \neq \int \overline{dE}_m \qquad (23)$$

since the $\bar{e}_{vm}$ components remain individually separated. That is, in macro-time a SHADOW force vector exists, made of microscopically ordered BUT UNJOINED (unintegrated) "virtual state" vector differentials of what would be an observable mass system force vector $\bar{E}_m$ if integrated.

Thus, the "$\bar{E}$-vector" $\bar{\epsilon}_s$ that exists in vacuum is a "shadow" vector as shown in figure 12.

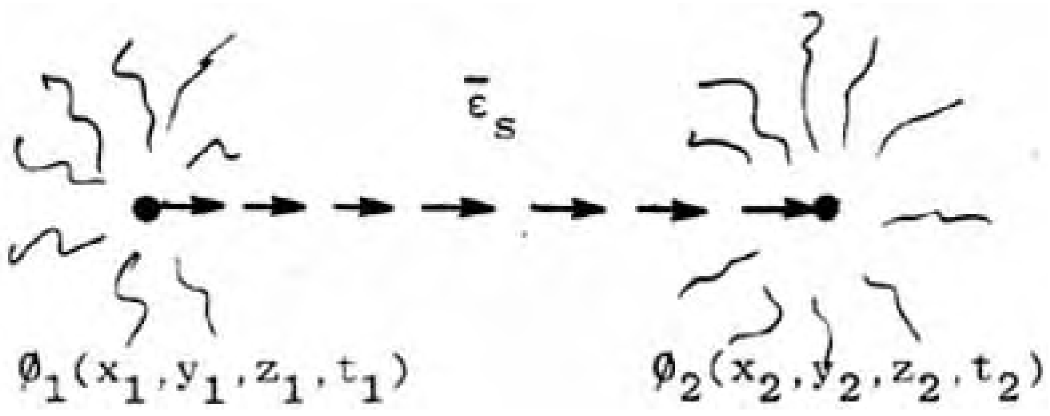

Figure 12. A "shadow vector" $\bar{\epsilon}_s$

We say that such a previous mass system vector, broken into ordered but unjoined virtual vectors by the absenting of all mass, is a SHADOW VECTOR, and we label it with a subscript vm, to represent "virtual mass" system. To the macro observer, this is the kind of "vector" that exists in vacuuo.

Note that, observably, the shadow vector merely represents a special ordering in $\nabla\phi$. It is NOT an OBSERVABLE (mass system) vector, but it IS an ordered series of consecutive virtual vectors.

With each virtual bit vector, a virtual time exists as well, and these "virtual time bits" are also ordered consecutively (in macrotime).

I point out that any observable vector must be finite, and so it must have a finite magnitude (finite length). In the simplest case, this length ΔL is related to a Δt by

$$\Delta L = c\Delta t \qquad (24)$$

What I am saying is that ANY observable spatial vector is actually a spatiotemporal vector, and the MAGNITUDE of any vector is related to TIME (to the existence of that vector in time) at the most fundamental level. Suffice it to say that, if the fundamental quantum level (Δt) aspect of a vector is interfered with, then the MAGNITUDE of the vector is interfered with.[24] That is, if we can make a time wave, we can change or affect ANY vector's magnitude, including the magnitude of mass system vectors and charged mass system vectors. Such a "time wave" can be made easily, and it has been.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- A Scalar is a Zero Vector --

Now let us look at the idea of a scalar.

A "scalar" may in a general sense be considered as the sum of the "absolute values" of the individual vector components of a system of vectors whose observable resultant is zero. That is, it represents the magnitude of the internal stress of a vector system, with the absence of a single observable directionality of the system. It also follows that every scalar is actually a stressed zero vector, and every zero vector is a scalar.

Thus we have four major types of scalars related to the four types of vectors:

(a) $S_1 = |\overline{v}_s|$ (25)

(b) $S_2 = |\overline{v}_m|$ (26)

(c) $S_3 = |\overline{v}_{cm}|$ (27)

(d) $S_4 = |\overline{v}_{cs}|$ (28)

where S stands for scalar, $\overline{V}$ for vector, and subscript s for spatial, m for mass, and c for charged.

For example, comparing equations (25) and (26), it can easily be seen that twice as many "point-motions" is not at all the same thing as twice as many "gram-mass-motions." The two resulting vector systems are quite different.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Virtual and Observable Aspects --

We must also examine some aspects of "virtual" and "observable."

For example, we construct several spatial vector summations in Figure 13. The "resultants" of these spatial vectors are all equal. However, the actual sums, even though equal, are quite different, because their internal "stresses" (substructure forms) are quite different.

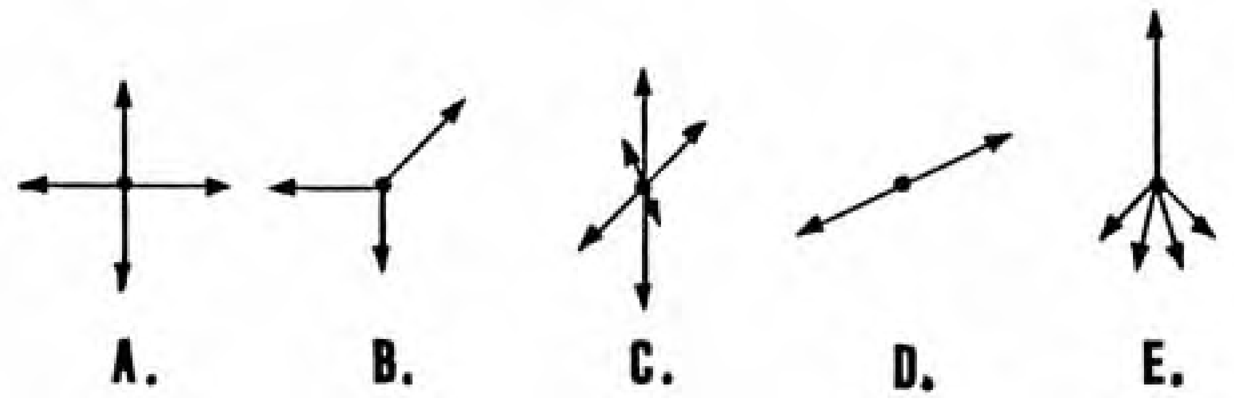

SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE.

When the time aspects of the vector systems of Figure 13 are considered, one can easily understand the problem. That is, the resultant of each of these "systems" is zero, and so one can say that the vectorial "magnitude" of the system is zero since the magnitude of the resultant vector is zero. However, in each case the "action" represented by each vector element actually occurs in a finite tiny Δt. So: (1) The zero resultant must exist for a finite Δt, and (2) all the actions indicated by the system component vectors actually occur in that Δt. The absolute value of the "activity per unit time per unit volume" of such a zero-resultant system thus has physical meaning, and one may refer to this notion as the "stress" on spacetime[25], or the "electrostatic scalar potential" of the system. Note that this differs from the present definition of electrostatic scalar potential, which becomes just a special case of the more fundamental potential defined here.[26]

The derivatives of this spatiotemporal stress also have physical meaning. The time derivative is indicative of the stress on the flow of macroscopic time at a fixed spatial point, and the spatial derivative is indicative of the stress on space. Here one is confronted with the fact that what we call "space" and "time" are continually being created, directly in the physical observing/detecting apparatus itself.[27] That is, rigorously, "detected physical reality" exists totally in and of the mass-changes of the observer's mass or his detecting instruments. In the fundamental detection process itself, there is a flow of the rate of creation of spatial lengths and a flow of the rate of creation of time lengths. Indeed, to a linear observer the stress on the creation of the flow of time controls the flow of the creation of space, and the stress on the creation of the flow of space controls the flow of the creation of time. The change in the stress on 4-space (ordinary Minkowskian space-time) controls the "curvature of that spacetime" in the fifth dimension. The change in the stress on 5-space controls the "curvature of that 5-space spacetime" in 6-space, and so on. Development of these facets of the new concepts is beyond the scope of this paper.)

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- SCALARS AND VECTORS HAVE SUBSTRUCTURES --

As can now be seen, the sum of each structure in figure 13 is observably zero. Therefore we might define the sum as a "zero spatial vector." We note, however, that it actually exists for a time Δt and is thus a spatiotemporal entity, rigorously.

If we define the internal stress action A in a region $\Delta s^3 \Delta t$ of spacetime as

$$A_{(s^3t)} \equiv \sum_{k=1}^{n} |\bar{i}_s| \qquad (29)$$

and the 4-space internal stress intensity or potential as

$$S \equiv \lim_{(\Delta s^3 \Delta t) \to 0} \left[\frac{A(s^3t)}{\Delta s^3 \Delta t} \right] \qquad (30)$$

where $\bar{i}_s$ is any internal vector in the substructure, Δs^3 is the spatial volume (about a point) containing vector $\bar{i}_s$, and Δt is the inseparable time during which these component actions occurred, then we see that, stress-wise, all the "zero-vectors" in figure 13 are quite different in their internal stresses, 4-space potentials, and internal substructures. For the five "zero sum" vectors, OBSERVABLY we have

$$\bar{0}_1 = \bar{0}_2 = \bar{0}_3 = \bar{0}_4 = \bar{0}_5 \qquad (31)$$

whether or not

$$\Delta s^3 \Delta t_m = \Delta s^3 \Delta t_n \; (m \neq n;\; 1 \leq m \leq 5;\; 1 \leq n \leq 5) \qquad (32)$$

But considering the substructures,

$$\bar{0}_m \neq \bar{0}_n \quad (1 \leq m \leq 5;\ 1 \leq n \leq 5;\ m \neq n) \qquad (33)$$

I now point out that a scalar can be regarded as a stressed zero-sum vector, where the magnitude S of the scalar represents the internal stress intensity caused by the substructure of the zero-vector.

Thus, generally,

$$S(s^n t) \equiv \lim_{\Delta s^n \Delta t \to 0} \left[\frac{\sum_{n=1}^{k} |\bar{i}_s|}{\Delta s^n \Delta t} \right] \qquad (34)$$

That is, in general any observable scalar has, consists of, and is comprised of a VIRTUAL (unobservable) substructure that is very real indeed. One must also consider the scalar as existing for some finite time Δt, (at least for the time of one quantum change), and the intensity of the virtual actions occurring in the spatiotemporal substructure of the scalar during that time Δt is proportional to the magnitude of the scalar.

Normally, the concept of a scalar -- as presently used -- makes no allowance for the scalar to exist in time, or for a virtual vector substructure, or for any patterning inside the substructure. This is equivalent to assuming that

$$A \equiv 0 \qquad (35)$$

and that all $\bar{i}_s$'s are evenly distributed. That is, from this new viewpoint, presently the mathematical theory assumes all scalars to have an equal density of virtual activity per spatiotemporal volume in its virtual substructure, and an isotropic virtual pattern distribution of an infinite number of equal virtual vectors in its 4-space substructure.[28]

In the new approach, neither of these two assumptions need hold -- though in special cases they can hold. Thus present orthodox theory is just a single special case of a more fundamental approach indicated here.

Note that, by directly affecting and changing the virtual substructures of scalars and vectors, we can directly perform virtual. state engineering, and this allows us to directly "engineer" the so-called "laws of nature" of the normal observable laboratory state and thus ENGINEER AND CHANGE PHYSICAL REALITY ITSELF.[29]

In the new approach, we can (observably) have

$$2 + 2 \neq 4 \qquad (36)$$

or

$$2 + 2 = 4 \qquad (37)$$

by the following means: In the first case (equation 36), we assume that the virtual substructures are patterned, and interact nonlinearly in such a way as to produce an extra observable. Thus we have a delta added to the normal observable scalar results of the interaction, as follows:

$$2_o + 2_o = 4_o + \Delta_{v\text{->}o} \qquad (38)$$

where subscript "o" means observable and "v" means virtual. Note that

$$\Delta_{v\text{->}o} \qquad (39)$$

indicates a delta due to virtual substructure interactions yielding an extra observable delta. This extra delta may be either scalar or vector in nature, depending on the circumstances and the particular interactions.

Note also that any vector or scalar must now be considered to HAVE, CONTAIN, and CONSIST OF an infinite substructure. And note that, similar to the scalar case, from the new viewpoint the present theory assumes each scalar (point) of the vector to have a structure similar to that of equation (34), except that now the scalars are ordered, with a linearly decreasing internal stress density per unit scalar along the line of the vector.

In the new approach, vector interaction (superposition, for example) can now violate present theory, if the two virtual substructures interact nonlinearly to produce a nonzero, observable delta. Observably (macroscopically) , this delta, again, may be either "scalar" or "vector."

This approach now becomes consistent with quantum mechanics at the foundation level.

Next Page

TOWARD A NEW ELECTROMAGNETICS PART III: CLARIFYING THE VECTOR CONCEPT

Help support the work

-- Substructures, Virtual Levels, and Hyperspaces --

In the new approach, our definitions and assumptions immediately drive us to a picture of an infinite set of nested levels of substructures in the virtual state. That is, anyone component (scalar or vector) in one level of virtual state has an infinite number of even finer virtual components, one level more subtle.

AND THAT IS WHAT PARTICLE PHYSICS AND QUANTUM MECHANICS ALREADY REVEAL ABOUT THE STRUCTURE OF PHYSICAL REALITY.

So these definitions and assumptions now provide the basis for a new vector mathematics that is in accord with, and fitted to, modern physical observations.

We have a picture such that any observable scalar or vector contains a virtual substructure (virtual level 1). Any scalar or vector in virtual level 1 also contains a finer virtual substructure, in virtual level 2. And so on ad infinitum.

Each succeedingly finer level of virtual state can be modeled as a hyperdimension (higher spatial dimension) as I pointed out in Appendix 1 to my book, The Excalibur Briefing.[30]

Thus this approach immediately ties into hyperdimensional or hyperspatial theory -- such as Elizabeth Rauscher's 8-dimensional theory[31] and C. Muses's hypernumber theory.[32]

The new definitions and assumptions are far richer than what is allowed by tensors, though there are many similarities. Muses's work, however, essentially can encompass most of these definitions and concepts, except the distinct types of vectors are not so clearly delineated in his theory (at least to my comprehension of it.) His theory does provide a nested, hyperdimensional structure of time, however, and thus allows "scalar" waves in the hyperspatial structure of time -- in other words, observably "scalar" waves in the virtual state structure of spacetime, or pure Tesla waves, or simply "time" waves.

These are the bare notes; from this approach, already new (proprietary) mechanisms and exact specifications to make scalar waves -- in essentially whatever quantity and degree desired -- appear to have been successfully accomplished by my close colleagues.

The new approach is real and it leads to a new physics. And I believe that the very beginnings of the new physics are already working on the laboratory bench.

Nikola Tesla discovered the most essential features of the new electromagnetics over eighty years ago and was simply suppressed for his efforts. Now, although it has been eighty years in the reborning, Tesla electromagnetics is once again loose in the Western world.

This time, let us hope that it fares better at the hands of orthodox science and large financial control groups than it did for Nikola Tesla.

Next Page

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

by Lt. Col. T.E. Bearden (retd.), 1983

Copyright

Title Page

Foreword

1. TOWARDS A NEW ELECTROMAGNETICS - ACKNOWLEDGEMENTS
2. OUTLINE
3. CHARACTERISTICS OF SELECTED CONCEPTS
4. QUOTE FROM NIKOLA TESLA
5. FORCE BETWEEN TWO LIKE CHARGES
6. REPETITIONS LEAD TO THE VECTOR "FIELD" CONCEPT
7. A SCALAR HAS A HIDDEN (VIRTUAL) VECTOR SUBSTRUCTURE
8. SCALAR AND VECTOR SUBSTRUCTURES
9. SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE
10. PHYSICAL VECTORS
11. THE FOUR VECTORS OF PHYSICS
12. FORCE DEFINITION
13. FORCE IS AN EFFECT, NOT A CAUSE
14. FORCE AND Ë-FIELD DO NOT EXIST IN VACUUM
15. COROLLARIES
16. ENERGY IS DERIVED, NOT FUNDAMENTAL
17. EXISTENCE IN, AND PASSAGE THROUGH, TIME
18. SOME IMPLICATIONS
19. CONSERVATION OF ANENERGY (AE): (AN EXPANDED CONSERVATION LAW)
20. FREE ENERGY IS ALLOWED

21. SCALAR POTENTIAL CAN FORCE THE RELATIVISTIC CASE
22. SCALAR STRESS WAVE (Ö-WAVE) (TESLA WAVE)
23. SCALAR Ö-WAVE PRODUCTION
24. A SIMPLIFIED CHARGED PARTICLE
25. A CHARGED PARTICLE IS A SPECIAL KIND OF "SPRAY NOZZLE"
26. AN OBSERVABLE CHARGED PARTICLE IS DIRECTLY CONNECTED TO VIRTUAL PARTICLE FLUX VACUUM
27. SPIN COUPLES VACUUM TO PARTICLE
28. ELECTRON PRECESSION (SIMPLIFIED)
29. VACUUM/SPACETIME IS PURE VIRTUAL PARTICLE FLUX (Ø$_0$)
30. VACUUM IS:
31. GRADIENT OF ELECTROSTATIC POTENTIAL
32. IN A ∇Ø, A CHARGED PARTICLE MOVES ITSELF
33. ELECTROMAGNETIC WAVES IN ELECTRON GAS IN A CONDUCTOR
34. ELECTROMAGNETIC WAVES IN VACUUM
35. DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES
36. CREATING A "RIVER" OF ANENERGY
37. A 2-WIRE SYSTEM IS CONSERVATIVE
38. A HOLLOW CHARGED SPHERE
39. ONE-WAY GATE VALVE
40. SIMPLE FREE ENERGY MOTOR
41. THE TRANSLATOR
42. TRANSVERSE WAVE INTERFERENCE
43. LONGITUDINAL WAVE INTERFERENCE (SCALAR WAVES)
44. CREATING ENERGY AT A DISTANCE
45. NONLINEARITIES CAUSE SHIFTING (PARTIAL TRANSLATION)
46. THE SCALAR AURA.
47. VIRTUAL FLUX AND MAGNETIC FIELD (LINEAR CASE)
48. PRODUCING MAGNETIC MONOPOLES
49. BREAKING METAL WITH MAGNETIC MONOPOLES
50. CREATING AN INERTIAL FIELD (ANTIGRAVITY)
51. EFFECTS ON HUMANS
52. INITIATING SPECIFIC BIOLOGICAL EFFECTS
53. GROUND-TO-AIR USE

54. TACTICAL AIR-TO-GROUND USAGE
55. INTEGRATED WEAPONS DISPLAY (MODIFIED JAMMER)
56. RADAR INVISIBILITY
57. NETWORK OF VIRTUAL TRANSMITTERS
58. CONCLUSION

Return to Homepage

TOWARD A NEW ELECTROMAGNETICS

PART 4: VECTORS AND MECHANISMS CLARIFIED

(©) 1983 T. E. Bearden

All rights reserved.

Delivered to the 1983 Annual Conference

of the

U. S. Psychotronics Association

July 20-24, 1983

at the

University of Portland

Portland, Oregon

--FOREWORD --

This paper is an adaptation of the oral/visual presentation given by the author at the 1983 Annual Conference of the U. S. Psychotronics Association on Saturday, July 23, 1983.

The presentation was delivered directly from the vugraphs alone, and not from a formal, written text.

The title of the presentation at the conference was "New Electromagnetics: Where the 'Old Math' Took the Wrong Turn."

As originally conceived, the lecture was to be given in a single period of some 45 minutes and cover only the first part of this paper, the clarification of the vector concept in physical theory. The Symposium director so graciously scheduled two periods instead, allowing time to present additional material further clarifying vector and scalar waves, and to present a variety of unusual physical mechanisms that apply the concepts.

After the conference, the author prepared this paper directly from the same slides, in the same order. However, no attempt was made to match the words in this paper to the words actually delivered in the formal presentation. The gist of the two, however, are the same -- except that, in the conference lecture, the author covered (as an aside) some important additional material: (1) the nature of what actually occurs in an "electrical ground," and (2) the author's explanation of the theory of Tesla's magnifying transmitter.

An audio tape and a videotape of the author's USPA presentation were made by the conference officials. Copies are available at nominal cost from Mr. Robert Beutlich, Secretary/ Treasurer, U. S. Psychotronics Association, 3459 Montrose Avenue, Chicago, Illinois 60618, U.S.A.,
phone (312) 478-7715.

The Tom Bearden Website

SLIDE 1.

TOWARD A NEW ELECTROMAGNETICS

© 1983 T.E. BEARDEN

ALL RIGHTS RESERVED

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to two of my fellow colleagues, Dr. William Tiller and Mr. Frank Golden, for many fruitful hours of discourse and interchange on these and related subjects.

Also, I wish to thank the Association of Distinguished American Scientists (A.D.A.S.) for support and assistance in the long and arduous struggle to revise the basic foundations of electromagnetics and mechanics.

Next Slide

SLIDE 2.

OUTLINE

- **MAIN BRIEFING**
  - **> TRANSLATOR EFFECTS**
  - **> POTENTIAL WEAPONS**
  - **> INCIDENTS**
  - **> CONCLUSION**
- **HISTORICAL BACKGROUND**
  - **> TESLA/MORAY**
  - **> WWII AND AFTER**
- **ELEMENTS OF THE EMERGING THEORY**
  - **> LONGITUDINAL/ TRANSVERSE WAVES**
  - **> VACUUM/MASS EFFECTS**
  - **> FLAWS IN VECTOR THEORY**
  - **> IMPLICATIONS**

This presentation is one part of a three part presentation, consisting of the main briefing on physical effects, the historical background, and the elements of the emerging new electromagnetics.

In this present briefing I will not cover the historical background. Instead, first I will address the concepts of the new theory, and then I will give an abbreviated main briefing, including some weapon aspects. I will present one simple free energy motor and show how it works, and where it gets its energy.

Much of this material I have known for some time, but have hesitated to release it because it enables one to develop -- in a fairly straightforward manner -- weapons and beams (ray) systems that can disable electronics, computers, communications, and people. Anyone handy in electronics, who has a few thousand dollars and access to bench equipment and an electronics parts supply house, can develop these weapons in his or her basement workshop. But because certain very radical elements in this country already know some of this information and may plan to utilize such devices in criminal activities, I am openly releasing the information as a warning, to be prepared.

Of course there are many good things that can be developed with this technology, including machines to heal a remarkable variety of ailments and illnesses. I strongly urge everyone to use this information wisely and only to benefit other human beings. There exists a cosmic or karmic law for those who misuse it against their fellow persons; if one does that, one will assuredly suffer the consequences.

Also, I strongly urge the development of Western defenses against the Soviet strategic and tactical weapons; already developed utilizing these principles. Unless the West wakes up to the extreme urgency of this task, in the near future we may be overwhelmed in short order by a much harsher, dictatorial system. If that should happen, it will plunge the human race into a new dark ages far more terrible than that dark period in history following the fall of Rome.

These are the final times, and this is the beginning of the final technology. The ability to directly engineer physical reality itself has been laid in our laps. Pandora's box is already spilled open, and there is no longer any stopping the new technology.

Let us prepare, then, for that which is to come.

With these introductory remarks, let us now turn to the main presentation.

Next Slide

The Tom Bearden Website

SLIDE 3.

CHARACTERISTICS OF SELECTED CONCEPTS

CONCEPT	CHARACTERISTIC
• ENERGY	**FLOW OF OBSERVABLE MASS PARTICLES**
• ANENERGY	**FLOW OF VIRTUAL PARTICLES**
• PRESENT "CHARGE"	**VIRTUAL PARTICLE FLUX TIED TO OBSERVABLE MASS PARTICLES**
• TRUE CHARGE	**VIRTUAL PARTICLE FLUX WITHOUT MASS**
• FLUX	**MULTI-DIRECTIONAL FLOW THROUGH A POINT**
• FLOW	**UNI-DIRECTIONAL MOVEMENT OF A FLUX [GRADIENT]**
• CURRENT	**UNI-DIRECTIONAL GRADIENT OF A [USUALLY CONSTRAINED] FLUX**

On this slide I show some fundamental defining characteristics of energy, "charge" or charged mass as presently in the theory, flux, flow, and current. These fundamental concepts are often somewhat confused in the literature.

Particularly note that the concept of energy, being tied to "the capacity to do work," is specifically tied to the concept of mass. Eventually, energy must be expressed in terms of moving masses -- specifically, in accelerating or decelerating masses.

I introduce the term "anenergy" here, specifically in relation to

the concept of a virtual particle. Anenergy is roughly equivalent to the present idea of "virtual energy,"
except it more precisely implies the flow or flux of virtual particles.

We also redefine charge as virtual particle flux, in order to remove the present error in EM theory where "charge" always implies a charged mass. The electrostatic scalar potential, phi (Ø) , then merely becomes the intensity of the massless charge -- that is, the intensity of the virtual particle flux -- at a point. Let me point out, however, that in the new view this "point" is always in n-dimensional space, where n is equal to or greater than four. The present 3-space formulation of Ø is thus a highly special case.

Next Slide

The Tom Bearden Website

SLIDE. 4

QUOTE FROM NIKOLA TESLA:

"THE HERTZ WAVE THEORY OF WIRELESS TRANSMISSION MAY BE KEPT UP FOR A WHILE, BUT I DO NOT HESITATE TO SAY THAT IN A SHORT TIME IT WILL BE RECOGNIZED AS ONE OF THE MOST REMARKABLE AND INEXPLICABLE ABERRATIONS OF THE SCIENTIFIC MIND WHICH HAS EVER BEEN RECORDED IN HISTORY."

Nikola Tesla
"The True Wireless,"
The Electrical Experimenter

One of the things we will find is that Hertz or transverse electromagnetic waves are strictly mass waves in a charged mass particle medium.

For example, transverse EM waves do not, and CAN not, exist in vacuum. Vacuum EM waves are longitudinal waves of compression and rarefaction. What is compressed and rarified is the virtual particle flux that comprises vacuum itself. Vacuum EM waves in this virtual flux are just like sound waves in air -- in the flux of molecules and atoms that comprise the atmospheric gas.

Nikola Tesla, of course, knew this and he frequently pointed out the falsity of the accepted transverse wave theory. Here is one Tesla quote, for example, taken from The Electrical Experimenter. (PAUSE)

Almost every scientist -- even the "Tesla bugs" -- have thought Tesla wrong in his adamant opposition to the idea of Hertzian waves in the vacuum.

As we shall see, Tesla was right and all the present scientists and textbooks are wrong .

Next Slide

The Tom Bearden
Website

SLIDE 5.

FORCE BETWEEN TWO LIKE CHARGES

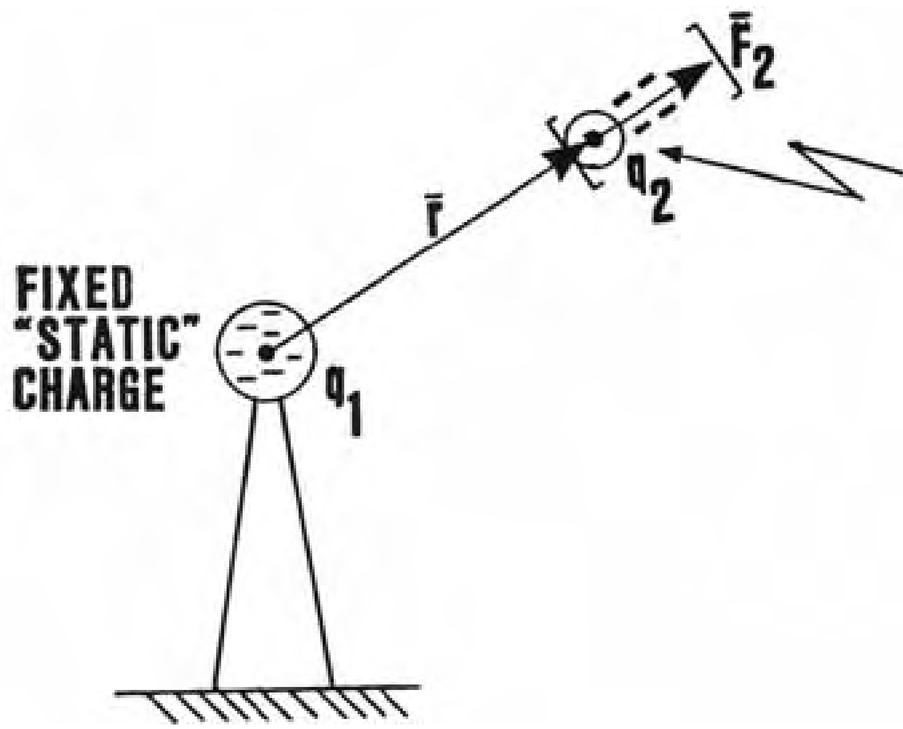

A NEGATIVELY CHARGED TEST MASS [e.g., AN ELECTRON] BROUGHT INTO THE VICINITY OF q_1 EXPERIENCES A FORCE $\bar{F}_2$ AS SHOWN. NOTE THE ELECTRON ACTUALLY IS SPINNING. THIS HAS NEVER BEEN DONE WITH A TRUE "STATIC CHARGE", FOR THERE IS NO SUCH THING IN NATURE.

In my presentation last year, I noted some 22 fundamental flaws in present electromagnetics theory. One of these flaws was in the actual vector theory itself. As we shall discover, one cannot simply plug geometrical vectors into physical systems without making fundamental errors. The vector concept itself must be adapted, for there are now four different kinds of vectors in physics, all confused as the same thing. We will come to that shortly.

To begin, let us look at the fundamental kind of experiment, shown on this slide, from which the idea of electrical force and electric field was taken.

We have a fixed static charged object q(l) in the laboratory, and we bring in a test charged mass q(2). The test charged mass may even be a single electron, as we illustrate here. We show in our example the fixed charge to be negative. When we release the test charge electron, it accelerates away from the fixed charged mass, as shown on the diagram. Note that we have no force at all until we bring in the test charged mass. Indeed, we shall find that the force is actually COMPRISED OF the accelerating test electron. Specifically, it is not something mysterious which appears in vacuum and PUSHES on the electron. We shall find that FORCE IS AN EFFECT, NOT A CAUSE. It CONSISTS OF the "smeared mass" that is accelerating -- in this case, it consists of the smeared, accelerating electron.

We summarize the experiment: in the absence of the charged test particle, no force exists in the vacuum. When we bring in the test charged mass, it accelerates away, and THE FORCE CONSISTS OF THAT ACCELERATING CHARGED MASS PARTICLE. The experiment does not at all address or deal with WHAT EXISTS IN THE VACUUM AROUND CHARGED MASS Q(l) IN THE ABSENCE OF TEST CHARGED MASS Q(2). Any assumption of a force existing in the vacuum is sheer speculation and not based on this experiment.

Next Slide

The Tom Bearden
Website

SLIDE 6.

REPETITIONS LEAD TO THE VECTOR "FIELD" CONCEPT

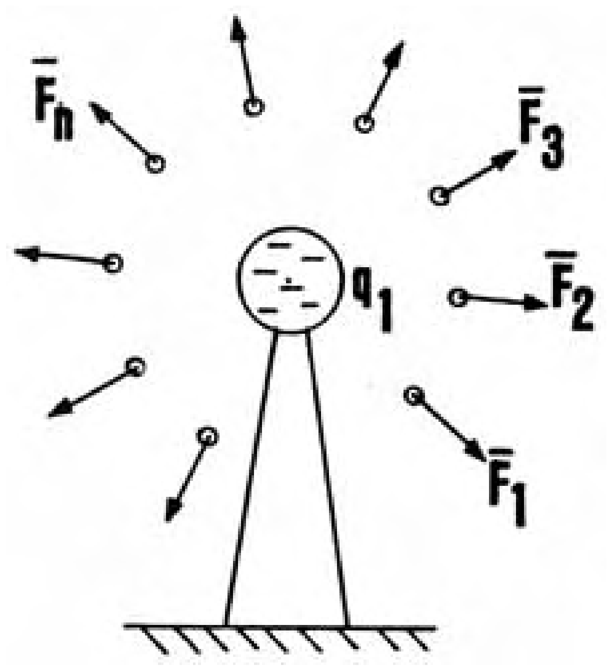

REPEATING FIGURE 4 A NUMBER OF TIMES LEADS TO SETS OF FORCE ARROWS, WHICH WE THEN VISUALIZE IN TOTAL AS A FIELD, Ë, OF FORCE(S) ABOUT q_1

RIGOROUSLY, THIS "FORCE FIELD" WOULD REQUIRE THAT A SPINNING CHARGED PARTICLE OF MASS EXIST AT EACH AND EVERY POINT IN SPACE ABOUT q_1

On this slide, we show that, as we repeat the experiment by bringing in the test charged mass from differing directions, the same thing happens around the fixed charged mass q(l). Test charged mass q (2) -- our electron -- accelerates radially away from q(l).

As we repeat the experiment and draw many of the little arrows radially outward, gradually the idea of a "vector field" emerges. That is, we form the idea that, at every point in space around fixed charge q (l), there exists one of our little vector arrows. Thus we assign a vector field to the space: to every point of it, there is assigned both a magnitude and a direction.

However, note that this does not describe the situation as it exists, at any time, in the "present." That is, rigorously any one arrow only exists whenever a little test charged mass is present and accelerating radially away.

We may take the view that the arrow WILL exist WHEN a test charged mass is brought to that point and released. In that case, the arrow WILL exist in the FUTURE, not the present. And we may refer to that as, it POTENTIALLY exists, when and if.. ..etc.

So already we see a fundamental logical problem in the idea of such a field. It WILL exist in the future, but only POTENTIALLY exists in the present.

Of course I am not the first one to point out such difficulties; the fact that the field concept is erroneous is already well-known to foundations scientists. But it is such a USEFUL concept that it is retained.

Now the Ë-field, or electric field, is defined as "the accelerating force on a charged mass per unit charged mass accelerated by it." If there is no charged mass present to be accelerated, then there can be no Ë-field present.

So Ë-field cannot exist in vacuum, for vacuum presupposes the absence of all observable mass.

So we can rigorously say that no Ë-field of force exists around the charged mass q(l) IN THE PRESENT. At best, it can only POTENTIALLY exist IN THE FUTURE.

Next Slide

The Tom Bearden Website

SLIDE 7.

A SCALAR HAS A HIDDEN [VIRTUAL] VECTOR SUBSTRUCTURE

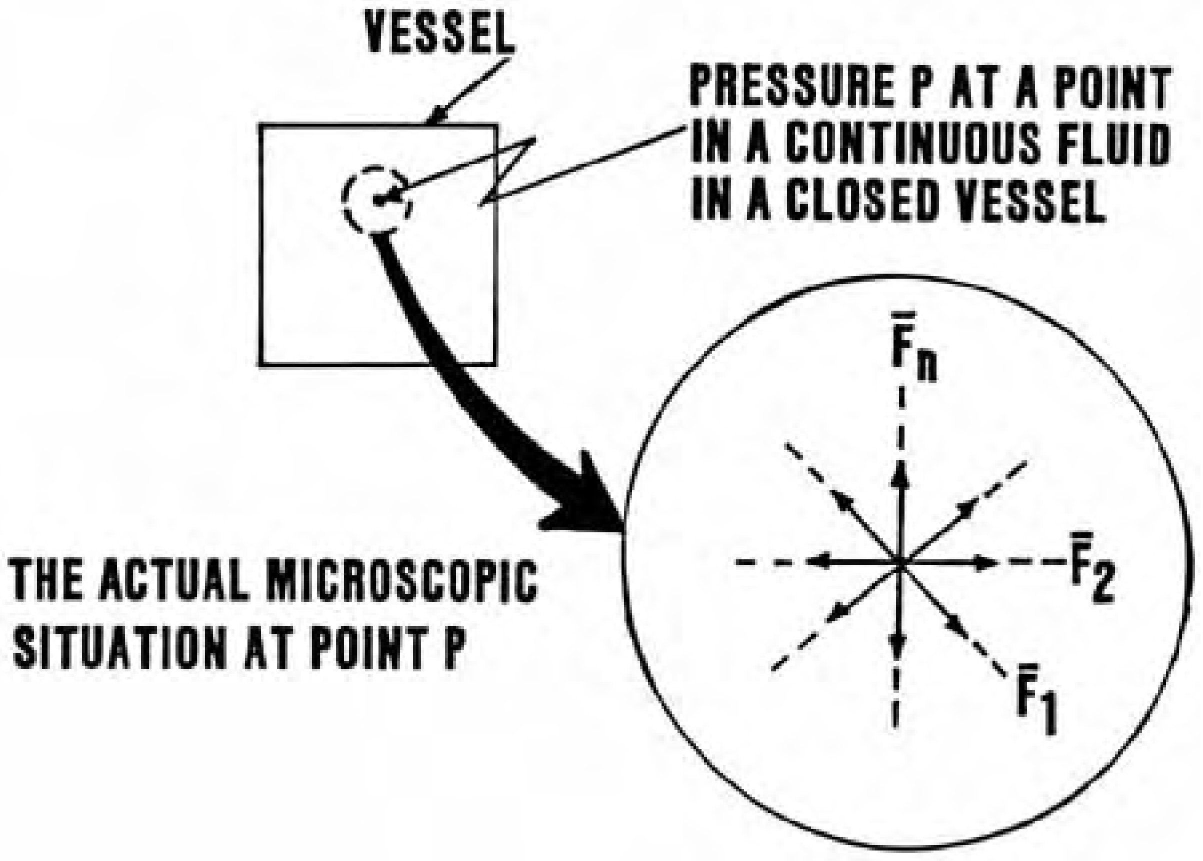
VESSEL
PRESSURE P AT A POINT IN A CONTINUOUS FLUID IN A CLOSED VESSEL
THE ACTUAL MICROSCOPIC SITUATION AT POINT P
$\bar{F}_n$
$\bar{F}_2$
$\bar{F}_1$

Our next slide shows that a physical scalar quantity has a hidden vector substructure.

For example, we show on the left a closed vessel containing a pressurized gas. The pressure at a point in the gas is a scalar quantity, since it has a magnitude but not a direction. However, physically what we have in the gas is a flux of individual molecules continually through any small volume, which we may let approach a point in the limit. Thus we have a flux in all directions through the point at which we specify pressure.

That is, the scalar "pressure" at a point may be said to contain a substructure of individual molecule flow vectors.

All physical process scalars have such hidden substructures. In the equilibrium case, the summation of all the individual flux vectors yields a resultant vector zero.

Thus we take the view that ANY PHYSICAL PROCESS SCALAR IS ALSO A ZERO VECTOR, AND ANY ZERO VECTOR IS A SCALAR AND CONTAINS A SUM-ZERO HIDDEN VECTOR SUBSTRUCTURE.

Next Slide

The Tom Bearden Website

SLIDE 8.

SCALAR AND VECTOR SUBSTRUCTURES

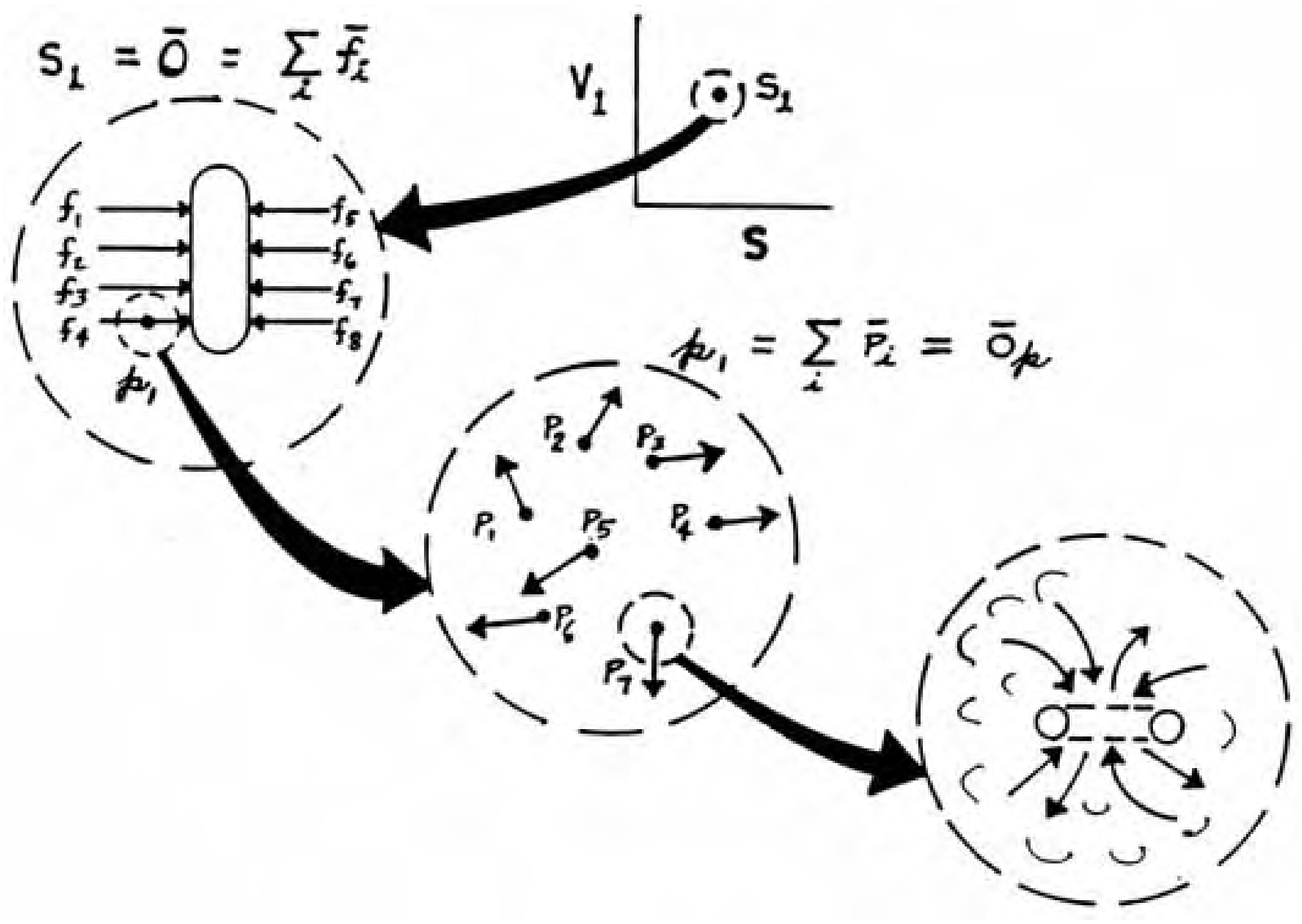
$S_1 = \bar{0} = \sum_i \bar{f}_i$
V_1
S_1
S
f_1
f_2
f_3
f_4
f_5
f_6
f_7
f_8
p_1
$p_1 = \sum_i \bar{P}_i = \bar{0}_p$
P_1
P_2
P_3
P_4
P_5
P_6
P_7

Now let us see if we can do some unusual thinking about the hidden vector substructure of a scalar or zero-vector.

On the leftmost figure in this slide, we show a plate with balanced opposing forces. The vector summation of all the forces' is zero, so the system has no resultant vector. The plate therefore does not move. However, the plate -- that is, the medium in which the forces act -- is stressed, much like the previous example of pressure in a gas. The difference is that now the stress is patterned in its substructure, and not random. Now suppose we rhythmically vary the forces, each the same, in a regular manner in time. Yet we maintain a continuous balance, so there is never a system resultant greater than zero. In this case, we maintain our vector zero, but we are producing regular stress waves in the plate medium..

I think you now can see that I can vary the individual opposing vector pairs so that rhythmic waves flow from top to bottom, or from bottom to top. That, of course, will describe a vector gradient traveling through the medium, and that will be an ordinary "vector wave." Is there any way I can make waves through the medium, and not have such vector waves ever appear?

Indeed there is, if I look further into the substructure. Take one point on one of the vectors shown acting on the plate in the left diagram. That point is a scalar, and it also has a substructure of even finer vectors, as shown in the lower middle figure. And any point in one of these second-order hidden vectors is a scalar and composed of a still finer hidden vector substructure, as shown in the lower right figure. This process repeats ad infinitum, and all is structures within structures. At least it can be modeled that way, for it already is treated that way in particle physics.

Now go back to the second order vector substructure. Here I can establish whatever patterns I wish, and not affect the first order pattern shown in the figure on the left. I can even rhythmically vary things in the second order substructure, and leave a blissfully unchanged first order substructure, to the first order observer.

So I can make as many waves; as I wish through the medium, as long as I stay second order or higher, and the external observer will never see any change at all unless he has a special detector of some sort to detect the lower substructure changes.

As I showed in the Appendix to my book, The Excalibur Briefing, these substructures are in reality the same thing as hyperspaces. And movements in them can proceed much faster than the speed of light.

So I can have higher order signals and waves through the stressed medium without changing the overall first order stress in the medium.

This is a sort of super-relativity; the present relativity is just a special case.

Superluminal communication systems are perfectly possible, but not with the present first order transmitters and receiver/detectors.

Note also that higher order stresses -- stresses in the higher order substructures -- represent special stresses in time, and time has an internal stress-pattern substructure. The higher order stress waves are phase waves or phase fluctuations in that respect.

Also note that Charles Musés has already created the hypernumber mathematics of the structure of time, and it is largely in his mathematics that these concepts will have to be mathematically modeled.

Next Slide

SLIDE 9.

SUPERPOSITION DOES NOT ELIMINATE THE VIRTUAL SUBSTRUCTURE

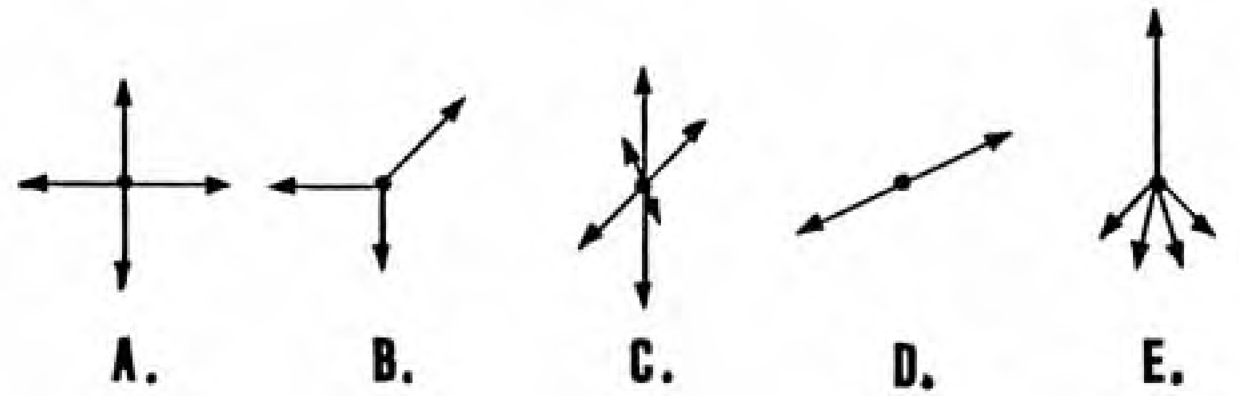

TWO OR MORE MACROSCOPICALLY "EQUAL" VECTORS OR SCALARS CAN DIFFER DRASTICALLY IN THEIR VIRTUAL SUBSTRUCTURE DISTRIBUTIONS. EVEN TWO "ZEROS" MAY BE DYNAMICALLY DIFFERENT DUE TO THEIR DIFFERING SUBSTRUCTURES.

Next we show some simple substructures of a zero vector. All of these vector systems sum to zero vectorially. Note that if we superpose two of these structures, all substructure vectors are still present., and the vector summation is still zero.

Also notice that, from the standpoint of substructuring, zeros can be and are drastically different.

We can even make rhythmic waves in the substructure by superposing patterned zero vectors in a regular fashion.

Next Slide

The Tom Bearden Website

SLIDE 10.

PHYSICAL VECTORS

- **VECTOR: A "SMEARED PARTICLE"**
- **FOUR MAJOR PARTICLES IN PHYSICS:**
  - POINT [MASSLESS, CHARGELESS]
  - MASS [CHARGELESS]
  - CHARGED MASS
  - CHARGED POINT [MASSLESS}
- **LEADS TO FOUR MAJOR VECTORS:**
  - GEOMETER'S
  - MECHANIC'S
  - ELECTRICIAN'S
  - ADVANCED ELECTRICIAN'S

In quantum mechanics, as we approach the quantum level, we find that the notions of a particle (static concept) and a wave become inseparable. That is, any particle becomes a "smear" in length and time, and it becomes inseparable from its smear. At the quantum level, there is no separate energy, length, mass, momentum, energy, or time -- all are welded together inextricably into a quantity called "action," having the units of angular momentum.

So at the fundamental quantum level, a moving point becomes a smear, and a vector represents a physical, undifferentiated entity.

To go quickly, we now define a vector as consisting of a smeared particle. That is, the particle is smeared in both time and length, and it is not differentiated from time or length.

There are four kinds of particles in physics which can be smeared to form vectors. First, there is a point in space or spacetime. This point may be approximated as massless and chargeless. At least that is the way the geometers of old conceived it in the abstract, and so the idea of a "point" in space assumes it to be chargeless and massless. Second, there is a tiny particle of mass, often abstracted to be a "point-mass," but visualized as having no charge. This is the mechanic's particle. Third, there is a tiny charged mass, such as an electron, and this is the electrician's particle. Again, it is often abstracted as a charged point-mass. Fourth, there is a charged point, having no mass, and this is the advanced electrician's particle, which he uses to form the idea of potentials and fields.

Smearing these four particles at the fundamental quantum level produces four different kinds of physical vectors: the geometer's, the mechanic's, the electrician's and the advanced electrician's vectors. They are all different.

Next Slide

The Tom Bearden Website

SLIDE 11.

THE FOUR VECTORS OF PHYSICS

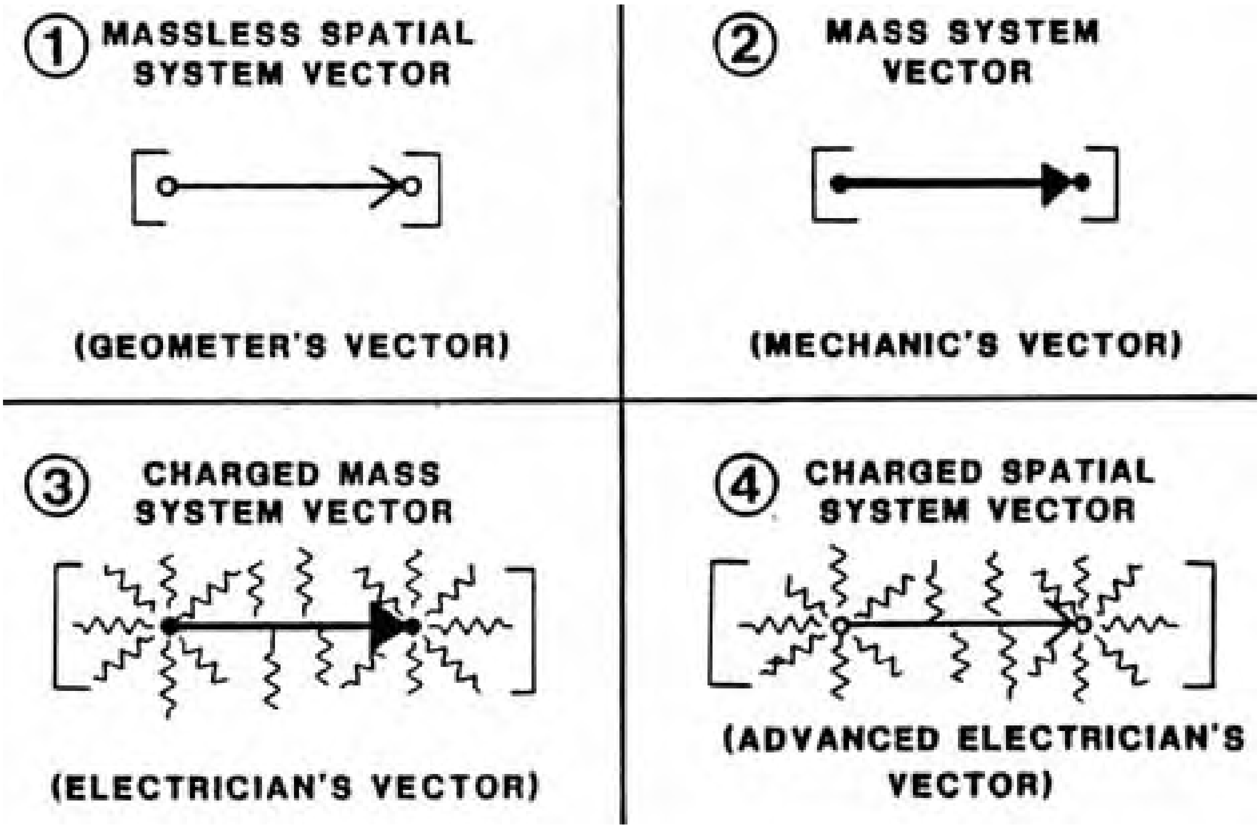
① MASSLESS SPATIAL SYSTEM VECTOR
(GEOMETER'S VECTOR)
② MASS SYSTEM VECTOR
(MECHANIC'S VECTOR)
③ CHARGED MASS SYSTEM VECTOR
(ELECTRICIAN'S VECTOR)
④ CHARGED SPATIAL SYSTEM VECTOR
(ADVANCED ELECTRICIAN'S VECTOR)

On this slide we show a graphic illustration of each of these vectors.

Note that they may be visualized as smeared particles, but the particle is not separate from the smear in time and length.

Each vector must be considered as a unitary entity -- it is a special kind of quantum, if you will.

Each is a special quantum of motion.

And Dewey Larson has it right when he says, "All is motion."

Next Slide

The Tom Bearden Website

SLIDE 12.

FORCE DEFINITION

- ***A TRUE DEFINITION IS AN IDENTITY.***

  $$F \equiv d/dt\,(mv)$$

- **FORCE CONSISTS OF THE TIME RATE OF CHANGE OF MOMENTUM**
- **MOMENTUM IS A MASS SMEARED THROUGH A LENGTH IN A CERTAIN TIME**
- **FORCE CONSISTS OF MASS-MOTION-CHANGE**
- ***FORCE IS AN EFFECT, NOT A CAUSE !***

We now look a little closer at the concept of a force.

I point out that a true definition is an identity. If a supposed definition of an entity is not an identity for that entity, then it is not a definition at all, but only a statement ABOUT the entity.

It may come as somewhat of a shock to the normal engineer and scientist that the foundations concepts of physics -- such as force, mass, energy, etc. -- are ALL in serious difficulty and contain many unresolved logical conflicts.

Force is in that kind of difficulty, and it has that kind of logical conflict.

In fact, in its force equations, physics has not even applied its own present definition of force.

As we show here, force is generally defined as the time rate of change of momentum.

If that is true, then as an identity the definition states that a force CONSISTS OF the time rate of change of momentum.

And of course, momentum is mass times velocity. Hence momentum is a mass smeared through a length and through a certain time.

Force then consists of the time rate of change of that smearing. That is, we smear the mass through time and length, faster or slower.

But if force CONSISTS OF that smearing change, it does not CAUSE it.

Force is thus an effect or result, not a cause.

This implies that there is a more fundamental mechanism that causes or generates force itself -- ANY force. If we can find that hidden mechanism, we may well find the long-sought unified field theory, since it will unify all forces on a common mechanism.

Next Slide

The Tom Bearden Website

SLIDE 13.

FORCE IS AN EFFECT NOT A CAUSE

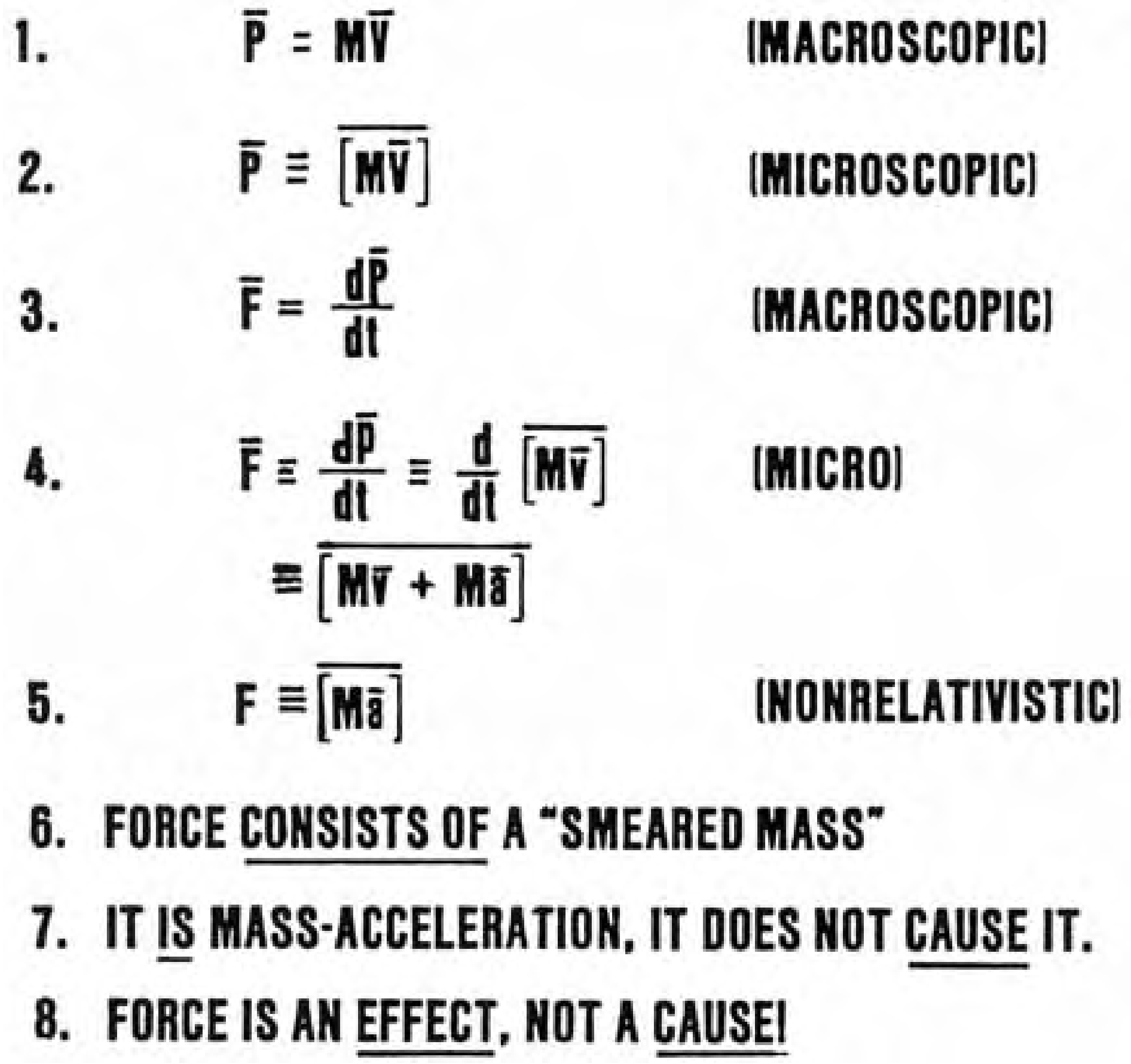

1. $\bar{P} = M\bar{V}$ (MACROSCOPIC)

2. $\bar{P} \equiv \overline{[M\bar{V}]}$ (MICROSCOPIC)

3. $\bar{F} = \frac{d\bar{P}}{dt}$ (MACROSCOPIC)

4. $\bar{F} \equiv \frac{d\bar{P}}{dt} \equiv \frac{d}{dt}\overline{[M\bar{v}]}$ (MICRO)

   $\equiv \overline{[M\bar{v} + M\bar{a}]}$

5. $F \equiv \overline{[M\bar{a}]}$ (NONRELATIVISTIC)

6. FORCE CONSISTS OF A "SMEARED MASS"

7. IT IS MASS-ACCELERATION, IT DOES NOT CAUSE IT.

8. FORCE IS AN EFFECT, NOT A CAUSE!

On this slide, we show briefly how one might express this idea in mathematical symbology.

Macroscopically, momentum consists of mass times velocity. Microscopically, the mass and the velocity are inextricable, and it is not proper to speak of a "mass in motion," but only of "mass-motion." We show this in equation 2.

We take the macro definition of momentum, as shown in equation 3, and express it for the microscopic, unseparated case, as shown in equation 4.

In the nonrelativistic case, we may consider m-dot to be zero, since the time rate of change of mass is almost zero. In that case, force may be defined for the microscopic case as a smeared accelerating mass, as shown in equation 5.

Force thus consists of the welded mass and acceleration, with no seam in the middle.

It IS mass-acceleration; it does not CAUSE it. Therefore force is an effect, not a cause.

Next Slide

The Tom Bearden Website

SLIDE. 14

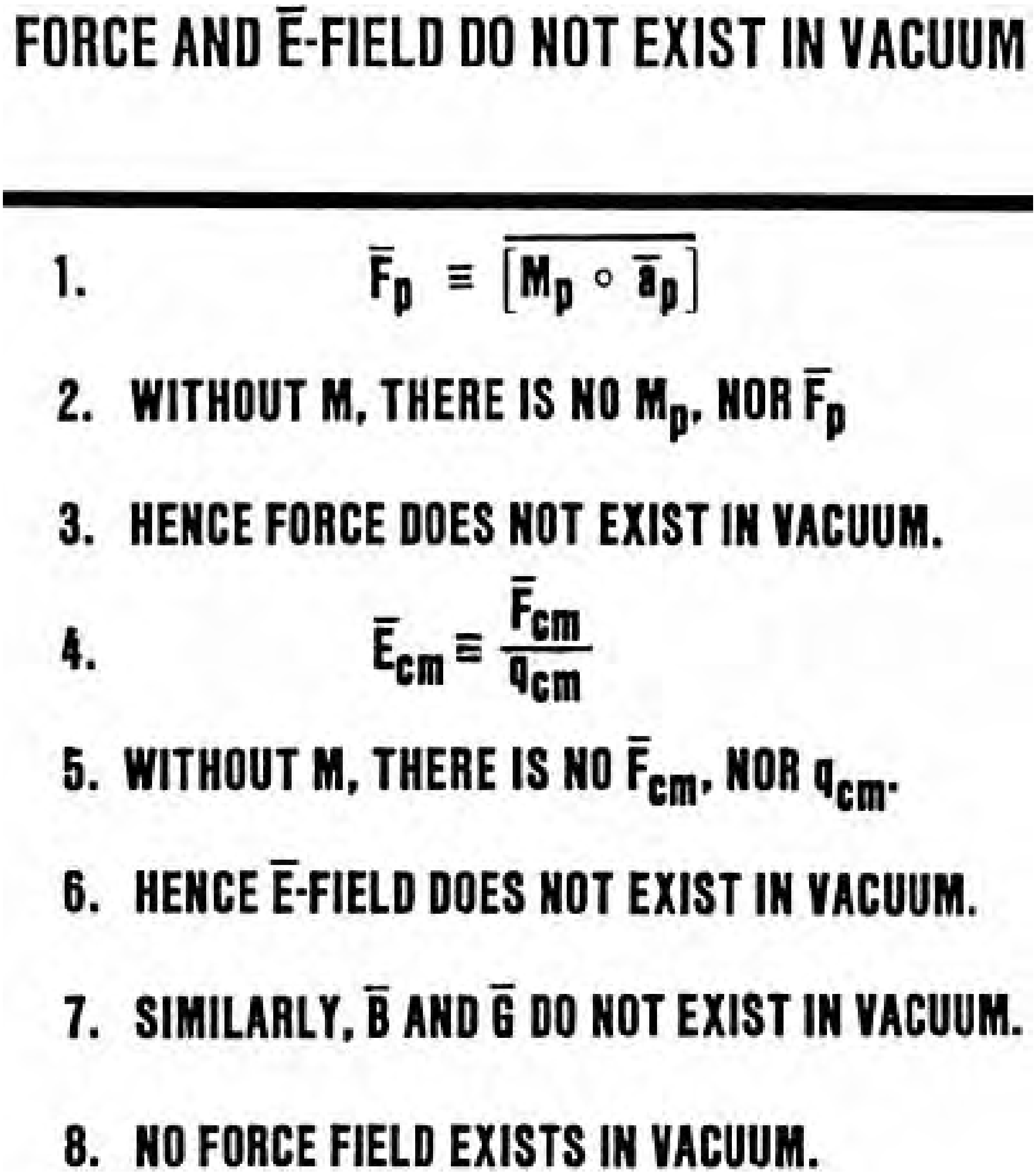

FORCE AND $\bar{E}$-FIELD DO NOT EXIST IN VACUUM

1. $$\bar{F}_p \equiv \overline{[M_p \circ \bar{a}_p]}$$

2. WITHOUT M, THERE IS NO M_p, NOR $\bar{F}_p$

3. HENCE FORCE DOES NOT EXIST IN VACUUM.

4. $$\bar{E}_{cm} \equiv \frac{\bar{F}_{cm}}{q_{cm}}$$

5. WITHOUT M, THERE IS NO $\bar{F}_{cm}$, NOR q_{cm}.

6. HENCE $\bar{E}$-FIELD DOES NOT EXIST IN VACUUM.

7. SIMILARLY, $\bar{B}$ AND $\bar{G}$ DO NOT EXIST IN VACUUM.

8. NO FORCE FIELD EXISTS IN VACUUM.

But if force CONSISTS OF mass welded to acceleration, it cannot exist in the absence of mass.

Hence force does not exist in vacuum.

Further, the Ë-field (Electric field) of a charged particle is defined as the force on a charged mass, divided by the mass. That is, the Ë-field is thought to represent the force per unit charged mass. And then it is assumed to exist in vacuum.

But vacuum is without mass or charged mass, observably. Without mass, neither force nor charged mass exists.

Therefore the Ë-field does not exist in the vacuum.

Magnetic field (B) and gravitational field (G) are similarly defined, and they have the same logical difficulties.

Therefore neither magnetic field nor gravitational field exist in vacuum.

A priori, NO force field exists in vacuum.

The greatest error in physics has been the assignment of a force as a cause, and thus assigning it to the vacuum. The force is an effect, and it never exists in vacuum, a priori.

Next Slide

The Tom Bearden
Website

SLIDE 15.

COROLLARIES

- MORE PRIMARY MECHANISM CAUSES/GENERATES FORCE -- ANY FORCE
- FORCE DOES NOT EXIST IN VACUUM, SINCE VACUUM IS THE ABSENCE OF MASS
- FORCE FIELDS DO NOT EXIST IN VACUUM
    - E-FIELD
    - B-FIELD
- TRANSVERSE [HERTZIAN] WAVES DO NOT EXIST IN VACUUM

Without belaboring the issue, I now point out and summarize some of these conclusions.

First, a more primary mechanism causes or generates force -- ANY force.

Second, force implies the presence of mass a priori. Force does not exist in vacuum, since vacuum is the absence of mass. _

Third, no force fields of any kind exist in vacuum, including Ë-field, B-field, and gravitational field. _

Fourth, since Hertzian waves are defined in terms of Ë and H force fields, and these fields do not exist in vacuum, then Hertzian waves do not exist in vacuum.

Tesla pointed out long ago that vacuum EM waves are longitudinal waves, not Hertz waves. Vacuum EM waves are similar to sound waves in a gas.

The Hertz wave exists as a matter wave in the electron gas in a transmitting antenna, and in the electron gas in a receiving antenna, but does not exist in the vacuum in between the two antennas. Between them, EM waves are longitudinal waves.

Next Slide

The Tom Bearden
Website

SLIDE 16.

ENERGY IS DERIVED, NOT FUNDAMENTAL

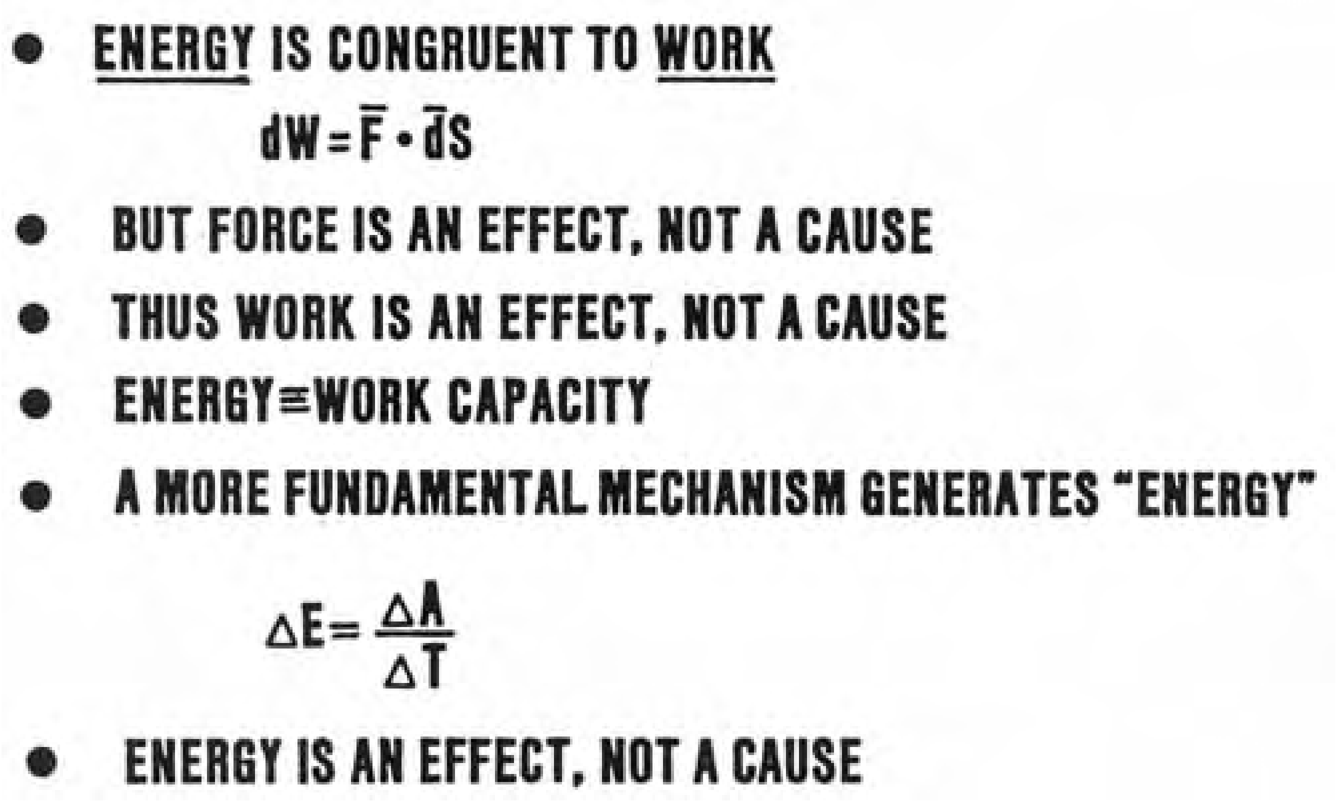

- **ENERGY** IS CONGRUENT TO **WORK**

$$dW = \bar{F} \cdot \bar{d}S$$

- BUT FORCE IS AN EFFECT, NOT A CAUSE
- THUS WORK IS AN EFFECT, NOT A CAUSE
- ENERGY≅WORK CAPACITY
- A MORE FUNDAMENTAL MECHANISM GENERATES "ENERGY"

$$\Delta E = \frac{\Delta A}{\Delta T}$$

- ENERGY IS AN EFFECT, NOT A CAUSE

Energy is usually taken as "the capacity to do work" at some future time. In other words, when we refer to the energy of something, we refer to HOW MUCH WORK IT CAN DO IN THE FUTURE, IF AND WHEN WE ARRANGE FOR IT TO DO IT. Rigorously, a body has no "work" at a particular instant in time, so we use "energy" to refer to potential work that can, or will, be done in the future. Thus energy is always congruent to work in the future.

And for that reason, the units of energy are the units of work; the only difference is in the time the work is done, relative to the "now" or "present" we are speaking of.

Energy is the potential for work to be done in the future.

But what is work?

Well, work is defined in terms of the movement of a force through a distance. And it is understood that only the projection of the force along the line of movement is "working." Thus work is defined differentially as shown on the slide. The differential of work is equal to the dot product of the moving force times the differential of displacement.

But as we have seen, force is an effect, not a cause.

That is, when we calculate the work done, it's always in the past, never in the present or future. It's always been DONE, and has been expended.

Being in the past, it's gone forever, so it certainly isn't the CAUSE of anything which happens in the future.

Thus work is an effect, not a cause.

That's interesting, because then it certainly does not directly generate "energy," since energy is something in the present which speaks of future work to be done.

Rigorously, the fundamental quantum -- which is what exists in the present, in unseparated and unquantized form -- consists of action. NOT energy, NOT time, NOT length, NOT momentum exclusively, but all of these INclusively.

And a change in energy occurs when a quantum is separated at a certain rate in time. Thus delta energy equals the time rate of change of action, or delta action divided by the delta time used in separating the quantum and fissioning it into parts (collapsing the wave function).

Thus energy itself is an effect or result of a more fundamental mechanism, and it is not a cause.

In quantum mechanics all this has been neatly "buried" in the mysterious "collapse of the wave function," which is supposedly what "causes" everything to happen. And then we swear that what results after the collapse is a totally statistical selection from the possible states of the system we are examining. We thus say that the microworld is totally statistical at base. Of course we also build a wave equation or wave function which propagates forward in time with rigorous causality, and that wave or function absolutely controls the basket of "possible states" allowed .

All that says is that the "wave function times time" is what exists in the present, when observation has not been made by "stopping time." And then we "stop time" in a rather blunt manner, somewhat like firing a billiard ball blindly into a group of billiard balls on a table, and we live with what gets hit and pops out.

If we examine this "time" that is flowing, look into its substructure, and arrange that substructure as we will, we can control which "billiard ball" is hit and which pocket it goes into. By working on the vector substructures of time, we can thus deliberately control the microworld and make it causal, not statistical. Quantum mechanics is in error in stopping where it has; it is presently only a first order approximation of the real quantum mechanics that awaits.

Einstein was correct after all; God does NOT play dice with the universe when one considers all levels.

One finds oneself in a statistical situation at one level only when the constituent sublevels of it are uncontrolled.

Next Chapter

The Tom Bearden Website

SLIDE 17.

EXISTENCE IN, AND PASSAGE THROUGH, TIME

- **IN RELATIVITY, TRANSLATION OF A BODY AND PASSAGE THROUGH TIME ARE RELATED**

  $$\Delta L = C \Delta T$$

- **"MASS" IS SPATIAL**
  - **DOES NOT EXIST IN TIME**
  - **"MASS-TIME" EXISTS IN TIME**

- **VIRTUAL FLUX TO AND FROM A BARE PARTICLE DETERMINES ITS "PASSAGE THRU TIME"**

- **SPHERICAL SPIN OF A MASS PARTICLE [L^3] COUPLES IT TO THE SPACETIME VACUUM [L^3T] VIRTUAL PARTICLE FLUX**

- **PASSAGE THROUGH ORDINARY TIME IS CONTROLLED BY CREATION AND ANNIHILATION OF QUANTA**

On this slide, we simply point out that translation of a body through space and its passage through time are related, according to relativity.

Now "mass" is spatial.

It does not exist in time.

Only things which contain the unit of time can "exist in time."

Otherwise, the thing only exists as such when the passage of time has been momentarily stopped -- usually by photon emission and the consequent carrying off of the time unit by the photon.

Since every mass that has temperature -- and every mass does -- is continually absorbing and emitting photons, then the mass is continually grabbing a time tail (absorption) and losing it (emission) shortly thereafter.

Masstime is what exists in time, not mass.

And every mass particle is continually turning from mass to masstime to mass, etc.

Indeed, it is the summation of the absorption and emission of virtual subquanta by a mass particle -- i.e., its "charge" **--** that determines its rate of flow through "time."

And the spherical spin of a mass particle couples the particle of mass to the spacetime vacuum; that is, to the virtual particle flux.

Passage through ordinary time is controlled by creation and annihilation of quanta, and these processes are controlled by substructure (subquantal) actions and interactions.

Next Slide

The Tom Bearden Website

SLIDE 18.

SOME IMPLICATIONS

- **MODIFIES CONSERVATION OF CHARGE**
- **MODIFIES CONSERVATION OF ENERGY**
    - ***MASS***
    - ***ENERGY***
    - ***ANENERGY (MASSLESS CHARGE)***
- **MODIFIES NEWTON'S LAW**
    - ***FOR EVERY ACTION, THERE IS AN OPPOSITE AND EQUAL REACTION***
    - ***REACTION NEED NOT BE LOCAL***
    - ***REACTION NEED NOT BE ANTIPARALLEL***
- **UNIFIED FIELD THEORY**

Without belaboring the issue, I now simply point out some implications of the new approach.

First, we modify two laws -- the conservation of charge and the conservation of energy -- by combining them into one law, a new conservation law in which anenergy is conserved.

Neither charge per se nor energy per se need be conserved; but the overall anenergy (massless charge) equivalency is what is conserved. This allows us to derive mechanisms for free energy from the vacuum, for example. These mechanisms readily violate the present conservation of energy law, but do not violate the overall conservation of anenergy law.

It is the total anenergy equivalency of mass, energy, and massless charge that is now to be conserved in general. The old laws are now just special cases.

Also, Newton's third law -- for every action there is an equal and opposite reaction -- is modified. It still applies, but in changed form.

The reaction still occurs, but it need not be local, and it need not be antiparallel to the original action.

Thus it is now possible to design and build a "locally reactionless" inertial space drive for a vehicle.

The final implication is that the common mechanism producing force -- any force -- provides a unified field theory. Specifically, it provides one which easily lends itself to direct engineering. It is not esoteric nor mysterious. It is practical and physical, and it can be done with some modifications to our present electrical circuits and devices.

Next Slide

The Tom Bearden
Website

SLIDE 19.

CONSERVATION OF ANENERGY (AE): (AN EXPANDED ENERGY CONSERVATION LAW)

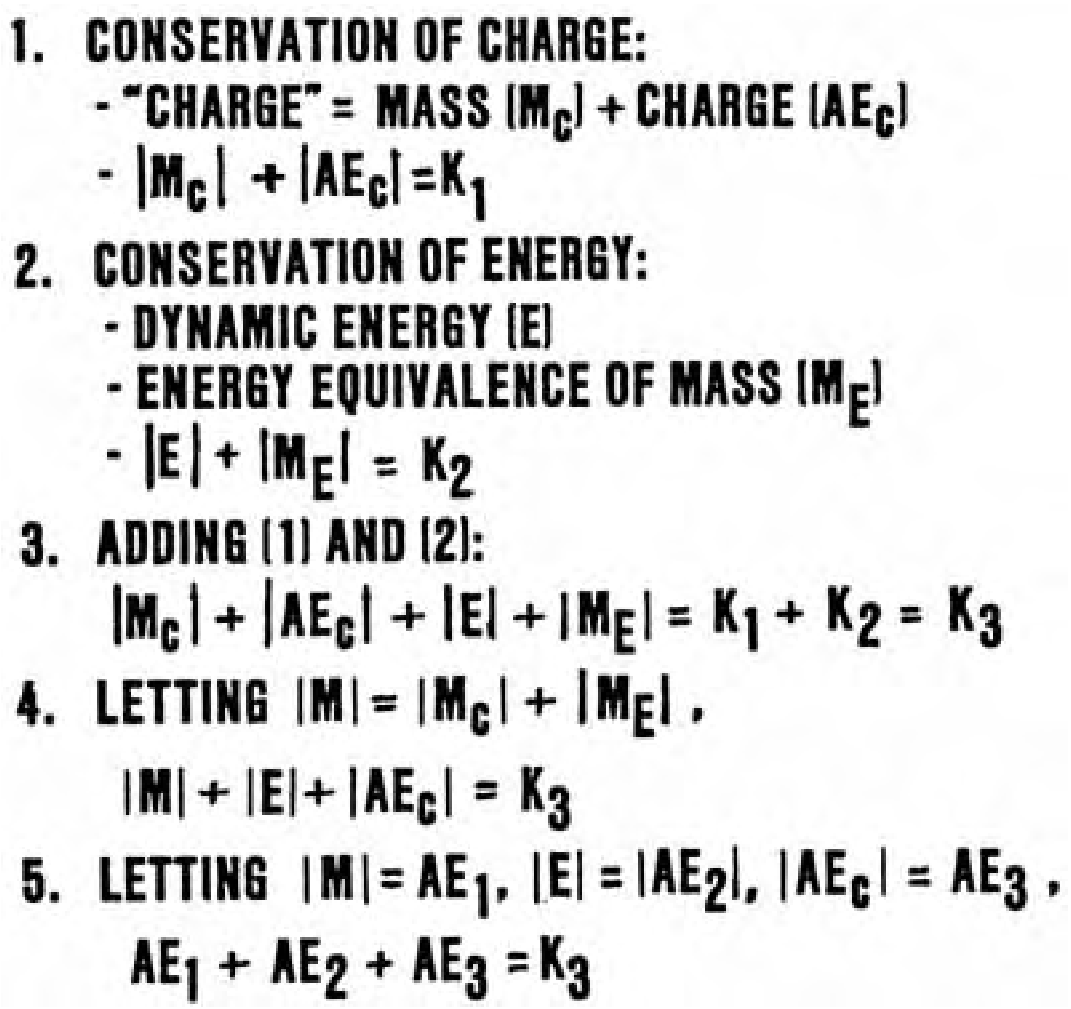

1. CONSERVATION OF CHARGE:
   - "CHARGE" = MASS (M_C) + CHARGE (AE_C)
   - $|M_C| + |AE_C| = K_1$
2. CONSERVATION OF ENERGY:
   - DYNAMIC ENERGY (E)
   - ENERGY EQUIVALENCE OF MASS (M_E)
   - $|E| + |M_E| = K_2$
3. ADDING (1) AND (2):

   $|M_C| + |AE_C| + |E| + |M_E| = K_1 + K_2 = K_3$
4. LETTING $|M| = |M_C| + |M_E|$,

   $|M| + |E| + |AE_C| = K_3$
5. LETTING $|M| = AE_1$, $|E| = |AE_2|$, $|AE_C| = AE_3$,

   $AE_1 + AE_2 + AE_3 = K_3$

We briefly show a simplified derivation of the new conservation of anenergy law.

We express the conservation of charge law by correcting the term "charge" to show its true meaning: charged mass. We use the absolute value symbol to mean the "anenergy equivalency."

As shown on the slide, we account separately :for the mass (m-sub-c) of the conserved charges and the virtual particle charge flux (the anenergy -- AE-sub-c) of the charges conserved. That law now says that the absolute values of the masses and the massless charge fluxes have a total constant summation.

For a dynamic mass, we do a similar thing, in this simplified case ignoring potential energy of position, and taking the energy equivalence of the dynamic energy (M-sub-E) and the rest mass energy of the particles, M-sub-E. The conservation of energy now states that these two terms have a constant absolute value summation.

We add equations 1 and 2, getting equation 3. We consolidate like terms to get equation 4.

Expressing all terms in their anenergy equivalencies (absolute values), we get equation 5, which is the overall conservation of anenergy equivalency.

Note that we now can change mass into energy or anenergy (massless charge), energy into mass or anenergy, and anenergy into mass or energy.

Thus we can turn everything into pure vacuum ether (pure massless charge, pure anenergy, pure electrostatic scalar potential, pure stress on spacetime) or we can turn part of the vacuum into energy or particles of mass.

Vacuum now becomes a very real thing indeed.

It is pure anenergy.

Pure massless charge, pure electrostatic scalar potential, pure "broken bits" (subquantal bits) of (quantized) energy.

The vacuum is identically the virtual state.

It is not an emptiness filled with virtual things; rather, it IS the sum total of those virtual things, for they are totally unseparated until some sort of "conceptual observation" is invoked.

Physics should have adopted Einstein's suggestion that the vacuum, complete with its EM and gravitational fields, be called the ether.

Next Slide

The Tom Bearden Website

SLIDE 20.

FREE ENERGY IS ALLOWED

- **ANY ONE-WAY GATING PROCESS FROM ANENERGY TO ENERGY YIELDS "FREE ENERGY" SINCE THE ANENERGY IS AUTOMATICALLY REPLENISHED BY THE UNIVERSE.**

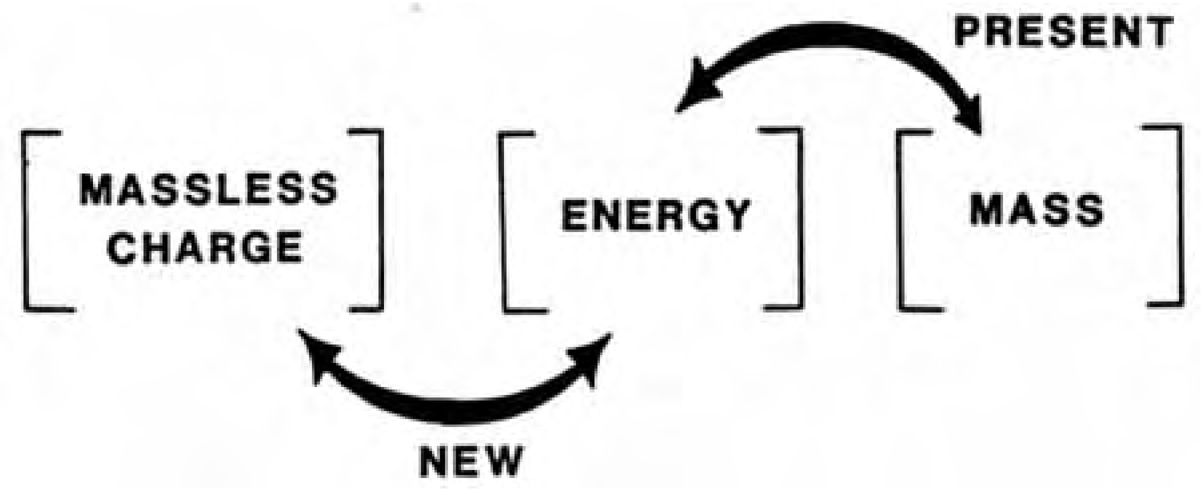
PRESENT
MASSLESS CHARGE
ENERGY
MASS
NEW

Now we can put the search for free energy devices on a rigorous scientific basis.

As shown here, mass and energy are translatable one into the other. So are massless charge and energy.

And the universe is filled with fantastic amounts of anenergy; ultimately, that is what the universal vacuum IS.

Any one-way gating process from anenergy to energy yields "free energy" since the anenergy is automatically replenished by the universe.

A sort of Maxwell's demon, after all, is what we need. Only we need one which works with the virtual state, not observable molecules.

Fortunately nature has provided several Maxwell's demons of this sort, if we but recognize them and learn to use them.

Next Slide

The Tom Bearden
Website

SLIDE 21.

SCALAR POTENTIAL CAN FORCE THE RELATIVISTIC CASE

- FOR SCALAR POTENTIAL ENERGY
  OF APPRECIABLE SIZE RELATIVE TO
  A PARTICLE'S REST ENERGY.
    - NEWTONIAN MECHANICS AND
    - THE SCHRODINGER EQUATION
- MAY BE INADEQUATE
    - EVEN IF V/C IS SMALL

Bloch & Crater, "Lorentz-invariant potentials and the non-relativistic limit," American Journal of Physics, Vol. 49, No. 1, Jan. 1981, pp. 67-75

First, let us now digress to point out that one does not have to have relativistic velocity to obtain relativistic effects.

One can use the common electrostatic scalar potential to drive the situation relativistic. On this slide we show just one of several references in the standard literature that address this fact.

Let's understand what we are saying.

Anything you get from a relativistic situation, you can get directly by cleverly applying electrostatic scalar potential.

You can get a change in the passage through time, you can get energy changes, mass changes, inertial resistance changes, etc.

You can bend, warp, and twist spacetime like a pretzel.

If you "wave" the scalar potential by simply varying it , you can create pure time waves. You can also produce pure inertial field waves, pure gravitational waves, etc.

The ability to engineer the curving of spacetime allows the direct ability to engineer physical reality itself.

Next Slide

The Tom Bearden Website

SLIDE 22.

SCALAR STRESS WAVE
[$\overline{\text{O}}$-WAVE] [TESLA WAVE]

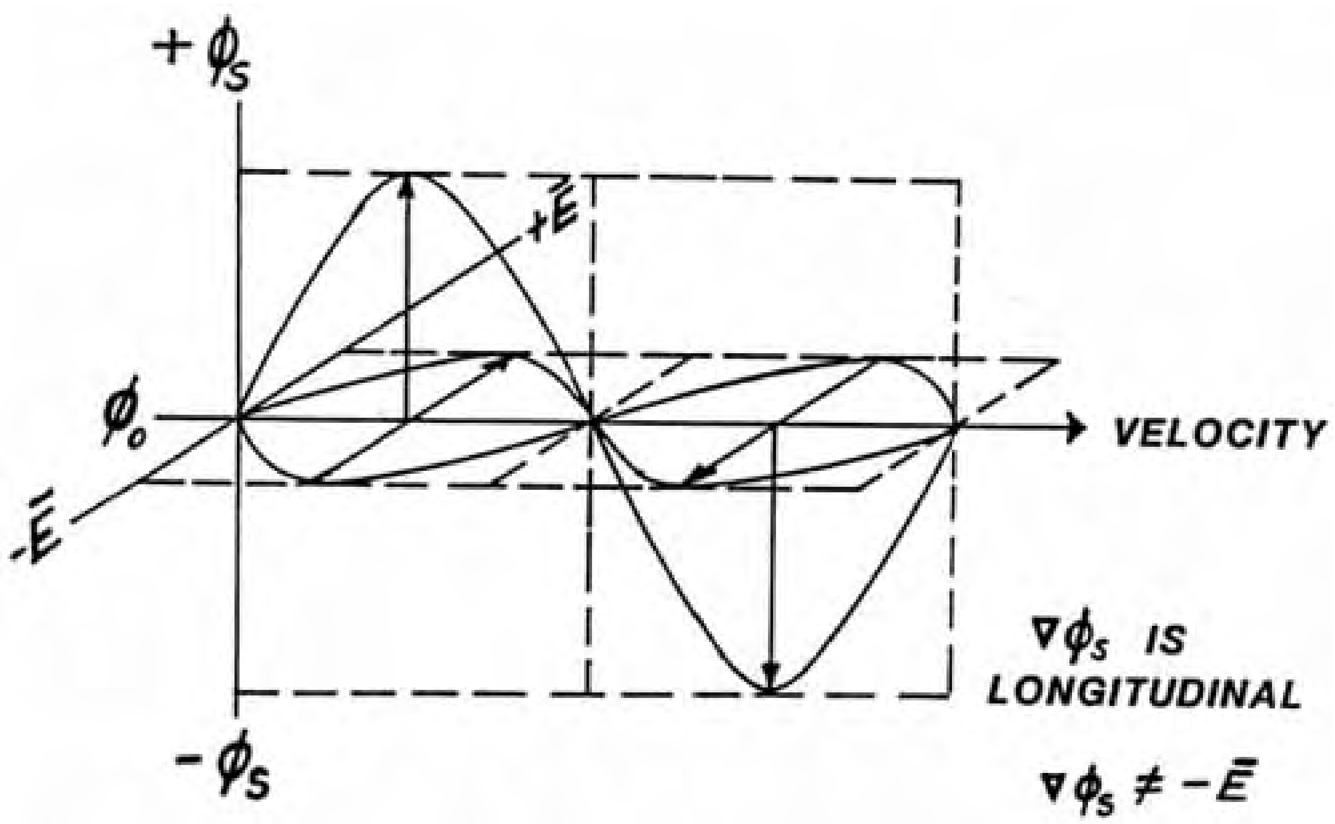
$+\phi_S$
$+\bar{E}$
ϕ_o
VELOCITY
$-\bar{E}$
$\nabla\phi_S$ IS LONGITUDINAL
$-\phi_S$
$\nabla\phi_S \neq -\bar{E}$

On this slide we now show a simple way to make a scalar, zero-vector wave -- the kind of wave originally discovered by Nikola Tesla.

It's simple. We just believe that a sum-zero vector substructure makes a scalar quantity, and we MAKE some scalars that way.

We also understand that a zero-gradient of a scalar is a zero vector, so that the scalar itself may be taken to be a zero vector.

The simplest explanation of this wave is as follows:

First, in physics we have two competing, mutually exclusive theories as to the nature of electromagnetic energy: the wave theory and the particle theory. Physicists argued for decades over these theories, for some experiments support one and some support the other. They never solved the problem; they just agreed to quit arguing. They formulated the "duality" principle to allow the saving of face to both sides.

Briefly, the duality principle implies that, whatever the nature of electromagnetic energy is before an interaction, in the interaction you can get it to act as a wave or as a particle. In other words, AS IT EXISTS, BEFORE THE INTERACTION, it is implicitly both particle and wave, joined together in some fashion, without being explicitly either one.

With the fourth law of logic, this becomes perfectly clear. With three-law Aristotlean logic, the problem is unresolvable.

Let us use this idea of "explicit duality without implicit duality" to analyze the wave shown on the slide.

First, from a wave aspect, the E-fields and the B-fields of the two waves do superpose and vectorially add. Since the waves are 180 degrees out of phase, the exterior resultant wave has a zero electric field and a zero magnetic field. Therefore it is a "zero-vector" wave, or "scalar" wave. It's a wave of pure stress in spacetime.

However, this scalar wave has a precisely determined substructure, consisting_of two ordinary sine waves, each of which comprises an ordinary E-H vector EM wave.

Now we apply the photon consideration (remember, before we interact with the wave, it must implicitly possess BOTH wave and particle natures combined, and we have so far only examined the implication of the wave nature.

The theory of photons' states that they are monocular critters. Photons pass right through other photons without interaction, in a linear situation. Therefore they can coexist without interaction, which is what we show here.

One photon, by the way, is one wavelength .

The photon theory requires that both substructure waves continue to exist as independent photons. Therefore we are assured that our substructure is intact.

However, notice that the totality of the two waves stresses spacetime. In other words, we have twice the stress on spacetime now as we would have from either wave separately.

This wave is therefore just a pure stress wave in spacetime itself.

This thing oscillates time, oscillates the relativity of the situation, and can affect energy, time flow rate, inertia, gravity, etc. aspects of an absorbing system.

Note that we have a rhythmic oscillation in phi (Ø), and we have a longitudinal stress wave, very similar to a sound wave. The MEDIUM for this wave is the virtual particle flux that identically comprises vacuum spacetime itself.

Next Slide

The Tom Bearden Website

SLIDE 23.

SCALAR Ö-WAVE PRODUCTION

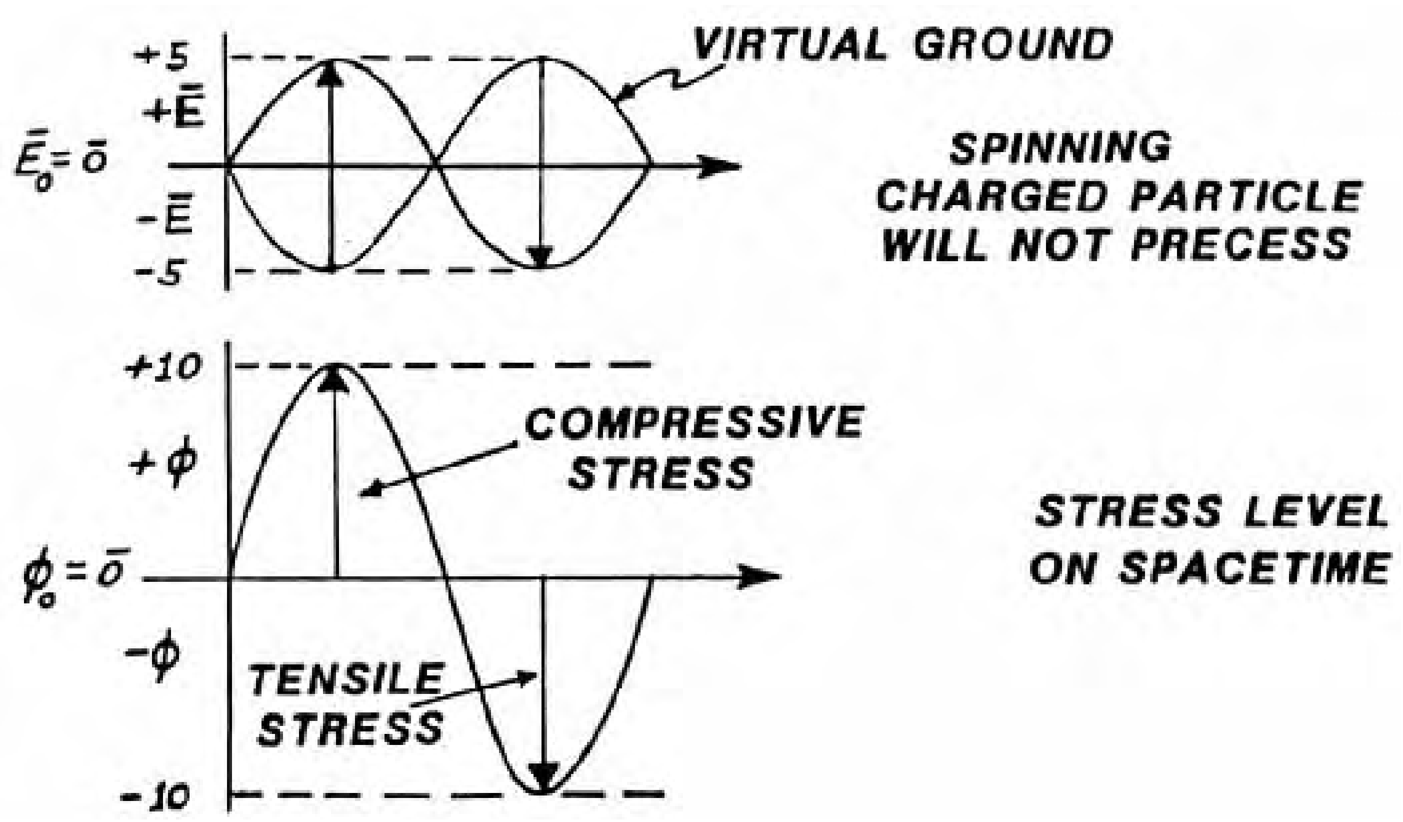
+5
+$\bar{E}$
$\bar{E}_o = \bar{o}$
-$\bar{E}$
-5
VIRTUAL GROUND
SPINNING
CHARGED PARTICLE
WILL NOT PRECESS
+10
+ϕ
$\phi_o = \bar{o}$
-ϕ
-10
COMPRESSIVE
STRESS
TENSILE
STRESS
STRESS LEVEL
ON SPACETIME

On this slide we show how to regard the magnitude of the stress, by using one wave envelope as "virtual ground." In the top diagram, the original E-field stress of each wave has a magnitude of 5, and the composite stress wave now has a magnitude of 10, in whatever units we choose to express them.

We show in the bottom diagram that during one half cycle we have compressive stress in the virtual particle vacuum flux, and during the second half cycle we have tensile stress.

This shows the wave is like a sound wave in the gaseous molecules of the air.

However, this wave has one difference. It also oscillates time, and thus has at least one additional degree of freedom, compared to ordinary EM waves.

In fact, this wave can be made n-dimensional and hyperspatial.

As a first order approximation, we can treat such a wave in a spatial fashion, if the wave is not too great in magnitude and the relativistic oscillation of time and inertia is not too large.

Next Slide

[The Tom Bearden Website](#)

SLIDE 24.

A SIMPLIFIED CHARGED PARTICLE

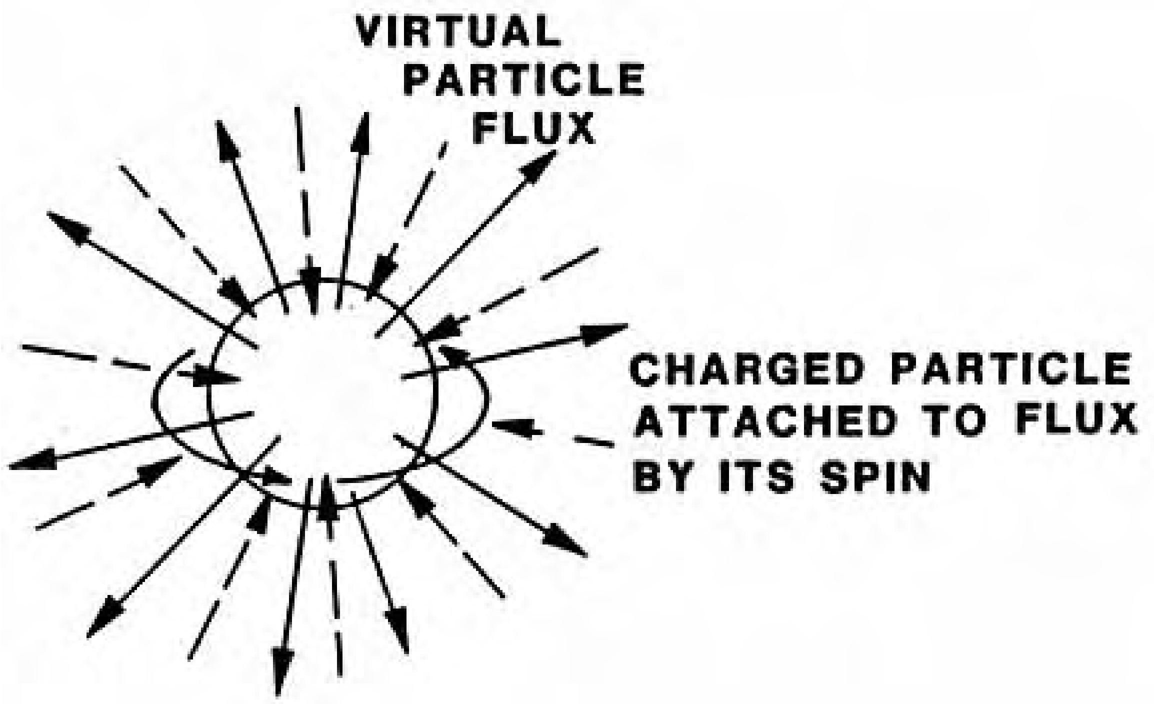

(THINK OF THE FLUX IN AND OUT, EXCEPT GRADIENTS ARE SPIRAL-LIKE, DUE TO PARTICLE'S SPIN)

We now digress slightly to pick up some more concepts.

On this slide, we show a simplified physical model of a charged particle. According to present particle physics, we
model it as an observable, spinning bare mass in the middle, connected to a spray of virtual particles on and off the mass in all directions.

It is the spin of the mass that "attaches" the spray of virtual particles to and from itself.

Note that, because of the spin, the gradient lines of flux will actually be spiral or vortex in nature.

Also note that, when EM theory was founded, no one knew that an electron was spinning. Electron spin was therefore omitted from the foundations of present electromagnetic theory, and compensated for later by strictly ad hoc measures.

Next Slide

The Tom Bearden Website

SLIDE 25.

A CHARGED PARTICLE IS A SPECIAL KIND OF "SPRAY NOZZLE"

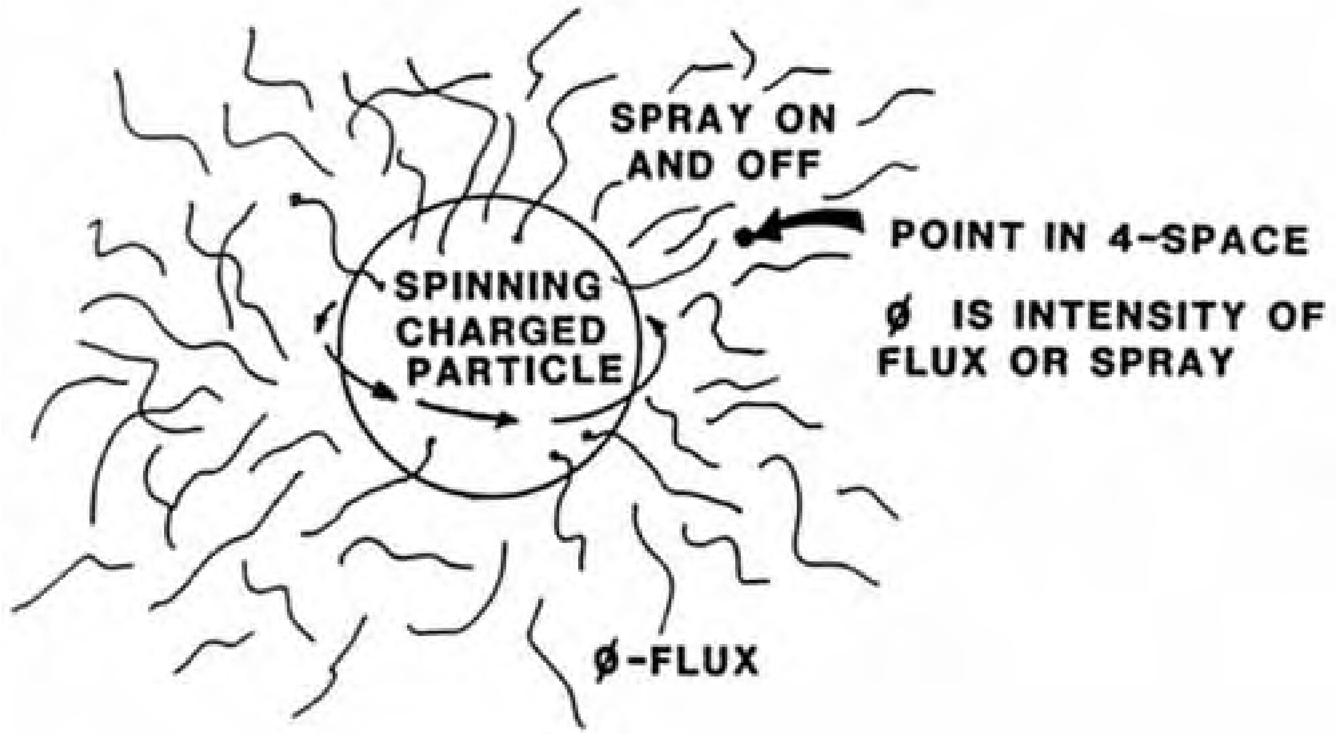
SPRAY ON AND OFF
POINT IN 4-SPACE
SPINNING CHARGED PARTICLE
ø IS INTENSITY OF FLUX OR SPRAY
ø-FLUX

(SPRAY IS THROUGH 4 DIMENSIONS, NOT JUST 3)

Here we show a very important concept, and a very important fact of nature.

A charged particle may be idealized as a special kind of spray nozzle.

The 3-dimensional particle of mass is spraying out virtual particle fluid, and sucking in virtual particle fluid, in four or more dimensions.

We never have to furnish or replenish the spray. Nature provides that. And it's inexhaustible. It comes from the entire virtual particle flux vacuum of the universe. It's already connected firmly to matter, by its connection to the spinning charged particle of mass.

I repeat and strongly stress: This tremendous, continuous spray of virtual particles is absolutely free, and absolutely furnished by the universe.

We don't have to create the basic pumps for the fluid. We just move the spray nozzles around, by moving the spinning charged masses around.

We never furnish any ultimate energy to or in our present EM circuits. We just pump around the spray nozzles, gating the spray furnished by the universe.

Next Slide

The Tom Bearden Website

SLIDE 26.

AN OBSERVABLE CHARGED PARTICLE IS DIRECTLY CONNECTED TO VIRTUAL PARTICLE FLUX VACUUM

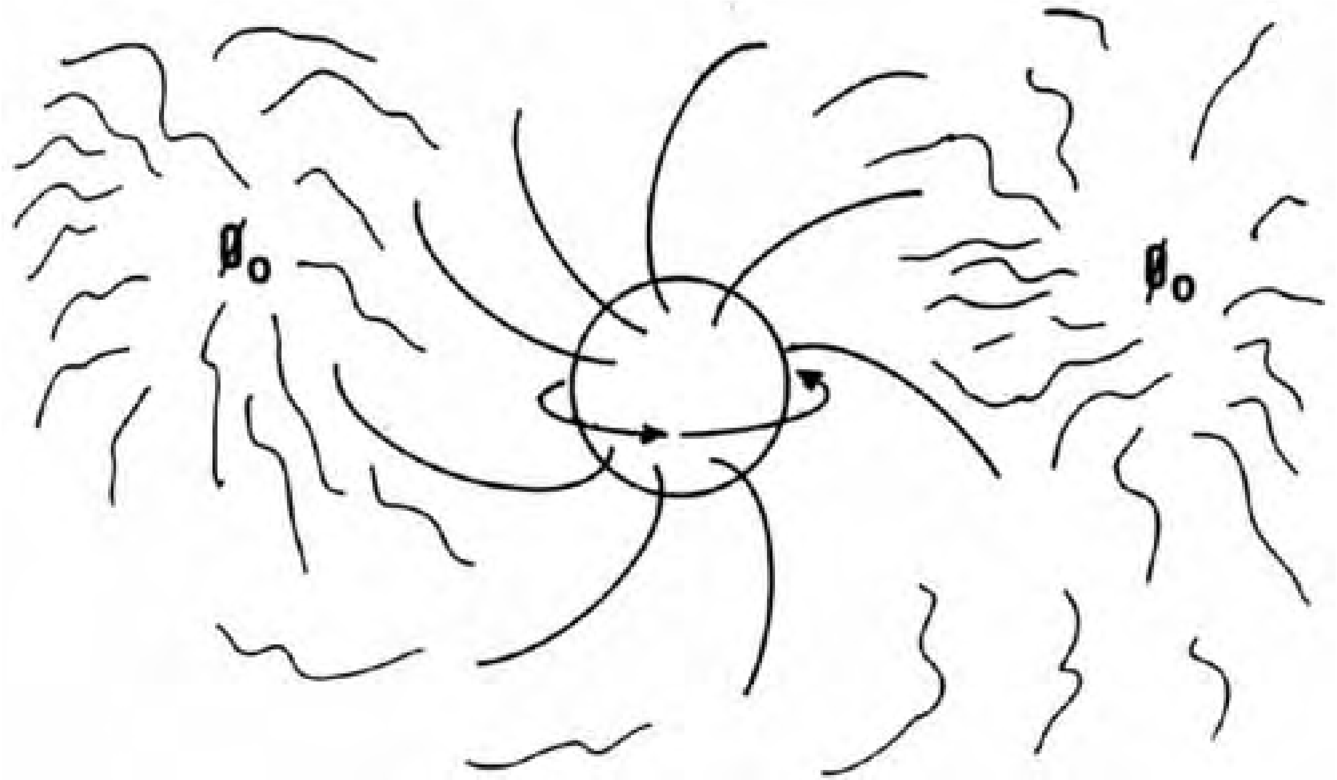

On this slide we accent that our 3-dimensional particle of mass is directly connected by its spin to the vacuum ether itself. To the virtual particle flux that identically IS the vacuum spacetime.

Interestingly, relativity states that the mass itself is nothing but a "kink" or curvature in spacetime .

The fact that the surface of the particle is spinning at right angles to the virtual particle vacuum flux "drags" part of the flux along with it, spiraling it, and certainly "kinking" or "bending" the vacuum flux .

That's the MECHANISM for the bending of spacetime. And that's what MAKES the particle of mass in the first place, and reduces its dimensionality to three.

Also note that the vacuum IS phi-sub-zero, and that is never a zero value. This is contrary to present EM theory, which erroneously assumes phi-sub-zero -- the charge of vacuum -- to be zero or non-existent.

Next Slide

The Tom Bearden Website

SLIDE 27.

SPIN COUPLES VACUUM TO PARTICLE

- **SPHERICAL ROTATION IS THE KEY**
- **A PARTICLE CAN BE MODELED AS A SPHERICALLY ROTATING VORTEX OF SPACETIME**
- **ITS MASS IS DUE TO ITS SPIN**
- **THE SPINNING OBJECT IS CONTINUALLY CONNECTED TO ITS ENVIRONMENT**
- **VERY HIGH VALUES OF ELECTROSTATIC POTENTIAL CAN INDUCE RELATIVISTIC CONDITIONS**
    - **CHANGE RATE OF FLOW OF TIME**
    - **EVEN THOUGH VELOCITY IS NONRELATIVISTIC**

E.P.BATTEY-PRATT AND T.J.RACEY
"GEOMETRIC MODEL FOR FUNDAMENTAL
PARTICLES,"
INTL. J. OF PHYS. 19, NO. 6, 437-475, 1980

In fact, the bits and pieces of many of the new concepts are strewed all through the scientific literature. Here we show one example, establishing a good model that spin of a particle couples the particle to the vacuum .

This paper is by E.P. Battey-Pratt and T.J. Racey, and was . published in the International Journal of Physics in 1980.

Note again that very high values of electrostatic potential
can induce relativistic conditions, even though relativistic velocity is not present.

Next Slide

The Tom Bearden Website

SLIDE 28.

ELECTRON PRECESSION (SIMPLIFIED)

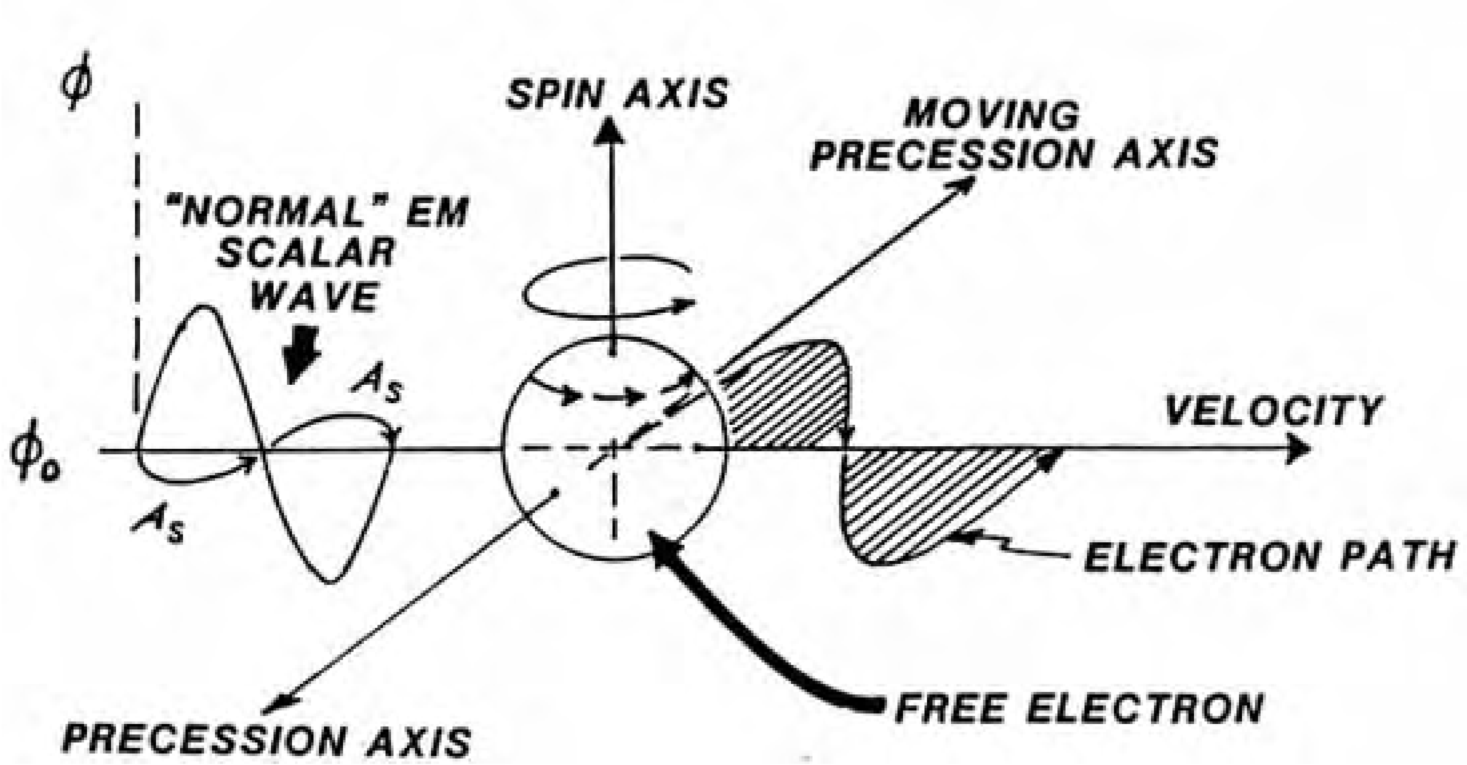

ϕ
SPIN AXIS
MOVING PRECESSION AXIS
"NORMAL" EM SCALAR WAVE
A_s
ϕ_0
A_s
VELOCITY
ELECTRON PATH
FREE ELECTRON
PRECESSION AXIS

On this slide I show the effects of precession of an electron when it encounters a longitudinal scalar wave that contains "spin vortex holes" for the electrons to fall into and mesh its spin with.

The basic idea here -- that electron precession accounts for the Hertzian waves in the electron gas in a transmitting antenna and in a receiving antenna, came from my close colleague and friend, Frank Golden, and I am most happy to give him full credit for this important insight.

As we have previously stated, transverse (force) waves cannot exist in vacuum in the absence of mass. Hertz waves therefore cannot exist in vacuum, just as Tesla stated.

Yet we know that Hertz waves exist in the electron gas in our transmitting antennas and in the electron gas in our receiving antennas. How then do we get Hertz waves here, if only longitudinal waves can exist in the vacuum in between?

Electron precession is the key.

We never measure what's happening in vacuum with our instruments. Instead, almost always we measure what is happening to and in the electron gas in our antenna or probe and feeding current to the instrument.

Here we show a "normal" EM vacuum wave -- which is a longitudinal wave containing spin vortexes from the electrons that generated it -- approaching and striking a spinning charged electron. As the peak and trough of the wave passes, it is as if we had a force pressing against the spinning electron, first along the line of wave travel, and then antiparallel to the wave travel. *(that is, "explaining" it in present concepts.)*

The electron acts as a little gyroscope, and precesses laterally, first in one direction and then the other.

Therefore the wave recovered in the electron gas in our receiving antenna or instrument probe is a transverse Hertzian matter wave.

Hertz waves are always matter vector waves.

Vacuum EM waves are always nonmaterial longitudinal scalar waves.

Longitudinal scalar waves in vacuum normally contain many spinning vortex "holes" of flux, created from the spinning electrons which launched the wave and stayed behind in the transmitting antenna. This kind of longitudinal wave is directly detectable by a normal free electron charged gas in a receiving antenna or probe. It also directly interacts with free electrons in a conducting metal shield, and so is shielded by Faraday cages.

On the other hand, our zero-vector longitudinal wave, made by opposing waves, contains opposing spin holes which annihilate or cancel each other.

In the absence of spin holes, the longitudinal wave will not mesh with spinning electrons in a conductor, and so it is not detectable in the normal fashion. It also does not interact with free electrons in a conducting metal shield, so it readily penetrates Faraday cages.

An easy way to see that electrons do not interact with that substructured longitudinal zero-vector wave is to visualize both substructure component waves interacting on the electron simultaneously, pushing in opposite directions equally. In that case the electron tries to precess in both directions, equally, and so it does not precess in either direction. Therefore it does not "detect" the passing wave.

The wave without "golf ball holes", however, is detectable

by any circuit having high nonlinearity actions occurring in it. Such highly nonlinear dynamic areas act to provide a phase shifting between the composite substructure waves. This phase shift results in violation of the sum-zero condition, producing a "normal" EM wave which deposits energy in the out-of-phase area.

From the spin vortex "golf ball hole" concept, the out-of-phase condition means that now we have an alternating preponderance of spin holes, spinning first in one direction and then in the other. Thus the electrons in the nonlinear, phase shift area are hooked and oscillated (precessed) to and fro, producing energy.

Solid state, highly doped transistors are particularly vulnerable to this effect, as are gas discharge tubes, spark and cascade ion discharges, plasmas, etc.

Next Slide

The Tom Bearden Website

SLIDE 29.

VACUUM/SPACETIME IS PURE VIRTUAL PARTICLE FLUX ($Ø_O$)

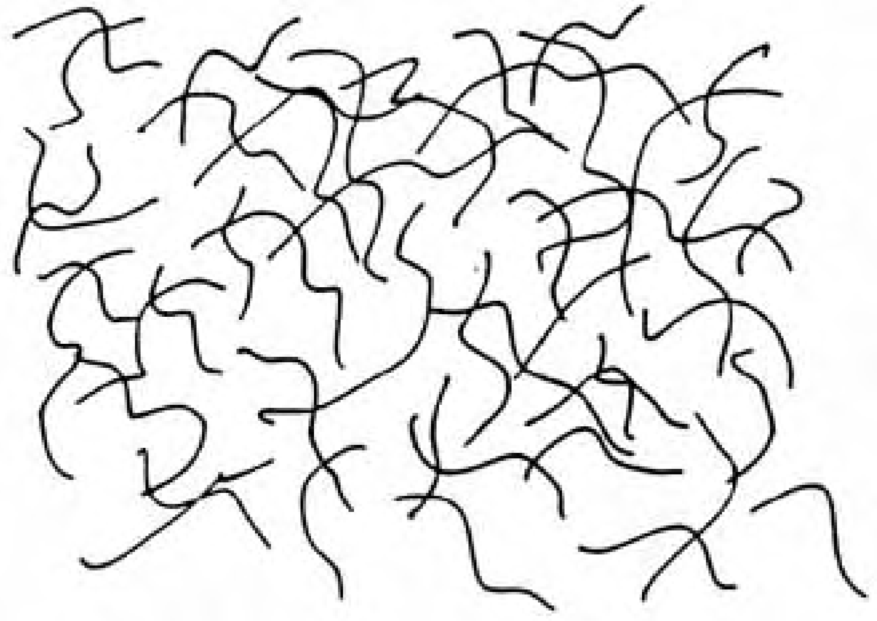

VACUUM IS IDENTICAL TO:

- **VIRTUAL PARTICLE FLUX**
- **SPACETIME**
- **ANENERGY**
- **$Ø_0$**
- **MASSLESS CHARGE**

VACUUM IS:

- **MADE OF UNQUANTIZED ACTION**
- **WITHOUT DEFINITE LENGTH INTERVALS**
- **WITHOUT DEFINITE TIME INTERVALS**
- **N-DIMENSIONAL (UNFIXED)**

We pause here to again summarize some important things we now know about vacuum.

Vacuum IS IDENTICAL TO virtual particle flux, spacetime, anenergy, Phi-subzero, massless charge, unquantized action, and infolded energy, momentum, length, time, power, velocity, acceleration, etc.

Vacuum is made of unquantized action, dimensionally.

It is without definite length or time, as Einstein stated.

It is n-dimensional, where n is always four or greater.

It is not an emptiness filled with things. Rather, it is identically a plenum filled with things which are totally unseparated, unfixed, and undetermined.

From vacuum: mass, energy, time, frames, and all the rest of physical reality are dynamically made.

Vacuum is the ether. The ether has never been eliminated from physics. Ether just has become much more subtle, and physicists have used other names for it.

Next Slide

The Tom Bearden Website

SLIDE 30.

VACUUM IS:

- SPACETIME (L^nT, WHERE $n \geq 3$)
- CHARGE (MASSLESS)
- ELECTROSTATIC SCALAR POTENTIAL $\text{Ø}_0 \neq 0$
  ERROR IN PRESENT THEORY
- BROKEN BITS OF ENERGY (SUBQUANTAL)
- PURE VIRTUAL PARTICLE FLUX
- O-WAVE FLUX
- MULTILEVEL, STRUCTURED, PATTERNED
- A VIRTUAL PLENUM
- AN OBSERVABLE EMPTINESS

Again, we list some of the things that vacuum identically is.

It is spacetime, massless charge, electrostatic scalar potential, broken bits (subquantal) of energy, pure virtual particle flux, zero-vector wave flux, multilevel, structured, patterned, a virtual plenum, and an observable emptiness.

It is all things and contains all things in potential state.

It is not, in that it is not observable. But from it comes all observables.

It is both ordered and disordered, simultaneously.

It involves four-law logic, not just three-law Aristotlean logic.

Next Slide

The Tom Bearden Website

SLIDE 31.

GRADIENT OF ELECTROSTATIC POTENTIAL

IN VACUUM

$$\nabla \mathbf{Ø} \neq \mathbf{\ddot{E}}$$

ON AND OF A MOVING CHARGED PARTICLE

$$\nabla \mathbf{Ø} = \mathbf{\ddot{E}}$$

Again I accent that, in vacuum in the absence of a charged spinning particle of observable mass, del-phi does not yield an Ë-field, and the conventional equations of EM are wrong in that respect.

In the presence of a spinning charged particle, in a del-phi which contains electron-vortex-holes to mesh with, the charged particle attaches itself to the moving del-phi flux gradient, moving itself with the river. This produces an Ë field. The Ë-field CONSISTS OF the smeared electron, it does not CAUSE THE MOVEMENT of the electron. It is an effect, not a cause. The conventional equation for del-phi equals E is correct for matter waves in electron gases; it is not correct in vacuum itself.

Next Slide

The Tom Bearden Website

SLIDE 32.

IN A ∇Ø, A CHARGED PARTICLE MOVES ITSELF

∇Ø>0 IMPLIES A MOVING RIVER, AND THE CHARGED PARTICLE IS "HOOKED TO" THE RIVER

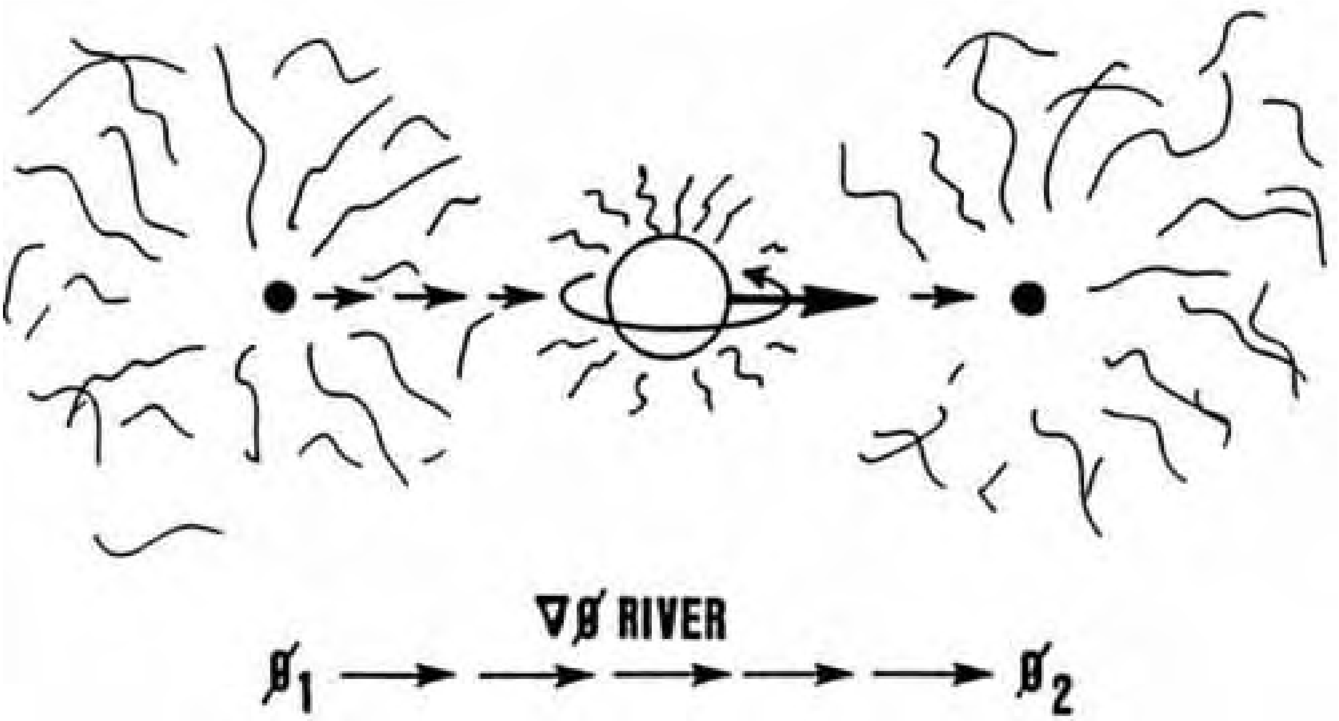

We show clearly on this slide that a spinning charged particle, when it hooks to a spin-hole in a del-phi river, MOVES ITSELF.

Note that a del-phi implies a "pressure difference" in phi-flux between two points. Therefore there is a moving river -- a gradient -- of phi-dot, moving from the higher pressure to the lower.

If a charged particle is attached to this moving river, it flows along with the river. It is just like putting a boat into a moving stream of water. If the boat "latches" to the moving water, it moves with the river.

Electron spin and spin vortexes in the del-phi mesh, to accomplish the "hooking" or "latching."

Once hooked into a moving river, the electron moves itself. One does not have to furnish work to move it.

If the river is accelerating, the electron will accelerate itself.

By cleverly gating such rivers into and onto sources of free electrons, one can directly produce free energy.

The extra energy comes from converting anenergy (massless charge, vacuum flux) to energy of a moving mass.

Next Slide

The Tom Bearden Website

SLIDE 33.

ELECTROMAGNETIC WAVES IN ELECTRON GAS IN A CONDUCTOR

- **VECTOR WAVES**
- **CONTAIN E, B FIELDS**
- **GENERATED BY SPIN COUPLING OF ELECTRONS WITH SCALAR WAVE "SPIN-HOLES"**
- **WAVES OF FORCE**
- **HERTZIAN WAVES**
- **TRANSVERSE WAVES**
- **WAVES OF ELECTRON PRECESSION**

Here we pause to summarize the characteristics of electromagnetic waves in the electron gas of a conductor.

First, these are vector EM waves in the accepted sense.

They contain E and B fields.

They are generated by spin coupling of electrons with spin-holes in a scalar wave delta-phi river. The electrons move themselves under that condition, producing work and forces.

They are waves of force or force field variations.

They are Hertzian waves.

They are transverse oscillatory, not longitudinal.

They are matter waves.

They are waves of electron precession.

Next Slide

SLIDE 34.

ELECTROMAGNETIC WAVES IN VACUUM

- **SCALAR LONGITUDINAL WAVES**
- **VECTOR ZERO [Ö] WAVES**
- **WITHOUT Ë, B FIELDS**
- **INTERNALLY STRUCTURED**
- **PATTERNED**
- **MAY CONTAIN ELECTRON "SPIN-HOLES"**

Here we summarize the characteristics of electromagnetic waves in vacuum.

They are scalar longitudinal waves of alternate compression and rarefaction of the vacuum virtual particle flux.

That is, they are waves of electrostatic potential.

They are zero-vector waves.

They are internally structured and patterned.

They usually contain electron "spin holes" unless made in a fashion so as to make opposing spin holes that cancel each other.

Since they are pure phi-waves, they need not be limited in velocity to the speed of light.

They are hyperspatial waves.

They are waves in virtual state itself.

Next Slide

The Tom Bearden Website

SLIDE 35.

DETECTION OF TRANSVERSE AND LONGITUDINAL WAVES

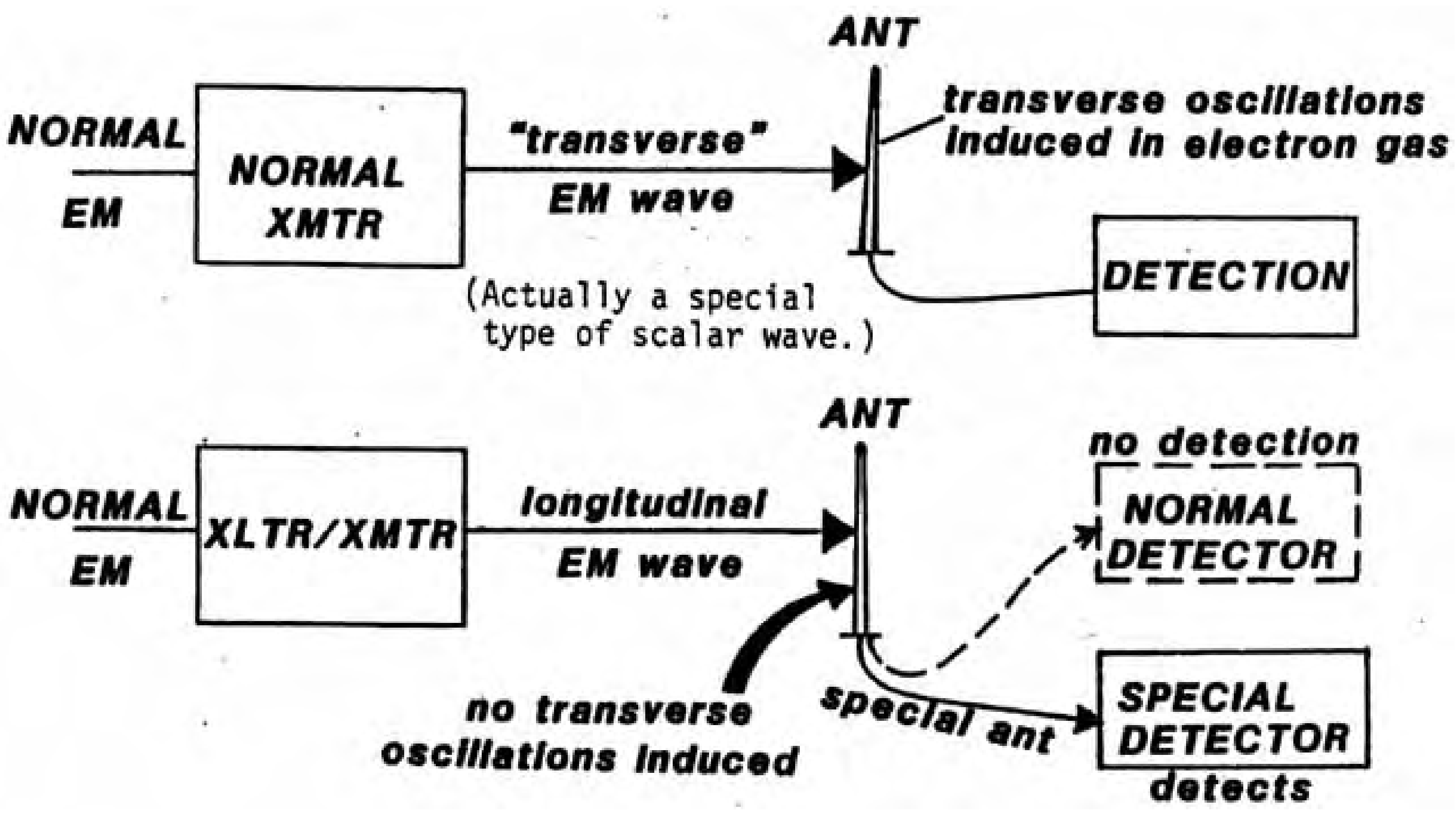
ANT
NORMAL
EM
NORMAL XMTR
"transverse"
EM wave
transverse oscillations induced in electron gas
DETECTION
(Actually a special type of scalar wave.)
ANT
no detection
NORMAL
EM
XLTR/XMTR
longitudinal
EM wave
NORMAL DETECTOR
no transverse oscillations induced
special ant
SPECIAL DETECTOR
detects

In the top drawing on this slide, we show a normal transmitter putting out a normal EM wave, which received in a normal antenna/receiver. What we call a "transverse" wave rigorously exists only in the electron gas in the transmitting antenna and in the electron gas in the receiving antenna. Specifically, a longitudinal wave exists in the vacuum in between the two antennas. However, due to the method of production, the longitudinal wave contains spin-holes for electrons, so the electrons in the receiving antenna readily couple with the wave by falling into and meshing their spins with the spins of the spin-vortex holes. Electron precession produces electron gas waves that are transverse oscillatory, hence Hertzian waves in nature.

In the bottom drawing, we show a translator/

transmitter. In other words, we oppose ordinary EM waves in a sum-zero substructure, deliberately producing longitudinal EM waves in vacuum with the spin-hole vortexes canceling each other. This type of wave does not "hook" spinning electrons in the normal receiving antenna, and thus it is not detected. The normal antenna/receiver system never sees it at all.

However, by means of a special antenna which generates nonlinear phase shifts in the composite substructure waves, "hooking" holes are restored in an oscillatory nature. Electrons then hook with this output and detect the wave.

Note that it requires a two-stage detector, operated in an interferometer fashion, to detect this pure scalar wave.

Next Slide

The Tom Bearden Website

SLIDE 36.

CREATING A "RIVER" OF ANENERGY

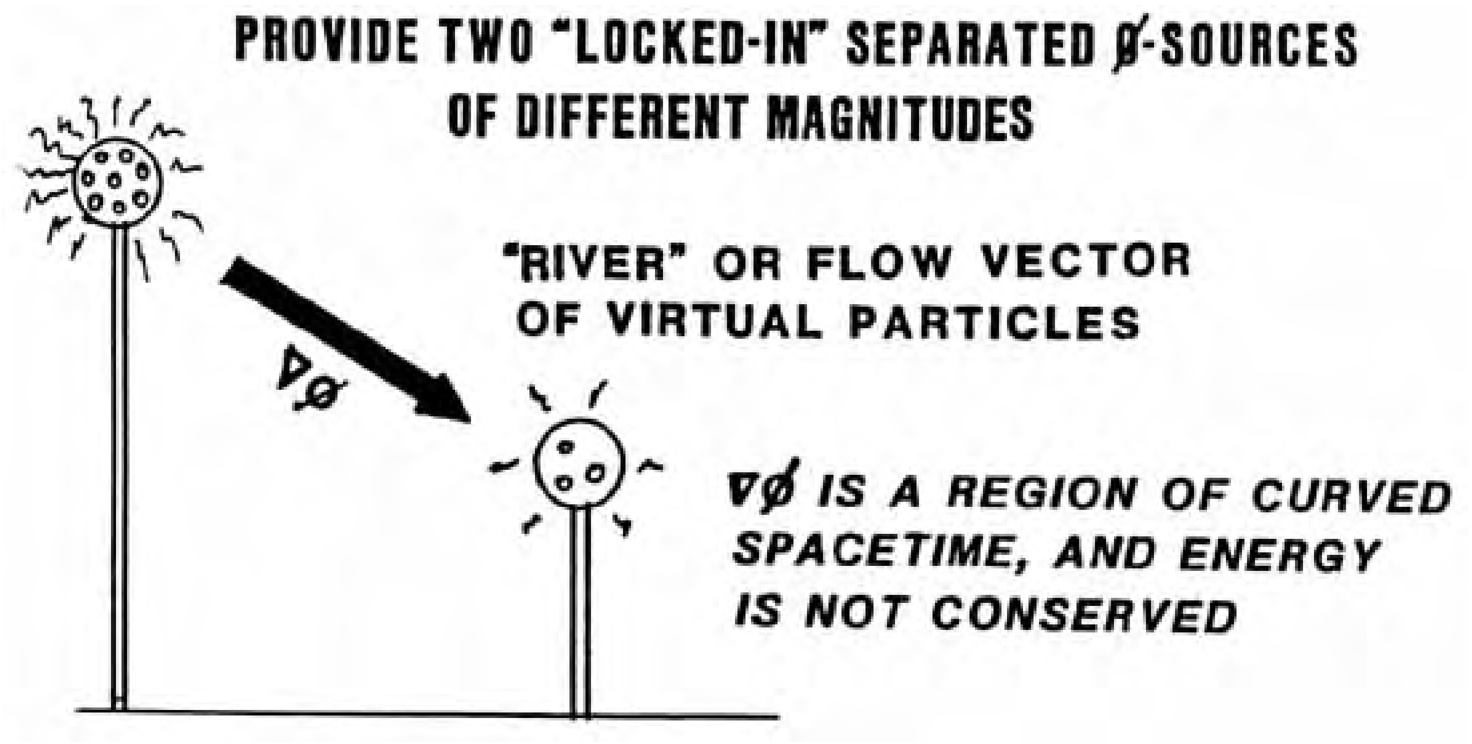

PROVIDE TWO "LOCKED-IN" SEPARATED ϕ-SOURCES OF DIFFERENT MAGNITUDES

"RIVER" OR FLOW VECTOR OF VIRTUAL PARTICLES

$\nabla\phi$ IS A REGION OF CURVED SPACETIME, AND ENERGY IS NOT CONSERVED

IN VACUUM, THIS IS NOT YET A VOLTAGE DROP (NO MASS FLOW),
NOR IS THERE AN Ë-FIELD

Now if we wish to get free energy, we are going to have to provide a continuous anenergy river, and a means of tapping it to produce mass movement.

This slide shows one easy way to make an anenergy river.

We simply pump some electrons (spray nozzles) onto an elevated charged ball, and we LEAVE them there. A second ball is connected to ground. A higher phi -- that is, a denser spray -- is now in the vicinity of the elevated ball on the left. A lower phi -- that is, a less dense spray -- is in the vicinity of the lower ball on the right. Between the two balls, now, there is a gradient in phi, and a virtual flux flowing from the "higher virtual pressure" to the "lower virtual pressure."

This del-phi river does NOT constitute an Ë-field, as we have previously pointed out.

The del-phi region, however, is definitely a region of curved spacetime. As is well known from general relativity, in such a region energy need not be conserved.

Therefore it is entirely possible -- consistent with ordinary physics -- to violate conservation of energy in this del-phi river, if one believes general relativity. If one argues adamantly that conservation of energy cannot be violated under any circumstances, then one must throw out general relativity. Also, one must throw out most of particle physics, whose explanations presently involve virtual interactions, each of which violates the conservation of energy.

Next Slide

The Tom Bearden Website

SLIDE 37

A 2-WIRE SYSTEM IS CONSERVATIVE

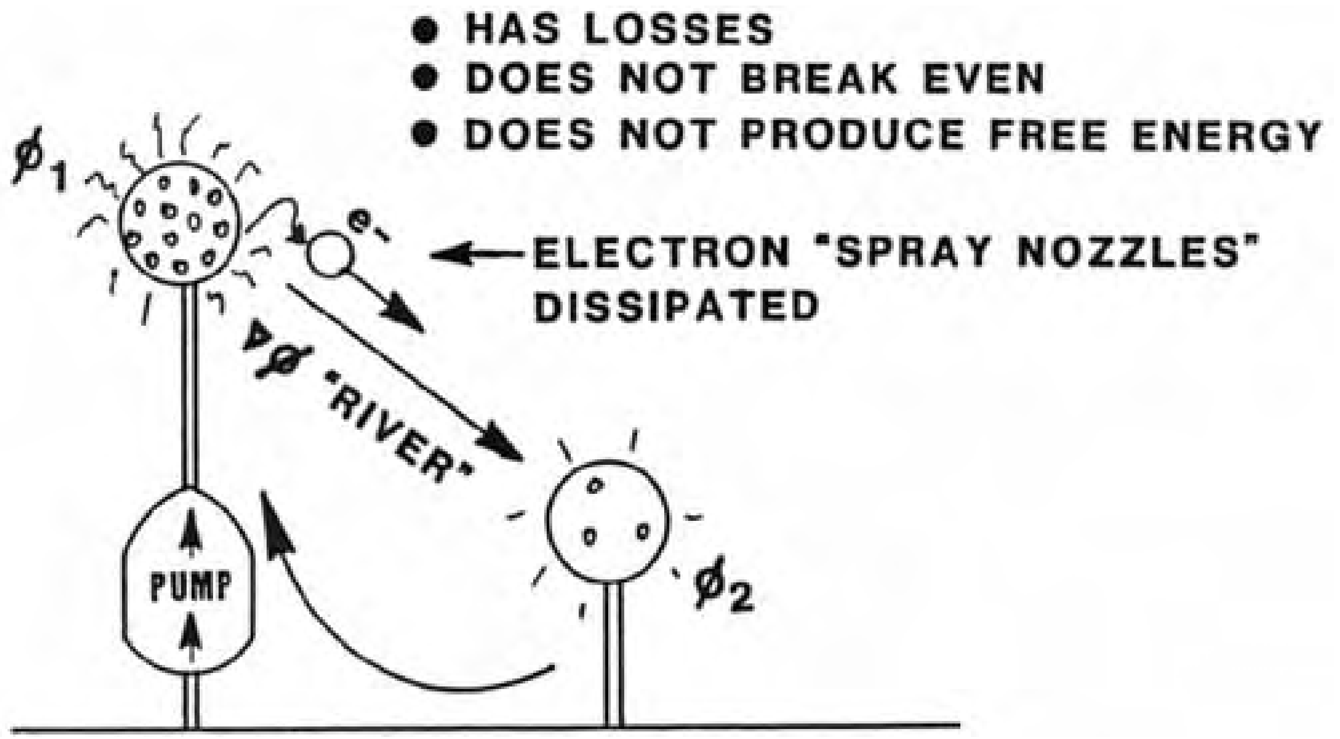

- HAS LOSSES
- DOES NOT BREAK EVEN
- DOES NOT PRODUCE FREE ENERGY

ELECTRON "SPRAY NOZZLES" DISSIPATED

REPLENISH SPRAY NOZZLES BY PUMPING MORE ELECTRONS. (MOVING ELECTRON MASSES AGAINST A "SPRAY PILEUP" REQUIRES WORK.

Here we show that, as engineers, we've all been tricked into releasing our virtual rivers so that we would not discover free energy.

While I cannot prove this as yet, I tentatively suspect that J. P. Morgan and Thomas Edison are directly responsible for this occurring. Tesla utilized a single-wire system; we have all been taught to utilize only two-wire systems.

What we in effect do is connect a wire or circuit to our charged ball, allowing our hose nozzles to drain away.

In other words, we first do some work to pump hose nozzles (electrons) into and onto our elevated reservoir, so we will obtain a denser flux, and a flux river to utilize. Then we release the nozzles themselves, and let THEM move on the flux river, deplenishing our river. Of course we can get back as much energy as we pumped up there in the first place. But this two-wire system is at best theoretically even. Since real circuits have losses, we can never even quite break even. Thus we have to continue to pump hose nozzles, and let them bleed off to provide energy, in a less-than-100% fashion.

By foolishly releasing our hose nozzles, we guarantee that we will never achieve free energy. And that's nice, because the people who control things, and for their own wealth and power depend on selling us energy at high prices, are guaranteed to stay in power and control.

One is never truly free unless one is energy independent, in a self-contained fashion.

Tesla nearly succeeded in giving us free or very cheap, inexhaustible energy -- and for that reason was quickly and dramatically suppressed.

Only three graduate schools of electrical engineering existed in the U.S. The only one of any real importance was at Columbia University.

Only one or two professors were necessary to "reach" to modify and subtly direct the entire foundations of the emerging electrical theory.

Professors were in general poor, and had no research funds unless some rich industrialist or philanthropist gave them a few dollars.

The setup then was certainly very ripe for Morgan and Edison to stop free energy, and suppress Tesla at the same time. They could have done so by influencing only one -- or at most two -- theoreticians at Columbia.

At any rate, the theory was developed in such fashion as to rule out the single wire circuit and the longitudinal wave. Free energy was suppressed, and the financial empires of Morgan and Edison stayed secure.

Next Slide

The Tom Bearden Website

SLIDE 38.

A HOLLOW CHARGED SPHERE

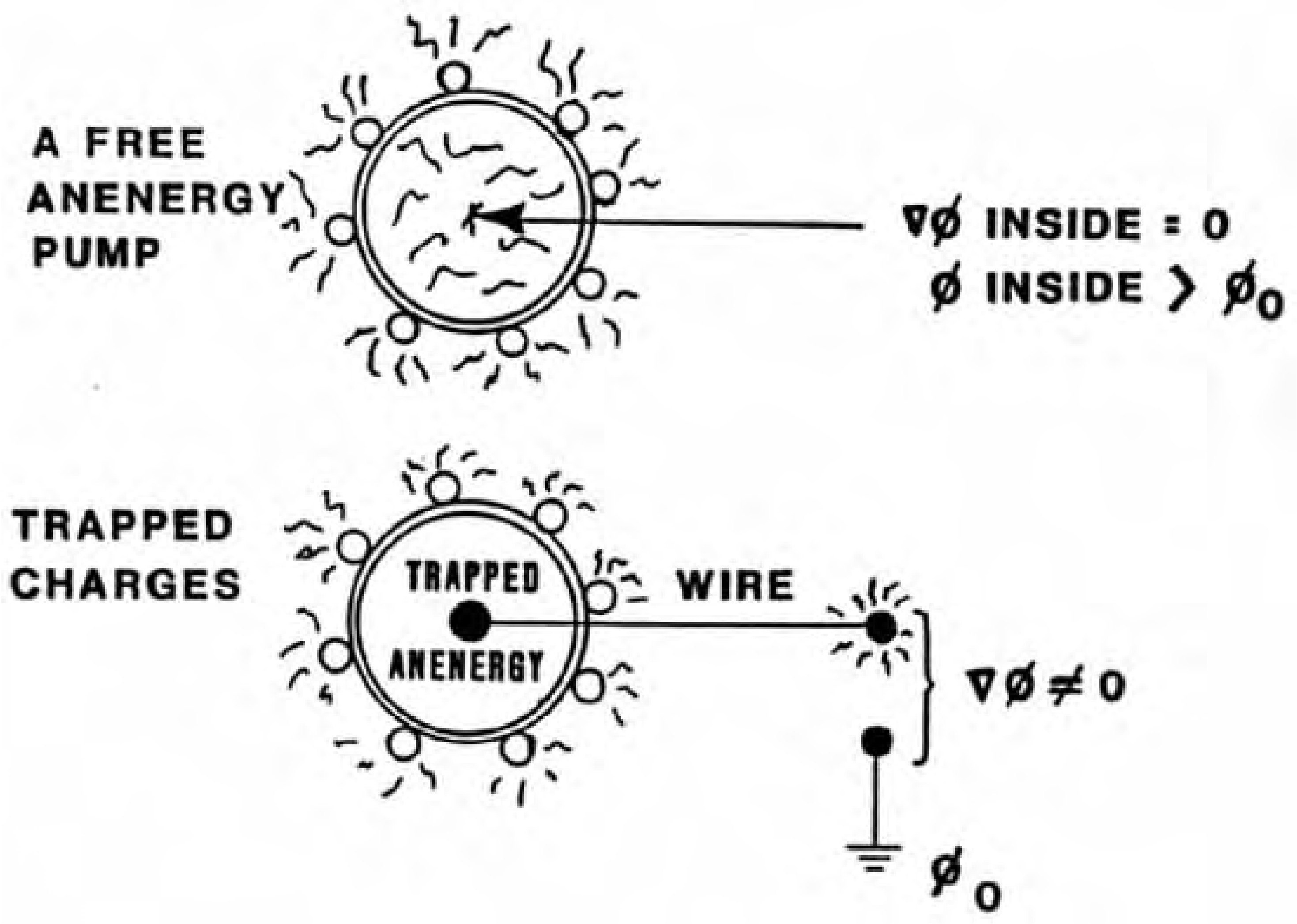
A FREE ANENERGY PUMP
$\nabla\phi$ INSIDE = 0
ϕ INSIDE > ϕ_0
TRAPPED CHARGES
TRAPPED ANENERGY
WIRE
$\nabla\phi \neq 0$
ϕ_0

Here we show another easy source of a higher vacuum charge, hence an easy source for a free river.

Simply charge up a sphere. As is well known, inside that sphere one has no del-phi. So one can introduce an electron therein and not produce an Ë-field .

However, since we've piled up additional spray nozzles, the density of the spray inside the sphere (as well as outside it) is certainly higher.

In other words, the inside of the sphere is also a source
for an increased del-phi river. In the bottom drawing, we drill
a tiny hole in the sphere, insulate a thin wire running into the inside of the sphere, and run the wire out to a terminal opposing another grounded terminal. Between the two terminals now there must exist a del-phi river.

And between those terminals there now exists a curvature of spacetime, hence the theoretical possibility to obtain free energy without violation of principles.

Next Slide

SLIDE 39.

ONE-WAY GATE VALVE

(ANENERGY TO ENERGY)

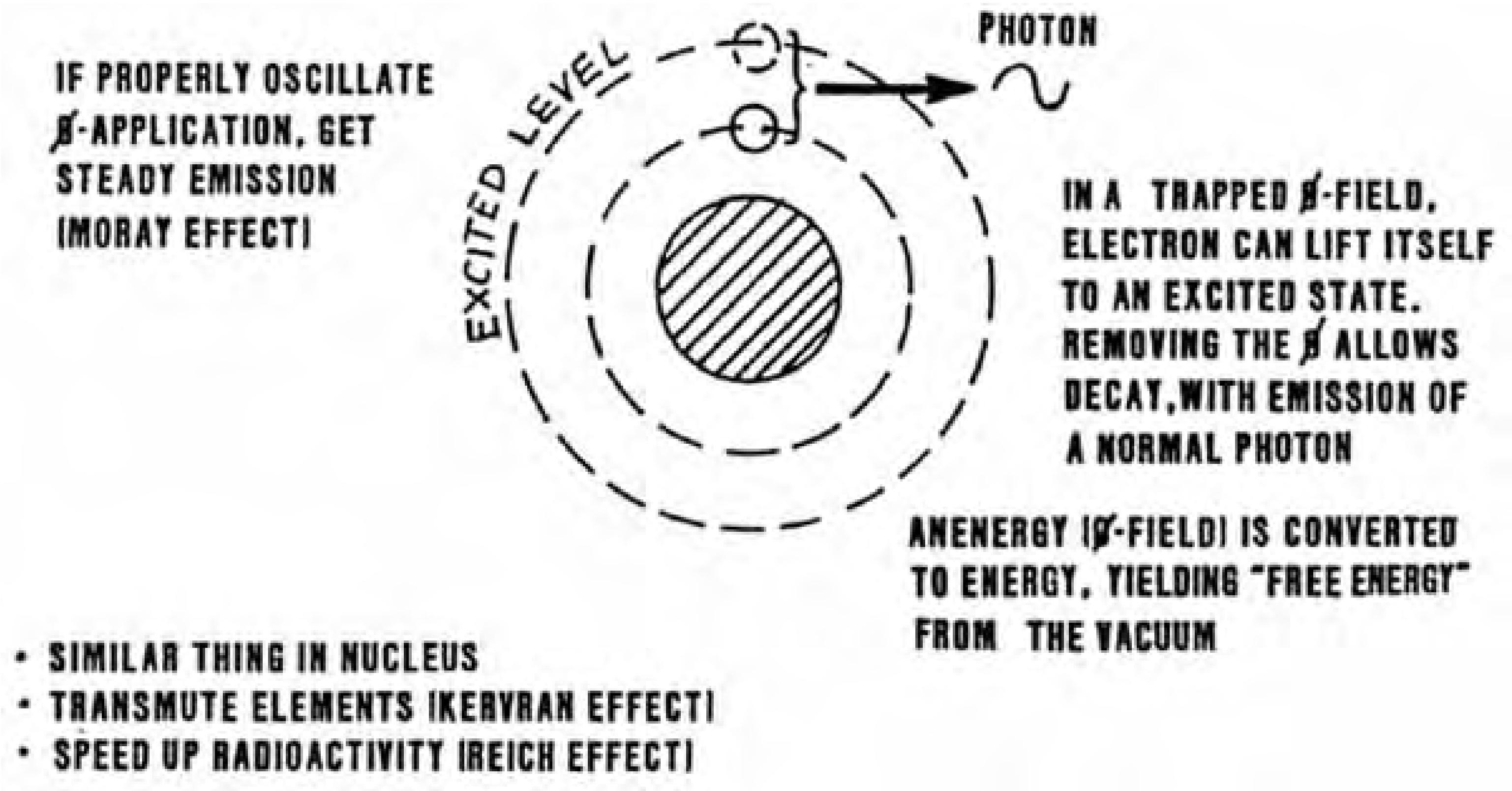

IF PROPERLY OSCILLATE Ø-APPLICATION, GET STEADY EMISSION (MORAY EFFECT)

IN A TRAPPED Ø-FIELD, ELECTRON CAN LIFT ITSELF TO AN EXCITED STATE. REMOVING THE Ø ALLOWS DECAY, WITH EMISSION OF A NORMAL PHOTON

ANENERGY (Ø-FIELD) IS CONVERTED TO ENERGY, YIELDING "FREE ENERGY" FROM THE VACUUM

- SIMILAR THING IN NUCLEUS
- TRANSMUTE ELEMENTS (KERVRAN EFFECT)
- SPEED UP RADIOACTIVITY (REICH EFFECT)

Here we show one way to provide a one-way gate valve from a del-phi river to energy released into the outside world.

By providing a properly tuned, oscillating phi wave, one can cause an orbital electron in an atom to raise itself to an excited state. Remember, in a del-phi, an electron that hooks to it raises itself.

We then switch out the phi application, and the excited electron decays in normal fashion, emitting a normal photon of energy.

By phase locking the processes, one produces a steady stream of photons.

By absorbing or directing the photons, one can produce heat, light, energy beams, etc.

By this means, one produces free energy by gating anenergy into energy.

And if one does not release spray nozzles in the phi-oscillations which feed the mechanism, one has an inexhaustible energy production.

A very similar thing can be done with the excited states of the nucleus. For example, by using a radioactive material with a predisposition to emit some particle -- say an electron -- the nucleus can be made to continually emit the particles WITHOUT ITSELF DECAYING INTO BYPRODUCTS. In other words, an electron is lifted from the Dirac sea of vacuum to instantly replace the electron emitted from the nucleus.

This provides a one-way gate valve from the Dirac sea to the external universe.

Theoretically, these mechanisms are no more mysterious than a hydraulic ram, which raises water higher than the stream that feeds the ram.

In the 20's and 30's, T. Henry Moray successfully built such gating devices for converting vacuum anenergy to external energy.

In the Kervran effect, living systems are able to directly transmute elements, even though they only possess feeble energies, by influencing the virtual particle interactions in the nucleus that furnish its binding energy.

Reich used an adaptation of a cloudbuster -- which emits scalar waves -- to drastically speed up radioactive decay of radioelements.

Both plants and animals transmute elements by similar processes.

Next Slide

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 40,

SIMPLE FREE ENERGY MOTOR

BATTERY RECHARGES ITSELF

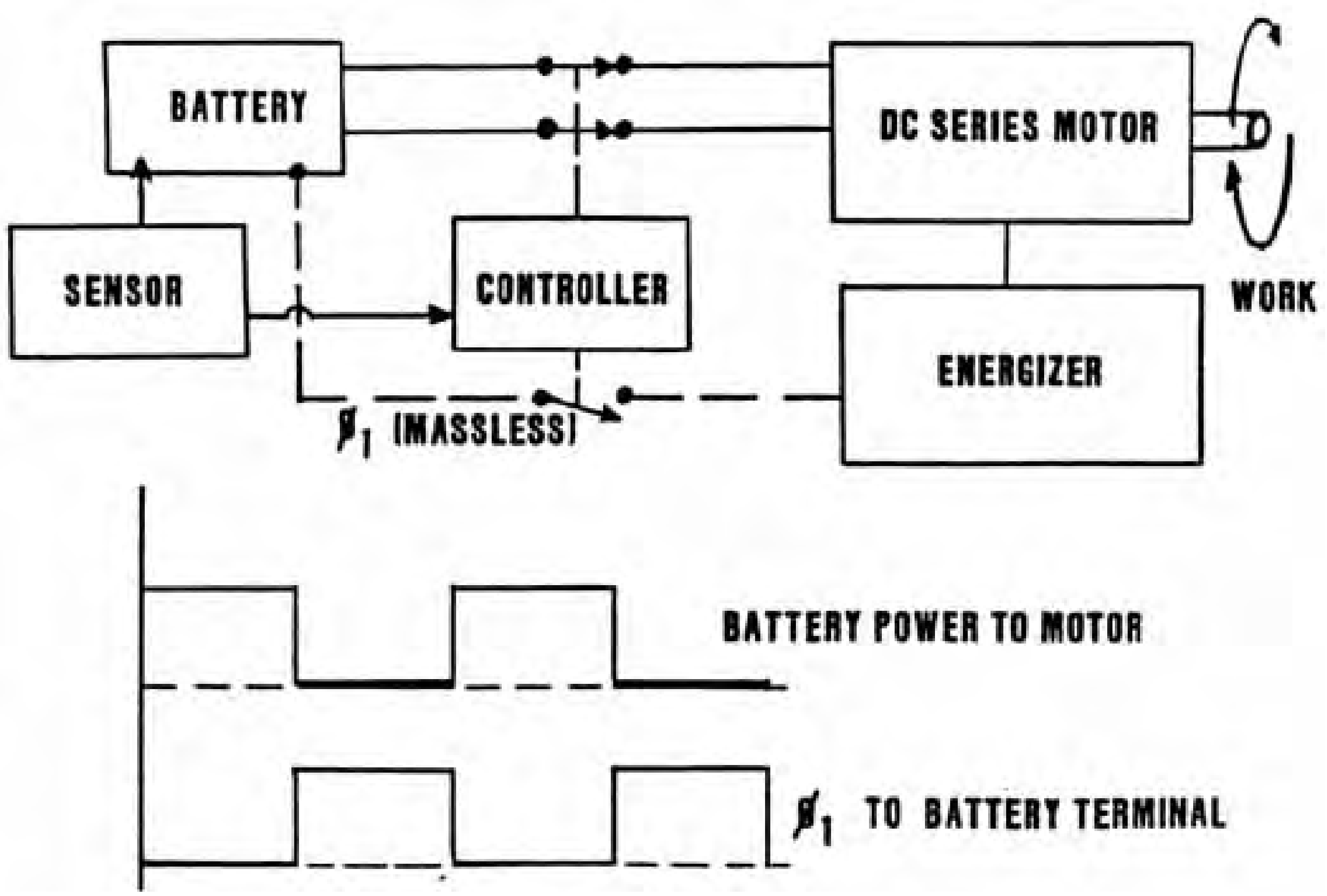
BATTERY
DC SERIES MOTOR
SENSOR
CONTROLLER
WORK
ENERGIZER
Ø1 (MASSLESS)
BATTERY POWER TO MOTOR
Ø1 TO BATTERY TERMINAL

On this slide, we show a theoretical scheme which several researchers have discovered and used to build simple free energy motors.

In this scheme, we drive an ordinary d.c. series motor by a two wire system from an ordinary battery. The motor produces shaft horsepower, at -- say -- some 30 or 40 percent efficiency, compared to the power drained from the battery. This much of the circuit is perfectly ordinary.

The trick here is to get the battery to recharge itself, without furnishing normal power to it, or expending work from the external circuit in the process.

To do this, recall that a charged particle in a "hooking" del-phi river moves itself. This is true for an ion, as well as for an electron. We need only make the del-phi in correct fashion and synchronize it; specifically, we must not release the hose nozzles we utilize to produce our del-phi river or waves.

The inventors who have discovered this have used various variations, but here we show a common one.

First, we add an "energizer" (often referred to by various other names) to the circuit. This device makes the del-phi waves we will utilize, but does NOT make currents of electron masses. In other words, it makes pure Ø-dot. It takes a little work to do this, for the energizer circuit must pump a few charges now and then. So the energizer draws a little bit of power from the motor, but not very much.

Now we add a switching device, called a controller, which breaks up power to the motor in pulses. During one pulse, the battery is connected and furnishes power to the motor; during the succeeding pulse, the battery is disconnected completely from the motor and the output from the energizer is applied across the terminals of the battery.

If frequency content, spin-hole content, etc. are properly constructed by the energizer, then the ion movements in the battery reverse themselves, recharging the battery. Again, remember that these ions MOVE THEMSELVES during this recharge phase. Specifically, we are NOT furnishing ordinary current to the battery, and we are not doing work on it from the energizer.

If things are built properly, the battery can be made to more than recover its charge during this pulse cycle.

To prevent excess charge of the battery and overheating and destroying it, a sensor is added which senses the state of charge of the battery, and furnishes a feedback signal to the controller to regulate the length of recharge time per "power off" pulse. In other words, the system is now self-regulating.

The relation between power pulses and recharge pulses is shown on the graphs at the bottom. Note that regulation may decrease the time of recharge application of the del-phi river.

This system, if properly built and tuned, will furnish "free shaft energy" continually, without violating conservation of anenergy. Remember that the del-phi condition across the battery terminals means that spacetime is suddenly curved there, and conservation of energy need no longer apply.

Again, this system is consistent with general relativity and with the fact that Ø-field alone can drive a situation relativistic. We have deliberately used these facts to do direct engineering. Our "extra energy" comes from shifting phi-flux -- the energy of the universal vacuum spacetime -- directly into ordinary energy for our use. Thus we draw on an inexhaustible source, and our device is no more esoteric than a paddlewheel in a river. The only difference is that, in this case, we have to be clever enough to make and divert the river in the right timing sequency.

This is a free energy device which an ordinary person, who knows a little electronics, can experiment with in the basement. To develop it, one is talking several thousands of dollars and a lot of persistence and tinkering; one is not talking millions.

WARNING: THE PRECEDING EXPERIMENTS ARE HAZARDOUS. DO NOT ATTEMPT THESE EXPERIMENTS UNDER ANY CIRCUMSTANCES UNLESS YOU ARE AN EXPERIENCED ELECTRICAL RESEARCHER, EXPERIENCED IN PERFORMING EXPERIMENTS WITH LEAD-ACID BATTERIES AND PULSE CHARGE AND DISCHARGE OF SAME, AND UNLESS YOU ALSO USE ALL SAFETY PRECAUTIONS SUCH AS GOGGLES

AND PROTECTIVE GLOVES, SLEEVES, AND APRON. YOU MUST NOT HAVE OTHER INFLAMMABLE LIQUIDS OR OTHER SUBSTANCES PRESENT WHICH COULD BE IGNITED AND BURN OR EXPLODE. SURGED LEAD-ACID BATTERIES PRODUCE HYDROGEN GAS, WHICH CAN EASILY EXPLODE SINCE SPARKING ALSO CAN OCCUR. THE ACID FROM SUCH AN EXPLOSION CAN EASILY BLIND YOU IF IT GETS IN YOUR EYES, AND IT CAN BURN YOUR SKIN. IN ADDITION, LEAD AND LEAD COMPOUNDS ARE POISONS, AND ARE TO BE HANDLED ONLY BY EXPERIENCED RESEARCHERS. THESE EXPERIMENTS ARE NOT FOR AMATEURS UNDER ANY CIRCUMSTANCE, BUT ONLY FOR EXPERIENCED PROFESSIONALS WITH PROPER KNOWLEDGE AND TRAINING, AND USING PROPER PRECAUTIONS. NEITHER THE AUTHOR NOR THE PUBLISHER IS REPONSIBLE OR LIABLE FOR ANY ACCIDENTS OR DAMAGE YOU MAY ENCOUNTER, AND ALL EXPERIMENTATION WITH THESE DEVICES AND PROCEDURES IS AT YOUR OWN ASSUMED PERSONAL RISK.

Next Slide

The Tom Bearden Website

SLIDE 41.

THE TRANSLATOR

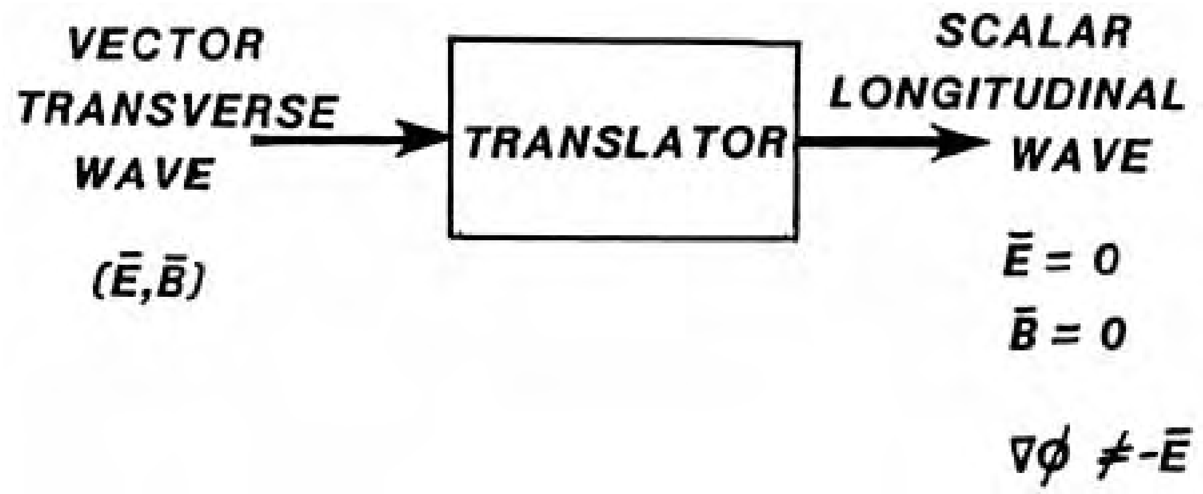

We now turn to a fundamental concept we need in order to understand some of the effects used by the Soviets in their Tesla weapons.

Note that one other country -- one not hostile to the U.S. -- also has such weaponry.

This slide shows the concept of a TRANSLATOR. Briefly, if we input a transverse matter wave, the translator outputs a scalar longitudinal wave without spin vortex hooks. If we input a scalar longitudinal wave -- even one without hooks -- the translator puts out a normal vector transverse matter wave.

We define a translator as anything which will accomplish one or both of those functions (usually both).

Basically, any device which is TOTALLY nonlinear to transverse waves will produce scalar waves from a transverse wave input. Any device which is totally nonlinear to longitudinal waves will produce transverse waves from a longitudinal wave input.

It is comparable to certain plasmas, which, as is well
known, produce such a translator effect. If you input longitudinal waves, you get transverse waves. If you input transverse waves, you get longitudinal waves. And the plasma is rather totally nonlinear.

Again, if our translator outputs non-hooking waves, it means that the component substructure waves oppose each other in sum zero fashion, and "kill" or nullify the vortex spin-holes.

Next Slide

The Tom Bearden Website

SLIDE 42.

TRANSVERSE WAVE INTERFERENCE

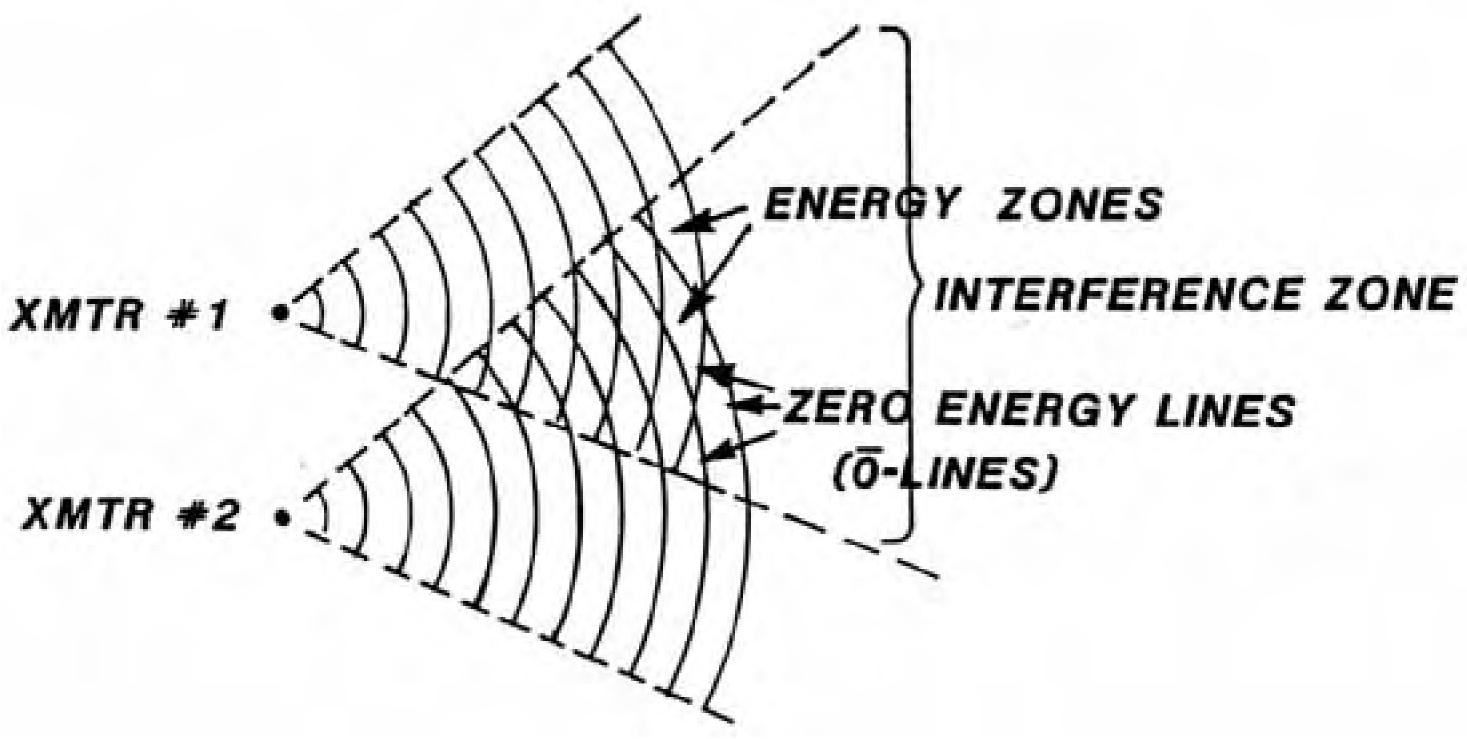

On this slide we show "normal " transverse wave interference, as we presently teach it in our textbooks.

We visualize two identical transmitters, each producing a beam in a pattern of about 55 degrees of so.

Where these beams overlap -- in the ideal, perfect case, we have constructive and destructive interference, with absolute zero-vector linear regions being formed. We put in the energy, and out-of-phase interference creates the zero zones, which contain no energy as such.

We point out, however, that these zero-lines are scalars,
and contain high stresses on spacetime.

Next Slide

The Tom Bearden Website

SLIDE 43.

LONGITUDINAL WAVE INTERFERENCE [SCALAR WAVES]

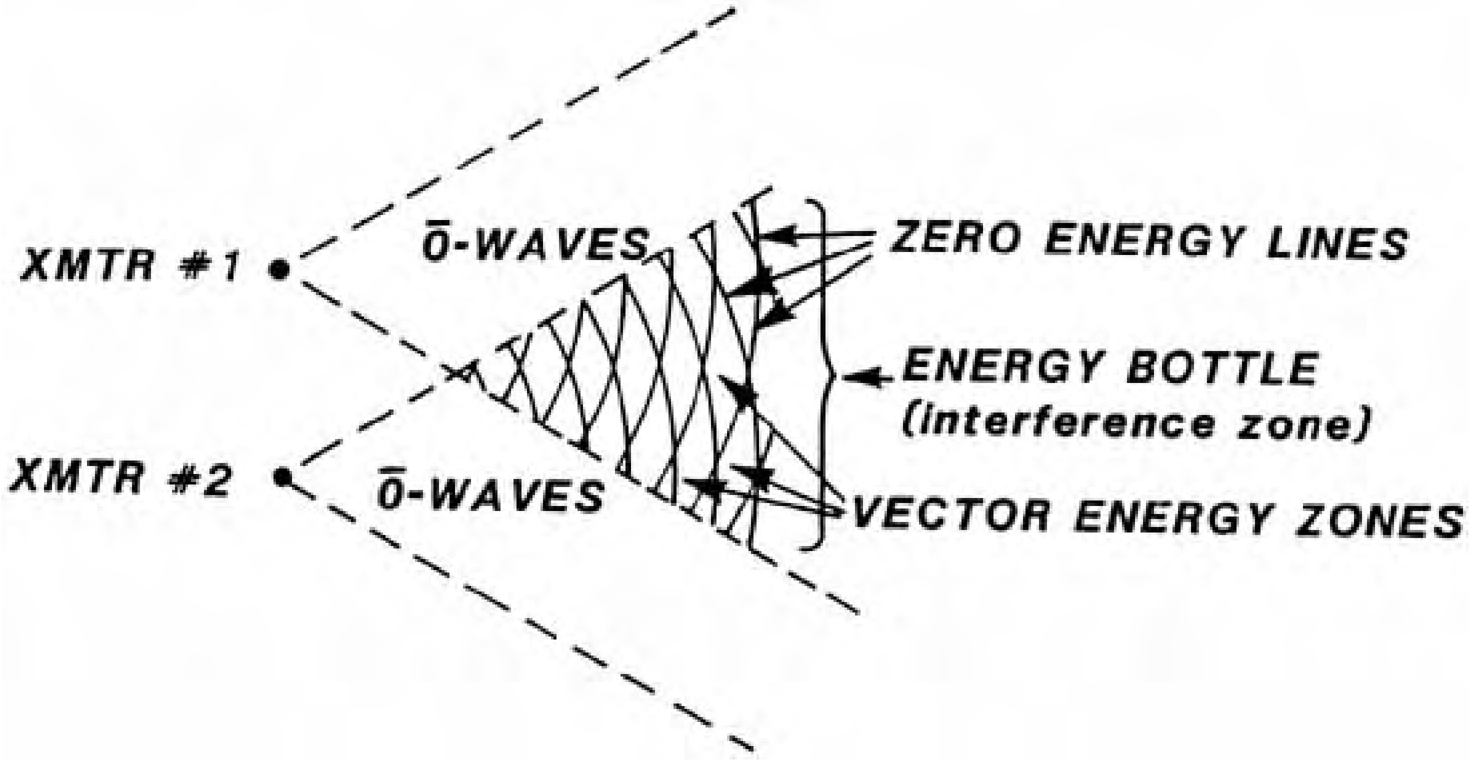
XMTR #1
XMTR #2
Ō-WAVES
Ō-WAVES
ZERO ENERGY LINES
ENERGY BOTTLE
(interference zone)
VECTOR ENERGY ZONES

On this slide we show the interference pattern that emerges from the interference of two identical scalar wave beams from translator/transmitters.

Now note we are interfering two zero-vector waves, each containing a substructure. In this case we assume similar substructures.

What happens now is that we get a similar interference pattern, but with some startling differences.

First, we put in the zero-lines. These represent in-phase conditions for the substructures.

. Second, the substructures themselves interfere, and form energy in the grid zones between intersecting zero-lines.

In the perfect case, however, this energy cannot radiate away. It is physically created and trapped as if in an "energy bottle." We are CREATING this energy at a distance, and storing it in a bottle. If the bottle were perfect, we would continue to accumulate energy in the grid zones -- notice we are continuing to pour energy into both transmitters, and there is zero ordinary energy anywhere else except in the grid zone "bottles."

If we turn off the transmitters smoothly, we extinguish the energy in the bottle completely. It is sucked right out of there, and the vacuum returns to its normal condition.

If we erratically and nonlinearly turn off the transmitters, we destroy the bottle and "dump" the energy suddenly. This is exactly like a sudden EMP from a nuclear weapon. A sudden pulse of energy is freed to radiate and interact in the region where the bottle was.

The high altitude booms off the East coast of the U.S. a few years ago were made in precisely this fashion, from such a weapon being adjusted and calibrated in the Soviet Union.

Next Slide

The Tom Bearden Website

SLIDE 44.

CREATING ENERGY AT A DISTANCE

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

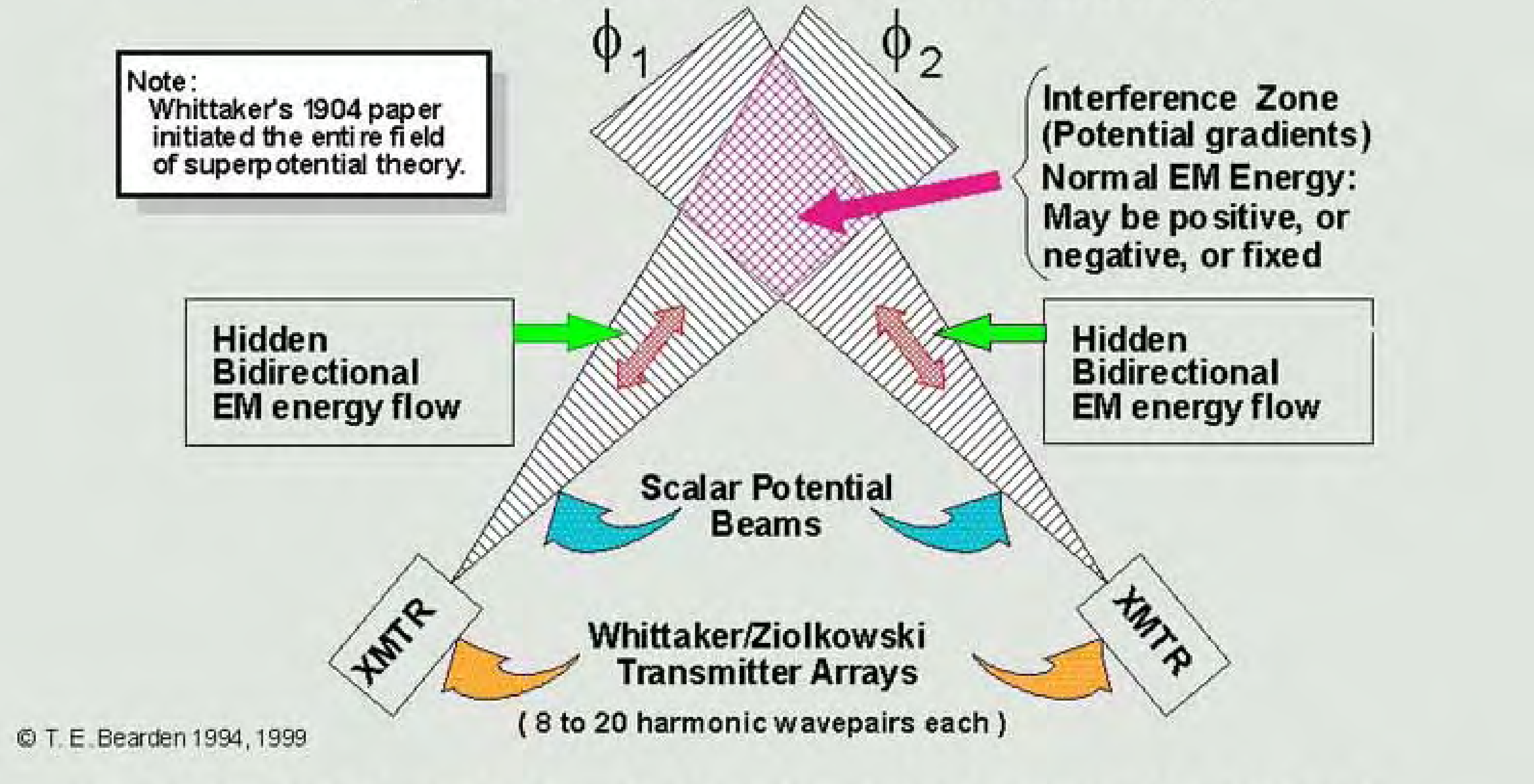
ϕ_1
ϕ_2
Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.
Interference Zone (Potential gradients)
Normal EM Energy: May be positive, or negative, or fixed
Hidden Bidirectional EM energy flow
Hidden Bidirectional EM energy flow
Scalar Potential Beams
XMTR
XMTR
Whittaker/Ziolkowski Transmitter Arrays
(8 to 20 harmonic wavepairs each)
© T. E. Bearden 1994, 1999

Figure 5. Scalar potential interferometry (between the two sets of bidirectional longitudinal EM wavepair functions) produces all EM force fields and waves.

But let us look at one way to do this sort of thing more efficiently.

Here we show two scalar transmitters (that is, two translator /transmitters) which form narrow beams, and which cross those beams at a distance.

In the crossing zone, scalar interference is established, making an energy bottle and producing trapped or locked in energy.

If I have a physical target in that crossed zone, I can literally "fry" the target, for I get all the power inputted to the transmitters contained in the bottle zone. I don't have any square law losses.

I can heat a metal object white hot at a distance, for example. Because all real bottles are slightly imperfect, the real bottle is leaky and some photons escape. I can observe the white hot steel in the bottle, and I can record and measure the temperature from the escaping photons, by simple radiometric methods.

Now if I smoothly turn off both transmitters linearly, all the energy in the bottle disappears immediately. I can immediately reach in and pick up a black, cool piece of steel. And that cannot be done with ordinary inductive or radiowave heating.

That has been done in a laboratory on the North American continent.

Next Slide

The Tom Bearden
Website

SLIDE 45.

NONLINEARITIES CAUSE SHIFTING [PARTIAL TRANSLATION]

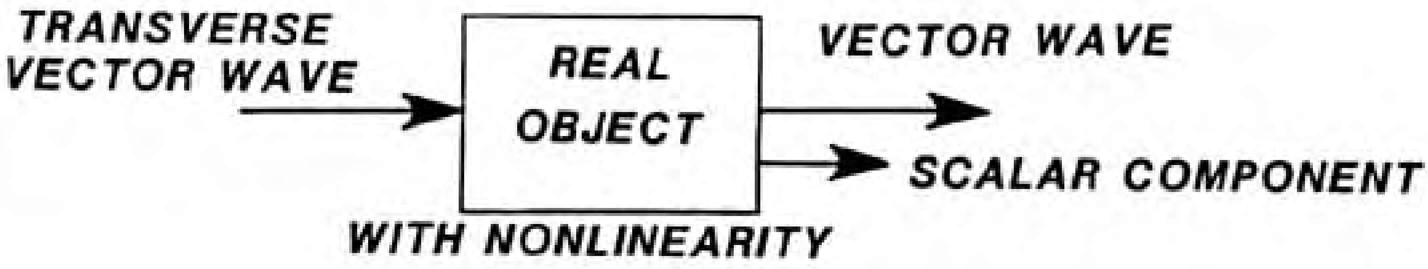
TRANSVERSE
VECTOR WAVE
REAL
OBJECT
WITH NONLINEARITY
VECTOR WAVE
SCALAR COMPONENT

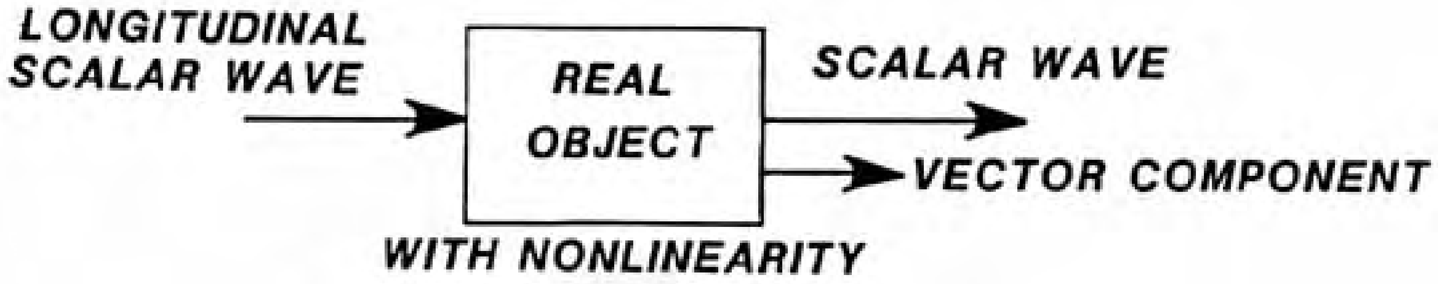
LONGITUDINAL
SCALAR WAVE
REAL
OBJECT
WITH NONLINEARITY
SCALAR WAVE
VECTOR COMPONENT

Here we show the results of electrical nonlinearities in an object which encounters a true translated (non-hooking) scalar wave. All objects have some electrical nonlinearities in them, and so all act partially as translators.

This means that any real object creates a small scalar aura around itself from its continual temperature interactions -- from absorbing and radiating ordinary EM waves. In addition, the object normally is subject to undetected scalar wave radiations, and so it also produces a small aura of transverse vector waves around itself by translation.

All real objects accomplish at least a little bit of translation.

Next Slide

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 46.

THE SCALAR AURA

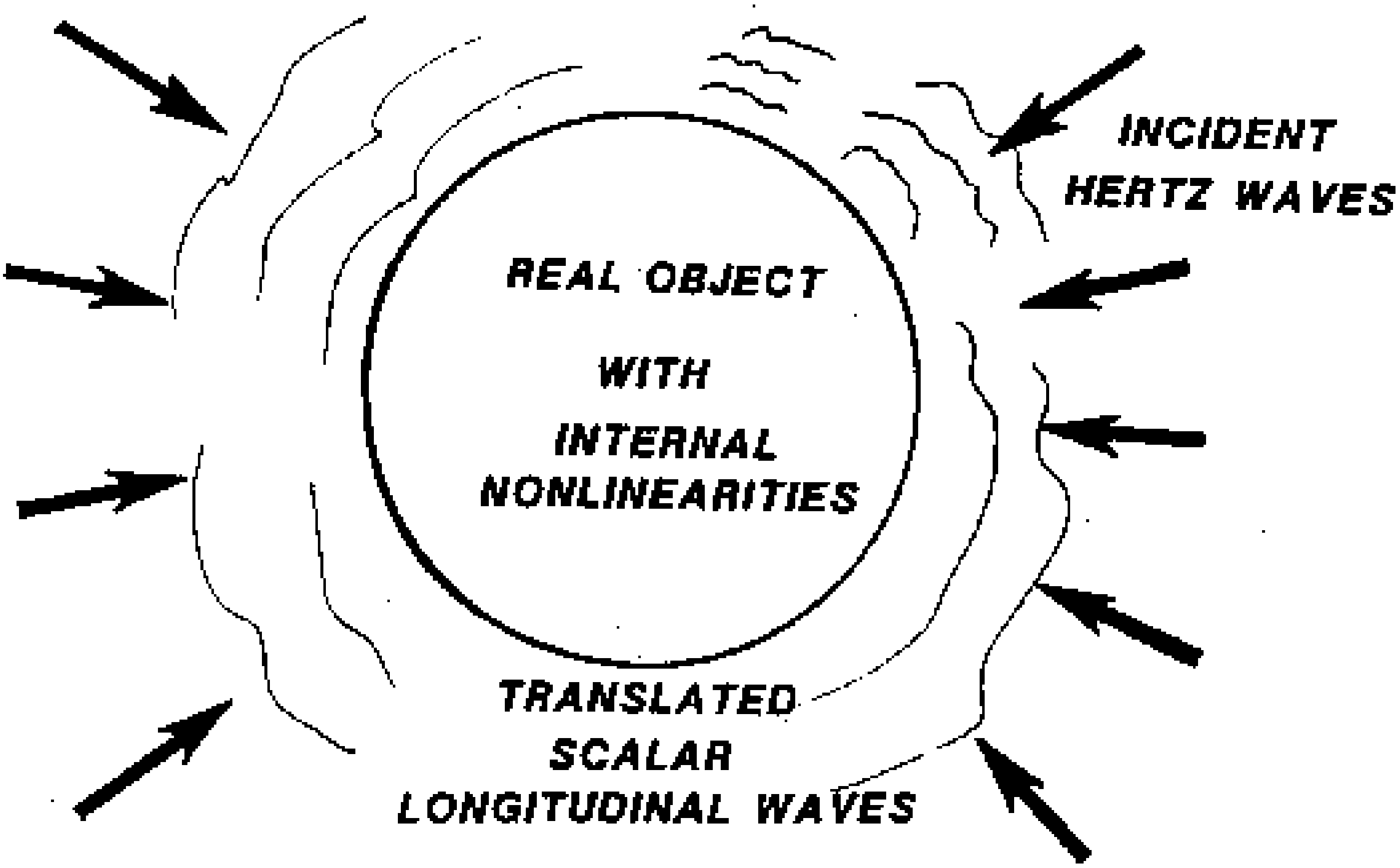
INCIDENT
HERTZ WAVES
REAL OBJECT
WITH
INTERNAL
NONLINEARITIES
TRANSLATED
SCALAR
LONGITUDINAL WAVES

We examine now an object with temperature, that is, a normal object continually absorbing ordinary EM radiation and continually emitting EM radiation.

The object has a small fraction of its incident energy expressed as a scalar aura around it, because of translation. Note that normal detectors will not detect this aura, but special two-stage interferometric detectors will detect it. Also note that the object replenishes- the ordinary radiation flow by its translation of some of its incident scalar radiation into the normal or "transverse" (so-called) radiation.

The entire internal pattern -- all interactions and functions of the body -- are precisely present in its scalar aura, down to the finest detail. This makes an x-ray pale by comparison.

True psychics actually use the nonlinearity of their nervous systems (note the avalanche ion discharges across dendrite gaps) and the interferometric aspects (note two cerebral hemispheres, connected by a corpus callosum) to detect the scalar aura. This is a physical fact and not esoteric delusion. To control this, however, is normally as difficult as learning to walk a high wire and do acrobatics on it. There are not many high wire walkers in the world, but it certainly is a skill within human control capability and purview.

T. Galen Hieronymus was absolutely correct when he stated that the entire internal working pattern of an object is radiated by that object in terms of an extraordinary energy, which he called "eloptic" energy. It is actually translated scalar wave energy, and entirely consistent with present EM theory when the glaring errors in its foundations are corrected.

Shortly I hope we will see instrumentation which can directly examine the scalar aura in the finest detail, seeing every illness, and simply setting dials to radiate the body with gentle, harmless scalar waxes to directly correct illnesses.

Next Slide

The Tom Bearden Website

SLIDE 47.

VIRTUAL FLUX AND MAGNETIC FIELD [LINEAR CASE]

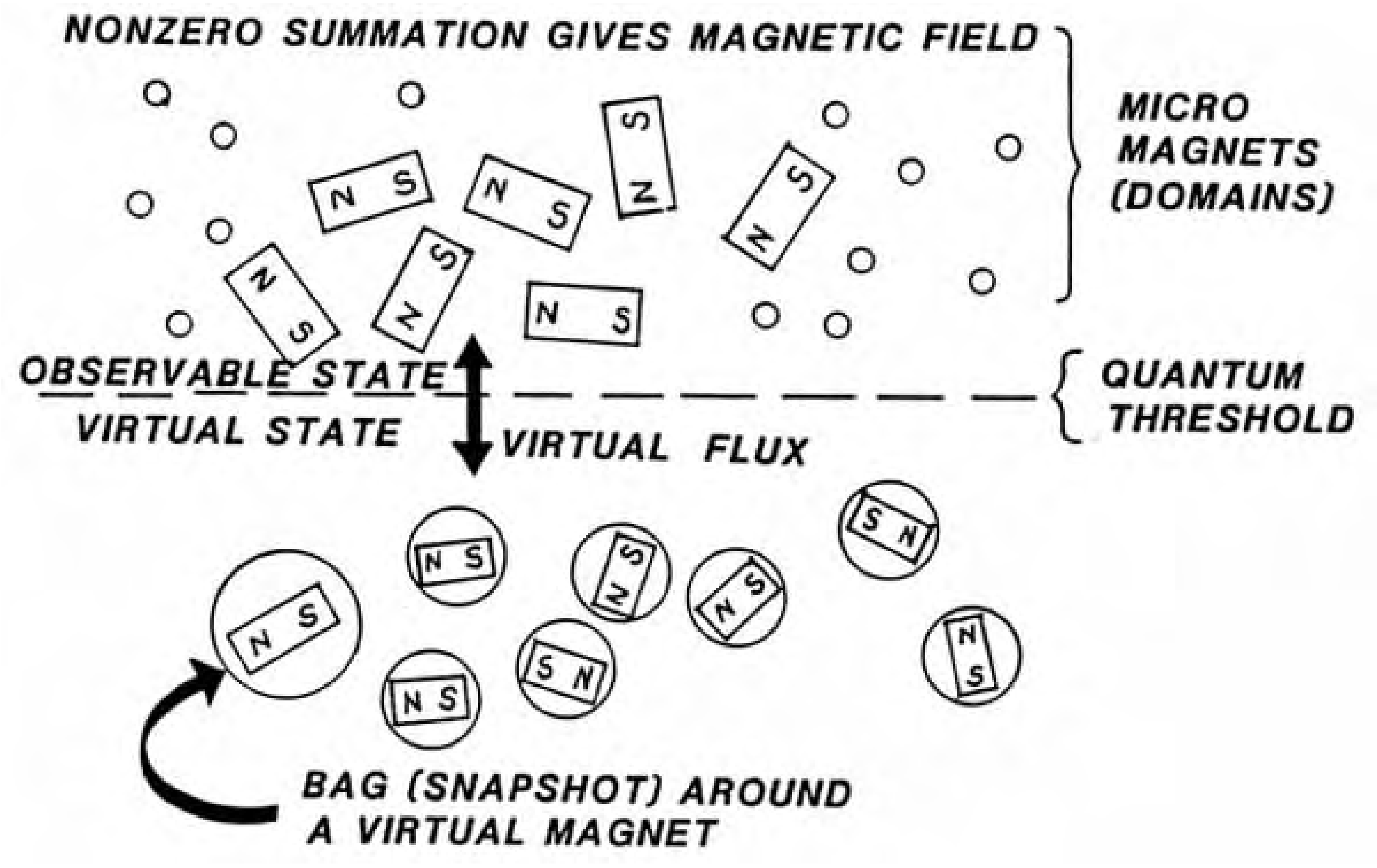
NONZERO SUMMATION GIVES MAGNETIC FIELD
MICRO MAGNETS (DOMAINS)
N
S
OBSERVABLE STATE
VIRTUAL STATE
VIRTUAL FLUX
QUANTUM THRESHOLD
BAG (SNAPSHOT) AROUND A VIRTUAL MAGNET

Briefly, let's look at some nice things we can do with magnetism. Specifically, let's find out how to make magnetic monopoles and use them to do things in the laboratory. This again has been done in a laboratory on the North American continent.

First, about all the present science can tell us about a magnet is that "a big magnet is made of littler magnets." And if we examine one of the smaller magnets, it's made of even smaller magnets. This is like saying big dirt piles are made of smaller dirt piles, and that explains what dirt is.

However, let's use that with our knowledge of the virtual flux vacuum.

In this slide, we show diagrammatically the situation for a common magnet. Note that domains in the magnetic material are themselves little magnets, and their alignment and vector summation determine whether or not there's said to be a magnetic field present. Nonzero summation states that there's a resultant magnetic vector, and hence an external magnetic field. Zero summation states there's no external nonzero resultant, hence no external magnetic field.

Actually, there's an external scalar magnetic potential field, even when the external vector magnetic field is zero. The substructure is still there.

If we pursue this "big magnets are composed of smaller magnets," eventually we reach the quantum threshold, and we have a substructure of virtual, subquantal magnets in vacuum, in the virtual particle flux. At least we conceive each little virtual particle as if it were spinning, and hence a little magnet.

We put a conceptual bag around each little virtual magnet.

In a linear situation, the north pole is as strong as the south pole, and so just outside the bag, the poles cancel or appear zero. Since the virtual bag appears to be zero length to an external macro observer, the poles seem to be directly superposed on top of each other, yielding no pole at all to the observer.

However, you and I now know that both poles are still in there, in the virtual substructure, and we certainly have remaining with us a translated scalar magnetic field.

There's a virtual flux to and from each observable particle of charged mass in the observable state as shown, but this flux is now scalar in virtual magnets, except for nonlinearities in the structures above the quantum level. Thus accelerated portions -- atoms with electrons in whirling orbit, spinning electrons, protons, etc. -- possess nonzero ordinary vector magnetic fields by translation. Again, notice the successive interlocking levels of reality, all the way from deep in the virtual state into interlocking levels in observable state.

Next Slide

The Tom Bearden Website

SLIDE 48.

PRODUCING MAGNETIC MONOPOLES

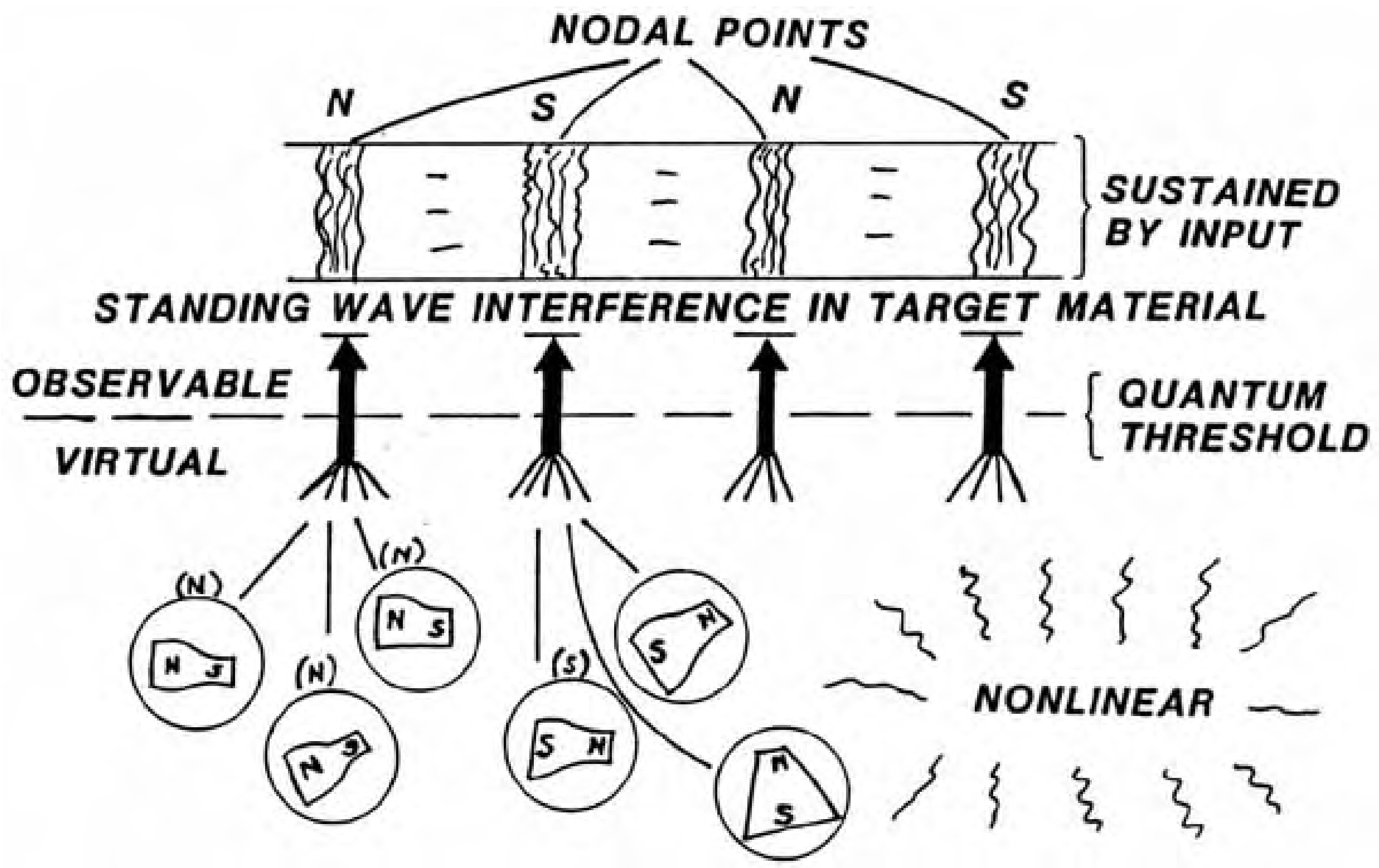
NODAL POINTS
N
S
N
S
SUSTAINED BY INPUT
STANDING WAVE INTERFERENCE IN TARGET MATERIAL
OBSERVABLE
VIRTUAL
QUANTUM THRESHOLD
(N)
N S
(N)
N S
(N)
N S
(S)
S N
S N
N S
NONLINEAR

Now let us look at a non-linear case of the same sort of thing. Now the bags are still around the little virtual magnets, but they are nonlinear. This means that, to an observer outside the bag, one pole seems bigger or more powerful than the other. In other words, to the quantum observer, this bag appears to be a magnetic monopole at a point.

In the flux still on and off each observable spinning charged mass, we now have a steady component of "monopoles". If we have a standing scalar wave present in a physical material in which the nonlinearity exists -- and the scalar waves can even be PRODUCING that nonlinearity -- we will have nodes available at which the monopoles will congregate and emerge and interact.

North monopoles will congregate at one node, while south monopoles will congregate at the next, and so on in alternating fashion.

That means that, at anyone node, monopoles of the same kind are steadily being "deposited" in the material. These monopoles strongly repel each other, and so the material at that node is increasingly stressed in a tensile fashion.

Eventually the material will be torn apart at the node,
stress relieving the situation. Movement of the material will release the nonlinear condition, stopping monopole production and deposit. However, at a break node, the same kind of magnetic pole will appear on each side of the break.

That is, the breaks will be N-N, S-S, N-N, S-S, etc.

An ordinary magnetic does not do that when it breaks. Instead, it breaks N-S, N-S, etc.

Next Slide

The Tom Bearden Website

SLIDE 49.

BREAKING METAL WITH MAGNETIC MONOPOLES

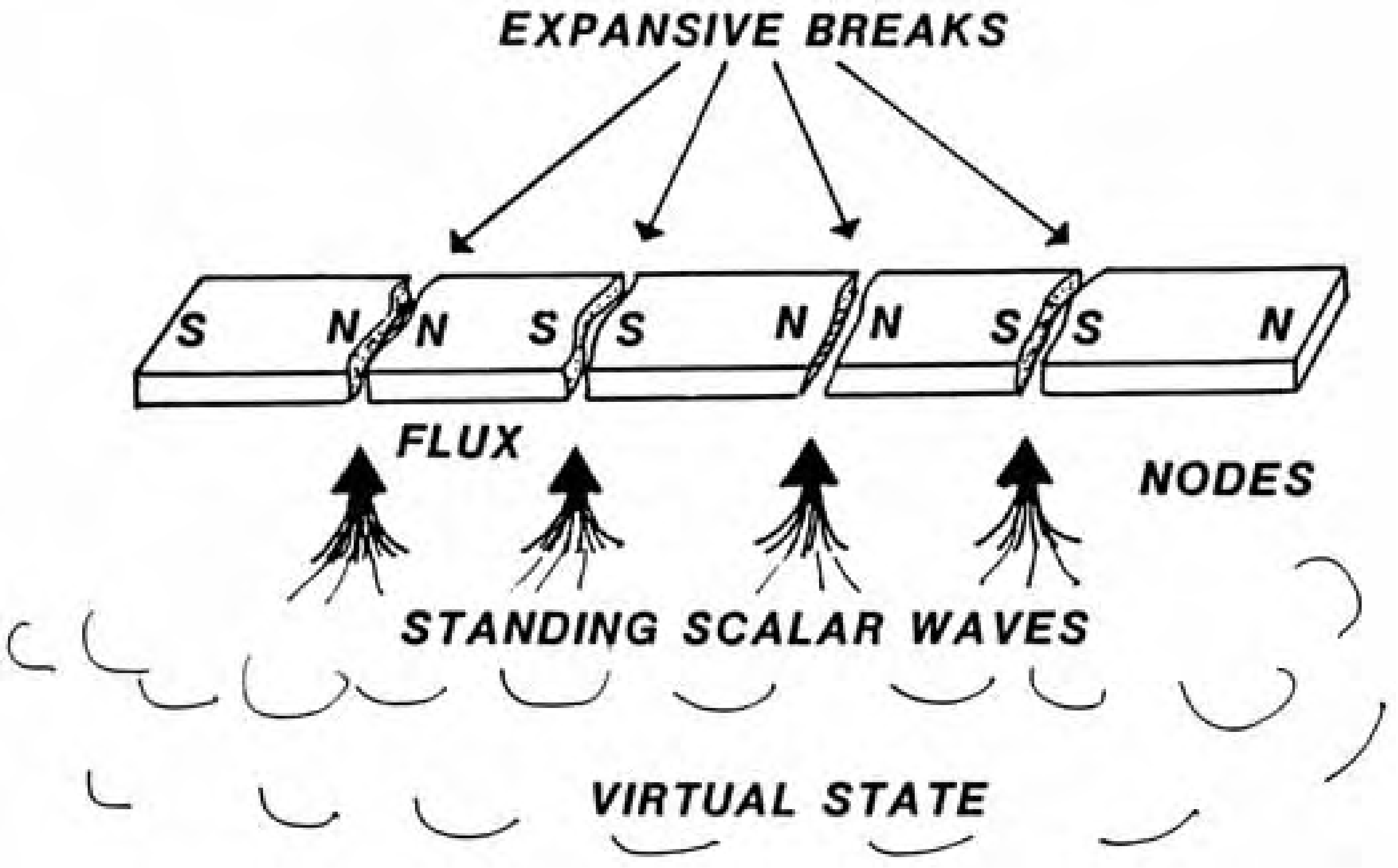

Again, metal has been broken in a North American laboratory utilizing the production and deposit of magnetic monopoles in this fashion.

On this slide we show what was done.

An ordinary bar of steel was exposed to standing scalar waves in a nonlinear situation, depositing magnetic monopoles at the wave nodal points in the metal bar. The breaks occurred as shown on this slide.

Much of what I am presenting is based on certain experiments accomplished in experimental laboratories here and abroad. Most of the mechanisms have actually been produced in the lab. However, these are proprietary labs and I cannot violate trust and reveal the precise construction details utilized. In some cases I have not seen the precise constructions, but only the direct results.

But ironically, the production of magnetic monopoles -- which has fiercely resisted the normal scientific approach -- yields readily to the new approach.

Next Slide

SLIDE 50.

CREATING AN INERTIAL FIELD (ANTIGRAVITY)

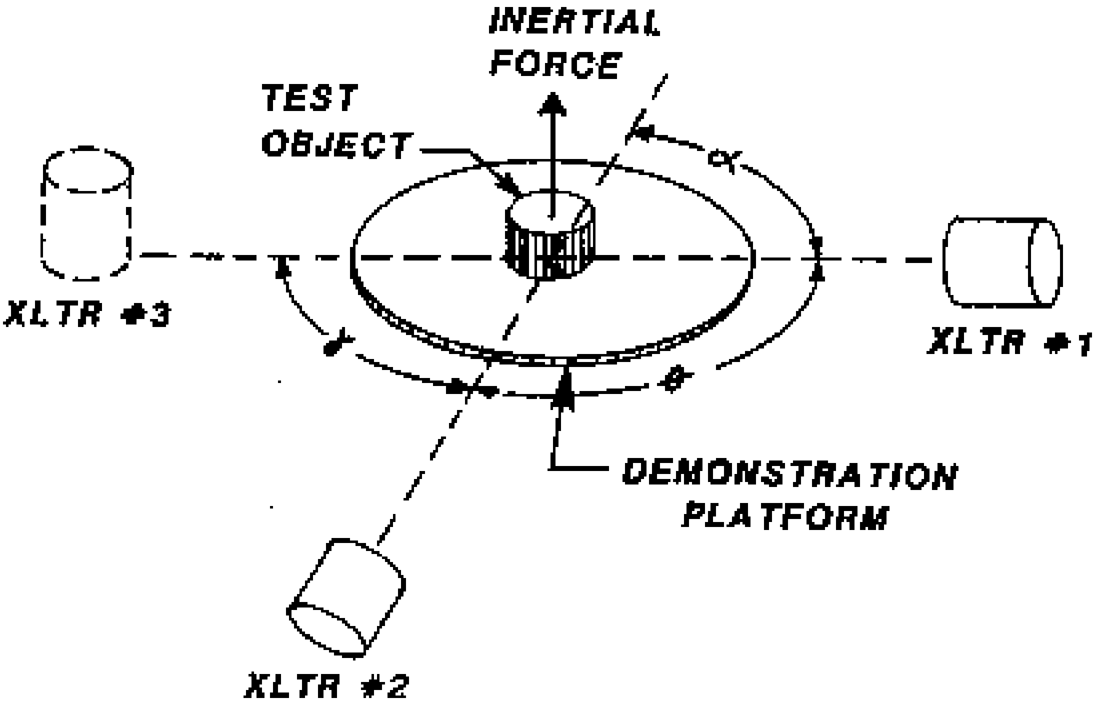

We now show schematically how to produce an inertial field, or antigravity. Also an inertial space drive can be produced in this manner.

This again has been done in a North American laboratory. In fact, inertial fields have been produced in more than one lab here in North America.

In the setup shown here, three translator/ transmitter projectors are used and oriented so as to give three-dimensional interference, or interference along all three spatial axes.

By controlling the substructures of the three projectors all together, one can control the precise phasing in all substructures, with respect to a common time reference.

By this means, a unilateral force can be created of each and every charged mass particle (that is, on each proton) in the nucleus.

Since we know that the neutrons and protons in a nucleus of an atom are continually interchanging, one into the other, back and forth, all of the nucleons can be treated as "diffuse protons" for mass purposes, and scalar wave interaction purposes.

Thus if we have the correct pattern to affect protons, we can control the phases of the substructure waves of the projectors so as to produce unilateral, parallel forces on and of each spinning proton in the nucleus.

This means we can produce unilateral thrust in an object exposed to the triad interference of the projectors.

In short, this produces antigravity. If sufficient force is produced, the test object in the interference zone simply lifts itself off against earth gravity. This has been done in North America.

If we phase the projectors so that counterbalancing forces are produced on each proton, we can either make the proton "lighter" inertially or "heavier."

In fact, we could "freeze" a common pistol so that it could not be moved and its trigger could not be pulled, simply by sufficiently increasing its inertia. This was done in experimental demonstrations in Toronto, Canada in 1969, and witnessed by the Chief of the Homicide Squad of the Toronto police. It was his pistol that was frozen for the demonstration.

If the projectors themselves are on board a space capsule, for example, they can produce such a unilateral thrust in and of a central mass or disc attached to the of the craft. By this means, an inertial spacedrive can be built.

Next Slide

SLIDE 51.

EFFECTS ON HUMANS

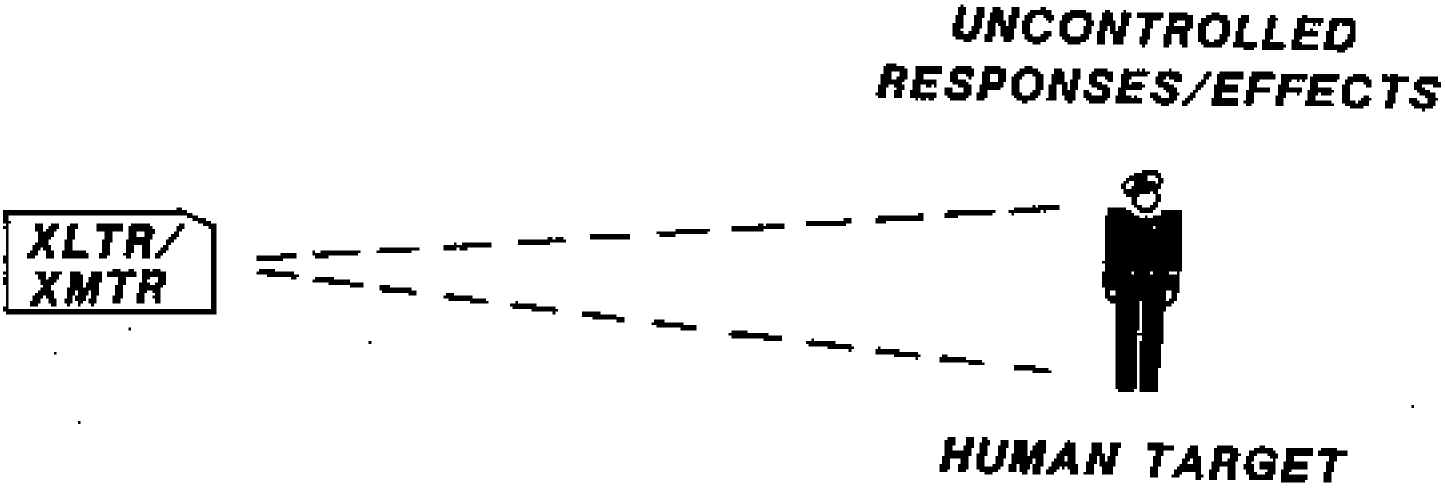

Now let us look at some of the effects on humans, when radiated with a true scalar wave beam from a translator/transmitter.

First, the human nervous system is highly nonlinear. Essentially, across billions of synapses, sudden ion discharges -- analogous to spark discharges -- are continually or periodically occurring. Brain waves, for example, are not waves along wires; instead, they are waves of avalanche discharges across tremendous numbers of nerve cell gaps.

The central nervous system is thus highly nonlinear, and the gap firings of the nerve endings act as translators.

These translators partially translate incident scalar waves, so that additional transverse matter waves -- or ordinary energy inputs -- occur in the ion discharges.

Thus additional energy is being added to the nervous system by incident scalar waves.

Many long-term effects of the environment on the body are gradually kindled in this fashion, even over a period of years.

For a stronger dosage of scalar waves, the nervous system experiences "jamming" of its signals. This leads to four basic "noise jamming" effects.

If a small dose is absorbed, the interference is slight. One feels as if one had taken a whiff of anaesthesia -- an unreal, dreamy feeling.

If a slightly larger dose is absorbed, the interference is moderate. Now the single channel system -- the motor system -- goes, and the individual is paralyzed. Mentally he is barely conscious, in a sort of hypnogogic, waking dream state. Any tendency toward epileptic seizure or other nervous disorder may surface here, as may heart palpitations or even heart failure.

For a heavier dosage yet, the interference is now strong. Now the multiple systems start failing. Consciousness is lost and the individual is in a deep coma. In addition, seizures and convulsions, and loss of control of some body functions, may occur. The autonomic nervous system is keeping the heart beating with difficulty and keeping the body gasping for breath.

For a very strong dosage, the entire nervous system fails and death results almost immediately. This is quite similar to the effects of nerve gas.

These are the four major symptoms of scalar wave radiation of the body and translating absorption by the nervous system in an uncoordinated fashion.

Next Slide

The Tom Bearden Website

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 52.

INITIATING SPECIFIC BIOLOGICAL EFFECTS

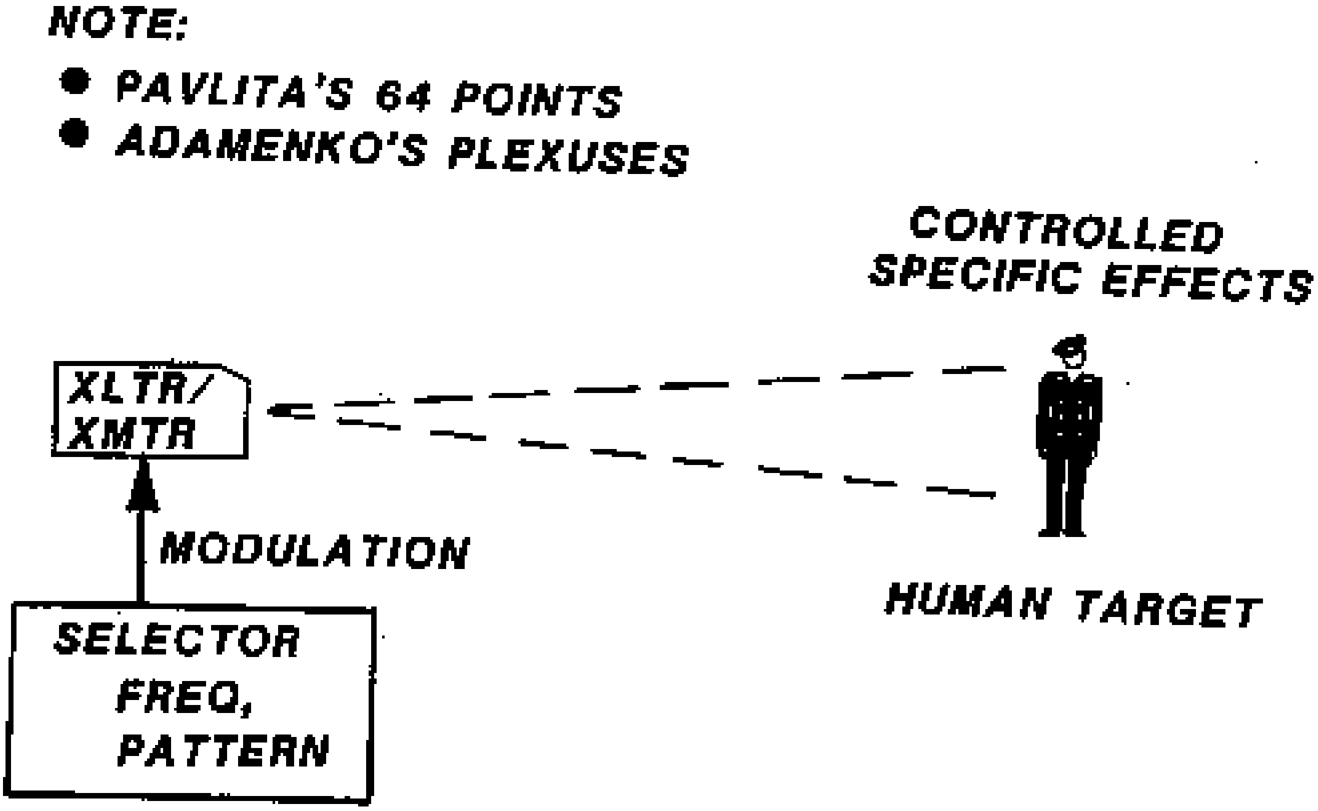
NOTE:
● PAVLITA'S 64 POINTS
● ADAMENKO'S PLEXUSES
CONTROLLED SPECIFIC EFFECTS
XLTR/ XMTR
MODULATION
HUMAN TARGET
SELECTOR FREQ, PATTERN

If we now add precisely patterned modulations to the substructure waves of the scalar zero vector beam, we can induce deliberate patterns in the dosage absorbed by a biological system.

As Kaznacheyev's experiments showed, any death or disease pattern can be induced upon an electromagnetic carrier -- specifically, upon the near-ultraviolet as an example. Some of his experiments were duplicated at the University of Marburg in West Germany, utilizing infrared carriers. These patterns can be dumped into cell cultures, eventually kindling the pattern of death or disease in the target cells.

The same can be done to a human from an external source. The major entry point is the acupuncture system, which reacts to scalar waves and transfers them to the nervous system for translation.

Pavlita, Czechoslovakian engineer, has built "psychotronic generators" for over 30 years that affect the human body. Adamenko, Soviet physicist, found that the acupuncture points formed groupings or plexuses that were frequency sensitive. In other words, certain frequency bands affect one plexus more than others. Further, these acupuncture points are known to be connected to effects produced at remote locations in the body.

Pavlita stated he had found 64 different sensitive "points" on the body, and had succeeded in building a generator for each. If by his points he meant Adamenko's plexuses, then frequency patterns can be created that affect certain acupuncture groupings -- and thereby certain specific body locations and effects -- more than others.

It thus appears reasonable that, by placing the proper frequency pattern in the modulations of the substructure waves feeding into the translator/ transmitter, the scalar ray or beam can be made highly selective as to body part affected and the effect generated or kindled in the body.

By this means, specifically tailored attacks on the body can be created.

Strong emotions such as sheer terror, overpowering rage, blind panic, etc. can be induced. Or physical effects such as epileptic seizure, loss of control of internal functions, loss of muscular control, severe weakness, specific failure of an individual organ such as the heart or liver, or blackouts and comas can be induced. In addition, patterns for specific toxins and diseases can be induced. With a strong absorption, these changes can be effected quickly. Imagine bubonic plague developing in 30 seconds, for example, or the effects of a cobra bite developing in 10 seconds.

If the individual is placed in crossed interference beams (that is, in the interference zone of a scalar interferometer), strong effects can be produced essentially instantly, including instantaneous death.

Of course BENEFICIAL effects can also be accomplished. As instrumentation is developed, one should be able to scan the aura, do a computerized readout to show the different medical problems either developed or developing in the body, have the computer compose a "cancellation or offsetting pattern," and radiate the body with a totally beneficial pattern, customized for that particular condition at that particular time. One should be able to quickly cure cancers and leukemias, for example, even of the brain and of the bone marrow. One may even be able to quickly reverse the effects of aging; at least there is no foreseeable reason why it should not be possible.

The new technology is a very powerful tool, and it can be used for either good or bad, depending upon the intent of the developer and the user.

Next Slide

The Tom Bearden Website

SLIDE 53.

GROUND-TO-AIR USE

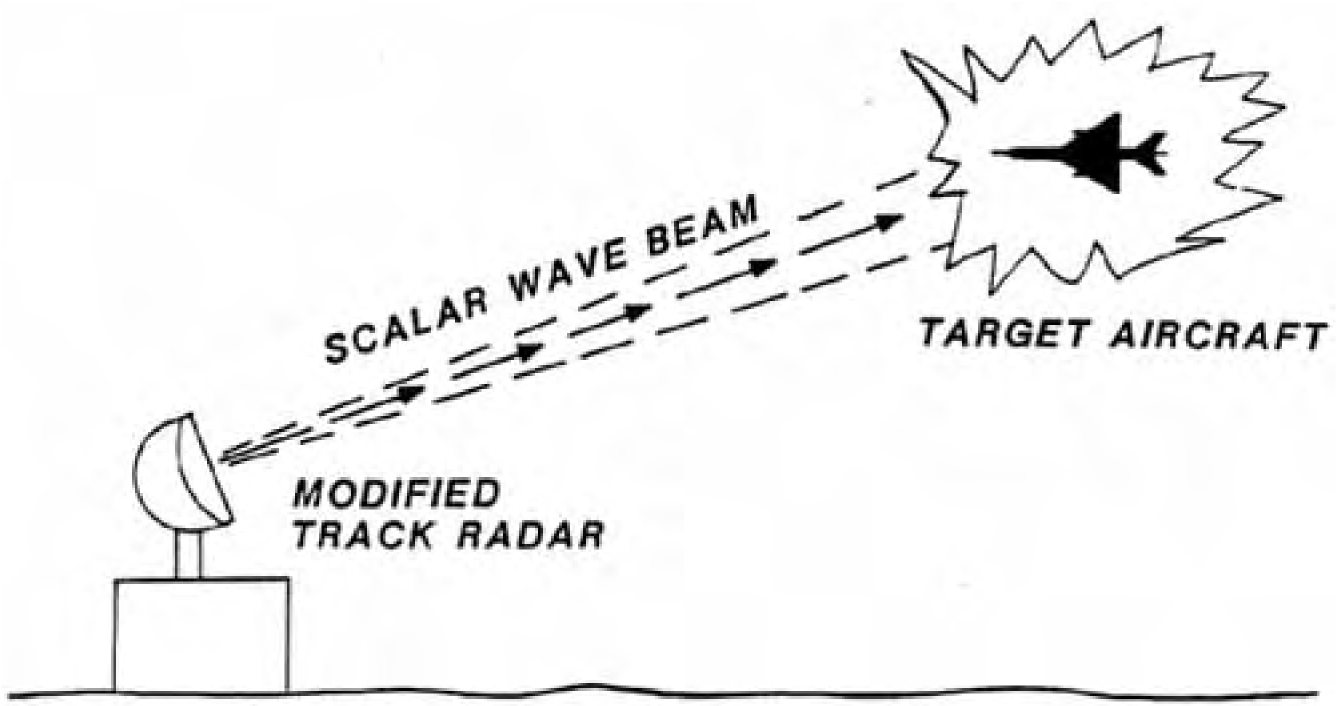

Here we show another use of the scalar beam technology already developed by foreign powers.

A modified track radar -- either continuous wave or pulsed -- transmits a scalar wave beam. It radiates an incoming hostile aircraft, keeping the beam on it, and kindling energy in the nonlinear electronic circuits of the aircraft by translation. Communications, navigation systems, avionics systems, electronic fuze systems, ignition and control systems, airborne radar, missile armament electronics, etc. are all vulnerable to this translation kindling of internally jamming energy from the incident scalar beam.

The electronics can be knocked out and the aircraft knocked down.

The pilot himself is also vulnerable because of his highly nonlinear central nervous system.

Some or all of the eight F-111's mysteriously

lost in
Vietnam may have been bagged by modified SA-2
FANSONG radars, using this effect.

Next Slide

The Tom Bearden
Website

SLIDE 54.

TACTICAL AIR-TO-GROUND USAGE

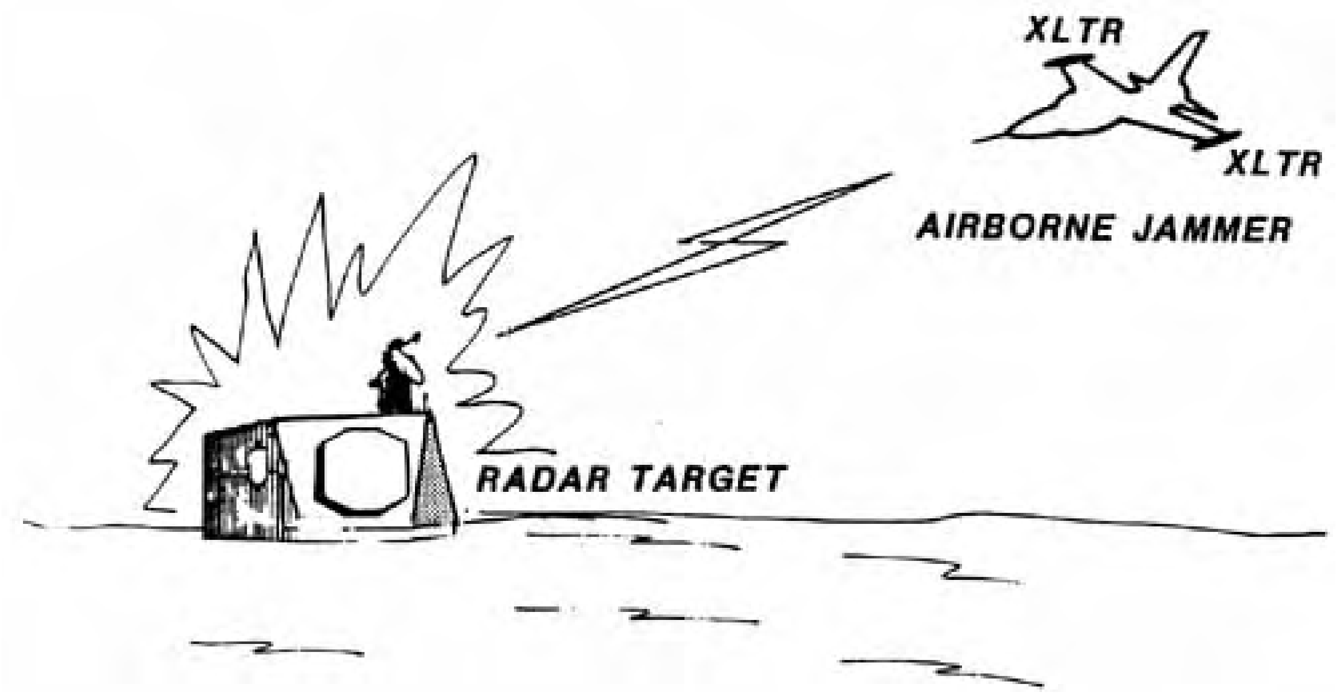

As shown on this slide, the technology is also usable from aircraft against ground targets, particularly ground-based radars and communications systems.

As shown, the aircraft has a modified jammer, able to emit a beam of scalar energy from one or more translator/transmitters.

By simply training a beam on the radar, eventually the radar can be knocked out. That is the simplest use.

A more sophisticated usage is provided by more than one translator/transmitter, acting as a scalar interferometer.

In this mode, each translator/transmitter emits multiple scalar carrier frequencies. Two Fourier expansions are established, so that a three dimensional "ball" or sphere of interference-zone energy is established. By pulsing or phasing, this ball is moved to strike the location of the radar, delivering a large burst of energy to it and knocking it out. Alternatively, the energy bottle is "dumped" when it reaches the targeted radar or its near vicinity.

A still better way is to fire one Fourier scalar pattern directly at the target radar, then fire a second pulse faster so that it overtakes the first directly upon the target. (The ensemble of such a pattern acts as a soliton, and its velocity is controllable by controlling the amplitude of the pattern generated).

When the second scalar overtakes the first, a sharp interference is formed, with a fast energy bottle, which is then nonlinearly dumped in microseconds or milliseconds. In this fashion, all the energy put out by the jammer is delivered directly upon and into the target in one sharp EMP pulse, much like that from a small nuclear explosion. The target is destroyed.

Note that such a weapon can be utilized against tanks, vehicles, explosives, missiles, ammunition, bridges (the airblast), bunkers (the blast can occur INSIDE the enclosure), ships, planes, personnel, and targets of varied types.

The Soviets have had such secret weapons for years.

One additional country has them also.

Next Slide

The Tom Bearden
Website

SLIDE 55.

INTEGRATED WEAPONS DISPLAY (MODIFIED JAMMER)

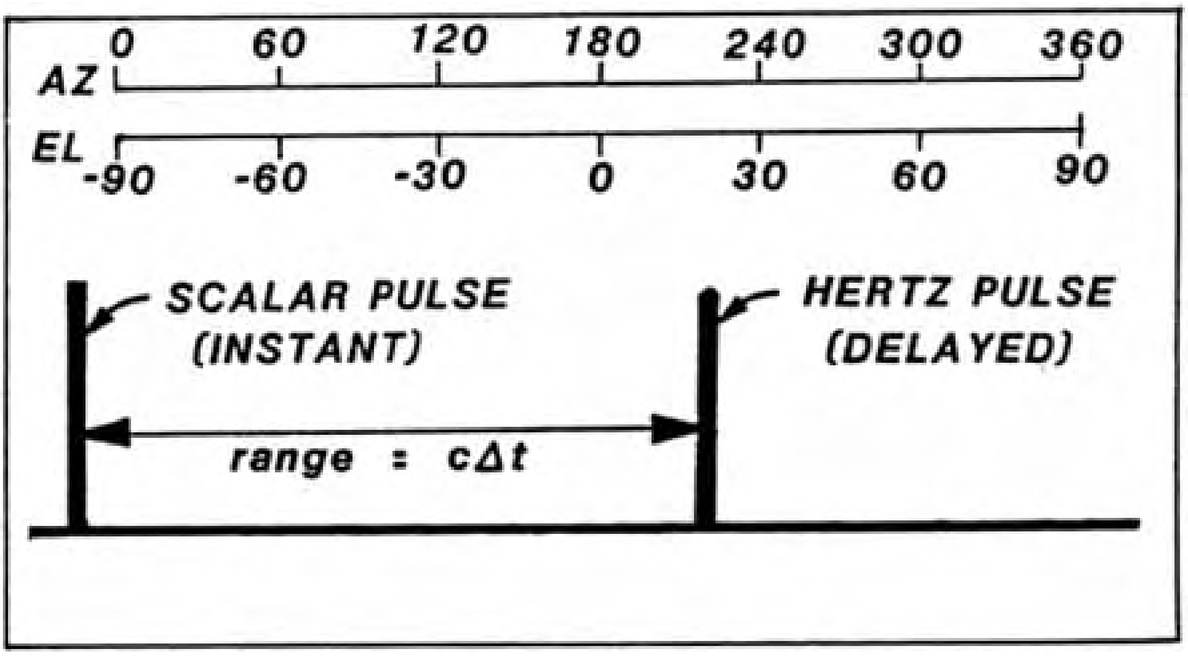

Here we show diagrammatically what a modified Soviet integrated weapons display for a modified jammer might be, to employ scalar interferometer weaponry.

As the "main bang" is emerging from the antenna of the
target radar on the ground, the phi-field from it reaches to infinity, including through the aircraft with a finite magnitude. In the a1rcraft, two channels act in interferometer fashion as a scalar detector for phi-dot. The main bang creation (phi-dot) is detected, and displayed as the leftmost pulse.

At that time, a computer time-base starts clocking off microseconds.

At some time later, the third channel, acting as a normal receiver, receives the ordinary "Hertz" pulse from the radar, which has traveled to the aircraft at the speed of light.

The computer converts elapsed time between the scalar pulse and the Hertz pulse to range.

Azimuth and elevation angle to the radar are obtained by normal angle tracking.

At this time, the computer has the exact coordinates of the radar, allowing for motion of the aircraft during the lapse time.

The computer switches the interferometer into transmit mode, and arranges amplitude patterns to target the radar.

It fires the first scalar pulse, then fires the second so as to create the full EMP directly upon the target.

The EMP flash occurs, destroying the radar.

At that time, another target can be processed.

In a more advanced weapon, scanning may be used to detect, locate, target, and engage multiple targets, including incoming air-to-air and surface-to-air missiles, one after another.

The same weapon is usable against a variety of aerial and ground based targets.

The major weapon is the electronic emitter, not a missile or ordinary ordnance.

Hordes of such special "jammer aircraft" now become a potent force on the battlefield, destroying all before them with annihilating ray weapons.

Star war is a reality now.

The High Frontier is already seriously obsolete.

Our entire defense establishment is completely vulnerable to these weapons at present, for we have not developed countermeasures or defenses.

Next Slide

The Tom Bearden
Website

TOWARDS A NEW ELECTROMAGNETICS
PART 4: VECTORS AND MECHANISMS CLARIFIED

SLIDE 56

RADAR INVISIBILITY

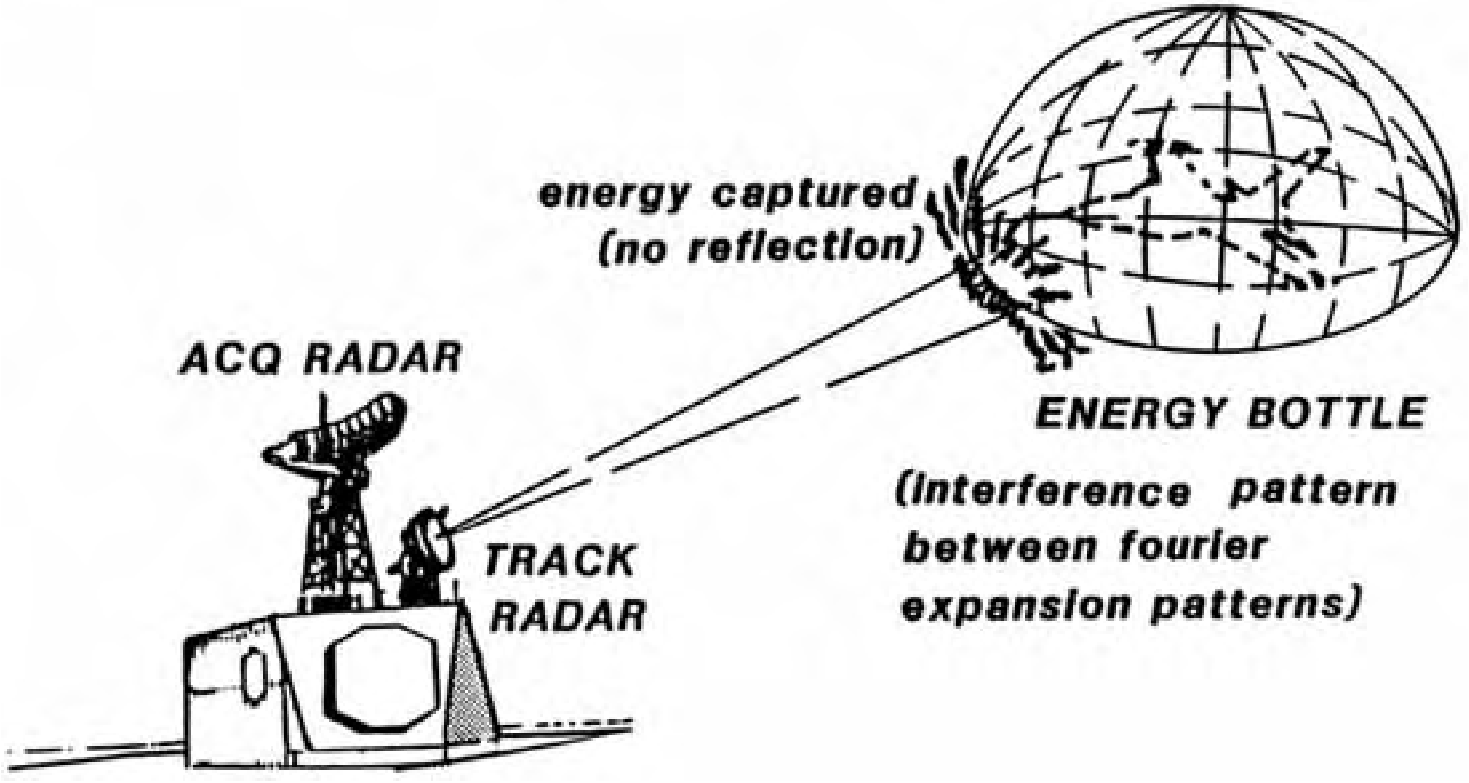
energy captured
(no reflection)
ACQ RADAR
TRACK
RADAR
ENERGY BOTTLE
(Interference pattern
between fourier
expansion patterns)

On this slide we show another effect which can be obtained. We can make an airplane invisible to ordinary radar.

Imagine now the same "jammer" type aircraft with three channels, one for use in scalar interferometry and one as an ordinary channel.

This time, a Fourier interference pattern is established around the aircraft, so this establishes an "energy bottle" around the plane. We specifically arrange this bottle in the frequency band of the hostile radar we wish to avoid. For multiple bands, we establish multiple bottles -- this just requires more frequencies or more translator/transmitters.

The beam from the ground radar now strikes the energy bottle, and its energy enters and is trapped in the bottle grids. Since the bottle is leaky, the incident energy gradually disperses and "leaks" out all around, with very little emerging energy.

The ground radar receives no return pulse or reflection.

Therefore it sees no target.

The same interferometer system is usable both in this defensive screen fashion and in the attack fashion previously discussed.

Use of this system has been demonstrated internationally.

An energy bottle has been created in a North American laboratory as well.

Next Slide

The Tom Bearden Website

SLIDE 57.

NETWORK OF VIRTUAL TRANSMITTERS

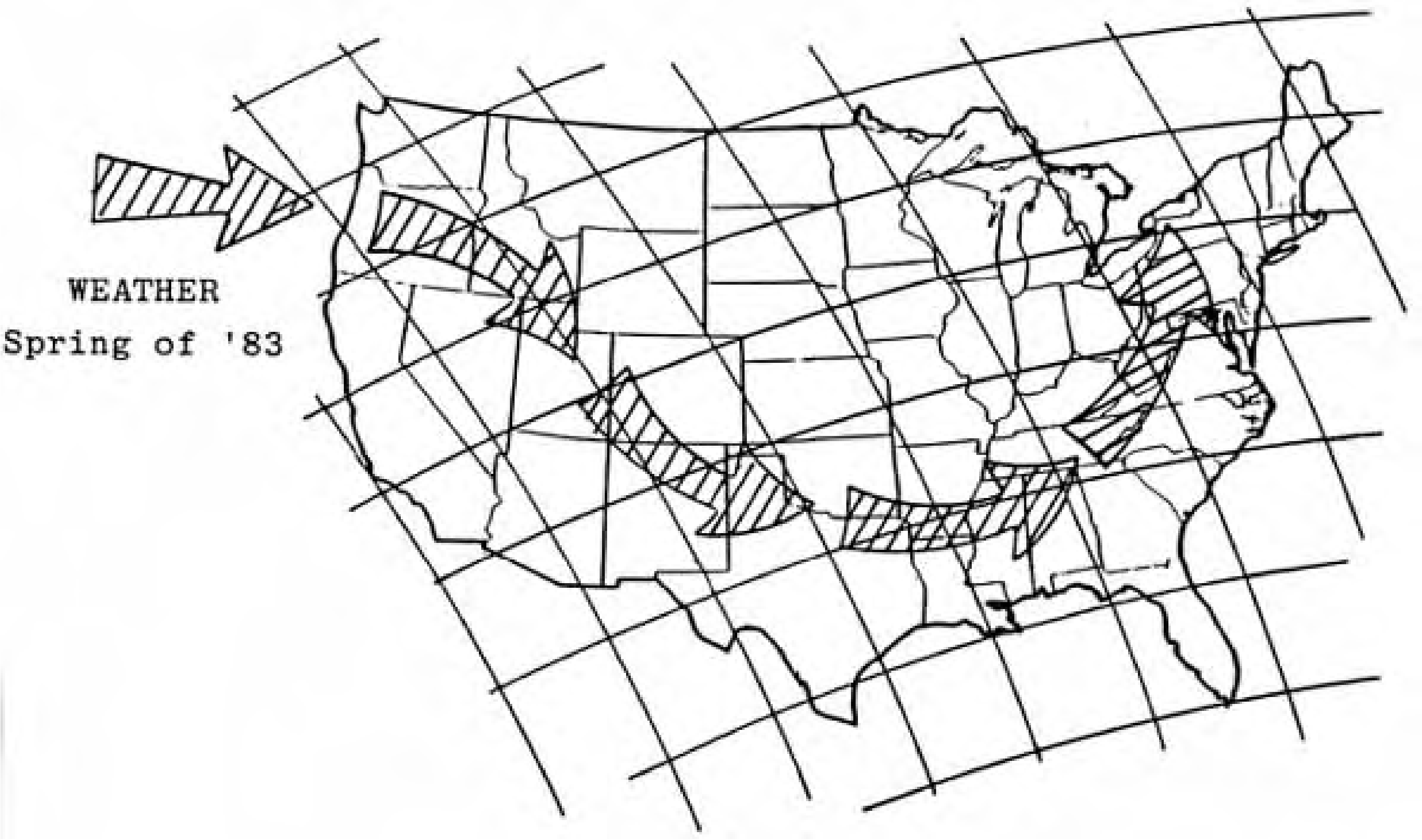

As is well-known, for years the Soviets have blanketed the world with large, powerful transmitter signals in the communications band, referred to as the "Woodpecker" signals because of the staccato noise the chirped signals make to the ear. The carriers have been modulated with known harmful bioactive ELF frequencies. At least some effects have been noted on humans from these signals, and other human effects are suspected but not proven.

In addition, evidence exists that such transmitters, utilized in the scalar transmission mode, may have been used to cause earthquakes and to control the weather. High altitude bursts off the coast of the Eastern United States and Canada, over

Timmons, Ontario, Canada, and off the coast from St. Petersburg, Florida have occurred also.

Mysterious "blackouts" of U. S .satellites -- and British satellites -- have occurred. These would be difficult to explain as the result of illumination from a ground laser, but easily explained as the result of illumination from a ground-based scalar beam or scalar interferometer.

Strange patterns of micro sky quakes and micro earthquakes occurred at various locations throughout the U.S. just before and shortly after the death of Leonid Brezhnev. These strongly appear to be connected with the adjustment of a scalar interference pattern -- a gigantic "energy bottle" across North America, as shown on this slide.

Shortly after this suspected adjustment and alignment, the underground net in the U.S. detected the emergence of numerous ELF-modulated signal carriers throughout the U.S., generally utilizing carriers in the vicinity of 40 MHz. Locations of the apparent transmitters were such places as Atlanta, North Georgia, St. Petersburg, Florida; Mississippi; Alabama; in the vicinity of Washington, D.C.; near Boston; around the Great Lakes; Michigan; in the Pacific Northwest (Washington and Oregon); and in both northern and southern California.

Researchers knowledgeable of the Soviet woodpecker signals and ELF modulations detected and reported these signals.

Physical effects were reported and correlated to the signals, including wakefulness at night, restlessness, fatigue, headaches, nausea, ringing in the ears, anxiety, disturbed dreams, etc.

What appears to have occurred is the adjustment and
alignment of a vast scalar interference pattern across the U.S. Recall that energy emerges in each grid, by phase interference between the composite substructure waves of the scalar beams.

Thus the Woodpecker "over-the-horizon radars," as accepted
by Western intelligence analysts, may have acquired a diabolical new strategic capability.

It is as if the Soviets had been permitted to build a vast network of transmitters directly on U.S. soil, for hostile employment against U.S. citizens.

Shortly after the net was aligned, a curious pattern of weather assaults on the U.S. began.

Storms assaulted California unceasingly. Then they roared across the Rockies, dipping sharply down to hit the Texas Panhandle, and bending northeast again to assault the South and the East.

The mighty jetstreams themselves were diverted and "kinked" into this looped, distorted pattern.

And the storms and floods assaulted the U.S. one after the other in this highly unusual fashion.

I maintain that the virtual transmitter net, just established by the Soviets and adjusted in on the U.S., was tested in the weather control mode. If so, it proved highly effective. Imagine what would have happened to the U.S. if it had been unleashed full force, and not just "gently tested."

And imagine what happens when disease, death, and nervous system disturbances are modulated upon these blanketing signals. The "transmitters" for the patterns are in each grid, all across the U.S.! And there are no physical transmitters there to destroy.

A counter is needed quickly. A simple one would consist of recording the signals in each grid zone, analyzing the substructure, reversing the signals and their structured patterns 180 degrees, and feeding the "cancellation signals" out in each area.

A great strategic Russian weapon of devastating power and potential is now operating directly on U.S. soil. The need for a crash countermeasures program is urgent.

In 1975 Brezhnev had introduced to the SALT talks the strange proposal that we should also consider outlawing the development of new weapons more frightful than the mind of man had ever imagined. We didn't know what the Russians were talking about. When we asked them later what weapons they meant, they referred vaguely to weapons of an electromagnetic nature.

The presentation I have just given you is part of what Brezhnev meant.

Next Slide

The Tom Bearden Website

SLIDE 58.

CONCLUSION

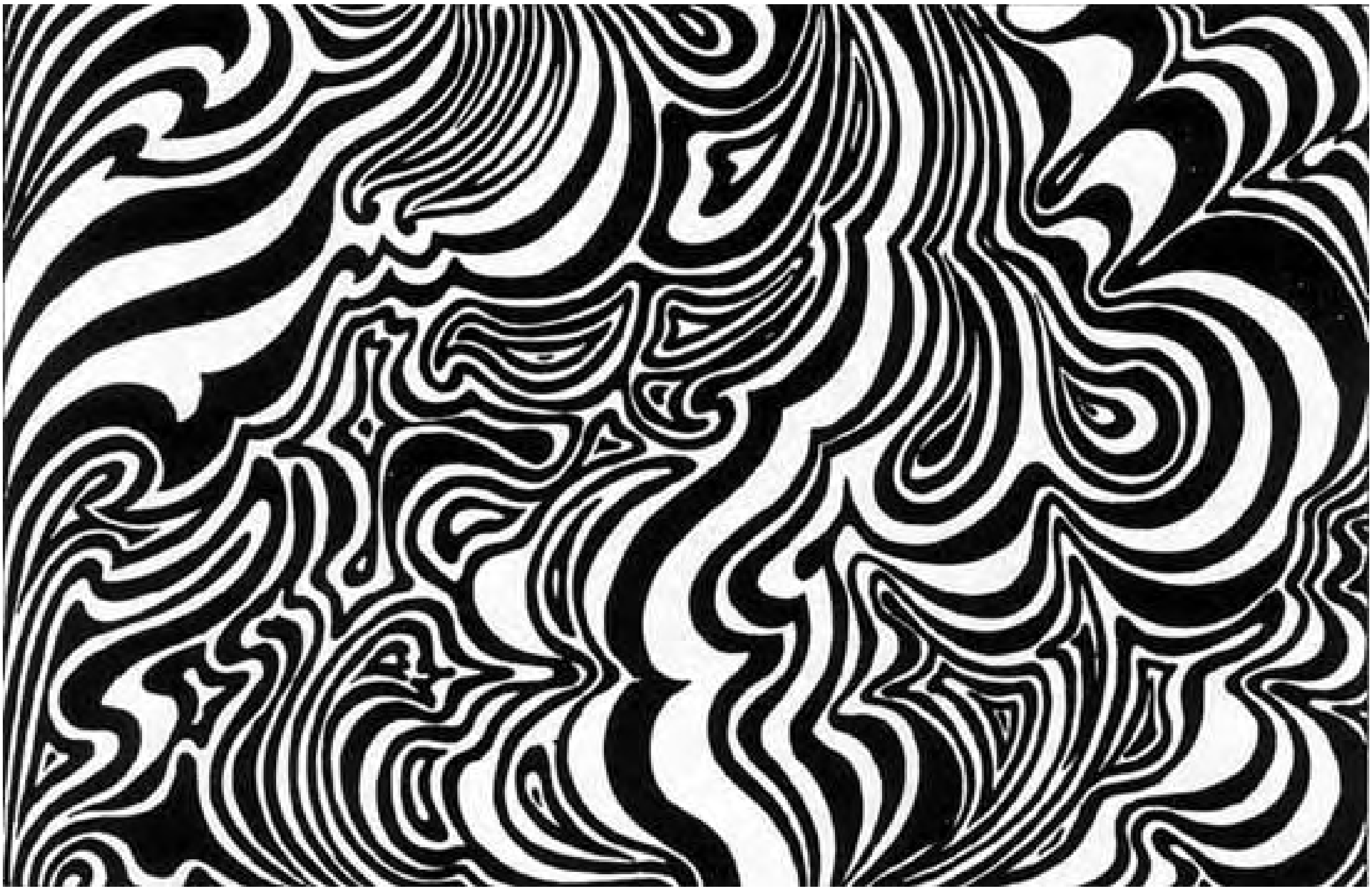

In conclusion, today I have released a number of powerful mechanisms that result from the new technological concepts. Many of these were originally discovered -- at least in rudimentary form -- by Nikola Tesla. In the West, the suppression of Tesla led to the suppression of these effects. And our knowledge of them went to the grave with Tesla.

After World War II, the Soviet Union mounted a massive campaign to obtain all the scientific literature of the West and
thoroughly digest it. Great centers were set up, and thousands

of PhD's devoted to this purpose. Gradually they synthesized this technology from the errors in EM foundations they uncovered, and from obscure, ignored papers published in the orthodox Western scientific literature.

They rediscovered the technology of Tesla, Moray, and Rogers. And they have highly weaponized the effects, and adjusted and aligned the weapons upon us.

Now the radical movements are onto part of this technology. As late as about six months ago, one such group discovered how to build a translator and a scalar beam projector. Excitedly, they planned to build them the size of a large pistol, powered by a portable battery pack, and sell them to street criminals to kill or stop the police. They planned to get rich by this illicit means. At that time, a personal friend of mine, who moves in underground circles, was attempting to purchase one of the prototypes for my analysis. Fortunately, the inventor stepped into the beam of one of his devices and was almost killed. This frightened the group, and at least for now they have desisted. But be forewarned: this technology is coming down the pipe, and the police establishments in the West are in for the shock of their lives if the scientific community does not wake up and prepare defenses.

Again I urge my fellow colleagues to experiment vigorously with these new concepts, but build equipment for the benefit of man. Insect control, disease elimination and control -- both are noble purposes. Defense against the hostile use of such devices is also a fitting purpose.

We are in the final days, and the final technology is upon us. It is a stupendous tool for either good or bad, and assuredly it will be utilized for both means. I call attention once again to the fact that my ultimate purpose is to use this technology benevolently to achieve mind linkage, and elevate man from his present brutish state to the lofty state that is his destiny and his true God-nature.

For better or worse, Tesla technology is once again loose in the world, this time for good, and on a massive scale.

Let us prepare ourselves for a wild canoe ride, for the real white water lies just ahead.

Thank you for your attention, and good evening.

LEADING EDGE

Volume 7, Issue No. 4

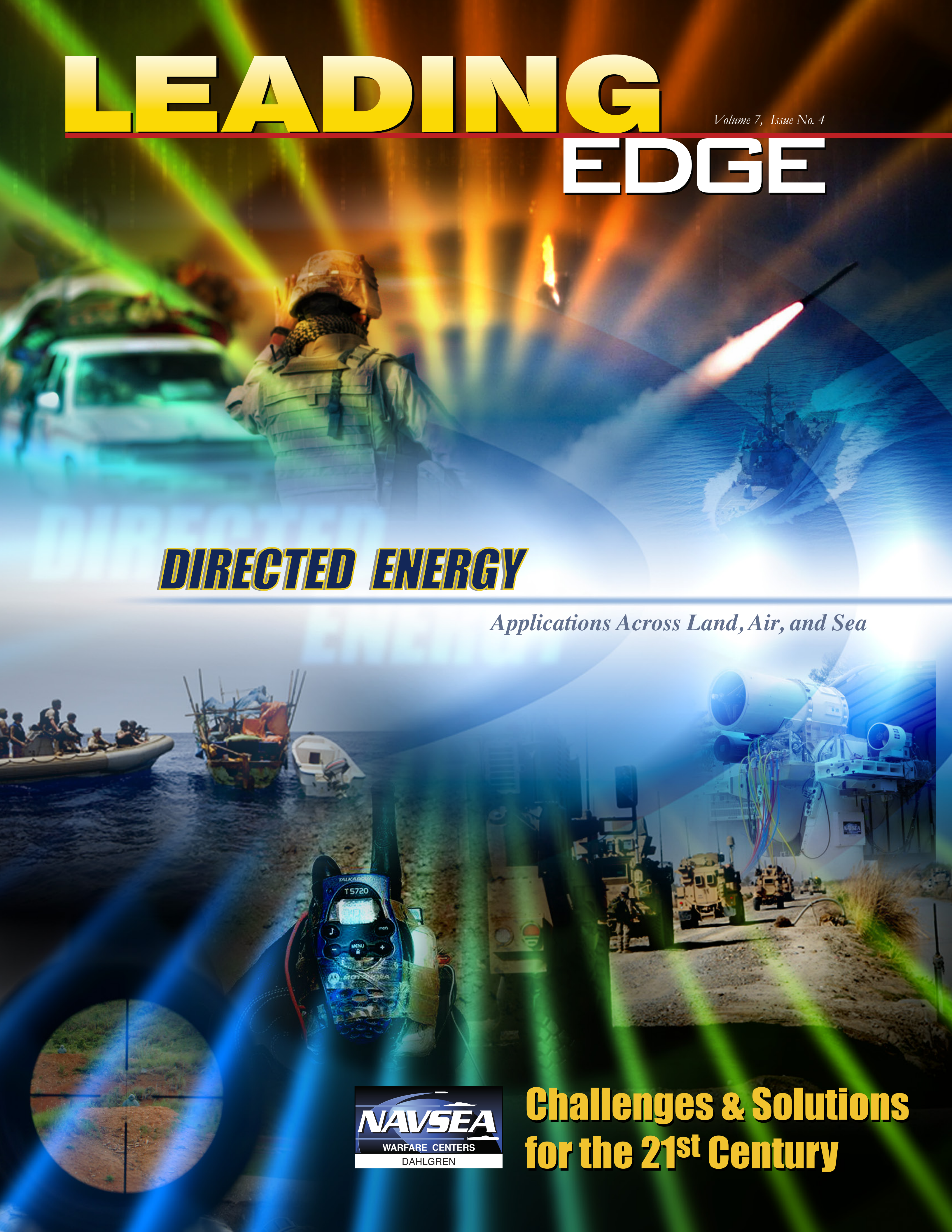

DIRECTED ENERGY

Applications Across Land, Air, and Sea

NAVSEA
WARFARE CENTERS
DAHLGREN

Challenges & Solutions for the 21st Century

LEADING EDGE
Volume 7, Issue No. 4

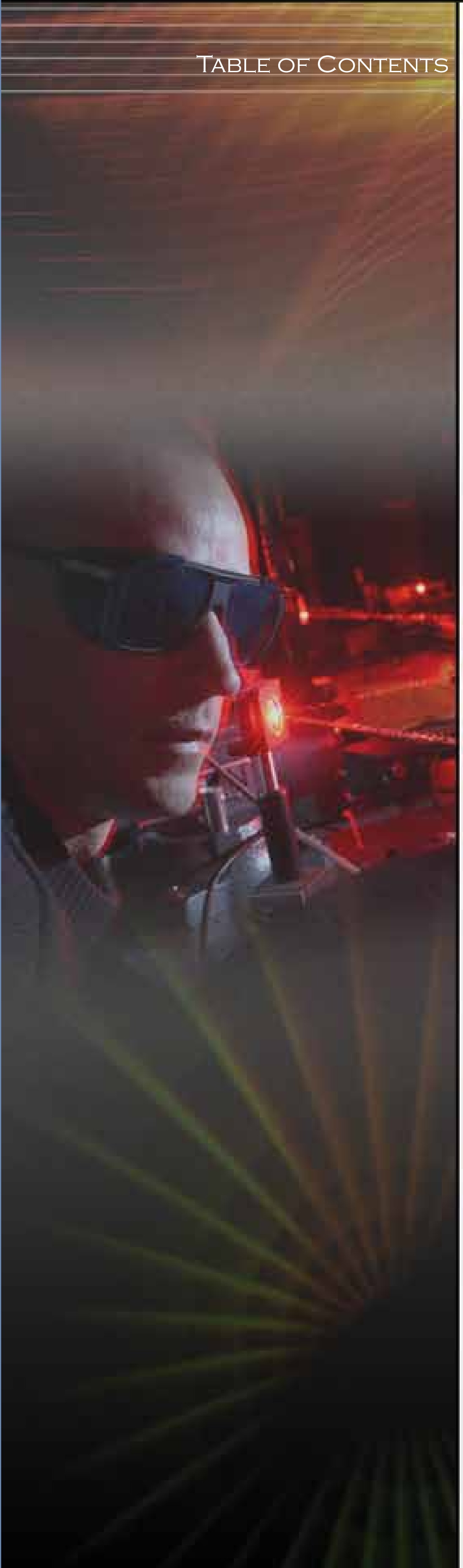

TABLE OF CONTENTS

Directed Energy

3 LEADING EDGE DIRECTED-ENERGY RDT&E
Captain Michael H. Smith

4 DIRECTED-ENERGY TOPICS IN THIS ISSUE
Mr. Dale Sisson

PAST, PRESENT, AND FUTURE

6 NAVAL DIRECTED-ENERGY WEAPONS – NO LONGER A FUTURE WEAPON CONCEPT
David C. Stoudt

12 HISTORICAL OVERVIEW OF DIRECTED-ENERGY WORK AT DAHLGREN
Stuart Moran

26 HISTORY OF LASER WEAPON RESEARCH
Melissa Olson

36 LASER WEAPON SYSTEM (LAWS) ADJUNCT TO THE CLOSE-IN WEAPON SYSTEM (CIWS)
Robin Staton and Robert Pawlak

RDT&E, ACQUISITION, AND WARFARE MANAGEMENT

44 THE ACQUISITION CHALLENGE ASSOCIATED WITH DIRECTED-ENERGY RDT&E
Mike Kotzian

TECHNOLOGY, MODELING, AND ASSESSMENT

50 THE BASICS OF ELECTRIC WEAPONS AND PULSED-POWER TECHNOLOGIES
Stuart Moran

58 SOLID MODELING OF DIRECTED-ENERGY SYSTEMS
Joseph F. Sharrow

64 A FUNDAMENTAL KEY TO NEXT-GENERATION DIRECTED-ENERGY SYSTEMS
Directed Energy Division, Electromagnetic and Sensor Systems Department

70 ACTIVE DENIAL ARRAY
Randy Woods and Matthew Ketner

74 DIRECTED ENERGY IN THE MILITARY ENVIRONMENT
LT Leedjia Svec, Jeremy Beer, and Dave Freeman

HIGH-POWER MICROWAVE

78 DIRECTED ENERGY USING HIGH-POWER MICROWAVE TECHNOLOGY
Jacob Walker and Matthew McQuage

HIGH-ENERGY LASER

82 LASER COUNTER ROCKET, ARTILLERY, AND MORTAR (C-RAM) EFFORTS
Michael Libeau

NONLETHAL CAPABILITIES

86 MULTIFREQUENCY RADIO-FREQUENCY (RF) VEHICLE STOPPER
Stephen A. Merryman

92 HIGH-POWER ELECTRICAL VEHICLE-STOPPING SYSTEMS
Jordan Chaparro and Melanie Everton

96 NONLETHAL SMALL-VESSEL STOPPING WITH HIGH-POWER MICROWAVE TECHNOLOGY
Jacob Walker

LEADING EDGE

NAVAL SURFACE WARFARE CENTER, DAHLGREN DIVISION (NSWCDD)
Captain Michael H. Smith, *Commander*
Carl R. Siel, Jr., *Technical Director*
David C. Stoudt, *Distinguished Engineer for Directed Energy (ST) and NAVSEA Technical Warrant for Directed Energy and Electric Weapon Systems*
Janice Miller, *Corporate Communications Director (Acting)*
Steve Zehring, *Managing Editor*
Margie Stevens, *Production Coordinator*
Patrice Waits, *Editor & Layout*
Clement Bryant, *Layout Design & Graphic Artist*
Kellie Yeatman, *Graphic Artist*
Trey Hamlet, *Graphic Artist/3-D Modeling*

ELECTROMAGNETIC & SENSOR SYSTEMS
Dale Sisson, *Head, Q Department*
Brandy Anderson, *Graphic Artist/Cover Design*

NSWCDD
Jordan Chaparro
Melanie Everton
Matthew Ketner
Michael Libeau
Matthew McQuage
Stephen A. Merryman
Stuart Moran
Melissa Olson
Robert Pawlak
Joseph F. Sharrow
Robin Staton
Jacob Walker
Randy Woods

DEFENSE ACQUISITION UNIVERSITY
Mike Kotzian

NAVAL MEDICAL RESEARCH UNIT – SAN ANTONIO
Jeremy Beer
Dave Freeman
LT Leedjia Svec

The Leading Edge *magazine is produced by the Naval Surface Warfare Center, Dahlgren, Virginia. The purpose of the publication is to showcase technical excellence across the Warfare Centers and promote a broader awareness of the breadth and depth of knowledge and support available to the Navy and DoD.*

Address all correspondence to Corporate Communications, C6
Email: dlgr_nswc_c6@navy.mil; or write to
Commander
Naval Surface Warfare Center, Dahlgren Division
Corporate Communications, C6
6149 Welsh Road, Suite 239
Dahlgren, VA 22448-5130

NSWCDD/MP-09/34
Approved for public release; distribution is unlimited.

NAVSEA
WARFARE CENTERS
DAHLGREN

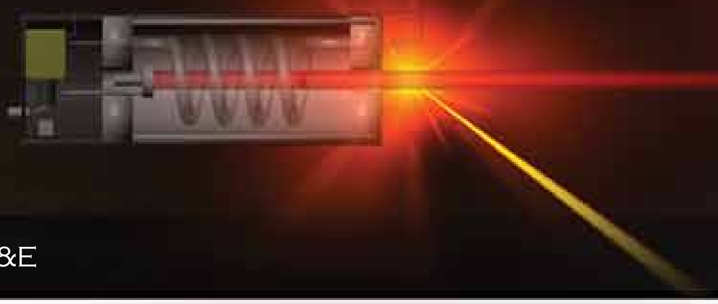

Introduction

Leading Edge Directed-Energy RDT&E

Captain Michael H. Smith, USN
Commander, NSWCDD

Dahlgren first launched research and development efforts devoted to harnessing the power of electromagnetic energy over 40 years ago. From early work with voltage multipliers and pulse-powered technology, to today's high-energy lasers and high-power microwave technologies, the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) has led, and continues to lead, cutting-edge directed-energy research, development, testing, and evaluation. Our commitment in this area only grows stronger—evidenced by our chartering of the Directed Energy Warfare Office (DEWO)—in order to provide increased focus on warfighting applications of these technologies.

Today's military forces face a wide array of challenges in diverse operating environments around the world. Directed energy offers unique and flexible options to address today's challenges, as traditional kinetic weapons are often of limited value in peacekeeping missions and in urban environments, where restricted rules of engagement typify the norm. Kinetic weapons can also be more costly or ineffective to employ against asymmetric threats.

The Chief of Naval Operations (CNO) recently placed added emphasis on directed energy and on expanding the range of directed-energy capabilities. In response, scientists and engineers at NSWCDD are actively developing prototype systems in a number of areas that you will read about in this issue—areas that have been successfully demonstrated and tested in our Navy laboratories and ranges.

In this issue of *The Leading Edge* magazine, you will trace the rich history of directed-energy work at Dahlgren, gain insight into directed-energy weapons already fielded or being readied for the field, and learn about prototypes that show real promise for providing incredibly effective offensive and defensive directed-energy solutions. For example, scientists and engineers at NSWCDD are leading the way toward realizing small, lightweight radio frequency (RF) transmitters using high-power, solid-state switching amplifiers for the development of counter-improvised explosive device detection and neutralization systems. You will also learn about diverse applications of directed-energy technology—such as research and testing of laser glare devices and laser eye protection—and have the opportunity to gain a better understanding of the Department of Defense (DoD) acquisition framework and the challenge of maintaining cost and schedule estimates while delivering weapons systems that are critical to the warfighter.

From lasers to high-power electrical vehicle-stopping systems, I am sure you will be fascinated and, along with me, be impressed with the advancements our scientists, engineers, and technical staff are achieving in the directed-energy arena to support of our men and women in uniform.

Directed-Energy Topics in This Issue

Dale Sisson
Head, Electromagnetic and
Sensor Systems Department
NSWCDD Dahlgren, Virginia

Welcome to our Directed Energy issue of the *Leading Edge* magazine. This issue represents the third in a trilogy of issues covering the truly fascinating and incredibly challenging area of naval warfare in the operational electromagnetic environment. In our first issue, we covered the full range of operational and readiness implications when operating in the electromagnetic environment. Then, in our second issue, we highlighted the complexities and dynamics of providing relevant and effective sensors and radars to our warfighters. Now, we focus on directed energy and relate how the Naval Sea Systems Command (NAVSEA) Warfare Centers, and the Naval Surface Warfare Center, Dahlgren Division's (NSWCDD's), in particular, are working on state-of-the-art directed-energy weapons capabilities for the warfighter.

In this issue, we first look back to the early years, decades ago, when directed-energy weapons research began. We examine the history of directed energy, and we cover significant discoveries and achievements made by NAVSEA Warfare Center scientists and engineers, and others in the scientific community. We then relate information about several of our current directed-energy initiatives, and about how we're working hard to solve some of the most complex technical challenges associated with directed-energy weapons. We highlight how others in the Navy, such as the Naval Medical Research Unit in San Antonio, Texas, are also conducting research into directed energy and how our forces can better protect themselves from the effects of directed energy. We show how directed energy can be employed in a variety of offensive and defensive, lethal and nonlethal situations. We explain how directed-energy weapons work and how they can be employed in various environments against a wide range of situations. Lastly, we look forward as we provide technical and strategic leadership for the efficient and effective development, acquisition, and fielding of directed-energy systems for the warfighter.

So, if you want to learn about what the NAVSEA Warfare Centers and others in the Navy are doing in the area of directed-energy weapons, look no further than this issue of the *Leading Edge* magazine. I'm confident that you will be impressed by the progress made in this most important technology field.

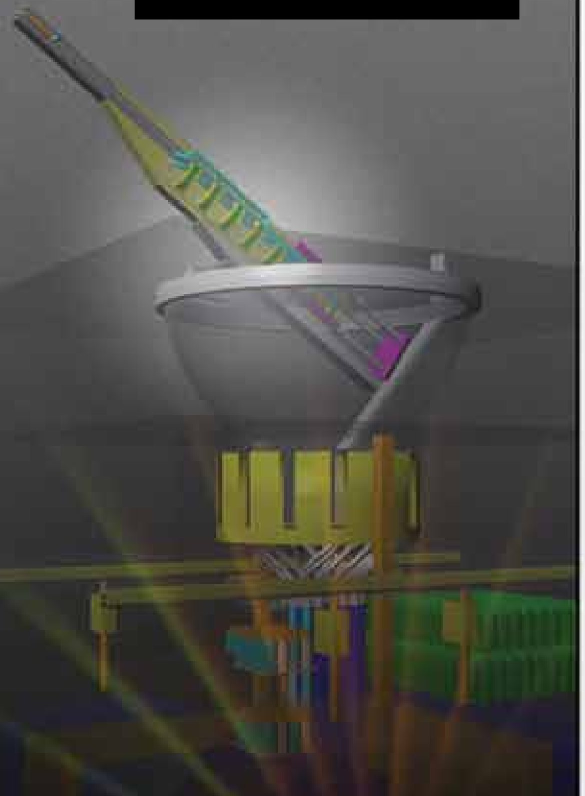

DIRECTED ENERGY

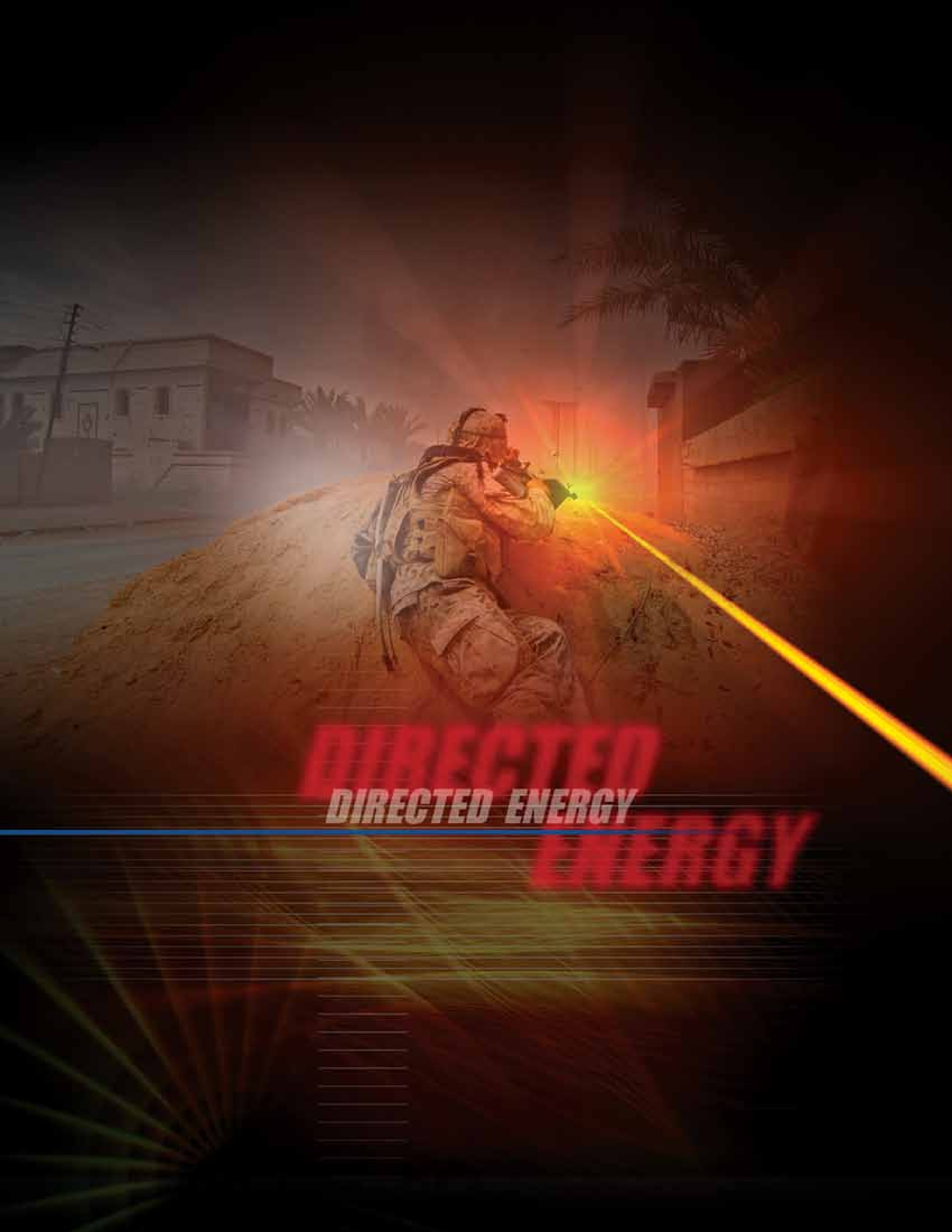
DIRECTED
ENERGY

NAVAL DIRECTED-ENERGY WEAPONS – NO LONGER A FUTURE WEAPON CONCEPT

By David C. Stoudt

Dr. Stoudt is the Distinguished Engineer for Directed Energy (ST) and the NAVSEA Technical Warrant for Directed Energy and Electric Weapon Systems.

Directed-energy weapon (DEW) technologies typically take the form of high-energy lasers (HELs), high-power microwaves (HPMs), and charged-particle beams. This article focuses on the first two technology areas, as they have reached the point of being ready for operational testing and evaluation, and in some cases, operational use on the battlefield. DEWs have been popularized in science-fiction writings for over a hundred years. The Department of Defense (DoD) has been investing in their development since the 1970s. This article will not go into technical depth regarding the various directed-energy (DE)-related efforts currently underway in the Navy, but rather, it will overview DE areas under development and relate recent Navy leadership activity. Other articles in this issue of *The Leading Edge* magazine will provide the reader with much greater technical and programmatic details on various DE efforts.

HIGH-ENERGY LASER WEAPONS

HEL weapon systems have been envisioned for a great many years, to include being referred to as Martian "Heat Ray" weapons in H.G. Wells' epic novel *The War of the Worlds*, originally published in 1898. In reality, a high-average-power laser weapon system is very similar to a "heat ray", or even a blow torch. During the early years of DoD investments in DE technology, the Navy led the development of HEL with the creation of the world's first megawatt-class, continuous-wave, Mid-Infrared Advanced Chemical Laser (MIRACL), located at White Sands Missile Range (WSMR). Roughly 80 years after the work of H.G. Wells, the U.S. Navy tested the MIRACL laser and ultimately used that laser system to engage static and aerial targets in the desert of WSMR in the following years. While that laser proved to be the wrong choice for the Surface Navy's self-defense mission, it did spawn work by the Air Force on the Airborne Laser (ABL), and the Army on the Tactical High-Energy Laser (THEL). In 2000 and 2001, the THEL successfully shot down 28 supersonic Katyusha artillery rockets and 5 artillery shells.

In 2010, the ABL successfully engaged and destroyed tactical ballistic missiles during the boost phase of their flight. All three of these laser systems—the MIRACL, the ABL, and the THEL—are chemical lasers that utilize toxic chemicals and operate in less than optimal wavelengths that make them a poor choice for most naval applications. The MIRACL is shown in Figure 1.

Recent advances in solid-state lasers, to include fiber lasers, have moved these electric lasers to the forefront of the Department's research and development (R&D) for near-term HEL applications in the services. The Navy has particular interest in electric lasers, to include the free-electron laser (FEL), for shipboard self-defense and force protection applications. The speed-of-light delivery of HEL energy can defeat the high-g maneuvers of newly developed foreign antiship cruise missiles (ASCMs). Thus, the Office of Naval Research (ONR) started an FEL Innovative Naval Prototype (INP) program in FY10, with a goal of reaching the output power of 100 kW. The eventual goal of the FEL program is to reach the multi-megawatt power level with wavelength selectivity. The Naval Sea Systems Command (NAVSEA) Directed Energy and Electric Weapons Program Office (PMS 405) has been actively developing a fiber laser-based Laser Weapon System (LaWS) that could be a retrofit to augment the current capabilities of the Close-In Weapon System (CIWS) currently deployed on many surface combatants. The Naval Surface Warfare Center, Dahlgren Division (NSWCDD), is the Technical Direction Agent and lead system integrator for PMS 405 on the LaWS program. The Naval Air Systems Command (NAVAIR) has interest in compact, solid-state HEL systems for aircraft self-protect and air-to-ground engagements, and will be starting a

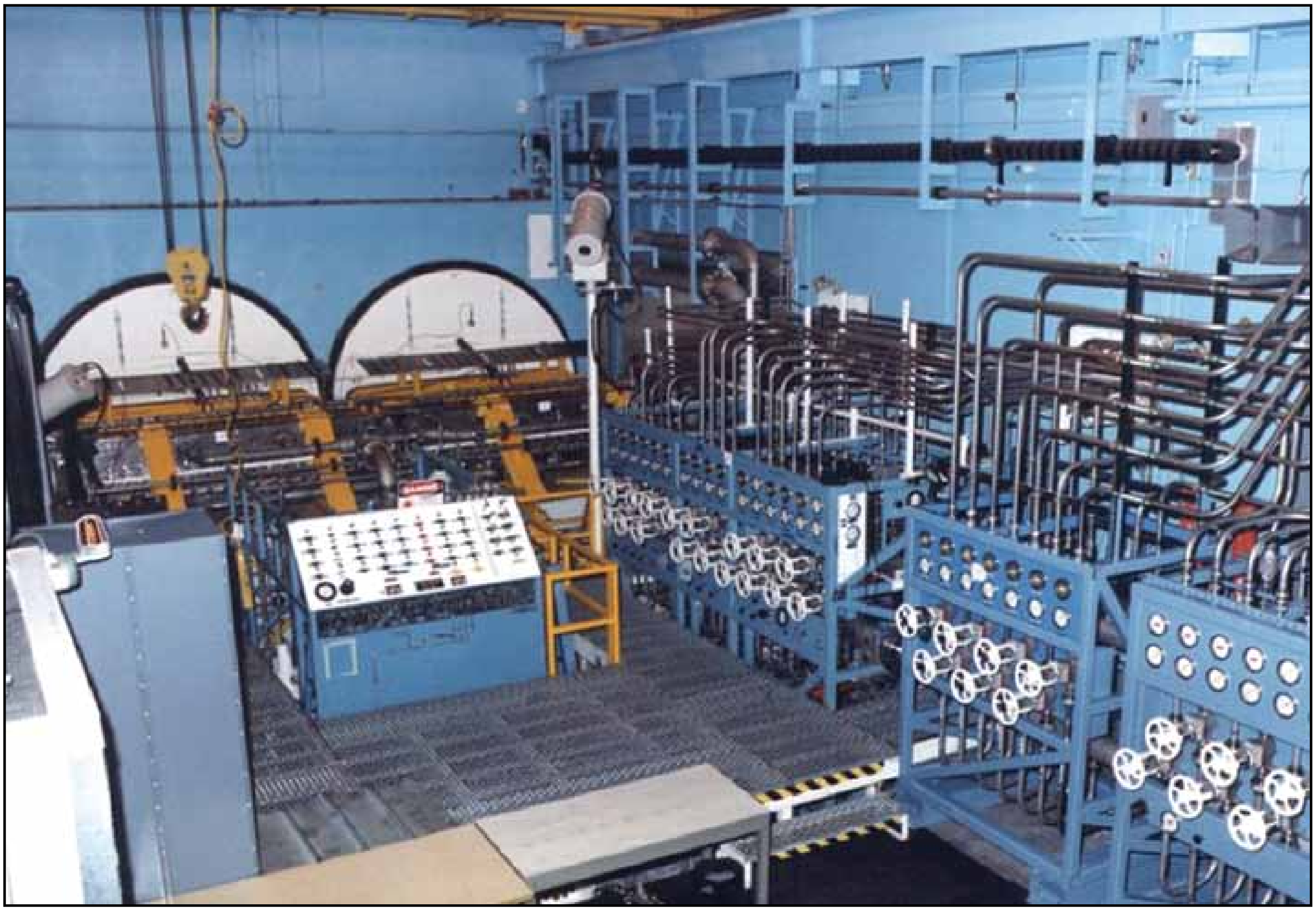

Figure 1. Mid-Infrared Advanced Chemical Laser (MIRACL)

fiber laser-based ONR Future Naval Capability effort in FY12. LaWS is shown in Figure 2.

High-Power Microwave Weapons

Like lasers, microwave weapons have been fantasized about ever since the invention of microwave power generators. In fact, in 1932 it was generally recognized by the British government that bombers, ostensibly German bombers, would be able to penetrate British air space and bomb its civilian population and infrastructures. In 1934, the Air Ministry initially asked Robert Watson-Watt, of the National Physical Laboratory, if he could build a "death ray" that could kill enemy pilots or detonate bombs while they are still on the planes of enemy aircraft. Such a "death ray" had been proposed to the Air Ministry by Harry Gindell-Mathews 10 years earlier in 1924. Watson-Watt, a former meteorologist who had become an expert on radio signals, suggested that energy reflected from an aircraft could be used to locate it. His experiments were successful and RADAR (radio detection and ranging), a name coined by the U.S. Navy in 1940, was born. While RADAR is not a DEW in the way they are thought of today, its roots can clearly be traced to the military's desire for such capabilities.

The Navy's HPM, or high-power radio-frequency (RF) systems, have been progressively increasing in power density to the point where it is now feasible to integrate the technology into weapon systems for deployment. While initial HPM applications suffered from their inability to obtain militarily useful outcomes, either due to technology limitations, difficult concept of operations (CONOPS), or inherent robustness of potential target systems, many feasible military applications for using HPM

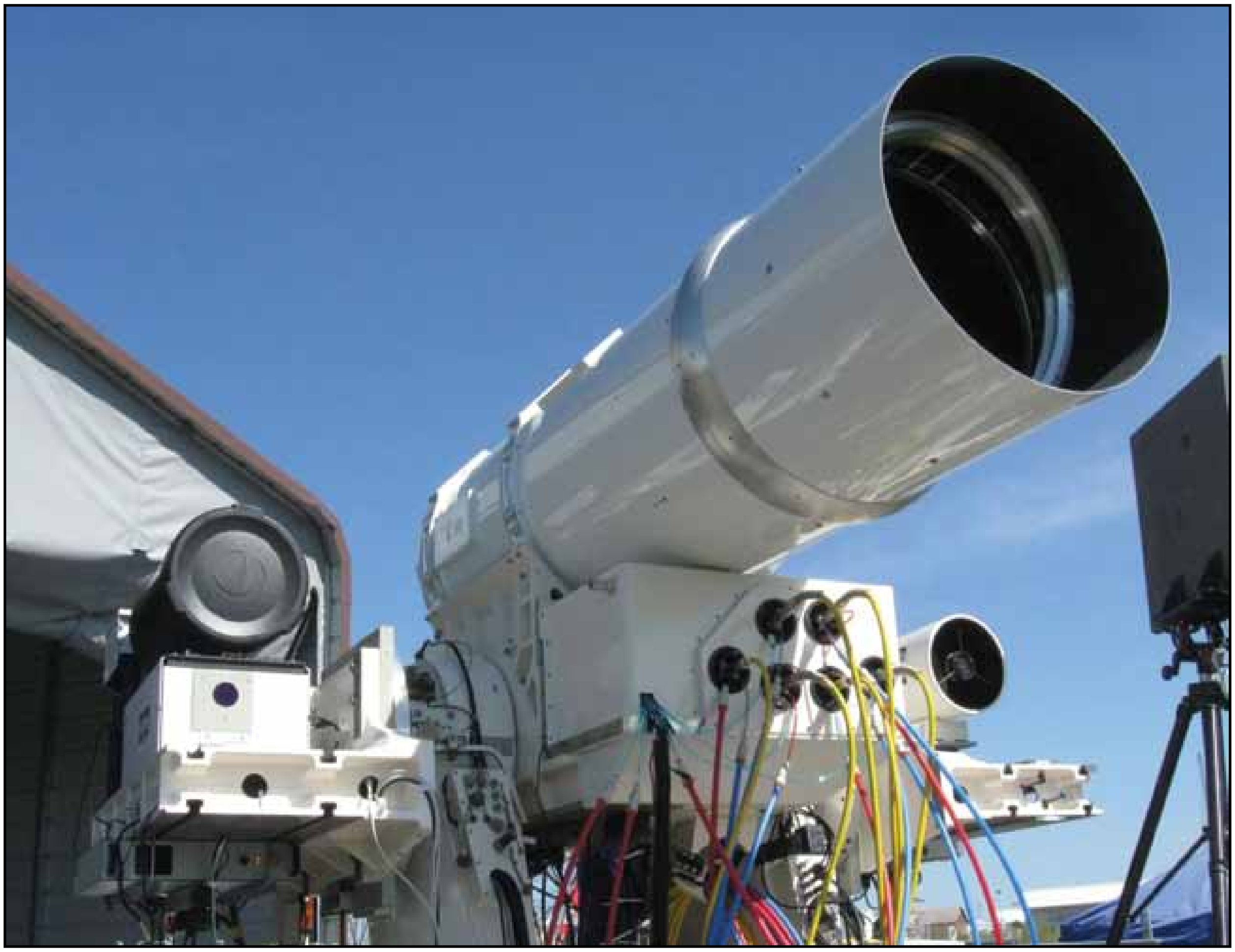

Figure 2. Laser Weapon System (LaWS)

devices have surfaced over recent years to include nonlethal, antipersonnel weapons and nonkinetic, antimateriel weapons. While these concepts offer unique capabilities to the warfighter due to the nonkinetic effects they generate, other warfighting concepts—such as stopping vehicles, or countering hidden roadside bombs or improvised explosive devices (IEDs)—are difficult to achieve by any other means. The multifrequency Radio-Frequency Vehicle Stopper (RFVS) system is shown in Figure 3.

In addition, the difficulty in overcoming the propagation losses associated with HPM has driven some concepts into platforms such as unmanned aerial vehicles (UAVs) or cruise missiles that deliver the HPM device to the target for a close-in engagement. Over the past 10 years, field-testable prototypes have been developed to demonstrate the operational utility of these concepts, and in some cases, those prototypes have or will be deployed operationally to support our troops in theater. It is only through the hard work and perseverance of the Naval Research Enterprise (NRE), as well as other DoD laboratories, that concepts that were once only laboratory curiosities are now making their way onto the battlefield and contributing to the fight.

Foreign Directed-Energy Weapon (DEW) Development

While the United States has been very active in this warfighting area, significant foreign DEW development also has elevated the need for the Navy to afford these threats a higher priority. This can be done either by incorporating the necessary DEW countermeasures into weapon systems, platforms, and critical infrastructures, or by adapting the CONOPS and tactics, techniques, and procedures (TTPs) employed by our armed forces to properly account for those foreign DEW systems. Materiel developers need to understand how this threat

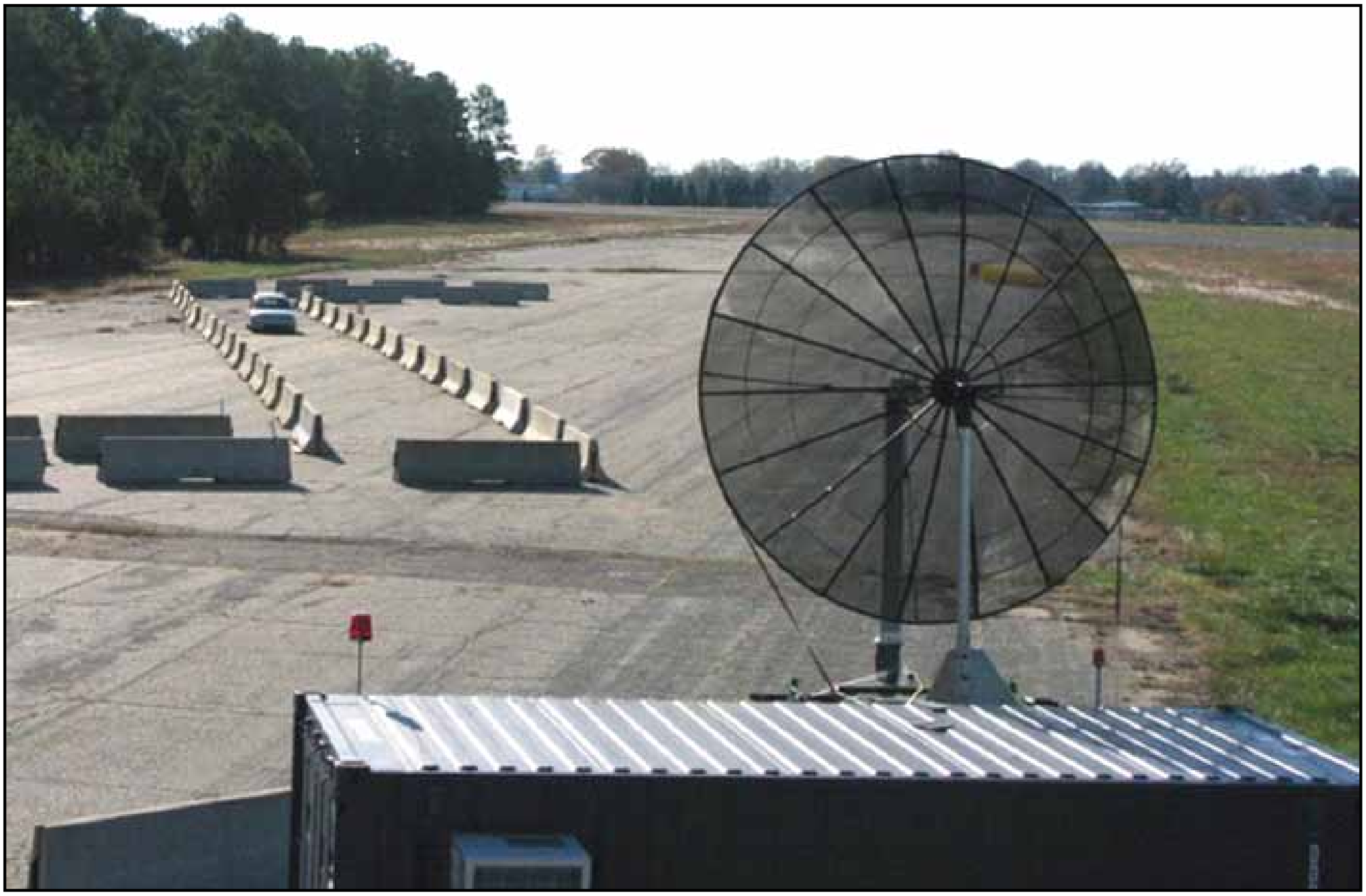

Figure 3. Multifrequency RF Vehicle-Stopper (RFVS) System

is evolving and properly address it during the design of their systems. They also need to address DE in the development of their system threat assessments. There has been movement on the HPM side to modify existing military standards, such as MIL STD 464[1] and others, to now include information on potential HPM threats. For example, in the HEL arena, work has been accomplished in the development of protective measures for eyes; however, this threat needs to be considered during the system development process. It is well known that building in countermeasures is much cheaper during the initial development of a system, vice trying to retrofit systems with countermeasures once a new threat is on the battlefield. As analysts evaluate the foreign development of DE technologies, and the trends become clearer, it is the responsibility of the acquisition community to take this threat into consideration and ensure that weapon systems, platforms, and infrastructures will be available and at full capability when needed. By accounting for foreign threat developments, assessing blue force susceptibilities and vulnerabilities, and adopting appropriate measures to negate or counter these threats, naval forces will avoid technological surprise on the battlefield in the future.

REQUIREMENTS

The DE programs briefly mentioned in this article, and covered more deeply in this and other publications, offer warfighters unique capabilities not currently found in their arsenal. The continuing problem, however, is matching those unique capabilities to vetted operational requirements. The DE technical community has made great strides in helping the operational community understand the capabilities of DE weapons and their potential military effects on targets. The lack of formal requirements, however, has yielded more of a technology push—rather than an operational pull—of various DE capabilities. Progress has been made, but more effort is required if DE capabilities are to be developed and transitioned between science and technology (S&T), and formal programs of record. Notwithstanding, the current outlook and trends are positive.

A RESURGENCE OF NAVY INTEREST IN DIRECTED ENERGY

The Navy's interest in DEWs for future maritime operations has increased in recent years due to a number of weapons development successes. Recognizing the importance and value of DEWs, NAVSEA reestablished the Navy Directed Energy Weapons Program Office (PMS 405) in 2004. Accordingly, PMS 405 was designated as the point of contact for matters related to DE and electric weapon systems (EWS) development and acquisition initiation for NAVSEA, and for matters being coordinated with other federal agencies and military services. PMS 405's mission is to transition technology from the laboratory to prototype/advanced development/testing for operational development and use.[2,3]

The Navy also established its first formal executive position for DE (ST-level), the Navy's Distinguished Engineer/Scientist for Directed Energy, at NSWCDD in August 2004. Following the establishment of this position, NAVSEA then formally established a Technical Authority Warrant for Directed Energy and Electric Weapon Systems (DE&EWS)—Surface Ships in July 2008. The scope of the warrant includes the transition of S&T development to weapon system development of lethal and nonlethal capabilities associated with the DE&EWS for Surface Ships.[4] This included, but was not limited to, the following:

- Laser Weapon Systems
  - High-Energy Lasers
  - Solid-State Lasers
  - Free-Electron Lasers
  - Femtosecond Ultrashort Pulse Lasers
  - Laser-Induced Plasma Channel
  - Lethality/Vulnerability
- Electromagnetic Rail Gun Weapon System
- High-Power Microwave
  - Active Denial System
  - Laser-Guided Energy
- Maritime Directed Energy Test Center
- Electromagnetic Launch of Weapons (excluding the Electromagnetic Aircraft Launch System (EMALS))

Then, within the NAVSEA Warfare Center Enterprise, Warfare Center leadership established two technical capabilities (TCs): an NSWCDD TC for DE systems research, development, test, and evaluation (RDT&E); and a Naval Surface Warfare Center, Port Hueneme Division (NSWCPHD) TC for in-service engineering, test and evaluation (T&E), and integrated logistics support to DE systems. NSWCDD leads all S&T and RDT&E for the development and weaponization of DE systems for surface, air, and ground environments. It also leads the development of offensive and defensive DE technologies needed to characterize and exploit vulnerabilities, provide weapons, and protect against attack. NSWCDD provides the technologies, devices, and systems designed to create or control electromagnetic energy used to cause persistent disruption or permanent damage by attacking target materials, electronics, optics, antennas, sensors, arrays, and personnel, including nonlethal applications. NSWCPHD provides in-service engineering, T&E, and integrated logistics support to DE systems throughout the system life cycle.

The Navy further demonstrated increased interest in DE when Assistant Secretary of the Navy (Research, Development & Acquisition (ASN(RDA)) designated NAVAIR offensive and defensive leads for naval aviation DE activities:

- Program Executive Officer for Unmanned Aviation and Strike Weapons (PEO(U&W)), assigned as the offensive DE lead for naval aviation
- PEO for Tactical Aircraft Programs (T), assigned as the defensive lead for naval aviation

Concerning future initiatives, the Chief of Naval Operations (CNO) tasked the Strategic Studies Group (SSG) to examine a topic entitled "Maritime Operations in the Age of Hypersonic and Directed-Energy Weapons."[5] The intent of the study was to provide Navy leadership with an understanding of where DE technologies and weapons are today and how they might influence future maritime operations. The theme of the study was completed during FY10, the results of which discuss many DE concepts, as well as tactics for the employment of DE capabilities. The study's findings are currently under review and consideration by senior Navy leadership.

CONCLUSION

While H.G. Wells' *The War of the Worlds* novel and television programs like *Star Trek* popularized the notion of using DE for weapons in years past, today— through persistent DEW RDT&E—Navy leadership is realizing the great potential that DEWs offer naval warfighters and homeland defenders. The scientific and technical advances the Navy has made in HEL and HPM in recent years have been nothing short of extraordinary. Moreover, future technological and engineering advances undoubtedly will result in profound differences in our nation's future warfighting capabilities. Naval DEWs, therefore, are no longer just a future weapon concept…they are here today.

REFERENCES

1. *Department of Defense Interface Standard, Electromagnetic Environmental Effects—Requirements for Systems*, MIL-STD-464, 18 March 1997.
2. NAVSEA Notice 5400, Ser 09B/240, Subject: Establishment of the Navy Directed Energy Weapons Program Office (PMS 405), dated 4 January 2002.
3. NAVSEA Instruction 5400.101, Ser SEA 06/058, Subject: Directed Energy and Electric Weapons Program Office (PMS 405) Charter, dated 21 July 2004.
4. NAVSEANOTE 5400 Ser 09B/240 of 4 January 2002 and NAVSEAINST 5400.101 Ser SEA 06/058 of 21 July 2004.
5. Letter from the CNO tasking the SSG XXIX Theme—"Maritime Operations in the Age of Hypersonic and Directed-Energy Weapons," dated 23 September 2009.

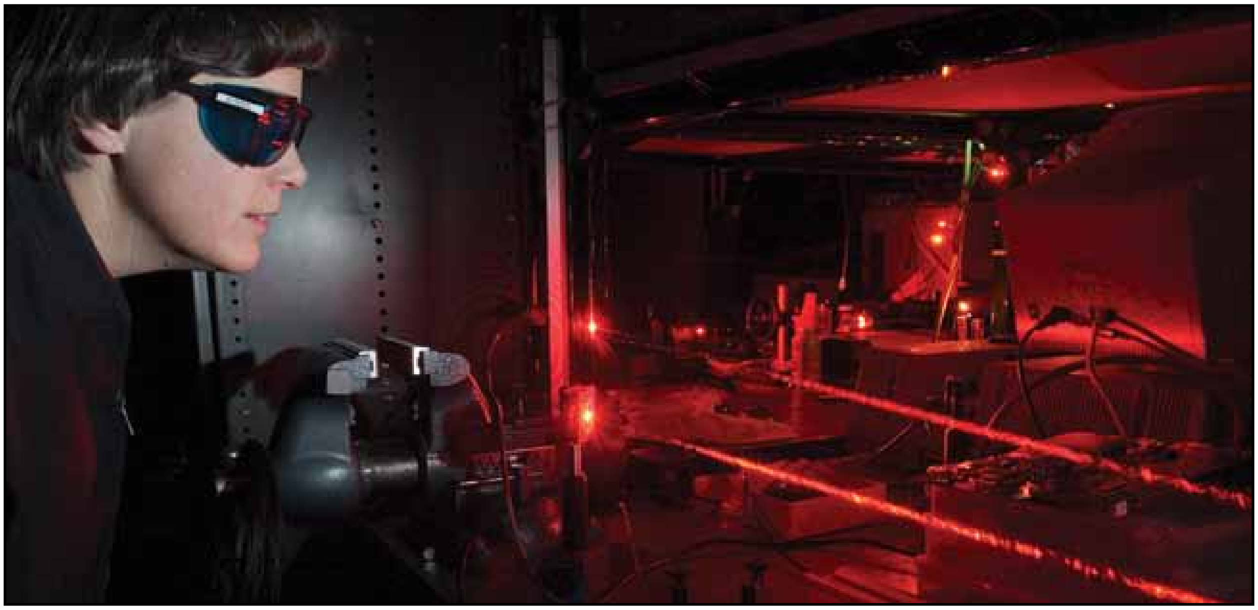

HISTORICAL OVERVIEW OF DIRECTED-ENERGY WORK AT DAHLGREN

By Stuart Moran

In 1962, the United States set off a megaton nuclear weapon 250 miles above the Pacific. The blast caused a large imbalance of electrons in the upper atmosphere that interacted with the Earth's magnetic field to create oscillating electric fields over a large area of the Pacific. These fields were strong enough to damage electronics in Hawaii, a thousand miles away, and clearly demonstrated the effects of an electromagnetic pulse (EMP). It didn't take long for the military to begin considering ways to create such pulses without using nuclear weapons.

In the late 1960s, the Special Applications Branch at the Naval Weapons Laboratory at Dahlgren began studying ways to generate high-power oscillating electric fields that could be used as a weapon to damage enemy electronics. These devices were basically high-power versions of the old spark-gap transmitters used in the early days of radio. To construct a device that could produce nuclear EMP-like fields, stored electrical energy was converted to radio-frequency (RF) energy that could be radiated from an antenna through the atmosphere to a target. These devices typically would store energy in a high-voltage capacitor and release the energy quickly using a spark-gap switch. This would then drive oscillating currents on an antenna, causing it to radiate. To achieve field strengths of thousands of volts per meter, typical of a nuclear EMP, devices operating at hundreds of thousands of volts or more were needed.

A number of radiating devices were studied in the early 1970s. Most belonged to a class of devices called Hertzian oscillators. A capacitor is charged to high voltage, the switch is closed, and current flows in the circuit, causing the stored energy to oscillate between the electric field of the capacitor and the magnetic field of the inductor. To charge the capacitor to extremely high voltages, a step-up transformer of some type must be used. One of the fastest voltage multipliers, the Marx generator, was frequently used. The losses from internal resistance and external radiation damp the oscillating waveform, typically after a few cycles. The radiated pulses are, therefore, short in time and broad in frequency content.[1] A simple diagram of the inductance-capacitance oscillator (L-C oscillator) is shown in Figure 1.

SINGLE-PULSE BURNOUT DEVICES

Many types of Hertzian devices were designed, constructed, and tested at Dahlgren during the 1970s. The transmission-line oscillator, or cavity oscillator, used a quarter-wavelength

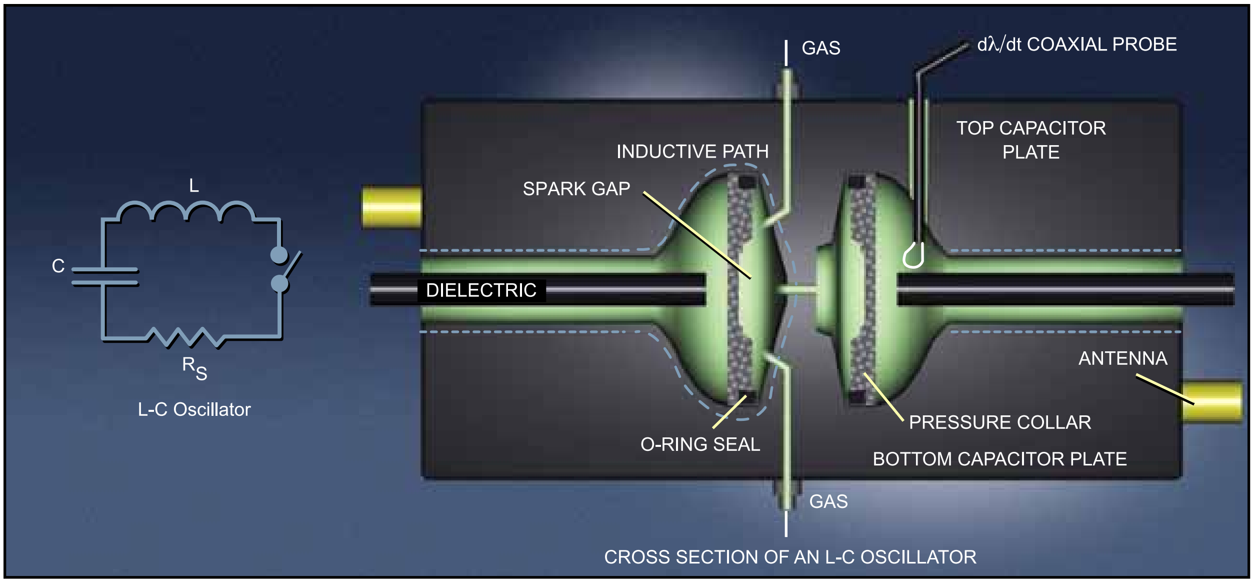

Figure 1. Inductance-Capacitance Oscillator (L-C Oscillator) Diagram

coaxial pipe, which was switched at one end, to create the oscillating waveform. A frozen wave generator, a different type, had quarter-wave sections of cable that were charged plus and minus to create a two-cycle waveform "frozen" in the cable. All sections were simultaneously switched, causing the wave to travel to an antenna. A special folded design was developed so one switch could be used, eliminating the multiswitch synchronization problem. A Ross circuit used a square wave pulse, which traveled down cable "tees," creating reflections, which were timed to create several RF cycles. In the Travetron, the turn-on time of a series of spark-gap switches was incorporated as a designed delay, creating reflections through a series of gaps to produce the waveform. This design allowed higher frequencies. All of these devices were designed, built, and tested to determine power and frequency capabilities, as well as efficiency.

Scientists and engineers at Dahlgren built and tested versions of Hertzian oscillators operating up to half a million volts. These devices powered relatively simple monopole or dipole antennas that could produce very high electric fields at hundreds of meters. In the early 1970s, a special outdoor field-measurement range was constructed. It housed high-voltage systems in underground trailers that fed antennas above ground on a specially-built, 100-m-long ground plane that was constructed for testing and field measurements. A picture of the ground place in a fielded measurement range is shown in Figure 2. Field probes were even carried aboard helicopters to make measurements above ground effects, as shown in Figure 3.

Other types of devices to produce pulses were constructed, too. Vector inversion generators used spiral-wound capacitive plates to generate high voltages without transformers.[2,3] The Landecker ring used a paddle-wheel arrangement of capacitors and inductors charged in parallel and discharged in series. The circular arrangement was designed so the entire system would radiate as a magnetic dipole, thus forming its own antenna.[4] Switch timing was critical, and Dahlgren engineers attempted to verify reports that Landecker developed a specific type that brought all capacitor leads into a single-center spark gap.

Scientists and engineers also looked at devices that used explosives to generate the electrical energy needed. These included explosive flux compressors of several types, which generated fields and then explosively squeezed the fields between conductors to amplify the peak power. In the early 1970s, a large (70-ft clear zone) anechoic chamber was constructed at Dahlgren with an explosive chamber in one end. Explosives would be set off in the chamber to drive various types of flux compressor schemes that would generate electrical pulses fed into an oscillator and antenna in the anechoic chamber. Pulse parameters and field strengths could be measured. Impedance-matching networks, matching transformers, and methods of improving efficiency were studied. Tests were performed at Dahlgren and at Los Alamos using large antennas suspended from balloons.[5] In other schemes, piezoelectric devices were developed, which could be compressed hydraulically and then quickly released to produce high voltages. The concept was to use explosives to generate the high

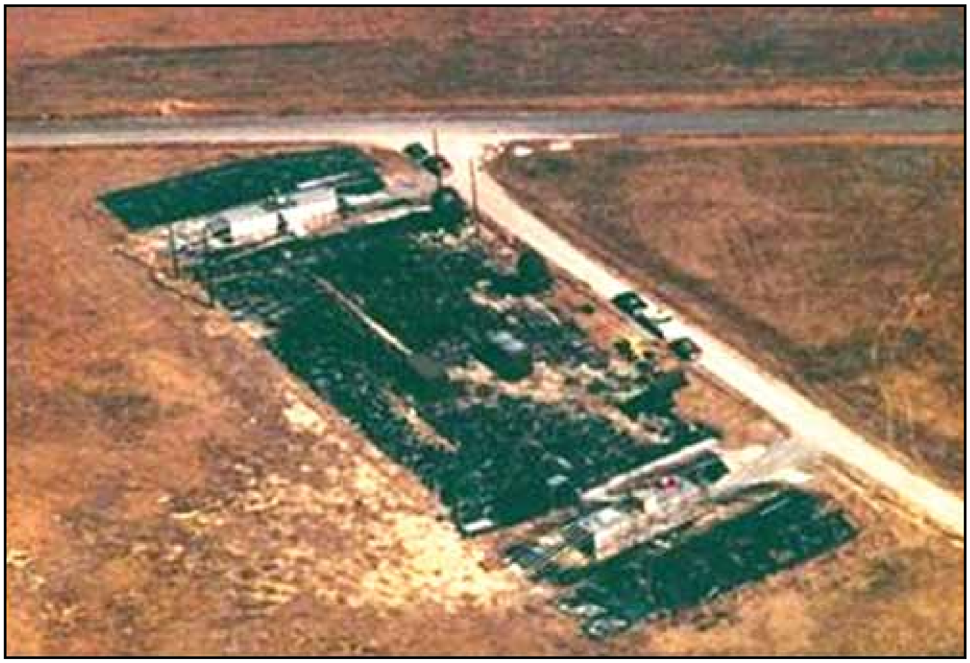

Figure 2. Field Measurement Range

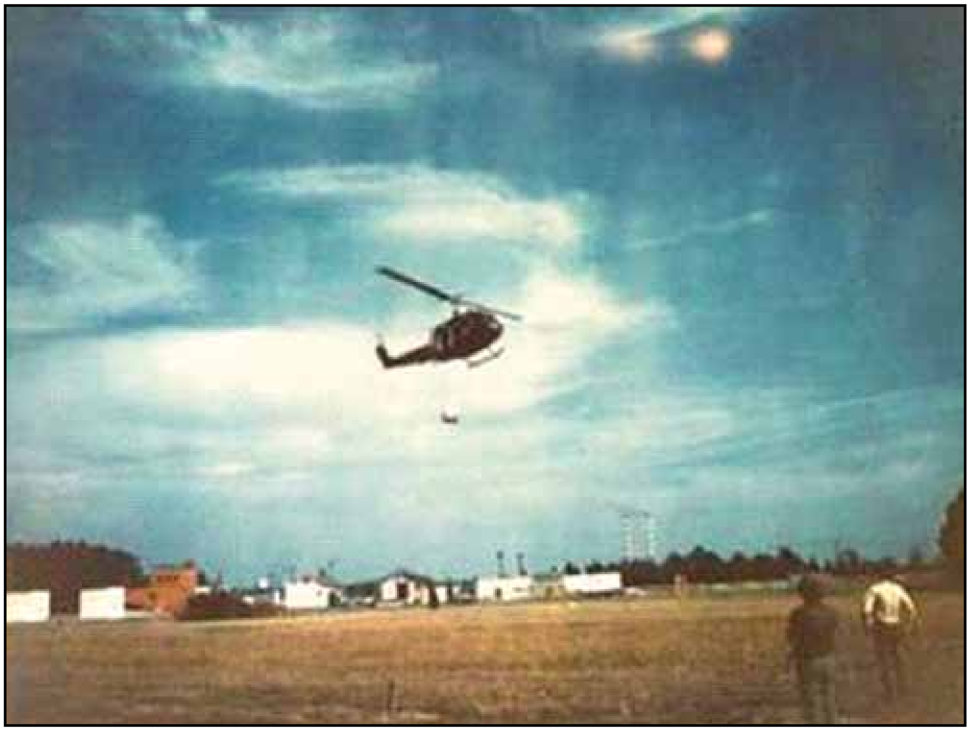

Figure 3. Airborne Electric Field Measurements

pressures. Ferroelectric and ferromagnetic transducers driven by explosives were also tested.[6]

Special Effects Warhead (SEW) Program

In 1973, Dahlgren began the SEW Program to look at the feasibility of "burning out" enemy radar and missile systems using single-shot, very high-peak-power EMPs. The program looked at the feasibility of constructing an electromagnetic warhead that could disable electronics beyond a normal hard-kill explosive range as far as a mile away. The program was funded at several million dollars a year through most of the 1970s.

A major thrust of the SEW Program was to better understand the effects of high fields on military electronics. Little information was available on the vulnerability of foreign or U.S. electronics, particularly entire systems. A trailer-based RF impulse system, employing a Marx-driven L-C oscillator charged at two million volts, was constructed at Dahlgren. This Transportable Oscillating Pulser System (TOPS) was connected to a large bounded-wave structure that produced uniform fields over a region large enough to place an entire radar or missile system. The electric field emitted from the throat of this system was so high that a special bag of high-voltage gas was needed until the radiating structure became large enough to transition to the normal atmosphere. A picture of TOPS is shown in Figure 4.

Since many important target systems were not available for testing, much of the vulnerability information was obtained from U.S. electronics, and estimates were then made for foreign systems. In addition to the tests done at Dahlgren, pulsers were also constructed in mobile trailers that could be transported to other sites for testing against simulated or actual targets. The Mobile Oscillating Pulser System (MOPS) was an example that was carried to test sites, such as China Lake, to perform tests against radars and simulated foreign systems.

A key requirement for the SEW Program was to demonstrate enforceable target vulnerability, which means that a high percentage of the time a large percentage of the targets are affected. One important finding was the broad difference between an electromagnetic safety concern—where a 1 percent vulnerability was far too great—and a weapon concern—where a 10 percent vulnerability was not good enough. The field strengths between the safety requirements and weapon requirements often were many orders of magnitude apart.

The SEW Program looked at many types of electronic component vulnerability, subsystem vulnerability, and complete system vulnerability. As a result, energy tables for burnout effects were developed. Subsequently, Dahlgren performed numerous field tests against radar and communications systems between 1973 and 1978, and funded component and subsystem testing on missiles.

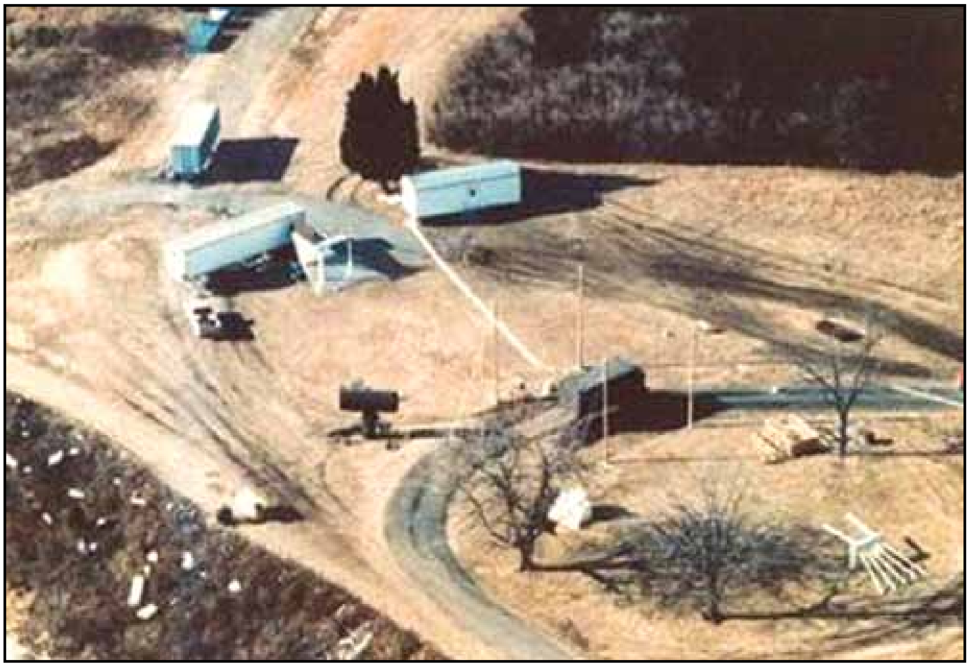

Figure 4. Transportable Oscillating Pulser System (TOPS)

REPETITIVE SYSTEMS FOR ELECTRONIC WARFARE

The electric fields required to damage military electronics in the 1970s often were very high, and ranges typically were limited. As a spinoff of programs trying to damage targets with a single pulse, some of these devices were reduced in size and power, and operated in a repetitive mode to generate noise pulses for the purpose of electronically jamming target systems. In 1976, the Naval Air Systems Command (NAVAIR) began the Electromagnetic Countermeasures Program to study the application of high-repetition-rate Hertzian devices for use as noise jammers. The initial targets were low-frequency radars.

In late 1976, Dahlgren performed effectiveness tests against various radars using helicopter-mounted Hertzian jammers. These devices were able to screen incoming target aircraft at useful ranges. The concept of a forward-launched rocket to deliver a parachute-suspended Hertzian jammer also was investigated. Dahlgren teamed with engineers at China Lake to study packaging concepts of utilizing an extended 5-inch Zuni rocket as a forward-fired delivery vehicle. A prototype is shown in Figure 5.

Similar Hertzian devices were considered for use as communications and data-link jammers. Several antenna deployment schemes were developed, and by fall 1978, successful ground launches had been performed in which the deployment sequence and jammer operation were demonstrated. The name Zuni Expendable Pulsed-Power Oscillator (ZEPPO) was given to the project. Dahlgren teamed with the Naval Avionics Center (NAC) to build the systems. By 1980, China Lake fired the first air-launched prototypes at both low and high altitudes. Devices, batteries, spark gaps, and antennas continued to be developed, and new targets—such as spread-spectrum systems—were tested. Other delivery systems besides rockets were also considered.

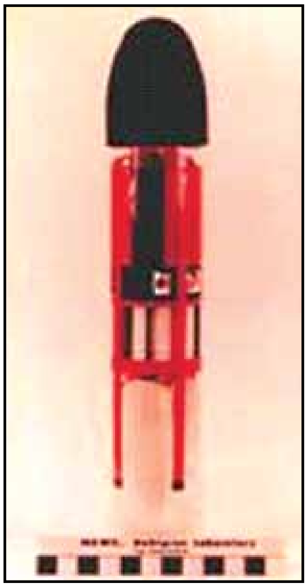

Figure 5. ZEPPO Payload

THE PULSED POWER TECHNOLOGY PROGRAM

Large directed-energy weapons (DEWs) often required megawatts or gigawatts of peak power, so methods of supplying and modifying this power were needed. As Dahlgren became involved in a broad range of DEW systems, one attribute became more and more obvious: the size, weight, and cost of a directed-energy (DE) system were dominated by the pulsed-power technologies needed to drive the system, not by the source device itself. Consequently, more effort began to be devoted to the power-delivery technologies needed for many of the weapon concepts. Pulsed-power components enabled energy to be stored over long periods of time (seconds) and released very quickly (nanoseconds) to obtain a billion times increase in peak power.

Dahlgren hosted a pulsed-power systems symposium and workshop in 1976 and helped initiate the International Pulsed Power Conferences, which began in 1977 and continues today under the Institute of Electrical and Electronics Engineers (IEEE). As Dahlgren's involvement with systems design increased, it became apparent that new technologies were needed in the prime-power and pulsed-power area to support a variety of new concepts. Dahlgren urged the Navy to initiate a Pulsed Power Technology Program to develop power sources, energy storage systems, high-power switches, and power conditioning systems needed for a variety of future weapons. This program was initiated in 1978 and was originally funded by NAVAIR and then by the Directed Energy Program Office (PMS 405) in the early 1980s. In addition to the Pulsed Power Technology Program, PMS 405 also began funding free-electron lasers (FELs), chemical lasers, high-power microwaves (HPMs), and charged-particle beams (CPBs). The Pulsed Power Technology Program at Dahlgren, in turn, funded many areas of research, both internal and external, over the next 10 years. Dahlgren served as the focal point for the Navy's science and technology (S&T) in pulsed power and funded many universities, government laboratories, and commercial companies under the Pulsed Power Technology Program.

To provide large amounts of electrical prime power, new types of rotating machines were studied, including flywheels, conventional alternators, homopolar generators, rotary flux compressors, and compensated pulsed alternators. These machines attempted to produce fast, high-power pulses using special materials to reduce losses, eddy currents, and mechanical stresses. MHD generators were developed using rocket-motor propellant that could be started and stopped. In the mid-1980s, a full-scale hybrid (solid fuel/liquid oxidizer) combustor was fabricated and tested at 10 MW, achieving world records for power-to-weight ratio and conductivity. By 1980, new types of energy storage systems were studied, including inductive storage and advanced capacitors using new types of insulating materials and geometries. During the late 1980s, programs such as the Mile-Run Capacitor Program reduced the capacitor size by a factor of 10 through better synthesis of polymer films.

Beginning with internal independent research funds, Dahlgren developed liquid dielectric materials based on water/glycol mixtures at low temperatures. These water-capacitor devices could hold energy for orders-of-magnitude longer time periods than ever before, allowing pulseforming lines to be constructed that could be charged directly from rotating machines. Dahlgren scientists developed a world-record high-voltage water capacitor that could hold pulses for milliseconds and became internationally recognized experts in water breakdown.[7,8]

High-power fast switching was another important area of research. Dahlgren funded companies to develop new types of multistage thyratrons that could operate at very high voltages. By the early 1980s, multistage thyratrons capable of operating at over 200 kV, 40 kA with 20 nsec risetimes were demonstrated. Vacuum switches, ignitrons, plasma pinch switches, pseudospark switches, back-lighted thyratrons, and e-beam switches all were studied, as well as a variety of spark-gap switches. Higher power solid-state switches were developed, too, using new geometries and substrate material. Superconducting coils were considered, both for energy storage and as opening switches. Dahlgren engineers developed exploding-wire opening switches, and several types of plasma pinch switches were funded. They also worked on stacked cable pulsers. Additionally, concepts for electromagnetic armor were developed. These systems used high-density capacitors to blunt penetrators. Inductive energy storage—which could be far denser than capacitors—was studied, including methods of generating the seed current and the problematic high-voltage opening switch. Opening switches—which were needed for inductive energy store systems—were studied, as well as magnetic switches, which used saturating magnetic material to sharpen pulses. Magnetic switches operating at 10 kHz were demonstrated by 1983.[9]

In 1985, Dahlgren used internal funds to upgrade a facility to provide controls, diagnostics, and 200 kW of average power at 50 kV to accommodate testing of new switches and water-based capacitors. This facility could control the power with a vacuum-tube pulser and could generate over a million volts with a rep-rated Marx generator. The facility was used to:

- Develop water-dielectric energy storage, rep-rated spark gaps, and pseudospark switches.
- Test a variety of switches developed by contractors, such as back-lighted thyratrons.[10, 11]

A picture of one system being tested—a water pulse-forming line and spark-gap switch—is shown in Figure 6.

Dahlgren concentrated in-house switching efforts in spark gaps. New types of gases were studied, as well as electrode materials, gas-flows, switch geometries, and triggering techniques to produce high-repetition-rate switches for electronic warfare, as well as particle-beam weapons.[12] Dahlgren scientists and engineers demonstrated 100-µs recovery of spark-gap switches after handling kilojoules of energy at hundreds of kilovolts, a world record.[13] The High Energy 2-Pulse System for fast recovery experiment is shown in Figure 7.

In 1986, Dahlgren ran a workshop on high-power switching for Navy tactical and Department of Defense (DoD) strategic applications and became involved with numerous DoD working groups on electromagnetic propulsion, high-power diagnostics, advanced energy conversion, power modulators, and pulsed power. Spark gaps were investigated to create underwater noise for submarines. Dahlgren also led four North Atlantic Treaty Organization (NATO) Advanced Study Institutes in Europe and the UK on various pulsed-power topics. International assessments of key pulsed-power technologies were also performed.

Particle-Beam Weapons

Particle-beam weapons were a major focus of DE work during the 1970s and 1980s. A CPB weapon takes subatomic particles, generally electrons, and accelerates them to near the speed of light before sending them toward a target. These fast electrons penetrate deeply into most materials, so they are difficult to counter. The high-current electron beam was to be accelerated by an induction-type

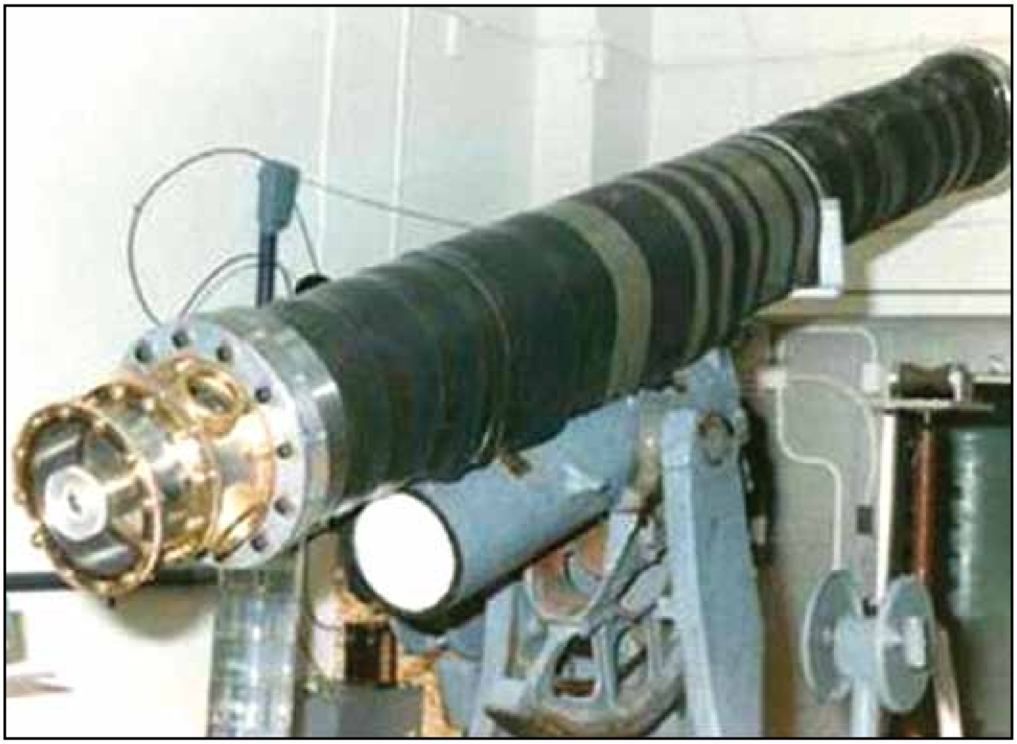

Figure 6. A Water Pulse-Forming Line and Spark-Gap Switch Test

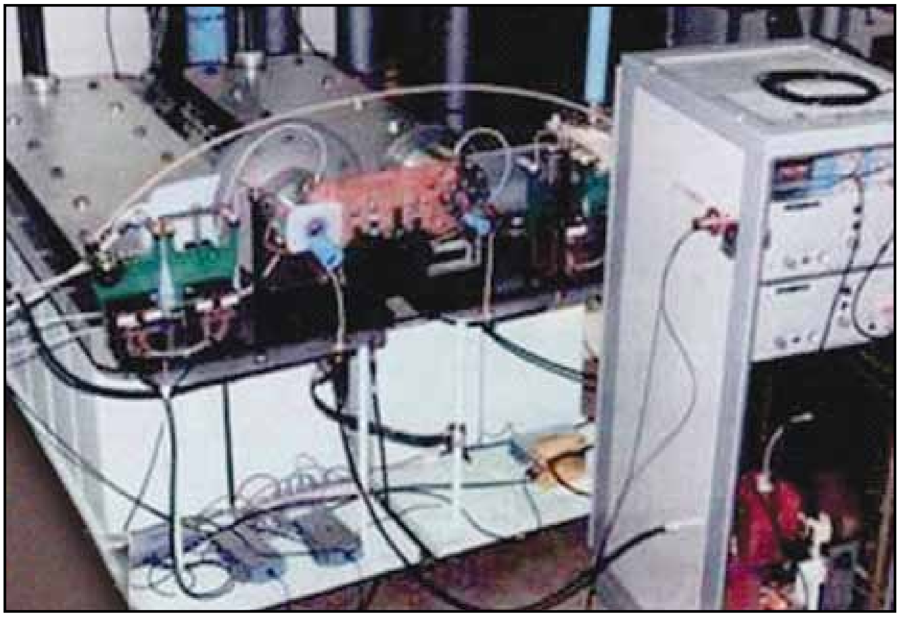

Figure 7. High Energy 2-Pulse System

accelerator, repetitively pulsed. High electron-beam currents (kiloamps) and a hole-boring series of pulses were anticipated to create a stable, long-range beam. Since the beam was capable of penetrating quickly and deeply into any target material, it had the potential to damage electronics and set off explosives before salvage fuzing could occur. The beam was predicted to be all-weather and essentially countermeasure-proof. Even a near miss could cause substantial damage from high fields and X-rays produced by the deceleration of electrons as they hit air molecules near the target. The CPB concept is shown in Figure 8.

Scientists and engineers from Dahlgren worked on the pulsed-power technologies needed to drive these machines beginning in 1980, and it became a major focus of the Pulsed Power Technology Program.[14] The White Oak Laboratory developed beam-steering concepts and looked at material interactions. By 1989, the program investigated:

- Propagation
- Compact Recirculating Accelerators
- Pointing and Tracking
- Prime Power
- Material Interaction
- Fratricide

For a compact shipboard system, recirculating accelerators were needed to make multiple passes of the electron beam past the accelerating cavities. This required a high-power, fast recovery switch, which Dahlgren began working on in 1988. Using patented hydrogen switches and special triggering techniques—efforts that had begun with internal research funds—Dahlgren demonstrated spark-gap switches, the only technology that could meet the current, voltage, and recovery requirements at that time.[15] The High-Voltage 5-Pulse System experiment is shown in Figure 9.

During these technology efforts, significant advances were achieved in all aspects of the program. These included:

- Generating high-current, high-energy beams (although still below weapons parameters)
- Demonstrating a 360º turn in a high-current beam
- Propagating a single pulse through the air
- Demonstrating beam steering on a small scale
- Performing target interaction measurements

Multipulse, long-range propagation was never demonstrated. A comprehensive tri-service summary called the Net Technical Assessment for CPB was sponsored by the Defense Advanced Research Projects Agency (DARPA) in 1987 to describe the accomplishments of the program. The report said compact accelerators were the most pressing technology need. As a result, most funding was directed toward this topic. Funding was stopped in the early 1990s, however, due to the high expense, stretched timelines, and changes in the threat.

PULSED POWER AND ELECTROMAGNETIC LAUNCHERS

During the 1980s, the Army and Air Force looked at short-range electromagnetic weapons to penetrate stronger armor with higher velocities. The Navy worked on concepts for a weapon that could be mounted on ships to intercept missile systems at line-of-sight distances. The Navy—then the biggest user of space systems—was also interested in studies showing that small satellites could

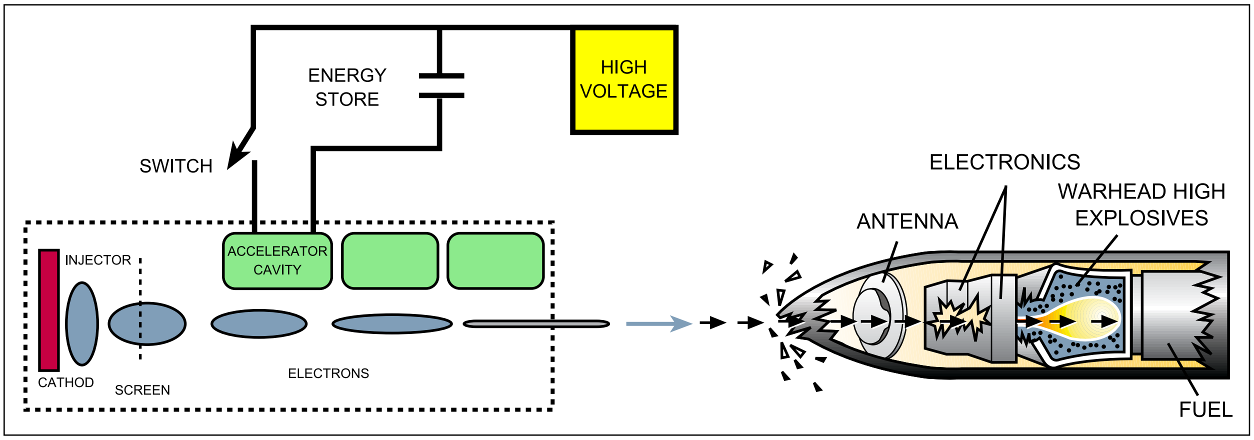

Figure 8. Charged-Particle Beam (CPB) Concept

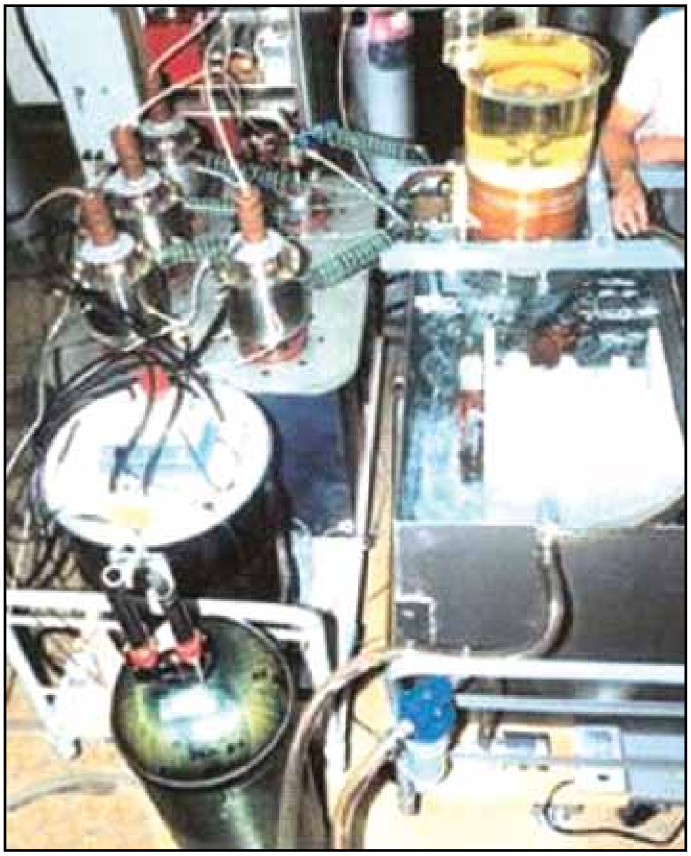

Figure 9. High-Voltage 5-Pulse System Experiment

be electromagnetically launched into low Earth orbit for the fraction of the cost for a normal launch.

Through the 1980s, electric guns were funded by independent research and independent exploratory development programs at Dahlgren, studying electric gun concepts for both rail guns and electrothermal (ET) guns. Kinetic energy weapons were also investigated as part of the Pulsed Power Technology Program. Under these programs, pure electric launchers were developed and tested at Dahlgren, including ones that self-formed projectiles.[16–18] Also studied were ET guns that used the discharge of electrical energy at the gun breech to generate a plasma jet. This plasma jet heated a low-molecular-weight working fluid, such as water, to produce a heated gas that accelerated the projectile to higher velocities than conventional explosives. The Electrothermal-Chemical (ETC) Gun concept augmented the electrical energy generating the plasma jet with a chemical reaction. A 127mm ETC gun was investigated, and a 60mm ETC gun was tested at Dahlgren, with the ability to fire short bursts at a rate of 100 rounds per minute.[19]

Early Dahlgren work on electromagnetic launchers—along with capacitor development and switch advances from the Pulsed Power Technology Program—allowed Dahlgren to provide the Navy with detailed conceptual designs in the late 1990s for near-term, long-range rail guns based on capacitor energy store. These efforts helped support the decision to begin a long-range rail-gun program at Dahlgren that continues today, resulting in world-record achievements. Capital investment funds were used to construct a high-energy facility in 2005 to test pulsed-power components and module designs for use in electromagnetic launcher programs. An early electromagnetic launcher is shown in Figure 10.

High-Energy Lasers (HELs)

In general, megawatts of continuous laser power are required to kill hard targets at long ranges. Laser technologies that can produce this much power are very limited. The Navy was a leader in developing powerful chemical lasers in the 1970s and 80s. These lasers burned chemical reactants to

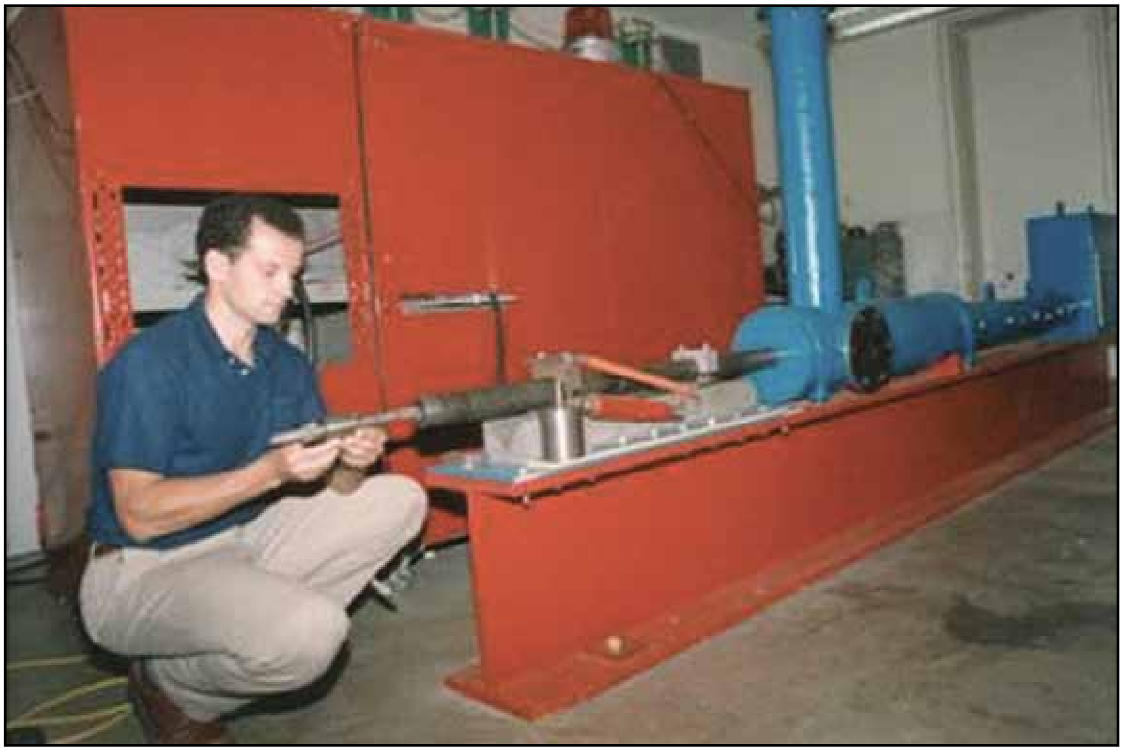

Figure 10. Early Electromagnetic Launcher at Dahlgren

generate the excited states for lasing, thus reducing the need for large amounts of electrical power. The Navy built an entire HEL system, including the Mid-Infrared Advanced Chemical Laser (MIRACL) and the Sea-Lite beam director. By 1990, this building-sized system demonstrated shooting boosters, missiles in flight, and supersonic vehicles. However, the system had drawbacks because it:

- Used hazardous, expensive chemicals
- Had propagation problems at the mid-infrared wavelength
- Was large in size and high in cost

FELs require electron accelerators similar to CPB weapons, so they also are large and complex. However, they can be designed to operate at optimum wavelengths and scale nicely to higher powers. The Strategic Defense Initiative began working on FELs in the late 1980s, funding the advanced test accelerator at LLNL, originally developed for CPBs. FELs were also studied under the Strategic Defense Initiative Organization (SDIO) to be used as an antisatellite weapon. These lasers went from milliwatts to watts under SDIO, and then to kilowatts more recently with work at the Thomas Jefferson National Accelerator Facility in Virginia. Space-based lasers and relay mirror systems were studied under SDIO funding, too, including the development of the Advanced Beam Control System for beam steering, beam control, rapid optical retargeting, and self-alignment.

Dahlgren engineers concentrated its internal laser efforts on medium-power soft-kill weapons. They performed tests against sensors and cameras, and investigated damage thresholds. In the late 1980s, Dahlgren engineers worked with optical augmentation to locate enemy optics for targeting and on green laser dazzlers for defense against small-boat attack. There were efforts to harden electro-optical equipment, including sights and night-vision systems for the Marines, and laser eye-protection filters for goggles and binoculars. Laser systems were also investigated for remotely cutting holes and wires to disable electronics. Lethality work continued under funding from the Joint Technology Office for High-Energy Lasers to look at alternative wavelengths and pulse shapes in addition to modern target materials.[20]

Dahlgren scientists continued to investigate laser-damage thresholds for materials, components, and subsystems for a variety of laser technologies. Near the start of the 21st century, commercial lasers based on pumping optical fibers with semiconductor lasers became common and more powerful. Dahlgren purchased the Navy's largest collection of fiber lasers in 2004 and began investigating ways to combine multiple beams into a laser weapon. These lasers have very high efficiencies, above 20 percent, and the fiber-optic output reduces the requirement for complex optical paths. In 2008, Dahlgren engineers demonstrated a laser capability to ignite spinning mortar rounds, and in 2009, engineers demonstrated the capability of fiber lasers in a shoot down of soft targets at China Lake, California.

Resurgence of Directed Energy

With the fall of the Soviet Union and a greatly altered threat, DoD funding (particularly technology funding) experienced an overall decline in the late 1980s and early 1990s. This caused Navy managers to emphasize near-term, lower risk, evolutionary concepts. The Pulsed Power Technology Program and the Navy's Charged Particle Beam Program both came to an end. Investigations into HPM weapons declined as the difficulty of burnout of military electronics—particularly analog components—became apparent. Problems with propagation and cost caused the Navy to greatly reduce efforts on chemical lasers. With the cancellation of major programs, Dahlgren used internal funding in 1990 to keep a core technical capability together, which was necessary for the Center to remain in the mainstream of tactical DE and its associated technologies. Efforts continued in water breakdown, testing of contractor-developed pulsed-power components, and electric guns. New talent and technologies from universities were brought in to jump-start new projects. Tunable waveform generators using unique semiconductor materials were developed. These used bulk semiconductor material, fabricated in-house, that could be used as a fast switch controlled by laser light for both on and off operation. This allowed faster repetition rates and better triggering than could be done with small spark gaps, as well as the ability to create specific waveforms.[21] "Green" technologies were also investigated using nonthermal plasmas and spark-gap shock waves for cleaning and pollution reduction.[22] New types of particle detectors and magnetic field sensors were developed, and new methods of infrastructure protection were investigated.[23] Soft-kill weapons, both optical and HPM, continued to be studied. Short-pulse jamming of spread-spectrum systems was investigated, as well as beat-wave coupling and special waveforms.[24]

A number of trends led to a resurgence of DEWs by the end of the 20th century. The DoD trend in using digital electronics and off-the-shelf commercial technologies increased dramatically. The pace of change in electronics and computers changed rapidly, too. Most of these new electronic systems had never been tested for vulnerability, and there was a question of how much they would increase military vulnerability to RF or HPM attack. The reduced emphasis on nuclear EMP shielding meant more military electronics were not as well protected from RF attack. Consequently, interest in protecting U.S. military and civilian infrastructure increased, including systems in foreign countries. Moreover, with the increasing reliance on civilian infrastructure, such as power, communications, and emergency and industrial systems—all of which were controlled by digital electronics—the potential that an adversary could attack infrastructure systems to affect or divert military operations became an increasing concern. Following several major terrorist attacks during this time period, there was also concern about the impact of an RF attack on airport towers, financial systems, alarm systems, and industrial plants. Human factors—such as a state of confusion experienced by humans—also played an important part in determining the overall effects of an RF attack.

The asymmetric threat—where large numbers of cheap weapons in a swarm attack could overrun a few sophisticated weapons—caused more concern. As the asymmetric threat to the surface Navy pushed the limits of conventional defensive systems, DE—with it speed-of-light propagation, soft-kill potential, and cheap rounds—offered tactical advantages, either as an adjunct to conventional systems or as stand-alone systems. Additionally, there was an increased emphasis on nonlethal, precise accuracy and graduated effects that could be used. Moreover, the idea that future battles would be fought together with civilians and friendly forces on the battlefield increased the importance of low collateral damage and antimateriel attacks.

The Joint Program Office for Special Technology Countermeasures (JPO/STC), located at Dahlgren, began efforts concerning the vulnerability of new digital systems to RF attack. The program also established a DoD-wide database of vulnerability data, source designs, and RF-effects information—bringing together much of the information collected by the services over the years. The program looked at the protection of modern digital infrastructure systems and funded a facility constructed in 1992 to test large-scale electromagnetic vulnerabilities to various methods of attack.

In the late 1990s and early 2000s, Dahlgren initiated programs regarding the potential for RF attack using nonkinetic disruption, with minimal collateral damage. Capital investment funds were used to construct a test facility for this effort in 1998. Dahlgren developed RF payloads for remotely piloted vehicles and demonstrated their effectiveness in field tests in 1999, and in similar tests in 2007. The successful completion of Project Guillotine was DoD's first demonstration of this type of HPM technology. As the need for statistical vulnerability to commercial digital systems became apparent, Dahlgren constructed instrumented test facilities in 1999 and 2002. Two multistory buildings could be reconfigured to reflect different types of building construction and electromagnetic shielding. Large complexes of electronics, computer networks, server systems, telephone systems, security systems, and various types of digital industrial controls could be assembled, instrumented and exposed to attack from an external device or technique. This program-funded complex—called the Maginot Open Air Test Site (MOATS) facility—continues to be used to test target systems, as well as a variety of RF weapon technologies developed internally and by external and international organizations. A picture of the MOATS facility is shown in Figure 11.

As the need for additional DE laboratory space and testing capabilities became apparent, Dahlgren applied for military construction funds, and in 2008, constructed the Naval Directed Energy Center (NDEC), with access to Dahlgren's overwater test range. Other construction funds were used to construct a remote facility at the Pumpkin Neck Explosive Test Range to serve as a laser backstop and measurement facility, as well as an explosive-test staging area. These facilities already have been used to develop and test fiber lasers against modern threat targets. Construction is currently underway to build an expansion of the NDEC and a 120-m laser test laboratory building using an existing tunnel structure. This collection of facilities represents very important capabilities to develop and test future DE systems.

CONCLUSION

For over 40 years, the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) has been a leader in developing DE devices, pulsed-power systems, and electric weapons. Its people have contributed many publications and patents, and set world records. DEWs tend to be complex and technically challenging to build. Regardless, these weapons offer important, powerful advantages, such as:

- Deep Magazines
- Cheap Rounds
- Fast Targeting
- Variable Lethality
- Pinpoint Targeting

As a result of NSWCDD's leadership, persistent scientific initiatives, and leading-edge engineering

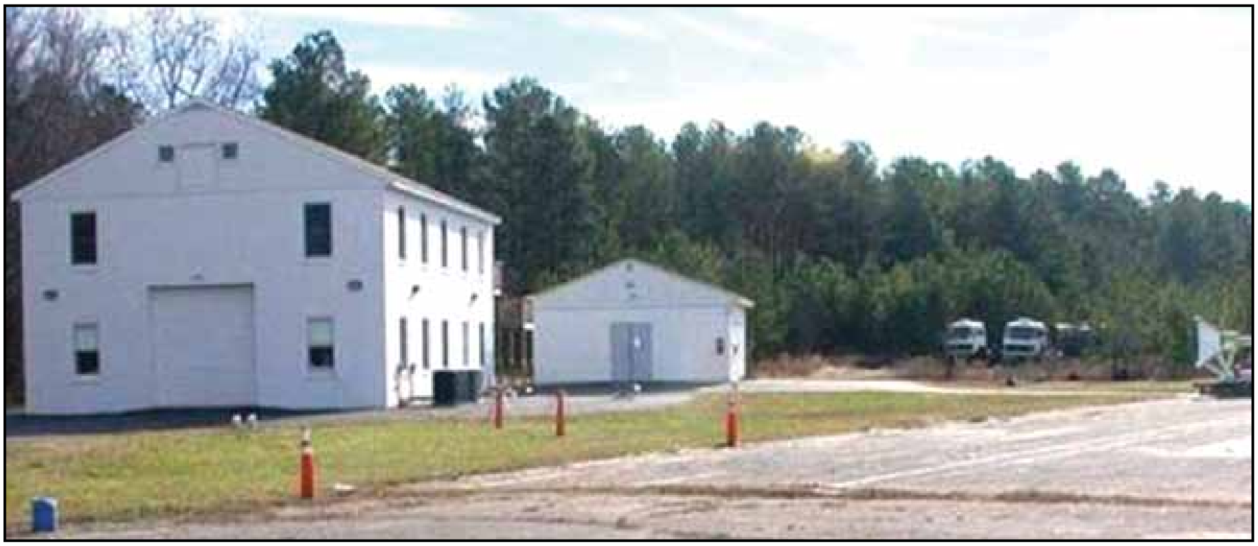

Figure 11. MOATS Facility Undergoing Testing with an RF Weapon (on right)

over the years, naval warfighters will increasingly find themselves turning to DEWs when dealing with situations spanning the spectrum of conflict.

References

1. Moran, S.L., "High Repetition Rate L-C Oscillator," *IEEE Transactions on Electron Devices*, Vol. ED-26 No. 10, October 1979.
2. Fitch, R.A. and Howell, V.T.S., "Novel Principle of Transient High-Voltage Generation," *Proceedings IEEE*, Vol. III, April 1964.
3. Ramrus, A. and Rose F., "High-Voltage Spiral Generators," 1*st International Pulsed Power Conference*, Lubbock Texas, pIIIC9, November 1976.
4. Landecker, K.; Skatterbol, L.V. and Gowdie, D.R., "Single Spark Ring Transmitter," *Proc. IEEE*, Vol. 59, No. 7, July 1971, pp. 1082–1090.
5. Erickson, D.J.; Fowler, C.M. and Caird, R., *EM Radiation from an Explosive Generator System Coupled to a Large Antenna*, Los Alamos Scientific Laboratory Report M-6-258, March 1978.
6. Fowler C.M. and Brooks, M.L., *Explosive Flux Compression Devices Project*, ARPA Order No. 2161, Semiannual Technical Report, 1 March 1972–31 August 1972.
7. Zahn, M.; Ohki, Y.; Fenneman, D.; Gripshover, R. and Gehman, V., "Dielectric Properties of Water and Water/Glycol Mixtures for Use in Pulsed Power System Design," *Proceedings of the IEEE*, September 1986, pp. 1182–1221.
8. Gehman, Jr., V.H.; Gripshover, R.J.; Berger, T.L.; Bowen, S.P. and Zia, R.K.P., "Effects of Filtration on the Impulse Breakdown Strength of High-Purity Water," *IEEE 1995 Tenth International Pulsed Power Conference*, Albuquerque, New Mexico, 10–13 July 1995.
9. Moran, S.L., *Pulsed-Power Technology Program: Accomplishments in Switching*, NAVSWC Tech Note TN-89-259, 1990.
10. Berger, T.L.; Gehman, Jr., V.H.; Lindberg, D.D. and Gripshover, R.J., "A 2.5 Gigawatt Liquid Dielectric Coaxial Pulse Forming Line," *Proceedings of the 6th IEEE International Pulsed Power Conference*, Arlington, Virginia, 29–30 June and 1 July 1987.
11. Bernardes, J.S., "A High Power Fast Response Switch Testing System," *7th IEEE Pulsed Power Conference*, June 1989.
12. Moran, S.L. and Hutcherson, R.K., *High-PRF High-Current Switch*, U.S. Patent No. 4,912,396, 27 March 1990.
13. Moran, S.L. and Hardesty, L.H., "High-Repetition-Rate Hydrogen Spark Gap," *IEEE Transactions on Electron Devices*, Vol. 38, No.4, April 1991.
14. Rose, M.F. and Huddleston, C.M., *Pulsed Power and Directed Energy Technology*, 10th Anniversary Technical Symposium, NSWC MP 84-560, October 1984.
15. Moran, S.L., "Hydrogen Spark Gap for High-Repetition-Rate Compact Accelerators," *NSWCDD Technical Digest*, NSWCDD/MP-92/104, September 1992, pp. 126–139.
16. Bernardes, J.S. and Merryman, S.,"Parameter Analysis of a Single Stage Induction Mass Driver," *Proceedings of the Fifth IEEE Pulsed Power Conference*, Arlington, Virginia, June 1985, pp. 552–555.
17. Bernardes, J.S., "Muzzle Velocity Control with Electrothermal Guns," *Proceedings First Navy Independent Research/Independent Exploratory Development Symposium*, Laurel, Maryland, June 1988, Vol. 2, pp. 123–131.
18. Bernardes, J. S., "Focusing Induction Accelerator," *Proceedings of the Sixth IEEE Pulsed Power Conference*, Arlington, Virginia, June 1987, pp. 735–738.
19. Bernardes, J.S., *Sliding Breech-Block for Repetitive Electronic Ignition*, Patent Number: 5,233,902, 10 August 1993.
20. Kim, H.S. et al., "Effects of Chopping Laser Penetration of Metal Targets," *Applied Physics Letters*, 55 (8), 21 August 1989.
21. Stoudt, D.C. and Schoenbach, K.H., "The Electrical Characteristics of Semi-Insulating GaAs for High-Power Switches," *7th IEEE Pulsed Power Conference Proceedings*, June 1989.
22. Grothaus, M.G. et. al., "Gaseous Effluent Treatment using a Pulsed Corona Discharge," *Proceedings of 10th IEEE Pulsed Power Conference*, Albuquerque, New Mexico, July 1995.
23. Gripshover, R.J. et. al., *Power Frequency Magnetic Field Detection Systems*, Patent No. 6,483,309, 19 November 2002.
24. *Spread Spectrum Countermeasures Final Report*, E-Systems, Inc., Melpar Division under NSWC contract # N68786-82-C-0109.

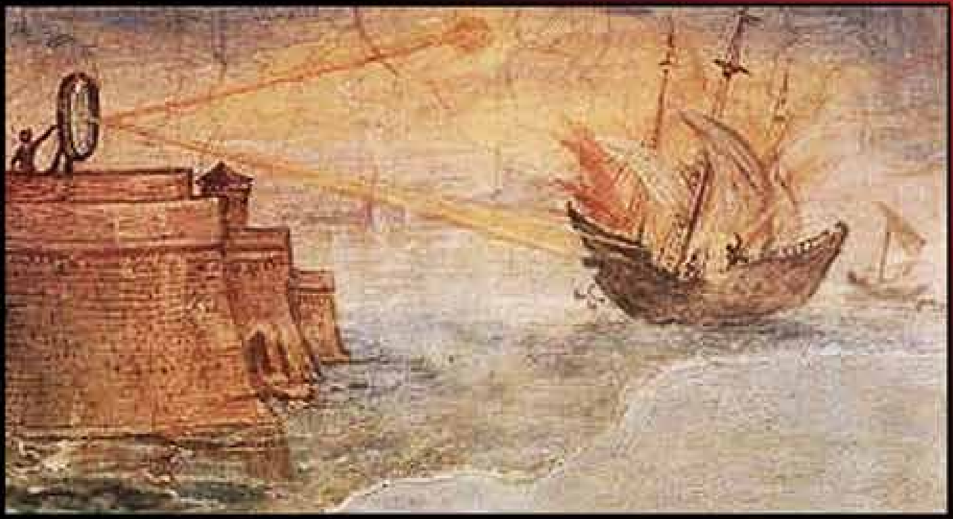

HISTORY OF LASER WEAPON RESEARCH

By Melissa Olson

The idea of using light as a weapon can be traced back to Hippocrates, commander of the Greek forces in 212 B.C. His forces supposedly set fire to the sails of the Roman fleet by focusing sunlight with mirrors. Weapons systems based on lasers and "ray guns," long a staple of science fiction, have captured the imagination of people everywhere. But with steady progress toward the development of lasers in the last 40 years, viable, state-of-the-art laser weapon systems have now become a reality.

The production of lasers in the modern scientific world is fairly new. The first laser was developed in the 1960s and represented the beginning of a drastic change in how the military viewed warfare. The late 1970s and 1980s, too, marked a busy time period for developing lasers into possible weapon systems. All branches of the military and industry were striving to master high power levels, beam control, and adaptive optics. In 1999, the Department of Defense (DoD) formally recognized lasers as future weapons and began research and development (R&D). In 2000, the Joint Technology Office for High Energy Lasers was formed to bring all laser technologies together to develop a complete laser weapon system that could be used by the warfighter.

ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum contains all the types of electromagnetic energy, including radio waves, microwaves, infrared, visible light, ultraviolet, and gamma rays. ***Laser*** is an acronym for "light amplification by stimulated emission of radiation." Light, therefore, is a type of electromagnetic radiation. Light is made up of tiny packets of energy called photons. The amount of energy is what determines the wavelength. Lasers are usually infrared (1 mm to 750 nm) and visible light (750- to 400-nm wavelength). Microwaves are mostly high-frequency radio waves (millimeters to centimeters), with wavelengths 10,000 times longer than lasers. Diffraction of any electromagnetic radiation beam is based on the wavelength and aperture size. For the same aperture size, lasers diffract 10,000 times less than microwaves. This allows the beam to reach farther ranges while maintaining a small spot size of concentrated energy on the target. Lasers are preferred in specific scenarios because of minimal diffraction. The electromagnetic spectrum is shown in Figure 1.

LASER FUNDAMENTALS

The quantum mechanical idea of stimulated emission of light was discovered by Albert Einstein in 1917 and is one of the fundamental ideas behind the laser. Einstein

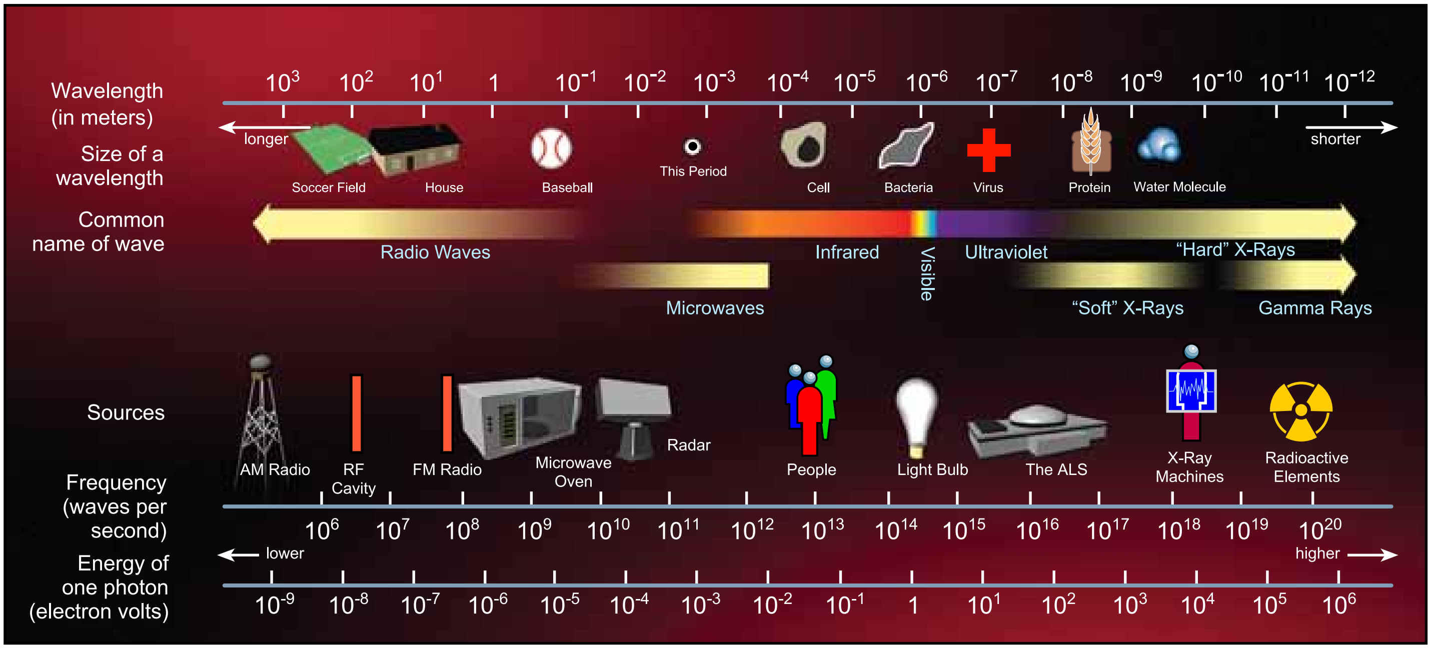
Wavelength (in meters)
10^{3} 10^{2} 10^{1} 1 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 10^{-6} 10^{-7} 10^{-8} 10^{-9} 10^{-10} 10^{-11} 10^{-12}
longer
shorter
Size of a wavelength
Soccer Field
House
Baseball
This Period
Cell
Bacteria
Virus
Protein
Water Molecule
Common name of wave
Radio Waves
Infrared
Visible
Ultraviolet
"Hard" X-Rays
Microwaves
"Soft" X-Rays
Gamma Rays
Sources
AM Radio
RF Cavity
FM Radio
Microwave Oven
Radar
People
Light Bulb
The ALS
X-Ray Machines
Radioactive Elements
Frequency (waves per second)
10^{6} 10^{7} 10^{8} 10^{9} 10^{10} 10^{11} 10^{12} 10^{13} 10^{14} 10^{15} 10^{16} 10^{17} 10^{18} 10^{19} 10^{20}
lower
higher
Energy of one photon (electron volts)
10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}

Figure 1. Electromagnetic Spectrum

theorized that when a photon interacts with an atom or molecule in an excited state, two photons are produced when the atom or molecule leaves the excited state. Population inversion occurs when the atoms or molecules are in the excited state. In order for molecules to come out of the normal "ground" state, a source of power must be introduced to the system energizing the atoms to the excited state. When many photons are passed through many excited atoms, more and more photons are produced. The photons are contained and reflected back and forth in a cavity, with mirrors usually on each end. The mirror on the output end is only partially reflective, allowing some photons to leak through, creating the laser beam.

The difference between an everyday light bulb and the light of a laser is temporal and spatial coherence. In a light bulb, the light emits photons equally in all directions. The light is random, out of phase, and multiwavelength. A laser emits coherent light, so photons travel in identical direction and phase. A laser is also monochromatic, i.e., light of one wavelength. Another significant difference is that laser light is highly collimated, which means the laser beam can travel long distances with minimum spreading.

The laser gain medium through which the photons travel to become amplified or magnified can vary. The source of power used to excite the medium, achieving population inversion, can be the result of a chemical reaction, an electric discharge, a flash lamp, another laser, or some other excitation mechanism. The type of the lasing medium determines the type of laser. The three categories in which lasers are usually classified are chemical, gas, and solid state. A laser can also be continuous wave (CW) or pulsed. Each type of laser produces a specific wavelength of radiation. It is important to note that different wavelengths of radiation interact with the atmosphere differently. A laser beam is either scattered or absorbed by air molecules, water vapor, or dust. Longer wavelengths scatter less and are absorbed more than shorter wavelengths; our sky is blue because the shorter blue wavelengths of light are scattered more than the longer wavelengths.[1] Gamma rays are so highly absorbed that they cannot propagate more than a few feet in the air. Thus, some laser wavelengths are scattered or absorbed more than others. This makes laser wavelengths with minimum absorption better for use as directed-energy weapons since they propagate through the atmosphere better than others. For example, the carbon-dioxide (CO_2) laser is strongly absorbed by water vapor, so any use near the ocean will be negatively affected. Near-infrared and infrared lasers have shorter wavelengths with negligible absorbance. The optimal laser choice, therefore, would be a wavelength-tunable laser that could vary depending on the atmospheric conditions, such as the free-electron laser (FEL).

Lasers have affected almost every type of modern technology. Most laser technologies use low powers and were mastered very quickly. They are used in many everyday appliances, such as scanning/inventory devices, surgery/medicine, hair removal, presentation pointers, law enforcement, ranging and sighting devices, welding applications, and much more. Using a laser as a weapon has many advantages. For example, a laser:

- Is unaffected by gravity
- Causes minimal collateral damage
- Travels at the speed of light
- Can precisely reach far distances
- Is capable of causing a specific, predetermined amount of damage to targets

The theory behind these capabilities makes the laser weapon a prime choice in multiple engagement scenarios. However, developing lasers with higher powers to use as a weapon has proven more difficult than first considered.

Military Laser History and Laser Types

Generally, a laser weapon is any laser used against the enemy with more than 50 kW to megawatts of power. This is much greater power than commercial lasers. Accordingly, they have greater support needs, including:

- Environmental and personnel safety
- Mirror coatings
- Chilling requirements
- Power requirements
- Laser fuel storage
- Alignment and tracking requirements

In 1960, the very first laser (a ruby laser) was built, producing minimal power. This event was followed by many other laser technology developments. The first chemical laser, hydrogen fluoride (HF), was built in 1965, producing 1 kW. It was then that DoD became interested in researching and developing a more powerful laser for weapon applications. Subsequently, in 1968, the Defense Advanced Research Projects Agency (DARPA) Baseline Demonstration Laser produced 100 kW, and the Navy-ARPA Chemical Laser (NACL) produced 250 kW in 1975. The very first laser is depicted in Figure 2.

Solid-State Lasers (SSLs)

An SSL uses a solid lasing medium, such as a rod made up of glass or crystal, or a gem, like the ruby laser. Along with the rod or host material is an active material, such as chromium, neodymium, erbium, holmium, or titanium. Chromium is the active material used in ruby lasers. Neodymium is the active material in the most widespread applications. A flash lamp, arc lamp, or another laser carries out the optical cavity pumping to achieve population inversion and stimulate the laser beam. The Neodymium Yttrium-aluminum garnet (Nd:YAG) laser is a popular SSL. It operates at a 1064.5-nm wavelength and can be pulsed wave or CW. A great advantage of these lasers is that the wavelength and pulse duration can be varied considerably.[1] The power level can reach up to megawatts when using Q-switching to achieve

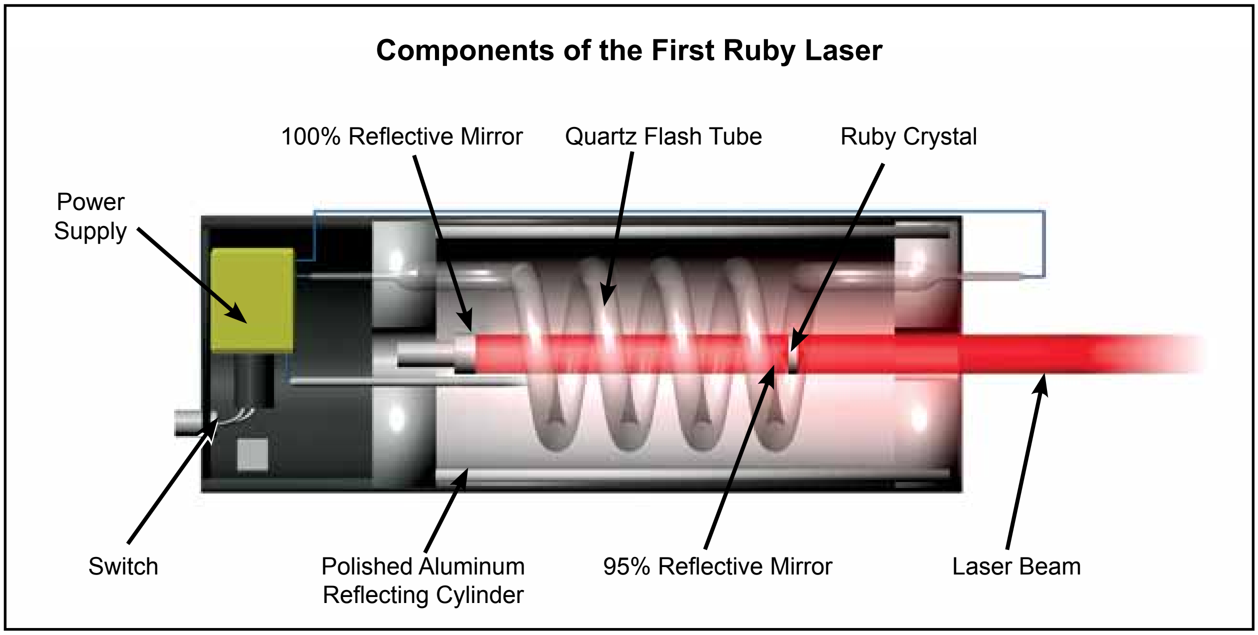

Figure 2. First Ruby Laser Developed in 1960 by Research Physicist Theodore H. Maiman

short pulse lengths. The various interactions with the laser and different crystalline materials can double the electromagnetic frequency, which will halve the wavelength, bringing the laser beam into the visible range, 532 nm (green). The wavelength can be further divided down three or four times, making this laser range from the near-infrared to the ultraviolet wavelength. These lasers are commonly used for rangefinders and target designators. Other advantages of these lasers are that they can be made very small, rugged, cheap, and battery-powered. Characteristics of SSLs are shown in Table 1.

Chemical Lasers

A chemical laser uses chemical reaction to create population inversion in the lasing medium. One example is the Mid-Infrared Advanced Chemical Laser (MIRACL) developed in the mid-1980s. The MIRACL is a continuous-wave, mid-infrared (3.8-μ) laser. Its operation is similar to a rocket engine in which a fuel (ethylene, C2H4) is burned with an oxidizer (nitrogen trifluoride, NF3).[2] Free, excited fluorine atoms are among the combustion products. Just downstream from the combustor, deuterium and helium are injected into the exhaust. Deuterium (U) combines with the excited fluorine to create excited deuterium fluoride (DF) molecules, while the helium stabilizes the reaction and controls the temperature.[2] The laser's resonator mirrors are wrapped around the excited exhaust gas, and optical energy is extracted. The cavity is actively cooled and can be run until the fuel supply is exhausted. The laser's megawatt-class output power can be varied over a wide range by altering the fuel flow rates and mixture. The laser beam in the resonator is approximately 21-cm high and 3-cm wide. Beam-shaping optics are used to produce a 14- × 14-cm (5.5- × 5.5-inch) square, which is then propagated through the rest of the beam train. Diagnostics for evaluating the beam shape, absolute power, and intensity profile are used on each firing of the laser. The beam can be directed to a number of different test areas or to the SEA LITE beam director.[2] The DF Chemical Laser (MIRACL) and the Sea Lite Beam are shown in Figure 3.

The laser and beam director were integrated in the mid-1980s at the Army's High Energy Laser Systems Test Facility (HELSTF) at White Sands Missile Range, New Mexico. Following integration, extensive tests were conducted in the areas of:

- High-power optical components and beam-path conditioning
- Beam-control techniques
- High-power propagation
- Target damage and vulnerability
- Target lethality[3]

Tests supported by the MIRACL included:

- The high-power dynamic with flying drone (BQM-34)
- Conventional defense initiative with flying drone
- High-velocity target test with Vandal Missile
- High-altitude target tests with flying drone
- Missile and plume tests using the 1.5-m aperture
- Radiometrically calibrated images and spectral radiometry

These successful tests are what made many believe that MIRACL was the first and only successful laser weapon system developed by the Navy prior to the Navy Laser Weapon System (LaWS).[3]

Gas Lasers

Gas lasers are a type of chemical laser that uses a pure gas or gas mixture to produce a beam. The typical gas laser contains a tube with mirrors on each end. One end transmits the beam out of the cavity. Most gas lasers use electron-collision pumping, with electric current passing through the gas. Some use optical pumping with flash lamps. The helium

Table 1. Characteristics of Solid-State Lasers

Name	Wavelength (nm)	Typical Power	Typical Operation
Alexandrite	700-830	5 watts	Pulsed/CW tunable
Erbium	850/1230/1540/1730/2900	8 watts	Pulsed
Holmium: glass	1950	milliwatts	Pulsed
Neodymium	1064/1123/1318/1370	megawatts	Pulsed
Neodymium: glass	1060	megawatts	Pulsed
Neodymium: YAG	1064.5	megawatts	Pulsed/CW
Ruby	694.3	10–15 watts	Pulsed
Titanium-sapphire	660–1060	15 watts	Pulsed/CW tunable

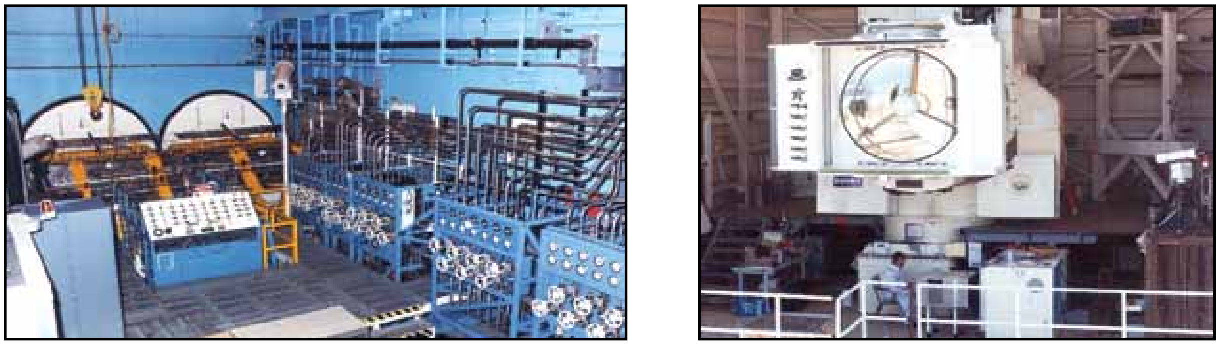
DF Chemical Laser (MIRACL)

Sea Lite Beam

Figure 3. DF Chemical Laser (MIRACL) and Sea Lite Beam

neon (HeNe) laser is a very well-known gas laser. It produces a bright red, continuous beam of low power. It is used for many applications such as scanning, alignment, measurement, and stabilization devices. University students use them in optical training laboratories. Many larger lasers contain a HeNe inside the beam path, as well to verify beam alignment. HeNe lasers are fairly cheap and very rugged. They can work continuously for thousands of hours.

CO_2 lasers are in the gas family. These lasers were the earliest, truly high-power lasers and have been among the most crucial lasers used in R&D for high-energy laser (HEL) weapons. In industry, the more powerful CO_2 lasers are used for welding, drilling, and cutting. There are many different types of CO_2 lasers that vary in pumping design. CO_2 lasers work by burning a hydrocarbon fuel (like kerosene or methane) in oxygen or nitrous oxide. The hot gas flows through a comb of nozzles, expands quickly, and achieves population inversion. The gas then flows through an optical resonator at supersonic speeds, resulting in stimulated emission and a laser beam.[4]

CO_2 lasers have been researched for use as nonlethal weapons. The wavelength produced by a CO_2 laser is also absorbed by glass. For example, the beam does not penetrate a windshield. Thus, shooting a CO_2 laser at a vehicle's windshield could deter a threat by damaging the windshield or by causing a dazzling effect to reduce the visibility of the driver, while not reaching the driver at all.

The gas dynamic laser (GDL) is a CO_2 laser based on differences in relaxation velocities of molecular vibrational states. The laser medium's gas has properties such that an energetically lower vibrational state relaxes faster than a higher vibrational state; thus, a population inversion is achieved in a particular time. A GDL is shown in Figure 4. Characteristics of chemical and gas lasers are identified in Table 2.

Fiber Lasers

Modern fiber lasers are considered SSLs. They are powered by electricity, making them highly mobile and supportable on the battlefield. Fiber lasers use optical fibers as the gain media. In most cases, the gain medium is a fiber doped with rare earth elements—such as erbium (Er3+), neodymium (Nd3+), ytterbium (Yb3+), thulium (Tm3+), or praseodymium (Pr3+)—and one or several laser diodes are used for pumping. Optical fibers have been used in industry, specifically for telecommunications to transport information via light. With developing technology, optical fibers have become high-energy, powerful laser energy sources. Fiber lasers have proven to have much benefit over traditional SSLs. They are rugged and do not require a clean room to operate or maintain, as most other laser systems do. They also are extremely efficient; however, they cannot operate well in all weather conditions. One example is the IPG CW fiber lasers, which produce moderate beam quality, causing damage to materials and components through thermal heating and burn-through. The Naval Surface Warfare Center, Dahlgren Division (NSWCDD) purchased eight commercially available 5.5-kW IPG lasers, where two multimode (seven fibers) lasers are housed per cabinet. This type of laser is easy to mount due to the flexible fibers. The IPG CW Fiber Laser is shown in Figure 5.

Miscellaneous Lasers

There are other types of lasers that do not necessarily fit into the chemical or solid-state categories. These include semiconductor lasers, used in:

- Television
- CD Players
- Dye Lasers
- Spectroscopy
- Radios
- Telecommunications
- Medicine
- Astronomy

There also are the FELs mentioned previously. The FEL is a completely different breed of laser.

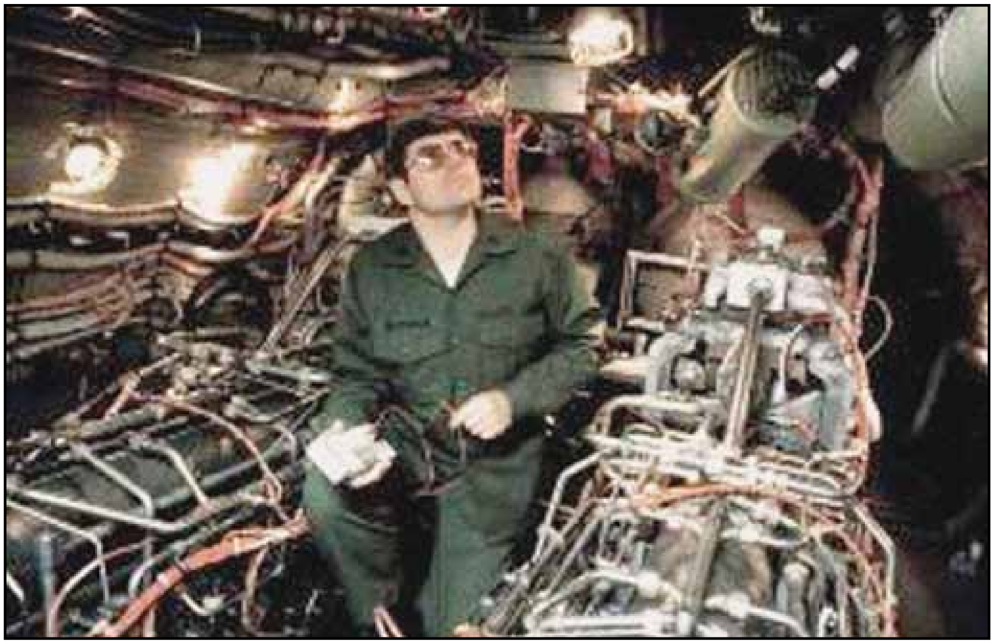

Figure 4. A laser engineer inspects a gas dynamic laser after installation aboard an NKC-135 airborne laser laboratory.

Table 2. Characteristics of Chemical/Gas Lasers

Name	Wavelength (nm)	Typical Power	Typical Operation
Helium-Neon	543/632.8	.0001–.001 watts	CW
Krypton	350–647	.0001–.05 watts	CW
Argon	350–514.5	.001–6.0 watts	CW
Xenon fluoride (excimer)	351	.001–20 watts	CW
Argon fluoride (excimer)	193	.05–30 watts	Pulsed
Krypton fluoride (excimer)	249	7–100 watts	Pulsed
Deuterium fluoride (chemical)	3,000–4,200	.01–100 megawatts	Pulsed/CW
Hydrogen fluoride (chemical)	2,600–3,000	.01–150 megawatts	Pulsed/CW
Carbon dioxide	9,000–12,000	.1–15,000 megawatts	Pulsed/CW
GaAlAs (semiconductor)	750–900	10–4,000 milliwatts	Pulsed

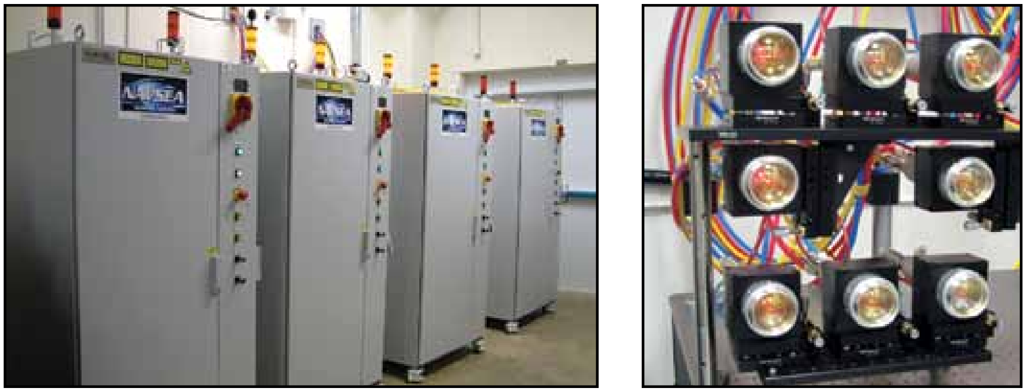

Figure 5. IPG CW Fiber Laser System

It uses electrons to create photons instead of some type of matter. The electrons are produced, collected, and directed to flow at very high speeds. To excite the electrons, they are passed through a "wiggler," i.e., a series of magnets positioned in such a way that electromagnetic radiation (light) is produced when the electrons release photons. The significant feature of the FEL is that the wavelength can be controlled, depending on the magnet positions and the speed of electrons. This versatility makes the FEL particularly appealing. However, the footprint of the FEL system is too large to transform into any ideal defense weapon. The Jefferson Laboratory in Newport News, Virginia, has an FEL and continues to maintain and test its capabilities and effects. This laser was new to the military in the late 1990s and received funding to optimize its capabilities and integrate as a defense weapon. Although great progress has been made, the required footprint could be much larger than desired. Consequently, some interest in the FEL has shifted to other HEL sources.

Many scientists foresee the probability of using the laser as a global weapon. This possibility is proven through basic laws of physics. Actually implementing such a system, however, can be more difficult. The global weapon concept uses a base laser with optics and is strategically positioned in space to be able to direct its beam multiple places on Earth at the speed of light with maximum power levels. This idea faces significant problems, including appropriate power levels, optics to handle such levels, propagation issues, and the ethical measures behind any global weapon. Still, the idea presents interesting possibilities.

Laser Weapon Development

The following paragraphs highlight some of the laser weapons that have been successfully developed over the last 40 years.

Baseline Demonstrator Laser (BDL) Hydrogen Fluoride (HF)

In 1973, TRW Inc. produced the world's first high-energy chemical laser, the Baseline Demonstration Laser, for DoD. After that, TRW Inc. produced and demonstrated six more HELs, including the MIRACL (1985) and Alpha (2000), the nation's only megawatt-class chemical lasers.

Navy-ARPA Chemical Laser (NACL) HF

The NACL was mated with the Navy Pointer Tracker at TRW Inc.'s San Juan Capistrano, California, facilities in the 1975–1978 time frame. This was the Navy's initial, integrated HEL system test bed and was used to provide the first demonstrated kill of an operational missile in 1978.

Alpha HF—Built for Strategic Defense Initiative (SDI) Space-Based Laser (SBL)

Alpha, an HF laser, was the baseline technology for the SBL readiness demonstration (SBLRD). In 1991, the Alpha laser demonstrated megawatt-class power levels similar to MIRACL, but in a low-pressure, space operation environment. Alpha demonstrated that multimegawatt, space-compatible lasers can be built and operated.

Tactical High-Energy Laser (THEL)

The THEL is a DF chemical laser developed by the Army. In 2000 and 2001, THEL shot down 28 Katyusha artillery rockets and 5 artillery shells. On 4 November 2002, THEL shot down an incoming artillery shell and a mobile version successfully completed testing. Subsequently, during a test conducted on 24 August 2004, the system successfully shot down multiple mortar rounds. These tests represented actual mortar threat scenarios in which both single mortar rounds and salvo were tested and intercepted. A photograph of THEL is shown in Figure 6.

Advanced Tactical Laser (ATL)

The ATL uses a closed-cycle, chemical oxygen-iodine laser (COIL) with beam control, which lases at a 1.315-µ wavelength. The ATL was developed to engage tactical targets from a moving platform at ranges of approximately 10 km. It can spot a 10-cm-wide beam on a distant target for up to 100 shots. This beam has enough power to slice through metal at a distance of 9 miles. The aircraft equipped with the ATL weapon system is shown in Figure 7.

A specially modified 46th Test Wing NC-130H aircraft equipped with the ATL weapon system fired its laser while flying over White Sands Missile Range, New Mexico, successfully hitting a target board located on the ground. Equipped with a chemical laser, a beam control system, sensors, and weapon-system consoles, the ATL is designed to damage, disable, or destroy targets with little or no collateral damage.

Airborne Laser (ABL) (CO_2) Chemical Oxygen

The ABL C-130H aircraft contains three laser beam systems: the powerful killing primary laser beam (ATL), a set of illuminating laser beams for infrared surveillance and high-speed target acquisition, and a beacon laser for a high-precision laser target tracking beam control system. The primary

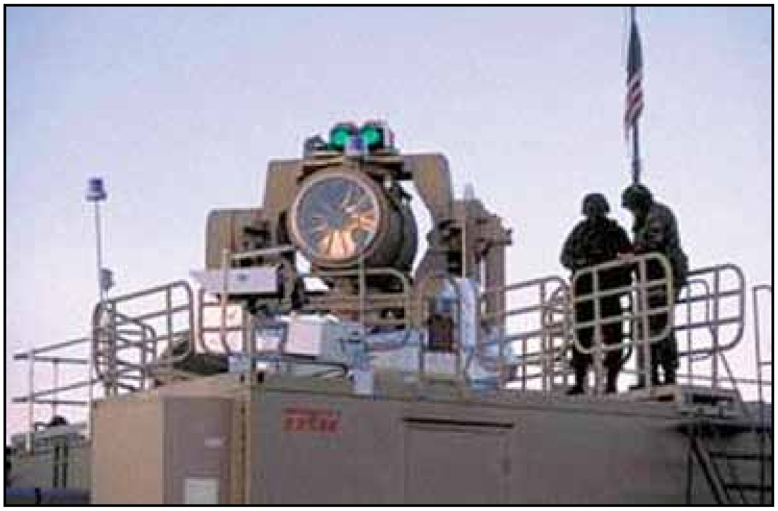

Figure 6. Tactical High-Energy Laser (THEL)

Figure 7. 46th Test Wing NC-130H Aircraft Equipped with the ATL Weapon System

laser beam is generated by a megawatt COIL located at the rear of the fuselage. The high-power laser beam travels towards the front of the aircraft through a pipe. The pipe passes through a Station 1000 bulkhead/airlock, which separates the rear fuselage from the forward cabins. The high-power beam passes through the fine beam control system mounted on a vibration-isolated optical bench. Beam pointing is achieved with very fast, lightweight steering mirrors, which are tilted to follow the target missile. The ABL finally destroyed a target while in flight at White Sands Missile Range in August 2009. The 12,000-lb ABL locked onto an unspecified ground target and fired the laser, making the target disappear. Although it was successful at this demonstration, using the ABL in the fleet has fallen out of favor due to affordability and technology problems. The ABL is shown on an aircraft in Figure 8.

Joint High-Power Solid-State Laser (JHPSSL)

In hopes of accelerating SSL technology for military uses, work is being performed by the U.S. Army Space and Missile Defense Command (SMDC) and the Army Test and Engineering Center at White Sands Missile Range. The technology uses an electric laser diode to shoot light into 32 garnet crystal modules that combine to create “laser amplifier chains” producing 15 kW. By using seven chains and by combining multiple beams, they have reached 105 kW in the laboratory operating in a clean room. The program’s ultimate goal is for a laser system to reach high powers outside a laboratory environment. Fielding such a delicate optical structure can present significant barriers for this laser system. Nonetheless, it will be a great accomplishment for a variety of force protection missions, such as shipboard defense against cruise missiles. The JHPSSL system is shown in Figure 9.

Navy Laser Weapon System (LaWS)

The Navy LaWS is the most recent, successful laser weapon. It uses an electric-fiber laser design, avoiding the problems that chemical lasers present. In the summer of 2009, the Naval Sea Systems Command (NAVSEA)—with support from NSWCDD—successfully tracked, engaged, and destroyed unmanned aerial vehicles (UAV)

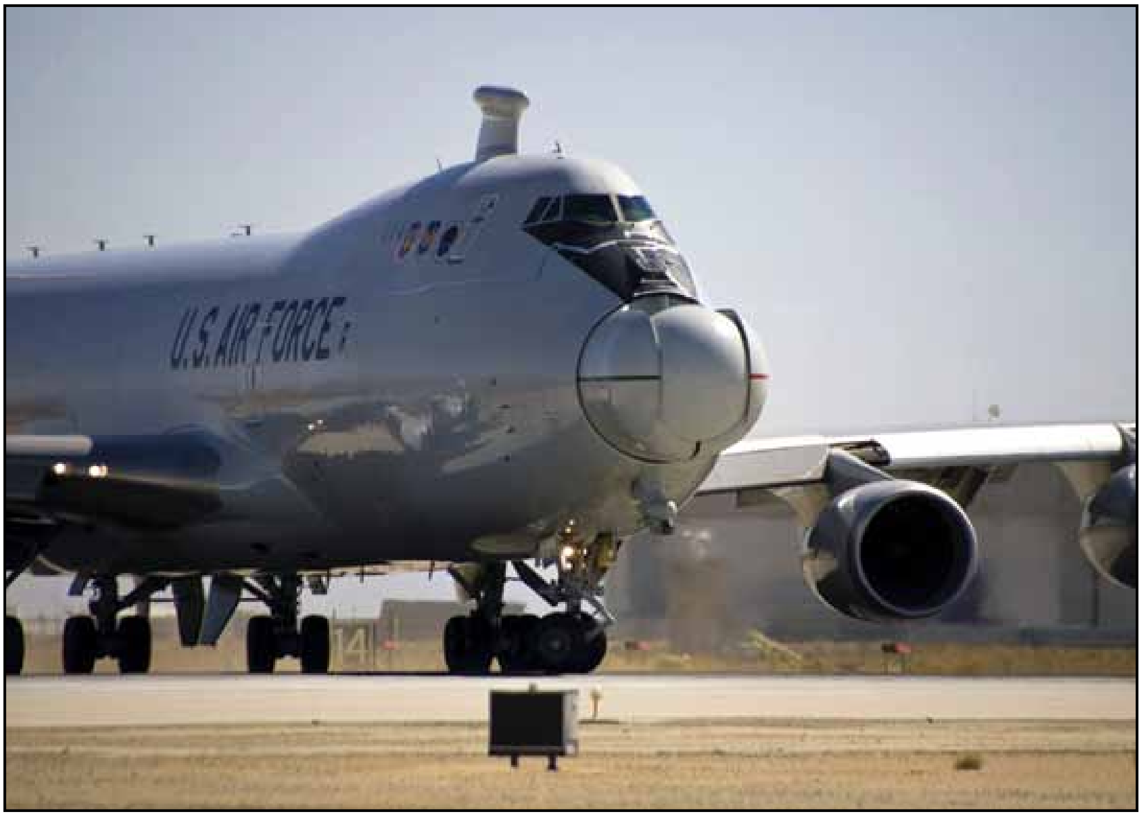

Figure 8. Airborne Laser (ABL)

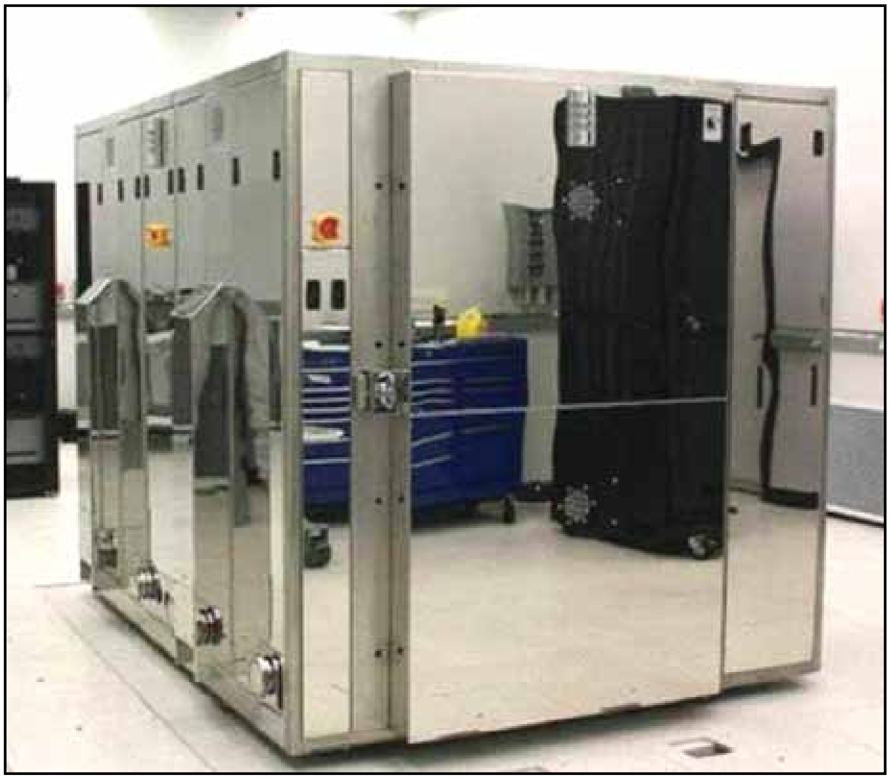

Figure 9. Joint High-Power Solid-State Laser (JHPSSL) System

in flight at the Naval Air Warfare Center, China Lake, California. A total of five targets were engaged and destroyed during the testing, which represented a first for the U.S. Navy. The laser was fired through a beam director on a Kineto Tracking Mount similar to the Sea Lite beam director. The system used fiber lasers in the configuration and has proven to be a rugged and dependable choice for the warfighter's needs. A photograph of LaWS is shown in Figure 10.

Laser weapon systems development in recent years has taken giant steps forward. Dedicated R&D has advanced the state of the art considerably. What was unimaginable only a few short years ago, today has become reality. Accordingly, given continued R&D, warfighters in the near term will have additional weapon options to choose from for dealing with a spectrum of threats and contingencies.

REFERENCES

1. Beason, Doug Ph.D., *The E-Bomb: How America's New Directed Energy Weapons Will Change The Way Future Wars Will Be Fought*, Da Capo Press, 2005.
2. Pike John, *Mid-Infrared Advanced Chemical Laser (MIRACL)*, FAS Space Policy Project, Military Space Programs, March 1998, http://www.fas.org/spp/military/program/asat/miracl.htm
3. Albertine, John R., "History of Navy HEL Technology Development and Systems Testing," (Proceedings Paper), *Proc. SPIE*, Vol. 4632, Laser and Beam Control Technologies, Santanu Basu and James F. Riker, Eds., pp. 32–37.
4. Anderberg, Major General Bengt and Dr. Myron L. Wolbarsht, *Laser Weapons: The Dawn of a New Military Age*, Plenum Press, New York, 1992.

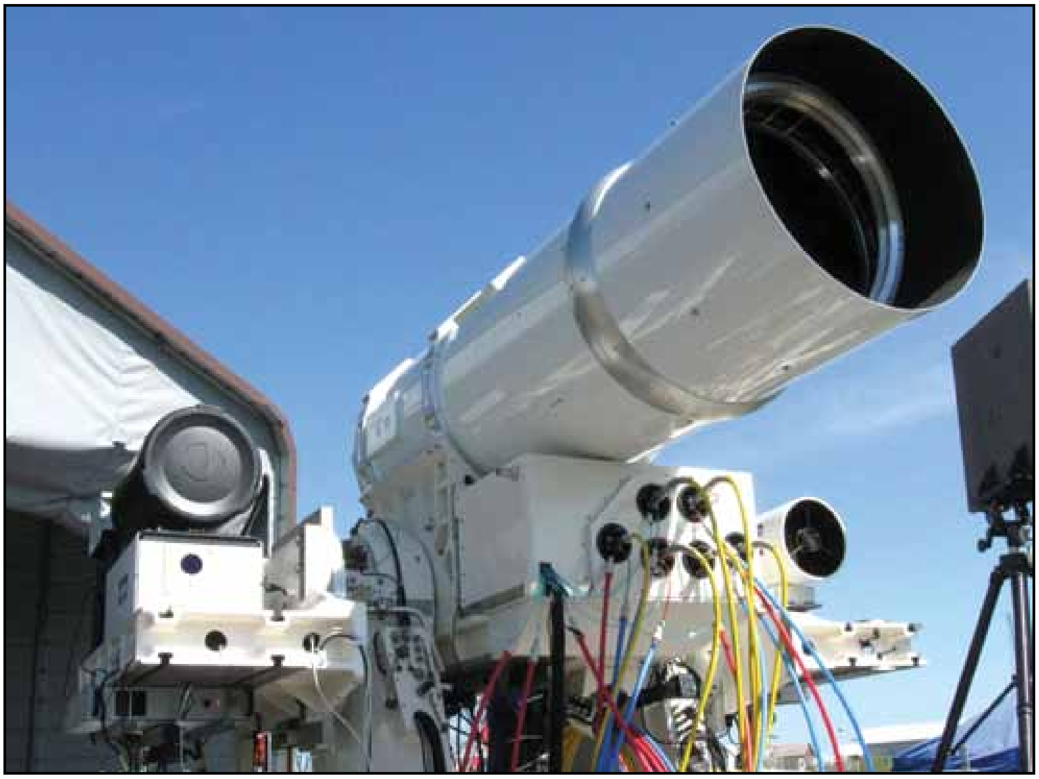

Figure 10. Navy Laser Weapon System (LaWS)

Laser Weapon System (LaWS) Adjunct to the Close-In Weapon System (CIWS)

By Robin Staton and Robert Pawlak

The Naval Sea Systems Command (NAVSEA) established the Navy Directed Energy Weapons Program Office in January 2002 and subsequently chartered the Directed Energy and Electric Weapon Systems Program Office (PMS 405) in July 2004.[1, 2] Its mission is to change the way the Navy fights in the 21st century by transitioning directed-energy and electric weapon technology, providing the warfighter with additional tools to fight today's and tomorrow's wars. In support of this mission, the Laser Weapon System (LaWS) was developed, which potentially adds a suite of tools for offensive and defensive operations.

The LaWS program is managed by PMS 405 in cooperation with the Program Executive Office Integrated Warfare Systems (PEO IWS), the Navy's Close-In Weapon System (CIWS) manager. A multilaboratory/multicontractor organization led by the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), has been executing the program since March 2007. The potential advantages of a lethal, precise, speed-of-light weapon are numerous and have been recognized for many years. However, even in light of these advantages, there are realities that need to be considered for any program to succeed to the point that an actual system is placed in the hands of the warfighters.

The LaWS system offers viable solutions for an important subset of threats while fitting into acceptable size and weight constraints. In addition, since LaWS is a fully electric laser, the operation of the system does not require the handling and storage of hazardous chemicals, such as hydrogen fluoride. As will be discussed later, due to the incorporation of high levels of commercial off-the-shelf (COTS) technology, the LaWS system also has advantages for topside design, logistic supportability, and cost. Thus, LaWS could enable the Navy to address adverse cost-exchange situations, which can occur when engaging proliferated inexpensive threats such as unmanned aerial vehicles (UAVs).

Background

Based on mission analysis work conducted prior to the LaWS program and additional work done as part of the program, it became clear that a number of factors require careful consideration. First, a high-power laser is not likely to replace anything on a ship in the next 5 years. For a new system to be added to a ship, a high-power laser must supplement current capabilities or provide new capabilities that clearly justify its addition. Second, because a laser provides such a diverse set of capabilities, conventional air-to-air warfare (AAW) models—such as the Fleet AAW Model for Comparison of Tactical Systems (FACTS), Antiair Warfare Simulation (AAWSIM), and Extended Air Defense Simulation (EADSIM), as well as other existing AAW analysis approaches—are not well suited for showcasing current or near-term laser-weapon capabilities. While they can (and have) been used for laser-weapon analysis, their application to a megawatt-class laser that could "instantly" destroy boats or cruise missiles (akin to missile engagements) is a more straightforward application of the existing models and techniques.

In November 1995, the Chief of Naval Operations requested that the National Research Council initiate, through its Naval Studies Board, a thorough examination of the impact of advancing technology on the form and capability of the naval forces to the year 2035. A major observation of the report is quoted below:

> Numerous laboratory and field-test versions of laser weapons have been developed and demonstrated. They have worked as expected and demonstrated suitable lethality against their intended targets. The primary factors that have inhibited the transition of the technology into deployed systems are size and weight. Generally, the conceptual designs of laser weapons that are scaled for combat effectiveness are too large to be appealing to users; conversely, weapons that are sized for platform convenience generally lack convincing lethality.[3]

Subsequently, an August 2006 U.S. Air Force (USAF) Scientific Advisory Board Study examined the increasing threat posed by UAVs in some detail. Key conclusions included:

> No single system can completely address the UAV threat. A single sensor solution is inadequate because of the size and speed challenges presented by small UAVs. A single-weapon-layer solution fails to provide for adaptability to multiple scenarios or adequate probability of kill.

Key recommendations of the USAF Advisory Scientific Board Study included:

> Develop and field longer-term upgrades to counter increased UAV threats. They include:…a small, multimission air/air and air/ground weapon; and directed-energy air defense weaponry.[4]

In addition to the USAF Scientific Advisory Board study, a 2007 OPNAV Deep Blue Study noted the potential advantage of nonkinetic defeat options and recommended that the Navy accelerate development of nonkinetic systems to include high-energy lasers (HELs).[5]

The laser power levels likely to be available in the near term, within reasonable size and cost, are in the neighborhood of 100 kW of radiated power. While this power level is not adequate to engage certain threats, such as cruise missiles or tactical ballistic missiles at tactically useful ranges, there is still a wide spectrum of threats that could be engaged at ranges that are comparable to many current ship-defense weapons, including minor-caliber guns and small missiles. The spectrum of threats includes:

- UAVs
- Missile Seekers
- Intelligence, Surveillance, and Reconnaissance Systems
- Rockets
- Man-Portable Air-Defense Systems (MANPADS)
- Mortar Rounds
- Floating Mines
- Artillery Rounds

LaWS on CIWS

The Mk 15 Phalanx CIWS can often detect, track, and (sometimes) identify potential threats at ranges well outside the effective range of the 20mm gun. These functions are accomplished using the search/track radar system and the Phalanx Thermal Imager (PTI). When added to the Phalanx mount and pointed in the same direction as the gun (see Figure 1), a laser weapon could potentially add a number of useful functions and capabilities to the mount, but technical challenges must be overcome. Preliminary analyses of the mechanical characteristics of the mount suggest that the additional weight that could be added to the mount must be kept under approximately 1200–1500 lb. Additionally, it is highly desirable that the addition of the laser weapon not substantially affect the train/elevation operation of the mount in angle, peak velocity, or acceleration. Consequently, use of rapidly evolving fiber laser technology appears to be the only currently foreseeable path to adding significant laser energy directly to the mount within these constraints.

One major driver in the genesis of the LaWS system was the availability of relatively low-cost COTS fiber-optic lasers. Because these fibers are flexible, they obviate the need for an expensive coudé path system (an optical mirror/lens assembly that turns radiation 90° and may also support rotation of the beam director), thus allowing the use of low-cost mount technology, as well as the retrofitting of the system on existing mounts. The last factor is extremely important because of the scarcity of topside real estate on today's ships. These fiber-optic lasers do have limitations in terms of power, although power levels are growing with advancing technology. The reality today is that, in order to get adequate lethality from a system based on this technology, the use of a beam-combining apparatus utilizing several individual fibers is necessary. (Figure 2 depicts combining multiple fibers in the same beam director.) Furthermore, a smaller beam size is desirable since this drives power density up—increasing the performance required for the tracking and pointing elements of the system. Thus, a high-resolution fine track sensor is needed, as well as an appropriately robust line-of-sight control.

A Potential Suite of LaWS-Related Capabilities

Potential added capabilities that an adjunct LaWS could contribute to the total ship combat system are briefly outlined in the following subsections.

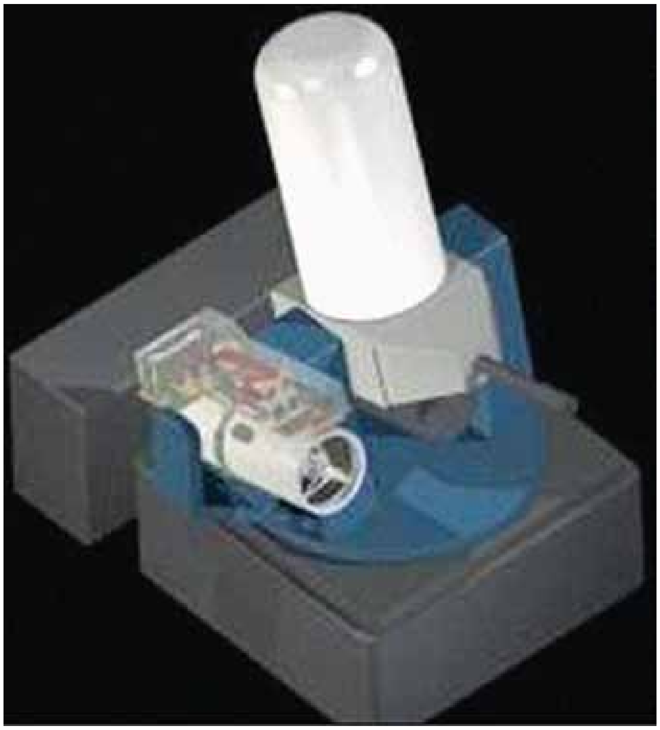

Figure 1. LaWS Mounted on CIWS

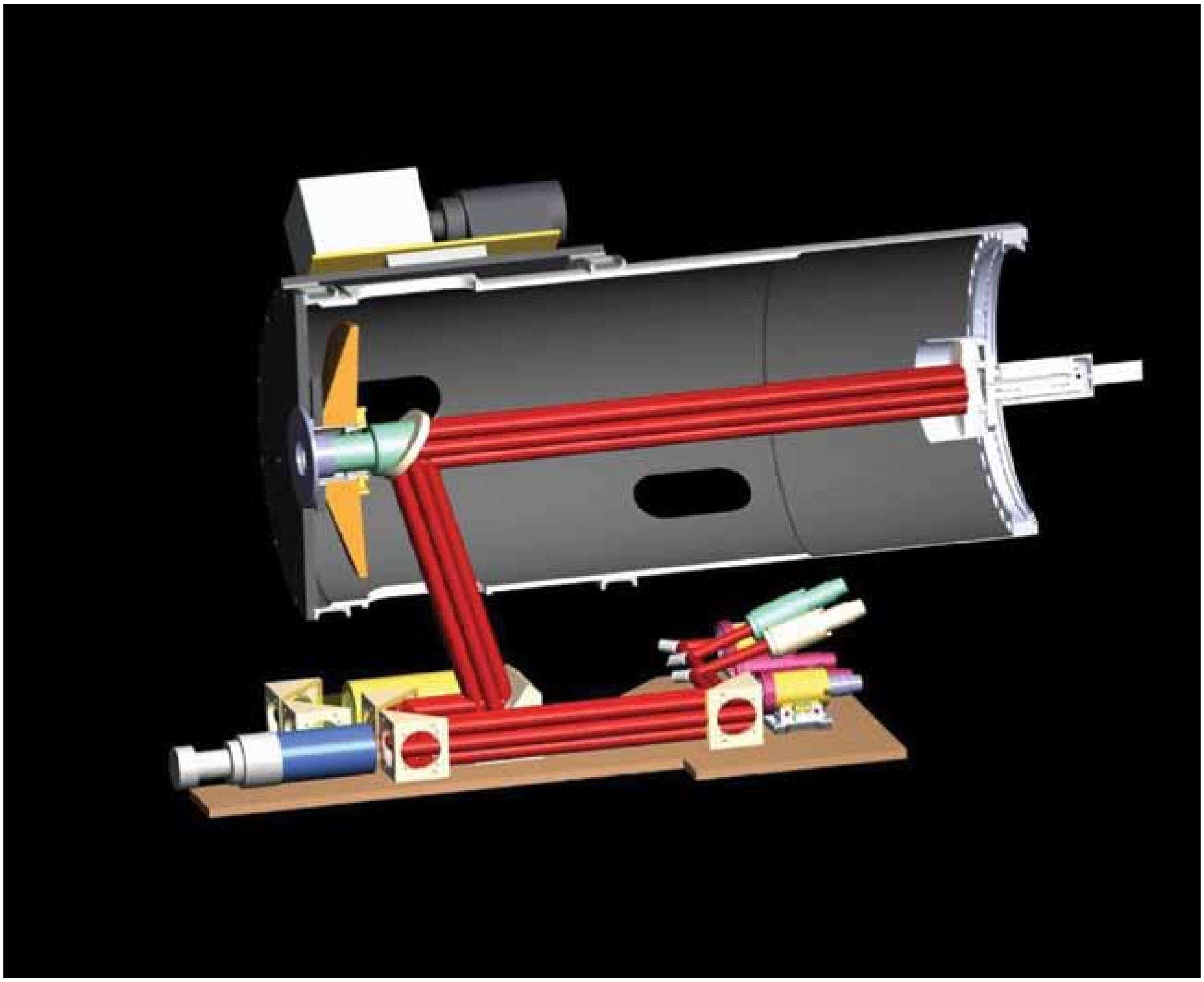

Figure 2. Cutaway View of the LaWS Beam Director

Target Identification, Tracking, and Intent Determination at Range

The optics that would be added for the laser to detect and track targets in support of a laser engagement would immediately contribute additional capabilities to the entire ship combat system even without operating the laser. A laser-gated illuminator, part of the tracking system, significantly increases the signal to the background level of tracked targets and provides good range resolution as well. The additional sensitivity and angle resolution provided by the LaWS optics would allow the identification, precision tracking, and "monitoring" (at high resolution) of potential threats or vehicles of interest at substantially greater ranges than could be achieved by the PTI alone. The Phalanx radar, or another source, would have to provide an initial, accurate cue to facilitate initial acquisition. Once acquired, the target could be examined and monitored with high resolution at range. This capability could make a substantial contribution to identification efforts—efforts to determine intent and potentially even to documenting target behavior to resolve issues with rules-of-engagement doctrine. It is widely recognized that rules-of-engagement issues, such as threat identification and intent determination, are among the most difficult problems faced by ship commanding officers.

Unambiguous Warning at Range

If a fraction of the laser energy is routed through a frequency-doubling crystal, an intense, visible beam can be projected to significant ranges to provide a clear, unambiguous warning that a potential target is about to be engaged unless an immediate change in behavior is observed. This feature also would have utility for dazzling aircraft, surface vehicle, or submarine sensors, and would provide exceptional long-range, unambiguous warning to boats or aircraft at night.

Sensor Destruction at Range

Many electro-optical (EO) sensors are quite susceptible to damage by laser energy in the fiber-laser band as is the case with infrared (IR) missile seekers with germanium optics. The frequency-doubling feature described in the previous paragraph also would be useful to ensure that a band-pass filter at a single frequency could not be applied as an effective countermeasure. The intent here would be to destroy the seeker or imager at ranges well beyond those achievable by the Phalanx 20mm gun. Other examples include intelligence, surveillance, reconnaissance, and targeting sensors on UAVs or unmanned surface vehicles (USVs).

IR Missile Assist at Range

Many targets of interest—including UAVs, USVs, and small boats—are somewhat "marginal" from a target-signature standpoint, particularly at the maximum range of existing IR guided missiles such as the FIM-92 Stinger, the FGM-148 Javelin, the RIM-116 RAM, and the AIM-9X Sidewinder. The CIWS laser adjunct could potentially "correct" this situation by laser heating target vehicles to enhance their signature to existing IR guided missiles. Note that this is NOT "conventional" semiactive-laser (SAL) guidance—the LaWS is not a coded illuminator, nor do the seekers in question rely on this coding. The IR missiles would be unmodified weapons taken from inventory. The LaWS adjunct would simply contribute laser energy that heats the target and enhances its signature for the missile. While, at the ranges envisioned, this laser heating alone would not be sufficient to "kill" the target, it could definitely heat the target. It should also be noted that the laser "illumination" could potentially be used to preferentially select a specific target from among a group of targets for engagement by a missile. It is expected that these engagements could occur at ranges of two to four times the effective Phalanx gun engagement ranges. Use of LaWS in this manner would be exactly analogous to the use of a SAL designator for a SAL guided missile, such as the AGM-114 Hellfire. It is expected that similar rules of engagement would apply.

Direct Target Destruction by Laser Heating

Some threats are known to be vulnerable to direct destruction by the application of laser energy for an appropriate period of time. The currently envisioned system would be able to destroy a subset of naval threats at ranges comparable to, and in some cases greater than, the ranges achieved with modern, stabilized guns using EO fire control systems and modern ammunition. In the case of a LaWS adjunct, the addition of the laser would open new options for a firing/engagement doctrine and would be expected to conserve CIWS rounds for use on threats that are not appropriate for this laser power level. While the laser is often quoted as having an "unlimited magazine," the true number of threats that can be engaged by the laser in any period of time is limited by the required illumination time and by the time required to evaluate a kill and transition to the next target. Thus, for particular target velocities and numbers, the "effective laser magazine" might be added to the CIWS magazine to increase the total number of targets engaged by the combined system.

LaWS Accomplishments

A government/industry team, led by government technical personnel, have achieved significant accomplishments since the start of the LaWS program in 2007; specifically, the team:

- Conducted mission analyses
- Developed threat lethality estimates
- Performed industry surveys for critical components and subsystems
- Performed extensive trade-off analyses
- Designed a prototype system
- Constructed the system—the prototype director and mount (see Figure 3)
- Performed numerous laboratory-based tests of subsystems and the complete prototype
- Validated system operation with a full-up field test at high power using BQM-147A UAV target drones

Additionally, the team was able to minimize the cost of the prototype by leveraging hardware that had already been developed or procured for other applications, including an L3-Brashear tracking mount, a 50-cm telescope, and high-performance IR sensors. Some components were commercially procured, such as the 5.4-kW fiber lasers. Figure 4 shows three laser cabinets, containing two lasers apiece, resulting in a total power output of 32.4 kW. Other components, such as the beam combiner and much of the system software required for operation and target tracking, had to be specifically designed, fabricated, and tested.

The LaWS program achieved a highly successful field test/demonstration in June 2009 when the prototype successfully engaged and destroyed five drone targets at tactically significant ranges at the China Lake, California, test range (see Figure 5).

Additional Work to be Done

Since the LaWS prototype sits on a dedicated gimbal, much additional work needs to be done

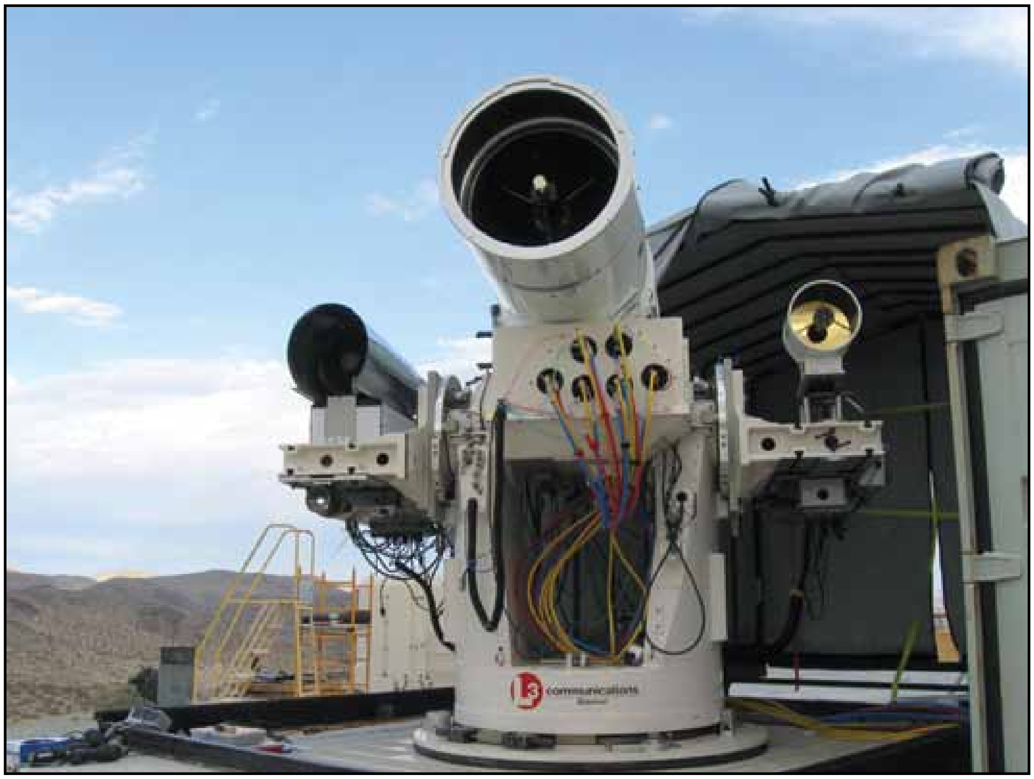
L3 communications

Figure 3. Photo of LaWS During Testing at the Naval Weapons Center, China Lake

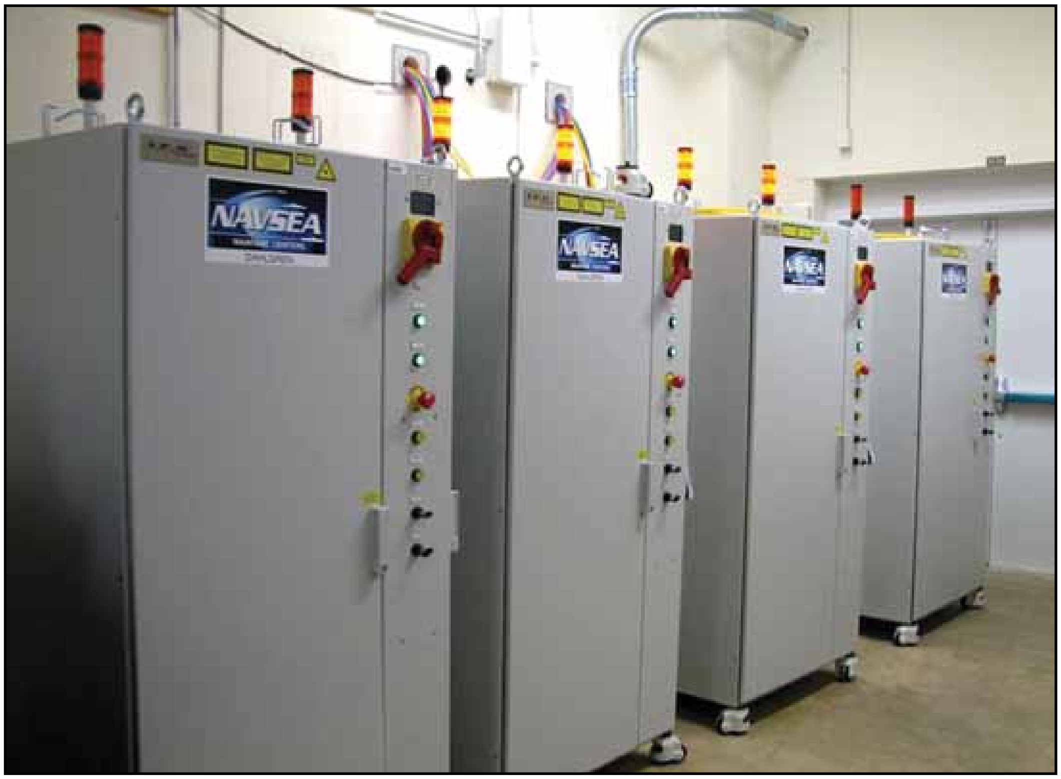

Figure 4. IPG Laser Cabinets

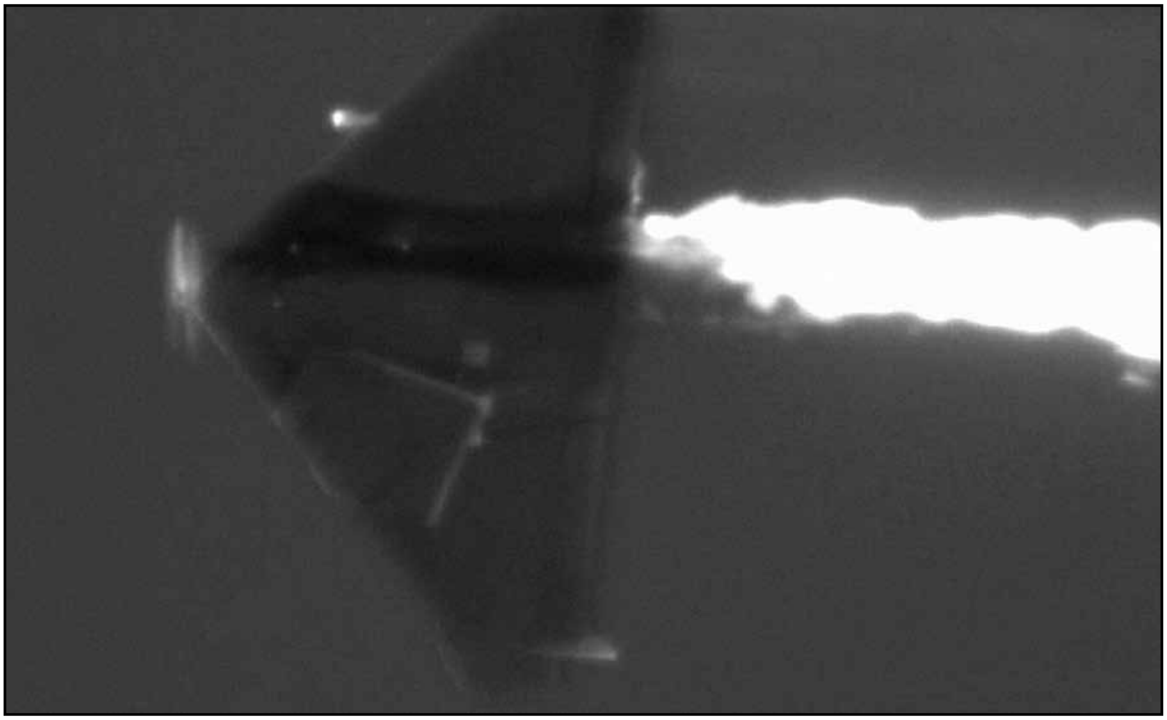

Figure 5. BQM-147A During LaWS Engagement

to place the weapon on the CIWS mount. The latter would require new control systems and optomechanical hardware for line-of-sight stabilization. Other aspects of the shipboard environment are also more stressful, and future mission areas may require an increasingly robust capability to deal with optical turbulence and the high-clutter environment of the ocean surface. Additional laser power might also be required. These modifications, depending on the level of capability desired, will require engineering modifications to the system. Engineering analysis and design to address these issues is currently underway at NSWCDD.

While the aforementioned engineering issues are important to address, there are additional technical issues that have yet to be analyzed. These issues are concerned with the potential utility of the system. Indeed, most of the detailed technical analyses and experiments performed thus far have focused on target destruction, with some effort expended on the issue of seeker damage/destruction. Developing credible lethality estimates for various potential threat targets is clearly very important, but one consequence of the lethality focus is that necessary, detailed, defendable technical analysis, analytic model development, and experiments have not been performed to explore the other functions/features that a CIWS Adjunct LaWS might provide to the overall ship combat system. Some of these contributions might become "routine" if the LaWS were available.

For example, a hard-kill engagement of a target by a Navy shipboard weapon is a relatively rare event, even during wartime conditions. On the other hand, ships in combat zones—and elsewhere—constantly have the problem of detecting potential threats, tracking them, identifying them, determining their intent, and providing warning. Thus, use of the LaWS system, at less than its full lethal potential, could become a daily, standard practice. It is still not clear how these potential benefits and capabilities could be measured or quantified to the satisfaction of key decision makers.

Likewise, other potential advantages of laser weapons—such as the potential for precision engagement, covert engagement, fire starting, graduated lethality, low cost per shot, and "unlimited" magazine—have not been subjected to rigorous technical analysis for feasibility, utility, and practicality. These investigations need to be performed and are gradually being addressed within the LaWS program.

Although the Phalanx CIWS system is currently installed on a number of Navy surface warships—either a single mount or a double mount—there are still significant numbers of ships that do not have a Phalanx system. It is highly desirable to make LaWS potentially available to virtually any ship that could benefit from the enhanced capabilities.

While the technical issues associated with the addition of LaWS to the Phalanx CIWS will be somewhat different from those associated with adding a LaWS system to other weapon systems—or the provision of a "stand-alone" LaWS—they do not appear to be insurmountable. For example, a LaWS beam director might be added to the stabilized Mk 38 Mod 2 25mm gun or the Mk 46 Mod 2 30mm gun. A LaWS beam director might be added to (or even substituted for) the Mk 46 EO Sight on DDGs or added to the trainable RAM launcher. Other options may exist as well.

The issue of defending combat logistics force ships, joint sealift ships, and certain support vessels from attacks from small boats or UAVs is also relevant. These ships often have little or no installed defensive capabilities for potential terrorist or pirate threats, and expeditionary security detachments do not have decisive warning or engagement capability. In addition, there are severe limitations placed on concept of operations (CONOPS) and rules of engagement due to the limited objectives/limited means of the various missions.

A system such as LaWS could provide graduated lethality from warning to destruction. It also could provide additional applications to minimize risk to sea base platforms and enhance sea shield capabilities against nonstate threats. If acceptable rules of engagement can be established, the advantages of graduated lethality might be extended to ships in port or entering/exiting harbors.

While considerable additional work needs to be done to produce a tactical system, the LaWS program's recent demonstration of capability provides strong evidence that a useful, tactical system could be produced within reasonable cost, volume, weight, and power constraints to provide the warfighter with a suite of additional tools to fight today's and tomorrow's wars.

References

1. NAVSEA Notice 5400, Ser 09B/240, "Establishment of the Navy Directed Energy Weapons Program Office (PMS 405)," 4 January 2002.
2. NAVSEA Instruction 5400.101, Ser SEA 06/058, "Directed Energy and Electric Weapon Systems Program Office (PMS 405) Charter," 21 July 2004.
3. *Technology for the United States Navy and Marine Corps 2000–2035: Becoming a 21st-Century Force*, Naval Studies Board, National Research Council, National Academy Press, 1997, Volume 5: Weapons, Chapter 6, "Navy and Marine Corps Applications for Laser Weapons."
4. *Report on the Air Defense Against Unmanned Vehicles*, USAF Scientific Advisory Board, SAB-TR-06-01, August 2006.
5. "Counter Unmanned Aerial Systems (UAS)," *OPNAV Deep Blue Report 08-01*, August 2007.

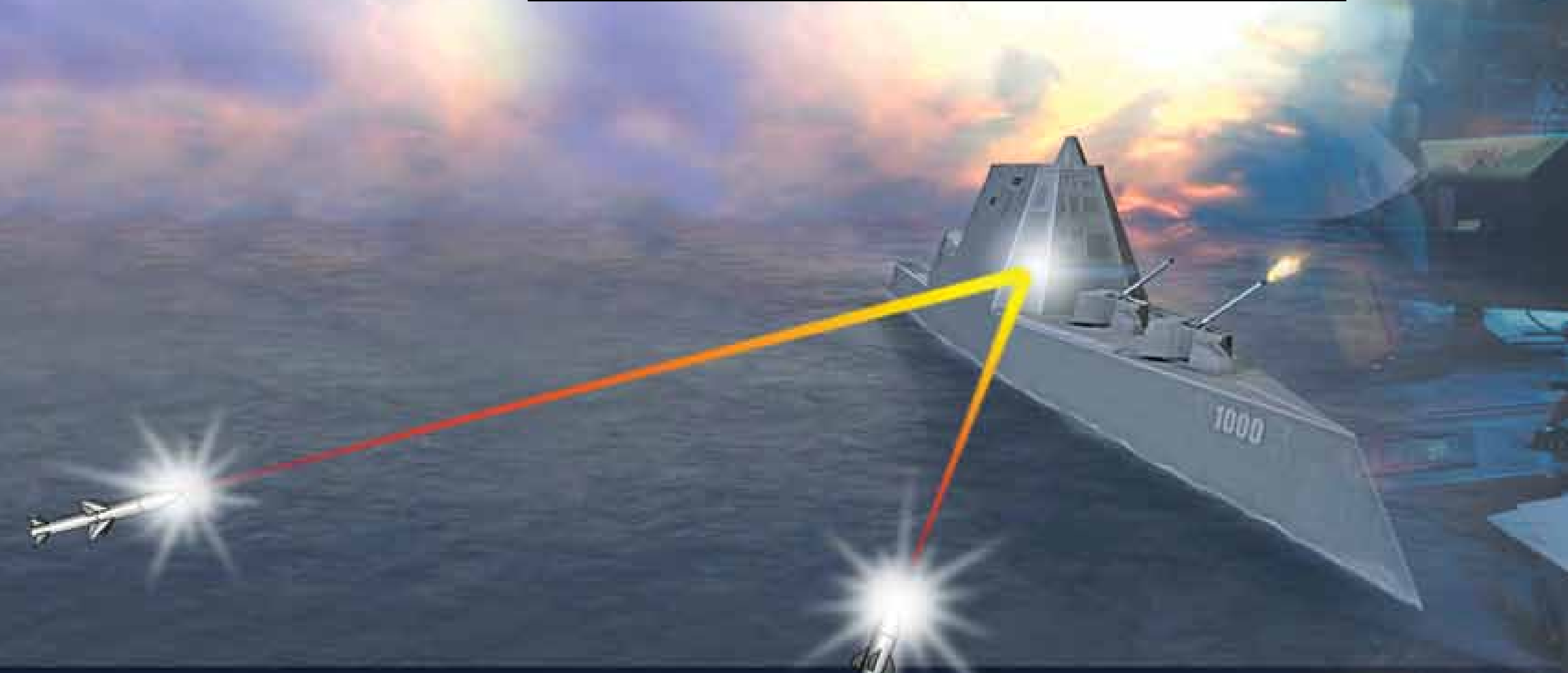

The Acquisition Challenge Associated With Directed-Energy RDT&E

By Mike Kotzian

An already tense situation quickly escalated. Everyone within the combat information center of the Navy's newest all-electric ship suddenly realized that two surface-skimming, antiship missiles were bearing down on their destroyer. With less than 30 seconds to impact, the tactical warfare officer gave the order to fire. Seconds later, the first surface-skimming missile vanished from all tracking consoles. Another order to fire closely followed, and the second missile threat was also destroyed. Consequently, within a matter of 10 seconds from threat recognition to threat elimination, the Navy's newest all-electric ship was able to destroy two incoming threats by using one of the Navy's newest weapon systems—the free-electron laser.

Does this scenario of a Navy all-electric ship, employing a high-energy laser to shoot down enemy surface-skimming antiship missiles, sound like inevitable reality or unattainable science fiction? For scientists and engineers working on directed-energy systems for the Navy, the answer does not lie solely in the advanced technical challenges associated with developing directed-energy weapons. Rather, the answer also lies in how well scientists and engineers understand and adhere to the Department of Defense's (DoD's) Defense Acquisition Management System (DAMS) framework governing the development of new weapon systems.

Evolution of Defense Acquisition

The way in which DoD identifies needs and subsequently develops, tests, procures, and sustains weapon systems has evolved over time. Today's acquisition foundation can be traced back to the Packard Commission report in 1986, where many of this report's recommendations became part of the Goldwater-Nichols DoD Reorganization Act of 1986. This evolution continued along three tracks:

1. Requirements moving from threat-based to capability-based
2. The resource allocation system adding execution reviews with concurrent program and budget reviews

3. The acquisition process attempting to incorporate a more flexible and tailored process

These three tracks form the Defense Support System organizational structure: the Joint Capabilities Integration and Development System (JCIDS) process; the Planning, Programming, Budgeting, and Execution (PPBE) process; and the DAMS process, respectively. These three processes operate as "systems of systems" and are referred to as the "Big A" acquisition process shown in Figure 1.[1]

While all three of these phases hold their own level of importance, the major focus for scientists and engineers at research and development (R&D) facilities is the "Little a" acquisition process. It is this "Little a" acquisition process, where the rules and processes are found, that governs how DoD goes about developing a new materiel solution to a validated warfighter requirement. These rules and processes are codified within DoD Instruction 5000.02, *Operation of the Defense Acquisition System*, which was issued in December 2008.

The acquisition framework associated with DoD Instruction 5000.02 is the DAMS structure. This framework, shown in Figure 2, consists of numerous strategically placed milestones and major program reviews to ensure proper programmatic oversight.[2] Each of the milestones has specific

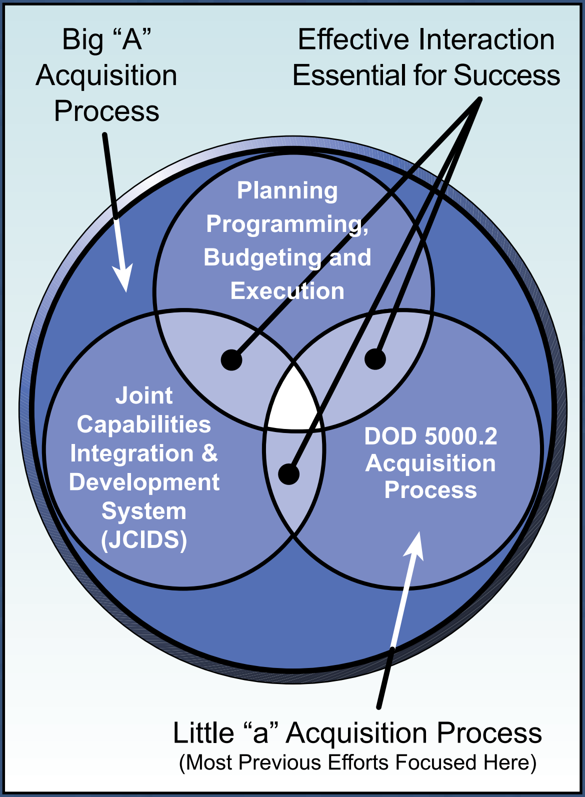
Big "A" Acquisition Process
Effective Interaction Essential for Success
Planning Programming, Budgeting and Execution
Joint Capabilities Integration & Development System (JCIDS)
DOD 5000.2 Acquisition Process
Little "a" Acquisition Process
(Most Previous Efforts Focused Here)

Figure 1. Defense Support System Organizational Structure

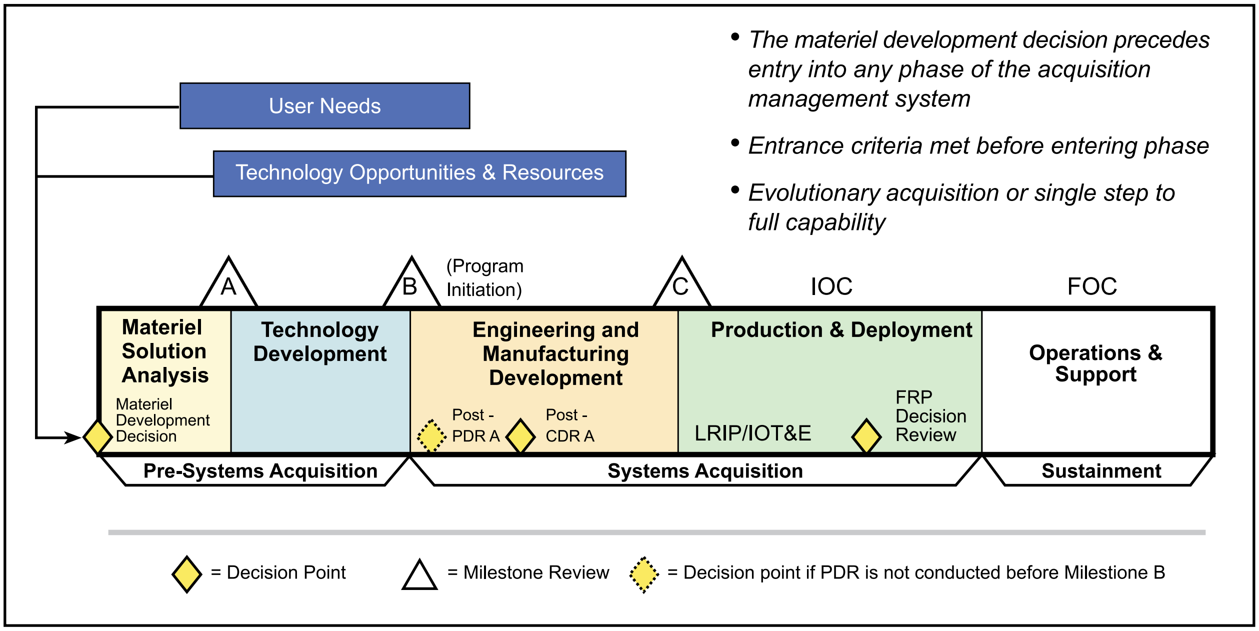
• The materiel development decision precedes entry into any phase of the acquisition management system
• Entrance criteria met before entering phase
• Evolutionary acquisition or single step to full capability
User Needs
Technology Opportunities & Resources
A
B
(Program Initiation)
C
IOC
FOC
Materiel Solution Analysis
Materiel Development Decision
Technology Development
Engineering and Manufacturing Development
Post - PDR A
Post - CDR A
Production & Deployment
LRIP/IOT&E
FRP Decision Review
Operations & Support
Pre-Systems Acquisition
Systems Acquisition
Sustainment
= Decision Point
= Milestone Review
= Decision point if PDR is not conducted before Milestone B

Figure 2. DoD Acquisition Framework

criteria that must be satisfied before a program is allowed to further proceed along the DAMS. The program's Milestone Decision Authority (MDA) rests with the individual responsible for deciding if the milestone criteria have been met and, if so, for allowing the program to proceed to the next phase of the acquisition process. Designation of a program's MDA depends on a program's level of research, development, test, and evaluation (RDT&E) and procurement funding. For example, an Acquisition Category (ACAT) I program is defined as an eventual total expenditure for RDT&E of more than $365 million in fiscal year (FY) 2000 constant dollars or, for procurement, of more than $2.19 billion in FY 2000 constant dollars. In this case, for an ACAT ID ("D" refers to the Defense Acquisition Board (DAB)) the Under Secretary of Defense for Acquisition, Technology and Logistics (USD(AT&L)) is the MDA; for an ACAT IC ("C" refers to Component or Service), the MDA is the Head of the DoD Component or, if delegated, the Component Acquisition Executive.[3]

In addition, civilian and military workforce members within the DoD whose job responsibilities are deemed acquisition-related find themselves with a training requirement necessary to carry out their acquisition-related job responsibilities. Specifically, these workforce members are required to gain acquisition training and education with the passage of the Defense Acquisition Workforce Improvement Act (DAWIA) signed into law in 1990. The current certification process comprises three levels covering 16 different career fields. Each of these 16 career fields has a set of specific training, education, and experience requirements that must be met in order for an individual to achieve Level 1, Level 2, or Level 3 certification. The Defense Acquisition University (DAU) provides the necessary training classes required for the certification. DAU identifies "core-plus" training classes and continuous learning modules for each level of certification. The core-plus classes and modules are not required for certification but are identified as additional sources of information to assist individuals in becoming more knowledgeable about their career field beyond the minimum standards required for certification. The most up-to-date certification frameworks for all 16 career fields can be found at the following DAU website: http://icatalog.dau.mil/onlinecatalog/CareerLvl.aspx

Defense Acquisition Reform

The DoD acquisition environment is undergoing continuous change. The issuance of DoD Instruction 5000.02 marked the opening salvo of what has become seemingly constant updates, modifications, and guidance impacting how DoD procures weapon systems to meet warfighter requirements. In addition to DoD's issuance of DoD Instruction 5000.02, the Government Accountability Office published a stream of reports and findings that indicate significant cost growth and schedule delays in major defense acquisition programs. In 2009, Secretary of Defense Robert M. Gates proclaimed

a new way of doing business within DoD when it comes to weapon systems acquisition. Pressures are building for every program to maintain cost and schedule estimates while delivering the technical requirements originally developed to support the warfighter.

Moreover, there have been two major policy issuances. As previously mentioned, the first was DoD Instruction 5000.02 in December 2008. This update of the rules and processes governing DoD weapon systems acquisition primarily impacted the early part of the DAMS framework. The problem was that weapon system programs were failing their initial operational test and evaluation phases at alarming rates—many times traced to program offices attempting to design weapon systems with immature technology. Such failures were preventing those programs from proceeding to a full-rate production decision review and, more importantly, causing a repeat of some of the DAMS framework, which translated to increased costs and delayed initial operational capability timelines.

DoD Instruction 5000.02 attempted to solve this problem with three main emphases. First, a mandatory requirement was inserted for competitive prototyping prior to program initiation at Milestone B. The intent was to ensure a competition among contractors competing for a contract award. The theory was that such a competition would reduce technical risk, validate designs, improve cost estimates, evaluate manufacturing processes, and refine requirements. Reducing technical risks was especially important because weapon system programs were expected to demonstrate a technology readiness level (TRL) of six—where the system/subsystem model or prototype is demonstrated in a relevant environment—by the time a program reached Milestone B. TRLs are categorized on a scale of 1 to 9. A TRL of 1 is the lowest level of technology readiness, where scientific research begins to be translated into applied R&D. A TRL of 9 is the highest level of technology readiness, where the actual system is proven through successful mission operations. A TRL of 6 represents a major step up in a technology's demonstrated readiness. Using TRLs enables consistent comparisons of technical maturity across different types of technologies, giving program decision makers a common benchmark to consider when assessing program risk. Note that TRLs are meant to capture a level of technical maturity, not the probability of occurrence (i.e., the likelihood of attaining a required maturity level) or the impact of not achieving a level of technical maturity.[4]

The second emphasis was on a stricter adherence to systems engineering processes and technical reviews. Too often weapon system programs were not closely following systems engineering processes or avoiding due diligence when it came to the definition of successful exit criteria for a technical review. Consequently, all technical efforts must be outlined in a program's systems engineering plan. The program manager will use the eight technical management processes—decision analysis, technical planning, technical assessment, requirements management, risk management, configuration management, technical data management, and interface management—to manage the technical development of the system increments, including the supporting or enabling systems.[5] The program manager will use the eight technical processes—stakeholders requirements definition, requirements analysis, architectural design, implementation, integration, verification, validation, and transition—to design the system, subsystems, and components, including the supporting or enabling systems required to produce, support, operate, or dispose of a system.[6] Figure 3 provides an overlay of the new DoD Instruction 5000.02 and Secretary of the Navy (SECNAV Instruction) 5000.2D (Implementation and Operation of the Defense Acquisition System and the JCIDS), and shows the timing of specific systems engineering technical reviews as a program matures through the DAMS.

The third emphasis was a more prominent role of the MDA, starting with a mandatory requirement that all weapon system programs seeking a full or partial materiel solution must hold a Materiel Development Decision chaired by the MDA. Thus, the old Design Readiness Review was replaced with the Post-Critical Design Review Assessment chaired by the MDA. In short, the MDA was to become a more prominent figure in the oversight of a weapon system program's progress.

The second relatively recent major policy issuance was the Weapon Systems Acquisition Reform Act (WSARA) of 2009, implemented by Directive-Type Memorandum (DTM) 09-027 in December 2009. This DTM amended DoD Instruction 5000.02, the *Defense Federal Acquisition Regulation Supplement (DFARS)*, and associated business practices within the *Defense Acquisition Guidebook (DAG)*. The WSARA implementation brought about changes to policies and procedures across 13 categories. Some of the WSARA changes most relevant to the Navy directed-energy community include:

- Analysis of alternatives study guidance
- Acquisition strategies to ensure competition
- Competitive prototyping
- Developmental test and evaluation
- Systems engineering
- Preliminary design reviews
- Critical cost growth

THE ACQUISITION IMPACT

So why should the directed-energy community care about these acquisition policy changes? Because these policy changes impact the community's

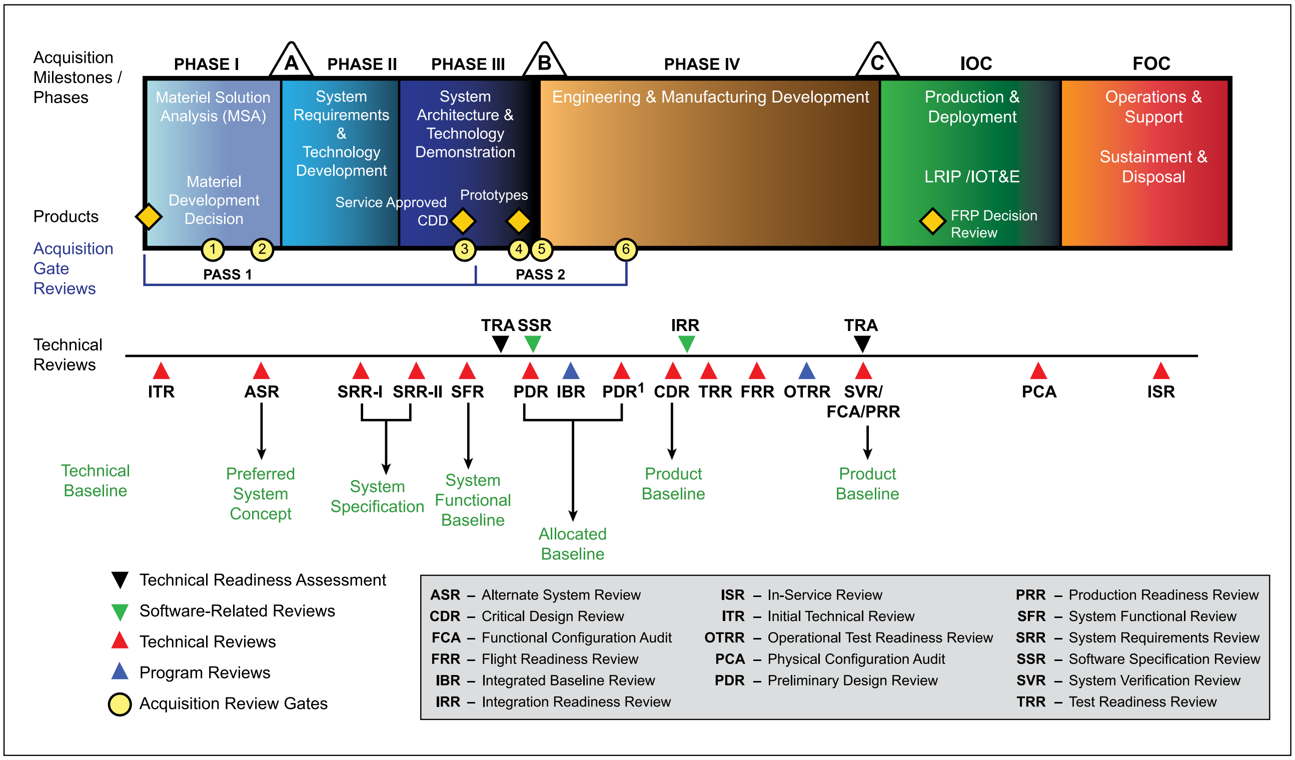
Acquisition Milestones / Phases
PHASE I
A
PHASE II
PHASE III
B
PHASE IV
C
IOC
FOC
Materiel Solution Analysis (MSA)
System Requirements & Technology Development
System Architecture & Technology Demonstration
Engineering & Manufacturing Development
Production & Deployment
Operations & Support
Materiel Development Decision
Service Approved CDD
Prototypes
LRIP /IOT&E
FRP Decision Review
Sustainment & Disposal
Products
Acquisition Gate Reviews
1
2
3
4
5
6
PASS 1
PASS 2
Technical Reviews
TRA
SSR
IRR
TRA
ITR
ASR
SRR-I
SRR-II
SFR
PDR
IBR
PDR1
CDR
TRR
FRR
OTRR
SVR/ FCA/PRR
PCA
ISR
Technical Baseline
Preferred System Concept
System Specification
System Functional Baseline
Allocated Baseline
Product Baseline
Product Baseline
▼ Technical Readiness Assessment
▼ Software-Related Reviews
▲ Technical Reviews
▲ Program Reviews
○ Acquisition Review Gates
ASR – Alternate System Review
CDR – Critical Design Review
FCA – Functional Configuration Audit
FRR – Flight Readiness Review
IBR – Integrated Baseline Review
IRR – Integration Readiness Review
ISR – In-Service Review
ITR – Initial Technical Review
OTRR – Operational Test Readiness Review
PCA – Physical Configuration Audit
PDR – Preliminary Design Review
PRR – Production Readiness Review
SFR – System Functional Review
SRR – System Requirements Review
SSR – Software Specification Review
SVR – System Verification Review
TRR – Test Readiness Review

Figure 3. Systems Engineering Technical Review Timing

ability to develop, produce, and/or sustain directed-energy weapon systems. The ultimate goal of the directed-energy community is to deploy directed-energy weapons to the fleet. Accordingly, regardless of which phase or phases an organization in the community supports, its actions are impacted by the language in DoD Instruction 5000.02 and the WSARA of 2009. The more scientists and engineers in the organization are aware of governing policy documents like DoD Instruction 5000.02, the better their chances are of meeting DoD leadership's expectations in terms of cost, schedule, and technical effectiveness.

Actions have shown that DoD senior leadership has come to expect all weapon system programs to adhere to the current acquisition-related policy and guidance changes. As mentioned earlier, major weapon system programs have recently been canceled or restructured for not meeting DoD senior leadership expectations—something that rarely occurred previously. In today's environment, technology alone will not carry the argument for a program's survivability. Directed-energy weapons definitely carry the allure of a "Star Wars-like" capability, but these same weapon systems will need to show sustainable cost and schedule compliance if they are to come to fruition. Resources are too limited, and the warfighter has too many needs to allow unsustainable weapon system programs to continue. Therefore, everyone involved with the development, procurement, and/or sustainment of a directed-energy weapon system needs to have an adequate understanding of the acquisition underpinnings now governing DoD.

Summary

The proverbial "winds of change" are blowing across the DoD acquisition landscape. The management of major weapon systems dependent upon cutting-edge technologies—such as those of directed energy—cannot afford to conduct business in a manner reminiscent of bygone days. Everyone involved with the development, production, or sustainment of a directed-energy weapon system needs to understand the "rules of engagement" laid down by the most recent DoD acquisition policy guidance. Highly skilled scientists and engineers typically already understand the need for a structured systems engineering approach to problem solving. Today, though, more than ever, cost and schedule must be factored in as potential tradespace to deliver the ultimate goal: a cost-effective, directed-energy weapon system delivered in a timely manner while meeting the warfighter's requirements. Scientists and engineers who adhere to these recent acquisition changes will help their organizations achieve this goal, thereby ensuring that warfighters will be armed with the most technologically superior weapons possible.

References

1. Congressional Research Service Report 7-5700, 10 July 2009, p. 3.
2. Ibid., p. 8.
3. DoD Instruction 5000.02, 8 December 2008, p. 33.
4. *Defense Acquisition Guidebook*, 5 May 2010, paragraph 10.5.2.2.
5. Ibid., paragraph 4.2.3.1.
6. Ibid., paragraph 4.2.3.2.

THE BASICS OF ELECTRIC WEAPONS AND PULSED-POWER TECHNOLOGIES

By Stuart Moran

WHAT ARE ELECTRIC WEAPONS?

Most conventional weapons rely on chemical energy (explosives) as their destruction mechanism, either to explode on target, like bombs, or to create kinetic energy, like a bullet. Electric weapons are different. Electric weapons use stored electrical energy, rather than explosives, to attack or destroy the target. Electric weapons generally fall into two categories: directed-energy weapons (DEWs) and electromagnetic (EM) launchers. DEWs send energy, instead of matter, toward a target, and can be separated into three types: laser weapons, particle-beam weapons, and high-power microwave (HPM) or radio-frequency (RF) weapons. EM launchers use electrical energy to throw a mass at a target, thus making them distinct from directed energy. There are also three types of EM launchers: rail guns, coil guns, and induction drivers. All involve the use of strong magnetic fields to push against projectiles. While electric guns are an electric weapon, they are not a DEW.

High electrical powers and large energies are needed for all these weapons. Technologies for storing and controlling electric power are needed and are commonly called pulsed-power technologies. Electric guns are often associated with DEWs due to their common reliance on pulsed-power technology. The types of electric weapons are shown in Figure 1.

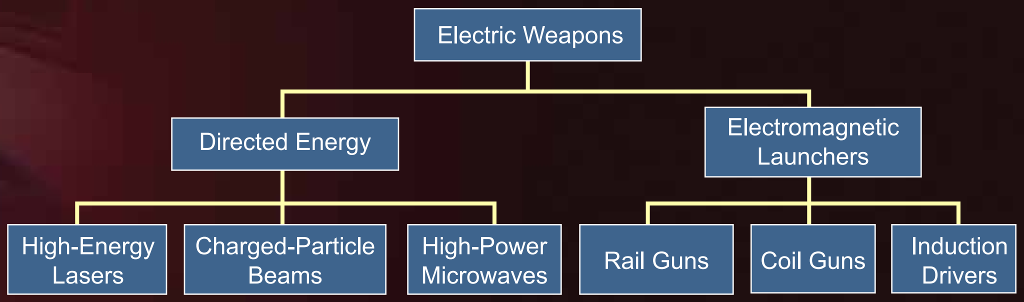
Electric Weapons
Directed Energy
Electromagnetic Launchers
High-Energy Lasers
Charged-Particle Beams
High-Power Microwaves
Rail Guns
Coil Guns
Induction Drivers

Figure 1. Types of Electric Weapons

There are a number of powerful advantages of electric weapons over conventional explosives:

- DEWs have a near-zero time of flight compared to conventional ordnance, allowing longer decision times and quicker reaction times.
- Electric weapons have a large "magazine" capacity, often limited only by the ability of the power source to recharge the system. The firing rate depends on how fast the system can be recharged, which in turn, depends on the available power source.
- The cost of engagement is greatly reduced. With increasingly sophisticated conventional weapons, the cost of practice rounds, such as a missile, can be millions. For an electric weapon, the cost per engagement is greatly reduced, making the attack of small targets (the asymmetric threat) less costly and training much more affordable.
- There is the potential for variable lethality, where the weapon effects can be controlled or attenuated to provide a warning or nonlethal effect. Otherwise, a full-power setting can be used to destroy the target.
- Electric weapons have the benefit of increased safety since less ordnance needs to be stored. Logistics costs less, and underway replenishment is easier since explosives are reduced or eliminated.
- Electric weapons can be used in conjunction with conventional weapons to heighten overall combat system effectiveness, such as knocking out electronics before engaging with a kinetic weapon.

Historically, the key Navy scenario for using directed-energy technologies has been close-in protection of naval vessels from antiship cruise missiles, particularly in a littoral environment. The ability of a DEW's speed-of-light engagement is particularly attractive under conditions of short warning times from supersonic stealthy missiles. However, increasingly difficult and problematic threats from nonmilitary aircraft and surface ships, countersurveillance platforms, fast patrol boats, unmanned aerial vehicles (UAVs), and terrorist inflatable boats or jet skis present different challenges. The threat has shifted from small numbers of expensive targets in open water to large numbers of small and cheap targets among neutral forces. The unique characteristics offered by DEWs, when compared to traditional weapon systems, allow them to be applied across a spectrum of threat roles, particularly in friendly or neutral-rich regions where precision pointing or less-than-lethal capability is paramount. The potential for HPM to counter electronics at levels below human effects makes them ideal nonlethal weapons. Electromagnetically launched projectiles allow longer range, shorter flight times, reduced reliance on air strikes and missiles, and safer storage and replenishment. With military budgets being squeezed, the low cost of directed-energy engagements, which often require just a few gallons of fuel, cannot be overemphasized. Instead of million-dollar missile shots, electric weapons allow new tactics, warning shots, and continual fire against large and small targets. They also allow inexpensive practice and training for improved readiness.

Pulsed Power for Electric Weapons

A useful rule of thumb is that a stick of TNT contains about a megajoule (MJ) of chemical energy, and this amount is often needed to destroy a military target. To destroy a target with an electric weapon, the electrical energy must also be deposited quickly. Surprisingly, a candy bar also has a megajoule of chemical energy, but it is released very slowly when we eat it. Many electric weapons require peak powers of more than a gigawatt (GW) or energies more than a megajoule. The time scales for delivery range from milliseconds to nanoseconds. As an example, delivering 1 MJ of energy in 10 μs requires 100 GW of power, which is more than a commercial power plant can produce. It is not practical to build continuous power supplies to directly drive most electric weapons. Consequently, pulsed-power technologies are needed to store energy at low power rates and release it quickly for weapon use. A pulsed-power system takes electrical power from a prime source (like a motor), stores it, and transforms the power to meet specific user requirements. The importance of a pulsed-power system is often underappreciated. For most electric weapon systems, the system size, weight, volume, and reliability are dominated by the pulsed-power chain. Pulsed-power components must be improved along with the weapon technology to make electric weapons systems practical. A block diagram of a pulsed-power system is shown in Figure 2.

Electrical energy can be stored in many ways, such as a battery (actually a chemical storage). A car battery has about a megajoule of energy, but it takes many seconds to drain it. A much faster method of storing electrical energy is in a capacitor, which can be discharged in milliseconds or faster. Inductive methods store the energy in the magnetic fields of a coil. This has the potential of achieving higher energy density than capacitors, but

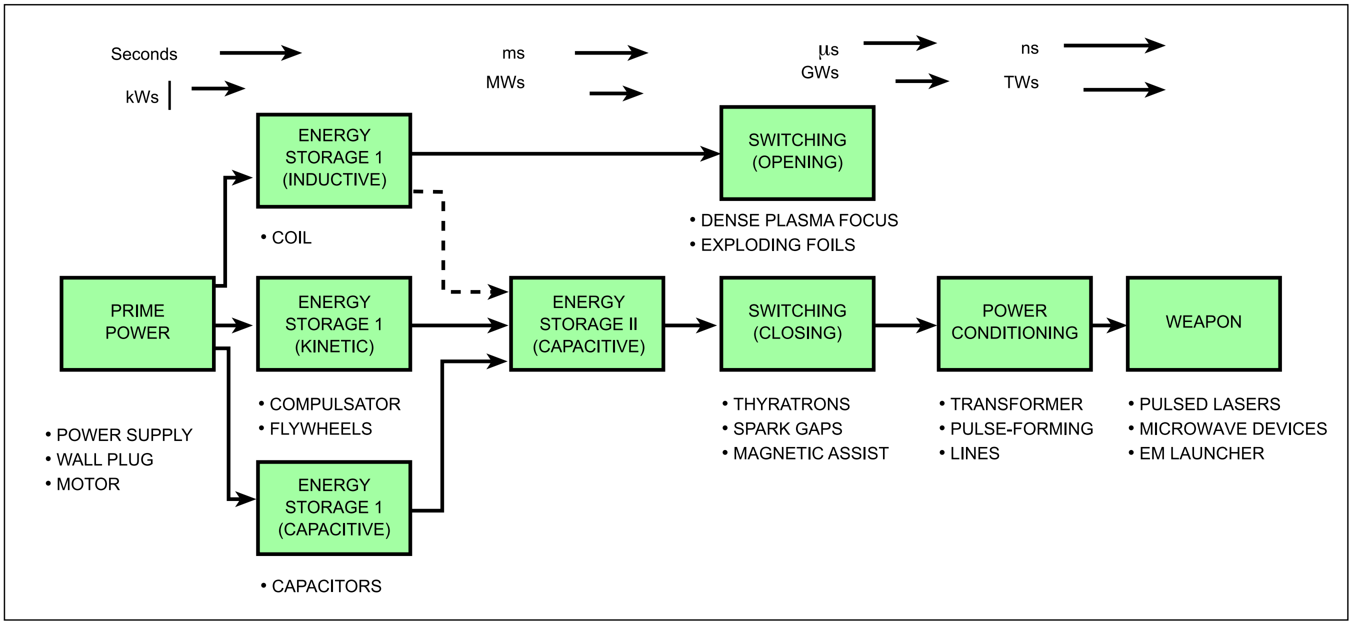
Seconds
kWs
ms
MWs
µs
GWs
ns
TWs
ENERGY STORAGE 1 (INDUCTIVE)
• COIL
SWITCHING (OPENING)
• DENSE PLASMA FOCUS
• EXPLODING FOILS
PRIME POWER
• POWER SUPPLY
• WALL PLUG
• MOTOR
ENERGY STORAGE 1 (KINETIC)
• COMPULSATOR
• FLYWHEELS
ENERGY STORAGE II (CAPACITIVE)
SWITCHING (CLOSING)
• THYRATRONS
• SPARK GAPS
• MAGNETIC ASSIST
POWER CONDITIONING
• TRANSFORMER
• PULSE-FORMING
• LINES
WEAPON
• PULSED LASERS
• MICROWAVE DEVICES
• EM LAUNCHER
ENERGY STORAGE 1 (CAPACITIVE)
• CAPACITORS

Figure 2. Block Diagram of a Pulsed-Power System

when the supporting systems are considered, the technology becomes less attractive. Energy storage for electric weapons can also be done with chemical explosive energy, where an explosive force is converted into electrical energy using techniques such as flux compression. Energy can be stored in the inertia of rotating machines and flywheels, but the energy can be released only as fast as the flywheel or motor can be stopped, usually in seconds. In many cases, several stages of energy store are used where each stage is faster than the last. Once the energy is stored, it must be released quickly using a high-power switch. There are many types of switches. Perhaps the most common type for electric-weapon applications has been the spark gap. Many types of controlled spark gaps exist, including pin-triggered, laser-triggered, field distortion, and simple overvolted. To achieve high repetition rates, flowing oil or gas can be used to flush the hot spark products, or sealed gaps using special fast-recovery gases, such as hydrogen, can be employed. Other switches, such as vacuum tubes and solid-state switches, can be used if they can handle the voltages and currents needed. Solid-state technologies, such as thyristers, have become very capable in recent years. Once the energy is switched out, there is usually some additional power conditioning, where transformers or pulse-forming networks are used to provide the desired pulse shape, voltage, and current required for the weapon. For rapid firing rates or continuous use, high average input powers are needed.

ALL-ELECTRIC SHIP

One of the major impediments to the development of electric weapons systems for Navy ships has been a lack of electrical prime power. Current surface combatant designs employ up to 90 percent of engine power mechanically dedicated solely to propulsion. These designs are unable to provide the tens to hundreds of megawatts (MW) of electrical power capacity required for many electric weapons. The solution is an electric-drive ship that uses all the engine power to generate electricity, enabling it to allocate power to weapons or propulsion as needed. In recent years, the Navy has been investigating cost-effective power-system options to meet future platform requirements.

HIGH-POWER MICROWAVE (HPM) AND RF WEAPONS

Microwave weapons are generally considered to use frequencies above a gigahertz, whereas lower frequencies are generally called RF weapons. These weapons are more powerful than electronic warfare systems and are designed to create extended disruption or permanent damage. An HPM weapon is considered to have a peak power of more than 100 MW, or energies above 1 J. The energy can enter a target through intended RF paths, such as target antennas (front door), or unintended paths, such as housing joints, cavities, and circuit wires (back door). Pulses ranging from a few nanoseconds to microseconds in duration can be sufficient to reset computers, cause loss of stored data, or cause microprocessors to switch operating modes. Nonlinear circuits and components can rectify signals and absorb energy outside of their normal operating parameters. Figure 3 illustrates some of the vulnerability areas on a missile body.

RF or HPM devices can be divided into narrowband or wideband systems, dependent upon the employed pulse length. Narrowband systems are similar to high-power radar pulses and produce RF radiation with a very narrow bandwidth (frequency coverage). The damage concept is to create enough energy in a target to overheat or overload electronic components. Wideband systems generally produce very short pulses (nanoseconds) and typically operate in lower frequency ranges. Wideband systems produce much lower average powers and rely on high-peak electric fields to produce reset or arcing of digital components. Creating short pulses—often only a few RF cycles long—generates a very broad frequency output to take advantage of a target's weak point. But, it also means that the energy is spread over many frequencies, so there may be very little energy at a specific vulnerable frequency. Vulnerability data is critical to estimate the effectiveness of HPM weapons. Ultimately, air breakdown will limit the amount of energy out of an antenna to around 1 MW/cm^2.

HPM devices can produce effects that range from denying the use of electronic-based equipment to disrupting, damaging, or destroying such equipment. HPM weapon advantages include all-weather capability, low precision pointing requirements, and effects persistence after the radiated EM energy "beam" has been turned off. One major advantage of HPM is that electronics are generally more vulnerable to high fields and high energies than humans. This provides the ability to attack electronics without harming people, which makes HPM an ideal choice for nonlethal applications.

Two major challenges of implementing HPM technologies into an operational weapon systems platform are:

1. Fratricide, or self-destruction, can be a problem because of the large areas affected by the sidelobes and near field of any meaningful HPM weapon system. Therefore, when attacking a target of interest with an HPM weapon, there is a greater risk of disruption to systems that were not intended to be targeted but fell within the sphere of influence. Host platforms, therefore, may need to undergo interference hardening.
2. With regard to battle damage assessment, kinetic weapons have the advantage of typically leaving visual evidence. HPM weapon systems do not leave large holes in a target but create more subtle influences as a result of attacking critical electronic components. Consequently, it can be more difficult to ascertain whether a target's capabilities have been sufficiently degraded or destroyed—and for how long—in determining whether a mission was successful.

For HPM system development, a fundamental challenge is the understanding of what it takes to affect the target. Coupling mechanisms, where EM energy enters and affects the target system, are extremely complex. The vulnerability of components is often vastly different if it is outside or inside a circuit board or enclosure. Effects depend upon the interactions with other components, connectors, and nearby conductors. The effects on a component can vary many orders of magnitude depending on frequency, orientation, cracks and seams, protective circuits, pulse energy, and duration. Research regarding effects on missiles has shown large variations not only between designs, but also between different serial numbers due to assembly methods, cable routing, and component variations. With the increasing use of commercial equipment by the military, such as computers and radios, effects are difficult to predict due to constant design and component changes. In general, electronics are getting smaller and operating at lower voltages, making them more sensitive to high fields. But smaller components often have lower pickup areas, and the proliferation of interfering signals has increased the amount of shielding on modern electronics. When

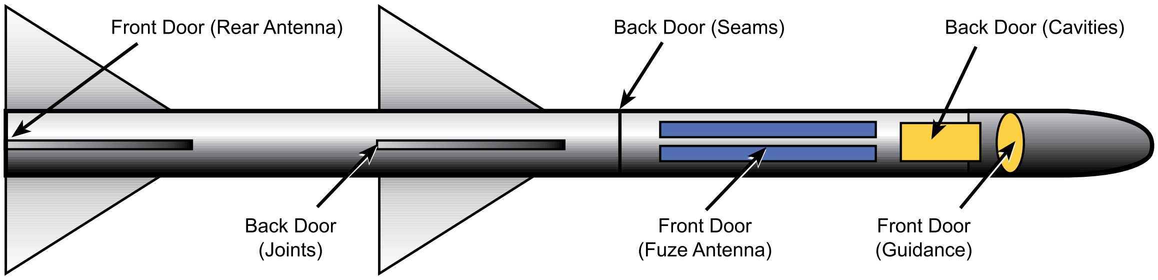

Figure 3. HPM Coupling Paths on Missile Body

target systems are located inside structures or buildings, it becomes even more difficult to predict. Efforts to predict reflections and interference inside complex structures become extremely complicated. Accordingly, generic electronics kill using universal waveforms is not likely. There continues to be a lot of hype about what RF weapons can do, but the idea that a backpack device can wipe out all electronics in a city is no more realistic than a hand-held laser cutting through a bank vault door.

High-Energy Lasers (HELs)

A laser generally produces a beam of coherent light at a specific wavelength dependent on the atomic structure of the lasing substance. Only certain substances have the atomic properties appropriate for producing laser light, and these are often limited in power. Lasers are characterized by the substance being lased (gas, liquid, or solid) and the "pumping" process (light energy, electricity, or chemical reaction). A resonant optical cavity provides the means for aligning the energy in the beam and extracting that energy. A military laser system also includes beam processing or beam-path conditioning, beam pointing and control and—for long-range applications—adaptive optics to compensate for the atmosphere.

Until recently, HELs have been driven by chemical energy, so very little electrical power or pulsed power was needed. Chemical lasers use the reactions of gases or liquids to create the excited energy states necessary for laser emission. Large chemical lasers and beam directors have been developed by the Navy in recent decades and have successfully ruptured fuel tanks and downed supersonic missiles. However, these lasers required high-velocity, chemical-reaction chambers and emitted hazardous gaseous by-products. They often operated at wavelengths where the atmosphere absorbed much of the energy. Absorption creates thermal blooming, whereby absorbed energy in the air creates a negative lens that defocuses the beam. Increasing the power of the laser increases the energy absorbed and worsens the problem. The Army and Air Force are developing chemical lasers for airborne applications, where atmospheric absorption is less of a problem. Recent Navy interest in HELs has concentrated on lasers that are electrically powered, rather than chemically powered, and that operate at shorter wavelengths to allow smaller optics and more efficient propagation near the water.

Small semiconductor (or diode) lasers use current flow through an electrical junction to excite electrons and create laser light. These lasers are very limited in power, so research has focused on using large numbers of lasers assembled into a coherent array. Semiconductor lasers also create efficient light to excite or "pump" other types of lasers. Solid-state lasers (SSLs) use crystalline materials mixed (doped) with elements needed for proper lasing. SSLs show strong promise for compact, medium-power HEL weapon systems. Scaling these systems up to megawatt levels creates extreme heat in the crystal material, making it very difficult to prevent internal damage. Forced cooling and the heat capacity of large masses are under study.

Fiber lasers—which use semiconductor diode lasers to pump a flexible, doped crystalline fiber (similar to a fiber-optic line)—have demonstrated high efficiency and relatively high power. The technology is being used in the welding and cutting industries. Methods of pumping large numbers of fiber-optic lasers and combining them are being investigated. An example is shown in Figure 4.

The free-electron laser (FEL) operates differently from a conventional laser. An FEL uses a high-voltage electron accelerator to push electrons through a magnetic "wiggler" to create light radiation across a tunable band of frequencies. The FEL is extremely complex and large, but scaling to very high powers may be possible. Perhaps the biggest promise of the FEL is the ability to design the laser at an ideal atmospheric propagation wavelength. Significant technical hurdles remain in reaching the status of a deployable FEL, in scaling the beam to megawatt powers and in providing the necessary engineering to turn a laboratory device into a weapon system of reasonable size. For Navy application, FELs will require improvements in areas of radiation shielding, high vacuum, high-current photo-injectors, and probably cryogenic cooling—all of which must be integrated into a ship's basic design.

Fiber lasers and SSLs are the leading-candidate Navy lasers for medium power, as FELs are for high power. All are electrically driven and can meet the requirement for shorter wavelength, capable of transmitting at the "maritime window" of approximately 1 μ.

HEL weapons' advantages include a highly directional and narrowly focused beam, providing:

- Minimal collateral damage
- Speed-of-light delivery
- Rapid retargeting
- Low cost of engagement

Disadvantages center on:

- Limited range due to atmospheric attenuation
- Weather limitations
- Low efficiency (often less than 10 percent)
- Need for eye protection
- Relatively large size and weight requirements

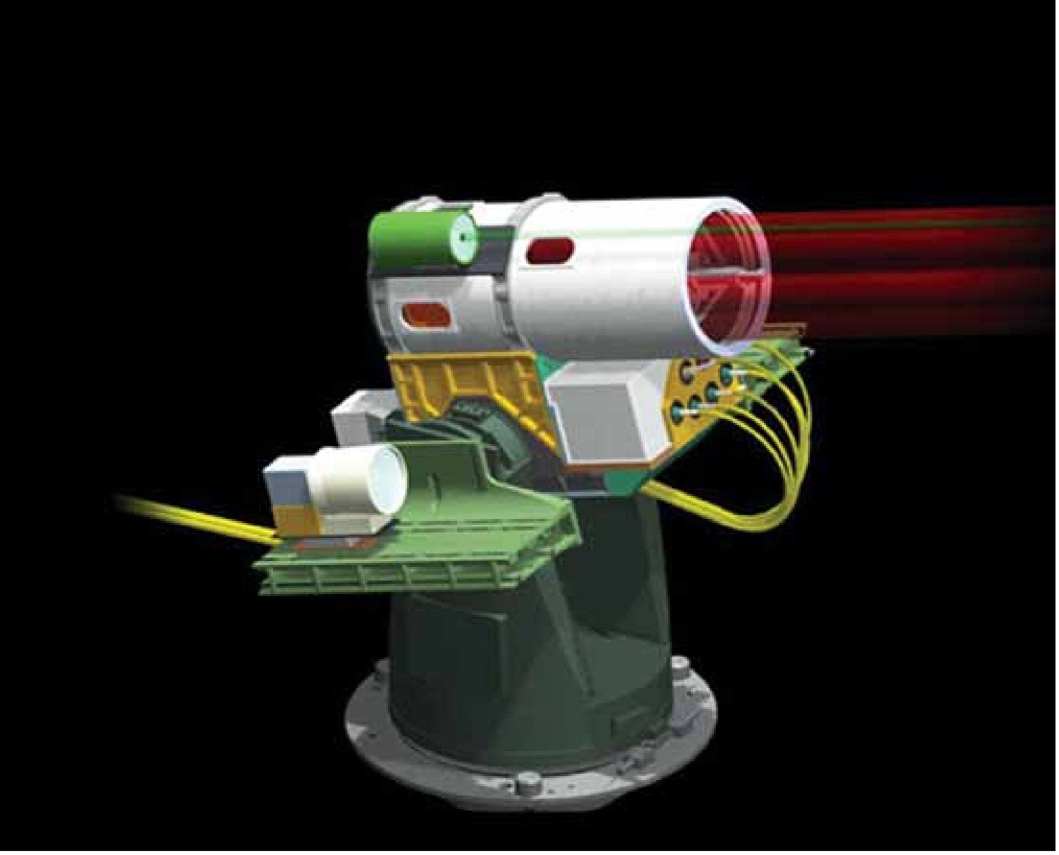

Figure 4. Drawing of Laser Weapon System (LaWS)

Long dwell times (seconds) will be needed for most targets. As with RF systems, there is a potential nonlethal or variable lethality capability since the energy can be easily defocused. A critical challenge is the understanding of a laser beam's propagation through a maritime boundary layer environment, where the sea and air interface creates turbulence and moisture gradients. Measuring the atmosphere and compensating for variations in real time may require adaptive optics or "rubber mirrors" that can be constantly adjusted to compensate for changes. Focusing a small spot at long range will require high beam quality and large optics, probably meter-size mirrors that are very highly reflective and very clean.

HELs in the future are expected to be able to focus energy to a spot size of much less than a meter at ranges of kilometers. This will necessitate very accurate target tracking systems, and precise stabilization and beam-pointing systems, both of which are difficult but should be feasible in the near term. Real-time atmospheric measuring systems will be needed for compensation techniques. Methods to protect the sensitive optical system from salt spray and corrosion will also be needed.

From a lethality perspective, three considerations need to be better understood before a HEL can be deemed a true weapon system:

1. Achievable spot size of beam on target at range
2. Amount of coupling into the target material
3. Subsequent effects of the damage inflicted

For the more severe threats, such as high-speed, antiship cruise missiles, HELs face the difficult task of engaging maneuverable, stealthy, inbound missiles. As such, a better quantitative understanding of the interactions among a laser beam's energy deposition, target material, and flight dynamics is needed.

Particle Beams

A particle-beam weapon is a directed flow of atomic or subatomic particles. These particles can be neutral or electrically charged. Neutral beams need to be used outside the atmosphere (in space), where charged particles would repel and fly apart. Charged-particle beams (CPBs) are easier to make and are used within the atmosphere, where air molecules can constrain the beam. A CPB weapon transmits matter—not just EM waves—like lasers and microwave weapons. The particles are near the speed of light and deposit their kinetic energy deeply into any target material. They have the potential to be highly destructive weapons and are very difficult to shield against.

Charged particles are produced by applying a strong electric field near a material that emits electrons. These electrons then pass through accelerating stages with high voltage gradients (often megavolts), which increase the electron's velocity. As the electrons pass each stage, the velocity increases until they approach the speed of light (become relativistic), at which point they have substantial energy to penetrate a target. The accelerating systems can be linear, but a recirculating design is more compact and can reuse stages. These systems are basically high-current versions of scientific particle accelerators.

Once the electron beam is produced, it must propagate to the target. High-velocity electrons will not go far before they collide with air molecules and lose energy. The fact that air molecules struck by the beam are heated and moved out of the way for a short period of time creates a rarified "hole" in the atmosphere through which a second pulse can travel farther. In this manner, a fast series of pulses can "hole-bore" to the target, each pulse going farther than the last. The final pulse must have enough energy to damage the target. The deceleration of electrons in the atmosphere causes

Bremsstrahlung radiation in the forward direction toward the target, creating gamma rays that, in turn, create X-rays and RF radiation.[a] These effects can cause electronic upset and "soft-kill" mechanisms even if the beam slightly misses the target.

The beam of electrons is typically a few centimeters in diameter. When a beam strikes a target, the energy is deposited deep in the material (the collision cross section is small because of the relativistic speeds) in microseconds (much faster than a laser), creating thermal shock that is very difficult to shield against. For an explosive target, there is also the possibility of causing a deflagration or low-order burn, disrupting the normal warhead mechanism.

Scientists studying CPB weapons made significant technical advancements in the 1980s, but the weapons are still far from being practical. A CPB weapon is technically very challenging and expensive to build. Studies project that the volume requirements necessary for a CPB system could be on the order of a 5-inch gun system. Advantages of a CPB weapon include rapid penetration, a deep magazine, all-weather capability, and soft-kill mechanisms for a near miss. Problems include complexity, size, limited range, and the need to demonstrate compact accelerators and propagation mechanisms.

ELECTROMAGNETIC (EM) LAUNCHERS

A number of technology concepts to launch projectiles exist using electrical energy. These systems rely on large currents in conductors, creating strong magnetic fields that drive a projectile. The velocity of a normal powder gun projectile is limited by the expansion speed of the explosive powder, and present military guns are reaching that limit. With an electric gun, the fields can push projectiles much faster, providing longer ranges and increased kinetic energies. The simplest version is an EM rail gun, shown in Figure 5.

In any conducting loop, the generated magnetic field tries to expand the loop. If everything is held in position, the only movable item is the conducting projectile, which moves down the rails in an attempt to expand the loop. Since megajouoles of projectile energy are needed for EM rail guns, energy storage

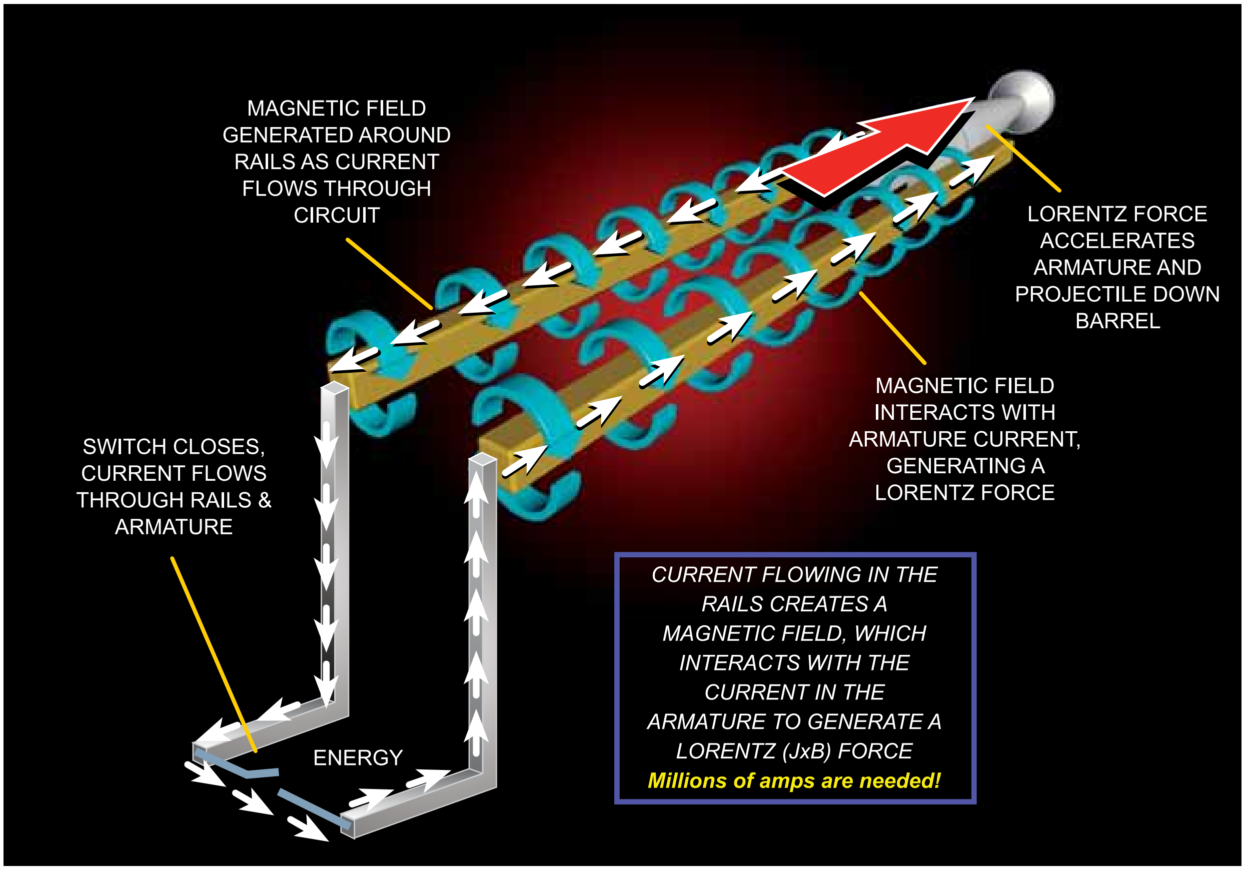
MAGNETIC FIELD GENERATED AROUND RAILS AS CURRENT FLOWS THROUGH CIRCUIT
LORENTZ FORCE ACCELERATES ARMATURE AND PROJECTILE DOWN BARREL
MAGNETIC FIELD INTERACTS WITH ARMATURE CURRENT, GENERATING A LORENTZ FORCE
SWITCH CLOSES, CURRENT FLOWS THROUGH RAILS & ARMATURE
ENERGY

CURRENT FLOWING IN THE RAILS CREATES A MAGNETIC FIELD, WHICH INTERACTS WITH THE CURRENT IN THE ARMATURE TO GENERATE A LORENTZ (JxB) FORCE
Millions of amps are needed!

Figure 5. Electromagnetic (EM) Rail Gun Concept

mechanisms that can store about 100 MJ are needed, along with the ability to discharge the energy in milliseconds. To generate useful forces, millions of amps of current are needed—a major challenge and significant loss mechanism. Large capacitor banks with very high-current switches are required. Spark gap switches have historically been the only option, but new high-current solid-state switches are now becoming available. Capacitor energy densities, too, have improved an order of magnitude in the last few decades. Rotating machines have also been considered because they are smaller than equivalent capacitor banks, but extracting the energy quickly, without tearing the machine apart, has been problematic. The launch energy of various projectiles is shown in Figure 6.

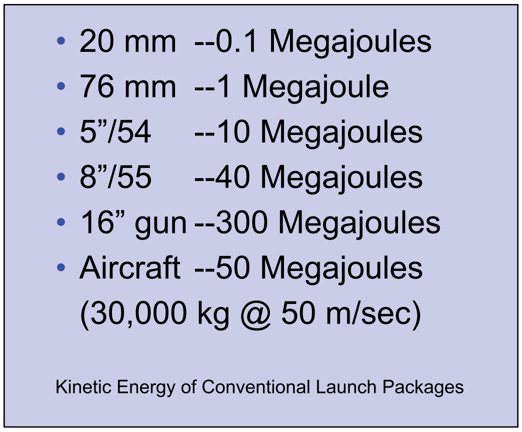

- 20 mm --0.1 Megajoules
- 76 mm --1 Megajoule
- 5”/54 --10 Megajoules
- 8”/55 --40 Megajoules
- 16” gun --300 Megajoules
- Aircraft --50 Megajoules (30,000 kg @ 50 m/sec)

Kinetic Energy of Conventional Launch Packages

Figure 6. Launch Energy of Various Projectiles

A rail gun is probably the most compact form of electric launcher. However, it requires direct electrical contact between the projectile and barrel rails, creating the potential for arcing, melting, and erosion. Coil guns use a series of sequentially fired coils around a “barrel” to push the projectile in stages. This does not require direct electrical contact, so it avoids rail erosion but requires a series of fast timed switches and more space. Linear induction motors are basically unrolled electric motors and have been used on electric trains and roller coasters, typically with magnetic levitating systems to avoid contact erosion. This concept is being developed by the Navy for launching aircraft. The energy to launch an aircraft is similar to a large-caliber projectile—more weight but less speed. The slower speeds are more suitable for rotating machines since the launch times are seconds rather than microseconds.[1] Electrothermal guns and electrothermal-chemical (ETC) guns use a combination of electricity and chemicals. Electrical energy is used to initiate chemical reactions that can produce lightweight driving gases, like steam, or allow more energetic propellants that are difficult to ignite in a conventional fashion.

Some advantages of electrically driven projectiles include:

- Higher projectile velocity (over conventional explosives)
- Very long range (>100 miles) with lower cost than missiles
- Time-critical delivery (because of shorter time of flight)
- Safer projectile stowage (minimal explosives)
- Potentially adjustable velocity levels, for better accuracy and controllable damage

The potential of having nonexplosive rounds and magazines is very attractive for the Navy. For long-range, large-caliber EM projectiles, the kinetic energy from the projectile velocity is greater than the chemical explosive energy in a conventional round traveling much slower. Therefore, damage can be equivalent even without explosives. System size and lifetime are still behind conventional systems, but getting close.

OUTLOOK

Challenges remain for many electric weapon concepts. These weapon systems appear promising to meet the increasingly important asymmetric threats with low-cost precision rounds. They also can be employed across the energy spectrum for nonlethal targeting. Electric weapon systems will, in many cases, continue to supplement existing kinetic weapon systems in the near term. Despite technology challenges, directed-energy and electric weapons hold great promise in offering the future warfighter unique combat capabilities not currently available.

ENDNOTE

a. Bremsstrahlung—a type of radiation emitted when high-energy electrons are decelerated. (German for ***braking radiation***)

REFERENCE

1. Shope, S. et al., *Long-Range Naval Fire Support with a Coilgun*, Sandia Report 2001-3832, February 2002.

SOLID MODELING OF DIRECTED-ENERGY SYSTEMS

By Joseph F. Sharrow

Not long ago, around the mid-1980s, development of most new mechanical systems—such as automobiles, consumer products, and military devices—was performed manually on a drafting table or drawing board, much like the present-day version shown in Figure 1. These tables and boards performed a necessary function, but they offered little assistance other than for drawing lines. Engineers used them to prepare layouts, or two-dimensional sketches of what they were designing. They then would take these layouts to a draftsman, who would create drawings of each part in the device. The drawings would subsequently be sent to a manufacturing facility.

This layout and drawing preparation process typically would need to be repeated multiple times because mistakes would be made, or design issues would be discovered late in the process. Similarly, the manufacturing process would sometimes require multiple iterations as well because of the inherent limitations in designing three-dimensional (3-D) devices on two-dimensional boards. This less-than-ideal process made it difficult to design and manufacture even mundane products and frequently resulted in things that just didn't work. With the emergence of early computerization, numerical analyses of more complex systems began to be performed. These analyses were conducted to ensure that the systems worked in the real world. For example, engineers might conduct a structural analysis of the forces in a loaded dump-truck bed to make sure that the frame wouldn't bend and fail. Because of the difficulty in performing these analyses, they would often require a specially trained group of structural engineers, expensive software, and large mainframe computers, limiting their use to only the largest, most well-funded companies or organizations.

EMERGENCE OF COMPUTER-AIDED DESIGN (CAD) AND SOLID MODELING

With the availability of smaller scale computers and more economical software in the mid-to-late 1980s, CAD was born, initiating a period of rapid improvement in the design process. This was driven, in part, by the introduction of software packages such as AutoCAD. Initially, these software packages only attempted to automate drawing lines by making wireframe (stick-figure) versions on the computer of what previously had been made by hand on the drafting board. This reduced the difficulty in making changes in the development process, but it still limited the engineer's pallet to a two-dimensional space. What was really needed was a 3-D method of design. Solid modeling

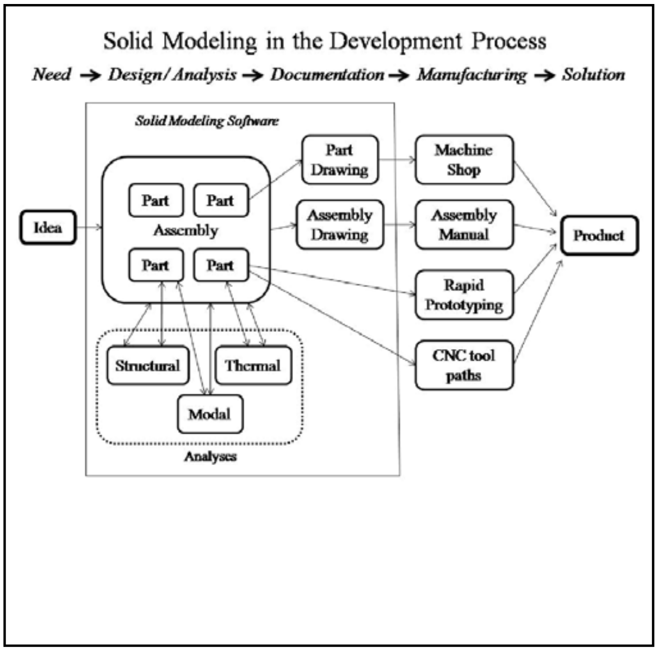
Solid Modeling in the Development Process
Need → Design/Analysis → Documentation → Manufacturing → Solution
Solid Modeling Software
Idea
Part
Part
Assembly
Part
Part
Part Drawing
Assembly Drawing
Machine Shop
Assembly Manual
Rapid Prototyping
CNC tool paths
Product
Structural
Thermal
Modal
Analyses

Figure 1. Drawing Board

addressed this need beginning in the late 1980s to early 1990s.

Solid modeling is analogous to taking blocks of clay and cutting and forming them into the shape of a solid part on a computer. These 3-D parts are then put together in an assembly, more accurately representing real-world devices. Though originally used only in a limited way for specialized applications in the aircraft and automobile industries, it wasn't until the 1990s that solid modeling experienced widespread availability and mainstream acceptance due to software packages such as Pro/ENGINEER. Figure 2 summarizes how Pro/ENGINEER and other similar packages fit into the development of new products. The general flow of the process moves from left to right.

Initially, nearly all 3-D solid modeling packages required significant computing and graphics display power, necessitating the use of large graphics workstations running the UNIX operating system. Rapid advances in computing and graphics power have since enabled nearly all packages to run efficiently on personal computers (PCs) and laptops, bringing solid modeling and analysis capability into the mainstream.

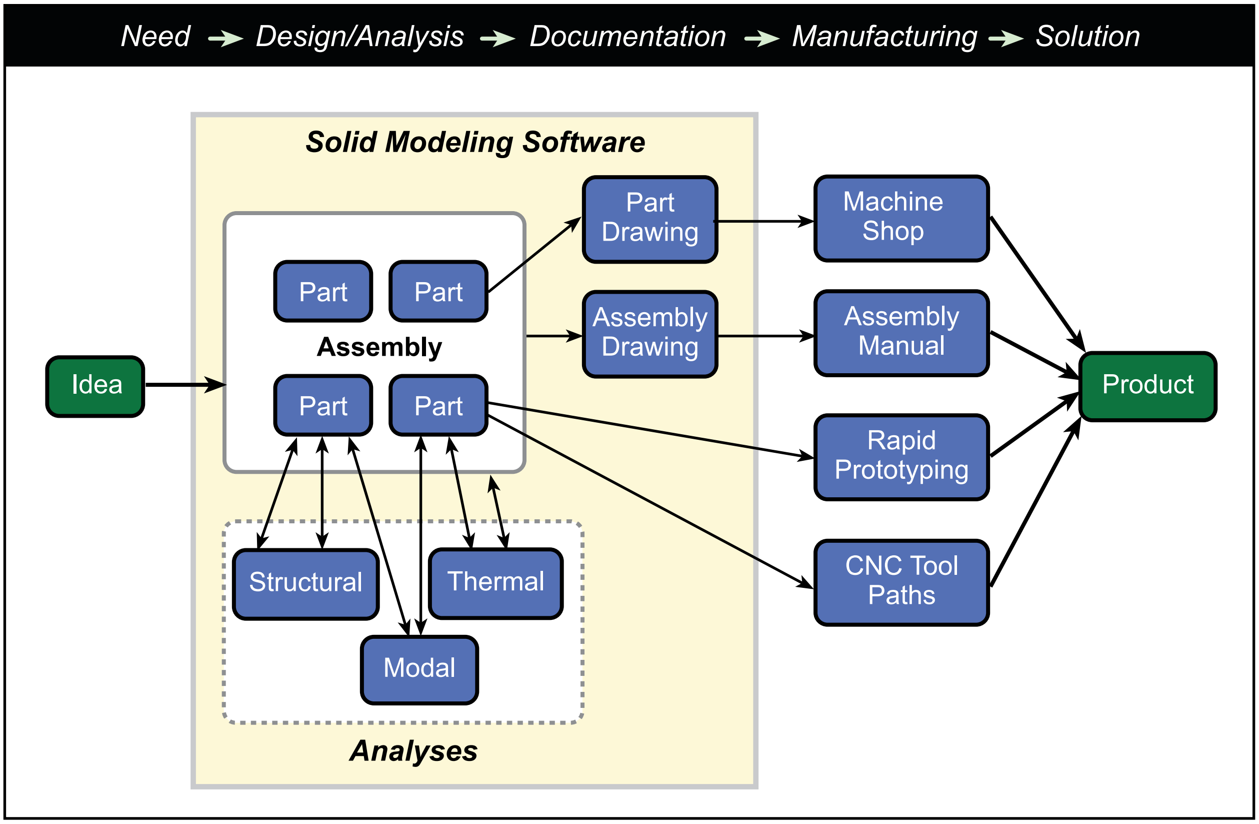
Need → Design/Analysis → Documentation → Manufacturing → Solution
Solid Modeling Software
Idea
Part
Part
Assembly
Part
Part
Part Drawing
Assembly Drawing
Machine Shop
Assembly Manual
Rapid Prototyping
CNC Tool Paths
Product
Structural
Thermal
Modal
Analyses

Figure 2. Solid Modeling in the Development Process

Solid Modeling of Directed-Energy Systems

Engineers working in the Directed Energy Division at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), use solid modeling to develop hardware for nearly all of its programs. Both Pro/ENGINEER and SolidWorks are used extensively to develop new products in virtual 3-D space. Additionally, the structural simulation package within Pro/ENGINEER is used to determine stresses and natural frequencies of parts and assemblies. Consequently, these packages have enabled a single mechanical engineer in the Directed Energy Division and a draftsman in the Engagement Systems Department at Dahlgren to perform the design and analysis work that would have required an entire group of engineers and draftsmen just a few years ago. Today, collaboration among many organizations using similar packages has become commonplace. Insofar as solid modeling has become an indispensable tool for development and collaboration, its successful implementation requires proper training and experience before engineers can use it effectively, just as medical surgeons require training in the use of advanced robotic surgical devices before they can be used effectively. Thus, while these high-tech modeling systems not only have reduced the number of personnel needed for design and development, they have enabled the Navy to get significantly more bang for its buck while supporting warfighting needs. An example of how solid modeling is currently being used is discussed below.

Navy Laser Weapon System (LaWS) Beam Director

The Directed Energy Warfare Office (DEWO) and Directed Energy Division at Dahlgren are currently developing the Navy LaWS for the Naval Sea Systems Command's Directed Energy and Electric Weapon Systems (DE&EWS) Program Office (PMS 405). The program's goal is to take advantage of currently available industrial laser technology and incorporate it into a future naval weapon system. As part of the development process, major subsystems have been integrated with a Kineto Tracking Mount (KTM) into a LaWS beam director. The KTM/beam director was modeled and analyzed using Pro/ENGINEER. Ultimately, the resulting LaWS will be installed on Navy ships on the Close-In Weapon System (CIWS) gun mount. During field testing in June 2009 at the Naval Air Warfare Center, China Lake, California, the prototype KTM/beam director successfully destroyed five unmanned aerial vehicles (UAVs). The actual beam director used in the China Lake testing is shown in Figure 3; the Pro/ENGINEER assembly model used for development is shown in Figure 4.

The LaWS effort took advantage of many aspects of solid modeling including collaboration, structural and modal analysis, and manufacturing drawing creation. The project required development of new, unique hardware, as well as the integration of electronic models from commercial vendors. The KTM model was provided by L-3 Brashear and was originally designed using Pro/ENGINEER. The beam-directing telescope model was provided by RC Optical Systems, Incorporated, and was originally made in SolidWorks. These models were combined with many new optical and structural components developed by the Directed Energy Division into a single, comprehensive assembly model. This model was instrumental in understanding the interaction of the many components, and its use increased accuracy and precision that would have been impossible with old-fashioned two-dimensional development processes. Figure 5 shows a cross section through the main portion of the beam director, revealing the complexity of the many parts and subassemblies required for such a device. In addition to modeling the mechanical components, the actual laser beams were also included to better understand their path through the various mirrors and optical devices in the beam director, and to better highlight any interference they might have with structural components within the KTM or telescope.

Numerous analyses were performed to make sure that everything worked the way it was intended. One major analysis addressed the telescope mount. To ensure that the beams were stable at range, the mount had to be extremely stiff. The best way to ensure this was to perform a structural analysis using the structural simulation package within Pro/ENGINEER. Figure 6 shows the results of that analysis: a displacement plot in which different colors represent how much the telescope will move when the KTM rotates at its maximum speed. The large cylindrical object simulates the mass of the telescope. The minimum amount of displacement is indicated by blue, and the maximum is shown in red. This analysis verified that the movement of the telescope, relative to the optical components within the optics breadboard, was acceptable and should perform well at the range specified by the program office.

After modeling and analysis were completed, manufacturing drawings of custom parts were created by the Engagement Systems Department to be

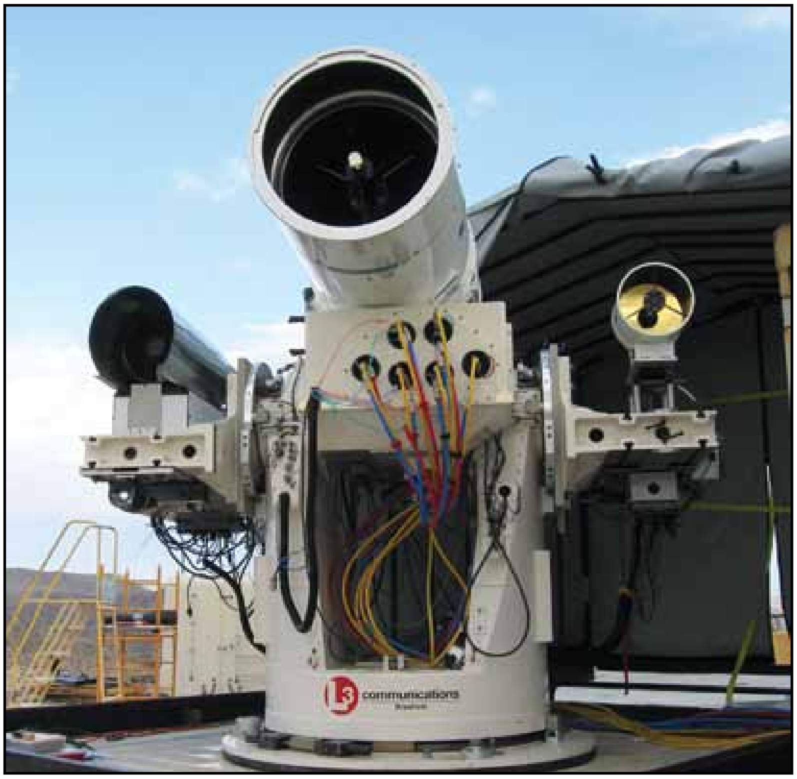
L3 communications

Figure 3. LaWS Beam Director

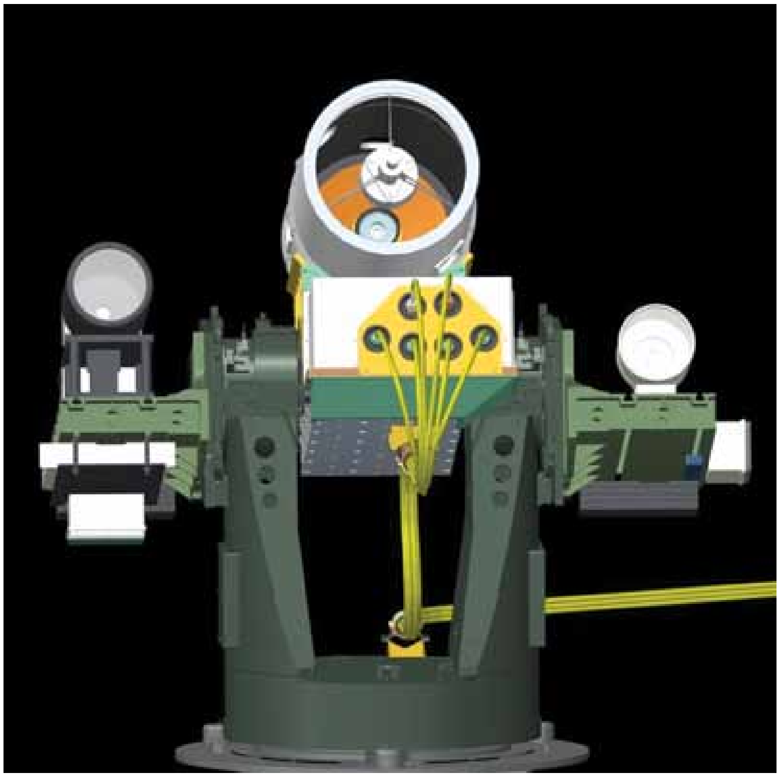

Figure 4. LaWS Beam Director Assembly Model

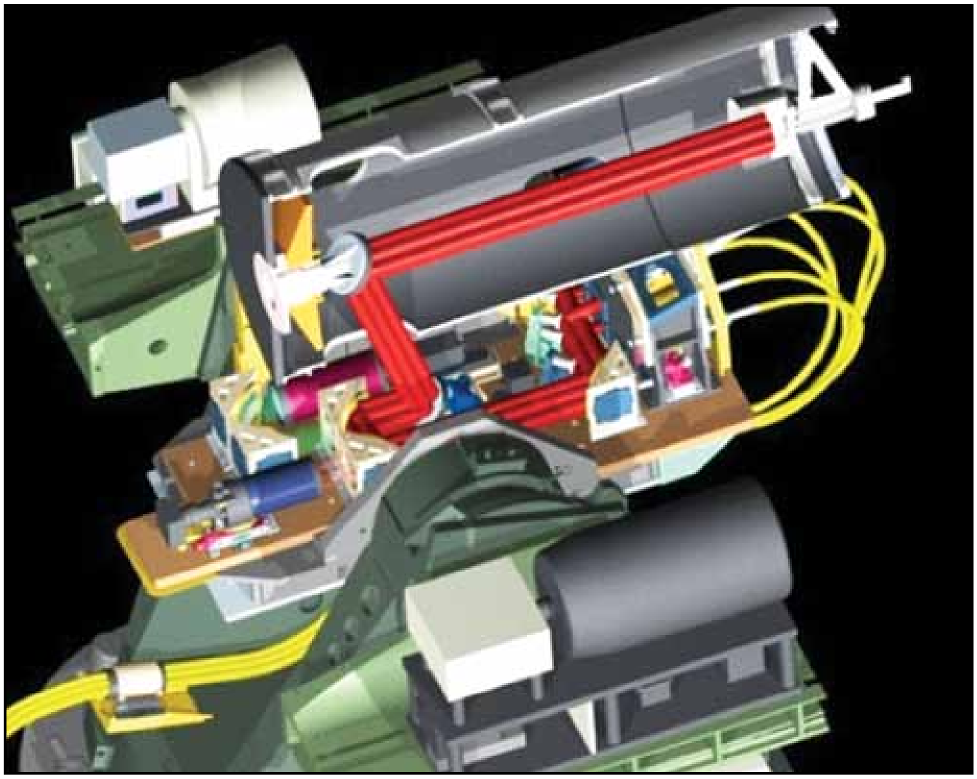

Figure 5. LaWS Beam Director Cutaway

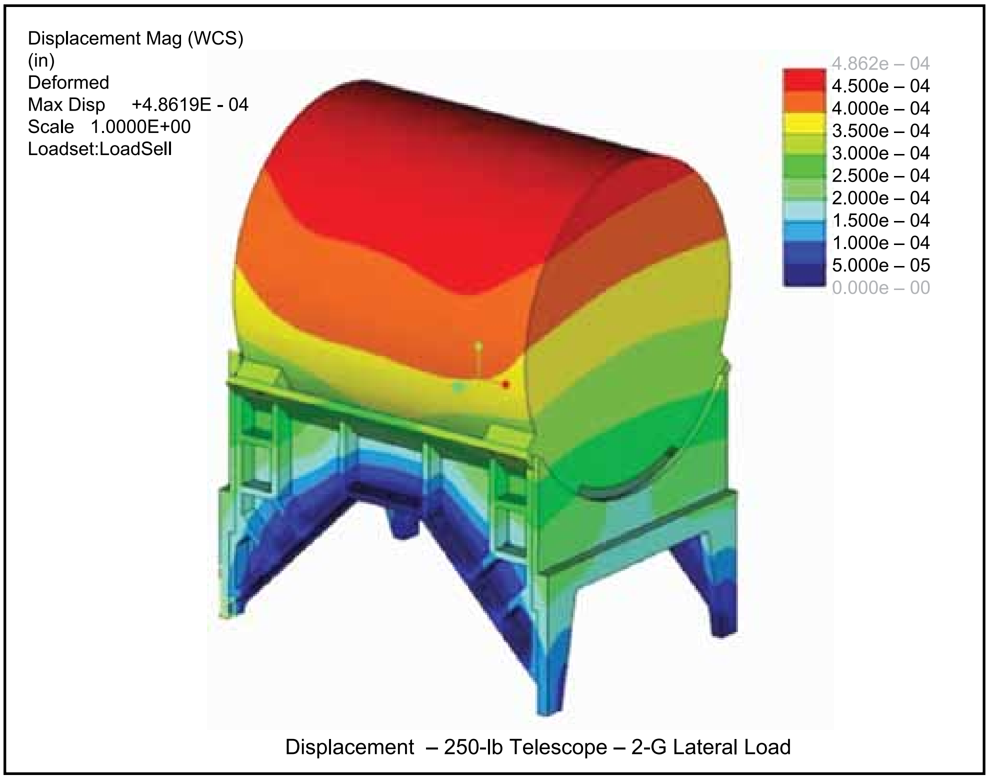

Figure 6. Displacement Plot

sent to manufacturing facilities, such as machine shops. One example is shown in Figure 7, which shows the first sheet of the multisheet drawing needed to manufacture the large plates that support the telescope from the center platform of the KTM. One of these plates is also shown in the displacement plot in Figure 6.

Even though it would be possible for one person to do all of the modeling, analyses, and drawings for a particular program, a more efficient process takes advantage of using the best skills available by collaborating with other experts. Collaboration enables assembly, part, and drawing files to be sent electronically, eliminating the need for collocating personnel. Drawings for the LaWS program, for instance, were made using noncollocated personnel across base at NSWCDD. They could just as easily have been made using personnel from across the country.

The LaWS program exemplifies how the Directed Energy Division uses solid modeling to enhance the quality and effectiveness of Navy directed-energy capabilities. As a result, warfighters will be better armed with more effective weapons and capabilities for future naval conflicts.

Acknowledgment

The author would like to acknowledge the valuable contributions of the Engagement Systems Department's drafting group in support of numerous DEWO and Directed Energy Division programs.

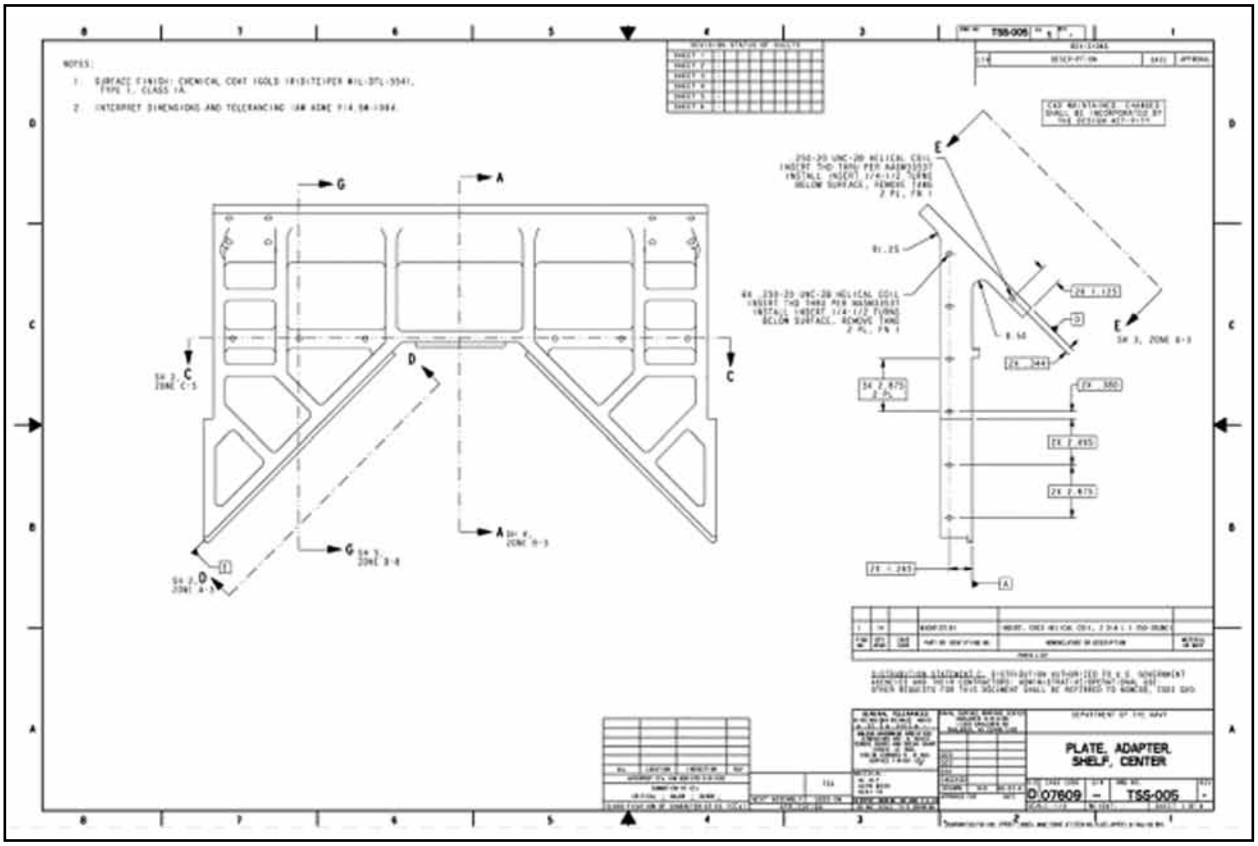

Figure 7. Manufacturing Drawing

A FUNDAMENTAL KEY TO NEXT-GENERATION DIRECTED-ENERGY SYSTEMS

By Directed Energy Division, Electromagnetic and Sensor Systems Department

Imagine an explosive ordnance disposal (EOD) unit on a routine scouting patrol deep in the notorious "Triangle of Death" south of Baghdad, where Marines, Sailors, and Soldiers frequently find themselves exposed to improvised explosive devices (IEDs). Fortunately, this newly outfitted unit is equipped with the latest unmanned, mobile, remote-controlled, radio frequency (RF) transmitter used as a directed-energy weapon (DEW). The integrated system provides comprehensive IED prediction, detection, prevention, and neutralization capabilities. Lightweight, pocket-sized transmitters carried by each warfighter constantly communicate sensor intelligence, key vital signs, critical conditions, and location telemetry to a geostationary satellite (GEOSAT). It intercepts, collects, and retransmits intelligence and situational awareness data simultaneously to any command post in the world and to each member of the unit on patrol. Highly efficient, miniature, switch-mode, RF amplifiers with high-power density (small size and weight with high-power output) enable these visions of future capabilities as their systems' transmitter backbone.

To civilians, the miniaturization of modern wireless (electromagnetic) devices is considered a mere convenience or luxury, i.e., Blackberries, mobile phones, and high-speed wireless network connections. To the next-generation warfighter, miniaturized, wireless, directed-energy (DE) systems open the door to the realization of a whole new set of effective and efficient wireless modalities. And while the capabilities mentioned in the above scenario are not yet available to warfighters, researchers believe they have uncovered the key to next-generation DE systems leading to the miniaturization of DE devices.

NEXT-GENERATION DE SYSTEM REQUIREMENTS

At the Naval Surface Warfare Center, Dahlgren Division (NSWCDD), key system requirements for effective next-generation DE systems are being researched and developed for applications to counter IEDs, to detect explosively formed penetrators (EFPs), to neutralize explosives, and to predict threat locations. Next-generation DE systems must yield a high probability of mission success and be inherently safe to operate. By design, they must minimize or eliminate the risk of hostile attack or collateral damage especially during screening missions. Considering the DEW example above, practical

080102-N-1132M-006 SHEIK SA'ID, Iraq (2 January 2008) U.S. Army Soldiers attached to 3rd Squadron, 2nd Cavalry Regiment patrol and search for weapons or Improvised Explosive Devices (IEDs) during a clearing mission. (U.S. Navy photo by Mass Communication Specialist 1st Class Sean Mulligan/Released)

next-generation DE systems must be physically characterized by:

- Low mass (weight)
- Small size (volume)
- High-power output with respect to size or high-power density
- High efficiency for extended mission use
- Minimized prime power and cooling support
- Portability
- Mobility
- Configurability

They must also ensure a high probability of mission effectiveness. The DEW must be easily transportable and agile, adapting to the immediate, local military mission requirements in various warfighting environments. Additionally, DE systems must be mechanically robust and able to withstand the shock and vibration of combat missions in rough and rugged environments. The key requirement—efficiency—fundamentally facilitates all required characteristics, including mass and size.

Moving Beyond Requirements

Scientists at NSWCDD, sponsored by the Office of Naval Research (ONR), are researching and developing key system requirements for effective next-generation DE systems to counter IEDs, to detect EFPs, to neutralize explosives, and to predict threat locations.

Researchers at NSWCDD are leading the way toward realizing small, lightweight, RF transmitters using high-power, solid-state, switch-mode amplifiers, theoretically 100 percent efficient. These practical switch-mode amplifier realizations are at least 1/100 the volume and weight of any commercially available linear solid-state amplifier of comparable power output. The challenges included assessing what type of active amplifier device and operation would provide the greatest power density (power output per unit volume and mass) with its necessary auxiliary systems, such as prime power generation and cooling of waste heat. Such a device also needed to provide sufficient output power based on required standoff range and IED system-coupling efficiency while also maintaining a manageably-sized, easily transportable system. Researchers initially considered tube-based systems, but large, heavy, direct-current (DC) power supplies are required, and typically 40 percent of the input power is dissipated in heat, which negates any possibility of miniaturization.

Upon a practical review of amplifier-class operations and suitable active amplifier devices, however, research pointed to contemporary switch-mode amplifier schemes (e.g., Class-E and Class-F) using solid-state technology—such as the high-electron mobility transistor (HEMT)—as satisfying the high-power density and abusive mechanical requirements for expected worst-case transportation and operation in a rugged environment. To significantly impact reduction of size and weight, practical, high-efficiency thresholds were defined for next-generation DE systems at 90 percent and greater. The key technology enabler to realize amplifier high efficiency in high-power amplifiers up to 60 kW was found in exploiting contemporary switch-mode amplifier architecture with efficient power combining. Particularly, switch-mode schemes in Class-E and Class-F operation as solid-state, active-hybrid planar topology designs were found to be necessary and sufficient for DE applications. These analyses led to a novel, Class-E RF switch-mode amplifier design. A Class-E RF switch-mode amplifier can theoretically operate at 100-percent efficiency. For every input watt supplied, an RF output watt is produced. The conductors and dielectric substrate of the hybrid planar load network and the commercial off-the-shelf (COTS) transistor all exhibit some small degree of power loss, suggesting an estimated practically realized efficiency of 90 percent.

Moreover, the amplifier under research consisted of a novel microwave load network operating with high-power output at ultrahigh frequency (UHF). This research led to the state of the art in Class-E designs leading by hundreds of watts, several hundred megahertz in frequency, and roughly 10 percentage points in efficiency. A common, solid-state, high-power amplifier design technique sums the phase and amplitude of smaller amplifier units to the large values required for DE systems. A practical hardware limitation exists that limits the theoretically infinite number of fixed RF output power units to a finite number. Approximately 60-kW RF output power sets the boundary as the largest hardware realization. By applying spatial power combining in the propagating medium, phased-array antennas can be employed with constructive wave interference in air that would allow sufficient RF power densities on target, based on the number of elements in the array. This technique eliminates the traditional hardware necessary to power combine the smaller power-amplifier elements, realizing a much simplified DE system with enhanced power density in the transmitter, and reduced mass and volume.

The key to ultrahigh efficiency in a switch-mode amplifier, such as Class-E or Class-F, is found in zero-voltage switching (ZVS). Here, the load network is not only designed to be resonant at and around a particular desired switching frequency, it must simultaneously act to force the voltage across the switch to be zero when current flows and when it switches off; hence, theory suggests that no power is dissipated because the product of current through, and voltage across, the switch is zero. It is this aspect of the design that makes the job of switch-mode amplifier realization difficult. Of course, in practice, a small voltage exists for a very short time during the switching action, resulting in a small amount of input power being dissipated in heat. This theoretical description also assumes that all components are ideal (i.e., no impedance to current flow exists in the switch when turned on). All realistic switches exhibit finite impedance when turned on, which does dissipate some wasted energy, but again, this is very small in modern HEMT devices using the ZVS technique.

Class-E switch-mode amplifier theory development began in the United States during the 1960s, with details published in 1975, although some earlier reports were published in Russia. Lumped element electrical components (RF choke inductors and metal film capacitors) were initially used in lower frequency (3 to 30 MHz) prototypes. As engineers attempted higher frequency designs in the very high frequency (VHF) range, solid-state transistor switch parasitic intrinsic and packaging elements found inside the transistor began to be used as some of the key components necessary for ZVS. These parasitic elements included stray capacitance caused by differences of potential between parts inside the transistor and inductance caused by bond wire length that is used to connect the transistor to accessible terminals in its packaging. At microwave frequencies, these parasitic elements become sensitive, invoking unintended significant changes to load networks designed to operate with the transistors. Intrinsic elements include drain-to-source breakdown voltage capability and peak current capability. As the need for higher frequency operation and higher power increased, constraints of key transistor parameters became difficult to produce in traditional silicon technology:

- High instantaneous transient (peak) current capability through the transistor
- Moderate breakdown potential across the transistor
- Low output capacitance

Only within the past few years have transistor manufacturers produced COTS transistors that meet the required capabilities necessary to operate in switch mode for microwave frequencies and

081107-N-1120L-072 RAMADI, Iraq (7 November 2008) Joint EOD Rapid Response Vehicles (JERRVs) assigned to Naval Mobile Construction Battalion (NMCB) 7's convoy security element are secured following an escort mission from a forward operating base. The Cougar-type JERRVs are employed by coalition forces for escort and logistics missions, and to protect personnel from IEDs. NMCB 7 is deployed to U.S. Forces Central Command to provide contingency construction support to coalition forces in support of Operations Enduring Freedom and Iraqi Freedom. (U.S. Navy photo by Mass Communication Specialist 2nd Class Michael B. Lavender/Released)

high-power output. Selection is still somewhat limited for designers.

New transistor technology known as gallium nitride (GaN) HEMTs—using state-of-the-art manufacturing processes with GaN on silicon carbide materials—now facilitates Class-E high-power amplifier (100-W) designs at ultrahigh frequencies. The design process for switch-mode amplifiers is radically different than linear amplifiers, so engineers have tended to continue using linear amplifier design techniques due to familiarity, rather than advance to the switch-mode designs. Today, the Class-E and Class-F unit power output (greater than 100 W) capability and upper frequency limitation is based on a lack of available HEMTs with the necessary parameter capabilities.

Most recently, transistor manufacturers have limited their investment in the Class-E amplifier solid-state switch market due to no commercial market mandate. An assortment of presently available HEMTs provides a low-power capability in terms of 1- to 10-W output power for Class-E amplifiers in the cell phone market. The need remains to continue advancing in commercially manufactured HEMTs with key capabilities necessary to realize larger unit power output, hundreds of watts to a thousand watts, for practical implementation in DE systems.

POSSIBLE MULTIPLE APPLICATIONS

Directed-Energy Weapon Systems

Expanding on the vision of the next-generation DEW system mentioned at the beginning of this article, further imagine that EOD scouts detect a laser fluorescence signature of C4 high explosive and chlorine outgasses in the vicinity of an abandoned vehicle 2-km north of their current position. An electronic support measure (ESM) team on board an approaching clearing vehicle initiates RF jamming and electromagnetic surveillance procedures. Electronic specialists also scan the area with ground-surface differential thermography—particularly to detect possible buried IEDs and EFPs or their tiny command wires, crush wires, or pressure plates—while clearing a pathway to the abandoned roadside vehicle.

Upon arrival at a 500-m safe distance, the EOD specialists command the RF transmitter's robotic platform, also equipped with sensitive gamma-ray planar and computed tomography (CT) imaging to navigate toward and around the vehicle, interrogating every possible hiding place. It discloses an IED in the fuel tank. The specialist lifts the transmitter arming safety and commands the remote transmitter to radiate a prescribed dose of RF energy directed at a carefully chosen component of the vehicle-borne IED (VBIED) system. Without entering the vehicle, the advanced screening system detects and defuses the deadly IED buried within the rusty, metal vehicle chassis. Within minutes, the suspected VBIED threat is entirely neutralized, with absolutely no wounded warfighters or casualties.

Pictured here is the National Aeronautics and Space Administration/National Oceanic and Atmospheric Administration (NASA/NOAA) Geostationary Operational Environmental Satellite-P (GOES-P) launching from Cape Canaveral Air Force Station, Florida, aboard a Delta IV rocket procured by Boeing Launch Services on 4 March 2010. Built by Boeing Space and Intelligence Systems, GOES-P will provide NOAA and NASA scientists with data to support weather, solar, and space operations, and will enable future science improvements in weather prediction and remote sensing. Additionally, GOES-P will provide data on global climate changes and capability for search and rescue.

Mobile Ad-Hoc Wireless Network (MANET)

Beyond IED detection and neutralization, imagine an expeditionary unit on patrol, with each member equipped with an RF transceiver about the size and weight of a cigarette pack with an ultrahigh-efficient switch-mode amplifier. The miniature transceiver constantly communicates sensor intelligence, key vital signs, critical conditions, and location telemetry to a GEOSAT. This small switch-mode amplifier has the needed output power to reach an altitude of 35786 km, where the GEOSAT intercepts, collects, and retransmits this intelligence and situational awareness data to any command post in the world and to each member of the unit on patrol simultaneously. The expeditionary unit, spread out over a wide area with large interspacing, shares the situational awareness and intelligence data of each other at the speed of light. Thus, near real-time, worldwide communications with ubiquitous secure access from the battlefield is possible in a multiple-input, multiple-output (MIMO) architecture. The same system could provide a soldier-to-soldier MANET.

Next-generation switch-mode RF amplifier designs could also optimize payload weight and volume on board new communication satellites while supplying higher power density and making efficient use of the solar power supply budget. Improved switch-mode amplifier power output, when combined with enhanced antenna design, would minimize Earth-station antenna size requirements. The recently launched satellite shown at left demonstrates an example of the latest antenna technology.

Looking Forward

Miniaturizing next-generation DE systems opens up a whole new world of applications to support warfighters in ways unimaginable just a few years ago. Reduction of transmitter mass and volume, accompanied with high efficiency, creates a welcome trickle-down effect. Low profile, small, lightweight DE systems means:

- Less vulnerability to attack
- Greater mobility and maneuverability
- Simplified logistics with less fuel-supply demands
- Less impact on the environment

Clandestine operations, too, could be executed with greater ease and simplified logistics support. In the case of MIMO MANETs, miniaturized high-power density transmitters could further expand capabilities for the warfighter, enabling them to carry high-power transmitters to communicate with satellites or other supporting platforms. The

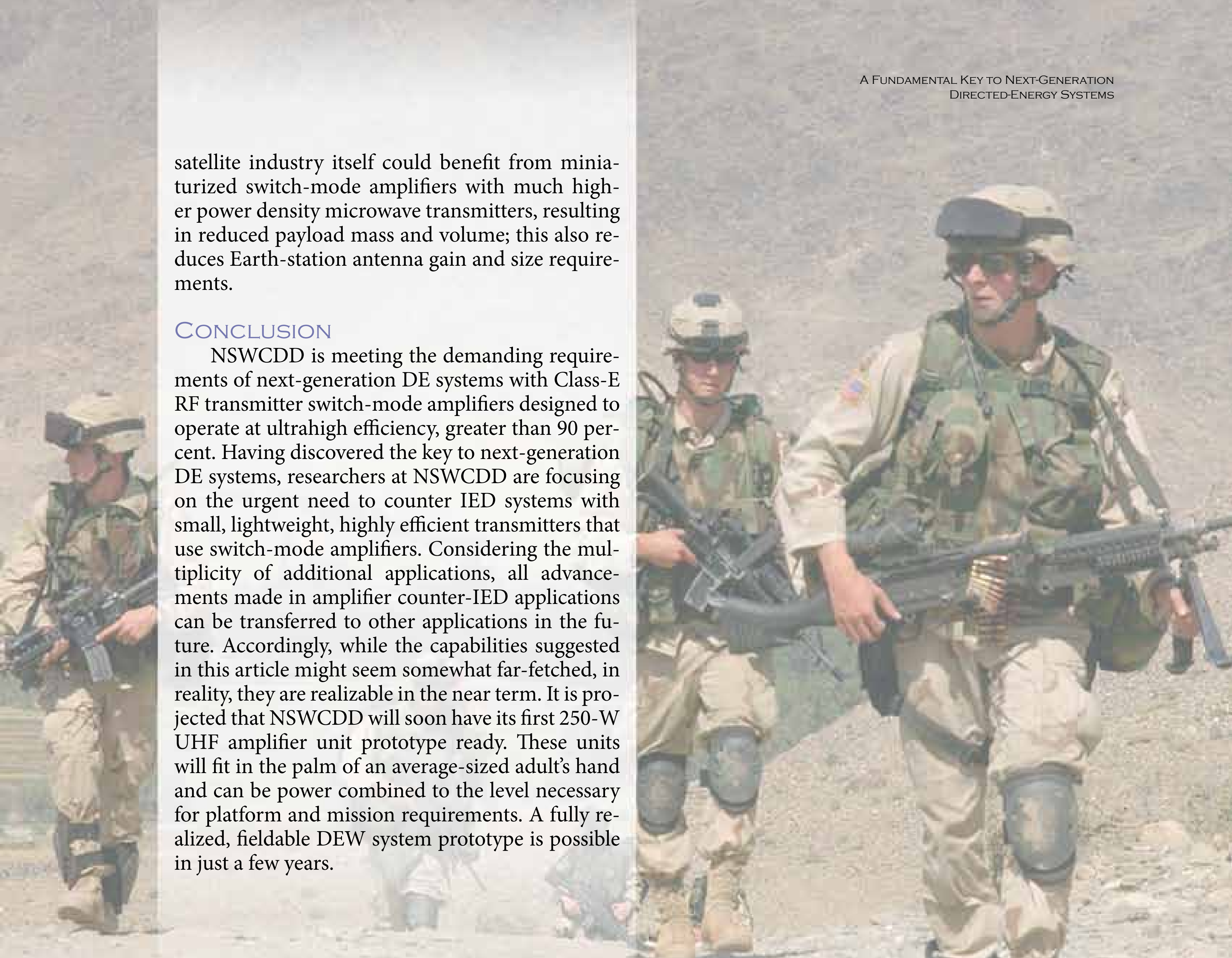

satellite industry itself could benefit from miniaturized switch-mode amplifiers with much higher power density microwave transmitters, resulting in reduced payload mass and volume; this also reduces Earth-station antenna gain and size requirements.

CONCLUSION

NSWCDD is meeting the demanding requirements of next-generation DE systems with Class-E RF transmitter switch-mode amplifiers designed to operate at ultrahigh efficiency, greater than 90 percent. Having discovered the key to next-generation DE systems, researchers at NSWCDD are focusing on the urgent need to counter IED systems with small, lightweight, highly efficient transmitters that use switch-mode amplifiers. Considering the multiplicity of additional applications, all advancements made in amplifier counter-IED applications can be transferred to other applications in the future. Accordingly, while the capabilities suggested in this article might seem somewhat far-fetched, in reality, they are realizable in the near term. It is projected that NSWCDD will soon have its first 250-W UHF amplifier unit prototype ready. These units will fit in the palm of an average-sized adult's hand and can be power combined to the level necessary for platform and mission requirements. A fully realized, fieldable DEW system prototype is possible in just a few years.

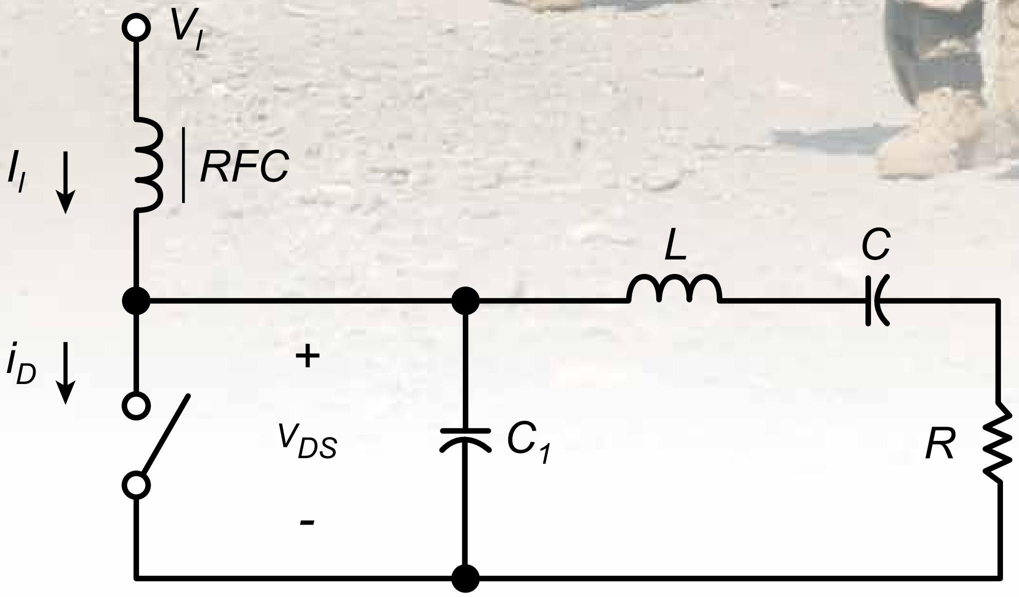

ACTIVE DENIAL ARRAY

By Randy Woods and Matthew Ketner

Active Denial Technology (ADT)—which encompasses the use of millimeter waves as a directed-energy, nonlethal, counterpersonnel weapon—has the potential to provide an important new escalation-of-force capability to U.S. operating forces. ADT projects a focused beam of 95-GHz millimeter waves to induce an intolerable heating sensation on an adversary's skin, repelling the individual with minimal risk of injury. More than a decade of research has established the biological and behavioral effects of ADT for large spot size systems, such as Active Denial System 1 (Figure 1). While the effects of this large spot size system have been successfully established, the technology that produces those effects has the potential to progress in a number of ways, particularly with the development of smaller, lighter, and lower-cost systems.

One research effort focuses on the development of smaller, lighter, and lower cost ADT demonstrators that produce commensurate "ADS-effects," with effective spot size and power densities on target. In support of this effort, the Joint Non-Lethal Weapons Program (JNLWP) sponsored the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) to develop a "smart target system," which measures the millimeter-wave beam using fast-response, 95-GHz diode detectors. NSWCDD subsequently developed and tested the W-Band Beam Diagnostic Array to characterize the system's beam with a temporal resolution of 30 Hz and a high spatial resolution of 1 inch.

The current method of measuring the 95-GHz beam is to use carbon-loaded Teflon (CLT) to produce an average power beam image. This method works as the CLT is exposed to the system's beam. The material heats, over a period of seconds, proportional to the magnitude of the radio frequency (RF) field, resulting in an image as shown in Figure 2. After the exposure, the specific heat capacity of the CLT can be used with the temperature increase in the CLT to provide an indication of the total energy deposited in the material. This method produces a good representation of the average RF field; however, any peak variations in the beam are averaged out.

To allow for high temporal-resolution measurements of the 95-GHz beam, a high-density, 95-GHz diode-detector array was commissioned by the Joint Non-Lethal Weapons Directorate (JNLWD), and was designed and built by NSWCDD, with support from Millitech, Inc. The array consists of a center 11 × 11 matrix (shown in Figure 3) with four removable arms that can be attached (shown in Figure 4), resulting in a measurement area of approximately 1 × 1 m.

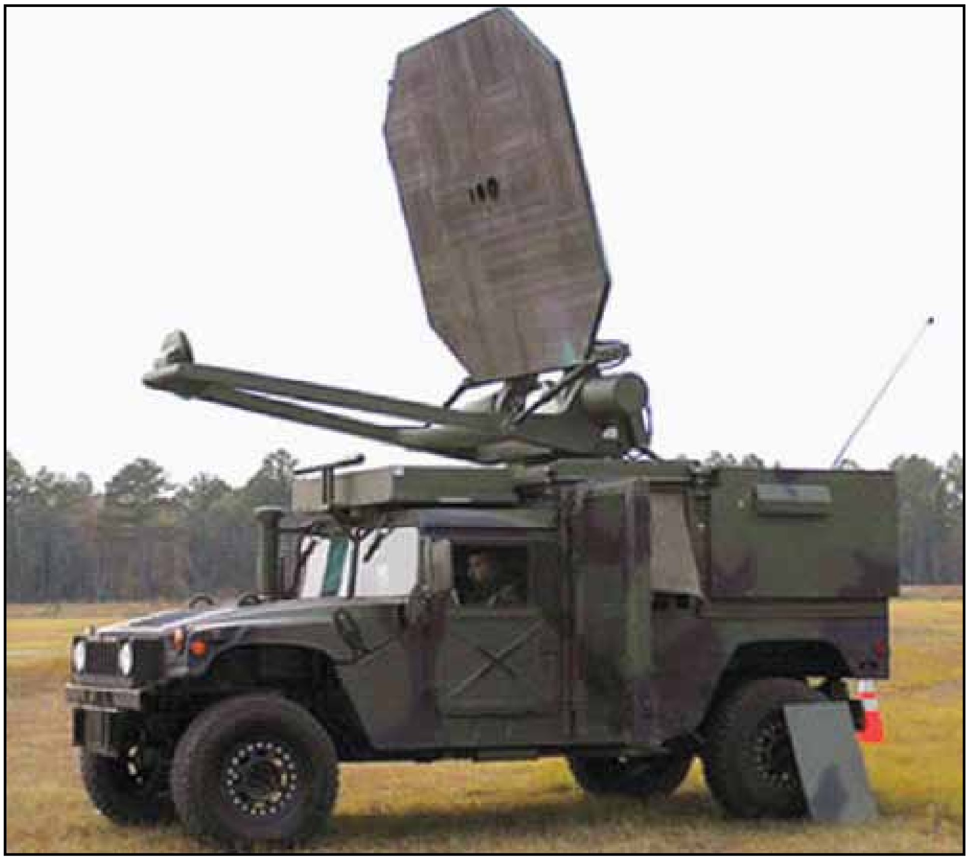

Figure 1. Active Denial System 1

Each element's profile consists of the individual horn antenna from the array, an attenuator, a detector, and a SubMiniature version A (SMA) connection to the digitizer circuitry. This configuration allows for the power received from the antenna to be attenuated and converted to a direct current (DC) output capable of being measured by an analog-to-digital converter. The machined antenna elements provide a uniform effective area for each element, allowing field strength (W/cm^2) to be converted into power received (W or dBm). The aperture antennas also provide an impedance match between free space and the waveguide system. A cross-sectional view of the array element is shown in Figure 5, followed by a signal flow diagram shown in Figure 6.

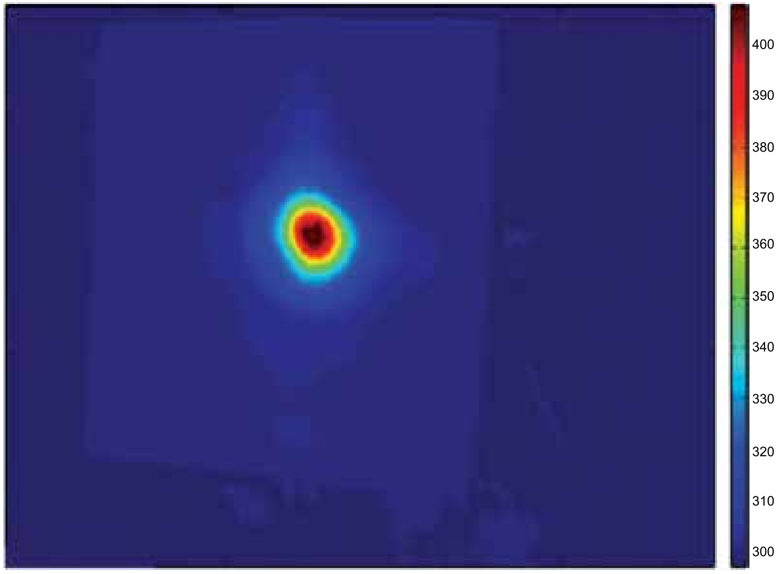

Figure 2. CLT Representation of Small, 95-GHz Spot

The basic principle of operation behind the array is that the derivative of the diode detector's power vs. output voltage curve is very repeatable between detector elements. Therefore, when the detector elements arrived at NSWCDD, each detector element was paired with a variable

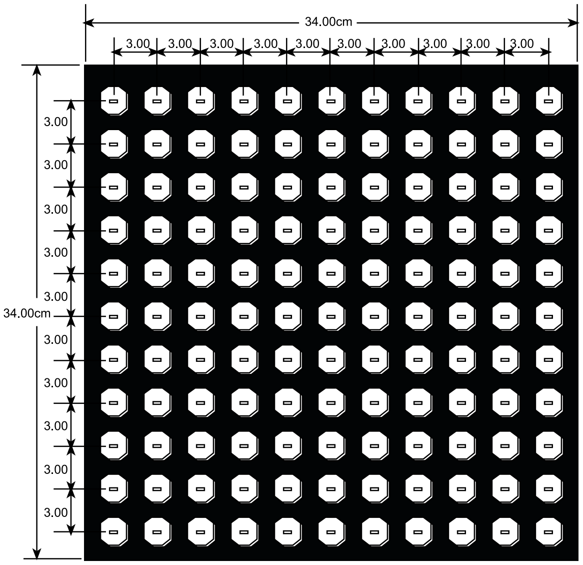
34.00cm
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
34.00cm
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00
3.00

Figure 3. Main Array Face

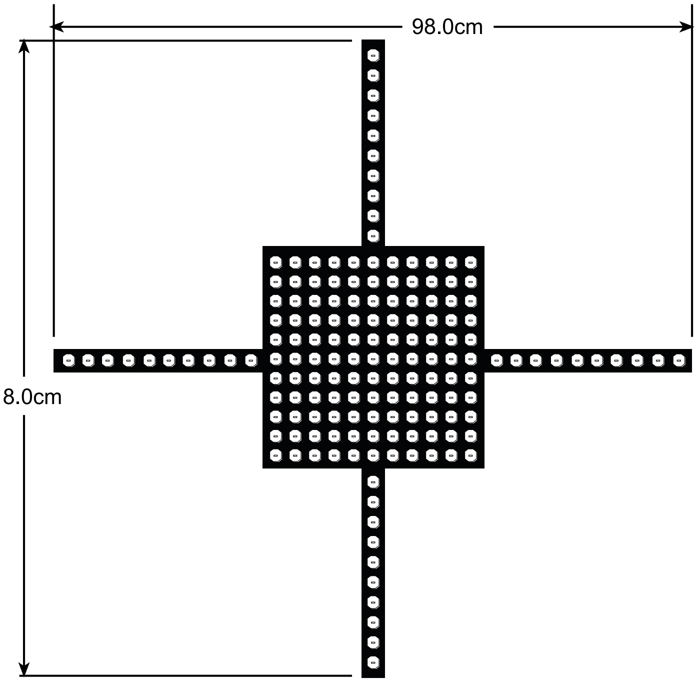

Figure 4. Full W-Band Array

attenuator and calibrated as a single unit. The calibration was accomplished by inserting a known input power of +5 dBm into the input of the attenuator and setting the DC output voltage at a predetermined millivolt (mV) output. This allowed the detector's individual offset voltages to be removed and caused the detectors to behave in a repeatable manner. The attenuator is able to be adjusted by varying the depth that the aluminum nickel card is inserted into the section of waveguide.

The final section of the electrical system converts the DC voltage output from the detectors to a digital signal to send back to the operator station. For this, it was determined that a 16-bit digitizer would be required to enable measuring the microvolts output by the detectors on the low end of their range, while still allowing the digitizer to measure the full output voltage of 1.8 V for high-input powers. Also, due to the proximity of the operator to the array and overall system flexibility, it was determined that Ethernet communications would provide a sufficient means of reading the system data.

To display the data to the operator, a two-dimensional array is populated and displayed for the user (shown in Figure 7). This allows values to be

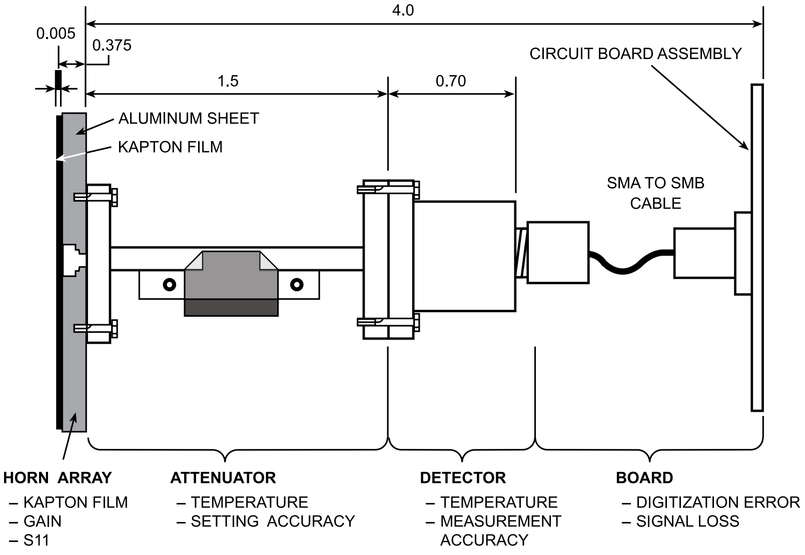
4.0
0.005
0.375
1.5
0.70
CIRCUIT BOARD ASSEMBLY
ALUMINUM SHEET
KAPTON FILM
SMA TO SMB CABLE
HORN ARRAY
– KAPTON FILM
– GAIN
– S11
ATTENUATOR
– TEMPERATURE
– SETTING ACCURACY
DETECTOR
– TEMPERATURE
– MEASUREMENT ACCURACY
BOARD
– DIGITIZATION ERROR
– SIGNAL LOSS

Figure 5. Cross-Sectional View of Array Element

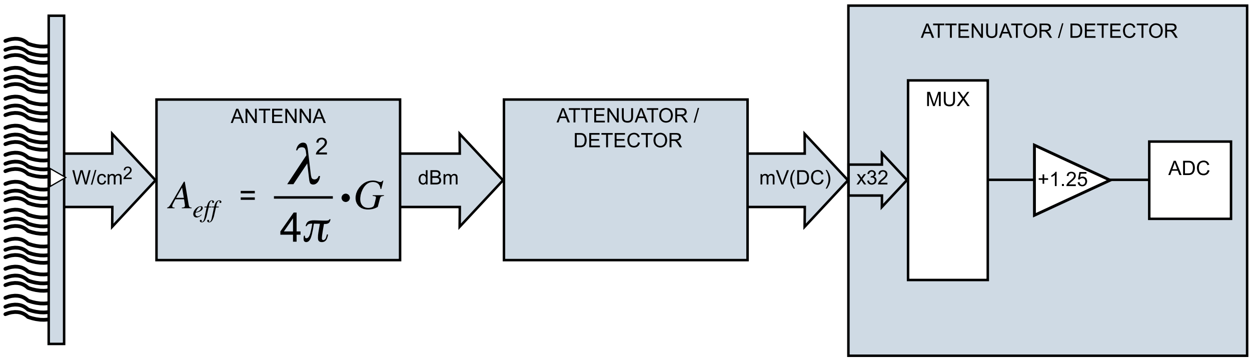
W/cm2
ANTENNA
$A_{eff} = \frac{\lambda^2}{4\pi} \cdot G$
dBm
ATTENUATOR / DETECTOR
mV(DC)
ATTENUATOR / DETECTOR
x32
MUX
+1.25
ADC

Figure 6. Signal Flow Diagram

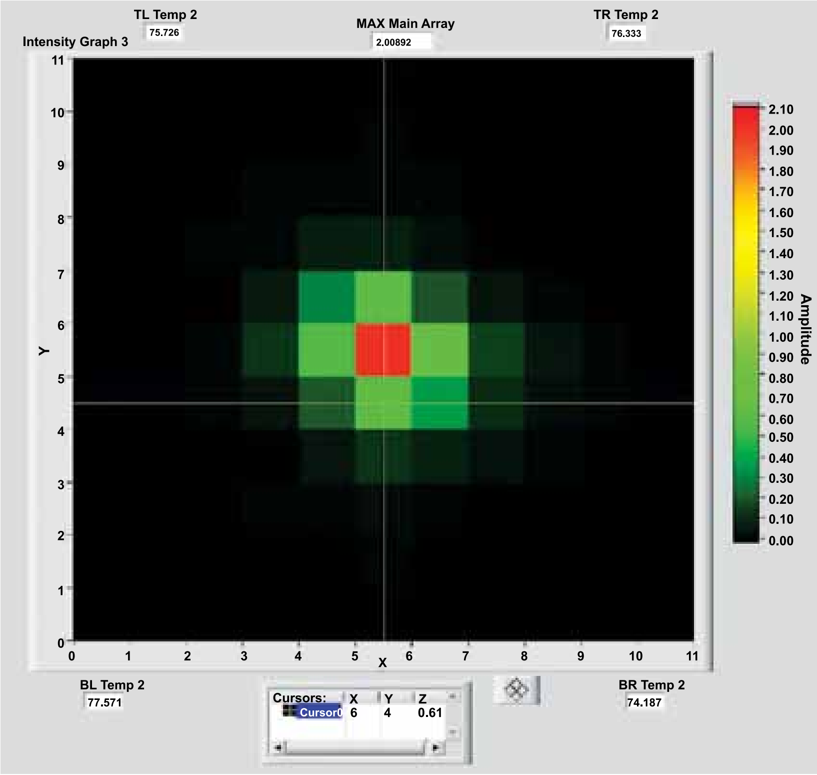

TL Temp 2 75.726

MAX Main Array 2.00892

TR Temp 2 76.333

BL Temp 2 77.571

BR Temp 2 74.187

Cursors: X 6 Y 4 Z 0.61

Intensity Graph 3
Y
11
10
9
8
7
6
5
4
3
2
1
0
X
0
1
2
3
4
5
6
7
8
9
10
11
Amplitude
2.10
2.00
1.90
1.80
1.70
1.60
1.50
1.40
1.30
1.20
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Cursor0

Figure 7. Array's Operator Interface Showing a Small Spot Source

read directly from the display corresponding to the watts per centimeter squared (W/cm^2) present at the array face. Data also is recorded so that it can be viewed later in a player application, such as a video file, or it can be viewed in a spreadsheet application, frame by frame. The data shown in Figure 7 is representative of small-source testing performed recently and very clearly shows the beam profile.

Conclusion

NSWCDD engineers successfully met the W-band array's design goals of providing a high temporal-resolution image of 95-GHz beams. The system has been tested against two active denial systems, providing good agreement with the currently accepted methods, as well as valuable information regarding the system's beam characteristics. These accomplishments will allow future system development to take advantage of this better understanding to possibly reduce system size and increase the effective range. A better understanding of the 95-GHz beam helps to facilitate future ADT development for this much-needed, nonlethal escalation-of-force capability for U.S. warfighters, homeland defenders, and law enforcement personnel.

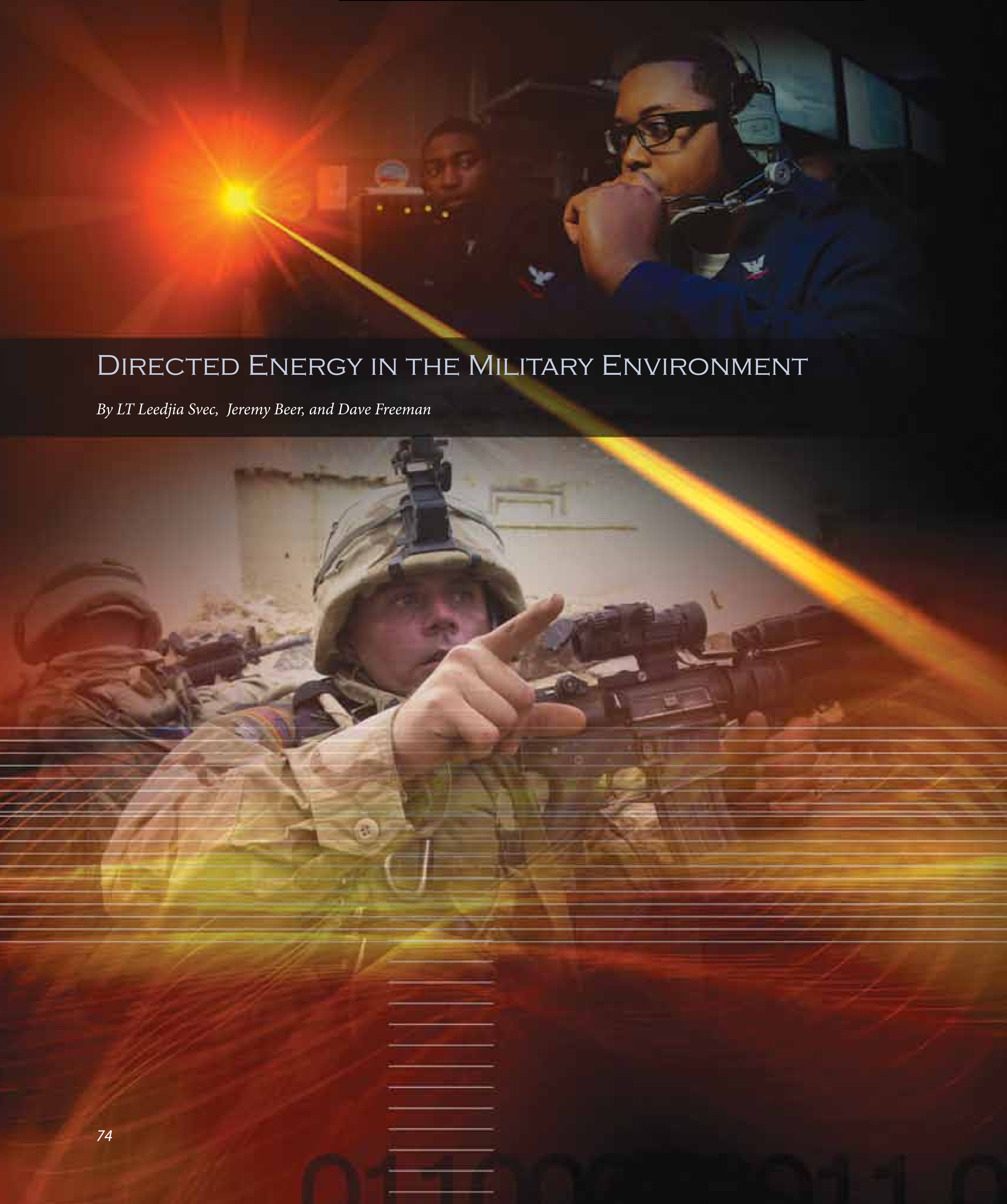

DIRECTED ENERGY IN THE MILITARY ENVIRONMENT

By LT Leedjia Svec, Jeremy Beer, and Dave Freeman

The military operates in the land, air, and maritime environments. In each of these environments, lasers and laser devices are increasingly being seen and used in a variety of ways. Accordingly, the military must protect itself and civilians from the potentially dangerous effects of lasers and other directed-energy devices.

Lasers are being used on the ground to determine the intentions of people who approach checkpoints and to dissuade aircraft from entering restricted airspace. Laser weapons are also being developed for use in the maritime environment. With the use of lasers comes the requirement for eye protection. The eye is particularly sensitive to lasers and its anatomy includes optical components that amplify the power of incoming light. Consequently, the potential for injury or blinding is great.

Naval Medical Research Unit – San Antonio (NAMRU-SA) is poised to lead the way in researching and testing laser glare devices and laser eye protection. The mission of the NAMRU-SA is to conduct medical, dental, and directed-energy biomedical research, which focuses on ways to enhance the health, safety, performance, and operational readiness of Navy and Marine Corps personnel, and addresses their emergent medical and dental problems in routine and combat operations. NAMRU-SA was officially commissioned on 6 May 2009 and is a subordinate command under the Naval Medical Research Center (NMRC) in Silver Spring, Maryland, reporting to Navy Medicine Support Command (NMSC) in Jacksonville, Florida. NAMRU-SA consolidates the Naval Health Research Center Detachment Directed Energy Bioeffects Laboratory, the Naval Institute for Dental and Biomedical Research in Great Lakes, and the NMRC Combat Casualty Care research function. As part of the Base Realignment and Closure (BRAC) 2005, NAMRU-SA has moved to Fort Sam Houston. Two new buildings that have been constructed are the Battlefield Health and Trauma Research Institute and the Tri Service Research Laboratory. A conceptual drawing of the NAMRU-SA Tri-Service Research Laboratory (to house directed-energy research) is shown in Figure 1.

Many factors must be considered when lasers operate in military environments. On the ground, lasers offer a greater likelihood of close contact exposure. In aviation and maritime environments, the mobility of lasers is limited to permanent fixtures on aircraft or ships, so target acquisition can be much more complicated. Often ignored, but just as important and common to all environments, are the psychological factors that need to be explored. These factors include clarifying intentions, communications, and effectiveness. In certain situations,

Figure 1. Naval Medical Research Unit – San Antonio Tri-Service Research Laboratory at Fort Sam Houston, San Antonio, Texas (artist's concept)

sometimes lasers are coupled with other modalities, such as auditory instructions.

On the ground, laser exposure has been shown to interfere with driving vehicles, making color judgments, and target shooting. In aviation, lasers can interfere with pilot vision, causing afterimages, glare, or temporary ocular injury, with attendant effects on navigation and control. In the maritime environment, lights frequently are used to signal a variety of messages, from direction (left, right, etc.) to more complicated messages such as "man overboard." More prolific use of lasers underscores the need for laser eye protection, a dynamic area of research, which must respond to changing threat wavelengths and changing environments. Figure 2 shows NAMRU-SA personnel executing an operational field test at Kennedy Space Center, July 2009.

Recent studies undertaken by NAMRU-SA have investigated the use of laser dazzlers on sailors in small boats.[a] In these studies, participants were exposed to the laser glare at different angles and distances, in both day and night conditions. Study protocols were approved in accordance with the Institutional Review Board in compliance with all applicable federal regulations governing the protection of human subjects. Participants were given a survey assessing their subjective response to the laser, as well as a more objective visual eye chart. The

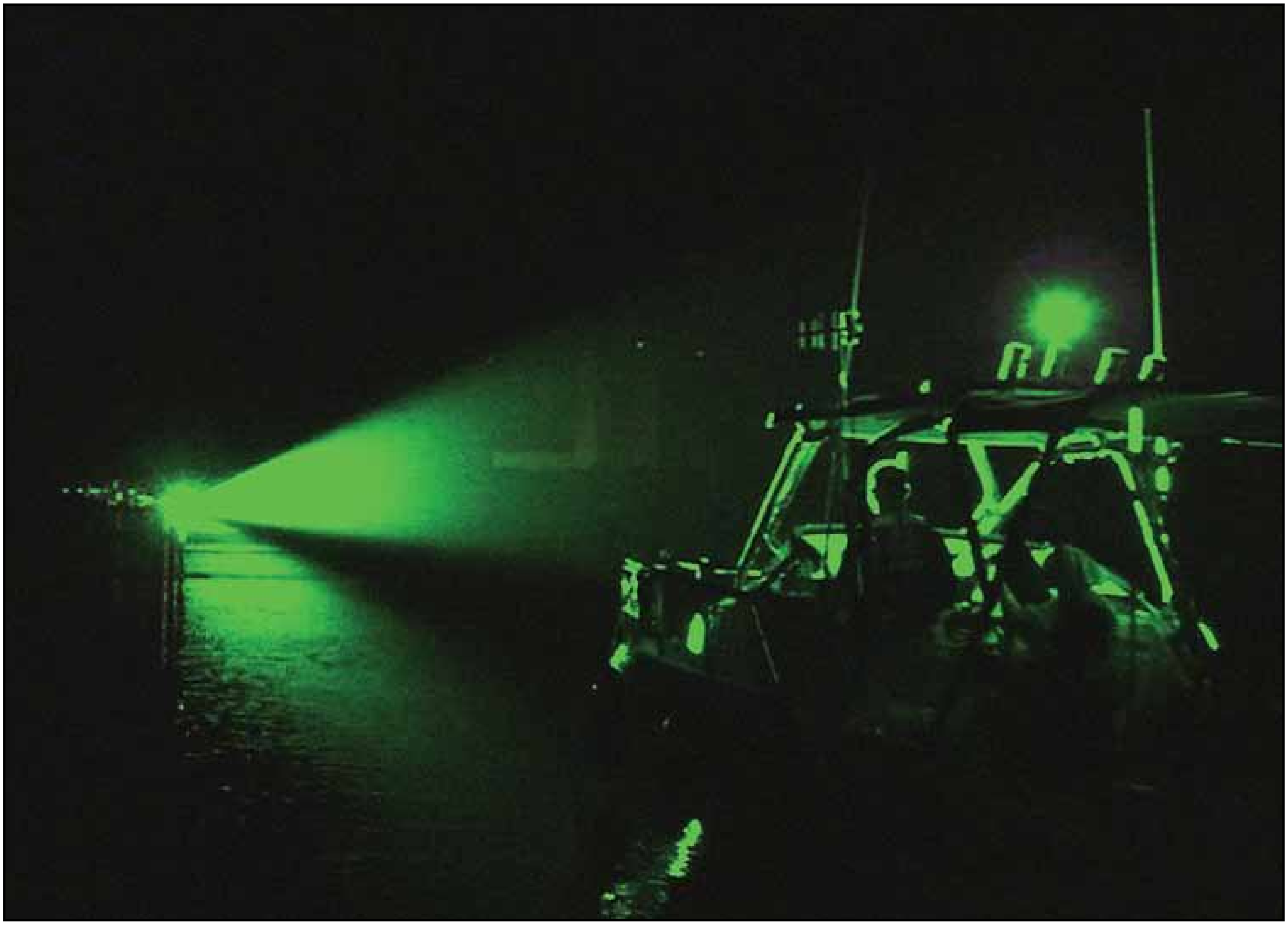

Figure 2. NAMRU-SA personnel execute operational field test at Kennedy Space Center, July 2009, in which a nonlethal laser prototype is evaluated for power delivery (stability and beam propagation) at range and human visual effectiveness aboard a maritime target.

results suggested that participants were most affected by the laser at night when they were looking straight at it (as opposed to many degrees away) and at the closest exposure distances. The most surprising finding, however, was that some participants reported being drawn to the laser rather than away from it, especially at farther distances. Participants remarked that they couldn't tell what the signal was, so they would want to go closer to find out. This illustrates that the assumption (by some)—that distant laser lights will deter and repel innocent mariners—might not always be true. Further research is needed to verify this finding, however, before employing laser glare devices in the maritime environment. Figure 3 shows NAMRU-SA personnel executing operational field tests, which were conducted at Cheatham Annex, Virginia, and Panama City, Florida, in 2008–2009.

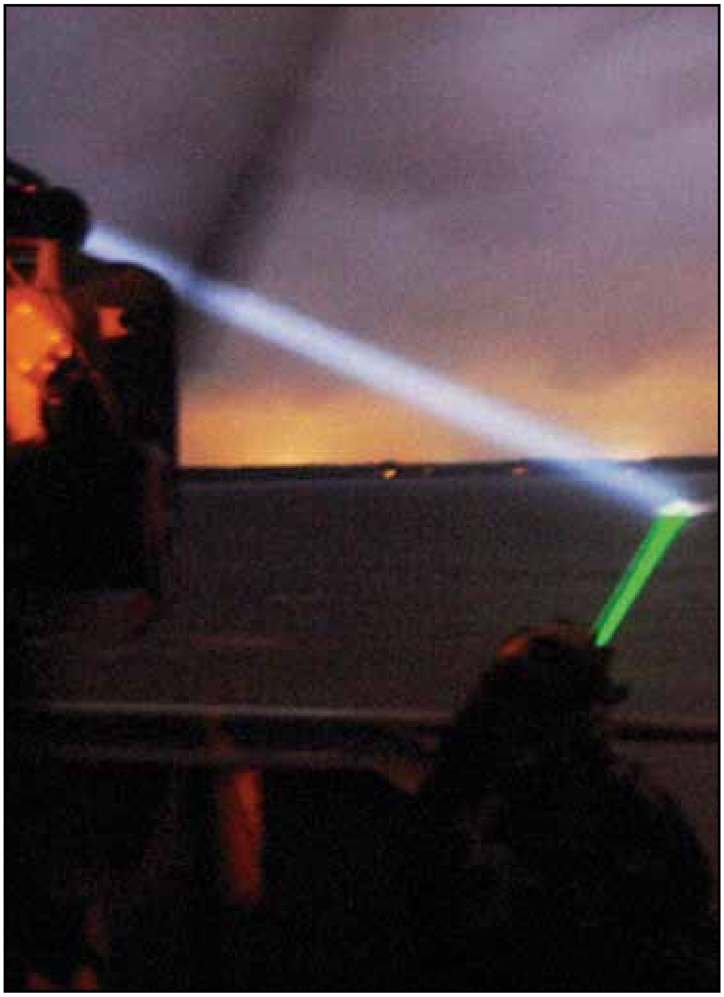

Figure 3. A compact hand-held laser is evaluated for effectiveness in maritime defense against small-boat attacks.

These studies also brought the factor of communication to light. Participants remarked that "green is not a threatening color," and some thought "it could be a signal for help." Many felt curious about the "blinking light" used in the study and would go closer or try to contact the vessel to determine the intent of the message. Green lasers are used because they are more visually salient; however, they may not be as psychologically salient. Participants remarked that if the signal were paired with another signal, such as an auditory one, then the message of "warning" or "do not come closer" might be clearer.

Lastly, these studies brought to light the matter of effectiveness. Laser glare devices are used to stop or alter the behavior of the recipient, but one study yielded mixed results. At close distances, participants noticed the signal, felt affected by it, and reported that their behavior changed in the manner desired by the person pointing the laser. But at greater distances, behavior might not change. Thus, these findings need to be replicated in different maritime scenarios in order to be truly useful in developing laser glare devices. This particular study was encouraging regarding the effectiveness and visual usefulness of glare devices, but it brought up new questions about their psychological impact on behavior. Resolving these questions must be an integral goal of technical research and development studies to determine the operational effectiveness of directed-energy devices, not just for the maritime environment, but for all military environments.

ENDNOTE

a. Results and technical reports are available upon request from the corresponding author or from the NAMRU-SA Public Affairs Officer.

DIRECTED ENERGY USING HIGH-POWER MICROWAVE TECHNOLOGY

By Jacob Walker and Matthew McQuage

The Directed Energy Warfare Office (DEWO) and Directed Energy Division at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) merge past research and data with continuous innovation in the field of high-power microwave(s) (HPM) to address the critical need for nonlethal, nonkinetic weapons. HPM weapons can be described as nonkinetic devices that radiate electromagnetic energy in the radio frequency (RF) or microwave spectrum. They are designed to disrupt, deny, degrade, damage, or destroy targets. In essence, this is achieved when high-power electromagnetic waves propagate through air and interdict targets by traveling through the exterior layers of structures and coupling energy to critical electronic components. Since effectiveness against a wide range of targets is the goal, HPM has become a collective term for various technologies: wave shapes, source frequencies, and the distribution of varying signal bandwidths. It is the objective of HPM research and assessment, therefore, to address targets for which no engagement option currently exists. NSWCDD is working to identify optimal HPM mission platforms and move relevant technologies into the field.

HPM INITIATIVES

NSWCDD has actively pursued HPM research since the advent of the field in the 1970s. Since then, scientists and engineers have conducted HPM research and development in many areas, including hydrogen spark-gap switching, spiral generators, and related technologies. More currently, the Directed Energy Division developed a variety of high-power wideband RF systems based on pulsed power and Marx generators (Figure 1). In addition to the extensive work accomplished in HPM and RF source development, NSWCDD contributed substantially to the area of counter-HPM vulnerability assessments. Researchers developed site assessment guides and threat brochures, as well as a number of wideband RF sources, to determine the susceptibility of electronic equipment to high-power RF interference. This latter effort involved assessing and exploiting the weaknesses of specified electronic targets to various HPM and RF threats. Data gleaned from these efforts was then used to support optimized prototypes and system designs employing effects-based design methodology. NSWCDD utilized these wideband RF sources to determine the susceptibility of a multitude of military

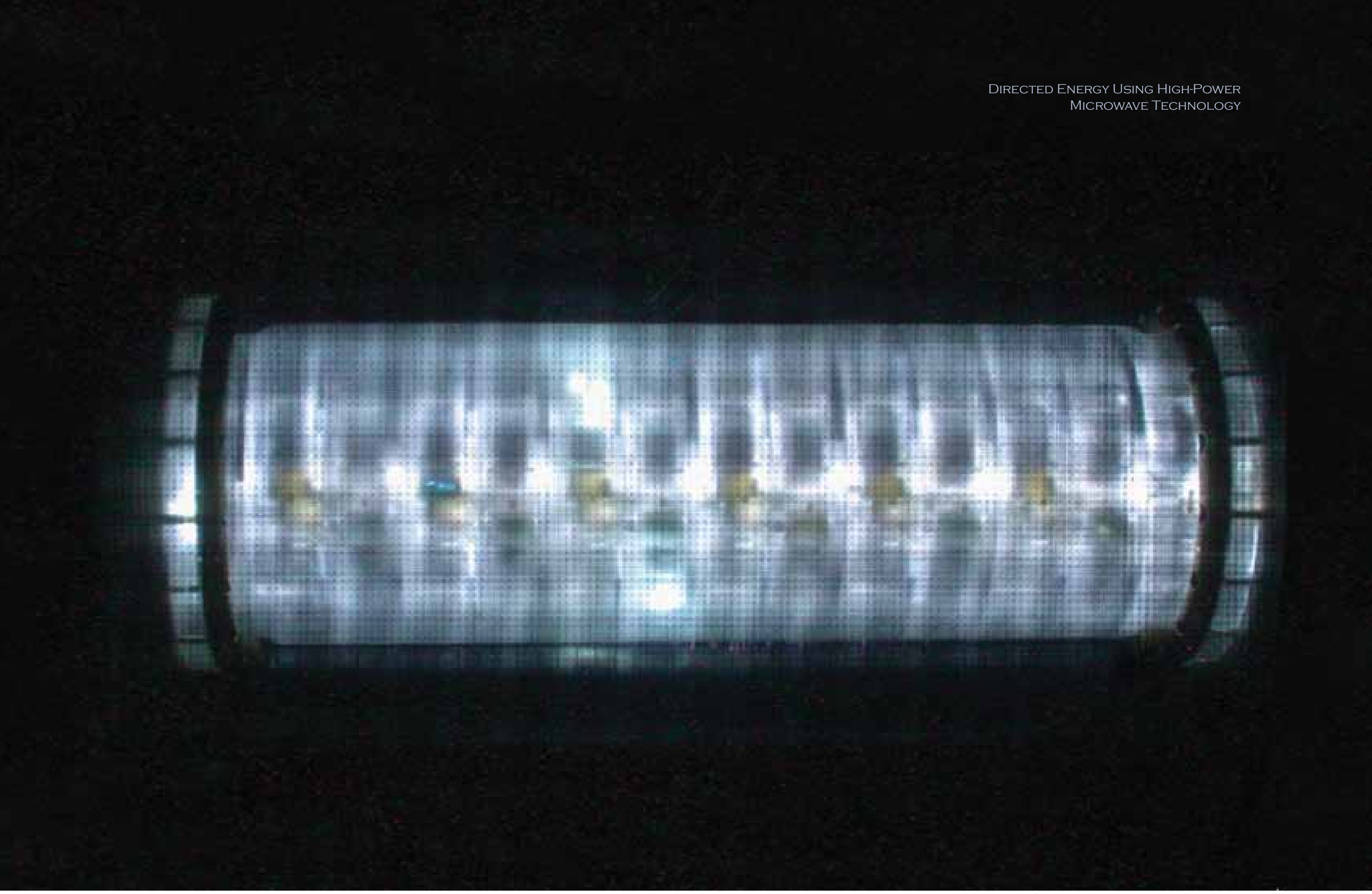

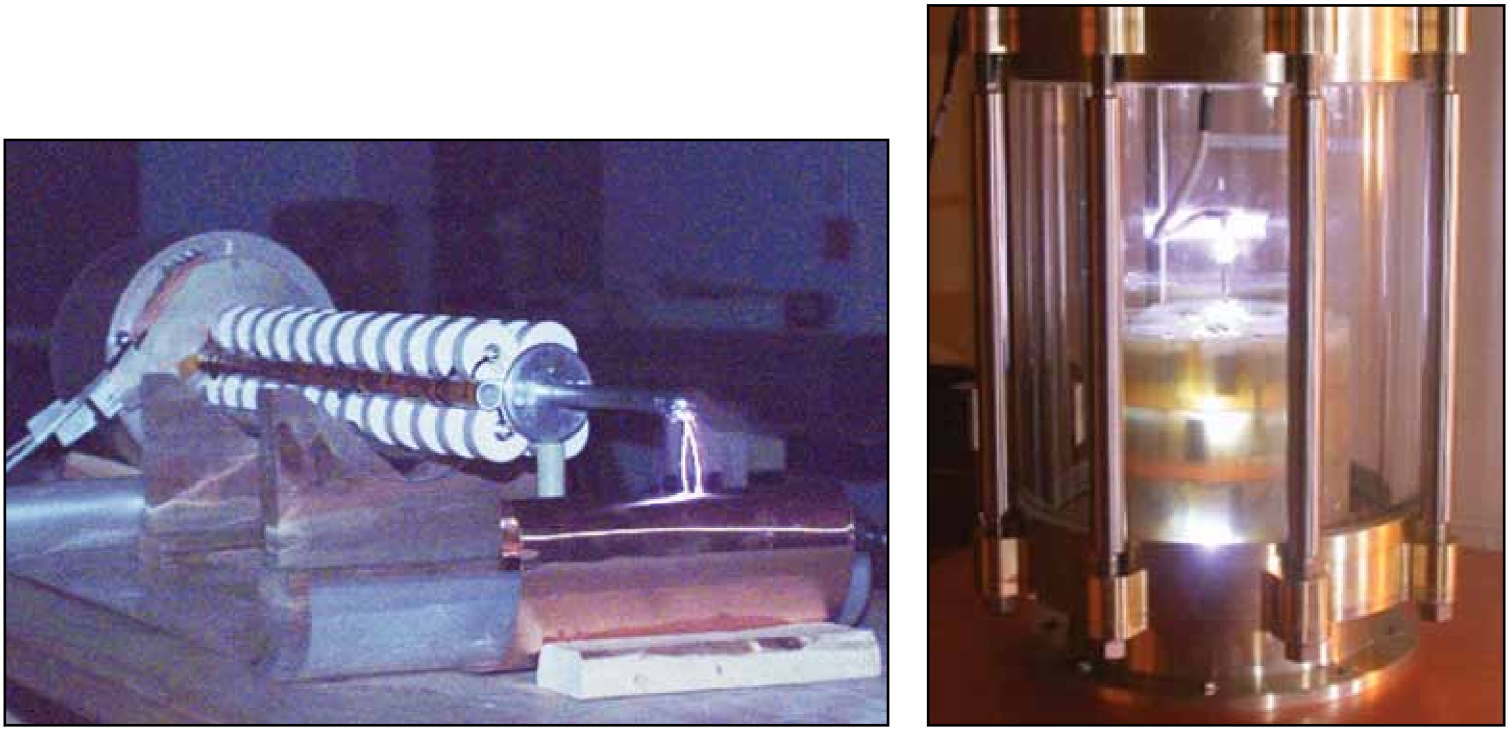

Figure 1. Examples of NSWCDD Marx Generators

and electronic infrastructure equipment to high-power RF interference.

HPM Counterattack Operational Overview

Research in support of HPM-driven electronic attack increased significantly as the demand for nontraditional warfare emerged. Traditional kinetic weapons often are of limited value in peace-keeping missions, for example, as today's enemies frequently are embedded within civilian populations and structures. This creates the need for novel HPM technologies that minimize the risk of collateral damage while effectively neutralizing threats. Dahlgren researchers conduct HPM system research and development—as well as lethality and weapon effectiveness assessments—to address this need while developing technologies against a wide variety of electronic targets. These projects leverage NSWCDD's assets, including the Maginot' Open Air Test Site (MOATS), state-of-the-art RF diagnostics, and modeling and simulation tools to identify applications and platforms in which HPM technologies can be employed. Figure 2 shows a computer model of the MOATS facility and a modeling and

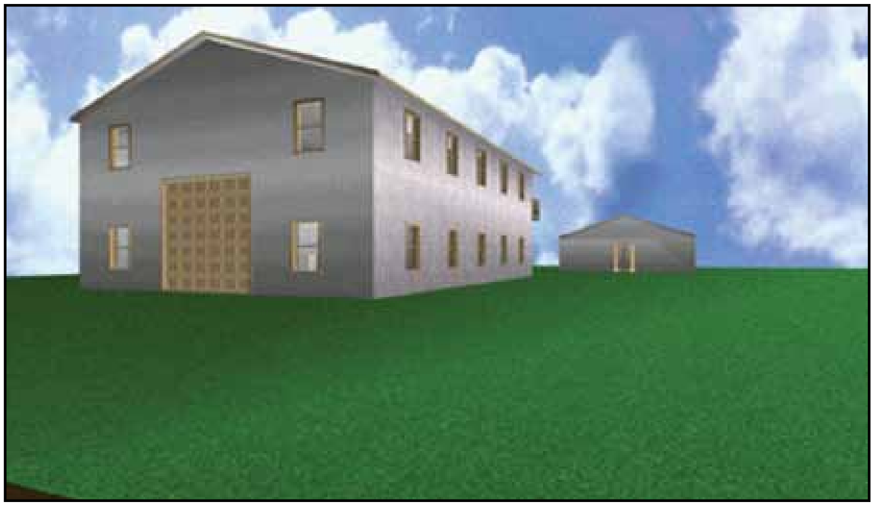

(a)

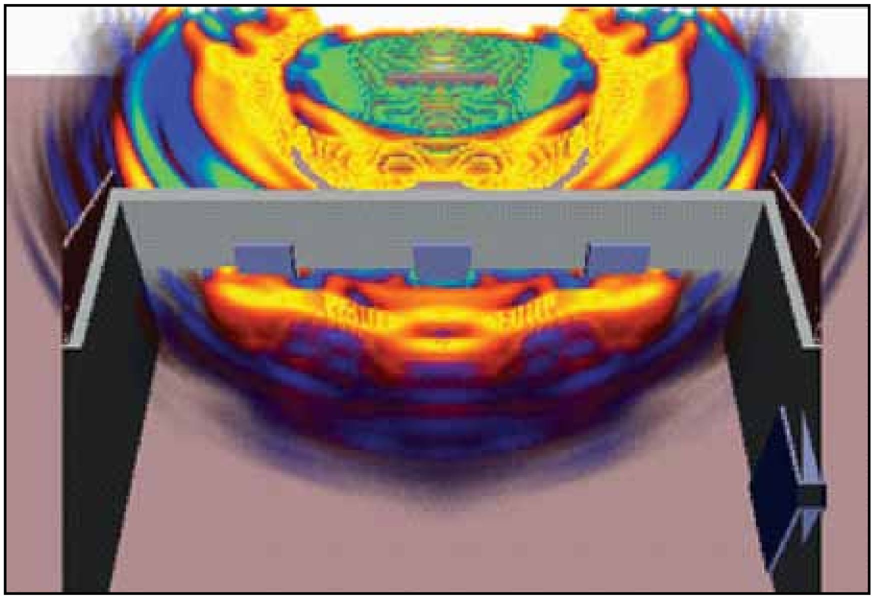

(b)

Figure 2. Modeling and Simulation Depicting (a) NSWCDD Test Facility and (b) Simulation of Radiated RF

simulation graphic depicting the RF emitted by an HPM dipole antenna.

Potential platforms for HPM integration include: man-portable, aerial, vehicle, and vessel-mounted systems. These platforms all provide unique methods for delivery of HPM sources. For example, aerial delivery—which, in many ways, is the most challenging due to size and weight constraints—can increase the effective range of these systems and can engage multiple targets at close range without endangering personnel. Likewise, vehicles and vessel-mounted HPM systems provide a way for law enforcement and the military to stop vehicles in chase scenarios almost as soon as they begin. The goal of all of these projects is to provide military forces with the ability to employ nonkinetic, electronic strike technologies against an adversary's electronics.

The DEWO and Directed Energy Division are uniquely positioned to provide numerous capabilities for in-house development while engaging with the private sector to test and provide feedback on HPM systems developed externally. In the past decade, NSWCDD has evaluated several HPM systems at Dahlgren to determine their effectiveness against various electronic targets while maintaining the Office of the Secretary of Defense's Tri-Service RF Directed Energy Weapon (DEW) Database. This database contains all effects data collected from directed-energy tests performed within the U.S. Air Force, Army, and Navy.

Conclusion

NSWCDD continues to pioneer HPM source development and lethality and integration studies, leading to the demonstration and delivery of prototype capabilities. It also is committed to researching and developing critical subsystems for HPM delivery. By leveraging numerous target assets and sophisticated diagnostic equipment—in conjunction with MOATS—NSWCDD has positioned itself at the forefront of HPM electronic attack, leading the way in the development and delivery of these capabilities to the warfighter.

Acknowledgment

Nancy Muncie, Bowhead, contributed to this article.

Laser Counter Rocket, Artillery, and Mortar (C-RAM) Efforts

By Michael Libeau

Mortars and rockets are common weapons confronting U.S. troops abroad. Insurgents fire the inexpensive projectiles into populated areas, intending to kill or injure service members and to inflict physical damage. While kinetic solutions like guns and missile interceptors are used to counter rockets and mortars, laser counter rocket, artillery, and mortar (C-RAM) systems present a promising solution to counter these challenging threats in the near future.

Scientists and engineers at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) have been researching, developing, testing, and evaluating laser C-RAM systems through collaboration, modeling and simulation, and experimentation. The Joint Technology Office (JTO) and the Directed Energy and Electric Weapons Program Office (PMS 405) sponsored the first year of these initiatives in 2007. Consecutive and current work has been sponsored by the Office of Naval Research (ONR) Expeditionary Maneuver Warfare and Combating Terrorism S&T Department.

Background

In preparation for the development of a laser C-RAM system, an understanding of the vulnerability of rockets and mortars to laser energy was crucial. Engineers from NSWCDD and the U.S. Army Space and Missile Defense Command (SMDC) collaborated on laser C-RAM efforts. Engineers analyzed the RAM threat and examined a variety of targets, accessing RAM vulnerabilities to laser energy by utilizing theoretical, numerical, and experimental work. They then developed theoretical models that captured the physics of the laser-induced failures of targets containing high explosives (HE). Additionally, NSWCDD engineers enhanced lethality simulations using a tool called the Effectiveness Toolbox to model engagements of RAM targets with laser energy. Figure 1 shows a screen capture from the Effectiveness Toolbox.

The resulting simulations included results from a laser atmospheric propagation model and a thermal model to determine the effect of the laser energy on the target. The simulations also incorporated target trajectories necessary for modeling the changing laser conditions on the target resulting from the engagement of a ballistic target. Subsequent to modeling these effects, live testing was performed. Figure 2 shows the lasing and destruction of a RAM target during live testing.

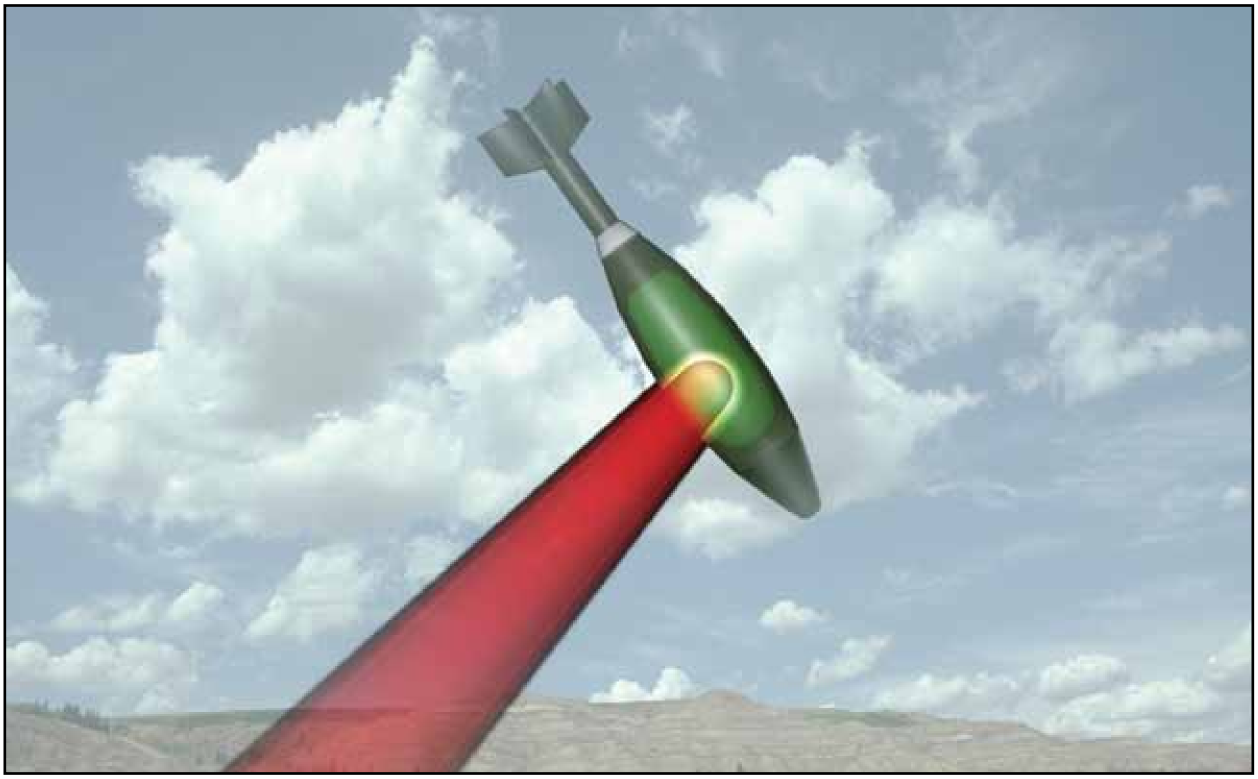

Figure 1. Screen Capture from the Effectiveness Toolbox Showing the Laser Engagement of a Mortar Target

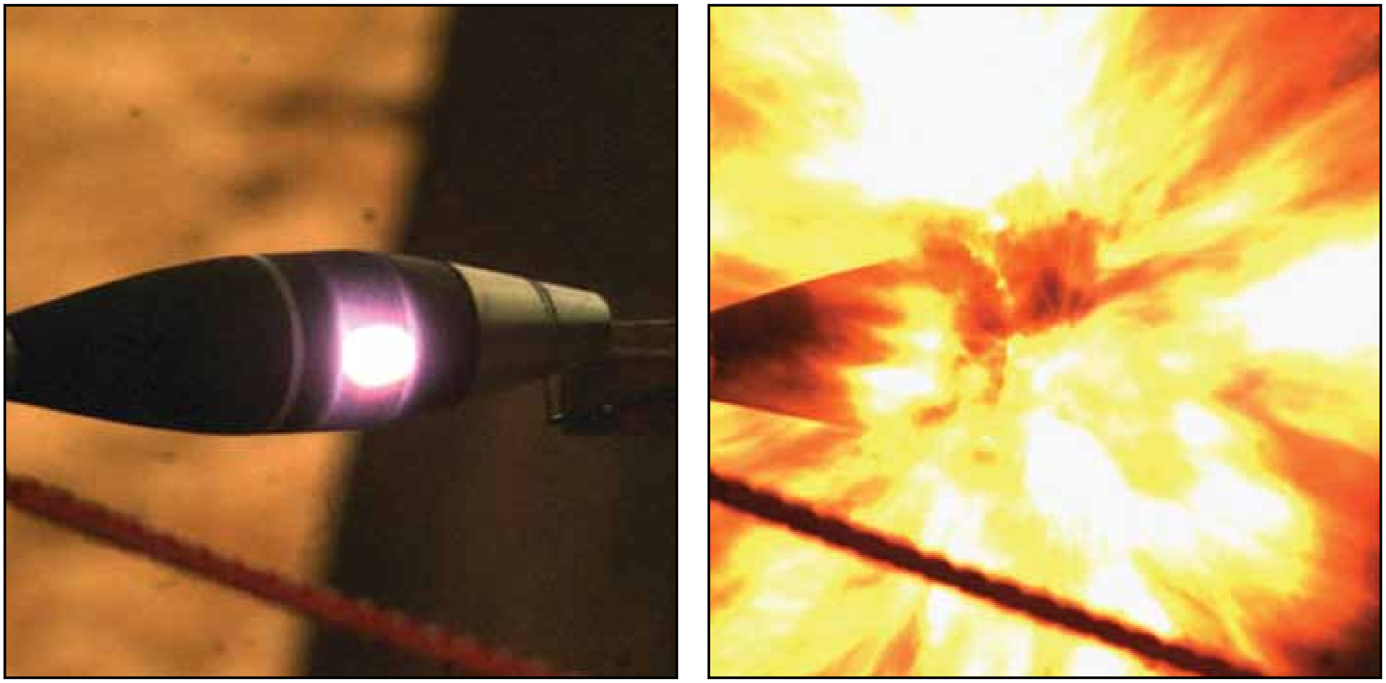

Figure 2. Explosive Target is Destroyed with NSWCDD's Fiber Lasers

NSWCDD engineers conducted two large experimental tests to determine the vulnerability of HE targets to laser energy using NSWCDD's High-Energy Fiber lasers. The first test was conducted jointly with SMDC. During these two tests, over 40 RAM targets were destroyed under different laser conditions, producing significant information on laser lethality. Researchers measured the failure times of multiple targets for different laser powers, spot sizes, incidence angles, and aimpoints. The experimental data yielded by the tests increased engineers' understanding of the vulnerability of targets containing energetic materials. This data was then used to benchmark predictive models.

Future tests are planned with additional HE targets to further the knowledge of RAM vulnerability. These tests are controlled and conducted carefully to ensure that good data is obtained. Accurate measurements of laser power on the target and the resulting target failure times must be made during the tests. To that end, NSWCDD engineers leverage Division-wide expertise in lasers and optics with its long history of explosives testing to achieve meaningful test results. NSWCDD personnel have been instrumental in improving techniques to measure the spatial profile of laser power on a target. The spatial distribution of laser power on a target is critical to understanding the target's failure. Figure 3 shows a laser beam's spatial power distribution measured during a test.

ONGOING LASER C-RAM INITIATIVES

Recent advances in fiber lasers have increased the power outputs of these rugged, solid-state devices. Both government and contractor efforts are examining the application of commercial off-the-shelf (COTS) lasers and other electric lasers for application into advanced weapon systems. The Navy's Laser Weapon System (LaWS) Program, for example, is examining a laser system built around efficient fiber lasers. This is significant because a high-energy fiber laser system offers two critical advantages over gun and missile interceptor C-RAM systems. First, the laser has a great depth of magazine since it requires only electricity for operation. Consequently, unlike a gun system, which has a limited supply of ammunition, a laser system is limited only by its supply of electrical energy. Second, a laser system offers a cost per kill that is significantly lower than alternative systems because only electricity is being expended instead of gun ammunition or a costly missile interceptor. This low cost per kill also better matches the low cost of the RAM target being engaged.

High-Energy Fiber Laser C-RAM systems will provide significant advantages in defeating the RAM threat while augmenting existing C-RAM solutions. More importantly, laser C-RAM systems will help protect members of the armed forces from the inexpensive, yet often deadly threats posed by rockets and mortars.

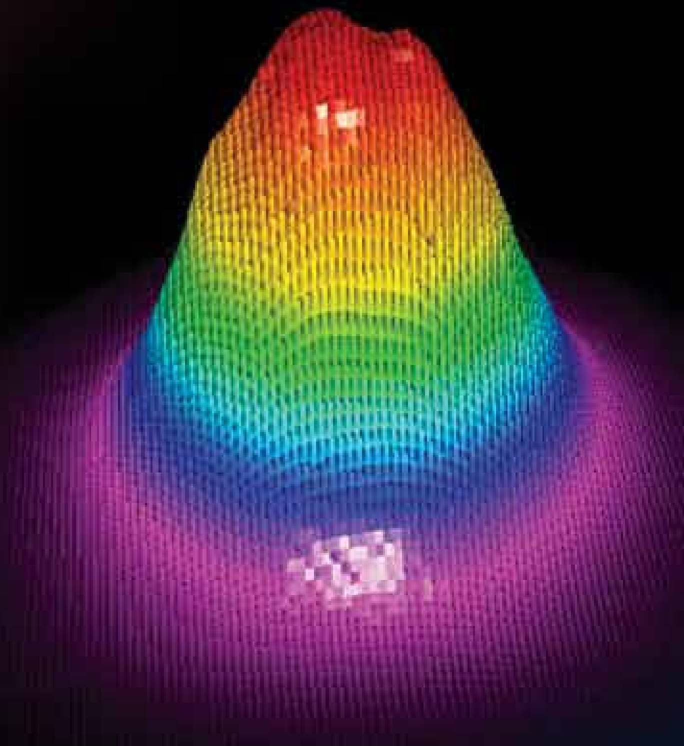

Figure 3. Laser Power Spatial Variation from a C-RAM Test

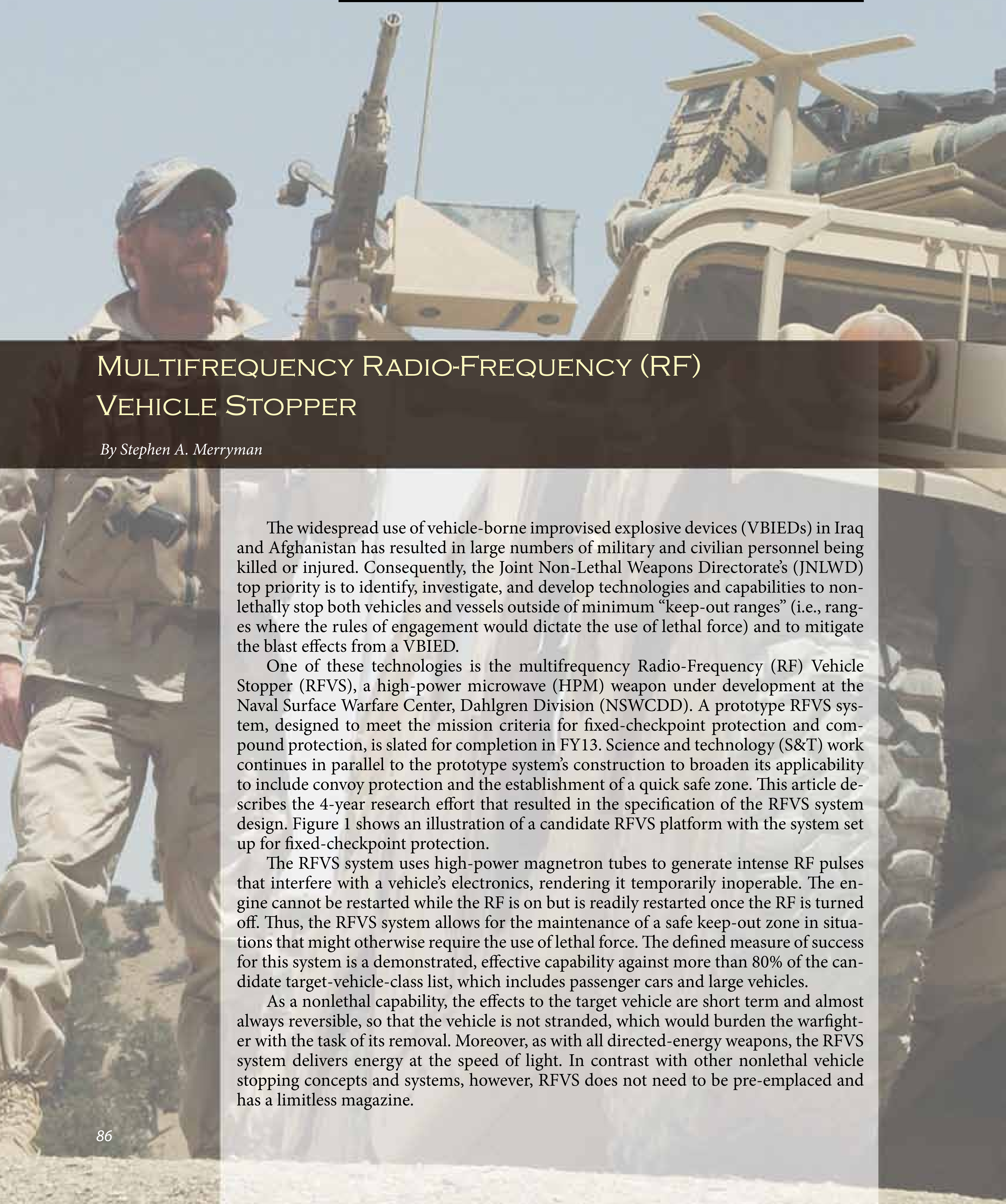

MULTIFREQUENCY RADIO-FREQUENCY (RF) VEHICLE STOPPER

By Stephen A. Merryman

The widespread use of vehicle-borne improvised explosive devices (VBIEDs) in Iraq and Afghanistan has resulted in large numbers of military and civilian personnel being killed or injured. Consequently, the Joint Non-Lethal Weapons Directorate's (JNLWD) top priority is to identify, investigate, and develop technologies and capabilities to non-lethally stop both vehicles and vessels outside of minimum "keep-out ranges" (i.e., ranges where the rules of engagement would dictate the use of lethal force) and to mitigate the blast effects from a VBIED.

One of these technologies is the multifrequency Radio-Frequency (RF) Vehicle Stopper (RFVS), a high-power microwave (HPM) weapon under development at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD). A prototype RFVS system, designed to meet the mission criteria for fixed-checkpoint protection and compound protection, is slated for completion in FY13. Science and technology (S&T) work continues in parallel to the prototype system's construction to broaden its applicability to include convoy protection and the establishment of a quick safe zone. This article describes the 4-year research effort that resulted in the specification of the RFVS system design. Figure 1 shows an illustration of a candidate RFVS platform with the system set up for fixed-checkpoint protection.

The RFVS system uses high-power magnetron tubes to generate intense RF pulses that interfere with a vehicle's electronics, rendering it temporarily inoperable. The engine cannot be restarted while the RF is on but is readily restarted once the RF is turned off. Thus, the RFVS system allows for the maintenance of a safe keep-out zone in situations that might otherwise require the use of lethal force. The defined measure of success for this system is a demonstrated, effective capability against more than 80% of the candidate target-vehicle-class list, which includes passenger cars and large vehicles.

As a nonlethal capability, the effects to the target vehicle are short term and almost always reversible, so that the vehicle is not stranded, which would burden the warfighter with the task of its removal. Moreover, as with all directed-energy weapons, the RFVS system delivers energy at the speed of light. In contrast with other nonlethal vehicle stopping concepts and systems, however, RFVS does not need to be pre-emplaced and has a limitless magazine.

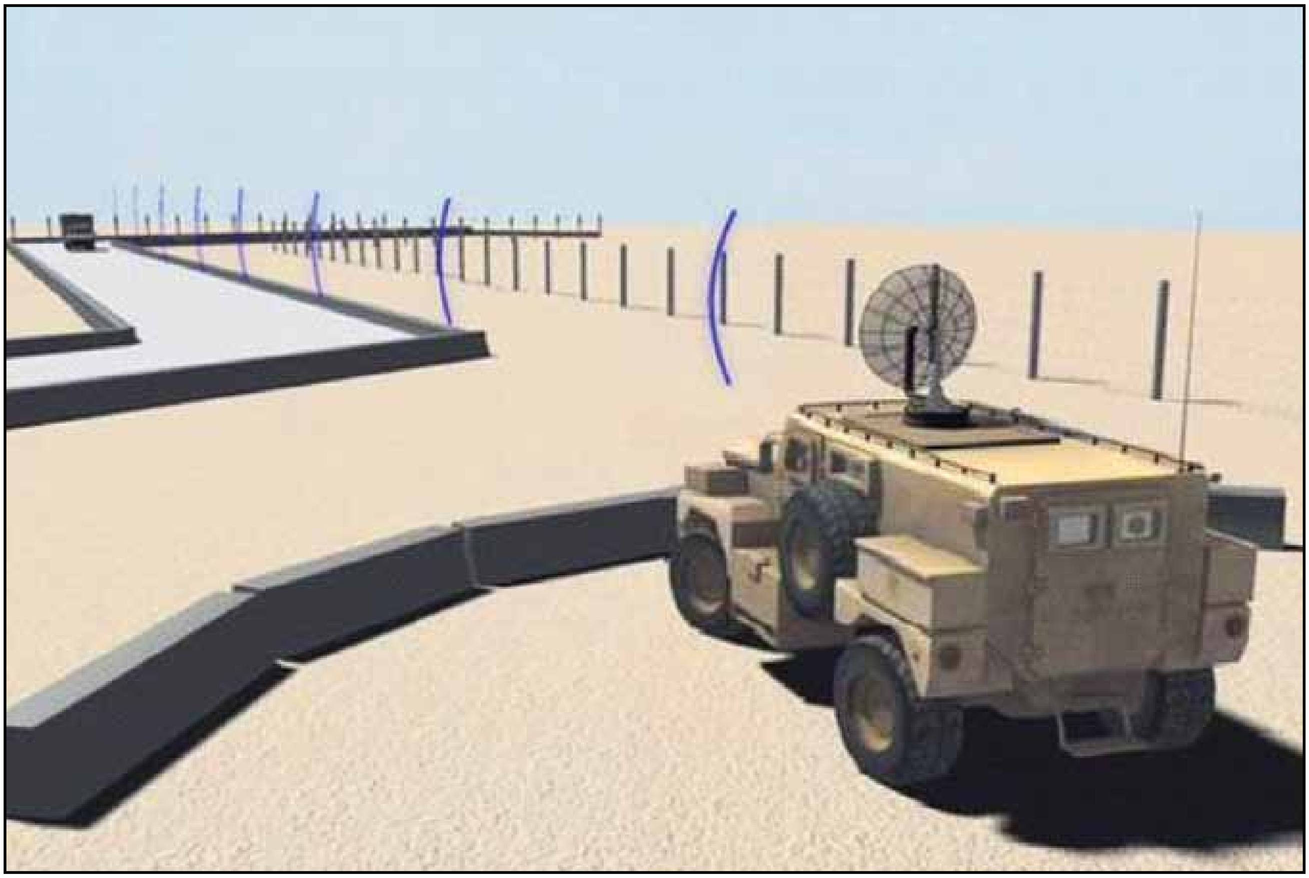

Figure 1. Illustration of Candidate RFVS System Setup for Checkpoint Protection

BACKGROUND

Using HPM or RF energy to stop an automobile engine is not a new concept; it has been under investigation for some time in private, academic, and military sectors. To that end, the RFVS program leveraged as much historic work as possible while collaborating with academic and military laboratories and while aggressively pursuing contacts in the automobile industry to gain knowledge of vehicle electronic design and function.

In 2005, the JNLWD funded the then-Directed Energy Technology Office (DETO) at NSWCDD to perform an extensive reverberation chamber test series to characterize the vulnerability of a representative cross section of automobiles to a wide range of HPM source frequencies.[a] The purposes of the tests were twofold. First, the applicability of the Army's Ground Vehicle Stopper (GVS) data set needed to be established for newer vehicles, and second, a thorough, source-technology independent assessment of vehicle vulnerabilities needed to be performed. The rationale behind the latter was to establish vehicle vulnerabilities without inadvertently biasing the process. Only after the full assessment was performed would factors such as concept of operations (CONOPS) and system requirements come into play. Figure 2 is a photograph of reverberation chamber testing.

Over the past decade, a significant number of private, academic, and military laboratories have investigated the susceptibility of automobiles to HPM energy. The range in approaches spans the gamut from isolated component testing, through direct injection and radiated testing of electronic control units (ECUs), and continuing through full vehicle radiated testing. Each of the different test methods has its strengths and weaknesses. Testing of isolated ECUs in controlled laboratory conditions is arguably the best way to determine exactly how a specific unit is responding to the RF. However, whether or not the identified susceptibilities continue to hold true when the unit is in place in a

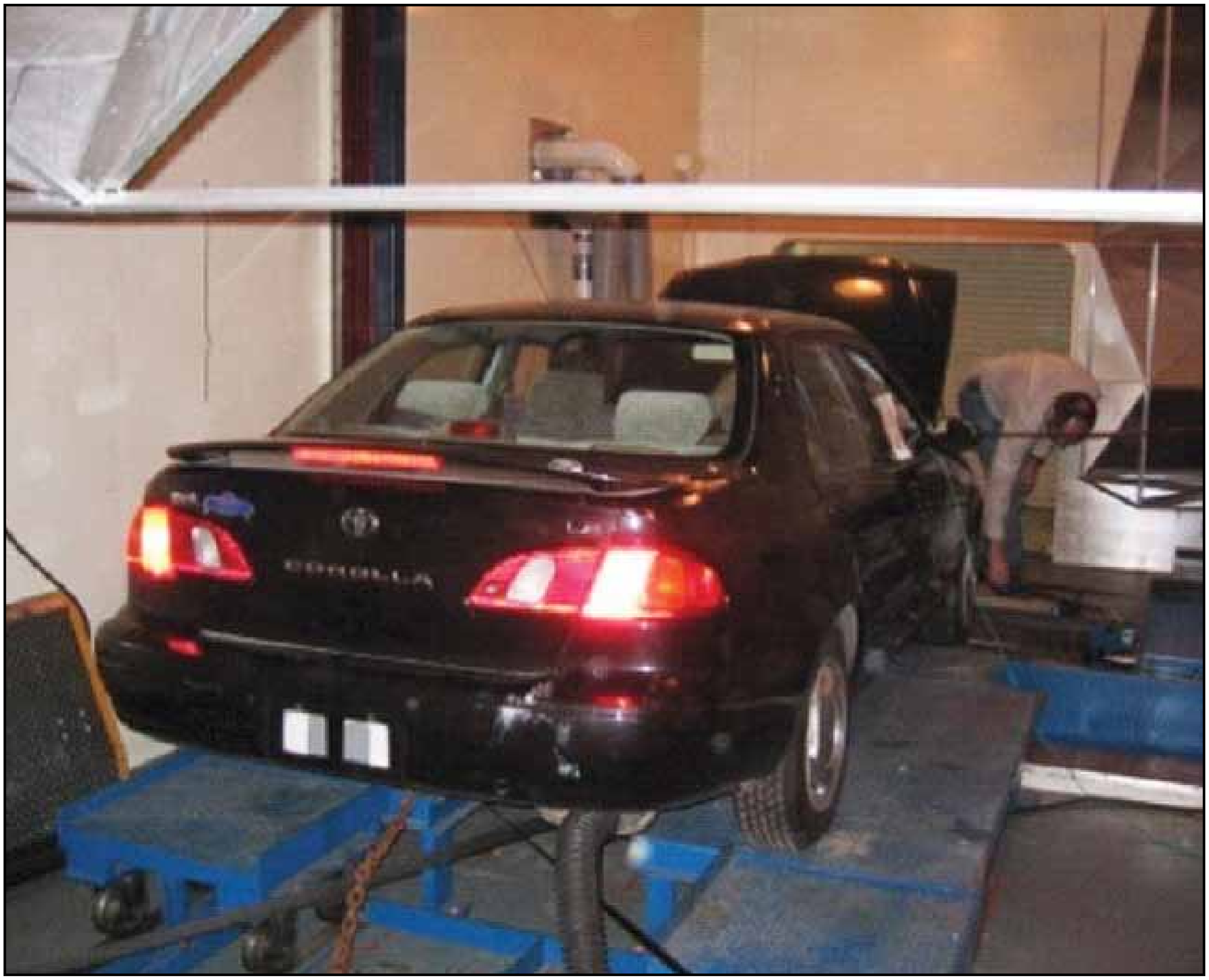

Figure 2. Vehicle on a Dynamometer in the NSWCDD Reverberation Chamber

vehicle, or whether the results apply to other vehicles' ECUs, remains a significant question that limits the applicability of the results.

Full vehicle testing and failure analysis of the ECU can be a daunting task. That said, full vehicle testing affords the advantage of ensuring that the response is commensurate with expectations of genuine engagements. The test approach one chooses to take depends upon resources, test facility availability, and most importantly, the objectives of the program. For the RFVS program, the objectives were to identify an HPM waveform that is effective against a broad class of the candidate target vehicles and to ensure that the identified waveform could be generated with a source that can be packaged in a footprint and cost amenable to military users. To meet the program's objectives, the RFVS program chose to invest the majority of its resources in full vehicle testing. While the focus of the effects testing portion of the RFVS program has remained on full vehicle testing, both time and resources have been devoted to fostering and maintaining connections with academia and the auto industry. There is concerted effort to keep abreast of the latest trends in automotive technology, to ensure that the current RFVS system design will continue to be effective against future vehicle designs, and to leverage all research that might aid in future RFVS designs.

SYSTEM OPERATION

The majority of current HPM system concepts employ a narrowband, single-frequency HPM source. In contrast, RFVS utilizes multiple HPM frequencies. The rationale for using multiple frequencies is associated with increased system effectiveness. Electromagnetic (EM) energy can be used to disrupt or damage an electronic target. In order for the energy to affect the electronics, however, it must be able to reach a critical component(s) inside the target. This involves a process referred to as coupling. Different EM waveforms are more or less effective against specific targets depending, in part, on their frequency, as different frequencies couple better or worse depending on varying target geometries. To be specific, each piece of electronics has specific resonance frequencies that most effectively facilitate coupling energy to the target. Unfortunately, these resonant frequencies can be unique to each piece of equipment. Consequently, a single-frequency waveform might be very effective against one target, but less effective against another target. Therefore, a system that utilizes either a sweep of frequencies or multiple frequencies will be more effective against a larger target set. This is not a novel idea, but rather one that has been readily acknowledged within the HPM community for some time and fervently embraced by the RFVS program. Current technology limitations prohibit high-power-swept frequency sources as viable options, leading to the idea of a multifrequency source. The more frequencies that are used, the more effective the system; however, a trade-off is made with system size and cost as the number of source frequencies is increased.

Brassboard System

After completion of the exhaustive vehicle effects characterization testing in 2006, the RFVS program identified the optimal number of frequencies needed to meet mission requirements. It then used this information in the design and construction of the Brassboard System. The purpose in constructing the Brassboard System was to demonstrate the benefit of the multifrequency approach and the ability to meet mission objectives with specified power on target requirements. Construction of the RFVS Brassboard System began in 2007 and was completed in 2008. The Brassboard System was not constructed with specific system footprints in mind. Thus, the antenna and conex used are significantly larger than those in the prototype design. The RFVS team collaborated with a Marine Corps service representative identified by the JNLWD to flesh out the specifics of the mock checkpoint to be used in the RFVS Brassboard System Demonstration. Figure 3 provides a diagram of the checkpoint setup used in the RFVS Brassboard Demonstration. Figure 4 provides photographs of the RFVS Brassboard System.

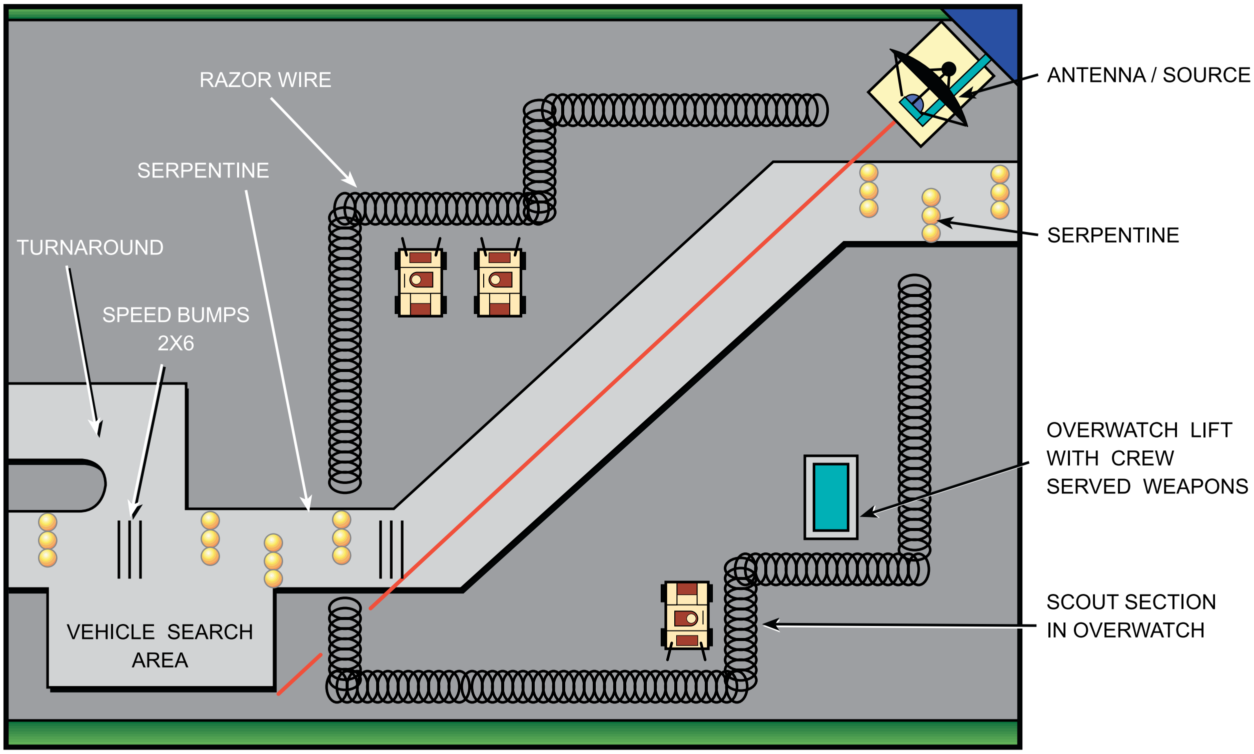

Figure 3. Schematic of Checkpoint Setup

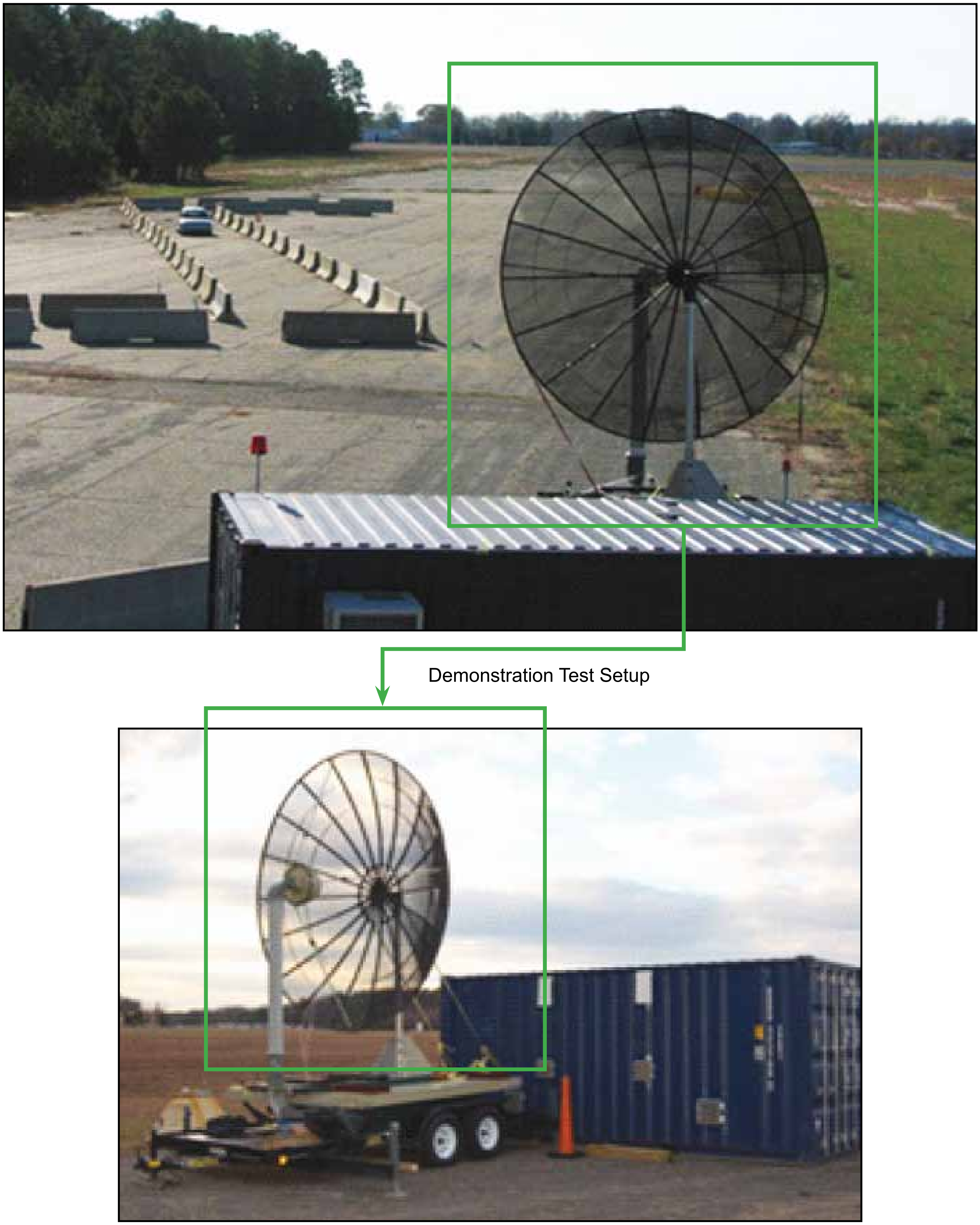
Demonstration Test Setup

Figure 4. Photograph of the RFVS Brassboard System and Demonstration Setup

The Brassboard System Demonstration was conducted in Spring 2008. The Demonstration was a success, and funding for the RFVS prototype was consequently approved. To date, 42 passenger vehicles (cars, pickup trucks, vans, and sport utility vehicles (SUVs)) and 3 large trucks (dump truck and tractors) have been tested as part of the RFVS program.

Way Ahead

The JNLWD continues to work with the Directed Energy Warfare Office (DEWO) toward the development of a fieldable multifrequency RFVS system. Once the capability is fully developed, tested, and certified ready for operational use, warfighters and civilians alike will benefit greatly. Lives will no doubt be saved using the ability to stop vehicles nonlethally and mitigate the blast effects from VBIEDs.

Acknowledgment

Dr. Cynthia Ropiak (SAQ Consulting) contributed to this article.

Endnote

a. The Directed Energy Technology Office (DETO) was renamed the Directed Energy Warfare Office (DEWO) in August 2009. For reference, see the charter for the DEWO, NSWCDD, 17 August 2009.

HIGH-POWER ELECTRICAL VEHICLE-STOPPING SYSTEMS

By Jordan Chaparro and Melanie Everton

The military needs devices that can safely and reliably stop or arrest vehicles. The primary concern is security at entry control points and vehicle check points similar to the one shown in Figure 1. In such scenarios, it is desirable to be able to stop unauthorized vehicles at predefined standoff ranges to protect personnel, equipment, and critical infrastructure.

Both the military and civilian law enforcement agencies face similar issues with chase scenarios, where concerns over bringing an offending vehicle to a stop without killing or injuring innocent civilians, or causing collateral damage, often prolongs high-speed pursuits. That said, currently employed nonlethal options for arresting vehicles have significant logistical limitations and carry a high cost per use.

The Naval Surface Warfare Center, Dahlgren Division's Directed Energy Warfare Office (DEWO), under the sponsorship of the Joint Non-Lethal Weapons Directorate (JNLWD), investigated compact systems designed to couple high-power electrical impulses to a target vehicle to stop its engine. Such systems are highly portable, can operate remotely, can be deployed quickly by a two-man team, and can engage hundreds of targets before requiring any significant maintenance.

SYSTEM OVERVIEW

Conceptually, electrical vehicle-stopping systems are fairly simple devices. The systems use several stages of energy compression to take a low-peak power source—like a battery pack—and create very intense, short-duration, oscillating electrical impulses. The block diagram, shown in Figure 2, illustrates the principal components of such a system.

A high-energy density, 300-V lithium battery pack, similar to what might be found in a hybrid vehicle, serves as the prime power source for the device. These batteries are capable of driving the system for hundreds of engagements before requiring recharge.

The direct current bus from the batteries is stepped up to several kilovolts in order to charge a capacitive voltage multiplier, such as a Marx Generator, Spiral Line Generator, or Tesla Transformer. Once triggered, these generators charge a resonant circuit to hundreds of kilovolts which, when switched, generate the desired oscillating waveform. Coupling this electrical pulse to a target may be accomplished by direct electrode contact, by radiating the waveform from a broadband antenna structure, or by a combination of both methods.

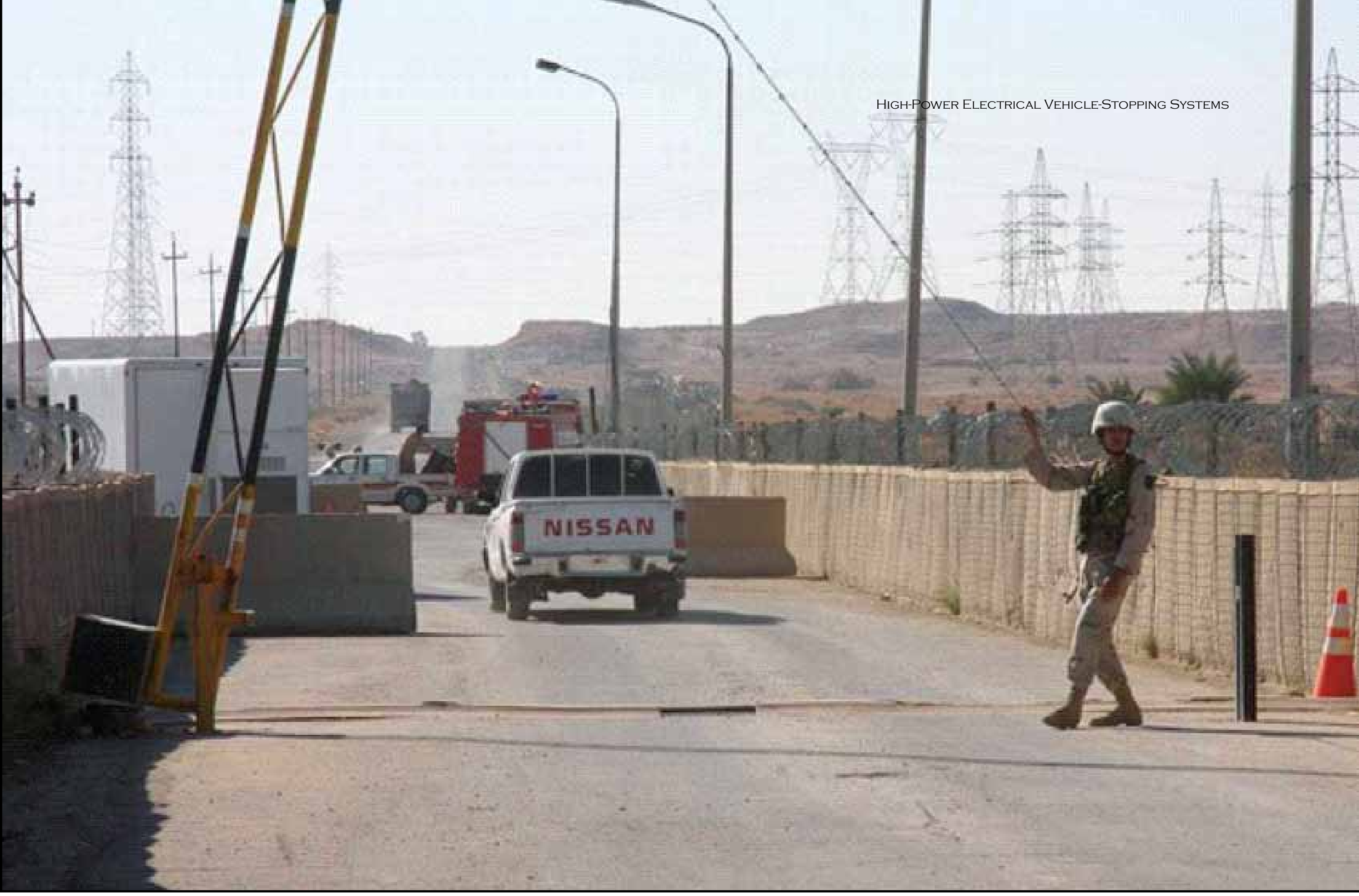

Figure 1. An Azerbaijani Soldier Guarding Entry Control Point 1 at the Haditha Dam in Support of Operation Iraqi Freedom

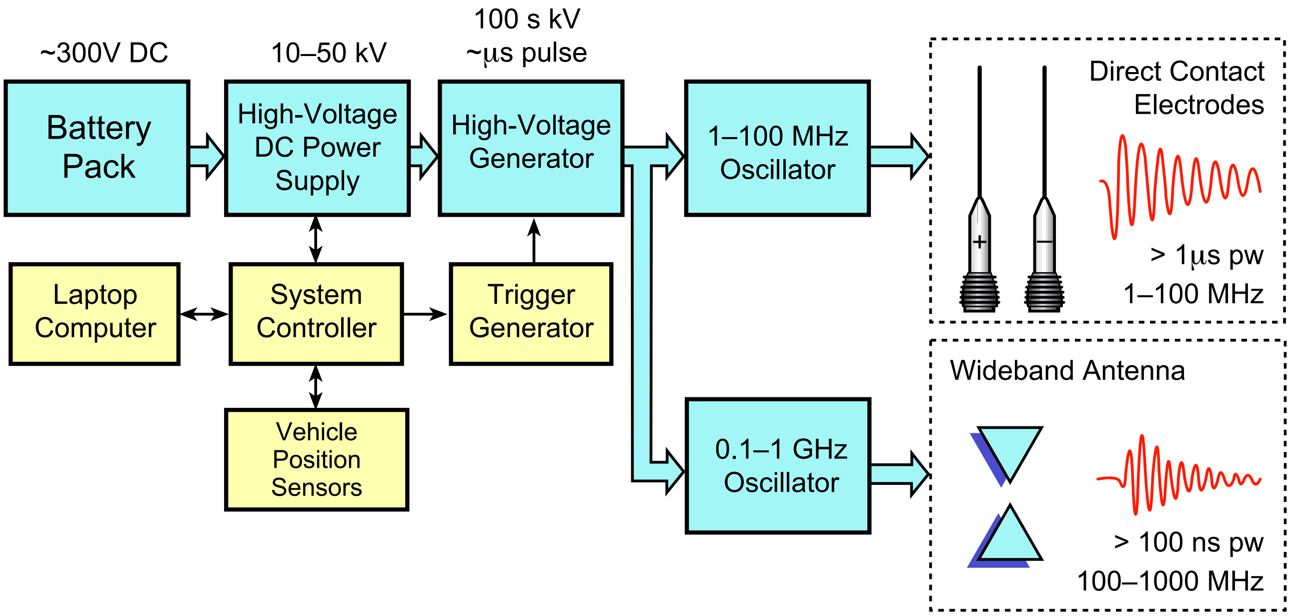
~300V DC
10–50 kV
100 s kV
~µs pulse
Battery Pack
High-Voltage DC Power Supply
High-Voltage Generator
1–100 MHz Oscillator
Direct Contact Electrodes
+
–
> 1µs pw
1–100 MHz
Laptop Computer
System Controller
Trigger Generator
Vehicle Position Sensors
0.1–1 GHz Oscillator
Wideband Antenna
> 100 ns pw
100–1000 MHz

Figure 2. System Block Diagram for Generic Electrical Vehicle-Stopping Systems

The system is monitored and controlled by an integrated system computer. A laptop computer, remotely connected to the system controller through either fiber optic or a wireless network, can be used to arm the system. At this point, motion detection sensors trigger the pulse train upon the targeted vehicle. The laptop can also be used to monitor the system's status, change system parameters, and receive data collected during the last engagement event.

A conceptual rendering of how such a system might look when in use is shown in Figure 3. Traffic would be funneled with barriers to a single lane. When not engaged, the system electrodes would sit flush with the roadway unit, with an exposed height of less than 3 inches. When required, the electrodes could be released to make contact with a vehicle's undercarriage and deliver the electrical impulses.

COMPARISON WITH EXISTING SYSTEMS

Tire spike systems are frequently employed but do not limit the momentum, drive, or control of a vehicle to an extent that could be useful in any type of control or checkpoint scenario. Consequently, while tire spike systems are primarily used in high-speed pursuit applications, they are limited, in that they cripple the target just enough to allow law enforcement to force the vehicle to a stop.

Restraining nets are most comparable to electrical vehicle stoppers with respect to their intended application and desired effect. Restraining net systems and electrical vehicle stoppers both completely arrest vehicles, although by different means. Restraining nets bind the front axle of the vehicle, causing it to forcibly lose momentum and skid to a stop. Thus, the vehicle operator loses the ability to steer the vehicle, further resulting in a lower potential for collateral damage. Electrical systems stop the engine of the vehicle, leaving the operator with control for the duration of the vehicle's momentum. Physical barrier structures can then be employed to force an affected vehicle to stop in a fairly short distance. Modern vehicles lose power steering when the engine is cut off, such that the maneuverability of the vehicle is limited enough to

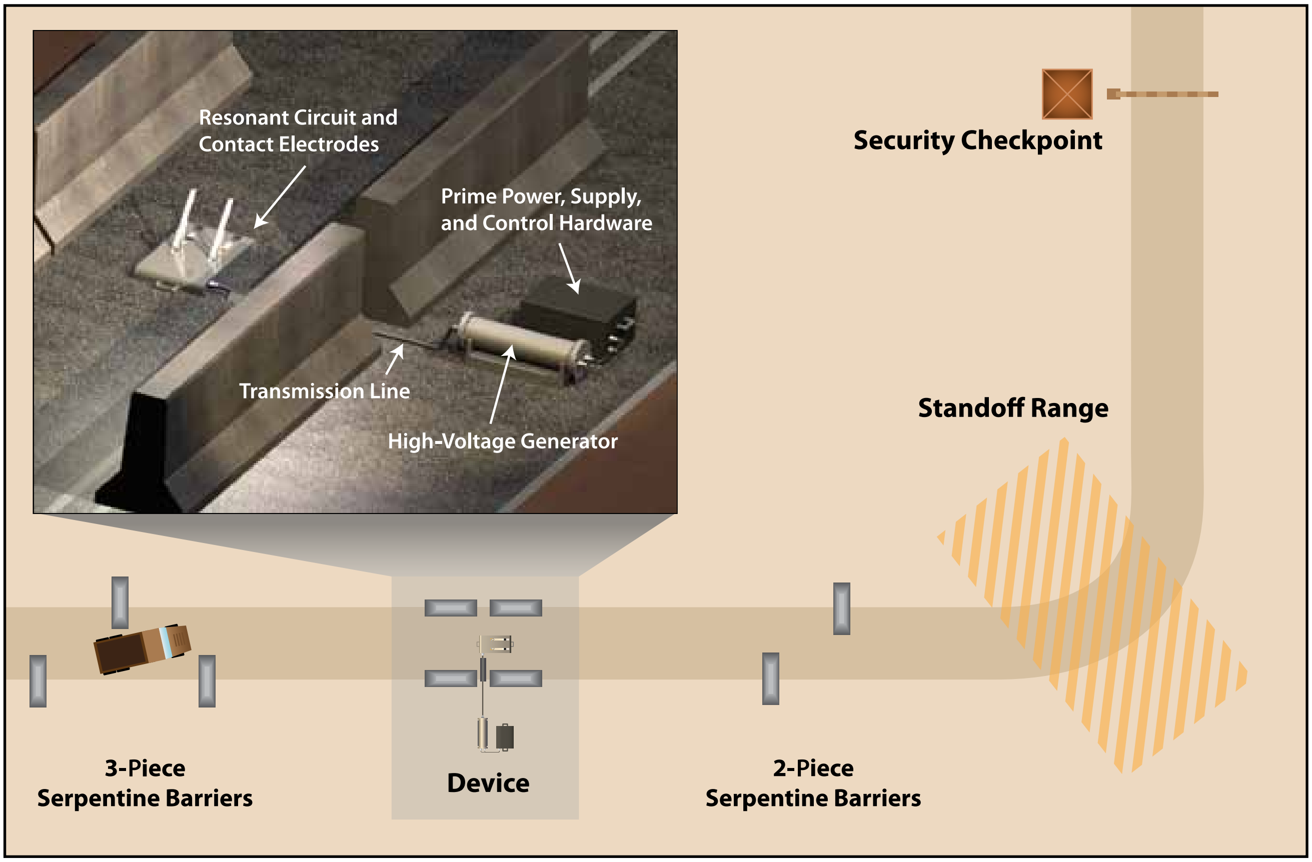
Resonant Circuit and Contact Electrodes
Prime Power, Supply, and Control Hardware
Transmission Line
High-Voltage Generator
Security Checkpoint
Standoff Range
3-Piece Serpentine Barriers
Device
2-Piece Serpentine Barriers

Figure 3. Conceptual Rendering of an Employed Electric Vehicle Stopping System

allow normally nonrestrictive serpentines to be effective at limiting roll-off distances.

One key logistical advantage of electrical vehicle stoppers, compared to restraining nets, is the average cost per engagement. Restraining net systems are one-time use devices that cost several thousand dollars each. Electrical systems initially cost tens of thousands of dollars but can perform thousands of stops within the expected lifetime of the device. Also, there is no requirement to physically reset or reload an electrical system, as with restraining nets. The maintenance required for electrical systems involves the occasional replacement of electrode arms and the inspection of the system connections and pressure levels.

Operationally, both systems have limitations on the types of targets that can be effectively stopped. Restraining nets are limited by vehicle momentum, which can be a product of high speeds or large vehicles. Electrical systems are not limited by vehicle size or speed, but they require additional support from structures—such as serpentines or speed bumps—to force the target to brake and dissipate its momentum once the engine has been stopped.

Both devices typically cause damage to targeted, stopped vehicles. Restraining nets almost always cause tire damage. Less commonly, brake lines, front axles, wheels, and transmissions also might be damaged. Electrical systems typically damage engine controllers, security modules, and engine sensors. In addition, noncritical parts—such as gauges, radios, and cabin fans—also might be damaged. Moreover, moving affected targets is much less of an issue with electrical stoppers than vehicles stopped by net systems, which must first have the net cut away and freed before the target is moved to the side of the roadway.

System Refinement and Look Forward

Previous attempts to field electrical vehicle-stopping systems have been hampered by limited success rates on a large population of vehicles. Many models of vehicles are easily affected by any type of large injected current, while others are fairly resistant. Through carefully designed and controlled experiments, and logistical regression modeling techniques, the DEWO team has been able to determine key waveform attributes that scale with stopping effectiveness rates on a representative population of vehicles. Successful stop rates exceeding 90 percent have been achieved on a diverse vehicle test set by engineering system resonators to enhance system performance The DEWO team, through continued research, testing, and evaluation, is continuing its work to increase the reliability and effectiveness of these systems to make them more compact and to improve their functionality for future military and law enforcement applications.

NONLETHAL SMALL-VESSEL STOPPING WITH HIGH-POWER MICROWAVE TECHNOLOGY

By Jacob Walker

The employment of small vessels to attack merchant ships and other seafaring units has emerged as a significant threat to international navigation and safe operations on the high seas. Along with swarm tactics, small vessels have been known to carry improvised explosive devices, help smuggle terrorists and weapons, and serve as attack platforms on the water for larger weapons. While kinetic solutions serve as the decisive option, alternative solutions that employ nonlethal means are being explored. A depiction of a swarm of small vessels ready to attack is shown in Figure 1.

The Naval Surface Warfare Center, Dahlgren Division's (NSWCDD's) Directed Energy Warfare Office (DEWO) is evaluating directed-energy (DE) concepts based on high-power microwave (HPM) technology for nonlethal vessel-stopping applications. Nonlethal weapons are defined by the Department of Defense (DoD) as weapons that are explicitly designed and primarily employed so as to incapacitate personnel or materiel while minimizing fatalities, permanent injury to personnel, and undesired damage to property and the environment.[1]

Several methodologies exist for using nonlethal means to stop small vessels. They include:

- Running-gear or prop entanglement systems
- Exhaust stack blockers
- A sea-anchor vessel-stopping system, which casts a net across the bow of a vessel to impart resistance
- Small-craft disablers, which insert a spear into the hull and deploy a fin that drags in the water, making steering impossible

Prop entanglement systems, exhaust stack blockers, and sea-anchor systems are useful and effective, but all are operationally difficult to deliver when deployment methods rely on positioning them in front of, or directly over, a vessel moving

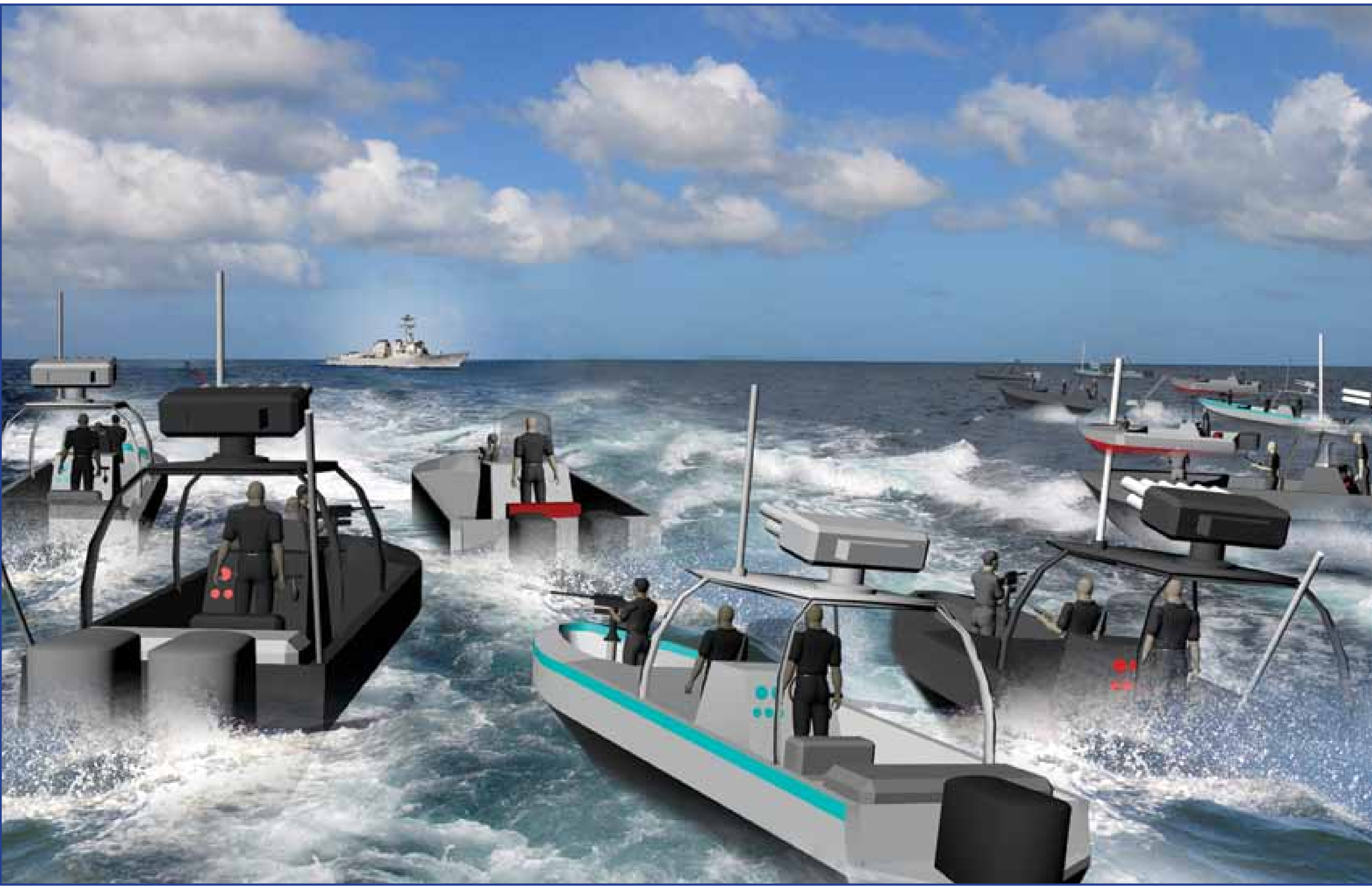

Figure 1. Depiction of a Small-Vessel Swarm Ready to Attack

at high speeds. Small-craft disablers also are a formidable vessel-stopping solution and may be easier to deploy, but they cause permanent damage to the vessel in question.

Under the direction of the Joint Non-Lethal Weapons Directorate (JNLWD), the DEWO is in the initial stages of a multiyear effort to evaluate DE concepts for nonlethal vessel-stopping applications. It is currently focusing on HPM technology. This technology uses HPM sources to radiate radio frequency (RF) pulses downrange to interfere with motor-control electronics and significantly impede or stop small-vessel motors with minimal collateral damage. These RF pulses can be generated using different technologies ranging from wideband LC oscillators and microwave tubes (e.g., magnetrons, klystrons, and backward wave oscillators) to emerging solid-state technologies (e.g., nonlinear transmission line and photo-conductive switching). An outboard motor on a test stand is shown in Figure 2.

In comparison to kinetic weapons or other nonlethal systems, HPM avoids gross physical destruction to the vessel while, more importantly, providing zero-to-low risk of human injury. HPM accomplishes this at safe distances using speed of light delivery, therefore making evasion difficult, if not impossible, with the added benefit of scalable effects ranging from disruption to damage. Despite its numerous advantages, the use of HPM technology as a nonlethal weapon presents challenges as well, including a trade-off between system size and standoff range. This is particularly important when considering the use of HPM systems in different environments.

Upfront HPM source development costs represent one of the biggest challenges. However, long-term savings associated with HPM technology can offset this challenge. For example, prop entanglement systems might be deployed only once before they are rendered useless. HPM sources integrated onto a ship or other military vehicle can be employed in potentially thousands of missions, therefore resulting in a lower cost per single use, bringing overall associated costs of the system down significantly. Priorities for HPM nonlethal weapons include developing a system effective against different types of small vessels.

NSWCDD's Directed Energy Division began HPM susceptibility testing to determine the effectiveness of HPM weapons against relevant outboard engines. This involves testing small vessels in

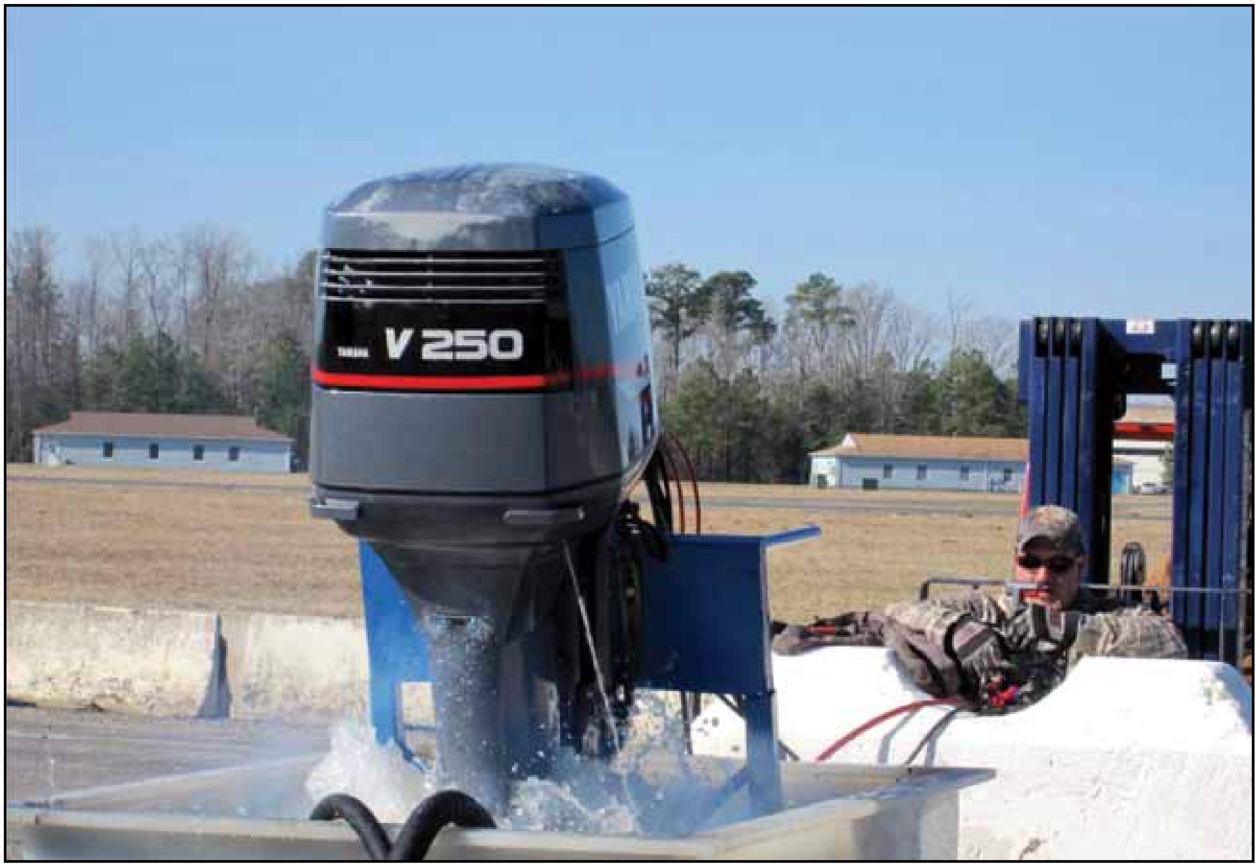

Figure 2. Outboard Motor Test Stand

a variety of environments, including reverberation and anechoic chambers, and open-air testing. All help identify different, effective waveform parameters such as frequency, pulse width, rise time, and required power or energy on target. They further facilitate the identification of design specifications necessary for an eventual HPM source. This source, once developed, will then be integrated into one of several potential platforms. Candidate concepts of deployment include U.S. Coast Guard and naval vessels in addition to unmanned surface or aerial vessels. Another potential application might be to supplement existing Coast Guard or Navy platforms used for fast-boat interdiction with an HPM vessel-stopping capability. A small-vessel test using an HPM source is shown in Figure 3.

Developing solutions for the growing threat that small vessels pose to navigation and safe operations in the world's oceans is one of JNLWD's top priorities. Using nonlethal HPM weapons to stop vessels will provide the warfighter with a viable option for swarm threat and fast-boat interdiction. DEWO is working diligently to accelerate this technology and provide a DE alternative to kinetic weapons and fulfill this long overdue capability gap.

REFERENCE

1. *Department of Defense Dictionary of Military and Associated Terms*, Joint Publication 1-02, 12 April 2001 (as amended through 19 August 2009).

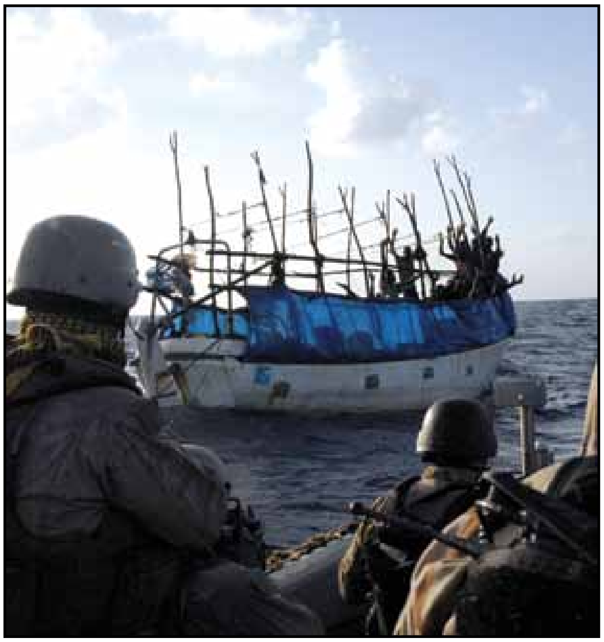

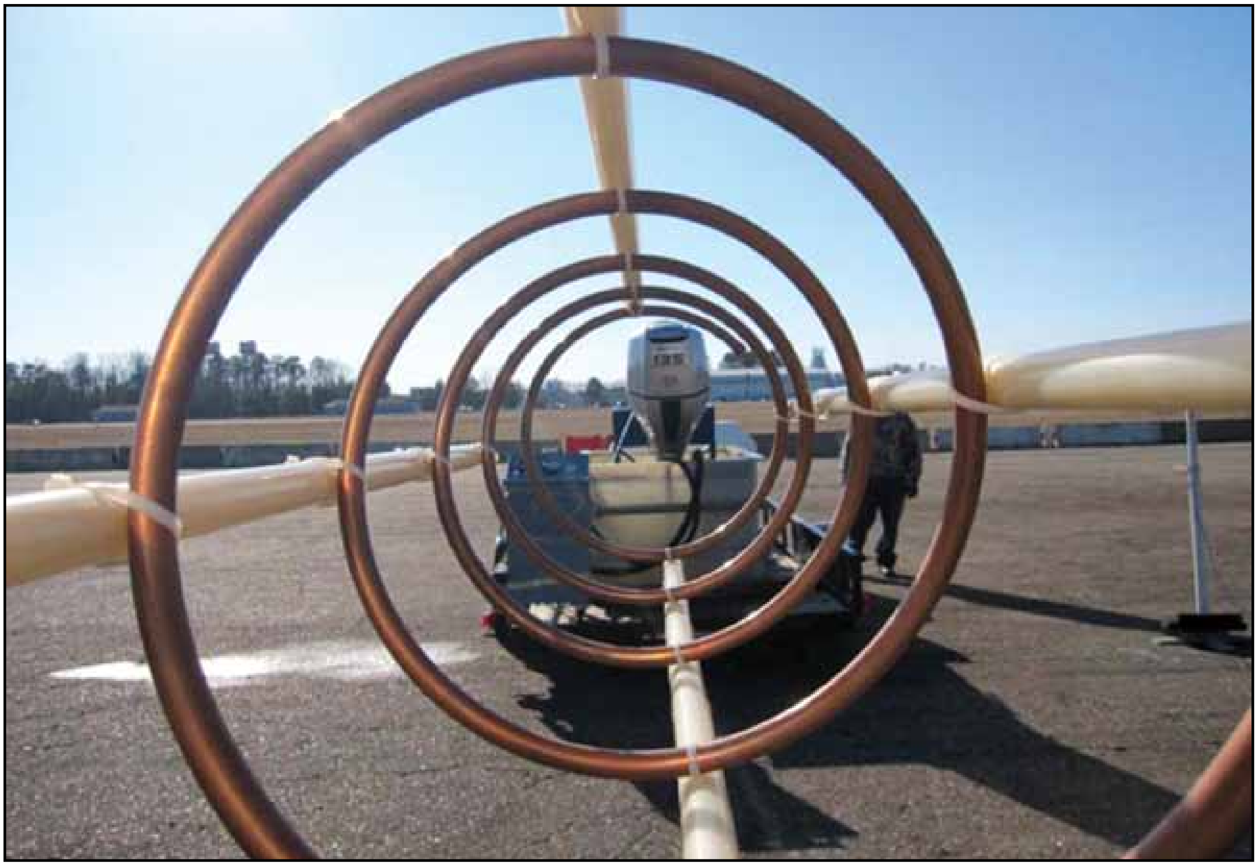

Figure 3. Small-Vessel Testing Using a High-Power Microwave Source

LEADING EDGE
Volume 7, Issue No. 4

Directed Energy

We look forward as we provide enterprise-wide technical and strategic leadership for the efficient and effective development, acquisition, and fielding of directed-energy systems for the warfighter.

Dale Sisson
Head, Electromagnetic and
Sensor Systems Department
NSWCDD, Dahlgren, Virginia

LEADING
EDGE
Volume 7, Issue No. 4

Fallen Warriors

Here we honor those who died while serving their country

NAVSEA
WARFARE CENTERS
DAHLGREN

NSWCDD/MP-09/34

Statement A: Approved for public release; distribution is unlimited

2013

NAWCWD TP 8347
Fourth Edition

Electronic Warfare and Radar Systems

Engineering Handbook

A Comprehensive Handbook for
Electronic Warfare and Radar Systems Engineers

E*WARFARE

Naval Air Warfare Center Weapons Division, Point Mugu, California

DISTRIBUTION STATEMENT A: Approved for public release.

NAWCWD TP 8347
Fourth Edition
NAVAIR

Electronic Warfare and Radar Systems Engineering Handbook

by
Avionics Department

OCTOBER 2013

NAVAL AIR WARFARE CENTER WEAPONS DIVISION
POINT MUGU, CA 93042

DEPARTMENT OF THE NAVY
UNITED STATES OF AMERICA

E*WARFARE

DISTRIBUTION STATEMENT A: Approved for public release.

Naval Air Warfare Center Weapons Division

FOREWORD

This handbook is designed to aid electronic warfare and radar systems engineers in making general estimations regarding capabilities of systems. This handbook is sponsored by the NAVAIR Director of Electronic Warfare/Combat Systems Department. This fourth edition updates technical information in Sections 3-7 and 3-8 from previous editions.

This document was reviewed for technical accuracy by Dr. Andrew Chen.

Approved by
DR. R. E. SMILEY, *NAVAIR Director*
EW/Combat Systems Department
30 September 2013

Under authority of
M. MORAN
RDML, U.S. Navy
Commander

Released for publication by
S. O'NEIL
Director for Research and Engineering

NAWCWD Technical Publication 8347

Published by..Technical Communication Office
Collation..Cover, 226 leaves, + Tabs
Previously published as NAWCWPNS TS 92-78 (1st through 3rd Edition) .....4,040 paper
First Edition ....105 paper
Second Edition ....500 paper
Third Edition..1,000 paper
Fourth Edition ....100 paper, 500 electronic

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE ***(DD-MM-YYYY)*** 01-10-2013	**2. REPORT TYPE** Technical publication	**3. DATES COVERED (From - To)** 1 January 2012 – 30 September 2013

4. TITLE AND SUBTITLE Electronic Warfare and Radar Systems Engineering Handbook (U)	**5a. CONTRACT NUMBER** N/A **5b. GRANT NUMBER** N/A **5c. PROGRAM ELEMENT NUMBER** N/A
6. AUTHOR(S) *Avionics Department*	**5d. PROJECT NUMBER** N/A **5e. TASK NUMBER** N/A **5f. WORK UNIT NUMBER** N/A
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Air Warfare Center Weapons Division 575 I Avenue, Suite 1 (Code 450000E) Point Mugu, California 93042	**8. PERFORMING ORGANIZATION REPORT NUMBER** NAWCWD TP 8347
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSOR/MONITOR'S ACRONYM(S)** N/A **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release.

13. SUPPLEMENTARY NOTES

None.

14. ABSTRACT

(U) This handbook is designed to aid electronic warfare and radar systems engineers in making general estimations regarding capabilities of systems. This handbook is sponsored by the NAVAIR Director of Electronic Warfare / Combat Systems Department.

15. SUBJECT TERMS

Aircraft Dynamics, Antenna, Data Transfer Busses, Electro-Optics and Infrared, Ethernet, Internet Protocol, Maneuverability, Microwave Components, Polarization, Radar Equations, Radiation Patterns, Receiver Characteristics, Receiver Sensitivity, Receiver Tests, Radio Frequency Components, RS-232 Interface, RS-422 Balanced Voltage Interface, RS-485 Interface, Transmission Control Protocol

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Timothy Stolsig
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	450	**19b. TELEPHONE NUMBER** *(include area code)* (805) 989-3452

CONTENTS

Section

ABBREVIATIONS AND ACRONYMS 1-1

FUNDAMENTALS
- Constants, Conversions, and Characters 2-1
- Mathematical Notation 2-2
- Frequency Spectrum 2-3
- Decibel (dB) 2-4
- Duty Cycle 2-5
- Doppler Shift 2-6
- Electronic Formulas 2-7
- Missile and Electronic Equipment Designations 2-8
- Radar Horizon / Line of Sight 2-9
- Propagation Time / Resolution 2-10
- Modulation 2-11
- Transforms / Wavelets 2-12

ANTENNAS
- Antenna Introduction / Basics 3-1
- Polarization 3-2
- Radiation Patterns 3-3
- Frequency / Phase Effects of Antennas 3-4
- Antenna Near Field 3-5
- Radiation Hazards 3-6
- Active Electronically Scanned Arrays (AESA) 3-7
- Fractal Antennas 3-8

RADAR EQUATIONS
- Field Intensity and Power Density 4-1
- Power Density 4-2
- One-Way Radar Equation / RF Propagation 4-3
- Two-Way Radar Equation (Monostatic) 4-4
- Alternate Two-Way Radar Equation 4-5
- Two-Way Radar Equation (Bistatic) 4-6
- Jamming to Signal (J/S) Ratio - Constant Power [Saturated] Jamming 4-7
- Burn-Through / Crossover Range 4-8
- Support Jamming 4-9
- Jamming to Signal (J/S) Ratio - Constant Gain (Linear) Jamming 4-10
- Radar Cross Section (RCS) 4-11
- Emission Control (EMCON) 4-12
- EW Jamming Techniques 4-13

RADAR AND RECEIVER CHARACTERISTICS & TEST
- RF Atmospheric Absorption / Ducting 5-1
- Receiver Sensitivity / Noise 5-2
- Receiver Types and Characteristics 5-3
- Radar Modes 5-4
- General Radar Display Types 5-5
- IFF - Identification - Friend or Foe 5-6
- Receiver Tests 5-7
- Signal Sorting and Direction Finding 5-8

MICROWAVE / RF COMPONENTS

Microwave Waveguides and Coaxial Cable 6-1
Voltage Standing Wave Ratio (VSWR) / Reflection Coefficient
    Return Loss / Mismatch Loss 6-2
Microwave Coaxial Connectors 6-3
Power Dividers and Directional Couplers 6-4
Attenuators / Filters / DC Blocks 6-5
Terminations / Dummy Loads 6-6
Circulators and Diplexers 6-7
Mixers and Frequency Discriminators 6-8
Detectors 6-9
RF / Microwave Amplifiers 6-10
Signal Generation 6-11
Digital Processing Components 6-12
Microwave Measurements 6-13

ELECTRO-OPTICS AND IR 7-1
Introduction 7-1.1
Optical Spectrum 7-1.1
Radiometric Quantities and Terminology 7-1.2
Photometric Quantities 7-1.3
Basic Principles 7-1.4
Basic Radiant Power Relationships 7-1.5
Infrared Source Characteristics 7-1.10
Atmospheric Transmission 7-1.14
EO Components and Sensors 7-1.16
IR Threats to Aircraft and Their Countermeasures 7-1.30
IRCM 7-1.36
Lasers 7-1.38
Fiber Optics 7-1.41
Laser Safety 7-2

AIRCRAFT DYNAMICS CONSIDERATIONS

Free Fall / Aircraft Drag 8-1
Mach Number and Airspeed vs Altitude 8-2
Maneuverability 8-3
EMP / Aircraft Dimensions 8-4

DATA TRANSFER BUSSES

Data Busses 9-1
RS-232 Interface 9-2
RS-422 Balanced Voltage Interface 9-3
RS-485 Interface 9-4
IEEE-488 Interface Bus (HP-IB/GP-IB) 9-5
MIL-STD-1553 & 1773 Data Bus 9-6
Ethernet 9-7
Transmission Control Protocol / Internet Protocol 9-8

GLOSSARY 10-1

ABBREVIATIONS AND ACRONYMS

a	Acceleration or atto (10^{-18} multiplier)
A	Ampere, Area, Altitude, Angstrom (Å), Antenna Aperture, or Aerial
A-799	No evidence of failure report
A/A, A-A, AA	Air-to-Air or Anti-Aircraft
AA-()	Air-to-Air missile number ()
AAA	Anti-Aircraft Artillery
AAAA	Army Aviation Association of America
AAED	Advanced Airborne Expendable Decoy
AAM	Air-to-Air Missile
AARGM	Advanced Anti-Radiation Guided Missile
AAW	Anti-Air Warfare
A-BIT	Automatic Built-in-Test
ABM	Air Breathing Missile or Anti-ballistic Missile
A/C	Aircraft (also acft.)
AC	Alternating Current
ACA	Associate Contractor Agreement or Airspace Coordination Area
ACAT	Acquisition Category
ACC	Air Combat Command
ACCB	Aircraft Configuration Control Board
Acft	Aircraft (also A/C)
ACLS	Aircraft Carrier Landing System
ACM	Advanced Cruise Missile or Air Combat Maneuvering
ACQ	Acquisition
ACS	Antenna Coupler Set
ACTD	Advanced Concept Technology Demonstration
A/D	Analog to Digital
ADM	Advanced Development Model
ADP	Automatic Data Processing or Advanced Development Program
AEA	Airborne Electronic Attack
AEC	Aviation Electronic Combat (Army)
AEGIS	Automatic Electronic Guided Intercept System
AEL	Accessible Emission Limit
AESA	Active Electronically Scanned Array
AEW	Airborne Early Warning
AF	Antenna Factor, Air Force, or Audio Frequency
AFB	Air Force Base or Airframe Bulletin
AFC	Automatic Frequency Control or Airframe Change
AFIPS	Automated Financial Information Processing System
AFOTEC	Air Force Operational T&E Center
A/G	Air-to-Ground
AGB	Autonomous Guided Bomb
AGC	Automatic Gain Control
AGI	Auxiliary General Intelligence (Intelligence-gathering Ship)
AGL	Above Ground Level
AGM	Air-to-Ground Missile
AGS	Angle Gate Stealer
AHWS	Advanced Helicopter Weapons System
AI	Artificial Intelligence, Air Intercept, or Airborne Interceptor
AIAA	American Institute of Aeronautics and Astronautics
AIC	Air Intercept Control
AIM	Air Intercept Missile
AIRLANT	Commander, U.S. Naval Air Forces, Atlantic Fleet
AIRPAC	Commander, U.S. Naval Air Forces, Pacific Fleet
AJ	Anti-jamming or Anti-Jam
A-Kit	Aircraft wiring kit for a system (includes cabling, racks, etc. excluding WRAs)
ALC	Air Logistics Center
AM	Amplitude Modulation
AMD	Aircraft Maintenance Department
AMES	Advanced Multiple Environment Simulator
AMLV	Advanced Memory Loader/Verifier
Amp	Amplifier
AMRAAM	Advanced, Medium-Range, Air-to-Air Missile
ANSI	American National Standards Institute
ANT	Antenna
Ao	Operational Availability
AO	Acousto-Optical
AOA	Angle of Arrival, Angle of Attack, or Analysis of Alternatives (similar to COEA)
AOC	Association of Old Crows (Professional EW Society) or Award of Contract
AOT	Angle Only Track, Angle Off Tail, or Acquisition-on-Target
APC	Amphenol Precision Connector or Armored Personnel Carrier
APN	Aircraft Procurement, Navy
APO	Armed Forces (or Army or Air) Post Office, Acquisition Program Office

APU	Auxiliary Power Unit
AR	Anti-reflection or Aspect Ratio
ARM	Anti-radiation Missile
ARO	After Receipt of Order
A/S, A-S, AS	Air-to-Surface
ASC	Air Systems Command
ASCM	Anti-ship Cruise Missile
ASE	Aircraft Survivability (or Survival) Equipment, Allowable Steering Error, or Automatic Support Equipment
ASIC	Application Specific Integrated Circuit
ASK	Amplitude Shift Keying
ASM	Air-to-Surface Missile
ASO	Aviation Supply Office
A-Spec	System Specification
ASPJ	Airborne Self-Protection Jammer
ASPO	Avionics Support (also Systems) Project Office (also Officer)
ASR	Advanced Special Receiver or Airport/Airborne Surveillance Radar
ASRAAM	Advanced Short Range Air-to-Air Missile
ASTE	Advanced Strategic and Tactical Expendables
ASW	Anti-submarine Warfare
ATA	Advanced Tactical Aircraft
ATARS	Advanced Tactical Air Reconnaissance System
ATC	Air Traffic Control
ATD	Advanced Technology Demonstration
ATE	Automatic Test Equipment
ATEDS	Advanced Technology Expendables and Dispenser Systems
ATF	Advanced Tactical Fighter (F-22)
ATIMS	Airborne Turret Infrared Measurement System or Airborne Tactical Information Management System
ATIRCM	Advanced Threat Infrared Countermeasures
ATP	Acceptance Test Procedure
ATR	Autonomous Target Recognition, Airborne Transportable Rack, Atlantic Test Range
ATRJ	Advanced Threat Radar Jammer
AUTODIN	Automatic Digital Network
AUX	Auxiliary
avdp.	Avoirdupois (system of measures)
Avg	Average
AWACS	Airborne Warning and Control System
AZ	Azimuth (also Az)
B	Bandwidth (also BW) or Magnetic inductance
BAFO	Best and Final Offer
BAU	Bus Adapter Unit
BC	Bus Controller
BDA	Battle Damage Assessment
BDI	Battle Damage Indication
BFO	Beat Frequency Oscillator
BI	Background Investigation
BIFF	Battlefield Identification, Friend, or Foe
BIT	Built-in-Test, Binary Digit or Battlefield Information Technology
BITE	Built-in-Test Equipment
BIU	Bus Interface Unit
B-Kit	Avionics "Black Box" WRAs
B/N	Bombardier/Navigator
BNC	Bayonet Navy Connector
BOA	Basic Ordering Agreement
BOL	Swedish chaff dispenser in a launcher
BPF	Band Pass Filter
BPS	Bits Per Second
BUMED	Bureau of Medicine (Navy)
BUNO	Bureau Number (aircraft)
BUR	Bottom Up Review
BVR	Beyond Visual Range
BW	Beamwidth (referring to an antenna) or sometimes Bandwidth
BWA	Backward Wave Amplifier
BWO	Backward Wave Oscillator
c	Speed of Light = 3x10^{8} meters/sec = 1.8x10^{12} furlongs per fortnight or 1.8 terafurlongs per fortnight, or centi (10^{-2}) multiplier
C	Electron Charge, Coulomb, Capacitance, Celsius, Centigrade, Confidential, Roman numeral for 100, or a programming language (also C+ and C++)
C^{2}	Command and Control
C^{3} (C^{4})	Command, Control, Communications (and Computers)
C^{3}I (C^{4}I)	Command, Control, Communications, (Computers) and Intelligence
CAD	Computer-Aided Design
CAE	Computer-Aided Engineering
CAG	Carrier Air Group
CAGE	Commercial and Government Entry
CAIV	Cost as an Independent Variable
CAL	Calibration

CAM	Computer-Aided Manufacturing or Constant Addressable Memory
CAO	Competency Aligned Organization or Contract Administrative Officer
CAP	Combat Air Patrol
CAS	Close Air Support or Calibrated Airspeed
CASS	Consolidated Automated Support System
CAT	Catapult or Cockpit Automation Technology
CB	Citizens Band (also see Seabee)
CBD	Commerce Business Daily
CBIT	Continuous Built-in-Test
CBO	Congressional Budget Office
CCA	Circuit Card Assembly
CCB	Configuration Control Board
CCD	Charge Coupled Device
CCM	Counter-Countermeasures
CCN	Contract Change Number or Configuration Change Notice
CCU	Cockpit Control Unit
cd	Candela (SI unit of luminous intensity)
CD	Compact Disk or Control and Display
CDC	Combat Direction Center
CDR	Critical Design Review
CDRL	Contract Data Requirements List
CE	Conducted Emission
CECOM	Communications and Electronics Command (Army)
CEESIM	Combat Electromagnetic Environment Simulator
CEP	Circular Error Probability
CFA	Cognizant Field Activity
CFAR	Constant False Alarm Rate
CFE	Contractor Furnished Equipment
CG	Center of Gravity, Commanding General, Command Guidance, or Cruiser
CI	Configuration Item
CIA	Central Intelligence Agency
CIC	Combat Information Center (now called CDC)
CID	Combat Identification or Charge Injection Device
CILOP	Conversion in Lieu of Procurement
CINC	Commander in Chief
CIP	Capital Improvement Program
CIS	Commonwealth of Independent States (11 of 15 former Soviet Union territories except Estonia, Georgia, Latvia, and Lithuania)
CIWS	Close-In Weapon System
CJ	Coherent Jamming
CLC	Command Launch Computer
cm	Centimeter
CM	Countermeasures or Configuration Management
CMC	Command Mission Computer or Commandant Marine Corps
CMDS	Countermeasure Dispensing System
CMOS	Complementary Metal-Oxide Semiconductor
CMP	Configuration Management Plan
CMWS	Common Missile Warning System
CNAF	Commander, Naval Air Forces
CNAL	Commander, Naval Air Forces Atlantic (also COMNAVAIRLANT)
CNAP	Commander, Naval Air Forces Pacific (also COMNAVAIRPAC)
CNI	Communications, Navigation, and Identification
CO	Commanding Officer, Contracting Officer, Change Order, or Carbon Monoxide
COB	Close of Business
COCOM	Combatant Command
COEA	Cost and Operational Effectiveness Analysis
COG	Center of Gravity or Cognizant
COMM	Communications
COMSEC	Communications Security
CONSCAN	Conical Scanning Radar
CONUS	Continental United States
CO-OP	Cooperative (countermeasures)
COR	Contracting Officers Representative
CORPORAL	Collaborative On-Line Reconnaissance Provider/Operationally Responsive Attack Link
Cos	Cosine
COSRO	Conical-Scan on Receive Only
COTS	Commercial Off-The-Shelf (hardware/software)
CP	Circularly Polarized (antenna), Central Processor, or Command Post
CPS	Computer or Control Power Supply (depends on application)
CPU	Central Processing Unit
CRC	Originally Chemical Rubber Company, now published reference books by CRC Press
CRFM	Coherent RF Memory
CRISD	Computer Resources Integrated Support Document
CRLCMP	Computer Resources Life Cycle Management Plan

CRO	Countermeasures Response Optimization
CRT	Cathode Ray Tube or Combat Rated Thrust (afterburner)
CSAR	Combat Search and Rescue
Crypto	Cryptographic
CS	Conducted Susceptibility
CSC	Commodity Software Change
CSCI	Computer Software Configuration Item
C-Spec	Product Specification
CSS	Contractor Support Services
CTR	Chesapeake Test Range
CV	Aircraft Carrier
CVN	Nuclear Powered Aircraft Carrier
CVR	Crystal Video Receiver
CW	Continuous Wave or Chemical Warfare
CWBS	Contract Work Breakdown Structure
CWI	Continuous Wave Illuminator
CY	Calendar Year
d	Distance, Diameter, or deci (10^{-1} multiplier)
D	Distance, Diameter, Electron displacement, Detectivity, Doppler, Density, or Roman numeral for 500
da	deca (10^{0} multiplier)
D/A	Digital-to-Analog
DAB	Defense Acquisition Board
DAC	Digital to Analog Converter or Dept of Army Civilian
DAR	Defense Acquisition Regulation
DARPA	Defense Advanced Research Projects Agency
DB	Database
dB	Decibel
dBc	dB referenced to the Carrier Signal
dBi	Decibel antenna gain referenced to an isotropic antenna
dBm	Decibel referenced to the power of one milliwatt
DBOF	Defense Business Operations Fund
dBsm	Decibel value of radar cross section referenced to a square meter
dBW	Decibel referenced to the power of one watt
DC	Direct Current, Discrete Circuit, or District of Columbia
DCE	Data Communication Equipment
DCS	Direct Commercial Sales or Distributed Control System
DDI	Digital Display Indicator
DDS	Direct Digital Synthesizers
DECM	Deceptive Electronic Countermeasures (also Defensive ECM)
deg	Degree
DEMVAL	Demonstration Validation (also DEM/VAL)
DET	Detachment
DF	Direction Finding
DFT	Discrete Fourier Transform
DI	Data Item
DIA	Defense Intelligence Agency or Diameter
DID	Data Item Description
DIRCM	Directed Infrared Countermeasures
DJ	Deceptive Jamming
D-Level	Depot Level Maintenance
DM	Data Management (also manager)
DMA	Direct Memory Address or Defense Mapping Agency
DME	Distance Measuring Equipment
DNA	Defense Nuclear Agency, Does Not Apply, or Deoxyribonucleic Acid
DOA	Direction of Arrival
DOD or DoD	Department of Defense
DoDISS	DoD Index of Specifications and Standards
DOM	Depth of Modulation
DON	Department of the Navy
DOS	Disk Operating System
DOT&E	Director, Operational Test & Evaluation
DPRO	Defense Plant Representative Office
DRB	Defense Review Board
DRFM	Digital RF Memory
DSARC	Defense Systems Acquisition (and) Review Council
DSN	Defense Switching Network
DSO	Dielectrically Stabilized Oscillator
DSP	Digital Signal Processor
D-Spec	Process Specification
DT (&E)	Development or Developmental Test (and Evaluation)
DTC	Design to Cost
DTE	Data Terminal Equipment
DTO	Digitally Tuned Oscillator or Defense Technology Objectives
DTRMC	Defense (or DoD) Test Recourse Management Center

e	Electron charge or base of natural logarithms (2.71828...)
E	Electric Field Intensity or Strength, Energy, East, or Exa (10^{18} multiplier)
E^3	Electromagnetic Environmental Effects
EA	Electronic Attack (similar to older term of ECM)
EC	Electronic Combat
ECAC	Electromagnetic Compatibility Analysis Center (DOD), now Joint Spectrum Center
ECCM	Electronic Counter-Countermeasures (similar to newer term of EP)
ECL	Emitter Coupled Logic
ECM	Electronic Countermeasures (similar to newer term of EA)
ECN	Engineering Change Notice
ECO	Engineering Change Order
ECP	Engineering Change Proposal or Egress Control Point
ECR	Electronic Combat Range (China Lake) or Electronic Combat & Reconnaissance
ECS	Environmental Control System
ECSEL	Electronic Combat Simulation and Evaluation Laboratory
ECU	Electronic Control Unit
EDM	Engineering Development Model
EED	Electro-Explosive Device
EEPROM	Electrically Erasable/Programmable Read-only Memory
EHF	Extremely High Frequency [30 to 300 GHz]
EIA	Electronic Industries Associates
EID	Emitter Identification Data
EIRP	Effective Isotropic Radiated power
EL	Elevation (also El)
ELF	Extremely Low Frequency [3 Hz to 3 kHz]
ELINT	Electronics Intelligence
ELNOT	Emitter Library Notation
EM	Electromagnetic
E-Mail	Electronic Mail
EMC	Electromagnetic Compatibility
EMCAB	EMC Advisory Board
EMCON	Emission Control
EMD	Engineering and Manufacturing Development
EME	Electromagnetic Environment
EMI	Electromagnetic Interference
EMP	Electromagnetic Pulse
EMR	Electromagnetic Radiation
EMS	Electromagnetic Susceptibility, Electromagnetic Spectrum
EMV	Electromagnetic Vulnerability
EO	Electro-Optic, Electro-Optical, or Engineering Order
EOB	Electronic Order of Battle or Expense Operating Budget
EOCM	Electro-Optic Countermeasures
EOF	Electro-Optical Frequency (300 to 3 x 10^7 GHz)
EP	Electronic Protection (similar to older terms of ECCM and DECM)
EPA	Environmental Protection Agency
EPROM	Electrically Programmable Read-only Memory
ERAM	Electronic Counter-Countermeasures (also Protection) Requirements and Assessment Manual
ERP	Effective Radiated Power
ES	Electronic Surveillance (similar to older term of ESM)
ESD	Electrostatic Discharge
ESM	Electronic Support Measures (similar to newer term of ES)
ESSM	Evolved Sea Sparrow Missile
ET	Electronics Technician
ETI	Elapsed Time Indicator
ETIRMS	EW Tactical Information and Report Management System
ETR	Estimated Time to Repair
EVM	Earned Value Management
EW	Electronic Warfare , Early Warning, or Expeditionary Warfare
EWBM	EW Battle Management
EWDS	EW Data Systems
EWIA	EW Intelligence Analysis
EWIR	Electronic Warfare Integration & Reprogramming (USAF database)
EWMP	Electronic Warfare Master Plan
EWO	Electronic Warfare Officer
EWRL	Electronic Warfare Reprogrammable Library (USN)
EWSI	EW Systems Integration
EWSSA	EW Software Support Activity
EXP	Expendable Countermeasure
f	femto (10^{-15} multiplier), Frequency (also F), or lens *f* number
F	Frequency (also *f*), Force, Farad, Faraday Constant, Female, Fahrenheit, Noise Figure, Noise Factor or “Friendly” on RWR display

F2T2EA	Find, Fix, Track, Target, Engage, Assess (targeting of hostile forces)
F/A	Fighter/Attack
FAA	Federal Aviation Administration
FAC	Forward Air Controller
FAR	Federal Acquisition Regulations or False Alarm Rate
f_c	Footcandle (unit of illuminance)
FCA	Functional Configuration Audit
FCR	Fire Control Radar
FDR	Frequency Domain Reflectometry
FEBA	Forward Edge of the Battle Area
FET	Field-Effect Transistor
FEWC	Fleet EW Center
FFT	Fast Fourier Transform
FIFO	First In / First Out
FIPR	Federal Information Processing Resources
fl	fluid
FLAK	AAA Shrapnel, from the German "Flieger Abwher Kanone" (AAA gun that fires fast and furiously)
FLIR	Forward Looking Infrared
FLPS	Flightline Payload Simulator
FLT	Flight
FM	Frequency Modulation or Failure Mode
FME	Foreign Material Exploitation
FMEA	Failure Mode and Effects Analysis
FMS	Foreign Military Sale(s)
FOC	Full Operational Capability
FOD	Foreign Object Damage
FORCECAP	Force Combat Air Patrol
FOT&E	Follow-On Test and Evaluation
FOTD	Fiber Optic Towed Device
FOUO	For Official Use Only
FOV	Field of View
FPA	Focal Plane Array
fps	feet per second
FRACAS	Failure, Reporting, Analysis, and Corrective Actions System
FRB	Failure Review Board
FRD	Functional Requirements Document
FSD	Full Scale Development
FSED	Full Scale Engineering Development
FSK	Frequency Shift Keying
FSU	Former Soviet Union
ft	Feet or Foot
FTC	Fast Time Constant
FTD	Foreign Technology Division (USAF)
FWD	Forward
FY	Fiscal Year
g	Gravity (also G)
G	Universal Gravitational Constant (also K), Giga (10^9 multiplier), Conductance, or Gain
G&A	General and Administrative (expense)
GaAs	Gallium Arsenide
GACIAC	Guidance and Control Information Analysis Center (DoD)
gal	Gallon
GAO	General Accounting Office
GBU	Guided Bomb Unit
GCA	Ground Controlled Approach
GCI	Ground Control Intercept
GENSER	General Service
GEN-X	Generic Expendable
GFE	Government Furnished Equipment
GHz	GigaHertz
GI	Government Issue
GIDEP	Government Industry Data Exchange Program
GIG	Global Information Grid
GIGO	Garbage In / Garbage Out
GOCO	Government Owned Contract Operated
GOFO	General Officer / Flag Officer
GP	General Purpose
GPI	Ground Plane Interference
GPIB	General Purpose Interface Bus
GPS	Global Positioning System
GSE	Ground Support Equipment
h	hours, hecto (10^2 multiplier), Plank's constant, or height (also H)
H	Height (also h), Henry (Inductance), or Irradiance
HARM	High-speed Anti-Radiation Missile
HAWK	Homing All the Way Killer
HDBK	Handbook
HDF	High Duty Factor
HE	High Explosive
HEF	High Energy Frequency (3x10^7 to 3x10^{14} GHz)
HEL	High Energy Laser
HELO	Helicopter
HERF	Hazards of Electromagnetic Radiation to Fuel
HERO	Hazards of Electromagnetic Radiation to Ordnance
HERP	Hazards of Electromagnetic Radiation to Personnel
hex	hexadecimal
HF	High Frequency [3 - 30 MHz]
HIL or HITL	Hardware-in-the-Loop

HOJ	Home-On-Jam
HOL	Higher Order Language
HPF	High-Pass Filter
HP-IB	Hewlett-Packard Interface Bus
HP-IL	Hewlett-Packard Interface Loop
HPM	High Powered Microwave
HPRF	High Pulse Repetition Frequency
hr	hour
HSDB	High Speed Data Bus
HUD	Heads-Up Display
HV	High Voltage
H/W	Hardware
HWCI	Hardware Configuration Item
HWIL	Hardware-in-the-loop
Hz	Hertz (Cycles per second)

i	current (also I)
I	Current (also i), Intensity, Irradiance, Intermediate, or Roman Numeral for One
IA	Information Assurance
IADS	Integrated Air Defense System
I&Q	In-Phase and Quadrature
IAS	Indicated Airspeed
IAW	In Accordance With
IBIT	Initiated Built-in-Test
IBU	Interference Blanker Unit
IC	Integrated Circuit
ICD	Interface Control Document Initial Capabilities Document
ICMD	Improved Countermeasure Dispenser
ICNIA	Integrated Communication, Navigation, Identification Avionics
ICS	Inverse Conical Scan or Intercommunications System (aircraft)
ICW	In Compliance With
ID	Identification
IDA	Institute For Defense Analysis
IDECM	Integrated Defensive Electronic Countermeasures
IEEE	Institute of Electrical and Electronic Engineers
IF	Intermediate Frequency
IFF	Identification Friend-or-Foe
IFM	Instantaneous Frequency Measurement
IFR	Instrument Flight Rules
IG	Inspector General
IIR	Imaging Infrared
I-Level	Intermediate Level of Repair (also "I" Level)
ILS	Integrated Logistic Support, Instrument Landing System, or Inertial Locator System
ILSMT	Integrated Logistic Support Management Team
IM	Intermodulation or Item Manager
IMA	Intermediate Maintenance Activity
in	Inch
INEWS	Integrated Electronic Warfare System
INS	Inertial Navigation System
INT	Intensity
IO	Information Operations
I/O	Input/Output
IOC	Initial Operational (also Operating) Capability
IOT&E	Initial Operational Test and Evaluation
IPO	International Projects (Program) Office
IPR	In-Progress/Process Review
IPT	Integrated Product (also Program) Team
IR	Infrared
IR&D	Independent Research and Development
IRCM	Infrared Countermeasures
IRDS	Infrared Detecting System
IREXP	IR Expendables
IRIG-B	Inter-range Instrumentation Group B
IRLS	Infrared Line Scanner
IRS	Interface Requirements Specification, IR Suppression or Internal Revenue Service
IRST	Infrared Search and Track
ISAR	Inverse Synthetic Aperture Radar
ISO	Derived from the Greek "isos" meaning "equal," the official title is International Organization for Standardization
ISP	Integrated Support Plan Intelligence Support Plan
ISR	Interference to Signal Ratio (also I/S)
ITU	International Telecommunications Union
IV&V	Independent Validation and Verification
IW	Information Warfare

J	Jamming, Radiance, Current Density, or Joules
JAAS	Joint Architecture for Aircraft Survivability
JAFF	Jammer (illuminating) Chaff
JAG	Judge Advocate General
JAMS	Jamming Analysis Measurement System
JARS	Jamming Aircraft & Radar Simulation
JASSM	Joint Air-to-Surface Standoff Missile
JAST	Joint Advanced Strike Technology
JATO	Jet Assisted Takeoff or JAmmer Technique Optimization
JC2WC	Joint Command and Control Warfare Center (now JIOWC)
JCTD	Joint Concept Technology Demonstration
JCS	Joint Chiefs of Staff or Joint Spectrum Center (formerly ECAC)
JDAM	Joint Direct Attack Munition
JEACO	Joint Electronic Attack and Compatibility Office
JED	Journal of Electronic Defense (Published by the Association of Old Crows)
JEM	Jet Engine Modulation
JETS	Joint Emitter Targeting System
JEWC	Joint EW Conference or Joint EW Center (then JC2WC & now JIOWC)
JEWEL	Joint Electronic Warfare Effects Laboratory
JIOWC	Joint Information Operations Warfare Command
JMEM	Joint Munitions Effectiveness Manual
JMR	Jammer
JOVIAL	Julius' Own Version of International Algorithmic Language (Air Force computer programming language)
JPATS	Joint Primary Aircraft Training System
J/S	Jamming to Signal Ratio
JSF	Joint Strike Fighter
JSGCC	Joint Services Guidance and Control Committee
JSIR	Joint Spectrum Interference Resolution (signal interference portion of MIJI)
JSOW	Joint Stand-Off Weapon (AGM-154A)
JSTARS	Joint Surveillance Target Attack Radar System
JTA	Jammer Threat Analysis
JTAT	JATO Techniques Analysis and Tactics
JTCG/AS	Joint Technical Coordinating Group for Aircraft Survivability
JTCG/ME	Joint Technical Coordinating Group for Munitions Effectiveness
JTIDS	Joint Tactical Information Distribution System
JV or J/V	Joint Venture

k	kilo (10^3 multiplier) or Boltzmann Constant
K	Kelvin, Cathode, Universal gravitational constant (also G), or Luminous efficacy
KCAS	Knots Calibrated Airspeed
kg	kilogram
kHz	KiloHertz
KIA	Killed in Action
KIAS	Knots Indicated Air Speed
km	Kilometer
KSLOC	Thousand Source Lines of Code (software)
kt	Knot (nautical miles per hour)
kW	Kilowatt

l	length (also L) or liter
L	Length (also l), Loss, inductance, Luminance, or Roman Numeral for fifty
LADAR	Laser Detection and Ranging (i.e., laser radar)
LAN	Local Area Network
LANTIRN	Low Altitude Navigation & Targeting Infrared for Night
LASER	Light Amplification by Stimulated Emission of Radiation
LAT	Latitude (0-90° N or S from equator)
lbs	pounds
LCC	Life Cycle Cost(s)
LCD	Liquid Crystal Display or Lowest Common Denominator
LCP or LHCP	Left-hand Circular Polarization
LDF	Low Duty Factor
LDS	Laser Detecting Set
LED	Light-Emitting Diode
LEX	Leading Edge Extension
LGB	Laser Guided Bomb
LF	Low Frequency [30 - 300 kHz]
LIC	Low Intensity Combat or Laser Intercept Capability

LISP	List Processing (A programming language used in artificial intelligence)
LLL	Low Light Level (as in LLL TV)
lm	lumen (SI unit of luminous flux)
ln	Natural Logarithm
LO	Local Oscillator or Low Observable
LOA	Letter of Agreement (or Acceptance)
LOB	Line of Bearing (see also AOA)
LOG	Logarithm to the base 10 (also log) or Logistician
LONG	Longitude (0-180° E or W from Greenwich, U.K.)
LOR	Level of Repair
LORA	Level of Repair Analysis
LORAN	Long Range Navigation
LORO	Lobe on Receive Only
LOS	Line-of-Sight
LPAR	Large Phased-Array Radar
LPD	Low Probability of Detection
LPF	Low Pass Filter
LPI or LPOI	Low Probability of Intercept
LPRF	Low Pulse Repetition Frequency
LR	Lethal Range
LRA	Line Replaceable Assembly
LRF	Laser Rangefinder
LRIP	Low Rate Initial Production
LRU	Line Replaceable Unit
LSA	Logistic Support Analysis
LSAR	Logistic Support Analysis Record
LSB	Least Significant Bit
LSI	Large Scale Integration
LSO	Landing Signal Officer
LSSO	Laser System Safety Officer
LTBB	Look Through Blanking Bus
LWIR	Long Wave Infrared
LWR	Laser Warning Receiver
lx	Lux (SI unit of illuminance)
LZ	Landing Zone

m	milli (10^{-3} multiplier), meter, or electron mass
M	Mega (10^{6} multiplier), Male, Mach number, or Roman numeral for 1,000
MA	Missile Alert or Missile Active
MAD	Magnetic Anomaly Detection (also Detector)
MADD	Microwave Acoustic Delay Device
MAF	Maintenance Action Form
MAG	Marine Aircraft Group or Magnetic
MAGTF	Marine Air-Ground Task Force
MANPADS	Man-portable Air Defense System
M&S	Modeling and Simulation
MASER	Microwave Amplification by Simulated Emission of Radiation
MATE	Modular Automatic Test Equipment
MAW	Missile Approach Warning system (also MAWS) or Marine Aircraft Wing
MAX	Maximum or Maximum aircraft power (afterburner)
MBFN	Multiple Beam Forming Network
MC	Mission Computer
MCIOC	Marine Corps Information Operations Center
MCP	Micro-Channel Plate
MDF	Mission Data File
MDI	Multiple Display Indicator or Miss Distance Indicator
MDG	Mission Data Generator
MDS	Minimum Discernible Signal or Minimum Detectable Signal
MDU	Multipurpose Display Unit
MF	Medium Frequency (300 kHz to 3 MHz)
MFD	Multifunction (video) Display
MG	Missile Guidance
MHz	MegaHertz (10^{6} Hz)
MIA	Missing in Action
MIC	Microwave Integrated Circuit or Management Information Center
MICRON	10^{-6} meter
MiG	Mikoyan-Gurevich (Soviet aircraft manufacturer)
MIGCAP	MiG Combat Air Patrol
MIJI	Meaconing, Intrusion, Jamming, & Interference (also see JSIR)
mil	One-thousandth of an inch
MIL	Military power (100%, no afterburner) or Military
MILCON	Military Construction
MILSPEC	Military Specification
MILSTRIP	Military Standard Requisitioning and Issue Procedure(s)
MIMIC	Microwave Monolithic Integrated Circuit (also MMIC)
MIN	Minimum
Mincon	Minimal Construction
MIPPLE	RWR display switching between ambiguous emitters
MIPS	Millions of (Mega) Instructions Per Second
ML	Missile Launch
MLC	Main Lobe Clutter
MLV	Memory Loader Verifier
MLVS	Memory Loader Verifier Set

mm	Millimeter
MM	Man Month
MMIC	Microwave Monolithic Integrated Circuit (also MIMIC)
MMW	Millimeter Wave (40 GHz or higher per IEEE, but commonly used down to 30 GHz)
MOA	Memorandum of Agreement
MOE	Measure of Effectiveness
MOM	Methods of Moments (also MoM) or Metal-Oxide-Metal
MOP	Modulation on Pulse or Measure of Performance
MOPS	Million Operations Per Second
MOS	Minimum Operational Sensitivity, Military Occupational Specialty, Metal-Oxide Semiconductor, or Measure of Suitability
MOSAIC	Modeling System for Advanced Investigation of Countermeasures
MOU	Memorandum of Understanding
MPD	Multi-Purpose Display or Microwave Power Device
MPE	Maximum Permissible Exposure
mph	Miles per Hour
MPLC	Multi-Platform Launch Controller
MPM	Microwave Power Module
MPPS	Million Pulses Per Second
MPRF	Medium Pulse Repetition Frequency
mr or mrad	Milliradian
MRC	Maintenance Requirement Card or Medium Range CAP
MRE’s	Meals Ready to Eat
ms	Milliseconds
MSB	Most Significant Bit
MSI	Multi-Sensor (also Source) Integration, Management Support Issues, or Medium Scale Integration
MSIC	Missile and Space Intelligence Center
MSL	Mean Sea Level (altitude) or Missile
MTBF	Mean Time Between Failures
MTI	Moving Target Indicator (or Indication)
MTTR	Mean Time To Repair
MUXBUS	Multiplex Bus
MVS	Minimum Visible Signal
mw	Microwave
mW	Milliwatt
MWIR	Mid Wave Infrared
MWS	Missile Warning Set
MY	Man Year
n	nano (10^{-9} multiplier) or number of elements
N	Noise, Newton (force), Radiance, North, or No
n/a	Not Applicable (also N/A)
NA	Numerical Aperture
NADEP	Naval Aviation Depot
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NATOPS	Naval Air Training and Operating Procedures Standardization
NAV	Navigation
NAVAIR	Naval Air Systems Command (also NAVAIRSYSCOM)
NavMPS	Naval Mission Planning System
NAVSEA	Naval Sea Systems Command (also NAVSEASYSCOM)
NAWCAD	Naval Air Warfare Center Aircraft Division
NAWCWD	Naval Air Warfare Center Weapons Division
NBC	Nuclear, Biological, Chemical
NCTR	Non-Cooperative Target Recognition
NDI	Non-Developmental Item or Non Destructive Inspection
NEI	Noise Equivalent Power
NEMP	Nuclear Electromagnetic Pulse
NEOF	No Evidence of Failure
NEP	Noise Equivalent Power
NF	Noise Figure or Noise Factor (also F)
NFO	Naval Flight Officer
NGJ	Next Generation Jammer
NIOC	Navy Information Operations Command
NIPO	Navy International Program Office
NIR	Near Infrared
nm	nanometer or Nautical Mile (also NM or NMI)
NM or NMI	Nautical Mile (also nm)
NMCI	Navy Marine Corps Intranet
NNWC	Naval Network Warfare Command
NOHD	Nominal Ocular Hazard Distance
NORAD	North American Air Defense Command
NPG or NPGS	Naval Post Graduate School
NRE	Non-Recurring Engineering
NRL	Naval Research Laboratory
NRZ	Non Return to Zero
NSA	National Security Agency
nsec or ns	Nanosecond
NSN	National Stock Number
NSWC	Naval Surface Weapons Center
nt	Nit (SI unit of luminance)

NUWC	Naval Undersea Warfare Center
NVG	Night Vision Goggles
NWIP	Naval Warfare Information Publication
NWP	Naval Warfare Publication
O	Optical
OADR	Originating Agency's Determination Required
OAG	Operational Advisory Group
O&MN	Operations and Maintenance, Navy (also O&M,N)
OBE	Overtaken (Overcome) By Events
OCA	Offensive Counter Air
OEWTPS	Organizational Electronic Warfare Test Program Set
OFP	Operational Flight Program
OJT	On-the-Job Training
O-Level	Organizational Level of Repair (also "O" Level)
OMA	Organizational Maintenance Activity
OMB	Office of Management and Budget
OMEGA	Optimized Method for Estimating Guidance Accuracy (VLF Navigation System)
ONR	Office of Naval Research
OOK	On-Off Keying
OPEVAL	Operational Evaluation
OPM	Office of Personnel Management
OPSEC	Operational Security
OPTEVFOR	Operational Test and Evaluation Force
OR	Operational Requirement or Operationally Ready
ORD	Operational Requirements Document
OSD	Office of the Secretary of Defense
OSHA	Occupational Safety and Health Act
OSIP	Operational Safety Improvement Program
OSM	Operating System Memory or SMA connector made by Omni-Spectra
OSRB	Operational Software Review Board
OT (&E)	Operational Test (and Evaluation)
OTD	Operational Test Director
OTH	Over the Horizon
OTH-B	Over-the-Horizon Backscatter
OTH-R	Over-the-Horizon Radar
OTH-T	Over-the-Horizon Targeting
OTRR	Operational Test Readiness Review
OUSD	Office of the Under Secretary of Defense
oz	ounce

p	pico (10^{-12} multiplier) or page
P	Power, Pressure, or Peta (10^{15} multiplier)
P^3I	Pre-Planned Product Improvement
Pa	Pascal (pressure)
PA	Public Address or Program Analyst
PBIT	Periodic Built-in-Test
PC	Pulse Compression, Personal Computer, or Photoconductive
PCA	Physical Configuration Audit
PCM	Pulse Code Modulation
P_d	Probability of Detection
PD	Pulse Doppler
PDI	PD Illuminator or Post Detection Integration
PDP	Plasma Display Panel
PDQ	Pretty Darn [sic] Quick
PDR	Preliminary Design Review
PDW	Pulse Descriptor Word
PEL	Personnel Exposure Limits
PEM	Photoelectromagnetic
PEO	Program Executive Officer
pf	Power Factor or Pico Farads
PFA	Probability of False Alarm
PGM	Precision Guided Munition
ph	Phot (unit of illuminance)
P_h	Probability of Hit
pi	Greek letter π
P_i	Probability of Intercept (also POI)
PID	Positive Identification
PIN	Personal Identification Number
PIP	Product Improvement Plan or Predicted Intercept Point
Pixel	Picture Element
P_k	Probability of Kill or Peak
PLSS	Precision Location Strike System
PM	Phase Modulation or Program Manager
PMA	Program (also Project) Manager, Air
PMAWS	Passive Missile Approach Warning System
PMS	Program Manager, Ship
PMT	Photomultiplier Tube
PMW	Program Manager, Warfare
P-N	Positive to Negative Junction (also p-n)
PN or P/N	Part Number
POC	Point of Contact
POET	Primed Oscillator Expendable Transponder
POI	Probability of Intercept (also PI)
POL	Polarization
POM	Program Objective Memorandum

POP	Pulse-on-Pulse or Product Optimization Program
POST	Passive Optical Seeker Technology (Stinger missile)
PPI	Plan Position Indicator
PPS	Pulses Per Second
PRF	Pulse Repetition Frequency
PRI	Priority or Pulse Repetition Interval
PROM	Programmable Read-only Memory
PRR	Production Readiness Review or Pulse Repetition Rate
PRT	Pulse Repetition Time
P_s	Probability of Survival
P’s & Q’s	Pints and Quarts (small details)
PSK	Phase-shift Keying
PUPS	Portable Universal Programming System
PV	Photovoltaic
pw or PW	Pulse Width
PWB	Printed Wiring Board

q	electron charge
Q	Quantity Factor (figure of merit), Quadrature, aerodynamic pressure, or Charge (coulomb)
QA	Quality Assurance
QC	Quality Control
QED	Quod Erat Demonstradum (end of proof)(Satirically “quite easily done”)
QML	Qualified Manufacturer Listing
QPL	Qualified Parts List
QRC	Quick-Reaction Capability
QRD	Quick Reaction Demonstration
QRT	Quick-Reaction test

r or R	Radius or Range
R	Resistance, Reliability, or Roentgen
rad	Radian
R&D	Research and Development
RADAR	Radio Detection and Ranging
RADHAZ	Radiation Hazard
RAM	Random Access Memory, Radar Absorbing Material, Rolling Airframe Missile, or Reliability, Availability, and Maintainability
R&M	Reliability and Maintainability
RAT	Ram Air Turbine
RBOC	Rapid Blooming Offboard Chaff
RCP or RHCP	Right-hand Circular Polarization
RCS	Radar Cross Section
RCVR	Receiver
RDT&E	Research, Development, Test, & Evaluation
RDY	Ready
RE	Radiated Emissions
REC	Receive
RET	Return
RF	Radio Frequency
RFEXP	RF Expendables
RFI	Radio Frequency Interference, Ready-For-Issue, or Request for Information
RFP	Request for Proposal
RFQ	Request for Quotation
RFSS	Radio Frequency Simulation System (Army)
RGPO	Range Gate Pull Off
RGS	Range Gate Stealer
RGWO	Range Gate Walk Off (see RGPO)
RHAW	Radar Homing and Warning Receiver or Radar Homing All the Way
RHAWS	Radar Homing and Warning System
RINT	Radiation Intelligence
RIO	Radar Intercept Officer
RM	Radar Mile
rms or RMS	Root Mean Square
RNG	Range
ROC	Required Operational Capability
ROE	Rules of Engagement
ROI	Return on Investment
ROM	Read-only Memory or Rough Order of Magnitude
ROR	Range Only Radar or Rate of Return (financial)
ROT	Rate of Turn
ROWG	Response Optimization Working Group
RPG	Receiver Processor Group
RPM	Revolutions per Minute
RPT	Repeat
RPV	Remotely Piloted Vehicle
RRT	Rapid Reprogramming Terminal (a type of MLVS)
RS	Radiated Susceptibility or Remote Station
RSDS	Radar Signal Detecting Set
RSO	Range Safety Officer or Receiver, Set-on
RST	Receiver Shadow Time
RT	Remote Terminal, Termination Resistance, or Receiver/Transmitter (also R/T)
RUG	Radar Upgrade
RWR	Radar Warning Receiver
Rx	Receive

s, S, or sec	seconds
S	Signal Power, Surface Area, Secret, Electrical conductance (siemens), South, Scattering (as in S-parameters), or Seconds
SA	Situational Awareness, Semi-Active, Spectrum Analyzer, or Surface-to-Air (also S/A or S-A)
SA-()	Surface-to-Air missile number ()
SAE	Society of Automotive Engineers
SAM	Surface-to-Air Missile
SA-N-()	Naval Surface-to-Air missile number
SAR	Synthetic Aperture Radar, Special Access Required, Semi-Active Radar, Search and Rescue, or Specific Absorption Rate
SATS	Semi-Active Test System
SAW	Surface Acoustic Wave
SBIR	Small Business Innovative Research
SCI	Sensitive Compartmented Information
SCIF	Sensitive Compartmented Information Facility
SCN	Specification Change Notice
SCR	Software Change Request
SCP	Software Change Proposal
SCRB	Software Configuration Review Board
SCUD	Soviet short-range surface-to-surface missile
SE	Support Equipment
SDLM	Standard Depot Level Maintenance
SDI	Strategic Defense Initiative
SEAD	Suppression of Enemy Air Defense (pronounced "seed" or "C add")
SEAL	Sea-Air-Land (Navy special forces)
sec	seconds (also S or s)
SECDEF	Secretary of Defense
SEI	Specific Emitter Identification
SEMA	Special Electronic Mission Aircraft
SERD	Support Equipment Recommendation Data
SHAPE	Supreme Headquarters Allied Powers Europe (NATO military command)
SHF	Super High Frequency (3 to 30 GHz)
SI	Special Intelligence or System International (Units)
SIF	Selective Identification Feature
SIGINT	Signals Intelligence
SIJ	Stand-In Jamming (also S/J)
SIM	Simulation
sin	Sine
SINCGARS	Single Channel Ground and Airborne Radio System
SIRCM	Suite of IR Countermeasures
SIRFC	Suite of Integrated RF Countermeasures
SJ	Support Jamming
S/J	Stand-In Jamming or Signal to Jamming Ratio
SL	Side lobe or Sea Level (also S.L.)
SLAM	Standoff Land Attack Missile
SLAR	Side-Looking Airborne Radar
SLC	Side Lobe Clutter
SLOC	Source Lines of Code or Sea Lines of Communication
SM	Statute Mile (also sm) or Standard Missile
SMA	Scheduled Maintenance Action or Sub-Miniature A connector
SMC	Sub-Miniature C connector
SME	Subject Matter Expert
SML	Support Material List
SMS	Stores Management Set or Status Monitoring System
S/N or SNR	Signal-to-Noise Ratio
SNORT	Supersonic Naval Ordnance Research Track
SNTK	Special Need to Know
SOF	Safety of Flight
SOJ	Stand-off Jammer
SONAR	Sound Navigation and Ranging
SOO	Statement of Objectives (replacing SOW)
SOP	Standard Operating Procedures
SORO	Scan-on-Receive Only
SOS	"Save Our Ship" (distress call with easy Morse code, i.e. • • • – – – • • •)
SOW	Statement of Work (being replaced by SOO)
SPAWAR	Space and Naval Warfare Systems Command
SPEC	Specification
SPIRITS	Spectral Infrared Imaging of Targets and Scenes
SPO	System Program Office
SPY	Radar on an AEGIS ship
sq	Square
sr	Steradian
SRA	Shop Replaceable Assembly
SRAM	Static Random Access Memory
SRB	Software Review Board
SRBOC	Super Rapid Blooming Offboard Chaff
SRD	Systems Requirements Document
SRS	Software Requirements Specification
SRU	Shop Replaceable Unit

SSA	Software (also Special or System) Support Activity, Source Selection Activity, or Solid State Amplifier
SSB	Single Side Band
SSI	Small Scale Integration
SSJ	Self Screening Jamming
SSM	Surface-to-Surface Missile
SSRO	Sector Scan Receive Only
SSW	Swept Square Wave
S&T	Science and Technology
STANAG	Standardization Agreement (NATO)
STAR	System Threat Assessment Report
stat	Statute
STBY	Standby
STC	Sensitivity Time Control, Short Time Constant or SHAPE Technical Center
STD	Software Test Description, Standard, or Sexually Transmitted Disease
STE	Standard Test Equipment
STOVL	Short Takeoff and Vertical Landing
STP	Software Test Plan, or Standard Temperature and Pressure (0°C at 1 atmosphere)
STR	Software (also System) Trouble Report
STT	Single Target Track
STU	Secure Telephone Unit
SUBSAM	Subsurface-to-Air Missile
SUT	System Under Test
S/W	Software (also SW)
SWAP	Size, Weight, And Power
SWC	Scan With Compensation
SWM	Swept Wave Modulation
SYSCOM	Systems Command

t	Time (also T)
T	Time (also t), tera (10^{12} multiplier), Temperature, or Telsa
TA	Target Acquisition or Terrain Avoidance
TAAF	Test, Analyze, and Fix
TAC	Tactical Air Command (now ACC)
TACAIR	Tactical Aircraft
TACAMO	Take Charge and Move Out (airborne strategic VLF communications relay system)
TACAN	Tactical Air Navigation
TACDS	Threat Adaptive Countermeasures Dispensing System
TACTS	Tactical Aircrew Combat Training System
TAD	Threat Adaptive Dispensing, Temporary Additional (also Active) Duty, or Tactical Air Direction
T&E	Test & Evaluation
TALD	Tactical Air Launched Decoy
TAMPS	Tactical Automated Mission Planning System
TAR	Target Acquisition Radar or Training Administrative Reserve
TAS	True Airspeed
TBA	To Be Announced
TBD	To Be Determined
TBMD	Theater Ballistic Missile Defense
TD	Technical Directive (also Director)
TDD	Target Detection Device
TDM	Time Division Multiplexing
TE	Transverse Electric
TEA	Technology Exchange Agreement
TEAMS	Tactical EA-6B Mission Support
TECHEVAL	Technical Evaluation
TEL	Transporter Erector Launcher
TEM	Transverse Electromagnetic
TEMP	Test and Evaluation Master Plan
TEMPEST	Not an acronym. Certification of reduced electromagnetic radiation for security considerations
TENA	Training Enabling Architecture
TERPES	Tactical Electronic Reconnaissance Processing and Evaluation System
TGT	Target
TIM	Technical Interchange Meeting
TM	Telemetry, Transverse Magnetic, or Technical Manual
TMD	Theater Missile Defense
TNC	Threaded Navy Connector
TOA	Time of Arrival
TOJ	Track on Jam
TOO	Target of Opportunity (HARM operating mode)
TOR	Tentative (also Tactical) Operational Requirement or Time of Receipt
TOS	Time on Station
TOT	Time on Target
TOW	Tube-Launched, Optically-Tracked, Wire-guided
TPI	Test Program Instruction
TPS	Test Program Set or Test Pilot School
TPWG	Test Plan Working Group
TQM	Total Quality Management
T/R	Transmit / Receive
TRB	Technical Review Board
TRD	Test Requirements Document
TREE	Transient Radiation Effects on Electronics

TRF	Tuned Radio Frequency
TRR	Test Readiness Review
TS	Top Secret
TSS	Tangential Sensitivity
TSSAM	Tri-Service Standoff Attack Weapon
TT	Target Track
TTI	Time To Impact/Intercept
TTG	Time-to-Go
TTL	Transistor-Transistor Logic
TTR	Target Tracking Radar
TV	Television
TVC	Thrust Vector Control
TWS	Track While Scan or Tail Warning System
TWSRO	Track While Scan on Receive Only
TWT	Traveling Wave Tube
TWTA	Traveling Wave Tube Amplifier
Tx	Transmit
TYCOM	Type Commander

u or µ	micron / micro (10^{-6} multiplier)
U	Unclassified, Unit, or Unknown (on RWR display)
UAS	Unmanned Aerial System
UAV	Unmanned (also uninhabited) Air (or Aerial) Vehicle
UCAV	Uninhabited Combat Air Vehicle (new USAF term for UAV)
UDF	User Data File
UDFG	User Data File Generator
UDM	User Data Module
UHF	Ultra High Frequency (300 MHz to 3 GHz)
ULF	Ultra Low Frequency (3 to 30 Hz)
µm	Micrometer
UN	United Nations
UNK	Unknown (also U)
UPC	Unique Planning Component
UPS	Uninterruptable Power Supply
us or µs	Microseconds
U.S.	United States
USA	United States of America or United States Army
USAF	United States Air Force
USMC	United States Marine Corps
USN	United States Navy
UTA	Uninhabited Tactical Aircraft
UUT	Unit Under Test
UV	Ultraviolet

v	Volts (also V), Velocity (also V or v_t)
V	Volts (also v), Velocity (also v or v_t), Volume, or Roman Numeral for five
VA	Veterans Administration, Volt-Amperes
VAQ	Prefix for Navy tactical EW squadron
V&V	Validation and Verification
VCO	Voltage Controlled Oscillator
Vdc or VDC	Volts Direct Current
VDT	Video Display Terminal
VECP	Value Engineering Change Proposal
VF	Prefix for Navy fighter squadron
VFO	Variable Frequency Oscillator
VFR	Visual Flight Rules
VGPO	Velocity Gate Pull Off
VGS	Velocity Gate Stealer
VGWO	Velocity Gate Walk Off
VHF	Very High Frequency (30 - 300 MHz)
VHSIC	Very High Speed Integrated Circuit
VID	Visual Identification
VLF	Very Low Frequency (3 to 30 kHz)
VLSI	Very Large Scale Integration
VLSIC	Very Large Scale Integrated Circuit
VMAQ	Prefix for Marine Tactical EW Squadron
VP	Prefix for Navy patrol squadron
VQ	Prefix for Navy special mission (usually reconnaissance) squadron
VRAM	Video Random Access Memory
VS or vs	Velocity Search or Versus (also vs.)
V/STOL	Vertical/Short Take-off and Landing (also VSTOL)
vt	Velocity (also V or v)
VTOL	Vertical Takeoff and Landing
VSWR	Voltage Standing Wave Ratio
VVA	Voltage Variable Attenuator

W	Watts, Weight, or West
W&T	Warning & Targeting
WARM	Wartime Reserve Mode
wb	Weber (magnetic flux)
WBS	Work Breakdown Structure
WC	Waveguide, circular
WFT	Windowed Fourier Transform
WGIRB	Working Group on Infrared Background
WORM	Write Once Read Many [times] (Refers to optical disks)
WOW	Weight on/off Wheels (also WonW or WoffW)
WPN	Weapons Procurement, Navy or Weapon

WR	Waveguide, rectangular
WRA	Weapon Replaceable Assembly
WRD	Waveguide, rectangular double ridged
WSSA	Weapons System Support Activity
WVR	Within Visual Range
x	Multiplication symbol
X	Reactance, Experimental, Extraordinary, Roman Numeral for ten, or X axis
X-EYE	Cross Eye
X-POL	Cross Polarization
XMIT	Transmit
XMTR	Transmitter
Y	Yes or Y-Axis
YAG	Yttrium-Aluminum Garnet
yd	Yard
YIG	Yttrium-Iron Garnet
Z	Impedance, Zenith, or Z-Axis
1xLR, 2xLR	One (or two or three etc.) Times Lethal Range
1v1 or 1-v-1	One versus One (Aerial engagement)
2D	Two Dimension
3D	Three Dimension
3M	Navy Maintenance and Material Management System

FUNDAMENTALS

Constants, Conversions, and Characters 2-1
Mathematical Notation 2-2
Frequency Spectrum 2-3
Decibel (dB) 2-4
Duty Cycle 2-5
Doppler Shift 2-6
Electronic Formulas 2-7
Missile and Electronic Equipment Designations 2-8
Radar Horizon / Line of Sight 2-9
Propagation Time / Resolution 2-10
Modulation 2-11
Transforms / Wavelets 2-12

This page intentionally left blank.

CONSTANTS, CONVERSIONS, and CHARACTERS

DECIMAL MULTIPLIER PREFIXES

Prefix	Symbol	Multiplier
exa	E	10^{18}
peta	P	10^{15}
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
hecto	h	10^{2}
deka	da	10^{1}
deci	d	10^{-1}
centi	c	10^{-2}
milli	m	10^{-3}
micro	µ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

EQUIVALENCY SYMBOLS

Symbol	Meaning
∝	Proportional
~	Roughly equivalent
≈	Approximately
≅	Nearly equal
=	Equal
≡	Identical to, defined as
≠	Not equal
>>	Much greater than
>	Greater than
≥	Greater than or equal to
<<	Much less than
<	Less than
≤	Less than or equal to
∴	Therefore
°	Degrees
′	Minutes or feet
“	Seconds or inches

UNITS OF LENGTH

1 inch (in) = 2.54 centimeters (cm)
1 foot (ft) = 30.48 cm = 0.3048 m
1 yard (yd) ≅ 0.9144 meter
1 meter (m) ≅ 39.37 inches

1 kilometer (km) ≅ 0.54 nautical mile
≅ 0.62 statute mile
≅ 1093.6 yards
≅ 3280.8 feet

1 statute mile ≅ 0.87 nautical mile
(sm or stat. mile) ≅ 1.61 kilometers
= 1760 yards
= 5280 feet

1 nautical mile ≅ 1.15 statute miles
(nm or naut. mile) ≅ 1.852 kilometers
≅ 2025 yards
≅ 6076 feet
1 furlong = 1/8 mi (220 yds)

UNITS OF SPEED

1 foot/sec (fps) ≅ 0.59 knot (kt)*
≅ 0.68 stat. mph
≅ 1.1 kilometers/hr

1000 fps ≈ 600 knots

1 kilometer/hr ≅ 0.54 knot
(km/hr) ≅ 0.62 stat. mph
≅ 0.91 ft/sec

1 mile/hr (stat.) ≅ 0.87 knot
(mph) ≅ 1.61 kilometers/hr
≅ 1.47 ft/sec

1 knot* ≅ 1.15 stat. mph
≅ 1.69 feet/sec
≅ 1.85 kilometer/hr
≅ 0.515 m/sec

*A knot is 1 nautical mile per hour.

UNITS OF VOLUME

1 gallon ≅ 3.78 liters
≅ 231 cubic inches
≅ 0.1335 cubic ft
≅ 4 quarts
≅ 8 pints

1 fl ounce ≅ 29.57 cubic centimeter (cc) or milliliters (ml)

1 in^3 ≅ 16.387 cc

UNITS OF AREA

1 sq meter ≅ 10.76 sq ft

1 sq in ≅ 645 sq millimeters (mm)
= 1,000,000 sq mil
1 mil = 0.001 inch

1 acre = 43,560 sq ft

UNITS OF WEIGHT

1 kilogram (kg) ≅ 2.2 pounds (lbs)

1 pound ≅ 0.45 Kg
= 16 ounce (oz)
1 oz = 437.5 grains
1 carat ≅ 200 mg
1 stone (U.K.) ≅ 6.36 kg

NOTE: These are the U.S. customary (avoirdupois) equivalents, the troy or apothecary system of equivalents, which differ markedly, was used long ago by pharmacists.

UNITS OF POWER / ENERGY

1 H.P. = 33,000 ft-lbs/min
= 550 ft-lbs/sec
≅ 746 Watts
≅ 2,545 BTU/hr
(BTU = British Thermal Unit)

1 BTU ≅ 1055 Joules
≅ 778 ft-lbs
≅ 0.293 Watt-hrs

SCALES

OCTAVES

"N" Octaves = Freq to Freq x 2^N
i.e., One octave would be 2 to 4 GHz
Two Octaves would be 2 to 8 GHz
Three octaves would be 2 to 16 GHz

DECADES

"N" Decades = Freq to Freq x 10^N
i.e., One decade would be 1 to 10 MHz
Two decades would be 1 to 100 MHz
Three decades would be 1 to 1,000 MHz

TEMPERATURE CONVERSIONS

°F = (9/5) °C + 32
°C = (5/9)(°F - 32)
°K = °C + 273.16
°F = (9/5)(°K - 273) + 32
°C = °K - 273.16
°K = (5/9)(°F - 32) + 273

UNITS OF TIME

1 year = 365.2 days
1 fortnight = 14 nights (2 weeks)
1 century = 100 years
1 millennium = 1,000 years

NUMBERS

1 decade = 10
1 Score = 20
1 Billion = 1 x 10^9 (U.S.) (thousand million)
= 1 x 10^{12} (U.K.) (million million)

RULE OF THUMB FOR ESTIMATING DISTANCE TO LIGHTNING / EXPLOSION:

km: Divide 3 into the number of seconds which have elapsed between seeing the flash and hearing the noise.
miles: Multiply 0.2 times the number of seconds which have elapsed between seeing the flash and hearing the noise.

Note: Sound vibrations cause a change of density and pressure within a media, while electromagnetic waves do not. An audio tone won't travel through a vacuum but can travel at 1,100 ft/sec through air. When picked up by a microphone and used to modulate an EM signal, the modulation will travel at the speed of light.

Physical Constant	Quoted Value	S*	SI unit	Symbol
Avogadro constant	6.0221367 x 10^{23}	36	mol^{-1}	N_A
Bohr magneton	9.2740154 x 10^{-24}	31	J•T^{-1}	μ_B
Boltzmann constant	1.380658 x 10^{-23}	12	J•K^{-1}	k(=R N_A)
Electron charge	1.602177 33 x 10^{-19}	49	C	-e
Electron specific charge	-1.758819 62 x 10^{11}	53	C•kg^{-1}	-e/m_e
Electron rest mass	9.1093897 x 10^{-31}	54	kg	m_e
Faraday constant	9.6485309 x 10^{4}	29	C•mol^{-1}	F
Gravity (Standard Acceleration)	9.80665 or 32.174	0	m/sec^{2} ft/sec^{2}	g
Josephson frequency to voltage ratio	4.8359767 x 10^{14}	0	Hz•V^{-1}	2e/hg
Magnetic flux quantum	2.06783461 x 10^{-15}	61	Wb	φ_o
Molar gas constant	8.314510	70	J•mol^{-1}•K^{-1}	R
Natural logarithm base	≅ 2.71828	-	dimensionless	*e*
Newtonian gravitational constant	6.67259 x 10^{-11}	85	m^{3}•kg^{-1}•s^{-2}	G or K
Permeability of vacuum	4π x 10^{-7}	d	H/m	μ_o
Permittivity of vacuum	≅ 8.8541878 x 10^{-12}	d	F/m	ε_o
Pi	≅ 3.141592654		dimensionless	π
Planck constant	6.62659 x 10^{-34}	40	J•s	h
Planck constant/2π	1.05457266 x 10^{-34}	63	J•s	h(=h2π)
Quantum of circulation	3.63694807 x 10^{-4}	33	J•s•kg^{-1}	h/2m_e
Radius of earth (Equatorial)	6.378 x 10^{6} or 3963		m miles	
Rydberg constant	1.0973731534 x 10^{7}	13	m^{-1}	R_χ
Speed of light	2.9979246 x 10^{8}	1	m•s^{-1}	c
Speed of sound (dry air @ std press & temp)	331.4	-	m•s^{-1}	-
Standard volume of ideal gas	22.41410 x 10^{-3}	19	m^{3}•mol^{-1}	V_m
Stefan-Boltzmann constant	5.67051 x 10^{-8}	19	W•K^{-4} •m^{-2}	σ

* S is the one-standard-deviation uncertainty in the last units of the value, d is a defined value.
(A standard deviation is the square root of the mean of the sum of the squares of the possible deviations)

THE SPEED OF LIGHT			
ACTUAL	UNITS	RULE OF THUMB	UNITS
$\cong 2.9979246 \times 10^8$	m/sec	$\cong 3 \times 10^8$	m/sec
$\cong 299.79$	m/μsec	$\cong 300$	m/μsec
$\cong 3.27857 \times 10^8$	yd/sec	$\cong 3.28 \times 10^8$	yd/sec
$\cong 5.8275 \times 10^8$	NM/hr	$\cong 5.8 \times 10^8$	NM/hr
$\cong 1.61875 \times 10^5$	NM/sec	$\cong 1.62 \times 10^5$	NM/sec
$\cong 9.8357105 \times 10^8$	ft/sec	$\cong 1 \times 10^9$	ft/sec
$\cong 9.8357105 \times 10^2$	ft/μsec	$\cong 1 \times 10^3$	ft/μsec

SPEED OF LIGHT IN VARIOUS MEDIUMS

The speed of EM radiation through a substance such as cables is defined by the following formula:

$$V = c/(\mu_r \varepsilon_r)^{1/2}$$

Where: μ_r = relative permeability
ε_r = relative permittivity

The real component of ε_r = dielectric constant of medium.

EM propagation speed in a typical cable might be 65-90% of the speed of light in a vacuum.

APPROXIMATE SPEED OF SOUND (MACH 1)

Sea Level (CAS/TAS)		**36,000 ft*** (TAS)	(CAS)
1230 km/hr	Decreases	1062 km/hr	630 km/hr
765 mph	Linearly	660 mph	391 mph
665 kts	To ⇒	573 kts	340 kts

* The speed remains constant until 82,000 ft, when it increases linearly to 1215 km/hr (755 mph, 656 kts) at 154,000 ft. Also see Section 8-2 for discussion of Calibrated Air Speed (CAS) and True Airspeed (TAS) and a plot of the speed of sound vs. altitude.

SPEED OF SOUND IN VARIOUS MEDIUMS

Substance	Speed (ft/sec)
Vacuum	Zero
Air	1,100
Fresh Water	4,700
Salt Water	4,900
Glass	14,800
Steel	20,000

DECIMAL / BINARY / HEX CONVERSION TABLE

Decimal	Binary	Hex	Decimal	Binary	Hex	Decimal	Binary	Hex
1	00001	01h	11	01011	0Bh	21	10101	15h
2	00010	02h	12	01100	0Ch	22	10110	16h
3	00011	03h	13	01101	0Dh	23	10111	17h
4	00100	04h	14	01110	0Eh	24	11000	18h
5	00101	05h	15	01111	0Fh	25	11001	19h
6	00110	06h	16	10000	10h	26	11010	1Ah
7	00111	07h	17	10001	11h	27	11011	1Bh
8	01000	08h	18	10010	12h	28	11100	1Ch
9	01001	09h	19	10011	13h	29	11101	1Dh
10	01010	0Ah	20	10100	14h	30	11110	1Eh

When using hex numbers it is always a good idea to use "h" as a suffix to avoid confusion with decimal numbers.

DECIMAL TO HEX CONVERSION

Both the following methods must use long division. Method one computes the digits from right to left while method two works from left to right.

Method one: To convert a decimal number above 16 to hex, divide the number by 16, then record the integer resultant and the remainder. Convert the remainder to hex and write this down - this will become the far right digit of the final hex number. Divide the integer you obtained by 16, and again record the new integer result and new remainder. Convert the remainder to hex and write it just to the left of the first decoded number. Keep repeating this process until dividing results in only a remainder. This will become the left-most character in the hex number (i.e., to convert 60 (decimal) to hex we have 60/16 = 3 with 12 remainder). 12 is C (hex) - this becomes the right most character. Then 3/16=0 with 3 remainder. 3 is 3 (hex). This becomes the next (and final) character to the left in the hex number, so the answer is 3C.

Method two: Use table of powers to work the digits from left to right:

For example: Here is your Decimal Number - 9379

Step 1 - Set up your chart:

65536	4096	256	16	1
16^4	16^3	16^2	16^1	16^0

Step 2 - Look in the table for the highest divisible number in the chart.
9379 / 4096 = 2 (the left-most Hex digit)
Must use long division to calculate the remainder (1187)

Step 3 - Divide the remainder with its highest divisible number in the chart:
1187 / 256 = 4 (the next digit to the right)
Must use long division to calculate the remainder (163)

Step 4 - Divide the remainder with its highest divisible number in the chart:
163 / 16 = 10 (or "A" from table L-1) (the next digit to the right)
Must use long division to calculate the remainder (3)

Step 5 - The remainder will not divide: remainder = 3 (the right-most Hex digit)

65536	4096	256	16	1
16^4	16^3	16^2	16^1	16^0
	2	4	A	3

HEX TO DECIMAL CONVERSION

To convert a hex number to decimal, multiply each hex digit converted to decimal by the decimal equivalent of the hex power represented and add the results.

For example: Here is your Hex Number - 24A3

Step 1 - Set up your chart:

4096	256	16	1
16^3	16^2	16^1	16^0

Step 2 - Place the numbers in a table:

4096	256	16	1
16^3	16^2	16^1	16^0
2	4	A	3

Step 3 - Multiply the Hex number times the power value:
2 x 4096 = 8192
4 x 256 = 1024
A(10) x 16 = 160
3 x 1 = 3

Step 4 - Add up your values:
Decimal value is 9379

GREEK ALPHABET

Case		Greek Alphabet Name	English Equivalent
Upper	Lower		
Α	α	alpha	a
Β	β	beta	b
Γ	γ	gamma	g
Δ	δ	delta	d
Ε	ε	epsilon	e
Ζ	ζ	zeta	z
Η	η	eta	e
Θ	θ , ‘	theta	th
Ι	ι	iota	i
Κ	κ	kappa	k
Λ	λ	lambda	l
Μ	μ	mu	m

Case		Greek Alphabet Name	English Equivalent
Upper	Lower		
Ν	ν	nu	n
Ξ	ξ	xi	x
Ο	ο	omicron	o
Π	π	pi	p
Ρ	ρ	rho	r
Σ	σ	sigma	s
Τ	τ	tau	t
Υ	υ	upsilon	u
Φ	φ	phi	ph
Χ	χ	chi	ch
Ψ	ψ	psi	ps
Ω	ω	omega	o

LETTERS FROM THE GREEK ALPHABET COMMONLY USED AS SYMBOLS

Symbol	Name	Use
α	alpha	space loss, angular acceleration, or absorptance
β	beta	3 dB bandwidth or angular field of view [radians]
Γ	Gamma	reflection coefficient
γ	gamma	electric conductivity, surface tension, missile velocity vector angle, or gamma ray
Δ	Delta	small change or difference
δ	delta	delay, control forces and moments applied to missile, or phase angle
ε	epsilon	emissivity [dielectric constant] or permittivity [farads/meter]
η	eta	efficiency or antenna aperture efficiency
Θ	Theta	angle of lead or lag between current and voltage
θ or ‘	theta	azimuth angle, bank angle, or angular displacement
Λ	Lambda	acoustic wavelength or rate of energy loss from a thermocouple
λ	lambda	wavelength or Poisson Load Factor
μ	mu	micro 10^{-6} [micron], permeability [henrys/meter], or extinction coefficient [optical region]

v	nu	frequency
π	pi	3.141592654+
ρ	rho	charge/mass density, resistivity [ohm-meter], VSWR, or reflectance
Σ	Sigma	algebraic sum
σ	sigma	radar cross section [RCS], Conductivity [1/ohm-meter], or Stefan-Boltzmann constant
Τ	Tau	VSWR reflection coefficient
τ	tau	pulse width, atmospheric transmission, or torque
Φ	Phi	magnetic/electrical flux, radiant power [optical], or Wavelet’s smooth function [low pass filter]
φ	phi	phase angle, angle of bank, or beam divergence [optical region]
Ψ	Psi	time-dependent wave function or Wavelet’s detail function [high pass filter]
ψ	psi	time-independent wave function, phase change, or flux linkage [weber]
Ω	Omega	Ohms [resistance] or solid angle [optical region]. Note: inverted symbol is conductance [mhos]
ω	omega	carrier frequency in radians per second

MORSE CODE and PHONETIC ALPHABET

A - alpha	• –	J - juliett	• – –	S - sierra	• • •	1	• – – – –
B - bravo	– • • •	K - kilo	– • –	T - tango	–	2	• • – – –
C - charlie	– • – •	L - lima	• – • •	U - uniform	• • –	3	• • • – –
D - delta	– • •	M - mike	– –	V - victor	• • • –	4	• • • • –
E - echo	•	N - november	– •	W - whiskey	• – –	5	• • • • •
F - foxtrot	• • – •	O - oscar	– – –	X - x-ray	– • • –	6	– • • • •
G - golf	– – •	P - papa	• – •	Y - yankee	– • – –	7	– – • • •
H - hotel	• • • •	Q - quebec	– – • –	Z - zulu	– – • •	8	– – – • •
I - india	• •	R - romeo	• – •	0	– – – – –	9	– – – – •

Note: The International Maritime Organization agreed to officially stop Morse code use by February 1999, however use may continue by ground based amateur radio operators. (The U.S. Coast Guard discontinued its use in 1995.)

BASIC MATH / GEOMETRY REVIEW

EXPONENTS

$a^x a^y = a^{x+y}$

$a^x / a^y = a^{x-y}$

$(a^x)^y = a^{xy}$

$a^0 = 1$

Example:

$$\frac{x}{\sqrt{x}} = x \bullet x^{-\frac{1}{2}} = x^{(1-\frac{1}{2})} = x^{\frac{1}{2}} = \sqrt{x}$$

LOGARITHMS

$\log(xy) = \log x + \log y$

$\log(x/y) = \log x - \log y$

$\log(x^N) = N \log x$

If $z = \log x$ then $x = 10^z$

Examples: $\log 1 = 0$
$\log 1.26 = 0.1$; $\log 10 = 1$

if $10 \log N = dB\#$,
then $10^{(dB\#/10)} = N$

TRIGONOMETRIC FUNCTIONS

$\sin x = \cos(x-90°)$

$\cos x = -\sin(x-90°)$

$\tan x = \sin x / \cos x = 1 / \cot x$

$\sin^2 x + \cos^2 x = 1$

A radian is the angular measurement of an arc which has an arc length equal to the radius of the given circle, therefore there are 2π radians in a circle. One radian = 360°/2π = 57.296....°

ELLIPSE

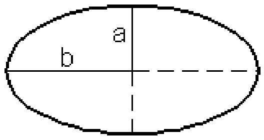

Area = π a b

Approx circumference

$$= 2\pi\sqrt{\frac{a^2 + b^2}{2}}$$

RECTANGLULAR SOLID

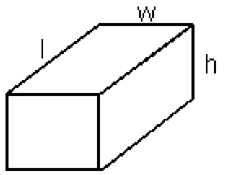

Area = lw

Volume = lwh

CYLINDER

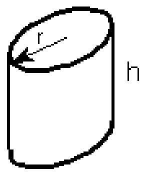

Volume − $\pi r^2 h$

Lateral surface area _ $2\pi rh$

ANGLES

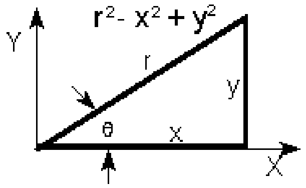
Y
r²- x² + y²
r
y
θ
x
X

Sin θ = y/r   Cos θ = x/r

Tan θ = y/x

SPHERE

Surface area = $4\pi r^2$

Volume = $4/3\ \pi r^3$

Cross Section (circle) Area = πr^2

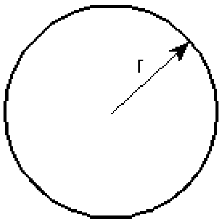

TRIANGLES

Angles: A + B + C = 180°

$c^2 = a^2 + b^2 - 2ab \cos C$

Area = 1/2 hc = 1/2 ab sin C

$b = \sqrt{e^2 + h^2}$

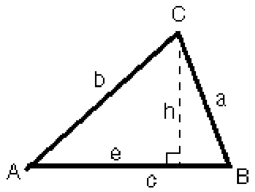

DERIVATIVES

Assume: a = fixed real #; u, v & w are functions of x

d(a)/dx = 0 ; d(sin u)/dx = du(cos u)/dx

d(x)/dx = 1 ; d(cos v)/dx = -dv(sin v)/dx

d(uvw)/dx = uvdw/dx + vwdu/dx + uwdv/dx +...etc

INTEGRALS

Note: All integrals should have an arbitrary constant of integration added, which is left off for clarity

Assume: a = fixed real #; u, & v are functions of x

$\int a\,dx = ax$ and $\int a\, f(x)dx = a\int f(x)dx$

$\int (u + v)dx = \int u\,dx + \int v\,dx$; $\int e^x dx = e^x$

$\int(\sin ax)dx = -(\cos ax)/a$; $\int(\cos ax)dx = (\sin ax)/a$

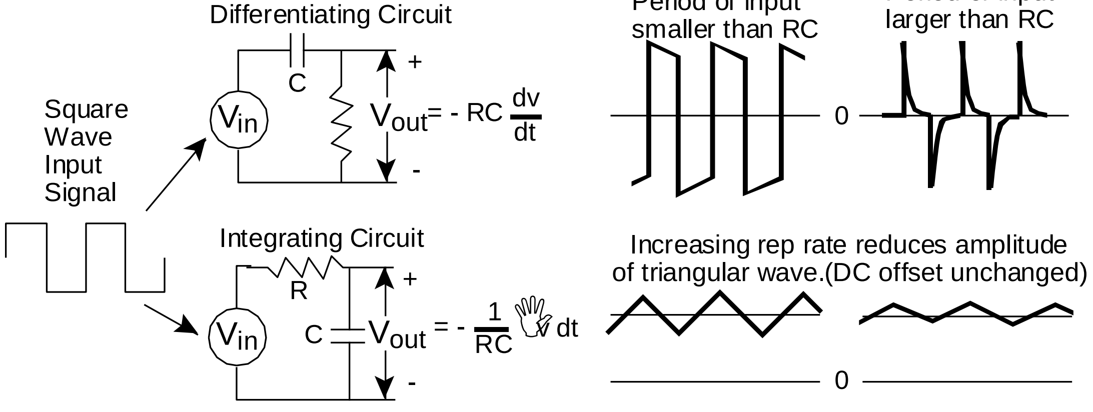
Differentiating Circuit
C
Vin
+
Vout = - RC dv/dt
-
Square Wave Input Signal
Integrating Circuit
R
Vin
C
+
Vout = - 1/RC ✋ dt
-
Period of input smaller than RC
Period of input larger than RC
0
Increasing rep rate reduces amplitude of triangular wave.(DC offset unchanged)
0

MATHEMATICAL NOTATION

The radar and EW communities generally accept some commonly used notation for the various parameters used in radar and EW calculations. For instance, "P" is almost always power and "G" is almost always gain. Textbooks and reference handbooks will usually use this notation in formulae and equations.

A significant exception is the use of "α" for space loss. Most textbooks do not develop the radar equation to its most usable form as does this reference handbook, therefore the concept of "α" just isn't covered.

Subscripts are a different matter. Subscripts are often whatever seems to make sense in the context of the particular formula or equation. For instance, power may be "P", "P_T", "P_t", or maybe "P_1". In the following list, generally accepted notation is given in the left hand column with no subscripts. Subscripted notation in the indented columns is the notation used in this handbook and the notation often used in the EW community.

α = Space loss
α_1 = One way space loss, transmitter to receiver
α_2 = Two way space loss, xmtr to target (including radar cross section) and back to the rcvr
α_{1t} = One way space loss, radar transmitter to target, bistatic
α_{1r} = One way space loss, target to radar receiver, bistatic

Other notation such as α_{tm} may be used to clarify specific losses, in this case the space loss between a target and missile seeker, which could also be identified as α_{1r}.

A = Antenna aperture (capture area)
A_e = Effective antenna aperture
Å = Angstrom

B = Bandwidth (to 3dB points)
B_{IF} = 3 dB IF bandwidth of the receiver (pre-detection)
B_J = Bandwidth of the jamming spectrum
B_{MHz} = 3 dB bandwidth in MHz
B_N = Equivalent noise bandwidth, a.k.a. B
B_V = 3 dB video bandwidth of the receiver (post-detection) (Subscript V stands for video)
BF = Bandwidth reduction factor (jamming spectrum wider than the receiver bandwidth)
BW = Beamwidth (to 3 dB points)

c = Speed of Light

f = Frequency (radio frequency)
f_c = Footcandle (SI unit of illuminance)
f_D = Doppler frequency
f_R = Received frequency
f_T = Transmitted frequency

G = Gain
G_t = Gain of the transmitter antenna
G_r = Gain of the receiver antenna
G_{tr} = Gain of the transmitter/receiver antenna (monostatic radar)
G_J = Gain of the jammer
G_{JA} = Gain of the jammer antenna
G_{JT} = Gain of the jammer transmitter antenna
G_{JR} = Gain of the jammer receiver antenna
G_σ = Gain of reflected radar signal due to radar cross section

h = Height or Planks constant
h_{radar} = Height of radar
h_{target} = Height of target

J = Jamming signal (receiver input)
J_1 = Jamming signal (constant gain jammer)
J_2 = Jamming signal (constant power jammer)
J/S = Jamming to signal ratio (receiver input)

k = Boltzmann constant
$K_{1,2,3,4}$ = Proportionality constants, see Sections 4-3, 4-4, 4-5, and 4-1 respectively.

λ = Lambda, Wavelength or Poisson factor

L = Loss (due to transmission lines or circuit elements)

N = Receiver equivalent noise input (kT_oB)
NF = Noise figure

P = Power
P_d = Probability of detection
P_D = Power density
P_J = Power of a jammer transmitter
P_n = Probability of false alarm
P_r = Power received
P_t = Power of a transmitter

R = Range (straight line distance)
R_1 = Bistatic radar transmitter to target range
R_2 = Bistatic radar target to receiver range
R_J = Range of jammer to receiver (when separate from the target)
R_{NM} = Range in nautical miles

σ = Sigma, radar cross section (RCS)

S = Signal (receiver input)
S_R = Radar signal received by the jammer
S_{min} = Minimum receiver sensitivity

t = Time
t_{int} = Integration time
t_r = Pulse Rise Time
τ = Pulse Width

V = Velocity
V_r = Radial velocity

This page intentionally left blank.

FREQUENCY SPECTRUM

Figure 1, which follows, depicts the electromagnetic radiation spectrum and some of the commonly used or known areas. Figure 2 depicts the more common uses of the microwave spectrum. Figure 3 shows areas of the spectrum which are frequently referred to by band designations rather than by frequency.

Section 7-1 provides an additional breakdown of the EO/IR spectrum.

To convert from frequency (f) to wavelength (λ) and vice versa, recall that f = c/λ, or λ = c/f; where c = speed of light.

$$\lambda_{meter} = \frac{3x10^8}{f_{Hz}} = \frac{3x10^5}{f_{kHz}} = \frac{300}{f_{MHz}} = \frac{0.3}{f_{GHz}} \qquad \text{or} \qquad f_{Hz} = \frac{3x10^8}{\lambda_{meter}} \quad f_{kHz} = \frac{3x10^5}{\lambda_{meter}} \quad f_{MHz} = \frac{300}{\lambda_{meter}} \quad f_{GHz} = \frac{0.3}{\lambda_{meter}}$$

Some quick rules of thumb follow:

Metric:

Wavelength in cm = 30 / frequency in GHz
For example: at 10 GHz, the wavelength = 30/10 = 3 cm

English:

Wavelength in ft = 1 / frequency in GHz
For example: at 10 GHz, the wavelength = 1/10 = 0.1 ft

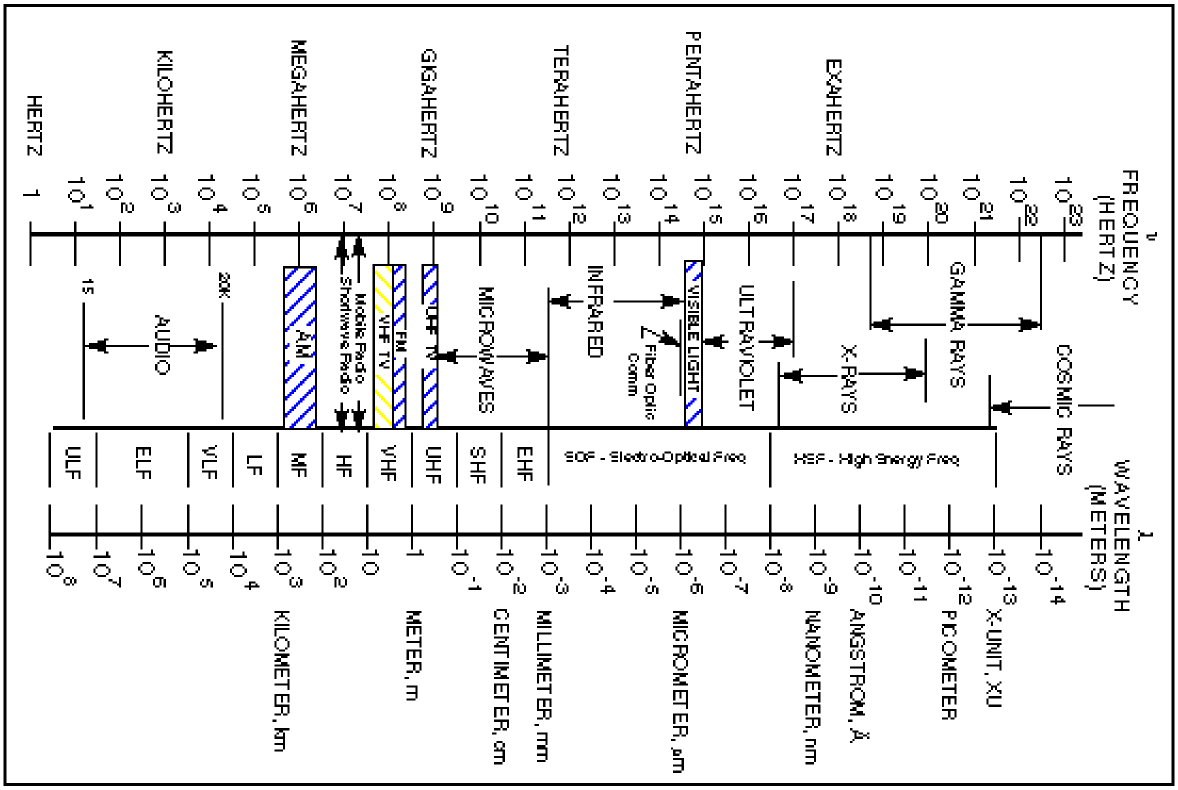

Figure 1. Electromagnetic Radiation Spectrum.

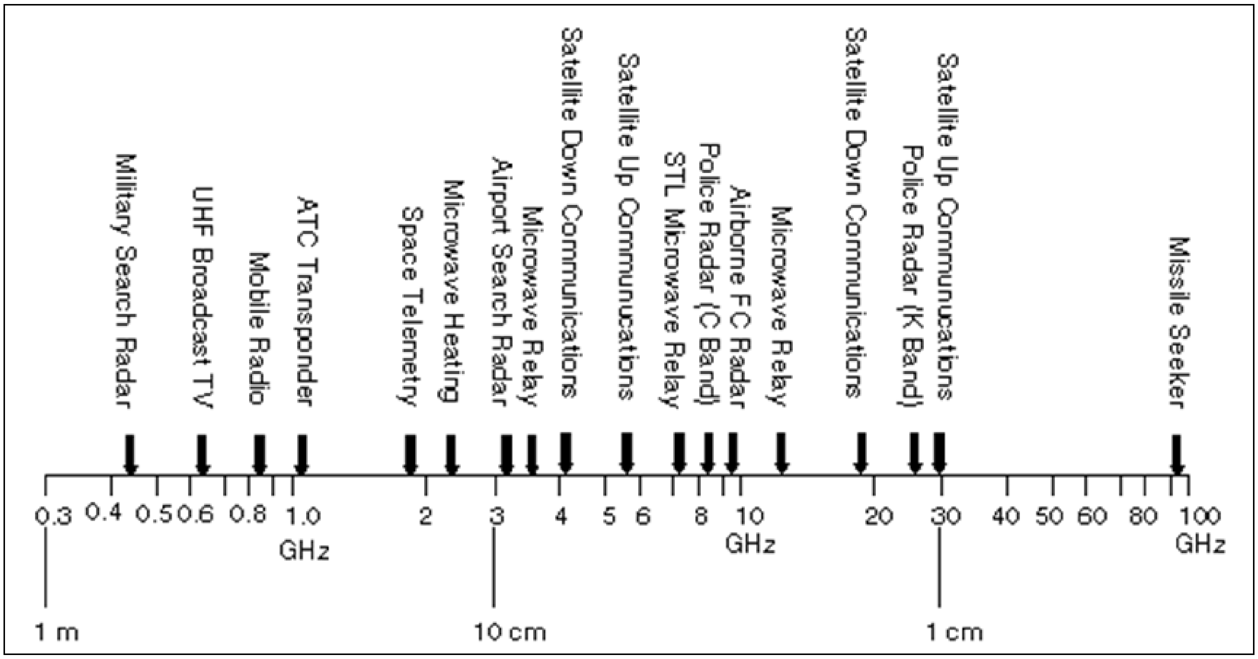
Military Search Radar
UHF Broadcast TV
Mobile Radio
ATC Transponder
Space Telemetry
Microwave Heating
Airport Search Radar
Microwave Relay
Satellite Down Communications
Satellite Up Communucations
STL Microwave Relay
Police Radar (C Band)
Airborne FC Radar
Microwave Relay
Satellite Down Communications
Police Radar (K Band)
Satellite Up Communucations
Missile Seeker
0.3
0.4
0.5
0.6
0.8
1.0
GHz
2
3
4
5
6
8
10
GHz
20
30
40
50
60
80
100
GHz
1 m
10 cm
1 cm

Figure 2. The Microwave Spectrum.

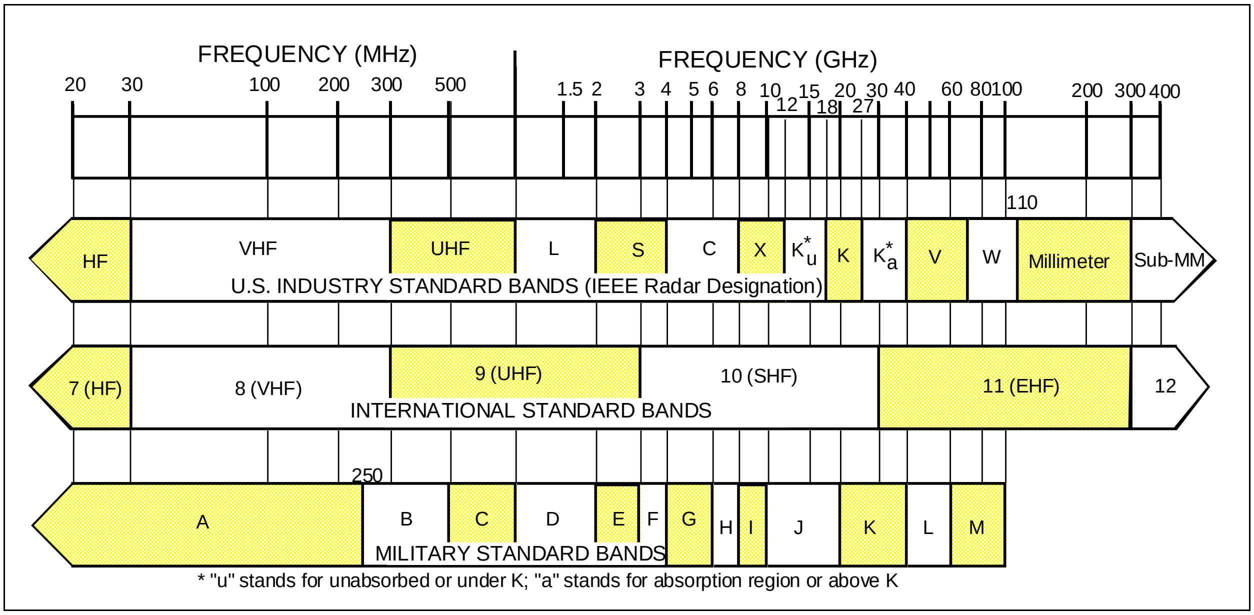
FREQUENCY (MHz)
20
30
100
200
300
500
FREQUENCY (GHz)
1.5
2
3
4
5
6
8
10
12
15
18
20
27
30
40
60
80
100
200
300
400
110
HF
VHF
UHF
L
S
C
X
K*u
K
K*a
V
W
Millimeter
Sub-MM
U.S. INDUSTRY STANDARD BANDS (IEEE Radar Designation)
7 (HF)
8 (VHF)
9 (UHF)
10 (SHF)
11 (EHF)
12
INTERNATIONAL STANDARD BANDS
250
A
B
C
D
E
F
G
H
I
J
K
L
M
MILITARY STANDARD BANDS
* "u" stands for unabsorbed or under K; "a" stands for absorption region or above K

Figure 3. Frequency Band Designations.

See Section 7, Figure 1 for a more detailed depiction of the UV and IR spectrum.

DECIBEL (dB)

The Decibel is a subunit of a larger unit called the bel. As originally used, the bel represented the power ratio of 10 to 1 between the strength or intensity i.e., power, of two sounds, and was named after Alexander Graham Bell. Thus a power ratio of 10:1 = 1 bel, 100:1 = 2 bels, and 1000:1 = 3 bels. It is readily seen that the concept of bels represents a logarithmic relationship since the logarithm of 100 to the base 10 is 2 (corresponding to 2 bels), the logarithm of 1000 to the base 10 is 3 (corresponding to 3 bels), etc. The exact relationship is given by the formula

$$\text{Bels} = \log(P_2/P_1) \qquad \textbf{[1]}$$

where P_2/P_1 represents the power ratio.

Since the bel is a rather large unit, its use may prove inconvenient. Usually a smaller unit, the Decibel or dB, is used. 10 decibels make one bel. A 10:1 power ratio, 1 bel, is 10 dB; a 100:1 ratio, 2 bels, is 20 dB. Thus the formula becomes

$$\text{Decibels (dB)} = 10 \log(P_2/P_1) \qquad \textbf{[2]}$$

The power ratio need not be greater than unity as shown in the previous examples. In equations [1] and [2], P_1 is usually the reference power. If P_2 is less than P_1, the ratio is less then 1.0 and the resultant bels or decibels are negative. For example, if P_2 is one-tenth P_1, we have

$$\text{bels} = \log(0.1/1) = -1.0 \text{ bels}$$

and $$\text{dB} = 10 \log(0.1/1) = -10 \text{ dB}.$$

It should be clearly understood that the term decibel does not in itself indicate power, but rather is a ratio or comparison between two power values. It is often desirable to express power levels in decibels by using a fixed power as a reference. The most common references in the world of electronics are the milliwatt (mW) and the watt. The abbreviation dBm indicates dB referenced to 1.0 milliwatt. One milliwatt is then zero dBm. Thus P_1 in equations [1] or [2] becomes 1.0 mW. Similarly, the abbreviation dBW indicates dB referenced to 1.0 watt, with P_2 being 1.0 watt, thus one watt in dBW is zero dBW or 30 dBm or 60 dBμW. For antenna gain, the reference is the linearly polarized isotropic radiator, dBLI. Usually the "L" and/or "I" is understood and left out.

dBc is the power of one signal referenced to a carrier signal, i.e., if a second harmonic signal at 10 GHz is 3 dB lower than a fundamental signal at 5 GHz, then the signal at 10 GHz is -3 dBc.

THE DECIBEL, ITS USE IN ELECTRONICS

The logarithmic characteristic of the dB makes it very convenient for expressing electrical power and power ratios. Consider an amplifier with an output of 100 watts when the input is 0.1 watts (100 milliwatts); it has an amplification factor of

$$P_2/P_1 = 100/0.1 = 1000$$

or a gain of:

$$10 \log(P_2/P_1) = 10 \log(100/0.1) = 30 \text{ dB}.$$

(notice the 3 in 30 dB corresponds to the number of zeros in the power ratio)

The ability of an antenna to intercept or transmit a signal is expressed in dB referenced to an isotropic antenna rather than as a ratio. Instead of saying an antenna has an effective gain ratio of 7.5, it has a gain of 8.8 dB (10 log 7.5).

A ratio of less than 1.0 is a loss, a negative gain, or attenuation. For instance, if 10 watts of power is fed into a cable but only 8.5 watts are measured at the output, the signal has been decreased by a factor of

$8.5/10 = .85$

or $10 \log(.85) = -0.7$ dB.

This piece of cable at the frequency of the measurement has a gain of -0.7 dB. This is generally referred to as a loss or attenuation of 0.7 dB, where the terms "loss" and "attenuation" imply the negative sign. An attenuator which reduces its input power by factor of 0.001 has an attenuation of 30 dB. The utility of the dB is very evident when speaking of signal loss due to radiation through the atmosphere. It is much easier to work with a loss of 137 dB rather than the equivalent factor of 2×10^{-14}.

Instead of multiplying gain or loss factors as ratios we can add them as positive or negative dB. Suppose we have a microwave system with a 10 watt transmitter, and a cable with 0.7 dB loss connected to a 13 dB gain transmit antenna. The signal loss through the atmosphere is 137 dB to a receive antenna with an 11 dB gain connected by a cable with 1.4 dB loss to a receiver. How much power is at the receiver? First, we must convert the 10 watts to milliwatts and then to dBm:

10 watts = 10,000 milliwatts

and

$10 \log (10{,}000/1) = 40$ dBm

Then

40 dBm - 0.7 dB + 13 dB - 137 dB + 11 dB - 1.4 dB = -75.1 dBm.

-75.1 dBm may be converted back to milliwatts by solving the formula:

$\text{mW} = 10^{(\text{dBm}/10)}$

giving: $10^{(-75.1/10)} = 0.00000003$ mW

Voltage and current ratios can also be expressed in terms of decibels, provided the resistance remains constant. First we substitute for P in terms of either voltage, V, or current, I. Since P=VI and V=IR we have:

$P = I^2R = V^2/R$

Thus for a voltage ratio we have:

$$\text{dB} = 10 \log[(V_2^2/R)/(V_1^2/R)] = 10 \log(V_2^2/V_1^2)$$
$$= 10 \log(V_2/V_1)^2 = 20 \log(V_2/V_1)$$

Like power, voltage can be expressed relative to fixed units, so one volt is equal to 0 dBV or 120 dBμV.

Similarly for current ratio: $\text{dB} = 20 \log(I_2/I_1)$

Like power, amperage can be expressed relative to fixed units, so one amp is equal to 0 dBA or 120 dBμA.

Decibel Formulas (where Z is the general form of R, including inductance and capacitance)

When impedances are equal: $dB = 10\log\frac{P2}{P1} = 20\log\frac{E2}{E1} = 20\log\frac{I2}{I1}$

When impedances are unequal: $dB = 10\log\frac{P2}{P1} = 20\log\frac{E2\sqrt{Z1}}{E1\sqrt{Z2}} = 20\log\frac{I2\sqrt{Z2}}{I1\sqrt{Z1}}$

SOLUTIONS WITHOUT A CALCULATOR

Solution of radar and EW problems requires the determination of logarithms (base 10) to calculate some of the formulae. Common “four function” calculators do not usually have a log capability (or exponential or fourth root functions either). Without a scientific calculator (or math tables or a Log-Log slide rule) it is difficult to calculate any of the radar equations, simplified or “textbook.” The following gives some tips to calculate a close approximation without a calculator.

DECIBEL TABLE

DB	Power Ratio	Voltage or Current Ratio	DB	Power Ratio	Voltage or Current Ratio
0	1.00	1.00	10	10.0	3.16
0.5	1.12	1.06	15	31.6	5.62
1.0	1.26	1.12	20	100	10
1.5	1.41	1.19	25	316	17.78
2.0	1.58	1.26	30	1,000	31.6
3.0	2.00	1.41	40	10,000	100
4.0	2.51	1.58	50	10^5	316
5.0	3.16	1.78	60	10^6	1,000
6.0	3.98	2.00	70	10^7	3,162
7.0	5.01	2.24	80	10^8	10,000
8.0	6.31	2.51	90	10^9	31,620
9.0	7.94	2.82	100	10^{10}	10^5

For dB numbers which are a multiple of 10

An easy way to remember how to convert dB values that are a multiple of 10 to the absolute magnitude of the power ratio is to place a number of zeros equal to that multiple value to the right of the value 1.
40 dB = 10,000 : 1 (for power)

Minus dB moves the decimal point that many places to the left of 1.
-40 dB = 0.0001 : 1 (for power)

For voltage or current ratios, if the multiple of 10 is even, then divide the multiple by 2, and apply the above rules.
40 dB = 100 : 1 (for voltage)
-40 dB = 0.01 : 1

If the power in question is not a multiple of ten, then some estimation is required. The following tabulation lists some approximations. Some would be useful to memorize.

DB RULES OF THUMB

Multiply Current / Voltage By			Multiply Power By:	
if +dB	if -dB	**dB**	if +dB	if -dB
1	1	0	1	1
1.12	0.89	1	1.26	0.8
1.26	0.79	2	1.58	0.63
1.4	0.707	3	2	0.5
2.0	0.5	6	4	0.25
2.8	0.35	9	8	0.125
3.16	0.316	10	10	0.1
4.47	0.22	13	20	0.05
10	0.1	20	100	0.01
100	0.01	40	10,000	0.0001

You can see that the list has a repeating pattern, so by remembering just three basic values such as one, three, and 10 dB, the others can easily be obtained without a calculator by addition and subtraction of dB values and multiplication of corresponding ratios.

Example 1:

A 7 dB increase in power (3+3+1) dB is an increase of (2 x 2 x 1.26) = 5 times whereas
A 7 dB decrease in power (-3-3-1) dB is a decrease of (0.5 x 0.5 x 0.8) = 0.2.

Example 2: Assume you know that the ratio for 10 dB is 10, and that the ratio for 20 dB is 100 (doubling the dB increases the power ratio by a factor of ten), and that we want to find some intermediate value.

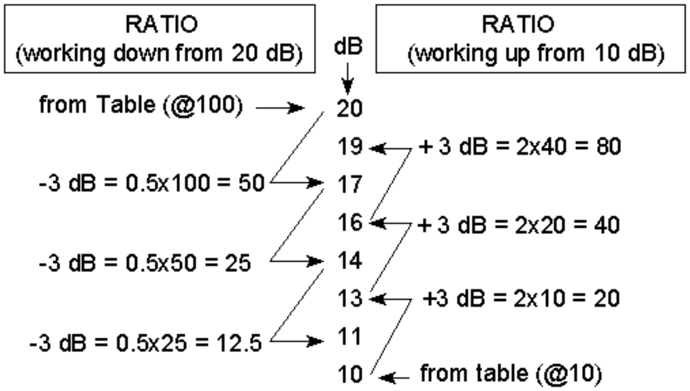

We can get more intermediate dB values by adding or subtracting one to the above, for example, to find the ratio at 12 dB we can:

work up from the bottom; 12 = 1+11 so we have 1.26 (from table) x 12.5 = 15.75
alternately, working down the top 12 = 13-1 so we have 20 x 0.8 (from table) = 16

The resultant numbers are not an exact match (as they should be) because the numbers in the table are rounded off. We can use the same practice to find any ratio at any other given value of dB (or the reverse).

dB AS ABSOLUTE UNITS

Power in absolute units can be expressed by using 1 Watt (or 1 milliwatt) as the reference power in the denominator of the equation for dB. We then call it dBW or dBm. We can then build a table such as the adjoining one.

From the above, any intermediate value can be found using the same dB rules and memorizing several dB values for determining the absolute power, given 48 dBm power output, we determine that 48 dBm = 50 dBm - 2 dB so we take the value at 50 dB which is 100W and divide by the value 1.58 (ratio of 2 dB) to get:

100 watts/1.58 = 63 W or 63,291 mW.

dB AS ABSOLUTE UNITS

dBμW	dBm	POWER	dBW
120	90	1 MW	60
90	60	1 kW	30
80	50	100 W	20
70	40	10 W	10
60	30	1 W (1000 mW)	0
50	20	100 mW	-10
40	10	10 mW	-20
33	3	2 mW	-27
32	2	1.58 mW	-28
31	1	1.26 mw	-29
30	0	1 mW	-30

Because dBW is referenced to one watt, the Log of the power in watts times 10 is dBW. The Logarithm of 10 raised by any exponent is simply that exponent. That is: $\mathrm{Log}(10)^4 = 4$. Therefore, a power that can be expressed as any exponent of 10 can also be expressed in dBW as that exponent times 10. For example, 100 kW can be written 100,000 watts or 10^5 watts. 100 kW is then +50 dBW. Another way to remember this conversion is that dBW is the number of zeros in the power written in watts times 10. If the transmitter power in question is conveniently a multiple of ten (it often is) the conversion to dBW is easy and accurate.

This page intentionally left blank.

DUTY CYCLE

Duty cycle (or duty factor) is a measure of the fraction of the time a radar is transmitting. It is important because it relates to peak and average power in the determination of total energy output. This, in turn, ultimately affects the strength of the reflected signal as well as the required power supply capacity and cooling requirements of the transmitter.

Although there are exceptions, most radio frequency (RF) measurements are either continuous wave (CW) or pulsed RF. CW RF is uninterrupted RF such as from an oscillator. Amplitude modulated (AM), frequency modulated (FM), and phase modulated (PM) RF are considered CW since the RF is continuously present. The power may vary with time due to modulation, but RF is always present. Pulsed RF, on the other hand, is bursts (pulses) of RF with no RF present between bursts. The most general case of pulsed RF consists of pulses of a fixed pulse width (PW) which come at a fixed time interval, or period, (T). For clarity and ease of this discussion, it is assumed that all RF pulses in a pulse train have the same amplitude. Pulses at a fixed interval of time arrive at a rate or frequency referred to as the pulse repetition frequency (PRF) of so many pulse per second. Pulse repetition interval (PRI) and PRF are reciprocals of each other.

$$PRF = 1/T = 1/PRI \qquad \textbf{[1]}$$

Power measurements are classified as either peak pulse power, P_p, or average power, P_{ave}. The actual power in pulsed RF occurs during the pulses, but most power measurement methods measure the heating effects of the RF energy to obtain an average value of the power. It is correct to use either value for reference so long as one or the other is consistently used. Frequently it is necessary to convert from P_p to P_{ave}, or vice versa; therefore the relationship between the two must be understood. Figure 1 shows the comparison between P_p and P_{ave}.

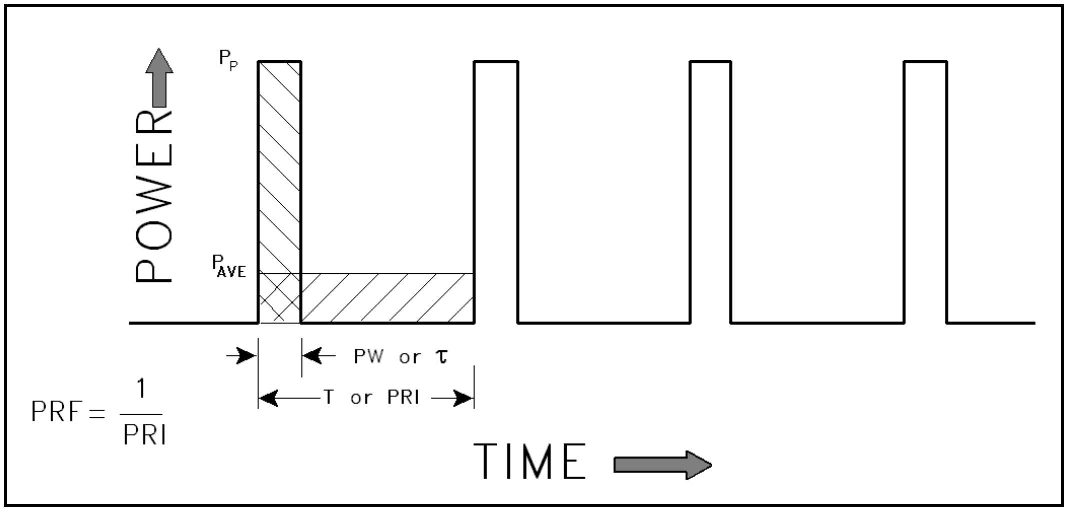

Figure 1. RF Pulse Train.

The average value is defined as that level where the pulse area above the average is equal to area below average between pulses. If the pulses are evened off in such a way as to fill in the area between pulses, the level obtained is the average value, as shown in Figure 1 where the shaded area of the pulse is used to fill in the area between pulses. The area of the pulse is the pulse width multiplied by the peak pulse power. The average area is equal to the average value of power multiplied by the pulse period.

Since the two values are equal:

P_{ave} x T = P_p x PW **[2]**

or

P_{ave}/P_p = PW/T **[3]**

Using [1]

P_{ave}/P_p = PW/T = PW x PRF = PW/PRI = duty cycle **[4]**

(note that the symbol τ represents pulse width (PW) in most reference books)

The ratio of the average power to the peak pulse power is the duty cycle and represents the percentage of time the power is present. In the case of a square wave the duty cycle is 0.5 (50%) since the pulses are present 1/2 the time, the definition of a square wave.

For Figure 1, the pulse width is 1 unit of time and the period is 10 units. In this case the duty cycle is:

PW/T = 1/10 = 0.1 (10%).

A more typical case would be a PRF of 1,000 and a pulse width of 1.0 microseconds. Using [4], the duty cycle is 0.000001 x 1,000 = 0.001. The RF power is present one-thousandth of the time and the average power is 0.001 times the peak power. Conversely, if the power were measured with a power meter which responds to average power, the peak power would be 1,000 time the average reading.

Besides expressing duty cycle as a ratio as obtained in equation [4], it is commonly expressed as either a percentage or in decibels (dB). To express the duty cycle of equation [4] as a percentage, multiply the value obtained by 100 and add the percent symbol. Thus a duty cycle of 0.001 is also 0.1%.

The duty cycle can be expressed logarithmically (dB) so it can be added to or subtracted from power measured in dBm/dBW rather than converting to, and using absolute units.

Duty cycle (dB) = 10 log(duty cycle ratio) **[5]**

For the example of the 0.001 duty cycle, this would be 10 log(0.001) = -30 dB. Thus the average power would be 30 dB less than the peak power. Conversely, the peak power is 30 dB higher than the average power.

For pulse radars operating in the PRF range of 0.25-10 kHz and PD radars operating in the PRF range of 10-500 kHz, typical duty cycles would be:

Pulse	~	0.1 - 3%	=	0.001 - .03	=	-30 to -15 dB
Pulse Doppler	~	5 - 50%	=	0.05 - .5	=	-13 to -3 dB
Continuous Wave	~	100%	=	1	=	0 dB

Intermediate Frequency Bandwidths of typical signals are:

Pulse	1 to 10 MHz
Chirp or Phase coded pulse	0.1 to 10 MHz
CW or PD	0.1 to 5 kHz

PRF is usually subdivided into the following categories: Low 0.25-4 kHz; Medium 8-40 kHz; High 50-300 kHz.

DOPPLER SHIFT

Doppler is the apparent change in wavelength (or frequency) of an electromagnetic or acoustic wave when there is relative movement between the transmitter (or frequency source) and the receiver.

Summary RF Equation for the Two-Way (radar) case

$$f_{Rec} = f_{Xmt} + f_D = f_{Xmt} + \frac{2(V_{Xmtr} + V_{Tgt})\, f_{Xmt}}{c}$$

Summary RF Equation for the One-Way (ESM) case

$$f_{Rec} = f_{Xmt} + f_D = f_{Xmt} + \frac{V_{Xmtr\ or\ Rec}\ \ f_{Xmt}}{c}$$

Rules of Thumb for two-way signal travel
(divide in half for one-way ESM signal measurements)
At 10 GHz, $f_D \cong$
35 Hz per Knot
19 Hz per km/Hr
67 Hz per m/sec
61 Hz per yd/sec
20 Hz per ft/sec

To estimate f_D at other frequencies, multiply these by:

$$\left[\frac{f_{Xmt}(GHz)}{10}\right]$$

The Doppler effect is shown in Figure 1. In everyday life this effect is commonly noticeable when a whistling train or police siren passes you. Audio Doppler is depicted, however Doppler can also affect the frequency of a radar carrier wave, the PRF of a pulse radar signal, or even light waves causing a shift of color to the observer.

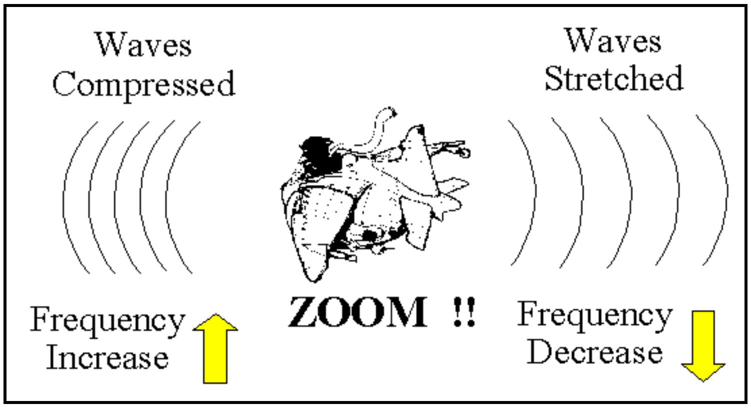

Figure 1. Doppler Frequency Creation From Aircraft Engine Noise.

How do we know the universe is expanding?

Answer: The color of light from distant stars is shifted to red (see Section 7-1: higher λ or lower frequency means Doppler shift is stretched, i.e., expanding).

A memory aid might be that the lights from a car (going away) at night are red (tail lights)!

Doppler frequency shift is directly proportional to velocity and a radar system can therefore be calibrated to measure velocity instead of (or along with) range. This is done by measuring the shift in frequency of a wave caused by an object in motion (Figure 2).

- Transmitter in motion
- Reflector in motion
- Receiver in motion
- All three

For a closing relative velocity:

- Wave is compressed
- Frequency is increased

For an opening relative velocity:

- Wave is stretched
- Frequency is decreased

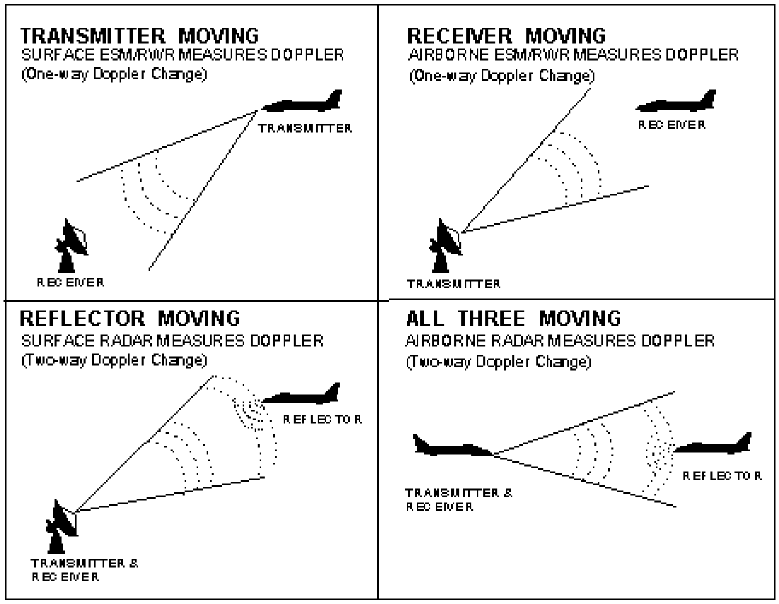
TRANSMITTER MOVING
SURFACE ESM/RWR MEASURES DOPPLER
(One-way Doppler Change)
TRANSMITTER
RECEIVER
RECEIVER MOVING
AIRBORNE ESM/RWR MEASURES DOPPLER
(One-way Doppler Change)
RECEIVER
TRANSMITTER
REFLECTOR MOVING
SURFACE RADAR MEASURES DOPPLER
(Two-way Doppler Change)
REFLECTOR
TRANSMITTER & RECEIVER
ALL THREE MOVING
AIRBORNE RADAR MEASURES DOPPLER
(Two-way Doppler Change)
REFLECTOR
TRANSMITTER & RECEIVER

Figure 2. Methods of Doppler Creation.

To compute Doppler frequency we note that velocity is range rate; V = dr/dt

For the reflector in motion case, you can see the wave compression effect in Figure 3 when the transmitted wave peaks are one wavelength apart. When the first peak reaches the target, they are still one wavelength apart (point a).

When the 2nd peak reaches the target, the target has advanced according to its velocity (vt) (point b), and the first reflected peak has traveled toward the radar by an amount that is less than the original wavelength by the same amount (vt) (point c).

As the 2nd peak is reflected, the wavelength of the reflected wave is 2(vt) less than the original wavelength (point d).

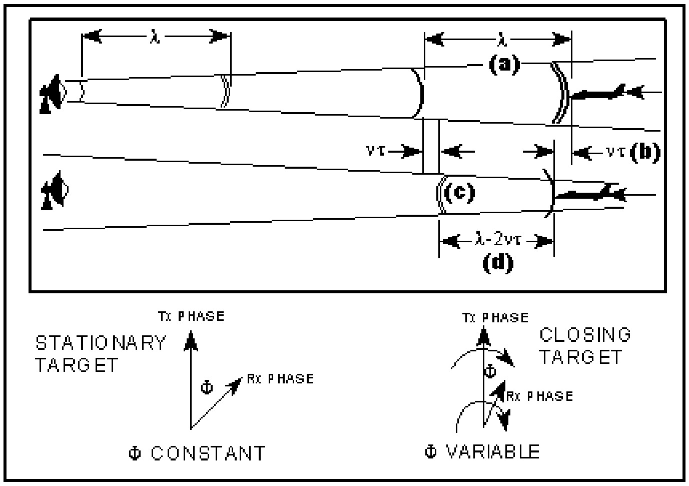
λ
λ
(a)
vτ
vτ
(b)
(c)
λ-2vτ
(d)
Tx PHASE
STATIONARY TARGET
Φ
Rx PHASE
Φ CONSTANT
Tx PHASE
CLOSING TARGET
Φ
Rx PHASE
Φ VARIABLE

Figure 3. Doppler Compression Equivalent to Variable Phase Shift.

The distance the wave travels is twice the target range. The reflected phase lags transmitted phase by 2x the round trip time. For a fixed target the received phase will differ from the transmitted phase by a constant phase shift. For a moving target the received phase will differ by a changing phase shift.

For the closing target shown in Figure 3, the received phase is advancing with respect to the transmitted phase and appears as a higher frequency.

Doppler is dependent upon closing velocity, not actual radar or target velocity as shown in Figure 4.

For the following equations (except radar mapping), we assume the radar and target are moving directly toward one another to simplify calculations (if this is not the case, use the velocity component of one in the direction of the other in the formulas).

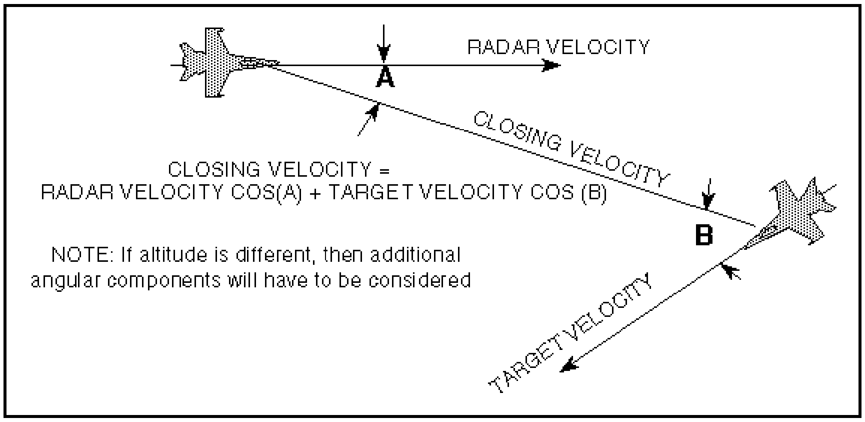
RADAR VELOCITY
A
CLOSING VELOCITY
CLOSING VELOCITY =
RADAR VELOCITY COS(A) + TARGET VELOCITY COS (B)
B
NOTE: If altitude is different, then additional
angular components will have to be considered
TARGET VELOCITY

Figure 4. Doppler Depends Upon Closing Velocity.

For the case of a moving reflector, doppler frequency is proportional to 2x the transmitted frequency:
Higher rf = higher doppler shift
$f_D = (2 \text{ x } V_{Target})(f/c)$

Speed of Light
Conversions
* * *
$c \cong 2.9979 \text{ x } 10^8$ m/sec
$c \cong 5.8275 \text{ x } 10^8$ nm/hr (knots)
$c \cong 9.8357 \text{ x } 10^8$ ft/sec

Likewise, it can be shown that for other cases, the following relationships hold:

For an airplane radar with an airplane target (The "all three moving" case, aircraft radar transmitter, target, and aircraft radar receiver)
$f_D = 2(V_{Radar} + V_{Target})(f/c)$

For the case of a semi-active missile receiving signals (Also "all three moving")
$f_D = (V_{Radar} + 2V_{Target} + V_{Missile})(f/c)$

For the airplane radar with a ground target (radar mapping) or vice versa.
$f_D = 2(V_{Radar} \cos\theta \cos\varphi)(f/c)$, Where θ and φ are the radar scan azimuth and depression angles.

For a ground based radar with airborne target - same as previous using target track crossing angle and ground radar elevation angle.

For the ES/ESM/RWR case where only the target or receiver is moving (One-way Doppler measurements)
$f_D = V_{Receiver\ or\ Target}\,(f/c)$

Note: See Figure 4 if radar and target are not moving directly towards or away from one another.

Figure 5 depicts the results of a plot of the above equation for a moving reflector such as might be measured with a ground radar station illuminating a moving aircraft.

It can be used for the aircraft-to-aircraft case, if the total net closing rate of the two aircraft is used for the speed entry in the figure.

It can also be used for the ES/ESM case (one-way doppler measurements) if the speed of the aircraft is used and the results are divided by two.

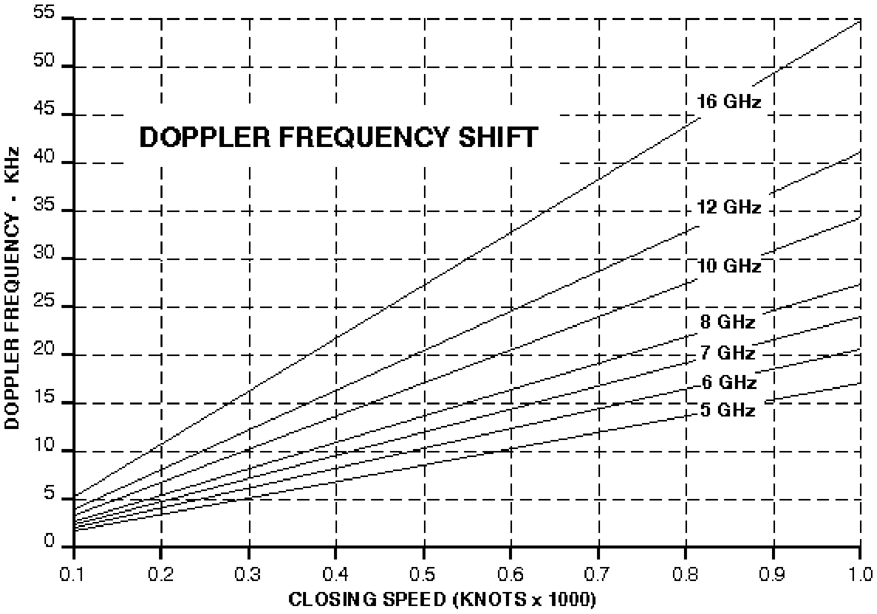

Figure 5. Two-Way Doppler Frequency Shift.

SAMPLE PROBLEMS:

(1) If a ground radar operating at 10 GHz is tracking an airplane flying at a speed of 500 km/hr tangential to it (crossing pattern) at a distance of 10 km, what is the Doppler shift of the returning signal?

Answer: Since the closing velocity is zero, the Doppler is also zero.

(2) If the same aircraft turns directly toward the ground radar, what is the Doppler shift of the returning signal?

Answer: 500 km/hr = 270 kts from Section 2-1. From Figure 4 we see that the Doppler frequency is about 9.2 KHz.

(3) Given that a ground radar operating at 7 GHz is Doppler tracking an aircraft 20 km away (slant range) which is flying directly toward it at an altitude of 20,000 ft and a speed of 800 ft/sec, what amount of VGPO switch would be required of the aircraft jammer to deceive (pull) the radar to a zero Doppler return?

Answer: We use the second equation from the bottom of page 2-6.3 which is essentially the same for this application except a ground radar is tracking an airplane target (vs. an airplane during ground mapping), so for our application we use a positive elevation angle instead of a negative (depression) angle.

$f_D = 2(V_r \text{ Cos } \theta \text{ Cos } \varphi)(f/c)$, where θ is the aircraft track crossing angle and φ is the radar elevation angle.

Since the aircraft is flying directly at the radar, $\theta = 0°$; the aircraft altitude = 20,000 ft = 6,096 meters.

Using the angle equation in Section 2-1, sin $\varphi = x/r$ = altitude / slant range, so:
$\varphi = \sin^{-1}$ (altitude/slant range) = $\sin^{-1}$ (6,096 m / 20,000 m) = 17.7°

$F_D = 2(800 \text{ ft/sec} \cos 0° \cos 17.7°)(7\text{x}10^9 \text{ Hz} / 9.8357 \text{ x } 10^9 \text{ ft/sec}) = \underline{10{,}845 \text{ Hz}}$

This page intentionally left blank.

ELECTRONIC FORMULAS

Ohm's Law Formulas for D-C Circuits. $E = IR = \frac{P}{I} = \sqrt{PR}$ $P = I^2 R = EI = \frac{E^2}{R}$

Ohm's Law Formulas for A-C Circuits and Power Factor.

$$E = IZ = \frac{P}{I \cos\Theta} = \sqrt{\frac{P\,Z}{\cos\Theta}} \qquad P = I^2 Z \cos\Theta = IE \cos\Theta = \frac{E^2 \cos\Theta}{Z}$$

In the above formulas Θ is the angle of lead or lag between current and voltage and cos Θ = P/EI = power factor or *pf*.

$$pf = \frac{\text{Active power (in watts)}}{\text{Apparent power (in volt-amps)}} = PoverE\,I \qquad pf = \frac{R}{Z}$$

Note: Active power is the "resistive" power and equals the equivalent heating effect on water.

Voltage/Current Phase Rule of Thumb Remember "ELI the ICE man"

ELI: Voltage (E) comes before (leads) current (I) in an inductor (L)
ICE: Current (I) comes before (leads) Voltage (E) in a capacitor (C)

Resistors in Series $R_{total} = R_1 + R_2 = R_3 + \ldots$

Two Resistors in Parallel $R_t = \frac{R_1 R_2}{R_1 + R_2}$

Resistors in Parallel, General Formula

$$R_{total} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots}$$

Resonant Frequency Formulas *Where in the second formula f is in kHz and L and C are in microunits.

$$f = \frac{1}{2\pi\sqrt{LC}}, \text{ or } f = \frac{159.2^*}{\sqrt{LC}} \quad L = \frac{1}{4\pi^2 f^2 C}, \text{ or } L = \frac{25{,}330^*}{f^2 C} \quad C = \frac{1}{4\pi^2 f^2 L}, \text{ or } C = \frac{25{,}330^*}{f^2 L}$$

Conductance $G = \frac{1}{R}$ (for D-C circuit) $G = \frac{R}{R^2 + X^2}$ (for A-C circuit)

Reactance Formulas $X_C = \frac{1}{2\pi f\,C}$ $C = \frac{1}{2\pi f\,X_C}$ $X_L = 2\pi fL$ $L = \frac{X_L}{2\pi f}$

Impedance Formulas $Z = \sqrt{R^2 + (X_L - X_C)^2}$ (for series circuit)

$Z = \frac{RX}{\sqrt{R^2 + X^2}}$ (for R and X in parallel)

Q or Figure of Merit $Q = \frac{X_L}{R}$ or $\frac{X_C}{R}$

Frequency Response

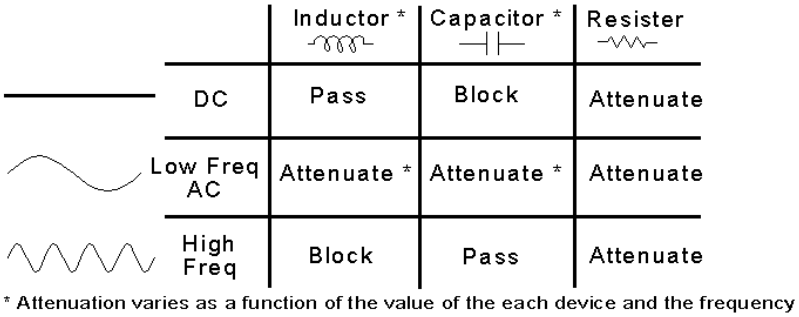

	Inductor *	Capacitor *	Resister
DC	Pass	Block	Attenuate
Low Freq AC	Attenuate *	Attenuate *	Attenuate
High Freq	Block	Pass	Attenuate

* Attenuation varies as a function of the value of the each device and the frequency

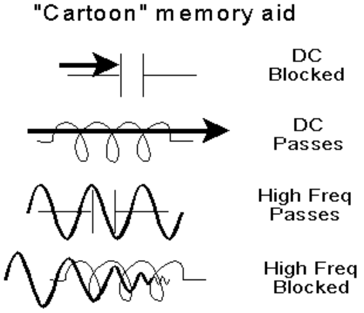

Sinusoidal Voltages and Currents

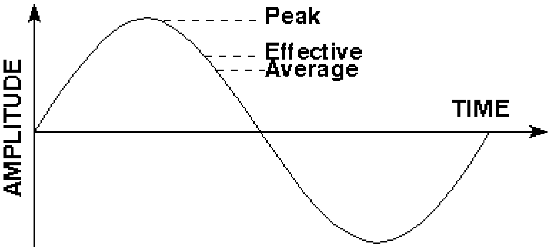

Effective value = 0.707 x peak value
[Also known as Root-Mean Square (RMS) value]

Half Cycle Average value = 0.637 x peak

Peak value = 1.414 x effective value

∴ Effective value = 1.11 x average value

Three-phase AC Configurations

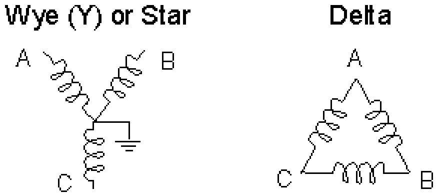

(120° phase difference between each voltage)
If the connection to a three phase AC configuration is miswired, switching any two of the phases will put it back in the proper sequence.
Electric power for ships commonly uses the delta configuration, while commercial electronic and aircraft applications commonly use the wye configuration.

Color Code for House Wiring:	**PURPOSE:**	**Color Code for Chassis Wiring:**
Black or red	HOT	Red
White	NEUTRAL (Return)	White
Green or bare	GROUND	Black

Color Code for Resistors: First and second band:
(and third band # of zeros if not gold/silver)

				Third band Multiplier		**Fourth band Tolerance**	
0	Black	5	Green	.1	Gold	5%	Gold
1	Brown	6	Blue	.01	Silver	10%	Silver
2	Red	7	Violet			20%	No color
3	Orange	8	Gray				
4	Yellow	9	White				

The third color band indicates number of zeros to be added after figures given by first two color bands. But if third color band is gold, multiply by 0.1 and if silver multiply by 0.01. Do not confuse with fourth color-band that indicates tolerance. Thus, a resistor marked blue-red-gold-gold has a resistance of 6.2 ohms and a 5% tolerance.

MISSILE AND ELECTRONIC EQUIPMENT DESIGNATIONS

Missiles are designated with three letters from the columns below plus a number (i.e., AIM-7M) Suffixes (M in this case) indicate a modification.

First Letter Launch Environment	**Second Letter** Mission Symbols	**Third Letter** Vehicle Type
A Air	D Decoy	M Guided Missile
B Multiple	E Special electronic	N Probe (non-orbital instruments)
C Coffin	G Surface attack	R Rocket (without installed or remote control guidance)
H Silo stored	I Intercept, aerial	
L Silo launched	Q Drone	
M Mobile	T Training	
P Soft Pad	U Underwater attack	
R Ship	W Weather	
U Underwater		

U.S. military electronic equipment is assigned an identifying alphanumeric designation that is used to uniquely identify it. This system is commonly called the “AN” designation system, although its formal name is the Joint Electronics Type Designation System (JETDS). The letters AN preceding the equipment indicators formerly meant “Army/Navy,” but now are a letter set that can only be used to indicate formally designated DOD equipment. The first three letters following the “AN/” indicate Platform Installation, Equipment Type, and Equipment Function, respectively. The appropriate meaning is selected from the lists below. The letters following the AN designation numbers provide added information about equipment. Suffixes (A, B, C, etc.) indicate a modification. The letter (V) indicates that variable configurations are available. The letter (X) indicates a development status. A parenthesis ( ) without a number within it indicates a generic system that has not yet received a formal designation, e.g., AN/ALQ( ). Quite often the () is pronounced “bow legs” since they look like the shape of cowboy legs.

First Letter Platform Installation	**Second Letter** Equipment Type	**Third Letter** Function or Purpose
A Piloted aircraft	A Invisible light, heat radiation	B Bombing
B Underwater mobile, submarine	C Carrier	C Communications
D Pilotless carrier	D Radiac	D Direction finder, reconnaissance and/or surveillance
F Fixed ground	F Photographic	E Ejection and/or release
G General ground use	G Telegraph or teletype	G Fire control or searchlight directing
K Amphibious	I Interphone and public address	H Recording and/or reproducing
M Mobile (ground)	J Electromechanical or inertial wire covered	K Computing
P Portable	K Telemetering	M Maintenance and/or test assemblies
S Water	L Countermeasures	N Navigation aids
T Ground, transportable	M Meteorological	Q Special or combination of purposes
U General utility	N Sound in air	R Receiving, passive detecting
V Vehicular (ground)	P Radar	S Detecting and/or range and bearing, search
W Water surface and underwater combination	Q Sonar and underwater sound	T Transmitting
Z Piloted-pilotless airborne vehicle combination	R Radio	W Automatic flight or remote control
	S Special or combinations of types	X Identification and recognition
	T Telephone (wire)	Y Surveillance and control
	V Visual and visible light	
	W Armament	
	X Facsimile or television	
	Y Data Processing	

This page intentionally left blank.

RADAR HORIZON / LINE OF SIGHT

There are limits to the reach of radar signals. At the frequencies normally used for radar, radio waves usually travel in a straight line. The waves may be obstructed by weather or shadowing, and interference may come from other aircraft or from reflections from ground objects (Figure 1).

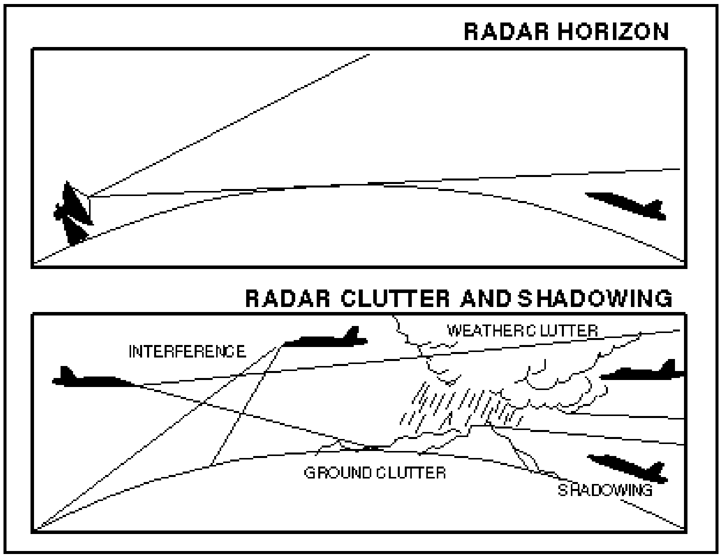

Figure 1. Radar Horizon and Shadowing.

As also shown in Figure 1, an aircraft may not be detected because it is below the radar line which is tangent to the earth's surface.

Some rules of thumb are:

Range (to horizon):

$$R_{NM} = 1.23\sqrt{h_{radar}} \quad with\ h\ in\ ft$$

Range (beyond horizon / over earth curvature):

$$R_{NM} = 1.23\left(\sqrt{h_{radar}} + \sqrt{h_{target}}\right) \quad with\ h\ in\ ft$$

In obtaining the radar horizon equations, it is common practice to assume a value for the Earth's radius that is 4/3 times the actual radius. This is done to account for the effect of the atmosphere on radar propagation. For a true line of sight, such as used for optical search and rescue, the constant in the equations changes from 1.23 to 1.06.

A nomograph for determining maximum target range is depicted in Figure 2. Although an aircraft is shown to the left, it could just as well be a ship, with radars on a mast of height "h." Any target of height (or altitude) "H" is depicted on the right side.

See also Section 5-1 on ducting and refraction, which may increase range beyond these distances.

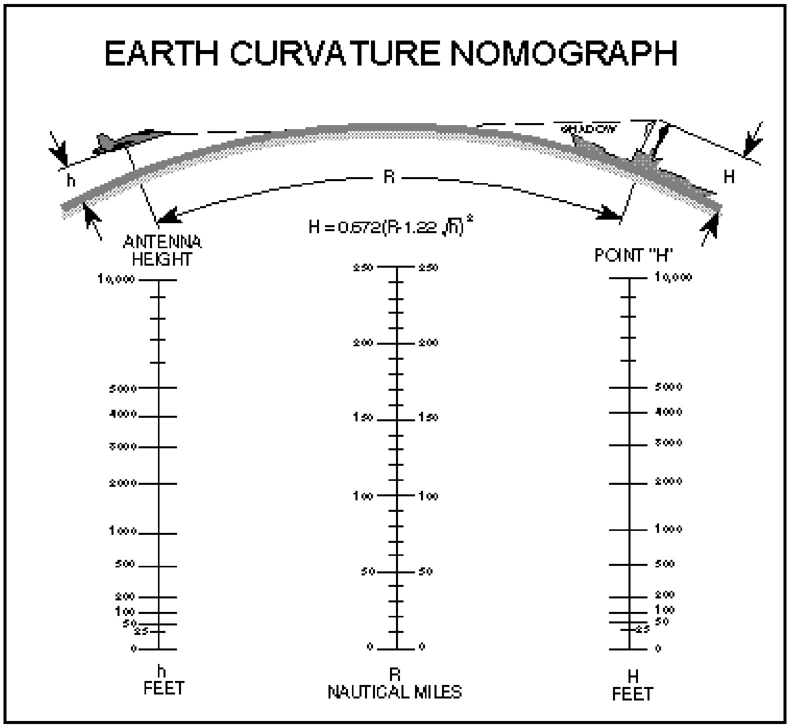

Figure 2. Earth Curvature Nomograph.

This data was expanded in Figure 3 to consider the maximum range one aircraft can detect another aircraft using:

$$R_{NM} = 1.23\left(\sqrt{h_{radar}} + \sqrt{h_{target}}\right)$$

(with h in feet)

It can be used for surface targets if $H_{target} = 0$. It should be noted that most aircraft radars are limited in power output, and would not detect small or surface objects at the listed ranges.

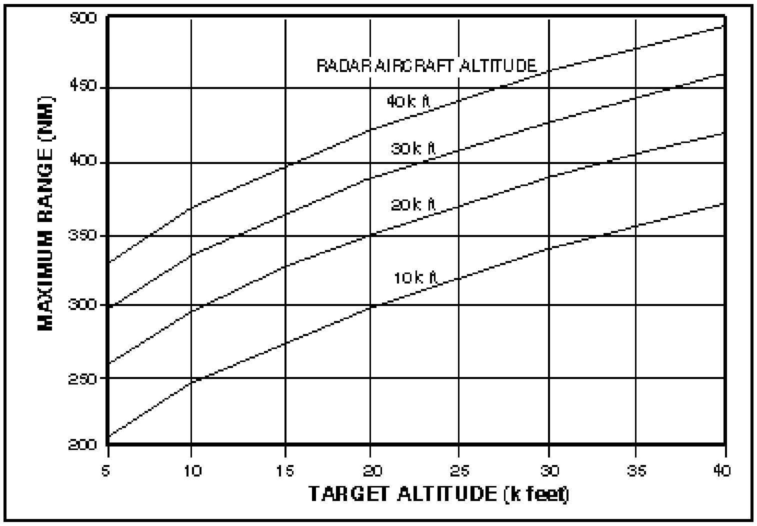

Figure 3. Aircraft Radar vs. Aircraft Target Maximum Range.

Figure 4 depicts the maximum range that a ship height antenna can detect a zero height object (i.e., rowboat).

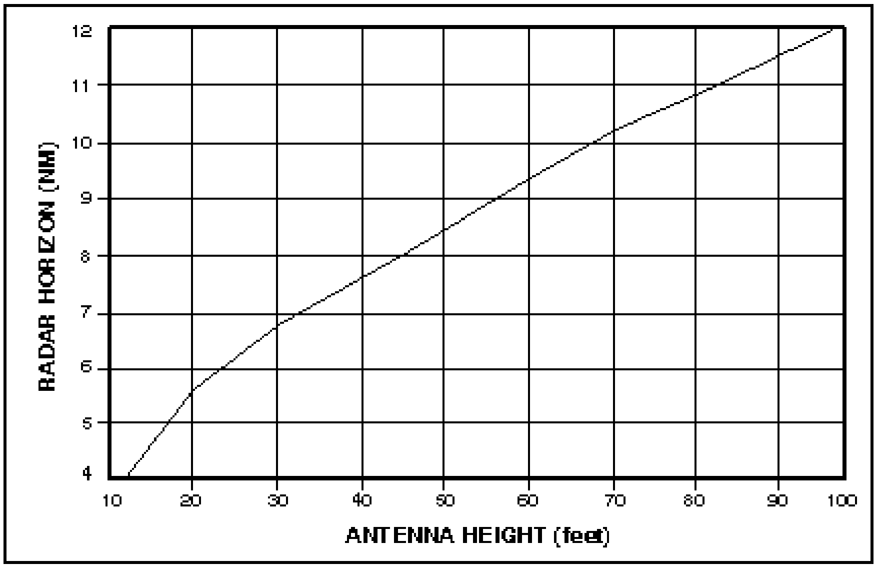

Figure 4. Ships Radar Horizon With Target on the Surface.

In this case "H" = 0, and the general equation becomes: $R_{\max}(NM) = 1.23\sqrt{h_r}$

Where h_r is the height of the radar in feet.

Figure 5 depicts the same for aircraft radars. It should be noted that most aircraft radars are limited in power output, and would not detect small or surface objects at the listed ranges.

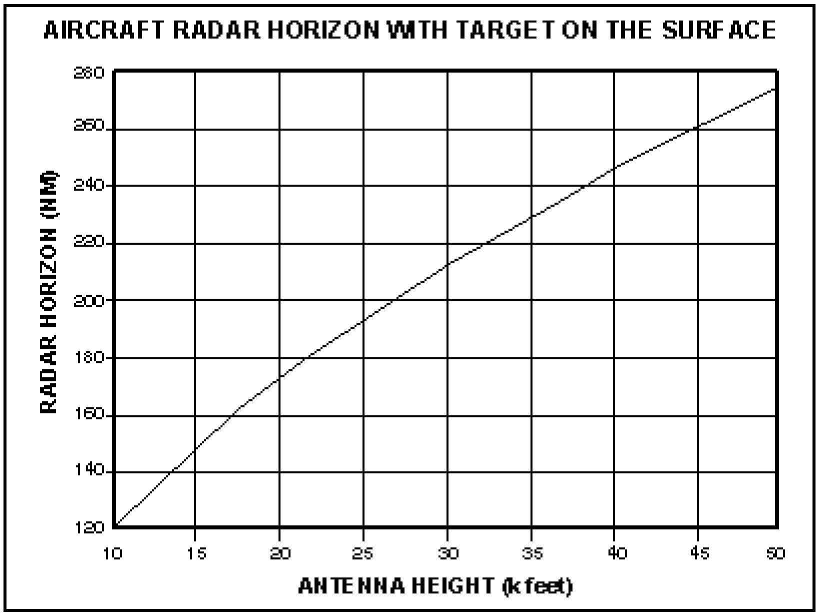

Figure 5. Aircraft Radar Horizon With Target on the Surface.

Other general rules of thumb for surface "targets/radars" are as follows:

For Visual SAR:

$$R_{Visual}(NM) = 1.05\sqrt{Acft\ \ Alt\ \ in\ \ ft}$$

For ESM:

$$R_{ESM}(NM) = 1.5\sqrt{Acft\ \ Alt\ \ in\ \ ft}$$

This page intentionally left blank.

PROPAGATION TIME / RESOLUTION

1. ROUND TRIP RANGE: $R = \frac{c\,t}{2}$ with t = time to reach target

Rules of Thumb

In one µsec round trip time, a wave travels to and from an object at a distance of:	The time it takes to travel to and from an object at a distance of:
≅ 150 m	1 m ≅ 0.0067 µsec
≅ 164 yd	1 yd ≅0.006 µsec
≅ 500 ft	1 ft ≅ 0.002 µsec
≅ 0.08 NM	1 NM ≅ 12.35 µsec
≅ 0.15 km	1 km ≅ 6.7 µsec

2. ONE WAY RANGE: ***R = ct*** with t = time to reach target

Time	Distance Traveled
1 milli sec (ms)	165 NM
1 micro sec (µs)	1000 ft
1 nano sec (ns)	1 ft

Distance	Time it Takes
1 NM	6.18 µsec
1 km	3.3 µsec
1 ft	1 nsec

3. UNAMBIGUOUS RANGE

(DISTANCE BETWEEN PULSES): $R = \frac{c \bullet PRI}{2}$

Normally a radar measures "distance" to the target by measuring time from the last transmitted pulse. If the inter-pulse period (T) is long enough that isn't a problem as shown in "A" to the right. When the period is shortened, the time to the last previous pulse is shorter than the actual time it took, giving a false (ambiguous) shorter range (figure "B").

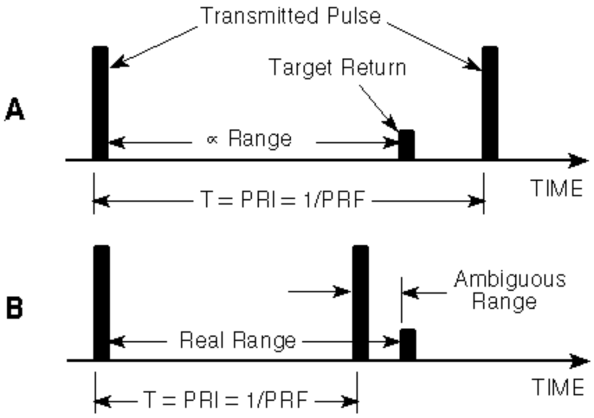

Rules of Thumb

RNM ≅ 81Pms

RKm ≅ 150Pms

Where Pms is PRI in milliseconds

4. RANGE RESOLUTION

Rules of Thumb

500 ft per microsecond of pulse width

500 MHz IF bandwidth provides 1 ft of resolution.

5. BEST CASE PERFORMANCE:

The atmosphere limits the accuracy to 0.1 ft

The natural limit for resolution is one RF cycle.

This page intentionally left blank.

MODULATION

Modulation is the process whereby some characteristic of one wave is varied in accordance with some characteristic of another wave. The basic types of modulation are angular modulation (including the special cases of phase and frequency modulation) and amplitude modulation. In missile radars, it is common practice to amplitude modulate the transmitted RF carrier wave of tracking and guidance transmitters by using a pulsed wave for modulating, and to frequency modulate the transmitted RF carrier wave of illuminator transmitters by using a sine wave.

Frequency Modulation (FM) - As shown in Figure 1, an unmodulated RF signal in the time domain has only a single spectral line at the carrier frequency (f_c) in the frequency domain. If the signal is frequency modulated, as shown in Figure 2 (simplified using only two changes), the spectral line will correspondingly shift in the frequency domain. The bandwidth can be approximated using Carson's rule: $BW = 2(\Delta f + fm)$, where Δf is the peak deviation of the instantaneous frequency from the carrier and fm is the highest frequency present in the modulating signal. There are usually many more "spikes" in the frequency domain than depicted. The number of spikes and shape of the frequency domain envelope (amplitude) are based on the modulation index β. The modulation index is related to the same two factors used in Carson's rule. A high Δf means a higher modulation index with many more "spikes" spread across a wider bandwidth.

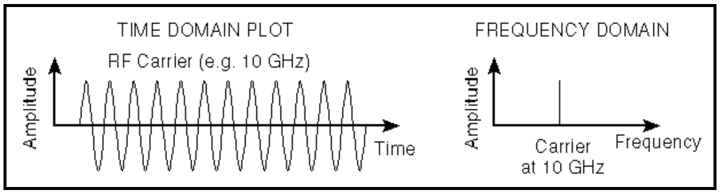
TIME DOMAIN PLOT
RF Carrier (e.g. 10 GHz)
Amplitude
Time
FREQUENCY DOMAIN
Amplitude
Carrier at 10 GHz
Frequency

Figure 1. Unmodulated RF Signal.

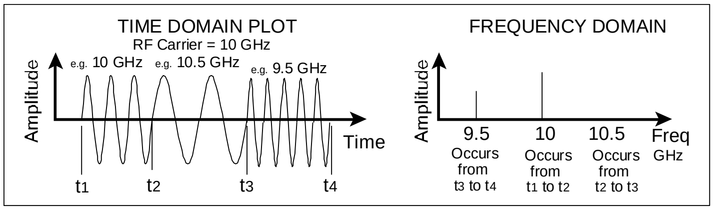
TIME DOMAIN PLOT
RF Carrier = 10 GHz
e.g. 10 GHz
e.g. 10.5 GHz
e.g. 9.5 GHz
Amplitude
Time
t1
t2
t3
t4
FREQUENCY DOMAIN
Amplitude
9.5
Occurs from t3 to t4
10
Occurs from t1 to t2
10.5
Occurs from t2 to t3
Freq GHz

Figure 2. RF Signal With Frequency Modulation.

Amplitude Modulation (AM) - If the signal in Figure 1 is amplitude modulated by a sinewave as shown in Figure 3, sidebands are produced in the frequency domain at $F_c \pm F_{AM}$. AM other than by a pure sine wave will cause additional sidebands normally at $F_c \pm nF_{AM}$, where n equals 1, 2, 3, 4, etc.

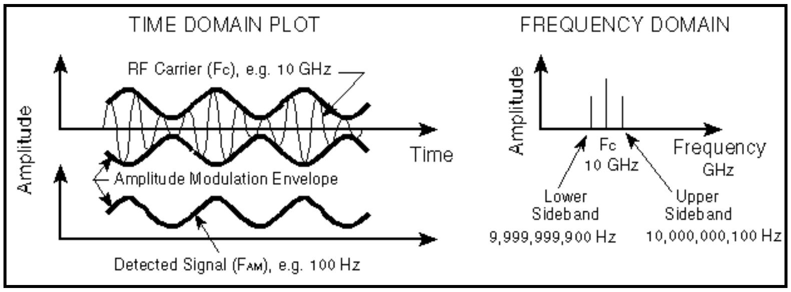
TIME DOMAIN PLOT
RF Carrier (Fc), e.g. 10 GHz
Amplitude
Time
Amplitude Modulation Envelope
Detected Signal (FAM), e.g. 100 Hz
FREQUENCY DOMAIN
Amplitude
Fc
10 GHz
Frequency GHz
Lower Sideband
9,999,999,900 Hz
Upper Sideband
10,000,000,100 Hz

Figure 3. Sinewave Modulated RF Signal.

Pulse modulation is a special case of AM wherein the carrier frequency is gated at a pulsed rate. When the reciprocal of the duty cycle of the AM is a whole number, harmonics corresponding to multiples of that whole number will be missing, e.g., in a 33.33% duty cycle, AM wave will miss the 3rd, 6th, 9th, etc. harmonics, while a square wave or

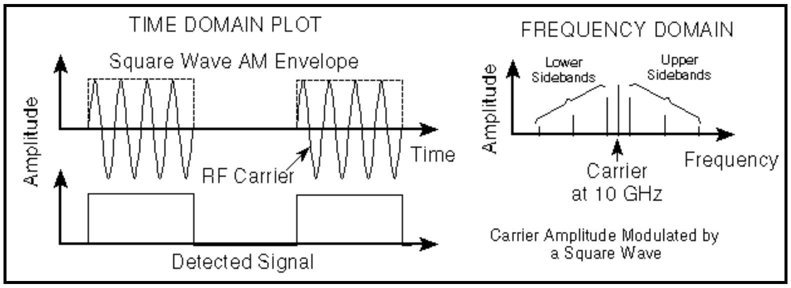
TIME DOMAIN PLOT
Square Wave AM Envelope
Amplitude
Time
RF Carrier
Detected Signal
FREQUENCY DOMAIN
Lower Sidebands
Upper Sidebands
Amplitude
Carrier at 10 GHz
Frequency
Carrier Amplitude Modulated by a Square Wave

Figure 4. Square Wave Modulated RF Signal (50% Duty Cycle).

50% duty cycle triangular wave will miss the 2nd, 4th, 6th, etc. harmonic, as shown in Figure 4. It has sidebands in the frequency domain at $F_c \pm nF_{AM}$, where n = 1, 3, 5, etc. The amplitude of the power level follows a sine x / x type distribution.

Figure 5 shows the pulse width (PW) in the time domain which defines the lobe width in the frequency domain (Figure 6). The width of the main lobe is 2/PW, whereas the width of a side lobe is 1/PW. Figure 5 also shows the pulse repetition interval (PRI) or its reciprocal, pulse repetition frequency (PRF), in the time domain. In the frequency domain, the spectral lines inside the lobes are separated by the PRF or 1/PRI, as shown in Figures 7 and 8. Note that Figures 7 and 8 show actual magnitude of the side lobes, whereas in Figure 4 and 6, the absolute value is shown.

The magnitude of each spectral component for a rectangular pulse can be determined from the following formula:

$$a_n = 2A\frac{\tau}{T}\,\frac{\sin(n\pi\tau/T)}{n\pi\tau/T} \qquad where: \begin{array}{c}\tau = pulse\ width\ (PW)\\ T = period\ (PRI)\end{array} \quad and\ \ A = Amplitude\ of\ rectangular\ pulse \qquad [1]$$

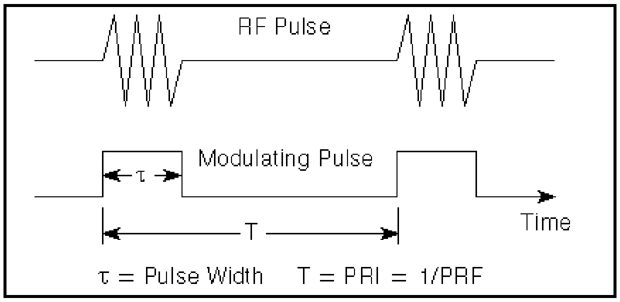

Figure 5. Pulse Width and PRI/PRF Waveforms.

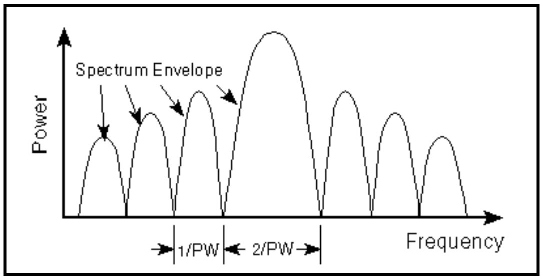

Figure 6. Sidelobes Generated by Pulse Modulation (Absolute Value).

Figure 7 shows the spectral lines for a square wave 50% duty cycle), while Figure 8 shows the spectral lines for a 33.33% duty cycle rectangular wave signal.

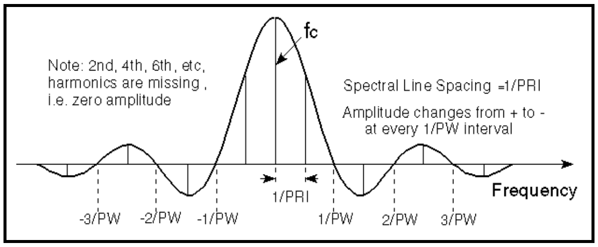

Figure 7. Spectral Times for a Square Wave Modulated Signal.

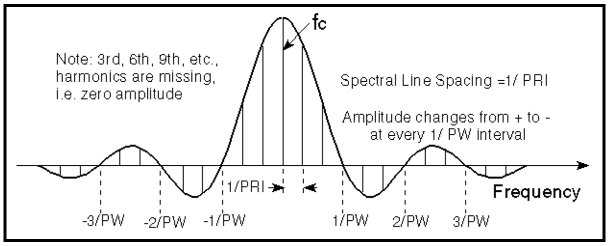

Figure 8. Spectral Lines for a 33.3% Duty Cycle.

Figure 9 shows that for square wave AM, a significant portion of the component modulation is contained in the first few harmonics which comprise the wave. There are twice as many sidebands or spectral lines as there are harmonics (one on the plus and one on the minus side of the carrier). Each sideband represents a sine wave at a frequency equal to the difference between the spectral line and f_c.

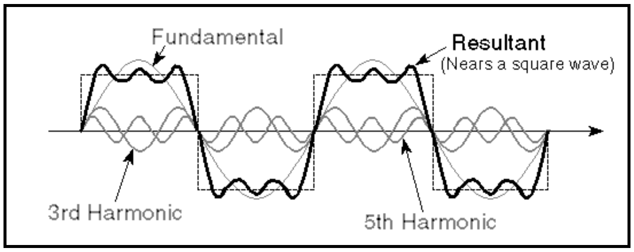

Figure 9. Square Wave Consisting of Sinewave Harmonics.

A figure similar to Figure 9 can be created for any rectangular wave. The relative amplitude of the time domain sine wave components are computed using equation [1]. Each is constructed such that at the midpoint of the pulse the sine wave passes through a maximum (or minimum if the coefficient is negative) at the same time. It should be noted that the "first" harmonic created using this formula is NOT the carrier frequency, f_c, of the modulated signal, but at $F_c \pm F_{AM}$.

While equation [1] is for rectangular waves only, similar equations can be constructed using Fourier coefficients for other waveforms, such as triangular, sawtooth, half sine, trapezoidal, and other repetitive geometric shapes.

PRI Effects - If the PW remains constant but PRI increases, the number of sidelobes remains the same, but the number of spectral lines gets denser (move closer together) and vice versa (compare Figures 7 and 8). The spacing between the spectral lines remains constant with constant PRI.

Pulse Width (PW) Effects - If the PRI remains constant, but the PW increases, then the lobe width decreases and vice versa. If the PW approaches PRI, the spectrum will approach "one lobe," i.e., a single spectral line. The spacing of the lobes remains constant with constant PW.

RF Measurements - If the receiver bandwidth is smaller than the PRF, the receiver will respond to one spectral line at a time. If the receiver bandwidth is wider than the PRF but narrower than the reciprocal of the PW, the receiver will respond to one spectral envelope at a time.

Jet Engine Modulation (JEM)

Section 2-6 addresses the Doppler shift in a transmitted radar signal caused by a moving target. The amount of Doppler shift is a function of radar carrier frequency and the speed of the radar and target. Moving or rotating surfaces on the target will have the same Doppler shift as the target, but will also impose AM on the Doppler shifted return (see Figure 10). Reflections off rotating jet engine compressor blades, aircraft propellers, ram air turbine (RAT) propellers used to power aircraft pods, helicopter rotor blades, and protruding surfaces of automobile hubcaps will all provide a chopped reflection of the impinging signal. The reflections are characterized by both positive and negative Doppler sidebands corresponding to the blades moving toward and away from the radar respectively.

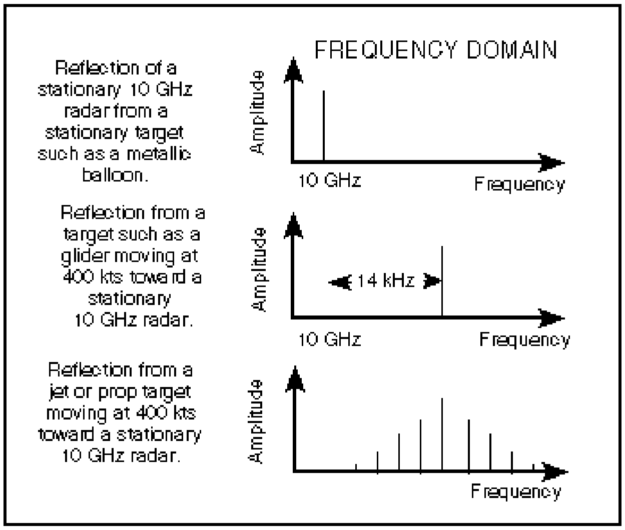
FREQUENCY DOMAIN
Reflection of a stationary 10 GHz radar from a stationary target such as a metallic balloon.
Amplitude
10 GHz
Frequency
Reflection from a target such as a glider moving at 400 kts toward a stationary 10 GHz radar.
Amplitude
14 kHz
10 GHz
Frequency
Reflection from a jet or prop target moving at 400 kts toward a stationary 10 GHz radar.
Amplitude
Frequency

Figure 10. Doppler Return and JEM.

Therefore, forward/aft JEM does not vary with radar carrier frequency, but the harmonics contained in the sidebands are a function of the PRF of the blade chopping action and its amplitude is target aspect dependent, i.e., blade angle, intake/exhaust internal reflection, and jet engine cowling all effect lateral return from the side. If the aspect angle is too far from head-on or tail-on and the engine cowling provides shielding for the jet engine, there may not be any JEM to detect. On the other hand, JEM increases when you are orthogonal (at a right angle) to the axis of blade rotation. Consequently for a fully exposed blade as in a propeller driven aircraft or helicopter, JEM increases with angle off the boresight axis of the prop/rotor.

This page intentionally left blank.

TRANSFORMS / WAVELETS

Transform Analysis

Signal processing using a transform analysis for calculations is a technique used to simplify or accelerate problem solution. For example, instead of dividing two large numbers, we might convert them to logarithms, subtract them, then look-up the anti-log to obtain the result. While this may seem a three-step process as opposed to a one-step division, consider that long-hand division of a four digit number by a three digit number, carried out to four places requires three divisions, 3-4 multiplication's, and three subtractions. Computers process additions or subtractions much faster than multiplications or divisions, so transforms are sought which provide the desired signal processing using these steps.

Fourier Transform

Other types of transforms include the Fourier transform, which is used to decompose or separate a waveform into a sum of sinusoids of different frequencies. It transforms our view of a signal from time based to frequency based. Figure 1 depicts how a square wave is formed by summing certain particular sine waves. The waveform must be continuous, periodic, and almost everywhere differentiable. The Fourier transform of a sequence of rectangular pulses is a series of sinusoids. The envelope of the amplitude of the coefficients of this series is a waveform with a Sin X/X shape. For the special case of a single pulse, the Fourier series has an infinite series of sinusoids that are present for the duration of the pulse.

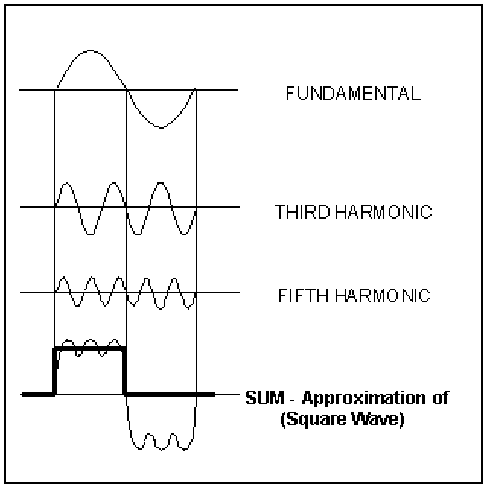

Figure 1. Harmonics.

Digital Sampling of Waveforms

To process a signal digitally, we need to sample the signal frequently enough to create a complete "picture" of the signal. The discrete Fourier transform (DFT) may be used in this regard. Samples are taken at uniform time intervals as shown in Figure 2 and processed.

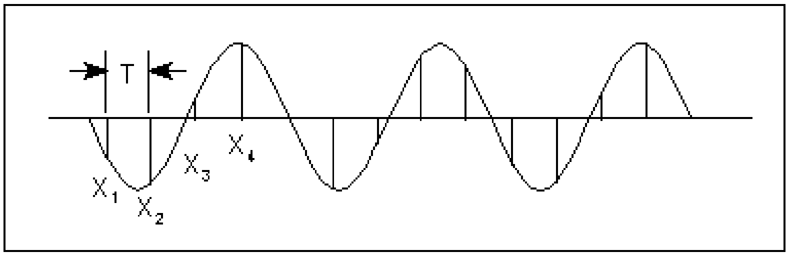

Figure 2. Waveform Sampling.

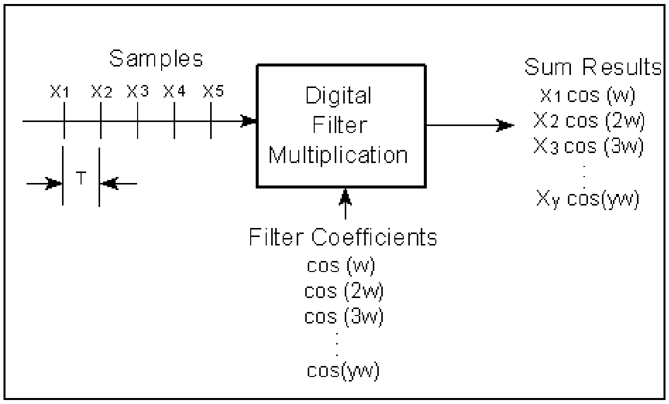

Figure 3. Digital Filtering.

If the digital information is multiplied by the Fourier coefficients, a digital filter is created as shown Figure 3. If the sum of the resultant components is zero, the filter has ignored (notched out) that frequency sample. If the sum is a relatively large number, the filter has passed the signal. With the single sinusoid shown, there should be only one resultant. (Note that being "zero" and relatively large may just mean below or above the filter's cutoff threshold.)

Figure 4 depicts the process pictorially: The vectors in the figure just happen to be pointing in a cardinal direction because the strobe frequencies are all multiples of the vector (phasor) rotation rate, but that is not normally the case. Usually the vectors will point in a number of different directions, with a resultant in some direction other than straight up.

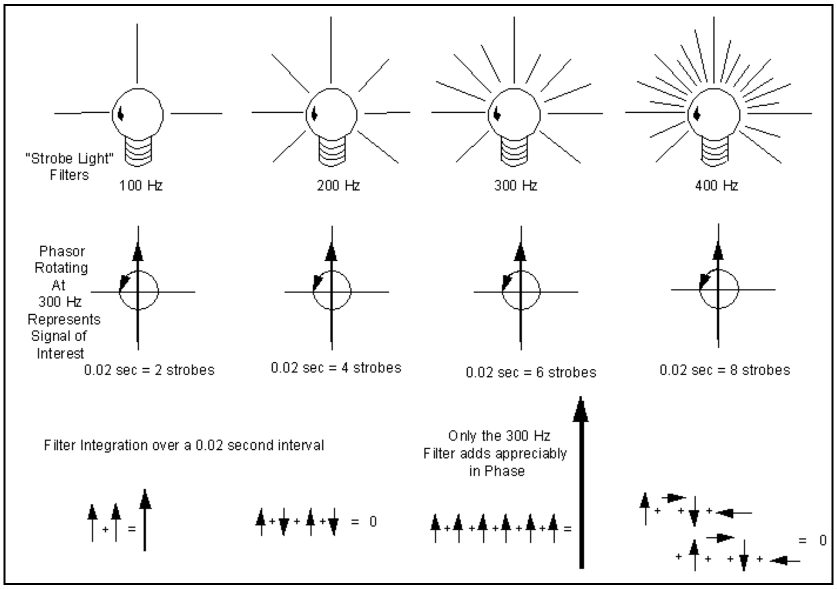
"Strobe Light"
Filters
100 Hz
200 Hz
300 Hz
400 Hz
Phasor
Rotating
At
300 Hz
Represents
Signal of
Interest
0.02 sec = 2 strobes
0.02 sec = 4 strobes
0.02 sec = 6 strobes
0.02 sec = 8 strobes
Filter Integration over a 0.02 second interval
Only the 300 Hz
Filter adds appreciably
in Phase
= 0
= 0

Figure 4. Phasor Representation.

In addition, sampling normally has to taken at or above twice the rate of interest (also known as the Nyquist rate), otherwise ambiguous results may be obtained. Figure 4 is under-sampled (for clarity) and consequently does not depict typical filtering.

Fast Fourier Transforms

One problem with this type of processing is the large number of additions, subtractions, and multiplications which are required to reconstruct the output waveform. The Fast Fourier transform (FFT) was developed to reduce this problem. It recognizes that because the filter coefficients are sine and cosine waves, they are symmetrical about 90, 180, 270, and 360 degrees. They also have a number of coefficients equal either to one or zero, and duplicate coefficients from filter to filter in a multibank arrangement. By waiting for all of the inputs for the bank to be received, adding together those inputs for which coefficients are the same before performing multiplications, and separately summing those combinations of inputs and products which are common to more than one filter, the required amount of computing may be cut drastically.

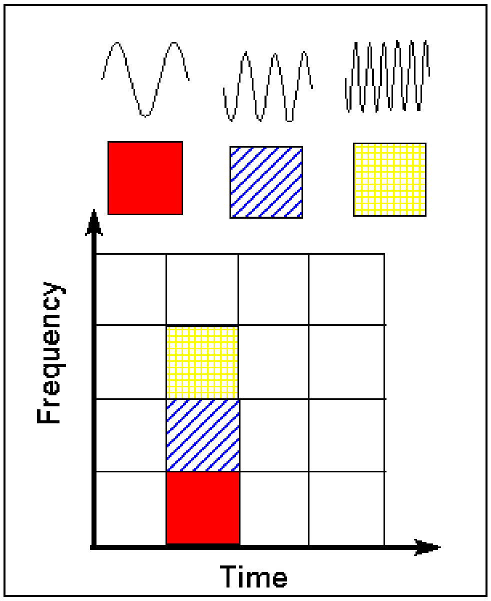
Frequency
Time

Figure 5. Windowed Fourier Transform.

- The number of computations for a DFT is on the order of N squared.

- The number of computations for a FFT when N is a power of two is on the order of N $\log_2$ N.

For example, in an eight filter bank, a DFT would require 512 computations, while an FFT would only require 56, significantly speeding up processing time.

Windowed Fourier Transform

The Fourier transform is continuous, so a windowed Fourier transform (WFT) is used to analyze non-periodic signals as shown in Figure 5. With the WFT, the signal is divided into sections (one such section is shown in Figure 5) and each section is analyzed for frequency content. If the signal has sharp transitions, the input data is windowed so that the sections converge to zero at the endpoints. Because a single window is used for all frequencies in the WFT, the resolution of the analysis is the same (equally spaced) at all locations in the time-frequency domain.

The FFT works well for signals with smooth or uniform frequencies, but it has been found that other transforms work better with signals having pulse type characteristics, time-varying (non-stationary) frequencies, or odd shapes.

The FFT also does not distinguish sequence or timing information. For example, if a signal has two frequencies (a high followed by a low or vice versa), the Fourier transform only reveals the frequencies and relative amplitude, not the order in which they occurred. So Fourier analysis works well with stationary, continuous, periodic, differentiable signals, but other methods are needed to deal with non-periodic or non-stationary signals.

Wavelet Transform

The Wavelet transform has been evolving for some time. Mathematicians theorized its use in the early 1900's. While the Fourier transform deals with transforming the time domain components to frequency domain and frequency analysis, the wavelet transform deals with scale analysis, that is, by creating mathematical structures that provide varying time/frequency/amplitude slices for analysis. This transform is a portion (one or a few cycles) of a complete waveform, hence the term wavelet.

The wavelet transform has the ability to identify frequency (or scale) components, simultaneously with their location(s) in time. Additionally, computations are directly proportional to the length of the input signal. They require only N multiplications (times a small constant) to convert the waveform. For the previous eight filter bank example, this would be about twenty calculations, vice 56 for the FFT.

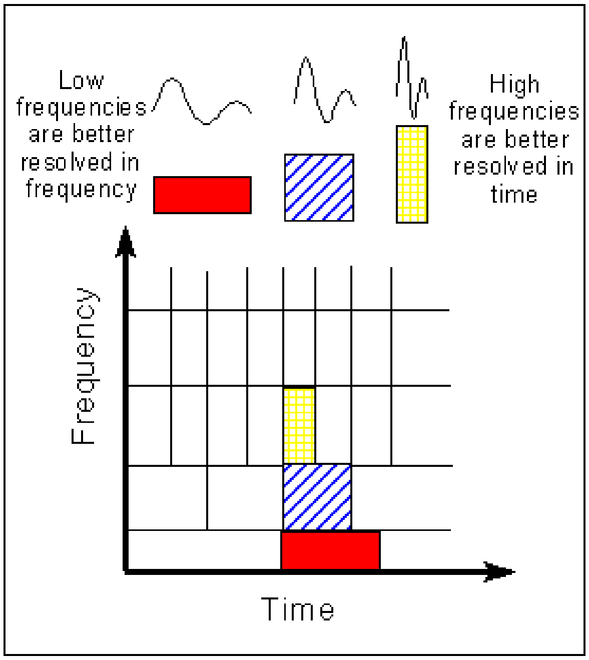
Low frequencies are better resolved in frequency
High frequencies are better resolved in time
Frequency
Time

Figure 6. Wavelet Transform.

In wavelet analysis, the scale that one uses in looking at data plays a special role. Wavelet algorithms process data at different scales or resolutions. If we look at a signal with a large "window," we would notice gross features. Similarly, if we look at a signal with a small "window," we would notice small discontinuities as shown in Figure 6. The result in wavelet analysis is to "see the forest and the trees." A way to achieve this is to have short high-frequency fine scale functions and long low-frequency ones. This approach is known as multi-resolution analysis.

For many decades, scientists have wanted more appropriate functions than the sines and cosines (base functions) which comprise Fourier analysis, to approximate choppy signals. (Although Walsh transforms work if the waveform is periodic and stationary). By their definition, sine and cosine functions are non-local (and stretch out to infinity), and therefore do a very poor job in approximating sharp spikes. But with wavelet analysis, we can use approximating functions that are contained neatly in finite (time/frequency) domains. Wavelets are well-suited for approximating data with sharp discontinuities.

The wavelet analysis procedure is to adopt a wavelet prototype function, called an "analyzing wavelet" or "mother wavelet." Temporal analysis is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the prototype wavelet. Because the original signal or function can be represented in terms of a wavelet expansion (using coefficients in a linear combination of the wavelet functions), data operations can be performed using just the corresponding wavelet coefficients as shown in Figure 7.

If one further chooses the best wavelets adapted to the data, or truncates the coefficients below some given threshold, the data is sparsely represented. This "sparse coding" makes wavelets an excellent tool in the field of data compression. For instance, the FBI uses wavelet coding to store fingerprints. Hence, the concept of wavelets is to look at a signal at various scales and analyze it with various resolutions.

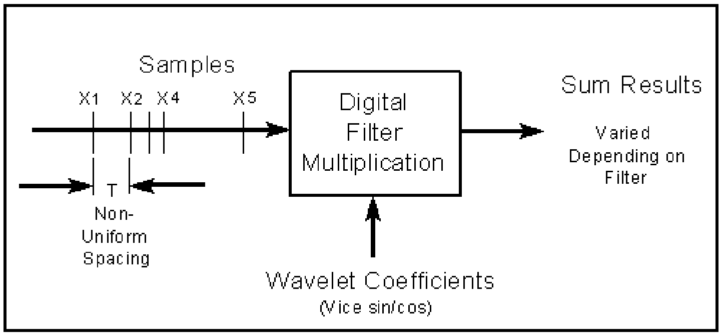
Samples
X1
X2
X4
X5
T
Non-Uniform Spacing
Digital Filter Multiplication
Sum Results
Varied Depending on Filter
Wavelet Coefficients
(Vice sin/cos)

Figure 7. Wavelet Filtering.

Analyzing Wavelet Functions

Fourier transforms deal with just two basis functions (sine and cosine), while there are an infinite number of wavelet basis functions. The freedom of the analyzing wavelet is a major difference between the two types of analyses and is important in determining the results of the analysis. The "wrong" wavelet may be no better (or even far worse than) than the Fourier analysis. A successful application presupposes some expertise on the part of the user. Some prior knowledge about the signal must generally be known in to select the most suitable distribution and adapt the parameters to the signal. Some

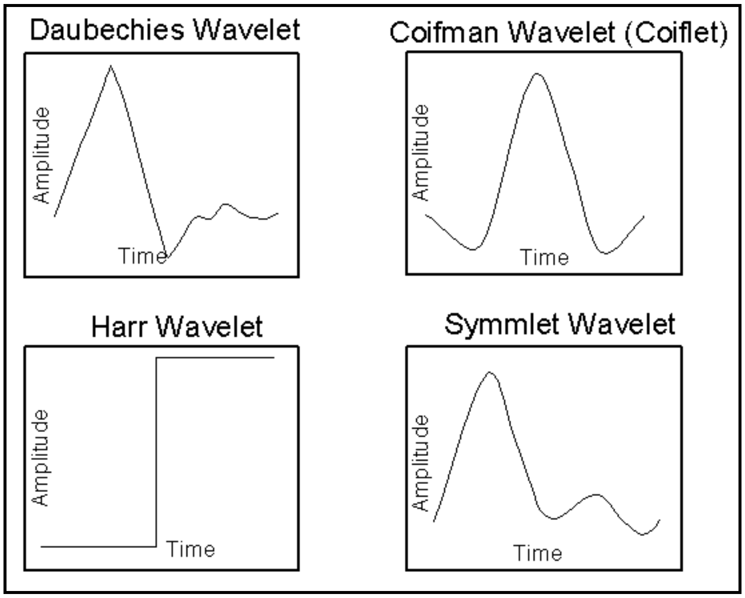
Daubechies Wavelet
Amplitude
Time
Coifman Wavelet (Coiflet)
Amplitude
Time
Harr Wavelet
Amplitude
Time
Symmlet Wavelet
Amplitude
Time

Figure 8. Sample Wavelet Functions.

of the more common ones are shown in Figure 8. There are several wavelets in each family, and they may look different than those shown. Somewhat longer in duration than these functions, but significantly shorter than infinite sinusoids is the cosine packet shown in Figure 9.

Wavelet Comparison With Fourier Analysis

While a typical Fourier transform provides frequency content information for samples within a given time interval, a perfect wavelet transform records the start of one frequency (or event), then the start of a second event, with amplitude added to or subtracted from, the base event.

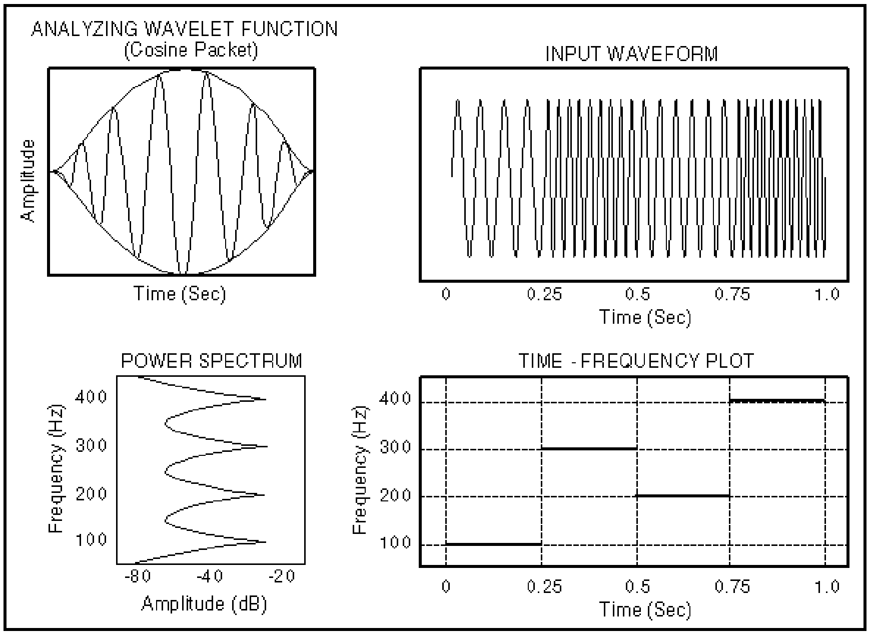

Figure 9. Sample Wavelet Analysis.

Example 1.

Wavelets are especially useful in analyzing transients or time-varying signals. The input signal shown in Figure 9 consists of a sinusoid whose frequency changes in stepped increments over time. The power of the spectrum is also shown. Classical Fourier analysis will resolve the frequencies but cannot provide any information about the times at which each occurs. Wavelets provide an efficient means of analyzing the input signal so that frequencies and the times at which they occur can be resolved. Wavelets have finite duration and must also satisfy additional properties beyond those normally associated with standard windows used with Fourier analysis. The result after the wavelet transform is applied is the plot shown in the lower right. The wavelet analysis correctly resolves each of the frequencies and the time when it occurs. A series of wavelets is used in example 2.

Example 2.

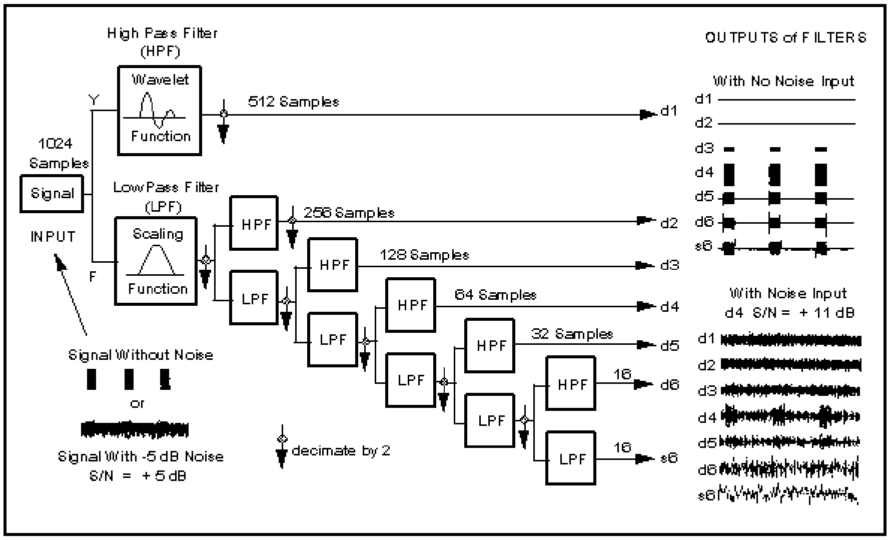
High Pass Filter (HPF)
Wavelet
Function
Y
512 Samples
d1
1024 Samples
Signal
INPUT
Low Pass Filter (LPF)
Scaling
Function
F
HPF
LPF
256 Samples
d2
128 Samples
d3
64 Samples
d4
32 Samples
d5
16
d6
16
s6
Signal Without Noise
or
Signal With -5 dB Noise S/N = + 5 dB
decimate by 2
OUTPUTS of FILTERS
With No Noise Input
d1
d2
d3
d4
d5
d6
s6
With Noise Input
d4 S/N = + 11 dB
d1
d2
d3
d4
d5
d6
s6

Figure 10. Example 2 Analysis Wavelet.

Figure 10 shows the input of a clean signal, and one with noise. It also shows the output of a number of “filters” with each signal. A 6 dB S/N improvement can be seen from the d4 output. (Recall from Section 4.3 that 6 dB corresponds to doubling of detection range.) In the filter cascade, the HPFs and LPFs are the same at each level. The wavelet shape is related to the HPF and LPF in that it is the “impulse response” of an infinite cascade of the HPFs and LPFs. Different wavelets have different HPFs and LPFs. As a result of decimating by 2, the number of output samples equals the number of input samples.

Wavelet Applications

Some fields that are making use of wavelets are: astronomy, acoustics, nuclear engineering, signal and image processing (including fingerprinting), neurophysiology, music, magnetic resonance imaging, speech discrimination, optics, fractals, turbulence, earthquake-prediction, radar, human vision, and pure mathematics applications.

See Andrew Bruce, David Donoho, and Hong-Ye Gao, “Wavelet Analysis,” *IEEE Spectrum*, Vol. 33 No. 10, October 1996.

ANTENNAS

Antenna Introduction / Basics 3-1
Polarization 3-2
Radiation Patterns 3-3
Frequency / Phase Effects of Antennas 3-4
Antenna Near Field 3-5
Radiation Hazards 3-6
Active Electronically Scanned Arrays (AESA) 3-7
Fractal Antennas 3-8

This page intentionally left blank.

ANTENNA INTRODUCTION / BASICS

Rules of Thumb:

1. The power gain of an antenna with losses, excluding input impedance mismatch, is given by:

$$G \approx \frac{4\pi\eta A}{\lambda^2} \quad \text{Where} \quad \begin{aligned} \eta &= Efficiency \\ A &= Physical\ aperture\ area \\ \lambda &= wavelength \end{aligned}$$

another is :

$$G = \frac{X\eta}{BW_\phi BW_\theta}$$

Where $BW_{\theta \text{ and } \phi}$ *are the elev & az beamwidths in degrees.*

For approximating an antenna pattern with :

(1) A rectangle; $X = 41253,\ \eta_{typical} = 0.7$

(2) An ellipsoid; $X = 52525,\ \eta_{typical} = 0.55$

2. Directive gain of rectangular X-Band Aperture
   G = 1.4 LW Where: Length (L) and Width (W) are in cm

3. Power gain of Circular X-Band Aperture
   $G = d^2\eta$ Where: d = antenna diameter in cm
   η = aperture efficiency

4. Directive gain of an imaginary isotropic antenna radiating in a uniform spherical pattern is one (0 dB).

5. Antenna with a 20 degree beamwidth has approximately a 20 dB directive gain.

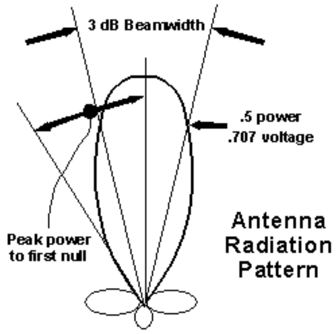

6. 3 dB beamwidth is approximately equal to the angle from the peak of the power to the first null (see figure at right).

7. Parabolic Antenna Beamwidth: $BW = \frac{70\lambda}{d}$

   Where: BW = antenna beamwidth; λ = wavelength; d = antenna diameter.

The antenna equations which follow relate to Figure 1 as a typical antenna. In Figure 1, BW_φ is the azimuth beamwidth and BW_θ is the elevation beamwidth. Beamwidth is normally measured at the half-power or -3 dB point of the main lobe unless otherwise specified. See Glossary.

The gain or directivity of an antenna is the ratio of the radiation intensity in a given direction to the radiation intensity averaged over all directions.

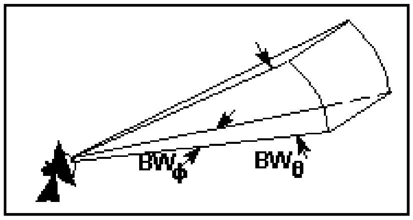

Figure 1. Antenna Aperture.

Quite often directivity and gain are used interchangeably and it sometimes leads to overly optimistic antenna performance estimations. The difference is that directivity is based solely on antenna pattern shape estimation where antenna losses such as dielectric, ohmic resistance, and polarization mismatch are neglected. If these losses are included in the antenna gain calculations, the antenna gain is then referred to as the power gain. Moreover, if additional impedance mismatch or VSWR losses are included in the antenna system gain estimation, the antenna gain calculations are then referred to as the realized gain. However, using directive gain (or directivity) calculations is very convenient in practice for a first order idealized antenna performance estimation.

Normalizing a radiation pattern by the integrated total power yields the directivity of the antenna. This concept in shown in equation form by:

$$D(\theta,\phi) = 10\,\mathrm{Log}\left[\frac{4\pi P\ (\theta,\phi)}{\iint P_{in}(\theta,\phi)\,\mathrm{Sin}\,\theta\, d\theta\, d\phi}\right] \qquad \begin{matrix} 0<\phi\le 360^\circ \\ 0<\theta\le 180^\circ \end{matrix} \tag{1}$$

Where $D(\theta,\varphi)$ is the directivity in dB, and the radiation pattern power in a specific direction is $P_d(\theta,\varphi)$, which is normalized by the total integrated radiated power. Another important concept is that when the angle in which the radiation is constrained is reduced, the directive gain goes up. For example, using an isotropic radiating source, the gain would be 0 dB by definition (Figure 2(a)) and the power density (P_d) at any given point would be the power in (P_{in}) divided by the surface area of the imaginary sphere at a distance R from the source. If the spacial angle was decreased to one hemisphere (Figure 2(b)), the power radiated, P_{in}, would be the same but the area would be half as much, so the gain would double to 3 dB. Likewise if the angle is a quarter sphere, (Figure 2(c)), the gain would be 6 dB. Figure 2(d) shows a pencil beam. The gain is independent of actual power output and radius (distance) at which measurements are taken.

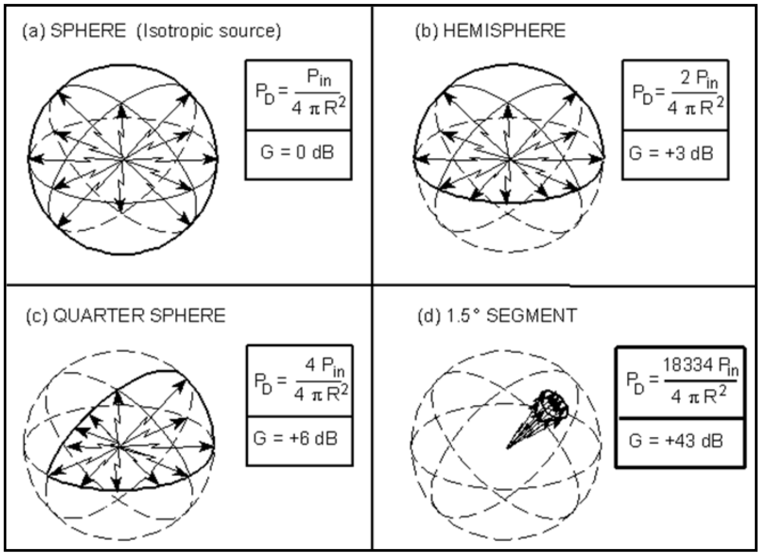
(a) SPHERE (Isotropic source)
$P_D = \frac{P_{in}}{4\pi R^2}$
G = 0 dB
(b) HEMISPHERE
$P_D = \frac{2P_{in}}{4\pi R^2}$
G = +3 dB
(c) QUARTER SPHERE
$P_D = \frac{4P_{in}}{4\pi R^2}$
G = +6 dB
(d) 1.5° SEGMENT
$P_D = \frac{18334P_{in}}{4\pi R^2}$
G = +43 dB

Figure 2. Notional Representation of Directive Antenna Gain.

Real antennas are different, however, and do not have an ideal radiation distribution. Energy varies with angular displacement and losses occur due to sidelobes. However, if we can measure the pattern, and determine the beamwidth we can use two (or more) ideal antenna models to approximate a real antenna pattern as shown in Figure 3.

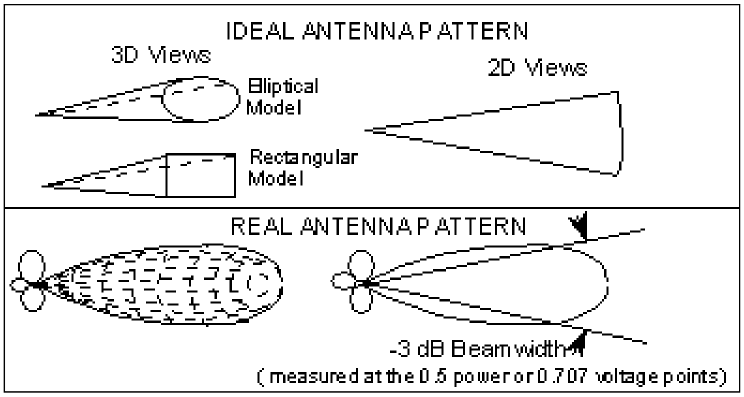
IDEAL ANTENNA PATTERN
3D Views
Elliptical Model
Rectangular Model
2D Views
REAL ANTENNA PATTERN
-3 dB Beamwidth
(measured at the 0.5 power or 0.707 voltage points)

Figure 3. Antenna Beamwidth.

Assuming the antenna pattern is uniform, the gain is equal to the area of the isotropic sphere ($4\pi r^2$) divided by the sector (cross section) area.

$$G = \frac{Area\ of\ Sphere}{Area\ of\ Antenna\ pattern} \tag{2}$$

It can be shown that:

$$G \approx \frac{4\pi}{BW_{\phi\,az}\,BW_{\theta\,el}} \quad or \quad \frac{4\pi}{\phi\theta\ (radians)} \qquad where: \begin{matrix} BW_{\phi\,az} = \text{Azmith beamwidth in radians} \\ BW_{\theta\,el} = \text{Elevation beamwidth in radians} \end{matrix} \tag{3}$$

From this point, two different models are presented:

(1) Approximating an antenna pattern using an elliptical area, and
(2) Approximating an antenna pattern using a rectangular area.

Approximating the antenna pattern as an **elliptical** area:

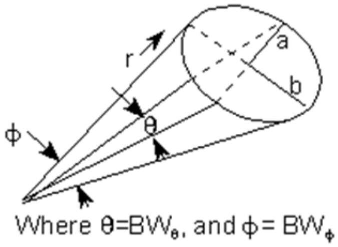

Area of ellipse = π a b = π[(r sin θ)/2][(r sin φ)/2]= (π r^2 sin θ sin φ)/4

$$G = \frac{\textit{Area of Sphere}}{\textit{Area of Antenna pattern}} = (4\,\pi\, r^2)\left(\frac{4}{\pi\, r^2\,\sin\theta\sin\phi}\right) = \frac{16}{\sin\theta\sin\phi}$$

For small angles, sin φ = φ in radians, so:

$$G = \frac{16}{\sin\phi\sin\theta} \approx \frac{16}{\phi\,\theta\,(radians)} = \frac{16}{\phi\,\theta}\left(\frac{360°\;360°}{2\,\pi\;2\,\pi}\right) = \frac{52525}{\phi\,\theta\,(degrees)}\;or\;\frac{52525}{BW_\phi\, BW_\theta\,(degrees)} \qquad \textbf{[4]}$$

The second term in the equation above is very close to equation [3].

For a very directional radar dish with a beamwidth of 1° and an average efficiency of 55%:

Ideally: G = 52525, or in dB form: 10 log G =10 log 52525 = 47.2 dB

With efficiency taken into account, G = 0.55(52525) = 28888, or in log form: 10 log G = 44.6 dB

Approximating the antenna pattern as a **rectangular** area:

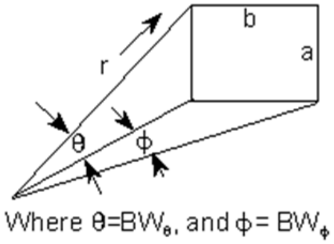

a = r sin θ ,   b = r sin φ, area = ab = r^2 sin θ sin φ

$$G = \frac{\textit{Area of Sphere}}{\textit{Area of Antenna pattern}} = \frac{4\,\pi\, r^2}{r^2\,\sin\theta\sin\phi} = \frac{4\,\pi}{\sin\theta\sin\phi}$$

For small angles, sin φ = φ in radians, so:
The second term in the equation above is identical to equation [3].

$$G = \frac{4\,\pi}{\sin\phi\sin\theta} = \frac{4\,\pi}{\phi\,\theta\,(radians)} = \frac{4\,\pi}{\phi\,\theta}\left(\frac{360°\;360°}{2\,\pi\;2\,\pi}\right) = \frac{41253}{\phi\,\theta\,(degrees)}\;or\;\frac{41253}{BW_\phi\, BW_\theta\,(degrees)} \qquad \textbf{[5]}$$

Converting to dB, $G_{\max}(dB) = 10\ \mathrm{Log}\left[\frac{41253}{BW_\phi\, BW_\theta}\right]$ *with* BW_ϕ *and* BW_θ *in degrees* **[6]**

For a very directional radar dish with a beamwidth of 1° and an average efficiency of 70%:
Ideally (in dB form): 10 log G =10 log 41253 = 46.2 dB.
With efficiency taken into account, G = 0.7(41253) = 28877, or in log form: 10 log G = 44.6 dB

Comparison between elliptical and rectangular areas for antenna pattern models:

By using the rectangular model there is a direct correlation between the development of gain in equation [5] and the ideal gain of equation [3]. The elliptical model has about one dB difference from the ideal calculation, but will yield the same real antenna gain when appropriate efficiencies are assumed.

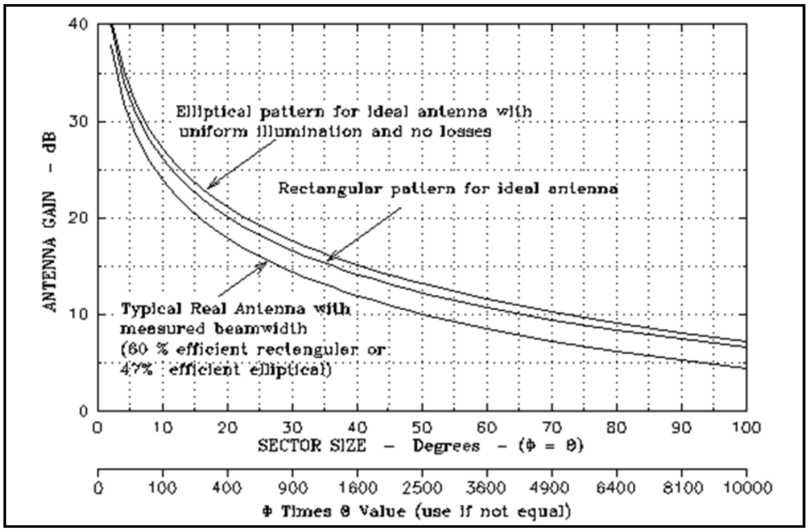

Figure 4. Antenna Sector Size vs. Gain.

The upper plot of Figure 4 shows the gain for an ideal antenna pattern using the elliptical model. The middle plot shows the gain for an ideal antenna using the rectangular model. The lower plot of Figure 4 shows the gain of a typical real antenna (rectangular model using an efficiency of 70% or elliptical model using an efficiency of 47%).

Gain as a function of λ:

When θ = 0, each wave source in Figure 5 is in phase with one another and a maximum is produced in that direction.

Conversely, nulls to either side of the main lobe will occur when the waves radiating from the antenna cancel each other. The first null occurs when there is a phase difference of λ/2 in the wave fronts emanating from the aperture. To aid in visualizing what happens, consider each point in the antenna aperture, from A to C in Figure 5, as a point source of a spherical wave front. If viewed from infinity, the electromagnetic waves from each point interfere with each other, and when, for a particular direction, θ in Figure 5, each wave source has a corresponding point that is one-half wavelength out of phase, a null is produced in that direction due to destructive interference.

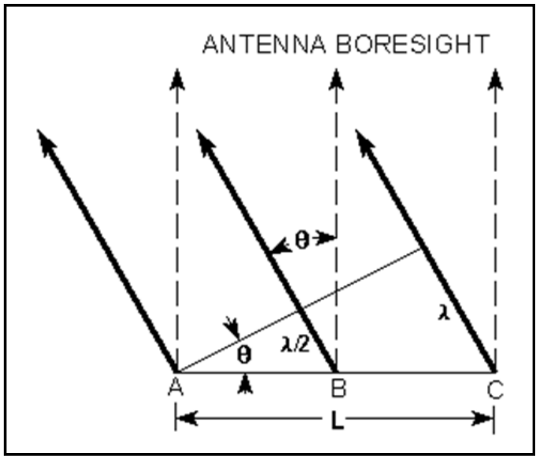

Figure 5. Directional Gain vs. Wavelength.

In Figure 5, the wave emanating from point A is out of phase with the wave from point B by one-half of a wavelength. Hence, they cancel. Similarly, a point just to the right of point A cancels with a point just to the right of point B, and so on across the entire aperture. Therefore, the first null in the radiation pattern is given by:

$$\text{Sin}\ \theta = \lambda/L \text{ and, in radians, } \theta = \lambda/L \text{ (for small angles)} \qquad \textbf{[7]}$$

As the angle off boresight is increased beyond the first null, the intensity of the radiation pattern rises then falls, until the second null is reached. This corresponds to a phase difference of two wavelengths between the left and right edges of the aperture. In this case, the argument proceeds as before, except now the aperture is divided into four segments (point A canceling with a point halfway between A and B, and so on).

The angle θ is the angle from the center (maximum) of the radiation pattern to the first null. The null-to-null beam width is 2θ. Generally, we are interested in the half-power (3 dB) beamwidth. It turns out that this beamwidth is approximately one-half of the null-to-null beamwidth, so that:

$$BW_{3\,dB} \approx (\tfrac{1}{2})(2\theta) = \lambda/L \qquad \textbf{[8]}$$

Therefore, beamwidth is a function of the antenna dimension "L" and the wavelength of the signal. It can be expressed as follows: Note: for circular antennas, L in the following equations = diameter

$$Bw_{\varphi(az)} = \lambda/L_{Az\ eff} \text{ and } BW_{\theta(el)} = \lambda/L_{El\ eff} \qquad \textbf{[9]}$$

Substituting the two variations of equation [9] into equation [3] and since $L_{Az\ eff}$ times $L_{El\ eff} = A_e$ (effective capture area of the antenna), we have:

$$G \approx \frac{4\pi}{BW_{\phi}\, BW_{\theta}\, (radians)} = \frac{4\pi\, L_{az}\, L_{el}}{\lambda^2} = \frac{4\pi\, A_e}{\lambda^2} \qquad \textbf{[10]}$$

Note: Equation is approximate since aperture efficiency isn't included as is done later in equation [12].

The efficiency (discussed later) will reduce the gain by a factor of 30-50%, i.e. real gain = .5 to .7 times theoretical gain.

Unity Gain Antenna.
If a square antenna is visualized and G=1, $A_e = \lambda^2 / 4\pi$. When a dimension is greater than $0.28\,\lambda$ ($\sim\frac{1}{4}\lambda$) it is known as an electrically large antenna, and the antenna will have a gain greater than one (positive gain when expressed in dB). Conversely, when the dimension is less than $0.28\,\lambda$ ($\sim\frac{1}{4}\lambda$)(an electrically small antenna), the gain will be less than one (negative gain when expressed in dB). Therefore, a unity gain antenna can be approximated by an aperture that is $\frac{1}{4}\lambda$ by $\frac{1}{4}\lambda$.

Beamwidth as a Function of Aperture Length
It can be seen from Figure 5, that the wider the antenna aperture (L), the narrower the beamwidth will be for the same λ. Therefore, if you have a rectangular shaped horn antenna, the radiation pattern from the wider side will be narrower than the radiation pattern from the narrow side.

APERTURE EFFICIENCY, η

The Antenna Efficiency, η, is a factor which includes all reductions from the maximum gain. η can be expressed as a percentage, or in dB. Several types of "loss" must be accounted for in the efficiency, η:

(1) Illumination efficiency which is the ratio of the directivity of the antenna to the directivity of a uniformly illuminated antenna of the same aperture size,
(2) Phase error loss or loss due to the fact that the aperture is not a uniform phase surface,
(3) Spillover loss (Reflector Antennas) which reflects the energy spilling beyond the edge of the reflector into the back lobes of the antenna,

(4) Mismatch (VSWR) loss, derived from the reflection at the feed port due to impedance mismatch (especially important for low frequency antennas), and
(5) RF losses between the antenna and the antenna feed port or measurement point.

The aperture efficiency, η_a, is also known as the illumination factor, and includes items (1) and (2) above; it does not result in any loss of power radiated but affects the gain and pattern. It is nominally 0.6-0.8 for a planer array and 0.13 to 0.8 with a nominal value of 0.5 for a parabolic antenna, however η can vary significantly. Other antennas include the spiral (.002-.5), the horn (.002-.8), the double ridge horn (.005-.93), and the conical log spiral (.0017-1.0).

Items (3), (4), and (5) above represent RF or power losses which can be measured. The efficiency varies and generally gets lower with wider bandwidths. Also note that the gain equation is optimized for small angles - see derivation of wavelength portion of equation [7]. This explains why efficiency also gets lower for wider beamwidth antennas.

EFFECTIVE CAPTURE AREA

Effective capture area (A_e) is the product of the physical aperture area (A) and the aperture efficiency (η) or:

$$A_e = \eta\, A = \frac{\lambda^2 G}{4\pi} \qquad \textbf{[11]}$$

GAIN AS A FUNCTION OF APERTURE EFFICIENCY

The Gain of an antenna with losses is given by:

$$G = \frac{4\pi\eta A}{\lambda^2} \quad \begin{array}{c} \eta = \textit{Aperture Efficiency} \\ \textit{Where}\, A = \textit{Physical aperture area} \\ \lambda = \textit{wavelength} \end{array} \qquad \textbf{[12]}$$

Note that the gain is proportional to the aperture area and inversely proportional to the square of the wavelength. For example, if the frequency is doubled, (half the wavelength), the aperture could be decreased four times to maintain the same gain.

BEAM FACTOR

Antenna size and beamwidth are also related by the beam factor defined by:

Beam Factor = $(D/\lambda)\cdot$(Beamwidth) where D = antenna dimension in wavelengths.

The beam factor is approximately invariant with antenna size, but does vary with type of antenna aperture illumination or taper. The beam factor typically varies from 50-70°.

APERTURE ILLUMINATION (TAPER)

The aperture illumination or illumination taper is the variation in amplitude across the aperture. This variation can have several effects on the antenna performance:

(1) reduction in gain,
(2) reduced (lower) sidelobes in most cases, and
(3) increased antenna beamwidth and beam factor.

Tapered illumination occurs naturally in reflector antennas due to the feed radiation pattern and the variation in distance from the feed to different portions of the reflector. Phase can also vary across the aperture which also affects the gain, efficiency, and beamwidth.

CIRCULAR ANTENNA GAIN

Solving equation [12] in dB, for a circular antenna with area $\pi D^2/4$, we have:

$$10 \text{ Log } G = 20 \text{ Log } (D/\lambda) + 10 \text{ Log } (\eta) + 9.94 \text{ dB}; \quad \text{where } D = \text{diameter} \qquad \textbf{[13]}$$

This data is depicted in the nomograph of Figure 6. For example, a six foot diameter antenna operating at 9 GHz would have approximately 44.7 dB of gain as shown by the dashed line drawn on Figure 6. This gain is for an antenna 100% efficient, and would be 41.7 dB for a typical parabolic antenna (50% efficient).

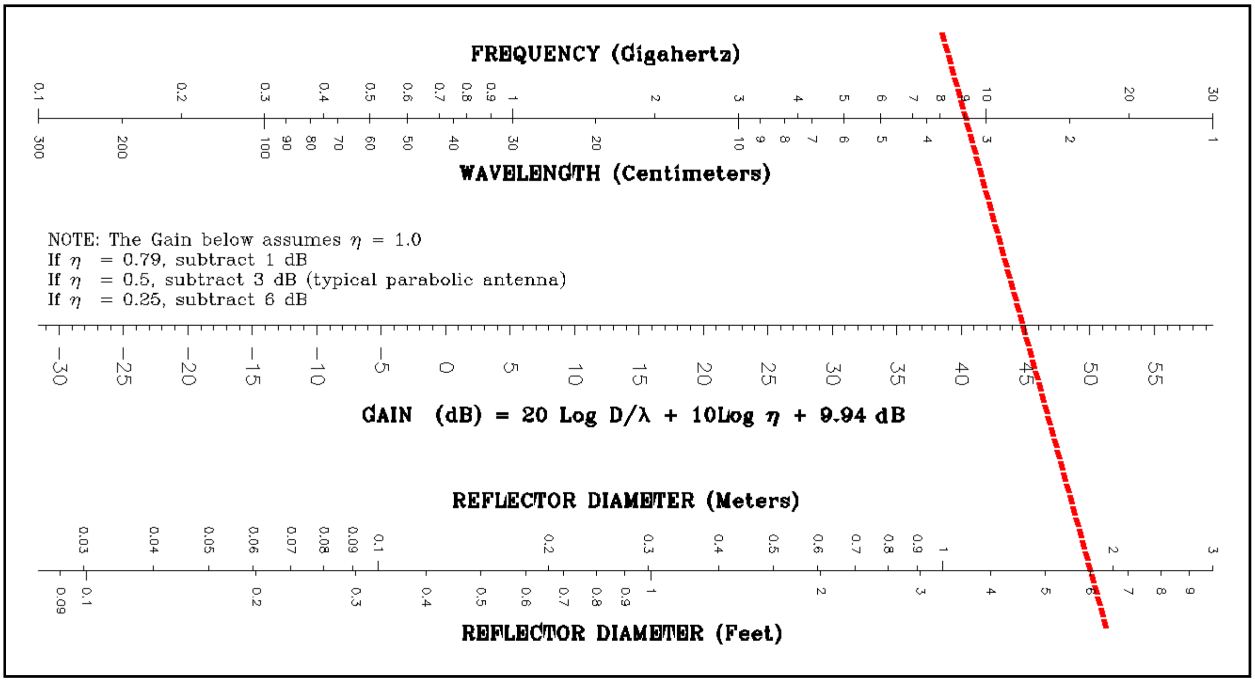

Figure 6. Antenna Gain Nomograph.

An example of a typical antenna (with losses) showing the variation of gain with frequency is depicted in Figure 7, and the variation of gain with antenna diameter in Figure 8. The circle on the curves in Figure 7 and 8 correspond to the Figure 6 example and yields 42 dB of gain for the 6 ft dish at 9 GHz.

Example Problem: If the two antennas in the drawing are "welded" together, how much power will be measured at point A? (Line loss $L_1 = L_2 = 0.5$, and $10\log L_1$ or $L_2 = 3$ dB)

Multiple choice:

A. 16 dBm b. 28 dBm c. 4 dBm d. 10 dBm e. < 4 dBm

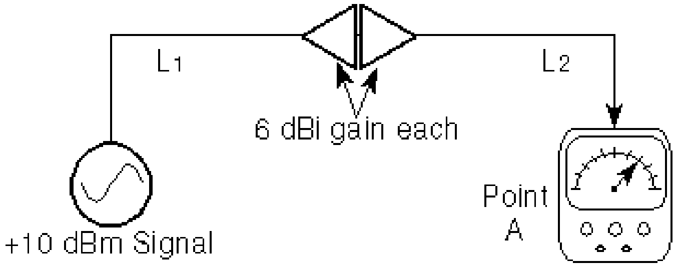

Answer:

The antennas do not act as they normally would since the antennas are operating in the near field. They act as inefficient coupling devices resulting in some loss of signal. In addition, since there are no active components, you cannot end up with more power than you started with. The correct answer is "e. < 4 dBm."

10 dBm - 3 dB - small loss -3 dB = 4 dBm - small loss

If the antennas were separated by 5 ft and were in the far field, the antenna gain could be used with space loss formulas to calculate (at 5 GHz): 10 dBm - 3 dB + 6 dB - 50 dB (space loss) + 6 dB -3 dB = -34 dBm (a much smaller signal).

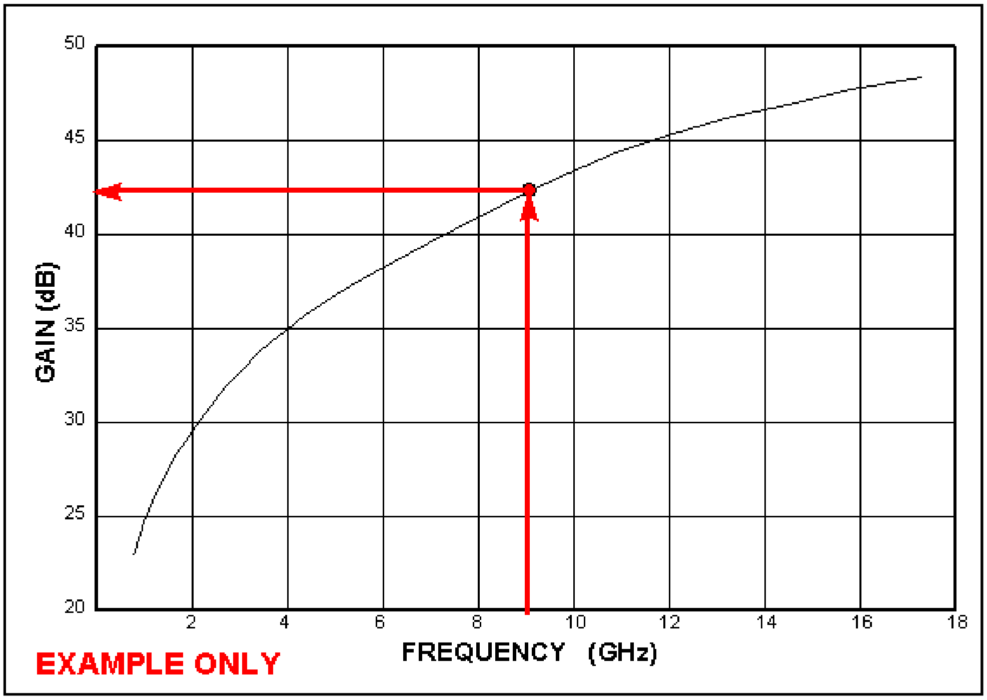

Figure 7. Gain of a Typical 6-Foot Dish Antenna (With Losses).

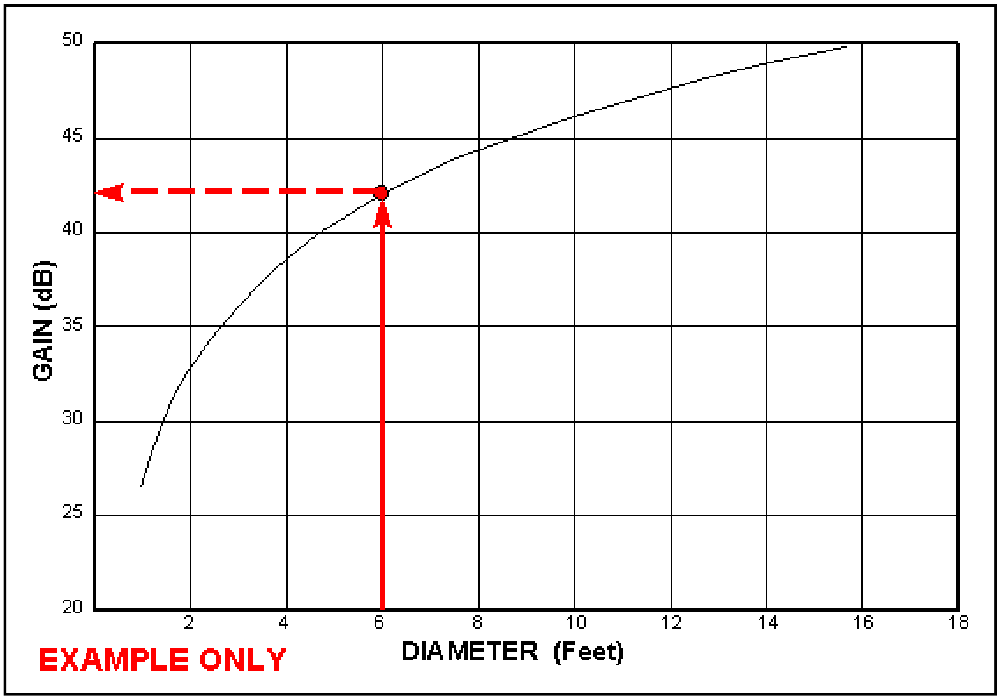

Figure 8. Gain of a Typical Dish at 9 GHz (With Losses).

This page intentionally left blank.

POLARIZATION

Table 1 shows the **theoretical** ratio of power transmitted between antennas of different polarization. These ratios are seldom fully achieved due to effects such as reflection, refraction, and other wave interactions, so some practical ratios are also included.

Table 1. Polarization Loss for Various Antenna Combinations.

Transmit Antenna Polarization	**Receive Antenna Polarization**	**Ratio of Power Received to Maximum Power**					
		Theoretical		Practical Horn		Practical Spiral	
		Ratio in dB	as Ratio	Ratio in dB	as Ratio	Ratio in dB	as Ratio
Vertical	Vertical	0 dB	1	*	*	N/A	N/A
Vertical	Slant (45° or 135°)	-3 dB	½	*	*	N/A	N/A
Vertical	Horizontal	- ∞ dB	0	-20 dB	1/100	N/A	N/A
Vertical	Circular (right-hand or left-hand)	-3 dB	½	*	*	*	*
Horizontal	Horizontal	0 dB	1	*	*	N/A	N/A
Horizontal	Slant (45° or 135°)	-3 dB	½	*	*	N/A	N/A
Horizontal	Circular (right-hand or left-hand)	-3 dB	½	*	*	*	*
Circular (right-hand)	Circular (right-hand)	0 dB	1	*	*	*	*
Circular (right-hand)	Circular (left-hand)	- ∞ dB	0	-20 dB	1/100	-10 dB	1/10
Circular (right or left)	Slant (45° or 135°)	-3 dB	½	*	*	*	*

* Approximately the same as theoretical.
Note: Switching transmit and receive antenna polarization will give the same results.

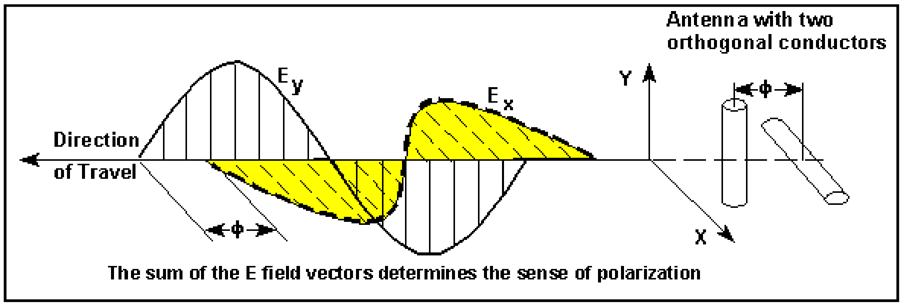
Antenna with two orthogonal conductors
Ey
Ex
Y
X
ϕ
Direction of Travel
ϕ
The sum of the E field vectors determines the sense of polarization

Figure 1. Polarization Coordinates.

The polarization of an electromagnetic wave is defined as the orientation of the electric field vector. Recall that the electric field vector is perpendicular to both the direction of travel and the magnetic field vector. The polarization is described by the geometric figure traced by the electric field vector upon a stationary plane perpendicular to the direction of propagation, as the wave travels through that plane. An electromagnetic wave is frequently composed of (or can be broken down into) two orthogonal components as shown in Figure 1. This may be due to the arrangement of power input leads to various points on a flat antenna, or due to an interaction of active elements in an array, or many other reasons.

The geometric figure traced by the sum of the electric field vectors over time is, in general, an ellipse as shown in Figure 2. Under certain conditions the ellipse may collapse into a straight line, in which case the polarization is called linear.

In the other extreme, when the two components are of equal magnitude and 90° out of phase, the ellipse will become circular as shown in Figure 3. Thus linear and circular polarization are the two special cases of elliptical polarization. Linear polarization may be further classified as being vertical, horizontal, or slant.

Figure 2 depicts plots of the E field vector while varying the relative amplitude and phase angle of its component parts.

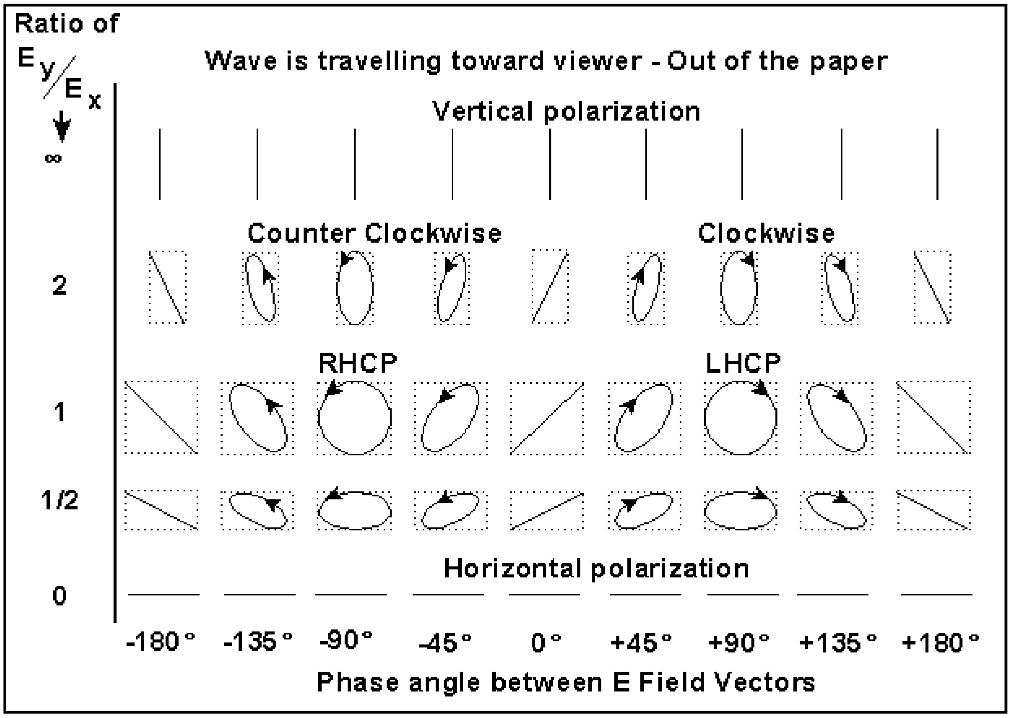
Ratio of
E y / E x
∞
2
1
1/2
0
Wave is travelling toward viewer - Out of the paper
Vertical polarization
Counter Clockwise
Clockwise
RHCP
LHCP
Horizontal polarization
-180°
-135°
-90°
-45°
0°
+45°
+90°
+135°
+180°
Phase angle between E Field Vectors

Figure 2. Polarization as a Function of E_y / E_x Ratio and Phase Angle.
Adopted from J.D. Kraus, "Antennas," 2nd ed., Figure 2-37

For a linearly polarized antenna, the radiation pattern is taken both for a co-polarized and cross polarized response. The polarization quality is expressed by the ratio of these two responses. The ratio between the responses must typically be great (30 dB or greater) for an application such as cross-polarized jamming. For general applications, the ratio indicates system power loss due to polarization mismatch. For circularly polarized antennas, radiation patterns are usually taken with a rotating linearly polarized reference antenna. The reference antenna rotates many times while taking measurements around the azimuth of the antenna that is being tested. The resulting antenna pattern is the linear polarized gain with a cyclic ripple. The peak-to-peak value is the axial ratio, and represents the polarization quality for a circular polarized antenna. The typical RWR antenna has a maximum 3 dB axial ratio within 45° of boresight.

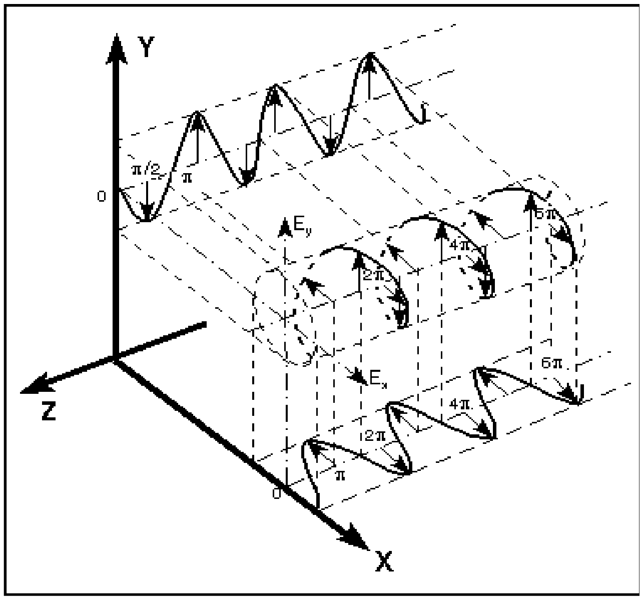
Y
Z
X
0
π/2
π
E_y
2π
4π
5π
E_x
6π
0
π
2π
4π

Figure 3. Circular Polarization – E Field.

For any antenna with an aperture area, as the aperture is rotated, the viewed dimension along the axis remains constant, while the other viewed dimension decreases to zero at 90° rotation. The axial ratio of an antenna will get worse as the antenna is rotated off boresight because the field contribution from the axial component will remain fairly constant and the other orthogonal component will decrease with rotation.

The sense of antenna polarization is defined from a viewer positioned behind an antenna looking in the direction of propagation. The polarization is specified as a transmitting, not receiving antenna regardless of intended use.

We frequently use “hand rules” to describe the sense of polarization. The sense is defined by which hand would be used in order to point that thumb in the direction of propagation and point the fingers of the same hand in the direction of rotation of the E field vector. For example, referring to Figure 4, if your thumb is pointed in the direction of propagation and the rotation is counterclockwise looking in the direction of travel, then you have left hand circular polarization.

Optics people view an aperture from the front and therefore use the opposite reference.

The polarization of a linearly polarized horn antenna can be directly determined by the orientation of the feed probe, which is in the direction of the E-field.

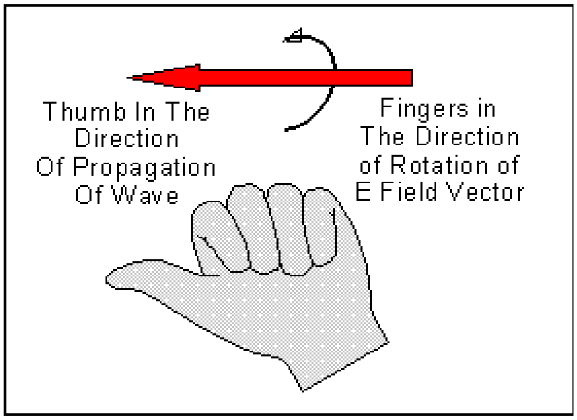
Thumb In The Direction Of Propagation Of Wave
Fingers in The Direction of Rotation of E Field Vector

Figure 4. Left Hand Polarization.

In general, a flat surface or sphere will reflect a linearly polarized wave with the same polarization as received. A horizontally polarized wave may get extended range because of water and land surface reflections, but signal cancellation will probably result in “holes” in coverage. Reflections will reverse the sense of circular polarization.

If the desired antenna is used for receiving a direct transmission as shown in Figure 5, the same polarization sense (specified if transmitting) is required for maximum signal reception in this situation. Buy two right-hand or two left-hand circularly polarized antennas for this case. When you procure antennas, remember that the polarization is specified as if transmitting, regardless of intended use.

Wave propagation between two identical antennas is analogous to being able to thread a nut from one bolt to an identical opposite facing bolt.

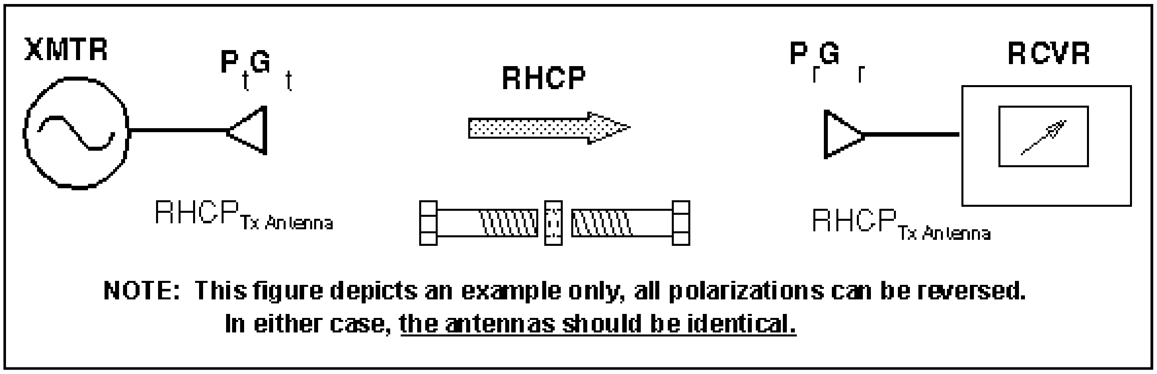

NOTE: This figure depicts an example only, all polarizations can be reversed. In either case, the antennas should be identical.

Figure 5. Same Circular Polarization.

If the desired antenna is used for a receiving a wave with a single or odd number of reflections, such as a bistatic radar where separate antennas are used for transmit and receive as shown in Figure 6, then opposite circularly polarized antennas would be used for maximum signal reception. In this case buy antennas of opposite polarization sense (one left hand and one right hand).

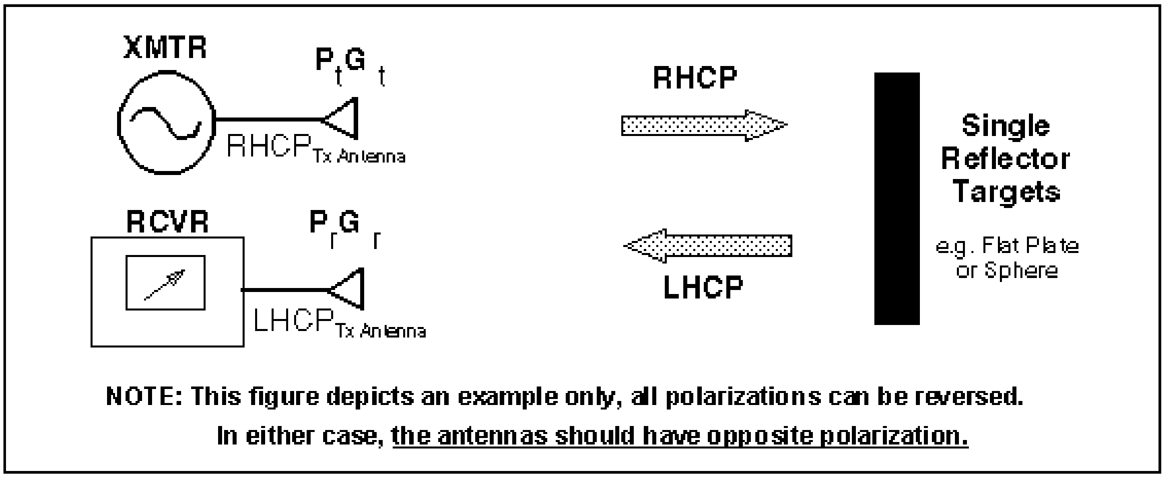
XMTR
P_tG_t
RHCP Tx Antenna
RCVR
P_rG_r
LHCP Tx Antenna
RHCP
LHCP
Single Reflector Targets
e.g. Flat Plate or Sphere
NOTE: This figure depicts an example only, all polarizations can be reversed.
In either case, the antennas should have opposite polarization.

Figure 6. Opposite Circular Polarization.

In a corner reflector, waves reflect twice before returning to the receiver as shown in Figure 7, consequently they return with the same sense as they were transmitted. In this case (or any even number of reflections) buy antennas of the same polarization sense.

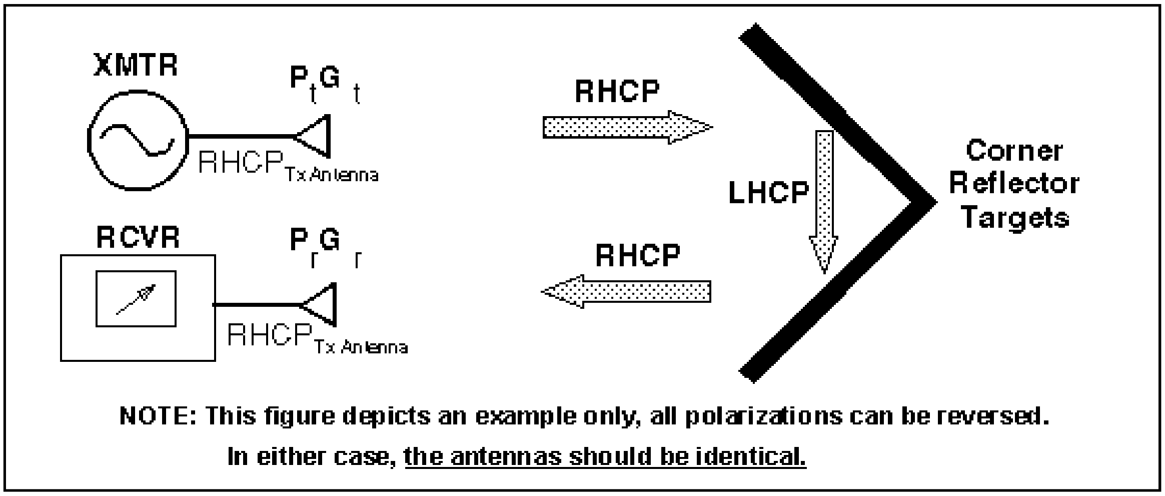
XMTR
P_tG_t
RHCP Tx Antenna
RCVR
P_rG_r
RHCP Tx Antenna
RHCP
LHCP
RHCP
Corner Reflector Targets
NOTE: This figure depicts an example only, all polarizations can be reversed.
In either case, the antennas should be identical.

Figure 7. Circular Polarization With Corner Reflector.

An aircraft acts as both a corner reflector and a "normal" reflector so the return has mixed polarization. Most airborne radars use the same antenna for transmitting and receiving in order to receive the corner reflections and help exclude receipt of reflections from rain (single polarization reversal), however in doing so there is about a 5-9 dB loss from the ideal receiver case. It should be noted that the return from raindrops is attenuated by approximately 20 dB.

RADIATION PATTERNS

The radiation pattern is a graphical depiction of the relative field strength transmitted from or received by the antenna. Antenna radiation patterns are taken at one frequency, one polarization, and one plane cut. The patterns are usually presented in polar or rectilinear form with a dB strength scale. Patterns are normalized to the maximum graph value, 0 dB, and a directivity is given for the antenna. This means that if the side lobe level from the radiation pattern were down -13 dB, and the directivity of the antenna was 4 dB, then the sidelobe gain would be -9 dB.

Figures 1 to 14 on the pages following depict various antenna types and their associated characteristics. The patterns depicted are those which most closely match the purpose for which the given shape was intended. In other words, the radiation pattern can change dramatically depending upon frequency, and the wavelength to antenna characteristic length ratio. See Section 3-4. Antennas are designed for a particular frequency. Usually the characteristic length is a multiple of $\lambda/2$ minus 2-15% depending on specific antenna characteristics.

The gain is assumed to mean directional gain of the antenna compared to an isotropic radiator transmitting to or receiving from all directions.

The half-power (-3 dB) beamwidth is a measure of the directivity of the antenna.

Polarization, which is the direction of the electric (not magnetic) field of an antenna is another important antenna characteristic. This may be a consideration for optimizing reception or jamming.

The bandwidth is a measure of how much the frequency can be varied while still obtaining an acceptable VSWR (2:1 or less) and minimizing losses in unwanted directions. See Glossary, Section 10.

A 2:1 VSWR corresponds to a 9.5 dB (or 10%) return loss - see Section 6-2.

Two methods for computing antenna bandwidth are used:

Narrowband by %, $B = \left(\frac{F_U - F_L}{F_C}\right)(100)$, where F_C = Center frequency

Broadband by ratio, $B = \frac{F_U}{F_L}$

An antenna is considered broadband if $F_U / F_L > 2$. The table at the right shows the equivalency of the two, however the shaded values are not normally used because of the aforementioned difference in broadband/narrowband.

Bandwidth	
%	Ratio
5	1.05 : 1
10	1.11 : 1
20	1.22 : 1
30	1.35 : 1
40	1.50 : 1
50	1.67 : 1
60	1.85 : 1
67	2 : 1
100	3 : 1
120	4 : 1
133	5 : 1
150	7 : 1
160	9 : 1
163	10 : 1

Should there be ever a need to express bandwidth of an antenna in one or the other alternative formats, a conversion between the two narrowband and broadband bandwidth quantities can be easily calculated using the following relationships:

Calculate broadband ratio B_{bb} given narrowband B%,

$$B_{bb} = (200 + B\%)/(200 - B\%) \qquad \textbf{[1]}$$

Calculate narrowband B% given broadband ratio B_{bb},

$$B\% = 200 * (Bbb - 1)/(Bbb + 1) \quad \textbf{[2]}$$

For an object that experiences a plane wave, the resonant mode is achieved when the dimension of the object is nλ/2, where n is an integer. Therefore, one can treat the apertures shown in the following figure as half wave length dipole antennas for receiving and reflecting signals. More details are contained in Section 8-4.

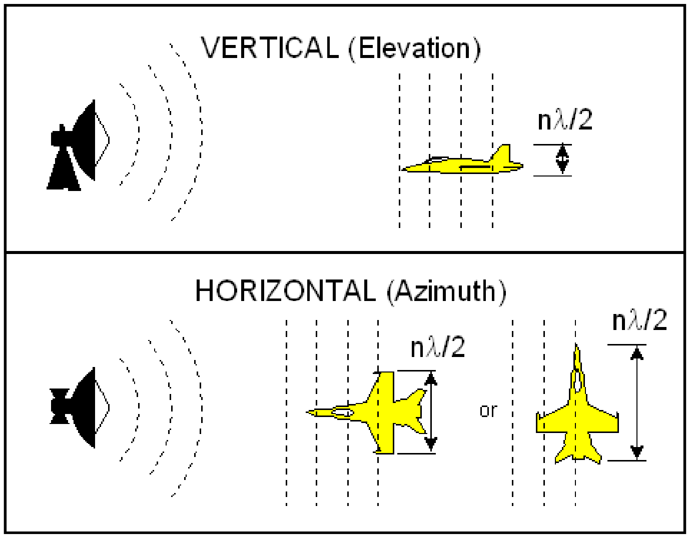

The following lists antenna types by page number. The referenced page shows frequency limits, polarizations, etc.

Type	Page	Type	Page
4 arm conical spiral	3-3.6	log periodic	3-3.8
alford loop	3-3.4	loop, circular	3-3.4
aperture synthesis	3-3.8	loop, alford	3-3.4
array	3-3.8	loop, square	3-3.4
axial mode helix	3-3.5	luneberg lens	3-3.9
biconical w/polarizer	3-3.6	microstrip patch	3-3.9
biconical	3-3.6	monopole	3-3.3
cavity backed circuit fed slot	3-3.5	normal mode helix	3-3.5
cavity backed spiral	3-3.9	parabolic	3-3.7
circular loop	3-3.4	patch	3-3.9
conical spiral	3-3.5	reflector	3-3.10
corner reflector	3-3.9	rhombic	3-3.3
dipole array, linear	3-3.8	sinuous, dual polarized	3-3.6
dipole	3-3.3	slot, guide fed	3-3.9
discone	3-3.4	slot, cavity backed	3-3.9
dual polarized sinuous	3-3.6	spiral, 4 arm conical	3-3.6
guide fed slot	3-3.9	spiral, conical	3-3.5
helix, normal mode	3-3.5	spiral, cavity backed	3-3.5
helix, axial mode	3-3.5	square loop	3-3.4
horn	3-3.7	vee	3-3.3
linear dipole array	3-3.8	yagi	3-3.8

Antenna Type	Radiation Pattern	Characteristics
MONOPOLE 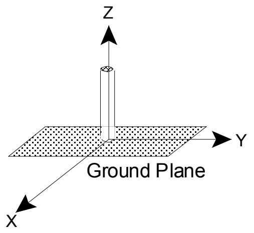 	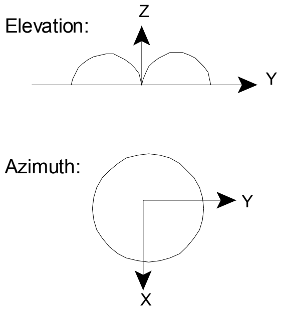 	**Polarization: Linear** Vertical as shown **Typical Half-Power Beamwidth** 45 deg x 360 deg **Typical Gain:** 2-6 dB at best **Bandwidth:** 10% or 1.1:1 **Frequency Limit** **Lower:** None **Upper:** None **Remarks:** Polarization changes to horizontal if rotated to horizontal
λ/2 DIPOLE 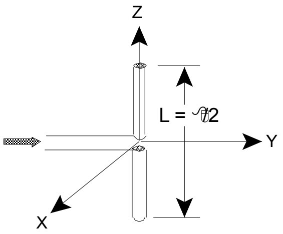 	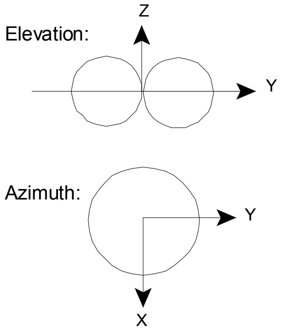 	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 80 deg x 360 deg **Typical Gain:** 2 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit** **Lower:** None **Upper:** 8 GHz (practical limit) **Remarks:** Pattern and lobing changes significantly with L/f. Used as a gain reference < 2 GHz.

Figure 1. Monopole and Dipole Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
VEE 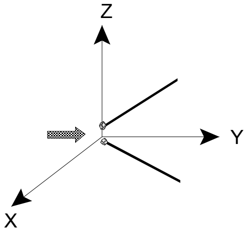 	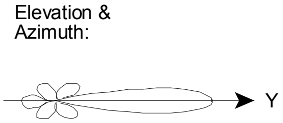 	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 60 deg x 60 deg **Typical Gain:** 2 to 7 dB **Bandwidth:** "Broadband" **Frequency Limit** **Lower:** 3 MHz **Upper:** 500 MHz (practical limits) **Remarks:** 24KHz versions are known to exist. Terminations may be used to reduce backlobes.
RHOMBIC 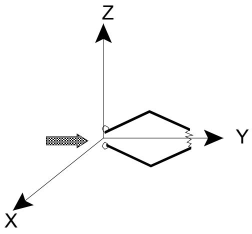 	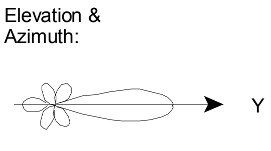 	**Polarization**: Linear Vertical as shown **Typical Half-Power Beamwidth** 60 deg x 60 deg **Typical Gain:** 3 dB **Bandwidth:** "Broadband" **Frequency Limit** **Lower:** 3 MHz **Upper:** 500 MHz **Remarks:** Termination resistance used to reduce backlobes.

Figure 2. Vee and Rhombic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CIRCULAR LOOP (Small) 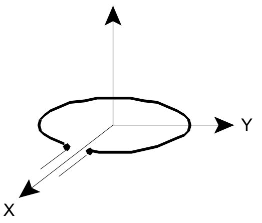 	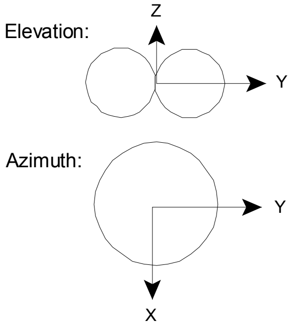 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 80 deg x 360 deg **Typical Gain:** -2 to 2 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 1 GHz
SQUARE LOOP (Small) 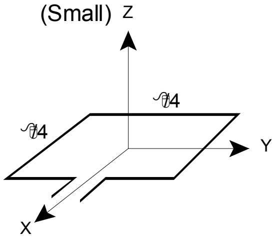 	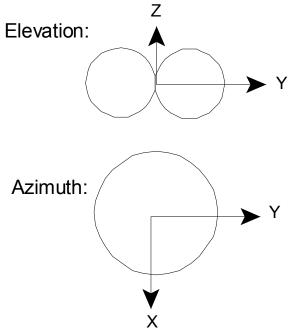 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 100 deg x 360 deg **Typical Gain:** 1-3 dB **Bandwidth:** 10% or 1.1:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 1 GHz

Figure 3. Circular Loop and Square Loop Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
DISCONE 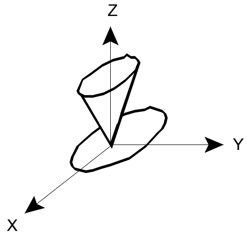 	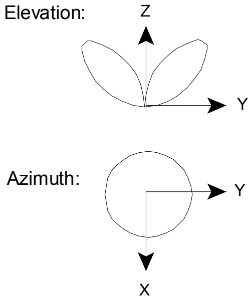 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Vertical as shown **Typical Half-Power Beamwidth:** 20-80 deg x 360 deg **Typical Gain:** 0-4 dB **Bandwidth:** 100% or 3:1 **Frequency Limit:** **Lower:** 30 MHz **Upper:** 3 GHz
ALFORD LOOP 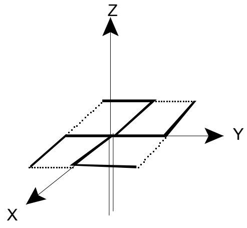 	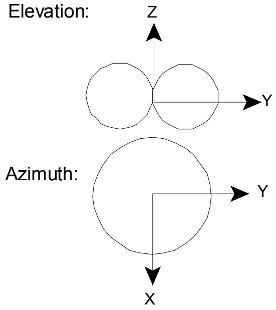 Elevation: Z Y Azimuth: Y X	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth:** 80 deg x 360 deg **Typical Gain:** -1 dB **Bandwidth:** 67% or 2:1 **Frequency Limit:** **Lower:** 100 MHz **Upper:** 12 GHz

Figure 4. Discone and Alford Loop Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
AXIAL MODE HELIX 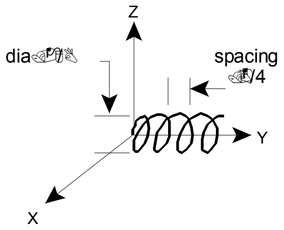 Z dia λ/π spacing λ/4 Y X	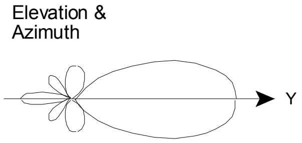 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 50 deg x 50 deg **Typical Gain:** 10 dB **Bandwidth:** 52% or 1.7:1 **Frequency Limit** **Lower:** 100 MHz **Upper:** 3 GHz **Remarks:** Number of loops >3
NORMAL MODE HELIX 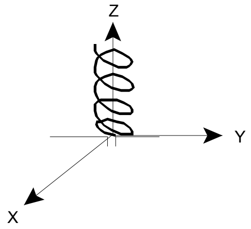 	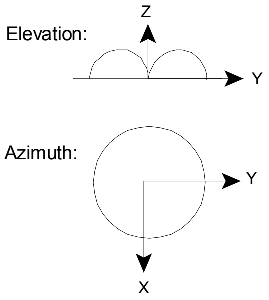 Elevation: Z Y Azimuth: Y X	**Polarization:** Circular - with an ideal pitch to diameter ratio. **Typical Half-Power Beamwidth:** 60 deg x 360 deg **Typical Gain:** 0 dB **Bandwidth:** 5% or 1.05:1 **Frequency Limit** **Lower:** 100 MHz **Upper:** 3 GHz

Figure 5. Axial Mode Helix and Normal Mode Helix Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CAVITY BACKED SPIRAL (Flat Helix) 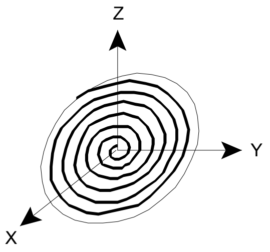 	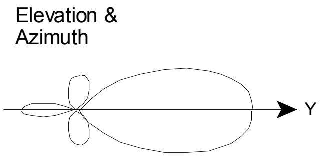 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 60 deg x 90 deg **Typical Gain:** 2-4 dB **Bandwidth:** 160% or 9:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz
CONICAL SPIRAL 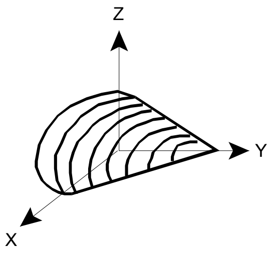 	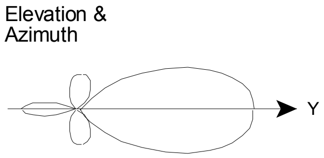 Elevation & Azimuth Y	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 60 deg x 60 deg **Typical Gain:** 5-8 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 18 GHz

Figure 6. Cavity Backed Spiral and Conical Spiral Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
4 ARM CONICAL SPIRAL 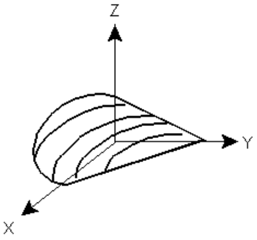 	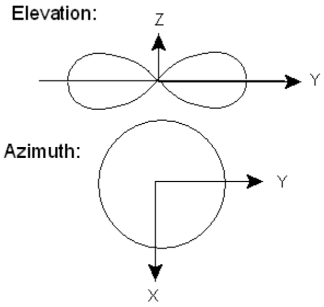 	**Polarization:** Circular Left hand as shown **Typical Half-Power Beamwidth:** 50 deg x 360 deg **Typical Gain:** 0 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz
DUAL POLARIZED SINUOUS 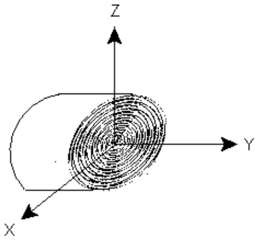 	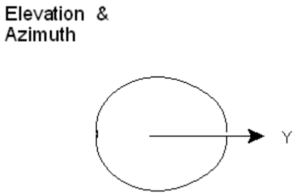 	**Polarization**: Dual vertical or horizontal or dual Circular right hand or left hand with hybrid **Typical Half-Power Beamwidth:** 75 deg x 75 deg **Typical Gain:** 2 dB **Bandwidth:** 163% or 10:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 18 GHz

Figure 7. Dual Polarized Sinuous Antenna Characteristics and 4 Arm Conical Spiral.

Antenna Type	Radiation Pattern	Characteristics
BICONICAL 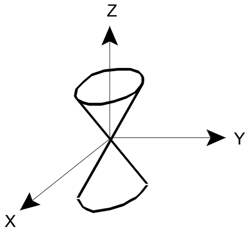 	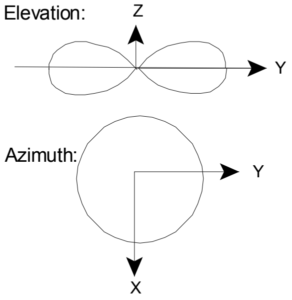 	**Polarization:** Linear, Vertical as shown **Typical Half-Power Beamwidth:** 20-100 deg x 360 deg **Typical Gain:** 0-4 dB **Bandwidth:** 120% or 4:1 **Frequency Limit:** **Lower:** 500 MHz **Upper:** 40 GHz
BICONICAL W/POLARIZER 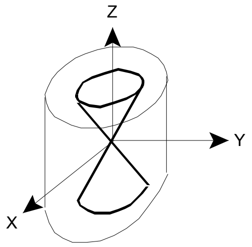 	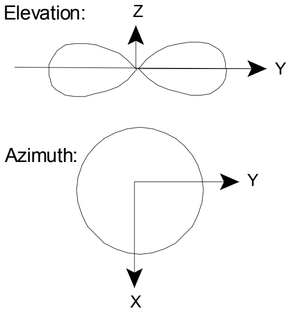 	**Polarization:** Circular, Direction depends on polarization **Typical Half-Power Beamwidth:** 20-100 deg x 360 deg **Typical Gain:** -3 to 1 dB **Bandwidth:** 100% or 3:1 **Frequency Limit:** **Lower:** 2 GHz **Upper:** 18 GHz

Figure 8. Biconical and Biconical With Polarizer Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
HORN 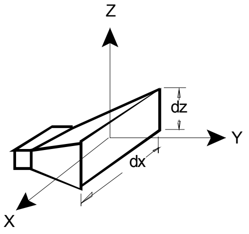 	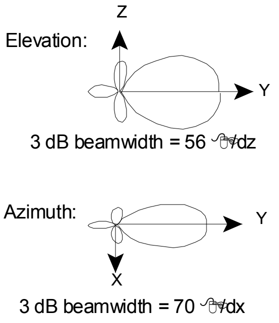 Z Elevation: Y Azimuth: Y X 3 dB beamwidth = 56 🖰/dz 3 dB beamwidth = 70 🖰/dx	**Polarization:** Linear **Typical Half-Power Beamwidth:** 40 deg x 40 deg **Typical Gain:** 5 to 20 dB **Bandwidth:** If ridged: 120% or 4:1 If not ridged: 67% or 2:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 40 GHz
HORN W / POLARIZER 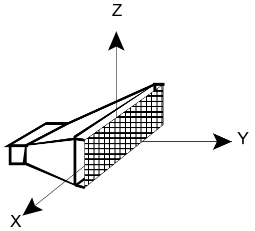 	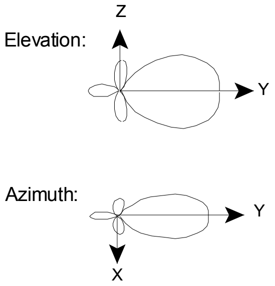 Z Elevation: Y Azimuth: Y X	**Polarization:** Circular, Depends on polarizer **Typical Half-Power Beamwidth:** 40 deg x 40 deg **Typical Gain:** 5 to 10 dB **Bandwidth:** 60% or 2:1 **Frequency Limit:** **Lower:** 2 GHz **Upper:** 18 GHz

Figure 9. Horn and Horn w/Polarizer Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
PARABOLIC (Prime) 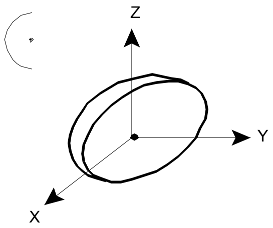 	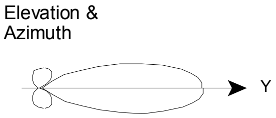 Elevation & Azimuth Y	**Polarization:** Takes polarization of feed **Typical Half-Power Beamwidth:** 1 to 10 deg **Typical Gain:** 20 to 30 dB **Bandwidth:** 33% or 1.4:1 limited mostly by feed **Frequency Limit:** **Lower:** 400 MHz **Upper:** 13+ GHz
PARABOLIC 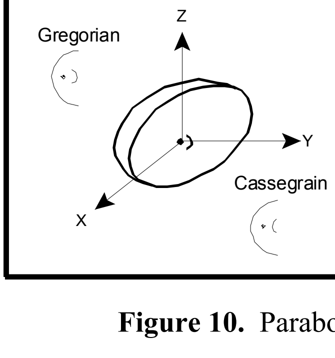 Z Gregorian Y Cassegrain X	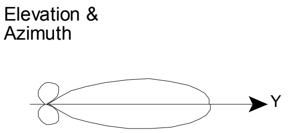 Elevation & Azimuth Y	**Polarization:** Takes polarization of feed **Typical Half-Power Beamwidth:** 1 to 10 deg **Typical Gain:** 20 to 30 dB **Bandwidth:** 33% or 1.4:1 **Frequency Limit:** **Lower:** 400 MHz **Upper:** 13+ GHz

Figure 10. Parabolic (Prime) and Parabolic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
YAGI 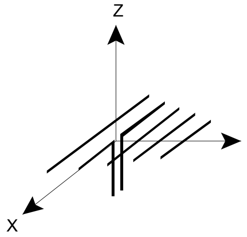	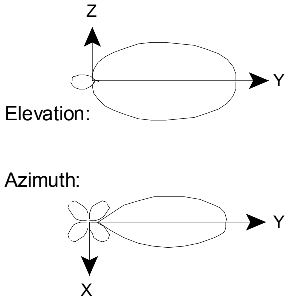	**Polarization:** Linear Horizontal as shown **Typical Half-Power Beamwidth** 50 deg X 50 deg **Typical Gain:** 5 to 15 dB **Bandwidth:** 5% or 1.05:1 **Frequency Limit:** **Lower:** 50 MHz **Upper:** 2 GHz
LOG PERIODIC 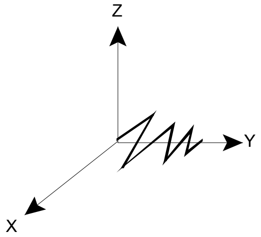	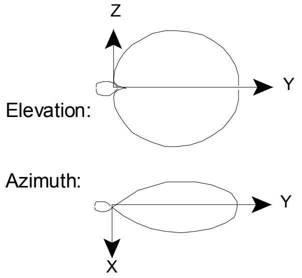	**Polarization:** Linear **Typical Half-Power Beamwidth:** 60 deg x 80 deg **Typical Gain:** 6 to 8 dB **Bandwidth:** 163% or 10:1 **Frequency Limit:** **Lower:** 3 MHz **Upper:** 18 GHz **Remarks:** This array may be formed with many shapes including dipoles or toothed arrays.

Figure 11. Yagi and Log Periodic Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
LINEAR DIPOLE ARRAY (Corporate Feed) 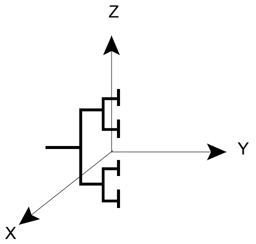	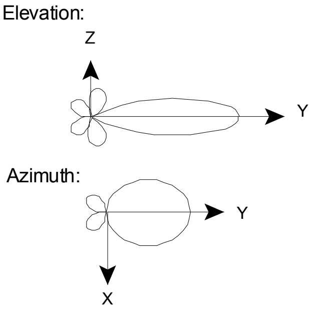	**Polarization:** Element dependent Vertical as shown **Typical Half-Power Beamwidth:** Related to gain **Typical Gain:** Dependent on number of elements **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 10 MHz **Upper:** 10 GHz
APERTURE SYNTHESIS 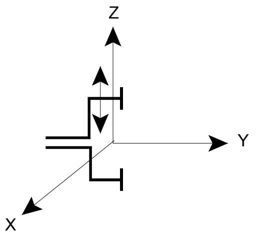	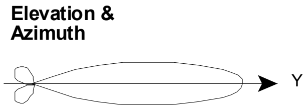	All characteristics dependent on elements **Remarks:** Excellent side-looking, ground mapping where the aircraft is a moving linear element.

Figure 12. Linear Dipole Array and Aperture Synthesis Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CAVITY BACKED CIRCUIT FED SLOT (and Microstrip Patch) 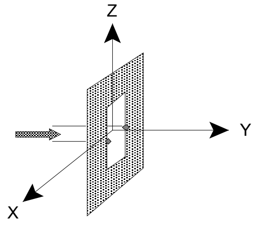	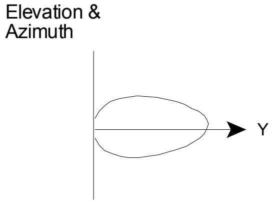	**Polarization:** Linear, vertical as shown **Typical Half-Power Beamwidth**: 80 deg x 80 deg **Typical Gain:** 6 dB **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 50 MHz **Upper:** 18 GHz **Remarks:** The feed line is sometimes separated from the radiator by a dialetric & uses capacititive coupling. Large conformal phased arrays can be made this way.
GUIDE FED SLOT 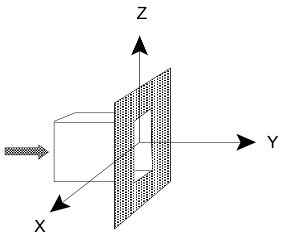	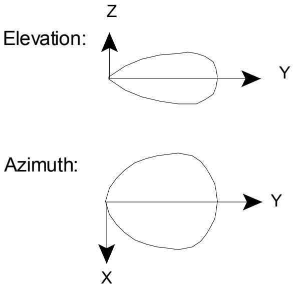	**Polarization:** Linear, **Typical Half-Power Beamwidth** Elevation: 45-50☜ Azimuth: 80☜ **Typical Gain:** 0 dB **Bandwidth:** Narrow **Frequency Limit:** **Lower:** 2 GHz **Upper:** 40 GHz **Remarks:** Open RF Waveguide

Figure 13. Cavity Backed Circuit Fed Slot and Guide Fed Slot Antenna Characteristics.

Antenna Type	Radiation Pattern	Characteristics
CORNER REFLECTOR 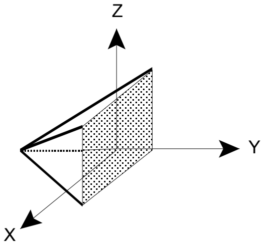	Elevation: (Z-Y) Azimuth: (X-Y) Dependent upon feed emitter	**Polarization**: Feed dependent **Typical Half-Power Beamwidth** 40 deg x variable **Typical Gain:** 10 dB above feed **Bandwidth:** Narrow **Frequency Limit** **Lower:** 1 GHz **Upper:** 40 GHz **Remarks:** Typically fed with a dipole or colinear array.
LUNEBURG LENS 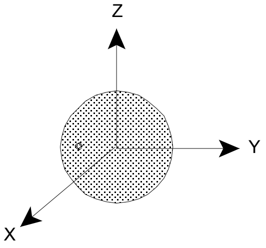	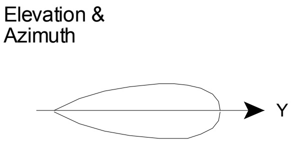	**Polarization:** Feed dependent **Typical Half-Power Beamwidth:** System dependent **Typical Gain:** System dependent **Bandwidth:** Narrow **Frequency Limit** **Lower:** 1 GHz **Upper:** 40 GHz **Remarks:** Variable index dielectric sphere.

Figure 14. Corner Reflector and Luneburg Lens Antenna Characteristics.

This page intentionally left blank.

FREQUENCY / PHASE EFFECTS OF ANTENNAS

The radiation patterns of the antennas presented in the previous section are for antenna geometries most commonly used. The antenna should be viewed as a matching network that takes the power from a transmission line (50 ohm, for example), and matches it to the free space "impedance" of 377 ohms. The most critical parameter is the change of VSWR with frequency. The pattern usually does not vary much from acceptable to the start of unacceptable VSWRs (> 2:1). For a given physical antenna geometric size, the actual radiation pattern varies with frequency.

The antenna pattern depicted in Figure 1 is for the dipole pictured in Section 3-3. The maximum gain is normalized to the outside of the polar plot and the major divisions correspond to 10 dB change. In this example, the dipole length (in wavelengths) is varied, but the same result can be obtained by changing frequency with a fixed dipole length. From the figure, it can be seen that side lobes start to form at 1.25λ and the side lobe actually has more gain than the main beam at 1.5λ. Since the radiation pattern changes with frequency, the gain also changes.

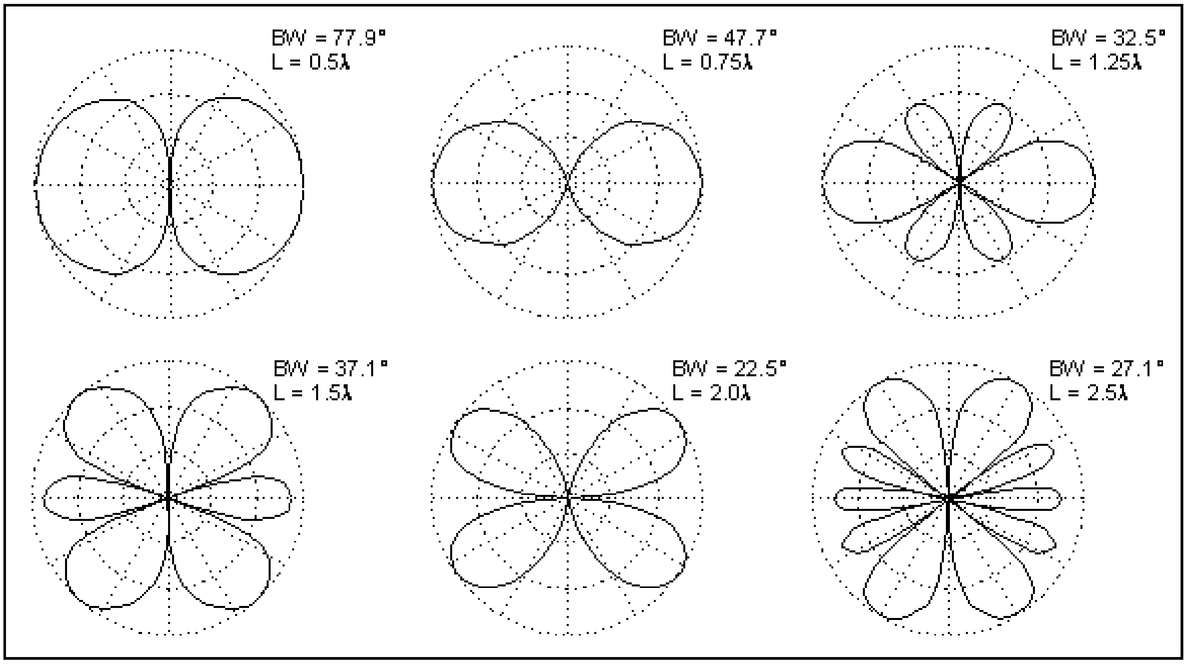

Figure 1. Frequency Effects.

Figure 2 depicts phase/array effects, which are yet another method for obtaining varied radiation patterns. In the figure, parallel dipoles are viewed from the end. It can be seen that varying the phase of the two transmissions can cause the direction of the radiation pattern to change. This is the concept behind phased array antennas. Instead of having a system mechanically sweeping the direction of the antenna through space, the phase of radiating components is varied electronically, producing a moving pattern with no moving parts. It can also be seen that increasing the number of elements further increases the directivity of the array. In an array, the pattern does vary considerably with frequency due to element spacing (measured in wavelengths) and the frequency sensitivity of the phase shifting networks.

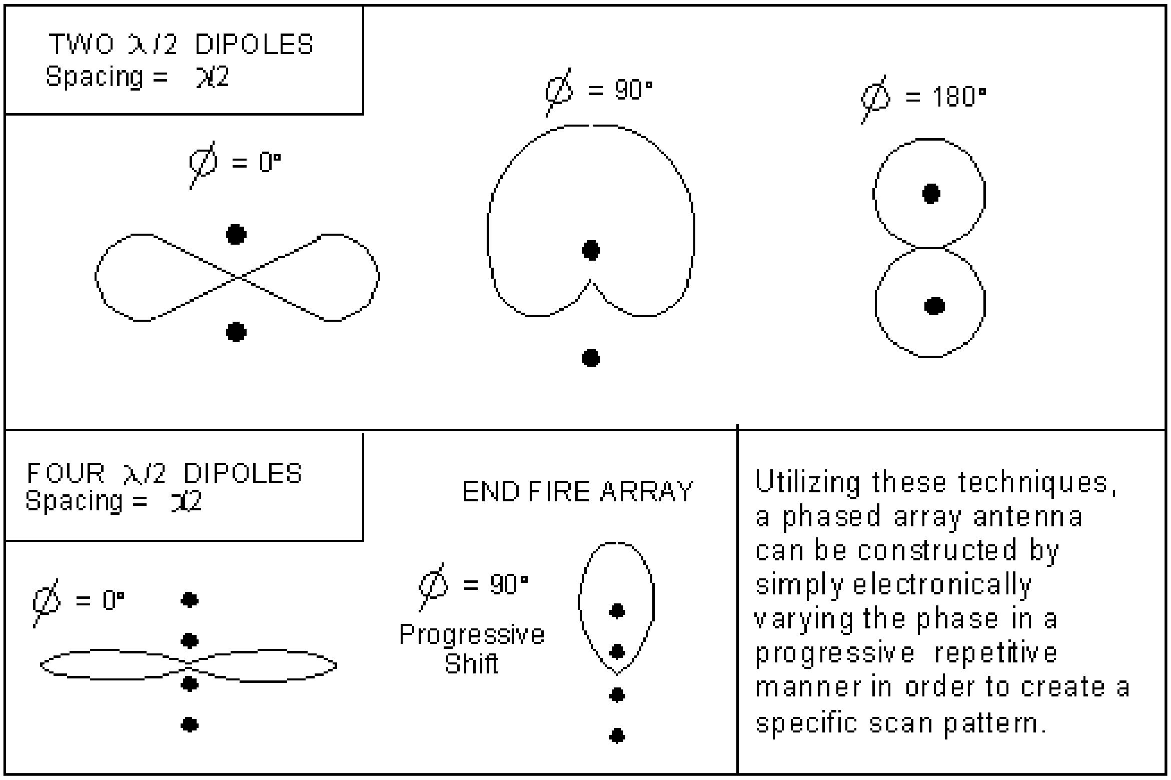

Utilizing these techniques, a phased array antenna can be constructed by simply electronically varying the phase in a progressive repetitive manner in order to create a specific scan pattern.

Figure 2. Phase / Array Effects*.

* Note: Assuming Figure 2 depicts x-y plane antenna pattern cross section, to achieve the indicated array patterns using dipole antennas, the dipole antenna elements must be aligned with the z-axis.

Two antennas that warrant special consideration are the phased array and the Rotman bootlace type lens. Both of these antennas find wide application in EW, RADAR, and Communications. The phased array will be described first.

LINEAR PHASED ARRAY

The linear phased array with equal spaced elements is easiest to analyze and forms the basis for most array designs. Figure 3 schematically illustrates a corporate feed linear array with element spacing d.

It is the simplest and is still widely used. By controlling the phase and amplitude of excitation to each element, as depicted, we can control the direction and shape of the beam radiated by the array. The phase excitation, $\varphi(n)$, controls the beam pointing angle, θ_o, in a phased array. To produce a broadside beam, $\theta_o=0$, requires phase excitation, $\varphi(n)=0$. Other scan angles require an excitation, $\varphi(n) = nkd \sin(\theta_o)$, for the nth element where k is the wave number ($2\pi/\lambda$). In this manner a linear phased array can radiate a beam in any scan direction, θ_o, provided the element pattern has sufficient beamwidth. The amplitude excitation, A_n, can be used to control beam shape and sidelobe levels. Often the amplitude excitation is tapered in a manner similar to that used for aperture antennas to reduce the sidelobe levels. One of the problems that can arise with a phased array is insufficient bandwidth, since the phase shift usually is not obtained through the introduction of additional path length. However, it should be noted that at broadside the corporate feed does have equal path length and would have good bandwidth for this scan angle.

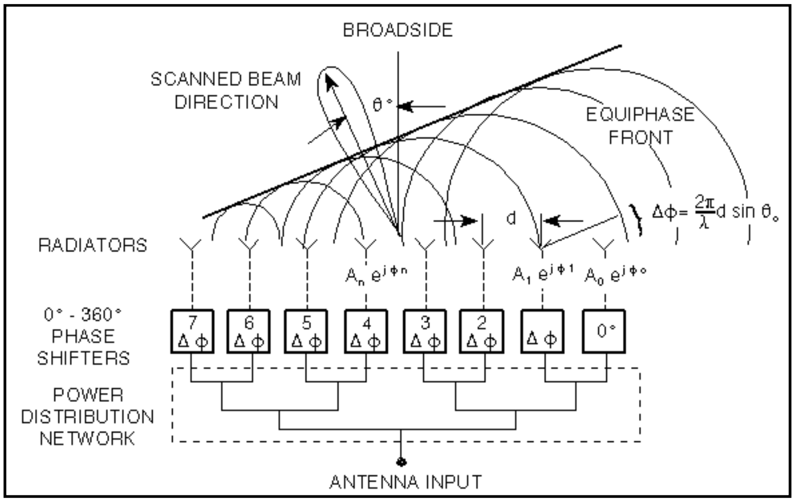

Figure 3. Corporate Fed Phased Array.

The linear array described above would yield a narrow fan beam in the plane normal to the plane containing the array and scan direction, with the narrow beamwidth in the plane of the array. To obtain a pencil beam it would be necessary to array several of these linear arrays in such a manner resulting in a planar array of radiating elements. A problem associated with all electronic scanning is beam distortion with scan angle. Figure 4 illustrates this phenomenon. It results in spread of the beam shape with a concomitant reduction in gain. This effect is known as "scan loss." For an ideal array element, scan loss is equal to the reduction in aperture size in the scan direction which varies as cos θ, where θ is the scan angle measured from the planar array normal.

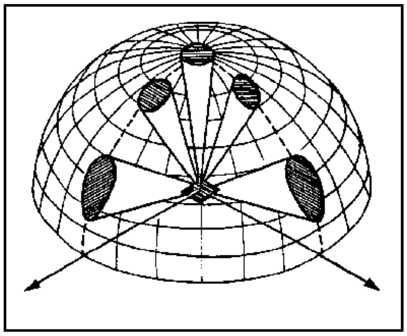

Figure 4. Beam Distortion.

When elements are spaced greater than λ/2 apart, grating lobes are possible when scanning. As the beam is scanned further from broadside, a point is reached at which a second symmetrical main lobe is developed at the negative scan angle from broadside. This condition is not wanted because antenna gain is immediately reduced by 3 dB due to the second lobe. Grating lobes are a significant problem in EW applications because the broad frequency bandwidth requirements mean that at the high end of the frequency band, the elements may be spaced greater than λ/2. Therefore in order to avoid grating lobes over a large frequency bandwidth, element spacing must be no greater than λ/2 at the highest frequency of operation.

There are many other factors to consider with a phased array such as coning, where the beam curves at large scan angles, and mutual coupling between elements that affect match and excitation. Excessive mutual coupling will invariably result in blind scan angles where radiation is greatly attenuated. These issues will not be covered in detail here.

Of interest is the gain of the array which is given by:

$$\text{Array Gain} = G_e(\theta) \bullet \sum_{n=1}^{N} A(n)\, e^{j\phi(n)}\, e^{j n k d \sin\theta}$$ Where each element is as described in Section 3-4.

$G_e(\theta)$ is the element gain which in this case has been taken the same for all elements. Note that if we set $A(n)=1$, and $\varphi(n)=0$, then at broadside where $\sin(\theta) = 0$, the gain would be $(N\ G_e)$. This represents the maximum gain of the array, which typically will not exceed $n\pi$, and is a familiar figure. It should be noted that in practical array design, the element pattern characteristics are greatly influenced by mutual coupling and that the characteristic of elements at the edge of the array can deviate significantly from those near the center.

ROTMAN BOOTLACE LENS

Another method of feeding an array of elements is to use a lens such as the Rotman (rhymes with rotten) Bootlace type shown in Figure 5. The lens consists of a parallel plate region (nowadays microstrip or stripline construction) and cables of specified length connecting the array of elements to the parallel plate region. The geometry of the lens and the cable lengths are designed so that all ray paths traced from a beam port on the right side to its associated wavefront on the left array port side, are equal. This tailoring of the design is accomplished at three focus points (beam ports 1, 4, and 7 in Figure 5). Departure from perfect focus at intermediate beam ports is negligible in most designs.

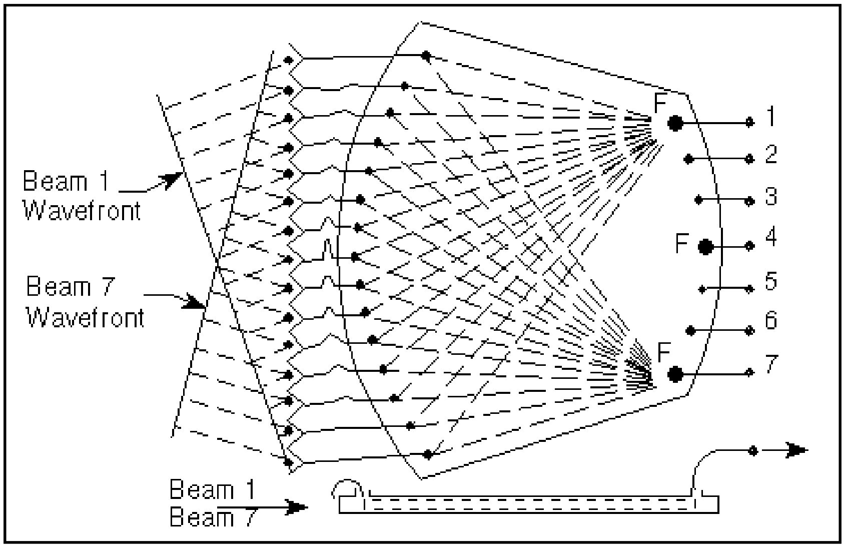

Figure 5. Rotman Bootlace Lens.

The Rotman lens provides both true time delay phase shift and amplitude taper in one lens component. The true time delay is one of the distinct advantages of the lens over the phase shifted array since that makes it independent of frequency. To understand how the taper is obtained requires knowledge of the parallel plate region. For a stripline design the unit would consist of a large flat plate-like center conductor sandwiched between two ground planes, and having a shape much like that of the plan view outline shown in Figure 5 with individual tapered launchers (connectors) attached to each beam port and array port. If the antenna is in the receive mode, the energy intercepted on the array port side can be controlled by the angle subtended by the tapered sections of the connector (launcher) much like a larger antenna would intercept a larger portion of energy from free space.

Unlike the phased array with its fine beam steering, the Rotman lens provides only a distinct set of beams. Fine steering is obtained by combining beams either equally or unequally to form intermediate beams. As can be seen in Figure 6, this results in a broader beam with less gain but lower side lobes than the primary beams.

High transmit power can be obtained using a Rotman lens by placing a low power amplifier between each lens output port and its antenna. In this case a separate Rotman lens would have to be used for receiving.

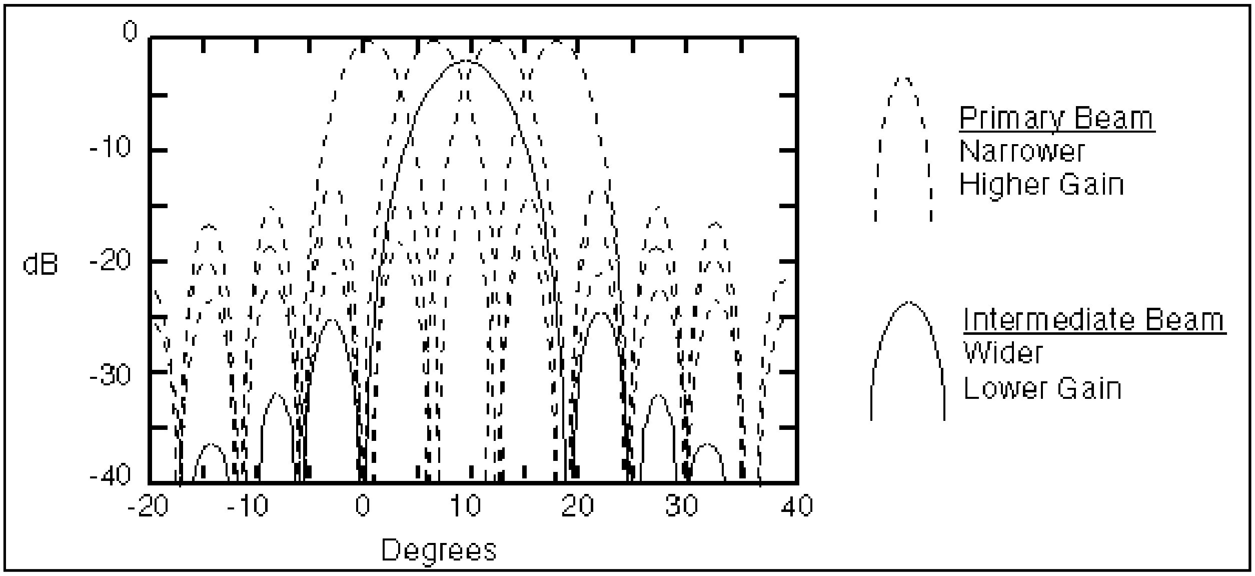

Figure 6. Primary and Intermediate Beam Formation in Lens Arrays.

This page intentionally left blank.

ANTENNA NEAR FIELD

As noted in the sections on RF propagation and the radar equation, electromagnetic radiation expands spherically (Figure 1) and the power density at a long range (R) from the transmitting antenna is:

$$P_D = \frac{P_t G_t}{4\pi R^2} \qquad \textbf{[1]}$$

When the range is large, the spherical surface of uniform power density appears flat to a receiving antenna which is very small compared to the surface of the sphere. This is why the far field wave front is considered planar and the rays approximately parallel. Also, it is apparent that at some shorter range, the spherical surface no longer appears flat, even to a very small receiving antenna.

The planer, parallel ray approximation is valid for distances greater than the distance where the phase error is 1/16 of a wavelength or 22.5 degrees. This distance is given by

$R_{ff} = \frac{2D^2}{\lambda}$ where λ is the wavelength and D is the largest dimension of the transmit antenna. **[2]**

Antenna measurements made at distances greater than R_{ff} generally result in negligible pattern error. Distances less than R_{ff} is termed the near-field.

If the same size antenna is used for multiple frequencies, R_{ff} will increase with increasing frequency. However, if various size antennas are used for different frequencies and each antenna is designed with D as a function of λ ($\lambda/2$ to 100λ), then R_{ff} will vary from $c/2f$ to $20000c/f$. In this case R_{ff} will decrease with increasing frequency. For example: a 10λ antenna at 3 GHZ has a D of 100 cm and corresponding R_{ff} of 20 m, while a 10λ antenna at 30 GHz has a D of 10 cm and corresponding R_{ff} of 2 m.

While the above analogy provides an image of the difference between the near and far fields, the relationship must be defined as a characteristic of the transmitting antenna.

Actual antennas, of course, are not ideal point source radiators but have physical dimensions. If the transmitting antenna placed at the origin of Figure 1 occupies distance D along the Z-axis and is boresighted along the Y-axis ($\varphi = 90$), then the geometry of point P on the sphere is represented in two dimensions by Figure 2. For convenience, the antenna is represented by a series of point sources in an array.

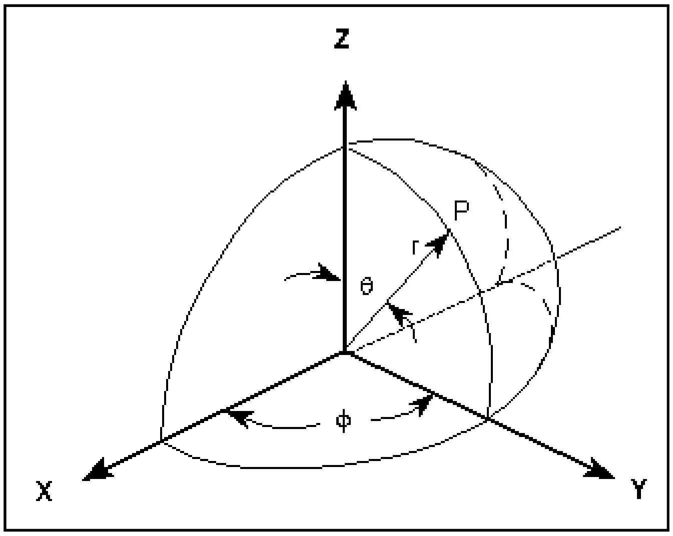

Figure 1. Spherical Radiation to Point "P" from an Ideal Point Source.

When point P is close to the antenna, as in Figure 2, then the difference in distance of the two rays r and R taken respectively from the center of the antenna and the outer edge of the antenna varies as point P changes.

Derivation of equation [2] is given as follows:

From Figure 2, the following applies:

$r^2 = z^2 + y^2$ **[3]**

$z = r \cos \theta$ **[4]**

$y = r \sin \theta$ and **[5]**

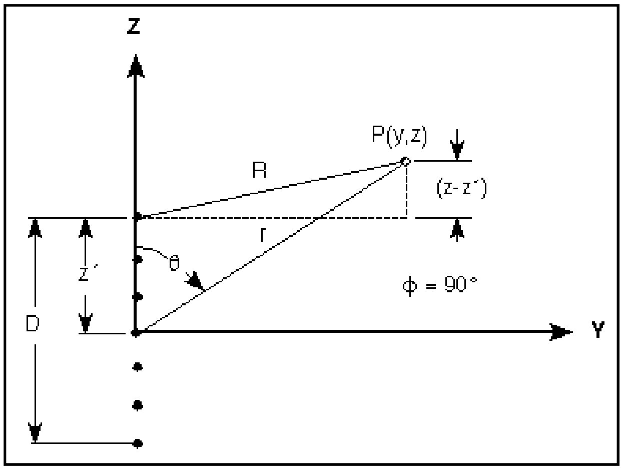

Figure 2. Near Field Geometry of Point "P" for a Non-Ideal Radiator With Dimension D.

$$R = \sqrt{y^2 + (z - z')^2} = \sqrt{y^2 + z^2 - 2z'z + (z')^2} \qquad [6]$$

Substituting [3] and [4] into [6] $R = \sqrt{r^2 + [-2(r \cos \theta) z' + (z')^2]}$ **[7]**
which puts point P into spherical coordinates.

Equation [7] can be expanded by the binomial theorem which for the first three terms, reduces to:

$$R = r - z' \cos \theta + \frac{(z')^2 \sin^2 \theta}{2r} + ... \qquad [8]$$

In the parallel ray approximation for far field calculations (Figure 3) the third term of [8] is neglected.

The distance where the far field begins (R_{ff}) (or where the near field ends) is the value of r when the error in R due to neglecting the third term of equation [8], equals 1/16 of a wavelength.

R_{ff} is usually calculated on boresight, so θ = 90° and the second term of equation [8] equals zero (Cos 90° = 0), therefore from Figure 3, where D is the antenna dimension, R_{ff} is found by equating the third term of [8] to 1/16 wavelength.

$$\frac{(z')^2 \sin^2 \theta}{2 R_{ff}} = \frac{\lambda}{16}$$

$\text{Sin}\, \theta = \text{Sin}\, 90 = 1 \text{ and } z' = D/2$ so: $\dfrac{\left(\frac{D}{2}\right)^2}{2 R_{ff}} = \dfrac{\lambda}{16}$

$$R_{ff} = \frac{16 (D/2)^2}{2\lambda} = \frac{2 D^2}{\lambda} \qquad [9]$$

Equation [9] is the standard calculation of far field given in all references.

Besides [9] some general rules of thumb for far field conditions are:

$$r >> D \quad \text{or} \quad r >> \lambda$$

If the sphere and point P are a very great distance from the antenna, then the rays are very nearly parallel and this difference is small as in Figure 3.

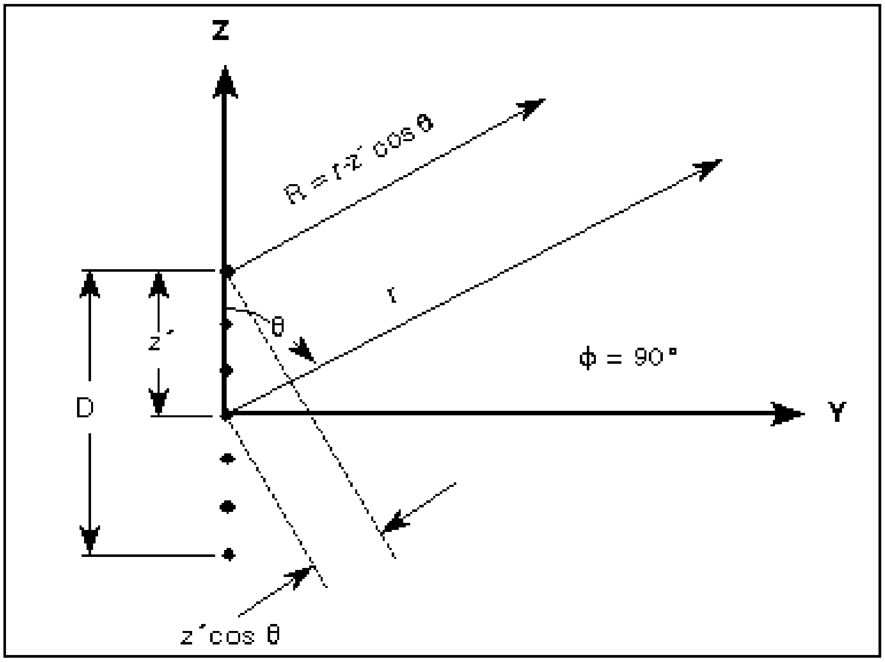

Figure 3. Far Field Parallel Ray Approximation for Calculations.

For reference purposes, a simplified alternative method to derive minimum far field distance approximation for an antenna or an antenna array with aperture size "d" is presented in Figure 4 without the need to resort to the spherical coordinate system and the Binomial Theorem. This approach illustrates a simple application of the Pythagorean Theorem. The P_{FF} symbol represents an arbitrary point in the far field at antenna boresight. Refer to relevant quantities shown in Figure 4.

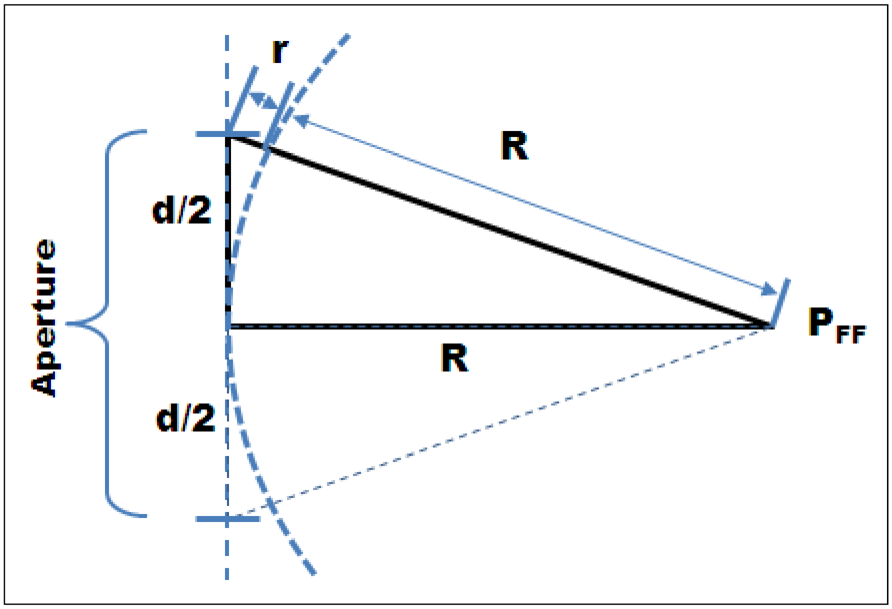

Figure 4. Alternative Geometry for Far Field Estimation.

From Figure 4, the distance “r” is the maximum propagation path difference across a given antenna aperture endpoint locations. Expressing this propagation path difference in terms of a wavelength provides means to derive an equation for generalized far field range relationship. Note a slight change of notation between Figures 3 and 4. A derivation of results already shown in Equations [2] & [9] is given as follows. From Figure 4, the following applies:

$$(R + r)^2 = (d/2)^2 + R^2 \qquad \textbf{[10]}$$

Collecting terms and cancelling,

$$2Rr + r^2 = d^2/4 \qquad \textbf{[11]}$$

Simplifying with a noted valid assumption,

$$2Rr \approx d^2/4,\ \text{ since } r << R \qquad \textbf{[12]}$$

Solving for “r” (see Figure 4)

$$r \approx d^2/(8R) \qquad \textbf{[13]}$$

For far field condition assume $r \leq \lambda/16$. Substituting “r” in terms of lambda,

$$\lambda/16 \geq r \approx d^2/(8R) \ \rightarrow \ P_{FF} \geq 2d^2/\lambda \qquad \textbf{[14]}$$

As shown, Equation [14] represents the same result as Equation [9] derived earlier.

The power density within the near field varies as a function of the type of aperture illumination and is less than would be calculated by equation [1]. Thus, in the antenna near field there is stored energy. (The complex radiation field equations have imaginary terms indicating reactive power.) Figure 5 shows normalized power density for three different illuminations.

Curve A is for reference only and shows how power density would vary if it were calculated using equation [1].

Curve B shows power density variations on axis for an antenna aperture with a cosine amplitude distribution. This is typical of a horn antenna in the H-plane.

Curve C shows power density variations on axis for a uniformly illuminated antenna aperture or for a line source. This is typical of a horn antenna in the E-plane.

Curve D shows power density variations on axis for an antenna aperture with a tapered illumination. Generally the edge illumination is approximately -10 dB from the center illumination and is typical of a parabolic dish antenna.

Point E For radiation safety purposes, a general rule of thumb for tapered illumination is that the maximum safe level of 10 mW/cm^2 (~200 V/m) is reached in the near field if the level at R_{ff} reaches 0.242 mW/cm^2 as can be verified by computing the power density at point E in Figure 5. (10 mW/cm^2 at point E extrapolates to 0.242 mW/cm^2 [16 dB lower] at $R=R_{ff}$, or Y axis value =1). Figure 1 in Section 3-6 depicts more precise values for radiation hazard exposure.

Point F Far Field Point. At distances closer to the source than this point (near field), the power density from any given antenna is less than that predicted using Curve A. At farther distances, (far field) power densities from all types of antennas are the same.

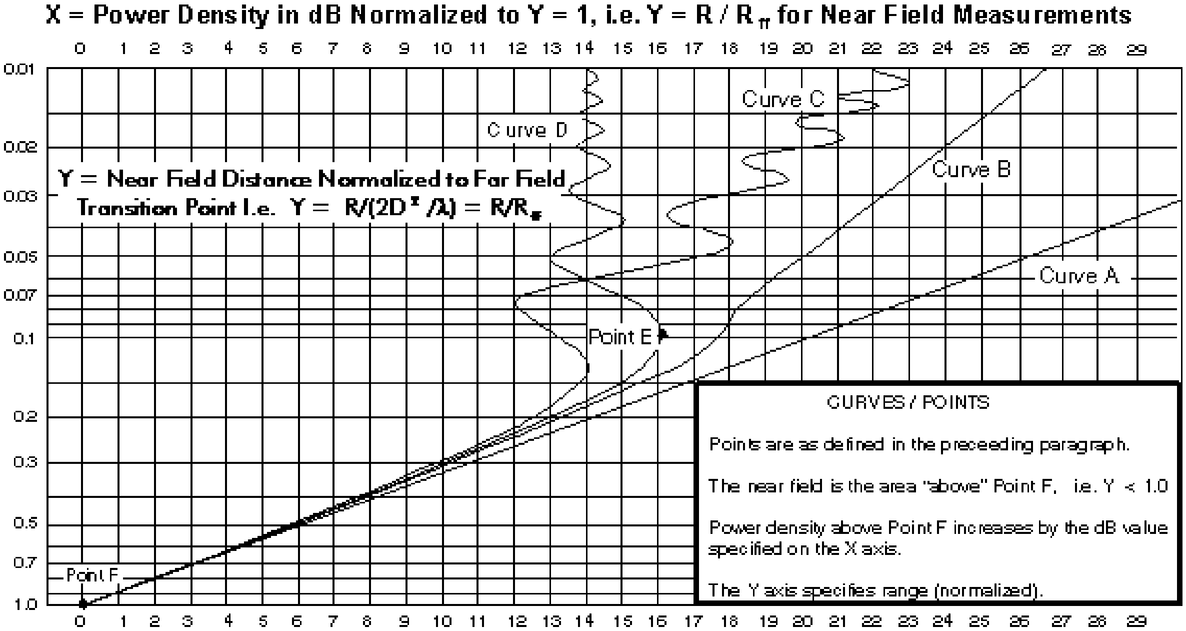

Figure 5. Antenna Near-Field On-Axis Power Density (Normalized) for Various Aperture Illuminations.

FOR FAR FIELD MEASUREMENTS:

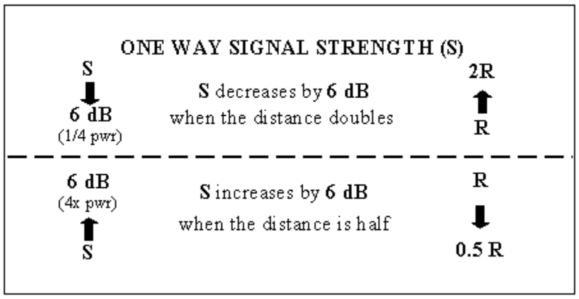
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB
when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB
when the distance is half
R
0.5 R

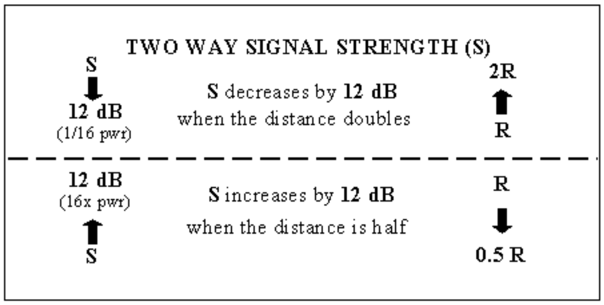
TWO WAY SIGNAL STRENGTH (S)
S
12 dB
(1/16 pwr)
S decreases by 12 dB
when the distance doubles
2R
R
12 dB
(16x pwr)
S
S increases by 12 dB
when the distance is half
R
0.5 R

When free space measurements are performed at a known distance from a source, it is often necessary to know if the measurements are being performed in the far field. As can be seen from Curve A on Figure 5, if the distance is halved (going from 1.0 to 0.5 on the Y axis), the power density will increase by 6 dB (going from 0 to 6 dB on the X axis). Each reduction in range by ½ results in further 6 dB increases. As previously mentioned, Curve A is drawn for reference only in the near field region, since at distances less than R_{ff} the power density increases less than 6 dB when the range is halved. In the far field, all curves converge and Equation [1] applies.

When a measurement is made in free space, a good check to ensure that is was performed in the far field is to repeat the measurement at twice the distance. The power should decrease by exactly 6 dB. A common error is to use 3 dB (the half power point) for comparison. Conversely, the power measurement can be repeated at half the distance, in which case you would look for a 6 dB increase, however the assumed extrapolation conclusion is not as certain, because the first measurement could have been made in the far field, and the second could have been made in the near field.

Care must be exercised in using the **$2d^2/\lambda$** far field measurement criterion. For antennas with moderate sidelobe levels (>-25 dB) pattern errors are negligible and the error in directivity is less than 0.1 dB and this measurement distance suffices in most cases. However low sidelobe antennas require longer measurement distances. For example, maintaining a 1 dB or less sidelobe error for a linear array with a -40 dB sidelobe level requires a measurement distance of $6d^2/\lambda$. (1). This corresponds to criterion for "r" in Equation [14] to be changed from $r \leq \lambda/16$ to $r \leq \lambda/48$ or 7.5 deg.

[1] R. C. Hansen. "Measurement Distance Effects on Low Sidelobe Patterns," *IEEE Transactions on Antennas and Propagation*, Vol. AP-32, No. 6, June 1984.

RADIATION HAZARDS

Radiation Hazard (RADHAZ) describes the hazards of electromagnetic radiation to fuels, electronic hardware, ordnance, and personnel. In the military these hazards are segregated as follows:

1) Hazards of Electromagnetic Radiation to Personnel (HERP)
2) Hazards of Electromagnetic Radiation to Ordnance (HERO)
3) Hazards of Electromagnetic Radiation to Fuel (HERF)

The current industrial specifications for RADHAZ are contained in ANSI/IEEE C95.1-1992 which was used as a reference to create the combined Navy regulation NAVSEA OP3565 / NAVAIR 16-1-529. Volume I contains HERP and HERF limits - its current version is REV 5. Volume II (REV 6) covers HERO. These limits are shown in Figure 1 although all values have been converted to average power density.

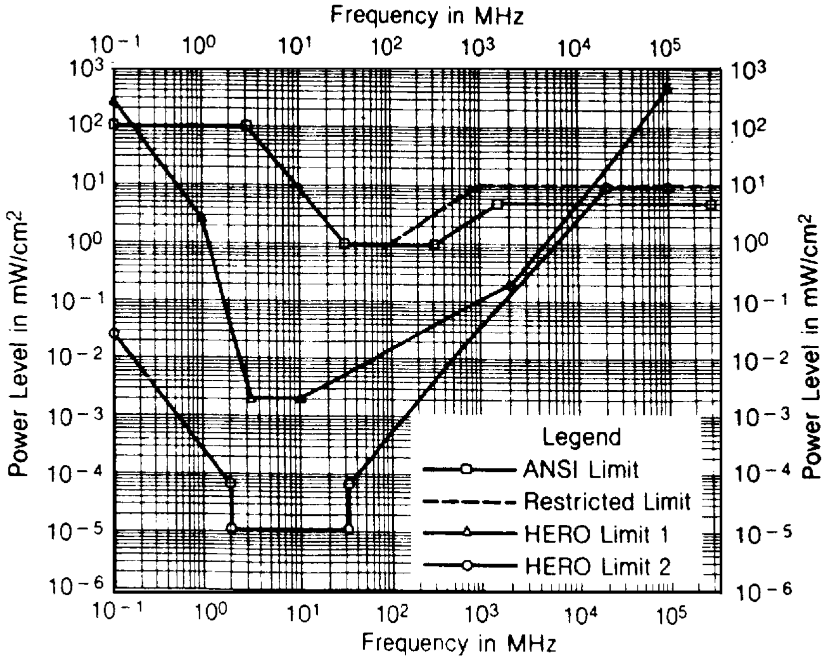

Figure 1. Radiation Hazards to Personnel and Ordnance.

OP 3565 specifies HERO RADHAZ levels at frequencies below 1 GHz in peak value of electric field strength (V/m), while levels above 200 MHz are specified in average power density (mW/cm^2) - note the overlapping frequencies. Since Figure 1 depicts power density as the limits, you must convert the average values to peak field strength for use at lower frequencies. Also many applications of EMC work such as MIL-STD-461 use limits based on the electric (E) field strength in volts/meter. Remember that $P=E^2/R$, and from Section 4-2, we note that $R=377\Omega$ for free space. It can also be shown that the magnetic field strength (H field in Amps/meter) = I/m where I=E/R. Don't forget that RMS = 0.707 Peak. With the units of P_D in mW/cm^2, E in V/m, and H in A/m, then P_D (mW/cm^2) = E^2 / 3770 = 37.7 H^2. It should thus be noted that a 100 times increase in power (mW/cm^2) is only a 10 times increase in V/m.

The potential dangers to ordnance and fuels are obvious because there could be an explosive "chain reaction" by exploding; consequently, these limits are generally lower than personnel limits. There are three **HERO** categories. The HERO limit 2 is for HERO "unsafe" or "unreliable" explosive devices with exposed wires arranged in optimum (most susceptible) receiving orientation. This usually occurs during the assembly/disassembly of ordnance, but also applies to new/untested ordnance until proven "safe" or "susceptible." The HERO limit 1 is for HERO susceptible ordnance fully assembled undergoing normal handling and loading operations. HERO safe ordnance requires no RF radiation precautions. A list of which specific ordnance (by NALC) falls into each category can be found in OP 3565 along with specific frequency restrictions for each piece of ordnance. For example, all missiles of one variety are susceptible

(HERO 1 limits), while another missile has both susceptible and safe variants (with no RADHAZ limits). Other ordnance may be HERO unsafe (HERO 2 limits).

The danger of **HERP** occurs because the body absorbs radiation and significant internal heating may occur without the individual's knowledge because the body does not have internal sensation of heat, and tissue damage may occur before the excess heat can be dissipated. As shown in Figure 1, the current "restricted" limit is for individuals more than 55" tall because they have more body mass. In other words, all people may be exposed to the lower limit, but only persons taller than 55" may be exposed to the higher limit of 10 mW/cm^2.

NAVSEA OP 3565 will be updated in the future to be compatible with DoD INST 6055.11 dated Feb 21, 1995 which supersedes it. The personnel radiation levels in Figures 2 and 3 were taken from the new release of DoD INST 6055.11.

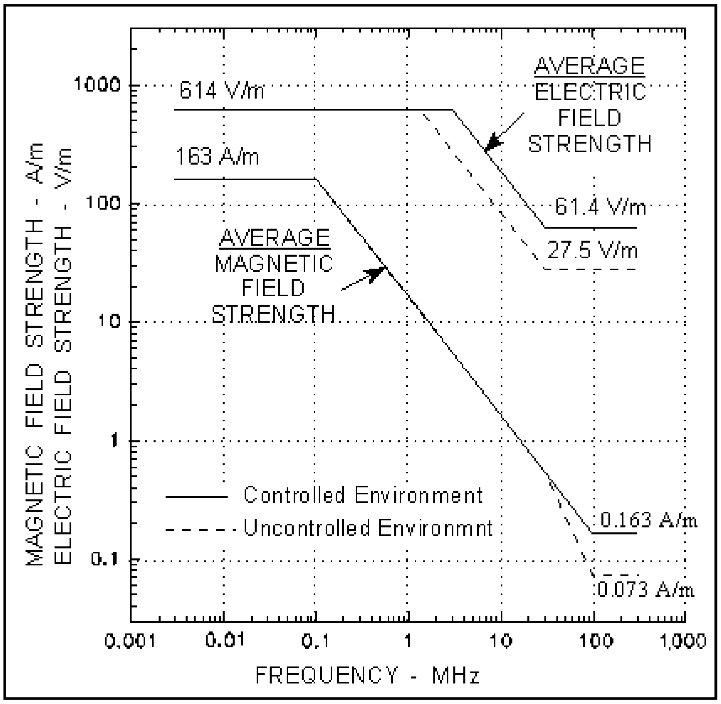

Figure 2. Lower Frequency HERP from DoD INST 6055.11.

Unlike the existing "restricted limit" of NAVSEA OP 3565 discussed above, in the revised DoD instruction for personnel radiation hazards, a different approach to exposure was taken.

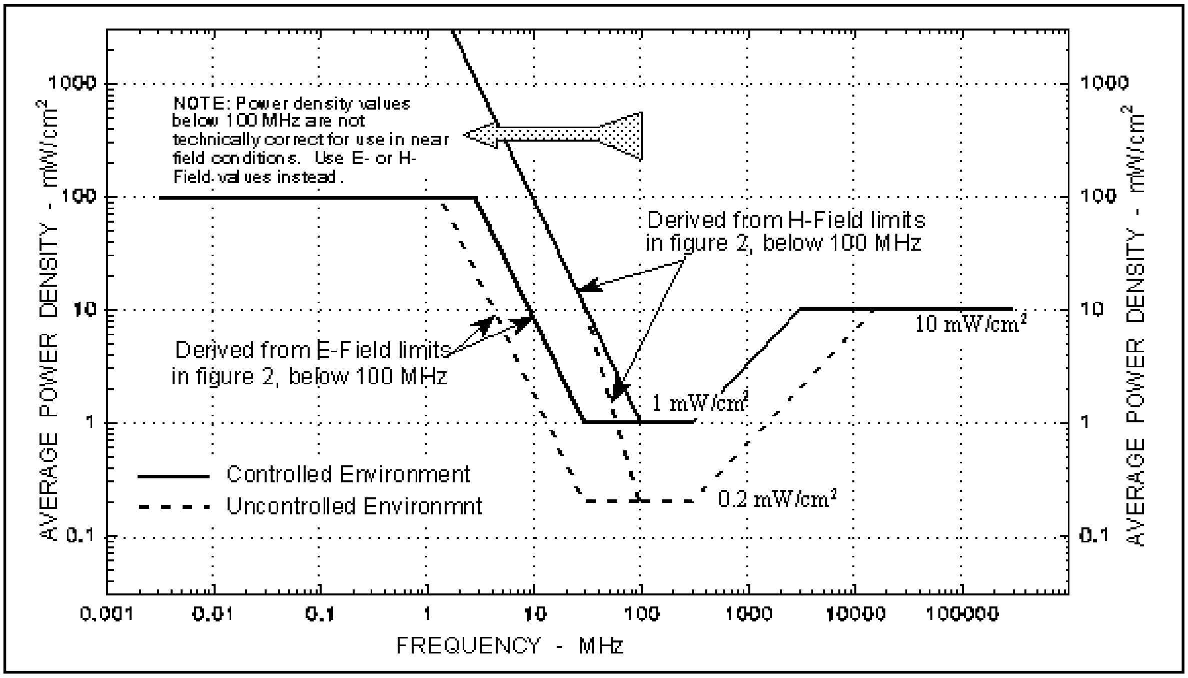

Figure 3. Radiation Hazards to Personnel from DoD INST 6055.11.

Two maximum hazard limits are defined;

1) Controlled Environments - where personnel are aware of the potential danger of RF exposure concurrently with employment, or exposure which may occur due to incidental transient passage through an area, and;
2) Uncontrolled Environments - A lower maximum level where there is no expectation that higher levels should be encountered, such as living quarters.

These Personnel Exposure Limits (PELs) are based on a safety factor of ten times the Specific Absorption Rate (SAR) which might cause bodily harm. The term PEL is equivalent to the terms "Maximum Permissible Exposure (MPE)" and "Radio Frequency Protection Guides (RFPG)" in other publications.

There are several exceptions to the max limits in Figs 2 and 3 (in some cases higher levels are permitted):

- High Power Microwave (HPM) system exposure in a controlled environment, which has a single pulse or multiple pulses lasting less than 10 seconds, has a higher peak E-Field limit of 200 kV/m.
- EMP Simulation Systems in a controlled environment for personnel who are exposed to broad-band (0.1 MHz to 300 GHz) RF are limited to a higher peak E-Field of 100 kV/m.
- The given limits are also increased for pulsed RF fields. In this case the peak power density per pulse for pulse durations < 100 msec and no more than 5 pulses in the period is increased to: PELPulse = PEL x TAVG / 5 x Pulse Width, and the peak E-field is increased to 100 kV/m. If there are more than 5 pulses or they are greater than 100 msec, a time averaged PD should not exceed that shown in Figure 3.
- A rotating or scanning beam likewise reduces the hazard, so although an on-axis hazard might exist, there may be none with a moving beam. The power density may be approximated with:
  PDscan = PDfixed (2 x Beam Width / scan angle)
- Many other special limitations also apply, such as higher limits for partial body exposure, so if in doubt, read the DoD Inst 6055.11 in detail. Field measurements may be measured in accordance with IEEE C95.3-1991.

The PELs listed in Figures 2 and 3 were selected for an average RF exposure time at various frequencies. In a controlled environment, this averaging time was selected as 6 minutes for 0.003 to 15,000 MHz. If the exposure time is less than 6 minutes, then the level may be increased accordingly. Similar time weighted averages apply to uncontrolled environments, but it varies enough with frequency such that DoD INST 6055.11 should be consulted.

NAVSEA OP 3565 contains a list of Navy avionics which transmit RF as well as radars along with their respective hazard patterns. Special training is required for individuals who work in areas which emit RF levels which exceed the uncontrolled levels. Warning signs are also required in areas which exceed either the controlled or uncontrolled limits.

Although E-Field, H-Field, and power density can be mathematically converted in a far-field plane wave environment, the relations provided earlier do not apply in the near field, consequently the E- or H-field strength must be measured independently below 100 MHz. It should be noted that the specifications in NAVSEA OP 3565 for lower frequency HERO limits are listed as peak E-field values, whereas lower RF limits in DoD INST 6055.11 on HERP are in average (RMS) E-field values. Upper frequency restrictions are based on average (RMS) values of power density in both regulations except for certain circumstances.

HERF precautions are of more general concern to fuel truck operators. However, some general guidelines include:

- Do not energize a transmitter (radar/comm) on an aircraft or motor vehicle being fueled or on an adjacent aircraft or vehicle.
- Do not make or break any electrical, ground wire, or tie down connector while fueling.
- Radars able to illuminate fueling areas with a peak power density of 5 W/cm2 should be shut off.
- For shore stations, antennas radiating 250 watts or less should be installed at least 50 ft from fueling areas (at sea 500 watts is the relaxed requirement).
- For antennas which radiate more than 250 watts, the power density at 50 ft from the fueling operation should not be greater than the equivalent power density of a 250 watt transmitter at 50 ft.

ACTIVE ELECTRONICALLY SCANNED ARRAYS (AESA)

Scanning phased arrays employing electronically controlled phase shifters (e.g., PiN diode, ferrite) have been used in high power radar applications since the early 1960s. These are planar arrays and generally use a corporate type feed structure to array rows of elements in one dimension of the array and use an analogous feed to array the rows in the orthogonal dimension of the array, resulting in a single feed point. A high power RF transmitter, usually employing some sort of liquid cooling, is used to excite the array. This architecture is cumbersome and difficult to package and the high power transmitter is a single point of failure for the system.

The maturation of solid state transmit/receive modules (T/R) using Gallium Arsenide (GaAs) technology or more recently Gallium Nitride (GaN) technology has made AESAs a practical reality. The T/R modules consist of a low noise amplifier (LNA) for the receive function and a solid state high power amplifier (SSPA) for realization of the transmit function. The T/R module may also contain the necessary phase control elements for beam scanning. A generic 8-element AESA for one dimension is shown in Figure 1.

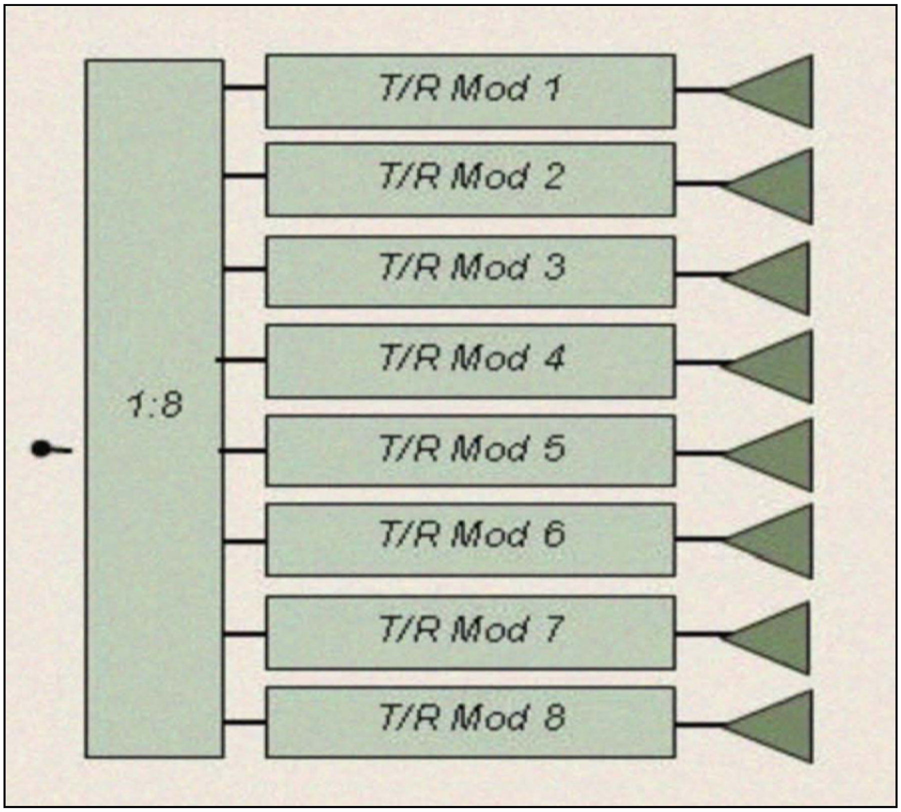

Figure 1. Eight Element AESA.

Electronically scanned phased array designs with decade bandwidths have been reported in the open literature for many years. However those reported are only lab versions and none have been implemented in a fielded system. In addition, the high average power requirement, particularly for EW systems, calls for the use of GaN SSPA technology and this, in conjunction with the dense array packaging requirement, poses severe heat dissipation/cooling issues.

AESAs for EW applications impose design considerations vastly different from an AESA for radar. In contradistinction with high **peak** power radar applications (e.g., AN/APG79), the high **average** power large duty factors (sometimes CW) requirements for EW impose severe design requirements and constraints for the wide bandwidth AESA. Incorporation of an AESA in airborne environment, with its concomitant restrictive volume, further exacerbates design constraints.

The heat dissipation problem is of major concern in modern high power AESAs. Due to the behavior of microwave transistor amplifiers, the power added efficiency (PAE) of a TR module transmitter is typically a relatively small fraction of the total prime power consumption. As a result, an AESA will dissipate a lot of heat, which must be extracted. The reliability of GaAs and GaN MMIC chips improves if the RF amplifier system operates at reduced temperatures. Traditional air cooling used in most established avionic hardware is ill suited to the high packaging density of an AESA. As a result modern AESAs are predominantly liquid cooled. A typical liquid cooling system will use pumps to drive the coolant through channels in the cooling plenum of the array, and then route it to a heat exchanger. In comparison, with a conventional air cooled fighter radar, the AESA will be more reliable but will require more electrical power and more cooling infrastructure, and typically can produce much higher average transmit power.

ADVANTAGES OF AESA RADIATORS

AESAs add many capabilities of their own to those of the Passive Electronically Steered Array (PESAs). Among these are: the ability to form multiple beams, to scan without mechanical steering, to use each transmit/receive module for different roles concurrently, like radar detection, and, more importantly, their multiple wave and scanning frequencies create multiple difficulties for traditional, correlation-type radar detectors.

LOW PROBABILITY OF INTERCEPT

Radar systems work by sending out a signal and then listening for its echo off distant objects. Each of these paths, to and from the target, is subject to the inverse square law of propagation. That means that a radar's received energy drops with the fourth power of distance, which is why radar systems require high powers, often in the megawatt range, to be effective at long range.

The radar signal being sent out is a simple radio signal, and can be received with a simple radio receiver. It is common to use such a receiver in the targets, normally aircraft, to detect radar broadcasts. Unlike the radar unit, which must send the pulse out and then receive its reflection, the target's receiver does not need the reflection and thus the signal drops off only as the square of distance. This means that the receiver is always at an advantage over the radar in terms of range. It will always be able to detect the signal long before the radar can see the target's echo. Since the position of the radar is extremely useful information in an attack on that platform, this means that radars generally must be turned off for lengthy periods if they are subject to attack; this is common on ships, for instance.

Turning that received signal into a useful display is the purpose of the RWR. Unlike the radar, which knows which direction it is sending its signal, the receiver simply gets a pulse of energy and has to interpret it. Since the radio spectrum is filled with noise, the receiver's signal is integrated over a short period of time, making periodic sources like a radar add up and stand out over the random background. Typically RWRs store the detected pulses for a short period of time, and compare their broadcast frequency and pulse repetition frequency against a database of known radars. The rough direction can be calculated using a rotating antenna, or similar passive array, and combined with symbology indicating the likely purpose of the radar - airborne early warning, surface to air missile, etc.

This technique is much less useful against AESA radars. Since the AESA can change its frequency with every pulse, and generally does so using a pseudo-random sequence, integrating over time does not help pull the signal out of the background noise. Nor does the AESA have any sort of fixed pulse repetition frequency, which can also be varied and thus hide any periodic brightening across the entire spectrum. Traditional RWRs suffered significantly decreased effectiveness against AESA radars.

HIGH JAMMING RESISTANCE

Jamming is likewise much more difficult against an AESA. Traditionally, jammers have operated by determining the operating frequency of the radar and then broadcasting a signal on it to confuse the receiver as to which is the “real” pulse and which is the jammer’s. This technique works as long as the radar system cannot easily change its operating frequency. When the transmitters were based on klystron tubes this was generally true, and radars, especially airborne ones, had only a few frequencies to choose among. A jammer could listen to those possible frequencies and select the one to be used to jam.

Since an AESA could change its operating frequency with every pulse, and spread the frequencies across a wide band even in a single pulse, jammers are much less effective. Although it is possible to send out broadband white noise against all the possible frequencies, this means the amount of energy being sent at any one frequency is much lower, reducing its effectiveness. In fact, AESAs can then be switched to a receive-only mode, and use these powerful jamming signals instead to track its source, something that required a separate receiver in older platforms.

AESA radars can be much more difficult to detect, and so much more useful in receiving signals from the targets, that they can broadcast continually and still have a very low chance of being detected. This allows such radar systems to generate far more data than traditional radar systems, which can only receive data periodically, greatly improving overall system effectiveness.

OTHER ADVANTAGES

Since each element in an AESA is a powerful radio receiver, active arrays have many roles besides traditional radar. One use is to dedicate several of the elements to reception of common radar signals, eliminating the need for a separate radar warning receiver. The same basic concept can be used to provide traditional radio support, and with some elements also broadcasting, form a very high bandwidth data link.

AESAs are also much more reliable than either a PESA or older designs. Since each module operates independently of the others, single failures have little effect on the operation of the system as a whole. Additionally, the modules individually operate at lower powers and voltages so the need for a large high-voltage power supply is eliminated.

Replacing a mechanically scanned array with a fixed AESA mount can help reduce an aircraft’s overall RCS, but some designs (such as the Eurofighter Typhoon) forgo this advantage in order to combine mechanical scanning with electronic scanning and provide a wider angle of total coverage.

This page intentionally left blank.

FRACTAL ANTENNAS

The term fractal, which means broken or irregular fragments, was originally coined by B.B. Mandelbrot **[1]** to describe a family of complex shapes that possess self-similarity in their geometrical structure. In-depth studies of the patterns of nature provided inspiration for the development of fractal geometry where it has been used to model numerous natural objects. Recent advances in the antenna art have led to the application of fractal geometry instead of convention Euclidean geometric concepts to the design of wide bandwidth, low profile, compact antennas.

A fractal antenna is an antenna that uses a fractal, self-similar design to maximize the length, or increase the perimeter (on inside sections or the outer structure), of material that can receive or transmit electromagnetic radiation within a given total surface area or volume. The earliest published reference to use of the term fractal radiators and fractal antennas to refer to fractal shaped antenna elements appeared May 1994 **[2]**.

Such fractal antennas are also referred to as multilevel and space filling curves, but the key aspect lies in their repetition of a motif over two or more scale sizes, or "iterations." For this reason, fractal antennas are very compact, multiband or wideband, and have useful applications in many commercial and military systems.

A good example of a fractal antenna as a space-filling curve is in the form of a Minkowski Island (Figure 1). Here, each line of copper is just a small fraction of a wavelength **[3]**.

A fractal antenna's response differs markedly from traditional antenna designs, in that it is capable of operating with good-to-excellent performance at many different frequencies simultaneously. Normally standard antennas have to be "cut" for the frequency for which they are to be used — and thus the standard antennas only work well at that frequency. This makes the fractal antenna an excellent design for wideband and multiband applications.

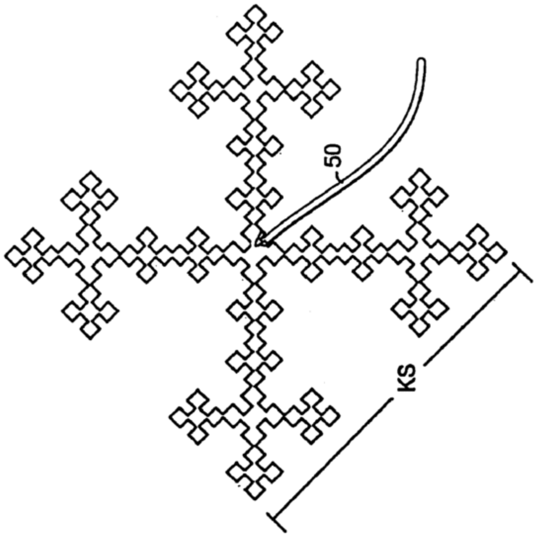

Figure 1. An Example of a Fractal Antenna: a Space-Filling Curve Called a Minkowski Island.

Antenna elements (as opposed to antenna arrays) made from self-similar shapes were first created by Nathan Cohen, then a professor at Boston University, starting in 1995 **[4].** Cohen's efforts with a variety of fractal antenna designs were first published in 1995 (thus the first scientific publication on fractal antennas), and a number of patents have been issued from the 1995 filing priority of invention. Most allusions to fractal antennas make reference to these "fractal element antennas."

Many fractal element antennas use the fractal structure as a virtual combination of capacitors and inductors. This makes the antenna have so that it has many different resonances that can be selected and adjusted by choosing the proper fractal design. Electrical resonances may not be directly related to a particular scale size of the fractal antenna structure. The physical size of the antenna is unrelated to its resonant or broadband performance. The general rule of antenna length being near target frequency wavelength does not apply itself in the same way with fractal antennas.

This complexity arises because the current on the structure has a complex arrangement caused by the inductance and self capacitance. In general, although their effective electrical length is longer, the fractal element antennas are physically smaller.

Fractal element antennas are shrunken compared to conventional designs and do not need additional components. In general, the fractal dimension of a fractal antenna is a poor predictor of its performance and application. Not all fractal antennas work well for a given application or set of applications. Computer search methods, optimization algorithms, and antenna simulations are commonly used to identify which fractal antenna designs best meet the need of the application.

Although the first validation of the technology was published as early as 1995 **[4]** recent independent studies show advantages of the fractal element technology in real-life applications, such as radio frequency identification and cell phones.

A different and also useful attribute of some fractal element antennas is their self-scaling aspect. In 1999, it was discovered that self-similarity was one of the underlying requirements to make antennas "invariant" (same radiation properties) at a number or range of frequencies. Previously, under Rumsey's Frequency Independent Antenna Principle, it was believed that antennas had to be defined by angles for this to be true; the 1999 analysis **[5]**, based on Maxwell's equations, showed this to be a subset of the more general set of self-similar conditions. Hence fractal antennas offer a closed-form and unique insight into a key aspect of electromagnetic phenomena to wit the invariance property of Maxwell's equations.

Antenna tuning units are typically not required on fractal antennas due to their wide bandwidth and complex resonance. However, if a transmitting antenna has deep nulls in its response or has electromagnetic structural issues that require equalization then an antenna tuning unit should be used.

In addition to their use as antennas, fractals have also found application in other antenna system components including loads, counterpoises, and ground planes. Confusion by those who claim "grain of rice"-sized fractal antennas arises, because such fractal structures serve the purpose of loads and counterpoises, rather than bona fide antennas.

Fractal inductors and fractal tuned circuits (fractal resonators) were also discovered and invented simultaneously with fractal element antennas. An emerging example of such is in metamaterials. A recent report demonstrates using close-packed fractal resonators to make the first wideband metamaterial "invisibility cloak" at microwave frequencies **[6]**. Fractal filters (a type of tuned circuit) are another example of fractal geometry in microwave componentry.

As fractals can be used as counterpoises, loads, ground planes, and filters, all parts that can be integrated with antennas, they are considered parts of some antenna systems and thus are discussed in the context of fractal antennas.

REFERENCES

[1] B. B. Mandelbrot. *The Fractal Geometry of Nature.* New York, W. H. Freeman, 1983.

[2] D. H. Wemer. “Fractal Radiators,” *Proceedings of the 4th Annual 1994 IEEE Mohawk Valley~Section Dual-Use Technologies* and *Applications Conference,* Volume I, SUNY Institute of Technology at UticalRomc, New York, 23-26 May 1994, pp. 478-482.

[3] J. P. Gianvittorio and Y. Rahmat-Samii. “Fractal Antennas: A Novel Antenna Miniaturization Technique, and Applications,” *Antennas and Propagation Magazine*, *IEEE*, Vol. 44, No. 1, February 2002, pp. 20-36.

[4] N. Cohen. “Fractal Antennas,” *Communications Quarterly*, Vol. 9, Summer 1995.

[5] R. Hohlfeld and N. Cohen. “Self-Similarity and the Geometric Requirements for Frequency Independence in Antenna,” *Fractals,* Vol. 7, No. 1, 1999, pp. 79–84. doi:10.1142/S0218348X99000098.

[6] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith. “Metamaterial Electromagnetic Cloak at Microwave Frequencies,” *Science,* Vol. 314, No. 5801, 10 November 2006, pp. 977-980. Published online 19 October 2006 [DOI:10.1126/science.1133628].

This page intentionally left blank.

RADAR EQUATIONS

Field Intensity and Power Density 4-1
Power Density 4-2
One-Way Radar Equation / RF Propagation 4-3
Two-Way Radar Equation (Monostatic) 4-4
Alternate Two-Way Radar Equation 4-5
Two-Way Radar Equation (Bistatic) 4-6
Jamming to Signal (J/S) Ratio - Constant Power [Saturated] Jamming 4-7
Burn-Through / Crossover Range 4-8
Support Jamming 4-9
Jamming to Signal (J/S) Ratio - Constant Gain [Linear] Jamming 4-10
Radar Cross Section (RCS) 4-11
Emission Control (EMCON) 4-12
EW Jamming Techniques 4-13

This page intentionally left blank.

FIELD INTENSITY and POWER DENSITY

Sometimes it is necessary to know the actual field intensity or power density at a given distance from a transmitter instead of the signal strength received by an antenna. Field intensity or power density calculations are necessary when estimating electromagnetic interference (EMI) effects, when determining potential radiation hazards (personnel safety), or in determining or verifying specifications.

Field intensity (field strength) is a general term that usually means the magnitude of the electric field vector, commonly expressed in volts per meter. At frequencies above 100 MHZ, and particularly above one GHz, power density (P_D) terminology is more often used than field strength.
Power density and field intensity are related by equation [1]:

$$P_D = \frac{E^2}{Z_0} = \frac{E^2}{120\pi} = \frac{E^2}{377} \qquad \textbf{[1]}$$

where P_D is in W/m^2, E is the RMS value of the field in volts/meter and 377 ohms is the characteristic impedance of free space. When the units of P_D are in mW/cm^2, then P_D (mW/cm^2) = E^2/3,770.

Conversions between field strength and power density when the impedance is 377 ohms, can be obtained from Table 1. It should be noted that to convert dBm/m^2 to dBμV/m add 115.76 dB. Sample calculations for both field intensity and power density in the far field of a transmitting antenna are in Section 4-2 and Section 4-8. Refer to chapter 3 on antennas for the definitions of near field and far field.

Note that the "/" term before m, m^2, and cm^2 in Table 1 mean "per," i.e., dBm per m^2, not to be confused with the division sign which is valid for the Table 1 equation $P=E^2/Z_o$. Remember that in order to obtain dBm from dBm/m^2 given a certain area, you must add the logarithm of the area, not multiply. The values in the table are rounded to the nearest dBW, dBm, etc. per m^2 so the results are less precise than a typical handheld calculator and may be up to ½ dB off.

VOLTAGE MEASUREMENTS

Coaxial cabling typically has input impedances of 50, 75, and 93Ω, (±2) with 50Ω being the most common. Other types of cabling include the following: TV cable is 75Ω (coaxial) or 300Ω (twin-lead), audio public address (PA) is 600Ω, audio speakers are 3.2 (4), 8, or 16Ω.

In the 50Ω case, power and voltage are related by:

$$P = \frac{E^2}{Z_0} = \frac{E^2}{50} = 50\, I^2 \qquad \textbf{[2]}$$

Conversions between measured power, voltage, and current where the typical impedance is 50 ohms can be obtained from Table 2. The dBμA current values are given because frequently a current probe is used during laboratory tests to determine the powerline input current to the system.

MATCHING CABLING IMPEDANCE

In performing measurements, we must take into account an impedance mismatch between measurement devices (typically 50 ohms) and free space (377 ohms).

Table 1. Conversion Table - Field Intensity and Power Density.
$P_D = E^2/Z_0$ (Related by free space impedance = 377 ohms)

E (Volts/m)	20 log 10^6 (E) (dBµV/m)	P$_D$ (watts/m^2)	10 Log P$_D$ (dBW/m^2)	Watts/cm^2	dBW/cm^2	mW/cm^2	dBm/cm^2	dBm/m^2
7,000	197	130,000	+51	13	+11	13,000	+41	+81
5,000	194	66,300	+48	6.6	+8	6,630	+38	+78
3,000	190	23,900	+44	2.4	+4	2,390	+34	+74
4,000	186	10,600	+40	1.1	0	1,060	+30	+70
1,000	180	2,650	+34	.27	-6	265	+24	+64
700	177	1,300	+31	.13	-9	130	+21	+61
500	174	663	+28	.066	-12	66	+18	+58
300	170	239	+24	.024	-16	24	+14	+54
200	166	106	+20	.011	-20	11	+10	+50
100	160	27	+14	.0027	-26	2.7	+4	+44
70	157	13	+11	1.3x10^{-3}	-29	1.3	+1	+41
50	154	6.6	+8	6.6x10^{-4}	-32	.66	-2	+38
30	150	2.4	+4	2.4x10^{-4}	-36	.24	-6	+34
20	146	1.1	+0	1.1x10^{-4}	-40	.11	-10	+30
10	140	.27	-6	2.7x10^{-5}	-46	.027	-16	+24
7	137	.13	-9	1.3x10^{-5}	-49	.013	-19	+21
5	134	.066	-12	6.6x10^{-6}	-52	66x10^{-4}	-22	+18
3	130	.024	-16	2.4x10^{-6}	-56	24x10^{-4}	-26	+14
2	126	.011	-20	1.1x10^{-6}	-60	11x10^{-4}	-30	+10
1	120	.0027	-26	2.7x10^{-7}	-66	2.7x10^{-4}	-36	+4
0.7	117	1.3x10^{-3}	-29	1.3x10^{-7}	-69	1.3x10^{-4}	-39	+1
0.5	114	6.6x10^{-4}	-32	6.6x10^{-8}	-72	66x10^{-4}	-42	-2
0.3	110	2.4x10^{-4}	-36	2.4x10^{-8}	-76	24x10^{-4}	-46	-6
0.2	106	1.1x10^{-4}	-40	1.1x10^{-8}	-80	11x10^{-4}	-50	-10
0.1	100	2.7x10^{-5}	-46	2.7x10^{-9}	-86	2.7x10^{-6}	-56	-16
70x10^{-3}	97	1.3x10^{-5}	-49	1.3x10^{-9}	-89	1.3x10^{-6}	-59	-19
50x10^{-3}	94	6.6x10^{-6}	-52	6.6x10^{-10}	-92	66x10^{-8}	-62	-22
30x10^{-3}	90	2.4x10^{-6}	-56	2.4x10^{-10}	-96	24x10^{-8}	-66	-26
20x10^{-3}	86	1.1x10^{-6}	-60	1.1x10^{-10}	-100	11x10^{-8}	-70	-30
10x10^{-3}	80	2.7x10^{-7}	-66	2.7x10^{-11}	-106	2.7x10^{-8}	-76	-36
7x10^{-3}	77	1.3x10^{-7}	-69	1.3x10^{-11}	-109	1.3x10^{-8}	-79	-39
5x10^{-3}	74	6.6x10^{-8}	-72	6.6x10^{-12}	-112	66x10^{-10}	-82	-42
3x10^{-3}	70	2.4x10^{-8}	-76	2.4x10^{-12}	-116	24x10^{-10}	-86	-46
2x10^{-3}	66	1.1x10^{-8}	-80	1.1x10^{-12}	-120	11x10^{-10}	-90	-50
1x10^{-3}	60	2.7x10^{-9}	-86	2.7x10^{-13}	-126	2.7x10^{-10}	-96	-56
7x10^{-4}	57	1.3x10^{-9}	-89	1.3x10^{-13}	-129	1.3x10^{-10}	-99	-59
5x10^{-4}	54	6.6x10^{-10}	-92	6.6x10^{-14}	-132	66x10^{-12}	-102	-62
3x10^{-4}	50	2.4x10^{-10}	-96	2.4x10^{-14}	-136	24x10^{-12}	-106	-66
2x10^{-4}	46	1.1x10^{-10}	-100	1.1x10^{-14}	-140	11x10^{-12}	-110	-70
1x10^{-4}	40	2.7x10^{-11}	-106	2.7x10^{-15}	-146	2.7x10^{-12}	-116	-76
7x10^{-5}	37	1.3x10^{-11}	-109	1.3x10^{-15}	-149	1.3x10^{-12}	-119	-79
5x10^{-5}	34	6.6x10^{-12}	-112	6.6x10^{-16}	-152	66x10^{-14}	-122	-82
3x10^{-5}	30	2.4x10^{-12}	-116	2.4x10^{-16}	-156	24x10^{-14}	-126	-86
2x10^{-5}	26	1.1x10^{-12}	-120	1.1x10^{-16}	-160	11x10^{-14}	-130	-90
1x10^{-5}	20	2.7x10^{-13}	-126	2.7x10^{-17}	-166	2.7x10^{-14}	-136	-96
7x10^{-6}	17	1.3x10^{-13}	-129	1.3x10^{-17}	-169	1.3x10^{-14}	-139	-99
5x10^{-6}	14	6.6x10^{-14}	-132	6.6x10^{-18}	-172	66x10^{-16}	-142	-102
3x10^{-6}	10	2.4x10^{-14}	-136	2.4x10^{-18}	-176	24x10^{-16}	-146	-106
2x10^{-6}	6	1.1x10^{-14}	-140	1.1x10^{-18}	-180	11x10^{-16}	-150	-110
1x10^{-6}	0	2.7x10^{-15}	-146	2.7x10^{-19}	-186	2.7x10^{-16}	-156	-116

NOTE: Numbers in table rounded off.

FIELD STRENGTH APPROACH

To account for the impedance difference, the antenna factor (AF) is defined as: AF=E/V, where E is field intensity which can be expressed in terms taking 377 ohms into account and V is measured voltage which can be expressed in terms taking 50 ohms into account. Details are provided in Section 4-12.

POWER DENSITY APPROACH

To account for the impedance difference, the antenna's effective capture area term, A_e relates free space power density P_D with received power, P_r, i.e. $P_r = P_D A_e$. A_e is a function of frequency and antenna gain and is related to AF as shown in Section 4-12.

SAMPLE CALCULATIONS

Section 4-2 provides sample calculations using power density and power terms from Tables 1 and 2, whereas Section 4-12 uses these terms plus field intensity and voltage terms from Table 1 and Table 2. Refer the examples in Section 4-12 for usage of the conversions while converting free space values of power density to actual measurements with a spectrum analyzer attached by coaxial cable to a receiving antenna.

Conversion Between Field Intensity (Table 1) and Power Received (Table 2).

Power received (watts or milliwatts) can be expressed in terms of field intensity (volts/meter or μv/meter) using equation [3]:

$$Power\ received\ (P_r) = \frac{E^2}{480\pi^2}\frac{c^2}{f^2}G \qquad [3]$$

or in log form: $$10 \log P_r = 20 \log E + 10 \log G - 20 \log f + 10 \log (c^2/480\pi^2) \qquad [4]$$

Then $$10 \log Pr = 20 \log E_1 + 10 \log G - 20 \log f_1 + K_4 \qquad [5]$$

$$Where\ K_4 = 10 \log\left[\frac{c^2}{480\pi^2} \bullet \left(\begin{matrix} conversions \\ as\ required \end{matrix}\ \frac{(Watts\ to\ mW)}{(volts\ to\ \mu v)^2\ (Hz\ to\ MHz\ or\ GHz)^2}\right)\right]$$

The derivation of equation [3] follows:

$P_D = E^2/120\pi$ Eq [1], Section 4-1, terms (v^2/Ω)

$A_e = \lambda^2 G/4\pi$ Eq [8], Section 3-1, terms (m^2)

$P_r = P_D A_e$ Eq [2], Section 4-3, terms $(W/m^2)(m^2)$

$\therefore\ P_r = (E2/120\pi)(\lambda^2 G/4\pi)$ terms $(v^2/m^2\Omega)(m^2)$

$\lambda = c/f$ Section 2-3, terms (m/sec)(sec)

$\therefore P_r = (E^2/480\pi^2)(c^2 G/f^2)$ which is equation [3]

terms $(v^2/m^2\Omega)(m^2/sec^2)(sec^2)$ or v^2/Ω = watts

Values of K_4 (dB)

P_r	E_1	f_1 (Hz)	f_1 (MHz)	f_1 (GHz)
Watts (dBW)	volts/meter	132.8	12.8	-47.2
	μv/meter	12.8	-107.2	-167.2
mW (dBm)	volts/meter	162.8	42.8	-17.2
	μv/meter	42.8	-77.2	-137.7

Table 2. Conversion Table - Volts to Watts and dBμA.
($P_x = V_x^2/Z$ - Related by line impedance of 50 Ω)

Volts	dBV	dBμV	Watts	dBW	dBm	dBμA
700	56.0	176.0	9800	39.9	69.9	142.9
500	53.9	173.9	5000	37.0	67.0	140.0
300	49.5	169.5	1800	32.5	62.5	135.5
200	46.0	166.0	800	29.0	59.0	132.0
100	40.0	160.0	200	23.0	53.0	126.0
70	36.9	156.9	98	19.9	49.9	122.9
50	34.0	154.0	50	17.0	47.0	120.0
30	29.5	149.5	18	12.5	42.5	115.5
20	26.0	146.0	8	9.0	39.0	112.0
10	20.0	140.0	2	3.0	33.0	106.0
7	16.9	136.9	0.8	0	29.9	102.9
5	14.0	134.0	0.5	-3.0	27.0	100.0
3	9.5	129.5	0.18	-7.4	22.5	95.6
2	6.0	126.0	0.08	-11.0	19.0	92.0
1	0	120.0	0.02	-17.0	13.0	86.0
0.7	-3.1	116.9	9.8×10^{-3}	-20.1	9.9	82.9
0.5	-6.0	114.0	5.0×10^{-3}	-23.0	7.0	80.0
0.3	-10.5	109.5	1.8×10^{-3}	-27.4	2.6	75.6
0.2	-14.0	106.0	8.0×10^{-4}	-31.0	-1.0	72.0
0.1	-20.0	100.0	2.0×10^{-4}	-37.0	-7.0	66.0
.07	-23.1	96.9	9.8×10^{-5}	-40.1	-10.1	62.9
.05	-26.0	94.0	5.0×10^{-5}	-43.0	-13.0	60.0
.03	-30.5	89.5	1.8×10^{-5}	-47.4	-17.7	55.6
.02	-34.0	86.0	8.0×10^{-6}	-51.0	-21.0	52.0
.01	-40.0	80.0	2.0×10^{-6}	-57.0	-27.0	46.0
7×10^{-3}	-43.1	76.9	9.8×10^{-7}	-60.1	-30.1	42.9
5×10^{-3}	-46.0	74.0	5.0×10^{-7}	-63.0	-33.0	40.0
3×10^{-3}	-50.5	69.5	1.8×10^{-7}	-67.4	-37.4	35.6
2×10^{-3}	-54.0	66.0	8.0×10^{-8}	-71.0	-41.0	32.0
1×10^{-3}	-60.0	60.0	2.0×10^{-8}	-77.0	-47.0	26.0
7×10^{-4}	-64.1	56.9	9.8×10^{-9}	-80.1	-50.1	22.9
5×10^{-4}	-66.0	54.0	5.0×10^{-9}	-83.0	-53.0	20.0
3×10^{-4}	-70.5	49.5	1.8×10^{-9}	-87.4	-57.4	15.6
2×10^{-4}	-74.0	46.0	8.0×10^{-10}	-91.0	-61.0	12.0
1×10^{-4}	-80.0	40.0	2.0×10^{-10}	-97.0	-67.0	6.0
7×10^{-5}	-84.1	36.9	9.8×10^{-11}	-100.1	-70.1	2.9
5×10^{-5}	-86.0	34.0	5.0×10^{-11}	-103.0	-73.0	0
3×10^{-5}	-90.5	29.5	1.8×10^{-11}	-107.4	-77.4	-4.4
2×10^{-5}	-94.0	26.0	8.0×10^{-12}	-111.0	-81.0	-8.0
1×10^{-5}	-100.0	20.0	2.0×10^{-12}	-117.0	-87.0	-14.0
7×10^{-6}	-104.1	16.9	9.8×10^{-13}	-120.1	-90.1	-17.1
5×10^{-6}	-106.0	14.0	5.0×10^{-13}	-123.0	-93.0	-20.0
3×10^{-6}	-110.5	9.5	1.8×10^{-13}	-127.4	-97.4	-24.4
2×10^{-6}	-114.0	6.0	8.0×10^{-14}	-131.0	-101.0	-28.0
1×10^{-6}	-120.0	0	2.0×10^{-14}	-137.0	-107.0	-34.0
7×10^{-7}	-124.1	-3.1	9.8×10^{-15}	-140.1	-110.1	-37.1
5×10^{-7}	-126.0	-6.0	5.0×10^{-15}	-143.0	-113.0	-40.0
3×10^{-7}	-130.5	-10.5	1.8×10^{-15}	-147.4	-117.4	-44.4
2×10^{-7}	-134.0	-14.0	8.0×10^{-16}	-151.0	-121.0	-48.0
1×10^{-7}	-140.0	-20.0	2.0×10^{-16}	-157.0	-127.0	-54.0

POWER DENSITY

Radio Frequency (RF) propagation is defined as the travel of electromagnetic waves through or along a medium. For RF propagation between approximately 100 MHz and 10 GHz, radio waves travel very much as they do in free space and travel in a direct line of sight. There is a very slight difference in the dielectric constants of space and air. The dielectric constant of space is one. The dielectric constant of air at sea level is 1.000536. In all but the highest precision calculations, the slight difference is neglected.

From chapter 3, Antennas, an isotropic radiator is a theoretical, lossless, omnidirectional (spherical) antenna. That is, it radiates uniformly in all directions. The power of a transmitter that is radiated from an isotropic antenna will have a uniform power density (power per unit area) in all directions. The power density at any distance from an isotropic antenna is simply the transmitter power divided by the surface area of a sphere ($4\pi R^2$) at that distance. The surface area of the sphere increases by the square of the radius, therefore the power density, P_D, (watts/square meter) decreases by the square of the radius.

$$\begin{array}{c} \text{Power density from} \\ \text{an isotropic antenna} \end{array} = P_D = \frac{P_t}{4\pi R^2} \qquad \begin{array}{r} \text{where : } P_t = \text{Transmitter Power} \\ R = \text{Range From Antenna (i.e., radius of sphere)} \end{array} \qquad \textbf{[1]}$$

P_t is either peak or average power depending on how P_D is to be specified.

Radars use directional antennas to channel most of the radiated power in a particular direction. The Gain (G) of an antenna is the ratio of power radiated in the desired direction as compared to the power radiated from an isotropic antenna, or:

$$G = \frac{\text{Maximum radiation intensity of actual antenna}}{\text{Radiation intensity of isotropic antenna with same power input}}$$

The power density at a distant point from a radar with an antenna gain of G_t is the power density from an isotropic antenna multiplied by the radar antenna gain.

Power density from radar, $$P_D = \frac{P_t G_t}{4\pi R^2} \qquad \textbf{[2]}$$

P_t is either peak or average power depending on how P_D is to be specified.

Another commonly used term is effective isotropic radiated power (EIRP), where **EIRP = P_t G_t.** ERP is also used but EIRP is preferred because it specifically defines the type of reference antenna as isotropic.

A receiving antenna captures a portion of this power determined by its effective capture Area (A_e). The received power available at the antenna terminals is the power density times the effective capture area (A_e) of the receiving antenna.

e.g., If the power density at a specified range is one microwatt per square meter and the antenna's effective capture area is one square meter then the power captured by the antenna is one microwatt.

For a given receiver antenna size the capture area is constant no matter how far it is from the transmitter, as illustrated in Figure 1. Also notice from Figure 1 that the received signal power decreases by 1/4 (6 dB) as the distance doubles. This is due to the R^2 term in the denominator of equation [2].

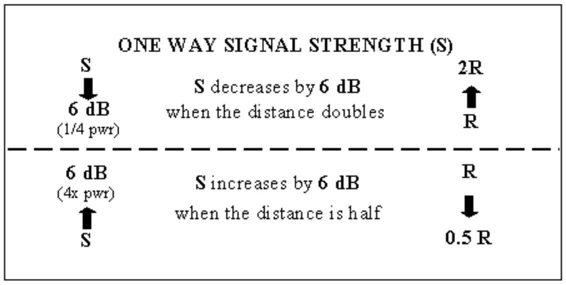
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB
when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB
when the distance is half
R
0.5 R

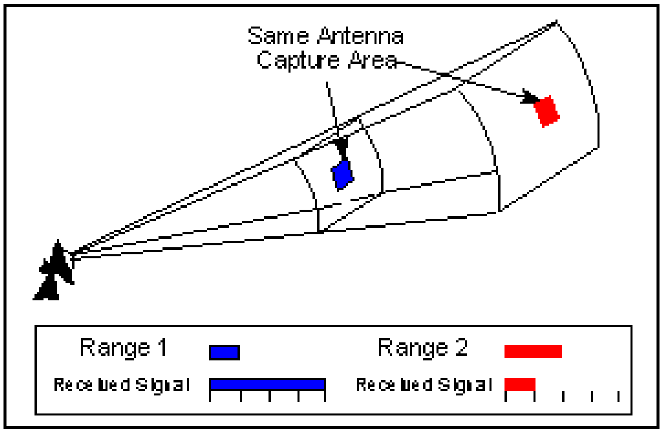
Same Antenna
Capture Area
Range 1
Received Signal
Range 2
Received Signal

Figure 1. Power Density vs. Range.

Sample Power Density Calculation - Far Field
(Refer to Section 3-5 for the definition of near field and far field)

Calculate the power density at 100 feet for 100 watts transmitted through an antenna with a gain of 10.

Given: P_t = 100 watts G_t = 10 (dimensionless ratio) R = 100 ft

This equation produces power density in watts per square range unit.

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(100 \text{ watts})(10)}{4\pi (100 \text{ ft})^2} = 0.0080 \text{ watts/ ft}^2$$

For safety (radiation hazard) and EMI calculations, power density is usually expressed in milliwatts per square cm. That's nothing more than converting the power and range to the proper units.

100 watts = 1×10^2 watts = 1×10^5 mW

100 feet = 30.4785 meters = 3047.85 cm.

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(10^5 \text{ mW}) \bullet (10)}{4\pi (3047.85 \text{cm})^2} = 0.0086 \text{ mW/ cm}^2$$

However, antenna gain is almost always given in dB, not as a ratio. It's then often easier to express EIRP in dBm.

$$P_t (dBm) = 10 \text{ Log}\left[\frac{P_t \text{ watts}}{1 \text{ mW}}\right] = 10 \text{ Log}\left[\frac{100}{.001}\right] = 50 \text{ dBm}$$

$$G_t (dB) = 10 \text{ Log}\left[\frac{G_t}{1}\right] = 10 \text{ Log}(10) = 10 \text{ dB}$$

EIRP (dBm) = P_t (dBm) + G_t (dB) = 50 + 10 = 60 dBm

To reduce calculations, the graph in Figure 2 can be used. It gives EIRP in dBm, range in feet and power density in mW/cm^2. Follow the scale A line for an EIRP of 60 dBm to the point where it intersects the 100 foot range scale. Read the power density directly from the A-scale x-axis as 0.0086 mW/cm^2 (confirming our earlier calculations).

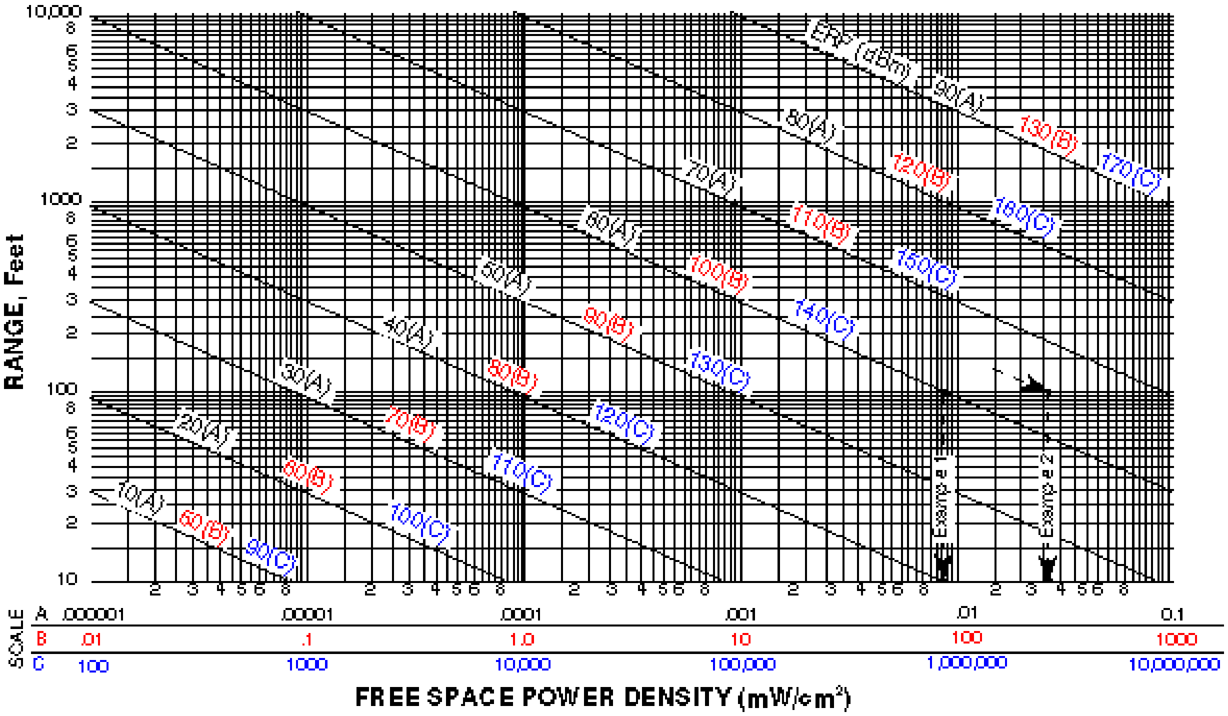

Figure 2. Power Density vs. Range and EIRP.

Example 2

When antenna gain and power (or EIRP) are given in dB and dBm, it is necessary to convert back to ratios in order to perform the calculation given in equation [2]. Use the same values as in example 1 except for antenna gain.

Suppose the antenna gain is given as 15 dB: **G_t (dB) = 10 Log (G_t)**

$$Therefore: \quad G_t = [10]^{\left(\frac{G_t(dB)}{10}\right)} = [10]^{\left(\frac{15}{10}\right)} = 31.6228$$

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{(10^5\ mW)(31.6228)}{4\pi\,(3047.85\ )^2} = 0.0271\ mW/\ cm^2$$

Follow the 65 dBm (extrapolated) EIRP line and verify this result on the A-scale X-axis.

Example 3 - Sample Real Life Problem

Assume we are trying to determine if a jammer will damage the circuitry of a missile carried onboard an aircraft and we cannot perform an actual measurement. Refer to the diagram at the right.

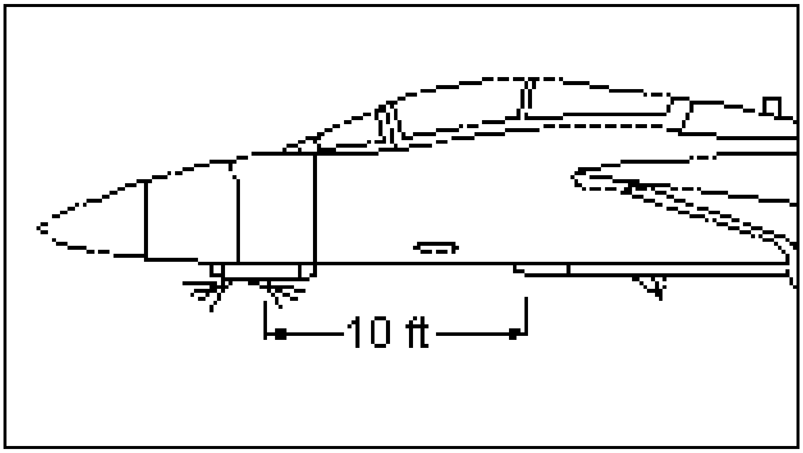

Given the following:

Jammer power: 500 W ($P_t = 500$)

Jammer line loss and antenna gain: 3 dB ($G_t = 2$)

Missile antenna diameter: 10 in

Missile antenna gain: Unknown

Missile limiter protection (maximum antenna power input): 20 dBm (100mW) average and peak.

The power density at the missile antenna caused by the jammer is computed as follows:

$$P_D = \frac{P_t G_t}{4\pi R^2} = \frac{500W\,(2)}{4\pi[(10ft)(.3048m/ft)]^2} = 8.56W/m^2$$

The maximum input power actually received by the missile is either:

$P_r = P_D A_e$ (if effective antenna area is known) or
$P_r = P_D G_m \lambda^2/4\pi$ (if missile antenna gain is known)

To cover the case where the missile antenna gain is not known, first assume an aperture efficiency of 0.7 for the missile antenna (typical). Then:

$P_r = P_D A \eta = 8.56 \text{ W/m}^2 (\pi)[(10/2 \text{ in})(.0254 \text{ m/in})]^2 (0.7) = 0.3$ watts

Depending upon missile antenna efficiency, we can see that the power received will be about 3 times the maximum allowable and that either better limiter circuitry may be required in the missile or a new location is needed for the missile or jammer. Of course if the antenna efficiency is 0.23 or less, then the power will not damage the missile's receiver.

If the missile gain were known to be 25 dB, then a more accurate calculation could be performed. Using the given gain of the missile (25 dB= numeric gain of 316), and assuming operation at 10 GHz ($\lambda = .03$m)

$P_r = P_D G_m \lambda^2 / 4\pi = 8.56 \text{ W/m}^2 (316)(.03)^2/ 4\pi = .19$ watts (still double the allowable tolerance)

ONE-WAY RADAR EQUATION / RF PROPAGATION

The one-way (transmitter to receiver) radar equation is derived in this section. This equation is most commonly used in RWR, communications, or ESM type of applications. The following is a summary of the important equations explored in this section:

ONE-WAY RADAR EQUATION

Peak Power at

Receiver Input, $P_r\,(or\ S) = P_D A_e = \frac{P_t G_t A_e}{4\pi R^2}$ and Antenna Gain, $G = \frac{4\pi A_e}{\lambda^2}$

or : Equivalent Area, $A_e = \frac{G\lambda^2}{4\pi}$

So the one-way radar equation is :

$$S\,(or\ P_r) = \frac{P_t G_t G_r \lambda^2}{(4\pi R)^2} = P_t G_t G_r \left[\frac{c^2}{(4\pi f R)^2}\right]^* \qquad (\ Note : \lambda = \frac{c}{f}\ )$$

* keep λ, c, and R in the same units

On reducing to log form this becomes:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - 20\log fR + 20\log (c/4\pi)$

or in simplified terms:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_1$ (in dB)

Where: α_1 = one-way free space loss = $20\log (f_1 R) + K_1$ (in dB) and:
$K_1 = 20\log [(4\pi/c)$(Conversion factors if units if not in m/sec, m, and Hz)]

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain

Values of K_1 (in dB)

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 & 5-1) are not included in any of these equations.

Recall from Section 4-2 that the power density at a distant point from a radar with an antenna gain of G_t is the power density from an isotropic antenna multiplied by the radar antenna gain.

Power density from radar, $P_D = \frac{P_t G_t}{4\pi R^2}$

If you could cover the entire spherical segment with your receiving antenna you would theoretically capture all of the transmitted energy. You can't do this because no antenna is large enough. (A two degree segment would be about a mile and three-quarters across at fifty miles from the transmitter.)

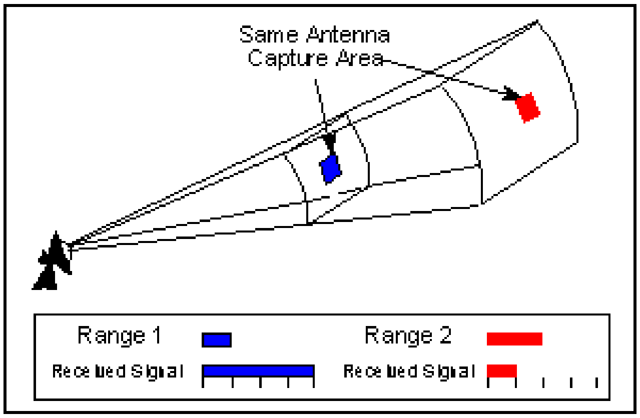

Figure 1. Power Density vs. Range.

A receiving antenna captures a portion of this power determined by its effective capture Area (A_e). The received power available at the antenna terminals is the power density times the effective capture area (A_e) of the receiving antenna.

For a given receiver antenna size the capture area is constant no matter how far it is from the transmitter, as illustrated in Figure 1. This concept is shown in the following equation:
Peak Power at Receiver input,

$$P_R \text{ (or S)} = P_D A_e = \frac{P_t G_t A_e}{4\pi R^2} \quad \text{which is known as the one-way (beacon) equation} \qquad [2]$$

In order to maximize energy transfer between an antenna and transmitter or receiver, the antenna size should correlate with frequency. For reasonable antenna efficiency, the size of an antenna will be greater than $\lambda/4$. Control of beamwidth shape may become a problem when the size of the active element exceeds several wavelengths.

The relation between an antenna's effective capture area (A_e) or effective aperture and its Gain (G) is:

$$\text{Antenna Gain, } G = \frac{4\pi A_e}{\lambda^2} \qquad [3]$$

$$\text{or : Equivalent Area, } A_e = \frac{G\lambda^2}{4\pi} \qquad [4]$$

Since the effective aperture is in units of length squared, from equation [3], it is seen that gain is proportional to the effective aperture normalized by the wavelength. This physically means that to maintain the same gain when doubling the frequency, the area is reduced by 1/4. This concept is illustrated in Figure 2.

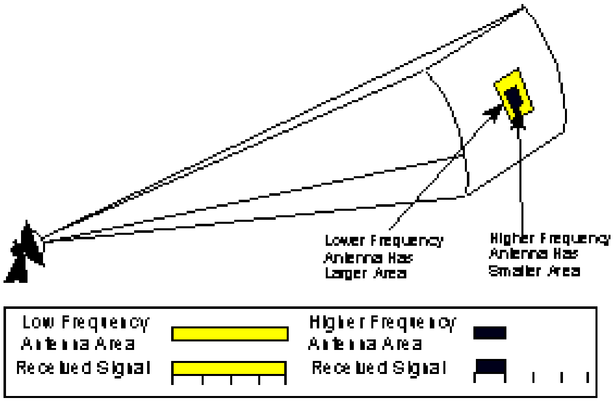
Lower Frequency Antenna Has Larger Area
Higher Frequency Antenna Has Smaller Area
Low Frequency Antenna Area
Received Signal
Higher Frequency Antenna Area
Received Signal

Figure 2. Capture Area vs. Frequency.

If equation [4] is substituted into equation [2], the following relationship results:

$$\text{Peak Power at Receiver Input} = S \text{ (or } P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2} = \frac{P_t G_t G_r \lambda^2}{(4\pi R)^2} \qquad [5]$$

This is the signal calculated one-way from a transmitter to a receiver. For instance, a radar application might be to determine the signal received by a RWR, ESM, or an ELINT receiver. It is a general purpose equation and could be applied to almost any line-of-sight transmitter to receiver situation if the RF is higher than 100 MHZ.

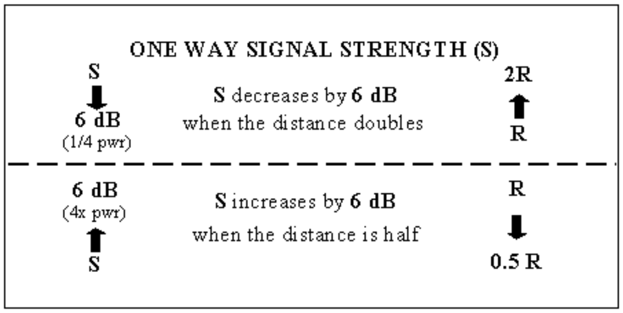
ONE WAY SIGNAL STRENGTH (S)
S
6 dB
(1/4 pwr)
S decreases by 6 dB
when the distance doubles
2R
R
6 dB
(4x pwr)
S
S increases by 6 dB
when the distance is half
R
0.5 R

Values of K_1 (dB)

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

The free space travel of radio waves can, of course, be blocked, reflected, or distorted by objects in their path such as buildings, flocks of birds, chaff, and the earth itself.

As illustrated in Figure 1, as the distance is doubled the received signal power decreases by 1/4 (6 dB). This is due to the R^2 term in equation [5].

To illustrate this, blow up a round balloon and draw a square on the side of it. If you release air so that the diameter or radius is decreased by 1/2, the square shrinks to 1/4 the size. If you further blow up the balloon, so the diameter or radius is doubled, the square has quadrupled in area.

The one-way free space loss factor (α_1), (sometimes called the path loss factor) is given by the term $(4\pi R^2)(4\pi/\lambda^2)$ or $(4\pi R/\lambda)^2$. As shown in Figure 3, the loss is due to the ratio of two factors (1) the effective radiated area of the transmit antenna, which is the surface area of a sphere ($4\pi R^2$) at that distance (R), and (2) the effective capture area (A_e) of the receive antenna which has a gain of one. If a receiving antenna could capture the whole surface area of the sphere, there would be no spreading loss, but a practical antenna will capture only a small part of the spherical radiation. Space loss is calculated using isotropic antennas for both transmit and receive, so α_1 is independent of the actual antenna. Using Gr = 1 in equation [11] in Section 3-1, $A_e = \lambda^2/4\pi$. Since this term is in the denominator of α_1, the higher the frequency (lower λ) the more the space loss. Since G_t and G_r are part of the one-way radar equation, S (or P_r) is adjusted according to actual antennas as shown in the last portion of Figure 3. The value of the received signal (S) is:

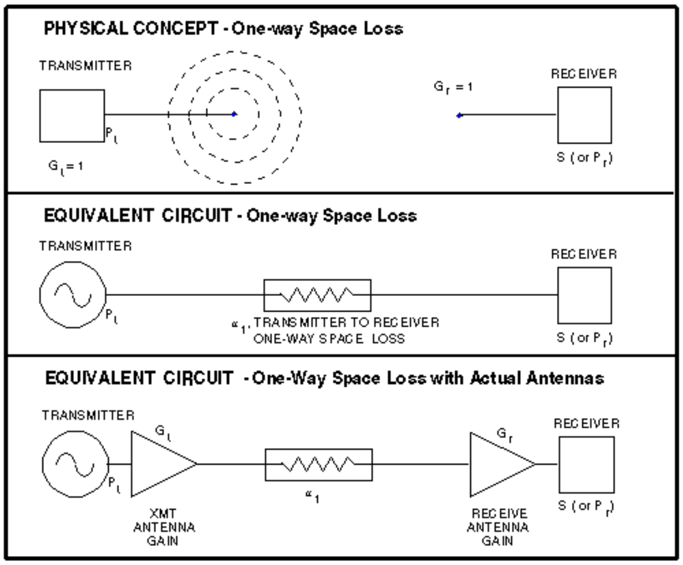
PHYSICAL CONCEPT - One-way Space Loss
TRANSMITTER
P_t
$G_t = 1$
$G_r = 1$
RECEIVER
S (or P_r)
EQUIVALENT CIRCUIT - One-way Space Loss
TRANSMITTER
P_t
α_1, TRANSMITTER TO RECEIVER ONE-WAY SPACE LOSS
RECEIVER
S (or P_r)
EQUIVALENT CIRCUIT - One-Way Space Loss with Actual Antennas
TRANSMITTER
G_t
P_t
XMT ANTENNA GAIN
α_1
G_r
RECEIVE ANTENNA GAIN
RECEIVER
S (or P_r)

Figure 3. Concept of One-Way Space Loss.

$$S\,(or\ P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2} = P_t G_t G_r \left[\frac{\lambda^2}{(4\pi R)^2}\right] \qquad \textbf{[6]}$$

To convert this equation to dB form, it is rewritten as:

$$10\,\text{Log}\,(\text{S or P}_r) = 10\,\text{Log}\,(\text{P}_t\text{G}_t\text{G}_r) + 20\,\text{Log}\left[\frac{\lambda}{4\pi f R}\right] \qquad \textbf{[7]}$$

Since $\lambda = c / f$, equation [7] can be rewritten as:

$$10\text{ Log }(\text{S or P}_r) = 10\text{ Log}(\text{P}_t\,\text{G}_t\,\text{G}_r) - \alpha_1 \qquad \textbf{[8]}$$

Where the one-way free space loss, α_1, is defined as:
$$\alpha_1 = 20\,\text{Log}\left[\frac{4\pi f R}{c}\right]^* \qquad \textbf{[9]}$$

The signal received equation in dB form is: $10\log (P_r \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r - \alpha_1$ **[10]**

The one-way free space loss, α_1, can be given in terms of a variable and constant term as follows:

$$\alpha_1 = 20 \,\text{Log}\left[\frac{4\pi f R}{c}\right]^* = 20 \,\text{Log}\, f_1 R + K_1 \quad (in\ dB) \qquad \textbf{[11]}$$

The value of f_1 can be either in MHz or GHz as shown with commonly used units of R in the adjoining table.

$$where\ K_1 = 20\,\text{Log}\left[\frac{4\pi}{c} \bullet (Conversion\ units\ if\ \ not\ in\ m/\text{sec}, m, and\ \ Hz)\right]$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

A value for the one-way free space loss (α_1) can be obtained from:

(a) The One-way Free Space Loss graph (Figure 4). Added accuracy can be obtained using the Frequency Extrapolation graph (Figure 5)

(b) The space loss nomograph (Figure 6 or 7)

(c) The formula for α_1, equation [11].

FOR EXAMPLE:

Find the value of the one-way free space loss, α_1, for an RF of 7.5 GHz at 100 NM.

(a) From Figure 4, find 100 NM on the X-axis and estimate where 7.5 GHz is located between the 1 and 10 GHz lines (note dot). Read α_1 as 155 dB. An alternate way would be to read the α_1 at 1 GHz (138 dB) and add the frequency extrapolation value (17.5 dB for 7.5:1, dot on Figure 5) to obtain the same 155 dB value.

(b) From the nomogram (Figure 6), the value of α_1 can be read as 155 dB (Note the dashed line).

(c) From the equation 11, the precise value of α_1 is 155.3 dB.

Remember, α_1 is a free space value. If there is atmospheric attenuation because of absorption of RF due to certain molecules in the atmosphere or weather conditions etc., the atmospheric attenuation is in addition to the space loss (refer to Section 5-1).

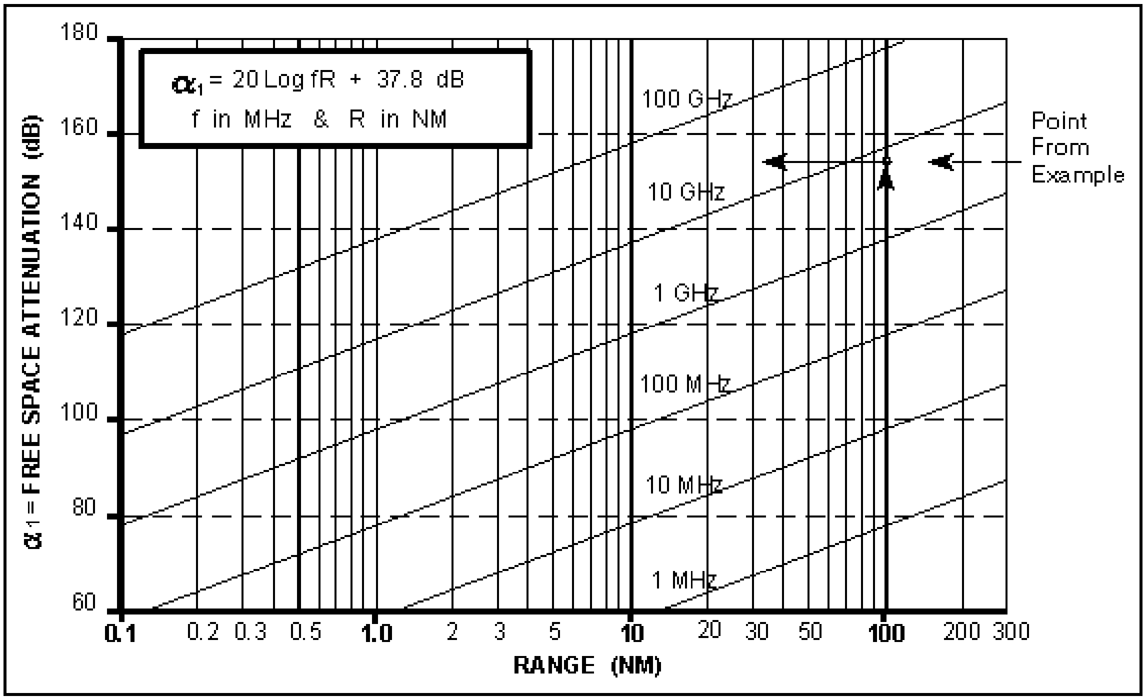

Figure 4. One-Way Free Space Loss.

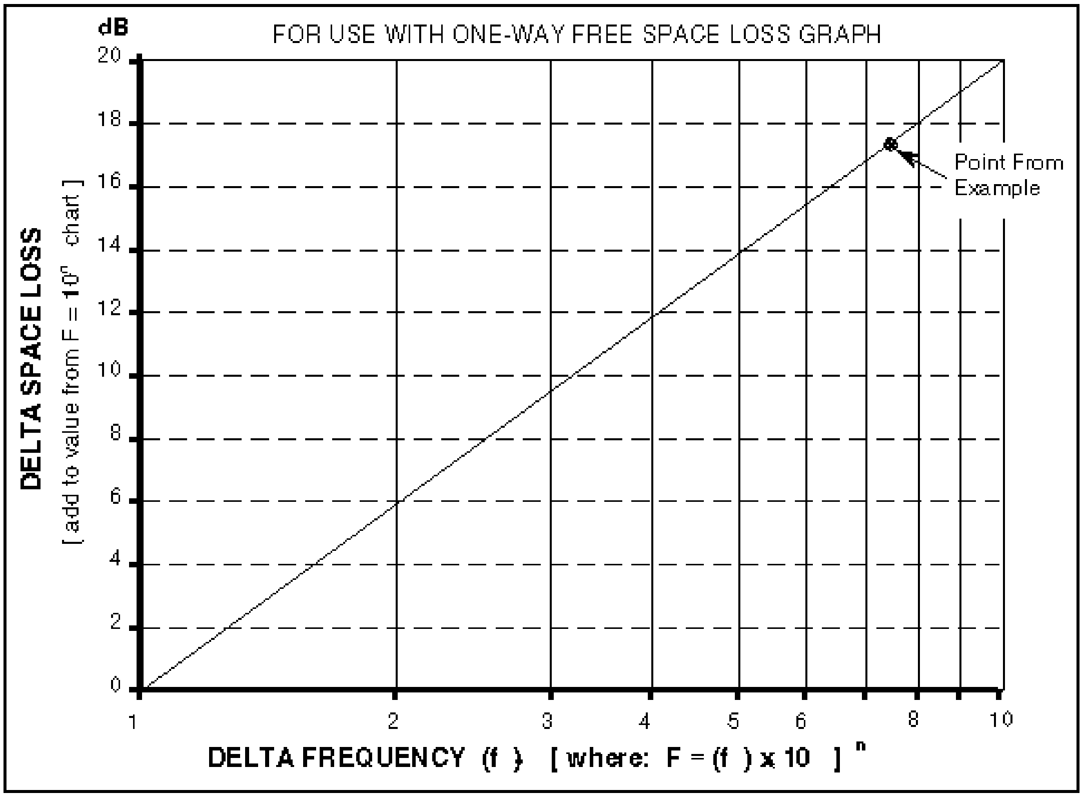

Figure 5. Frequency Extrapolation.

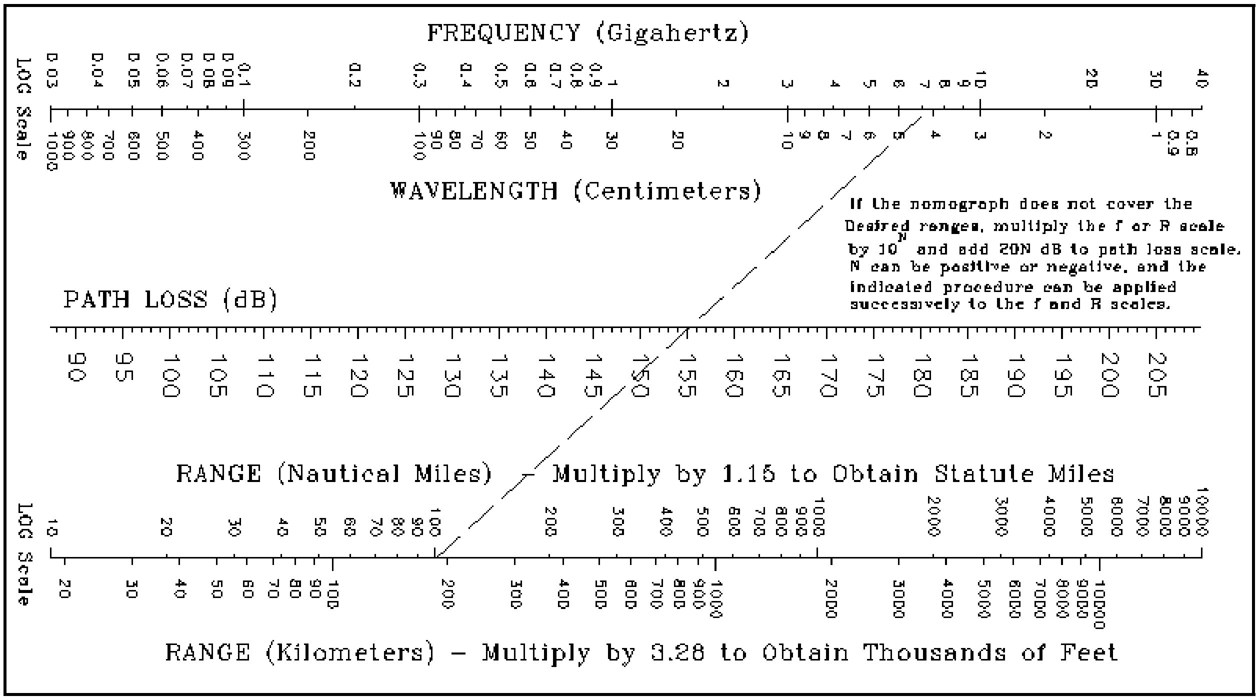

If the nomograph does not cover the Desired ranges, multiply the f or R scale by 10^N and add 20N dB to path loss scale. N can be positive or negative, and the indicated procedure can be applied successively to the f and R scales.

Figure 6. One-Way Space Loss Nomograph for Distances Greater Than 10 Nautical Miles.

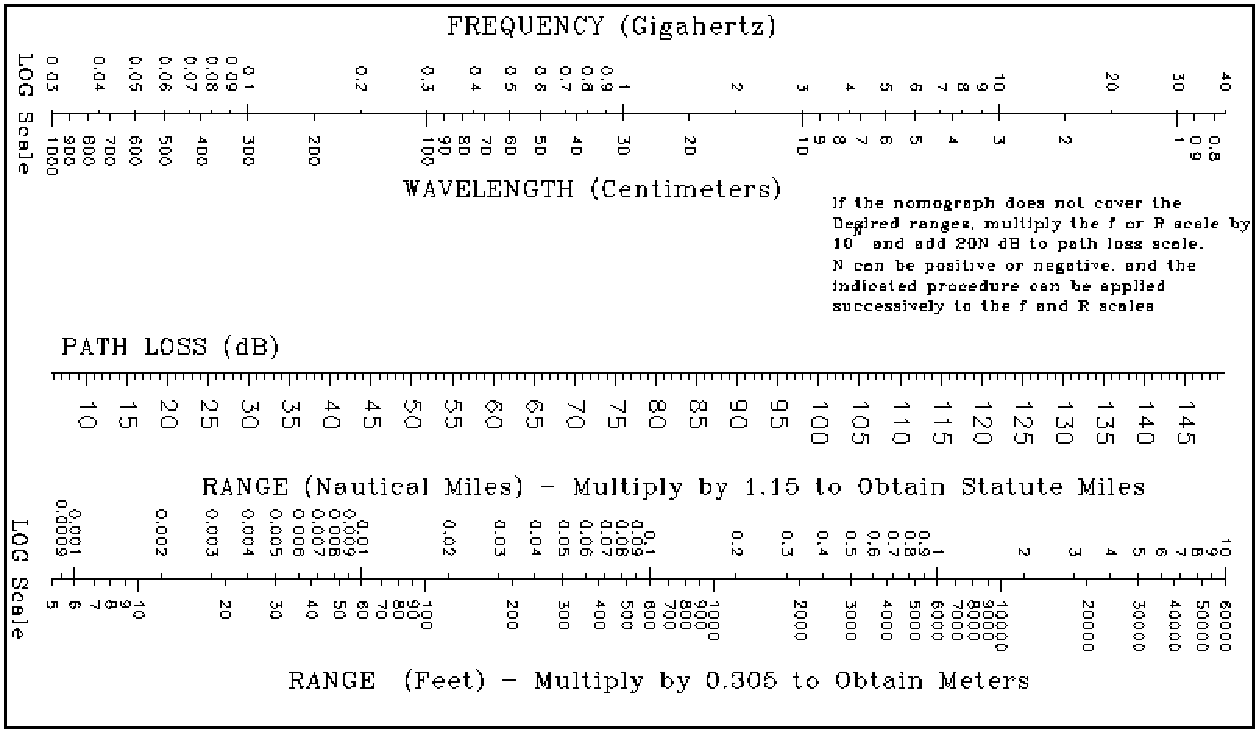

If the nomograph does not cover the Desired ranges, multiply the f or R scale by 10^N and add 20N dB to path loss scale. N can be positive or negative, and the indicated procedure can be applied successively to the f and R scales

Figure 7. One-Way Space Loss Nomograph for Distances Less Than 10 Nautical Miles.

Figure 8 is the visualization of the losses occurring in one-way radar equation. Note: To avoid having to include additional terms, always combine any transmission line loss with antenna gain. Losses due to antenna polarization and atmospheric absorption also need to be included.

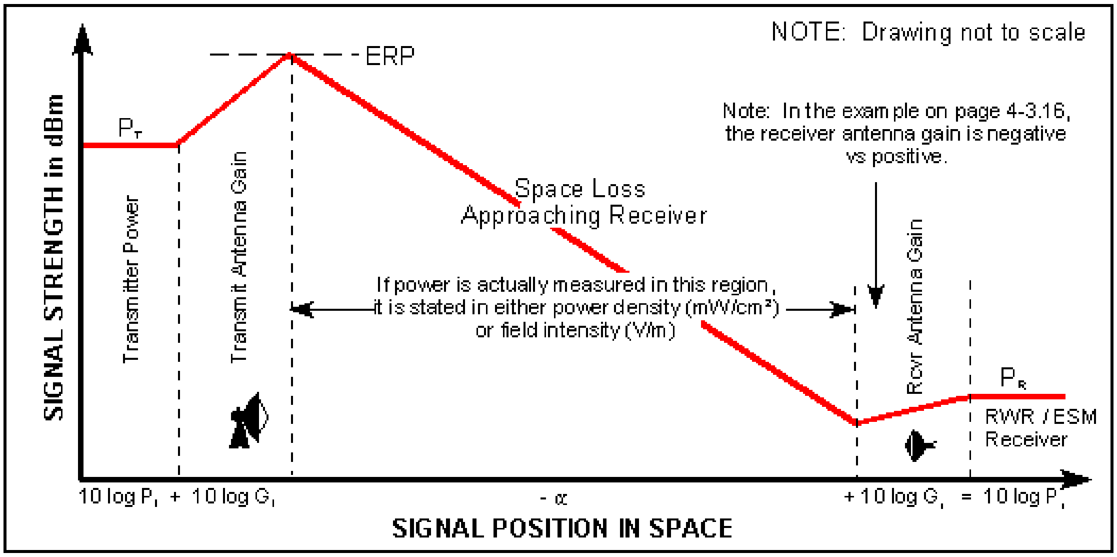
NOTE: Drawing not to scale
ERP
SIGNAL STRENGTH in dBm
PT
Transmitter Power
Transmit Antenna Gain
Space Loss
Approaching Receiver
Note: In the example on page 4-3.16, the receiver antenna gain is negative vs positive.
If power is actually measured in this region, it is stated in either power density (mW/cm²) or field intensity (V/m)
Rcvr Antenna Gain
PR
RWR / ESM
Receiver
10 log Pt
+ 10 log Gt
- α
+ 10 log Gr
= 10 log Pr
SIGNAL POSITION IN SPACE

Figure 8. Visualization of One-Way Radar Equation.

RWR/ESM RANGE EQUATION (One-Way)

The one-way radar (signal strength) equation [5] is rearranged to calculate the maximum range R_{max} of RWR/ESM receivers. It occurs when the received radar signal just equals S_{min} as follows:

$$R_{max} \cong \left[\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 S_{min}}\right]^{\frac{1}{2}} \text{ or } \left[\frac{P_t G_t G_r c^2}{(4\pi f)^2 S_{min}}\right]^{\frac{1}{2}} \text{ or } \left[\frac{P_t G_t A_e}{4\pi S_{min}}\right]^{\frac{1}{2}} \qquad [12]$$

In log form:

$$20\log R_{max} = 10\log P_t + 10\log G_t - 10\log S_{min} - 20\log f + 20\log(c/4\pi) \qquad [13]$$

and since $K_1 = 20\log\{4\pi/c$ times conversion units if not in m/sec, m, and Hz$\}$
(Refer to Section 4-3 for values of K_1).

$$10\log R_{max} = \tfrac{1}{2}[\,10\log P_t + 10\log G_t - 10\log S_{min} - 20\log f - K_1] \quad \text{(keep } P_t \text{ and } S_{min} \text{ in same units)} \qquad [14]$$

If you want to convert back from dB, then $R_{max} \cong \frac{10^{M\ dB}}{20}$, where M dB is the resulting number in the brackets of equation 14.

From Section 5-2, Receiver Sensitivity / Noise, S_{min} is related to the noise factor S:

$$S_{min} = (S/N)_{min}\,(NF)KT_oB \qquad [15]$$

The one-way RWR/ESM range equation becomes:

$$R_{max} \cong \left[\frac{P_t G_t G_r \lambda^2}{(4\pi)^2 (S/N)_{min} (NF) K T_o B} \right]^{\frac{1}{2}} \; or \; \left[\frac{P_t G_t G_r c^2}{(4\pi f)^2 (S/N)_{min} (NF) K T_o B} \right]^{\frac{1}{2}} \; or \; \left[\frac{P_t G_t A_e}{4\pi (S/N)_{min} (NF) K T_o B} \right]^{\frac{1}{2}} \qquad \textbf{[16]}$$

RWR/ESM RANGE INCREASE AS A RESULT OF A SENSITIVITY INCREASE

As shown in equation [12] $S_{min}^{-1} \propto R_{max}^2$. Therefore, $-10 \log S_{min} \propto 20 \log R_{max}$ and the following table results:

% Range Increase: Range + (% Range Increase) x Range = New Range
i.e., for a 6 dB sensitivity increase, 500 miles +100% x 500 miles = 1,000 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 6 dB sensitivity increase 500 miles x 2 = 1,000 miles

dB Sensitivity Increase	% Range Increase	Range Multiplier
+ 0.5	6	1.06
1.0	12	1.12
1.5	19	1.19
2	26	1.26
3	41	1.41
4	58	1.58
5	78	1.78
6	100	2.0
7	124	2.24
8	151	2.51
9	182	2.82
10	216	3.16
11	255	3.55
12	298	3.98
13	347	4.47
14	401	5.01
15	462	5.62
16	531	6.31
17	608	7.08
18	694	7.94
19	791	8.91
20	900	10.0

RWR/ESM RANGE DECREASE AS A RESULT OF A SENSITIVITY DECREASE

As shown in equation [12] $S_{min}^{-1} \propto R_{max}^2$. Therefore, $-10 \log S_{min} \propto 20 \log R_{max}$ and the following table results:

% Range Decrease: Range - (% Range decrease) x Range = New Range
i.e., for a 6 dB sensitivity decrease, 500 miles - 50% x 500 miles = 250 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 6 dB sensitivity decrease 500 miles x .5 = 250 miles

dB Sensitivity Decrease	% Range Decrease	Range Multiplier	dB Sensitivity Decrease	% Range Decrease	Range Multiplier
- 0.5	6	0.94	-10	68	0.32
- 1.0	11	0.89	- 11	72	0.28
- 1.5	16	0.84	- 12	75	0.25
- 2	21	0.79	- 13	78	0.22
- 3	29	0.71	- 14	80	0.20
- 4	37	0.63	- 15	82	0.18
- 5	44	0.56	- 16	84	0.16
- 6	50	0.50	- 17	86	0.14
- 7	56	0.44	- 18	87	0.13
- 8	60	0.4	- 19	89	0.11
- 9	65	0.35	- 20	90	0.10

Example of One-Way Signal Strength: A 5 (or 7) GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an aft EW antenna with -1 dB gain and a 5 dB line loss to the EW receiver (assume all antenna polarizations are the same).

Note: The respective transmission line losses will be combined with antenna gains, i.e.:

-5 +45 = 40 dB, -5 - 1 = -6 dB, -10 + 5 = -5 dB.

(1) What is the power level at the input of the EW receiver?

Answer (1): P_r at the input to the EW receiver = Transmitter power - xmt cable loss + xmt antenna gain - space loss + rcvr antenna gain - rcvr cable loss.

Space loss (from Section 4-3) @ 5 GHz = $20 \log f R + K_1 = 20 \log (5x31) + 92.44 = 136.25$ dB. Therefore:

$$P_r = 70 + 40 - 136.25 - 6 = -32.25 \text{ dBm @ 5 GHz } (P_r = -35.17 \text{ dBm @ 7 GHz since } \alpha_1 = 139.17 \text{ dB})$$

(2) If the received signal is fed to a jammer with a gain of 60 dB, feeding a 10 dB loss transmission line which is connected to an antenna with 5 dB gain, what is the power level from the jammer at the input to the receiver of the 5 (or 7) GHz radar?

Answer (2): P_r at the input to the radar receiver = Power at the input to the EW receiver+ Jammer gain - jammer cable loss + jammer antenna gain - space loss + radar rcvr antenna gain - radar rcvr cable loss. Therefore:

$$P_r = -32.25 + 60 - 5 - 136.25 + 40 = -73.5 \text{ dBm @ 5 GHz. } (P_r = -79.34 \text{ dBm @ 7 GHz since } \alpha_1 = 139.17 \text{ dB and } P_t = -35.17 \text{ dBm}).$$

This problem continues in Sections 4-4, 4-7, and 4-10.

This page intentionally left blank.

TWO-WAY RADAR EQUATION (MONOSTATIC)

In this section the radar equation is derived from the one-way equation (transmitter to receiver) which is then extended to the two-way radar equation. The following is a summary of the important equations to be derived here:

TWO-WAY RADAR EQUATION (MONOSTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} = P_t G_t G_r \left[\frac{\sigma c^2}{(4\pi)^3 f^2 R^4} \right]^*$$

Note : $\lambda = c/f$ and $\sigma = RCS$ ⊥

** keep λ or c, σ, and R in the same units*

On reducing the above equation to log form we have:
$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f - 40\log R - 30\log 4\pi + 20\log c$

or in simplified terms: $10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1$ (in dB)

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K₂ Values (dB)

RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

One-way free space loss, $\alpha_1 = 20\log (f_1 R) + K_1$ (in dB)

K₁ Values (dB)

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
Km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

Figure 1 illustrates the physical concept and equivalent circuit for a target being illuminated by a monostatic radar (transmitter and receiver co-located). Note the similarity of Figure 1 to Figure 3 in Section 4-3. Transmitted power, transmitting and receiving antenna gains, and the one-way free space loss are the same as those described in Section 4-3. The physical arrangement of the elements is different, of course, but otherwise the only difference is the addition of the equivalent gain of the target RCS factor.

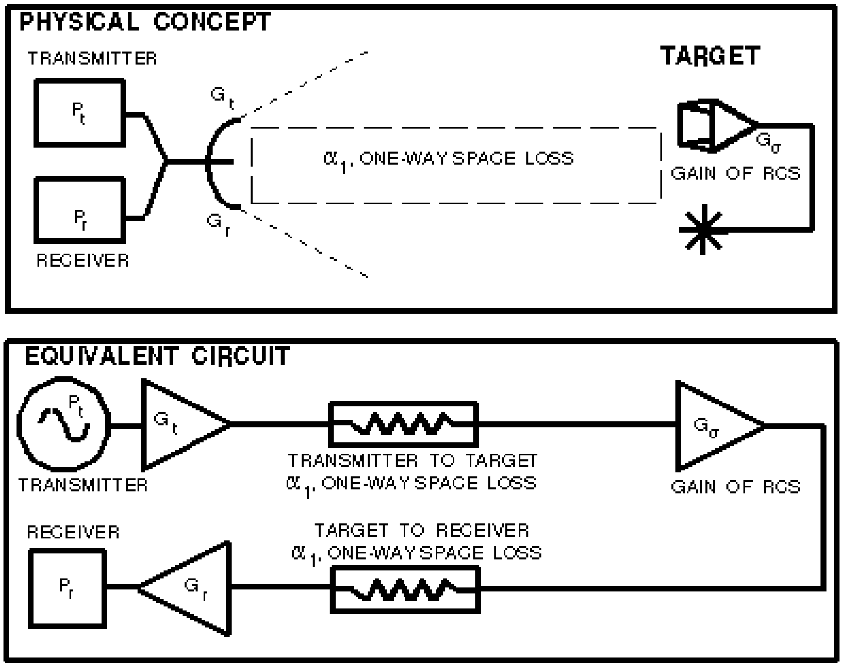
PHYSICAL CONCEPT
TRANSMITTER
P_t
G_t
TARGET
α_1, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
P_r
G_r
RECEIVER
EQUIVALENT CIRCUIT
P_t
G_t
TRANSMITTER
TRANSMITTER TO TARGET
α_1, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
RECEIVER
TARGET TO RECEIVER
α_1, ONE-WAY SPACE LOSS
P_r
G_r

Figure 1. The Two-Way Monostatic Radar Equation Visualized.

From Section 4-3, One-Way Radar Equation / RF Propagation, the power in the receiver is:

$$\begin{matrix}\textit{Received Signal}\\ \textit{at Target}\end{matrix} = \frac{P_t\,G_t\,G_r\,\lambda^2}{(4\pi R\,)^2} \qquad \textbf{[1]}$$

From equation [3] in Section 4-3: $$\textit{Antenna Gain}\,,\,G = \frac{4\pi\;A_e}{\lambda^2} \qquad \textbf{[2]}$$

Similar to a receiving antenna, a radar target also intercepts a portion of the power, but reflects (reradiates) it in the direction of the radar. The amount of power reflected toward the radar is determined by the Radar Cross Section (RCS) of the target. RCS is a characteristic of the target that represents its size as seen by the radar and has the dimensions of area (σ) as shown in Section 4-11. RCS area is not the same as physical area. But, for a radar target, the power reflected in the radar's direction is equivalent to re-radiation of the power captured by an antenna of area σ (the RCS). Therefore, the effective capture area (A_e) of the receiving antenna is replaced by the RCS (σ).

$$G_r = \frac{4\pi\sigma}{\lambda^2} \qquad \textbf{[3]}$$

so we now have:

$$\begin{matrix}\textit{Reflected Signal}\\ \textit{from target}\end{matrix} = \frac{P_t\,G_t\,\lambda^2\,4\pi\sigma}{(4\pi R\,)^2\,\lambda^2} \qquad \textbf{[4]}$$

The equation for the power reflected in the radar's direction is the same as equation [1] except that P_t G_t, which was the original transmitted power, is replaced with the reflected signal power from the target, from equation [4]. This gives:

$$\begin{matrix}\textit{Reflected Signal Received Back}\\ \textit{at Input to Radar Receiver}\end{matrix} = \frac{P_t\,G_t\,\lambda^2\,4\pi\sigma}{(4\pi R\,)^2\,\lambda^2}\;x\;\frac{G_r\,\lambda^2}{(4\pi R\,)^2} \qquad \textbf{[5]}$$

If like terms are cancelled, the two-way radar equation results. The peak power at the radar receiver input is:

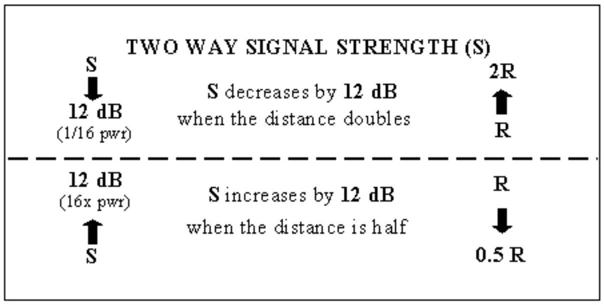

$$P_r = \frac{P_t\,G_t\,G_r\,\lambda^2\,\sigma}{(4\pi\,)^3\,R^4} = P_t\,G_t\,G_r\left[\frac{\sigma\,c^2}{(4\pi\,)^3\,f^{\,2}\,R^4}\right]^{*} \qquad \textbf{[6]}$$

* Note: λ=c/*f* and σ = RCS. Keep λ or c, σ, and R in the same units.

On reducing equation [6] to log form we have:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f - 40\log R - 30\log 4\pi + 20\log c \qquad \textbf{[7]}$$

Target Gain Factor

If Equation [5] terms are rearranged instead of cancelled, a recognizable form results:

$$S\,(or\ P_r) = (P_t G_t G_r) \bullet \left[\frac{\lambda^2}{(4\pi R)^2}\right] \bullet \left[\frac{4\pi\sigma}{\lambda^2}\right] \bullet \left[\frac{\lambda^2}{(4\pi R)^2}\right] \qquad \textbf{[8]}$$

In log form:

$$10\log[S\,(or\ P_r)] = 10\log P_t + 10\log G_t + 10\log G_r + 20\log\left[\frac{\lambda}{4\pi R}\right] + 10\log\left[\frac{4\pi\sigma}{\lambda^2}\right] + 20\log\left[\frac{\lambda}{4\pi R}\right] \qquad \textbf{[9]}$$

The fourth and sixth terms can each be recognized as -α, where α is the one-way free space loss factor defined in Section 4-3. The fifth term containing RCS (σ) is the only new factor, and it is the "Target Gain Factor."

In simplified terms the equation becomes:

$$10\log\,[S\ (or\ P_r)] = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \qquad (in\ dB) \qquad \textbf{[10]}$$

Where α_1 and G_σ are as follows:

From Section 4-3, equation [11], the space loss in dB is given by:

$$\alpha_1 = 20\log\left[\frac{4\pi f\,R}{c}\right]^* = 20\log f_1 R + K_1 \quad where\ K_1 = 20\log\left[\frac{4\pi}{c} \bullet (Conversion\ units\ if\ not\ in\ m/sec,\ m,\ and\ Hz)\right] \qquad \textbf{[11]}$$

* Keep c and R in the same units. The table of values for K_1 is again presented here for completeness. The constant, K_1, in the table includes a range and frequency unit conversion factor.

One-way free space loss, α_1 = 20log (f_1R) + K_1 (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

While it's understood that RCS is the antenna aperture area equivalent to an isotropically radiated target return signal, the target gain factor represents a gain, as shown in the equivalent circuit of Figure 1. The Target Gain Factor expressed in dB is G_σ as shown in equation [12].

$$G_\sigma = 10\log\left[\frac{4\pi\sigma}{\lambda^2}\right] = 10\log\left[\frac{4\pi\sigma f^2}{c^2}\right] = 10\log\sigma + 20\log f_1 + K_2 \quad (in\ dB) \qquad \textbf{[12]}$$

$$where: K_2 = 10\log\left[\frac{4\pi}{c^2} \bullet \left(\begin{matrix} Frequency\ and\ RCS \\ conversions\ as\ required \end{matrix} \ \frac{(Hz\ to\ MHz\ or\ GHz)^2}{(meters\ to\ feet)^2}\right)\right]$$

The “Target Gain Factor” (G_σ) is a composite of RCS, frequency, and dimension conversion factors and is called by various names: “Gain of RCS”, “Equivalent Gain of RCS”, “Gain of Target Cross Section”, and in dB form “Gain-sub-Sigma”.

If frequency is given in MHz and RCS (σ) is in m^2, the formula for G_σ is:

$$G_\sigma = 10\log\sigma + 20\log f_1 + 10\log\left[4\pi \bullet \left(\frac{\sec}{3\,x\,10^8\,m}\right)^2 \bullet m^2 \bullet \left(\frac{1x\,10^6}{\sec}\right)^2\right] \qquad \textbf{[13]}$$

or:

$$G_\sigma = 10\log\sigma + 20\log f_1 - 38.54 \quad (in\ dB) \qquad \textbf{[14]}$$

Target gain factor, G_σ = 10log σ + 20log f_1 + K_2 (in dB)

K_2 Values (dB) (units)	RCS (σ) K_2 =	f_1 in MHz K_2 =	f_1 in GHz
m^2	-38.54	21.46	
ft^2	-48.86	11.14	

For this example, the constant K_2 is -38.54 dB. K_2 values for various area and frequency and frequency units are summarized in the adjoining table.

In the two-way radar equation, the one-way free space loss factor (α_1) is used twice, once for the radar transmitter to target path and once for the target to radar receiver path. The radar illustrated in Figure 1 is monostatic so the two path losses are the same and the values of the two α_1’s are the same.

If the transmission loss in Figure 1 from P_t to G_t equals the loss from G_r to P_r, and $G_r = G_t$, then equation [10] can be written as:

$$10\log\,[S \text{ or } P_r] = 10\log P_t + 20\log G_{tr} - 2\alpha_1 + G_\sigma \quad (\text{in dB}) \qquad \textbf{[15]}$$

The space loss factor (α_1) and the target gain factor (G_σ) include all the necessary unit conversions so that they can be used directly with the most common units. Because the factors are given in dB form, they are more convenient to use and allow calculation without a calculator when the factors are read from a chart or nomograph.

Most radars are monostatic. That is, the radar transmitting and receiving antennas are literally the same antenna. There are some radars that are considered “monostatic” but have separate transmitting and receiving antennas that are co-located. In that case, equation [10] could require two different antenna gain factors as originally derived:

$$10\log\,[S \text{ or } P_r] = 10\log P_t + 10\log G_t + 10\log G_r - 2\alpha_1 + G_\sigma \quad (\text{in dB}) \qquad \textbf{[16]}$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

Figure 2 is the visualization of the path losses occurring with the two-way radar equation. **Note:** to avoid having to include additional terms, always combine any transmission line loss with antenna gain. Losses due to antenna polarization and atmospheric absorption also need to be included.

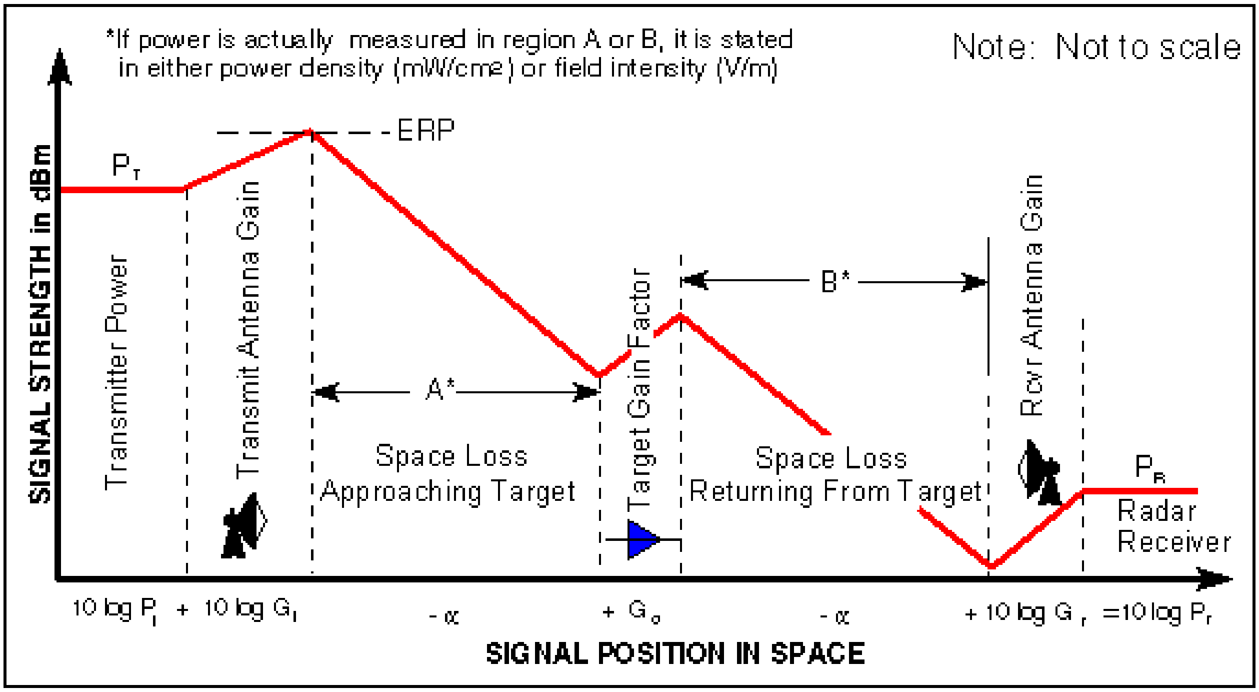
*If power is actually measured in region A or B, it is stated in either power density (mW/cm²) or field intensity (V/m)
Note: Not to scale
ERP
P_T
Transmitter Power
Transmit Antenna Gain
A*
Space Loss Approaching Target
Target Gain Factor
B*
Space Loss Returning From Target
Rcvr Antenna Gain
P_R
Radar Receiver
SIGNAL STRENGTH in dBm
10 log P_t + 10 log G_t
- α
+ G_σ
- α
+ 10 log G_r =10 log P_r
SIGNAL POSITION IN SPACE

Figure 2. Visualization of Two-Way Radar Equation.

RADAR RANGE EQUATION (Two-Way Equation)

The Radar Equation is often called the "Radar Range Equation." The Radar Range Equation is simply the Radar Equation rewritten to solve for maximum Range. The maximum radar range (R_{max}) is the distance beyond which the target can no longer be detected and correctly processed. It occurs when the received echo signal just equals S_{min}.

The Radar Range Equation is then:

$$R_{\max} \cong \left[\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 S_{\min}}\right]^{\frac{1}{4}} \; or \; \left[\frac{P_t G_t G_r c^2 \sigma]}{(4\pi)^3 f^2 S_{\min}}\right]^{\frac{1}{4}} \; or \; \left[\frac{P_t G_t A_e \sigma}{(4\pi)^2 S_{\min}}\right]^{\frac{1}{4}} \qquad \textbf{[17]}$$

The first equation, of the three above, is given in Log form by:

$$40\log R_{max} \cong 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 10\log S_{min} - 20\log f - 30\log 4\pi + 20\log c \qquad \textbf{[18]}$$

As shown previously, Since $K_1 = 20\log [(4\pi/c)$ times conversion units if not in m/sec, m, and Hz], we have:

$$10\log R_{max} \cong \tfrac{1}{4}\,[10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 10\log S_{min} - 20\log f_1 - K_1 - 10.99\text{ dB}] \qquad \textbf{[19]}$$

If you want to convert back from dB, then

$$R_{max} \cong 10^{\frac{M\,dB}{40}}$$

Where M dB is the resulting number within the brackets of equation 19.

One-way free space loss, $\alpha_1 = 20\log(f_1 R) + K_1$ (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

From Section 5-2, Receiver Sensitivity / Noise, S_{min} is related to the noise factors by:

$$S_{min} = (S/N)_{min}(NF)\,kT_0 B \qquad \textbf{[20]}$$

The Radar Range Equation for a tracking radar (target continuously in the antenna beam) becomes:

$$R_{max} \cong \left[\frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 (S/N)_{min}(NF)\,kT_0 B}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t G_r c^2 \sigma}{(4\pi)^3 f^2 (S/N)_{min}(NF)\,kT_o B}\right]^{\frac{1}{4}} \text{ or } \left[\frac{P_t G_t A_e \sigma}{(4\pi)^2 (S/N)_{min}(NF)\,kT_o B}\right]^{\frac{1}{4}} \qquad \textbf{[21]}$$

P_t in equations [17], [19], and [21] is the peak power of a CW or pulse signal. For pulse signals these equations assume the radar pulse is square. If not, there is less power since P_t is actually the average power within the pulse width of the radar signal. Equations [17] and [19] relate the maximum detection range to S_{min}, the minimum signal which can be detected and processed (the receiver sensitivity). The bandwidth (B) in equations [20] and [21] is directly related to S_{min}. B is approximately equal to 1/PW. Thus a wider pulse width means a narrower receiver bandwidth which lowers S_{min}, assuming no integration.

One cannot arbitrarily change the receiver bandwidth, since it has to match the transmitted signal. The "widest pulse width" occurs when the signal approaches a CW signal (see Section 2-11). A CW signal requires a very narrow bandwidth (approximately 100 Hz). Therefore, receiver noise is very low and good sensitivity results (see Section 5-2). If the radar pulse is narrow, the receiver filter bandwidth must be increased for a match (see Section 5-2), i.e. a 1 μs pulse requires a bandwidth of approximately 1 MHz. This increases receiver noise and decreases sensitivity.

If the radar transmitter can increase its PRF (decreasing PRI) and its receiver performs integration over time, an increase in PRF can permit the receiver to "pull" coherent signals out of the noise thus reducing S/N_{min} thereby increasing the detection range. Note that a PRF increase may limit the maximum range due to the creation of overlapping return echoes (see Section 2-10).

There are also other factors that limit the maximum practical detection range. With a scanning radar, there is loss if the receiver integration time exceeds the radar's time on target. Many radars would be range limited by line-of-sight/radar horizon (see Section 2-9) well before a typical target faded below S_{min}. Range can also be reduced by losses due to antenna polarization and atmospheric absorption (see Sections 3-2 and 5-1).

Two-Way Radar Equation (Example)

Assume that a 5 GHz radar has a 70 dBm (10 kilowatt) signal fed through a 5 dB loss transmission line to a transmit/receive antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an RCS of 9 m^2. What is the signal level at the input to the radar receiver? (There is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). This problem continues in Sections 4-3, 4-7, and 4-10.

Answer:
Starting with: $10\log S = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1$ (in dB)

We know that: $\alpha_1 = 20\log fR + K_1 = 20\log(5x31) + 92.44 = 136.25$ dB

and that: $G_\sigma = 10\log \sigma + 20\log f_1 + K_2 = 10\log 9 + 20\log 5 + 21.46 = 44.98$ dB (see Table 1)
(Note: The aircraft transmission line losses (-5 dB) will be combined with the antenna gain (45 dB) for both receive and transmit paths of the radar)

So, substituting in we have: $10\log S = 70 + 40 + 40 + 44.98 - 2(136.25) =$ **-77.52 dBm** @ 5 GHz

The answer changes to -80.44 dBm if the tracking radar operates at 7 GHz provided the antenna gains and the aircraft RCS are the same at both frequencies.

$\alpha_1 = 20\log(7x31) + 92.44 = 139.17$ dB, $G_\sigma = 10\log 9 + 20\log 7 + 21.46 = 47.9$ dB (see Table 1)

$10\log S = 70 + 40 + 40 + 47.9 - 2(139.17) =$ **-80.44 dBm** @ 7 GHz

Table 1. Values of the Target Gain Factor (G_σ) in dB for Various Values of Frequency and RCS.

Frequency (GHz)	RCS - Square meters						
	0.05	5	9	10	100	1,000	10,000
0.5 GHz	2.44	22.42	24.98	25.44	35.44	45.44	55.44
1 GHz	8.46	28.46	31.0	31.46	41.46	51.46	61.46
5 GHz	22.44	42.44	44.98	45.44	55.44	65.44	75.44
7 GHz	25.36	45.36	47.9	48.36	58.36	68.36	78.36
10 GHz	28.46	48.46	51.0	51.46	61.46	71.46	81.46
20 GHz	34.48	54.48	57.02	57.48	67.48	77.48	87.48
40 GHz	40.50	60.48	63.04	63.5	73.5	83.5	93.5

Note: Shaded values were used in the examples.

TWO-WAY RADAR RANGE INCREASE AS A RESULT OF A SENSITIVITY INCREASE

As shown in equation [17] $S_{min}^{-1} \propto R_{max}^4$. Therefore, $-10\log S_{min} \propto 40\log R_{max}$ and the table below results:

% Range Increase: Range + (% Range Increase) x Range = New Range
i.e., for a 12 dB sensitivity increase, 500 miles +100% x 500 miles = 1,000 miles
Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 12 dB sensitivity increase 500 miles x 2 = 1,000 miles

Table 2. Effects of Sensitivity Increase.

dB Sensitivity Increase	% Range Increase	Range Multiplier	dB Sensitivity Increase	% Range Increase	Range Multiplier
+ 0.5	3	1.03	10	78	1.78
1.0	6	1.06	11	88	1.88
1.5	9	1.09	12	100	2.00
2	12	1.12	13	111	2.11
3	19	1.19	14	124	2.24
4	26	1.26	15	137	2.37
5	33	1.33	16	151	2.51
6	41	1.41	17	166	2.66
7	50	1.50	18	182	2.82
8	58	1.58	19	198	2.98
9	68	1.68	20	216	3.16

TWO-WAY RADAR RANGE DECREASE AS A RESULT OF A SENSITIVITY DECREASE

As shown in equation [17] $S_{min}^{-1} \propto R_{max}^4$ Therefore, $-10 \log S_{min} \propto 40 \log R_{max}$ and the table below results:

% Range Decrease: Range - (% Range Decrease) x Range = New Range
i.e., for a 12 dB sensitivity decrease, 500 miles - 50% x 500 miles = 250 miles

Range Multiplier: Range x Range Multiplier = New Range
i.e., for a 12 dB sensitivity decrease 500 miles x 0.5 = 250 miles

Table 3. Effects of Sensitivity Decrease.

dB Sensitivity Decrease	% Range Decrease	Range Multiplier	dB Sensitivity Decrease	% Range Decrease	Range Multiplier
- 0.5	3	0.97	-10	44	0.56
- 1.0	6	0.94	- 11	47	0.53
- 1.5	8	0.92	- 12	50	0.50
- 2	11	0.89	- 13	53	0.47
- 3	16	0.84	- 14	55	0.45
- 4	21	0.79	- 15	58	0.42
- 5	25	0.75	- 16	60	0.40
- 6	29	0.71	- 17	62	0.38
- 7	33	0.67	- 18	65	0.35
- 8	37	0.63	- 19	67	0.33
- 9	40	0.60	- 20	68	0.32

ALTERNATE TWO-WAY RADAR EQUATION

In this section the same radar equation factors are grouped differently to create different constants as is used by some authors.

TWO-WAY RADAR EQUATION (MONOSTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} = \frac{P_t G_t G_{rsigma} c^2 *}{(4\pi)^3 f^2 R^4} \quad (\ Note : \lambda = \frac{c}{f} \ and\ \sigma\ is\ RCS\ ) \qquad \textbf{[1]}$$

* Keep λ or c, σ, and R in the same units. On reducing the above equation to log form we have:

or: $\mathbf{10log\ P_r = 10log\ P_t + 10log\ G_t + 10log\ G_r - \alpha_2}$ (in dB)

Where: $\alpha_2 = 20\log f_1 R^2 - 10\log \sigma + K_3$, and $K_3 = -10\log c^2/(4\pi)^3$

Note:
Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations

K_3 Values: (dB)	Range Units	f_1 in MHz σ in m²	f_1 in GHz σ in m²	f_1 in MHz σ in ft²	f_1 in GHz σ in ft²
	NM	114.15	174.15	124.47	184.47
	km	103.44	163.44	113.76	173.76
	m	-16.56	43.44	-6.24	53.76
	yd	-18.1	41.9	-7.78	52.22
	ft	-37.2	22.8	-26.88	33.12

In the last section, we had the basic radar equation given as equation [6] and it is repeated as equation [1] in the table above.

In Section 4-4, in order to maintain the concept and use of the one-way space loss coefficient, α_1, we didn't cancel like terms which was done to form equation [6] there. Rather, we regrouped the factors of equation [5]. This resulted in two minus α_1 terms and we defined the remaining term as G_σ, which accounted for RCS (see equation [8] & [9]).

Some authors take a different approach, and instead develop an entirely new single factor α_2, which is used instead of the combination of α_1 and G_σ.

If equation [1] is reduced to log form, (and noting that $f = c/\lambda$) it becomes:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r - 20\log (f R^2) + 10\log \sigma + 10\log (c^2/(4\pi)^3) \qquad \textbf{[2]}$$

We now call the last three terms on the right minus α_2 and use it as a single term instead of the two terms α_1 and G_σ. The concept of dealing with one variable factor may be easier although we still need to know the range, frequency, and radar cross section to evaluate α_2. Additionally, we can no longer use a nomograph like we did in computing α_1 and visualize a two-way space loss consisting of two times the one-way space loss, since there are now 3 variables vs. two.

Equation [2] reduces to:

$$\mathbf{10log\ P_r = 10log\ P_t + 10log\ G_t + 10log\ G_r - \alpha_2} \qquad \text{(in dB)} \qquad \textbf{[3]}$$

Where $\alpha_2 = 20\log(f_1R^2) - 10\log\sigma + K_3$ and where f_1 is the MHz or GHz value of frequency and $K_3 = -10\log(c^2/(4\pi)^3) + 20\log$ (conversion for Hz to MHz or GHz)+ 40log (range unit conversions if not in meters) - 20log (RCS conversions for meters to feet)

The values of K_3 are given in the table above.

Comparing equation [3] to equation [10] in Section 4-4, it can be seen that $\alpha_2 = 2\alpha_1 - G_\sigma$.

TWO-WAY RADAR EQUATION (BISTATIC)

The following table contains a summary of the equations developed in this section.

TWO-WAY RADAR EQUATION (BISTATIC)

Peak power at the radar receiver input is:

$$P_r = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R_{Tx}^2\, R_{Rx}^2} = P_t\, G_t\, G_r \left[\frac{\sigma\, c^2}{(4\pi)^3\, f^{\,2}\, R_{Tx}^2\, R_{Rx}^2} \right]^*$$

Note : $\lambda = c/f$ and $\sigma = RCS$

** keep λ or $c, \sigma,$ and R in the same units*

On reducing the above equation to log form we have:

$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f + 20\log c - 30\log 4\pi - 20\log R_{Tx} - 20\log R_{Rx}$

or in simplified terms: $10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx}$ (in dB)

Where α_{Tx} corresponds to transmitter to target loss and α_{Rx} corresponds to target to receiver loss.

Note: Losses due to antenna polarization and atmospheric absorption (Sections 3-2 and 5-1) are not included in these equations.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss, $\alpha_{Tx\text{ or }Rx} = 20\log (f_1 R_{Tx\text{ or }Rx}) + K_1$ (in dB)

K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	Km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

BISTATIC RADAR

There are also true bistatic radars - radars where the transmitter and receiver are in different locations as is depicted in Figure 1. The most commonly encountered bistatic radar application is the semi-active missile. The transmitter is located on, or near, the launch platform (surface or airborne), and the receiver is in the missile which is somewhere between the launch platform and the target.

The transmitting and receiving antennas are not the same and are not in the same location. Because the target-to-radar range is different from the target-to-missile range, the target-to-radar and target-to-missile space losses are different.

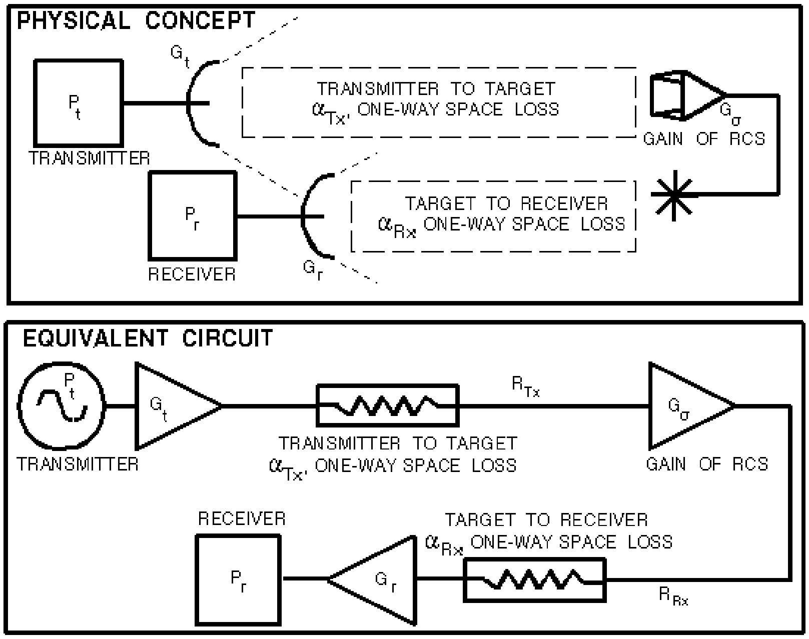
PHYSICAL CONCEPT
G_t
P_t
TRANSMITTER
TRANSMITTER TO TARGET
α_{Tx}, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
P_r
RECEIVER
G_r
TARGET TO RECEIVER
α_{Rx} ONE-WAY SPACE LOSS
EQUIVALENT CIRCUIT
P_t
G_t
TRANSMITTER
R_{Tx}
TRANSMITTER TO TARGET
α_{Tx}, ONE-WAY SPACE LOSS
G_σ
GAIN OF RCS
RECEIVER
P_r
G_r
TARGET TO RECEIVER
α_{Rx} ONE-WAY SPACE LOSS
R_{Rx}

Figure 1. Bistatic Radar Visualized.

The peak power at the radar receiver input is:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R_{Tx}^2 R_{Rx}^2} = P_t G_t G_r \left[\frac{\sigma c^2}{(4\pi)^3 f^2 R_{Tx}^2 R_{Rx}^2} \right] \qquad (\text{Note}: \lambda = \frac{c}{f} \text{ and } \sigma = RCS) \qquad \textbf{[1]}$$

Keep λ or c, σ, and R in the same units.

On reducing the above equation to log form we have:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + 10\log \sigma - 20\log f + 20\log c - 30\log 4\pi - 20\log R_{Tx} - 20\log R_{Rx} \quad \textbf{[2]}$$

or in simplified terms:

$$10\log P_r = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx} \quad \text{(in dB)} \qquad \textbf{[3]}$$

Where α_{Tx} corresponds to transmitter to target loss and α_{Rx} corresponds to target to receiver loss, or:

$\alpha_{Tx} = 20\log(f_1 T_{Tx}) + K_1$ (in dB) and $\alpha_{Rx} = 20\log(f_1 T_{Rx}) + K_1$ (in dB)

with K_1 values provided on page 4-6.1 and with f_1 being the MHz or GHz value of frequency.

Therefore, the difference between monostatic and bistatic calculations is that two α's are calculated for two different ranges and different gains may be required for transmit and receive antennas.

To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain.

As shown in Figure 2, it should also be noted that the bistatic RCS received by the missile is not always the same as the monostatic RCS. In general, the target's RCS varies with angle. Therefore, the bistatic RCS and monostatic RCS will be equal for receive and transmit antennas at the same angle to the target (but only if all three are in a line, as RCS also varies with elevation angle).

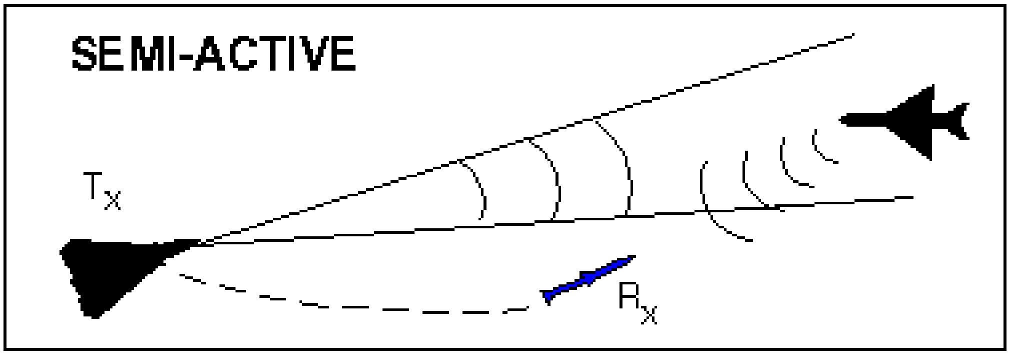

Figure 2. Bistatic RCS Varies.

JAMMING TO SIGNAL (J/S) RATIO - CONSTANT POWER [SATURATED] JAMMING

The following table contains a summary of the equations developed in this section.

JAMMING TO SIGNAL (J/S) RATIO (MONOSTATIC)

$J/S = (P_j G_{ja} 4\pi R^2) / (P_t G_t \sigma)$ (ratio form)* or:

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log\sigma^* + 10.99\text{ dB} + 20\log R^*$

Note (1): Neither f nor λ terms are part of these equations

If simplified radar equations developed in previous sections are used:

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1$ (in dB)

Note (2): the $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1

JAMMING TO SIGNAL (J/S) RATIO (BISTATIC)

R_{Tx} is the range from the radar transmitter to the target. See note (1).

$J/S = (P_j G_{ja} 4\pi R_{Tx}^2) / (P_t G_t \sigma)$ (ratio form) * or:

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log\sigma^* + 10.99\text{ dB} + 20\log R_{Tx}^*$

If simplified radar equations developed in previous sections are used: see note (2).

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_{Tx}$ (in dB)

* Keep R and σ in same units

Target gain factor, (in dB)

$G_\sigma = 10\log\sigma + 20\log f_1 + K_2$

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m2	-38.54	21.46
ft2	-48.86	11.14

One-way free space loss (dB)

α_1 or α_{Tx} =
$20\log(f_1 R) + K_1$

K_1 Values (dB):

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
ft	-37.87	22.13

This section derives the J/S ratio from the one-way range equation for J and the two-way range equation for S, and deals exclusively with active (transmitting) Electronic Attack (EA) devices or systems. Furthermore, the only purpose of EA is to prevent, delay, or confuse the radar processing of target information.

By official definition, EA can be either Jamming or Deception. This may be somewhat confusing because almost any type of active EA is commonly called "jamming," and the calculations of EA signal in the radar compared to the target signal in the radar commonly refer to the "jamming-to-signal" ratio ("J-to-S" ratio). Therefore this section uses the common jargon and the term "jammer" refers to any EA transmitter, and the term "jamming" refers to any EA transmission, whether Deception or Concealment.

Jamming: "Official" jamming should more aptly be called Concealment or Masking. Essentially, Concealment uses electronic transmissions to swamp the radar receiver and hide the targets. Concealment (Jamming) usually uses some form of noise as the transmitted signal. In this section, Concealment will be called "noise" or "noise jamming."

Deception: Deception might be better called Forgery. Deception uses electronic transmissions to forge false target signals that the radar receiver accepts and processes as real targets.

"J" designates the EA signal strength whether it originates from a noise jammer or from a deception system.

Basically, there are two different methods of employing active EA against hostile radars:

Self Protection EA
Support EA

For most practical purposes, Self Protection EEA is usually Deception and Support EA is usually noise jamming. As the name implies, Self Protection EA is EA that is used to protect the platform that it is on. Self Protection EA is often called "self screening jamming," "Defensive EA" or historically "Deception ECM." The top half of Figure 1 shows self-screening jamming.

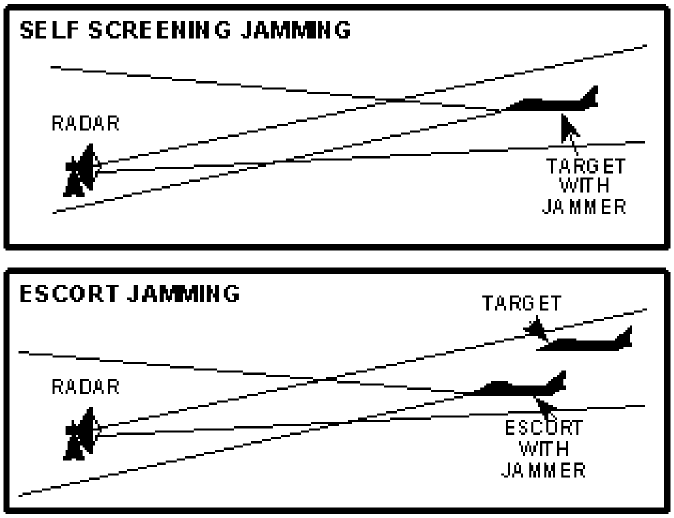

Figure 1. Self Protection and Escort Jamming.

The bottom half of Figure 1 illustrates escort jamming which is a special case of support jamming. If the escort platform is sufficiently close to the target, the J-to-S calculations are the same as for self protection EA.

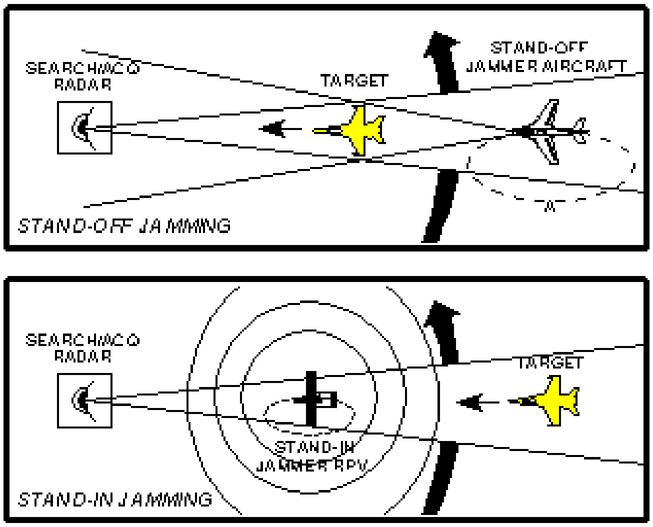

Figure 2. Support Jamming.

Support EA is electronic transmissions radiated from one platform and is used to protect other platforms or fulfill other mission requirements, like distraction or conditioning. Figure 2 illustrates two cases of support jamming protecting a striker - stand-off jamming (SOJ) and stand-in jamming (SIJ). For SOJ the support jamming platform is maintaining an orbit at a long range from the radar - usually beyond weapons range. For SIJ, a remotely piloted vehicle is orbiting very close to the victim radar. Obviously, the jamming power required for the SOJ to screen a target is much greater than the jamming power required for the SIJ to screen the same target.

When factoring EA into the radar equation, the quantities of greatest interest are "J-to-S" and Burn-Through Range.

"J-to-S" is the ratio of the signal strength of the jammer signal (J) to the signal strength of the target return signal (S). It is expressed as "J/S" and, in this section, is always in dB. J usually (but not always) must exceed S by some amount to be effective, therefore the desired result of a J/S calculation in dB is a positive number. Burn-through Range is the radar to target range where the target return signal can first be detected through the jamming and is usually slightly farther than crossover range where J=S. It is usually the range where the J/S just equals the minimum effective J/S (See Section 4-8).

The significance of "J-to-S" is sometimes misunderstood. The effectiveness of EA **is not** a direct mathematical function of "J-to-S." The magnitude of the "J-to-S" required for effectiveness is a function of the particular EA technique and of the radar it is being used against. Different EA techniques may very well require different "J-to-S" ratios against the same radar. When there is sufficient "J-to-S" for effectiveness, increasing it will rarely increase the effectiveness at a given range. Because modern radars can have sophisticated signal processing and/or EP capabilities, in certain radars too much "J-to-S" could cause the signal processor to ignore the jamming, or activate special anti-jamming modes. Increasing "J-to-S" (or the jammer power) does, however, allow the target aircraft to get much closer to the threat radar before burn-through occurs, which essentially means more power is better if it can be controlled when desired.

IMPORTANT NOTE: If the signal S is CW or PD and the Jamming J is amplitude modulated, then the J used in the formula has to be reduced from the peak value (due to sin x/x frequency distribution). The amount of reduction is dependent upon how much of the bandwidth is covered by the jamming signal. To get an exact value, integrals would have to be taken over the bandwidth. As a rule of thumb however:

- If the frequency of modulation is less than the BW of the tracking radar reduce J/S by 10 Log (duty cycle).
- If the frequency of modulation is greater than the BW of the tracking radar reduce J/S by 20 Log(duty cycle).

For example; if your jamming signal is square wave chopped (50% duty cycle) at a 100 Hz rate while jamming a 1 kHz bandwidth receiver, then the J/S is reduced by 3 dB from the maximum. If the duty cycle was 33%, then the reduction would be 4.8 dB. If the 50% and 33% duty cycle jamming signals were chopped at a 10 kHz (vice the 100 Hz) rate, the rule of thumb for jamming seen by the receiver would be down 6 dB and 9.6 dB, respectively, from the maximum since the 10 kHz chopping rate is greater than the 1 kHz receiver BW.

J/S for SELF PROTECTION EA vs. MONOSTATIC RADAR

Figure 3 is radar jamming visualized. The Physical concept of Figure 3 shows a monostatic radar that is the same as Figure 1, Section 4-4, and a jammer (transmitter) to radar (receiver) that is the same as Figure 3, Section 4-3. In other words, Figure 3 is simply the combination of the previous two visual concepts where there is only one receiver (the radar's).

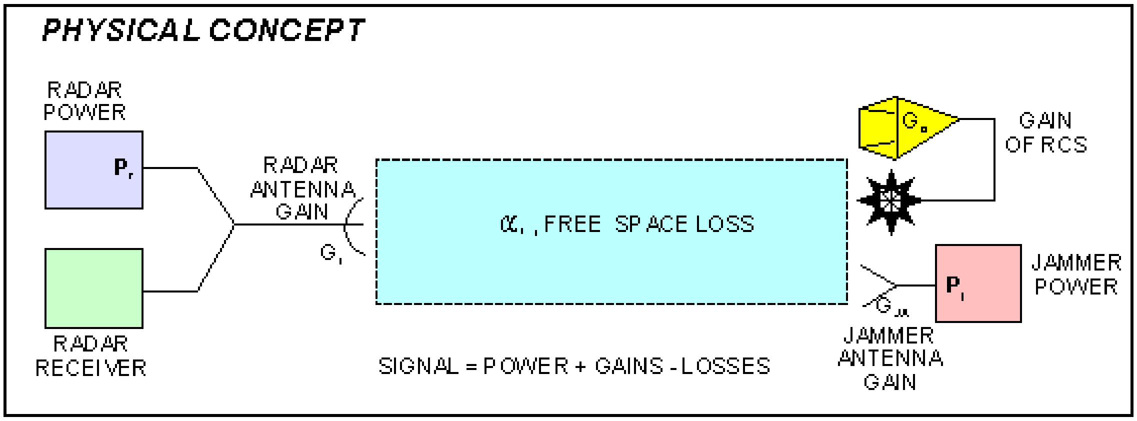

Figure 3. Radar Jamming Visualized.

The equivalent circuit shown in Figure 4 applies to jamming monostatic radars with either self protect EA or support EA. For self protect (or escort) vs. a monostatic radar, the jammer is on the target and the radar receive and transmit antennas are collocated so the three ranges and three space loss factors (α's) are the same.

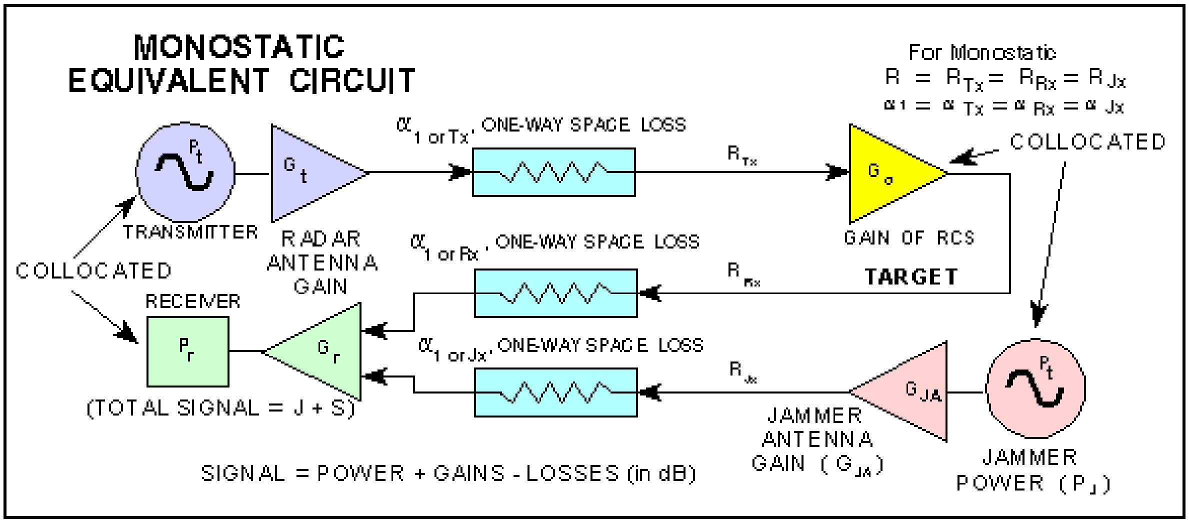
MONOSTATIC EQUIVALENT CIRCUIT
For Monostatic
R = R_Tx = R_Rx = R_Jx
α1 = α Tx = α Rx = α Jx
P_t
G_t
α1 or Tx, ONE-WAY SPACE LOSS
R_Tx
G_σ
COLLOCATED
TRANSMITTER
RADAR ANTENNA GAIN
α1 or Rx, ONE-WAY SPACE LOSS
R_Rx
GAIN OF RCS
TARGET
COLLOCATED
RECEIVER
P_r
G_r
α1 or Jx, ONE-WAY SPACE LOSS
R_Jx
G_JA
P_t
(TOTAL SIGNAL = J + S)
JAMMER ANTENNA GAIN (G_JA)
JAMMER POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 4. Monostatic Radar EA Equivalent Circuit.

J-S Ratio (Monostatic) - The ratio of the power received (P_{r1} or J) from the jamming signal transmitted from the target to the power received (P_{r2} or S) from the radar skin return from the target equals J/S.

From the one way range equation in Section 4-3: $P_r 1 \text{ or } J = \frac{P_j G_{ja} G_r \lambda^2}{(4\pi R)^2}$ **[1]**

From the two way range equation in Section 4.4: $P_r 2 \text{ or } S = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}$ **[2]**

so $$\frac{J}{S} = \frac{P_j G_{ja} G_r \lambda^2 (4\pi)^3 R^4}{P_t G_t G_r \lambda^2 \sigma (4\pi R)^2} = \frac{P_j G_{ja} 4\pi R^2}{P_t G_t \sigma}^{*} \qquad \textit{(ratio form)}$$ **[3]**

* Keep R and σ in the same units.

On reducing the above equation to log form we have:

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10\log 4\pi + 20\log R$ **[4]**

or $10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99 \text{ dB} + 20\log R$ **[5]**

Note: Neither *f* nor λ terms are part of the final form of equation [3] and equation [5].

J/S Calculations (Monostatic) Using a One Way Free Space Loss - The simplified radar equations developed in previous sections can be used to express J/S.

From the one way range equation Section 4-3:

$$10\log (P_{r1} \text{ or } J) = 10\log P_j + 10\log G_{ja} + 10\log G_r - \alpha_1 \quad \text{(in dB)} \qquad \textbf{[6]}$$

From the two way range equation in Section 4.4:

$$10\log (P_{r2} \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \quad \text{(in dB)} \qquad \textbf{[7]}$$

$$10\log (J/S) = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1 \quad \text{(in dB)} \qquad \textbf{[8]}$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. The $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values (dB) RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

One-way free space loss, $\alpha_1 = 20\log (f_1 R) + K_1$ (in dB)

K_1 Values (dB) Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

J/S for SELF PROTECTION EA vs. BISTATIC RADAR

The semi-active missile illustrated in Figure 5 is the typical bistatic radar which would require the target to have self protection EA to survive. In this case, the jammer is on the target and the target to missile receiver range is the same as the jammer to receiver range, but the radar to target range is different. Therefore, only two of the ranges and two of the α's (Figure 6.) are the same.

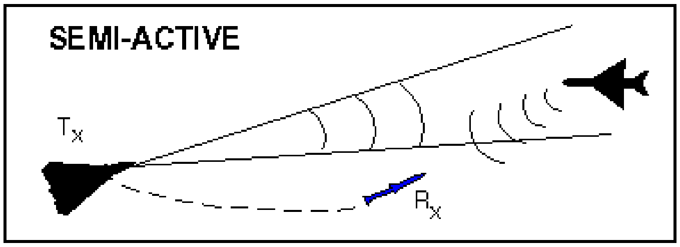

Figure 5. Bistatic Radar.

In the following equations:

α_{Tx} = The one-way space loss from the radar transmitter to the target for range R_{Tx}
α_{Rx} = The one-way space loss from the target to the missile receiver for range R_{Rx}

Like the monostatic radar, the bistatic jamming and reflected target signals travel the same path from the target and enter the receiver (missile in this case) via the same antenna. In both monostatic and bistatic J/S equations this common range cancels, so both J/S equations are left with an R_{Tx}^2 or 20 log R_{Tx} term. Since in the monostatic case $R_{Tx} = R_{Rx}$ and $\alpha_{Tx} = \alpha_{Rx}$, only R or α_1 is used in the equations. Therefore, the bistatic J/S equations [11], [13], or [14] will work for monostatic J/S calculations, but the opposite is only true if bistatic R_{Tx} and α_{Tx} terms are used for R or α_1 terms in monostatic equations [3], [5], and [8].

The equivalent circuit shown in Figure 6 applies to jamming bistatic radar. For self protect (or escort) vs. a bistatic radar, the jammer is on the target and the radar receive and transmit antennas are at separate locations so only two of the three ranges and two of the three space loss factors (α's) are the same.

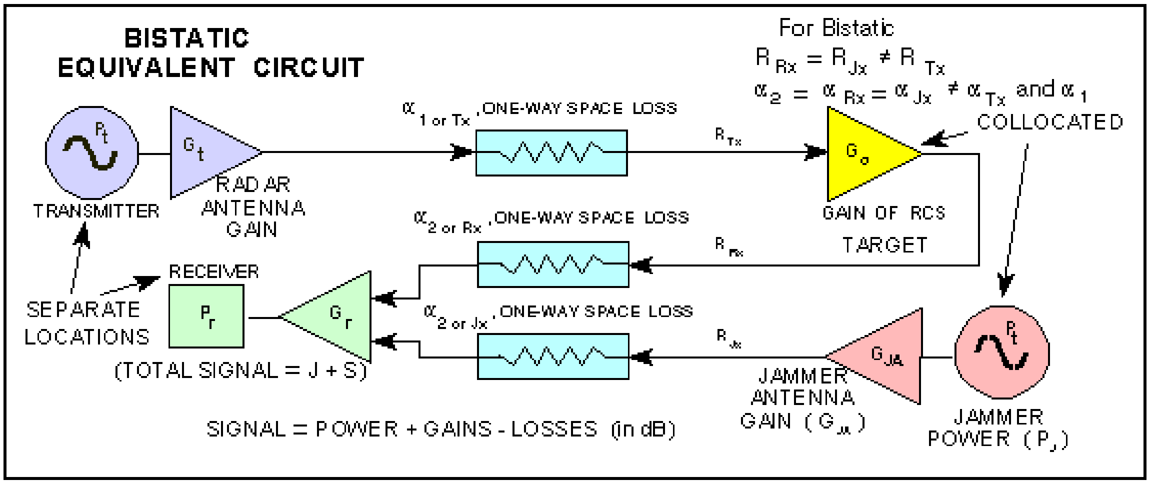
BISTATIC EQUIVALENT CIRCUIT
For Bistatic
$R_{Rx} = R_{Jx} \neq R_{Tx}$
$\alpha_2 = \alpha_{Rx} = \alpha_{Jx} \neq \alpha_{Tx}$ and α_1
COLLOCATED
P_t
G_t
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
G_σ
TRANSMITTER
RADAR ANTENNA GAIN
$\alpha_{2 \text{ or } Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
GAIN OF RCS
TARGET
RECEIVER
SEPARATE LOCATIONS
P_r
G_r
$\alpha_{2 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
P_t
(TOTAL SIGNAL = J + S)
JAMMER ANTENNA GAIN (G_{JA})
JAMMER POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 6. Bistatic Radar EA Equivalent Circuit.

J-to-S Ratio (Bistatic) When the radar's transmit antenna is located remotely from the receiving antenna (Figure 6), the ratio of the power received (P_{r1} or J) from the jamming signal transmitted from the target to the power received (P_{r2} or S) from the radar skin return from the target equals J/S. For jammer effectiveness J normally has to be greater than S.

From the one way range equation in Section 4-3: $$P_r 1 \text{ or } J = \frac{P_j G_{ja} G_r \lambda^2}{(4\pi R_{Rx})^2} \qquad (R_{Jx} = R_{Rx}) \qquad \textbf{[9]}$$

From the two way range equation in Section 4.4: $$P_r 2 \text{ or } S = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R_{Tx}^2 R_{Rx}^2} \qquad \textbf{[10]}$$

so $$\frac{J}{S} = \frac{P_j G_{ja} G_r \lambda^2 (4\pi)^3 R_{Tx}^2 R_{Rx}^2}{P_t G_t G_r \lambda^2 \sigma (4\pi R_{Rx})^2} = \frac{P_j G_{ja} 4\pi R_{Tx}^2}{P_t G_t \sigma}^{*} \qquad \textit{(ratio form)} \qquad \textbf{[11]}$$

* Keep R and σ in the same units.

On reducing the above equation to log form we have:

$$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10\log 4\pi + 20\log R_{Tx} \qquad \textbf{[12]}$$

or $$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99 \text{ dB} + 20\log R_{Tx} \qquad \textbf{[13]}$$

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. Neither *f* nor λ terms are part of the final form of equation [11] and equation [13].

Bistatic J/S Calculations (Bistatic) Using a One Way Free Space Loss - The simplified radar equations developed in previous sections can be used to express J/S.

From the one way range equation in Section 4-3:

$10\log (P_{r1} \text{ or } J) = 10\log P_j + 10\log G_{ja} + 10\log G_r - \alpha_{Rx}$ (all factors dB) **[14]**

From the two way range equation in Section 4-4:
$10\log (P_{r2} \text{ or } S) = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - \alpha_{Tx} - \alpha_{Rx}$ (all factors dB) **[15]**

$10\log (J/S) = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_{Tx}$ (all factors dB) **[16]**

Note: To avoid having to include additional terms for these calculations, always combine any transmission line loss with antenna gain. The $20\log f_1$ term in $-G_\sigma$ cancels the $20\log f_1$ term in α_1.

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)			
K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

One-way free space loss $\alpha_{Tx\text{ or }Rx} = 20\log f_1 R_{Tx\text{ or }Rx} + K_1$ (in dB)			
K_1 Values (dB)	Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
	NM	37.8	97.8
	km	32.45	92.45
	m	-27.55	32.45
	yd	-28.33	31.67
	ft	-37.87	22.13

Saturated J/S (Monostatic) Example (Constant Power Jamming)

Assume that a 5 GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft is flying 31 km from the radar. The aft EW antenna has -1 dB gain and a 5 dB line loss to the EW receiver (there is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). The aircraft has a jammer that provides 30 dBm saturated output if the received signal is above -35 dBm. The jammer feeds a 10 dB loss transmission line which is connected to an antenna with a 5 dB gain. If the RCS of the aircraft is 9 m^2, what is the J/S level received by the tracking radar?

Answer: The received signal at the jammer is the same as the example in Section 4-3, i.e. answer (1) = -32.3 dBm @ 5 GHz. Since the received signal is above -35 dBm, the jammer will operate in the saturated mode, and equation [5] can be used. (See Section 4-10 for an example of a jammer in the linear region.)

$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99 \text{ dB} + 20\log R$

Note: the respective transmission line losses will be combined with antenna gains,
i.e. $-5 + 45 = 40$ dB & $-10 + 5 = -5$ dB.

$10\log J/S = 30 - 5 - 70 - 40 - 9.54 + 10.99 + 89.8 = 6.25$ dB @ 5 GHz*

* The answer is still 6.25 dB if the tracking radar operates at 7 GHz provided the antenna gains and the aircraft RCS are the same at both frequencies.

In this example, there is inadequate jamming power at each frequency if the J/S needs to be 10 dB or greater to be effective. One solution would be to replace the jammer with one that has a greater power output. If the antenna of the aircraft and the radar are not the proper polarization, additional power will also be required (see Section 3-2).

BURN-THROUGH / CROSSOVER RANGE

The burn-through equations are derived in this section. These equations are most commonly used in jammer type of applications. The following is a summary of the important equations explored in this section:

J/S CROSSOVER RANGE (MONOSTATIC) (J = S)

$$R_{J=S} = [\ (P_t\ G_t\ \sigma)\ /\ (P_j\ G_{ja}\ 4\pi)\ ]^{1/2} \qquad \text{(dB Ratio)}$$

or $20\log R_{J=S} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99$ dB

If simplified radar equations already converted to dB are used:

$20\log R_{J=S} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1$ (dB)

* Keep P_t & P_j in same units
Keep R and σ in same units

K_1 Values (dB):

Range (units)	f_1 in MHz K_1 =	in GHz K_1 =
m	-27.55	32.45
ft	-37.87	22.13

BURN-THROUGH RANGE (MONOSTATIC)

The radar to target range where the target return signal (S) can first be detected through the EA (J).

$$R_{BT} = [\ (P_t\ G_t\ \sigma\ J_{\text{min eff}})\ /\ (P_j\ G_{ja}\ 4\pi\ S)\ ]^{1/2} \qquad \text{(dB Ratio)}$$

or $20\log R_{BT} = 10\log P_t + 10\log G_t + 10\log\sigma - 10\log P_j - 10\log G_{ja} + 10\log(J_{\text{min eff}}/S) - 10.99$

If simplified radar equations already converted to dB are used:

$20\log R_{BT} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log(J_{\text{min eff}}/S) - 20\log f_1$(in dB)*

f_1 is MHz or GHz value of frequency

Target gain factor (dB)

$G_\sigma = 10\log\sigma + 20\log f_1 + K_2$

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

BURN-THROUGH RANGE (BISTATIC)

R_{Tx} is the range from the radar transmitter to the target and is different from R_{Rx} which is the range from the target to the receiver. Use Monostatic equations and substitute R_{Tx} for R

CROSSOVER RANGE and BURN-THROUGH RANGE

To present the values of J and S, (or J/S) over a minimum to maximum radar to target range of interest, equation [1], Section 4-7. which has a slope of 20 log for J vs. range and equation [2], Section 4-7, which has a slope of 40 log for S vs. range are plotted. When plotted on semi-log graph paper, J and S (or J/S) vs. range are straight lines as illustrated in Figure 1.

Figure 1 is a sample graph - it cannot be used for data.

The crossing of the J and S lines (known as crossover) gives the range where J = S (about 1.29 NM), and shows that shorter ranges will produce target signals greater than the jamming signal.

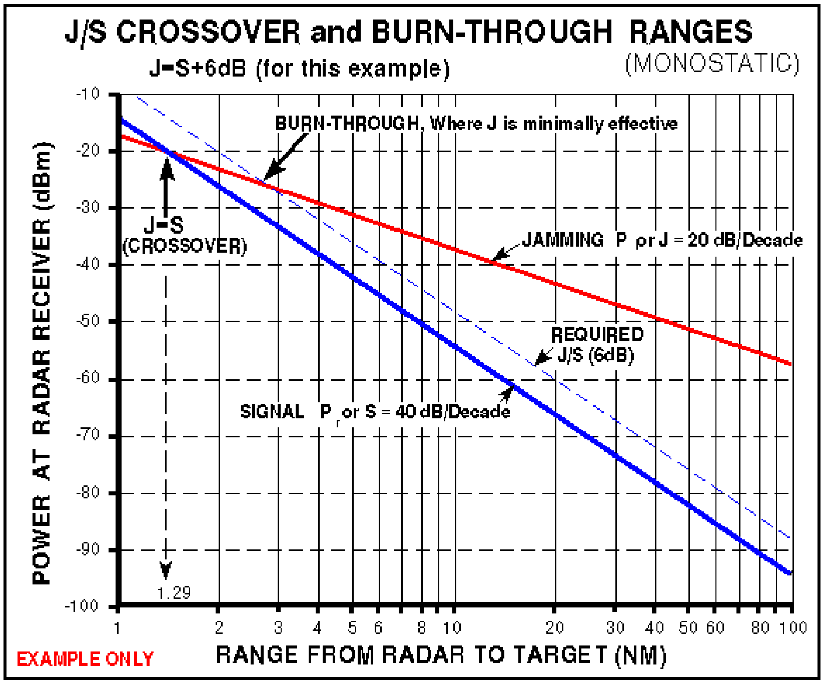

Figure 1. Sample J and S Graph.

The point where the radar power overcomes the jamming signal is known as burn-through. The crossover point where J = S could be the burn-through range, but it usually isn't because normally J/S > 0 dB to be effective due to the task of differentiating the signal from the jamming noise floor (see receiver sensitivity section). For this example, the J/S required for the EA to be effective is given as 6 dB, as shown by the dotted line. This required J/S line crosses the jamming line at about 2.8 NM which, in this example, is the burn-through range.

In this particular example, we have:

P_t = 80 dBm	G_t = 42 dB
P_j = 50 dBm	G_{ja} = 6 dB
σ = 18 m^2	f = 5.9 GHz (not necessary for all calculations)

A radar can be designed with higher than necessary power for earlier burn-through on jamming targets. Naturally that would also have the added advantage of earlier detection of non-jamming targets as well.

Note: To avoid having to include additional terms for the following calculations, always combine any transmission line loss with antenna gain.

CROSSOVER AND BURN-THROUGH RANGE EQUATIONS (MONOSTATIC)

To calculate the crossover range or burn-through range the J/S equation must be solved for range. From equation [3], Section 4-7:

$$\frac{J}{S} = \frac{P_j\, G_{ja}\, 4\pi\, R^2}{P_t\, G_t\, \sigma} \quad \textit{(ratio form)} \qquad \text{Solving for R:} \quad R = \sqrt{\frac{P_t\, G_t\, \sigma\, J}{P_j\, G_{ja}\, 4\pi\, S}} \qquad \textbf{[1]}$$

BURN-THROUGH RANGE (MONOSTATIC) - Burn-through Range (Monostatic) is the radar to target range where the target return signal (S) can first be detected through the jamming (J). It is usually the range when the J/S just equals the minimum effective J/S.

$$R_{BT} = \sqrt{\frac{P_t\, G_t\, \sigma\, J_{\min\, eff}}{P_j\, G_{ja}\, 4\pi\, S}} \quad \text{(burn-through range)} \qquad \textbf{[2]}$$

or in dB form, (using 10log 4π = 10.99 dB):

$$20\log R_{BT} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} + 10\log (J_{\min\, eff}/S) - 10.99 \text{ dB} \qquad \textbf{[3]}$$

RANGE WHEN J/S CROSSOVER OCCURS (MONOSTATIC) - The crossover of the jammer's 20 dB/decade power line and the skin return signal's 40 dB/decade power line of Figure 1 occurs for the case where J = S in dB or J/S=1 in ratio. Substituting into equation [1] yields:

$$R_{(J=S)} = \sqrt{\frac{P_t\, G_t\, \sigma}{P_j\, G_{ja}\, 4\pi}} \quad \text{(Crossover range)} \qquad \textbf{[4]}$$

or in dB form:

$$20\log R_{J=S} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99 \text{ dB} \qquad \textbf{[5]}$$

Note: keep R and σ in same units in all equations.

CROSSOVER AND BURN-THROUGH EQUATIONS (MONOSTATIC)
USING α - ONE WAY FREE SPACE LOSS

The other crossover burn-through range formulas can be confusing because a frequency term is subtracted (equations [6], [7] and [8]), but both ranges are independent of frequency. This subtraction is necessary because when J/S is calculated directly as previously shown, λ^2 or $(c/f)^2$ terms canceled, whereas in the simplified radar equations, a frequency term is part of the G_σ term and has to be cancelled if one solves for R. From equation [8], Section 4-7:

$$10\log J/S = 10\log P_j + 10\log G_{ja} - 10\log P_t - 10\log G_t - G_\sigma + \alpha_1 \quad \text{(factors in dB)}$$

or rearranging: $\alpha_1 = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} + 10\log (J/S)$

from Section 4-4: $\alpha_1 = 20\log f_1 R_1 + K_1$ or $20\log R_1 = \alpha_1 - K_1 - 20\log f_1$

then substituting for α_1:

$$20\log R_1 = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J/S) - 20\log f_1 \quad \text{(in dB)} \qquad \textbf{[6]}$$

EQUATION FOR BURN-THROUGH RANGE (MONOSTATIC) - Burn-through occurs at the range when the J/S just equals the minimum effective J/S. G_σ and K_1 are as defined on page 4-8.1.

$$20\log R_{BT} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J_{min\ eff}/S) - 20\log f_1 \quad \text{(in dB)} \qquad \textbf{[7]}$$

EQUATION FOR THE RANGE WHEN J/S CROSSOVER OCCURS (MONOSTATIC)

The J/S crossover range occurs for the case where J = S, substituting into equation [6] yields:

$$20\log R_{J=S} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1 \quad \text{(factors in dB)} \qquad \textbf{[8]}$$

BURN-THROUGH RANGE (BISTATIC)

Bistatic J/S crossover range is the radar-to-target range when the power received (S) from the radar skin return from the target equals the power received (J) from the jamming signal transmitted from the target. As shown in Figure 6, Section 4-7, the receive antenna that is receiving the same level of J and S is remotely located from the radar's transmit antenna. Bistatic equations [11], [13], and [14] in Section 4-7 show that J/S is only a function of radar to target range, therefore J/S is not a function of wherever the missile is in its flight path provided the missile is in the antenna beam of the target's jammer. The missile is closing on the target at a very much higher rate than the target is closing on the radar, so the radar to target range will change less during the missile flight.

It should be noted that for a very long range air-to-air missile shot, the radar to target range could typically decrease to 35% of the initial firing range during the missile time-of-flight, i.e. A missile shot at a target 36 NM away, may be only 12 NM away from the firing aircraft at missile impact.

Figure 2 shows both the jamming radiated from the target and the power reflected from the target as a function of radar-to-target range. In this particular example, the RCS is assumed to be smaller, 15 m^2 vice 18 m^2 in the monostatic case, since the missile will be approaching the target from a different angle. This will not, however, always be the case.

In this plot, the power reflected is:

$$P_{ref} = \frac{P_t\, G_t\, 4\,\pi\,\sigma}{(4\pi R\,)^2}$$

Substituting the values given previously in the example on page 4-8.1, we find that the crossover point is at 1.18 NM (due to the assumed reduction in RCS).

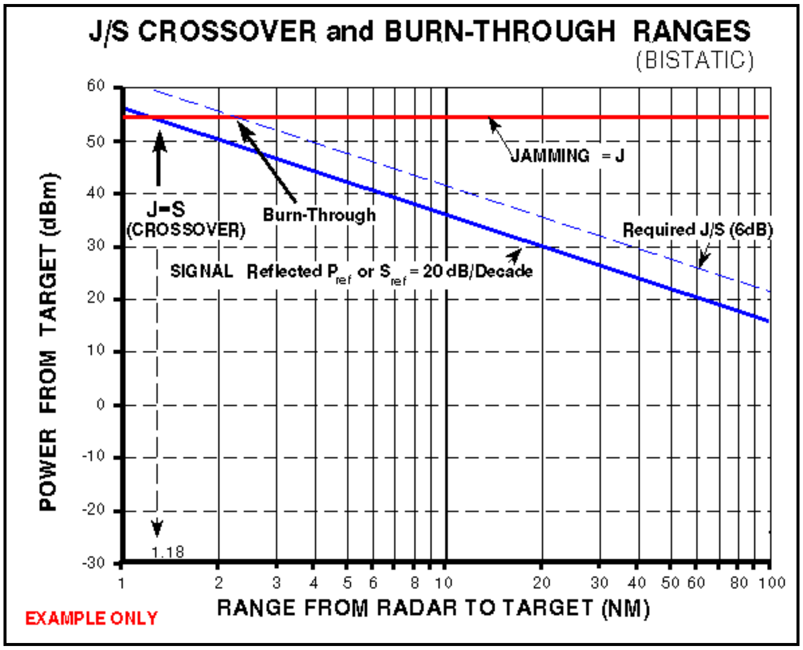

Figure 2. Bistatic Crossover and Burnthrough.

CROSSOVER AND BURN-THROUGH RANGE EQUATIONS (BISTATIC)

To calculate the radar transmitter-to-target range where J/S crossover or burn-through occurs, the J/S equation must be solved for range. From equation [11] in Section 4-7:

$$\frac{J}{S} = \frac{P_j\, G_{ja}\, 4\pi\, R_{Tx}^2}{P_t\, G_t\, \sigma} \qquad \textit{(ratio form)}$$

Solving for R_{Tx}:

$$R_{Tx} = \sqrt{\frac{P_t\, G_t\, \sigma\, J}{P_j\, G_{ja}\, 4\pi\, S}} \qquad \textbf{[9]}$$

Note:
Bistatic equation [10] is identical to monostatic equation [1] except R_{Tx} must be substituted for R and a bistatic RCS (σ) will have to be used since RCS varies with aspect angle. The common explanations will not be repeated in this section.

BURN-THROUGH RANGE (BISTATIC) - Burn-through Range (Bistatic) occurs when J/S just equals the minimum effective J/S. From equation [9]:

$$R_{Tx(BT)} = \sqrt{\frac{P_t\, G_t\, \sigma\, J_{\min eff}}{P_j\, G_{ja}\, 4\pi\, S}} \qquad \textit{(ratio form)} \qquad \textbf{[10]}$$

or in dB form:

$$20\log R_{Tx(BT)} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} + 10\log (J_{min\ eff}/S) - 10.99\ dB \quad \textbf{[11]}$$

If using the simplified radar equations (factors in dB):

$$20\log R_{Tx(BT)} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 + 10\log (J_{min\ eff}/S) - 20\log f_1 \quad \textbf{[12]}$$

Where G_σ and K_1 are defined on page 4-8.1

RANGE WHEN J/S CROSSOVER OCCURS (BISTATIC)

The crossover occurs when J = S in dB or J/S = 1 in ratio.

$$R_{Tx(J=S)} = \sqrt{\frac{P_t\, G_t\, \sigma}{P_j\, G_{ja}\, 4\pi}} \qquad \textit{(ratio)} \quad \textbf{[13]}$$

or in log form:

$$20\log R_{Tx(J=S)} = 10\log P_t + 10\log G_t + 10\log \sigma - 10\log P_j - 10\log G_{ja} - 10.99\ dB \quad \textbf{[14]}$$

If simplified equations are used (with G_σ and K_1 as defined on page 4-8.1) we have:

$$20\log R_{Tx(J=S)} = 10\log P_t + 10\log G_t + G_\sigma - 10\log P_j - 10\log G_{ja} - K_1 - 20\log f_1 \qquad \text{(factors in dB)} \quad \textbf{[15]}$$

Note: keep R and σ in same units in all equations.

DETAILS OF SEMI-ACTIVE MISSILE J/S

Unless you are running a large scale computer simulation that includes maneuvering, antenna patterns, RCS, etc., you will seldom calculate the variation in J/S that occurs during a semi-active missile's flight. Missiles don't fly straight lines at constant velocity. Targets don't either - they maneuver. If the launch platform is an aircraft, it maneuvers too. A missile will accelerate to some maximum velocity above the velocity of the launch platform and then decelerate.

The calculation of the precise variation of J/S during a missile flight for it to be effective requires determination of all the appropriate velocity vectors and ranges at the time of launch, and the accelerations and changes in relative positions during the fly out. In other words, it's too much work for too little return. The following are simplified examples for four types of intercepts.

In these examples, all velocities are constant, and are all along the same straight line. The missile velocity is 800 knots greater than the launch platform velocity which is assumed to be 400 kts. The missile launch occurs at 50 NM.

	J/S (dB)	ΔJ/S (dB)
At Launch:	29	n/a
Intercept Type	At 2 sec. to Intercept:	
AAM Head-on:	23	-6
SAM Incoming Target:	25	-4
AAM Tail Chase:	29	0
SAM Outbound Target:	35	+6

For the AAM tail chase, the range from the radar to the target remains constant and so does the J/S. **In these examples** the maximum variation from launch J/S is ± 6 dB. That represents the difference in the radar to target range closing at very high speed (AAM head on) and the radar to target range opening at moderate speed (SAM outbound target). The values shown above are examples, not rules of thumb, every intercept will be different.

Even for the simplified linear examples shown, graphs of the J and S will be curves - not straight lines. Graphs could be plotted showing J and S vs. radar to target range, or J and S vs. missile to target range, or even J/S vs. time of flight. If the J/S at launch is just barely the minimum required for effectiveness, and increasing it is difficult, then a detailed graph may be warranted, but in most cases this isn't necessary.

This page intentionally left blank.

SUPPORT JAMMING

The following table contains a summary of equations developed in this section:

MAIN LOBE JAMMING TO SIGNAL (J/S) RATIO (For SOJ/SIJ)

$$J/S = (P_j\, G_{ja}\, 4\pi\, R_{Tx}^4) / (P_t\, G_t\, \sigma\, [BW_J/BW_R]\, R_{Jx}^2) \quad \text{(ratio form)*}$$

$$10\log J/S = 10\log P_j - 10\log[BW_J/BW_R] + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + \downarrow$$
$$10.99\text{ dB} + 40\log R_{Tx} - 20\log R_{Jx}\ * \qquad \leftarrow\leftarrow$$

or if simplified radar equations are used:

$$10\log J/S = 10\log P_j - BF + 10\log G_{ja} - \alpha_{jx} - 10\log P_t - 10\log G_t - G_\sigma + 2\alpha_1 \quad \text{(in dB)*}$$

Target gain factor,

$G_\sigma = 10\text{Log}\sigma + 20\text{Log} f_1 + K_2$ (in dB)

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

SIDE LOBE JAMMING TO SIGNAL (J/S) RATIO (For SOJ/SIJ)

$$J/S = (P_j\, G_{ja}\, G_{r(SL)}\, 4\pi\, R_{Tx}^4) / (P_t\, G_t\, G_{r(ML)}\, \sigma\, [BW_J/BW_R]\, R_{Jx}^2) \quad \text{(ratio form)*}$$

$$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} + \rightarrow\downarrow$$
$$10.99\text{ dB} - 10\log \sigma + 40\log R_{Tx} - 20\log R_{Jx}\ * \qquad \leftarrow\leftarrow$$

or if simplified radar equations are used (in dB)*:

$$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{jx} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - G_\sigma + 2\alpha_1$$

One-way free space loss,

α_1 or $\alpha_{Tx} = 20\text{Log}(f_1 R) + K_1$ (in dB)

K_1 Values (dB):

Range (units)	f_1 in MHz K_1 =	f_1 in GHz K_1 =
NM	37.8	97.8
Km	32.45	92.45
m	-27.55	32.45
yd	-28.33	31.67
ft	-37.87	22.13

R_{Jx}	Range from the support jammer transmitter to the radar receiver
R_{Tx}	Range between the radar and the target
BF	10 Log of the ratio of BW_J of the noise jammer to BW_R of the radar receiver
$G_{r(SL)}$	Side lobe antenna gain
$G_{r(ML)}$	Main lobe antenna gain
α_{JX}	One way free space loss between SOJ transmitter and radar receiver
α_1	One way space loss between the radar and the target

* Keep R and σ in same units

Support jamming adds a few geometric complexities. A SOJ platform usually uses high gain, directional antennas. Therefore, the jamming antenna must not only be pointed at the victim radar, but there must be alignment of radar, targets, and SOJ platform for the jamming to be most effective. Two cases will be described, main lobe-jamming and side-lobe jamming.

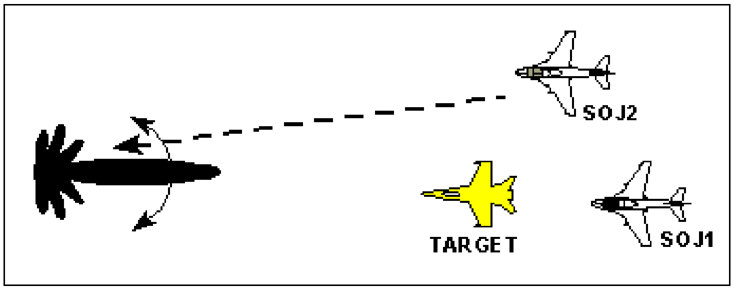

Figure 1. Radar Antenna Pattern.

Support jamming is usually applied against search and acquisition radars which continuously scan horizontally through a volume of space. The scan could cover a sector or a full 360°. The horizontal antenna pattern of the radar will exhibit a main lobe and side lobes as illustrated in Figure 1. The target is detected when the main lobe sweeps across it. For main lobe jamming, the SOJ platform and the target(s) must be aligned with the radar's main lobe as it sweeps the target(s).

For side lobe jamming, the SOJ platform may be aligned with one or more of the radar's side lobes when the main lobe sweeps the target. The gain of a radar's side lobes are many tens of dB less (usually more than 30 dB less) than the gain of the main lobe, so calculations of side lobe jamming must use the gain of the side lobe for the radar receive antenna gain, not the gain of the main lobe. Also, because many modern radars employ some form of side lobe blanking or side lobe cancellation, some knowledge of the victim radar is required to predict the effectiveness of side lobe jamming.

All radar receivers are frequency selective. That is, they are filters that allow only a narrow range of frequencies into the receiver circuitry. Deceptive EA, by definition, creates forgeries of the real signal and, ideally, are as well matched to the radar receiver as the real signal. On the other hand, noise jamming probably **will not match** the radar receiver bandwidth characteristics. Noise jamming is either spot jamming or barrage jamming. As illustrated in Figure 2, spot jamming is simply narrowing the bandwidth of the noise jammer so that as much of the jammer power as possible is in the radar receiver bandwidth. Barrage jamming is using a wide noise bandwidth to cover several radars with one jammer or to compensate for any uncertainty in the radar frequency. In both cases some of the noise power is "wasted" because it is not in the radar receiver filter.

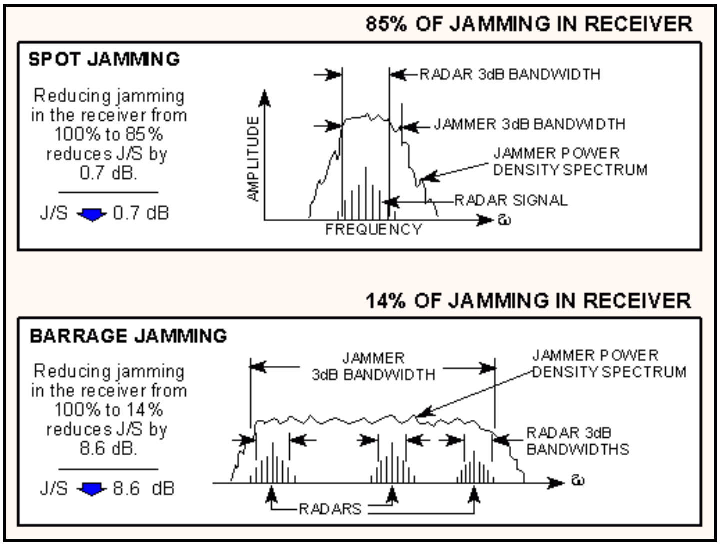
85% OF JAMMING IN RECEIVER
SPOT JAMMING
Reducing jamming in the receiver from 100% to 85% reduces J/S by 0.7 dB.
J/S ▼ 0.7 dB
RADAR 3dB BANDWIDTH
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR SIGNAL
AMPLITUDE
FREQUENCY
ω
14% OF JAMMING IN RECEIVER
BARRAGE JAMMING
Reducing jamming in the receiver from 100% to 14% reduces J/S by 8.6 dB.
J/S ▼ 8.6 dB
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR 3dB BANDWIDTHS
RADARS
ω

Figure 2. Noise Jamming.

In the past, noise jammers were often described as having so many "watts per MHz." This is nothing more than the power of the noise jammer divided by the noise bandwidth. That is, a 500 watt noise jammer transmitting a noise bandwidth of 200 MHz has 2.5 watts/MHz. Older noise jammers often had noise bandwidths that were difficult, or impossible, to adjust accurately. These noise jammers usually used manual tuning to set the center frequency of the noise to the radar frequency. Modern noise jammers can set on the radar frequency quite accurately and the noise bandwidth is selectable, so the noise bandwidth is more a matter of choice than it used to be, and it is possible that all of the noise is placed in the victim radar's receiver.

If, in the example above, the 500 watt noise jammer were used against a radar that had a 3 MHz receiver bandwidth, the noise jammer power applicable to that radar would be:

$$3\ MHz\ x\ 2.5\ watts/MHz = 7.5\ watts\ \_\ 38.75\ dBm \qquad \textbf{[1]}$$

The calculation must be done as shown in equation [1] - multiply the watts/MHz by the radar bandwidth first and then convert to dBm. You can't convert to dBm/MHz and then multiply. (See derivation of dB in Section 2-4)

An alternate method for dB calculations is to use the bandwidth reduction factor (BF).

The BF is: $$BF_{dB} = 10\ \text{Log}\left[\frac{BW_J}{BW_R}\right] \qquad \textbf{[2]}$$

where: BW_J is the bandwidth of the noise jammer, and BW_R is the bandwidth of the radar receiver.

The power of the jammer in the jamming equation (P_J) can be obtained by either method. If equation [1] is used then P_J is simply 38.75 dBm. If equation [2] is used then the jamming equation is written using (P_J - BF). All the following discussion uses the second method. Whichever method is used, it is required that $BW_J \geq BW_R$. If $BW_J < BW_R$, then all the available power is in the radar receiver and equation [1] does not apply and the BF = 0.

> Note: To avoid having to include additional terms for the following calculations, always combine any transmission line loss with antenna gain.

MAIN LOBE STAND-OFF / STAND-IN JAMMING

The equivalent circuit shown in Figure 3 applies to main lobe jamming by a stand-off support aircraft or a stand-in RPV. Since the jammer is not on the target aircraft, only two of the three ranges and two of the three space loss factors (α's) are the same. Figure 3 differs from the J/S monostatic equivalent circuit shown in Figure 4 in Section 4-7 in that the space loss from the jammer to the radar receiver is different.

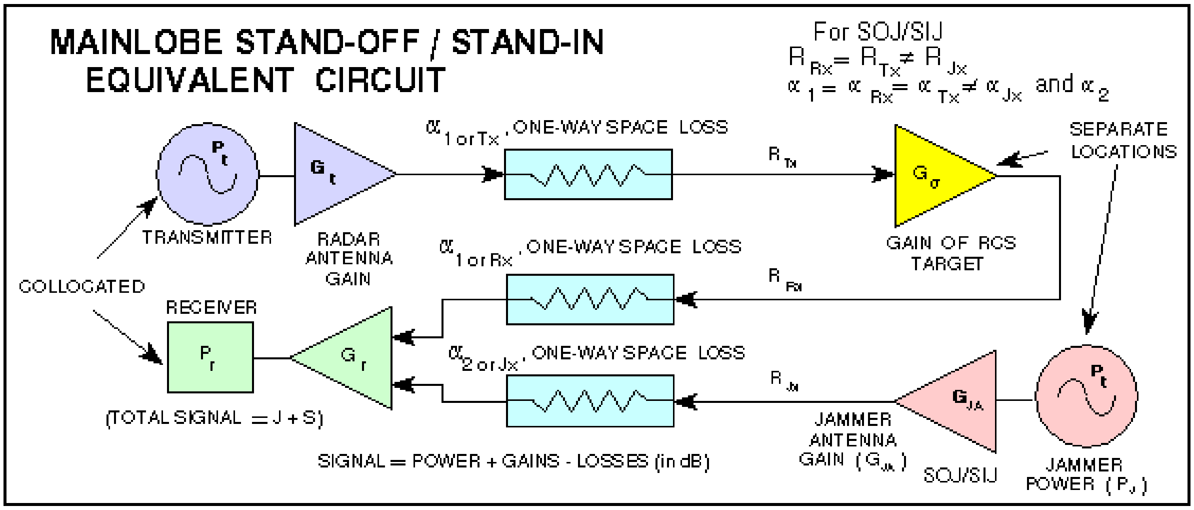
MAINLOBE STAND-OFF / STAND-IN EQUIVALENT CIRCUIT
For SOJ/SIJ
$R_{Rx} = R_{Tx} \neq R_{Jx}$
$\alpha_1 = \alpha_{Rx} = \alpha_{Tx} \neq \alpha_{Jx}$ and α_2
P_t
TRANSMITTER
G_t
RADAR ANTENNA GAIN
$\alpha_{1 \, or \, Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
G_σ
GAIN OF RCS TARGET
SEPARATE LOCATIONS
COLLOCATED
RECEIVER
P_r
G_r
$\alpha_{1 \, or \, Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
$\alpha_{2 \, or \, Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
(TOTAL SIGNAL = J + S)
G_{JA}
JAMMER ANTENNA GAIN (G_{JA})
SOJ/SIJ
P_t
JAMMER POWER (P_J)
SIGNAL = POWER + GAINS - LOSSES (in dB)

Figure 3. Main Lobe Stand-Off / Stand-In EA Equivalent Circuit.

The equations are the same for both SOJ and SIJ. From the one way range equation in Section 4-3, and with inclusion of BF losses:

$$P_r 1 \; or \; J = \frac{P_j \, G_{ja} \, G_r \, \lambda^2 \, BW_R}{(4\pi \, R_{Jx})^2 \, BW_J} \qquad [3]$$

From the two way range equation in Section 4.4: $$P_r 2 \; or \; S = \frac{P_t \, G_t \, G_r \, \lambda^2 \, \sigma}{(4\pi)^3 \, R_{Tx}^4} \qquad [4]$$

so $$\frac{J}{S} = \frac{P_j \, G_{ja} \, G_r \, \lambda^2 (4\pi)^3 \, R_{Tx}^4 \, BW_R}{P_t \, G_t \, G_r \, \lambda^2 \, \sigma (4\pi \, R_{Jx})^2 \, BW_J} = \frac{P_j \, G_{ja} \, 4\pi \, R_{Tx}^4 \, BW_R}{P_t \, G_t \, \sigma \, R_{Jx}^2 \, BW_J} \qquad \textit{(ratio form)} \qquad [5]$$

Note: Keep R and σ in the same units. Converting to dB and using 10 log 4π = 10.99 dB:

$$10\log J/S = 10\log P_j - 10\log [BW_j/BW_R] + 10\log G_{ja} - 10\log P_t - 10\log G_t - 10\log \sigma + 10.99 \text{ dB} + 40\log R_{Tx} - 20\log R_{Jx} \quad \textbf{[6]}$$

If the simplified radar equation is used, the free space loss from the SOJ/SIJ to the radar receiver is α_{Jx}, then equation [7] is the same as monostatic equation [6] in Section 4-7 except α_{Jx} replaces α, and the bandwidth reduction factor [BF] losses are included:

$$10\log J = 10\log P_j - BF + 10\log G_{ja} + 10\log G_r - \alpha_{Jx} \qquad \text{(factors in dB)} \qquad \textbf{[7]}$$

Since the free space loss from the radar to the target and return is the same both ways, $\alpha_{Tx} = \alpha_{Rx} = \alpha_1$, equation [8] is the same as monostatic equation [7] in Section 4-7.

$$10\log S = 10\log P_t + 10\log G_t + 10\log G_r + G_\sigma - 2\alpha_1 \qquad \text{(factors in dB)} \qquad \textbf{[8]}$$

and $$10\log J/S = 10\log P_j - BF + 10\log G_{ja} - \alpha_{Jx} - 10\log P_t - 10\log G_t - G_\sigma + 2\alpha_1 \qquad \text{(factors in dB)} \qquad \textbf{[9]}$$

Notice that unlike equation [8] in Section 4-7, there are two different α's in [9] because the signal paths are different.

SIDE LOBE STAND-OFF / STAND-IN JAMMING

The equivalent circuit shown in Figure 4. It differs from Figure 3, (main lobe SOJ/SIJ) in that the radar receiver antenna gain is different for the radar signal return and the jamming.

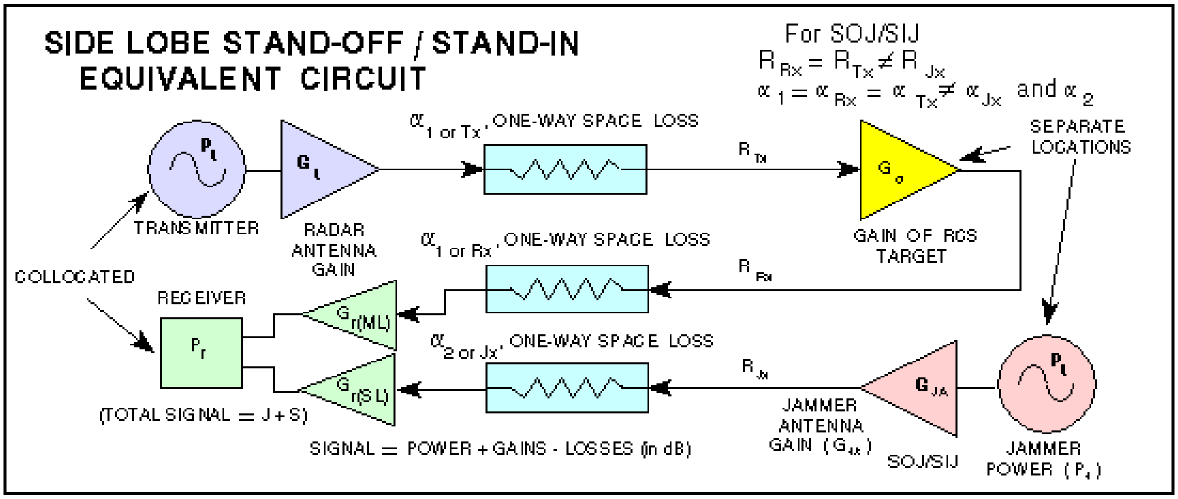
SIDE LOBE STAND-OFF / STAND-IN EQUIVALENT CIRCUIT
For SOJ/SIJ
$R_{Rx} = R_{Tx} \neq R_{Jx}$
$\alpha_1 = \alpha_{Rx} = \alpha_{Tx} \neq \alpha_{Jx}$ and α_2
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
SEPARATE LOCATIONS
P_t
G_t
G_σ
TRANSMITTER
RADAR ANTENNA GAIN
GAIN OF RCS TARGET
$\alpha_{1 \text{ or } Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
COLLOCATED
RECEIVER
P_r
$G_{r(ML)}$
$G_{r(SL)}$
$\alpha_{2 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
P_t
(TOTAL SIGNAL = J + S)
SIGNAL = POWER + GAINS - LOSSES (in dB)
JAMMER ANTENNA GAIN (G_{JA})
SOJ/SIJ
JAMMER POWER (P_J)

Figure 4. Side Lobe Stand-Off / Stand-In EA Equivalent Circuit.

To calculate side lobe jamming, the gain of the radar antenna's side lobes must be known or estimated. The gain of each side lobe will be different than the gain of the other side lobes. If the antenna is symmetrical, the first side lobe is the one on either side of the main lobe, the second side lobe is the next one on either side of the first side lobe, and so on. The side lobe gain is G_{SLn}, where the 'n' subscript denotes side lobe number: 1, 2, ..., n.

The signal is the same as main lobe equations [4] and [8], except $G_r = G_{r(ML)}$

$$P_r 2 \text{ or } S = \frac{P_t\, G_t\, G_{r(ML)}\, \lambda^2\, \sigma}{(4\pi)^3\, R_{Tx}^4} \quad \textit{(ratio form)} \qquad \textbf{[10]}$$

If simplified radar equations are used:

$10\log S = 10\log P_t + 10\log G_t + 10\log G_{r(ML)} + G_\sigma - 2\alpha_1$ (factors in dB)

The jamming equation is the same as main lobe equations [3] and [7] except $G_r = G_{r(SL)}$:

$$J = \frac{P_j\, G_{ja}\, G_{r(SL)}\, \lambda^2\, BW_R}{(4\pi\, R_{Jx})^2\, BW_J} \qquad \textbf{[11]}$$

$10\log J = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{Jx}$ (factors in dB) **[12]**

so
$$\frac{J}{S} = \frac{P_j\, G_{ja}\, G_{r(SL)}\, 4\pi\, R_{Tx}^4\, BW_R}{P_t\, G_t\, G_{r(ML)}\, \sigma\, R_{Jx}^2\, BW_J} \quad \textit{(ratio form)} \qquad \textbf{[13]}$$

Note: keep R and σ in same units. Converting to dB and using $10\log 4\pi = 10.99$ dB:

$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - 10\log\sigma + 10.99 \text{ dB} + 40\log R_{Tx} - 20\log R_{Jx}$
(factors in dB) **[14]**

If simplified radar equations are used:

$10\log J/S = 10\log P_j - BF + 10\log G_{ja} + 10\log G_{r(SL)} - \alpha_{Jx} - 10\log P_t - 10\log G_t - 10\log G_{r(ML)} - G_\sigma + 2\alpha_1$ **[15]**
(in dB)

This page intentionally left blank.

JAMMING TO SIGNAL (J/S) RATIO - CONSTANT GAIN [LINEAR] JAMMING

JAMMING TO SIGNAL (J/S) RATIO (MONOSTATIC)

$$\frac{J}{S} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, c^2}{4\pi\sigma\, f^2} \quad \text{(ratio form)}$$

$G_{ja(Rx)}$ = The Gain of the jammer receive antenna
G_j = The gain of the jammer
$G_{ja(Tx)}$ = The Gain of the jammer transmit antenna

or:

$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - 10\log (4\pi\sigma/\lambda^2)$

or if simplified radar equations developed in previous sections are used:

$$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma \quad \text{(dB)}$$

* Keep λ and σ in same units. Note: λ = c/f

Target gain factor,
$G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (dB)

K_2 Values (dB):

RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
m^2	-38.54	21.46
ft^2	-48.86	11.14

JAMMING TO SIGNAL (J/S) RATIO (BISTATIC)

Same as the monostatic case except G_σ will be different since RCS (σ) varies with aspect angle.

Since the jammer on the target is amplifying the received radar signal before transmitting it back to the radar, both J and S experience the two way range loss. Figure 1 shows that the range for both the signal and constant gain jamming have a slope that is 40 dB per decade. Once the jammer output reaches maximum power, that power is constant and the jamming slope changes to 20 dB per decade since it is only a function of one way space loss and the J/S equations for constant power (saturated) jamming must be used.

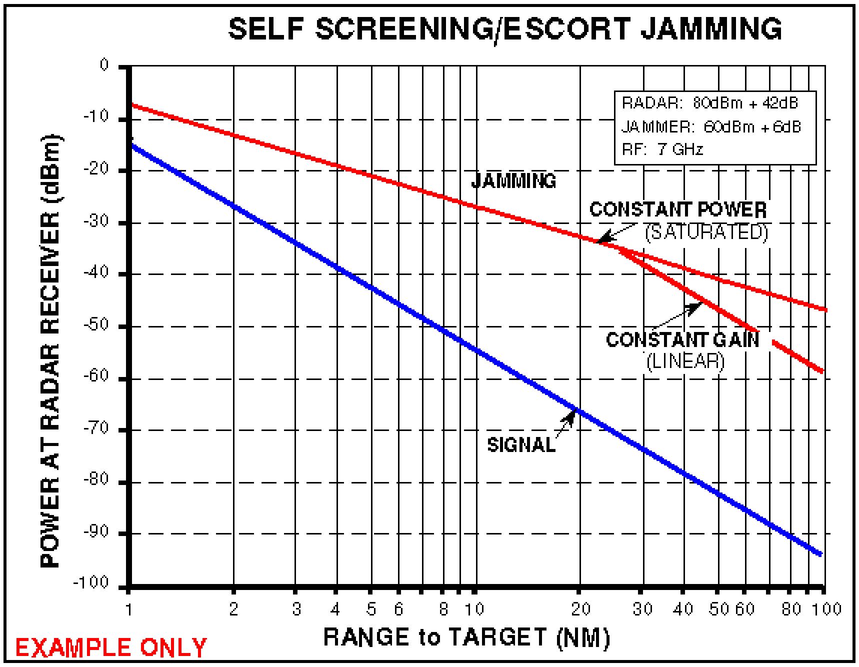

Figure 1. Sample Constant Gain / Constant Power Graph.

Normally the constant gain (linear) region of a repeater jammer occurs only at large distances from the radar and the constant power (saturated) region is reached rapidly as the target approaches the radar. When a constant gain jammer is involved it may be necessary to plot jamming twice - once using J from the constant power (saturated) equation [1] in Section 4-7 and once using the constant gain (linear) equation [4], as in the example shown in Figure 1.

CONSTANT GAIN SELF PROTECTION EA

Most jammers have a constant power output - that is, they always transmit the maximum available power of the transmitter (excepting desired EA modulation). Some jammers also have a constant gain (linear) region. Usually these are coherent repeaters that can amplify a low level radar signal to a power that is below the level that results in maximum available (saturated) power output. At some radar to target range, the input signal is sufficiently high that the full jammer gain would exceed the maximum available power and the jammer ceases to be constant gain and becomes constant power.

To calculate the power output of a constant gain jammer where:

S_{Rj} = The Radar signal at the jammer input (receive antenna terminals)
$G_{ja(Rx)}$ = The Gain of the jammer receive antenna
G_j = The gain of the jammer
α_{Tx} = The one-way free space loss from the radar to the target
P_{jCG} = The jammer constant gain power output
P_j = The maximum jammer power output
L_R = The jammer receiving line loss; combine with antenna gain $G_{ja(Rx)}$

From equation [10], Section 4-3, calculate the radar power received by the jammer.

$$10\log S_{Rj} = 10\log P_t + 10\log G_t - \alpha_{Tx} + 10\log G_{ja(Rx)} \qquad \text{(factors in dB)} \qquad \textbf{[1]}$$

The jammer constant gain power output is: $$10\log P_{jCG} = 10\log S_{Rj} + 10\log G_{ja} \qquad \textbf{[2]}$$

and, by definition: $$P_{jCG} \leq P_j \qquad \textbf{[3]}$$

MONOSTATIC

The equivalent circuit shown in Figure 2 is different from the constant power equivalent circuit in Figure 4 in Section 4-7. With constant gain, the jamming signal experiences the gain of the jammer and its antennas plus the same space loss as the radar signal.

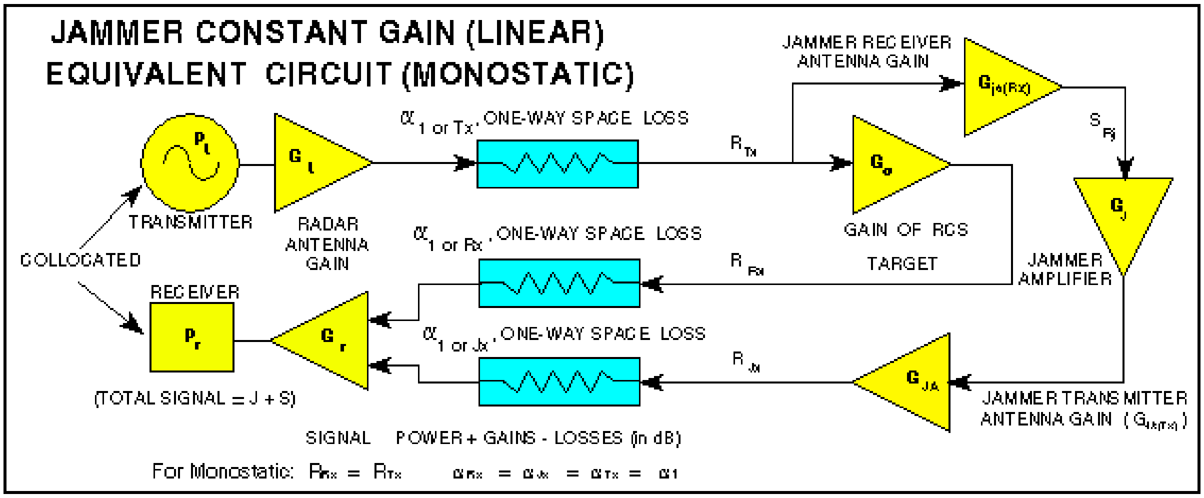
JAMMER CONSTANT GAIN (LINEAR)
EQUIVALENT CIRCUIT (MONOSTATIC)
JAMMER RECEIVER ANTENNA GAIN
$G_{ja(Rx)}$
P_t
G_t
$\alpha_{1 \text{ or } Tx}$, ONE-WAY SPACE LOSS
R_{Tx}
G_σ
S_{Rj}
G_j
TRANSMITTER
RADAR ANTENNA GAIN
$\alpha_{1 \text{ or } Rx}$, ONE-WAY SPACE LOSS
R_{Rx}
GAIN OF RCS
TARGET
JAMMER AMPLIFIER
COLLOCATED
RECEIVER
P_r
G_r
$\alpha_{1 \text{ or } Jx}$, ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
(TOTAL SIGNAL = J + S)
JAMMER TRANSMITTER ANTENNA GAIN ($G_{ja(Tx)}$)
SIGNAL POWER + GAINS - LOSSES (in dB)
For Monostatic: $R_{Rx} = R_{Tx}$ $\alpha_{Rx} = \alpha_{Jx} = \alpha_{Tx} = \alpha_1$

Figure 2. Jammer Constant Gain EA Equivalent Circuit (Monostatic).

To calculate J, the one way range equation from Section 4-3 is used twice:

$$J = \frac{P_t\, G_t\, G_{ja(Rx)}\, \lambda^2}{(4\pi R\,)^2} \frac{G_j\, G_{ja(Tx)}\, G_r\, \lambda^2}{(4\pi R\,)^2} \qquad \textbf{[4]}$$

From the two way range equation in Section 4-4: $S = \dfrac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi\,)^3\, R^4}$ **[5]**

Terms cancel when combined: $\dfrac{J}{S} = \dfrac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma}$ *Keep λ and σ in same units* **[6]**

Or in dB form: $10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - 10\log (4\pi\sigma/\lambda^2)$ **[7]**

Since the last term can be recognized as minus G_σ from equation [10] in Section 4-4, where the target gain factor, $G_\sigma = 10\log (4\pi\sigma/\lambda^2) = 10\log (4\pi\sigma f^2/c^2)$, it follows that:

$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma$ (factors in dB) **[8]**

Target gain factor, $G_\sigma = 10\log \sigma + 20\log f_1 + K_2$ (in dB)

K_2 Values

(dB)	RCS (σ) (units)	f_1 in MHz $K_2 =$	f_1 in GHz $K_2 =$
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

BISTATIC

The bistatic equivalent circuit shown in Figure 3 is different from the monostatic equivalent circuit shown in Figure 2 in that the receiver is separately located from the transmitter, $R_{Tx} \neq R_{Rx}$ or R_{Jx} and G_σ will be different since the RCS (σ) varies with aspect angle.

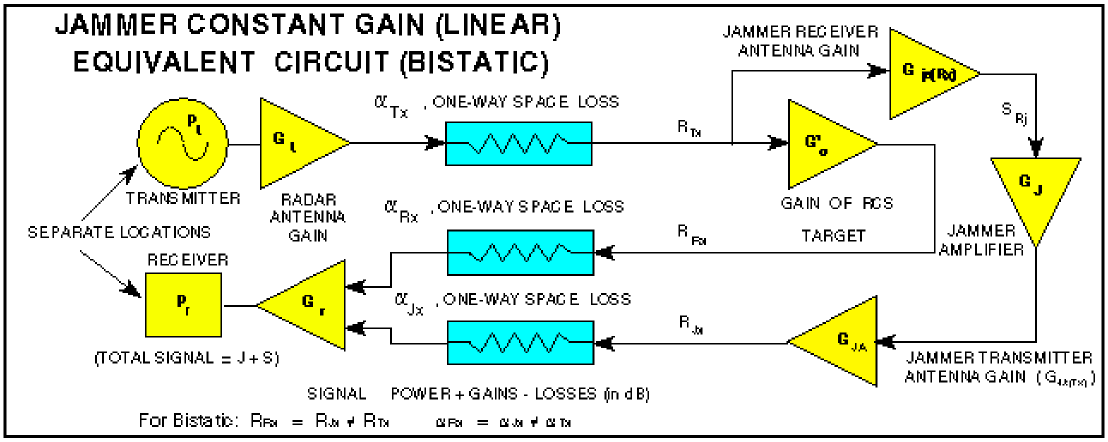
JAMMER CONSTANT GAIN (LINEAR)
EQUIVALENT CIRCUIT (BISTATIC)
JAMMER RECEIVER ANTENNA GAIN
$G_{ja(Rx)}$
α_{Tx} , ONE-WAY SPACE LOSS
P_t
G_t
R_{Tx}
G'_σ
S_{Rj}
G_J
TRANSMITTER
RADAR ANTENNA GAIN
α_{Rx} , ONE-WAY SPACE LOSS
GAIN OF RCS
TARGET
SEPARATE LOCATIONS
R_{Rx}
JAMMER AMPLIFIER
RECEIVER
P_r
G_r
α_{Jx} , ONE-WAY SPACE LOSS
R_{Jx}
G_{JA}
(TOTAL SIGNAL = J + S)
JAMMER TRANSMITTER ANTENNA GAIN ($G_{ja(Tx)}$)
SIGNAL POWER + GAINS - LOSSES (in dB)
For Bistatic: $R_{Rx} = R_{Jx} \neq R_{Tx}$ $\alpha_{Rx} = \alpha_{Jx} \neq \alpha_{Tx}$

Figure 3. Jammer Constant Gain EA Equivalent Circuit (Bistatic).

To calculate J, the one way range equation from Section 4-3 is used twice:

$$J = \frac{P_t\, G_t\, G_{ja(Rx)}\, \lambda^2}{(4\pi\, R_{Tx})^2} \frac{G_j\, G_{ja(Tx)}\, G_r\, \lambda^2}{(4\pi\, R_{Rx})^2} \qquad (R_{Jx} = R_{Rx}) \qquad \textbf{[9]}$$

From the two way range equation in Section 4-4: $S = \frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma'}{(4\pi)^3\, R_{Tx}^2\, R_{Rx}^2}$ (σ' is bistatic RCS) **[10]**

Terms cancel when combined: $\frac{J}{S} = \frac{G_{ja(Rx)}\, G_j\, G_{ja(Tx)}\, \lambda^2}{4\pi\sigma'}$ *Keep λ and σ in same units* **[11]**

Or in dB form: 10log J/S = 10log $G_{ja(Rx)}$ + 10log G_j + 10log $G_{ja(Tx)}$ - 10log $(4\pi\sigma'/\lambda^2)$ **[12]**

Since the last term can be recognized as minus G_σ from equation [10] in Section 4-4, where the target gain factor, G_σ = 10log $(4\pi\sigma'/\lambda^2)$ = 10log $(4\pi\sigma' f^2/c^2)$, it follows that:

10log = 10log $G_{ja(Rx)}$ + 10log G_j + 10log $G_{ja(Tx)}$ - G_σ' (factors in dB) **[13]**

Target gain factor, G_σ = 10log σ + 20log f_1 + K_2 (in dB)

K_2 Values (dB)	RCS (σ) (units)	f_1 in MHz K_2 =	f_1 in GHz K_2 =
	m^2	-38.54	21.46
	ft^2	-48.86	11.14

Linear J/S (Monostatic) Example (Linear Power Jamming)

Assume that a 5 GHz radar has a 70 dBm signal fed through a 5 dB loss transmission line to an antenna that has 45 dB gain. An aircraft that is flying 31 km from the radar has an aft EW antenna with -1 dB gain and a 5 dB line loss to the EW receiver (there is an additional loss due to any antenna polarization mismatch but that loss will not be addressed in this problem). The received signal is fed to a jammer with a gain of 60 dB, feeding a 10 dB loss transmission line which is connected to an antenna with 5 dB gain.

If the RCS of the aircraft is 9 m^2, what is the J/S level received at the input to the receiver of the tracking radar?

Answer:

$10\log J/S = 10\log G_{ja(Rx)} + 10\log G_j + 10\log G_{ja(Tx)} - G_\sigma$

$G_\sigma = 10\log \sigma + 20\log f_1 + K_2 = 10\log 9 + 20\log 5 + 21.46 = 44.98$ dB

Note: The respective transmission line losses will be combined with antenna gains, i.e. -1 -5 = -6 dB and -10 + 5 = -5 dB

$10\log J/S = -6 + 60 - 5 - 44.98 = 4.02$ dB @ 5 GHz

The answer changes to 1.1 dB if the tracking radar operates at 7 GHz provided the antenna gains and aircraft RCS are the same at both 5 and 7 GHz.

$G_\sigma = 10\log 9 + 20\log 7 + 21.46 = 47.9$ dB

$10\log J/S = -6 + 60 - 5 - 47.9 = 1.1$ dB @ 7 GHz

Separate J (-73.5 dBm @ 5 GHz and -79.34 dBm @ 7 GHz) and S (-77.52 dBm @ 5 GHz and -80.44 dBm @ 7 GHz) calculations for this problem are provided in Sections 4-3 and 4-4, respectively. A saturated gain version of this problem is provided in Section 4-7.

This page intentionally left blank.

RADAR CROSS SECTION (RCS)

Radar cross section is the measure of a target's ability to reflect radar signals in the direction of the radar receiver, i.e. it is a measure of the ratio of backscatter power per steradian (unit solid angle) in the direction of the radar (from the target) to the power density that is intercepted by the target.

The RCS of a target can be viewed as a comparison of the strength of the reflected signal from a target to the reflected signal from a perfectly smooth sphere of cross sectional area of 1 m^2 as shown in Figure 1.

The conceptual definition of RCS includes the fact that not all of the radiated energy falls on the target. A target's RCS (σ) is most easily visualized as the product of three factors:

σ = Projected cross section x Reflectivity x Directivity.

RCS(σ) is used in Section 4-4 for an equation representing power reradiated from the target.

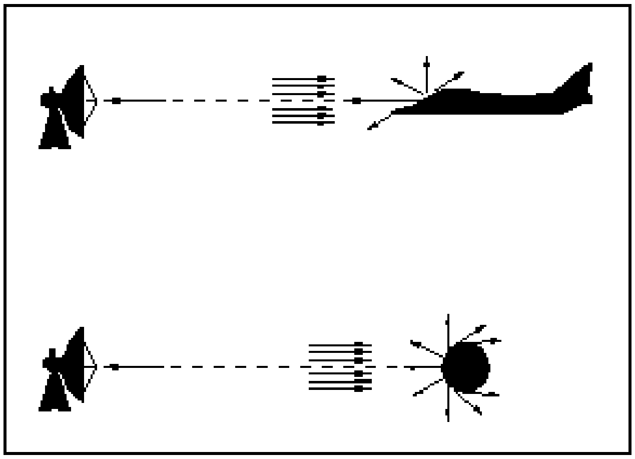

Figure 1. Concept of Radar Cross Section.

Reflectivity: The percent of intercepted power reradiated (scattered) by the target.

Directivity: The ratio of the power scattered back in the radar's direction to the power that would have been backscattered had the scattering been uniform in all directions (i.e. isotropically).

Figures 2 and 3 show that RCS does not equal geometric area. For a sphere, the RCS, **$\sigma = \pi r^2$**, where r is the radius of the sphere.

The RCS of a sphere is independent of frequency if operating at sufficiently high frequencies where λ<<Range, and λ<< radius (r). Experimentally, radar return reflected from a target is compared to the radar return reflected from a sphere which has a frontal or projected area of one square meter (i.e. diameter of about 44 in). Using the spherical shape aids in field or laboratory measurements since orientation or positioning of the sphere will not affect radar reflection intensity measurements as a flat plate would. If calibrated, other sources (cylinder, flat plate, or corner reflector, etc.) could be used for comparative measurements.

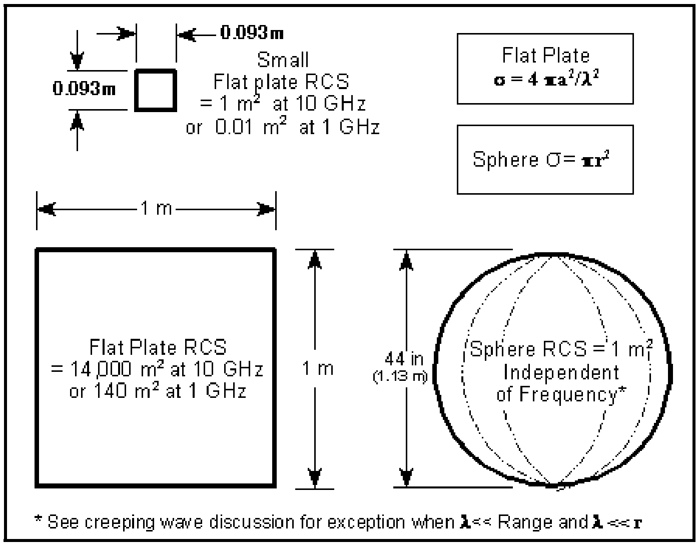
0.093m
0.093m
Small
Flat plate RCS
= 1 m² at 10 GHz
or 0.01 m² at 1 GHz
Flat Plate
σ = 4 πa²/λ²
Sphere σ= πr²
1 m
Flat Plate RCS
= 14,000 m² at 10 GHz
or 140 m² at 1 GHz
1 m
44 in
(1.13 m)
Sphere RCS = 1 m²
Independent
of Frequency*
* See creeping wave discussion for exception when λ<< Range and λ << r

Figure 2. RCS vs. Physical Geometry.

To reduce drag during tests, towed spheres of 6", 14", or 22" diameter may be used instead of the larger 44" sphere, and the reference size is 0.018, 0.099, or 0.245 m^2 respectively instead of 1 m^2. When smaller sized spheres are used for tests you may be operating at or near where λ~radius. If the results are then scaled to a 1 m^2 reference, there may be some perturbations due to creeping waves. See the discussion at the end of this section for further details.

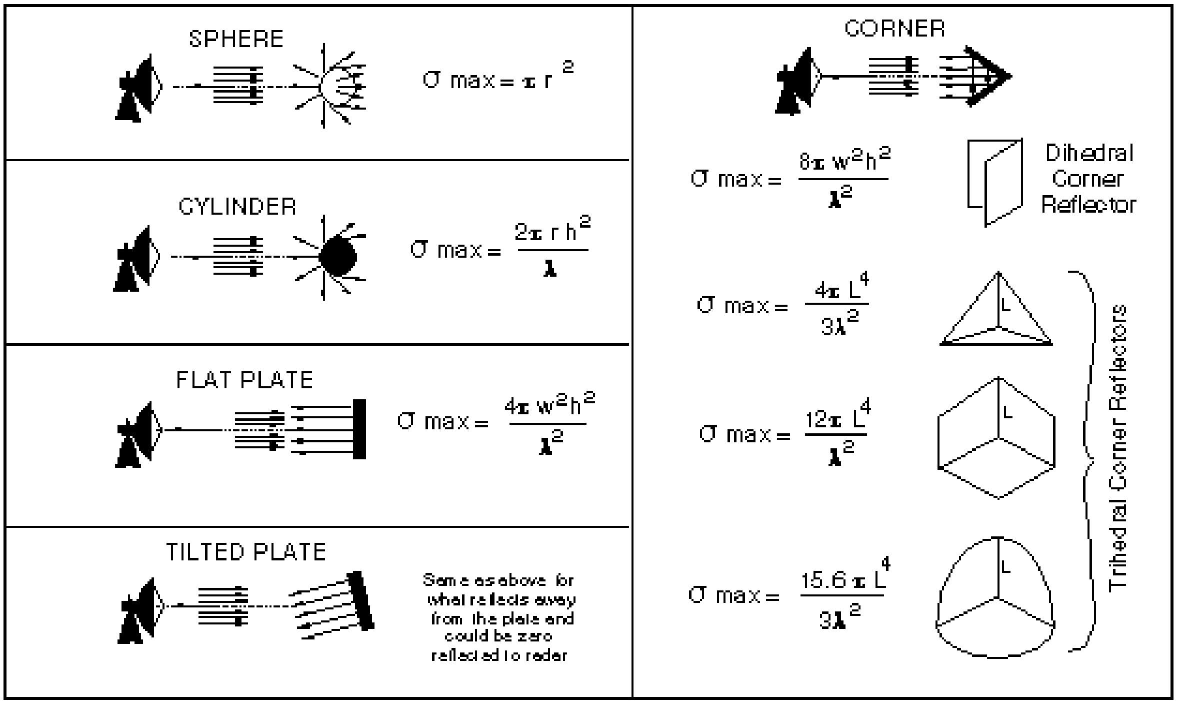
SPHERE
$\sigma \max = \pi r^2$
CYLINDER
$\sigma \max = \frac{2\pi r h^2}{\lambda}$
FLAT PLATE
$\sigma \max = \frac{4\pi w^2 h^2}{\lambda^2}$
TILTED PLATE
Same as above for what reflects away from the plate and could be zero reflected to radar
CORNER
$\sigma \max = \frac{8\pi w^2 h^2}{\lambda^2}$
Dihedral Corner Reflector
$\sigma \max = \frac{4\pi L^4}{3\lambda^2}$
$\sigma \max = \frac{12\pi L^4}{\lambda^2}$
$\sigma \max = \frac{15.6\pi L^4}{3\lambda^2}$
L
Trihedral Corner Reflectors

Figure 3. Backscatter From Shapes.

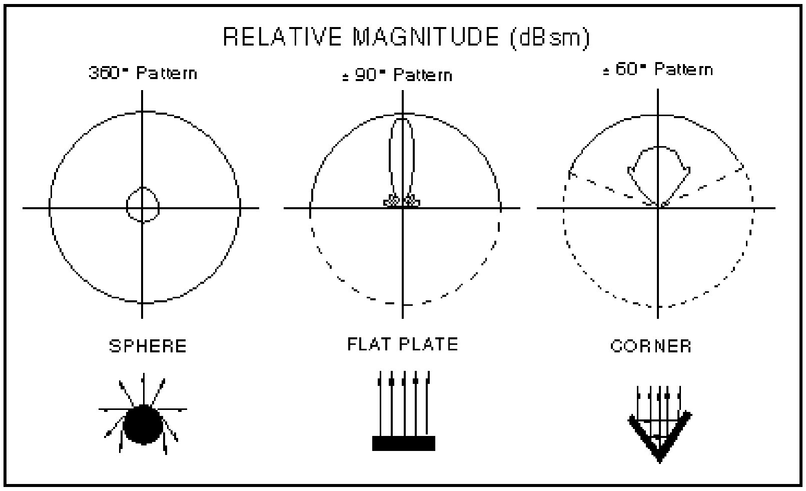
RELATIVE MAGNITUDE (dBsm)
360° Pattern
± 90° Pattern
± 60° Pattern
SPHERE
FLAT PLATE
CORNER

Figure 4. RCS Patterns.

In Figure 4, RCS patterns are shown as objects are rotated about their vertical axes (the arrows indicate the direction of the radar reflections).

The sphere is essentially the same in all directions.

The flat plate has almost no RCS except when aligned directly toward the radar.

The corner reflector has an RCS almost as high as the flat plate but over a wider angle, i.e., over ±60°. The return from a corner reflector is analogous to that of a flat plate always being perpendicular to your collocated transmitter and receiver.

Targets such as ships and aircraft often have many effective corners. Corners are sometimes used as calibration targets or as decoys, i.e. corner reflectors.

An aircraft target is very complex. It has a great many reflecting elements and shapes. The RCS of real aircraft must be measured. It varies significantly depending upon the direction of the illuminating radar.

Figure 5 shows a typical RCS plot of a jet aircraft. The plot is an azimuth cut made at zero degrees elevation (on the aircraft horizon). Within the normal radar range of 3-18 GHz, the radar return of an aircraft in a given direction will vary by a few dB as frequency and polarization vary (the RCS may change by a factor of 2-5). It does not vary as much as the flat plate.

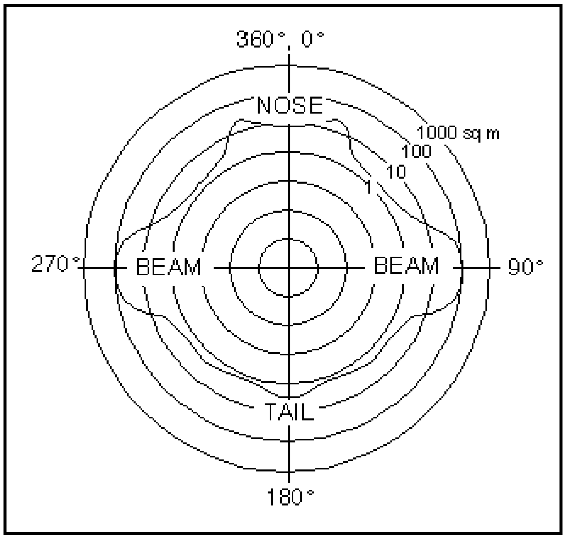

Figure 5. Typical Aircraft RCS.

As shown in Figure 5, the RCS is highest at the aircraft beam due to the large physical area observed by the radar and perpendicular aspect (increasing reflectivity). The next highest RCS area is the nose/tail area, largely because of reflections off the engines or propellers. Most self-protection jammers cover a field of view of +/- 60 degrees about the aircraft nose and tail, thus the high RCS on the beam does not have coverage. Beam coverage is frequently not provided due to inadequate power available to cover all aircraft quadrants, and the side of an aircraft is theoretically exposed to a threat 30% of the time over the average of all scenarios.

Typical radar cross sections are as follows: Missile 0.5 sq m; Tactical Jet 5 to 100 sq m; Bomber 10 to 1000 sq m; and ships 3,000 to 1,000,000 sq m. RCS can also be expressed in decibels referenced to a square meter (dBsm) which equals 10 log (RCS in m^2).

Again, Figure 5 shows that these values can vary dramatically. The strongest return depicted in the example is 100 m^2 in the beam, and the weakest is slightly more than 1 m^2 in the 135°/225° positions. These RCS values can be very misleading because other factors may affect the results. For example, phase differences, polarization, surface imperfections, and material type all greatly affect the results. In the above typical bomber example, the measured RCS may be much greater than 1000 square meters in certain circumstances (90°, 270°).

SIGNIFICANCE OF THE REDUCTION OF RCS

If each of the range or power equations that have an RCS (σ) term is evaluated for the significance of decreasing RCS, Figure 6 results. Therefore, an RCS reduction can increase aircraft survivability. The equations used in Figure 6 are as follows:

Range (radar detection):
From the 2-way range equation in Section 4-4:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}$$

Therefore, $R^4 \propto \sigma$ or $\sigma^{1/4} \propto R$

Range (radar burn-through):
The crossover equation in Section 4-8 has:

$$R_{BT}^2 = \frac{P_t G_t \sigma}{P_j G_j 4\pi}$$

Therefore, $R_{BT}^2 \propto \sigma$ or $\sigma^{1/2} \propto R_{BT}$

Power (jammer): Equating the received signal return (P_r) in the two way range equation to the received jammer signal (P_r) in the one way range equation, the following relationship results:

$$P_r = \underset{\displaystyle S}{\underset{\uparrow}{\frac{P_t\, G_t\, G_r\, \lambda^2\, \sigma}{(4\pi)^3\, R^4}}} = \underset{\displaystyle J}{\underset{\uparrow}{\frac{P_j\, G_j\, G_r\, \lambda^2}{(4\pi R)^2}}}$$

Therefore, $P_j \propto \sigma$ or $\sigma \propto P_j$

Note: jammer transmission line loss is combined with the jammer antenna gain to obtain G_t.

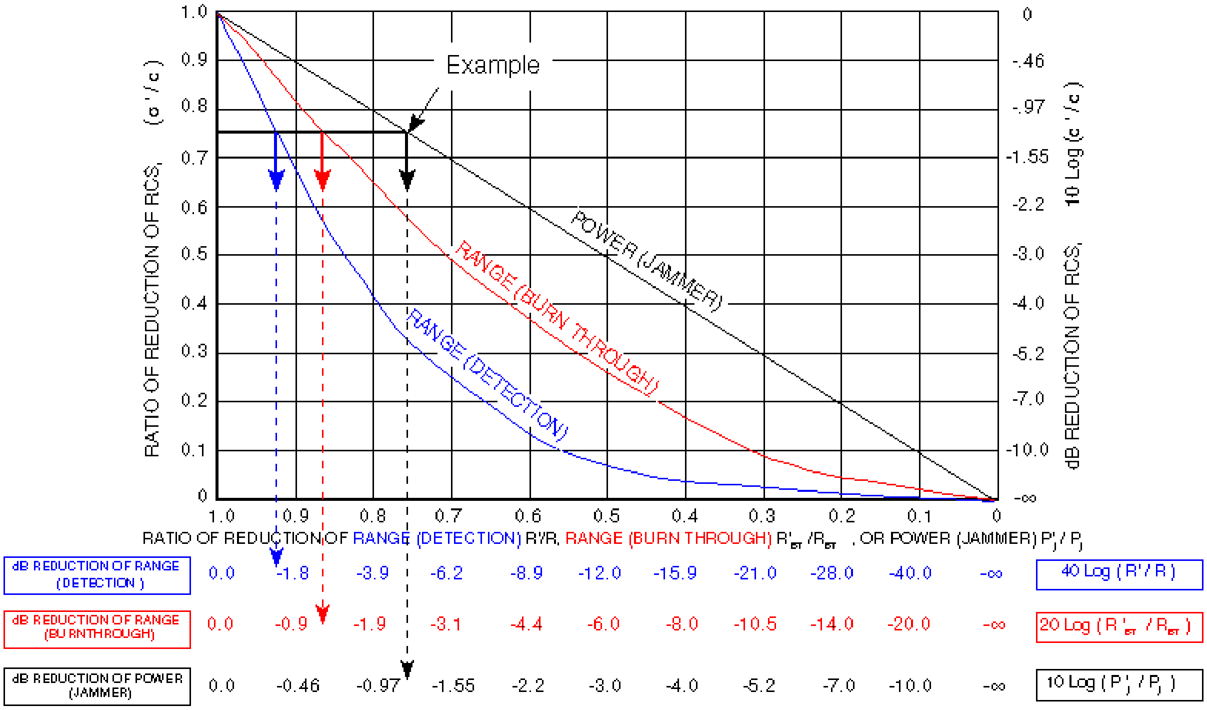

Figure 6. Reduction of RCS Affects Radar Detection, Burn-through, and Jammer Power.

Example of Effects of RCS Reduction - As shown in Figure 6, if the RCS of an aircraft is reduced to 0.75 (75%) of its original value, then (1) the jammer power required to achieve the same effectiveness would be 0.75 (75%) of the original value (or -1.25 dB). Likewise, (2) If Jammer power is held constant, then burn-through range is 0.87 (87%) of its original value (-1.25 dB), and (3) the detection range of the radar for the smaller RCS target (jamming not considered) is 0.93 (93%) of its original value (-1.25 dB).

OPTICAL / MIE / RAYLEIGH REGIONS

Figure 7 shows the different regions applicable for computing the RCS of a sphere. The optical region ("far field" counterpart) rules apply when $2\pi r/\lambda > 10$. In this region, the RCS of a sphere is independent of frequency. Here, the RCS of a sphere, $\sigma = \pi r^2$. The RCS equation breaks down primarily due to creeping waves in the area where λ~2πr. This area is known as the Mie or resonance region. If we were using a 6” diameter sphere, this frequency would be 0.6 GHz. (Any frequency ten times higher, or above 6 GHz, would give expected results). The largest positive perturbation (point A) occurs at exactly 0.6 GHz where the RCS would be 4 times higher than the RCS computed using the optical region formula. Just slightly above 0.6 GHz a minimum occurs (point B) and the actual RCS would be 0.26 times the value calculated by using the optical region formula. If we used a one meter diameter sphere, the perturbations would occur at 95 MHz, so any frequency above 950 MHz (~1 GHz) would give predicted results.

CREEPING WAVES

The initial RCS assumptions presume that we are operating in the optical region (λ<<Range and λ<<radius). There is a region where specular reflected (mirrored) waves combine with back scattered creeping waves both constructively and destructively as shown in Figure 8. Creeping waves are tangential to a smooth surface and follow the “shadow” region of the body. They occur when the circumference of the sphere ~ λ and typically add about 1 m^2 to the RCS at certain frequencies.

RAYLEIGH REGION
$\sigma = [\pi r^2][7.11(kr)^4]$

where: $k = 2\pi/\lambda$

MIE (resonance)
$\sigma = 4\pi r^2$ at Maximum (point A)
$\sigma = 0.26\pi r^2$ at Minimum (pt B)

OPTICAL REGION
$\sigma = \pi r^2$
(Region RCS of a sphere is independent of frequency)

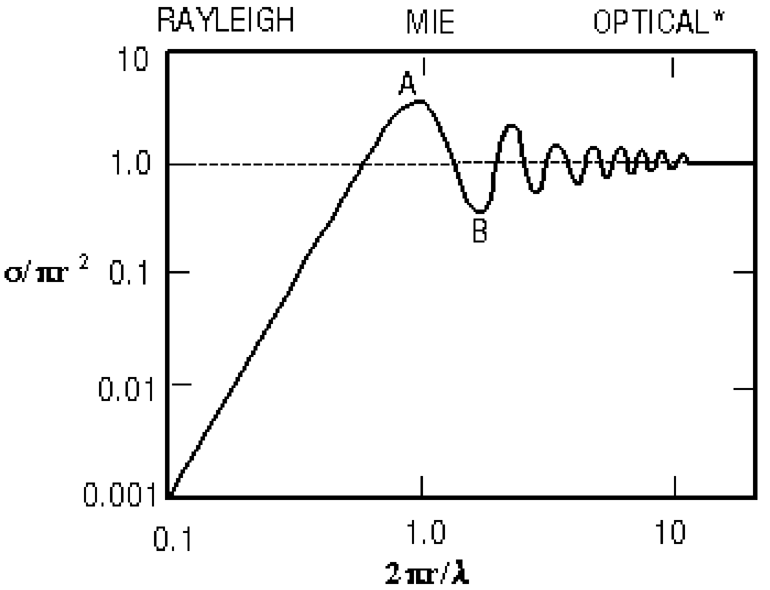

* "RF far field" equivalent
Courtesy of Dr. Allen E. Fuhs, Ph.D.

Figure 7. Radar Cross Section of a Sphere.

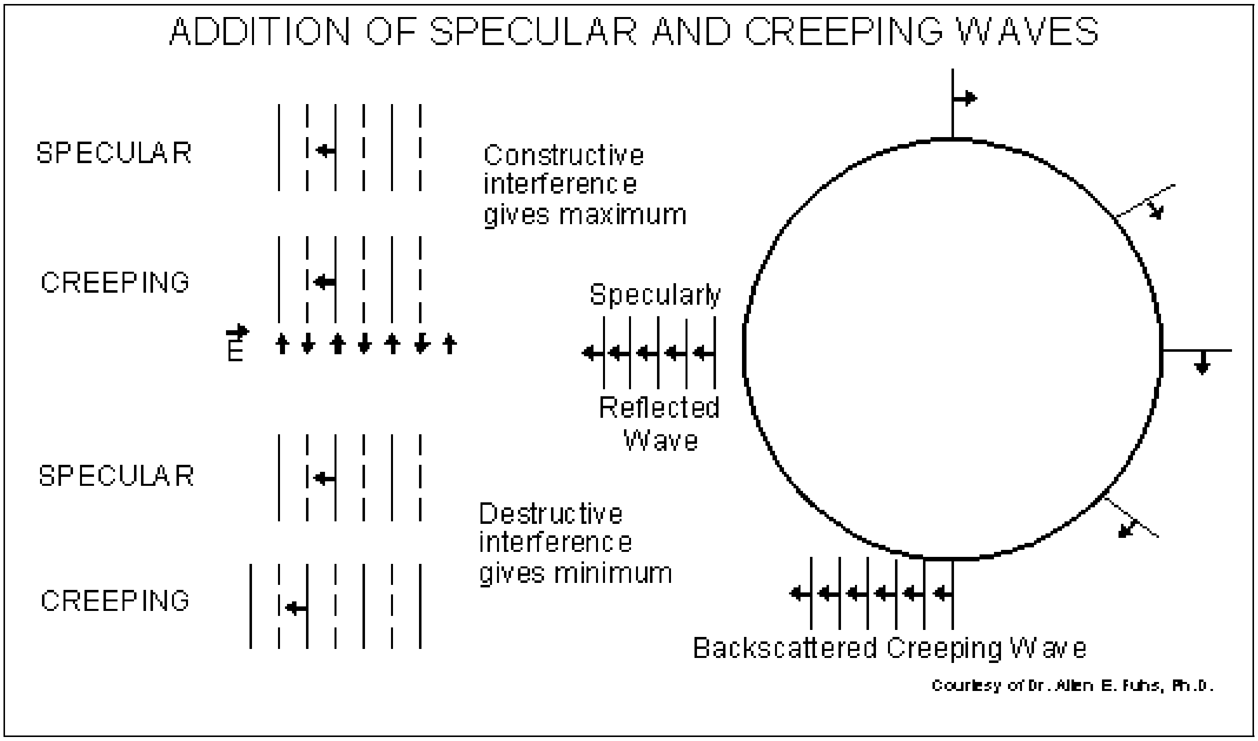
ADDITION OF SPECULAR AND CREEPING WAVES
SPECULAR
CREEPING
E
Constructive interference gives maximum
SPECULAR
CREEPING
Destructive interference gives minimum
Specularly
Reflected Wave
Backscattered Creeping Wave
Courtesy of Dr. Allen E. Fuhs, Ph.D.

Figure 8. Addition of Specular and Creeping Waves.

EMISSION CONTROL (EMCON)

When EMCON is imposed, RF emissions must not exceed -110 dBm/meter^2 at one nautical mile. It is best if systems meet EMCON when in either the Standby or Receive mode versus just the Standby mode (or OFF). If one assumes antenna gain equals line loss, then emissions measured at the port of a system must not exceed -34 dBm (i.e. the stated requirement at one nautical mile is converted to a measurement at the antenna of a point source - see Figure 1). If antenna gain is greater than line loss (i.e. gain 6 dB, line loss 3 dB), then the -34 dBm value would be lowered by the difference and would be -37 dBm for the example. The opposite would be true if antenna gain is less.

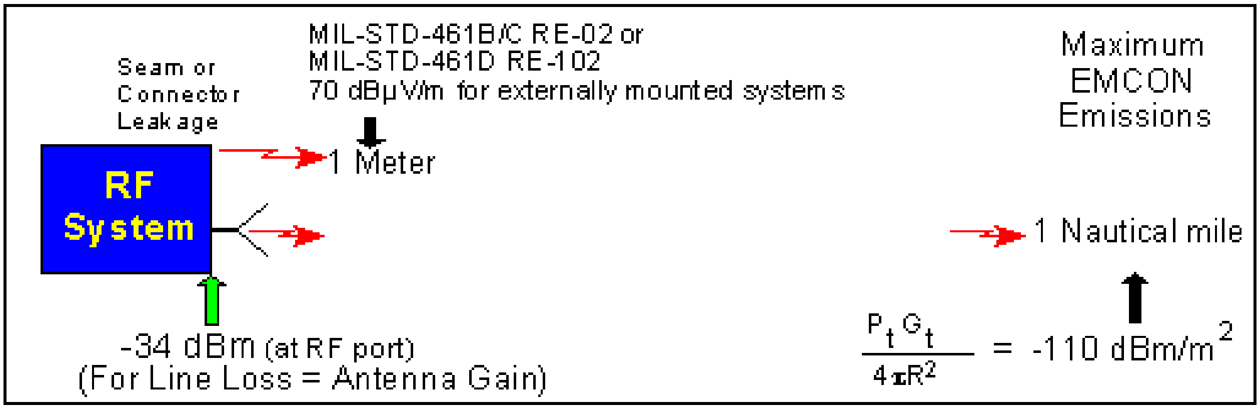

Seam or Connector Leakage
MIL-STD-461B/C RE-02 or
MIL-STD-461D RE-102
70 dBµV/m for externally mounted systems
1 Meter
Maximum EMCON Emissions
RF System
1 Nautical mile
-34 dBm (at RF port)
(For Line Loss = Antenna Gain)
$\frac{P_t G_t}{4\pi R^2} = -110\ \text{dBm/m}^2$

Figure 1. EMCON Field Intensity / Power Density Measurements.

To compute the strength of emissions at the antenna port in Figure 1, we use the power density equation (see Section 4-2)

$$P_D = \frac{P_t G_t}{4\pi R^2} \quad \textbf{[1]} \qquad \text{or rearranging} \qquad P_t G_t = P_D\,(4\pi R^2) \quad \textbf{[2]}$$

Given that $P_D = -110\ \text{dBm/m}^2 = (10)^{-11}\ \text{mW/m}^2$, and R = 1 NM = 1852 meters.

$$P_t G_t = P_D\,(4\pi R^2) = (10^{-11}\text{mW/m}^2)(4\pi)(1852\text{m})^2 = 4.31(10)^{-4}\ \text{mW} = -33.65 \approx -34\ \text{dBm}$$

at the RF system antenna as given.

or, the equation can be rewritten in Log form and each term multiplied by 10:

$$10\log P_t + 10\log G_t = 10\log P_D + 10\log\,(4\pi R^2) \quad \textbf{[3]}$$

Since the m^2 terms on the right side of equation [3] cancel, then:

$$10\log P_t + 10\log G_t = -110\ \text{dBm} + 76.35\ \text{dB} = -33.65\ \text{dBm} \approx -34\ \text{dBm as given in Figure 1.}$$

If MIL-STD-461B/C RE02 (or MIL-STD-461D RE-102) measurements (see Figure 2) are made on seam/connector leakage of a system, emissions below 70 dBµV/meter which are measured at one meter will meet the EMCON requirement. Note that the airframe provides attenuation so portions of systems mounted inside an aircraft that measure 90 dBµV/meter will still meet EMCON if the airframe provides 20 dB of shielding (note that the requirement at one nm is converted to what would be measured at one meter from a point source).

The narrowband emission limit shown in Figure 2 for RE02/RE102 primarily reflect special concern for local oscillator leakage during EMCON as opposed to switching transients which would apply more to the broadband limit.

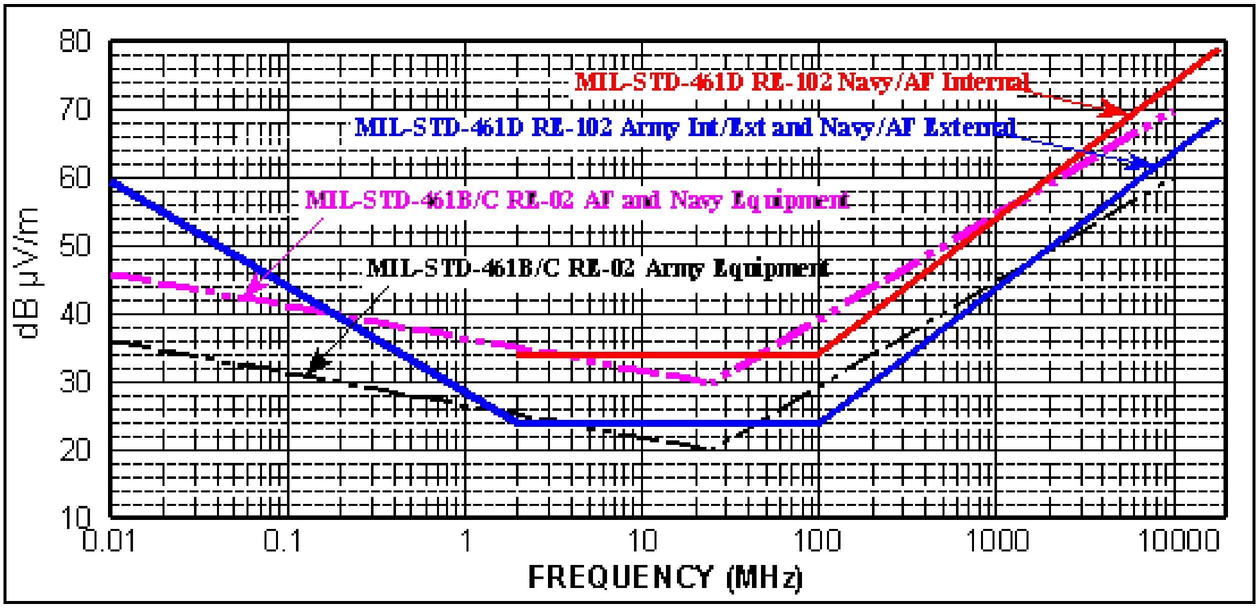

Figure 3. MIL-STD-461 Narrowband Radiated Emissions Limits.

Note that in MIL-STD-461D, the narrowband radiated emissions limits were retitled RE-102 from the previous RE-02 and the upper frequency limit was raised from 10 GHz to 18 GHz. The majority of this section will continue to reference RE02 since most systems in use today were built to MIL-STD-461B/C.

For the other calculation involving leakage (to obtain 70 dBµV/m) we again start with: $P_D = \frac{P_t G_t}{4\pi R^2}$

and use the previous fact that: 10log (P_tG_t) = -33.6 dBm = 4.37x10^{-4} mW (see Section 2-4).

The measurement is at one meter so $R^2 = 1\ m^2$

we have: $\frac{4.37x\,10^{-4}}{4\pi}\ mW/\ m^2 = .348x\,10^{-4}\ mW/\ m^2 = -\,44.6\ dBm/\ m^2 = P_D$ @ 1 meter

Using the field intensity and power density relations (see Section 4-1)

$$E = \sqrt{P_D Z} = \sqrt{3.48x\,10^{-8} \bullet 377\Omega} = 36.2x\,10^{-4}\ V/m$$

Changing to microvolts (1V = 10^6 µV) and converting to logs we have:

20 log (E) = 20 log (10^6 x 36.2x10^{-4}) = 20 log (.362x10^4) = 71.18 dBµV/m ≈ 70 dBµV/m as given in Figure 1.

Some Words of Caution

A common error is to only use the one-way free space loss coefficient α_1 directly from Figure 6, Section 4-3 to calculate what the output power would be to achieve the EMCON limits at 1 NM. This is incorrect since the last term on the right of equation [3] (10 Log($4\pi R^2$)) is simply the Log of the surface area of a sphere - **it is NOT** the one-way free space loss factor α_1. You cannot interchange power (watts or dBW) with power density (watts/m^2 or dBW/m^2).

The equation uses power density (P_D), **NOT** received power (P_r). It is independent of RF and therefore varies only with range. If the source is a transmitter and/or antenna, then the power-gain product (or EIRP) is easily measured and it's readily apparent if 10log (P_t G_t) is less than -34 dBm. If the output of the measurement system is connected to a power meter in place of the system transmission line and antenna, the -34 dBm value must be adjusted. The measurement on the power meter (dBm) minus line loss (dB) plus antenna gain (dB) must not be higher than -34 dBm.

However, many sources of radiation are through leakage, or are otherwise inaccessible to direct measurement and P_D must be measured with an antenna and a receiver. The measurements must be made at some RF(s), and received signal strength is a function of the antenna used therefore measurements must be scaled with an appropriate correction factor to obtain correct power density.

RE-02 Measurements

When RE-02 measurements are made, several different antennas are chosen dependent upon the frequency range under consideration. The voltage measured at the output terminals of an antenna is not the actual field intensity due to actual antenna gain, aperture characteristics, and loading effects. To account for this difference, the antenna factor is defined as:

$$AF = E/V \qquad \textbf{[4]}$$

where E = Unknown electric field to be determined in V/m (or μV/m)
V = Voltage measured at the output terminals of the measuring antenna

For an antenna loaded by a 50 Ω line (receiver), the theoretical antenna factor is developed as follows:

$$P_D A_e = P_r = V^2/R = V_r^2/50 \text{ or } V_r = \sqrt{50\, P_D A_e}$$

From Section 4-3 we see that $A_e = G_r\lambda^2/4\pi$, and from Section 4-1, $E^2 = 377\ P_D$ therefore we have:

$$AF = \frac{E}{V} = \frac{\sqrt{377\ P_D}}{\sqrt{50\ P_D(\lambda^2 G_r / 4\pi)}} = \frac{9.73}{\lambda\sqrt{G_r}} \qquad \textbf{[5]}$$

Reducing this to decibel form we have:

$$20 \log AF = 20 \log E - 20 \log V = 20 \log\left[\frac{9.73}{\lambda\sqrt{GSUBr}}\right] \quad \textit{with } \lambda \textit{ in meters and Gain numeric ratio (not dB)} \qquad \textbf{[6]}$$

This equation is plotted in Figure 3.

Since all of the equations in this section were developed using far field antenna theory, use only the indicated region.

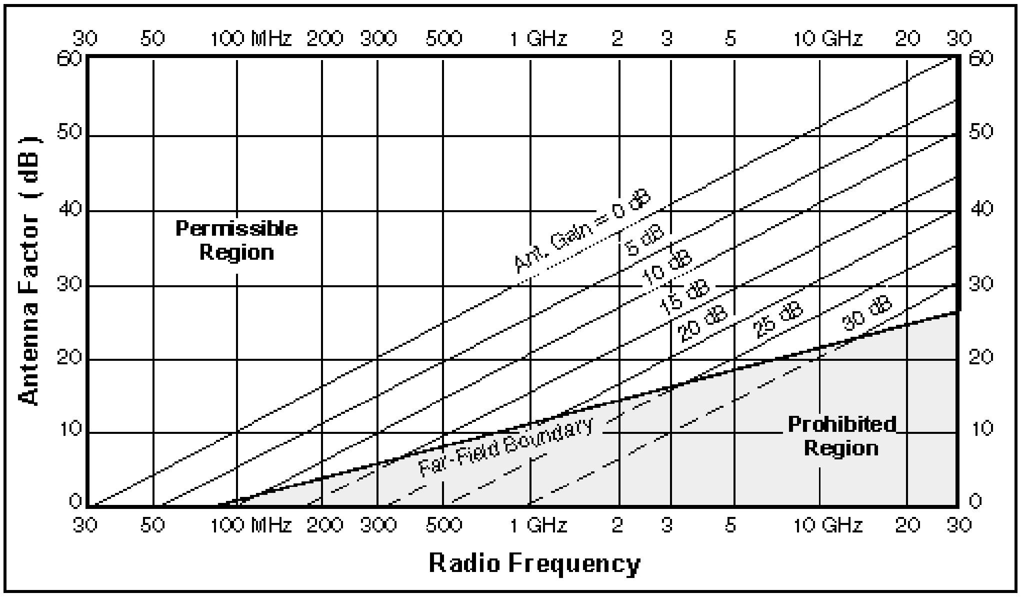

Figure 3. Antenna Factor vs. Frequency for Indicated Antenna Gain.

In practice the electric field is measured by attaching a field intensity meter or spectrum analyzer with a narrow bandpass preselector filter to the measuring antenna, recording the actual reading in volts and applying the antenna factor.

$$20\log E = 20\log V + 20\log AF \qquad \textbf{[7]}$$

Each of the antennas used for EMI measurements normally has a calibration sheet for both gain and antenna factor over the frequency range that the antenna is expected to be used. Typical values are presented in Table 1.

Table 1. Typical Antenna Factor Values.

Frequency Range	Antenna(s) Used	Antenna Factor	Gain(dB)
14 kHz - 30 MHz	41” rod	22-58 dB	0 - 2
20 MHz - 200 MHz	Dipole or Biconical	0-18 dB	0 - 11
200 MHz - 1 GHz	Conical Log Spiral	17-26 dB	0 - 15
1 GHz - 10 GHz	Conical Log Spiral or Ridged Horn	21-48 dB	0 - 28
1 GHz - 18 GHz	Double Ridged Horn	21-47 dB	0 - 32
18 GHz - 40 GHz	Parabolic Dish	20-25 dB	27 - 35

The antenna factor can also be developed in terms of the receiving antenna's effective area. This can be shown as follows:

$$AF = \frac{E}{V} = \frac{\sqrt{377\, P_D}}{\sqrt{50\, P_D\, A_e}} = \frac{2.75}{\sqrt{A_e}} \qquad \textbf{[8]}$$

Or in log form:

$$20 \log AF = 20 \log E - 20 \log V = 20 \log \left[\frac{2.75}{\sqrt{A_e}} \right] \qquad \textbf{[9]}$$

While this relation holds for any antenna, many antennas (spiral, dipole, conical etc.) which do not have a true "frontal capture area" do not have a linear or logarithmic relation between area and gain and in that respect the parabolic dish is unique in that the antenna factor does not vary with frequency, only with effective capture area. Consequently a larger effective area results in a smaller antenna factor.

A calibrated antenna would be the first choice for making measurements, followed by use of a parabolic dish or "standard gain" horn. A standard gain horn is one which was designed such that it closely follows the rules of thumb regarding area/gain and has a constant antenna factor. If a calibrated antenna, parabolic dish, or "standard horn" is not available, a good procedure is to utilize a flat spiral antenna (such as the AN/ALR-67 high band antennas). These antennas typically have an average gain of 0 dB (typically -4 to +4 dB), consequently the antenna factor would not vary a lot and any error would be small.

EXAMPLE:

Suppose that we want to make a very general estimation regarding the ability of a system to meet EMCON requirements. We choose to use a spiral antenna for measurements and take one of our samples at 4 GHz. Since we know the gain of the spiral is relatively flat at 4 GHz and has a gain value of approximately one (0 dB) in that frequency range. The antenna is connected to a spectrum analyzer by 25 feet of RG9 cable. We want to take our measurements at 2 meters from the system so our setup is shown as follows:

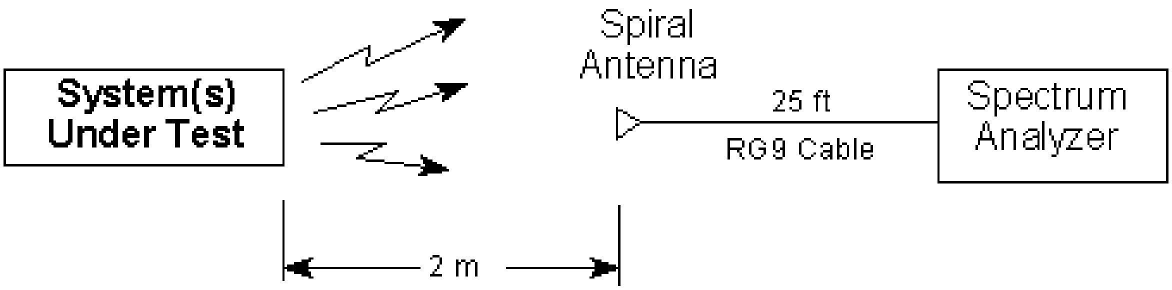

Our RG9 cable has an input impedance of 50Ω, and a loss of 5 dB (from Figure 5, Section 6-1).

First, let's assume that we measure -85 dBm at the spectrum analyzer and we want to translate this into the equivalent strength at 1 NM. Our power received by the antenna is: P_r = -85 dBm + 5 dB line loss = -80 dBm

also $P_D = P_r/A_e$ and $A_e = G\lambda^2/4\pi = (G/4\pi)\bullet(c/f)^2 = (1/4\pi)\bullet(3x10^8/4x10^9)^2 = 4.47x10^{-4}$ m^2

in log form: $10 \text{ Log } P_D = 10 \text{ Log } P_r - 10 \text{ Log } A_e$ = -80 dBm + 33.5 = -46.5 dBm/m^2
at our 2 meter measuring point

To convert this to a value at 1 NM, we use

$P_t\, G_t = P_{D@1\ nm}\, 4\pi R_1^{\,2} = P_{D@2\ m}\, 4\pi R_2^{\,2}$ and we solve for PD@1 $_{nm}$

in log form after cancelling the 4π terms:

$10 \text{ Log } P_{D@1\ nm} = 10 \text{ Log } P_{D@2\ m} + 10 \text{ Log } (R_{2m}/R_{1nm})^2$ = -46.5 dBm/m^2 - 59.3 dB = -105.8 dBm/m^2 which is more power than the maximum value of -110 dBm/m^2 specified.

If we are making repetitive measurement as we might do when screening an aircraft on the flight line with numerous systems installed, or when we want to improve (reduce) the leakage on a single system by changing antennas, lines, connectors, or EMI gaskets or shielding, this mathematical approach would be unnecessarily time consuming since it would have to be repeated after each measurement. A better approach would be to convert the -110 dBm/m^2 value at 1 NM to the maximum you can have at the measuring instrument (in this case a spectrum analyzer), then you could make multiple measurements and know immediately how your system(s) are doing. It should be noted that -90 to -100 dBm is about the minimum signal level that can be detected by a spectrum analyzer, so you couldn't take measurements much further away unless you used an antenna with a much higher gain.

In order not to exceed EMCON, the power density must not exceed -110 dBm/m^2 at 1 NM, which is 10^{-11} mW/m^2.

$$P_t\, G_t = P_{D@1\ nm}\, 4\pi R_1^{\,2} = P_{D@2\ m}\, 4\pi R_2^{\,2}$$

we solve for $P_{D@2\ m} = 10^{-11}(1852m)^2/(2m)^2 = 8.57 \text{ x } 10^{-6}$ mW/m^2 = -50.7 dBm/m^2

We'll be using a spectrum analyzer, so we want to compute what the maximum power or voltage may be.

Method 1 - Using the Power Density Approach

Using logs/dB and the values of $P_{D@2\ m}$ and A_e determined previously:
$10 \text{ Log } P_r = 10 \text{ Log } P_D + 10 \text{ Log } A_e$ = -50.7 - 33.5 = -84.2 dBm

taking line loss into account we have: -84.2 - 5 dB = - 89.2 dBm as the maximum measurement reading.

If we wanted to calculate it in volts, and take into account our line impedance we would have the following:

$P_r = P_D A_e = V^2/R = V^2/50\Omega$ also $A_e = G\lambda^2/4\pi$ so solving for V we have:

$$V = \sqrt{P_D\left[\frac{G_r\lambda^2}{4\pi}\right]R} = \sqrt{P_D\left[\frac{G_r}{4\pi}\left(\frac{c}{f}\right)^2\right]R} = \sqrt{8.57x10^{-9}\left[\frac{1}{4\pi}\left(\frac{3x10^8}{4x10^9}\right)^2\right]50\Omega} = 1.38x10^{-5}\ volts\quad (before\ line\ loss)$$

since our line loss is 5 dB, we have -5 dB = 20 Log V_2/V_1. Solving for V_2 we get $\underline{7.79x10^{-6}\text{ volts}}$ or -89 dBm as a maximum at our measurement device input. We can see immediately that our value of -85 dBm that we measured on the previous page would not meet specifications, and neither would any signal with more power than -89 dBm.

Method 2 - Using the Antenna Factor Approach

Starting with the same value of power density that we obtained above ($8.57x10^{-9}$ W/m^2), we find the field intensity from Table 1, Section 4-1 to be approximately 65 dBμv/m. Also from Figure 3 in this section, AF = 43 dB @ 4 GHz (by calculating with equation [6], the exact value is 42.3 dB).

From equation [6]:
20log V = 20log E - 20log AF
20log V = 65 - 43 = 22 dBμv/m.

Since dBμv/m = 20 log (V)(10^6) = 20 log V + 20 log 10^6 = 20 log V + 120, we see that to get an answer in dBv we must subtract 120 from the dBμv/m value so: V_{dB} = 22 - 120 = -98dBv. We then subtract our line loss (-5 dB) and we have:

V = -98 - 5 = -103 dBv = 17 dBμv = $\underline{7.1x10^{-6}\text{ volts}}$

using the fact that P = V^2/R and for the input line R = 50Ω, P = $1x10^{-12}$ W = -120 dBW = -90 dBm

Although this method is just as accurate as that obtained using method 1, the values obtained in Table 1, Section 4-1, and Figure 3 must be interpolated, and may not result in values which are as precise as the appropriate formulas would produce.

Sample Problem: What is the approximate transmit power from a receiver?

A. 1 nanowatt (nW)
B. 10 nW
C. 100 nW
D. 1 microwatt (μW)
E. 10 μW
F. 100 μW
G. 1 milliwatt (mW)
H. 10 mW
I. 100 mW
J. 1 watt (W)
K. 10 W
L. 100 W
M. 1 kilowatt (kW)
N. 10 kW
O. 100 kW

The question may seem inappropriate since a receiver is supposedly a passive device which only receives a signal. If the receiver was a crystal video receiver as shown in Section 5-3, it wouldn't transmit power unless a built-in-test (BIT) signal was injected after the antenna to periodically check the integrity of the microwave path and components. The potential exists for the BIT signal to leak across switches and couple back through the input path and be transmitted by the receiver's antennas.

If the receiver uses a local oscillator (LO) and a mixer to translate the signal to an intermediate frequency (IF) for processing (such as a superhet shown in Section 5-3), there is the potential for the CW LO signal to couple back through the signal input path and be transmitted by the receiver's antenna. Normally a mixer has 20 dB of rejection for the reverse direction. In addition, the LO may be further attenuated by receiver front end filters.

In both cases, the use of isolators described in Section 6-7 could be used to further attenuate any signals going in the reverse direction, i.e. back to the antenna. A good receiver design should ensure that any RF leakage radiated by the receiver will not exceed the EMCON level.

In answer to the initial question, "transmit" leakage power should be less than -34 dBm (0.4 μW) to meet EMCON. Therefore, the real answer may be "A," "B," or "C" if EMCON is met and could be "D" through possibly "G" if EMCON is not met.

EW JAMMING TECHNIQUES

INTRODUCTION

Electronic jamming is a form of Electronic Attack where jammers radiate interfering signals toward an enemy's radar, blocking the receiver with highly concentrated energy signals. The two main technique styles are noise techniques and repeater techniques. The three types of noise jamming are spot, sweep, and barrage. Repeater techniques can be further subdivided into categories as shown in Figure 1.

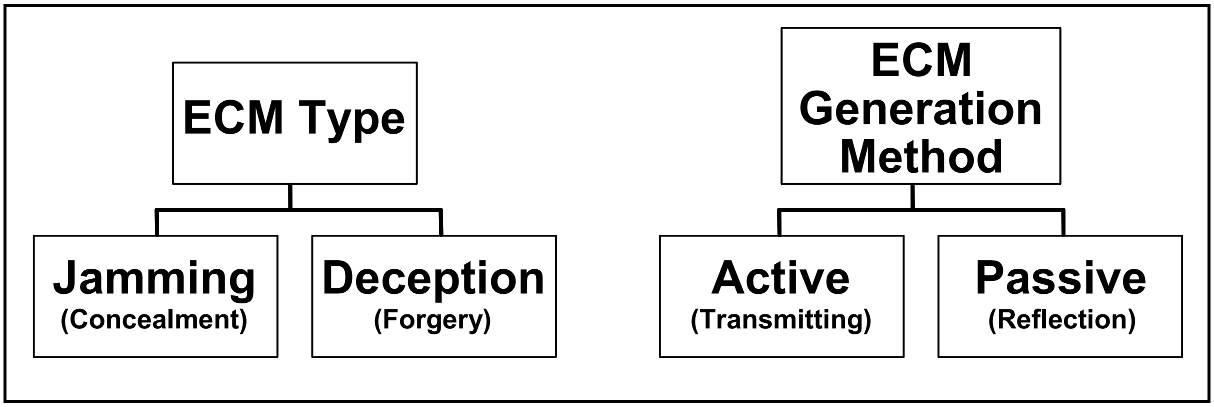
ECM Type
Jamming
(Concealment)
Deception
(Forgery)
ECM Generation Method
Active
(Transmitting)
Passive
(Reflection)

Figure 1. EA Repeater Technique Divisions.

ASYNCHRONOUS SWEPT WAVE MODULATION (ASWM)

ASWM is synonymous with A-SWM. Asynchronous indicates that the waveform is free running - also see SSWM. A swept wave modulation is essentially a swept amplitude modulation (SAM). It is a waveform that is swept between two frequencies that are usually chosen to bracket a radar's passive angle scanning rate. The modulation amplitude can be either down modulated or On-Off Keyed (OOK). The down modulated shape can be square wave, rectangular wave, linear (e.g. a sine wave), or a combination. The OOK modulated shape can be square wave or rectangular wave.

BARRAGE JAMMING

The jamming of multiple frequencies at once by a single jammer. The advantage is that multiple frequencies can be jammed simultaneously; however, the jamming effect can be limited because this requires the jammer to spread its full power between these frequencies. So the more frequencies being jammed, the less effectively each is jammed.

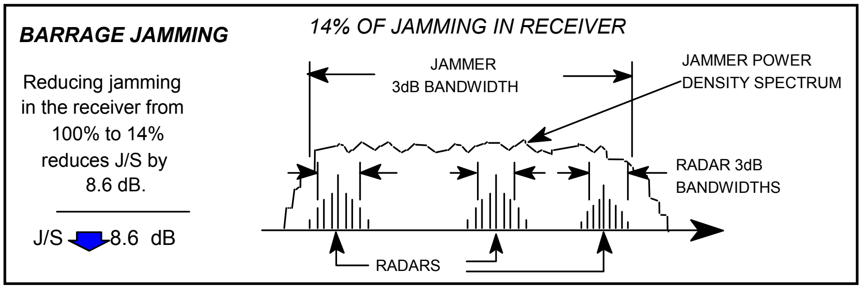
BARRAGE JAMMING
14% OF JAMMING IN RECEIVER
Reducing jamming in the receiver from 100% to 14% reduces J/S by 8.6 dB.
J/S 8.6 dB
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR 3dB BANDWIDTHS
RADARS

Figure 2. Barrage Jamming.

CROSS POLARIZATION (X-POL)

(1) A self-screening or support EA technique that causes angle errors in tracking radars and sensing errors in jamming suppression EP systems of surveillance radars by radiating a signal that is orthogonally polarized to the principal polarization of the victim radar. (2) A technique used against monopulse and other passive lobe tracking radars. Requires a strong jam-to-signal ratio or the skin echo will show up in the pattern nulls.

HOME ON JAM (HOJ)

A means whereby a missile guidance receiver utilizes the self-screening target jamming signal to develop angular steering information so that the missile can home on that target.

IMAGE JAMMING

Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and image frequency of the radar. Image jamming inverts the phase of the response and is thereby useful as an angle deception technique. Not effective if the radar uses image rejection.

INVERSE CON SCAN (ICS)

One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna or antenna dipole scan rate and then modulating repeater amplifier gain so that the weapons system will fire at some bearing other than the true target bearing.

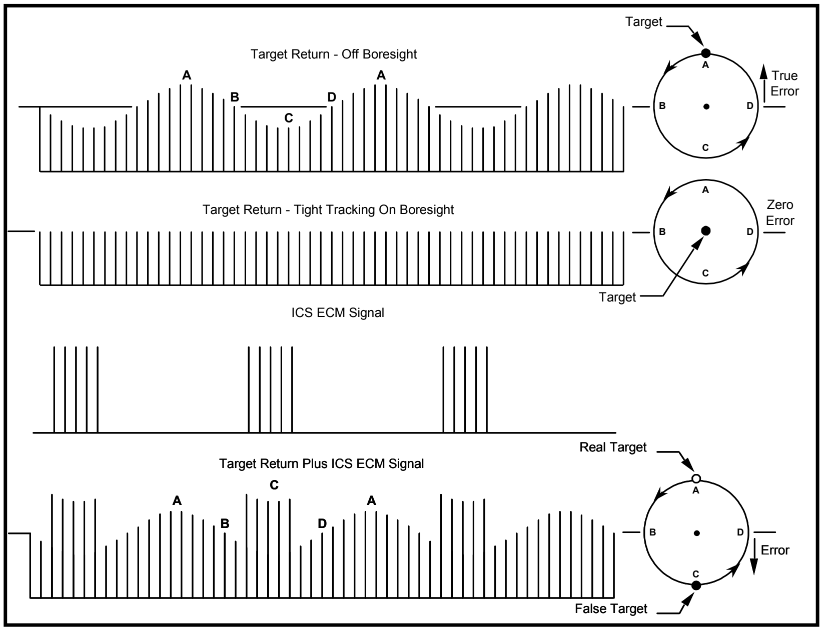
Target
Target Return - Off Boresight
A
B
C
D
A
A
B
D
C
True Error
Target Return - Tight Tracking On Boresight
A
B
D
C
Zero Error
Target
ICS ECM Signal
Real Target
Target Return Plus ICS ECM Signal
C
A
B
D
A
A
B
D
C
Error
False Target

Figure 3. Inverse Con Scan.

J/S REQUIREMENT

The jamming to signal ratio for effective coverage of the true target. Usually on the order of zero dB (J/S=1).

LINEAR MODULATION

A modulation technique in which the output varies in a straight line (linear) manner with the input (modulating) waveform. This is different from a discontinuous (On-Off) modulation. Typically the modulating waveform is a ramp or triangular wave but could also be a sine wave modulating input.

LOBE ON RECEIVE ONLY (LORO)

Predicting the future location of a target (to track) requires that the radar look at areas where the target is not located. When this scanning is accomplished with the radiated beam, large angle targets such as chaff clouds can create complete radar "white-out." RWR indications can be obtained before actual target lock-on. These problems can be overcome by scanning only the receiving antennas and using a separate transmitting antenna pointed only at the target.

In a LORO system, a transmitting antenna emits a few "exploratory" pulses along a direction obtained from an acquisition radar. These exploratory pulses are the acquisition mode of the TTR. That is, in its acquisition mode the small beamed TTR must scan the large location segment provided by the acquisition radar. In radars equipped with Fast Time Constant, the return pulse is applied to a differentiator of extremely short time constant. When the pulse is received, it is "cut-off" on the leading edge and only that portion is fed to the computer. This allows the radar to effectively track on the leading edge of the target. FTC does not improve the range resolution but it can prevent any countermeasures aft of the target which are in the same resolution cell as the target (such as chaff) from interfering with the radar receiver.

The receiving antennas scan their sector for the target return due to these exploratory pulses; as the power centroid is located, the center of the receiving pattern is brought onto the target. The transmitting antenna, which is slaved to the receiving antenna, is then pointing directly at the desired target and only that target is radiated during tracking. This approach allows a very small radiated beam, but the resolution cell of the system is still that of the receiving antenna.

NOISE JAMMING

The transmission of noise-like signals in the target system's radar receiver bandpass. At low power levels, noise jamming has the characteristics of receiver noise and can be mistaken by the radar operator as a problem with the radar. The object of noise jamming is to introduce a disturbing signal into the hostile electronic equipment so that the actual signal is obscured by the interference. The victim of this disturbance might be a radar receiver, a communications network, or a data link.

See also Barrage Jamming and Spot Jamming.

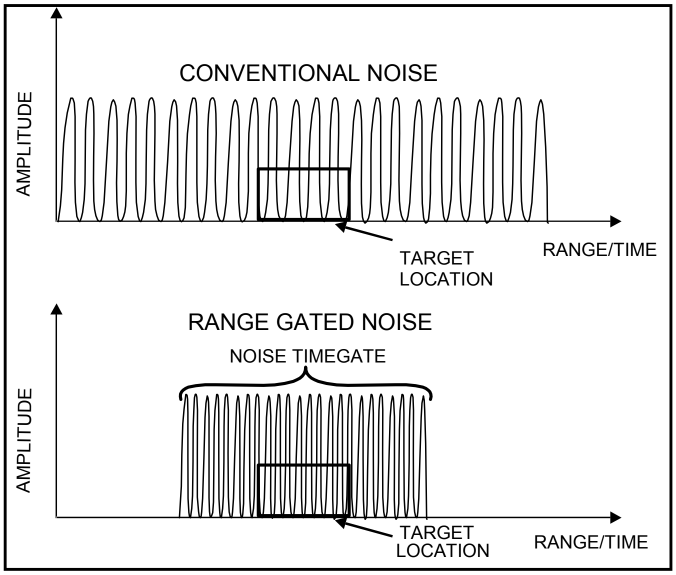

Figure 4. Noise Jamming.

ON-OFF MODULATION

On-Off Modulation is any modulation which switches rapidly between two states. This definition includes pulse radar operation.

On-Off Keying (OOK) is the envelope modulation of a jamming signal with a rectangular wave. The modulation rate and duty cycle are adjusted commensurate with the victim radar's processing time constants. These can be related to AGC time constants, logic time-outs, data sampling cycles, or any other data processing response times.

An important distinction must be made between the terms On-Off Keying and Blinking. While both terms involve envelope modulations that turn a jamming signal on and off, the term OOK is used as the envelope modulation of a single jamming signal, and the term blinking is used as the tactical application of OOK involving two or more cooperative jamming platforms.

PSEUDO RANDOM NOISE (PRN)

A controlled, noise-like, pulse pattern repeated in synchronism with the victim radar pulse repetition frequency. Synonymous with quasi-noise jamming.

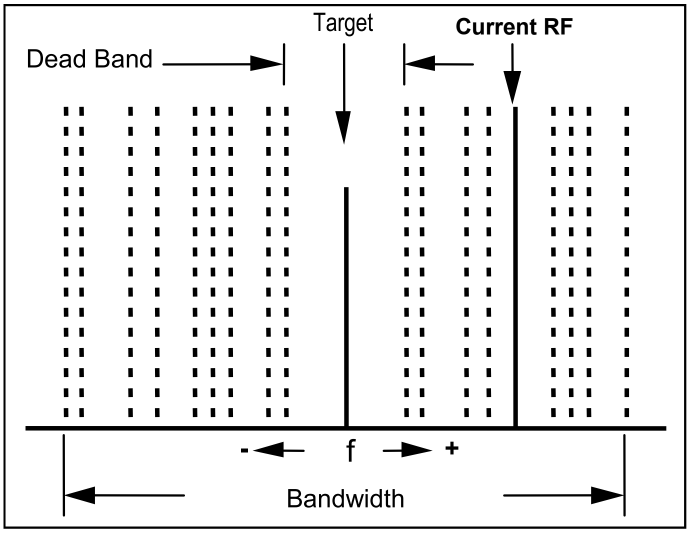

Figure 5. Random Dual Line.

RANDOM DUAL LINE (RDL)

RDL is a coherent repeater technique that is essentially the same as velocity noise (VN), narrow band noise (NBN), and pseudo random noise (PRN) except that no false Doppler frequencies are stepped directly over the target return. The objective is to prevent the momentary additions of the EA signal and target signal that might highlight the presence of a coherent return.

RANDOM RANGE PROGRAM (RANRAP)

A dynamic False Target Jamming technique program to create multiple realistic targets of varying size and distance from the jamming plane.

RANGE GATE PULL OFF (RGPO)

Once a tracking radar has detected a target, it will place *range gates* to either side of it. Range gates essentially blank out all signals which originate from ranges outside a narrow window, substantially increasing the signal-to-noise ratio and protecting the radar against unsynchronized jamming pulses. The radar 'concentrates' on a short range interval which encloses the target's location, and it no longer looks out for other targets. This state is known as 'lock-on'. But range gates can be 'stolen', and it is the objective of the Range Gate Pull-Off (RGPO) technique to *break lock* and escape from out of the window.

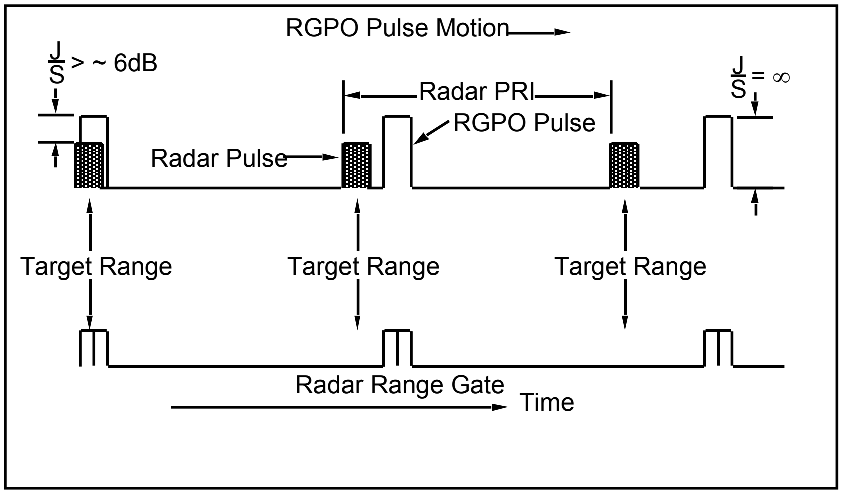

Figure 6. Range Gate Pull Off.

RGPO works as follows:

Upon detection (or assumption) that a tracking radar has locked on, the on-board jammer is switched on and starts to work in a couple of phases:

1. First, a sample of the illuminating pulse signal is taken and the radar's pulse repetition frequency (PRF) is determined. This sample is amplified and retransmitted simultaneously when further pulses are received. The aircraft actually highlights itself on the radar screen. The jamming power is steadily increased, and this continues until the replica is much stronger than the echo from the aircraft skin return. At this time, the sensitivity of the tracking radar's receiver is usually reduced in order to avoid overload. This causes the skin echo vanishes below the noise floor.

2. *Another* replica is transmitted after each of the 'dummy' skin echoes. The power of the second replica is increased while the dummy is made weaker.

3. Next, the tracker has locked on to the *delayed replica*, whereas the skin return has decreased into the noise. With respect to each of the radar's pulses, the replica is now being delayed by small, but increasing amounts of time. The range gates, of course, follow the dummy target which appears to be receding. This continues until the range gates have been moved away from the target's real position. The result is that the radar is tracking a *phantom target* and the skin return is being blanked out by the range gates.

4. Finally, the jammer is switched off and leaves the radar with just *nothing* but noise inside the window between its range gates. Break-lock was successfully achieved and the "tracking" radar needs to

switch back into a search or acquisition mode and loses time. The whole cycle will start again if the target is still within range and is reacquired.

As described above, RGPO creates only false targets which appear at greater ranges than the real target because the deceptive signal is transmitted *after* the skin echo. However, if the victim radar's PRF is constant then the time of incidence of the next radar pulse can be calculated and jamming pulses can be placed such that false targets at closer ranges are also produced.

SCINTILLATION (SCINT)

Scintillation is not an EA technique by itself, it is an implementation of an EA technique. Scintillation is simply superimposing a small, pseudo random amplitude modulation on the EA signal to make it appear more realistic to a manual operator.

SPOT JAMMING

Occurs when a jammer focuses all of its power on a single frequency. While this would severely degrade the ability to track on the jammed frequency, a frequency agile radar would hardly be affected because the jammer can only jam one frequency. While multiple jammers could possibly jam a range of frequencies, this would consume a great deal of resources to have any effect on a frequency-agile radar, and would probably still be ineffective

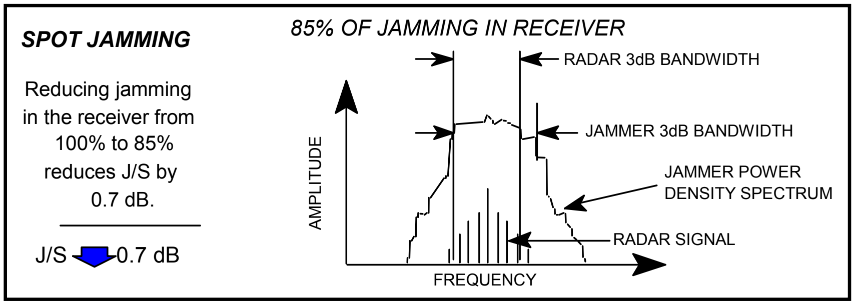
SPOT JAMMING
Reducing jamming in the receiver from 100% to 85% reduces J/S by 0.7 dB.
J/S ↓ 0.7 dB
85% OF JAMMING IN RECEIVER
RADAR 3dB BANDWIDTH
JAMMER 3dB BANDWIDTH
JAMMER POWER DENSITY SPECTRUM
RADAR SIGNAL
AMPLITUDE
FREQUENCY

Figure 7. Spot Jamming.

SWEPT JAMMING

This happens when a jammer's full power is shifted from one frequency to another. While this has the advantage of being able to jam multiple frequencies in quick succession, it does not affect them all at the same time, and thus limits the effectiveness of this type of jamming. Although, depending on the error checking in the receiver(s) this can render a wide range of receivers effectively useless.

SWEPT AMPLITUDE MODULATION (SAM)

The OOK frequency is linearly varied in a sawtooth fashion between preset frequency limits while the duty factor is held constant.

SWEPT WAVE MODULATION (SWM)

A swept wave modulation.2 (SWM.2) is essentially a swept amplitude modulation (SAM). It is a waveform that is swept between two frequencies that are usually chosen to bracket a radar's passive angle scanning rate. SWM can be either Synchronous Swept Wave Modulation.2-; (S-SWM) or Asynchronous Swept Wave Modulation.2-; (A-SWM.2-;). The modulation amplitude can be either down modulated or On-Off Keyed (OOK). The down modulating shape can be square wave, rectangular wave, linear (e.g. a sine wave), or a combination. The OOK modulating shape can be square wave or rectangular wave.

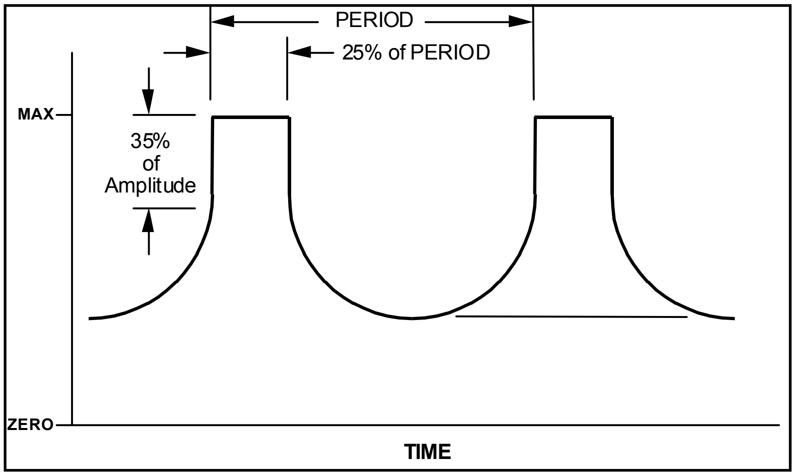

Figure 8. Typical Linear SWM Modulation.

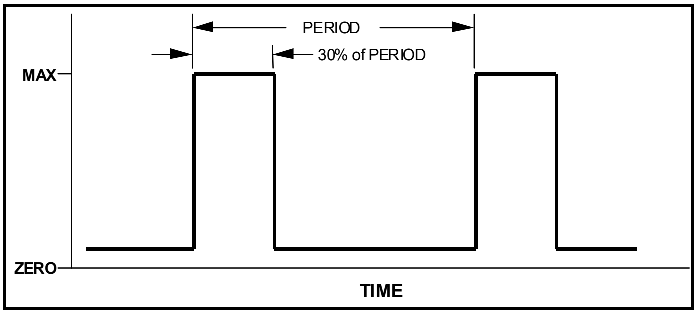

Figure 9. Rectangular SWM.

SYNCHRONOUS SWEPT WAVE MODULATION (SSWM)

SSWM is synonymous with S-SWM. S-SWM and A-SWM are essentially the same except that asynchronous means that the waveform is free running and synchronous means that when a radar scan or TWS beam can be detected, the modulation waveform is synchronized to the detected beam. For programming purposes A-SWM sets sweep limits and rate by frequency (Hz), while S-SWM sets them by period (mSec).

Active Con-Scan radars will not have a detectable modulation if the target is being tightly tracked in the center of the beam. Therefore, a SWM can "jog" the tracking sufficiently to detect the modulation and allowing subsequent synchronization of the SWM waveform.

TRACK WHILE SCAN JAMMING

The technique of shifting or "walking" EA pulses off target. Many angle jamming techniques are effective.

VELOCITY FALSE TARGETS (VFT)

VFT is a pseudo-random false Doppler target concealment technique. It is designed for use against radars that acquire Doppler targets with a bank of contiguous narrow band filters. A false Doppler target is programmed to remain in a Doppler filter long enough for the radar processing to declare it a valid target return, but not long enough for the radar processing to establish tracking.

The false Doppler target is then switched to the next pseudo-randomly selected frequency and repeated. It is intended to overload the radar processing and/or the operator's ability to identify an actual target. In the illustration, the VFTs jump around in the indicated numerical order.

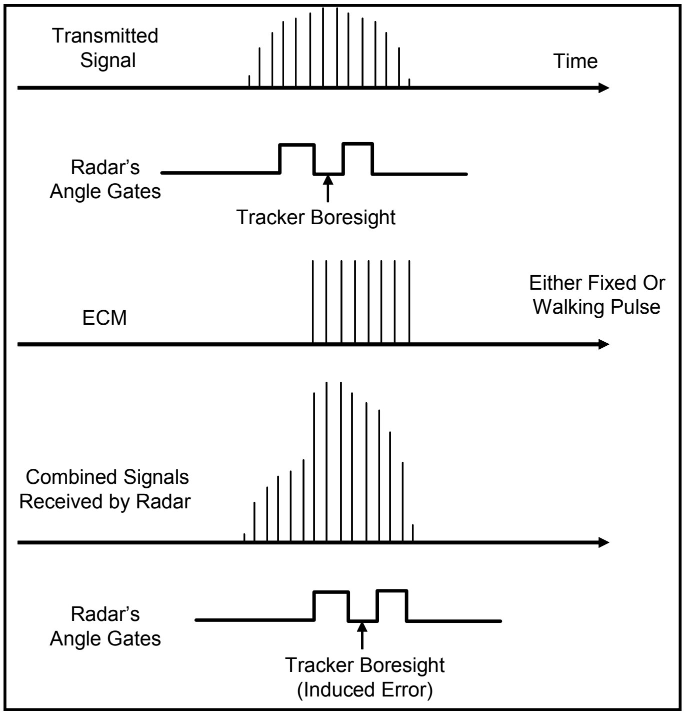
Transmitted Signal
Time
Radar's Angle Gates
Tracker Boresight
ECM
Either Fixed Or Walking Pulse
Combined Signals Received by Radar
Radar's Angle Gates
Tracker Boresight (Induced Error)

Figure 10. TWS Jamming.

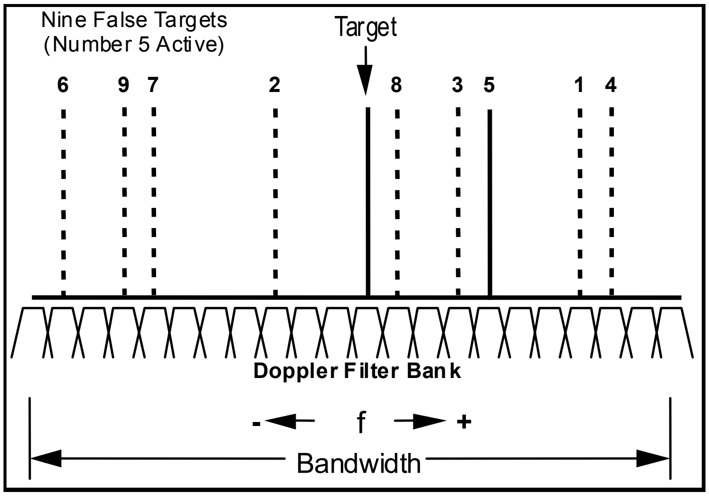
Nine False Targets (Number 5 Active)
Target
6
9
7
2
8
3
5
1
4
Doppler Filter Bank
- f +
Bandwidth

Figure 11. Velocity False Targets.

VELOCITY GATE PULL OFF

This is a method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against CW or Doppler velocity tracking radar systems. The CW or pulse doppler frequency, which is amplified and retransmitted, is shifted in frequency (velocity) to provide an apparent rate change or Doppler shift.

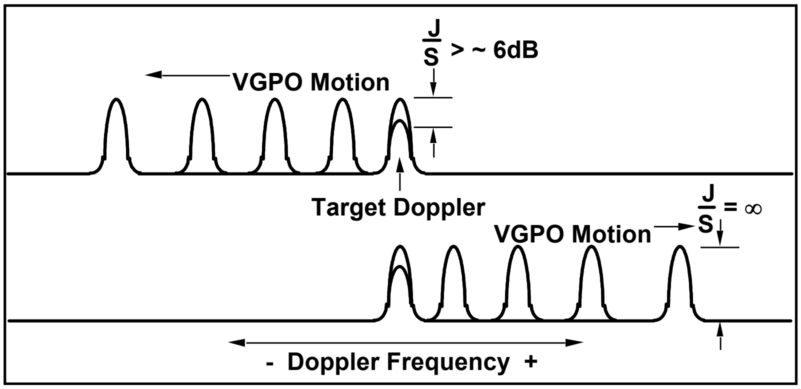

Figure 12. Velocity Gate Pull Off.

VELOCITY NOISE (VN)

VN is a coherent repeater technique. The objective is to create noise centered on a coherent radar's RF, with a noise bandwidth that is close to, or less than, the radar's bandwidth and conceal the target, or destroy target signal coherency. VN is generated by pseudo randomly stepping a frequency over the victim radar's bandwidth. The dashed RF lines represent possible frequencies and the solid line represents the frequency currently active.

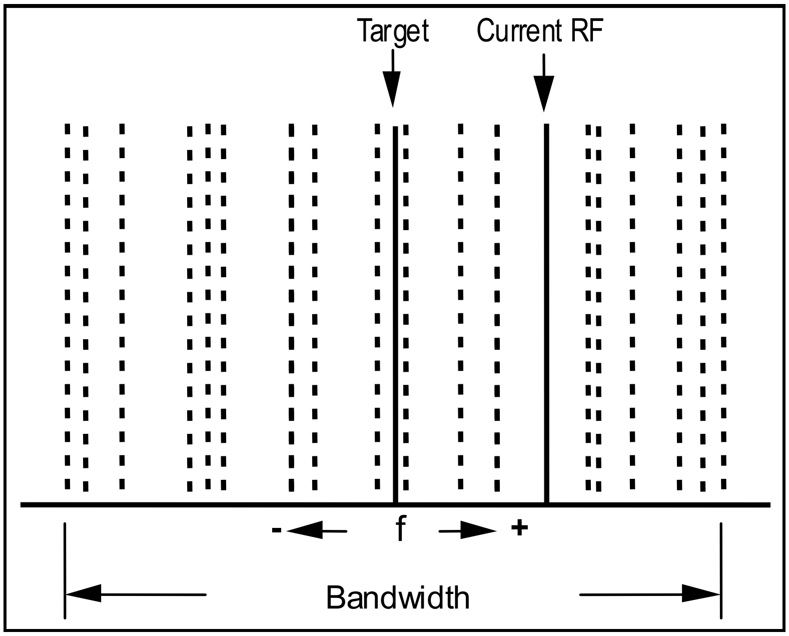

Figure 13. Velocity Noise.

This page intentionally left blank.

RADAR AND RECEIVER CHARACTERISTICS & TEST

RF Atmospheric Absorption / Ducting 5-1
Receiver Sensitivity / Noise 5-2
Receiver Types and Characteristics 5-3
Radar Modes 5-4
General Radar Display Types 5-5
IFF - Identification - Friend or Foe 5-6
Receiver Tests 5-7
Signal Sorting and Direction Finding 5-8

This page intentionally left blank.

RF ATMOSPHERIC ABSORPTION / DUCTING

Signal losses are associated with each stage of signal processing in both the transmitting and receiving portions of the system. The transmitting losses include power transmission efficiency, waveguide and antenna losses, and duplexer losses. In the receiver, losses include antenna, waveguide, RF amplifier, mixer, and IF amplifier.

In addition to these losses, energy traveling through the atmosphere suffers from atmospheric attenuation caused primarily by absorption by the gasses. For lower frequencies (below 10 GHz), the attenuation is reasonably predictable. For high frequencies in the millimeter wave range, the attenuation not only increases, but becomes more dependent upon peculiar absorbing characteristics of H_2O, O_2, and the like.

Figure 1 shows the areas of peak absorption in the millimeter wave spectrum. Figure 2 shows how the intensity of precipitation can affect atmospheric attenuation.

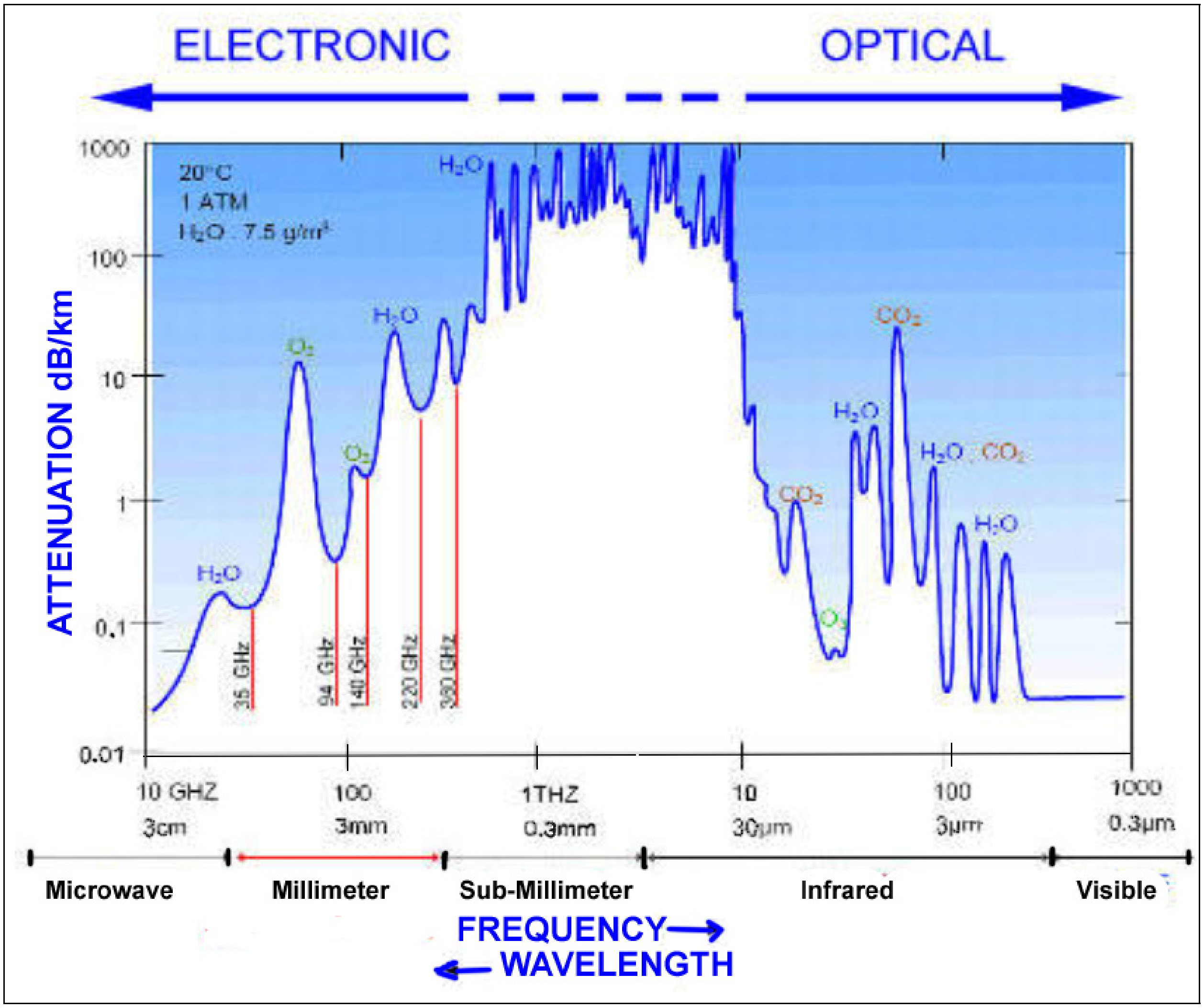

Figure 1. Atmospheric Absorption of Millimeter Waves.

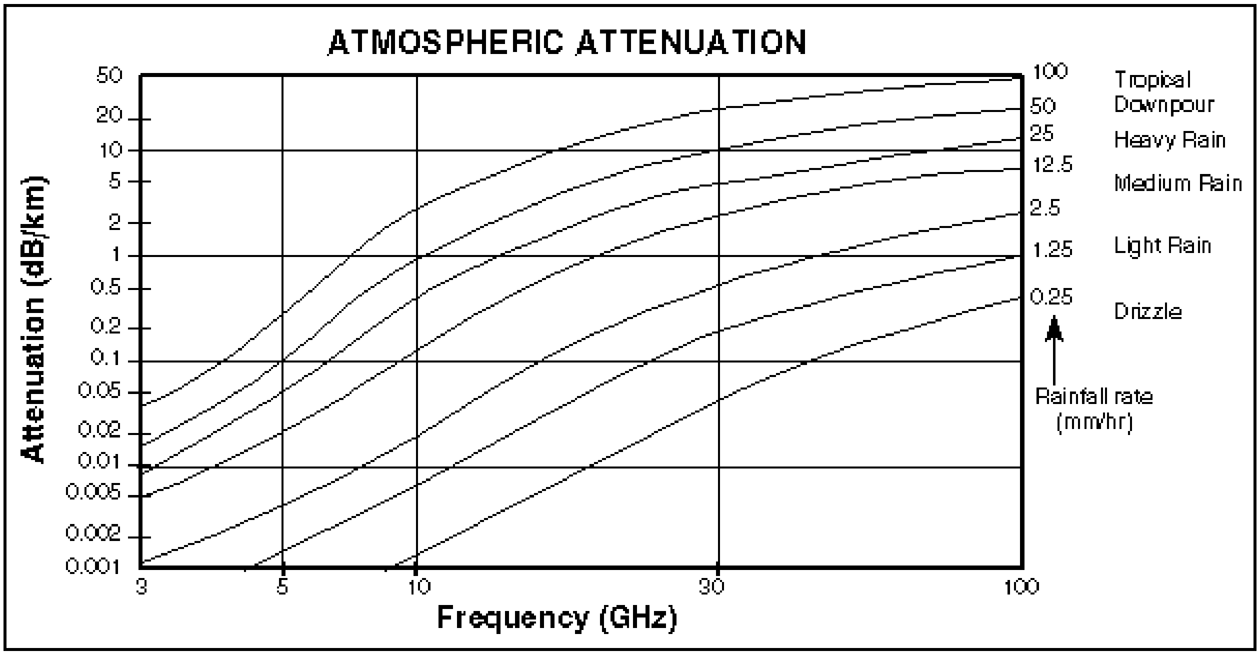

Figure 2. Atmospheric Attenuation.

Ducting is an increase in range that an electromagnetic wave will travel due to a temperature inversion of the lower atmosphere (troposphere) as shown in Figure 3. The temperature inversion forms a channel or waveguide (duct) for the waves to travel in, and they can be trapped, not attenuating as would be expected from the radar equation. Ducting may also extend range beyond what might be expected from limitations of the radar horizon (see Section 2-9).

The ducting phenomena is frequency sensitive. The thicker the duct, the lower the minimum trapped frequency.

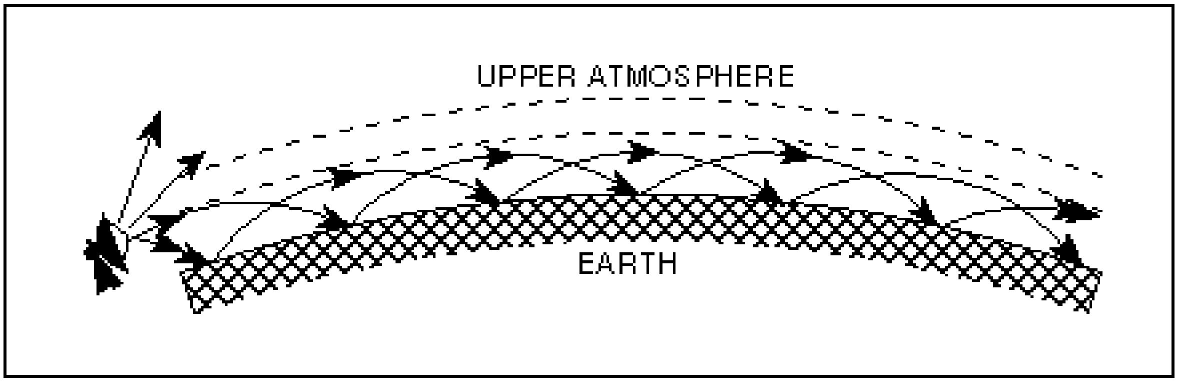

Figure 3. Ducting.

A similar occurrence takes place with ionospheric refraction, however the greatest increase in range occurs in the lower frequencies. This is familiar to amateur radio operators who are able to contact counterparts “around the world.”

RECEIVER SENSITIVITY / NOISE

RECEIVER SENSITIVITY

Sensitivity in a receiver is normally taken as the minimum input signal (S_{min}) required to produce a specified output signal having a specified signal-to-noise (S/N) ratio and is defined as the minimum signal-to-noise ratio times the mean noise power, see equation [1]. For a signal impinging on the antenna (system level) sensitivity is known as minimum operational sensitivity (MOS), see equation [2]. Since MOS includes antenna gain, it may be expressed in dBLi (dB referenced to a linear isotropic antenna). When specifying the sensitivity of receivers intended to intercept and process pulse signals, the minimum pulse width at which the specified sensitivity applies must also be stated. See the discussion of post-detection bandwidth (B_V) in Section 5-2 for significance of minimum pulsewidth in the receiver design.

$\mathbf{S_{min} = (S/N)_{min} k T_o B (NF)}$ receiver sensitivity ("black box" performance parameter) **[1]**

or $\mathbf{MOS = (S/N)_{min} k T_o B (NF)/G}$ system sensitivity i.e. the receiver is connected to an antenna **[2]**
(transmission line loss included with antenna gain)

where: S/N_{min} = Minimum signal-to-noise ratio needed to process (vice just detect) a signal
NF = Noise figure/factor
k = Boltzmann's Constant = 1.38×10^{-23} Joule/°K
T_o = Absolute temperature of the receiver input (°Kelvin) = 290°K
B = Receiver Bandwidth (Hz)
G = Antenna/system gain

We have a lower MOS if temperature, bandwidth, NF, or S/N_{min} decreases, or if antenna gain increases. For radar, missile, and EW receivers, sensitivity is usually stated in dBm. For communications and commercial broadcasting receivers, sensitivity is usually stated in micro-volts or dBμv. See Section 4-1.

There is no standard definition of sensitivity level. The term minimum operational sensitivity (MOS) can be used in place of S_{min} at the system level where aircraft installation characteristics are included. The "black box" term minimum detectable signal (MDS) is often used for S_{min} but can cause confusion because a receiver may be able to detect a signal, but not properly process it. MDS can also be confused with minimum discernable signal, which is frequently used when a human operator is used to interpret the reception results. A human interpretation is also required with minimum visible signal (MVS) and tangential sensitivity (discussed later). To avoid confusion, the terms S_{min} for "black box" minimum sensitivity and MOS for system minimum sensitivity are used in this section. All receivers are designed for a certain sensitivity level based on requirements. One would not design a receiver with more sensitivity than required because it limits the receiver bandwidth and will require the receiver to process signals it is not interested in. In general, while processing signals, the higher the power level at which the sensitivity is set, the fewer the number of false alarms which will be processed. Simultaneously, the probability of detection of a "good" (low-noise) signal will be decreased.

Sensitivity can be defined in two opposite ways, so discussions can frequently be confusing. It can be the ratio of response to input or input to response. In using the first method (most common in receiver discussions and used herein), it will be a negative number (in dBm), with the more negative being "better" sensitivity, e.g., -60 dBm is "better" than -50 dBm sensitivity. If the second method is used, the result will be a positive number, with higher being "better." Therefore the terms low sensitivity or high sensitivity can be very confusing. The terms S_{min} and MOS avoid confusion.

SIGNAL-TO-NOISE (S/N) RATIO

The Signal-to-Noise Ratio (S/N) (a.k.a. SNR) in a receiver is the signal power in the receiver divided by the mean noise power of the receiver. All receivers require the signal to exceed the noise by some amount. Usually if the signal power is less than or just equals the noise power it is not detectable. For a signal to be detected, the signal energy plus the noise energy must exceed some threshold value. Therefore, just because N is in the denominator doesn't mean it can be increased to lower the MOS. S/N is a required minimum ratio, if N is increased, then S must also be increased to maintain that threshold. The threshold value is chosen high enough above the mean noise level so that the probability of random noise peaks exceeding the threshold, and causing false alarms, is acceptably low.

Figure 1 depicts the concept of required S/N. It can be seen that the signal at time A exceeds the S/N ratio and indicates a false alarm or target. The signal at time B is just at the threshold, and the signal at time C is clearly below it. In the sample, if the temperature is taken as room temperature (T_o = 290°K), the noise power input is -114 dBm for a one MHz bandwidth. Normally S/N_{min} may be set higher than S/N shown in Figure 1 to meet false alarm specifications.

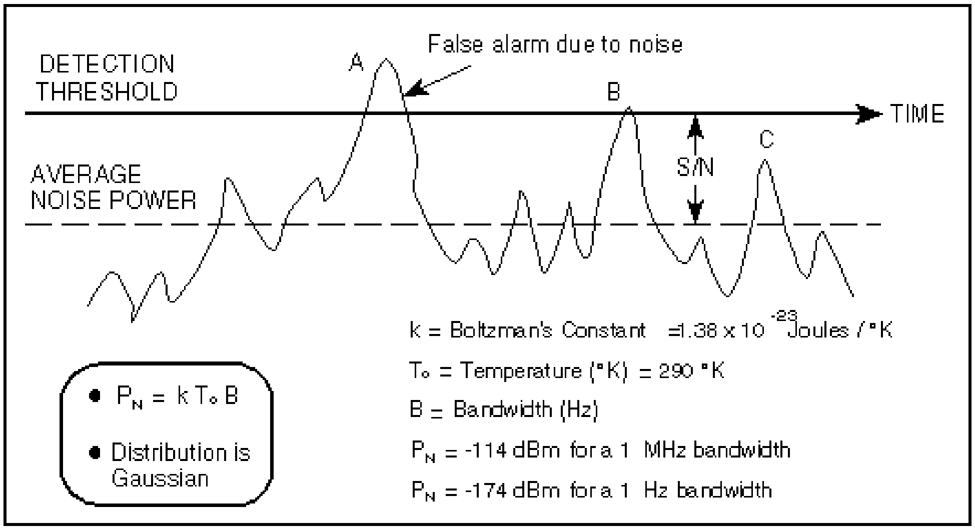
False alarm due to noise
A
DETECTION THRESHOLD
B
TIME
C
S/N
AVERAGE NOISE POWER
k = Boltzman's Constant =1.38 x 10^{-23} Joules / °K
T_o = Temperature (°K) = 290 °K
B = Bandwidth (Hz)
P_N = -114 dBm for a 1 MHz bandwidth
P_N = -174 dBm for a 1 Hz bandwidth
● $P_N = k T_o B$
● Distribution is Gaussian

Figure 1. Receiver Noise Power at Room Temperature.

The acceptable minimum Signal-to-Noise ratio (or think of it as Signal **above** Noise) for a receiver depends on the intended use of the receiver. For instance, a receiver that had to detect a single radar pulse would probably need a higher minimum S/N than a receiver that could integrate a large number of radar pulses (increasing the total signal energy) for detection with the same probability of false alarms. Receivers with human operators using a video display may function satisfactorily with low minimum S/N because a skilled operator can be very proficient at picking signals out of a noise background. As shown in Table 1, the setting of an acceptable minimum S/N is highly dependent on the required characteristics of the receiver and of the signal.

Table 1. Typical Minimum S/N Required.

Skilled Operator	Auto-Detection	Auto-detection with Amplitude, TOA, and Frequency Measurements	AOA Phase Interferometer	AOA Amplitude Comparison
3 to 8 dB	10 to 14 dB	14 to 18 dB	14 to 18 dB	16 to 24 dB

A complete discussion of the subject would require a lengthy dissertation of the probability and statistics of signal detection, which is beyond the scope of this handbook, however a simplified introduction follows. Let's assume that we have a receiver that we want a certain probability of detecting a single pulse with a specified false alarm probability. We can use Figure 2 to determine the required signal-to-noise ratio.

S/N EXAMPLE

If we are given that the desired probability of detecting a single pulse (P_d) is 98%, and we want the false alarm rate (P_n) to be no more than 10^{-3}, then we can see that S/N must be 12 dB (see Figure 2).

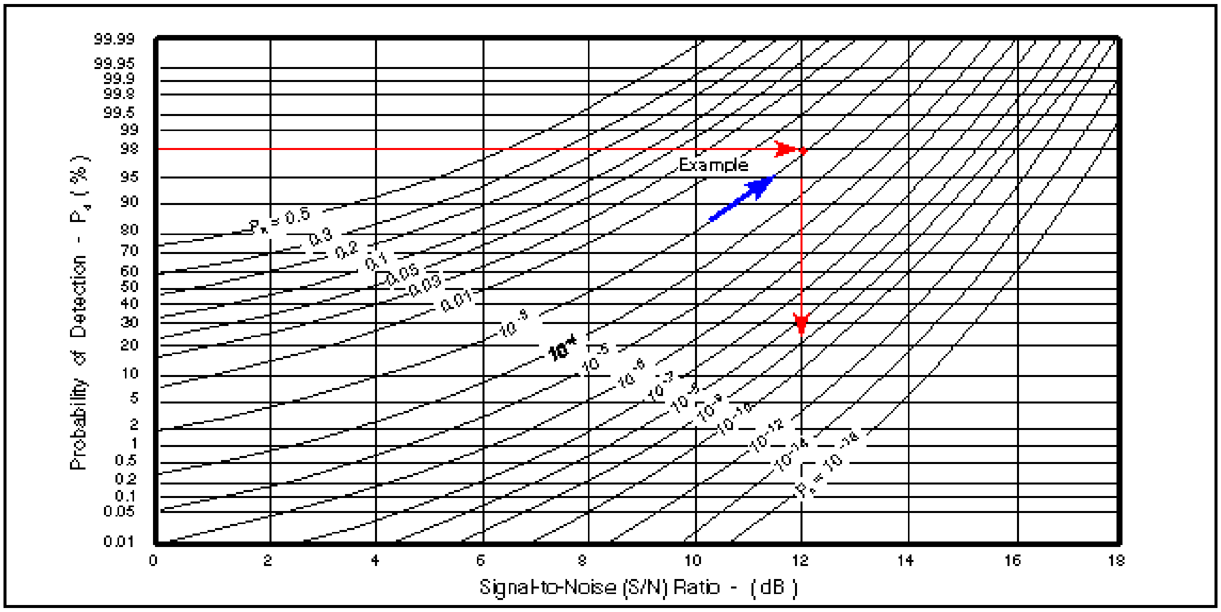

Figure 2. Nomograph of Signal-to-Noise (S/N) Ratio as a Function of Probability of Detection (P_d) and Probability of False Alarm Rate (P_n).

MAXIMUM DETECTION RANGE (ONE-WAY)

From Section 4-3, the one way signal strength from a transmitter to a receiver is:

$$S\ (or\ P_R) = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 R^2}$$

For calculations involving receiver sensitivity the "S" can be replaced by S_{min}. Since $S_{min} = (S/N)_{min}\, kT_oB(NF)$, given by equation [1], the one-way radar equation can be solved for any of the other variables in terms of receiver parameters. In communication, radar, and electronic warfare applications, you might need to solve for the maximum range (R_{max}) where a given radar warning receiver could detect a radiated signal with known parameters. We would then combine and rearrange the two equations mentioned to solve for the following one-way equation:

$$R_{max} \cong \sqrt{\frac{P_t\ G_t\ G_r\ \lambda^2}{(4\pi)^2\,(S/N)_{min}\,k\,T_o\,B\,(NF)}} \quad or \quad \sqrt{\frac{P_t\ G_t\ G_r\ c^2}{(4\pi f)^2\,(S/N)_{min}\,k\,T_o\,B\,(NF)}} \quad or \quad \sqrt{\frac{P_t\ G_t\ A_e}{4\pi\,(S/N)_{min}\,k\,T_o\,B\,(NF)}} \qquad \textbf{[3]}$$

We could use standard room temperature of 290° K as T_o, but NF would have to be determined as shown later.

In this calculation for receiver R_{max} determination, P_t, G_t, and λ are radar dependent, while G_r, S/N_{min}, NF, and B are receiver dependent factors.

Equation [3] relates the maximum detection range to bandwidth (B). The effects of the measurement bandwidth can significantly reduce the energy that can be measured from the peak power applied to the receiver input. Additional bandwidth details are provided in Sections 4-4, 4-7, and in other parts of this section.

NOISE POWER, kT_oB

Thermal noise is spread more or less uniformly over the entire frequency spectrum. Therefore the amount of noise appearing in the output of an ideal receiver is proportional to the absolute temperature of the receiver input system (antenna etc) times the bandwidth of the receiver. The factor of proportionality is Boltzmann's Constant.

Mean noise power of ideal receiver = kT_oB = P_N (Watts)
Mean noise power of a real receiver = $(NF)kT_oB$ (Watts)

The convention for the temperature of T_o is set by IEEE standard to be 290°K, which is close to ordinary room temperature. So, assuming $T_o = 290°K$, and for a bandwidth B = 1 Hz, $kT_oB = 4x10^{-21}$ W = -204 dBW = -174 dBm.

For any receiver bandwidth, multiply $4x10^{-21}$ W by the bandwidth in Hz, or if using dB;

$$10 \log kT_oB = -174 \text{ dBm} + 10 \text{ Log (actual BW in Hz)} \quad \text{or} \quad -114 \text{ dBm} + 10 \text{ Log (actual BW in MHz)}$$

and so on, as shown by the values in Table 2.

Table 2. Sample Noise Power Values (kT_oB).

Bandwidth	Bandwidth Ratio (dB)	Watts	dBW	dBm
1 Hz	0	$4x10^{-21}$	-204	-174
1 kHz	30	$4x10^{-18}$	-174	-144
1 MHz	60	$4x10^{-15}$	-144	-114
1 GHz	90	$4x10^{-12}$	-114	-84

Typical values for maximum sensitivity of receivers would be:

RWR	-65 dBm
Pulse Radar	-94 dBm
CW Missile Seeker	-138 dBm

If antenna contributions are ignored (see note in Table 4) for a CW receiver with a 4 GHz bandwidth, the ideal mean noise power would be -174 dBm + 10 Log($4x10^9$) = -174 dBm + 96 dB = -78 dBm. A skilled operator might only be able to distinguish a signal 3 dB above the noise floor (S/N=3 dB), or -75 dBm. A typical radar receiver would require a S/N of 3 to 10 dB to distinguish the signal from noise, and would require 10 to 20 dB to track. Auto tracking might require a S/N of approximately 25 dB, thus, a receiver may only have sufficient sensitivity to be able to identify targets down to -53 dBm. Actual pulse receiver detection will be further reduced due to sin x/x frequency distribution and the effect of the measurement bandwidth as discussed in Sections 4-4 and 4-7. Integration will increase the S/N since the signal is coherent and the noise is not.

Noise Bandwidth

Equivalent Noise Bandwidth (B_N) - Set by minimum pulse width or maximum modulation bandwidth needed for the system requirements. A choice which is available to the designer is the relationship of pre- and post-detection bandwidth. Pre-detection bandwidth is denoted by B_{IF}, while post-detection is denoted B_V, where V stands for video. The most affordable approach is to set the post-detection filter equal to the reciprocal of the minimum pulse width, then choose the pre-detection passband to be as wide as the background interference environment will allow. Recent studies suggest that pre-detection bandwidths in excess of 100 MHz will allow significant loss of signals due to “pulse-on-pulse” conditions. Equations [4] and [5] provide B_N relationships that don’t follow the Table 3 rules of thumb.

Table 3. Rules of Thumb for B_N a.k.a. B (Doesn’t apply for S/N between 0 and 10 to 30 dB).

S/N out	Linear Detector	Square Law Detector
High S/N (>15 to 20 dB)	$B_N = B_V$ (> 20 to 30 dB)	$B_N = 4\,B_V$ (> 10 to 15 dB)
Low S/N (< 0 dB)	$B_N = \sqrt{(2\,B_{IF}\,B_V - B_V^2)\,/\,4\,(S/N)_{out}}$	$B_N = \sqrt{(2\,B_{IF}\,B_V - B_V^2)\,/\,(S/N)_{out}}$

For a square law detector: (1)

$$B_N = B_V\left[\,2 + \sqrt{4 + \frac{(2\,B_{IF}\,/\,B_V) - 1}{(S/N)_{out}}}\,\right] \qquad \textbf{[4]}$$

At high $(S/N)_{out}$, the $1/(S/N_{out})$ term goes to zero and we have: $B_N = B_V\,[\,2 + \sqrt{4}\,] = 4\,B_V$

At low $(S/N)_{out}$, the $1/(S/N_{out})$ term dominates, and we have:

$$B_N = B_V\sqrt{\frac{(2\,B_{IF}\,/\,B_V) - 1}{(S/N)_{out}}} = \sqrt{\frac{2\,B_{IF}\,B_V - B_V^2}{(S/N)_{out}}}$$

For a linear detector: (1)

$$B_N = \frac{B_V}{2} + \frac{1}{4}\bullet\sqrt{B_V\left(4\,B_V + \frac{H^2(2\,B_{IF} - B_V)}{(S/N)_{out}}\right)} \qquad \textbf{[5]}$$

H is a hypergeometric (statistical) function of $(S/N)_{in}$

H = 2 for $(S/N)_{in} << 1$
H = 1 for $(S/N)_{in} >> 1$

At high $(S/N)_{out}$, the $1/(S/N_{out})$ term goes to zero and we have: $B_N = \frac{B_V}{2} + \frac{1}{4}\sqrt{B_V\,(4\,B_V)} = B_V$

At low $(S/N)_{out}$, the $1/(S/N_{out})$ term dominates, and we have:

$$B_N = \frac{1}{4} \bullet \sqrt{\frac{B_V H^2 (2 B_{IF} - B_V)}{(S/N)_{out}}} = \sqrt{\frac{2 B_{IF} B_V - B_V^2}{4 (S/N)_{out}}}$$

Note (1): From Klipper, "Sensitivity of Crystal Video Receivers With RF Pre-amplification," *The Microwave Journal*, August 1965.

TRADITIONAL "RULE OF THUMB" FOR NARROW BANDWIDTHS (Radar Receiver Applications)
Required IF Bandwidth for Matched Filter Applications:

$$B_{IF} = \frac{1}{PW_{\min}} \quad \text{Where :} \quad \begin{array}{l} B_{IF} = \text{Pre - detection RF or IF bandwidth} \\ PW_{\min} = \text{Specified minimum pulse width} = \tau \end{array}$$

Matched filter performance gives maximum probability of detection for a given signal level, but: (1) Requires perfect centering of signal spectrum with filter bandwidth, (2) Time response of matched pulse does not stabilize at a final value, and (3) Out-of-band splatter impulse duration equals minimum pulse width. As a result, EW performance with pulses of unknown frequency and pulse width is poor.

Required Video Bandwidth Post - Detection
Traditional " Rule of Thumb"

$$B_V = \frac{0.35}{PW_{\min}} \quad \text{Where : } B_V = \text{Post - detection bandwidth}$$

Some authors define B_V in terms of the minimum rise time of the detected pulse, i.e., $B_V = (0.35 \text{ to } 0.5)/t_r$ min, where t_r = rise time.

REVISED "RULE OF THUMB" FOR WIDE BANDWIDTHS (Wideband Portion of RWRs)

$$B_{IF} = \frac{2 \text{ to } 3}{PW_{\min}} \quad \text{and} \quad B_V = \frac{1}{PW_{\min}}$$

The pre-detection bandwidth is chosen based upon interference and spurious generation concerns. The post-detection bandwidth is chosen to "match" the minimum pulse width. This allows (1) Half bandwidth mistuning between signal and filter, (2) Half of the minimum pulse width for final value stabilization, and (3) The noise bandwidth to be "matched" to the minimum pulse width. As a result, there is (1) Improved EW performance with pulses of unknown frequency and pulse width, (2) Measurement of in-band, but mistuned pulses, and (3) Rejection of out-of-band pulse splatter.

NOISE FIGURE / FACTOR (NF)

Electrical noise is defined as electrical energy of random amplitude, phase, and frequency. It is present in the output of every radio receiver. At the frequencies used by most radars, the noise is generated primarily within the input stages of the receiver system itself (Johnson Noise). These stages are not inherently noisier than others, but noise generated at the input and amplified by the receiver's full gain greatly exceeds the noise generated further along the receiver chain. The noise performance of a receiver is described by a figure of merit called the noise figure (NF). The term noise factor is synonymous, with some authors using the term "factor" for numeric and "figure" when using dB notation. (The notation "F_n" is also sometimes used instead of "NF.") The noise figure is defined as:

$$NF = \frac{Noise\ output\ of\ actual\ receiver}{Noise\ output\ of\ ideal\ receiver} = \frac{N_{out}}{GN_{in}} \quad or\ in\ dB: \quad 10\ \text{Log}\left[\frac{Noise\ output\ of\ actual\ receiver}{Noise\ output\ of\ ideal\ receiver}\right] = 10\ \log \frac{N_{out}}{GN_{in}}$$

A range of NF values is shown in Table 4.

Table 4. Typical Noise Figure / Factor Value.	Decimal	dB
Passive lossy network (RF transmission line, attenuator, etc.) Example: 20 dB attenuator (gain = 0.01)	Same as reciprocal of gain value ex: 100	Same as dB value ex: 20
Solid State Amplifier (see manufacturers specifications)	4	6
Traveling Wave Tube (see manufacturers specifications)	10 to 100	10 to 20
Antennas (Below ≈ 100 MHz, values to 12 dB higher if pointed at the sun) Note: Unless the antenna is pointed at the sun, its negligible NF can be ignored. Antenna gain is not valid for NF calculations because the noise is received in the near field.	1.012 to 1.4	0.05 to 1.5

An ideal receiver generates no noise internally. The only noise in its output is received from external sources. That noise has the same characteristics as the noise resulting from thermal agitation in a conductor. Thermal agitation noise is caused by the continuous random motion of free electrons which are present in every conductor. The amount of motion is proportional to the conductor's temperature above absolute zero. For passive lossy networks, the noise factor equals the loss value for the passive element:

$$NF = \frac{N_{out}}{G\,N_{in}} = \frac{kTB}{\frac{1}{L}kTB} = L \qquad \begin{array}{c} Where\ L = Ratio\ Value\ of\ Attenuation \\ i.e.\ For\ a\ 3\ dB\ attenuator, G = 0.5\ and\ L = 2 \\ \therefore NF = 2\ \ and\ \ 10\log NF = 3\ dB \end{array}$$

A typical series of cascaded amplifiers is shown in Figure 3.

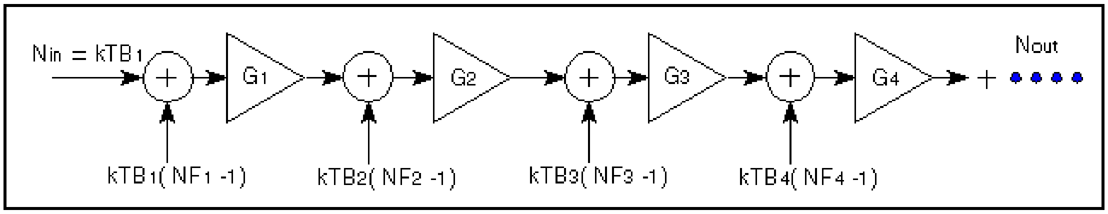

Figure 3. Noise Factors for Cascaded Amplifiers (NF_{CA}).

Loss (negative gain) can be used for the gain value of attenuators or transmission line loss, etc to calculate the noise out of the installation as shown in the following equation:

$$N_{out} = N_{in}\,G\,NF_{CA} = k\,T\,B_1(\,G_1G_2G_3\ldots)\left(NF_1 + \frac{B_2(NF_2 - 1)}{B_1G_1} + \frac{B_3(NF_3 - 1)}{B_1G_1G_2} + \frac{B_4(NF_4 - 1)}{B_1G_1G_2G_3} + \ldots\right) \quad (ratio\ form) \qquad \textbf{[6]}$$

If the bandwidths of the amplifiers are the same, equation [6] becomes:

$$N_{out} = N_{in}\, G\, NF_{CA} = k\, T\, B\, (\, G_1 G_2 G_3 \ldots) \left(NF_1 + \frac{NF_2 - 1}{G_1} + \frac{NF_3 - 1}{G_1 G_2} + \frac{NF_4 - 1}{G_1 G_2 G_3} + \ldots \right) \quad \textit{(ratio form)} \qquad [7]$$

Pre-amplifier Location Affects Receiver Input Noise

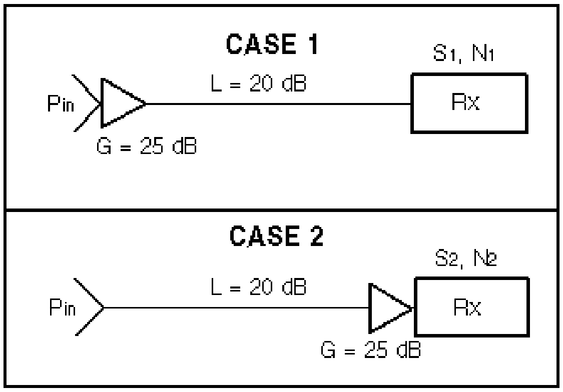

Figure 4. Pre-Amp S/N.

As shown in Figure 4, if a 2 to 12 GHz receiver installation doesn't have enough sensitivity, it is best to install an additional amplifier closer to the antenna (case 1) instead of closer to the receiver (case 2). In both cases, the line loss (L) and the amplifier gain (G) are the same, so the signal level at the receiver is the same. For case 1, $S_1 = P_{in} + G - L$. In case 2, $S_2 = P_{in} - L + G$, so $S_1 = S_2$. The noise generated by the passive transmission line when measured at the receiver is the same in both cases. However, the noise generated inside the amplifier, when measured at the receiver input, is different.

For this example, case 2 has a noise level at the input to the receiver which is 19.7 dB higher than case 1 (calculations follow later).

Table 5a	Case 1 Gain		Case 1 NF	
	Amp	L	Amp	L
dB	25	-20	6 *	20
ratio	316.2	0.01	4 *	100

Table 5b	Case 2 Gain		Case 2 NF	
	L	Amp	L	Amp
dB	-20	25	20	6 *
ratio	0.01	316.2	100	4 *

* Amplifier NF value from Table 4.

Using equation [3] and the data in Tables 5a and 5b, the noise generated by the RF installation is shown in Tables 6a and 6b (the negligible noise contribution from the antenna is the same in both cases and is not included) (also see notes contained in Table 4):

Table 6a. Case 1	**Table 6b.** Case 2
$G(NF) = 316.2\,(0.01)\left(4 + \frac{100 - 1}{316.7} \right) = 13.64$	$G(NF) = 0.01\,(316.2)\left(100 + \frac{4 - 1}{0.01} \right) = 1264.8$
10 log G(NF) = 11.34 dB	10 log G(NF) = 31 dB
Noise at receiver:	
$N_{out\,1}$ = -74 dBm + 11.34 dB = -62.7 dBm	$N_{out\,2}$ = -74 dBm + 31 dB = -43 dBm

$N_{out\,2} - N_{out\,1}$ = 19.7 dB. The input noise of -74 dBm was calculated using 10 log (kTB), where B = 10 GHz.

Note that other tradeoffs must be considered: (1) greater line loss between the antenna and amplifier improves (decreases) VSWR as shown in Section 6-2, and (2) the more input line loss, the higher the input signal can be before causing the pre-amplifier to become saturated (mixing of signals due to a saturated amplifier is addressed in Section 5-7).

Combining Receive Paths Can Reduce Sensitivity

If a single aircraft receiver processes both forward and aft signals as shown in Figure 5, it is desirable to be able to use the receiver's full dynamic range for both directions. Therefore, one needs to balance the gain, so that a signal applied to the aft antenna will reach the receiver at the same level as if it was applied to the forward antenna.

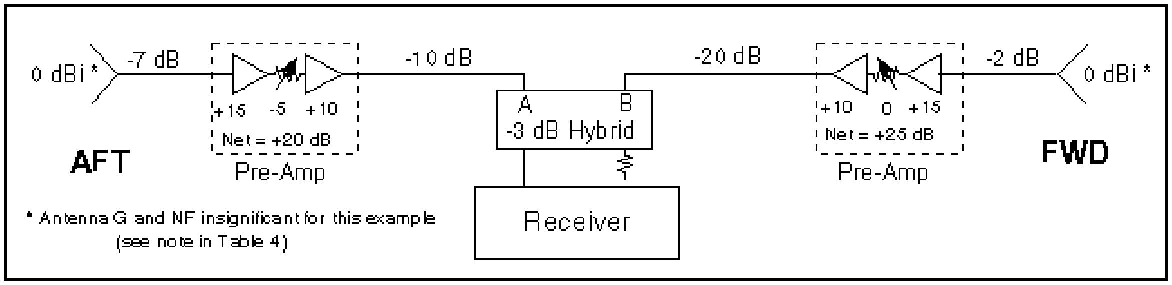

Figure 5. Example of Pre-Amplifier Affecting Overall Gain / Sensitivity.

Common adjustable preamplifiers can be installed to account for the excessive transmission line loss. In this example, in the forward installation, the level of the signal at the receiver is the same as the level applied to the antenna. Since the aft transmission line has 5 dB less attenuation, that amount is added to the preamplifier attenuator to balance the gain. This works fine for strong signals, but not for weaker signals. Because there is less loss between the aft preamplifier and the receiver, the aft noise dominates and will limit forward sensitivity. If the bandwidth is 2-12 GHz, and if port A of the hybrid is terminated by a perfect 50Ω load, the forward noise level would be -65.3 dBm. If port B is terminated, the aft noise level would be -60.4 dBm. With both ports connected, the composite noise level would be -59.2 dBm (convert to mw, add, then convert back to dBm). For this example, if the aft preamplifier attenuation value is changed to 12 dB, the gain is no longer balanced (7 dB extra loss aft), but the noise is balanced, i.e., forward = -65.6 dBm, aft = -65.3 dBm, and composite -62.4 dBm. If there were a requirement to see the forward signals at the most sensitive level, extra attenuation could be inserted in the aft preamplifier. This would allow the forward noise level to predominate and result in greater forward sensitivity where it is needed. Calculations are provided in Tables 7 and 8.

Table 7. Summary of Gain and NF Values for Figure 5 Components.

		Aft					Fwd				
		RF Line	Amp	Attn	Amp	RF Line & hybrid	RF Line	Amp	Attn	Amp	RF Line & hybrid
Gain	dB	-7	15	-5	10	-13	-2	15	0	10	-23
	ratio	0.2	31.6	0.32	10	0.05	0.63	31.6	0	10	0.005
NF	dB	7	6	5	6	13	2	6	0	6	23
	ratio	5	4	3.16	4	20	1.585	4	0	4	200

Aft NF = 22.79 therefore 10 log NF = 13.58 dB. Input noise level = -74 dBm + 13.58 dB = -60.42 dBm ≅ -60.4 dBm
Fwd NF = 7.495 therefore 10 log NF = 8.75 dB. Input noise level = -74 dBm + 8.75 dB = -65.25 dBm ≅ -65.3 dBm
The composite noise level at the receiver = -59.187 dBm ≅ -59.2 dBm.

Table 8. Effect of Varying the Attenuation (shaded area) in the Aft Preamplifier Listed in Table 7.							
Aft Attn NF	Aft Attn Gain	Aft Noise	Fwd Noise	Composite Noise	Min Signal Received ***	Aft Input	Fwd Input
0 dB	0 dB	-55.8 dBm	-65.3 dBm	-55.4 dBm	-43.4 dBm	-48.4 dBm	-43.4 dBm
5	-5	-60.4	-65.3	-59.2	-47.2 *	-47.2 *	-47.2 *
10	-10	-64.4	-65.3	-61.8	-49.8	-44.8	-49.8
12	-12	-65.6 **	-65.3 **	-62.4	-50.4	-43.4	-50.4
15	-15	-67.1	-65.3	-63.1	-51.1	-41.1	-51.1

* Gain Balanced ** Noise Balanced *** S/N was set at 12 dB

TANGENTIAL SENSITIVITY

Tangential sensitivity (TSS) is the point where the top of the noise level with no signal applied is level with the bottom of the noise level on a pulse as shown in Figure 6. It can be determined in the laboratory by varying the amplitude of the input pulse until the stated criterion is reached, or by various approximation formulas.

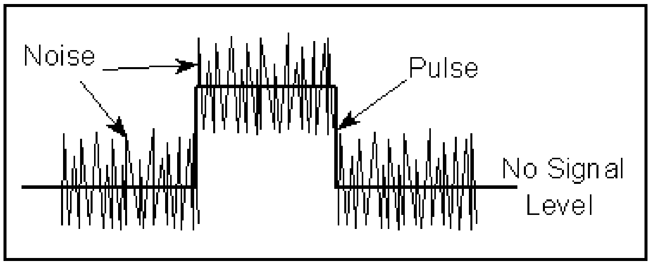

Figure 6. Tangential Sensitivity.

The signal power is nominally 8±1 dB above the noise level at the TSS point. TSS depends on the RF bandwidth, the video bandwidth, the noise figure, and the detector characteristic.

TSS is generally a characteristic associated with receivers (or RWRs), however the TSS does not necessarily provide a criterion for properly setting the detection threshold. If the threshold is set to TSS, then the false alarm rate is rather high. Radars do not operate at TSS. Most require a more positive S/N for track (> 10 dB) to reduce false detection on noise spikes.

SENSITIVITY CONCLUSION

When all factors effecting system sensitivity are considered, the designer has little flexibility in the choice of receiver parameters. Rather, the performance requirements dictate the limit of sensitivity which can be implemented by the EW receiver.

1. Minimum Signal-to-Noise Ratio (S/N) - Set by the accuracy which you want to measure signal parameters and by the false alarm requirements.

2. Total Receiver Noise Figure (NF) - Set by available technology and system constraints for RF front end performance.

3. Equivalent Noise Bandwidth (B_N) - Set by minimum pulse width or maximum modulation bandwidth needed to accomplish the system requirements. A choice which is available to the designer is the relationship of pre- (B_{IF}) and post-detection (B_V) bandwidth. The most affordable approach is to set the post-detection filter equal to the reciprocal of the minimum pulse width, then choose the pre-detection passband to be as wide as the background interference environment will allow. Recent studies suggest that pre-detection bandwidths in excess of 100 MHz will allow significant loss of signals due to “pulse-on-pulse” conditions.

4. Antenna Gain (G) - Set by the needed instantaneous FOV needed to support the system time to intercept requirements.

This page intentionally left blank.

RECEIVER TYPES AND CHARACTERISTICS

Besides the considerations of noise and noise figure, the capabilities of receivers are highly dependent on the type of receiver design. Most receiver designs are trade-offs of several conflicting requirements. This is especially true of the Electronic Support Measures (ESM) receivers used in Electronic Warfare.

This section consists of a figure and tables that provide a brief comparison of various common ESM receiver types. Figures 1 and 2 show block diagrams of common ESM receivers. Table 1 is a comparison of major features of receivers. Table 2 shows the receiver types best suited for various types of signals and Tables 3 and 4 compare several direction of arrival (DOA) and emitter location techniques. Table 5 shows qualitative and quantitative comparisons of receiver characteristics.

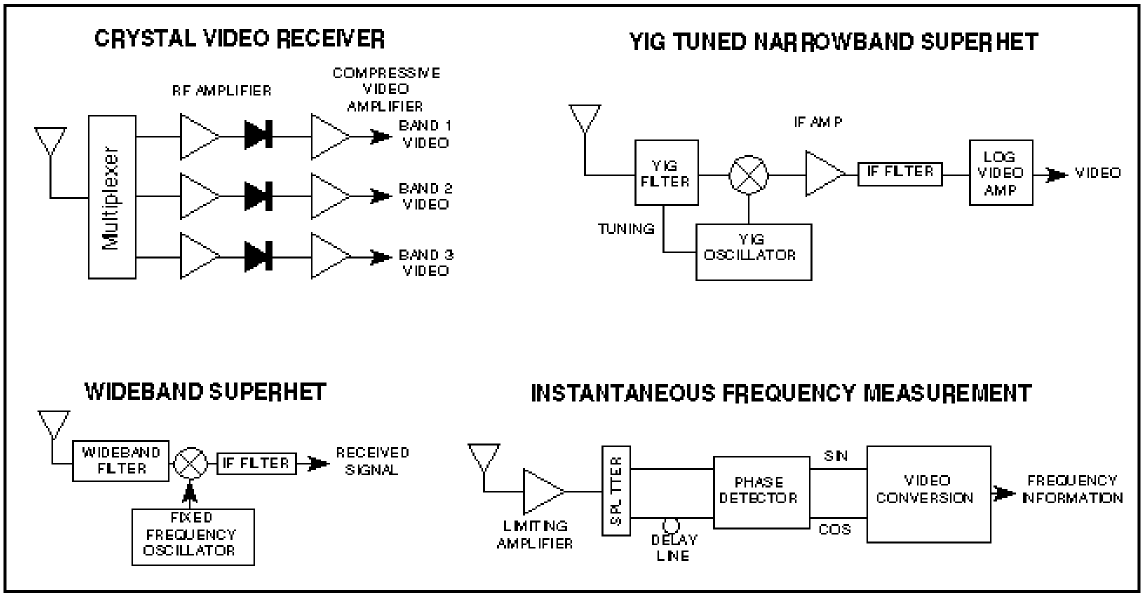
CRYSTAL VIDEO RECEIVER
RF AMPLIFIER
COMPRESSIVE VIDEO AMPLIFIER
Multiplexer
BAND 1 VIDEO
BAND 2 VIDEO
BAND 3 VIDEO
YIG TUNED NARROWBAND SUPERHET
IF AMP
YIG FILTER
IF FILTER
LOG VIDEO AMP
VIDEO
TUNING
YIG OSCILLATOR
WIDEBAND SUPERHET
WIDEBAND FILTER
IF FILTER
RECEIVED SIGNAL
FIXED FREQUENCY OSCILLATOR
INSTANTANEOUS FREQUENCY MEASUREMENT
LIMITING AMPLIFIER
SPLITTER
DELAY LINE
PHASE DETECTOR
SIN
COS
VIDEO CONVERSION
FREQUENCY INFORMATION

Figure 1. Common ESM Receiver Block Diagrams.

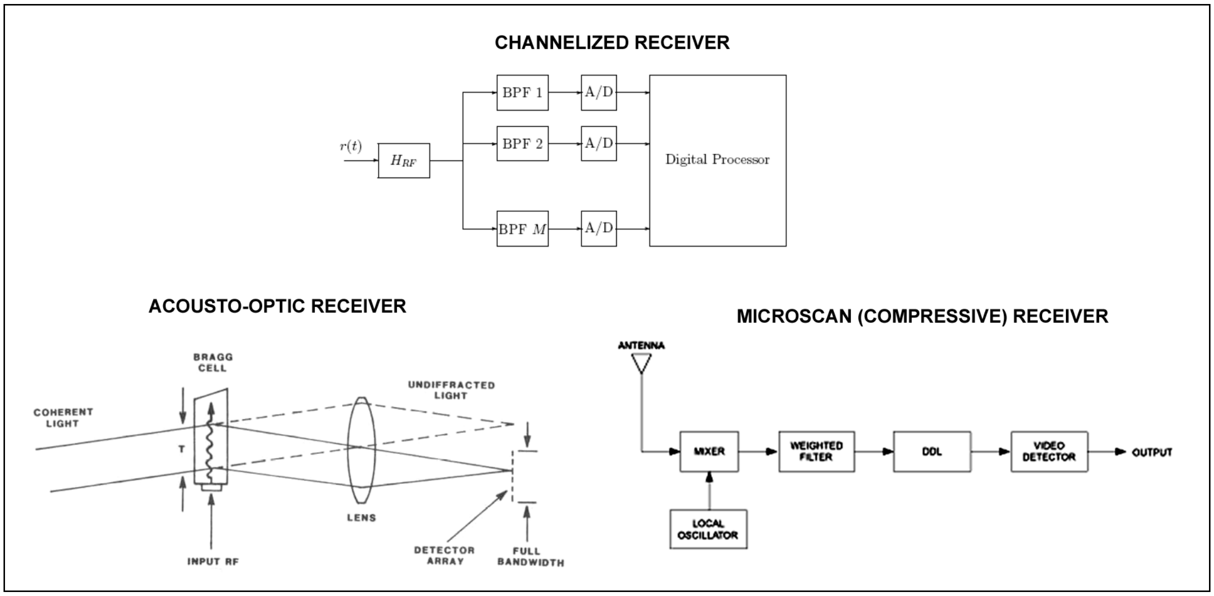
CHANNELIZED RECEIVER
r(t)
H_RF
BPF 1
BPF 2
BPF M
A/D
A/D
A/D
Digital Processor
ACOUSTO-OPTIC RECEIVER
BRAGG CELL
COHERENT LIGHT
T
UNDIFFRACTED LIGHT
LENS
INPUT RF
DETECTOR ARRAY
FULL BANDWIDTH
MICROSCAN (COMPRESSIVE) RECEIVER
ANTENNA
MIXER
WEIGHTED FILTER
DDL
VIDEO DETECTOR
OUTPUT
LOCAL OSCILLATOR

Figure 2. Common ESM Receiver Block Diagrams (Continued).

Table 1. Comparison of Major Features of Receivers.

Receiver	Advantages	Disadvantages	Principal Applications
Wideband crystal video	Simple, inexpensive, instantaneous, High POI in frequency range	No frequency resolution Poor sensitivity and Poor simultaneous signal performance	RWR
Tuned RF Crystal Video	Simple, Frequency measurement Higher sensitivity than wideband	Slow response time Poor POI	Option in RWR, Frequency measurement in hybrid
IFM	Relatively simple Frequency resolution Instantaneous, high POI	Cannot sort simultaneous signals Relatively poor sensitivity	Shipboard ESM, Jammer power management, SIGINT equipment
Narrow-band scanning Superhet	High sensitivity Good frequency resolution Simultaneous signals don’t interfere	Slow response time Poor POI Poor against frequency agility	SIGINT equipment Air and ship ESM Analysis part of hybrid
Wide-band Superhet	Better response time and POI	Spurious signals generated Poorer sensitivity	Shipboard ESM Tactical air warning
Channelized	Wide bandwidth, Near instantaneous, Moderate frequency resolution	High complexity, cost; Lower reliability; limited sensitivity	SIGINT equipment Jammer power management
Microscan	Near instantaneous, Good resolution and dynamic range, Good simultaneous signal capability	High complexity, Limited bandwidth No pulse modulation information Critical alignment	SIGINT equipment Applications for fine freq analysis over wide range
Acousto-optic	Near instantaneous, Good resolution, Good simultaneous signal capability Good POI	High complexity; new technology	

Note: The Microscan receiver is also known as a compressive receiver

Table 2. Receiver Types vs. Signal Types.

Signal Type	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
CW	Special design for CW	Special design for CW	Yes, but interferes with pulsed reception	Yes	Yes	Yes	Yes	Yes
Pulsed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Frequency	No	No	No	Yes, but won’t recognize as same source	No	Yes	Yes	Yes
Frequency Agile	Yes, doesn’t measure frequency	No	Yes	No	Yes (within passband)	Yes	Yes	No/Yes, depending on readout time
PRI Agile	Yes	Yes	Yes	No/Yes, depending on scan rate	Yes	Yes	No/Yes, imprecision in TOA	No/Yes, depending on readout time
Chirped	Yes, within acceptance BW	No	Yes	No/Yes, depending on BW	Yes	Yes (reduced sensitivity)	No/Yes, depending on scan rate	Yes (reduced sensitivity)
Spread Spectrum	Yes, within acceptance BW	No	Yes	No	No/Yes, depending on BW	Yes (reduced sensitivity)	Yes (reduced sensitivity)	Yes (reduced sensitivity)

Table 3. Direction of Arrival Measurement Techniques.

	Amplitude Comparison	Phase Interferometer
Sensor Configuration	Typically 4 to 6 Equal Spaced Antenna Elements for 360° Coverage	2 or more RHC or LHC Spirals in Fixed Array
DF Accuracy	$DF_{ACC} \approx \frac{\Theta_{bW}^{2}\,\Delta C_{dB}}{24\,S}$ (Gaussian Antenna Shape)	$DF_{ACC} \approx \frac{\lambda}{2\,\pi\,d\,\cos\theta}\Delta\theta$
DF Accuracy Improvement	Decrease Antenna BW; Decrease Amplitude Mistrack; Increase Squint Angle	Increase Spacing of Outer Antennas; Decrease Phase Mistrack
Typical DF Accuracy	3° to 10° rms	0.1° to 3° rms
Sensitivity to Multipath/Reflections	High Sensitivity; Mistrack of Several dB Can Cause Large DF Errors	Relatively Insensitive; Interferometer Can be Made to Tolerate Large Phase Errors
Platform Constraints	Locate in Reflection Free Area	Reflection Free Area; Real Estate for Array; Prefers Flat Radome
Applicable Receivers	Crystal Video; Channelizer; Acousto-Optic; Compressive; Superheterodyne	Superheterodyne
ΔC_{dB}= Amplitude Monopulse Ratio in dB S= Squint Angle in degrees θ_{BW}= Antenna Beamwidth in degrees		

Table 4. Emitter Location Techniques.

Measurement Technique	Advantages	Disadvantages
Triangulation	Single Aircraft	Non-instantaneous location Inadequate accuracy for remote targeting Not forward looking
Azimuth/elevation	Single Aircraft Instantaneous location possible	Accuracy degrades rapidly at low altitude Function of range
Time Difference of Arrival (Pulsed signals)	Very high precision Can support weapon delivery position requirements Very rapid, can handle short on-time threat	Very complex, diverse systems required, at least 3 aircraft High quality receivers, DME (3 sites) very wideband data link Very high performance control processor; requires very high reliability subsystems

Table 5. Qualitative Comparison of Receivers. (From NRL Report 8737)

Feature	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
Instantaneous Analysis Bandwidth	Very wide	Narrow	Very wide	Narrow	Moderate	Wide	Wide	Moderate
Frequency Resolution	Very poor	Fair	Good	Very good	Poor	Fair	Good	Good
Sensitivity	Poor (No preamp) Fair (preamp)	Fair/ good	Poor (No preamp) Fair (preamp)	Very good	Fair	Fair/ good	Very good	Good
Dynamic Range	Fair	Fair/ good	Good	Very good	Fair	Good	Fair	Poor
Speed of Acquisition	Very Fast	Slow	Very Fast	Slow	Fast	Very Fast	Very Fast	Fast
Short pulse Width Capability	Good	Good	Good	Good	Very good	Good	Fair	Fair
Retention of Signal Character-istics	Fair	Fair	Poor	Good	Fair/ good	Good	Poor	Fair/ good
Applicability to Exotic Signals	Poor/ fair	Poor	Good	Poor	Fair/ good	Good	Fair/ good	Fair/ good
High signal Density Performance	Poor (high false alarm rate from background)	Fair/ good	Good	Poor	Fair (depending on BW)	Fair/good, depending on architecture & processing	Good	Poor
Simultaneous Signal Capability	Poor	Fair/ good	Poor	Good	Fair (depending on BW)	Good	Good	Good
Processing Complexity	Moderate depending on application	Moderate depending on application	Moderate	Moderate	Moderate	Low-high depending on architecture	Complex	Simple signal processing complex data processing
Immunity to Jamming	Poor	Fair	Poor/ Fair	Good	Poor/ Fair	Good	Good	Good
Power Requirements	Low	Low/ Moderate	Moderate	Moderate	Moderate	High	Moderate	Moderate/ High
RF Range (GHz)	Multi-octave (0.5-40)	0.15-18 separate	>0.5 to 40	<0.01 to 40	0.5 to 18	0.5 to 60	<0.5 to 8	0.5-4 (0.5-18 channelized and down conversion)
Max Instantane-ous Analysis Bandwidth	Multi-octave (to 17.5 GHz)	As high as desired with equivalent reduction in resolution	Multi-octave (1 octave per unit)	50 MHz	500 MHz	~2 GHz without degradation, 17.5 GHz with degradation	0.5 to 2 depending on PW limitation	1 GHz
Frequency Accuracy	Measurement accuracy no better than analysis BW	Measurement accuracy no better than analysis BW	5-10 MHz	0.5% to 1%	0.5 to 3 MHz	±1 MHz	10 KHz	±1 MHz

Feature	Receiver Type							
	Wide-Band Crystal Video	TRF Crystal Video	IFM	Narrow-Band Superhet	Wide-Band Superhet	Channelized	Microscan	Acousto-optic
Pulse Width Range	CW to 50 ns	CW to 50 ns	CW to ~20 ns (depending on resolution)	CW to 100 ns with 20 MHz resolution	CW to 4 ns with 500 MHz resolution	CW to 30 ns (depending on resolution)	CW to 250 ns	CW to 0.5 µs
Frequency Resolution	~400 MHz (no better than BW)	25 MHz	1 MHz	<0.1 MHz	100-500 MHz	10-125 MHz (less with freq vernier)	1 MHz	0.5 to 1 MHz
Sensitivity (dBm)	-40 to -50 (no preamp) -80 (with preamp)	Better than -80 with preamp	-40 (no preamp) -75 (preamp) 4 GHz BW	-90, 1 MHz BW	-80, 500 MHz BW	-70, 10-50 MHz BW	-90, 5-10 MHz BW	-70 to -80
Maximum Dynamic Range (dB)	70	70-80	80 (w/preamp) 100+ (saturated)	90	60	50-80	40-60	25-35
Tuning Time	-	50 ms	-	1.0 s (1 octave)	.12 s (200 MHz band)	-	0.3 µs LO scan time	0.5 ms (integration time)
Signal ID Time	100 ns	50 ms	2-10 ms	~0.1 s	-	2.10 ms	~1 µs	-
Minimum Weight (lb)	20 (with processor)	30	<20 (octave unit) 65-75 (full coverage)	60-75	35 (tuner only)	1309-200 for 0.5 to 18 GHz coverage	25	29-55
Size / Minimum Volume (in³)	Small 300 (w/processor)	Small 375	Sm/Moderate 600-1000 ~100 miniaturized	Moderate 1500-3000	Moderate Several thousand	Large 4000-8000 (0.5-18 GHz coverage	Moderate 1200-2000	Small 800-1900
Minimum Power (W)	100 (with processor) <10 without processor	60 (without processor)	~50 (octave unit)	150	150 (tuner only)	350 to 1200 for 0.5 to 18 GHz coverage	70-80	200
Cost	Low	Low/ Moderate	Moderate	Moderate/ High	Moderate/ High	High	Moderate/ High	Low/ Moderate

This page intentionally left blank.

RADAR MODES

Typical Radar modes are listed below in the general functional category for which they were designed. Not all of these modes are applicable to all radars and certain radars have additional modes.

• NAVIGATION

Terrain avoidance - A mode in which the radar is set at a fixed depression angle and short range to continuously sweep the ground area directly in front of the aircraft in order to avoid mountains. This is particularly useful during flight into unfamiliar territory when clouds, haze, or darkness obscure visibility.

Ground mapping - A mode in which the radar uses a variety of techniques to enhance ground features, such as rivers, mountains, and roads. The mode is unlike air-to-air modes where ground return is rejected from the display.

Precision velocity update / Doppler navigation - A mode in which the radar again tracks ground features, using Doppler techniques, in order to precisely predict aircraft ground speed and direction of motion. Wind influences are taken into account, such that the radar can also be used to update the aircraft inertial navigation system.

• FIGHTER MISSIONS

Pulse search - Traditional pulse techniques are used to accurately determine range, angle, and speed of the target. Limitations are easy deception by enemy jamming, and less range when compared to other modes.

Velocity search - A high PRF Pulse Doppler waveform is used for long range detection primarily against nose aspect targets, giving velocity and azimuth information. Although velocity search can work against tail-on targets, the Doppler return is weaker, consequently the maximum detection range is also much less. When the target is in the beam (flying perpendicular to the fighter), the closure (Doppler) is the same as ground return and target return is almost zero.

Track While Scan (TWS) - A system that maintains an actual track on several aircraft while still searching for others. Since the radar is sharing its computing time between targets, the accuracy is less precise than for a single target track (STT) mode of operation.

Raid assessment - A mode in which the radar has an STT on a single target, but is routinely driven off by a small amount in order to determine if multiple aircraft exists in the immediate vicinity of the target aircraft.

Single-Target-Track (STT) (including air combat maneuvering modes) - Highly precise STT modes are used to provide the most accurate information to the fire control computer so that accurate missile or gun firing can be accomplished. The fire control radar continuously directs energy at the target so that the fired missile locates and tracks on the reflected energy from the target. Air combat maneuvering modes are automatic modes in which the radar has several sweep patterns fixed about the aircraft axis, such that little or no work is required of the pilot in order to lock up a target.

• AIR-TO-GROUND MISSIONS

Weapons delivery - A mode in which ground features are tracked, and particular emphasis is placed on determining range to the ground target, angle of dive, weapons ballistic tables, and aircraft speed.

Surveillance/tracking of ground forces/targets - Similar to the above with emphasis on multiple ground features and less on weapons delivery data.

Reconnaissance - A specific navigational mode to aid in identifying specific targets.

• AIR-TO-SURFACE MISSIONS

ASW - Navigational techniques specializing in specific search patterns to aid in detection of enemy submarines.

• TECHNIQUES USED FOR MULTIPLE APPLICATIONS

Synthetic Aperture Radar (SAR) - A form of radar that uses the relative motion between an antenna and its target region, to provide coherent-signal variations, in order to obtain finer spatial resolution than is possible with conventional beam-scanning means. SAR is usually implemented by mounting a single beam-forming antenna on a moving platform such as an aircraft from which a target scene is repeatedly illuminated with pulses of radio waves at wavelengths anywhere from a meter down to millimeters. The many echo waveforms received successively at the different antenna positions are coherently detected and stored and then post-processed together to resolve elements in an image of the target region.

Over-The-Horizon Radar (OTHR) - uses the refraction of high frequency radiation through the ionosphere in order to detect targets beyond the line-of-sight. The complexities of the ionosphere can produce multipath propagation, which may result in multiple resolved detections for a single target. When there are multipath detections, an OTHR tracker will produce several spatially separated tracks for each target. Information conveying the state of the ionosphere is required in order to determine the true location of the target and is available in the form of a set of possible propagation paths, and a transformation from measured coordinates into ground coordinates for each path. Since may be no other information as to how many targets are in the surveillance region, or which propagation path gave rise to which track, there is a joint target and propagation path association ambiguity which must be resolved using the available track and ionospheric information.

GENERAL RADAR DISPLAY TYPES

There are two types of radar displays in common use today.

RAW VIDEO

Raw video displays are simply oscilloscopes that display the detected and amplified target return signal (and the receiver noise). Raw video displays require a human operator to interpret the various target noise and clutter signals.

On the left hand display of Figure 1, an operator could readily identify three targets and a ghost (a ghost is a phony target that usually fades in and out and could be caused by birds, weather, or odd temporary reflections - also referred to as an angel). Target 3 is a weak return and hidden in the noise - an operator can identify it as a target by the “mouse under the rug” effect of raising the noise base line.

SYNTHETIC VIDEO

Synthetic video displays use a computer to clean up the display by eliminating noise and clutter and creating its own precise symbol for each target.

On the right hand display target 1 comes and goes because it is barely above the receiver noise level - notice that it is quite clear on the raw video. Target 3 wasn’t recognized by the computer because it’s too far down in the noise. The computer validated the ghost as a target. The ghost might be a real target with glint or ECM characteristics that were recognized by the computer but not the operator.

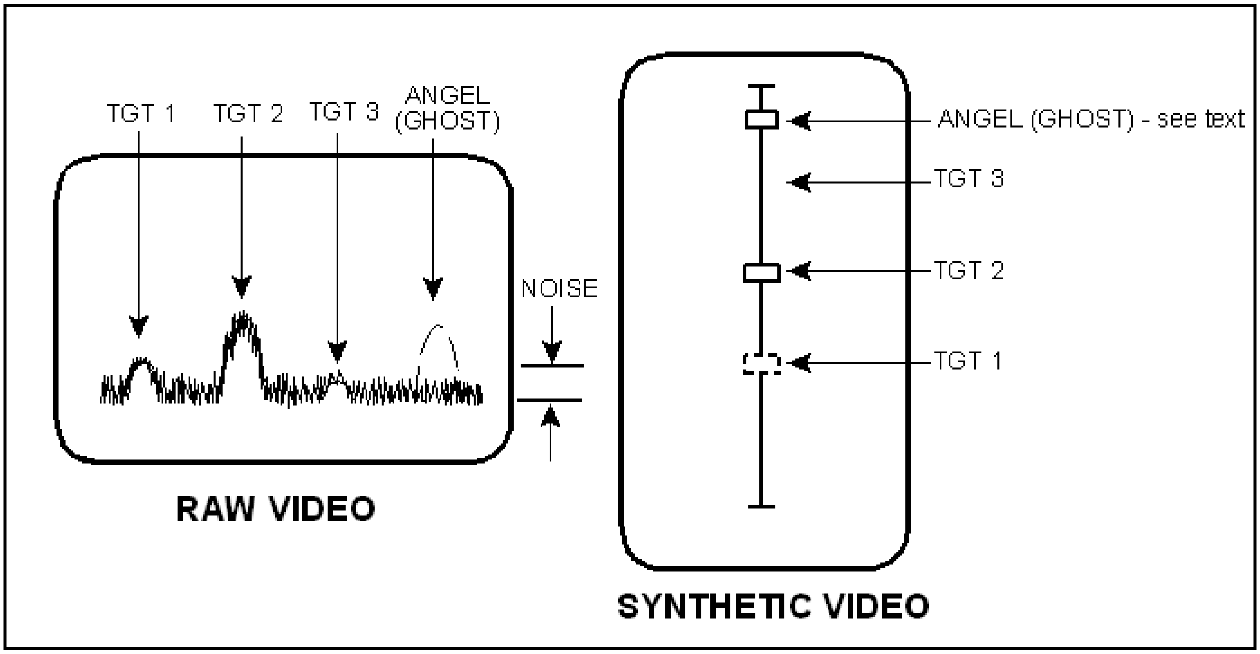

Figure 1. Radar Display Types.

SEARCH AND ACQUISITION RADARS

They generally use either a PPI or a sector PPI display as shown in Figure 2. PPI displays can be either raw video or synthetic video.

PPI scope (plan position indicator).
    Polar plot of direction and distance.
    Displays all targets for 360 degrees.

Sector PPI scope.
    Polar plot of direction and distance.
    Displays all targets within a specific sector.
    Origin may be offset so that “your” radar position may be off the scope.

TRACKING RADARS

Usually use some combination of A, B, C, or E scope displays. There are many other types of displays that have been used at one time or another - including meters - but those listed here are the most common in use today.

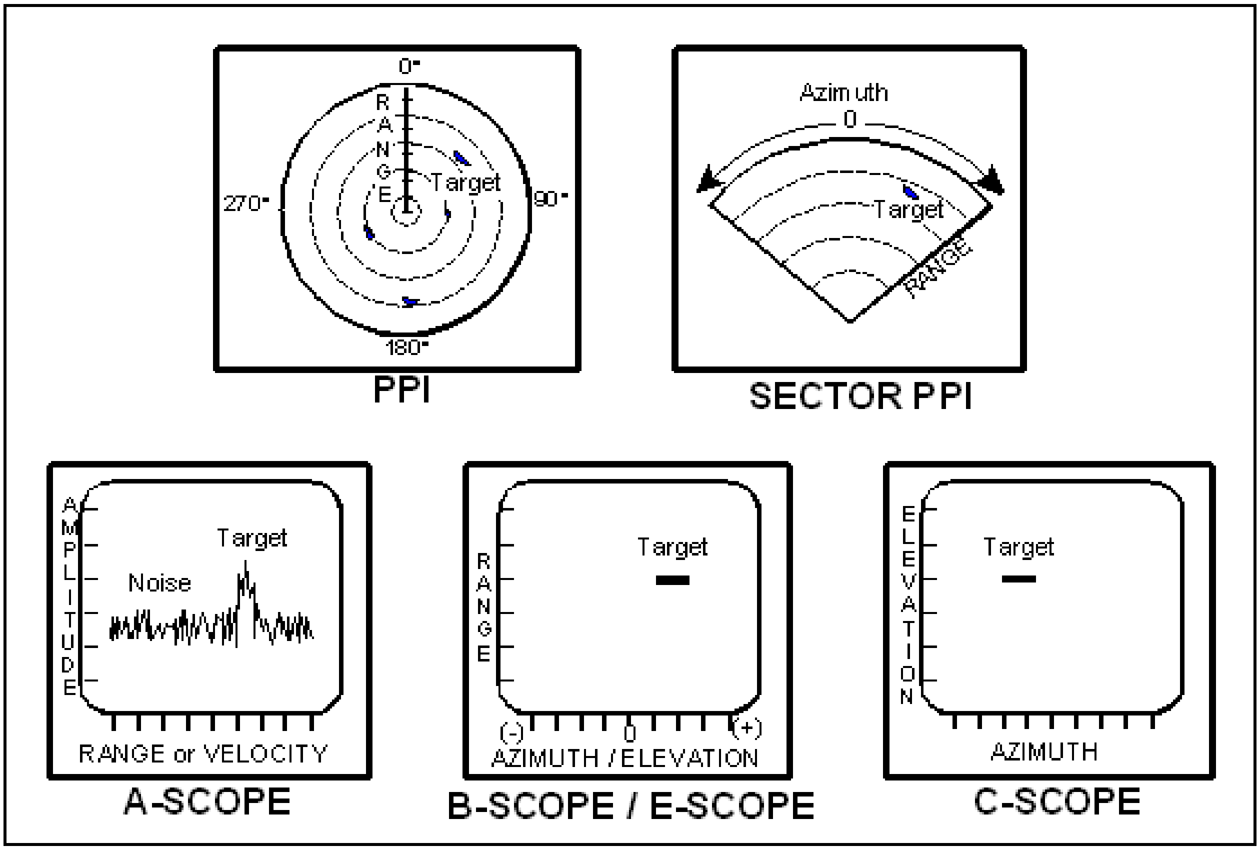
0°
RANGE
Target
270°
90°
180°
PPI
Azimuth
0
Target
RANGE
SECTOR PPI
AMPLITUDE
Target
Noise
RANGE or VELOCITY
A-SCOPE
RANGE
Target
(-)
0
(+)
AZIMUTH / ELEVATION
B-SCOPE / E-SCOPE
ELEVATION
Target
AZIMUTH
C-SCOPE

Figure 2. Common Radar Displays.

A-SCOPE

Target signal amplitude vs. range or velocity.

Displays all targets along pencil beam for selected range limits.

Displays tracking gate. Usually raw video. Some modern radars have raw video a-scopes as an adjunct to synthetic video displays.

Must be used with a separate azimuth and elevation display of some sort.

Also called a range scope (R-Scope).

B-SCOPE

Range vs. azimuth or elevation. Displays targets within selected limits.

Displays tracking gate. May be raw or synthetic video.

Surface radars usually have two. One azimuth/one elevation which can result in confusion with multiple targets.

C-SCOPE

Azimuth vs. elevation. Displays targets within selected limits of az and el.

Displays tracking gate. May display bull's-eye or aim dot.

May have range indicator inserted typically as a marker along one side. Usually synthetic video.

Pilots eye view and very common in modern fighter aircraft heads up displays for target being tracked.

Could be used in any application where radar operator needs an "aiming" or "cross hair" view like a rifle scope.

E-SCOPE

Elevation vs. Range similar to a B-scope, with elevation replacing azimuth.

This page intentionally left blank.

IFF - IDENTIFICATION - FRIEND OR FOE

Originated in WWII for just that purpose - a way for our secondary radars to identify U.S. aircraft from enemy aircraft by assigning a unique identifier code to U.S. aircraft transponders.

The system is considered a secondary radar system since it operates completely differently and independently of the primary radar system that tracks aircraft skin returns only, although the same CRT display is frequently used for both.

The system was initially intended to distinguish between enemy and friend but has evolved such that the term "IFF" commonly refers to all modes of operation, including civil and foreign aircraft use.

There are five major modes of operation currently in use by military aircraft plus two sub-modes.

- Mode 1 is a non-secure low cost method used by ships to track aircraft and other ships.
- Mode 2 is used by aircraft to make carrier controlled approaches to ships during inclement weather.
- Mode 3 is the standard system also used by commercial aircraft to relay their position to ground controllers throughout the world for air traffic control (ATC).
- Mode 4 is secure encrypted IFF (the only true method of determining friend or foe) Military only
- Mode 5 – provides a cryptographically secured version of Mode S and ADS-B GPS position. (military only). Mode 5 is divided into two levels. Both are crypto-secure with Enhanced encryption, Spread Spectrum Modulation, and Time of Day Authentication. Level 1 is similar to Mode 4 information but enhanced with an Aircraft Unique PIN. Level 2 is the same as Mode 5 level one but includes additional information such as aircraft position and other attributes
- Mode "C" is the altitude encoder (military and civilian).
- Mode S – provides multiple information formats to a selective interrogation. Typically aircraft are assigned a unique 24-bit Mode S address. The Mode S address is partitioned and a group of address ranges are allocated to each country. Some countries change the assigned address for security reasons, and thus it might not be a unique address. (military and civilian)

The non-secure codes are manually set by the pilot but assigned by the air traffic controller.

A cross-band beacon is used, which simply means that the interrogation pulses are at one frequency and the reply pulses are at a different frequency. 1030 MHz and 1090 MHz is a popular frequency pair used in the U.S.

The secondary radar transmits a series of selectable coded pulses. The aircraft transponder receives and decodes the interrogation pulses. If the interrogation code is correct, the aircraft transponder transmits a different series of coded pulses as a reply.

The advantage of the transponder is that the coded pulses "squawked" by the aircraft transponders after being interrogated might typically be transmitted at a 10 watt ERP, which is much stronger than the microwatt skin return to the primary radar. Input power levels may be on the order of several hundred watts.

The transponder antenna is low gain so that it can receive and reply to a radar from any direction.

An adjunct to the IFF beacon is the altitude encoding transponder known as mode C - all commercial and military aircraft have them, but a fair percentage of general aviation light aircraft do not because of cost. The number of transponder installations rises around many large metropolitan areas where they are required for safety (easier identification of aircraft radar tracks).

Air traffic control primary radars are similar to the two-dimensional search radar (working in azimuth and range only) and cannot measure altitude.

The expanded display in figure 1 is typical of an air traffic control IFF response. The aircraft was told to squawk a four digit number such as "4732." The altitude encoded transponder provides the aircraft altitude readout to the ground controllers display along with the coded response identifying that particular aircraft.

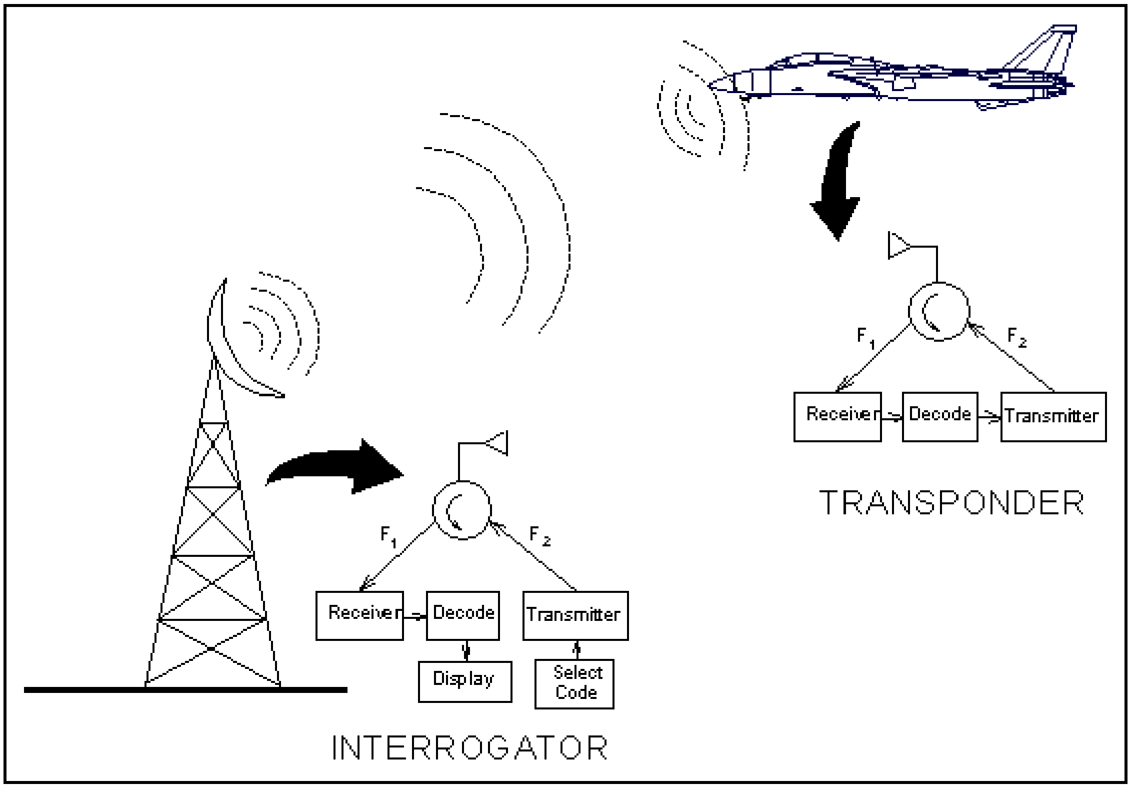

Figure 1. IFF Transponder.

In addition to systems with active electronic data interchange between airborne and ground equipment, some military surveillance systems can provide targeting in tactical applications. The development of automated techniques for use against ground targets is typically referred to as Automatic Target Recognition (ATR). When used against air targets, it is typically referred to as Non-Cooperative Target Recognition (NCTR). The requirements for radar target recognition are complex since typical targets have background clutter and often multiple targets types exist.

RECEIVER TESTS

Two tone and spurious response (single signal) receiver tests should be performed on EW and radar receivers to evaluate their spurious free dynamic range. A receiver should have three ranges of performance: (1) protection from damage, (2) degraded performance permitted in the presence of a strong interfering signal(s) and no degradation when only a strong desired signal is present, and (3) full system performance.

The original MIL-STD-461A design requirement and its companion MIL-STD-462 test requirement specified four receiver tests. These standards allowed the interfering signal(s) to be both inband and out of band, which is meaningful for design and test of EW receivers, however inband testing generally is not meaningful for narrowband communications receivers. These standards were difficult to follow and had to be tailored to properly evaluate the EW and radar system. MIL-STD-461B/C still allowed the interfering signal(s) to be both inband and out of band but deleted the single signal interference test (CS08 Conducted Susceptibility test). MIL-STD-461D/-462D leave the pass/fail criteria entirely up to what is listed in the individual procurement specification. It also places all interfering signals out of band, redesignates each test number with a number "100" higher than previously used, and combines "CS08" as part of CS104. Therefore, to provide meaningful tests for EW and radar systems, the procurement specification must specify the three ranges of performance mentioned in the beginning of this section and that the tests are to be performed with the interfering signal(s) both inband and out of band. The four tests are as follows (listed in order of likelihood to cause problems):

Test Name	MIL-STD-461A	MIL-STD-461D
Undesired, Single signal interference test	CS08	Part of CS104
Desired with undesired, two signal interference tests	CS04	CS104
Two signal intermodulation test	CS03	CS103
Two signal cross modulation test	CS05	CS105

The rest of this section explains the application of these tests and uses the names of the original MIL-STD-461A tests to separate the tests by function.

TEST SETUP

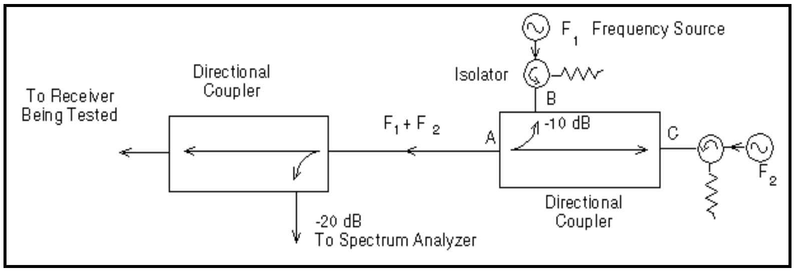

Figure 1. Receiver Test Setup When Antenna Can Be Removed.

A directional coupler used backwards (as shown here in Figure 1) is an easy way to perform two signal tests. The CW signal should be applied to the coupling arm (port B) since the maximum CW signal level is -10 dBm. The pulse signal should be applied to the straight-through path (port C) since the maximum pulse level is +10 dBm peak. These power levels are achievable with standard laboratory signal generators, therefore one doesn't have to resort to using amplifiers which may distort the signals. Always monitor the output signal to verify spectrally pure signals are being applied to the test unit. This can be accomplished by another directional coupler used in the standard configuration. Dissimilar joints or damaged or corroded microwave components can cause mixing. This can also result if the two signal generators are not isolated from one another. Therefore, even if a directional coupler is used to monitor the signal line, it is still advisable to directly measure the input to the receiver whenever there is a suspected receiver failure. This test does not need to be performed in an EMI shielded room and is more suitable for a radar or EW lab where the desired signals are readily available.

If the receiver's antenna is active or cannot be removed, a modified test as shown in Figure 2 should be performed. The monitoring antenna which is connected to the spectrum analyzer should be the same polarization as the antenna for the receiver being tested. Amplifiers may be required for the F_1 and F_2 signals. It is desirable to perform this test in an anechoic chamber or in free space.

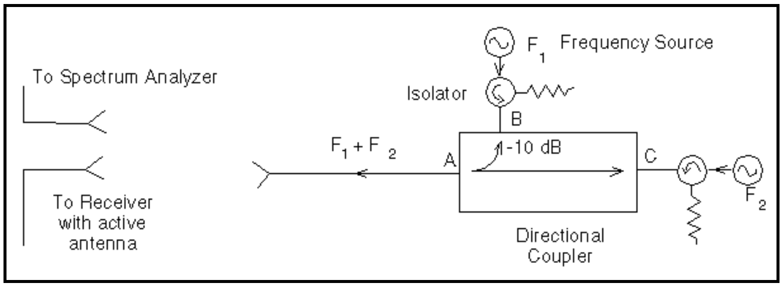

Figure 2. Receiver Test Setup When Antenna Is Active.

In the following discussion of CS08, CS04, CS03, and CS05 tests, it is assumed that when the receive light illuminates, the receiver identifies a signal that matches parameters in the User Data File (UDF) or pre-programmed list of emitter identification parameters. If a receiver is different, the following procedures will have to be appropriately tailored. If the UDF does not have entries for very low level signals in the 10% and 90% regions of each band, complete testing is not possible. Most problems due to higher order mixing products and adjacent band leakage are only evident in these regions. In the following tests, the lowest level where the receive light is constantly on is used to identify the minimum receive level. If a receiver has a receive level hysteresis or other idiosyncrasy, then using a 50% receive light blinking indicator may be more appropriate. Whatever technique is appropriate, it should be consistently used during the remainder of the test. The maximum frequency for testing is normally 20 GHz. If a millimeter wave receiver is being tested, the maximum frequency should be 110 GHz.

CS08 - UNDESIRED, SINGLE SIGNAL INTERFERENCE TEST

MIL-STD-461B/C (EMI design requirements) deleted this test. MIL-STD-461D allows a single signal test as part of CS104 (CS04) but specifies it as an out of band test. The original CS08 inband and out of band test is still needed and is the most meaningful test for wide band EW receivers which have a bandwidth close to an octave. This test will find false identification problems due to 1) lack of RF discrimination, 2) higher order mixing problems, 3) switch or adjacent channel/band leakage, and 4) cases where the absence of a desired signal causes the receiver to search and be more susceptible. In this latter case, a CS04 two signal test could pass because the receiver is captured by the desired signal, whereas a CS08 test could fail. Examples of the first three failures are as follows.

EXAMPLE 1

A 2 to 4 GHz receiver which uses video detection (e.g., crystal video) and doesn't measure RF is used for this example. This receiver assumes that if the correct Pulse Repetition Interval (PRI) is measured, it is from a signal in the frequency band of interest. Three cases can cause false identification. Refer to Figure 3.

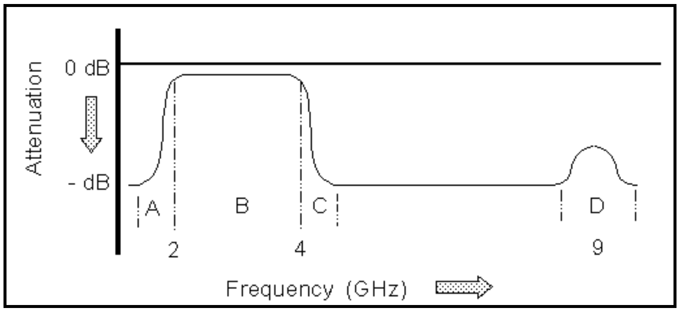

Figure 3. Frequency Areas in a Sample 2-4 GHz Receiver.

(1) Region A&C. The 2 to 4 GHz band pass filter will pass strong signals in regions A&C. If they have the correct PRI, they will also be identified.

(2) Region B. Any other signal besides the desired signal in the 2 to 4 GHz region that has the correct PRI will also be identified as the signal of interest.

(3) Region D. Band pass filters with poor characteristics tend to pass signals with only limited attenuation at frequencies that are three times the center frequency of the band pass filter. If these signals have the correct PRI, they will be incorrectly identified.

High duty cycle signals (CW or pulse doppler) in regions A, B, C, and D may overload the processing of signals, saturate the receiver, or desensitize the receiver. This case is really a two signal CS04 test failure and will be addressed in the CS04 section.

EXAMPLE 2

A receiver measuring the carrier frequency of each pulse (i.e., instantaneous frequency measurement (IFM)) and the PRI is used for this example. False signal identification can occur due to higher order mixing products showing up in the receiver pass bands. These unwanted signals result from harmonics of the input RF mixing with harmonics of the Local Oscillator (LO). Refer to Figures 4 and 5.

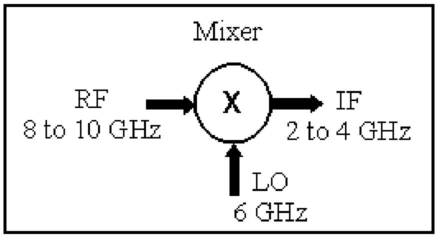

Figure 4. Low Side Mixing.

Mixers are nonlinear devices and yield the sum, difference, and the original signals. Any subsequent amplifier that is saturated will provide additional mixing products.

If a 8.5 GHz signal with a 1 kHz PRI is programmed to be identified in the UDF, measurements are made at the 2.5 GHz Intermediate Frequency (IF), i.e., RF-LO = IF = 8.5-6 = 2.5 GHz.

The same 2.5 GHz signal can result from an RF signal of 9.5 GHz due to mixing with the second harmonic of the LO i.e., 2 X 6 - 9.5 = 2.5 GHz. This signal will be substantially attenuated (approximately 35 dB) when compared to the normal IF of 9.5 - 6 = 3.5 GHz. If the receiver has filters at the IF to reduce the signal density and a filter has minimum insertion loss at 2.5 GHz and maximum insertion loss at 3.5 GHz, then only the low level 2.5 GHz signal will be measured and assumed to be due to a 8.5 GHz input signal whereas the input is really at 9.5 GHz.

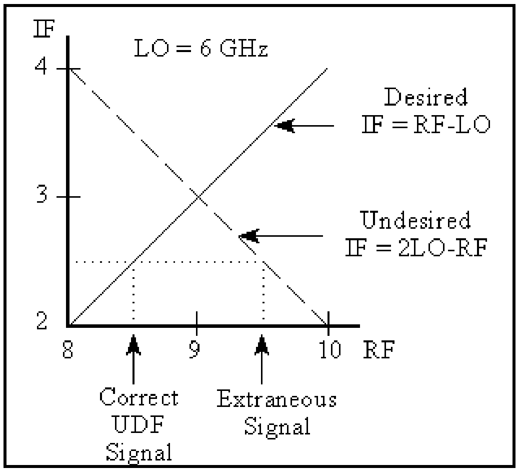

Figure 5. Low Side Mixing Results.

Spurious intermodulation products can also result from high side mixing, but generally the suppression of undesired signals is greater. In this case, the LO is at a frequency higher than the RF input. This is shown in Figures 6 and 7.

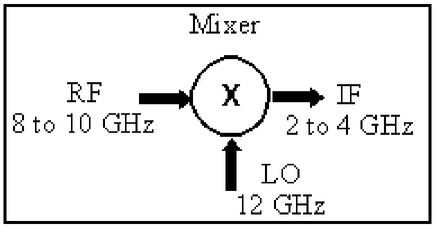

Figure 6. High Side Mixing.

Table 1. Intermodulation Product Suppression.

Harmonic of		
LO	RF	Suppression
1	1	0
1	2	ΔP-41
1	3	$2\Delta P$-28
2	1	-35
2	2	ΔP-39
2	3	$2\Delta P$-44
3	1	-10
3	2	ΔP-32
3	3	$2\Delta P$-18
4	1	-35
4	2	Δ-39
5	1	-14
5	3	$2\Delta P$-14
6	1	-35
6	2	ΔP-39
7	1	-17
7	3	$2\Delta P$-11

Courtesy Watkins-Johnson

As previously mentioned, the amplitude of intermodulation products is greatly reduced from that of the original signals. Table 1 shows rule of thumb approximate suppression (reduction), where $\Delta P = P_{RF}(\mathrm{dBm}) - P_{LO}(\mathrm{dBm})$. As can be seen, the strength of the LO is a factor. The higher the LO power, the more negative the suppression becomes.

If one assumes the maximum RF power for full system performance is +10 dBm and the LO power level is +20 dBm, then ΔP = -10 dB minimum. Therefore in this example, the 3RF-2LO mixing product would be $2\Delta P - 44 = -20 - 44 = -64$ dB when compared to the desired mixing product.

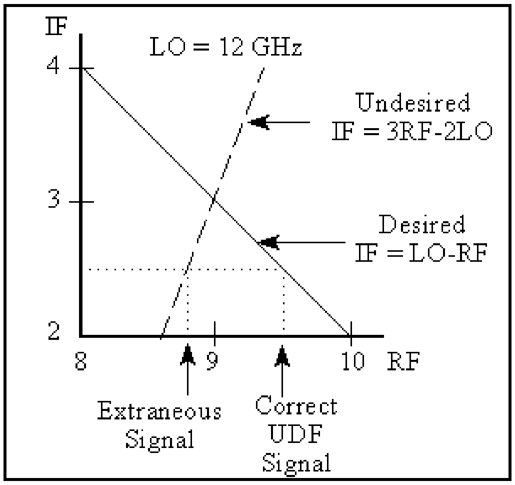

Figure 7. High Side Mixing Byproducts.

The use of double mixing, as shown in Figure 8, can significantly reduce unwanted signals but it is more expensive. For a 8 GHz signal in, one still generates a 2 GHz IF but by mixing up, then down, unwanted signals are not generated or significantly suppressed.

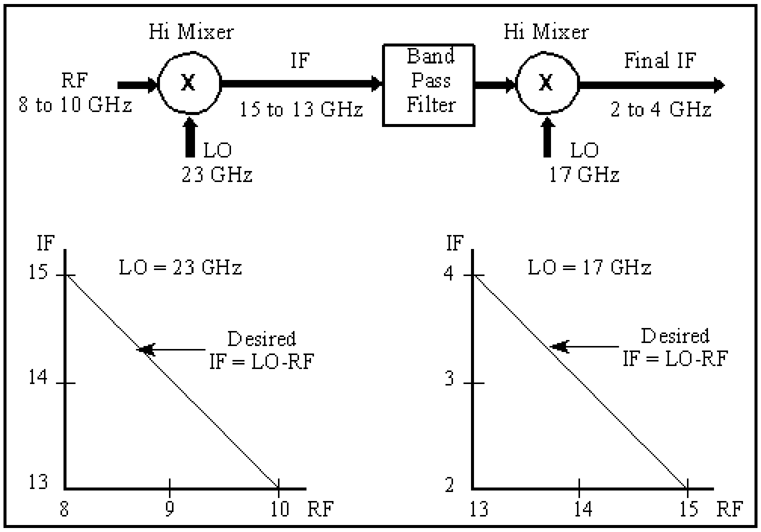
Hi Mixer
RF
8 to 10 GHz
X
IF
15 to 13 GHz
LO
23 GHz
Band
Pass
Filter
Hi Mixer
X
Final IF
2 to 4 GHz
LO
17 GHz
IF
15
14
13
8
9
10 RF
LO = 23 GHz
Desired
IF = LO-RF
IF
4
3
2
13
14
15 RF
LO = 17 GHz
Desired
IF = LO-RF

Figure 8. Double Mixing.

Some of these problems can be corrected by:

(1) always having LOs on the high side versus low side of the input RF (but this is more expensive),

(2) using double mixing

(3) software programming the receiver to measure for the potential stronger signal when a weak signal is measured in a certain IF region, and

(4) improved filtering of the LO input to the mixer and the output from the mixer.

EXAMPLE 3

If the same receiver discussed in example 2 had additional bands (Figure 9) and used a switch at the IF to select individual bands, a strong signal in an adjacent band could be inadvertently measured because:

(1) the switch, which may have 80 dB of isolation when measured outside the circuit, may only have 35 dB isolation when installed in a circuit because of the close proximity of input and output lines,

(2) the strong signal in one band may have the same IF value that is being sought in an adjacent band, and

(3) the additional parameters such as PRI may be the same.

As shown in Figure 9, assume that in band 2 we are looking for a 4.5 GHz signal that has a PRI of 1 kHz. Measurements are made at an IF of 3.5 GHz since LO-RF = IF = 8-4.5 = 3.5 GHz. If a 6.5 GHz signal is applied to band 3, its IF also equals 3.5 since LO-RF = 10-6.5 = 3.5 GHz. If this is a strong signal, has a PRI of 1 kHz, and there is switch leakage, a weak signal will be measured and processed when the switch is pointed to band 2. The receiver measures an IF of 3.5 GHz and since the switch is pointed to band 2, it scales the measured IF using the LO of band 2 i.e., LO-IF = RF = 8-3.5 = 4.5 GHz. Therefore, a 4.5 GHz signal is assumed to be measured when a 6.5 GHz signal is applied. Similarly this 6.5 GHz signal would appear as a weak 3.5 GHz signal from band 1 or a 9.5 GHz signal from band 4.

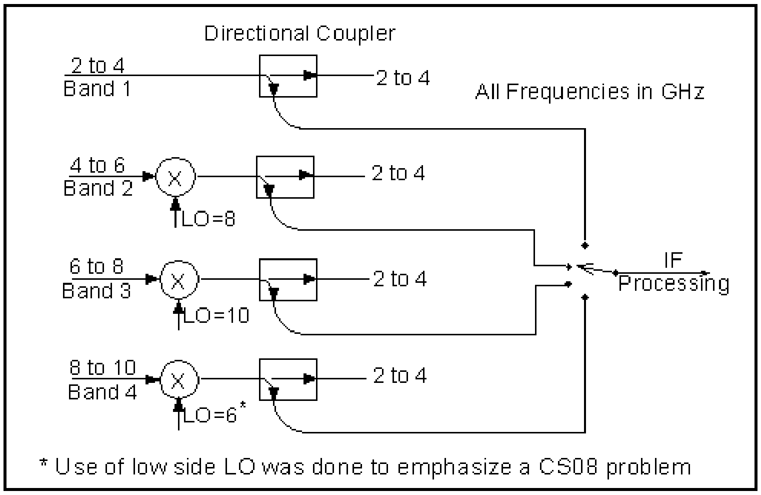

Figure 9. Multi Band Receiver With Common IF.

In performing this test it is important to map the entries of the UDF for each band i.e., show each resulting IF, its PRI, and the sensitivity level that the receive light is supposed to illuminate, i.e., if a test in one band used a PRI corresponding to a PRI in another band where the receive threshold is programmed to not be sensitive this will negate the effectiveness of a cross coupling test. Mapping the UDF will facilitate applying a strong signal to one band using the PRI of a desired signal in an adjacent band.

CS08 TEST PROCEDURE

Assume that the receiver band is 2 to 4 GHz as shown in Figure 10. Pick the UDF entry that has the greatest sensitivity. UDF #1 entry is for a 3±.05 GHz signal with a PRI of 1 kHz. If the test signal is set for the UDF #1 PRI, a receive light will also occur at the frequencies of UDF #2 if it also has the same PRI (this is not a test failure). If adjacent bands don't also have entries with the same PRI, then the test should be repeated for the band being tested with at least one of the adjacent band PRI values.

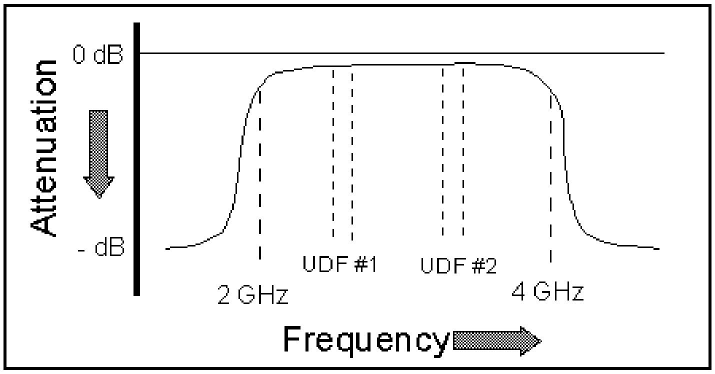

Figure 10. Receiver Band With Multiple UDF Entries.

(1) Set the receiver or jammer to the receive mode, verify it is working for UDF #1 and record P_o, the minimum signal level where the receive light is constantly on.

(2) Raise this signal to its maximum specified level for full system performance. If a maximum level is not specified, use +10 dBm peak for a pulse signal or -10 dBm for a CW signal.

(3) Tune this strong RF signal outside the UDF #1 range and record any RF frequency where the receive light comes on. If another inband UDF has the same PRI, this is not a failure.

(4) This test is performed both inband and out of band. Out of band tests should be performed on the high end to five times the maximum inband frequency or 20 GHz, whichever is less, and on the low end to IF/5 or 0.05 F0, whichever is less, unless otherwise specified. The out of band power level is +10 dBm peak for a pulse signal or -10 dBm for a CW signal, unless otherwise specified.

(5) If a receive light comes on when it is not supposed to, record the RF and reduce the power level to where the receive light just stays on constantly. Record this level P1. The interference rejection level is P1-P0= PIR

(6) Repeat this test for each type of signal the receiver is supposed to process, i.e., pulse, PD, CW, etc.

CS04 - DESIRED WITH UNDESIRED, TWO SIGNAL INTERFERENCE TEST

The intent is for a weak desired signal to be received in the presence of an adjacent CW signal. The desired signal is kept tuned at minimal power level and a strong unmodulated signal is tuned outside the UDF region. Radar and EW receivers without preselectors are likely to experience interference when this test is performed inband. Receivers with nonlinear devices before their passive band pass filter, or filters that degrade out of band, are likely to experience susceptibility problems when this test is performed out of band.

Tests performed inband - An unmodulated CW signal is used. If the receiver is supposed to handle both pulsed and CW signals, this test is performed inband. If the pulse receiver is supposed to desensitize in order to only process pulse signals above the CW level, then only this limited function is tested inband i.e., normally the levels correspond, if a CW signal of -20 dBm is present, then the receiver should process pulse signals greater than -20 dBm.

CS04 TEST PROCEDURE

(1) As shown in Figure 11, initially the pulse signal is tuned to F_0 and the minimum receive level P_0 is recorded, i.e., minimum level where the receive light is constantly on.

(2) The pulse signal is raised to the maximum specified level for full system performance and tuned on either side of F_0 to find the frequencies on both sides (F_{High} and F_{Low}) where the receive light goes out. If a maximum pulse power level is not specified, then +10 dBm peak is used. In some receivers F_L and F_H are the band skirts.

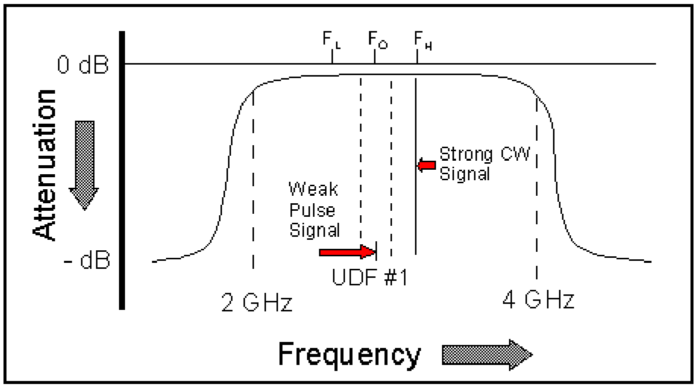

Figure 11. CS04 Test Signals.

(3) The pulse signal is returned to the level found in step 1. A CW signal at the maximum specified CW power level for full system performance is tuned above F_H and below F_L. If a maximum CW power level is not specified, then -10 dBm is used. Anytime the receive light is lost, the tuned CW RF value is recorded. The CW signal should be turned off to verify that the pulse signal can still be received in the absence of

interference. If the pulse signal is still being received, then the interfering CW signal should be reapplied and decreased to the lowest power level where the receive light stays on constantly. Record this level P_1. The interference rejection level is $P_1 - P_0 = P_{IR}$.

(4) Out of band tests should be performed to five times the maximum inband frequency or 20 GHz, whichever is less, and on the low end to IF/5 or 0.05 F_0, whichever is less, unless otherwise specified. The out of band CW power level is -10 dBm unless otherwise specified.

Failures - Out of band test

(1) If a non-linear device such as a limiter is placed before a band pass filter, a strong out of band signal can activate the limiter and cause interference with the inband signal. The solution is to place all non-linear or active devices after a passive band pass filter.

(2) Band pass filters with poor characteristics tend to pass signals with only limited attenuation at frequencies that are three times the center frequency of the band pass filter. Passage of a CW or high duty cycle signal that is out of band may desensitize or interfere with the processing of a weak inband signal.

CS03 INTERMODULATION TEST

This two signal interference test places a pulse signal far enough away (Δf) from the desired UDF frequency (F_0) that it won't be identified. A CW signal is initially placed 2Δf away. If an amplifier is operating in the saturated region, these two signals will mix and produce sum and difference signals. Subsequent mixing will result in a signal at the desired UDF frequency F_0 since $F_1 - (F_2 - F_1) = F_0$. These two signals are raised equally to strong power levels. If no problem occurs, the CW signal is tuned to the upper inband limit and then tuned out of band. A similar test is performed below F_0.

CS03 TEST PROCEDURES

(1) Set the receiver or jammer to the receive mode. Verify it is working at a desired signal frequency, (F_0), and record the minimum signal level i.e., lowest level where the receive light is constantly on (record this level P_0).

(2) The modulated signal is raised to the maximum specified level for full system performance and tuned on either side of F_0 to find the frequency F_1 on both sides where the receive light goes out. If a maximum power level is not specified, +10 dBm peak is used. The difference between F_1 and F_0 is Δf as shown in Figure 12.

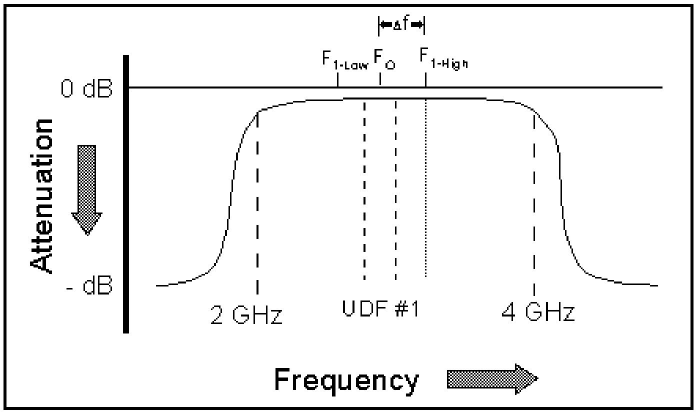

Figure 12. Initial CS03 Test Signal.

(3) As shown in Figure 13, a pulse signal is tuned to F_1 and a CW signal is tuned to F_2 where $F_2 = F_1 + \Delta f$ on the high side. The power level of the two signals is initially set to P_0 and raised together until the maximum specified levels for full system performance are reached. If maximum power levels are not specified, then +10 dBm peak is used for the pulse signal and -10 dBm is used for the CW signal. Whenever the receive light comes on, the two signals should be turned off individually to verify that the failure is due to a combination of the two signals versus (1) a single signal (CS08) type failure or (2) another inband UDF value has been matched. If the failure is due to the two signal operation, then the power level (P_1 and P_2) of F_1 and F_2 should be recorded. If P_1=P_2, the intermodulation rejection level is P_1-P_0=P_{IM}. If $P_1 \neq P_2$, it is desirable to readjust them to be equal when the receive light just comes on.

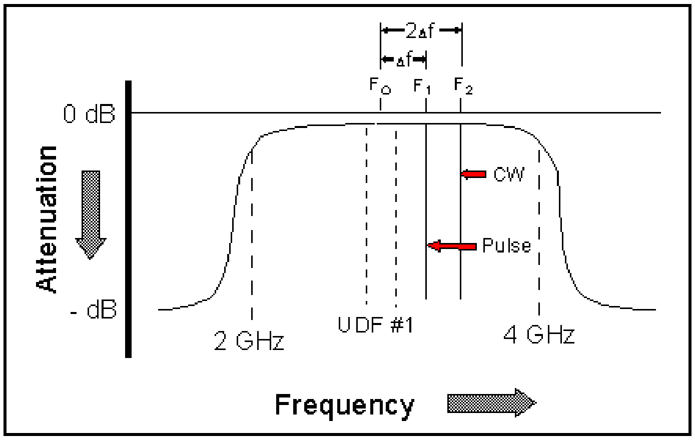

Figure 13. CS03 Testing Signal.

(4) Once the $F_1 + F_2$ signals are raised to the maximum power test levels described in step 3 without a failure, then F_2 is tuned to the upper limit of the band. F_2 should also be tuned out of band to five times the maximum inband frequency or 20 GHz whichever is less unless otherwise specified. The out of band power level is -10 dBm unless otherwise specified. Whenever the receive light comes on, F_2 should be turned off to verify that the failure is due to a two signal test. If it is, turn F_2 back on and equally drop the power levels of F_1 and F_2 to the lowest level where the receive light just comes on. Record the power levels (P_1 and P_2).

(5) Step 3 is repeated where F_1 is Δf below F_0 and F_2=F_1-Δf. Step 4 is repeated except F_2 is tuned to the lower limit of the band. F_2 should also be tuned out of band down to 0.1 F_0, unless otherwise specified.

(6) Normally if a failure is going to occur it will occur with the initial setting of F_1 and F_2. Care must be taken when performing this test to ensure that the initial placements of F_1 and F_2 do not result in either of the signals being identified directly.

As shown in Figure 14, if F_1 was placed at 3.2 GHz it would be identified directly and if F_2 was placed at 3.4 GHz it would be identified directly. Whereas, if F_1 was at 3.1 GHz and F_2 was at 3.2 GHz neither interfering signal would be identified directly but their intermodulation may result in an improper identification at F_0. Later when F_2 is tuned higher, the receive light will come on around 3.4 GHz and 3.6 GHz. This is not a test failure just a case of another inband UDF value being matched.

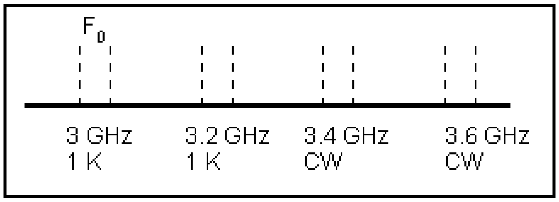

Figure 14. Sample UDF Entries.

CS05 - CROSS MODULATION

This two signal interference test places a weak CW signal where the receiver is programmed for a pulse signal and tunes a strong pulse signal elsewhere. As shown in Figure 15, when an amplifier is saturated, lower level signals are suppressed. When an amplifier is operated in the linear region all signals receive the rated linear gain. In this test the pulse signal will cause the amplifier to kick in and out of saturation and modulate the weak CW signal. The receiver may measure the modulation on the CW signal and incorrectly identify it as a pulse signal.

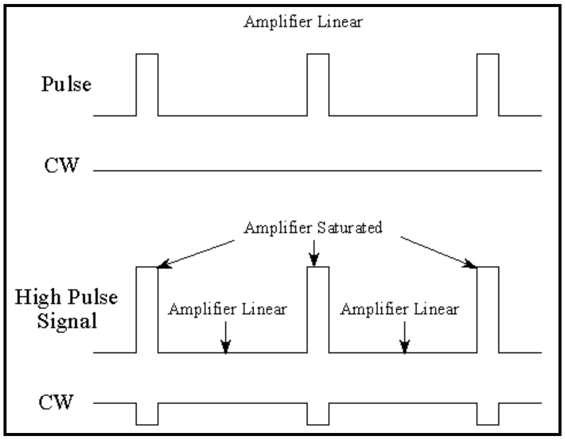
Amplifier Linear
Pulse
CW
Amplifier Saturated
High Pulse Signal
Amplifier Linear
Amplifier Linear
CW

Figure 15. Cross Modulation Example.

CS05 TEST PROCEDURE

(1) Initially the pulse signal is tuned to F_0 and the minimum power level P_0 where the receive light is constantly on is recorded.

(2) As shown in Figure 16, the signal is raised to the maximum specified level for full system performance for a pulse signal and tuned on either side of F_0 to find the frequencies on both sides, (F_{High} and F_{Low}) where the receive light goes out. If a maximum pulse power level is not specified, then +10 dBm peak is used.

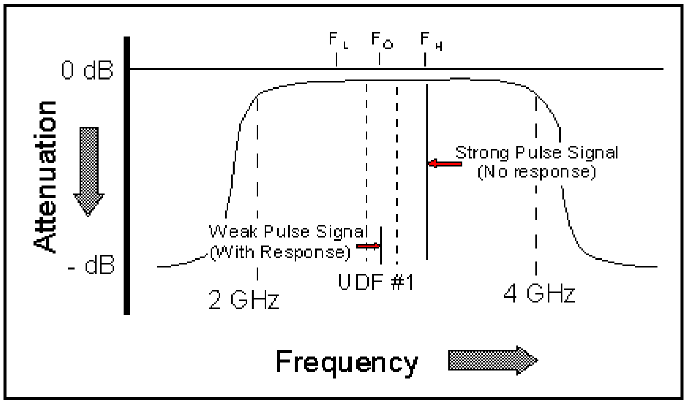
F_L F_0 F_H
0 dB
Attenuation
- dB
Strong Pulse Signal (No response)
Weak Pulse Signal (With Response)
2 GHz
UDF #1
4 GHz
Frequency

Figure 16. Initial CS05 Test Signals.

(3) The pulse signal from step 2 is turned off and a second signal is placed at F_0. It is a CW signal that is 10 dB stronger than the peak power level (P_0) measured is step 1. The receive light should not come on.

(4) As shown in Figure 17, the strong pulse signal of step 2 is turned back on and tuned above F_H and then tuned below F_L. Out of band tests should be performed to the maximum RF of the system + maximum IF or 20 GHz whichever is less and on the low end to the minimum RF of the system minus the maximum IF, unless otherwise specified.

(5) If a receive light occurs, turn off the weak CW signal since the “failure” may be due to the tuned pulsed signal, i.e., a CS08 failure or another inband UDF value has been matched.

If the light extinguishes when the weak CW signal is turned off, then turn the signal back on, reduce the value of the high level pulse signal until the minimum level is reached where the light stays on constantly. Record this level as P_1.

The cross modulation rejection level is P_1-P_0-10 dB = P_{CM}.

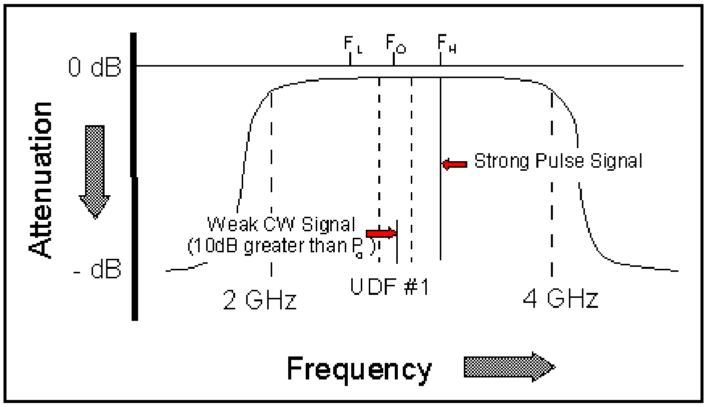
F_L
F_O
F_H
0 dB
Attenuation
Strong Pulse Signal
Weak CW Signal
(10dB greater than P_0)
- dB
2 GHz
UDF #1
4 GHz
Frequency

Figure 17. Final CS05 Test Signals.

This page intentionally left blank.

SIGNAL SORTING METHODS and DIRECTION FINDING

As shown in Figure 1, signal processing is basically a problem of signal detection, emitter parameter measurement and correlation, emitter sorting, identification, and operator notification.

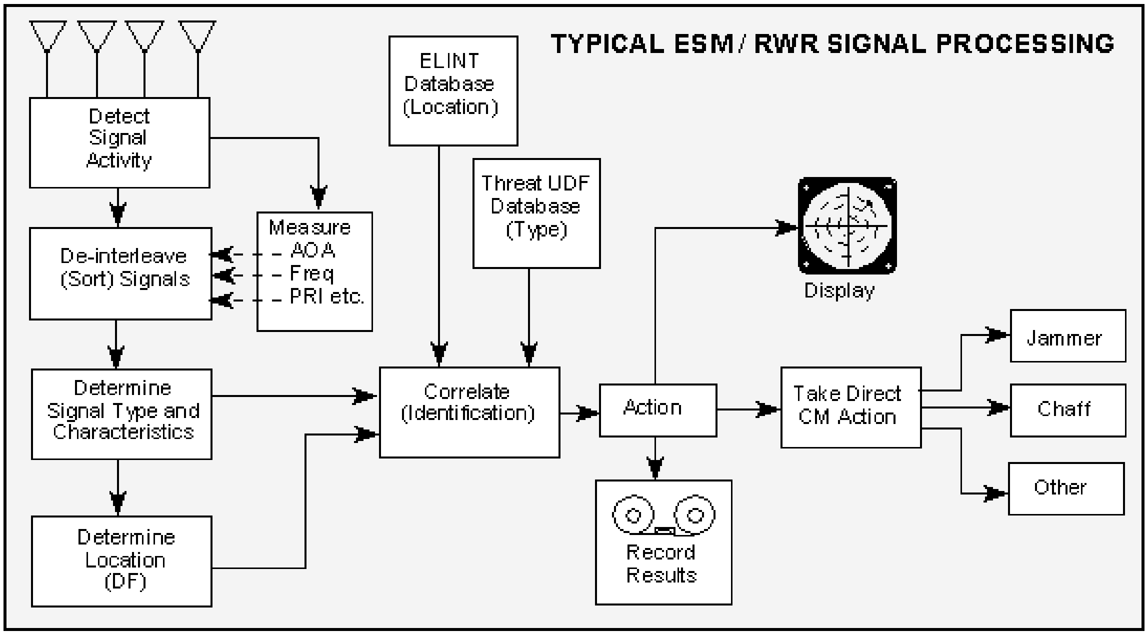
TYPICAL ESM / RWR SIGNAL PROCESSING
Detect Signal Activity
De-interleave (Sort) Signals
Measure
– AOA
– Freq
– PRI etc.
Determine Signal Type and Characteristics
Determine Location (DF)
ELINT Database (Location)
Threat UDF Database (Type)
Correlate (Identification)
Action
Display
Record Results
Take Direct CM Action
Jammer
Chaff
Other

Figure 1. Signal Processing Steps.

The ultimate goal of this processing is to classify radar signals by their unique characteristics and to use this data to identify enemy radars operating in the environment, determine their location or direction, assess their threat to friendly forces, and display this information to the operator.

While not all electronic support measures (ESM) or radar warning receiver (RWR) systems perform every step in this process, each completes some of them. For example, ESM systems seldom initiate direct CM action, while RWRs sometimes do. Also ESM systems frequently record electronic data for future use, but few RWRs do. ESM systems place more emphasis on accurate emitter location and hence direction finding capabilities, while RWRs usually give a rough estimate of position/distance.

The typical emitter characteristics that an ESM system can measure for a pulse radar include the following data:

1. Radio Frequency (RF)
2. Amplitude (power)
3. Direction of Arrival (DOA) - also called Angle of Arrival (AOA)
4. Time of Arrival (TOA)
5. Pulse Repetition Interval (PRI)
6. PRI type
7. Pulse Width (PW)
8. Scan type and rate
9. Lobe duration (beam width)

However, this list is not comprehensive. Other emitter parameters are available which may be necessary to characterize the threat system.

More sophisticated ESM systems can measure additional parameters, such as PRI modulation characteristics, inter- and intra-pulse Frequency Modulation (FM), missile guidance characteristics (e.g., pattern of pulse spacing within a pulse group), and Continuous Wave (CW) signals.

Still other parameters which can describe an electromagnetic wave but are currently not commonly used for identification include polarization and phase. However, as threat emitters begin to use this data more frequently to avoid jamming the more important they may become in identifying signals.

Some of the emitter characteristics which describe an electromagnetic wave are shown in Figure 2.

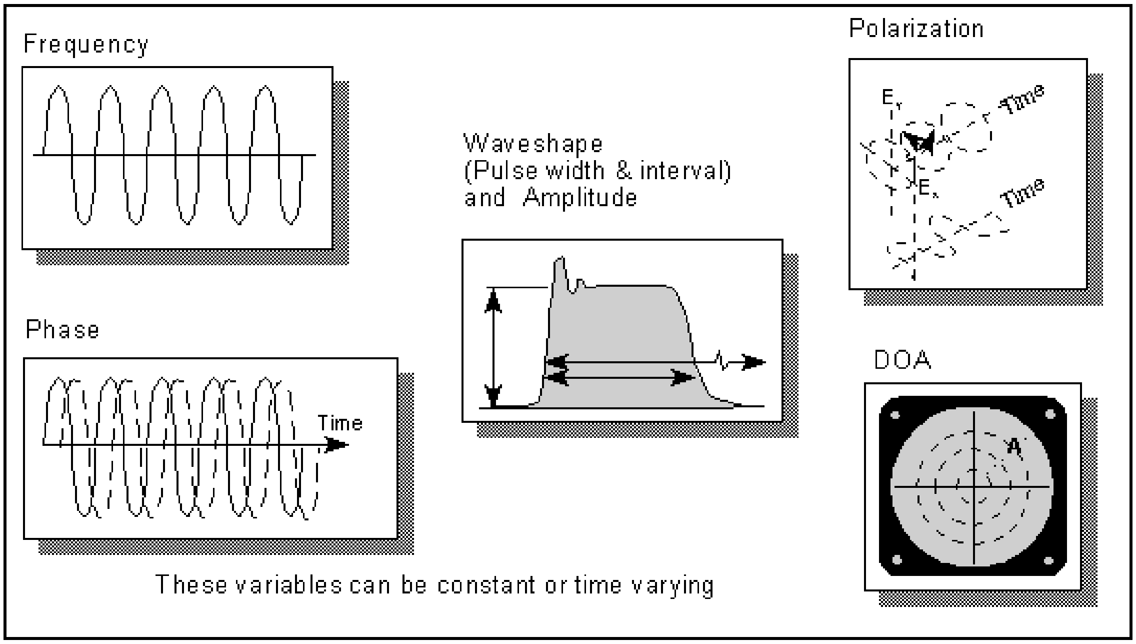
Frequency
Polarization
E_Y
Time
E_X
Time
Waveshape
(Pulse width & interval)
and Amplitude
Phase
Time
DOA
A
These variables can be constant or time varying

Figure 2. Information Content of an Electromagnetic Wave.

Table 1 illustrates the relative importance of several measured parameters during various stages of signal processing.

Table 1. Importance of Emitter Parameters During Signal Processing.

Parameter	Pulse Train De-interleavement	Emitter Identification	Intercept Correlation
Frequency	2	2	2
Amplitude	1	0	1
Angle of Arrival	2	0	2
TOA	0	0	1
PRI	2	2	2
PRI type	2	2	2
PW	2	1	1
Scan rate and type	0	2	1
Lobe Duration	0	1	1
0 Not Useful 1 Some Use 2 Very Useful			

Some emitter parameters can be measured using a single pulse; these parameters are referred to as monopulse parameters. The monopulse parameters include RF, PW, DOA, amplitude and TOA. RF can be determined on a pulse-by-pulse basis by receivers that can measure frequency. Frequency is very useful for emitter identification since most radars operate at a single frequency. Most real-time systems measure pulse width instead of pulse shape because the latter is much more difficult to characterize mathematically. Unfortunately, the apparent pulse width can be severely distorted by reflections, and consequently, its usefulness for emitter identification is limited. DOA cannot be used for emitter identification, but is excellent for sorting signals. A number of ESM systems use both frequency and DOA information to distinguish the new signals from the old (that is, known) ones. Amplitude also cannot be used for emitter identification. However, it can be used for sorting and for gross distance estimation using precompiled emitter's effective radiated power. Moreover, amplitude in conjunction with TOA can be used to determine the emitter's scan characteristics.

Other emitter parameters such as PRI, guidance and scan characteristics can be determined only by analyzing a group of pulses. All these parameters are useful for emitter identification; unfortunately, they require time for data collection and analysis, and call for sophisticated signal processing algorithms.

The problem of signal recognition in real-time is complicated by two factors: modulation of the signals and the very high pulse densities expected in the environment. Complex modulations (for example, inter-pulse RF modulation, intra-pulse RF modulation and agile Pulse Repetition Frequencies (PRFs)) present a significant pattern recognition problem for a number of ESM systems. It is expected that during some missions, hundreds of emitters will be transmitting simultaneously in the same vicinity. Wide-open antenna/receiver combination systems may have to cope with up to a million PPS. Even narrow-band receivers can expect data rates up to 100,000 PPS. At these rates, a single modern computer cannot be expected to process all the pulses, derive the characteristics for all emitters and identify the emitters in real-time. Other factors which encumber signal recognition include missing pulses, atmospheric noise and multiple reflections of pulses.

Present RWRs are designed primarily to cope with stable emitters. A stable emitter is one whose frequency and pulse repetition interval (PRI) remain relatively constant from pulse to pulse. The future threat will move steadily away from the stable emitter towards agile emitters which vary their frequency and PRI characteristics. The first change in this direction is towards the patterned agile emitter which varies its pulse and frequency parameters in accordance with a specific pattern. Examples of patterned agile emitters are MTI radars which use staggered PRFs, pulse Doppler radars which change frequency and PRF on a block-to-block basis, and certain frequency-agile radars whose transmitter frequency is mechanically modulated in a systematic pattern (e.g., spin-tuned magnetron). The next step in this evolution is towards truly agile emitters which change their frequency and PRF in a random manner on a pulse-to-pulse basis. One tempering factor in this evolution is that radars which process Doppler must maintain a constant frequency for at least two consecutive pulses.

In addition to agile frequency and PRI parameters, the future threat will be composed of a number of high-PRF pulsed Doppler, burst-frequency, CW, pulse-compression, agile-beam, and LPI radars, which use pseudo-noise waveforms. This conglomeration of radar types will cause a high signal density which must be segmented into a manageable data stream by the use of both frequency and spatial filtering in the RWR. While frequency and PRI are good parameters for sorting present-day non-agile emitters, they are poor or useless parameters for sorting agile emitters.

Angle of arrival is generally regarded as the best initial sorting parameter because it cannot be varied by the emitter from pulse to pulse.

PASSIVE DIRECTION FINDING AND EMITTER LOCATION

Direction finding (DF) systems provide several important functions in modern EW systems. We have already discussed the importance of measuring the emitter's bearing, or angle of arrival (AOA), as an invariant sorting parameter in the deinterleaving of radar signals and in separating closely spaced communication emitters. In addition, the conservation of jamming power in power-managed ECM systems depends on the ability of the associated ESM system to measure the direction to the victim emitter. A function which is becoming increasingly important in defense suppression and weapon delivery systems involves locating the emitter's position passively. This can be accomplished from a single moving platform through successive measurements of the emitter's angular direction, or from multiple platforms which make simultaneous angular measurements.

The emitter identification function requires identifying and associating consecutive pulses produced by the same emitter in angle of arrival (AOA) and frequency. The AOA is a parameter which a hostile emitter cannot change on a pulse-to-pulse basis. However, to measure the AOA of pulses which overlap in the time domain first requires them to be separated in the frequency domain. The advanced ESM receivers which accomplish this function must operate over several octaves of bandwidth while providing RMS bearing accuracies on the order of at least 2 degrees with high POI and fast reaction time in dense signal environments.

There are basically three methods, depicted in Figure 3, which allow the passive location of stationary ground-based emitters from airborne platforms.

These are:

1. The azimuth triangulation method where the intersection of successive spatially displaced bearing measurements provides the emitter location.

2. The azimuth/elevation location technique, which provides a single-pulse instantaneous emitter location from the intersection of the measured azimuth/elevation line with the earth's surface.

3. The time difference of arrival (TDOA), or precision emitter location system (PELS) method, which measures the difference in time of arrival of a single pulse at three spatially remote locations.

Additional methods include:

1. Phase rate of change, which is similar to triangulation, except it makes calculations using the phase derivative.

2. Angle distance techniques, where the distance from the emitter is derived from the signal strength (with known "threat" characteristics).

3. RF Doppler processing, which measures Doppler changes as the aircraft varies direction with respect to the "target" radar.

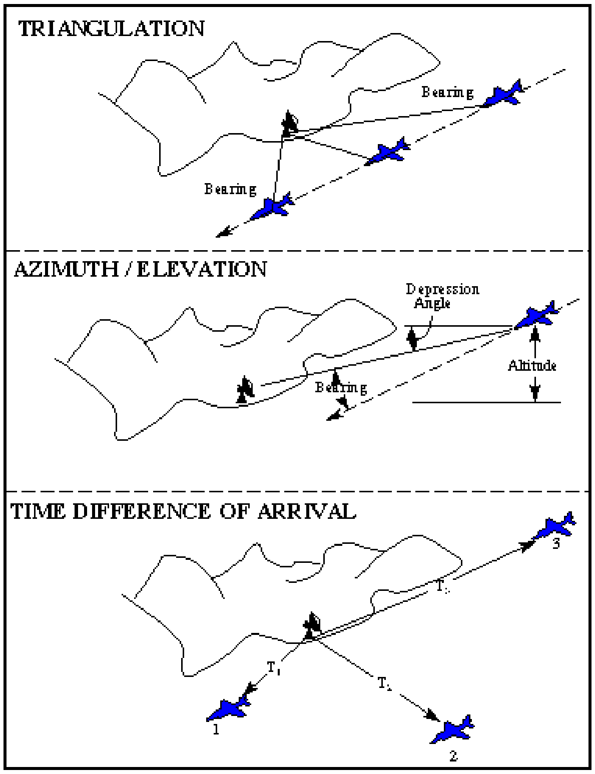

Figure 3. Passive Emitter Location Techniques.

The relative advantages and disadvantages of each are given in Table 2.

Table 2. Emitter Location Techniques.

Measurement Technique	Advantages	Disadvantages
Triangulation	Single Aircraft	Non-Instantaneous Location; Inadequate Accuracy for Remote Targeting; Not Forward Looking
Azimuth/Elevation	Single Aircraft Instantaneous Location Possible	Accuracy Degrades Rapidly at Low Altitude; Function of Range
Time Difference of Arrival (Pulsed Signals)	Very High Precision Can Support Weapon Delivery Position Requirements Very Rapid, Can Handle Short On-Time Threat	Very Complex, At Least 3 Aircraft; High Quality Receivers; DME (3 Sites); Very Wideband Data Link; Very High Performance Control Processor; Requires Very High Reliability Subsystems. Requires common time reference and correlation operation for non-pulse signals.

The triangulation method has the advantage of using a single aircraft, and its accuracy is greatest for a long baseline and the broadside geometry. The accuracy degenerates as the aircraft heading line approaches the boresight to the emitter.

The azimuth/elevation technique also has the advantage of using a single aircraft, but suffers from the difficultness of making an accurate elevation measurement with limited vertical aperture and in the presence of multipath effects.

The TDOA technique requires multiple aircraft and is complex, but has high potential accuracy. The determination of the location of the site involves the solution of at least two simultaneous second order equations for the intersection of two hyperbolas which represent $T_2 - T_1$ = Constant #1 and $T_3 - T_2$ = Constant #2. This method can be used to obtain a fix for an emitter which radiates only a single pulse.

ANGLE-OF-ARRIVAL (AOA) MEASUREMENTS

Several of the above DF measurements require AOA determination. Threat AOA measurements are also required to inform the aircrew in order to position the aircraft for optimal defense.

As shown in Figure 4, angle-of-arrival measuring systems fall into three main system categories of:

1. Scanning beam
2. Amplitude comparison or Simultaneous-multiple-beam
3. Phased Interferometer techniques

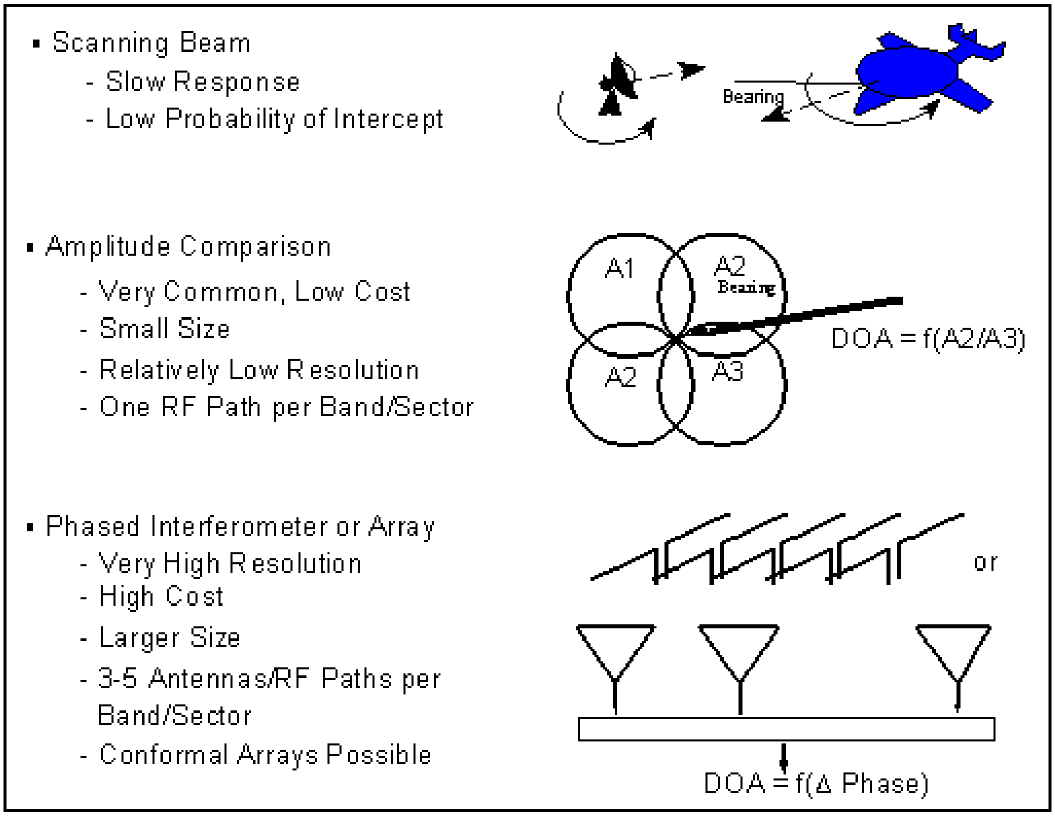

- Scanning Beam
  - Slow Response
  - Low Probability of Intercept
- Amplitude Comparison
  - Very Common, Low Cost
  - Small Size
  - Relatively Low Resolution
  - One RF Path per Band/Sector
- Phased Interferometer or Array
  - Very High Resolution
  - High Cost
  - Larger Size
  - 3-5 Antennas/RF Paths per Band/Sector
  - Conformal Arrays Possible

Figure 4. Angle-of-Arrival Measurement Techniques.

Scanning Beam

The mechanically scanning beam, or "spinner," requires only a single receiver and also exhibits high sensitivity due to the use of a directive antenna. The disadvantage is that the "spinner" usually exhibits slow response because it must rotate through the coverage angle (e.g., 360 degrees) to ensure that it intercepts an emitter. Also, if the emitter uses a scanning directional antenna, both beams must point at each other for maximum sensitivity, which is a low probability occurrence. Both of these effects cause the mechanically scanning beam technique to have a low probability of intercept (POI).

Amplitude Comparison

The two primary techniques used for direction finding are the amplitude-comparison method and the interferometer or phase-comparison method. The phase-comparison method generally has the advantage of greater accuracy, but the amplitude-comparison method is used extensively due to its lower complexity and cost. Regardless of which technique is used, it should be emphasized that the ultimate rms angular accuracy is given by:

$$\Delta\theta = \frac{k\,\theta_B}{\sqrt{SNR}}$$

where θ_B is the antenna's angular beamwidth, or interferometer lobe width, and SNR is the signal-to-noise ratio.

Thus, phase interferometers that typically use very widebeam antennas require high signal-to-noise ratios to achieve accurate angle-of-arrival measurements. Alternately, a multi-element array antenna can be used to provide relatively narrow interferometer lobes, which require modest signal-to-noise ratios.

Virtually all currently deployed radar warning receiving (RWR) systems use amplitude-comparison direction finding (DF). A basic amplitude-comparison receiver derives a ratio, and ultimately angle-of-arrival or bearing, from a pair of independent receiving channels, which utilize squinted antenna elements that are usually equidistantly spaced to provide an instantaneous 360° coverage. Typically, four or six antenna elements and receiver channels are used in such systems, and wideband logarithmic video detectors provide the signals for bearing-angle determination. The monopulse ratio is obtained by subtraction of the detected logarithmic signals, and the bearing is computed from the value of the ratio.

Amplitude comparison RWRs typically use broadband cavity-backed spiral antenna elements whose patterns can be approximated by Gaussian-shaped beams. Gaussian-shaped beams have the property that the logarithmic output ratio slope in dB is linear as a function of angle of arrival. Thus, a digital look-up table can be used to determine the angle directly. However, both the antenna beamwidth and squint angle vary with frequency over the multi-octave bands used in RWRs. Pattern shape variations cause a larger pattern crossover loss for high frequencies and a reduced slope sensitivity at low frequencies. Partial compensation of these effects, including antenna squint, can be implemented using a look-up table if frequency information is available in the RWR. Otherwise, gross compensation can be made, depending upon the RF octave band utilized.

Typical accuracies can be expected to range from 3 to 10 degrees rms for multi-octave frequency band amplitude-comparison systems which cover 360 degrees with four to six antennas.

The four-quadrant amplitude-comparison DF systems employed in RWRs have the advantage of simplicity, reliability, and low cost. Usually, only one antenna per quadrant is employed which covers the 2 to 18 GHz band. The disadvantages are poor accuracy and sensitivity, which result from the broad-beam antennas employed. Both accuracy and sensitivity can be improved by expanding the number of antennas employed. For example, expanding to eight antennas would double the accuracy and provide 3 dB more gain. As the number of antennas increases, it becomes appropriate to consider multiple-beam-forming antennas rather than just increasing the number of individual antennas. The geometry of multiple-beam-forming antennas is such that a conformal installation aboard an aircraft is difficult. Therefore, this type of installation is typically found on naval vessels or ground vehicles where the space is available to accommodate this type of antenna.

Simultaneous-multiple-beam (amplitude comparison)

The simultaneous-multiple-beam system uses an antenna, or several antennas, forming a number of simultaneous beams (e.g., Butler matrix or Rotman lens), thereby retaining the high sensitivity of the scanning antenna approach while providing fast response. However, it requires many parallel receiving channels, each with full frequency coverage. This approach is compatible with amplitude-monopulse angular measuring techniques which are capable of providing high angular accuracy.

A typical example of a multiple-beam antenna is a 16-element circular array developed as part of a digital ESM receiver. This system covers the range from 2 to 18 GHz with two antenna arrays (2 to 7.5 GHz and 7.5 to 18 GHz), has a sensitivity of -55 to -60 dBm and provides an rms bearing accuracy of better than 1.7 degrees on pulsewidths down to 100 ns.

Phased Interferometer Techniques

The term interferometer generally refers to an array type antenna in which large element spacing occurs and grating lobes appear.

Phase interferometer DF systems are utilized when accurate angle-of-arrival information is required. They have the advantage of fast response, but require relatively complex microwave circuitry, which must maintain a precise phase match over a wide frequency band under extreme environmental conditions. When high accuracy is required (on the order of 0.1 to 1°), wide baseline interferometers are utilized with ambiguity resolving circuitry. The basic geometry is depicted in Figure 5, whereby a plane wave arriving at an angle is received by one antenna earlier than the other due to the difference in path length.

The time difference can be expressed as a phase difference:

$$\varphi = \omega\Delta\tau = 2\pi a(f/c) = 2\pi\,(d \sin\theta)/\lambda,$$

where θ is the angle of arrival,
d is the antenna separation, and
λ is the wavelength in compatible units.

The unambiguous field of view (FOV) is given by $\theta = 2\sin^{-1}(\pi/2d)$, which for $\lambda/2$ spacing results in 180° coverage. This spacing must be established for the highest frequency to be received.

Interferometer elements typically use broad antenna beams with beamwidths on the order of 90°. This lack of directivity produces several adverse effects. First, it limits system sensitivity due to the reduced antenna gain. Secondly, it opens the system to interference signals from within the antenna's broad angular coverage. The interference signals often include multipath from strong signals which can limit the accuracy of the interferometer.

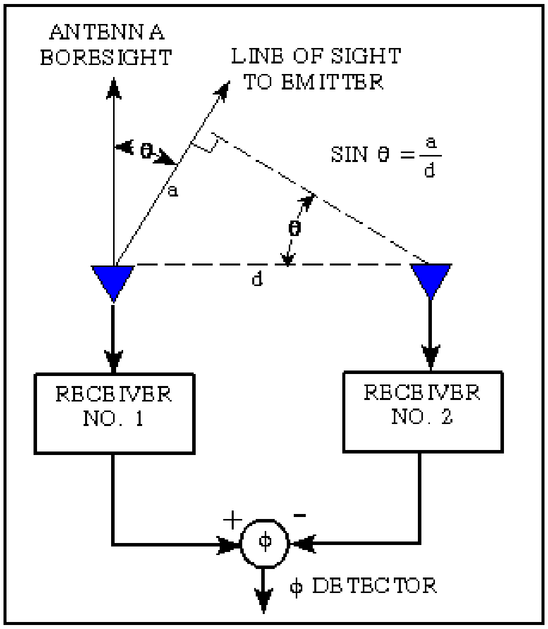

Figure 5. Phase Interferometer Principle.

In an interferometer, the locus of points that produce the same time or phase delay forms a cone. The indicated angle is the true azimuth angle multiplied by the cosine of the elevation angle. The error in assuming the incident angle to be the azimuth angle is negligible for signals near the antenna's boresight. At 45° azimuth and 10° elevation, the error is less than 1°, increasing to 15° for both at 45°. Two orthogonal arrays, one measuring the azimuth angle and the other the elevation angle can eliminate this error. For targets near the horizon, the depression angle is small, thereby requiring only horizontal arrays.

The rms angular accuracy of an interferometer in radians is given by:

$$\sigma_\theta = \Delta\alpha \,/\, (\pi \bullet \sqrt{SNR}),$$ where $\Delta\alpha = \lambda/(d\cdot\cos\theta)$ is the separation between adjacent nulls.

For a two-element interferometer, the spacing (d) must be $\lambda/2$ or less to provide unambiguous, or single lobe ± 90°, coverage. This, in effect, sets a wide interferometer (or grating) lobe which must be split by a large factor to achieve high accuracy. This, in turn, imposes a requirement for high SNR to achieve the large beam-splitting factor. For example, if 0.1° accuracy is required from an unambiguous two-element interferometer, then a SNR of about 50 dB is required to achieve this accuracy. This may be difficult to achieve considering the inherently low sensitivity of an interferometer system.

When high accuracy is required from an interferometer system, it is usual to employ separations greater than $\lambda/2$. The increased separation sets up a multi-grating-lobe structure through the coverage angle which requires less SNR to achieve a specified accuracy. For example, a two-element interferometer with 16λ spacing would set up a 33-grating-lobe structure (including the central lobe) throughout the ± 90° coverage angle. Within each of the 33 grating lobes, it would only require a SNR on the order of 20 dB to achieve 0.1° accuracy. However, there would be 33 ambiguous regions within the ± 90° angular coverage and also 32 nulls (where the phase detector output is zero), about which the system would be insensitive to an input signal. The ambiguities could be resolved by employing a third antenna element with $\lambda/2$ spacing, which would provide an accuracy on the order of 3° with 20 dB SNR. This accuracy is sufficient to identify which of the 33 lobes contains the signal. Providing coverage in the null regions requires additional antenna elements.

Interferometers employing multiple antenna elements are called multiple-baseline interferometers. In a typical design, the receiver consists of a reference antenna and a series of companion antennas. The spacing between the reference element and the first companion antenna is $\lambda/2$; other secondary elements are placed to form pairs separated by 1, 2, 4, and 8 wavelengths. The initial AOA is measured unambiguously by the shortest-spaced antenna pair. The next greatest spaced pair has a phase rate of change which is twice that of the first, but the information is ambiguous due to there being twice as many lobes as in the preceding pair. A greater phase rate of change permits higher angular accuracy while the ambiguity is resolved by the previous pair. Thus, the described multiple-baseline interferometer provides a binary AOA measurement where each bit of the measurement supplies a more accurate estimate of the emitter's AOA.

Harmonic multiple-baseline interferometers use elements which are spaced at $2^n \cdot \lambda/2$, with $n = 0, 1, 2, 3$. In nonharmonic interferometers, no pair of antennas provides a completely unambiguous reading over the complete field of view. For example, the initial spacing in the nonharmonic interferometer might be λ, while the next companion element spacing is $3\lambda/2$. Ambiguities are resolved by truth tables, and hence the accuracy is set by the spacing of the widest baseline antenna pair. Nonharmonic interferometers have been implemented over 9:1 bandwidths (2 to 18 GHz) with rms accuracies from 0.1 to 1° and with no ambiguities over ± 90°. The principal advantage of the nonharmonic over the harmonic interferometer is the increased bandwidth for unambiguous coverage.

Interferometer DF accuracy is determined by the widest baseline pair. Typical cavity-backed spirals, track to 6 electrical degrees, and associated receivers track to 9°, resulting in an rms total of 11°. At a typical 16 dB SNR, the rms phase noise is approximately 9 electrical degrees. For these errors and an emitter angle of 45°, a spacing of 25λ is required for 0.1° rms accuracy while a spacing of 2.5λ is needed for 1° accuracy. For high accuracy, interferometer spacings of many feet are required. In airborne applications, this usually involves mounting interferometer antennas in the aircraft's wingtips.

The characteristics of typical airborne amplitude comparison and phase interferometer DF systems are summarized in Table 3. The phase interferometer system generally uses superheterodyne receivers which provide the necessary selectivity and sensitivity for precise phase measurements.

Table 3. Direction of Arrival Measurement Techniques.

	Amplitude Comparison	Phase Interferometer
Sensor Configuration	Typically 4 to 6 Equispaced Antenna Elements for 360° Coverage	2 or more RHC or LHC Spirals in Fixed Array
DF Accuracy	$\mathrm{DF}_{\mathrm{ACC}} \approx \frac{\theta_{BW}^{2}\,\Delta C_{dB}}{24S}$ (Gaussian Shape)	$\mathrm{DF}_{\mathrm{ACC}} = \frac{\lambda}{2\pi d\cos\theta}\Delta\theta$
DF Accuracy Improvement	Decrease Antenna BW Decrease Amplitude Mistrack Increase Squint Angle	Increase Spacing of Outer Antennas; Decrease Phase Mistrack
Typical DF Accuracy	3° to 10° rms	0.1° to 3° rms
Sensitivity to Multipath/ Reflections	High Sensitivity Mistrack of Several dB Can Cause Large DF Errors	Relatively Insensitive; Interferometer Can Be Made to Tolerate Large Phase Errors
Platform Constraints	Locate in Reflection Free Area	Reflection Free Area; Real Estate For Array; Prefers Flat Radome
Applicable Receivers	Crystal Video; Channelizer; Acousto-Optic; Compressive; Superheterodyne	Superheterodyne

ΔC_{dB} = Amplitude Monopulse Ratio in dB
S = Squint Angle in degrees
θ_{BW} = Antenna Beamwidth in degrees

This page intentionally left blank.

MICROWAVE / RF COMPONENTS

Microwave Waveguides and Coaxial Cable 6-1
Voltage Standing Wave Ratio (VSWR) / Reflection Coefficient
    Return Loss / Mismatch Loss 6-2
Microwave Coaxial Connectors 6-3
Power Dividers and Directional Couplers 6-4
Attenuators / Filters / DC Blocks 6-5
Terminations / Dummy Loads 6-6
Circulators and Diplexers 6-7
Mixers and Frequency Discriminators 6-8
Detectors 6-9
RF / Microwave Amplifiers 6-10
Signal Generation 6-11
Digital Processing Components 6-12
Microwave Measurements 6-13

This page intentionally left blank.

MICROWAVE WAVEGUIDES and COAXIAL CABLE

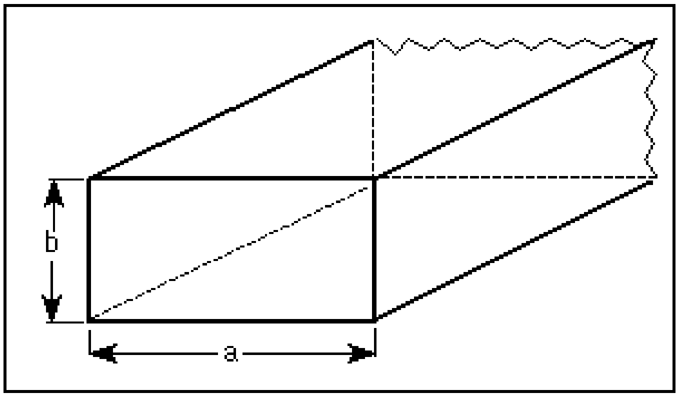

Figure 1. The Rectangular Waveguide.

In general, a waveguide consists of a hollow metallic tube of arbitrary cross section uniform in extent in the direction of propagation. Common waveguide shapes are rectangular, circular, and ridged. The rectangular waveguide has a width a and height b as shown in Figure 1. Commonly used rectangular waveguides have an aspect ratio b/a of approximately 0.5. Such an aspect ratio is used to preclude generation of field variations with height and their attendant unwanted modes. Waveguides are used principally at frequencies in the microwave range; inconveniently large guides would be required to transmit radio-frequency power at longer wavelengths. In the X-Band frequency range of 8.2 to 12.4 GHz, for example, the U.S. standard rectangular waveguide, WR-90, has an inner width of 2.286 cm (0.9 in.) and an inner height of 1.016 cm (0.4 in.).

In waveguides the electric and magnetic fields are confined to the space within the guides. Thus no power is lost to radiation. Since the guides are normally filled with air, dielectric losses are negligible. However, there is some I^2R power lost to heat in the walls of the guides, but this loss is usually very small.

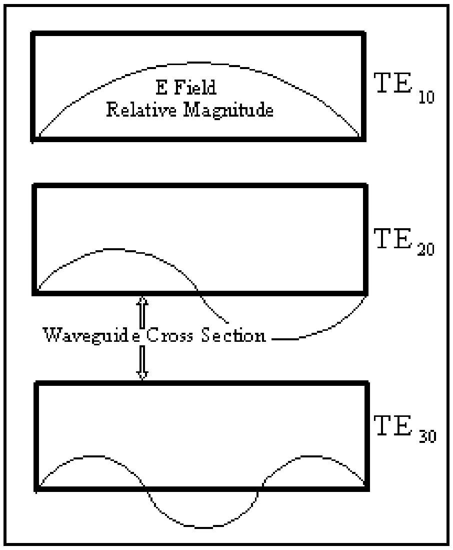

Figure 2. TE Modes.

It is possible to propagate several modes of electromagnetic waves within a waveguide. The physical dimensions of a waveguide determine the cutoff frequency for each mode. If the frequency of the impressed signal is above the cutoff frequency for a given mode, the electromagnetic energy can be transmitted through the guide for that particular mode with minimal attenuation. Otherwise the electromagnetic energy with a frequency below cutoff for that particular mode will be attenuated to a negligible value in a relatively short distance. This grammatical use of cutoff frequency is opposite that used for coaxial cable, where cutoff frequency is for the highest useable frequency. The dominant mode in a particular waveguide is the mode having the lowest cutoff frequency. For rectangular waveguide this is the TE_{10} mode. The TE (transverse electric) signifies that all electric fields are transverse to the direction of propagation and that no longitudinal electric field is present. There is a longitudinal component of magnetic field and for this reason the TE_{mn} waves are also called H_{mn} waves. The TE designation is usually preferred. Figure 2 shows a graphical depiction of the E field variation in a waveguide for the TE_{10}, TE_{20}, and TE_{30} modes. As can be seen, the first index indicates the number of half wave loops across the width of the guide and the second index, the number of loops across the height of the guide - which in this case is zero. It is advisable to choose the dimensions of a guide in such a way that, for a given input signal, only the energy of the dominant mode can be transmitted through the guide. For example, if for a particular frequency, the width of a rectangular guide is too large, then the TE_{20} mode can propagate causing a myriad of problems. For rectangular guides of low aspect ratio the TE_{20} mode is the next higher order mode and is harmonically related to the cutoff frequency of the TE_{10} mode. It is this relationship together with attenuation and propagation considerations that determine the normal operating range of rectangular waveguide.

The discussion on circular waveguides will not be included because they are rarely used in the EW area. Information regarding circular waveguides can be found in numerous textbooks on microwaves.

CHARACTERISTICS OF STANDARD RECTANGULAR WAVEGUIDES

Rectangular waveguides are commonly used for power transmission at microwave frequencies. Their physical dimensions are regulated by the frequency of the signal being transmitted. Table 1 tabulates the characteristics of the standard rectangular waveguides. It may be noted that the number following the EIA prefix "WR" is in inside dimension of the widest part of the waveguide (i.e., WR90 has an inner dimension of 0.90").

DOUBLE RIDGE RECTANGULAR WAVEGUIDE

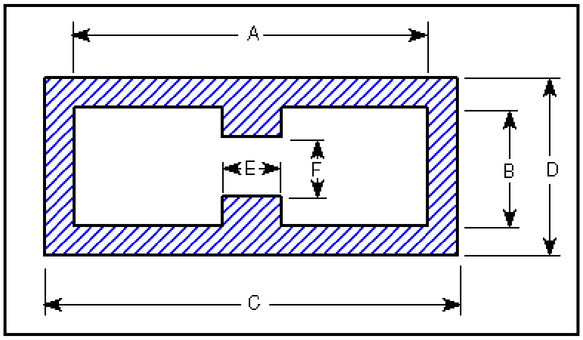

Figure 3. Double Ridge Waveguide.

Another type of waveguide commonly used in EW systems is the double ridge rectangular waveguide. The ridges in this waveguide increase the bandwidth of the guide at the expense of higher attenuation and lower power-handling capability. The bandwidth can easily exceed that of two contiguous standard waveguides. Introduction of the ridges mainly lowers the cutoff frequency of the TE_{10} mode from that of the unloaded guide, which is predicated on width alone. The reason for this can easily be explained when the field configuration in the guide at cutoff is investigated. At cutoff there is no longitudinal propagation down the guide. The waves simply travel back and forth between the side walls of the guide. In fact the guide can be viewed as a composite parallel plate waveguide of infinite width where the width corresponds to the direction of propagation of the normal guide. The TE_{10} mode cutoff occurs where this composite guide has its lowest-order resonant frequency. This occurs when there is only one E field maximum across the guide which occurs at the center for a symmetrical ridge. Because of the reduced height of the guide under the ridge, the effective TE_{10} mode resonator is heavily loaded as though a shunt capacitor were placed across it. The cutoff frequency is thus lowered considerably. For the TE_{20} mode the fields in the center of the guide will be at a minimum. Therefore the loading will have a negligible effect. For guides of proper aspect ratio, ridge height, and ridge width, an exact analysis shows that the TE_{10} mode cutoff can be lowered substantially at the same time the TE_{20} and TE_{30} mode cutoffs are raised slightly. Figure 3 shows a typical double ridged waveguide shape and Table 2 shows double ridged waveguide specifications. In the case of ridged waveguides, in the EIA designation, (WRD350 D36) the first "D" stands for double ridged ("S" for single ridged), the 350 is the starting frequency (3.5 GHz), and the "D36" indicates a bandwidth of 3.6:1. The physical dimensions and characteristics of a WRD350 D24 and WRD350 D36 are radically different. A waveguide with a MIL-W-23351 dash number beginning in 2 (i.e., 2-025) is a double ridge 3.6:1 bandwidth waveguide. Likewise a 1- is a single ridge 3.6:1, a 3- is a single ridge 2.4:1, and a 4- is a double ridge 2.4:1 waveguide.

Figure 4 shows a comparison of the frequency /attenuation characteristics of various waveguides. The attenuation is based on real waveguides which is higher than the theoretical values listed in Tables 1 and 2. Figure 5 shows photographs of waveguides with some common connectors. Figure 6 shows attenuation characteristics of various RF coaxial cables.

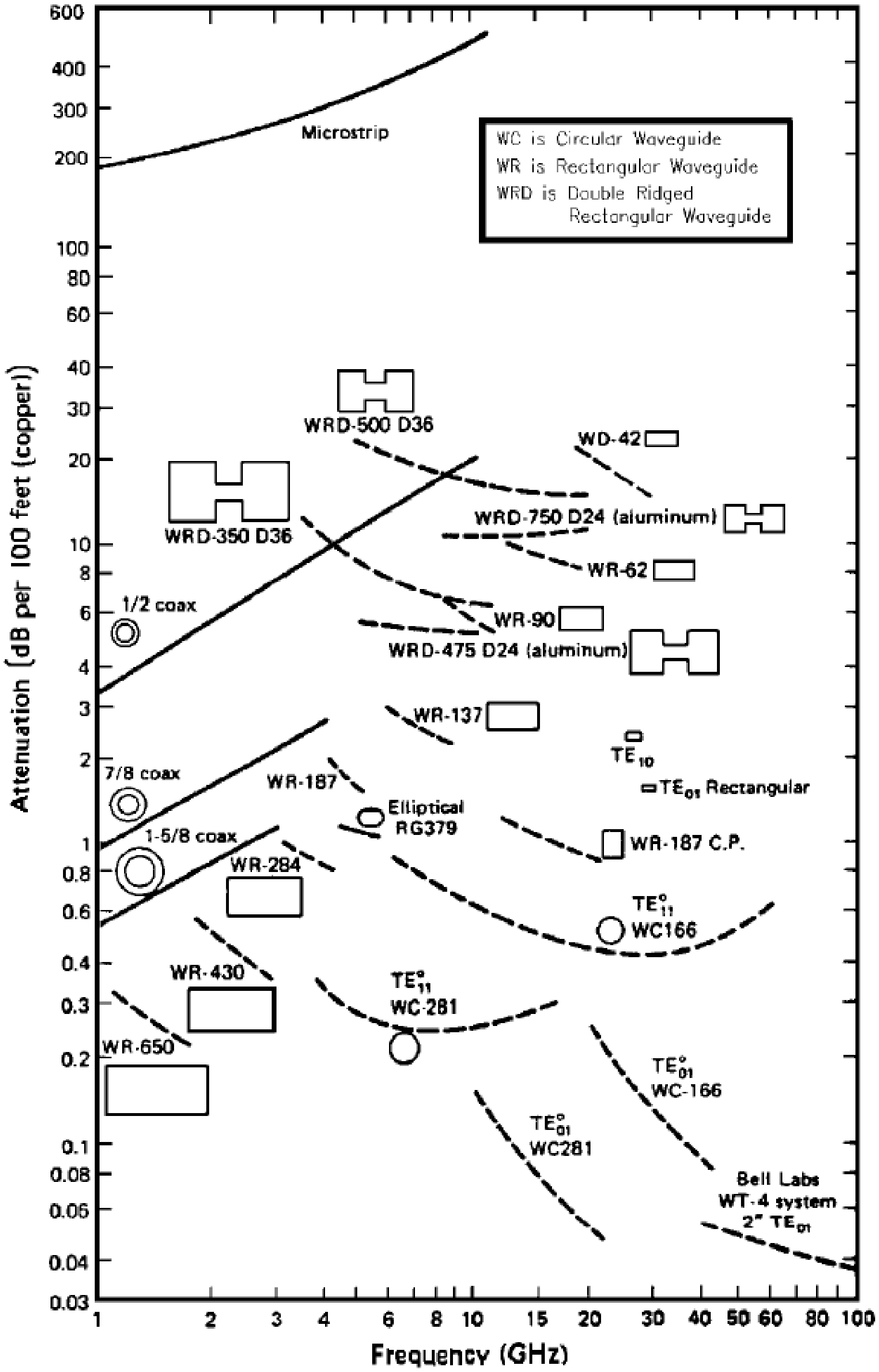

Figure 4. Attenuation vs. Frequency for a Variety of Waveguides and Cables.

Table 1. Rectangular Waveguide Specifications.

Waveguide Size	JAN WG Desig	MIL-W-85 Dash #	Material	Freq Range (GHz)	Freq Cutoff (GHz)	Power (at 1 Atm)		Insertion Loss (dB/100ft)	Dimensions (Inches)	
						CW	Peak		Outside	Wall Thickness
WR284	RG48/U RG75/U	1-039 1-042	Copper Aluminum	2.60 - 3.95	2.08	45 36	7650	.742-.508 1.116-.764	3.000x1.500	0.08
WR229	RG340/U RG341/U	1-045 1-048	Copper Aluminum	3.30 - 4.90	2.577	30 24	5480	.946-.671 1.422-1.009	2.418x1.273	0.064
WR187	RG49/U RG95/U	1-051 1-054	Copper Aluminum	3.95 - 5.85	3.156	18 14.5	3300	1.395-.967 2.097-1.454	1.000x1.000	0.064
WR159	RG343/U RG344/U	1-057 1-060	Copper Aluminum	4.90 - 7.05	3.705	15 12	2790	1.533-1.160 2.334-1.744	1.718x0.923	0.064
WR137	RG50/U RG106/U	1-063 1-066	Copper Aluminum	5.85 - 8.20	4.285	10 8	1980	1.987-1.562 2.955-2.348	1.500x0.750	0.064
WR112	RG51/U RG68/U	1-069 1-072	Copper Aluminum	7.05 - 10.0	5.26	6 4.8	1280	2.776-2.154 4.173-3.238	1.250x0.625	0.064
WR90	RG52/U RG67/U	1-075 1-078	Copper Aluminum	8.2 - 12.4	6.56	3 2.4	760	4.238-2.995 6.506-4.502	1.000x0.500	0.05
WR75	RG346/U RG347/U	1-081 1-084	Copper Aluminum	10.0 - 15.0	7.847	2.8 2.2	620	5.121-3.577 7.698-5.377	0.850x0.475	0.05
WR62	RG91/U RG349/U	1-087 1-091	Copper Aluminum	12.4 - 18.0	9.49	1.8 1.4	460	6.451-4.743 9.700-7.131	0.702x0.391	0.04
WR51	RG352/U RG351/U	1-094 1-098	Copper Aluminum	15.0 - 22.0	11.54	1.2 1	310	8.812-6.384 13.250-9.598	0.590x0.335	0.04
WR42	RG53/U	1-100	Copper	18.0 - 26.5	14.08	0.8	170	13.80-10.13	0.500x0.250	0.04
WR34	RG354/U	1-107	Copper	2.0 - 33.0	17.28	0.6	140	16.86-11.73	0.420x0.250	0.04
WR28	RG271/U	3-007	Copper	26.5 - 40.0	21.1	0.5	100	23.02-15.77	0.360x0.220	0.04

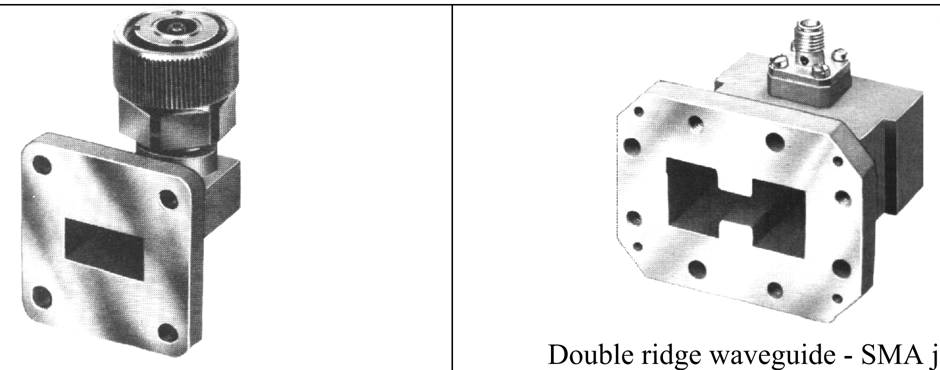

Standard waveguide - 7 mm

Double ridge waveguide - SMA jack

Figure 5. Waveguides With Some Common Connections.

Table 2. Double Ridge Rectangular Waveguide Specifications.

Waveguide Size	MIL-W-23351 Dash #	Material	Freq Range (GHz)	Freq Cutoff (GHz)	Power (at 1 Atm)		Insertion Loss (dB/ft)	Dimensions (Inches)					
					CW	Peak		A	B	C	D	E	F
WRD250		Alum Brass Copper Silver Al	2.60 - 7.80	2.093	24	120	0.025 0.025 0.018 0.019	1.655	0.715	2	1	0.44	0.15
WRD350 D24	4-029 4-303 4-031	Alum Brass Copper	3.50 - 8.20	2.915	18	150	0.0307 0.0303 0.0204	1.48	0.688	1.608	0.816	0.37	0.292
WRD475 D24	4-033 4-034 4-035	Alum Brass Copper	4.75 - 11.00	3.961	8	85	0.0487 0.0481 0.0324	1.09	0.506	1.19	0.606	0.272	0.215
WRD500 D36	2-025 2-026 2-027	Alum Brass Copper	5.00 - 18.00	4.222	4	15	0.146 0.141 0.095	0.752	0.323	0.852	0.423	0.188	0.063
WRD650		Alum Brass Copper	6.50 - 18.00	5.348	4	25	0.106 0.105 0.07	0.720	0.321	0.820	0.421	0.173	0.101
WRD750 D24	4-037 4-038 4-039	Alum Brass Copper	7.50 - 18.00	6.239	4.8	35	0.0964 0.0951 0.0641	0.691	0.321	0.791	0.421	0.173	0.136
WRD110 D24	4-041 4-042 4-043	Alum Brass Copper	11.00 - 26.50	9.363	1.4	15	0.171 0.169 0.144	0.471	0.219	0.551	0.299	0.118	0.093
WRD180 D24	4-045 4-046 4-047	Alum Brass Copper	18.00 - 40.00	14.995	0.8	5	0.358 0.353 0.238	0.288	0.134	0.368	0.214	0.072	0.057

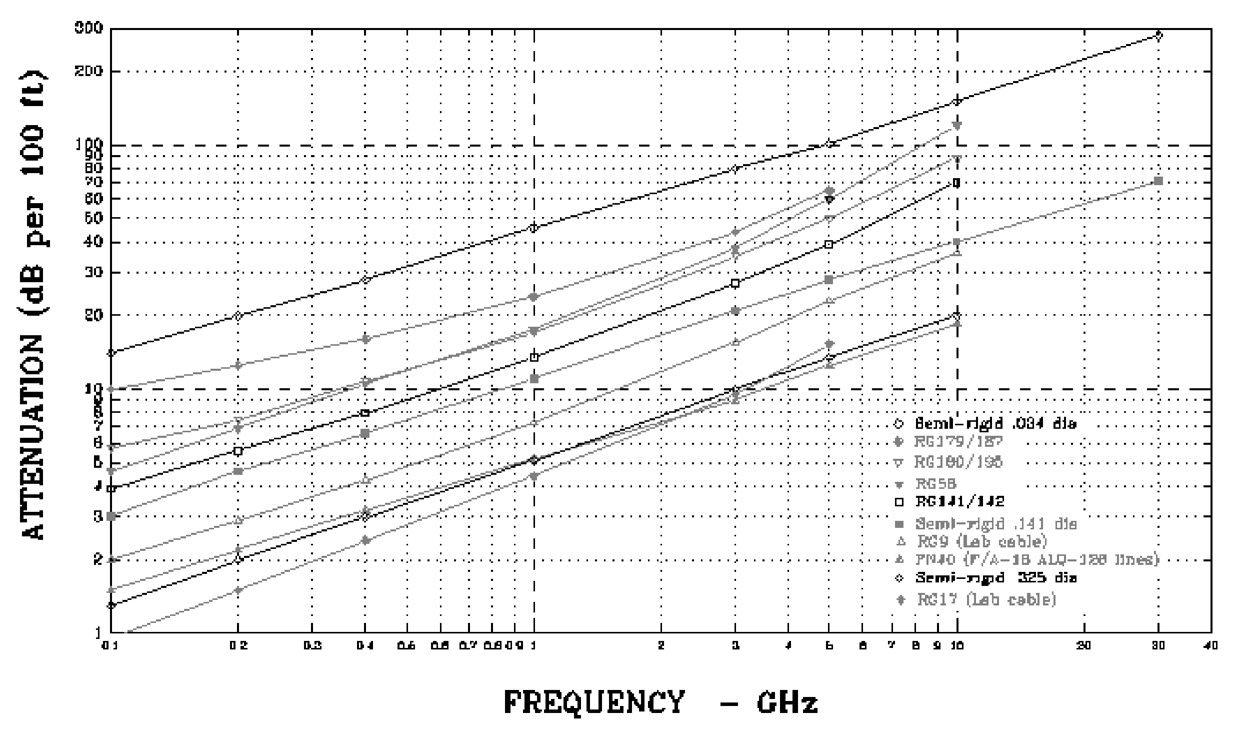

Figure 6. Attenuation vs. Frequency for a Variety of Coaxial Cables.

This page intentionally left blank.

VOLTAGE STANDING WAVE RATIO (VSWR) / REFLECTION COEFFICIENT RETURN LOSS / MISMATCH LOSS

When a transmission line is terminated with an impedance, Z_L, that is not equal to the characteristic impedance of the transmission line, Z_O, not all of the incident power is absorbed by the termination. Part of the power is reflected back so that phase addition and subtraction of the incident and reflected waves creates a voltage standing wave pattern on the transmission line. The ratio of the maximum to minimum voltage is known as the Voltage Standing Wave Ratio (VSWR) and successive maxima and minima are spaced by 180° (λ/2).

$$VSWR = \frac{E_{\max}}{E_{\min}} = \frac{E_i + E_r}{E_i - E_r}$$

where E_{max} = maximum voltage on the standing wave
E_{min} = minimum voltage on the standing wave
Ei = incident voltage wave amplitude
Er = reflected voltage wave amplitude

The reflection coefficient, ρ, is defined as E_r/E_i and in general, the termination is complex in value, so that ρ will be a complex number.

Additionally we define: $\Gamma = \frac{Z_L - Z_O}{Z_L + Z_O}$ The refection coefficient, ρ, is the absolute value of the magnitude of Γ. If the equation for VSWR is solved for the reflection coefficient, it is found that:

$$\begin{matrix}\textit{Reflection} \\ \textit{Coefficient}\end{matrix} = \rho = |\Gamma| = \frac{VSWR - 1}{VSWR + 1} \qquad \text{Consequently, } VSWR = \frac{1 + \rho}{1 - \rho}$$

The return loss is related through the following equations:

$$\begin{matrix}\textit{Return} \\ \textit{Loss}\end{matrix} = 10 \log\left[\frac{P_i}{P_r}\right] = -20 \log\left[\frac{E_r}{E_i}\right] = -20 \log\left[\frac{VSWR - 1}{VSWR + 1}\right] = -20 \log \rho$$

Return loss is a measure in dB of the ratio of power in the incident wave to that in the reflected wave, and as defined above always has a positive value. For example if a load has a Return Loss of 10 dB, then 1/10 of the incident power is reflected. The higher the return loss, the less power is actually lost.

Also of considerable interest is the Mismatch Loss. This is a measure of how much the transmitted power is attenuated due to reflection. It is given by the following equation:

Mismatch Loss = $-10 \log (1 - \rho^2)$

For example, an antenna with a VSWR of 2:1 would have a reflection coefficient of 0.333, a mismatch loss of 0.51 dB, and a return loss of

VSWR	Return Loss (dB)	% Power / Voltage Loss	Reflection Coefficient	Mismatch Loss (dB)
1	∞	0 / 0	0	0.000
1.15	23.1	0.49 / 7.0	0.07	.021
1.25	19.1	1.2 / 11.1	0.111	.054
1.5	14.0	4.0 / 20.0	0.200	.177
1.75	11.3	7.4 / 27.3	0.273	.336
1.9	10.0	9.6 / 31.6	0.316	.458
2.0	9.5	11.1 / 33.3	0.333	.512
2.5	7.4	18.2 / 42.9	0.429	.880
3.0	6.0	25.1 / 50.0	0.500	1.25
3.5	5.1	30.9 / 55.5	0.555	1.6
4.0	4.4	36.3 / 60.0	0.600	1.94
4.5	3.9	40.7 / 63.6	0.636	2.25
5.0	3.5	44.7 / 66.6	0.666	2.55
10	1.7	67.6 / 81.8	0.818	4.81
20	0.87	81.9 / 90.5	0.905	7.4
100	0.17	96.2 / 98.0	0.980	14.1
∞	.000	100 / 100	1.00	∞

* Divide % Voltage loss by 100 to obtain ρ (reflection coefficient)

9.54 dB (11% of your transmitter power is reflected back). In some systems this is not a trivial amount and points to the need for components with low VSWR.

If 1000 watts (60 dBm/ 30 dBW) is applied to this antenna, the return loss would be 9.54 dB. Therefore, 111.1 watts would be reflected and 888.9 watts (59.488 dBm/29.488 dBW) would be transmitted, so the mismatch loss would be 0.512 dB.

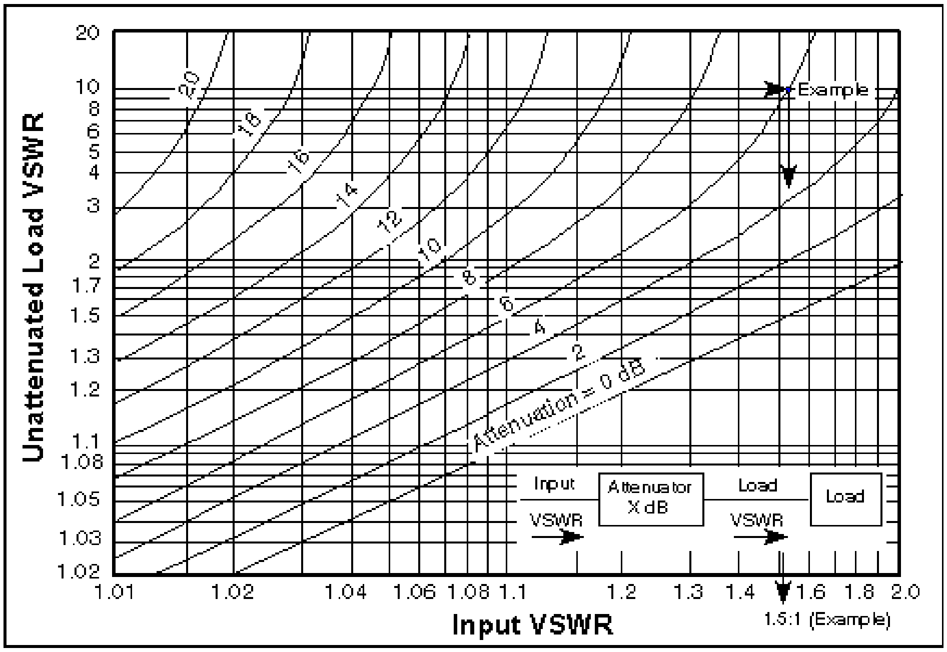

Figure 1. Reduction of VSWR by Attenuation.

Transmission line attenuation improves the VSWR of a load or antenna. For example, a transmitting antenna with a VSWR of 10:1 (poor) and a line loss of 6 dB would measure 1.5:1 (okay) if measured at the transmitter. Figure 1 shows this effect.

Therefore, if you are interested in determining the performance of antennas, the VSWR should always be measured at the antenna connector itself rather than at the output of the transmitter. Transmit cabling will load the line and create an illusion of having a better antenna VSWR. Transmission lines should have their insertion loss (attenuation) measured in lieu of VSWR, but VSWR measurements of transmission lines are still important because connection problems usually show up as VSWR spikes.

Historically VSWR was measured by probing the transmission line. From the ratio of the maximum to minimum voltage, the reflection coefficient and terminating impedance could be calculated. This was a time consuming process since the measurement was at a single frequency and mechanical adjustments had to be made to minimize coupling into circuits. Problems with detector characteristics also made the process less accurate. The modern network analyzer system sweeps very large frequency bandwidths and measures the incident power, P_i, and the reflected power, P_r. Because of the considerable computing power in the network analyzer, the return loss is calculated from the equation given previously, and displayed in real time. Optionally, the VSWR can also be calculated from the return loss and displayed real time.

If a filter is needed on the output of a jammer, it is desirable to place it approximately half way between the jammer and antenna. This may allow the use of a less expensive filter, or a reflective filter vs. an absorptive filter.

Special cases exist when comparing open and shorted circuits. These two conditions result in the same ∞ VSWR and zero dB return loss even though there is a 180° phase difference between the reflection coefficients. These two conditions are used to calibrate a network analyzer.

MICROWAVE COAXIAL CONNECTORS

For high-frequency operation, the average circumference of a coaxial cable must be limited to about one wavelength in order to reduce multimodal propagation and eliminate erratic reflection coefficients, power losses, and signal distortion.

Except for the sexless APC-7 connector, all other connectors are identified as either male (plugs) which have a center conductor that is a probe or female (jacks) which have a center conductor that is a receptacle. Sometimes it is hard to distinguish them as some female jacks may have a hollow center “pin” which appears to be male, yet accepts a smaller male contact.

An adapter is an approximately zero loss interface between two connectors and is called a barrel when both connectors are identical. A number of common of coaxial connectors are described below, however other special purpose connectors exist, including blind mate connectors where spring fingers are used in place of threads to obtain shielding (desired connector shielding should be at least 90 dB).

APC-2.4 (2.4mm) - The 50 Ω APC-2.4 (Amphenol Precision Connector-2.4 mm) is also known as an OS-50 connector. It was designed to operate at extremely high microwave frequencies (up to 50 GHz).	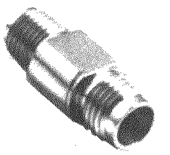 2.4 mm jack - 3.5 mm jack
APC-3.5 (3.5mm) - The APC-3.5 connector provides repeatable connections and has a very low VSWR. Either the male or female end of this 50 Ω connector can mate with the opposite type of SMA connector. The APC-3.5 connector can work at frequencies up to 34 GHz.	
APC-7 (7mm) - The APC-7 was developed by HP, but has been improved and is now manufactured by Amphenol. The connector provides a coupling mechanism without male or female distinction and is the most repeatable connecting device used for very accurate 50 Ω measurement applications. Its VSWR is extremely low up to 18 GHz. Other companies have 7mm series available.	 7mm - 3.5mm plug
BNC (OSB) - The BNC (Bayonet Neill-Concelman or Bayonet Navy Connector) was originally designed for military system applications during World War II. The connector operates best at frequencies up to about 4 GHz; beyond that it tends to radiate electromagnetic energy. The BNC can accept flexible cables with diameters of up to 6.35 mm (0.25 in.) and characteristic impedance of 50 to 75 Ω. It is now the most commonly used connector for frequencies under 1 GHz. Other names the BNC has picked up over the years include: “Baby Neill-Concelman”, “Baby N connector”, “British Naval Connector”, and “Bayonet Nut Connector”	 BNC male plug (above) BNC female jack

C - The C (Concelman) coaxial connector is a medium size, older type constant 50 Ω impedance. It has a bayonet (twist and lock) connection. It is larger than the BNC, but about the same as Type N. It has a frequency range of 0-11 GHz	C Plug (top) 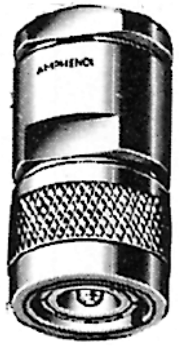 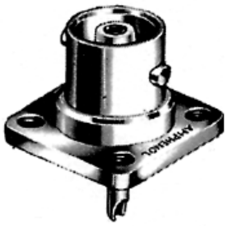 C jack (bottom)
SC (OSSC) - The SC type connector is a screw version of the "C" connector.	
SMA (OSM/3mm) - The SMA (Sub-Miniature A) connector was originally designed by Bendix Scintilla Corporation, but it has been manufactured by the Omni-Spectra division of M/ACOM (as the OSM connector) and many other electronic companies. It is a 50 Ω threaded connector. The main application of SMA connectors is on components for microwave systems. The connector normally has a frequency range to 18 GHz, but high performance varieties can be used to 26.5 GHz.	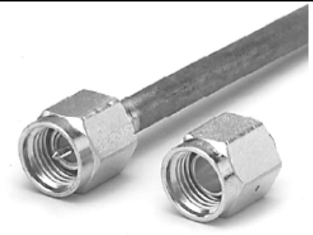 SMA Plug (above) 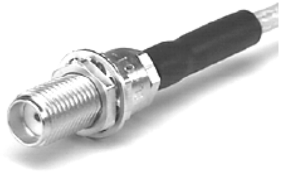 SMA Jack
SSMA (OSSM) - The SSMA is a microminiature version of the SMA. It is also 50 Ω and operates to 26.5 GHz with flexible cable or 40 GHz with semi-rigid cable.	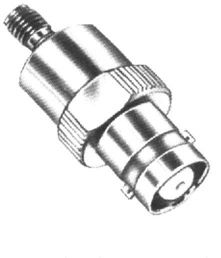 SSMA jack - BNC jack

SMC (OSMC) - The SMC (Sub-Miniature C) is a 50 Ω or 75 Ω connector that is smaller than the SMA. The connector can accept flexible cables with diameters of up to 3.17 mm (0.125 in.) for a frequency range of up to 7-10 GHz.	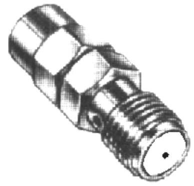 SMC Plug - SMA Jack (Some call this a SMC Jack, even though it has a female connector)
SMB (OSMB) - The SMB is like the SMC except it uses quick disconnect instead of threaded fittings. It is a 50 / 75 Ω connector which operates to 4 GHz with a low reflection coefficient and is useable to 10 GHz.	 SMB Plug (above) SMB Jack
TNC (OST) - The TNC (Threaded Neill-Concelman or threaded Navy Connector) is merely a threaded BNC. The function of the thread is to stop radiation at higher frequencies, so that the connector can work at frequencies up to 12 GHz (to 18 GHz when using semi-rigid cable). It can be 50 or 75 Ω. For size comparison here is a SMA plug - TNC plug 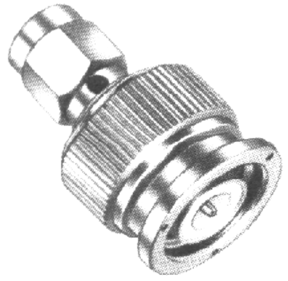	 TNC Plug (above) TNC Jack

Type N (OSN) - The 50 or 75 Ω Type N (Navy) connector was originally designed for military systems during World War II and is the most popular measurement connector for the frequency range of 1 to 11 GHz. The precision 50 Ω APC-N and other manufacturers high frequency versions operate to 18 GHz.

For size comparison here is a picture of Type N plug - TNC jack

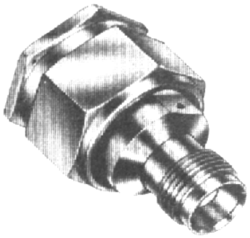

N Plug (above)

N Jack

Note: Always rotate the movable coupling nut of the plug, not the cable or fixed connector, when mating connectors. Since the center pin is stationary with respect to the jack, rotating the jack puts torque on the center pin. With TNC and smaller connectors, the center pin will eventually break off.

An approximate size comparison of these connectors is depicted below (not to scale).

Large ======================= Medium =========================== Small
SC 7mm N TNC/BNC 3.5mm SMA 2.4mm SSMA SMC

Figure 1 shows the frequency range of several connectors.

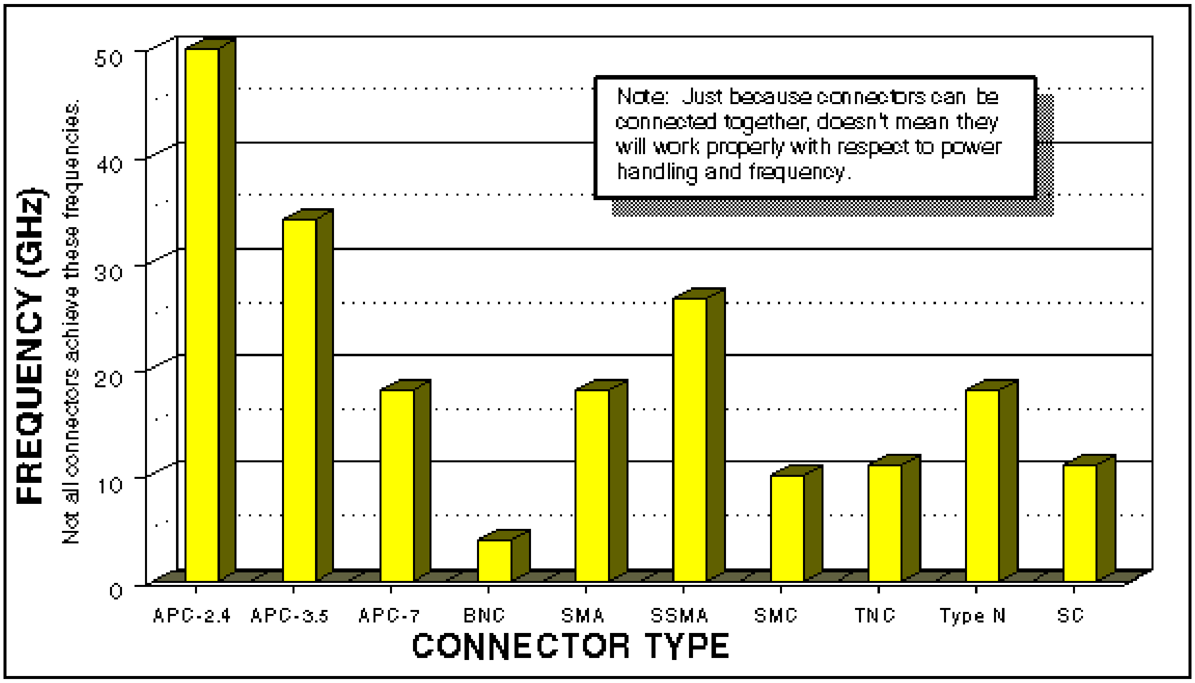

Note: Just because connectors can be connected together, doesn't mean they will work properly with respect to power handling and frequency.

Figure 1. Frequency Range of Microwave Connectors.

POWER DIVIDERS AND DIRECTIONAL COUPLERS

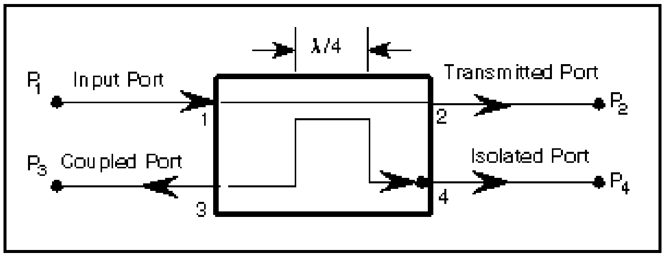

Figure 1. Directional Coupler.

A directional coupler is a passive device which couples part of the transmission power by a known amount out through another port, often by using two transmission lines set close enough together such that energy passing through one is coupled to the other. As shown in Figure 1, the device has four ports: input, transmitted, coupled, and isolated. The term "main line" refers to the section between ports 1 and 2. On some directional couplers, the main line is designed for high power operation (large connectors), while the coupled port may use a small SMA connector. Often the isolated port is terminated with an internal or external matched load (typically 50 ohms). It should be pointed out that since the directional coupler is a linear device, the notations on Figure 1 are arbitrary. Any port can be the input, (as in Figure 3) which will result in the directly connected port being the transmitted port, adjacent port being the coupled port, and the diagonal port being the isolated port.

Physical considerations such as internal load on the isolated port will limit port operation. The coupled output from the directional coupler can be used to obtain the information (i.e., frequency and power level) on the signal without interrupting the main power flow in the system (except for a power reduction - see Figure 2). When the power coupled out to port three is half the input power (i.e., 3 dB below the input power level), the power on the main transmission line is also 3 dB below the input power and equals the coupled power. Such a coupler is referred to as a 90 degree hybrid, hybrid, or 3 dB coupler. The frequency range for coaxial couplers specified by manufacturers is that of the coupling arm. The main arm response is much wider (i.e., if the spec is 2-4 GHz, the main arm could operate at 1 or 5 GHz - see Figure 3). However it should be recognized that the coupled response is periodic with frequency. For example, a λ/4 coupled line coupler will have responses at nλ/4 where n is an odd integer.

Common properties desired for all directional couplers are wide operational bandwidth, high directivity, and a good impedance match at all ports when the other ports are terminated in matched loads. These performance characteristics of hybrid or non-hybrid directional couplers are self-explanatory. Some other general characteristics will be discussed below.

COUPLING FACTOR

The coupling factor is defined as: $\mathit{Coupling\ factor\ (dB)} = -10 \log \frac{P_3}{P_1}$

where P_1 is the input power at port 1 and P3 is the output power from the coupled port (see Figure 1).

The coupling factor represents the primary property of a directional coupler. Coupling is not constant, but varies with frequency. While different designs may reduce the variance, a perfectly flat coupler theoretically cannot be built. Directional couplers are specified in terms of the coupling accuracy at the frequency band center. For example, a 10 dB coupling ± 0.5 dB means that the directional coupler can have 9.5 dB to 10.5 dB coupling at the frequency band center. The accuracy is due to dimensional tolerances that can be held for the spacing of the two coupled lines. Another coupling specification is frequency sensitivity. A larger frequency sensitivity will allow a larger frequency band of operation. Multiple quarter-wavelength coupling sections are used to obtain wide frequency bandwidth directional couplers. Typically this type of directional coupler is designed to a frequency bandwidth ratio and a maximum coupling ripple within the frequency band. For example a typical 2:1 frequency bandwidth coupler design that produces a 10 dB coupling with a ±0.1 dB ripple would, using the previous accuracy specification, be said to have 9.6 ± 0.1 dB to 10.4 ± 0.1 dB of coupling across the frequency range.

LOSS

In an ideal directional coupler, the main line loss port 1 to port 2 (P_1 - P_2) due to power coupled to the coupled output port is:

$$Insertion\ loss\ (dB) = 10 \log \left[1 - \frac{P_3}{P_1} \right]$$

The actual directional coupler loss will be a combination of coupling loss, dielectric loss, conductor loss, and VSWR loss. Depending on the frequency range, coupling loss becomes less significant above 15 dB coupling where the other losses constitute the majority of the total loss. A graph of the theoretical insertion loss (dB) vs. coupling (dB) for a dissipationless coupler is shown in Figure 2.

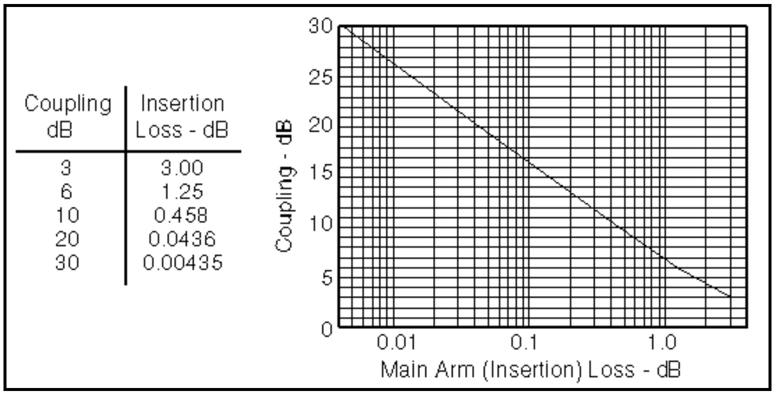

Coupling dB	Insertion Loss - dB
3	3.00
6	1.25
10	0.458
20	0.0436
30	0.00435

Figure 2. Coupling Insertion Loss.

ISOLATION

Isolation of a directional coupler can be defined as the difference in signal levels in dB between the input port and the isolated port when the two output ports are terminated by matched loads, or:

$$Isolation\ (dB) = -10 \log \frac{P_4}{P_1}$$

Isolation can also be defined between the two output ports. In this case, one of the output ports is used as the input; the other is considered the output port while the other two ports (input and isolated) are terminated by matched loads.

Consequently: $Isolation\ (dB) = -10 \log \frac{P_3}{P_2}$

The isolation between the input and the isolated ports may be different from the isolation between the two output ports. For example, the isolation between ports 1 and 4 can be 30 dB while the isolation between ports 2 and 3 can be a different value such as 25 dB. If both isolation measurements are not available, they can assumed to be equal. If neither are available, an estimate of the isolation is the coupling plus return loss (see VSWR section). The isolation should be as high as possible. In actual couplers the isolated port is never completely isolated. Some RF power will always be present. Waveguide directional couplers will have the best isolation.

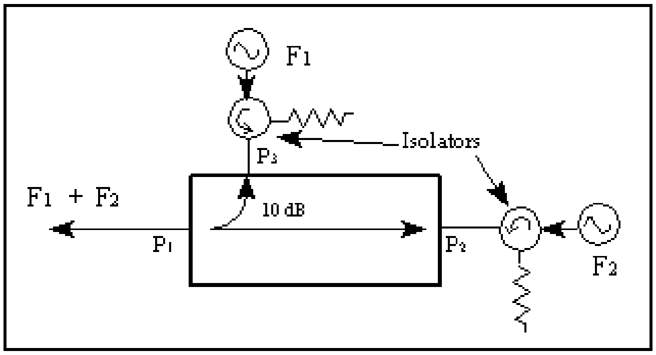

Figure 3. Two-Tone Receiver Tests.

If isolation is high, directional couplers are excellent for combining signals to feed a single line to a receiver for two-tone receiver tests. In Figure 3, one signal enters port P_3 and one enters port P_2, while both exit port P_1. The signal from port P_3 to port P_1 will experience 10 dB of loss, and the signal from port P_2 to port P_1 will have 0.5 dB loss. The internal load on the isolated port will dissipate the signal losses from port P_3 and port P_2. If the isolators in Figure 3 are neglected, the isolation measurement (port P_2 to port P_3) determines the amount of power from the signal generator F_2 that will be injected into the signal generator F_1. As the injection level increases, it may cause modulation of signal generator F_1, or even injection phase locking. Because of the symmetry of the directional coupler, the reverse injection will happen with the same possible modulation problems of signal generator F_2 by F_1. Therefore the isolators are used in Figure 3 to effectively increase the isolation (or directivity) of the directional coupler. Consequently the injection loss will be the isolation of the directional coupler plus the reverse isolation of the isolator.

DIRECTIVITY

Directivity is directly related to Isolation. It is defined as:

$$Directivity\ (dB) = -10 \log \frac{P_4}{P_3} = -10 \log \frac{P_4}{P_1} + 10 \log \frac{P_3}{P_1}$$

where: P_3 is the output power from the coupled port and P_4 is the power output from the isolated port.

The directivity should be as high as possible. Waveguide directional couplers will have the best directivity. Directivity is not directly measurable, and is calculated from the isolation and coupling measurements as:

Directivity (dB) = Isolation (dB) - Coupling (dB)

HYBRIDS

The hybrid coupler, or 3 dB directional coupler, in which the two outputs are of equal amplitude takes many forms. Not too long ago the quadrature (90 degree) 3 dB coupler with outputs 90 degrees out of phase was what came to mind when a hybrid coupler was mentioned. Now any matched 4-port with isolated arms and equal power division is called a hybrid or hybrid coupler. Today the characterizing feature is the phase difference of the outputs. If 90 degrees, it is a 90 degree hybrid. If 180 degrees, it is a 180 degree hybrid. Even the Wilkinson power divider which has 0 degrees phase difference is actually a hybrid although the fourth arm is normally imbedded.

Applications of the hybrid include monopulse comparators, mixers, power combiners, dividers, modulators, and phased array radar antenna systems.

AMPLITUDE BALANCE

This terminology defines the power difference in dB between the two output ports of a 3 dB hybrid. In an ideal hybrid circuit, the difference should be 0 dB. However, in a practical device the amplitude balance is frequency dependent and departs from the ideal 0 dB difference.

PHASE BALANCE

The phase difference between the two output ports of a hybrid coupler should be 0, 90, or 180 degrees depending on the type used. However, like amplitude balance, the phase difference is sensitive to the input frequency and typically will vary a few degrees.

The phase properties of a 90 degree hybrid coupler can be used to great advantage in microwave circuits. For example in a balanced microwave amplifier the two input stages are fed through a hybrid coupler. The FET device normally has a very poor match and reflects much of the incident energy. However, since the devices are essentially identical the reflection coefficients from each device are equal. The reflected voltage from the FETs are in phase at the isolated port and are 180° different at the input port. Therefore, all of the reflected power from the FETs goes to the load at the isolated port and no power goes to the input port. This results in a good input match (low VSWR).

If phase matched lines are used for an antenna input to a 180° hybrid coupler as shown in Figure 4, a null will occur directly between the antennas. If you want to receive a signal in that position, you would have to either change the hybrid type or line length. If you want to reject a signal from a given direction, or create the difference pattern for a monopulse radar, this is a good approach.

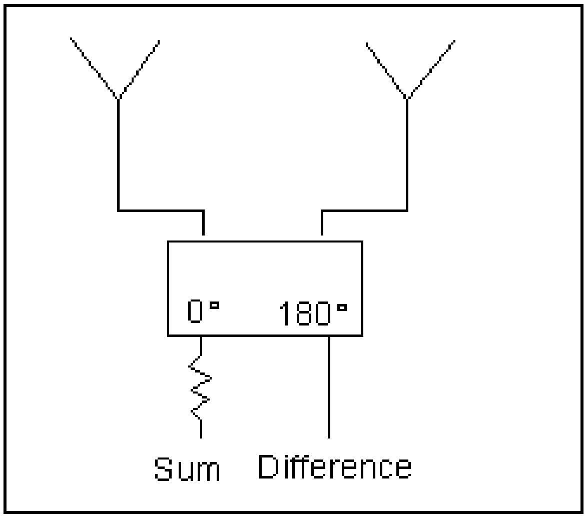

Figure 4. Balanced Antenna Input.

OTHER POWER DIVIDERS

Both in-phase (Wilkinson) and quadrature (90°) hybrid couplers may be used for coherent power divider applications. The Wilkinson's power divider has low VSWR at all ports and high isolation between output ports. The input and output impedances at each port is designed to be equal to the characteristic impedance of the microwave system. A typical power divider is shown in Figure 5. Ideally, input power would be divided equally between the output ports. Dividers are made up of multiple couplers, and like couplers, may be reversed and used as multiplexers. The drawback is that for a four channel multiplexer, the output consists of only 1/4 the power from each, and is relatively inefficient. Lossless multiplexing can only be done with filter networks.

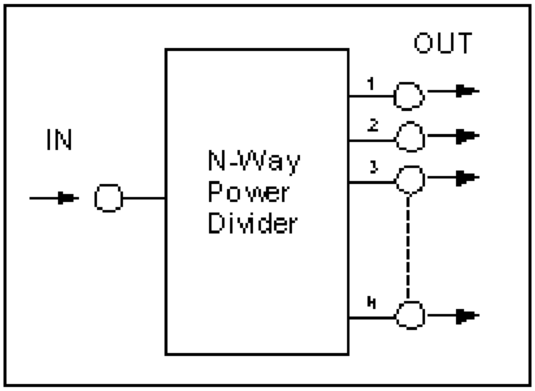

Figure 5. Power Divider.

Coherent power division was first accomplished by means of simple Tee junctions. At microwave frequencies, waveguide tees have two possible forms - the H-Plane or the E-Plane. These two junctions split power equally, but because of the different field configurations at the junction, the electric fields at the output arms are in-phase for the H-Plane tee and are anti-phase for the E-Plane tee. The combination of these two tees to form a hybrid tee allowed the realization of a four-port component which could perform the vector sum (Σ) and difference (Δ) of two coherent microwave signals. This device is known as the magic tee.

POWER COMBINERS

Since hybrid circuits are bi-directional, they can be used to split up a signal to feed multiple low power amplifiers, then recombine to feed a single antenna with high power as shown in Figure 6. This approach allows the use of numerous less expensive and lower power amplifiers in the circuitry instead of a single high power TWT. Yet another approach is to have each solid state amplifier (SSA) feed an antenna and let the power be combined in space or be used to feed a lens which is attached to an antenna. (See Section 3-4.)

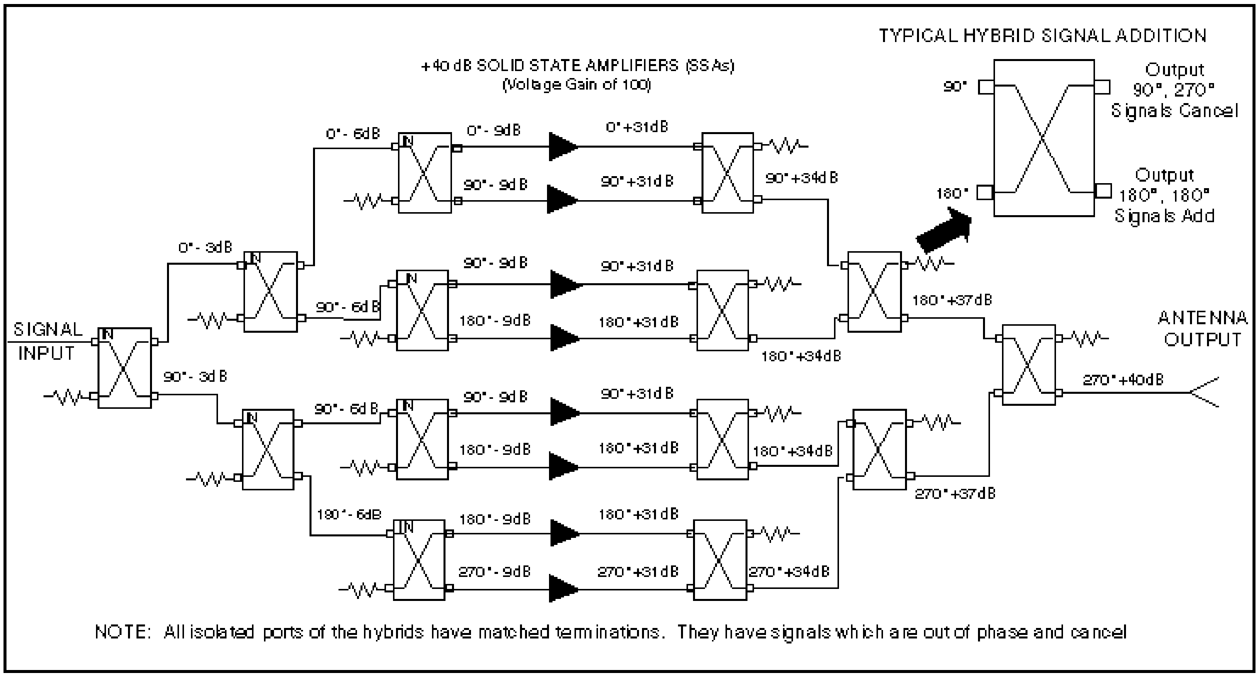
TYPICAL HYBRID SIGNAL ADDITION
+40 dB SOLID STATE AMPLIFIERS (SSAs)
(Voltage Gain of 100)
90°
180°
Output 90°, 270° Signals Cancel
Output 180°, 180° Signals Add
0°- 6dB
0°- 9dB
0°+31dB
90°- 9dB
90°+31dB
90°+34dB
0°- 3dB
90°- 9dB
90°+31dB
90°- 6dB
180°- 9dB
180°+31dB
180°+34dB
180°+37dB
SIGNAL INPUT
IN
90°- 3dB
90°- 6dB
90°- 9dB
90°+31dB
180°- 9dB
180°+31dB
180°+34dB
ANTENNA OUTPUT
270°+40dB
270°+37dB
180°- 6dB
180°- 9dB
180°+31dB
270°- 9dB
270°+31dB
270°+34dB
NOTE: All isolated ports of the hybrids have matched terminations. They have signals which are out of phase and cancel

Figure 6. Combiner Network.

Sample Problem:

If two 1 watt peak unmodulated RF carrier signals at 10 GHz are received, how much peak power could one measure?

A. 0 watts

B. 0.5 watts

C. 1 watt

D. 2 watts

E. All of these

The answer is all of these as shown in Figure 7.

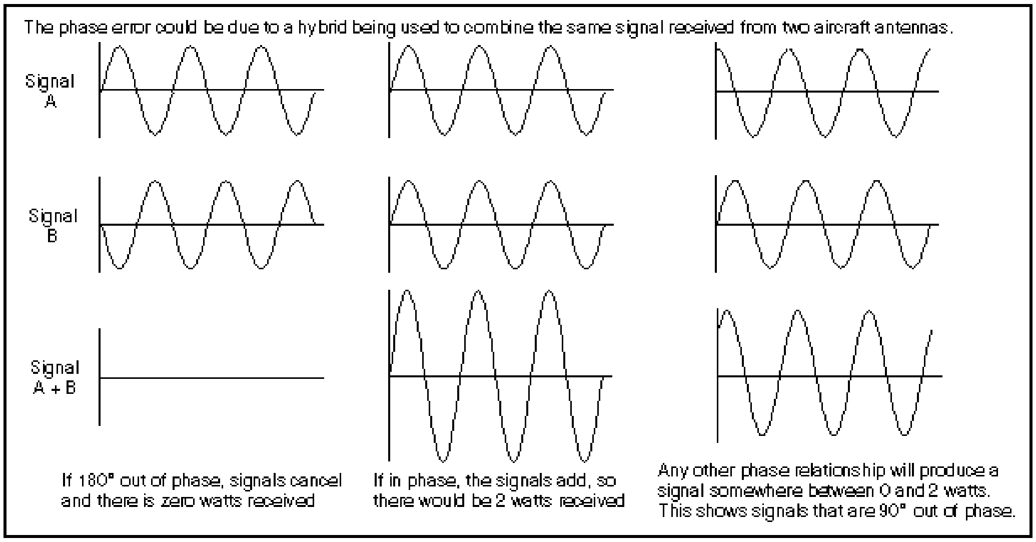
The phase error could be due to a hybrid being used to combine the same signal received from two aircraft antennas.
Signal A
Signal B
Signal A + B
If 180° out of phase, signals cancel and there is zero watts received
If in phase, the signals add, so there would be 2 watts received
Any other phase relationship will produce a signal somewhere between 0 and 2 watts. This shows signals that are 90° out of phase.

Figure 7. Sine Waves Combined Using Various Phase Relationships.

ATTENUATORS / FILTERS / DC BLOCKS

ATTENUATORS

An attenuator is a passive microwave component which, when inserted in the signal path of a system, reduces the signal by a specified amount. They normally possess a low VSWR which makes them ideal for reducing load VSWR in order to reduce measurement uncertainties. They are sometimes used simply to absorb power, either to reduce it to a measurable level, or in the case of receivers to establish an exact level to prevent overload of following stages.

Attenuators are classified as either fixed or variable and either reflective or non-reflective. The fixed and variable attenuators are available in both waveguide and coaxial systems. Most of the receivers under 20 GHz use coaxial type attenuators.

FIXED

The performance characteristics of a fixed attenuator are:

1. input and output impedances
2. flatness with frequency
3. average and peak power handling capability
4. temperature dependence

VARIABLE

The variable attenuator can be subdivided into two kinds: step attenuator and continuously variable attenuator. In a step attenuator, the attenuation is changed in steps such as 10 dB, 1 dB or 0.5 dB. In a continuously variable attenuator, the attenuation is changed continuously and a dial is usually available to read the attenuation either directly or indirectly from a calibration chart.

For a variable attenuator, additional characteristics should be considered, such as:

1. amount or range of attenuations
2. insertion loss in the minimum attenuation position
3. incremental attenuation for step attenuator
4. accuracy of attenuation versus attenuator setting
5. attenuator switching speed and switching noise.

REFLECTIVE

A reflective attenuator reflects some portion of the input power back to the driving source. The amount reflected is a function of the attenuation level. When PIN diodes are zero or reverse biased, they appear as open circuits when shunting a transmission line. This permits most of the RF input power to travel to the RF output. When they are forward biased, they absorb some input, but simultaneously reflect some back to the input port. At high bias current, most RF will be reflected back to the input resulting in a high input VSWR and high attenuation.

ABSORPTIVE

The VSWR of a non-reflective (absorptive) PIN diode attenuator remains good at any attenuation level (bias state). This is accomplished by configuring the diodes in the form of a Pi network that remains matched for any bias state or by use of a 90° hybrid coupler to cancel the waves reflected to the input connector.

MICROWAVE FILTERS

INTRODUCTION

Microwave filters are one of the most important components in receivers. The main functions of the filters are: (1) to reject undesirable signals outside the filter pass band and (2) to separate or combine signals according to their frequency. A good example for the latter application is the channelized receiver in which banks of filters are used to separate input signals. Sometimes filters are also used for impedance matching. Filters are almost always used before and after a mixer to reduce spurious signals due to image frequencies, local oscillator feedthrough, and out-of-frequency band noise and signals. There are many books which are devoted to filter designs. There are many kinds of filters used in microwave receivers, so it is impossible to cover all of them.

If a filter is needed on the output of a jammer, it is desirable to place it approximately half way between the jammer and antenna vs. adjacent to either. The transmission line attenuation improves the VSWR of the filter at the transmitter. This may allow use of a less expensive filter, or use of a reflective filter vs. an absorptive filter.

A filter is a two-port network which will pass and reject signals according to their frequencies. There are four kinds of filters according to their frequency selectivities. In the examples that follow, f_L = low frequency, f_M = medium frequency, and f_H = high frequency. Their names reflect their characteristics, and they are:

1. A low-pass filter which passes the low frequency signals below a predetermined value as shown in Figure 1.

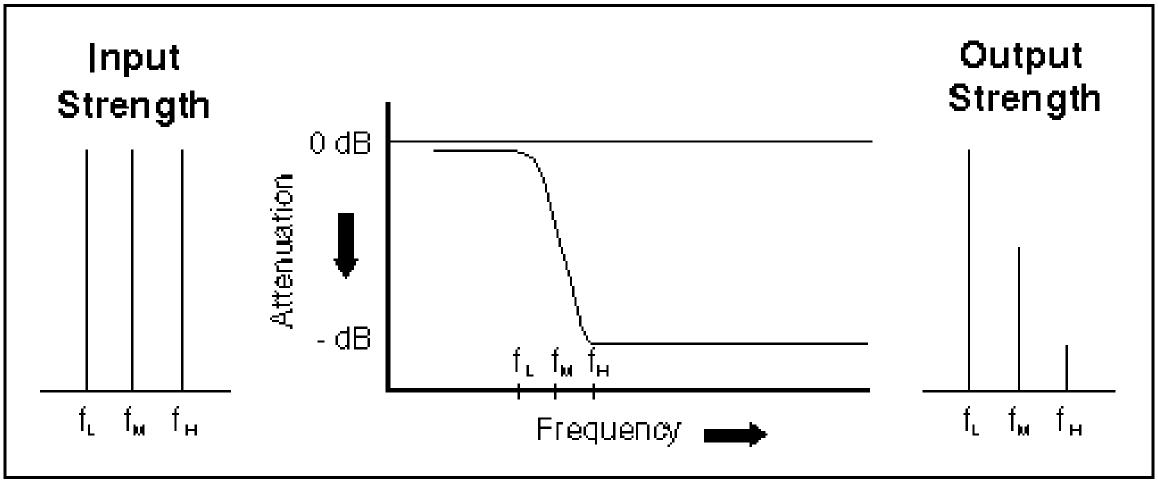

Figure 1. Low-Pass Filter.

2. A high-pass filter which passes the high frequency signals above a predetermined value as in Figure 2.

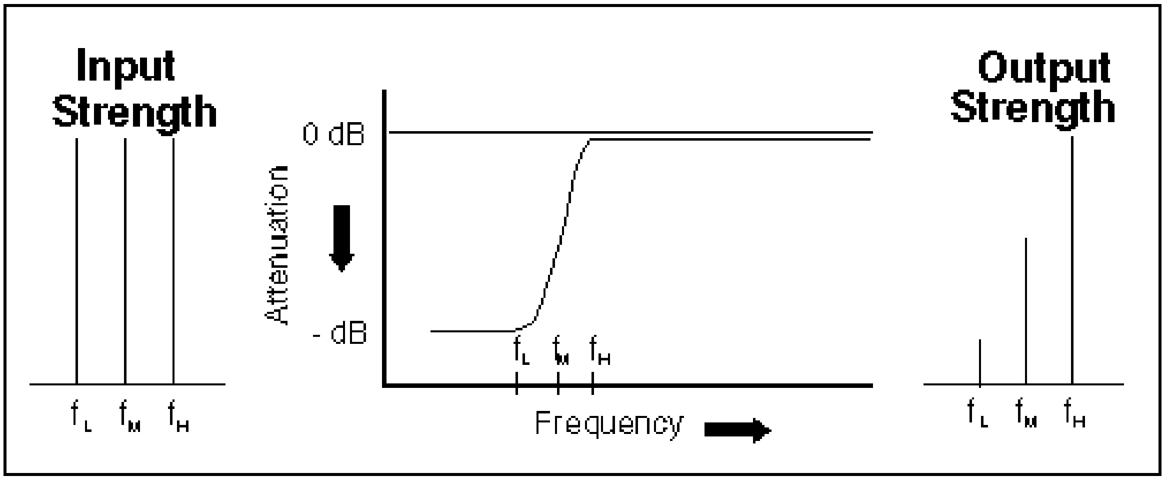
Input
Strength
f_L f_M f_H
0 dB
Attenuation
- dB
f_L f_M f_H
Frequency
Output
Strength
f_L f_M f_H

Figure 2. High-Pass Filter.

3. A band-pass filter which passes signals between two predetermined frequencies as shown in Figure 3.

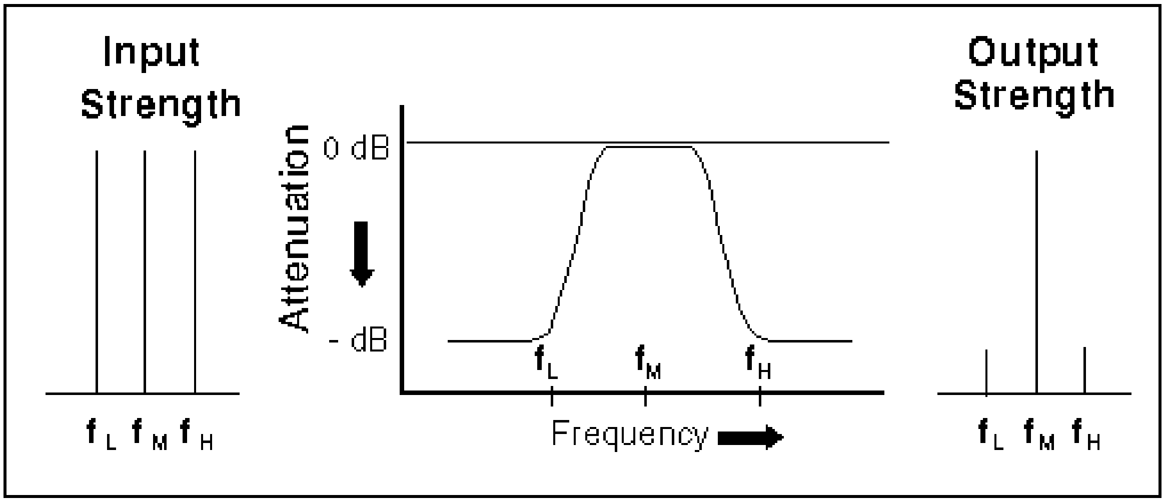
Input
Strength
f_L f_M f_H
0 dB
Attenuation
- dB
f_L f_M f_H
Frequency
Output
Strength
f_L f_M f_H

Figure 3. Band-Pass Filter.

A band-pass filter with different skirt slopes on the two sides of the pass band is sometimes referred to as an asymmetrical filter. In this filter the sharpness of the rejection band attenuation is significantly different above and below the center frequency. One additional note regarding band-pass filters or filters in general, their performance should always be checked in the out-of-band regions to determine whether or not they possess spurious responses. In particular they should be checked at harmonics of the operating frequency.

4. A band reject filter (sometimes referred to as a bandstop or notch filter) which rejects signals between two predetermined frequencies such as high power signals from the aircraft's own radar as shown in Figure 4.

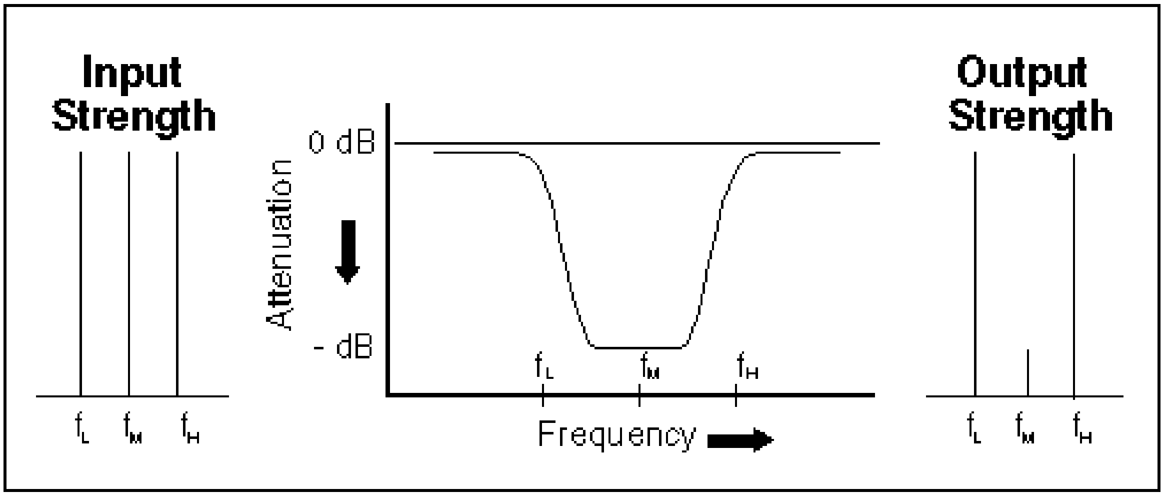

Figure 4. Band-Reject Filter.

In general, filters at microwave frequencies are composed of resonate transmission lines or waveguide cavities that, when combined, reflect the signal power outside the filter frequency pass band and provide a good VSWR and low loss within the frequency pass band. As such, specifications for filters are maximum frequency, pass band loss, VSWR, and rejection level at a frequency outside of the pass band. The trade-offs for filters are a higher rejection for a fixed frequency pass band or a larger frequency pass band for a fixed rejection, which requires a filter with more resonators, which produce higher loss, more complexity, and larger size.

DC BLOCKS

DC Blocks are special connectors which have a capacitor (high pass filter) built into the device. There are three basic types:

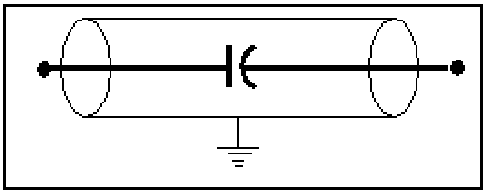

Figure 5. Inside DC Block.

1. INSIDE - The high pass filter is in series with the center conductor as shown in Figure 5. DC is blocked on the center conductor.
2. OUTSIDE - The high pass filter is in series with the cable shield as shown in Figure 6.
3. INSIDE/OUTSIDE - A high pass arrangement is connected to both the inner and outer conductors.

DC Blocks are ideal for filtering DC, 60 Hz, and 400 Hz from the RF line.

In general, capacitors with a large value of capacitance do not have the least loss at microwave frequencies. Also, since capacitance is proportional to size, a large size produces more capacitance with more inductance. Because of these reasons, D.C. blocks are typically available with a high pass frequency band starting in the region of 0.1 to 1 GHz.

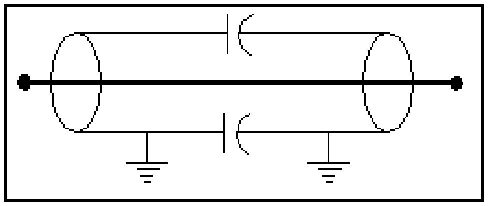

Figure 6. Outside DC Block.

TERMINATIONS / DUMMY LOADS

A termination is a one-port device with an impedance that matches the characteristic impedance of a given transmission line. It is attached to a certain terminal or port of a device to absorb the power transmitted to that terminal or to establish a reference impedance at that terminal. Important parameters of a termination are its VSWR and power handling capacity. In a receiver, terminations are usually placed at various unconnected ports of components such as hybrid and power dividers to keep the VSWR of the signal path low. It is extremely important that the isolated port in a directional coupler and the unused port of a power divider (i.e., only three ports of a four-way power divider are used) be properly terminated. All of the design considerations of directional couplers and power dividers are based on the fact that all ports are terminated with matched loads. If an unused port is not properly terminated, then the isolation between the output ports will be reduced which may severely degrade the performance of the receiver.

A termination is the terminology used to refer to a low power, single terminal device intended to terminate a transmission line. Similar devices designed to accommodate high power are generally termed dummy loads.

TERMINATIONS

Terminations are employed to terminate unconnected ports on devices when measurements are being performed. They are useful as dummy antennas and as terminal loads for impedance measurements of transmission devices such as filters and attenuators.

The resistive elements in most terminations are especially fabricated for use at microwave frequencies. Two types are commonly employed: (1) resistive film elements, and (2) molded resistive tapers. The resistive film is very thin compared to the skin depth and normally very short relative to wavelength at the highest operating frequency. The molded taper consists of a dissipative material evenly dispersed in a properly cured dielectric medium. Both forms of resistive elements provide compact, rugged terminations suitable for the most severe environmental conditions with laboratory stability and accuracy.

Terminations should be properly matched to the characteristic impedance of a transmission line. The termination characteristics of primary concern are:

a. operating frequency range
b. average power handling capability
c. operating temperature range
d. VSWR
e. size
f. weight

Many microwave systems employ directional couplers which require terminations on at least one port, and most have various modes of operation or test where terminations are needed on certain terminals.

A matched termination of a generalized transmission line is ideally represented by an infinite length of that line having small, but non-zero loss per unit length so that all incident energy is absorbed and none is reflected.

Standard mismatches are useful as standards of reflection in calibrating reflectometer setups and other impedance measuring equipment. They are also used during testing to simulate specific mismatches which would be encountered on the terminals of components once the component is installed in the actual system.

The following table shows common mismatches with the impedance that can provide the mismatch.

Common Mismatches (Z_O = 50 Ω)		
Ratio	Z_L (higher)	Z_L (lower)
1.0 : 1	50 Ω (matched)	50 Ω (matched)
1.25 : 1	62.5 Ω	40 Ω
1.50 : 1	75 Ω	33.3 Ω
2.00 : 1	100 Ω	25 Ω

DUMMY LOADS

A dummy load is a high power one port device intended to terminate a transmission line. They are primarily employed to test high power microwave systems at full power capacity. Low power coaxial loads are generally termed terminations and typically handle one watt or less.

Most radars or communications systems have a dummy load integrated into them to provide a non-radiating or EMCON mode of operation, or for testing (maintenance).

Three types of dissipative material are frequently employed in dummy loads: (1) lossy plastic, (2) refractory, and (3) water.

The lossy plastic consists of particles of lossy material suspended in plastic medium. This material may be designed to provide various attenuations per unit length but is limited as to operating temperature. It is employed primarily for low power applications.

The refractory material is a rugged substance that may be operated at temperatures up to 1600°F. It is virtually incapable of being machined by ordinary means but is often fabricated through diamond wheel grinding processes. Otherwise material must be fired in finished form. Such material is employed in most high power applications.

The dissipative properties of water are also employed for dummy load applications. Energy from the guide is coupled through a leaky wall to the water which flows alongside the main guide. Water loads are employed for extremely high power and calorimetric applications.

While dummy loads can operate over full waveguide bands, generally a more economical unit can be manufactured for use over narrower frequency ranges.

The power rating of a dummy load is a complex function dependent upon many parameters, including average and peak power, guide pressure, external temperature, guide size, air flow, and availability of auxiliary coolant. The average and peak powers are interrelated in that the peak power capacity is a function of the operating temperature which in turn is a function of the average power.

CIRCULATORS AND DIPLEXERS

A microwave circulator is a nonreciprocal ferrite device which contains three or more ports. The input from port n will come out at port n + 1 but not out at any other port. A three-port ferrite junction circulator, usually called the Y-junction circulator, is most commonly used. They are available in either rectangular waveguide or strip- line forms. The signal flow in the three-port circulator is assumed as 1→2, 2→3, and 3→1 as shown in Figure 1.

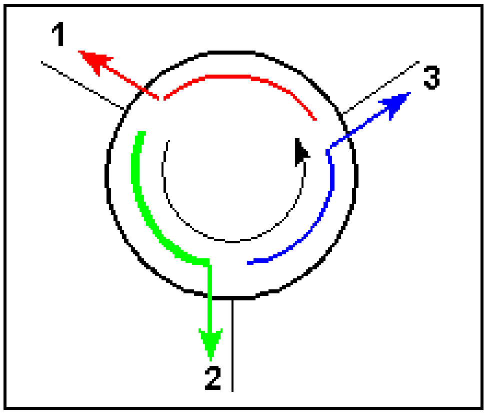

Figure 1. Symbolic Expression for a Y-Junction Circulator.

If port 1 is the input, then the signal will come out of port 2; in an ideal situation, no signal should come out of port 3 which is called the isolated port. The insertion loss of the circulator is the loss from 1 to 2, while the loss from 1 to 3 is referred to as isolation. A typical circulator will have a few tenths of a dB insertion loss from port 1 to 2 and 20 dB of isolation from port 1 to 3 for coaxial circulators (30 dB or more for waveguide circulators). When the input is port 2, the signal will come out of port 3 and port 1 is the isolated port. Similar discussions can be applied to port 3.

> Since circulators contain magnets, they should not be mounted near ferrous metals since the close proximity of metals like iron can change the frequency response.

As shown in Figure 2, if one port of a circulator is loaded, it becomes an isolator, i.e., power will pass from ports one to two, but power reflected back from port two will go to the load at port three versus going back to port one.

As shown in Figure 3 this circulator is made into a diplexer by adding a high pass filter to port two. Frequencies from port one that are below 10 GHz will be reflected by port two. Frequencies above 10 GHz will pass through port two. At the 10 GHz crossover frequency of the diplexer, a 10 GHz signal will be passed to both ports two and three but will be half power at each port. Diplexers or triplexers (one input and three output bands), must be specifically designed for the application.

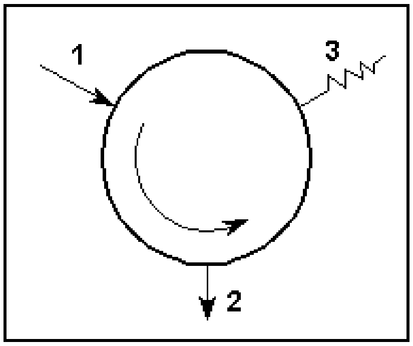

Figure 2. Isolator From A Circulator.

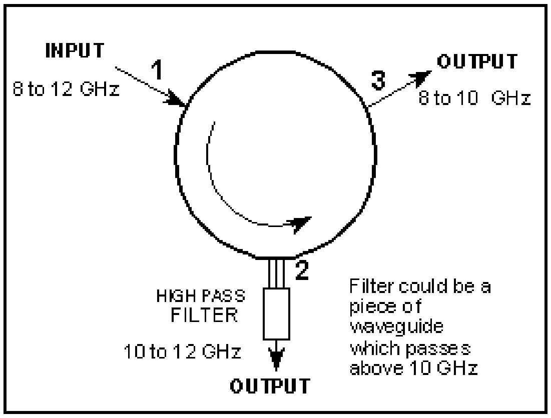

Figure 3. Diplexer From A Circulator.

Another useful device is the 4-port Faraday Rotator Circulator shown symbolically in Figure 4. These waveguide devices handle very high power and provide excellent isolation properties. It is useful when measurements must be made during high power application as shown. A water load is used to absorb the high power reflections so that a reasonable power level is reflected to the receiver or measurement port.

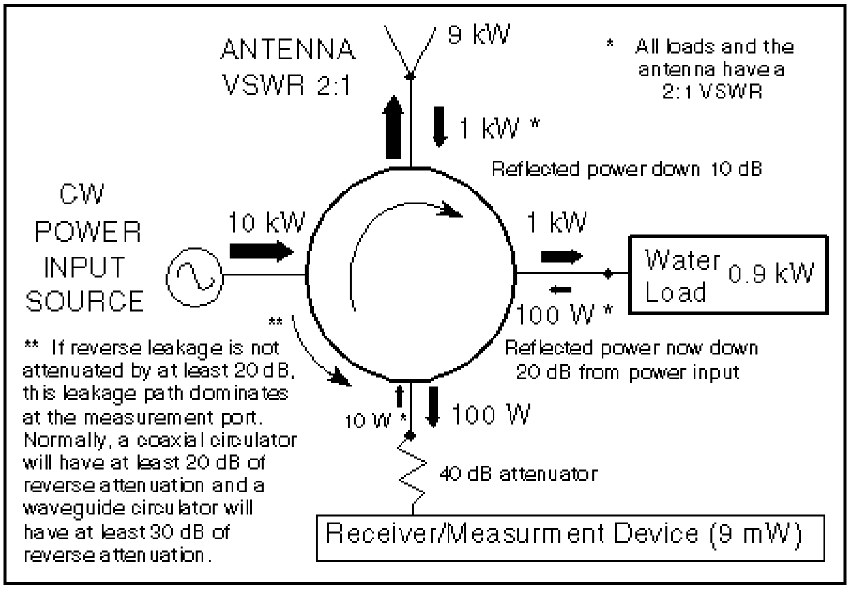
ANTENNA
VSWR 2:1
9 kW
1 kW *
Reflected power down 10 dB
CW
POWER
INPUT
SOURCE
10 kW
1 kW
Water
Load
0.9 kW
100 W *
**
10 W *
100 W
40 dB attenuator
Receiver/Measurment Device (9 mW)

* All loads and the antenna have a 2:1 VSWR

Reflected power now down 20 dB from power input

** If reverse leakage is not attenuated by at least 20 dB, this leakage path dominates at the measurement port. Normally, a coaxial circulator will have at least 20 dB of reverse attenuation and a waveguide circulator will have at least 30 dB of reverse attenuation.

Figure 4. Faraday Rotator Circulator.

The Maximum Input Power to a Measurement Device - The ideal input to a measurement device is in the 0 to 10 dBm (1 to 10 mW) range. Check manufacturer's specification for specific maximum value.

If the RF transmission lines and their components (antenna, hybrid, etc.) can support the wider frequency range, circulators could be used to increase the number of interconnecting RF ports from two as shown in Figure 5, to four as shown in Figure 6. Figure 7 shows an alternate configuration using diplexers which could actually be made from circulators as shown previously in Figure 3.

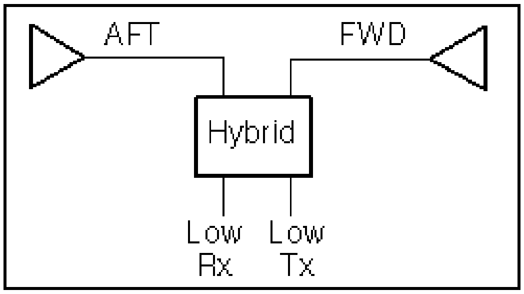
AFT
FWD
Hybrid
Low
Rx
Low
Tx

Figure 5. Low Band Configuration.

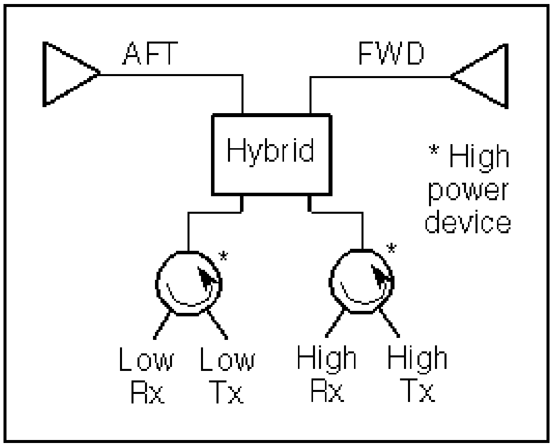
AFT
FWD
Hybrid
* High power device
*
*
Low
Rx
Low
Tx
High
Rx
High
Tx

Figure 6. Low/High Band Configuration.

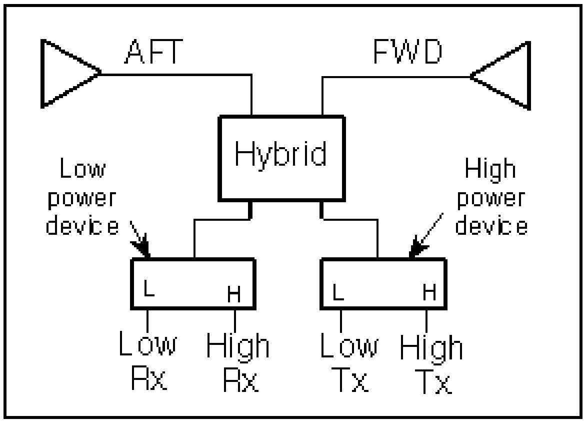
AFT
FWD
Hybrid
Low power device
High power device
L
H
L
H
Low
Rx
High
Rx
Low
Tx
High
Tx

Figure 7. Alternate Low/High Band Configuration.

MIXERS AND FREQUENCY DISCRIMINATORS

Mixers are used to convert a signal from one frequency to another. This is done by combining the original RF signal with a local oscillator (LO) signal in a non-linear device such as a Schottky-barrier diode.

The output spectrum includes:

- The original inputs, LO and RF
- All higher order harmonics of LO and RF
- The two primary sidebands, LO ± RF (m,n = 1)
- All higher order products of mLO ± nRF (where m,n are integers)
- A DC output level

The desired output frequency, commonly called the intermediate frequency (IF), can be either the lower (LO-RF) or upper (LO+RF) sideband. When a mixer is used as a down converter, the lower sideband is the sideband of interest.

A microwave balanced mixer makes use of the 3 dB hybrid to divide and recombine the RF and LO inputs to two mixing diodes. The 3 dB hybrid can be either the 90° or 180° type. Each has certain advantages which will be covered later. The critical requirement is that the LO and RF signals be distributed uniformly (balanced) to each mixer diode.

Figure 1 is a typical balanced mixer block diagram. The mixer diodes are reversed relative to each other; the desired frequency (IF) components of each diode are then in-phase while the DC outputs are positive and negative respectively.

The two diode outputs are summed in a tee where the DC terms cancel and only the desired IF component exists at the IF port.

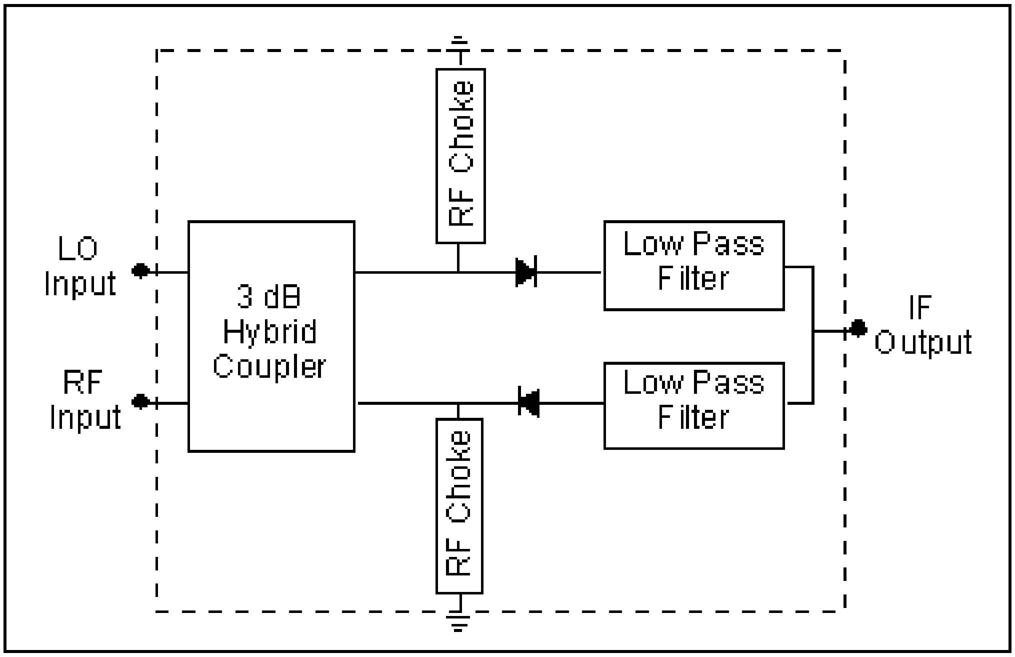

Figure 1. Mixer Block Diagram.

Other types of mixers exist, including the double-balanced mixer, and the Ortho-Quad® (quadrature fed dual) mixer. The relative advantages and disadvantages of each of the four types are summarized in Table 1.

Table 1. Mixer Comparison.

Mixer Type	VSWR [1]	Conversion Loss [2]	LO/RF Isolation [3]	Harmonic Suppression [4]	Dynamic Range	IF Bandwidth
90° Hybrid	good	lowest	poor	poor-fair	high	wide
180° Hybrid	poor	low	good	good	high	wide
Double-Balanced	poor	low	Very good -excellent	very good	high	extremely wide
Ortho Quad	good	low	very good	fair	high	wide

NOTES:

(1) Poor = 2.5:1 typical; Good = 1.3:1 typical
(2) Conversion loss: lowest: 5-7 dB typical; Low 7-9 dB typical
(3) Poor: 10 dB typical; Good: 20 dB typical; Very Good: 25-30 dB typical; Excellent: 35-40 dB typical
(4) Poor: partial rejection of LO/RF even harmonics
Fair: slightly better
Good: can reject all LO even harmonics
Very Good: can reject all LO and RF even harmonics

Used in various circuits, mixers can act as modulators, phase detectors, and frequency discriminators.

The phase discriminators can serve as a signal processing network for systems designed to monitor bearing, polarization, and frequency of AM or FM radiated signals.

A frequency discriminator uses a phase discriminator and adds a power divider and delay line at the RF input as shown in Figure 2. The unknown RF signal "A" is divided between a reference and delay path. The differential delay (T) creates a phase difference (θ) between the two signals which is a linear function of frequency (f) and is given by $\theta = 2\pi fT$.

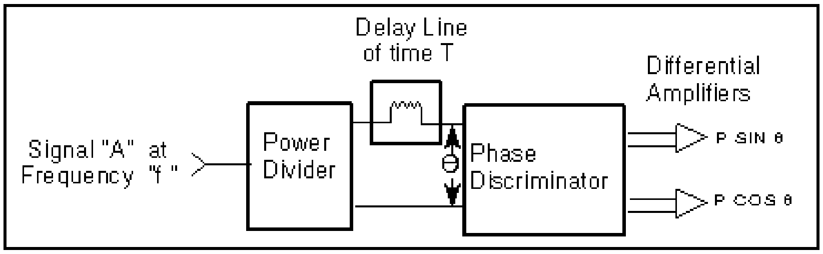

Figure 2. Frequency Discriminator.

When the two output signals are fed to the horizontal and vertical input of an oscilloscope, the resultant display angle will be a direct function of frequency.

DETECTORS

A detector is used in receiver circuits to recognize the presence of signals. Typically a diode or similar device is used as a detector. Since this type of detector is unable to distinguish frequency, they may be preceded by a narrow band-pass filter.

A typical simplistic circuit is shown in Figure 1.

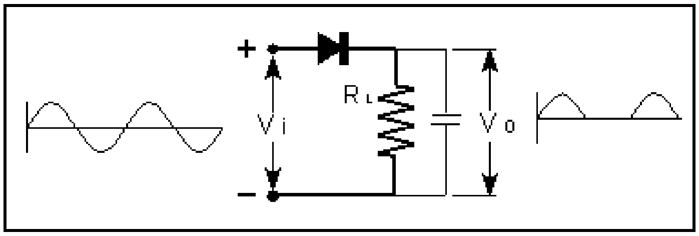

Figure 1. Typical Diode Detector Circuit.

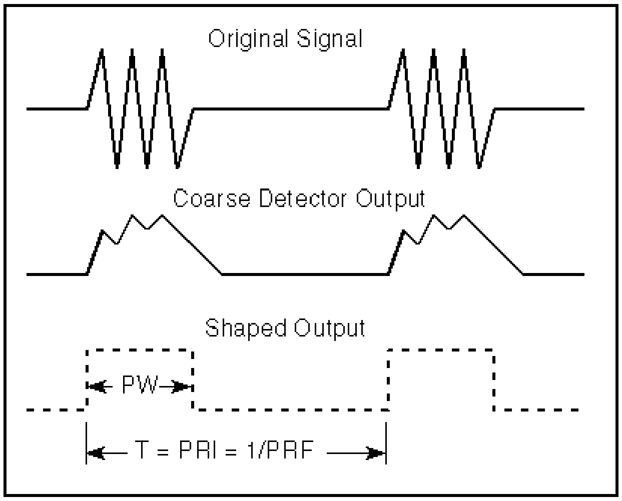

Figure 2. Demodulated Envelope Output.

To integrate a pulse radar signal, we can add capacitance to the circuit in parallel with the output load R_L to store energy and decrease the bleed rate. Figure 2 shows a typical input/output waveform which detects the envelope of the pulse radar signal. From this information pulse width and PRF characteristics can be determined for the RWR UDF comparison.

When the diode is reverse biased, very little current passes through unless the reverse breakdown voltage is exceeded. When forward biased and after exceeding the cut-in voltage, the diode begins to conduct as shown in Figure 3. At low voltages, it first operates in a square law region. Detectors operating in this region are known as small signal type. If the voltage is higher, the detector operates in a linear region, and is known as the large signal type.

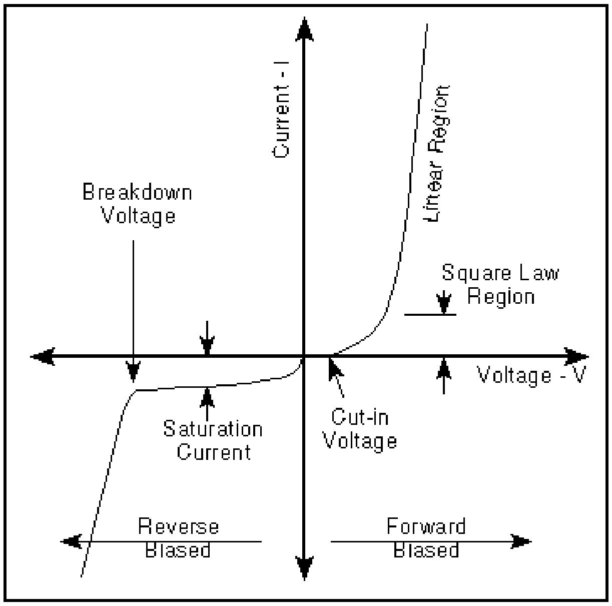

Figure 3. Diode Electrical Characteristics.

The power/voltage characteristics for a typical diode detector is shown in Figure 4.

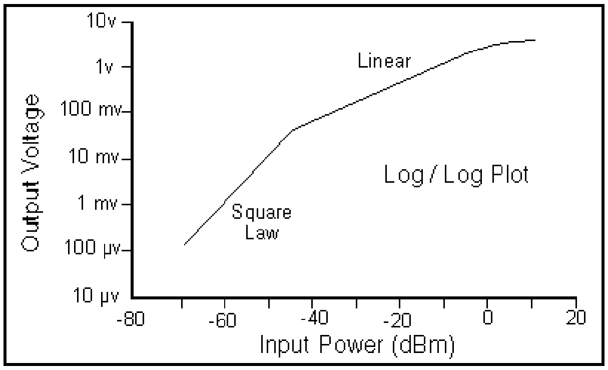

Figure 4. Diode Power/Voltage Characteristic.

Square Law Detector

In the square law region, the output voltage V_o is proportional to the square of the input voltage V_i, thus V_o is proportional to the input power.

$V_o = nV_i^2 = nP_i$ or $P_i \propto V_o$
Where n is the constant of proportionality

Linear Detector

In the linear detection region, the output voltage is given by:

$V_o = mV_i$ and since $P=V^2/R$, $P_i \propto V_o^2$
Where m is the constant of proportionality

Log Detector Amplifier

Another type of detector arrangement is the Log detector amplifier circuit shown in Figure 5. It is formed by using a series of amplifiers and diode detectors. Due to the nature of the amplifier/diode characteristics, the output voltage is related to the power by:

$P_i \propto 10^{pV_o + q}$
Where p and q are constants of proportionality

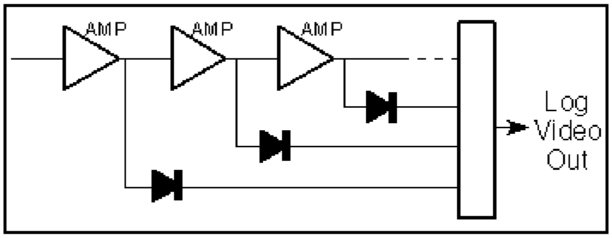

Figure 5. Log Detector.

The Log detector has good range, but is hampered by large size when compared to a single diode detector.

Pulse Width Measurements

If the pulse width of a signal was specified at the one-half power point, the measurements of the detected signal on an oscilloscope would vary according to the region of diode operation. If the region of operation is unknown, a 3 dB attenuator should be inserted in the measurement line. This will cause the power to decrease by one-half. The (temporary) peak amplitude on the oscilloscope becomes the amplitude reference point for measuring the pulse width when the external 3 dB attenuator is removed.

These voltage levels for half power using the three types of detectors are shown in Table 1.

Table 1. Detector Characteristics.

	Square Law	Linear	Log
Output Voltage When Input Power is reduced by Half (3 dB)	0.5 V_{in}	0.707 V_{in}	A very small value. ~ 0.15 V_{in} for typical 5 stage log amplifier
Sensitivity & Dynamic Range	Good sensitivity Small dynamic range	Less sensitivity Greater dynamic range	Poorest sensitivity Greatest dynamic range (to 80 dB)

Also see the Microwave Measurements section subsection entitled “Half Power or 3 dB Measurement Point.”

This page intentionally left blank.

RF / MICROWAVE AMPLIFIERS

An amplifier is one of the most essential and ubiquitous elements in modern RF and microwave systems. Fundamentally, an amplifier is a type of electronic circuit used to convert low-power signals into ones of significant power. The specific requirements for amplification are as varied as the systems where they are used. Amplifiers are realized using a wide range of different technologies, and are available in many form factors. While the performance of an amplifier can be measured by wide range of attributes, several important ones include: gain, power added efficiency (PAE), input and output return loss, 1-dB compression point (P_{1dB}), stability, linearity, intermodulation distortion, and noise figure.

Traveling Wave Tubes (TWTs)

A traveling-wave tube (TWT) is a specialized vacuum tube used in electronics to amplify radio frequency (RF) signals to very high power. A TWT is a component of an electronic assembly known as a traveling-wave tube amplifier (TWTA), often pronounced "Tweet-uh." While trades between these parameters exist, modern TWTAs are capable of providing very high gains (>50 dB), multiple octaves of bandwidth, high efficiency, and output powers that range from tens to thousands of watts. Pulsed TWTAs can reach even higher output powers. Figure 1 presents a simplified diagram of a helix-type TWT.

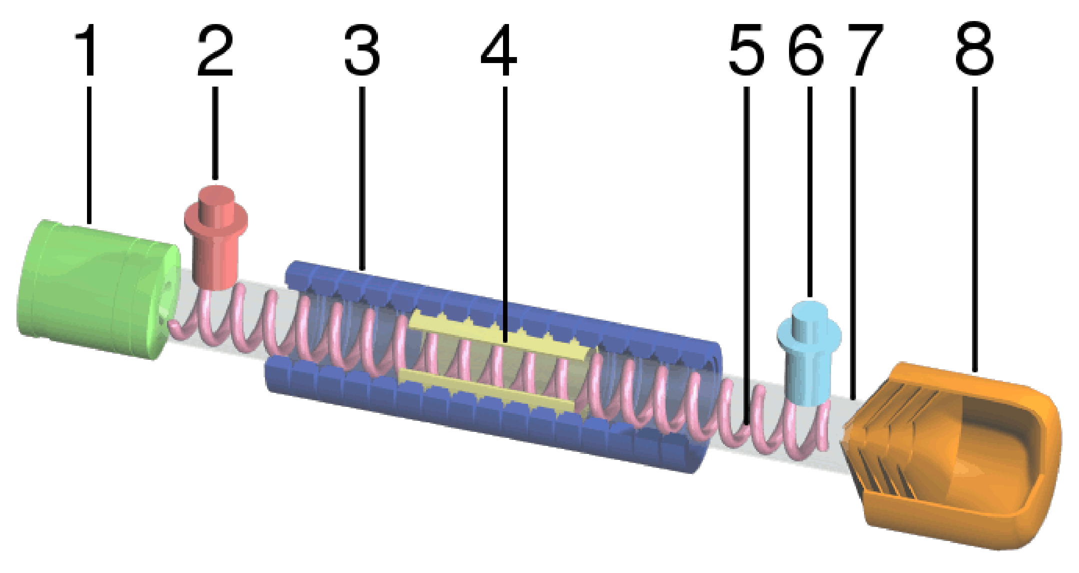

Figure 1. Cutaway view of a helix TWT.
(1) Electron gun; (2) RF input; (3) Magnets; (4) Attenuator;
(5) Helix coil; (6) RF output; (7) Vacuum tube; (8) Collector.
Reproduced from Wikipedia.

The device is an elongated vacuum tube with an electron gun (a heated cathode that emits electrons) at one end. A solenoid coil wrapped around the tube creates a magnetic field which focuses the electrons into a beam, which then passes down the middle of a wire helix that stretches the length of the tube, finally striking a collector at the other end. An input signal is coupled into the helix near the emitter, is amplified by the electron beam as it travels down the length of the tube, and the amplified signal is coupled to an external port near the collector.

The helix, which acts as a RF transmission line, delays the signal to near the same propagation speed as the electron beam. The speed at which the electromagnetic wave travels down the tube can be varied by changing the number or diameter of the turns in the helix. While propagating along the tube, the EM wave interacts with the electron beam. Since the electromagnetic wave effectively propagates slower than the electron beam, the electrons "bunch up" and modulate the input signal, giving up energy in the process - an effect known as velocity modulation. Thus, the traveling wave progressively grows in amplitude as it propagates down the tube towards the collector.

Figure 2 features a photo of a high voltage power supply and TWT.

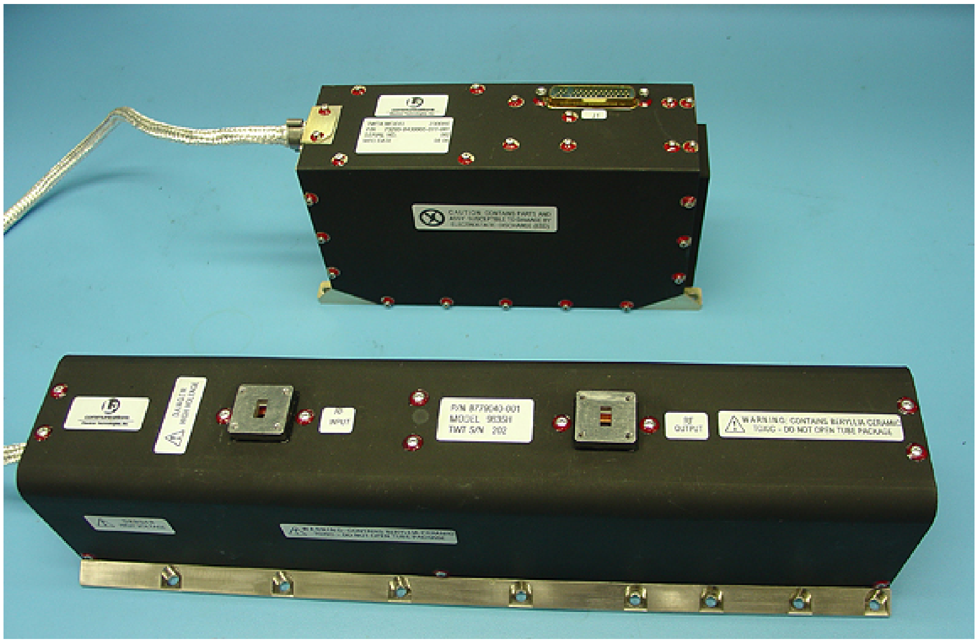

Figure 2. TWTA Including High Voltage Power Supply (top) and TWT (bottom).

Historically, TWTAs have been used in satellite transponders, radars, and in electronic warfare and self-protection systems. Recently, with the advent of wideband, high power solid-state amplifier solutions, however, TWTAs are slowly being replaced due to the higher reliability of their solid state counterparts.

Microwave Power Modules (MPMs)

A Microwave Power Module (MPM) is a hybrid solution between solid-state and vacuum tube electronics, which aims to take advantage of the best features of both technologies. They feature a solid-state pre-amplifier, miniaturized TWTA, and a high-density power supply, all integrated into a unit much more compact and lightweight than the traditional TWTA. While MPMs generally don't provide as much power as their larger TWTA counterparts, their lighter weight, compact form factor, and relaxed power supply requirements (often 28 VDC or 270 VDC) enable use in applications where a TWTA would not be possible.

Similar to TWTAs, MPMs are capable of providing very high gains (>50 dB), multiple octaves of bandwidth, high efficiency. Typical power levels range from tens to hundreds of watts, with ~1 kW capability for pulsed MPMs.

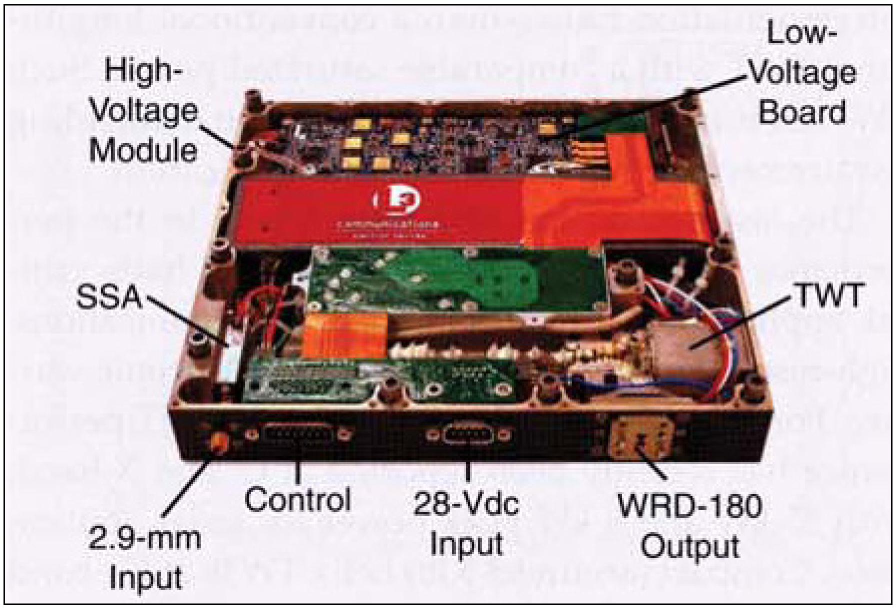
High-
Voltage
Module
Low-
Voltage
Board
SSA
TWT
2.9-mm
Input
Control
28-Vdc
Input
WRD-180
Output

Figure 3. 40W Ka-band MPM With Components Identified.
Image from *IEEE Magazine*, December 2009, page 42.

They have found applications in phased array antennas, lower-power radar transmitters, satellite communications, EW systems, and UAVs.

Solid State

Solid-state electronics, including amplifiers, are built entirely from solid materials and in which the electrons, or other charge carriers, are confined entirely within the solid material. The building material is most often a crystalline semiconductor. Solid-state power amplifiers, or SSPAs, are fabricated on many different semiconductor technologies, some of which include GaAs, GaN, InP, SiGe, CMOS. A photo of a GaN SSPA is presented in Figure 4.

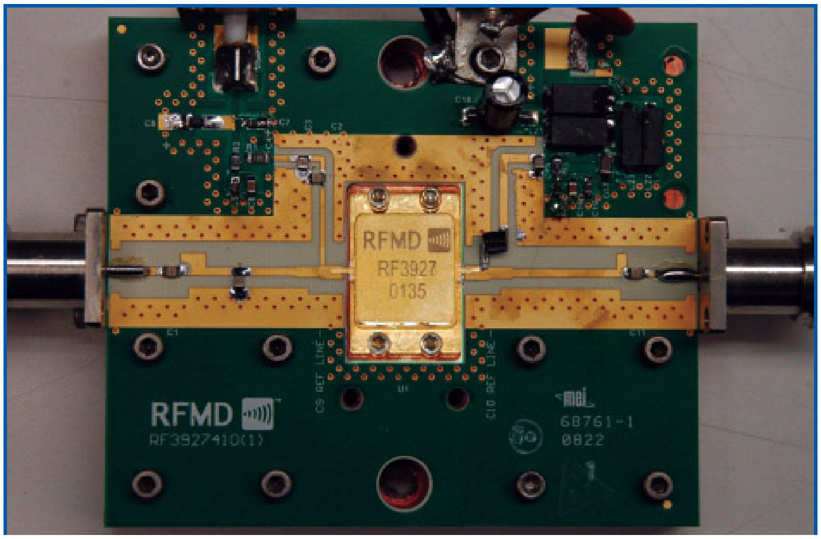
RFMD
RF3927
0135
RFMD
RF3927410(1)
68761-1
0822

Figure 4. Typical SSPA. Note the GaN die and external impedance matching networks.

In general, SSPAs are heralded as having a higher reliability than TWTAs. While high power SSPAs are available on the market, achieving high efficiency has been a challenge as many amplifier stages often need to be combined to meet to achieve power levels in the tens to hundreds of watts. Multi-stage designs increase size, weight, and power (SWAP) and decrease PAE (due to ohmic losses in the combining networks), which is one reason why TWTs and MPMs still provide excellent alternatives for medium to high power applications.

Below is a table that attempts to compare and contrast different solid-state technologies in the context of microwave networks and amplifiers. It is by no means complete, but provides a general overview.

Table 1. Comparison Between Different Semiconductor Process Technologies.

Technology	**Advantages**	**Disadvantages**
Silicon variants (CMOS, SiGe)	• Cheapest substrate due to CPU industry • Can be fabricated with nanometer accuracy	• Not a good overall microwave substrate • Results in lossy, high noise figure, low power components • Crystal is fragile • Junction temperatures limited to ~110C
Indium Phosphide Variants (InP/InGaP)	• High performance at low voltage • Good thermal stability • Results in high efficiency PAs, particularly at lower operating voltage • Extremely low noise figure • Useful through W-band and beyond	• Less mature, “niche” fabrication houses • Brittle, fragile material • Higher cost than GaAs • Low breakdown voltage – not good for high power
Gallium Arsenide variants (GaAs)	• Most mature, widespread technology • Many transistor variants (MESFET, PHEMT, MHEMT, HBT, etc) • High reliability • Fairly low cost (but more than silicon) • Great microwave substrate • Low loss, high ε_r. • 16-20V breakdown possible • Junction temperatures up to 150C	• High noise figure, Noise figure and power performance • Difficult to summarize - depends on transistor type used
Gallium Nitride (GaN)	• Up to 10x the power density of GaAs • High breakdown voltage (100V possible) • Results in high efficiency, high frequency, wide bandwidth PAs • Can operate hotter than GaAs, Si, or SiGe	• Currently more expensive than GaAs, but costs are decreasing • Difficult to fabricate • High power density leads to thermal challenges

SIGNAL GENERATION

Signal generators, also known variously as function generators, RF and microwave signal generators, pitch generators, arbitrary waveform generators, digital pattern generators or frequency generators are electronic devices that generate repeating or non-repeating electronic signals (in either the analog or digital domains). They are generally used in designing, testing, troubleshooting, and repairing electronic or electro-acoustic devices; though they often have artistic uses as well.

There are many different types of signal generators, with different purposes and applications (and at varying levels of expense); in general, no device is suitable for all possible applications.

Analog Signal Generators

RF signal generators are capable of producing CW (continuous wave) tones. The output frequency can usually be tuned anywhere in their frequency range. Many models offer various types of analog modulation, either as standard equipment or as an optional capability to the base unit. This could include AM, FM, ΦM (phase modulation) and pulse modulation. Another common feature is a built-in attenuator which makes it possible to vary the signal's output power. Depending on the manufacturer and model, output powers can range from -135 to +30 dBm. A wide range of output power is desirable, since different applications require different amounts of signal power. For example, if a signal has to travel through a very long cable out to an antenna, a high output signal may be needed to overcome the losses through the cable and still have sufficient power at the antenna. But when testing receiver sensitivity, a low signal level is required to see how the receiver behaves under low signal-to-noise conditions.

RF signal generators are required for servicing and setting up analog radio receivers, and are used for professional RF applications.

Arbitrary Waveform Generator

An arbitrary waveform generator (AWG) is a piece of electronic test equipment used to generate electrical waveforms. These waveforms can be either repetitive or single-shot (once only) in which case some kind of triggering source is required (internal or external). The resulting waveforms can be injected into a device under test and analyzed as they progress through the device, confirming the proper operation of the device or pinpointing a fault in the device.

Unlike function generators, AWGs can generate any arbitrarily defined wave shape as their output. The waveform is usually defined as a series of "waypoints" (specific voltage targets occurring at specific times along the waveform) and the AWG can either jump to those levels or use any of several methods to interpolate between those levels.

Because AWGs synthesize the waveforms using (baseband) digital signal processing techniques, their maximum frequency is usually limited to no more than a few gigahertz (~10 GHz being the latest state-of-the-art in 2012).

A major difficulty in generating non-repetitive waveforms at higher frequencies with AWGs is the large amount of data required to describe high-frequency baseband signals. For example, a 20 gigasample/sec arbitrary waveform with 8 bits of resolution requires 20 GB of data to represent every 1 second of signal, regardless of that signal's nature. Generating such large digital data streams and delivering them to the DAC in the AWG is an increasingly difficult problem as DAC upper frequencies continue to grow.

AWGs, like most signal generators, may also contain an attenuator, various means of modulating the output waveform, and often contain the ability to automatically and repetitively "sweep" the frequency of the output waveform (by means of a voltage-controlled oscillator) between two operator-determined limits. This capability makes it very easy to evaluate the frequency response of a given electronic circuit.

AWGs can operate as conventional function generators. These would include standard waveforms such as sine, square, ramp, triangle, noise, and pulse. Some units include additional built-in waveforms such as exponential rise and fall times, sinx/x, cardiac. Some AWGs allow users to retrieve waveforms from a number of digital and mixed-signal oscilloscopes. Some AWGs may display a graph of the waveform on their screen - a graph mode. Some AWGs have the ability to generate a pattern of words from multiple bit output connector to simulate data transmission, combining the properties of both AWGs and digital pattern generators.

One feature of direct digital synthesizer (DDS) based arbitrary waveform generators is that their digital nature allows multiple channels to be operated with precisely controlled phase offsets or ratio-related frequencies. This allows the generation of polyphase sine waves, I-Q constellations, or simulation of signals from geared mechanical systems such as jet engines. Complex channel-channel modulations are also possible.

Vector Signal Generators

Modern vector signal generators can be seen as a hybrid of the arbitrary waveform generator and the analog signal generator, combining a lower-bandwidth AWG with analog upconversion, allowing a moderate-bandwidth (~100 MHz, circa 2012) digital signal to be upconverted to any frequency (~50 GHz typically upper frequency typically).

With the advent of digital communications systems, it is no longer possible to adequately test these systems with traditional analog signal generators. This has led to the development of vector signal generators, also known as digital signal generators. These signal generators are capable of generating digitally-modulated radio signals that may use any of a large number of digital modulation formats such as QAM, QPSK, FSK, BPSK, and OFDM. In addition, since modern commercial digital communication systems are almost all based on well-defined industry standards, many vector signal generators can generate signals based on these standards. Examples include GSM, W-CDMA (UMTS), CDMA2000, LTE, Wi-Fi (IEEE 802.11), and WiMAX (IEEE 802.16). In contrast, military communication systems such as JTRS, which place a great deal of importance on robustness and information security, typically use very proprietary methods. To test these types of communication systems, users will often create their own custom waveforms and download them into the vector signal generator to create the desired test signal.

This page intentionally left blank.

DIGITAL SIGNAL PROCESSING COMPONENTS

The goal of DSP is usually to measure, filter and/or compress continuous real-world analog signals. The first step is usually to convert the signal from an analog to a digital form, by sampling and then digitizing it using an analog-to-digital converter (ADC), which turns the analog signal into a stream of numbers. However, often, the required output signal is another analog output signal, which requires a digital-to-analog converter (DAC). Even if this process is more complex than analog processing and has a discrete value range, the application of computational power to digital signal processing allows for many advantages over analog processing in many applications, such as error detection and correction in transmission as well as data compression. With the increasing bandwidth and dynamic range of ADC and digital components, common analog RF signal processing operations such as filtering, threshold detection, and pulse compression are being carried out in the digital domain. Additionally new capabilities such as synthetic aperture radar (SAR), digital RF memory (DRFM), and space-time adaptive processing (STAP) are wholly enabled by the increasing power of digital processing components.

Analog-to-Digital Converter

An analog-to-digital converter (abbreviated ADC, A/D or A to D) is a device that converts a continuous quantity to a discrete time digital representation. An ADC may also provide an isolated measurement. The reverse operation is performed by a digital-to-analog converter (DAC).

Typically, an ADC is an electronic device that converts an input analog voltage or current to a digital number proportional to the magnitude of the voltage or current. However, some non-electronic or only partially electronic devices, such as rotary encoders, can also be considered ADCs.

Resolution

The resolution of the converter indicates the number of discrete values it can produce over the range of analog values. The values are usually stored electronically in binary form, so the resolution is usually expressed in bits. In consequence, the number of discrete values available, or "levels," is a power of two. For example, an ADC with a resolution of 8 bits can encode an analog input to one in 256 different levels, since $2^8 = 256$. The values can represent the ranges from 0 to 255 (i.e. unsigned integer) or from −128 to 127 (i.e., signed integer), depending on the application.

Resolution can also be defined electrically, and expressed in volts. The minimum change in voltage required to guarantee a change in the output code level is called the least significant bit (LSB) voltage. The resolution Q of the ADC is equal to the LSB voltage. The voltage resolution of an ADC is equal to its overall voltage measurement range divided by the number of discrete values:

$$Q = \frac{E_{FSR}}{2^M}$$

where M is the ADC's resolution in bits and E_{FSR} is the full scale voltage range (also called 'span').

E_{FSR} is given by

$$E_{FSR} = V_{RefHi} - V_{RefLow}$$

where V_{RefHi} and V_{RefLow} are the upper and lower extremes, respectively, of the voltages that can be coded.

Normally, the number of voltage intervals is given by

$$N = 2^M - 1$$

where M is the ADC's resolution in bits.

That is, one voltage interval is assigned in between two consecutive code levels.

A typical 3 bit (2^8 = 8-level) coding scheme is depicted in Figure 1.

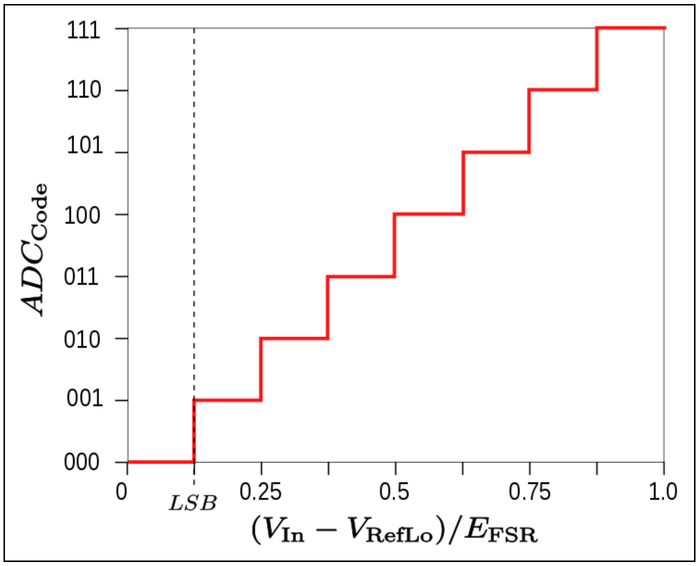

Figure 1. An 8-level (3 bit) ADC Coding Scheme.

Example:

Assume input signal x(t) = A cos(t), A = 5V
Full scale measurement range = -5 to 5 volts
ADC resolution is 8 bits: 2^8 - 1 = 256 - 1 = 255 quantization levels (codes)
ADC voltage resolution, Q = (5 V − [-5] V) / 255 = 10 V / 255 ≈ 0.039 V ≈ 39 mV.

In practice, the useful resolution of a converter is limited by the best signal-to-noise ratio (SNR) that can be achieved for a digitized signal. An ADC can resolve a signal to only a certain number of bits of resolution, called the effective number of bits (ENOB). One effective bit of resolution changes the signal-to-noise ratio of the digitized signal by 6 dB, if the resolution is limited by the ADC. If a preamplifier has been used prior to A/D conversion, the noise introduced by the amplifier can be an important contributing factor towards the overall SNR.

Accuracy

An ADC has several sources of errors. Quantization error and (assuming the ADC is intended to be linear) non-linearity are intrinsic to any analog-to-digital conversion. There is also a so-called aperture error which is due to a clock jitter and is revealed when digitizing a time-variant signal (not a constant value).

These errors are measured in a unit called the least significant bit (LSB). In the above example of an eight-bit ADC, an error of one LSB is 1/256 of the full signal range, or about 0.4%.

Sampling Rate

The analog signal is continuous in time and it is necessary to convert this to a flow of digital values. It is therefore required to define the rate at which new digital values are sampled from the analog signal. The rate of new values is called the sampling rate or sampling frequency of the converter.

A continuously varying band limited signal can be sampled (that is, the signal values at intervals of time T, the sampling time, are measured and stored) and then the original signal can be exactly reproduced from the discrete-time values by an interpolation formula. The accuracy is limited by quantization error. However, this faithful reproduction is only possible if the sampling rate is higher than twice the highest frequency of the signal. This is essentially what is embodied in the Shannon-Nyquist sampling theorem.

Oversampling

Usually, for economy, signals are sampled at the minimum rate required, with the result that the quantization noise introduced is white noise spread over the whole pass band of the converter. If a signal is sampled at a rate much higher than the Nyquist frequency and then digitally filtered to limit it to the signal bandwidth there are the following advantages:

- Digital filters can have better properties (sharper roll-off, phase) than analog filters, so a sharper anti-aliasing filter can be realized and then the signal can be down-sampled giving a better result
- A 20-bit ADC can be made to act as a 24-bit ADC with 256× oversampling
- The signal-to-noise ratio due to quantization noise will be higher than if the whole available band had been used. With this technique, it is possible to obtain an effective resolution larger than that provided by the converter alone.
- The improvement in SNR is 3 dB (equivalent to 0.5 bits) per octave of oversampling which is not sufficient for many applications. Therefore, oversampling is usually coupled with noise shaping. With noise shaping, the improvement is 6L+3 dB per octave where L is the order of loop filter used for noise shaping. e.g. - a 2nd order loop filter will provide an improvement of 15 dB/octave.

Digital to Analog Converter (DAC)

Digital-to-analog converters (DACs) perform the inverse function of the ADC. As such, the aforementioned principles of accuracy, resolution, sampling and oversampling, etc. apply equivalently.

Field-Programmable Gate Array (FPGA)

A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by the customer or designer after manufacturing—hence "field-programmable." The FPGA configuration is generally specified using a hardware description language (HDL), similar to that used for an application-specific integrated circuit (ASIC). Circuit diagrams were previously used to specify the configuration, as they were for ASICs, but this is increasingly rare. FPGAs can be used to implement any logical function that an ASIC could perform. The ability to update the functionality after shipping, partial re-configuration of a portion of the design and the low non-recurring engineering costs relative to an ASIC design (notwithstanding the generally higher unit cost), offer advantages for many applications.

FPGAs contain programmable logic components called "logic blocks," and a hierarchy of reconfigurable interconnects that allow the blocks to be "wired together"—somewhat like many (changeable) logic gates that can be inter-wired in (many) different configurations. Logic blocks can be configured to perform complex combinational functions, or merely simple logic gates like AND and XOR. In most FPGAs, the logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory.

FPGAs are becoming increasingly more popular for implementing RF signal processing functions such as signal compression (matched filtering), channel selection, modulation / demodulation, etc.

MICROWAVE MEASUREMENTS

Measurement Procedures

Calculate your estimated power losses before attempting to perform a measurement. The ideal input to a measurement device is in the 0 to 10 dBm (1 to 10 mW) range.

Linearity Check

To verify that a spectrum measurement is accurate and signals are not due to mixing inside the receiver, a linearity check should be performed, i.e., externally insert a 10 dB attenuator - if measurements are in the linear region of the receiver, all measurements will decrease by 10 dB. If the measurements decrease by less than 10 dB, the receiver is saturated. If the measurements disappear, you are at the noise floor.

Half-Power or 3 dB Measurement Point

To verify the half power point of a pulse width measurement on an oscilloscope, externally insert a 3 dB attenuator in the measurement line, and the level that the peak power decreases to is the 3 dB measurement point (Note: you cannot just divide the peak voltage by one-half on the vertical scale of the oscilloscope).

VSWR Effect on Measurement

Try to measure VSWR (or reflection coefficient) at the antenna terminals. Measuring VSWR of an antenna through its transmission line can result in errors. Transmission lines should be measured for insertion loss not VSWR.

High Power Pulsed Transmitter Measurements

When making power measurements on a high power pulsed transmitter using a typical 40 dB directional coupler, an additional attenuator may be required in the power meter takeoff line, or the power sensor may be burnt out.

For example, assume we have a 1 megawatt transmitter, with PRF = 430 pps, and PW = 13 μs. Further assume we use a 40 dB directional coupler to tap off for the power measurements. The power at the tap would be:

$10 \log(P_p) - 10 \log(DC) - \text{Coupler reduction} =$

$10 \log(10^9 \text{mW}) - 10 \log(13 \times 10^{-6})(430) - 40 \text{ dB} =$

$90 \text{ dBm} - 22.5 \text{ dB} - 40 \text{ dB} = 27.5 \text{ dBm}$ (too high for a power meter)

Adding a 20 dB static attenuator to the power meter input would give us a value of 7.5 dBm or 5.6 mW, a good level for the power meter.

High Power Measurements With Small Devices

When testing in the presence of a high power radar, it is normally necessary to measure the actual field intensity. The technique shown in Figure 4, in Section 6-7, may not be practical if the measurement device must be small. An alternate approach is the use of a rectangular waveguide below its cutoff frequency. In this manner, the "antenna" waveguide provides sufficient attenuation to the frequency being measured so it can be coupled directly to the measurement device or further attenuated by a low power attenuator. The attenuation of the waveguide must be accurately measured since attenuation varies significantly with frequency.

ELECTRO-OPTICS AND IR

ELECTRO-OPTICS AND IR 7-1
    Introduction 7-1.1
    Optical Spectrum 7-1.1
    Radiometric Quantities and Terminology 7-1.2
    Photometric Quantities 7-1.3
    Basic Principles 7-1.4
    Basic Radiant Power Relationships 7-1.5
    Infrared Source Characteristics 7-1.10
    Atmospheric Transmission 7-1.14
    EO Components and Sensors 7-1.16
    IR Threats to Aircraft and Their Countermeasures 7-1.30
    IRCM 7-1.36
    Lasers 7-1.38
    Fiber Optics 7-1.41

Laser Safety 7-2

This page intentionally left blank.

ELECTRO-OPTICAL SYSTEMS AND EW COUNTERMEASURES

INTRODUCTION

The development of infrared countermeasures and the evaluation of their performance against threat missiles is a broad, complex technical field. Much of the detailed information about the threats to be countered and the characteristics of the countermeasures themselves is understandably sensitive, beyond the limitations of this document. More detailed information can be requested, provided the requesting organization has the necessary clearance and need to know. Such requests will be considered on a case-by-case basis. Information will be provided only upon written concurrence of the controlling Government organization.

There are many electro-optical (EO) electronic warfare (EW) systems, which are analogous to radio frequency (RF) EW systems. These EO EW systems operate in the optical portion of the electromagnetic spectrum. Electro-optics (EO), as the name implies, is a combination of electronics and optics. By one definition EO is the science and technology of the generation, modulation, detection and measurement, or display of optical radiation by electrical means. Most infrared (IR) sensors, for example, are EO systems. In the popularly used term "EO/IR," the EO is typically used to mean visible or laser systems. The use of EO in this context is a misnomer. Actually, almost all "EO/IR" systems are EO systems as defined above. Another often-used misnomer is referring to an EO spectrum. EO systems operate in the optical spectrum, which is from 0.01 to 1000 micrometers. EO systems include, but are not limited to, lasers, photometry, infrared, and other types of visible, and UV imaging systems.

Within the broad field of Electronic Warfare, electro-optical systems are prevalent for communication systems and offensive and defensive applications. Lasers have been used extensively for weapons guidance purposes, warhead fuzing applications, targeting systems and other offensive weapons related purposes. Understanding electro-optical and radiometric principles and sensors is critical to the development of vehicle survivability systems. These principles range from signature reduction and camouflage to active countermeasure systems such as lamp-based and laser jammers to passive threat warning systems and expendable flare decoys.

Although military systems operate in many portions of the electro-optical spectrum, the infrared is of paramount importance for remote detection systems and weapons applications. Missile seekers, Forward Looking Infrared (FLIR) systems, and Infrared Search and Track Systems all operate in the infrared portion of the spectrum.

OPTICAL SPECTRUM

The optical spectrum is that portion of the electromagnetic spectrum from the extreme ultraviolet (UV) through the visible to the extreme IR. Figure 1 shows the optical spectrum in detail. The end points of the optical spectrum are somewhat arbitrary.

IR spectrum terminology has also varied through the years, with near (near visible), or short-wavelength infrared (SWIR) being on the high frequency end. Then as frequency decreases the spectrum is followed by intermediate or mid-wavelength infrared (MWIR), then far or long-wavelength IR (LWIR), and finally extreme IR.

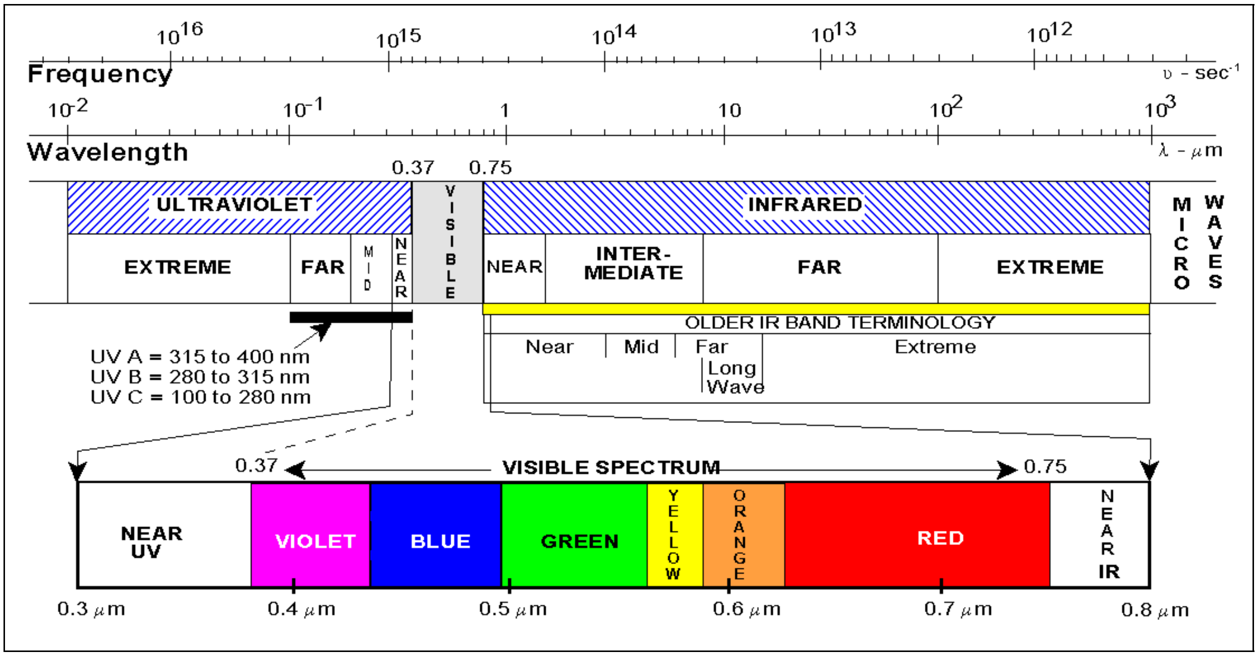

Frequency
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
υ - sec^{-1}
Wavelength
10^{-2}
10^{-1}
1
10
10^{2}
10^{3}
λ - μm
0.37
0.75
ULTRAVIOLET
EXTREME
FAR
MID
NEAR
VISIBLE
INFRARED
NEAR
INTER-MEDIATE
FAR
EXTREME
MICRO WAVES
OLDER IR BAND TERMINOLOGY
Near
Mid
Far
Long Wave
Extreme
UV A = 315 to 400 nm
UV B = 280 to 315 nm
UV C = 100 to 280 nm
0.37
VISIBLE SPECTRUM
0.75
NEAR UV
VIOLET
BLUE
GREEN
YELLOW
ORANGE
RED
NEAR IR
0.3 μm
0.4 μm
0.5 μm
0.6 μm
0.7 μm
0.8 μm

Figure 1. Optical Spectrum.

RADIOMETRIC QUANTITIES AND TERMINOLOGY

The common terms used to describe optical radiation are the source parameters of power, radiant emittance (older term) or radiant exitance (newer term), radiance, and radiant intensity. They refer to how much radiation is given off by a body. The parameter measured by the detector (or collecting object/surface) is the irradiance. Any of these quantities can be expressed per unit wavelength in which case the subscript is changed from e (meaning energy derived units) to λ and the term is then called "Spectral ...X...", i.e. I_e is radiant intensity, while I_λ is spectral radiant intensity. These quantities in terms of currently preferred "Système International d'Unités" (SI units) are defined in Table 1.

Table 1. Radiometric SI Units.

Symbol	Name	Description	Units
Q	Radiant Energy		J (joules)
Φ_e	Radiant Power (or flux)	Rate of transfer of radiant energy	W (watts)
M_e	Radiant Exitance	Radiant power per unit area emitted from a surface	W m^{-2}
L_e	Radiance	Radiant power per unit solid angle per unit projected area	W m^{-2}sr^{-1}
I_e	Radiant Intensity	Radiant power per unit solid angle from a point source	W sr^{-1}
E_e	Irradiance	Radiant power per unit area incident upon a surface	W m^{-2}
X_λ	Spectral ...X..	(Quantity) per unit wavelength interval	(Units) nm^{-1} or μm^{-1}

Where X_λ is generalized for each unit on a per wavelength basis; for example, L_λ would be called "spectral radiance" instead of radiance.

In common usage, irradiance is expressed in units of Watts per square centimeter and wavelengths are in **µ**m instead of nanometers (nm). Other radiometric definitions are shown in Table 2.

Table 2. Other Radiometric Definitions.

Symbol	Name	Description	Units
α	Absorptance[1]	α = (*) absorbed / (*) incident	numeric
ρ	Reflectance	ρ = (*) reflected / (*) incident	numeric
τ	Transmittance	τ = (*) transmitted / (*) incident	numeric
$\in$	Emissivity	$\in$ = (*) of specimen / (*) of blackbody @ same temperature	numeric

Where (*) represents the appropriate quantity Q, Φ, M, E, or L

Note (1) Radiant absorptance should not be confused with absorption coefficient.

PHOTOMETRIC QUANTITIES

Whereas the radiometric quantities Φ_e, M_e, I_e, L_e, and E_e have meaning throughout the entire electromagnetic spectrum, their photometric counterparts Φ_v, M_v, I_v, L_v, and E_v are meaningful only in the visible spectrum (0.38 **µ**m thru 0.78 **µ**m).

The "standard candle" was redefined as the new candle or candela (cd). One candela is the luminous intensity of 1/60th of 1 cm^2 of the projected area of a blackbody radiator operating at the temperature of the solidification of platinum (2045 ºK). By definition, the candela emits one lumen (lm) per steradian.

Table 3 displays the photometric quantities and units. These are used in dealing with optical systems such as aircraft television camera systems, optical trackers, or video recording.

Table 3. Photometric SI Units.

Symbol	Name	Description	Units
Q_v	Luminous energy		lumen sec (lm s)
F or Φ_v	Luminous flux or Luminous Power	Rate of transfer of luminant energy	lumen (lm)
M_v	Luminous Exitance or flux density (formerly luminous emittance)	Luminant power per unit area emitted from a surface	lm m^{-2}
L_v	Luminance (formerly brightness)	Luminous flux per unit solid angle per unit projected area	nit (nt) or candela / m^2 or lm/sr·m^2
I_v	Luminous Intensity (formerly candlepower)	Luminous power per unit solid angle from a point source	candela (cd) or lm/sr
E_v	Illuminance (formerly illumination)	Luminous power per unit area incident upon a surface	lux or lx or lm/m^2
K	Luminous efficacy	$K = \Phi_v / \Phi_e$	lm / w

Table 4 displays conversion factors for commonly used illuminance quantities.

Table 4. Illuminance Conversion Units.

		Lux (lx)	Footcandle (fc)	Phot (ph)
1 lux (lm m^{-2})	=	1	0.0929	1×10^{-4}
1 footcandle (lm ft^{-2})	=	10.764	1	0.001076
1 phot (lm cm^{-2})	=	1×10^{4}	929	1

THE BASIC PRINCIPLES

The processes of absorption, reflection (including scattering), and transmission account for all incident radiation in any particular situation, and the total must add up to one in order that energy be conserved:

$\alpha + \rho + \tau = 1$, as shown in Figure 2.

If a material is opaque (no transmission), then: absorption + reflection = 1

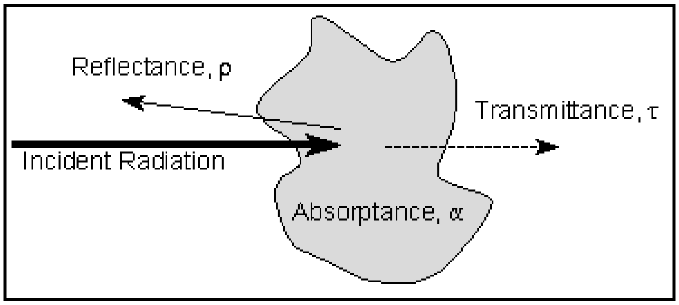

Figure 2. Radiation Incident on a Body.

In addition to the above processes, optical (including IR) radiation interacts with matter the same as radiation in any other part of the spectrum, including:

1. **Diffraction** – around edges
2. **Emission** – from matter by conversion from another form of energy
3. **Interference** – constructive and destructive
4. **Refraction** – bends when passing between two media with different propagation speeds (Snell's Law)
5. **Scattering** – when interacts with particles whose size approaches length of the wave
6. **Polarized** - electric field is partially polarized by reflection from dielectric

STERADIAN – SOLID ANGLE

Of significance to many terms and units in radiometric calculations is the *solid angle*. Figure 3 is a pictorial depicting the relationship of area, distance, and solid angle. By definition, the ratio of area on the surface of a sphere to the square of distance (the radius) is the unit less parameter *solid angle*, or steradian in the SI system of units. Solid angle is usually abbreviated as "sr" or given the Greek letter, Ω. The steradian is a dimensionless quantity in radiometric calculations.

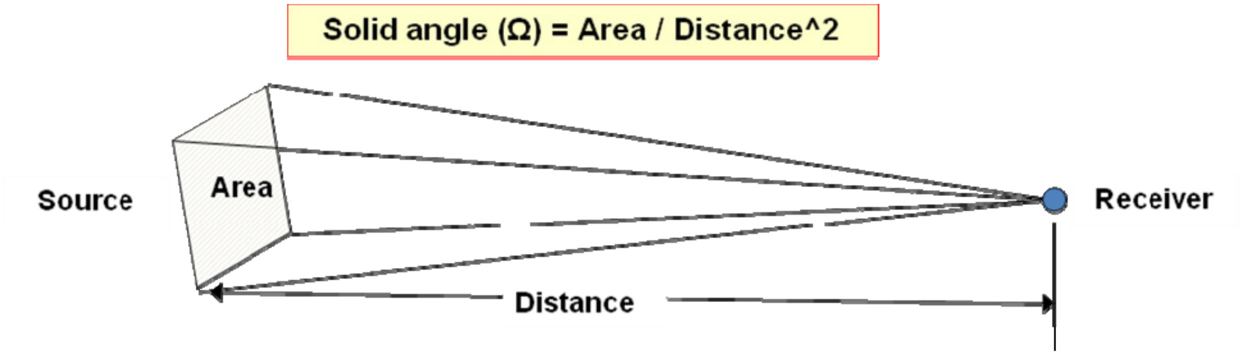

Figure 3. Solid Angle.

A sphere contains a solid angle of 4π steradians; a hemisphere contains 2π steradians, and so on. The area is a curved surface, but in most applications, the solid angles are sufficiently small that the area can be approximated as a plane. Also, for small angles, the solid angle in steradians is approximately equal to the product of two plane angles in radians.

CONVERSIONS

IR wavelengths are typically expressed in **µ**m, visible wavelengths in **µ**m or nm, and UV wavelengths in nm or angstroms. Table 5 lists conversion factors for converting from one unit of wavelength to another. The conversion is from column to row. For example, to convert from **µ**m to nm, multiply the value expressed in **µ**m by 10^3. IR wavelengths are also sometimes expressed in a frequency-like unit called wavenumbers or inverse centimeters. A wavenumber value can be found by dividing 10,000 by the wavelength expressed in **µ**m. For example, 2.5 **µ**m converts to a wavenumber of 4000 or 4000 inverse centimeters (cm^{-1}).

Table 5. Wavelength Conversion Units.

From -> To get ↓	Angstroms - Å	Nanometers - nm	Micrometers - **µ**m
	Multiply by		
Angstroms - Å	1	10	10^4
Nanometers - nm	10^{-1}	1	10^3
Micrometers - **µ**m	10^{-4}	10^{-3}	1

BASIC RADIANT POWER RELATIONSHIPS

Radiant intensity is the most commonly used term to describe the radiant power of a source per unit solid angle. Radiant Intensity offers the advantage of being a source term, like radiance, that is not related to the size of the radiating source. In practice, radiant intensity is a derived term and is not directly measurable. If the Instantaneous Field of View (IFOV), which represents the smallest optical resolution element of a remote sensor, subtends an angle smaller than the size of the radiating source, the sensor responds directly to *radiance*. If the IFOV of the sensor subtends an angle larger than the radiating source, the sensor responds to *irradiance*. The relationship among radiant intensity, radiance, and irradiance is shown in the following equation:

$$I = ED^2 = E\frac{A}{\Omega} = LA = L\Omega D^2$$

Where:

I = Intensity. Radiant intensity is the target source power per unit solid angle, in Watts/steradian.

E = Irradiance. Irradiance is the received power density in Watts/cm2 incident on a distant sensor. Irradiance is the quantity measured or detected by a distant sensor where the target is not spatially resolved (i.e., where the target subtends an angle smaller than the resolution or field of view of the instrument.)

L = Radiance. Radiance is the intensity per unit area of source, in W/cm^2/steradians. Radiance is the quantity measured by a camera or imaging system, where the target is optically resolved (i.e., where the target subtends an angle much greater than the resolution or instantaneous field of view of the imager.)

A = Area. Area is the cross-sectional or projected area of the target in cm^2.

D = Distance. Distance between the target and sensor in cm.

Ω = Solid angle subtended by the target in steradians. Solid angle appears directly or indirectly in many infrared quantities. The solid angle subtended by a source is the ratio of the source area to the square of distance.

All remote sensors receive energy from a source through an atmospheric path. The atmosphere attenuates the propagation of energy due to scattering, absorption, and molecular rotation and vibration, depending on the wavelengths involved. The atmosphere itself also radiates. All incident energy on a remote sensor, except as received under vacuum conditions, is detected through the atmosphere. The term *apparent* is applied to a radiometric quantity to acknowledge the presence of atmospheric effects. A remote sensor, responding the radiance of a source over an atmospheric path, responds to *apparent radiance*.

Real sensors do not respond uniformly to energy. Sensors have non-uniform spectral and spatial response due to many factors, including detector and optical characteristics. All energy received by the detector is spectrally weighted by the spectral response of the instrument. The term *effective* is applied to acknowledge the non-uniform response of the instrument.

Outside of a vacuum, all radiation incident on a remote sensor is attenuated by atmospheric effects such as scattering and absorption. Where the atmosphere absorbs, it also emits. Path radiance is the term applied to the contribution of the atmospheric path to the received radiation. The term *apparent* is applied to radiometric quantities to acknowledge the influence of the atmosphere on the received radiation.

DOMAINS, DISTRIBUTIONS, AND SENSOR RESPONSE

Radiant power that is emitted or reflected from a source is distributed across multiple dimensions or domains. The three domains that are most important to radiometric applications are spectral (Figure 4), spatial (Figure 5), and temporal (Figure 6). Knowing how the radiant power from a source is distributed across each domain is important to understanding performance of an EW system, whether it be a sensor, a weapon, or a countermeasure.

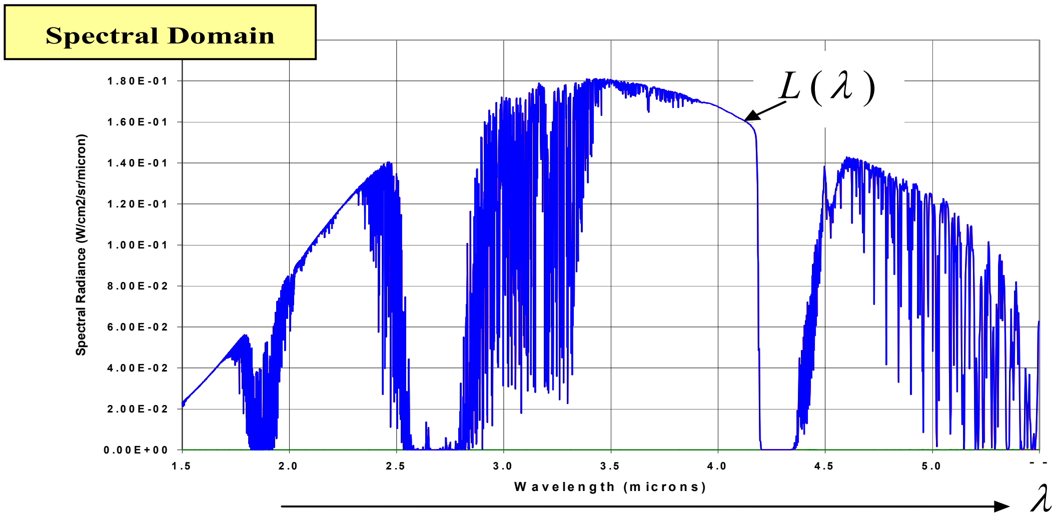

In the spectral domain, radiation is distributed as a function of wavelength (or wavenumber). Radiation from solid materials is distributed as a continuum in accordance with Planck's formula. The graph above shows the spectral distribution of radiation from a 600 degree Celsius blackbody after passage through the atmosphere. Radiation from gases appears at specific wavelength "lines" corresponding to molecular resonances.

Figure 4. Sample Spectral Distribution of a Solid Material.

Spatial Domain

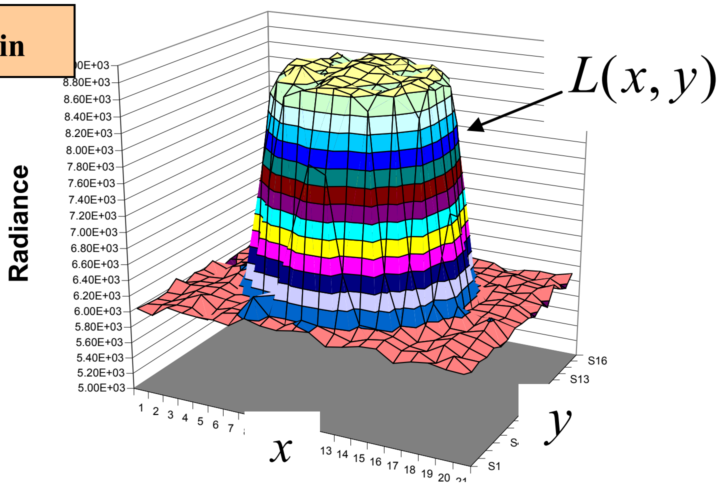

In the spatial domain, radiation is distributed as a function of angle, position, size, shape, or orientation. The graph above shows the radiance of a circular target as viewed by an imaging system.

Figure 5. Sample Spatial Distribution of a Circular Target Image.

Temporal Domain

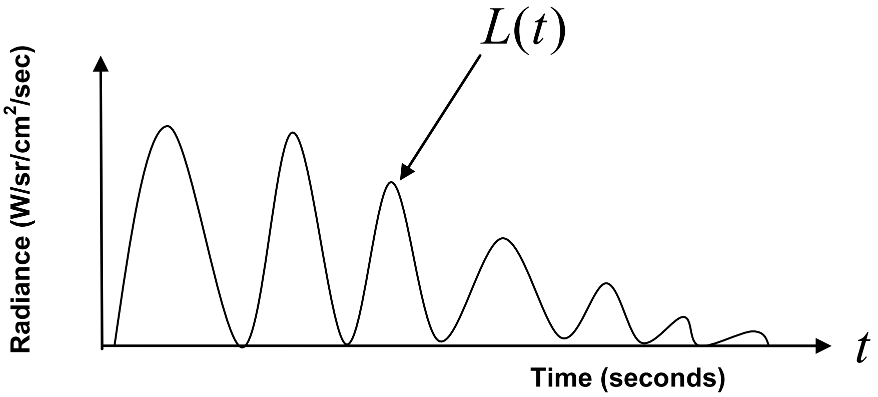

In the temporal domain, radiation is distributed as a function of time (or frequency). Most IR target sources, such as aircraft and ground vehicles, change slowly with time and can be considered steady state for measurement purposes. However, IR countermeasures systems and devices, such as jammers and decoy flares, have high frequency content that must be considered in their design and effectiveness assessment.

Figure 6. Sample Temporal Distribution of a Modulated Source.

All electro-optical sensors have some sensitivity to how radiation is distributed across the different domains. This sensitivity or response takes the following forms:

1. **Spectral response**:
    a. May be broad in the case of a radiometer or imager, or narrow in the case of a spectrometer where response takes the form of resolution.
    b. Spectral response is chosen to exploit particular spectral features of interest in a target.
    c. Non-uniformity of spectral response can have a significant effect on the performance of sensor.

2. **Spatial response**:
    a. May be broad in the case of a radiometer or spectrometer, where spatial response takes the form of sensitivity across the sensor's FOV. In the case of an imager, the spatial response may be narrow because spatial response takes the form of the resolution or IFOV.
    b. Required spatial resolution is defined by the size and radiant power of the source, the distance between the source and the target, and by the sensitivity of the sensor.
    c. Non-uniformity of spatial response can have a significant effect on the performance of sensor.
    d. Response to radiation from the background within the FOV is primarily a spatial response issue and can be significant to sensor performance.

3. **Temporal response**:
    a. Takes the form of bandwidth or, in the case of a digital system, of sampling frequency.
    b. Unlike spectral and spatial response, temporal response includes the response time of the detector and all of the electronics, data transfer time, and processing time for a sensor.
    c. Non-uniformity across a region is usually not a concern. Temporal response is usually uniform for target frequencies that fall well below the low-pass or Nyquist breakpoints.

Sensors respond to received radiant power by integrating the power distribution in each domain under the corresponding response curve. Sensors respond in multiple domains, so output is proportional to a multi-dimensional, weighted integral. Figure 7 illustrates response-weighting, or convolution, of radiation with instrument response in the spectral domain. The concept applies equally in the spatial domain.

Response Weighting in the Spectral Domain

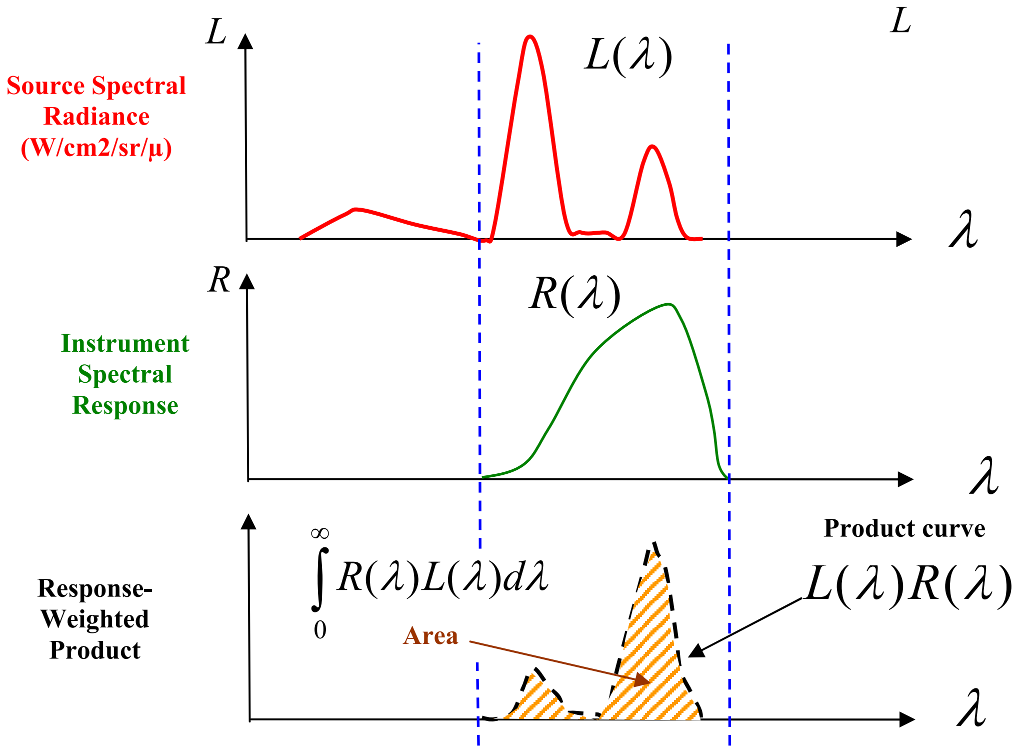
L
Source Spectral Radiance (W/cm2/sr/µ)
$L(\lambda)$
L
λ
R
$R(\lambda)$
Instrument Spectral Response
λ
Product curve
$\int_0^\infty R(\lambda)L(\lambda)d\lambda$
$L(\lambda)R(\lambda)$
Response-Weighted Product
Area
λ

Non-uniform sensor response across a domain distorts the distribution of received radiant power by weighting the radiation at some wavelengths differently than others. Instrument output is proportional to the integral of this weighted product, which makes instrument response shape an embedded part of both the measurement and the calibration. Some consequences of non-uniform response are:

- Two different sources with identical power, but different distributions will produce different measurement values.
- If two different sources (such as the calibration and the target) have similar relative power distributions, the relative weighting by the instrument will be the same for both sources, so the shape and degree of non-uniformity has little or no effect. This will be discussed in more detail later as a key strategy in calibration to reduce measurement uncertainty.

Figure 7. Response-Weighting in the Spectral Domain.

The consequence of instrument output being proportional to multiple weighted integrals is that the characteristics of the instrument affect the measured value in ways that cannot be easily extracted or corrected. Understanding non-uniformity in sensor response is critical to understanding performance and evaluating the effectiveness of the system.

INFRARED SOURCE CHARACTERISTICS

Reflectance

Reflectance is generally categorized as either *specular* (mirror-like) or *diffuse* (scattered by reflection from a rough surface). Most surfaces exhibit both types of reflection, but one typically dominates. Reflections from smooth surfaces are specular. Reflectance is a unitless quantity between zero and one that is the ratio of reflected power to the incident power. Reflectance varies with angle from the normal and with wavelength.

Diffuse reflectors scatter incident radiation broadly. Johann Heinrich Lambert described an ideal diffuse reflector in which the intensity of reflected rays is distributed as a cosine of the angle from the normal, regardless of the angles of the arriving incident rays. Such a surface is described as *Lambertian*. The same relationship applies to a diffuse source. If the source is perfectly diffuse, the radiance is independent of the viewing angle because the projected area of the source also varies as a cosine function of the angle from normal.

While no surface is perfectly diffuse, standard paint on most aircraft, for example, is near enough for at least a first approximation. Significant effort has gone into measuring and describing surface properties of military vehicles to support modeling and simulation activities. Bidirectional Reflectance Distribution Function (BRDF) is defined as the ratio of the reflected radiance to the incident irradiance at a wavelength, λ.

The BRDF provides a complete description of the reflectance properties of a surface. Measurements of BRDF are made in specially equipped optical laboratories. For aircraft, a typical use of BRDF data is in computer models that predict the reflections of the earth and sun from the aircraft fuselage and wings.

Emissivity

Emissivity is a unitless quantity that is a measure of the efficiency of a surface as an absorber or emitter. Emissivity is expressed as a number between 0.0 and 1.0. According to Kirchoff's Law, for an opaque object in thermal equilibrium, i.e., no net heat transfer, the emissivity equals the absorptance. In other words, a perfect emitter is also a perfect absorber. This ideal emitter is known as a *blackbody*. A surface with an emissivity of 1.0 emits the maximum radiation that Planck's Law allows. A body for which the emissivity is constant with wavelength but less than 1.0 is commonly known as a *graybody*. Planck's equation is typically expressed for an ideal blackbody emitter, but multiplying the blackbody expression by the emissivity term expresses the spectral distribution of power for a graybody.

Electromagnetic (EM) radiation, including infrared radiation, can be characterized by amplitude, frequency (or wavelength), coherency, and polarization properties. Amplitude refers to the magnitude of the electromagnetic wave. Coherency refers to the degree in which the electromagnetic waves maintain a constant phase difference, both spatially and temporally. Polarization is a description of the orientation of the wave's propagation perpendicular to the direction of travel. Like all electromagnetic radiation, infrared radiation travels as an EM field with the propagation velocity of the speed of light in a vacuum, and slower through air and other dielectrics. , Infrared radiation also exhibits both wave and particle properties. Which property is used depends on the application: Waves are used for applications involving propagation and geometrical optics, while particles (photons) are used for most detector applications. The energy of a photon is inversely proportional to its wavelength (*hc*/λ) where *h* is the Planck's constant, and *c* is the speed of light.

Infrared radiation also interacts with matter in a variety of ways:

- Reflects – wave is reflected from a surface.
- Refracts – direction of wave bends when passing between two transparent media with different propagation speeds.
- Scatters – when interacts with particles whose size approaches the wavelength of the radiation.
- Diffracts – around edges of an obstruction.
- Interferes – constructively and destructively.
- Absorbs – when absorbed by matter, radiation is (usually) converted into heat energy.
- Emits – radiation is emitted from matter by conversion from another form of energy.
- Transmitted – propagates through a transparent medium (or vacuum).

Plank's Law:

Infrared is directly related to the heat radiated from matter. Anything with a temperature above absolute zero (-273.15 degrees Celsius) radiates in the infrared. Planck's Law, discovered by German Physicist Max Planck, mathematically describes the distribution of radiant power across the spectrum for a given temperature. The form presented below calculates source radiance in Watts/sr/cm^2/ μ:

$$L_\lambda = \frac{C_1}{\pi\lambda^5\left[e^{\left(\frac{C_2}{\lambda T}\right)} - 1\right]}$$

Where: $C_1 = 2\pi c^2 h = 3.7415 \times 10E{+}04 \ \mathrm{W\ cm^{-2}}\ \mu^4$
$C_2 = ch/k = 1.4389E{+}04 \ \mu\ \mathrm{K}$
c = speed of light; h = Plank's constant; k = Boltzmann's constant
With λ in μ and T in Kelvin (= ºC + 273.15)

The behavior of Planck's curve with temperature is fundamental to every infrared detection scenario. Figure 8 shows Planck's equation for a single temperature. The shape of this curve resembles a wave of water, with a steep rise in power on the short wavelength side of the peak and a tail off on the long wavelength side.

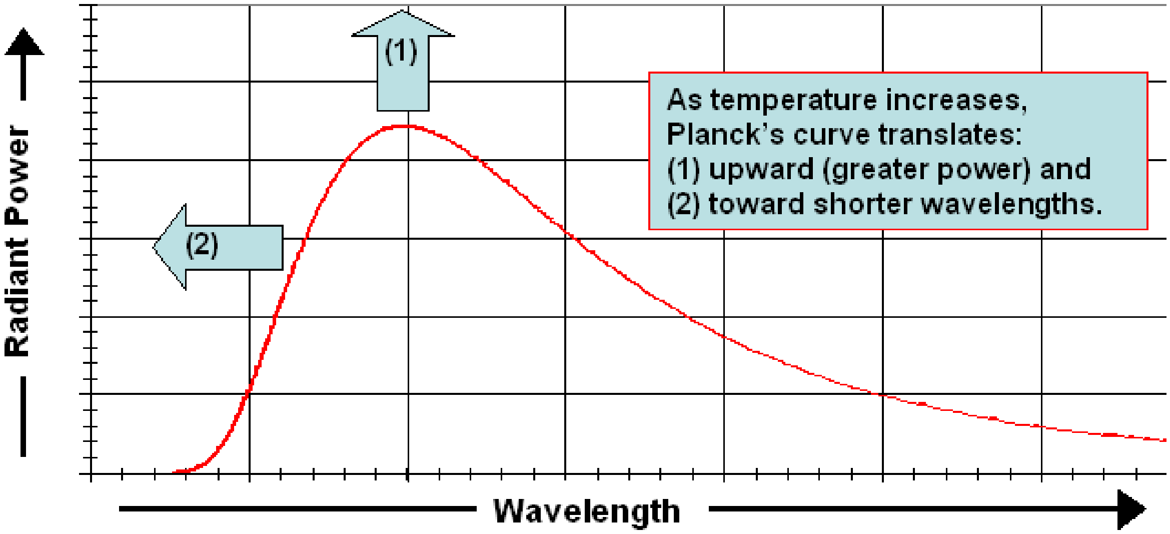

Figure 8. Temperature Effects and Power Distribution.

When temperature increases, two significant changes in power distribution occur and are governed by the following two relationships:

Stefan-Boltzmann Law:

The total power under the curve increases proportionally to the fourth power of the temperature. According to the Stefan-Boltzmann law, the total radiant emittance (power) of a blackbody is:

$$M = \sigma T^4 \qquad \text{Where: } \sigma = \frac{2\pi^5 k^4}{15 c^2 h^3} = 5.67 \times 10^{-12} \, Watts \, cm^{-2} \, {}^\circ K^{-4}$$

and T is in Kelvin

This is Plank's radiation law integrated over all values of λ.

Wein's Displacement Law

The peak emission translates toward shorter wavelengths.

Wein's displacement law takes the derivative of the Plank's law equation to find the wavelength for maximum spectral exitance (emittance) at any given temperature:

$$\lambda_{max} = 2897.8 / \mathbf{T} \qquad \text{where T = temperature is in Kelvin}$$

The surface of the sun radiates with a spectral distribution like that of a 5900 Kelvin source. In accordance with Wein's Law, the wavelength at which the radiation for a 5900 Kelvin source peaks is approximately 490 nanometers. The maximum sensitivity of the day-adapted human eye (photopic) occurs at about 555 nanometers, which happens to be in a highly transparent region of the atmosphere. In other words, the temperature of our sun and the transmission of our atmosphere are conveniently matched to the response of the human eye. Figure 9 shows the relationship between the response of the human eye and the spectral distribution of solar irradiance.

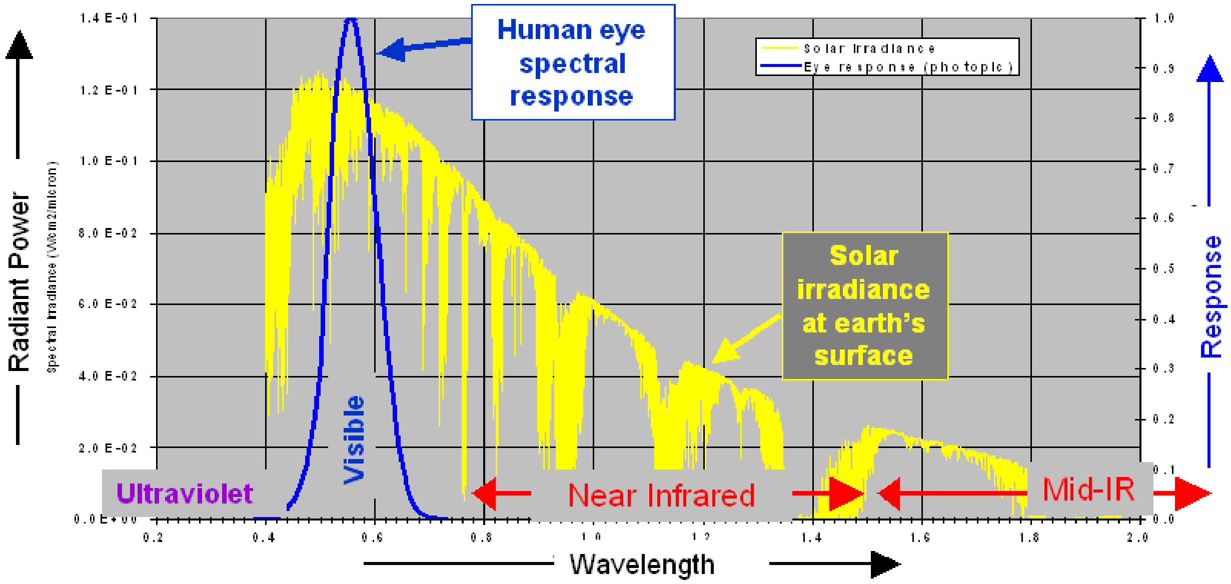

Figure 9. Power Distribution of the Sun Related to the Human Eye.

Unlike other electromagnetic radiation, the peak radiation from objects at the temperature of the earth, and of humans, vehicles, aircraft, etc., lies at infrared wavelengths. This makes the infrared portion of the spectrum militarily important.

For detection of an object, whether it is with the human eye or with an electro-optical sensor, it is not where the maximum emissions occur in the spectrum that is important. It is where the maximum difference between the target and background lies that is most significant.

The *absolute* signature of a source or target is the power radiated and reflected by the target without influence from background radiation. Absolute signature can be measured with an instrument such as an imaging camera at close range, producing highly resolved IR imagery. The difference between radiation from a target and its background is its *contrast*. Contrast is the target signature relative to a specific background. Contrast irradiance is the quantity detected by any remote sensor or missile. Contrast varies greatly with background conditions.

Figure 10 shows several different types of source distributions encountered by infrared sensors. Most targets of military interest are a complex combination of spectral contributions from a variety of sources. For example, the signature of an aircraft can typically be described by the contributions from airframe, hot engine parts, and plume.

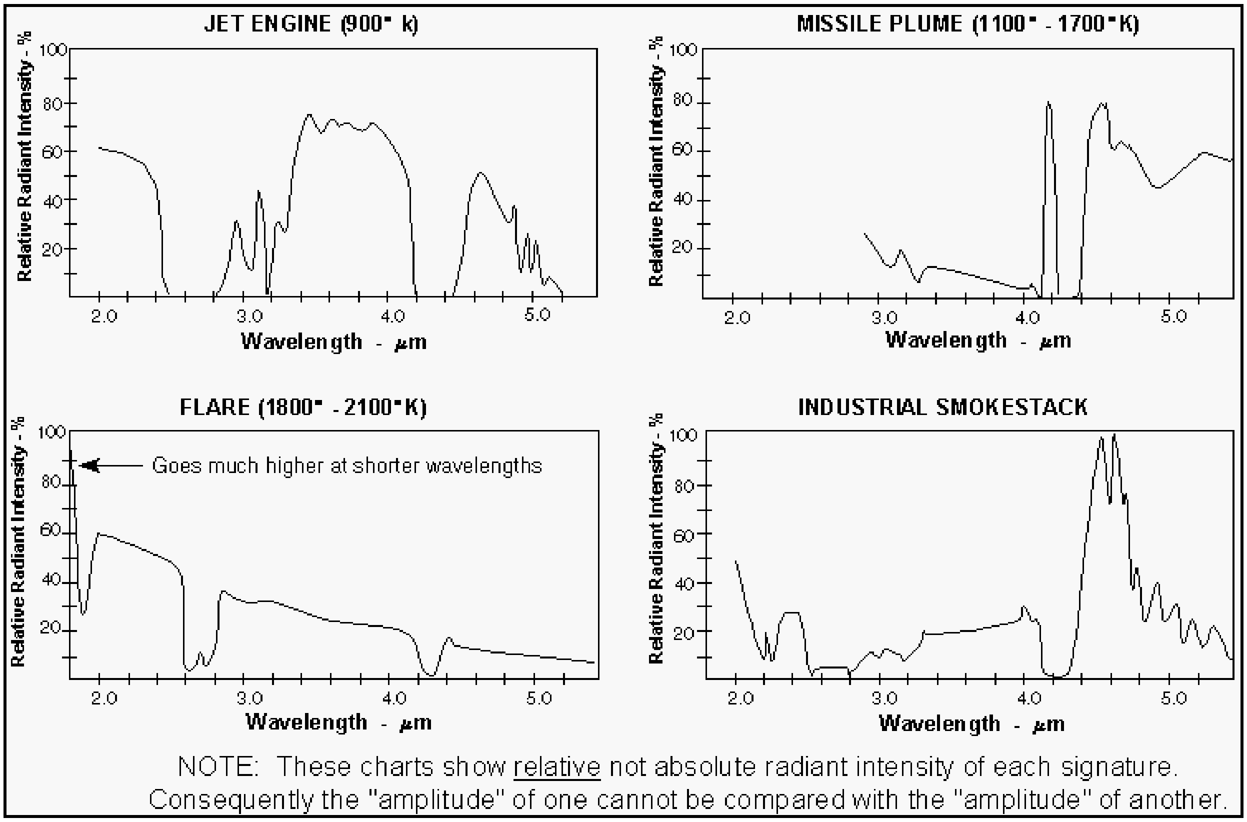

NOTE: These charts show relative not absolute radiant intensity of each signature. Consequently the "amplitude" of one cannot be compared with the "amplitude" of another.

Figure 10. Spectral Distribution of Various Targets.

ATMOSPHERIC TRANSMISSION

The radiation emitted or reflected from the targets and backgrounds must pass through the intervening atmosphere before reaching the detection system. The radiation is absorbed and re-emitted by molecular constituents of the atmosphere and scattered into and out of the path by various aerosol components.

Figure 11 reveals the presence of atmospheric windows, i.e. regions of reduced atmospheric attenuation. IR detection systems are designed to operate in these windows. Combinations of detectors and spectral bandpass filters are selected to define the operating region to conform to a window to maximize performance and minimize background contributions.

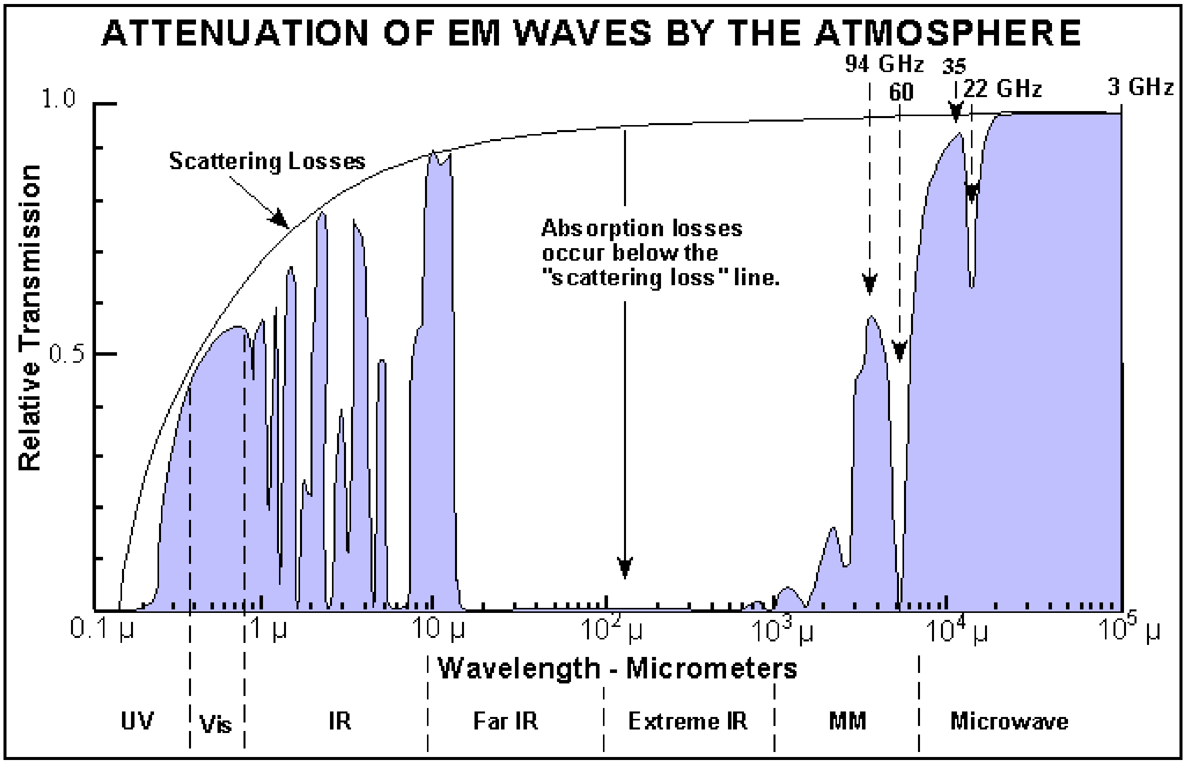

Figure 11. Atmospheric Transmission Over 1 NM Sea Level Path.

The molecules that account for most of the absorption in the IR region are water, carbon dioxide, nitrous oxide, ozone, carbon monoxide, and methane. Figure 12 shows an expanded view of the infrared portion of the spectrum.

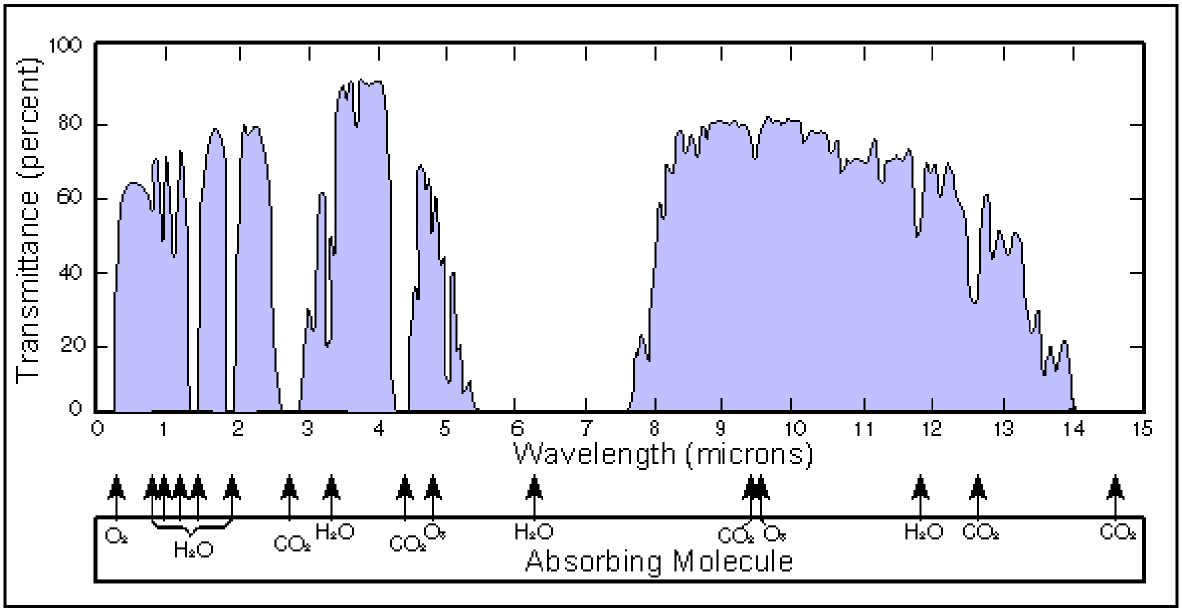

Figure 12. Transmittance of Atmosphere Over 1 NM Sea Level Path (Infrared Region).

The transmission in a window is greatly dependent on the length and characteristics of the path. As path length increases, absorption gets deeper and broader. Water vapor also has a significant effect overall on transmission through the atmosphere. High relative humidity attenuates transmission at all optical wavelengths.

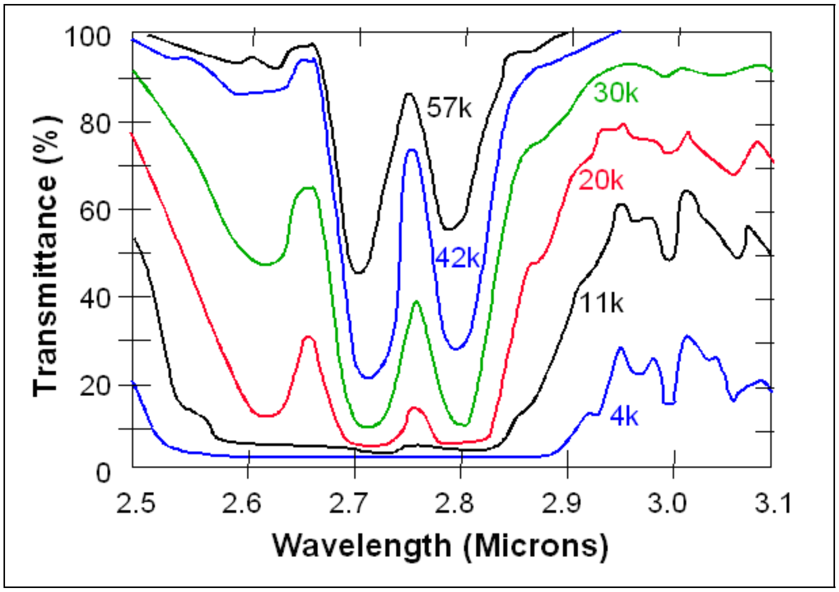

Figure 13. Transmittance at Various Altitudes.

Since water vapor generally decreases with altitude, transmission generally increases and path length becomes the determining factor. However, path length does not affect transmission of all wavelengths the same.

The altitude effects on transmittance are shown in Figure 13.

A great deal of work has gone into developing high fidelity atmospheric models. One of the most commonly used tools is MODTRAN™ (MODerate Resolution Atmospheric TRANsmission). MODTRAN™ models atmospheric propagation of EM radiation from 0.2 to 100 um. MODTRAN was developed by Spectral Sciences Inc. and the United States Air Force Research Laboratory.

ELECTRO-OPTICAL COMPONENTS AND SENSORS

Almost all IR instruments, missiles, search systems, etc. have similar functional components. Basic components typically include:

- Optics - Reflective or refractive lenses to:
  - Collect radiation. Irradiance (power density) is increased by collecting radiation over a large area and focusing down to a small area.
  - Form or focus an image that will be used to extract information about the target.

- Filter(s) - Spectral and spatial filters to distinguish target from background and to extract
  - Target information.
  - Spectral filters restrict sensitive wavelength range.
  - A spatial filter separates image information by features such as size or position.

- Detector - A transducer to convert received radiation to an electrical signal for processing.

- Electronics - Used to amplify and condition the detector signal and perform some action, such as controlling a servo for tracking and guidance or recording the received information. In addition to the above, the optical head may also contain a window to protect the electronics and an output unit consisting of indicators or displays.

Windows / Domes / Lens Materials

For most applications of EO systems in EW the detection system is protected from the environment by a window or dome of optically transmissive material. The window operates both as a weather seal and, in some cases, helps to define the spectral response region of the system. In some cases the window functions as a lens.

IR energy interacts with matter in ways we associate with light (reflection, refraction, and transmission). Lower energy of IR photons results in different optical properties than light. For example:

- Glass and water are not transparent to wavelengths longer than about 3 microns.
- Silicon and germanium are highly transparent to infrared radiation, but are opaque to visible light.

Transmission bands of representative window or lens materials are shown in Figure 14. The end points depicted are for the 10% transmission wavelengths. Not shown in Figure 14 are the various UV transmissive glasses such as Pyrex, Corex, and Vycor or Amorphous Material Transmitting IR radiation (AMTIR) which are various combinations of Ge-As-Se (AMTIR-1), As-Se (AMTIR-2, 4, & 5), Ge-Sb-Se (AMTIR-3), As-S (AMTIR-6).

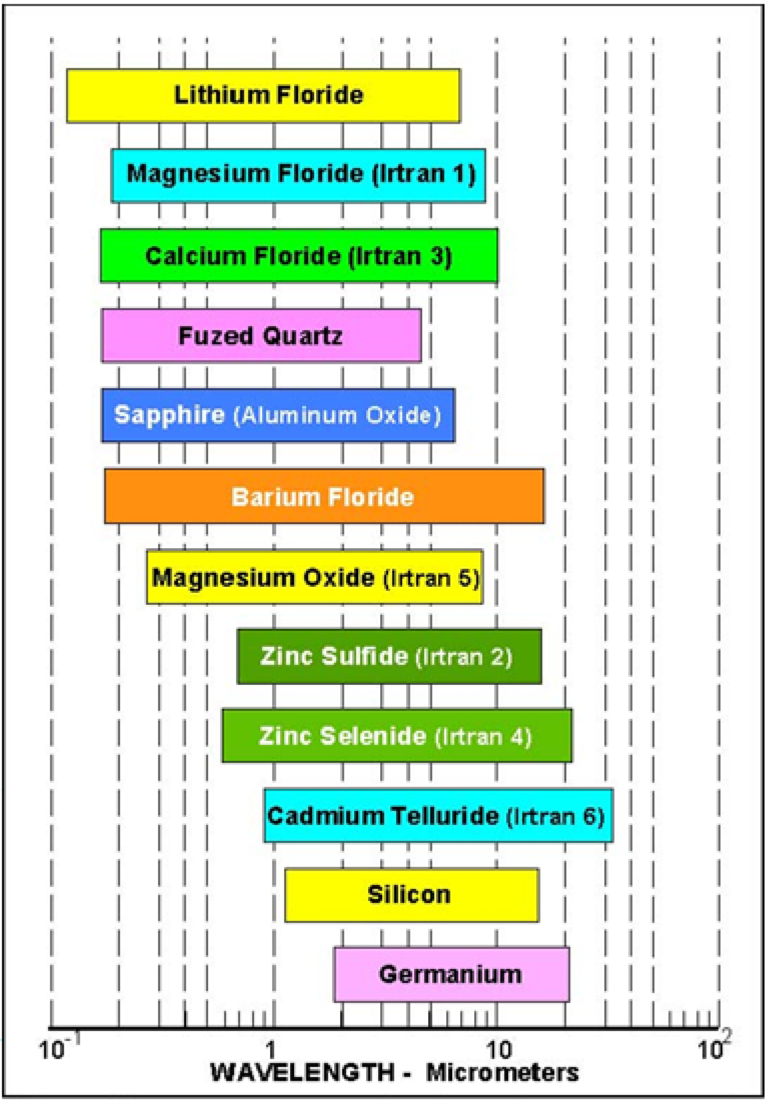

Figure 14. Transmission of Selected Window / Lens Materials.

Objective Lens

The objective lens is the first optical element in a lens in a typical sensor or missile seeker. The objective lens serves two main functions:

- Collect radiation (i.e., multiply irradiance (power density) by collecting over a large area and focusing onto a small area)
- Form an image of the target scene onto a filter and detector array.

Lens Types

Lenses for the IR can be either refractive or reflective. Refractive optics are “straight through” lenses, with the light never making large bends as shown in Figure 15.

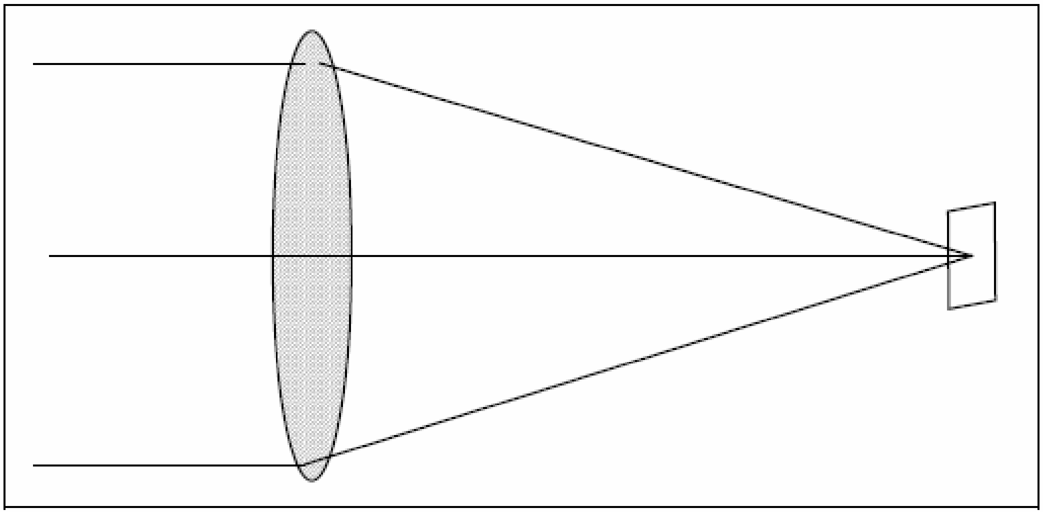

Figure 15. Refractive Lens.

A common reflective design used in many missiles is the Cassegrain. The Cassegrain design shown in Figure 16 is compact in size for its focal length.

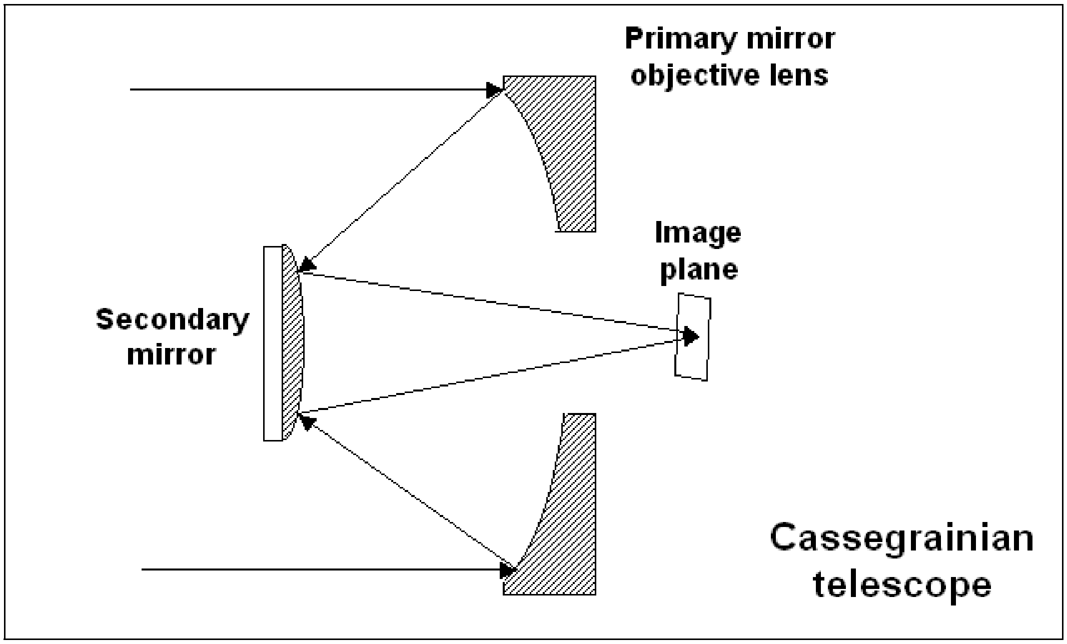
Primary mirror
objective lens
Image
plane
Secondary
mirror
Cassegrainian
telescope

Figure 16. Cassegrain Lens.

Optical Filters

Filters may be divided into two major categories: (1) spectral and (2) spatial.

Most optical radiation detectors have a wider sensitivity band than desired for the particular application. Spectral filters restrict sensitive wavelength range. Reasons for filtering include: enhancement of target-to-background contrast, avoidance of unwanted plume emissions and/or atmospheric absorption regions, and extraction and measurement of target spectral features.

To further define the system sensitivity, band interference filters or absorption filters are used. An absorption filter is a bulk material with a sharp cut-on or cut-off in its transmission characteristic. A cut-on and a cut-off filter can be combined to make a bandpass filter. By selecting absorption characteristics of absorption filters combined with the response of a detector, the desired system response can be obtained. An interference filter is composed of dielectric coatings on an appropriate substrate combined in such a way to produce cut-on, cut-off, or bandpass filters. Interference filters allow more control of the final response characteristics and smaller elements. See Figure 17.

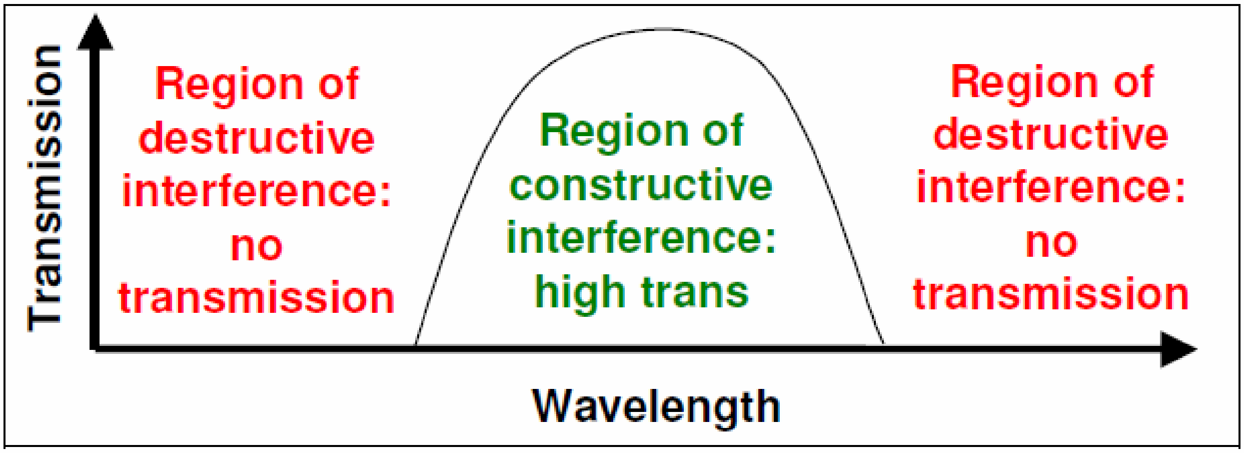
Transmission
Region of destructive interference: no transmission
Region of constructive interference: high trans
Region of destructive interference: no transmission
Wavelength

Figure 17. Spectral Bandpass Filter.

Most spectral filters are of the thin-film interference type. Layers of dielectric material are vacuum deposited on a substrate window material. Typical substrate materials in IR are sapphire, silicon, and germanium. Thickness of deposited layers designed to have constructive interference to pass desired radiation at desired wavelengths and destructive interference to block undesired wavelengths Besides spectral filters, EO system optics often have antireflection (or AR) coatings to eliminate or greatly reduce unwanted reflections between optical elements.

A spatial filter separates information in a scene image by features such as size or position. Spatial filters take a variety of forms. Some common types and their functions include:

- Field stop: limits an instrument's field of view. Blocks unwanted sources (such as sun) outside nominal field of view.
- Mechanical modulator or "chopper."
- Reticle: A mechanical modulator used in many missile designs. Usually discriminates against extended sources (such as background) in favor of "point" target sources and provides target directional information from modulation phase.

Detector Coolers

Many IR detectors have to be cooled for proper operation. Most systems use closed-cycle coolers or thermoelectric coolers. Thermoelectric coolers use the Peltier effect, which produces a reduced temperature by passing a d-c current through a thermoelectric junction. Multi-stage coolers can cool a detector down to below 200ºK. Closed-cycle coolers typically are of the Stirling cycle design and utilize the expansion of a gas (helium) to cool a cold finger attached to the detector. These generally operate at liquid nitrogen temperature (77ºK).

Detectors

A detector is a transducer that transforms electromagnetic radiation into a form, which can be more easily detected. In the detectors of interest to EW the electromagnetic radiation is converted into an electrical signal. In some systems the signal is processed entirely within the system to perform its function. In others the signal is converted to a form to allow the human eye to be used for the final detection and signal analysis.

Detectors are transducers than convert optical radiation into electrons. The physical effects by which electromagnetic radiation is converted to electrical energy are divided into two categories: photon effects and thermal effects. EW systems primarily use detectors dependent on photon effects. These effects can be divided into internal photo effects and external photo effects.

The external photo effect is known as photoemission. In the photoemissive effect, photons impinging on a photocathode drive electrons from its surface. These electrons may then be collected by an external electrode and the photocurrent thus obtained is a measure of the intensity of the received radiation.

Internal photoeffects of interest are the photoconductive effect and the photovoltaic effect. In the photoconductive effect, absorbed photons cause an increase in the conductivity of a semiconductor. The change is detected as a decrease in the resistance in an electrical circuit. In the photovoltaic effect, absorbed photons excite electrons to produce a small potential difference across a p-n junction in the semiconductor. The photovoltage thus produced may be amplified by suitable electronics and measured directly.

Thermal detectors respond directly to heat. Examples of these devices include bolometers, thermopiles, and pyroelectric detectors. The pyroelectric effect is an example of the thermal effect. The pyroelectric effect is a change in polarization in a crystal due to changes in temperature. Radiation falling on such a crystal is detected by observing the change in polarization as a build up of surface charge due to local heating. When coated with a good black absorber, the crystal will be sensitive to a wide band of wavelengths.

Figure 18 shows the spectral sensitivity range of typical detectors using these effects.

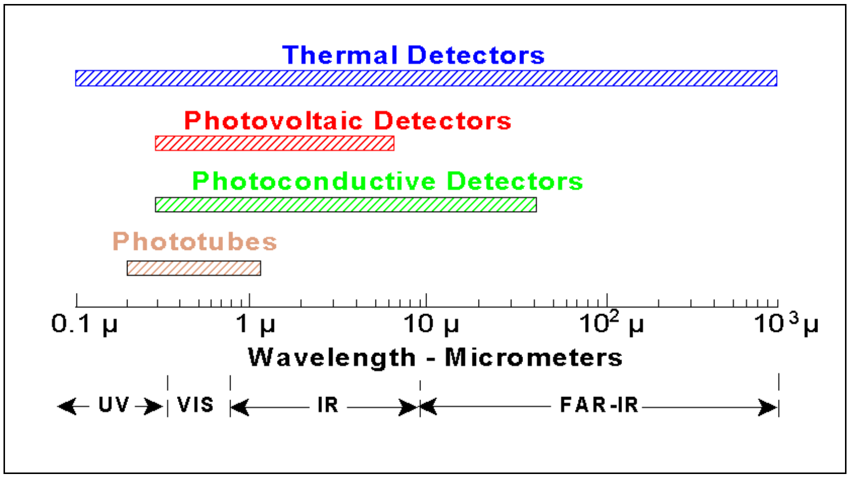

Figure 18. Spectral Range of Various Detectors.

Detector Types

Photoconductive detectors operate as resistors in a circuit. The resistance of the detector changes as the radiation incident on its surface changes. For EW applications, the most photoconductive detector types include: Indium Antimonide (InSb), which can also be operated in photovoltaic mode; Gallium Arsenide (GaAs); Lead Sulfide (PbS); and Lead Selenide (PbSe).

Photovoltaic detectors, the most common detectors used in modern EW and military sensor applications, produce a voltage that is proportional to the incident radiation. Common examples of photovoltaic detectors are Indium Antimonide (InSb) and Mercury Cadmium Telluride (HgCdTe). Both of these detector types offer high sensitivity when cryogenically cooled.

Diode phototubes and photomultipliers are commonly used detectors for UV systems including many operational missile-warning systems. These types of tubes offer the advantage of operating uncooled which can significantly reduce the complexity of a sensor system and offer increased reliability. Most of the modern IR sensors require cooled detectors. InSb, for example, requires cooling to 77 Kelvin to achieve the necessary sensitivity. Depending on the application, HgCdTe can be operated at somewhat higher temperature conditions.

A Photoelectromagnetic (PEM) detector has a junction that generates a current when exposed to light in a magnetic field.

Some detectors (such as InSb) have multiple modes of operation, including: Photoconductive (PC), Photovoltaic (PV), or Photoelectromagnetic (PEM) modes of operation.

Detector Parameters and Figures of Merit

The important parameters in evaluating a detector are the spectral response, time constant, the sensitivity, and the noise figure.

The spectral response determines the portion of the spectrum to which the detector is sensitive.

The time constant is a measure of the speed of response of the detector. It is also indicative of the ability of the detector to respond to modulated radiation. When the modulation frequency is equal to one over the time constant, the response has fallen to 70.7 % of the maximum value. The time constant is related to the lifetime of free carriers in photoconductive and photovoltaic detectors and to the thermal coefficient of thermal detectors. The time constant in photoemissive devices is proportional to the transit time of photoelectrons between the photocathode and anode.

The sensitivity of a detector is related to its responsivity. The responsivity is the ratio of the detected signal output to the radiant power input. For photoconductive and photovoltaic detectors the responsivity is usually measured in volts per watt -- more correctly, RMS volts per RMS watt. However, the sensitivity of a detector is limited by detector noise. Responsivity, by itself, is not a measure of sensitivity. Detector sensitivity is indicated by various figures of merit, which are analogous to the minimum detectable signal in radar. Such a quantity is the noise equivalent power (NEP). The NEP is a measure of the minimum power that can be detected. It is the incident power in unit bandwidth, which will produce a signal voltage equal to the noise voltage. That is, it is the power required to produce a signal-to-noise ratio of one when detector noise is referred to unit bandwidth. The units of NEP are usually given as watts, but more correctly, are watts/Hz$^{1/2}$ or watts·sec$^{1/2}$.

Another figure of merit is the noise equivalent irradiance (NEI). The NEI is defined as the radiant power per unit area of the detector required to produce a signal-to-noise ratio of one. The units of NEI are watts per square centimeter.

Noise equivalent power (NEP) is the radiant power required to produce a signal to noise ratio of one for a detector. Detectivity (D) of a detector is defined as the reciprocal of the NEP. The units of D are watts $^{-1}$·sec $^{-½}$. A higher value of detectivity indicates an improvement in detection capability. Since D depends on detector area, an alternate figure of merit, known as D-star (D*). D* is the detectivity measured with a bandwidth of one hertz and reduced to a responsive area of one square centimeter. The units of D* are cm·watts $^{-1}$·sec $^{-½}$. D* is the detectivity usually given in detector specification sheets. Typical spectral detectivity characteristics for various detectors are shown in Figure 19.

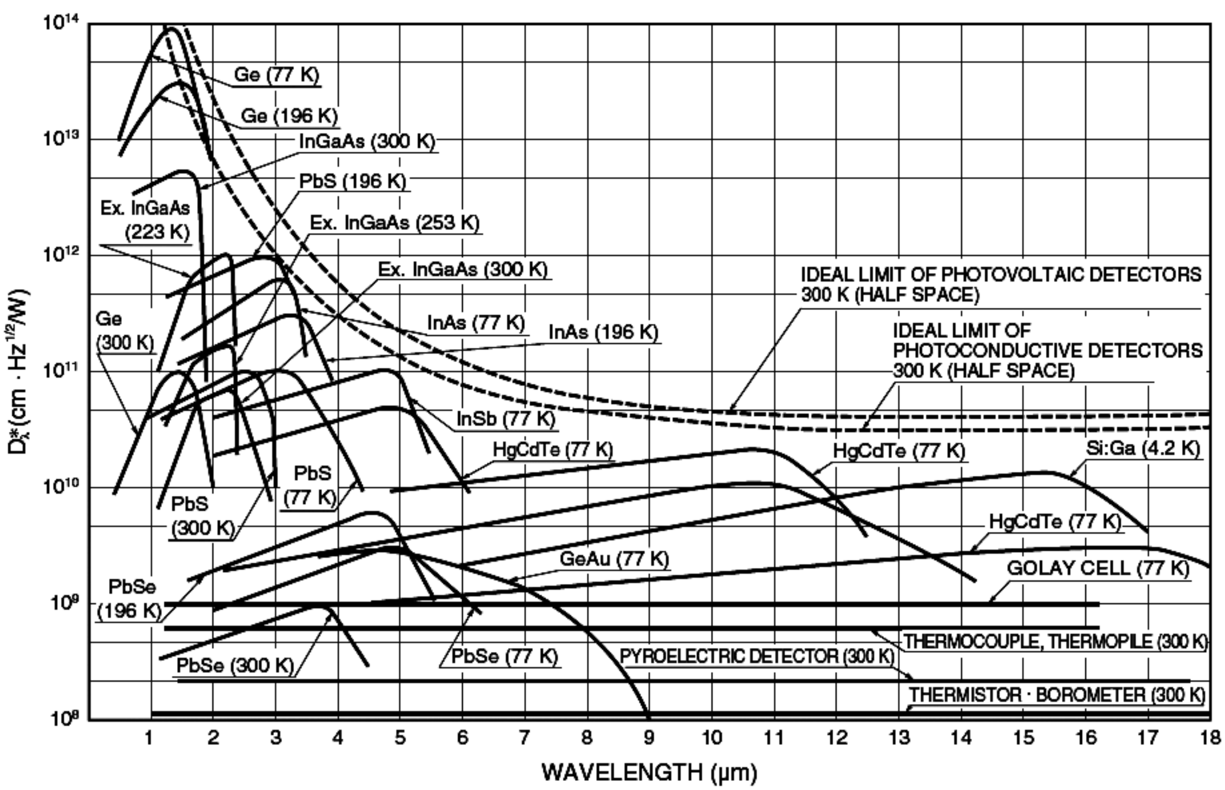

Figure 19. Spectral Detectivity of Various Detectors.

Besides the NEI mentioned above, the quantum efficiency of the photocathode is also a figure of merit for photoemissive devices. Quantum efficiency is expressed as a percent -- the ratio of the number of photoelectrons emitted per quantum of received energy expressed as a percent. A quantum efficiency of 100 percent means that one photoelectron is emitted for each incident photon.

There are other figures of merit for television cameras. The picture resolution is usually described as the ability to distinguish parallel black and white lines and is expressed as the number of line pairs per millimeter or TV lines per picture height. The number of pixels in the scene also defines the quality of an image. A pixel, or picture element, is a spatial resolution element and is the smallest distinguishable and resolvable area in an image. CCD cameras with 512 x 512 elements are common.

Another resolution quantity is the gray scale, which is the number of brightness levels between black and white a pixel can have.

Noise in Detectors

The performance of a detector is limited by noise. The noise is the random currents and voltages that compete with or obscure the signal or information content of the radiation. Five types of noise are most prominent in detectors: (1) thermal, (2) temperature, (3) shot, (4) generation-recombination, and (5) 1/f noise.

Thermal noise, also known as Johnson noise or Nyquist noise, is electrical noise due to random motions of charge carriers in a resistive material.

Temperature noise arises from radiative or conductive exchange between the detector and its surroundings, the noise being produced by fluctuations in the temperature of the surroundings. Temperature noise is prominent in thermal detectors.

Shot noise occurs due to the discreetness of the electronic charge. In a photoemissive detector shot noise is due to thermionic emission from the photocathode. Shot noise also occurs in photodiodes and is due to fluctuations in the current through the junction.

Generation-recombination noise is due to the random generation and recombination of charge carriers (holes and electrons) in semiconductors. When the fluctuations are caused by the random arrival of photons impinging upon the detector, it is called photon noise. When it is due to interactions with phonons (quantized lattice vibrations), it is called generation-recombination noise. Johnson noise is predominant at high frequencies, shot noise predominates at low frequencies, and generation-recombination and photon noise are predominant at intermediate frequencies.

As the name implies, 1/f noise has a power spectrum that is inversely proportional to frequency. It is dominant at very low frequencies. In photoemissive detectors it is called flicker noise and has been attributed to variation in the emission from patches of the photocathode surface due to variation in the work function of the surface. In semiconductors 1/f noise is also called modulation noise. Here it is apparently due to surface imperfections and ohmic contacts (which are a form of surface imperfection).

Infrared Spectral Region and Features of Interest

Different portions of the infrared spectrum are common for particular applications. The reasons for the selection of a specific window are often sensitive and beyond the scope of this document, but selections are typically based on several key considerations:

- Target characteristics such as size and spectral distribution of signature.
- Background radiance and clutter.
- Atmospheric effects (transmission, path radiance, scintillation, etc.).
- Distinguishing characteristics between natural and man-made sources.

Table 6 describes some of the types of characteristics that are prevalent in the short-wavelength (SWIR) infrared (0.7 to 3.0 microns), mid-wavelength (MWIR) infrared (3.0 to 6.0 microns) and long-wavelength (LWIR) infrared (7.0 to 14.0 microns) along with some types of systems that operate in these regions.

Table 6. Infrared Features, Regions, and Types of Systems.

Near or Short Wave IR (SWIR)	**Dominant natural source:**	**Sun**
	Atmospheric:	
	Transmission:	High
	Path radiance:	Scattered sunlight
	Dominant aircraft IR component:	Sunlit airframe
	Anti-aircraft threat:	Vehicle-launched SAM
Mid-wave IR (MWIR)	**Dominant natural source:**	**Sun**
	Atmospheric:	
	Transmission:	High transmission "windows" between H_2O and CO_2 absorption
	Path radiance:	Scattered sunlight below 3 microns Thermal at longer than 3 microns
	Dominant aircraft IR component:	Engine hot parts and plume
	Anti-aircraft threat:	All AAMs and SAMs
Far or Long-wave IR (LWIR)	**Dominant natural source:**	**Earth**
	Atmospheric:	
	Transmission:	High
	Path radiance:	Low: small thermal emission from ozone
	Dominant aircraft IR component:	Airframe direct emission and terrestrial illumination
	Anti-aircraft threat:	Airborne IRST. No anti-aircraft missiles

Sensors and Detection

Figure 20 shows a generalized detection problem. On the left of the diagram are the radiation sources - the sun, background, and the target of interest. In the middle is the intervening atmosphere, which attenuates the radiation as it travels to the detection system shown on the right of the diagram.

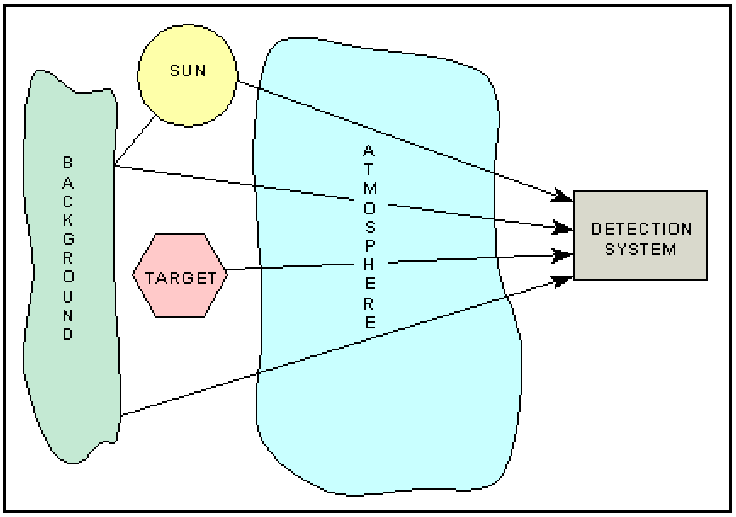

Figure 20. Generalized Detection Problem.

Figure 21 shows the basic relationships that are critical to detection of a target in the infrared. The figure is based on a generalized aircraft, but the principles apply whether the target is in fact an aircraft against a sky background or a ground vehicle being viewed from above against a terrain background. At detection, most targets are unresolved. The sensor's ability to detect the target against background in this case is driven primarily by noise equivalent irradiance.

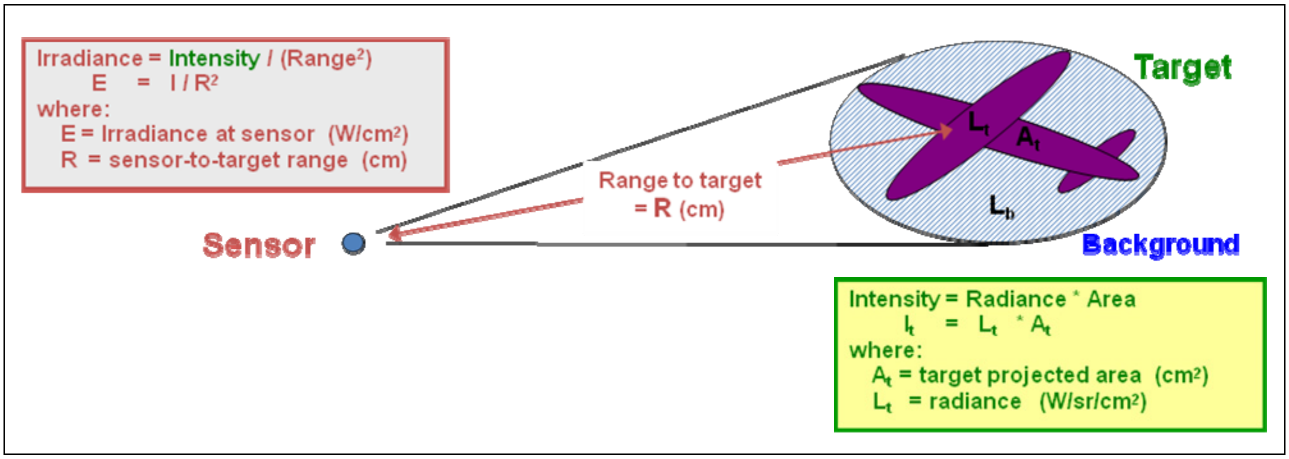
Irradiance = Intensity / (Range^2)
E = I / R^2
where:
E = Irradiance at sensor (W/cm^2)
R = sensor-to-target range (cm)
Target
L_t
A_t
L_b
Range to target
= R (cm)
Sensor
Background
Intensity = Radiance * Area
I_t = L_t * A_t
where:
A_t = target projected area (cm^2)
L_t = radiance (W/sr/cm^2)

Figure 21. Detection of a Target with a Remote Sensor.

Each of the equations shown in Figure 21 in reality has atmospheric effects and attenuation due to transmission losses and contributes to path radiance. Just as the power distribution of the target and background vary with wavelength, atmospheric effects are also spectrally selective.

Figure 22 shows the roll off of irradiance as a function of range for two different aircraft. Detection occurs at the point of intersection with the sensors noise equivalent irradiance. In the case of threat missiles, there is often a signal-to-noise threshold required for launch of the missile to ensure target quality prior to launch. The product of noise equivalent irradiance and the threshold for these systems is known typically as the minimum trackable irradiance (MTI). This is the figure used to calculate detection range for such systems.

In an effort to simplify calculations, band average atmospheric transmission values are often applied during analysis of detection scenarios. The target itself is non-uniformly distributed as a function of wavelength, and the atmospheric effect is non-uniform, so this approach is mathematically incorrect since it pulls a non-constant term out of an integral. The degree of error introduced by the band average approach depends on the spectral distribution of the source and the overall transmission of the band in question, but caution should be applied when applying band averages. All calculations involving atmospheric propagation should be done spectrally and then integrated to provide the in-band value.

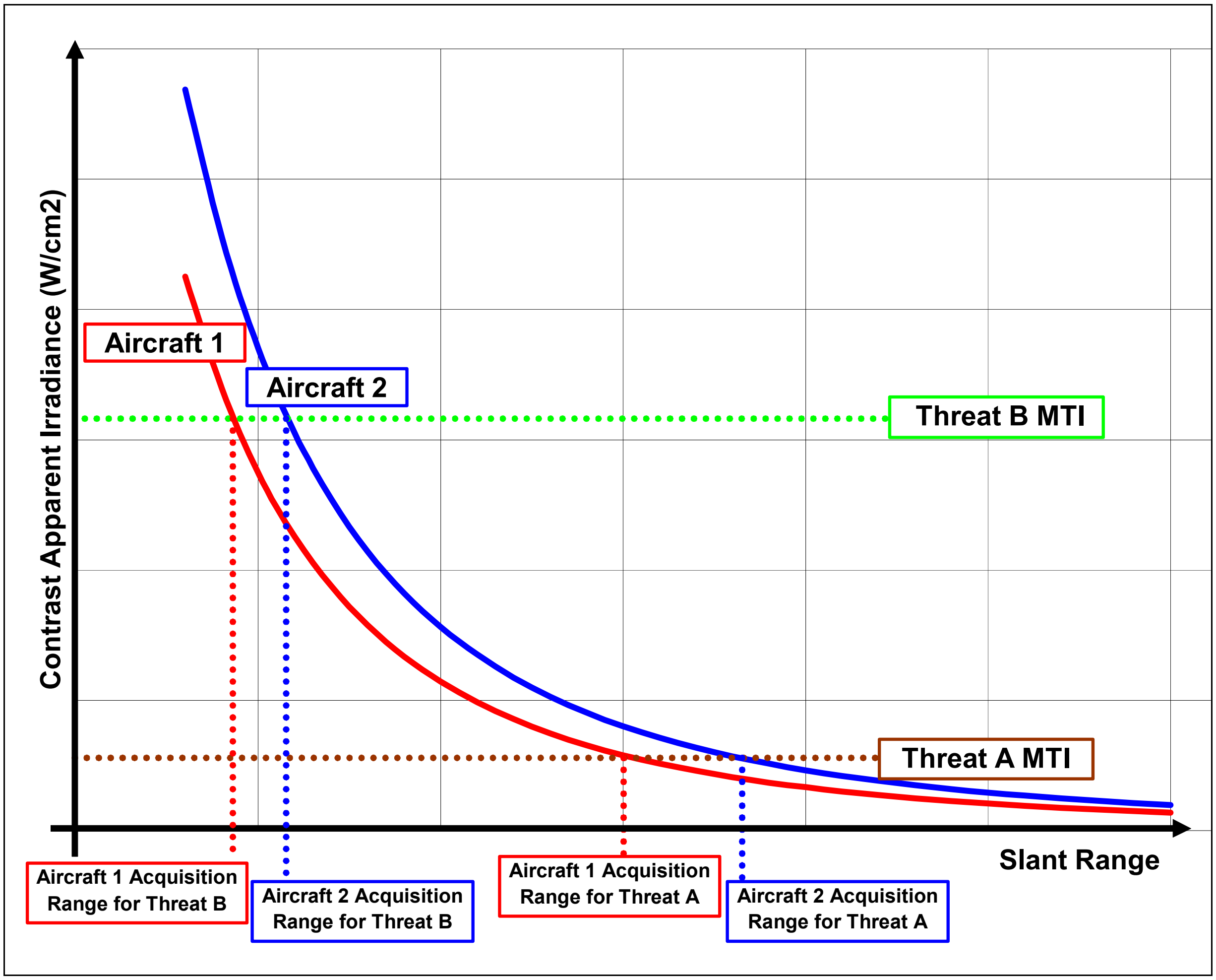

Figure 22. Detection Range Calculation.

Sensor Characterization

As described in a previous section, the output of every sensor is proportional to an integral of the received radiation weighted by the instrument's response function in that domain. Every sensor responds to radiation in accordance with the characteristics of the sensor and its components.

Characterization quantifies the sensor's response shape. Knowledge of response shape is essential to the design of a sensor and understanding its performance in with changing ambient conditions and against various types of targets in real environments.

Normalization

Calibration of a sensor, which describes its response to known input sources, and characterization of the sensor would ideally be the same process. Ideally, the absolute instrument response would be mapped over a domain with a traceable standard laboratory source that was tunable across the range of interest. In the spectral domain, for example, this would require a tunable monochromater whose output beam provided a level of spectral irradiance traceable to a radiation standard and that also had a cryogenically-cooled background source. The result would be an absolute spectral response function in units of output reading per unit radiance (or irradiance) as a function of wavelength.

In practice, it is sufficient and more practical to separate the characterization and calibration processes, so characterization determines only the relative response shape rather than absolute. Calibration then incorporates the results of the shape characterization to determine the absolute instrument response. Both, however, are important to understanding performance of a sensor.

Characterization uses a variety of different methods and sources to map relative response shapes. The response curve is then normalized and this normalized curve is used in the later derivation of the calibration coefficients. Different normalizations, such as normalization to an average value, are possible, but the convention throughout most of the measurement community today is to normalize response curves to unity at the peak.

When the contributions of all the components are combined into one curve, the result is then peak normalized and this, now unitless, curve is used in the calibration calculations.

Calibration

Sensor calibration, which is the process of relating the known input power to the output of a sensor, requires the use of standard sources, typically National Institute of Standards and Technology (NIST) traceable blackbodies and various other laboratory equipment such as collimators which make all of the rays coming from the source parallel to each other, thus representing a source at infinity. Figure 23 is a pictorial illustration of the calibration of a sensor.

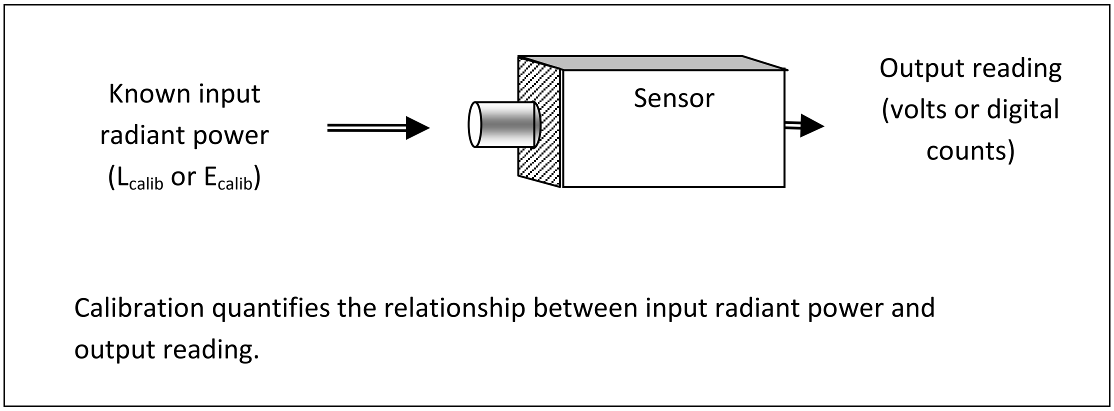
Known input radiant power (L_{calib} or E_{calib})
Sensor
Output reading (volts or digital counts)

Calibration quantifies the relationship between input radiant power and output reading.

Figure 23. Sensor Calibration Relates Input Power to Output.

Calibration of a sensor usually involves two major steps. Responsivity is the change in output of the sensor to changing input. For a sensor that responds linearly, for example, responsivity represents the slope of the curve when source radiance or incident irradiance is plotted against output voltage or counts for a digitized system. Figure 24 represents a calibration curve for a sensor that has a linear change in output over its dynamic range with changing input power. The slope of the curve is the “m” in the linear equation. Not all sensors respond linearly with power. Higher order response coefficients are common, especially for bolometers and infrared focal plane array sensors operated at short integration times. For these the process of determining the response of the instrument is the same, the curve just yields higher order terms.

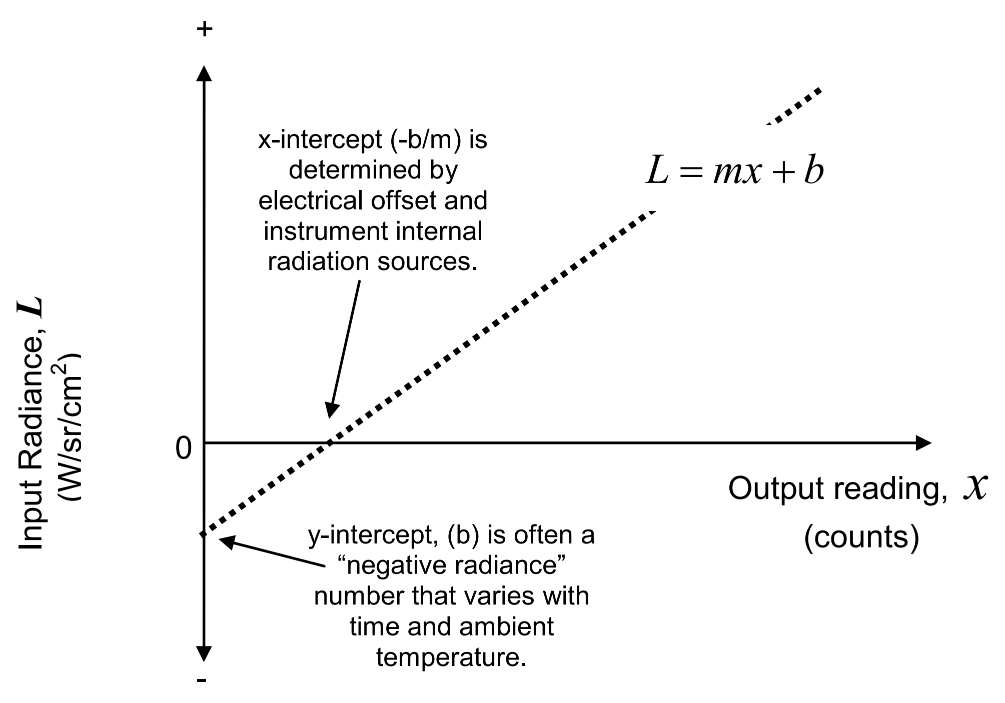

Figure 24. Calibration of a Sensor to Determine Responsivity and Offset.

For most electro-optical sensors, the responsivity does not change with ambient temperature. In other words, the non-constant terms in the calibration equation, whether it is linear or higher order, do not change with changes in temperature. Over time, however, as detectors decay, responsivity decreases. This would show itself on the Figure 24 as an increase in the slope value, i.e., higher input power is required for the same output as the detector becomes less responsive.

Offset

Offset is another important parameter for instrument calibration. For any real (non-ideal) instrument, the response curve does not pass through zero. There are several reasons for this; one being that except in a complete vacuum, zero radiance does not exist. Additionally, contributions from detector noise and radiation from the optical elements in the lens, which cause the offset to drift with changes in ambient temperature, contribute to the offset term. Some amount of offset is designed into the system as well. All electronic circuits have some amount of DC drift. To prevent clipping of the signal if this drift should go below the lower limit of the analog-to-digital converter, the "bottom end" level is adjusted up to some offset level. The consequence of this offset voltage is the addition of a y-intercept term (b), which also must be quantified by the calibration if the sensor is a laboratory instrument. For a sensor that is used for contrast detection, the intercept value is unimportant since it subtracts out in the contrast calculation.

Sensitivity

Sensitivity for a sensor is determined to a large extent by the noise level in the detector output. For focal plane array detector, pixel-to-pixel non-uniformity also limits the sensitivity of the system since detection is determined by contrast with surrounding pixels. In practice, at least for cooled infrared sensors, detection is typically limited by background and not noise limits.

Instrument Response Uniformity and Non-Uniformity Correction

In reality, all sensors exhibit non-uniform response in all of the domains referenced previously. For example, in the spatial domain, the raw output of an infrared focal plane array detector exhibits pixel-to-pixel offset differences and response differences across the field-of-view. The response change is the result of two primary factors. Since each pixel is essentially a unique detector, it exhibits unique response because of manufacturing tolerances, slight differences in crystal structure, etc. Additionally, most electro-optical sensors implement an aperture or "field stop" in the case of infrared sensors, that limits the radiation that can reach the detector outside of the sensor's desired field-of-view. Radiation entering at angles off of normal to the detector shows a cosine roll-off in incident power. The result is a reduction in responsivity for pixels that are radially separated from the center of the detector. For the majority of systems, an optical gain correction can be applied to compensate for the change in response. The typical method involves using an extended blackbody source that fills the FOV of the sensor. Reference images are collected with the source at two temperatures that are well separated across the sensor's dynamic range. This process is typically called a "2-point" correction. Actual temperature is unimportant. Slope corrections can be determined for each pixel. The result is a gain map that can be stored in the sensor electronics that can be applied to each image to correct for the non-uniform spatial response across the detector array. Pixel-to-pixel offset maps can be determined using one of the same reference images. Pixel slope and offset corrections are typically derived as normalized quantities relative to a center pixel, average of center pixels, or maximum value. The application of the correction maps to the images is commonly referred to as "non-uniformity correction." Figure 25 shows the transition from a raw image to a non-uniformity corrected image.

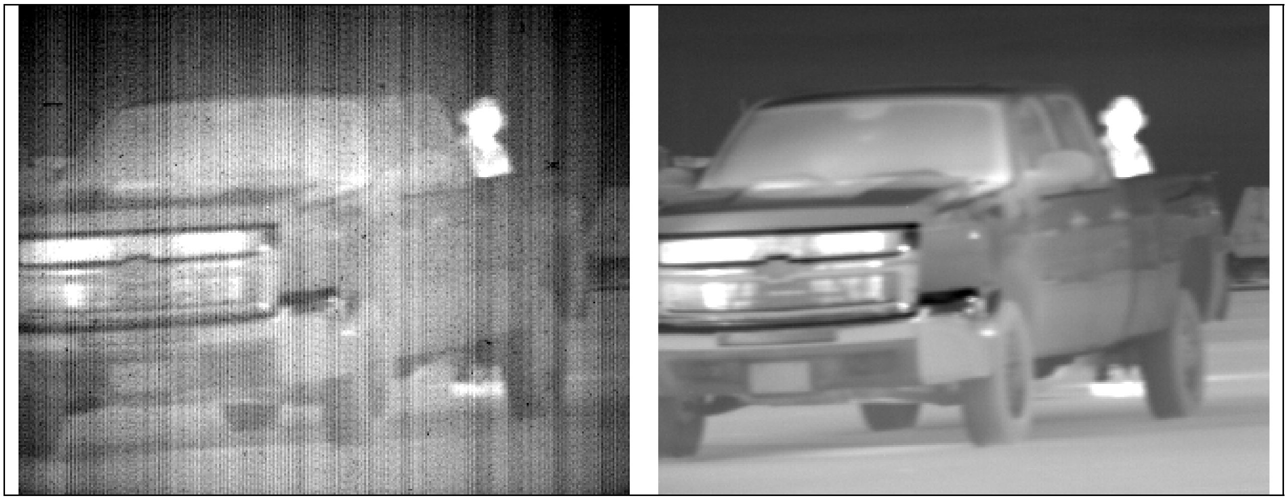

Figure 25. Non-Uniformity Correction of a Mid-IR Image.

Bad Pixel Replacement

Focal plane arrays have pixels that are either unresponsive or responsive outside of useful limits. Figure 25 shows some of the bad pixels that appear as small black spots in an image from an InSb IRFPA imager. Bad pixels can be identified during laboratory calibration or with a sensor mounted reference source. There are many approaches to replacement of bad pixels and the best approach often depends on the sensor characteristics and its application. One common approach is simply to replace the pixel with the average of its nearest neighbors.

INFRARED THREATS TO AIRCRAFT AND THEIR COUNTERMEASURES

IR guided missiles are the largest single cause for aircraft losses since the start of the 1991 Gulf war. All missiles designed within the last 20 years have counter-countermeasures circuitry. Every missile can be defeated with IR countermeasures given time to develop and test devices and techniques, but many missiles have not been exploited and the variety and complexity of the different designs present formidable challenges to the US countermeasure community.

The IR “signature” of any aircraft has three main components:

- Engine exhaust plumes
- Engine hot parts (tailpipe, etc)
- Airframe (aerodynamic heating & reflection from sun, earth, etc)

Infrared guided missiles modulate the signal produced by the aircraft in contrast with its background. Previous generations of missiles used reticles to produce signal waveforms that would provide spatial and temporal information from which signal processing could produce trackable information.

IR Missile Operation

Aircraft (or any other object) can be intercepted using several different types of guidance. The simplest type is pure pursuit, where the missile is always pointed directly at the target location. This is not aerodynamically efficient since the missile would follow a longer (curving) flight path when following a crossing target.

Most missile guidance systems are designed to lead the target so that intercept occurs at the point where the target will be at the time the missile arrives. This requires that the missile fly a course so the relative bearing to the target stays constant (constant “line of sight” angle). The LOS angle is determined by missile speed relative to the target (higher closing speed = smaller angle). The size of the angle isn’t important; only that it be constant (zero line of sight rate) as shown in Figure 26.

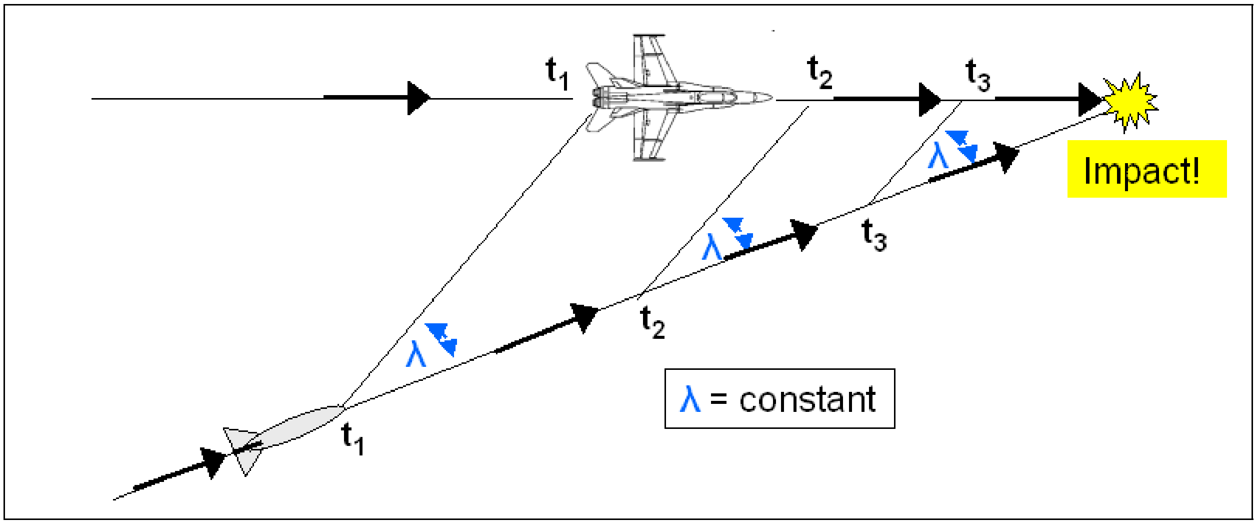

Figure 26. Missile Proportional Navigation.

This intercept course ("proportional navigation") requires that the missile have two separate servo loops: (1) a target tracker and (2) a wing control servo to control direction of flight

For a missile to guide to its intended target, it needs a tracker, which contains the following elements:

- Optics to collect and focus IR from target.
- Gimbals to allow movement to point the optics.
- Gyro stabilization to isolate optics from missile body.
- Detector to convert the received IR to electrical signal.
- Stabilization (gyro) to isolate from missile body.
- A method to determine target direction to enable closed-loop tracking.
- A method to distinguish the target from natural background.

The target tracker is the "window" into the missile's guidance through which it can be deceived by countermeasures.

The problem of determining target direction with a single detector was solved by forming an image of the target scene onto the center of a reticle disk that spun with the optics. Unlike, for example, the reticle in a rifle telescope that superimposes cross hairs, the reticle in a missile acts as a kind of shutter that blocks the passage of IR through part of the reticle and allows IR to pass through the other part. A target image falling on the opaque portion is blocked and produces no detector signal. A target image falling on the transparent portion is passed on to the detector. When the reticle is spun, IR from a target off center is alternately passed and blocked, resulting in amplitude modulation (AM). The phase of this modulation relative to a spin reference is used to tell target direction from center. A closed servo loop moves the optics to keep the target centered on the reticle. This is depicted in Figure 27.

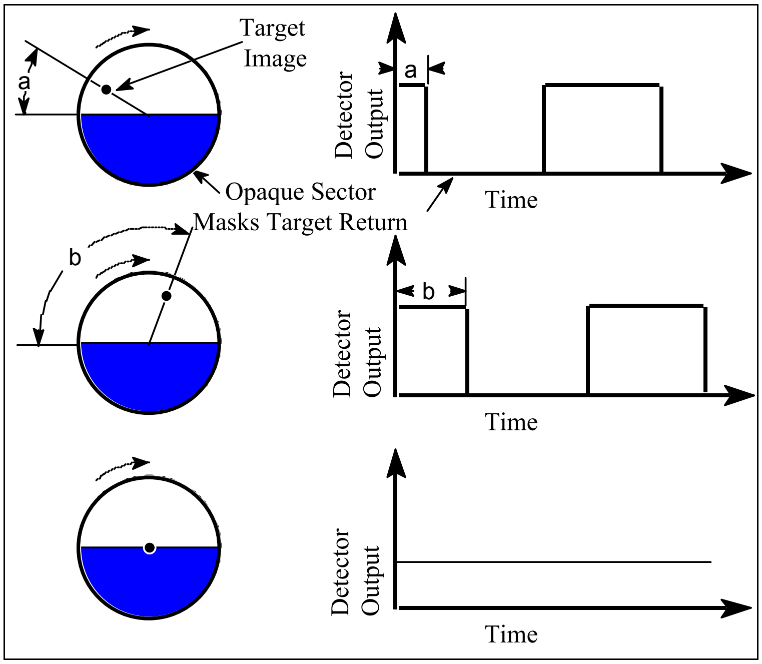
Target
Image
a
Opaque Sector
Masks Target Return
b
Detector
Output
Time
a
b

Figure 27. Basic Reticle Design.

Target trackers have another problem: The aircraft target must be distinguished from natural background sources, such as sunlit clouds and terrain. To solve this, they look for features in the spectral, spatial, and temporal domains where the target is different from background.

- Temporal: There is no difference. Neither clouds nor aircraft signature are time varying.
- Spectral: Some difference. Choice of wavelength band yields helpful differences between target and background, but this is not sufficient by itself.
- Spatial: The most viable option. Aircraft are smaller than clouds and terrain. Background radiation can be greatly reduced by spatial filtering.

If half of the reticle is made with opaque "spokes," then some irradiance from targets with small images (such as aircraft) will be modulated more completely and generate a stronger signal at a faster modulating rate than large images (clouds) as shown in Figure 28.

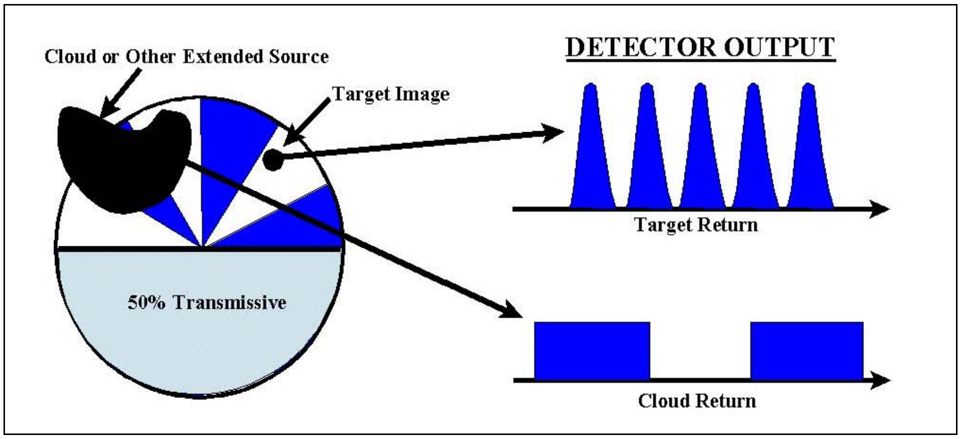
Cloud or Other Extended Source
Target Image
DETECTOR OUTPUT
50% Transmissive
Target Return
Cloud Return

Figure 28. IR Seeker Design for Background Discrimination.

The past figure and the following two figures (29 and 30) depict a spin-scan reticle used on the early Sidewinder designs. After the detector preamp, signal goes through a narrow bandpass filter to improve S/N. The AM waveform is then rectified and filtered. Target direction is determined from AM envelope phase.

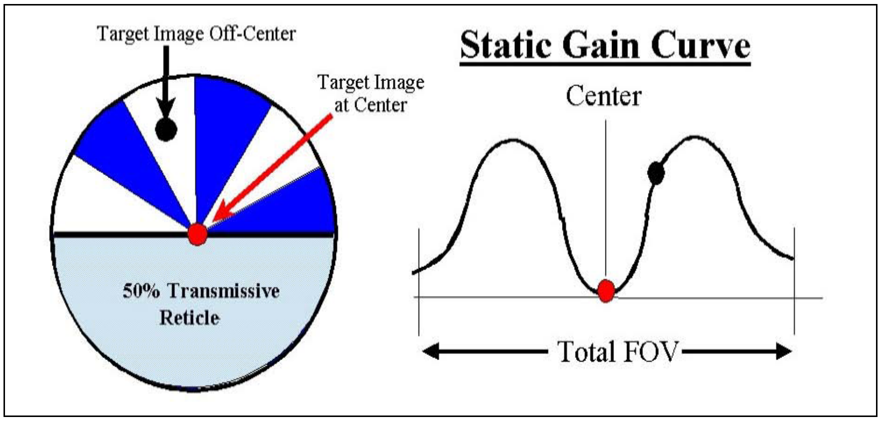
Target Image Off-Center
Target Image at Center
50% Transmissive Reticle
Static Gain Curve
Center
Total FOV

Figure 29. Spin Scan Seeker.

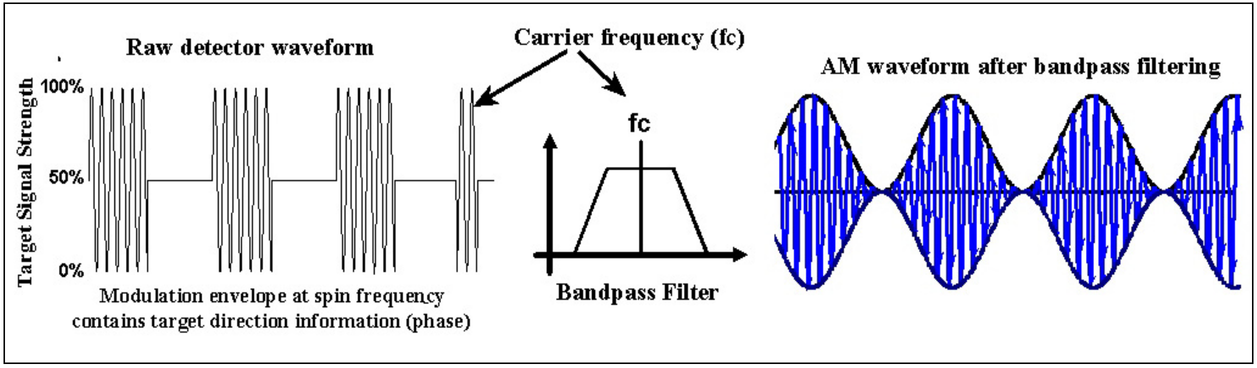
Raw detector waveform
Carrier frequency (fc)
AM waveform after bandpass filtering
Target Signal Strength
100%
50%
0%
fc
Modulation envelope at spin frequency contains target direction information (phase)
Bandpass Filter

Figure 30. Spin Scan Waveforms for Off-Center Target.

Spin-scan has the following characteristics that are important to countermeasures:

- The tracker loop drives to null the signal to zero. This occurs when the target is on the optical axis and the target image is at the center of the reticle.
- If the target is off-center, an AM carrier "error signal" is generated, where the phase of the modulation envelope indicates the target direction.

With spin scan, the missile is always looking at the target. This vulnerability to jammers led to the next evolution in target trackers: conical scan.

Conical scan borrows concept from early fire-control radars, which used a nutating feed horn. A con scan tracker is shown in Figure 31. With con scan:

- The secondary mirror of the Cassegrain is canted so the field of view seen by the detector sweeps out a pattern of overlapping circles.
- A target image at boresight falls near the edge of the reticle instead of center.
- Reticle pattern is same all the way around. (Usually tapered spokes.)

- Modulation of target near boresight is FM rather than AM. This allows tighter tracking.
- For larger angles off boresight, the target image falls outside the FOV of the detector for part of the scan. The modulation then becomes AM.

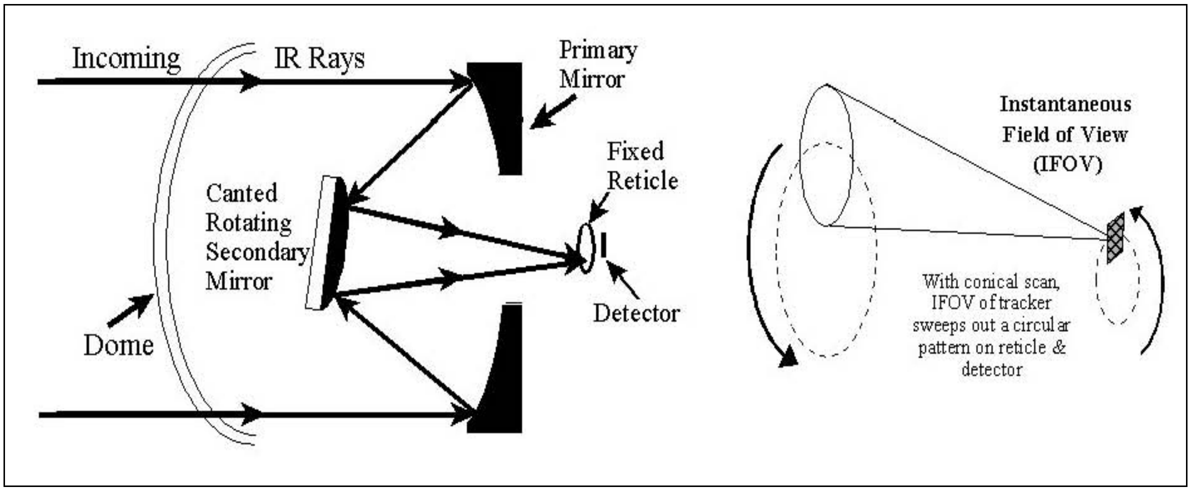
Incoming
IR Rays
Primary Mirror
Fixed Reticle
Canted Rotating Secondary Mirror
Detector
Dome
Instantaneous Field of View (IFOV)
With conical scan, IFOV of tracker sweeps out a circular pattern on reticle & detector

Figure 31. Conical Scan Tracker.

In the con scan tracker, as the missile instantaneous field of view nutates about a target on boresite, (moving through positions at t_1 through t_5 shown in Figure 32), the apparent position of a target image on the reticle sweeps out the circular pattern shown.

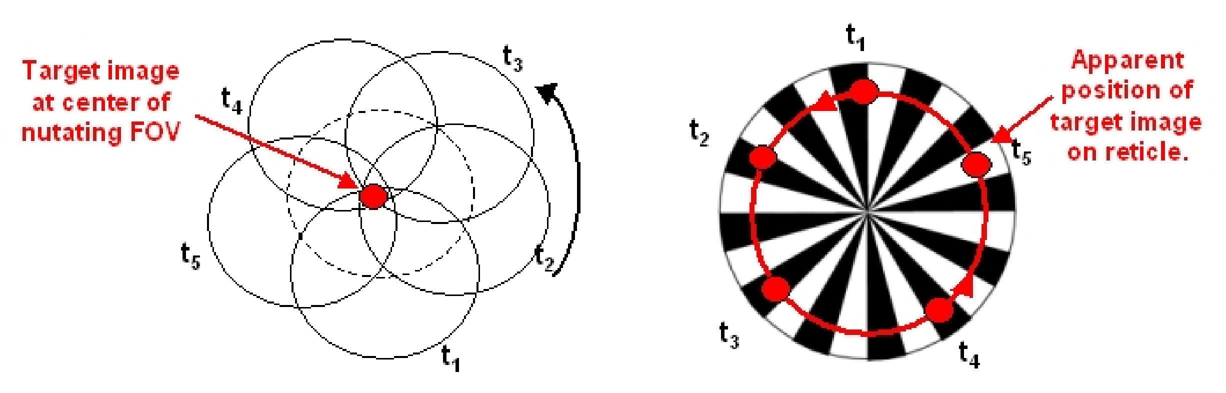
Target image at center of nutating FOV
t_1
t_2
t_3
t_4
t_5
Apparent position of target image on reticle.

Figure 32. Image on a Con Scan Reticle: Target at Boresite.

If the target is off boresight as shown in Figure 33, the detector receives a signal of varying pulse widths.

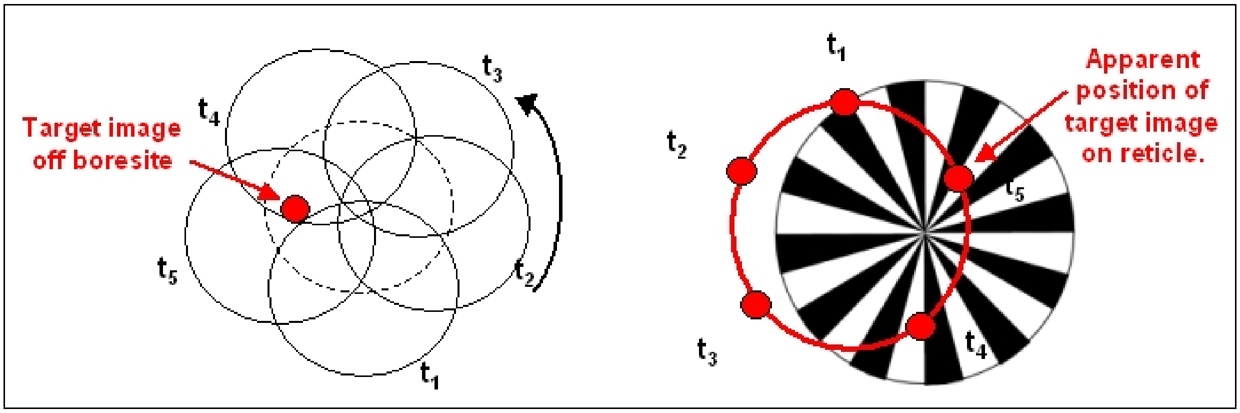
Target image off boresite
t_1
t_2
t_3
t_4
t_5
t_1
t_2
t_3
t_4
t_5
Apparent position of target image on reticle.

Figure 33. Image on a Con Scan Reticle: Target off Boresite**.**

The waveform produced by a target on boresite is a constant amplitude carrier at the reticle chopping frequency as shown in Figure 34. A target slightly off boresite produces a constant amplitude carrier that is frequency modulated at spin frequency.

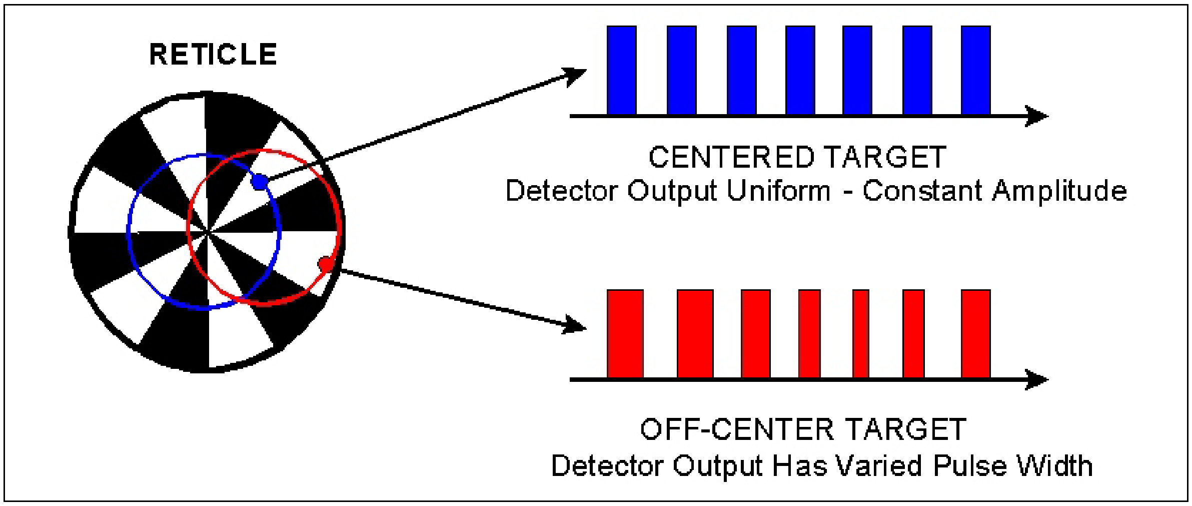
RETICLE
CENTERED TARGET
Detector Output Uniform - Constant Amplitude
OFF-CENTER TARGET
Detector Output Has Varied Pulse Width

Figure 34. Conical Scan Seeker Output.

A target further off boresite leaves the missile field of view during part of the scan, producing an amplitude-modulated waveform similar to that of a spin scan tracker. The important difference is that with a spin scan tracker, the target never leaves the missile field of view. With con scan, the target may fall outside the missile FOV at certain times during the scan. Because con-scan trackers do not necessarily view the target continuously, they can have high resistance to jammers.

Other types of seeker scan patterns now exist. The Rosette scan pattern shown in Figure 35 is one such example. It has an even higher resistance to countermeasures.

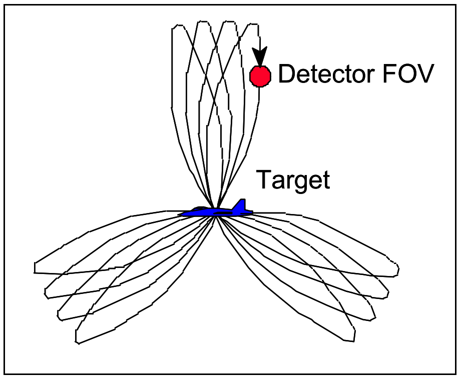

Figure 35. Rosette Scan Pattern.

Imaging arrays of detectors without reticles are newer yet. They may be classified as either staring (every pixel sees the entire scene), or scanning arrays, where the optics plays a role in determining which "pixels" are exposed to optical / IR radiation.

INFRARED COUNTERMEASURE (IRCM)

Flares

Figure 36 shows a Navy F/A-18E Super Hornet aircraft dispensing IRCM flares from its internal flare dispensers. IRCM have been the staple of countermeasures protection for military aircraft more than four decades. Flares are designed to transfer the track of an attacking infrared missile by exhibiting characteristics that confuse the tracking and guidance algorithms built into the missile. Modern missiles incorporate Counter-Countermeasures (CCM) capabilities including hardware configurations, circuitry, and logic to help identify countermeasures and reject them from processing. CCM capabilities can be based on spectral, spatial, and temporal features of the target scene. As missiles continue to improve in their sensitivity, range, maneuverability, and CCM capabilities, flares continue to evolve in order to keep pace with the evolving threat.

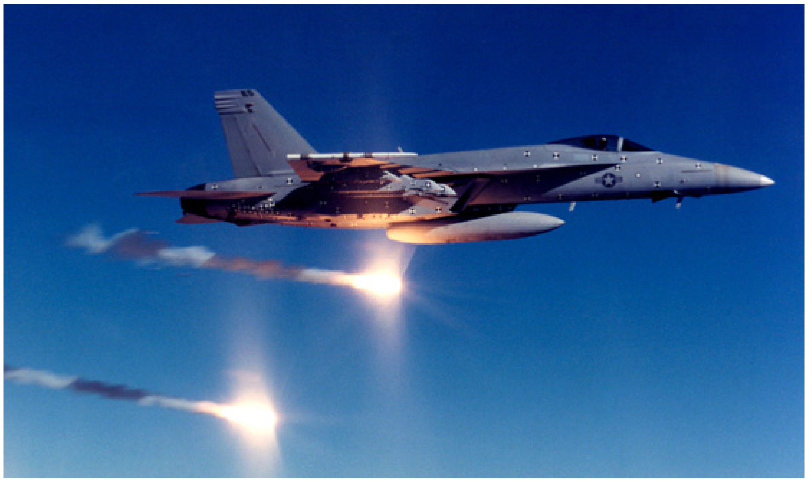

Figure 36. Navy F/A-18E Aircraft Dispensing IRCM Flares.

Over the years the Navy has fielded many flare types including the MJU-2/B, MJU-8, MJU-32, MJU-38, MJU-27, MJU-49, and many other improved versions of these flares and other types as well. IRCM flares continue to be the prevailing countermeasure for military aircraft protection by offering cost effective and robust protection.

Dispenser Systems

Most Navy fixed wing and rotary wing aircraft are equipped with countermeasures dispenser systems. These systems are critical to the survivability of the aircraft in a hostile threat environment. Modern dispensers such as the AN/ALE-47 offer high reliability and substantial programming capability that allows flare dispenses to be tailored for maximum protection of the host aircraft type. AN/ALE-47 is highly integrated into the aircraft over the 1553 data buses. The dispenser can incorporate information from several aircraft systems including missile-warning systems to improve its responses to threats and provide vital situational awareness to the aircrew.

Impulse Cartridges

IRCM flares are dispensed from the aircraft flare dispensers with electrically initiated impulse cartridges. Impulse cartridges incorporate energetic materials within a small confined canister. Upon application of a voltage to the electrical leads, a bridge wire in contact with the energetic materials burns through, igniting the propellant materials. The expanding gases push the flare from its case, held captive in the aircraft dispenser. Impulse cartridges for Navy use have been designed to withstand the extreme electro-magnetic environments encountered around aircraft carriers and other combat ships. Examples of impulse cartridges include the CCU-63 and CCU-136.

Infrared Jammers

Several lamp-based and mechanically modulated jammers have been developed over the years for protection of aircraft. Examples include the AN/ALQ-144 and AN/ALQ-157, predominantly used on helicopters and cargo aircraft. These jammers offer some level of protection over a broad field-of-regard and offer the advantage of continuous operation.

In principle, these jammers produce a modulated signal in the track band of the threat that corrupts the target tracking pulses in the missile seeker.

Passive Missile Warning Systems

Infrared-guided weapons provide passive attack capabilities against military vehicles. Unlike a radar-guided weapon that actively emits radiation and tracks the reflected pulses from the target, infrared weapons track radiation already being emitted from the target. Attacking missiles fly at very high speeds, and they are exceptionally maneuverable. Missile warning systems must be capable of detecting the threat, alerting the aircrew, and cueing a countermeasures response within sufficient time to counter the attacking missile. The time from launch to impact can be very short, making timely detection critical. Active warning systems have been developed in the past that use Doppler Radar capabilities to detect missiles, but passive missile warning systems have been preferred because of the desire to minimize emissions from the aircraft under attack.

Several passive missile-warning systems have been developed over the years for military aircraft. These systems operate in a variety of different parts of the electro-optical spectrum, but the most common are ultraviolet and infrared sensor systems. Examples of passive missile warning systems include AN/AAR-47, AN/AAR-54, AN/AAR-57, AN/AAQ-24 (both passive and active components), and the Joint and Allied Threat Awareness System (JATAS), currently under development by the Navy. Passive missile warning sensors continue to improve with advances in detector technologies, particularly with imaging detectors. These sensors provide excellent angle-of-arrival information, necessary to support cueing of laser based countermeasures, and advanced processing to detect and declare threat missiles in cluttered environments.

Laser Countermeasures

Laser-based infrared countermeasures have been in development for many years. Several systems have been fielded over the past fifteen years including the AN/AAQ-24 system on Air Force cargo aircraft and helicopters and a derivative system for Marine Corps helicopters.

Although configurations vary, most of these systems incorporate a single multi-band laser or several single-band lasers that produce modulated waveforms designed to corrupt a missile's guidance target tracking. The laser optics are located in a tracking gimbal that provides agile and rapid pointing over a broad field-of-regard. Laser based countermeasures require a relatively high angle of arrival accuracy from the host aircraft's missile warning sensor. Upon declaration of the threat, the missile warning system hands-off track to the tracking gimbal and cues the lasers to lase. A tracking camera in the tracking gimbal with high optical resolution helps to maintain track on the threat missile through the engagement period.

LASERS

The word laser comes from Light Amplification by Stimulated Emission of Radiation. A laser system emits light that is generated through a process of stimulated emission. The radiation produced by a laser exhibits high temporal and spatial coherence. In order to begin the process of stimulated emission, the lasing medium absorbs the energy from a pump source. The atoms in the lasing medium are excited to a higher energy state. These atoms will eventually return to their ground state. A large number of atoms that are excited to higher states create a population inversion. Population inversion describes the number of atoms in excited state versus the number of atoms in the ground state. In order for the atoms to return to their ground state, they must release energy. This energy is released in the form of photons. Energy of a photon is expressed as

$$E = \frac{hc}{\lambda}$$

Where

E = Energy, generally electron volts (eV)
h = Planck's constant = 4.136 x 10^{-15} (eV·s)
c = speed of light = 2.998 x 10^{8} (m/s)
λ = wavelength of light in meters

The energy that must be released by the atom to return to the ground state will direct the wavelength of the photon emitted since h and c are constants. If all the excited atoms released the same amount of energy to return to their ground state, the released photons would all have the same wavelength and would be considered fully monochromatic. Most lasers do not emit a single wavelength but a range of slightly differing wavelengths (Δλ).

The lasing medium may be a solid, a gas, liquid, or plasma. Some laser types include gas, chemical, dye, fiber-based, solid-state and semiconductor lasers. The laser radiation can be output in a continuous wave (CW) or in a pulsed wave. A continuous wave laser emits light that maintains a steady amplitude and frequency. A pulse wave will vary in amplitude and is also characterized by the systems pulse repetition frequency (PRF). The PRF is defined as the number of pulses emitted during a unit of time. Figure 37 shows the spectral output of several laser types.

The first laser was constructed by Theodore Maiman at Hughes Research Laboratories in Malibu, California. This laser was a pulsed, solid-state ruby laser. The ruby laser uses a synthetic ruby crystal as the lasing medium. A xenon flash lamp is used to excite the atoms in a ruby rod to higher energy levels. The highly polished and mirrored ends of the rod form a resonant cavity. One end of the rod has a slightly lower reflectivity. The lamp excitation produces an inverted population of excited atoms, which are stimulated to relax to lower energy levels releasing their extra energy as photons. Repeated reflections off the mirrored ends of the rod causes the photons to bounce back and forth through the rod stimulating further emissions at the same wavelength and phase producing a highly coherent beam, which finally passes through the lower reflectivity end.

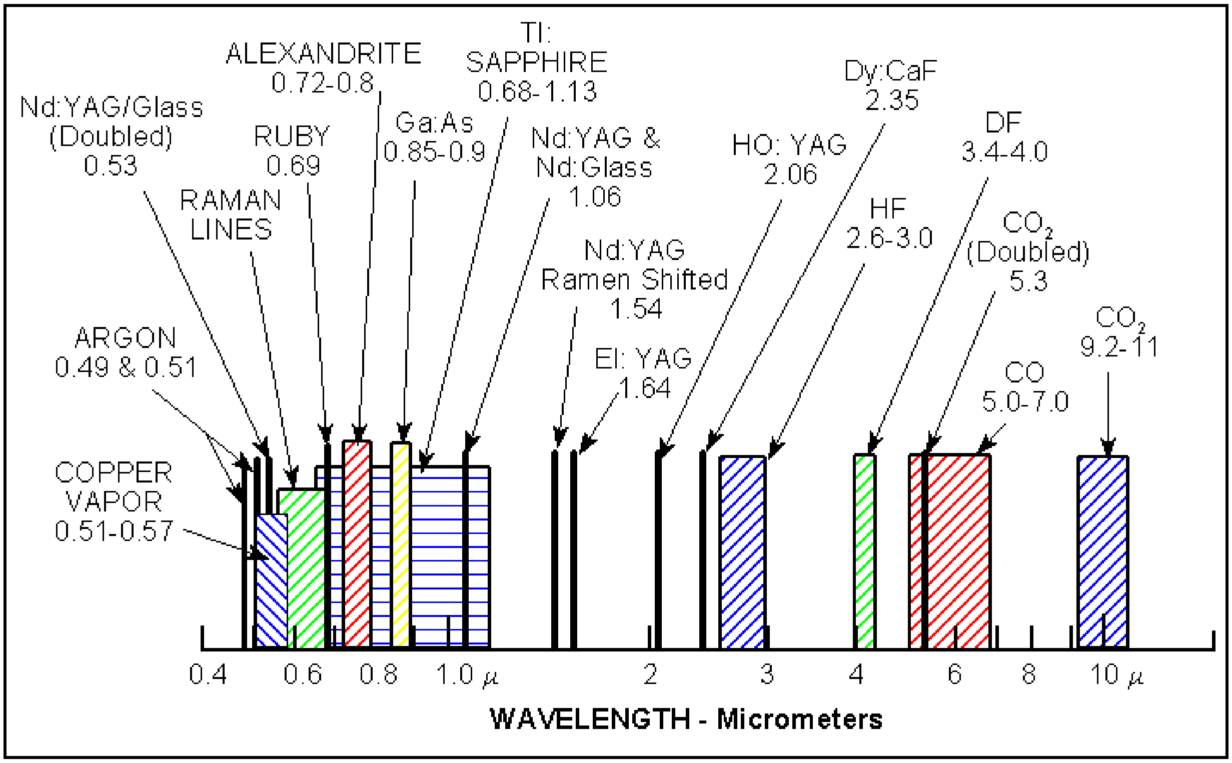

Figure 37. Spectral Lines / Ranges of Available Lasers.

The typical laser rangefinder uses a solid-state laser with a neodymium-YAG crystal lasing at 1.06 μm.

Gas lasers can be pulsed or CW. The gas dynamic laser obtains its inverted population through a rapid temperature rise produced by accelerating the gas through a supersonic nozzle. In chemical lasers the inversion is produced by a chemical reaction. In the electric discharge laser the lasing medium is electrically pumped. The gas can also be optically pumped. In an optically pumped gas laser the lasing medium is contained in a transparent cylinder. The cylinder is in a resonant cavity formed by two highly reflective mirrors.

Many gas lasers use carbon dioxide as the lasing medium (actually a mixture of CO_2 and other gases). These are the basis for most high energy or high power lasers. The first gas laser was an optically pumped CW helium-neon laser. The common laser pointer is a helium-neon laser operating at 0.6328 μm. The lasing medium is a mixture of helium and neon gas in a gas discharge or plasma tube.

The dye laser is an example of a laser using a liquid for the lasing medium. The lasing medium is an organic dye dissolved in a solvent such as ethyl alcohol. Dye lasers operate from the near UV to the near IR, are optically pumped, and are tunable over a fairly wide wavelength range.

Another type of laser is the semiconductor or injection laser, also known as a laser diode. The junctions of most semiconductor diodes will emit some radiation if the devices are forward biased. This radiation is the result of energy released when electrons and holes recombine in the junction. There are two kinds of semiconductor diode emitters: (1) the light emitting diode (LED), which produces incoherent spontaneous emission when forward biased and which has a broad (800 angstrom) spectral output, and (2) the laser diode, which maintains a coherent emission when pulsed beyond a threshold current and which has a narrow spectral width (< 10 angstrom). In the laser diode the end faces of the junction region are polished to form mirror surfaces. They can operate CW at room temperatures, but pulsed operation is more common. Figure 38 shows a typical diode laser structure.

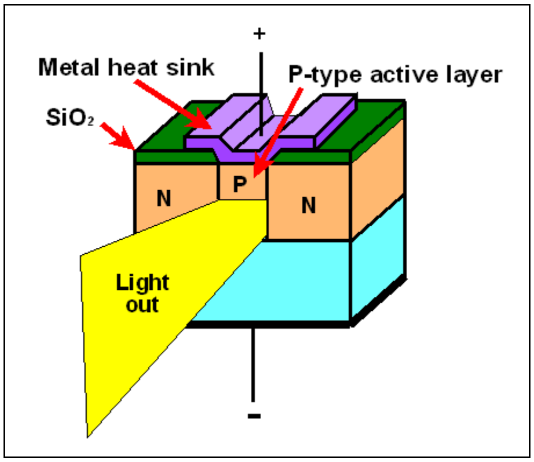

Figure 38. Diode Laser Construction.

Fiber lasers use fibers that are doped with rare-earth elements as the pumping medium. These rare-earth elements include elements such as Erbium (most common), Ytterbium, and Neodymium. There are other elements such as Thulium that are used for doping purposes. Erbium doped fiber lasers can emit in the 1.5 to 1.6 micron wavelength, which is important due to eye safety concerns in this part of the spectrum. Other wavelength emissions for Erbium include 2.7 and 0.55 microns.

Fiber based laser systems are beneficial in many ways. The fiber gain medium is compact compared to many other types of gain medium and is highly efficient. The fiber gain medium can also be physically manipulated to save space. Fiber based lasers are able to achieve high output powers. The gain medium of a fiber laser can extend for several kilometers to achieve these higher power outputs. The fact that the light is already propagating in a flexible fiber can also allow for system designs that implement a gain cavity in one location and then deliver the output in another location.

Q-switching is a common means of obtaining short intense pulses from lasers. The Q-switch inhibits lasing until a very large inverted population builds up. The switch can be active or passive. A passive Q-switch switches at a predetermined level. An active Q-switch is controlled by external timing circuits or mechanical motion. The switch is placed between the rod (or lasing medium) and the 100 percent mirror. Figure 39 shows an arrangement using a Pockels cell as an active Q-switch.

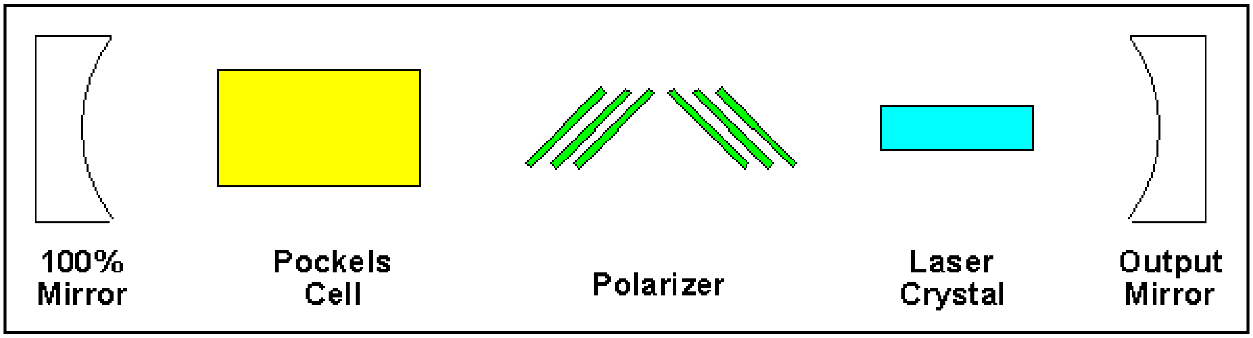

Figure 39. Q-Switch Arrangement.

Other methods of obtaining pulsed operation include using pump sources that are pulsed and mode-locking.

FIBER OPTICS

Fiber optic cables are the optical analog of RF waveguides. Fiber optic cables are made from transparent dielectrics. The fiber optic cable acts as an optical waveguide allowing light to propagate along the length of the fiber by using the principle of total internal reflection. This phenomenon can only occur under certain conditions relating to the material indices of refraction and the light ray's angle of incidence. Some benefits of fiber optic fiber include low losses, bandwidth, electromagnetic interference immunity, size, and weight.

Consider the physical construction of a bare optical fiber, depicted in Figure 40. A bare optical fiber is simply the inner glass core and the surrounding glass sleeve. The core must have a higher index of refraction than the cladding, $n_1 > n_2$. When $n_1 > n_2$, light impinging the boundary between the core and the cladding will totally internally reflect if the incident angle at each reflection is greater than the critical angle, θ_c. $\sin \theta_c = (n_2 / n_1)$

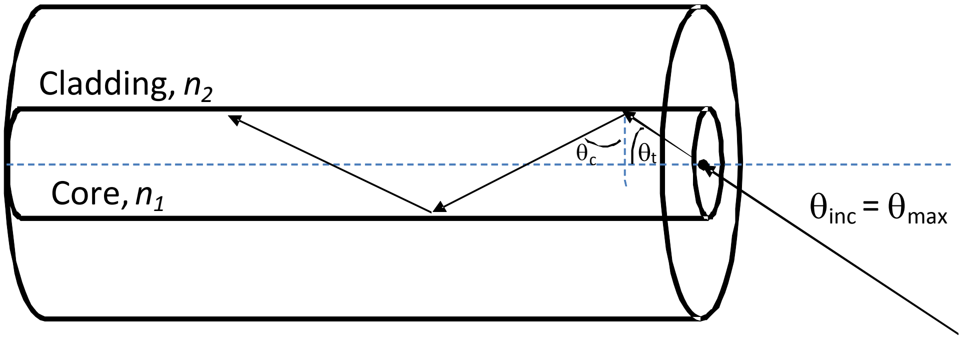

Figure 40. Bare Fiber Optic Cable.

Incident rays on the face of the fiber must intersect at angles less than θ_{max} for the internal ray, θ_t, to intersect at θ_c. When rays intersect the front face of the fiber at angles greater than θ_c, they are only partially reflected in the core and will leak out.

There are many varieties of optical fibers. Optical fibers can either be single mode or multimode. Single mode fibers are fibers which propagate a single mode down the length of the fiber while multimode fibers can propagate many modes. Single mode fibers typically have a much smaller core diameter, typically around 8 to 10 µm. Their cladding is usually 125 µm. Multimode fibers typically have core diameters around 62.5 µm with 125 µm claddings. These diameters can vary depending on the application. Loss in multimode fibers over a 1 kilometer distance is typically around 1 dB at 1310 nm. This value will vary some with changes in wavelength. Single mode fibers can maintain the quality of a light pulse over longer distances than multimode fibers due to modal dispersion effects that occur in multimode fibers. Typical losses for a single mode fiber over 1 kilometer is approximately .3 dB at 1310 nm. Again, this value will vary some with changes in wavelength. However, multimode fiber is much less expensive than single-mode and can have a lower connection loss due to the larger core diameter. Multimode fiber is commonly used in communications.

In addition to single mode or multimode, a fiber can have a step index profile or a graded index profile. Figure 41 depicts the two profiles. The step index profile maintains a uniform index of refraction within the core. A graded index profile has a peak index of refraction at the center of the core. The index of refraction value rolls off from the center to lower values closer to the cladding interface. This profile assists with the modal dispersion issue found in multimode fiber.

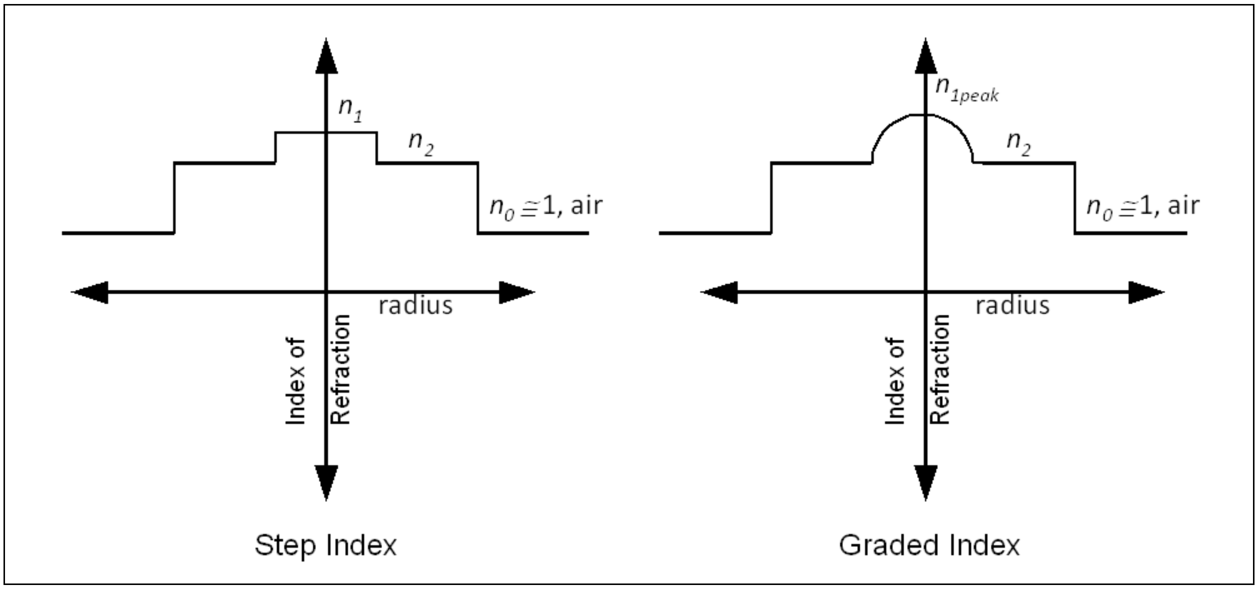

Figure 41. Fiber Profiles for a Single Mode Step Index and a Multimode Graded-Index Fiber.

Most fiber is not used in a bare form and has some additional layers of protection around the cladding. These layers can include a 250 µm buffer with a 900 µm PVC tight buffer. Some fiber will also contain aramid yarn followed by a 3 mm PVC furcation tube. Buffer tubes are often used to assist with identification and provide damage protection. The outer layers can provide additional isolation from environmental factors and lower optical crosstalk.

There are also more specialized types of fiber that include polarization maintaining and photonic crystal fibers. Polarization maintaining fibers are not constructed with a cylindrical core but instead use elliptical, bow-tie styled cores or stress rods located in the cladding (PANDA style). These are shown in Figure 42.

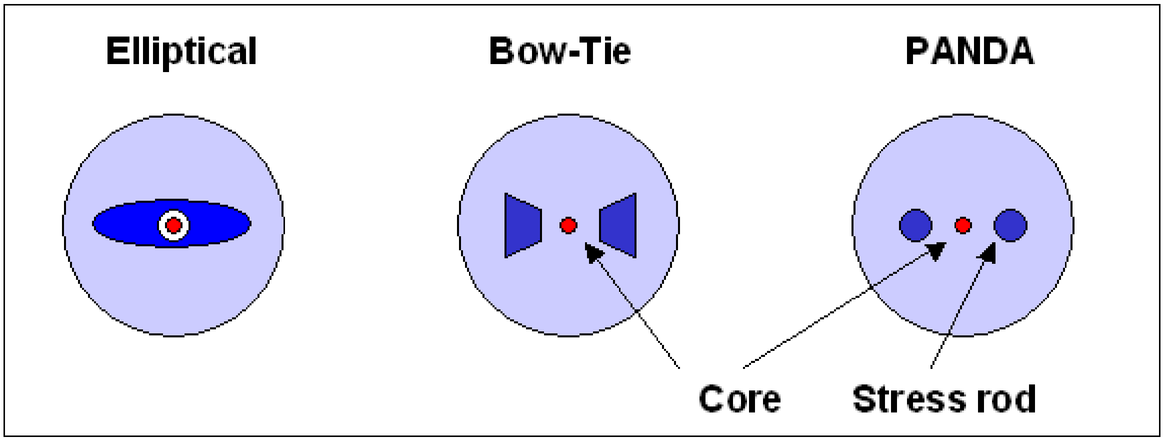

Figure 42. Polarization Maintaining Optical Fiber Types.

Polarization maintaining fibers maintain the state of the linearly polarized light propagating through the fiber. This type of fiber is used when the polarization state of the light cannot vary within a system. Photonic crystals propagate light by an arrangement of very small and closely spaced air holes that are maintained throughout the length of the fiber. Applications of photonic crystal fibers are varying and can be used in fiber lasers, amplifiers, sensors, and telecom.

It is well known that fiber optics has many communication applications; however, improvements in fiber optic technology have lent themselves to many EO applications. Many EO components are now fiber based and can interface with the tremendous advancements in fiber-based laser systems as well as other EO systems. An example of the use of fiber optics in an EW system is the AN/ALE-50 and 55 Fiber-Optic Towed Decoy (FOTD). The FOTD uses fiber optic cabling to communicate with the jammer.

This page intentionally left blank.

LASER SAFETY

Lasers are divided into the following classes:

Class 1	Low power / non-hazardous
Class 2/2a	Low power / minor controls necessary Emit less than 1 mW visible CW radiation. Not considered hazardous for momentary (<0.25 sec) unintentional exposure. Class 2a lasers are those class 2 lasers not intended to be viewed, i.e. supermarket scanners.
Class 3a/3b	Medium power / direct viewing hazard / little diffuse reflection hazard.
Class 3a	Visible lasers with 1-5 mW power output, invisible lasers, and those having 1-5 times the Accessible Emission Limit (AEL) of class 1 lasers.
Class 3b	All other class 3 lasers at all wavelengths which have a power output less than 500 mW.
Class 4	High power / eye & skin hazard / potential diffuse reflection hazard or fire hazard

There are several pertinent instructions and guidelines regarding laser use. They are:

- OPNAVINST 5100.27B Navy Laser Hazards Control Program
  (which replaced OPNAVINST 5100.27A and SPAWARINST 5100.12B)
- MIL-HDBK-828B, Range Laser Safety
- ANSI Z136.1-2007, American National Standard for the Safe Use of Lasers (Parent)

Every Navy command which uses lasers must have a Laser System Safety Officer (LSSO). All LSSOs must attend a Navy LSSO course.

There are four categories of LSSOs.

- Administrative Laser Safety Officer (ALSO)
- Technical Laser Safety Officer (TLSO)
- Laser Safety Specialist (LSS)
- Range Laser Safety Specialist (RLSS)

See OPNAVINST 5100.27B for details of their qualifications and responsibilities.

The hazard ranges of interest are the NOHD for direct viewing of a beam and the $r_{1(safe)}$ or $r_{2(safe)}$ for viewing a beam reflected off an object such as a wall. These are depicted in Figure 1. The Maximum Permissible Exposure (MPE) values present laser safety levels as a function of exposure time, laser PRF, pulse duration, and wavelength. Different tables are used for eye safety while directly viewing a beam, for viewing a diffusely reflected beam, and for skin exposure.

For repeated pulses the following equation is used to calculate the maximum permissible exposure (MPE).

$$\text{MPE (repeated pulse)} = \frac{MPE(single\ pulse)}{(PRF \ x\ t_e)^{1/4}} \qquad \textbf{[1]}$$

Where PRF is the pulse repetition frequency of the laser and t_e is the exposure duration.

For visible lasers t_e is usually taken as 1/4 second and for non-visible lasers a value of 10 seconds is used.

Figure 1 depicts some laser hazard distances.

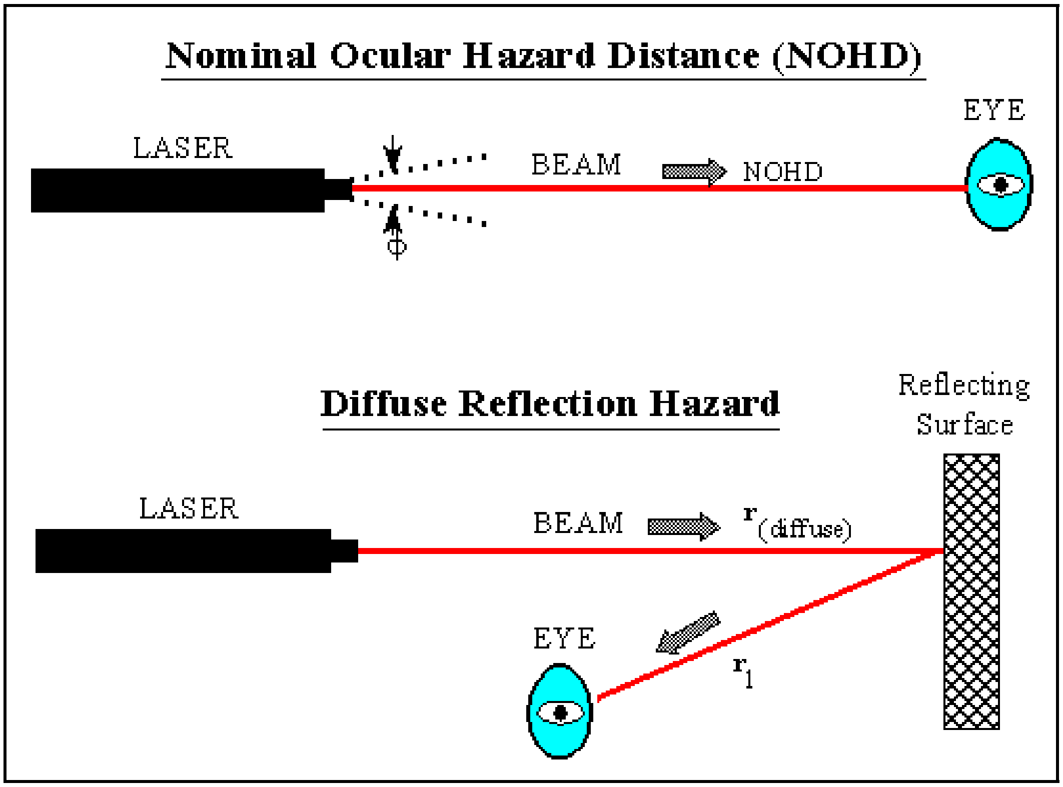
Nominal Ocular Hazard Distance (NOHD)
EYE
LASER
BEAM
NOHD
Φ
Diffuse Reflection Hazard
Reflecting Surface
LASER
BEAM
r(diffuse)
EYE
r1

Figure 1. Laser Hazard Distances.

Range laser safety specialists shall be designated for external operations. Range test plans shall specify:

- Permissible aircraft flight paths, and ship or vehicle headings.
- Hazard areas to be cleared.
- Operational personnel locations.
- Types of surveillance to be used to ensure a clear range.
- Radio / communications procedures.

During laser operations no portion of the laser beam may extend beyond the controlled target area unless adequate surveillance can prevent radiation of unprotected areas. Class 3 and class 4 lasers shall not be directed above the horizon unless coordinated with those responsible for the given airspace (FAA, Navy, Air Force, etc).

In an industrial environment, warning and hazard signs and lights will be posted, a hazard zone shall be designated when lasers are in operation, and training shall be provided to operators in the proper eye and body (skin) protection required. Interlocks to laser operation shall be provided when there is the possibility of unauthorized personnel entering the hazard area.

Fiber optic cables usually have laser power sources so appropriate warnings or labels need to be applied to connections or possible breakage points.

This page intentionally left blank.

AIRCRAFT DYNAMICS CONSIDERATIONS

Free Fall / Aircraft Drag........8-1
Mach Number and Airspeed vs. Altitude........8-2
Maneuverability 8-3
EMP / Aircraft Dimensions........8-4

This page intentionally left blank.

FREE FALL / AIRCRAFT DRAG

The purpose of this section is to get an awareness of the distance traveled by a flare or other object such as a bomb, which is jettisoned or dropped by an aircraft. This will give the reader an appreciation for the significance of aircraft tactical altitude.

From Newton's second law of motion:

$$F = m_o a$$

where:
F = Force
m_o = Mass of object
a = Acceleration

and the law of gravitation: $$F = G\frac{m_o m_e}{r^2}$$

where:		English Units	SI Units
	F = Force of attraction	lb_f	Newton
	G = universal gravitational constant	$3.44\text{x}10^{-8}$ $ft^4/lb\text{-}sec^4$	$6.67\text{x}10^{-11}$ $m^3/kg\text{-}sec^2$
	m_o, m_e= Masses (not weight) of object & earth	slug	kg
	r = distance between center of gravity of objects	feet	meter

Combining the two equations and solving for "a":

$$a = \frac{G\,m_e}{r^2} = g$$, the familiar constant acceleration due to gravity.

Since G and m_e are fixed and the variation in r (the distance from the earth's center) is small except for satellites, "g" is considered fixed at 32.2 ft/sec^2.

For objects with a constant acceleration (g), it can be shown that:

$$d = v_i t + \frac{1}{2} g t^2$$

where
d = distance traveled
v_i = Initial velocity
t = time
g = acceleration

For a falling object, Figure 1 on the following page may be used to estimate time/distance values.

- The upper curve is for an object shot upward with an initial velocity of 50 ft/sec.

- The middle curve is for an object shot horizontally with an initial velocity of 50 ft/sec or one that is a free-falling object dropped with no initial vertical velocity.

- The lower curve is for an object with a downward initial velocity of 50 ft/sec.

Notes:

1) 50 ft/sec is the typical cartridge ejection velocity of a flare/chaff expendable.
2) The top curve actually goes up 39 feet before starting back down, but this is difficult to see due to the graph scale.
3) This simplification ignores the effects of air drag or tumbling effects on a falling object which will result in a maximum terminal velocity, with resultant curve straightening.

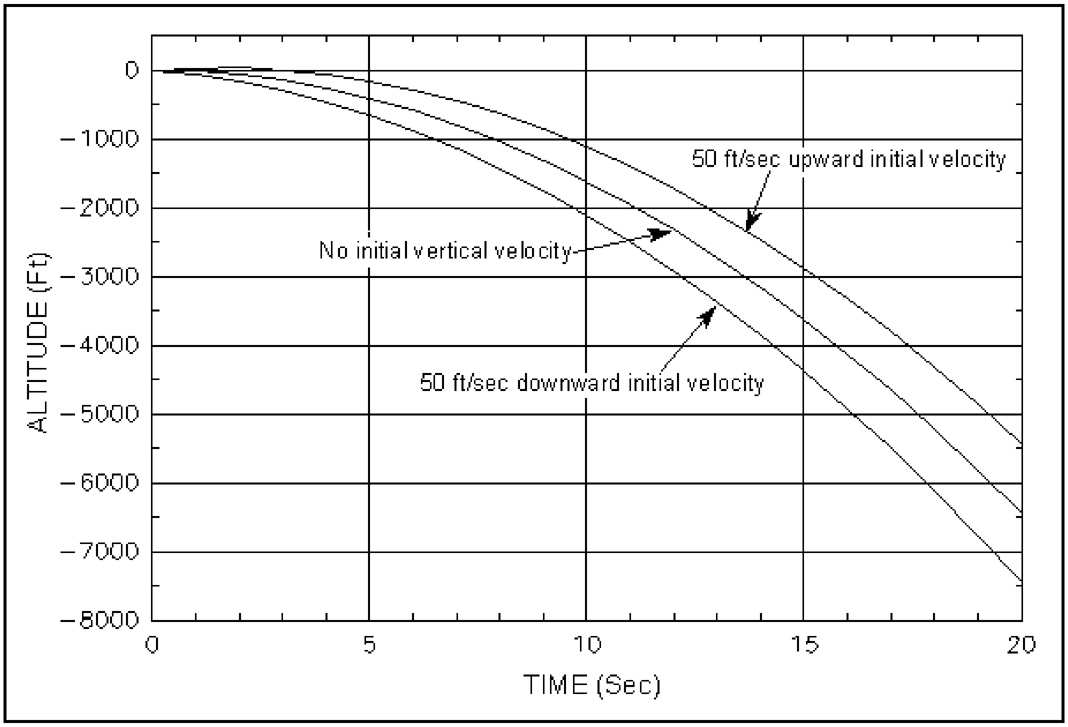

Figure 1. Object Fall Rate.

SAMPLE CALCULATIONS

Let us assume that we want to know how far a bomb or other object has fallen after 13 seconds if it had been dropped from an aircraft traveling at 450 kts which was in a 40° dive.

Our initial vertical velocity is: 450 kts (Sin 40°) (1.69 ft/sec per knot) = 489 ft/sec downward

$d = V_i t + \frac{1}{2}gt^2 = -489(13) + \frac{1}{2}(-32.2)(13)^2 = -6355 - 2721 = -9{,}076$ ft.

Remember to keep the signs (+/-) of your calculations in agreement with whatever convention you are using. Gravity pulls downward, so we used a minus sign for acceleration. Also the initial velocity was downward.

In reality, any object may well have reached terminal velocity before the time indicated using the above formula or Figure 1. In this example, the actual distance determined from ballistics tables would have been 8,000 ft, which is about 13% less than the above calculation would indicate. The drag characteristics of the object determine how much shorter the distance will be. In any case, it will not have dropped farther.

AIRCRAFT DRAG INDEX POINTS

Tactical aircraft carry stores in various combinations depending upon the mission. Each store has a different drag load which affects range. The pilot needs to know the total drag load in order to determine his aircraft range on a particular mission. Adding up the total drag in pounds of force for wind resistance would be cumbersome. Therefore, the drag of the stores is compared to a known reference drag (usually the aircraft), and expressed as a percentage of aircraft drag multiplied by some constant. This ratio is variously called drag count, drag index, or drag points. For instance, if a missile has 100 pounds of drag and the reference aircraft drag is 50,000 pounds, the ratio is 100/50,000 = 0.002. Multiply this by a constant of 100 (for example) and the drag index point is 0.2. The pilot only needs to look on a chart to see what the drag index points are for his stores, add up the drag points, and look on a chart to see what his aircraft range and best range (or endurance) speed will be.

This page intentionally left blank.

MACH NUMBER and AIRSPEED vs. ALTITUDE

MACH NUMBER is defined as a speed ratio, referenced to the speed of sound, i.e.

$$MACH\ NUMBER = \frac{Velocity\ of\ Interest}{Velocity\ of\ Sound} \quad (at\ the\ given\ atmospheric\ conditions) \qquad \textbf{[1]}$$

Since the temperature and density of air decreases with altitude, so does the speed of sound, hence a given true velocity results in a higher MACH number at higher altitudes.

AIRSPEED is a term that can be easily confused. The unqualified term airspeed can mean any of the following:

a. Indicated airspeed (IAS) - the airspeed shown by an airspeed indicator in an aircraft. Indicated airspeed is expressed in knots and is abbreviated KIAS.

b. Calibrated airspeed (CAS) - indicated airspeed corrected for static source error due to location of pickup sensor on aircraft. Calibrated airspeed is expressed in knots and is abbreviated KCAS. Normally it doesn't differ much from IAS.

c. True airspeed (TAS) - IAS corrected for instrument installation error, compressibility error, and errors due to variations from standard air density. TAS is expressed in knots and is abbreviated KTAS. TAS is approximately equal to CAS at sea level but increases relative to CAS as altitude increases. At 35,000 ft, 250 KIAS (or KCAS) is approximately 430 KTAS.

IAS (or CAS) is important in that aircraft dynamics (such as stall speed) responds largely to this quantity. TAS is important for use in navigation (True airspeed ± wind speed = ground speed).

Figures 1 and 2 depict relations between CAS and TAS for various altitudes and non-standard temperature conditions. The first graph depicts lower speed conditions, the second depicts higher speeds.

As an example of use, consider the chart on the next page. Assume we are in the cockpit, have read our IAS from the airspeed indicator, and have applied the aircraft specific airspeed correction to obtain 370 KCAS. We start at point “A” and go horizontally to our flight altitude at point “B” (25,000 ft in this case). To find our Mach, we go down vertically to point “C” to obtain 0.86 Mach. To get our TAS at our actual environmental conditions, we go from point “B” vertically until we hit the Sea Level (S.L.) reference line at point “D”, then travel horizontally until we reach our actual outside air temperature (-20°C at altitude) at point “E”, then go up vertically to read our actual TAS from the scale at point “F” (535 KTAS). If we wanted our TAS at “standard” temperature and pressure conditions, we would follow the dashed lines slanting upward from point “B” to point “G” and read 515 KTAS from the scale. Naturally, we could go into the graph at any point and go “backwards” to find CAS from true Mach or TAS.

Figure 3 shows a much wider range of Mach numbers. It contains only TAS and Mach, since aircraft generally do not fly above Mach 2, but missiles (which don't have airspeed indicators) do. The data on this graph can be obtained directly from the following formula for use at altitudes of 36,000 ft and below:

$$Speed\ of\ Sound\ (KTAS) = 29.06\sqrt{518.7 - 3.57\ A} \qquad Where\ A = altitude\ (K\ ft) \qquad \textbf{[2]}$$

The speed of sound calculated from this formula can be used with the equation on the first page to obtain Mach number. This equation uses the standard sea level temperature of 59° F and a lapse rate of -3.57°/1000 ft altitude. Temperature stabilizes at -69.7° F at 36,000 ft so the speed of sound stabilizes there at 573 knots. See the last page of this section for a derivation of equation [2].

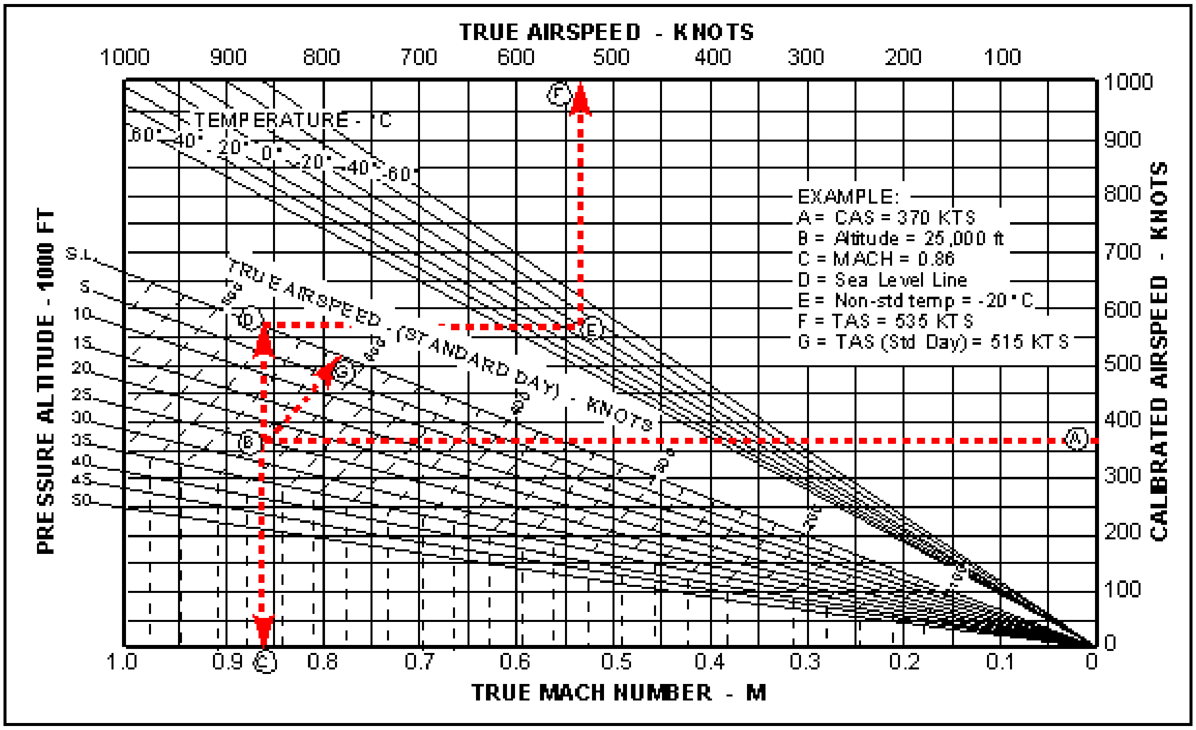

Figure 1. TAS and CAS Relationship With Varying Altitude and Temperature.

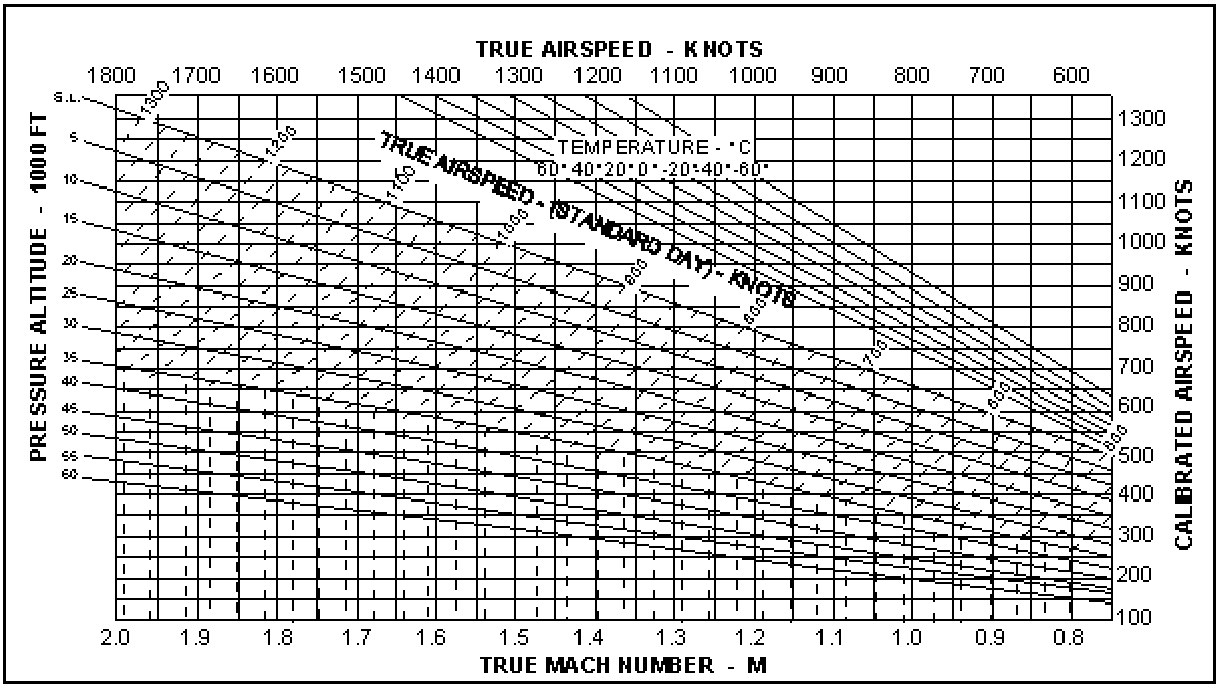

Figure 2. TAS and CAS Relationship With Varying Altitude and Temperature (Continued).

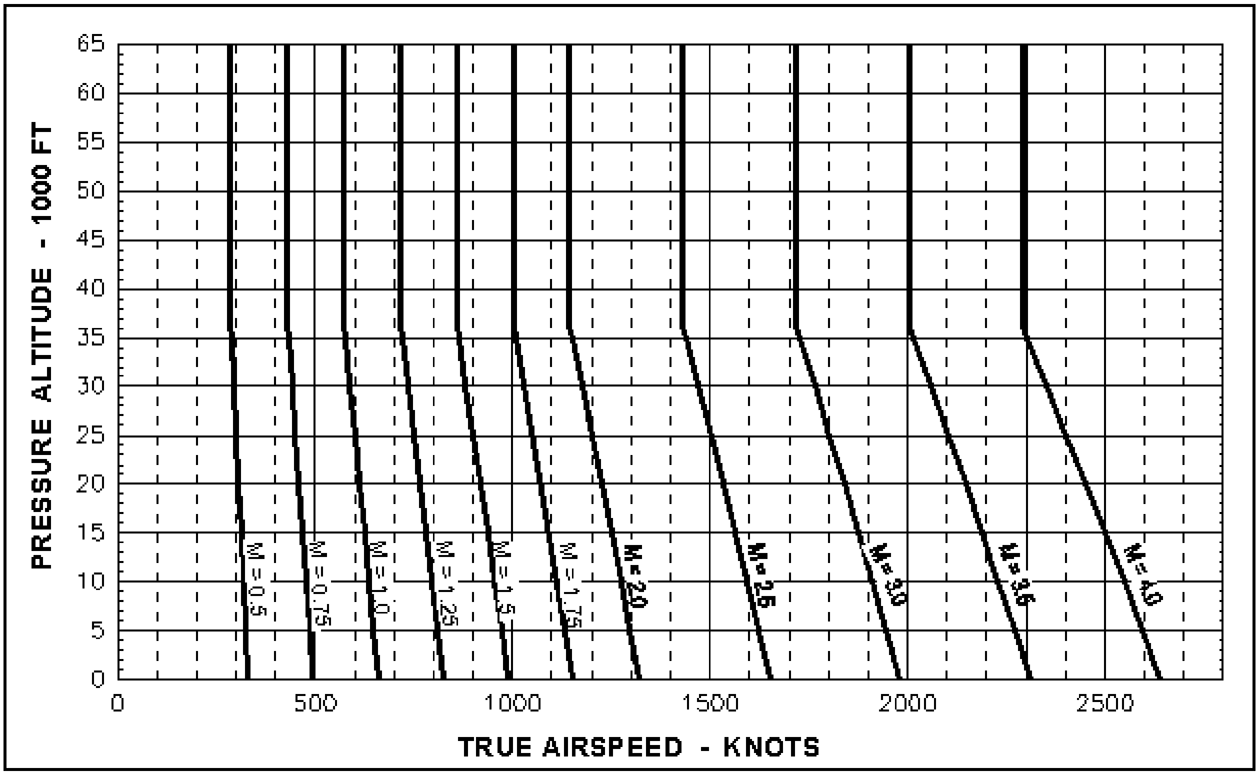

Figure 3. Mach Number vs. TAS Variation With Altitude.

The following is a derivation of equation [2] for the speed of sound:

Given: p = pressure (lb/ft^2)
v = specific volume (ft^3/lb)
R = a constant (for air: R = 53.3)
T = absolute temperature (°Rankine) = °F + 459.7
w = specific weight (lb/ft^3) = $1/v$
ρ = density = $w/g = 1/gv \;\therefore\; v = 1/g\rho$

From Boyle's law of gasses: $pv = RT$, therefore we have: $p/\rho = gRT = (32.2)(53.3)T = 1718\ T$ **[3]**

It can also be shown that: p/ρ^γ = constant; for air $\gamma = 1.4$ **[4]**

From the continuity equation applied to a sound wave: $\rho AV_a = (\rho + dp)A(V_a + dV_a)$ **[5]**

Expanding and dropping insignificant terms gives: $dV_a = -V_a\, d\rho/\rho$ **[6]**

Using Newton's second law ($p + \rho V_a/2$ = a constant) and taking derivatives: $dp = -\rho V_a dV_a$

Substituting into [6] gives: $V_a^2 = dp/d\rho$ **[7]**

Then taking derivatives of [4] and substituting in [7] gives: $V_a = \sqrt{\dfrac{\gamma p}{\rho}}$ **[8]**

Then using [3] gives: $V_a = \sqrt{\gamma gRT} = \sqrt{1.4(1718)T} = 49\sqrt{T}$ **[9]**

Using a “Standard” atmosphere of 59° F @ Sea Level (S.L.) and a lapse rate of -3.57°/1000 ft altitude:

$$V_a = 49\sqrt{459.7 + 59 - 3.57\,A}\ \frac{ft}{\text{sec}}\left[\frac{3600\,\text{sec}}{hr}\frac{nm}{6076\ ft}\right] = 29.06\sqrt{518.7 - 3.57A} \quad \text{which is equation [2]}$$

MANEUVERABILITY

A useful function is to determine how many "G's" an aircraft might require to make a given turn without altitude loss. From Newton's laws, $F \cos \varphi = W$, where: F = force applied to an aircraft, W = weight, and φ = bank angle. By definition "G's" is the ratio of the force on an object to its weight, i.e., $G = F/W = 1/\cos \varphi$.

Simple calculations will show the results presented in Table 1, to the right.

Given that the average structural limit of an aircraft is about 7 G's, the maximum bank angle that can be achieved in level (non-descending) flight is 81.8°.

Table 1. G vs. Angle of Bank.
(No altitude loss)

G	φ
1.0	0
1.4	45
2.0	60
3.9	75
7.2	82
11.5	85

Figure 1 can be used to determine the turn radius and rate-of-turn for any aircraft, given speed and angle of bank (assuming the aircraft maintains level flight). It may also be used in the reverse context. It should be noted that not all aircraft can fly at the speeds depicted - they may stall beforehand or may be incapable of attaining such speeds due to power/structural limitations.

In the example shown on Figure 1, we assume an aircraft is traveling at 300 kts, and decides to make a 30° angle of bank turn. We wonder what his turn radius is so we can approximate his flight path over the ground, and what his rate of turn will be. We enter the chart at the side at 300 kts and follow the line horizontally until we intercept the 30° "bank angle for rate of turn" line. We then go down vertically to determine the 2.10°/sec rate of turn. To get radius, we continue horizontally to the 30° "bank angle for turn radius" line. We can then go down vertically to determine the radius of 13,800 ft.

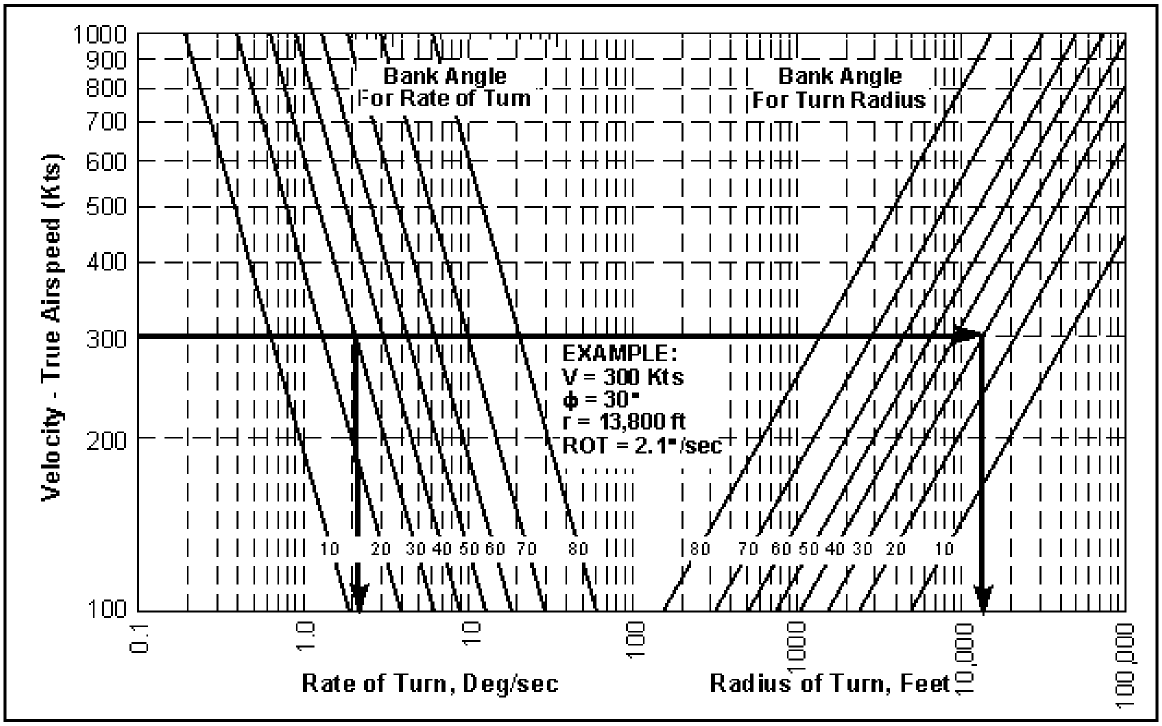

Figure 1. Aircraft Turn Rate / Radius vs. Speed.

The exact formulas to use are:

$$Rate\ of\ Turn = \frac{1091\tan(\phi)}{V} \qquad Radius\ of\ Turn = \frac{V^2}{11.26\tan(\phi)} \qquad Where: \begin{array}{c} V = Velocity\ (Knots) \\ and\ \phi = Angle\ of\ Bank \end{array}$$

Another interesting piece of information might be to determine the distance a typical aircraft might travel during a maneuver to avoid a missile.

Figure 2 shows a birds-eye view of such a typical aircraft in a level (constant altitude) turn.

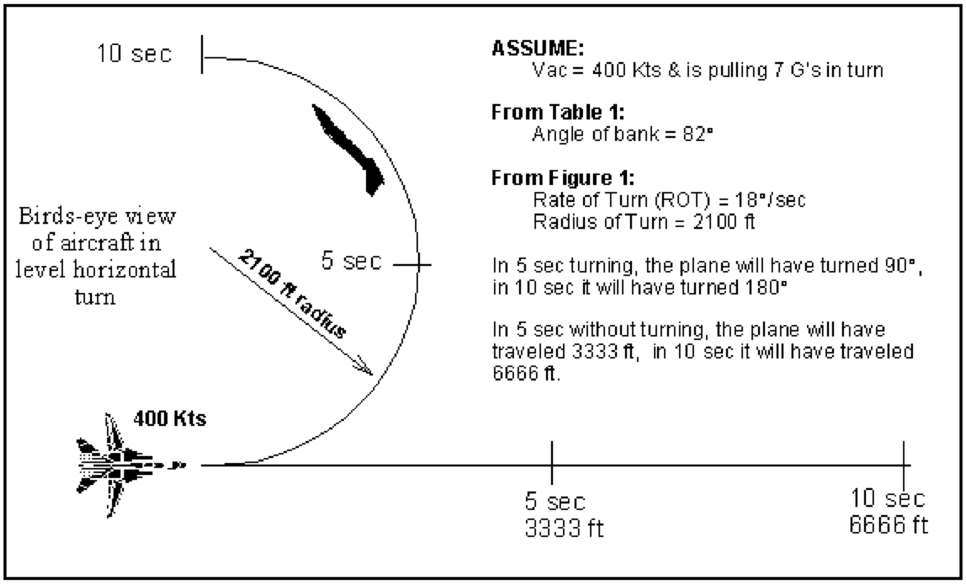

ASSUME:
Vac = 400 Kts & is pulling 7 G's in turn

From Table 1:
Angle of bank = 82°

From Figure 1:
Rate of Turn (ROT) = 18°/sec
Radius of Turn = 2100 ft

In 5 sec turning, the plane will have turned 90°, in 10 sec it will have turned 180°

In 5 sec without turning, the plane will have traveled 3333 ft, in 10 sec it will have traveled 6666 ft.

Figure 2. Maneuvering Aircraft.

To counter many air-to-air missiles the pilot might make a level turn, however in countering a SAM, altitude is usually lost for two reasons: (1) the direction of maneuvering against the missile may be downward, and (2) many aircraft are unable to maintain altitude without also losing speed. These aircraft may have insufficient thrust for their given weight or may be at too high an altitude. The lighter an aircraft is (after dropping bombs/burning fuel), the better the performance. Likewise, the higher the altitude, the poorer the thrust-to-weight ratio. Maximum afterburner is frequently required to maintain altitude at maximum "G" level.

REFERENCE AXES (Roll, Pitch, Yaw)

The rotational or oscillating movement of an aircraft, missile, or other object about a longitudinal axis is called roll, about a lateral axis is called pitch, and about a vertical axis is called yaw as shown in Figure 3.

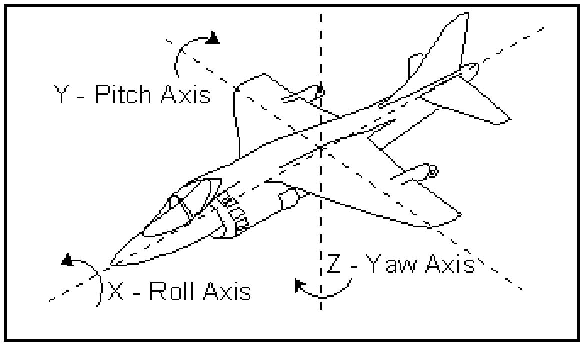

Figure 3. Reference Axes.

SAMPLE CALCULATIONS

If we want to determine the rate of turn or turn radius more precisely than can be interpolated from the chart in Figure 1, we use the formulas. For our initial sample problem with an aircraft traveling 300 kts, in a 30° angle of bank turn, we have:

$$Rate\ of\ Turn = \frac{1091\tan(\phi)}{V} = \frac{1091\tan(30)}{300} = 2.1^\circ/\sec$$

$$Radius\ of\ Turn = \frac{V^2}{11.26\tan(\phi)} = \frac{300^2}{11.26\tan(30)} = 13{,}844\ ft$$

These are the same results as we determined using Figure 1.

This page intentionally left blank.

EMP / AIRCRAFT DIMENSIONS

An aircraft flying in the vicinity of an electromagnetic pulse (EMP) acts like a receiving antenna and picks up EMP radiation in relation to size like a dipole (or half-wavelength dipole). The electromagnetic pulse spectrum decreases above 1 MHz as shown in Figure 1, so an F-14 aircraft that is an optimum ½ wavelength antenna at $\approx$8 MHz will pick up less EMP voltage than a B-52 or an aircraft with a trailing wire antenna. A rule of thumb for the voltage picked up is:

V_{EMP} = 8.1 volts/ft times the maximum dimension of the aircraft in feet

This rule of thumb was generated because a single linear relationship between voltage and aperture seemed to exist and compared favorably with more complex calculations for voltage picked up by various aircraft when subjected to EMP.

Table 1 shows various aircraft and the frequencies they would be most susceptible to, using $f = c/\lambda$, where λ matches the selected aircraft dimension for maximum "antenna reception effect." This should be a design consideration when trying to screen onboard avionics from the effects of EMP.

The following is a partial listing of aircraft types vs. identifying prefix letters (several are used in Table 1):

A	Attack	K	Tanker	T	Trainer
B	Bomber	O	Observation	U	Utility
C	Cargo	P	Patrol	V	Vertical or Short Takeoff and Landing (V/STOL)
E	Electronic Surveillance	Q	Special mission		
F	Fighter	R	Reconnaissance	X	Experimental
H	Helicopter	S	Anti Sub/Ship	Y	Prototype

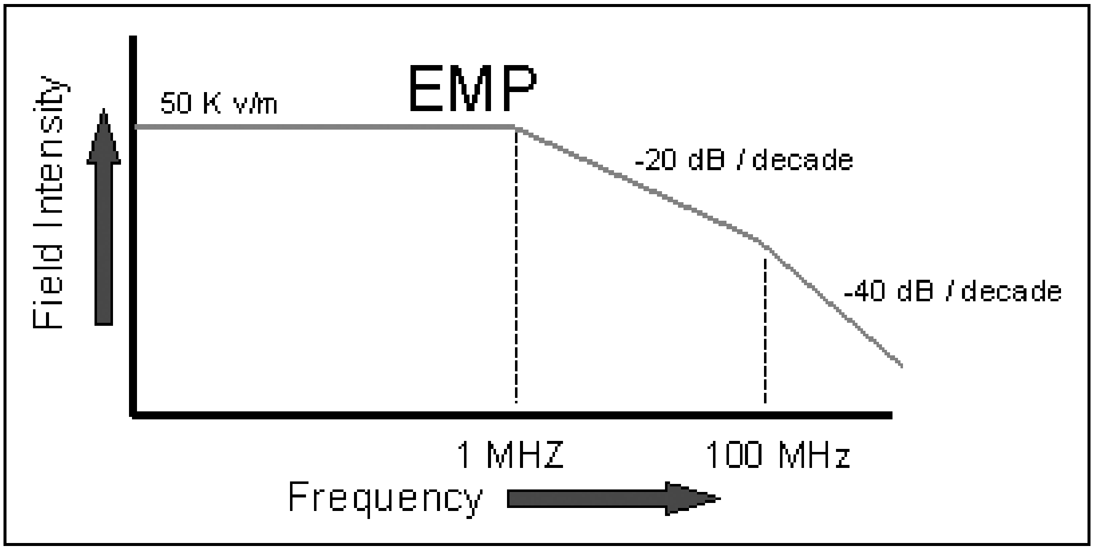

Figure 1. EMP as a Function of Frequency.

Table 1. Aircraft Dimensions and Equivalent Antenna Aperture.

Mission	Aircraft Type	Height (ft.)	Frequency (MHz)		Length (ft.)	Frequency (MHz)		Wing Span (ft.)	Frequency (MHz)	
		A	f	f/2	A	f	f/2	A	f	f/2
ATTACK	A-10	14.66	67.05	33.52	53.33	18.43	9.21	57.5	17.1	8.55
ELECTRONIC WARFARE	EA-6B	16.50	59.64	29.82	59.34	16.58	8.29	53.0	18.57	9.29
	EA-18	16.0	61.47	30.74	60.17	16.35	8.17	44.67	22.02	11.01
FIGHTER	F-15	18.4	53.42	26.71	63.75	15.42	7.71	42.8	22.97	11.48
	F-16	16.66	59.00	29.5	49.25	19.96	9.98	31.0	31.71	15.85
	F/A-18C/D	15.3	64.31	32.16	56.0	17.57	8.79	40.70	24.18	12.09
	F/A-18E/F	16	61.47	30.7	60.17	16.34	8.17	44.67	22.02	11.01
	F-22	17.75	55.41	27.7	64.17	15.33	7.66	43	22.87	11.43
	F-35A	14.2	69.27	34.63	51.4	19.1	9.57	35	28.1	14.05
	F-35C	15.5	63.45	31.73	51.4	19.1	9.57	43	22.8	11.43
	F-117	12.42	79.15	39.57	65.92	14.91	7.46	43.33	22.69	11.34
ASW	P-3C	33.75	29.16	14.58	116.42	8.45	4.23	99.67	9.87	4.94
	S-3A	22.75	43.25	21.63	54.34	18.45	9.23	68.67	14.33	7.17
	P-8	42.1	23.36	11.68	129.5	7.6	3.8	123.6	7.96	3.98
AEW	E-2C	18.4	53.48	26.74	56.50	17.42	8.71	80.58	12.21	6.11
V/STOL	AV-8B	11.64	84.45	42.23	46.3	21.23	10.62	30.3	32.44	16.22
	V-22	18.1	54.3	27.2	57.3	17.17	8.58	84.5	11.64	5.82
	F-35B	15.0	65.6	32.78	51.2	19.2	9.61	35	28.1	14.05
HELICOPTERS	AH-1W	14.16	69.46	34.73	58.25	16.89	8.44	48.0	20.49	10.24
	UH-1E	12.75	71.18	35.59	52.91	18.60	9.30	44.0	22.36	11.18
	U/SH-2	15.41	63.85	31.93	52.5	18.74	9.37	44.0	22.36	11.18
	SH-3D	16.42	59.93	29.97	72.67	13.54	6.77	62.00	15.87	7.84
	CH-46D	16.75	58.75	29.38	84.34	11.67	5.84	50.0	19.68	9.84
	CH-47	18.75	52.45	26.23	99.0	9.94	4.98	60.0	16.39	8.19
	CH-53A	24.91	39.50	19.75	88.16	11.16	5.58	72.25	13.62	6.81
	SH-60	17.0	57.85	28.93	64.83	15.17	7.59	53.83	18.27	9.14
	AH-64	15.25	64.5	32.25	58.25	16.89	8.44	48.0	20.49	10.24
TRANSPORT	C-2A	15.92	61.81	30.91	56.6	17.39	8.70	80.58	12.21	6.11
TANKERS	KC-130F	38.1	25.83	12.92	97.8	10.06	5.03	132.5	7.43	3.72
SPECIAL ELECTRONICS	EC-13OQ	38.5	25.56	12.78	99.34	9.91	4.96	132.58	7.42	3.71
TRAINER	T-2B	14.8	66.49	33.25	38.7	25.43	12.72	37.85	26.00	13.0
	T-39D	16.0	61.50	30.75	43.75	22.49	11.25	44.34	22.19	11.10
	T-45	13.5	72.86	36.43	39.33	25.0	12.5	30.8	31.93	15.97

DATA TRANSFER BUSSES

Data Busses 9-1
RS-232 Interface 9-2
RS-422 Balanced Voltage Interface 9-3
RS-485 Interface 9-4
IEEE-488 Interface Bus (HP-IB/GP-IB) 9-5
MIL-STD-1553 & 1773 Data Bus 9-6
Ethernet 9-7
Transmission Control Protocol / Internet Protocol 9-8

This page intentionally left blank.

DATA BUSSES

INTRODUCTION

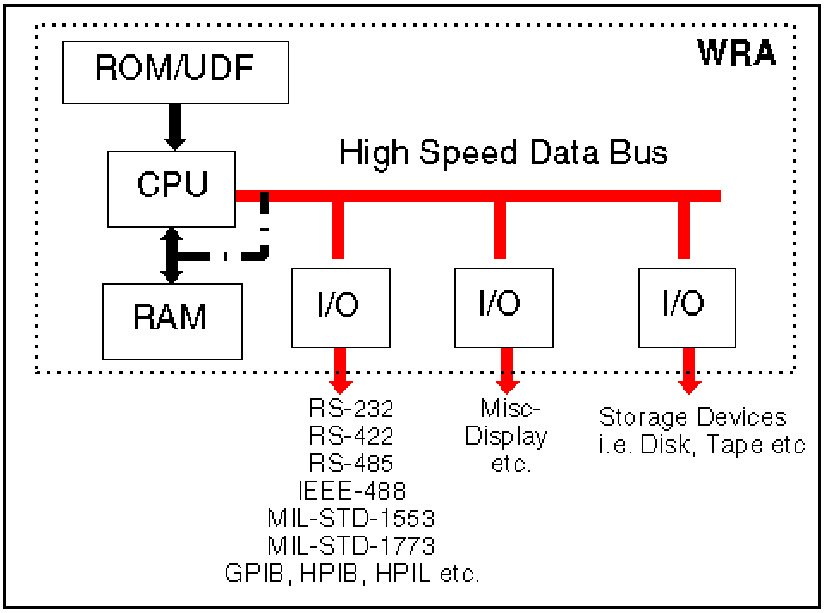

Figure 1. Avionics Block Diagram.

The avionics systems on aircraft frequently contain general purpose computer components which perform certain processing functions, then relay this information to other systems. Some common examples are the mission computers, the radar processors, RWRs, and jammers. Each system is frequently laid out as shown in Figure 1.

The Input/Output (I/O) modules will vary in function, but all serve the same purpose - to translate the electrical signals from one protocol to one of another in order to exchange information. I/O modules are used similarly in general purpose computers in laboratories to test equipment and/or tie computers together via a local area network (LAN) to exchange information. Some of the methodologies include a star, ring, or bus type network (see terminology at the end of this section).

A number of network "models" exist for describing the functions, interfaces and protocols involved in network data interchange. Regardless of the descriptive model used, all strive toward the same end and none actually changes the detailed implementation. Table 1 shows the layer names and the number of layers of networking models presented in Internet Engineering Task Forces (IETF) Request For Comments (RFCs) and in common use by textbooks.

Table 1. Some Common Network Architecture Models.

IETF-RFC 1122	Kurose, Forouzan	Comer, Kozierok	Stallings	Cisco Academy
Four layer Internet model	*Five-layer Internet model or TCP/IP protocol suite*	*TCP/IP 5-layer reference model*	*Five layer TCP/IP model*	*Four layer Internet model*
Application	Application	Application	Application	Application
Transport	Transport	Transport	Host-to-host or transport	Transport
Internet	Network	Internet	Internet	Internetwork
Link	Data link	Data link (Network interface)	Network access	Network interface
	Physical	(Hardware)	Physical	

These textbooks are secondary sources that may be contrary to the intent of RFC 1122 and other primary sources such as the Open Systems Interconnection (OSI) Reference Model developed by the International Organization for Standardization (ISO).

The OSI Reference Model is a more general description for layered communications and computer network protocol design. The IETF makes no effort to follow the OSI model although RFCs sometimes refer to it. The description of the OSI layers is shown in Table 2.

Table 2. OSI Reference Model.

	Name	Use	Misc Example	IP Suitte
7	APPLICATION	Meaning of data	HL7, Modbus	DHCP, DNS, FTP, Gopher, HTTP, NTP, SMTP, SNMP, Telnet
6	PRESENTATION	Building blocks of data and encryption	ASCII, EBCDIC, MIDI	MIME, XDR, SSL
5	SESSION	Opening and closing of specific communication paths	NetBIOS, SAP, Half Duplex, Simplex, SDP	NetBIOS, RTP, SAP
4	TRANSPORT	Error checking	NBF	PPTP, SCTP, TCP, UDP
3	NETWORK	Determination of data paths within the network	NBF, Q.931, IS-IS	IP, ICMP
2	DATA LINK	Data transmission, source, destination, and checksum	802.3 (Ethernet), PPP 802.11a/b/g/n MAC, FDDI	ARP, PPP, PPTP, SLIP
1	PHYSICAL	Voltage levels, signal connections, wire, or fiber	RS-232, 10Base-T, 802.11a/b/g/n Physical	

A layer is a collection of conceptually similar functions that provide services to the layer above it and receives service from the layer below it. On each layer an instance provides services to the instances at the layer above and requests service from the layer below. For example, a layer that provides error-free communications across a network provides the path needed by applications above it, while it calls the next lower layer to send and receive packets that make up the contents of the path. Conceptually two instances at one layer are connected by a horizontal protocol connection on that layer.

Most networks do not use all layers. For example, RS-232 is only a physical layer. Ethernet is only layers 1 and 2. TCP/IP is a protocol, not a network, and uses layers 3 and 4 regardless of whether layers 1 and 2 are a telephone line, wireless connection, or 10Base-T Ethernet cable.

Most of the sections in this division discuss the lowest (physical) layer of communication. There are, however, several more areas of general interest which are included in later sections such as Ethernet and TCP/IP. These are used in general purpose computers like the desktop PC or lab networks, and are not commonly used in aircraft.

The typical high-speed data busses on avionics/computers do not operate as fast as the CPU clock speed, but they are much faster than the interface busses they connect to. There are a number of interface busses (physical layer in network model) which are widely used by aircraft, avionics systems and test equipment. The most common include the RS-232, the RS-422, the RS-485, the IEEE-488 (GP-IB/HP-IB) and the MIL-STD-1553A/B. The MIL-STD-1773 bus is a fiber optic implementation of the 1553 bus and may be used in the future when technology requires it to reduce susceptibility to emissions or other reasons.

A summary of these more common types follows in Table 3, which includes a brief descriptive comparison, while a section covering each in more detail is provided later.

Table 3. Summary of Physical Bus Characteristics.

Bus	Max Length	Max Number of Terminals(1)	Type	# of Lines(2)	Data Rate	Rise Time(3)	Data Format
RS-232C	100 feet max 50 ft at 20k bps	1	Serial	3-20	150 - 19,200 baud per sec		5- to 8- bit serial
RS-422	1.2 km(4)	10(5)	Serial	3	see figure in RS-232 section	$<0.1\ T_b$	unspecified
RS-485	unspecified	32	Serial	3	10 MHz	$<0.3\ T_b$	unspecified
IEEE-488 (GP-IB/HP-IB)	20 meters	14	Parallel	16	500 kHz(6)		8-bit parallel
HP-IL	100 meters	30	Serial	2	20 k BPS		serial
MIL-STD-1553B MIL-STD-1773	300 feet N/A	32(7)	Serial	3	1 MHz	100-300 ns	20-bit serial

NOTES FROM TABLE:

(1) Max Number of Terminals does not include the bus controller.

(2) Including ground/shield

(3) T_b = time duration of the unit interval at the applicable data signaling rate (pulse width)

(4) Length is function of data signaling rate influenced by the tolerable signal distortion, amount of longitudinally coupled noise and ground potential difference introduced between the controller and terminal circuit grounds as well as by cable balance. See RS-422 section for graph.

(5) Physical arrangement of multiple receivers involves consideration of stub line lengths, fail-safe networks, location of termination resistors, data rate, grounding, etc.

(6) Rate can go up to 1 MHz if special conventions are followed.

(7) Max Number of Terminals includes terminal reserved for broadcast commands.

BUS TERMINOLOGY

10BASE-T: Standard "Plain Vanilla" Ethernet based on Unshielded Twisted Pair wire

10BASE-F: 10Mbps fiber optic Ethernet

100BASE-T: Standard "Fast Ethernet" based on twisted pair copper wire

ADDRESS: A unique designation for the location of data or the identity of an intelligent device; allows each device on a single communications line to respond to its own message.

ASCII (American Standard Code for Information Interchange): Pronounced asky. A seven-bit-plus-parity code established by ANSI to achieve compatibility between data services.

ASYNCHRONOUS OPERATION: Asynchronous operation is the use of an independent clock source in each terminal for message transmission. Decoding is achieved in receiving terminals using clock information derived from the message.

BAUD: Unit of signaling speed. The speed in baud is the number of discrete events per second. If each event represents one bit condition, baud rate equals bits per second (BPS). When each event represents more than one bit, baud rate does not equal BPS.

BIT: Contraction of binary digit: may be either zero or one. A binary digit is equal to one binary decision or the designation of one or two possible values of states of anything used to store or convey information.

BIT RATE: The number of bits transmitted per second.

BRIDGE: A network bridge connects multiple network segments at the data link layer (layer 2) of the OSI model, and the term layer 2 switch is very often used interchangeably with bridge. Bridges are similar to repeaters or network hubs, devices that connect network segments at the physical layer; however, with bridging, traffic from one network is managed rather than simply rebroadcast to adjacent network segments.

BROADCAST: Operation of a data bus system such that information transmitted by the bus controller or a remote terminal is addressed to more than one of the remote terminals connected to the data bus.

BUS CONTROLLER: The terminal assigned the task of initiating information transfers on the data bus.

BUS MONITOR: The terminal assigned the task of receiving bus traffic and extracting selected information to be used at a later time.

BYTE: A binary element string functioning as a unit, usually shorter than a computer "word." Eight-bits per byte are most common. Also called a "character."

COMMAND/RESPONSE: Operation of a data bus system such that remote terminals receive and transmit data only when commanded to do so by the bus controller.

CRC: Cyclic Redundancy Check; a basic error-checking mechanism for link-level data transmissions; a characteristic link-level feature of (typically) bit-oriented data communications protocols. The data integrity of a received frame or packet is checked by an algorithm based on the content of the frame and then matched with the result that is performed by a sender and included in a (most often, 16-bit) field appended to the frame.

CROSSOVER CABLE: Cable with transmit/receive pairs reversed so one computer or hub or switch can link directly to another.

DATA BUS: Whenever a data bus or bus is referred to in MIL-STD-1553B, it shall imply all the hardware including twisted shielded pair cables, isolation resistors, transformers, etc., required to provide a single data path between the bus controller and all the associated remote terminals.

DCE (Data Communications Equipment): Devices that provide the functions required to establish, maintain, and terminate a data-transmission connection; e.g., a modem.

DHCP: Dynamic Host Configuration Protocol - permits auto-assignment of temporary IP addresses for new devices logging in

DNS: Domain Name Server - associates names with IP addresses

DTE (Data Terminal Equipment): Devices acting as data source, data sink, or both.

DUPLEX: Communication traveling between two nodes in both directions

DYNAMIC BUS CONTROL: The operation of a data bus system in which designated terminals are offered control of the data bus.

EIA (Electronic Industries Association): A standards organization in the U.S.A. specializing in the electrical and functional characteristics of interface equipment.

FDM (Frequency-Division Multiplexer: A device that divides the available transmission frequency range into narrower banks, each of which is used for a separate channel.

FDX (Full Duplex): Simultaneous, two-way, independent transmission in both directions (4-wire).

FTP: File Transfer Protocol - the most popular mechanism for bulk movement of files on TCP/IP.

GATEWAY: Device which links Ethernet to dissimilar networks and transfers data at the application layer level. Interface cards link the PC to Ethernet via the PCI, ISA, PCMCIA, PC/104, or other buses.

GPIB: General Purpose Interface Bus (see section 9-5),

HALF DUPLEX: Operation of a data transfer system in either direction over a single line, but not in both directions on that line simultaneously.

HANDSHAKING: Exchange of predetermined signals between two devices establishing a connection. Usually part of a communications protocol.

HPIB / HPIL: Hewlett-Packard Interface Bus / Hewlett-Packard Interface Loop

HUB: The simplest method of redistributing data, are “dumb,” not interpreting or sorting messages that pass through them. A hub can be as simple as an electrical buffer with simple noise filtering. It isolates the impedances of multiple spokes in a star topology. Some hubs also have limited store-and-forward capability. They indiscriminately transmit data to all other devices, which are still on the same collision domain, connected to the hubs. They are not assigned MAC addresses or IP addresses.

IEEE (Institute of Electrical and Electronic Engineers): An international professional society that issues its own standards and is a member of ANSI and ISO.

IP: Internet Protocol portion of TCP/IP. It is a protocol used for communicating data across a packet-switched network.

IP ADDRESS: Address of a TCP/IP enabled device on an Intranet or Internet – in the form xxx.xxx.xxx.xxx, where xxx is an integer between 0 and 255.

LAN: Local Area Network

MAC: Media Access Control - the physical components which disassemble Ethernet message fames

MANCHESTER ENCODING: Digital encoding technique (specified for the IEEE 802.3 Ethernet baseband network standard) in which each bit period is divided into two complementary halves; a negative-to-positive (voltage) transition in the middle of the bit period designates a binary “1,” while a positive-to-negative transition represents a “0.” The encoding technique also allows the receiving device to recover the transmitted clock from the incoming data stream (self-clocking).

MESSAGE: A single message is the transmission of a command word, status word, and data words if they are specified. For the case of a remote terminal to remote terminal (RT to RT) transmission, the message shall include the two command words, the two status words, and data words.

MODE CODE: A means by which the bus controller can communicate with the multiplex bus related hardware, in order to assist in the management of information flow.

MODEM (Modulator-Demodulator): A device used to convert serial digital data from a transmitting terminal to a signal suitable for transmission over a telephone channel, or to reconvert the transmitted signal to serial digital data for acceptance by a receiving terminal.

MULTIPLEXOR (also Multiplexer): A device used for division of a transmission into two or more subchannels, either by splitting the frequency band into narrower bands (frequency division) or by allotting a common channel to several different transmitting devices one at a time (time division).

NETWORK: An interconnected group of nodes; a series of points, nodes, or stations connected by communications channels; the assembly of equipment through which connections are made between data stations.

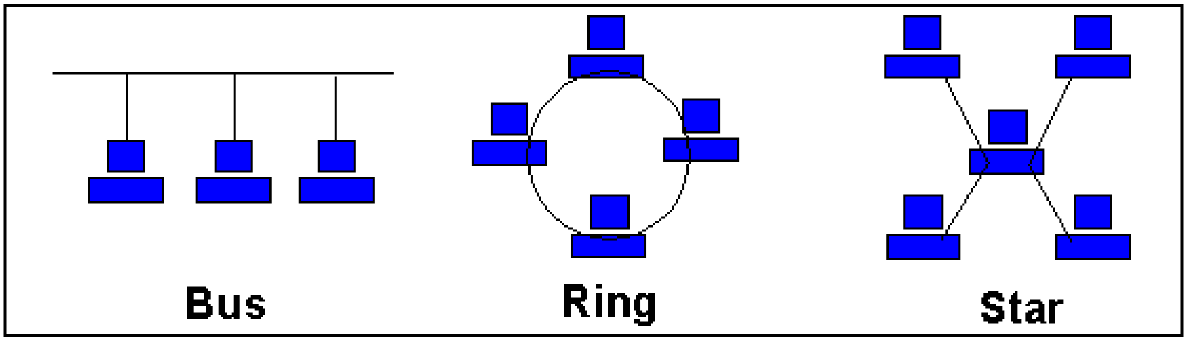

NODE: A point of interconnection to a network. Normally, a point at which a number of terminals or tail circuits attach to the network.

PARALLEL TRANSMISSION: Transmission mode that sends a number of bits simultaneously over separate lines (e.g., eight bits over eight lines) to a printer. Usually unidirectional.

PHASE MODULATION: One of three ways of modifying a sine wave signal to make it “carry” information. The sine wave or “carrier” has its phase changed in accordance with the information to be transmitted.

PING: Packet Inter Net Groper - very useful utility which probes for the existence of a TCP/IP host

POLLING: A means of controlling devices on a multipoint line.

PORT: A number in TCP/IP to which services are assigned; e.g. FTP is port 21; SMTP is port 25; HTTP is port 80.

PROTOCOL: A formal set of conventions governing the formatting and relative timing of message exchange between two communicating systems.

PULSE CODE MODULATION (PCM): The form of modulation in which the modulation signal is sampled, quantized, and coded so that each element of information consists of different types or numbers of pulses and spaces.

REMOTE TERMINAL (RT): All terminals not operating as the bus controller or as a bus monitor.

REPEATER: Buffer which cleans up, strengthens and re-transmits a signal.

ROUTER: Repeater which selectively re-distributes messages based on IP address

SERIAL TRANSMISSION: The most common transmission mode; in serial, information bits are sent sequentially on a single data channel.

SNMP: Simple Network Management Protocol; allows monitoring and management of a network.

SOCKET: Specific instance of an IP address and Port number that represents a single connection between two applications.

STAR TOPOLOGY: Topology which allows only one device at each end of a wire and requires repeaters for more than two devices.

STUBBING: Stubbing is the method wherein a separate line is connected between the primary data bus line and a terminal. The direct connection of stub line causes a mismatch which appears on the waveforms. This mismatch can be reduced by filtering at the receiver and by using bi-phase modulation. Stubs are often employed not only as a convenience in bus layout but as a means of coupling a unit to the line in such a manner that a fault on the stub or terminal will not greatly affect the transmission line operation. In this case, a network is employed in the stub line to provide isolation from the fault. These networks are also used for stubs that are of such length that the mismatch and reflection degrades bus operation. The preferred method of stubbing is to use transformer coupled stubs. The method provides the benefits of DC isolation, increased common mode protection, a doubling of effective stub impedance, and fault isolation for the entire stub and terminal. Direct coupled stubs should be avoided if at all possible. Direct coupled stubs provide no DC isolation or common mode rejection for the terminal external to its subsystem. Further, any shorting fault between the subsystems' internal isolation resistors (usually on the circuit board) and the main bus junction will cause failure of that entire bus. It can be expected that when the direct stub length exceeds 1.6 feet, that it will begin to distort the main bus waveforms. Note that this length includes the cable runs internal to a given subsystem.

SUBSYSTEM: The device or functional unit receiving data transfer service from the data bus.

SWITCH: Repeater which selectively re-distributes messages based on hardware MAC address

SYNCHRONOUS TRANSMISSION: Transmission in which data bits are sent at a fixed rate, with the transmitter and receiver synchronized. Synchronized transmission eliminates the need for start and stop bits.

TCP: Transmission Control Protocol - mechanism in TCP/IP that ensures that data arrives intact and in correct order

TELNET: Standard interface through which a client may access a host as though it were local

TERMINAL: The electronic module necessary to interface the data bus with the subsystem and the subsystem with the data bus. Terminals may exist as separate units or be contained within the elements of the subsystem.

TIME DIVISION MULTIPLEXING (TDM): The transmission of information from several signal sources through one communication system with different signal samples staggered in time to form a composite pulse train.

UDP: User Datagram Protocol - lower overhead alternative to TCP protocol which does not guarantee message delivery

WORD: A set of bits or bytes comprising the smallest unit of addressable memory. In MIL-STD-1553B, a word is a sequence of 16 bits plus sync and parity.

RS-232 INTERFACE

INTRODUCTION

The RS-232 interface is the Electronic Industries Association (EIA) standard for the interchange of serial binary data between two devices. It was initially developed by the EIA to standardize the connection of computers with telephone line modems. The standard allows as many as 20 signals to be defined, but gives complete freedom to the user. Three wires are sufficient: send data, receive data, and signal ground. The remaining lines can be hardwired on or off permanently. The signal transmission is bipolar, requiring two voltages, from 5 to 25 volts, of opposite polarity.

COMMUNICATION STANDARDS

The industry custom is to use an asynchronous word consisting of: a start bit, seven or eight data bits, an optional parity bit and one or two stop bits. The baud rate at which the word sent is device-dependent. The baud rate is usually 150 times an integer power of 2, ranging from 0 to 7 (150, 300, 600, ..., 19,200). Below 150 baud, many system-unique rates are used. The standard RS-232-C connector has 25 pins, 21 pins which are used in the complete standard. Many of the modem signals are not needed when a computer terminal is connected directly to a computer, and Figure 1 illustrates how some of the “spare” pins should be linked if not needed. Figure 1 also illustrates the pin numbering used in the original DB-25 connector and that now commonly used with a DB-9 connector normally used in modern computers

Specifying compliance to RS-232 only establishes that the signal levels in two devices will be compatible and that if both devices use the suggested connector, they may be able to be connected. Compliance to RS-232 does not imply that the devices will be able to communicate or even acknowledge each other’s presence.

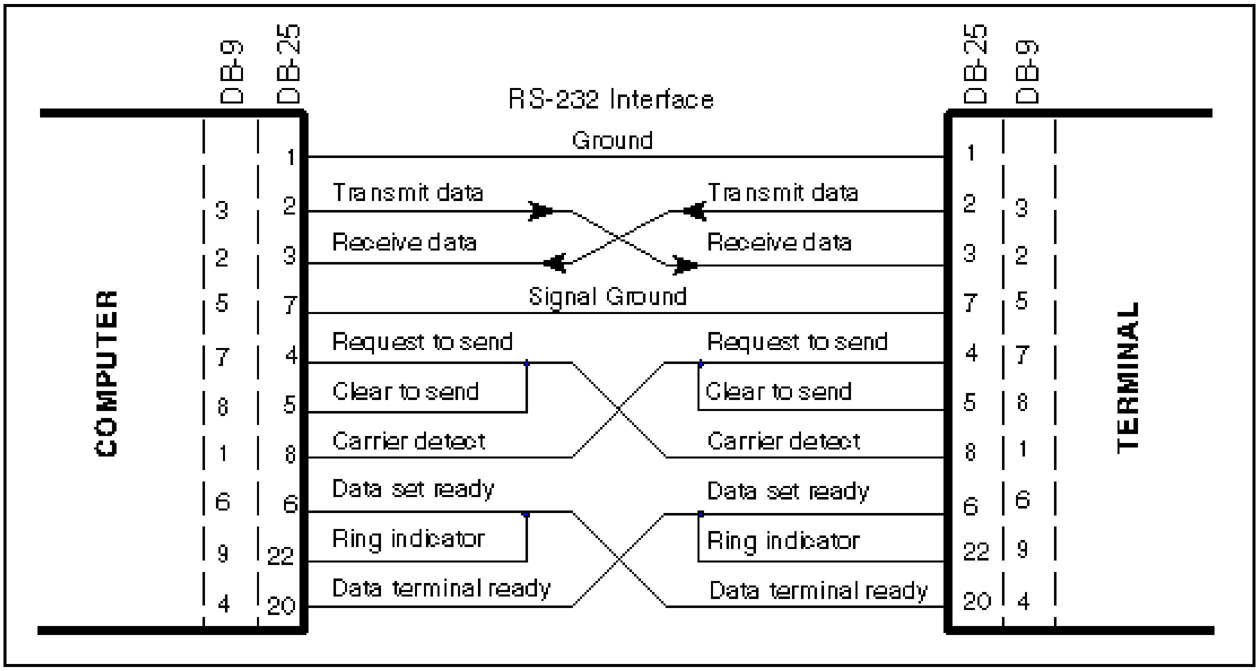
DB-9
DB-25
RS-232 Interface
Ground
Transmit data
Receive data
Signal Ground
Request to send
Clear to send
Carrier detect
Data set ready
Ring indicator
Data terminal ready
COMPUTER
TERMINAL
DB-25
DB-9
1
2
3
7
4
5
8
6
22
20
3
2
5
7
8
1
6
9
4

Figure 1. Direct-to-Computer RS-232 Interface.

Table 1 shows the signal names, and functions of the RS-232 serial port pinout. Table 2 shows a complete pin description.

Table 1. RS-232 Serial Port Pinout.

Name	Pin	Signal Name	Function
AA	1	PG Protective Ground	This line is connected to the chassis ground of the GPIB-232CV. Since the GPIB-232CV chassis ground is not connected to earth ground, pin 1 should be connected on both serial devices.
BA	2	TxD Transmit Data	This line carries serial data from the GPIB-232CV to the serial host.
BB	3	RxD Receive Data	This line carries serial data from the serial host to the GPIB-232CV.
CA	4	RTS Request to Send	This signal line is driven by the GPIB-232CV and when asserted indicates that the GPIB-232CV is ready to accept serial data. The GPIB-232CV un-asserts RTS when it is no longer ready to accept serial data because of a buffer full condition.
CB	5	CTS Clear to Send	This signal line is asserted by the serial host and sensed by the GPIB-232CV. When asserted, it indicates that the serial host is ready to accept serial data. When unasserted, it indicates that data transmission should be disabled.
AB	7	SG Signal Ground	This line establishes a reference point for all interface voltages.
CD	20	DTR Data Terminal Ready	This signal line is asserted by the GPIB-232CV to signal that it has been powered on, and is ready to operate.

Table 2. RS-232C Interface Signals.

Pin	Description	Pin	Description	Pin	Description
1	Protective Ground	10	(Reserved for Data Set Testing)	19	Secondary Request to Send
2	Transmitted Data	11	Unassigned	20	Data Terminal Ready
3	Received Data	12	Sec. Rec'd. Line Sig. Detector	21	Signal Quality Detector
4	Request to Send	13	Sec. Clear to Send	22	Ring Indicator
5	Clear to Send	14	Secondary Transmitted Data	23	Data Signal Rate Selector (DTE/DCE Source)
6	Data Set Ready	15	Transmission Signal Element Timing (DCE Source)	24	Transmit Signal Element Timing (DTE Source)
7	Signal Ground (Common Return)	16	Secondary Received Data	25	Unassigned
8	Received Line Signal Detector	17	Receiver Signal Element Timing (DCE Source)		
9	(Reserved for Data Set Testing)	18	Unassigned		

Electrical Characteristics: The RS-232-C specifies the signaling rate between the DTE and DCE, and a digital signal is used on all interchange circuits. The RS-232 standard specifies that logic "1" is to be sent as a voltage in the range -15 to -5 V and that logic "0" is to sent as a voltage in the range +5 to +15 V. The standard specifies that voltages of at least 3 V in amplitude will always be recognized correctly at the receiver according to their polarity, so that appreciable attenuation along the line can be tolerated. The transfer rate is rated > 20 kbps and a distance of < 15 m. Greater distance and data rates are possible with good design, but it is reasonable to assume that these limits apply in practice as well as in theory. The load impedance of the terminator side of the interface must be between 3000 and 7000 ohms, and not more than 2500 pF.

Table 3, summarizes the functional specifications of the most important circuits.

Table 3. RS-232-C Circuit Definitions.

Name	Direction to:	Function
Data Signals		
Transmitted Data (BA)	DCE	Data generated by DTE
Received Data (BB)	DTE	Data Received by DTE
Timing signals		
Transmitter Signal Element Timing (DA)	DCE	Clocking signal, transitions to ON and OFF occur at center of each signal element
Transmitter Signal Element Timing (DB)	DTE	Clocking signal, as above; both leads relate to signals on BA
Receiver Signal Element Timing (DD)	DTE	Clocking signal, as above, for circuit BB
Control Signals		
Request to Send (CA)	DCE	DTE wishes to transmit
Clear to Send (CB)	DTE	DCE is ready to transmit; response to request to send
Data Set Ready (CC)	DTE	DCE is ready to operate
Data Terminal Ready (CD)	DCE	DTE is ready to operate
Ring Indicator (CE)	DTE	Indicates that DCE is receiving a ringing signal on the communication channel
Carrier Detect (CF)	DTE	Indicates that DCE is receiving a carrier signal
Signal Quality Detector (CG)	DTE	Asserted when there is reason to believe there is an error in the received data
Data Signal Rate Selector (CH)	DCE	Asserted to select the higher of two possible data rates
Data Signal Rate Selector (CI)	DTE	Asserted to select the higher of two possible data rates
Ground		
Protective Ground (AA)	NA	Attached to machine frame and possibly external grounds
Signal Ground (AB)	NA	Establishes common ground reference for all circuits

Range: The RS-232-C standard specifies that the maximum length of cable between the transmitter and receiver should not exceed 100 feet. Although in practice many systems are used in which the distance between transmitter and receiver exceeds this rather low figure. The limited range of the RS-232C standard is one of its major shortcomings compared with other standards which offer greater ranges within their specifications. One reason why the range of the RS-232C standard is limited is the need to charge and discharge the capacitance of the cable connecting the transmitter and receiver.

Mechanical Characteristics: The connector for the RS-232-C is a 25 pin connector with a specific arrangement of wires. In theory, a 25 wire cable could be used to connect the Data Terminal Equipment (DTE) to the Data Communication Equipment (DCE). The DTE is a device that is acting as a data source, data sink, or both, e.g. a terminal, peripheral, or computer. The DCE is a device that provides the functions required to establish, maintain, and terminate a data-transmission connecting, as well as the signal conversion, and coding required for communication between data terminal equipment and data circuit; e.g. a modem. Table 4, shows the complete summary of the RS-232-C, e.g., descriptor, sponsor, data format, etc.

Table 4. Summary of the RS-232-C.

Data Format	5- to 8- bit serial
Transfer Type	Asynchronous
Error Handling	Optional Parity Bit
Connector	25-pin female connector on DCE; 25-pin male connector on DTE
Length	20 meters
Speed	20 kb/s
Remarks	RS-232 is used in the microcomputer world for communications between two DTEs. The null-modem is included into one or both connecting devices, and/or cable and is seldom documented. As a result, establishing an RS-232 connection between two DTEs is frequently a difficult task.

RS-422 BALANCED VOLTAGE INTERFACE

Specifying compliance to RS-422 only establishes that the signal between the specified devices will be compatible. It does not indicate that the signal functions or operations between the two devices are compatible. The RS-422 standard only defines the characteristic requirements for the balanced line drivers and receivers. It does not specify one specific connector, signal names, or operations. RS-422 interfaces are typically used when the data rate or distance criteria cannot be met with RS-232. The RS-422 standard allows for operation of up to 10 receivers from a single transmitter. The standard does not define operations of multiple tri-stated transmitters on a link.

The RS-422-A interfaces between the Data Terminal Equipment (DTE) and Data Communication Equipment (DCE) or in any point-to-point interconnection of signals between digital equipment. It employs the electrical characteristics of balanced-voltage digital interface circuits.

The balanced voltage digital interface circuit will normally be utilized on data, timing, or control circuits where the data signaling rate is up to 10 Mbit/s. While the balanced interface is intended for use at the higher data signaling rate, it may (in preference to the unbalanced interface circuit) generally be required if any of the following conditions prevail:

- The interconnecting cable is too long for effective unbalanced operation.
- The interconnecting cable is exposed to an extraneous noise source that may cause an unwanted voltage in excess of + 1 volt measured differentially between the signal conductor and circuit common at the load end of the cable with a 50 ohm resistor substituted for the generator.
- It is necessary to minimize interference with other signals.
- Inversion of signals may be required, i.e. plus to minus MARK may be obtained by inverting the cable pair.

Applications of the balanced voltage digital interface circuit are shown in Figure 1.

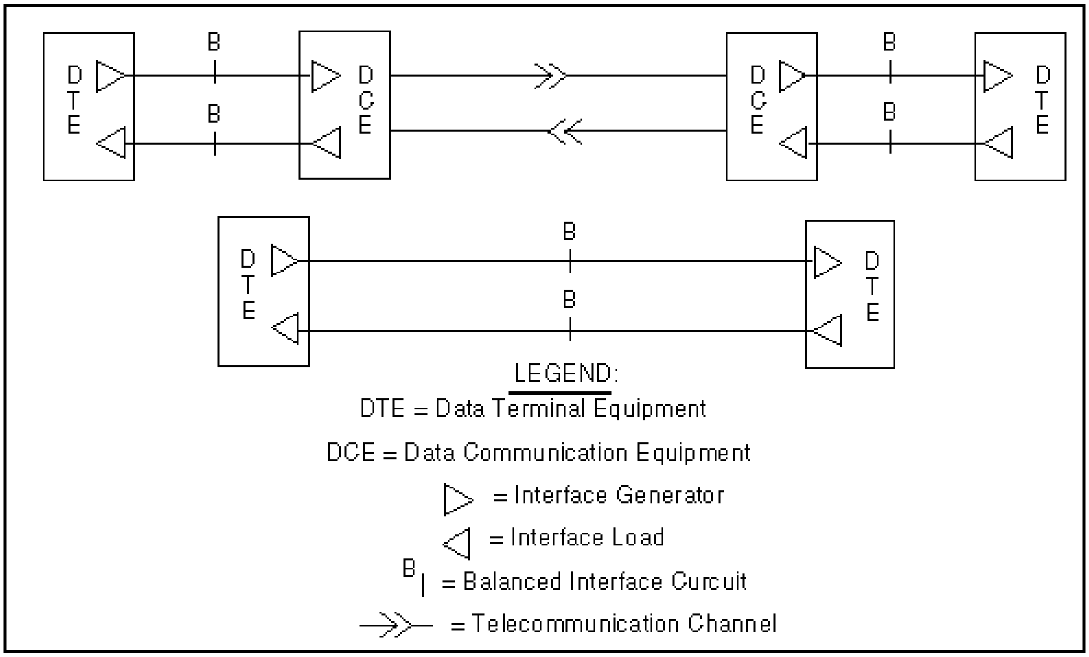
D T E
B
B
D C E
D C E
B
B
D T E
D T E
B
B
D T E
LEGEND:
DTE = Data Terminal Equipment
DCE = Data Communication Equipment
= Interface Generator
= Interface Load
B = Balanced Interface Curcuit
= Telecommunication Channel

Figure 1. Application of a RS-422 Circuit.

While a restriction on maximum cable length is not specified, guidelines are given later with respect to conservative operating distances as function of data signaling rate.

For a binary system in which the RS-422-A is designed, the data signaling rate in bit/s and the modulation in bauds are numerically equal when the unit interval used in each determination is the minimum interval.

Electrical Characteristics:

The balanced voltage digital interface circuit consists of three parts: the generator (G), the balanced interconnecting cable, and the load. The load is comprised of one or more receivers (R) and an optional cable termination resistance (RT). The balanced voltage interface circuit is shown in Figure 2.

Environmental Constraints:

Balanced voltage digital interface conforming to this standard will perform satisfactorily at data signaling rates up to 10 Mbit/s providing that the following operational constraints are satisfied:

- The interconnecting cable length is within that recommended for the applicable data signaling rate (see Figure 3) and the cable is appropriately terminated.
- The common mode voltage at the receiver is less than 7 volts (peak). The common mode voltage is defined to be any uncompensated combination of generator-receiver ground potential difference, the generator offset voltage (Vos), and longitudinally coupled peak noise voltage measured between the received circuit ground and cable within the generator ends of the cable short-circuited to ground.

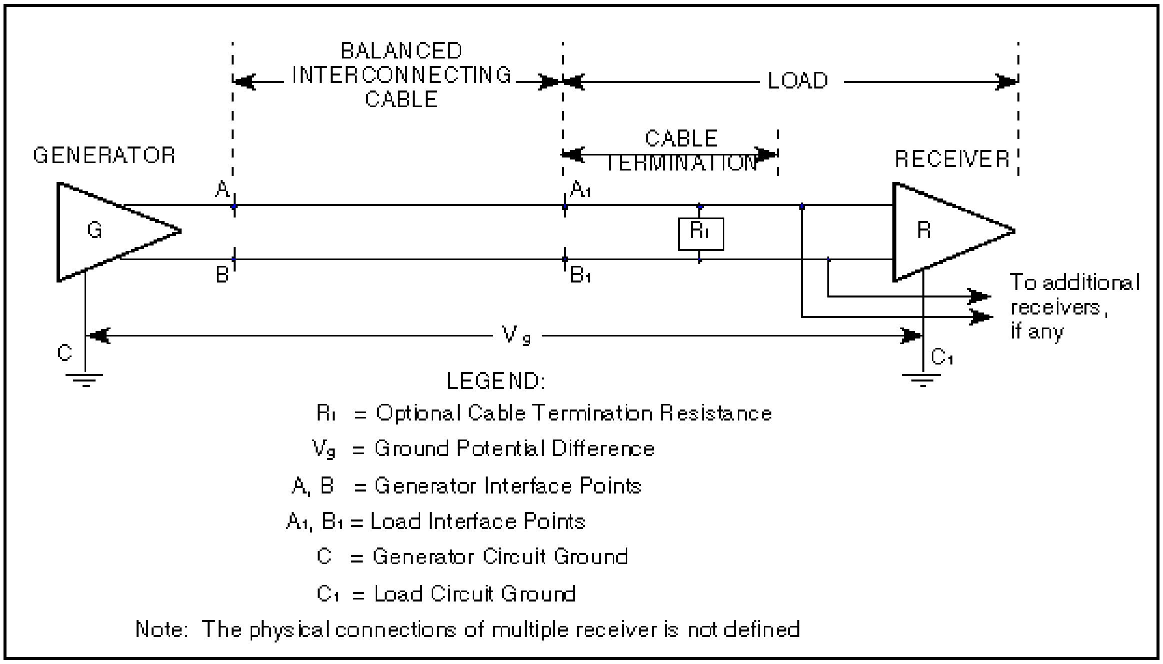
BALANCED INTERCONNECTING CABLE
LOAD
CABLE TERMINATION
GENERATOR
RECEIVER
A
A1
G
R1
R
B
B1
To additional receivers, if any
C
Vg
C1
LEGEND:
R1 = Optional Cable Termination Resistance
Vg = Ground Potential Difference
A, B = Generator Interface Points
A1, B1 = Load Interface Points
C = Generator Circuit Ground
C1 = Load Circuit Ground
Note: The physical connections of multiple receiver is not defined

Figure 2. Balanced Digital Interface Circuit.

Interconnecting Cable Guidelines:

The maximum permissible length of cable separating the generator and the load is a function of data signaling rate and is influenced by the tolerable signal distortion, the amount of coupled noise and ground potential difference introduced between the generator and load circuit as well as by cable balance. The curve of cable length versus signaling rate is given in Figure 3. This curve is based upon using 24 AWG copper, twisted-pair cable with a capacitance of 52.5 pF/meter terminated in a 100 ohm load. As data signaling rate is reduced below 90 kbit/s, the cable length has been limited at 1200 meters by the assumed maximum allowable 6 dBV signal loss.

Industry customs are not nearly as well established for RS-422 interfaces as they are for RS-232. The standard specifies use of the 37-pin "D"; the 9-pin "D" is specified for use with the secondary channel. Most data communications equipment uses the 37-pin "D"; many computer applications use a 9-pin "D" only. Some equipment applications use the 25-pin "D" defined for RS-232.

Compatibility With Other Interfaces:

Since the basic differential receivers of RS-423-A and RS-422-A are electrically identical, it is possible to interconnect an equipment using RS-423-A receivers and generators on one side of the interface with an equipment using RS-422-A generators and receivers on the other side of the interface, if the leads of the receivers and generators are properly configured to accommodate such an arrangement and the cable is not terminated.

This circuit is not intended for interoperation with other interface electrical circuits such as RS-232-C, MIL-STD-188C, or CCITT (Comite Consultatif Internationale Telegraphique et Telephonique), recommendations V.28 and V.35. Under certain conditions, the above interfaces may be possible but may require modification of the interface or equipment; therefore satisfactory operation is not assured and additional provisions not specified herein may be required.

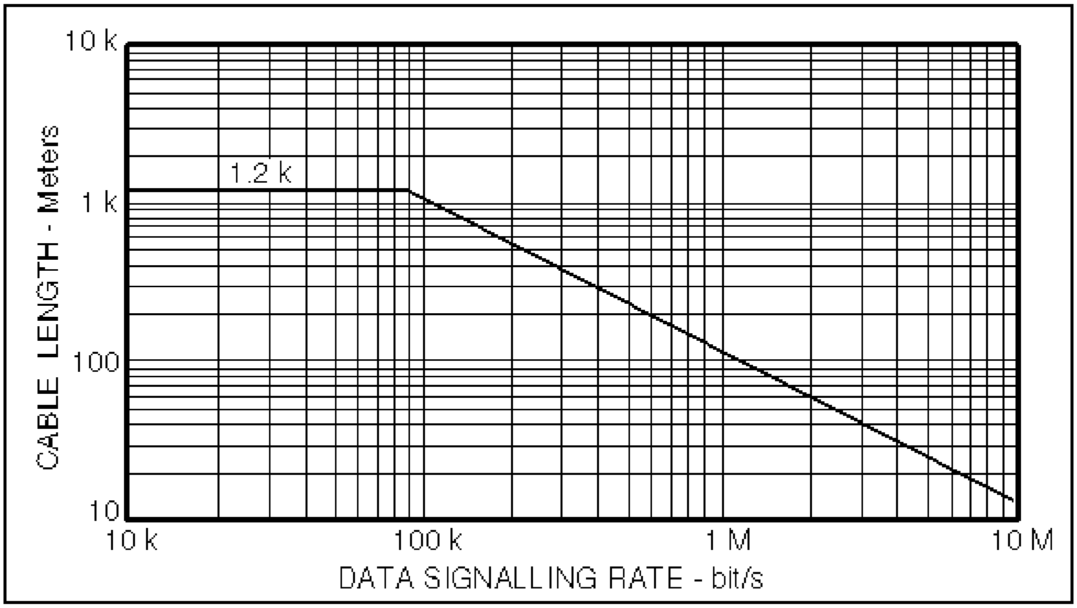

Figure 3. Data Signaling Rate vs. Cable Length.

This page intentionally left blank.

RS-485 INTERFACE

STANDARD FOR ELECTRICAL CHARACTERISTICS OF GENERATORS AND RECEIVERS FOR USE IN BALANCED DIGITAL MULTIPOINT SYSTEMS

Introduction: The RS-485 is the recommend standard by the Electronic Industries Association (EIA) that specifies the electrical characteristics of generators and receivers that may be employed for the interchange of binary signals in multipoint interconnection of digital equipments. When implemented within the guidelines, multiple generators and receivers may be attached to a common interconnecting cable. An interchange system includes one or more generators connected by a balanced interconnecting cable to one or more receivers and terminating resistors.

Electrical Characteristics: The electrical characteristics that are specified are measured at an interconnect point supplied by the devices manufacturer. Figure 1 shows an interconnection application of generators and receivers having the electrical parameters specified. The elements in the application are: generators, receivers, transmission cables, and termination resistances (Rt). The loads on the system caused by each receiver and passive generator shall be defined in terms of unit loads. Each generator can drive up to 32 unit loads consisting of both receivers and generators in the passive state. The loading caused by receivers and passive generators on the interconnect must be considered in defining the device electrical characteristics. Two areas are of concern: the DC load and the AC load characteristics. The DC load is defined as a number or fractions of “unit loads.” The AC loading is not standardized but must be considered in the design of a system using the devices meeting this standard.

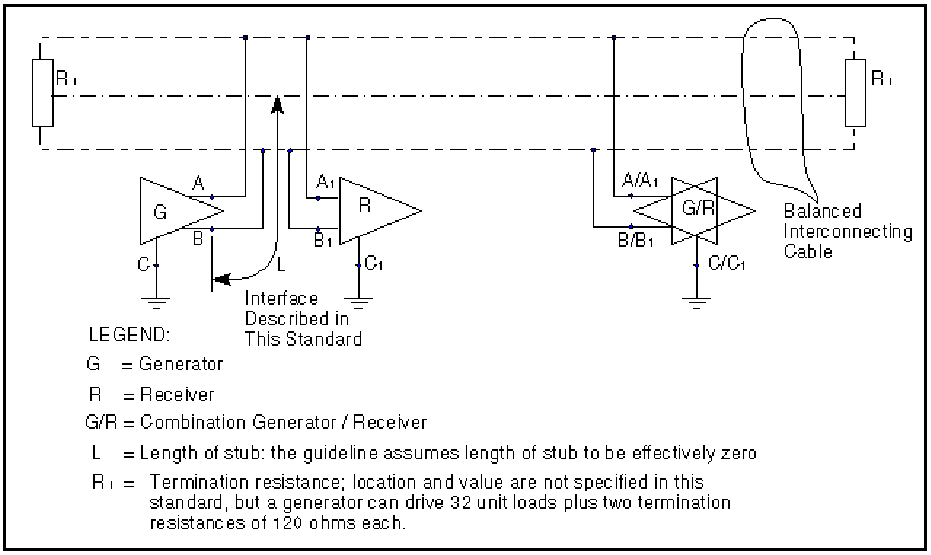
R_t
R_t
A
B
C
G
A_1
B_1
C_1
R
L
Interface Described in This Standard
A/A_1
B/B_1
C/C_1
G/R
Balanced Interconnecting Cable
LEGEND:
G = Generator
R = Receiver
G/R = Combination Generator / Receiver
L = Length of stub: the guideline assumes length of stub to be effectively zero
R_t = Termination resistance; location and value are not specified in this standard, but a generator can drive 32 unit loads plus two termination resistances of 120 ohms each.

Figure 1. Multipoint Interconnect Application.

General System Configuration: The generators and receivers conforming to the RS-485 standard can operate with a common mode voltage between -7 volts and +7 volts (instantaneous). The common mode voltage is defined to be any uncompensated combination of generator-receiver ground potential difference and longitudinally coupled peak noise voltage measured between the receiver circuit ground and cable with the generator ends of the cable short circuited to ground, plus the generator offset voltage (Vos).

Grounding Arrangements: Proper operation of the generator and receiver circuits requires the presence of a signal return path between the circuit grounds of the equipment at each end of the interconnection. The grounding arrangements are shown in Figure 2. Where the circuit reference is provided by a third conductor, the connection between circuit common and the third conductor must contain some resistance (e.g., 100 ohms) to limit circulating currents when other ground connections are provided for safety. Some applications may require the use of shielded interconnecting cable for EMI or other purposes. The shield shall be connected to frame ground at either or both ends, depending on the application.

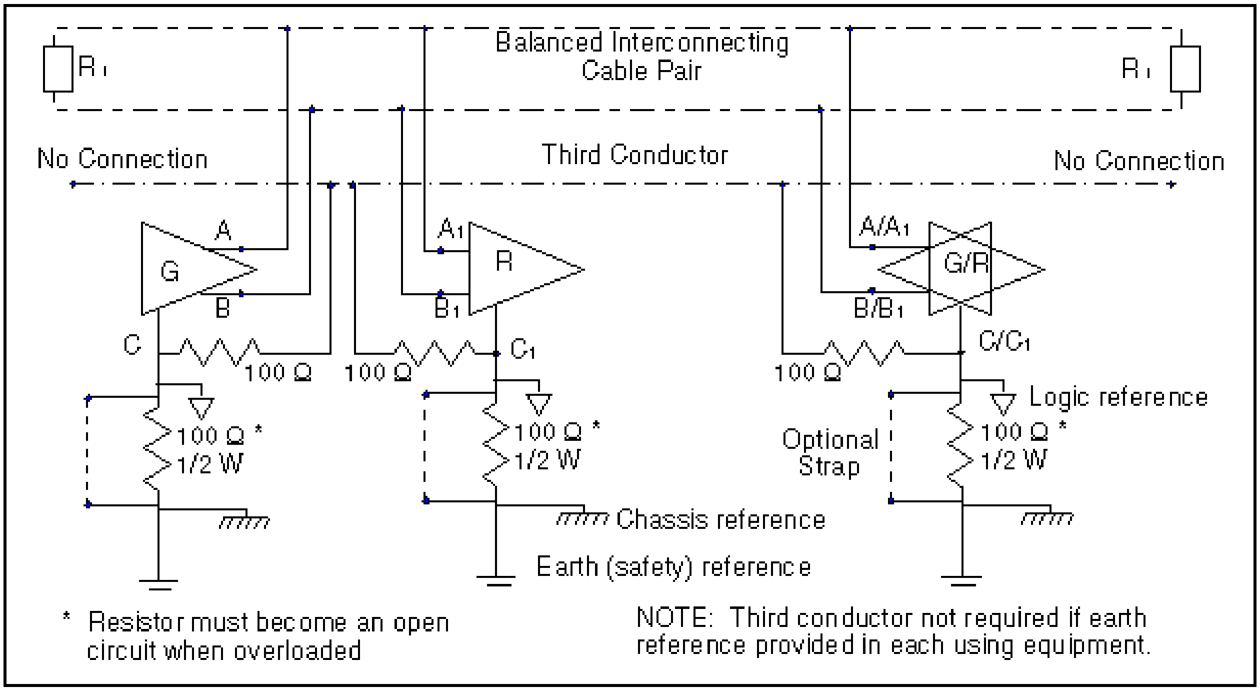
Balanced Interconnecting Cable Pair
R₁
R₁
No Connection
Third Conductor
No Connection
A
G
B
A₁
R
B₁
A/A₁
G/R
B/B₁
C
100 Ω
100 Ω
C₁
100 Ω
C/C₁
Logic reference
100 Ω *
1/2 W
100 Ω *
1/2 W
Optional Strap
100 Ω *
1/2 W
Chassis reference
Earth (safety) reference
* Resistor must become an open circuit when overloaded
NOTE: Third conductor not required if earth reference provided in each using equipment.

Figure 2. Grounding Arrangements.

Similarity with RS-422-A:

In certain instances, it may be possible to produce generators and receivers that meet the requirements of both RS-422-A and of RS-485. Table 1 depicts the differences in parameter specifications which exist between the two documents.

Table 1. Comparison of RS-422-A and RS-485 Characteristics.

Characteristic	RS-422-A	RS-485
Min. output voltage	2V into 100 ohm > 1/2 open circuit V	1.5 V into 54 ohms
I_{short} to ground	150 mA maximum	
I_{short} to -7, +12 volts		250 mA peak
t_{rise} time	< 0.1 t_b , 100 ohm load	< 0.3 t_b , 54 ohm, 50 pF load

Where t_b = time duration of the unit interval at the applicable data signaling rate (pulse width).

IEEE-488 INTERFACE BUS (HP-IB/GP-IB)

In the early 1970's, Hewlett-Packard came out with a standard bus (HP-IB) to help support their own laboratory measurement equipment product lines, which later was adopted by the IEEE in 1975. This is known as the IEEE Std. 488-1975. The IEEE-488 Interface Bus (HP-IB) or general purpose interface bus (GP-IB) was developed to provide a means for various instruments and devices to communicate with each other under the direction of one or more master controllers. The HP-IB was originally intended to support a wide range of instruments and devices, from the very fast to the very slow.

DESCRIPTION

The HP-IB specification permits up to 15 devices to be connected together in any given setup, including the controller if it is part of the system. A device may be capable of any other three types of functions: controller, listener, or talker. A device on the bus may have only one of the three functions active at a given time. A controller directs which devices will be talkers and listeners. The bus will allow multiple controllers, but only one may be active at a given time. Each device on the bus should have a unique address in the range of 0-30. The maximum length of the bus network is limited to 20 meters total transmission path length. It is recommended that the bus be loaded with at least one instrument or device every 2 meter length of cable (4 meters is maximum). The use of GP-IB extenders may be used to exceed the maximum permitted length of 20 meters.

ELECTRICAL INTERFACE

The GP-IB is a bus to which many similar modules can be directly connected, as is shown in Figure 1. A total of 16 wires are shown in the figure - eight data lines and eight control lines. The bus cables actually have 24 wires, providing eight additional for shielding and grounds.

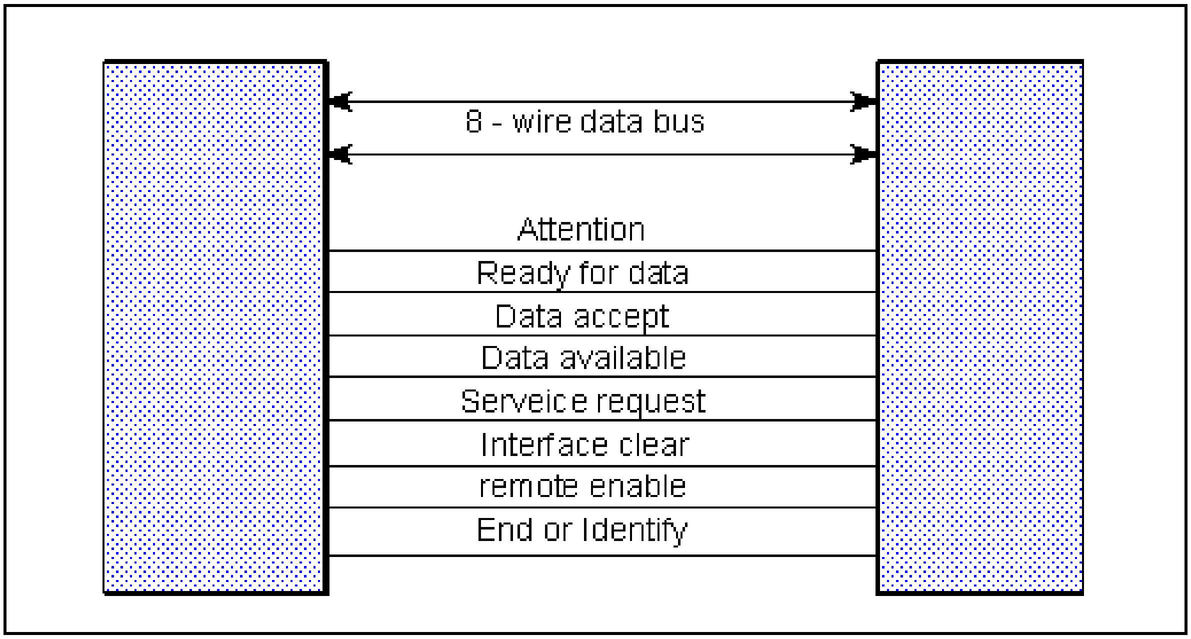

Figure 1. IEEE-488 (HP-IB / GP-IB) Bus Configuration.

The GP-IB defines operation of a three-wire handshake that is used for all data transfers on the bus. The bus operation is asynchronous in nature. The data-transfer rate of the GP-IB is 500 kHz for standard applications and can go up 1 MHz if special conventions are followed. Each transaction carries 8 bits, the maximum data bandwidth is on the order of 4 to 8 megabits (1 M byte) per second. The bus is a two way communications channel and data flows in both directions. Figure 2 illustrates the structure of the GP-IB bus and identifies the 16 connections of the interconnecting cable.

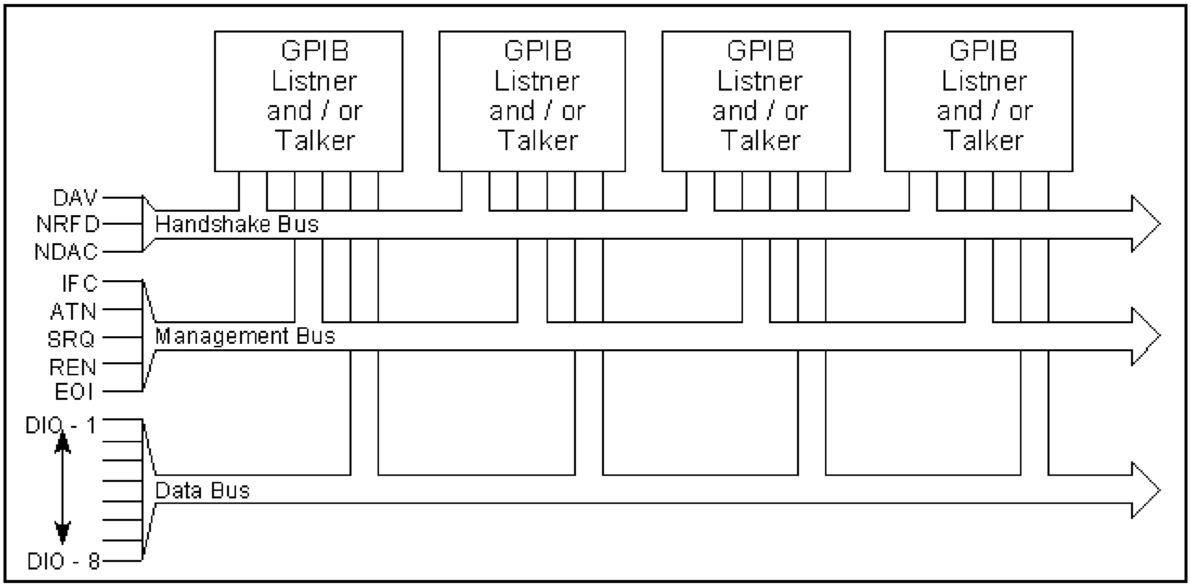
GPIB Listner and / or Talker
GPIB Listner and / or Talker
GPIB Listner and / or Talker
GPIB Listner and / or Talker
DAV
NRFD
NDAC
Handshake Bus
IFC
ATN
SRQ
REN
EOI
Management Bus
DIO - 1
Data Bus
DIO - 8

Figure 2. GP-IB Instrumentation Bus Structure.

The cabling limitations make it a less-than-ideal choice for large separation between devices. These limitations can be overcome with bus extenders. Those attempting to use bus extenders should be aware that few extenders are as transparent as claimed. This is especially true in handling of continuous data and interrupts. In nonextended environments, it provides an excellent means for high-speed computer control of multiple devices.

The following table shows the various interface functions, the mnemonics and the descriptions.

Table 1. GP-IB Interface Functions.

Interface Function	Mnemonic	Description
Talker (extended talker)	T (TE)	Device must be able to transmit
Listener (Extended listener)	L (LE)	Device must receive commands and data
Source Handshake	SH	Device must properly transfer a multiline message
Acceptor Handshake	AH	Device must properly receive remote multiline messages
Remote/Local	RL	Device must be able to operate from front panel and remote information from bus
Service Request	SR	Device can asynchronously request service from the controller
Parallel Poll	PP	Upon controller request, device must uniquely identify itself if it requires service
Device Clear	DC	Device can be initialized to a predetermined state
Device Trigger	DT	A device function can be initiated by the talker on the bus
Controller	C	Device can send addresses, universal commands, address commands, and conduct polls
Drivers	E	This code describes the type of electrical drivers in a device

The cabling specifications of the GP-IB interface system permit interconnecting all devices together in a star or linear configuration. The GP-IB connector is a 24-pin ribbon-type connector.

In summary, Table 2 on this page and the next shows the complete description of the GP-IB data bus.

Table 2. GP-IB Data Bus Description.

IEEE-488, GP-IB, HP-IB, or IEC-625					
Descriptor	8-bit parallel, monodirectional, multi-master (token passing) One controller, one talker, several listeners	Arbitration	Token passing: the controller addresses the next controller SRQ Service request when the controller assigns modes	Connector	24-pin Amphenol Female connector on equipment chassis. DIO1 1 ■ ■ 13 DIO5 DIO2 2 ■ ■ 14 DIO6 DIO3 3 ■ ■ 15 DIO7 DIO4 4 ■ ■ 16 DIO8 EOI 5 ■ ■ 17 REN DAV 6 ■ ■ 18 Gnd NRFD 7 ■ ■ 19 Gnd NDAC 8 ■ ■ 20 Gnd IFC 9 ■ ■ 21 Gnd SRQ 10 ■ ■ 22 Gnd ATN 11 ■ ■ 23 Gnd Shld 2 ■ ■ 24 Gnd
Sponsor	Hewlett-Packard	Error handling	Parity bit DI07 when 7-bit ACSII characters		
Standard	IEEE 488, IEC 625	Bus length	15 m		
Address space	31 devices	Driver	Special 24 mA drivers		
Data format	8-bit parallel	Speed	1 MByte/s		
Transfer type	Write only, talker toward listener(s) or commander toward all others			Remarks	The 488 is most commonly used for data acquisition of H-P peripherals. Programmable interfaces and drivers exist and simplify the development of microprocessor interfaces.
Timing	Handshaken 3-wire broadcast transfer: DAV data valid NDAC Not data accepted NRFD Not ready for data	References	IEEE Computer Society		

HP-IL VARIATION

Since introduction of the IEEE-488, technology produced a generation of medium-speed, low-power, instrumentation which had a need to operate in an automatic test system such as the GP-IB. The HP-IL (Hewlett-Packard Interface Loop), was introduced to meet this need. The HP-IL is a low-cost, low-power alternative to the GP-IB system. The HP-IL and GP-IB provide the same basic functions in interfacing controllers, instruments, and peripherals, but they differ in many other respects. HP-IL is suitable for use in low-power, portable applications (typically used for interface of battery-power systems). The GP-IB is not practical to operate from battery power. The HP-IL maximum data rate is 20K bytes per second. This is a high rate compared to the RS-232C, but much slower than GP-IB. The HP-IL can operate over distances of up to 100 meters between any two devices. Since it is a loop environment, there is no maximum system cable restriction. The basic device-addressing scheme allows for up to 30 devices on a loop.

MIL-STD-1553 & 1773 DATA BUS

PURPOSE

In recent years, the use of digital techniques in aircraft equipment has greatly increased, as have the number of avionics subsystems and the volume of data processed by them.

Because analog point-to-point wire bundles are inefficient and cumbersome means of interconnecting the sensors, computers, actuators, indicators, and other equipment onboard the modern military vehicle, a serial digital multiplex data bus was developed. MIL-STD-1553 defines all aspects of the bus, therefore, many groups working with the military tri-services have chosen to adopt it.

The 1553 multiplex data bus provides integrated, centralized system control and a standard interface for all equipment connected to the bus. The bus concept provides a means by which all bus traffic is available to be accessed with a single connection for testing and interfacing with the system. The standard defines operation of a serial data bus that interconnects multiple devices via a twisted, shielded pair of wires. The system implements a command-response format.

MIL-STD-1553, "Aircraft Internal Time-Division Command/Response Multiplex Data Bus," has been in use since 1973 and is widely applied. MIL-STD-1553 is referred to as "1553" with the appropriate revision letter (A or B) as a suffix. The basic difference between the 1553A and the 1553B is that in the 1553B, the options are defined rather than being left for the user to define as required. It was found that when the standard did not define an item, there was no coordination in its use. Hardware and software had to be redesigned for each new application. The primary goal of the 1553B was to provide flexibility without creating new designs for each new user. This was accomplished by specifying the electrical interfaces explicitly so that compatibility between designs by different manufacturers could be electrically interchangeable.

The Department of Defense chose multiplexing because of the following advantages:

- Weight reduction
- Simplicity
- Standardization
- Flexibility

Some 1553 applications utilize more than one data bus on a vehicle. This is often done, for example, to isolate a Stores bus from a Communications bus or to construct a bus system capable of interconnecting more terminals than a single bus could accommodate. When multiple buses are used, some terminals may connect to both buses, allowing for communication between them.

MULTIPLEXING

Multiplexing facilitates the transmission of information along the data flow. It permits the transmission of several signal sources through one communications system.

BUS

The bus is made up of twisted-shielded pairs of wires to maintain message integrity. MIL-STD-1553 specifies that all devices in the system will connect to a redundant pair of buses. This provides a second path for bus traffic should one of the buses be damaged. Signals are only allowed to appear on one of the two buses at a time. If a message cannot be completed on one bus, the bus controller may switch to the other bus. In some applications more than one 1553 bus may be implemented on a given vehicle. Some terminals on the bus may actually connect to both buses.

BUS COMPONENTS

There are only three functional modes of terminals allowed on the data bus: the bus controller, the bus monitor, and the remote terminal. Devices may be capable of more than one function. Figure 1 illustrates a typical bus configuration.

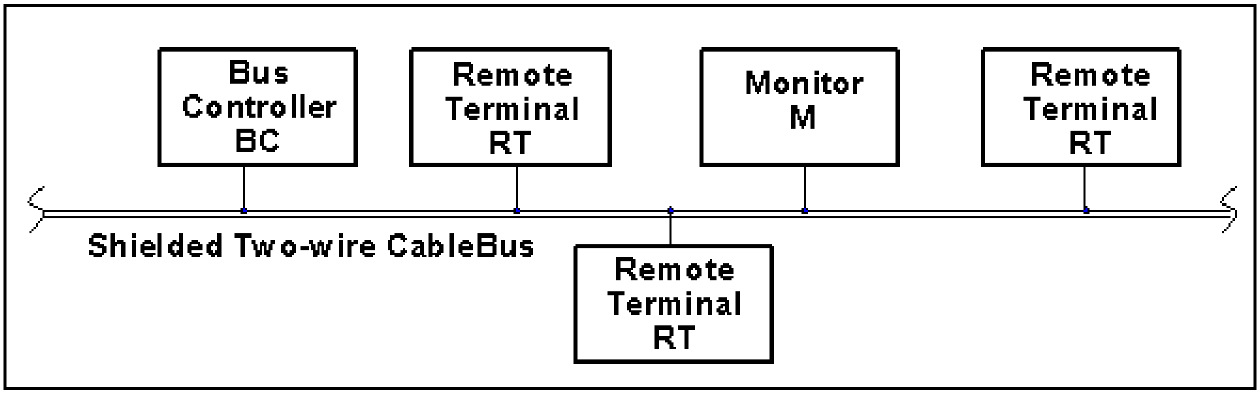

Figure 1. 1553 Bus Structure.

- Bus Controller - The bus controller (BC) is the terminal that initiates information transfers on the data bus. It sends commands to the remote terminals which reply with a response. The bus will support multiple controllers, but only one may be active at a time. Other requirements, according to 1553, are: (1) it is "the key part of the data bus system," and (2) "the sole control of information transmission on the bus shall reside with the bus controller."

- Bus Monitor - 1553 defines the bus monitor as "the terminal assigned the task of receiving bus traffic and extracting selected information to be used at a later time." Bus monitors are frequently used for instrumentation.

- Remote Terminal - Any terminal not operating in either the bus controller or bus monitor mode is operating in the remote terminal (RT) mode. Remote terminals are the largest group of bus components.

MODULATION

The signal is transferred over the data bus using serial digital pulse code modulation.

DATA ENCODING

The type of data encoding used by 1553 is Manchester II biphase.

- A logic one (1) is transmitted as a bipolar coded signal 1/0 (in other words, a positive pulse followed by a negative pulse).
- A logic zero (0) is a bipolar coded signal 0/1 (i.e., a negative pulse followed by a positive pulse).

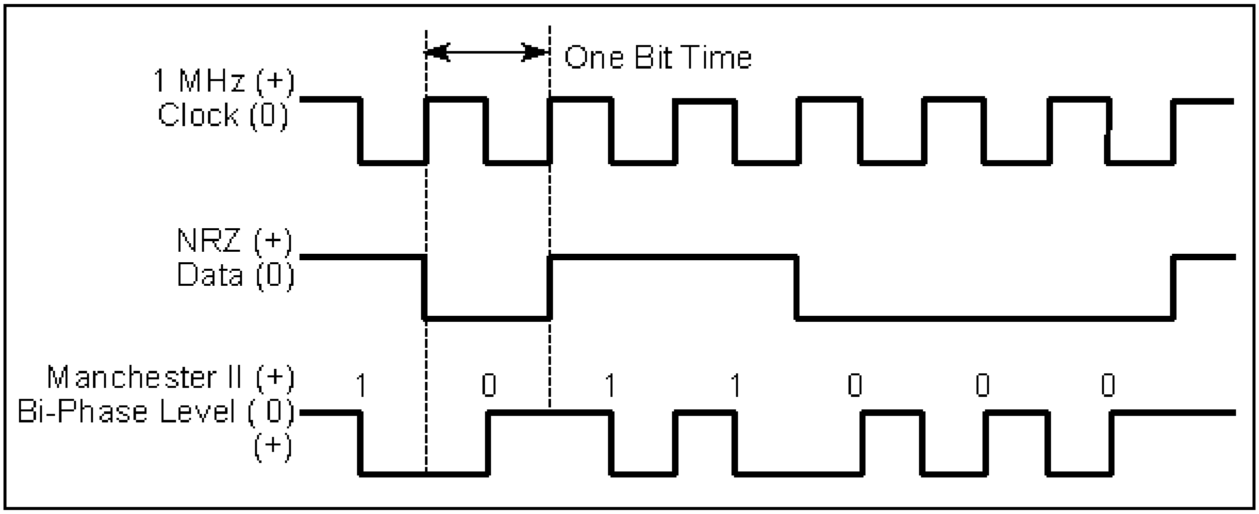

Figure 2. Data Encoding.

A transition through zero occurs at the midpoint of each bit, whether the rate is a logic one or a logic zero. Figure 2 compares a commonly used Non Return to Zero (NRZ) code with the Manchester II biphase level code, in conjunction with a 1 MHz clock.

BIT TRANSMISSION RATE

The bit transmission rate on the bus is 1.0 megabit per second with a combined accuracy and long-term stability of +/- 0.1%. The short-term stability is less than 0.01%.

There are 20 1.0-microsecond bit times allocated for each word. All words include a 3 bit-time sync pattern, a 16-bit data field that is specified differently for each word type, and 1 parity check bit.

WORD FORMATS

Bus traffic or communications travels along the bus in words. A word in MIL-STD-1553 is a sequence of 20 bit times consisting of a 3 bit-time sync wave form, 16 bits of data, and 1 parity check bit. This is the word as it is transmitted on the bus; 1553 terminals add the sync and parity before transmission and remove them during reception. Therefore, the nominal word size is 16 bits, with the most significant bit (MSB) first.

There are three types of words: command, status, and data. A packet is defined to have no inter-message gaps. The time between the last word of a controller message and the return of the terminal status byte is 4-12 microseconds. The time between status byte and the next controller message is undefined. Figure 3 illustrates these three formats.

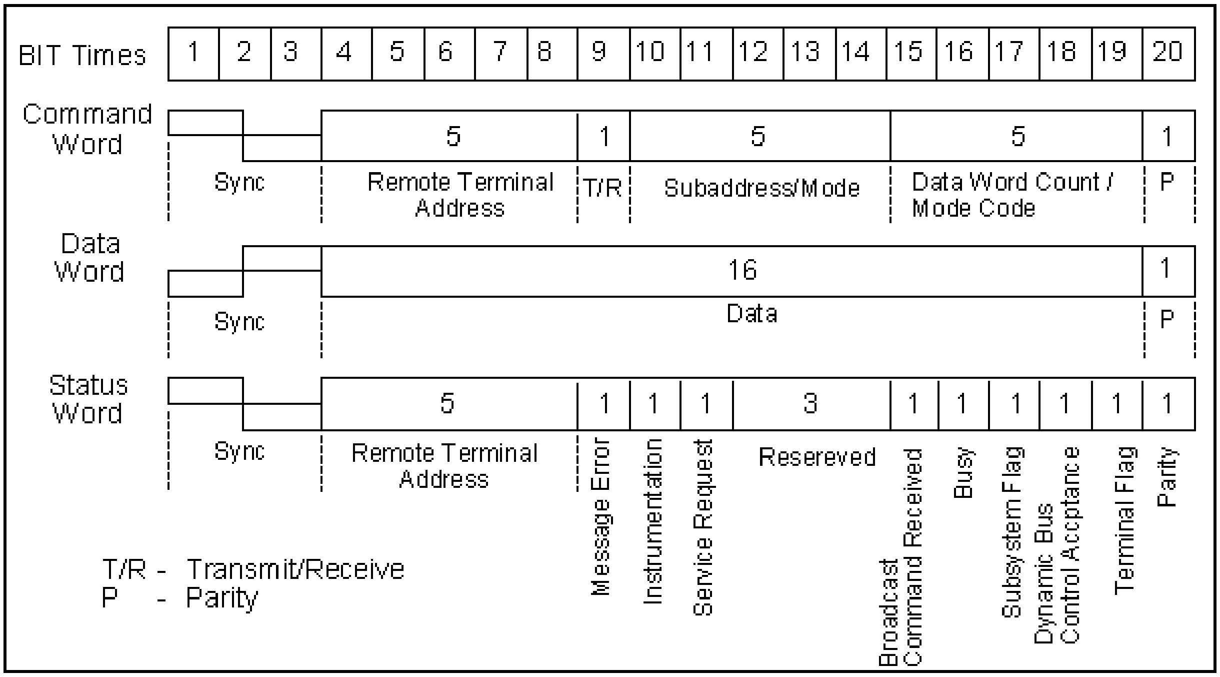
BIT Times
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Command Word
5
1
5
5
1
Sync
Remote Terminal Address
T/R
Subaddress/Mode
Data Word Count / Mode Code
P
Data Word
16
1
Sync
Data
P
Status Word
5
1
1
1
3
1
1
1
1
1
1
Sync
Remote Terminal Address
Message Error
Instrumentation
Service Request
Resereved
Broadcast Command Received
Busy
Subsystem Flag
Dynamic Bus Control Accptance
Terminal Flag
Parity
T/R - Transmit/Receive
P - Parity

Figure 3. 1553 Word Formats.

COMMAND WORD

Command words are transmitted only by the bus controller and always consist of:

- 3 bit-time sync pattern
- 5 bit RT address field
- 1 Transmit/Receive (T/R) field
- 5 bit subaddress/mode field
- 5 bit word count/mode code field
- 1 parity check bit.

DATA WORD

Data words are transmitted either by the BC or by the RT in response to a BC request. The standard allows a maximum of 32 data words to be sent in a packet with a command word before a status response must be returned. Data words always consist of:

- 3 bit-time sync pattern (opposite in polarity from command and status words)
- 16 bit data field
- 1 parity check bit.

STATUS WORD

Status words are transmitted by the RT in response to command messages from the BC and consist of:

- 3 bit-time sync pattern (same as for a command word)
- 5 bit address of the responding RT
- 11 bit status field
- 1 parity check bit.

The 11 bits in the status field are used to notify the BC of the operating condition of the RT and subsystem.

INFORMATION TRANSFERS

Three basic types of information transfers are defined by 1553:

- Bus Controller to Remote Terminal transfers
- Remote Terminal to Bus Controller transfers
- Remote Terminal to Remote Terminal transfers

These transfers are related to the data flow and are referred to as messages. The basic formats of these messages are shown in Figure 4.

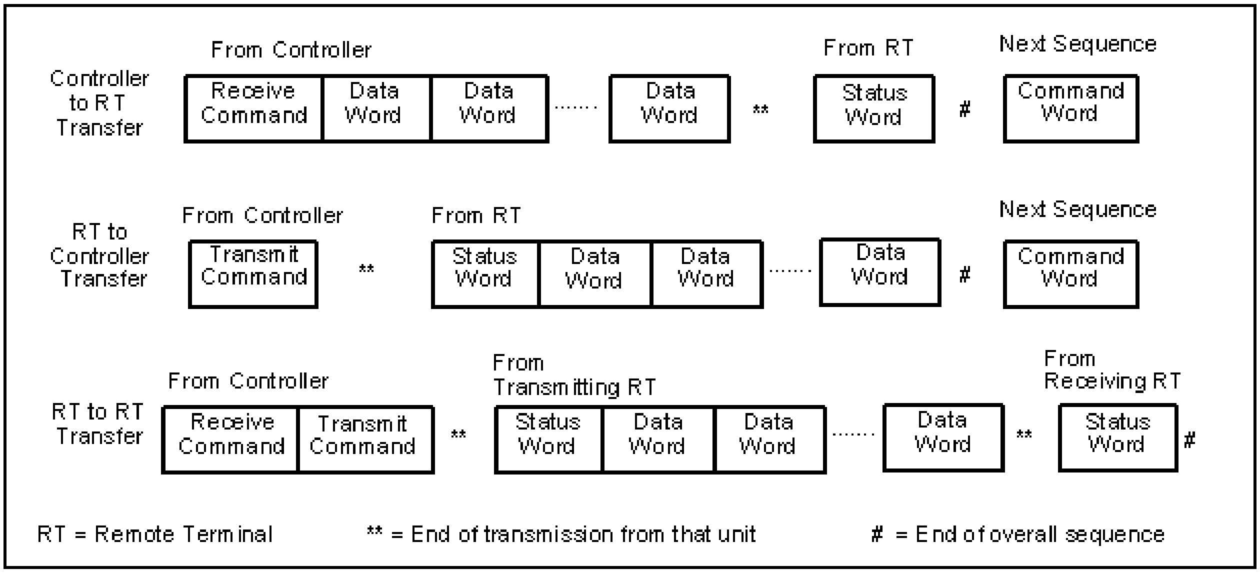
Controller to RT Transfer
From Controller
Receive Command
Data Word
Data Word
....
Data Word
**
From RT
Status Word
#
Next Sequence
Command Word
RT to Controller Transfer
From Controller
Transmit Command
**
From RT
Status Word
Data Word
Data Word
....
Data Word
#
Next Sequence
Command Word
RT to RT Transfer
From Controller
Receive Command
Transmit Command
**
From Transmitting RT
Status Word
Data Word
Data Word
....
Data Word
**
From Receiving RT
Status Word
#
RT = Remote Terminal
** = End of transmission from that unit
# = End of overall sequence

Figure 4. 1553 Data Message Formats.

The normal command/response operation involves the transmission of a command from the BC to a selected RT address. The RT either accepts or transmits data depending on the type (receive/transmit) of command issued by the BC. A status word is transmitted by the RT in response to the BC command if the transmission is received without error and is not illegal.

Figure 5 illustrates the 1553B Bus Architecture in a typical aircraft.

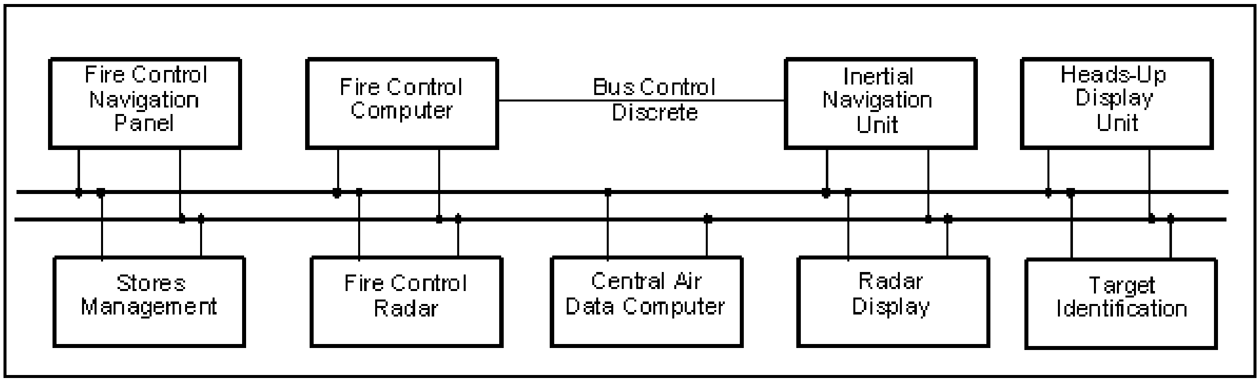
Fire Control Navigation Panel
Fire Control Computer
Bus Control Discrete
Inertial Navigation Unit
Heads-Up Display Unit
Stores Management
Fire Control Radar
Central Air Data Computer
Radar Display
Target Identification

Figure 5. Typical Bus Architecture.

MIL-STD-1773

MIL-STD-1773 contains the requirements for utilizing a fiber optic “cabling” system as a transmission medium for the MIL-STD-1553B bus protocol. As such, the standard repeats MIL-STD-1553 nearly word-for-word. The standard does not specify power levels, noise levels, spectral characteristics, optical wavelength, electrical/optical isolation or means of distributing optical power. These must be contained in separate specifications for each intended use.

Data encoding and word format are identical to MIL-STD-1553, with the exception that pulses are defined as transitions between 0 (off) and 1 (on) rather than between + and - voltage transitions since light cannot have a negative value.

Since the standard applies to cabling only, the bus operates at the same speed as it would utilizing wire. Additionally, data error rate requirements are unchanged.

Different environmental considerations must be given to fiber optic systems. Altitude, humidity, temperature, and age affects fiber optics differently than wire conductors. Power is divided evenly at junctions which branch and connectors have losses just as wire connectors do.

ETHERNET

HISTORY

Ethernet was originally developed by Digital, Intel, and Xerox (DIX) in the early 1970's at the Xerox Palo Alto Research Center (PARC). The two primary inventors were Robert Metcalf and David Boggs.

IEEE Project 802 was set up in February 1980 to provide standards for Local Area Network (LAN) and Metropolitan Area Network (MAN) Architecture. The following committees were established:

- IEEE 802.1: Standards related to network management.
- IEEE 802.2: General standard for the data link layer in the OSI Reference Model. The IEEE divides this layer into two sub-layers -- the logical link control (LLC) layer and the media access control (MAC) layer. The MAC layer varies for different network types and is defined by standards IEEE 802.3 through IEEE 802.5.
- IEEE 802.3: Defines the MAC layer for bus networks that use CSMA/CD. **This is the basis of what is most commonly referred to as the Ethernet standard**.
- IEEE 802.4: Defines the MAC layer for token-bus networks.
- IEEE 802.5: Defines the MAC layer for token-ring networks.
- IEEE 802.6: Standard for Metropolitan Area Networks (MANs).
- IEEE 802.7-9 Inactive Standards:
  - IEEE 802.7: Broadband local area networks. The working group is currently inactive.
  - IEEE 802.8: The Fibre Optic Technical Advisory Group was to create a LAN standard for fiber optic media used in token passing computer networks like FDDI.
  - IEEE 802.9: Integrated Services (IS) LAN Interface at the MAC and Physical Layers.
- IEEE 802.10: A standard for security functions that can be used in both LAN and MAN.
- IEEE 802.11: Wireless LAN MAC and physical layer
- IEEE 802.12: Demand Priority Access Method, Physical Layer and Repeater used in star topology, 100VG-AnyLAN, etc
- IEEE 802.16: Fixed broadband wireless LAN

Ethernet versions 1.0 and 2.0 followed after the original development until the IEEE 802.3 committee altered the structure of the Ethernet II packet to form the Ethernet 802.3 packet. You will currently see either Ethernet II (DIX) format or Ethernet 802.3 format being used.

IEEE 802.3 Ethernet uses Manchester Phase Encoding (MPE) for coding the data bits on the outgoing signal.

INTRODUCTION

Ethernet is a family of frame-based computer networking technologies for local area networks (LANs). The 'Ether' part of Ethernet denotes that the system is not meant to be restricted for use on only one medium type. Copper cables, fiber cables and even radio waves can be used.

Ethernet is standardized as IEEE 802.3. The combination of the twisted pair versions of Ethernet for connecting end systems to the network, along with the fiber optic versions for site backbones, is the most widespread wired LAN technology. It has largely replaced competing LAN standards such as token ring, FDDI, and ARCNET.

Ethernet defines a number of wiring and signaling standards for the Physical Layer of the OSI networking model, through means of network access at the Media Access Control (MAC) / Data Link Layer, and a common addressing format. Therefore, Ethernet defines the lower two layers of the OSI Reference Model (see Data Busses Section intro). Ethernet sometimes implies an attached protocol—such as TCP/IP, however TCP/IP really defines the transport and network layers, respectively, of the OSI model.

Ethernet was designed as a ‘broadcast’ system, i.e. stations on the network can send messages whenever and wherever they want. All stations may receive the messages, however only the specific station to which the message is directed responds. To handle simultaneous demands, the system uses Carrier Sense Multiple Access with Collision Detection (CSMA/CD).

CARRIER SENSE MULTIPLE ACCESS WITH COLLISION DETECTION (CSMA/CD)

When an Ethernet station is ready to transmit, it checks for the presence of a signal on the cable. If no signal is present then the station begins transmission. However, if a signal is present then the station delays transmission until the cable is not in use. If two stations detect an idle cable and at the same time transmit data, then a collision occurs. On a star-wired, unshielded twisted pair (UTP) network, if the transceiver of a sending station detects activity on both its receive and transmit pairs before it has completed transmitting, then it decides that a collision has occurred. On a coaxial system, a collision is detected when the DC signal level on the cable is the same or greater than the combined signal level of the two transmitters. Line voltage drops dramatically if two stations transmit at the same time and the first station to notice this sends a high voltage jamming signal around the network as a signal. The two stations involved with the collision quit transmitting again for a random time interval.

A Collision Domain is that part of the network where each station can ‘see’ other stations’ traffic both unicast and broadcasts. The Collision Domain is made up of one segment of Ethernet coax (with or without repeaters) or a number of UTP shared hubs. A network is segmented with bridges that create two segments, or two Collision Domains where a station on one segment cannot see traffic between stations on the other segment unless the packets are destined for itself. If it can still see all broadcasts on the segmented network(s), no matter the number of segments, it is a Broadcast Domain.

ETHERNET FRAME

Figure 1 depicts the structure of the older DIX (Ethernet II) and the now standard 802.3 Ethernet frames. The numbers above each field represent the number of bytes.

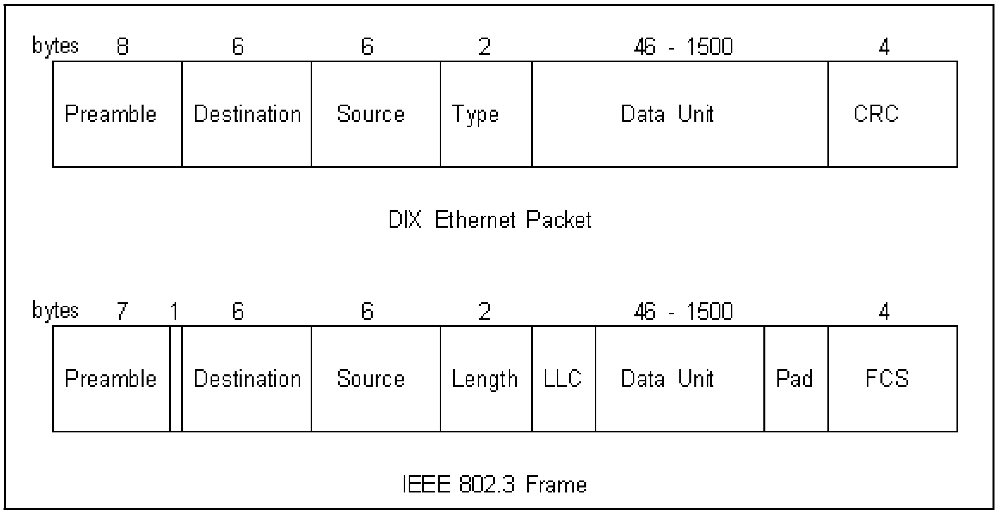

Figure 1. Ethernet Frame.

- **Preamble field:** Establishes bit synchronization and transceiver conditions so that the circuitry synchs with the received frame timing. The DIX frame has 8 bytes for the preamble rather than 7, as it does not have a Start Frame Delimiter (or Start of Frame).
- **Start Frame Delimiter:** Sequence 10101011 in a separate field, only in the 802.3 frame.
- **Destination address:** Hardware address (MAC address) of the destination station (usually 48 bits i.e. 6 bytes).
- **Source address:** Hardware address of the source station (must be of the same length as the destination address, the 802.3 standard allows for 2 or 6 byte addresses, although 2 byte addresses are never used, N.B. Ethernet II *only* uses 6 byte addresses).
- **Type:** Specifies the protocol sending the packet such as IP or IPX (only applies to DIX frame).
- **Length:** Specifies the length of the data segment, actually the number of LLC data bytes, (only applies to 802.3 frame and replaces the Type field).
- **LLC:** Logical Length Control is a data communication protocol layer which provides multiplexing and flow control mechanisms that make it possible for several network protocols (IP, IPX) to coexist within a multipoint network and to be transported over the same network media.
- **Data:** Actual data which is allowed anywhere between 46 to 1500 bytes within one frame.
- **Pad:** Zeros added to the data field to ‘Pad’ a short data field to 46 bytes (applies to 802.3 frame).
- **CRC:** Cyclic Redundancy Check to detect errors during transmission (DIX version of FCS).
- **FCS:** Frame Check Sequence to detect errors that occur during transmission (802.3 version of CRC). This 32-bit code has an algorithm applied to it, which will give the same result as the other end of the link, provided that the frame was transmitted successfully.

From the above we can see that the maximum 802.3 frame size is 1518 bytes and the minimum size is 64 bytes. Packets that have correct CRCs (or FCSs) but are smaller than 64 bytes are known as ‘Runts’.

The hardware address, or MAC address is transmitted and stored in Ethernet network devices in **Canonical** format i.e. Least Significant Bit (LSB) first. The expression **Little-Endian** is used to describe the LSB format in which Ethernet is transmitted. On the other hand, Token Ring networks and the subset, Fiber

Distributed Data Interface (FDDI) networks, transmit the MAC address with the Most Significant Bit (MSB) first, or **Big-Endian**, This is known as **Non-Canonical** format. Note that this applies on a byte-by-byte basis i.e. the bytes are transmitted in the same order, it is just the bits in each of those bytes that are reversed. The storage of the MAC addresses in Token Ring and FDDI devices however, may sometimes still be in Canonical format so this can sometimes cause confusion. The reference to, the distribution of MAC addresses and the Organizationally Unique Identifier (OUI) destinations are always in Canonical format.

The Logical Link Control (LLC) protocol data unit (PDU) is defined in IEEE 802.2 and operates with 802.3 Ethernet as seen in Figure 2.

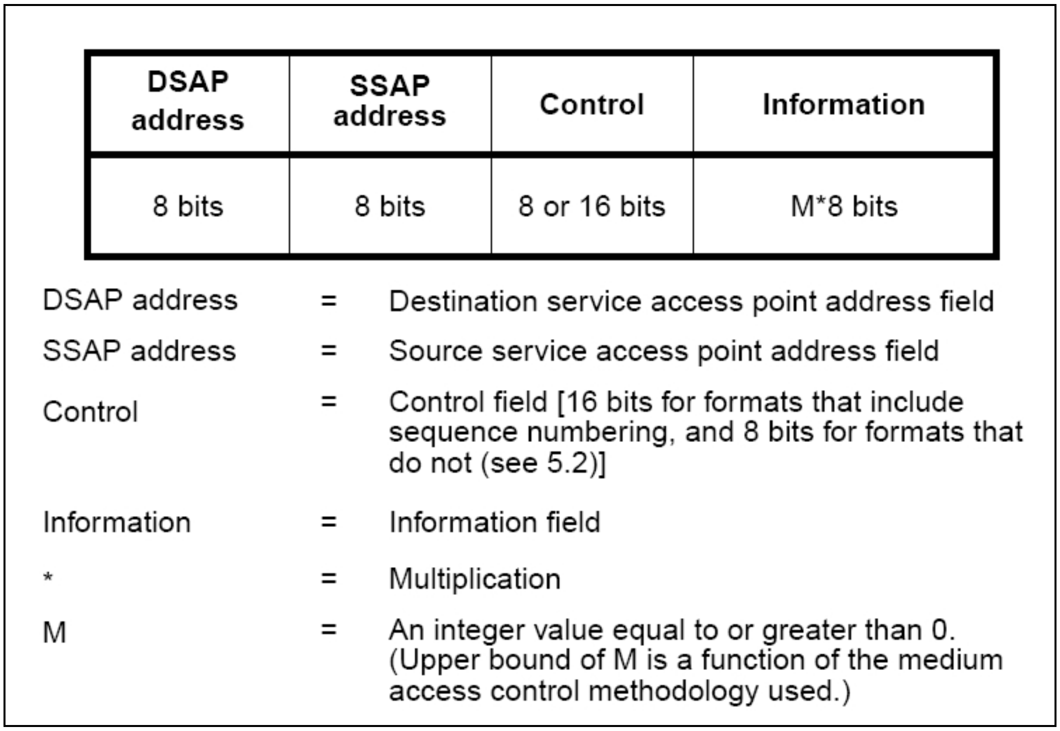

DSAP address	SSAP address	Control	Information
8 bits	8 bits	8 or 16 bits	M*8 bits

DSAP address = Destination service access point address field
SSAP address = Source service access point address field
Control = Control field [16 bits for formats that include sequence numbering, and 8 bits for formats that do not (see 5.2)]
Information = Information field
* = Multiplication
M = An integer value equal to or greater than 0. (Upper bound of M is a function of the medium access control methodology used.)

Figure 2. LLC Format.

LLC is based on the High-Level Data Link Control (HDLC) data protocol format. Whereas Ethernet II (2.0) combines the MAC and the Data link layers restricting itself to connectionless service in the process, IEEE 802.3 separates out the MAC and Data Link layers. IEEE 802.2 is also required by Token Ring and FDDI but cannot be used with the Novell 'Raw' format. There are three types of LLC; Type 1, which is connectionless, Type 2 which is connection-oriented, and Type 3 for Acknowledged Connections.

The **Service Access Point (SAP)** is used to distinguish between different data exchanges on the same end station and basically replaces the Type field for the older Ethernet II frame. The **Source Service Access Point (SSAP)** indicates the service from which the LLC data unit is sent, and the **Destination Service Access Point (DSAP)** indicates the service to which the LLC data unit is being sent. As examples, NetBIOS uses the SAP address of **F0** whilst IP uses the SAP address of **06**. The following lists some common SAPs:

00 - Null LSAP
02 - Individual LLC Sublayer Management Function
03 - Group LLC Sublayer Management Function
05 - IBM SNA Path Control (group)

06 - Internet Protocol (IP)
18 - Texas Instruments
42 - IEEE 802.1 Bridge Spanning Tree Protocol
80 - Xerox Network Systems (XNS)
86 - Nestar
AA - SubNetwork Access Protocol (SNAP)
E0 - Novell NetWare
F0 - IBM NetBIOS
F5 - IBM LAN Management (group)
FE - ISO Network Layer Protocol

The **Control Field** identifies the type of LLC, of which there are three:

- **Type 1** - Data (PDUs) shall be exchanged between LLCs without the need for the establishment of a data link connection. These PDUs shall not be acknowledged, nor shall there be any flow control or error recovery. This is called **Unsequenced Information (UI)**.
- **Type 2** – Needs data link connection. Uses **Information (I)** frames and maintains the sequence numbers during an acknowledged connection-oriented transmission.
- **Type 3** – No data link connection required. Uses **Acknowledged Connection (AC)** frames in an acknowledged connectionless service.

MAC ADDRESS

With an Ethernet MAC address, the first octet uses only the lowest significant bit as the Individual/Group (I/G) bit address identifier. If it is zero it indicates the address is individual. If it is one, it indicates a group address. The Universally or Locally administered (U/L) address bit is the bit of octet 0 adjacent to the I/G address bit. This bit indicates whether the address has been assigned by a local or universal administrator. Universally administered addresses have this bit set to 0. If this bit is set to 1, the entire address (i.e., 48 bits) has been locally administered.

The first 3 octets of the MAC address form the Organizational Unique Identifier (OUI) assigned by IEEE to organizations that requires their own group of MAC addresses.

SUBNETWORK ACCESS PROTOCOL (SNAP)

The SNAP protocol was introduced to allow an easy transition to the new LLC frame format for vendors. SNAP allows older frames and protocols to be encapsulated in a Type 1 LLC header making any protocol ‘pseudo-IEEE compliant’. SNAP is described in RFC 1042. Figure 3 shows how it looks:

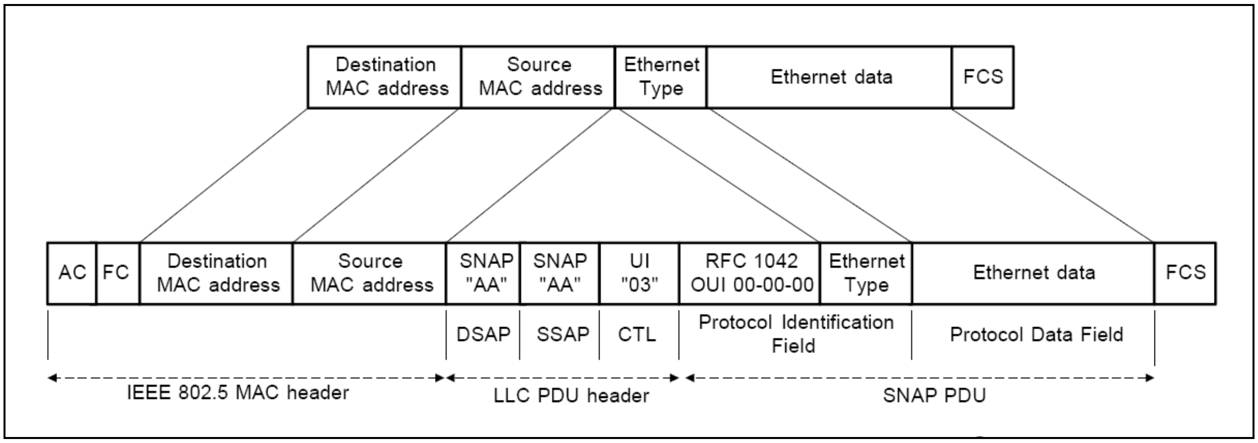
Destination MAC address
Source MAC address
Ethernet Type
Ethernet data
FCS
AC
FC
Destination MAC address
Source MAC address
SNAP "AA"
SNAP "AA"
UI "03"
RFC 1042 OUI 00-00-00
Ethernet Type
Ethernet data
FCS
DSAP
SSAP
CTL
Protocol Identification Field
Protocol Data Field
IEEE 802.5 MAC header
LLC PDU header
SNAP PDU

Figure 3. Ethernet Frame SNAP-encapsulated in IEEE 802.5 Frame.

It can be seen that it is an LLC data unit (sometimes called a Logical Protocol Data Unit (LPDU)) of Type 1 (indicated by 03). The DSAP and SSAP are set to AA to indicate that this is a SNAP header coming up. The SNAP header then indicates the vender via the Organizational Unique Identifier (OUI) and the protocol type via the Ethertype field. In the example above the OUI is 00-00-00 which means that there is an Ethernet frame. An ethernet type (Ethertype) of 08-00 would indicate Internet IP is the protocol. The official list of types can be found at IEEE. Although vendors tend to be are moving to LLC1 on the LAN, SNAP still remains.

TRANSMISSION MEDIA

10Base-5 - largely obsolete, but may still be in use.

Traditionally, Ethernet is used over ‘thick’ coaxial cable called 10Base-5 (the ‘10’ denotes 10Mbps, base means that the signal is baseband i.e. takes the whole bandwidth of the cable (so that only one device can transmit at one time on the same cable), and the ‘5’ denotes 500 m maximum length). The minimum length between stations is 2.5 m.

The cable is run in one long length forming a ‘Bus Topology’. Stations attach to it by way of inline N-type connections or a transceiver, which is literally screwed into the cable (by way of a ‘Vampire Tap’) providing a 15-pin Attachment Unit Interface (AUI) connection (also known as a DIX connector or a DB-15 connector) for a drop lead connection (maximum of 50 m length) to the station. The segments are terminated with 50 ohm resistors and the shield grounded at one end only.

The segment could be appended with up to a maximum of 4 repeaters; therefore 5 segments (total length of 2,460 m) can be connected together. Of the 5 segments only 3 can have devices attached (100 per segment). This is known as the 5-4-3 rule. A total of 300 devices can be attached on a Thicknet broadcast domain.

10Base-2 – dominant for many years

It was common to see the Thick coax used in Risers to connect Repeaters, which in turn provide ‘Thin Ethernet’ coaxial connections for runs up to 30 workstations. Thin Ethernet (Thinnet) uses RG-58 cable and is called 10Base-2 (The ‘2’ now denoting 200 m maximum length - strictly speaking this is 185 m). The minimum length between stations is 0.5 m.

Each station connects to the thinnet by way of a Network Interface Card (NIC), which provides a BNC (British Naval Connector). At each station the thinnet terminates at a T-piece and at each end of the thinnet run (or ‘Segment’) a 50-ohm terminator is required to absorb stray signals, thereby preventing signal bounce. The shield is grounded at one end only.

A segment can be appended with other segments using up to 4 repeaters, i.e. 5 segments in total. Two of these segments however, cannot be tapped; they can only be used for extending the length of the broadcast domain (to 925 m). What this means is that 3 segments with a maximum of 30 stations on each can give you 90 devices on a Thinnet broadcast domain.

10Base-T – Most widespread today

The most common media used in of Ethernet today involves use of Unshielded Twisted Pair (UTP) or Shielded Twisted Pair (STP). For instance, Category 5 UTP can be installed in a ‘Star-wired’ format, with runs recommended at no greater than 100 m (including patch leads, cable run and flyleads) and Ethernet Hubs with UTP ports (RJ45) centrally located. It has been found though that runs of up to 150 m are feasible, the limitations being signal strength. Also, there should be no more than a 11.5 dB signal loss and the minimum distance between devices is 2.5 m. The maximum delay for the signal in a 10Mbps network is 51.2 microseconds. This comes from the fact that the bit time (time to transmit one bit) is 0.1 microseconds and that the slot time for a frame is 512 bit times.

In order to connect to Ethernet in this ‘Star Topology’, each station has a NIC, which contains an RJ45 socket which is used by a 4-pair droplead to connect to another nearby RJ45 socket.

Each port on the hub sends a ‘Link Beat Signal’ which checks the integrity of the cable and devices attached, a flickering LED on the front of the port of the hub tells you that the link is running fine. The maximum number of hubs (or, more strictly speaking, repeater counts) that you can have in one segment is four and the maximum number of stations on one broadcast domain is 1024.

The advantages of the UTP/STP technology are gained from the flexibility of the system, with respect to moves, changes, fault finding, reliability, and security. The following table shows the RJ45 pin-outs for 10Base-T:

RJ45 Pin	Function	Color
1	Transmit	White/Orange stripe
2	Transmit	Orange
3	Receive	White/Green stripe
4		Blue
5		White/Blue stripe
6	Receive	Green
7		White/Brown stripe
8		Brown

If you wish to connect hub to hub, or a NIC directly to another NIC, then the following 10Base-T crossover cable should be used:

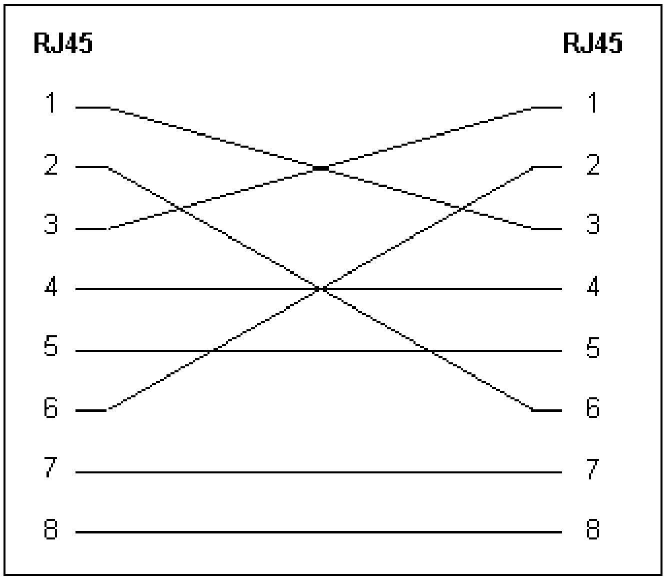

Figure 4. 10Base-T Crossover.

The 4-repeater limit manifests itself in 10/100Base-T environments where the active hub/switch port is in fact a repeater, hence the name multi-port repeater. Generally, the hub would only have one station per port but you can cascade hubs from one another up to the 4-repeater limit. The danger here is that you will have all the traffic from a particular hub being fed into one port so care would need to be taken on noting the applications being used by the stations involved, and the likely bandwidth that the applications will use.

There is a pre-standard called Lattisnet (developed by Synoptics), which runs 10 MHz Ethernet over twisted pair but instead of bit synchronization occurring at the sending (as in 10Base-T) the synchronization occurs at the receiving end.

10Base-F

The 10Base-F standard defines the use of fiber for Ethernet. 10Base-FL describes the standards for the fiber optic links between stations and repeaters, allowing up to 2 km per segment on multi-mode fiber. It replaces the 1987 FOIRL (Fiber Optic Inter-Repeater Link), which provides the specification for a fiber optic Media Attachment Unit (MAU) and other interconnecting components. Two related standards are essentially "dead": 10Base-FB allows up to 2 km per segment (on multi-mode fiber) and is designed for backbone applications such as cascading repeaters. In addition, there is the 10Base-FP (Passive components) standard

The 10Base-F standard allows for 1024 devices per network.

Fast Ethernet - 100Base-T

Fast Ethernet is the most popular of the newer standards and is an extension to 10Base-T. The '100' denotes 100 Mbps data speed and it uses the same two pairs as 10Base-T and must only be used on Category 5 UTP cable installations with provision for it to be used on Type 1 STP. The actual data throughput increases by between 3 to 4 times that of 10Base-T.

Fast Ethernet is specified in 802.3u and uses the same frame formats and CSMA/CD technology as normal 10Mbps Ethernet. The difference is that the maximum delay for the signal across the segment is 5.12 microseconds instead of 51.2 microseconds. This stems from the fact that the bit time (time to transmit one bit) is 0.01 microseconds and that the slot time for a frame is 512 bit times. The Inter-Packet Gap (IPG) for 802.3u is 0.96 microseconds as opposed to 9.6 microseconds for 10Mbps Ethernet.

Whereas 10Base-T uses Normal Link Pulses (NLP) for testing the integrity of the connection, 100Base-T uses Fast Link Pulses (FLP) which are backwardly compatible with NLPs but contain more information. FLPs are used to detect the speed of the network (e.g. in 10/100 switchable cards and ports). The distance must not exceed 100 m.

The IEEE uses the term 100Base-X to refer to both 100Base-Tx and 100Base-Fx and the Media-Independent Interface (MII) allows a generic connector for transceivers to connect to 100Base-Tx, 100Base-Fx and 100Base-T4 LANs.

There is no such thing as the 5-4-3 rule in Fast Ethernet. All 10Base-T repeaters are considered to be functionally identical. Fast Ethernet repeaters are divided into two classes of repeater: Class I and Class II. A Class I repeater has a repeater propagation delay value of 140 bit times, whilst a Class II repeater is 92 bit times. The Class I repeater (or Translational Repeater) can support different signaling types such as 100Base-Tx and 100Base-T4. A Class I repeater transmits or repeats the incoming line signals on one port to the other ports by first translating them to digital signals and then retranslating them to line signals. The translations are necessary when connecting different physical media (media conforming to more than one physical layer specification) to the same collision domain. Any repeater with an MII port would be a Class I device. Only one Class I repeater can exist within a single collision domain, so this type of repeater cannot be cascaded.

A Class II repeater immediately transmits or repeats the incoming line signals on one port to the other ports: it does not perform any translations. This repeater type connects identical media to the same collision domain (for example, TX to TX). At most, two Class II repeaters can exist within a single collision domain.

The cable used to cascade the two devices is called and unpopulated segment or IRL (Inter-Repeater Link). The Class II repeater (or **Transparent Repeater**) can only support one type of physical signaling.

100Base-T4

100Base-T4 was an early implementation of Fast Ethernet. It requires four twisted copper pairs, but those pairs were only required to be category 3 rather than the category 5 required by TX. One pair is reserved for transmit, one for receive, and the remaining two will switch direction as negotiated. A very unusual 8B6T code is used to convert 8 data bits into 6 base-3 digits (the signal shaping is possible as there are three times as many 6-digit base-3 numbers as there are 8-digit base-2 numbers). The two resulting 3-digit base-3 symbols are sent in parallel over 3 pairs using 3-level pulse-amplitude modulation (PAM-3).

100Base-T2

This little known version of Fast Ethernet is for use over two pairs of Category 3 cable and uses PAM-5 for encoding. There is simultaneous transmission and reception of data in both pairs and the electronics uses DSP technology to handle alien signals in adjacent pairs. 100Base-T2 can run up to 100 m on Category 3 UTP.

100Base-Fx

100Base-Fx is a version of Fast Ethernet over optical fiber. It uses a 1300 nm near-infrared (NIR) light wavelength transmitted via two strands of optical fiber, one for receive (RX) and the other for transmit (TX). Maximum length is 400 meters (1,310 ft) for half-duplex connections (to ensure collisions are detected) or 2 kilometers (6,600 ft) for full-duplex over multimode optical fiber. Longer distances are possible when using single-mode optical fiber. 100Base-Fx uses the same 4B5B encoding and NRZI line code that 100Base-TX does. 100Base-Fx should use the following connectors: SC, straight tip (ST), or media independent connectors (MIC), with SC being the preferred option. 100Base-Fx is not compatible with 10Base-Fl, the 10 MBit/s version over optical fiber.

100Base-Sx

100Base-Sx is a version of Fast Ethernet over optical fiber. It uses two strands of multi-mode optical fiber for receive and transmit. It is a lower cost alternative to using 100Base-Fx, because it uses short wavelength optics (850 nm) which are significantly less expensive than the long wavelength optics used in 100Base-Fx. 100Base-Sx can operate at distances up to 300 meters (980 ft).

100Base-Sx uses the same wavelength as 10Base-Fl, the 10 MBit/s version over optical fiber. Unlike 100Base-Fx, this allows 100Base-Sx to be backwards-compatible with 10Base-Fl. Because of the shorter wavelength used and the shorter distance it can support, 100Base-Sx uses less expensive optical components (LEDs instead of lasers) which makes it an attractive option for those upgrading from 10Base-Fl and those who do not require long distances.

100Base-Bx

100Base-Bx is a version of Fast Ethernet over a single strand of optical fiber (unlike 100Base-Fx, which uses a pair of fibers). Single-mode fiber is used, along with a special multiplexer which splits the signal into transmit and receive wavelengths. The two wavelengths used for transmit and receive are either 1310/1550 nm or 1310/1490 nm. Distances can be 10, 20, or 40 km.

100VG-AnyLAN

Based on 802.12 (Hewlett Packard), 100VG-AnyLAN uses an access method called Demand Priority. The 'VG' stands for 'Voice Grade' as it is designed to be used with Category 3 cable. This is where the repeaters (hubs) carry out continuous searches around all of the nodes for those that wish to send data. If two devices cause a 'contention' by wanting to send at the same time, the highest priority request is dealt with first, unless the priorities are the same, in which case both requests are dealt with at the same time (by alternating frames). The hub only knows about connected devices and other repeaters so communication is only directed at them rather than broadcast to every device in the broadcast domain (which could mean 100's of devices!). This is a more efficient use of the bandwidth. This is the reason why the new standard was developed as it is not strictly Ethernet. Standard 802.12 is designed to better support both Ethernet and Token Ring.

The encoding techniques used are 5B/6B and NRZ. All four pairs of UTP are used. On Cat3 the longest cable run is 100 m but this increases to 200 m on Cat5. The clock rate on each wire is 30 MHz; therefore 30 Mbits per second are transmitted on each pair giving a total bit rate of 120 Mbps. Since each 6-bits of data on the line represents 5 bits of real data due to the 5B/6B encoding, the rate of real data being transmitted is 25 Mbps on each pair, giving a total real data rate of 100 Mbps. For 2-pair STP and fiber, the bit rate is 120 Mbps on the transmitting pair, for a real data transmission rate of 100 Mbps.

Gigabit Ethernet

Although the functional principles of Gigabit Ethernet are the same as Ethernet and Fast Ethernet i.e. CSMA/CD and the Framing format, the physical setup is very different. One difference is the slot time. The standard Ethernet slot time required in CSMA/CD half-duplex mode is not long enough for running over 100 m of copper, so Carrier Extension is used to guarantee a 512-bit slot time.

1000Base-X (802.3z)

The 802.3z committee is responsible for formalizing the standard for Gigabit Ethernet. The 1000 refers to 1 Gbps data speed. The existing Fiber Channel interface standard (ANSI X3T11) is used and allows up to 4.268 Gbps speeds. The Fiber Channel encoding scheme is 8B/10B.

Gigabit Ethernet can operate in half or full duplex modes and there is also a standard 802.3x which manages XON/XOFF flow control in full duplex mode. With 802.3x, a receiving station can send a packet to a sending station to stop it sending data until a specified time interval has passed.

There are three media types for 1000Base-X. 1000Base-Sx, 1000Base-Lx, and 1000Base-Cx.

1000Base-Sx (‘S’ is for Short Haul) uses short-wavelength laser (850 nm) over multi-mode fiber. 1000Base-Sx can run up to 300 m on 62.5/125 µm multimode fiber and up to 550 m on 50/125 µm multimode fiber.

1000Base-Lx (‘L’ is for Long Haul) uses long wavelength laser (1300 nm) and can run up to 550 m on 62.5/125 µm multi-mode fiber or 50/125 µm multi-mode fiber. In addition, 1000Base-Lx can run up to 5 km (originally 3 km) on single-mode fiber using 1310 nm wavelength laser.

1000Base-Cx is a standard for STP copper cable and allows operation up to 25 m over STP cable.

1000Base-T (802.3ab)

Many cable manufacturers are enhancing their cable systems to ‘enhanced Category 5’ (Cat 5e) standards in order to allow Gigabit Ethernet to run at up to 100 m on copper. The Category 6 standard has yet to be ratified, and is not being pursued very strongly.

In order to obtain the 1000 Mbps data bit rate across the UTP cable without breaking the FCC rules for emission, all 4 pairs of the cable are used. Hybrid circuits at each end of each pair are used to allow simultaneous transmission and reception of data (full-duplex) by separating the transmission signal from the receiving signal. Because some transmission signal still manages to couple itself to the receiving side there is an additional echo canceller built in, this is called a Near End Crosstalk (NEXT) canceller.

Encoding is carried out with PAM-5.

TRANSMISSION CONTROL PROTOCOL / INTERNET PROTOCOL

Transmission **C**ontrol **P**rotocol / **I**nternet **P**rotocol (TCP/IP) is the communication protocol for the Internet.

Whereas IP handles lower-level transmissions from computer to computer as a message makes its way across the Internet, TCP operates at a higher level, concerned only with the two end systems, for example a Web browser and a Web server. In particular, TCP provides ordered delivery of a stream of bytes from a program on one computer to another program on another computer. Besides the Web, other common applications of TCP include e-mail and Internet file transfer like FTP. Among its other management tasks, TCP controls message size, the rate at which messages are exchanged, and network traffic congestion. Your computer Internet address is a part of the standard TCP/IP protocol and so is your domain name.

Inside the TCP/IP standard there are several protocols for handling data communication:

- TCP (Transmission Control Protocol) communication between applications
- UDP (User Datagram Protocol) simple communication between applications
- IP (Internet Protocol) communication between computers
- ICMP (Internet Control Message Protocol) for errors and statistics
- DHCP (Dynamic Host Configuration Protocol) for dynamic addressing

TCP: If one application wants to communicate with another via TCP, it sends a communication request. This request must be sent to an exact address. TCP is responsible for breaking data down into IP packets before they are sent, and for assembling the packets when they arrive. After a “handshake” between the two applications, TCP will set up a “full-duplex” communication between the two applications. The “full-duplex” communication will occupy the communication line between the two computers until it is closed by one of the two applications. UDP is very similar to TCP, but simpler and less reliable.

IP: IP is responsible for sending the packets to the correct destination. IP does not occupy the communication line between two computers and hence reduces the need for network lines. Each line can be used for communication between many different computers at the same time. With IP, messages (or other data) are broken up into small independent “packets” and sent between computers via the Internet. IP is responsible for “routing” each packet to the correct destination.

When an IP packet is sent from a computer, it arrives at an IP router. The IP router is responsible for “routing” the packet to the correct destination, directly or via another router. The path the packet will follow might be different from other packets in the same communication chain. The router is responsible for the right addressing, depending on traffic volume, errors in the network, or other parameters.

Each computer must have a unique IP address before it can connect to the Internet. Each IP packet must also have an address before it can be sent to another computer. TCP/IP uses 32 bits (4 bytes), of four numbers between 0 and 255, separated by a period, to address a computer. A typical IP address is of the format: 192.255.30.50

Domain Names: A name is much easier to remember than a 12 digit number. The names used for TCP/IP addresses are called domain names. “**navy.mil**” is a domain name. Subdomains are identified by prefixes to the main domain name (separated by a period). When you address a web site, like **http://www.navair.navy.mil** the name is translated to a number by a Domain Name System (DNS) Server. DNS servers are connected to the Internet all over the world. When a new domain name is registered together with a TCP/IP address, all DNS servers are updated with this information.

BROWSER AND OTHER APPLICATION LAYER PROTOCOLS

DHCP - Dynamic Host Configuration Protocol. DHCP is used for allocation of dynamic IP addresses to computers in a network.

FTP - File Transfer Protocol. FTP takes care of transmission of data files between computers.

HTTP - Hyper Text Transfer Protocol - takes care of the communication between a web server and a web browser. It is used for sending requests from a web client (a browser) to a web server, returning web content (web pages) from the server back to the client.

HTTPS - Secure HTTP. HTTPS typically handles credit card transactions and other sensitive data.

IMAP - Internet Message Access Protocol. The IMAP protocol is used by email programs (like Microsoft Outlook) just like the POP protocol (discussed below). The main difference between the IMAP protocol and the POP protocol is that the IMAP protocol will not automatically download all your emails each time your email program connects to your email server. The IMAP protocol allows you to look through your email messages at the email server before you download them. With IMAP you can choose to download your messages or just delete them. This way IMAP is useful if you need to connect to your email server from different locations, but only want to download your messages when you are back in your office.

NTP - Network Time Protocol. NTP is used to synchronize the time (the clock) between computers.

POP - Post Office Protocol. The POP protocol is used by email programs (like Microsoft Outlook) to retrieve emails from an email server. If your email program (also called email client), uses POP, all your emails are downloaded to your computer each time it connects to your email server.

SMTP - Simple Mail Transfer Protocol. The SMTP protocol is used for the transmission of e-mails to another computer. Normally your email is sent to an email server (SMTP server), and then to another server or servers, and finally to its destination. SMTP can only transmit pure text. It cannot transmit binary data like pictures, sounds, or movies directly. SMTP uses the MIME protocol to send binary data across TCP/IP networks.

SNMP - Simple Network Management Protocol. SNMP is used for administration of computer networks.

PRESENTATION LAYER PROTOCOLS

MIME - Multi-purpose Internet Mail Extensions protocol. Converts binary data to pure text. The MIME protocol lets SMTP transmit multimedia files including voice, audio, and binary data.

SSL - Secure Sockets Layer. The SSL protocol is used for encryption of data for secure data transmission.

LOWER LAYER PROTOCOLS

ARP - Address Resolution Protocol. ARP is used by IP to find the hardware address of a computer network card based on the IP address.

BOOTP - Boot Protocol. BOOTP is used for booting (starting) computers from the network.

ICMP - Internet Control Message Protocol. ICMP takes care of error-handling in the network.

LDAP - Lightweight Directory Access Protocol. LDAP is used for collecting information about users and e-mail addresses from the internet.

PPTP - Point to Point Tunneling Protocol. PPTP is used for setting up a connection (tunnel) between private networks.

RARP - Reverse Address Resolution Protocol. RARP is used by IP to find the IP address based on the hardware address of a computer network card.

This page intentionally left blank.

GLOSSARY

ACCEPTABLE DEGRADATION - The allowable reduction in system performance. For a fire control radar, the acceptable degradation is usually expressed as a reduction in range; for example, the maximum lock-on range might be degraded by 25 percent without loss of essential defense capability.

ACQUISITION - A procedure by which a fire control tracking radar attains initial lock-on. Usually, the approximate target coordinates are supplied to the tracking radar and it searches a predetermined volume of space to locate the target.

AEROSOLS - Solid particles dispersed in the atmosphere having resonant size particles with a high index of refraction. The particles both scatter and absorb visual and laser directed energy so as to cut down on weapon systems directed by these techniques.

AFC (AUTOMATIC FREQUENCY CONTROL) - An arrangement whereby the frequency of an oscillator or the tuning of a circuit is automatically maintained within specified limits with respect to a reference frequency. A magnetron drifts in frequency over a period of time. The AFC of a radar makes the local oscillator shift by an equal amount so the IF frequency will remain constant.

AGC (AUTOMATIC GAIN CONTROL) - A method for automatically obtaining an essentially constant receiver output amplitude. The amplitude of the received signal in the range gate determines the AGC bias (a DC voltage) which controls the receiver gain so as to maintain a nearly constant output even though the amplitude of the input signal changes.

AMPLIFIER - An electronic device used to increase signal magnitude or power. See also GaAs FET Amplifier, Klystron Amplifier, Traveling-Wave Tube Amplifier.

AMPLITUDE MODULATION (AM) - A method of impressing a message upon a carrier signal by causing the carrier amplitude to vary proportionally to the message waveform.

AMPLITUDE SHIFT KEYING (ASK) - A method of impressing a digital signal upon a carrier signal by causing the carrier amplitude to take different values corresponding to the different values of the digital signal.

ANGLE JAMMING - ECM technique, when azimuth and elevation information from a scanning fire control radar is jammed by transmitting a jamming pulse similar to the radar pulse, but with modulation information out of phase with the returning target angle modulation information.

ANGULAR SEPARATION - This term is frequently used to indicate a protective (from EMI) zone for a missile. The interfering antenna axis must be separated, throughout the critical portion of the missile flight, from the missile by the specified angle. The vertex of the angle is at the interference source antenna.

ANTENNA BEAMWIDTH - The angle, in degrees, between the half-power points (-3 dB) of an antenna beam. This angle is also nearly that between the center of the mainlobe and the first null. The angle is given for both horizontal and vertical planes unless the beam is circular. When so indicated, the term may refer to the angular width of the mainlobe between first nulls [beamwidth between first nulls (BWFN)]. See also Antenna Pattern. The figure illustrates vertical profile for antenna displaying a 10-degree beamwidth characteristic. The values can vary dramatically with frequency.

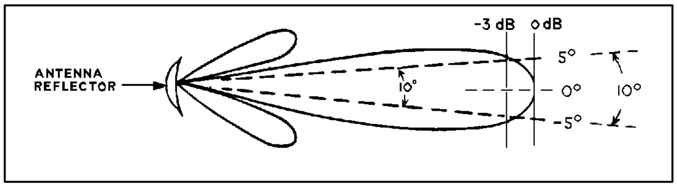

ANTENNA CROSS TALK - A measure of undesired power transfer through space from one antenna to another. Ratio of power received by one antenna to power transmitted by the other, usually expressed in decibels.

ANTENNA ISOLATION - The ratio of the power input to one antenna to the power received by the other. It can also be viewed as the insertion loss from transmit antenna input to receive antenna output to circuitry.

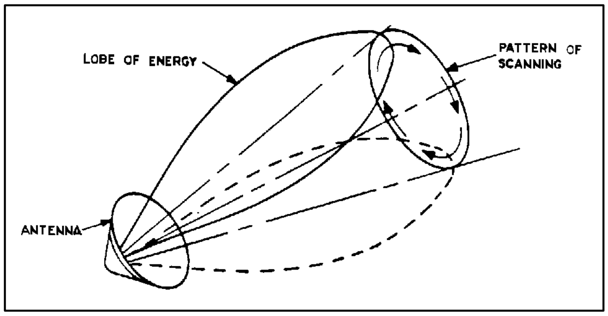

ANTENNA LOBING - Two lobes are created that overlap and intercept at -1 to -3 dB. The difference between the two lobes produces much greater spatial selectivity than provided by either lobe alone. (See also Lobe, Antenna.)

ANTENNA NUTATING - An antenna, as used in automatic-tracking radar systems, consisting of a parabolic reflector combined with a radiating element which is caused to move in a small circular orbit about the focus of the antenna with or without change of polarization. The radiation pattern is in the form of a beam that traces out a cone centered on the reflector axis. The process is also known as nutating conical scanning.

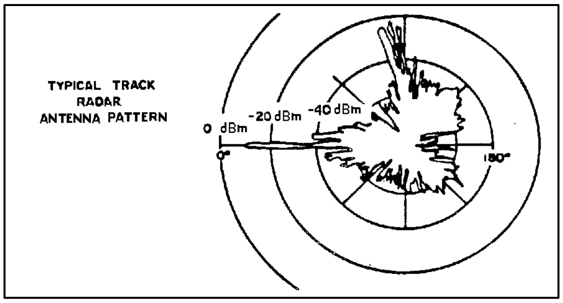

ANTENNA PATTERN - A cross section of the radiating pattern (representing antenna gain or loss) in any plane that includes the origin (source reference point) of the pattern. Both horizontal and vertical polar plots are normally used to describe the pattern. Also, termed “polar diagram” and “radiation pattern.”

ANTENNA, PENCIL-BEAM - A highly directional antenna designed that cross sections of the major lobe are approximately circular, with a narrow beamwidth.

ANTI-CLUTTER CIRCUITS (IN RADAR) - Circuits which attenuate undesired reflections to permit detection of targets otherwise obscured by such reflections.

APERTURE - In an antenna, that portion of the plane surface area near the antenna perpendicular to the direction of maximum radiation through which the major portion of the radiation passes. The effective and/or scattering aperture area can be computed for wire antennas which have no obvious physical area.

A-SCOPE - A cathode-ray oscilloscope used in radar systems to display vertically the signal amplitude as a function of time (range) or range rate. Sometimes referred to as Range (R)-Scope.

ASYNCHRONOUS PULSED JAMMING - An effective form of pulsed jamming. The jammer nearly matches the pulse repetition frequency (PRF) of the radar; then it transmits multiples of the PRF. It is more effective if the jammer pulsewidth is greater than that of the radar. Asynchronous pulsed jamming is similar to synchronous jamming except that the target lines tend to curve inward or outward slightly and appear fuzzy in the jammed sector of a radar scope.

ATTENUATION - Decrease in magnitude of current, voltage, or power of a signal in transmission between two points. May be expressed in decibels.

AUTOMATIC FREQUENCY CONTROL - See AFC.

AUTOMATIC GAIN CONTROL - See AGC.

BACKWARD WAVE OSCILLATOR (BWO) - A cross-field device in which an electron stream interacts with a backward wave on a nonreentrant circuit. This oscillator may be electronically tuned over a wide range of frequencies, is relatively unaffected by load variations and is stable. BWO is commonly pronounced "be woe".

BALANCED MIXERS - The two most frequently encountered mixer types are single-balanced and double-balanced. In a double-balanced mixer, four Schottky diodes and two wideband transformers are employed to provide isolation of all three ports.

BALLISTIC MISSILE - Any missile which does not rely upon aerodynamic surfaces to produce lift and consequently follows a ballistic trajectory when thrust is terminated.

BANDPASS FILTER - A type of frequency discrimination network designed to pass a band or range of frequencies and produce attenuation to all other frequencies outside of the pass region. The figure illustrates a typical bandpass filter, incorporating a bandpass region of (F_h)-(F_l), offering no rejection (0 dB) to desired signal (F_m) and much higher rejection to the adjacent undesired signals F_h, and F_l. The upper and lower frequencies are usually specified to be the half power (-3 dB) or half voltage points (-6 dB).

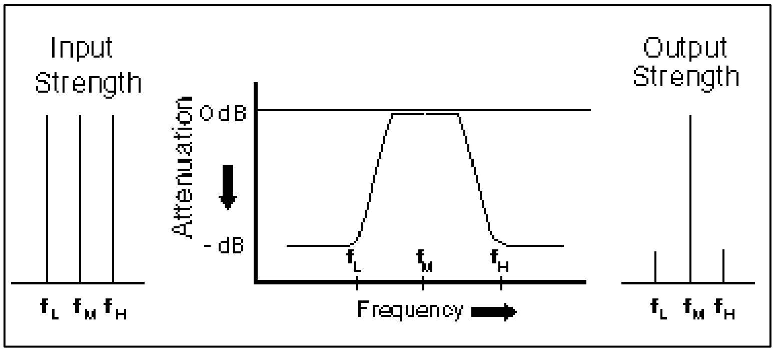

BANDWIDTH - An expression used to define the actual operational frequency range of a receiver when it is tuned to a certain frequency. For a radar receiver, it is the difference between the two frequencies at which the receiver response is reduced to some fraction of its maximum response (such as 3 dB, 6 dB, or some other specified level). The frequencies between which "satisfactory" performance is achieved. Two equations are used:

$$\textit{Narrowband by \%}\ \left(\frac{F_u - F_l}{F_c}\right)(100) \ ; \quad \textit{Broadband by ratio}\ \frac{F_u}{F_l}$$

$$\textit{Where } F_u = \textit{Upper} \ ; F_l = \textit{lower} \ ; F_c = \textit{center} = (F_u + F_l) \div 2$$

See also Receiver Bandwidth and Spectrum Width.

BARRAGE NOISE JAMMING - Noise jamming spread in frequency to deny the use of multiple radar frequencies to effectively deny range information. Although this is attractive because it enables one jammer to simultaneously jam several radars of different frequencies, it does have the inherent problem that the wider the jamming spread, the less jamming power available per radar, i.e. the watts per MHz bandwidth is low.

BATTERY, MISSILE - A missile battery consists of a missile launcher and its associated missile fire control systems (such as a MK 11 MOD 0 Missile Launcher and two MK 74 MOD 0 Missile Fire Control Systems).

BEACON - A system wherein a transponder in a missile receives coded signals from a shipboard radar guidance transmitter and transmits reply signals to a shipboard radar beacon receiver to enable a computer to determine missile position. The missile beacon transmitter and shipboard radar beacon receiver are tuned to a frequency different from that of the guidance transmitter.

BEAM - See Lobe, antenna. The beam is to the side of an aircraft or ship.

BEAM, CAPTURE - See Capture Beam.

BEAM-TO-BEAM CORRELATION (BBC) - BBC is used by frequency scan radars to reject pulse jamming and jamming at a swept frequency. Correlation is made from two adjacent beams (pulses). The receiver rejects those targets (signals) that do not occur at the same place in two adjacent beams.

BEAMWIDTH - See Antenna Beamwidth.

BEAT FREQUENCY OSCILLATOR (BFO) - Any oscillator whose output is intended to be mixed with another signal to produce a sum or difference beat frequency. Used particularly in reception of CW transmissions.

BINGO - The fuel state at which an aircraft must leave the area in order to return and land safely. Also used when chaff/flares reach a preset low quantity and automatic dispensing is inhibited.

BIPOLAR VIDEO - Unrectified (pre-detection) IF (both positive and negative portions of the RF envelope) signals that arise from the type of detection and console display employed in pulse Doppler and MTI receivers.

BISTATIC RADAR - A radar using antennas at different locations for transmission and reception.

BLANKING - The process of making a channel, or device non-effective for a certain interval. Used for retrace sweeps on CRTs or to mask unwanted signals such as blanking ones own radar from the onboard RWR.

BOGEY - Unknown air target.

BURN-THROUGH RANGE - The ability of a radar to see through jamming. Usually, described as the point when the radar's target return is a specified amount stronger than the jamming signal (typical values are 6 dB manual and 20 dB automatic). See Section 4-8.

BUTT LINE - Line used for reference in measurement of left/right location. One of several aircraft references. See also fuselage station and water line.

CAPTURE BEAM - A wide beam incorporated in capture transmitters of beam rider (command guided) missile systems to facilitate gaining initial control of a missile immediately after launch. Upon capture, the system then centers the missile in the narrow guidance beam. The figure illustrates a launched missile at point of capture.

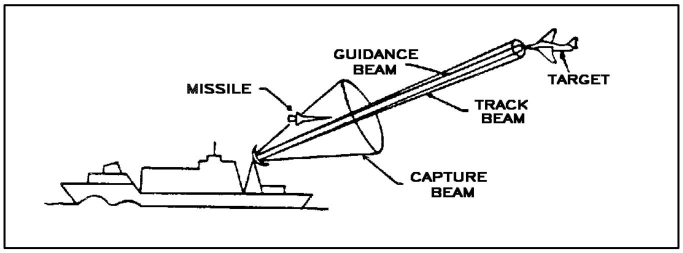

CAPTURE TRANSMITTER - A transmitter employing a wide beam antenna to gain initial control of in-flight missile for the purpose of centering the missile in the guidance transmitter antenna beam. See also Capture Beam.

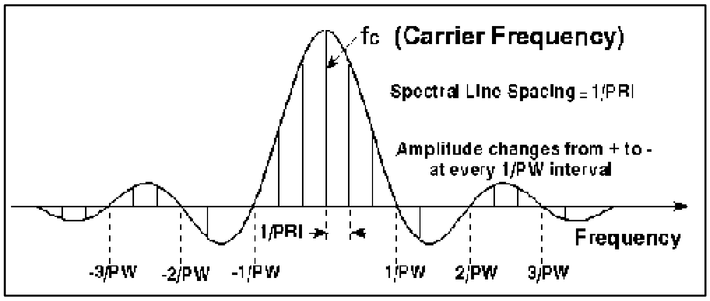

CARRIER FREQUENCY - The basic radio frequency of the wave upon which modulations are impressed. Also called “Carrier” or f_c. See figure at right.

CATCH-22 - A lose-lose situation, from the book of the same name.

CAVITY - A space enclosed by a conducting surface used as a resonant circuit at microwave frequencies. Cavity space geometry determines the resonant frequency. A storage area for oscillating electromagnetic energy.

CENTER FREQUENCY - The tuned or operating frequency. Also referred to as center operating frequency. In frequency diversity systems, the midband frequency of the operating range. See also Carrier Frequency.

CHAFF - Ribbon-like pieces of metallic materials or metalized plastic which are dispensed by aircraft or ships to mask or screen other “targets.” The radar reflections off the chaff may cause a tracking radar to break lock on the target. The foil materials are generally cut into small pieces for which the size is dependent upon the radar interrogation frequency (approximately 1/2 wave length of the victim radar frequency). Being this length, chaff acts as a resonant dipole and reflects much of the energy back to the radar. Also see rainbow, rope, stream chaff, and window.

CHANNEL - A frequency or band of frequencies. In guided missile systems, an assigned center frequency and a fixed bandwidth around it. Designates operating frequency of track radars and frequency/code assignments of X-band CW illuminators.

CHIRP - A pulse compression technique which uses frequency modulation (usually linear) on pulse transmission.

CHIRP RADAR - See PC.

CIRCULARLY POLARIZED JAMMING - The techniques of radiating jamming energy in both planes of polarization simultaneously. With this method, there is a loss of 3 dB of effective power in either linear plane, and substantial loss if the opposite sense of circular polarization is used (i.e. left vs. right). See Section 3-2.

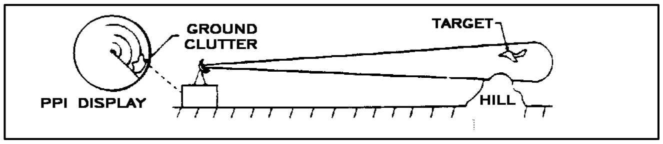

CLUTTER, RADAR - Undesired radar returns or echoes resulting from man-made or natural objects including chaff, sea, ground, and rain, which interfere with normal radar system observations. The figure illustrates a target being masked by ground clutter

CO-CHANNEL - This term is used to indicate that two (or more) equipments are operating on the same frequency.

COHERENT - Two signals that have a set (usually fixed) phase relationship.

COINCIDENCE DETECTOR - This radar video process requires more than one hit in a range cell before a target is displayed. This prevents video interference from pulses coming from another radar, because such interference is unlikely to occur twice in the same range cell.

COLLIMATION - The procedure of aligning fire control radar system antenna axes with optical line of sight, thereby ensuring that the radars will provide for correct target illumination and guidance beam positioning.

COMMAND CODE - Modulations superimposed upon transmitter carrier signals to provide electronic instructions to an airborne guided missile or pilotless aircraft. The receiver of the remotely guided vehicle is preset to accept only a selected transmitter code to eliminate the possibility of the vehicle responding to commands of extraneous signals. Missile command codes include instructions such as arm, warhead detonate, and self destruct.

COMMAND GUIDANCE - A guidance system wherein intelligence transmitted to the missile from an outside source causes the missile to traverse a directed flight path.

CONICAL SCAN - See Antenna, Nutating.

CONTINUOUS WAVE and CONTINUOUS WAVE ILLUMINATOR - See CW and CWI.

COOPERATIVE COUNTERMEASURES - (CO-OP) Generic term for jamming the same threat radar from two or more separate platforms that are in the same radar resolution cell.

COUPLING FACTOR - A multiplying factor, expressed in dB, used to express the change in EM energy intensity from a radar transmitter to a receiver. The factor includes the antenna gains and the loss (basic transmission loss) caused by the distance between the antennas. The factor will usually be a negative dB figure (a reduction in intensity) because basic transmission loss is always a large negative value. The antenna gains may be positive (pointed toward each other) or negative (no main beam interactions).

CROSS MODULATION - Intermodulation caused by modulation of the carrier by an undesired signal wave.

CROSS POLARIZATION - or “Cross Pole,” is a monopulse jamming technique where a cross-polarized signal is transmitted to give erroneous angle data to the radar. The component of the jamming signal with the same polarization as the radar must be very small.

CW (CONTINUOUS WAVE) - In radar and EW systems this term means that the transmitter is on constantly; i.e., not pulsed (100% duty cycle). These systems may frequency or phase modulate the transmitter output. A CW radar has the ability to distinguish moving targets against a stationary background while conserving spectrum bandwidth compared to pulsed radar requirements. A CW radar extracts accurate target range-rate data but cannot determine target range.

CWI (CONTINUOUS WAVE ILLUMINATOR) - A surface or aircraft-based CW transmitter employed in semiactive homing missile systems where the transmitter illuminates the target and the missile senses the reflected energy. The transmitter also provides a reference signal to the missile rear receiver to allow determination of range-rate data and target identification. CW transmitter emissions are sometimes coded corresponding to the associated missile receiver codes to reduce the possibility of the “missile accepting commands of extraneous signals.

DECIBEL (dB) - A dimensionless unit for expressing the ratio of two values of power, current, or voltage. The number of decibels being equal to: $dB = 10 \log P_2/P_1 = 20 \log V_2/V_1 = 20 \log I_2/I_1$ Normally, used for expressing transmission gains, losses, levels, and similar quantities. See Section 2-4.

DECEPTION - The deliberate radiation, reradiation, alteration, absorption or reflection of electromagnetic energy in a manner intended to mislead the enemy interpretation or use of information received by his electronic systems.

dB - See Decibel, or Decibel Section 2-4.

dBc - Decibels referenced to the carrier signal.

dBi - Decibels referenced to an isotropic radiator. (dBLi indicating linear isotropic radiator is sometimes used).

dBm - Decibels relative to 1 mW. dBm is calculated by using the ratio of some power (expressed in mW) to 1 mW. For example, 1 mW is 0 dBm and 10 mW is +10 dBm.

dBsm - Decibel referenced to one square meter.

dBv / dBµv - Decibels referenced to one volt or microvolt, i.e. 0 dBv is 1 volt or 120 dBµv.

dBW / dBµW - Decibels referenced to 1 watt or one microwatt, i.e. 0 dBW is 1 watt or 30 dBm or 60 dBµW.

DEMODULATOR - A device employed to separate the modulation signal from its associated carrier, also called Second Detector. See also Detection.

DESIGNATION - The assignment of a fire control radar to a specific target by supplying target coordinate data to the radar system.

DETECTION - Usually refers to the technique of recovering the amplitude modulation signal (envelope) superimposed on a carrier. See figure at right.

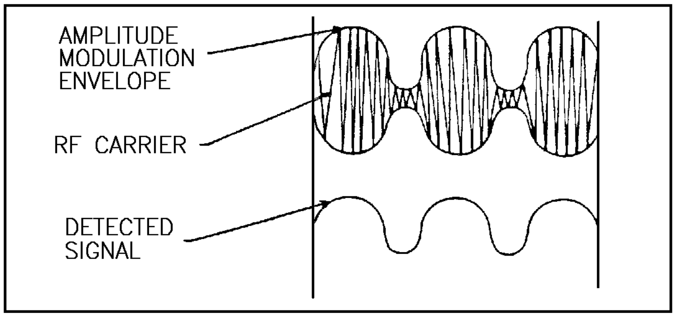

DICKE FIX - This type of radar processing occurs in the IF amplifier. A limiter follows a wideband amplifier, and then the signal goes to a matched filter amplifier. This discriminates against pulses that are too long (clutter) or too short (interference). The “DICKE FIX” is a technique that is specifically designed to protect the receiver from ringing caused by noise, fast-sweep, or narrow pulse jamming. The basic configuration consists of a broadband limiting IF amplifier, followed by an IF amplifier of optimum bandwidth. The limit level is preset at approximately the peak amplitude of receiver noise, the bandwidth may vary from 10 to 20 MHz, depending on the jamming environment. This device provides excellent discrimination against fast sweep jamming (10-500 MHz), usually something on the order of 20 to 40 dB, without appreciable loss in sensitivity. However, strong CW jamming will seriously degrade the performance of the DICKE FIX because the CW signal captures the radar signal in the limiter.

DIELECTRICALLY STABILIZED OSCILLATOR - The DSO uses a dielectric resonator as the frequency determining element. When the dielectric material is properly selected and used, the variations in dielectric constant vs. temperature and the dimensions of the resonant structure vs. temperature tend to cancel out, providing relatively good frequency vs. temperature stability. The DSO offers frequency accuracy and stability, low power consumption and high reliability. Some of the commonly used materials are barium, zirconium, or tin tinates. The composition of these materials may be controlled to achieve any frequency variation with temperature with close tolerances.

DIODE - An electronic device which restricts current flow chiefly to one direction. See also Gunn diode, IMPATT diode, PIN diode, point contact diode, Schottky barrier diode, step recovery diode, tunnel diode, varactor diode.

DIODE SWITCH - PIN-diode switches provide state-of-the-art switching in most present-day microwave receivers. These switches are either reflective or nonreflective in that the former reflect incident power back to the source when in the isolated state. While both types of switches can provide high isolation and short transition times, the reflective switch offers multi octave bandwidth in the all shunt diode configuration, while the non-reflective switch offers an octave bandwidth.

DIPLEX - The simultaneous transmission or reception of two signals using a common feature such as a single antenna or carrier. Typically, two transmitters operate alternately at approximately the same RF and using a common antenna. See Section 6-7 for a discussion of diplexers.

DIRECTIONAL COUPLER - A 4-port transmission coupling device used to sample the power traveling in one direction through the main line of the device. There is considerable isolation (typically 20 dB) to signals traveling in the reverse direction. Because they are reciprocal, the directional coupler can also be used to directively combine signals with good reverse isolation. The directional coupler is implemented in waveguide and coaxial configurations. See Section 6-4.

DIRECTIVITY - For antennas, directivity is the maximum value of gain in a particular direction. (Isotropic point source has directivity = 1). For directional couplers, directivity is a measure (in dB) of how good the directional coupling is and is equal to the isolation minus the coupling. See Section 6-4.

DISH - A microwave reflector used as part of a radar antenna system. The surface is concave and is usually parabolic shaped. Also called a parabolic reflector.

DOPPLER EFFECT - The apparent change in frequency of an electromagnetic wave caused by a change in distance between the transmitter and the receiver during transmission/reception. The figure illustrates the Doppler increase that would be realized by comparing the signal received from a target approaching the radar site to the transmitted reference signal. An apparent frequency decrease would be noted for targets departing the radar location. Differences can be calibrated to provide target range-rate data.

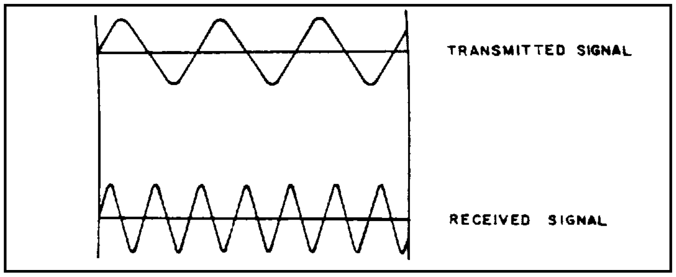

DRY RUN - A test run with aircraft/ship armament and/or EW switches off.

DUCTING - The increase in range that an electromagnetic wave will travel due to a temperature inversion of the atmosphere. The temperature inversion forms a channel or waveguide (duct) for the waves to travel in, and they can be trapped, not attenuating as would be expected from the radar equation.

DUMMY LOAD (Radio Transmission) - A dissipative but essentially nonradiating substitute device having impedance characteristics simulating those of the antenna. This allows power to be applied to the radar transmitter without radiating into free space. Dummy loads are commonly used during EMCON conditions or when troubleshooting a transmitter at a workbench away from its normal environment.

DUPLEXER - A switching device used in radar to permit alternate use of the same antenna for both transmitting and receiving.

DUTY CYCLE - The ratio of average power to peak power, or ratio of pulse length to interpulse period for pulsed transmitter systems. Interpulse period is equal to the reciprocal of the pulse repetition rate.
See Section 2-5.

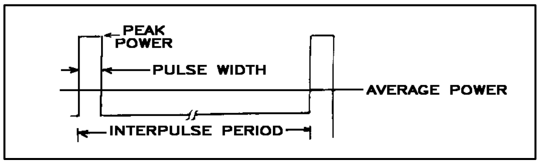

The duty cycle of a radar having a pulse length of 0.3 μsec and a PRF of 2000 pulses/sec is computed as follows:

Interpulse Period, T = PRI = 1/PRF = 500 μsec

$$\text{Duty Cycle} = \frac{\text{Pulse length}}{\text{Interpulse Period}} = \frac{0.3\ \mu\text{sec}}{500\ \text{musec}} = 0.0006\ (\text{or } 0.06\%)$$ or Duty Cycle (dB) = 10log(Duty cycle) = -32.2 dB

An output tube providing an average power of only 90 watts for such a system would, therefore, provide a peak power of:

$$\text{Peak Power} = \frac{\text{Average Power}}{\text{Duty Cycle}} = \frac{90}{0.0006} = 150{,}000\ W \text{ or } 52\ dBW \text{ or } 82\ dBm$$

EFFECTIVE RADIATED POWER (ERP) - Input power to the antenna in watts times the gain ratio of the antenna. When expressed in dB, ERP is the transmitter power (P_T), in dBm (or dBW) plus the antenna gain (G_T) in dB. The term EIRP is used sometimes and reiterates that the gain is relative to an isotropic radiator.

EGRESS - Exit the target area.

ELECTROMAGNETIC COUPLING - The transfer of electromagnetic energy from one circuit or system to another circuit or system. An undesired transfer is termed EMI (electromagnetic interference).

EMC (ELECTROMAGNETIC COMPATIBILITY) - That condition in which electrical/electronic systems can perform their intended function without experiencing degradation from, or causing degradation to other electrical/electronic systems. More simply stated, EMC is that condition which exists in the absence of EMI. See also Intersystem and Intrasystem EMC tests.

EME (ELECTROMAGNETIC ENVIRONMENT) - The total electromagnetic energy in the RF spectrum that exists at any given location.

EMI (ELECTROMAGNETIC INTERFERENCE) - Any induced, radiated, or conducted electrical emission, disturbance, or transient that causes undesirable responses, degradation in performance, or malfunctions of any electrical or electronic equipment, device, or system. Also synonymously referred to as RFI (Radio Frequency Interference).

EMI MODEL - Usually a set of equations or logical concepts designed to illustrate the interactions, the detailed parameters considerations, and mathematical procedures necessary for proper analysis of a given EMI situation.

EMITTER - Any device or apparatus which emits electromagnetic energy.

EMP (ELECTROMAGNETIC PULSE) - The generation and radiation of a very narrow and very high-amplitude pulse of electromagnetic noise. It is associated with the high level pulse as a result of a nuclear detonation and with intentionally generated narrow, high-amplitude pulse for ECM applications. In the case of nuclear detonations, EMP consists of a continuous spectrum with most of its energy distributed through the low frequency band of 3 KHz to 1 MHz.

ERROR SIGNAL - In servomechanisms, the signal applied to the control circuit that indicates the degree of misalignment between the controlling and the controlled members. In tracking radar systems, a voltage dependent upon the signal received from a target whose polarity and magnitude depend on the angle between the target and the center axis of the scanning beam.

FAST TIME CONSTANT - See FTC.

FEET DRY / WET - Aircraft has crossed from water to shore / aircraft has crossed from shore to water.

FERRET - An aircraft, ship, or vehicle especially equipped for the detection, location, recording, and analyzing of electromagnetic radiations.

FIELD STRENGTH - The magnitude of a magnetic or electric field at any point, usually expressed in terms of ampere turns per meter or volts per meter. Sometimes called field intensity and is expressed in volts/meter or dBμv/meter. Above 100 MHz, power density terminology is used more often. See Section 4-1.

FIRST HARMONIC - The fundamental (original) frequency.

FREQUENCY AGILITY - A radar's ability to change frequency within its operating band, usually on a pulse-to-pulse basis. This is an ECCM technique employed to avoid spot jamming and to force the jammer to go into a less effective barrage mode.

FREQUENCY AGILITY RADAR - A radar that automatically or semiautomatically tunes through a discrete set of operating frequencies in its normal mode of operation.

FREQUENCY DIVERSITY RADAR - A radar system technique, employed primarily as an antijamming feature, where the transmitter output frequency varies randomly from pulse to pulse over a wide frequency range.

FREQUENCY RANGE - (1) A specifically designated portion of the frequency spectrum; (2) of a device, the band of frequencies over which the device may be considered useful with various circuit and operating conditions; (3) of a transmission system, the frequency band in which the system is able to transmit power without attenuating or distorting it more than a specified amount.

FREQUENCY SHIFT KEYING (FSK) - A form of FM where the carrier is shifted between two frequencies in accordance with a predetermined code. In multiple FSK, the carrier is shifted to more than two frequencies. FSK is used principally with teletype communications.

"FRUIT" - In a radar beacon system, there is a type of interference called "FRUIT," caused by beacon replies to interrogation asynchronous with the observer's interrogator. The largest amount of this interference is received through the sidelobes of the interrogating antenna, but it can become dense enough to cause false target indications.

FTC (FAST TIME CONSTANT) - An antijam feature employed in radar systems where receiver circuits may be selected to provide a short time constant to emphasize signals of short duration to produce discrimination against the low frequency components of clutter.

FUNDAMENTAL FREQUENCY - Used synonymously for tuned frequency, carrier frequency, center frequency, output frequency, or operating frequency.

FUSELAGE STATION or just STATION - A reference point (usually the nose of an aircraft) used to measure or identify fore and aft locations. One of several aircraft location designations - also see butt line and water line.

GaAs FET AMPLIFIER - Because of their low noise, field-effect transistors are often used as the input stage of wideband amplifiers. Their high input resistance makes this device particularly useful in a variety of applications. Since the FET does not employ minority current carriers, carrier storage effects are eliminated giving the device faster operating characteristics and improved radiation resistant qualities.

GAIN - For antennas, the value of power gain in a given direction relative to an isotropic point source radiating equally in all directions. Frequently expressed in dB (gain of an isotropic source = 0 dB). The formula for calculating gain is:

$$G = \frac{4\pi\, P(\theta,\phi)}{P_{in}} \quad ;\ where$$

$P(\theta,\phi)$ = *Radiation intensity in given direction*

P_{in} = *Power into lossless antenna radiating uniformly in all directions*

Note: (1) If radiation efficiency is unity, then gain = directivity i.e. if directivity = 2, then gain = 3 dB, etc.
(2) Interference losses within an array also affect gain
(3) See Section 3-1 for further details

For amplifiers, gain is the ratio of the output power to input power (usually in dB).

GATE (RANGE) - A signal used to select radar echoes corresponding to a very short range increment. Range is computed by moving the range gate or marker to the target echo; an arrangement which permits radar signals to be received in a small selected fraction of the time period between radar transmitter pulses.

GATING - (1) The process of selecting those portions of a wave which exist during one or more selected time intervals; (2) the application of a square waveform of desired duration and timing to perform electronic switching; (3) the application of receiver operating voltages to one or more stages only during that portion of a cycle of operation when reception is desired. See also Gate (Range).

GCI (GROUND-CONTROLLED INTERCEPT) - Vectoring an interceptor aircraft to an airborne target by means of information relayed from a ground-based radar site which observes both the interceptor and the target.

GIGA - A prefix meaning 10^9 (times a billion). For example, gigahertz (GHz).

GLINT (In Radar) - 1. The random component of target location error caused by variations in the phase front of the target signal (as contrasted with Scintillation Error). Glint may affect angle, range of

Doppler measurements, and may have peak values corresponding to locations beyond the true target extent in the measured coordinate. 2. Electronic countermeasures that uses the scintillating, or flashing effect of shuttered or rotating reflectors to degrade tracking or seeking functions of an enemy weapons system.

GUARDBAND - A frequency band to which no other emitters are assigned as a precaution against interference to equipments susceptible to EMI in that band.

GUIDANCE, BEAM RIDER - A missile guidance technique which is dependent on the missile's ability to determine its positions with reference to the center of scan of the guidance radar beam and thus correct its trajectory on the basis of detected errors.

GUIDANCE CODE - A technique of modulating guidance transmitter carriers with coded pulses compatible with the receiver code of the missile assigned that system, thus reducing the possibility of the missile accepting erroneous commands of other transmissions.

GUIDANCE, COMMAND - A guidance system wherein intelligence transmitted to the missile from an outside source causes the missile to traverse a directed path in space.

GUIDANCE, HOMING, ACTIVE - A system of homing guidance wherein both the transmitter and receiver are carried within the missile.

GUIDANCE, HOMING, PASSIVE - A form of homing guidance, which is dependent on a missile's ability to detect energy emitted by the target. Frequently termed Home-On-Jam (HOJ).

GUIDANCE, HOMING, SEMIACTIVE - A system of homing guidance wherein the missile uses reflected signals from the target which has been illuminated by a source other than within the missile. See also CWI.

GUIDANCE, INERTIAL - A self-contained system independent of information obtained from outside the missile, usually using Newton's second law of motion.

GUNN DIODE - The Gunn diode is a transferred electron device which because of its negative resistance can be used in microwave oscillators or amplifiers. When the applied voltage exceeds a certain critical value, periodic fluctuations in current occur. The frequency of oscillation depends primarily upon the drift velocity of electrons through the effective length of the device. This frequency may be varied over a small range by means of mechanical tuning.

HARMONIC - A sinusoidal component of a periodic wave or quantity having a frequency that is an integral multiple of the fundamental frequency. For example, a component which is twice the fundamental frequency is called the second harmonic. (The fundamental is the first harmonic, which is frequently misunderstood.)

HERTZ - The unit of frequency equal to one cycle per second.

HOME-ON-JAM (HOJ) - See Guidance, Homing, Passive.

HORN ANTENNA - A flared, open-ended section of waveguide used to radiate the energy from a waveguide into space. Also termed "horn" or "horn radiator." Usually linearly polarized, it will be vertically polarized when the feed probe is vertical, or horizontally polarized if the feed is horizontal. Circular polarization can be obtained by feeding a square horn at a 45° angle and phase shifting the vertical or horizontal excitation by 90°.

HYPERABRUPT VARACTOR OSCILLATOR - Due to a non-uniform concentration of N-type material (excess electrons) in the depletion region, this varactor produces a greater capacitance change in tuning voltage and a far more linear voltage-vs.-frequency tuning curve. As a result, this device has an improved tuning linearity and low tuning voltage.

IF (INTERMEDIATE FREQUENCY) - The difference frequency resulting from mixing (beating) the received signal in a superheterodyne receiver with the signal from the local oscillator. The difference frequency product provides the advantages inherent to the processing (amplification, detection, filtering, and such) of low frequency signals. The receiver local oscillator may operate either below or above the receiver tuned frequency. A single receiver may incorporate multiple IF detection.

$IF = F_{LO} - F_O$ (for a local oscillator operating above the fundamental)

where:
F_O = Received fundamental frequency
F_{LO} = Local oscillator frequency

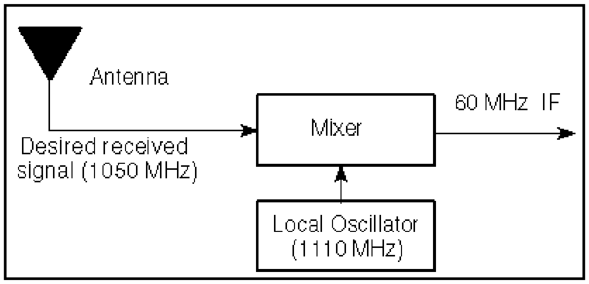
Antenna
Desired received signal (1050 MHz)
Mixer
60 MHz IF
Local Oscillator (1110 MHz)

The simplified block diagram illustrates a typical mixing procedure employed in radar systems to obtain desired IF frequencies. The local oscillator is tuned above the fundamental frequency in this example. It should be noted that an undesired signal received at the receiver image frequency of 1170 MHz will also produce the desired 60 MHz IF frequency; this relationship provides the receiver image. See also Image Frequency.

IFF (IDENTIFICATION FRIEND OR FOE) - A system using radar transmission to which equipment carried by friendly forces automatically responds by emitting a unique characteristic series of pulses thereby distinguishing themselves from enemy forces. It is the "Mode IV" for the aircraft transponder. See also transponder.

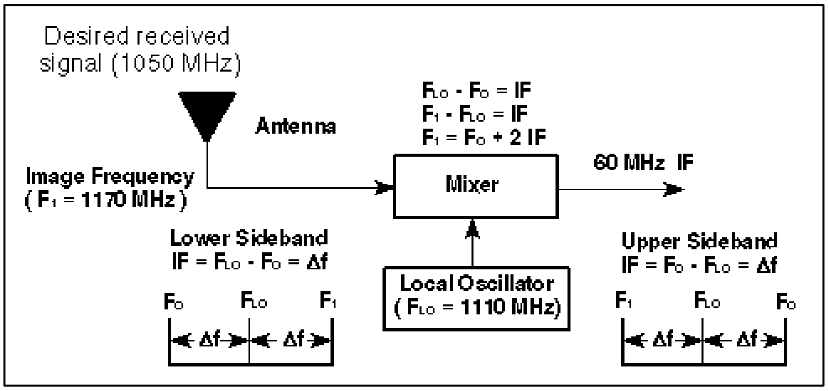
Desired received signal (1050 MHz)
Antenna
$F_{LO} - F_O = IF$
$F_I - F_{LO} = IF$
$F_I = F_O + 2\ IF$
Image Frequency (F_I = 1170 MHz)
Mixer
60 MHz IF
Lower Sideband
$IF = F_{LO} - F_O = \Delta f$
F_O F_{LO} F_I
Δf Δf
Local Oscillator (F_{LO} = 1110 MHz)
Upper Sideband
$IF = F_O - F_{LO} = \Delta f$
F_I F_{LO} F_O
Δf Δf

IMAGE FREQUENCY - That frequency to which a given superheterodyne receiver is inherently susceptible, thereby rendering such a receiver extremely vulnerable to EMI at that frequency. The image frequency is located at the same frequency difference (Δf) to one side of the local oscillator as the tuned (desired) frequency is to the other side. An undesired signal received at the image frequency by a superheterodyne receiver not having preselection would, therefore, mix (beat) with the oscillator, produce the proper receiver IF, and be processed in the same manner as a signal at the desired frequency. See also receiver selectivity.

IMAGE JAMMING - Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and image frequency of the radar.

Image jamming inverts the phase of the response and is thereby useful as an angle deception technique. Not effective if the radar uses image rejection.

IMPATT DIODE - The IMPATT (IMPact Avalanche and Transit Time) diode acts like a negative resistance at microwave frequencies. Because of this property, Impatt diodes are used in oscillators and amplifiers. Usually the frequency range is in the millimeter wave region where other solid state devices cannot compete.

INGRESS - Go into the target area.

INSERTION LOSS - The loss incurred by inserting an element, device, or apparatus in an electrical/electronic circuit. Normally expressed in decibels determined as 10 log of the ratio of power measured at the point of insertion prior to inserting the device (P_1) to the power measured after inserting the device (P_2). Insertion loss (dB) = 10 log P_1/P_2.

INTEGRATION EFFECT - Pulse radars usually obtain several echoes from a target. If these echoes are added to each other, they enhance the S/N ratio, making a weak target easier to detect. The noise and interference do not directly add from pulse to pulse, so the ratio of target strength to undesired signal strength improves making the target more detectable. Random noise increases by the square root of the number of integrations, whereas the signal totally correlates and increases directly by the number of integrations, therefore the S/N enhancement is equal to the square root of the number of integrations.

INTERFERENCE - See EMI.

INTERFERENCE PARAMETERS - Equipment and propagation characteristics necessary for the proper evaluation of a given EMI situation.

INTERFERENCE/SIGNAL RATIO - See I/S Ratio.

INTERFERENCE THRESHOLD - The level of interference normally expressed in terms of the I/S (interference/signal) ratio at which performance degradation in a system first occurs as a result of EMI.

INTERFEROMETER - When two widely spaced antennas are arrayed together, they form an interferometer. The radiation pattern consists of many lobes, each having a narrow beamwidth. This antenna can provide good spatial selectivity if the lobe-to-lobe ambiguity can be solved such as using amplitude comparison between the two elements.

INTERMODULATION - The production, in a nonlinear element (such as a receiver mixer), of frequencies corresponding to the sums and differences of the fundamentals and harmonics of two or more frequencies which are transmitted through the element; or, the modulation of the components of a complex wave by each other, producing frequencies equal to the sums and differences of integral multiples of the component frequencies of the complex wave.

INTERSYSTEM EMC - EMC between the external electromagnetic environment (EME) and an aircraft with its installed systems. Generally, only system BIT must operate properly on the carrier deck while all system functions must operate properly in the operational EME.

INTRASYSTEM EMC - EMC between systems installed on an aircraft, exclusive of an external environment.

INVERSE CON SCAN - One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna scan rate and then modulating repeater amplifier gain so the weak portion of the radar signal is amplified by the jammer, while the strong portion is not, so the weapons systems will fire at some bearing other than the true target bearing. The angle deception technique is used to break lock on CONSCAN radars.

INVERSE GAIN - Amplification, inverse modulation, and re-radiation of a radar's pulse train at the rotation rate of the radar scan. Deceives a conical scanning radar in angle.

ISOTROPIC ANTENNA - A hypothetical antenna which radiates or receives energy equally in all directions.

I/S RATIO (INTERFERENCE-TO-SIGNAL RATIO) (ISR) - The ratio of electromagnetic interference level to desired signal level that exists at a specified point in a receiving system. The ratio, normally expressed in dB, is employed as a tool in prediction of electronic receiving system performance degradation for a wide range of interference receiver input levels. Performance evaluations compare actual I/S ratios to minimum acceptable criteria.

JAFF - Expression for the combination of electronic and chaff jamming.

JAMMING - The deliberate radiation, reradiation, or reflection of electromagnetic energy with the object of impairing the use of electronic devices, equipment, or systems by an enemy.

JATO - JAmmer Technique Optimization group -- Organization coordinates EA technique development and evaluation efforts in support of naval airborne EA (ALQ-99, USQ-113, ALQ-227, ALQ-231, ALE-43). The group is run out of PMA-234 and comprised of elements at NAWCWD-Point Mugu, Johns Hopkins University/ Applied Physics Lab, Naval Research Lab, and NSWC-Crane. Efforts include M&S, Lab test, ground test, flight test, direct fleet support, and requirements and test support to development programs.

JINK - An aircraft maneuver which sharply changes the instantaneous flight path but maintains the overall route of flight. More violent than a weave.

JITTERED PRF - An antijam feature of certain radar systems which varies the PRF consecutively, and randomly, from pulse to pulse to prevent enemy ECM equipment from locking on, and synchronizing with, the transmitted PRF. PRF is synonymous with pulse repetition rate (PRR).

JTAT - JATO Techniques Analysis and Tactics -- Analysis documents produced by JATO to advise EA systems operators on the best ways to employ their systems by region, mission and equipment configuration. Currently covering ALQ-99 and USQ-113 on EA-6B; ALQ-99 and ALQ-227 on EA-18G, and ALQ-231 on AV-8B.

KILO - A prefix meaning 10^3 (times one thousand). For example, kilohertz.

KLYSTRON AMPLIFIER - An electron beam device which achieves amplification by the conversion of periodic velocity variations into conduction-current modulation in a field-free drift region. Velocity variations are launched by the interaction of an RF signal in an input resonant cavity and are coupled out through an RF output cavity. Several variations including reflex and multi cavity klystrons are used.

KLYSTRON, MULTICAVITY - An electron tube which employs velocity modulation to generate or amplify electromagnetic energy in the microwave region. Since velocity modulation employs transit time of the electron to aid in the bunching of electrons, transient time is not a deterrent to high frequency operations as is the case in conventional electron tubes. See also Velocity Modulation.

KLYSTRON, REFLEX - A klystron which employs a reflector (repeller) electrode in place of a second resonant cavity to redirect the velocity-modulated electrons through the resonant cavity. The repeller causes one resonant circuit to serve as both input and output, which simplifies the tuning operation. This type of klystron is well adapted for use as an oscillator because the frequency is easily controlled by varying the position of the repeller. See also Velocity Modulation.

LEAKAGE - Undesired radiation or conduction of RF energy through the shielding of an enclosed area or of an electronic device.

LENS, RADAR (MICROWAVE) - The purpose of any such lens is to refract (focus) the diverging beam from an RF feed into a parallel beam (transmitting) or vice versa (receiving). The polarization is feed dependent.

LIGHT AMPLIFICATION BY STIMULATED EMISSION OF RADIATION (LASER) - A process of generating coherent light. The process utilizes a natural molecular (and atomic) phenomenon whereby molecules absorb incident electromagnetic energy at specific frequencies, store this energy for short but usable periods, and then release the stored energy in the form of light at particular frequencies in an extremely narrow frequency-band.

LIMITING - A term to describe that an amplifier has reached its point of saturation or maximum output voltage swing. Deliberate limiting of the signal is used in FM demodulation so that AM will not also be demodulated.

LITTORAL - Near a shore.

LOBE, ANTENNA - Various parts of the antenna's radiation pattern are referred to as lobes, which may be subclassified into major and minor lobes. The major lobe is the lobe of greatest gain and is also referred to as the main lobe or main beam. The minor lobes are further subclassified into side and back lobes as indicated in the figure to the right. The numbering of the side lobes are from the main lobe to the back lobe.

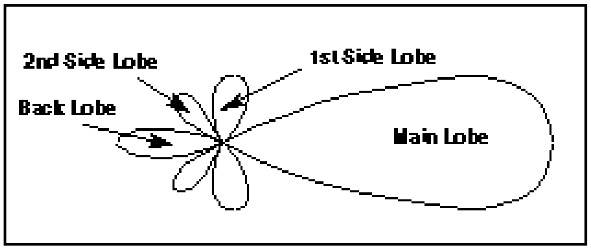

LOCAL OSCILLATOR FREQUENCY - An internally generated frequency in a superheterodyne receiver. This frequency differs from the receiver operating frequency by an amount equal to the IF of the receiver. The local oscillator frequency may be designed to be either above or below the incoming signal frequency.

LOG VIDEO - This receiver process, generally implemented in the IF, compresses the dynamic range of the signal so both weak and strong signals are displayed without changing the gain setting. Output voltage can be calibrated in volts/dB of input power.

LONG PULSE MODE - Many pulsed radars are capable of transmitting either long or short pulses of RF energy. When the long pulses of RF energy are selected manually (or sometimes automatically), the radar is said to be operating in the long pulse mode. In general, "long pulse mode" is used to obtain high average power for long-range search or tracking, and "short pulse mode" gives low average power for short-range, high-definition, tracking or search.

LOOSE DEUCE - General term for two aircraft working in mutual support of each other.

LORO (LOBE-ON-RECEIVE-ONLY) - A mode of operation generally consisting of transmitting on one non-scanning antenna system and receiving the reflected energy on another scanning system (The receiver could be TWS, Conical, or monopulse).

MACH NUMBER - The ratio of the velocity of a body to the speed of sound in the medium that is being considered. In the atmosphere, the speed of sound varies with temperature and atmospheric pressure, hence, so does mach number.

MAGNETIC ANOMALY DETECTOR - A means of detecting changes in the earth's magnetic field caused by the presence of metal in ships and submarines.

MAGNETRON - A magnetron is a thermionic vacuum tube which is constructed with a permanent magnet forming a part of the tube and which generates microwave power. These devices are commonly used as the power output stage of radar transmitters operating in the frequency range above 1000 MHz and are used less commonly down to about 400 MHz. A magnetron has two concentric cylindrical electrodes. On a conventional magnetron, the inner one is the cathode and the outer one is the anode. The opposite is true for a coaxial magnetron.

MAGNETRON OSCILLATOR - A high-vacuum tube in which the interaction of an electronic space charge and a resonant system converts direct current power into ac power, usually at microwave frequencies. The magnetron has good efficiency, is capable of high power outputs, and is stable.

MATCHED FILTER - This describes the bandwidth of an IF amplifier that maximizes the signal-to-noise ratio in the receiver output. This bandwidth is a function of the pulsewidth of the signal.

MDS (MINIMUM DETECTABLE/DISCERNIBLE SIGNAL) - The receiver input power level that is just sufficient to produce a detectable/discernible signal in the receiver output. The detectable term is interchangeable with S_{min} and the discernable term is interchangeable with MVS. See Section 5-2.

MEACONING -- A system receiving radio signals and rebroadcasting them (or just transmitting) on the same frequency to confuse navigation. The meaconing station attempts to cause aircraft to receive inaccurate range or bearing information.

MEATBALL - Visual light "ball" seen in Fresnel lens optical landing system (FLOLS) by pilot during carrier or Navy field landing. Used as a reference to determine if flight path is high or low.

MEGA - A prefix meaning 10^6 (times one million). For example megahertz (MHz)

MICROVOLT PER METER - A commonly used unit of field strength at a given point. The field strength is measured by locating a standard receiving antenna at that point, and the "microvolts per meter" value is then the ratio of the antenna voltage in microvolts to the effective antenna length in meters. Usually used below 100 MHz. Above 100 MHz, power density terminology is normally used.

MICROWAVE AMPLIFICATION BY STIMULATED EMISSION OF RADIATION (MASER) - A low-noise radio-frequency amplifier. The emission of energy stored in a molecular or atomic system by a microwave power supply is stimulated by the input signal.

MISS DISTANCE - Used variously in different contexts. The distance from the missile to the geometric center of the aircraft, or the closest point of approach (CPA) of the missile to any portion of the aircraft such as the aircraft nose or telemetry pod, etc.

MISSILE SYSTEMS FUNCTIONS - Examples of missile system functions are: "acquisition" (ability to lock-on a desired target); "tracking" of a target; "guidance" of a missile toward a target; "illumination" of a target so that a homing missile can home on the reflected RF illumination; and "command" signal transmission to a missile to cause it to arm, to detonate, to commence homing, or to destroy itself.

MIXERS - See Balanced and Schottky Diode Mixers.

MODULATION - The process whereby some characteristic of one wave is varied in accordance with some characteristic of another wave. The basic types of modulation are angle modulation (including the special cases of phase and frequency modulation) and amplitude modulation. In missile radars, it is common practice to amplitude modulate the transmitted RF carrier wave of tracking and guidance transmitters by using a pulsed wave for modulating, and to frequency module the transmitted RF carrier wave of illuminator transmitters by using a sine wave.

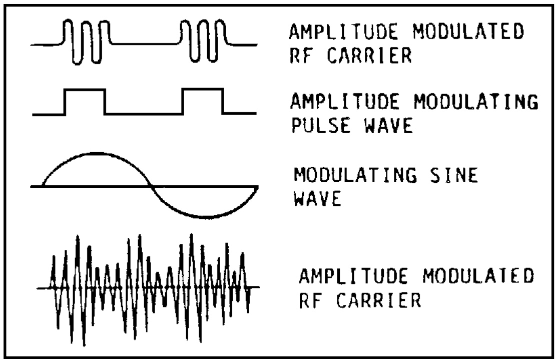
AMPLITUDE MODULATED RF CARRIER
AMPLITUDE MODULATING PULSE WAVE
MODULATING SINE WAVE
AMPLITUDE MODULATED RF CARRIER

MODULATION, AMPLITUDE - This type of modulation changes the amplitude of a carrier wave in responses to the amplitude of a modulating wave. This modulation is used in radar and EW only as a switch to turn on or turn off the carrier wave; i.e., pulse is a special form of amplitude modulation.

MODULATION, FREQUENCY - The frequency of the modulated carrier wave is varied in proportion to the amplitude of the modulating wave and therefore, the phase of the carrier varies with the integral of the modulating wave. See also Modulation.

MODULATION, PHASE - The phase of the modulated carrier is varied in proportion to the amplitude of the modulating wave. See also Modulation.

MONOPULSE - (See figure to right) A type of tracking radar that permits the extracting of tracking error information from each received pulse and offers a reduction in tracking errors as compared to a conical-scan system of similar power and size. Multiple (commonly four) receiving antennas or feeds are placed symmetrically about the center axis and operate simultaneously to receive each RF pulse reflected from the target. A comparison of the output signal amplitude or phase among the four antennas indicates the location of the target with respect to the radar beam center line. The output

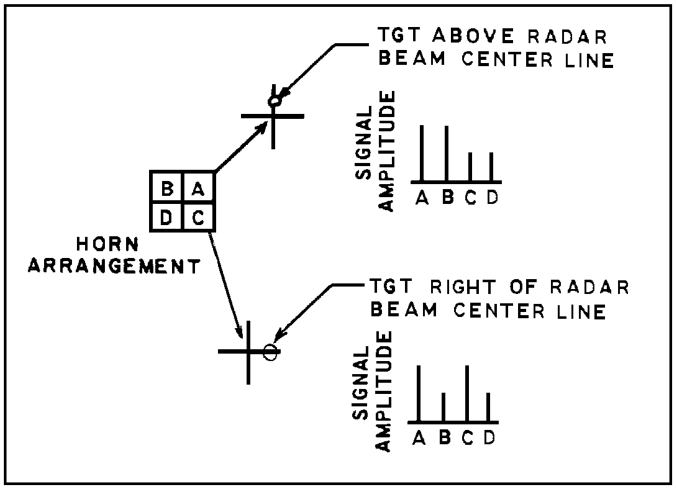

of the comparison circuit controls a servo system that reduces the tracking error to zero and thereby causes the antenna to track the target.

MOS (MINIMUM OPERATIONAL SENSITIVITY) - The minimum signal which can be detected and automatically digitally processed by a radar without human discrimination.

MTI (MOVING TARGET INDICATOR) - This radar signal process shows only targets that are in motion. Signals from stationary targets are subtracted out of the return signal by a memory circuit.

MULTIPATH - The process by which a transmitted signal arrives at the receiver by at least two different paths. These paths are usually the main direct path, and at least one reflected path. The signals combine either constructively or destructively depending upon phase, and the resultant signal may be either stronger or weaker than the value computed for free space.

MULTIPLEX - Simultaneous transmission of two or more signals on a common carrier wave. The three types of multiplex are called time division, frequency division, and phase division.

MULTIBAND RADAR - A type of radar which uses simultaneous operation on more than one frequency band through a common antenna. This technique allows for many sophisticated forms of video processing and requires any jammer to jam all channels at the same time in order to be effective.

MVS (MINIMUM VISIBLE SIGNAL) - The minimum input pulse signal power level which permits visibility of the output pulse, such as on a radar A-scope display. This level is determined by initially setting the input level above the visible detection threshold, and then slowly decreasing the amplitude.

NOISE FIGURE, RECEIVER - A figure of merit (NF or F) of a system given by the ratio of the signal-to-noise ratio at the input, S_i / N_i, divided by the signal-to-noise ratio at the output, S_o/N_o. It essentially expresses the ratio of output noise power of a given receiver to that of a theoretically perfect receiver which adds no noise.

$$\text{Noise Figure} = \frac{S_i / N_i}{S_o / N_o} = \frac{N_o}{G\,N_i}$$

Where $S_o = GS_i$ and G is the gain of the system.

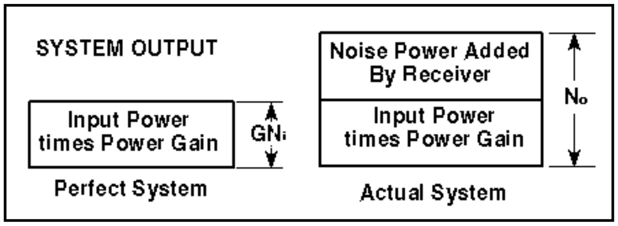
SYSTEM OUTPUT
Input Power times Power Gain
GNi
Perfect System
Noise Power Added By Receiver
Input Power times Power Gain
No
Actual System

Noise figure is usually expressed in dB and given for an impedance matched condition. Impedance mismatch will increase the noise figure by the amount of mismatch loss. NF is usually given at room temperature; 17°C or 290°K. See Section 5-2.

NOISE JAMMING - A continuous random signal radiated with the objective of concealing the aircraft echo from the enemy radar. In order for it to be effective, it must have an average amplitude at least as great as the average amplitude of the radar echo. There are three major categories of noise jamming which are grouped by how jamming power is concentrated: Spot, barrage, and swept jamming. (See individual definitions)

NONCOHERENT - Two signals that have no set phase relationship.

NOTCH - The portion of the radar velocity display where a target disappears due to being notched out by the zero Doppler filter. If not filtered (notched), ground clutter would also appear on the display. A notch filter is a narrow band-reject filter. A "notch maneuver" is used to place a tracking radar on the beam of the aircraft so it will be excluded.

NULL, ANTENNA PATTERN - The directions of minimum transmission (or reception) of a directional antenna. See also Lobe, Antenna.

NULL FILL - The nulls in an antenna pattern may be reduced (filled) by using a second ancillary (spoiler) antenna whose pattern is such that it fills in the nulls of the main antenna pattern.

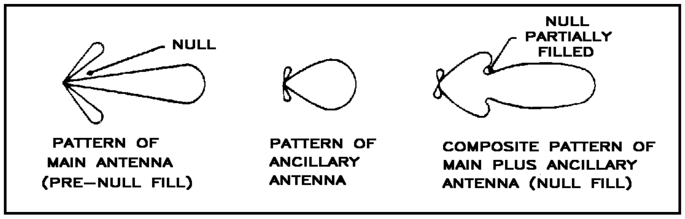
NULL
NULL PARTIALLY FILLED
PATTERN OF MAIN ANTENNA (PRE–NULL FILL)
PATTERN OF ANCILLARY ANTENNA
COMPOSITE PATTERN OF MAIN PLUS ANCILLARY ANTENNA (NULL FILL)

NUTATION - As applied to current missile system radars, this term refers to the mechanical motion of an antenna feed to produce a conical scan (fixed polarization) by the main beam of a tracking antenna, thus providing a means of developing tracking error signals. See also Antenna, Nutating. By analogy, "Nutation" also is used to denote the electrical switching of the quadrants of a seeker antenna. See also Interferometer. The effect is similar to that of a conical scan.

NUTATOR - A motor-driven rotating antenna feed used to produce a conical scan for a tracking radar. See also Antenna, Nutating. Also, the electrical circuits necessary to effect nonmechanical conical scans. See also Nutation.

OPERATIONAL CONSTRAINTS - Limitations on operating procedures in order to prevent interference between missile systems on a ship or between missile systems in a formation of ships under operational conditions. These limitations consist of such things as limited frequency bands or channels in which the radars may be tuned, limited sectors of space into which radar beams may be pointed, limits on minimum spacing between ships, limits on what codes may be used by radars and missiles on each ship, and limits on minimum interval between firing of certain missiles.

OSCILLATORS - Devices which generate a frequency. See also Backward Wave, Dielectrically Stabilized Oscillator, Hyperabrupt Varactor Oscillator, Magnetron Oscillator, Varactor Tuned Oscillator, and YIG tuned oscillator.

OSCILLATOR, LOCAL - See Local Oscillator Frequency.

PALMER SCAN - Conical scan superimposed on another type of scan pattern - usually a spiral pattern.

PARAMETER - A quantity which may have various values, each fixed within the limits of a stated case or discussion. In the present case, some examples of parameters; would be: radar frequency, limited by the tuning range of the radar; missile range, limited by the maximum operating range of the missile; or a missile code, limited by the number of codes available and by the codes that the ship radars are set up to operate on.

PASSIVE ANGLE TRACKING - Tracking of a target using radiation from the target (such as jamming), with no radiation from the radar itself. Only angular tracking is possible under these conditions since no measurement of time of travel of radiation to the target is possible, as is required to obtain target range.

PC (PULSE COMPRESSION) - The process used in search and tracking pulse radars whereby the transmitted pulse is long, so as to obtain high average transmitter output power, and the reflected pulse is processed in the radar receiver to compress it to a fraction of the duration of the transmitted pulse to obtain high definition and signal strength enhancement. Pulse compression may be accomplished by sweeping the transmitted frequency (carrier) during the pulse. The returned signal is then passed through a frequency-dependent delay line. The leading edge of the pulse is therefore delayed so that the trailing edge catches up to the leading edge to produce effectively a shorter received pulse than that transmitted. Pulse compression radars are also referred to as CHIRP radars. Other more sophisticated pulse compression techniques are also possible and are becoming more popular.

PENCIL BEAM - A narrow circular radar beam from a highly directional antenna (such as a parabolic reflector).

PHASED ARRAY RADAR - Radar using many antenna elements which are combined in a controlled phase relationship. The direction of the beam can be changed as rapidly as the phase relationships (usually less than 20 microseconds). Thus, the antenna typically remains stationary while the beam is electronically scanned. The use of many antenna elements allows for very rapid and high directivity of the beam(s) with a large peak and/or average power. There is also a potential for greater reliability over a conventional radar since the array will fail gracefully, one element at a time.

PIN DIODE - A diode with a large intrinsic (I) region sandwiched between the P- and N- doped semiconducting regions. The most important property of the PIN diode is the fact that it appears as an almost pure resistance at RF. The value of this resistance can be varied over a range of approximately one-10,000 ohms by direct or low frequency current control. When the control current is varied continuously, the PIN diode is useful for attenuating, leveling and amplitude modulation of an RF signal. When the control current is switched on and off or in discrete steps, the device is useful in switching, pulse modulating, and phase shifting an RF signal.

POINT CONTACT DIODE - This was one of the earliest semiconductor device to be used at microwave frequencies. Consisting of a spring-loaded metal contact on a semiconducting surface, this diode can be considered an early version of the Schottky barrier diode. Generally used as a detector or mixer, the device is somewhat fragile and limited to low powers.

POLARIZATION - The direction of the electric field (E-field) vector of an electromagnetic (EM) wave. See Section 3-2. The most general case is elliptical polarization with all others being special cases. The E-field of an EM wave radiating from a vertically mounted dipole antenna will be vertical and the wave is said to be vertically polarized. In like manner, a horizontally mounted dipole will produce a horizontal electric field and is horizontally polarized. Equal vertical and horizontal E-field components produce circular polarization.

PORT - The left side of a ship or aircraft when facing the bow (forward)

POWER (AVERAGE) FOR PULSED RADARS - Average power for a pulse radar is the average power transmitted between the start of one pulse and the start of the next pulse (because the time between pulses is many times greater than the pulse duration time, the average power will be a small fraction of peak power).

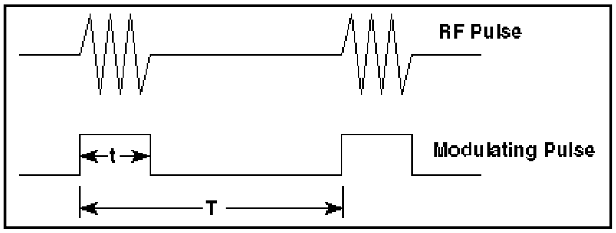

For this example: Peak Power = 1 MW, Pulse Time (t) = 0.5 microsecond, and Interval Between Pulses (T) = 1000 microseconds (1000 pps).

Peak Power = Pwr during pulse time (t) = 1 MW = 10^6 Watts = 90 dBm.

Avg Power = Average Power During Time (T) = 10^6 x t/T
= 10^6 x 0.5/1000 = 0.5 x 10^3 = 0.5 kilowatt = 57 dBm or 27 dBW

POWER OUTPUT - Power output of a transmitter or transmitting antenna is commonly expressed in dBW or dBm. One megawatt would be expressed as 60 dBW or 90 dBm:

10 log (1 megawatt / 1 watt) = 10 log ($10^6/10^0$) = 10 x 6 = 60 dBW

10 log (1 megawatt / 1 milliwatt) = 10 log ($10^6/10^{-3}$) = 10 x 9 = 90 dBm

POWER (PEAK) FOR PULSED RADARS - Peak power for a pulsed radar is the power radiated during the actual pulse transmission (with zero power transmitted between pulses).

POWER FOR CW RADARS - Since the power output of CW transmitters (such as illuminator transmitters) usually have a duty cycle of one (100%), the peak and average power are the same.

POWER DENSITY - The density of power in space expressed in Watts/meter², dBW/m², etc. Generally used in measurements above 100 MHz. At lower frequencies, field intensity measurements are taken. See Section 4-1.

PPI-SCOPE - A radar display yielding range and azimuth (bearing) information via an intensity modulated display and a circular sweep of a radial line. The radar is located at the center of the display.

PRESELECTOR - A device placed ahead of the mixer in a receiver, which has bandpass characteristics such that the desired (tuned) RF signal, the target return, is allowed to pass, and other undesired signals (including the image frequency) are attenuated.

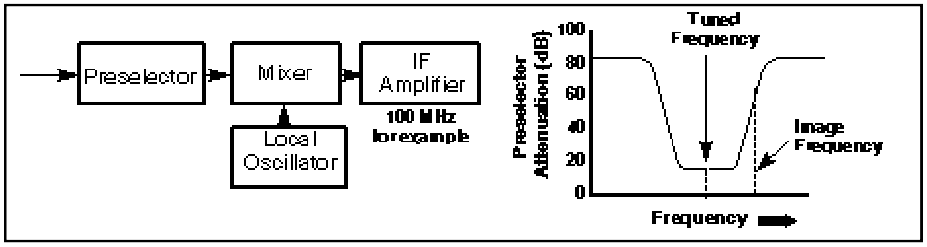

PROPAGATION - In electrical practice, the travel of waves through or along a medium. The path traveled by the wave in getting from one point to another is known as the propagation path (such as the path through the atmosphere in getting from a transmitting antenna to a receiving antenna, or the path through the waveguides and other microwave devices in getting from an antenna to a receiver).

PULSE COMPRESSION - See PC.

PULSED DOPPLER (PD) - A type of radar that combines the features of pulsed radars and CW Doppler radars. It transmits pulses (instead of CW) which permits accurate range measurement. This is an inherent advantage of pulsed radars. Also, it detects the Doppler frequency shift produced by target range rate which enables it to discriminate between targets of only slightly different range rate and also enables it to greatly reduce clutter from stationary targets. See also Doppler Effect.

PULSE LENGTH - Same meaning as Pulsewidth.

PULSE MODULATION - A special case of amplitude modulation wherein the carrier wave is varied at a pulsed rate. Pulse Modulation - The modulation of a carrier by a series of pulses generally for the purpose of transmitting data. The result is a short, powerful burst of electromagnetic radiation which can be used for measuring the distance from a radar set to a target.

PULSE REPETITION FREQUENCY (PRF) - The rate of occurrence of a series of pulses, such as 100 pulses per second. It is equal to the reciprocal of the pulse spacing (T) or PRT. (PRF = 1/T = 1/PRI). Sometimes the term pulse repetition rate (PRR) is used.

PULSE REPETITION FREQUENCY (PRF) STAGGER - The technique of switching PRF (or PRI) to different values on a pulse-to-pulse basis such that the various intervals follow a regular pattern. This is useful in compensating for blind speeds in pulsed MTI radars. Interpulse intervals which differ but follow a regular pattern.

PULSE REPETITION INTERVAL (PRI) or TIME (PRT) - Time between the beginning of one pulse and the beginning of the next.

PULSE SPACING - The interval of time between the leading edge of one pulse and the leading edge of the next pulse in a train of regularly recurring pulses. See also Pulse Repetition Frequency. Also called “the interpulse period.”

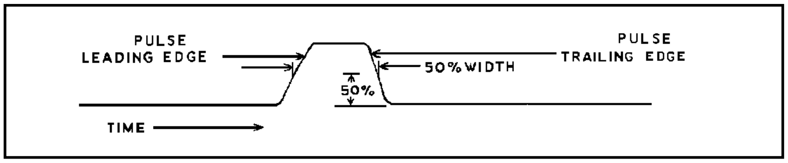

PULSEWIDTH - The interval of time between the leading edge of a pulse and the trailing edge of a pulse (measured in microseconds for the short pulses used in radar). Usually measured at the 3 dB midpoint (50-percent power or 70% voltage level) of the pulse, but may be specified to be measured at any level. See Section 6-10 for measurement techniques.

QUANTIZE - The process of restricting a variable to a number of discrete values. For example, to limit varying antenna gains to three levels.

RADAR - Radio detection and ranging.

RADAR CROSS SECTION - A measure of the radar reflection characteristics of a target. It is equal to the power reflected back to the radar divided by power density of the wave striking the target. For most targets, the radar cross section is the area of the cross section of the sphere that would reflect the same energy back to the radar if the sphere were substituted. RCS of sphere is independent of frequency if operating in the far field region. See Section 4-11.

RADAR RANGE EQUATION - The radar range equation is a basic relationship which permits the calculation of received echo signal strength, if certain parameters of the radar transmitter, antenna, propagation path, and target are known.

Given: $P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4}$ *(freespace)* as the basic two-way radar equation (see Sections 4-4 thru 4-6)

where:
P_r = Peak power at receiver input
λ = Wavelength of signal (length) = c/f

P_t = Transmitted signal level (power)
R = Range of target to radar (distance)
G_t = Gain of transmitting antenna (dimensionless ratio)
σ = Radar cross section of target
G_r = Gain of receiving antenna (dimensionless ratio)

In practical use, the radar range equation is often written in logarithmic form, all terms expressed in decibels, so that the results can be found by simple processes of addition and subtraction. Using the above equation and $\lambda = c/f$.

$$10 \log P_r = 10 \log P_t + 10 \log G_t + 10 \log G_r + 10 \log \sigma - 40 \log R - 20 \log f + 20 \log c - 30 \log 4\pi$$

where: f = Signal frequency (cycles {dimensionless}/time) c = Speed of light (length/time)

$$10 \log P_r = 10 \log P_t + 10 \log G_t + 10 \log G_r + G_\sigma - 2\alpha_1$$

where α_1 and G_σ are factors containing the constants and conversion factors to keep the equations in consistent units.

Refer to Sections 4-4 through 4-6.

RADAR TRIGGER KILL - see Trigger Kill, Radar.

RADIATION EFFICIENCY - $E = P_{radiated}/P_{in}$ (ideal=1)

RADIATION PATTERN - See Antenna Pattern.

RADIO FREQUENCY - See RF.

RADIO FREQUENCY INTERFERENCE - See RFI.

RAIL KEEPING - Ability of countermeasures to keep the missile on the launch rail, i.e., prevent launch.

RAINBOW - A technique which applies pulse-to-pulse frequency changing to identifying and discriminating against decoys and chaff.

RANGE CELL - In a radar, a range cell is the smallest range increment the radar is capable of detecting. If a radar has a range resolution of 50 yards and a total range of 30 nautical miles (60,000 yds), there are: 60000/50 = 1,200 range cells.

RANGE GATE - A gate voltage used to select radar echoes from a very short range interval.

RANGE GATE PULL OFF (RGPO) - Deception technique used against pulse tracking radars using range gates. Jammer initially repeats the skin echo with minimum time delay at a high power to capture the AGC circuitry. The delay is progressively increased, forcing the tracking gates to be pulled away ("walked off") from the target echo. Frequency memory loops (FML's), or transponders provide the variable delay.

RANGE RATE - The rate at which a radar target is changing its range with respect to the radar (in feet per second for example). Note that this rate is not the same as target velocity unless the target is moving straight toward or straight away from the radar.

RANGE SCOPE - See A-Scope or PPI.

RECEIVER BANDWIDTH - The difference between the limiting frequencies within which receiver performance in respect to some characteristic falls within specified limits. (In most receivers this will be the difference between the two frequencies where the intermediate frequency (IF) amplifier gain falls off 3 dB from the gain at the center IF frequency.) See also Receiver Selectivity.

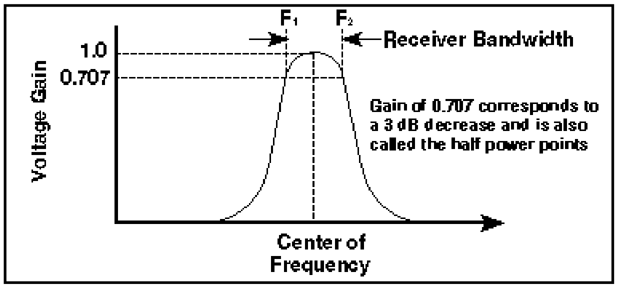

RECEIVER SELECTIVITY - The degree to which a receiver is capable of differentiating between the desired signal and signals or interference at other frequencies. (The narrower the receiver bandwidth, the greater the selectivity.)

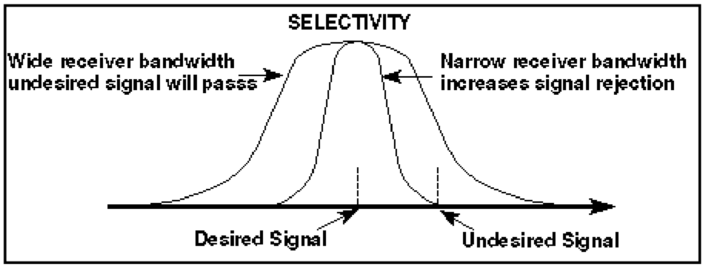
SELECTIVITY
Wide receiver bandwidth undesired signal will pass
Narrow receiver bandwidth increases signal rejection
Desired Signal
Undesired Signal

REFLECTION - The turning back (or to the side) of a radio wave as a result of impinging on any conducting surface which is at least comparable in dimension to the wavelength of the radio wave.

RESOLUTION - In radar, the minimum separation in angle or in range between two targets which the radar is capable of distinguishing.

RF (RADIO FREQUENCY) - A term indicating high frequency electromagnetic energy.

RFI (RADIO FREQUENCY INTERFERENCE) - Any induced, radiated, or conducted electrical disturbance or transient that causes undesirable responses or malfunctioning in any electrical or electronic equipment, device, or system. Same as EMI. Not to be confused with the logistic term ready for issue (also RFI).

RING AROUND - A condition in which a repeater jammer’s total gain, from receiver antenna to transmitter antenna, exceeds the antenna isolation resulting in the repeater amplifying its own internal noise. Akin to positive feedback in an amplifier that causes unwanted oscillations.

RING AROUND (RADAR-TO-MISSILE) - The condition where radio frequency interference signals from a transmitter of one missile radar enter the receiving circuits of a missile under the control of another missile radar.

RING AROUND (RADAR-TO-RADAR) - The condition where radio frequency interference signals from a transmitter of one radar enter the receiving circuits of another radar.

ROPE - An element of chaff consisting of a long roll of metallic foil or wire which is designed for broad, low-frequency response. See Chaff.

R-SCOPE - (RANGE SCOPE) See A-scope or PPI.

SAFETY OF FLIGHT (SOF) TEST - A flight test to verify that a new or modified subsystem will not cause a major problem with the aircraft, i.e., interference can occur, but will not be such that required navigational systems will fail or which might potentially cause the loss of an aircraft under all normally expected weather conditions.

SCAN - To transverse or sweep a sector or volume of airspace with a recurring pattern, by means of a controlled directional beam from a radar antenna. See also "Antenna, nutating."

SCHOTTKY BARRIER DIODE - The Schottky barrier diode is a simple metal-semiconductor boundary with no P-N junction. A depletion region between the metal contact and the doped semiconductor region offers little capacitance at microwave frequencies. This diode finds use as detectors, mixers, and switches.

SCHOTTKY DIODE MIXER - The mixer is a critical component in modern RF systems. Any nonlinear element can perform the mixing function, but parameters determining optimal mixing are noise figure, input admittance, and IF noise and impedance. The Schottky diode is particularly effective because of its low noise figure and nearly square law characteristics.

SCHOTTKY DIODE SWITCH - Standard P-N diodes are limited in switching ability at high frequencies because of capacitance provided by the minority carriers. The Schottky diode overcomes this problem by use of the metal-semiconductor junction with inherently low carrier lifetimes, typically less than 100 picoseconds.

SEARCH RADAR - A radar whose prime function is to scan (search) a specified volume of space and indicate the presence of any targets on some type of visual display, and, in some cases, to provide coordinates of the targets to a fire control system to assist in target acquisition and tracking.

SEEKER - The seeker consists of circuitry in a homing missile which detects, electronically examines, and tracks the target; provides data for controlling the flight path of the missile; and provides signals for destroying the missile or for detonating it at intercept. (The seeker function is similar to that of an interferometer.)

SELF-SCREENING JAMMING (SSJ) - Each aircraft carries its own jamming equipment for its own protection.

SENSITIVITY - The sensitivity of a receiver is taken as the minimum signal level required to produce an output signal having a specified signal-to-noise ratio. See also Minimum Visible Signal and Minimum Discernible Signal (MDS).

SENSITIVITY TIME CONTROL - See STC.

SENSOR - The receiver portion of a transmitter/receiver pair used to detect and process electromagnetic energy.

SHIELDING - The physical arrangement of shields for a particular component, equipment, or system, (A shield is a housing, screen, or other material, usually conducting, that substantially reduces the effect of electric or magnetic fields on one side of the shield upon devices or circuits on the other side.) Examples are tube shields, a shielded enclosure or cabinet for a radar receiver, and the screen around a screen room.

SHORT PULSE MODE - See Long Pulse Mode.

SIDEBAND - A signal either above or below the carrier frequency, produced by the modulation of the carrier wave by some other wave. See figure at right ⇒

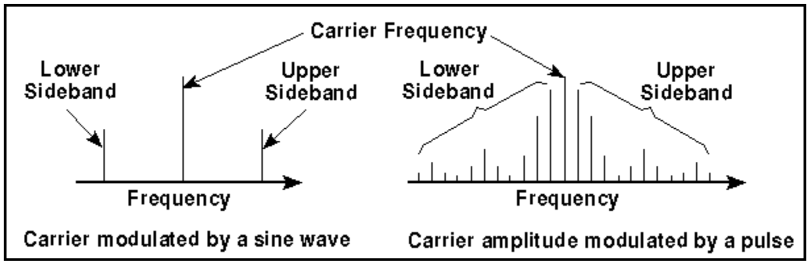

SIDELOBE - See Lobe, Antenna.

SIGNAL STRENGTH - The magnitude of a signal at a particular location. Units are volts per meter or dBV/m.

SIGNATURE - The set of parameters which describe the characteristics of a radar target or an RF emitter and distinguish one emitter from another. Signature parameters include the radio frequency of the carrier, the modulation characteristics (typically the pulse modulation code), and the scan pattern.

SILICON CONTROLLED SWITCH - A P-N-P-N device able to operate at sub-microsecond switching speeds by the application of gate signals. Because it is a four layer device, this switch is also known as a tetrode thyristor.

SLANT POLARIZATION - Technique of rotating a linear antenna 45° so it can receive or jam both horizontal and vertical polarization although there is a 3 dB loss. See Section 3.2.

SOLID STATE STAMO - A stable master oscillator constructed using transistors and other solid state devices as opposed to vacuum tubes. See also STAMO.

SPECTRUM - The distribution of power versus frequency in an electromagnetic wave. See also Spectrum Signature Analysis and illustrations under Sideband.

SPECTRUM ANALYZER - An electronic device for automatically displaying the spectrum of the electromagnetic radiation from one or more devices. A cathode ray tube display is commonly used to display this power-versus frequency spectrum. For examples of two types of displays, see illustrations under Sideband.

SPECTRUM SIGNATURE ANALYSIS - The analysis of the electromagnetic radiation from an electronic device to determine the relative power in each sideband, harmonic, and spurious emission compared to the carrier frequency. This particular distribution (or spectrum) is peculiar to the device and can identify this type of device, thereby acting as an identifying "signature."

SPECTRUM WIDTH (TRANSMITTER) - The difference between the frequency limits of the band which contains all the spectrum frequency components of significant magnitude.

SPOILER ANTENNA - An antenna used to change (spoil) the antenna pattern of a second antenna so as to reduce the nulls in the pattern of the second antenna. See also Null Fill.

SPOKING (RADAR) - Periodic flashes of the rotating radial display. Sometimes caused by mutual interference.

"SPOOFING" - A type of deception by using an electronic device to transmit a "target" echo. The spoofing transmitter must operate at the same frequency and PRF as the radar to be deceived. The radar main pulse triggers the spoofing transmitter which, after a delay, transmits a false echo.

SPOT JAMMING - Narrow frequency band jamming concentrated against a specific radar at a particular frequency. The jamming bandwidth is comparable to the radar bandpass. Can deny range and angle information.

SPURIOUS EMISSION - Electromagnetic radiation transmitted on a frequency outside the bandwidth required for satisfactory transmission of the required waveform. Spurious emissions include harmonics, parasitic emissions, and intermodulation products, but exclude necessary modulation sidebands of the fundamental carrier frequency.

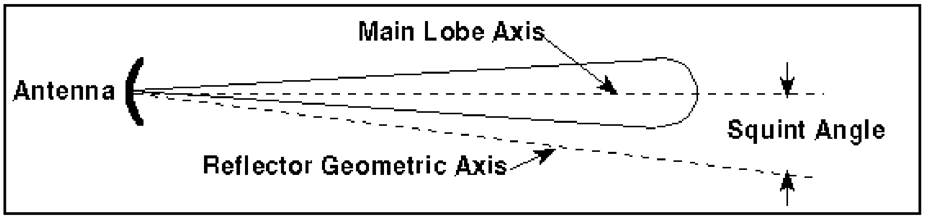

SQUINT ANGLE - The angular difference between the axis of the antenna mainlobe and the geometric axis of the antenna reflector, such as the constant angle maintained during conical scan as the mainlobe rotates around the geometric axis of the reflector.

STAGGERED PRF - Staggered PRF allows an increase in MTI blind speeds such that no zeros exist in the velocity response at lower velocities. In a two-period mode, the usual "blind speed" or occurrence of a zero in the velocity response is multiplied by a factor which is a function of the ratio of the two repetition periods.

STAMO (STABLE MASTER OSCILLATOR) - A very stable (drift free) oscillatory used to provide a precise frequency for transmission and for comparison with the reflected radar signal returned to the receiver, such as in a Doppler radar where a precise difference between transmitted and received signals must be measured to determine accurately the Doppler frequency.

STAND-FORWARD JAMMING - A method which places the jamming vehicle between the enemy sensors and attack aircraft.

STAND-IN JAMMING (SIJ) - Similar to stand-forward jamming but usually using an UAV with a lower powered jammer instead of a jammer aircraft.

STAND-OFF JAMMING (SOJ) - An ECM support aircraft orbits in the vicinity of the intended target. As the fighter-bomber pilot starts his strike penetration, the ECM aircraft directs jamming against all significant radars in the area. This technique provides broad frequency band ECM without affecting performance of the strike aircraft.

STARBOARD - The right side of a ship or airplane when facing the bow (forward).

STC (SENSITIVITY TIME CONTROL) - Gain control that reduces the radar receiver gain for nearby targets as compared to more distant targets. STC prevents receiver saturation from close-in targets.

STEP RECOVERY DIODE - A charge-controlled switch which ceases current conduction so rapidly that it can be used to produce an impulse. Cyclic operation of the diode can produce a train of impulses which when used with a resonant circuit can produce a single frequency output at any harmonic of the pulse frequency.

STERADIAN - Unit of solid angle. An entire sphere has 4π steradians.

STREAM CHAFF - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a “ribbon” or “stream” of returns many miles in lengths on radarscopes. The penetrating strike force can then use the resulting chaff corridor to mask their penetration.

SUBHARMONIC - A frequency which is an integral submultiple of another frequency. For example, a sine wave whose frequency is one-third of the frequency of another sine wave is called the third subharmonic. (3 MHz is the third subharmonic of 9 MHz).

SUPERHETERODYNE RECEIVER - A receiver that mixes the incoming signal with a locally generated signal (local oscillator) to produce a fixed, low intermediate frequency (IF) signal for amplification in the IF amplifiers.

SUPPRESSION - Elimination or reduction of any component of an emission, such as suppression of a harmonic of a transmitter frequency by band rejection filter.

SUPPRESSION OF ENEMY AIR DEFENSES (SEAD) - Activity which neutralizes, destroys, or temporarily degrades enemy air defense systems by using physical attack or electronic means (SEAD pronounced “seed” or “C add”).

SUSCEPTIBILITY - The degree to which an equipment or a system is sensitive to externally generated interference.

SWEPT JAMMING - Narrowband jamming which is swept through the desired frequency band in order to maximize power output. This technique is similar to sweeping spot noise to create barrage jamming, but at a higher power.

SWITCHES - See also Diode Switch, Silicon Controlled Switch, Schottky Diode Switch.

SYNCHRODYNE - A klystron mixer amplifier stage in a transmitter, where two signal frequencies are applied as inputs and a single amplified signal is taken out.

TARGET SIZE - A measure of the ability of a radar target to reflect energy to the radar receiving antenna. The parameter used to describe this ability is the “radar cross section” of the target. The size (or radar cross section) of a target, such as an aircraft, will vary considerably as the target maneuvers and presents different views to the radar. A side view will normally result in a much larger radar cross section than a head-on view. See also Radar Cross Section.

TERMINAL IMPEDANCE - The equivalent impedance as seen by the transmitter/receiver.

TERRAIN BOUNCE - Term for jamming that is directed at the earth’s surface where it is reflected toward the threat radar. Reflected jamming creates a virtual image of the jamming source on the earth as a target for HOJ missiles.

THERMISTOR - A resistor whose resistance varies with temperature in a defined manner. The word is formed from the two words “thermal” and “resistor,”

THRESHOLD ISR - The interference to signal ratio (ISR) at which the performance of a receiver starts undergoing degradation. It must be determined by tests.

TRACKING RADAR - A radar whose prime function is to track a radar target and determine the target coordinates (in range and angular position) so that a missile may be guided to the target, or a gun aimed at the target.

TRACKING RADAR RECEIVER - These are of two primary types: conical scan and monopulse. (1) The conical scan system directs the radar signal in a circle around the target. The radar paints this circle 15 to 40 times per second. As the target moves out of the center of this circle, the radar develops aim error voltages and re-aims the antenna. (2) The monopulse system directs four beams at the target simultaneously. The target is in the middle of the four beams. If the target is not in the center, the radar return develops an aim error voltage to re-aim the antenna.

TRACK WHILE SCAN (TWS) RADAR - Although it is not really a tracking radar in the true sense of the word, it does provide complete and accurate position information for missile guidance. In one implementation it would utilize two separate beams produced by two separate antennas on two different frequencies. The system utilizes electronic computer techniques whereby raw datum is used to track an assigned target, compute target velocity, and predict its future position, while maintaining normal sector scan. Most aircraft use only a single antenna.

TRADE-OFF TABLES - A set of tables showing the various combinations of two or more variables that are related in that making one variable better will make the other variable worse. The trade-off helps find the best solution considering all combinations. (For example, how a no-interference condition can be maintained if two emitter platforms are brought close together, if at the same time the frequency separation between their radar transmitters is increased.)

TRANSIENT - A phenomenon (such as a surge of voltage or current) caused in a system by a sudden change in conditions, and which may persist for a relatively short time after the change (sometimes called ringing).

TRANSPONDER - A transmitter-receiver capable of accepting the electronic challenge of an interrogator and automatically transmitting an appropriate reply. There are four modes of operation currently in use for military aircraft. Mode 1 is a nonsecure low cost method used by ships to track aircraft and other ships. Mode 2 is used by aircraft to make carrier controlled approaches to ships during inclement weather. Mode 3 is the standard system used by commercial aircraft to relay their position to ground controllers throughout the world. Mode 4 is IFF. See also IFF.

TRAVELING-WAVE TUBE AMPLIFIER - The TWT is a microwave amplifier capable of operation over very wide bandwidths. In operation, an electron beam interacts with a microwave signal which is traveling on a slow wave helical structure. The near synchronism of the beam and RF wave velocities results in amplification. Bandwidths of 3:1 are possible. Operation at high powers or at millimeter wavelengths is possible at reduced bandwidths.

TRIGGER KILL (RADAR) - A method employed to momentarily disable certain radar system circuits to reduce or eliminate RF emissions which may cause an EMI/EMC or RADHAZ situation such as on the deck of a ship.

TUNNEL DIODE - The tunnel diode is a heavily doped P-N junction diode that displays a negative resistance over a portion of its voltage-current characteristic curve. In the tunneling process, electrons from the p-side valence bands are able to cross the energy barrier into empty states in the N-side conduction band when a small reverse bias is applied. This diode is used as a microwave amplifier or oscillator.

UPLINK - The missile guidance signal which passes midcourse correction command guidance intelligence from the guidance radar site to the missile.

VARACTOR DIODE - A P-N junction employing an external bias to create a depletion layer containing very few charge carriers. The diode effectively acts as a variable capacitor.

VARACTOR TUNED OSCILLATOR - A varactor diode serves as a voltage-controlled capacitor in a tuned circuit to control the frequency of a negative resistance oscillator. The major feature of this oscillator is its extremely fast tuning speed. A limiting factor is the ability of the external voltage driver circuit to change the voltage across the varactor diode, which is primarily controlled by the driver impedance and the bypass capacitors in the tuning circuit.

VELOCITY GATE PULL-OFF (VGPO) - Method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against CW or Doppler velocity tracking radar systems. The CW or pulse doppler frequency, which is amplified and retransmitted, is shifted in frequency (velocity) to provide an apparent rate change or Doppler shift.

VELOCITY MODULATION - Velocity modulation is modification of the velocity of an electron beam by alternately accelerating and decelerating the electrons at a frequency equal to the input frequency. Thus, the electrons are segregated in bunches, each bunch causing a cycle or current as it passes an output electrode. The velocity of the electrons is thus a function of the modulation voltage. See also Klystron, Multicavity and Klystron, Reflex.

VICTIM - A receiver (radar or missile) that suffers degradation due to ECM or EMI effects.

VIDEO - Receiver RF signals that have been converted (post detection) into a pulse envelope that can be seen when applied to some type of radar visual display; also used to describe the actual display itself (such as the video on an A-scope).

WARM - Acronym for Wartime Reserve Mode. Any mode of operation of a radar or ECM that is held in reserve, and never used, except in actual combat.

WATER LINE - A reference line used for vertical measurements. When used with an aircraft it is usually the ground with the landing gear extended normally. One of several aircraft location designations, also see butt line and fuselage station.

WAVEGUIDE - A transmission line consisting of a hollow conducting tube of arbitrary geometry (usually rectangular, but may be circular) within which electromagnetic waves may propagate.

WAVELENGTH (λ) - The distance traveled by a wave in one period (the period is the time required to complete one cycle). $\lambda = c/f$. In the atmosphere, electromagnetic waves travel at c, the speed of light (300 million meters per second or 30 cm/nsec). At 5 GHz, one wavelength = 6 cm. At 10 GHz, one wavelength = 3 cm.

WAVEMETER - An instrument for measuring the frequency of a radio wave. The wavemeter is a mechanically tunable resonant circuit. It must be part of a reflection of transmission measurement system to measure the maximum response of a signal. Below 20 GHz, the wavemeter has been replaced by the frequency counter with much greater accuracy and ease of use.

WEAVE - An aircraft maneuver that smoothly changes the instantaneous flight path but maintains the overall route of flight. Not as violent as a jink.

WET RUN - A test run with ship / aircraft armament and/or EW switches on.

WILD WEASEL - USAF aircraft (F-4Gs during Desert Storm) used for suppression of enemy air defense (SEAD) mission.

WINDOW - WWII name for chaff

YIG TUNED OSCILLATOR - A YIG (yttrium iron garnet) sphere, when installed in the proper magnetic environment with suitable coupling will behave like a tunable microwave cavity with Q on the order of 1,000 to 8,000. Since spectral purity is related to Q, the device has excellent AM and FM noise characteristics.

ZENER DIODE - A diode that exhibits in the avalanche-breakdown region a large change in reverse current over a very narrow range in reverse voltage. This characteristic permits a highly stable reference voltage to be maintained across the diode despite a wide range of current.

THE REALITY OF PSYCHOTRONIC WEAPONS TODAY!

by
Preston T. Bailey, PhD.

Poll: Most Americans Feel Threatened By the Government

A new survey from Pew Research Center finds more than half of Americans believe the government threatens their personal rights. The national survey marks the first time a majority of Americans have viewed the government as a threat since Pew began tracking public opinion on the subject in 1995.

Welcome to
IDAHO

ATTENTION CRIMINALS & TERRORISTS

Over 170,000 Idaho Residents Have A Permit to Legally Carry A Concealed Weapon, About 60% Of The rest of the Population Is Armed But Have Not Bothered to Purchase the License As It Is Not A Requirement To Carry A Firearm. Understand, A Substantial Portion Of The Population Is Armed And Prepared To Defend themselves And Others Against Acts of Criminal Violence.

YOU HAVE BEEN WARNED!

However, California, New York and Illinois have disarmed their citizens for your convenience.

The outgoing Homeland Security Secretary has a warning for her successor: A massive and “serious” cyber attack on the U.S. homeland is coming, and a natural disaster — the likes of which the nation has never seen — is also likely on its way.

ELECTROMAGETIC ENERGY WEAPONS

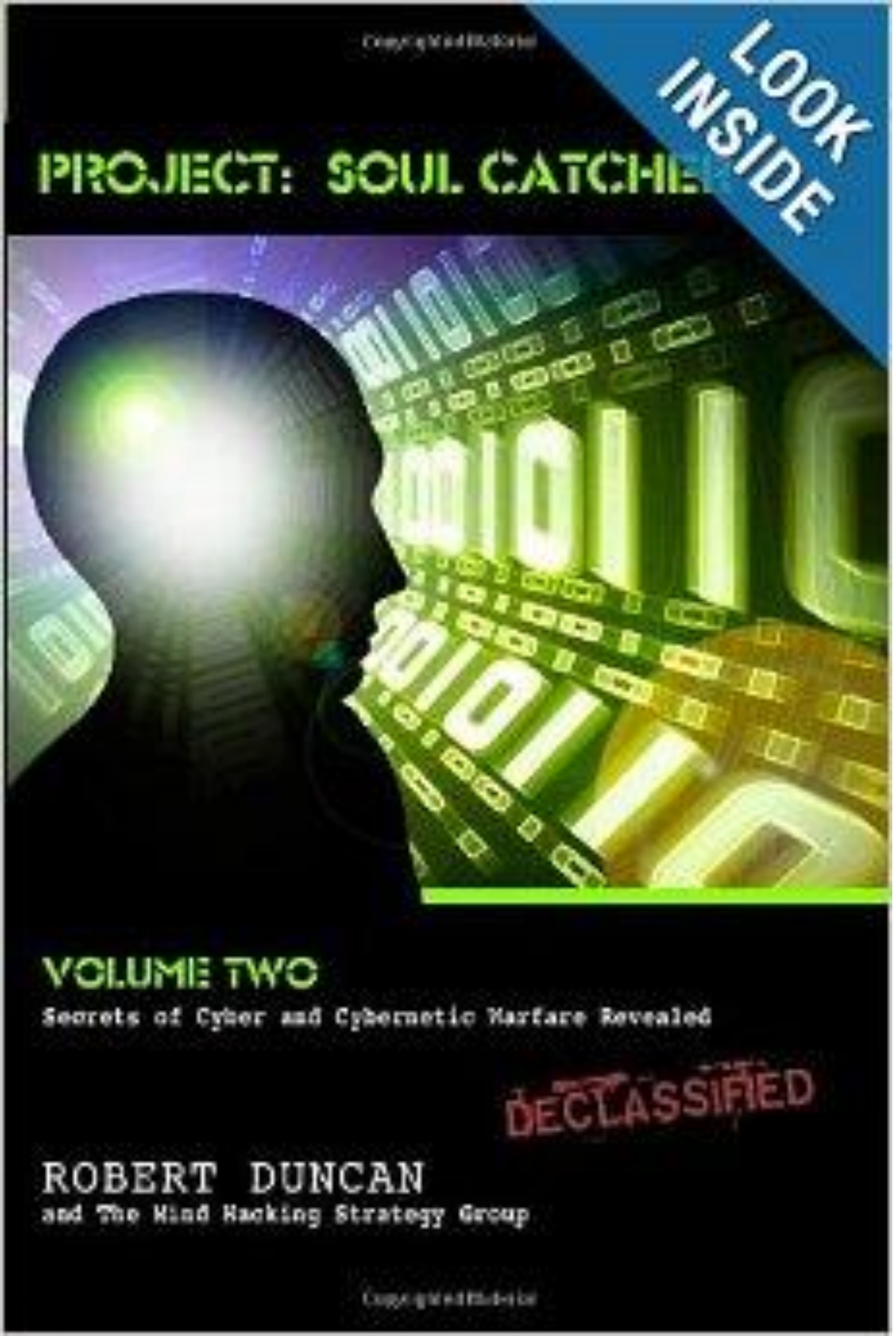
Copyrighted Material
LOOK INSIDE
PROJECT: SOUL CATCHE
VOLUME TWO
Secrets of Cyber and Cybernetic Warfare Revealed
DECLASSIFIED
ROBERT DUNCAN
and The Mind Hacking Strategy Group
Copyrighted Material

Project Soul Catcher: Secrets of Cyber and Cybernetic Warfare Revealed reveals the cybernetic mind control by the CIA and their use of psychotronic weapons and their pursuit to weaponize neurpsychology.

ELECTROMAGNETIC SATELLITE TECHNOLOGY

THE NEUROPHONE:US Patent # 3,393,279.
July 16th, 1968

It's a device that converts sound to electrical impulses sometimes called the “Voice of God” or makes the person think he is crazy. When the signals are sent to the brain, it makes a person think that God is speaking to him/her or even an “alien.” It targets the unique bioelectrical resonance frequency using his “vibrational pattern” in the brain called a “signature” and it acts like an audio receiver.

THE PSYCHO-ACOUSTIC PROJECTOR:

U.S. patent #3,566,347, (23/2/71)
A device/weapon which can actually deafen the target.

SILENT SUBLIMAL MESSAGES: US Patent # 5,159,703. October 27th, 1992

Non aural carriers in extreme audio frequency ranges are amplified or modulated with the desired material and propagated acoustically for direct inducement into the brain. This is an excellent method of influencing people without their knowledge and is used in advertising.

METHODS AND SYSTEMS OF ALTERING CONSICOUSNESS:

US Patent # 5,123,844. June 23rd, 1992
US Patent # 5,289,438. February 22nd, 1994
These systems stimulate the brain with different frequencies and wave forms to alter the subject's state of consciousness.

Electro Magnetic Field (EMF) monitoring/interference is one of the most insidious and secretive of all methods used by the agencies.

MICROWAVE WEAPONS: "A hearing system" U.S. patent #4,877,027, 31/10/89 & "A hearing device" U.S. patent #4,858,612, 22/8/89.

The first known experiment with microwaved voices was conducted by Sharp and Grove in the early 70's. However, the Defence Intelligence Agency and ARPA (The Advanced Research Projects Agency) are principally to blame for the abuse of such technologies since. eg Project Pandora etc. The CIA's Langley Research Center as well as numerous scientists working in Energy/Defence department labs across the U.S. are also responsible.

BRAIN WAVE MONITORS/ANALYSERS

When the brain wave pattern of an individual is detected from the use of one of numerous types of detectors such as a satellite, these brainwave signals are sent to a monitor with a computer. These brain wave patterns are interpreted through a brain wave vocabulary that records the thoughts into words. The NSA monitors the brainwaves of their targets by satellite and decode the evoked potentials (3.50Hz 5 milliwatts) that the brain emits. So, using lasers / satellites and high-powered computers, the agencies have now gained the ability to decipher human thoughts - and from a considerable distance almost instantaneously.

Psychotronic Weapons

Russian former Prime Minister Dmitry Medvedev with a psychotronic gun

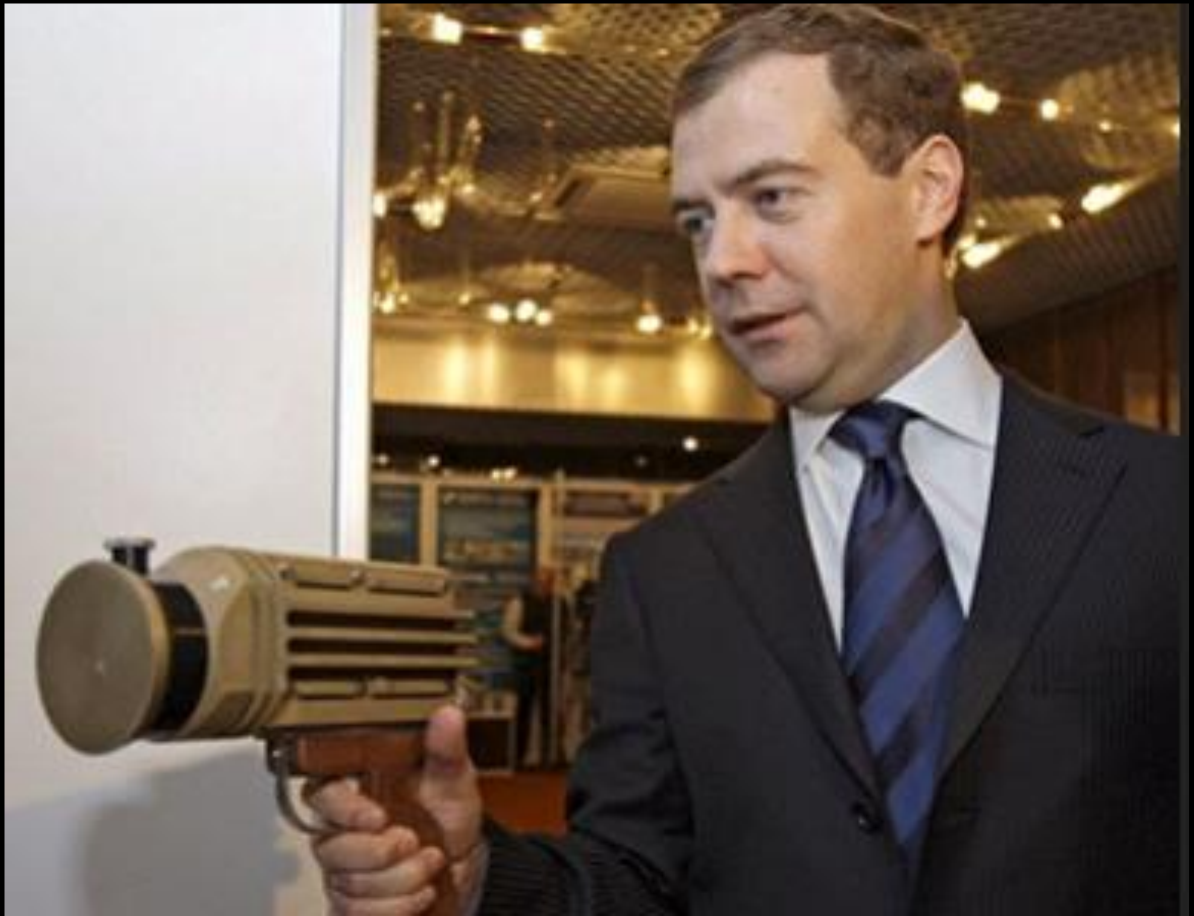

Psychotronic Guns by Thor Jensen

This is still an in-development project, but the potential for mayhem is unimaginable. Russia is currently working on weapons that employ electromagnetic radiation to directly affect the central nervous system – the root of all sensory processing – in a variety of unpleasant ways. The weapon will most likely build on existing research that shows the ability of low-frequency microwaves to affect brain function, alter moods and transmit false sensory data. Although this may all seem like *X-Files* stuff, the scientific basis is there. It’s up to some unscrupulous weapons designer to put it into practice.

Long Range Acoustic Device

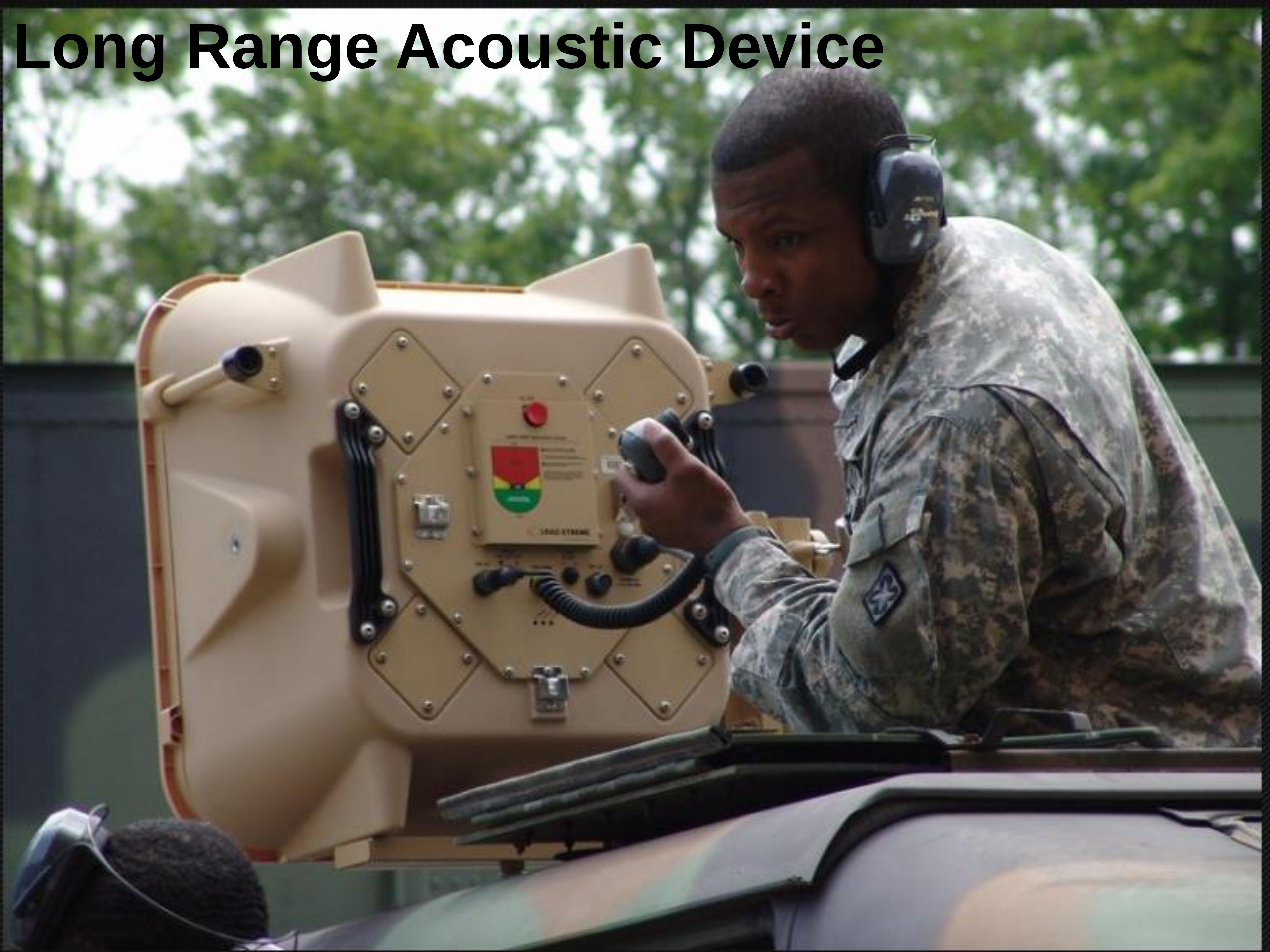

Sound waves have the potential to deal tremendous damage at loud enough amplitudes – think about a sonic boom, for instance. But you don’t need a jet engine to turn sound into a weapon. The Long Range Acoustic Device is a creation of the American technology Corporation that was originally designed to defend safe zones around boats. It produces an incredibly loud noise that can travel almost 1000 feet in a targeted beam, and if the frequency is amplified enough it can cause legitimate physical pain in its victims.

Dazer Laser

Of all the senses you use on the battlefield, vision is probably the most important. So disrupting that can turn the tide in war. One of the most effective sensory weapons on the market right now is the Dazer Laser, made by Laser Energetics. Using a green laser source that emits both a continuous and pulsed wave, the range on the unit can be adjusted to as far as 2400 meters. One shot of the laser targeted at the eye area can completely incapacitate the victim for several minutes, long enough for them to be subdued non-lethally.

Distributed Sound And Light Array Debilitator

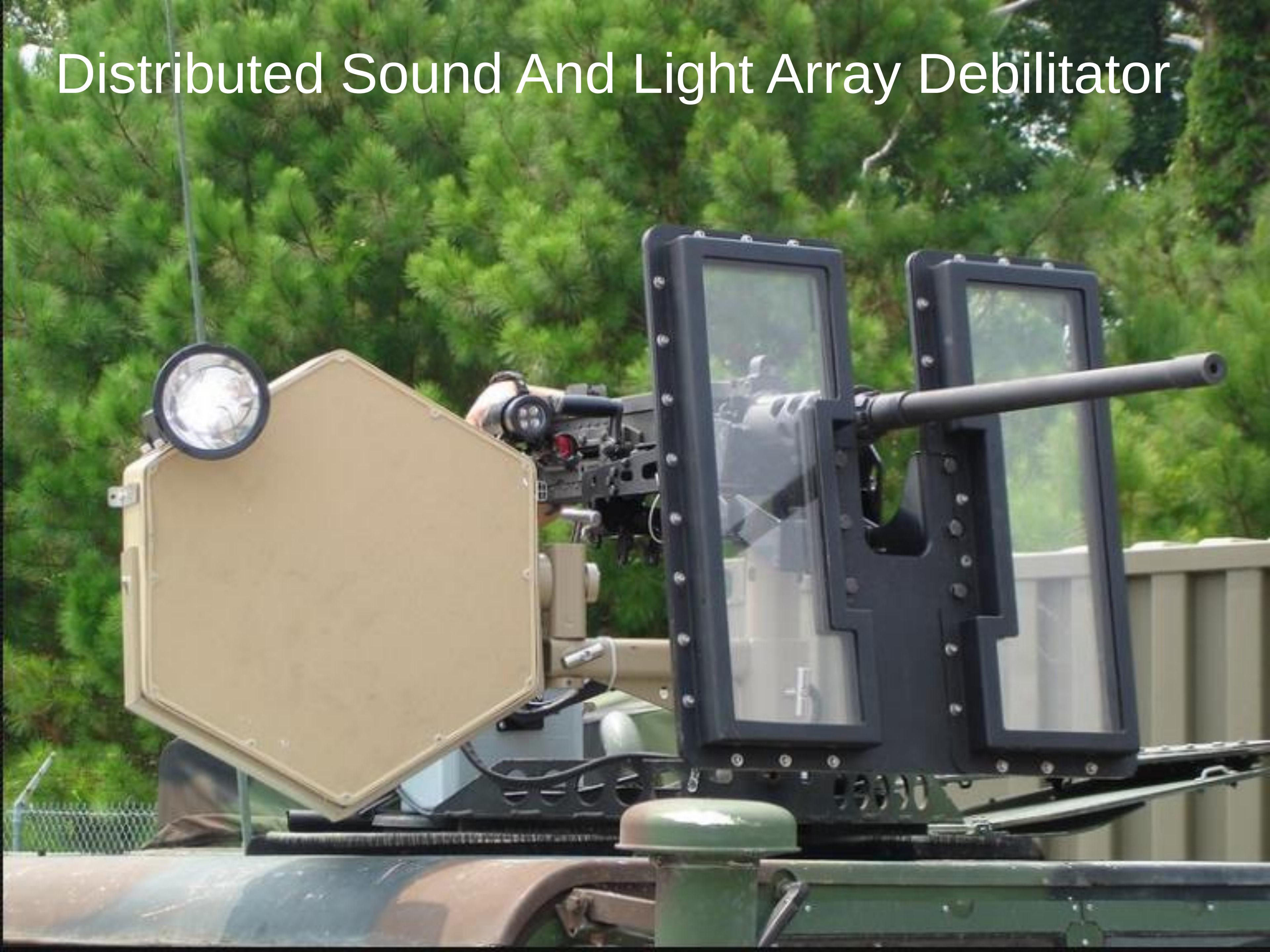

The most effective sensory weapons are ones that can engage and disorient multiple methods of input as well. Colloquially known as the “puke ray,” the Distributed Sound And Light Array Debilitator generates sound and light in a specific sequence designed to overload the brain’s processors, generating feelings of unease and discomfort. Prolonged exposure can result in even more unpleasant effects, including nausea and vomiting. Some scientists are even working towards producing a portable version the size of a flashlight that can be used on an individual basis.

Pulsed Energy Projectiles

E. P. Industries, Inc./ Mission Research Corp.
04-18-2001

The sensory organ with the largest active area in the body is, of course, the sense of touch. The human body has millions of individual nerves that are used to transmit sensory data to the brain, and a new experimental weapon from the U.S. government aims to really get them going. Pulsed energy projectiles are a specially-modulated laser that generates a burst of expanding plasma when they hit a solid object. That plasma in turn emits an electromagnetic pulse that causes immediate, searing pain to nerve cells all over the target's body. They're not in use yet, but research has been very promising.

MEDUSA

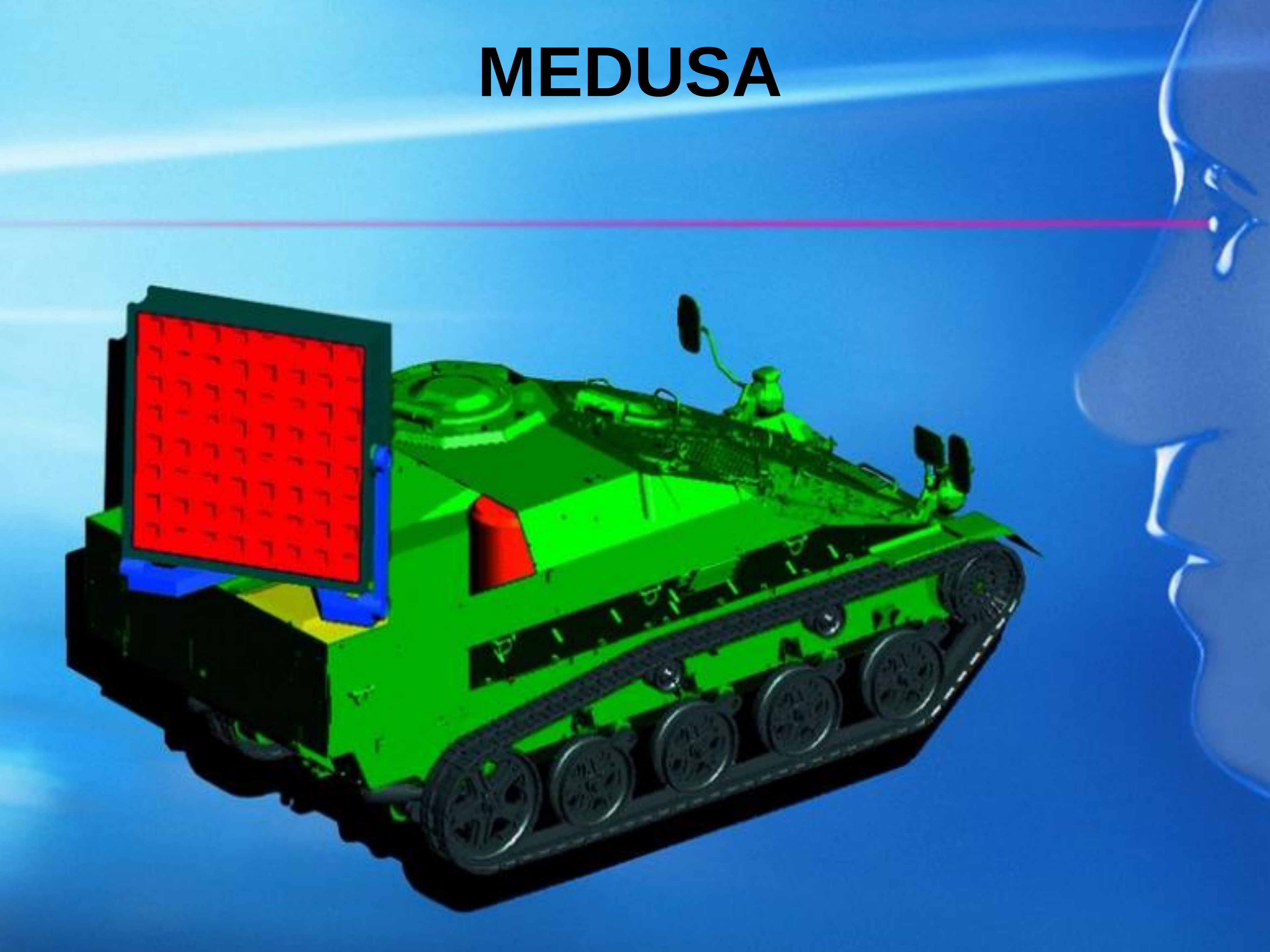

One of the most interesting sensory weapons on the market uses audio to disorient and confuse targets, but it skips the ear entirely. MEDUSA, a product of the Sierra Nevada Corporation, doesn’t create sound in a conventional way. Instead, it uses pulses of microwave radiation to generate shockwaves inside the skull that are detected by the ears from within – so earplugs won’t help you at all. It can generate painful white noise and even recognizable speech that sounds like it’s coming from the inside of your brain. Scary stuff.

"The Mind Has No Firewall" Army article on psychotronic weapons (Parameters)

The human body, much like a computer, contains myriad data processors. They include, but are not limited to, the chemical-electrical activity of the brain, heart, and peripheral nervous system, the signals sent from the cortex region of the brain to other parts of our body, the tiny hair cells in the inner ear that process auditory signals, and the light-sensitive retina and cornea of the eye that process visual activity.

We are on the threshold of an era in which these data processors of the human body may be manipulated or debilitated. Examples of unplanned attacks on the body’s data-processing capability are well-documented. Strobe lights have been known to cause epileptic seizures. Not long ago in Japan, children watching television cartoons were subjected to pulsating lights that caused seizures in some and made others very sick.

http://dprogram.net/2009/07/01/the-mind-has-no-firewall-army-article-on-psychotronic-weapons/

TYPES OF PSYCHOTRONIC WEAPONS

A psychotronic generator produces a powerful electromagnetic emanation capable of being sent through telephone lines, TV, radio networks, supply pipes, and incandescent lamps.

An autonomous generator is a device that operates in the 10-150 Hertz band. At the 10-20 Hertz band, it forms an infrasonic oscillation that is destructive to all living creatures.

A nervous system generator is designed to paralyze the central nervous systems of insects which could have the same applicability to humans.

Ultrasound emanation devices using ultrasound emanations are capable of carrying out bloodless internal operations without leaving a mark on the skin. They can also be used to kill people from a distance.

Noiseless cassettes is the ability to place infra-low frequency voice patterns over music, patterns that are detected by the subconscious. These messages are then stored in a person’s sub-conscious mind which they may later act upon.

MICHIGAN LAW AGAINST ELECTROMAGNETIC DEVICES

House Bill 4513 would amend the penal code (MCL 750.200h) to define “harmful electronic or electromagnetic device” as a device designed to emit or radiate an electronic or electromagnetic pulse or signal or microwave intended to cause harm to others or cause damage to, destroy, or disrupt any electronic or telecommunications system or device including a computer or computer network. The bill would also include a harmful electronic or electromagnetic device in the definition of “imitation harmful substance or device.”

The Nuremberg Code opposes all forms of human experiments

The state of Missouri has a law against involuntary chipping or implanting of a human

Many countries have laws against human experimentation and the use of electromagnetic devices to control a person either their mind or their bodies

SENATOR JOHN GLENN

Senator Glenn's publicly-expressed outrage that this government has been found to be engaging in brutal forms of involuntary human experimentation, and that any and all forms of this type of experimentation be exposed. He stated that there are experiments with non-ionizing, so-called "non-lethal," directed-energy weapons, surveillance and psychotronic systems. He opposes Microwave Harassment and Mind-Control Experimentation especially connected to Los

Alamos National Laboratory, the Department of Energy, the Department of Defense, CIA, and the National Reconnaissance Office. He stated the DOD has increased in the numbers of Ground Wave Emergency Network (GWEN) towers and satellites that qualify as directed-energy emitters, microwave towers, and antennae arrays. John Glenn is also a member of the Electronic Surveillance Project that is trying to stop these forms of harassment and human experimentation.

CITIZENS AGAINST HUMAN RIGHTS ABUSE

Scientists, many of which are Nobel Prize winners, in this Human Rights organization have documented for years the numerous abuses of electromagnetic weapons and electromagnetic radiation (EMR) mind control weapons. Most of the perpetrators of this abuse has been Russia and the United States. The International Red Cross has also verified these offenses.

http://mindjustice.org/emr13.htm

RUSSIAN DEFENSE MINISTER

In March 2012 the Russian defense minister Anatoli Serdjukov said:

“The development of weaponry based on new physics principles; direct-energy weapons, geophysical weapons, wave-energy weapons, genetic weapons, psychotronic weapons, etc., is part of the state arms procurement program for 2011-2020,”Voice of Russia

EUROPEAN PARLIAMENT

A directed energy weapon system is designed to match a radio frequency source to interfere with human brain activity at a synapse level.” Since 1999 those weapons have been upgraded and should be banned in the world.

UNITED NATIONS

The United Nations also agreed with this report and stated the “United States is the main country of the world to use directed energy weapons and electromagnetic weapons.

The former United States Secretary of Defense William Cohen stated in a press briefing, while commenting on new technological threats possibly held by terrorist organizations: *"Others are engaging in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, (and) volcanoes remotely, using the use of electromagnetic waves."*

A TARGETED INDIVIDUAL

Targeted Individuals are people who are illegally and covertly stalked, tortured and harassed by the use of hi-tech, Modern-day Military type Technologies (DAARPA, NASA, CIA etc.). Targets/Victims are usually implanted with **RFID** *(Radio-Remote Frequency Identification Device)* type transmitter chips ***(PositiveID*** or VERi Chips) and sometimes miniature Nano-Biochips and Stimulators/Transmitters mostly for GPS tracking purposes. Targets can be remotely connected to via radio frequency signals sent to the chip/s thus the mind and body being in theory "wirelessly" connected.. controlled and manipulated.

There are targeted individuals, victims of Psychotronic (or Mind Control, Electronic, Electromagnetic, Directed Energy, Neurological, Non-lethal) Weapons, who are being targeted unjustly or used as human subjects in experiments without their informed consent. Victims of mind/body control detail the most extreme and totalitarian violations of human rights in human history. Criminals may implant people with microchips or nanomaterials and place them under continuous surveillance. They monitor the human brain, continuously alter consciousness and behavior, directly assault and torture mind and bodies of the human subjects.

These are some of the symptoms that targeted individuals may experience:

a). Microwave hearing, torture with loud noise.
b). Visual hallucinations, artificial dreams.
c). Reading thoughts, retrieving memories.
d). Manipulation of thoughts, will, emotions, desires and perceptions.
e). Manipulation of human behavior: forced speech, involuntary body movements, transmission of specific commands into the subconscious, compulsory execution of these commands (i.e. the Manchurian Candidate effect).

f). Debilitation of mental acuity, loss of memory.

g). Sleep deprivation or uncontrollable sleep

h). Cramps, spasms, excruciating artificial pain in any part of the body, manipulation of body systems, including heart attacks and other serious medical conditions.

TARGETED INDIVIDUAL SUPPORT GROUPS

http://www.freedomfchs.com/id19.html

http://www.stopcovertwar.com/support.html

https://www.facebook.com/victimsoforganizedstalking

http://www.targetedinamerica.com/hitlist.html

http://targetedindividualsua.weebly.com/

http://targetedindividuals.blogspot.com/2011/08/tis-support-group.html

Dennis M. Bushnell
Chief Scientist
NASA Langley Research Center

Future Strategic Issues/Future Warfare

© Getty Images

CURRENTLY

- Order of 70% of Worlds Research conducted outside of U.S. (to first order, a % of GDP, U.S. produces order of 18% of worlds GDP)
- Order of 70% of U.S. Research now "Commercial" (as opposed to Government sponsored)

TYPES OF WARFARE

- **Hunter-Gatherer - “Nature Provided”**
- **Agriculture - Controlled Nature (Plants/Animals)**
- **Industrial - Mechanized Agriculture**
- **IT/BIO/Nano - Automating Industry/Agriculture**
- **Virtual - Robotization of IT/Bio/Nano/Industry/Agriculture**

Worldwide IT Revolution

- **Comms/Computing/Sensors/Electronics**
- **U.S. Commercial IT R&D ~ $100B/yr.**
- **Factor of 1 Million further improvement [Silicon,Molecular,Quantum,Bio,Optical]**
- **Beyond Human AI?**
- **Automatics/Robotics "in the large"**
- **Immersive multi-sensory VR/"Holodecks"**
- **Ubiquitous multi physics/hyperspectral sensors [land/sea/air/space]**
- **Micro/Nano sats/GNC/sensors,etc.**

Current Competitive Landscape

- U.S. produces only 18% of Worlds GDP
- ~70% of Research conducted offshore
- $300B/yr trade deficit
- 32 other nations devote a larger % of their GDP to Research
- 5th in No of R&D personnel/labor unit
- 3% savings rate vs. 30% in Asia
- Proliferation of IT,bio,nano,Space Technology etc.

Bio Revolution Applications

- “Pharm Animals” [drugs, spare parts]
- Fast Growing plants on/near sea surface & sea water irrigated plants for biomass energy/closed CO2 cycle
- Polymer growing plants
- Spider genes in goats allow spider silk spinning from goat milk for “Biosteel”, 3.5X strength of aramid fibers for Armor
- Binary Bio-weaponry

A Real Transgenic Spider Goat

Carbon Nanotubes

- C1,000,000, Buckminister Fullerine Carbon
- 100X strength, 1/6 weight of steel
- 8X better Armor
- Low energy Molecular/Petaflop Computing
- Ultra Capacitor/High Temperature SC
- Non-Cryo H2 storage

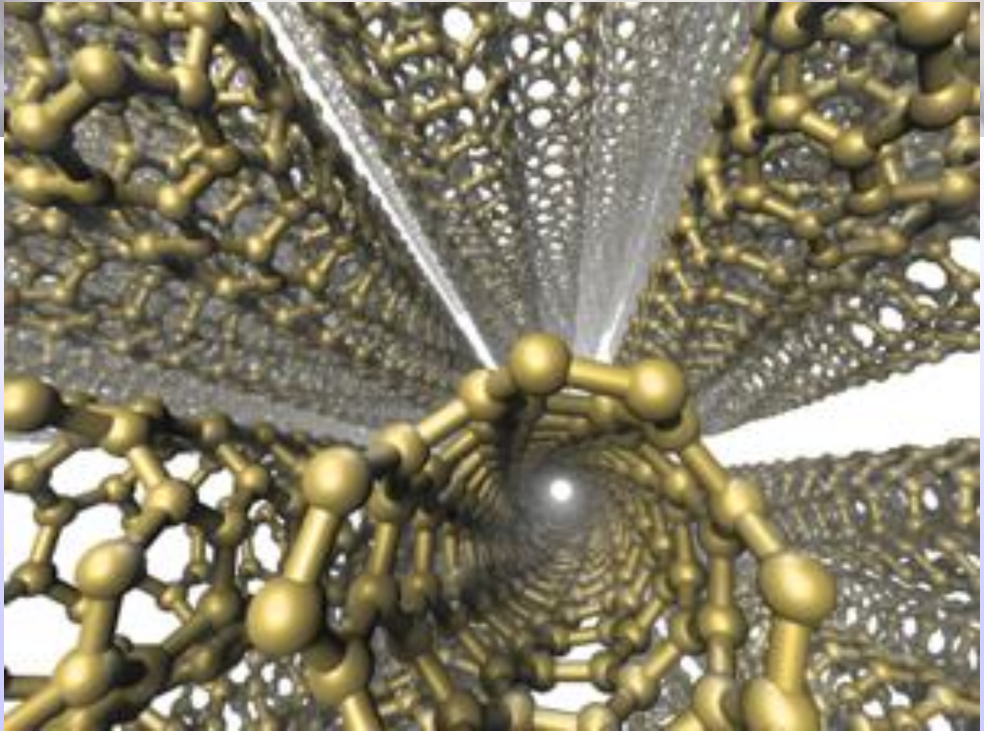

Some Sensor "Swarms"

- SMART DUST
  - Cubic mm or less
  - Combined sensors, comms and power supply
  - Floats in air currents for up to 2 years
- NANOTAGS
  - Placed on everything/everywhere
  - Identification and Status Info
- Co-opted INSECTS

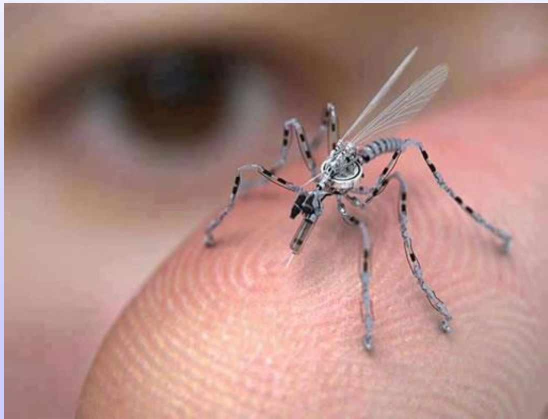

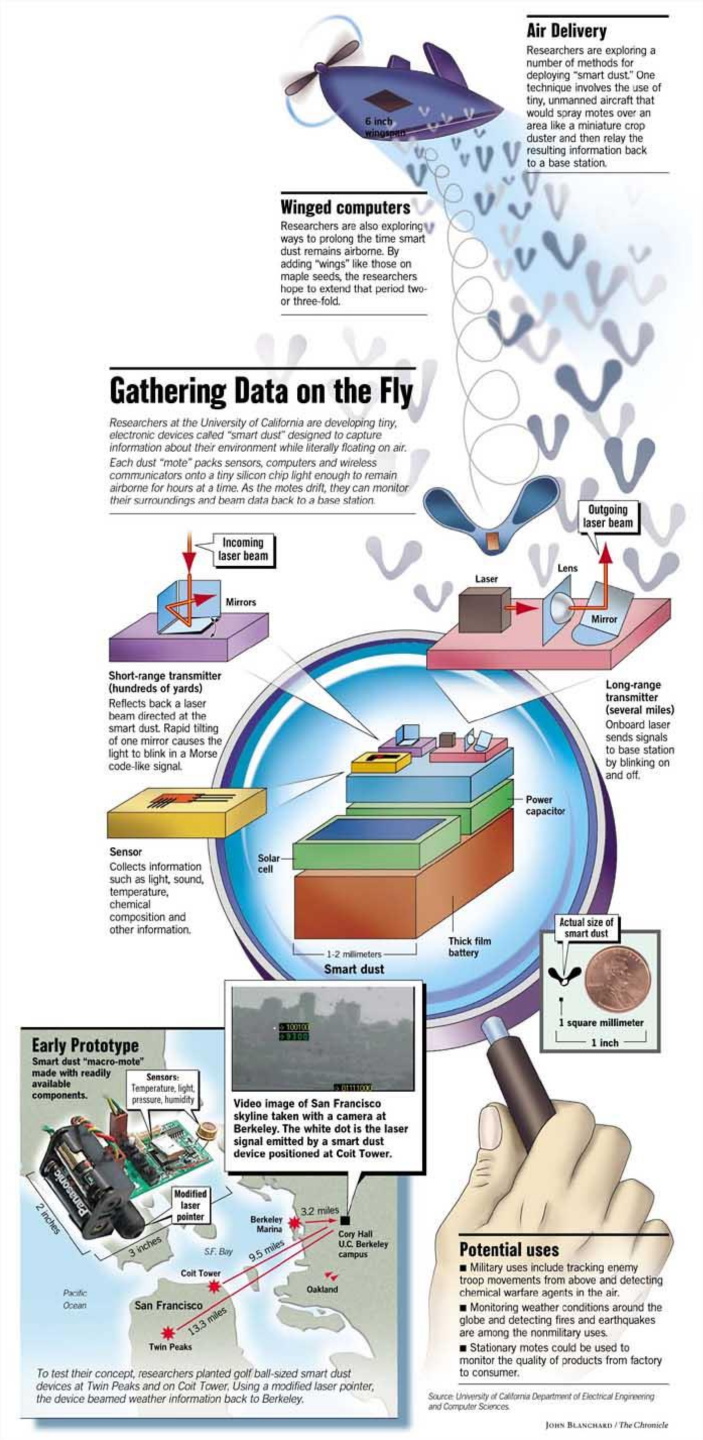

Air Delivery

Researchers are exploring a number of methods for deploying "smart dust." One technique involves the use of tiny, unmanned aircraft that would spray motes over an area like a miniature crop duster and then relay the resulting information back to a base station.

Winged computers

Researchers are also exploring ways to prolong the time smart dust remains airborne. By adding "wings" like those on maple seeds, the researchers hope to extend that period two- or three-fold.

Gathering Data on the Fly

Researchers at the University of California are developing tiny, electronic devices called "smart dust" designed to capture information about their environment while literally floating on air.

Each dust "mote" packs sensors, computers and wireless communicators onto a tiny silicon chip light enough to remain airborne for hours at a time. As the motes drift, they can monitor their surroundings and beam data back to a base station.

Short-range transmitter (hundreds of yards)
Reflects back a laser beam directed at the smart dust. Rapid tilting of one mirror causes the light to blink in a Morse code-like signal.

Long-range transmitter (several miles)
Onboard laser sends signals to base station by blinking on and off.

Sensor
Collects information such as light, sound, temperature, chemical composition and other information.

Early Prototype

Smart dust "macro-mote" made with readily available components.

Video image of San Francisco skyline taken with a camera at Berkeley. The white dot is the laser signal emitted by a smart dust device positioned at Coit Tower.

To test their concept, researchers planted golf ball-sized smart dust devices at Twin Peaks and on Coit Tower. Using a modified laser pointer, the device beamed weather information back to Berkeley.

Potential uses

- Military uses include tracking enemy troop movements from above and detecting chemical warfare agents in the air.
- Monitoring weather conditions around the globe and detecting fires and earthquakes are among the nonmilitary uses.
- Stationary motes could be used to monitor the quality of products from factory to consumer.

Source: University of California Department of Electrical Engineering and Computer Sciences.

JOHN BLANCHARD / *The Chronicle*

6 inch wingspan
Incoming laser beam
Mirrors
Outgoing laser beam
Laser
Lens
Mirror
Power capacitor
Solar cell
Thick film battery
1-2 millimeters
Smart dust
Actual size of smart dust
1 square millimeter
1 inch
Sensors: Temperature, light, pressure, humidity
Modified laser pointer
2 inches
3 inches
Panasonic
> 100100
> 0111100
Berkeley Marina
3.2 miles
Cory Hall U.C. Berkeley campus
S.F. Bay
9.5 miles
Coit Tower
Oakland
Pacific Ocean
San Francisco
13.3 miles
Twin Peaks

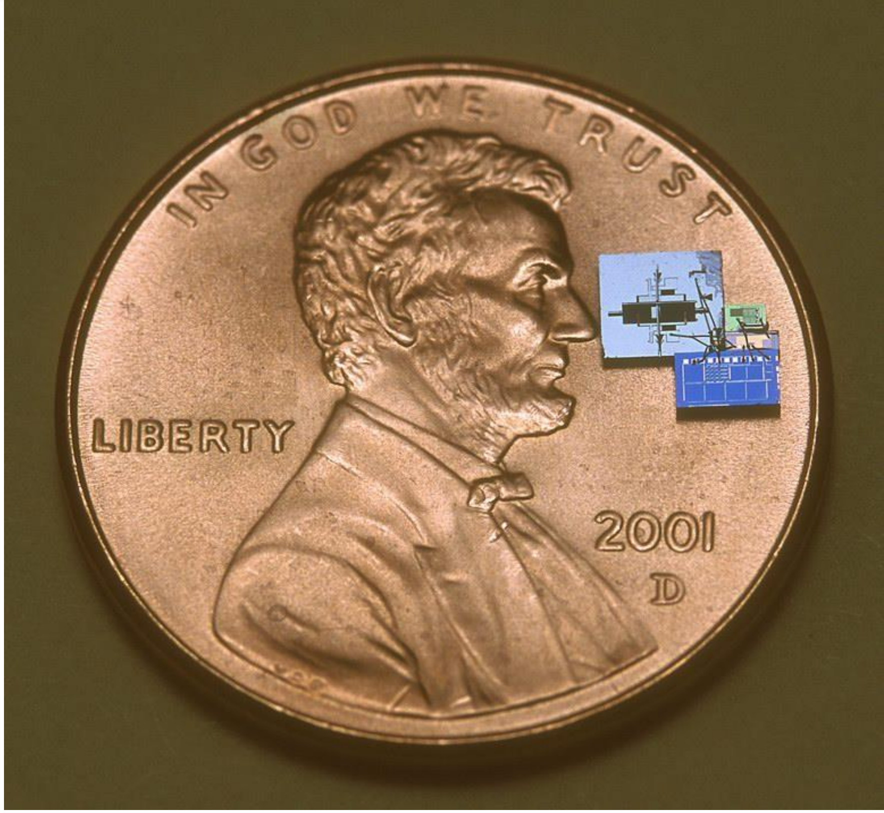
IN GOD WE TRUST
LIBERTY
2001
D

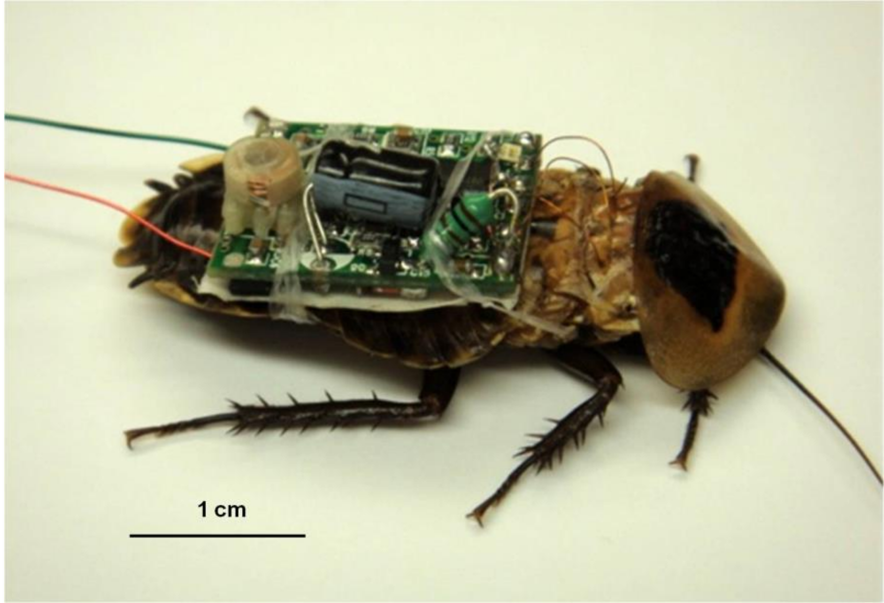

“Volumetric” Weaponry

[Alternatives to HE]

- EMP
- Info/Net/Psy warfare
- Miniature brilliant sensor/mine combo’s
- Fuel/air & dust/air
- RF
- Chem/bio Antifunctionals/antifauna
- Isomers, Strained Bond Energy Release, etc.
- Carbon fibers/Acoustics etc.

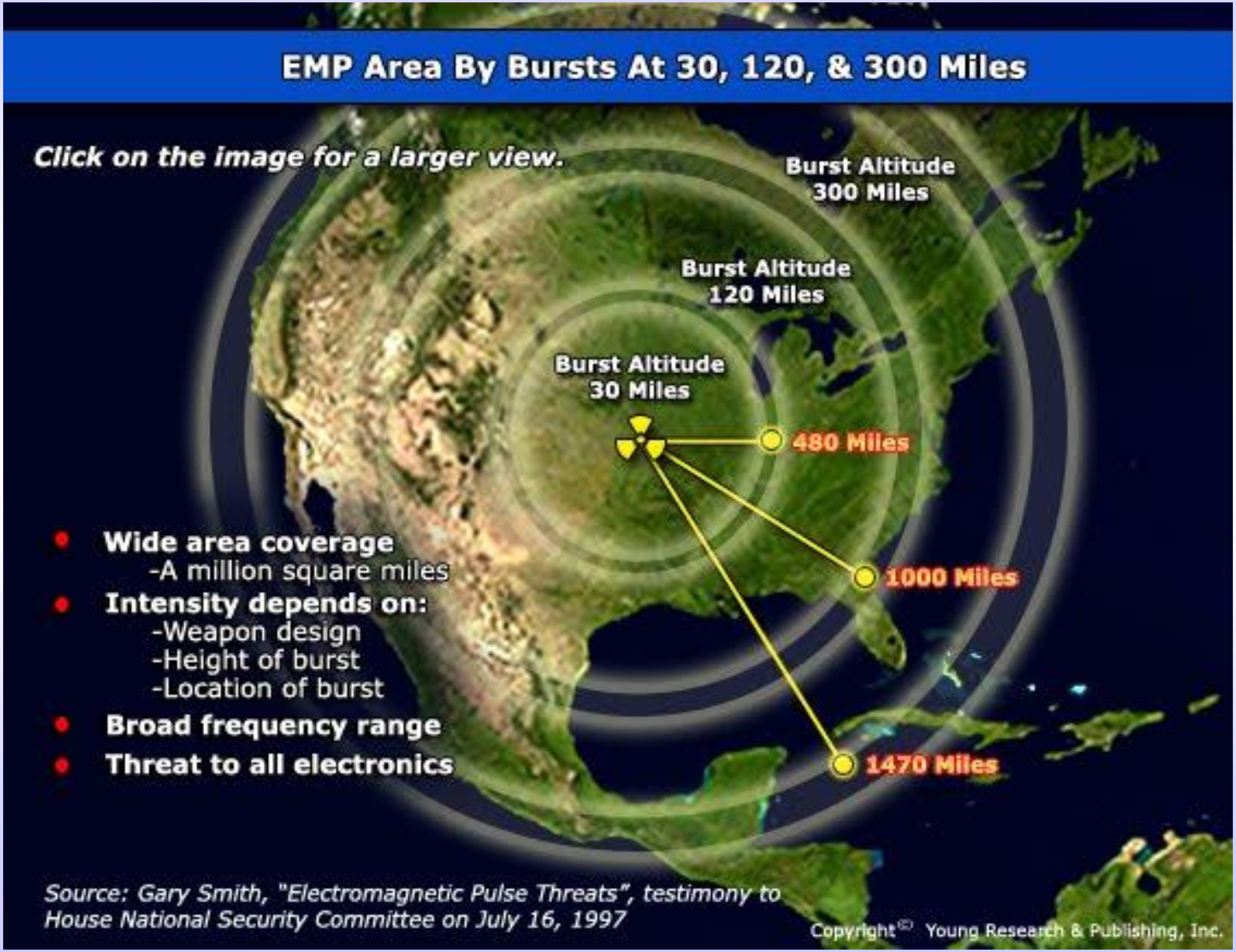

EMP Area By Bursts At 30, 120, & 300 Miles

Click on the image for a larger view.

- **Wide area coverage**
  - -A million square miles
- **Intensity depends on:**
  - -Weapon design
  - -Height of burst
  - -Location of burst
- **Broad frequency range**
- **Threat to all electronics**

Burst Altitude 300 Miles
Burst Altitude 120 Miles
Burst Altitude 30 Miles
480 Miles
1000 Miles
1470 Miles

Source: Gary Smith, "Electromagnetic Pulse Threats", testimony to House National Security Committee on July 16, 1997

Copyright© Young Research & Publishing, Inc.

“Slingatron” for Global Precision Strike

- 10Kg projectiles, up to thousands/minute
- Global, or less, range
- $20M/device
- Mechanical “on-the-ground” propulsion via Gyrating Spiral Guide Tube (a multiple “hula hoop”
- “Poor Mans” Global Precision Strike/“Takedown Weapon”

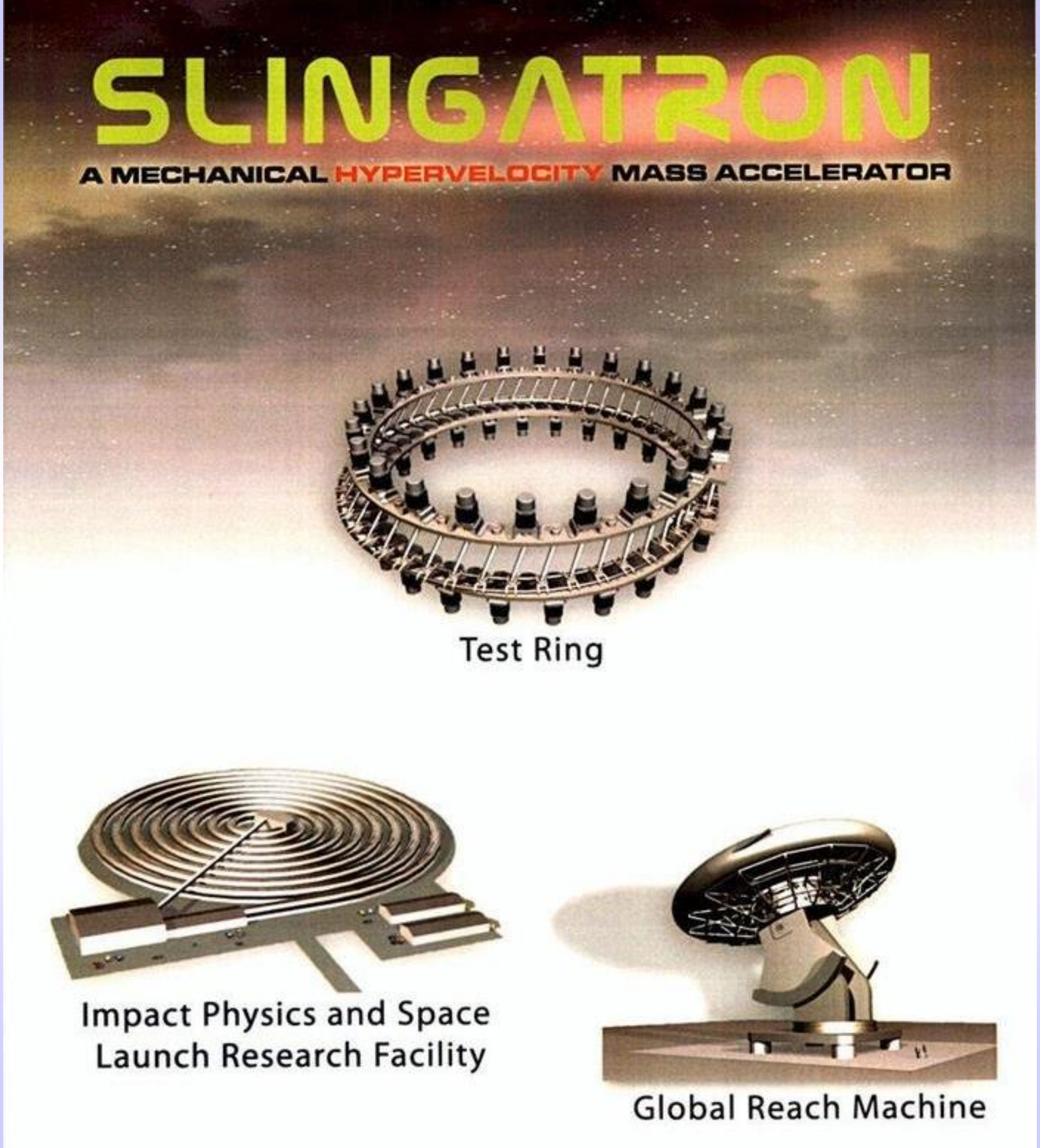

SLINGATRON

A MECHANICAL HYPERVELOCITY MASS ACCELERATOR

Test Ring

Impact Physics and Space
Launch Research Facility

Global Reach Machine

Potential Future “Orders of Magnitude” Increases in Overall Weapon Effectiveness/Availability at Orders of Magnitude Reduced Cost(s)

- Bio/Chem/Molec./Nano Computing - (E6)
- Ubiquitous Optical Comms - (E4)
- Micro/Nano/Ubiquitous Sensors - (E4)
- BioWeaponry - (EN)
- Co-operative Swarms of Cheap/Small Weapons/Sensors - (E4)
- Volumetric Weaponry - (E4)
- Cyber/Artificial Life (Beyond AI) - (E?)

Chemtrails

Inside of a US Chemtrail Plane

Major U.S. Future Warfare Issues

- **“Non-explosive Warfare” (psywar, biowar IT/net war, “anti-operability war,” Beam weaponry including RF, Spoofing/Cammo**
- **Robotic Warfare “in the large”/better than human AI/“Cyber life”**

Worth1000.com

DARPA Laser Weapon

ROBOTIC SOLDIERS

Robotic soldiers are being created right now in the world to build a huge army against humans and God. The New World Order is not going to reveal their plans on the evening news!

SINGULARITY

The **technological singularity** is a theoretical moment in time when artificial intelligence will have progressed to the point of a greater-than-human intelligence that will "radically change human civilization” according to Amnon Eden. Ray Kurzweil says “revolutions in genetics, nanotechnology and robotics will usher in the beginning of the Singularity.” Robots and cybernetic organisms are developing at a very fast pace that is blurring the line between humans and robots.

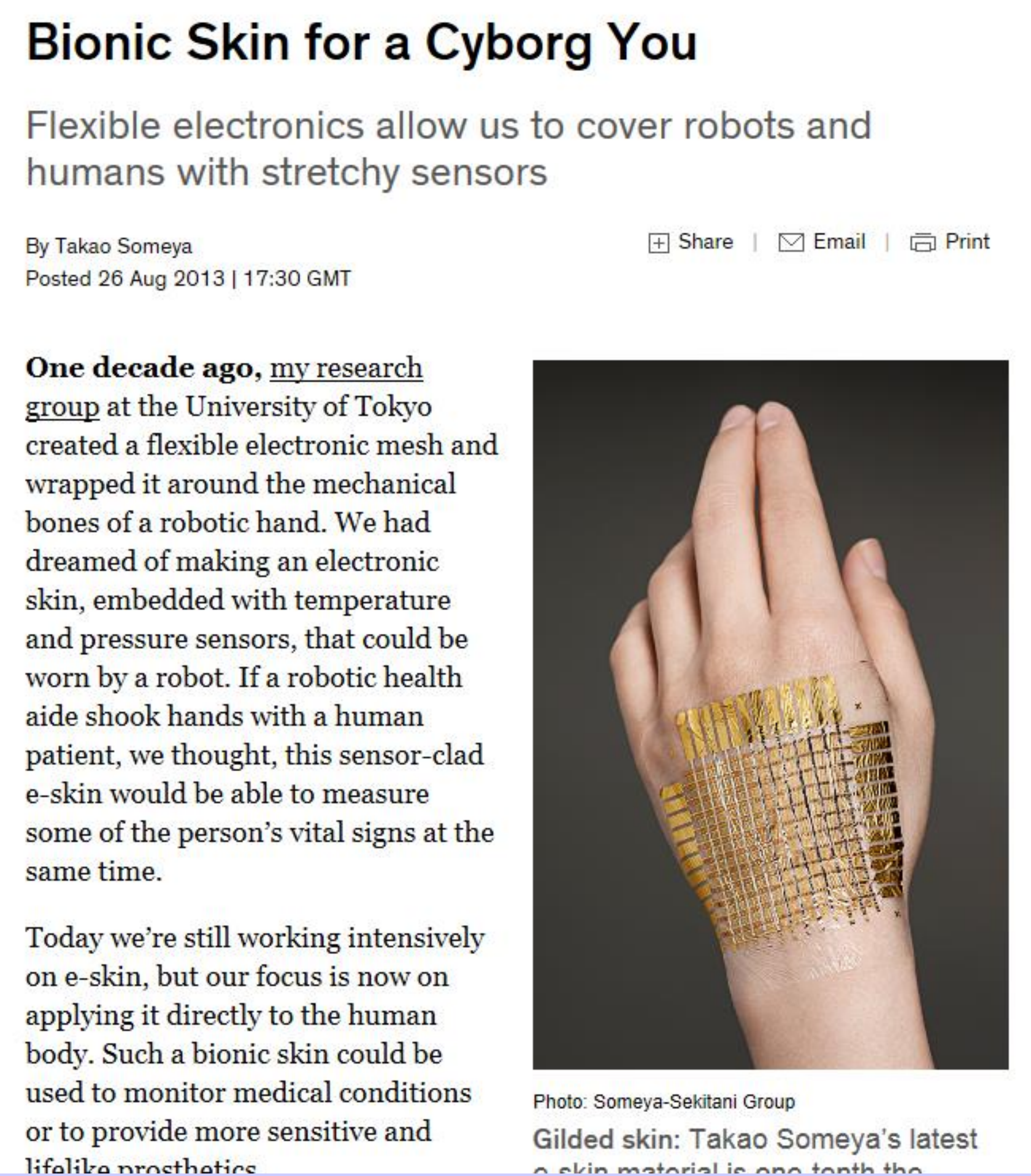

Bionic Skin for a Cyborg You

Flexible electronics allow us to cover robots and humans with stretchy sensors

By Takao Someya
Posted 26 Aug 2013 | 17:30 GMT

One decade ago, my research group at the University of Tokyo created a flexible electronic mesh and wrapped it around the mechanical bones of a robotic hand. We had dreamed of making an electronic skin, embedded with temperature and pressure sensors, that could be worn by a robot. If a robotic health aide shook hands with a human patient, we thought, this sensor-clad e-skin would be able to measure some of the person's vital signs at the same time.

Today we're still working intensively on e-skin, but our focus is now on applying it directly to the human body. Such a bionic skin could be used to monitor medical conditions or to provide more sensitive and lifelike prosthetics

Photo: Someya-Sekitani Group

Gilded skin: Takao Someya's latest e-skin material is one-tenth the

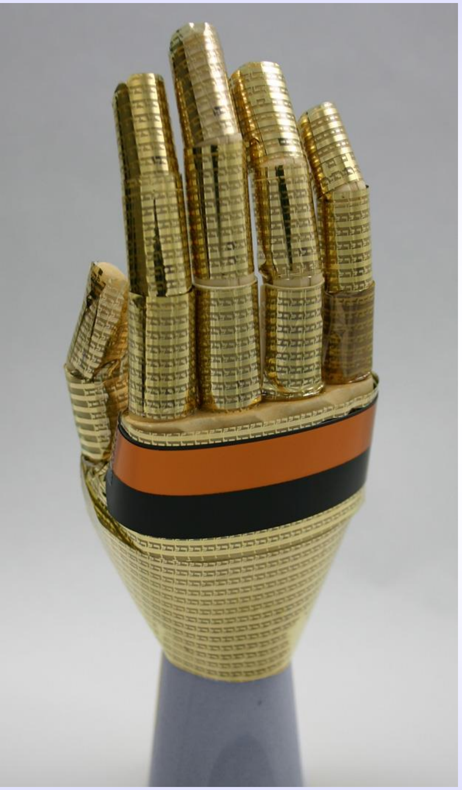

Today we’re still working intensively on e-skin but our focus is now on applying it directly to the human body. Such a bionic skin could be used to monitor medical conditions or to provide more sensitive and lifelike prosthetics.

Gilded skin: Takao Someya’s latest e-skin material is one-tenth the thickness of plastic kitchen wrap, and it can conform to any body shape.

Thin-film transistors don’t just allow electronics to be flexible—they can also help an e-skin mimic the sensitivity of real skin. Consider this: There are more than 2 million pain receptors in a person’s skin, which is equivalent to the number of pixels found in a typical high-definition TV. Synthetic skin covers the circuits that looks like human skin.

http://spectrum.ieee.org/biomedical/bionics/bionic-skin-for-a-cyborg-you

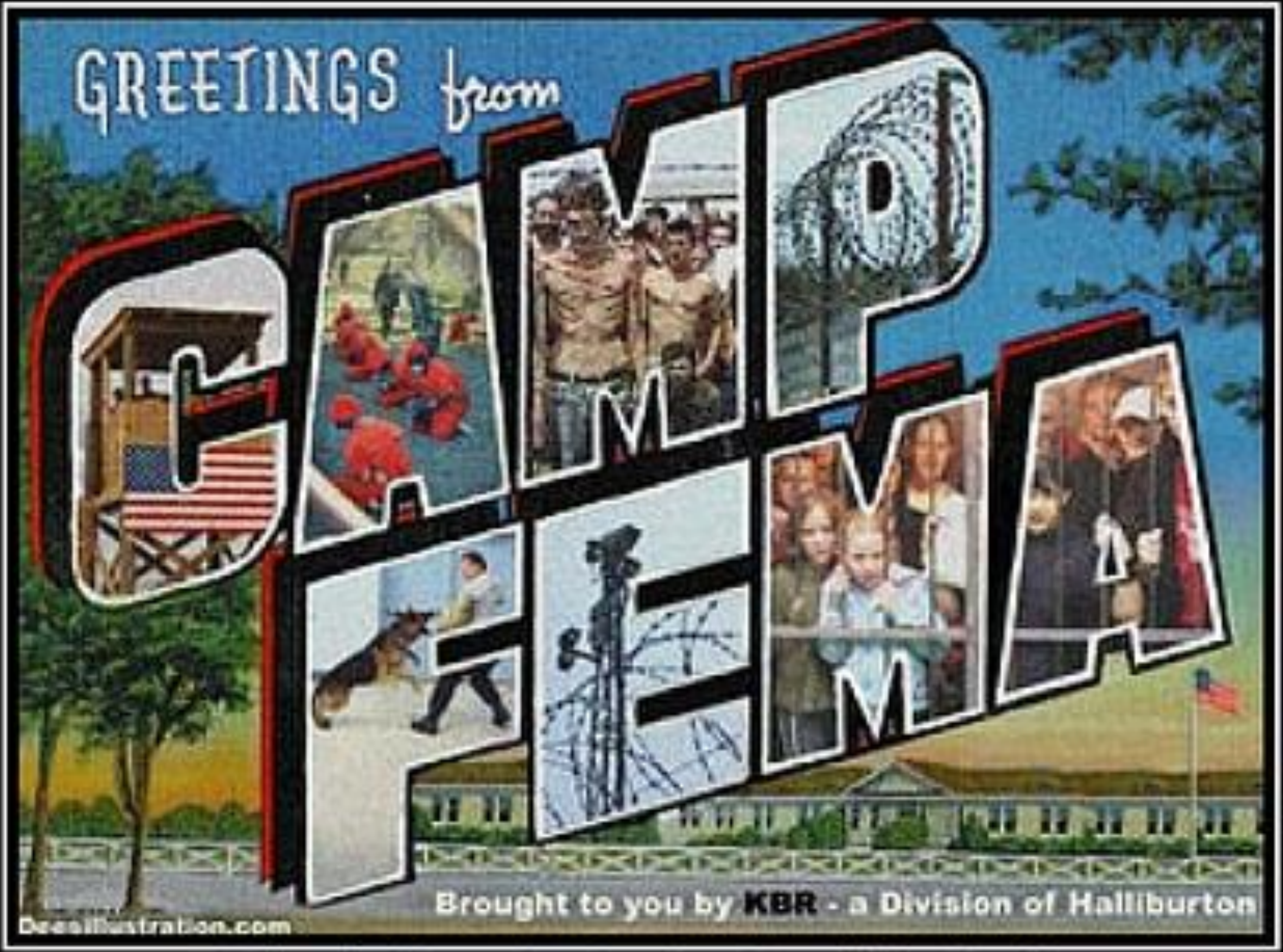
GREETINGS from
CAMP
FEMA
Brought to you by KBR - a Division of Halliburton
Deesillustration.com

CAMP FEMA

Coming to your city soon!

motifake.com

FEMA CAMPS in the United States

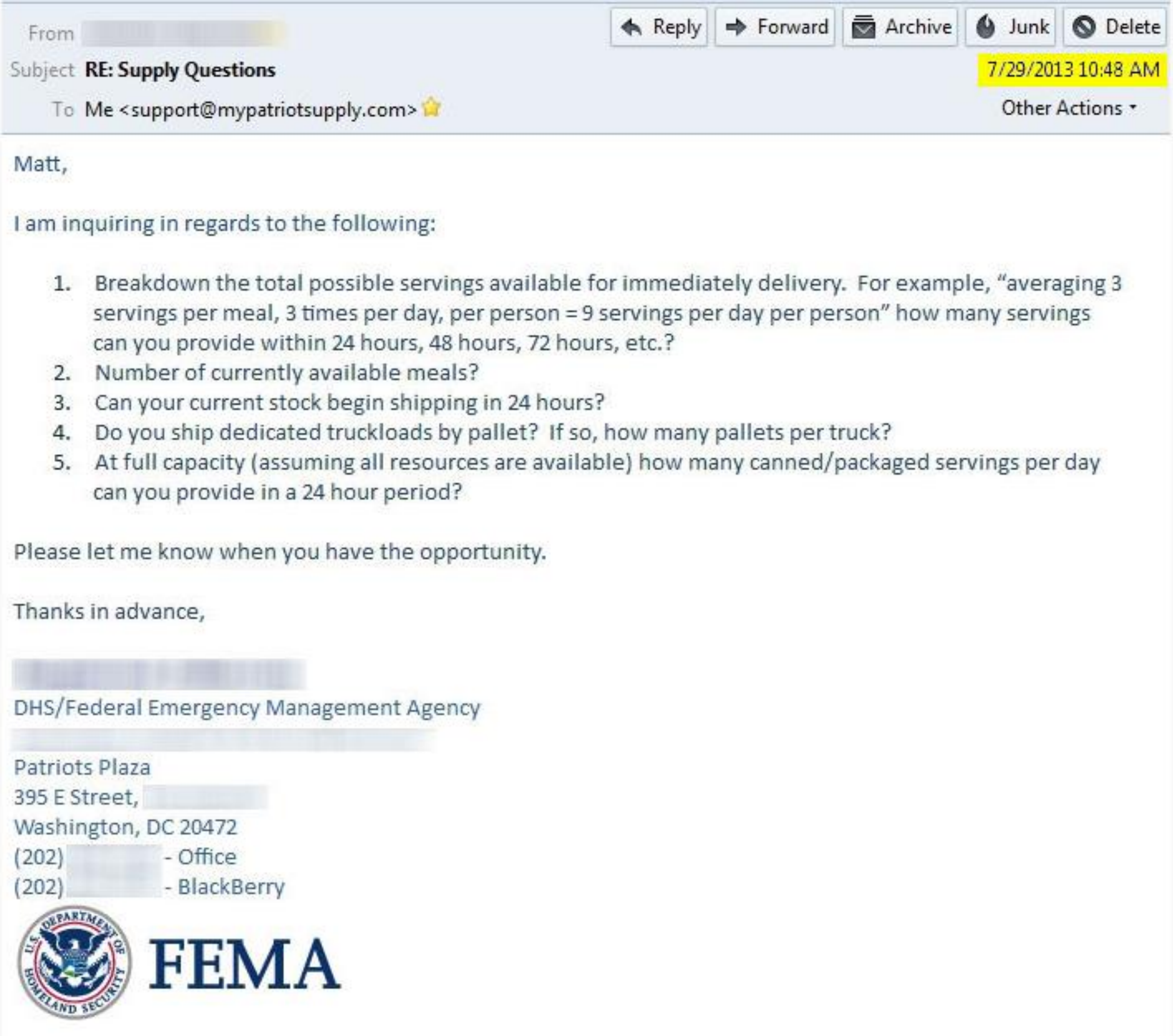

From [redacted]

Subject **RE: Supply Questions**

To Me <support@mypatriotsupply.com>

7/29/2013 10:48 AM

Matt,

I am inquiring in regards to the following:

1. Breakdown the total possible servings available for immediately delivery. For example, "averaging 3 servings per meal, 3 times per day, per person = 9 servings per day per person" how many servings can you provide within 24 hours, 48 hours, 72 hours, etc.?
2. Number of currently available meals?
3. Can your current stock begin shipping in 24 hours?
4. Do you ship dedicated truckloads by pallet? If so, how many pallets per truck?
5. At full capacity (assuming all resources are available) how many canned/packaged servings per day can you provide in a 24 hour period?

Please let me know when you have the opportunity.

Thanks in advance,

DHS/Federal Emergency Management Agency
Patriots Plaza
395 E Street, [redacted]
Washington, DC 20472
(202) [redacted] - Office
(202) [redacted] - BlackBerry

FEMA

FM 3-39.40

There is over 600 FEMA camps

INTERNMENT AND RESETTLEMENT OPERATIONS

FEMA has divided the U.S. into 10 regions with Governors assigned over each region

February 2010

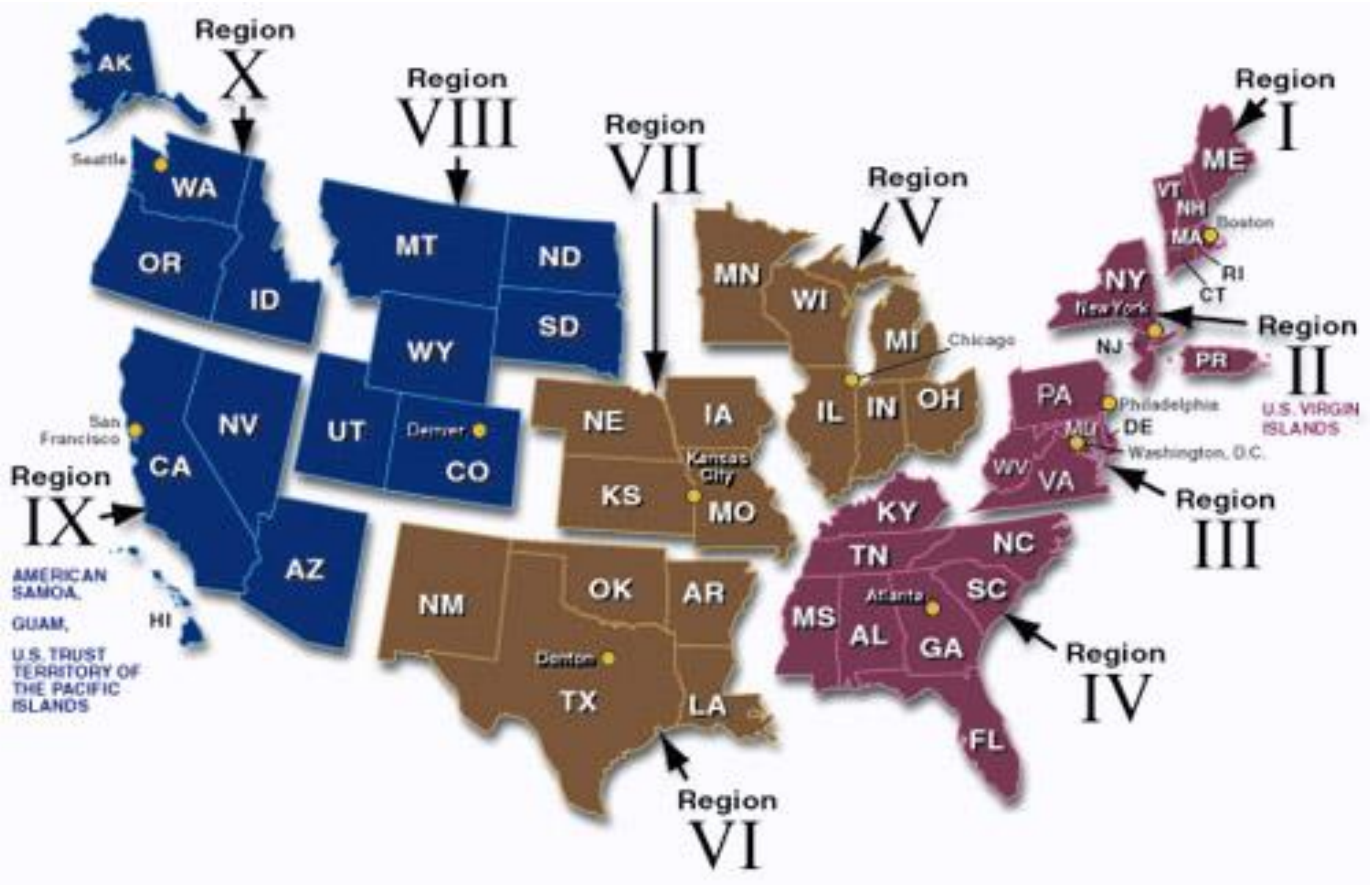
Region X
AK
Seattle
WA
OR
ID
Region VIII
MT
ND
SD
WY
UT
Denver
CO
Region IX
San Francisco
NV
CA
AZ
HI
AMERICAN SAMOA,
GUAM,
U.S. TRUST TERRITORY OF THE PACIFIC ISLANDS
Region VII
NE
IA
Kansas City
KS
MO
Region VI
NM
OK
AR
Denton
TX
LA
Region V
MN
WI
MI
Chicago
IL
IN
OH
Region IV
KY
TN
NC
SC
Atlanta
MS
AL
GA
FL
Region III
PA
Philadelphia
MD
DE
Washington, D.C.
WV
VA
Region II
NY
New York
NJ
PR
U.S. VIRGIN ISLANDS
Region I
ME
VT
NH
MA
Boston
RI
CT

10 FEDERAL REGIONS

ent: Military Internment Camps in U.S to be Used for Political Dissidents

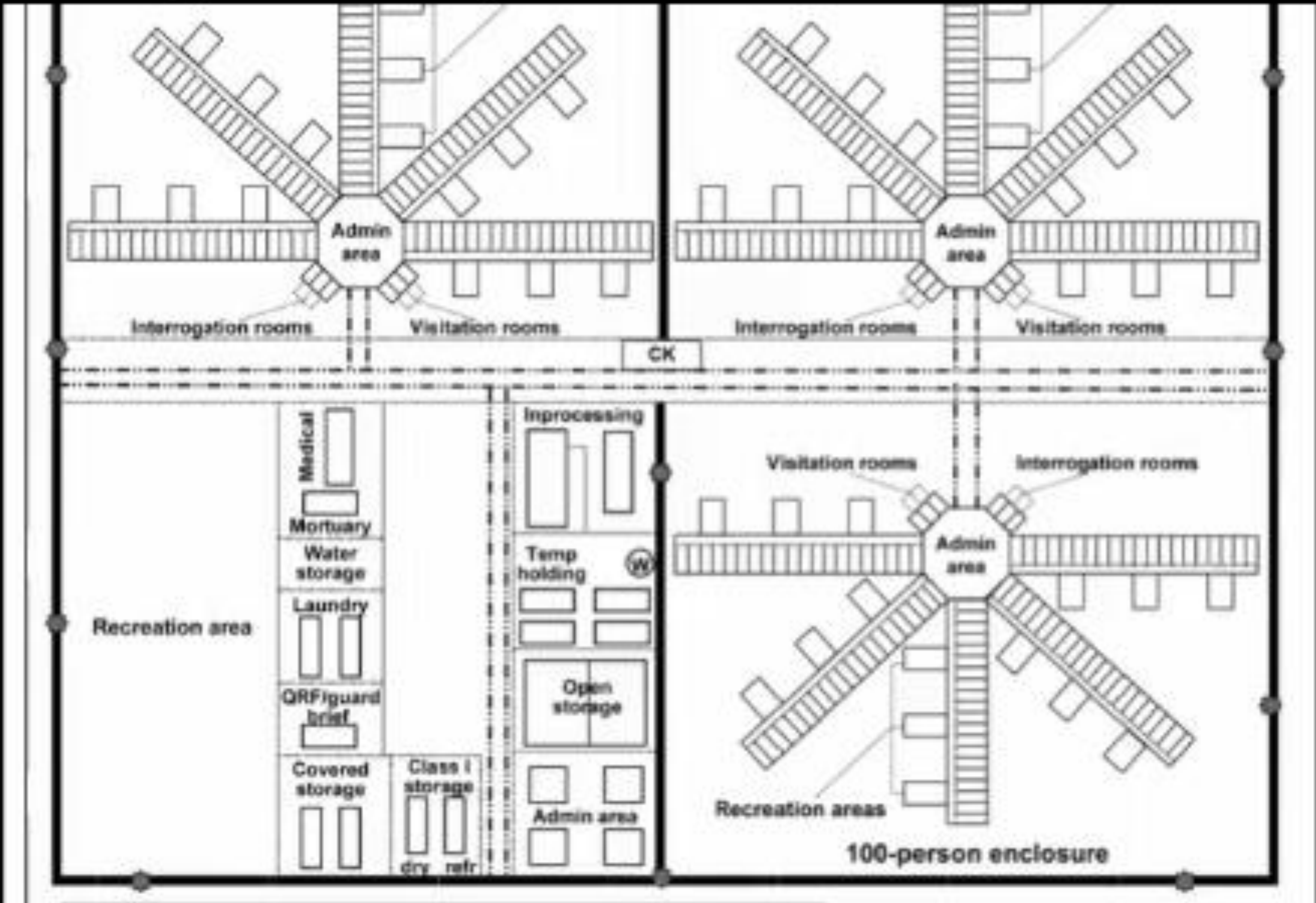
Admin area
Admin area
Interrogation rooms
Visitation rooms
Interrogation rooms
Visitation rooms
CK
Medical
Mortuary
Water storage
Laundry
QRF/guard brief
Covered storage
Class I storage
dry
refr
Recreation area
Inprocessing
Temp holding
W
Open storage
Admin area
Visitation rooms
Interrogation rooms
Admin area
Recreation areas
100-person enclosure

FM 3-39.40

INTERNMENT AND RESETTLEMENT OPERATIONS

February 2010

Internment and Resettlement Operations

Contents

Page

PREFACE viii

INTRODUCTION xi

Chapter 1 **INTERNMENT AND RESETTLEMENT AND THE OPERATIONAL ENVIRONMENT** 1-1

Conduct 1-1

Principles 1-3

Personnel Categories 1-5

Status Determination 1-7

Article 5 Tribunals 1-8

Appeals and Periodic Reviews of Civilian Internees 1-9

General Protection and Care of Detainees, U.S. Military Prisoners, and Dislocated Civilians 1-10

Agencies Concerned With Internment and Resettlement 1-12

Protecting Power 1-13

Planning Considerations for Internment and Resettlement Operations 1-14

Military Police Capabilities 1-16

Chapter 2 **INTERNMENT AND RESETTLEMENT IN SUPPORT OF THE SPECTRUM OF OPERATIONS** 2-1

Support to Combat Operations 2-1

Support to Stability Operations 2-3

Support to Civil Support Operations 2-8

Army Command and Support Relationships 2-8

Considerations Within the Operational Area and the Area of Operations 2-9

Chapter 3 **COMMAND AND STAFF ROLES AND RESPONSIBILITIES** 3-1

National and Theater Reporting Agencies 3-1

Roles and Responsibilities 3-2

Chapter 4 **CAPTURE, INITIAL DETENTION, AND SCREENING** **4-1**
Detainee Flow 4-1
Detainee Processing 4-5
Custody and Accountability of Property, Evidence, and Intelligence Information 4-12
Detainee Movement 4-15
Methods of Transportation 4-16
Detainee Release 4-19

Chapter 5 **DETAINEE OPERATIONS** **5-1**
Command and Control 5-1
Planning Considerations 5-2
Intelligence and Interrogation 5-3
Medical Support 5-6
Dental Support 5-8
Specific Detainee Support Requirements 5-9
Detainee Deaths 5-13
Legal Considerations 5-14

Chapter 6 **DETAINEE FACILITIES** **6-1**
General Considerations 6-1
Detainee Collection Point 6-4
Detainee Holding Area 6-10
Fixed Detainee Internment Facilities 6-14
Theater Internment Facility 6-17
Strategic Internment Facility 6-37
Transfers or Releases 6-38

Chapter 7 **CONFINEMENT OF U.S. MILITARY PRISONERS** **7-1**
U.S. Battlefield Confinement Operations Principles 7-1
Planning Process for U.S. Military Prisoners 7-1
Battlefield Facilities 7-2
Processing, Classification, and Identification Requirements 7-3
Clothing, Meals, and Dining Facilities 7-4
Medical Care and Sanitation 7-5
Discipline, Control, and Administration 7-6
Emergency Planning and Investigations 7-11
Rules of Interaction 7-12
Use of Force 7-12
Escape

If there are not real FEMA camps for detainment of American citizens, then why:

- Did Obama sign so many Executive Orders taking away the freedoms of American citizens?
- Did Obama say that anyone who is a threat to American security could be “detained indefinitely?”
- Does Homeland Security avoid answering a direct question about the hundreds of camps that exist?
- Are so many camps totally empty but staffed 24 hours a day?

- Why did they not house a single illegal immigrant of the 12 million that are here in any of these camps even though they say that is what they were built for?
- Why didn’t any of those effected by hurricanes go to a FEMA camp?
- Why does Homeland Security keep on building more FEMA camps?
- Why are a majority of these camps designed to hold 20,000 inmates?

- Why aren’t the people paid to station these camps informed about the purpose of the camps?
- Why are these camps hiring so many guards recently?
- Why aren’t the victims of fires, flooding, tornadoes, and hurricanes taken to any of these camps?
- Why did the sponsors of the bill years ago to build these camps even know what these camps were designed for?

LIST OF ALL OF THE CAMPS:

http://www.sianews.com/modules.php?name=News&file=article&sid=1062

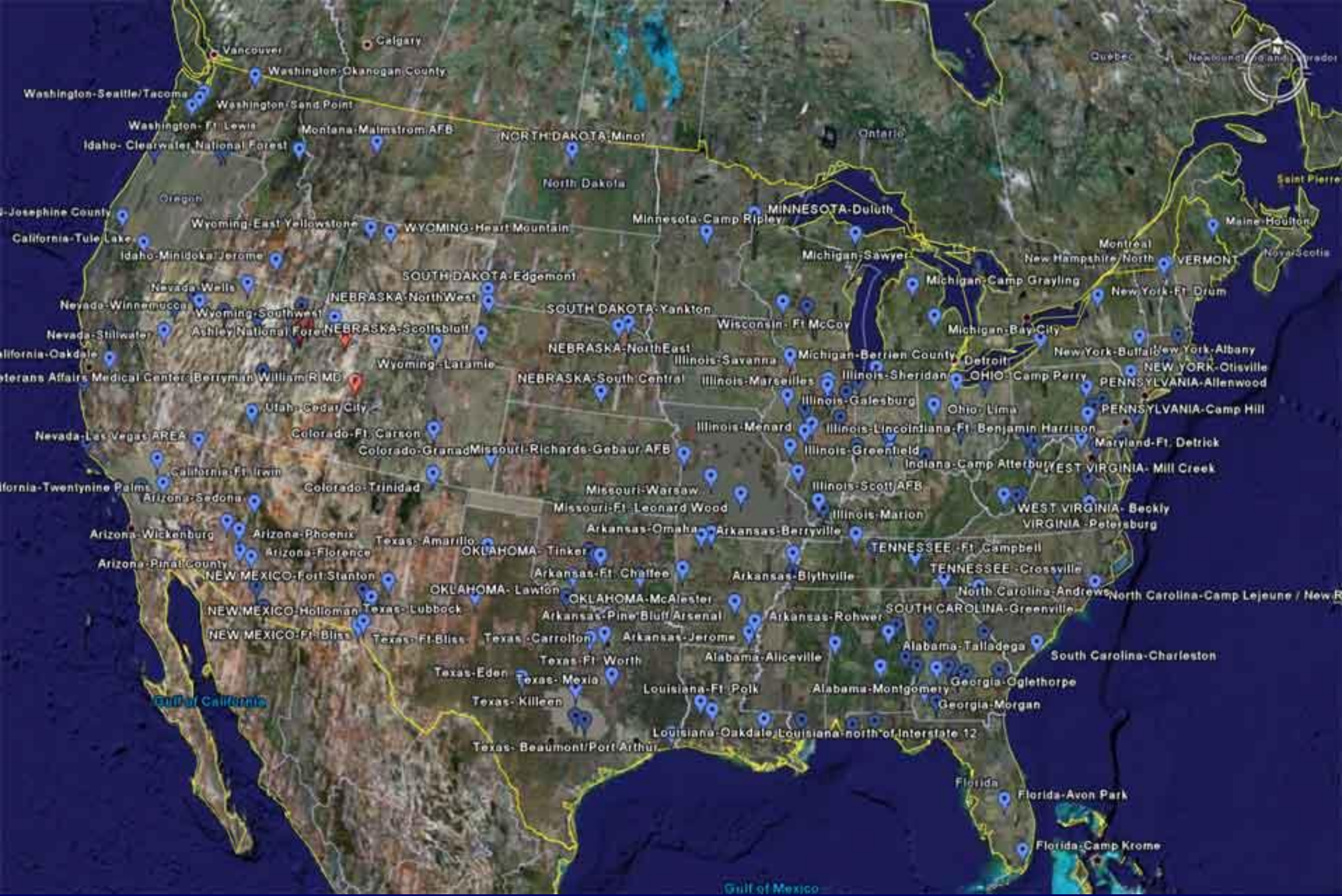
Calgary
Vancouver
Washington-Okanogan County
Washington-Seattle/Tacoma
Washington Sand Point
Washington- Ft Lewis
Montana-Malmstrom AFB
NORTH DAKOTA-Minot
Idaho- Clearwater National Forest
Quebec
Newfoundland and Labrador
Ontario
Saint Pierre
North Dakota
Oregon
-Josephine County
Wyoming-East Yellowstone
WYOMING-Heart Mountain
Minnesota-Camp Ripley
MINNESOTA-Duluth
Maine-Houlton
California-Tule Lake
Idaho-Minidoka Jerome
Michigan-Sawyer
Montreal
New Hampshire North
VERMONT
Nova Scotia
SOUTH DAKOTA-Edgemont
Michigan-Camp Grayling
Nevada-Wells
NEBRASKA-NorthWest
New York-Ft Drum
Nevada-Winnemucca
Wyoming-Southwest
SOUTH DAKOTA-Yankton
Wisconsin- Ft McCoy
Ashley National Forest
NEBRASKA-Scottsbluff
Michigan-Bay City
Nevada-Stillwater
NEBRASKA-NorthEast
alifornia-Oakdale
Wyoming-Laramie
Illinois-Savanna
Michigan-Berrien County
Detroit
New York-Buffalo
New York-Albany
eterans Affairs Medical Center: Berryman William R MD
NEBRASKA-South Central
Illinois-Marseilles
Illinois-Sheridan
OHIO-Camp Perry
NEW YORK-Otisville
PENNSYLVANIA-Allenwood
Illinois-Galesburg
Ohio- Lima
Utah-Cedar City
PENNSYLVANIA-Camp Hill
Illinois-Menard
Illinois-Lincoln
Indiana-Ft Benjamin Harrison
Nevada-Las Vegas AREA
Colorado-Ft. Carson
Colorado-Granada
Missouri-Richards-Gebaur AFB
Illinois-Greenfield
Maryland-Ft. Detrick
Indiana-Camp Atterbury
WEST VIRGINIA- Mill Creek
California-Ft. Irwin
Colorado-Trinidad
Illinois-Scott AFB
ifornia-Twentynine Palms
Missouri-Warsaw
Arizona-Sedona
Missouri-Ft. Leonard Wood
WEST VIRGINIA- Beckly
Illinois-Marion
VIRGINIA -Petersburg
Arkansas-Omaha
Arkansas-Berryville
Arizona- Wickenburg
Arizona-Phoenix
Texas- Amarillo
TENNESSEE -Ft. Campbell
Arizona-Florence
OKLAHOMA- Tinker
Arizona-Pinal County
TENNESSEE -Crossville
Arkansas-Ft. Chaffee
Arkansas-Blythville
NEW MEXICO-Fort Stanton
OKLAHOMA- Lawton
North Carolina-Andrews
North Carolina-Camp Lejeune / New R
OKLAHOMA-McAlester
SOUTH CAROLINA-Greenville
NEW MEXICO-Holloman
Texas- Lubbock
Arkansas-Pine Bluff Arsenal
Arkansas-Rohwer
NEW MEXICO-Ft. Bliss
Texas- Ft Bliss
Texas -Carrolton
Arkansas-Jerome
Alabama-Talladega
Texas-Ft. Worth
Alabama-Aliceville
South Carolina-Charleston
Texas-Eden
Texas- Mexia
Louisiana-Ft. Polk
Alabama-Montgomery
Georgia-Oglethorpe
Gulf of California
Texas- Killeen
Georgia-Morgan
Louisiana-Oakdale
Louisiana-north of Interstate 12
Texas- Beaumont/Port Arthur
Florida
Florida-Avon Park
Florida-Camp Krome
Gulf of Mexico

Googles list of FEMA Camps

Complete 326 page manual

http://info.publicintelligence.net/USArmy-InternmentResettlement.pdf

Leaked Document: Military Internment Camps in U.S to be Used for Political Dissidents

http://www.youtube.com/watch?v=FfkZ1yri26s

Military ad as it appeared in Government Pub:

Correction Officer Internment / Resettlement Specialist

As an Internment/Resettlement Specialist for the Army National Guard, you will ensure the smooth running of military confinement/ correctional facility or detention/internment facility, similar to those duties conducted by civilian Corrections Officers. This will require you to know proper procedures and military law; and have the ability to think quickly in high-stress situations. Specific duties may include assisting with supervision and management operations; providing facility security; providing custody, control, supervision, and escort; and counseling individual prisoners in rehabilitative programs.

So, why is the National Guard hiring “Internment / Resettlement Specialists?”

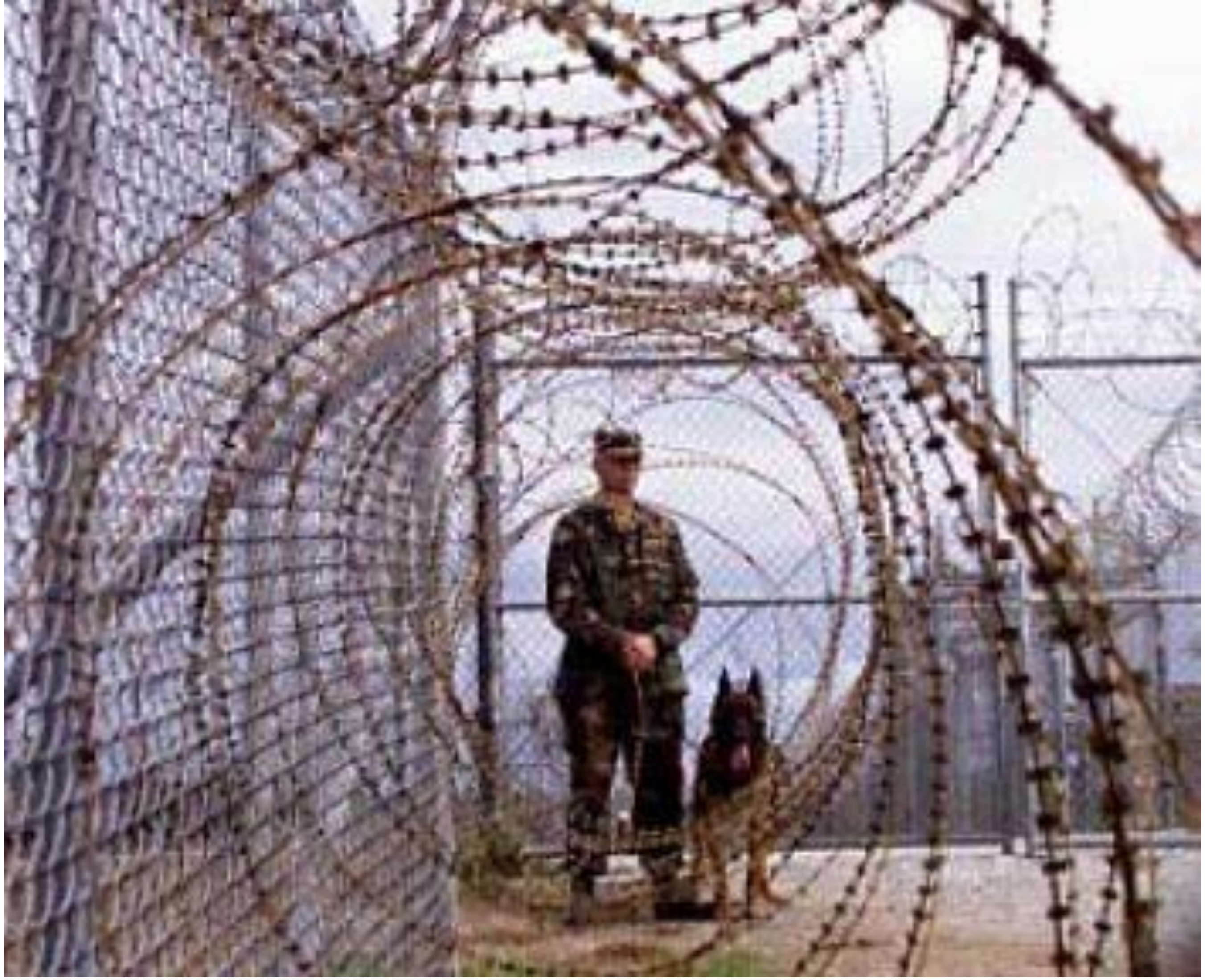

http://intellihub.com/2013/06/13/find-the-fema-camp-nearest-you/

The White House quietly issued National Security Presidential Directive 51 (NSPD-51), to ensure "continuity of government" in the event of what the document vaguely calls a "catastrophic emergency." Should the president determine that such an emergency has occurred, he and he alone is empowered to do whatever he deems necessary to ensure "continuity of government." This could include everything from canceling elections to suspending the Constitution to launching a nuclear attack.

REX 84

The Rex 84 Program was established on the reasoning that if a mass exodus of illegal aliens crossed the Mexican/US border, they would be quickly rounded up and detained in detention centers by FEMA. Rex 84 allowed many military bases to be closed down and to be turned into prisons. With half a million entering the US every year, not one of them has been put in a FEMA camp yet! What are they waiting for?

NATIONAL DEFENSE AUTHORIZATION ACT

Sect. 1042 of the 2007 National Defense Authorization Act (NDAA), "Use of the Armed Forces in Major Public Emergencies," gives the executive the power to invoke martial law. For the first time in more than a century, the president is now authorized to use the military in response to "a natural disaster, a disease outbreak, a terrorist attack or any other condition in which the President determines that domestic violence has occurred to the extent that state officials cannot maintain public order."

If the US government spends billions of dollars on Homeland Security, they scan all of our emails, record all of our telephone calls, and watch almost everything we do; because they expect a Muslim terrorist to blow up our country any minute, then why did they allow 9 million Muslims to come into our country?

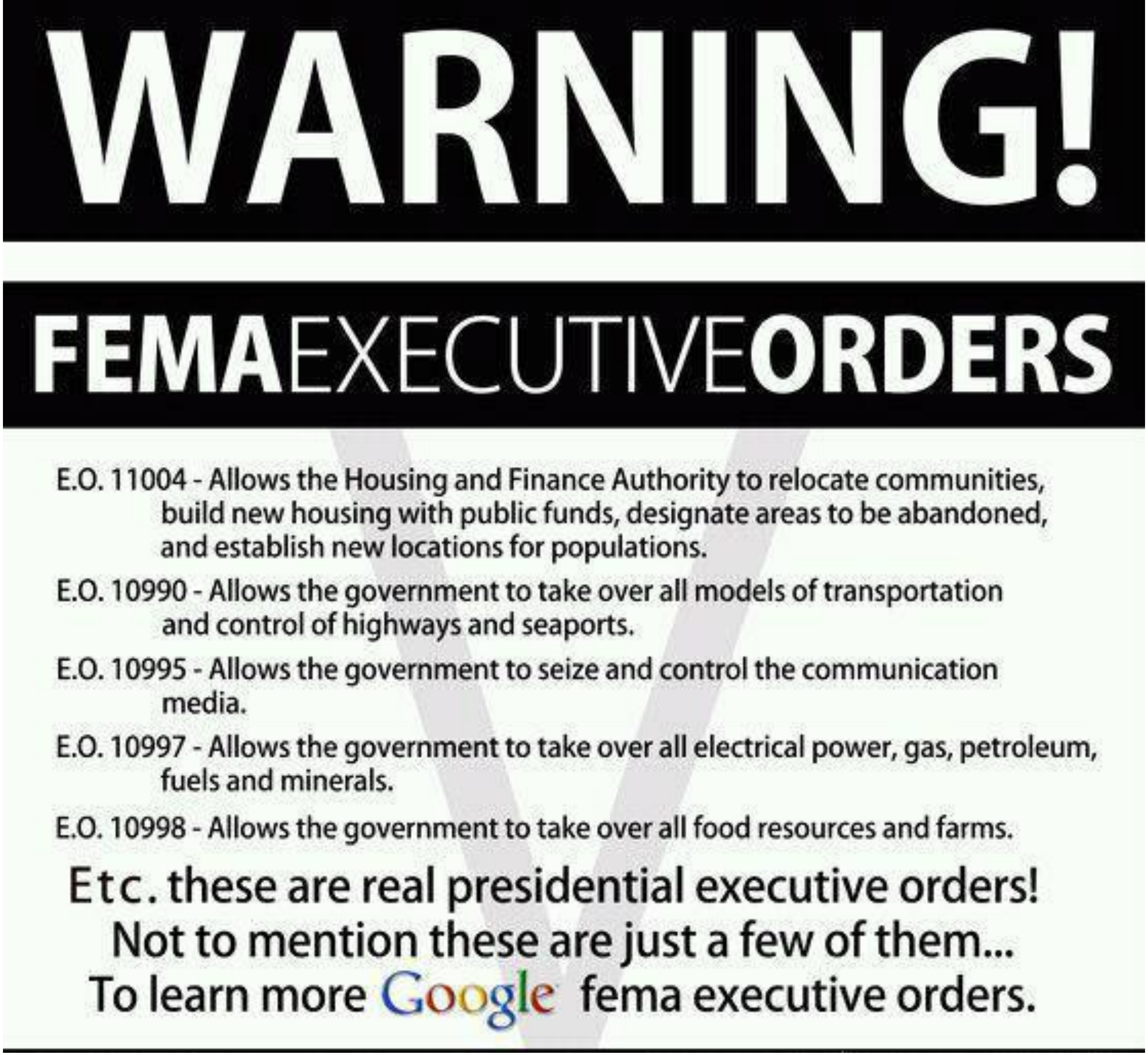

WARNING!

FEMA EXECUTIVE ORDERS

E.O. 11004 - Allows the Housing and Finance Authority to relocate communities, build new housing with public funds, designate areas to be abandoned, and establish new locations for populations.

E.O. 10990 - Allows the government to take over all models of transportation and control of highways and seaports.

E.O. 10995 - Allows the government to seize and control the communication media.

E.O. 10997 - Allows the government to take over all electrical power, gas, petroleum, fuels and minerals.

E.O. 10998 - Allows the government to take over all food resources and farms.

Etc. these are real presidential executive orders!
Not to mention these are just a few of them...
To learn more Google fema executive orders.

ARMY TRAINING MANUAL

The manual states, “These operations (camp) may be performed as domestic civil support operations,” and adds that “The authority to approve resettlement such operations within U.S. territories,” would require a “special exception” to “The Posse Comitatus Act,” which can be obtained via “the President invoking his executive authority.” The document also makes reference to identifying detainees using their “social security number.”

Mock Military Invasions

UNITED STATES ARMY

http://www.independentsentinel.com/2013/03/military-drills-armored-tanks-zombies-are-coming-to-a-city-near-you/

DRONES

- The U.S. has over 8,000 drones
- They are used by police for surveillance on Americans
- They are used along the borders
- Some states have passed laws against flying drones

CHRISTIAN PERSECUTION

Media Silent as Muslims Ethnically Cleanse 60,000 Christians in Philippines

http://frontpagemag.com/2013/dgreenfield/media-silent-as-muslims-ethnically-cleanse-60000-christians-in-philippines/

PERSECUTION AT UNPRECEDENTED LEVELS

The Body of Christ today is suffering a degree of persecution not seen previously in our lifetime. In Central African Republic a serious humanitarian crisis is unfolding as the predominantly Christian population suffers terror at the hands of Islamic rebels. Afghan converts risk death not only in Afghanistan but also in India because the Afghan media is inciting violence against them, particularly Pastor Obaid. In Somalia, al-Shabaab routinely kills Somali converts.

Mustaf (13), who has already lost two Christian relatives to al-Shabaab, was recently kidnapped by militants seeking to draw his Christian parents out of hiding. For example on 30 August eleven Laos families (50 people) were expelled from their village and sent out into the bush for refusing to renounce Christ. On 1 September nine Catholics were massacred by Fulani Muslim gunmen in Nigeria's Kaduna State. On 4 September jihadists killed eight Syrian soldiers before seizing control of the Aramaic-speaking, Christian village of Maaloula in western Syria.

"The power of the Executive to cast a man into prison without formulating any charge known to the law, and particularly to deny him the judgment of his peers, is in the highest degree odious and is the foundation of all totalitarian government whether Nazi or Communist."

- Winston Churchill, Nov. 21, 1943

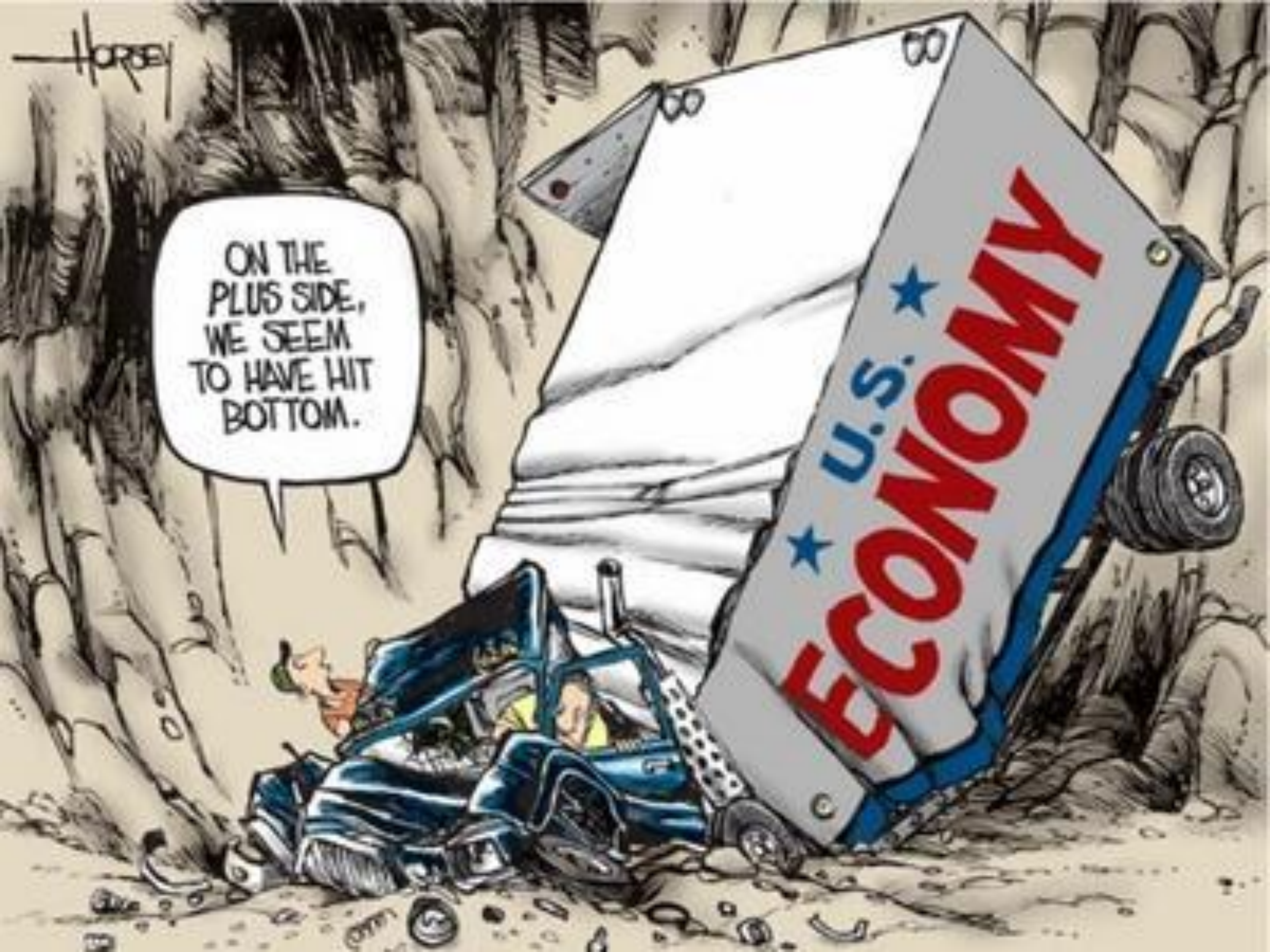
HORSEY
ON THE PLUS SIDE, WE SEEM TO HAVE HIT BOTTOM.
U.S. ECONOMY

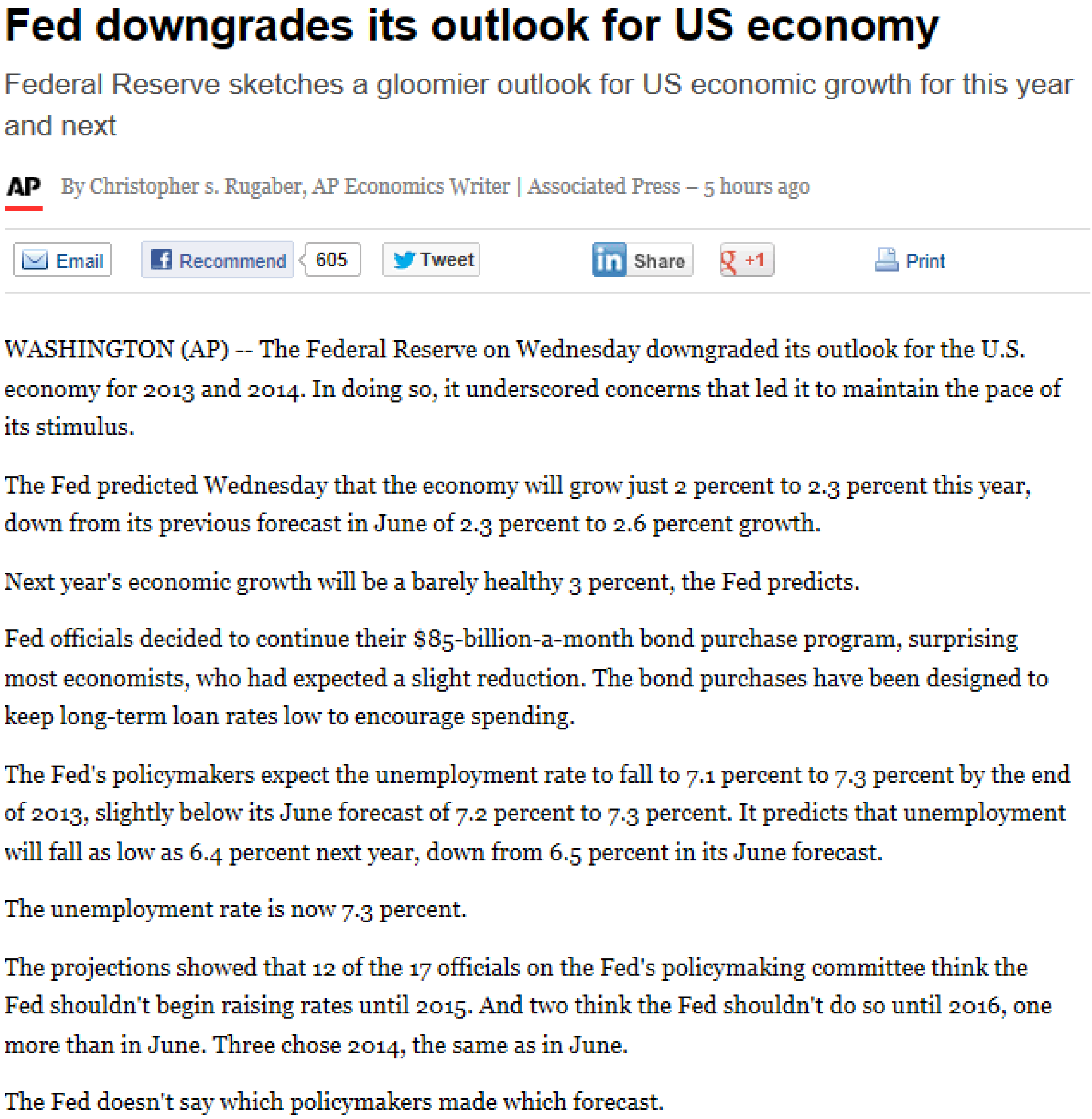

Fed downgrades its outlook for US economy

Federal Reserve sketches a gloomier outlook for US economic growth for this year and next

AP By Christopher s. Rugaber, AP Economics Writer | Associated Press – 5 hours ago

WASHINGTON (AP) -- The Federal Reserve on Wednesday downgraded its outlook for the U.S. economy for 2013 and 2014. In doing so, it underscored concerns that led it to maintain the pace of its stimulus.

The Fed predicted Wednesday that the economy will grow just 2 percent to 2.3 percent this year, down from its previous forecast in June of 2.3 percent to 2.6 percent growth.

Next year's economic growth will be a barely healthy 3 percent, the Fed predicts.

Fed officials decided to continue their $85-billion-a-month bond purchase program, surprising most economists, who had expected a slight reduction. The bond purchases have been designed to keep long-term loan rates low to encourage spending.

The Fed's policymakers expect the unemployment rate to fall to 7.1 percent to 7.3 percent by the end of 2013, slightly below its June forecast of 7.2 percent to 7.3 percent. It predicts that unemployment will fall as low as 6.4 percent next year, down from 6.5 percent in its June forecast.

The unemployment rate is now 7.3 percent.

The projections showed that 12 of the 17 officials on the Fed's policymaking committee think the Fed shouldn't begin raising rates until 2015. And two think the Fed shouldn't do so until 2016, one more than in June. Three chose 2014, the same as in June.

The Fed doesn't say which policymakers made which forecast.

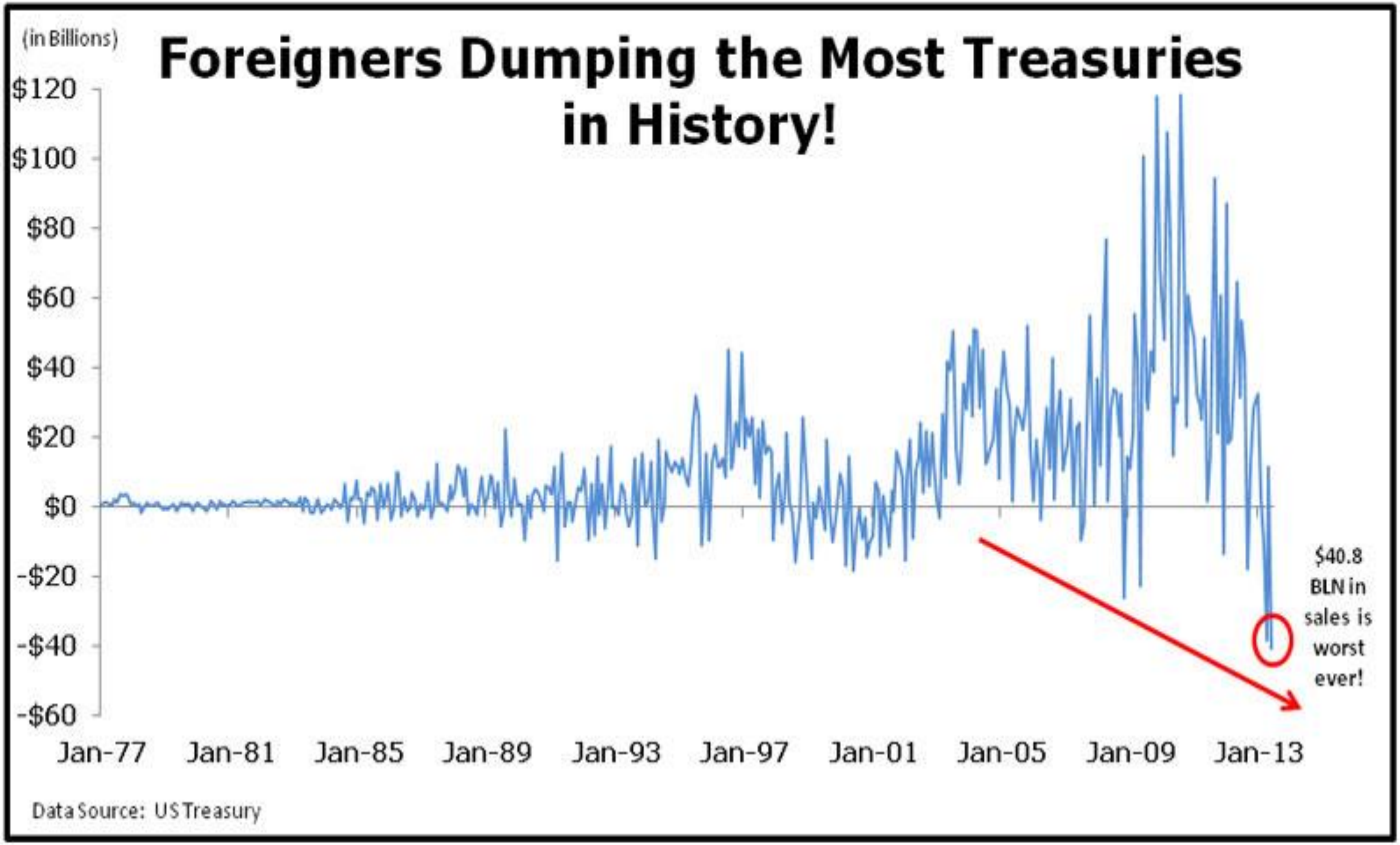

Foreigners Dumping the Most Treasuries in History!

$40.8 BLN in sales is worst ever!

Data Source: US Treasury

FINANCES IN THESE DIFFICULT TIMES

We Are Now On The Verge Of A Historic Meltdown & Collapse

On the heels of a tremendous rally in gold and silver, today a man who has been involved in the financial markets for 50 years shocked King World News when he said that we are on the verge of a historic and catastrophic global financial "meltdown." He also spoke about events that are unfolding behind the scenes at the White House right now. This is without question one of John Embry's most powerful interviews ever.

Embry: ***"I guess I'm always unnerved as a result of what happened in April, the last time the President of the United States had a meeting with all of the bank heads, and two days later the price of gold was smashed for over $200. Now, the President is meeting with all of the heads of the various agencies, institutions, the Fed, and all of the other key money entities in the United States today. What's that all about?***

But clearly if the President is having this meeting, there is a crisis unfolding somewhere in the background, and it could very well relate to the dollar, interest rates, and the massive derivatives market associated with interest rates....

“Rising interest rates are a killer in an over-levered economy, and that’s exactly what we’ve been seeing in the United States.

This surge in interest rates may have already seriously destabilized the entire financial system, and that’s why there is this meeting taking place in the White House today. The fact is that the vast majority of derivatives in the global financial system are related to interest rates.

Now, the entire financial system may be on the precipice of some sort of catastrophic event unfolding because of what we have already seen in the bond market, and how the derivatives are so heavily intertwined. Meaning, we may be on the verge of another disastrous derivatives meltdown.

We have an unbelievable amount of interest rate derivatives in the financial system. So the winners theoretically take the profits on them, and the losers simply misprice them on their books. But as you get higher interest rates it becomes even more destabilizing.

I firmly believe the reason the President has called this meeting today is because if interest rates in the U.S. continue to rise, it could really unleash something disastrous. We are talking here about the possibility of a meltdown.
http://kingworldnews.com/kingworldnews/KWN_DailyWeb/Entries/2013/8/20_We_Are_Now_On_The_Verge_Of_A_Historic_Meltdown_%26_Collapse.html

Financial Crisis Alert: Half A Million People Drop Off Workforce In One Month, Bank of America To Cut 2100 Jobs, Shut 16 offices, HP Has Chopped 22,700 Jobs Since Last Year, Polled Unemployment Soars To March 2012 Levels!

Bank of America
350

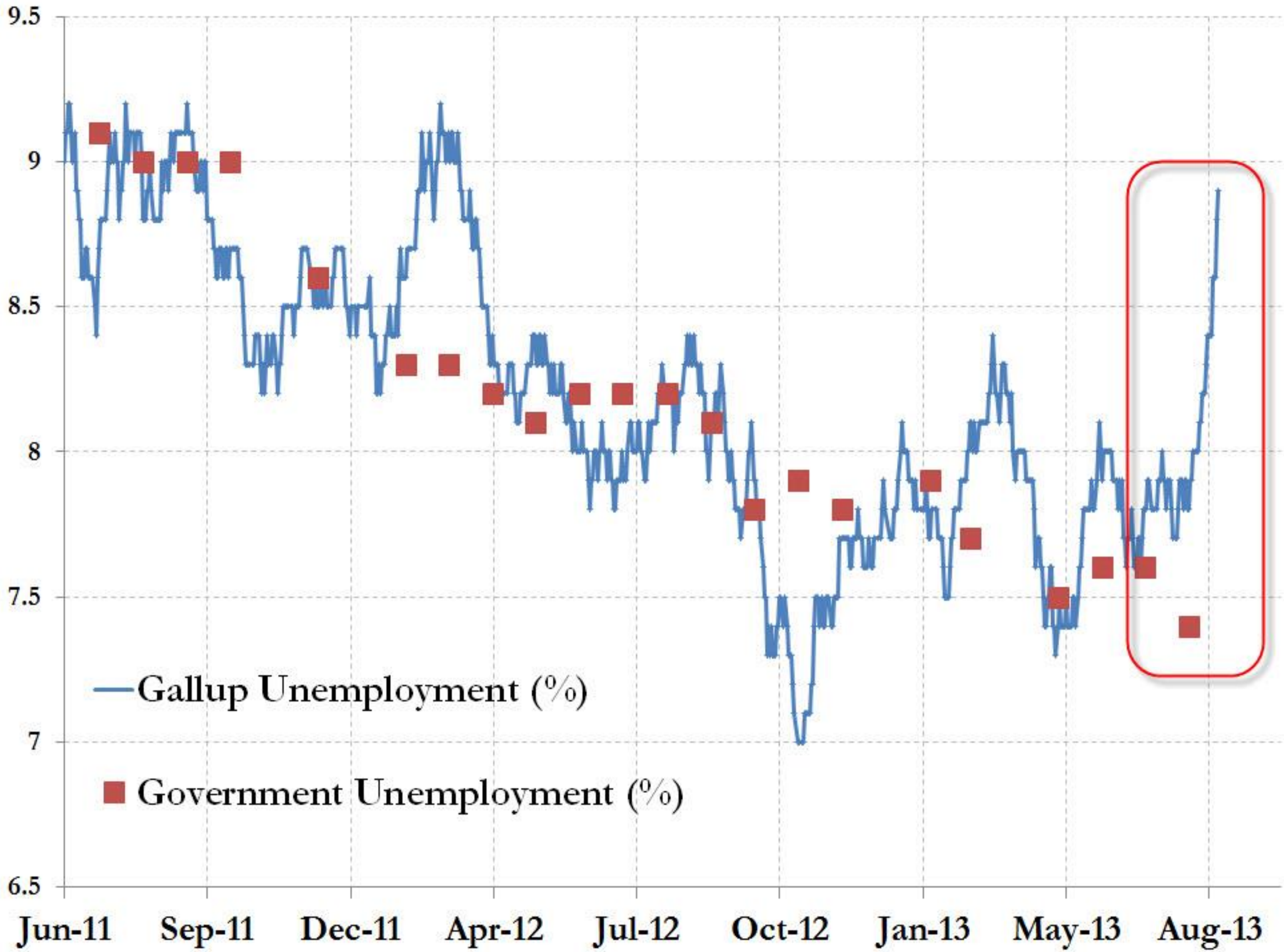
9.5
9
8.5
8
7.5
7
6.5
Jun-11
Sep-11
Dec-11
Apr-12
Jul-12
Oct-12
Jan-13
May-13
Aug-13
Gallup Unemployment (%)
Government Unemployment (%)

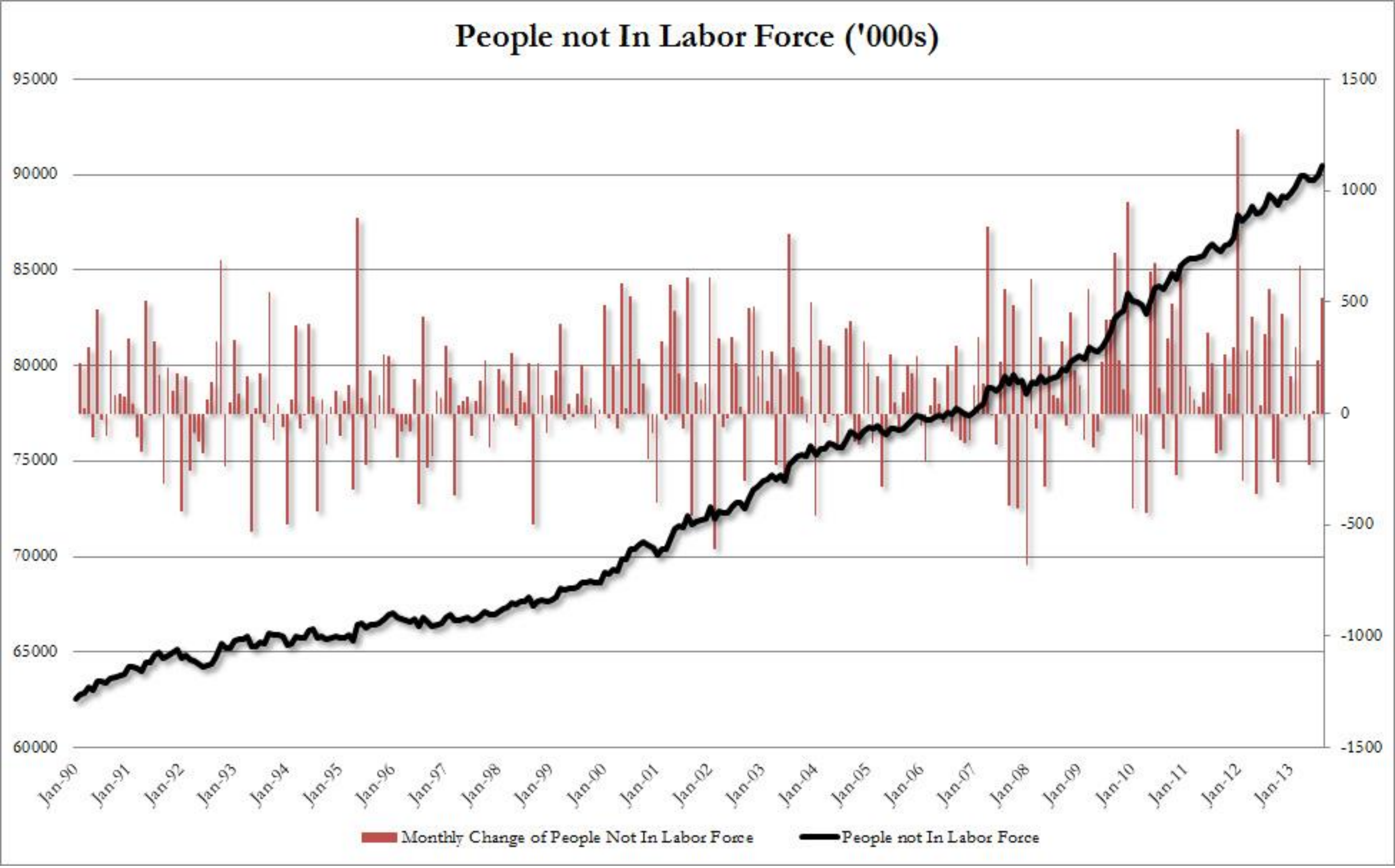

http://investmentwatchblog.com/financial-crisis-alert-half-a-million-people-drop-off-workforce-in-one-month-bank-of-america-to-cut-2100-jobs-shut-16-offices-hp-has-chopped-22700-jobs-since-last-year-jpmorgan-exits-commodity-a/

US Pension Funding Ratios by State

(Ranked in Descending Order, All 50 States)

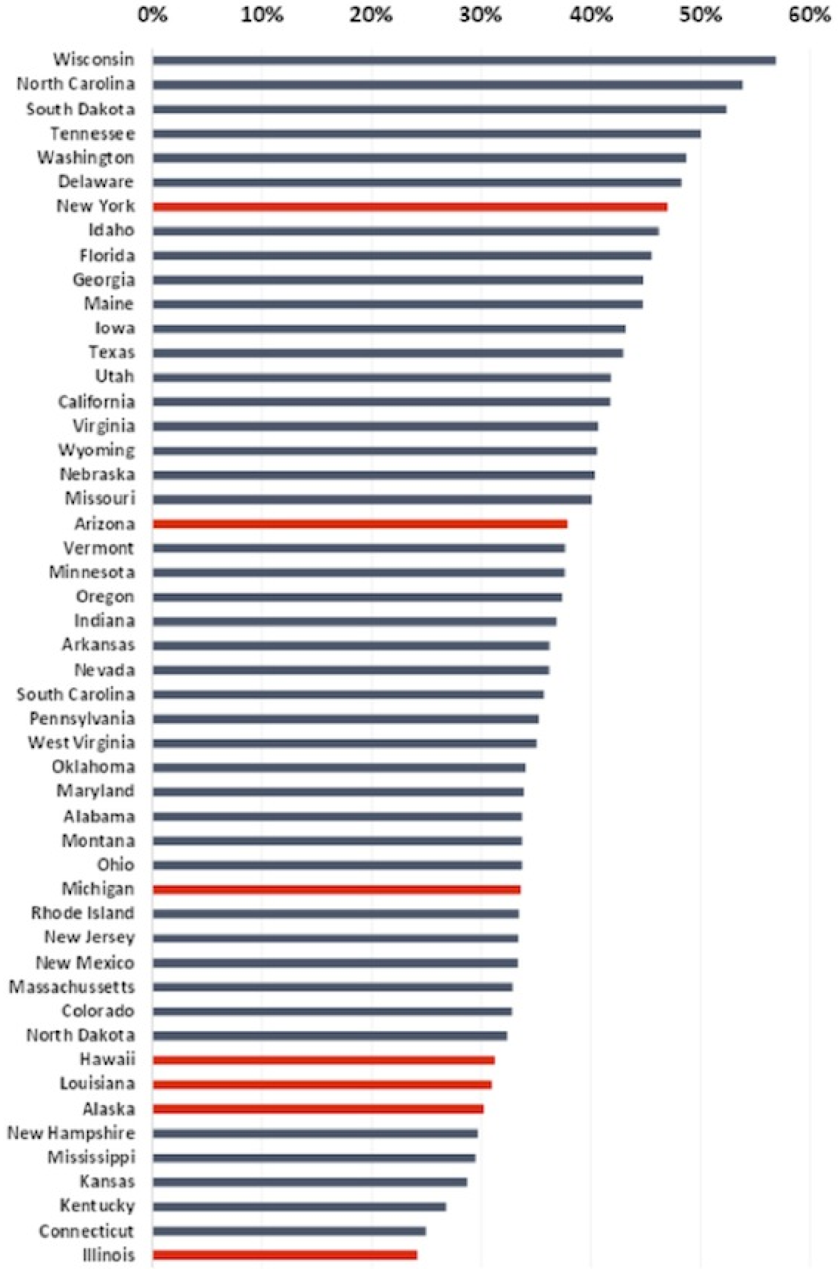

Source: Mauldin Economics; State Budget Solutions, September 2013; Manhattan Institute for Policy Research, August 2013

International Investors Dump $40.8 Billion in Treasuries, the Most Ever

Money and Markets

Foreign holders dumped a whopping $40.8 billion in long-term Treasuries, the biggest exodus from bonds in the history of the U.S. Worse, June was actually the third month of mass dumping in the past four, for a total of $79 billion. China, the biggest holder of our bonds, unloaded $21.5 billion, while Japan, the second-largest holder, dumped $20.3 billion.

It wasn't just government bonds, either. Foreigners dumped $116 million of bonds made up of packaged U.S. mortgages. They sold $5.2 billion of Fannie Mae, Freddie Mac and Ginnie Mae bonds, and $5 billion in corporate bonds. And they unloaded $26.8 billion of U.S. stocks. All told, more international capital flowed *out* of U.S. markets than at any time in history, worse even than at the depths of the 2008 credit crisis.

State's unfunded pension debt will grow to more than $200 billion under new reporting rules

	Unfunded liability as currently reported	Unfunded liability under Moody's
TRS	$43.5 billion	$120.6 billion
SERS	$20.2 billion	$46.2 billion
SURS	$17.6 billion	$39.3 billion
JRS	$1.3 billion	$2.6 billion
GARS	$0.2 billion	$0.4 billion
Combined	**$82.9 billion**	**$209.0 billion**

Source: Commission on Government Forecasting and Accountability; Illinois Policy Institute

Market Week Pullout Dow Falls 1.3% on Syria Fears; Time to Buy Tyson Foods Page M3

BARRON'S

The Dow Jones Business and Financial Weekly Vol. XCIII No. 35 barrons.com September 2, 2013 $5.00

Good Hands: Allstate Could Rise 50%
Page 16

RANDALL W. FORSYTH • 5
Obama's Hobson's choice on Syria

STREETWISE • 9
Investors pause but don't panic

SHOT IN THE ARM • 14
Onyx deal will give Amgen a boost

CLOUDY OUTLOOK • 17
Woes at Elon Musk-backed SolarCity

PLACE YOUR BETS • 20
Regional casino stock could rise 30%

INTERVIEW • 21
Cliff Asness sees risk everywhere

TAKEOVER BAIT? • M8
A flush Vodafone could attract AT&T

THE BULL'S IN CHARGE

Despite recent stock-market declines and the possibility of an autumn correction, Wall Street's top strategists see the market headed higher over the longer term, with stocks rising 20% or more in 18 months.

DOW JONES

Subscriptions/Customer Service: 800 544 0422

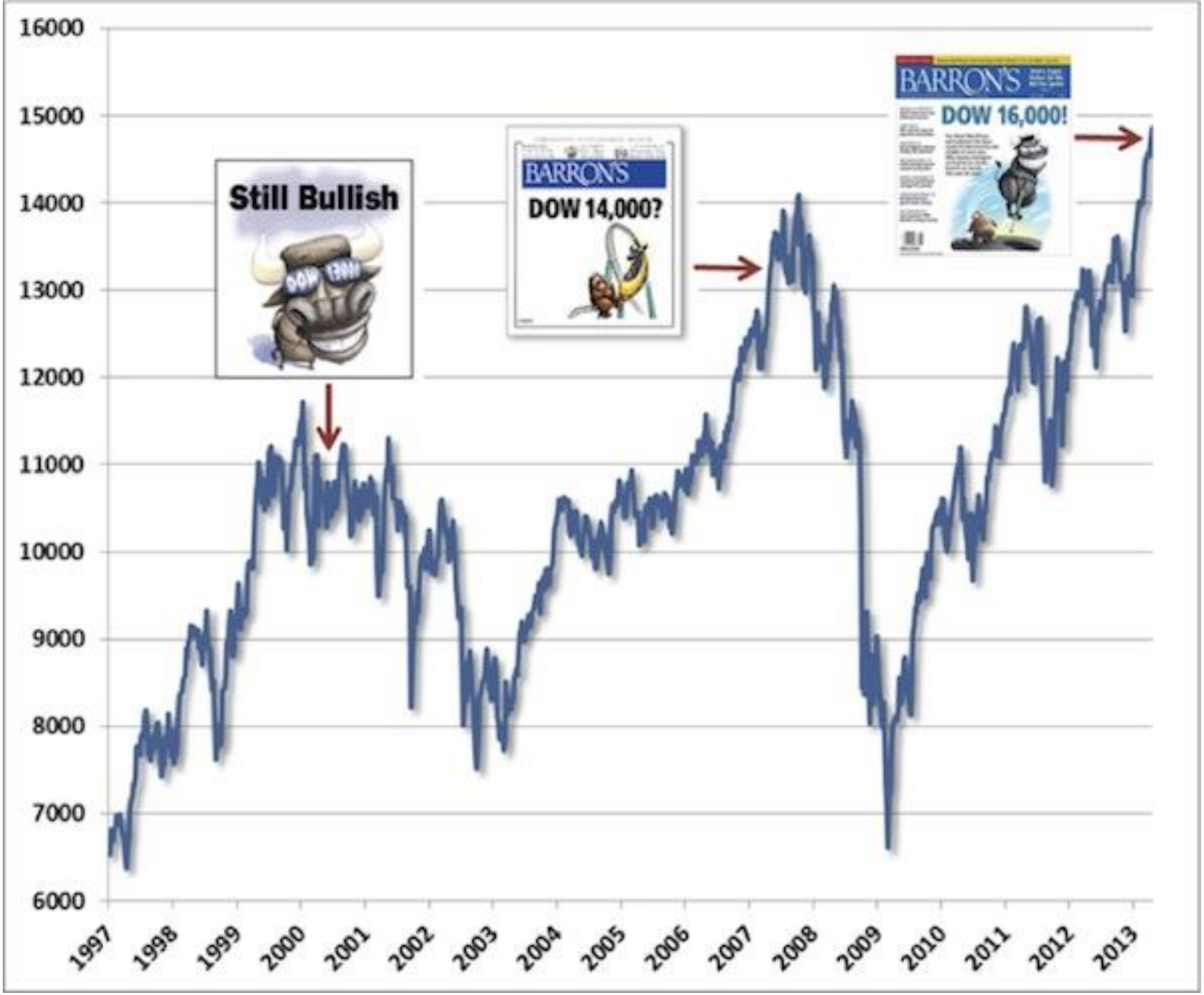
16000
15000
14000
13000
12000
11000
10000
9000
8000
7000
6000
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Still Bullish
DOW 17000
BARRON'S
DOW 14,000?
BARRON'S
DOW 16,000!

UNITED STATES DOLLAR

China and Japan a few weeks ago sold $48 billion worth of U.S. Treasury Bonds in one day more than ever in history. Bernanke and the Fed are printing $85 billion per month ($40 billion to buy Mortgage Backed Securities and $45 billion to buy Treasuries). All foreign countries are buying less US Treasury bonds since June of this year. This is very bad news.

WHAT TO DO TO PREPARE FOR THIS?

1. Get a job and keep it
2. Start a job on something that makes money
3. Store cash at your home
4. Buy gold and silver eagles and keep it at home
5. Pay off your bills
6. Store food
7. Store water

8. Store guns & ammunition

9. Store toilet paper

10. Store tooth brush, tooth paste, shampoo, soap, etc.

11. Have something to cook on such as an outside gas cooker, fireplace, or solar cooker.

12. Store cooking oil, condiments, bowls, cups, etc.

13. Have a water filter- Big Berkey

Preston T. Bailey, PhD., D.Min.

“That's all Folks!”

www.SaveAmerica Foundation.com

Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

Empowerment by Virtue of Golden Truth
Human Cloning: Specifically,
REM Driven Human Cloning
FULL DISCLOSURE

Donald Marshall

29 October 2015

https://www.facebook.com/donald.marshall.148
http://donaldmarshall.proboards.com/
Vincent Eastwood Interview
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab

Contents

List of Images 5

Introduction 6

Whistle Blowing: Qualifying Disclosure 6

The Serious Crime Act 2015 7

Fundamentals 8

Key Figures 9

- *Phil Schneider* 9
- *George Green* 11
- *Aaron Russo* 13
  - *Income Tax* 13
  - *Radio Frequency Identification (RFID) Microchips* 14
  - *Habeas Corpus* 15
  - *Fraudulent Manipulation of Election Results* 15
  - *The War on Terrorism is the War on Your Freedom* 16
  - *Whoever Makes the Money Makes the Rules* 16
  - *Russo's Message to Mankind* 16
  - *Let's Consolidate our Problems* 17
- *Aaron Russo: Reflections and Warnings* 17
  - *The Falsehood of September 11th 2001* 17
  - *The New World Order Agenda* 19
  - *Women's Liberation from the Perspective of the Elite* 20
  - *America is a Republic* 20
  - *Depopulation* 21
  - *Borrowing Money from Private Banks causes Inflation and Debt* 21
  - *The Deception of the Council of Foreign Relations (CFR)* 23
  - *Combining America, Mexico and Canada into one Country* 24
  - *Russo's Vision on How to Bring These People Down* 25

The Trans-humanist / Post Human Agenda 26

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep 27

REM Sleep 28

Current Advances in Technology 28

- *Memory Suppression Technologies* 29
- *Mind-Voice Technology* 29
- *H.A.A.R.P. Technology* 29
- *H.A.A.R.P. and Weather Control* 30
- *H.A.A.R.P. and Mind Control* 31
- *RFID Technology* 32

  RFID Technology and the Medical Profession 32
  Further Implications of RFID Technology 33
  Mind Control: MK Ultra Technology 33
  Human Cloning 35
    Cloning Centre and Cloning Technology 36
Drip Feed Disclosure and Evaluative Conditioning 36
  Drip Feed Disclosure 36
  Evaluative Conditioning 37
Consciousness Transfer 38
Disclosure: Save the Victims through your Diligence 40
  Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell 41
  The Science of Hidden Consciousness 41
Disclosure: Donald Marshall's Message to the World 42
  The Two Types of Cloning Techniques 42
  The Different Grades of Clones: "Mark 1" to "Mark 4" Clones 43
  Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones 43
  Consciousness Transfer Happens When the Original Reaches REM Sleep 43
  Donald Marshall is an Original as he sits at Home Typing this Disclosure 44
  When Donald Marshall's REM Driven Cloned Experiences Began 45
  The Science of REM Driven Cloning Began in 1945 46
  Help Donald Marshall STOP REM Driven Cloning NOW! 47
  Do Not Panic, Riot, or Cause Chaos. This is Very Important. 47
  Spread this Disclosure Document until it reaches the Armed Forces 47
  Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour' 48
  How the Nostradamus Prophecies shapes the lives of the Illuminati 49
  HIV/AIDS has been cured; so have many forms of cancer 49
  The Behaviour of High Profile People at the Cloning Centre 50
  The Ring Leaders of the Cloning Centre 50
  The Royal Family of England 53
  Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family 57
  The Deplorable Nature of the Illuminati 59
  What Happened to Britney Spears? 60
  NEVER EVER SELL YOUR SOUL 62
    In This Physical World Selling Your Soul Is Selling Your REM Driven Clone 62
  Fake Alien Abductions conducted through REM Driven Cloning Technology 64
  I am THE Spy for Humanity against the Corruption of the Illuminati 65
  The Threat to Humanity when the H.A.A.R.P Grid is Complete 66
  REM Driven Cloning is the Most Terrible Nightmare Situation Ever 68

  - *The Meek Shall Inherit the Earth* 70
  - *Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents* 72
- *Contact Information* 73
- *Interviews* 74
  - *Links to Donald Marshall Interviews* 76
    - *Vincent Eastwood* 76
    - *Jeanice Barcelo* 76
    - *Lisa Phillips (CFR)* 76
- *Radio Presenters –Contact Donald Marshall* 77
- *Professionals who understand “Consciousness Transfer” –Contact Donald Marshall* 77
- *Email* 77
- *How to Learn More about Donald Marshall’s Whistle Blowing Disclosures* 77
- *How Much Do You Know About Post-humanism / Trans-humanism?* 78
- *This IS your ONLY chance to end these monsters. There is a deadline!* 79
- *Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1* 80
- *Noteworthy Frequently Asked Question* 81
- *Donald Marshall’s Mission and Vision on How to Bring Down The Illuminati* 81
- *Empowerment by Virtue of Golden Truth* 83
- *Appendices* 84
  - *Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript* 84
  - *Appendix B: MK Ultra* 87
  - *Appendix C: The Illuminati* 89
- *References* 99
- *Legislation* 111

List of Images

Image 1: Stages of sleep.. 27
Image 2: The ionosphere's relationship to earth.. 30
Image 3: Megadeth Album Cover: *The World Needs a Hero* (2001)... 58
Image 4: Donald Marshall’s original body compared with his REM Driven Duplicate Clone........... 58
Image 5: Britney Spears Japanimation: Britney Spears grown as a duplicate clone......................... 61
Image 6: Britney Spears Japanimation: Britney Spears walks into the Cloning Centre..................... 61
Image 7: Britney Spears Japanimation: The Cloning Centre depicted from an aerial view................ 61
Image 8: Possible H.A.A.R.P. Locations. 67

Introduction

Just read it. Every word in this document is true, and it is the most important document to ever reach the internet. This document is written to empower good people of the world against the tyranny which exists all around us in our world today. Effectively, the document presented here discloses concealed information about the extent to which medicine, science, and technology has advanced in our **current** era. Moreover, I am disclosing this information because I have been extensively abused, and I am victim of **current** advances in medicine, science, and technology, which I describe in detail below. There are many, many, other victims of abuse, including myself, and therefore it is our plea that you read this document with an open mind and you investigate **ALL** statements which may initially irk you and appear 'suspect' to begin with (the circumstantial evidence is available all around you); because as a victim of abuse, I can honestly say, there is nothing worse than experiencing abuse, and those who have the power to protect you from your abusers, ignore you.

I will now discuss what qualifies as whistle blowing disclosure. After detailing the criteria which qualifies for whistle blowing disclosure, I will detail aspects of the Serious Crime Act, and preface a foundation which defines, explains, and references key information for the reader to understand as well as explore the extent to which medicine, science, and technology has advanced today; all for the express purpose of compelling the reader that medicine, science, and technology, has advanced at an incredible rate; that **current** advances in medicine, science and technology are withheld from public knowledge; and finally these advancements in medicine, science, and technology, have been and are **currently** used to commit **serious unspeakable** crimes against humanity and our environment here on earth. Moreover, it is my aim to present this information as coherently as possible to assist the reader to save me, and save many, many others who experience abuse daily resulting from concealed advances in technology.

Whistle Blowing: Qualifying Disclosure

In the United Kingdom under the Public Interest Disclosure Act (1998) (PIDA) section 43B, a "qualifying disclosure" (whistle blowing disclosure) implies "any disclosure of information which in the reasonable belief of the worker, tends to show one of the following:

a) That a criminal offence has been committed, is being committed or is likely to be committed,
b) That a person has failed, is failing or is likely to fail to comply with any legal obligation to which he is subject
c) That a miscarriage of justice has occurred, is occurring or is likely to occur,
d) That the health or safety of any individual has been, is being or is likely to be endangered,
e) That the environment has been, is being, or is likely to be damaged, or
f) That information tending to show any matter falling within any one of the preceding paragraphs has been, or is likely to be deliberately concealed"

Furthermore, under the Enterprise and Regulatory Reform Act (ERRA) 2013 section 17 (*The new 'public interest' test)* is added for whistle blowing disclosures, amending section 43B of the Employment Rights Act (ERA) 1996 and therefore now reads as such (new words underlined):

"(1) In this part a 'qualifying disclosure' means any disclosure of information which, in the reasonable belief of the worker making the disclosure, *is made in the public interest and* tends to show one or more of the following-- [criminal offence, breach of legal obligation, etc]." (Halliday 2013, p. 2).

The following information is presented **with the express intent** that it is made in the public interest, and meets the all criteria set out in section 43B (a-f) of PIDA as well as the 'public interest test', ERRA section 17. Furthermore, in order for the reader to perceive the credence of my statements, so that (s)he has impetus to research the validity of my statements and not dismiss it as folly supplied by an internet troll who has nothing better to do than waste peoples' time, I will cite the Serious Crime Act 2015 here for the express purpose of convincing the reader that all statements made in this document and my original document are indeed FACTUAL! THIS IS NO HOAX. THIS IS NO JOKE. This is simply what is happening in the world today.

The Serious Crime Act 2015

The Serious Crime Act 2015 (UK) section 41, 3ZA, amends section 3A of the Computer Misuse Act 1990 which **previously** read [under Computer Misuse Offences]:

[[F1] 3A Making, supplying or obtaining articles for use in offence under section 1 or 3

1) [That] A person is guilty of an offence if he makes, adapts, supplies or offers to supply any article intending it to be used to commit, or to assist in the commission of, an offence under section 1 or 3.

To: [The Serious Crime Act 2015, section 41, 3ZA]

3ZA Unauthorised acts causing, or creating risk of, serious damage

(1) A person is guilty of an offence if—

  (a) the person does any unauthorised act in relation to a computer;

(2) Damage is of a "material kind" for the purposes of this section if it is—

  (a) damage to human welfare in any place;

  (b) damage to the environment of any place;

  (c) damage to the economy of any country; or

  (d) damage to the national security of any country.

3) For the purposes of subsection (2)(a) an act causes damage to human welfare only if it causes—

(a) loss to human life;

(b) human illness or injury;

(4) It is immaterial for the purposes of subsection (2) whether or not an act causing damage—

(a) does so directly;

(b) is the only or main cause of the damage.

(5) In this section—

(c) a reference to a country includes a reference to a territory, and to any place in, or part or region of, a country or territory.

(6) A person guilty of an offence under this section is (unless subsection (7) applies) liable, on conviction on indictment, to imprisonment for a term not exceeding 14 years, or to a fine, or to both.

(7) Where an offence under this section is committed as a result of an act causing or creating a significant risk of—

(a) serious damage to human welfare of the kind mentioned in subsection (3)(a) or (3)(b), or

(b) serious damage to national security,

a person guilty of the offence is liable, on conviction on indictment, to imprisonment for life, or to a fine, or to both."

In laymen's terms it is a **serious crime** (Serious Crime Act 2015), punishable by imprisonment, a fine, or both, to cite hatred or spread hoaxes by use of a computer which would cause mass hysteria or public dissension. I have been condemning public figures, high profile individuals, celebrities, singers, movie stars, directors, politicians, world leaders, royalty, anti-secret society 'truthers' (who are really misinformation agents, and agents of deception) etc. –the list is long - since 2011 and **NOT** a single one of these high profile people has issued a public statement saying "I do not like how you have tarnished my name" or filed a law suit against me. Why you may ask? **Because** every word I have stated and will state again (in this document) is true.

Fundamentals

Before I describe the heinous crimes committed against me and many others (including unsuspecting civilians of the world) it is important that I explain, define, and reference:

- Key figures such as Phil Schneider, George Green and Aaron Russo, so that the reader has reference points to verify the extent to which technology has advanced in our **present** era (and continues to advance);
- Define and briefly explain the transhumanist / post human agenda (for those who may be unfamiliar with their aims).
- Define and explain rapid eye movement (R.E.M) sleep, the phases of sleep, and what happens to the (original's) body during sleep.

- Reference advances in technology, with particular attention to: Memory suppression technologies; Mind-voice technology; H.A.A.R.P technology, RFID microchip technology, and MK Ultra technology (CIA's mind control program);
- Define and explain cloning; the different types of clones; cloning centres and cloning technology;
- Define drip feed disclosure / evaluative conditioning and explain why it occurs;
- Explain what 'consciousness transfer' is to the best of my knowledge, and cite drip feed disclosure articles with the express intent to allow the reader to grasp the abuse I have suffered daily at the hands of my tormentors.

Key Figures

Phil Schneider

Phil Schneider (pictured) had 17 years experience working in government black projects carrying a level three security clearance. He was a geologist and engineer who worked in the black projects underground bases at Area 51, S-4, and Los Alamos.

He is most notable for disclosing (Schneider 1995; 1996; Open Minds 2011):

- The 'black budget' expenditure of the United States, which Schneider claims to be between 1.023 trillion U.S. dollars every 2 years (over $500 billion per year);
- Deep Underground Military Bases (also known as D.U.M.Bs -"dumbs"), and at the time of his lecture (Schneider, 1995), –that there are 131 **active** Deep Underground Military Bases present in the United States, and 1477 Deep Underground Military Bases worldwide;
- Each D.U.M.B costs on average 17-19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build D.U.M.Bs with sophisticated methods.
- That military technology outstrips the general public's technology at a rate of 44 to 45 years of technology for every calendar year which passes. In other words for every 12 months which passes military technology will have advanced by 44 years than what we as the general public is currently accustomed to. Therefore as a rough example if we were to take the year Facebook was founded (2004) as a base year, then the military technology since the creation of Facebook will have outstripped what the

general public is accustomed to by as much as **484 years(!)** (2015 - 2004 = 11. 11 x 44 = 484). Similarly, if we are to use the year which YouTube was founded as a base year (2005), once again, military technology would be 440 years more advanced than what the public is currently accustomed to today.

For the purpose of disclosing current advances in medicine, science and technology, and how such advances are used against me and the people of the world to commit monstrous crimes I will use 1945 as a base year. Everything will become apparent including why I use 1945 as a base year for my disclosure, but for the express purpose of compelling the reader to investigate my disclosure I must present everything, logically, sequentially, methodologically, and provide references for the reader in a coherent way to enable him or her to **pay serious attention** to my eye witness accounts. Therefore, at this present stage keep in mind the year 1945, and the fact that military technology outstrips the general public's technology at a rate of 44 years for every 12 months which passes.

George Green

George Green (pictured) was affiliated with U.S. Presidential candidates, and was once asked to be the Finance Chairman for the next President of the United States. Green would later declined the offer of Finance Chairman when a comment made by Ted Kennedy regarding sleeping with Green's 14 year old daughter caused George Green to re-evaluate his associations with this in-group.

During his 2008 interview with Project Camelot (2008a; 2008b), Green disclosed the following noteworthy information:

- U.S. presidents are "selected" and not elected. This is clearly expressed in a story Green recounts, when Green asks:
  "Who is going to be the next President of the United States [for him to overlook their finances]?"
  And the reply was: "Jimmy Carter"
  Green responds: "Jimmy who?"
  The reply Green received was: "Well, he's the Democratic Governor of Georgia."
  George Green: "But I've been voting Republican."
  Green was then confronted by a tall man, Paul Volcker (American Economist, and Chairman of Federal Reserve under Jimmy Carter and Ronald Regan), who walked over and said, "Son, don't worry about it [Republican or Democrat], we control 'em both."
- U.S. scientists learned how to make people (clones) since 1938 -walking talking ones –and the scientists call these people "synthetics" or "the others".
- Cloning technology is relatively advanced. All that has to be done is take two cells from the original, give the cells a small electrical charge (retain a fertilised egg), then all one needs is a receiver (a womb / artificial womb for the fertilised egg to grow).
- Scientist were excited by the synthetic technology because it meant that spare organ parts could be grown for an "original" human without rejection, because theoretically speaking, the DNA of the synthetic is the same as the original.
- Green gives an example of a cloned Politician: George Walker Bush. Green advices the viewer to seek old video recordings of George W. Bush, and compare the old George W. Bush, with the George W. Bush during Bush's second term,

-in terms of mannerism, speech pattern, body language etc. (old videos of George Walker Bush (Boringest 2006; Fox 4 News - Dallas-Fort Worth 2014) are presented in the References section).

- Scientist have also learned how to make these synthetic people within a few months, to the point where the synthetic can be a walking, talking duplicate of the original, intact with all the memories and experiences of the original. The only problem is that the memory, experiences, and functioning capabilities of these synthetics are like: "A DVD recorder. Sometimes you have glitches, and you have to take the synthetics to Camp David occasionally [every 6 months to a year] to get them tuned up."
- Remember these synthetics are people too, they can think and act just like you can, but they do not have a soul.
- Most of the world leaders have been bought and paid for [by men in the background] and are created to think a certain way –to meet the agendas of these men who remain in the background.
- The global elite plans on depopulating the current human population of **over 7 billion** to **500 million** people. This is corroborated by the "Georgia Guidestones" (WorldTruth 2014).
- Vladimir Putin is executing plans to bring the United States down and into a massive depression. Furthermore, China agrees with Vladimir Putin's plans; stating that the Chinese government has not been a Superpower for over 5000 years and "It is their turn to run the world."
- China has the capacity to set back / shut down all of U.S. computing and electrical systems within two days.

At this present stage, the most **salient** points to keep in mind from Green's accounts are that: U.S. presidents are selected and not elected; and **most importantly**, that scientists have been capable of creating synthetics (cloning people) since 1938. If you have kept in mind Schneider's statement (1995; 1996), (that for every 12 months which passes military technology increases by the equivalent of 44 years compared to what the general public is accustomed to) –then learning that scientists have been capable of cloning humans since 1938 should not come as much surprise. Again, everything regarding my disclosure will become apparent, and I thank these men for disclosing their information, because it helps the reader to corroborate my disclosure.

Aaron Russo

Aaron Russo (pictured) was an American businessman, director, and political activist. He is best known for directing blockbuster films such as *Trading Places*, *The Rose* and *Wise Guys*. Russo believed that a human being "Should stand for something, and do the right thing when the time calls to act". Consequently, during the latter part of his career he did his best to warn the American public by producing documentaries such as *Mad as Hell* (1996), *Freedom to Fascism* (Russo 2006), and *Reflections and Warnings* (Jones 2008).

Russo disclosed the following key information in *Freedom to Fascism* (Russo 2006):

Income Tax

- There is no law requiring Americans to pay income tax. **Although this is a truth, I DO NOT advocate Americans NOT to pay income tax (and neither did Russo).** This is simply because the Federal Reserve can imprison you and seize your possessions for not paying income tax (although there is no law requiring Americans to do so). So for the sake of avoiding hassle for the mean time –pay your income taxes.
- To further illustrate the above point, Bob Shultz speaking at "We the People Foundation" said the following: "Most people believe that the income tax system is legal and that the revenue from the tax is used in the public interest. However, there is a substantial conclusive body of evidence that proves that our income tax system represents the most pernicious form of tyranny. It is the greatest hoax ever perpetrated by government against the working men and women of America."
- Charlie Beall: "The federal government itself refuses to provide, the American people, who are coercively being subjected to this extraction of their private property, without any underlying legal justification. There is no law. There is no law that requires the average American worker in the private sector to pay a direct un-apportioned tax on their labour and compensation for services. There is no law."

- State Representative – Phil Hart –(R-Idaho): "You can look through the statutes, and look for the law that requires you to pay, and when you do that, you **cannot** identify a law that requires the average person in America who earns a wage and works in private business to pay an income tax."
- Peter Gibbons (Tax Attorney) –It's actually very simple. Congress tried to enact an income tax in 1894 –The Supreme Court said that is unconstitutional. When The Supreme Court says something is unconstitutional, it's unconstitutional. They (Congress) tried again in 1913 –and The Supreme Court said –the 16th amendment- "No new power of taxation" –so if they (Congress) didn't have it then (1913) and they didn't get it; they **DON'T** have it. There is no constitutional base for a tax on the wages for Americans living and working in the 50 States of the union. Period. End of argument.

Radio Frequency Identification (RFID) Microchips

- The latest technology for identifying people at the point of self, when they make purchases –is actually the implantable (RFID) microchip. There are microchips that can actually be imbedded directly into human flesh... It's a tiny glass capsule about the size of a grain of rice... it contains an RFID computer microchip with a coiled antenna and it can transmit information also at a distance.
- Katherine Albrecht –Author of "*Spychips*" (2005) makes the following comment: "RFID is a technology that uses tiny computer technology the size of a grain of sand or smaller; hooked up to miniature antennas to transmit information about items at a distance. Back in 1999, Procter and Gamble, Gillette, and MIT got together to find a way to commercialise this technology and make it small enough, make it efficient enough and make it low cost enough to essentially ---their dream is to put these tiny computer chips on every physical item manufactured on planet earth."
- Radio waves can travel through walls, they can travel through wood, and they can travel through things we normally rely on to protect our privacy, for example your purse, your bag-pack, your pocket, –anything you are wearing or carrying.
- One of the most worrisome applications of RFID microchips, are proposals to put them into cash, meaning that it would be able to track every bank note where it had been, who it had been issued to, and create in essence an audit trail. That would essentially take away the anonymity of cash we now enjoy today.
- Once everything you do is tied down to a single number and there is no longer the ability to pay with cash, then all it takes to render you a non citizen is to simply turn that microchip off. **You will no longer be able to participate in any function in society including buying food. Once money becomes digitised through RFID technology, the elite can deduct whatever amount of money they want out of your microchip, whenever they want. They can trace you whenever they want. You will be at their mercy.**

Habeas Corpus

- Habeas Corpus is "[A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention" (Legal-dictionary 2015).
- During the Bush Presidency [2001-2009 (History 2015)], President Bush signed executive orders giving him sole authority to impose martial law, and suspend Habeas Corpus. This gives him doctorial power over the people without any 'checks and balances.'

  In other words, "The government can jail you for life without charges, without a trial, without a lawyer" (Russo 2006).
- Furthermore: "The National Defense Authorization Act [4] signed by President Obama on the 31st December 2011 authorises the indefinite detention, without trial or indictment, of any US citizens designated as enemies by the executive." See Paye (2013) for further discussion regarding the suspension of Habeas Corpus.

Fraudulent Manipulation of Election Results

- In 2004 at the "Forum of Presidential Election" Clinton Eugene Curtis, a former Computer Programmer for NASA and ExxonMobil testified under oath that election results can be rigged using electronic programs.
- In 2004 he gave the following abridged testimony (Russo 2006):

  Judge: Mr Curtis, are there programs which can be used to secretly fix elections?

  Curtis: Yes.

  Judge: How do you know that to be the case?

  Curtis: Because in October of 2000 I wrote a prototype for present congressman Tom Feeney, and the company I worked for in Oviedo Florida –it did just that.

  Judge: And when you said, "it did just that" –it would rig an election?

  Curtis: It would flip the vote 51:49 –whoever you wanted it to go to, and whichever race you wanted to win.

  Judge: And would that program that you designed be something that election officials that might be on county boards would actually, could detect?

  Curtis: They'd never see it.

  Judge: So how would such a program, a secret program that, fixes the election –how could it be detected?

  Curtis: You would have to view it either with source code, or you would have to have a receipt and then count the hard paper against the actual vote total –other than that you won't see it.
- **Judge:** Given the availability of such vote rigging software and the testimony that has been given under oath of substantial statistical anomalies and gross differences – between exit polling data and the actual tabulated results, do you have an opinion whether or not Ohio elections, the Ohio Presidential elections was hacked?

  Curtis: Yes. I would say it was.

Person in the audience: ---So in other words there is absolutely no assurance whatsoever in anything in regard to these machines?
Curtis: Absolutely none.

- See Truthstream (2006) for further details regarding the full account of Curtis' testimony on how elections can be manipulated.
- Moreover, voting machine manufactures refuse to allow anyone to see the source code. Without paper ballots, the honesty of any election cannot be verified.

The War on Terrorism is the War on Your Freedom

- It is time to wake up America. These ID cards are not about defeating terrorism, but they are **ALL** about controlling the American people.
- The (mainstream) media controls the information that a person gets in various ways. They can make sure that the average American watching T.V. or reading the newspaper is going to come out with a certain mindset. She is going to say this is good, that's bad –and that is all they (media/elites) have to do.

Whoever Makes the Money Makes the Rules

- As Mayer Rothschild said "Give me control of a nation's money supply, and I care not who makes its laws". Mayer Rothschild, private banker, –knew that he and the other bankers would now control the laws of the nation. Government gave these bankers one of its most important powers, and now had to borrow money from the bankers and pay interest to finance the government.
- America has gone from people owning their own property, owning their own businesses –to a nation in debt because all the money is created by borrowing (from private banks) and this country has become one where people just live by borrowing.

Russo's Message to Mankind

- If you are in the military or law enforcement, remember you swore an oath [to uphold the law of the land]. You did not swear an oath to promote world government, or corporations.
- Now that you do understand what happened [to America, and the monetary system based on debt slavery] and how it is leading to a tyrannical one world government – the future of mankind depends on you –will you choose freedom or slavery?
- Stop being passive liberals. Stop being passive conservatives. Stop being passive centralists. Stop being passive human beings! When the media starts telling you that the country will fall apart if this is done... Do not be fooled!!
- Remember these are sick, malevolent and twisted people we are dealing with, trying to save themselves. Squash their agendas and stay on course.
- "I believe the time for mankind is time to give all or perish. Grow up or die. Grow up and become adults. Act like adults. Take some responsibility. The world which you have perceived is childlike; –and now the curtain has been pulled back for all to see."

- Unless people get active [get fully informed, and learn how governments of the world are conspiring against their people] and say –I am going to help shut down the Federal Reserve System; I am going to shut down the powers that be –the whole human race is doomed otherwise.
- We are coming near the end game and things are starting to accelerate... –and people are looking at the world leaders to declare martial law [which in turn will rid Habeas corpus].

Let's Consolidate our Problems

- If we can all get focused on how to win the game instead of all these different objectives –such as people are fighting for a better environment; people are fighting to keep their guns/gun control; there are all these different issues which are going on around the world -and which are all important on their own –but if we can consolidate on that and focus –and take all those people –and say –hey let's shut down the Federal Reserve, then we'll deal with those issues. Let's go to the agenda, the objective, the root cause of man's problems, first, then we'll go back to the other stuff.
- That is the priority right now. You have to cut off the head of the beast. And the head of the beast is the Federal Reserve System and the people behind it. You see, and that is what will save the world, and if people understand that –and they **stop** being passive liberals, and **stop** being passive republicans, but rather become **active** human beings –that is what will save the world!

Aaron Russo: Reflections and Warnings

In this documentary Russo recalls his relationship with Nick Rockefeller. Russo shares critical information Rockefeller discussed with him.

The Falsehood of September 11th 2001

- Russo met Nick Rockefeller through a female attorney who telephoned Russo and said "One of the Rockefellers would like to meet you." Russo made a documentary called *Mad as Hell* (1996), and Rockefeller had watched the video and knew Russo was running for Governor of Nevada and wanted to meet him. Russo said "Sure I'd like to meet him", and the two met and talked. Rockefeller proved to be a very smart man and shared ideas with Russo and was the person to tell Russo 11 months before 9/11 happened "There was going to be an event..."
- Rockefeller never told Russo what the event was going to be; but, there was going to be an event –and out of that event, we [America] were going to invade Afghanistan, to run pipelines from the Caspian Sea; we were going to invade Iraq, to take over the oil fields and establish a base in the Middle East, and make it all part of the New World Order; and we would go after (Hugo) Chavez and Venezuela –and sure enough, later when 9/11 happened –and I remember he was telling me how you are going to see soldiers looking in caves for people in Afghanistan and Pakistan and all these places and there is going to be this war on terror which has no real enemy and

the whole thing is a giant HOAX, but it is a way for the government to take over the American people (and the world).

- There is no question about it [that 9/11 is a hoax]. Nick Rockefeller said to Russo: "There is going to be a "War on Terror"" and he was laughing. Who are we fighting against...? Why do you think 9/11 happened and then nothing has happened since then? Do you think that our security is so great here that these people who pulled off 9/11, who were able to lock down another plane...? Come on it is ridiculous... 9/11 was done by people in our own government and our own banking system to perpetuate the fear of the American people into subordinating themselves to anything the government wants them to do. That is what it is about and to create this endless war on terror.
- Nick Rockefeller was laughing when he said, "We are going to be sending men into caves in Afghanistan and Pakistan" –and it was just cynical, he kept laughing and saying "Look how stupid everyone is! We can do whatever we want!"
- 9/11 was the first lie; and the next lie was to go into Iraq, to get Saddam Hussein out with his weapons of mass destruction (when the real issue was control of the oil fields) –that was the next lie. 9/11 created an endless war on terror that would go on and on and you can never define a real winner. There is no one to defeat and so it goes on and on forever. And they can do whatever they want; because they scared the hell out of the American public.
- This whole war on terror is [perpetuated on] a fraud. It is a farce. It is very difficult to say it out loud because people are intimidated in saying it. Because if you say it they want to make you into a nutcase –but the truth has to be... and the truth has to come out. The fact of the matter happens to be the whole war on terror is a fraud, it is a farce. Yes. There is a war going on in Iraq, because we invaded Iraq, and people over there are fighting... but the 'war on terror' –it is a JOKE –you know, and until we discover what really happened in 9/11 and who was responsible for 9/11 –because that is where the war on terror emanates from. That is where it comes from. It was 9/11 which allowed this war on terror to begin and until we get to the bottom root of 9/11 –the truth of 9/11 we'll never know about the war on terror.
- Russo was in Tahiti when 9/11 happened and he got a call from his son – and his son said –"The twin towers, they were just attacked and they are falling down or something..." Russo was in Tahiti and had just woken up from sleep. Russo didn't realise what it was immediately (11 months after Rockefeller had told him about 9/11 and it had happened –because Rockefeller said there was going to be an event –he wasn't specific) –but after Russo saw that America was going to go to Iraq and Afghanistan, that is when he realised, and equated it to what Rockefeller had said.
- 9/11 was only a manifestation to create a fear in the American public. So that we would obey and do what they want us to do. Take for example, Richard Reed 'The shoe bomber' –now here is a guy who is 6 feet 6, ugly as can be; I heard he smelled... He sits on a plane, lights a match in a non-smoking area, to put his shoe on fire... surrounded by people... That is idiotic! If you were going to blow up a plane... you would go into the bathroom... you close the door... and you put your shoe on fire...

you are not going to sit there.... surrounded by people, lighting matches in a no smoking flight... they (elite) want you to believe this nonsense. That is ridiculous!

- The war on terrorism is to keep people in fear. It is an endless war without a real enemy (‘terrorists’) –so that people would submit and do whatever the government wants them to. Submit to searches, give you ID cards, put Radio Frequency Identification (RFID) chips in you etc. You become servants to the elite –that is what this is all about...
- Freedom of liberty is what people really want, and it is time to stop the duplicity of the government from lying to us. You see many people know the truth of what is happening in this country; like 9/11 but they are afraid to stand up. People have to stand up and find their courage and say “I’m not going to take this anymore, I know the truth” –and they create a situation where if you tell the truth, you are considered a lunatic.
- In other words if someone goes on a T.V. show and says that 9/11 was an inside job –immediately the person is labelled an idiot or crazy. They call you names. **You cannot be afraid of that.**
- If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

The New World Order Agenda

- The whole agenda is to create a one world government where everyone has an RFID chip implanted in them. All the money is to be in those chips [a cashless world]. This information came straight from Nick Rockefeller himself. That is what the ultimate plans of the global elite, banking industry and Rockefeller wants to accomplish.
- The agenda is to implant everyone with RFID microchips. All money is transferred to those microchips. There is no more cash. Money would be in microchips. Instead of having cash, you would have money in your microchips, but whenever they want, they (elite) could take whatever amount out of your microchip whenever they want to. Total control. If you are a protestor they just turn off your chip; you cannot buy food; you cannot do anything; it is total control of the people.
- So they want a one world government controlled by them. Everyone being chipped, all the money in those chips, and they control the chips and they control people, and you become a slave. You become a serf to these people –that is their goal. That is their goal; that is their intentions.
- Russo did not believe in enslaving people and Rockefeller would question him in the following manner: “Why do you care about them? Why do you care about those people? What difference does it make to you [Aaron]? Take care of your own life.
  Do the best you can for you and your family. What do the rest of the people mean to you? They don’t mean anything to you. They are just serfs. They are just people...”

- Rockefeller asked –"Why are you fighting for the people for, what is it all about? The people have to be ruled. The constitution, what you are standing for is only for a few people, it's only for a few individuals who can live that way and we believe that it is best for society to be ruled by an elite people who control everything." Russo told Rockefeller he does not believe that. Russo believes: "God put me on this earth to be best person I could be and put everyone on this earth to be the best they can be, and NOT to be a slave and a sheep to YOU and these people (elite) –and I do not understand why you want to control everything. What is the need for that?"
- It was just a lack of caring [from Rockefeller's part], and that is just not who Russo was. It was just sort of like cold you know, and Russo used to say to Rockefeller, "What is the point, of all this? You have all the money in the world you need; you have all the power you need, what is the point? What is the end goal? Rockefeller said "The end goal is to get everyone chipped. To control the whole society. To have the elite people [the bankers and government] controlling the world."

Women's Liberation from the Perspective of the Elite

- Women's' Liberation was founded by the Rockefellers.
- Rockefeller asked Russo: "What do you think women's liberation was about?" At the time Russo had a pretty conventional thinking about it and he said "It's about women having the right to work; get equal pay with men; just like they won the right to vote." At this point Rockefeller started to laugh and he said to Russo "You're an idiot"
- Rockefeller said –"Let me tell you what that was about. We the Rockefellers funded women's' liberation. We are the ones who got it all over the newspapers and television. The Rockefeller foundation –and you want to know why? There were two primary reasons:"
- 1) "We couldn't tax half the population before Women's Liberation and the second reason was"
- 2) "Now we get the kids at an early age [because both parents are away from home working] –we can indoctrinate the kids how to think, so it breaks up their family –the kids start looking at the state as the family, as the school as the officials as their family, not as their parents teaching them, and so those are the two primary reasons for women's liberation."
- Russo thought, up to that point, Women's Liberation was a noble thing; however when he saw the Rockefellers' intentions behind it, where they were coming from when they created Women's Liberation; the thought process of it; Russo saw the evil behind what he thought was a noble venture.

America is a Republic

- America is a constitution republic / it is supposed to be a constitution republic –and **NOT** a democracy. The majority should not rule, and nor should the majority take over the inalienable rights of the minority.

- Americans pledge allegiance to the "Flag and the **Republic** which it stands" –and **NOT** democracy.
- Democracy is the worst form of government you can have because it is majority rule. Therefore the government can tell you what to do because 'the majority' wants it. It is irrelevant what the majority wants. Decisions should not take away the inalienable rights of an individual.
- Russo also said "It doesn't matter who you vote for, republican or democrat; –they are the same –neither one of them is stopping the Federal Reserve or paying income taxes.
- The elite (Federal Reserve) have taken over the American government; there is no difference between republicans and democrats. There is no difference between the two parties. The duality is manufactured. They (Federal Reserve; central banks) control both parties. It doesn't matter to the elite which one wins, because whoever is running for President will be someone they anoint. Whoever runs for President, will do whatever the elite want them to do. The fact of the matter happens to be that you cannot win an election unless you have enough money to win; they (Federal Reserve; central banks) make sure who gets the money.

Depopulation

- Russo and Rockefeller discussed many things –and one of the things Rockefeller brought up in conversation was reducing the world population. Rockefeller felt that there are too many people in the world. In a way Russo agreed that there are too many people in the world, but he **does not** think he has the authority to say who dies and who lives; but the elite felt that they want to reduce world population and Rockefeller felt it should be reduced by half.
- Rockefeller even mentioned to Russo in conversation, that they were having a real problem trying to solve the Israel/Palestine problem –and they were playing with the idea of bringing Israel to Arizona –and taking everybody from Israel and giving everybody a million dollars, and setting up Israel in the state of Arizona –because that is a problem that they are not in charge of.

Borrowing Money from Private Banks causes Inflation and Debt

- These people (elite) control the money so they make all the rules, and therefore they put the rules in which they want into effect, and the truth is America has really become a socialist communistic country. Everyone says it is a capitalistic country, but how can it be a capitalistic country when you have a central bank? That is the first question people should ask. It can't be. It is a planned economy, it is a phony!
- If they want to create prosperity, they just print dollars, or put digits into the economy. Now you have prosperity. You do not have real prosperity, you do not have real manufacturing, you just have money being injected in which is infusion of credit. This makes the government go into more debt.

- Whoever makes the money makes the rules. Rothschild said that. Why are we allowing these private bankers to make the money for our country? It is nonsense. Why are we paying interest to these banks to make money for us when the government can do it itself without paying interest, without all that debt? There is no answer to that question and it is a question no politician will raise. Everybody talks about America’s debt... We are in debt because we borrow money... but we don’t HAVE to borrow money. They designed it so that we can go into debt. We can create the money, and back it by gold so that they cannot create too many of it, so that you do not have the inflation, and do what the Founding Fathers told us.
- Why in the world does the American government borrow money from the banks when they have the ability to create it themselves without borrowing it, and paying interest on it? Why? Nobody can answer that question –not one politician ever raises that. Why does the American government ever borrow money, when they can create it without paying interest?

  –And people say that –well if the American government creates it, it will cause inflation. And that is their answer. And Russo says well let’s look at it: the American government has the Federal Reserve do it, which creates the same inflation as if they did it, but also with the inflation –now you are getting massive debt –so with the Federal Reserve you have inflation and debt. Now if the American government made the money, backed by gold which would limit the amount they could make –you wouldn’t have debt and you wouldn’t have inflation.
- It wasn’t until 1913 when the Federal Reserve came in, that America had inflation. Before then there was no inflation for 100 years. There were points and spikes, mostly during the Civil War –but basically there was no inflation other than during that short period of time. I mean a loaf of bread would cost the same thing. People could plan their lives.
- Today, they have planned inflation, and now you have two parents working, they cannot afford to take and pay for their family anymore; the kids are going to state run schools now, the kids are being indoctrinated how to think; they are being given Ritalin, they are being given all these drugs, the whole country is being dumb down, it is all because of the Federal Reserve System; and the Federal Reserve system and these bankers are responsible for the demise of America. And if we ever want to win this battle you must shut down the Federal Reserve System, and we must shut down these bankers and restore sound money to this country.
- If you analyse the situation and if you realise that since the Federal Reserve has come into being since 1913, illegally, without a constitutional amendment, by bribing a few senators during Christmas vacation, they turned over the most important power that the American government has, the creation and issuance of money to a private bank.
- Through that private bank issuing money they have destroyed this country. They have destroyed the purchasing power of money in this country; they have created social programs that are destroying this country.
- The Federal Reserve has created massive inflation in America which means the American worker has to keep on making more money to keep up with the cost of

living. The more money they make to keep up with the cost of living, the less competitive they become in the world economy. So now what happens is that we have to pay our workers so much to keep up with the cost of living; and then they (elite; government; corporations etc.) say screw the American worker; let's go overseas now to get the cheap labour.

- The inflation the Federal Reserve has created has now allowed other countries to outcompete us. Other countries do not have to pay as much money as we have to pay to our workers to survive. So now we are not competitive anymore and we have lost our manufacturing base. We have lost our competitive edge.
- "Freedom to Fascism" is a documentary that everybody should see. Russo and his team show the fraud of the income tax; they show how Judges put people into jail for no reason; they show the corruption of the justice system. They show how the Federal Reserve came into being and how it is controlling society and how all the central banks are working together through the bank of International Settlements, in Switzerland which is the central bank for all central banks and how all are working together to create this one world government; this one world order; which is what they are trying to do.

The Deception of the Council of Foreign Relations (CFR)

- Russo was interested in joining the Council of Foreign Relations, but he found out from Rockefeller himself, that part of the end goal of the CFR is to get everyone RFID chipped. To control the whole society and have the elite people (bankers, government etc.) controlling the world. Russo asked Rockefeller "Do all the people in the CFR believe the way that you do?" Rockefeller said "No, no, no. Most of them believe they are doing the right thing. A lot of them believe it is better off being socialistic. We have to convince people that socialism is really capitalism." Because America is becoming a socialistic country; it is a communist country today.
- Russo's friendship with Rockefeller became one where they would share thoughts, ideas and philosophies and Rockefeller wanted Russo to become a part of what they were doing [enslavement agenda], and for Russo to become a member of the CFR; Rockefeller offered various business opportunities for Russo to get involved in the CFR and to **not** take up the fight or the battle that Russo had been taking up in the past. Rockefeller wanted Russo to drop the idea of helping the people; because "What was the point in Russo fighting for the people?" Rockefeller would question.
- Russo asked Rockefeller do all the people of the Council of Foreign Relations feel the same way you feel? Rockefeller said "A lot of them think they are doing the right thing, they think that socialism is the best way to go (but this form of socialism involves redistributing the wealth for the elites and not to everyone), they think that they are doing the right thing." But the people at the (very) top they all know the truth of what is happening. (Therefore, the good people in the CFR are also under an illusion, and do not know that effectively they are working for evil men who have the overall goal to control the populace of the world).

- So it is compartmentalised within the elite structure as well. All the people in the CFR -2000 to 3000 people like Dan Raddler –they don’t know what is going on--- they join the CFR because it is prestigious. They think it is good for business, it is good for this; they don’t know what is really happening –the evil that comes out of it –that is emanating out of it.
- In terms of the CFR, in terms of compartmentalisation, there are many good people which Russo believes are part of these organisations who do not even understand what these organisations are really about. For example, when Russo was in Germany, doing cancer treatment, there was a gentleman there who was visiting a friend of his with cancer. The gentleman visiting his friend was a member of the CFR, and him and Russo were talking and Russo showed him the movie (*Freedom to Fascism* (2006)); --and he said, “Oh my god! I’m going to resign.” “I had no idea this is what the CFR is about”. He had no idea; he is just a nice guy, who thought he was joining a prestigious organisation.
- A lot of people join the CFR because they think it is a prestigious organisation; it will help them in business; make good business contacts etc. They do not have an understanding that the CFR is really about world domination. How they, and the Trilateral Commission, The Bilderberger [Group], the banks, all work together to control the people –a lot of them do not understand that. They do not see the big picture. They think: Oh the CFR is a prestigious organisation. I’ll make this, I’ll make that, and I can do business deals. It is just business to them. The CFR wants to get the people in there that have influence and power, and so they are part of that (enslavement agenda), and so they are not opposed to them. So the whole country is becoming the haves and the have not. You are getting the very, very wealthy and the middle class being destroyed and you are getting the poor people.
- In Russo’s words: “You can call the CFR what you like, but it is a criminal organisation. Run by criminals. But people do not think of it as a criminal organisation, because it has ‘class’; ‘style’; ‘prestige’; –and it is ‘respected’; so people do not look at it as being a criminal organisation. That is what a great job they have done.”

Combining America, Mexico and Canada into one Country

- Imagine this... here you are in America, and they (elite) are combining American, Canada and Mexico into one country. The North American union. And the American people do not know anything about it. It is not even in the press. They would rather talk about Rosy O’Donnell and Donald Trump calling each other names than discussing the fact that we are merging into one country. This isn’t even reported.
- The fact of the matter happens to be, that tells you how controlled the media is. The elite control the media, and they control governments and they are all in bed together. Here you are combining American, Canada and Mexico into one country, and you do not see it in the press. You do not see it in the press. Why? This should be one of the top stories everywhere –and the elite are not worried about it. That tells you –there is the evidence that [media] it is controlled. They do not want the American people to

know what is going on, that is why they do not protect our borders. That is why we are losing our constitution, the very document that secures our freedoms.

Russo's Vision on How to Bring These People Down

- There is no question we are in tyranny, there is no question that the American citizen is no longer a free individual human being, to do the things that they wish to do. We're slaves and it is getting worse.
- We are dealing with complete evil; and until the American people wake up and say, we do not want this evil in our country anymore and we want to come back to a country of decency and goodness, integrity and honour, we are going down that road and that is what it is going to take? It is going to take people to stand up and say we do not want to live in this kind of world anymore. I believe we should pull all of our troops out of Iraq, I believe we should leave other countries alone. Let other countries live their lives the way they choose to. Stop trying to spread 'democracy' around the world, which is the worst form of government there is anyway (because 51 percent rule over 49%), restore our republic to what it is supposed to be and go back to what the founding fathers gave us. Restore the republic.
- In Russo's opinion: The populace must shut down the Federal Reserve System –and there has to be an uprising. There has to be an uprising. People have to stand up. One person cannot do it alone. You cannot do it alone. People do not seem to have the courage to do what they have to do.
- A lot of people in Hollywood know the truth, they do not want to stand up and speak about it; I know many of them have seen my movie (*Freedom to Fascism* (2006)) and they know I am right, and they want to talk about it because everybody is afraid. Everybody is afraid because they think that the money they get from the Federal Reserve is really money and they have a comfortable lifestyle and they are afraid of change. They are afraid to stand up for what is right, and until people are willing to stand up and have the courage to do what they need to do, it is not going to change; and hopefully we can affect change when people stand up and say "Hey, I've had enough".
- We (the populace) have one advantage. They (elite) need us to cooperate. See, if we do not cooperate with them, they cannot win. They always need our cooperation with them to go along with their programs. They try to "sell us". Democracy; this majority says this; believe in this; do this, do that; the war on terror; we have to be scared... They are always trying to do things to "sell to us" to go along with them, and once we learn not to cooperate with them; then we win the game.
- That is the point, do not cooperate with them, and do not go along with the program anymore. Stop it. Join forces, and bring freedom back to this country. It is going to take people who believe in freedom; The Constitution and the Founding Fathers, Thomas Jefferson, to make this country whole again, because right now it is in the grip of the evil ones, and the only way to stop that is for good men to stand up.

- We have to stop being scared. We have to do what is necessary to take back what is ours. We have to stop these bankers, these elite, full of liars, congressmen full of liars. They are destroying our borders.
- So through these bankers, attempting to take over America, knowing that America was the freest nation on this earth, it was necessary for them to takeover America, take away gun rights, freedom to bear arms, and create a country where we become slaves, because once they take over America, the rest of the world becomes a lot easier for them. And so by creating 9/11, an event to terrify the American people that we are being 'attacked by terrorist', you create a world where there is an enemy that can **never** be pinpointed. You can never win the battle. It is 100 year war – a never ending war on terrorism. So you are always fighting this war, and through the war on terrorism, through 9/11 which is the first lie, then you create the war on terrorism which is the next lie, then you create the war in Iraq through weapons of mass destruction, which is the next lie –so you get one lie, to the next lie, to the next lie, -- now it is going to be Iran the next lie –and sending more troops and insurgents into Iraq.
- Restore America's Republic back to what it is supposed to be. Get the bankers out of our government. Get government to stop borrowing money from the banks. Government should make its own money; restore the Republic. Restore individual freedoms. That is what this country is about -and until we do that we are going to be slaves.
- You have to take away the creation of money away from the private bankers and you will solve 95% of your problems.
- Americans, mobilise, stand tall, stand together, tell the government you are "Mad as hell!" Do not cooperate with the government do not accept a National ID card. Do everything in your power to restore freedom and your individuality back to America. Stop being a country run by the institutions for the institutions. Let's go back to "We the people, by the people for the people", as opposed to, we the institutions, by the institution, for the institution. Stand up for your individual rights. Stand up for the God leaders that are in each and every one of us!

The Trans-humanist / Post Human Agenda

Transhumanism is a cultural and intellectual movement that believes we can, and should, improve the human condition through the use of advanced technologies. One of the core concepts in transhumanist thinking is life extension: through genetic engineering, nanotech, cloning, and other emerging technologies, eternal life may soon be possible. Likewise, transhumanists are interested in the ever-increasing number of technologies that can boost our physical, intellectual, and psychological capabilities beyond what humans are naturally capable of (thus the term *trans*human) (Anthony 2013). Transcranial direct current stimulation (tDCS), for example, which speeds up reaction times and learning speed by running a very weak electric current through your brain, (Anthony 2012) has already been used by the US military to train snipers. On the more extreme side, transhumanism deals with the concepts of mind uploading (to a computer), and what happens when we finally craft a computer with greater-than-human intelligence (the technological singularity) (See: “How to create a mind, or die trying”, Hewitt 2012) (Anthony 2013).

Moreover, put simply, “posthumanism” can be defined as that condition in which humans and intelligent technology are becoming increasingly intertwined (TheNanoAge 2015). For readers interested in learning more about the posthuman / transhuman agenda see BT Soul Catcher 2025 (BEAMS 2007); Avatar Project 2045 (2045 Initiative 2015; Borghino 2012) and Mind Clone Robot (Bloomberg Business 2015; RT 2015) which have all been disclosed as methods of transferring the human consciousness to a computer.

The Phases of Sleep and Rapid Eye Movement (R.E.M) Sleep

Sleepers pass through five stages of sleep: 1, 2, 3, 4, and REM (rapid eye movement) sleep. These stages progress cyclically from stage 1 through REM then begin again with stage 1. A complete sleep cycle takes an average 90 to 110 minutes (Sleepdex 2015). In other words, after falling asleep, it takes approximately 90-110 minutes to enter REM sleep.

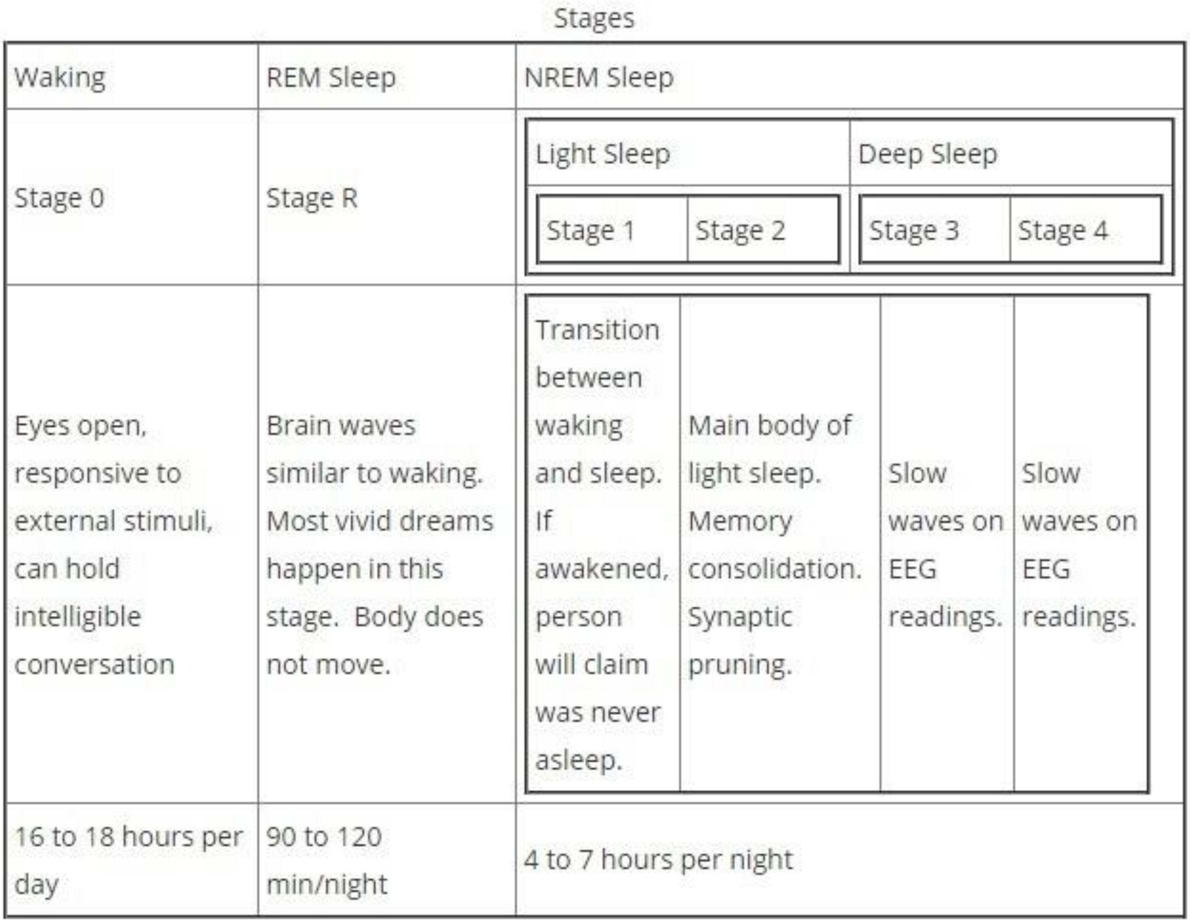

Stages

Waking	REM Sleep	NREM Sleep			
Stage 0	Stage R	Light Sleep		Deep Sleep	
		Stage 1	Stage 2	Stage 3	Stage 4
Eyes open, responsive to external stimuli, can hold intelligible conversation	Brain waves similar to waking. Most vivid dreams happen in this stage. Body does not move.	Transition between waking and sleep. If awakened, person will claim was never asleep.	Main body of light sleep. Memory consolidation. Synaptic pruning.	Slow waves on EEG readings.	Slow waves on EEG readings.
16 to 18 hours per day	90 to 120 min/night	4 to 7 hours per night			

Image 1: Stages of sleep. Source: Sleepdex: (2015)

Any sufficiently advanced technology is indistinguishable from magic.
Arthur C. Clarke

REM Sleep

Most dreaming occurs during Stage Five, known as REM. REM sleep is characterized by eye movement, increased respiration rate, and increased brain activity. REM sleep is also referred to as paradoxical sleep because, while the brain and other body systems become more active, your muscles become more relaxed, or paralyzed. Dreaming occurs because of increased brain activity, but voluntary muscles become paralyzed. Voluntary muscles are those that you need to move by choice, for example, your arms and legs. Involuntary muscles are those that include your heart and gut. They move on their own (Sleepdex 2015; Walcutt 2013).

Rapid eye movement, or REM sleep, is when you typically dream. You may have images float by in earlier stages, particularly when you are going through Alpha or Theta (brain waves), but the actual dream state occurs in REM (Walcutt 2013).

This period of paralyzation is a built-in protective measure to keep you from harming yourself. When you are paralyzed, you can’t leap out of bed and run. Do you ever feel like you can’t escape during a dream? Well, the truth is, you can’t. You can breathe, and your heart is working, but you really can’t move (Walcutt 2013).

The reader is also advised to see the Horizon documentary “Why Do We Dream?” (BBC Horizon 2009). This video is available, on YouTube, and details sleep, the phases of sleep, REM sleep, the period of paralysation and more.

Current Advances in Technology

This is a section some readers may have difficulty with, because reading about these current technologies alone, one cannot help but feel that such technologies described below sound like the stuff of science fiction and fantasy -although they have been corroborated through public drip-feed disclosure (explained below). Nevertheless, I urge the reader to keep in mind Arthur C. Clarke’s quote (above), as well as Phil Schneider’s testimony regarding the advancements of military technology in comparison to the general public’s technology (for every 12 months, military technology outstrips the technology the public is accustomed to by a rate of 44 years). Moreover, for the reader who finds difficulty understanding the technologies described below, seek the supporting articles / videos.

Memory Suppression Technologies

Memory suppression technologies are any scientifically advanced technologies which are used to suppress memory. Examples of how memories can be suppressed can be found by reading Winter's (2014) article which details how memories can be suppressed using light; and Greenberg's (2013) article which discusses memory suppression through gene / chemical modification.

Mind-Voice Technology

Mind-voice technology is an advanced technology which is capable of reading, listening, hearing or broadcasting your inner voice / thoughts. Examples of articles which discuss mind-voice technology are: Prigg (2014) details software which can read the inner voice; and New Scientist (2014) which also discusses a brain decoder which can eavesdrop on one's inner voice.

H.A.A.R.P. Technology

The High Frequency Active Auroral Research Program (HAARP) is a radio transmitting system that can bounce signals off the ionosphere (a region of the Earth's upper atmosphere 60km (37 miles) to 1000km (620 miles) altitude) and back to earth to probe deep into the earth or sea, its proponents say. The system could locate minerals or communicate with submarines. (Begich & Manning 1997; Sheen, Begich & Robbins 2005).

HAARP can also:

- Disrupt human mental processes.
- Knock out all global communications systems.
- Manipulate global weather.
- Change weather patterns over large areas.
- Interfere with wildlife migration patterns.
- Hurt ecosystems.
- Negatively affect your health, moods, and mental states.
- Unnaturally impact the Earth's upper atmosphere.

This illustration (below) shows the ionosphere's relationship to the Earth. The illustration appeared in the HAARP Environmental Impact Statement on page 10-125 of Volume II.

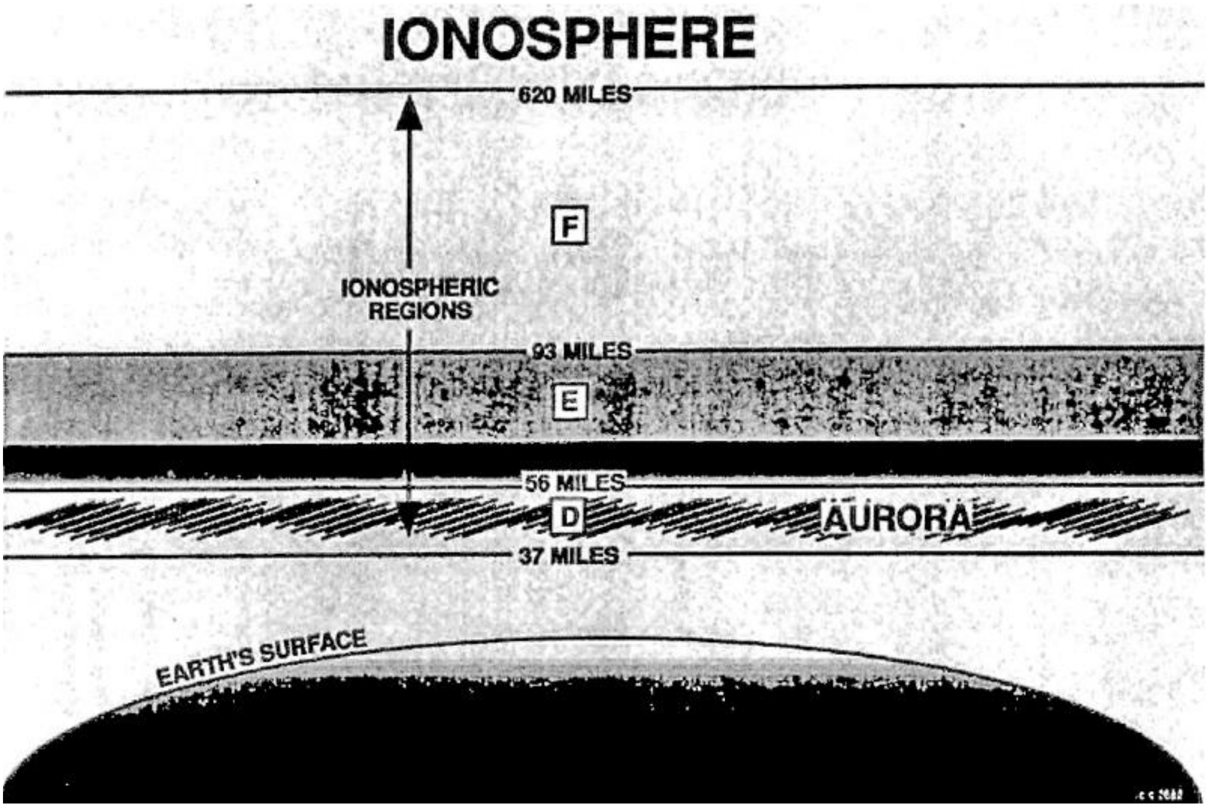

Image 2: The ionosphere's relationship to earth. Source: (Begich & Manning 1997)

The ionosphere protects the earth. HAARP (High frequency Active Auroral Research Program) is made to beam more than 1.7 gigawatts (billion watts) of radiated power into the ionosphere -the electrically charged layer above Earth‘s atmosphere. Put simply, the apparatus is a reversal of a radio telescope - just transmitting instead of receiving. It will boil the upper atmosphere. After disturbing the ionosphere, the radiations will bounce back onto the earth in the form of long waves which penetrate our bodies, the ground, and the oceans.

HAARP represents a technology which could lead to a new class of weapons that could change our world profoundly - an all-purpose military tool. If misused, the tool could mess up the weather. It could be used against humanity in a way that would change what people think, believe and feel. It could be used for good or evil, just as a harp can produce the music of Mozart or the melody of a death march.

H.A.A.R.P. and Weather Control

- “The theoretical implication [of Dr. Robert Helliwell and John Katsufrakis of Stanford University in 1974] suggested by their work is that global weather control can be attained by the injection of relatively small 'signals' into the Van Allen belts (radiation belts around Earth) - something like a super-transistor effect” said Frederic Jueneman.
- Yes. The weather can be controlled using HAARP technology.

- A series of weather disasters began in 1960, according to a CIA report mentioned in the editorial, but at the time climatologists couldn‘t look ahead and see that droughts, floods and abnormal temperatures would continue beyond that decade. As if natural disasters weren‘t bad enough, the CIA reported that national governments were already able to manipulate weather for military purposes [using HAARP technology].

As far back as 1958, the chief White House advisor on weather modification, Captain Howard T. Orville, said the U.S. Department of Defence (DoD) was studying ways to manipulate the charges of the earth and sky and so affect the weather by using an electronic beam to ionize or deionize the atmosphere over a given area. In 1966, Professor Gordon J. F. MacDonald, associate director of the Institute of Geophysics and Planetary Physics at the University of California, Los Angeles, was a member of the President‘s Science Advisory Committee, and later a member of the President's Council on Environmental Quality.

Gordon J. F. MacDonald published papers on the use of environmental control technologies for military purposes. MacDonald made a revealing comment: ―The key to geophysical warfare is the identification of environmental instabilities to which the addition of a small amount of energy would release vastly greater amounts of energy. MacDonald had a number of ideas for using the environment as a weapon system and he contributed to what was, at the time, the dream of a futurist. When he wrote his chapter, ―"How to Wreck the Environment" for the book "Unless Peace Comes" he was not kidding around.

In the text MacDonald describes the use of weather manipulation, climate modification, polar ice cap melting or destabilization, ozone depletion techniques, earthquake engineering, ocean wave control and brain wave manipulation utilizing the planet‘s energy fields. He also said that these types of weapons would be developed and, when used, would be virtually undetectable by their victims. He was not some wire haired fanatic when he made these observations in 1966 -he had the credentials of a world recognized scientist. What his futuristic concepts became, are the things which projects like HAARP are made of...

H.A.A.R.P. and Mind Control

- Radio frequency radiation, acting as a carrier for extremely low frequencies (ELF), can be used to wirelessly entrain (adjust) brain waves.
- We are talking about very, very low power requirements. The trick for influencing brain activity is in the combination of frequency, power level and wave form.
- As Dr. Patrick Flanagan, one of America's most gifted inventors noted in an interview, the HAARP project could be not only the biggest ionospheric heater in the world, but also the biggest brain-entrainment (brain adjustment) device ever conceived.
- According to HAARP records, when the device is built to full power it can send very low frequency (VLF) and extremely low frequency (ELF) waves using many wave forms at energy levels sufficient to affect the mental states of entire regional populations.

- The HAARP transmitting system could be used unintentionally or intentionally to alter mental functions.
- If HAARP is tuned to the right frequency, using just the right wave forms, mental disruption throughout a region could occur intentionally or as a side effect of the radio frequency transmissions [in other words: “mind control”].

RFID Technology

Radio Frequency Identification (RFID) microchips are microchips that can be directly imbedded into human flesh. This section discusses the dangers of RFID chips. See Rense (2001) for full review.

- RFID technology links the brains of people via implanted microchips to satellites controlled by ground-based super-computers.
- Today they are small enough to be inserted into the neck or back, and also intravenously (through a vein) in different parts of the body during surgical operations, with or without the consent of the subject. It is now almost impossible to detect or remove them.
- Implanted human beings can be followed anywhere.
- Today's microchips operate by means of low-frequency radio waves that target them. With the help of satellites, the implanted person can be tracked anywhere on the globe.

RFID Technology and the Medical Profession

- One reason the dangers of implantable microchip technology has remained a state secret is the widespread prestige of the psychiatric DIAGNOSTIC STATISTICAL MANUAL IV produced by the U.S. American Psychiatric Association (APA), and printed in 18 languages. Psychiatrists working for U.S. intelligence agencies no doubt participated in writing and revising this manual. This psychiatric "bible" covers up the secret development of Mind Control technologies by labelling some of their effects as symptoms of paranoid schizophrenia.
- The Psychiatric Diagnostic Statistical Manual (DSM) for mental disorders has been a brilliant cover up operation in 18 languages to hide the atrocities of military and intelligence agencies' actions towards their targets. THE MANUAL LISTS ALL MIND CONTROL ACTIONS AS SIGNS OF PARANOID SCHIZOPHRENIA.
- If a target is under surveillance with modern technology via TV, radio, telephone, loudspeakers, lasers, microwaves, poisoned with mind altering drugs via air-ducts, giving familiar smells which cause headache, nausea and so forth, if s/he claims her/his clothes are poisoned, her/his food or tap water as well --- all medical schools teach their students that the person is paranoid, ESPECIALLY if s/he believes intelligence agencies are behind it all.
- Never is the medical profession told that these are routine actions all over the world by intelligence agencies against their targets. Thus, victims of mind control are falsely considered mentally ill and get no help since they are not believed and their suffering is doubled by misinformed health professionals.

Further Implications of RFID Technology

- How many people realize what the implications of implantable chips actually mean? It means total loss of privacy and total outside control of the person's physical body functions, mental, emotional and thought processes, including the implanted person's subconscious and dreams! For the rest of his / her life!
- It sounds like science fiction but it is secret military and intelligence agencies' mind control technology, which has been experimented with for over half a century (since 1950s). Totally without the knowledge of the general public and even the general academic population.
- Supercomputers in Maryland, Israel and elsewhere with a speed of over 20 BILLION bits/sec can monitor millions of people simultaneously. In fact, the whole world population can be totally controlled by these secret brain-computer interactions, however unbelievable it sounds for the uninformed.
- Neuro-electromagnetic involuntary human experimentation has been going on with the so-called "vulnerable population" for over 50 years, in the name of "science" or "national security" contrary to all human rights. It happens today in the **USA, Japan, and Europe**. With few exceptions, the mass media suppresses all information about the entire topic.
- Only increased public awareness of the microchip implants, their frightful consequences to privacy by influencing of individuals' thoughts and actions, causing people to become biological robots with physical and emotional pain whenever the supercomputer technician so wishes, is enough reason to refuse to take the microchip into your body for whatever reason.
- It is the biggest threat to humanity and the most sinister plan to enslave the human race forever.

If you have a choice and want to remain a normal human being with privacy, **DO NOT** have your children implanted **NOR** yourself implanted with RFID microchip(s) (or any other type of implantable microchip). Otherwise your vision, hearing, sensing, thoughts, dreams and subconscious will be influenced by an outsider, who does not have your best interests in mind. For the rest of your life!

Mind Control: MK Ultra Technology

MK Ultra today has evolved from the 1950s variation of mind control (MK Ultra Compendium 1980), whereby drugs such as LSD, and interrogation were used as methods to weaken the mind of the individual to force confessions through mind control. Today, mind control is achieved through the implantable microchip (mentioned above; Rense 2001).

- The brain functions of an implanted person can then be remotely monitored by supercomputers and even altered through the changing of frequencies.
- Once implanted, the U.S. National Security Agency's (NSA) 20 billion bits / second supercomputers could now "see and hear" what you are experiencing with a remote monitoring system (RMS).

- Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded into thoughts, pictures and voices. MK Ultra technology is therefore capable of Mental Video and Audio Projection, as well as Artificial Telepathy.
- The mass media have not reported that an implanted person's privacy vanishes for the rest of his or her life. S/he can be manipulated in many ways. Using different frequencies, the secret controller of MK Ultra technology can even change a person's emotional life. S/he can be made aggressive or lethargic. Sexuality can be artificially influenced. Thought signals and subconscious thinking can be read, dreams affected and even induced, all without the knowledge or consent of the implanted person by using MK ultra technology.
- Memory suppression technologies are used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses.
- Mind control techniques, such as MK Ultra, can be used for political purposes. The goal of mind controllers today is to induce the targeted persons or groups to act against his/her own convictions and best interests. Zombified individuals can even be programmed using MK Ultra technology, to murder and remember nothing of their crime afterward.
- The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. The purpose of mind control, using MK Ultra technology is to disrupt memory, discredit people through unusual behaviour, to make them insane or to commit suicide or murder.

See Mind-Computer (2012) which discusses how artificial telepathy is achieved. Artificial telepathy, also known as ‘brain to brain communication’, is also possible using MK Ultra technology. Furthermore, review: Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013). The above mentioned articles and videos describe, demonstrate, and corroborate how audio and video projection of the brain is achieved; how dreams can be recorded and projected digitally; how the brain can be scanned to reveal hidden information personal to an individual; and what a person’s underlying intentions are, by using brain scans. Moreover, the articles and videos mentioned above confirm the functionalities of MK Ultra technology.

Furthermore, when our brain functions are connected to supercomputers by means of RFID technology, MK Ultra technology and other implantable microchips, it will be too late for protest. This threat can be defeated only by educating the public, using available literature on biotelemetry (electronic equipment that receives signals from radio transmitters) and information exchanged at international congresses.

Human Cloning

There are currently five different types of clones, concealed from public knowledge. There are Mark 1 clones; Mark 2 Clones; Mark 3 Clones; and Mark 4 clones and reanimated clones. Moreover, there are two types of cloning techniques: duplication cloning and replication cloning.

Mark 1 clones are REM sleep driven clones. However, the technology used for Mark 1 clones causes many side effects and therefore Mark 1 REM sleep driven clones are no longer a preferred choice.

Mark 2 clones are also REM sleep driven clones. Mark 2 clones have fewer side effects than Mark 1 clones, and therefore Mark 2 clones are currently the preferred method for REM sleep driven clones.

Mark 3 clones are independent clones which operate on microchip containing the entire consciousness of an individual.

Mark 4 clones are also independent clones, operating on a microchip which contains the entire consciousness of an individual and is an advanced version of a Mark 3 clone.

These independent clones (Mark 3 and Mark 4) have a lifespan of 6 months to 12 months and require adjustments after this period to run efficiently once more. Without these adjustments, their functionality weakens.

To "reanimate" means "to restore to life; resuscitate; revive". Consequently, reanimated clones are clones which are genetically identical to that of a person who once lived.

Replication cloning is what the public is most familiar with. Replication cloning involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

However, duplication cloning is a current concealed advanced from of cloning, and it involves taking as little as two cells from an individual, adding a constant electrical charge to the cells until a complete human being is formed.

Duplication cloning is similar to the process of regenerative medicine as demonstrated by Dr. Stephen Badylak's video "How to grow a New Fingertip" (CBS 2008; Science Channel 2014) where Badylak states: 'A whole human can be grown within 9 months.' On average it takes 5 months to grow a duplicate clone of an original by means of advanced scientific and technological regenerative procedures.

Cloning Centre and Cloning Technology

A cloning centre is a place where clones are produced. Cloning technology are the advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Drip Feed Disclosure and Evaluative Conditioning

Drip Feed Disclosure

Drip feed disclosure is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the "complete truth" for some time.

Drip feed disclosure is also a method to gauge public reaction used by governments, the media, multinational corporations and organisations as well as high ranking officials to "test" whether the general public is acceptant of the concealed information or not. When the public reacts favourably to the drip feed disclosure, more information is revealed and made public, and it appears to the unsuspecting observer that the people involved in making the disclosure are taking positive steps towards a favourable goal for all. However, when the public reacts adversely; information contradicting the drip feed disclosure is presented, and an expert is presented to the public who voices the concerns of the general public, and therefore it appears that the opinions of the public have been noted and research and development will not continue in the initial stated direction. Nevertheless, the truth remains concealed and research and development continues despite the aversions of the public.

Examples of drip feed disclosure involve articles such as mind uploading / downloading (BEAMS 2007), Mind Clone Robots (Bloomberg Business 2015; RT 2015), The 2045 Avatar Project (2045 Initiative 2015; Borghino 2012). I am here to tell you that I have been a spy for over 30 years and such technological accomplishments which are posted under trans-humanism / post-humanism genre have been realised many years ago and are **available** today. They are just concealed from the public.

Furthermore, Dolly the sheep was announced as the first publicly cloned mammal (Animal Research 1996), but how many readers can say they know that, four years later, a monkey (BBC News 2000), our closest primate, was cloned? This is publicly disclosed knowledge, but I suspect not many people know of a cloned monkey four years after Dolly. This is because there was such a worldwide adverse reaction to genetic cloning when Dolly was made public, that the disclosure of the cloned Rhesus monkey (BBC News 2000) was not made public knowledge on a grand scale.

Evaluative Conditioning

Evaluative conditioning is defined as a change in liking, which occurs due to an association with a positive or negative stimulus (see De Houwer et al., 2001). Simply put, this means that our preferences for brands, products, people and other things can be influenced and even modified by the presence of something we like or dislike strongly (Hale 2012).

In many settings, a neutral stimulus, called a "conditioned stimulus", often coincides with some desirable or undesirable object, called the unconditioned stimulus. An unknown brand, for example, might appear in a commercial that also depicts a happy child. Over time, stimuli that often coincide with desirable objects are perceived more favourably, whereas stimuli that often coincide with undesirable objects are perceived less favourably--called evaluative conditioning (De Houwer, Thomas, & Bauyens, 2001; Walter, Nagengast, & Trassilli, 2005; Moss 2009). An "unconditioned response" is a response to a neutral stimulus we have no / little control over. It is a natural automatic response.

In other words, our preferences for liking or disliking brands, products etc. (the neutral stimulus) can be influenced by embedding (implicitly placing) the brand etc. (neutral stimulus) with positive or negative associations. Overtime our conditioned response becomes one of conditioned favourable or negative response towards the brand, product etc. (neutral stimulus) when we are faced with the brand, product etc. at a future date. Our preferences have been guided overtime. Evaluative conditioning can change our preferences when carried out subliminally or implicitly; it does not have to be explicit.

In media; music; movies; and other forms of popular culture and entertainment, images and symbols (of stimuli which cause undesirable consequences) are embedded subliminally and implicitly and are paired with positive associations as a form of evaluative conditioning. This causes individuals who have no preconceived judgements of the stimulus to be guided to have positive associations with a stimulus which causes undesirable consequences.

In popular culture; media; advertisements; movies; music and other forms of entertainment, evaluative conditioning is used as a method of hinting (that something is wrong); showing off of power (i.e. nothing can be done to stop the negative stimulus) and as a form of gloating (i.e. we've pulled off the negative stimulus / we are pulling it off). In over the 30 years I have been a spy, I have witnessed methods of evaluative conditioning used to hint, show off power, and gloat more often than not, and evaluative conditioning is not used just to guide social order.

The fortunate aspect is that once a person learns that his / her preferences are being guided by methods of evaluative conditioning, in order to influence the person to have positive associations with negative stimuli (or feel powerless towards the negative stimuli); and that the stimulus (brand / product) does in fact cause undesirable consequences –then the 'spell' is broken. The person can now choose how he or she responds to the brand etc. Usually, once all is known: a negative stimulus is associated with negative associations; despite it being portrayed as positive through evaluative conditioning.

Consciousness Transfer

Consciousness is defined as "the state of being aware of and responsive to one's surroundings; a person's awareness or perception of something" (Dictionary Reference 2015). Consciousness can also be described as: individual awareness of a person's unique thoughts, memories, feelings, sensations and environment (Cherry 2015).

John Locke (1632-1704) was an English philosopher, Oxford academic and medical researcher who argues that it is sameness of consciousness rather than sameness of substance that constitutes personal identity. Consequently, if the psychological life is transferred from the body of a prince to the body of a cobbler (shoe mender), Locke argues, the resulting person will be the prince and not the cobbler. He would be responsible for the prince's actions and not the cobbler's; those who were close to the prince could continue their relationships with him but those who had relationships with the cobbler could not, and so on (Schechtman 2012, p.334). Moreover, basic Lockean intuition has proved to be that "consciousness transfer" can be thought of as the feat in which the person moves from one body into another (Schechtman 2012, p.334).

Consciousness transfer can also be thought of as the process of transferring or copying the mental content (including long-term memory and "self") from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. The computation device, robot or the clone, will then respond essentially the same way as the original brain (as suggested by Lockean theory on consciousness transfer) and therefore the computational device, robot or clone experiences having a conscious mind and essentially the behaviour of the computational device, robot or clone, can be attributed as belonging to that of the original.

It is **OF MONUMENTAL IMPORTANCE** that the reader understands the above statement regarding consciousness transfer is a method of **CURRENT** concealed advances in technology and not a concept relegated only to the genre of science fiction. At the very least the reader should be **open** to the possibility of consciousness transfer and the implications of consciousness transfer, in order to truly begin to understand the extent to which current concealed advances in technology are used to commit monstrous crimes.

I understand the above sounds unsettling, but there are also many wonderful concealed advances in technology (detailed in the disclosure section). The marvels in technological feats and human accomplishment will be released for the benefit of mankind once the world learns about the monstrous crimes which are committed against the them, and the earth, through the use of highly concealed technological advances and the good people of the world band together to put an end to the tyranny around us.

Now is also a good time to remind the reader that if military technology advances by a rate of 44 years for every 12 months which passes, then since 1945 military technology has advanced by a rate of more than 3000 years compared to the technology the public is currently accustomed to (44 times 70 = 3080). Now ask yourself honestly, in 3000 years from now (the year is 5015) don't you think humans will have been capable of developing a method which allows them to transfer their consciousness from one body to the next and more?

I am here to tell you that I have been a spy for over 30 years and that from the intelligence I have gathered over the years it has been illustrated to me countless times that consciousness transfer is indeed fact and currently exists. Consciousness transfer has been achieved since 1945, although this accomplishment in human advancement (and more) is concealed from public knowledge. Moreover, consciousness transfer is a highly advanced form of concealed technology used by high ranking members of society to commit unspeakable crimes against the public.

Furthermore, for the reader who wishes to understand how consciousness transfer is possible, see Petkova and Ehrsson (2008) and Ehrsson (2013). In the video Professor Henrik Ehrsson (2013) demonstrates:

1) Consciousness transfer from one body to the next (owning another body other than ones original)
2) Physiological evidence for owning the new body. In other words when consciousness has been transferred and the person perceives the new body as that of his or her own; when the new body is threatened as the mannequin (new body) was in the experiment, individuals still perceived the mannequin body as that of their own and became frightened (displayed biological and physiological responses).
3) Visual perception and stimulation causes us to perceive ownership of a new body.
4) The new body we inhabit must be similar to that of our own (it does not matter whether the new body is smaller or larger than our original so long as the measurements are proportional) for consciousness transfer to occur.

5) That it is possible for consciousness to be transferred to another person's body and have (or perceive) ownership of other person's body while being localised in your own body. In other words, dual consciousness is possible.

Petkova and Ehrsson's (2008) and Ehrsson's (2013) research gives one of the clearest publicly disclosed explanations of consciousness transfer and I urge the reader to watch the video in order to better understand my disclosure. One of the main findings from his research is that perception is not rigid; and perception does shape reality. Sight and synchronous (going on at the same time) stimulation can, and does alter the brains perception of reality. Sight and synchronous stimulation activates certain parts of the brain (the sensory parts of the brain and the motor (movement) parts of the brain). The match between the two (sight and stimulation) "convinces" the brain that 'hey, I'm no longer in this body; I'm in that one or that I have a third arm (although it is a false limp) etc. Accordingly, "perception is reality".

Disclosure: Save the Victims through your Diligence

I've done my best to preface this information as logically, sequentially and methodologically for the reader as possible. I have stated what the law is on 'whistle blowing'; I have given references and definitions where appropriate and I have provided many sources in order for the reader to research, corroborate, and better understand my disclosure, so that he or she becomes compelled to help me and victims like me, defeat tyranny and save the future of mankind; therefore if there are any areas of my disclosure which still appears suspect –and I understand, after all my efforts –things may still not appear clear to some readers because we are dealing with highly advanced concealed technologies which are not available in the world the general public lives in; and therefore it is difficult to fathom and furthermore conceive that men can be so evil to their fellow humans through the use of advances in science and technology. Nevertheless, I promise you, I have no reason to lie, I am victimised daily because of these technologies, and therefore my only option is to tell the truth, and nothing but the truth so that the people of the world can put an end to this grave injustice.

For the reader who still finds it difficult to accept my disclosure after first read, I want you to do two things:

1) **Give me the benefit of the doubt.** I am asking you to do this because as I stated earlier, one of the hardest things to do as a victim of abuse is to come forward about the abuse you have suffered. It is even more heartbreaking when nobody believes you. Consider all the victims of paedophiles you have heard about on the news. I cannot speak for their individual cases but I can attest to the fact it affects my mental psyche beyond comprehension; there is nothing worse than being ignored and that nobody believes you. So on that premise, please give me the benefit of the doubt, because I'm sure the majority of good human beings out there will feel worse when it is proven that I am a victim of severe abuse and you ignored me, when you could have done everything in your power to save me.

2) **Set out on a genuine quest to debunk my disclosure.** If my disclosure appears suspect to you to begin with, then set out on a genuine and investigative quest to debunk my disclosure because, if my disclosure initially appears circumspect to believe, then I should be proven false within minutes of your research of my disclosure, because even professionals make mistakes. There will be something about the disclosure which will just not 'add up'. However, when a person is telling the truth, then it is very difficult to debunk that person.
I promise you, it will be worth your time to know the real truth about the world, simply because the truth is liberating.

I would also like to remind you that the truth about the world has been hidden from the public for many decades, therefore when you do hear the truth of course it is going to sound absurd to begin with; but keep an open mind and pursue till you know what truth is; because "to "know" is to understand with certainty".

Words of Inspiration: Personalized from a Theoretical Physicist –Thomas Campbell

The following gives an introduction to many aspects which occur in this world behind closed doors or in plain sight depending on how familiar the reader is with the topics. Evidently, a lot of what is written predominantly affects the people of the world and it is time the people of the world stood up and reclaimed their world. I cannot stress this importance enough! The disclosure is also a basis which explains many of the anomalies which exist in the world today. I apologise in advance if the disclosure is too shocking to be believable (initially) and it shakes the foundations of your belief systems... but as I've said, rather than lash out at the information, you must question your own belief system... and probe into the mechanisms which have influenced you to adopt your current belief system... only then does the shock subside... and you can move forward with an open mind through investigation, which should lead you to either corroborate or debunk the disclosure. Keep in your heart: scepticism, optimism and open mindedness (Campbell 2008). Scepticism keeps you from believing fallacies. It keeps you from falling into that which you have not yet proven for yourself. Scepticism is a requirement for breakthroughs. Optimism gives you the ability to give merit to that which you have not yet proven for yourself and Open Mindedness gives you the ability to "see" that which has not yet been proven.

The Science of Hidden Consciousness

Furthermore, "Consciousness Science Kept Hidden" (YesEthan 2013) illustrates:

- "The heart is an electrical organ. It produces by far the strongest form of bio electricity in our body; up to 40-60 times stronger than the second most powerful source –which is the brain. This electrical energy travels through every single cell in our body and in a sense, binds the cells together. The bio-electricity field is strong enough that it can even be detected outside of the body; out into space; beyond the skin. It is very measurable electromagnetic energy, much like radio waves. What we found is that the heart produces an electromagnetic field that surrounds our entire body 360 degrees; and it can be detected about 3-4 feet with megatron meters outside of the body. Researchers at the medical facility in Kansas –say they have detected it 10-12 feet. So regardless of how far it goes, what is interesting is that we produce this electromagnetic field which can be detected; it can be measured, with sensitive mainstream medical equipment." -Howard Martin (co Author) "The Heart Math Solution""
- Scientific studies have shown: that the heart is not only the most powerful centre in the body, it is the most intelligent and has the ability to receive precognition (knowledge of a future event or situation, especially outside one's normal sense of perception). If we can create coherence with our heart and mind centre we can access our intuition more frequently.

When you read my disclosure, remember that it is the heart and NOT the brain which is the most powerful centre in the body; the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**; therefore if you find that your heart is in resonance with the themes I disclose; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

Make an effort and reach your own conclusions to the point where you feel comfort in the fact that you **KNOW** the information contained in this disclosure is indeed factual. This will be the point you will have become empowered with golden truth and compelled to fulfil your duty as a human being and also spread the truth far and wide. Yes, we are all living in very interesting times.

Disclosure: Donald Marshall's Message to the World

My name is Donald Marshall. I have been cloned by a large secretive cult know as "The Freemasons" and "The Vril Society" and "Scientologist" together called the Illuminati. For readers unfamiliar with The Illuminati see "The New World Order" (1990) by A. Ralph Epperson and Appendix C in this document; an introductory guide to the Illuminati and their agenda for the world is presented.

Furthermore, I have to tell you that human cloning has been done since 1945. Now is also a very good time to recall Phil Schneider's comment (Schneider was a Geologist and Engineer who worked on black projects and Deep Underground Military Bases (DUMBs)): "That for every 12 months which passes, military technology advances by a rate of 44 years compared to the technology the public is currently accustomed to" (Schneider 1995; 1996; Open Minds 2011). George Green (a former Financier affiliated with U.S. Presidential candidates) also attests to the fact that humans have been cloned since 1938 (Green 2008a; 2008b). Therefore if we take 1945 as a base year, then military technology or hidden technology has advanced by the rate equivalent to 3080 years than the technology the general public is currently accustomed to (2015 – 1945 = 70 years; 70 multiplied by 44 = 3080). So yes, humans have been cloned since 1945 and continue to be cloned today and that current hidden technology in the year 2015 is as much as 3000 years ahead of the technology the public is currently accustomed to. I'll repeat myself because it is **very important** the reader understands the above statement, even if, the rest of the disclosure is hard to fathom after first read. The two **most important** things to keep in mind are:
1) Humans have been cloned for over 70 years (since 1945);
2) **Present** hidden technology is more advanced -as much as 3080 years more advanced- than what the public is currently accustomed to and technology continues to advance at an incredible rate.

The Two Types of Cloning Techniques

Now there are two different types of cloning techniques, there is replication cloning and duplication cloning. Replication cloning is the type of cloning the public has generally heard of. Replication cloning involves taking the nucleus (the DNA) out of a donor egg, and replacing it with new DNA from the person to be cloned. After a few days the resulting embryo can be implanted for pregnancy. The newborn starts life off as a genetic copy of an original.

Duplication cloning, on the other hand, is the second hidden method and type of cloning where the DNA from the person is grown in a big thick tank full of (salty) water. It involves a method of regenerative technology where the cells are agitated and agitated and over the course of 5 months, a fully formed duplicate clone body of an original is developed. The Illuminati used to have to use a tissue sample from the original or the cells from women's pap smears because this contained rich cells to make duplicate clones; they also used children's foreskins which were discarded at the hospital that got removed. That is what happened to me. I had my foreskin removed at age 4 –and by age 5 the Illuminati grew duplicate clones of me. Now the Illuminati say they have upgraded the technology since 2000 and now all they need is blood from the original. They then agitate the blood cells over and over again through regenerative technology, until a fully formed duplicate clone body of an original is produced. For readers who are unfamiliar with regenerative procedures, see Dr Stephen Badylak's video on "How to grow a new fingertip" (Science Channel 2014) and Carmichael (2013) which discuses how scientists cloned a mouse from a blood sample (this is another example of drip-feed disclosure). Regenerative science and medicine (CBS 2008) does work; I've seen it many times.

The Different Grades of Clones: "Mark 1" to "Mark 4" Clones

There are a few different grades of clones that I know of. There are: "Mark 1" to "Mark 4" clones. Mark 1 clones were available at the end of World War II (1945). Mark 1 clones were rapid eye movement (REM) driven clones but they were primitive grades of cloning with lots of side effects caused to the original. Mark 1, REM driven clones were called "Organic Robotoids", even though there are no robotic parts to the clone at all. See Dr Peter Beter's discussion on "Organic Robotoids" (Beter 2011).

Donald Marshall's Dilemma with "Mark 2" Sleep Driven REM Clones

The problem I am having is with Mark 2 REM driven clones. "Mark 2" is a rapid eye movement (REM) driven clone. In other words, Mark 2 clones are sleep driven clones. REM sleep is the fifth stage of sleep (Sleepdex 2015). The first REM cycle usually happens 90 minutes to 110 minutes **after** we fall asleep (Sleepdex 2015). What that means is: **currently,** this is the Illuminati's main form of communication. They do not call people on the phone; they do not meet at the Bohemian Grove anymore (they only meet there once a year for traditional purposes). Since they discovered the science of sleep driven cloning, they meet at the cloning centre **WHEN THEY GO TO SLEEP** (The cloning centre is a physical location, located 5 / 6 hours radius from the Robert Pickton Farm (Port Coquitlam, British Columbia, Canada) at a nature reserve).

Consciousness Transfer Happens When the Original Reaches REM Sleep

The Illuminati can transfer the consciousness of an original once the person reaches REM sleep if there is a duplicate clone of the original at the cloning centre, once the person goes to sleep (90 minutes to 110 minutes after falling asleep). So what happens is that the consciousness of the original is transferred to the duplicated clone body (the duplicated clone body is grown within 5 months by regenerative technology) once the original reaches REM sleep. The original's consciousness is transferred from the original's body, although the original's body is still in their bed, asleep at home; to a duplicated REM driven clone body at the cloning centre and the original 'wakes up' as a cloned version of himself / herself at the cloning centre. It is a great marvel of science and one man's greatest achievements, but it is kept hidden and used for sinister purposes which I discuss below. I must address the readers at this point who are lost at the thought of duplication cloning, and consciousness transfer.

For readers who may have difficulty understanding the process of duplication cloning, see the video which features Dr Stephen Badylak (CBS 2008; Science Channel 2014). Badylak describes and illustrates how a new fingertip can be grown within 4 weeks, by the process of regenerative technology. Duplication cloning works a similar way; involving regenerative technology whereby the cells are agitated and agitated, and over the course of 5 months a fully formed duplicate clone body of an original is grown. Moreover, readers should also view the Horizon documentary "Why Do We Dream?" (2009); the documentary clearly explains the phases of sleep; particularly REM sleep. The documentary also clearly explains that during REM phase sleep the whole body shuts down, and it is only the brain which is active. The fact that the brain is still active, although the body is inactive, is a perfect opportunity for consciousness transfer; and this is usually the moment the Illuminati transfer the consciousness of a person who is asleep, to a cloned version of himself / herself. The reader should also explore Ehrsson's (2013) lecture on consciousness transfer. In the video, Ehrsson (2013) clearly explains how through the first person visual perspective (seeing through the eyes), and through synchronous (occurring at the same time) stimulation of a body part, individuals can perceive and see the world from another body different from their original. In other words, so long as people can see through the eyes, and the body is stimulated, the match between the visual perspective and feeling body sensations allows the individual to see and perceive the world through another body different from their original. The consciousness has been transferred to another body.

Ehrsson (2013) also demonstrates that consciousness is linked. Once consciousness is transferred (from the original's body to the new body), even though the new body is not the persons original body; when the new body is attacked, because the person perceives the world through the new body, everything feels "very real". When the new body (for which the person's consciousness has been transferred to) is attacked, the original still perceives the threat as "real", and therefore the person displays a biological and physiological fear response in their original body; the heart rate increases, as does respiration (breathing) rate, as well as anxiety.

Donald Marshall is an Original as he sits at Home Typing this Disclosure

This is how advanced technology is today. This is also exactly what happens to me. **I am an original as I sit at home here typing this disclosure,** but when I enter REM sleep, my entire original body shuts down, and only my brain is active. The Illuminati have linked my consciousness to a REM driven duplicate clone body of me. Therefore, as soon as I enter REM sleep, they are alerted to the fact that I have entered stage five of the sleep cycle, and in REM sleep, because they have a green and red light above duplicate clones at the cloning centre. The red light indicates that the original is awake, and the green light lets them know that original has entered REM phase sleep and it is time to transfer their consciousness to a duplicate clone version. They cannot transfer the consciousness of an original before REM phase or when the original is awake. When they try and transfer consciousness before an original is in REM phase it makes the original have intense headaches. Setting an alarm every 90 minutes to wake up and then go back to sleep also does not work in terms of avoiding REM sleep and having your consciousness transferred, because the body requires REM sleep, and therefore you become very tired, quickly, if you do not get REM sleep. If you stay awake for four days without REM sleep you begin to hallucinate; after seven days without REM sleep you will die.

When Donald Marshall's REM Driven Cloned Experiences Began

This whole sordid episode began when I was 5 years old, after going to the doctors to get my foreskin removed, at age 4. The tissues from my discarded foreskin were used to grow multiple duplicate clones of me. Within months, I was having my consciousness transferred to a REM driven, 5 year old clone version of me every time I entered REM sleep. The reason the Illuminati did this was because they wanted to use me as what they term "a diddle kid" – in other words they wanted to have sex with an REM driven, 5 year old, duplicate clone version of me. This is what they do. They clone people. They make duplicate clone bodies of people to victimise in terrible ways, and they clone children for men with undeveloped penises to have sex with.

However, I could never remember any of these REM driven clone experiences when I woke up for 25 years; because another aspect of REM driven cloning is that the Illuminati have the advantage of being able to suppress the experiences of people who have their consciousness transferred to their REM driven duplicate clone versions through advanced memory suppression technologies. For the reader who wishes to understand how memory suppression works see Winter (2014). Memories are stored and retrieved from different parts of the brain. Interactions between the cerebral cortex (the outer layer of the brain) and hippocampus need to work together in order to bring the memories out of mental storage to be re-experienced by the mind (Winter 2014). Therefore all that has to be done is for them to turn off the nerve cells which communicate with the cerebral cortex and hippocampus for the evening which you went to sleep and had your consciousness transferred to a REM driven duplicate clone version of you, and you will not remember anything. The technology is so advanced; they can, and do suppress memories at a push of a button. Therefore, the victim wakes up the next morning and will not remember a single REM driven duplicate clone experience. All that they will be aware of is that they did not have a dream the previous night.

This is what happened to me for over 25 years. I wasn't allowed to remember these REM driven duplicate clone experiences until I turned 30. My memories were suppressed for 25 years. I would wake up many days, after a night's sleep, thinking 'I didn't have a dream last night'; or I would wake up the next morning and I would be sick, although I went to sleep the previous night feeling completely healthy. Throughout the years, I thought I had terminal illness; because whenever I went to see a doctor, the diagnosis would always come back that 'there was nothing wrong with me'.

However, over the course of 25 years when I was memory suppressed, and not allowed to remember my REM sleep driven clone experiences, the Illuminati were using me to produce songs, lots and lots of songs, and whenever I could not produce a song, or did not want to, they would stab my REM driven clone body; sodomize my REM driven clone body; chain me to a crucifix and burn my REM driven clone and torture me for all kinds different reasons; and even when I was compliant and did what they told me to, I was attacked, simply because some members of the Illuminati wanted to know what it feels like to stab someone. Now because consciousness is linked (Ehrsson 2013), all these REM clone torture experiences affected me in my original body, and I have a weak heart today because of it, at the age of 39. The memories of my REM clone driven experiences also had to be released to me slowly (see Alford (2014) and FW: Thinking (2014) –for an example of how memories can be wiped and restored) over the course of many months because I had been through so much horror and torture that if the memories was released all at once, I would have had a heart attack in my real body and died.

This is one of the main reasons the Illuminati currently use REM sleep driven cloning (besides sex and torture): to plagiarise talented people under duress. There are many people including me who have been, and are plagiarised on a daily basis. Their consciousness is transferred to their REM driven clone alternates, and night after night and under duress they are forced to reveal their money making ideas or face torture or clone death; clone death after clone death. That is what the movie *Inception* (2010) is about: REM sleep driven cloning, and stealing ideas from the minds of unsuspecting individuals. The Illuminati actually made the film *Inception* (2010) as well as *The Island* (2005), *Avatar* (2009) and *The 6th Day* (2000); the films mentioned above contain references of cloning and REM driven cloning to show off the Illuminati's 'power' and throw in the world's face and laugh.

The Science of REM Driven Cloning Began in 1945

I am one of the few people fortunate, or unfortunate depending on how you look at it, who remembers all of my REM driven cloning experiences because these memories were released to me at the age of 30. I remember many people who have attended and still attend the cloning centre which they bring me to over the course of 34 years. I am currently 39 years old and was brought there since the age of 5. The science of REM sleep driven cloning was first discovered in 1945, and at first it was just a political thing which leaders of the world did. Through the science of REM sleep driven cloning, political figures, heads of state and royalty met each other as REM driven clones versions of their selves when they went to sleep and discussed worldly affairs in complete secret with each other. They also did whatever they wanted with each other as REM driven clones, which they could not do in their original bodies such as sex, fighting each other to the death, jumping off cliffs as their REM driven clones, and just about anything you can think of when one has the ability of pseudo-immortality as a sleep driven clone version of themselves hidden from the guise of the world and general public.

Well they soon got bored of each other very quickly, and started to make REM driven duplicate clones of movie stars, musicians, celebrities and public figures from all walks of life to hang with as REM driven clones and now they all get together for a disgusting time as REM driven clones when they go to sleep. Most of the G20 political leaders meet at the cloning centre as REM driven clones when they go to sleep to discuss worldly matters and watch gruesome things done to innocent and unsuspecting civilians, who are also REM sleep driven clones at the cloning centre, but civilians have their memories suppressed. This is what the films *Hostel* (2006); *Hostel: Part II* (2007); and *Hostel: Part III* (2011) (Illuminati made films) -are about: -rich people paying to torture and kill innocent people at a secluded location for sport; just that at the cloning centre, the Illuminati members torture and kill REM driven clones of innocent people.

Therefore, depending on what was done to your REM driven clone, you could wake up the next day with a very intense headache in your **original body**, if your REM clone was repeatedly punched in the head; an upset stomach in your **original body** if your internal organs were tortured; achy limbs; and worst of all, feel sick all over; although the previous evening you went to sleep feeling healthy. Remember consciousness is linked (Ehrsson 2013), therefore whatever happens to your REM driven clone will affect you in your original body. Even when they suppress your memory; the previous night's experiences still affects you in your original body. You wake up with the knowledge that you did not dream the previous evening when you went to sleep, and with certain illnesses depending on what was done to your REM driven clone.

Help Donald Marshall STOP REM Driven Cloning NOW!

Everything I discuss in this disclosure is firsthand witness accounts. I have seen and experienced horror spectacles done to innocent civilians and to me. For the sake of humanity and the innocent children brought to the cloning centres all over the world; I must tell the whole world about this and the good people of this world **MUST HELP ME TO BRING THIS TO A STOP.** My sense of duty, moral correctness, and conscience cannot allow me to stay silent about this evil. REM sleep driven cloning, and the torture of innocent civilians as REM driven clones; having REM driven clone sex with beautiful, innocent unsuspecting civilians of the world; the paedophilia of children through the process of REM driven cloning; REM driven sex slavery; the plagiarism of talented individuals through the process of REM driven cloning; and REM driven idea slavery, whereby talented individuals must produce new ideas and concepts daily (or whenever they sleep) or be stabbed or tortured in their sleep as REM clones –is the most vicious and pernicious form of tyranny ever to occur against humanity and **IT MUST BE STOPPED NOW!**

Do Not Panic, Riot, or Cause Chaos. This is Very Important.

Now I must preface this carefully, because members of the Illuminati have told me that I have to put this in an eloquent fashion, in a way which does not make people panic because people finding out about REM driven cloning, and the extent of the evil it has been used for, and continues to be used for; it could cause loss of social order, riots and anarchy in the streets. Moreover, I too **DO NOT** want riots and anarchy in the streets, despite the fact that I am vehemently angry considering the extent to which they plagiarised my talents over the years. If you are a good person reading this, and you want to help and you want social order restored for the benefit of mankind, promise yourself; me; all the innocent children they have affected through REM driven cloning, and the children of the future that you will **NOT** riot, and destroy a world for children who are going to inherit the world. **This is very important**. Remember the Illuminati have highly advanced technologies (kept hidden and secret), including weaponry, and they are just waiting for any excuse to use it on the populace. In all revolutions, the populace always win, and I want this disclosure and the end of REM driven sleep cloning to go smoothly. I want the good people of the world to keep spreading this information, keep spreading this disclosure all over social media, tell your close friends, your family, and as many people on social media platforms as you can. The internet, and radio shows with small audiences are the only forms of communication the Illuminati do not control. They own all forms of television networks, even including the Aboriginal People's Television Network (APTN), and therefore a message as important as this will never reach the average man or woman through television because it does not serve their agenda.

Spread this Disclosure Document until it reaches the Armed Forces

I want this message to reach the armed forces, because currently people in the armed forces are following orders, yet they do not know how corrupt their governments and people in high profile places are. They are unknowingly defending these corrupt people. People in the armed forces will **NOT** defend these corrupt people nor will they harm civilians when they realise that these corrupt people have been growing duplicate clone bodies of civilians, transferring civilians consciousness to their duplicate clone versions when the victim reaches REM sleep, and torturing civilians in their sleep; having sex with under-aged REM driven cloned children; having sex with innocent and unsuspecting adults as REM driven sleep clones, against their will; torturing REM driven clones for sport; and for money-making ideas to benefit their own pockets while they make innocent civilians sick and have side effects in their original bodies from REM sleep driven cloning technology. No. The armed forces will not accept that. Therefore keep spreading and sharing this information for everyone to see so that it reaches personnel in the armed forces **ASAP!** I want the armed forces to overthrow these corrupt people. The Illuminati think they are very sly and untouchable because together they hold most of the wealth in the world. We must show them that they are not untouchable. Please do all you can, to share and spread this message. It is the most important message to ever reach the internet because it affects all our liberties.

I must also address the readers who may feel afraid in spreading my message because they may want to do the right thing, but they are terrified the Illuminati may degrade or end their lives. As unbelievable as the next thing I am going to share sounds, it is a belief system which shapes the reality of the Illuminati and therefore they themselves are afraid to do anything which may degrade their lives or 'eternal soul'.

They believe that they are the 'fornicators' mentioned in The Bible, and if they were to degrade or harm the lives of anyone aiding me, they would suffer the wrath of God when they die. Notice I said **when they die**, but so long as they are still alive they can do all the evil they like, with little consequence, as they do in their REM driven clone versions at the cloning centre; worshipping Lucifer and doing all sorts of ungodly things. They even say God does not exist, only science and technology; yet they are afraid of dying. Very afraid of dying; because they believe they'll meet God's judgement. I'm just relaying what they have told me. They are weird like that. They also follow Hopi Indian Prophecy, Mayan Prophecy and Nostradamus Prophecy. They mix and match those three prophecies and come up with their own religion of what may or may not happen in the future.

Donald Marshall is NOT "The New Age Saviour". Nobody wants to be a 'saviour'

Furthermore, **THEY** believe I am a new age saviour for the end times; that I was going to save the world from something, according to the Nostradamus prophecy (quatrains); the Illuminati consider Nostradamus the greatest prophet ever to walk the earth. They are in awe of him.

I **DO NOT** endorse their claims. I just want REM driven cloning to stop and these people face punishments for their crimes against humanity. I really do not endorse such claims, but if it saves my life and yours, so be it. In reality and practicality, nobody wants to be 'the saviour'; not with such tyranny and pernicious evil such as this. We'd rather all just be living normal lives without the knowledge and practice of REM driven sleep cloning which is used for torture, clone death spectacles and sex with children. My friends on Facebook do not want to be heroes or 'saviours' either; nobody really does, but we do it out of a sense of duty, moral correctness and our conscience would never rest knowing the true extent of evil in this world, and doing nothing to end it. So all that saviour and Nostradamus prophecy talk, it is just the Illuminati; they like to quote prophecies, and make them come true to make themselves feel special.

Like I said, they are weird. I'm just relaying what they say, and I **do not** endorse the new age saviour claims. I'm just a normal guy, who was unfortunate to be cloned when he had his foreskin removed at age 4. This could have happened to anyone. Thankfully I had a very active imagination at the age of 5 and with the use of the technologies such as Mind-voice technology (that is what the Illuminati call it) it allowed me to make songs to keep the perverts off me. Mind-voice technology is just a technology which can tune into the inner voice in your head (See articles by Prigg (2014) and New Scientist (2014) –which discuss how scientist can listen to your inner voice; –it is similar to how Mind-voice technology works). Therefore, when you hear the sound of a guitar; the sound of drums; or any instrument for that matter; after hearing the sound, when you imagine the sound you heard, your REM driven clone can replicate this sound with Mind-voice technology **exactly**; and I used Mind-voice technology to make many, many songs over the years. As I've said, and I'll say it again: technology today is very advanced.

How the Nostradamus Prophecies shapes the lives of the Illuminati

Also another reason they have not killed me is because the Illuminati said "If they do, everyone is going to know I am telling the truth", -which I am with my right hand to God and I cannot lie about this stuff; the things I have seen are too sinister and diabolical. It would make me as bad as them if I did. Furthermore, because they follow the Nostradamus prophecies religiously, they have told me that they believe anyone helping me is considered part of "The Army of Light" and again, to hurt or degrade anyone helping me, considered part of The Army of Light would incur the wrath of God; and that to hurt anyone considered part of The Army of Light will degrade their lives; they will suffer misfortune and entire ruin as will their eternal soul. These are just some of the interpretations of the Nostradamus quatrains they have relayed to me. These people are beyond crazy, zealot and religious fanatics. Nevertheless, the main point to remember by all this is that **you are safe**, and **they cannot hurt you**, **they are scared to hurt you**, as well as kill me because prophecies have shaped their lives and the lives of their ancestors for hundreds of years.

Furthermore, and practically, I have told too many people that this is the extent, to which technology has developed in our current day, and it is kept secret and hidden from the public for nefarious purposes and this is what the Illuminati do: REM sleep driven cloning. I have been spreading this information **since 2011** and not one of these high profile people, government officials or celebrity figures has issued a (public) statement against me or taken me to court over libel charges because I'm telling the truth; and their statements would not hold in court.

I have also told over a million Arabs; I have appeared on radio interviews with Vinny Eastwood (Vincent Eastwood 2013); Jeanice Barcelo (Jeanice Barcelo 2013); and Lisa Phillips on the Cry Freedom Radio show (Astral 7ight 2013a – 2013h). I have reached audiences as many as 280,000 views on the Vinny Eastwood Show (Vincent Eastwood 2013) alone. However, I must stress that YouTube reduces the view count of my interviews every so often in order to suppress the truth so that my interviews do not get featured on their platform. Nevertheless, there is safety in numbers and rest assured there are new people waking up to the disclosure of REM driven sleep cloning every day; and therefore you can share and spread this information **without any worry** about facing reprisals from the Illuminati, even if you do not believe that it is in fact their beliefs in Nostradamus prophecies which is saving you and I from the Illuminati.

HIV/AIDS has been cured; so have many forms of cancer

Now is also a good time to share with the reader that this situation is not all 'doom and gloom.' In their quest to perfect technologies, science and medicine, the Illuminati discovered the cure for HIV/AIDS; cancer (except pancreatic cancer); Alzheimer's; Dementia and many more debilitating diseases which humanity suffers with. They relayed the fact that they found the cure for HIV/AIDS by stimulating the cells in an oxygenated rich environment because diseases/viruses cannot thrive in an oxygenated rich environment (in layman's terms). Magic Johnson was cured from AIDS this way and they use him as a spokesperson to sell the retrovirus pills which they know is not as effective. However, the greed these people have, knows no bounds, and because they receive too many donations for HIV/AIDS as well as cancer donations they do not announce they have cured HIV/AIDS or cancer publicly. As I've told you, and will continue to tell you: technology is far, far, FAR advanced than what you currently see around you. With the help of the good people of earth I promise to release these technologies for the benefit of mankind, and all reading this disclosure can hold me to this. You **must** hold me to this. It is time we ended corrupt governments and corrupt individuals in high positions of power in this world. Hold me to this.

I hope you are now beginning to understand why this is a **world emergency** and why these corrupt people must be ousted and overthrown. I also hope you understand why you must not riot or cause chaos because as I said at the beginning of this disclosure "I write this to empower good people of the world against the tyranny which exists all around us in our world today"; I **did NOT** write this to cause chaos, public dissension or anarchy. Good people of the world will inherit the earth once these corrupt people are overthrown. Therefore, under no circumstance must you destroy the world when you are going to inherit it. No matter how angry this disclosure makes you; no matter how angry the diabolical people who commit these crimes against humanity make you feel. Let's make sure this goes smoothly. Keep spreading and sharing this information, until this disclosure reaches the armed forces; until the armed forces bring these corrupt people to their knees. We are about to inherit the earth; by all means let's ensure everything goes smoothly. Now that you understand how important this disclosure is and that the overthrow of these corrupt individuals must progress smoothly, I'll continue.

The Behaviour of High Profile People at the Cloning Centre

The high profile individuals who attend these REM driven clone gatherings in their sleep have nothing better to do than to show off in disgusting ways. They have no shame and it seems nothing embarrasses them. Some REM sleep driven, cloning centre attendees, sit in the stands of an unused arena, smaller than a hockey rink, but it still has the capacity to seat between approximately 300 to 400 people. There is dirt in the centre of the rink where ice would be. They have frightened REM driven clones of children walk into the middle of the dirt rink to be victimised for a bizarre and disgusting spectacle. Sometimes, they have animals like dogs have sex with the REM driven clone children, while a man holds the dogs on leash so that it wouldn't bite the child on the back of the neck; which I have seen happen before. They all try to outdo each other in their levels of depravity; to be evil is to be "cool" to them.

The Ring Leaders of the Cloning Centre

I will now discuss the ring leaders, but before I name names, please understand that I am not someone who has collated stories and evidence together over the internet. I REALLY **do not** need to do this. Every time I enter REM sleep, the ring leaders at the cloning centre transfer my consciousness to a duplicate clone version of me. I have seen and talked with these monsters over the course of 34 years as REM driven clones at the cloning centre. However, I do understand that for the average reader, without understanding how far medicine, science and technology has advanced; without seeing or hearing people, other than me, discuss the crimes the ring leaders of the cloning centres commit, it is very difficult to imagine that this is what these high profile people are involved with, and do. Nevertheless, I guarantee it; this is the true extent and prevalence of evil in our world.

Their public personas, and public images is well managed by their PR teams, and therefore for the average person it is difficult to imagine these people in any other way but "positive"; but these people behind closed doors are the most evil, sick and twisted people I have ever had the misfortune of encountering. It is also difficult to take in what I am saying because I am not a public figure, nor do I have the "influence" or public image these people have; the general populace have never met, or heard of me, or known who I am; if this was the case my eye witness testimony alone will be enough; however, I do understand that we are dealing with highly advanced technologies, thousands of years advanced compared to what the general public is accustomed to, and because of this I provide references for the reader; I provide references for the reader to have something to corroborate what I am disclosing; and I provide references for the reader to see past the illusions these people have created.

Moreover, as "Consciousness Science Kept Hidden" demonstrates (YesEthan 2013):
remember that it is the heart and not the brain which is the most powerful centre in the body, the heart is the most intelligent; it has the ability to receive precognition and intuition about a future event, **before it actually occurs**, therefore if you find that your heart is in resonance with the names mentioned; do not fight it with the ego of the mind; when our egos get in the way it overrides our heart's intuition; your heart is capable of discerning the truth of an event before your brain becomes conscious of it; by all means listen to your heart, and allow your brain to naturally develop its intuition. It will be the coherence between your heart first, and then your mind centre which will allow you to access your intuition.

The references I provide for the reader to corroborate what I am about to say includes: "Appeal from Survivors of Canadian Genocide" (inifiniLor 2013); Tila Tequila –"Missing Children and Cloning Centres" (Astral 7ight 2013i); "Royal Babylon by Heathcote Williams" (MrCowshedder 2012); and "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014).

Watch all the above videos and you will easily see all that I disclose. In Appeal from Survivors of Canadian Genocide interview (inifiniLor 2013): Listen to the indigenous people Stee-mas and Wahtsek speak about the pain and torment they have suffered under the Crown, the Vatican, the government, and churches of Canada. 50,000 to 150,000 indigenous children have gone missing because of the involvement of the Crown, the Vatican, the government, and the churches of Canada; and all these factions involved know this!

Tila Tequila is a television personality who has attended the cloning centres as a REM driven clone since childhood. In Tila Tequila's radio phone in (Astral 7ight 2013i), the audio is not very clear for the majority of her disclosure; she was also frightened and therefore she sounds most animated; however her words are clear at the following points of the audio (I also provide a transcript in the appendices section: "Appendix A"):

4.00 min: Cloning Centres and they take your children...
4.25 min: –There are these CLONING CENTRES –I'm not even talking about child molestation here...
5.40 min –They take your children –they not only molest them; men f*** them; and make them shoot each other...
7.59 min- There is a point where you cannot just turn the other way, you know this stuff is going on and you go 'oh well, you know... that's their problem; let's just turn the other cheek. How long are you going to turn the other cheek until it happens to your own freaking children?

10.34 min - Do you know why they love children? Because they are innocent souls; they are innocent... they are the most innocent, pure beings in this planet. They are not harmed by anything. They are new to the world; bright-eyed, pure innocent children. That is why these disgusting paedophile... and these clone rings, cloning centres, satanic rituals, Brownsville Texas... there are many of the of the cloning centres where they take your children that go missing.

Royal Babylon by Heathcote Williams (MrCowshedder 2012) addresses the fact that for Brits, despite their inordinate pride about their tradition, it is often revealed they know little about it than anyone. The documentary then goes on to address the fact that Prince Phillip the Duke of Edinburgh, has a lust for blood-sports which includes shooting endangered wildlife and mammals which he considers 'culling'. The documentary is also notable for demonstrating to the viewer that Queen Elizabeth II redistributed as much as 2.1 million for her private income from 276 people from Merseyside and Lancashire. These people had no will, so their property was grabbed. The same is true for the Duchy of Cornwall. If you die without making a will, everything you possess will go to Prince Charles. Furthermore, Queen Elizabeth II is the largest land owner on earth. Queen Elizabeth II: head of state of the United Kingdom, and 31 other states and territories is the legal owner of about 6600 million acres of land; one six of the earth's non ocean surface. She is the only person on earth who owns whole countries, and who owns countries that are not her own domestic territory. This land ownership is separate from her role as head of state; is different from other monarchies where no such claim is made –Norway, Belgium, Denmark etc. The value of her land holding is £17,600,000,000,000 (17.6 trillion Pounds) (approximately)).

In "THIS MOVIE WILL BLOW YOUR F%SNG MIND" (KafkaWinstonWorld 2014), readers with a deeper knowledge of the Illuminati and the agenda of the Illuminati will easily spot the minor inaccuracies in the documentary. However, the documentary provides a good reference and foundation for anyone new to the topic of the Illuminati and their agendas for the world. The documentary demonstrates how the world is currently a deception within a deception coated in reverse psychology. It gives widespread insight into the Royal family of England, and their lineage.

Furthermore, the documentary provides information that the world's most popular religions have been infiltrated with Illuminati dogma, in secret. The documentary details how (some) modern Christians, Jews, and Muslims do not (yet) recognise their common enemy because that enemy is invisible (until now). The enemy hides within their own religions. The enemy also hides within the activist community (anti-Illuminati movements), it finances big budget documentaries and movements. The end game is to sell a divisive (creating strong disagreement), anti-capitalistic, atheistic message and deliver trusting followers right into the lap of a communist (one government) New World Order. The enemy is the Illuminati.

The documentary also details that: "Today, the bloodlines of this unholy alliance are on the final stages of establishing one world religion, one world government, and one world ruler [--that is **IF**, the good people of this world do nothing to stop them]. The Windsor Royals are at the control of this web of deceit. It is a known fact that Adam Weishaupt (the founder of the Illuminati) took refuge and got assistance from the German Saxe-Coburg Gotha Royal family. The Saxe-Coburg Gotha's changed their German name to Windsor (in 1917). The Windsor's public image is a façade, to hide the cesspool from which they operate behind the scenes; their power and control cascades out of Buckingham Palace to the elite bloodlines who run world affairs."

For anyone who thinks they know they Royal family of England, did you know that still to this day: they have a person appointed called “Groom of the Stool”? What is a “Groom of the stool”, you may ask? A Groom of the stool is responsible for stools. The kind you sit on and the kind you expel. In other words a Groom of the stool wipes Royal bottoms. The above has been confirmed to me by Prince Charles of Wales himself as REM driven clones at the cloning centre, and I quote: “It gives you the feeling of being completely taken care of” when I asked Prince Charles (as a REM driven clone) why they still have that. But don’t just take my word for this; research “Groom of the stool”.

The Royal Family of England

As most of readers have already guessed; and as unbelievable as it may currently sound to some readers, the ring leaders of these cloning centres includes the royal family of England. As I’ve said REM driven cloning started as far back as 1945, and it was just a political thing, but Queen Elizabeth II has been involved with REM driven cloning since its inception! Yes! Queen Elizabeth II; Phillip, Duke of Edinburgh and Prince Charles of Wales are the worst of them. They are unbelievable depraved perverts all showing off to the celebrities who attend these REM driven clone gatherings. Elizabeth has the REM sleep driven cloned children at the cloning centre call her “Lilli-bet” and Elizabeth as a REM driven clone herself does ungodly things to these REM driven cloned children. Some she fakes being nice to; some she is terrible to; and as a REM sleep driven clone Elizabeth cuts these REM driven clone children with swords while they scream. The decent people at these cloning centres who have had their consciousness transferred into their duplicate REM driven clones, and are there against their will, like I am; they are terrified to say anything against Elizabeth and her cronies. Most of these decent people also have their children’s consciousness transferred into their REM driven duplicate clones, and Elizabeth and her cronies hold these decent peoples’ children as hostage; to be torn apart; clone death after clone death if they have the slightest inkling of informing anyone, but as Elizabeth and her cohort have been torturing me as REM driven sleep clones anyway, I will disclose all that they do.

Vladimir Putin is also there, as an REM driven clone, and he loves to put the fear of torture and death into people, but he is essentially a perverted coward himself. Most of the famous people who attended the cloning centre as REM driven clones are ashamed to speak or be seen by me there very much; because they are ashamed of the perverse and disgusting gatherings. I am a decent person, even when my consciousness is transferred to my REM duplicate clone alternative, and I will not participate in these depraved acts; so they use me as an example and torture me for being a good person. They have drugged my REM driven clone, and as I’ve mentioned because consciousness transfer is real (Ehrsson 2013) and because consciousness is linked, therefore everything feels “real as real”; when I’m drugged as a REM driven clone, I behave exactly as any person would on hard drugs (although I have never taken any drugs in my original body); my vision is blurry and my balance affected; they have used this as method of having me do disgusting things in a blurred state, videotaping these dirty acts and editing a highlight reel which also depicts me doing disgusting things to show to the world when the world learns of their depraved natures and their public personas are shattered so that I can be despised as much as they will be when the entire world learns about this. Furthermore, because of the microchip inserted in my REM driven clone; they mind controlled me like a “Manchurian Candidate” (*The Manchurian Candidate* 2004) using MK Ultra technology, which is far more advanced than the declassified MK Ultra techniques (MK Ultra Compendium 1980) mentioned online today which references MK Ultra techniques in the 1970s. See Rense (2001) for a detailed explanation of how implanted microchips work, and how these chips allow the programmer control victims. C.S.I.S Canadian intelligence are all involved; they have REM driven clones too; including a lot of Commissionaires and Canadian Prime Minister Steven Harper himself; They do what Elizabeth says, and seem to follow her every whim without question. It doesn't matter who you are; if you have a cute child, or a talented child with imagination, creativity and moneymaking potential (such as business smarts; song making ability; literary / oratory ability etc. –you name it so long as your child has moneymaking value to them) they will send for the blood record of your child; agitate the blood cells over and over until a duplicate clone of your child is grown; they will then transfer your child's duplicate clone body from the tank they were grown onto a steel rack; the steel racks contain 5 rows and each row holds a duplicate cloned body; once your child goes to sleep in his or her original body and once they enter REM sleep (90 minutes to 110 minutes after they first fall asleep) the Illuminati are alerted to this fact by a green light above your child's duplicate clone at the cloning centre and they then transfer your child's consciousness to the REM driven duplicate clone; this is the point where the prime evils molest your children; and so your children are being molested in their sleep as REM driven clones and then they suppress the child's memory or implant false memories (see Kim (2013) and Alford (2015) –on how false memories can be implanted) so your children wake up not remembering dreams from the previous night, or if they do remember anything, the memory is false. These depraved acts very much stunt children's development; it causes them to have learning disabilities; unexplained depression; loneliness; suicidal thoughts, and all kinds of side effects.

The cloning centre is a paedophiles paradise and **it must be stopped**! Elizabeth secretly owes a few music companies too (Universal and others) with bands under contract; she forces me under a knife to compose music for them; if I can't she will stab my REM driven clone; burn my REM driven clone and have her pervert thugs smash my REM driven clone body there. As Ehrsson (2013) demonstrated, because consciousness is linked; and consciousness transfer leads the person who has had their consciousness transferred to still perceive reality as "real as real" in the new body; I have the same biological and physiological responses as I would have if these attacks were occurring to my original body. This has caused me to have heart damage and severe debilitating headaches. Because my memories were released to me; I can now also remember all my past REM driven clone memories since age 5; as well as the previous night's REM driven cloned memories (when I have my consciousness transferred to the cloning centre) which causes severe stress to my mental psyche and mental well being.

As unbelievable as the above statement sounds regarding Elizabeth's ownership of music companies, just go and listen to the song "Abadon" performed by Boondox (RainmanJhof 2011) on YouTube. There are two "Abadon" versions; the version you seek can be found in the Reference section uploaded by RainmanJhof (2011); skip to 2:46 min (or listen to the whole song; pay particular attention at 2:46 min) and you will hear loud and clearly, Elizabeth say "Drop the mother f****ing base". That is Elizabeth's voice; you can compare it to any of her Christmas speeches. That is her. Elizabeth said the above as a REM driven clone at the cloning centre. I know. I was there. They got so confident, that nobody would ever find out about their REM driven cloning shenanigans, that they even had me sing on a song I produced: "Where'd You Go" –Fort Minor (Murdok Dubstep Remix) (MurdokDubstep 2010), available on YouTube. Tila Tequila and I sing on this remixed track as REM driven clones (MurdokDubstep 2010); I start to sing from 1:44 min onwards. You can compare the voice you hear in the song with the Vinny Eastwood interview (Vincent Eastwood 2013). That is me, in both cases, and that song was recorded at the cloning centre.

One good person at the cloning centre; Bernie Mac (also as an REM clone version of himself); who was an actor and comedian, tried to stand up for me and speak up for me saying ‘Cloning and torturing people here, it’s not right to do to a human being...’ and “This place is the devil”’... Bernie Mac also tried to have others join him in speaking out against the REM driven clone torture zone while he was having his consciousness transferred to his REM clone driven version. Elizabeth and Philip tortured him for 4 to 5 hours continuously at the cloning centre as an example of their ‘power’, Bernie Mac had an aneurysm (brain bleed) the next day and died. This is one of the aspects about REM driven clone torture: if you torture a REM driven clone continuously for hours on end, you will cause the person to die in their original body from a heart attack or aneurysm (while they sleep); even if they do manage to wake up the next day; because of the constant pain their consciousness has recorded; and because consciousness is linked they will wake up only to have a heart attack or aneurysm in their original bodies the following day. If you apply a constantly electrical current to a REM driven clone, that person will either have a heart attack or die from aneurysm in their original bodies while they sleep.

This form of murder is similar to the CIA’s heart attack gun (Non Mirage Truth Vision 2015); which was declassified in the 1970s and was a weapon used for causing heart attacks in victims, undetectable as the cause of heart attack under autopsy. The difference with REM driven clone death as a form of weaponry is that hardly anyone has heard of death by REM driven clone torture. Moreover, the perpetrator can be thousands of miles away from the victim; the culprit just needs a duplicate REM driven clone of the victim to torture or apply a constant electrical current to; this is exactly what is going on in our world every day. When Elizabeth and Philip don’t like someone and they want to get rid of that person, they clone that person; torture his or her REM driven clone till that person has a heart attack or aneurysm in their original body and dies.

This is what Elizabeth and Duke Phillip did to the indigenous people in Canada. The video reference (inifiniLor 2013) provided above, regarding the genocide of the indigenous people of Canada discusses this very topic (inifiniLor 2013). As Elizabeth and Duke Phillip visited Canada as clones, when they killed the native children; they both had solid alibis of being in England at the time. They then cloned all the witnesses of the Canadian missing children’s case and had them remotely killed by applying constant electric currents to their REM driven clones; one by one and each witness either died of a heart attack or aneurysm. They just dispatched the last witness by also applying a constant electric current to the persons REM driven clone because the missing children’s case was building momentum.

After Bernie Mac’s death everyone is afraid to speak up; they all became afraid of suffering the same fate. All it will take is a few lie detector tests from an independent source; and I will volunteer for them; they cover up for themselves so carefully that lie detector tests from an independent source unconnected to them in anyway will be the only viable method. I know many know non-famous people who have also attended these clone centre gatherings as their REM driven clone duplicate versions and also remember the clone gathering experiences in their original bodies (because they were not memory suppressed) that could be given lie detector tests to prove this. When it comes to the sexual exploitation of a child, lie detectors are admissible in Canadian courts. I need decent people with integrity to speak about this filthy business; to testify about who is involved. There is much to tell about the Illuminati and I am a wealth of information about this topic.

Elizabeth also had Princess Dianna Spencer killed by having someone shine the brightest light known to mankind through the window side on her car; they served away to avoid the light and hit the divider; it was no paparazzi. I first wrote about this in my original disclosure in **2011**; this was corroborated by an article 2 years **AFTER** I had made my original disclosure. See Radar Online (2013) for the corroborating article. Furthermore, "Candle in the wind" performed by Elton John (SadSongs4You 2010) was not Dianna's 'death song'; it was "Bigger than us" performed by "White Lies" (WhiteLiesVevo 2010). The Illuminati released this song, deliberately, 13 years after Dianna's death, again as a show of their 'power'. I did not want to compose this song but Elizabeth and her thugs threatened to stab and burn my REM driven clone repeatedly if I did not. Listen to the lyrics. I'll provide some of the lyrics here for those of you who are feeling too shocked by these revelations:

White Lies Lyrics "Bigger than us"

You took the tunnel route home.
You've never taken that way with me before.
Did you feel a need for change?
Apologies on your fingernails,
Love flickered in the city of lights
Like internet and radio waves.

I don't need your tears
I don't want your love
I've just got to get home

And I feel like I'm breaking up
But I wanted to stay.
Headlights on the hillside
Don't take me this way.
I don't want you to hold me
I want you to pray,
'cause it's bigger than us.

Are you beginning to see why these people must be stopped and stopped immediately? Are you beginning to see why it is no longer okay to 'turn a blind eye' or 'bury your head in the sand', and hope it all goes away? Can you see why I need the good people of this world to keep sharing this disclosure (online); to tell your friends and family until this information reaches the armed forces so that the armed forces can bring these people to their knees. I also do not want innocent civilians to get hurt despite the monstrosities these people cause to humanity on a daily basis. I want the overthrow to progress smoothly so that the good people of the earth can inherit the world, with all technologies and medical advancements which benefit mankind, made available to all. This is why I maintain my composure and resolve, in spite of the evil I see on a daily basis, and you the reader must also maintain your composure and resolve in spite of the evil which I am disclosing, so that we do not inherit a torn down world after we have completely overthrown these malevolent, sinister and corrupt hypocrites!

Furthermore, Diana survived the car crash; and while, either on her way to hospital or in hospital, Dianna was injected with a high concentration of salt, which is near undetectable in an autopsy and she died. Elizabeth did this because Diana was going to marry an Arab named Dodi Fayed. The Illuminati hate Arabs because for the most part, Arabs have not broken their relationship with God. Dianna hated her association with these people; Dianna knew about REM driven, sleep cloning and wanted nothing to do with it; but she was afraid of being killed and so said nothing. I know a lot about this and many more deplorable things the Illuminati have done over the course of 34 years and I will be more than happy to reveal all the evil they have done; as well as all the wonderful discoveries and progresses they have made in medicine, science and technology, which they have withheld from the general public. I am very happy to disclose all, because I known once the world learns of all they have done, we truly will begin to see the end of such evil on this planet. Just make sure you keep spreading this disclosure and it reaches the armed forces soon. Soon, soon, soon!

Help Donald Marshall by Sharing this Disclosure with Your Close Friends and Family

If you are reading this, then you have to help me. I know some of you may be reading this thinking, but I am just an ordinary man or woman Donald, I know you are telling the truth; I can feel it in my heart and my head that what you say is true and that you are telling the truth (which I promise you, with my right hand to God I am) but I'm just an ordinary man / woman Donald; I don't have any power; I don't have any influence; or because of the position I am currently in I can't speak out against these people; and I am just an ordinary man / woman I cannot be expected to take on the world's problems. You are right if you are thinking and feeling this, and I do not expect or want you to take on the world's problems. All I want you to do is share this information with close friends and family initially, because when you do that, at least that is one more person who will know about the true extent of evil in this world. When your close friend or family member also shares my disclosure with another close friend, the disclosure will eventually spread exponentially. It will also be one more person who has your well being in their thoughts because I understand how scary it is to know this. Therefore remember **there is safety in numbers;** and the more people you tell in your social circle, the more you can all look out for each other. So feel free to share my disclosure with close friends and family. Do not talk about my disclosure in public (for now) if it makes you feel uncomfortable. Moreover, if you want to reach more people (other than your initial social circle) the best way, is to do it online: through social media, such as Facebook, Twitter, and YouTube etc. As I've already mentioned, the Illuminati controls all forms of media except the internet.

The Illuminati, being so 'confident' at the time with the evil they wield onto the world, put my face on a Megadeth album cover titled *The World Needs a Hero* (2001) (to mock me) when I was 23. I posed for this album cover at the cloning centre when I was 23, and the album was released with a picture of my face, as a REM driven clone, 3 years later when I was 26. They did this because they were extremely confident that nothing would ever be proven. Google image it, although I am 39 now, you can still tell that it is me. The image shows what I end up looking like after an REM driven clone gathering with a skeleton popping out from my chest. I understand most of you are going to continue reading because although it is "liberating", it is also "paralysing" at the same time (Laurence Mountford) to realise what the truth of the world is, so I will display two images below, but promise yourself, that you will Google search this image in your spare time.

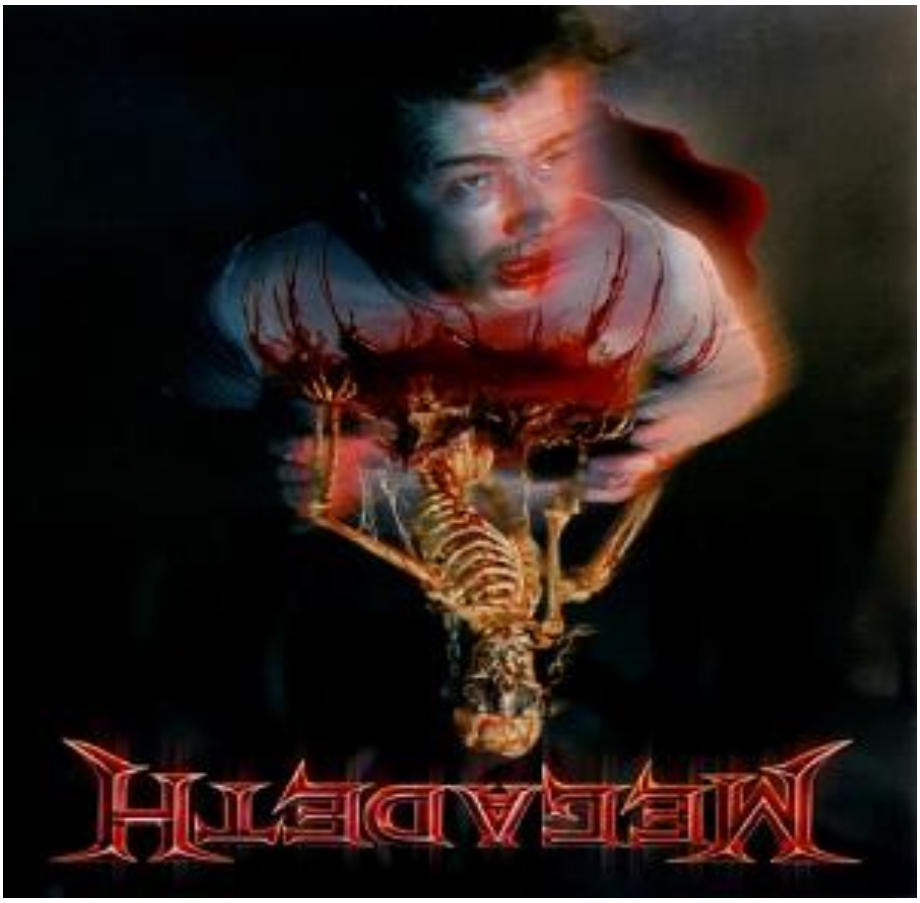

Image 3: Megadeth Album Cover: *The World Needs a Hero* (2001)

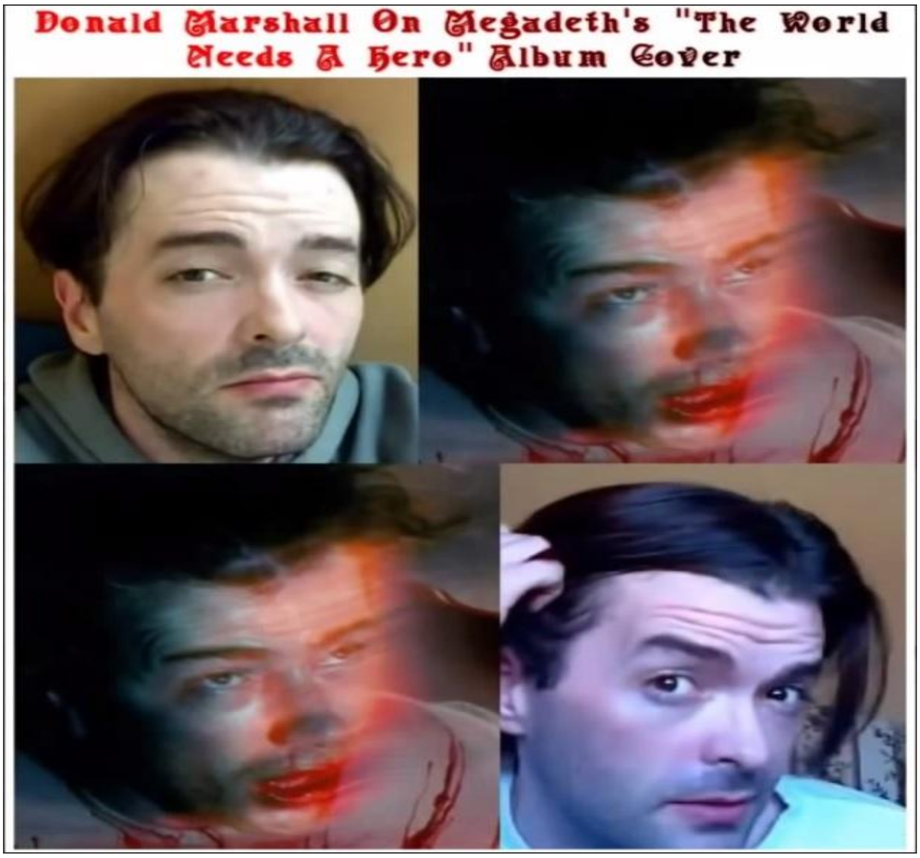
Donald Marshall On Megadeth's "The World Needs A Hero" Album Cover

**Image 4: My original body: top left (age 37) and bottom right (age 37);
compared to my 23 year old REM driven duplicate clone body (top right and bottom left) at the cloning centre.**

You can clearly see from the images that, it is in fact me. In Image 4, you can also compare what my real body (top left and bottom right) looks like compared to my duplicate clone body. Duplicate clones do not come out of the tank looking 100% like the original; it’s 99% - there are always some slight variations as some have commented that the eyebrows and nose in the picture, bottom right, is not the same (as the duplicate clone on the album cover). However, to people who make such comments, I say: my eyebrows and nose is the same when you compare it to the pictures of my original body (top left and bottom right). Also, remember that the image on the album cover is a duplicate REM driven clone of me. REM duplicate clones do not come out of the tank looking 100% like the original. There are always slight variations; but the wrinkles on my 37 year old forehead (when I took the picture in my original body; top left and bottom right) is similar to my 23 year old wrinkles, the lips are the same; and the face which you see, is definitely, the same! Yes. REM duplicate cloning is real; going to sleep and waking up as a cloned version of version of you is real and **we must put a stop to it now!**

The Deplorable Nature of the Illuminati

My own family has been turned against me one by one. The Illuminati have told my family members when they were REM driven clones at the cloning centre, things such as I secretly hit or molested their children; poisoned their pets when the pets died of sickness and other such things to make my own family hate me and not want to help me. Even when my family side with me there at the cloning centre, as REM driven clones of themselves; and my family members say ‘They do not believe that Donny did all those horrid things’ -which they have been told; Elizabeth then says “Are you cooling me a liyah?!” in her disgusting croaking voice made as low as low and evil sounding as she can make it, complete with psychotic and malevolent glare; and as terrified REM driven clones my family say “No, no, of course [what you are saying] it’s true”, stuttering and afraid as REM clones. My mother Catherine McMahon sold me into this REM driven sex and torture cloning slavery when she remarried a man named Gordon Cohoon; Gordon’s whole family is in this scummy secret society, his brothers Tom and Tony Cohoon; his sisters Darlene and Bernadette Cohoon; they fear the new and improved lie detector tests; and all it would take, is for an independent source, far, far removed from any association with the Illuminati to administer these lie detector tests on me **AND** the above named people to prove this unequivocally.

Half way through the making of my original disclosure document (in **2011**), when I went to sleep that evening, the Illuminati transferred my consciousness to my REM driven clone and introduced me to a man (who was also a REM driven clone) named TROY LANDRY. He is an alligator trapper from Louisiana on the TV program named “Swamp People”. Troy Landry said “If I send my [original] disclosure out to the public then he would take a power drill to my shin bone at the cloning centre and suck the marrow from my [REM driven clone’s] bones.” It is similar to spinal tap, and one of the WORST things you can do to a REM driven clone besides burning a REM driven clone. I told Troy that I will be sending out my disclosure to the public, because I have to escape the cloning centre, and so, Troy Landry did just that. It was excruciatingly painful; all the REM driven clones that were sat in the stands of the arena just watched; slack-jawed. Troy Landry is an insatiable child molester and an extra retarded REM driven clone. A side effect of REM driven cloning is that as a REM driven clone version of yourself, your emotions are heightened and you are less smart as you are when you are experiencing the world in your original body. Nevertheless, Troy Landry is extra retarded as a REM driven clone.

In Louisiana, when Troy sees a young boy that he likes, Troy asks of the boy's name and tells the boy "He is an alligator hunter"; Troy then proceeds to shake the boy's hand; Troy then has his paedophile friends at the local cloning centre find the boy's blood records to grow duplicate clones of the boy. Five months later, there are multiple REM driven duplicate clone bodies of the boy grown for Troy Landry to victimise before the crowd of onlookers. All members of the Illuminati do this; it's just a way of life to them; they consider themselves 'the privileged people' in the world power organisation. Almost all the cast on the "Swamp People" television program attended the cloning centre as REM driven duplicate clones of themselves, **when they go to sleep**. After putting a power drill to my shin bone, Troy Landry asked me again, "Will I send my [original] disclosure out?" I said "Yes" to which he replied "I'll do the same to your pelvic bone everyday!" The pain on my pelvic bone was far worse than when Troy put the power drill to my REM driven clone shin bone. I told Troy "I have no choice but to inform the public" – and so Troy put a power drill to my REM driven pelvic bone, everyday... and it is the worst pain ever... you forget your own name; where you are; all you know is pain; and you beg God or anything to save you from it...

Another deplorable thing which the Illuminati did in real life is: The Canadian government were trying to lower the amount of prostitutes on the streets (Elizabeth hates prostitutes) so they had a man named Robert Pickton start killing the women and feeding the women to pigs on his farm. The Illuminati had a camera set up in the upper corner of a room in Pickton's house and recorded Pickton hitting these women over the head with a hammer (a ball-peen hammer). They took the recordings, all 49 of these murders, and they all watch them at the cloning centre. To be seen as 'tough' at the cloning centre is to watch all 49 of the Pickton murders without having it affect your psyche. Elizabeth loves watching the Pickton murders and she says she has a macabre fascination with death. Canadian Prime minster Steven Harper knows all about the conduct of the Pickton murders; he has seen all the recordings and cheers on as the rest of them do. If Mr. Pickton says anything about his involvement with the Illuminati, or the recordings, they will apply a constant electrical current to his REM driven clone until Pickton has a heart attack or aneurysm in his original body.

What Happened to Britney Spears?

The Pickton murders are also the reason Britney Spears 'lost it' and shaved her head bald. She had watched too many Pickton murders as a REM driven clone at the cloning centre and it drove her 'over the edge'. I'll also reference the following songs performed by Britney Spears so that the reader has more corroborating evidence to add to the ever increasing pile. Search for the video "Break the Ice" (BritneySpearsVevo 2009) on YouTube. In the Japanimation video; at the beginning of the video; that is a cloned **cartoon** version of Britney Spears being grown in a big thick thank full of water (1 second to 7 seconds); and at 1:32 min to 1:43 min Britney walks into the cloning centre. This is **EXACTLY** how the cloning tubes are in real life. They are stacked one on top of the other, each filled with salty water. Britney Spears made this video because she said a fantasy of hers is to "Blow up a cloning centre". The Illuminati allowed her to release this video because at the time they figured nobody would ever figure it out. I have provided the images on the next page.

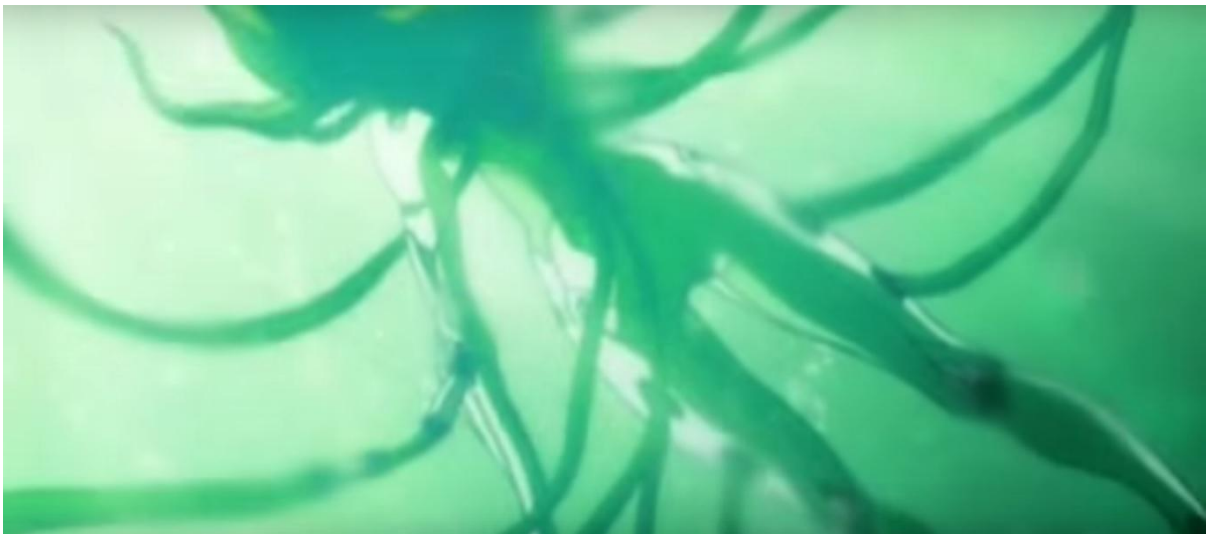

Image 5: Britney Spears is being grown as a duplicate clone in a cloning tube. The depiction is EXACTLY how the duplicate clones are grown in the Cloning Centre... with wires extending from the body; in a big thick tank full of salty water.

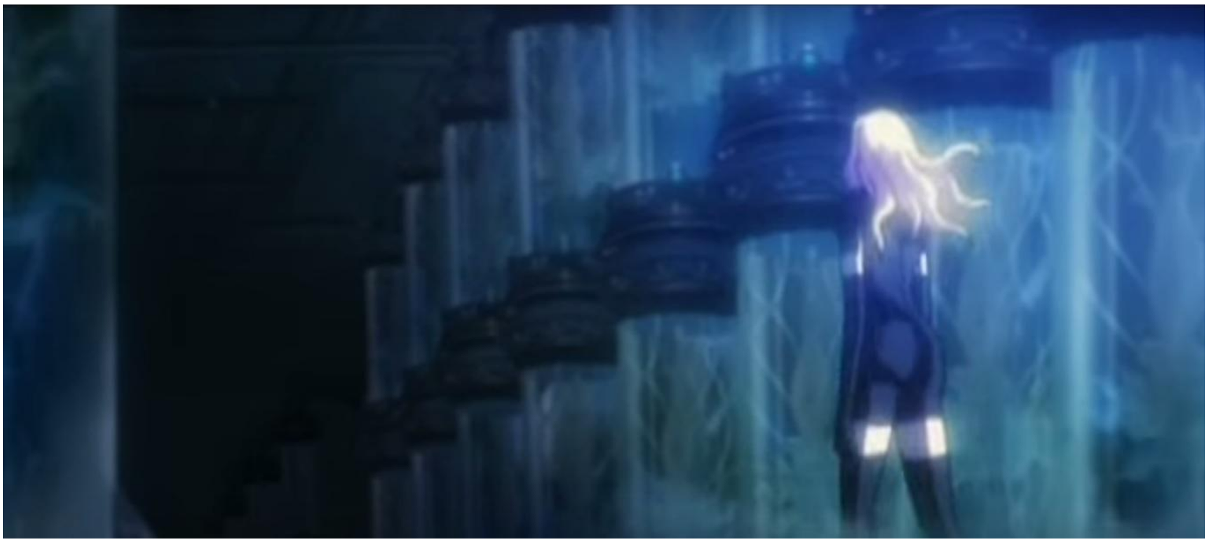

Image 6: Britney Spears walks into the cloning centre. The cloning tubes are stacked one on top of another. This is EXACTLY how these cloning tubes are in real life at the Cloning Centre.

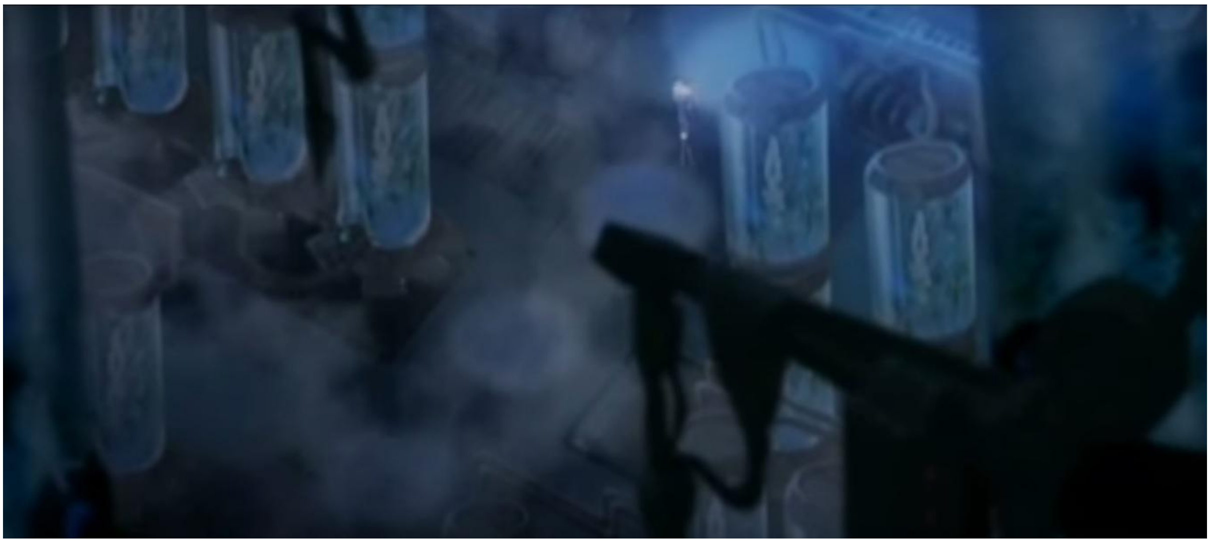

Image 7: The Cloning Centre depicted from an aerial view.

In "Hold it Against Me" (BritneySpearsVevo 2011) performed by Britney Spears; Britney is depicted fighting against her clone at 2:46 min till the end of the video. In "Mona Lisa" (SimpleGirl4ewer 2007) later remade by Britney Spears she changed the words in the song from "gone" to "cloned". I'll provide the lyrics to chorus here because they have not changed it on A-Z lyrics etc., but listen to the song and you will clearly hear the following:

Britney Spears Lyrics "Mona Lisa"

She's the original (yeah, yeah)
She's unforgettable (yeah yeah)
She wants you to know (yeah)
She's been cloned…
It's kind of incredible (yeah yeah)
She's so unpredictable (yeah yeah)
She wants you to know (yeah)
She's been cloned,
She's been cloned,
She's been cloned…

Yes. Britney Spears and many other celebrities want you to know they have been cloned. The celebrities are depending on you, the ordinary man / woman; the populace, as I am depending on you to spread this message fast and far. In fact Elizabeth has said that "If any public person speaks out, they will die!" Therefore, public figures cannot openly discuss this topic to warn the populace –so they hint. It is **NOW** up to the ordinary man and woman to end this. Furthermore, Elizabeth has a nuclear bomb hooked up at the cloning centre which she can press at any time to wreck havoc on the world unsuspectingly. Thankfully, I managed to convince Elizabeth not to do such a silly thing. I hope you can now understand why everyone is terrified of Elizabeth at the cloning centre, and why this is a **world emergency** and we are depending on you to share and spread my disclosure!

NEVER EVER SELL YOUR SOUL

In This Physical World Selling Your Soul Is Selling Your REM Driven Clone

Furthermore I must discuss an entrapment many up and coming 'stars' fall into. You have most likely heard many public figures say "I sold my soul to the devil" or some variation of "I sold my soul". I am now going to address the aspect of "selling souls" or "selling your soul" which I am sure you have heard many of these celebrities talk about. I'll address this as coherently as I can because the on the surface selling your soul talk is difficult to contemplate especially if you consider the soul to be a metaphysical, ethereal or mystical part separate from your earth body. Moreover, we live in a physical world, so selling a metaphysical aspect of you sounds incomprehensible (at first). For the most part we have little understanding of what the soul is. However, we are told that the soul is a part of us which never dies and it is eternal. It is all that we are, ever were, and ever will be. Now our faiths and beliefs come into question, because if you are an atheist: there is no God, selling your soul is a bunch of crazy nonsense.

From a practical perspective it also seems difficult to fathom, because if you go on what you have been told, you are selling an intangible part of yourself to be used by someone else for whatever they wish; and so if you get into the headspace of viewing your soul as something separate from you or that you are in control of your mind and your body, then again, ‘selling an intangible aspect of you’ again sounds like nonsense because right now nobody controls you; you are controlling you; so what are you really selling? Thin air?

Many up and coming ‘stars’ think this way when they are confronted with “selling their soul” aspect. The up and coming ‘stars’ are blinded by the prospect of being super famous and super rich; and in the moment they are confronted with “selling their soul” they believe these Illuminati people, are a group of over religious zealots and that they, -the up and coming ‘stars’ are taking advantage of the Illuminati because the ‘soon to be ‘stars’’ do not believe (in that moment) it is possible to sell your soul, and only consider being super rich and super famous; so the ‘stars’ sign their ‘soul’ over.

At this point, it is usually too late for the up and coming ‘star’; because what they have actually sold, in this physical world, is the use of their REM driven clone to the Illuminati. They have sold their REM driven clone to be used for REM driven sex by perverse old men and women whenever they enter REM sleep.

These celebrities are not memory suppressed either; so whenever they enter REM sleep, their consciousness is transferred to their REM driven clone, to be used for whatever the Illuminati wish: usually sex as REM driven clones with dirty and perverse old men in return for a promotion in the entertainment industry; such as movie role, advertisements, concert gigs etc. It is a very, very, sick business, and once these up and coming ‘stars’ get into the business they are trapped. Now you can begin to understand why so many celebrities, although on the surface it appears they have it all, are depressed; and do random and crazy things. It is a cry for help. Their illogical behaviours, (not all) for the most part, can be attributed to REM driven cloning. Furthermore, the Illuminati consider ‘selling your soul’ (the use of your REM clone) **a very serious business.** They do not joke about this; and there are NO RETURNS. Once you have sold the use of your REM driven clone for promotions etc., that’s it. If you ever make a fuss and want your ‘soul’ back, in other words: you no longer wish to have your REM driven clone to be used by dirty old men for sex; the Illuminati will kill you by applying a constant electric current to your REM driven clone until you have a heart attack or aneurysm in your original body and die. It is as simple as that.

For example, Whitney Houston wanted nothing to do with the cloning centres, she said she would keep quiet; she never wants to be in the spotlight; she never wants to do radio or television. Let bygones be bygones, I’ll go my way and you can go your way. The Illuminati were extremely angry! They had bank-rolled her, and now she didn’t want to be their ‘friend’ or ever associate with them again. She was sacrificed; she was killed with a constant electric current to her REM clone. She had sold her soul, in the beginning of her career; in other words, the “use of her REM clone”. Earlier in Whitney Houston’s career, she had signed that they can use her REM driven clone however they wish.

I have also seen people say "No" –when confronted with 'selling their souls'. These people are then memory suppressed, and will not remember the proposal of their REM driven experience. The people who also refuse to sell the use of their REM driven clone ('sell their souls') are blacklisted secretly, and they will never receive a high paying movie role, concert gig, advertisements etc. If you have ever wondered why so many talented people have not 'made it' –they refused to sell their souls, in other words, the use of their REM driven clones in exchange for promotions. This is how the entertainment industry really works. Another aspect of selling your soul, in other words, **selling your REM driven clone** is that you are also agreeing to sell your children, and your children's children etc.; your whole lineage to the Illuminati; for the Illuminati to use your unborn children for paedophilia, torture, sex sports and other grotesque and ungodly things when they sleep. It is a nasty business. **NEVER EVER** sell your soul.

The lesson is very important. NEVER EVER, 'sell your soul'; NEVER, EVER! You will experience the worst nightmare situation ever, and in this physical world, that is the use of your REM driven clone to be used for sex sports. You'll never have a decent night's sleep ever again, and you will always retain the full memory of your REM driven clone experiences; as well as, you will be selling your unborn children into this dirty business. Even when confronted with this as a joke: never sell your soul. Never Ever! The physical aspect you are confronted with when asked to 'sell your soul' is: "YOU", and everything which makes you, "you"; your ideas, your creativity, your privacy, your biology; everything which makes you, you. Never sell your soul, which in this life is currently the use of your REM driven clone! Never, ever! I hope that is well received.

Fake Alien Abductions conducted through REM Driven Cloning Technology

I must add they grow duplicate REM driven clones of people from all walks of life and then they transfer the persons consciousness to their REM driven clone version at the cloning centre while the person's REM driven duplicate clone is chained down to stainless steel corpse tables; or they pre-inject the REM driven duplicate clones with drugs before transferring the consciousness so that once the consciousness has been transferred, the REM driven clone cannot move. This is the point when they usually send people into the room dressed as aliens. It is a fake alien abduction. It's just these Illuminati people dressed in Hollywood quality accessories and they make the REM driven duplicate clones which have either been chained to the steel corpse tables or pre-injected with drugs, believe they are having an 'alien abduction'. Members of the Illuminati even dye chicken skin grey, and stretch it over a mask for realism and perform "experiments" on the immobile REM driven clones; the perverts anally probe them; they rape their limb bodies and do not suppress the victims memories; so these victims wake up thinking they have been abducted by aliens but in reality, they had their consciousness transferred to their duplicate clone versions when they entered REM sleep and they were violated. The victims do not know where to turn and naturally, they are embarrassed. Some victims even try and do install cameras all around their bedrooms in order to videotape themselves while they sleep; to prove they have been 'taken'; but to no avail they have not been taken; they have been cloned and had their consciousness transferred to their REM driven duplicate copies once they enter REM sleep by the most disgusting perverts in the world. The scum even videotape these 'abductions' to watch later as sick demented porn. These fake abductions are conducted by the Illuminati more often than you can imagine. They just keep doing it to random people over and over and over...

I am THE Spy for Humanity against the Corruption of the Illuminati

Over the years the Illuminati have offered me: a priceless katana (samurai sword); a rigged jackpot winning lottery ticket; REM driven clone slaves to keep me quiet, any of the prettiest women I've ever seen; any girl from high school; Vladimir Putin's daughter; even children; I do the right thing and spit in their REM driven clone faces; jam my fingers into their eyes but they have the technology to turn off the pain sensors of a REM driven clone, and therefore they have their pain receptors turned off as REM clones, while my pain receptors are fully on (sometimes even beyond the threshold of what is normal pain for a human). See the online articles: Science Daily (2014) and RT (2014) which discuss how pain receptors can be turned on and off in humans. Not much will hurt their REM driven clones other than getting dirt or vomit in a wound or bleeding too much. However, they have perfected the science and technology of REM driven cloning and therefore it only takes time and approximately $30 to grow new duplicate clone bodies. I started to smear excrement in their faces as a REM driven clone at the cloning centre in an attempt to deter them; I have **never** handled faeces in my original body, but I do this as a REM driven clone at the cloning centre, and even that does not deter them from their depraved ways. I composed a song which was later performed by "One Republic" referring to it called "All the Right Moves" ("All the right friends in all the right places... All the right moves in all the right faces"). I've made so many songs with MK Ultra technology (Rense 2001; Jim Cristea 2009; Berekley News 2011; Mind-Computer 2012 nature video 2013) and Mind-voice technology (New Scientist 2014; Prigg 2014) it is ridiculous. You see it is very easy to make music using Mind-voice technology; Flo-Rida used Mind-voice technology to produce the song "Whistle" when I first made my original disclosure (in 2011). The song is a taunt at me about my "whistle blowing disclosure" i.e. 'Can Donald blow the whistle? Etc. etc.' The thing is, anyone with a talent for music can make music much easier with the technologies which they have (MK ultra technology; Mind-voice technology); however, the Illuminati made more money from the music I composed, and so over the years, they just kept using me, over and over and over again, for rock, pop, rap, country etc. There are so many people involved in this REM driven cloning business, it is staggering; the organisation is vast. The only things the Illuminati fear are nuclear war; the new and improved lie detector tests and the general populace spreading my DISCLOSURE DOCUMENTS.

For those of you thinking Queen Elizabeth II looks like a kindly little old lady and you still cannot believe that all this is true, although I have provided plenty of evidence to the contrary as well as evidence of her saying "Drop the mother f***ing base" as a REM driven clone on a song track (RainmanJhof 2011); you could not be more wrong. She is the worst human being I have ever seen or heard of. It is so sad to see these REM driven women and children having their consciousness transferred to their REM driven clone duplicates and brought to the centre of the dirt rink, sitting there naked, afraid and crying and raped and beaten for sport, for the rich and famous. REM driven cloned women and children display the most range of emotions, and it is the saddest thing to see there. Therefore, of course I cannot bring myself to be a part of this despite any riches which the Illuminati offer me.

Over the 31 years when I was having my consciousness transferred to my REM driven clone at the cloning centre; Elizabeth, Vladimir and many, many Illuminati members and cloning centre attendees discussed many topics with me. These Illuminati people, such as Elizabeth and Vladimir, as well as many others confided in me because I had composed songs since I was 5 years old.

For readers struggling with the idea that a 5 year old can compose such wonderful songs at such a young age I will remind you that, it is easier to compose songs with technologies such as MK Ultra and Mind-voice technology. One is only limited by their imagination and creativity. Moreover, research talented children. I won't provide references of talented children here because I want to protect children from the Illuminati, and not have children exploited like I was; and for the readers who are parents; I am sure you know how talented, creative, and imaginative your children are.

Accordingly, since the age of 5, Elizabeth, Vladimir and many, others confided me in me because they thought I was special, and wanted me to hang with them; some of these Illuminati people, I had composed their favourite songs; I was also memory suppressed for many years so they thought it was harmless to tell me everything that exists or is possible. They were trying to impress me and amaze me, and therefore they showed me plenty of their technologies and what they had. They wanted me to me a willing, full Illuminati member; so over the years I feigned interest and pretended to like these Illuminati people, so that I could learn what they do and how they do it. I was looking for a way to block the consciousness transfer to a clone during REM sleep, or hoping they would reveal this part to me; but they did not, and I did not find a method to block the consciousness transfer. However, I bided my time, until I knew just about everything they do; so that I could reveal to the world everything they do; to give humanity a fighting chance against everything they do and as a consequence, I became a spy for humanity against the corruption of the Illuminati.

The Threat to Humanity when the H.A.A.R.P Grid is Complete

One of the most important topics which they discussed with me that I must add to this disclosure is HAARP technology and their plans for HAARP technology. I have discussed the basics of HAARP above. HAARP radio transmitters have to be placed at certain points across earth to complete a 'HAARP grid'. Once all the installations are completed over the earth HAARP will have more functions as a grid than the basics described. The Illuminati have told me, there is a possibility the HAARP grid will be capable of time travel, although there is a 50/50 chance using HAARP for time travel could destroy the earth. They also told me that once the HAARP installations are completed, the HAARP grid will be able to bend space and make the distance between, for example, Earth and Pluto approximately 1000 miles for a few seconds, and then the distance will go back to normal.

The Illuminati also said, once the HAARP grid is completed, they would be able to mind control people like never before without the use of RFID microchips or MK Ultra. Life on earth will be much worse once all the HAARP installations are up and working together in unison. The grid has to be complete though. HAARP cannot achieve time travel manipulation and mind control if it is just located in one or a few locations around the earth. There must be a certain number of HAARP installations around the world, in certain locations around the world; and the HAARP radio transmitters must be working in unison in order to achieve time travel manipulation and mind control of the populace of earth.

This is what they told me as REM driven clones at the cloning centre. This disclosure could be false, or a half truth. I do not know. However, it is very important that I add this to the disclosure because the Illuminati are working twice as fast to complete the HAARP grid to achieve their aim of mind control over the entire population since I the time I sent out my first disclosure (in **2011**). Whether or not they will achieve their aim of total mind control once all the HAARP installations are completed is NOT something we can afford to debate. They also told me that once the HAARP installations are completed; they will have the capability to mind control anyone remotely like never before, at the push of a button. HAARP will also be capable of setting back time; which the Illuminati will use if they are ever in danger of being exposed. However, the Illuminati will remember the previous time, and the general populace will NOT remember the previous time, so that the Illuminati can correct the mistake(s) which led to their downfall. **Once the HAARP grid is completed, humanity will be slaves forever.** This is why we cannot afford to debate. Furthermore, from the book "Angels don't play this HAARP" (Begich & Manning 1997, p. 8), it is estimated that the HAARP grid will be completed by approximately 2017; others estimate by the year 2020. The important thing to note is that humanity does not have long, until the HAARP grid is completed.

I must also add the positive side of HAARP, because as I have said, there is nothing inherently evil about technology. Technology is just a tool; it depends on the person using it. HAARP used correctly, and in the hands of honest men and women, will tame the weather with no adverse effects. Therefore humanity is on a timeline and a deadline to bring the Illuminati to an end. **HAARP is one of the main reasons, this disclosure must spread, and spread fast.**

See below for possible HAARP locations around the earth (Rense 2011). The HAARP sites are located globally where the richest mineral belts are located. Sub-surface mineral exploration has been done from satellites by radio tomography that is 100% accurate.

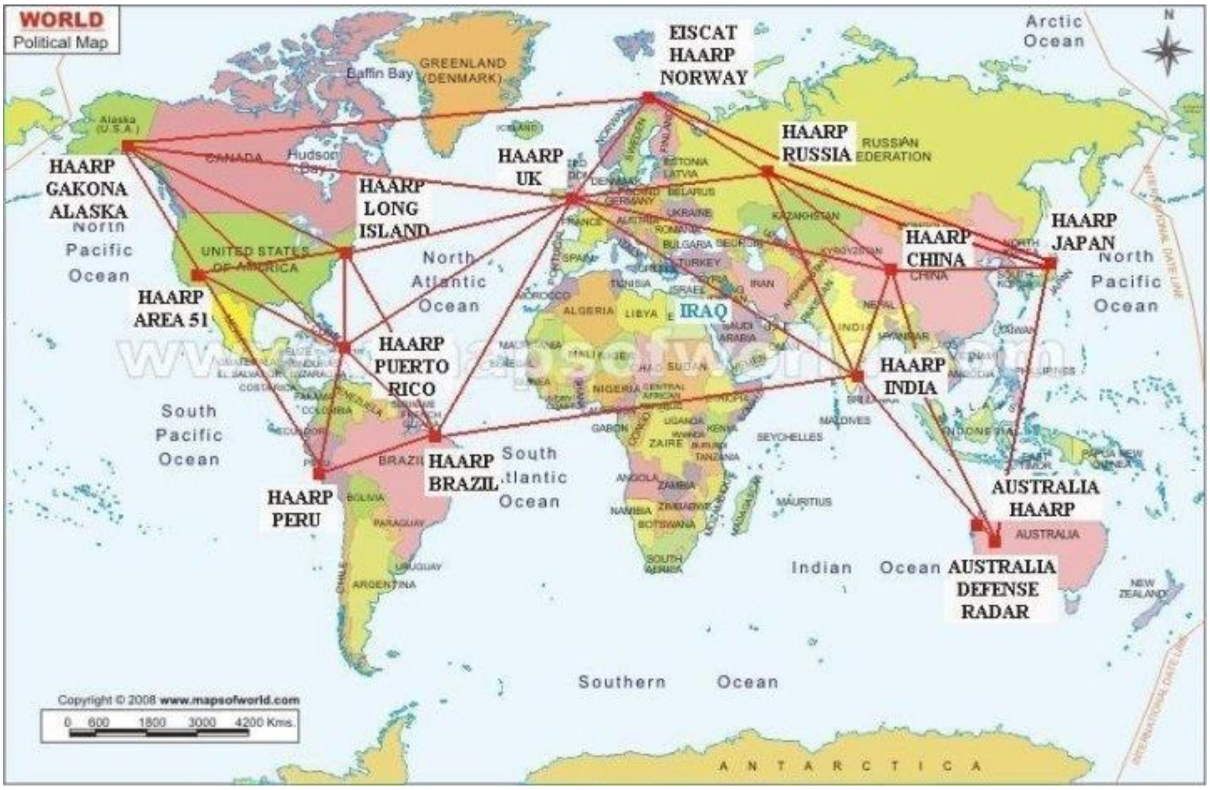
WORLD
Political Map
EISCAT HAARP NORWAY
HAARP RUSSIA
HAARP GAKONA ALASKA
HAARP UK
HAARP LONG ISLAND
HAARP CHINA
HAARP JAPAN
HAARP AREA 51
IRAQ
HAARP PUERTO RICO
HAARP INDIA
HAARP BRAZIL
HAARP PERU
AUSTRALIA HAARP
AUSTRALIA DEFENSE RADAR
Arctic Ocean
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
ANTARCTICA
GREENLAND (DENMARK)
Baffin Bay
Hudson Bay
CANADA
UNITED STATES OF AMERICA
RUSSIAN FEDERATION
Copyright © 2008 www.mapsofworld.com
0 600 1800 3000 4200 Kms

Image 8: Possible H.A.A.R.P. Locations. Source: Rense (2011)

REM Driven Cloning is the Most Terrible Nightmare Situation Ever

Furthermore, in terms of REM driven cloning; Fefe Dobson (singer) and Kurt Russell (actor) told me during my original disclosure (**2011**) specifically, to include in my disclosure that they do not like the cloning centre and they didn't torture me. Then they said not to mention them, but I will mention them again. Mila Kunis (actress) from "That 70s Show" stabbed in my REM driven clone body multiple times when I made my original disclosure and I was rendered immobile for saying that "She is a slimy scumbag for hanging with these people" and for also saying "She has enormous eyeballs and looks like a lemur". Mila Kunis (as a REM driven clone at the cloning centre) then begged me not to mention her (in my original disclosure, but I'll mention her again too) and then Mila Kunis said "She doesn't want to go to the clone zone anyway" and the Illuminati then deactivated her consciousness from her REM driven clone body, and she left. Her REM driven clone body went limp at the point where her consciousness was deactivated from it; her REM driven clone body looked 'dead', and she was gone. Some of the privileged people at the cloning centre are allowed **NOT** to have their consciousness transferred to their REM driven, clone duplicate versions (or they attend the cloning centre willingly when they want), but not me. I am an imprisoned slave in the worst nightmare situation. I told Nicole Leone (Madonna) also a REM driven clone at the cloning centre, that I was going to tell everyone that she coerced me to compose songs for her throughout my life (as REM driven clones). Nicole Leone (Madonna) told me VERY specifically to say in my disclosure here "She is not afraid!" and "No one will believe me and nothing will come of this". I beg you to help me prove her wrong.

Moreover, having your consciousness transferred to a REM driven clone alternative inhibits the person's ability to dream for that evening. Although your original body is immobile and asleep in your bed, when they transfer your consciousness you have actually 'woken up' as a REM driven clone version of you at a physical location here on earth; therefore you walk and talk as a clone version of you. You are not dreaming; you are just a REM driven clone for the evening. Furthermore, when the Illuminati restore your memories (as they have with me) or do not suppress the memory of your REM driven cloning experience for the evening (such as with the fake alien abductions), you will remember the experience as 'clear as daylight on a summer's day'; the experience is 'clear as a bell' as life is in your original body.

It is unlike dreaming, where dreams are fuzzy, incoherent and with random conjecture. The Illuminati know this as a fact, and therefore pre-inject the REM driven clones of random unsuspecting civilians for the evening with drugs before they transfer the person's consciousness. They tell the person who has had their consciousness transferred to their REM driven clone all sorts of nonsense: such as they are in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping etc. and all sorts of fallacies so that the person will not have a coherent picture of what has happened or what is happening. The truth is that person has had their consciousness transferred to their REM driven duplicate clone when they went to sleep and entered REM sleep; their original body still remains asleep in their bed; their consciousness however is now transferred to a clone duplicate alternative, and the person is 'awake', walking and talking as a cloned version of themselves in a physical location. The Illuminati allow some people to retain some of these REM driven clone experiences to affect their mental psyche.

The physical location where the populace will be able to verify these REM driven clones is somewhere within 5/6 hours drive, in a radius of the Robert Pickton farm, Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve. This is the cloning centre all the rich and famous attend because it is above ground and the rich and famous don't like the 'hospital smell' of underground cloning centres. The above ground cloning centre located in Canada is the cloning centre I have my consciousness transferred to when I sleep. It is guarded by military personnel. There are many cloning centres all over the world; as Phil Schneider (1995; 1996) disclosed as far back as the mid 1990s when he spoke: there were 1477 Deep Underground Military Bases (DUMBs) worldwide and 131 active DUMBs in the United States alone. Each DUMB costs on average 17 -19 billion U.S. dollars; paid for by the taxpayer; and it takes approximately a year-and-a-half to 2 years to build DUMBs with sophisticated methods. DUMBs contain cloning floors, and have an entire floor dedicated to human cloning, where duplicate clones are grown of unsuspecting civilians (remember they just need your blood / DNA); and unsuspecting civilians have their consciousness transferred to their REM driven clones at a physical location, which is usually these Deep Underground Military Bases when they sleep; unsuspecting civilians are memory suppressed, also to be used for gruesome torture and sex sports; as well as the plagiarism of talented individuals. Schneider gave the above statistics in the mid 1990s; chances are there are more DUMBS and cloning centres worldwide today.

Are you beginning to understand why this is the worst nightmare situation ever? Can you understand how: you, your friends, your family, your brothers, your sisters, aunties, uncles, children, work colleagues and anyone else you have contact with in your network, may be involved in this REM driven cloning business and will not even be fully aware of the situation or remember any experiences which they may have had once they wake up from sleep and they are back in their original bodies? I remember all of the REM driven clone experiences as clear as daylight because my memories are fully restored. Can you understand why the Illuminati must be stopped and stopped immediately? This REM driven clone business is vast, and it affects everyone, not just me.

This is why I mention that I am fortunate or unfortunate to remember all these experiences clearly depending on how you view it. I am unfortunate in the sense that I can recall all the depraved things these people do as REM driven cloned duplicates as clear as daylight and it affects my mental psyche. However, the fortunate part is: because I can recall all these experiences clearly, I can inform the world about the evil committed against humanity in a clear and logical way for the average reader to understand and verify, so that the populace can bring these people to justice for the crimes they have perpetrated against humanity and continue to inflict on humanity. For anyone who has had their intellectual property stolen by the Illuminati as a consequence of the technologies they have used on you unsuspectingly, you will be able to sue and regain monetary reward for the theft of intellectual property as well as invasion of your privacy.

I do not want to get too far ahead, because at this present moment, it is **very important** (I cannot say it enough) that this disclosure spreads (this is the stage we are currently at); it reaches the armed forces; to enable the armed forces to realise the crimes committed against humanity because of these people (the armed forces are kept 'out of the loop'; they just 'follow orders'); the armed forces will have to perform a military coup and bring this people to their knees. This is the point where the ordinary man / woman can have his REM driven cloning experiences restored, (that is if you want to) and sue / rightfully reclaim damages and legal reward for any crimes committed against him or her.

The Meek Shall Inherit the Earth

As you can see if everything goes smoothly; the meek shall inherit the earth. This is why you must maintain your composure and resolve, as I do every single day. You must not riot or cause chaos, damage property or undertake any act which will cause public dissension. The good people of this earth are about to inherit the earth; under no circumstance must you destroy anything. At the end of this currently dire and complicated situation; humanity will inherit the cure for HIV/AIDS, cancer, Alzheimer's, dementia, you name it –many debilitating diseases known to man; and other such wonderful technologies like hypersonic trains which run on magnets and are capable of doing Mach 5 (Mach 5 is the equivalent of 3806 miles per hour or 6125 kilometres per hour; Schneider (1995; 1996) also mentions the availability of Mach 5 trains in his disclosure, back in the 1990s). These trains can be used to deliver foods which are discarded in More Economically Developed Countries (MEDCs) to Less Economically Developed Countries (LEDCs) because these trains can travel thousands of miles within minutes. Yes. The Illuminati have all these wonderful technologies and more, but withhold it, and in turn withhold the progress of humanity. Yes. Plenty of man's problems can be solved when we address the root causes of our problems, and one of the root causes of humanity's problems is: REM driven human cloning which is currently in the hands of evil men. Even REM driven cloning can be used for the benefit of mankind when this tool is in the hands of righteous men. It can be used learn new skills in your sleep: such as gain extra qualifications; learn how to speak a new language; play instruments etc. –so long as memories are not suppressed, you will be able to perform the skills you learned as a REM clone in your original body when you wake up. As you see there is nothing wrong with most technologies; it all depends on the person / people using it. Nevertheless, although they may be beneficial uses to REM driven cloning not yet realised, we must be shut down REM cloning before we can even consider how REM driven cloning may benefit mankind, because at this present moment all that REM driven human cloning is used for is to commit evil, particularly against unsuspecting civilians.

It is OK to feel angry about all these crimes committed against humanity daily. But do not let these evil crimes rule you, in fact, let it motivate you; let it motivate you to the point where you will stand up and speak out to ensure that these crimes end. I think about the children and the future of the world every day; that is what motivates me and I view my situation as a God-given mission to bring these people to justice. I view things this way to maintain my sanity and composure. This is what keeps me going despite the extreme evil I experience, and it helps me to maintain my resolve and composure. The fact that I do not want a single child to grow up in a world, where they are being messed with in their sleep is what keeps me going despite this extreme evil. Some children are even kidnapped and brought to the cloning centre in their original bodies, again to be messed with. This is the root cause of why so many children go missing across the world. I'm sure any rational thinking human being feels the same way I do, and does not want their children, or any child to inherit a sick world governed by sick individuals.

I am a baptised Roman Catholic and a God fearing man. I believe there is a Creator of the universe but the Illuminati tell me there is no God, only science and technology. The Illuminati are also responsible for the death of the young beauty pageant girl Jean-Benet Ramsay; her parents are in this REM driven cloning business. Casey Anthony's daughter; as well as, Casey Anthony (before she died) AND Casey Anthony's parents also attend the cloning centre as REM driven clones. Many others; a man had his wife killed and this was mentioned on the news; it was an Asian woman whose rich husband had killed her; the husband said the wife was responsible for their son’s death. Elizabeth’s response was that “Rich people in this organisation (The Illuminati) do not go to jail, they are covered for.”

Stephen Spielberg and George Lucas are Illuminati members and they willing attend the cloning centre as REM driven clones when they sleep. George Lucas made other Illuminati members light my REM driven clone body on fire for the end of his movie *Star Wars Episode III: Revenge of the Sith* (2005).George Lucas said (as a REM driven clone) he wanted Hayden Christensen “To scream realistically”. Hayden Christensen (also a REM driven clone at the cloning centre) watched me, and listened to my screams and groans of pain and copied the sounds coming from my burning REM driven clone exactly. Hayden Christensen knows he is a REM driven clone when he goes to sleep and Hayden Christensen knows his consciousness is transferred when he reaches REM sleep. Hayden Christensen knows all about REM driven human cloning, and he is a “privileged” Illuminati member. Natalie Portman also knows all there is to know about REM driven human cloning; as does approximately a quarter of the “Star Wars” cast. Many directors have attended the cloning centre as REM driven clones too, and have used me in a similar role playing situation; they have caused me to have excruciating pain as a REM driven clone, to see what kind of squeak or screech I will make, as if I were less than a dog; and when I am a bloody mess on the floor as a REM driven clone; that is when they usually crawl on my broken REM driven clone body and sodomize me saying something like “They love me and they cannot control themselves because I made all their favourite songs and I’m so “special””.

My ribs in my REM driven clone body will be broken, I will have suffered internal bleeding and will be crying or screaming, if I am able, but they just continue to sodomize me, and videotape such a depraved act of wickedness so that they can view it again like “evil pornography”.

The REM driven cloned children, who have their consciousness transferred while they sleep at home in bed, in their original bodies, need a familiar face to talk to as REM clones or all they do is scream and cry at the cloning centre; that is where Joy Geizer comes in. Joy Geizer is married to my half brother from my father’s first marriage; Joy Geizer is a REM driven clone girl guide leader, and when the Illuminati clone young girls, Joy Geizer speaks to these REM driven cloned girls; Joy Geizer keeps the REM driven cloned girls calm and “pimps” them out for free, knowing that these young girls will have their memories suppressed, and therefore they will not remember the experience and will not talk about it when they wake up. All the Geizers are in this REM driven, human cloning business; there are many people who remember these REM driven cloning experiences in their ‘awake state’ and in their original bodies who could be polygraph tested by independent polygraph testers. The police, and polygraph testers in my City of Halifax, Nova Scotia are compromised (they attend cloning centres as REM driven clones too); they cover up for child molesters and therefore cannot be trusted to fairly administer a polygraph test; as I said, commissionaires and C.S.I.S are heavily involved. All it would take is a few polygraph tests to prove these things unequivocally, because for those of you who have ‘eyes’, you can already see that REM driven cloning is real, and that this is actually the way of the world. I will take these polygraph tests publicly too, and demand that my mother and step father submit to them. They have told me, they **wouldn’t** even attempt to lie on a polygraph.

Human cloning, particularly REM driven human cloning, is one of the absolute truths of this world; hidden from the world for over 70 years. My life and the freedoms of many, many people, including the reader (you) depend on this disclosure. The Illuminati say that the world finding out about REM driven human cloning will set the stage for the end of the world; people panicking and destroying property because of the evil these people have committed, but as I've said: prove them wrong, because we are going to inherit the earth, so there is no need to riot or cause public dissension. There will be no end of this world; just the end of the crazy Illuminati people. They always like to think and behave so negatively, don't they? *Sigh*. Furthermore, the whole talk about 2012; December 21st 2012 and the end of the world Mayan Prophecy; that was really about the world finding out REM driven human cloning, and the sick and sadistic nature it has been used for by these sick and sadistic people. Under no circumstance must you riot; you **must** prove you are better than these savage barbarians. I'll say it again, **because it is important**: we are going to inherit the earth, and therefore the smart and right thing to do, is not to destroy anything, especially the earth which you shall inherit.

Furthermore, I hope you are beginning to understand that the good people of this world must bring these people to justice and stop their wicked ways. I sincerely wrote this disclosure so that the good people of earth would be empowered to put a stop to their wicked ways. I wrote this disclosure to empower the good people of earth against tyranny and not so that the world would end. Therefore you must understand that the downfall of these Luciferians (Illuminati) must progress smoothly, because unless you have been to the cloning centre as a REM driven clone and retained the full memories of your experience, it is difficult to understand the level of depravity and sub human cruelty; it is beyond anything that has ever been heard of, and there is much more to tell.

Ignore this disclosure and you will condemn me to a horrible eventual death and you will encourage them that 'they are all 'powerful''. This is not an exaggeration. There is no need to exaggerate even President Barrack Obama is involved in the REM driven cloning business and attends the cloning centre as a REM driven clone version of himself, when he goes to sleep. He and his wife Michelle Obama attend these REM driven clone gatherings. Barrack Obama has even told me as a REM driven clone, at the cloning centre "Donny, we're all powerful. You are a slave and the people here [as REM driven clones at the cloning centre] won't speak up for you for fear of torture or death. Now make us a new song or we'll gut you like fish and leave you to writhe in agony!"

Be Wary of: Shills; Trolls; Disinformation Agents and Double Agents

There are many, many disinformation agents whose aim it is, is to denounce or debunk me. Be very wary of them. They DO NOT, and cannot debunk me, because after all, everything I have disclosed is truth. Although the truth concerns highly advanced concealed technologies, so it is a bit difficult to fathom at first, but keep reading, and you will realise that all which I have disclosed is true. Be very wary of the disinformation agents.

The most common thing these people say is that "Donald is a paranoid schizophrenic" –yet, ask them to detail the ins and outs of what a paranoid schizophrenic is and what specifically qualifies Donald to be a 'paranoid schizophrenic'; there is a high probability they will not be able to explain this to you, logically, sequentially and methodologically. Their main goal is just to make it as difficult for newcomers to understand what the real truth of this world is and as I have disclosed; and one of the root causes of man's problems is REM driven cloning.

Some disinformation agents are paid by the Illuminati, to behave this way, and are all too happy to echo empty statements (without any substantial evidence for their claims) because it provides them with a pay check. I will provide the following reference for the reader which discusses the signs and symptoms of a paranoid schizophrenic; all for the express intent for the reader to further understand paranoid schizophrenia. You can read more about paranoid schizophrenia on Medical News Today (2015). The link is provided in the reference section. You are welcome to read my Facebook wall and within days, you will realise I converse with others quite naturally. You can also ask individuals who have spoken to me through Skype or met me; but if this is not a viable option for you; all you have to do is read my Facebook wall and within days, you'll see that whoever makes such harsh and untrue remarks on my character is in fact a disinformation agent.

Furthermore, remember we live in a 3D **physical** world; and there is always a physical aspect to how something manifests. Therefore, when you hear others discussing concepts such as 'Tila Tequila demonstrated an 'energy ball' right on camera; Why didn't Donald discuss that?!' Please understand that these people are trying to deceive you with misinformation. There are also professional disinformation agents to be wary of. I do not perceive anything extra-ordinary in the world when I am awake in my original body beyond the five senses. I don't see orbs, auras or anything of the like. All I will ever discuss are practical things which have been unequivocally proven to me as fact; because in the situation the earth is currently in, I cannot afford to speculate.

Contact Information

My Facebook is: https://www.facebook.com/donald.marshall.148
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

I have a public wall on Facebook which starts from March 2012. In other words, all posts are made public since March 2012 and you do not have to join my friends list or subscribe me to see what I post (although you will have to have a Facebook account). I suggest that newcomers start reading from March 2012, and be patient, and read everything. You can read everything and learn the real truth of the world free of charge. I don't wish to write a book, I am not looking to turn a profit from this; I want to crush these guys and shut down the cloning centres!

It also worthwhile to add that: I am **not** asking for donations. I do not want **ANY** donations. Ever! These disclosures are far more important than any donations. This is **NOT** my job. I have a job. I am an independent contractor; carpentry is my trade and I earn a living this way. If you sincerely want to help; the best way to help is to spread and share my disclosures. That is all I ever ask for, so that eventually, the world knows about REM driven human cloning, and the armed forces can bring these people to their knees. **That is all I want; spread, spread, and spread this disclosure.** I will **never ever** ask for donations. Ever! Please keep this in mind, and anyone who asks for donations in my name or on my behalf should not be trusted. I, Donald Marshall, will **NEVER EVER** ask for any donations. I hope that is well received.

You can also view Proboards which has all my Facebook posts and have been archived by Celine O'Carroll and Astral 7ight by visiting:

http://donaldmarshall.proboards.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

There is a search function on Proboards, and you can use this to search for and read all the disclosures I have made regarding REM driven clones, the people involved and more. I understand that it is human nature to want to know, which people have been to the cloning centre as REM driven clones; therefore, use the search function to read about any public figure which I have already covered that you have an inkling about. You can also post anonymously on Proboards and Celine and other Administrators will transfer your question onto Facebook which I'll answer.

Donald Marshall Revolution is a website which details a brief overview of the Illuminati.
http://donaldmarshallrevolution.com/
You can also press Ctrl+Click (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the image to your right).

DONALD MARSHALL
REVOLUTION

Interviews

You can listen to the radio interviews I have done. Listen for consistency; particularly anything which you do not hear me, pronounce clearly for the first time; the best thing to do is to pause the recording at that particular point and replay it. You should also research the statement you do not understand. Sometimes reading helps comprehension a lot faster.

One of the main reasons you should pause and replay the recordings is because: the truth has been kept hidden for so long that a lot of what I discuss in my interviews are beyond most people's current world view; so at some points I may speak too fast for you; my audio/ microphone may not be so clear so you may miss what I say etc.

A friend has told me that when he first watched the Vinny Eastwood interview; he did not hear me say the word "scars" (When Vinny asked: 'How do I know I'm the real me?') although he replayed that particular point in the video 8 times. Everything was just beyond his current comprehension, at the time. No matter how many times he replayed that part he really could not hear me say the word "scars" –so he let that part go, and played the rest of the interview, pausing, and replaying points which he did not understand, especially to comprehend whether I was talking about my original body or my REM driven clone duplicate body. He also listened to all my interviews for consistency, to note any 'slip ups', or any parts of my testimonies which do not 'add up'. He would listen to all my interviews, pausing and replaying parts he did not understand and he would reserve his judgements until he felt everything I was saying was for example as ordinary as: 'I woke up today, brushed my teeth, and took the dog for a walk'. He was also patient to realise the truth. After listening to my interviews he would just let it 'sink in'. A week later he would come back and listen to the same interviews, to test whether his comprehension on the topics I discuss has improved, and whether he can understand what I am saying without having to pause and repeat at certain points in my interview; and soon enough he could now hear me say "scars" at that particular point of the Vinny Eastwood interview. He had reached the point where all topics I discuss sounded to him like I'm saying everyday common place stuff that people have heard, such as: 'I woke up today, brushed my teeth, and took the dog for a walk'. For anyone who may struggle to understand the topics I discuss: I strongly recommend you take the above approach as my friend did; soon enough, you too will realise the real truth of the world like he has: REM driven cloning, kept secret and used for sinister purposes.

I cannot say the following is true for everybody, however, an unproductive venture a complete newcomer can do is to listen to my interviews first time, all the way through, without pausing or replaying parts which they do not fully understand; If you do this and if there is just a single part of my interviews which does not make sense to you; this will interfere with your understanding of the entire interview. Remember, all I am discussing is technology, thousands of years advanced compared to what you currently use; available today, hidden and secret. If things start to get too complex for you, reduce it to its bare minimums: (advanced) science and technology. I hope that helps.

Another thing which I do in my interviews, that friends have picked up on, is: -because REM driven cloning has been my reality for many years; I don't differentiate between my original body and my REM driven clone body. I just say: I did this, I did that, and Elizabeth did this and that, therefore it can become very confusing for newcomers. Please bare with me; although I'm more emotional as a REM driven clone, and not as smart as I am in my original body (this is a side effect of cloning: REM driven clones are more emotional than normal, and dumber than they are in their original bodies) I'm still "me" when my consciousness is transferred; I have all the experiences and knowledge which makes me, "me" and therefore I naturally do not differentiate between my REM driven clone version, and my original body as an outsider discussing these concepts would. I understand it helps comprehension so I have painstakingly done this throughout this disclosure.

I hope this helps; and I hope this helps to better understand my disclosure as well as the interviews which I have done. On the next page you can find the links to my interviews. You can copy and paste the links to your web browser or press Ctrl+Click – (hold "Ctrl" on your keyboard and left "Click" with your mouse, on the images below) to direct you to the interviews.

Links to Donald Marshall Interviews

Vincent Eastwood

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab_channel=VincentEastwood
OR Press Ctrl+Click (on the image below) to follow the link.

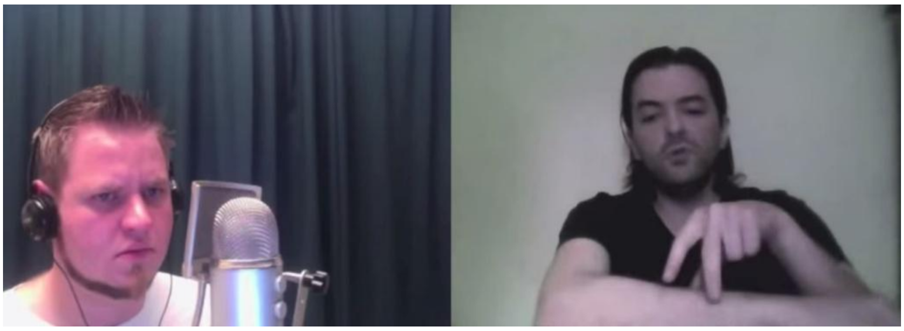

Jeanice Barcelo

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=3uzgu4ekT3c&ab_channel=JeaniceBarcelo
OR Press Ctrl+Click (on the image below) to follow the link.

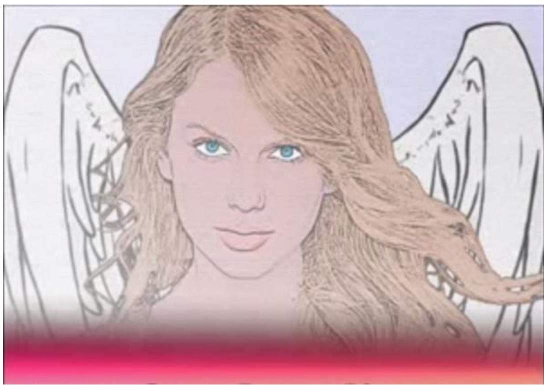

Lisa Phillips (CFR)

Copy and paste the link below to your web browser.
https://www.youtube.com/watch?v=UonnFuHLJKc&ab_channel=Astral7ight
OR Press Ctrl+Click (on the image below) to follow the link.
Listen to parts 1 through 8.

Donald Marshall
1st Part Of Interview
April 11, 2013

Radio Presenters –Contact Donald Marshall

Anyone who sincerely wants to contact me for radio interviews on their show is welcome to do this. Please contact me through Proboards by leaving a message for me to contact you.

Professionals who understand "Consciousness Transfer" –Contact Donald Marshall

Any neuroscientists, engineers or professionals who understand how consciousness transfer works, and can provide me with a detailed methodology of how to block the consciousness transfer to my REM driven clone; please message me on Proboards, and this will be greatly appreciated.

Email

I currently do not have a contactable email address. In my original disclosure, the email address has been compromised (hacked), and so has any other email accounts I created: Yahoo, hotmail, Gmail, AOL etc. It doesn't matter; they eventually get hacked; for whatever reason, the Illuminati do not want me to have an email account.

How to Learn More about Donald Marshall's Whistle Blowing Disclosures

The best ways to learn about the truth of the world is through my Facebook page and Proboards. There is a mountain of evidence which has been collated over the years, and corroborates all that I have disclosed here. I also understand that for some people it is impractical to read three years worth of Facebook posts, especially when some post are repetitive; contain unnecessary comments from trolls, shills and other disinformation agents. Therefore, I am currently compiling 5 documents which should make it easier for others to read and understand everything quickly; as well as, for others to share and spread those 5 documents including this main disclosure document –so 6 documents about the real truth of the world. All for free!!

These documents include:

- "Frequently Asked Donald Marshall Questions" –This will contain all the general FAQs which I have been asked over the years; all in one place.
- "Experiences from the cloning centre" –This document will contain the full disclosure of my REM driven cloning experiences from the cloning centre when I go to sleep; all in chronological order.
- "List of people to avoid" This will contain a whole list of trolls, shills, and disinformation agents and double agents in alphabetical order, to avoid at all costs, in real life and online; -all complied in one place; and the reasons why they should be avoided.
- "Public Figures and their relationship with Donald Marshall as REM driven clones" – this will be a complete list of every public figure I have ever come into contact with as a REM driven clone; and you will be able to read about all the public figures I have met as REM clones, as well as the experiences, all in one place.
- "The subterranean underground colonists: The Vril" –this will be everything about the biological parasites the Illuminati harbour. Yes. The Illuminati harbour biological parasites, which they use against unsuspecting humans. I told you the Illuminati were

> the biggest depraved perverts ever to grace God’s earth. REM driven human cloning, kept secret from the world for over 70 years is difficult to fathom as it is, as are these biological parasites. These parasites are also one of the main root causes of humanity’s problems here on earth, and must be rendered extinct. In this document you will be able to learn everything about their, strengths, weaknesses, psychology, biology, ecology, sexuality and arm yourself with knowledge which ensures that humanity exterminates these parasites from our earth, forever. The secret of Vril; a secret no more!

I hope after reading my eye witness statements presented in this document that you have now began to understand that this is simply what is happening in our world today.

Heed my disclosure, very diligently. At the start of my disclosure I asked the reader to do two things:

1) Give me the benefit of the doubt –because there is nothing worse than being a victim of abuse, reporting it and the people who have the power to stop it, ignore you;
2) Set out on a genuine quest to debunk my disclosure. I understand that for some people, plenty of what I have disclosed will still sound ‘too out there’ beyond their current scope of reality to be believable. For such people, the honest thing to do is to start with the first topic which you do not fully understand and explore it, either by finding that corresponding topic on Proboards (Donald Marshall Proboards 2015), and continuing your research from there; or just Google any topic you currently do not understand and review the topic on websites such as New Scientist (2015), Gizmag (2015), Motherboard (2015), BBC News (2015a; 2015b), BBC Future (2015), The Guardian (2015a; 2015b) The Independent (2015a; 2015b), Daily Mail Online (2015) and start reading more articles in the Science and Technology columns because these people have been telling you what they have been doing for decades. Consequently, keep in mind that articles which discuss ‘future’ technologies and ‘improvements’ in medicine or science, are articles, which are **really** discussing present technologies, and achievements in science and medicine already realised; because most of these achievements have been realised decades ago. It just hasn’t been fully disclosed publicly.

I have also saved and backed-up every reference (except the homepages of the websites referenced) in this document; therefore if a link or video is ever deleted let me know.

How Much Do You Know About Post-humanism / Trans-humanism?

Furthermore, you have to be honest here: because if you do not know much or anything about post-humanism / trans-humanism then I am afraid to say, -and this is no fault of yours – that you are behind in this the world; simply because these Illuminati people are Post-humanists; and that is the direction the want the rest of the world to go. They want the world to reach a destination where human cloning is common place; a world where downloading your mind / entire life experience onto a computer microchip and living on as a cloned version of you with the microchip running the consciousness to be common place; and these Illuminati Post-humanists want the world to head in a direction where it is commonplace to molest children through science and technology. No. This really is not Science-Fiction, this is the world we live in today; this is why such marvels in science, medicine and technology has been withheld from the general public and kept hidden by an ‘elite’ group of people because if these people honestly came forward, and said “We have invented this technology and we want to use it for this “negative” purpose” –the populace would not accept it, and these people will be lynched in the streets before lunchtime. Therefore, through concealed advancements in medicine, science and technology, the Illuminati can live out their inner depraved cravings and commit crimes against humanity and the world, for which most people will be too blind to see or even fathom, because the hidden science and technology is beyond the average person’s current comprehension.

Therefore, please attain an education in post-humanism / trans-humanism because these Illuminati people believe they can rival creation, nature and the marvels of the universe through science, medicine and technology. The Illuminati go **against** creation; they go **against** nature; and they go **against** humans. They don’t consider themselves humans. They consider themselves post-humans and believe they can become “gods” through advancements in medicine, science and technology. See LawOfIdentity (2014) and Mark Dice (2014) for Richard Seed’s comments on ‘becoming gods’ through transhumanism. Richard Seed is a Physicist with a Ph.D. from Harvard University; he is well known in the controversial cloning debate and declares his aspirations to "become god" saying 'We are going to become Gods, period. If you don't like it, get off. You don't have to contribute, you don't have to participate, but if you are going to interfere with me becoming a God, you're going to have trouble. There'll be warfare.' If you value humanity, do your best to attain an education in the trend of post-humanism / trans-humanism; otherwise the future of mankind will be bleak. We must do everything we can to stop these people.

This is of uttermost importance. REM driven cloning of is the most terrible thing in the world, especially when it involves unsuspecting civilians, and worst of all, innocent children. If you choose NOT to do anything you allow the Illuminati to continue to clone your children, sisters, wives, and sons. You allow the Illuminati to continue to hijack the minds of your children, sisters, wives, and sons while they sleep, through concealed advancements in science and technology. You allow the Illuminati to transfer the consciousness of your children, sisters, wives and sons to their REM driven duplicate clone versions while they sleep, whereby the Illuminati molest your children, sisters, wives and sons, which will cause them to have learning disabilities, unexplained depression and suicidal thoughts, as well as all kinds of side effects.

This IS your ONLY chance to end these monsters. There is a deadline!

This is your **ONLY** chance to do something to end these monsters. We are on a timeline and there is a deadline. As I have already mentioned, as well as, the top scientists in the field of physics and geophysics have mentioned: HAARP is not only capable of controlling the weather; once the HAARP grid is completed and working at full power, it will be capable of mind control over entire populations. We will all be slaves to the Illuminati FOREVER. Our freewill will be gone forever. Furthermore, if they achieve their aim of time travel with HAARP technology, future generations of humanity will never be able to stop them, because as they have told me: ‘They will always be able to go back to a previous time, and correct the mistake(s) which led to their downfall’. This is the Illuminati’s end game and what they wish to achieve. This is why this disclosure is so heavy, and so important. It goes beyond REM driven cloning; although it is important the world understands REM driven cloning. The future of humanity is at stake; and because of the 2-5 year deadline until the HAARP grid is completed, it is time humanity stood up and brought an end to the Illuminati. This REALLY is your ONLY chance to end these monsters!

The only way for evil to triumph is for good people to do nothing.
Edmund Burke

Remember for every 12 months which passes military / concealed technology outstrips the technology the general public is accustomed to by 44 years (Schneider 1995; 1996). Close your eyes and think 44 years into the future. The year is 2059. What kind of technologies do you expect mankind to have? Do it for real; don't just read the words. Close your eyes and imagine. The year is 2059. What kind of technologies does mankind have? You see. It is 'out of this world' technology, right. This is exactly what is going on today in the year 2015. It's just that it is hidden. That's all. Not so difficult to fathom now.

Do NOT Be Afraid to Help. We Outnumber the Illuminati by 1,000,000:1

Don't be afraid to help. These people are easily defeated when good people stand together as one. These Illuminati people total no more than 10,000 people. 10,000 people against 7,000,000,000 (7 billion) people –that's less than 0.000001 % of the world's population. Now can you begin to understand how prevalent evil can be when a small organised group of individuals, as little as 10,000 of the world's most evil and tyrannical people all work together in unison to exert their influence over the world? The world does not have to be this way. We outnumber the Illuminati people by a 1,000,000 to 1. For every Illuminati person there is, there are **one million** people who are not Illuminati, therefore do not to be afraid to help; all you will be doing is helping humanity rid itself of its sickness. So please feel free to share and spread my disclosure far and wide.

For those of you whom this applies to: remember the Illuminati (and I know it sounds ridiculous despite the evil they do) are deeply religious. They believe anyone aiding me is part of the "Army of Light" prophesised by Nostradamus; and to harm or degrade the life of anyone assisting me will bring them to their ultimate end and they will incur the wrath of God; they will suffer utter ruin and demise in their lives if they are to hurt you; they are very scared of people who can see through the lies and deceptions they have inflicted on the world. So please, stay calm, do not stress your heart, and know that **you are safe**, and **they cannot hurt you**. You can feel safe in the knowledge that you can do the right thing by helping me. They told me this as REM driven clones on the night of 21st of February 2014 when I went to sleep. The above is exactly what they said.

For anyone interested in understanding the interpretations of the Nostradamus Prophecies, Crystal Links (2015) provides a good source for all 942 of the quatrains. Delores Cannon is also an author who has written three volumes called "Conversations with Nostradamus" and for anyone interested you can read these online (Galactic 2012a; Galactic 2012b) as well as watch her YouTube videos (CreativeForceVideo 2014; Disclose TruthTV 2015).

Yes. This is the state of the world today: a deception within a deception coated in reverse psychology; and fact is indeed stranger than fiction. Now after everything I have said, if you are still sceptical and 'in-between' on this issue then the best thing for you to do is to pay attention to your dreams; or lack of dreams –they are your own experiences and you cannot deny your experiences.

In other words, you may be having REM driven cloned experiences too; for some people they have a ‘dream’, where they are in the same environment, over and over and over again; or the theme being discussed in their ‘dreams’ is the same over and over and over again; or the environment their ‘dreams’ take place always appear to happen in the same 3 or 4 environments over and over again –no matter what the theme is. In each scenario there is a likely probability that these are not just dreams, and in fact these people have had their consciousness transferred to the REM driven clone duplicates and the Illuminati is trying to extract something from them, once that is done they implant a false memory (See Kim (2013) and Alford (2015) for discussions on how false memories are implanted) so you wake up with the feeling of remembering something, but in actuality it is false; or you wake up with no memory of dreams from the previous evening, which means there is a high probability that you had your memories suppressed when your consciousness was transferred to your REM driven clone. This is why I say you can be sceptical about everything I have said (for now), but pay attention to your dreams, or lack of dreams, they are your own experiences, and you cannot deny your own experiences.

Noteworthy Frequently Asked Question

A Frequently Asked Question I receive, which is worth a mentioning here, is as follows: **Q:** Donald, if everything you have said is true...; in other words human clones walk among us now in a multilayered conspiracy which reaches the highest levels of government; armed with this knowledge what does one do exactly? What happens now, Donald? You obviously want to spread the word and make people aware but to what end?

DM: In short: I want to bring a complete end to the Illuminati and usher in a “Golden Age” of mankind. This question is best answered by detailing my ‘Mission and Vision’ for ending the Illuminati completely.

Donald Marshall’s Mission and Vision on How to Bring Down The Illuminati

- This disclosure must spread, and **spread FAST and FAR!**
- If you now understand everything I have disclosed in this document as truth, then do not waste any more time. Share this disclosure with your wives, husbands, brothers, sisters’ aunties, uncles, friends, co-workers, and children. We all have a part to play in saving the world from a premature doom.
- Call friends who have not heard from you in a while and tell them you have important news to share. Share this document. Share it on social media such as Facebook; Twitter; Instagram; Dropbox; Slideshare etc.
- Keep sharing this document until it comes to the attention of the Armed Forces. The Armed Forces will have to intervene. Once the Armed Forces intervene we will have reached the middle stage of this vision and will be witnessing a complete end to the Illuminati. Until then, we’re in the beginning stages, so please; spread this document faster and further. The quicker it is spread; the quicker the Illuminati are ended.
- Once the Armed Forces have intervened a Military coup can be orchestrated against these vile people.
- Cloning centres can then be shut down once these people are overthrown.

- The freeing of any missing people (children; teens; adults) trapped in the cloning centres can commence (once the Armed Forces intervene).
- HAARP (as well as other highly advanced technologies) can now be contained and not used for adverse effects against the world (once the world learns of this disclosure).
- These Illuminati people will then HAVE to appear in court for their crimes against humanity.
- Suing the Illuminati members in court (once court proceedings commence... we will be past the midpoint of my vision, and closer towards witnessing a complete end to the Illuminati).
- The populace (and it will be your choice) can then have their REM driven clone experiences restored and also sue and claim any legal reward / compensation.
- After the populace has sued the Illuminati for their crimes against humanity; the punishments can commence.
- Punishments will include: imprisonment and executions of these sick and malevolent people. The imprisonment and executions of evil Illuminati members will bring an end to the Illuminati.
- After the punishments and executions; Governments worldwide can now be replaced with incorrupt individuals, worldwide.
- The structure of Governance will also have to change. The reason for this because future generations will always be able to check their leaders and governance more appropriately; so that the depraved and subhuman acts I described can NEVER be committed against humanity again; nor will world leaders be capable of committing such monstrous crimes in such secrecy ever again.
- Ensuring that the laws of the land always progresses in line with advancements in medicine, science and technology; as well as, ensuring law progresses in the directions of Research and Development (R&D) regarding future technologies, science and medicine.
- Release of technologies, science and medicine which benefit mankind.
- Commence a “Golden Age” of mankind.

If you do not fight the corruption and you do not stand up for what is right in life, you end up being a serf and a slave and you are leaving your children a world in which you would not want to live in yourself, so how can you in decency behave that way? You have to stand up for what is right in life, and unless you do that you are nothing.

Aaron Russo

Empowerment by Virtue of Golden Truth

As you can see, despite all the horror of the world I currently present: if everything goes smoothly, the good people of this world, truly will inherit it. This is why despite any anger, sadness, or fears you may have; you must not riot, damage property or cause chaos or bring about any other form of public dissension. You cannot stay silent, or ignore the issue in the hope that the threat removes itself. You are called to act; you must take action to help bring the Illuminati to an end; you must act while at the same time, you must maintain your composure and resolve to ensure that the whole procedure goes smoothly, and we all inherit a world we want to live in.

Your life, **at this very moment**, is more important than you may have probably ever imagined. You have purpose. Through you, and other good people around the world, together we can bring an end to such unspeakable forms of tyranny in this world. It is my humble stance that you have now become truly empowered with golden truth and you are now compelled to bring this tyranny to an end.

Do not waste the knowledge you have obtained from this disclosure. It is my only hope to escape this man made living hell. It is my only hope, as well as, the hope of many REM clones imprisoned there, as well as, real people who go missing daily, and are trapped in the cloning centres.

We sincerely beg you.

Donald Marshall

Appendices

Appendix A: Thien Thanh Thi Nguyen (Tila Tequila) Transcript

Link: https://www.youtube.com/watch?v=7mRZ7ItF9ls&ab_channel=Astral7ight

00.00 – 1.10 min: You know what... since you f****** with my program darling Queen Elizabeth [II] and the paedophilia ring and the cloning centres, and the cloning centres. That's right darling the cloning centres. Parents listen to me right now, they are blocking me but that is quite alright. They are blocking me but that is quite alright. Because I have many, many other forces; I shall not say their names right now, but I have many big plans to expose all of you disgusting, sadistic f***s! Okay? That is all. I shall save that for another time. But, however, I shall REPEAT: that was just an introduction to the reptilian family, leading all the way back. They call themselves the "The Black Nobility". Now that is just one part of it; alright?

1.11 -2.47 min: The Black Nobility; the reptilian family; all the way back from ancient times; so which they think... they feel like they are the divine chosen ones... from whom may I ask? Definitely. Definitely not God. Our creator of the world. Reptilians: they feed on blood; children; the paedophilia ring; recently busted. Oh! It has been going on for centuries. Parents listen to me carefully. I don't care if I'm cutting out. I will continue this and I am not alone on this battle. Believe THAT! Believe THAT! I am not alone on this battle. I started out alone but I am marching on with MILLIONS; okay? So sit your old a** down okay. Because you are gonna roll over, but it doesn't matter anyway because you are all ancestral f***s! Who interbreed... ancestral f*****g... and then... and then... Oh! Only going to talk about cloning those children and... oh! And all those many children's parents listen to me carefully.

2.48 -4.24 min: Hundreds of missing children come up every year. You wonder... why? How? How could this be? And then there is so called CPS [Child Protection Services] or whatever they are called; they come and take your children, just, just for nothing; right? Not to discredit everybody, because not everybody is bad. I'm talking about the bad people. And they take your kids... they take them as this... they treat them like... I can't even say the word. It's disgusting, it's sadistic. They... they... they take your kids... they toss them out like little [inaudible] cause they are so f*****g... Pardon my French, but then again when I'm dealing with these evil cum-buckets I have no holy words coming out of my mouth, because these are the only words they resonate to. So therefore that's how I refer to them. Because they can only [inaudible] ...their masters whiplash on them with these [inaudible] words of cursing, vileness and slaving and that is not what the true God is; okay? The God of Hell...; Anyways...

4.25 – 5.30 min: There are these CLONING CENTRES where they take your children and do sadistic things to them. I'm not even talking about child molestation here; not to mention; uh there is one of them that got caught, flying out to Florida; to meet up with a four year old little girl, to have sex with a four year old girl. That's right. Google it because it is so highly sinful... We have commercials about... you know starving kids; you know save the starving kids and it is heartbreaking. We can have commercials about it... why? Because it is a horrible thing and people can have commercials about it. WHY do we not have commercials about... 'Daddy please don't, don't. Mommy please don't let daddy let daddy touch me?' – because it is disgusting! It is so disgusting; beyond sin that no one, NO ONE, can even make a commercial about that because that's how sinful it is.

5.31 – 7.19 min: Do you understand that? Do you understand how sinful that is? These people, I'm not even going to call them people; alright? They take your children; they not only molest them; men f*** them and make them shoot each other. They give them guns. It's either you shoot him or I'm gonna shoot you. They are... mind you, they are children. Children. Yes. I'm speaking out, because I... uh! Who else is doing this? You're all just [inaudible]... shame on you... And actually the most recent paedophilia... got taken down... WOW! How long did that take? Really?! Do you know how long this has been going on? *Sigh* Alright I'm gonna calm down... but as a parent and I love parents out there. My heart goes out to all the parents out there who have missing children. You know, we all pray for them; every day. And I put on a bold face in public every day; because there needs to be someone strong, believe that. But my strength comes from somewhere... that I have a very vulnerable emotional side where I feel very strongly for these children and innocent people. So therefore I want to speak out.

7.20 – 9.10 min: I have and I have my passion too; and in the end you shall all know why I am so passionate about exposing every single one of these scumbags; okay? The truth shall prevail and you all will know why. So, as for you parents... ah... there are no words to describe. But let me just expose because you can't just... there is a point where you CANNOT just turn the other way. You know this stuff is going on; and you go 'Oh well... that's their problem. Let's just turn the other cheek.' How long are you going to turn the other cheek, until it happens to your own freaking children? When, when, when your own child; three years old, get's run out [kidnapped] and gets blasted right in the head with a gun... yeah... there's more to that people; okay? And I'm not just saying that coming from some... I mean, actually, mothers, fathers out there... If you found out... I'm sure you would do way worse than what I am just saying. I'm just using voc. I'm just annihilating them vocally. I'm sure the parents out there who find out what their children have gone through, their missing children. I am pretty positive more than just a vocal annihilation of these scumbags that do this to your children; okay?

9.11 – 10.32 min: So keep on turning the other cheek folks. Hey, go and turn the other cheek; you are with the others. I am not. I AM NOT. Like I said you... There is only two ways to go about this: you're either with us: the good guys, or you're with the others. There is no in-between. Cause if you are in-between, hiding like cowards, turning the other cheek; doing whatever; well then you are a freaking coward and you are just a sheep. And sheep end up dying because you know what you are owned by "The Others". So pick one: you're either one of us: the good guys or you're part of the others. That is simple as that! Two choices: good guys; bad guys. In-between you're dead because the bad guys are going to suck your soul out [transfer your consciousness to your REM driven clone] and do some sadistic stuff to you and watch, and make you watch while they pretty sadistic stuff to your children as well.

10.33 – 11.33 min: Do you know why they love children? Because they are innocent souls; they're innocent... they're, they're the most innocent pure beings in this planet. They're not harmed by anything. They're new to the world; bright-eyed pure innocent children. That is why these disgusting paedophile and these clone rings; cloning centres; satanic rituals; Brownsville Texas... There are many other cloning centres where they take your children that go missing. And you wonder why? What happens to them? I'm sorry to break this to you but that is what happens. Now either do you want to know what happens to them or do you want to turn the other freaking cheek?

11.34 – 13.09 min: Tune into my next show. I'm gonna upload stuff; I have an arsenal ready to blow up! Okay. And I have reason behind this. You all should know my personal reason soon; but this has nothing to do with me right now. But know that I'm back, I'm back with a vengeance and I'm back with an army full of people around the world who are sick and tired of treated like animals; or quote, unquote "COWS". We all know what that means. For the outside world (the masses) we all know the term "sheep", sheeple. But for the insiders we know what the cows are don't we? You know what "The Others" like to do with the cows, right? They start to herd you in... and to... yeah...I'm gonna end it at that. And to all the parents, families and children out there, I love you so much. I... have to maintain composure, because that is what I do. That is all. Over and out.

Appendix B: MK Ultra

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct, and to control the absolute behaviour and thought patterns of the individual.

.

Modern-day MK Ultra involves an implantable microchip which is inserted into a Mark 2 clone of the victim. Every thought, reaction, hearing and visual observation causes a certain neurological potential, spikes, and patterns in the brain and its electromagnetic fields, which can now be decoded by the implantable microchip into thoughts, pictures and voices. The thoughts, pictures and voices of the implanted Mark 2 clone (victim) can now be displayed visually and heard on any system capable of converting visual images, such as a television or a computer. These images are usually displayed on a giant screen at the cloning centre.

The purpose of MK Ultra is to elicit a specific conditioned response in a victim (desired by the programmer) to an otherwise neutral stimulus. In other words, through the process of "classical conditioning" (which involves learning a new behaviour via the process of association. In simple terms two stimuli are linked together to produce a new learned response in a person) the implanted Mark 2 REM duplicate clone (victim) is placed thorough many similar recurring scenarios which are experienced (or perceived) by the victim as "real", because the programmer manipulates the victim's audio and visual field, and continuously pairs the victim with the neutral stimulus, so that either a positive or negative conditioned response is elicited in the victim towards the neutral stimulus.

In other words an implanted victim can be placed in many recurring situations deemed as frightful, where a mystery stranger saves the victim from the frightful situation over and over again. The victim will now have positive associations towards the mystery stranger and therefore the conditioned response is now one of 'positive associations' (towards the mystery stranger who did not elicit such a reaction before).

MK Ultra techniques are administered through Mark 2 (sleep driven) clones. In other words classical conditioning happens when the person sleeps. Memory suppression technologies are also used in conjunction with MK Ultra technology, which enables the programmer to control certain memories the victim remembers. The use of memory suppression technologies and MK Ultra technology allows the programmer to reinforce behaviour and elicit specific conditioned responses (in an original).

Therefore, in our example above, when the original meets the mystery stranger in real life, the original (victim) will be predisposed to the mystery stranger and automatically 'feel safe' around this person and may even believe that he / she should 'date this mystery stranger' – depending on what was programmed; because the victim was conditioned to elicit such an emotional response towards the mystery stranger in his / her sleep as a Mark 2 REM driven clone through MK Ultra functionality.

MK Ultra can also be used to create zombified individuals who are programmed to murder and remember nothing of their crime afterward. MK Ultra (technology) can also be used to disrupt the memory of the original, discredit people through unusual behaviour, make the person insane or commit suicide and murder.

MK Ultra has much functionality. Furthermore, because of its video and audio projection functionality, MK Ultra can also be used to compose music. MK Ultra is capable of relaying the Mark 2 REM driven clone's subconscious and conscious mind as visual images and audio, projected onto a screen. Therefore thoughts, pictures and voices which have been experienced consciously and subconsciously in one's life can be displayed visually and audibly on a television or computer screen. The Mark 2 clone now has the option of harmonising these thoughts, pictures, and voices into a coherent order which produces music through the thought process of his or her imagination. See Jim Cristea (2009); Berkeley News (2011); UC Berkeley Campus Life (2011) CTForecaster (2013); nature video (2013) and Stromberg (2013) for examples and discussions and the capability of MK Ultra technology.

Appendix C: The Illuminati

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies.

Today's Illuminati trace their roots back to Professor Adam Weishaupt who found the Illuminati on 1st May 1776. Since the inception of the Illuminati the intent has always been, and remains: "to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer." This intent is still prevalent today. The following excerpt is derived from A. Ralph Epperson (1990) *The New World Order:*

> Weishaupt was a teacher of Cannon Law (law governing the affairs of a Christian Church, especially the law created or recognised by the Papal authority in the Roman Catholic Church) at the University of Ingolstadt in Bavaria, now part of Germany.
>
> He even told the world, in his writings, where he would conceal the Order: "None is fitter than the three lower degrees of Free Masonry; the public is accustomed to it, expects little from it, and therefore takes little notice of it." He felt that this secrecy would lead him to success because he felt no one would be able to break into it. He wrote: "Our secret Association works in a way that nothing can withstand ...."
>
> Weishaupt accepted the fact that all secret associations and secret orders had two doctrines, one concealed and reserved for the Masters... the other public ...." and the Illuminati was [and are to this day] a secret society with two doctrines.
>
> Professor Weishaupt, its founder, boasted of his organization's secrecy. He realized that this secrecy would enable them to decide the fate of nations and because their deliberations were secret, no outsider could interfere. He wrote: "The great strength of our Order lies in its concealment; let it never appear in its own name, but always covered by another name, and another occupation." Weishaupt later wrote about that secrecy in a letter to a fellow member of the Illuminati: "Nothing can bring this about [the new world order] but hidden societies. Hidden schools of wisdom are the means which will one day free men from their bonds [the "bonds" of religion] Princes and nations shall vanish from the earth." So the secret societies were created to bring the world to the new society known as the New World Order. The members of these organizations obviously feel that their goals are so noble that they may perform whatever tasks are required of them to bring that goal to fruition. This means that murder, plunder, and lying all become acceptable as long as these methods assist its members in obtaining their goal.
>
> Adam Weishaupt, the founder of the Illuminati, wrote over and over and over again, that "the ends justified the means." Weishaupt also told initiates to use whatever means, which included murder, to achieve the goals of the association that he was joining. And that the major goal of the Illuminati, was the destruction of all religion, including Christianity. That meant that if Christians physically stood in the way, they could be removed by simply murdering them. Weishaupt even went so far as to say that anyone not willing to take the life of another was unfit to join the Illuminati. He wrote the following in a letter to a fellow member in 1778: "No man is fit for our Order who is not ... ready to go to every length ...."
>
> Another reason, that Weishaupt felt that the Illuminati would succeed, was the fact that he was offering his members worldwide power. He felt that this inducement would enable him to draw into his organization only those who would do anything to satisfy that desire for power. He wrote: "The true purpose of the Order was to rule the world. To achieve this it was necessary for the Order to destroy all religions, overthrow all governments and abolish private property."

> But his religion had a different base than the traditional religion: his was based upon a worship of reason: "... then will Reason rule with unperceived sway." "... Reason will be the only code of Man. This is one of our greatest secrets." "When at last Reason becomes the religion of man, then will the problem be solved." Weishaupt's dedication of his organization to "reason" makes some sense when the reader recalls that "reason" has been defined as the "unbridled use of man's mind to solve man's problems without the involvement of God." The Bible calls this "the fruit of the tree of the knowledge of good and evil." It was this knowledge that God wanted man not to have, and it was the promise made to man by Lucifer that man could have it by eating of "the fruit." In addition, Weishaupt's religion offered its believers a reward not offered by any other religion: worldwide power!
>
> Weishaupt wrote: "The pupils [members of the Illuminati] are convinced that the Order will rule the world. Every member therefore becomes a ruler." Weishaupt's religion not only offered power to his believers, but he offered them something else not guaranteed by any other religion: worldly success. He said that once a candidate had achieved the exalted degree of Illuminatus Minor, the fourth of the thirteen inside his Order, his superiors would: "assist him [the member] in bringing his talents into action, and [would] place him in situations most favourable for their exertion, so that he may be assured of success." Finally, the goal of the Illuminati was "man made perfect as a god - without God."

The ideology of "man made perfect as a god –without God" still remains to this day, and it is practiced by today's Illuminati members. The above phrase is what ties in Luciferian worship and trans-humanism. Lucifer is idolised by Illuminati members as the deity who gave man 'knowledge' and therefore is worthy of worship; God, -according to Luciferians, -did not want man to have knowledge and therefore is despised by Luciferians. Ingrained in the trans-humanism doctrine is the believe that: 'man can become 'god' through science and technology and in turn overthrow the Creator of the universe: God'. See LawOfIdentity (2014) and Mark Dice (2014).

These are the basics of Luciferianism. Therefore everything which is natural or pertains to nature must be contended or destroyed by Luciferians. This is why Illuminati members endorse having sex with children, killing first born sons, and drink blood. All the above go against nature and according to the 'edicts of Lucifer': paedophilia makes the person committing the act younger (it doesn't, it is just an excuse to act perverse on children because they know children are vulnerable); killing your first born son gives you good luck and fortune in this life (so yes, some Luciferians have sacrificed their first bon sons); and Illuminati members believe drinking blood / cannibalism is a 'purifying agent' (although in reality it causes spongiform encephalitis (holes in the brain)).

Modern-day Illuminati members still retain the goal of its founder "to bring about a NEW World Order that writes God out of the picture and deifies (glorifies) Lucifer." Infiltration through secrecy, still remains their mode of operation, for the current Illuminati and therefore they have secretly infiltrated all the major religions on earth; government and education – where each successive generation is being dumb down; they have continually diminished the ability for individuals to own private property, or claim inheritance; divided people against each other to continually diminish patriotism; and have continually diminished family values.

They also compartmentalise their knowledge between members. Until I fully exposed the Illuminati, many people who have been REM clones at the cloning centre did not know they were in fact REM driven clones, and thought they were in the 5th Dimension; the astral plane; a singularity; the spiritual realm; Valhalla; quantum hopping; a time stutter etc. or whatever else the Illuminati told them.

Today's Illuminati members also meet in secret (just like the founding members) as REM driven clones, **when they go to sleep**. Furthermore, because Illuminati meetings are in secret and not many people know the exact location (because knowledge is compartmentalised) of the cloning centre; or the fact that their consciousness has been transferred to Mark 2 REM driven clone bodies at the cloning centre (and they are not in their original bodies); as well as the fact that unsuspecting civilians have their memories suppressed; the points mentioned above are the reasons the Illuminati believe 'they are all powerful and untouchable'; and as a consequence, today's Illuminati members do all the disgusting things they want, because they believe no outsider can interfere.

The ring leaders of the Illuminati today also believe "the ends justify the means". This is why they clone, torture, molest, murder, and rape unsuspecting civilians as REM driven clones in their sleep. The Illuminati of today offer their members incitements to go along with their agenda and not oppose them (or face death). Another popular method is to entice their target with many, many wonderful prospects, and have the target believe they are joining a noble and prosperous venture, so that the target fulfils the objectives of the Illuminati unknowingly; promoting the Illuminati in a positive way, because the target has been deceived to perceive the nature of the Illuminati as 'positive'; by the time the target finds out the true intent of the Illuminati, and the evil which emanates from it, it is too late. Those who rise up in the ranks of the Illuminati are the men and women who have an insatiable lust for power, and most importantly: the men and women who want to rule the world.

Modern-day Illuminati members also wish to become gods (through technology); overthrow the Creator, and achieve their overall aim of enslaving mankind. This is why they clone people, and clone people in high rank society from all walks of life (movie stars, musicians, politicians etc.; whether the person willing wants to be part of the Illuminati or not) so long as that person is in a position of power and influence, the Illuminati clone that person, and threaten that person, for example -"Hey, you're going to hang with us –or else" –through such coercion, the people in high rank society who have power and influence will not oppose the Illuminati's plans to become gods; enslave mankind forever; and rule the world. Another reason for cloning high rank society is to include these people into the Illuminati (willingly or unwillingly) to ensure the world's populace remains in ignorance (because once all the world leaders and high rank society are cloned and under the coercion of the Illuminati nobody in a position of influence or power can warn the populace against the Illuminati) until the Illuminati's plan is completed and they have enslaved the world forever.

The Illuminati's overall aim of 'becoming gods' and ruling the world, as well as, mind controlling all the inhabitants of earth is also the reason the Illuminati:

- Administer drip feed disclosure through media, by telling some truths mixed with lies in order to conceal their true intentions and overall aim, and prevent the betrayed partner (the public) from ever discovering "the complete truth";
- Administer evaluative conditioning, by placing their symbols and ideology in popular media with positive associations, so that the unknowing and unsuspecting public will eventually become predisposed to the Illuminati and unsuspectingly have a positive or neutral response towards the Illuminati;
- Because they want to become 'gods' is also the reason the Illuminati is promoting RFID microchips and only discussing RFID microchips positively, while at the same time placing suppression (gagging) orders on anyone who speaks negatively about RFID microchips –which implies an unsuspecting public will willingly accept the microchip; and at the point of transaction, the person will have (unknowingly) given up their privacy to a third party (the Illuminati) for the rest of his / her life;
- Their aim to become 'gods' is also the reason the Illuminati are hurriedly trying to complete the HAARP grid across earth –because a complete HAARP grid will allow them to achieve their goal of mind control over the entire world; which fulfils their objective of becoming gods; because a complete HAARP grid will be capable of time travel, and therefore the Illuminati will be capable of going back to a previous time to correct the mistake(s) which led to their downfall; the Illuminati members will retain the knowledge of the previous time, and the rest of humanity will have no recollection of such a memory.

The Illuminati is **not** a joke. It is **not** fiction. They are very real, and part of humanity's reality; and through advanced concealed technologies the Illuminati aim to enslave humanity forever. The ring leaders of the present Illuminati includes Queen Elizabeth II, Prince Philip Duke of Edinburgh, Prince Charles of Wales and Vladimir Putin. I have also detailed the actions of the ringleaders in the main text in this disclosure, as well as, detailed other modern-day evil Illuminati members on my Facebook and Proboards. It is time the good people of earth, stopped being afraid, do the right thing, put a stop to this evil, and save themselves, as well as their children's children from being slaves forever. Spread and share this disclosure.

Glossary

Aneurysm -*noun:* An excessive swelling of the wall of an artery at a fixed point in the body. A brain aneurysm is therefore a: bulge or ballooning in a blood vessel in the brain. It often looks like a berry hanging on a stem. A brain aneurysm can leak or rapture, causing bleeding into the brain.

Brain Aneurysm -*noun:* see Aneurysm.

Clone –*noun:* a cell, group of cells, an organism produced asexually from a single ancestor and is genetically identical to a single ancestor.

Concealed Technology -*noun:* hidden machinery and devices undisclosed and currently unavailable for public consumption. Concealed (or military) technology develops at a rate of 44 years for every 12 months which passes in comparison to the technology the public is currently accustomed to. Origin: Phil Schneider.

Cloning Centre –*noun:* a physical location (on earth) where duplicate and replicate clones are produced. These physical locations are usually Deep Underground Military Bases (DUMBs). DUMBs have an entire floor dedicated to cloning. It is also a place where Illuminati members meet with each other as REM driven clones. The above-ground cloning centre where many high profile people attend can be found within a radius of 5/6 hours drive from the Robert Pickton Farm Port Coquitlam, British Columbia, in Canada, somewhere at a remote nature reserve.

Cloning Technology -*noun:* the technological advancements in medicine, science and technology used to produce duplicate and replicate copies of originals.

Conditioned Response -*noun:* an automatic response established by training to an ordinarily neutral stimulus.

Conditioned Stimulus –*noun:* A previously neutral stimulus that, after repeated association with an unconditioned stimulus, elicits the response produced by the unconditioned stimulus itself.

Consciousness –*noun:* the state of being aware of and responsive to one’s surroundings; a person’s awareness or perception of something.

Consciousness Transfer -*noun or verb:* the process of transferring or copying the mental content (including long-term memory and “self”) from a particular brain and copying it to a computational device; artificial body or avatar body such as that of a robot or clone version of the original. It is also the feat in which the person’s mental content (long term memory and “self”) moves from one body into another.

Depopulation *verb:* to remove or reduce the population of, as by destruction or force.

Drip Feed Disclosure *noun or verb:* is the process of supplying information but in small amounts overtime. Drip feed disclosure is also the process of revealing information slowly overtime, possibly telling lies to conceal certain aspects of the truth until the source administering the drip feed disclosure has adequate time to let out the truth in a slow and controlled way, thereby delaying the betrayed partner (in this disclosure, the public) from having the “complete truth” for some time.

DUMB *–noun:* [Deep Underground Military Base] a facility directly owned and operated by or for the military or one of its branches that shelters military equipment and personnel, and facilitates training and operations beneath the surface of the earth.

Duplicate Clone *-noun:* a fully formed human body which is a genetic copy of original developed through the process of regenerative technology. Duplicate clones are grown in a big thick tank full of (salty) water.

Duplication Cloning *-verb:* involves agitating the cells of on an original repetitively until a fully formed human body of the original is developed. Duplicate clones take an average of 5 months to form into a fully developed human body of the original through the process of regenerative medicine and technology.

Evaluative Conditioning *-noun:* is a change in liking, which occurs due to an association with a positive or negative stimulus.

H.A.A.R.P. Technology *–noun:* [High frequency Active Auroral Research Program] a radio transmitting system that can bounce signals off the Earth’s upper atmosphere, (60km (37 miles) to 1000km (620 miles) high) back to probe deep into the earth or sea. HAARP is also capable of: disrupting human mental processes; knocking out all global communication systems; changing weather patterns over large areas; interfering with wildlife and migration patterns; hurting ecosystems; negatively affect human beings health, moods and mental states; and unnaturally ‘boil’ the earth’s upper atmosphere. HAARP used correctly will control the weather without any adverse effects.

H.A.A.R.P. Grid *–noun:* a network of radio transmitters which can bounce signals off the earth’s upper surface. Each transmitter is located at a specific point across earth and communicates in unison with other radio transmitters across the earth. At this present time of writing, a HAARP grid has **not** been completed, although the Illuminati are working twice as fast to complete a HAARP grid. The threat to humanity once a HAARP grid is completed includes: mind control over the entire world’s inhabitants. A completed HAARP grid will also be capable of time travel, and therefore the Illuminati will always be able to go back to a pervious point in time to correct the mistake(s) which led to their downfall. Humanity will be slaves forever.

Habeas Corpus *-noun:* is [A] writ [formal document] requiring a person under arrest to be brought before a judge or into court, especially to secure the person's release unless lawful grounds are shown for their detention.

Heart Attack *-noun:* A sudden occurrence of a blockage of the flow of blood to the heart.

Human Clone *–noun:* The creation of a genetically identical copy of a human.

Ionosphere *–noun:* the layer of the earth's atmosphere which contains a high concentration of ions and free electrons and is able to reflect radio waves. It lies above the mesosphere and extends from about 60km (37 miles) to 1,000 km (620 miles) above the earth's surface.

Illuminati *–noun:* A modern-day criminal organisation operated by reprobate (depraved, unprincipled and wicked person) criminals. Their main agenda: is to enslave the whole world through advanced concealed technologies. See Appendix C for further discussion.

Mark 2 Clone *–noun:* Is a sleep driven clone; specifically, a REM sleep driven clone. A Mark 2 Clone is activated by transferring the consciousness of an original into a Mark 2 Clone **only when** the original reaches REM sleep (usually 90- 110 minutes **after** the original falls asleep). Once the consciousness of the original is transferred from the original's body to the Mark 2 Clone, the Mark 2 clone is now capable of motion: such as walking, talking etc. The Mark 2 clone 'drops limp' and becomes motionless once the original wakes up from sleep. The original's consciousness no longer resides in the Mark 2 clone (once the original is awake) and therefore the Mark 2 clone is now incapable of motion. Mark 2 Clones are also known as "REM Driven Clones" and "REM Duplicate Clones".

Memory Suppression *–noun or verb:* is the selective removal of memories or associations with the mind using memory suppression technology.

Memory Suppression Technology *–noun:* any scientifically advanced technology used selectively to remove memories from the conscious mind.

Mind-Voice Technology *–noun:* an advanced technology capable of reading, listening, hearing and broadcasting a person's inner voice or thoughts. It is capable of replicating sounds **exactly**. Therefore an individual can hear the sound of drums, a guitar or any instrument and replicate that sound exactly just by thinking about it. Consequently, Mind-voice technology has the functionality of producing music.

Military Technology *–noun:* machinery and devices developed from scientific knowledge used by the Armed Forces which advance at a rate of 44 years for every 12 months which passes, compared to the technology the public is accustomed to. Origin: Phil Schneider

MK Ultra *-noun:* [Manufacturing Killers Utilizing Lethal Tradecraft Requiring Assassinations] The goal of mind control, using MK Ultra technology is to program an individual to carry out any task against their will and self-preservation instinct and to control the absolute behaviour and thought patterns of the individual. See Appendix B for further details.

Neutral Stimulus *–noun:* is a stimulus which initially produces no specific response other than focusing attention. In classical conditioning, when used together with an unconditioned stimulus, the neutral stimulus becomes a conditioned stimulus.

Negative Association *-noun:* is an undesirable experience or perception.

Negative Stimulus *-noun:* a stimulus with undesirable consequences.

New World Order *-noun:* [NWO] Agenda. The whole NWO agenda is to turn humanity into mindless slaves forever; whereby the post-humans / trans-humans mind control the entire world's populace either through RFID microchips or a completed HAARP grid. Another aspect of the NWO agenda is to depopulate the world's current population of 7.3 billion people to 500 million people (and never exceed a world population of 500 million people afterwards); ruled by a one world government; a one world ruler; with a one world religion. See Appendix C for more details.

Original *–noun:* A person who is **not** a clone.

Pain Receptor *-noun:* Any one of the many nerve endings throughout the body that warn of harmful changes in the environment such as excessive pressure or temperature.

Positive Association *-noun:* is a desirable experience or perception.

Positive stimulus *-noun:* a stimulus with desirable consequences.

Posthumanism *–noun:* seeks to rewrite the very definition of being human. It is the condition in which humans and intelligent technology become intertwined. In the Posthuman there are no essential differences or absolute demarcations between bodily existence and computer stimulation, cybernetic mechanism and biological organism, robot technology and human goals.

Posthuman *-noun:* see Posthumanism.

Project MK Ultra *-noun:* see MK Ultra.

Regenerative Medicine *-noun:* (of a living organism) the process of re-growing new tissues after loss or damage.

Regenerative Technology *-noun:* any machinery or device developed from scientific knowledge which has the capability to re-grow new tissues after loss or damage.

REM Sleep *–noun:* [Rapid Eye Movement] is the fifth stage of sleep in the sleep cycle. It takes 90 - 110 minutes to reach REM sleep after we fall asleep. REM sleep is also known as the "period of paralysation". The involuntary muscles such as the brain become more active whereas voluntary muscles (those that you move by choice) such as your arms and legs become more relaxed or paralysed. REM sleep is a kind of sleep that occurs at intervals during sleep, and it is characterised by rapid eye movements.

REM Driven Clone *-noun:* [Rapid Eye Movement Driven Clone] a clone that can only become activated, once the original is in REM sleep. See Mark 2 Clone.

REM Duplicate Clone *-noun:* [Rapid Eye Movement Driven Clone] A clone developed by regenerative medicine and technology and is therefore an identical copy of an original. REM duplicate clones can only become activated when the original is in REM sleep. See Mark 2 Clone.

REM Driven Clone Death *–noun:* the process where an original dies because of constant torture to their REM driven clone or where a constant electrical current is applied to the REM driven clone resulting in death of an original usually in the form of an aneurysm or heart attack (because consciousness is linked) in the original's body.

REM Driven Clone Torture *—noun or verb:* the action or practice of inflicting serve pain on a REM driven clone. REM driven clone torture causes biological and physiological responses in the original's body because consciousness is linked. Intermittent REM driven clone torture (depending on what is done) causes the original to experience severe headaches, an upset stomach, achy limbs, sickness; a weakened heart. Continuous REM driven clone torture will lead to the death of the original; usually in the form of aneurysm or heart attack in the original's body.

Replication Cloning *–verb:* involves giving birth to a genetic identical of an original where the newborn starts life off as a baby and matures. The newborn is referred to as a clone.

RFID Technology [Radio Frequency Identification] *-noun:* are electronic microchips the size of a grain of sand that can be directly embedded into the human flesh. RFID microchips communicate wirelessly through the use of electromagnetic fields to transfer data. RFID microchips link the brains of people via the implanted microchip to satellites controlled by ground base super-computers. The dangers of RFID microchips to the implanted person are: total loss of privacy and total control of the person's physical body functions, mental and emotional thought processes, including the implanted person's subconscious or dreams –for the rest of that person's life! RFID microchips are also tracking devices, and the implanted person can be tracked anywhere on the globe.

Selling One's Soul *-verb:* to sell the use one's "Mark 2" REM driven clone to the Illuminati, for the Illuminati to use the individual's Mark 2 REM driven clone in whatever manner the Illuminati wishes. There are no returns once the individual has signed over his / her (soul) Mark 2 REM driven clone. When an individual sells their (soul) Mark 2 REM driven clone, the person has also entered into a contract to sell the Mark 2 REM driven clone(s) of their current children (if they have any) as well as any unborn children the person may have later in life. The person sells all their descendants (souls) Mark 2 REM driven clones to the Illuminati, once the individual sells their (soul) Mark 2 REM driven clone to the Illuminati. Selling one's (soul) Mark 2 REM driven clone is considered **a serious business transaction** to the Illuminati. There are no returns. If the person ever makes a fuss and wants their (soul) Mark 2 REM driven clone back, the Illuminati will either torture the person's Mark 2 Clone, or apply a constant electric current to the person's Mark 2 Clone until the person either has a heart attack or aneurysm in their original body. This is what public figures are hinting at when they say "They have sold their soul". They have sold the use of their Mark 2 REM driven clone to the Illuminati. Selling one's soul is not a joke. Never sell your soul.

Stimulus *–noun:* is something that causes a reaction, especially interest, excitement or energy. It is also an energy change registered by the senses. For example a stimulus can be a shinny object for a baby.

Technology *–noun:* machinery and devices developed from scientific knowledge.

Technological Advancement *–noun:* is incorporating, by means of experimental development, a characteristic or capability not previously existing or available in standard practice, into a new or existing process or product that enhances a product's performance. Novelty, uniqueness, or innovation alone does not indicate a technological advancement.

Transhumanism *–noun:* the belief or theory that the human race can evolve beyond its current physical and mental limitations, especially by means of science and technology.

Unconditioned Response *–noun:* is a response to a neutral stimulus we have no / little control over. It is a natural automatic response. For example, food is an unconditioned stimulus for a hungry animal, and salivation is the unconditioned response.

Unconditioned Stimulus *–noun:* A stimulus that elicits an unconditioned response.

All references in this document have been saved and backed-up; therefore if any link in this document, is ever deleted, modified etc. (online), let me know.

References

2045 Initiative (2015) *2045 Strategic Social Initiative* [Online] Available from: http://2045.com/ [Accessed 1st June 2015]

Albrecht, K., and McIntyre (2005) *Spychips: How Major Corporations and Government Plan to Track Your Every Move with RFID.* Plume

Alford, J., (2014) *Scientist Selectively Erase and Restore Memories* [Online] Available from: http://www.iflscience.com/brain/scientists-selectively-erase-and-restore-memories [Accessed 8th July 2015]

Alford, J., (2015) *Scientist Implant False Memories Into Sleeping Mice* [Online] Available from: http://www.iflscience.com/brain/scientists-implant-false-memories-sleeping-mice [Accessed 28th July 2015]

Animal Research (1996) *Cloning Dolly the Sheep* [Online] Available from: http://www.animalresearch.info/en/medical-advances/timeline/cloning-dolly-the-sheep/ [Accessed 28th June 2014]

Anthony, S., (2012) *GoFlow: a DIY tDCS brain-boosting kit* [Online] Available from: http://www.extremetech.com/extreme/121861-goflow-a-diy-tdcs-brain-boosting-kit [Accessed 24th May 2015]

Anthony, S., (2013) *What is transhumanism, or, what does it mean to be human?* [Online] Available from: http://www.extremetech.com/extreme/152240-what-is-transhumanism-or-what-does-it-mean-to-be-human [Accessed 24th May 2015]

Astral 7sight (2013a) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 1/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=UonnFuHLJKc&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral 7sight (2013b) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 2/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=QPAXCwu5MIo&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral 7sight (2013c) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 3/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=5zlJ0VQP444&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013d) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 4/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=dCGhrDEl-q8&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013e) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 5/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=EDyerA-k8Ic&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013f) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 6/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=iqQZPpXl2yg&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013g) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 7/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=WYZSOnyWwP8&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013h) *Illuminati Exposed 2013: Chemtrails, Atlantis, Clones, Drones & Vril Part 8/8*. [Online video]. June 23rd 2013. Available from: https://www.youtube.com/watch?v=Y18m0gPLQhM&ab_channel=Astral7ight [Accessed 12th July 2015]

Astral7sight (2013i) *Celebrity Tila Tequila: Missing Children and Cloning Centers* [Online video]. June 7th 2013. Available from: https://www.youtube.com/watch?v=7mRZ7ItF9ls&ab_channel=Astral7ight [Accessed 12th July 2015]

Avatar (2009) Film. Directed by James Cameron. [DVD]. UK: 20th Century Fox

BBC Future (2015) *BBC Future* [Online] Available from: http://www.bbc.com/future [Accessed 28th July 2015]

BBC Horizon (2009) *Why Do We Dream BBC Horizon* [Online video]. May 15th 2013. Available from: https://www.youtube.com/watch?v=E8MZ1twv0cU&ab_channel=AHOlearning [Accessed: 7th June 2015]

BBC News (2000) *Scientist 'clone' monkey* [Online] Available from: http://news.bbc.co.uk/1/hi/sci/tech/602027.stm [Accessed: 28th June 2015]

BBC News (2015a) *BBC News* [Online] Available from: http://www.bbc.co.uk/news/science_and_environment [Accessed 28th July 2015]

BBC News (2015b) *BBC News* [Online] Available from: http://www.bbc.co.uk/news/technology [Accessed 28th July]

BEAMS (2007) *BT's 'Soul Catcher2025' Implants* [Online] Available from: http://www.beamsinvestigations.org/BT's%20'Soul%20Catcher%202025'%20Implants.htm [Accessed 1st June 2015]

Begich, N., & Manning, J., (1997) *ANGELS DON'T PLAY THIS HAARP. Advances in Tesla Technology* [Online] Available from: http://www.alachuacounty.us/Depts/epd/EPAC/Angels%20Dont%20Play%20This%20HAARP%20by%20Nick%20Begich%201997.pdf [Accessed 15th June 2015]

Berkeley News (2011) *Scientists use brain imaging to reveal the movies in our mind* [Online] Available from: http://news.berkeley.edu/2011/09/22/brain-movies/ [Accessed: 22nd July 2015]

Beter, P., (2011) *"Organic Robotoids are real" by Dr Peter Beter* [Online video]. April 15th 2011. Available from: https://www.youtube.com/watch?v=nc0m5UMPwtU&ab_channel=youlittlerocket [Accessed 6th July 2015]

Bloomberg Business (2015) *See Future of Artificial Intelligence in Mind Clones Right Now!* [Online video] Available from: https://www.youtube.com/watch?v=4bqZp9TPYVk&ab_channel=BloombergBusiness [Accessed 27th June 2015]

Borghino, D., (2012) *"Avatar" project aims for human immortality by 2045* [Online] Available from: http://www.gizmag.com/avatar-project-2045/23454/ [Accessed 3rd June 2015]

Boringest (2006) *Bush Video 10 years ago!* [Online video]. January 3rd 2006. Available from: https://www.youtube.com/watch?v=pw4Bhmm22xo&ab_channel=boringest [Accessed: 11th May 2015]

BritneySpearsVevo (2009) *Britney Spears –Break The Ice* [Online video] October 24th 2009. Available from: https://www.youtube.com/watch?v=eQFIKP9rGhQ&ab_channel=BritneySpearsVEVO [Accessed 18th July 2015]

BritneySpearsVevo (2011) *Britney Spears –Hold it Against me* [Online video] February 17th 2011. Available from: https://www.youtube.com/watch?v=-Edv8Onsrgg&ab_channel=BritneySpearsVEVO [Accessed 18th July 2015]

Campbell, T., (2008) *Physics, Metaphysics and the nature of Consciousness* [Dr Thomas Campbell –My Big TOE (1 of 18)] [Online video] May 25th 2008. Available from: https://www.youtube.com/watch?v=MxECb7zcQhQ&list=PLBFFCEB1CAEDF9E6C&ab_channel=akn0ledge [Accessed 4th July 2015]

Carmichael, J., (2013) *Mouse Cloned From A Mere Drop of Blood* [Online] Available from: http://www.popsci.com/science/article/2013-06/mouse-cloned-mere-drop-blood [Accessed 19th July 2015]

CBS (2008) *Regeneration of Cells – Regrowing finger* [Online video] May 15th 2010. Available from: https://www.youtube.com/watch?v=ITxx2sOLW2Y&ab_channel=Weltenspur2 [Accessed: 25th June 2015]

Cherry, K., (2015) *What is Consciousness* [Online] Available from: http://psychology.about.com/od/statesofconsciousness/f/consciousness.htm [Accessed 1st July 2015]

CreativeForceVideo (2014) *Speaking with Nostradamus, Quantum Hypnosis with Delores Cannon* [Online video] August 2nd 2014. Available from: https://www.youtube.com/watch?v=CBOml6LTbpM&ab_channel=CreativeForceVideo [Accessed 25th July 2015]

Crystal Links (2015) *Nostradamus* [Online] Available from: http://www.crystalinks.com/nostradamus.html [Accessed 25th July 2015]

CTForecaster (2013) *Japanese Dream Recording Machine –Update* [Online video]. April 12th 2013. Available from: https://www.youtube.com/watch?v=gQueU9a8URw&ab_channel=CTForecaster [Accessed: 22nd July 2015]

Daily Mail Online (2015) *Mail Online* [Online] Available from: http://www.dailymail.co.uk/sciencetech/index.html [Accessed 28th July 2015]

De Houwer, J., Thomas, S., & Baeyens, F. (2001) *Associative Learning of Likes and Dislikes: A Review of 25 years of Research on Human Evaluative Conditioning*. Psychological Bulletin, Vol. 127, No.6, 853-869 In Hale, J., (2012) *What influences are Food Likes and Dislikes?* [Online] Available from: http://psychcentral.com/blog/archives/2012/02/15/what-influences-our-food-likes-and-dislikes/ [Accessed 28th June 2015]

Dictionary Reference (2015) *Consciousness* [Online] Available from: http://dictionary.reference.com/browse/consciousness?s=t [Accessed 30th June 2015]

Disclose TruthTV (2015) *Delores Cannon on WW3, Antichrist Nostradamus Prophecies [Full Video]* [Online video] March 19th 2015. Available from: https://www.youtube.com/watch?v=l0aCHrY0ObI&ab_channel=DiscloseTruthTV [Accessed 26th July 2015]

Donald Marshall Proboards (2015) *Donald Marshall Proboards* [Online] Available from: http://donaldmarshall.proboards.com/ [28th July 2015]

Ehrsson, H.H., (2013) *Inspirational Lecture –Professor Henrik Ehrsson* [Online video] October 3rd 2013. Available from: https://www.youtube.com/watch?v=iR7HissYN2U&ab_channel=karolinskainstitutet [Accessed 2nd July 2013]

Epperson, A. R., (1990) *The New World Order*, Publius Press. [Online] Available from: https://ia700406.us.archive.org/27/items/TheNewWorldOrder_342/TheNewWorldOrder.pdf [Accessed 7th July 2015]

Fox 4 News –Dallas Fort Worth (2014) *Aug. 15, 1988 – George W. Bush –KDFW* [Online video]. June 23rd 2014. Available from: https://www.youtube.com/watch?v=zwrl2axvPmY&ab_channel=FOX4News-Dallas-FortWorth [Accessed: 15th May 2015]

FW: Thinking (2014) *Erase and Restore your memories* [Online video]. Jun 18th 2014. Available from: https://www.youtube.com/watch?v=PpzzD_103jc&ab_channel=FW:Thinking [Accessed 8th July 2015]

Galactic (2012a) *Conversations With Nostradamus Volume One* [Online] Available from: http://galactic.no/rune/DoloresCannon_books/Dolores-Cannon-Conversations-With-Nostradamus-Volume-1.pdf [Accessed 25th July 2015]

Galactic (2012b) *Conversations With Nostradamus Volume Two* [Online] Available from: http://galactic.no/rune/DoloresCannon_books/Dolores-Cannon-Conversations-With-Nostradamus-V2.pdf [Accessed 25th July 2015]

Gizmag (2015) *Gizmag* [Online] Available from: http://www.gizmag.com/ [Accessed 28th July 2015]

Greenberg, P., (2013) *No More Bad Flashback: Scientist Find Gene That Erases Memories* [Online] Available from: http://mashable.com/2013/09/25/erase-bad-memories/ [Accessed 11th June 2015]

Hale, J., (2012) *What influences are Food Likes and Dislikes?* [Online] Available from: http://psychcentral.com/blog/archives/2012/02/15/what-influences-our-food-likes-and-dislikes/ [Accessed 28th June 2015]

Halliday, P., (2013) *Whistleblowing: the new 'public interest' test and other developments.* [Online] Available from: http://www.11kbw.com/uploads/files/PHPaper.pdf [Accessed 8th May 2015], p. 2

Hewitt, J., (2012) *How to create a mind or die trying* [Online] Available from: http://www.extremetech.com/extreme/141507-how-to-create-a-mind-or-die-trying [Accessed 24th May 2015]

History (2015) *George W. Bush* [Online] Available from: http://www.history.com/topics/us-presidents/george-w-bush [Accessed 14th May 2015]

Hostel (2006) Film. Directed by Eli Roth. [DVD]. USA: Lionsgate

Hostel: Part II (2007) Film. Directed by Eli Roth. [DVD]. USA: Lionsgate

Hostel: Part III (2011) Film. Directed by Scott Spiegel. [DVD]. USA: Sony Pictures Home Entertainment

Inception (2010) Film. Directed by Christopher Nolan. [DVD]. UK: Warner Bros. Pictures

inifiniLor (2013) *Appeal from Survivors of Canadian Genocide* [Online video]. April 8th 2013. Available from: https://www.youtube.com/watch?v=cVYkctM1k90&ab_channel=infiniLor [Accessed 13th July 2015]

Jeanice Barcelo (2013) *BNE Radio Show w/ guest Donald Marshall* [Online video]. March 9th 2013. Available from: https://www.youtube.com/watch?v=3uzgu4ekT3c&ab_channel=JeaniceBarcelo [Accessed 11th July 2015]

Jim Cristea (2009) *Mind Reading - FMRI - Machine that Reads Your Thoughts - 60 Minutes* [Online video]. January 9th 2009. Available from: https://www.youtube.com/watch?v=Cwda7YWK0WQ&ab_channel=JimCristea [Accessed: 22nd July 2015]

Jones, A., (2008) *Reflections and Warnings: An Interview with Aaron Russo* [Online video]. June 1st 2009. Available from:
https://www.youtube.com/watch?v=YGAaPjqdbgQ&ab_channel=rohstyles23
[Accessed: 12th May 2015]

KafkaWinstonWorld (2014) *IS THIS WHERE WE ARE GOING? THIS MOVIE WILL BLOW YOUR F%$NG MIND(mirrored)* [Online video]. December 11th 2014. Available from:
https://www.youtube.com/watch?v=4mUII1HcsRg&ab_channel=KafkaWinstonWorld
[Accessed 15th July 2015]

Kim, M., (2013) *MIT scientists implant a false memory into a mouse's brain* [Online] Available from: http://www.washingtonpost.com/national/health-science/inception-mit-scientists-implant-a-false-memory-into-a-mouses-brain/2013/07/25/47bdee7a-f49a-11e2-a2f1-a7acf9bd5d3a_story.html [Accessed 28th July 2015]

LawOfIdentity (2014) *"We Will Become Gods"* [Online video]. August 20th 2014. Available from: https://www.youtube.com/watch?v=5MmXHMaati0&ab_channel=LawOfIdentity
[Accessed 23 July 2015]

Legal-dictionary (2015) *Habeas Corpus* [Online] Available from: http://legal-dictionary.thefreedictionary.com/habeas+corpus [Accessed 14th May 2015]

Mark Dice (2014) *Top Transhumanist Claims He Will Become God and Kill Anyone Who Tries to Stop Him!!!* [Online video] May 14th 2014. Available from:
https://www.youtube.com/watch?v=KPJARo-5VXE&ab_channel=MarkDice [Accessed 24th July 2015]

Medical News Today (2015) *Paranoid Schizophrenia: Causes, Symptoms and Treatments* [Online] Available from: http://www.medicalnewstoday.com/articles/192621.php [Accessed 21st July 2015]

Megadeth (2001) *The World Needs a Hero* [CD] USA: Sanctuary Records Group Ltd

MK Ultra Compendium (1980) *Secret CIA Human Experiments in the United States: MK Ultra Mind Control Research Program.* [online video] May 26th 2012. Available from:
https://www.youtube.com/watch?v=c4f9Hs0s1jQ&ab_channel=TheFilmArchives
[Accessed: 22nd June 2015]

Mind-Computer (2012) *Synthetic telepathy "Artificial Telepathy"* [Online] Available from:
http://mind-computer.com/2012/05/15/synthetic-telepathy-artificial-telepathy/ [Accessed 22nd June 2015]

Moss, S., (2009) *Evaluative conditioning* [online] Available from: http://www.psych-it.com.au/Psychlopedia/article.asp?id=312 [Accessed 28th June 2015]

Motherboard (2015) *Motherboard* [Online] Available from: http://motherboard.vice.com/en_uk [Accessed 28th July 2015]

MrCowShedder (2012) *Royal Babylon by Heathcote Williams (rough cut).* [Online video]. May 11th 2012. Available from: https://www.youtube.com/watch?v=jIukrdRhnpw&ab_channel=MrCowshedder [Accessed 14th July 2015]

MurdokDubstep (2010) *Where'd You Go – Fort Minor (Murdok Dubstep Remix)* [Online video]. July 22nd 2010. Available from: https://www.youtube.com/watch?v=7VdAvIf1Nc4&ab_channel=MurdokDubstep [Accessed 15th July 2015]

nature video (2013) *Reading minds* [Online video]. October 23rd 2013. Available from: https://www.youtube.com/watch?v=z8iEogscUl8&ab_channel=naturevideo [Accessed: 22nd July 2015]

New Scientist (2014) *Brain Decoder can eavesdrop on your inner voice* [Online] Available from: https://www.newscientist.com/article/mg22429934-000-brain-decoder-can-eavesdrop-on-your-inner-voice/ [Accessed 12th June 2015]

New Scientist (2015) *New Scientist* [Online] Available from: https://www.newscientist.com/ [Accessed 28th July 2015]

Non Mirage Truth Vision (2015) *CIA's Heart Attack Gun* [Online video]. Available from: https://www.youtube.com/watch?v=Uwy56QTV4cs&ab_channel=NonMirageTruthVision [Accessed 16th July 2015]

Open Minds (2011) *Phil Schneider's incredible ET claims.* [Online] Available from: http://www.openminds.tv/phil-schneiders-incredible-et-claims/9982 [Accessed 9th May 2015]

Paye, J-C., (2013) *The Suspension of Habeas Corpus in America* [Online] Available from: http://www.globalresearch.ca/the-suspension-of-habeas-corpus-in-america/5311701 [Accessed 20th May 2015]

Petkova, V. I., and Ehrsson, H.H., (2008) *If I Were You: Perceptual Illusion of Body Swapping*, PLoS ONE, Volume 3, Issue 12, pp. 1-9

Prigg, M., (2014) *Mindreading software could eavesdrop on your secret inner voice* [Online] Available from: *http://www.dailymail.co.uk/sciencetech/article-2814896/The-mindreading-machine-listen-voices-head-let-paralysed-speak-again.html* [Accessed 13th June 2015]

Project Camelot (2008a) *Project Camelot Interviews George Green – Part 1 of 2* [Online video]. April 16th 2008. Available from: https://www.youtube.com/watch?v=sSYXrWIA618&ab_channel=ProjectCamelot [Accessed: 11th May 2015]

Project Camelot (2008b) *Project Camelot Interviews George Green – Part 2 of 2* [Online video]. April 16th 2008. Available from: https://www.youtube.com/watch?v=6zSrg0IxHzI&ab_channel=ProjectCamelot [Accessed: 11th May 2015]

RainmanJhof (2011) *Boondox-Abadon* [Online video]. October 31st 2011. Available from: https://www.youtube.com/watch?v=C2bH4B5_83Q&ab_channel=RainmanJhof [Accessed 15th July]

Radar Online (2013) *Princess Dianna 'Killed By Bright Light Shone By Special Forces Soldiers Into Car She Was In'* [Online] Available from: http://radaronline.com/exclusives/2013/09/princess-diana-killed-died-death-forces-light-special-air-service/ [Accessed 17th July 2015]

Rense (2001) *Microchip Mind Control, Implants and Cybernetics* [Online] Available from: http://www.rense.com/general17/imp.htm [Accessed 20th June 2015]

Rense (2011) *Possible HAARP Locations Around the World* [Online] Available from: http://www.rense.com/general92/haarp.htm [Accessed 20th July 2015]

Russo, A., (1996) *Mad as Hell* [Online video]. May 10th 2011. Available from: https://www.youtube.com/watch?v=Zz4TI75MszQ&ab_channel=XRepublicTV [Accessed: 13th May 2015]

Russo, A., (2006) *Freedom to Fascism* [Online video]. May 10th 2011. Available from: https://www.youtube.com/watch?v=uNNeVu8wUak&ab_channel=RevolutionistsUnited [Accessed: 12th May 2015]

RT (2014) *Revolutionary way to 'switch off' pain discovered* [Online] Available from: http://www.rt.com/news/210031-revolutionary-painkiller-discovered-scientists/ [Accessed 19th July 2015]

RT (2015) *Living forever as a robot? Prototype lets humans upload their mind into mechanised 'heads'* [Online] Available from: http://www.rt.com/usa/229811-mind-clones-robot-afterlife/ [Accessed 27th June 2015]

SadSongs4You (2010) *Elton John –Candle In The Wind. With lyrics* [Online video]. August 23rd 2010. Available from: https://www.youtube.com/watch?v=80rHyABCb20&ab_channel=SadSongs4You [Accessed 17th July 2015]

Schechtman (2012) *The Story of my (Second) Life: Virtual Worlds and Narrative Identity,* Philosophy and Technology, Volume 25, Issue 3, pp. 329-343

Schneider, P., (1995) *Phil Schneider Documentary of truth about Aliens & UFO's & our Government*. [Online video]. September 21st 2013. Available from: https://www.youtube.com/watch?v=Oljrjxnixtw&ab_channel=AliensAmongUs [Accessed: 10th May 2015]

Schneider, P., (1996) *Phil Schneider's Last Speech ~ Two Months Before His Assassination ~ Aliens & Underground Bases* [Online video]. November 24th 2013. Available from: https://www.youtube.com/watch?v=Slgb5U-OqFM&ab_channel=FallofMedia [Accessed: 10th May 2015]

Science Channel (2014) *How to Grow a New Fingertip | World's Strangest*. [Online video] June 16th 2014. Available from: https://www.youtube.com/watch?v=DtBUM51t4iw&ab_channel=ScienceChannel [Accessed 23rd June 2015]

Science Daily (2014) *'Off switch' for pain discovered: Activating the adenosine A3 receptor subtype is key to powerful pain relief* [Online] Available from: http://www.sciencedaily.com/releases/2014/11/141126132639.htm [Accessed 19th July 2015]

Sheen, M., Begich, N., & Robbins, W., (2005) *Holes in Heaven? H.A.A.R.P. & Advances in Tesla Technology* [Online video]. October 8th 2014. Available from: https://www.youtube.com/watch?v=SWVU6DKcjyA&ab_channel=documentary2014 [Accessed: 15th June 2015]

SimpleGirl4ewer (2007) *Britney Spears – Mona Lisa* [Online video]. March 18th 2007. Available from: https://www.youtube.com/watch?v=xEecXKUxl1s&ab_channel=SimpleGirl4ewer [Accessed 18th July 2015]

Sleepdex (2015) *Stages of Sleep* [Online] Available from: http://www.sleepdex.org/stages.htm [Accessed: 5th June 2015]

Star Wars Episode III: Revenge of the Sith (2005) Film. Directed by George Lucas. [DVD]. USA: 20 Century Fox

Stromberg, J., (2013) *Scientist Figure Out What You See While You're Dreaming* [Online] Available from: http://www.smithsonianmag.com/science-nature/scientists-figure-out-what-you-see-while-youre-dreaming-15553304/?no-ist [Accessed 22nd July 2015]

The 6th Day (2000) Film. Directed by Roger Spottiswoode. [DVD]. USA: Columbia Pictures

The Guardian (2015a) *The Guardian* [Online] Available from: http://www.theguardian.com/science [Accessed 28th July 2015]

The Guardian (2015b) *The Guardian* [Online] Available from: http://www.theguardian.com/uk/technology [Accessed 28th July 2015]

The Independent (2015a) The Independent [Online] Available from: http://www.independent.co.uk/news/science/ [Accessed 28th July 2015]

The Independent (2015b) The Independent [Online] Available from: http://www.independent.co.uk/life-style/gadgets-and-tech/ [Accessed 28th July 2015]

The Island (2005) Film. Directed by Michael Bay. [DVD]. USA: DreamWorks Pictures

The Manchurian Candidate (2004) Film. Directed by Jonathan Demme. [DVD]. USA: Paramount Pictures

TheNanoAge (2015) *Transhumanism and Posthumanism: The Future of Us (Humanity Plus)* [Online] Available from: *http://www.thenanoage.com/transhumanism-posthumanism.htm#transhuman* [Accessed: 30th May 2015]

Truthstream (2006) *Rigged USA Elections Exposed.* March 2nd 2006. Available from: https://www.youtube.com/watch?v=JEzY2tnwExs&ab_channel=truthstream [Accessed: 22th May 2015]

UC Berkeley Campus Life (2011) *Vision Reconstruction* [Online video]. December 11th 2011 Available from: https://www.youtube.com/watch?v=6FsH7RK1S2E&feature=youtu.be&ab_channel=UCBerkeleyCampusLife [Accessed 22nd July 2015]

Vincent Eastwood (2013) *Illuminati Cloning Programs, Sex and Murder Cults and Reptilians! 26Feb2013* [Online video] February 26th 2013. Available from: https://www.youtube.com/watch?v=M_1UiFeV5Jg&ab_channel=VincentEastwood [10th July 2015]

Walcutt, D. L., (2013) *Stages of Sleep*. [Online] Available from: http://psychcentral.com/lib/stages-of-sleep/ [Accessed 5th June 2015]

Walther, E., Nagengast, B., & Trasselli, C., (2005). *Evaluative conditioning in social psychology: Facts and speculations.* Cognition and Emotion, 19, 175–196 In Moss, S., (2009) *Evaluative conditioning* [online] Available from: http://www.psych-it.com.au/Psychlopedia/article.asp?id=312 [Accessed 28th June 2015]

WhiteLiesVevo (2010) *White Lies – Bigger Than Us* [Online video]. November 18th 2010 Available from: https://www.youtube.com/watch?v=JW0yynlDmqQ&ab_channel=WhiteLiesVEVO [Accessed 18th July]

Winter, L., (2014) *Scientist Erase Memories with Light* [Online] Available from: http://www.iflscience.com/brain/scientists-erased-memories-light [Accessed 10th June 2015]

WorldTruth (2014) *The Georgia Guidestones* [Online] Available from: http://worldtruth.tv/the-georgia-guidestones-2/ [Accessed 16th May 2015]

YesEthan (2013) *Consciousness Science Kept Hidden* [Online Video] January 12th 2013. Available from: https://www.youtube.com/watch?v=LFSRTsLOiv0&ab_channel=YesEthan [Accessed 4th July 2015]

Legislation

Computer Misuse Act (1990) Section 3A, *Making, supplying or obtaining articles for use in offence under section 1 or 3* [Online] Available from: http://www.legislation.gov.uk/ukpga/1990/18/section/3A [Accessed 8th May 2015]

Enterprise and Regulatory Reform Act (2013) Section 17, *Disclosures not protected unless believed to be made in the public interest.* [Online] Available from: http://www.legislation.gov.uk/ukpga/2013/24/section/17/enacted [Accessed 8th May 2015]

Employment Rights Act (1996) Section 43B, *Disclosures qualifying for protection.* [Online] Available from: http://www.legislation.gov.uk/ukpga/1996/18/section/43B [Accessed 8th May 2015]

Public Interest Disclosure Act (1998) Section 43B, *Disclosures qualifying for protection.* [Online] Available from: http://www.legislation.gov.uk/ukpga/1998/23/section/1 [Accessed 8th May 2015]

Serious Crime Act (2015) Section 41 3ZA, *Unauthorised acts causing, or creating risk of, serious damage.* [Online] Available from: http://www.legislation.gov.uk/ukpga/2015/9/section/41/enacted [Accessed 8th May 2015]

Tetryonic Theory - an overview (1) 2
Tetryonic theory..... challenging and correcting the foundations of quantum th 13
tetryonics merged pdf 21
Response to an analysis of Tetryonic Chemistry by Robert J Brotherus (1) 1675
Planck mass-energy templates [Tetryonic theory] 1694

Tetryonic Theory

The unified quantum field geometry of charged EM mass-ENERGY-Matter in motion

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them." *[Albert Einstein]*

Given the range and applied scope of Quantum Physics in today's modern technologic World there remains a driving desire to rationalise our numerous disparate scientific theories into one coherent theory with a intuitive model that can be applied equally to the Quantum and Cosmological scales of our Universe.

Such a theory would need to preserve the currently observed outcomes and present established theories in a new light, offering additional testable predictions of its own, and ideally do so in a manner that is simpler than that of the established theories and hypothesises.

Many foundational properties of Quantum Mechanics remain unaddressed by scientific theory and in the following pages an overview of the key quantum properties challenging our current scientific advancement will be highlighted, including a number of assumptions that currently impede the development of a fully realised, coherent solution to all of our current scientific questions.

While Mathematics is the language of Science it remains a language that lacks a well-defined physical model on which to test it and further its many and varied solutions to Quantum & Cosmological scale physics. It is this lack of any rigid, enforceable GEOMETRY that has allowed the development of numerous disjointed statistical and probabilistic solutions to physical problems, in turn impeding our scientific understanding and advancement of quantum processes.

The Standard Model has many observed and testable components to it but more recently new theories have emerged to contest it without being rigorously testable themselves. They rely on the established foundation provided by the Standard Model but try to explain its various deficiencies ad-hoc with various suppositions without any solid footing of their own.

Often the only way to progress further in our scientific endeavours is to retrace our footsteps in science and to develop new physical models on which we can discern our known results and observations thus excluding any false or misleading assumptions (mathematical of otherwise).

In doing so there exists the promise that a simple underlying foundation can be found to the physics of our Universe, revealing new and exciting advances in Science that will allow us to usher in a new age of scientific and technological advancement for the betterment of humanity as a whole.

'Tetryonics – The charged geometry of EM mass-ENERGY-Matter' whose founding principle is that **EQUILATERAL Planck energies are** the foundation geometry for all quantum mechanical processes is presented here as just such a solution to the current quandaries of Quantum Mechanics.

ENERGY

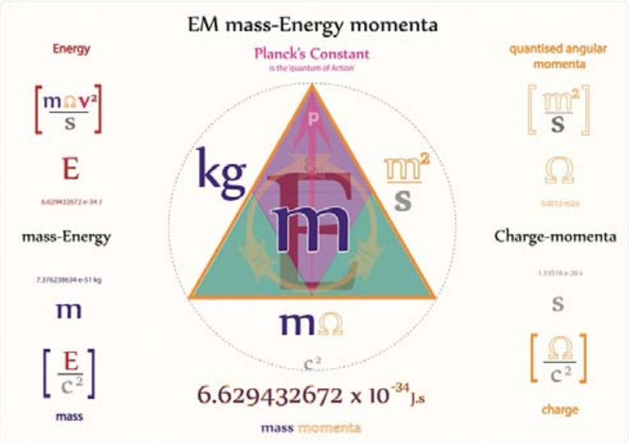
EM mass-Energy momenta
Energy
Planck's Constant
is the 'quantum of Action'
quantised angular momenta
mΩv²/s
E
kg
m²/s
Ω
mass-Energy
m
Charge-momenta
p
s
m
mΩ
c²
E/c²
Ω/c²
6.629432672 x 10-34 J.s
mass
charge
mass momenta

Energy, in Physics, is an indirectly observed quantity of a system that imbues it with the ability to exert a Force or do Work over a distance.

It has been measured and quantified through numerous methodologies over the centuries, most notably through its associated characteristics such as mass, velocity and ElectroMagnetic fields.

The most recent attempt to quantify its characteristics (with respect to heat and light) led to the discovery of Planck's constant and the development of Quantum Mechanics

The application of a simple proposition (that Energy has an equilateral geometry) opens the door on a greater understanding of the mechanics of the quantum world, a realm that will be forever beyond the reach of our physical eyes.

The myriad of perplexing properties of quantum properties (such as Charge, mass & Matter) and the astonishing outcomes of quantum experiments (Interference and Wave-Particle duality) can now all be readily modelled and explained rationally on a solid geometric footing.

Paving the way for new discoveries and a greater understanding of our Universe and its mechanics

Quantised Angular Momenta

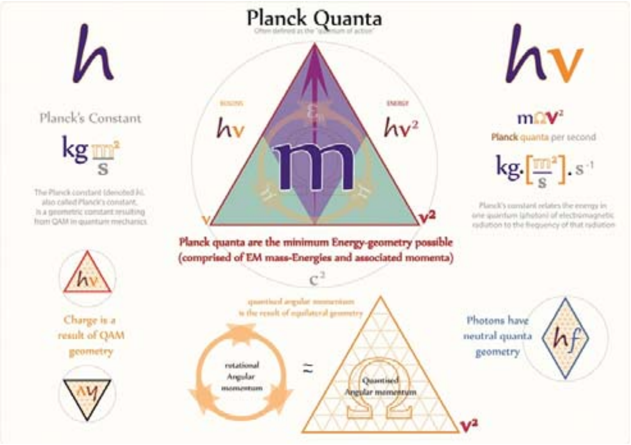
Planck Quanta
h
hv
Planck's Constant
kg m²/s
hv
hv²
m
mΩv²
Planck quanta per second
kg.[m²/s].s-1
v
v²
Planck quanta are the minimum Energy-geometry possible (comprised of EM mass-Energies and associated momenta)
c²
hv
Charge is a result of QAM geometry
rotational Angular momentum
≈
Quantised Angular momentum
Ω
v²
Photons have neutral quanta geometry
hf

Building on the priori definition of equilateral energy quantisation it can be quickly demonstrated that any energy measurement per unit of Time results in a geometric unit of measurement that has historically be taken to represent rotational motion (specifically Angular momentum)

This unit of motion is found throughout Quantum Mechanics and is directly related to the square energy levels in nuclear processes.

The equilateral EM fields of energy quanta are constrained by its geometry and this geometry lies at the heart of understanding the quantum world in all its beauty.

It determines properties such as Electric permittivity and Magnetic permeability, the vector direction of linear momentum and relates the scalar property of electromagnetic mass directly to velocity.

CHARGE

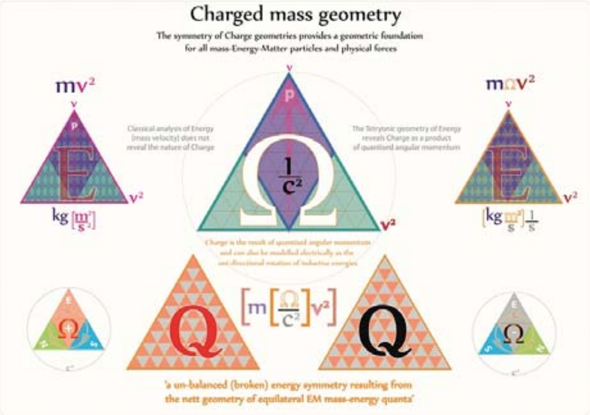
Charged mass geometry
The symmetry of Charge geometries provides a geometric foundation for all mass-Energy-Matter particles and physical forces
mv²
kg [m²/s²]
mΩv²
[kg m²/s] 1/s
E
Ω
1/c²
v
v²
p
Classical analysis of Energy (mass velocity) does not reveal the nature of Charge
The Tetryonic geometry of Energy reveals Charge as a product of quantised angular momentum
Charge is the result of quantised angular momentum and can also be modelled electrically as the anti-directional rotation of inductive energies
Q
[m[Ω/c²]v²]
Q
'a un-balanced (broken) energy symmetry resulting from the nett geometry of equilateral EM mass-energy quanta'

A fundemental property of all mass-ENERGY-Matter that gives rise to the Forces of electrical Interaction has eluded physical explanation by Physicists since its discovery and led to many 'band aid' additions to the Standard model in attempts to explain its observed properties.

The most recent attempt at an explanation was Special relativity theory which in turn led to our current models of charge from contractions of Matter due to motion.

SR explains Charge as an invariant property of of electrostatic bodies and that the motion of charges creates an additional magnetic moment through the relativistic distortion of spherical bodies of Matter.

Equilateral Planck energy momenta geometries and Matter topologies offer a completely different explanation for the source of electrostatic charges and their associated magnetic moments.

Positive and Negative charges are revealed to be opposite side of the same quantum energy 'quoin', and can be modelled electrically as ideal quantum inductive loops. It is equilateral geometry that gives rise to the physical properties of quantised angular momenta, inertial mass, elemental charges and even the geometric topology of Matter itself.

In a planar 2D form they form a neutral EM energy momenta geometry but it is when they form the topology of 3D Matter they are expressed as either Positive or Negative electric charge fascia.

ODD number energy quanta [W Bosons] combine via their magnetic bases in fixed quantum steps to create the 'squared' nuclear energy levels so familiar to quantum mechanics and form the basis for ElectroMagnetic Induction.

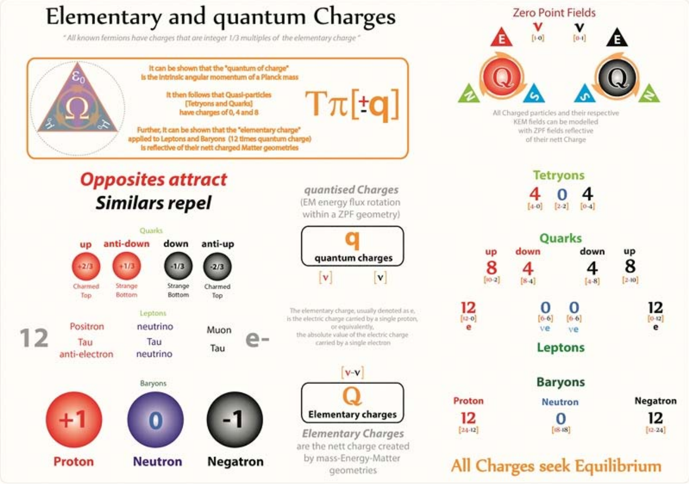
Elementary and quantum Charges
"All known fermions have charges that are integer 1/3 multiples of the elementary charge"
It can be shown that the "quantum of charge" is the intrinsic angular momentum of a Planck mass
It then follows that Quasi-particles [Tetryons and Quarks] have charges of 0, 4 and 8
Further, it can be shown that the "elementary charge" applied to Leptons and Baryons (12 times quantum charge) is reflective of their nett charged Matter geometries
Tπ[±q]
Opposites attract
Similars repel
Quarks
up
anti-down
down
anti-up
+2/3
+1/3
-1/3
-2/3
Charmed Top
Strange Bottom
Strange Bottom
Charmed Top
Leptons
12
Positron
Tau
anti-electron
neutrino
Tau
neutrino
Muon
Tau
e-
Baryons
+1
0
-1
Proton
Neutron
Negatron
quantised Charges
(EM energy flux rotation within a ZPF geometry)
q
quantum charges
[v]
[v]
The elementary charge, usually denoted as e, is the electric charge carried by a single proton, or equivalently, the absolute value of the electric charge carried by a single electron
[v-v]
Q
Elementary charges
Elementary Charges are the nett charge created by mass-Energy-Matter geometries
Zero Point Fields
v [1-0]
v [0-1]
E
Q
N
S
All Charged particles and their respective KEM fields can be modelled with ZPF fields reflective of their nett Charge
Tetryons
4 [4-0]
0 [2-2]
4 [0-4]
Quarks
up
8 [10-2]
down
4 [8-4]
down
4 [4-8]
up
8 [2-10]
12 [12-0] e
0 [6-6] ve
0 [6-6] ve
12 [0-12] e
Leptons
Baryons
Proton
12 [24-12]
Neutron
0 [18-18]
Negatron
12 [12-24]
All Charges seek Equilibrium

It is their rigid equilateral geometry that provides the basis for relativistic charge invariance and the quantum mechanism for photo-electron transitions in atoms while the net energy momenta quanta in charged geometries provides the electromotive force known as Voltage.

Equilateral geometry reveals an intrinsic connection between EM Energy, mass and Charge.

As separated charges seek equilibrium they provide motive energies & the quantum scaffolding for large scale Matter topologies and their force interactions throughout our Universe.

mass-ENERGY-Matter

All Energy has an EM mass equivalence and equally any object with EM mass has energy equivalence

But there has never been a formal scientific definition and associated equation to distinguish the property of EM mass from that of Matter.

Using Tetryonic geometry it can be clearly demonstrated that ElectroMagnetic mass is NOT Matter – it is the Energy content of a system per unit of time [E / c^2] in total agreement with Einstein (and all before him) however, lacking a formal definition of either the two terms have become mired with each other so much that they are often used for each other inappropriately even in scientific peer-reviewed literature.

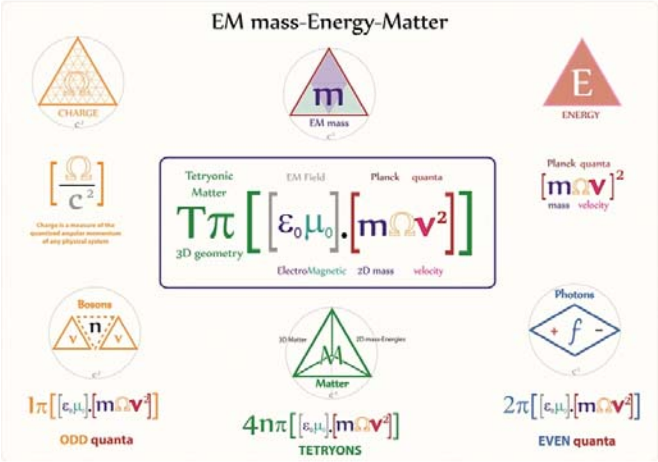
EM mass-Energy-Matter
CHARGE
c²
m
EM mass
c²
E
ENERGY
[Ω / c²]
Tetryonic
Matter
EM Field
Planck
quanta
Tπ[[ε₀μ₀].[mΩv²]]
3D geometry
ElectroMagnetic
2D mass
velocity
Planck quanta
[mΩv]²
mass
velocity
Bosons
n
v
v
c²
3D Matter
2D mass-Energies
M
Matter
c⁴
Photons
+ f -
c²
1π[[ε₀μ₀].[mΩv²]]
ODD quanta
4nπ[[ε₀μ₀].[mΩv²]]
TETRYONS
2π[[ε₀μ₀].[mΩv²]]
EVEN quanta

ElectroMagnetic mass is the two dimensional (planar) measurement of Energy per unit of Time whereas Matter is a measure of the tetrahedral energies found in a spherical volume of 3D space.

Mass-energy can be viewed as the paper from which 3D Matter is created when it is folded into a tetrahedral shape (the quantum canvas covering the topology of Matter)

The often used tem of 'massless' is now shown to be a misnomer that should be removed from the scientific vocabulary except where it specifically refers to empty space (devoid of any energies).

Matter is now formally defined geometrically as 4nπ tetrahedral standing wave of EM energy and it is the tetrahedral topology of Tetryons that forms the foundation for all large scale Matter in the Universe (not spherical point particles).

Charge allows EM masses to form a Tetryonic topology giving volume to Matter and its energy momenta provides the interactive forces between them imbuing motion into our Universe.

Matter at rest is comprised of EM mass-energies that are always in motion and propagating in a tetrahedral EM standing wave.

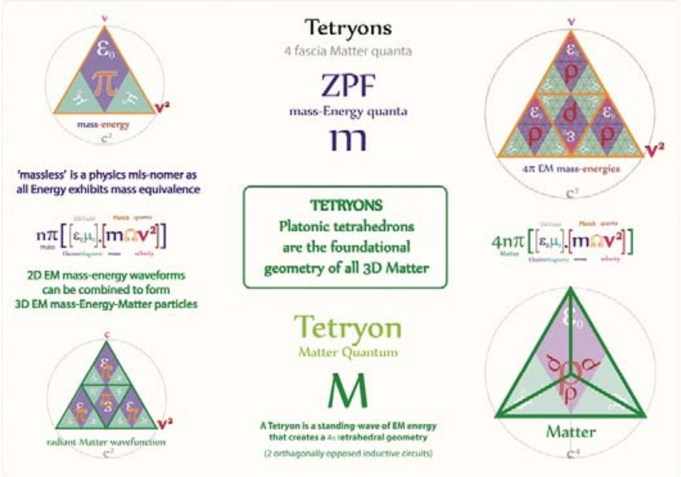
v
ε₀
π
v²
mass-energy
c²
Tetryons
4 fascia Matter quanta
ZPF
mass-Energy quanta
m
v
ε₀
ρ
d
3
v²
4π EM mass-energies
c²
'massless' is a physics mis-nomer as all Energy exhibits mass equivalence
nπ[[ε₀μ₀].[mΩv²]]
2D EM mass-energy waveforms can be combined to form 3D EM mass-Energy-Matter particles
TETRYONS
Platonic tetrahedrons are the foundational geometry of all 3D Matter
4nπ[[ε₀μ₀].[mΩv²]]
c
v²
radiant Matter wavefunction
c²
Tetryon
Matter Quantum
M
A Tetryon is a standing-wave of EM energy that creates a 4π tetrahedral geometry
(2 orthagonally opposed inductive circuits)
ε₀
ρ
Matter
c⁴

To cause at will the birth and death of matter would be man's grandest deed, which would give him the mastery of physical creation, make him fulfil his ultimate destiny."

[Nikola Tesla]

Sub-Atomic Particles

Using equilateral energy all the mass-energy geometries & mass-Matter topologies of EM fields and particles can be physically modelled revealing the known Standard Model particle sand a few more.

Bosons are shown to be transverse EM fields that facilitate EM induction and Photons are revealed as longitudinal dual-charge pairs possessing a neutral EM charge and inherent magnetic moment.

The net unidirectional momentum of Bosons then distinguishes them from Photons that possess a bi-directional momentum

Closer examination of the geometries highlights a long-standing error in the mathematical formulation of QM energy formulas, namely the mistaken interchanging of Planck's quanta [*v*] for Einstein's frequency [*f*] in relation to Energy.

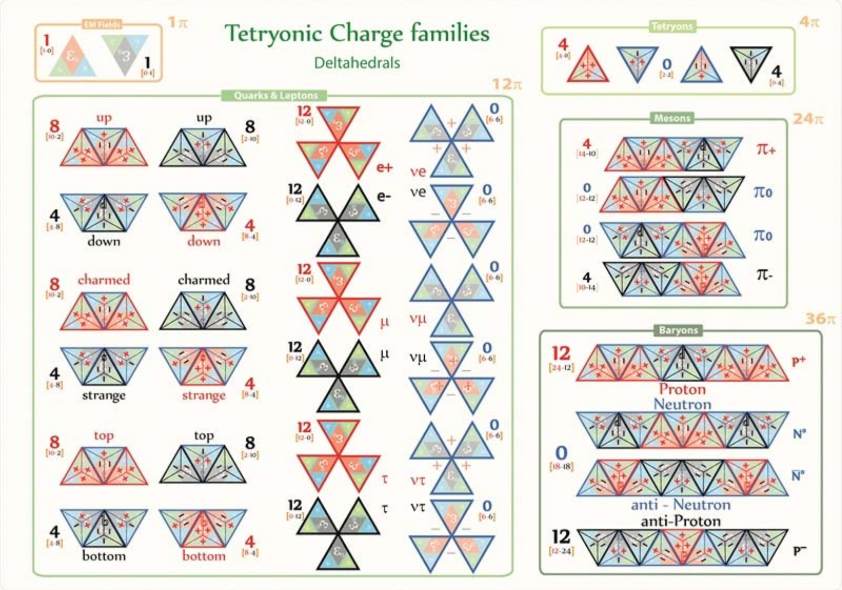
Tetryonic Charge families
Deltahedrals
EM Fields
1π
1 [1-0]
1 [0-1]
Quarks & Leptons
12π
8 [10-2] up
up 8 [2-10]
4 [4-8] down
down 4 [8-4]
8 [10-2] charmed
charmed 8 [2-10]
4 [4-8] strange
strange 4 [8-4]
8 [10-2] top
top 8 [2-10]
4 [4-8] bottom
bottom 4 [8-4]
12 [12-0] e+
12 [0-12] e-
ve 0 [6-6]
ve 0 [6-6]
12 [12-0] μ
12 [0-12] μ
νμ 0 [6-6]
νμ 0 [6-6]
12 [12-0] τ
12 [0-12] τ
ντ 0 [6-6]
ντ 0 [6-6]
Tetryons
4π
4 [4-0]
0 [2-2]
4 [0-4]
Mesons
24π
4 [14-10] π+
0 [12-12] π0
0 [12-12] π0
4 [10-14] π-
Baryons
36π
12 [24-12] P+
Proton
Neutron
N°
0 [18-18]
N̄°
anti - Neutron
anti-Proton
12 [12-24] P⁻

2D EM masses [bosons-photons] do not have the physical property of 3D topology possessed by Matter [Fermions]

Fermions can now be defined as the standing wave energy topologies that create all the known sub-atomic particles – the building blocks of large scale Matter and structure in the macro world.

Historically these particles have been classified according to their charges and masses, Tetryonic geometry now provides a 3D physical model of all the particles highlighting the physical source of the 1/3 charges of Quarks and revealing charge to be the foundational geometry of all Matter.

Tetryons are 4π charge topologies that are the foundational quanta of Matter and surprisingly have a mass-charge ratio identical to that of Leptons explaining how these particles have remained hidden from accelerator experiments.

Quarks are 12π charge geometries where the attractive strong charge interactions between their fascia form 8π particle topologies with entirely different properties to that of Leptons

Leptons also have 12π charge geometries but with repulsive fascia that result in them forming the quantum equivalent of a 6 loop rotor

Protons and Neutrons are 36π charge geometries [that result in 20π Baryon topologies] and have identical masses (against the current model of baryonic masses which derived from an entirely different process where Neutrons are formed by Protons absorbing electrons).

The Particle Zoo

Baryons – the building blocks of atomic nuclei can be formed from many tri- quark combinations and energy levels and this is of particular note when modern accelerators are used to probe atomic structures.

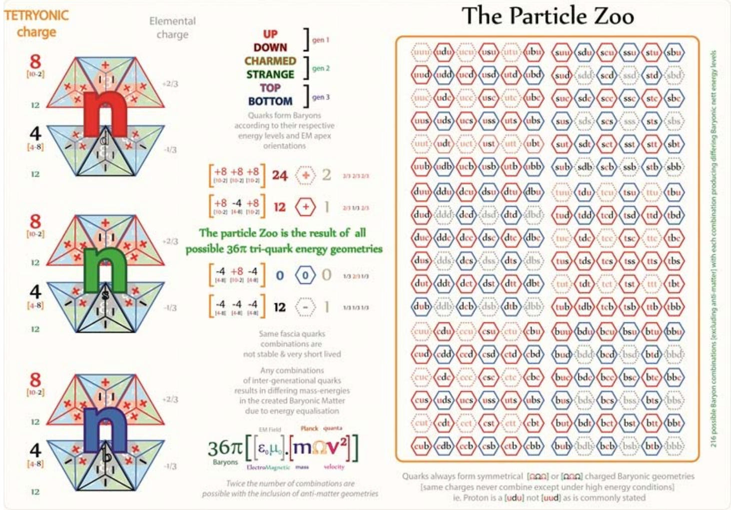
TETRYONIC charge
Elemental charge
8
[10-2]
12
4
[4-8]
12
+2/3
-1/3
UP
DOWN
gen 1
CHARMED
STRANGE
gen 2
TOP
BOTTOM
gen 3
Quarks form Baryons according to their respective energy levels and EM apex orientations
+8 +8 +8 24 + 2 2/3 2/3 2/3
+8 -4 +8 12 + 1 2/3 1/3 2/3
The particle Zoo is the result of all possible 36π tri-quark energy geometries
-4 +8 -4 0 0 0 1/3 2/3 1/3
-4 -4 -4 12 - 1 1/3 1/3 1/3
Same fascia quarks combinations are not stable & very short lived
Any combinations of inter-generational quarks results in differing mass-energies in the created Baryonic Matter due to energy equalisation
36π[[ε₀μ₀].[mΩv²]]
EM Field
Planck quanta
Baryons
ElectroMagnetic
mass
velocity
Twice the number of combinations are possible with the inclusion of anti-matter geometries
The Particle Zoo
216 possible Baryon combinations [excluding anti-matter] with each combination producing differing Baryonic nett energy levels

uuu	udu	ucu	usu	utu	ubu	suu	sdu	scu	ssu	stu	sbu
uud	udd	ucd	usd	utd	ubd	sud	sdd	scd	ssd	std	sbd
uuc	udc	ucc	usc	utc	ubc	suc	sdc	scc	ssc	stc	sbc
uus	uds	ucs	uss	uts	ubs	sus	sds	scs	sss	sts	sbs
uut	udt	uct	ust	utt	ubt	sut	sdt	sct	sst	stt	sbt
uub	udb	ucb	usb	utb	ubb	sub	sdb	scb	ssb	stb	sbb
duu	ddu	dcu	dsu	dtu	dbu	tuu	tdu	tcu	tsu	ttu	tbu
dud	ddd	dcd	dsd	dtd	dbd	tud	tdd	tcd	tsd	ttd	tbd
duc	ddc	dcc	dsc	dtc	dbc	tuc	tdc	tcc	tsc	ttc	tbc
dus	dds	dcs	dss	dts	dbs	tus	tds	tcs	tss	tts	tbs
dut	ddt	dct	dst	dtt	dbt	tut	tdt	tct	tst	ttt	tbt
dub	ddb	dcb	dsb	dtb	dbb	tub	tdb	tcb	tsb	ttb	tbb
cuu	cdu	ccu	csu	ctu	cbu	buu	bdu	bcu	bsu	btu	bbu
cud	cdd	ccd	csd	ctd	cbd	bud	bdd	bcd	bsd	btd	bbd
cuc	cdc	ccc	csc	ctc	cbc	buc	bdc	bcc	bsc	btc	bbc
cus	uds	ucs	uss	uts	ubs	bus	bds	bcs	bss	bts	bbs
cut	cdt	cct	cst	ctt	cbt	but	bdt	bct	bst	btt	bbt
cub	cdb	ccb	csb	ctb	cbb	bub	bdb	bcb	bsb	btb	bbb

Quarks always form symmetrical [ΩΩΩ] or [ΩΩΩ] charged Baryonic geometries
[same charges never combine except under high energy conditions]
ie. Proton is a [udu] not [uud] as is commonly stated

The higher energies they employ when colliding particles together results in higher 2D kinetic energies (and increased energy quanta) which recombine after collisions into a plethora of similar particles called the Particle Zoo.

A sound understanding of EM mass-energy geometries & the charged topologies of all sub-atomic Matter particles reveals the true genesis of all of the Baryons to be found in the particle zoo.

Tetryonic geometries facilitate the precise 3D modelling of all the periodic elements, allotropes and compounds enabling the development of new classes of materials and medicines providing us with many new insights into Quantum Chemistry and large scale Matter in general.

Quantum Electro-Dynamics [QED]

Adding to the mysteries of quantum mechanics is a number of well-known QED effects that have also eluded physical explanation – to date

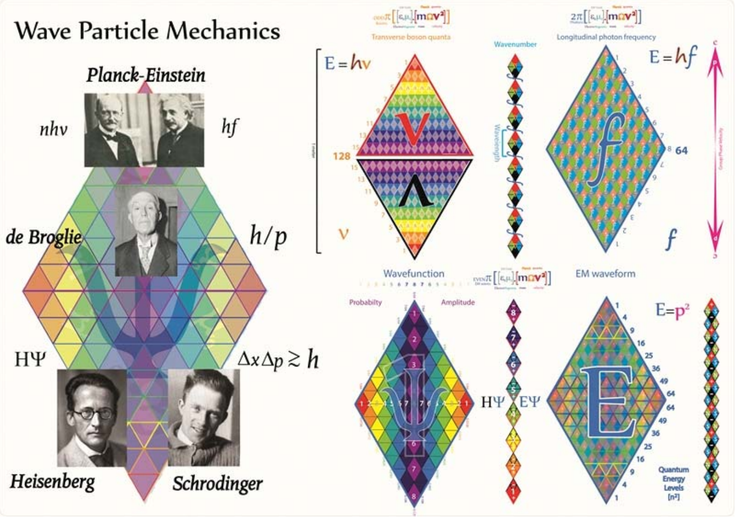
Wave Particle Mechanics
Planck-Einstein
nhν
hf
de Broglie
h/p
HΨ
ΔxΔp ≳ h
Heisenberg
Schrodinger
Transverse boson quanta
E = hν
128
ν
Wavenumber
Wavelength
Longitudinal photon frequency
2π[[ε₀μ₀].[mΩv²]]
E = hf
64
f
Wavefunction
Probabilty
Amplitude
HΨ
EΨ
EM waveform
E=p²
Quantum Energy Levels [n²]

In 1865, James Maxwell Clerk unified the then disparate theories of Electric and Magnetic fields into a theory of Electromagnetism and related then to the velocity of light.

The exact quantum geometry for EM wave geometries (and the photons comprising them) that explains their Wavefunctions and observed inference patterns represents one of the greatest challenges to developing a concise fully realised quantum theory of EM radiation.

Wave-Particle duality dating back to the 17th century is perhaps the best known example of quantum behaviour that has defied precise modelling despite intense effort by scientists around the World for over two centuries

The application of equilateral energy momenta to the charged geometries of Photons and EM waves not only explains these long standing mysteries but also removes the quantum fuzziness introduced by Heisenberg's Uncertainty principle, in turn clearing the way for Science to develop an advanced understanding of Electricity, its role in Quantum ElectroDynamics and provide new clean forms of energy from quantum processes.

Photo-electrons and spectral lines

The physical relationship between Spectral lines, Rydberg's Constant and the Kinetic energies of Photo-electrons as they interact with photons and atomic nuclei is quickly revealed when equilateral energy geometries are employed

It is the rigid geometry of equilateral Planck energy momenta that gives rise to the invisible forces of Faraday's Electric and Gauss' Magnetic fields as well as Newton's Action-at-a-Distance.

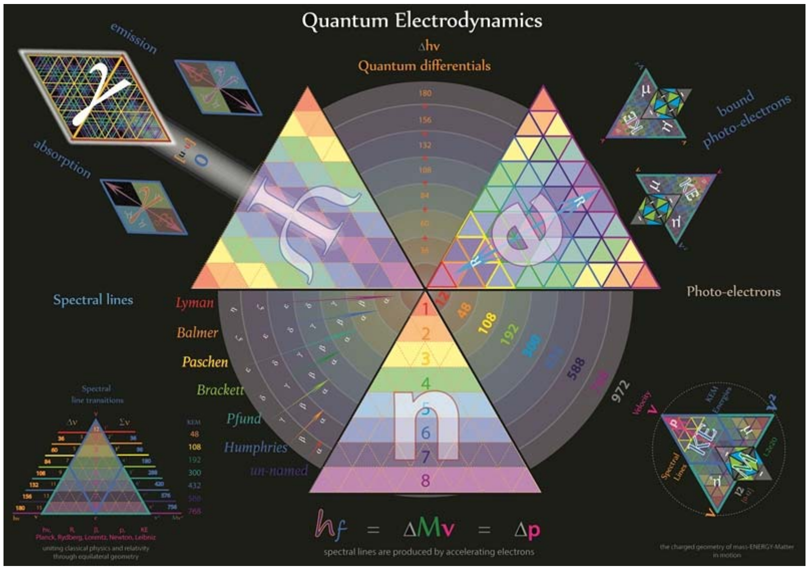
Quantum Electrodynamics
Δhv
Quantum differentials
emission
absorption
bound photo-electrons
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
1
2
3
4
5
6
7
8
12
48
108
192
588
972
Spectral line transitions
Δv
Σv
KEM
Velocity
KEM Energies
Spectral Lines
KE
μ
hf = ΔMv = Δp
spectral lines are produced by accelerating electrons
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity through equilateral geometry
the charged geometry of mass-ENERGY-Matter in motion

Quantised Angular Momentum, a direct measure of the long hidden equilateral geometry of Energy, is revealed as the source of Charge, the physical constants and even the geometry of EM mass-energy & Matter itself.

The geometry of Nuclear Forces and Constants

Charge interactions resulting from equilateral EM energy geometries have been mistakenly developed into three disparate nuclear forces:

The EM Force is the result of $2n\pi$ charge geometries acting along transverse or longitudinal directions (or in superposition) in the forms of Bosons and Photons

The Weak Force is the inductive coupling of the magnetic permeability of adjacent energy geometries

The Strong Force is the attractive force between opposite charged fascia of Tetryonic Matter. It can also form a repulsive force between similar charged fasciae resulting in charged Leptons [electrons].

Gravitation (mathematically identical to Coulomb's Force save for strength and source) can also be modelled as the geometric mean of super-positioned EM waves.

All of which are determined by the Fine Structure Constant – the mysterious hand of GOD – setting the strength of EM interactions and determining Charge on the quantum level.

Applying equilateral geometry to energy momenta quanta, the Tetryonic model of EM mass-ENERGY-Matter quickly evolves to explain many additional electrical properties such as Voltage, Current and Power

Quantum Chemistry

Applying tetrahedral Matter topologies to quantum chemistry leads to a number of significant advances in the understanding of chemical processes most notably:

- **Accurate models for all periodic elements and their associated allotropes**
- **Advanced molecular orbitals plots**
- **Exact rest masses for all elements and**
- **A new Periodic table based on the charged topologies of Matter**

Improving upon and extending the Aufbau principle with charged geometries leads to many refinements in chemical engineering and facilitates the building of accurate 3D models of all elements, isotopes and compounds.

Every element can be analysed from the quantum level upwards revealing its exact 3D topology, quark configuration, rest mass-Matter-energies and chemical properties.

And number of misconceptions can be quickly clarified and corrected in turn advancing our understanding to the physical mechanics underlying chemical processes like chemical bonding, element family properties and the structures of complex chemical compounds.

Hydrocarbons and Carbohydrates have already been modelled using equilateral charge topology with the results matching the established chemical models of Nobel gases, Core electrons, Lewis diagrams and the most recent electron photographs of these chemical properties of atoms.

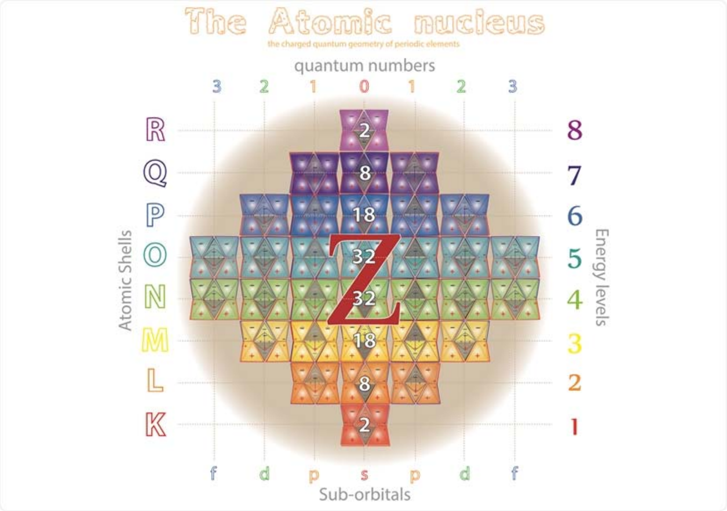

Even DNA can be modelled in 3 dimensions leading to a greater understanding of the role of quantum mechanics in biological processes and the interaction of EM radiation on organic compounds.

A highly developed quantum model of chemical elements and compounds will allow us to model chemical and biological structures in hitherto unimagined detail and facilitate the building of complex (and increasingly accurate) models of all chemical compounds and biological processes.

A Unified field equation

In physics a Unified Field Theory is a type of theory that allows all that is usually thought of as a fundamental force or elementary particle to be formulated in terms of a single equation that explains all their disparate properties.

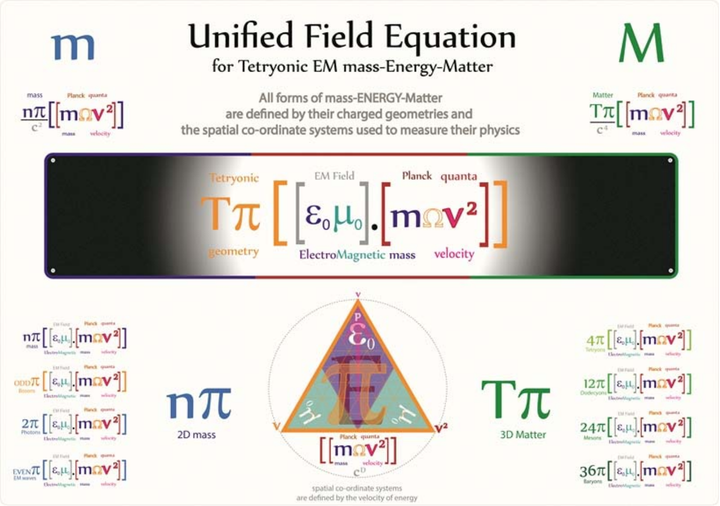
m
Unified Field Equation
for Tetryonic EM mass-Energy-Matter
M
mass
Planck quanta
$\frac{n\pi}{c^2}[[m\Omega v^2]]$
mass
velocity
All forms of mass-ENERGY-Matter
are defined by their charged geometries and
the spatial co-ordinate systems used to measure their physics
Matter
Planck quanta
$\frac{T\pi}{c^4}[[m\Omega v^2]]$
mass
velocity
Tetryonic
EM Field
Planck quanta
$T\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
geometry
ElectroMagnetic mass
velocity
$n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
$\text{ODD}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Bosons
$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Photons
$\text{EVEN}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
EM waves
$n\pi$
2D mass
v
P
ε_0
π
v
v²
Planck quanta
$[[m\Omega v^2]]$
mass
velocity
c^D
spatial co-ordinate systems
are defined by the velocity of energy
$T\pi$
3D Matter
$4\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Tetryons
$12\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Dodecyons
$24\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Mesons
$36\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$
Baryons

It is considered the holy grail of Science, offering the promise of advances in Technology, Energy and Medicine unparalleled in Human history.

Equilateral energy provides the foundation for the creation of a single equation that expresses EM mass-ENERGY and Matter as the result of geometric & topological charge interactions.

Applying equilateral geometry to Planck energy momenta quanta has been shown to firmly resolve many of the current mysteries plaguing the Standard model and provides a firm foundation for the development of a single unified quantum theory uniting all quantum and macroscopic forces.

It effortlessly merges all the tested features of Classical mechanics with the statistical probabilities of quantum mechanics and scales up to the cosmological scale of General Relativity.

Equilateral charged energy geometries provide a quantifiable foundation for advances in:

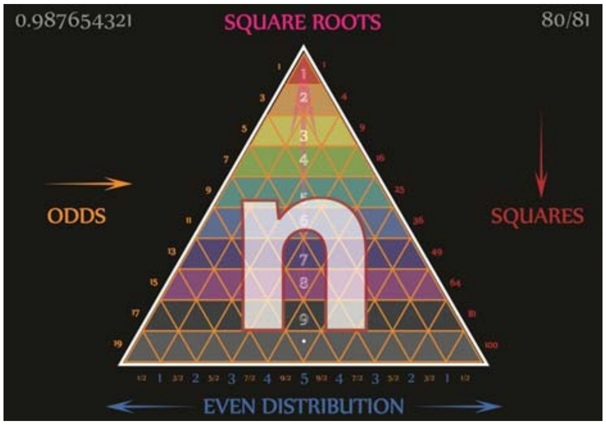
0.987654321
SQUARE ROOTS
80/81
ODDS
SQUARES
n
EVEN DISTRIBUTION

- **Quantum Mechanics**
- **Quantum Electrodynamics**
- **Quantum Chemistry**
- **Quantum Cosmology**

Explaining the quantum mysteries of mass & Matter, Wave-Particle duality and gravitational Singularities by challenging the very foundational assumptions of Relativity and the role of charge in our Universe.

Mathematical principles such as re-normalisation, probabilities and square roots of negative numbers can all be geometrically modelled through Tetryonics leading to further advances in fields such as biology, medicine, quantum computing and telecommunications

Through the charged geometry of Tetryonics many of the erroneous theories that have been developed during the last 1oo years can be falsified at last providing a solid foundation for further advances in science and mathematics.

Gravitation

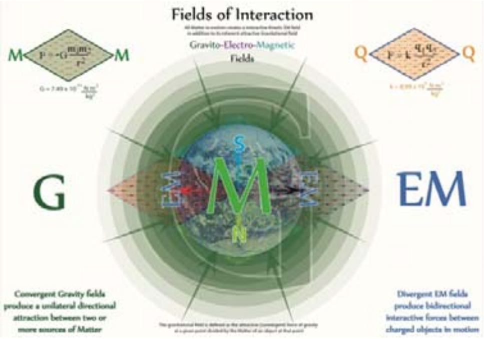
Fields of Interaction
Gravito-Electro-Magnetic
Fields
M
M
Q
Q
G
EM
S
M
N
EM
EM
Convergent Gravity fields produce a unilateral directional attraction between two or more sources of Matter
Divergent EM fields produce bidirectional interactive forces between charged objects in motion

The force of Gravity has been long-held to be the mysterious force that holds our Universe together, shaping the Stars and holding us to the Earth as we orbit our own SUN.

Tetryonic geometry reveals its quantum mechanics to be the result of electromagnetic fields and the effect produced by the geometry of Matter in these fields.

Newton saw Gravity as a force that acts instantaneously on distant bodies of Matter while Einstein dismissed the force between the bodies and explained it as the result of the 'curvature of Spacetime'.

In fact the Gravitational field is comprised of 3 separate components, each of which produces differing forces of interaction, but all combine to produce the nett Gravitational field that we observe and describe as strictly convergent Gravity.

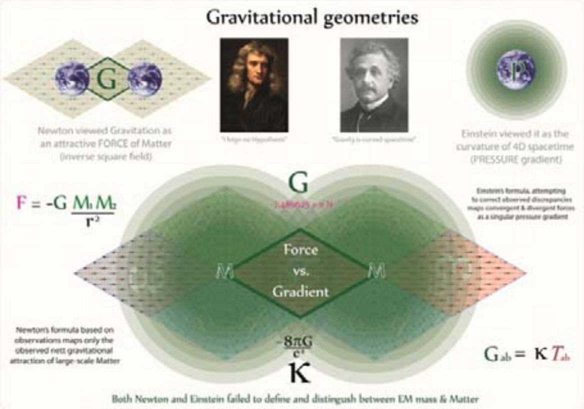
Gravitational geometries
G
Newton viewed Gravitation as an attractive FORCE of Matter (inverse square field)
Einstein viewed it as the curvature of 4D spacetime (PRESSURE gradient)
Einstein's formula, attempting to correct observed discrepancies maps convergent & divergent forces as a singular pressure gradient
G
F = -G M₁M₂ / r²
Force vs. Gradient
M
M
Newton's formula based on observations maps only the observed nett gravitational attraction of large-scale Matter
-8πG / c⁴
K
Gab = K Tab
Both Newton and Einstein failed to define and distingush between EM mass & Matter

Gravity fields are created by the topology of Matter itself as it displaces and distorts the Vacuum energies surrounding it.

Electric fields diverge radially from all Matter as it is heated, or placed in motion by other forces creating an interactive field with both divergent and convergent equatorial regions of force and acceleration.

Dipole Magnetic fields located around the Polar Regions of all Matter also create a perturbative force on bodies close to them

Although Newton's and Einstein's theories were both correct in their respective analysis of the forces at work they both failed to define and distinguish between EM masses and Matter leading to the continued debate over the true nature and physical formulation of quantum gravity.

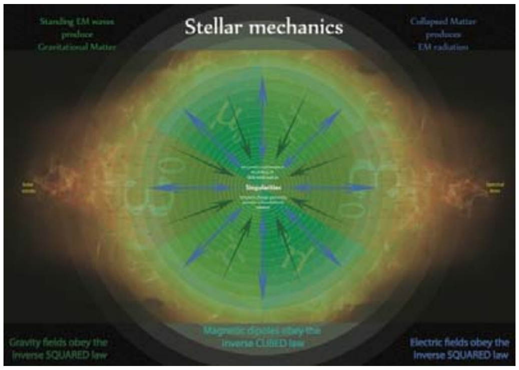
Stellar mechanics
Standing EM waves produce Gravitational Matter
Collapsed Matter produces EM radiation
Singularities
Gravity fields obey the inverse SQUARED law
Magnetic dipoles obey the inverse CUBED law
Electric fields obey the inverse SQUARED law

These three quantum forces [G+E+M] combine to produce a field of gravitational acceleration that was so eloquently modelled by Newton back in the 16th century through his inverse square law.

And these 3 distinct forces also combine to produce the stellar mechanics that power the Stars – a Gravito-Electro-Magnetic pinch.

The force of Gravity, created by the geometry of Matter draws Matter toward the core of Stars.

And the GEM pinch singularity at the core of the Star destroys Matter (collapsing its 3D geometry), turning it into radiant 2D ElectroMagnetic radiation (Light).

Energy from the Stars

Virtual particles, Dark Energy and Dark Matter (were all developed to explain the observed mechanics of our Universe) and can now be revealed to be differing manifestations of the electromagnetic force interacting at various angles to, and with, the standing-wave topologies that create it.

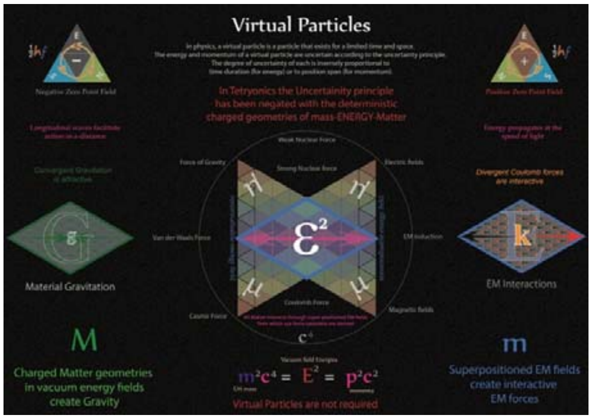
Virtual Particles
Negative Zero Point Field
Positive Zero Point Field
In Tetryonics the Uncertainity principle
has been negated with the deterministic
charged geometries of mass-ENERGY-Matter
Material Gravitation
EM Interactions
Weak Nuclear Force
Strong Nuclear force
Force of Gravity
Electric fields
Van der Waals Force
EM Induction
Coulomb Force
Magnetic fields
Divergent Coulomb forces are interactive
M
Charged Matter geometries
in vacuum energy fields
create Gravity
m
Superpositioned EM fields
create interactive
EM forces
Vacuum field Energies
m²c⁴ = E² = p²c²
Virtual Particles are not required

Quantum fields of Planck energies combine to create charged 2D geometries of radiative ElectroMagnetic fields that in turn spread out interacting with each other to give us the fundamental laws of Attraction and Repulsion.

The 3D standing-wave topologies of Matter provide a rigid structure that displaces the surrounding Vacuum energies and creates a pressure gradient which we call Gravity.

And the destruction of these Matter topologies in turn creates radiant energies that propagate out into the surrounding space to weaken and form the Vacuum energies that permeate all of Space.

The eternal cycle of Matter creation and destruction within Stars and Galaxies is what drives the Universe we see today causing it to contract as it forms gravitational Matter and to expand as it restructures the standing-wave Matter into radiant forms of Energy (Light and heat).

The same process offers us clean, limitless energy for all our needs as well as unlimited resources from the restructuring of this energy into its varied elemental and compound forms.

Our toxic stockpiles of radioactive wastes can be completely eliminated from the environment and electrical power can be safely stored as mass within large-scale models of the deuterium nuclei and distributed worldwide to any location without the need for transmission lines.

Even hydro-electric storage schemes can be eliminated, returning water courses to their natural states of flow in turn reducing conflicts stemming from the flow of this precious resource worldwide.

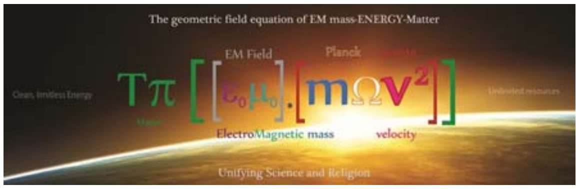
The geometric field equation of EM mass-ENERGY-Matter
Clean, limitless Energy
EM Field
Planck
Tπ [[ε₀μ₀].[mΩv²]]
Unlimited resources
ElectroMagnetic mass
velocity
Unifying Science and Religion

Through the geometries of equilateral Planck energies and the application of the Tetryonic unified field equation of mass-ENERGY-Matter in motion to our needs, for the first time in Humanity's history, we have the chance to advance our civilization to new pinnacles of technology and understanding in turn leaving our World a better place than we found it to be.

The age old dilemma of technological advancement versus environmental pollution is at an end

We stand on the threshold of realising one of the greatest advances in scientific understanding ever witnessed but we must also endeavour to redress our current social challenges with an equally rigorous determination to ensure all of Humanity benefits from this discovery

…………………….……… ABRAHAM [2012]

Tetryonics …. Challenging and correcting the foundations of quantum theory

Electric field permittivity has a 2pi (rhombic) field geometry while magnetic field permeability is always the result of two separate equilateral magnetic fields that form a magnetic dipole moment

Where two or more rhombic electric fields interact a Coulombic –charge force of interaction results
Where two or more equilateral magnetic fields interact a Gaussian-Ampere forces of interaction results *[opposite fields attract - same fields repel]*

The classical model of the unidirectional electrostatic force fields of point charged particles is corrected to show a bidirectional arrangement of equilateral Planck energy momenta geometries within the electrostatic field that creates the bi-directional forces of charge interactions

Where two similar coupled equilateral energy momenta fields combine an electrostatic field is created
Where two opposite coupled equilateral energy momenta fields combine a magnetic dipole is created

The arrangement and asymmetry of Planck quantum charges with in any EM field determines the motion and acceleration of charged particles within the field of electrostatic force

The classical forces of attraction should be renamed forces of interaction
(as they result from bidirectional quantum vector forces)

All electromagnetic fields are comprised of a squared number of equilateral Planck energy momenta quanta (whose asymmetrical distribution of equilateral Planck quanta results in a net divergent Force) with a convergent vector force component

When modelled with equilateral Planck geometries electromagnetic fields follow the Biot-Savart Law:

- Rhombic electric fields follow the inverse squared law of propagation
- Equilateral magnetic fields follow the inverse squared law of propagation
- Magnetic dipole fields follow an inverse cubed law of propagation

The equilateral geometry of quantised angular momentum creates chirality at the quantum level in physics (rotation does not affect the charge of plank quanta, but mirroring and/or spatial coordinate translations do)

All positive charged fields can be modelled as having a clockwise inductive flux geometry
All negative charged fields can be modelled as having an anticlockwise inductive flux geometry
(with OMEGA representing equilateral geometries per second and PI reflecting their scalar areas]

All W bosons are comprised of an odd number of equilateral Planck quanta in a trapezoidal geometry
All Z bosons are comprised of an even number of equilateral Planck quanta in a dual-trapezoidal geometry (neutral Z bosons are equivalent in terms of Planck mass energy momenta to photons save their net geometry)

All charged fascia [Higgs bosons] are comprised of squared numbers of equilateral Planck quanta in an equilateral geometry.

The odd quanta W bosons within electromagnetic fields form quantum levels of charged mass-energy momenta

The Planck equation [E=n.hv] of quantised energy is a generalised equation of Planck geometries (and applies to all energy momenta arrangements equally)

Bosons and Quantum levels are a transverse measure of mass energy momenta in EM fields
Photons are a longitudinal measure of mass energy momenta within the same electromagnetic fields

Equilateral Planck energy momenta per second [E/c^2] equates to mass geometries
Equilateral Planck energies momenta per second squared [E/c^4] equates to Matter topologies

Interactive **mass energy geometries** can combine via the weak force to form standing wave Matter topologies

Standing wave **Matter topologies** can combine via the strong force to form elementary subatomic particles

The fundamental Matter quantum of all physics is the tetryon (a tetrahedral mass-energy Matter topology) which is formed from photons of electrostatic energy and can be found as positive, neutral or negative charges topologies

These three forms of charged tetryons form the foundation of all the physical Matter of our universe as they seek a state of equilibrium, Tetryons are not spherical point charges – they are tetrahedral, likewise the subatomic particles they form are not spherical point charges - they are **[T]**π Tetryonic charge topologies

The Tetryonic topologies created by tetryons as they combine to form elementary particles creates new forms of closed volume topologies that follow Euler's characteristic [V+F-E = 2] for Platonic solids

Quarks and leptons are comprised of 12 charged mass energy geometries
- quarks form *[8π]* octahedral Matter topologies
- leptons form *[12π]* dodecahedral Matter topologies

and

Baryons are comprised of 36 charged mass energy geometries
- forming *[20π]* Icosahedral Matter topologies

Anti-UP quark's are comprised of 2 positive and 8 negative charges
creating a net charge of **-8** *[-2/3 elementary charge]*
DOWN quark's are comprised of 4 positive and 8 negative charges
creating a net charge of **-4** *[-1/3 elementary charge]*
Anti-DOWN quark's are comprised of 8 positive and 4 negative charges
creating a net charge of **4** *[1/3 elementary charge]*
UP quark's are comprised of 10 positive and 2 negative charges
creating a net charge of **8** *[2/3 elementary charge]*

Positrons are comprised of 12 positive charges
creating a net charge of **+12** *[positive integer elementary charge]*
neutrinos are comprised of 6 positive and 6 negative charges
creating a net charge of **0** *[no elementary charge]*
Electrons are comprised of 12 negative charges
creating a net charge of **-12** *[negative integer elementary charge]*

Protons are comprised of 24 positive charges and 12 negative charges
creating a net charge of **+12** *[positive integer elementary charge]*
Neutrons are comprised of 18 positive and 18 negative charges
creating a net charge of **0** *[no elementary charge]*
anti-Neutrons are comprised of 18 positive and 18 negative charges
creating a net charge of **0** *[no elementary charge]*
anti -Protons are comprised of 12 positive charges and 24 negative charges
creating a net charge of **-12** *[negative integer elementary charge]*

Rest lepton topologies have a completely neutralised magnetic dipole structure and all leptons (charged or neutral) have identical physical Matter topologies
Their rest masses are the result of the number of plank quanta comprising each Matter topology with applied forces creating additional asymmetrical KEM fields resulting in kinetic energies, magnetic moments and vector velocities of motion

As the energies of motion are added to equilateral KEM fields the Planck quanta themselves are physically WAVE-length Lorentz contracted

(Lorentz contractions apply to mass energy geometries only - NOT to Matter topologies)

Matter topologies are Lorentz invariant to velocity changes (or acceleration)
The relativistic distortion of spherical point charges in physics is no longer applicable, in turn undermining a basic tenet of special relativity as postulated by Albert Einstein

It is the equilateral secondary KEM fields of motion created by applied forces to material Matter topologies (and the asymmetrical distribution of plank charges therein) that creates the magnetic moments associated with charged particles in motion, not the relativistic distortion of spherical bodies of Matter as postulated by Einstein in special relativity

These secondary KEM fields of motion also contain the kinetic EM energies and vector [square root] linear momentum of material particles in motion - the mass energy momenta density and physical geometries of Matter topologies is unaffected by applied forces or changes to velocity-momenta

The changing Lorentz contracted mass energies of kinetic EM fields contributes to the total relativistic mass energy of Matter topologies but in no way changes the mass energy distribution of the Matter component of particles in motion itself - leading in turn to varying generations of subatomic particles with in the particle families themselves

The net asymmetric KEM field geometry of each particles in motion is reflective of the charged distribution of the Matter topology of the particle itself.

Mesons are created when quark and anti-quark particles combine via the strong force interaction between their charged fascias to create unique 14 pi Matter topologies (neutral or charged pions)

These strong interactions between opposite charge fascia of material particles and the resulting Matter topologies are evidence of the fact that Matter-antimatter combinations do not necessarily result in the explosive annihilation of the particles concerned (in fact the combining of opposite charge quarks is a necessary step in the formation of the baryons)

All Baryons are Matter topologies created from the combination of tri-quark topologies
The quark's combine via their opposite charged [Higgs] fascia to create neutral and charged elementary topologies
Due to this opposite charge interaction the charged distribution of baryonic topologies does not follow the textbook given arrangements of [UUD] & [DDU] instead they are arranged as follows:

Protons	**UP**	**+**	**DOWN**	**+**	**UP**
Neutrons	**DOWN**	**+**	**UP**	**+**	**DOWN**
anti-Neutrons	**anti-DOWN**	**+**	**anti-UP**	**+**	**anti-DOWN**
anti-Protons	**anti-UP**	**+**	**anti-DOWN**	**+**	**anti-UP**

The specific arrangement of charged fascia within these respective Matter topologies goes on to facilitate nuclear bonding via the residual strong force between atomic nuclei, following the charge arrangement given in modern textbooks does not allow the formation of baryons or their subsequent bonding to form larger scale atoms and molecules

Protons and neutrons are revealed to have quark arrangements that are mirror symmetries of each other, revealing the neutral charge neutron to have a complementary negative charge arrangement to that of protons.

The increased Planck mass energy contents of various families of quarks facilitates the creation of a huge number of possible Baryon topologies resulting in the particle zoo of subatomic elements created in particle collisions within accelerator experiments - all of these elements can be easily modelled and accounted for using Tetryonic mass-energy Matter topologies

Additionally, the particles often referred to as gluons within the standard model can be easily accounted for and explained as being neutral tetryons within Tetryonic theory - where these neutral charge Matter topologies form a neutral Matter dielectric between charged tetryons to facilitate the strong force interaction and the formation of larger particles

Neutral tetryons (gluons) can also form particles in their own right (neutrinos and glue-balls), as can charged tetryons (charge-ball particles), most of which are short lived

Tetryonics reveals the previously hidden geometry and topology of electromagnetic fields and subatomic particles and the geometric relationship and interplay between electric permittivity and magnetic permeability within all EM field that results on the speed of light in any medium.

As well as the revealing a geometric source of all physical constants within physics itself, Tetryonic charge geometries show the weak force to be a magnetic dipole [edge] interaction between mass energy geometries and Matter topologies and the strong force to be a charge [fascia] interaction between Matter topologies themselves

The equilateral [quantised angular momentum] of ***inductive*** Planck mass energy geometries within squared [HIGGS] fascia of Matter topologies is what creates ***inertial*** properties of Matter at the quantum level

The strong force between charged parallel Higgs fascia of Matter topologies results in 'hidden' partitions of mass energy with in the Matter topologies of quark's and baryons leading to the inability of mathematics online to accurately model and understand the physics of creating these particles at the quantum level

Tetryonic charged mass energy geometries allows for the accurate 2D and 3D modelling of all Matter topologies and fields of Force interactions within physics

It reveals that the deuterium atom [not the hydrogen atom] is the quantum building block of all elementary Matter - the unique charge topology of deuterium (consisting of a proton, neutron and electron) can be related to the electromechanical topology and function of quantum synchronous converters.

These quantum synchronous converters absorb and emit mass energies via the rotating electrons [quantum rotors] bound to the nuclei by way of radiant bosons and photons [heat and light]

The net charge of Deuteron nuclei facilitates their binding together to form helium and larger atoms, while stored mass energy quanta within their Matter topology results in raised quantum grounds states that can all be modelled mathematically using Schrodinger’s wavenumbers

The simpler geometric charge topology reveals visually how the mass energy content of baryons directly affects the angular momentum and energy levels of photo electrons bound to atomic nuclei, in turn revealing that larger elements be created in two ways - namely by the addition of a neutron to a Hydrogen atom (as classically modelled in modern chemical physics) or through the raising of Planck mass-energy quanta within the deuterium Matter topology itself

Deuterium nuclei without any bound photo-electrons form quantum batteries with EM mass energies can be stored indefinitely without being released.
These nuclei can be arranged atomically in either a series, parallel or empty parallel configuration each of which affects the ground energy states of the respective nuclei created.
Where photo-electrons (quantum rotors) are attracted and bind to these nuclei (quantum batteries) the stored mass energies can then be released in a controlled manner via spectral line emissions, or energy can be added via spectral line absorption.

As photo-electrons are bound to the deuteron nuclei the energy levels are affected directly by the mass energy levels of the nuclei that they bind to, in turn affecting their KEM field energies
The square root linear momentum of KEM field energies normally creates a vector motion but in the case of bound electrons this vector motion is turned into an angular motion resulting in electron spin within the nucleus.

The direction of the spin is always reference with respect to the nuclear magnetic moment resulting in up and down electron spins and their associated magnetic dipole moments

The kinematics of KEM fields provides an easy visual explanation of Newton's 2nd law of motion - namely that deceleration can be seen as the removal of odd number quanta from the squared kinetic energy fields and acceleration can be viewed as the addition of odd numbered quanta to squared kinetic energy fields.

The addition or subtraction of odd numbered Planck quanta from squared energy fields results in changing square root linear momentum for the particles that these fields are associated with - this changing linear momentum over time can be equated to acceleration leading to Newton's formulation of Force = mass x acceleration [F = ma = dp/t = dmv/t]

Using the equilateral geometry of Planck's constant we can now unify classical mechanics with quantum mechanics and reveal the underlying dynamics of relativistic mass energies of motion (as well is providing an electromechanical explanation of inertia in the physics of Matter in motion)

Newton's vector linear momentum [p=mv] forces are revealed to be the geometric square root of Leibnitz's scalar energies [E/mv^2] with their associated inertial mass being a measure of the scalar energies per second [m=E/v^2]

Linear momentum [mv] is differentiated from vector velocities [v] through the equilateral geometry of energy per second [m] in tetryonics, with the total linear momentum of any EM field or system of particles being the vector some of the square root Planck linear momentum of the quanta in the fields comprising the system

Once the equilateral field geometry of KEM fields is revealed it is easy to show that the total energy of the field [E=mv^2] is the result of the kinetic energies [$_{1/2}$Mv^2] plus the magnetic moment [$_{1/2}$Mv^2] of any material topology in motion

This in turn allows for a rigid geometric differentiation of quantised angular momentum [Ω] in Planck quanta, the linear momentum [p] of any system and the classical angular momentum [vector rotation about a point] in physics

These geometric definitions of mass energy momenta in physical systems as portrayed in Tetryonic theory reveal a differing geometric relationship to that commonly portrayed in modern physics]i.e. neutralising right angled Pythagorean triangles]

Additionally Lorentz corrections factors are restricted to planar 2-D mass energies of KEM fields associated to particles in motion where changes of velocity force result in the addition or subtraction of equilateral Planck quanta from the KEM field geometry resulting in the physical expansion or contraction of Planck geometries within the field

Tetryonic theory's geometric definition and application of quantised angular momentum with respect to mass energy momenta in physics allows for the physical modelling of mass energy and Matter at all scales of physics and their unification of classical, quantum and relativistic mechanics

The 'squared' energies of quantum mechanics and statistical distributions of the math associated with quantum mechanics is now revealed in an entirely new light facilitating the macro scale modelling of all hitherto unseen quantum particles and processes.

The geometric relationships between charge, mass geometries and Matter topologies is easily visualised and applied to all physics revealing quantised angular momenta [QAM] to be the hidden variable of Planck's constant [h].......

The modelling of quantised angular momentum as an equilateral geometry allows for the elimination of Heisenberg's uncertainty principle and Bell's inequality from the foundational tenants of quantum mechanics

Both the position and momentum of particles and fields can be now modelled at any stage of a system's spatio-temporal development and evolution using these simple equilateral charged geometries of mass energy momenta

Precise rest mass energies of any Matter topology can be calculated from 1st principles allowing for accurate modelling of all quantum systems and even periodic elements - expanding our knowledge of chemical systems and processes

The Planck-Compton frequency, de Broglie wavelength, absolute rest mass and charge Matter topology of all periodic elements, compounds and molecules can now be calculated with absolute certainty and precision and can be related back to Avogadro's number for all elements in chemistry

The exponential energy levels of atomic nuclei and atomic processes of energy release can also now be shown to be the result of an underlying geometric Planck scale geometry inherent to all mass energy Matter within physics

Einstein's mass energy equivalence formula is shown to be a geometric relationship between measurements of energy per second and inertial mass........ combined with Tetryonics new definition of Matter as energy per second squared we can now show that Einstein's famous formulation is an incomplete formulation of the mass-ENERGY-Matter relationship that applying only to specific processes within our physical universe

The Lorentz corrections on which special relativity is based are shown to be WAVE-length and QAM corrections for the changing physical geometries of scalar mass energies per unit of time in spatial co-ordinate systems (c^2) and that additional time-based dimensional analysis must be taken into account in order to define and differentiate between electromagnetic mass and Matter in physics

The application of squared (equilateral) geometries to physics in the role of quantised angular momentum has been overlooked since the inception of physics itself, but with the inclusion of this geometry a rigid geometric grammar can now be imposed upon the mathematics of physics leading to the correction of many erroneous mathematical assumptions

An exact value from first principles for Planck's constant can be calculated and Avogadro's number can be confirmed and related to the absolute rest mass of hydrogen atoms.

A geometric representation of Coulomb's charged field interactions and the impedance of free space can also be provided for all physical processes allowing for a fast and accurate visualisation of energy mechanics at all scales of physics

Tetryonics theory’s equilateral charged mass energy and polyhedral Matter topologies allow for the modelling and differentiation of energy in all its forms within any spatial coordinate system providing a vast improvement over relativity's use of stress energy tensors and non-Euclidean geometries in modelling physical systems

All fields and particles can be defined and differentiated in terms of their charged mass energy geometries and Matter topologies respectively and the differentiation of elemental families into their generations can be easily accounted for through the addition of mass energies of motion to secondary KEM fields associated with each particle grouping

Inertial resistance to force can be modelled and explained through the equilateral geometry of quantised angular momentum when viewed as an electromechanical equivalent to an inductive loop of energy

Electromagnetic inertial mass can now be clearly defined as a measure of the inductive EM energy density of any charged 2D planar Planck geometry in any spatial co-ordinate system of measurement

3D Matter can now be viewed and modelled as the closed topology created by 2D mass energies through charge interactions in any spatial co-ordinate system of measurement per second squared

Additionally physical terms such as 'mass-less' particles are shown to be a misnomer as any field or particle must by definition contain energy per unit of time - and should be more appropriately re-termed Matter-less or weightless particles

Tetryonics theory defines, differentiates and unifies energy in all its forms (mass, Matter & Force) through the equilateral geometry of quantised angular momentum at the quantum level, allowing for the development of a unified field equation that models physics on all scales and reveals a common geometric thread to all the disparate fields of physics that to date have defied unification

Equilateral (SQUARED) energy fields of Planck quanta comprised of statistical distributions of (ODD) transverse bosons and (EVEN) longitudinal photons account for all the quantum mechanical processes of physics including EM field geometries and material particle topologies and electron spin

Principle quantum levels [1st Schrodinger wave-numbers] are shown to be the result of the series addition of deuterium nuclei in atomic elements with their resulting energy levels determining the squared energy levels of bound photoelectron KEM fields – in turn forming the foundation of discrete quantum jumps of electrons bound to atomic nuclei

The bound photo-electron can only transition between allowed squared energy levels permitted by the energy levels of deuterium nuclei to which they are bound and once free from the atomic nuclei the electron can emit or absorbed a continuous spectra of light photons - but while bound to the deuterium nuclei in atoms it can only have specific [squared] KEM energy levels

The complex, asymmetric tri-quark (9 tetryon) Matter topology of baryons directly affects the magnetic moment of protons and neutrons in motion making them considerably weaker than the symmetric axial Bohr magnetron of electrons in motion

As the Bohr magneton of photo-electrons is determined by the squared energies of its secondary KEM field confusion has arisen with respect to the allocation of spins to subatomic particles, Tetryonic geometries now correct for this and suggest that the electron historically viewed as a spin 1/2 particle is in fact a spin 3 particle under the formal definitions of spin rotations i.e. rotating at 120° can bring it back to an identical quantum state as the initial state measured]

Again Tetryonic theory and its energy geometries in addition to explaining and correcting electron spins also shows that the term electron spin can now be taken literally when modelled with Tetryonic geometries to provide an accurate description of the origin of magnetic moments for electrons, in turn undermining the foundational postulate of special relativity that these magnetic moments are the result of the distortion of spherical point charges and that relativistic Lorentz contractions apply to Matter topologies

Dirack's constant, or Planck's reduced constant [h/2π], is revealed to be a measure of the electric permittivity of a relativistic electron's KEM field, and the measurement of the associated magnetic dipole of the same KEM field results in the mistaken concept of 1/2 spins in the measurement of Bohr magnetons

The magnetic vector created by the Bohr magneton [or KEM field dipole] when measured with respect to the nuclear magneton [Proton magnetic moment] produces either a higher energy [parallel] or lower energy [anti-parallel] state for all atomic nuclei

It is the arrangement of these parallel or anti-parallel magnetic dipoles within atomic nuclei that determines the exact frequency-wavelength (energy level) of emitted spectral lines with the interaction of a particle’s magnetic dipole moment with external electromagnetic fields creating Lorentz forces

As electrons and protons have identical quantum Matter topologies [elementary charges] but differing mass energy contents the electron is able to create a much larger secondary KEM field (with associated magnetic dipole moment) for the same velocity [$_{1/2}$Mv^2]

The energy level differences created by the parallel or anti-parallel spin coupling of electrons with protons or deuterium nuclei within atomic elements creates the hyperfine splitting and Zeemen effects observed and spectral lines

As the electron is now shown to be a complex non-spherical particle topology when it binds with protons to form hydrogen the unique Matter topology of both the proton and the electron [Hydrogen] create a small but detectable precession in the motion of the spinning electron as compared to that of the motion of the same electron in a deuterium nuclei

Additionally, as all periodic elements, compounds and molecules are comprised of deuterium nuclei (not hydrogen as currently supposed) the orientation and spin direction of electrons bound within these nuclei results in the diamagnetic and paramagnetic properties of various macro-scale materials

Tetryonics theory reveals that the mass charge ratio is of particles are measure of the constituent mass energy geometries that go into making the particle's final Matter topology that we observe a measure

But unlike the current mathematical approach the geometric approach of Tetryonics reveals the hidden mass energy partitions within the Matter topology of all subatomic particles and that particles such as tetryons (the quantum building blocks of all Matter) are in fact hidden within the debris of current collider results.

4 quantum charge tetryons have the same1/3 partial elementary charge as some quarks, and can be measured as having the same mass charge ratio's as charged leptons such as positrons and electrons in collider experiments.

………. To be continued [QED, Chemistry, Cosmology, Mathematics, Biology and economic theory]

answers@tetryonics.com

ABRAHAM [2014]

principia geometrica [3b] - quantum chemistry ________ 2
principia geometrica [4] - quantum cosmology ________ 230
principia geometrica [5] - quantum mathematics ________ 461
Principia Scientifica [5] - Quantum Mathematics ________ 624
Principia Scientifica [3a] - Quantum Chemistry ________ 787
principia geometrica [1] - quantum mechanics ________ 991
principia geometrica [2] - quantum electrodynamics ________ 1232
principia geometrica [3a] - quantum chemistry ________ 1451

Copyright ABRAHAM [2008] - All rights reserved

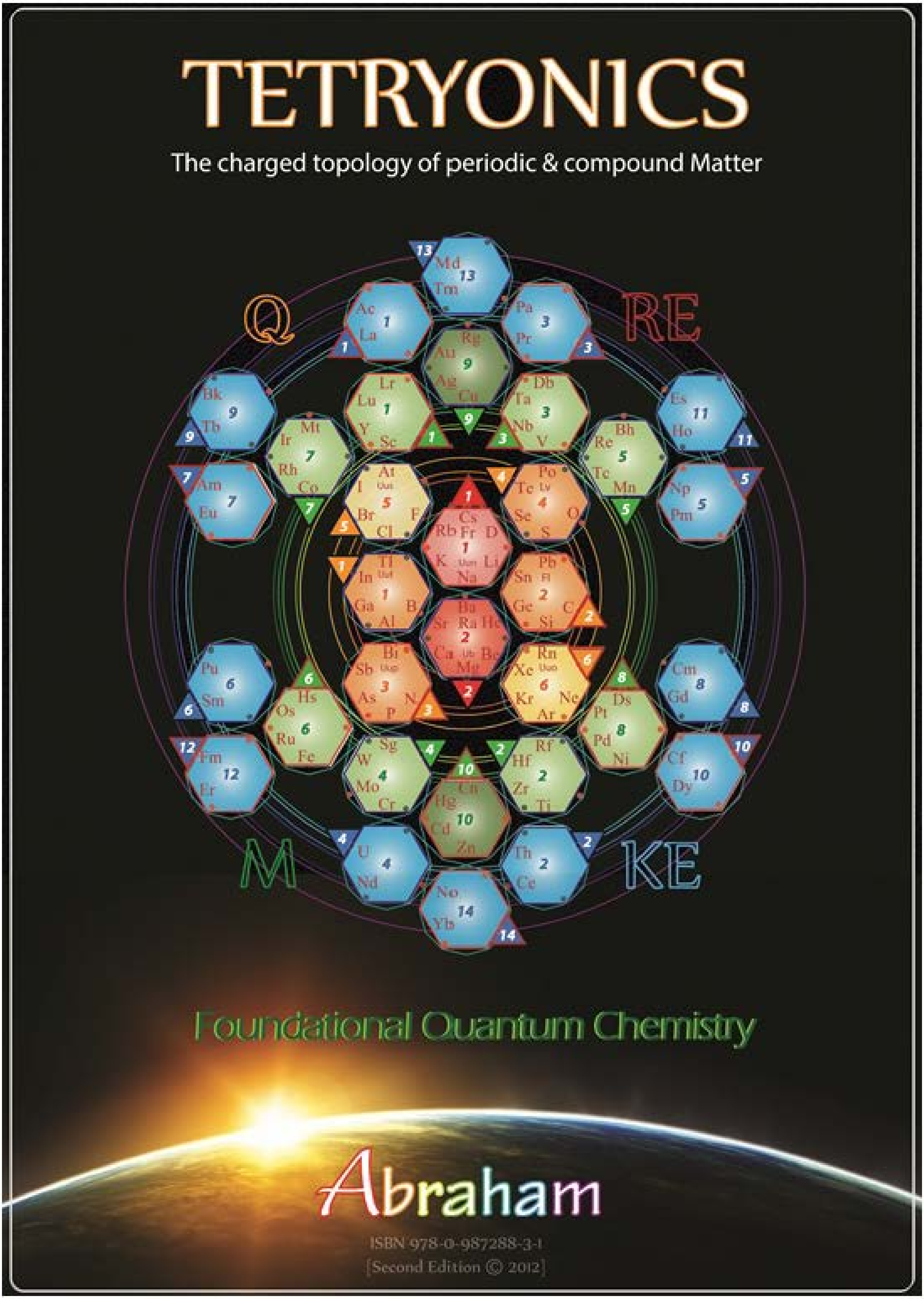

TETRYONICS

The charged topology of periodic & compound Matter

Q
RE
M
KE
Md 13 Tm
Ac 1 La
Pa 3 Pr
Rg Au 9 Ag Cu
Lr Lu 1 Y Sc
Db Ta 3 Nb V
Bk 9 Tb
Mt Ir 7 Rh Co
Bh Re 5 Tc Mn
Es 11 Ho
Am 7 Eu
At I Uus 5 Br F Cl
Cs Rb Fr D 1 K Uun Li Na
Po Te Lv 4 Se O S
Np 5 Pm
Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si
Ba Sr Ra He 2 Ca Ub Be Mg
Bi Sb Uup 3 As N P
Rn Xe Uuo 6 Kr Ne Ar
Pu 6 Sm
Hs Os 6 Ru Fe
Ds Pt 8 Pd Ni
Cm 8 Gd
Fm 12 Er
Sg W 4 Mo Cr
Cn Hg 10 Cd Zn
Rf Hf 2 Zr Ti
Cf 10 Dy
U 4 Nd
Th 2 Ce
No 14 Yb

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1
[Second Edition © 2012]

Periodic element Families

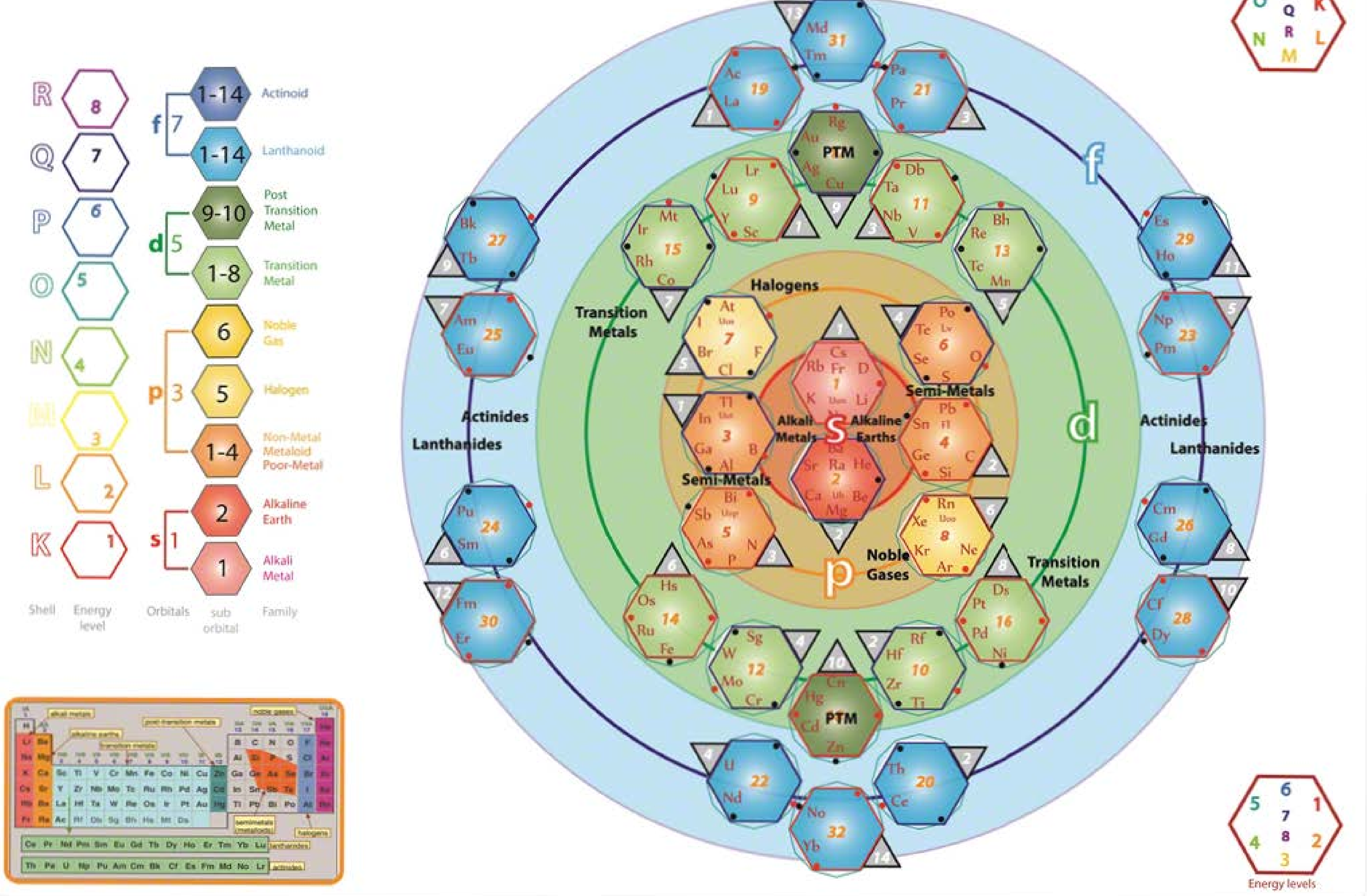
Atomic Shells
O
P
Q
K
N
R
L
M
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
f 7
1-14
Actinoid
1-14
Lanthanoid
d 5
9-10
Post Transition Metal
1-8
Transition Metal
p 3
6
Noble Gas
5
Halogen
1-4
Non-Metal Metaloid Poor-Metal
s 1
2
Alkaline Earth
1
Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family
f
d
s
p
PTM
PTM
Halogens
Transition Metals
Transition Metals
Semi-Metals
Semi-Metals
Alkali Metals
Alkaline Earths
Noble Gases
Actinides
Lanthanides
Actinides
Lanthanides
Energy levels
5
6
1
7
8
4
2
3

Copyright ABRAHAM [2008] - All rights reserved

Alkali Metals

The alkali metals are silver-colored (caesium has a golden tinge), soft, low-density metals, which react readily with halogens to form ionic salts, and with water to form strongly alkaline (basic) hydroxides.

These elements all have one electron in their outermost shell, so the energetically preferred state of achieving a filled electron shell is to lose one electron to form a singly charged positive ion.

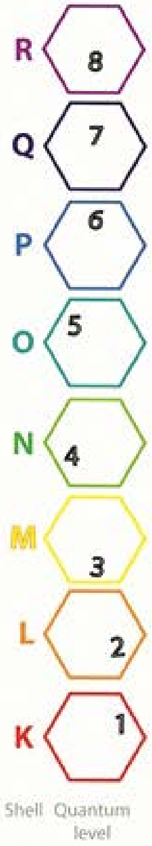
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell
Quantum level

S
1
Alkali Metal

s Orbital

n1-8

s1 sub-orbital

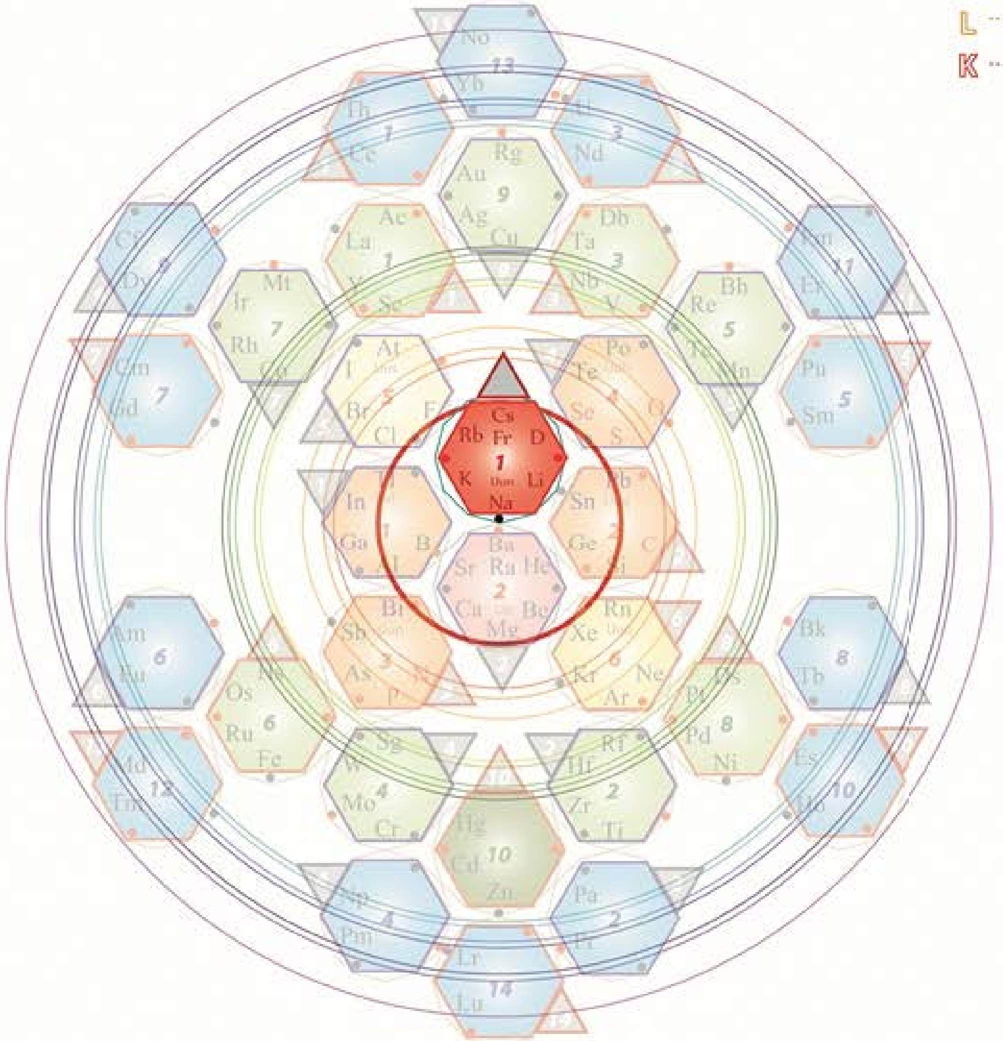
Cs
Fr
D
Rb
1
K
Li
Na
Ba
Sr
Ra
He
2
Ca
Be
Mg
No
13
Yb
Th
1
Ce
U
3
Nd
Rg
Au
9
Ag
Cu
Ac
La
1
Y
Sc
Db
Ta
3
Nb
V
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
Cf
8
Dy
Fm
11
Er
Cm
7
Gd
Pu
5
Sm
At
I
Br
Cl
F
Po
Te
Se
S
O
In
Ga
B
Al
Sn
Ge
C
Si
Bi
Sb
As
P
N
Rn
Xe
Kr
Ar
Ne
Am
6
Eu
Bk
8
Tb
Hs
Os
6
Ru
Fe
Ds
Pt
8
Pd
Ni
Md
12
Tm
Es
10
Ho
Sg
W
4
Mo
Cr
Rf
Hf
2
Zr
Ti
Hg
10
Cd
Zn
Np
4
Pm
Pa
2
Pr
Lr
14
Lu

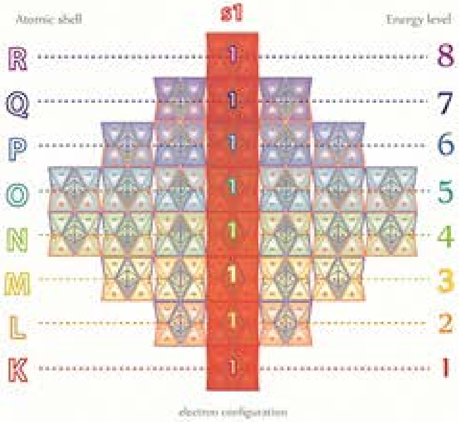
Atomic shell
s1
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
1
1
1
1
1
1
1
electron configuration

8

Alkaline Metals

6
5
7
1
4
8
2
3
Quantum level

The alkali metals are all highly reactive
and are never found in elemental form in nature.

As a result, in the laboratory they are stored under mineral oil.
They also tarnish easily and have low melting points and densities.
Potassium and rubidium possess a weak radioactive characteristic
(harmless) due to the presence of long duration radioactive isotopes.

Alkaline Earths

The alkaline earth metals are silver colored, soft metals, which react readily with halogens to form ionic salts, and with water, though not as rapidly as the alkali metals, to form strong alkaline (basic) hydroxides.

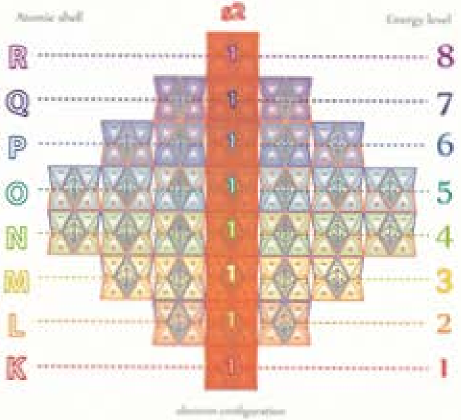
Atomic shell
s2
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
electron configuration

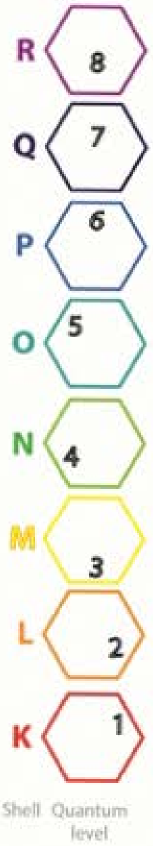
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell
Quantum level

S
2
Alkaline Earth

s Orbital

n1-8

s2 sub-orbital

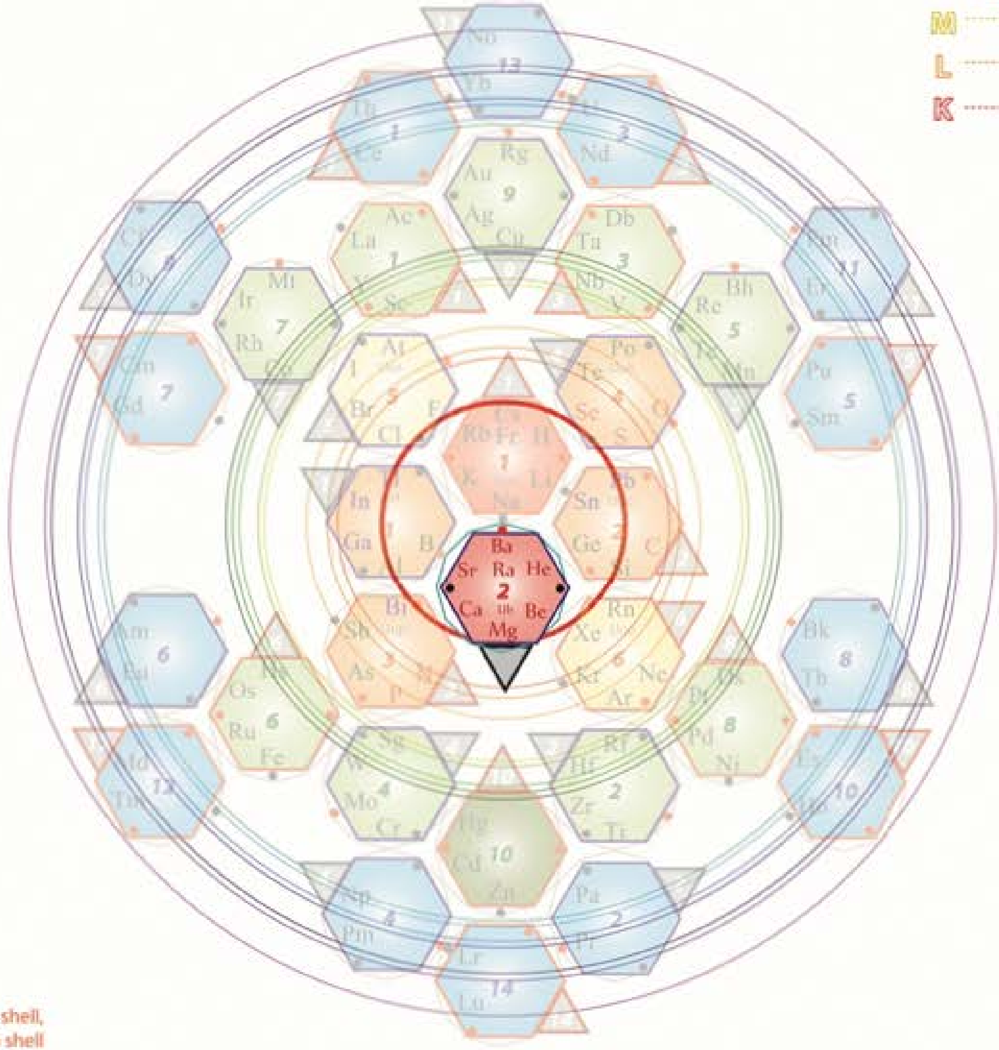
Ba
Sr
Ra
He
2
Ca
Be
Mg
Rb
Fe
H
K
Li
Na
1
Sn
Ge
In
Ga
Xe
Kr
Ar
Ne
Rn
Au
Ag
Cu
Rg
9
Hg
Cd
Zn
10
Lr
Lu
14
Yb
13

8
Alkaline Earths

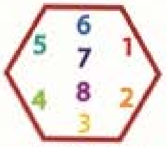
6
7
8
5
1
4
2
3
Quantum level

All the alkaline earth metals have two electrons in their valence shell, so the energetically preferred state of achieving a filled electron shell is to lose two electrons to form doubly charged positive ions.

Metalloids

A metaloid is a material with a small overlap in the energy of the conduction band and valence bands.

Unlike a regular metal, metaloids have been described electrically as charge carriers of both types (holes and elecrons), so it could be argue that they should be called 'double-metals' rather than metaloids.

As metaloids have fewer charge carriers than metals, they typically have lower electrical and thermal conductivities.

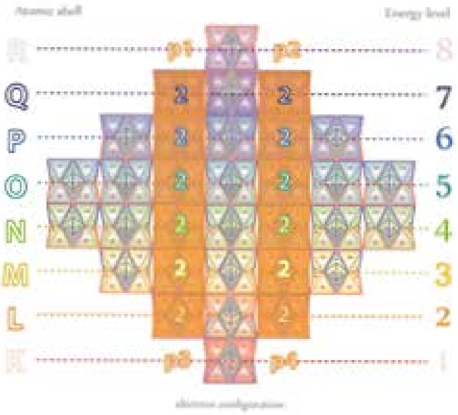
Atomic shell
Energy level
p1
p2
p3
p4
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
electron configuration

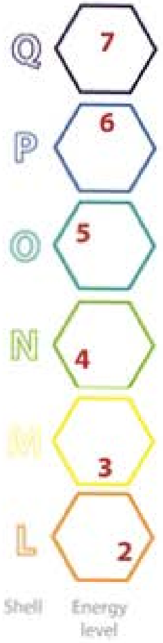
Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Energy level

p
1-4
Non-Metal
Metaloid
Poor-Metal

2 p-Orbitals

n2-7

p1-4 sub-orbitals

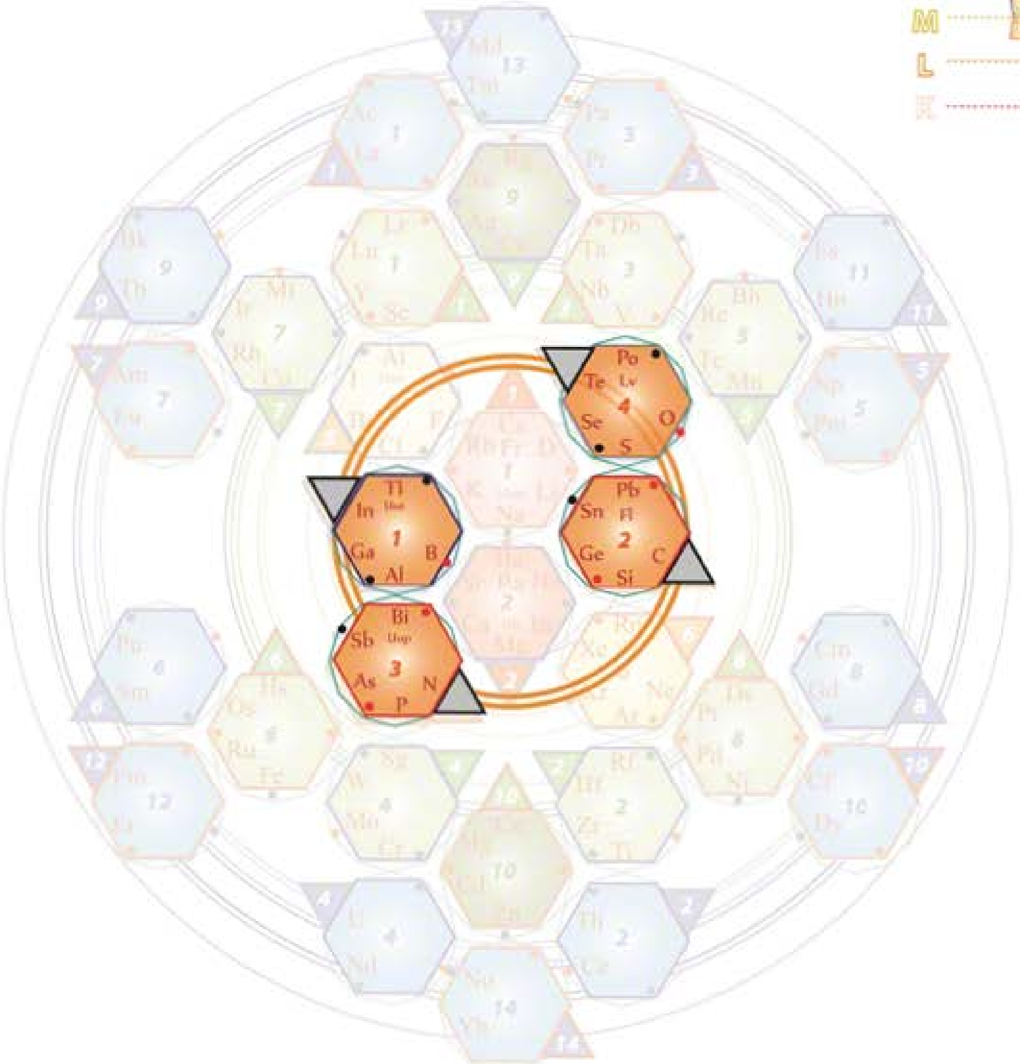
Po
Te
Lv
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P

24 Semi-Metals

6
5
7
4
2
3
Energy levels

Charge carriers typically occur in much smaller numbers than in a real metal. In this respect they resemble degenerate semiconductors more closely.
This explains why the electrical properties of metaloids are partway between those of metals and semiconductors.

Halogens

Owing to their high reactivity, the halogens are found in the environment only in compounds or as ions.

Halide ions and oxoanions such as iodate (IO3−) can be found in many minerals and in seawater.

Halogenated organic compounds can also be found as natural products in living organisms. In their elemental forms, the halogens exist as diatomic molecules, but these only have a fleeting existence in nature and are much more common in the laboratory and in industry.

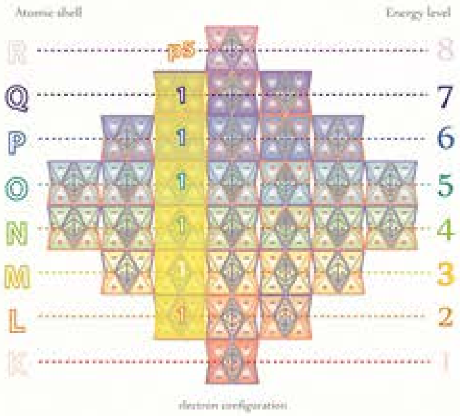
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
p5
1
1
1
1
1
1
electron configuration

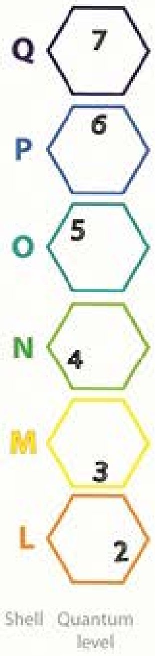
Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Quantum level

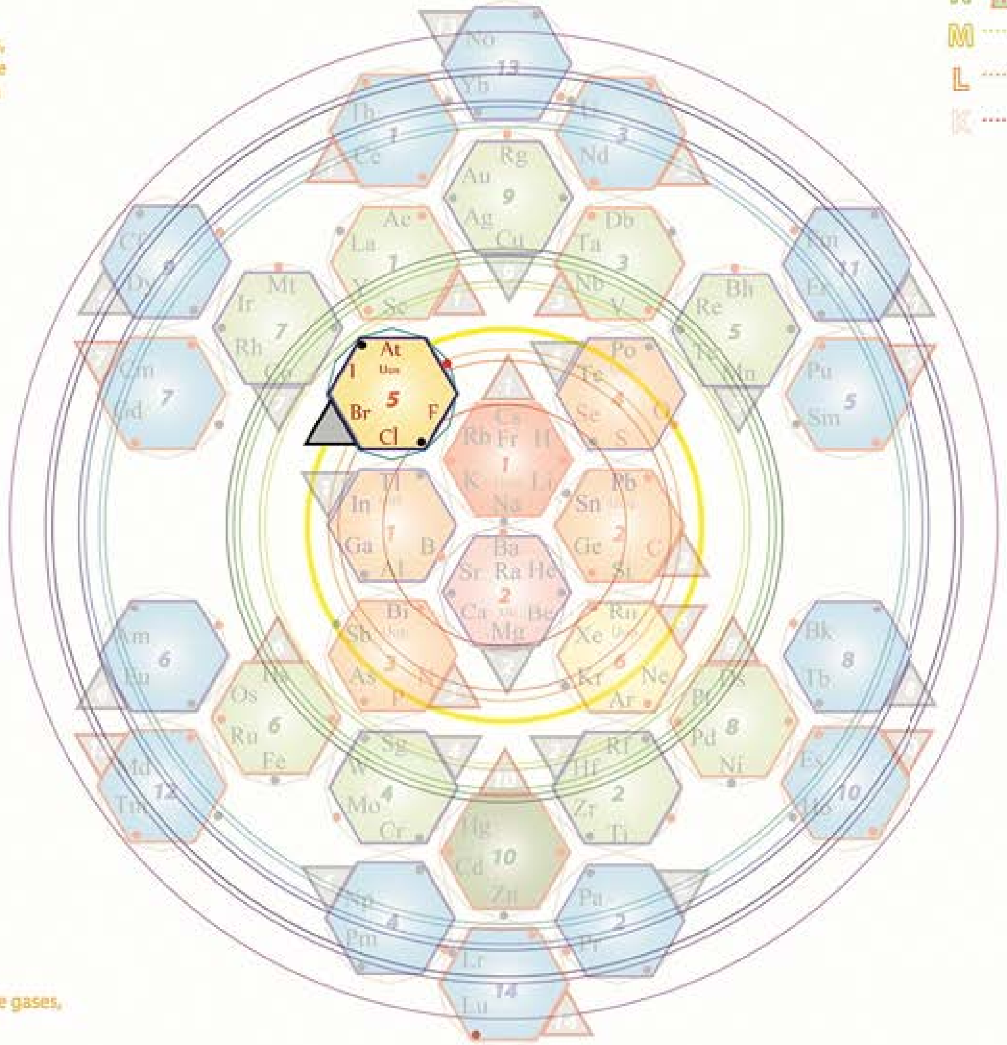
At
Uus
I
5
Br
F
Cl
No
Yb
13
Tb
Ce
1
Nd
3
Rg
Au
9
Ag
Cu
Ac
La
Db
Ta
Nb
V
Sc
Mt
Ir
Rh
7
Co
Bh
Re
Tc
Mn
5
Cf
Dy
Fm
Er
11
Cm
Gd
7
Pu
Sm
5
Po
Te
Se
S
O
Rb
Fr
H
K
Li
Na
1
Tl
In
Ga
B
Al
Pb
Sn
Ge
Si
C
2
Ba
Sr
Ra
He
Ca
Be
Mg
2
Bi
Sb
As
P
N
3
Rn
Xe
Kr
Ar
Ne
6
Am
Eu
6
Bk
Tb
8
Hs
Os
Ru
Fe
6
Ds
Pt
Pd
Ni
8
Md
Tm
12
Es
Ho
10
Sg
W
Mo
Cr
4
Rf
Hf
Zr
Ti
2
Hg
Cd
Zn
10
Np
Pm
4
Pa
Pr
2
Lr
Lu
14

p3 Orbital

n2-7

p5 sub-orbital

6

Halogens

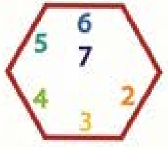

Quantum level

At room temperature and pressure, fluorine and chlorine are gases, bromine is a liquid and iodine and astatine are solids.

Group 17 is therefore the only periodic table group exhibiting all three states of matter at room temperature

Copyright ABRAHAM [2008] - All rights reserved

Nobel Gases

The properties of the noble gases can be well explained by modern theories of atomic structure: their outer shell of valence electrons is considered to be "full", giving them little tendency to participate in chemical reactions, and only a few hundred noble gas compounds have been prepared.

The melting and boiling points for each noble gas are close together, differing by less than 10 °C (18 °F); consequently, they are liquids only over a small temperature range.

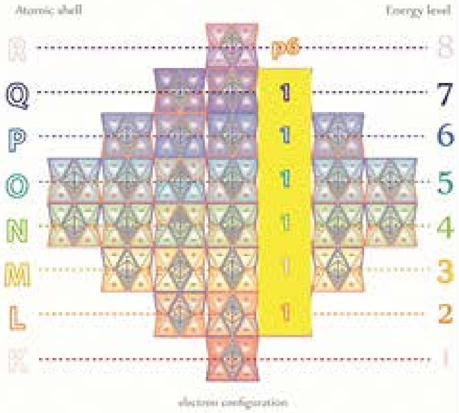
Atomic shell
Energy level
R
p6
8
Q
1
7
P
1
6
O
1
5
N
1
4
M
1
3
L
1
2
K
1
electron configuration

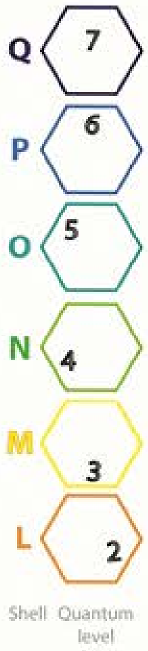
Q
7
P
6
O
5
N
4
M
3
L
2
Shell
Quantum level

p
6
Noble Gas

completed p3 Orbital

n2-7

p6 sub-orbital

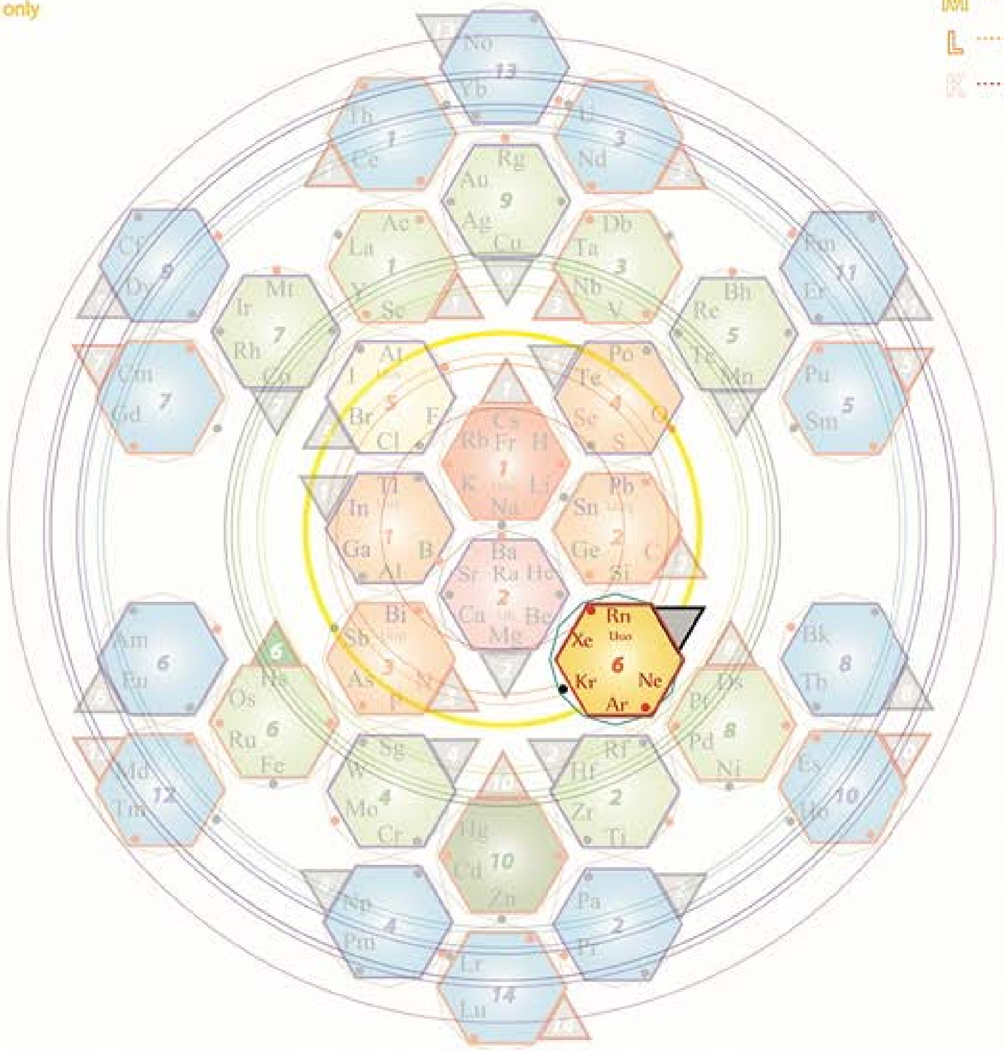
No
13
Yb
Th
1
Ce
U
3
Nd
Rg
Au
9
Ag
Cu
Ac
La
1
Y
Sc
Db
Ta
3
Nb
V
Cf
9
Dy
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
Fm
11
Er
Cm
7
Gd
At
I
5
Br
F
Cl
Po
Te
4
Se
O
S
Cs
Rb
Fr
H
1
K
Li
Na
Pu
5
Sm
Tl
In
1
Ga
B
Al
Pb
Sn
2
Ge
C
Si
Ba
Sr
Ra
He
2
Ca
Be
Mg
Bi
Sb
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar
Am
6
Eu
Hs
Os
6
Ru
Fe
Ds
Pt
8
Pd
Ni
Bk
8
Tb
Md
12
Tm
Sg
W
4
Mo
Cr
Hg
10
Cd
Zn
Rf
Hf
2
Zr
Ti
Es
10
Ho
Np
4
Pm
Pa
2
Pr
Lr
14
Lu

6

Nobel Gases

Neon, argon, krypton, and xenon are obtained from air using the methods of liquefaction of gases and fractional distillation.

Helium is typically separated from natural gas, and radon is usually isolated from the radioactive decay of dissolved radium compounds.

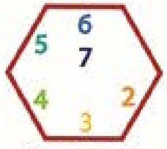

Quantum level

Transitional Metals

In chemistry, the term transition metal commonly refers to any element in the d-block of the periodic table, including the group 12 elements zinc, cadmium and mercury.

This corresponds to groups 3 to 12 on the periodic table, which are all metals.

More strictly, IUPAC defines a transition metal as "an element whose atom has an incomplete d sub-shell, or which can give rise to cations with an incomplete d sub-shell."

The first definition is simple and has traditionally been used. However, many interesting properties of the transition elements as a group are the result of their partly filled d subshells.

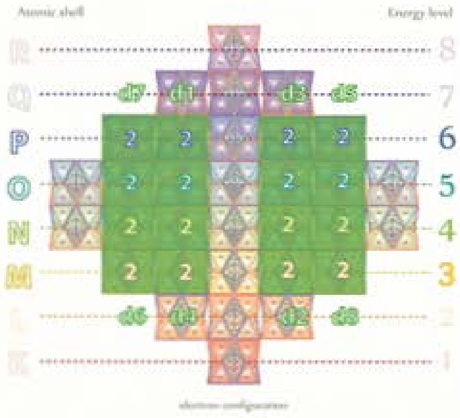
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
d7
d1
d3
d5
d6
d4
d2
d8
2
electron configuration

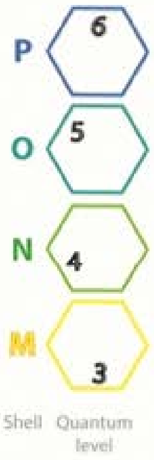
P
6
O
5
N
4
M
3
Shell
Quantum level

d
1-8
Transition Metal

4 d-Orbitals

n3-6

d1-8 sub-orbitals

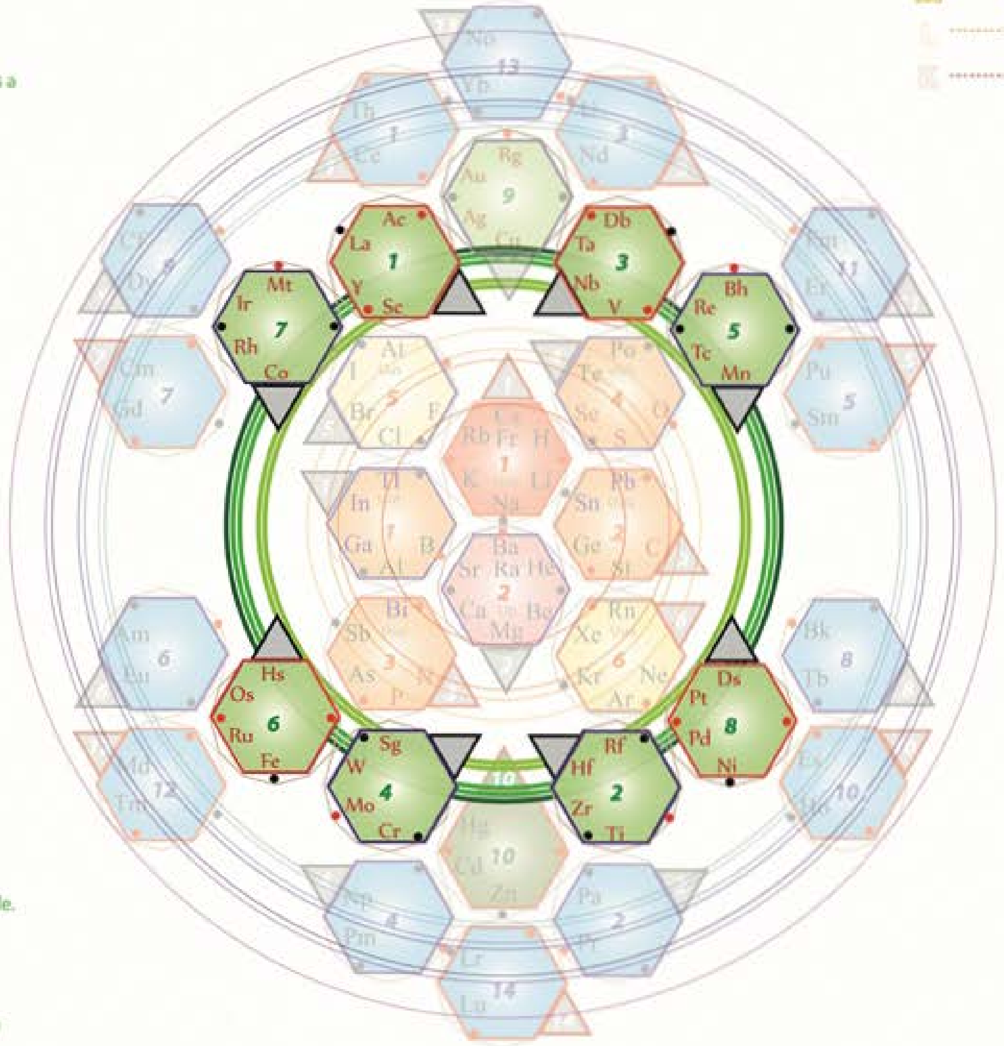
Ac
La
Y
Sc
1
Db
Ta
Nb
V
3
Mt
Ir
Rh
Co
7
Bh
Re
Tc
Mn
5
Hs
Os
Ru
Fe
6
Ds
Pt
Pd
Ni
8
Sg
W
Mo
Cr
4
Rf
Hf
Zr
Ti
2
Rg
Au
Ag
Cu
9
Hg
Cd
Zn
10

32 Transition Metals

6
5
4
3
Quantum level

The name transition comes from their position in the periodic table. In each of the four periods in which they occur, these elements represent the successive addition of electrons to the d atomic orbitals of the atoms.

In this way, the transition metals represent the transition between group 2 elements and group 13 elements.

Post-Transition Metals

In chemistry, the term post-transition metal is used to describe the category of metallic elements to the right of the transition elements on the periodic table.

There are two IUPAC definitions of "transition element" that have been in apparent conflict with one another since September 2007.

According to the first definition, transition metals are elements in group 3 through group 11.
In this case, post-transition metals include all of group 12—zinc, cadmium, mercury, and ununbium

According to the second definition, transition elements either have an incomplete d-subshell or have the ability to form an incomplete d-subshell.

Post Transitional metals are normally defined as completed d orbitals

Tetryonics suggests it may be appropriate to include all the d5 orbital elements in this grouping

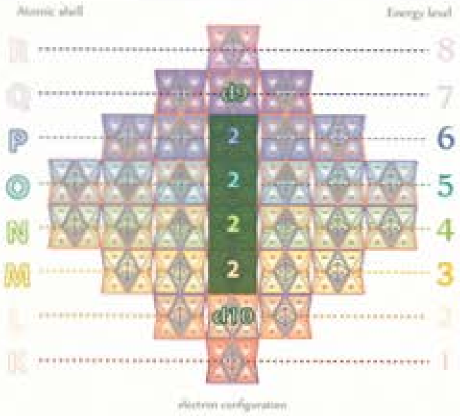

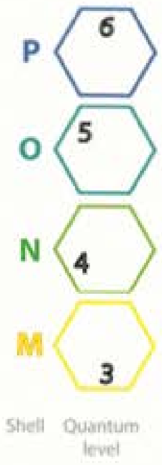

d 9-10 Post Transition Metal

d5 Orbital

n3-6

d9-10 sub-orbitals

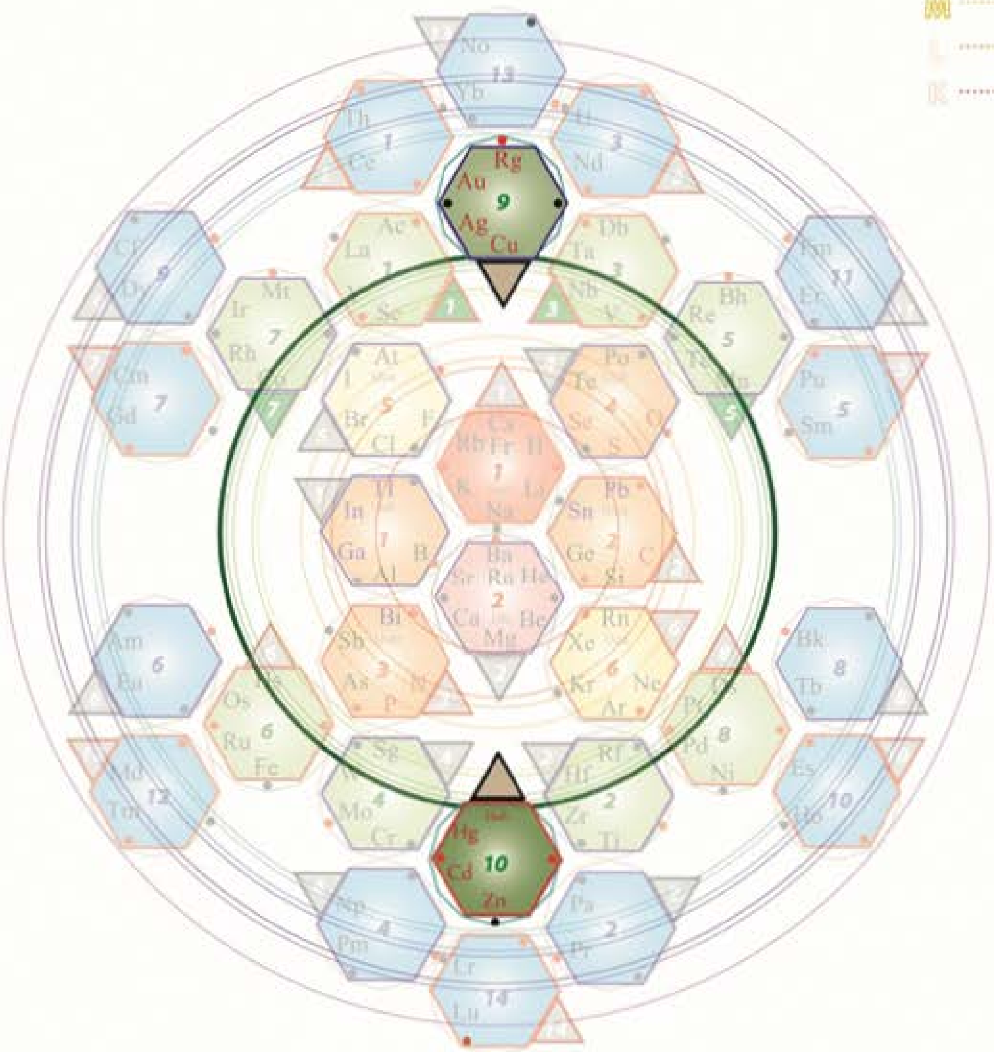

8 Post-transitional Metals

Quantum level

In 2007, mercury(IV) fluoride was synthesized.[2][3] This compound contains a mercury atom with an incomplete d-subshell, and ununbium is predicted to have the capacity to form a similar electronic configuration.

In this case, post-transition metals include only zinc and cadmium within group 12.

Copyright ABRAHAM [2008] - All rights reserved

Lanthanoids

All lanthanoids are f-block elements, corresponding to the filling of the 4f electron shell

The lanthanoid series (Ln) is named after Lanthanum.

The trivial name "rare earths" is sometimes used to describe all the lanthanoids together with scandium and yttrium.

These elements are in fact fairly abundant in nature, although rare as compared to the "common" earths such as lime or magnesia. Cerium is the 26th most abundant element in the Earth's crust, neodymium is more abundant than gold and even thulium (the least common naturally-occurring lanthanoid) is more abundant than iodine.

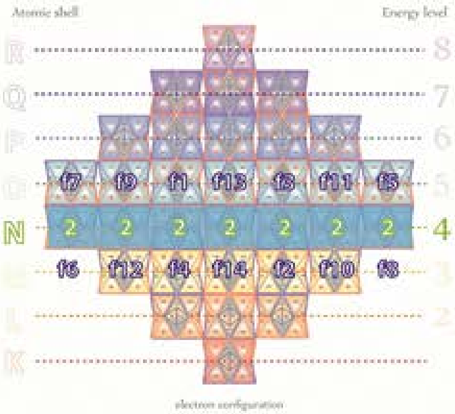
Atomic shell
Energy level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K
f7 f9 f1 f13 f3 f11 f5
2 2 2 2 2 2 2
f6 f12 f4 f14 f2 f10 f8
electron configuration

f
1-14
Lanthanoid

N
4
Shell
Quantum level

7 f-Orbitals

n4

f1-14 sub-orbitals

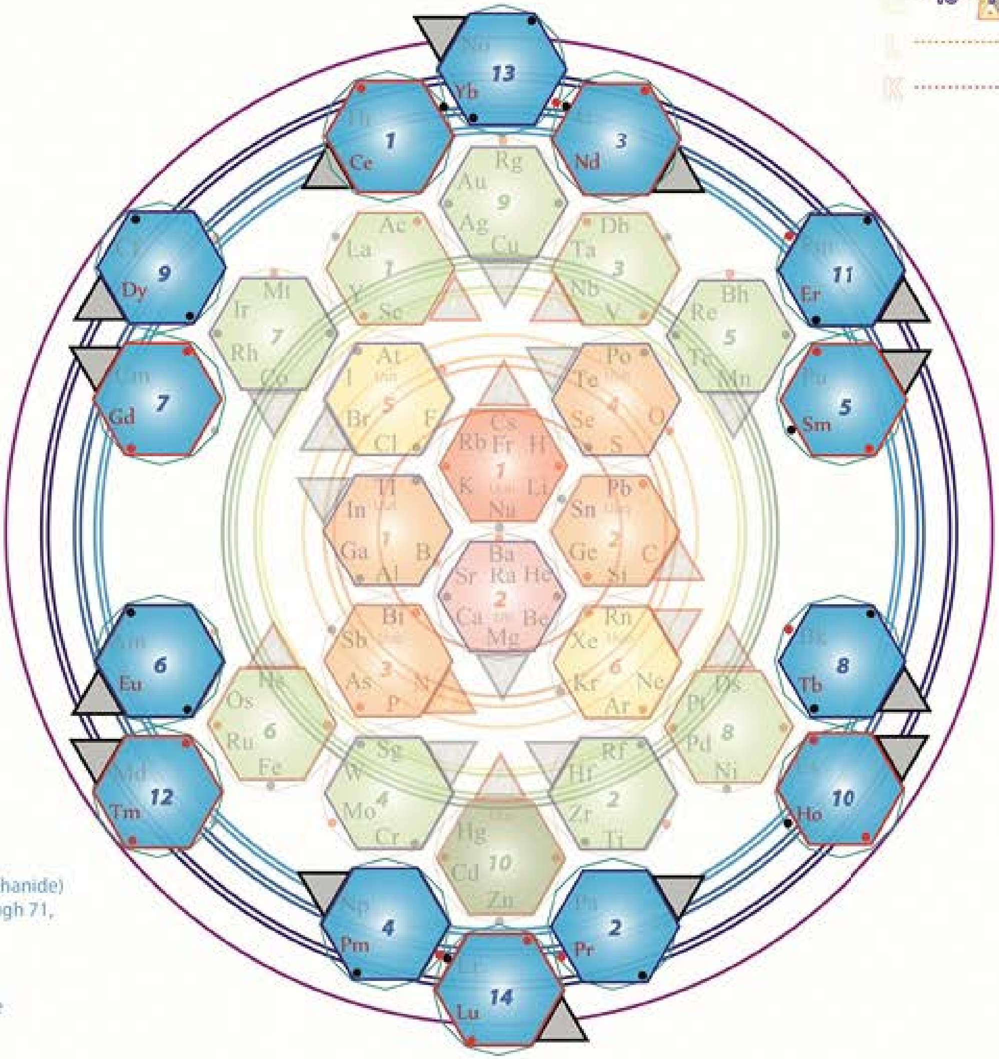
13 Yb
1 Ce
3 Nd
9 Dy
11 Er
7 Gd
5 Sm
6 Eu
8 Tb
12 Tm
10 Ho
4 Pm
2 Pr
14 Lu

14 Lanthanoids

The term "rare earths" arises from the minerals from which they were isolated, which were uncommon oxide-type minerals.

According to the IUPAC terminology, the lanthanoid (previously lanthanide) series comprises the fifteen elements with atomic numbers 57 through 71, from Lanthanum to Lutetium.

The use of this name is deprecated by IUPAC, as they are neither rare in abundance nor "earths" (an obsolete term for water-insoluble strongly basic oxides of electropositive metals incapable of being smelted into metal using late 18th century technology).

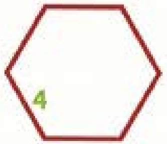

Quantum level

Actinoids

According to IUPAC nomenclature, the actinoid (previously actinide) series encompasses the 15 chemical elements that lie between actinium and lawrencium included on the periodic table, with atomic numbers 89 - 103.

The actinoid series derives its name from the first element in the series, actinium, and ultimately from the Greek ακτις (aktis), "ray," reflecting the elements' radioactivity.

The actinoids display less similarity in their chemical properties than the lanthanoid series (Ln), exhibiting a wider range of oxidation states, which initially led to confusion as to whether actinium, thorium, and uranium should be considered d-block elements. All actinoids are radioactive.

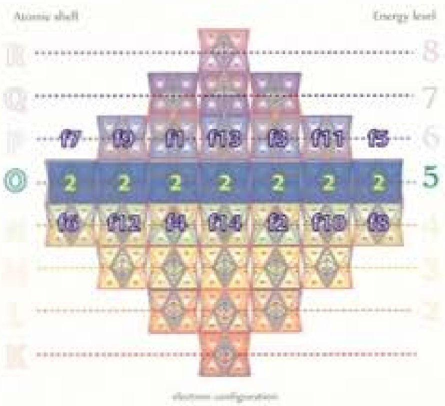
Atomic shell
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
f7 f9 f1 f13 f3 f11 f5
2 2 2 2 2 2 2
f6 f12 f4 f14 f2 f10 f8
electron configuration

f 1-14 Actinoid

o 5

Shell Quantum level

7 f-Orbitals

n5

f1-14 sub-orbitals

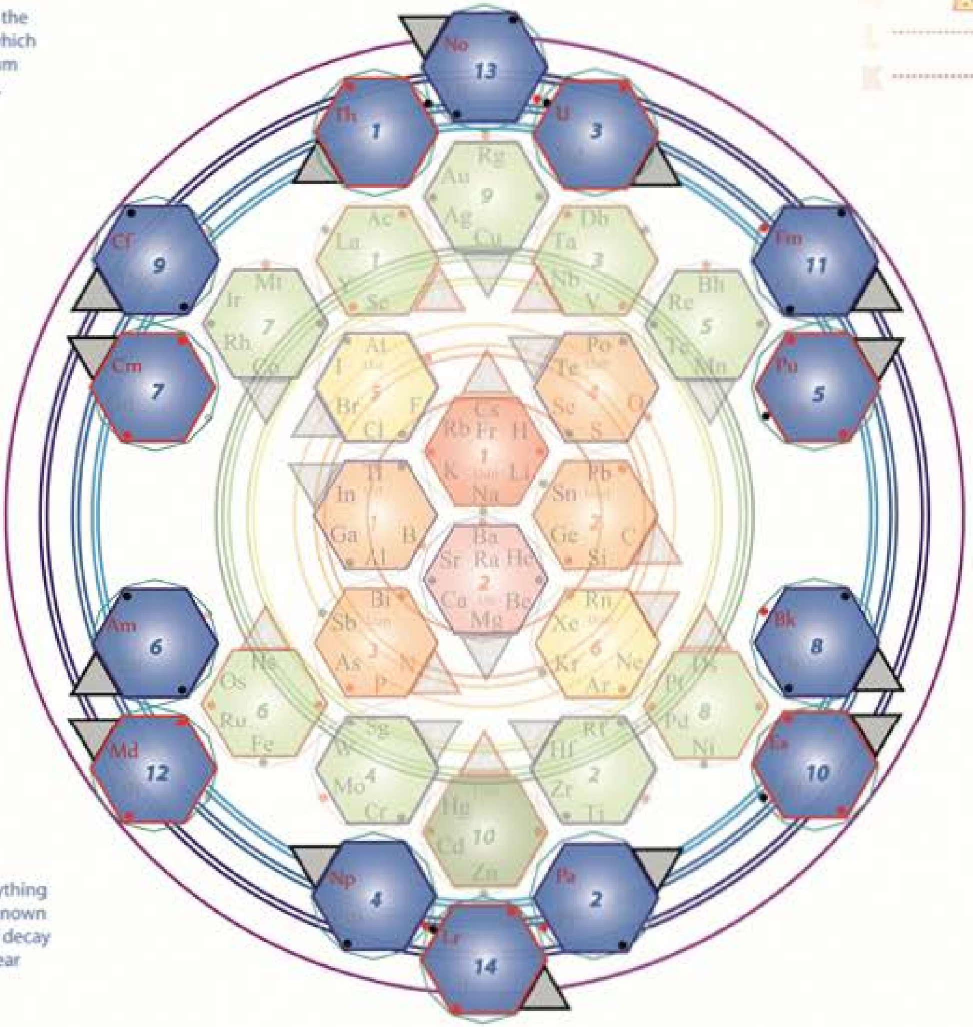
No 13
Th 1
U 3
Cf 9
Fm 11
Cm 7
Pu 5
Am 6
Bk 8
Md 12
Es 10
Np 4
Pa 2
Lr 14

14 Actinoids

5

Quantum level

Only thorium and uranium occur naturally in the earth's crust in anything more than trace quantities. Neptunium and plutonium have been known to show up naturally in trace amounts in uranium ores as a result of decay or bombardment. The remaining actinoids were discovered in nuclear fallout, or were synthesized in particle colliders.

The latter half of the series possess exceedingly short half-lives.

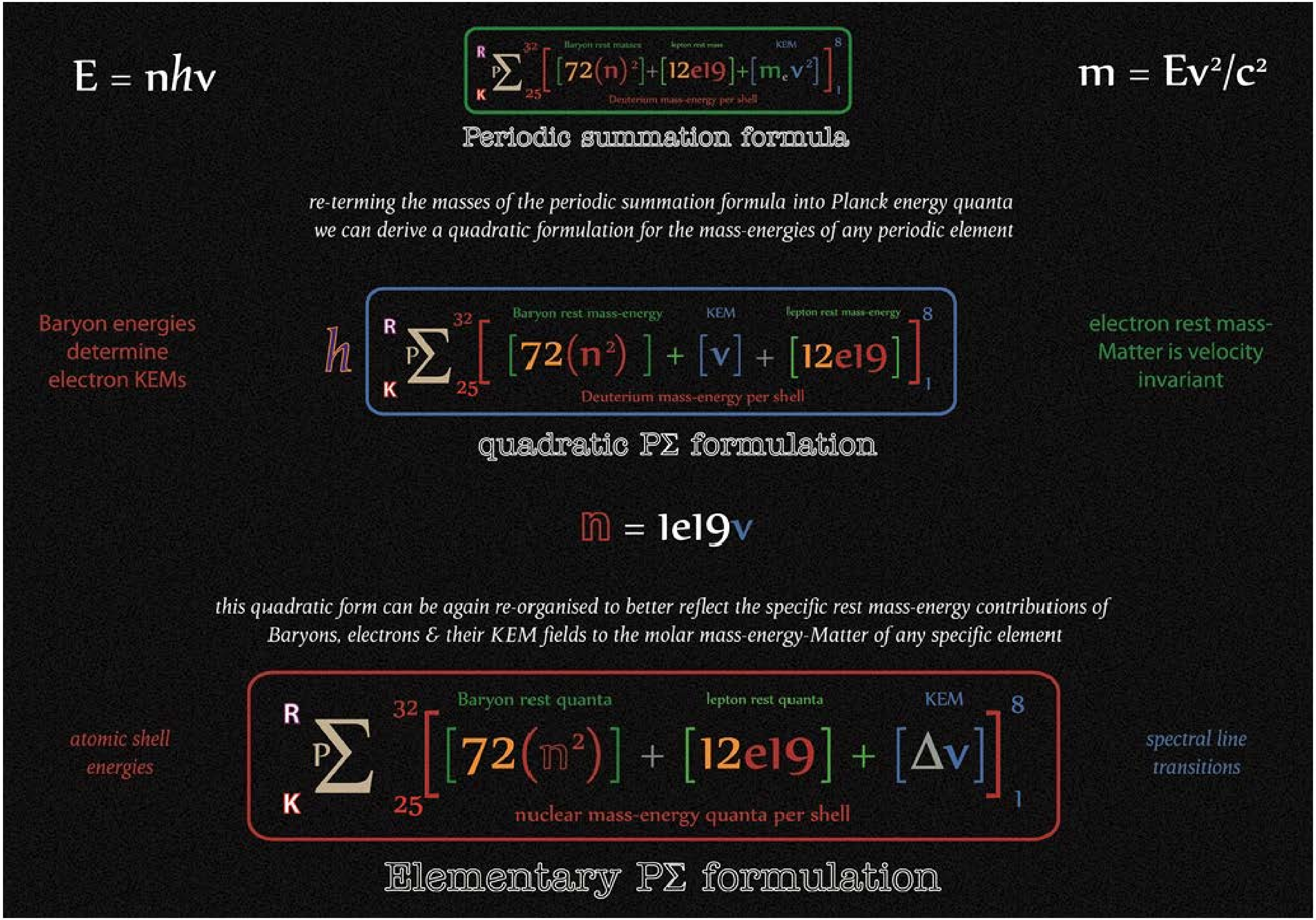

$$E = nhv$$

Baryon rest masses

lepton rest mass

KEM

$$ {}^{R}_{K}\,{}^{}_{P}\sum_{25}^{32}\Big[[72(n)^{2}]+[12e19]+[m_{e}v^{2}]\Big]_{1}^{8} $$

Deuterium mass-energy per shell

Periodic summation formula

$$m = Ev^{2}/c^{2}$$

re-terming the masses of the periodic summation formula into Planck energy quanta we can derive a quadratic formulation for the mass-energies of any periodic element

Baryon energies determine electron KEMs

Baryon rest mass-energy

KEM

lepton rest mass-energy

$$ h\;{}^{R}_{K}\,{}^{}_{P}\sum_{25}^{32}\Big[\,[72(n^{2})\,] + [v] + [12e19]\Big]_{1}^{8} $$

Deuterium mass-energy per shell

electron rest mass-Matter is velocity invariant

quadratic PΣ formulation

$$n = 1e19v$$

this quadratic form can be again re-organised to better reflect the specific rest mass-energy contributions of Baryons, electrons & their KEM fields to the molar mass-energy-Matter of any specific element

atomic shell energies

Baryon rest quanta

lepton rest quanta

KEM

$$ {}^{R}_{K}\,{}^{}_{P}\sum_{25}^{32}\Big[[72(n^{2})] + [12e19] + [\Delta v]\Big]_{1}^{8} $$

nuclear mass-energy quanta per shell

spectral line transitions

Elementary PΣ formulation

Tetryonics 52.13 - Periodic element summation formula

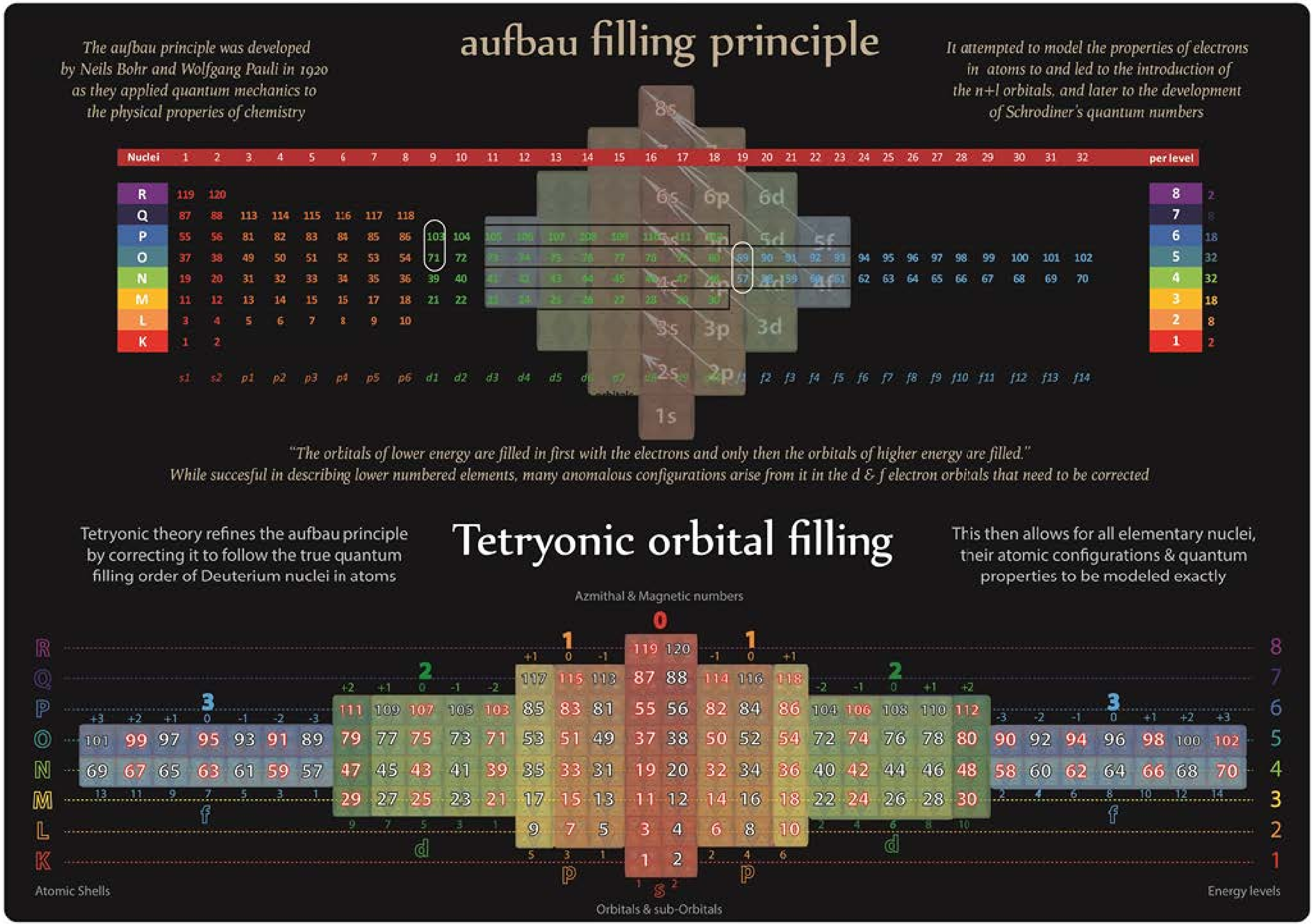

aufbau filling principle

The aufbau principle was developed by Neils Bohr and Wolfgang Pauli in 1920 as they applied quantum mechanics to the physical properies of chemistry

It attempted to model the properties of electrons in atoms to and led to the introduction of the n+l orbitals, and later to the development of Schrodiner's quantum numbers

"The orbitals of lower energy are filled in first with the electrons and only then the orbitals of higher energy are filled."

While succesful in describing lower numbered elements, many anomalous configurations arise from it in the d & f electron orbitals that need to be corrected

Tetryonic orbital filling

Tetryonic theory refines the aufbau principle by correcting it to follow the true quantum filling order of Deuterium nuclei in atoms

This then allows for all elementary nuclei, their atomic configurations & quantum properties to be modeled exactly

Nuclei
per level
R
Q
P
O
N
M
L
K
1s
2s
2p
3s
3p
3d
4s
4p
4d
4f
5s
5p
5d
5f
6s
6p
6d
7s
7p
8s
s1
s2
p1
p2
p3
p4
p5
p6
d1
d2
d3
d4
d5
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
f14
Azmithal & Magnetic numbers
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 52.14 - Tetryonic aufbau orbital filling

Copyright ABRAHAM [2008] - All rights reserved

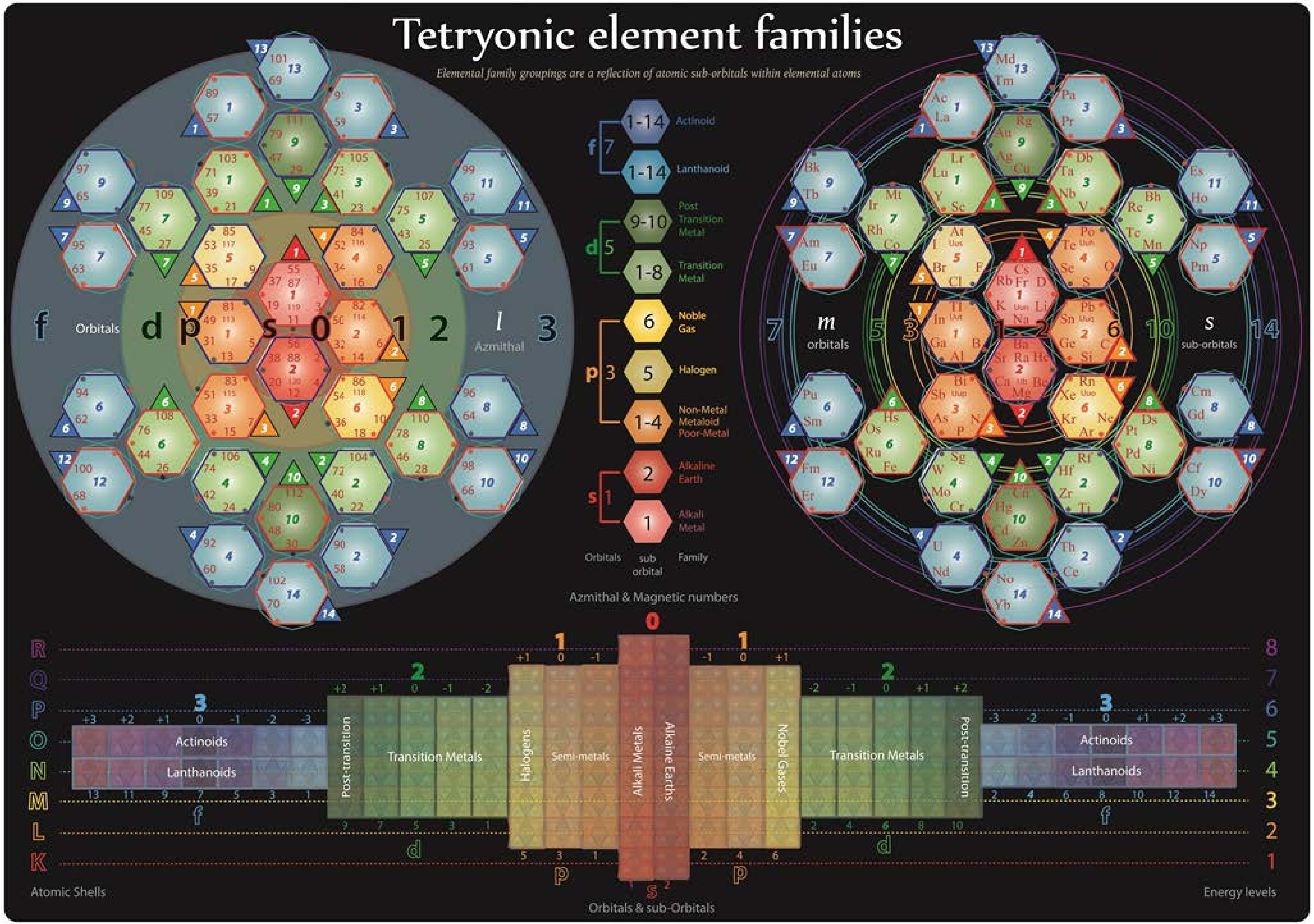
Tetryonic element families
Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Orbitals
sub orbital
Family
f
Orbitals
d
p
s
0
1
2
l
Azmithal
3
7
m
orbitals
5
3
1
2
6
10
14
s
sub-orbitals
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Actinoids
Lanthanoids
Post-transition
Transition Metals
Halogens
Semi-metals
Alkali Metals
Alkaine Earths
Semi-metals
Nobel Gases
Transition Metals
Post-transition
Actinoids
Lanthanoids
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 52.15 - Tetryonic elemental families

Copyright ABRAHAM [2008] - All rights reserved

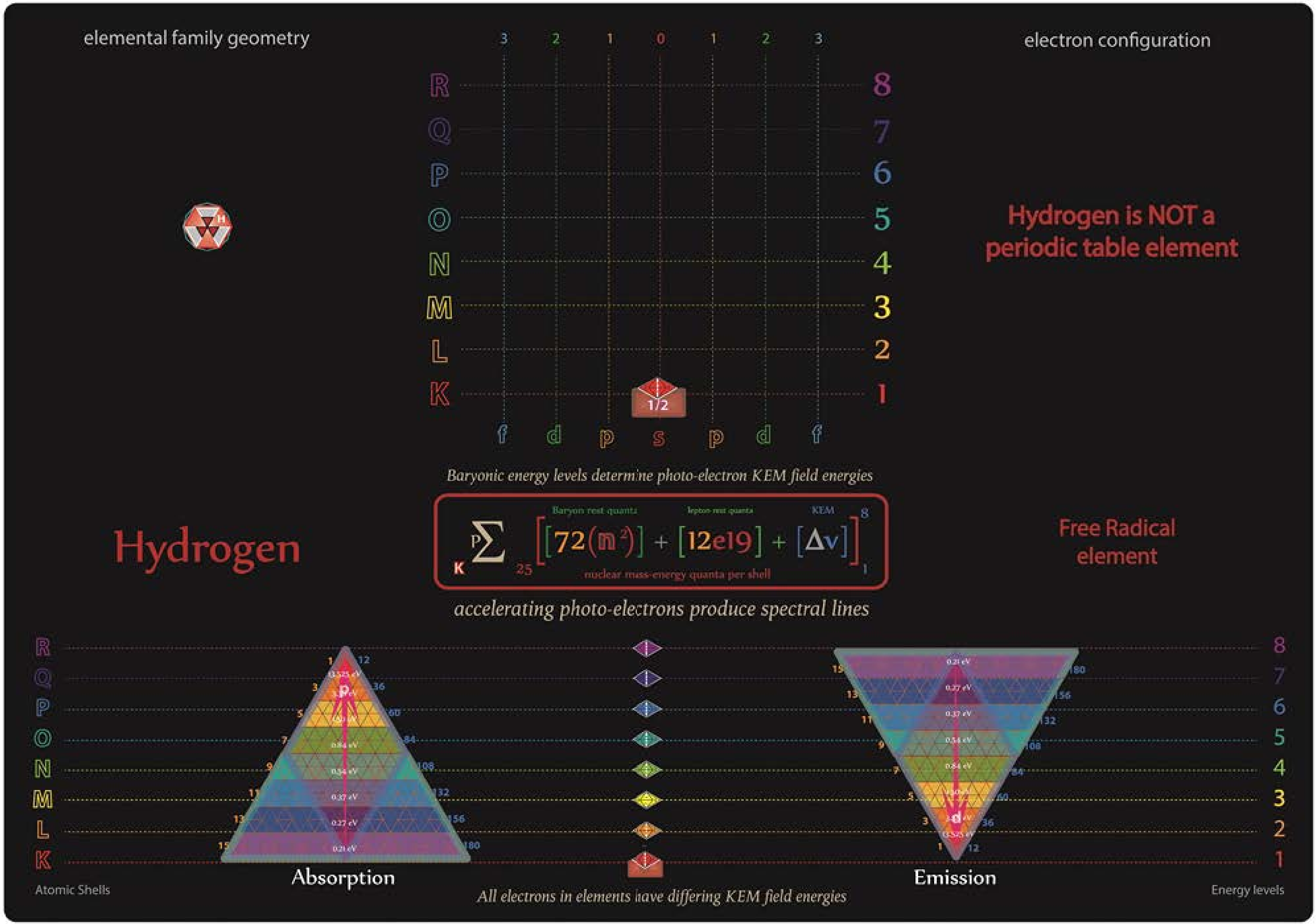
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1/2
f
d
p
s
p
d
f
H
1
12
13.525 eV
3
36
5
60
7
84
0.84 eV
9
108
0.54 eV
11
132
0.37 eV
13
156
0.27 eV
15
180
0.21 eV
Atomic Shells
Energy levels

elemental family geometry

electron configuration

Hydrogen is NOT a periodic table element

Hydrogen

Baryonic energy levels determine photo-electron KEM field energies

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{P}{}_{25}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{8}$$

nuclear mass-energy quanta per shell

accelerating photo-electrons produce spectral lines

Free Radical element

Absorption

Emission

All electrons in elements have differing KEM field energies

Tetryonics 53.00 - Hydrogen atomic config

Copyright ABRAHAM [2008] - All rights reserved

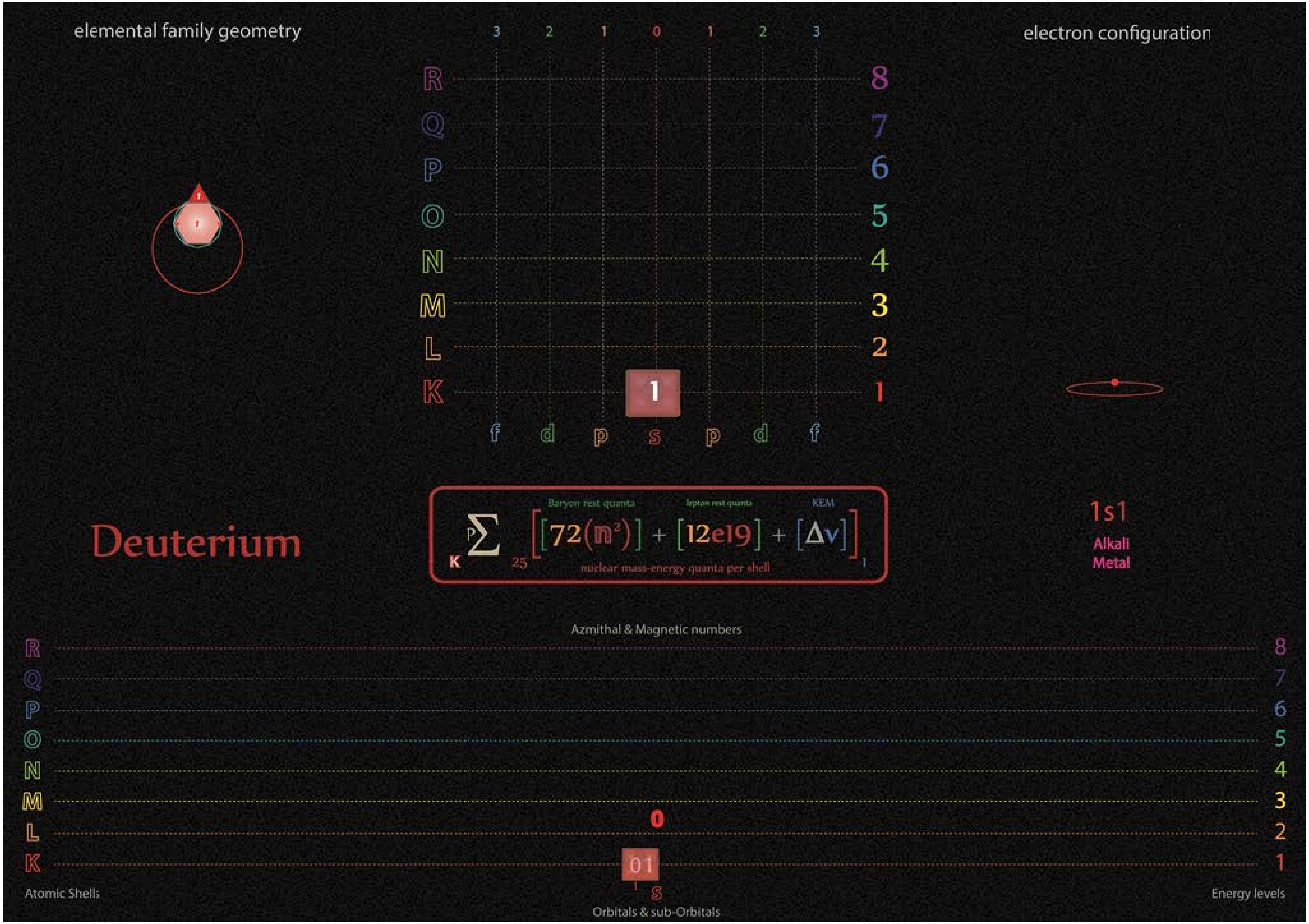

elemental family geometry

electron configuration

Deuterium

$${}_{K}\sum^{P}{}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}$$

nuclear mass-energy quanta per shell

1s1

Alkali Metal

Azmithal & Magnetic numbers

Atomic Shells

Orbitals & sub-Orbitals

Energy levels

Tetryonics 53.01 - Deuterium atomic config

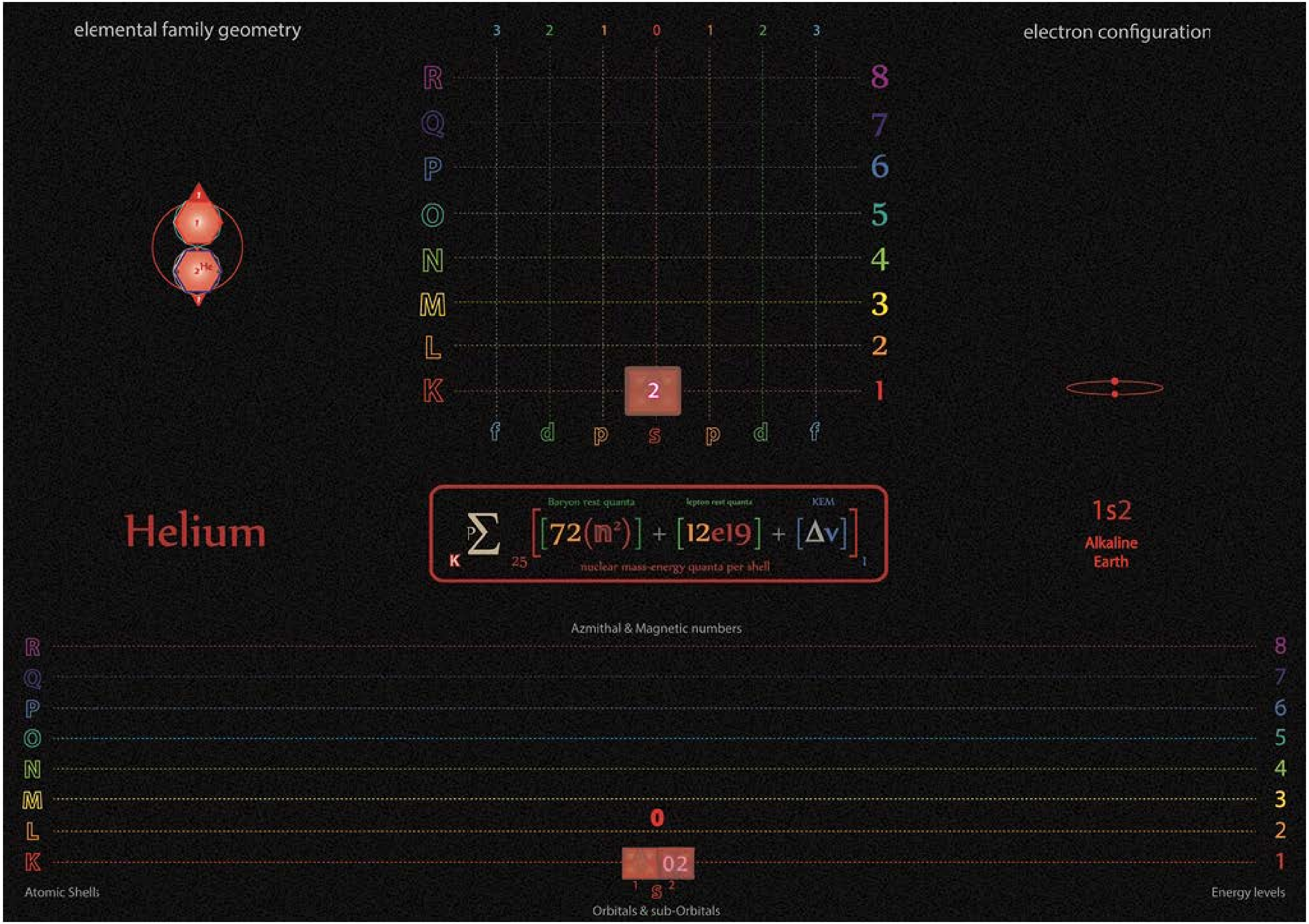
elemental family geometry
1
2He
Helium
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
f
d
p
s
p
d
f
electron configuration
1s2
Alkaline
Earth
Baryon rest quanta
lepton rest quanta
KEM
K
P
Σ
25
[[72(m²)] + [12e19] + [Δv]]
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
02
1
s
2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.02 - Helium atomic config

Copyright ABRAHAM [2008] - All rights reserved

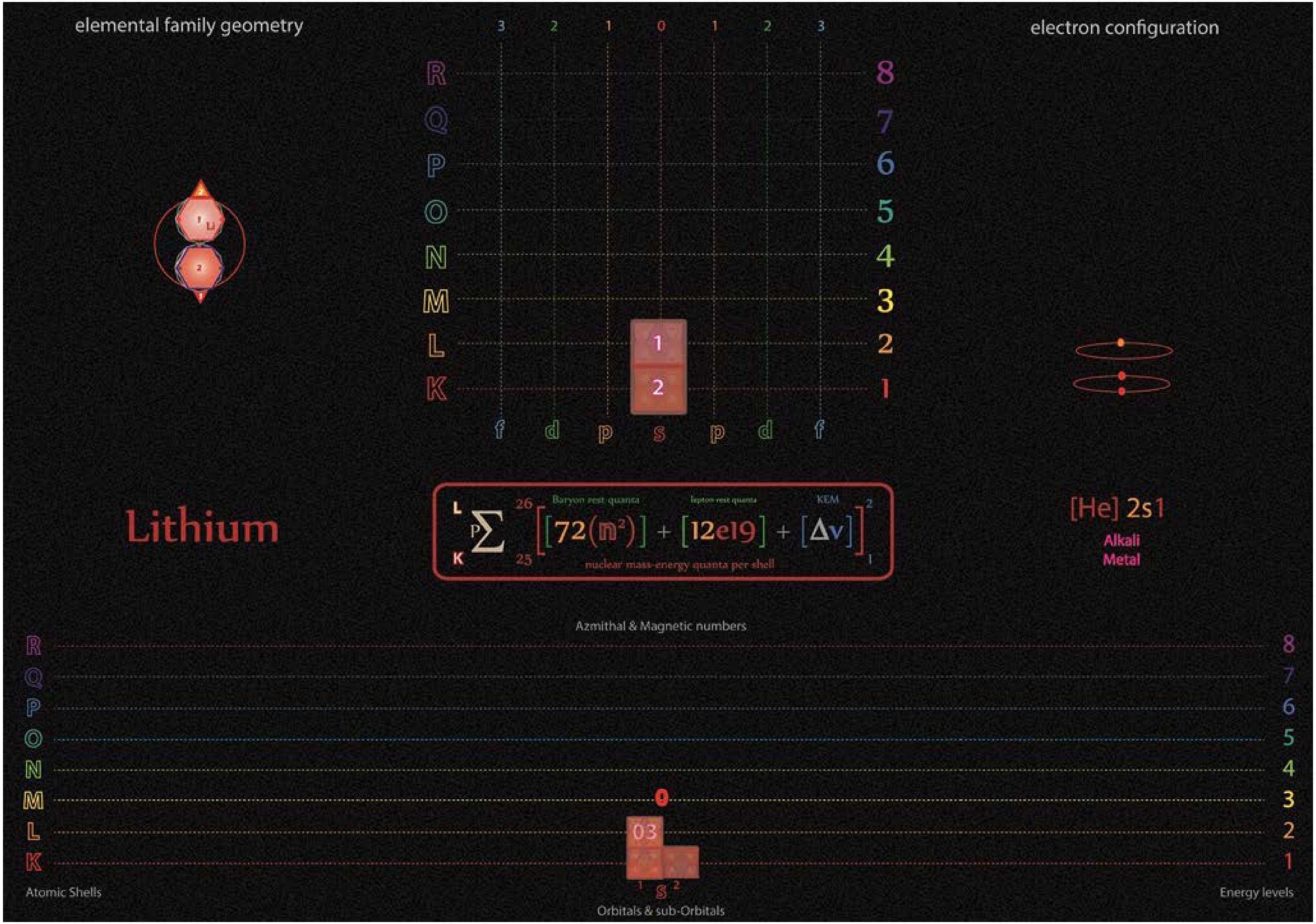
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
1
2
f d p s p d f
Lithium
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{L} P \, {}_{25}^{26}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_1^2$
nuclear mass-energy quanta per shell
[He] 2s1
Alkali
Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
03
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.03 - Lithium atomic config

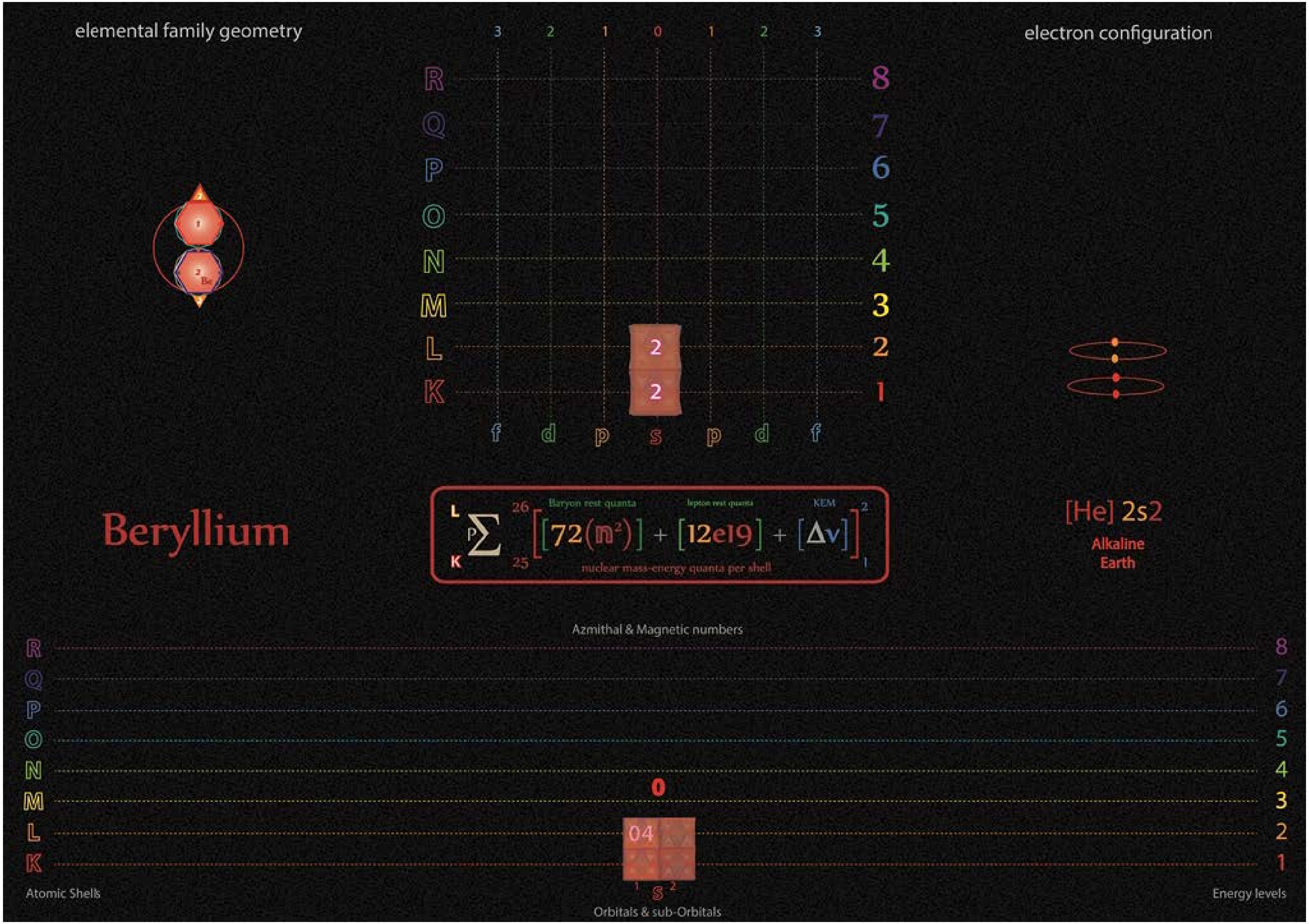
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
2
f d p s p d f
Beryllium
L K P Σ 26 25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[He] 2s2
Alkaline Earth
Azmithal & Magnetic numbers
0
04
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.04 - Beryllium atomic config

elemental family geometry

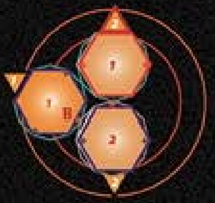

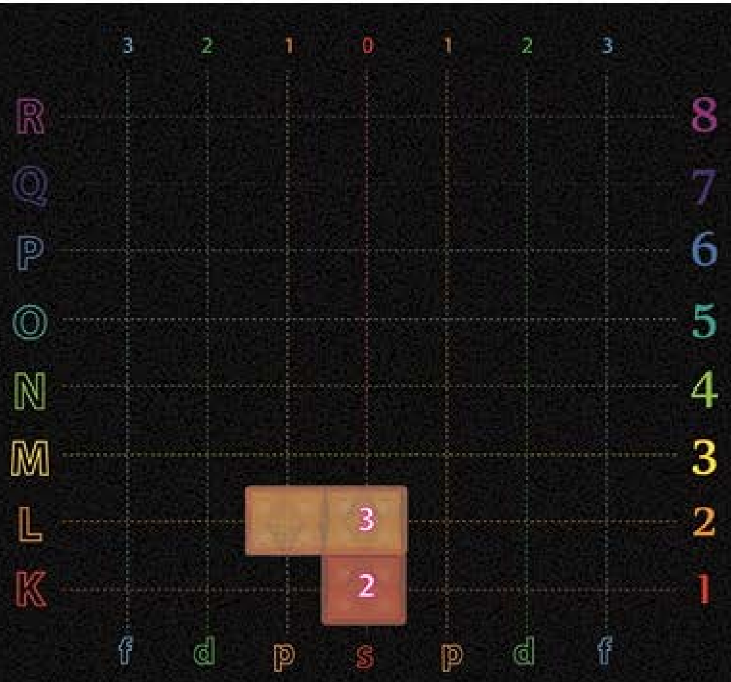
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
f
d
p
s
p
d
f

electron configuration

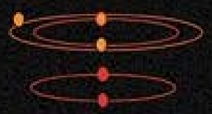

Boron

$$\sum_{K}^{L} P \; {}_{25}^{26}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{2}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[He] 2s2 2p1

Metaloid

Azmithal & Magnetic numbers

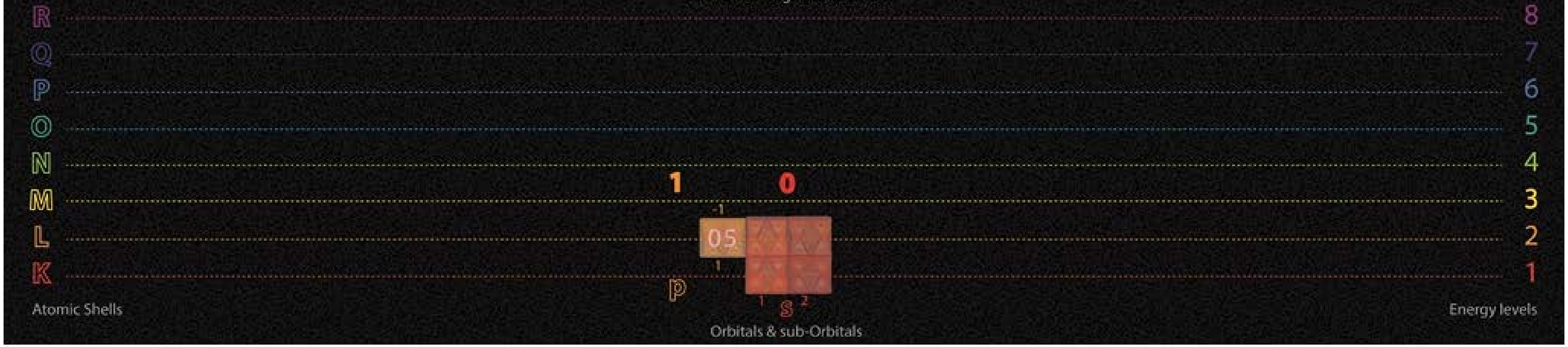
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
-1
05
1
p
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.05 - Boron atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

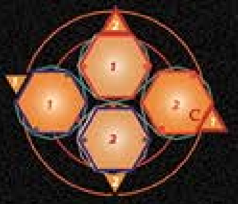

Carbon

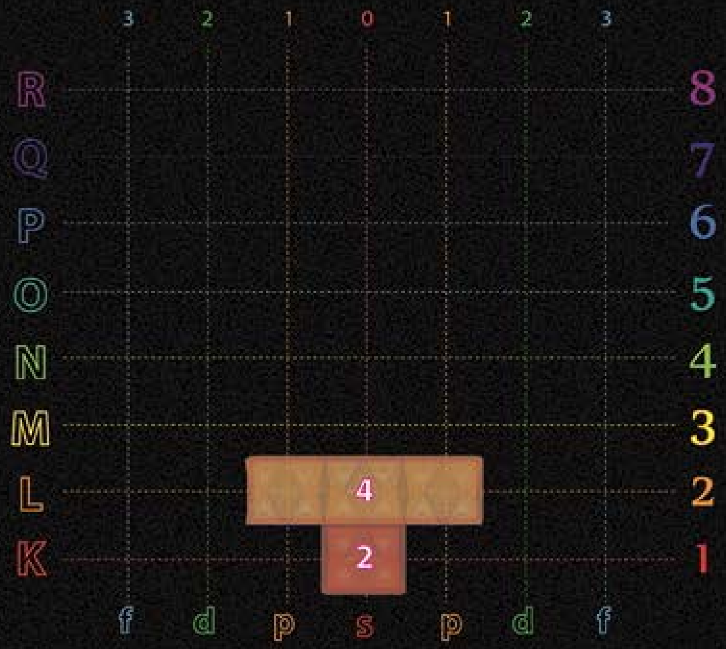
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
2
f
d
p
s
p
d
f

$$\sum_{K}^{L} P\sum_{25}^{26}\left[[72(n^2)] + [12e19] + [\Delta v]\right]_1^2$$

Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell

electron configuration

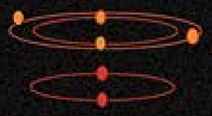

[He] 2s2 2p2

Metaloid

Azmithal & Magnetic numbers

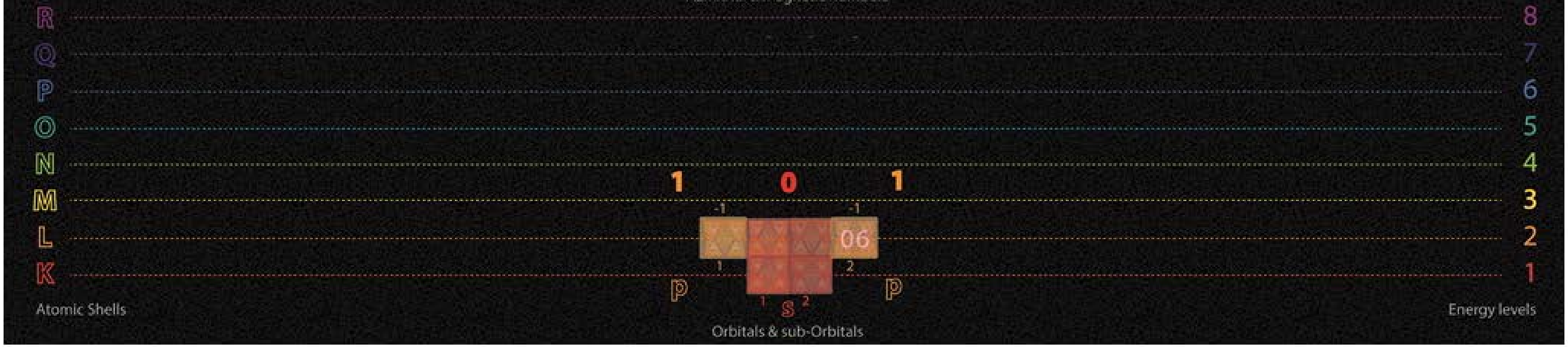
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
-1
-1
06
1
2
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.06 - Carbon atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

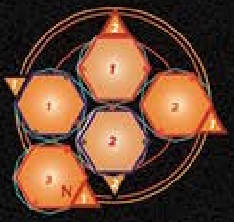

Nitrogen

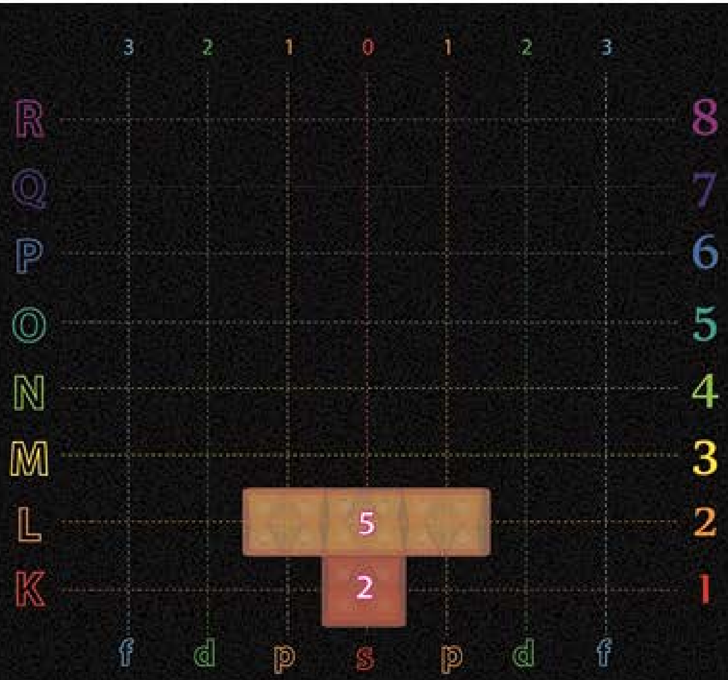
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
5
2
f
d
p
s
p
d
f

electron configuration

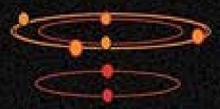

[He] 2s2 2p3

Metaloid

$$\sum_{K}^{L} P \left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{25\ \ 1}^{26\ \ 2}$$

nuclear mass-energy quanta per shell

Azmithal & Magnetic numbers

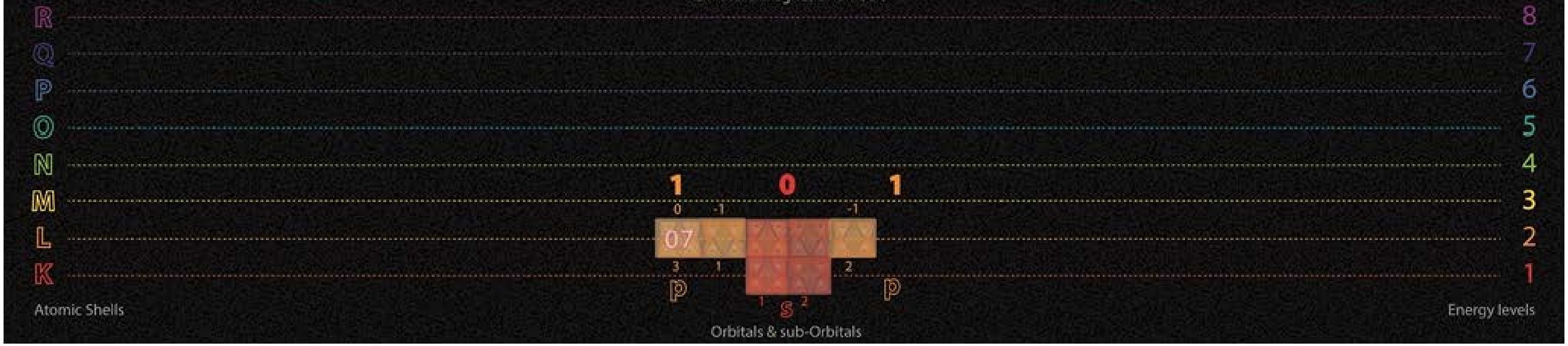
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
0
-1
-1
07
3
1
2
p
1
s
2
p

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.07 - Nitrogen atomic config

Copyright ABRAHAM [2008] - All rights reserved

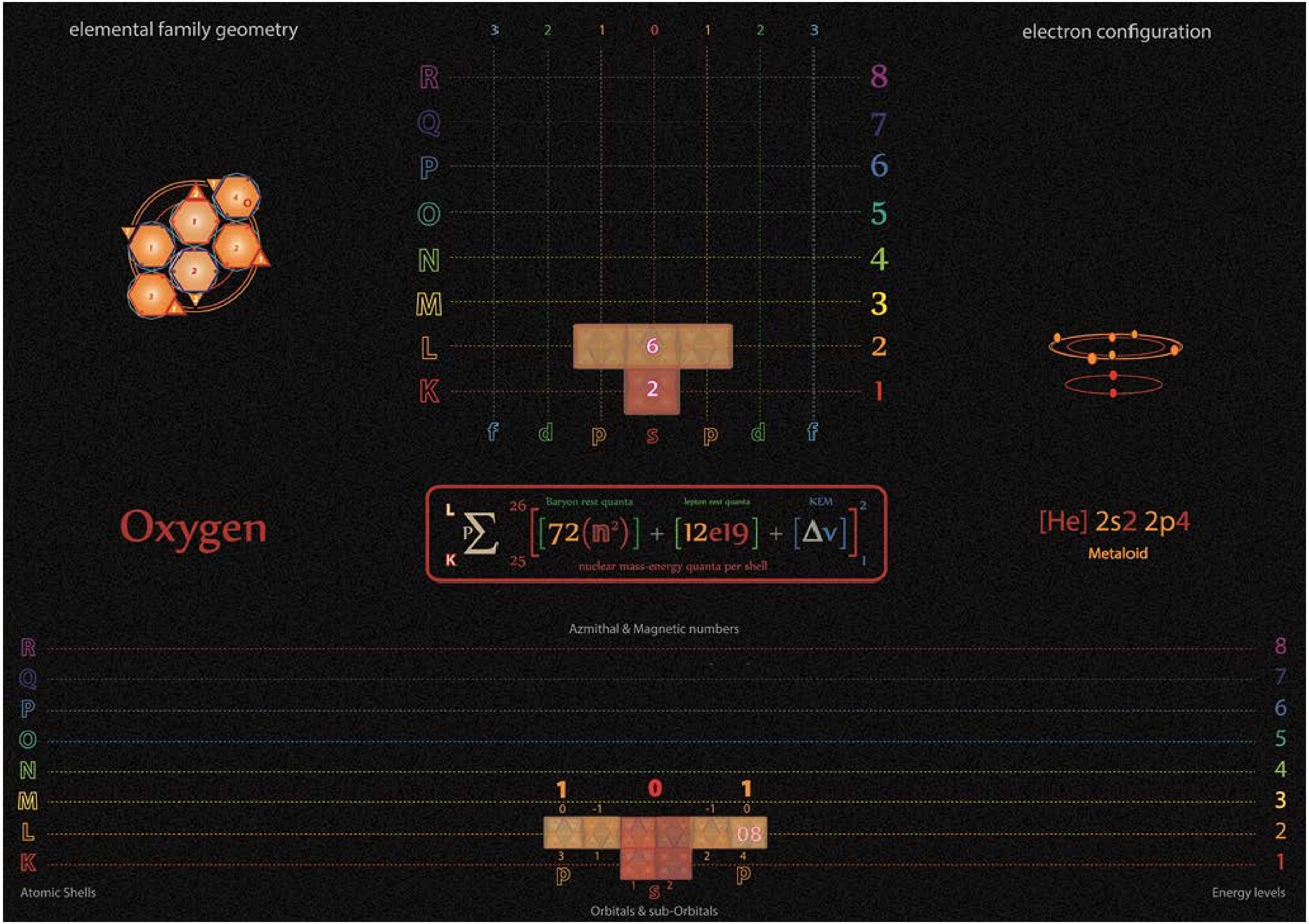

elemental family geometry

electron configuration

Oxygen

Baryon rest quanta · lepton rest quanta · KEM

$${}^{L}_{K}\,{}_{P}\!\sum_{25}^{26}\left[[72(n^{2})]+[12e19]+[\Delta v]\right]_{1}^{2}$$

nuclear mass-energy quanta per shell

[He] 2s2 2p4

Metaloid

Azmithal & Magnetic numbers

3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
2
f d p s p d f
1 0 1
0 -1 -1 0
08
3 1 2 4
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.08 - Oxygen atomic config

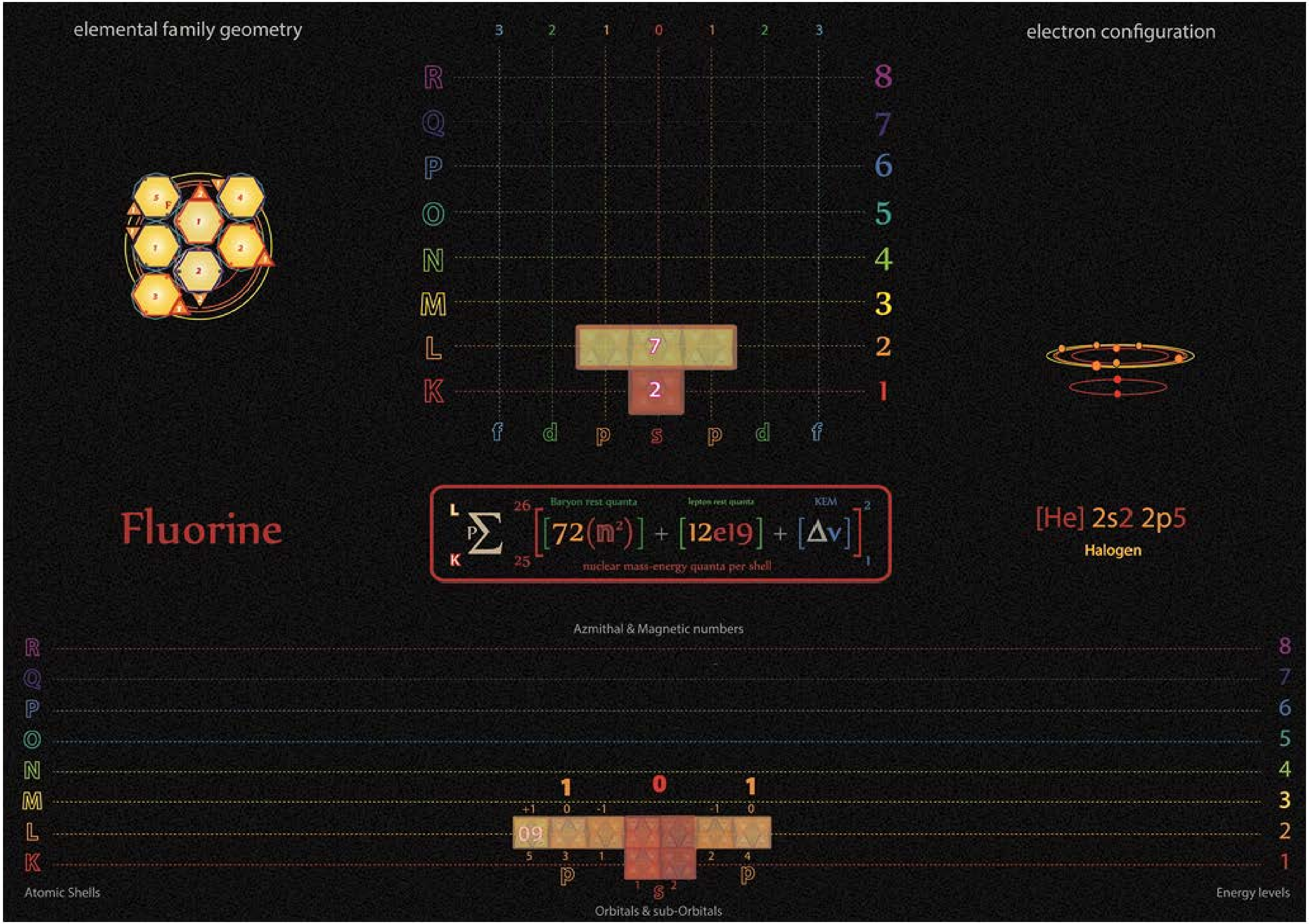
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
7
2
f d p s p d f
electron configuration
5 F 4 1 1 2 2 3
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0
09
5 3 1 2 4
p s p
1 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Fluorine

$$\sum_{K}^{L} {}_{P} \sum_{25}^{26} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_1^2$$

Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell

[He] 2s2 2p5

Halogen

Tetryonics 53.09 - Fluorine atomic config

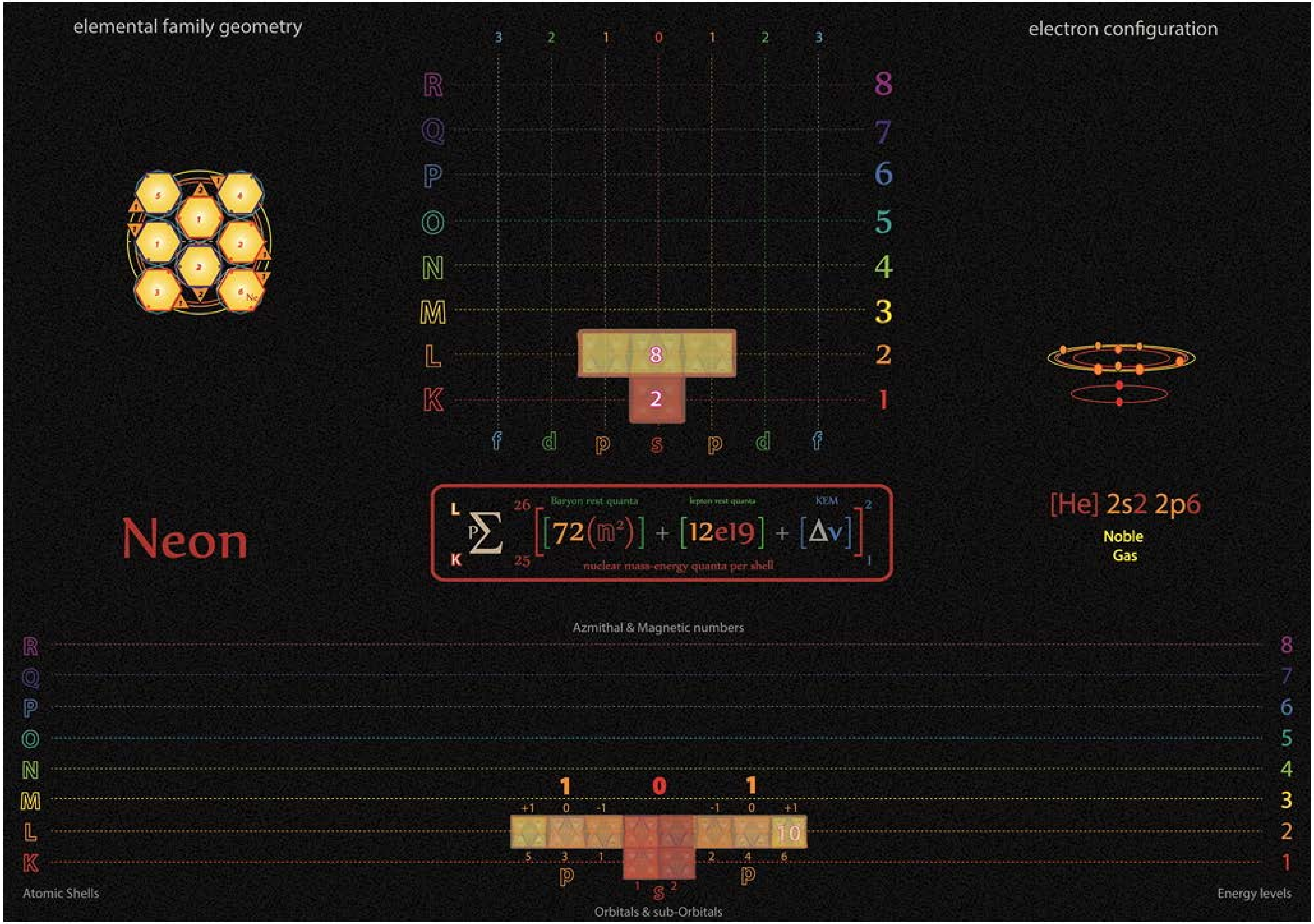
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
8
2
f d p s p d f
Neon
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[He] 2s2 2p6
Noble Gas
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
10
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{L} P \;{}^{26}_{25}\Big[\,[72(n^2)] + [12e19] + [\Delta v]\,\Big]^{2}_{1}$$

Tetryonics 53.10 - Neon atomic config

Copyright ABRAHAM [2008] - All rights reserved

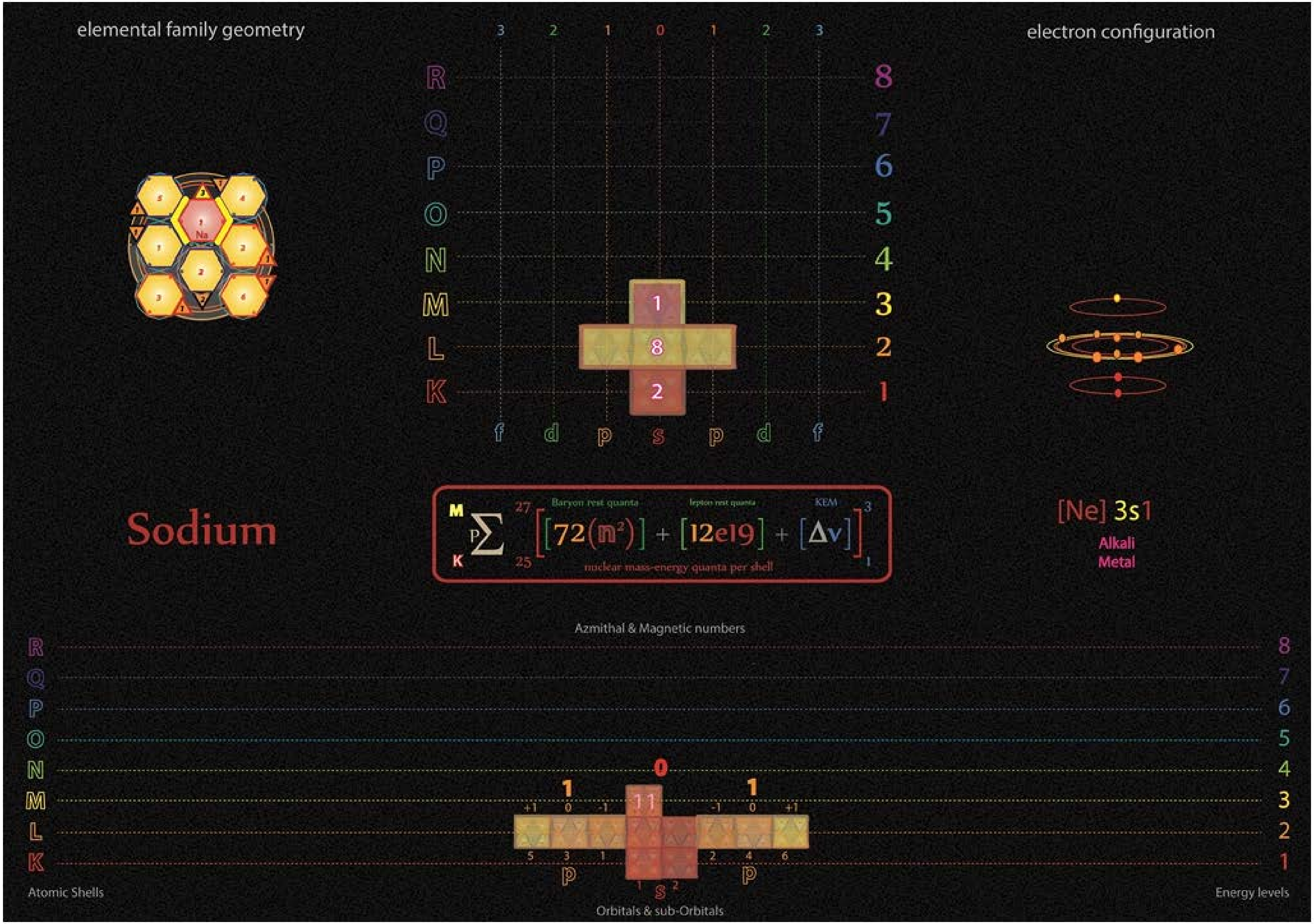
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
2
f d p s p d f
Na
Sodium
M
K
P
27
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
3
1
nuclear mass-energy quanta per shell
[Ne] 3s1
Alkali
Metal
Azmithal & Magnetic numbers
0
1
1
+1 0 -1
-1 0 +1
1 1
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.11 - Sodium atomic config

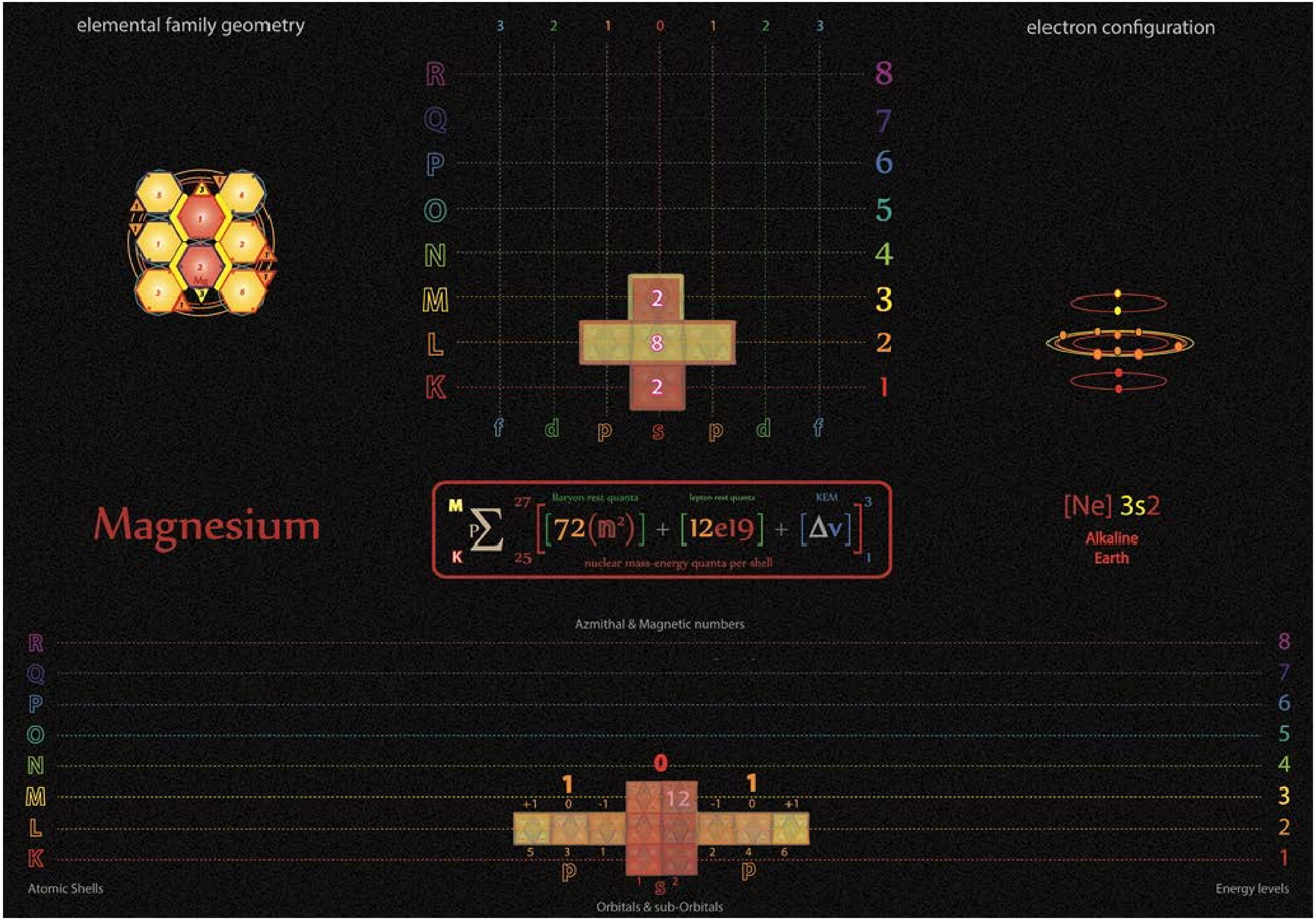
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
2
f d p s p d f
Magnesium
[Ne] 3s2
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
12
+1 0 -1
-1 0 +1
5 3 1
2 4 6
p
s
p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{M} {}_{p}\sum_{25}^{27} \left[\underbrace{\overset{\text{Baryon rest quanta}}{\left[72(m^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]}}_{\text{nuclear mass-energy quanta per shell}} \right]_1^3$$

Tetryonics 53.12 - Magnesium atomic config

Copyright ABRAHAM [2008] - All rights reserved

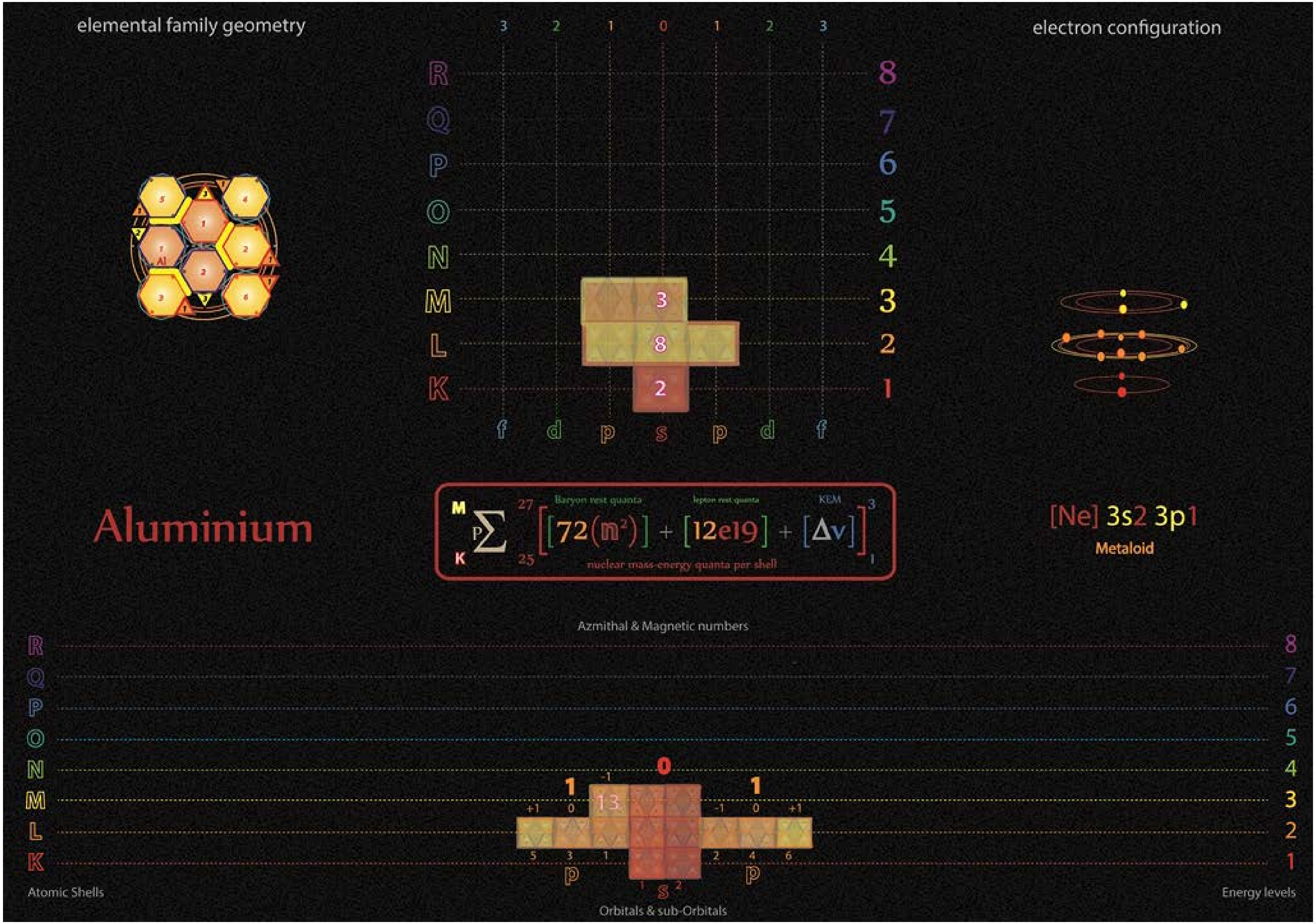
elemental family geometry
Al
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
-1
+1 0
13
-1 0 +1
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Aluminium

Baryon rest quanta — lepton rest quanta — KEM

$$ {}^{M}_{K}P\sum_{25}^{27}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{3} $$

nuclear mass-energy quanta per shell

[Ne] 3s2 3p1

Metaloid

Tetryonics 53.13 - Aluminium atomic config

Copyright ABRAHAM [2008] - All rights reserved

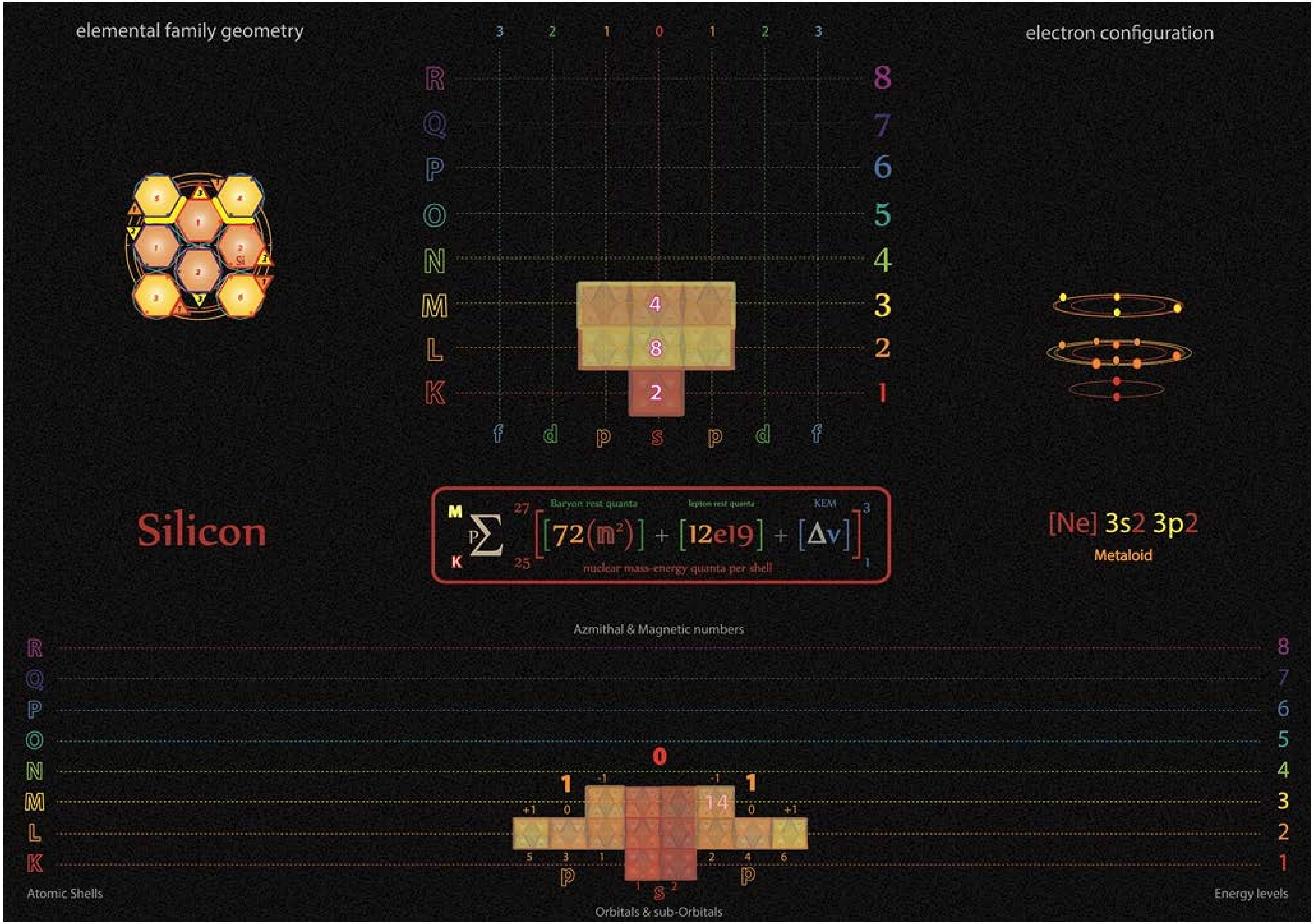
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
4
8
2
f d p s p d f
Si
Silicon
M
K
P
27
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
3
1
nuclear mass-energy quanta per shell
[Ne] 3s2 3p2
Metaloid
Azmithal & Magnetic numbers
0
1 1
-1 -1
+1 0 0 +1
14
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.14 - Silicon atomic config

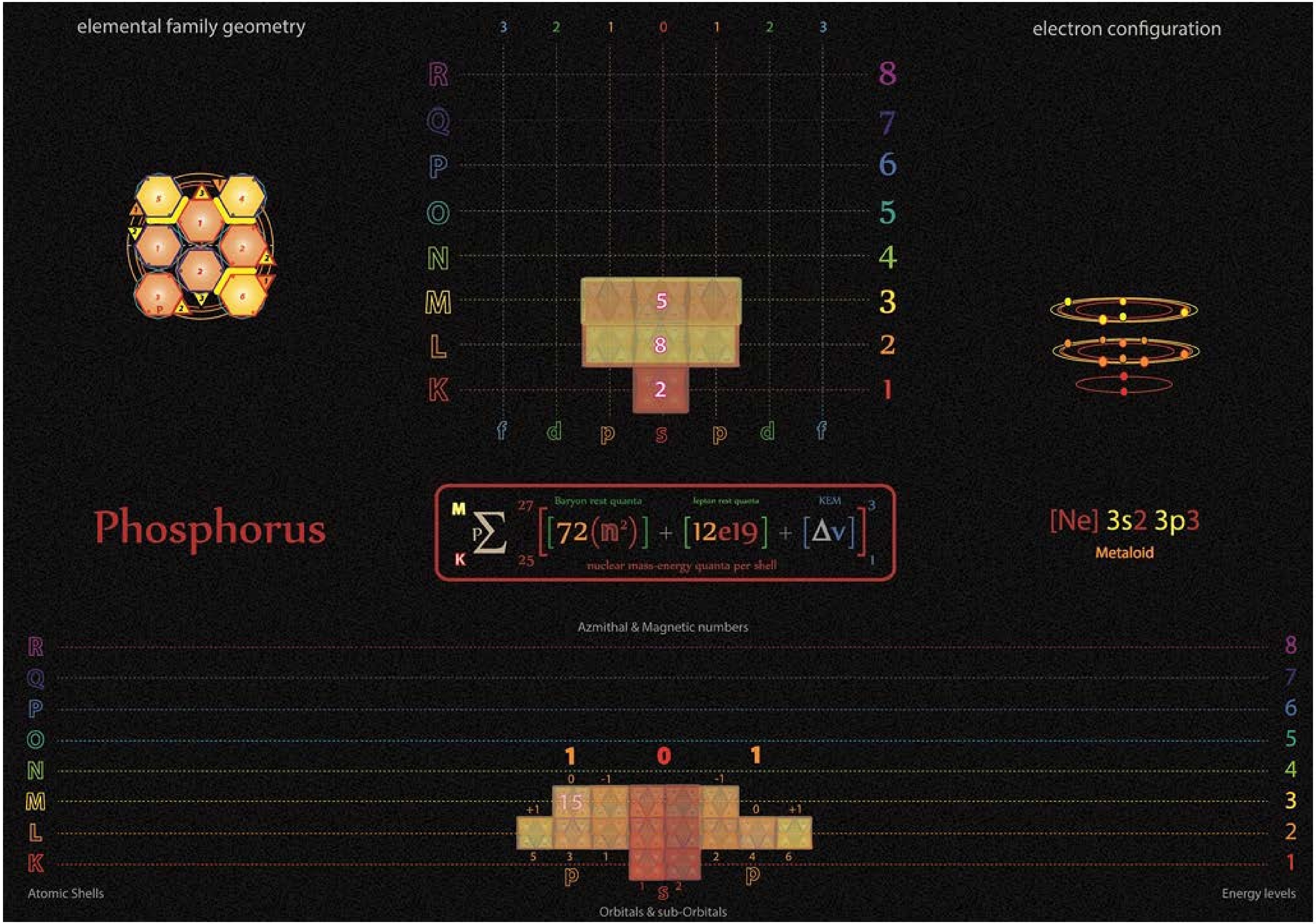

elemental family geometry

electron configuration

Phosphorus

$$\sum_{K}^{M}{}_{P}\;{}^{27}_{25}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]^{3}_{1}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Ne] 3s2 3p3

Metaloid

3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
5
8
2
f
d
p
s
p
d
f
Azmithal & Magnetic numbers
1
0
1
0
-1
-1
+1
15
0
+1
5
3
1
2
4
6
p
s
1
2
p
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.15 - Phosphorus atomic config

Copyright ABRAHAM [2008] - All rights reserved

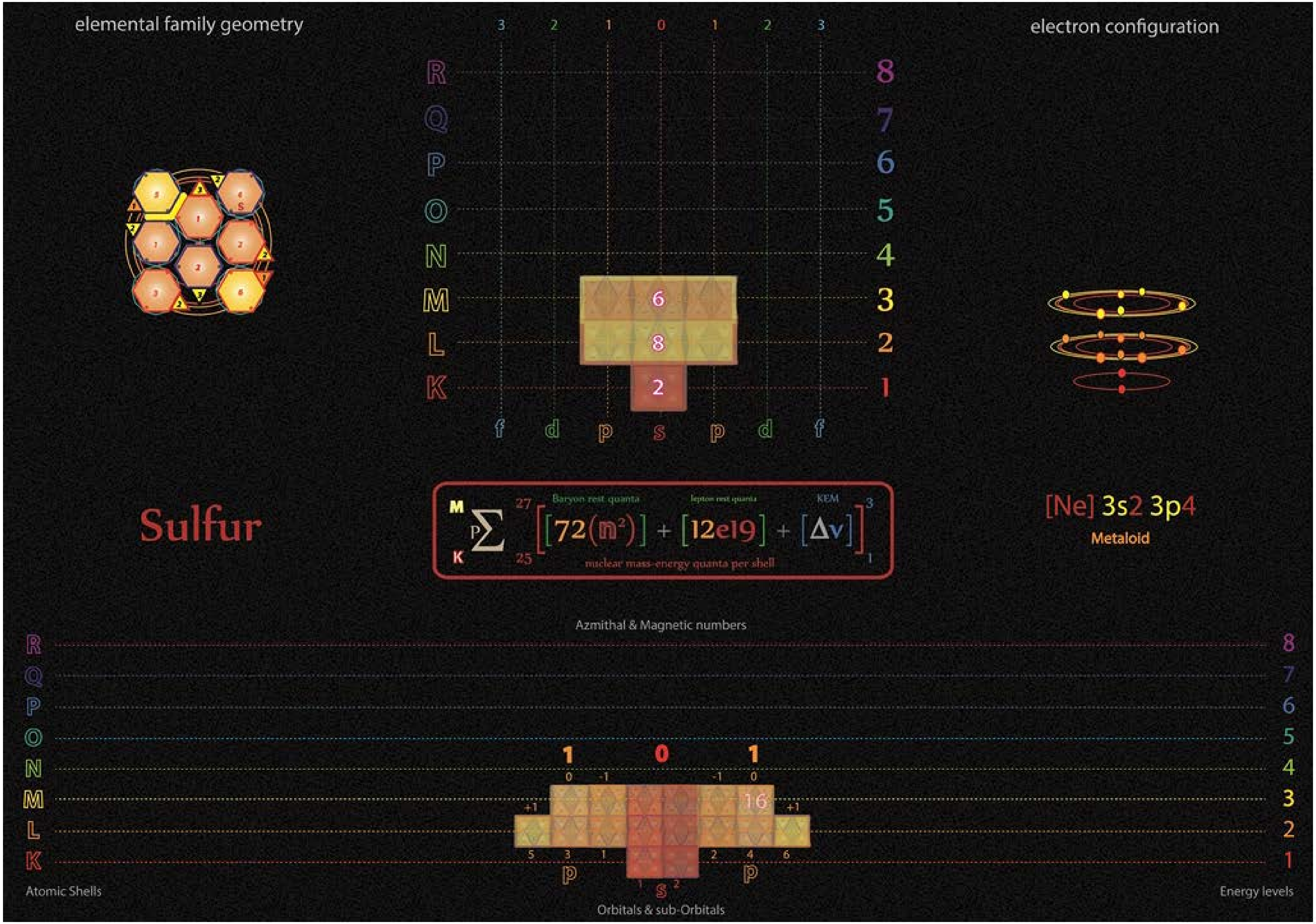
elemental family geometry
electron configuration
Sulfur
[Ne] 3s2 3p4
Metaloid
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
8
2
f d p s p d f
M
K
P∑
27
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(m²)] + [12e19] + [Δv]]
3
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
+1 +1
16
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.16 - Sulfur atomic config

elemental family geometry

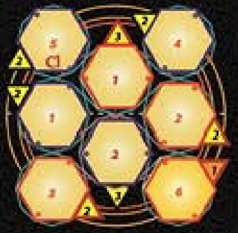

Chlorine

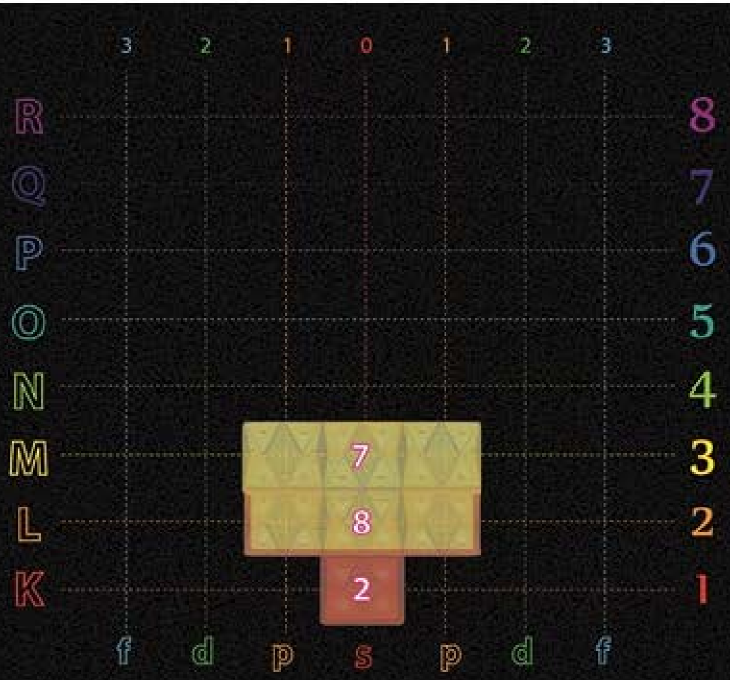
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
8
2
f
d
p
s
p
d
f

electron configuration

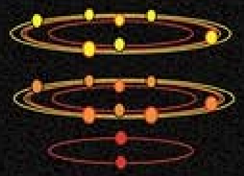

[Ne] 3s2 3p5

Halogen

$$\sum_{K}^{M} {}_{p}\left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}^{3}$$

27

25

nuclear mass-energy quanta per shell

Azmithal & Magnetic numbers

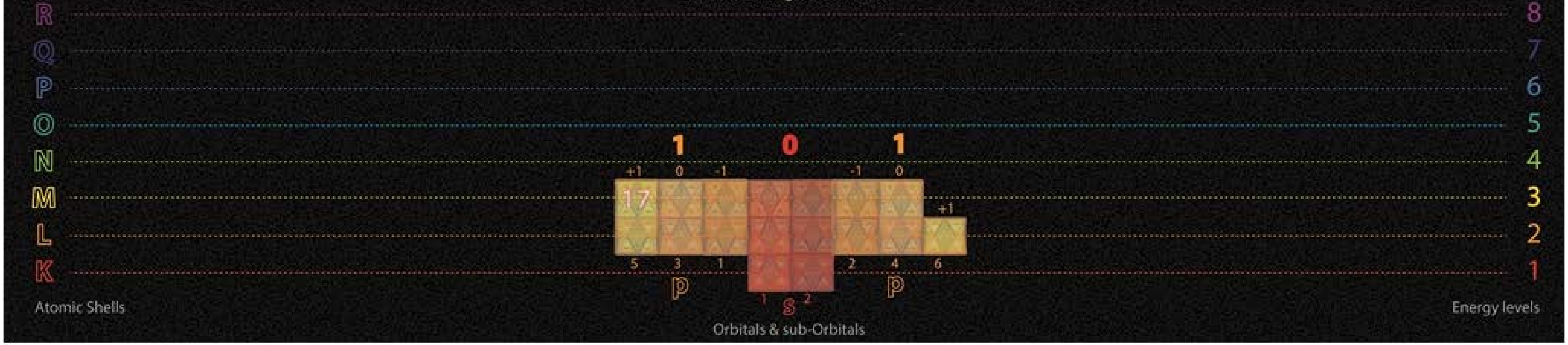
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
+1
0
-1
-1
0
+1
17
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.17 - Chlorine atomic config

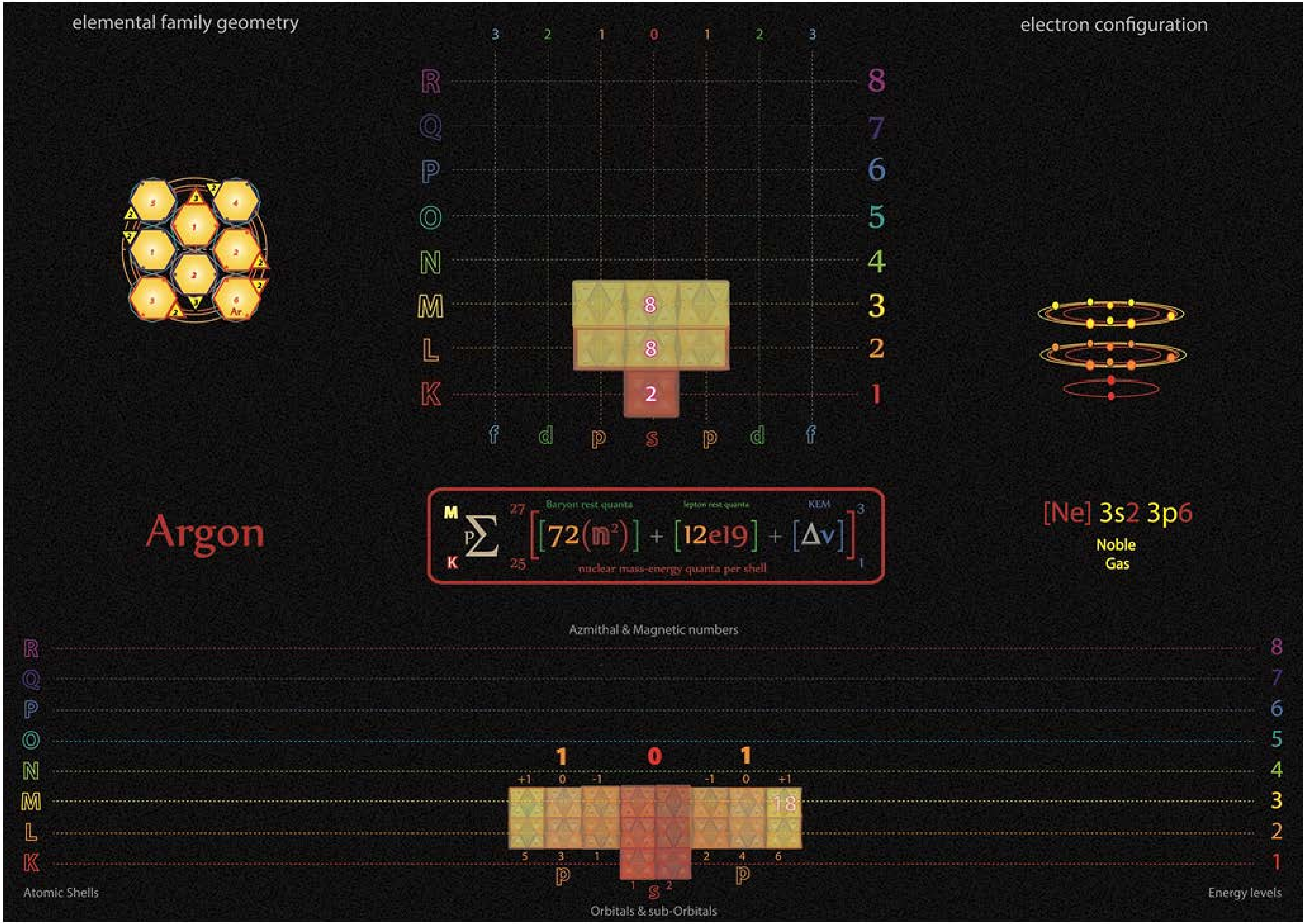

elemental family geometry

Argon

electron configuration

[Ne] 3s2 3p6

Noble Gas

$$ {}^{\mathrm{M}}_{\mathrm{K}}{}_{\mathrm{P}}\sum_{25}^{27}\left[\underset{\text{Baryon rest quanta}}{[72(\mathrm{m}^2)]}+\underset{\text{lepton rest quanta}}{[12e19]}+\underset{\text{KEM}}{[\Delta v]}\right]_1^3 $$

nuclear mass-energy quanta per shell

3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
8
8
2
f d p s p d f
Ar
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
18
5 3 1 2 4 6
p s p
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.18 - Argon atomic config

Copyright ABRAHAM [2008] - All rights reserved

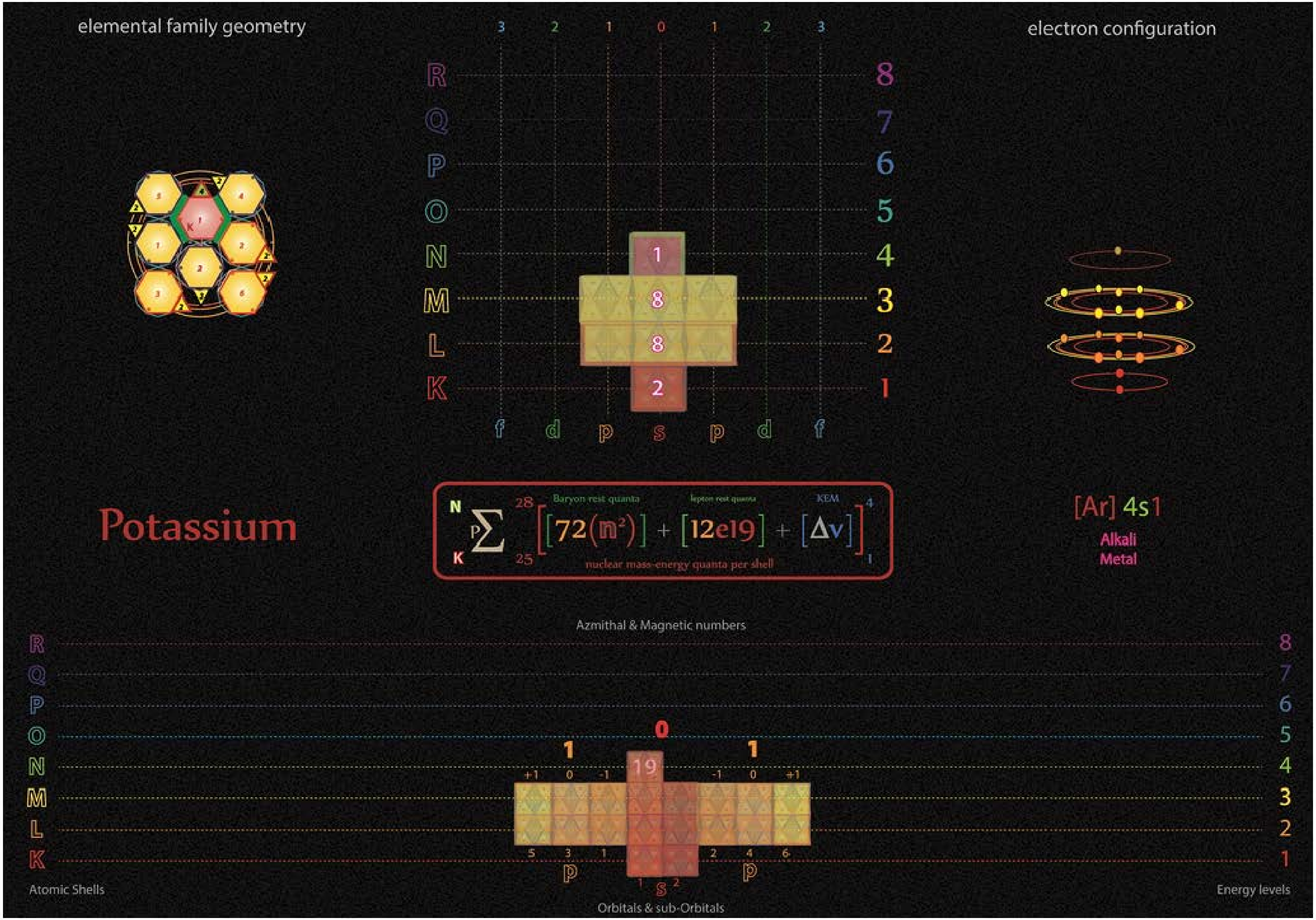
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
8
2
f d p s p d f
electron configuration
Potassium
$\sum_{K}^{N} {}_{P} \sum_{25}^{28} \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{4}$
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Ar] 4s1
Alkali
Metal
Azmithal & Magnetic numbers
0
1
1
19
+1 0 -1
-1 0 +1
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.19 - Potassium atomic config

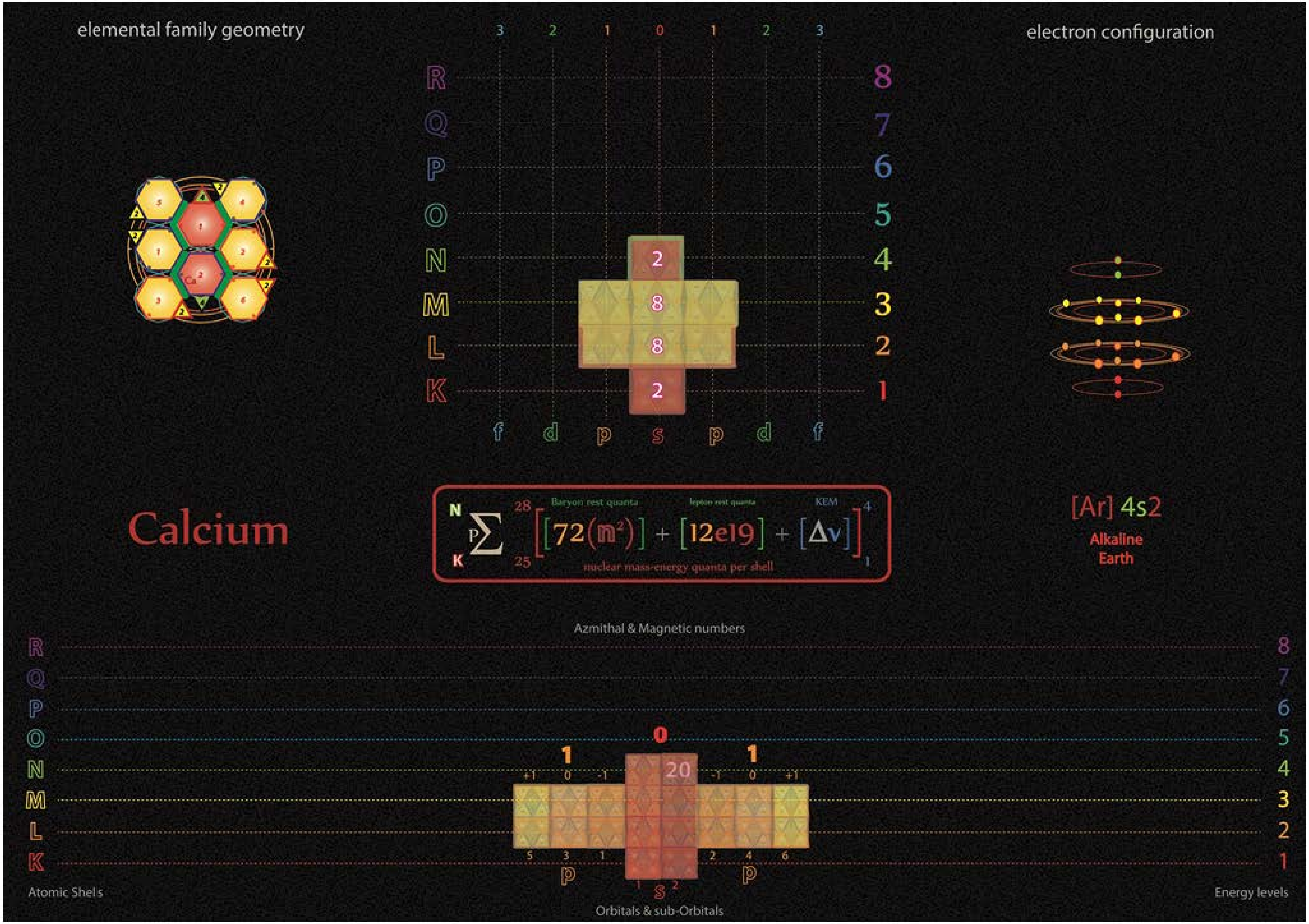
elemental family geometry
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
8
2
f d p s p d f
electron configuration
Calcium
N
K
28
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{N} p \sum_{25}^{28} \left[[72(n^2)] + [12e19] + [\Delta v] \right]_1^4$
nuclear mass-energy quanta per shell
[Ar] 4s2
Alkaline
Earth
Azmithal & Magnetic numbers
0
1
1
20
+1 0 -1
-1 0 +1
5 3 1
2 4 6
p
p
1 s 2
Atomic Shels
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.20 - Calcium atomic config

Copyright ABRAHAM [2008] - All rights reserved

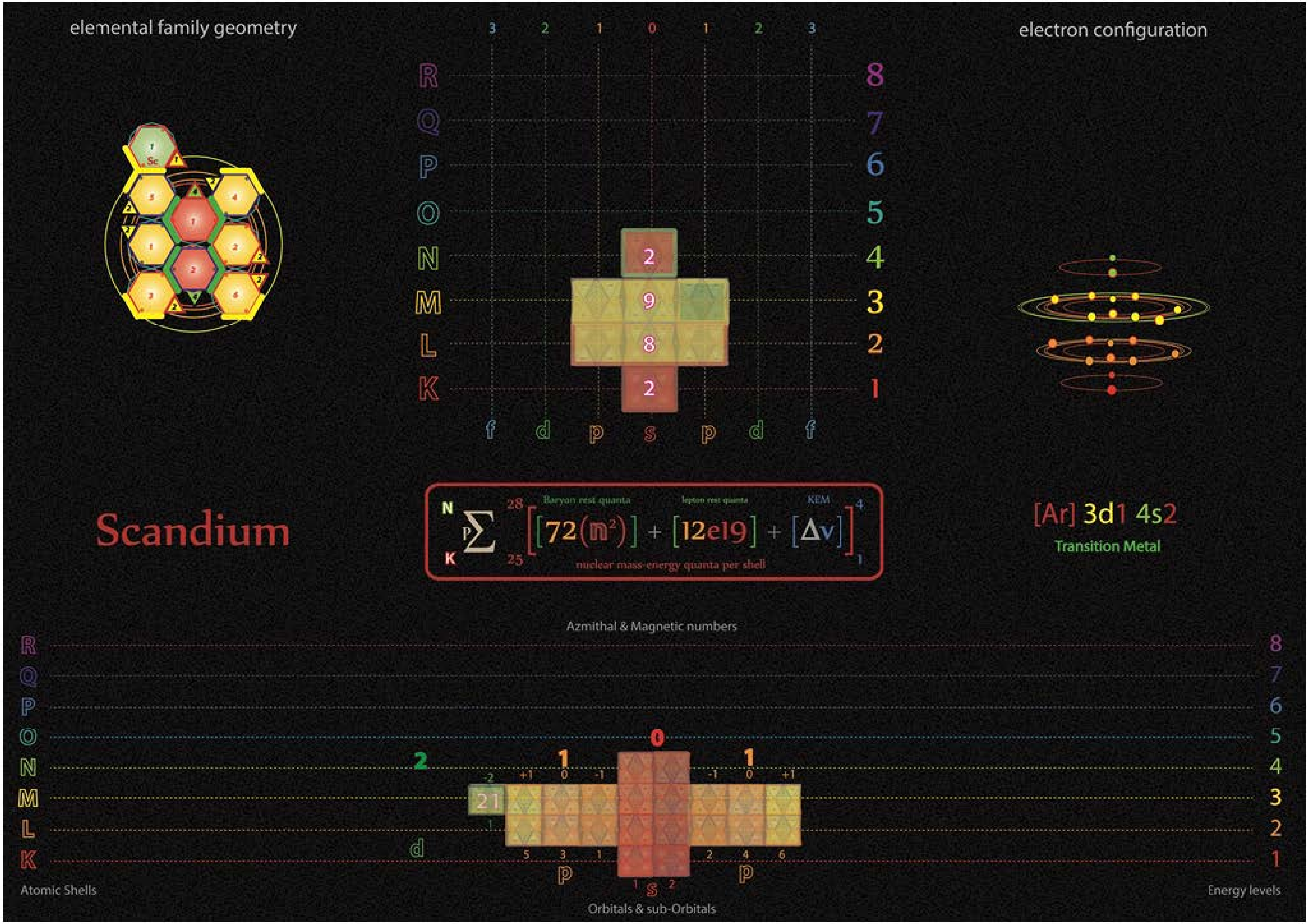
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
9
8
2
f
d
p
s
p
d
f
1
Sc
2
-2
+1
0
-1
-1
0
+1
21
1
d
5
3
1
2
4
6
p
p
1
s
2

elemental family geometry

electron configuration

Scandium

$$\sum_{K}^{N} {}_{P}\sum_{25}^{28}\left[\,[72(\mathfrak{m}^2)] + [12e19] + [\Delta v]\,\right]_1^4$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Ar] 3d1 4s2

Transition Metal

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.21 - Scandium atomic config

elemental family geometry

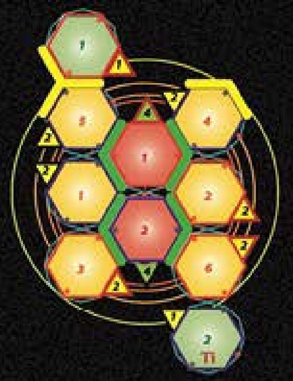

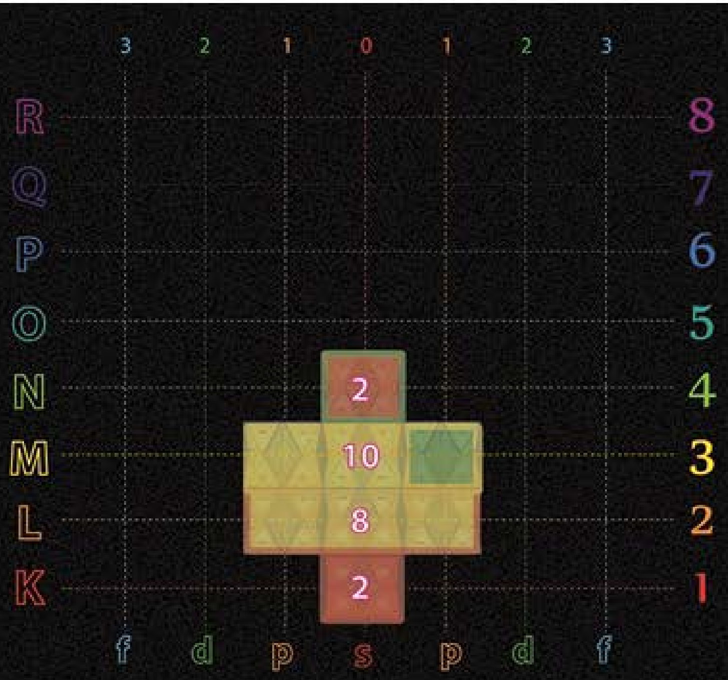
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
10
8
2
f
d
p
s
p
d
f

electron configuration

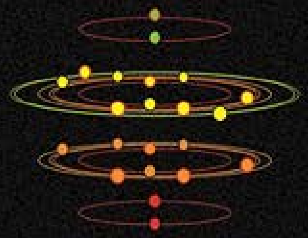

Titanium

$$\sum_{K}^{N} {}^{P}\,{}_{25}^{28}\Big[[72(m^2)] + [12e19] + [\Delta v]\Big]_{1}^{4}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Ar] 3d2 4s2

Transition Metal

Azmithal & Magnetic numbers

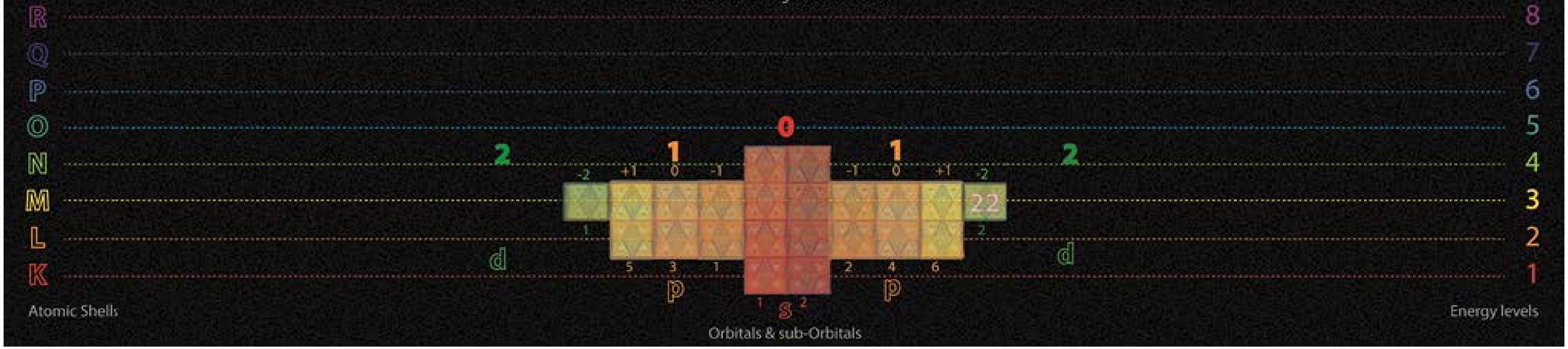
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
-2
+1
0
-1
-1
0
+1
-2
22
1
2
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.22 - Titanium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

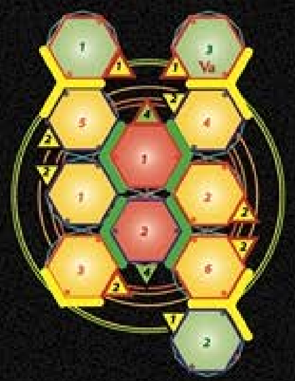
1
3
Va
5
4
1
1
2
2
3
6
2

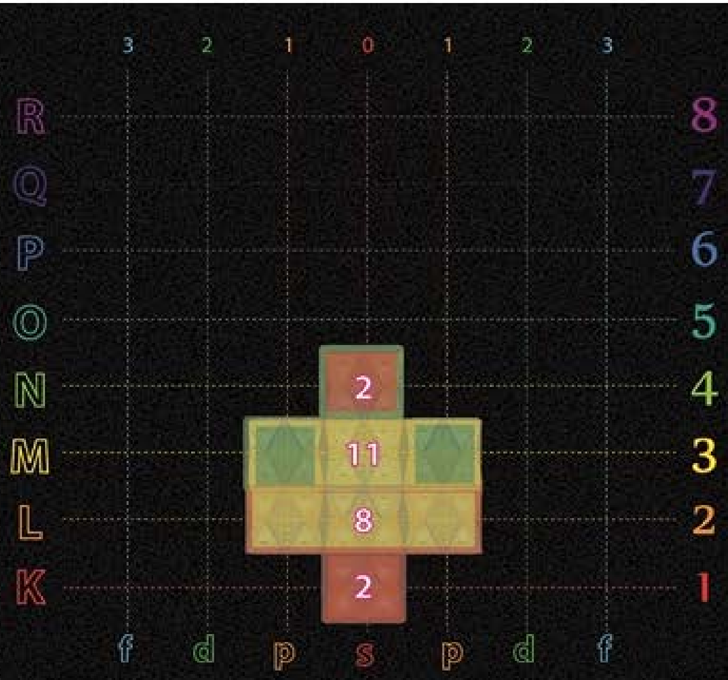
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
11
8
2
f
d
p
s
p
d
f

electron configuration

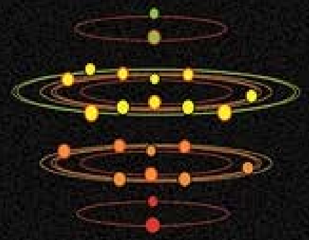

Vanadium

$$ {}^{N}_{K}\;{}_{P}\sum^{28}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(n^{2})]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]^{4}_{1} $$

nuclear mass-energy quanta per shell

[Ar] 3d3 4s2

Transition Metal

Azmithal & Magnetic numbers

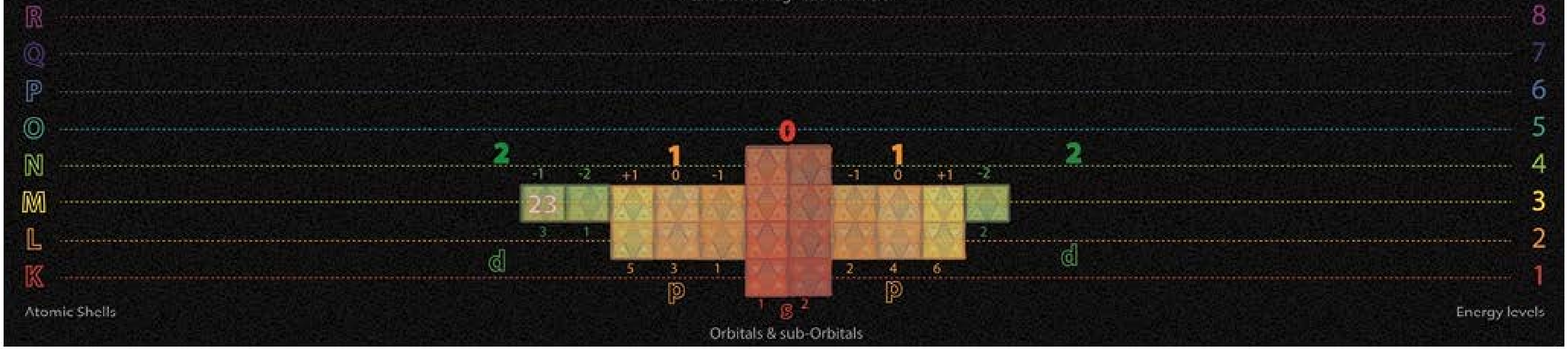
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
-1
-2
+1
0
-1
-1
0
+1
-2
23
3
1
2
5
3
1
2
4
6
d
p
s
1
2
p
d
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.23 - Vanadium atomic config

elemental family geometry

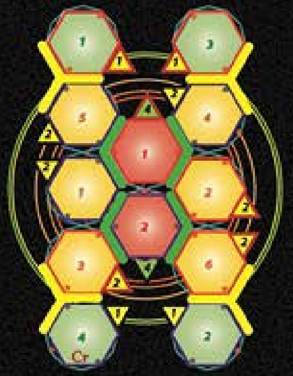

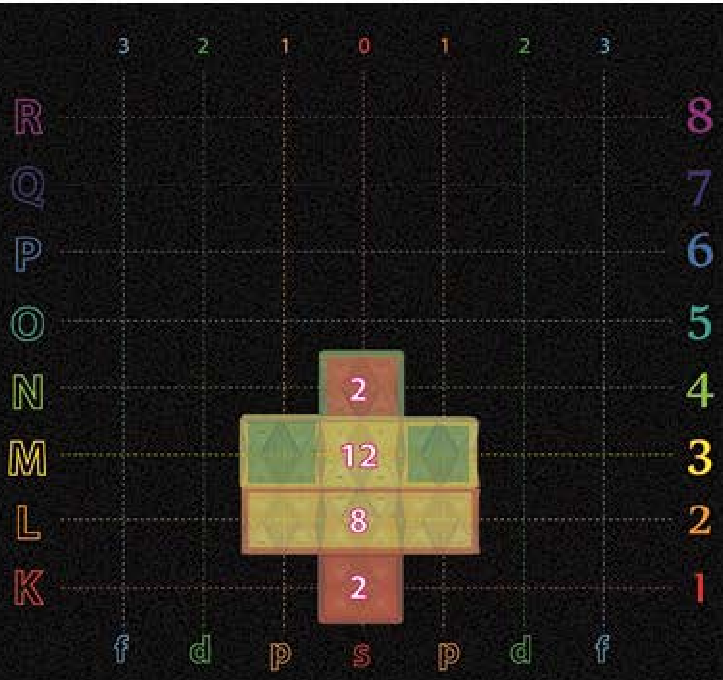
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
12
8
2
f
d
p
s
p
d
f

electron configuration

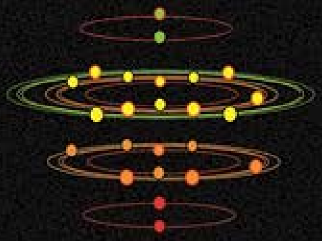

Chromium

$$\sum_{K}^{N} {}_{P} \, {}_{25}^{28}\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{4}$$

nuclear mass-energy quanta per shell

[Ar] 3d4 4s2

Transition Metal

Azmithal & Magnetic numbers

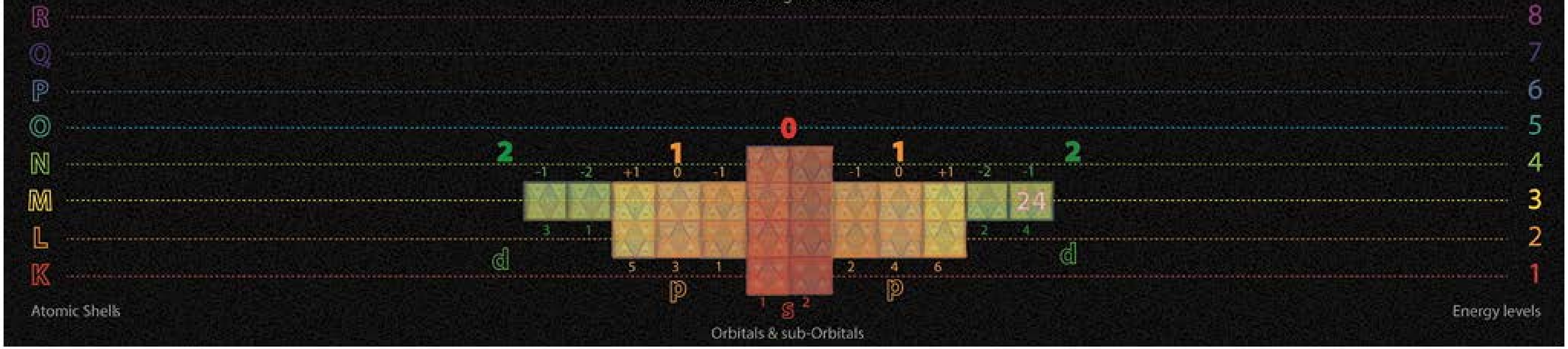
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
-1
-2
+1
0
-1
-1
0
+1
-2
-1
24
3
1
2
4
d
d
5
3
1
2
4
6
p
p
1
s
2

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.24 - Chromium atomic config

elemental family geometry

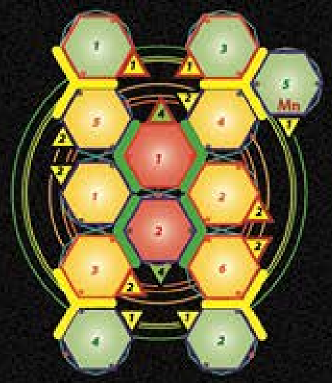

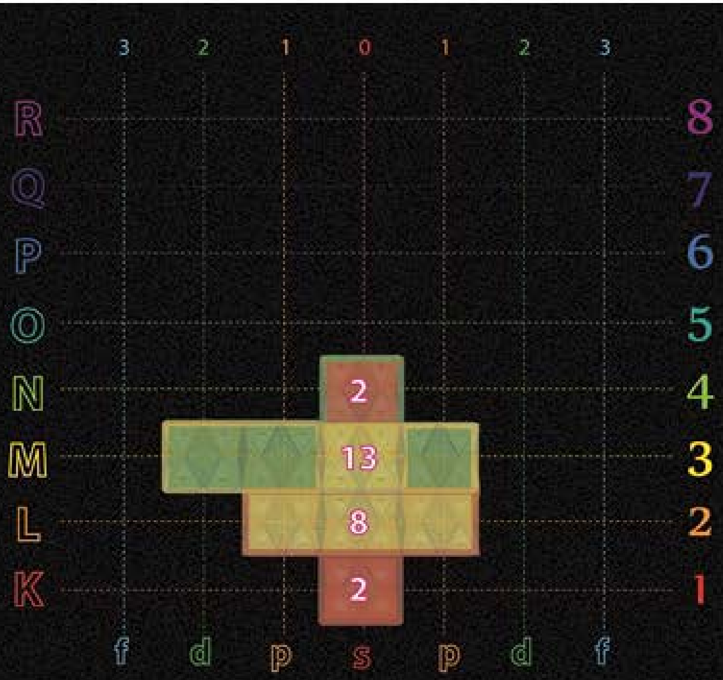
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
13
8
2
f d p s p d f

electron configuration

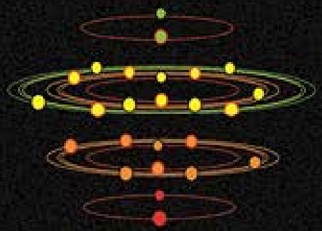

Manganese

Baryon rest quanta — lepton rest quanta — KEM

$$\overset{N}{\underset{K}{{}_{p}\sum}}\;{}^{28}_{25}\left[\,[72(m^{2})] + [12e19] + [\Delta v]\,\right]^{4}_{1}$$

nuclear mass-energy quanta per shell

[Ar] 3d5 4s2

Transition Metal

Azmithal & Magnetic numbers

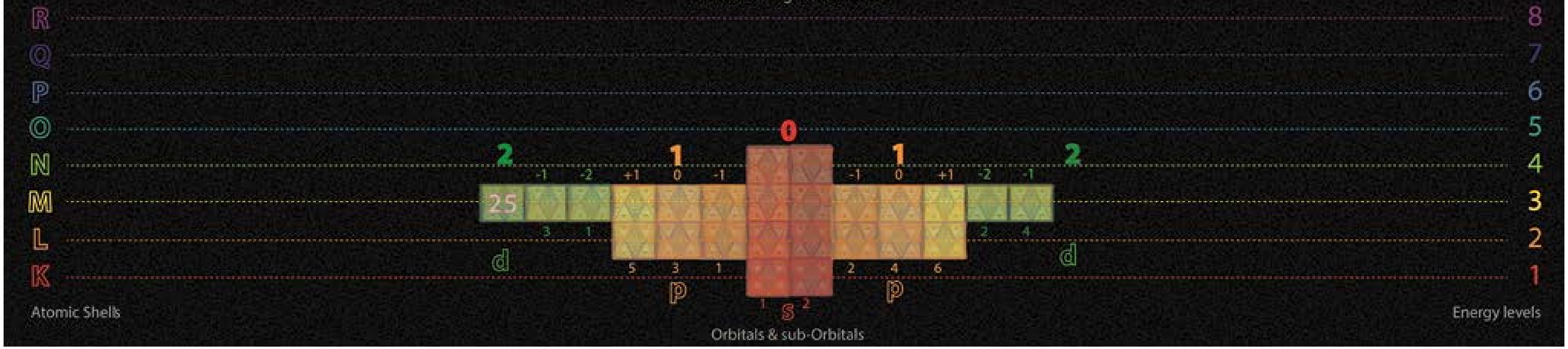
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
-1 -2 +1 0 -1 -1 0 +1 -2 -1
25
3 1 2 4
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.25 - Manganese atomic config

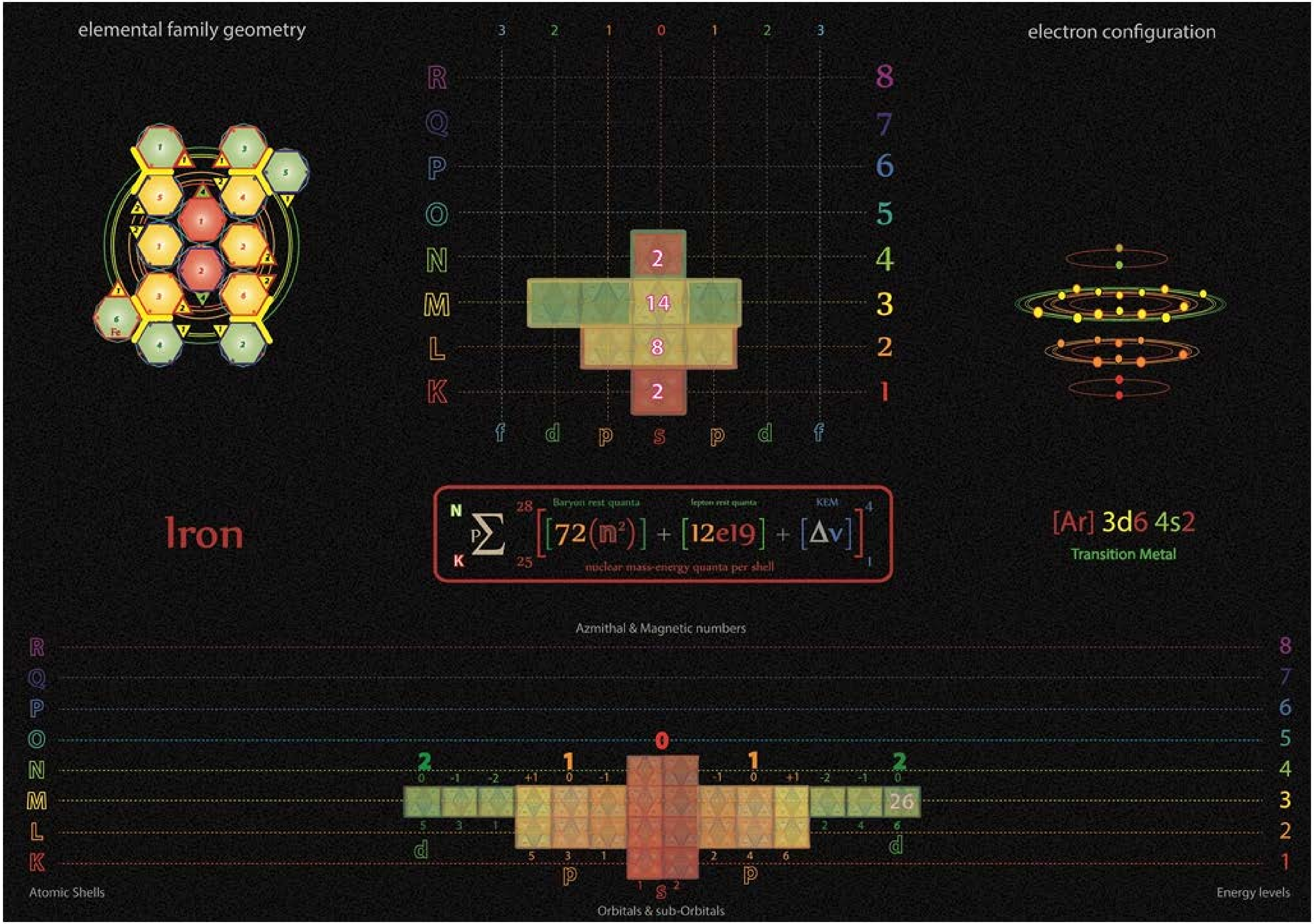

elemental family geometry

electron configuration

Iron

$$\sum_{K}^{N} {}_{P}\,{}_{25}^{28}\Big[[72(n^2)] + [12e19] + [\Delta v]\Big]_1^4$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Ar] 3d6 4s2

Transition Metal

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Fe
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
8
2
f d p s p d f
2 1 0 1 2
0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
26
5 3 1 2 4 6
d d
5 3 1 2 4 6
p p
1 s 2

Tetryonics 53.26 - Iron atomic config

elemental family geometry

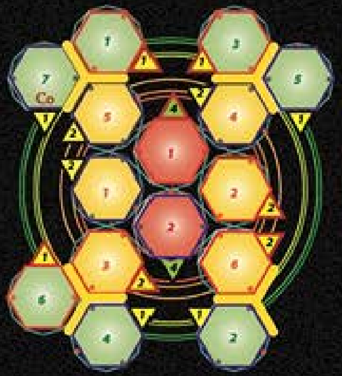
Co
1
3
7
5
5
4
1
1
2
2
6
3
8
6
4
2

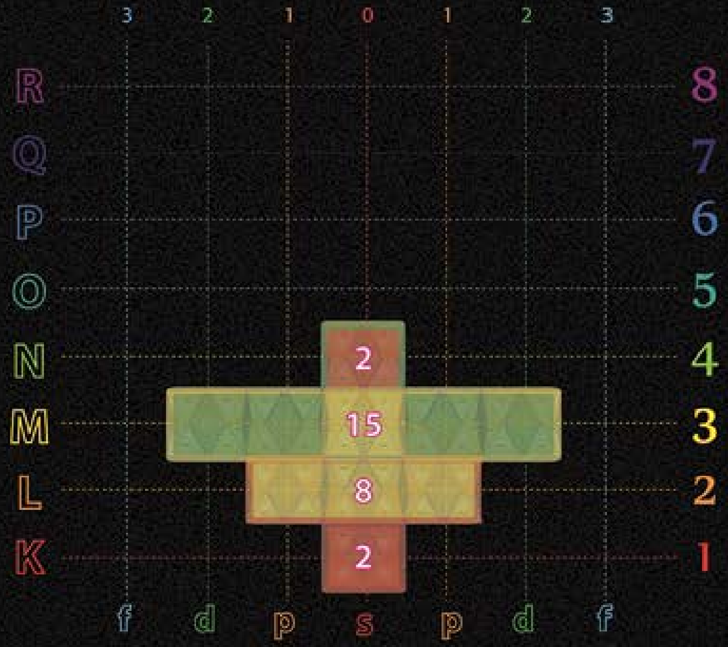
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
15
8
2
f
d
p
s
p
d
f

electron configuration

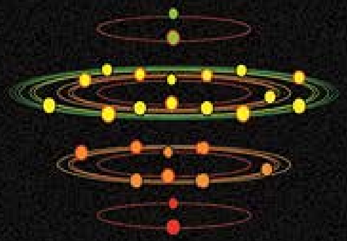

Cobalt

$$ {}^{N}_{K}\,{}_{P}\sum_{25}^{28}\left[\overset{\text{Baryon rest quanta}}{[72(\mathfrak{m}^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]}\right]_{1}^{4} $$

nuclear mass-energy quanta per shell

[Ar] 3d7 4s2

Transition Metal

Azmithal & Magnetic numbers

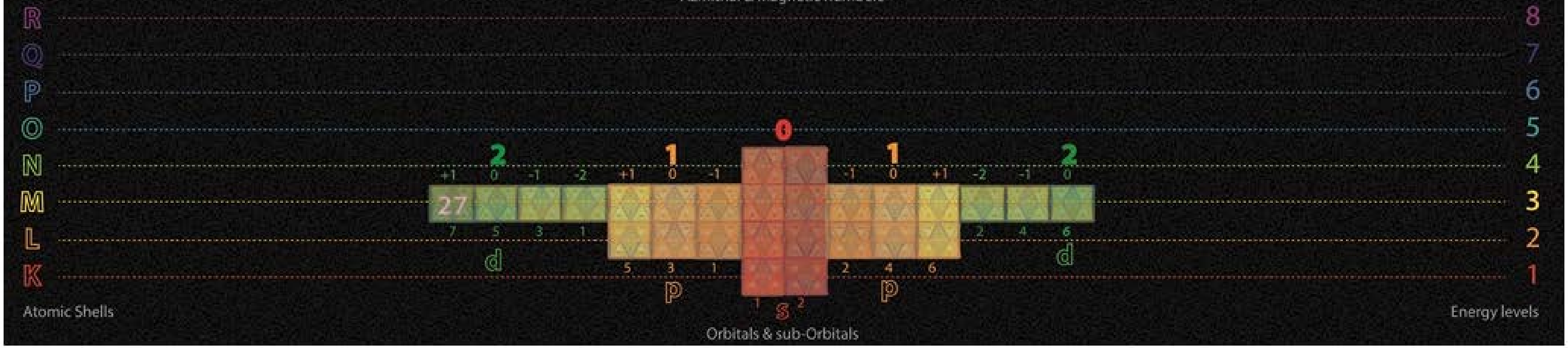
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
27
7
5
3
1
2
4
6
d
5
3
1
2
4
6
p
p
1
s
2
d

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.27 - Cobalt atomic config

Copyright ABRAHAM [2008] - All rights reserved

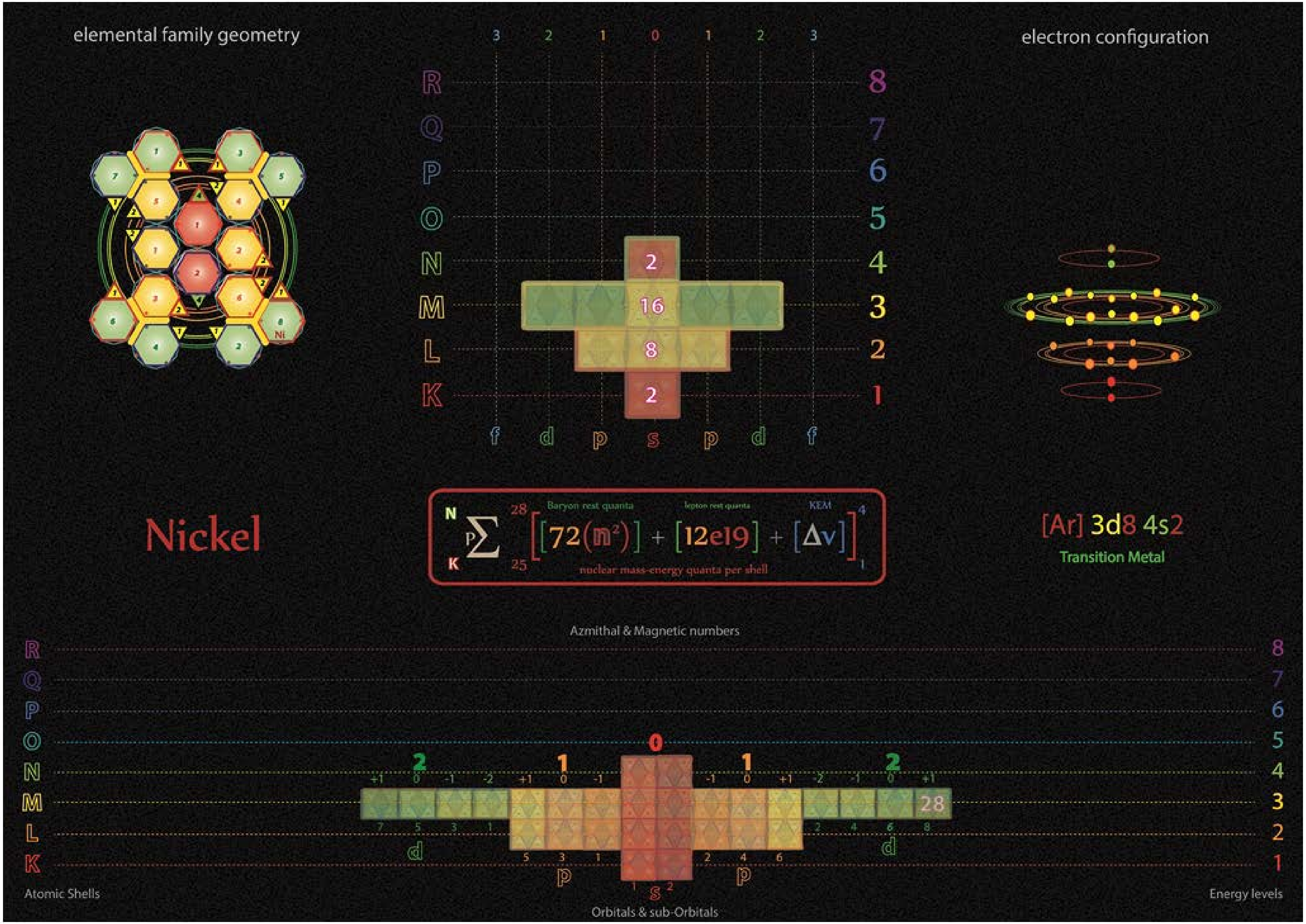
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2
16
8
2
f d p s p d f
Ni
Nickel
N
K
P
Σ
28
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
4
1
nuclear mass-energy quanta per shell
[Ar] 3d8 4s2
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
28
7 5 3 1 2 4 6 8
5 3 1 2 4 6
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.28 - Nickel atomic config

Copyright ABRAHAM [2008] - All rights reserved

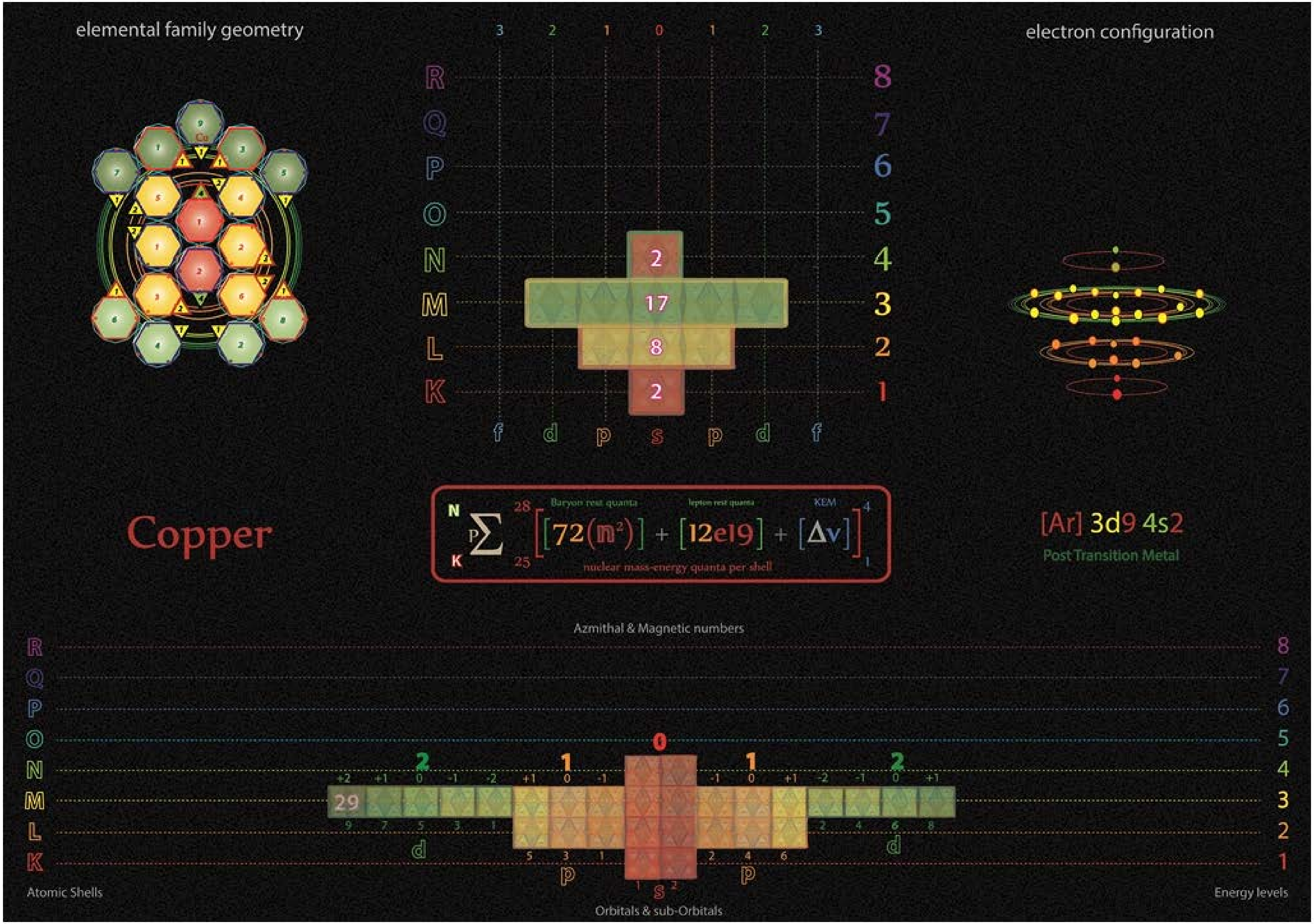
elemental family geometry
Cu
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
17
8
2
f d p s p d f
Copper
N
P
K
28
25
Baryon rest quanta
lepton rest quanta
KEM
4
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Ar] 3d9 4s2
Post Transition Metal
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
29
9 7 5 3 1 2 4 6 8
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.29 - Copper atomic config

Copyright ABRAHAM [2008] - All rights reserved

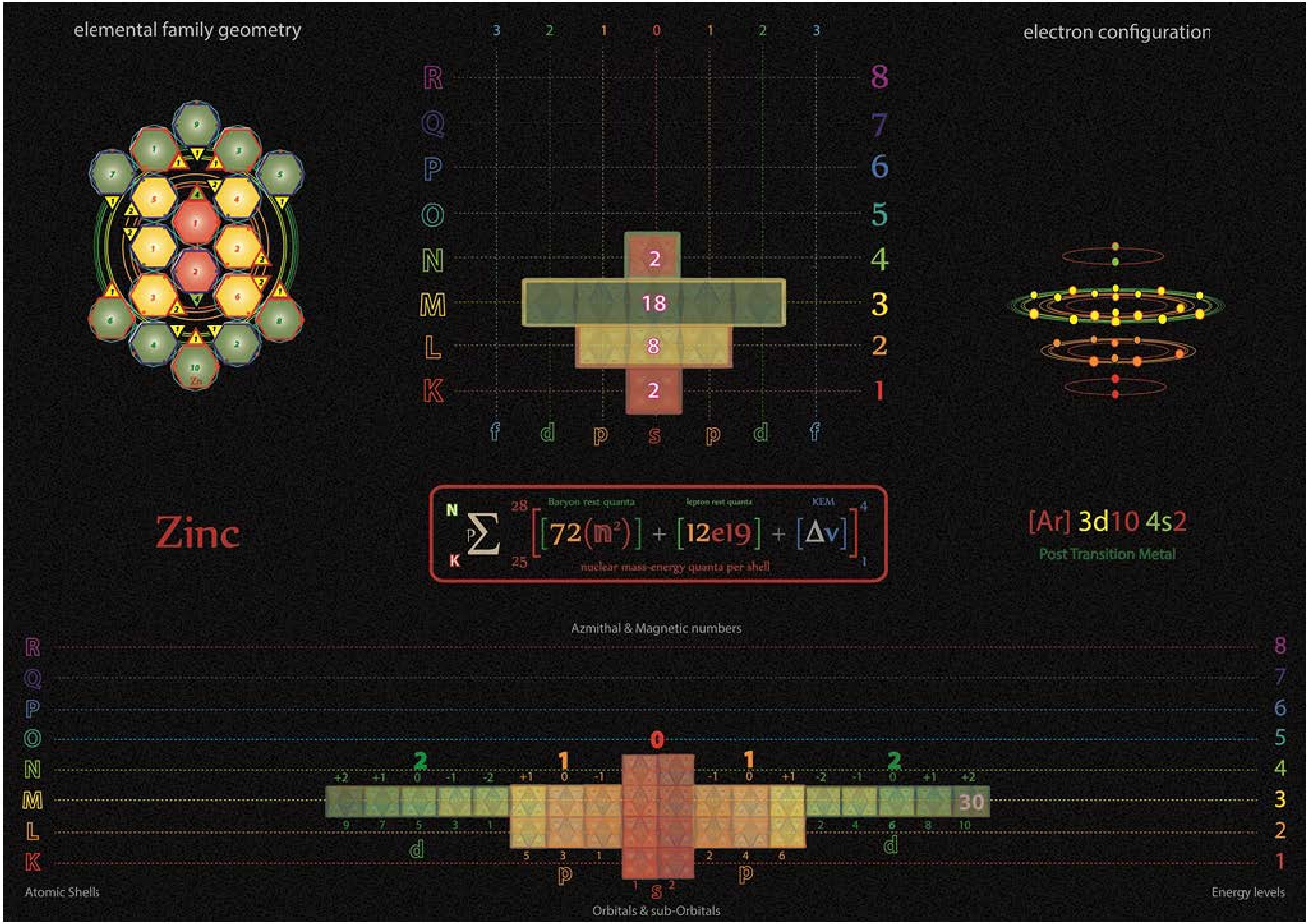
elemental family geometry
electron configuration
Zinc
[Ar] 3d10 4s2
Post Transition Metal
N
K
28
25
P
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
30
9 7 5 3 1 5 3 1 2 4 6 2 4 6 8 10
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.30 - Zinc atomic config

Copyright ABRAHAM [2008] - All rights reserved

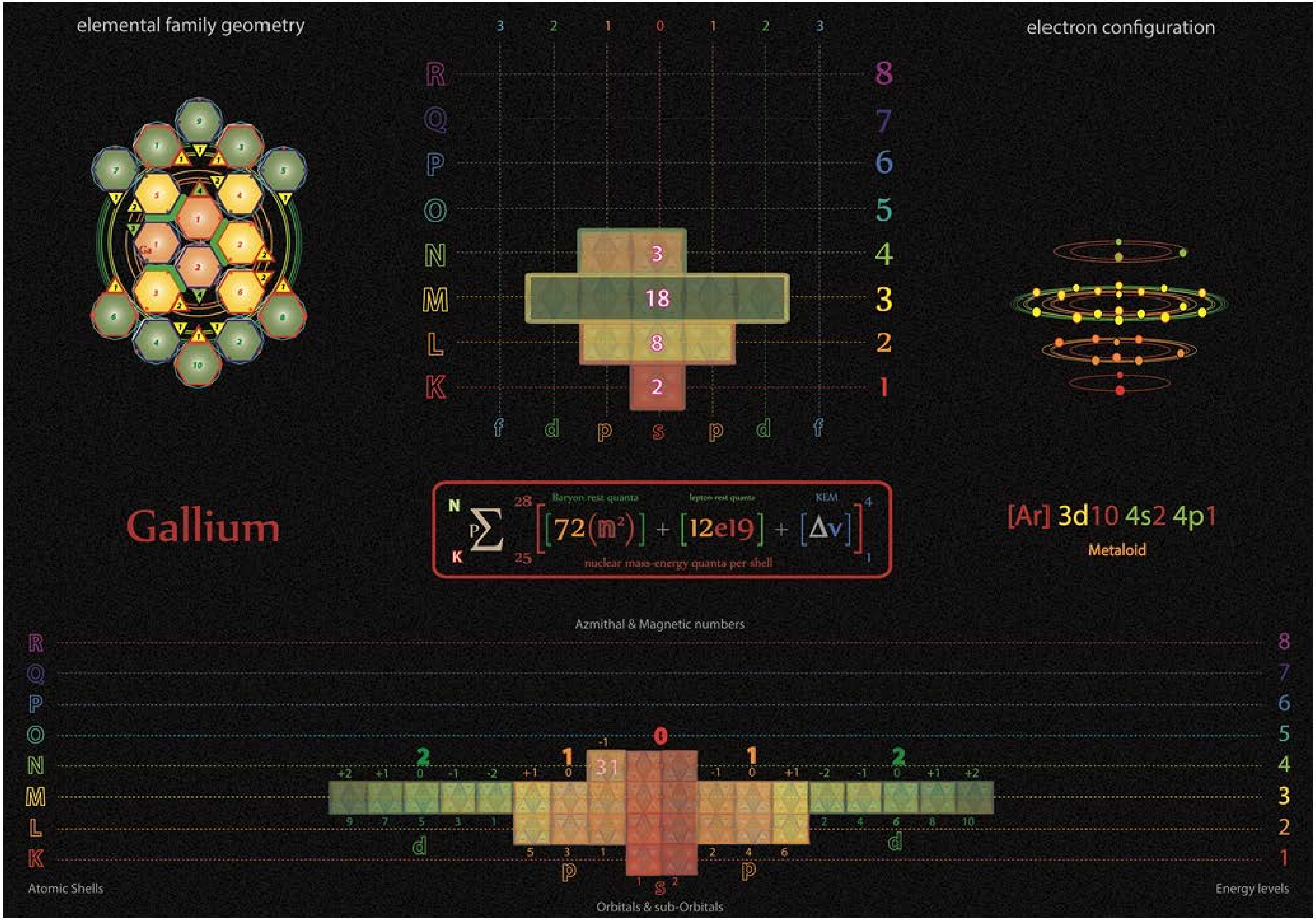
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
3
18
8
2
Gallium
N
K
P
28
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Ar] 3d10 4s2 4p1
Metaloid
Azmithal & Magnetic numbers
31
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.31 - Gallium atomic config

elemental family geometry

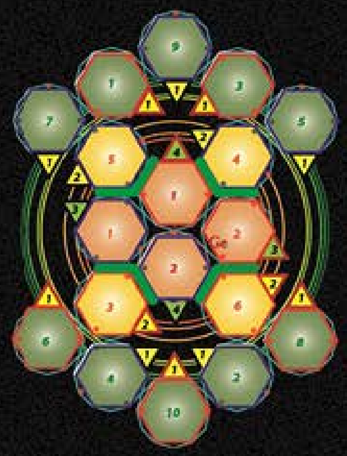

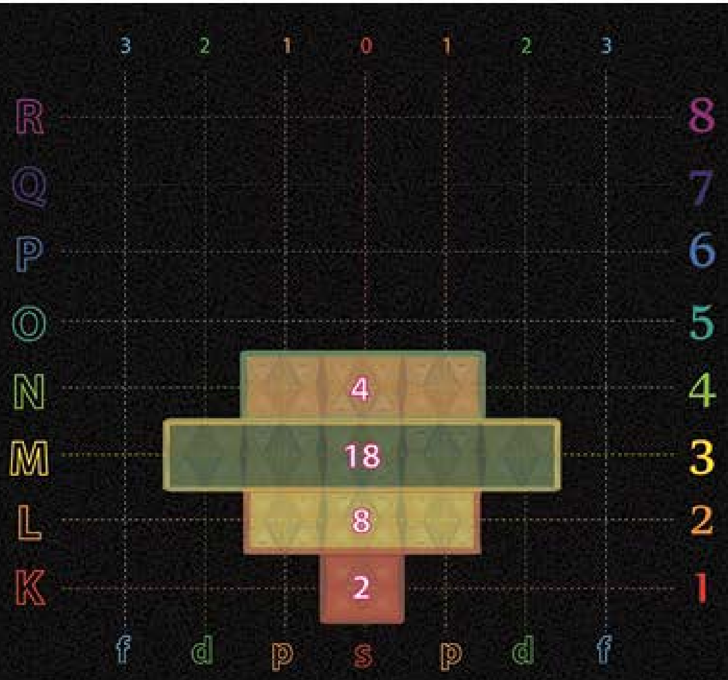
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
18
8
2
f
d
p
s
p
d
f

electron configuration

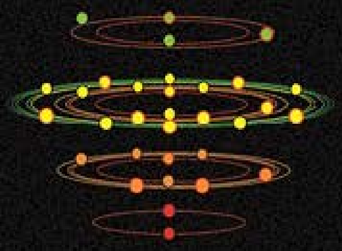

Germanium

Baryon rest quanta · lepton rest quanta · KEM

$$ {}_{K}^{N}\sum_{p}{}_{25}^{28}\left[\left[72(m^2)\right] + \left[12e19\right] + \left[\Delta v\right] \right]_{1}^{4} $$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p2

Metaloid

Azmithal & Magnetic numbers

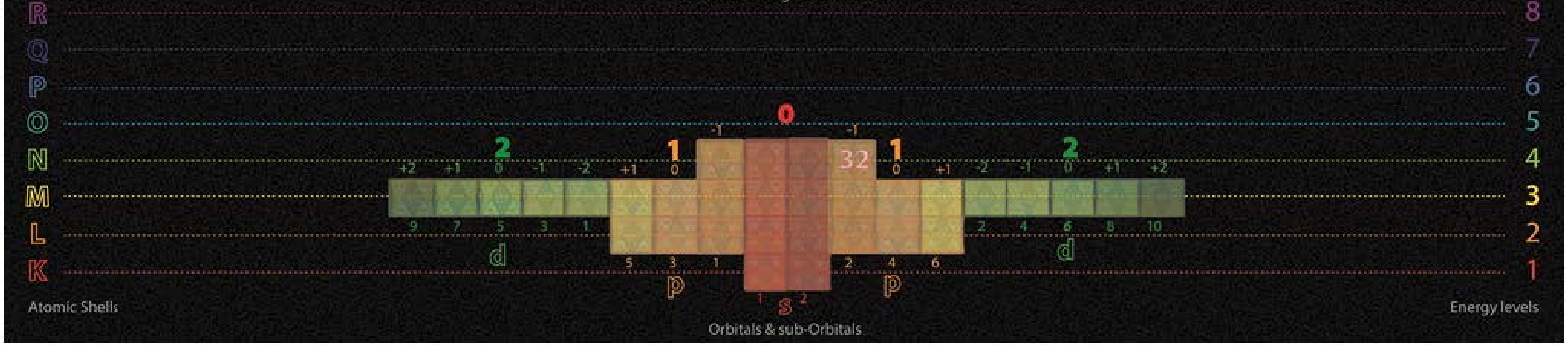
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
-1
-1
32
+2
+1
0
-1
-2
+1
0
0
+1
-2
-1
0
+1
+2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.32 - Germanium atomic config

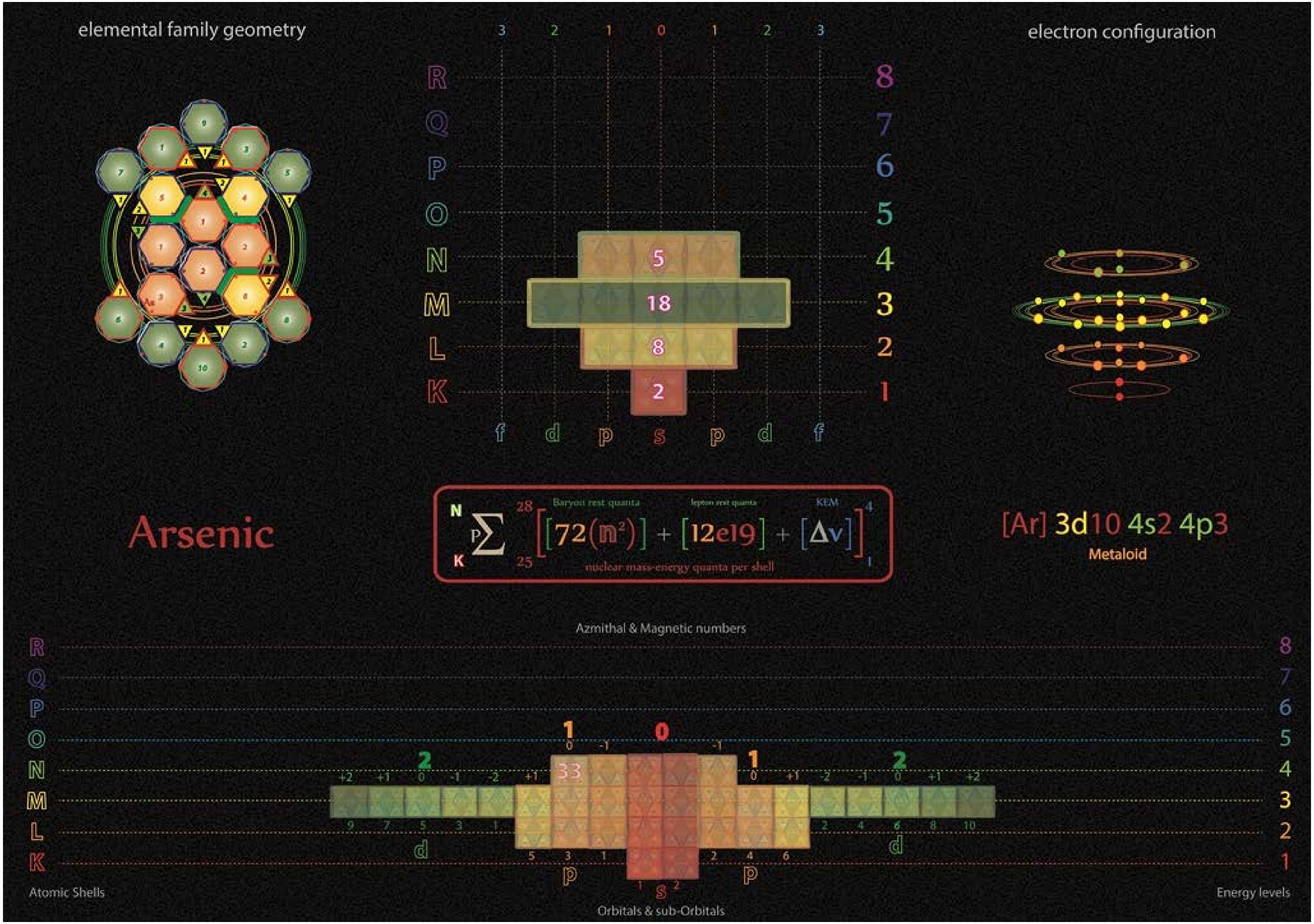
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
1 0 1 2 1 2
0 -1 -1 0
+2 +1 0 -1 -2 +1 +1 -2 -1 0 +1 +2
33
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Arsenic

Baryon rest quanta — lepton rest quanta — KEM

$${}^{N}_{K}\,{}_{P}\sum^{28}_{25}\left[[72(n^{2})]+[12e19]+[\Delta v]\right]^{4}_{1}$$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p3

Metaloid

Tetryonics 53.33 - Arsenic atomic config

Copyright ABRAHAM [2008] - All rights reserved

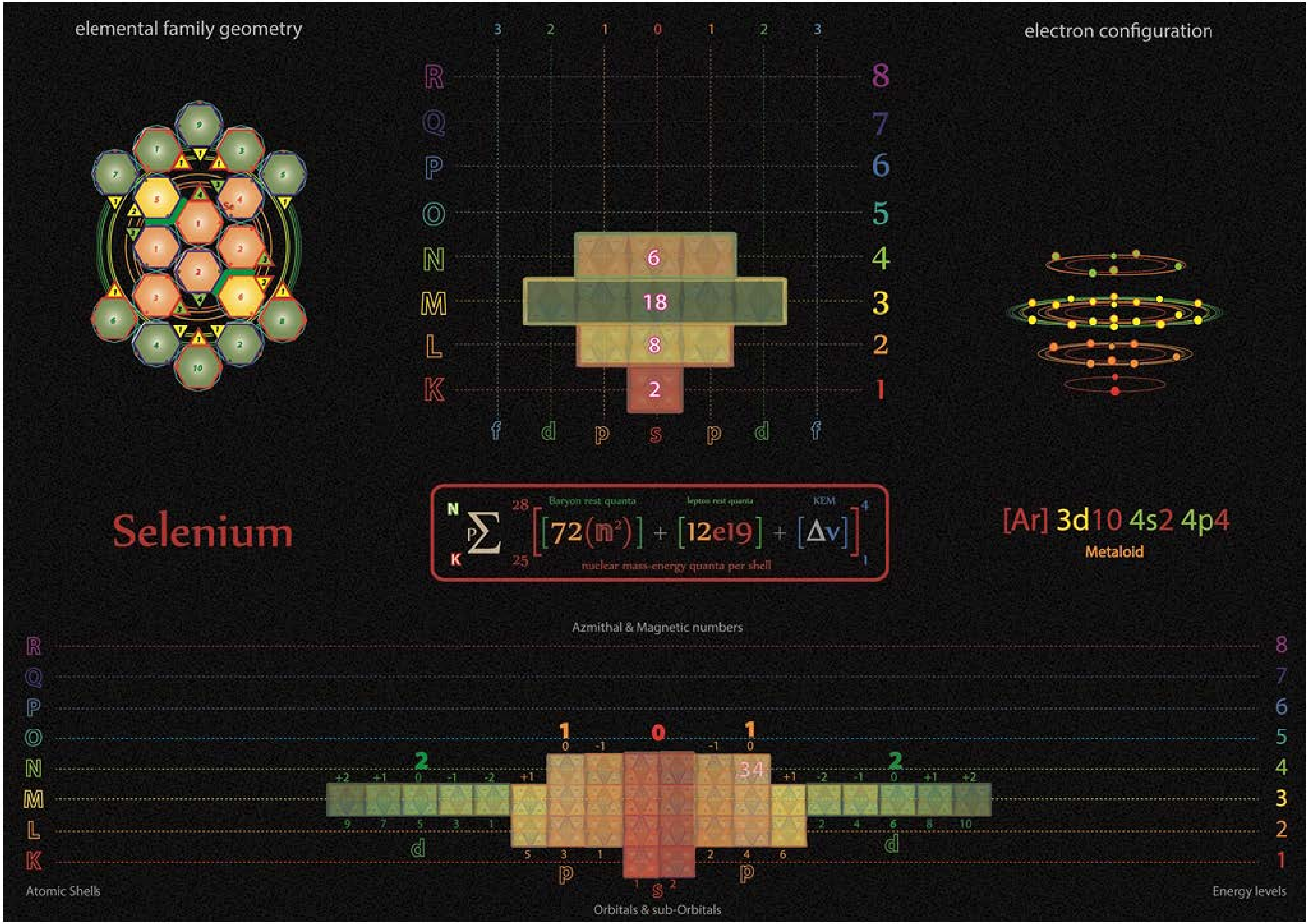

elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
2 2
+2 +1 0 -1 -2 +1 34 +1 -2 -1 0 +1 +2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Selenium

$$\sum_{K}^{N} {}_{p}\sum_{25}^{28}\left[\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{4}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p4

Metaloid

Tetryonics 53.34 - Selenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

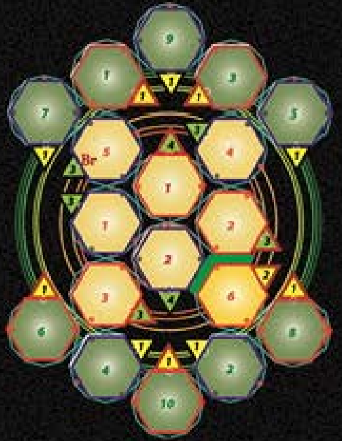

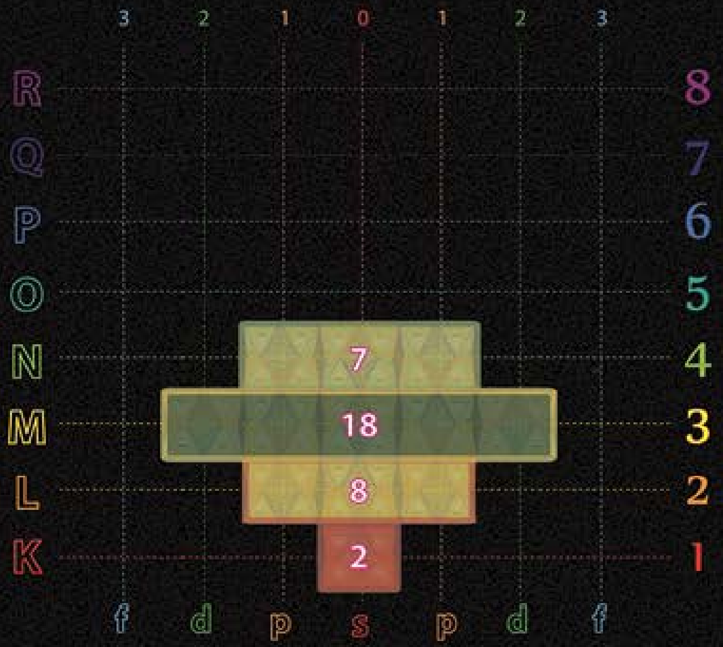
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
7
18
8
2
f d p s p d f

electron configuration

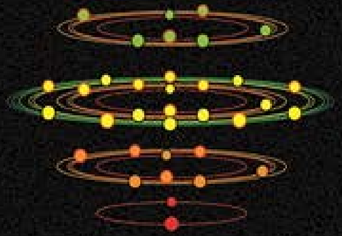

Bromine

$$ {}_{K}^{N}\,{}_{P}\sum_{25}^{28}\left[\overset{\text{Baryon rest quanta}}{[72(n^{2})]}+\overset{\text{lepton rest quanta}}{[12e19]}+\overset{\text{KEM}}{[\Delta v]}\right]_{1}^{4} $$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p5

Halogen

Azmithal & Magnetic numbers

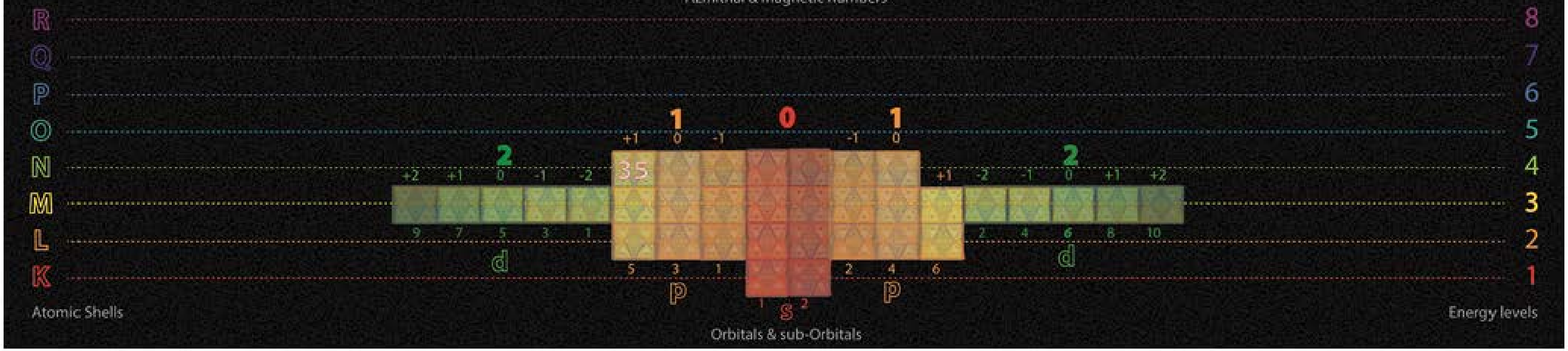
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
35
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.35 - Bromine atomic config

elemental family geometry

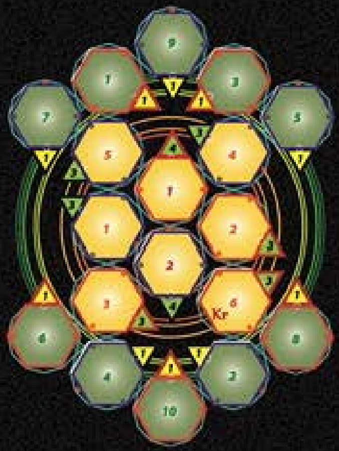

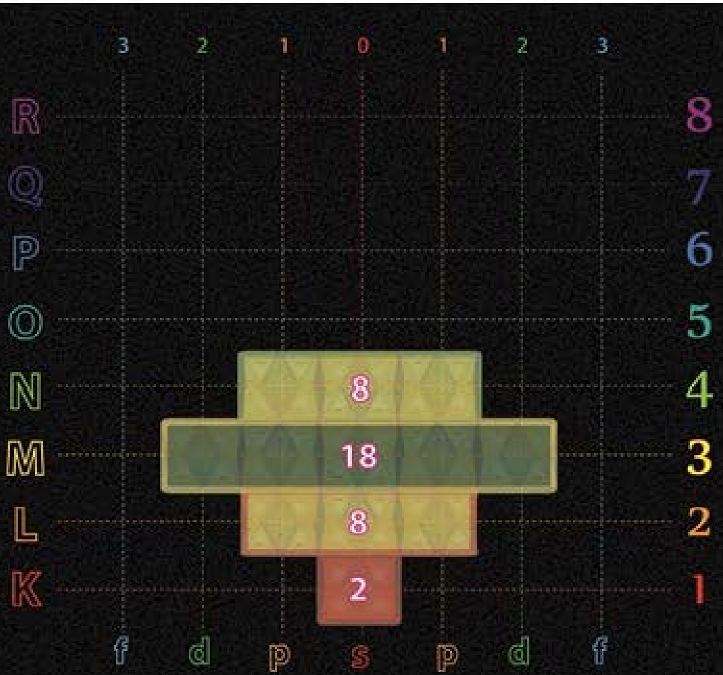
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
18
8
2
f
d
p
s
p
d
f

electron configuration

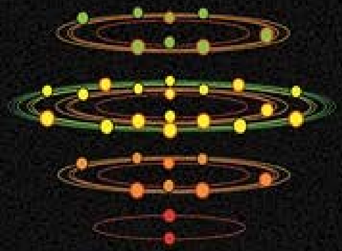

Krypton

Baryon rest quanta · lepton rest quanta · KEM

$$ {}_{K}^{N}\,{}^{P}\!\sum_{25}^{28}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]_{1}^{4} $$

nuclear mass-energy quanta per shell

[Ar] 3d10 4s2 4p6

Noble Gas

Azmithal & Magnetic numbers

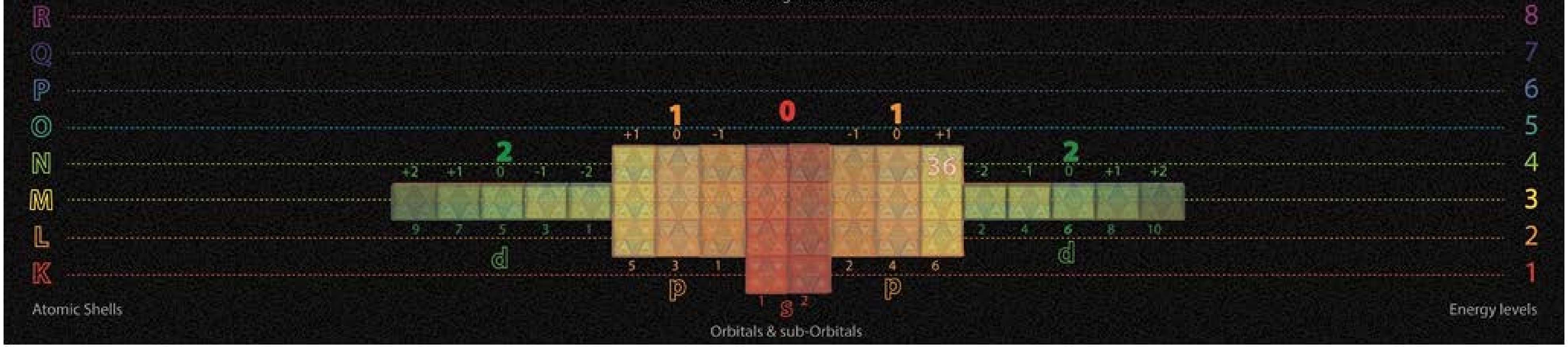
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
2
2
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
36
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
p
s
p
d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.36 - Krypton atomic config

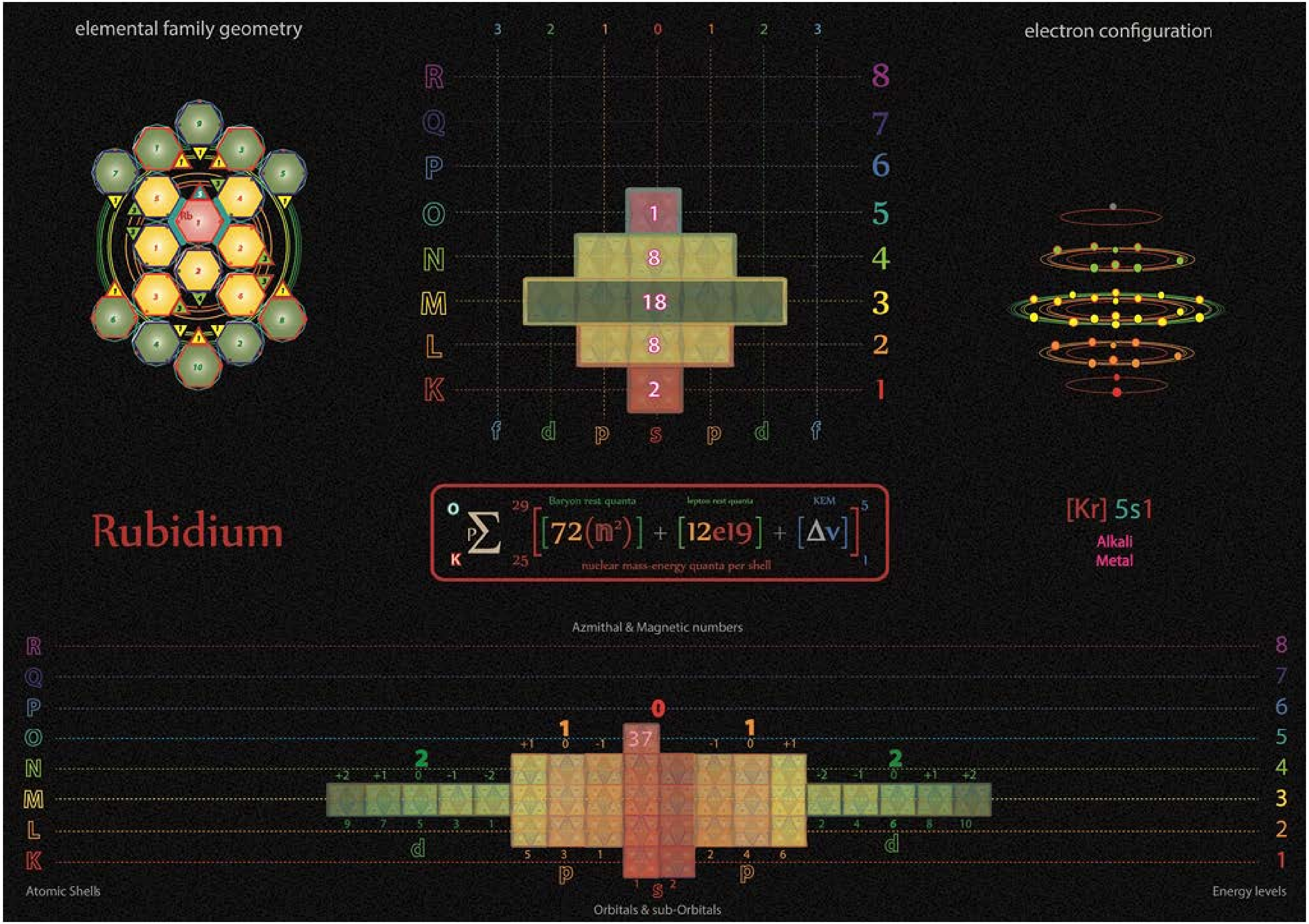
elemental family geometry
Rb
Rubidium
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
18
8
2
f d p s p d f
electron configuration
[Kr] 5s1
Alkali
Metal
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
0
1
1
2
2
37
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
p
s
p
d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.37 - Rubidium atomic config

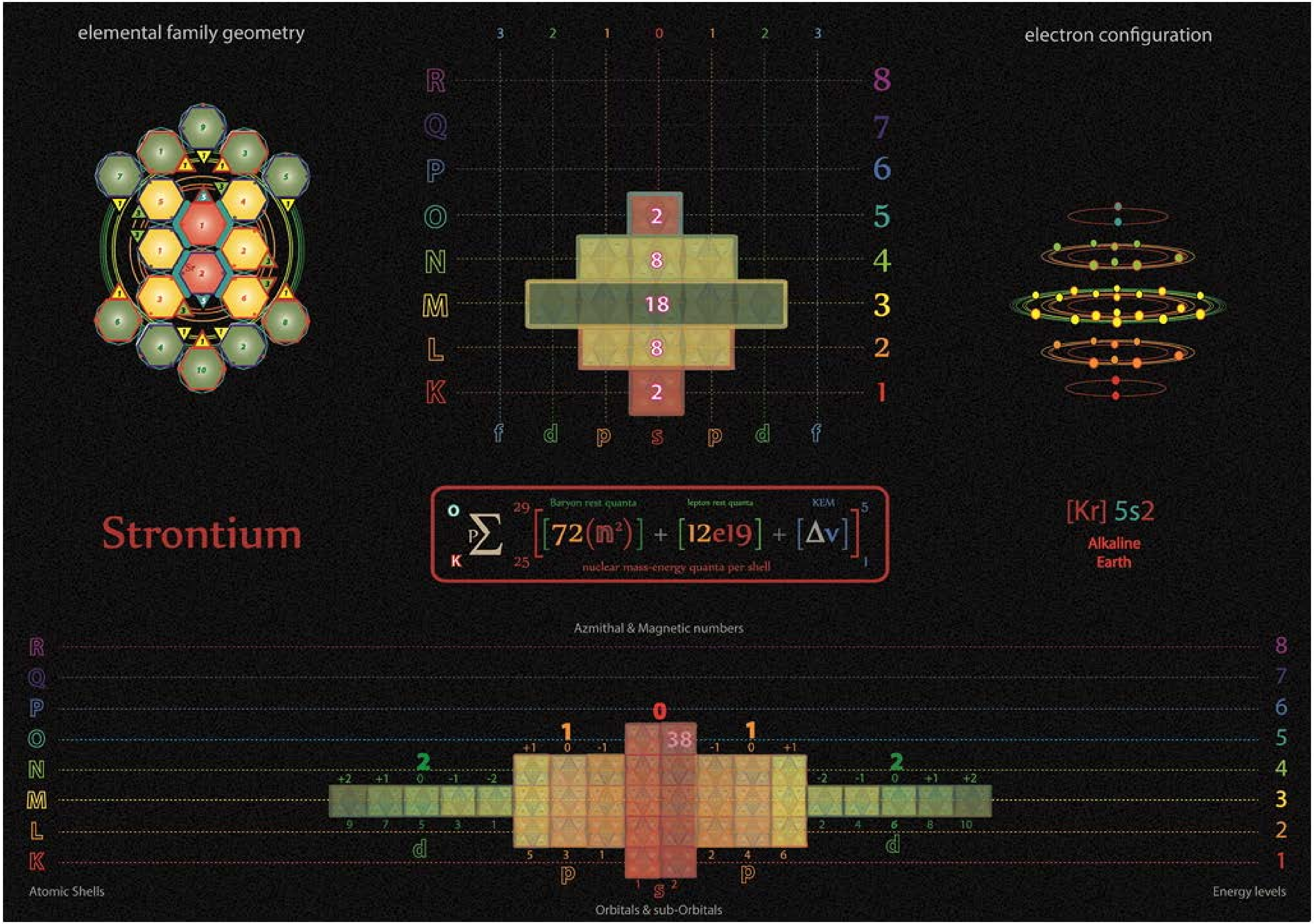
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
38
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
9 7 5 3 1 5 3 1 2 4 6 2 4 6 8 10
d p s p d
1 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Strontium

$$ {}^{O}_{K}\,{}_{P}\sum_{25}^{29}\left[\left[72(\mathrm{m}^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{5} $$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Kr] 5s2

Alkaline Earth

Tetryonics 53.38 - Strontium atomic config

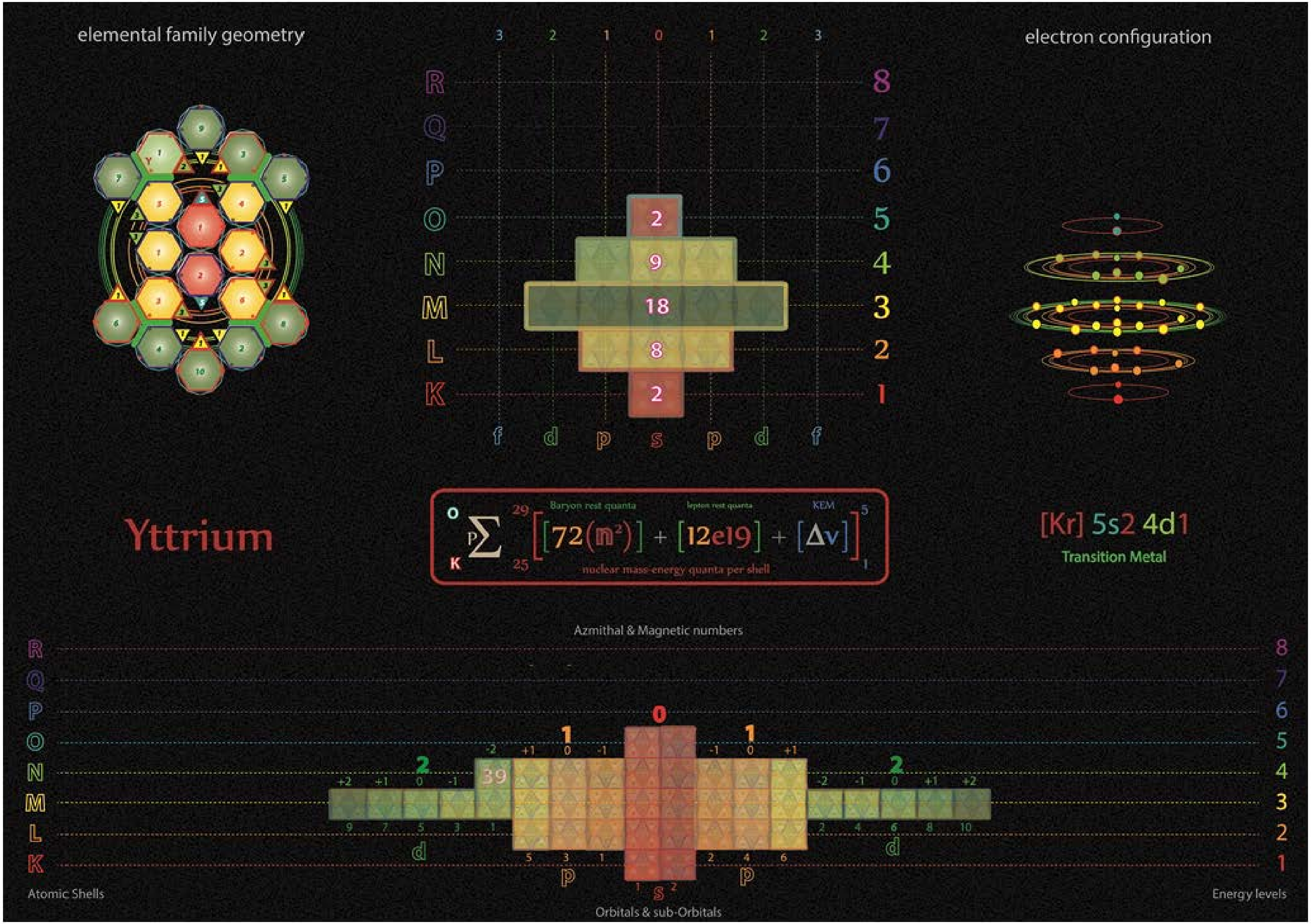
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
9
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
-2 +1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -1 0 +1 +2
39
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Yttrium

$$\sum_{K}^{O} P \sum_{25}^{29} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{5}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d1

Transition Metal

Tetryonics 53.39 - Yttrium atomic config

Copyright ABRAHAM [2008] - All rights reserved

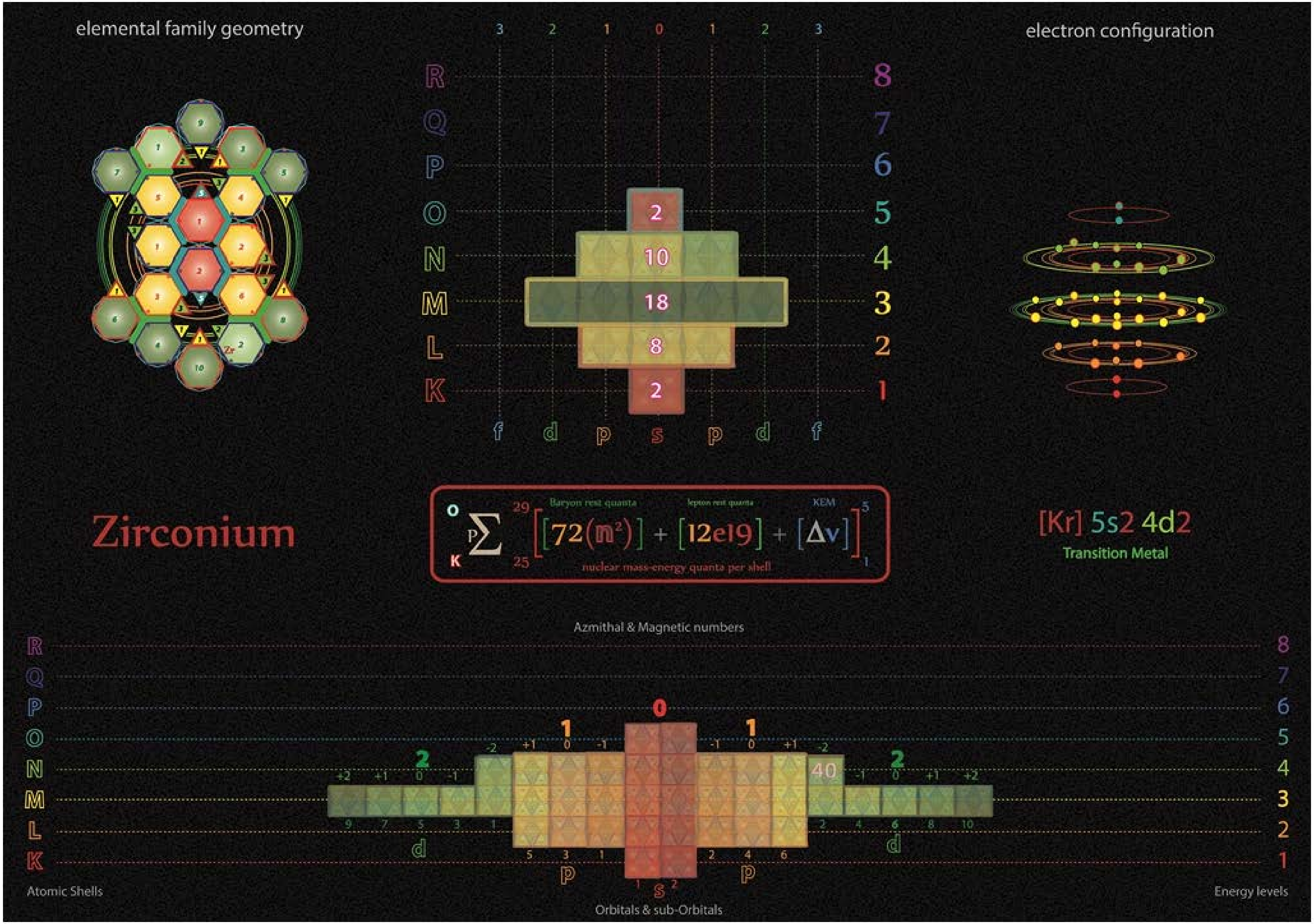

elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
10
18
8
2
f d p s p d f
Zirconium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d2
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
-2 +1 0 -1
-1 0 +1 -2
+2 +1 0 -1
40
-1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.40 - Zirconium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

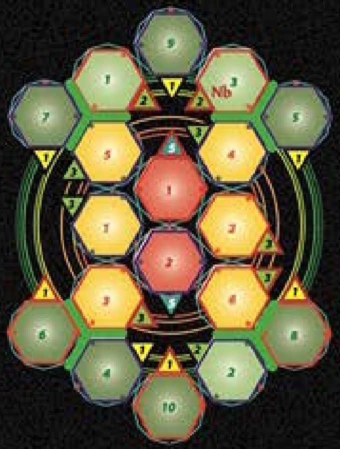
Nb
1
2
3
4
5
6
7
8
9
10

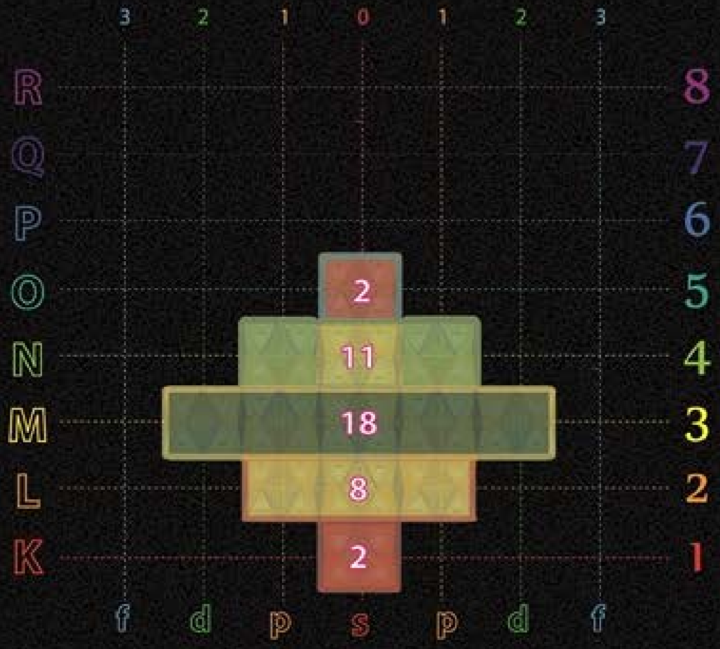
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
11
18
8
2
f
d
p
s
p
d
f

electron configuration

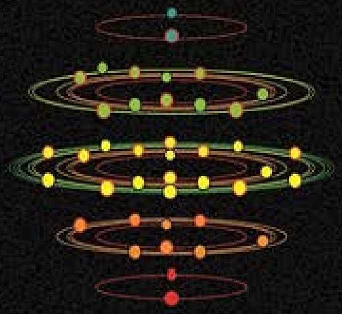

Niobium

$$\sum_{K}^{O} P \left[\overset{\text{Baryon rest quanta}}{\left[72(m^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]} \right]_{1}^{5} \quad {}^{29}_{25}$$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d3

Transition Metal

Azmithal & Magnetic numbers

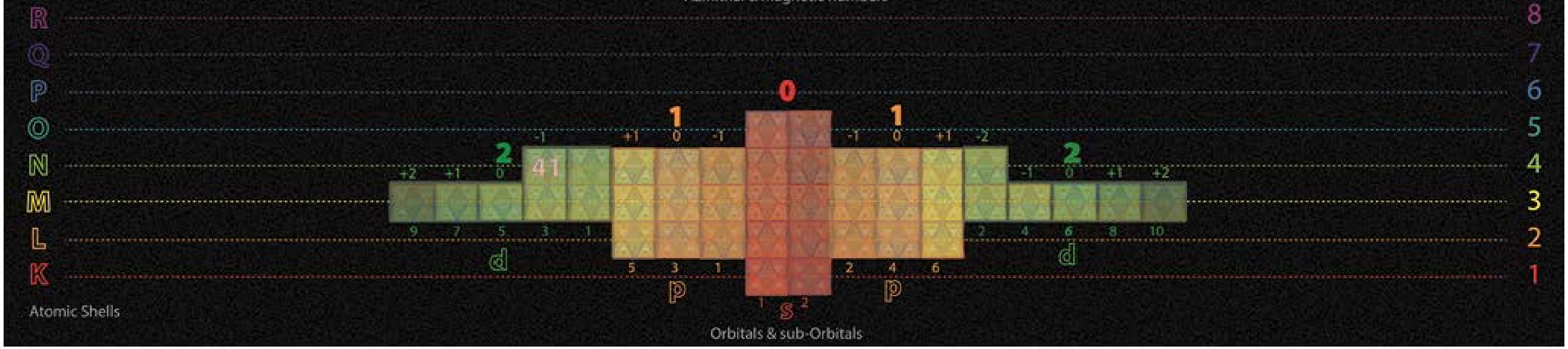
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
41
+2
+1
0
-1
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
9
7
5
3
1
5
3
1
2
4
6
2
4
6
8
10
d
p
s
1
2
p
d
Atomic Shells

Orbitals & sub-Orbitals

Tetryonics 53.41 - Niobium atomic config

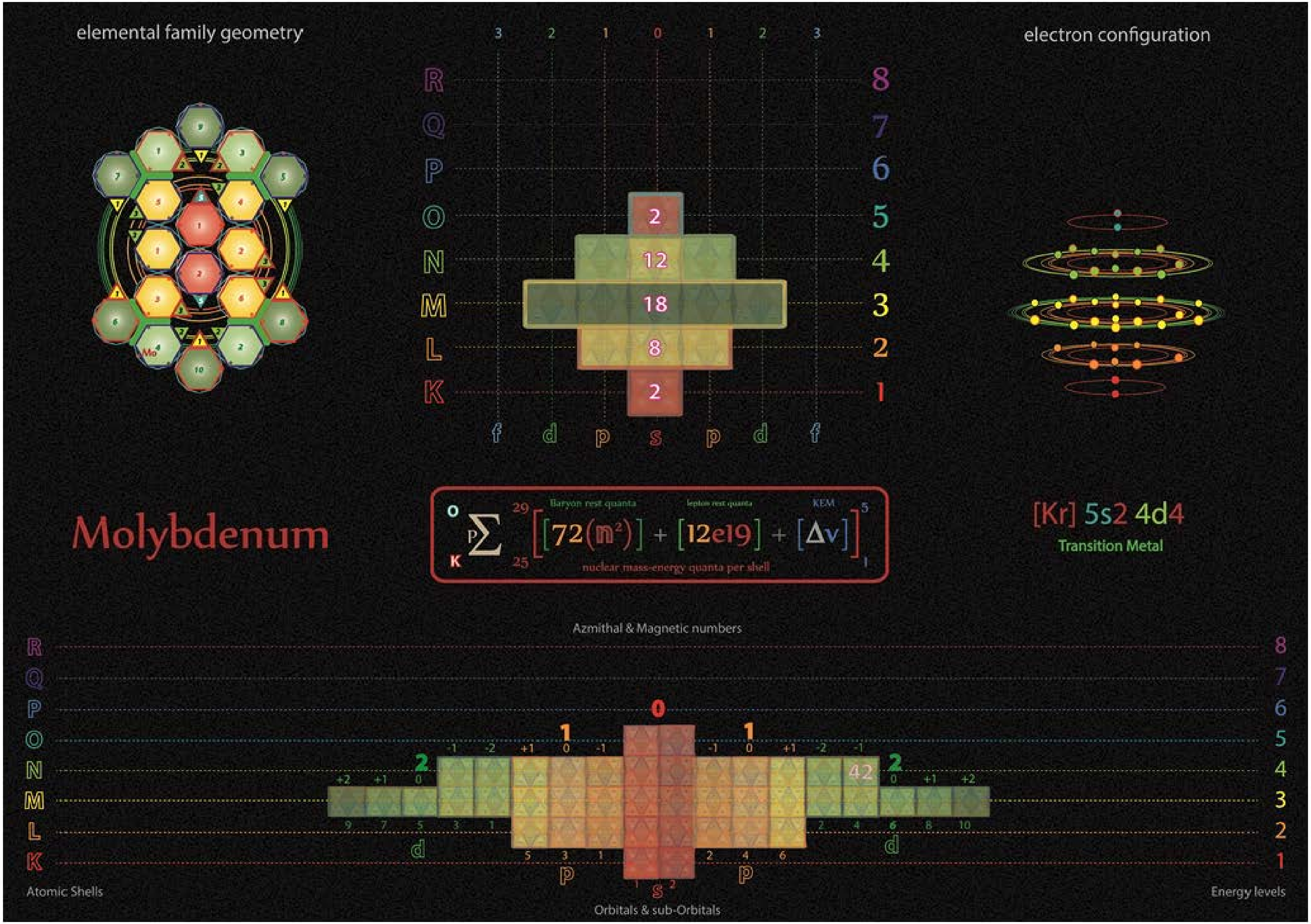
elemental family geometry
Mo
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
12
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
-1 -2 +1 0 -1 -1 0 +1 -2 -1
+2 +1 0 0 +1 +2
42
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d d
p p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Molybdenum

Baryon rest quanta · lepton rest quanta · KEM

$$ {}_{K}^{O}\, {}_{P}\sum_{25}^{29} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{5} $$

nuclear mass-energy quanta per shell

[Kr] 5s2 4d4

Transition Metal

Tetryonics 53.42 - Molybdenum atomic config

elemental family geometry

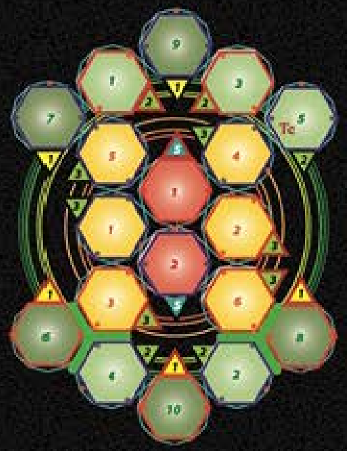

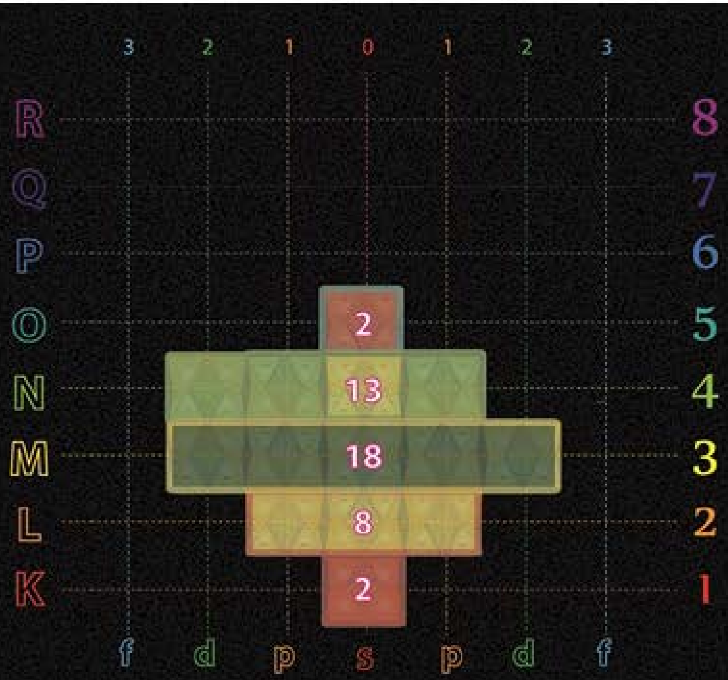
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
13
18
8
2
f
d
p
s
p
d
f

electron configuration

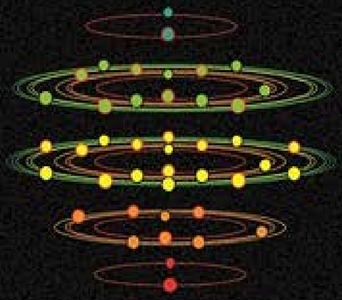

Technetium

$$\sum_{K}^{O} \mathrm{p} \sum_{25}^{29} \left[\left[72(n^2)\right] + \left[12e19\right] + \left[\Delta v\right] \right]_{1}^{5}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d5

Transition Metal

Azmithal & Magnetic numbers

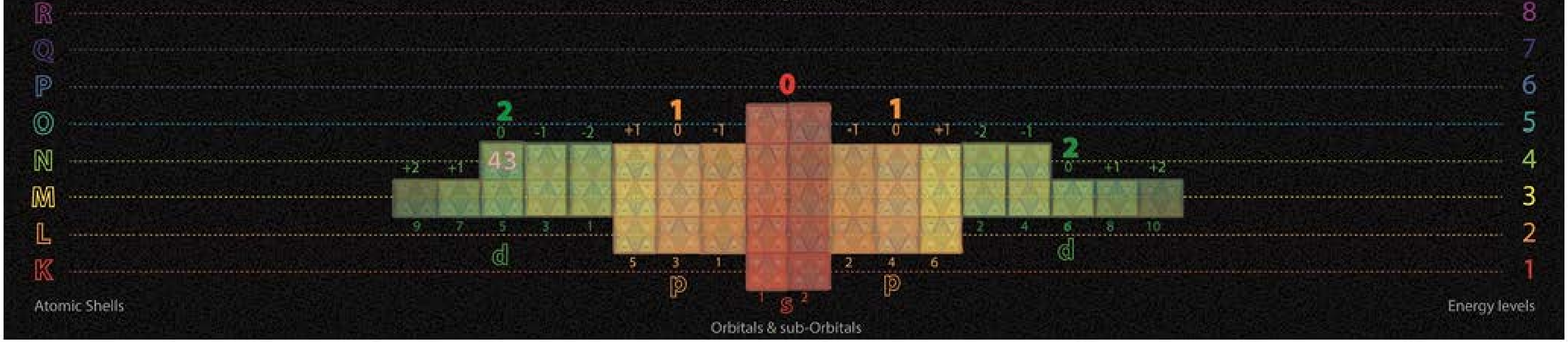
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
43
d
p
s
p
d
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

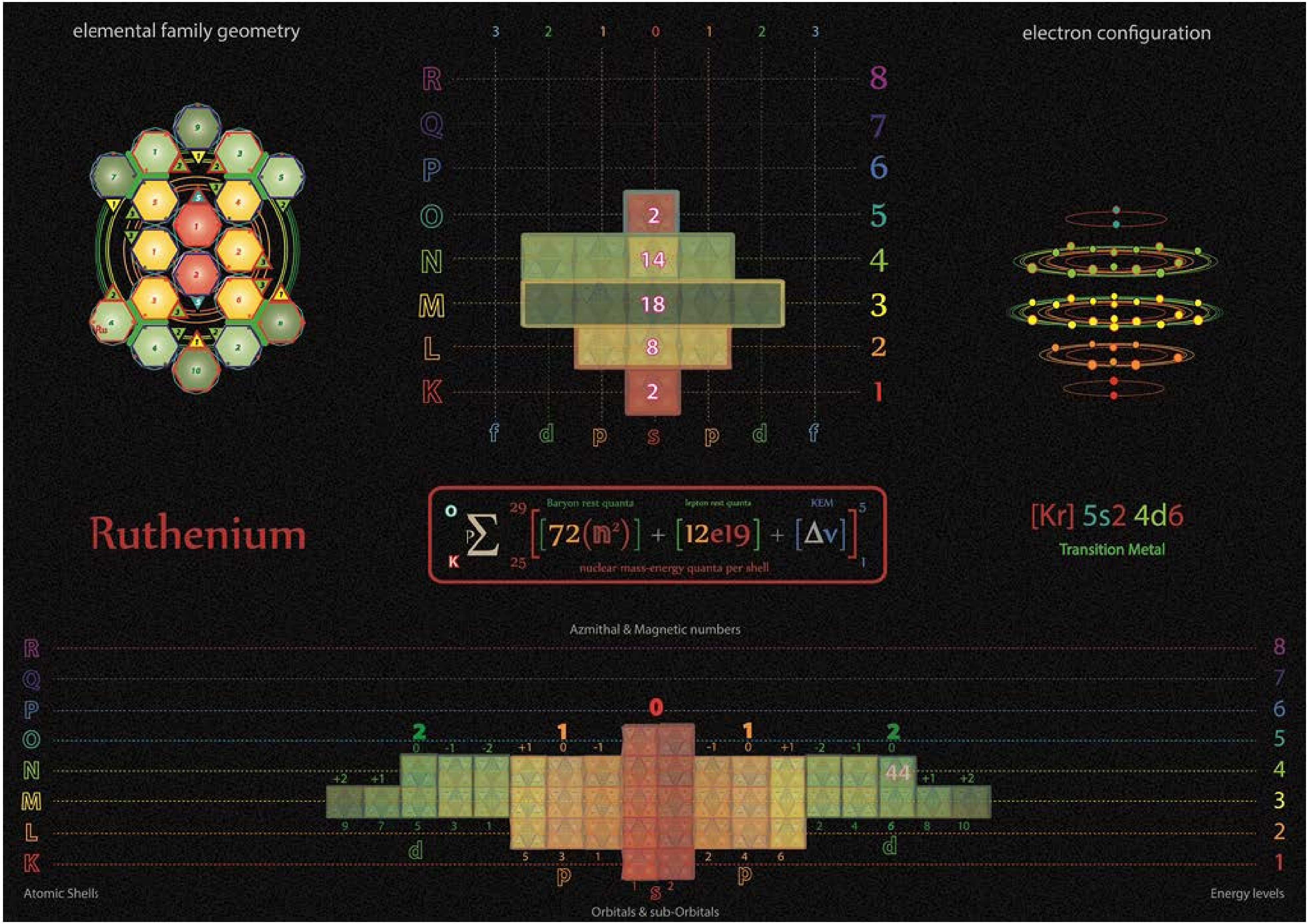
elemental family geometry
Ruthenium
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
18
8
2
f d p s p d f
electron configuration
O
K
p
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d6
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
+2 +1 +1 +2
44
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.44 - Ruthenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

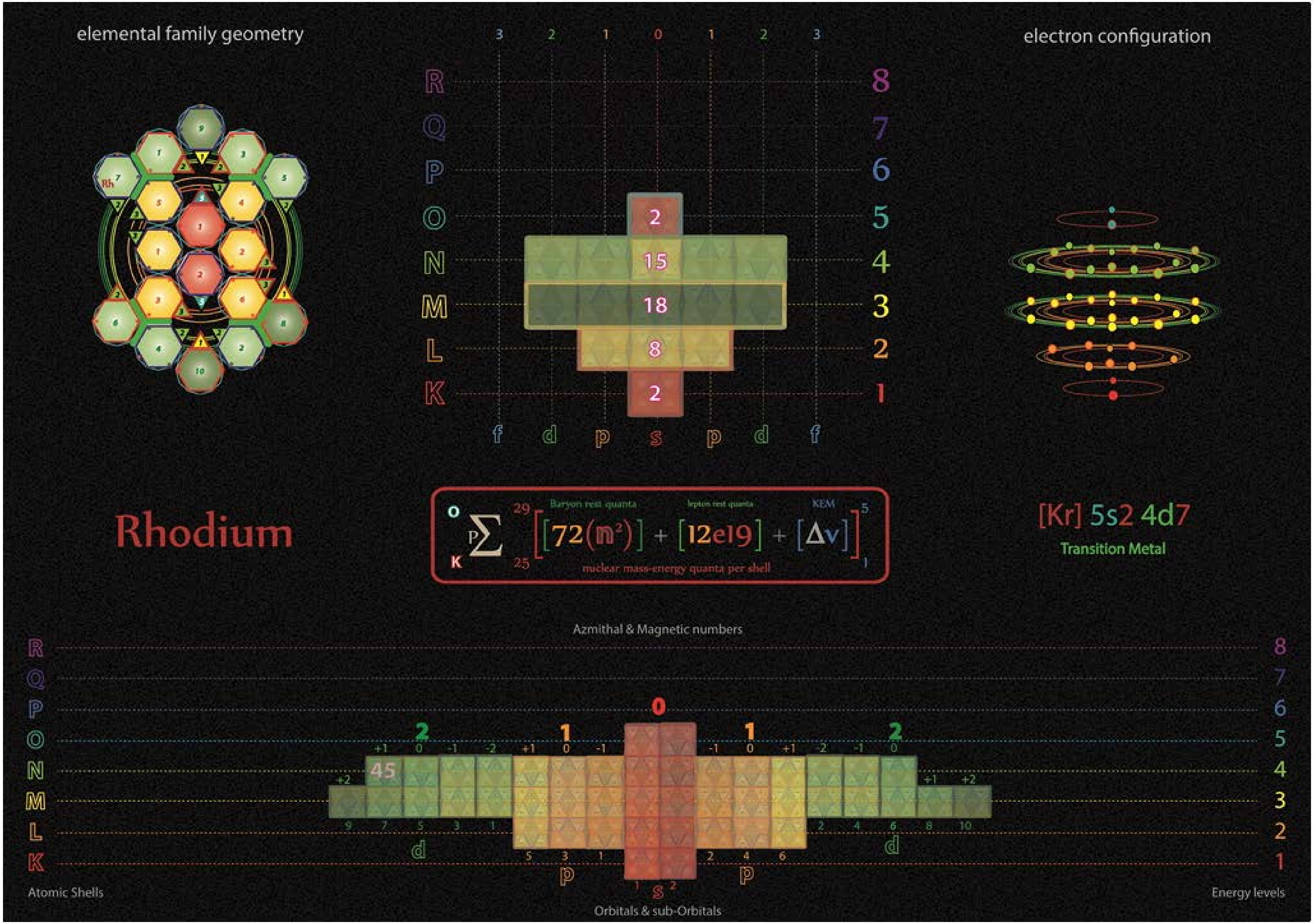
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2
15
18
8
2
f d p s p d f
Rhodium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
5
1
[Kr] 5s2 4d7
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
+2 +1 +2
45
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.45 - Rhodium atomic config

Copyright ABRAHAM [2008] - All rights reserved

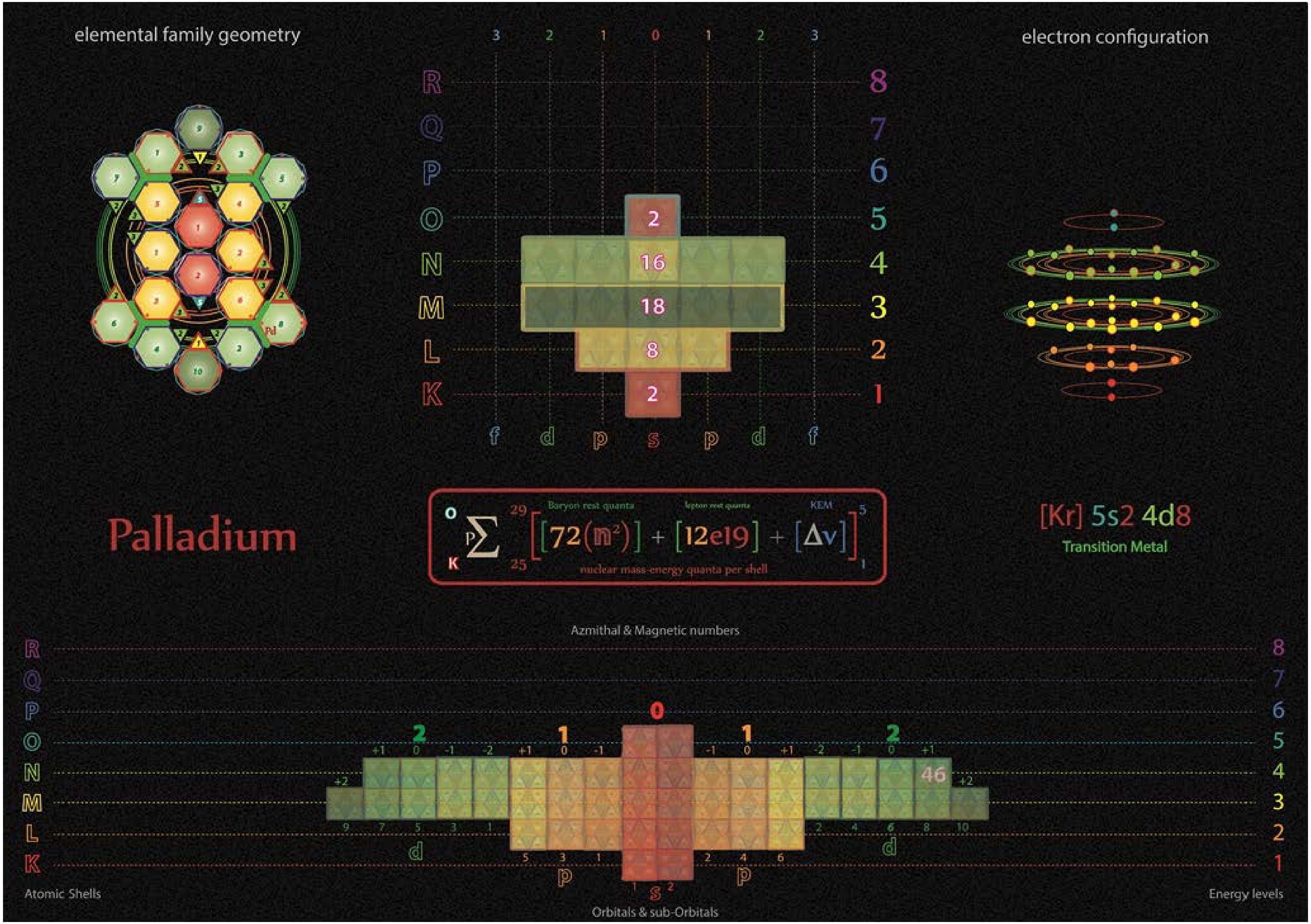
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
16
18
8
2
f d p s p d f
Pd
Azmithal & Magnetic numbers
0
2 1 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
+2 +2
46
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d d
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Palladium

$$\sum_{K}^{O}{}_{p}\sum_{25}^{29}\left[\underbrace{\overset{\text{Baryon rest quanta}}{[72(n^2)]}+\overset{\text{lepton rest quanta}}{[12e19]}+\overset{\text{KEM}}{[\Delta v]}}_{\text{nuclear mass-energy quanta per shell}}\right]_{1}^{5}$$

[Kr] 5s2 4d8

Transition Metal

Tetryonics 53.46 - Palladium atomic config

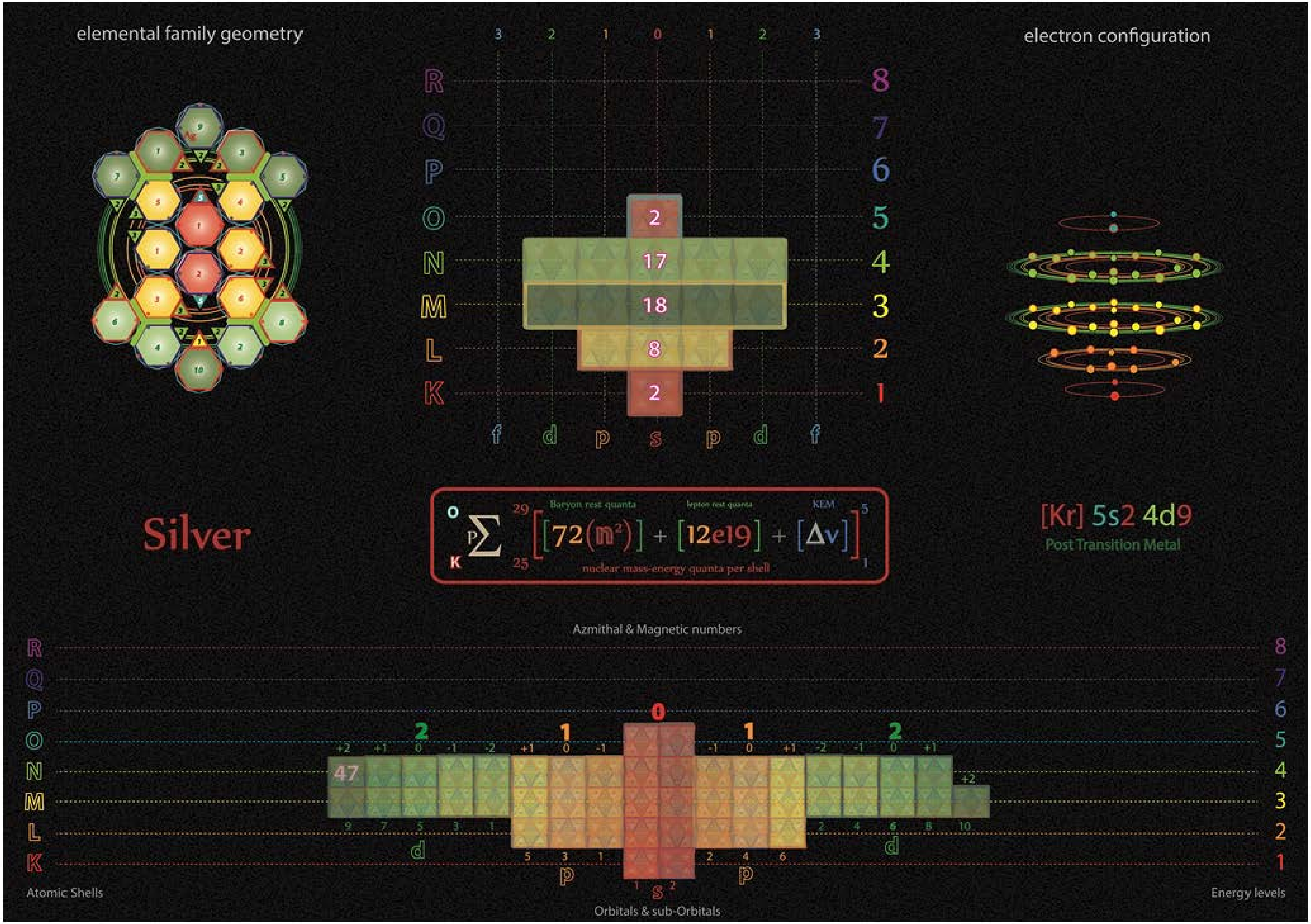
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
17
18
8
2
f d p s p d f
Silver
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
[Kr] 5s2 4d9
Post Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
47
9 7 5 3 1 5 3 1 1 2 2 4 6 2 4 6 8 10
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.47 - Silver atomic config

elemental family geometry

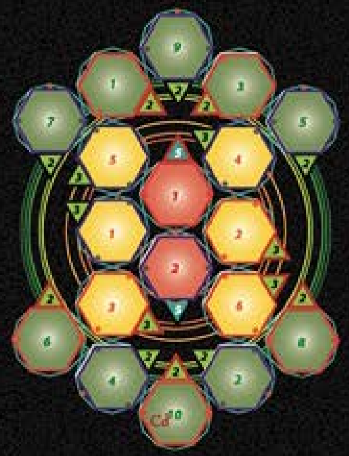
1
2
3
4
5
6
7
8
9
Cd 10

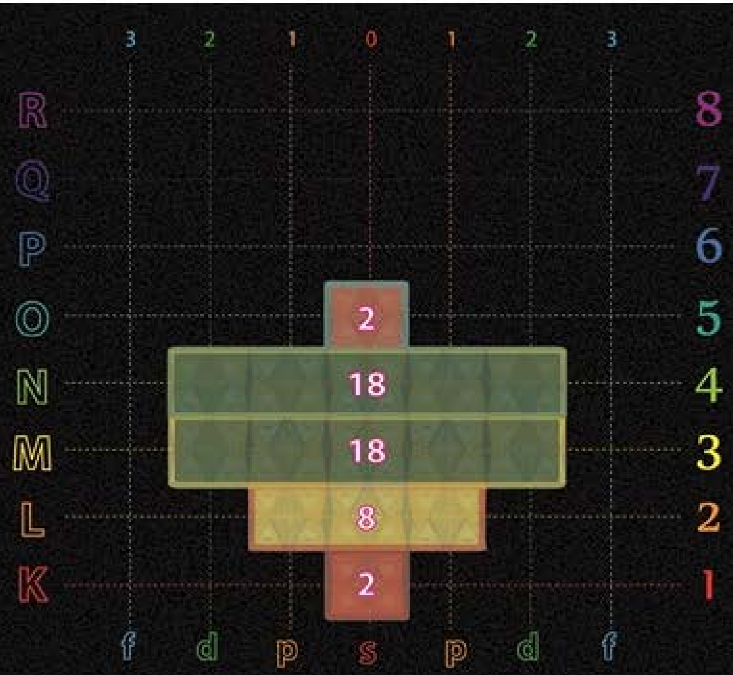
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
18
18
8
2
f d p s p d f

electron configuration

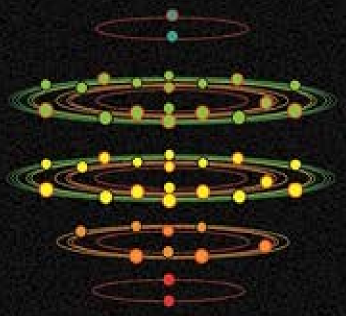

Cadmium

$$ {}^{O}_{K}\;P\sum_{25}^{29}\left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\right]_{1}^{5} $$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10

Post Transition Metal

Azmithal & Magnetic numbers

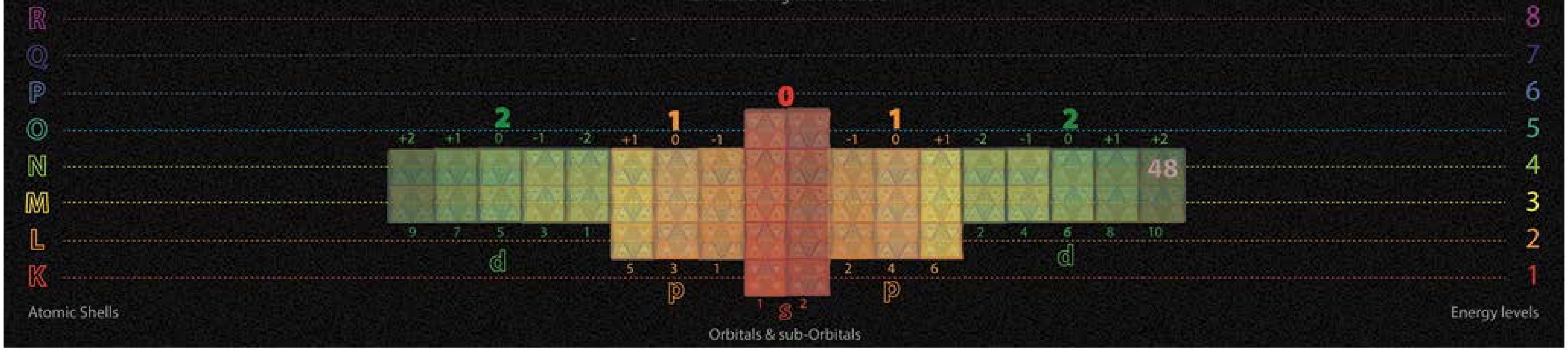
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
48
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.48 - Cadmium atomic config

elemental family geometry

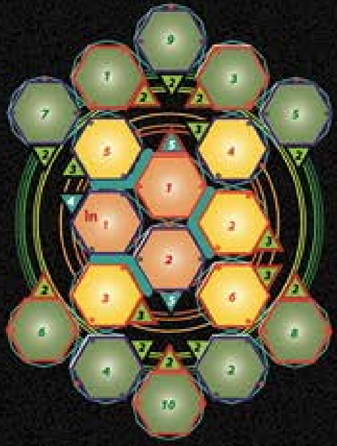

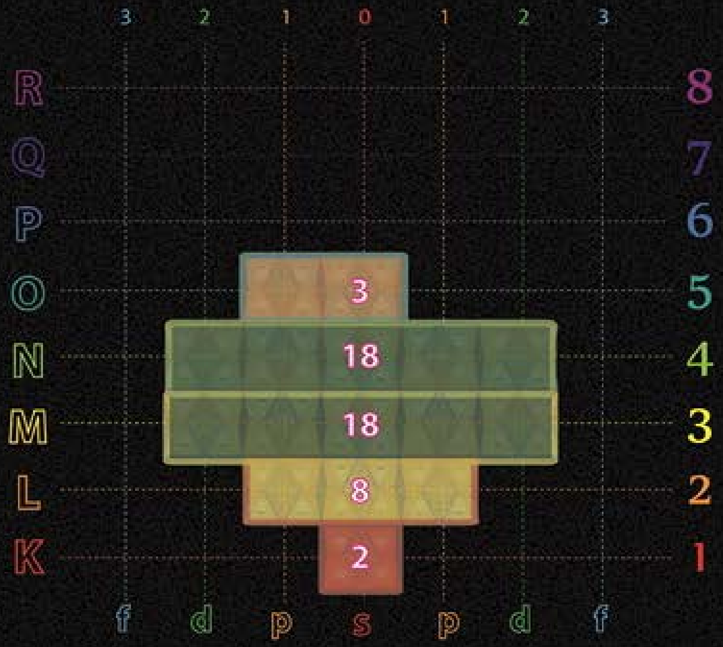
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
18
8
2
f d p s p d f

electron configuration

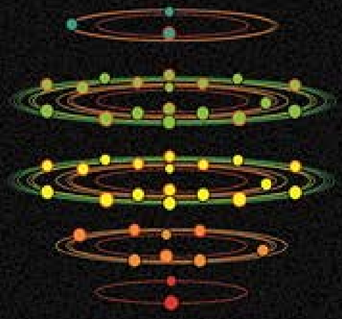

Indium

$$ {}^{O}_{K}\,{}^{P}\!\sum_{25}^{29}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]_{1}^{5} $$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10 5p1

Metaloid

Azmithal & Magnetic numbers

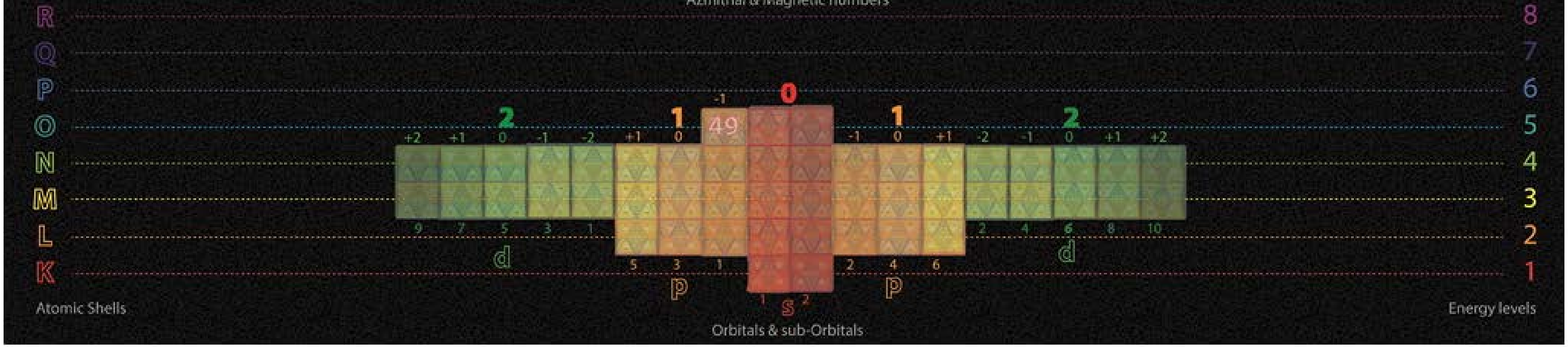
R Q P O N M L K
8 7 6 5 4 3 2 1
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
49
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.49 - Indium atomic config

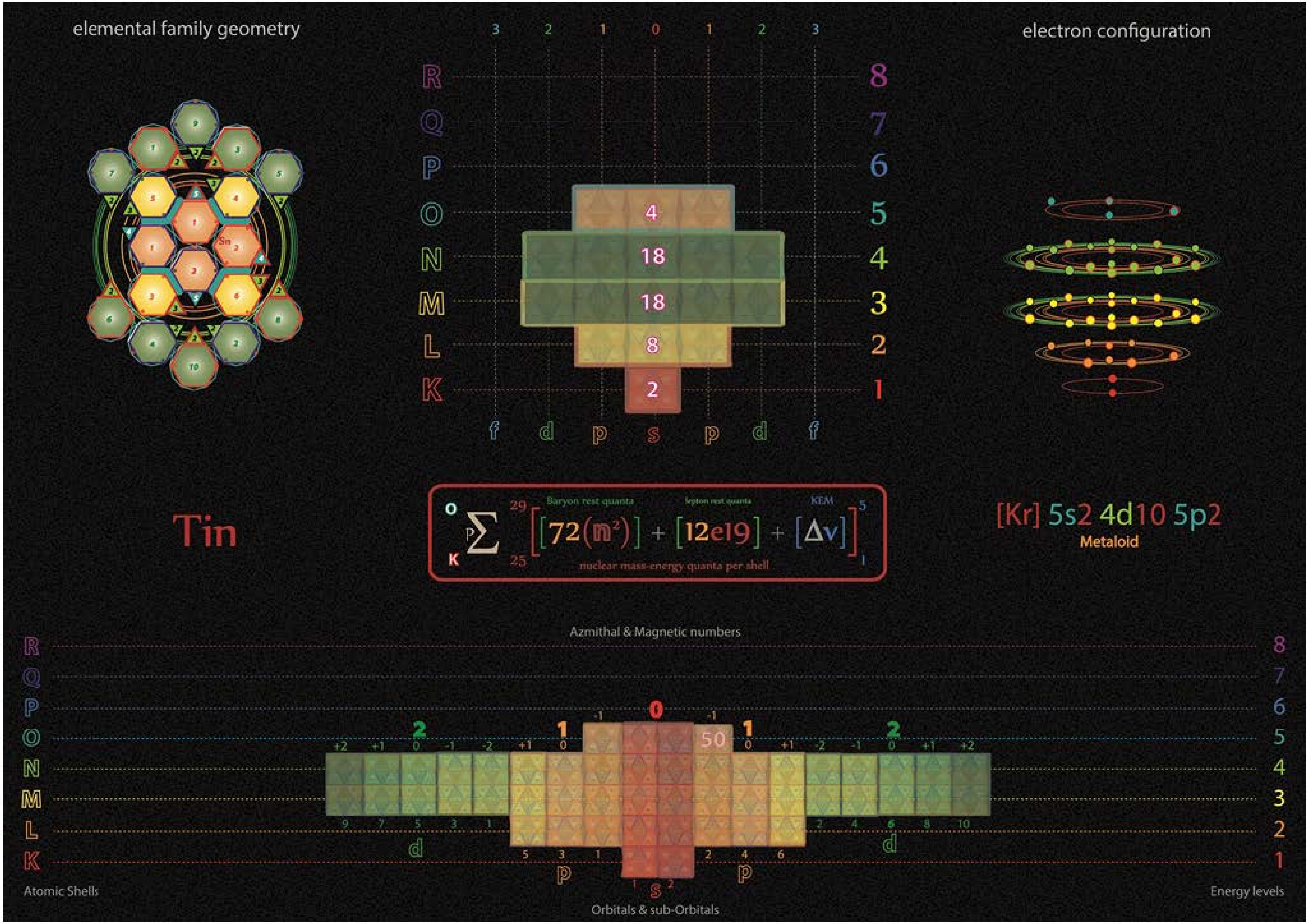

elemental family geometry

electron configuration

Tin

$$ {}^{\text{O}}_{\text{K}}{}_{p}\sum_{25}^{29}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]_{1}^{5} $$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Kr] 5s2 4d10 5p2

Metaloid

Azmithal & Magnetic numbers

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Sn
1
2
3
4
5
6
7
8
9
10
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
4
18
18
8
2
f d p s p d f
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 +1 -2 -1 0 +1 +2
50
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d
p
s
p
d

Tetryonics 53.50 - Tin atomic config

Copyright ABRAHAM [2008] - All rights reserved

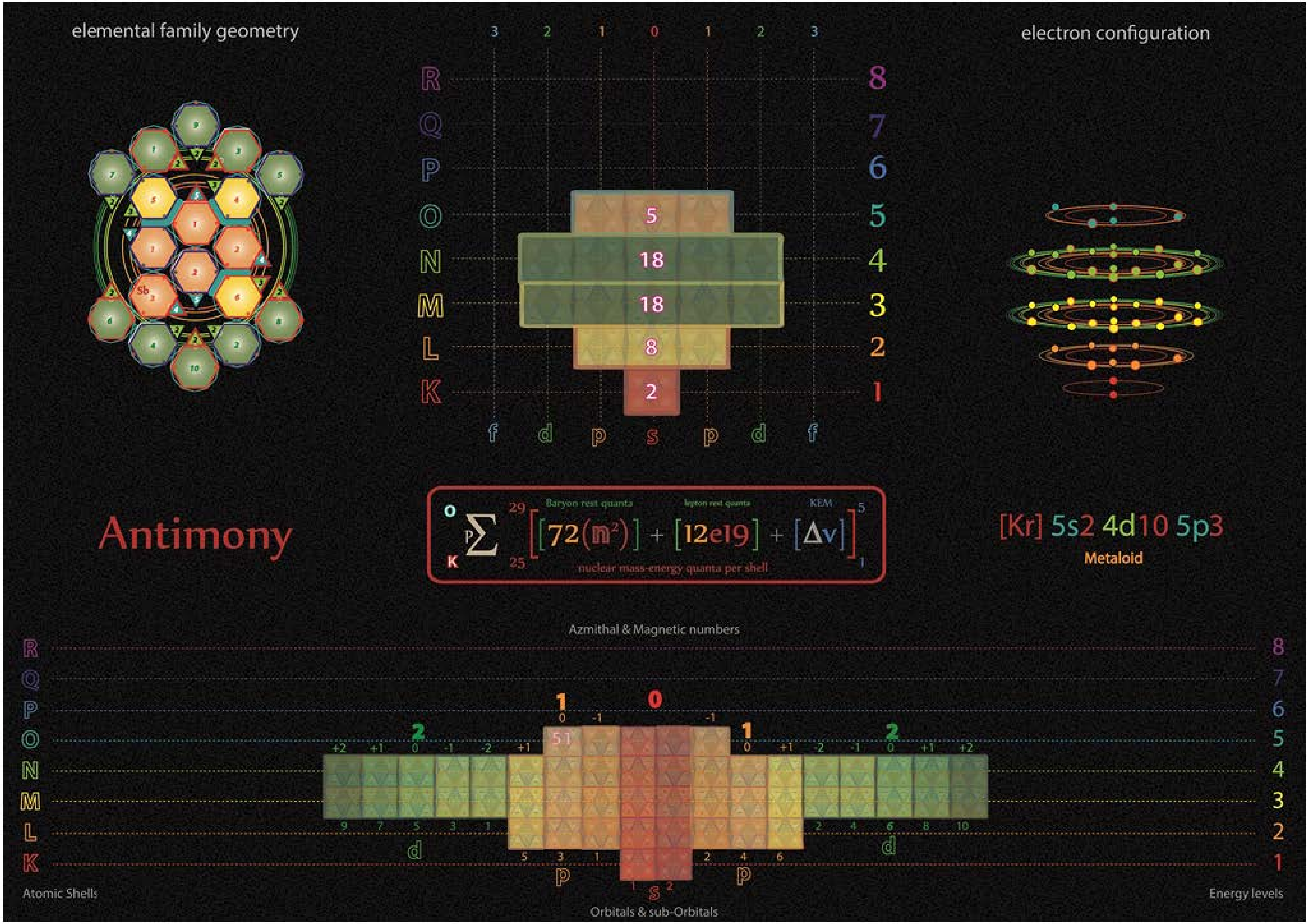
elemental family geometry
Sb
Antimony
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
18
8
2
f d p s p d f
electron configuration
[Kr] 5s2 4d10 5p3
Metaloid
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
5
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 0 -1 0 +1 -2 -1 0 +1 +2
51
9 7 5 3 1 5 3 1 1 2 2 4 6 2 4 6 8 10
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.51 - Antimony atomic config

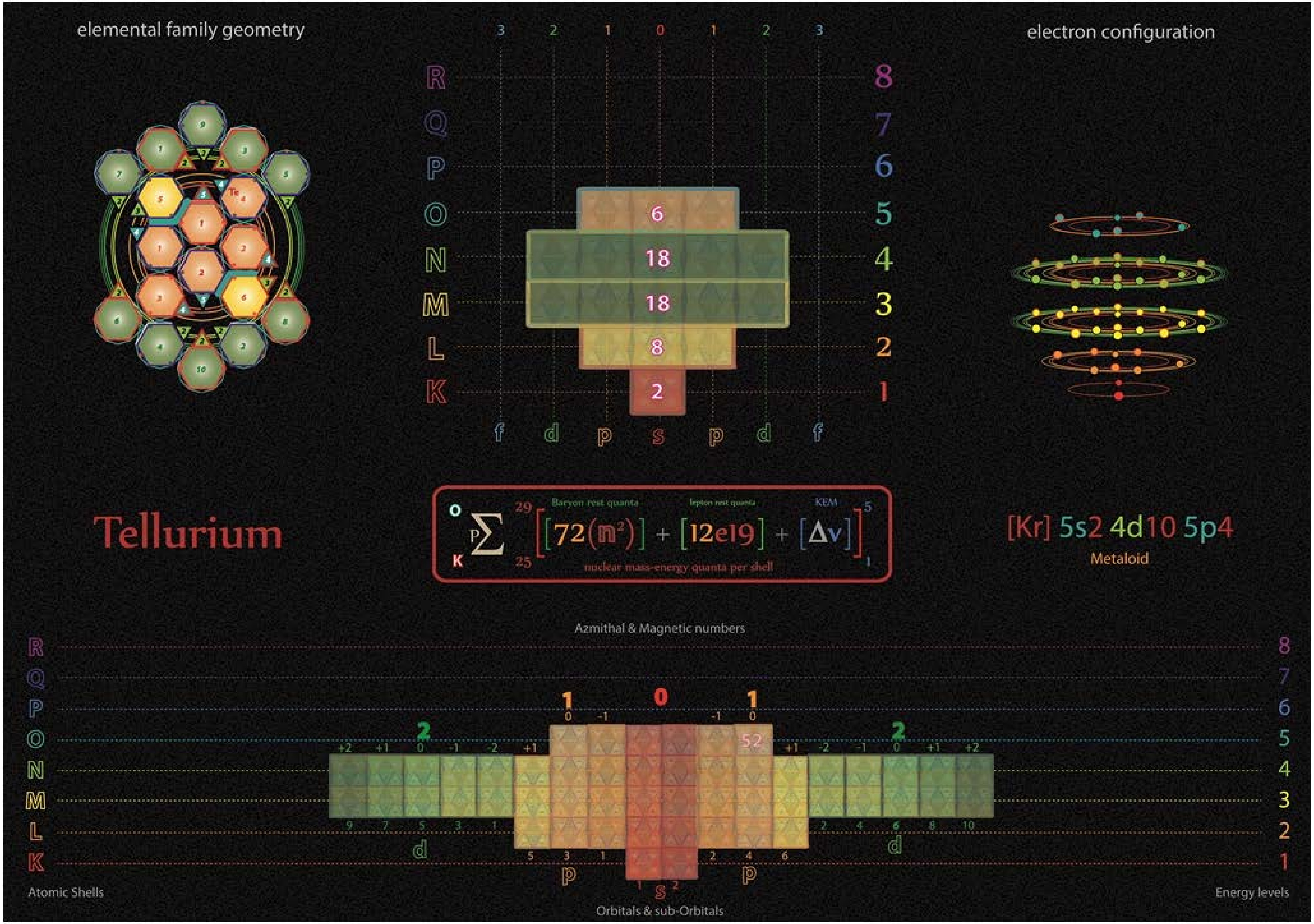
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
18
8
2
f d p s p d f
electron configuration
Tellurium
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{O} {}_{25}^{29}\left[[72(n^2)] + [12e19] + [\Delta v] \right]_1^5$
nuclear mass-energy quanta per shell
[Kr] 5s2 4d10 5p4
Metaloid
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
52
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
1 2
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.52 - Tellurium atomic config

elemental family geometry

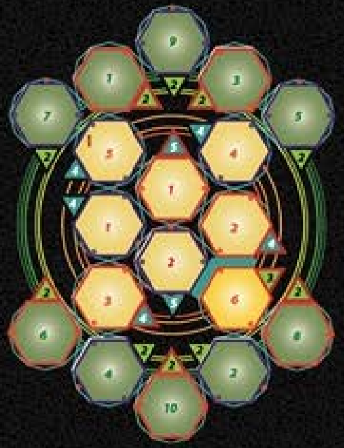

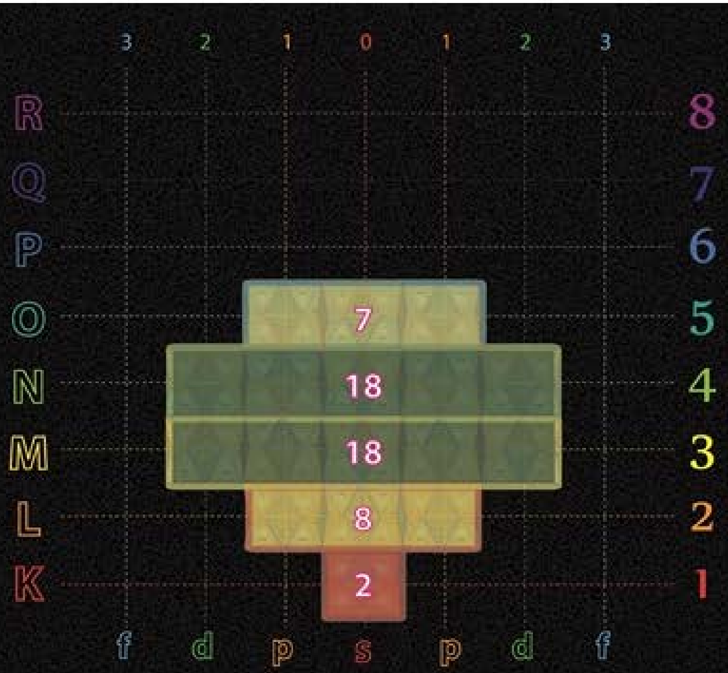
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
18
18
8
2
f
d
p
s
p
d
f

electron configuration

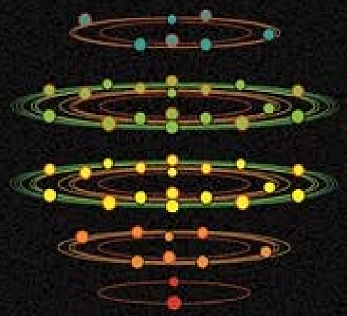

Iodine

$$\overset{O}{\underset{K}{}}\ {}^{P}\sum_{25}^{29}\underbrace{\left[\overset{\text{Baryon rest quanta}}{72(\mathrm{n}^2)}\right]+\left[\overset{\text{lepton rest quanta}}{12e19}\right]+\left[\overset{\text{KEM}}{\Delta v}\right]_1^5}_{\text{nuclear mass-energy quanta per shell}}$$

[Kr] 5s2 4d10 5p5

Halogen

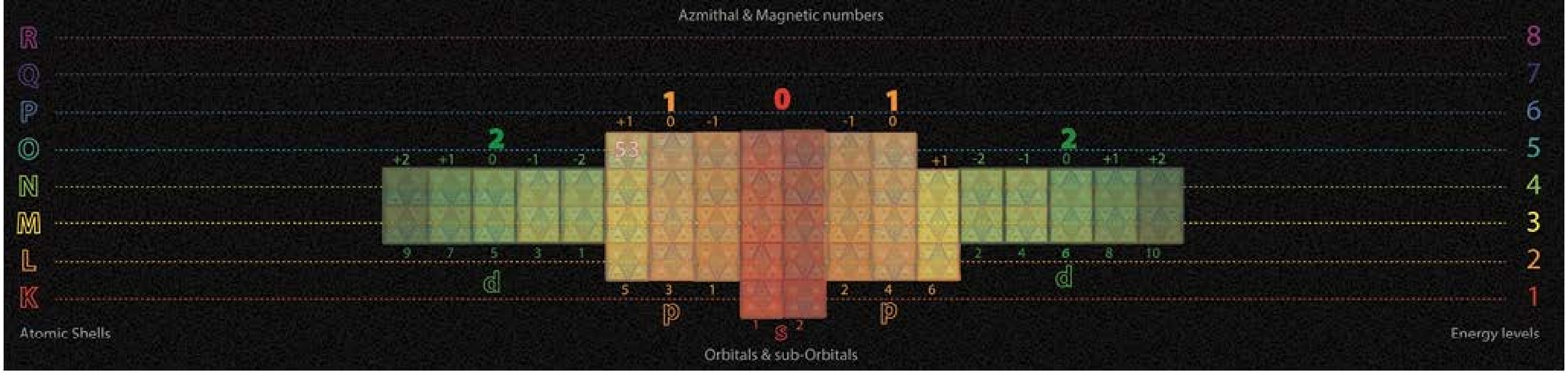
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
1
0
1
2
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
53
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
2
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.53 - Iodine atomic config

Copyright ABRAHAM [2008] - All rights reserved

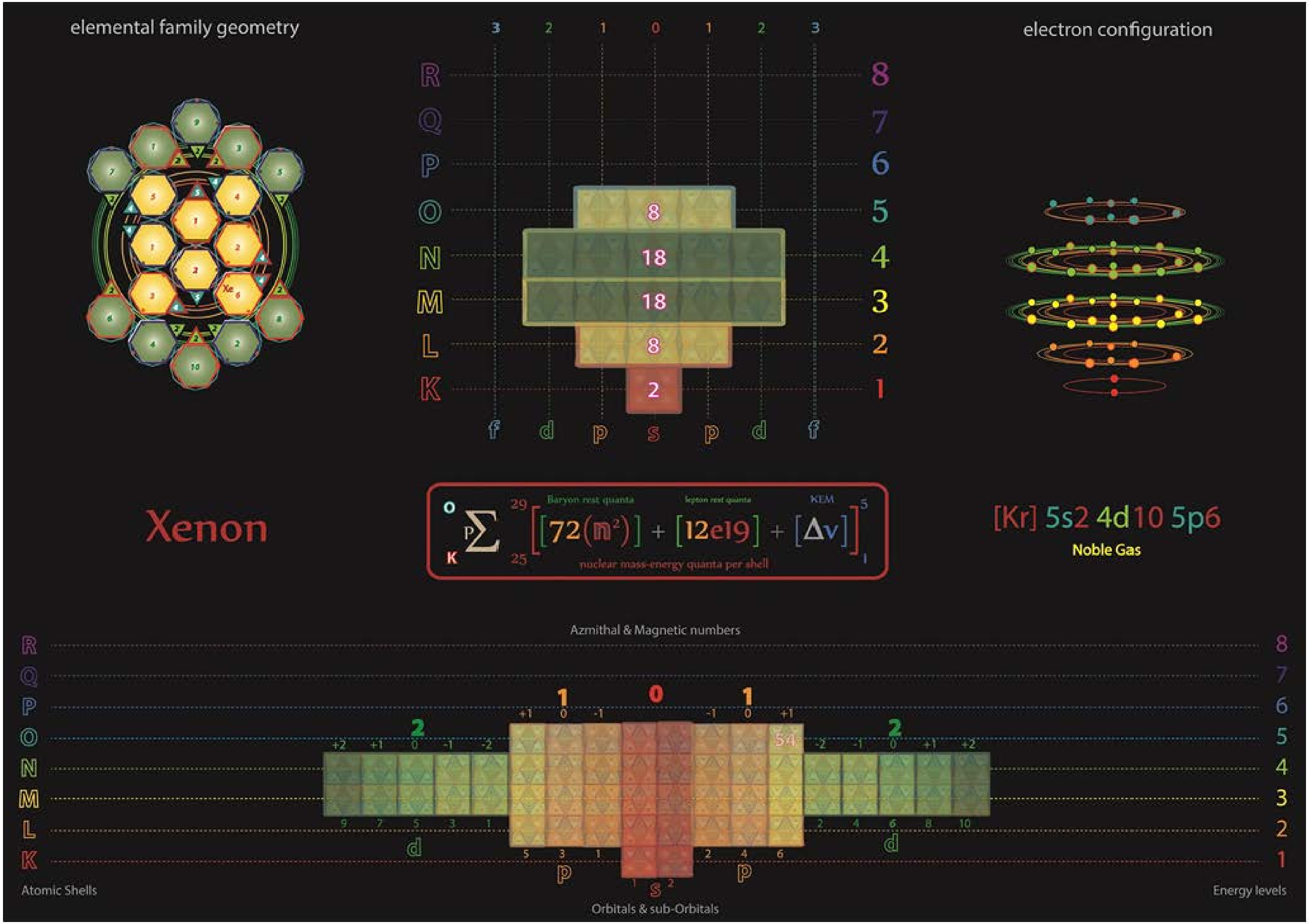
elemental family geometry
Xe
Xenon
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
8
18
18
8
2
f d p s p d f
electron configuration
[Kr] 5s2 4d10 5p6
Noble Gas
O
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
5
1
$\sum_{K}^{O} {}_{P}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{5}$
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
54
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d p s p d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.54 - Xenon atomic config

elemental family geometry

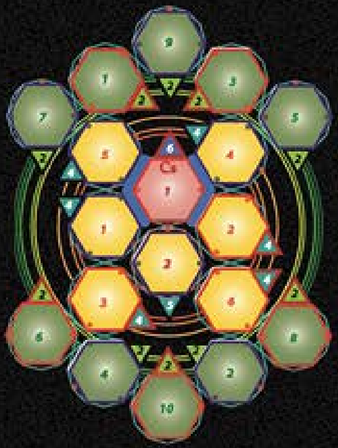
Cs
1
2
3
4
5
6
7
8
9
10

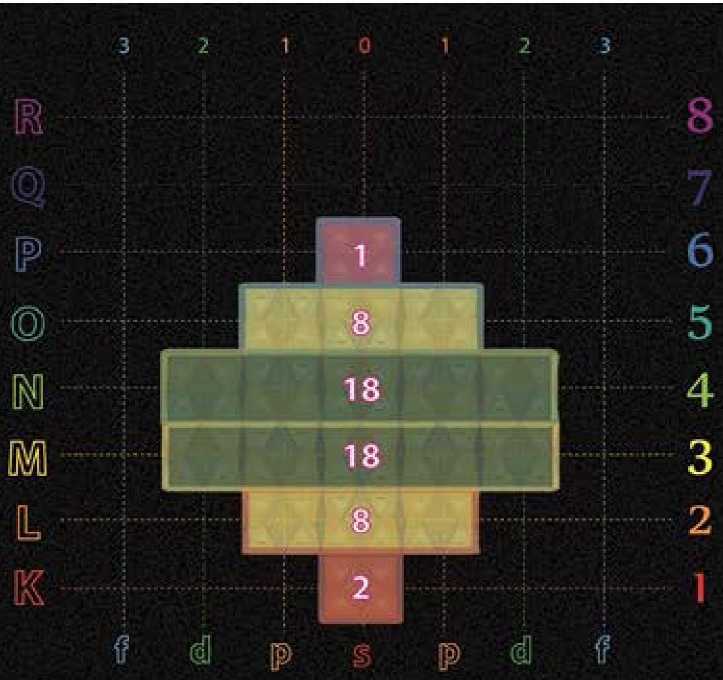
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
1
8
18
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

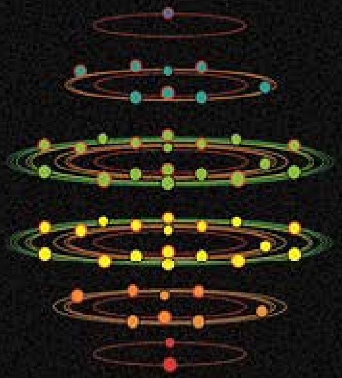

Caesium

$$ {}^{\mathrm{P}}_{\mathrm{K}}{}_{p}\sum_{25}^{29} \left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{6} $$

nuclear mass-energy quanta per shell

[Xe] 6s1

Alkali Metal

Azmithal & Magnetic numbers

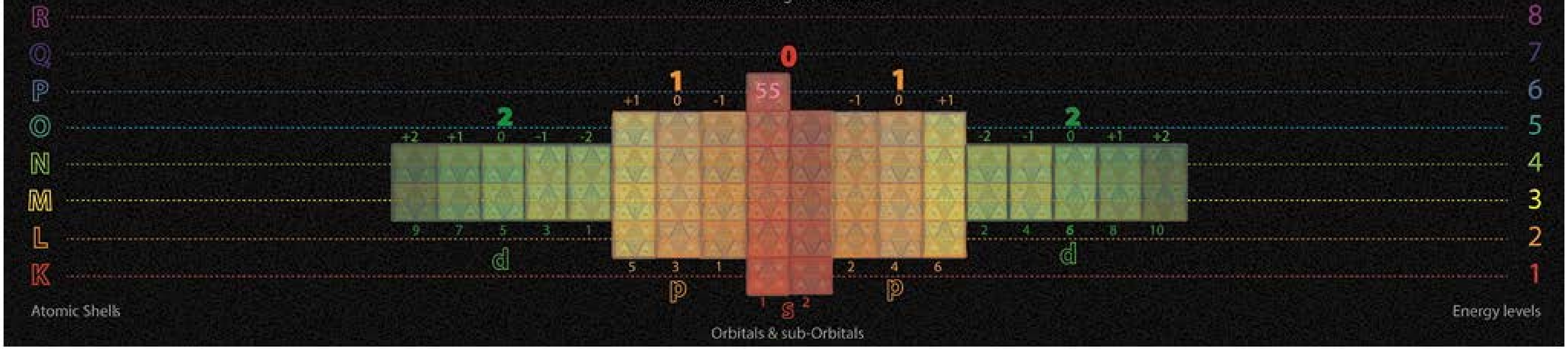
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
55
1
+1
0
-1
1
-1
0
+1
2
+2
+1
0
-1
-2
2
-2
-1
0
+1
+2
9
7
5
3
1
d
5
3
1
p
1
s
2
2
4
6
p
2
4
6
8
10
d
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Copyright ABRAHAM [2008] - All rights reserved

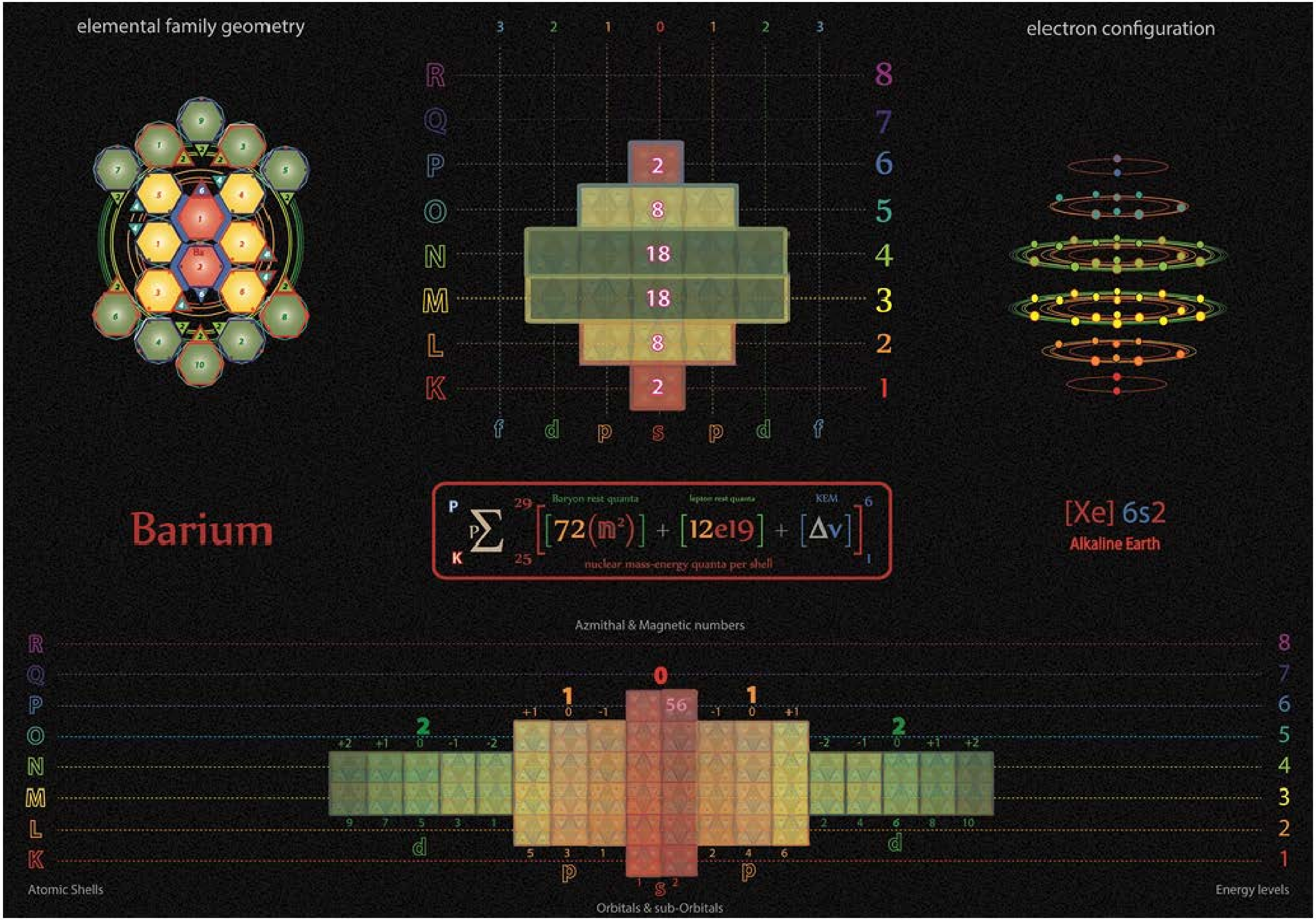
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
18
18
8
2
f d p s p d f
Barium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 6s2
Alkaline Earth
Azmithal & Magnetic numbers
0
56
1
+1 0 -1
-1 0 +1
2
+2 +1 0 -1 -2
-2 -1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.56 - Barium atomic config

Copyright ABRAHAM [2008] - All rights reserved

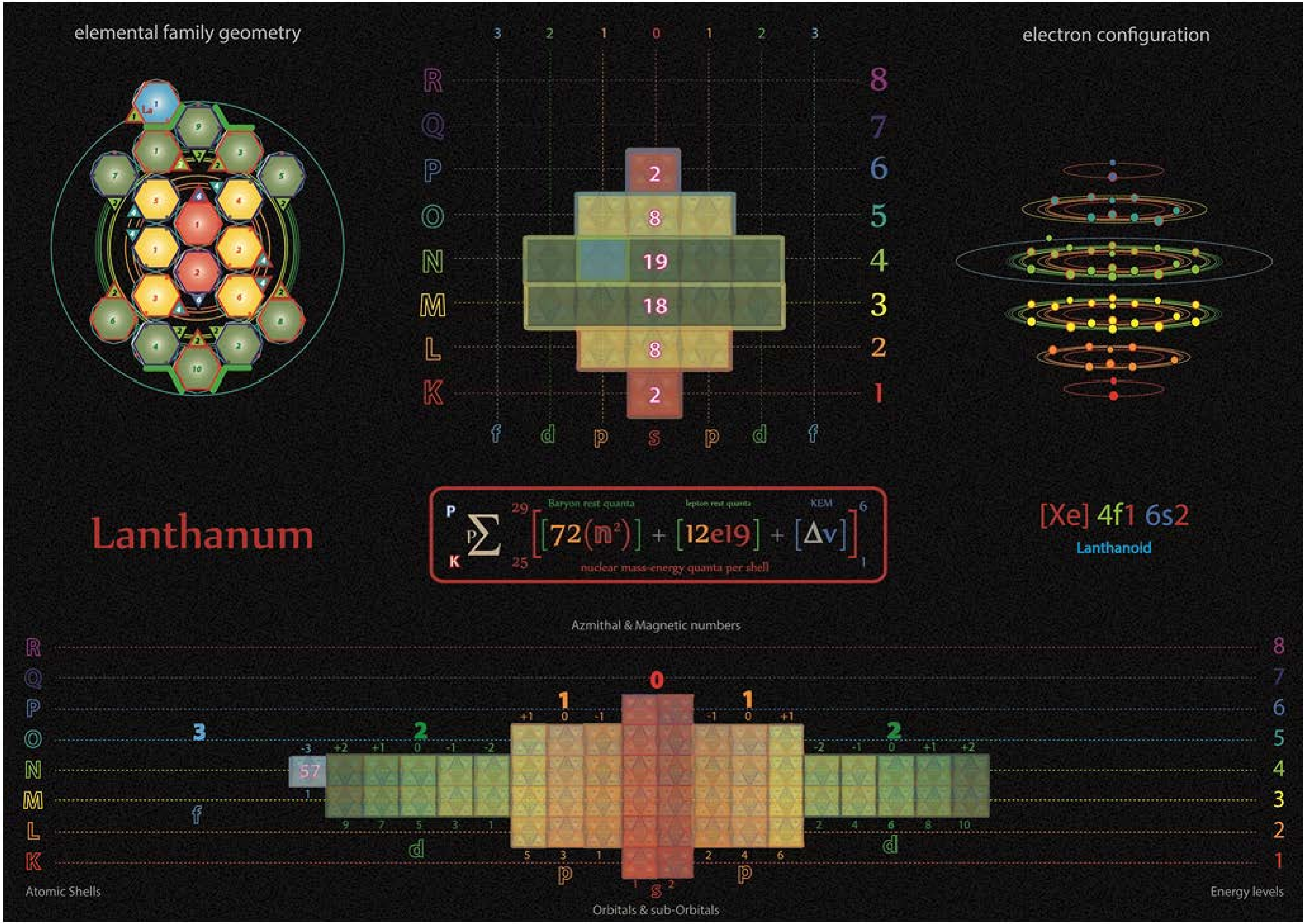
elemental family geometry
La 1
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
19
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3
-3 +2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
57
1
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Lanthanum

$$\sum_{K}^{P} {}_{P}\left[\,{}^{29}_{25}\;[72(m^2)] + [12e19] + [\Delta v]\,\right]^{6}_{1}$$

Baryon rest quanta, lepton rest quanta, KEM

nuclear mass-energy quanta per shell

[Xe] 4f1 6s2

Lanthanoid

Tetryonics 53.57 - Lanthanum atomic config

elemental family geometry

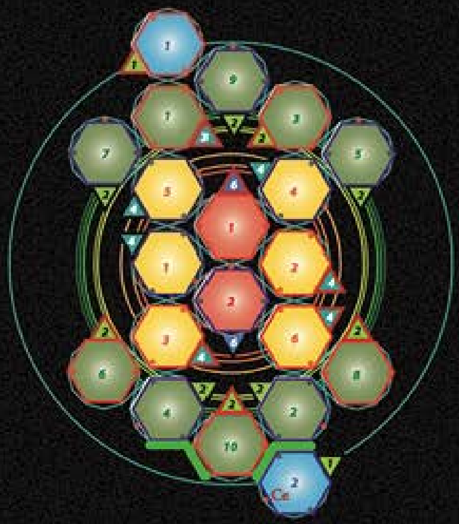

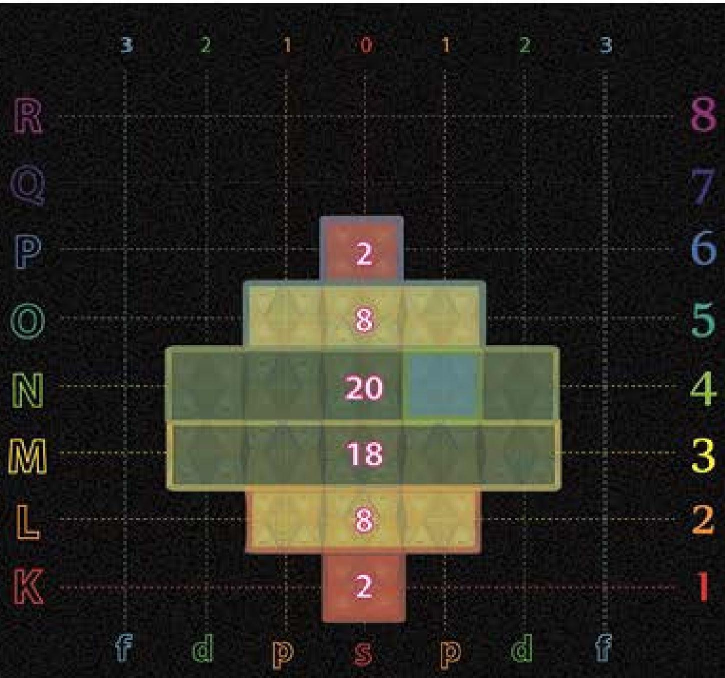
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
20
18
8
2
f d p s p d f

electron configuration

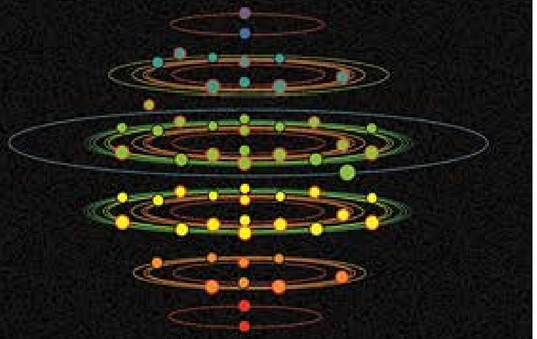

Cerium

$$\sum_{K}^{P} \, {}^{29}_{25}\left[72(m^2)\right] + \left[12e19\right] + \left[\Delta v\right]^{6}_{1}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f2 6s2

Lanthanoid

Azmithal & Magnetic numbers

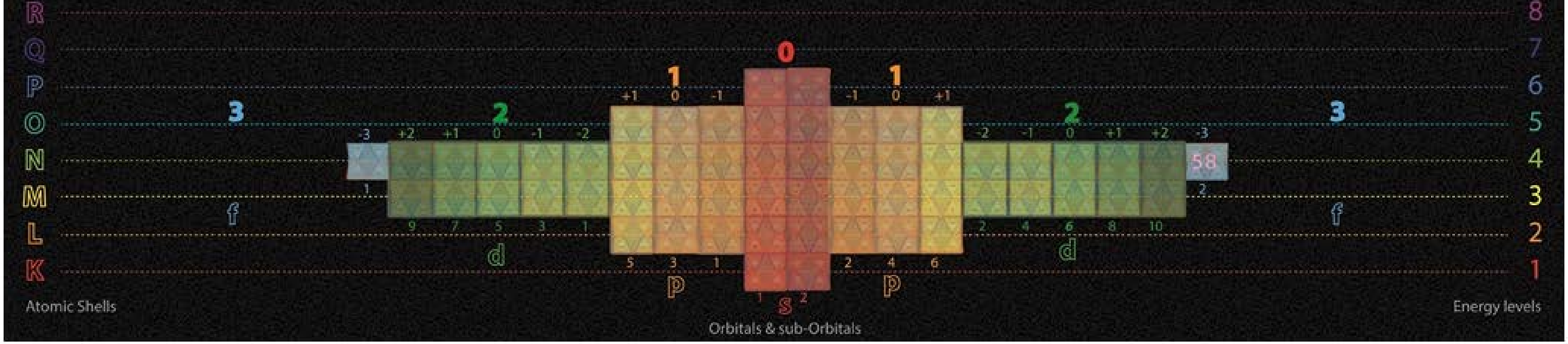
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
-3 +2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2 -3
58
1 9 7 5 3 1 5 3 1 1 2 2 4 6 2 4 6 8 10 2
f d p s p d f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.58 - Cerium atomic config

Copyright ABRAHAM [2008] - All rights reserved

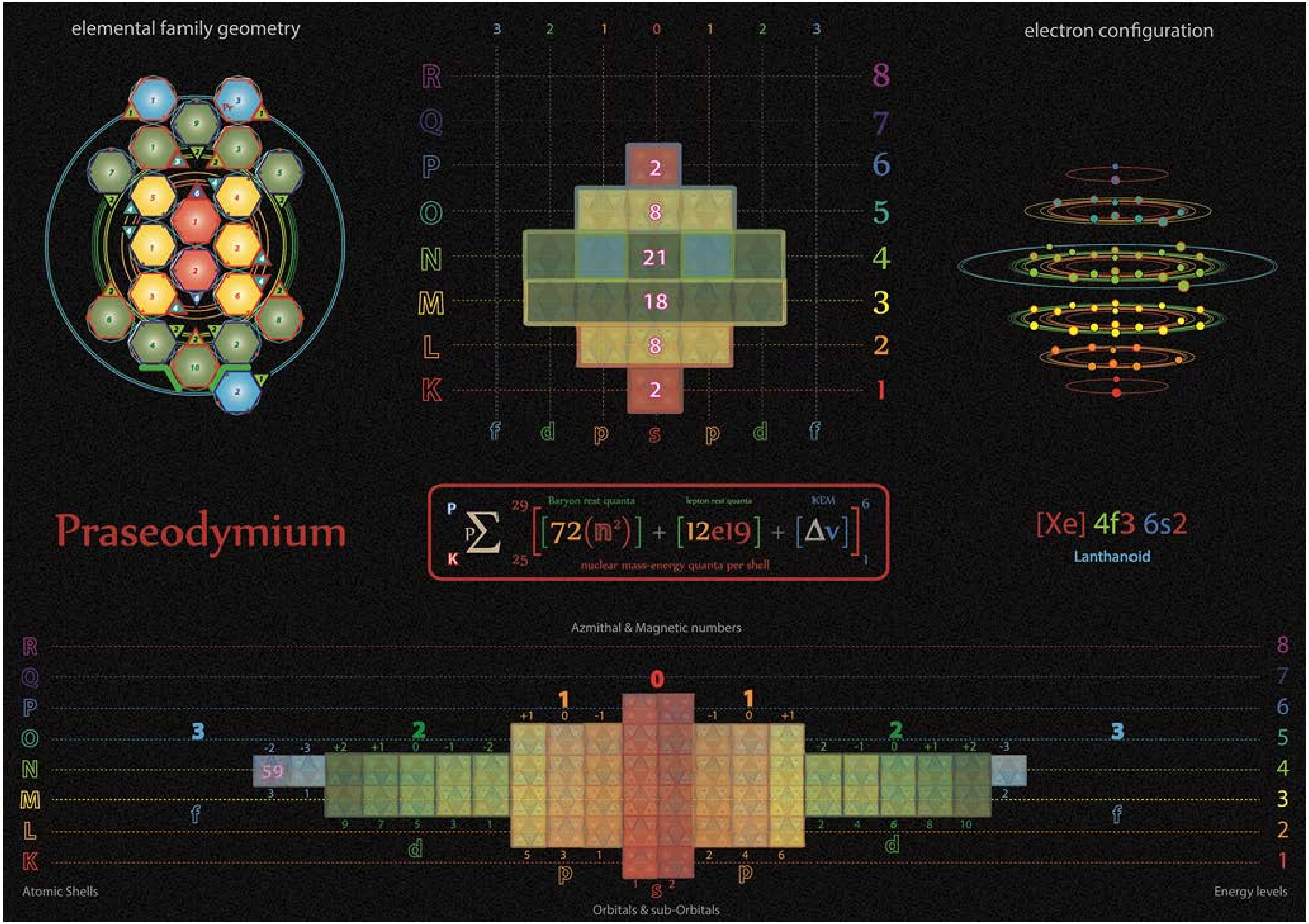
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
21
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
59
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Praseodymium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{P} {}_{P}\sum_{25}^{29} \left[[72(\mathfrak{n}^2)] + [12e19] + [\Delta v] \right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f3 6s2

Lanthanoid

Tetryonics 53.59 - Praseodymium atomic config

Copyright ABRAHAM [2008] - All rights reserved

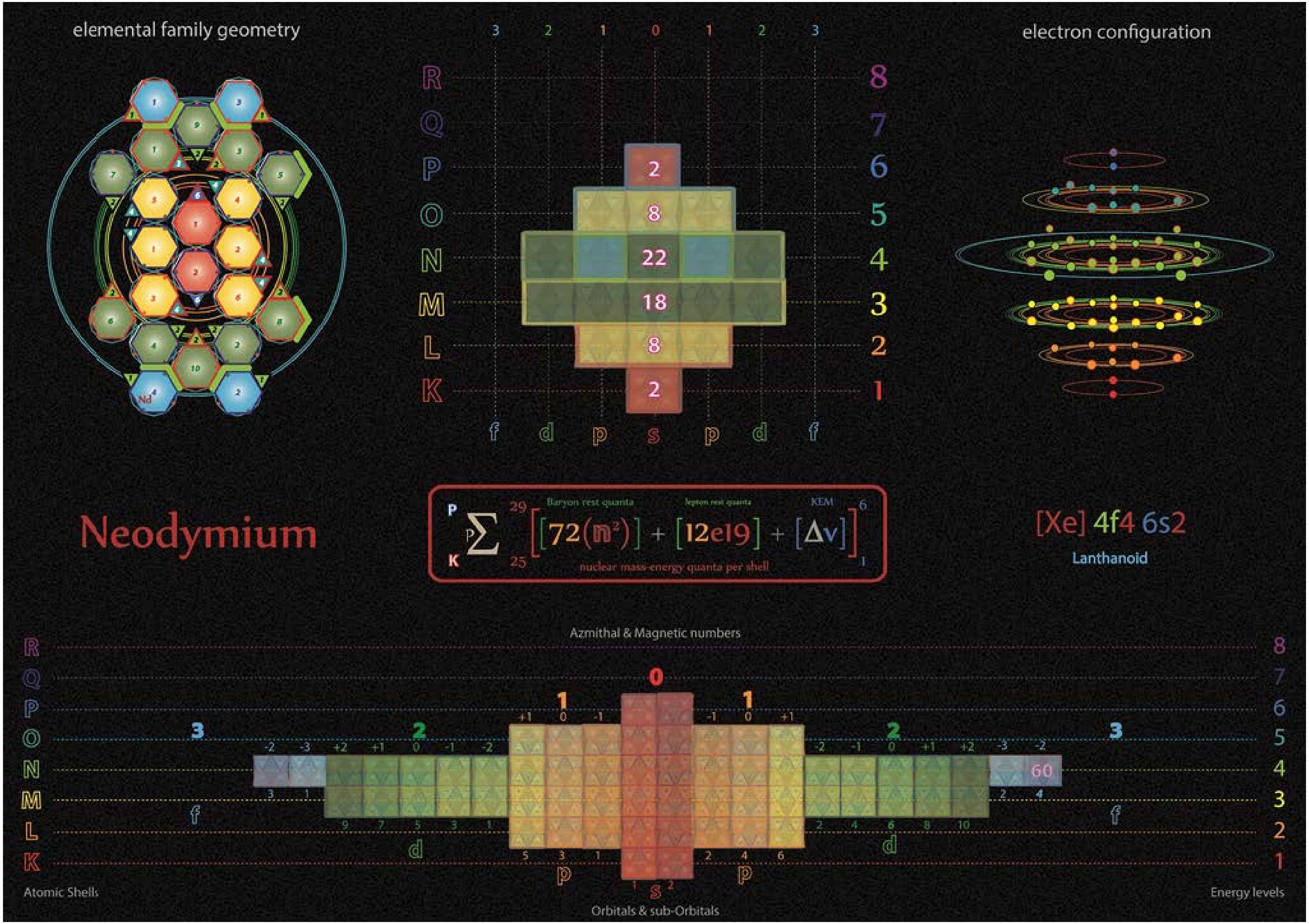
elemental family geometry
Nd
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
22
18
8
2
f d p s p d f
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1
-1 0 +1
-2 -3 +2 +1 0 -1 -2
-2 -1 0 +1 +2 -3 -2
60
3 1
2 4
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Neodymium

$$\sum_{K}^{P} {}_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{6}$$

[Xe] 4f4 6s2

Lanthanoid

Tetryonics 53.60 - Neodymium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

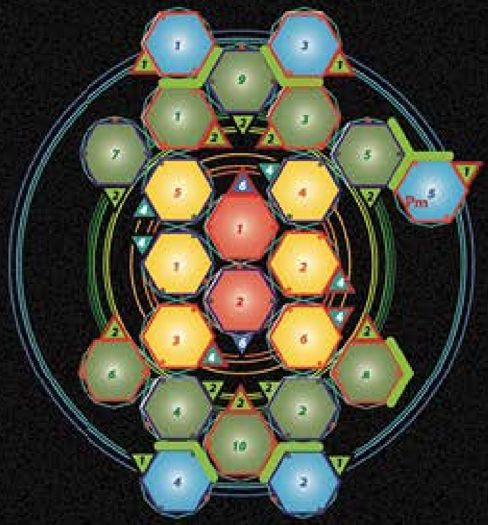

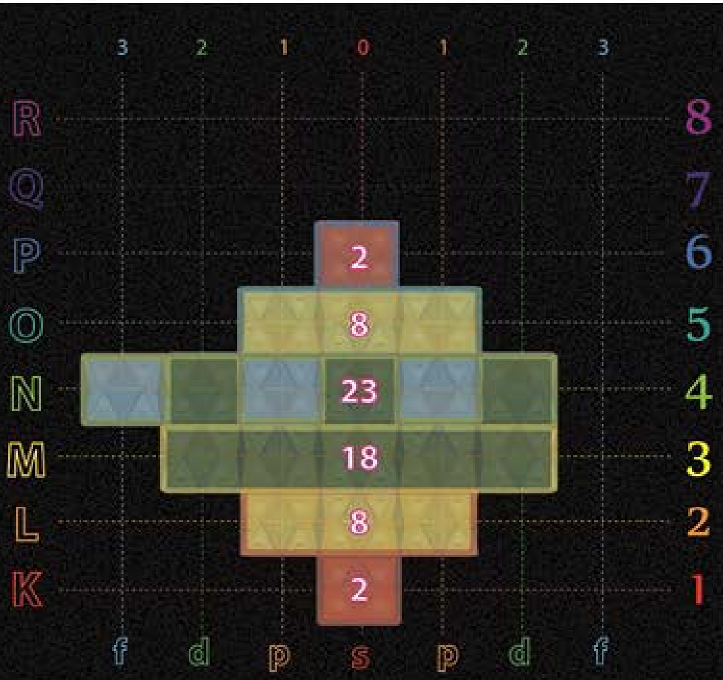
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
23
18
8
2
f
d
p
s
p
d
f

electron configuration

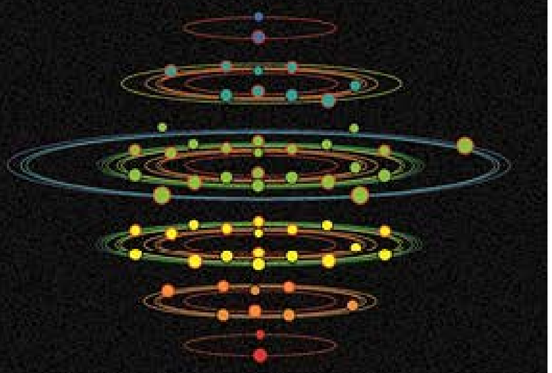

Promethium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{P} {}^{29}_{25}\left[72(m^2)\right] + \left[12e19\right] + \left[\Delta v\right]^{6}_{1}$$

nuclear mass-energy quanta per shell

[Xe] 4f5 6s2

Lanthanoid

Azmithal & Magnetic numbers

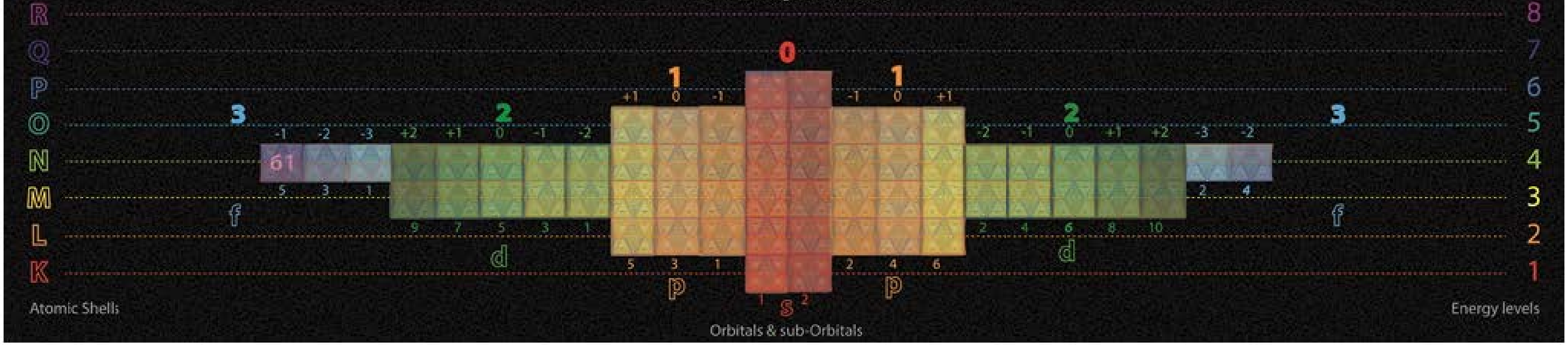
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
61
f
d
p
s
p
d
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.61 - Promethium atomic config

Copyright ABRAHAM [2008] - All rights reserved

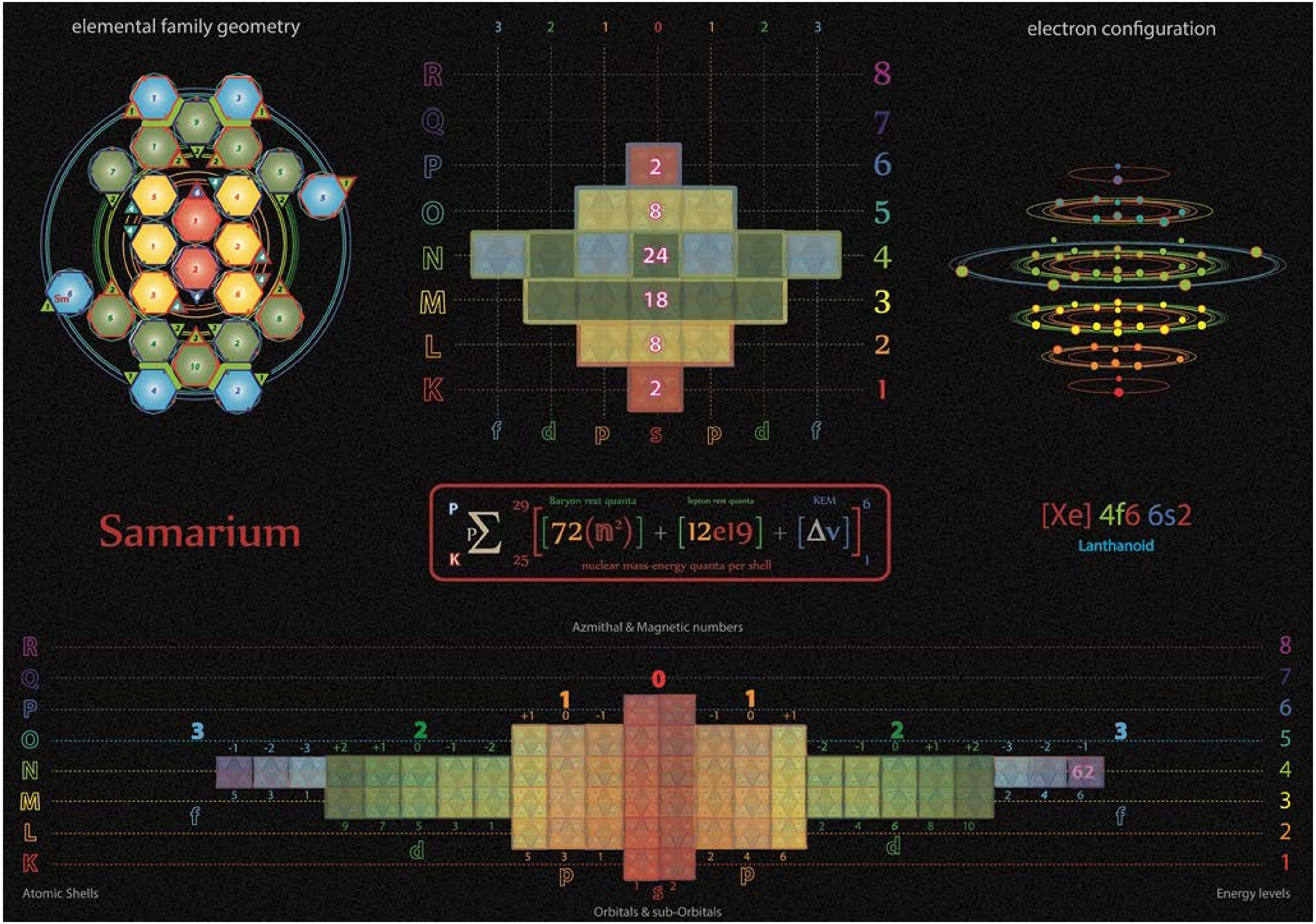
elemental family geometry
Sm
Samarium
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
8
24
18
8
2
f d p s p d f
electron configuration
[Xe] 4f6 6s2
Lanthanoid
P K
Baryon rest quanta
lepton rest quanta
KEM
29 25 6 1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
-1 -2 -3 +2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2 -3 -2 -1
62
5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6
f d p p d f
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Tetryonics 53.62 - Samarium atomic config

Copyright ABRAHAM [2008] - All rights reserved

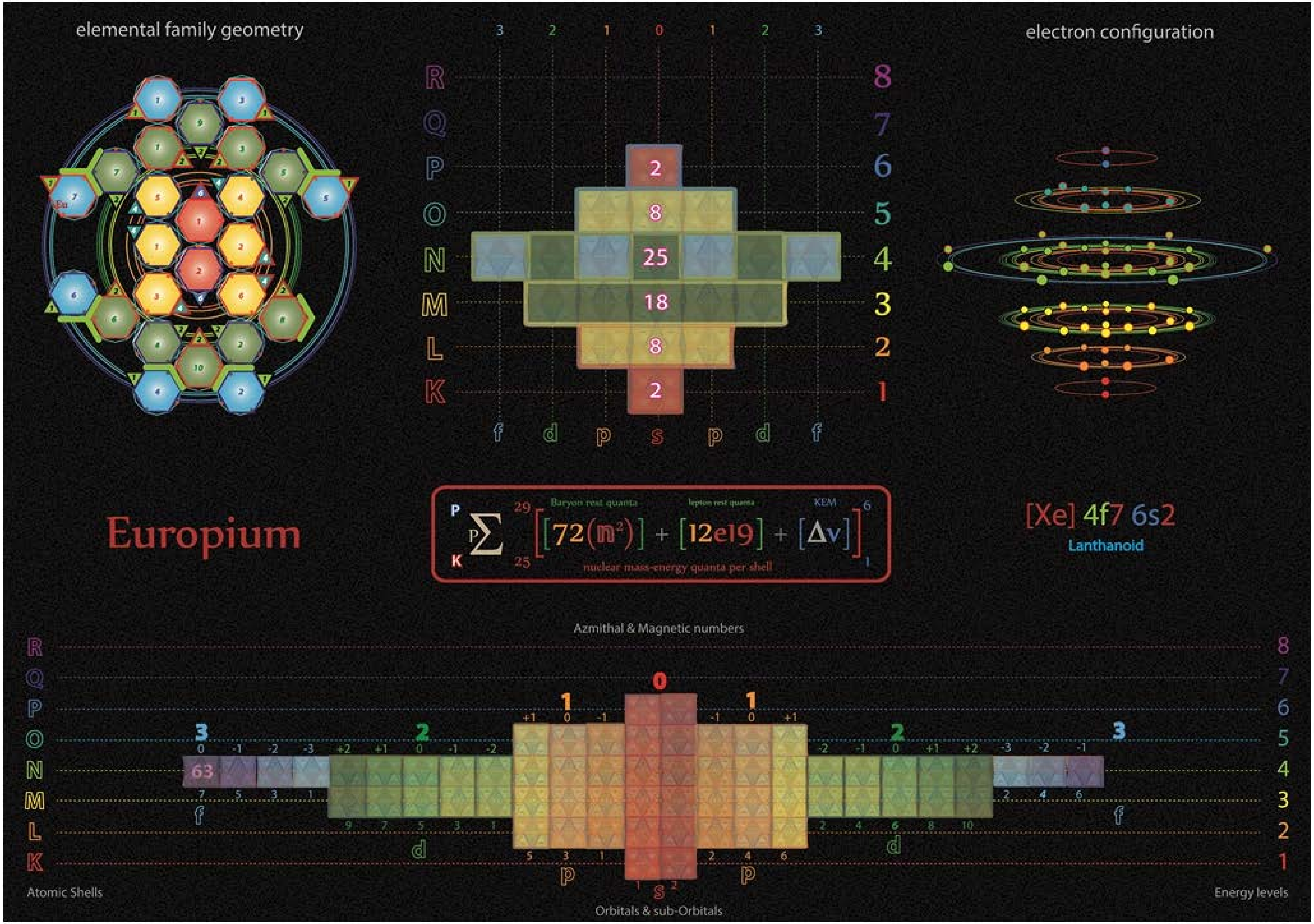
elemental family geometry
Eu
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
25
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
63
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Europium

$$\sum_{K}^{P} {}_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{6}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f7 6s2

Lanthanoid

Tetryonics 53.63 - Europium atomic config

Copyright ABRAHAM [2008] - All rights reserved

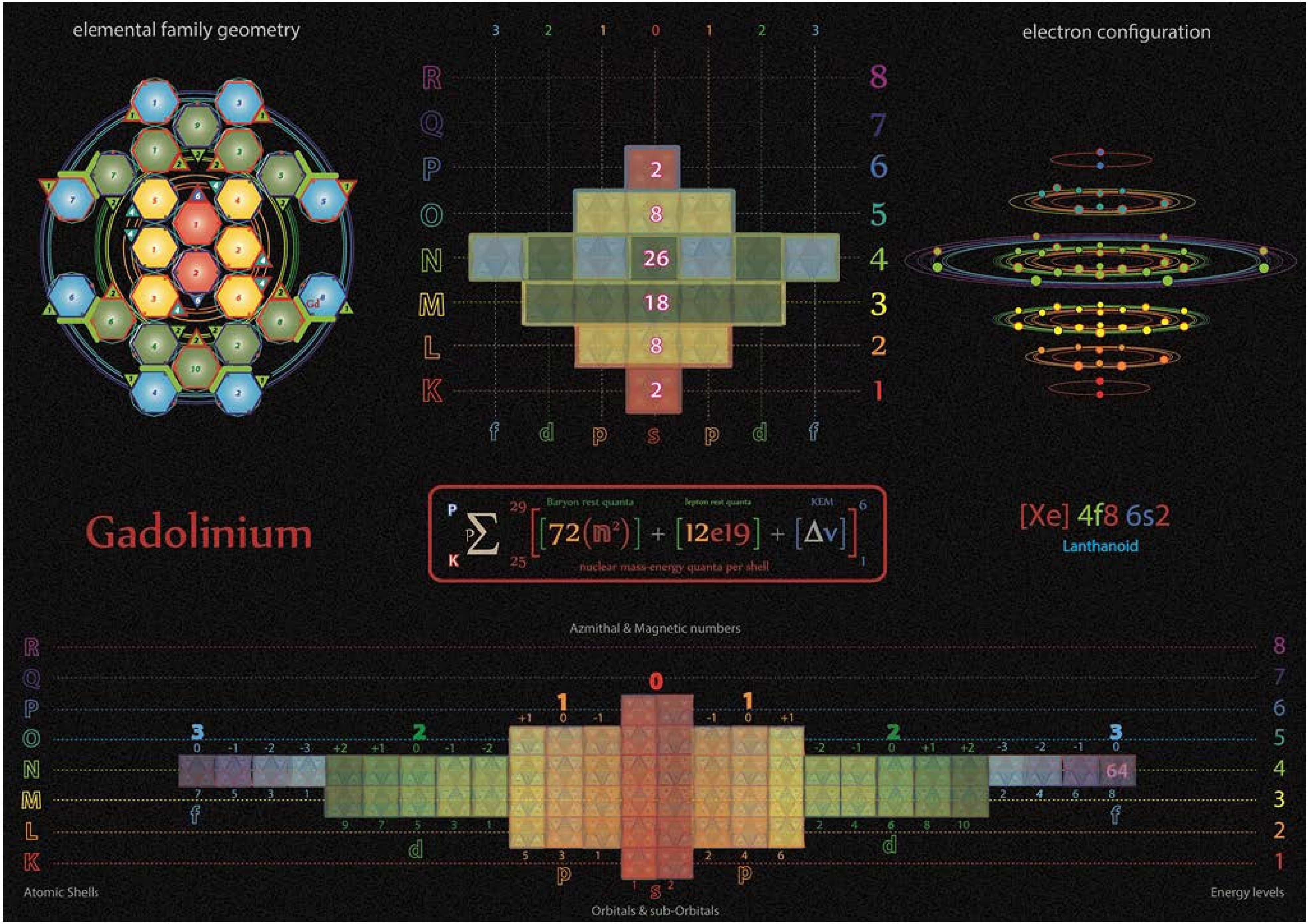
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
26
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
64
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Gadolinium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{P} {}_{p}\sum_{25}^{29}\left[\,[72(m^2)] + [12e19] + [\Delta v]\,\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f8 6s2

Lanthanoid

Tetryonics 53.64 - Gadolinium atomic config

Copyright ABRAHAM [2008] - All rights reserved

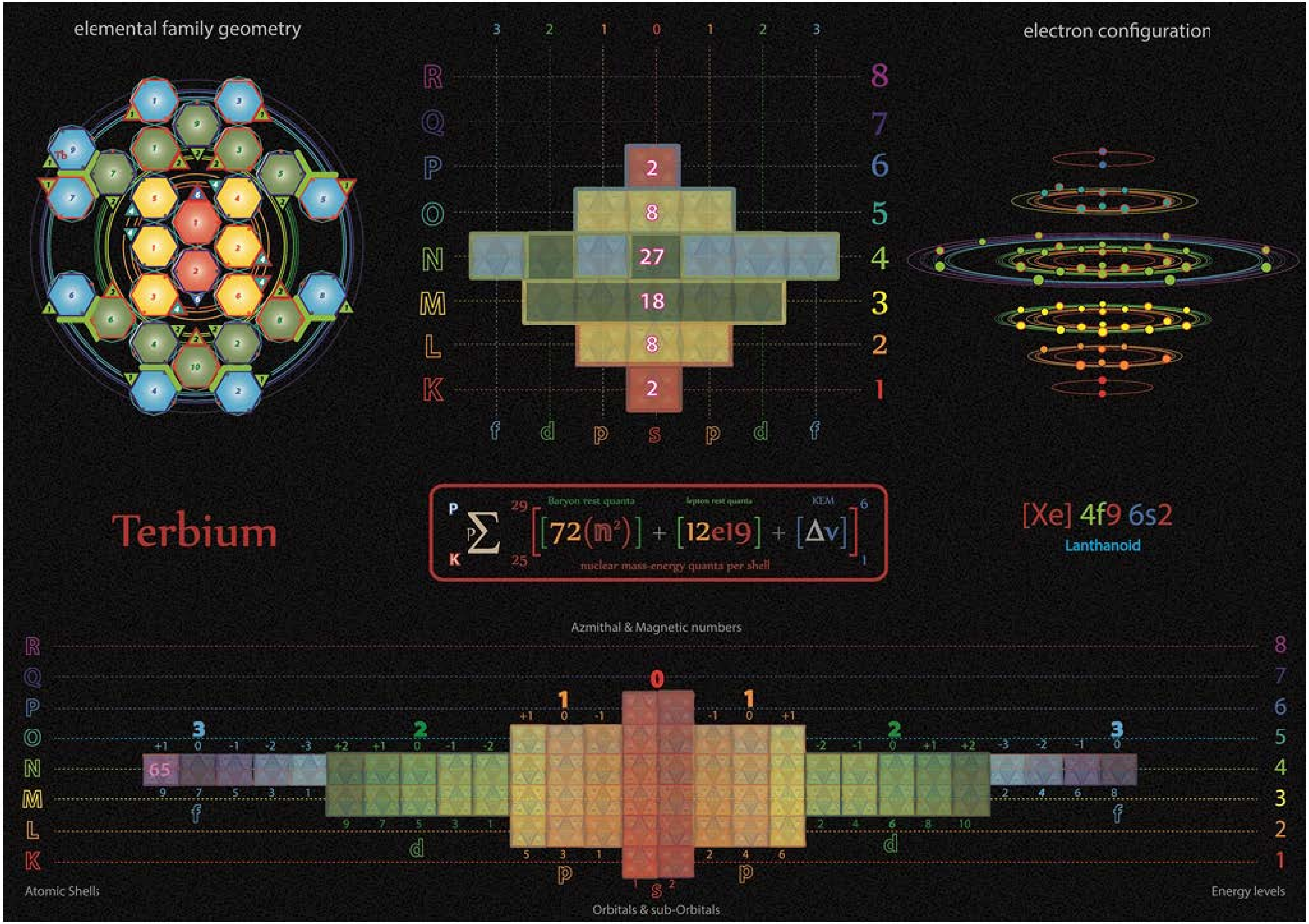
elemental family geometry
Tb
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
27
18
8
2
f d p s p d f
Terbium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f9 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
65
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.65 - Terbium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

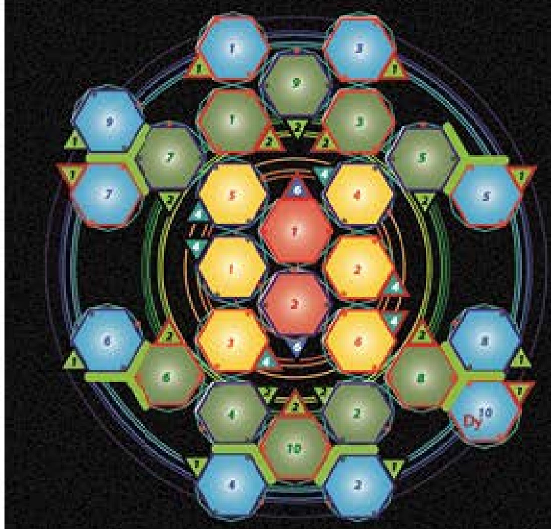

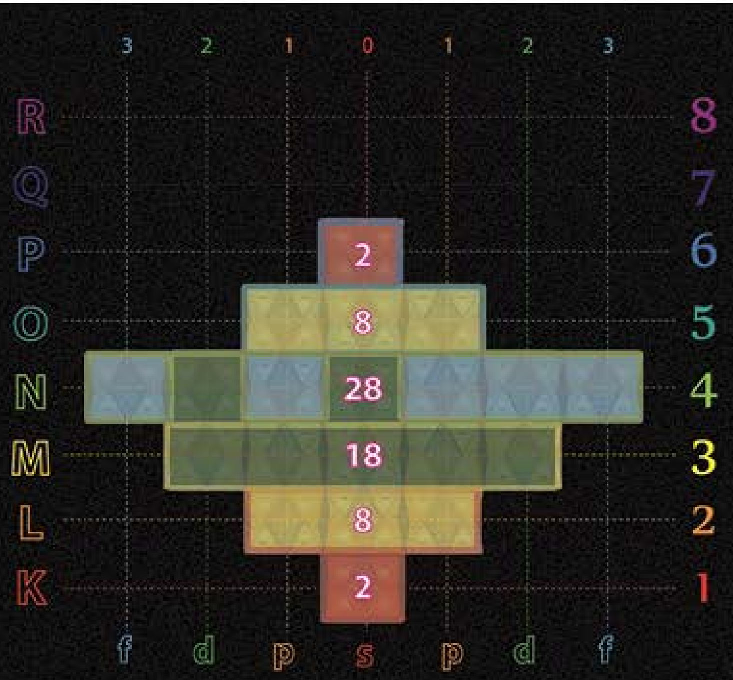
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
28
18
8
2
f d p s p d f

electron configuration

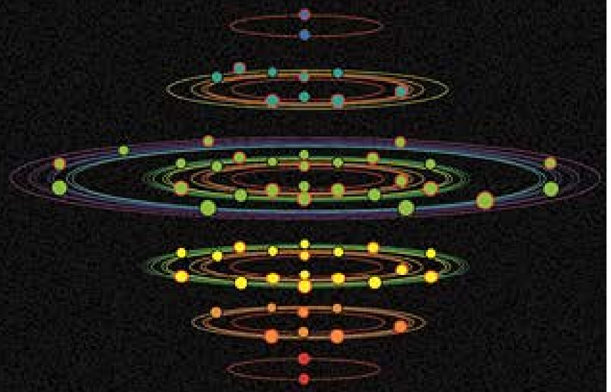

Dysprosium

$$\sum_{K}^{P} {}_{25}^{29}\left[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\right]_{1}^{6}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f10 6s2

Lanthanoid

Azmithal & Magnetic numbers

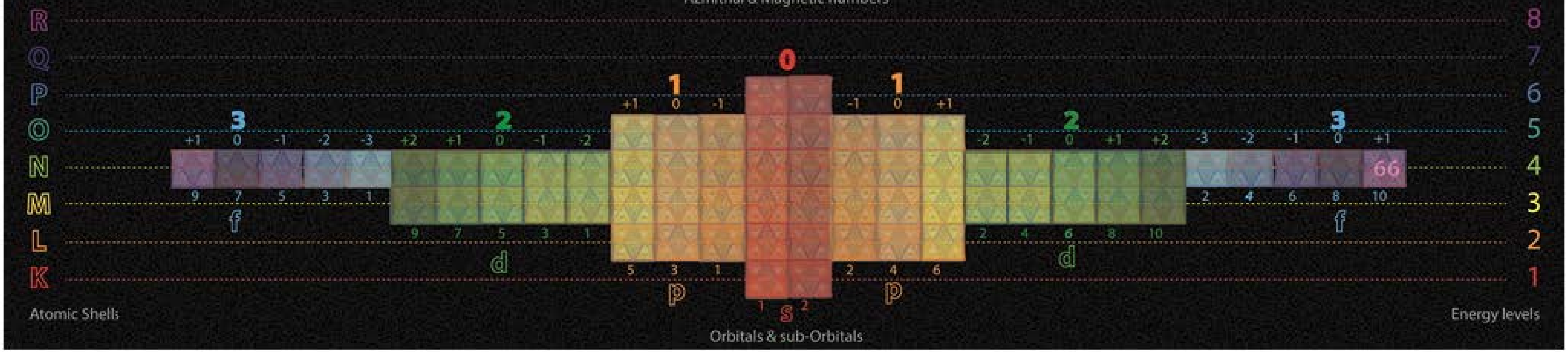
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
2 2
3 3
+1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1
66
9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.66 - Dysprosium atomic config

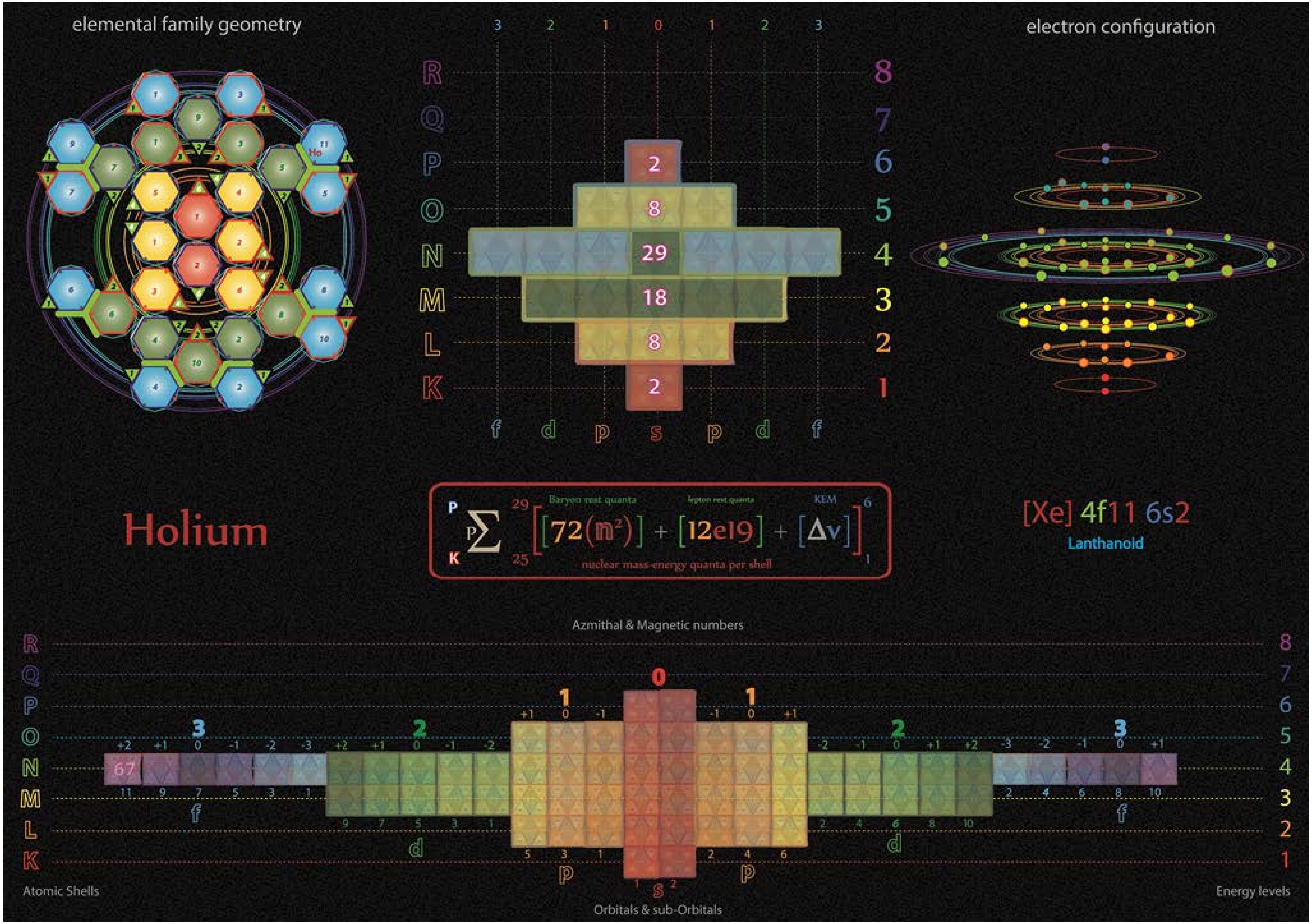
elemental family geometry
Ho
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
29
18
8
2
f d p s p d f
electron configuration
Holium
P K 29 25 Baryon rest quanta lepton rest quanta KEM 6 1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f11 6s2
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
3 2 2 3
+2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1 0 +1
67
11 9 7 5 3 1 2 4 6 8 10
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.67 - Holium atomic config

Copyright ABRAHAM [2008] - All rights reserved

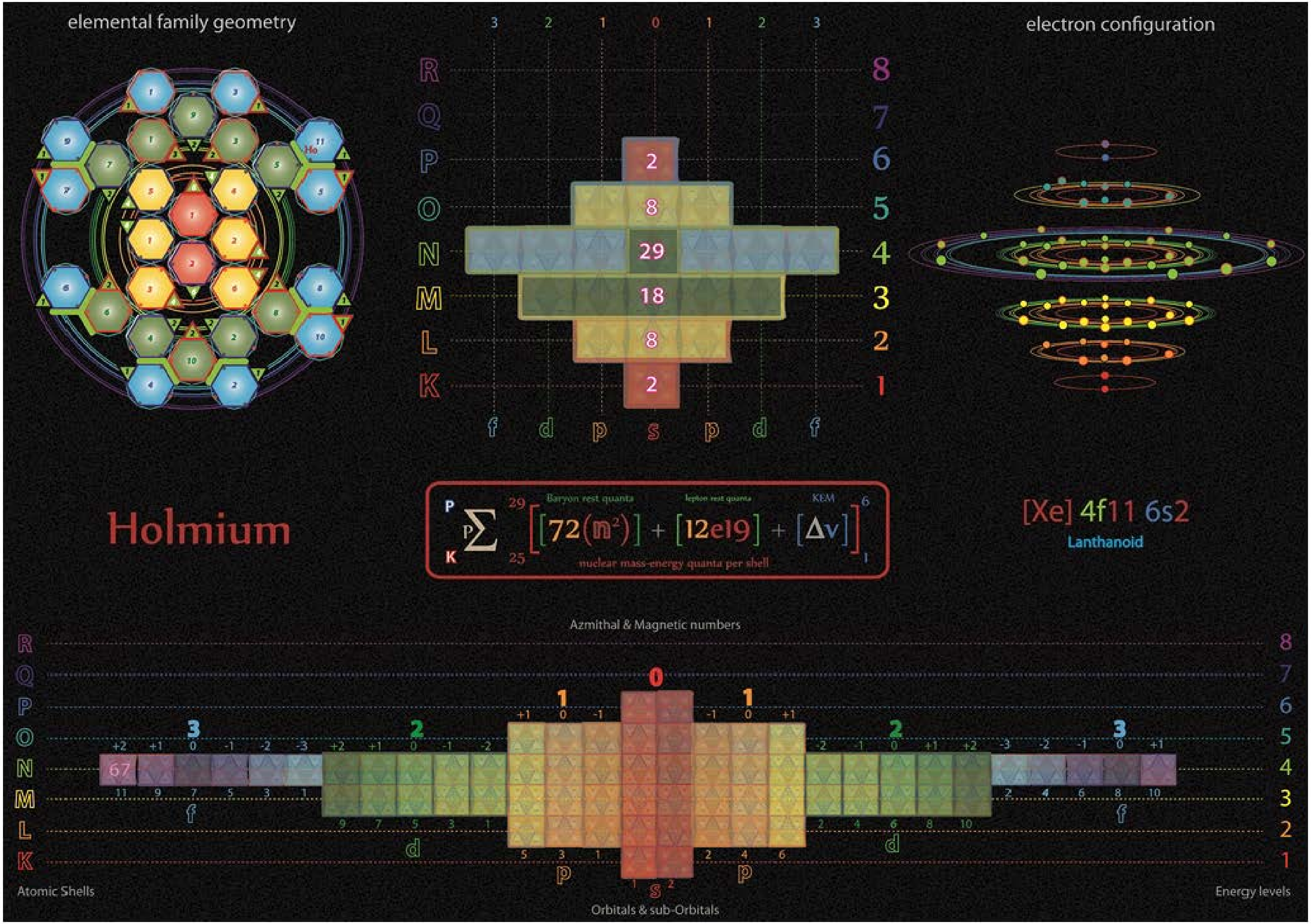

Holmium

$$\sum_{K}^{P} {}_{P}\!\!\left[\,[72(n^2)] + [12e19] + [\Delta v]\,\right]_{25\;1}^{29\;6}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f11 6s2

Lanthanoid

elemental family geometry
Ho
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
29
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 -2 -1 0 +1 +2 -3 -2 -1 0 +1
67
11 9 7 5 3 1
2 4 6 8 10
9 7 5 3 1
2 4 6 8 10
5 3 1 2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.67 - Holmium atomic config

elemental family geometry

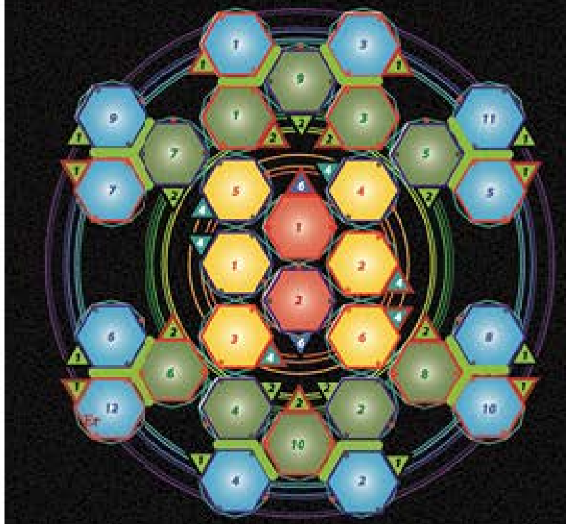

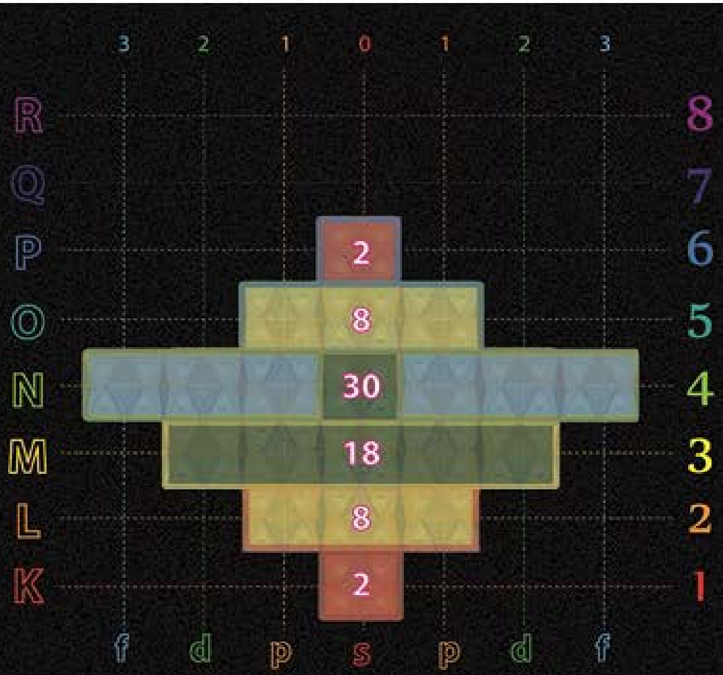
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
30
18
8
2
f
d
p
s
p
d
f

electron configuration

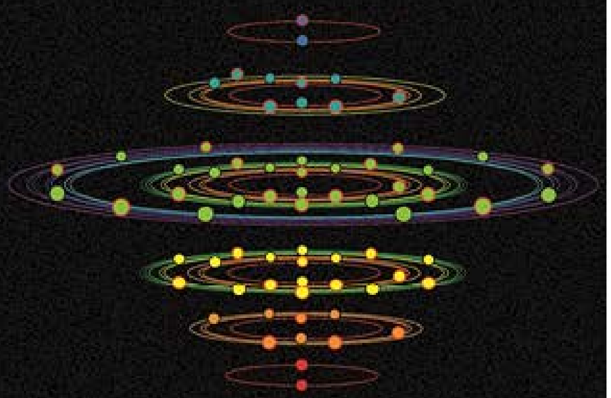

Erbium

$$\sum_{K}^{P}{}_{p}\sum_{25}^{29}\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]}+\overset{\text{lepton rest quanta}}{[12e19]}+\overset{\text{KEM}}{[\Delta v]}\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f12 6s2

Lanthanoid

Azmithal & Magnetic numbers

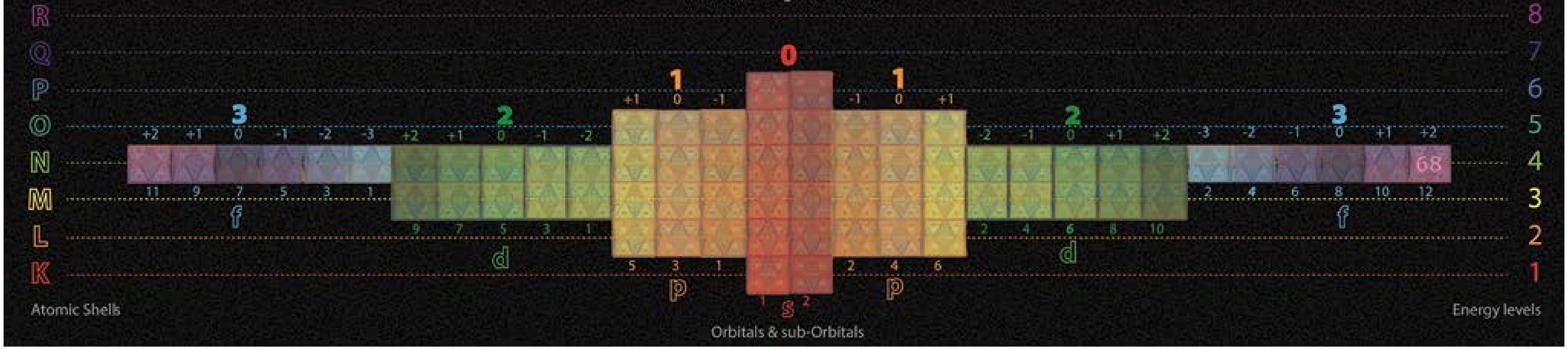
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
+2 +1 0 -1 -2 -3
11 9 7 5 3 1
f
2
+2 +1 0 -1 -2
9 7 5 3 1
d
1
+1 0 -1
5 3 1
p
0
1 2
s
1
-1 0 +1
2 4 6
p
2
-2 -1 0 +1 +2
2 4 6 8 10
d
3
-3 -2 -1 0 +1 +2
2 4 6 8 10 12
68
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.68 - Erbium atomic config

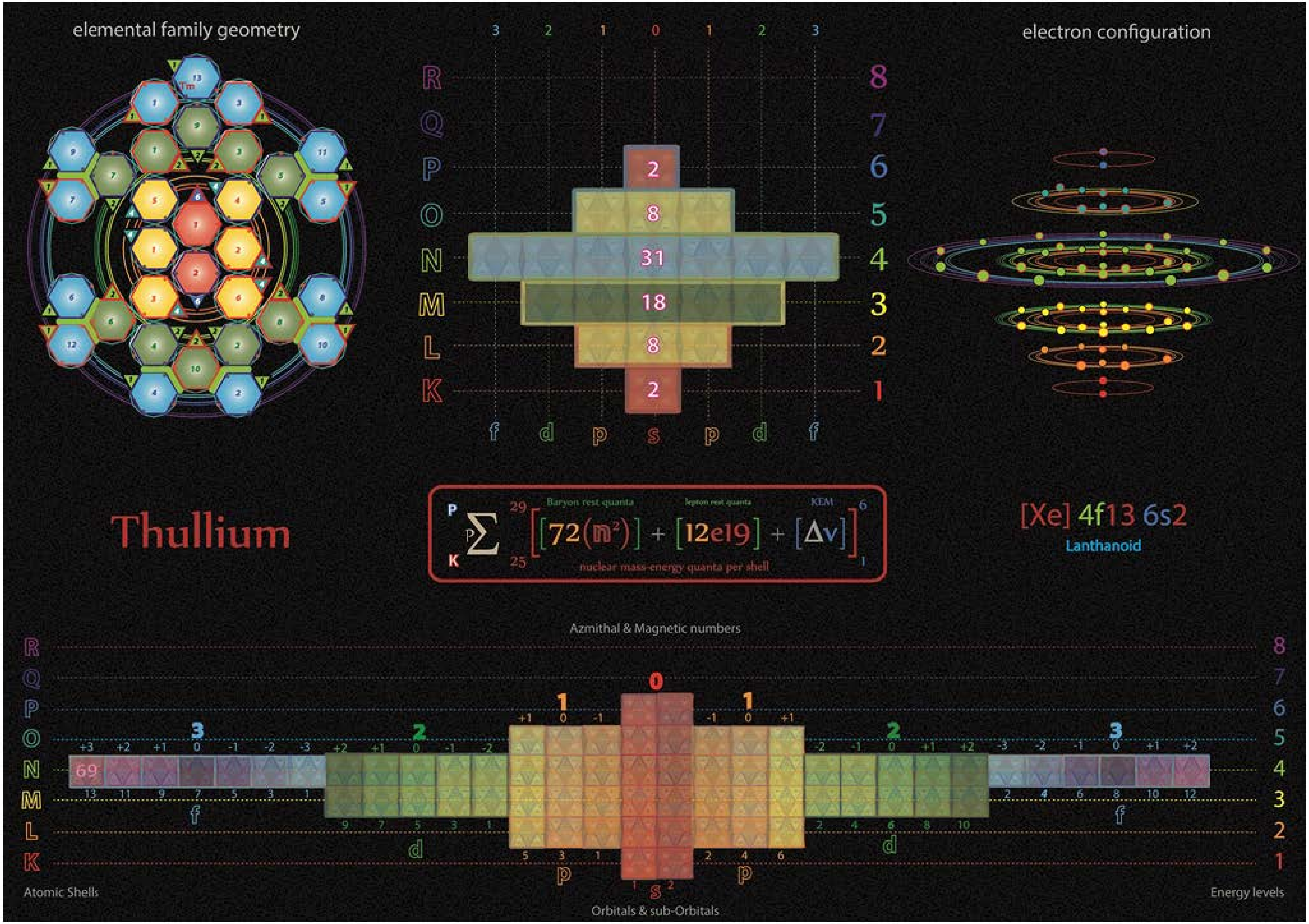
elemental family geometry
Tm
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2
8
31
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
69
Atomic Shells
Orbitals & sub-Orbitals
Energy levels

Thullium

$$\sum_{K}^{P} {}_{p}\left[[72(\mathfrak{m}^2)] + [12e19] + [\Delta v] \right]_{1}^{6} \quad 29 \quad 25$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Xe] 4f13 6s2

Lanthanoid

Tetryonics 53.69 - Thulium atomic config

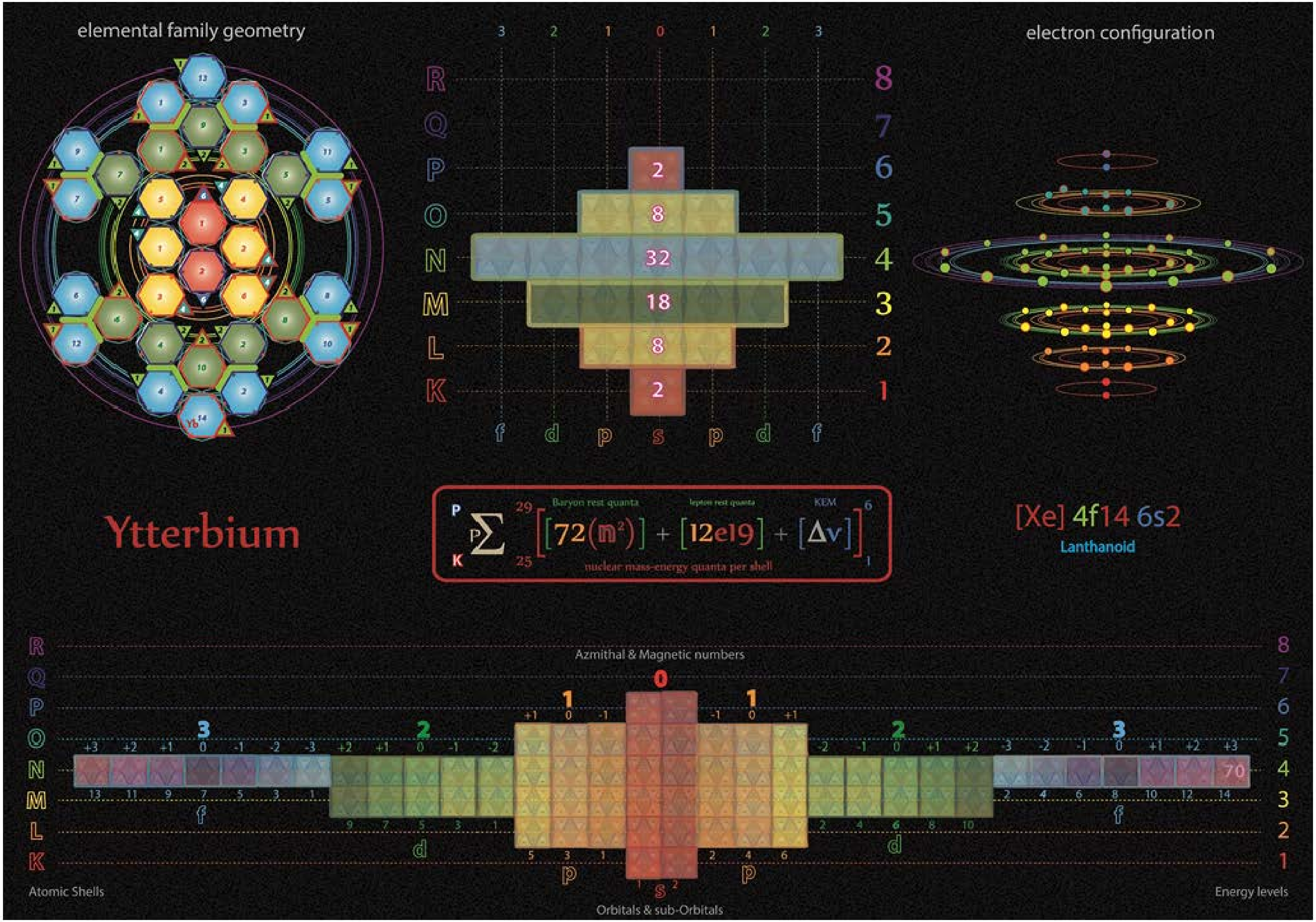
elemental family geometry
electron configuration
Ytterbium
Yb
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
32
18
8
2
f
d
p
s
p
d
f
P
K
Σ
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f14 6s2
Lanthanoid
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
70

Tetryonics 53.70 - Ytterbium atomic config

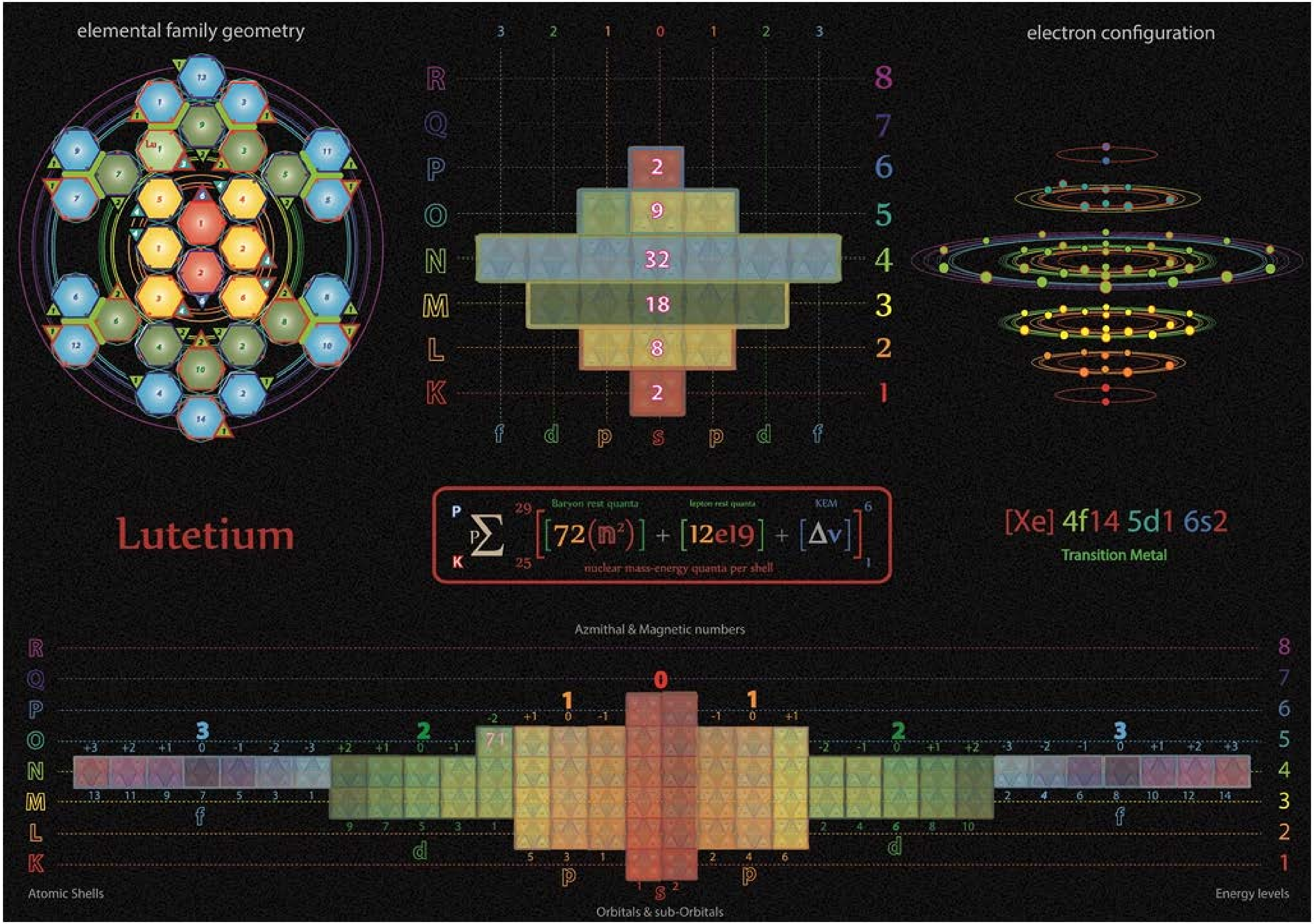
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2 9 32 18 8 2
f d p s p d f
Lutetium
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d1 6s2
Transition Metal
Azmithal & Magnetic numbers
71
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.71 - Lutetium atomic config

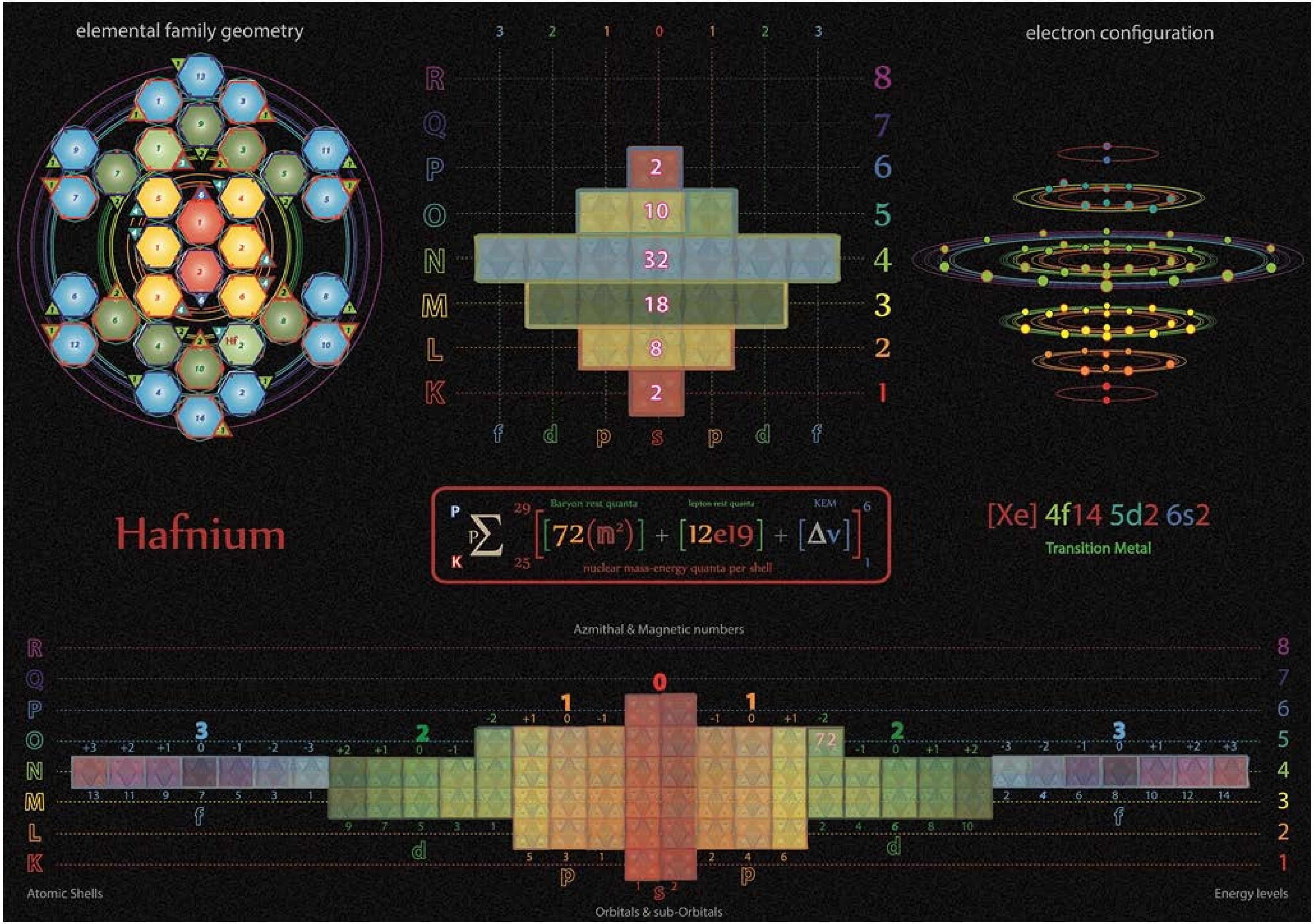
elemental family geometry
Hf
Hafnium
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
2
10
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
[Xe] 4f14 5d2 6s2
Transition Metal
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
$$\sum_{K}^{P} {}^{P}\sum_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_1^6$$
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
72
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f

Tetryonics 53.72 - Hafnium atomic config

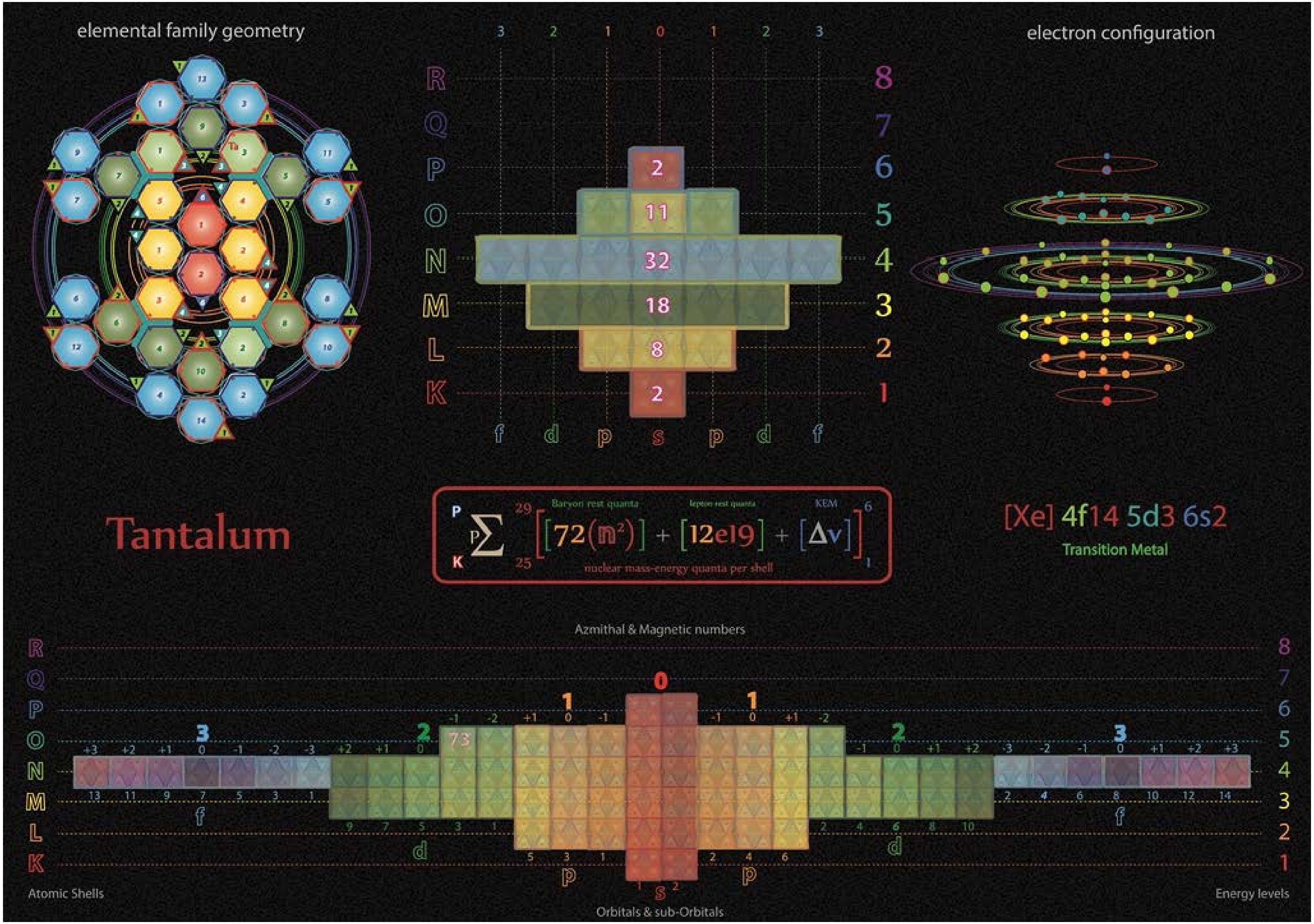
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
11
32
18
8
2
f d p s p d f
electron configuration
Tantalum
Baryon rest quanta
lepton rest quanta
KEM
P
K
29
25
6
1
[72(n²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d3 6s2
Transition Metal
Azmithal & Magnetic numbers
73
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.73 - Tantalum atomic config

Copyright ABRAHAM [2008] - All rights reserved

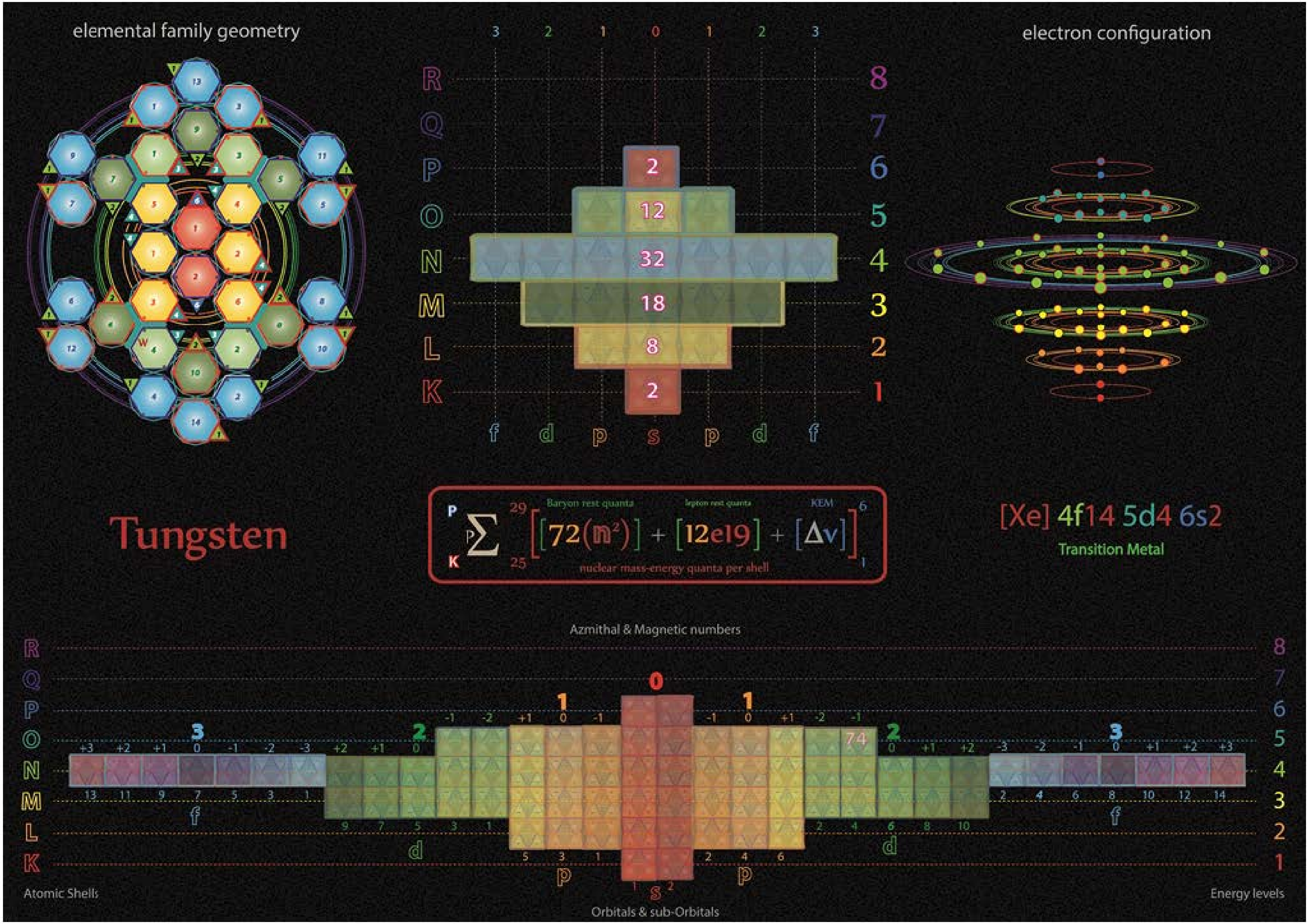
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
12
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
74
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tungsten

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{P} {}_{p}\sum_{25}^{29}\left[[72(\mathrm{n}^2)] + [12e19] + [\Delta v]\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f14 5d4 6s2

Transition Metal

Tetryonics 53.74 - Tungsten atomic config

Copyright ABRAHAM [2008] - All rights reserved

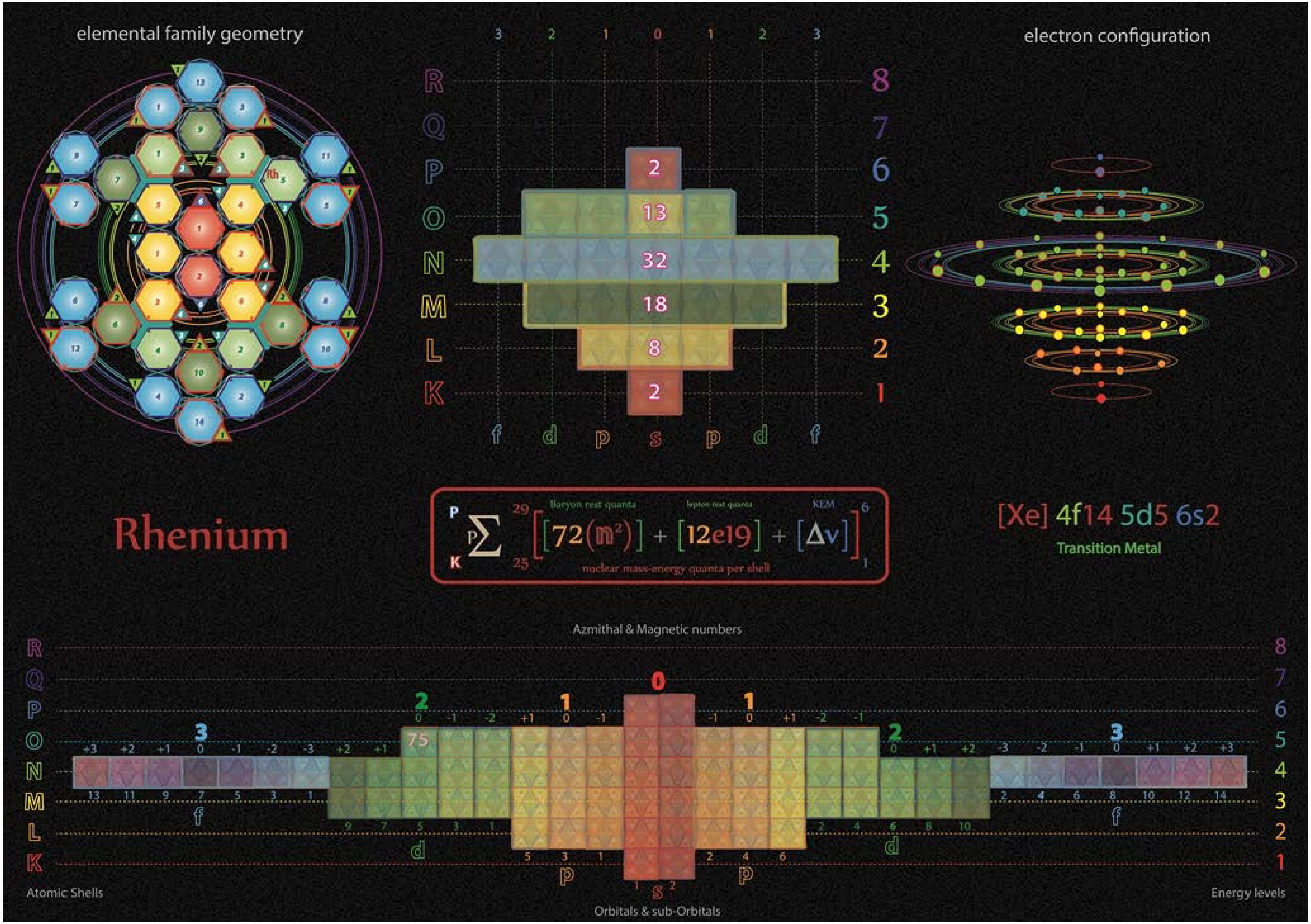
elemental family geometry
Rh
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
13
32
18
8
2
f d p s p d f
electron configuration
Rhenium
P
K
pΣ
29
25
Baryon rest quanta
lepton rest quanta
KEM
[[72(m²)] + [12e19] + [Δv]]
6
1
nuclear mass-energy quanta per shell
[Xe] 4f14 5d5 6s2
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
75
Atomic Shells
Energy levels
f d p s p d f
Orbitals & sub-Orbitals

Tetryonics 53.75 - Rhenium atomic config

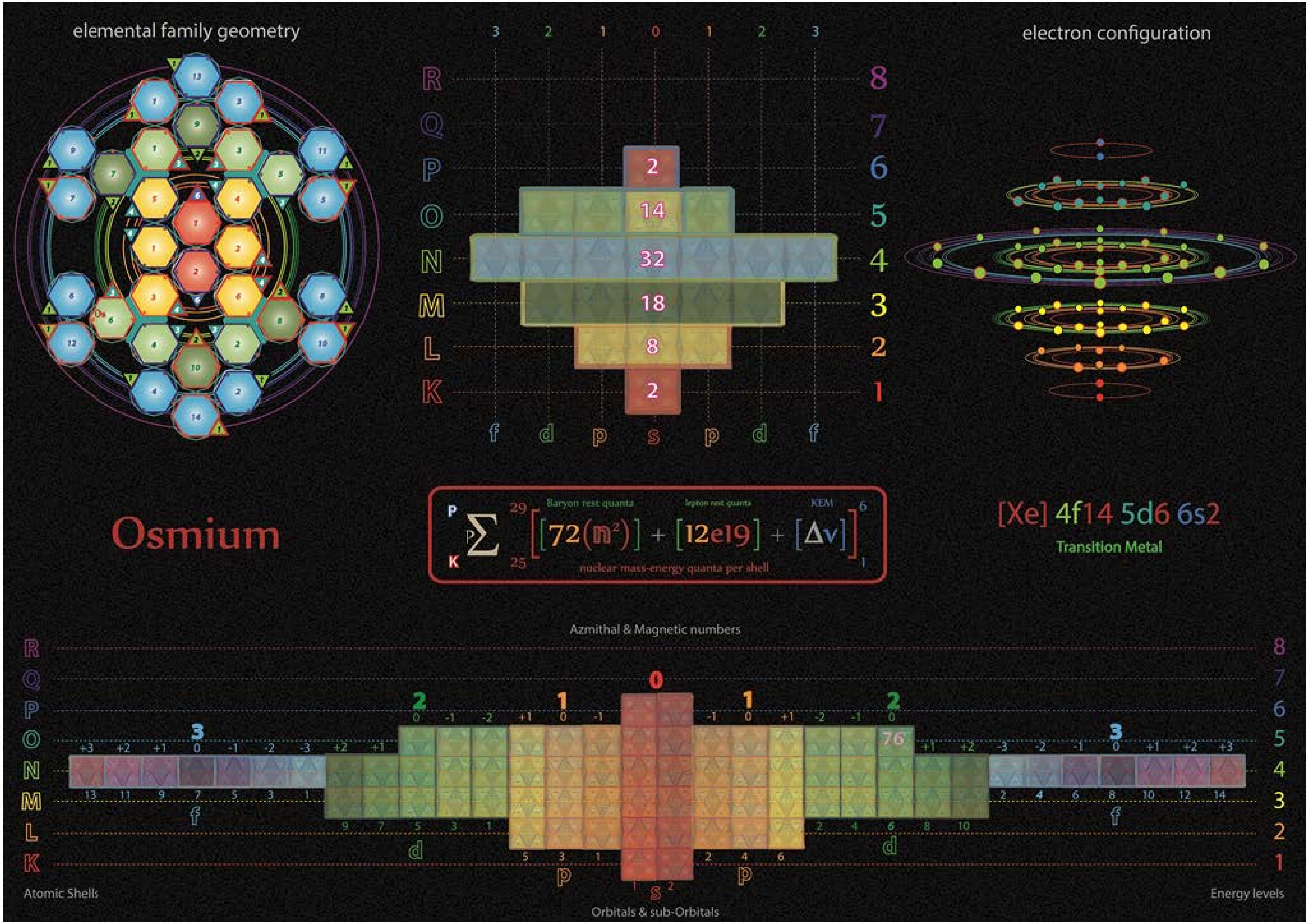
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
14
32
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
76
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Osmium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{P}{}_{25}^{29}\left[[72(m^2)] + [12e19] + [\Delta v]\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f14 5d6 6s2

Transition Metal

Tetryonics 53.76 - Osmium atomic config

Copyright ABRAHAM [2008] - All rights reserved

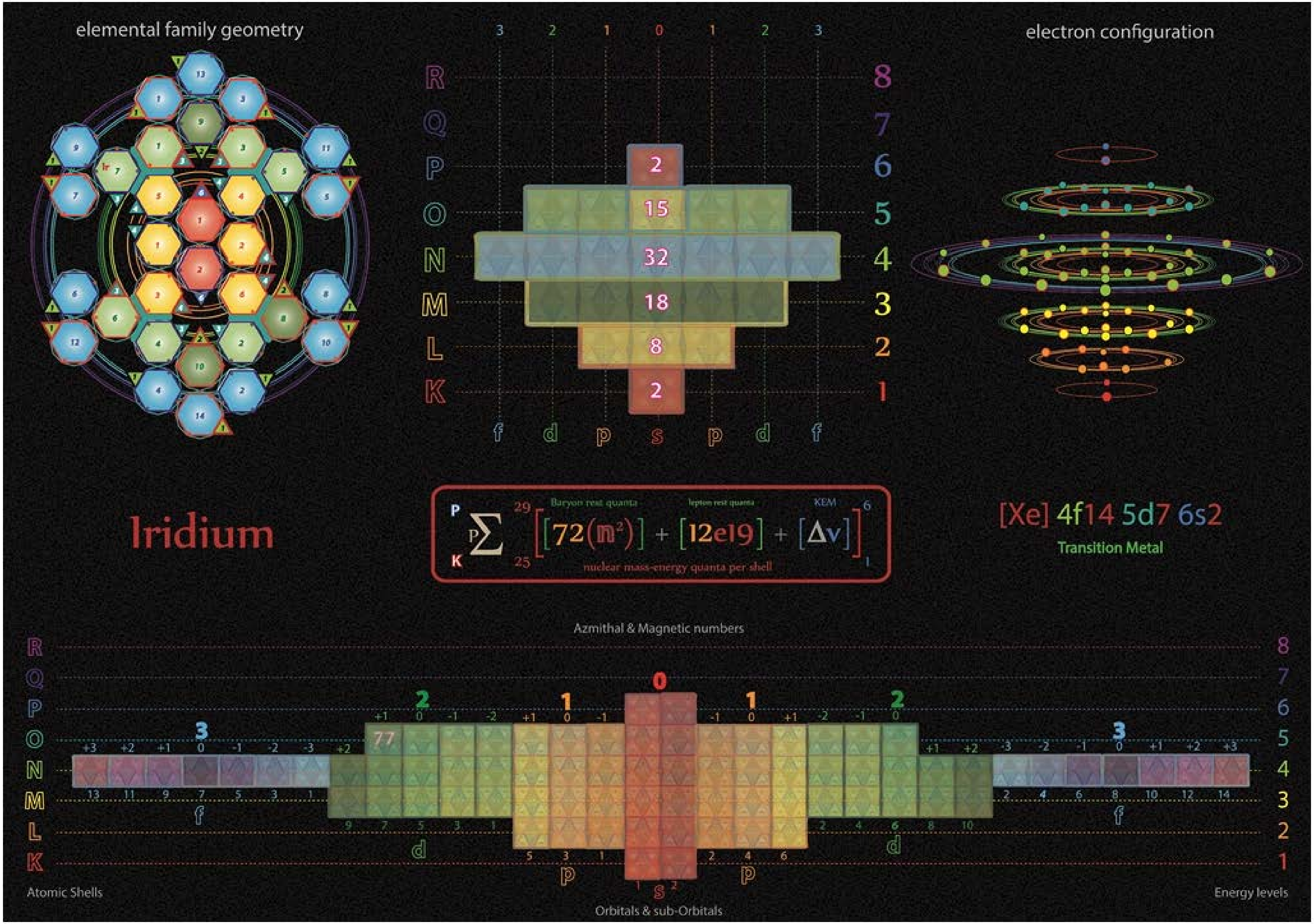
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
15
32
18
8
2
f d p s p d f
electron configuration
Iridium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
$\sum_{P} [72(m^2)] + [12e19] + [\Delta v]$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d7 6s2
Transition Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
2 2
3 3
77
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 3 1
p
1 s 2
2 4 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.77 - Iridium atomic config

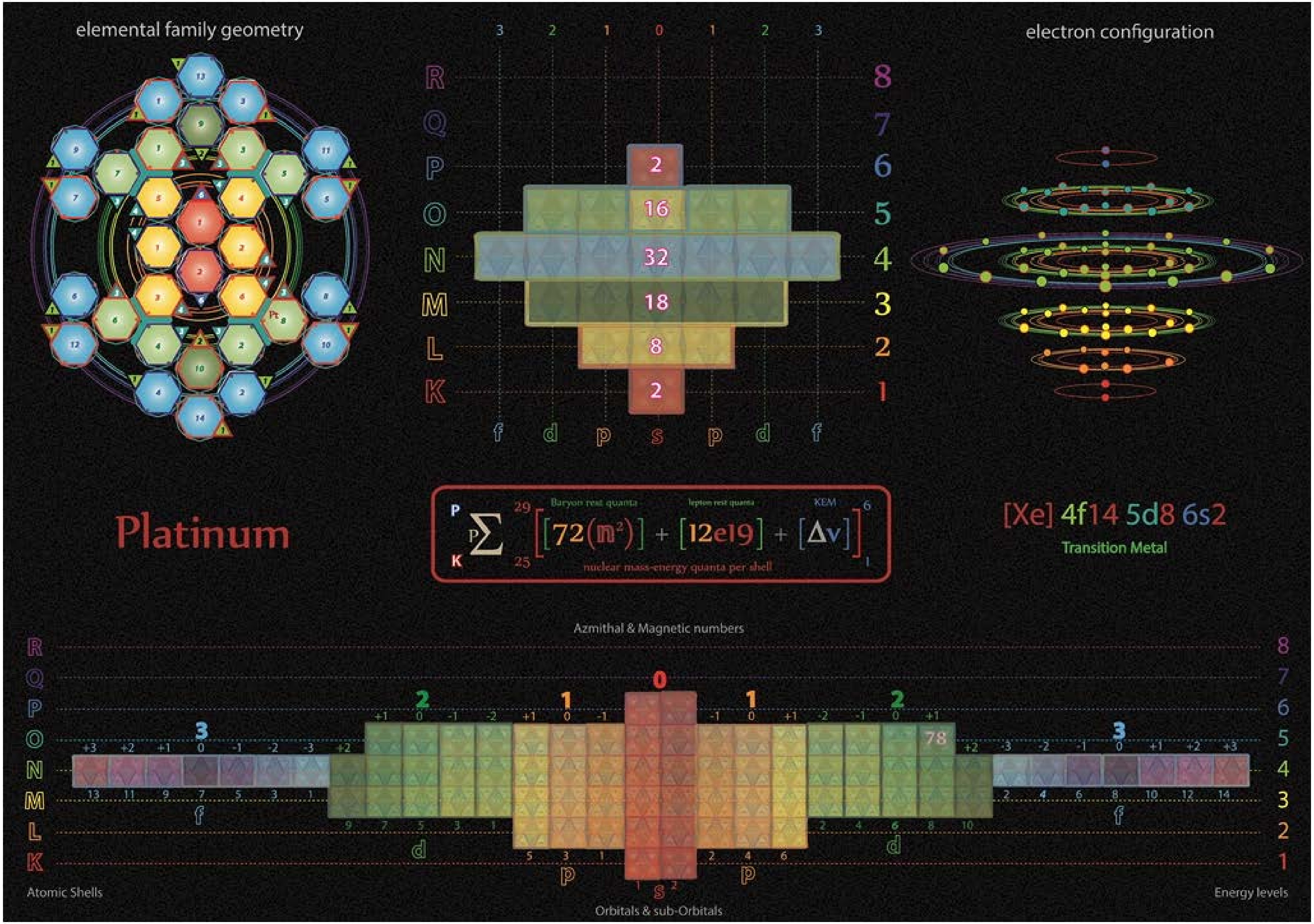

elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 16 32 18 8 2
f d p s p d f
Azmithal & Magnetic numbers
0 1 2 3
+3 +2 +1 0 -1 -2 -3
78
f d p s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Platinum

$$ {}^{P}_{K}\sum{}_{P}\ {}^{29}_{25}\Big[\,[72(m^2)] + [12e19] + [\Delta v]\,\Big]^{6}_{1} $$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Xe] 4f14 5d8 6s2

Transition Metal

Tetryonics 53.78 - Platinum atomic config

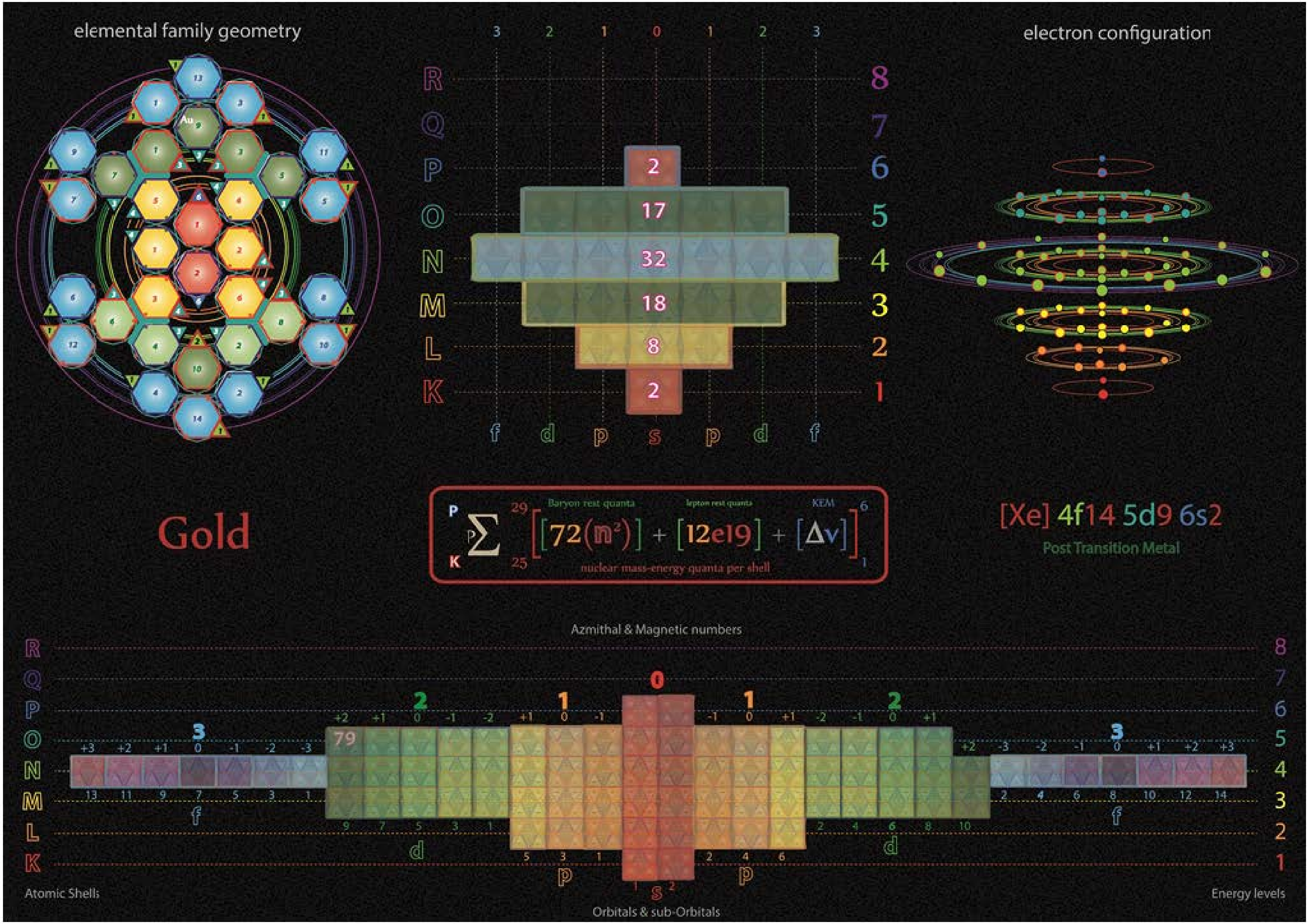
elemental family geometry
Au
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
17
32
18
8
2
f d p s p d f
3 2 1 0 1 2 3
Gold
Baryon rest quanta
lepton rest quanta
KEM
P
K
29
25
6
1
[72(n²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d9 6s2
Post Transition Metal
Azmithal & Magnetic numbers
79
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.79 - Gold atomic config

Copyright ABRAHAM [2008] - All rights reserved

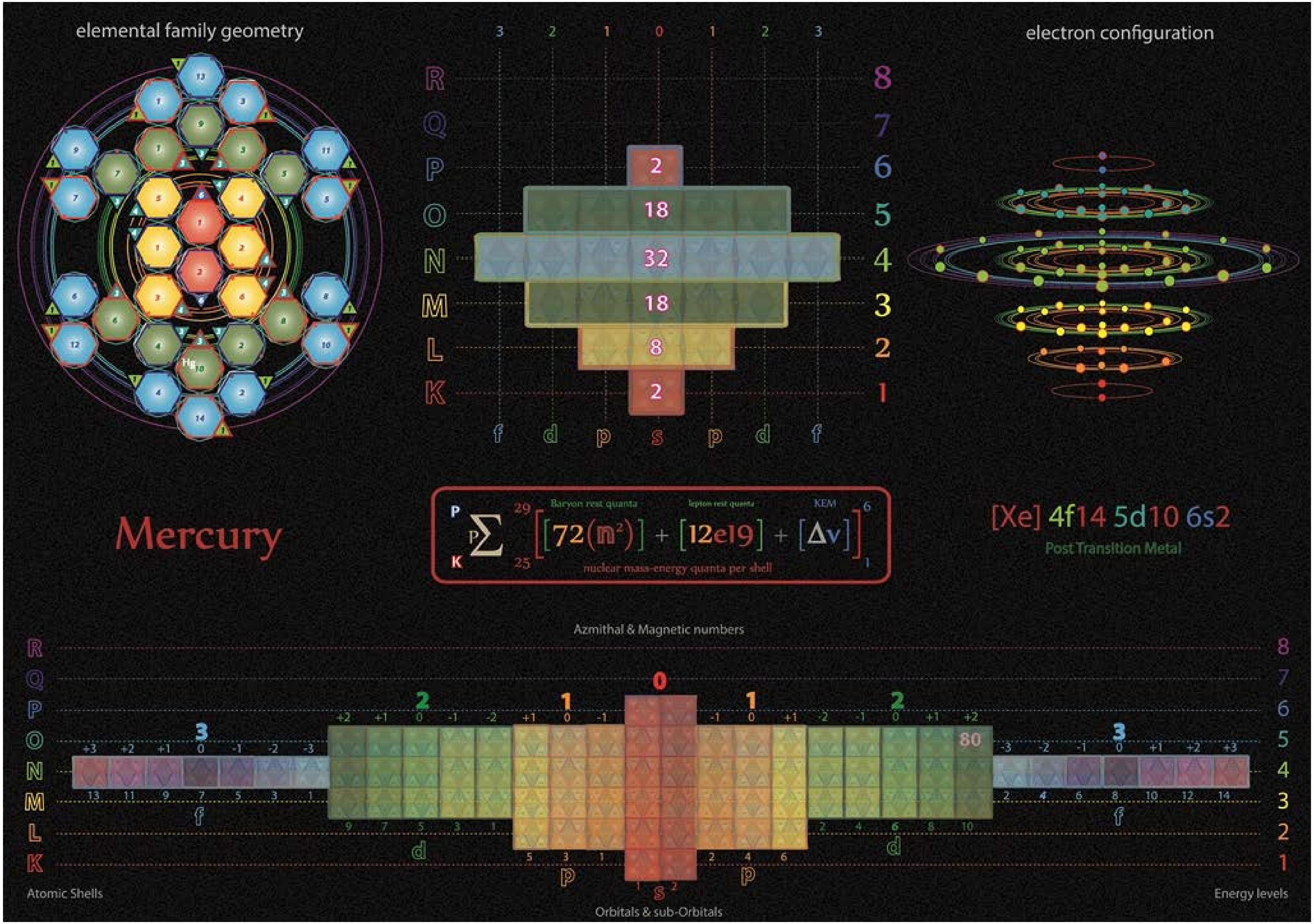
elemental family geometry
Hg
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
18
32
18
8
2
f d p s p d f
electron configuration
Mercury
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2
Post Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
80
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.80 - Mercury atomic config

Copyright ABRAHAM [2008] - All rights reserved

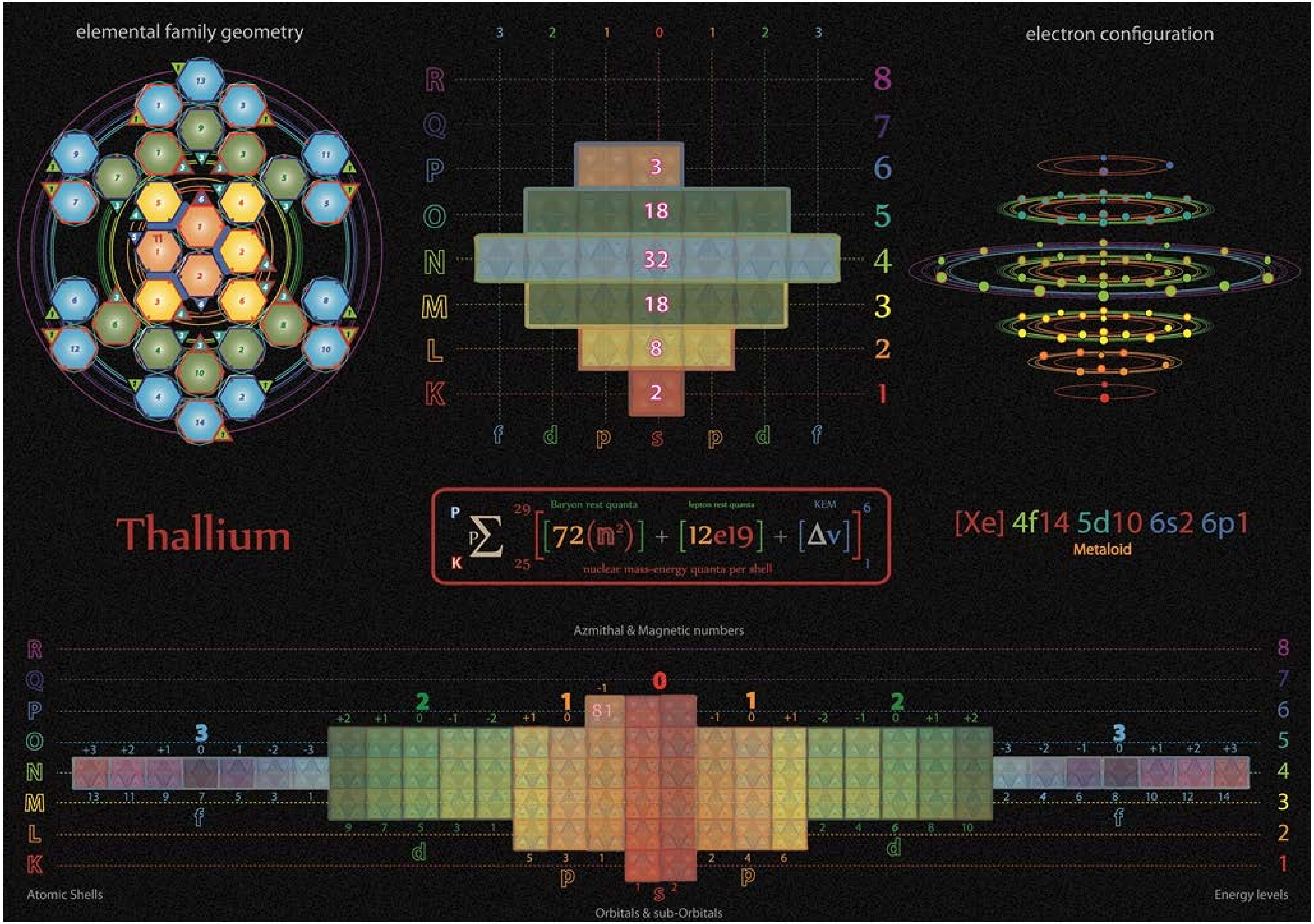
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
32
18
8
2
f d p s p d f
Thallium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
$\sum_{K}^{P} {}_{25}^{29}\left[72(\mathrm{m}^2)\right] + \left[12e19\right] + \left[\Delta v\right]_1^6$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p1
Metaloid
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0
-1
81
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.81 - Thallium atomic config

Copyright ABRAHAM [2008] - All rights reserved

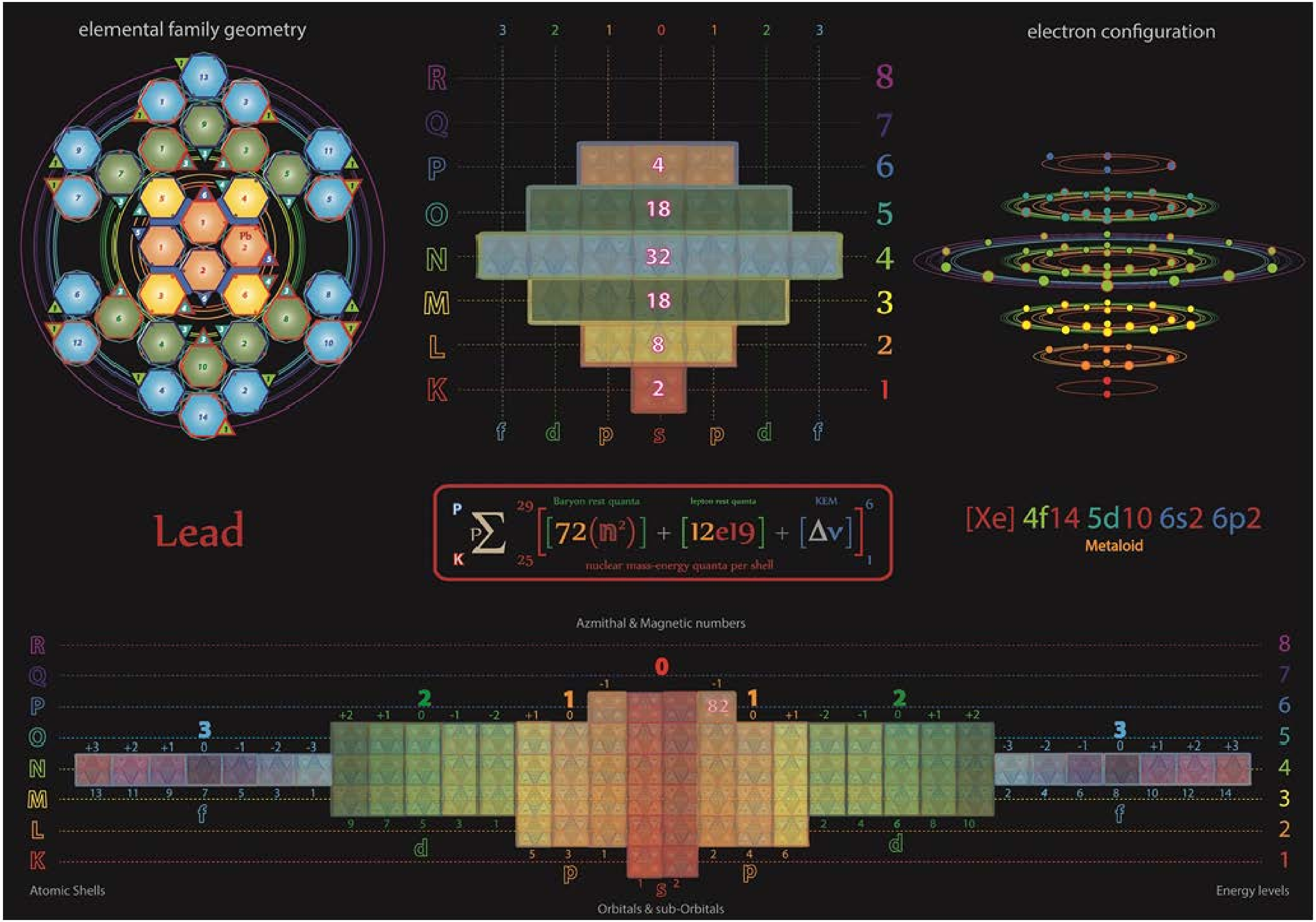
elemental family geometry
Pb
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
4
18
32
18
8
2
f d p s p d f
Lead
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p2
Metaloid
Azmithal & Magnetic numbers
0
1
2
3
82
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.82 - Lead atomic config

Copyright ABRAHAM [2008] - All rights reserved

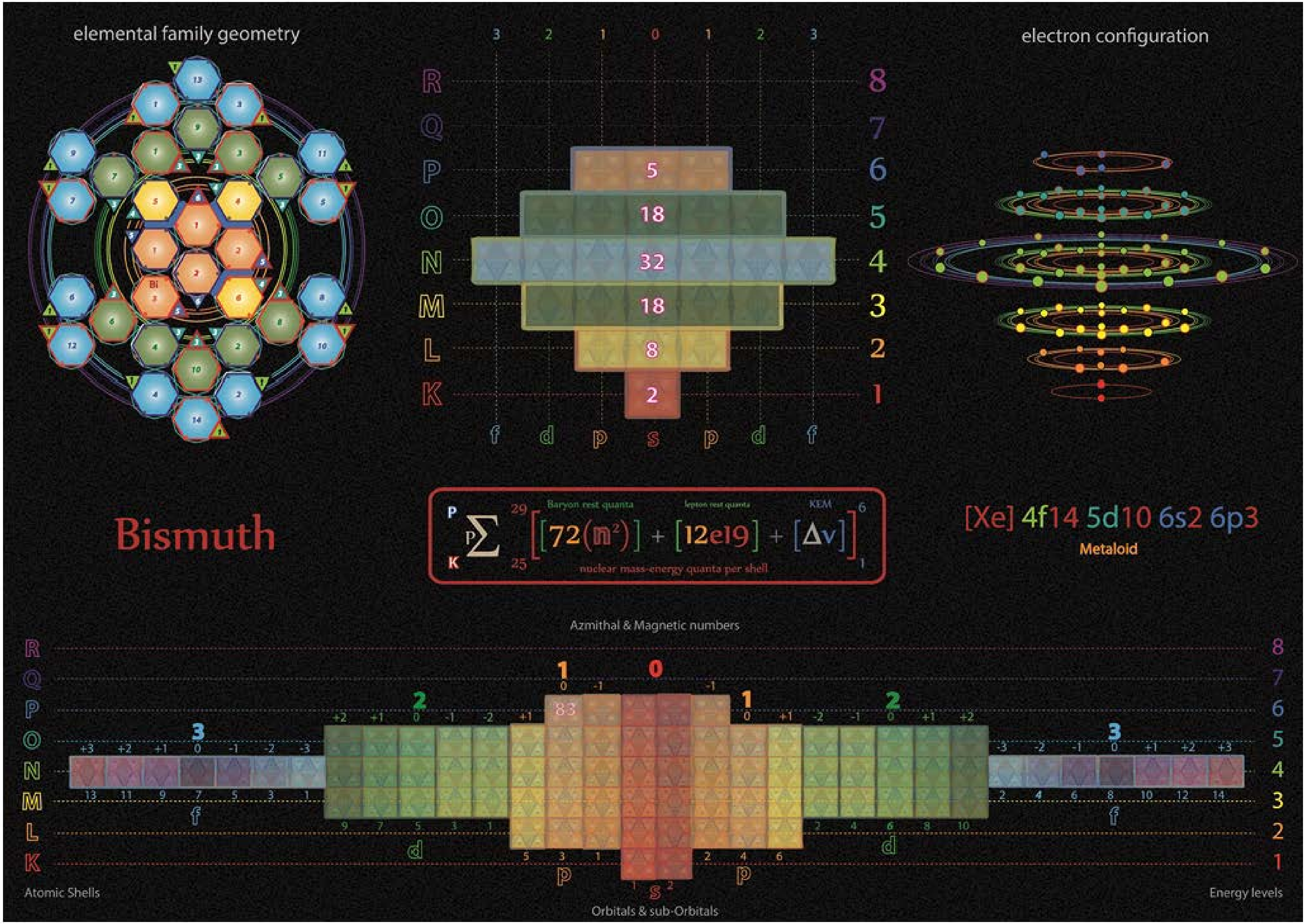
elemental family geometry
Bi
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
5
18
32
18
8
2
f d p s p d f
electron configuration
Bismuth
P
K
29
25
6
1
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p3
Metaloid
Azmithal & Magnetic numbers
83
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{P} {}^{29}_{25}\Big[[72(m^2)] + [12e19] + [\Delta v]\Big]^{6}_{1}$$

Tetryonics 53.83 - Bismuth atomic config

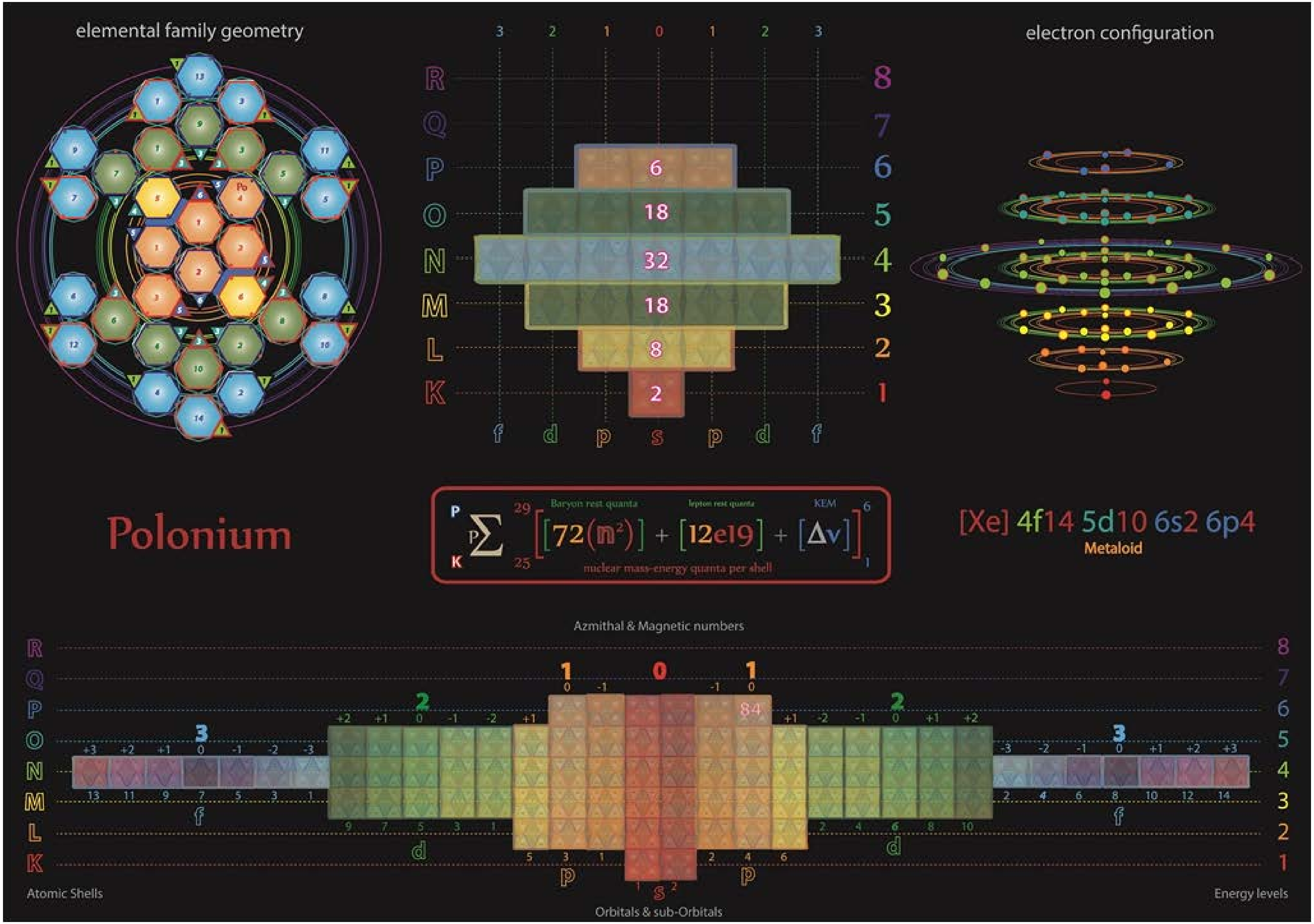
elemental family geometry
Po
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
32
18
8
2
f d p s p d f
electron configuration
Polonium
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
6
1
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p4
Metaloid
Azmithal & Magnetic numbers
84
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$${}^{P}_{K}\sum_{25}^{29}\left[\left[72(m^2)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{6}$$

Tetryonics 53.84 - Polonium atomic config

Copyright ABRAHAM [2008] - All rights reserved

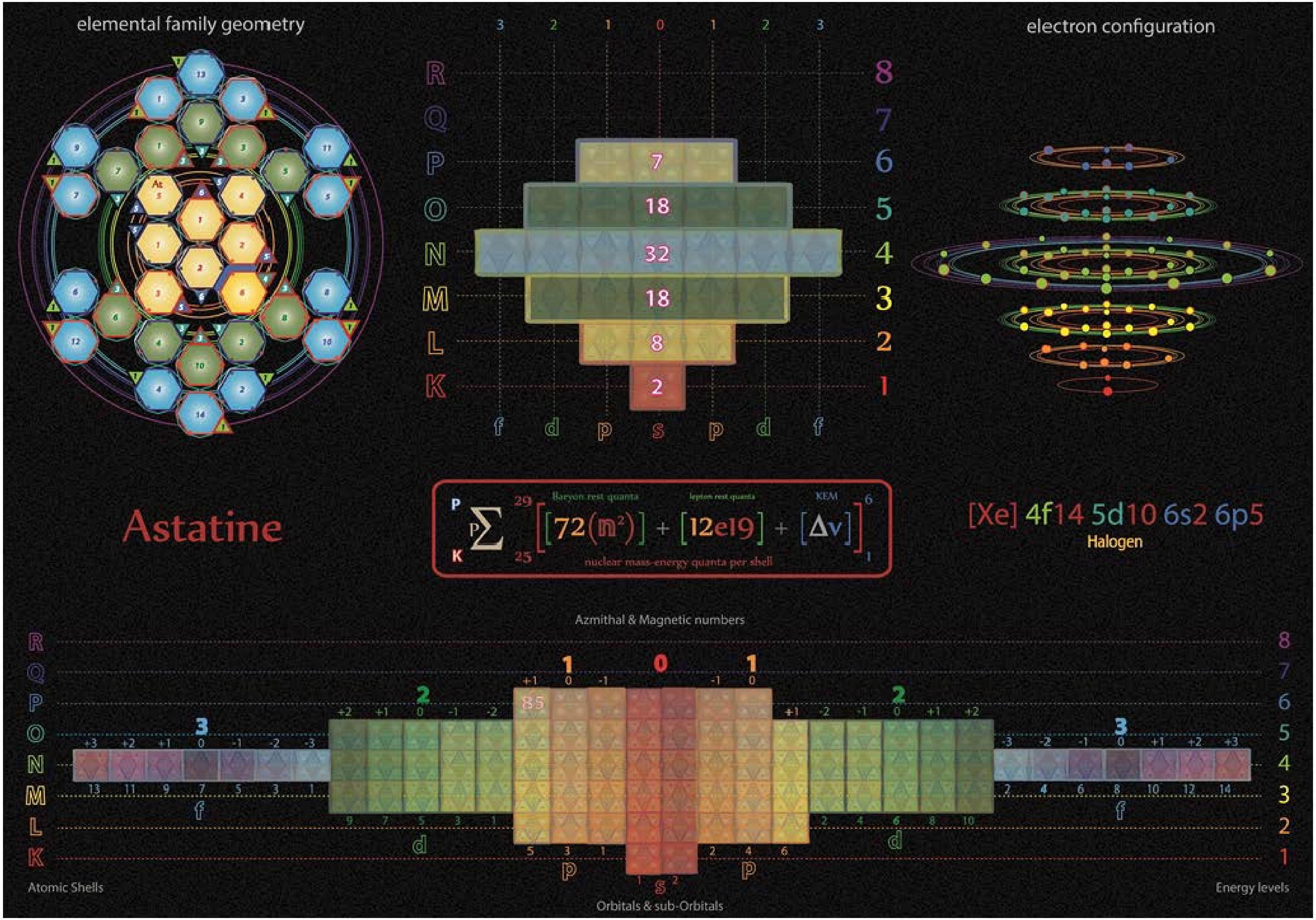
elemental family geometry
At
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
7
18
32
18
8
2
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
3
2
1
0
1
2
3
85
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Astatine

Baryon rest quanta — lepton rest quanta — KEM

$${}^{P}_{K}\;{}_{P}\sum_{25}^{29}\left[\left[72\left(\mathrm{m}^{2}\right)\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{6}$$

nuclear mass-energy quanta per shell

[Xe] 4f14 5d10 6s2 6p5

Halogen

Tetryonics 53.85 - Astatine atomic config

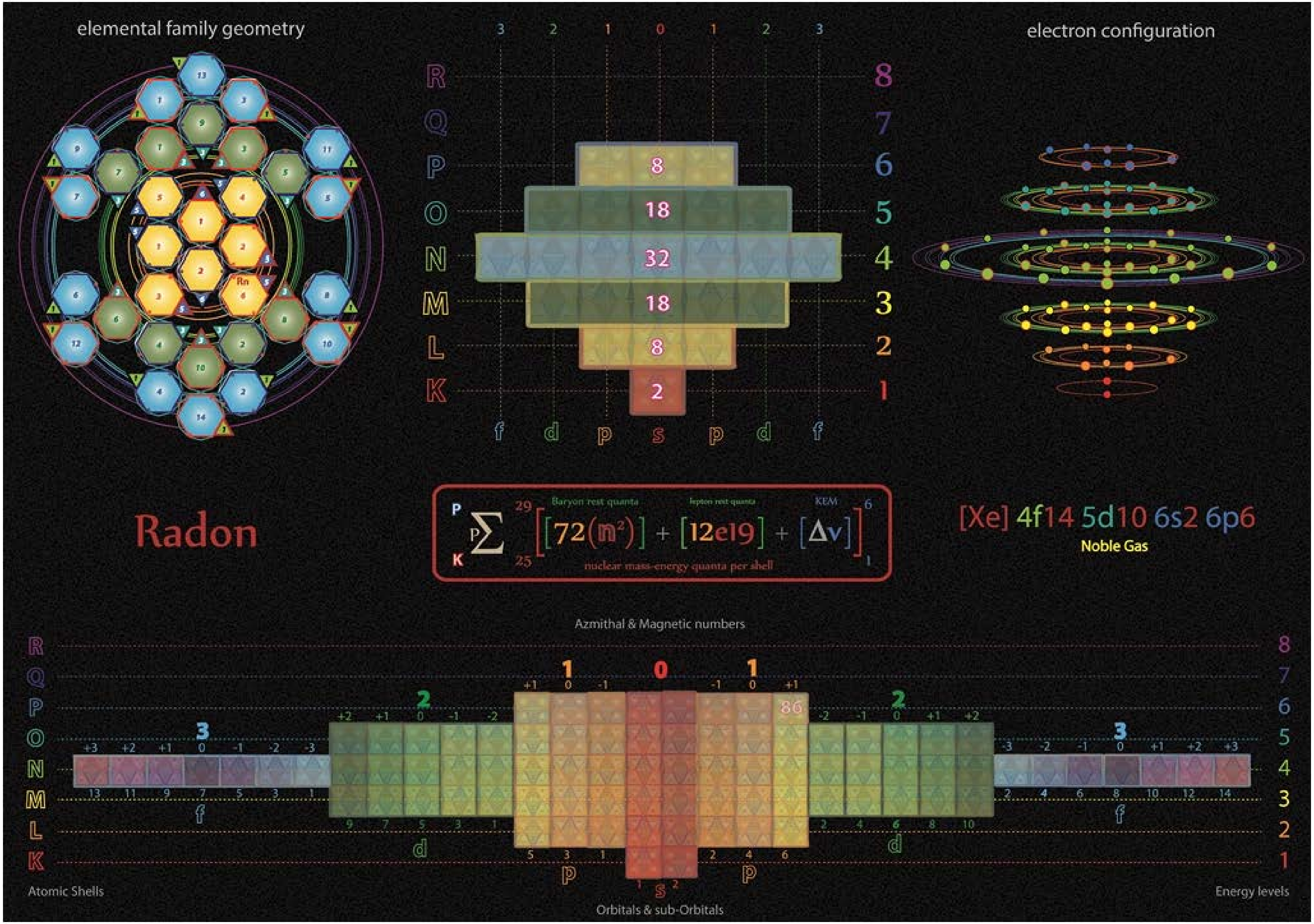
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
8
18
32
18
8
2
f d p s p d f
Radon
P
K
$\sum_{K}^{P}$ 29 25
Baryon rest quanta
lepton rest quanta
KEM
$\left[[72(\mathrm{n}^2)] + [12e19] + [\Delta v] \right]_1^6$
nuclear mass-energy quanta per shell
[Xe] 4f14 5d10 6s2 6p6
Noble Gas
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
86
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.86 - Radon atomic config

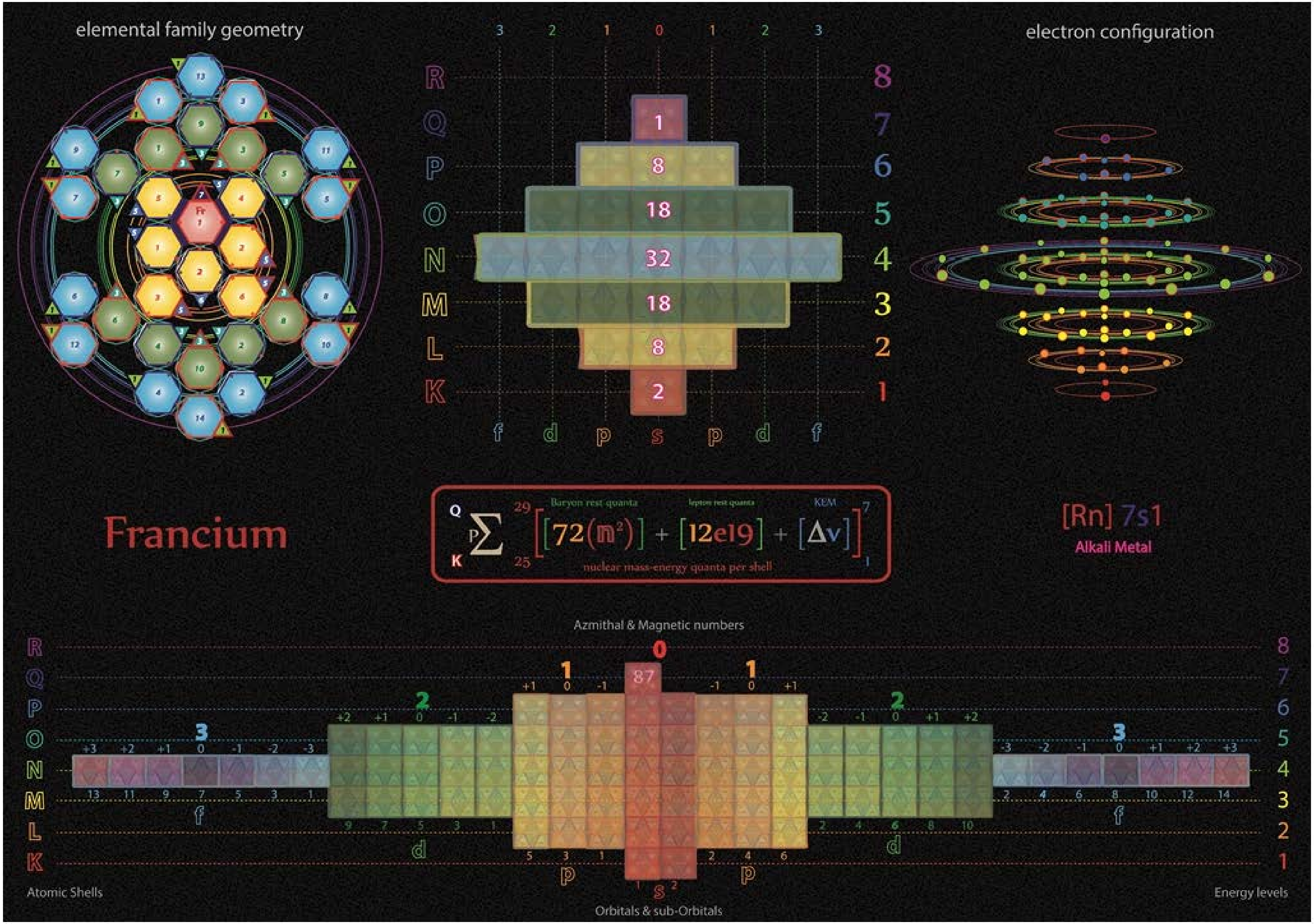
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
1
8
18
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
87
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Francium

$$\sum_{K}^{Q} P \; {}_{25}^{29}\Big[[72(\mathrm{m}^2)] + [12e19] + [\Delta v] \Big]_{1}^{7}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 7s1

Alkali Metal

Tetryonics 53.87 - Francium atomic config

Copyright ABRAHAM [2008] - All rights reserved

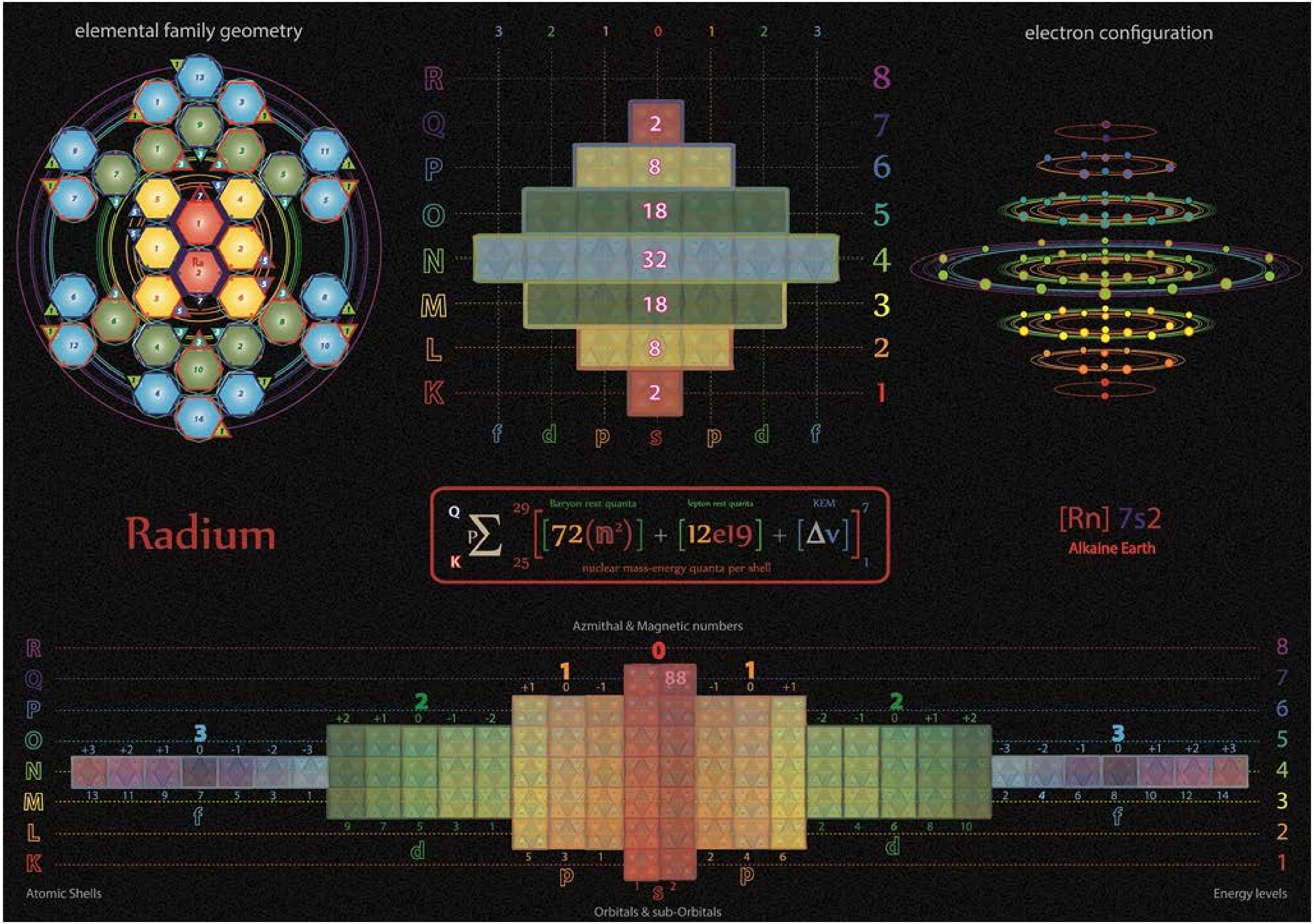
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
18
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
88
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Radium

$$\sum_{K}^{Q} P \sum_{25}^{29} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{7}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 7s2

Alkaine Earth

Tetryonics 53.88 - Radium atomic config

Copyright ABRAHAM [2008] - All rights reserved

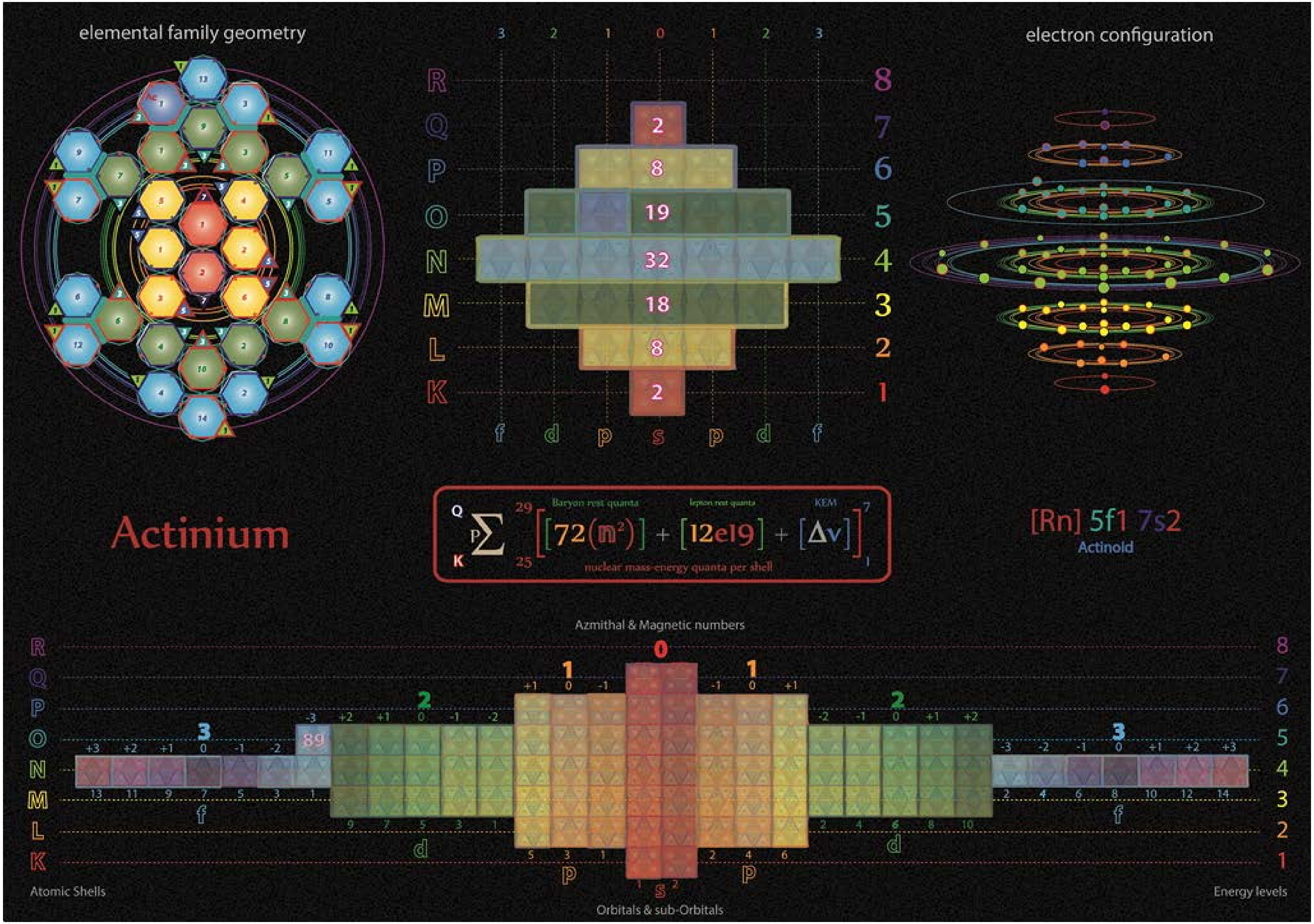
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 19 32 18 8 2
f d p s p d f
Actinium
$$\sum_{K}^{Q} P \sum_{25}^{29} \left[[72(m^2)] + [12e19] + [\Delta v] \right]_1^7$$
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f1 7s2
Actinoid
Azmithal & Magnetic numbers
89
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.89 - Actinium atomic config

Copyright ABRAHAM [2008] - All rights reserved

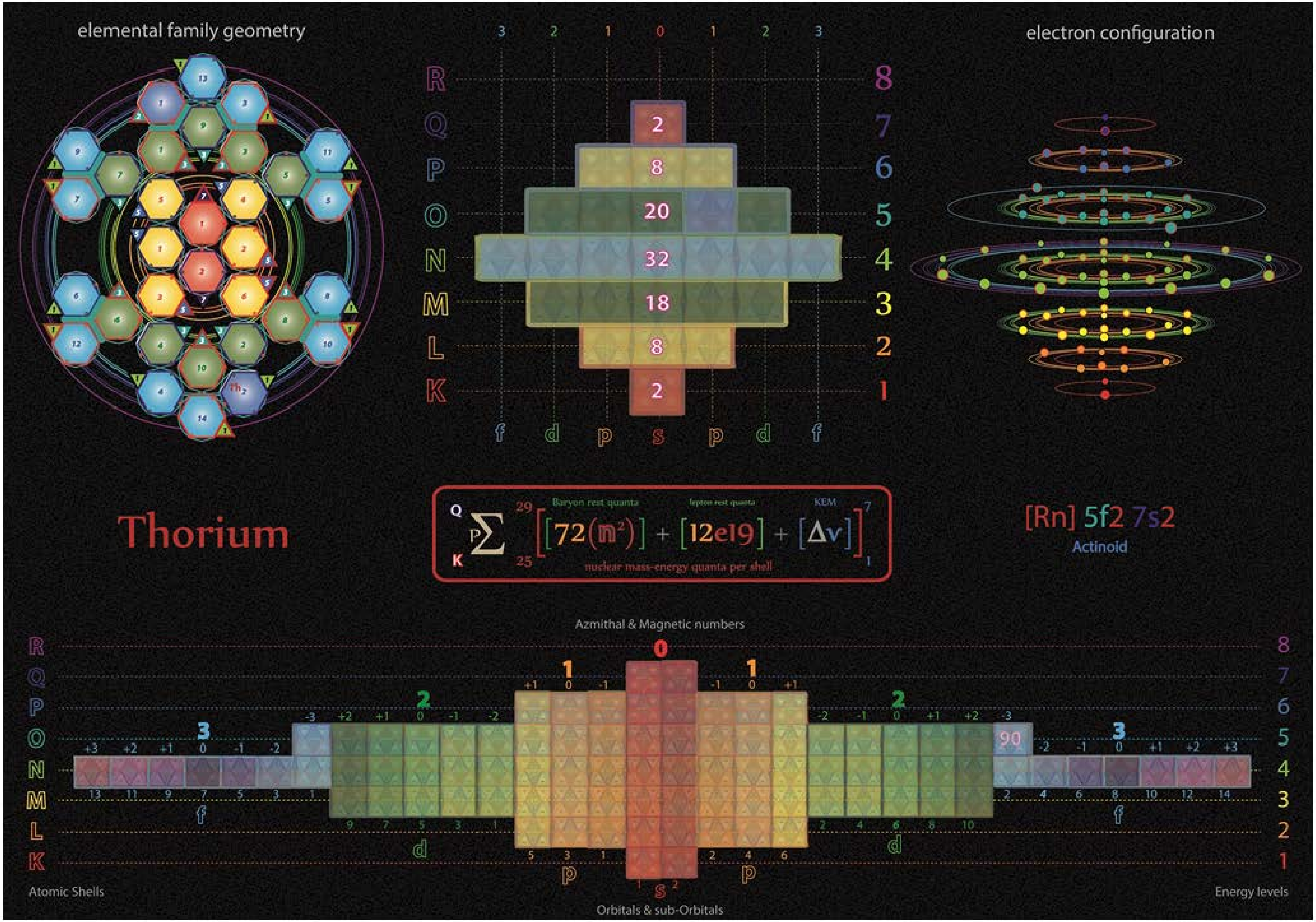
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 20 32 18 8 2
f d p s p d f
Th
Azmithal & Magnetic numbers
0 1 1 2 2 3 3
+1 0 -1
-3 +2 +1 0 -1 -2
+3 +2 +1 0 -1 -2
90
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Thorium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{Q} P \left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v] \right]_{25}^{29}\,{}_{1}^{7}$$

nuclear mass-energy quanta per shell

[Rn] 5f2 7s2

Actinoid

Tetryonics 53.90 - Thorium atomic config

elemental family geometry

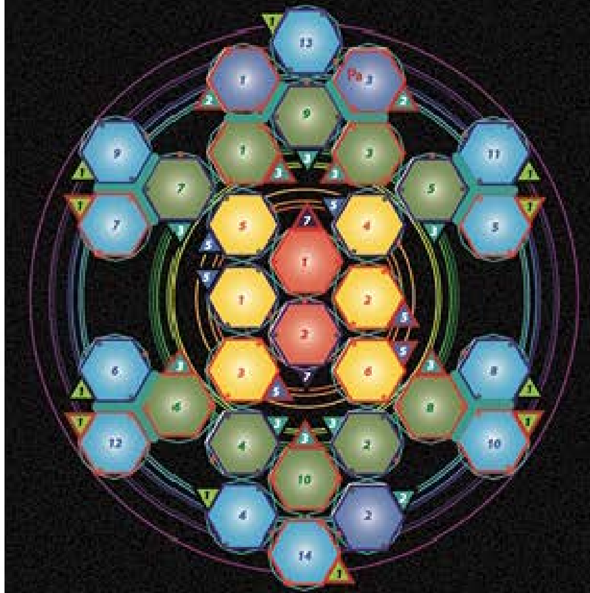
13
Pa 3
9
11
7
5
1
2
12
10
14
4

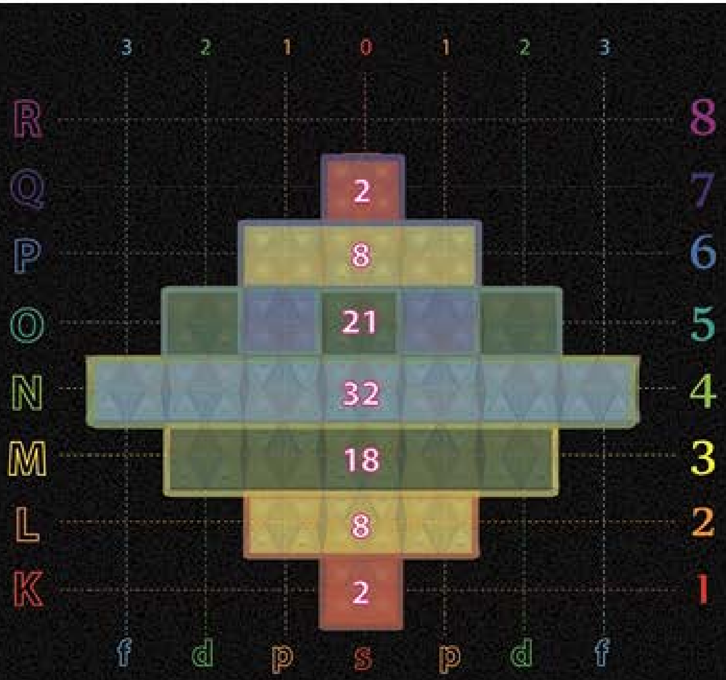
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
21
32
18
8
2
f
d
p
s
p
d
f

electron configuration

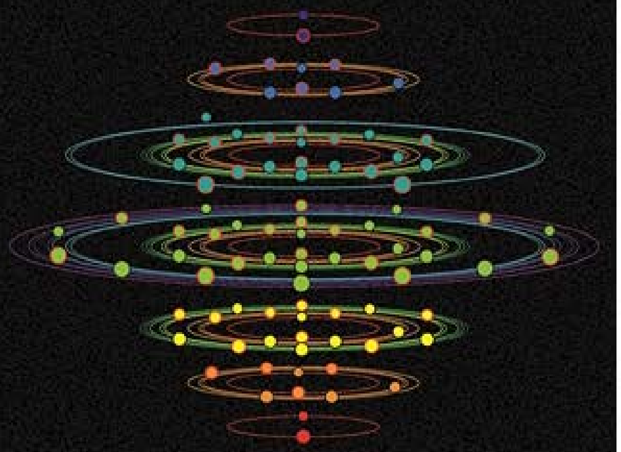

Protactinium

$$\sum_{K}^{Q} {}_{P} \; {}^{29}_{25}\left[\left[72(\mathrm{m}^2)\right] + \left[12\mathrm{e}19\right] + \left[\Delta v\right] \right]^{7}_{1}$$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Rn] 5f3 7s2

Actinoid

Azmithal & Magnetic numbers

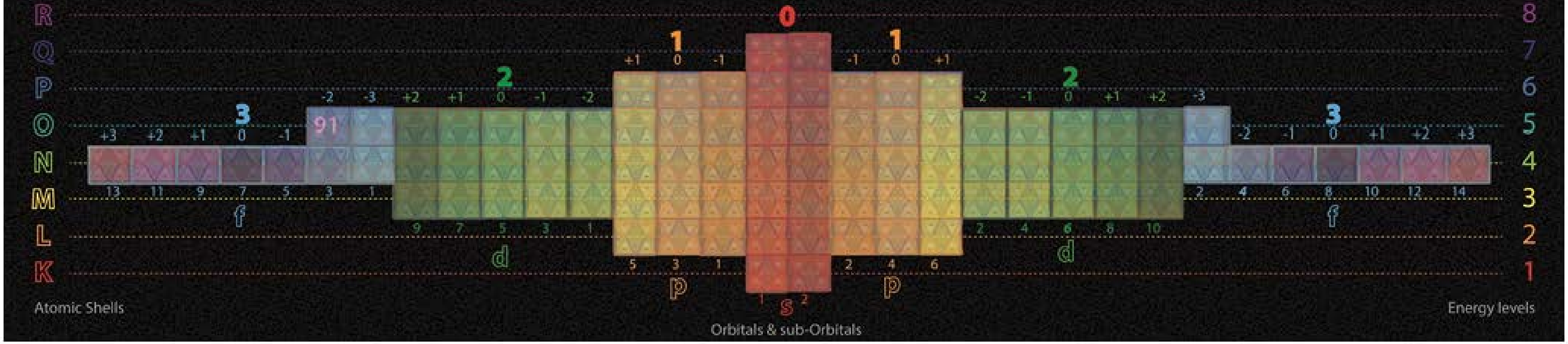
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
91
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
p
d
f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.91 - Protactinium atomic config

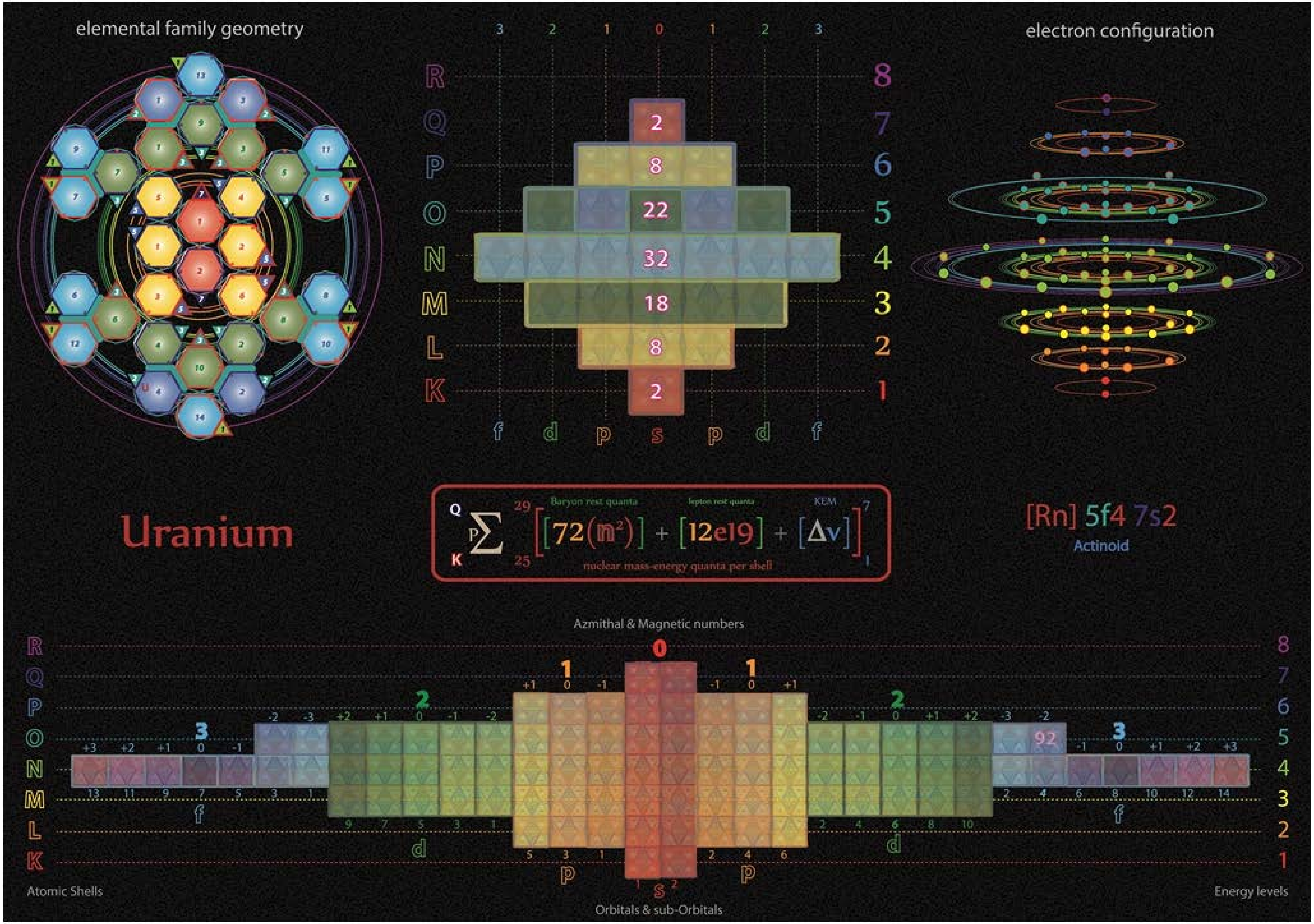
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
2
8
22
32
18
8
2
Uranium
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f4 7s2
Actinoid
Azmithal & Magnetic numbers
92
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{Q} {}_{P}\;{}_{25}^{29}\left[[72(\mathrm{m}^2)] + [12e19] + [\Delta v] \right]_{1}^{7}$$

Tetryonics 53.92 - Uranium atomic config

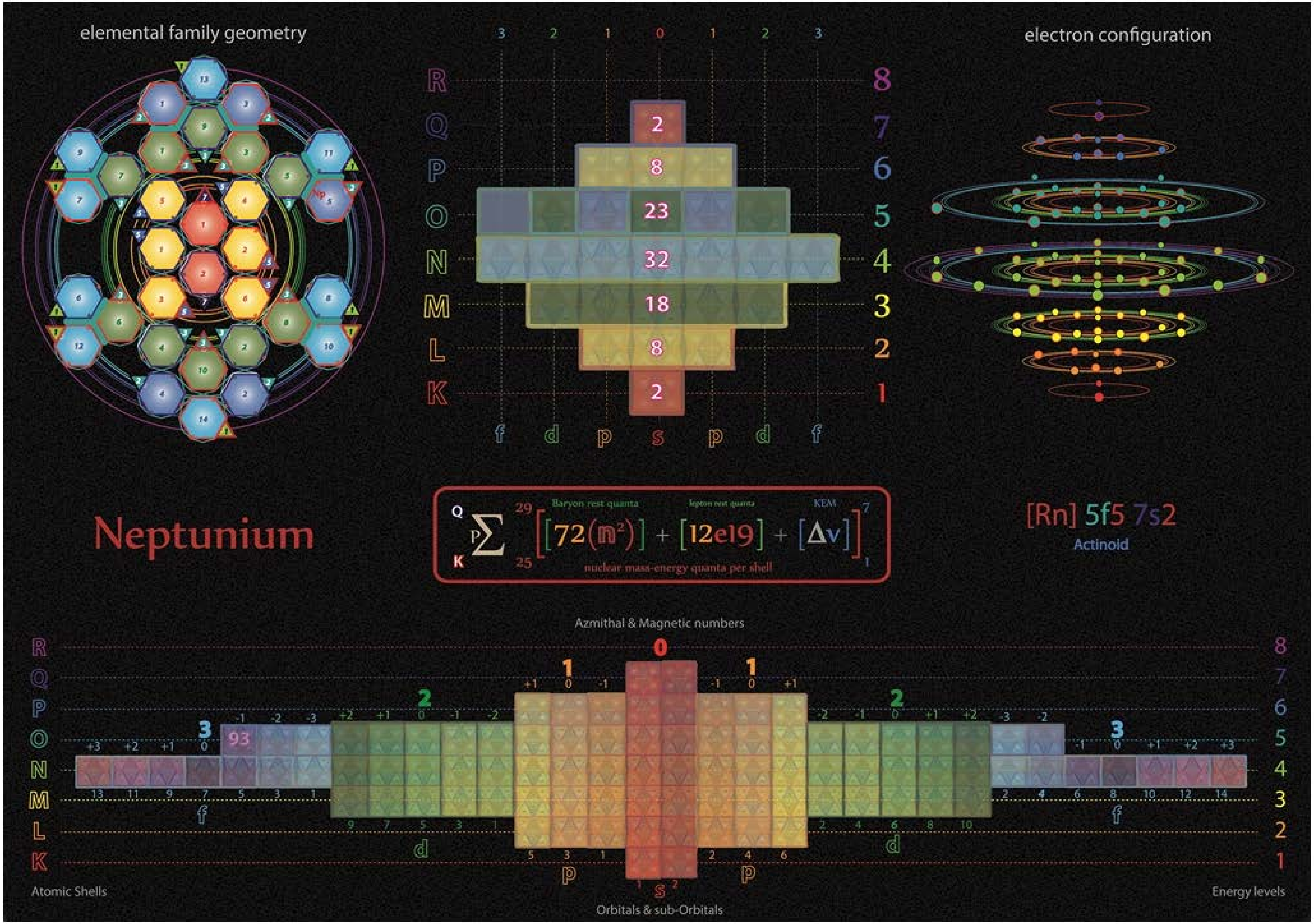
elemental family geometry
Np
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
2
8
23
32
18
8
2
f
d
p
s
p
d
f
Neptunium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f5 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
93
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.93 - Neptunium atomic config

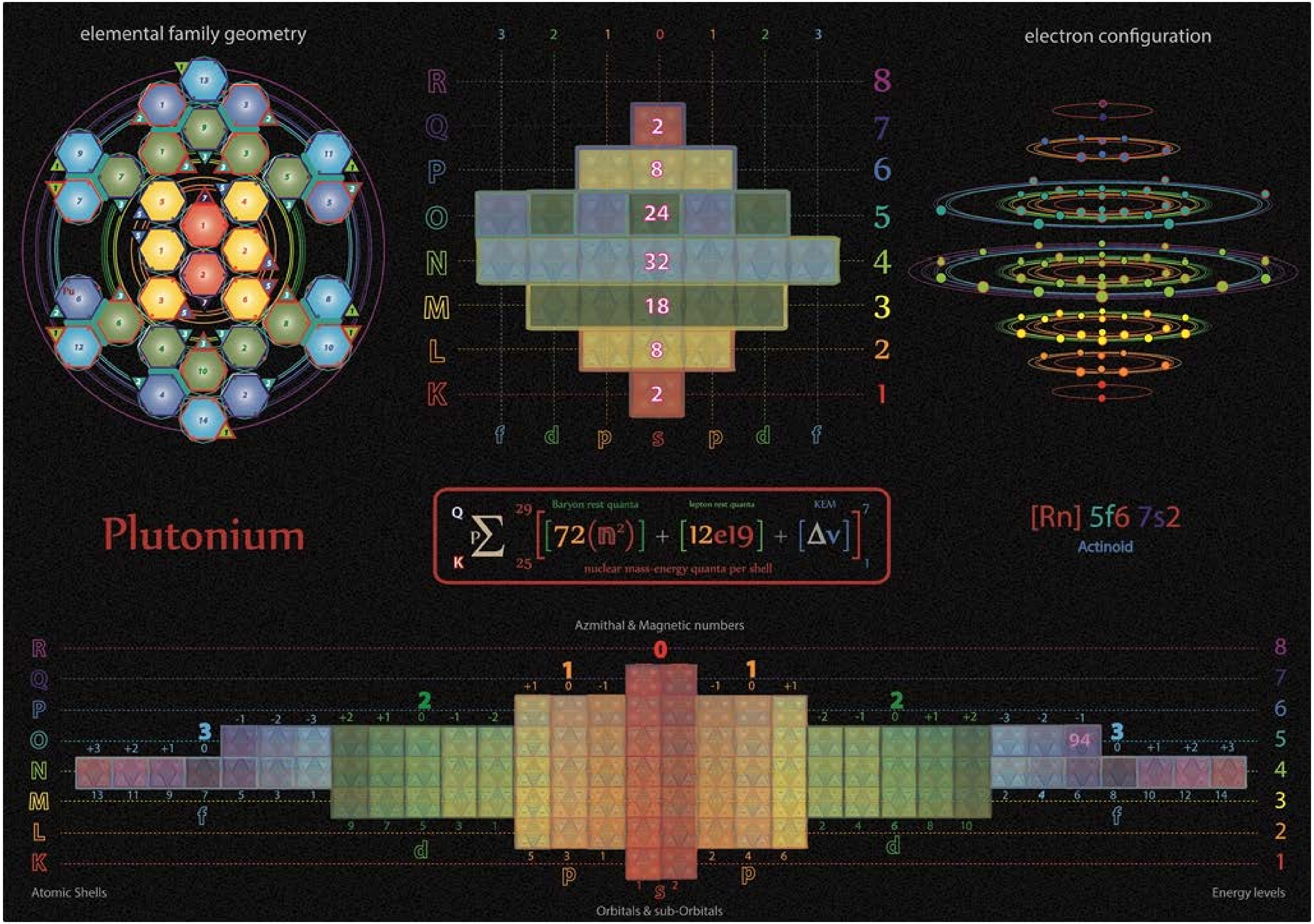
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2
8
24
32
18
8
2
f d p s p d f
Plutonium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f6 7s2
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
94
f
d
p
s
p
d
f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.94 - Plutonium atomic config

Copyright ABRAHAM [2008] - All rights reserved

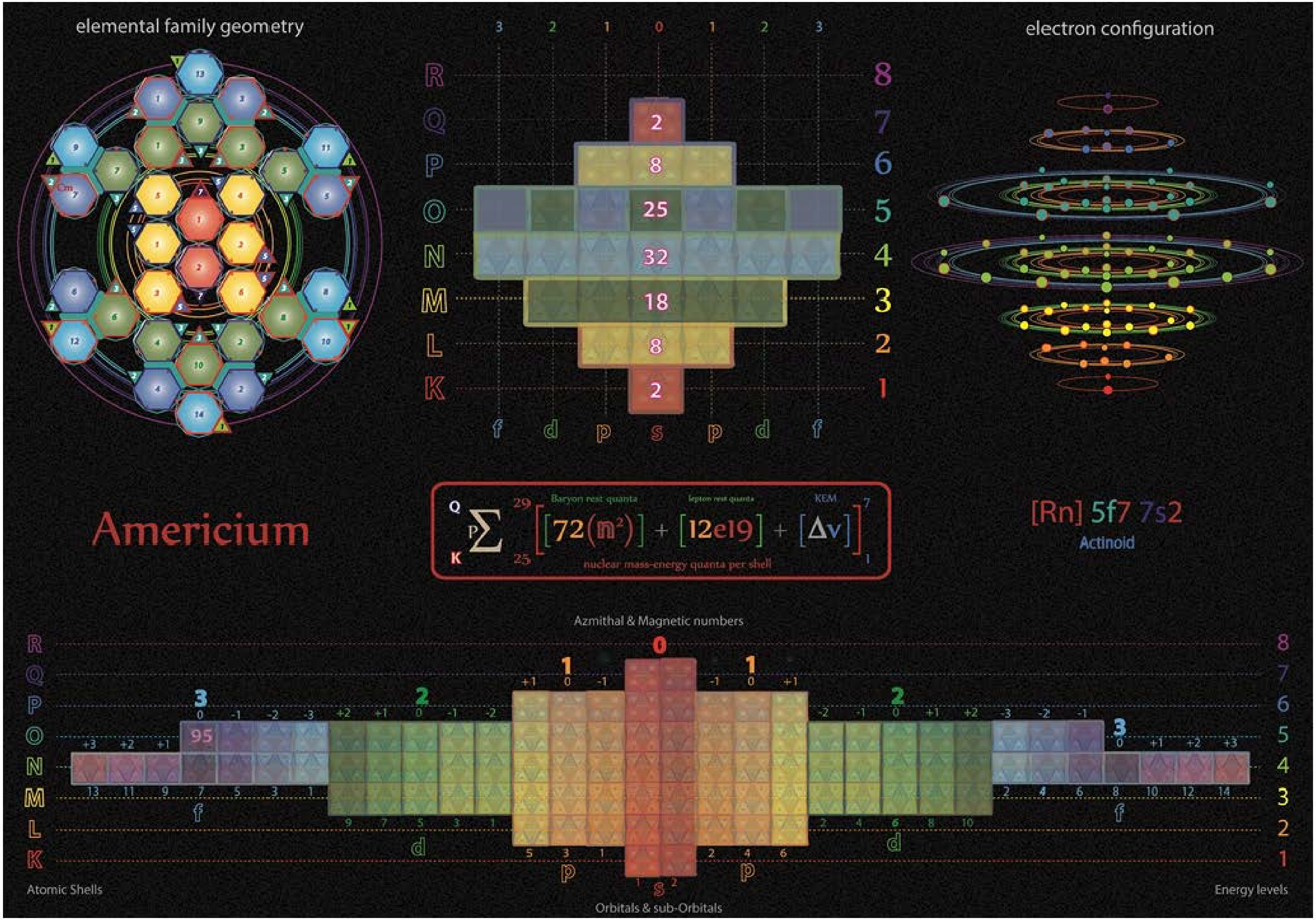
elemental family geometry
3 2 1 0 1 2 3
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
25
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
95
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Americium

Baryon rest quanta · lepton rest quanta · KEM

$$\sum_{K}^{Q} P \; {}^{29}_{25}\left[72(m^2)\right] + \left[12e19\right] + \left[\Delta v\right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f7 7s2

Actinoid

Tetryonics 53.95 - Americium atomic config

Copyright ABRAHAM [2008] - All rights reserved

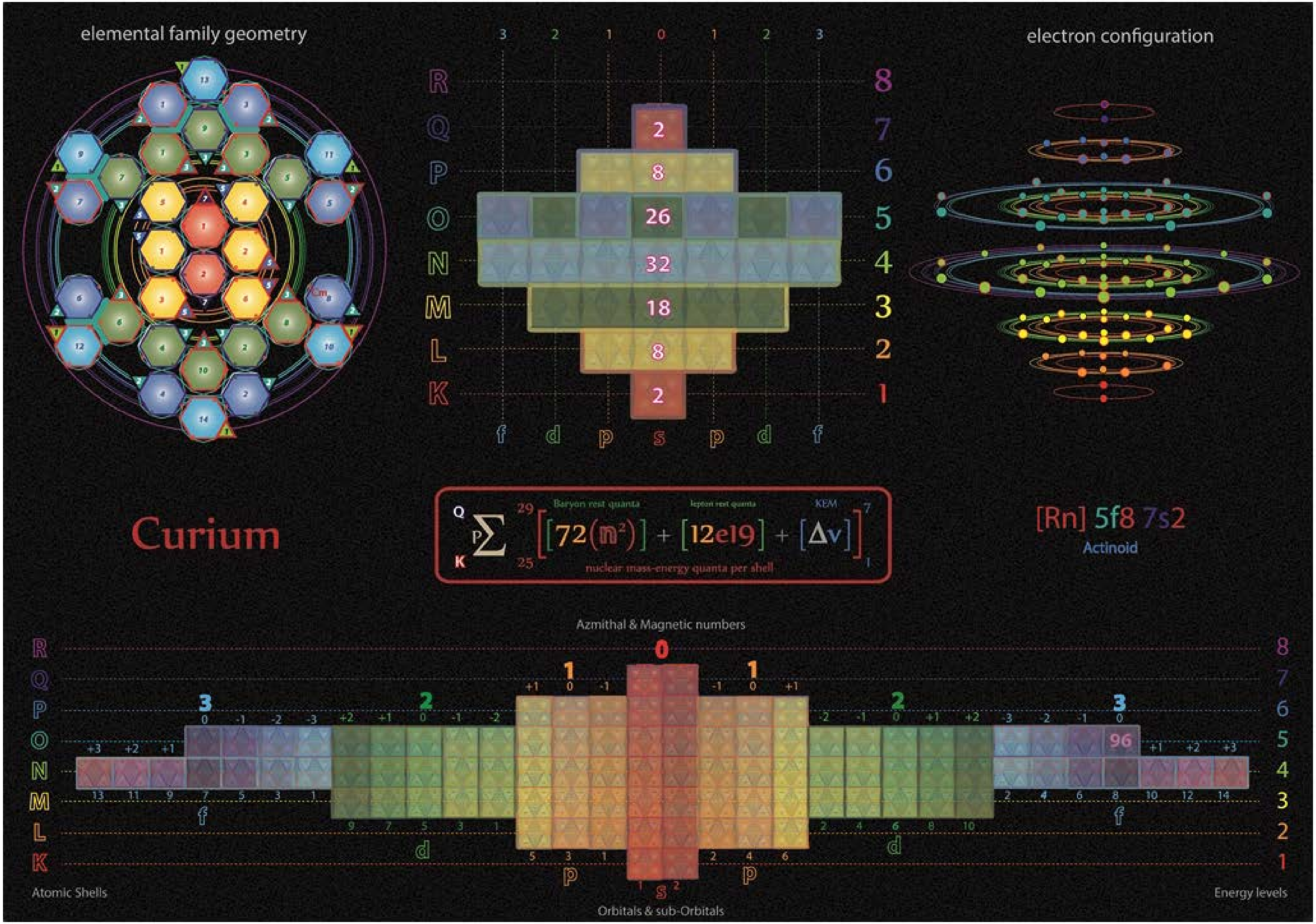
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
26
32
18
8
2
f d p s p d f
Curium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f8 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
0 -1 -2 -3 -3 -2 -1 0
+3 +2 +1 +1 +2 +3
96
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.96 - Curium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

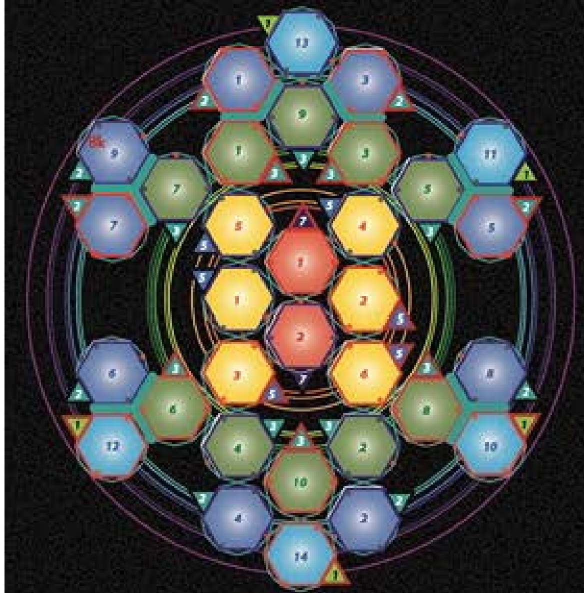
13
1
3
9
11
7
5
14
10
12
Bk

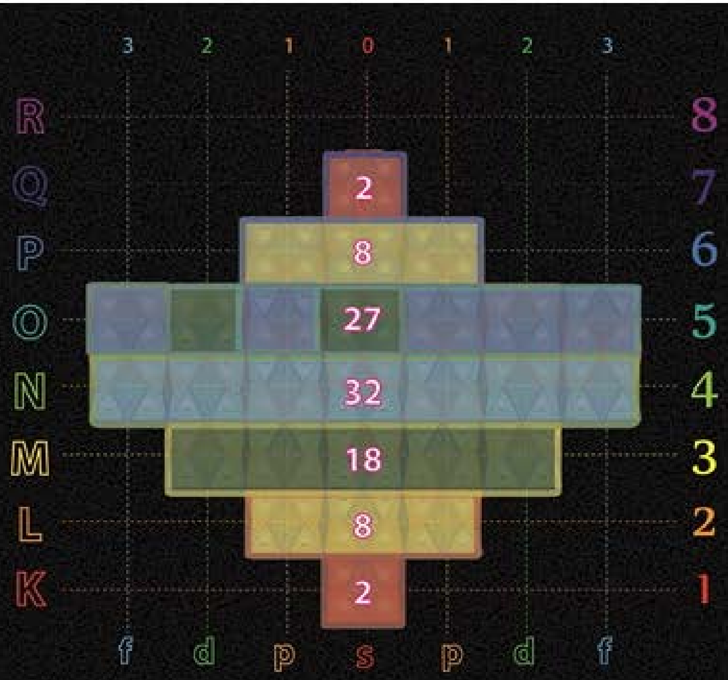
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
27
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

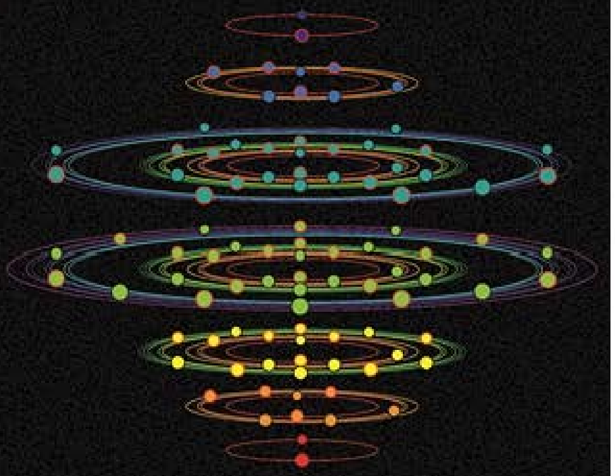

Berkelium

$$\sum_{K}^{Q} {}_{P}\,{}_{25}^{29}\Big[\,[72(\mathrm{m}^2)] + [12e19] + [\Delta v]\,\Big]_{1}^{7}$$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Rn] 5f9 7s2

Actinoid

Azmithal & Magnetic numbers

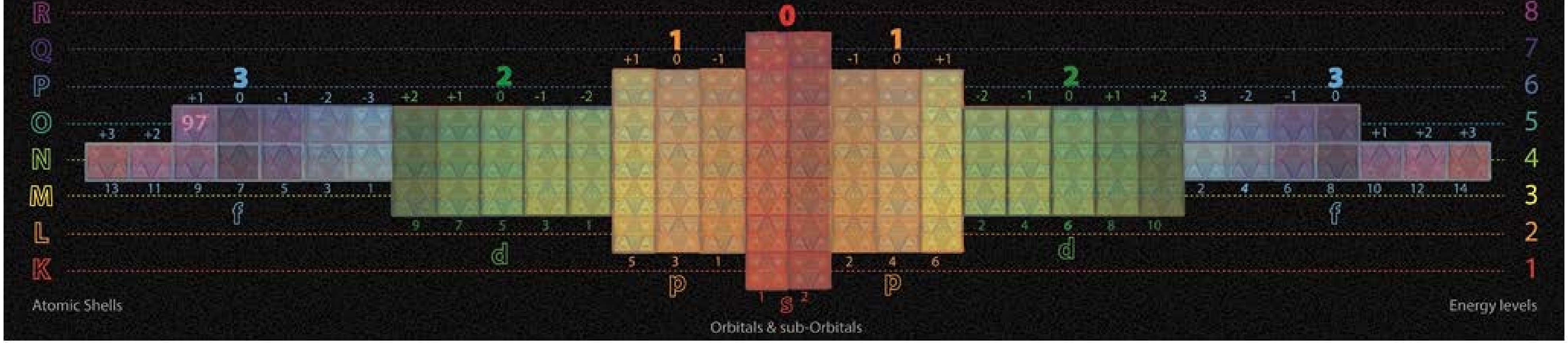
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+3
+2
+1
+2
+3
97
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
f
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.97 - Berkelium atomic config

Copyright ABRAHAM [2008] - All rights reserved

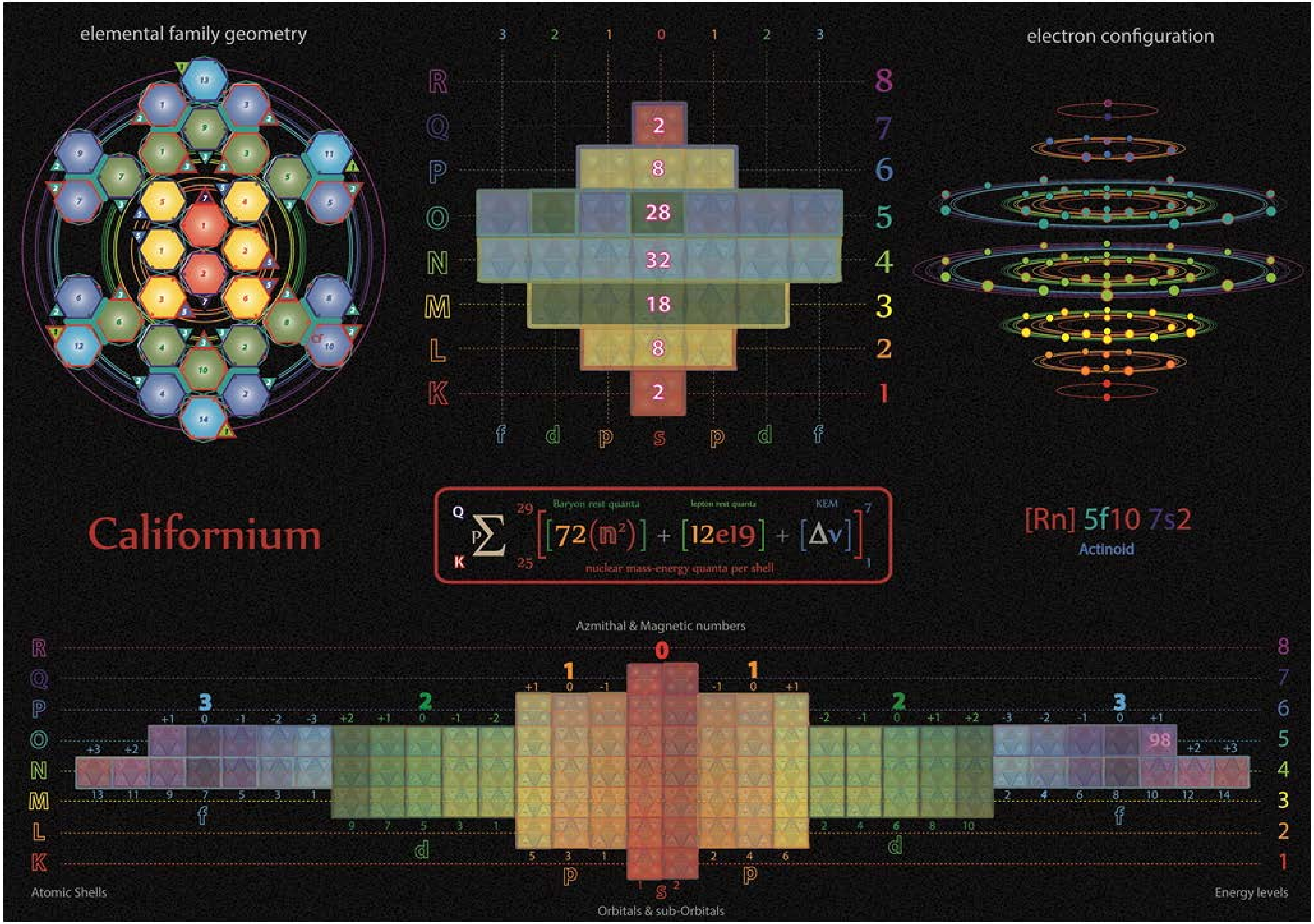
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
28
32
18
8
2
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
98
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Californium

$$ {}_{K}^{Q}\,{}_{P}\sum_{25}^{29}\left[\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]} + \overset{\text{lepton rest quanta}}{\left[12e19\right]} + \overset{\text{KEM}}{\left[\Delta v\right]}\right]_{1}^{7} $$

nuclear mass-energy quanta per shell

[Rn] 5f10 7s2

Actinoid

Tetryonics 53.98 - Californium atomic config

Copyright ABRAHAM [2008] - All rights reserved

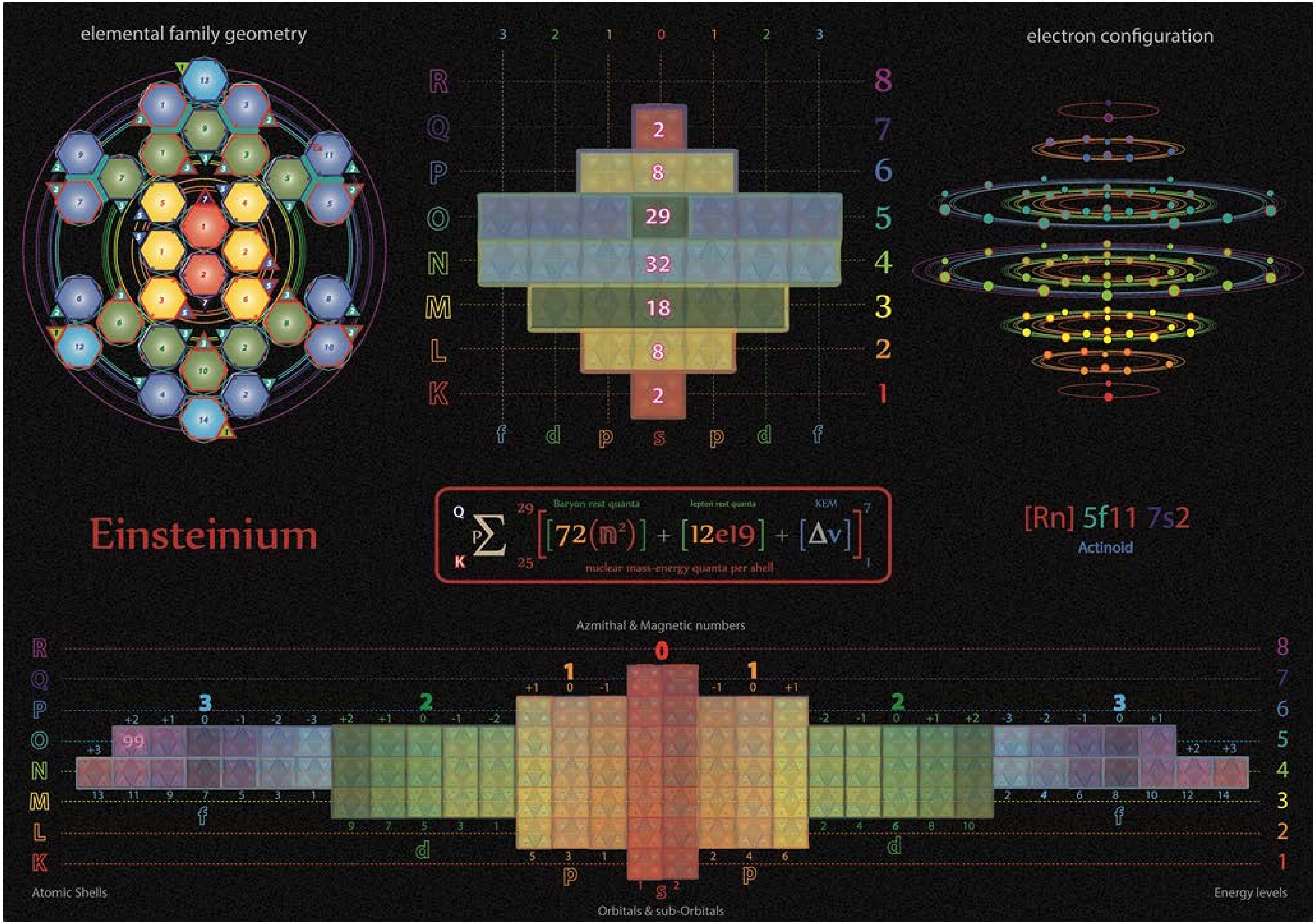
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
29
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1 -1 0 +1
99
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Einsteinium

$$ {}^{Q}_{K}\sum_{25}^{29} P \left[[72(m^2)] + [12e19] + [\Delta v] \right]_{1}^{7} $$

Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell

[Rn] 5f11 7s2
Actinoid

Tetryonics 53.99 - Einsteinium atomic config

Copyright ABRAHAM [2008] - All rights reserved

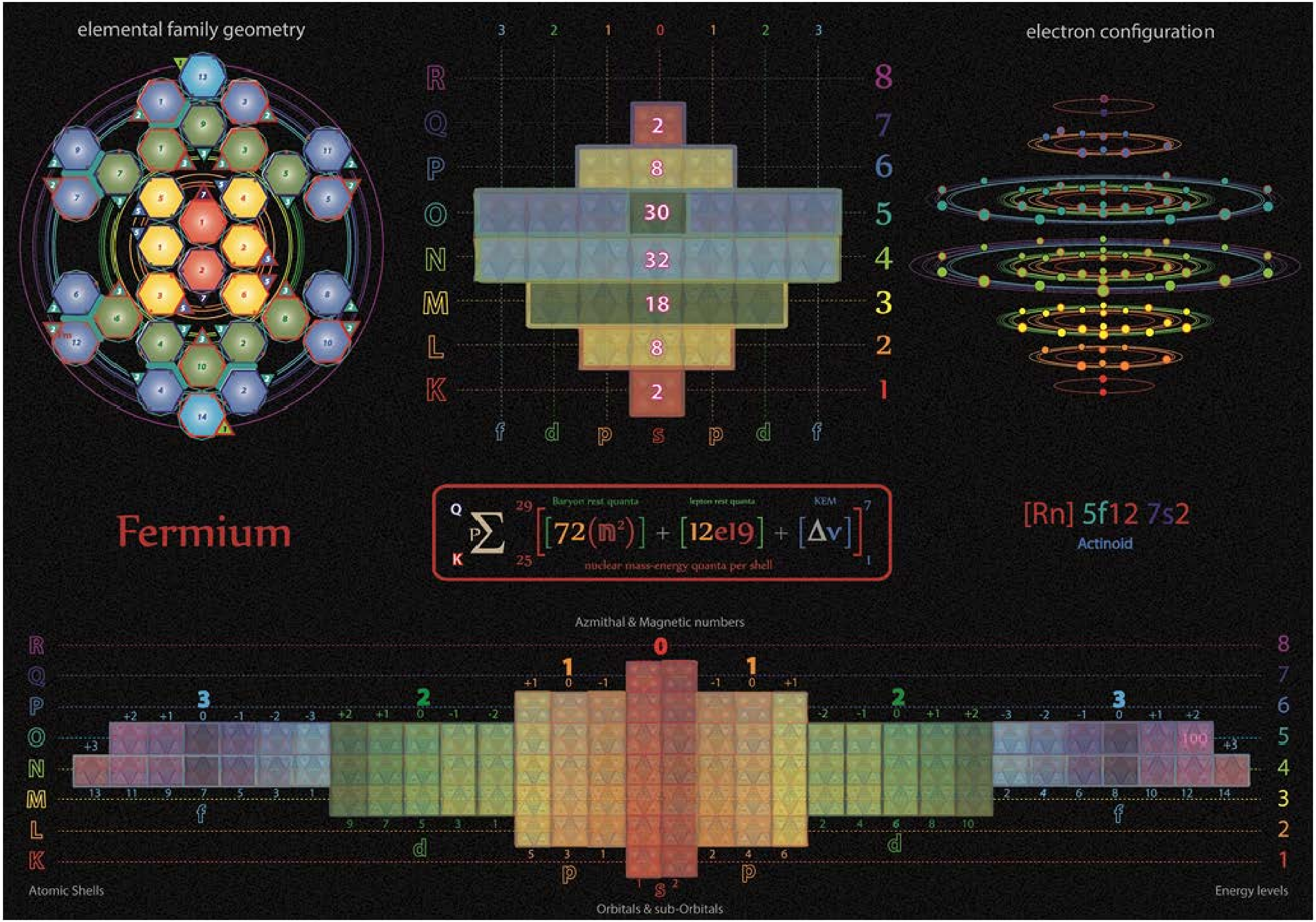
elemental family geometry
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
8
30
32
18
8
2
f d p s p d f
electron configuration
Fermium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{Q} P \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{25\,1}^{29\,7}$
nuclear mass-energy quanta per shell
[Rn] 5f12 7s2
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2
+3
+3
100
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.100 - Fermium atomic config

Copyright ABRAHAM [2008] - All rights reserved

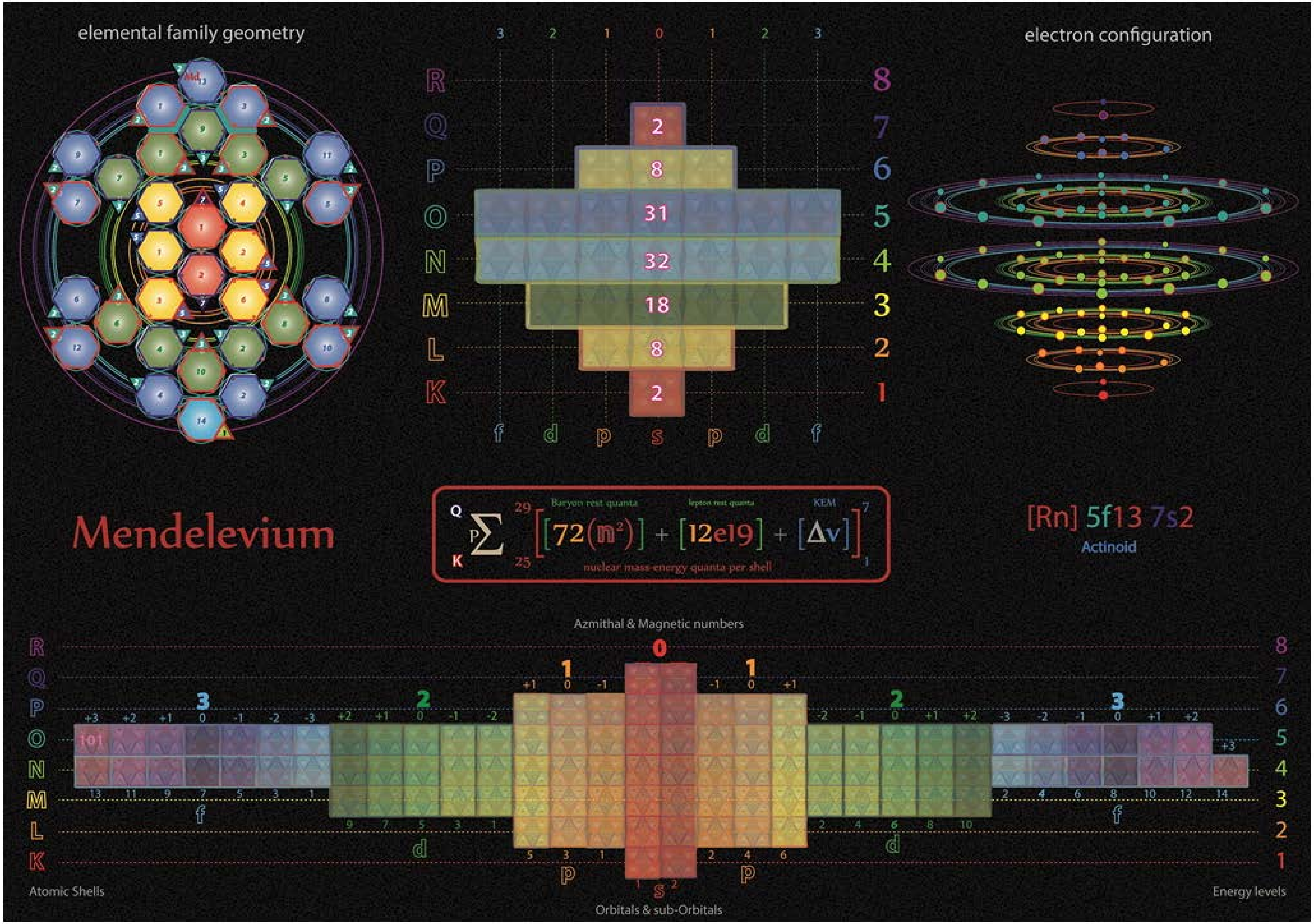
elemental family geometry
electron configuration
Mendelevium
[Rn] 5f13 7s2
Actinoid
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
2
8
31
32
18
8
2
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
Azmithal & Magnetic numbers
101
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.101 - Mendelevium atomic config

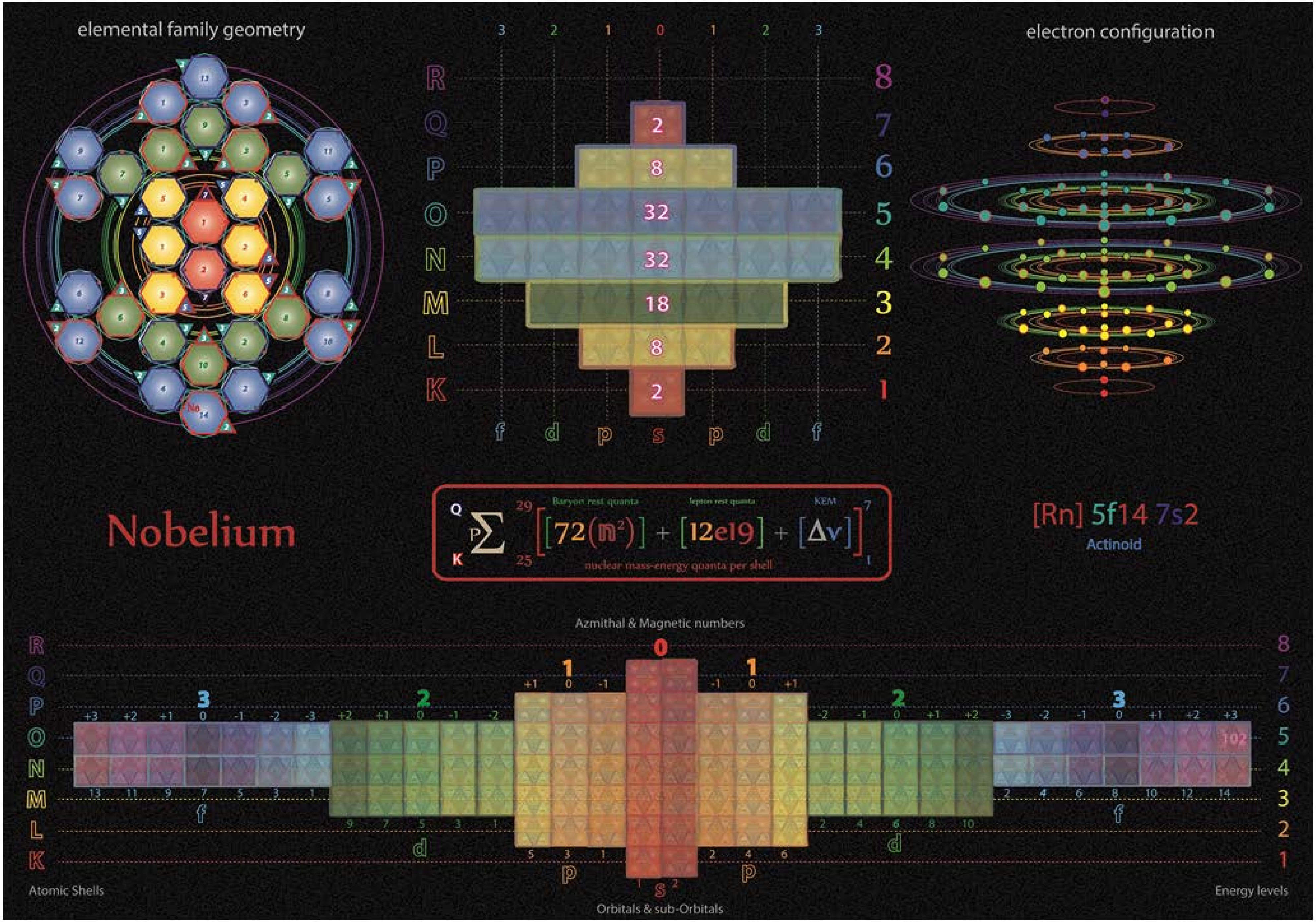
elemental family geometry
electron configuration
Nobelium
Q
K
P
Σ
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 7s2
Actinoid
R Q P O N M L K
8 7 6 5 4 3 2 1
2 8 32 32 18 8 2
f d p s p d f
3 2 1 0 1 2 3
Azmithal & Magnetic numbers
102
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.102 - Nobelium atomic config

Copyright ABRAHAM [2008] - All rights reserved

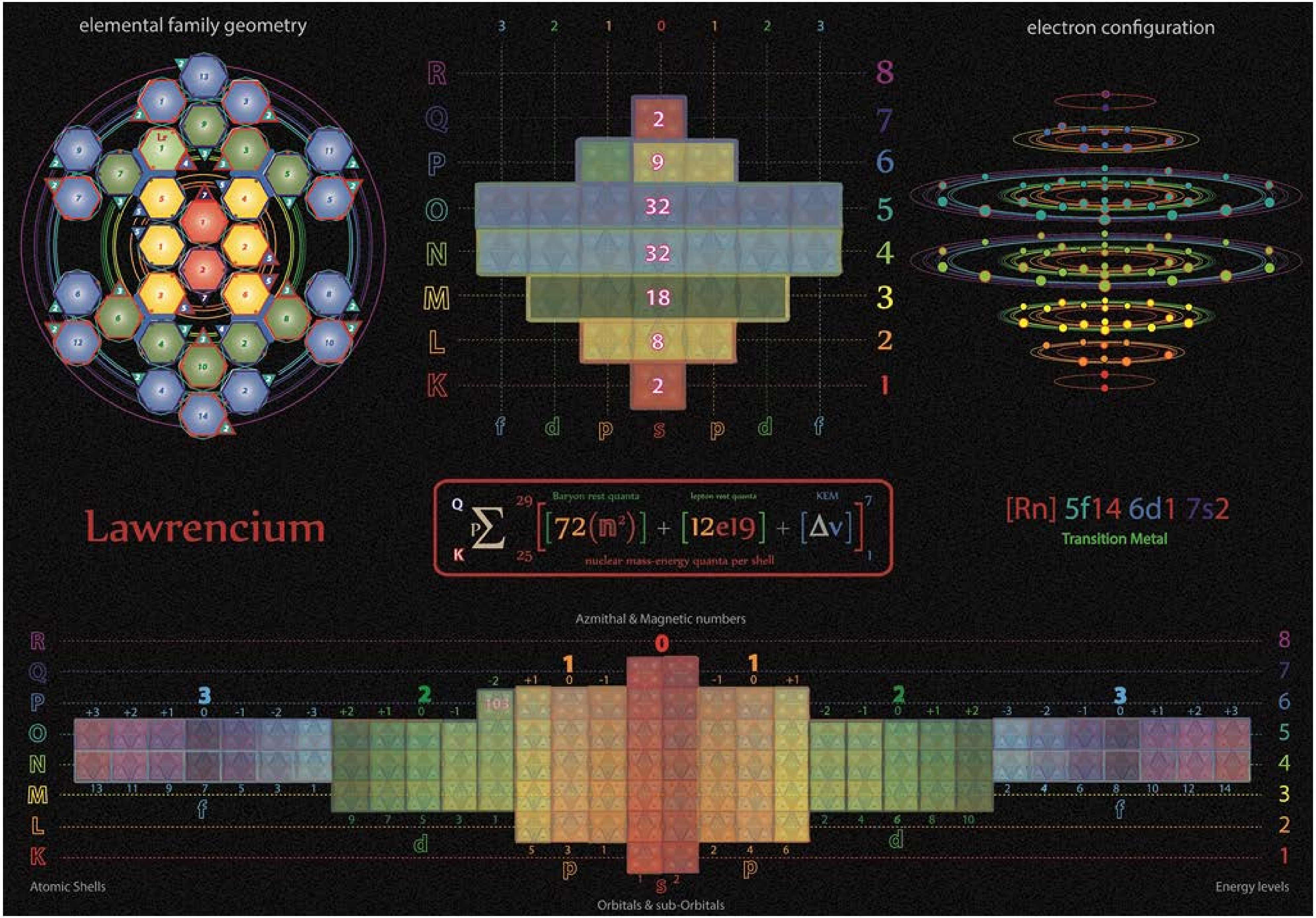
elemental family geometry
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
2 9 32 32 18 8 2
f d p s p d f
Lawrencium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(m²)] + [12e19] + [Δv]
7
1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d1 7s2
Transition Metal
Azmithal & Magnetic numbers
103
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.103 - Lawrencium atomic config

elemental family geometry

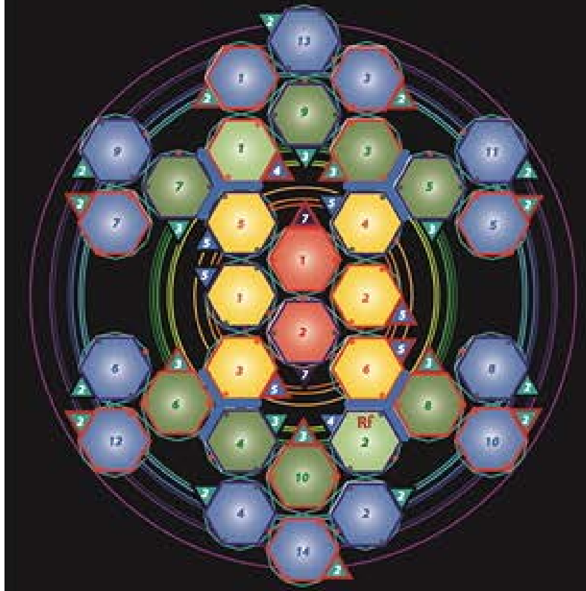

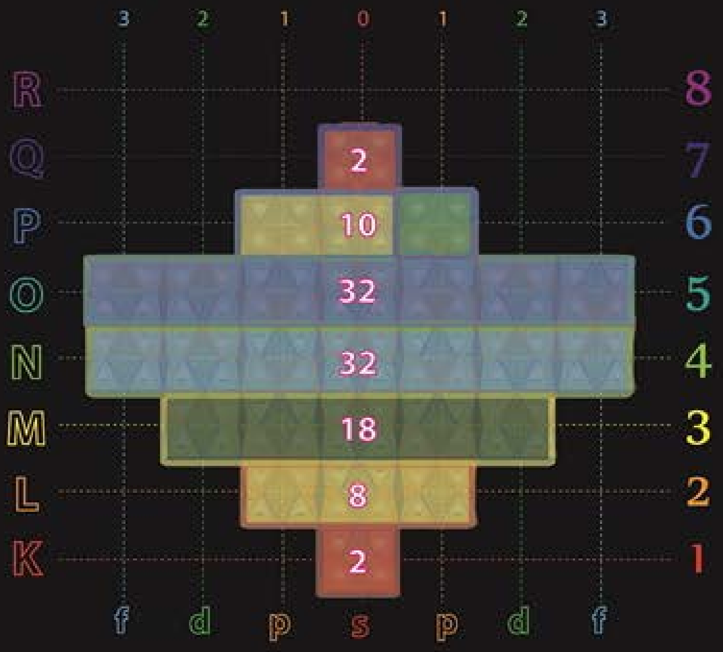
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
10
32
32
18
8
2
f d p s p d f

electron configuration

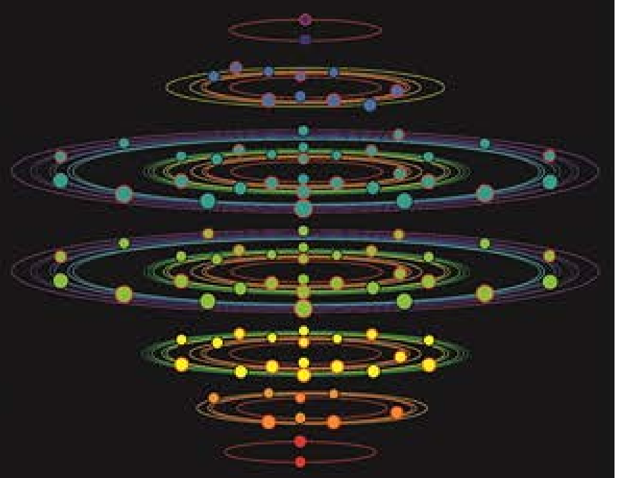

Rutherfordium

$$\sum_{K}^{Q} {}^{P}\;{}^{29}_{25}\left[\overset{\text{Baryon rest quanta}}{[72(m^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d2 7s2

Transition Metal

Azmithal & Magnetic numbers

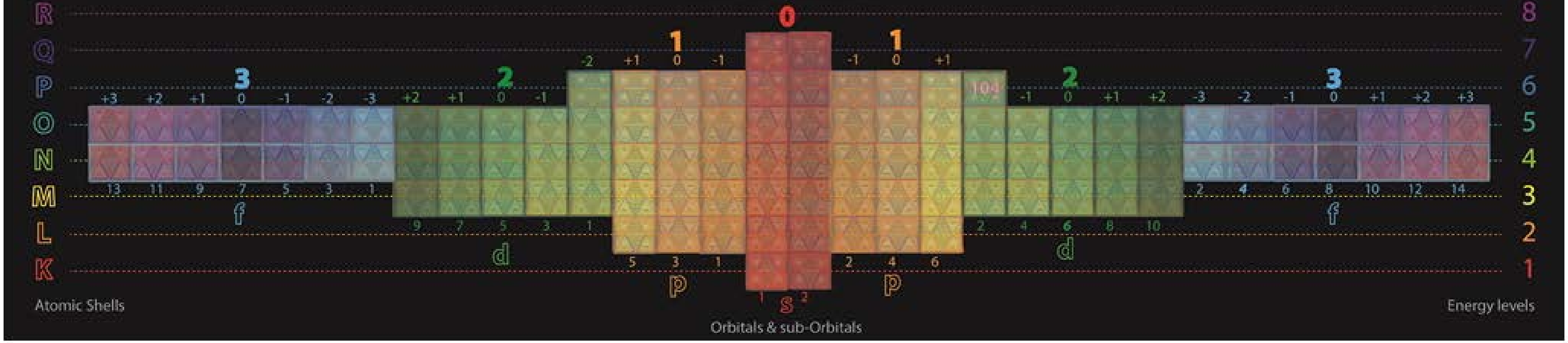
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
-2 +1 0 -1
-1 0 +1
-1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
104
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.104 - Rutherfordium atomic config

Copyright ABRAHAM [2008] - All rights reserved

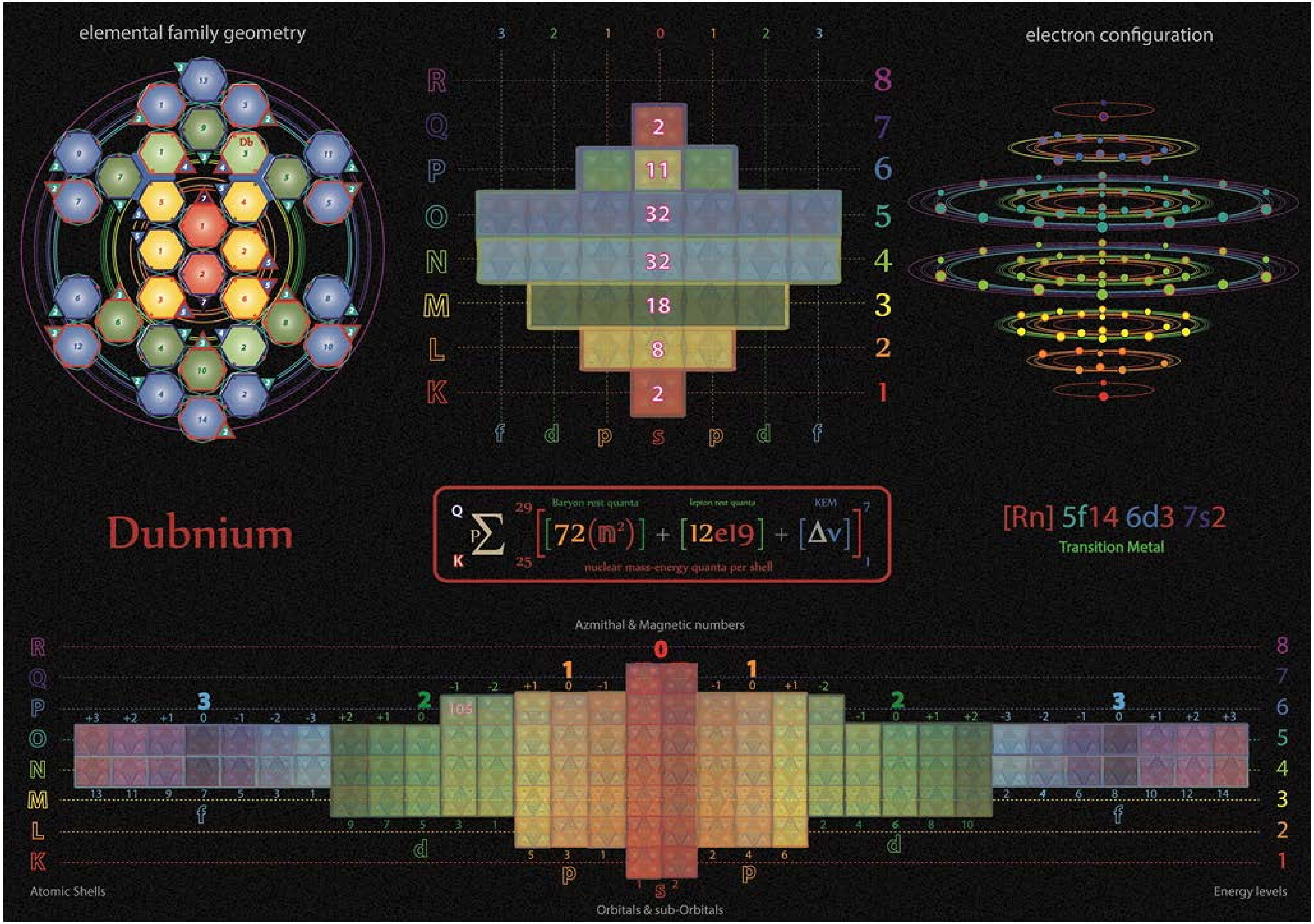
elemental family geometry
Db
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
11
32
32
18
8
2
f d p s p d f
Dubnium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f14 6d3 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
105
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

$$\sum_{K}^{Q}\,{}^{P}\!\!\sum_{25}^{29}\Big[[72(m^2)] + [12e19] + [\Delta v]\Big]_1^7$$

Tetryonics 53.105 - Dubnium atomic config

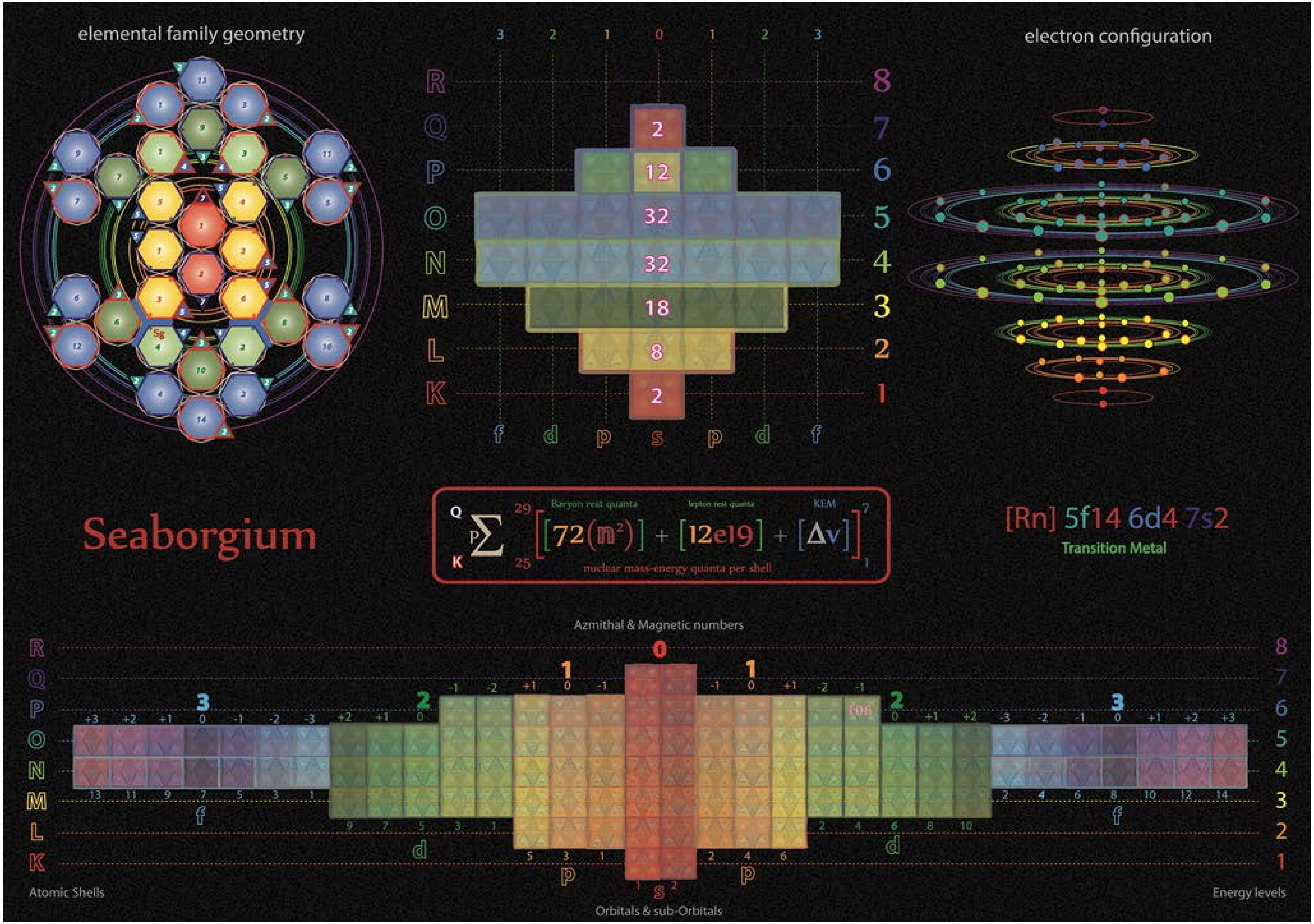
elemental family geometry
Sg
electron configuration
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
2
12
32
32
18
8
2
f d p s p d f
Seaborgium
Q
P
K
29
25
Baryon rest quanta
lepton rest quanta
KEM
[72(n²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
7
1
[Rn] 5f14 6d4 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
106
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.106 - Seaborgium atomic config

Copyright ABRAHAM [2008] - All rights reserved

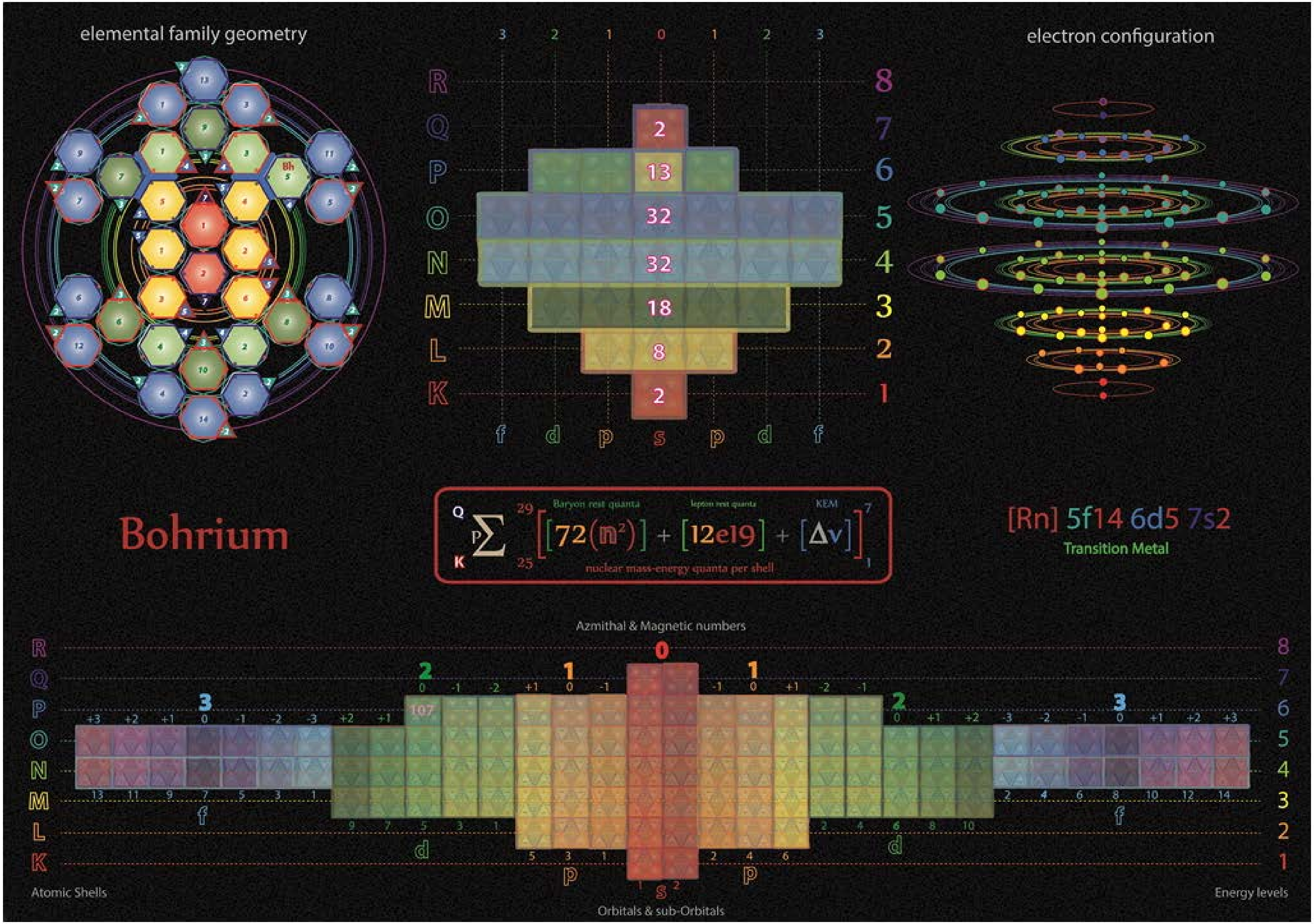
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2
13
32
32
18
8
2
f d p s p d f
Azmithal & Magnetic numbers
0
1
2
3
107
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Bohrium

$$\sum_{K}^{Q} P \; {}^{29}_{25}\left[[72(n^2)] + [12e19] + [\Delta v] \right]^{7}_{1}$$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d5 7s2

Transition Metal

Tetryonics 53.107 - Bohrium atomic config

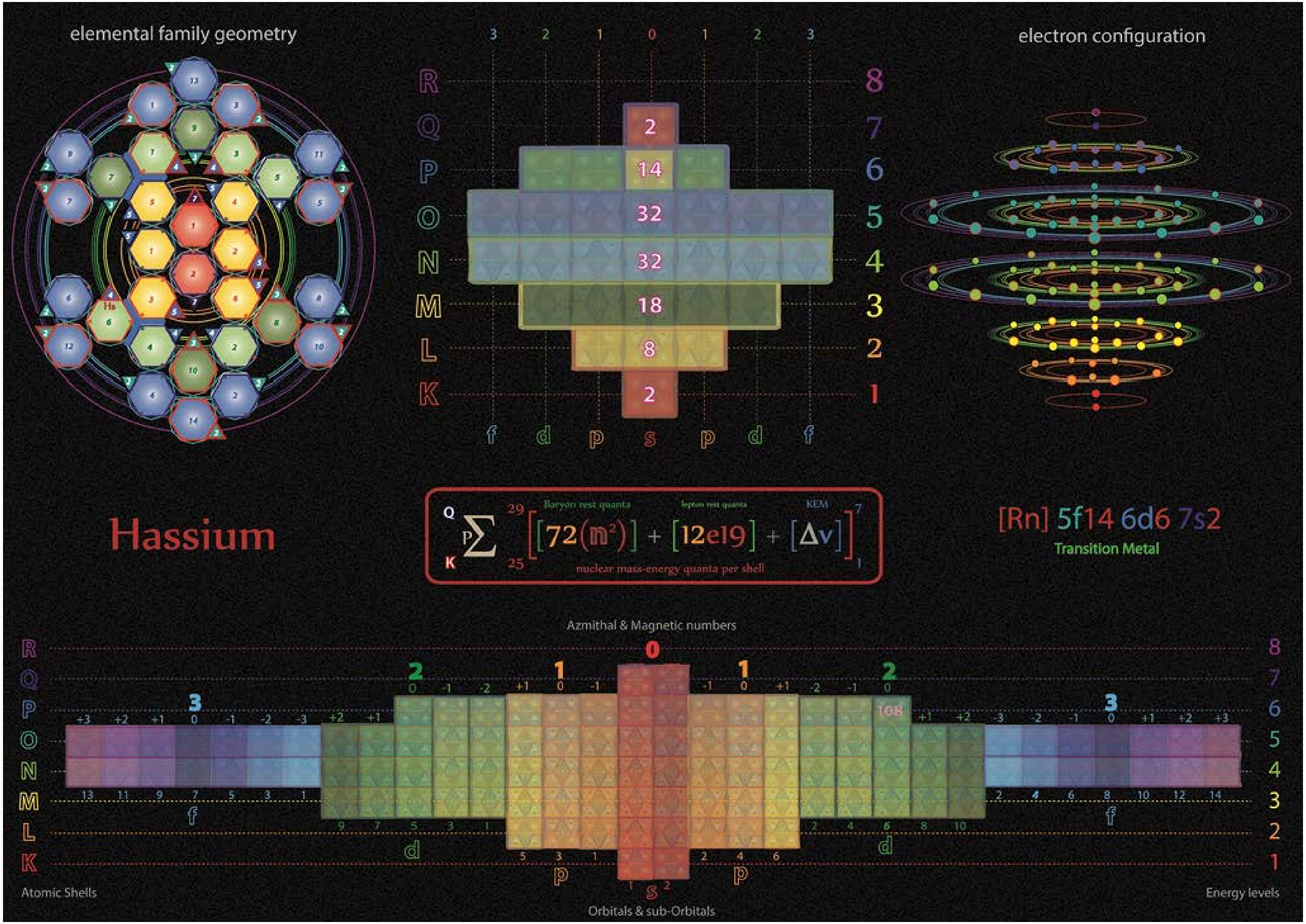

Hassium

$$ {}_{K}^{Q}\,{}_{P}\sum_{25}^{29}\left[\left[72(\mathrm{m}^{2})\right]+\left[12e19\right]+\left[\Delta v\right]\right]_{1}^{7} $$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d6 7s2

Transition Metal

elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
2 14 32 32 18 8 2
f d p s p d f
Azmithal & Magnetic numbers
0 1 2 3 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
108
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.108 - Hassium atomic config

elemental family geometry

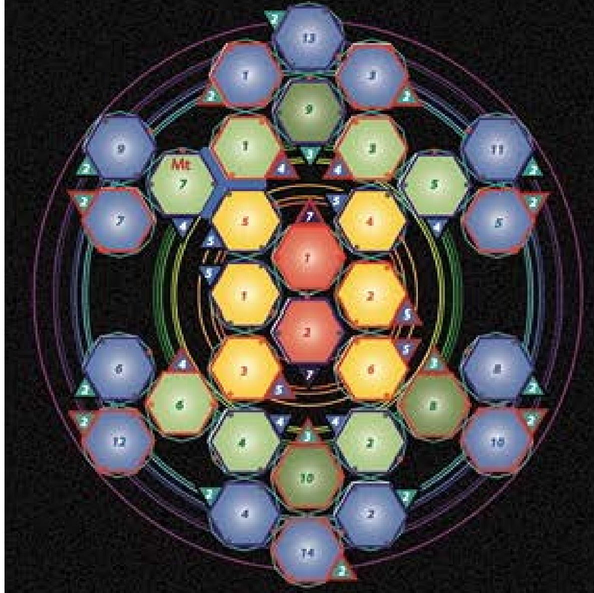
Mt
13
1
3
9
11
7
5
1
2
6
8
12
10
4
14

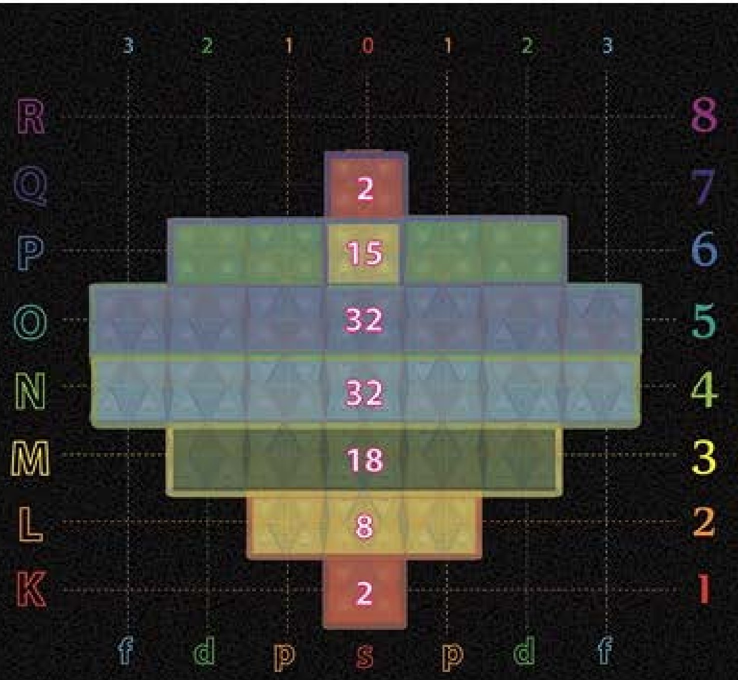
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
15
32
32
18
8
2
f d p s p d f

electron configuration

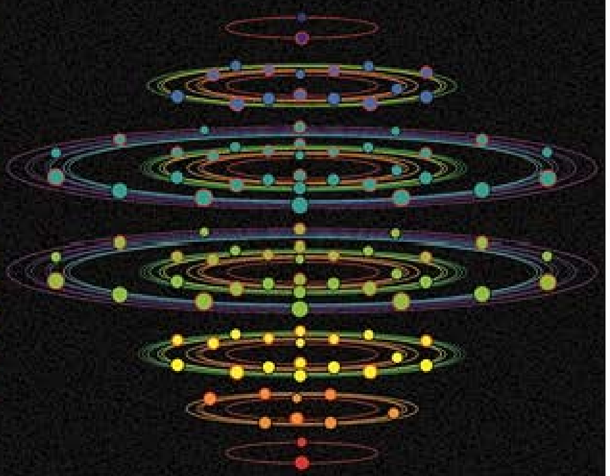

Meitnerium

$$ {}^{Q}_{K}\;{}^{P}\!\sum^{29}_{25}\left[\left[72(\mathrm{m}^2)\right] + [12e19] + [\Delta v]\right]^{7}_{1} $$

Baryon rest quanta

lepton rest quanta

KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d7 7s2

Transition Metal

Azmithal & Magnetic numbers

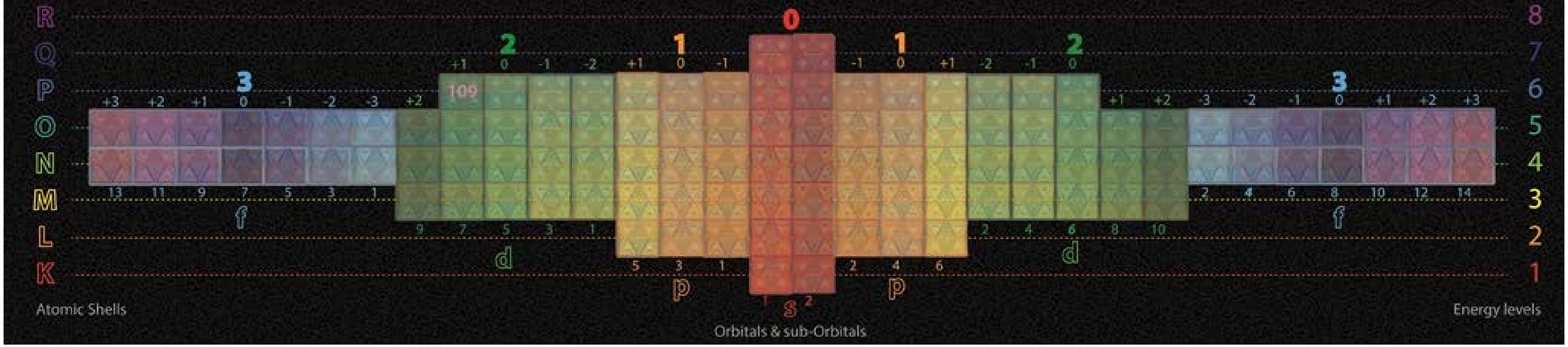
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
109
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f
d
p
s
p
d
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.109 - Meitnerium atomic config

Copyright ABRAHAM [2008] - All rights reserved

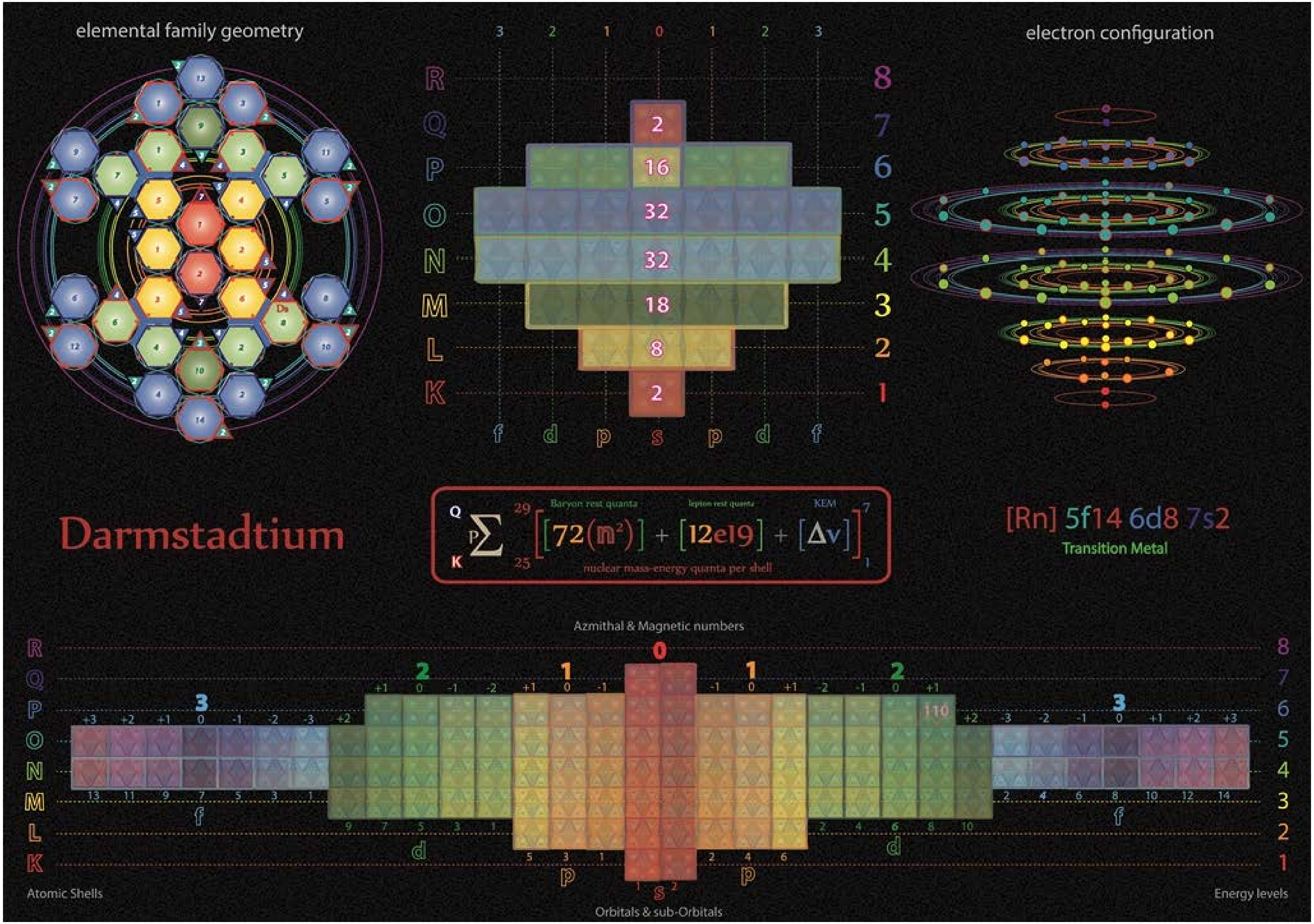
elemental family geometry
Da
3 2 1 0 1 2 3
R Q P O N M L K
2
16
32
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Darmstadtium
Q
K
P
29
25
Baryon rest quanta
lepton rest quanta
KEM
7
1
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
[Rn] 5f14 6d8 7s2
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
110
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.110 - Darmstadtium atomic config

Copyright ABRAHAM [2008] - All rights reserved

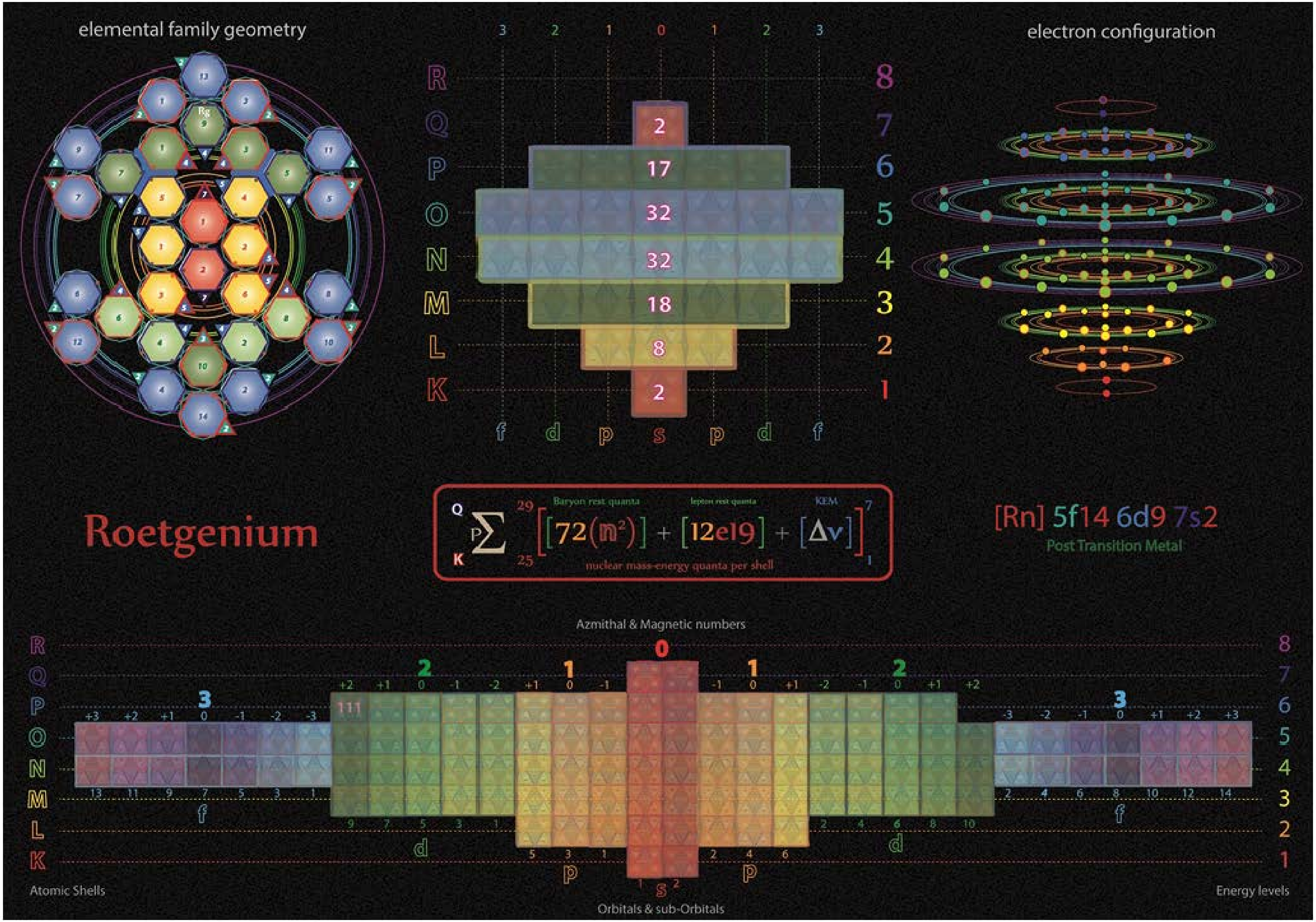
elemental family geometry
Rg
3 2 1 0 1 2 3
R Q P O N M L K
2
17
32
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
111
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Roetgenium

$$ {}_{K}^{Q}\,{}_{P}\sum_{25}^{29} \left[[72(n^2)] + [12e19] + [\Delta v] \right]_{1}^{7} $$

Baryon rest quanta · lepton rest quanta · KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d9 7s2

Post Transition Metal

Tetryonics 53.111 - Roetgenium atomic config

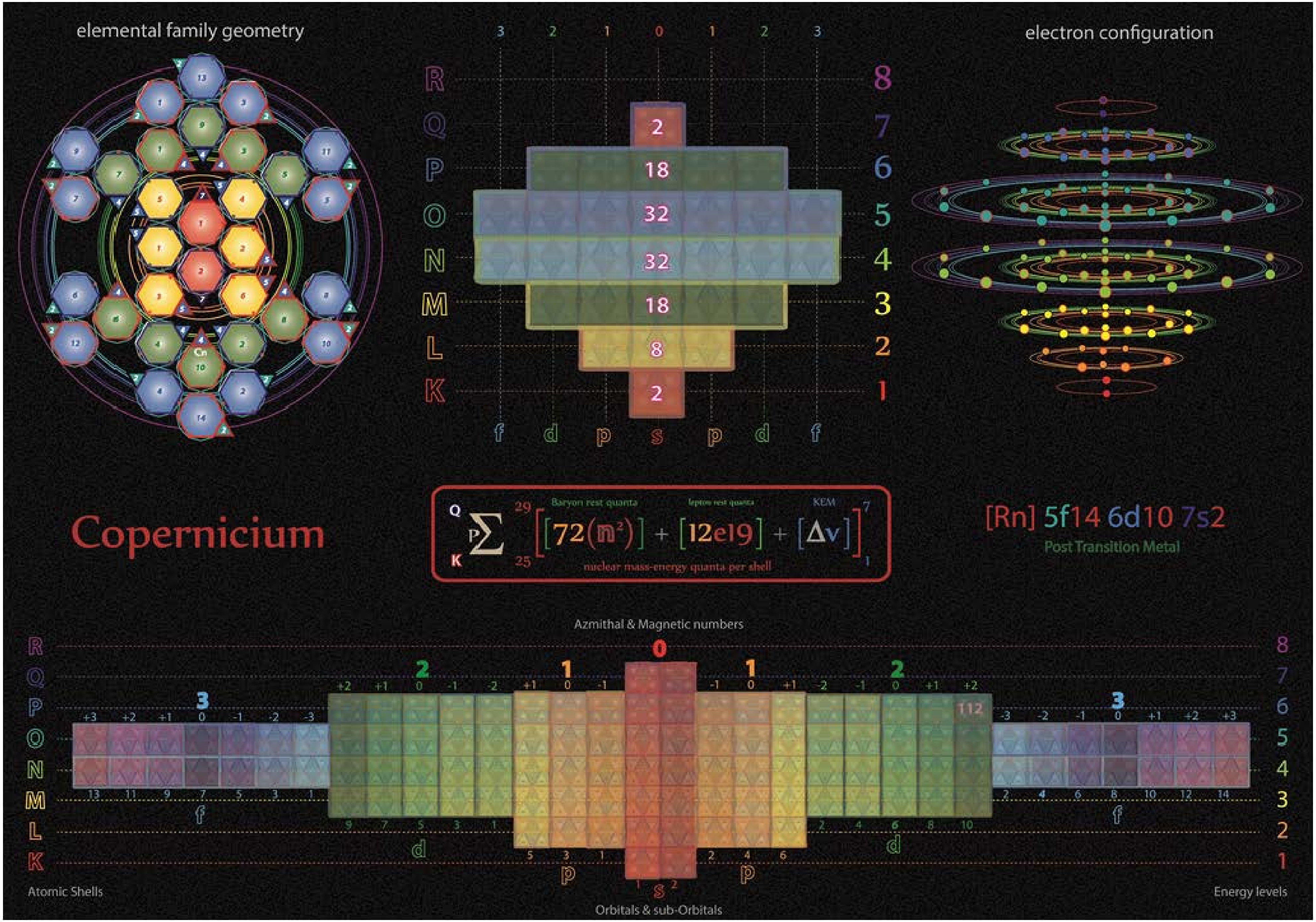

elemental family geometry
Cn
electron configuration
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
Azmithal & Magnetic numbers
0
1
2
3
1
2
3
+3
+2
+1
0
-1
-2
-3
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
112
-3
-2
-1
0
+1
+2
+3
13
11
9
7
5
3
1
9
7
5
3
1
5
3
1
1
2
2
4
6
2
4
6
8
10
2
4
6
8
10
12
14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Copernicium

Baryon rest quanta
lepton rest quanta
KEM

$${}^{Q}_{K}{}_{P}\sum_{25}^{29}\left[[72(\mathrm{n}^{2})]+[12e19]+[\Delta v]\right]_{1}^{7}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2

Post Transition Metal

Tetryonics 53.112 - Copernicium atomic config

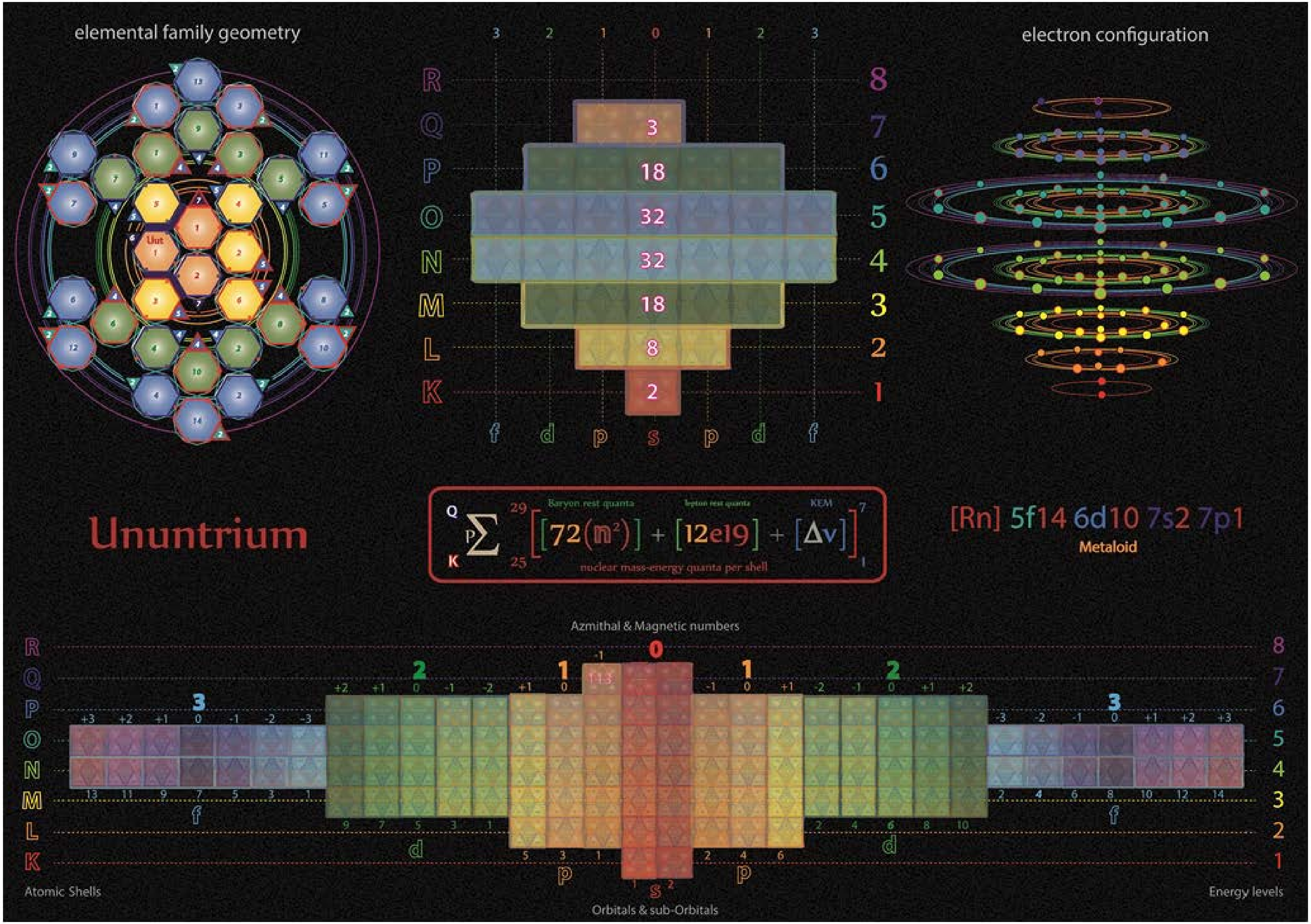
elemental family geometry
Uut
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
3
18
32
32
18
8
2
f d p s p d f
electron configuration
Ununtrium
Baryon rest quanta
lepton rest quanta
KEM
nuclear mass-energy quanta per shell
[Rn] 5f14 6d10 7s2 7p1
Metaloid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
113
Atomic Shells
Energy levels
f d p s p d f
Orbitals & sub-Orbitals

$$\sum_{K}^{Q}{}_{P}\;{}^{29}_{25}\left[[72(m^2)]+[12e19]+[\Delta v]\right]^{7}_{1}$$

Tetryonics 53.113 - Ununtrium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

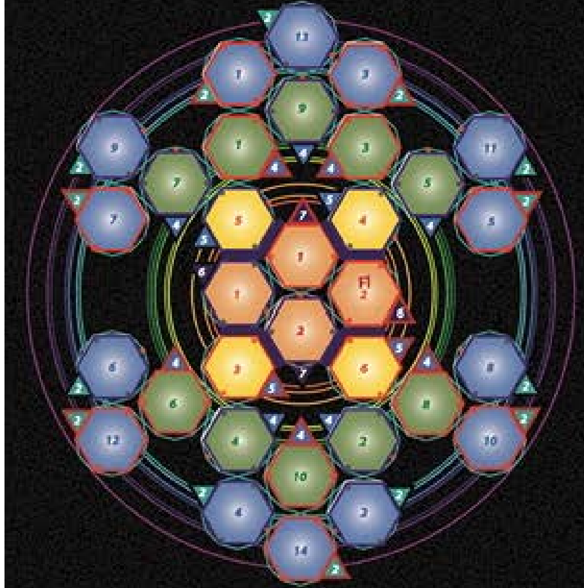

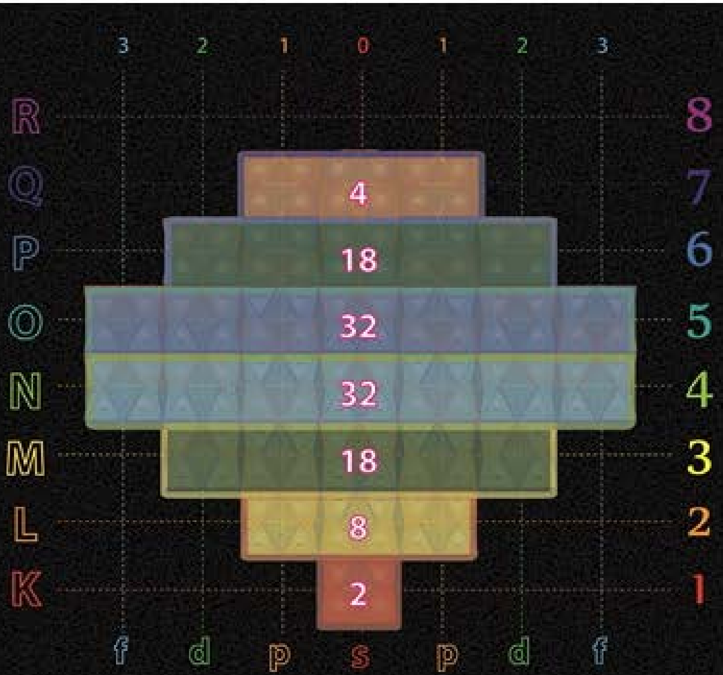
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
4
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

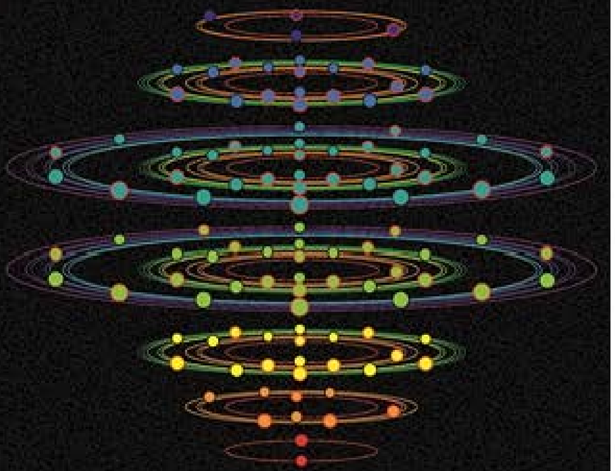

Flerovium

$$\overset{Q}{\underset{K}{}}{}_{P}\sum_{25}^{29}\left[\underbrace{\overbrace{[72(n^2)]}^{\text{Baryon rest quanta}}+\overbrace{[12e19]}^{\text{lepton rest quanta}}+\overbrace{[\Delta v]}^{\text{KEM}}}_{\text{nuclear mass-energy quanta per shell}}\right]_{1}^{7}$$

[Rn] 5f14 6d10 7s2 7p2

Metaloid

Azmithal & Magnetic numbers

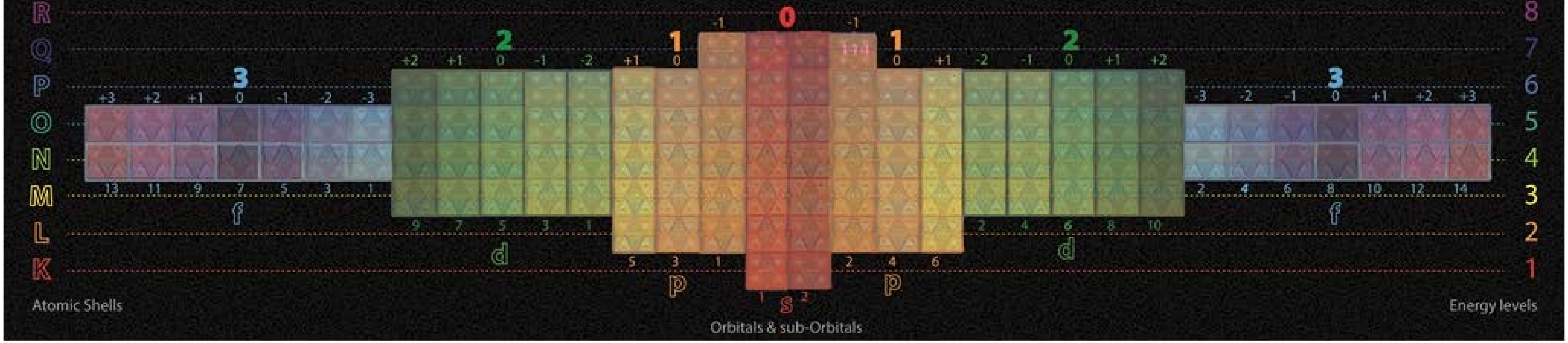
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
-1
-1
1
1
2
2
3
3
+2
+1
0
-1
-2
+1
0
0
+1
-2
-1
0
+1
+2
+3
+2
+1
0
-1
-2
-3
-3
-2
-1
0
+1
+2
+3
114
13
11
9
7
5
3
1
f
9
7
5
3
1
d
5
3
1
p
1
s
2
2
4
6
p
2
4
6
8
10
d
2
4
6
8
10
12
14
f
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.114 - Flerovium atomic config

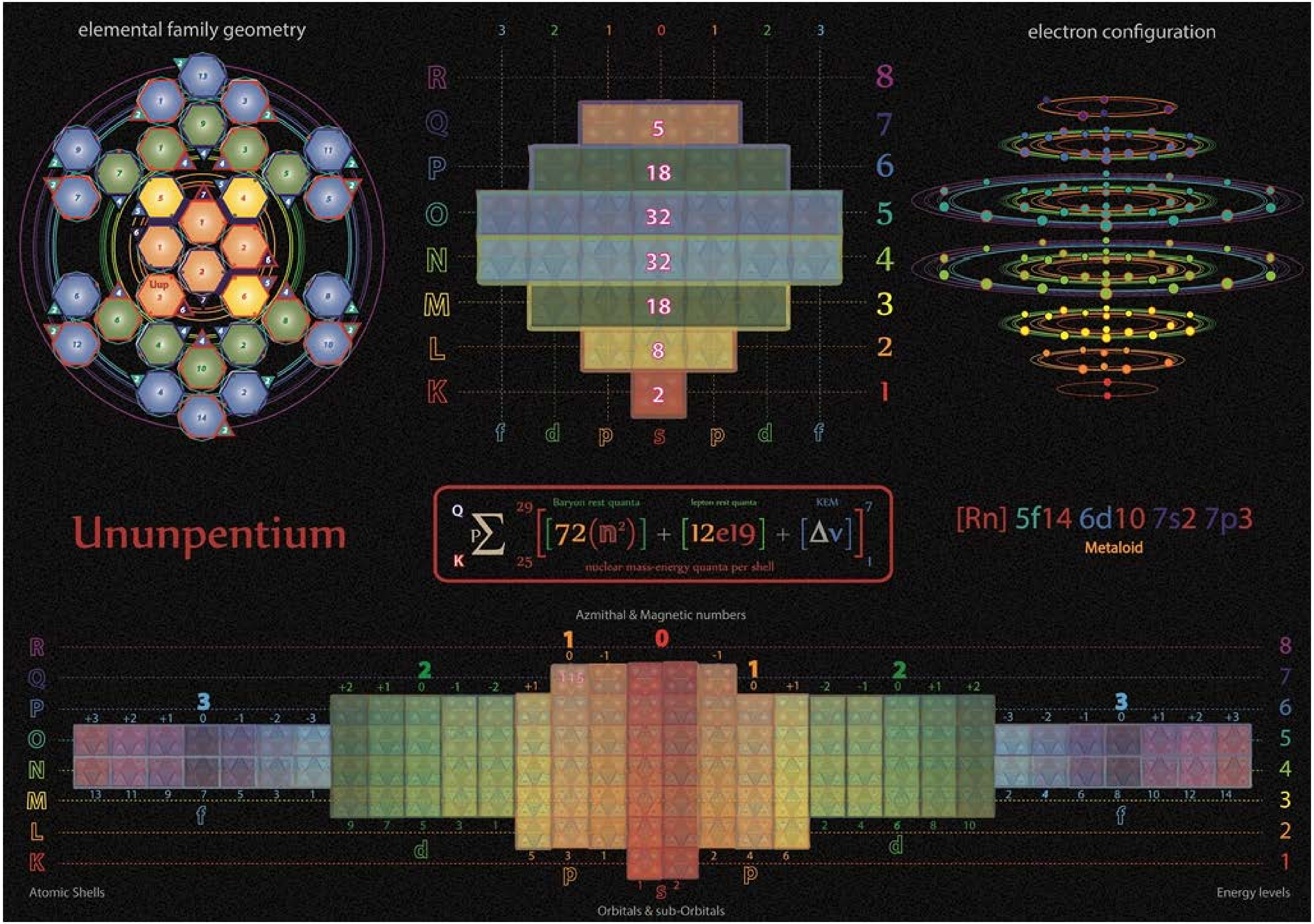
elemental family geometry
Uup
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
5
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Azmithal & Magnetic numbers
1
0
2
1
2
3
3
+3
+2
+1
0
-1
-2
-3
-1
115
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Ununpentium

$$\sum_{K}^{Q} P \sum_{25}^{29} \left[\overset{\text{Baryon rest quanta}}{[72(n^2)]} + \overset{\text{lepton rest quanta}}{[12e19]} + \overset{\text{KEM}}{[\Delta v]} \right]_{1}^{7}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p3

Metaloid

Tetryonics 53.115 - Ununpentium atomic config

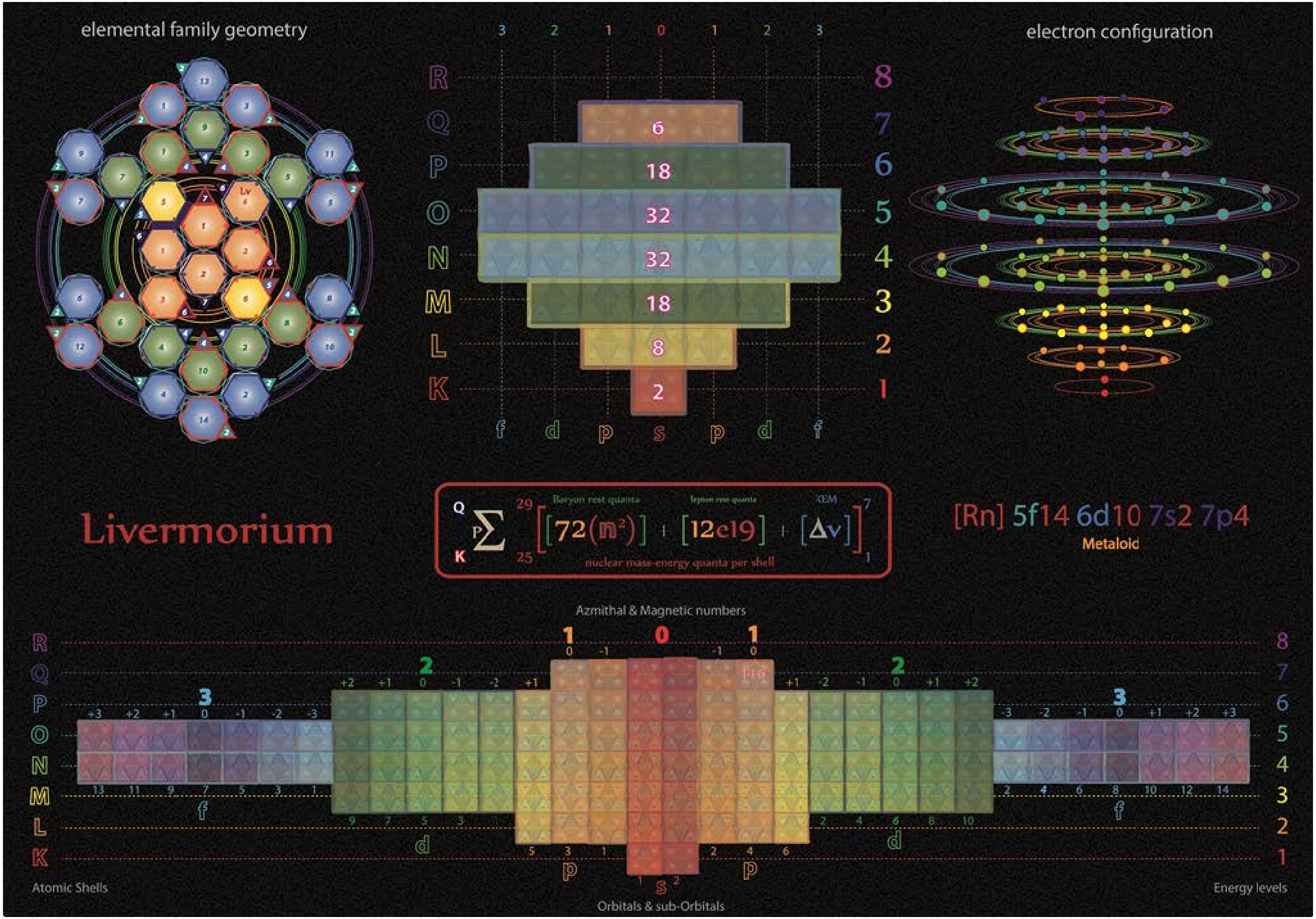
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
6
18
32
32
18
8
2
f d p s p d f
Livermorium
Q P K Σ 29 25
Baryon rest quanta
lepton rest quanta
ΣEM
[72(n²)] + [12e19] + [Δv]
7 1
nuclear mass-energy quanta per shell
[Rn] 5f14 6d10 7s2 7p4
Metaloid
Azmithal & Magnetic numbers
3 2 1 0 1 2 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
116
13 11 9 7 5 3 1
9 7 5 3
5 3 1
1 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p s p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.116 - Livermorium atomic config

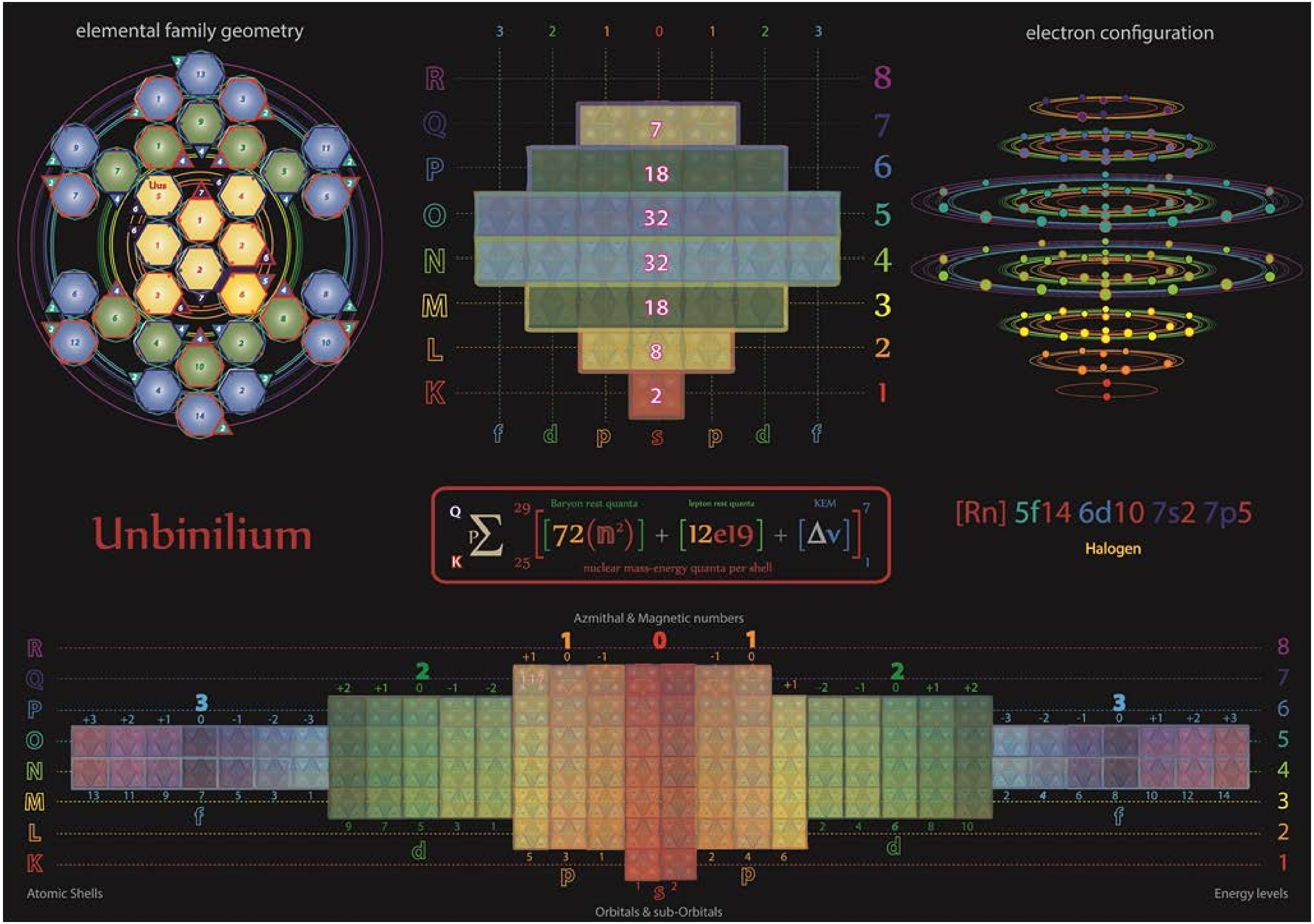
elemental family geometry
Uus
3 2 1 0 1 2 3
R Q P O N M L K
7
18
32
32
18
8
2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
1 0 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Unbinilium

$$ {}_{K}^{Q}\,{}_{P}\!\sum_{25}^{29}\left[\,[72(\mathfrak{m}^2)] + [12e19] + [\Delta v]\,\right]_1^7 $$

Baryon rest quanta — lepton rest quanta — KEM

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p5

Halogen

Tetryonics 53.117 - Ununseptium atomic config

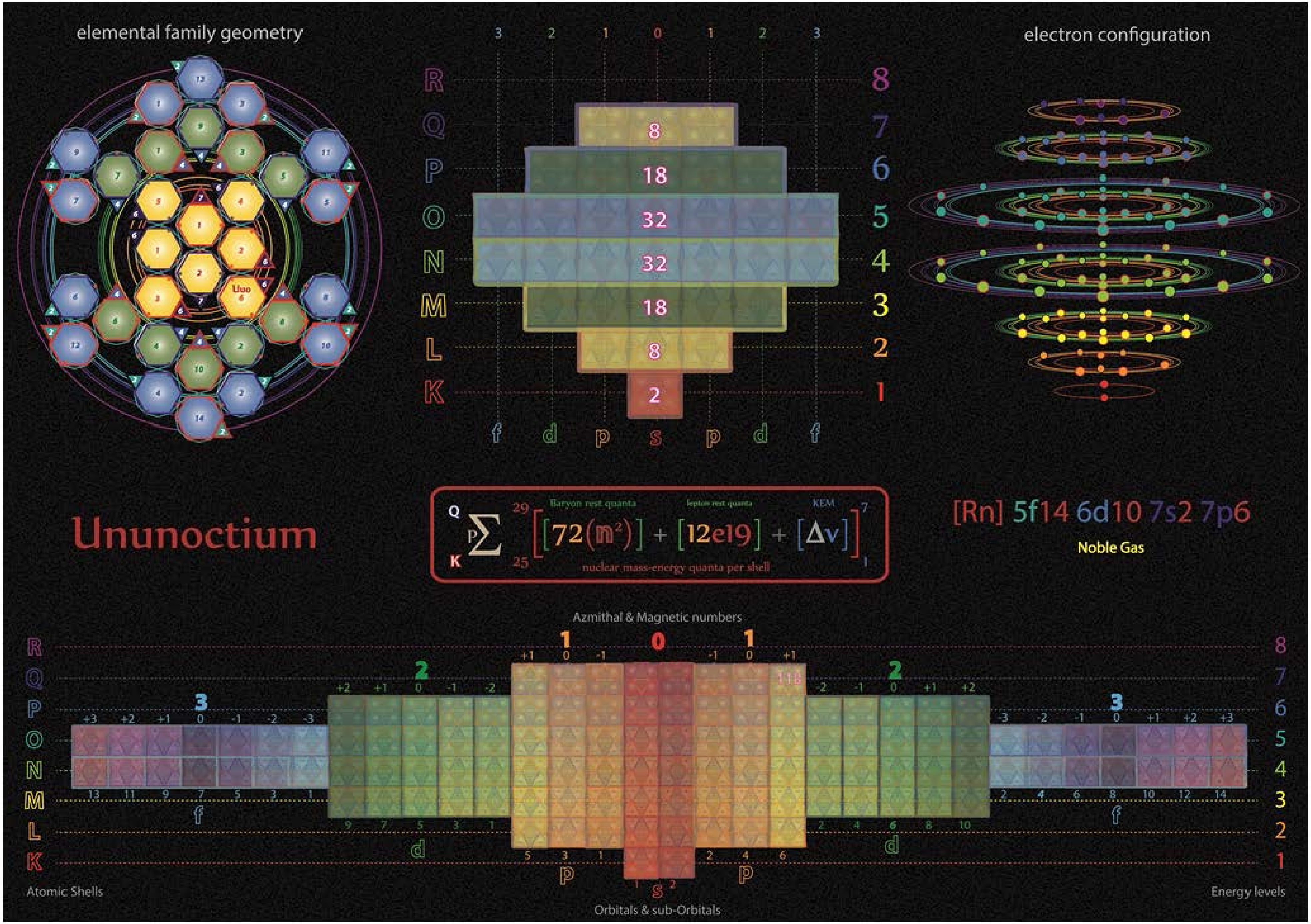

elemental family geometry
Uuo
3 2 1 0 1 2 3
R Q P O N M L K
8 18 32 32 18 8 2
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
118
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Ununoctium

Baryon rest quanta — lepton rest quanta — KEM

$$\sum_{K}^{Q} P \; {}^{29}_{25}\left[[72(n^2)] + [12e19] + [\Delta v] \right]^{7}_{1}$$

nuclear mass-energy quanta per shell

[Rn] 5f14 6d10 7s2 7p6

Noble Gas

Tetryonics 53.118 - Ununoctium atomic config

Copyright ABRAHAM [2008] - All rights reserved

elemental family geometry

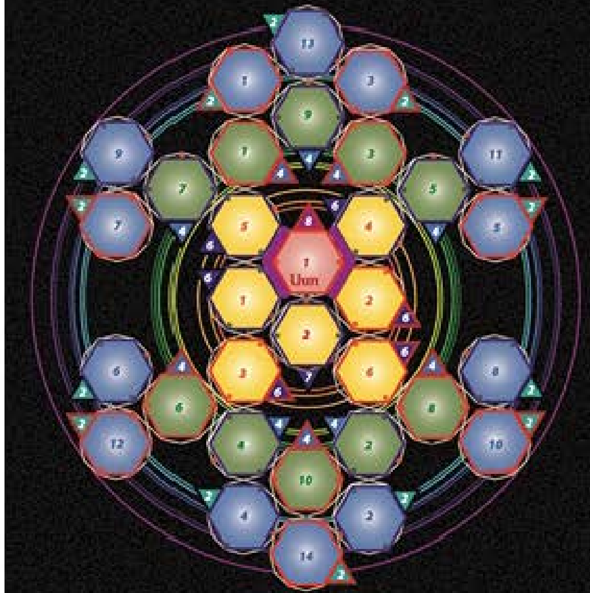
Uun
1
2
3
4
5
6
7
8
9
10
11
12
13
14

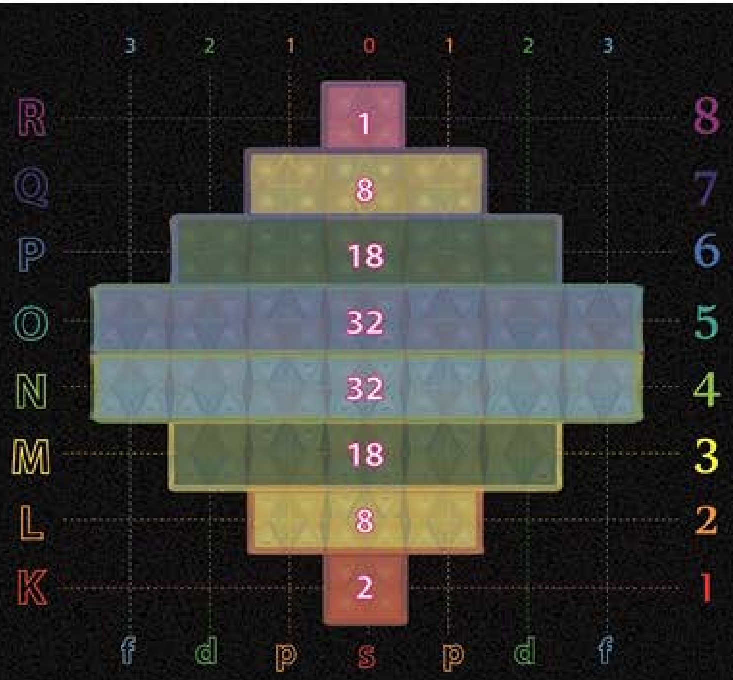
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
1
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

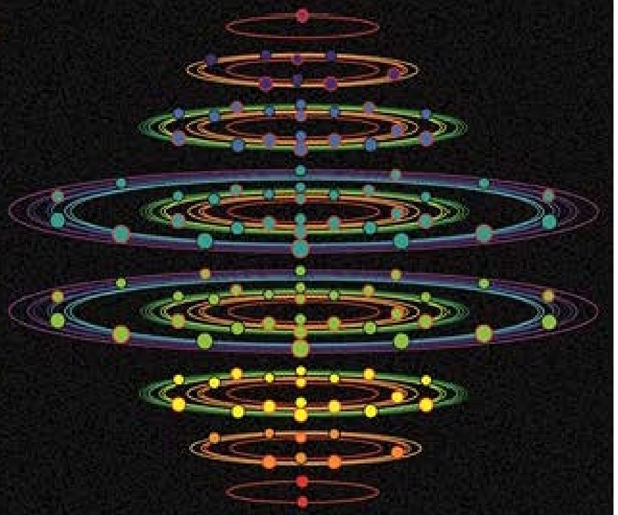

Ununnenium

$$\overset{R}{\underset{K}{\vphantom{\sum}}}\,{}_{P}\!\sum_{25}^{29}\underbrace{\left[\overset{\text{Baryon rest quanta}}{\left[72(n^2)\right]}+\overset{\text{lepton rest quanta}}{\left[12e19\right]}+\overset{\text{KEM}}{\left[\Delta v\right]}\right]_{1}^{8}}_{\text{nuclear mass-energy quanta per shell}}$$

[Uuo] 8s1

Alkali Metal

Azmithal & Magnetic numbers

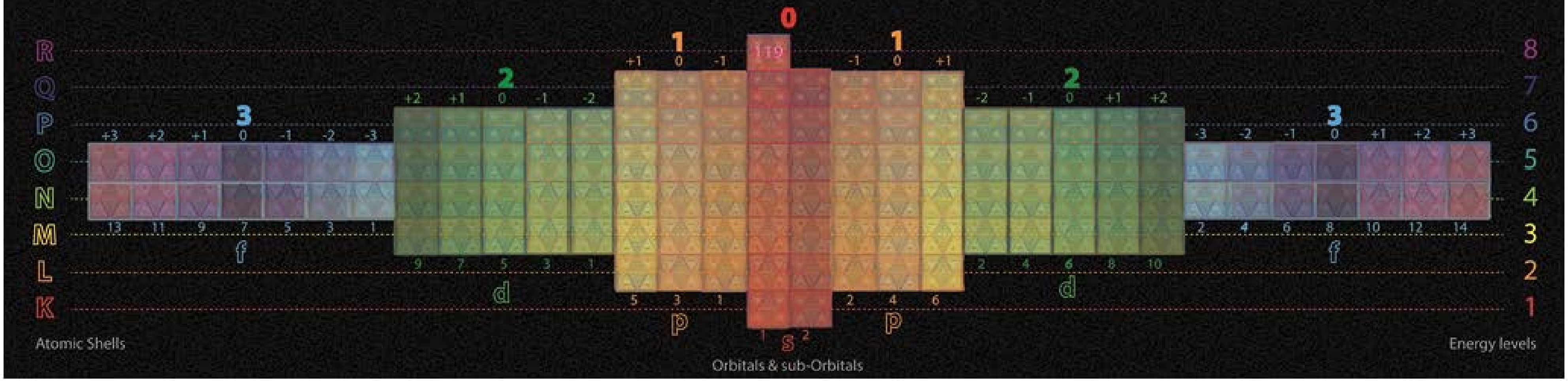
0
1
1
2
2
3
3
119
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
+3 +2 +1 0 -1 -2 -3
13 11 9 7 5 3 1
f
+2 +1 0 -1 -2
9 7 5 3 1
d
+1 0 -1
5 3 1
p
1
s
2
-1 0 +1
2 4 6
p
-2 -1 0 +1 +2
2 4 6 8 10
d
-3 -2 -1 0 +1 +2 +3
2 4 6 8 10 12 14
f

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Tetryonics 53.119 - Ununnenium atomic config

Copyright ABRAHAM [2008] - All rights reserved

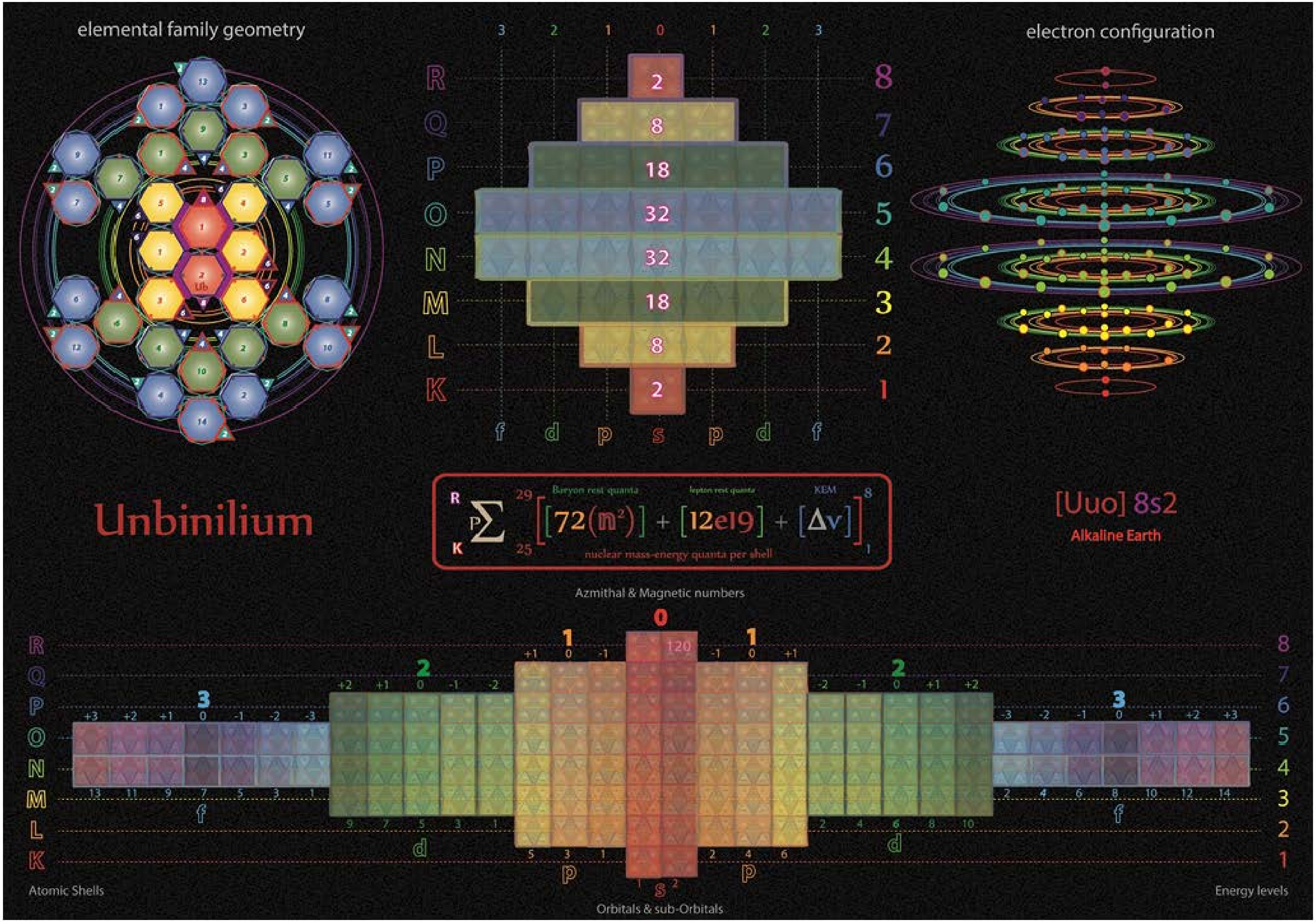
elemental family geometry
electron configuration
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2
8
18
32
32
18
8
2
f d p s p d f
Unbinilium
Baryon rest quanta
lepton rest quanta
KEM
$\sum_{K}^{R} P \; {}^{29}_{25}\left[[72(m^2)] + [12e19] + [\Delta v]\right]^{8}_{1}$
nuclear mass-energy quanta per shell
[Uuo] 8s2
Alkaline Earth
Azmithal & Magnetic numbers
0
120
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-1 0 +1
-2 -1 0 +1 +2
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
9 7 5 3 1
5 3 1
1 s 2
2 4 6
2 4 6 8 10
2 4 6 8 10 12 14
f d p p d f
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 53.120 - Unbinilium atomic config

Copyright ABRAHAM [2008] - All rights reserved

Chemical element bonding

All Deuterium nuclei pack in hexagonal patterns
to form larger atoms & molecules

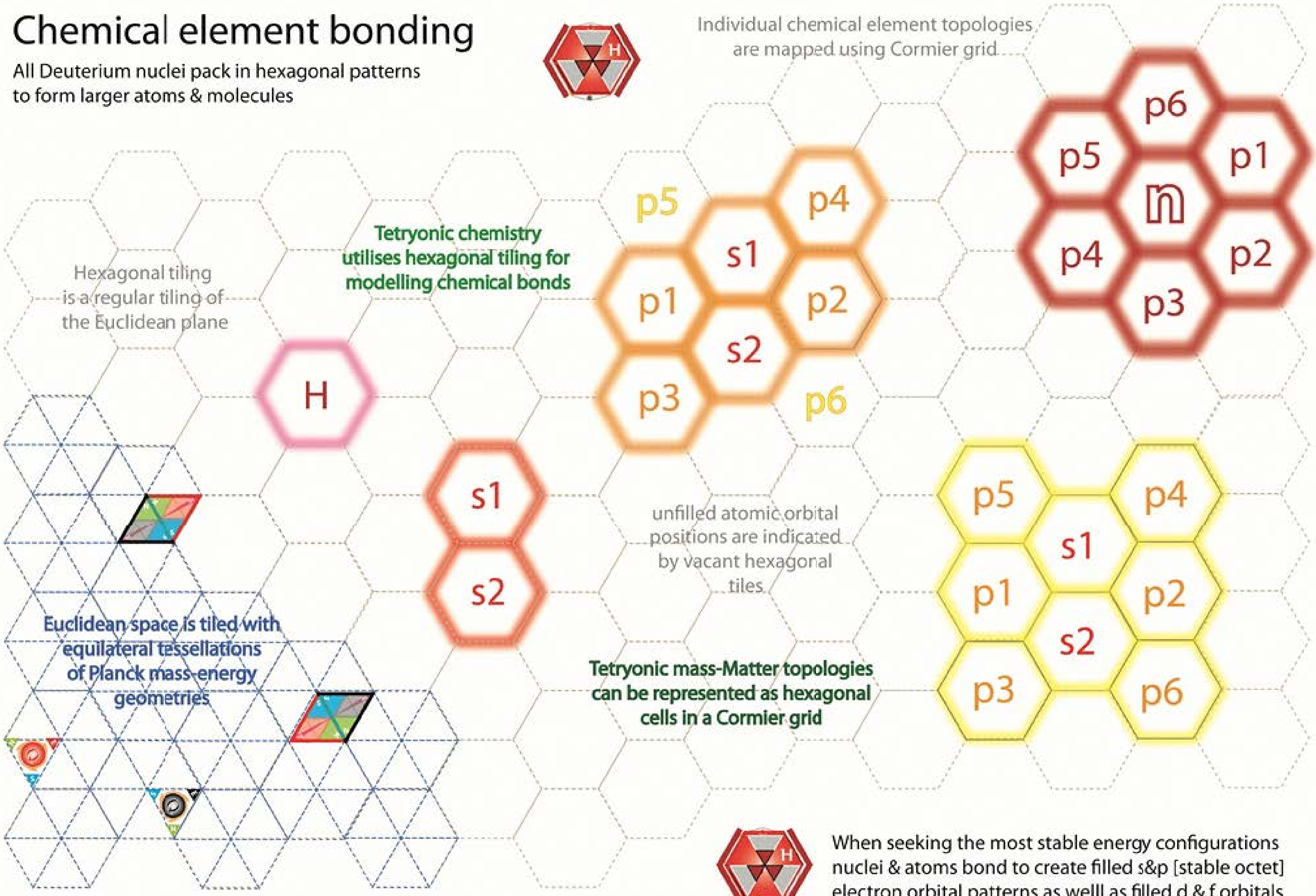

Individual chemical element topologies
are mapped using Cormier grid

Hexagonal tiling
is a regular tiling of
the Euclidean plane

**Tetryonic chemistry
utilises hexagonal tiling for
modelling chemical bonds**

**Euclidean space is tiled with
equilateral tessellations
of Planck mass-energy
geometries**

unfilled atomic orbital
positions are indicated
by vacant hexagonal
tiles

**Tetryonic mass-Matter topologies
can be represented as hexagonal
cells in a Cormier grid**

When seeking the most stable energy configurations
nuclei & atoms bond to create filled s&p [stable octet]
electron orbital patterns as welll as filled d & f orbitals.

Copyright ABRAHAM [2008] - All rights reserved

Periodic Table Atomic configurations

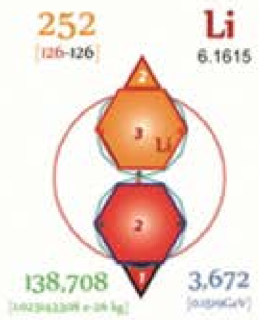
252
[126-126]
Li
6.1615
138,708
3,672

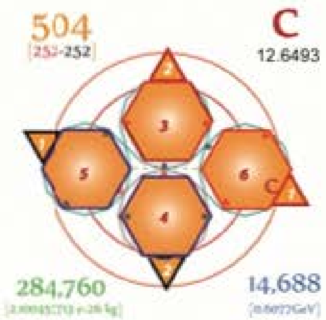
504
[252-252]
C
12.6493
284,760
14,688

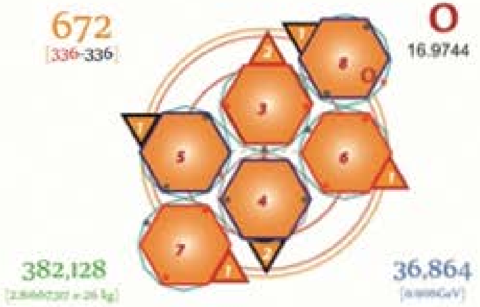
672
[336-336]
O
16.9744
382,128
36,864

Allotropes are different structural forms of the same element and can exhibit quite different physical properties and chemical behaviours

Lithium

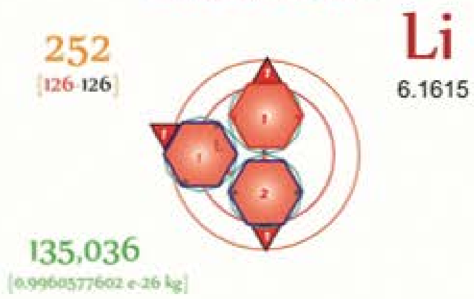
252
[126-126]
Li
6.1615
135,036
[0.9960577602 e-26 kg]

Allotropes

The Periodic Table, although useful in identifying Elements via their atomic and quantum numbers, does not reflect all the charged topologes that Deuterium nuclei can form as they combine.

Various atomic configurations with the same Tetryonic charge, but differing in their final mass-Matter topologies and properties can be formed - they are the Allotropes

Some of the more plentiful chemical elements form this way

Oxygen

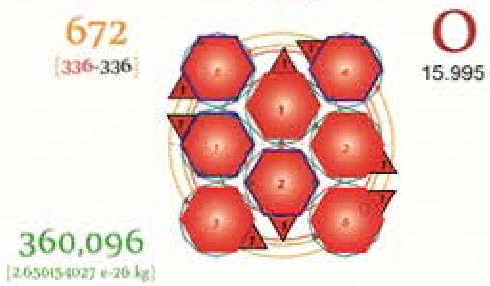
672
[336-336]
O
15.995
360,096
[2.656154027 e-26 kg]

Carbon

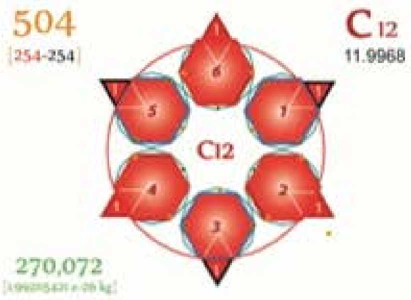
504
[254-254]
C12
11.9968
C12
270,072

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1

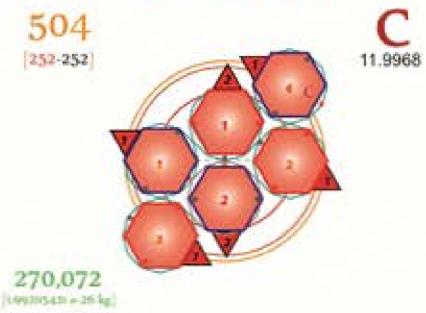
504
[252-252]
C
11.9968
270,072

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1

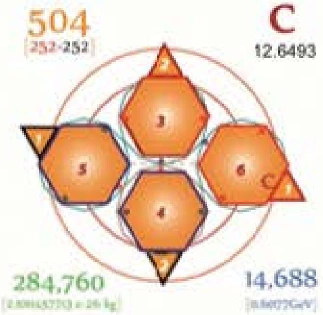
504
[252-252]
C
12.6493
284,760
14,688

6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2

Allotropes are elements created from the same number of Deuterium nuclei as periodic elements but possess a differing mass-Matter topology

Q

same component charge

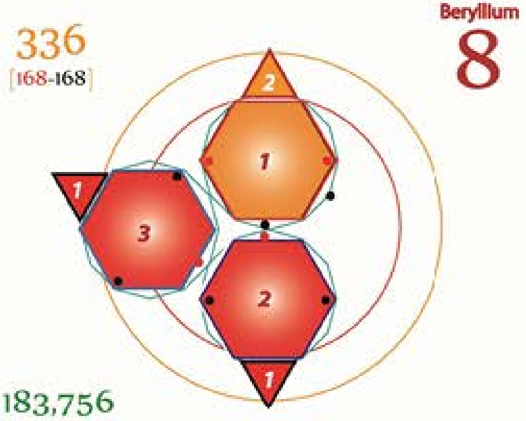
336
[168-168]
Beryllium
8
2
1
1
3
2
1
183,756

different Matter geometries

M

The varying material geometries allows what is the same chemical element to possess vastly different bonding points and chemical attributes

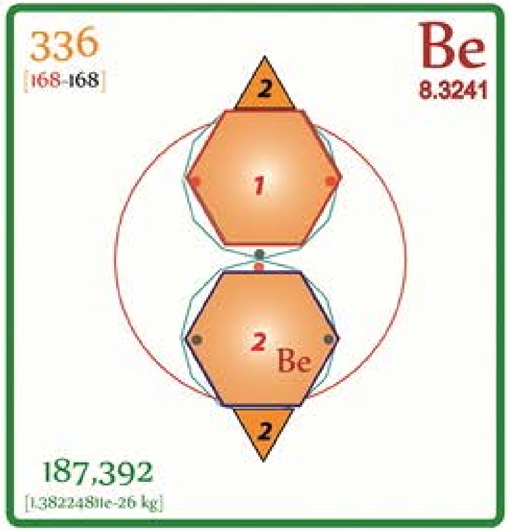
336
[168-168]
Be
8.3241
2
1
2 Be
2
187,392
[1.3822481e-26 kg]

Allotropic geometries

[charge vs Matter]

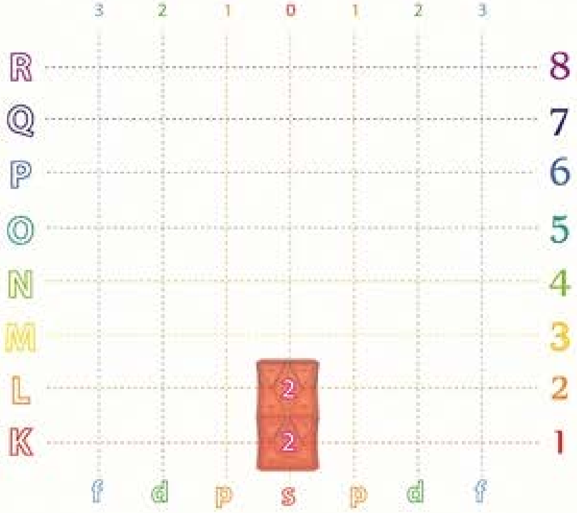
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
2
f
d
p
s
p
d
f

Some element allotropes have different Matter topologies that persist in different phases

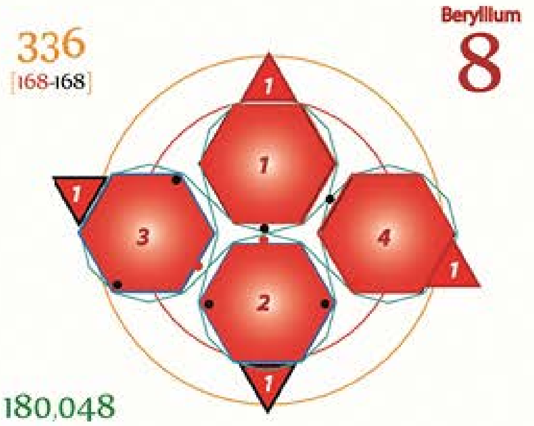
336
[168-168]
Beryllium
8
1
1
1
3
4
1
2
1
180,048

Allotropes vs. Isotopes

Isotopes are elementary atoms with the same number of nuclei, but with differing energy levels, resulting in different mass-energies

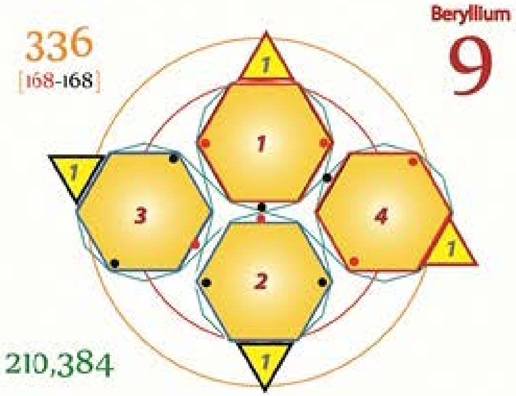
336
[168-168]
Beryllium
9
1
1
1
3
4
1
2
1
210,384

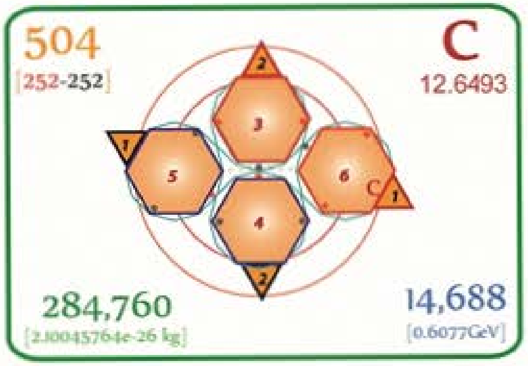
504
[252-252]
C
12.6493
3
5
6
4
1
2
1
284,760
[2.10045764e-26 kg]
14,688
[0.6077GeV]

Carbon Allotropes

Carbon is capable of forming many allotropes due to its valency.

Well known forms of carbon include:

Carbon
Diamond
Graphite
Graphene
Amorphous carbon
Buckminsterfullerenes
Carbon nanotubes
Glassy carbon
atomic & diatomic carbon

Tetryonic theory affords us the ability to model the charged mass-energy geometries & 3D Matter topologies of each element along with its bonding points

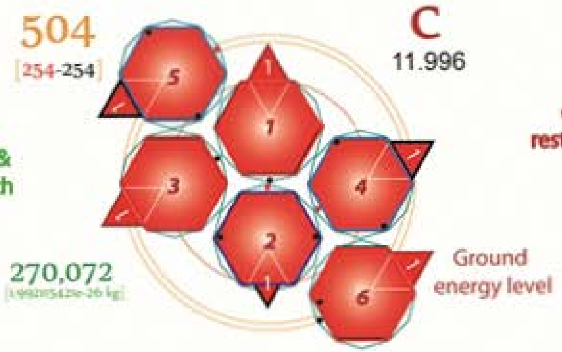
504
[254-254]
C
11.996
5
1
3
4
2
6
270,072
Ground energy level

It also allows for the calculation of the molar rest mass-energy geometry of any Matter topology

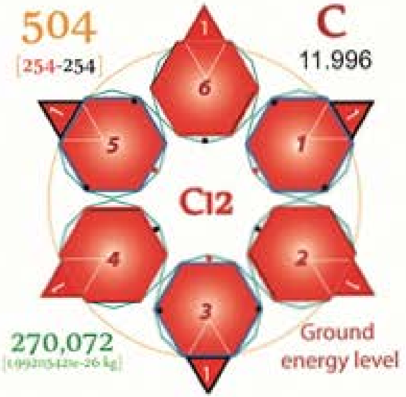
504
[254-254]
C
11.996
6
5
1
C12
4
2
3
270,072
Ground energy level

Graphene has a unique p orbital arrangement of its electrons

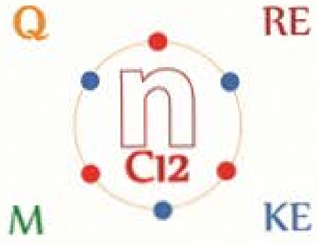
Q
RE
n
C12
M
KE

Electron Configuration

[n] p6

Shell			Molar meM	Nuclei per shell
R	8	446,976	19.855	6
Q	7	418,680	18.598	6
P	6	391,392	17.385	6
O	5	365,112	16.218	6
N	4	339,840	15.095	6
M	3	315,576	14.018	6
L	2	292,320	12.985	6
K	1	270,072	11.996	6

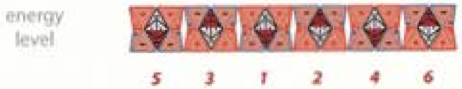
energy level
5
3
1
2
4
6

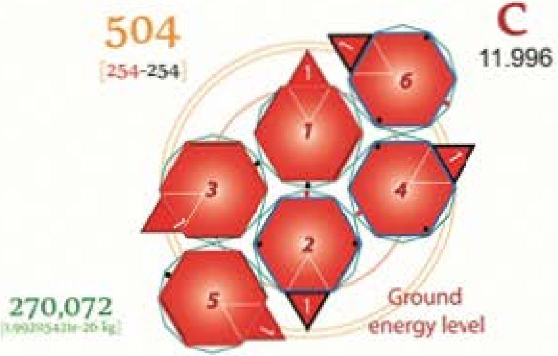
504
[254-254]
C
11.996
6
1
3
4
2
5
1
270,072
Ground energy level

Carbon-14, a radioactive isotope of carbon with a half-life of 5,730 years, is used to find the age of formerly living things through a process known as radiocarbon dating.

In 1961 the international unions of physicists and chemists agreed to use the mass of the isotope carbon-12 as the basis for atomic weight.

There are nearly ten million known carbon compounds and an entire branch of chemistry, known as organic chemistry, is devoted to their study.

Graphene

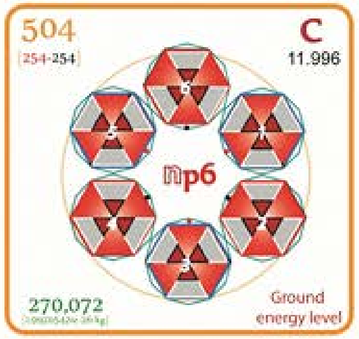

Graphene is also an allotrope of carbon.

Its structure is one-atom-thick planar sheets of sp2-bonded carbon atoms that are densely packed in a honeycomb crystal lattice

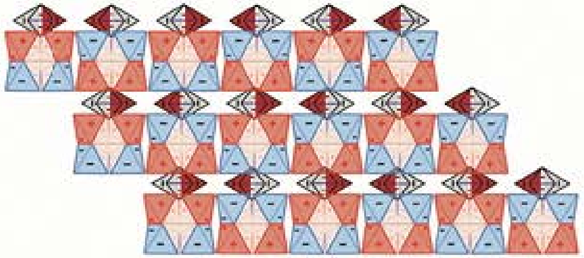

Graphite electrons can act as atomic bearings between the planar sheets of graphene

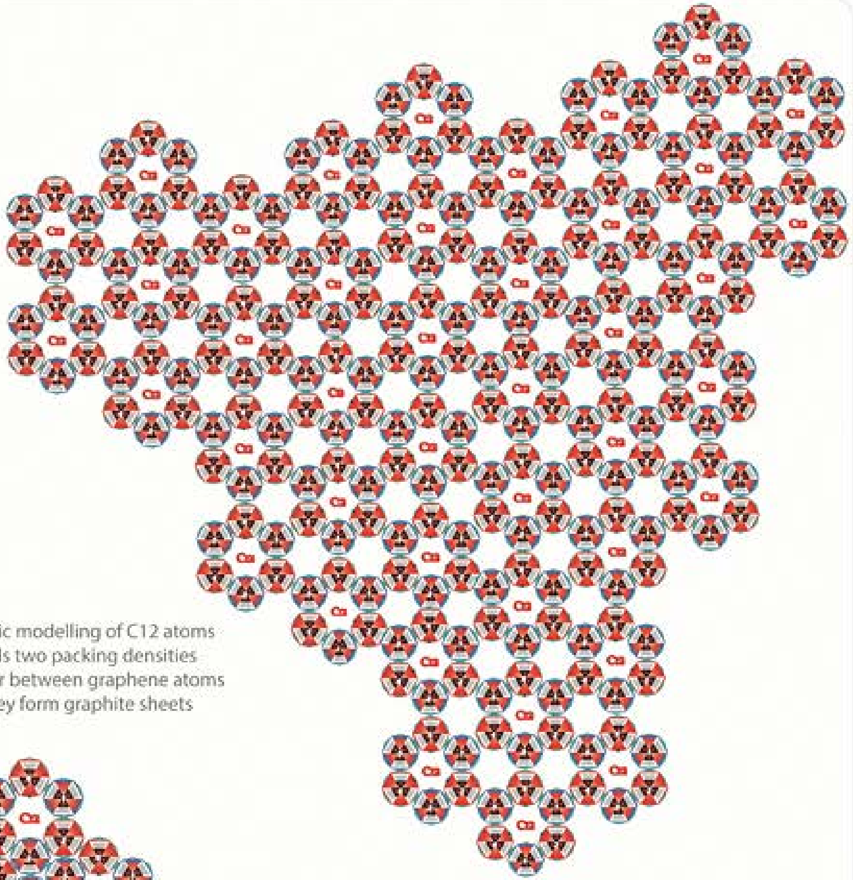

Graphite

The mineral graphite is an allotrope of carbon.

Unlike periodic table carbon [diamond], graphite is an electrical conductor, a semi-metal.

And Graphite is the most stable form of carbon under standard conditions.

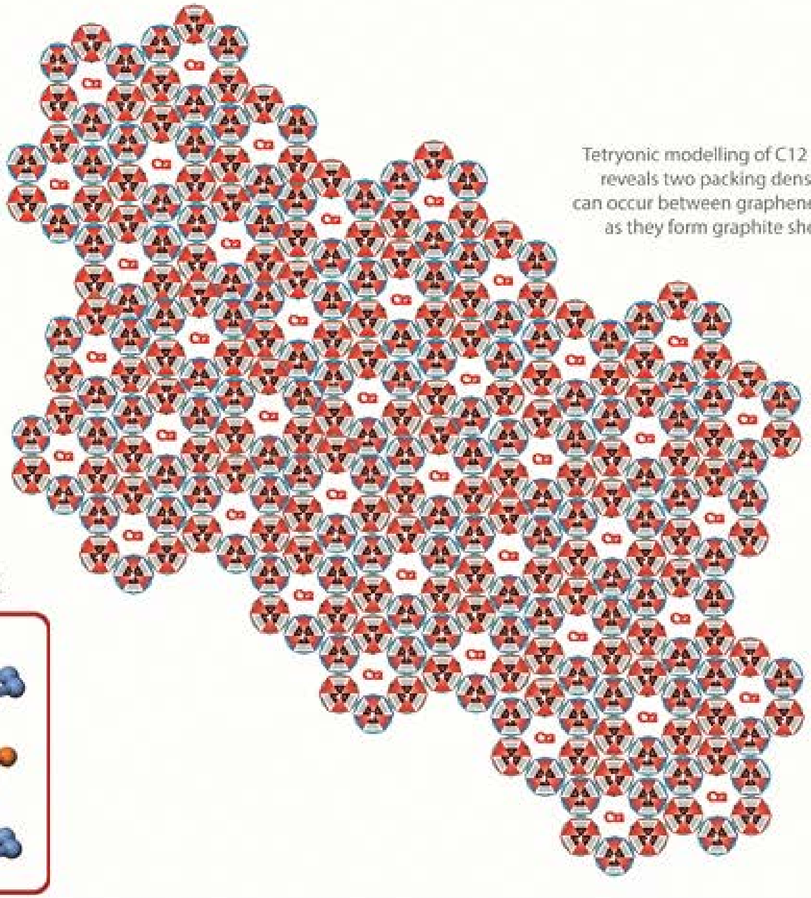

Tetryonic modelling of C12 atoms reveals two packing densities can occur between graphene atoms as they form graphite sheets

Graphite has a layered, planar structure. In each layer, the carbon atoms are arranged in a hexagonal lattice

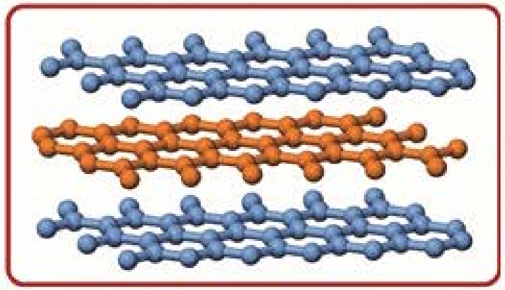

Unlike periodic carbon where electrons are tightly bound, graphite is highly conductive, with electrical energy being propagated via the electrons & their KEM fields

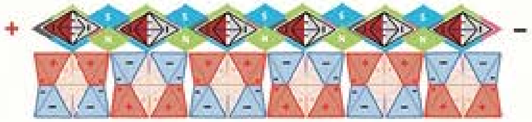

Graphene electrons are loosley bound to the top of the Deuterium nuclei

Copyright ABRAHAM [2008] - All rights reserved

Carbon Nanotubes

Carbon nanotubes (CNTs) are allotropes of carbon with a cylindrical nanostructure.

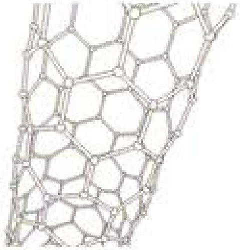

Large molecules consisting only of carbon, known as buckminsterfullerenes, or buckyballs, have recently been discovered and are currently the subject of much scientific interest.

A single buckyball consists of 60 or 70 carbon atoms (C60 or C70) linked together in a structure that looks like a soccer ball.

They can trap other atoms within their framework, appear to be capable of withstanding great pressures and have magnetic and superconductive properties.

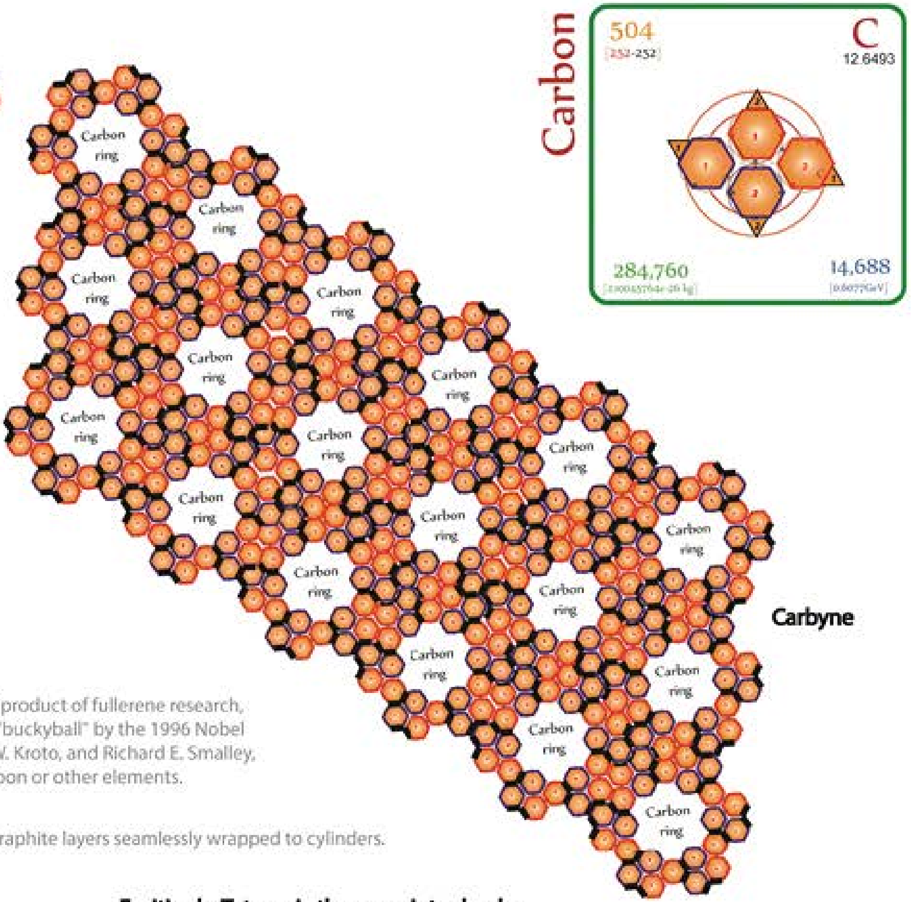
C36
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbon ring
Carbyne
Carbon
504
[252-252]
C
12.6493
284,760
14,688

They can be thought of as a sheet of graphite (a hexagonal lattice of carbon) rolled into a cylinder.

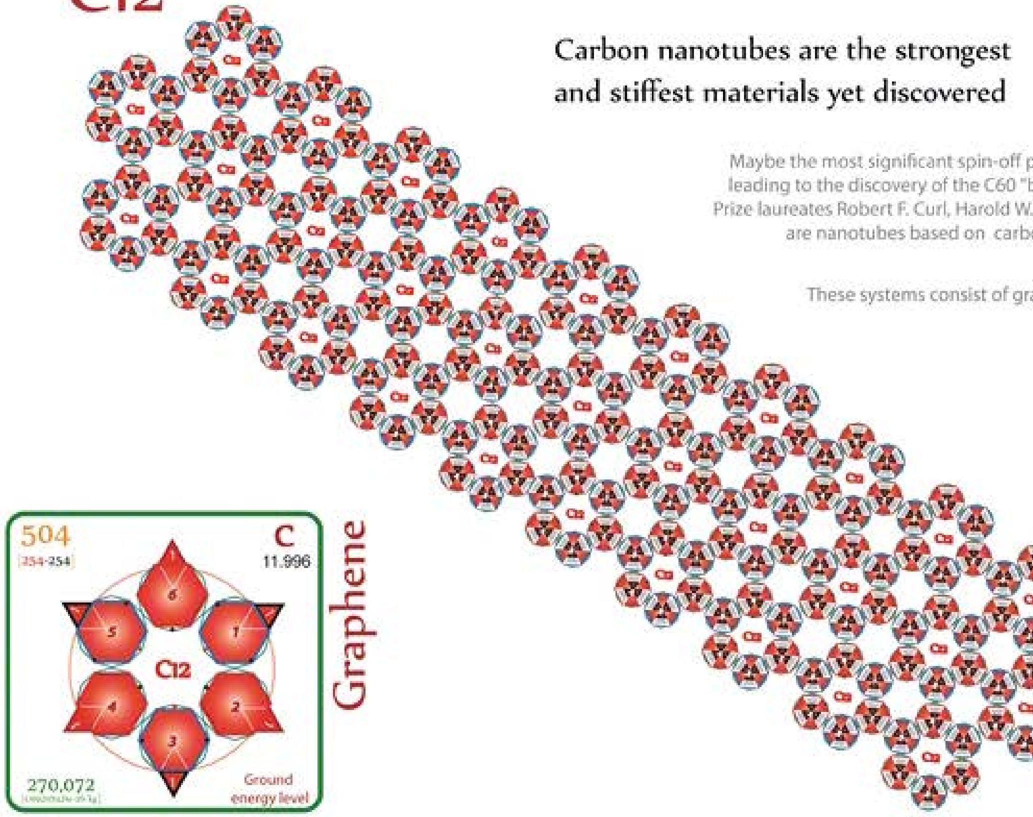
C12
504
[254-254]
C
11.996
6
5
1
C12
4
2
3
270,072
Ground energy level
Graphene

Carbon nanotubes are the strongest and stiffest materials yet discovered

Maybe the most significant spin-off product of fullerene research, leading to the discovery of the C60 "buckyball" by the 1996 Nobel Prize laureates Robert F. Curl, Harold W. Kroto, and Richard E. Smalley, are nanotubes based on carbon or other elements.

These systems consist of graphite layers seamlessly wrapped to cylinders.

Excitingly, Tetryonic theory points clearly to the possibility that organo-carbon nanotubes can be created.

Where C36 periodic carbon could be used insetad of conventional C12 graphene atoms

Carbyne

C60

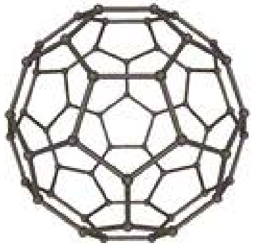

Nanotubes are members of the fullerene structural family, which also includes the spherical buckyballs, and the ends of a nanotube may be capped with a hemisphere of the buckyball structure.

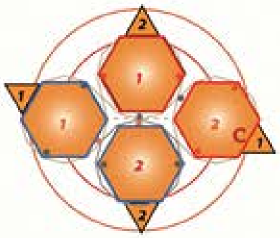

Periodic [C6] Carbon

Diamonds are renowned as a material with superlative physical qualities, most of which originate from the strong fascia bonds between its atoms.

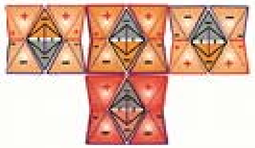

Diamonds

are a metastable allotrope of carbon, where the carbon atoms are arranged in cubic crystal structure called a diamond lattice.

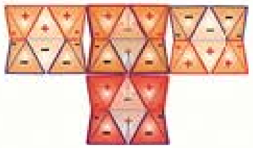

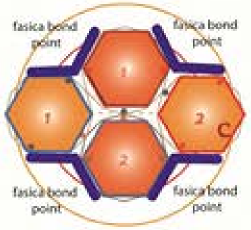

Ionised [C6] Carbon

Most natural diamonds are formed at high temperature and pressure at depths of 140 to 190 kilometers in the Earth's mantle [resulting in the loss of electrons form bound positions, making them poor conductors compared to grahphene atoms].

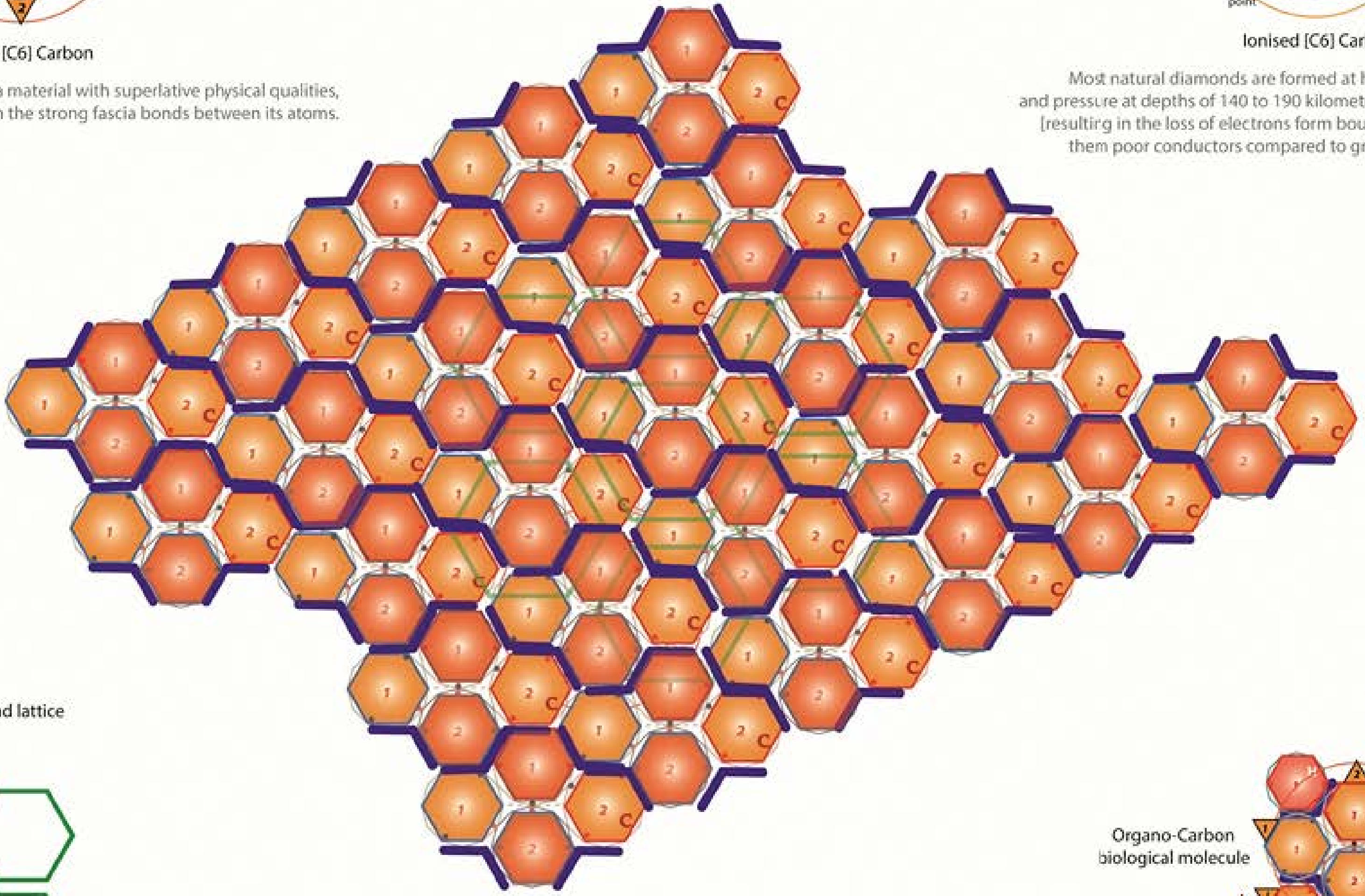

filled n[C7] Carbon diamond lattice

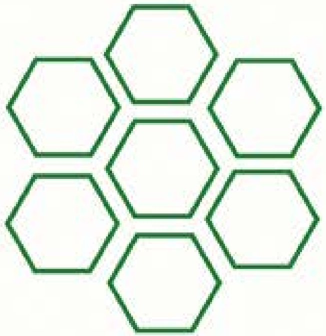

Organo-Carbon biological molecule

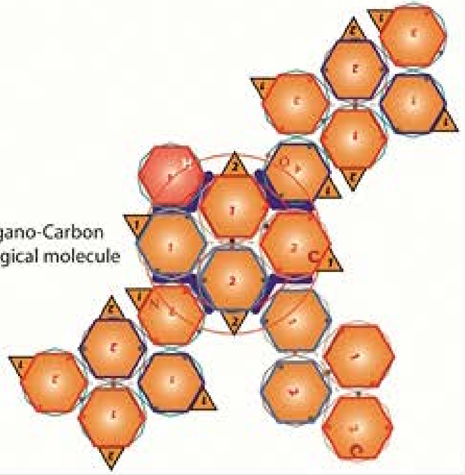

The chemical bonds that hold the carbon atoms in diamonds together are stronger than those in graphite.

In diamonds, the bonds form an inflexible three-dimensional lattice, whereas in graphite, the atoms are tightly bonded into sheets, which can slide easily over one another, making the overall structure weaker

Oxygen Allotropes

There are several known allotropes of oxygen.
The most familiar is molecular oxygen (O2), present at significant levels
in Earth's atmosphere and also known as dioxygen or triplet oxygen.
Another is the highly reactive ozone (O3).

Others include:
Atomic oxygen (O1, a free radical)
Singlet oxygen (O2), either of two metastable states of molecular oxygen
Tetraoxygen (O4), another metastable form
Solid oxygen, existing in six variously colored phases, of which
one is O8 and another one metallic

Atomic Oxygen also forms bonds easily with Hydrogen
to create Hydroxy compounds

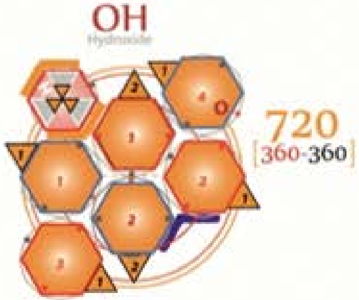
OH
Hydroxide
720
[360-360]

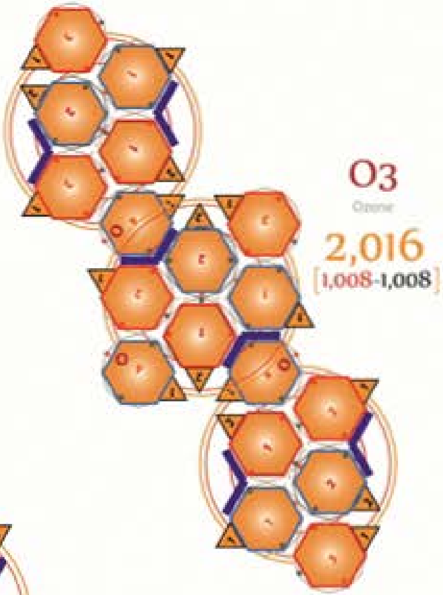
O3
Ozone
2,016
[1,008-1,008]

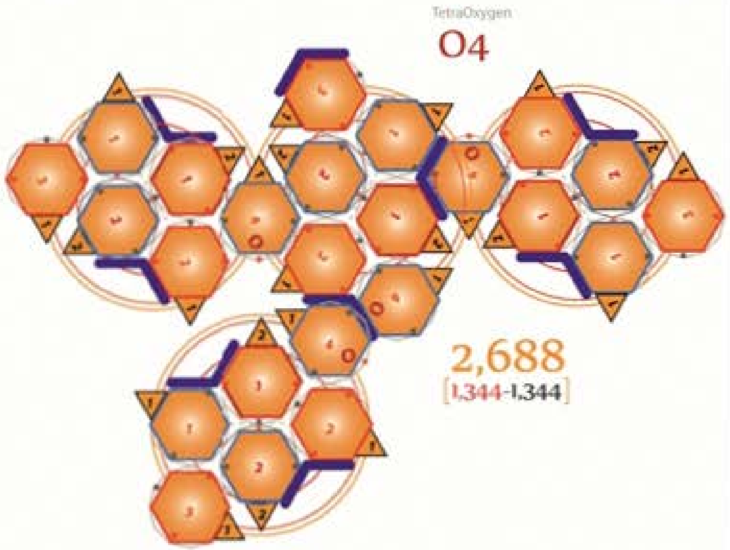
TetraOxygen
O4
2,688
[1,344-1,344]

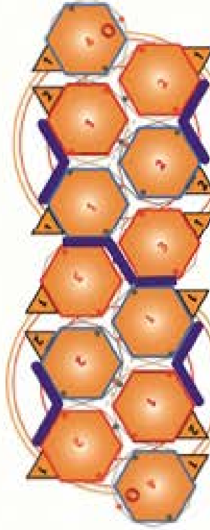
O2
Oxygen
1,344
[672-672]

All charge fascia interact
and bind where possible to form
neutral chemical bonds

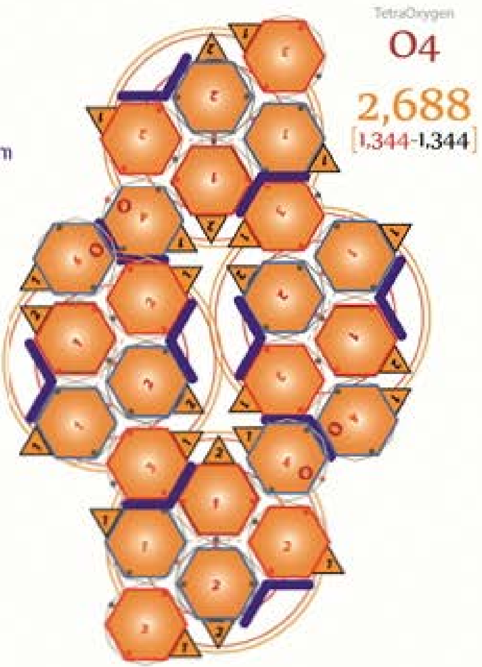
TetraOxygen
O4
2,688
[1,344-1,344]

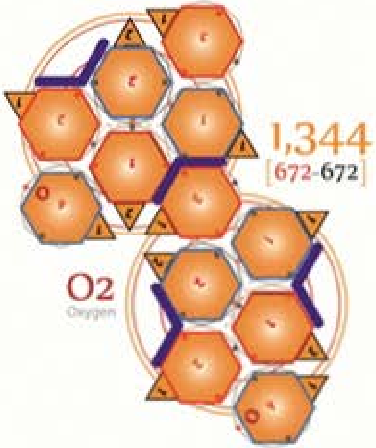
1,344
[672-672]
O2
Oxygen

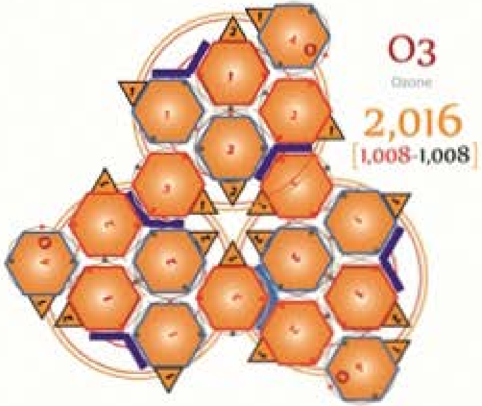
O3
Ozone
2,016
[1,008-1,008]

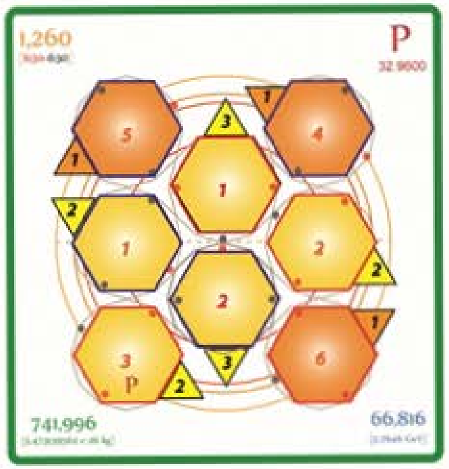
1,260
P
32.9600
5
4
3
1
1
1
2
1
2
2
2
3
6
1
3
P
2
741,996
66,816

Phosphor Allotropes

Elemental phosphorus can exist in several allotropes; the most common of which are white and red solids.

Solid violet and black allotropes are also known.

Gaseous phosphorus exists as diphosphorus and atomic phosphorus.

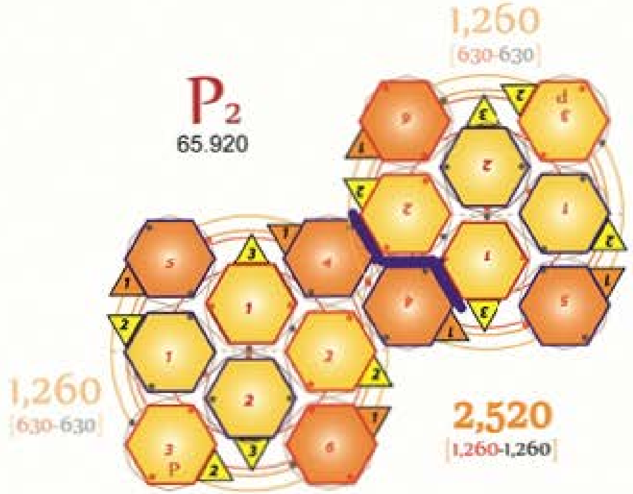
P_2
65.920
1,260
[630-630]
1,260
[630-630]
2,520
[1,260-1,260]

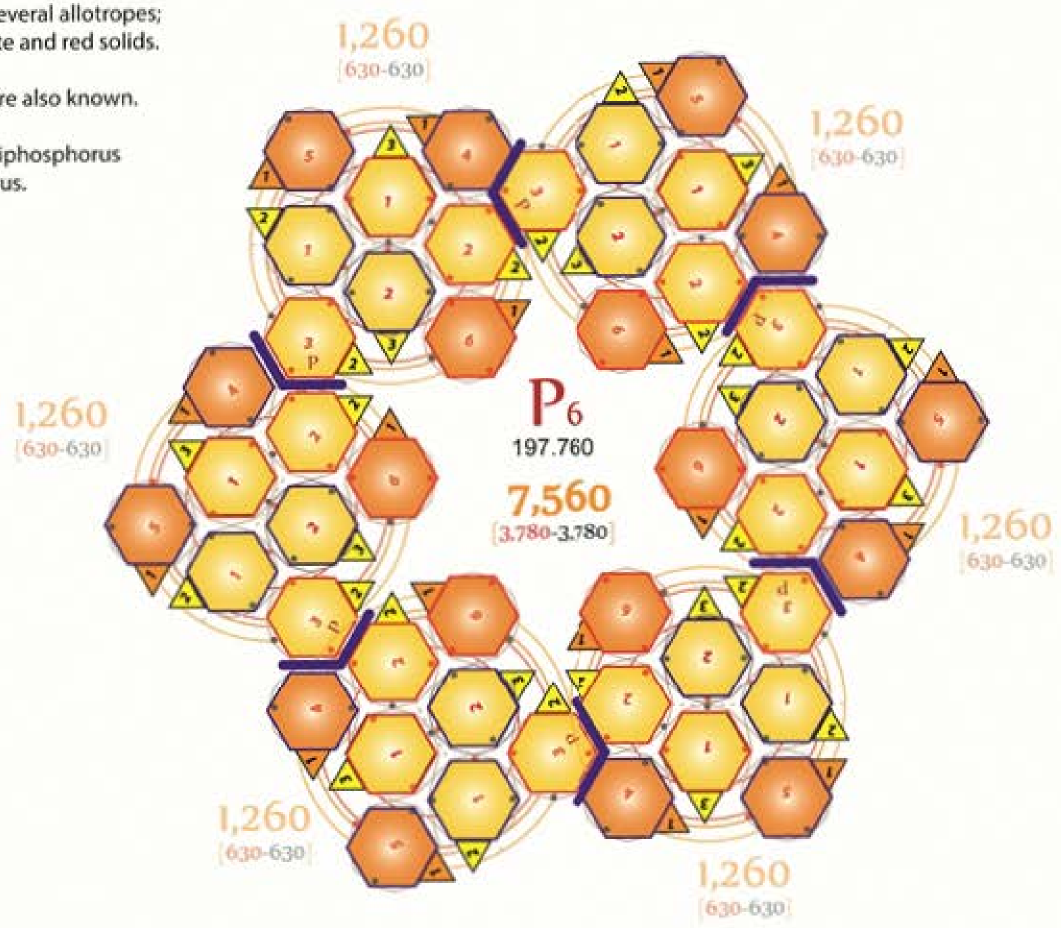
P_6
197.760
7,560
[3,780-3,780]
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]
1,260
[630-630]

The diphosphorus allotrope (P2) can be obtained normally only under extreme conditions (for example, from P4 at 1100 kelvin).

Nevertheless, some advancements were obtained in generating the diatomic molecule in homogenous solution, under normal conditions with the use by some transition metal complexes (based on for example tungsten and niobium).

Black phosphorus is the thermodynamically stable form of phosphorus at room temperature and pressure. It is obtained by heating white phosphorus under high pressures (12,000 atmospheres).

In appearance, properties and structure it is very like graphite, being black and flaky, a conductor of electricity, and having puckered sheets of linked atoms

Copyright ABRAHAM [2008] - All rights reserved

Sulfur Allotropes

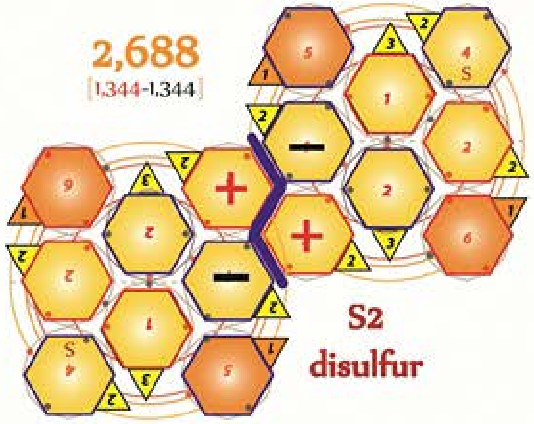

2,688
[1,344-1,344]

S2
disulfur

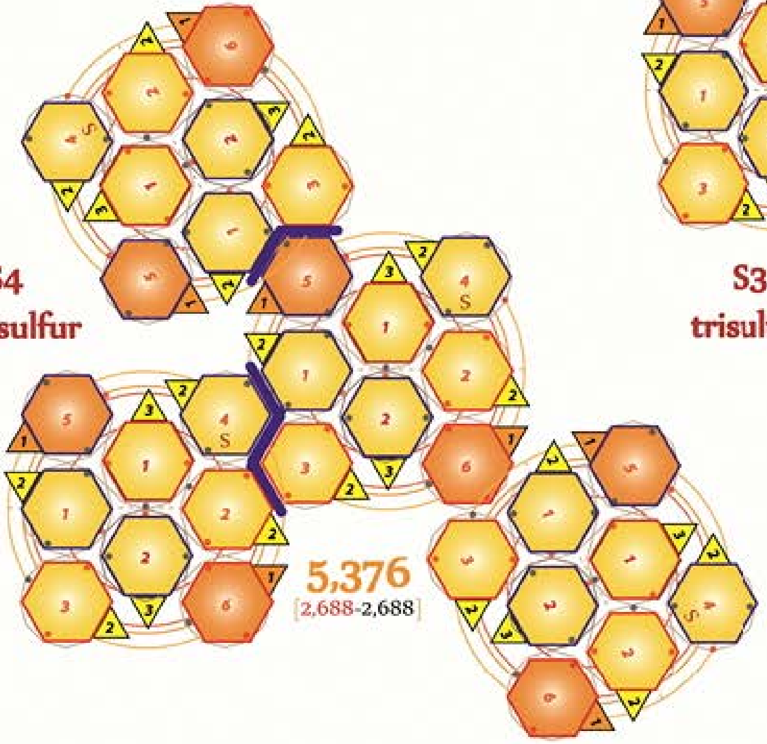

S4
tetrasulfur

5,376
[2,688-2,688]

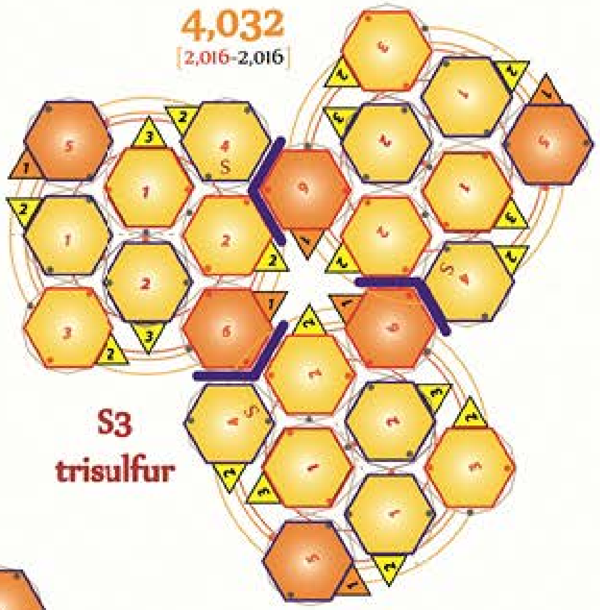

4,032
[2,016-2,016]

S3
trisulfur

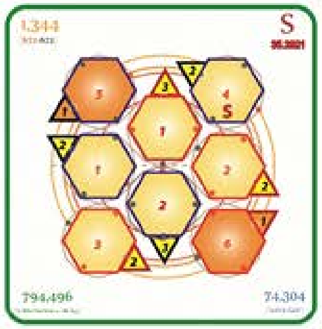

Charged fascia interactions
between differing elements
result in the familiar
chemical bonds

Sulfur [like Carbon]
can form Chains
or Ring compounds

There are a large number of allotropes of sulfur.
In this respect, sulfur is second only to carbon.

cyclo-hexasulfur

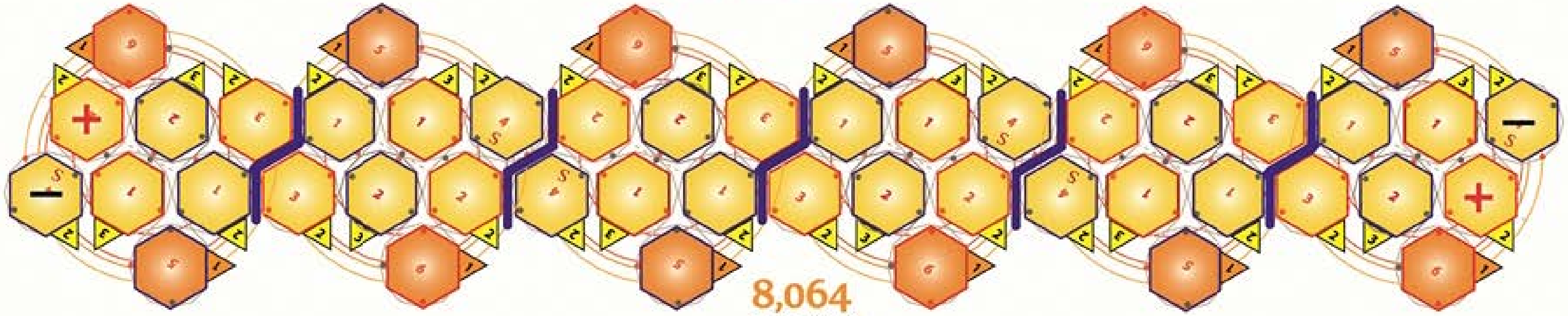

8,064
[4,032-4,032]

Tetryonics 54.10 - Sulfur allotrope

Isotopes

In addition to forming varying allotropic elements

Atoms can also absorb energy directly and create numerous elemental isotopes as a result of their differing nuclear energy levels

Deuterium nuclei with bound photo-electrons form quantum-scale synchronous converters

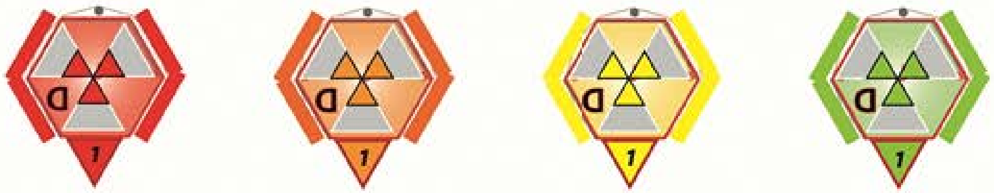

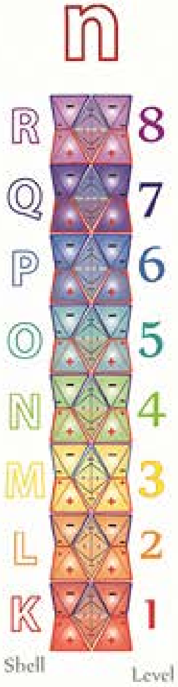

0	Hydrogen	22,512	24,384	26,352	28,416	30,576	32,832	35,184	37,632
Element		1	2	3	4	5	6	7	8
1	Deuterium	45,012	48,720	52,596	56,640	60,852	65,232	69,780	74,496
2	Helium	90,024	97,440	105,192	113,280	121,704	130,464	139,560	148,992
3	Lithium	135,036	146,160	157,788	169,920	182,556	195,696	209,340	223,488
4	Berylium	180,048	194,880	210,384	226,560	243,408	260,928	279,120	297,984
5	Boron	225,060	243,600	262,980	283,200	304,260	326,160	348,900	372,480
6	Carbon	270,072	292,320	315,576	339,840	365,112	391,392	418,680	446,976
7	Nitrogen	315,084	341,040	368,172	396,480	425,964	456,624	488,460	521,472
8	Oxygen	360,096	389,760	420,768	453,120	486,816	521,856	558,240	595,968
9	Fluorine	405,108	438,480	473,364	509,760	547,668	587,088	628,020	670,464
10	Neon	450,120	487,200	525,960	566,400	608,520	652,320	697,800	744,960
11	Sodium	495,132	535,920	578,556	623,040	669,372	717,552	767,580	819,456
12	Magnesium	540,144	584,640	631,152	679,680	730,224	782,784	837,360	893,952
13	Aluminium	585,156	633,360	683,748	736,320	791,076	848,016	907,140	968,448
14	Silicon	630,168	682,080	736,344	792,960	851,928	913,248	976,920	1,042,944
15	Phosphorus	675,180	730,800	788,940	849,600	912,780	978,480	1,046,700	1,117,440
16	Sulfur	720,192	779,520	841,536	906,240	973,632	1,043,712	1,116,480	1,191,936
17	Chlorine	765,204	828,240	894,132	962,880	1,034,484	1,108,944	1,186,260	1,266,432
18	Argon	810,216	876,960	946,728	1,019,520	1,095,336	1,174,176	1,256,040	1,340,928
19	Potassium	855,228	925,680	999,324	1,076,160	1,156,188	1,239,408	1,325,820	1,415,424
20	Calcium	900,240	974,400	1,051,920	1,132,800	1,217,040	1,304,640	1,395,600	1,489,920

Level
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
Shell

It is the increased nucleonic energy levels that creates isotopes (not extra Neutrons within an atomic nucleus)

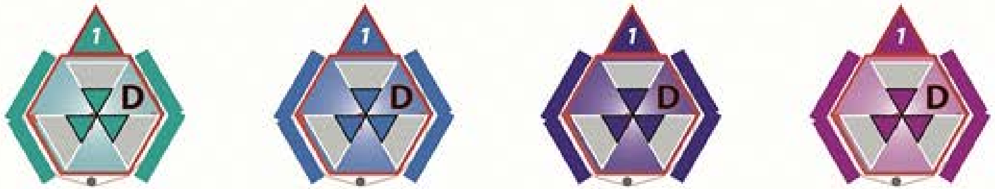

The energy level of Baryons determines the KEM energies of photo-electrons bound to them

Most isotopes are considered
to be radioactive as a result of
the nuclei seeking to release excess
energy in the form of photons of energy
or particles with high kinetic energies

Copyright ABRAHAM [2008] - All rights reserved

Carbon Isotopes

It is widely held in the scientific community tha: Carbon-14, 14C, or radiocarbon, is a radioactive isotope of carbon with a nucleus containing 6 protons and 8 neutrons.

It is in fact a nucleus comprised of 6 deuterium nuclei [with 6 Protons, 6 Neutrons & 6 electrons]

The mistaken belief in 'extra' neutrons being present in the nucleus stems from the fact that electrons and protons combine in equal numbers in the atomic nucleus and historically attributing the mass in excess of this as being the result of the mass contribution of 'extra' neutrons

Tetryonics finally corrects this erroneous assumption

Carbon 12-14

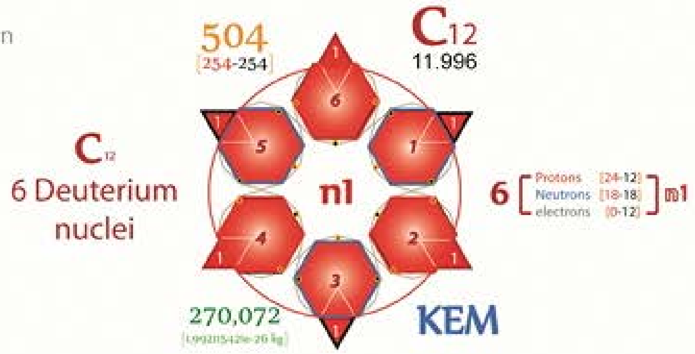
C_{12}
6 Deuterium
nuclei
504
[254-254]
C_{12}
11.996
6
5
1
n1
4
2
3
1
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1
270,072
[1.992115426e-26 kg]
KEM

There are NO extra neutrons (in excess of the element's Z#) in the nuclei of atomic isotopes.

The measured 'excess mass' is the direct result of the raised quantum levels of the Deuterium nuclei that comprise each atomic element

And is compreletely accounted for in Tetryonic theory by calculating for the total rest mass-energies in each elementary Matter topology.

The 'extra' mass historically attrtibuted to neutron numbers above that of the elemental number are now reflected as stored kinetic 'chemical' energies as they always were.

ALL elements & isotopes have equal numbers of Protons, electrons & Neutrons

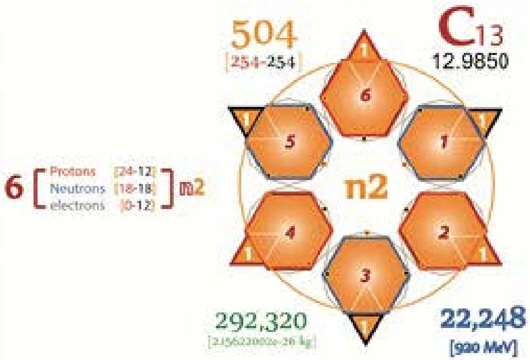
504
[254-254]
C_{13}
12.9850
6
5
1
n2
4
2
3
1
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m2
292,320
[2.156220002e-26 kg]
22,248
[920 MeV]

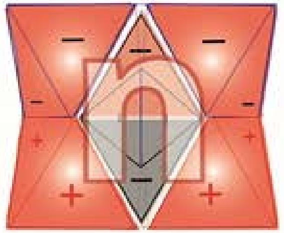

Quantum levels of atomic nuclei contribute to the molar mass [Isotopes are higher energy nuclei]

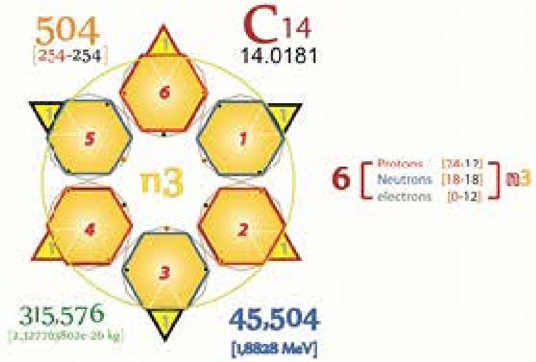
504
[254-254]
C_{14}
14.0181
6
5
1
n3
4
2
3
1
6 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m3
315,576
[2.327703802e-26 kg]
45,504
[1,8828 MeV]

	n1	n2	n3	n4	n5	n6	n7	n8
C12	270,072	292,320	315,576	339,840	365,112	391,392	418,680	446,976
KEM diff	0	22,248	45,504	69,768	95,040	121,320	148,608	176,904
neutron #		.98	2.0	3.1	4.2	5.4	6.6	7.8

This applies equally to all atomic nuclei

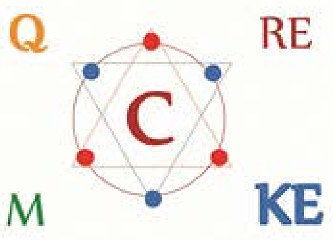

Electron Configuration
[n] p1

Atoms store mass-energies
in their standing wave Matter topologies
and release it via radioactive decay paths

It is increased kinetic energy level geometries that create nuclear isotopes (not extra Neutrons within a nucleus)

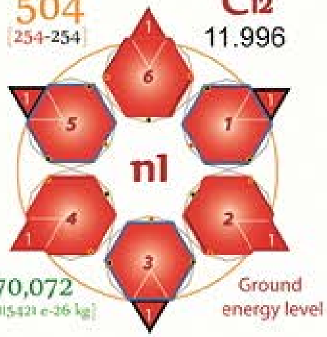

504
[254-254]
C_{12}
11.996
270,072
[1.992115421 e-26 kg]
Ground energy level

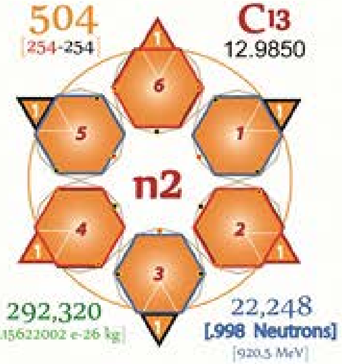

504
[254-254]
C_{13}
12.9850
292,320
[2.15622002 e-26 kg]
22,248
[.998 Neutrons]
[920.5 MeV]

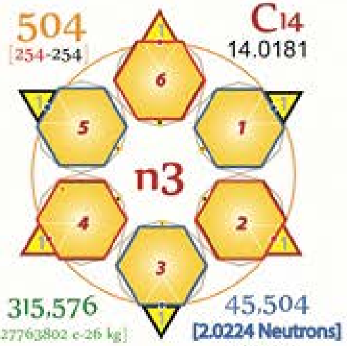

504
[254-254]
C_{14}
14.0181
315,576
[2.327763802 e-26 kg]
45,504
[2.0224 Neutrons]
[1.882 GeV]

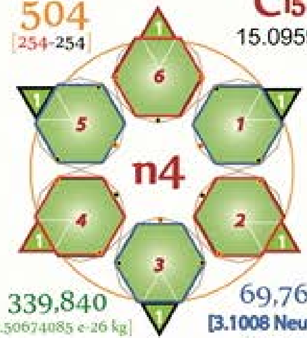

504
[254-254]
C_{15}
15.0959
339,840
[2.50674085 e-26 kg]
69,768
[3.1008 Neutrons]
[2.886 GeV]

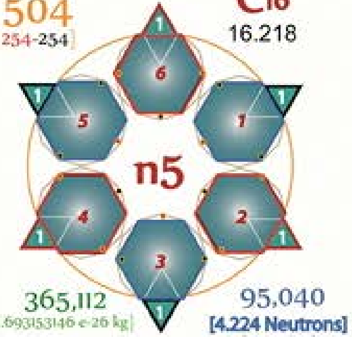

504
[254-254]
C_{16}
16.218
365,112
[2.693153146 e-26 kg]
95,040
[4.224 Neutrons]
[3.925 GeV]

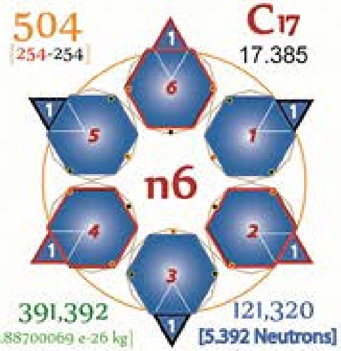

504
[254-254]
C_{17}
17.385
391,392
[2.88700069 e-26 kg]
121,320
[5.392 Neutrons]
[5.019 GeV]

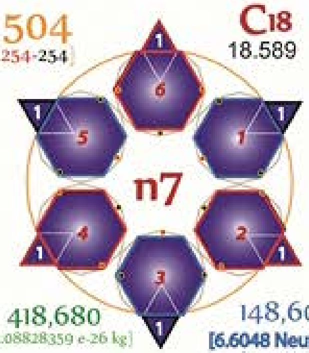

504
[254-254]
C_{18}
18.589
418,680
[3.08828359 e-26 kg]
148,608
[6.6048 Neutrons]
[6.145 GeV]

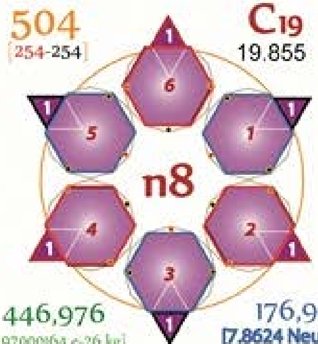

504
[254-254]
C_{19}
19.855
446,976
[23.29700064 e-26 kg]
176,904
[7.8624 Neutrons]
[7.319 GeV]

Shell			Molar meM	Nuclei per shell
R	8	446,976	19.855	6
Q	7	418,680	18.598	6
P	6	391,392	17.385	6
O	5	365,112	16.218	6
N	4	339,840	15.095	6
M	3	315,576	14.018	6
L	2	292,320	12.985	6
K	1	270,072	11.996	6

energy level

Deuterium nuclei with bound photo-electrons fom quantum-scale synchronous converters

22,500
[1 Neutron]

n

[930.97 MeV]

Kinetic mass-energy geometries
stored in 3D Matter topologies
is released as Chemical energies

Copyright ABRAHAM [2008] - All rights reserved

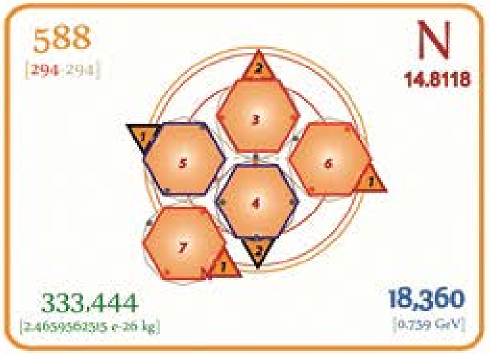
588
[294-294]
N
14.8118
3
5
6
4
7
2
1
333,444
[2.4659562315 e-26 kg]
18,360
[0.759 GeV]

It Is Increased Nucleonic energy levels that creates Isotopes (not extra Neutrons within any elementary Nucleus)

Nitrogen gas

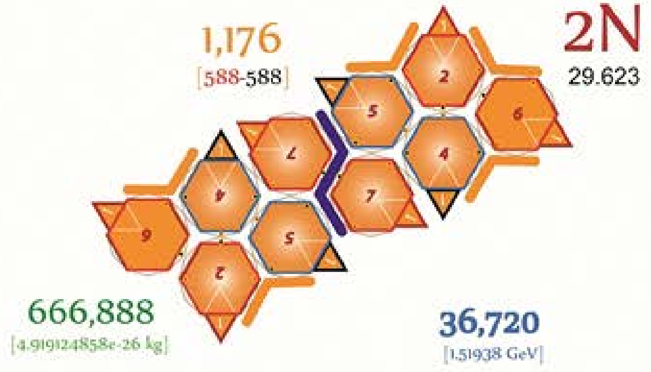
1,176
[588-588]
2N
29.623
2
5
6
4
7
5
9
2
666,888
[4.919124858e-26 kg]
36,720
[1.51938 GeV]

Nitrogen Allotropes & Isotopes

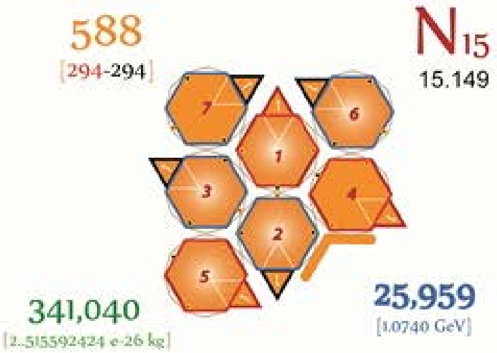
588
[294-294]
N_{15}
15.149
7
6
1
3
4
2
5
341,040
[2.515592424 e-26 kg]
25,959
[1.0740 GeV]

Isotopes are created by increasing the number of Planck mass-energies stored in the standing-wave geometries of chemical elements

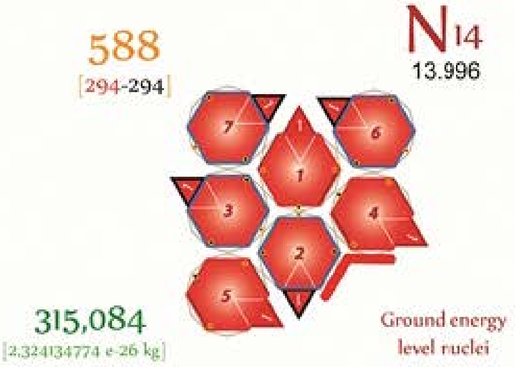
588
[294-294]
N_{14}
13.996
7
6
1
3
4
2
5
315,084
[2.324134774 e-26 kg]
Ground energy level nuclei

The charged mass-energy geometries of differing allotropic topologies create various chemical bond points

Shell			Molar meM	Nuclei per shell
R	8	521,472	23.164	7
Q	7	488,460	21.697	7
P	6	456,624	20.283	7
O	5	425,964	18.921	7
N	4	396,480	17.612	7
M	3	368,172	16.354	7
L	2	341,040	15.149	7
K	1	315,084	13.996	7
				7

energy level

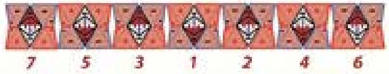

Copyright ABRAHAM [2008] - All rights reserved

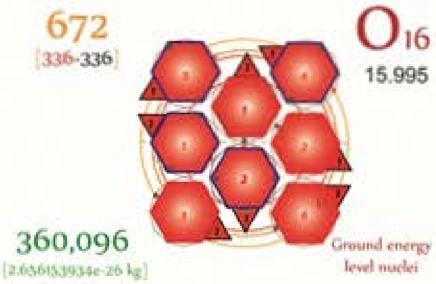
672
[336-336]
O16
15.995
360,096
[2.656153934e-26 kg]
Ground energy
level nuclei

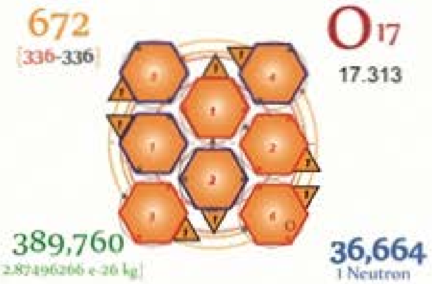
672
[336-336]
O17
17.313
389,760
2.87496266 e-26 kg]
36,664
1 Neutron

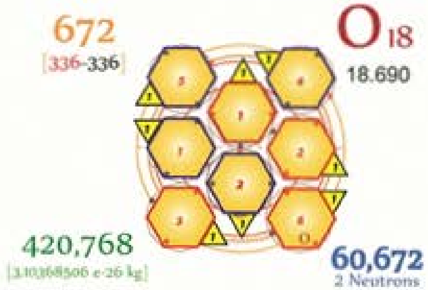
672
[336-336]
O18
18.690
420,768
[3.10368506 e-26 kg]
60,672
2 Neutrons

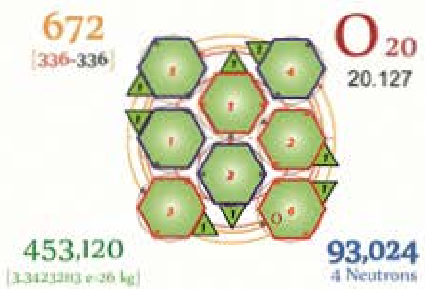
672
[336-336]
O20
20.127
453,120
[3.3423283 e-26 kg]
93,024
4 Neutrons

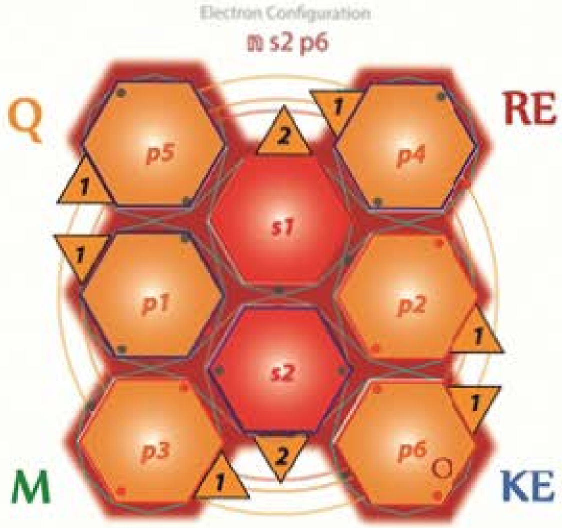
Electron Configuration
𝔪 s2 p6
Q
RE
M
KE
p5
p4
s1
p1
p2
s2
p3
p6
1
2
1
2

This oxygen allotrope's nuclei arrangement is very stable and inert like Neon [due to its filled sp electron orbitals]

Oxygen Allotropes & Isotopes

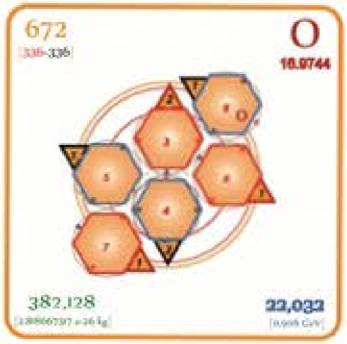
672
[336-336]
O
18.9744
382,128
[illegible]
22,032
[illegible]

There exists an allotrope of Oxygen formed when 8 nuclei combine in a single level SP orbital arrangement

Shell			Molar meM	Nuclei per shell
R	8	595,968	26.473	8
Q	7	558,240	24.797	8
P	6	521,856	23.181	8
O	5	486,816	20.825	8
N	4	453,120	20.127	8
M	3	420,768	18.690	8
L	2	389,760	17.313	8
K	1	360,096	15.995	8
				8

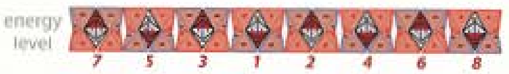
energy
level
7
5
3
1
2
4
6
8

It Is Increased Nucleonic energy levels that creates Isotopes (not extra Neutrons within a Nucleus)

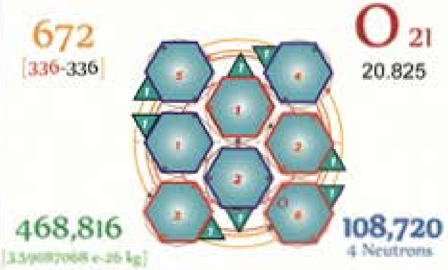
672
[336-336]
O21
20.825
468,816
[3.59087068 e-26 kg]
108,720
4 Neutrons

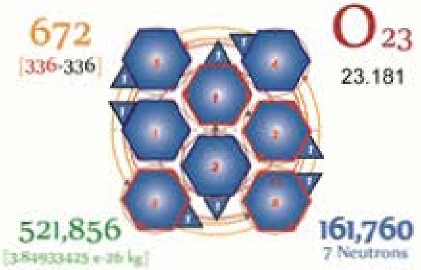
672
[336-336]
O23
23.181
521,856
[3.84933425 e-26 kg]
161,760
7 Neutrons

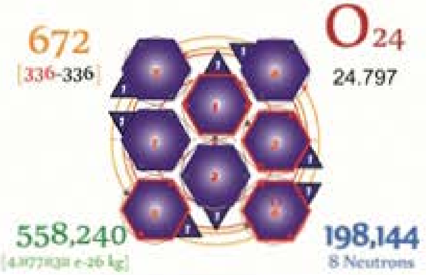
672
[336-336]
O24
24.797
558,240
[4.1177832 e-26 kg]
198,144
8 Neutrons

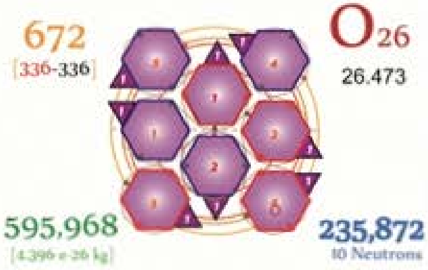
672
[336-336]
O26
26.473
595,968
[4.396 e-26 kg]
235,872
10 Neutrons

Tetryonics 54.15 - Oxygen isotopes

Copyright ABRAHAM [2008] - All rights reserved

Bose–Einstein condensates exist at the low energy extremes

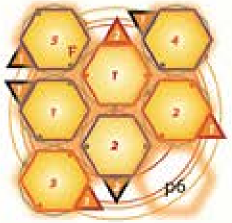

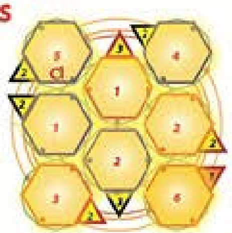

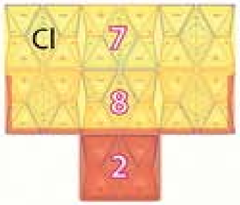

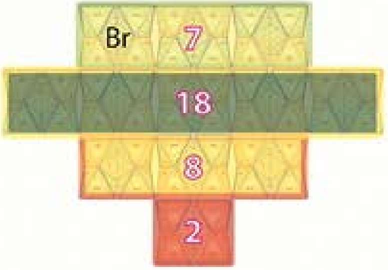

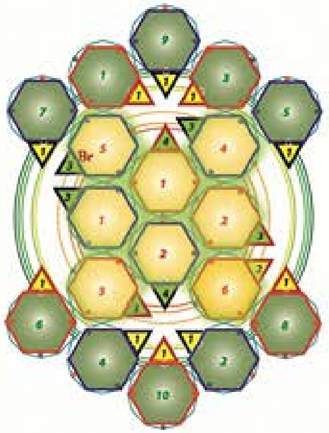

Gaseous

Liquid

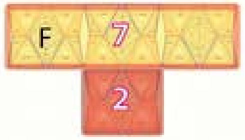

The Natural states of Matter

(room temperature)

Elements appear in their natural phase at room temperature as a gas, liquid, solid or synthetic

The halogen element family provides a great example of all 4 states of Matter at room temperature

As energy is added or removed from any element's charged Matter topology or their baryonic KEM field geometry their phase states & Brownian motion changes

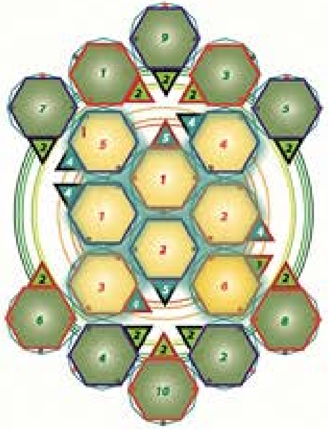

Solid

Radioactive

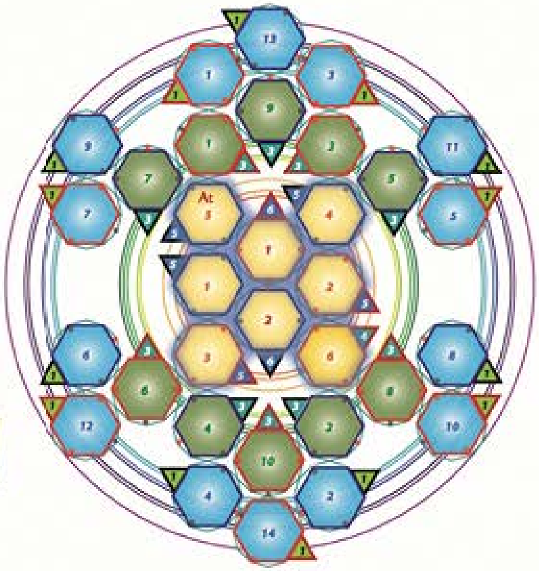

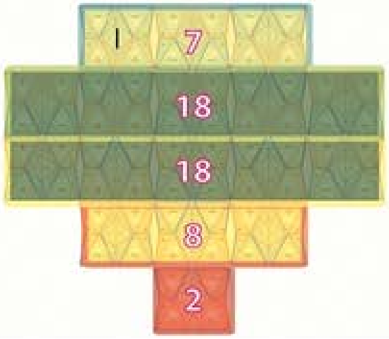

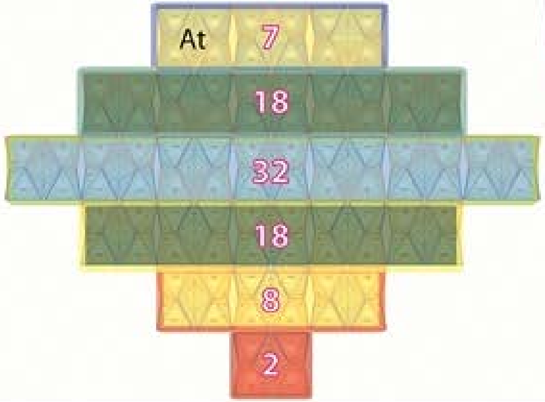

At the high energy extremes exist the Plasma states of Matter

Chemical bonds

A chemical bond is an attraction between atoms that allows
the formation of chemical substances that contain two or more atoms.

Particle charge & Inter-fascia electric interaction provides the foundation of chemical bonds

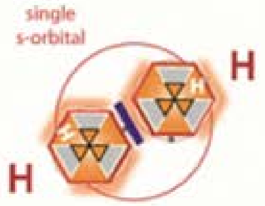
single
s-orbital
H
H

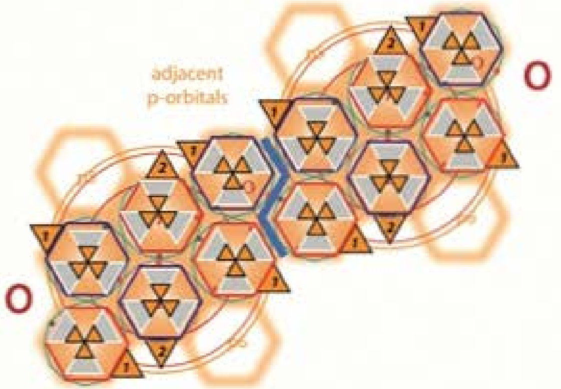
adjacent
p-orbitals
O
O

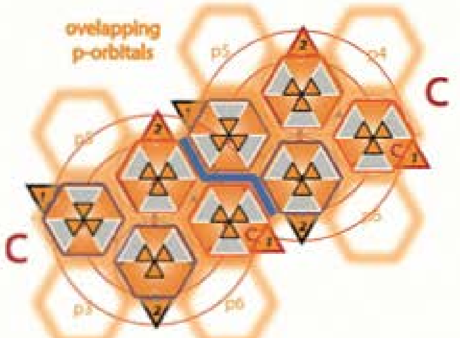
ovelapping
p-orbitals
C
C
p5
p4
p3
p6

It is usually taught that based on electron sharing principles a single bond contains 2 paired electrons, a double bond has 4 electrons and a triple bond has 6 shared electrons

filled electron shells
[weak KEM'ical bonds]

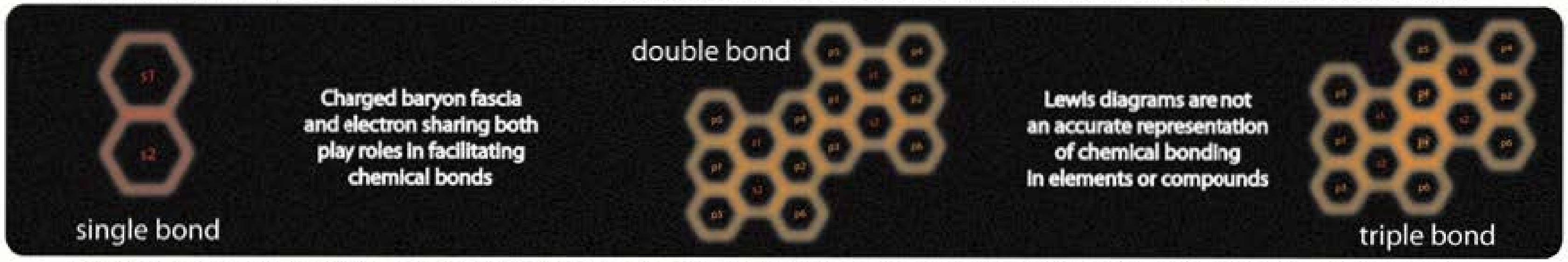
single bond
double bond
triple bond

Charged baryon fascia and electron sharing both play roles in facilitating chemical bonds

Lewis diagrams are not an accurate representation of chemical bonding in elements or compounds

unfilled electron shells
[stronger KEM'ical bonds]

Tetryonic geometry provides an advanced quantum topology for all atoms and compounds that is superior to the older Lewis diagrams and even molecular orbital theory,
facilitating a clear understanding of the quantum geometry of each element, its nuclear bonds and electron sharing in all compound chemical structures;
along with the core & valence electron configurations & interactive electric charge fascia bonding positions for each compound created.

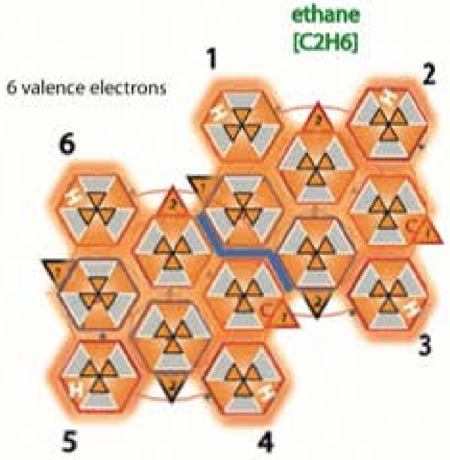
ethane
[C2H6]
6 valence electrons
1
2
3
4
5
6

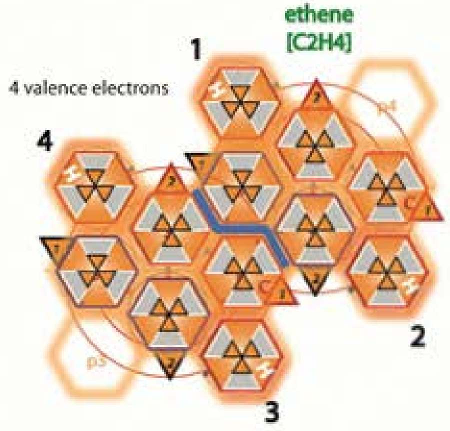
ethene
[C2H4]
4 valence electrons
1
2
3
4

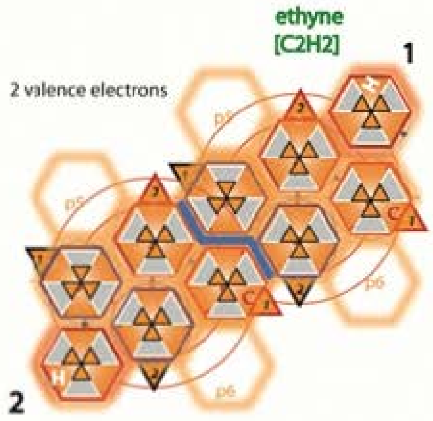
ethyne
[C2H2]
2 valence electrons
1
2

Valance electrons also play an important role in determining the strength of chemical bonds

Van der Waals Forces

Van der Waals forces include attractions between atoms, molecules, and surfaces. They differ from covalent and ionic bonding in that they are caused by correlations in the fluctuating polarizations of nearby particles

All charges seek equilibrium

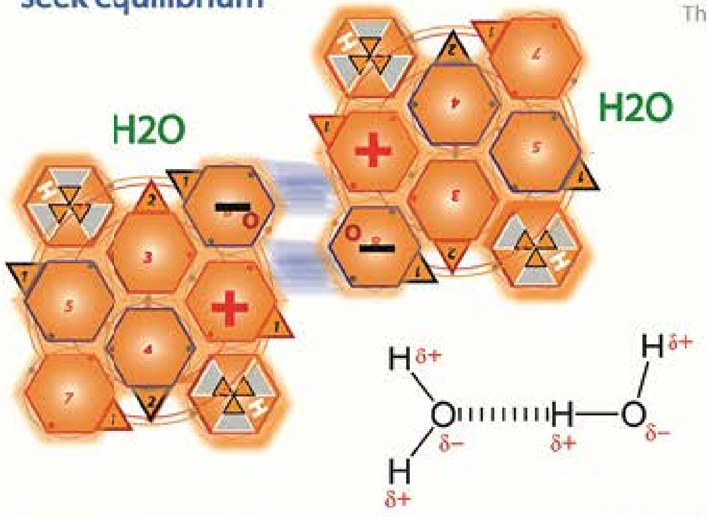

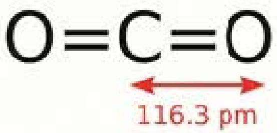

Van der Walls and London forces are the geometric fields of interaction between charged Matter topologies

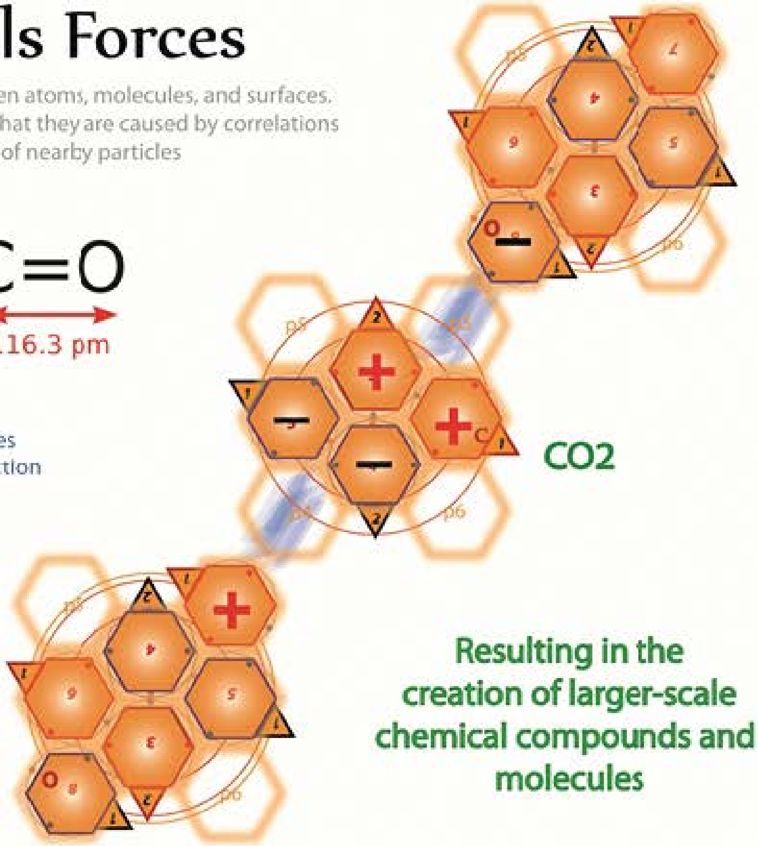

The Forces of interaction can be attractive or repulsive

Residual interactive bonding force between the charged fascia of atoms

Resulting in the creation of larger-scale chemical compounds and molecules

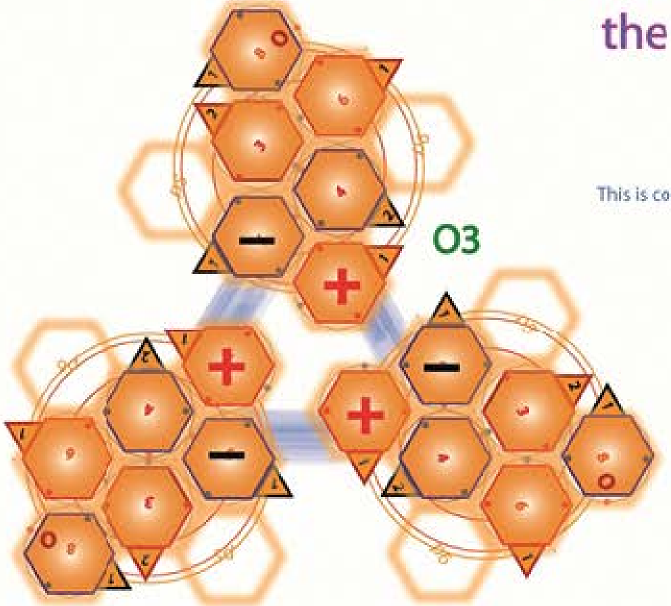

This is considered to be the only attractive intermolecular force present between neutral atoms (e.g., a noble gas configurations and charge neutral atomic topologies).

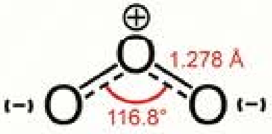

Without London forces, there would be no attractive force between noble gas atoms, and they wouldn't exist in liquid form.

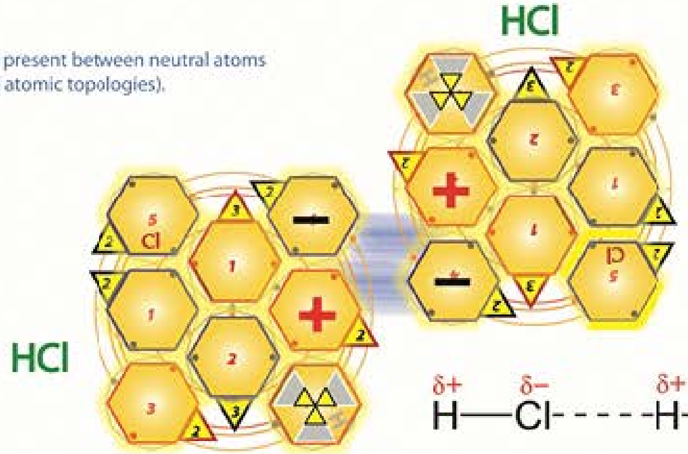

$$\overset{\delta+}{H}-\overset{\delta-}{Cl}\cdots\overset{\delta+}{H}-\overset{\delta-}{Cl}$$

Lewis Structures

The Lewis structure was named after Gilbert N. Lewis,
who introduced it in his 1916 article
The Atom and the Molecule.

Lewis structures, also called Lewis-dot diagrams,
are diagrams that show the bonding between the atoms of any molecule,
and the lone pairs of electrons that may exist in the molecule.

BeCl2

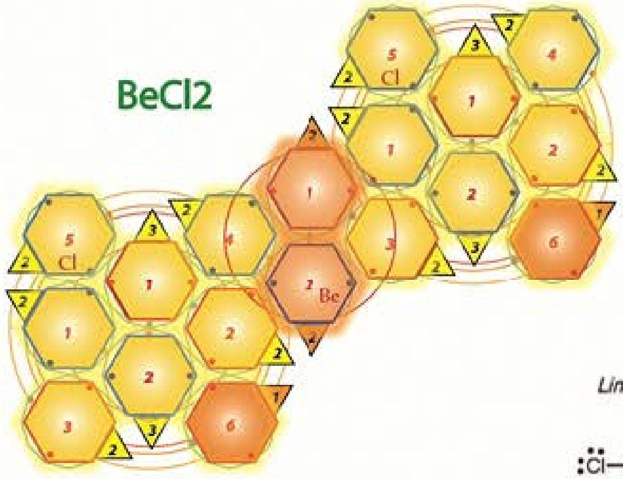

HF

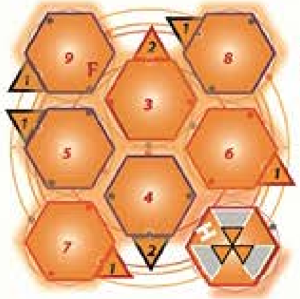

Line drawings can be used to depict molecular geometry:

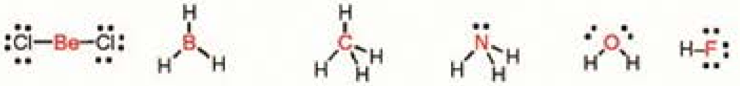

Beryllium dichloride $BeCl_2$ *linear*	Borane BH_3 *trigonal planar*	Methane CH_4 *tetrahedral*	Ammonia NH_3 *trigonal pyrimidal*	Water H_2O *bent*	Hydrofluoric acid HF *linear*

Also applies to molecules with multiple bonding:

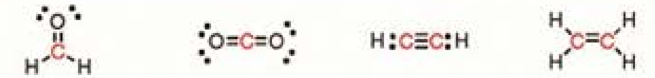

methanal (formaldehyde) *trigonal planar*	Carbon dioxide *linear*	ethyne (acetylene) *linear*	ethane (ethylene) *both carbons are trigonal planar*

Hydrogen is a 'free radical' atom whose energy can be changed to facilitate chemical bonding between elements

Tetryonic geometry uses the charged geometry of the element fascia themselves

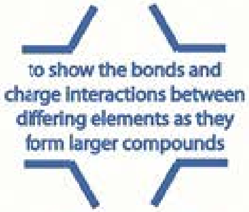

to show the bonds and charge interactions between differing elements as they form larger compounds

as well as the actual final quantum topology of compound elements and molecules

They are similar to electron dot diagrams in that the valence electrons in lone pairs are represented as dots, but they also contain lines to represent shared pairs in a chemical bond (single, double, triple, etc.).

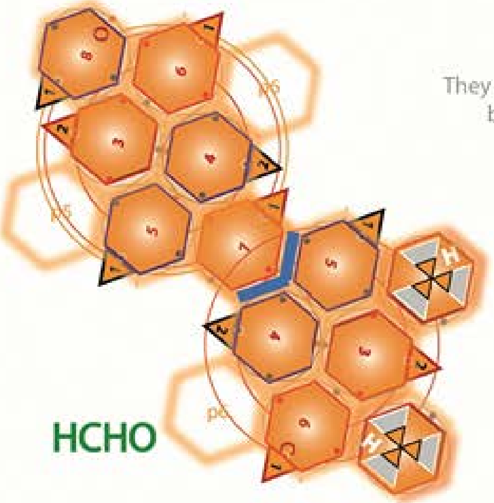

HCHO

single fascia bond	double fascia bonds	triple fascia bond

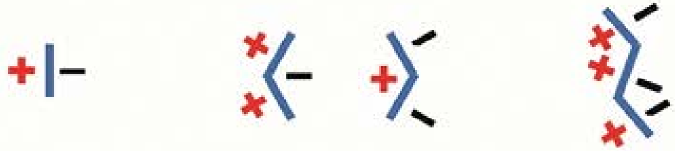

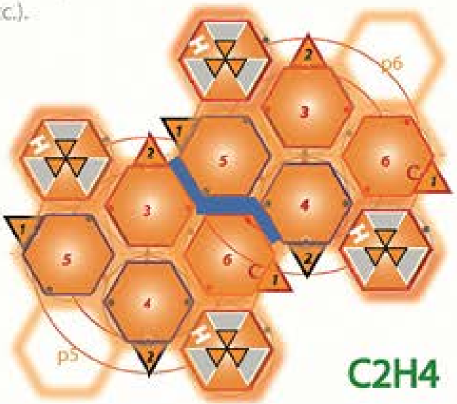

C2H4

it is the electric field fascia of baryons that facilitates chemical bonds

Atomic bonds

All atoms, elements and compounds seek stable core energy & configurations where their electron orbitals are filled

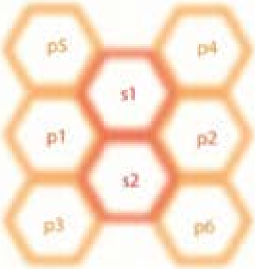

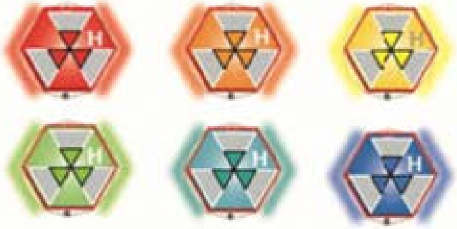

Hydrogen σ bonds
Outward presenting electic fascia bonds
Facilitate bonding between molecules

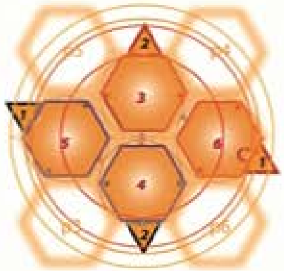

Molecular π bonds
Inward recieving bonds
capable of accepting Hydrogen
or extra-orbital elctric fascia bonds

Bonds fill in order of orbital filling
ie p1-6, d1-10

OH

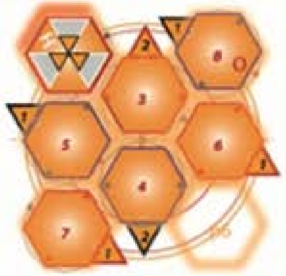

Hydroxide bonds
Oxygen-Hydrogen compound
creates a halogen-like topology
which seeks to fill its p6 orbital
in order to reach a stable electronic
configuration

Extremely reactive

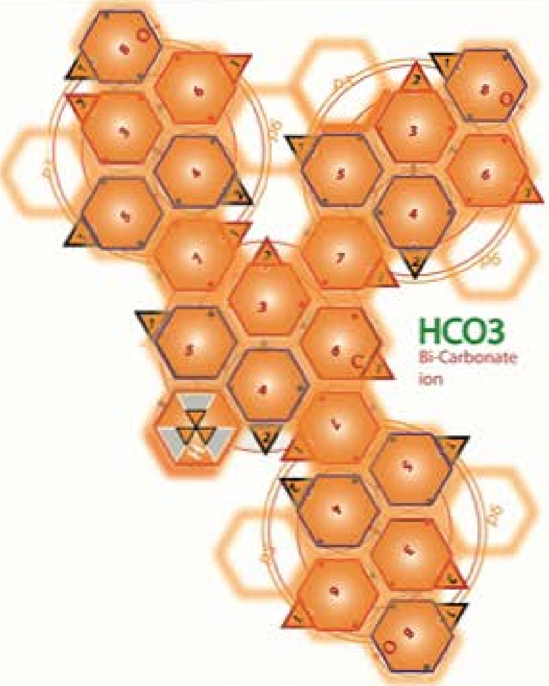

HCO3
Bi-Carbonate
ion

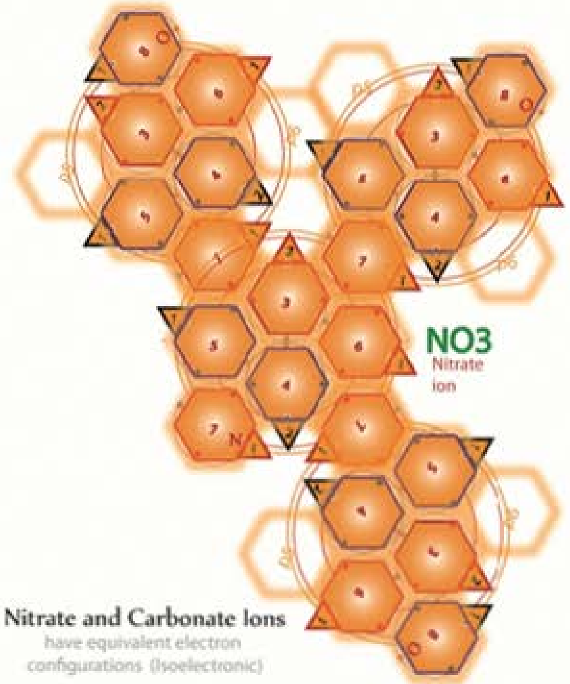

NO3
Nitrate
ion

Nitrate and Carbonate Ions
have equivalent electron
configurations (Isoelectronic)

CH4

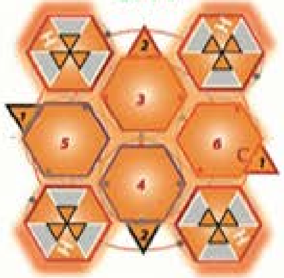

Covalent bonds
Intra-orbital bonding between
elements and compounds where
electrons exchange is the main
mechanisium resulting in
stable electronic configurations

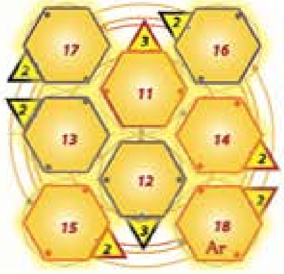

Core electron configuration
Unreactive [non-valence]
electron configuration

Ionic bonds
Extra-orbital bonding between
elements and compounds where
element charge attraction is the
predominent mechanism
with electrons sharing resulting in
stable electronic configurations

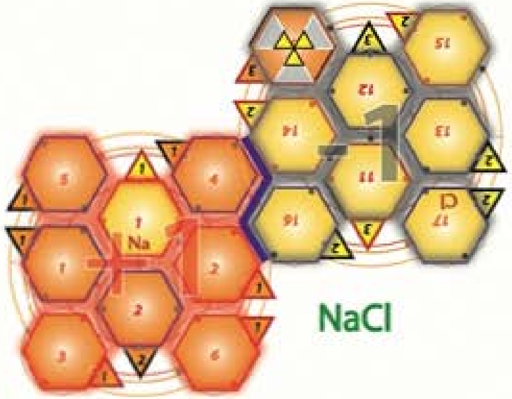

NaCl

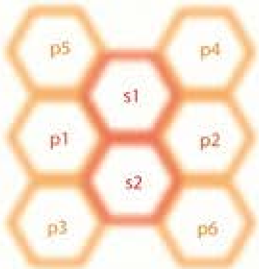

Atomic bonds

All atoms, elements and compounds seek stable core electron configurations where their electron orbitals are filled

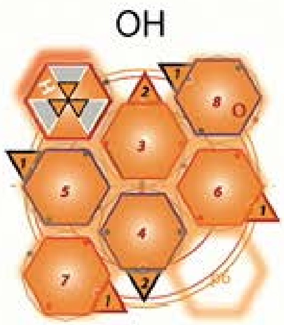

Hydroxide bonds
Oxygen-Hydrogen compound creates a halogen-like geometry which seeks to fill its p6 orbital in order to reach a stable electronic configuration

Extremely reactive

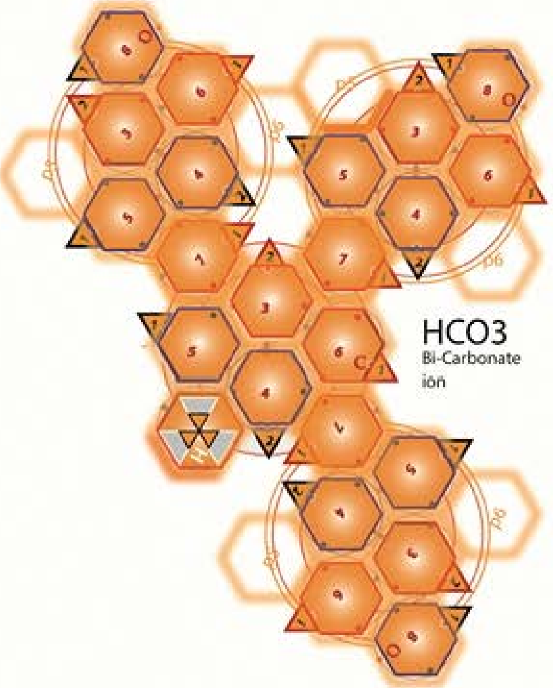

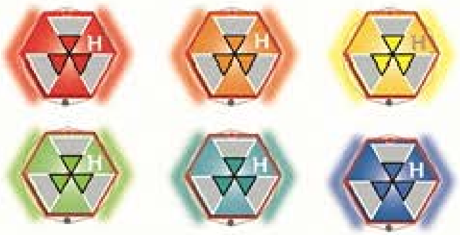

Hydrogen Bonds
Outward presenting electic fascia bonds
Facilitate bonding between molecules

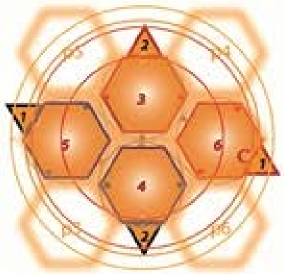

Molecular bonds
Inward recieving bonds
capable of accepting Hydrogen
or extra-orbital elctric fascia bonds

Bonds fill in order of orbital filling
ie p1-6, d1-10

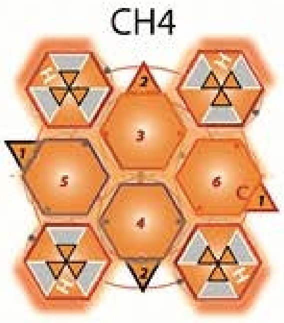

Ionic bonds
Extra-orbital bonding between elements and compounds where element charge attraction is the predominent mechanism with electrons sharing resulting in stable electronic configurations

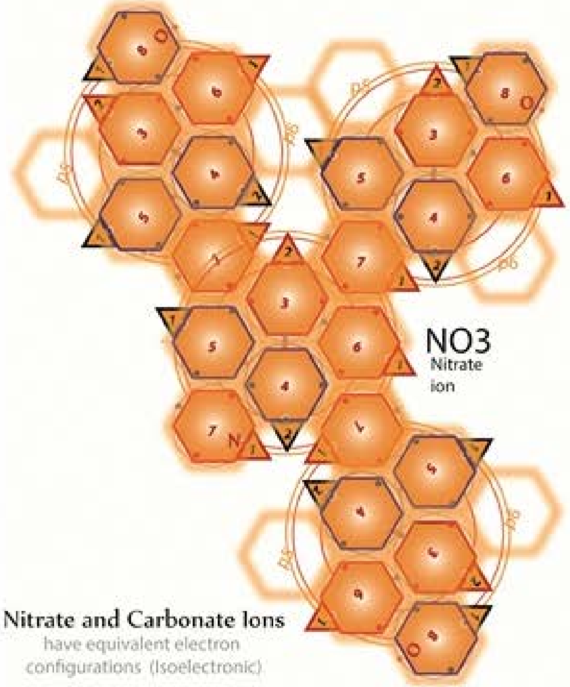

Nitrate and Carbonate Ions
have equivalent electron configurations (Isoelectronic)

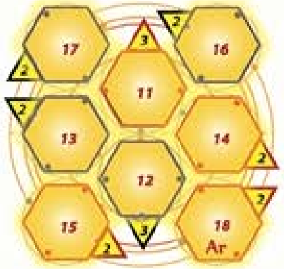

Core electron configuration
Unreactive [non-valence]
electron configuration

Covalent bonds
Intra-orbital bonding between elements and compounds where electrons exchange is the main mechanisim resulting in stable electronic configurations

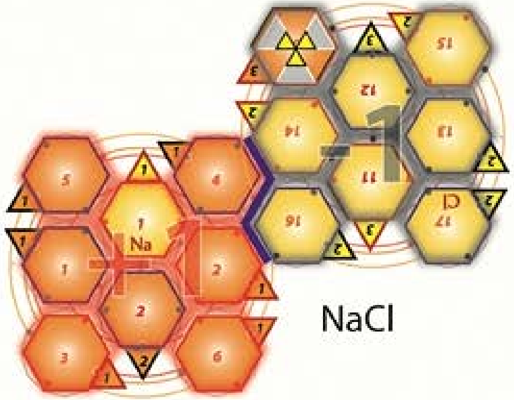

CORE ELECTRON configurations

			[He]
L	1s2	2s2 2p6	[Ne]
M	1s2 2s2 2p6	3s2 3p6	[Ar]
N	1s2 2s2 2p6 3s2 3p6 3d10	4s2 4p6	[Kr]
O	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10	5s2 5p6	[Xe]
P	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10	6s2 6p6	[Rn]
Q	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10 5f14 6s2 6p6 6d10	7s2 7p6	[Uuo]

The term "core electrons" or "noble gas core" refers to the electrons within the atom which have the same electron configuration as the nearest noble gas of lower atomic number and contain filled s&p orbitals

'Nobel Gas' configuration
full s,p orbitals

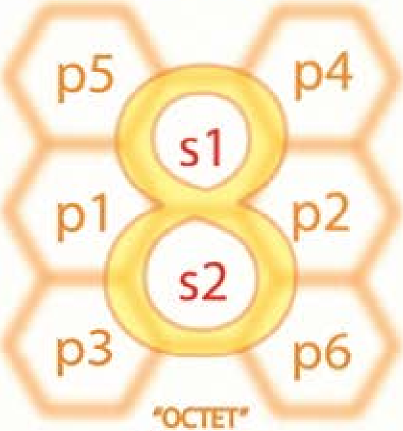
p5
p4
s1
p1
p2
s2
p3
p6

'core electron' configurations are present in energy levels L [2] - [7] Q

electrons not in filled or complete atomic orbitals are valence electrons

"OCTET"
No valance

Using core electron notation greatly reduces complexity of the electron configuration notations of larger elements in the periodic table

[Ub] 1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 5s2 4d10 5p6 4f14 5d10 6s2 6p6 5f14 6d10 7s2 7p6 8s2
120

2 8 18 32 32 18 8 2

They are the electrons in the inner part of the atom that are not valence electrons and therefore do not participate in bonding

[Uuo] 8s2
118

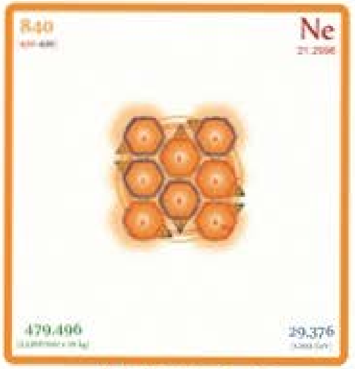
840
Ne
479,496
29,376

[He] 2s2 2p6
2

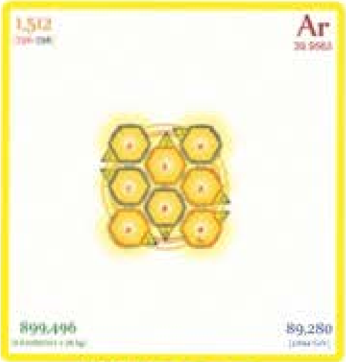
1,512
Ar
899,496
89,280

[Ne] 3s2 3p6
10

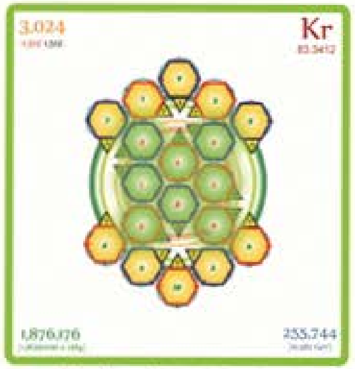
3,024
Kr
1,876,176
255,744

[Ar] 3d10 4s2 4p6
18

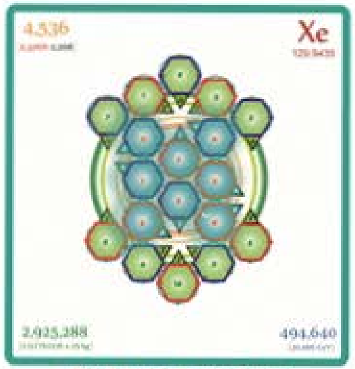
4,536
Xe
2,925,288
494,640

[Kr] 4d10 5s2 5p6
36

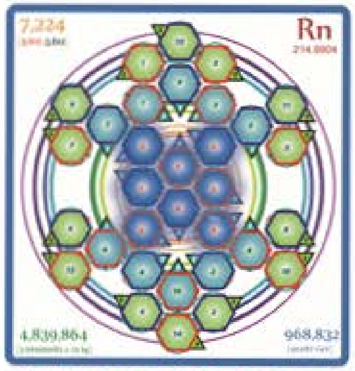
7,224
Rn
4,839,864
968,832

[Xe] 4f14 5d10 6s2 6p6
54

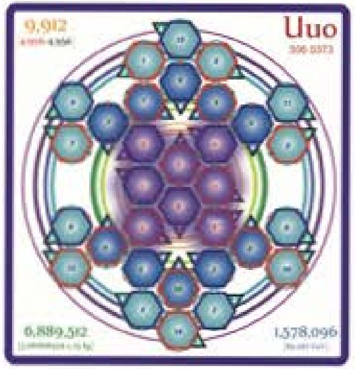
9,912
Uuo
6,889,512
1,578,096

[Rn] 5f14 6d10 7s2 7p6
86

Core electron configurations

Core electrons are the electrons in an atom that are not valence electrons and therefore do not participate in bonding.

Reactive

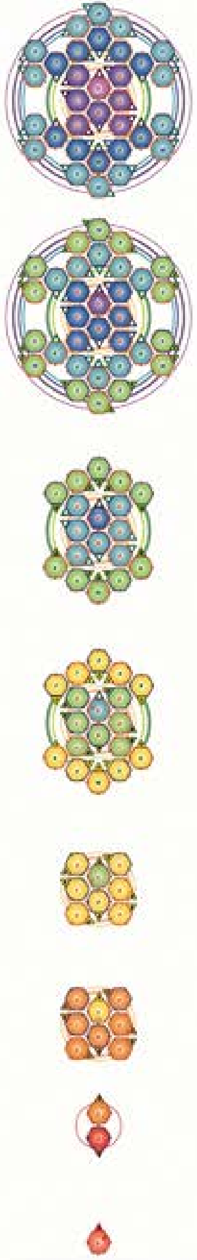

Elements with more nuclei than the core configurations are reactive

Reactive elements combine with other elements seeking core electron configurations

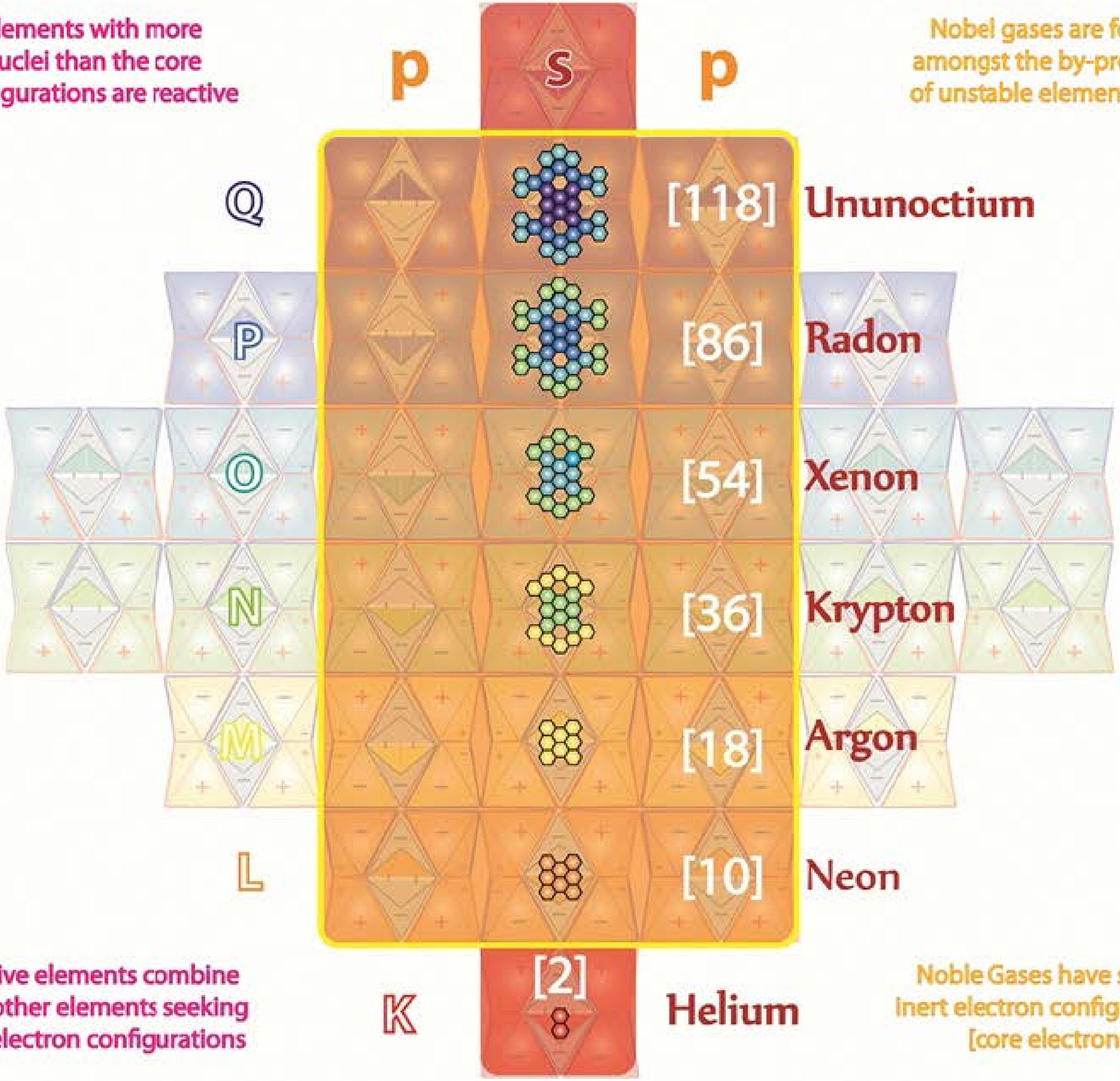
p
s
p
Q
[118]
Unununoctium
P
[86]
Radon
O
[54]
Xenon
N
[36]
Krypton
M
[18]
Argon
L
[10]
Neon
K
[2]
Helium

Inert

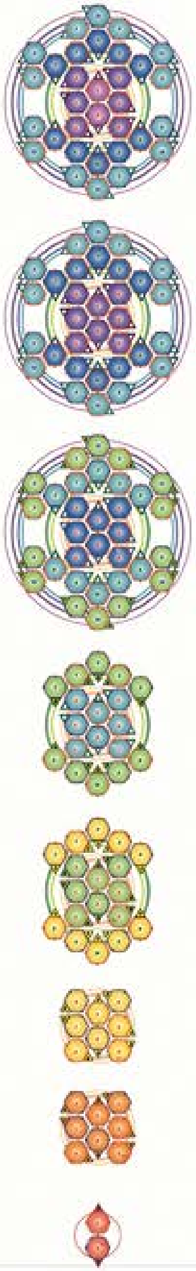

Nobel gases are found amongst the by-products of unstable element decay

Noble Gases have stable inert electron configurations [core electrons]

Copyright ABRAHAM [2008] - All rights reserved

!

Helium is a
unique noble gas
in that it only has
filled s-orbitals

suggesting that it
be more appropriate
to designate it as an
alkaline earth gas

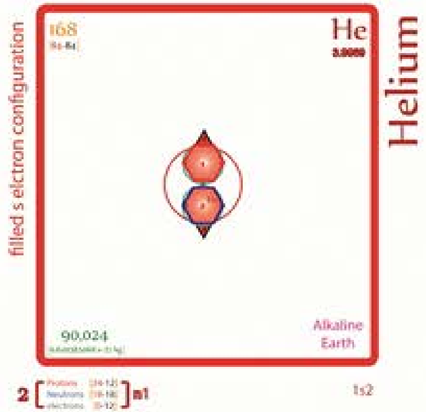
filled s elctron configuration
168
He
Helium
90,024
Alkaline
Earth
2
Protons
Neutrons
electrons
n1
1s2

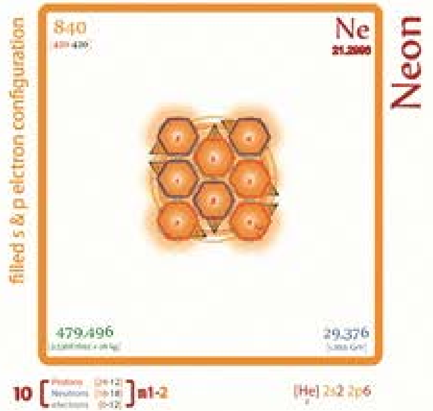
filled s & p elctron configuration
840
Ne
Neon
479,496
29,376
10
Protons
Neutrons
electrons
n1-2
[He] 2s2 2p6

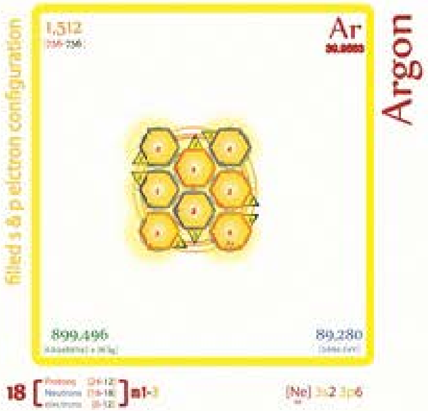
filled s & p elctron configuration
1,512
Ar
Argon
899,496
89,280
18
Protons
Neutrons
electrons
n1-3
[Ne] 3s2 3p6

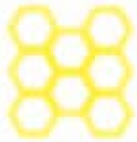

Nobel gases all have
inert core sp electron
configurations

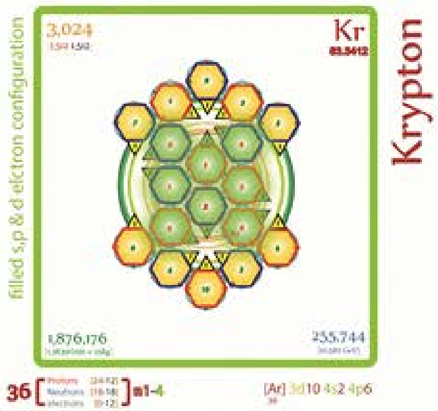
filled s,p & d elctron configuration
3,024
Kr
Krypton
1,876,176
255,744
36
Protons
Neutrons
electrons
n1-4
[Ar] 3d10 4s2 4p6

The Nobel Gases

are nuclei whose electron orbitals are completely filled
(and consequently have no valence electrons for bonding)

2			[He]
10	1s2	2s2 2p6	[Ne]
18	1s2 2s2 2p6	3s2 3p6	[Ar]
36	1s2 2s2 2p6 3s2 3p6 3d10	4s2 4p6	[Kr]
54	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10	5s2 5p6	[Xe]
86	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10	6s2 6p6	[Rn]
118	1s2 2s2 2p6 3s2 3p6 3d10 4s2 4p6 4d10 4f14 5s2 5p6 5d10 5f14 6s2 6p6 6d10	7s2 7p6	[Uuo]

All nobel gases have filled electron orbitals creating inert, chemically un-reactive elements

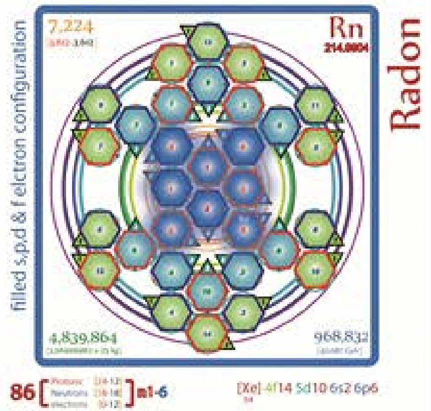
filled s,p,d & f elctron configuration
7,224
Rn
Radon
4,839,864
968,832
86
Protons
Neutrons
electrons
n1-6
[Xe] 4f14 5d10 6s2 6p6

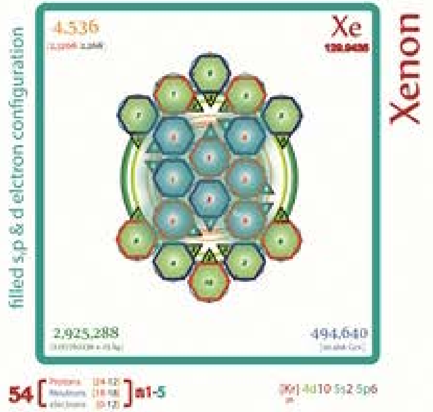
filled s,p & d elctron configuration
4,536
Xe
Xenon
2,925,288
494,640
54
Protons
Neutrons
electrons
n1-5
[Kr] 4d10 5s2 5p6

The noble gases are a group of chemical elements
with very similar properties: under standard conditions,
they are all odorless, colorless, monatomic gases, with
a very low chemical reactivity.

Whilst some of the nobel gases are often termed core electron configurations
the large elements also have filled d & f orbitals in them

The six noble gases that occur naturally are
Helium (He), Neon (Ne), Argon (Ar), Krypton (Kr),
Xenon (Xe), and the radioactive noble gases are
Radon (Rn) and Ununioctium (Uuo).

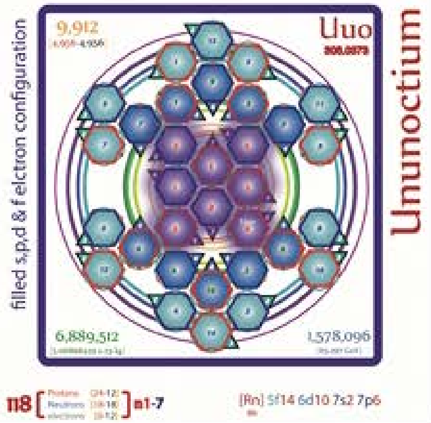
filled s,p,d & f elctron configuration
9,912
Uuo
Ununoctium
6,889,512
1,578,096
118
Protons
Neutrons
electrons
n1-7
[Rn] 5f14 6d10 7s2 7p6

Molecular Bonds

Hydrogen bonds

A weak, primarily electrostatic, bond between a hydrogen atom bound to a highly electronegative element in a given molecule and a second highly electronegative atom in another molecule

Diatomic bonds are fascia interactions between molecules with the same nuclei number [O2, F2, I2, Cl2]

Sigma bonds

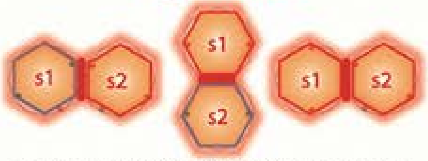
s1
s2
s1
s2
s1
s2

The residual EM force between opposite charge fascia in nuclei holds individual nuclei together in turn resulting in larger molecules and compounds

Pi bonds

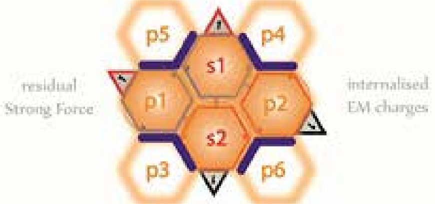
p5
p4
s1
residual Strong Force
p1
p2
internalised EM charges
s2
p3
p6

Molecular bonds

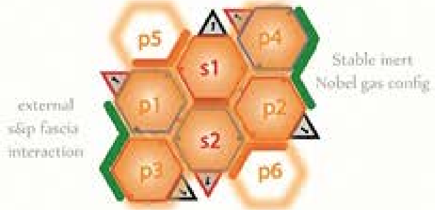
p5
p4
s1
Stable inert Nobel gas config
external s&p fascia interaction
p1
p2
s2
p3
p6

A molecular bond is an attractive force between nuclei that allows two or more elements or atoms to combine into a singular, complex compound structure

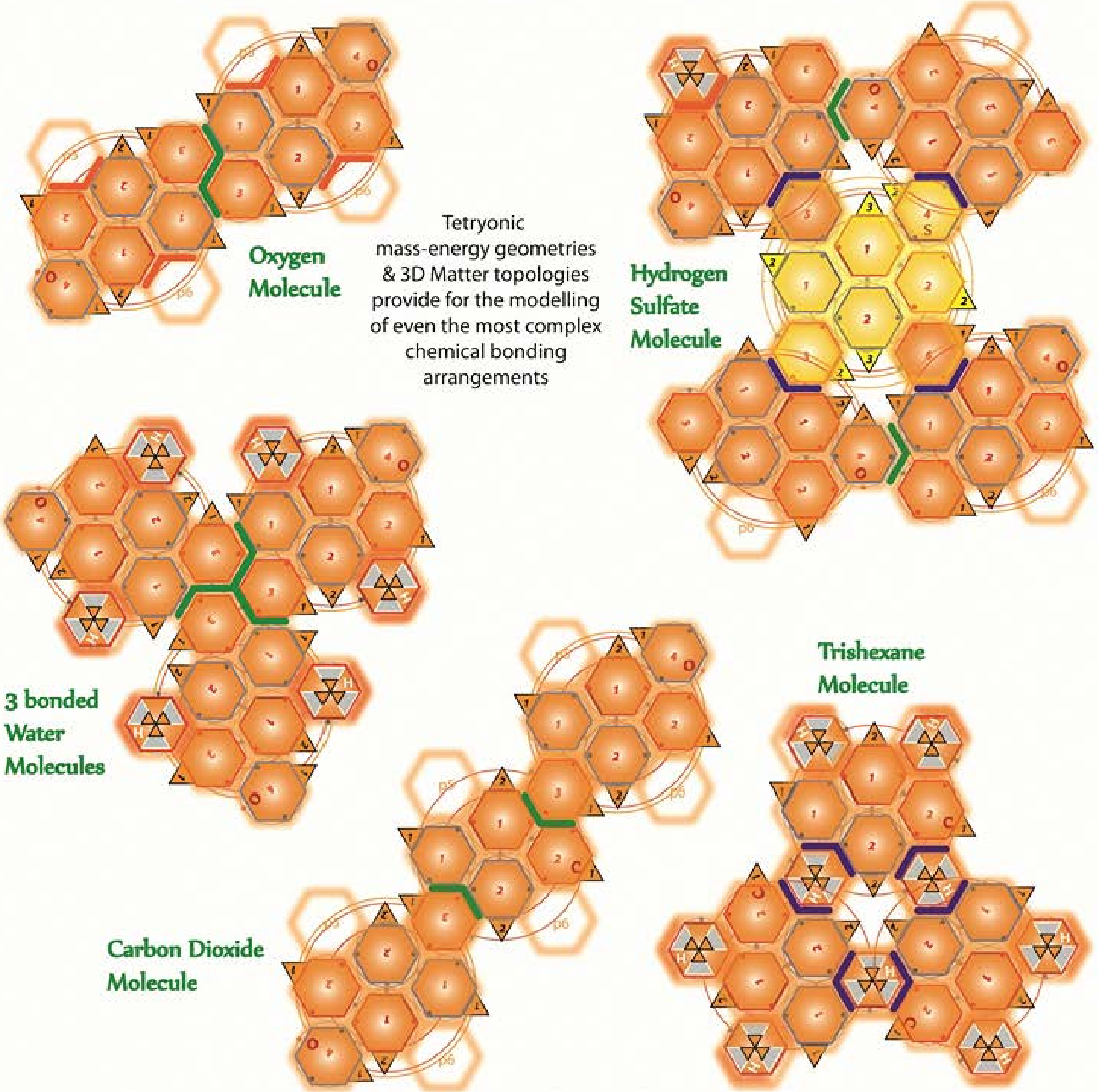

Oxygen Molecule

Tetryonic mass-energy geometries & 3D Matter topologies provide for the modelling of even the most complex chemical bonding arrangements

Hydrogen Sulfate Molecule

3 bonded Water Molecules

Carbon Dioxide Molecule

Trishexane Molecule

Covalent bonds are chemical links between two atoms in which electrons are shared between them.

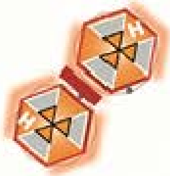

Sigma bonds are covalent bonds formed by direct overlapping of two adjacent atom's outermost orbitals.

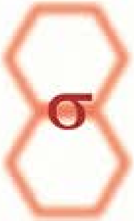

Note: sigma bonds can be formed by the bonding of either s-orbitals [or two adjacent p-orbitals]

Lewis electron dot diagrams fail to illustrate reality in that molecules exist as 3D objects and not as a two dimensional systems as shown by them.

Tetryonic geometry & topologies provide a polar view of 3D atomic nuclei that can be viewed as an exact representation of what a molecule & its bonds would look like when viewed from above.

Sigma & Pi covalent bonds

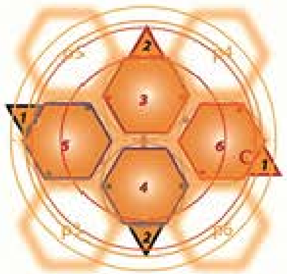

A pi bond is a covalent bond formed between two neighboring atom's unbonded p-orbitals.

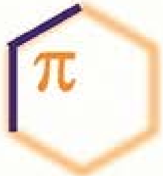

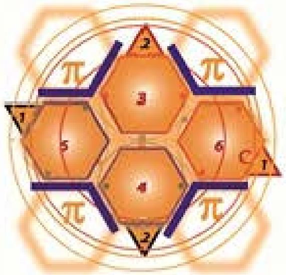

The single electrons from each atom's p orbital combine to form an electron pair creating the sigma bond.

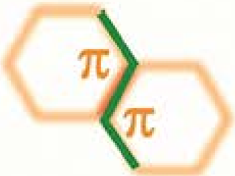

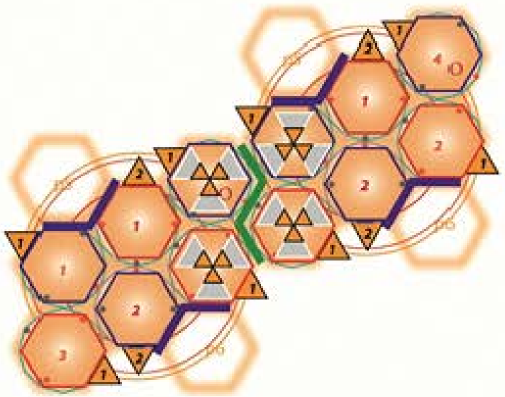

Double and triple bonds between atoms are usually made up of a single sigma bond and one or two pi bonds

π—σ—π

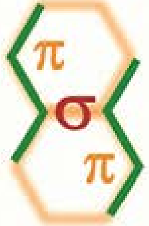

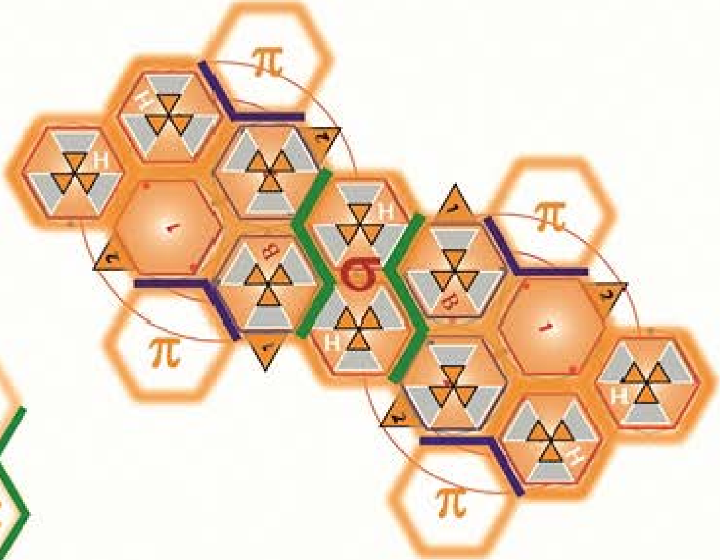

Alkali Metals

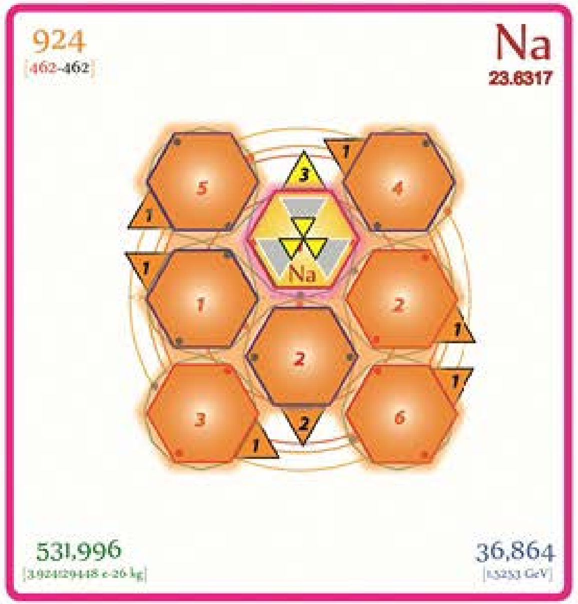
924
[462-462]
Na
23.6317
5
4
3
1
1
Na
1
2
2
3
2
6
1
531,996
[3.924129448 e-26 kg]
36,864
[1.5253 GeV]

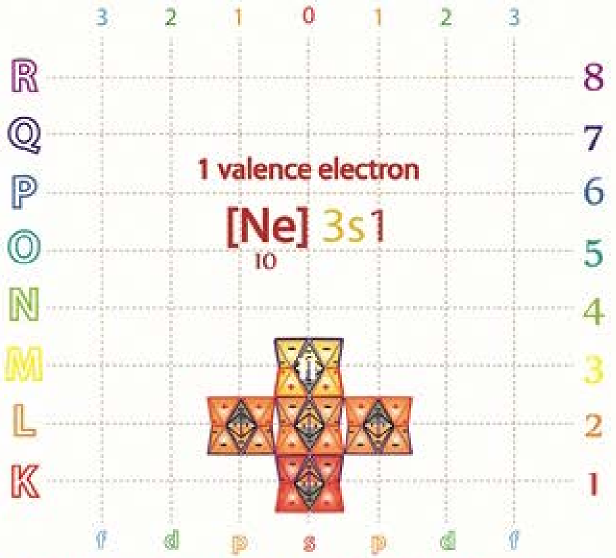
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1 valence electron
[Ne] 3s1
10
1
f d p s p d f

Atoms with more electrons than a closed shell are also highly reactive, as the extra valence electrons are easily removed from that orbital (to form a positive ion)

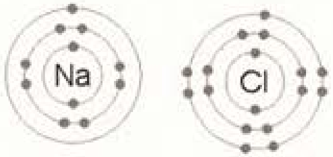

Valence electrons

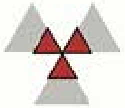

are the highest energy electrons in an atom forming the outermost electrons of an atom, and are important in determining how the atom reacts chemically with other atoms

Historically, the number of valence electrons was reflected by the element's group number in the Mendeleev table and formed the basis of elemental families

Tetryonic topologies now replace the older, incorrect models of valence electron configurations with the full 3D modelling of all atoms, elements & compounds

[zz] are core electron configurations

providing a superior visual means of accurately determining the energies and position of any electron in chemical compounds and determining the valence numbers

Halogens

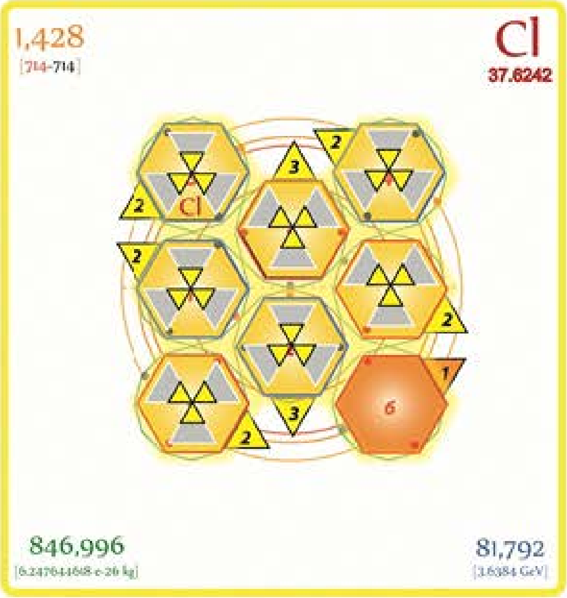
1,428
[714-714]
Cl
37.6242
2
3
2
Cl
2
2
6
1
3
2
846,996
[6.247644648 e-26 kg]
81,792
[3.6384 GeV]

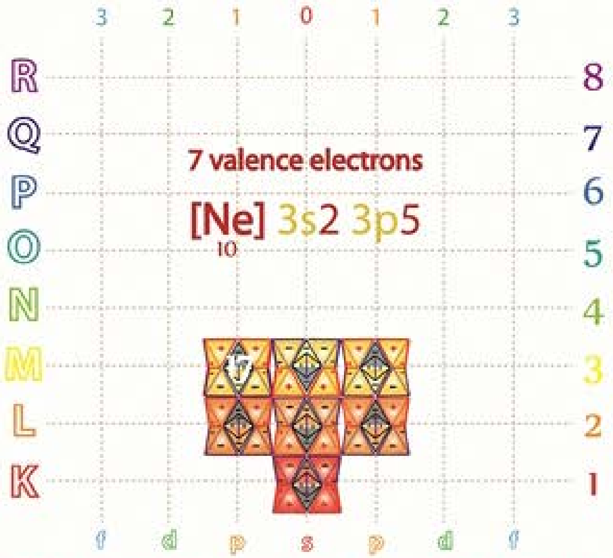
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7 valence electrons
[Ne] 3s2 3p5
10
7
f d p s p d f

Atoms one or two valence electrons short of a closed shell are highly reactive, due to their tendency to seek to gain the missing valence electrons (thereby forming a negative ion)

Copyright ABRAHAM [2008] - All rights reserved

Hydro-Chloric acid

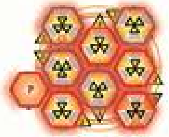

HCl

HNO3

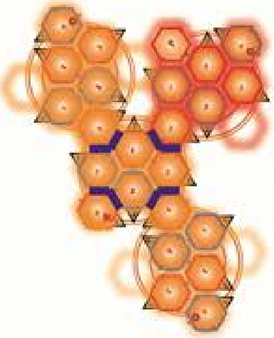

Nitric acid

Carbonic acid

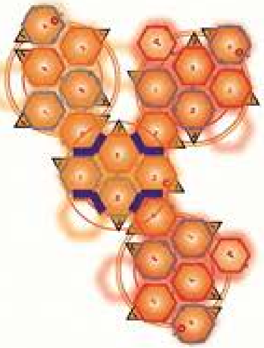

H2CO3

H2SO4

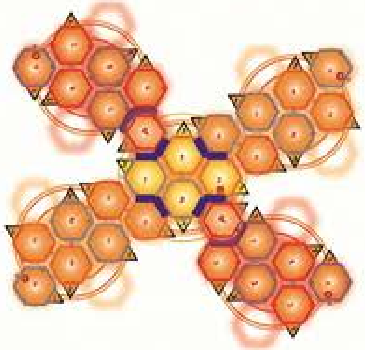

Sulfuric acid

Common Acids

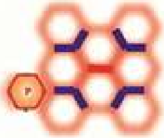

The Arrhenius definition defines acids as substances which increase the concentration of hydrogen ions (H+), or more accurately, hydronium ions (H3O+), when dissolved in water.

The Brønsted-Lowry definition is the most widely used definition where acid-base reactions are assumed to involve the transfer of a proton (H+) from an acid to a base.

A Lewis acid is a substance that can accept a pair of electrons to form a covalent bond.

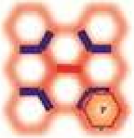

HC2H3O2

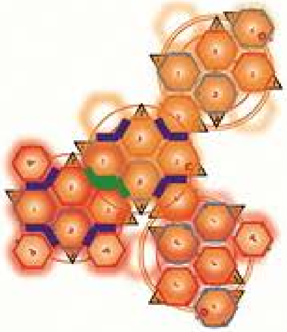

Acetic acid

Phosphoric acid

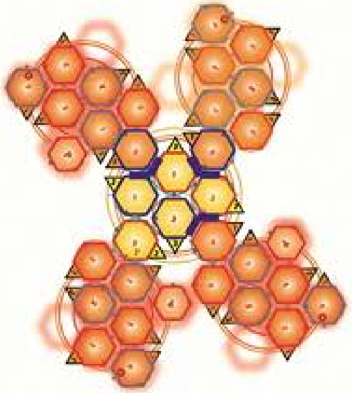

H3PO4

H2C2O4

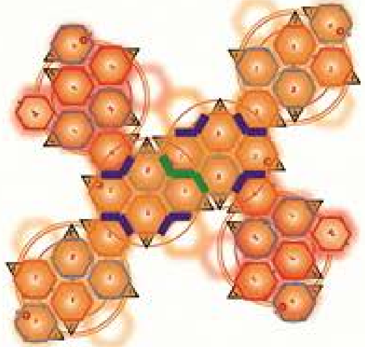

Oxalic acid

Citric acid

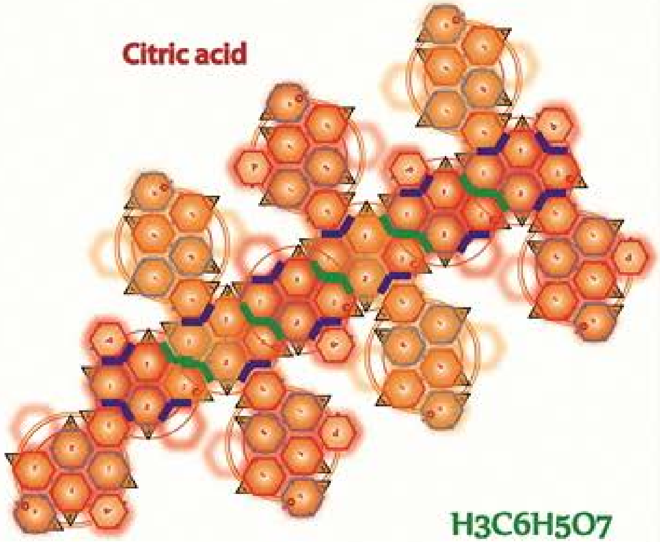

H3C6H5O7

Chemical reactions

Valence Bonds & Molecular orbitals

A chemical reaction is a process that leads to the transformation of one set of chemical substances to another.

Valence bond theory focuses on how the atomic orbitals of the dissociated atoms combine in molecular formations to give rise to individual chemical bonds.

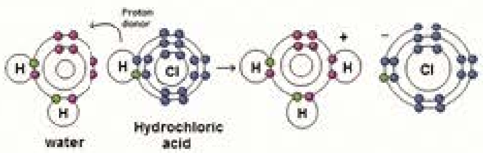

In contrast, molecular orbital theory has orbitals that are designed to cover the whole molecule but is mathematically more complex in application

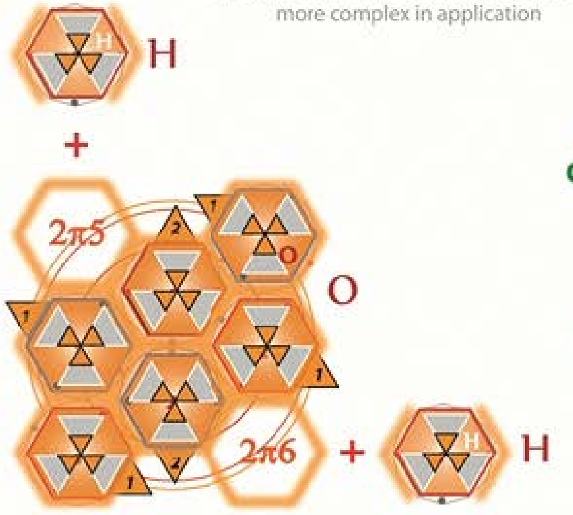

6 valence electrons
2 [n2] valence bonds available
(8 electron stable configuration)

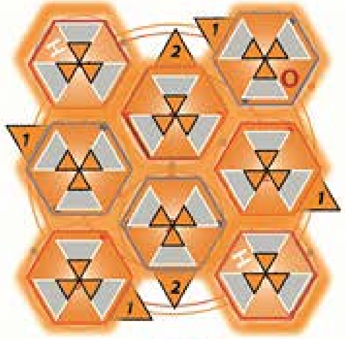

H_2O

water

Hydrogen Chloride gas

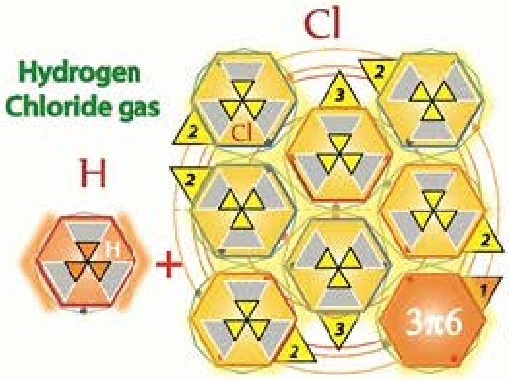

7 valence electrons
1 [n3] valence bond available
(8 electron stable configuration)

In order to form Hydro-chloric acid [HCl aq] and bond with similar water [H2O] molecules, the n2 Hydrogen atom's energy must be raised so it can migrate to the [3π6] orbital and bond

[n2] Hydrogen atom cannot bond in the [3π6] valence position so in HCl gas it forms a [n2] di-atomicbond instead and its electron migrates to the [π6] orbital

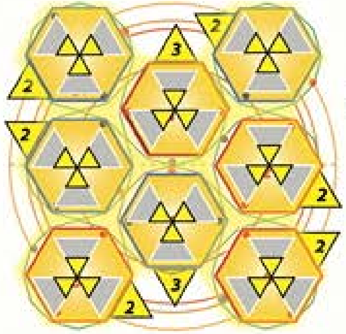

$HCl_{[aq]}$

Hydrochloric acid

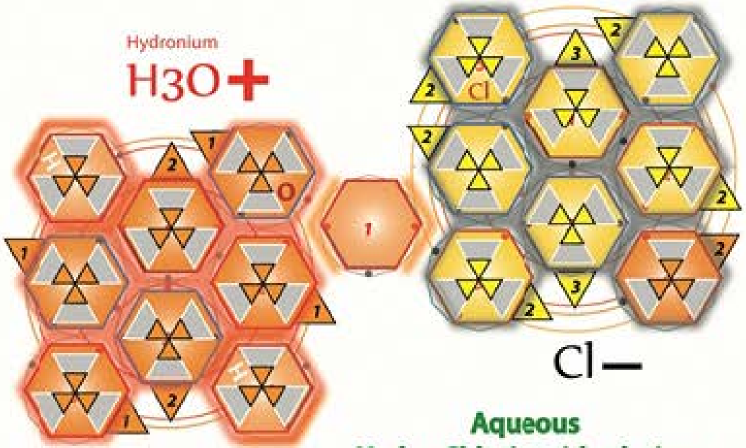

Aqueous Hydro-Chloric acid solution

Tetryonic theory replaces both VB & MO theories by modelling the physical quantum charge topologies and rest mass-energy geometry of all elements and compounds

Valence configurations seek to create stable filled obital configurations

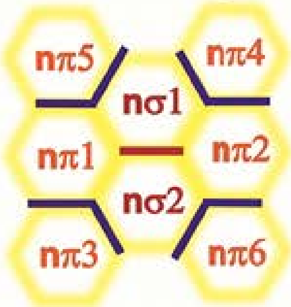

Building on exisiting electron configuration nomenclature tetryonic theory provides a new bonding schema for elements and complex compounds alike that is fully reflective of the physical quantum charge interactions

Ionic Bonding

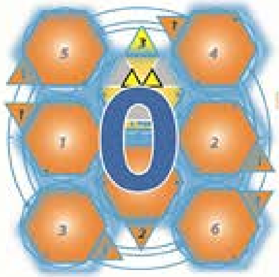
0
960
[480-480]

Alkaline Metal

Sodium atom gives up an electron in order to create a stable [Ne] nobel gas configuration

Sodium atom develops into a nett POSITIVE charge ion

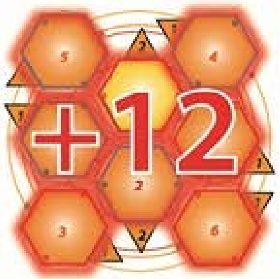
+12
948
[480-468]

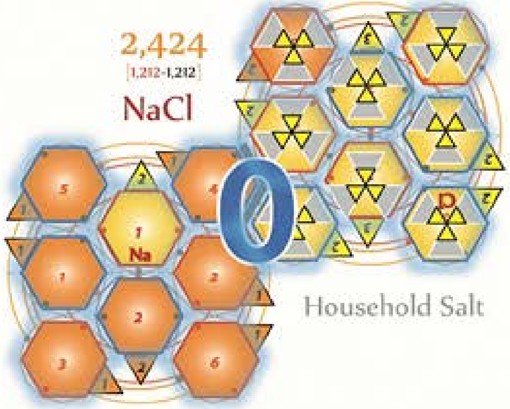
2,424
[1,212-1,212]
NaCl
0
Household Salt

Resultant salt [NaCl] molecule has a NEUTRAL charge with the n3 electron migrating to fill the 3p6 orbital

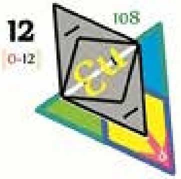

OPPOSITES ATTRACT

Chlorine atom seeks an electron to order to create a stable [Ar] nobel gas configuration

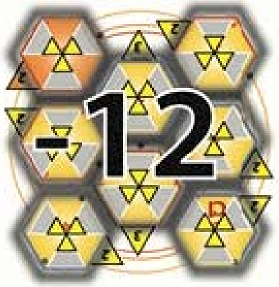
-12
1,476
[732-744]

Chlorine atom develops into a nett NEGATIVE charge ion

Halogen

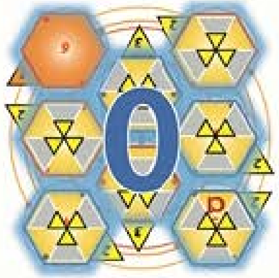
0
1,464
[732-732]

All atoms and molecules seek equilibrium via stable electron configurations

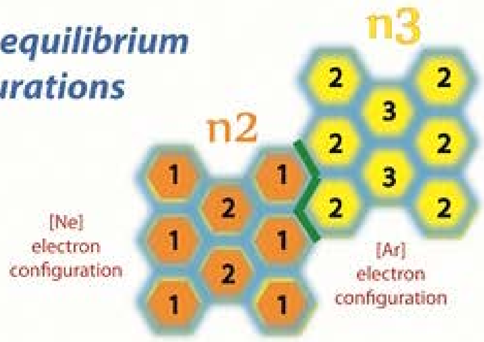
n3
n2
2
2
3
2
2
3
2
2
1
1
2
1
1
2
1
1
[Ne] electron configuration
[Ar] electron configuration

In short, it is a bond formed by the attraction between two oppositely charged ions.

An ionic bond (or electrovalent bond) is a type of chemical bond that can often form between metal and non-metal ions (or polyatomic ions such as ammonium) through electrostatic attraction.

The metal donates one or more electrons, forming a positively charged ion or cation with a stable electron configuration.
These electrons then enter the non metal, causing it to form a negatively charged ion or anion which also has a stable electron configuration.
The electrostatic attraction between the oppositely charged ions causes them to come together and form a bond.

Valence electron rules

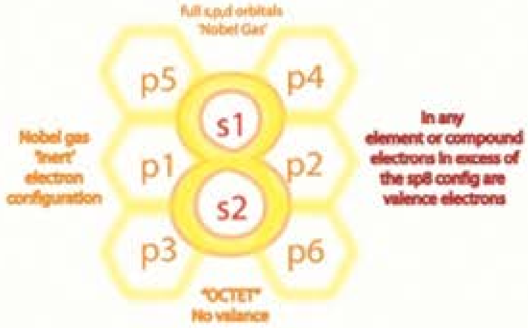

In any
element or compound
electrons in excess of
the sp8 config are
valence electrons

sp8 - core electron orbital grouping

*Tetryonic quantum charge topologies
provide the complete picture*

Valence electrons are the highest quantum level electrons in elemental or compound atom nuclei in excess of the core sp8 'noble gas' electron configuration

They are located in the outer most shells (ie K-Q or Quantum 1-7) making them easily ionised from the elemental nuclei

s 1-2 [electron orbital pair]
p 1-2, 3-4, 5-6 [electron orbital pairs]
d 1-2, 3-4, 5-6, 7-8 , 9-10 [electron orbital pairs]
f 1-2, 3-4, 5-6, 7-8, 9-10, 11-12, 13-14 (electron orbital pairs]

Valance electron congigurations always include 's' orbital electrons due to their higher quantum levels

Developed from Lewis diagrams
is usually only useful for elements Z<20

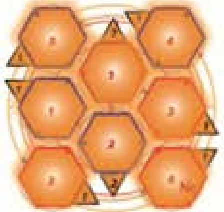

Neon
[He] 2s2, 2p6

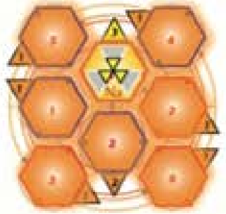

Sodium
[Ne] 3s1
Valence - 1

Fluorine
[He] 2s2, 2p5
Valence - 7

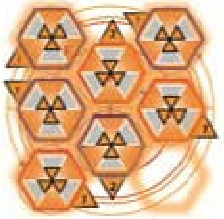

Scandium
[Ar] 3d1, 4s2
Valence - 3

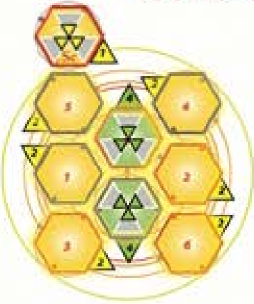

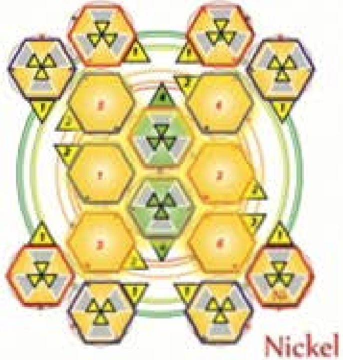

Nickel
[Ar] 3d8, 4s2
Valence - 10

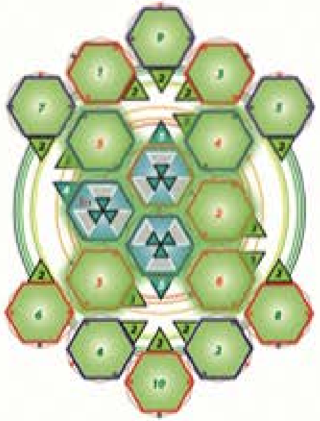

Indium
[Kr] 4d10, 5s2, 5p1
Valence - 3

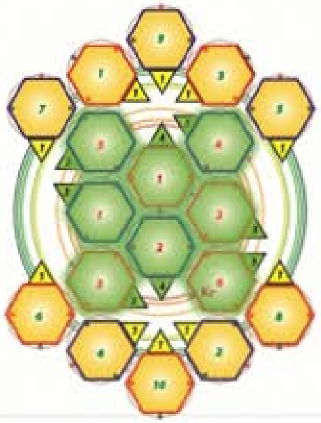

Krypton
[Ar] 4s2, 4p6

Gold
[Xe] 4f14, 5d9, 6s2
Valence - 11

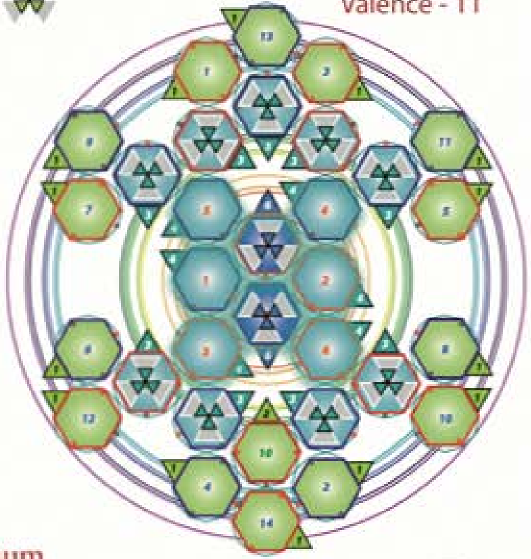

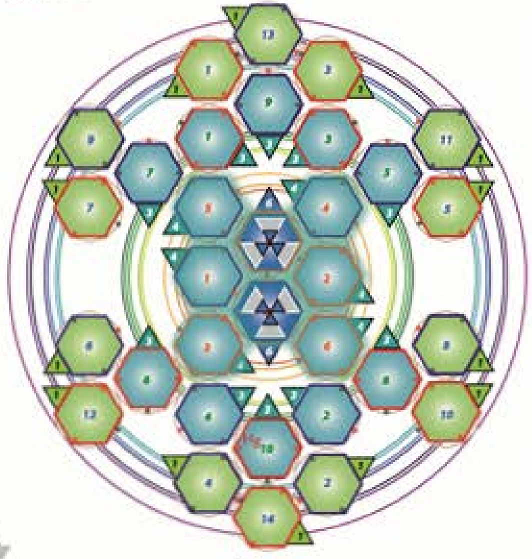

Mercury
[Xe] 4f14, 5d10, 6s2

Filled electron orbitals

Once 8 valence electrons are reached a stable valence cofiguration is created

Once sub-orbit 8 (P7-8) is reached sequential valence numbering switches
to paired valence numbering (as per orbital energies)

Valence numbers can be calculated by adding the 2 highest energy orbitals together
(and subtracting 8 - if the total is higher than 8)

The octet rule is a simple chemical rule of thumb that states that atoms tend to combine in such a way that they each have eight electrons in their valence shells, giving them the same electronic configuration as a noble gas.

The rule is applicable to the main-group elements, especially carbon, nitrogen, oxygen, and the halogens, but also to metals such as sodium or magnesium. In simple terms, molecules or ions tend to be most stable when the outermost electron shells of their constituent atoms contain eight electrons.

In short, an element's valence shell is full and most stable when it contains eight electrons, corresponding to an s2p6 election configuration.

CORE ELECTRONS

This stability is the reason that the noble gases are so unreactive, for example neon with electron configuration 1s2 2s2 2p6. (Helium is an exception as explained).

Note that a "full shell" means that there are the eight electrons in the valence shell when the next shell starts filling, even though higher subshells (d, f, etc.) have not been filled.

There can be at most eight valence electrons in a ground-state atom because p subshells are always followed by the s subshell of the next shell.

OCTET RULE

This means that once there are 8 valence electrons (when the p subshell is filled), the next additional electron goes into the next shell, which then becomes the valence shell.

A consequence of the octet rule is that atoms generally react by gaining, losing, or sharing electrons in order to achieve a complete octet of valence electrons. Reaction of atoms occurs primarily in two ways: ionically and covalently.

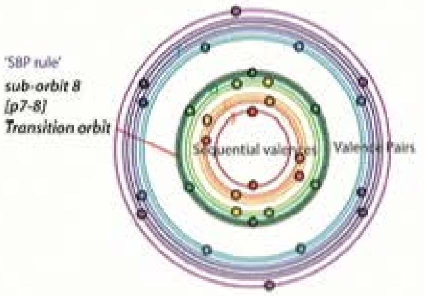
'S8P rule'
sub-orbit 8
[p7-8]
Transition orbit
Sequential valences
Valence Pairs

Once 8 valence electrons are reached the sub-orbits are stable and form a stable, non-reactive valence configuration.

Additionally, valence numbers proceed sequentially (1,2,3,4,5,6,7) up to sub-orbit 8 (P7-8) at which point all sub-orbit valences number in valence pairs (1-2, 3-4, 5-6, 7-8, 9-10, 11-2, 13-14)

The siginificance is that sub-orbit 8 is the middle sub-orbit (of 16 total sub-orbits possible)

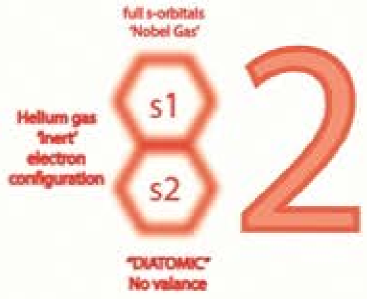
full s-orbitals
'Nobel Gas'
Helium gas
'Inert'
electron
configuration
s1
s2
2
"DIATOMIC"
No valance

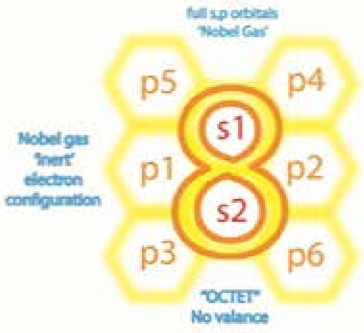
full s,p orbitals
'Nobel Gas'
Nobel gas
'Inert'
electron
configuration
p5
p4
s1
p1
8
p2
s2
p3
p6
"OCTET"
No valance

free
radical
atom
reactive
Hydrogen
bond
Hydrogen bond

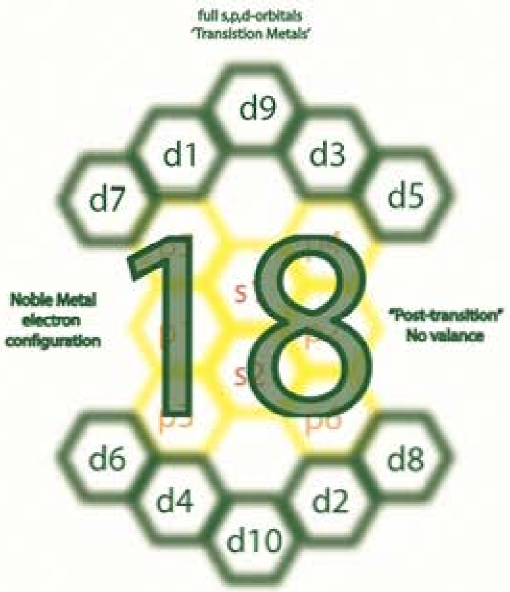
full s,p,d-orbitals
'Transistion Metals'
d9
d1
d3
d7
d5
Noble Metal
electron
configuration
18
"Post-transition"
No valance
d6
d8
d4
d2
d10

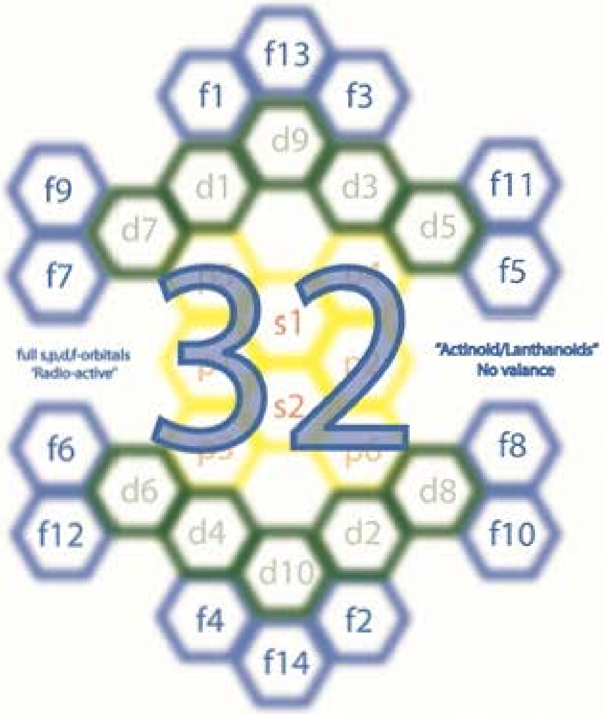
f13
f1
f3
d9
f9
d1
d3
f11
d7
d5
f7
f5
32
s1
s2
full s,p,d,f-orbitals
'Radio-active'
"Actinoid/Lanthanoids"
No valance
f6
f8
d6
d8
f12
d4
d2
f10
d10
f4
f2
f14

Tetryonics 56.08 - Filled electron orbitals

ACID

Hydrochloric acid is a clear, colourless solution of hydrogen chloride in water, it is a highly corrosive, strong mineral acid

Aqueous Hydro-Chloric acid solution

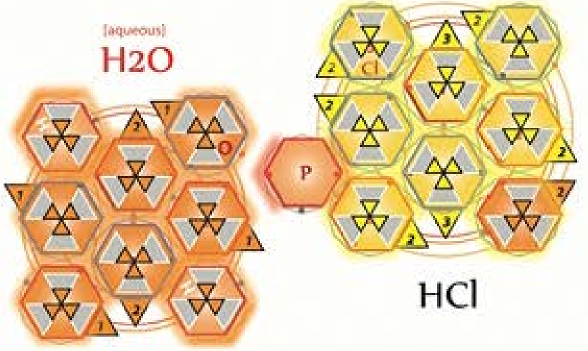
[aqueous]
H2O
P
Cl
HCl

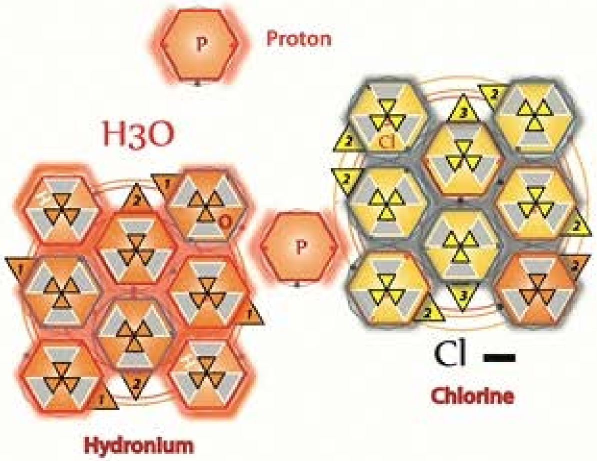
P
Proton
H3O
P
Cl
Cl −
Chlorine
Hydronium

The H(+) cation of the acid combines with the OH(-) anion of the base to form water.

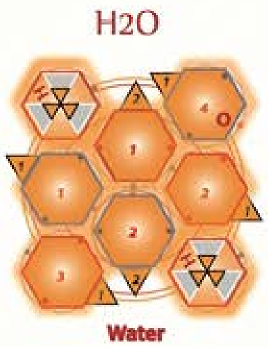

combining the free Protons & electrons

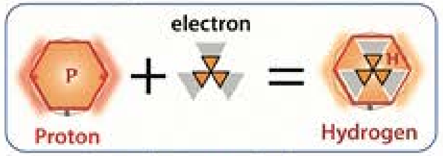
electron
P
+
=
H
Proton
Hydrogen

neutralises the ions and releases energy

+

all charges seek equilibrium

When an acid and a base are placed together, they react to neutralize the acid and base properties, producing a salt.

HCl(aq) + NaOH(aq) ----------> NaCl(aq) + H2O(l)

=>

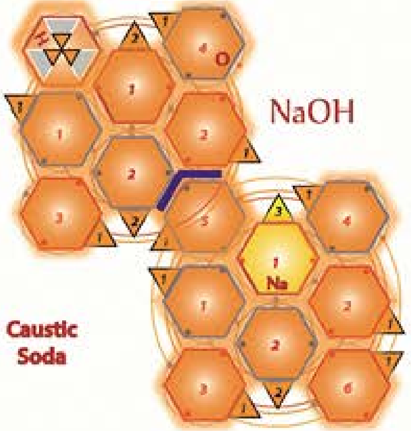
O
NaOH
Na
Caustic Soda

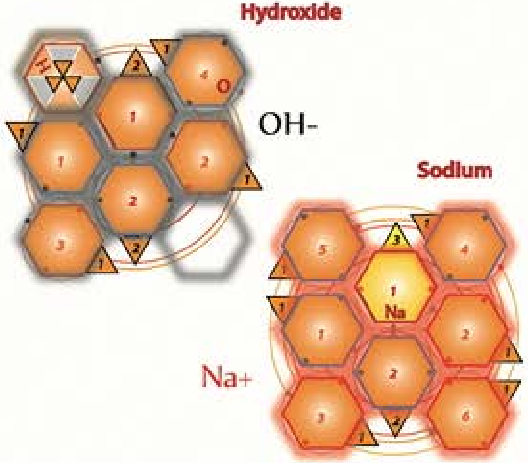
Hydroxide
O
OH-
Sodium
Na
Na+

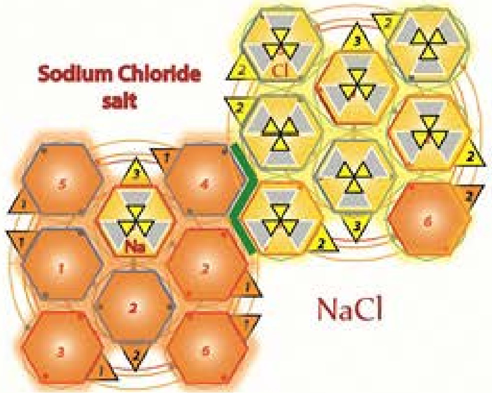
Sodium Chloride salt
Cl
Na
NaCl

The compound formed by the cation of the base and the anion of the acid is called a salt.

BASE

Sodium hydroxide, also known as caustic soda, or lye, has the molecular formula NaOH and is a highly caustic metallic base and alkali salt.

Geometric Molecular Topology

Molecules are most often held together with covalent bonds involving single, double, and/or triple bonds,
where a "bond" is a pair of electrons shared between elements as they seek equilibrium
(another method of bonding is ionic bonding and involves a positive cation and a negative anion).

Hydrogen

A Hydrogen bond is a chemical bond in which
a hydrogen atom of one molecule is attracted
to an electronegative atom, especially nitrogen,
oxygen, or flourine atoms, usually of another molecule

The hydrogen bond is often described as
an electrostatic dipole-dipole interaction.

However, it also has some features of covalent bonding:
it is directional and strong, acting over interatomic distances
shorter than that of van der Waals radii

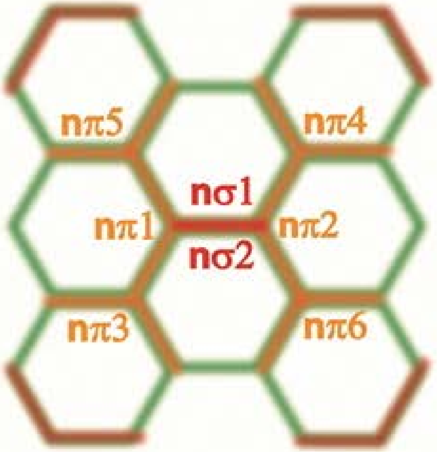

Oxygen

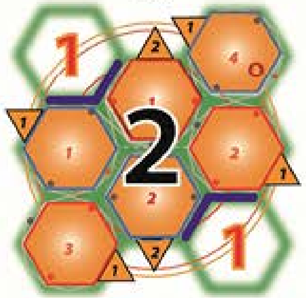

The oxidation state of oxygen is −2
in almost all known compounds

Oxides of Oxygen & oxygen molecules
are found throughout the range of
Organic & inorganic compounds

**All molecules seek
KEM'ical charge and
energy equilibrium**

*Geometric Molecular Topology is the overall arrangement of the atoms in a molecule,
where the bonded atoms in a molecule are responsible for determining
the final molecular topology of a chemical system of bonded elements.*

*As the numbers of atoms in molecules increases, the quantum molecular topology
of a system grows increasingly complex and can only be modelled accurately
using Tetryonic charged geometries*

**by bonding together
and forming larger
complex molecules**

Common nitrogen functional groups include:
amines, amides, nitro groups, imines, and enamines.

Compounds of Carbon form the 'backbones'
of Organic & inorganic compounds

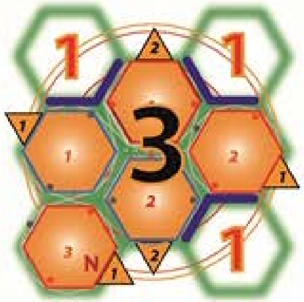

Nitrogen

Molecular Octets

The concept of the Expanded Octet occurs in any system that has an atom with more than four electron pairs attached to it.

Most commonly, atoms will expand their octets to contain a total of five or six electron pairs,
in total. In theory, it is possible to expand beyond those number.

The large amounts of negative charge concentrated in small volumes of space
prevent those larger expanded octets from forming.

When an atom expands its octet, it does so by making use of empty d-orbitals
that are available in the valence level of the atom doing the expanding.

The atom that expands its octet in a structure will usually be located in the center of the structure
and the system will not use any multiple bonds in attaching atoms to the central atom.

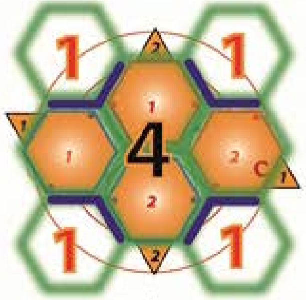

Carbon

Copyright ABRAHAM [2008] - All rights reserved

48
[24-24]

Hydrogen bonds

H

96
[48-48]

1 Proton
1 electron

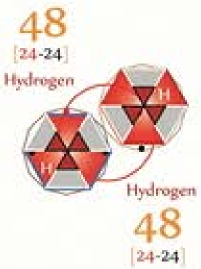
48
[24-24]
Hydrogen
H
H
Hydrogen
48
[24-24]

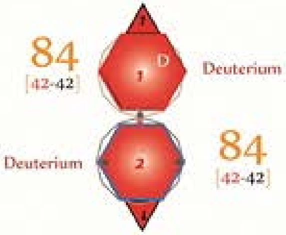
84
[42-42]
D
1
Deuterium
Deuterium
2
84
[42-42]

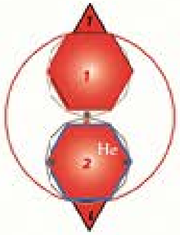

168
[84-84]

2 Protons
2 Neutrons
2 electrons

Hydrogen is a 'free radical' atom

H2 has a smaller s-orbital [lower energy level] compared to Helium

Deuterium is the building block of elements

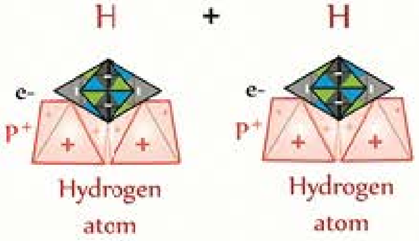
H
+
H
e-
P+
+
+
Hydrogen
atom
e-
P+
+
+
Hydrogen
atom

Hydrogen atoms → form Covalent bonds

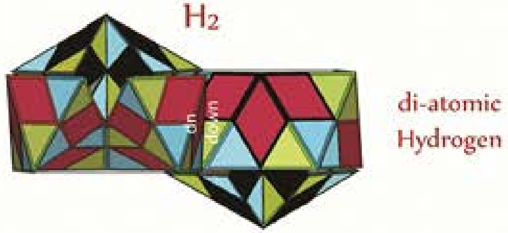
H2
up
down
di-atomic
Hydrogen

Hydrogen, bound mostly to carbon and nitrogen, is part of almost every molecule in your body: DNA, proteins, sugars, fats.
Hydrogen bonds - which form between atoms that "share" a hydrogen atom, is one of the most important interactions that makes biological molecules behave as they do.

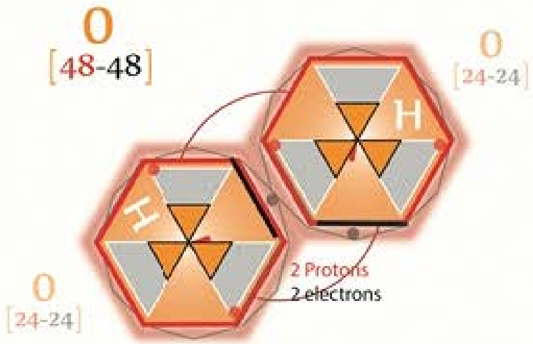
0
[48-48]
0
[24-24]
H
H
0
[24-24]
2 Protons
2 electrons

Hydrogen bonds are sigma bonds

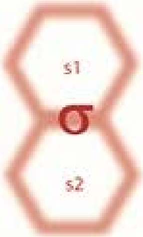

note: n2 energy level Hydrogen is illustrative only
[Hydrogen bonds can be of any energy level]

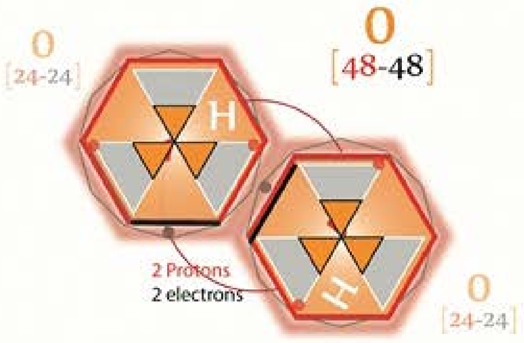
0
[24-24]
0
[48-48]
H
H
2 Protons
2 electrons
0
[24-24]

The charged fascia of Baryonic quarks facilitates chemical [hydrogen] bonding in molecules

Carbon Compounds

504
[252-252]

Carbon can bond in four places, and it can bond to itself so it's easy to make lots of different molecules.

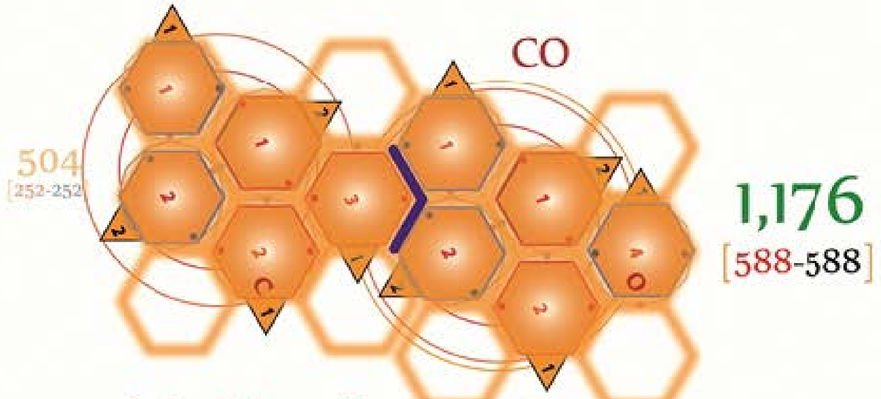
CO
504
[252-252]
1,176
[588-588]
C
O
672
[336-336]

Carbon Monoxide Molecule

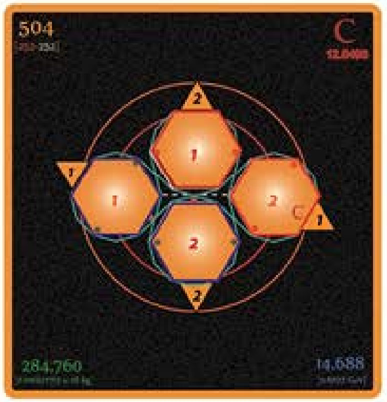
504
C
284,760
14,688

[He]$_2$ 2s2 2p2

Many carbon compounds are essential for life as we know it.
Some of the most common carbon compounds are: carbon dioxide (CO2), carbon monoxide (CO), carbon disulfide (CS2), chloroform (CHCl3), carbon tetrachloride (CCl4), methane (CH4), ethylene (C2H4), acetylene (C2H2), benzene (C6H6), ethyl alcohol (C2H5OH) and acetic acid (CH3COOH) to name just a few.

Acetylene Molecule

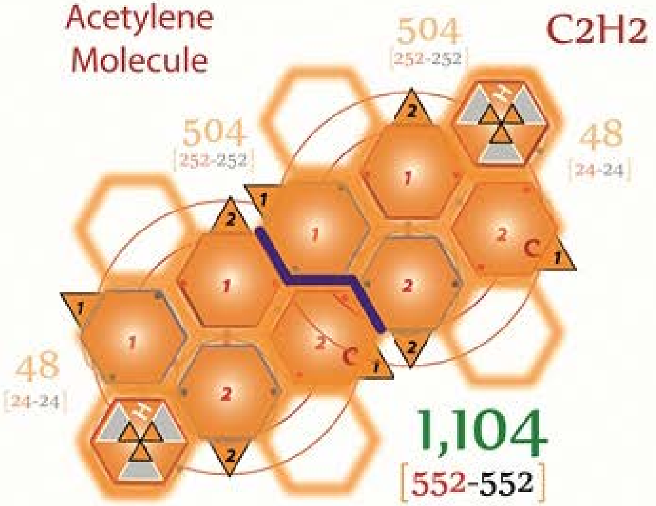
504
[252-252]
C2H2
504
[252-252]
48
[24-24]
48
[24-24]
C
C
H
H
1,104
[552-552]

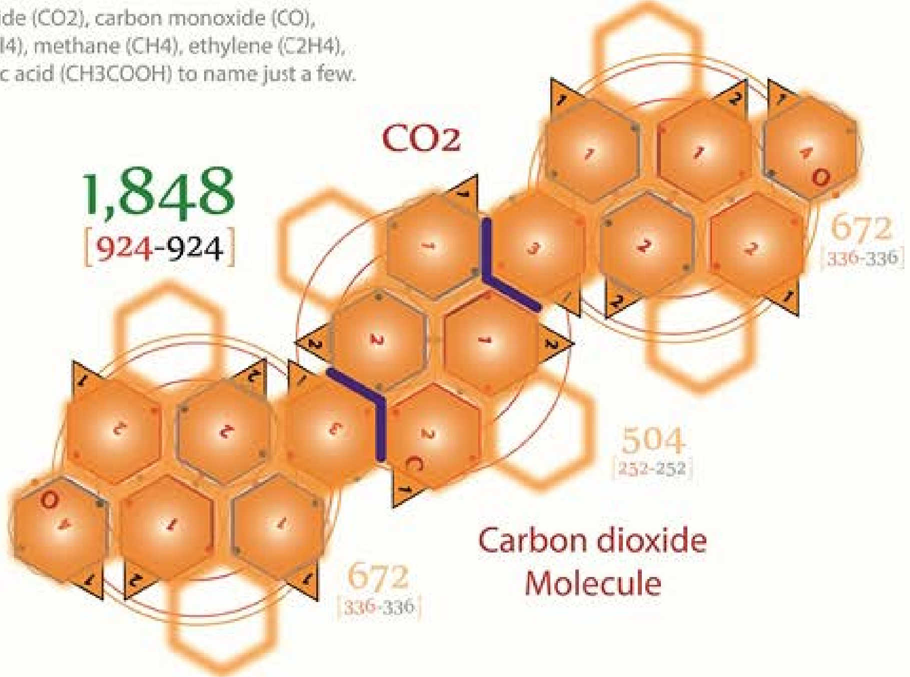
CO2
1,848
[924-924]
672
[336-336]
504
[252-252]
672
[336-336]
O
O
C

Carbon dioxide Molecule

C

There are nearly ten million known carbon compounds and an entire branch of chemistry, known as organic chemistry, is devoted to their study.

Nitrogen Compounds

588
[294-294]

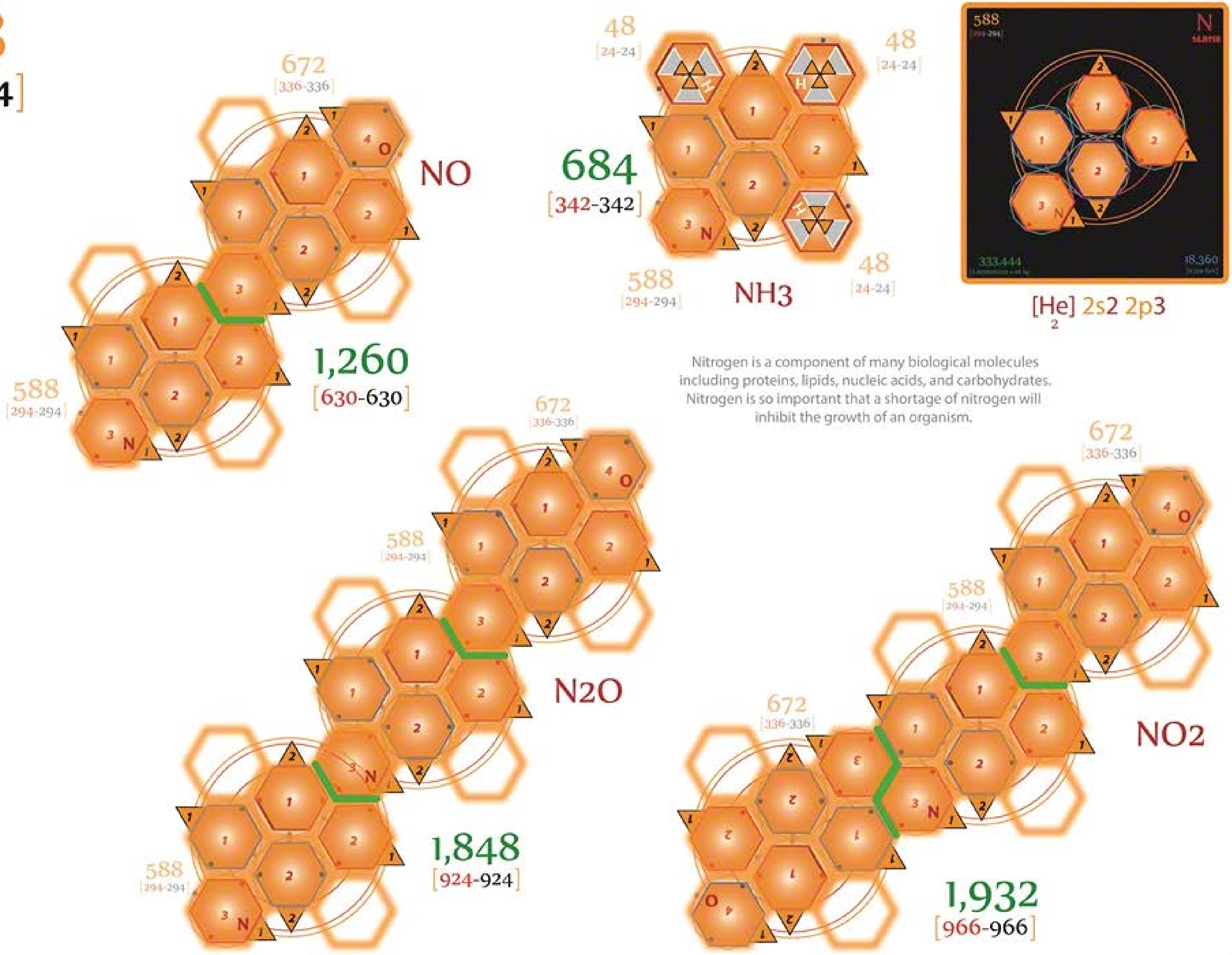
672
[336-336]
NO
588
[294-294]
1,260
[630-630]
48
[24-24]
48
[24-24]
684
[342-342]
588
[294-294]
NH3
48
[24-24]
588
[294-294]
N
333.444
18,360
[He] 2s2 2p3
672
[336-336]
588
[294-294]
N2O
588
[294-294]
1,848
[924-924]
672
[336-336]
588
[294-294]
672
[336-336]
NO2
1,932
[966-966]

Nitrogen is a component of many biological molecules including proteins, lipids, nucleic acids, and carbohydrates. Nitrogen is so important that a shortage of nitrogen will inhibit the growth of an organism.

N

Copyright ABRAHAM [2008] - All rights reserved

672

[336-336]

Oxygen Compounds

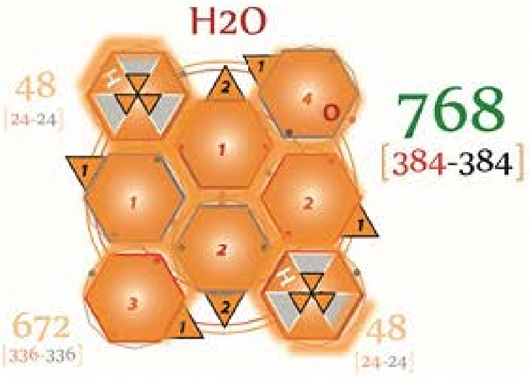
H2O
48
[24-24]
768
[384-384]
672
[336-336]
48
[24-24]

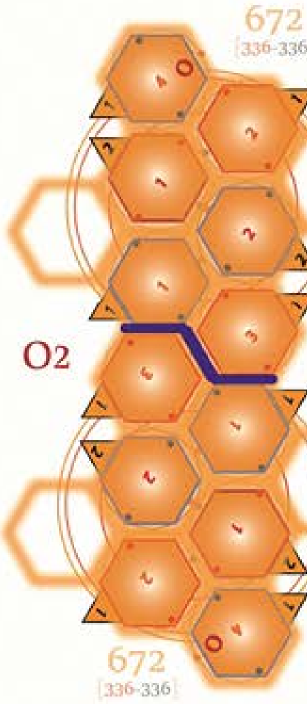
672
[336-336]
O2
1,504
[252-252]
672
[336-336]

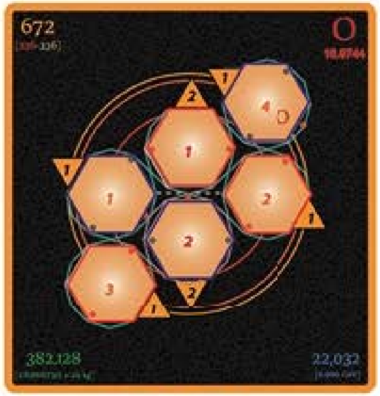
672
O
382,128
22,032

[He]$_2$ 2s2 2p4

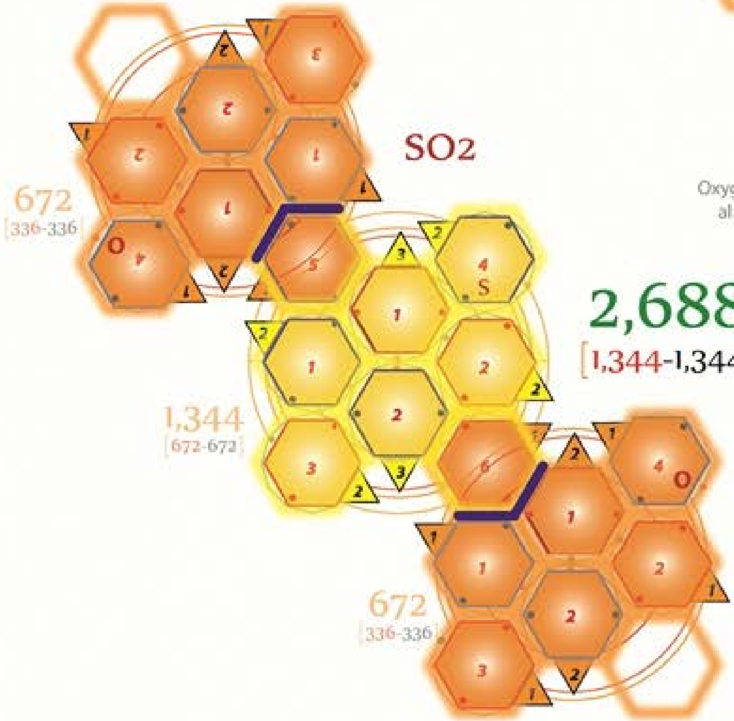
SO2
672
[336-336]
2,688
[1,344-1,344]
1,344
[672-672]
672
[336-336]

Oxygen forms oxide compounds with
almost all of the known elements

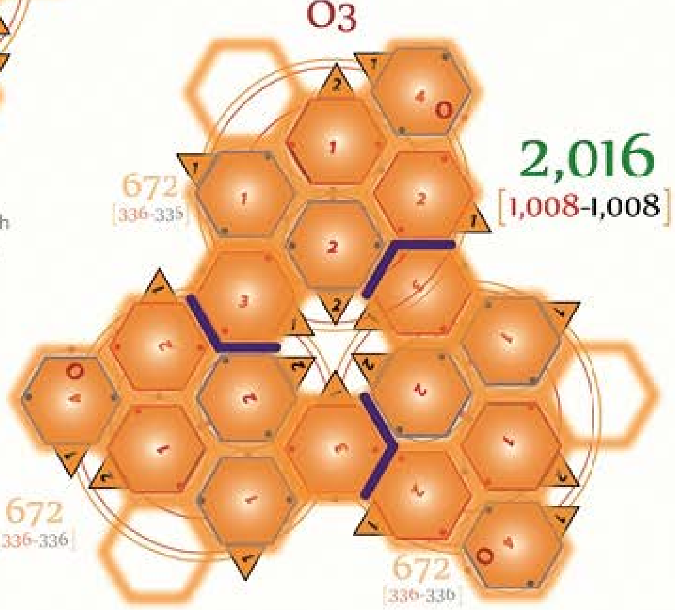
O3
672
[336-335]
2,016
[1,008-1,008]
672
[336-336]
672
[336-336]

As soon as the oxygen enters your blood, a passing protein molecule called hemoglobin picks it up. Each molecule of hemoglobin can transport four molecules of oxygen to almost anywhere in the body.

The hemoglobin transport the oxygen to your cells where another protein, called cytochrome C oxidase makes two molecules of water out of every molecules of oxygen delivered to it

O

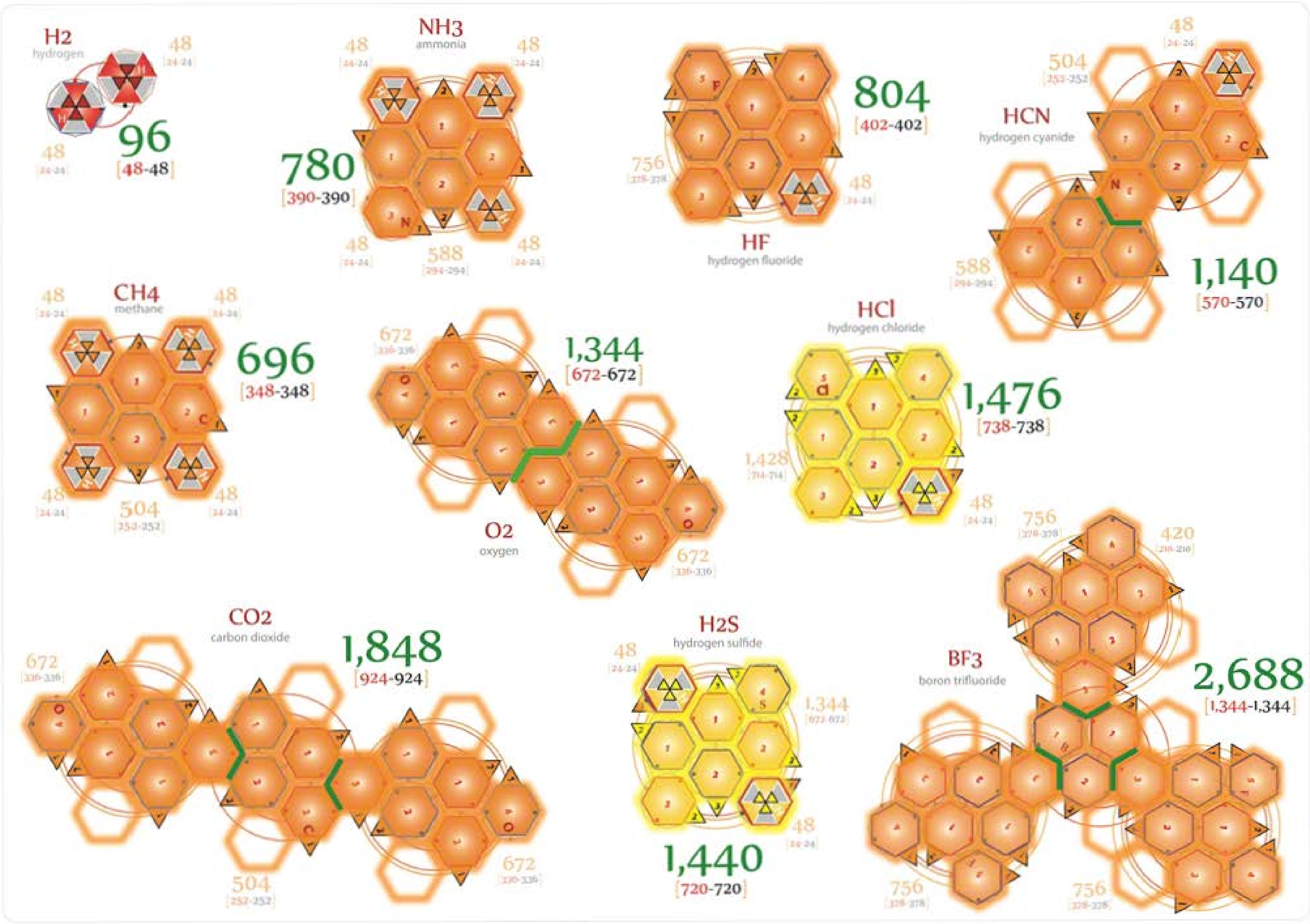
H2
hydrogen
48
[24-24]
48
[24-24]
96
[48-48]
NH3
ammonia
48
[24-24]
48
[24-24]
780
[390-390]
48
[24-24]
588
[294-294]
48
[24-24]
804
[402-402]
756
[378-378]
48
[24-24]
HF
hydrogen fluoride
504
[252-252]
48
[24-24]
HCN
hydrogen cyanide
588
[294-294]
1,140
[570-570]
CH4
methane
48
[24-24]
48
[24-24]
696
[348-348]
48
[24-24]
504
[252-252]
48
[24-24]
672
[336-336]
1,344
[672-672]
O2
oxygen
672
[336-336]
HCl
hydrogen chloride
1,476
[738-738]
1,428
[714-714]
48
[24-24]
756
[378-378]
420
[210-210]
CO2
carbon dioxide
672
[336-336]
1,848
[924-924]
504
[252-252]
672
[336-336]
H2S
hydrogen sulfide
48
[24-24]
1,344
[672-672]
48
[24-24]
1,440
[720-720]
BF3
boron trifluoride
2,688
[1,344-1,344]
756
[378-378]
756
[378-378]

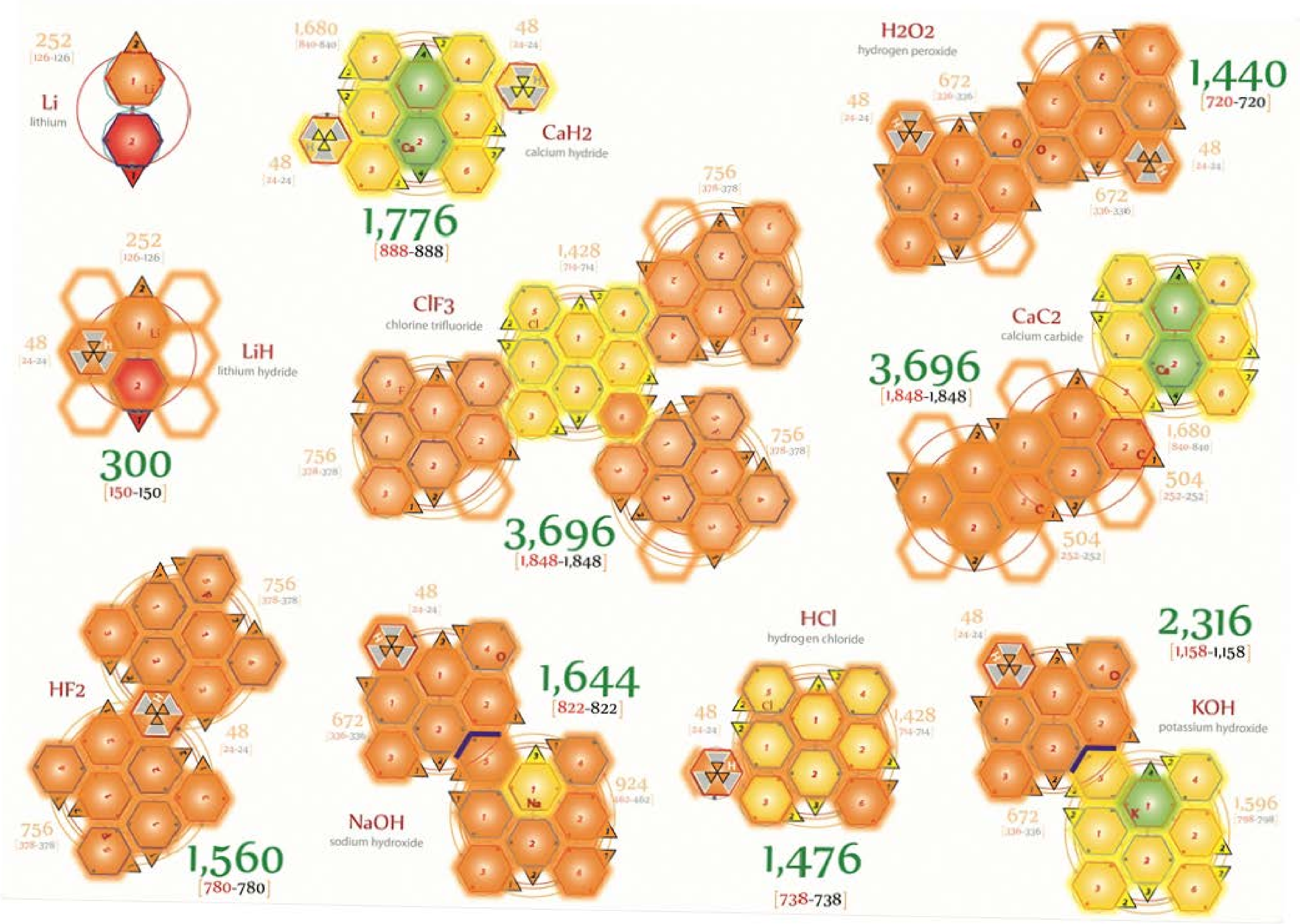
252
[126-126]
Li
lithium
1,680
[840-840]
48
[24-24]
CaH2
calcium hydride
48
[24-24]
1,776
[888-888]
H2O2
hydrogen peroxide
48
[24-24]
672
[336-336]
1,440
[720-720]
48
[24-24]
672
[336-336]
252
[126-126]
48
[24-24]
LiH
lithium hydride
300
[150-150]
756
[378-378]
1,428
[714-714]
ClF3
chlorine trifluoride
756
[378-378]
756
[378-378]
3,696
[1,848-1,848]
CaC2
calcium carbide
3,696
[1,848-1,848]
1,680
[840-840]
504
[252-252]
504
[252-252]
756
[378-378]
HF2
48
[24-24]
756
[378-378]
1,560
[780-780]
48
[24-24]
672
[336-336]
1,644
[822-822]
924
[462-462]
NaOH
sodium hydroxide
HCl
hydrogen chloride
48
[24-24]
1,428
[714-714]
1,476
[738-738]
48
[24-24]
2,316
[1,158-1,158]
KOH
potassium hydroxide
672
[336-336]
1,596
[798-798]

Copyright ABRAHAM [2008] - All rights reserved

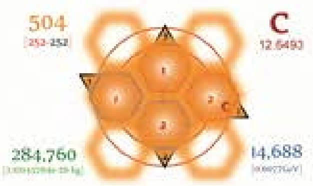
504
[252-252]
C
12.6493
284,760
14,688

Chemistry

is a branch of physical science, concerning the study of the composition, properties and behavior of Matter

Compounds containing bonds between carbon and a metal are called organometallic compounds.

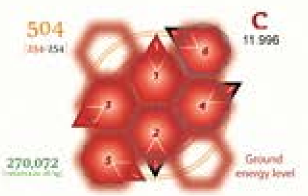
504
[254-254]
C
11.996
270,072
Ground energy level

Inorganic compounds are produced by non-living natural processes or in the laboratory.

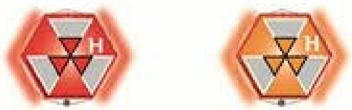

Inorganic compounds can form salts.

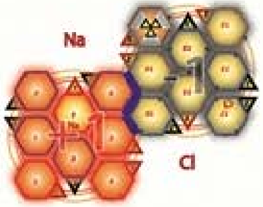

Inorganic compounds contain metal atoms.

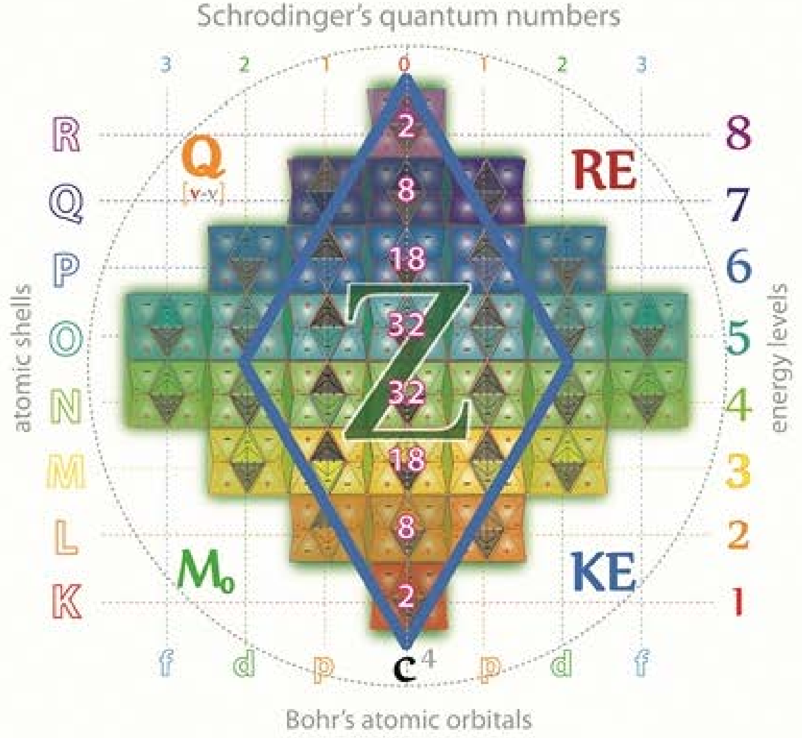
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
atomic shells
Q
[v-v]
RE
2
8
18
32
32
18
8
2
Z
M_0
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
C^4
p
d
f
Bohr's atomic orbitals

Organic compounds are produced by living things.

Organic compounds can't form salts.

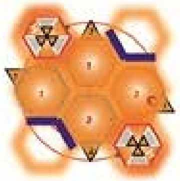

Organic compounds contain carbon-hydrogen bonds.

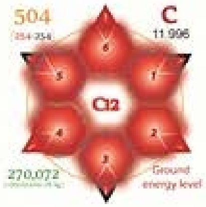
504
[254-254]
C
11.996
Cl2
270,072
Ground energy level

Tetryonic theory unifies and expands upon the currently disjointed physical and chemical theories through the application of 2D equilateral charged mass-energy geometries in 3D standing-wave mass-Matter topologies

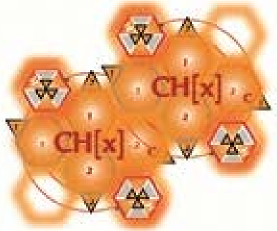

Inorganic **Typically the difference is defined as being whether or not a substance contains carbon or carbon-hydrogen bonds** **Organic**

A chemical compound is a collection of elements bonded together in a way that the resultant ions, atoms or molecules form a 3D material geometric structure. Tetryonic chemical geometries, along with its firm definition and distinction between EM mass & Matter provide a clear visual path for the differentiation between both branckes of modern chemistry - as well as the source of animation in living Matter

Tetryonics 57.08 - Compound molecular geometries

Functional Groups

A functional group is a reactive portion of a molecule.

carbonyl group

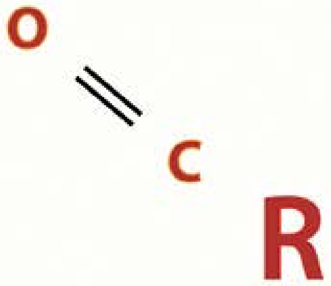

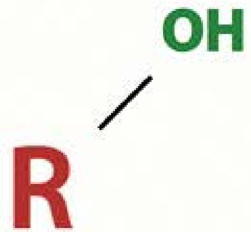

hydroxyl group

Organic molecules containing a hydroxyl group are known as alcohols.

carboxyl group

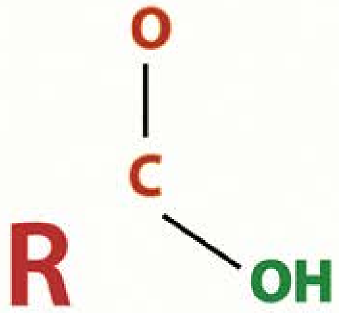

The combinations of functional groups with hydrocarbons produce a vast number of compounds.

Carbon monoxide

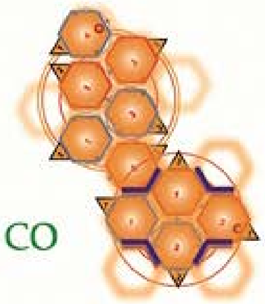

carbonyl group is composed of a carbon atom double-bonded to an oxygen atom

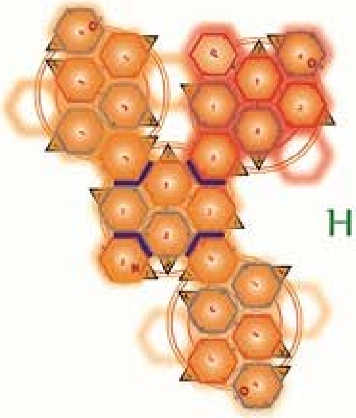

Nitric acid

Carbonic acid

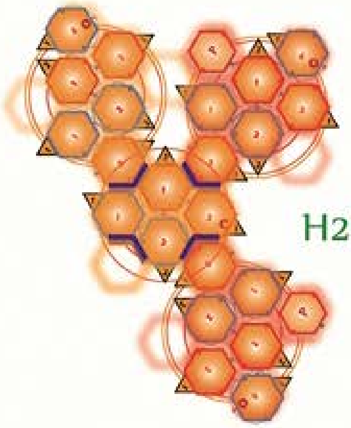

The carboxyl group is present in amino acids and carboxylic acids.

Derivatives of Hydrocarbons

An almost unlimited number of carbon compounds can be formed by the addition of a functional group to a hydrocarbon

Esters

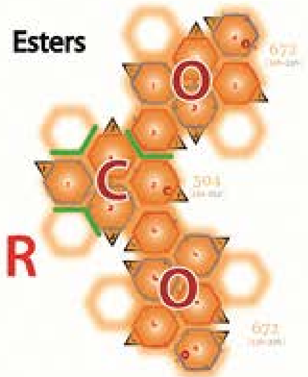

Most esters have pleasant odors.
Esters are responsible for the fragrances of
many flowers & the tastes of ripened fruits.

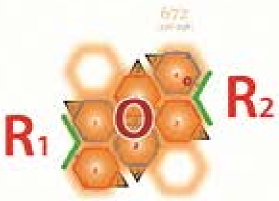

Ethers

The best known ether is diethyl ether.
It is a volatile, highly flammable liquid
that was used as an anesthetic in the past.

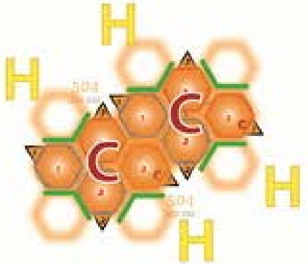

Alkyl Halides - haloalkanes

Common alkyl halides include medical anesthetics,
chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs)
and hydrofluorocarbons (HFCs).

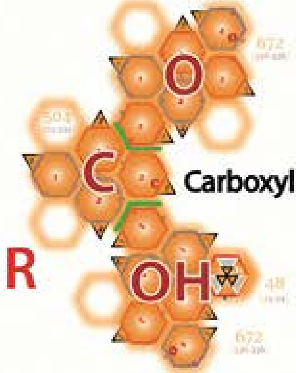

Carboxylic Acids

The simplest of the carboxylic acids is formic acid
and is a constituent of bee stings and the bites of
other insects including mosquitos.

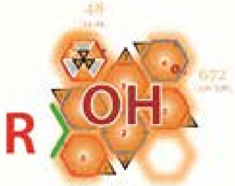

Alcohols

Alcohols are organic compounds containing a
hydroxyl group, [OH], substituted for a hydrogen atom.
Ethanol is the alcohol in alcoholic beverages and
it is also widely used as a solvent.

H
hydro

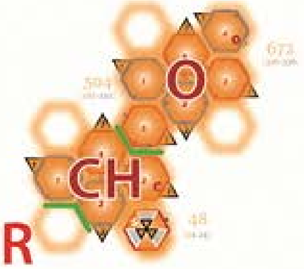

Aldehydes

An aldehyde is a compound containing a
carbonyl group with at least one hydrogen attached to it.
With a Hydrogen in place of the R group it forms Formaldehyde

C
carbons

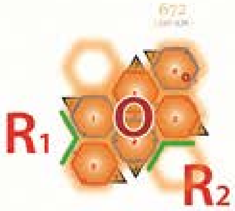

Ketones

Acetone is the simplest of the ketones.
Acetone is a commonly used solvent and is
the active ingredient in nail polish remover
and some paint thinners.

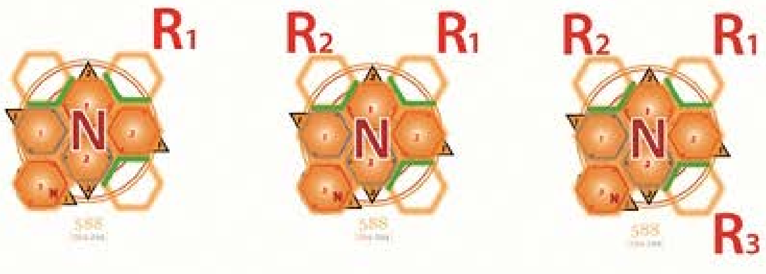

Amines

Amines are organic compounds that contain nitrogen,
they are basic compounds with strong odors,
often described as "fishy"

"R's" stand for carbon substituents or hydrogen atoms.

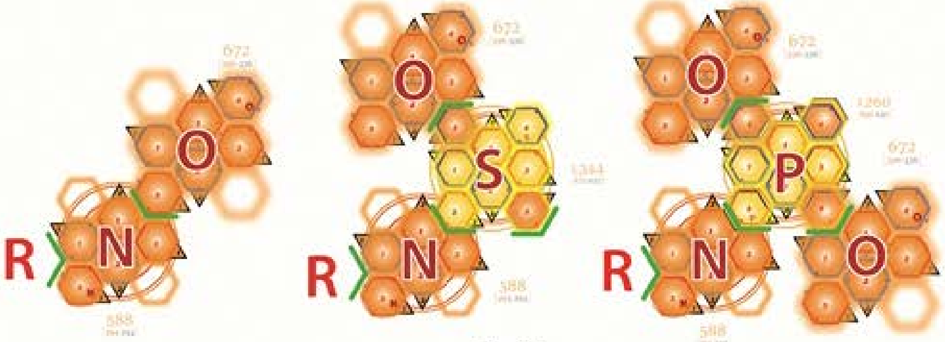

Amides

Amides are nitrogen-containing organic compounds
and are formed when amino acids react to form proteins.

Copyright ABRAHAM [2008] - All rights reserved

Hydrocarbons

In organic chemistry, a methylene group is any part of a molecule that consists of two hydrogen atoms bound to a carbon atom, which is connected to the remainder of the molecule

600
[300-300]

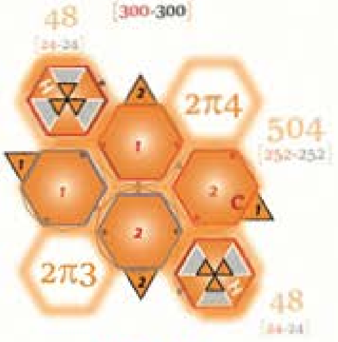
48
[24-24]
2π4
504
[252-252]
2π3
48
[24-24]

CH2
Methylene

The methylene group should be distinguished from the free methylene radical, also called carbene, whose molecule is a methylene group all by itself.

H2C
Carbene

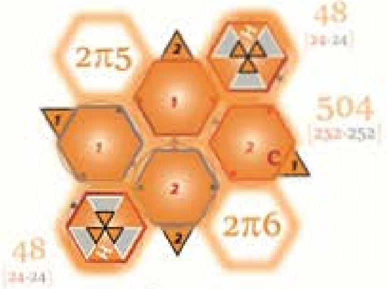
2π5
48
[24-24]
504
[252-252]
2π6
48
[24-24]

600
[300-300]

CH4
Methane

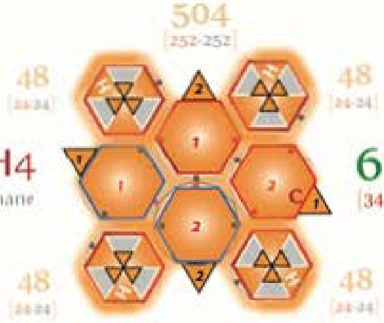
504
[252-252]
48
[24-24]
48
[24-24]
48
[24-24]
48
[24-24]

696
[348-348]

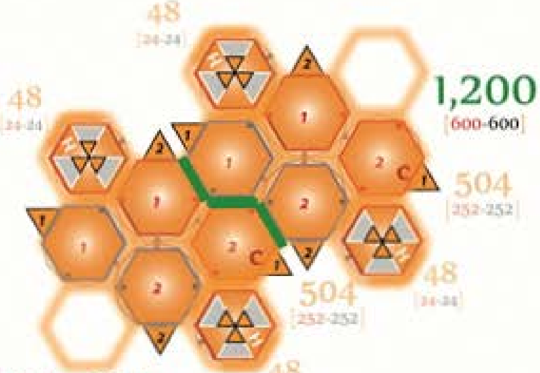
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
48
[24-24]

1,200
[600-600]

CH2-CH2
Ethylene

CH3-CH3
Ethane
C2H6

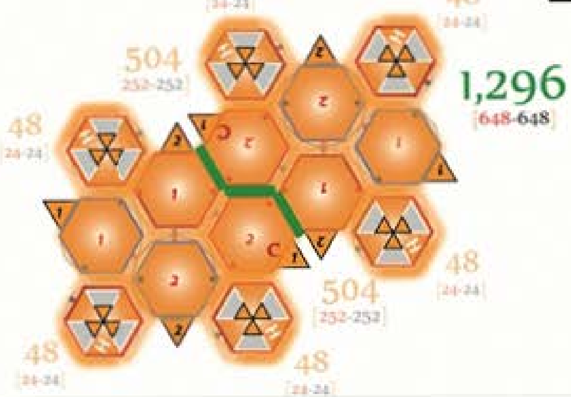
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]

1,296
[648-648]

2

2,304
[1,152-1,152]

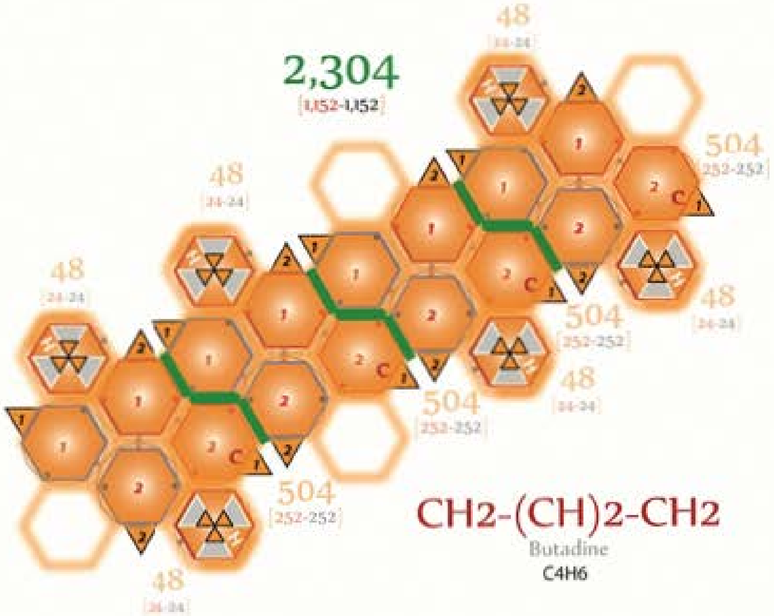
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]
504
[252-252]
504
[252-252]
48
[24-24]

CH2-(CH)2-CH2
Butadine
C4H6

CH3-(CH2)-CH3
Propane
C3H8

3

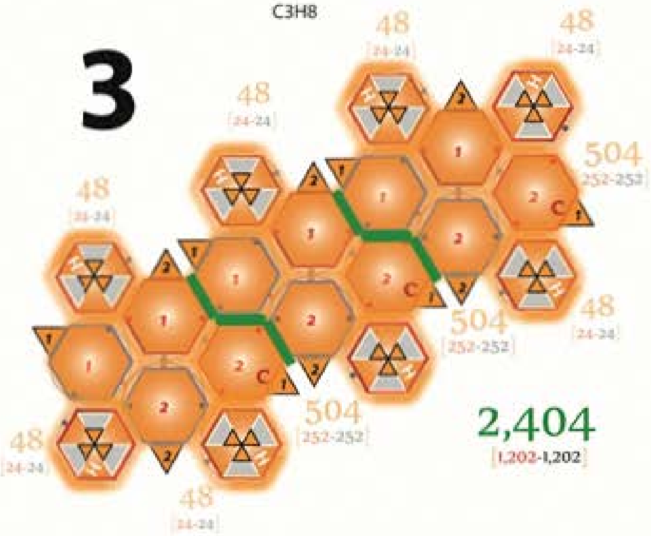
48
[24-24]
48
[24-24]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]

2,404
[1,202-1,202]

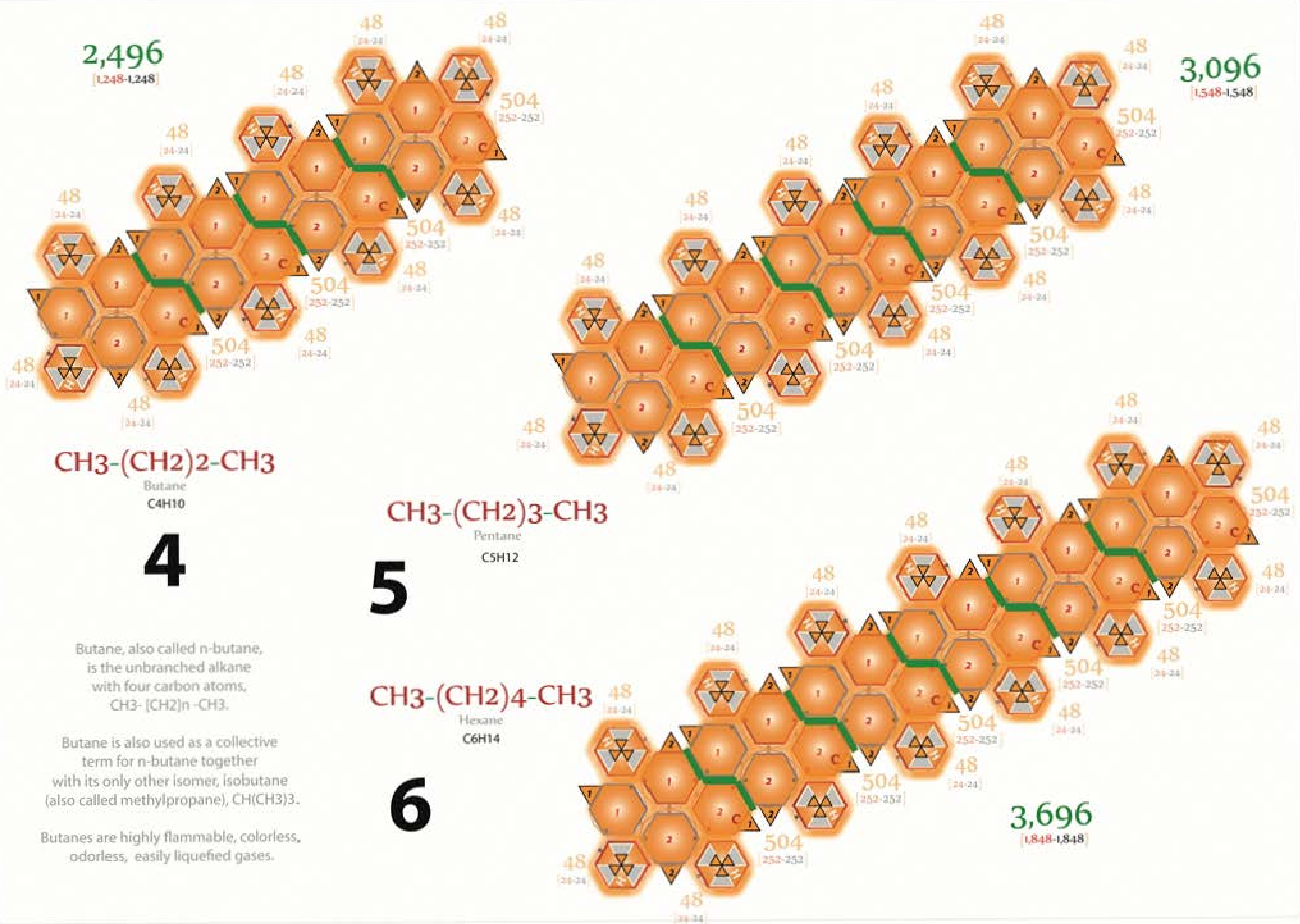
2,496
[1,248-1,248]
3,096
[1,548-1,548]
3,696
[1,848-1,848]
48
[24-24]
504
[252-252]
1
2
C

CH3-(CH2)2-CH3

Butane

C4H10

4

Butane, also called n-butane,
is the unbranched alkane
with four carbon atoms,
CH3- [CH2]n -CH3.

Butane is also used as a collective
term for n-butane together
with its only other isomer, isobutane
(also called methylpropane), CH(CH3)3.

Butanes are highly flammable, colorless,
odorless, easily liquefied gases.

CH3-(CH2)3-CH3

Pentane

C5H12

5

CH3-(CH2)4-CH3

Hexane

C6H14

6

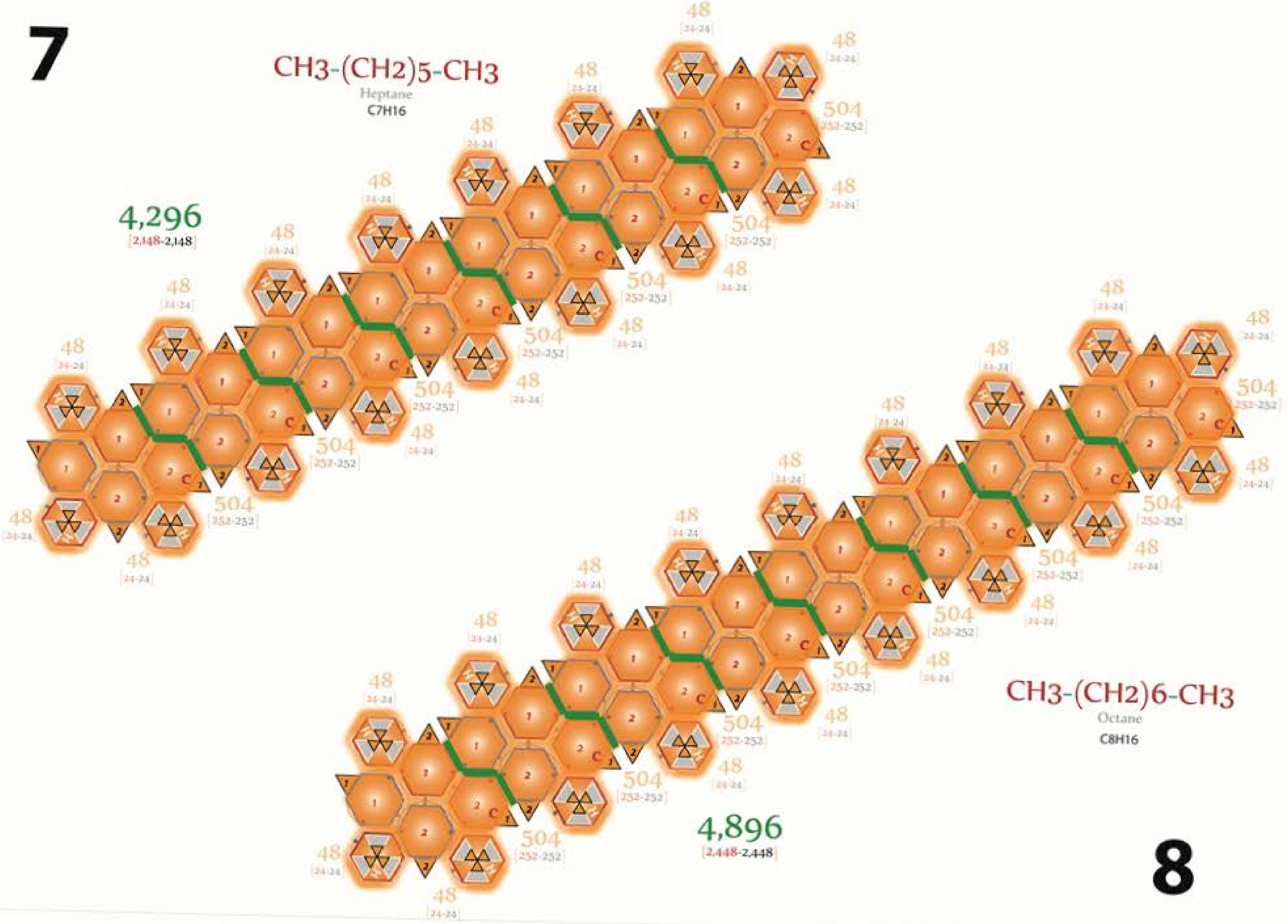
48
[24-24]
504
[252-252]
1
2
C

7

CH3-(CH2)5-CH3

Heptane

C7H16

4,296

[2,148-2,148]

CH3-(CH2)6-CH3

Octane

C8H16

4,896

[2,448-2,448]

8

Just like periodic elements, compounds can form allotropic compound structures

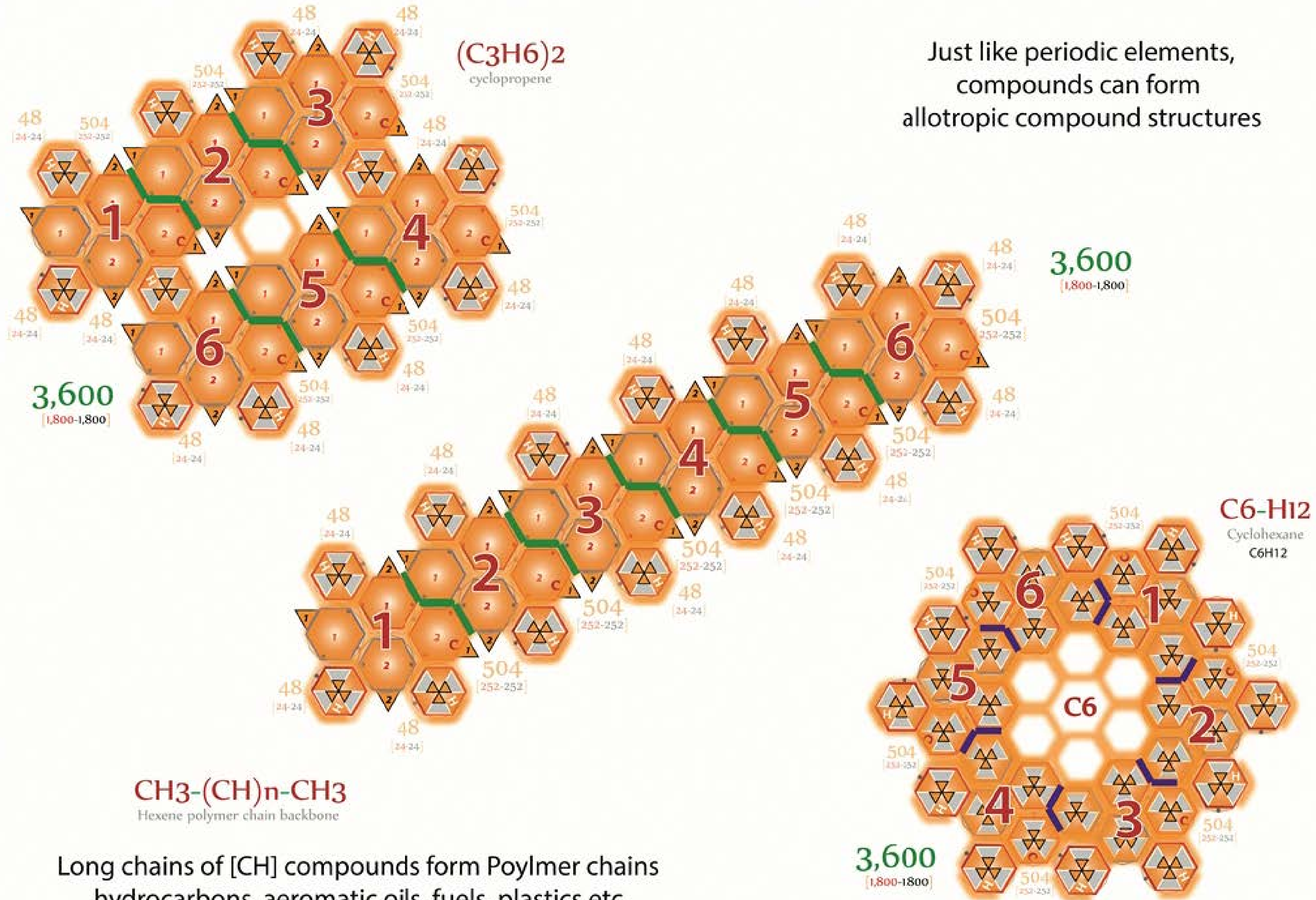
(C3H6)2
cyclopropene
48
[24-24]
504
[252-252]
1
2
3
4
5
6
C
3,600
[1,800-1,800]
C6-H12
Cyclohexane
C6H12
C6
[1,800-1800]
H
CH3-(CH)n-CH3
Hexene polymer chain backbone

Long chains of [CH] compounds form Poylmer chains hydrocarbons, aeromatic oils, fuels, plastics etc.

Copyright ABRAHAM [2008] - All rights reserved

Rings of atoms are also common in organic structures. You may have heard the famous story of Auguste Kekulé first realizing that benzene has a ring structure when he dreamt of snakes biting their own tails.

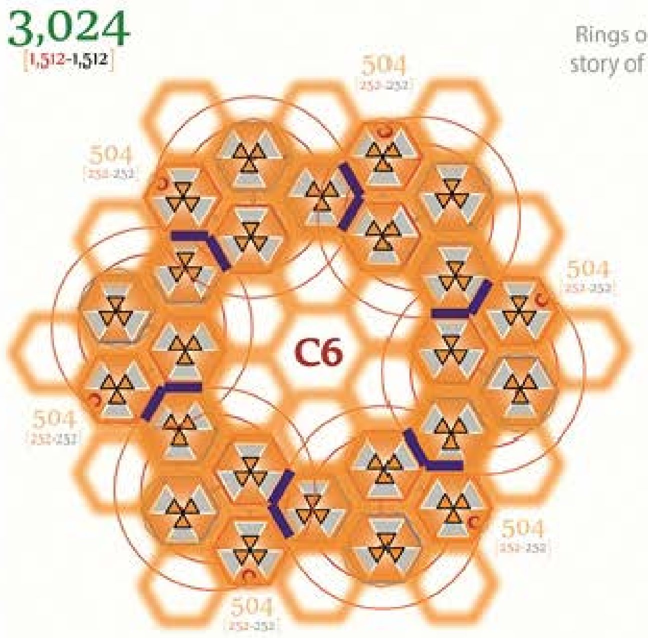
3,024
[1,512-1,512]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
C6

Carbon ring

Friedrich August Kekul

(7 September 1829 - 13 July 1896)

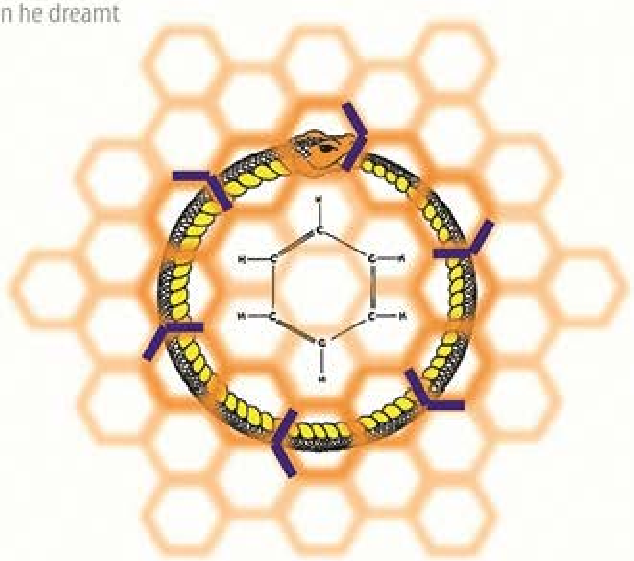

Organic Chemistry

In 1865, August Kekulé presented a paper at the Académie des Sciences in Paris suggesting a cyclic structure for benzene, the inspiration for which he ascribed to a dream. However, was Kekulé the first to suggest that benzene was cyclic. Some credit an Austrian schoolteacher, Josef Loschmidt with the first depiction of cyclic benzene structures.

In 1861, 4 years before Kekulé's dream, Loschmidt published a book in which he represented benzene as a set of rings. It is not certain whether Loschmidt or Kekulé—or even a Scot named Archibald Couper—got it right first

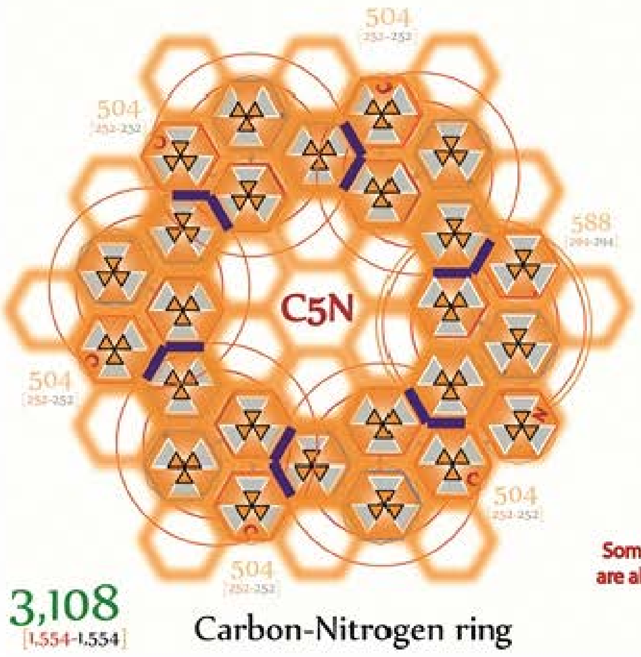
504
[252-252]
504
[252-252]
588
[294-294]
C5N
504
[252-252]
504
[252-252]
504
[252-252]
3,108
[1,554-1,554]

Carbon-Nitrogen ring

Some non-benzene-based compounds called heteroarenes, which follow Hückel's rule, are also aromatic compounds. In these compounds, at least one carbon atom is replaced by one of the heteroatoms oxygen, nitrogen, or sulfur.

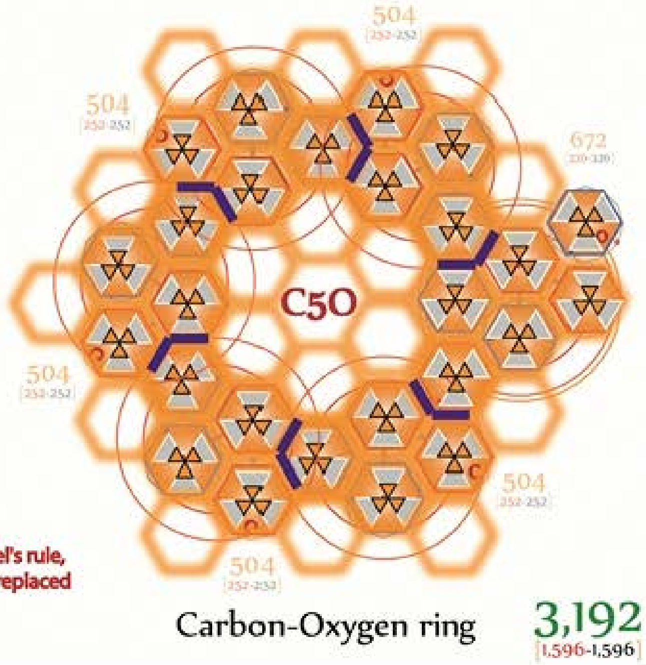
504
[252-252]
504
[252-252]
672
[336-336]
C5O
504
[252-252]
504
[252-252]
504
[252-252]
3,192
[1,596-1,596]

Carbon-Oxygen ring

Copyright ABRAHAM [2008] - All rights reserved

Benzene molecules

Benzene, or benzol, is an organic chemical compound and a known carcinogen with the molecular formula C6H6.

It is sometimes abbreviated Ph–H.
Benzene is a colorless and highly flammable liquid with a sweet smell and a relatively high melting point.

Because it is a known carcinogen, its use as an additive in gasoline is now limited, but it is an important industrial solvent and precursor in the production of drugs, plastics, synthetic rubber, and dyes. Benzene is a natural constituent of crude oil, and may be synthesized from other compounds present in petroleum.

Benzene is an aromatic hydrocarbon and the second [n]-annulene ([6]-annulene), a cyclic hydrocarbon with a continuous pi bond.

Cyclic hydrocarbon compounds are often referred to as Aromatic compounds due to their sweet smell

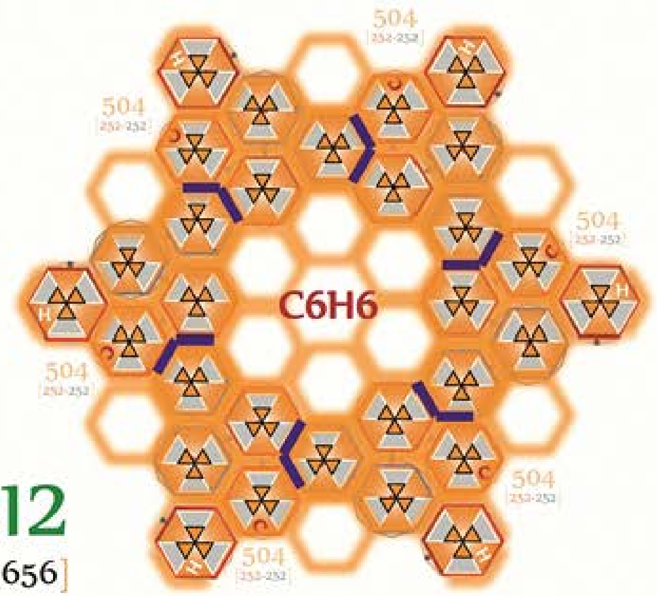
504
252-252
504
252-252
504
252-252
C6H6
504
252-252
504
252-252
504
252-252
H
H
C

3,312
[1,656-1,656]

(CH)6
Benzene ring
C6H6

A aromatic hydrocarbon is formed when CH compounds form a cyclic molecule

Many important additional chemical compounds are derived from benzene by replacing one or more of its hydrogen atoms with another functional group.

Examples of simple benzene derivatives are phenol, toluene, and aniline, abbreviated PhOH, PhMe, and PhNH2, respectively

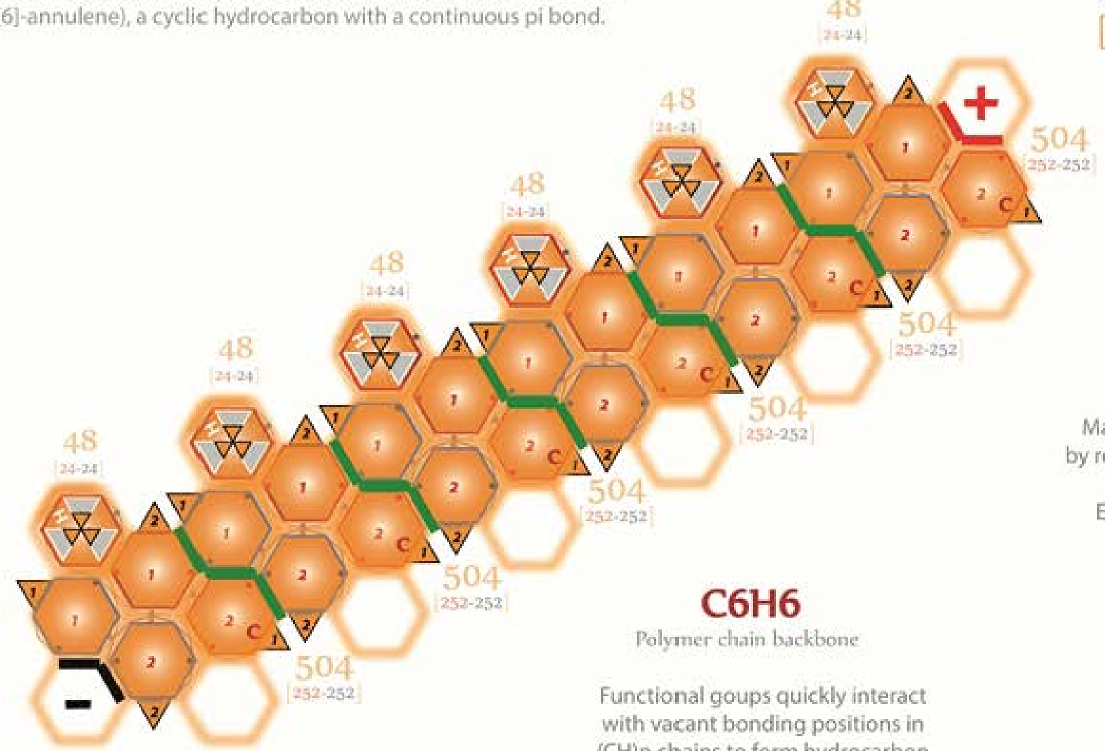
48
24-24
48
24-24
48
24-24
48
24-24
48
24-24
48
24-24
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
504
252-252
+
–
1
2
C

C6H6
Polymer chain backbone

Functional goups quickly interact with vacant bonding positions in (CH)n chains to form hydrocarbon compounds

Linear (CH)n chains are rarely found in nature as the Positive and Negative tail ends of Hydrocarbon chain interact and bond to form cyclic compounds.

Carbon-Nitrogen Chains

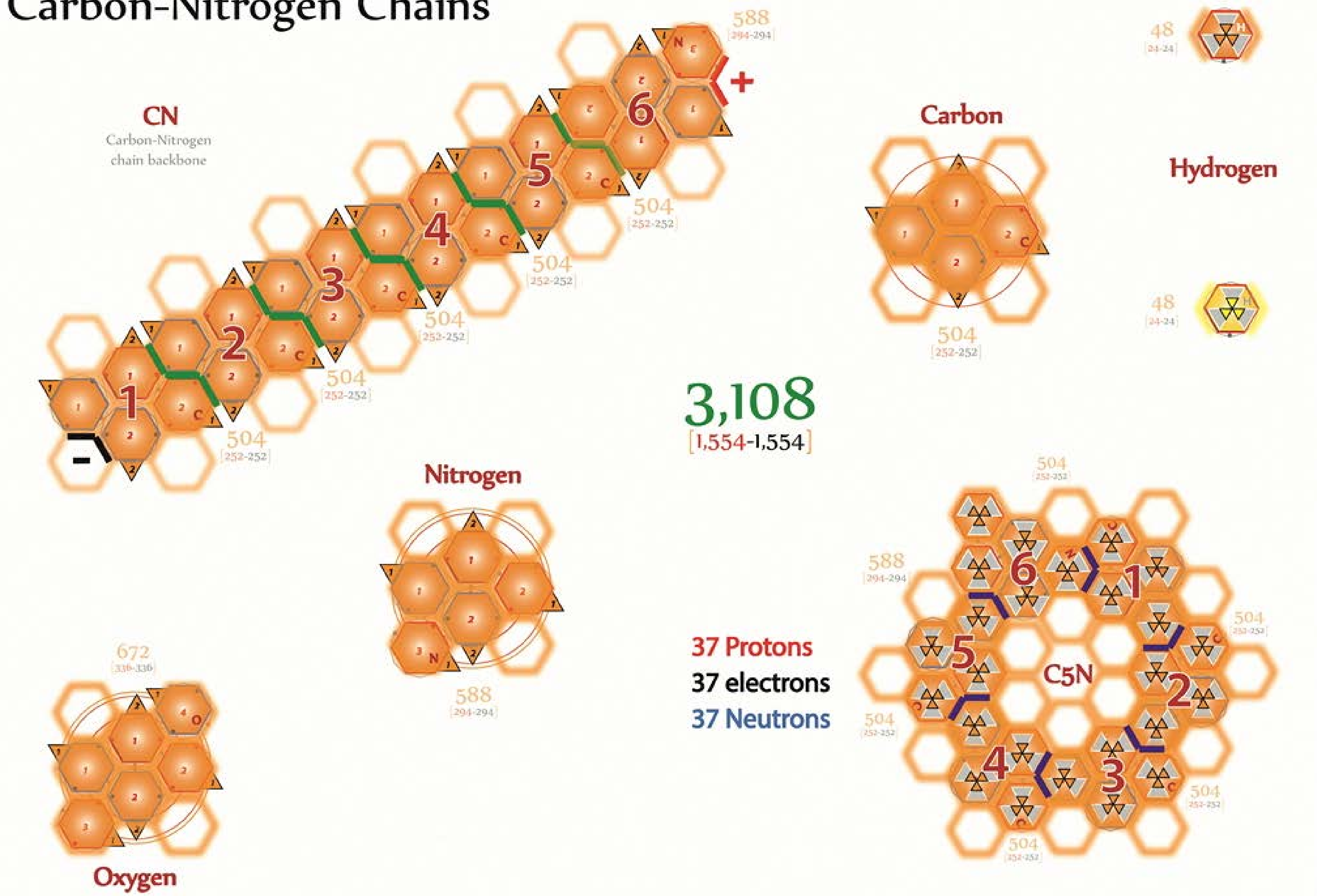
CN
Carbon-Nitrogen
chain backbone
1
2
3
4
5
6
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
588
[294-294]
-
+
Carbon
504
[252-252]
Hydrogen
48
[24-24]
48
[24-24]
Nitrogen
588
[294-294]
Oxygen
672
[336-336]
C5N
588
[294-294]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]

3,108
[1,554-1,554]

37 Protons
37 electrons
37 Neutrons

Copyright ABRAHAM [2008] - All rights reserved

Carbon-Oxygen Chains

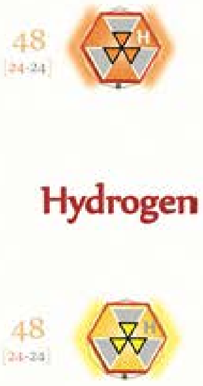
48
[24-24]
H
48
[24-24]
H

Hydrogen

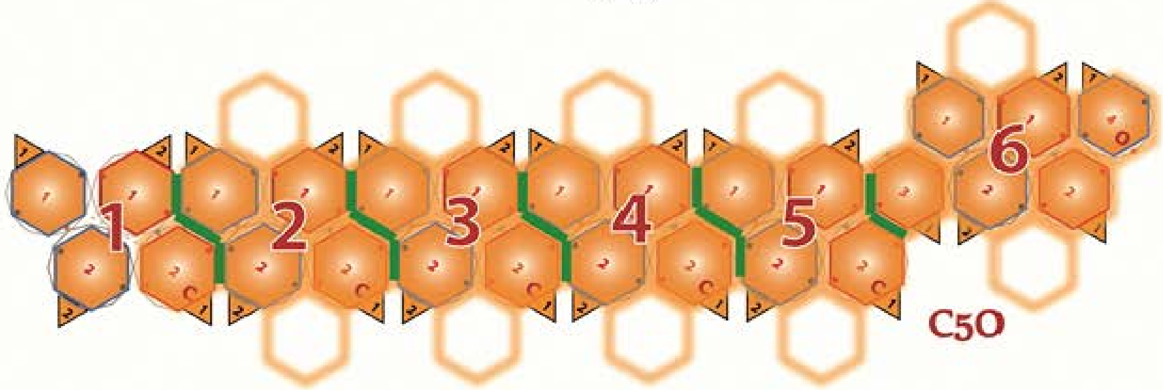
1
2
3
4
5
6
C
O
C5O

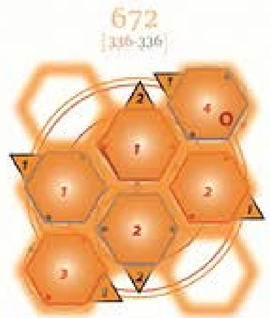
672
[336-336]
O

Oxygen

Heterocyclic Rings

Tetrahydropyran is the organic compound consisting of a saturated six-membered ring containing five carbon atoms and one oxygen atom.

Carbon

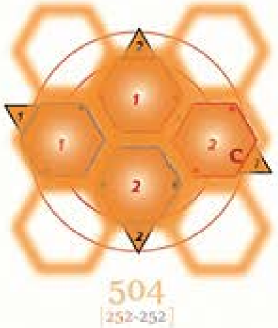
C
504
[252-252]

3,192
[1,596-1,596]

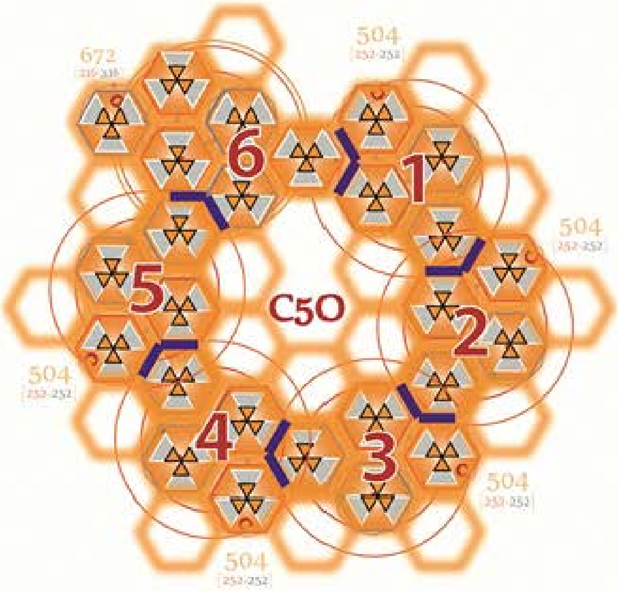
672
[336-336]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
504
[252-252]
1
2
3
4
5
6
C5O

38 Protons
38 electrons
38 Neutrons

Nitrogen

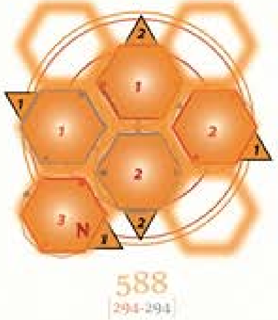
N
588
[294-294]

C6H2(NO2)3CH3

Trinitrotoluene

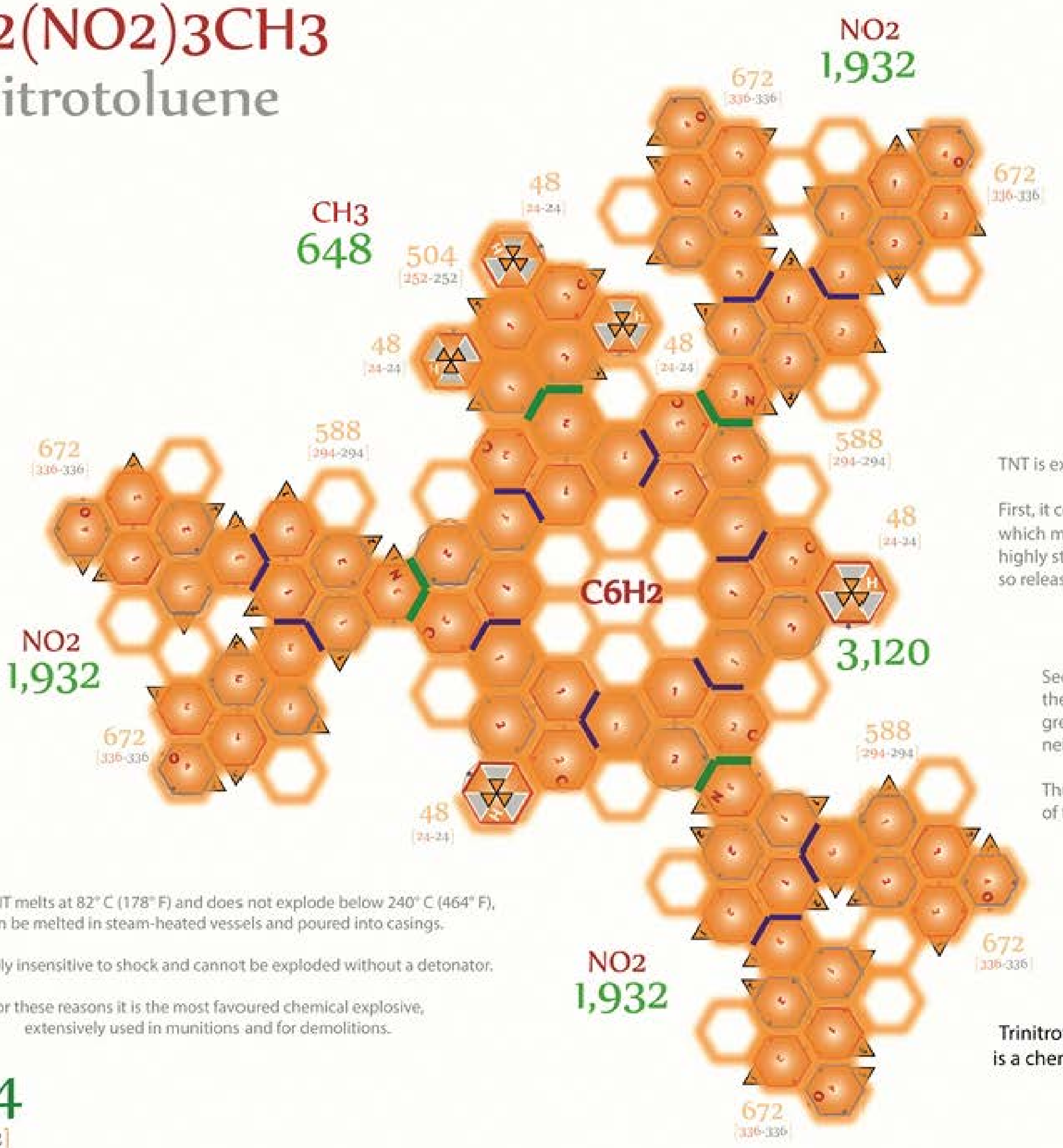
NO2
1,932
672
[336-336]
672
[336-336]
CH3
648
48
[24-24]
504
[252-252]
48
[24-24]
48
[24-24]
588
[294-294]
672
[336-336]
588
[294-294]
48
[24-24]
NO2
1,932
C6H2
3,120
672
[336-336]
588
[294-294]
48
[24-24]
672
[336-336]
NO2
1,932
672
[336-336]

2,4,6-trinitrotoluene is better known by its initials, TNT. It is an important explosive, since it can very quickly change from a solid into hot expanding gases.

TNT is explosive for two reasons;

First, it contains the elements carbon, oxygen and nitrogen, which means that when the material burns it produces highly stable substances (CO, CO2 and N2) with strong bonds, so releasing a great deal of energy.

Secondly, TNT is chemically unstable - the nitro groups are so closely packed that they experience a great deal of strain and hindrance to movement from their neighbouring groups.

Thus it doesn't take much of an initiating force to break some of the strained bonds, and the molecule then flies apart.

Because TNT melts at 82° C (178° F) and does not explode below 240° C (464° F), it can be melted in steam-heated vessels and poured into casings.

It is relatively insensitive to shock and cannot be exploded without a detonator.

For these reasons it is the most favoured chemical explosive, extensively used in munitions and for demolitions.

Trinitrotoluene, or more specifically, 2,4,6-trinitrotoluene, is a chemical compound with the formula C6H2(NO2)3CH3

9,564

[4,782-4,782]

(C6)4-H10
Pyrene

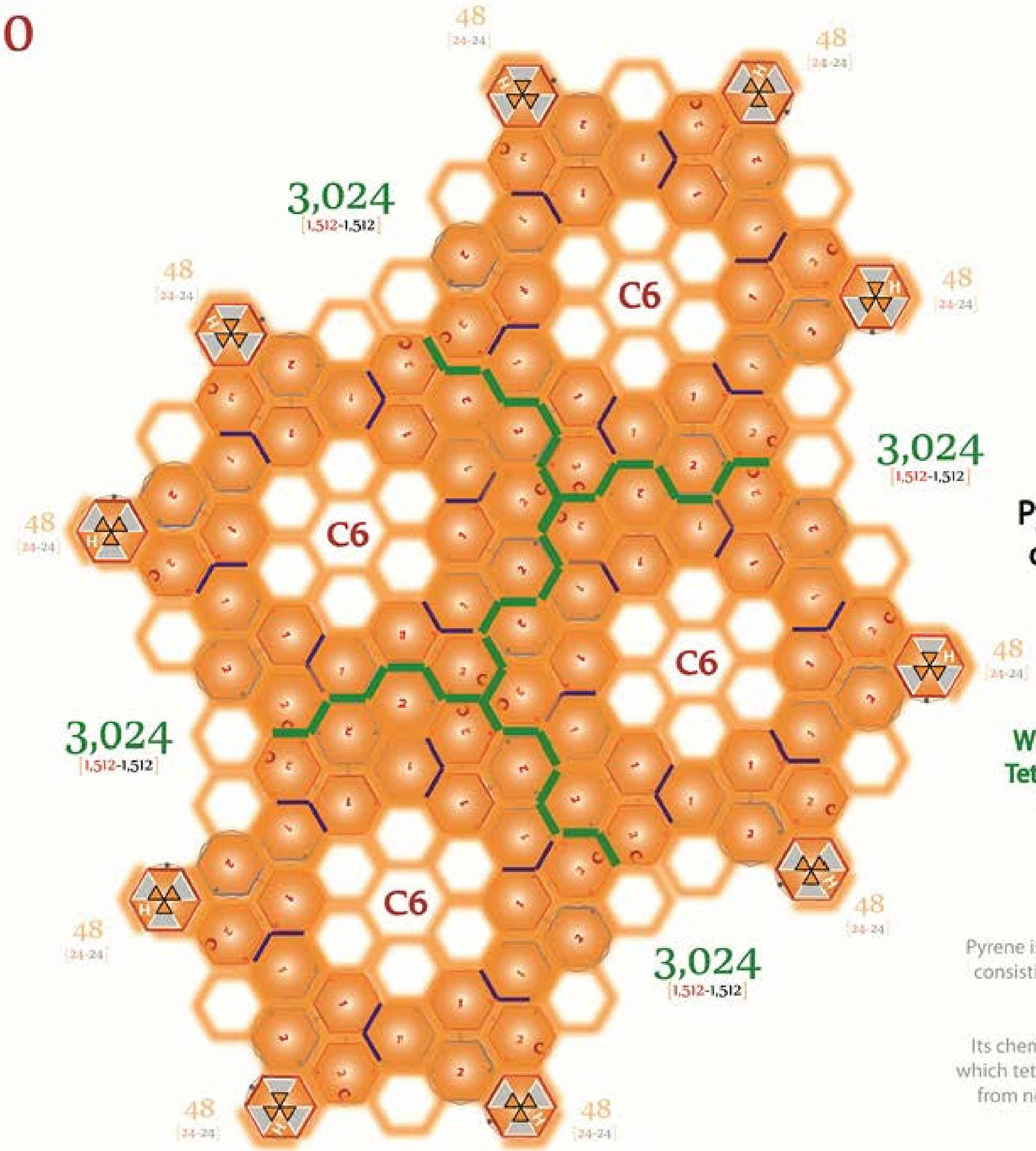
48
[24-24]
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
C6
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
C6
C6
48
[24-24]
3,024
[1,512-1,512]
48
[24-24]
48
[24-24]
C6
3,024
[1,512-1,512]
48
[24-24]
48
[24-24]

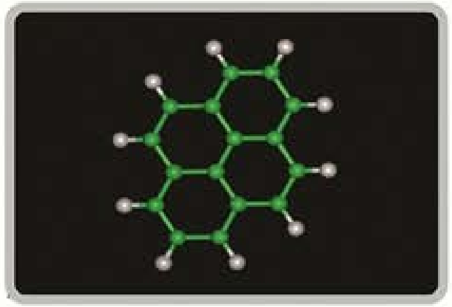

12,576
[6,288-6,288]

Pyrene highlights the failings of Lewis diagram structures

Which can be rectified using charged Tetryonic geometric Matter topologies for all elements and compound interactions & modelling

Pyrene is a polycyclic aromatic hydrocarbon (PAH) consisting of four fused benzene rings, resulting in a flat aromatic system.

Its chemcial formula as is often stated as C16H10, which tetryonic geometry shows is an error resulting from not using equilateral quantum geometries in chemical topology modelling.

Four cyclic carbon rings cannot be formed with only 16 C-atoms

Ammonium Nitrate

3,384

NH4NO3,

Ammonium nitrate is composed of two polyatomic ions:
1) Ammonium Ion (NH4+)
2) Nitrate Ion (NO3-)

The bond between these ions is an ionic bond meaning the ammonium ion transfers an electron to the nitrate ion.

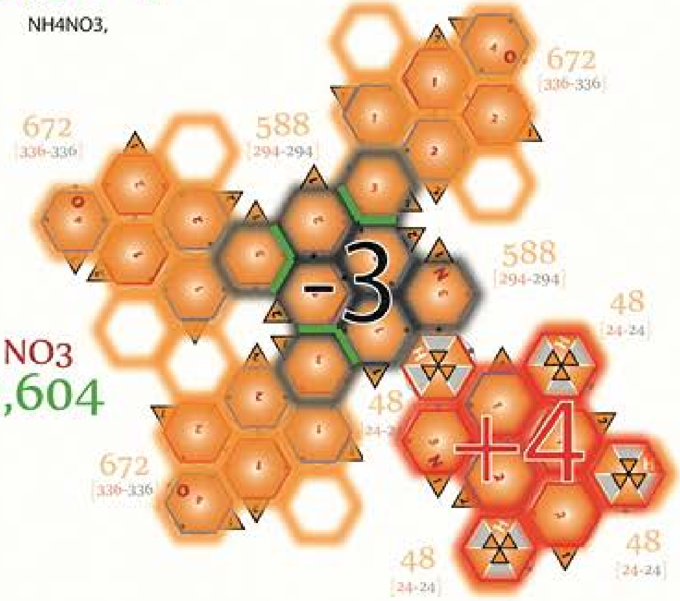
672
[336-336]
672
[336-336]
588
[294-294]
588
[294-294]
-3
+4
NO3
2,604
NH4
780
48
[24-24]
48
[24-24]
672
[336-336]
48
[24-24]
48
[24-24]

The bonds in a polyatomic atomic ion are covalent because they take place between gases.

This means that the hydrogens of the ammonium ion are bonded to the nitrogen covalently and that the oxygens of the nitrate are bonded to the nitrogen covalently

Ammonium nitrate decomposes into the gases nitrous oxide and water vapor when heated (non-explosive reaction); however, ammonium nitrate can be induced to decompose explosively by detonation.

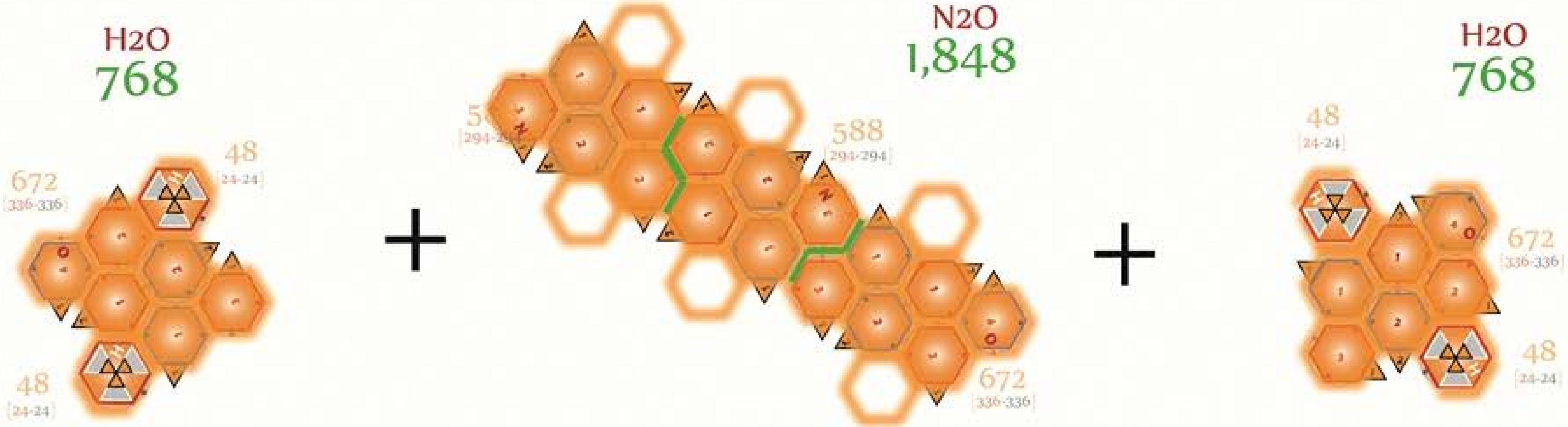
H2O
768
672
[336-336]
48
[24-24]
48
[24-24]
+
N2O
1,848
588
[294-294]
588
[294-294]
672
[336-336]
+
H2O
768
48
[24-24]
672
[336-336]
48
[24-24]

Carbohydrates

A carbohydrate is an organic compound that consists only of carbon, hydrogen, and oxygen (with a hydrogen:oxygen ratio of 2:1) in other words, with the empirical formula $C_m(H_2O)_n$

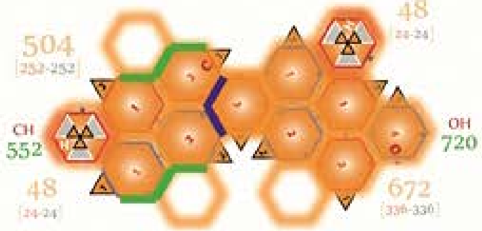
504
[252-252]
48
[24-24]
CH
552
48
[24-24]
OH
720
672
[336-336]

Monosaccharides

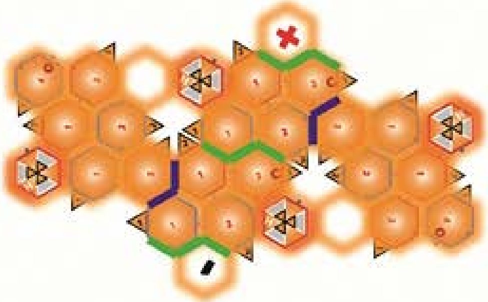

Disaccharides

Ribose

Gluclose

Lactose

$$C_m(H_2O)_n$$

Fructose

Sucrose

Deoxyribose

Oligosaccharides

Polysaccharides

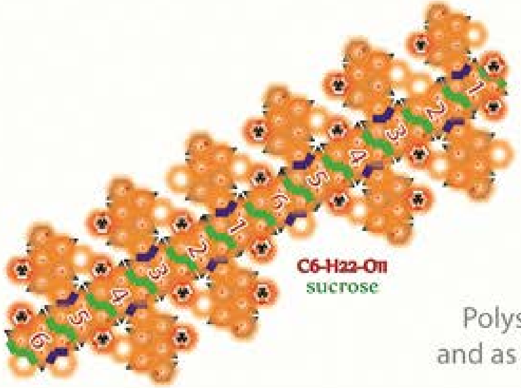
C6-H22-O11
sucrose
1
2
3
4
5
6

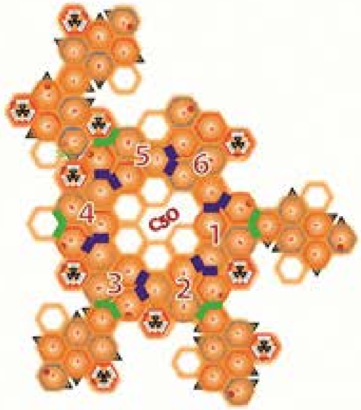

Carbohydrates perform numerous roles in living organisms.
Polysaccharides serve for the storage of energy (e.g., starch and glycogen),
and as structural components (e.g., cellulose in plants and chitin in arthropods)

Copyright ABRAHAM [2008] - All rights reserved

Monosaccharides

Monosaccharides are the simplest carbohydrates in that they cannot be hydrolyzed to smaller carbohydrates

The Carbon nuclei in Monosaccharides join together to form chains of biologically important carbohydrates

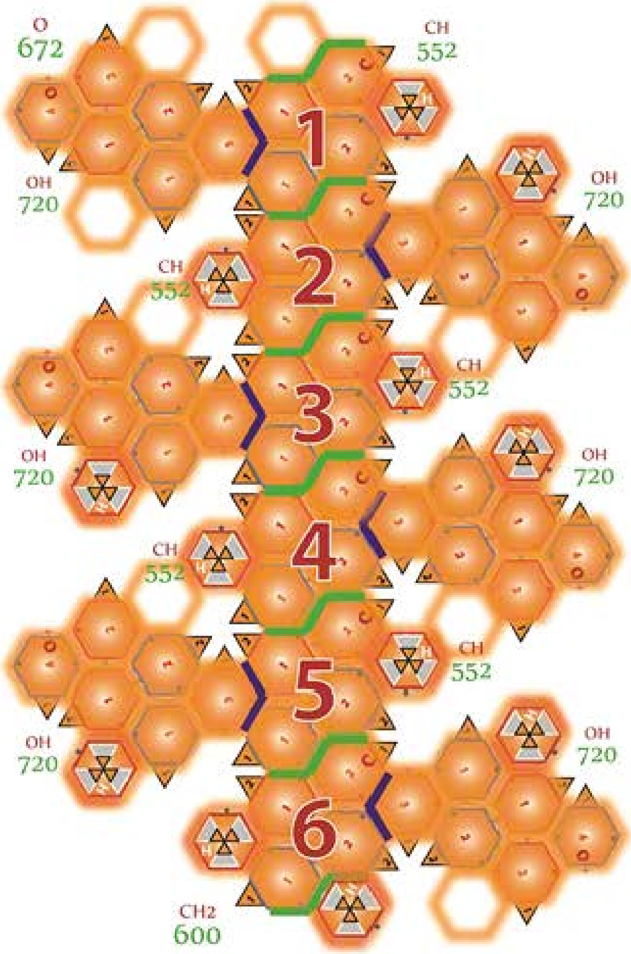
O
672
CH
552
1
OH
720
OH
720
CH
552
2
CH
552
OH
720
3
CH
552
OH
720
OH
720
CH
552
4
CH
552
5
OH
720
OH
720
6
CH2
600

C(H2O)6

D-glucose

C6-H12-O6

1,272

[636-636]

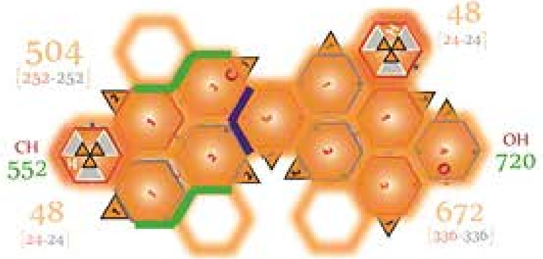
504
[252-252]
48
[24-24]
CH
552
OH
720
48
[24-24]
672
[336-336]

C(H2O)

The general chemical formula of an unmodified monosaccharide is (C•H2O) n, literally a "carbon hydrate."

CH
552

CH-OH

OH
720

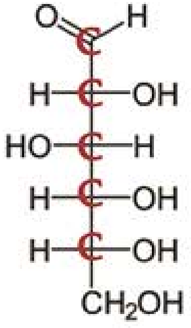
O
H
H
OH
HO
H
H
OH
H
OH
CH2OH

D-Glucose

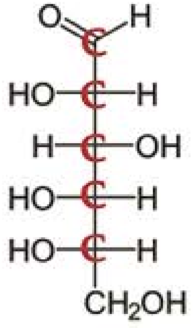
O
H
HO
H
H
OH
HO
H
HO
H
CH2OH

L-Glucose

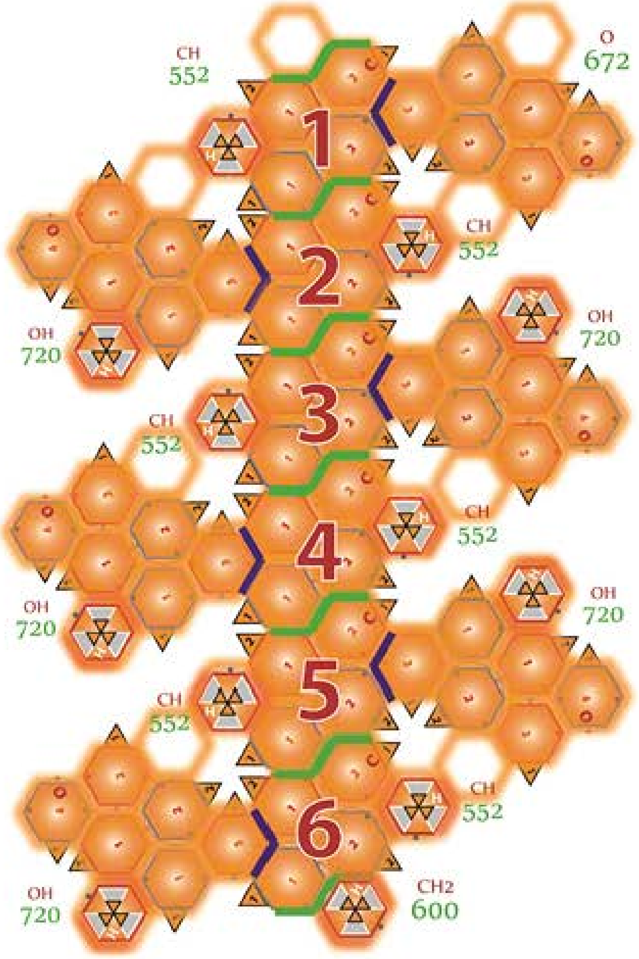
CH
552
O
672
1
CH
552
2
OH
720
OH
720
3
CH
552
CH
552
4
OH
720
OH
720
CH
552
5
CH
552
6
OH
720
CH2
600

C(H2O)6

L-glucose

C6-H12-O6

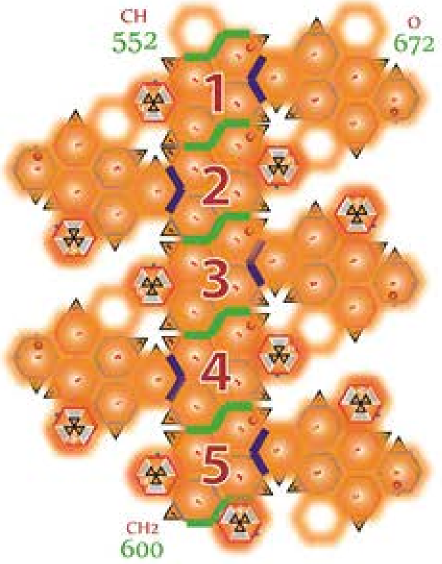
CH
552
O
672
1
2
3
4
5
CH2
600

C(H2O)5
ribose
C5-H10-O5

C(H2O)6
fructose
C6-H12-O6

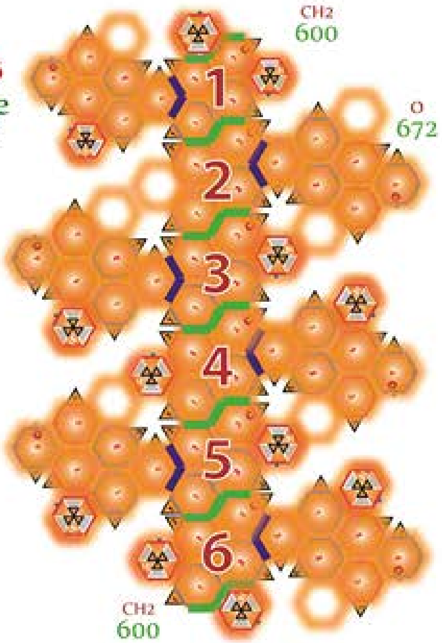
CH2
600
O
672
1
2
3
4
5
6
CH2
600

C(H2O)n

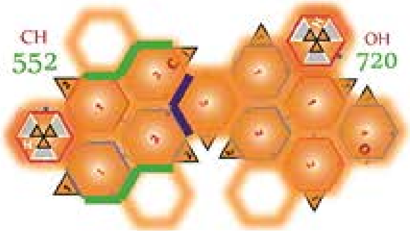
CH
552
OH
720

Major Monosaccharides

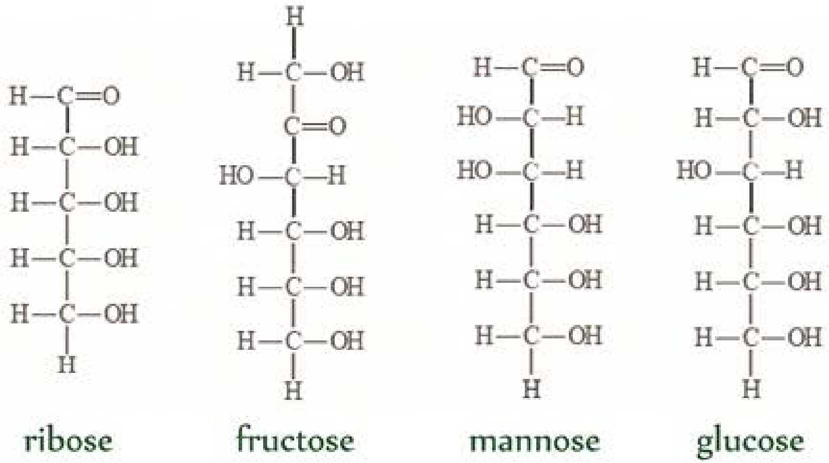
ribose
fructose
mannose
glucose

D D & L sugars are chiral, mirror images of one another and have the same name. **L**

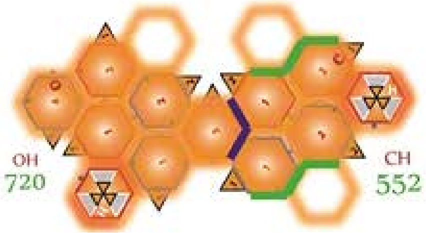
OH
720
CH
552

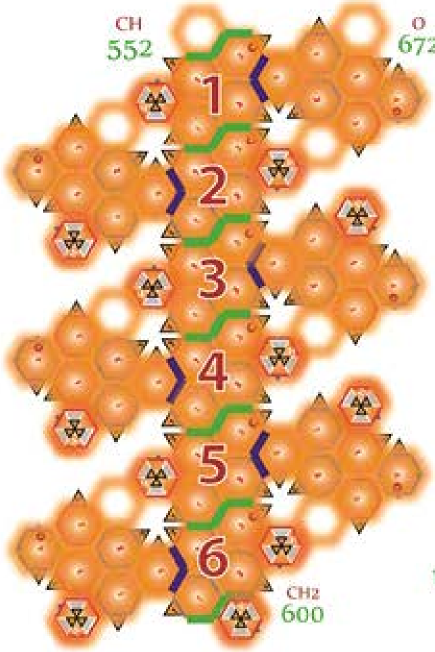
CH
552
O
672
1
2
3
4
5
6
CH2
600

C(H2O)6
mannose
C6-H12-O6

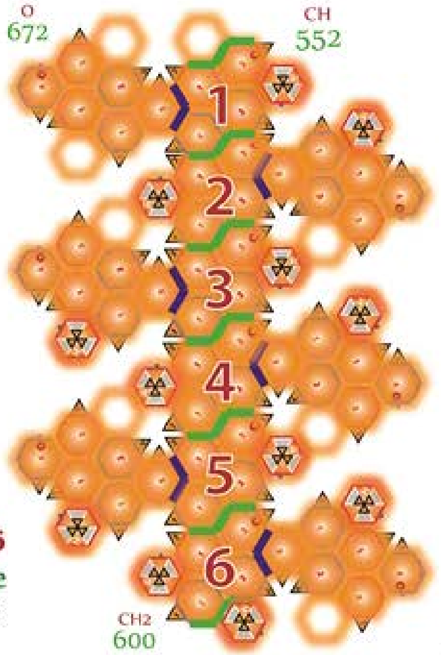
O
672
CH
552
1
2
3
4
5
6
CH2
600

C(H2O)6
glucose
C6-H12-O6

cyclic monosaccharides

A monosaccharide often switches from the acyclic (open-chain) form to a cyclic form, through a nucleophilic addition reaction between the carbonyl group and one of the hydroxyls of the same molecule.

For many monosaccharides (including glucose), the cyclic forms predominate, and therefore the same name commonly is used for forms of isomers.

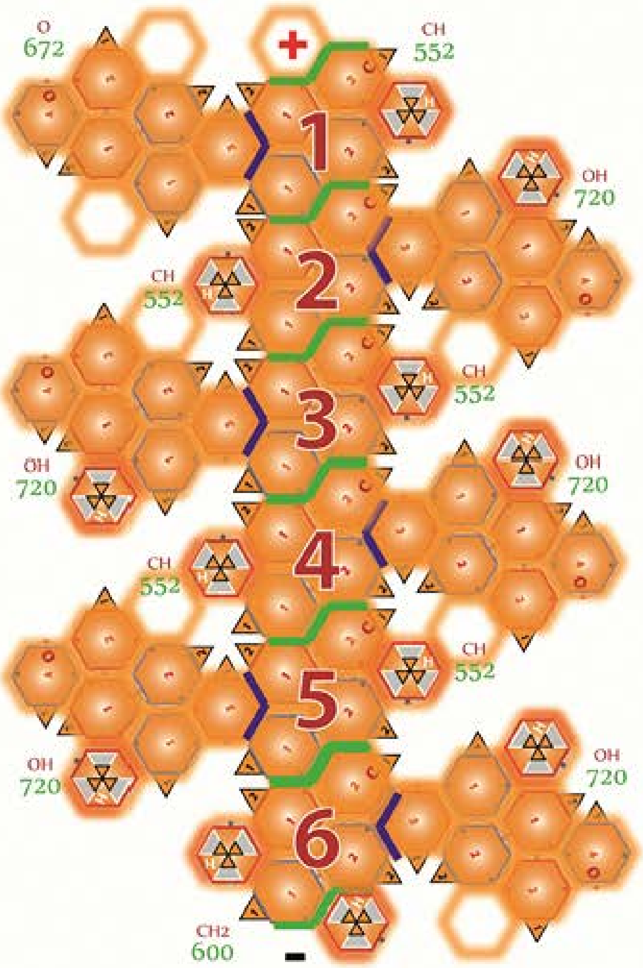
O
672
CH
552
+
1
OH
720
2
CH
552
CH
552
3
OH
720
OH
720
4
CH
552
CH
552
5
OH
720
OH
720
6
CH2
600
−

C(H2O)6
glucose
C6-H12-O6

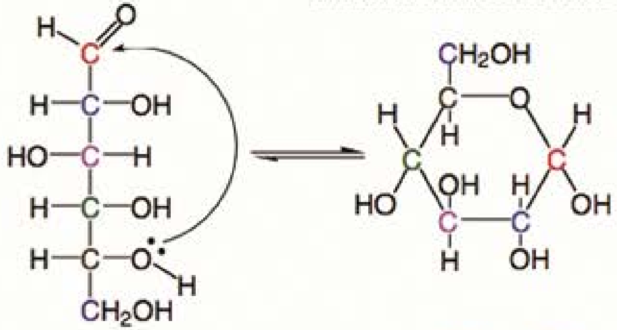
H
C
O
H
C
OH
HO
C
H
H
C
OH
H
C
O
H
CH2OH
CH2OH
C
O
H
H
H
C
C
H
HO
OH
H
OH
C
C
H
OH

C(H2O)6
glucopyranose
C6-H12-O6

cyclic Oxygen-Carbon ring

Allotropic cyclic [Carbon-Oxygen] forms of polycarbohydrates can also be formed via other element bonding processes

C_6-H_{12}-O_6

7,632

[3,816-3,816]

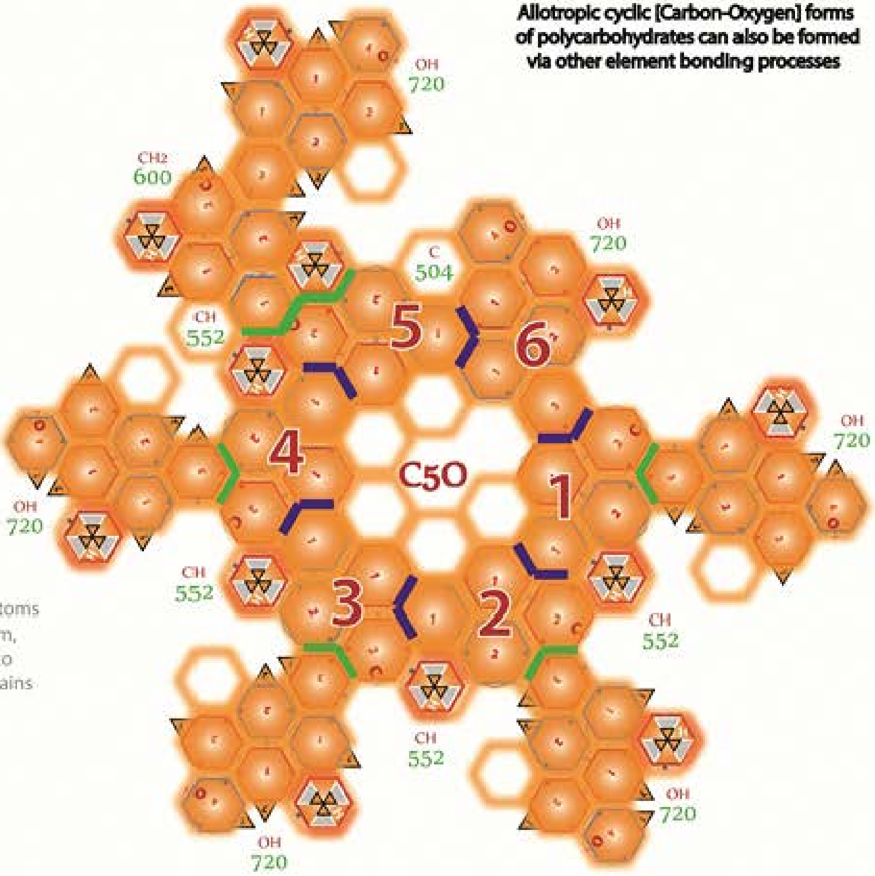
OH
720
CH2
600
OH
720
C
504
CH
552
5
6
4
C50
1
OH
720
OH
720
CH
552
3
2
CH
552
CH
552
OH
720
OH
720

The reaction creates a ring of carbon atoms closed by one bridging oxygen atom, which in turn can be easily broken to restore the linear monosaccharide chains

linear saccharide chain

The Carbon nuclei in Monosaccharides chains can join together to form cyclic [Carbon-Oxygen] rings or polycarbohydrates through condensation

acyclic monosaccharides

Condensed Disaccharides

Condensed Disaccharides are formed when two monosaccharides are joined together and a molecule of water is removed

For example, milk sugar (lactose) is made from glucose and galactose
whereas cane sugar (sucrose) is made from glucose and fructose

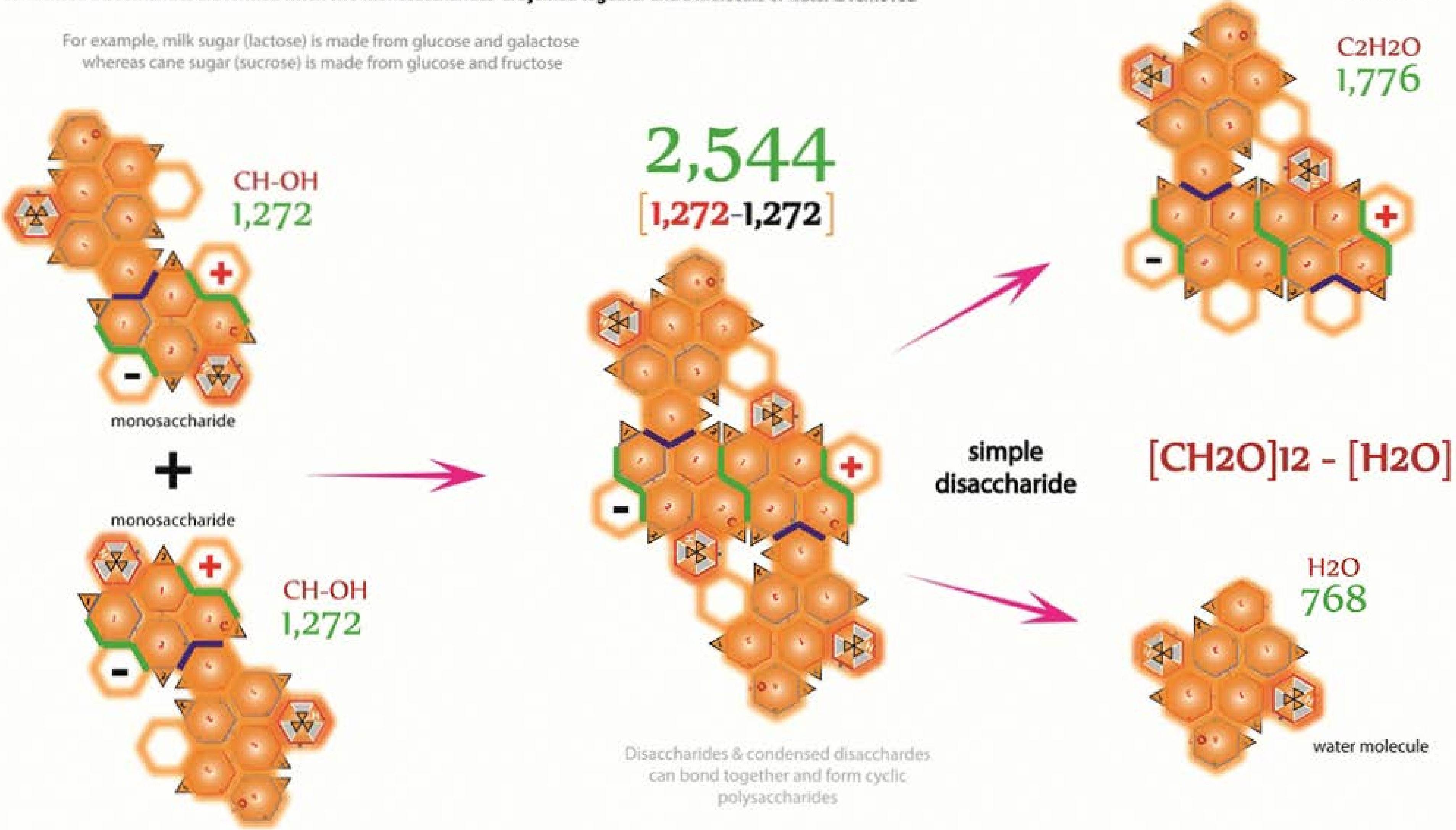

CH-OH
1,272

monosaccharide

+

monosaccharide

CH-OH
1,272

2,544
[1,272-1,272]

condensed
disaccharide

C_2H_2O
1,776

simple disaccharide

$[CH_2O]_{12} - [H_2O]$

H_2O
768

water molecule

Disaccharides & condensed disaccharides
can bond together and form cyclic
polysaccharides

The reverse of this reaction, the formation of two monosaccharides from one disaccharide, is called a hydrolysis reaction and requires one water molecule to supply the H and OH to the sugars formed.

Sucrose is used in many plants for transporting food reserves, often from the leaves to other parts of the plant.
Lactose is the sugar found in the milk of mammals and maltose is the first product of starch digestion
and is further broken down to glucose before absorption in the human gut.

Copyright ABRAHAM [2008] - All rights reserved

14,496

[7,248-7,248]

Disaccharides

Disaccharides are one of the four basic categories of carbohydrates,
(the others are monosaccharides, oligosaccharides, and polysaccharides)

Monosaccharides, such as glucose, are the monomers out of which disaccharides are constructed.

C12-(H2O)11

Disaccharides & condensed disaccharides can bond together and form cyclic polysaccharides

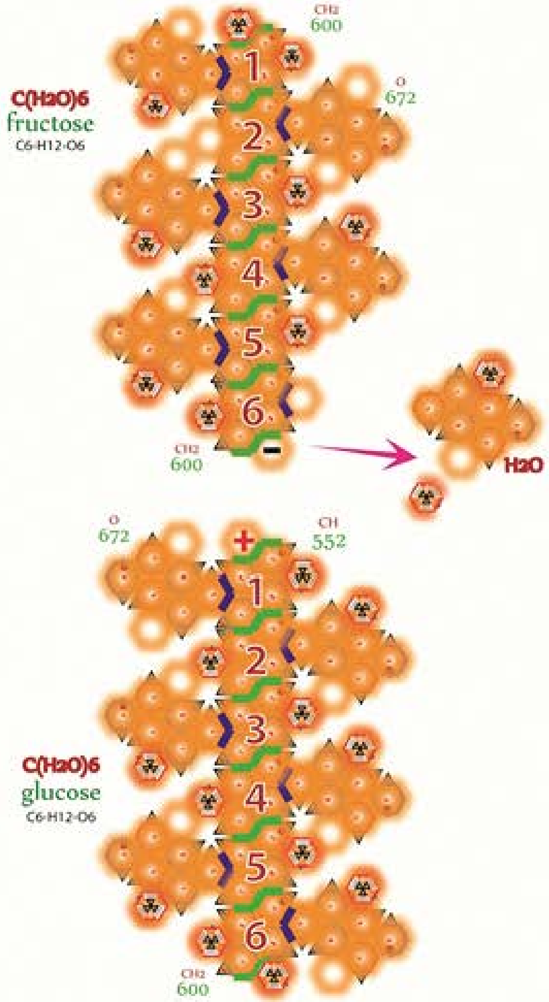
CH2
600
O
672
C(H2O)6
fructose
C6-H12-O6
1
2
3
4
5
6
CH2
600
H2O
O
672
CH
552
1
2
3
4
5
6
C(H2O)6
glucose
C6-H12-O6
CH2
600

sucrose
C12-H22-O11

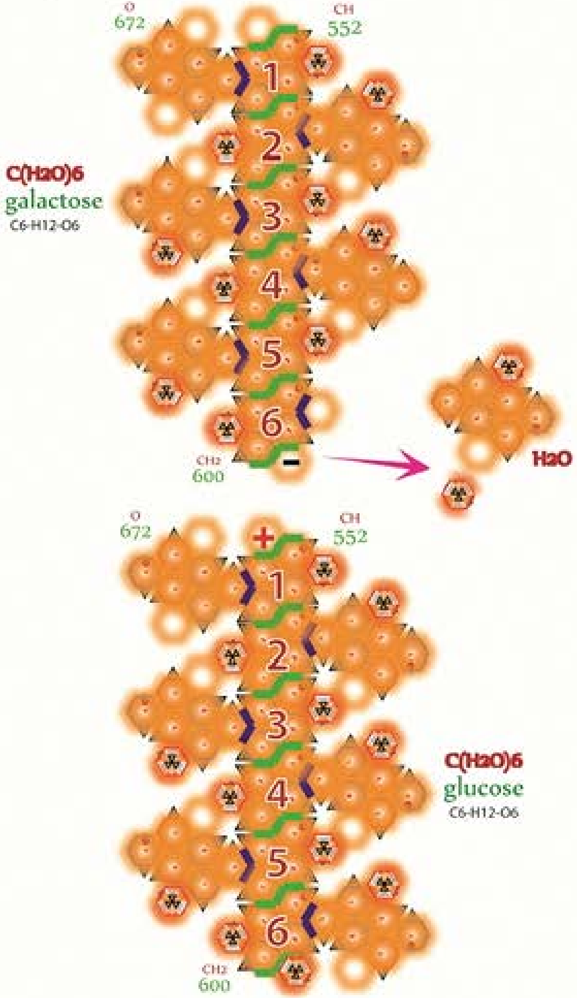
O
672
CH
552
C(H2O)6
galactose
C6-H12-O6
1
2
3
4
5
6
CH2
600
H2O
O
672
CH
552
1
2
3
4
5
6
C(H2O)6
glucose
C6-H12-O6
CH2
600

lactose
C12-H22-O11

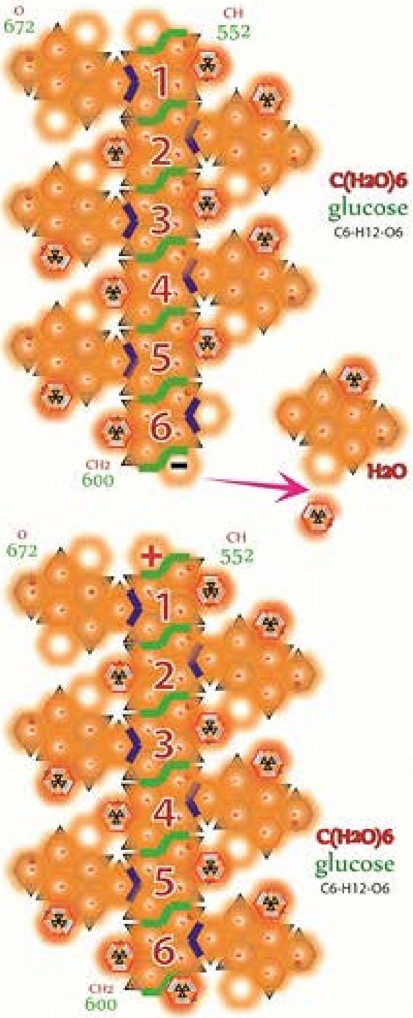
O
672
CH
552
C(H2O)6
glucose
C6-H12-O6
1
2
3
4
5
6
CH2
600
H2O
O
672
CH
552
1
2
3
4
5
6
C(H2O)6
glucose
C6-H12-O6
CH2
600

maltose
C12-H22-O11

The general chemical formula for carbohydrates, C(H2O),
gives the relative proportions of carbon, hydrogen, and oxygen in a monosaccharide
(the proportion of these atoms are 1:2:1)-(H2O)

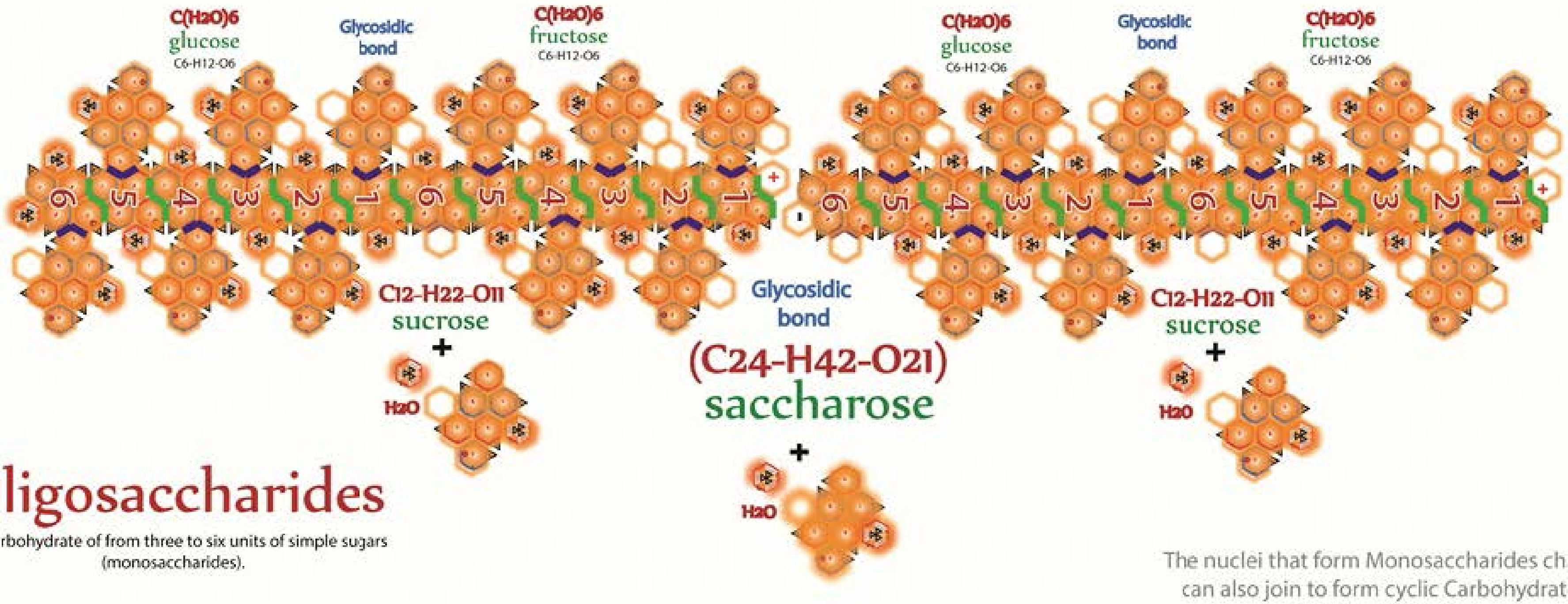
C(H2O)6
glucose
C6-H12-O6
Glycosidic
bond
C(H2O)6
fructose
C6-H12-O6
C(H2O)6
glucose
C6-H12-O6
Glycosidic
bond
C(H2O)6
fructose
C6-H12-O6
6 5 4 3 2 1 6 5 4 3 2 1
6 5 4 3 2 1 6 5 4 3 2 1
C12-H22-O11
sucrose
+
H2O
Glycosidic
bond
(C24-H42-O21)
saccharose
+
H2O
C12-H22-O11
sucrose
+
H2O

Oligosaccharides

any carbohydrate of from three to six units of simple sugars (monosaccharides).

The nuclei that form Monosaccharides chains can also join to form cyclic Carbohydrates [Polysaccharides]

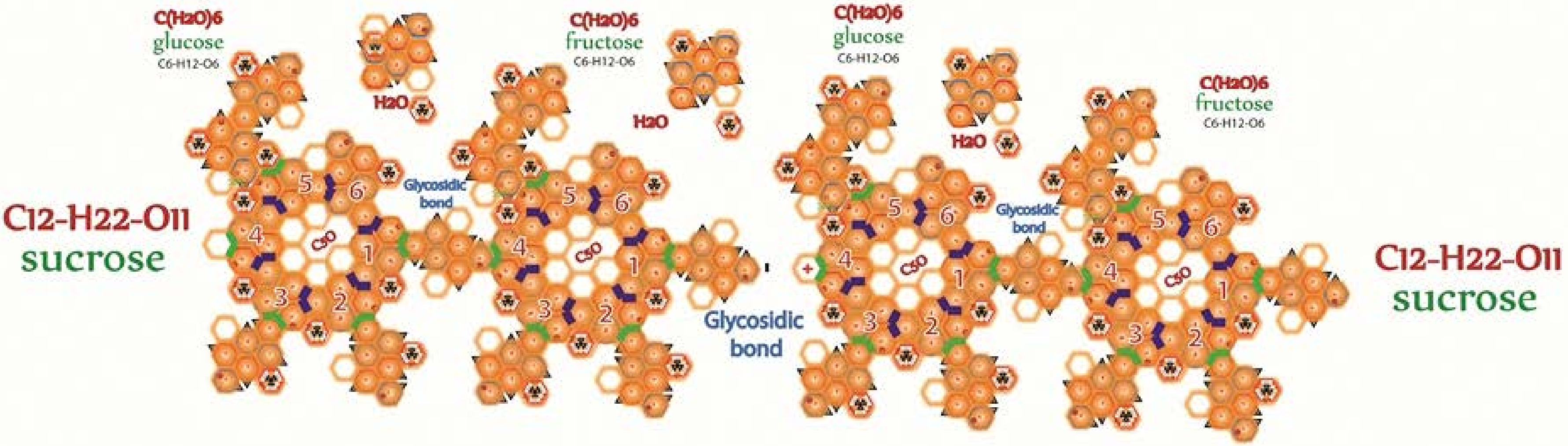
C(H2O)6
glucose
C6-H12-O6
H2O
C(H2O)6
fructose
C6-H12-O6
H2O
C(H2O)6
glucose
C6-H12-O6
H2O
C(H2O)6
fructose
C6-H12-O6
C12-H22-O11
sucrose
Glycosidic
bond
Glycosidic
bond
Glycosidic
bond
C12-H22-O11
sucrose
5 6 4 C6O 1 3 2

An oligosaccharide is a saccharide polymer containing a small number (typically two to ten) component sugars (monosaccharides). Oligosaccharides can have many functions; for example, they are commonly found on the plasma membrane of animal cells where they can play a role in cell–cell recognition.

Starch or amylum is a carbohydrate consisting of a large number of glucose units joined by glycosidic bonds. This polysaccharide is produced by most green plants as an energy store.

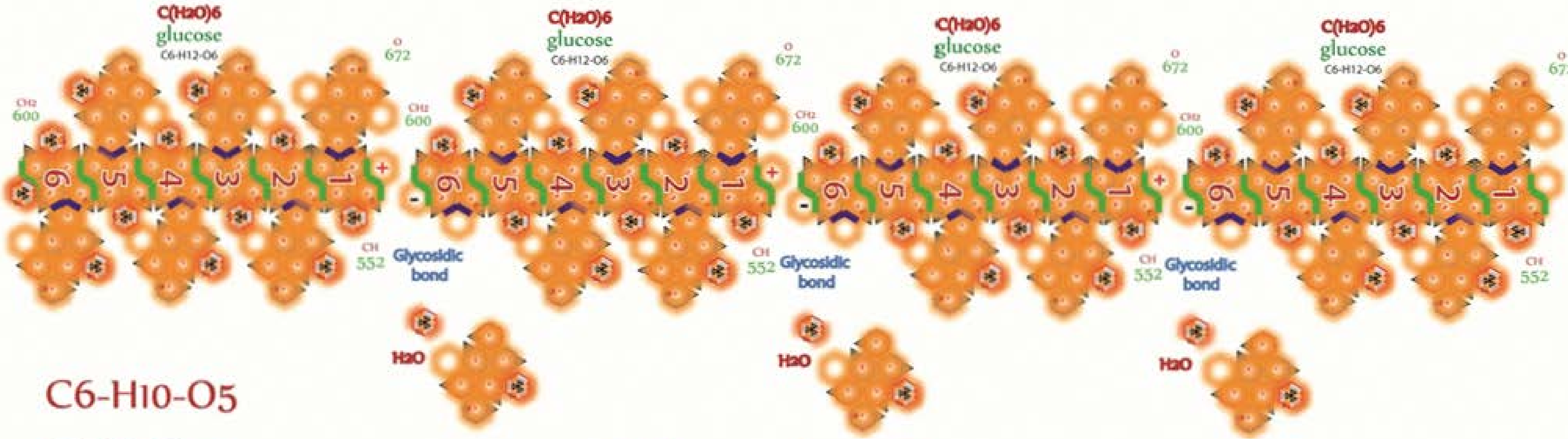
C(H2O)6
glucose
C6-H12-O6
O
672
CH2
600
CH
552
Glycosidic bond
6
5
4
3
2
1
+
-
H2O

C6-H10-O5

6,864 n

[3,432-3,432]

Polysaccharides

Polysaccharides have a general formula of Cx(H2O)y where x is usually a large number between 200 and 2500.

$C_x[H2O]_y$

The repeating units in a Polysaccharide polymer backbone are often six-carbon monosaccharides, their general formula can be represented as (C6-H10-O5)n

Polysaccharides are generally comprised of chains of cyclic monosaccharides

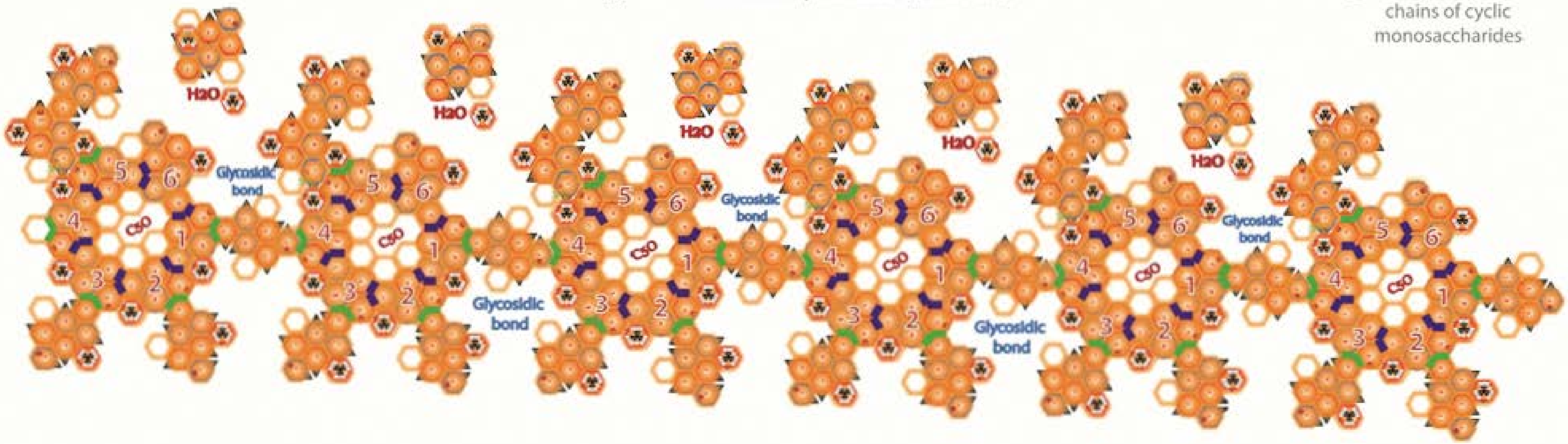
H2O
Glycosidic bond
5
6
4
C5O
1
3
2

Cellulose is an organic compound with the formula (C6H10O5)n, a polysaccharide consisting of a linear chain of several hundred to over ten thousand linked D-glucose sugars.

Amines

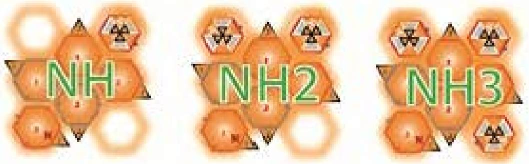

Amines are organic compounds that contain Nitrogen

NH2-[HC-R]-COOH + NH2-[HC-R]-COOH

As both the amine and carboxylic acid groups of amino acids can react to form amide bonds, one amino acid molecule can react with another and become joined through an amide linkage. This polymerization of amino acids is what creates proteins.

OH + NH2 -----> NH + H2O

This yields a peptide bond and a molecule of water via a condensation reaction

Carboxylic Acids

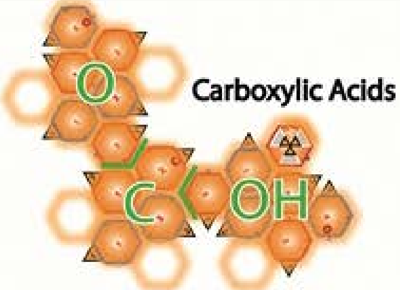

The general formula of a carboxylic acid is R-COOH

Amino Acids

Amino acids are biologically important organic compounds made from amine and carboxylic acid functional groups, along with a side-chain specific to each amino acid

un-ionized amino acid

NH2-[HC-R]-COOH

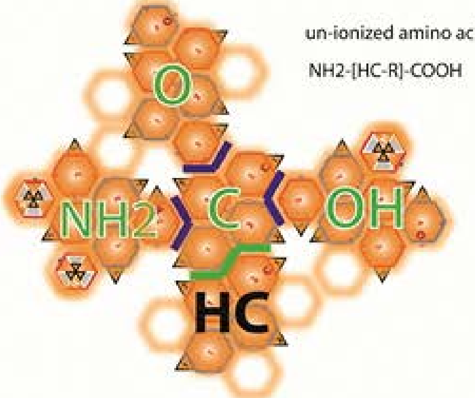

zwitterionic amino acid

NH3-[HC-R]-COO

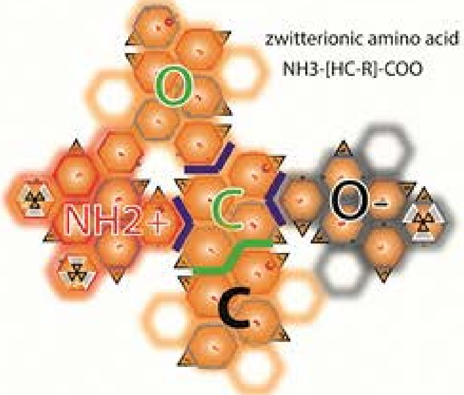

The α-carbon.

The α-amino acids in peptides and proteins consist of a carboxylic acid (-COOH) and an amino (-NH2) functional group attached to the same tetrahedral carbon atom

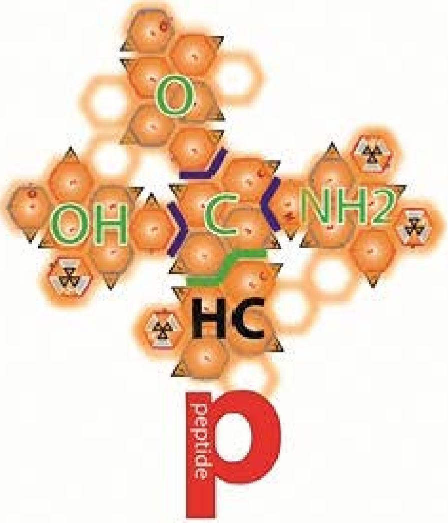

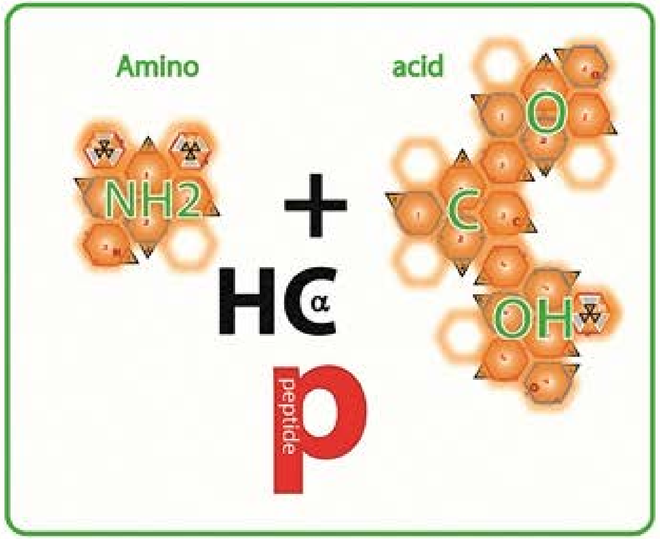

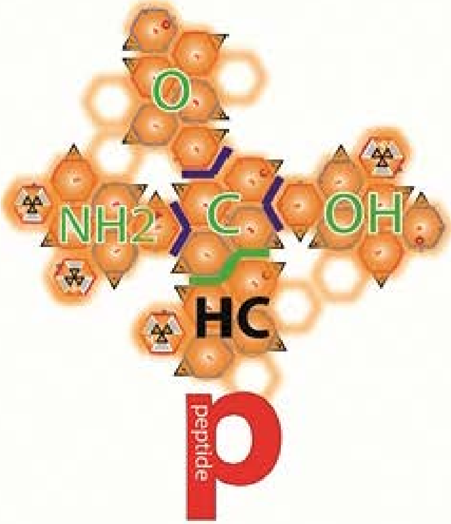

The peptide group attached to the alpha carbon distinguishes one amino acid from another (Tetryonic theory defines these compound side-chains of atoms as peptides [p])

Copyright ABRAHAM [2008] - All rights reserved

Peptide side chains

Amino acids are the structural units (monomers) that join together
to form short polymer chains called peptides;
or longer chains called either polypeptides or proteins

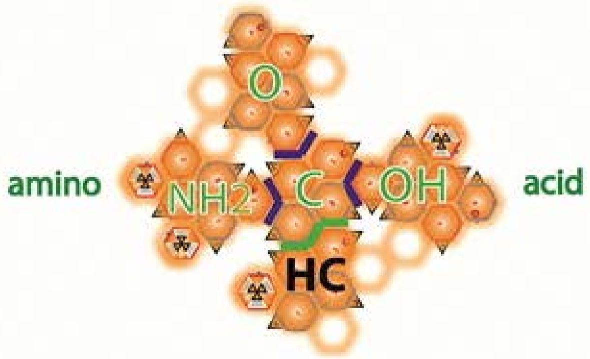

AAA
AAG
AAC
AAU
AGA
AGG
AGC
AGU
ACA
ACG
ACC
ACU
AUA
AUG
AUC
AUU

A
C5-H5-N5
Adenine

C
C4-H5-N3-O
Cytosine

CAA
CAG
CAC
CAU
CGA
CGG
CGC
CGU
CCA
CCG
CCC
CCU
CUA
CUG
CUC
CUU

3,180
Gly
C2H5NO2

3,780
Ala
C3H7NO2

4,452
Ser
C3H7NO3

4,884
Pro
C5H9NO2

4,980
Val
C5H11NO2

5,052
Thr
C4H9NO3

5,124
Cys
C3H7NO2S

5,580
Lle
C6H13NO2

5,580
Leu
C6H13NO2

5,592
Asn
C4H8N2O3

5,628
Asp
C4H7NO4

p
peptide side chain

6,192
Gln
C5H10N2O3

6,216
Lys
C6H14N2O2

6,228
Glu
C5H9NO4

6,324
Met
C5H11NO2S

6,564
His
C6H9N3O2

6,996
Phe
C9H11NO2

7,392
Arg
C6H14N4O2

7,668
Tyr
C9H11NO3

8,640
Trp
C11H12N2O2

current chemical theory hypothesizes that all peptides and proteins are the result of codon triplet side chains on amino acids

Tetryonic theory reveals serious flaws in this line of thought

GAA
GAG
GAC
GAU
GGA
GGG
GGC
GGU
GCA
GCG
GCC
GCU
GUA
GUG
GUC
GUU

G
C5-H5-N5-O
Guanine

U
C4-H4-N2-O2
Uracil

UAA
UAG
UAC
UAU
UGA
UGG
UGC
UGU
UCA
UCG
UCC
UCU
UUA
UUG
UUC
UUU

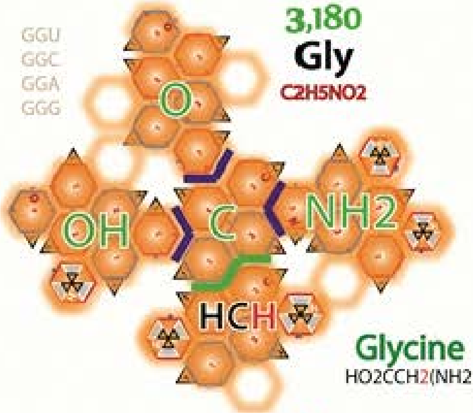
O
OH
C
NH2
HCH

GGU
GGC
GGA
GGG

3,180
Gly
$C_2H_5NO_2$

Glycine
HO2CCH2(NH2)

HCH

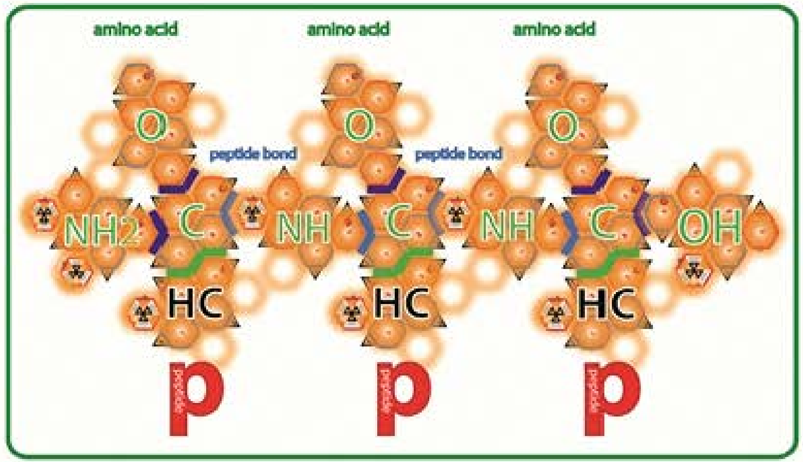
amino acid
amino acid
amino acid
O
O
O
peptide bond
peptide bond
NH2
C
NH
C
NH
C
OH
HC
HC
HC
p
p
p

poly-peptide chains

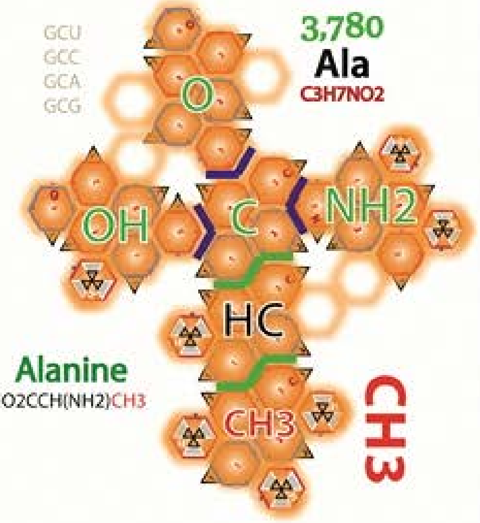
O
OH
C
NH2
HC
CH3
CH3

GCU
GCC
GCA
GCG

3,780
Ala
$C_3H_7NO_2$

Alanine
HO2CCH(NH2)CH3

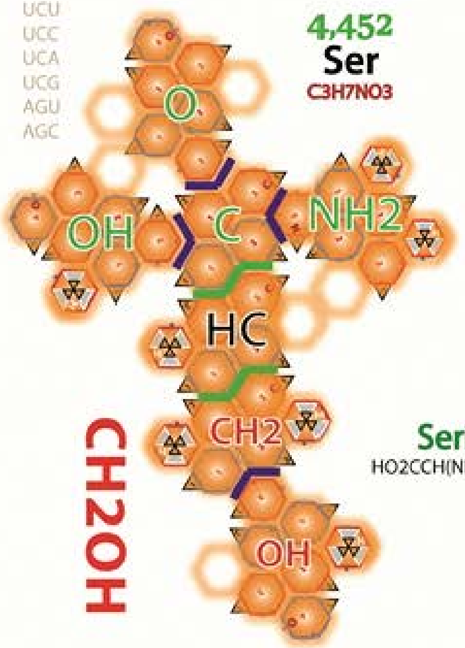
O
OH
C
NH2
HC
CH2
OH
CH2OH

UCU
UCC
UCA
UCG
AGU
AGC

4,452
Ser
$C_3H_7NO_3$

Serine
HO2CCH(NH2)CH2OH

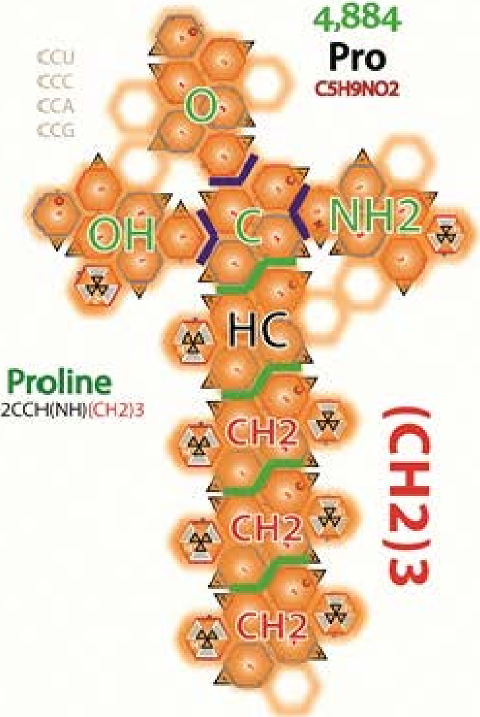
O
OH
C
NH2
HC
CH2
CH2
CH2
(CH2)3

CCU
CCC
CCA
CCG

4,884
Pro
$C_5H_9NO_2$

Proline
HO2CCH(NH)(CH2)3

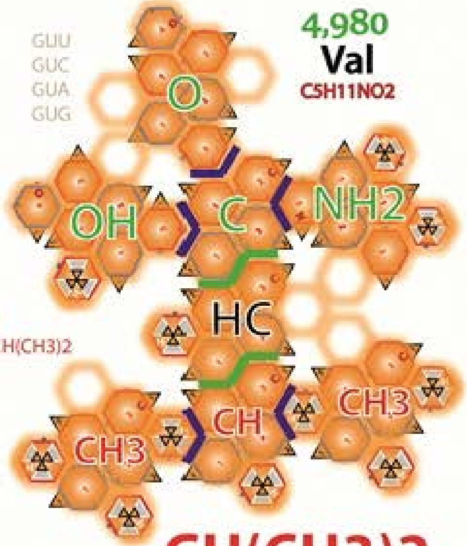
O
OH
C
NH2
HC
CH3
CH
CH3

GUU
GUC
GUA
GUG

4,980
Val
$C_5H_{11}NO_2$

Valine
HO2CCH(NH2)CH(CH3)2

CH(CH3)2

Copyright ABRAHAM [2008] - All rights reserved

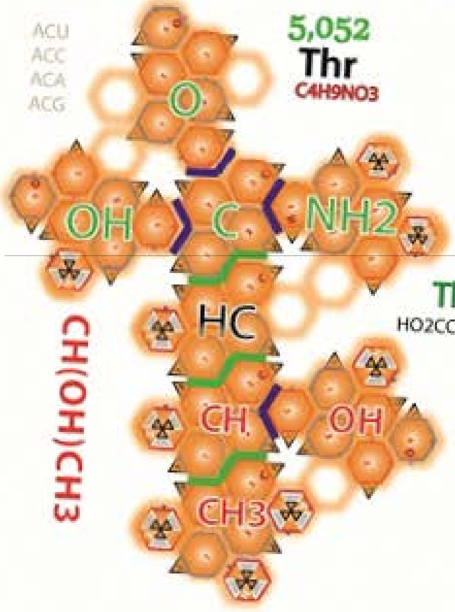
ACU
ACC
ACA
ACG
5,052
Thr
C4H9NO3
O
OH
C
NH2
HC
CH
OH
CH3
CH(OH)CH3

Threonine
HO2CCH(NH2)CH(OH)CH3

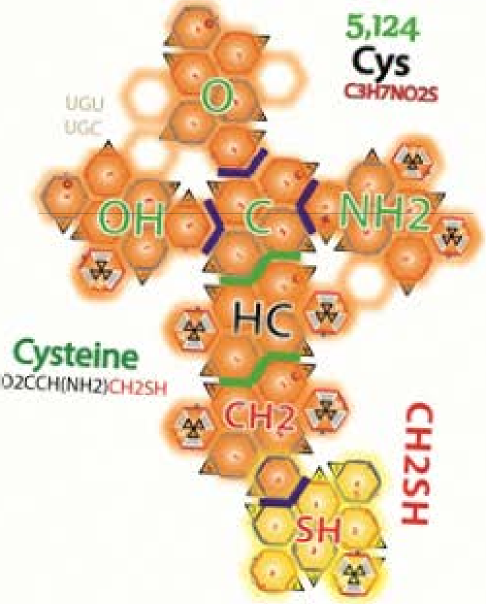
UGU
UGC
5,124
Cys
C3H7NO2S
O
OH
C
NH2
HC
CH2
SH
CH2SH

Cysteine
HO2CCH(NH2)CH2SH

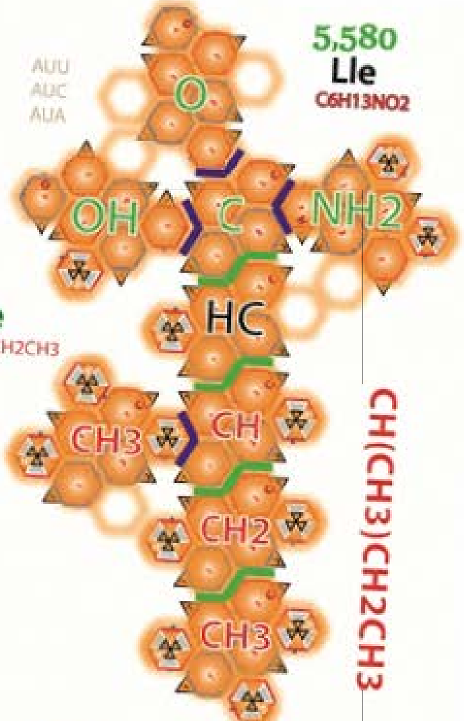
AUU
AUC
AUA
5,580
Lle
C6H13NO2
O
OH
C
NH2
HC
CH3
CH
CH2
CH3
CH(CH3)CH2CH3

Isoleucine
HO2CCH(NH2)CH(CH3)CH2CH3

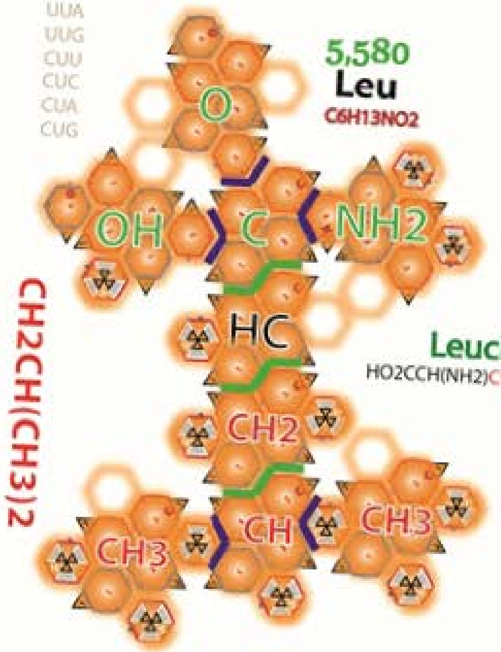
UUA
UUG
CUU
CUC
CUA
CUG
5,580
Leu
C6H13NO2
O
OH
C
NH2
HC
CH2
CH3
CH
CH3
CH2CH(CH3)2

Leucine
HO2CCH(NH2)CH2CH(CH3)2

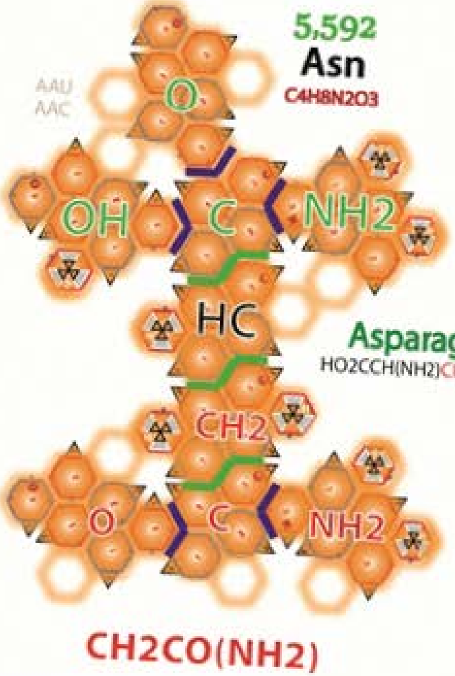
AAU
AAC
5,592
Asn
C4H8N2O3
O
OH
C
NH2
HC
CH2
O
C
NH2
CH2CO(NH2)

Asparagine
HO2CCH(NH2)CH2CO(NH2)

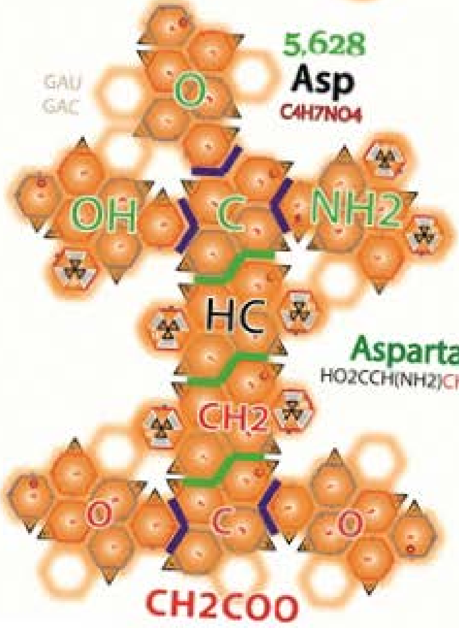
GAU
GAC
5,628
Asp
C4H7NO4
O
OH
C
NH2
HC
CH2
O
C
O
CH2COO

Aspartate
HO2CCH(NH2)CH2COO

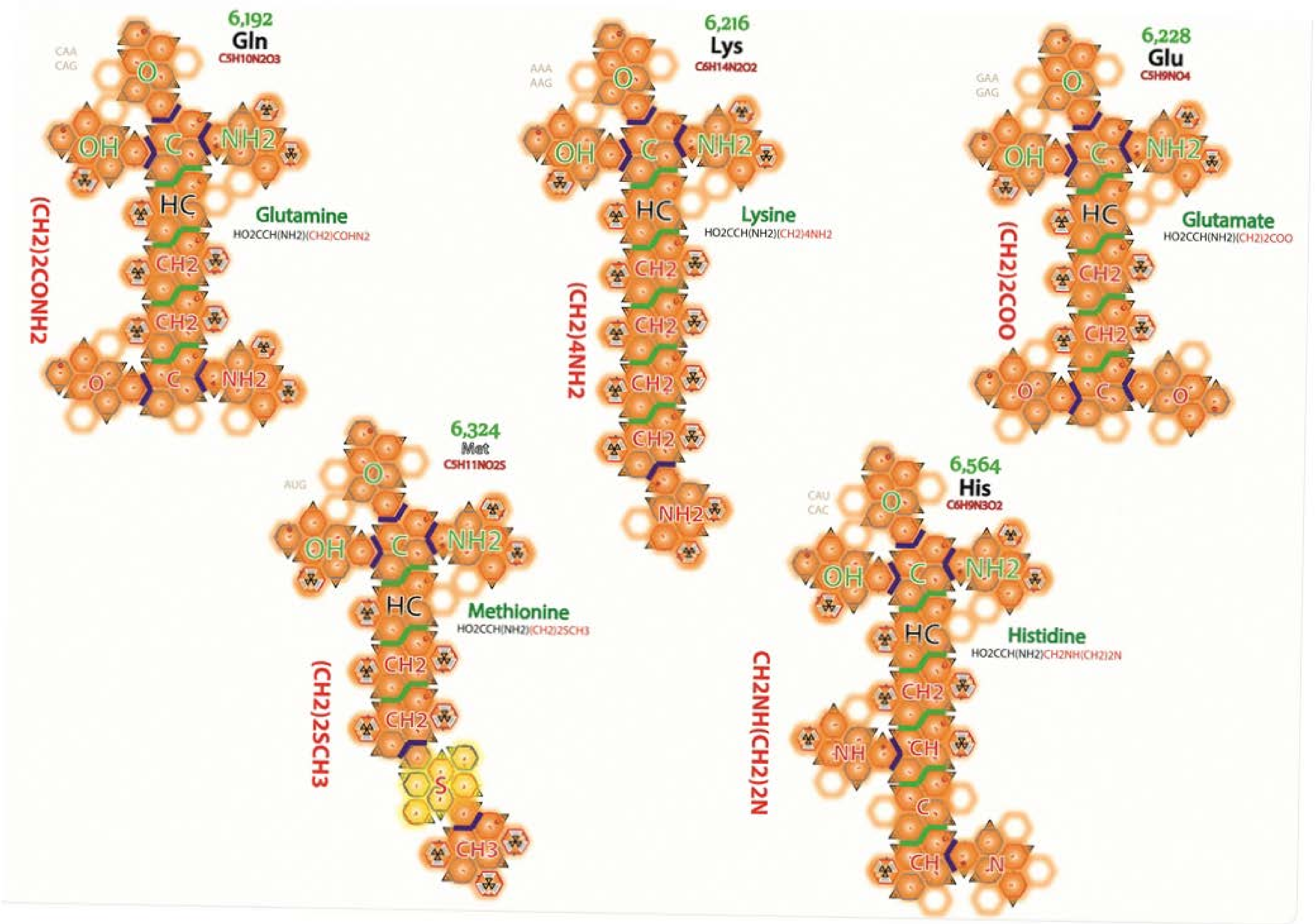
6,192
Gln
C5H10N2O3
CAA
CAG
O
OH
C
NH2
HC
Glutamine
HO2CCH(NH2)(CH2)COHN2
CH2
CH2
O
C
NH2
(CH2)2CONH2
6,216
Lys
C6H14N2O2
AAA
AAG
O
OH
C
NH2
HC
Lysine
HO2CCH(NH2)(CH2)4NH2
CH2
CH2
CH2
CH2
NH2
(CH2)4NH2
6,228
Glu
C5H9NO4
GAA
GAG
O
OH
C
NH2
HC
Glutamate
HO2CCH(NH2)(CH2)2COO
CH2
CH2
O
C
O
(CH2)2COO
6,324
Met
C5H11NO2S
AUG
O
OH
C
NH2
HC
Methionine
HO2CCH(NH2)(CH2)2SCH3
CH2
CH2
S
CH3
(CH2)2SCH3
6,564
His
C6H9N3O2
CAU
CAC
O
OH
C
NH2
HC
Histidine
HO2CCH(NH2)CH2NH(CH2)2N
CH2
NH
CH
C
CH
N
CH2NH(CH2)2N

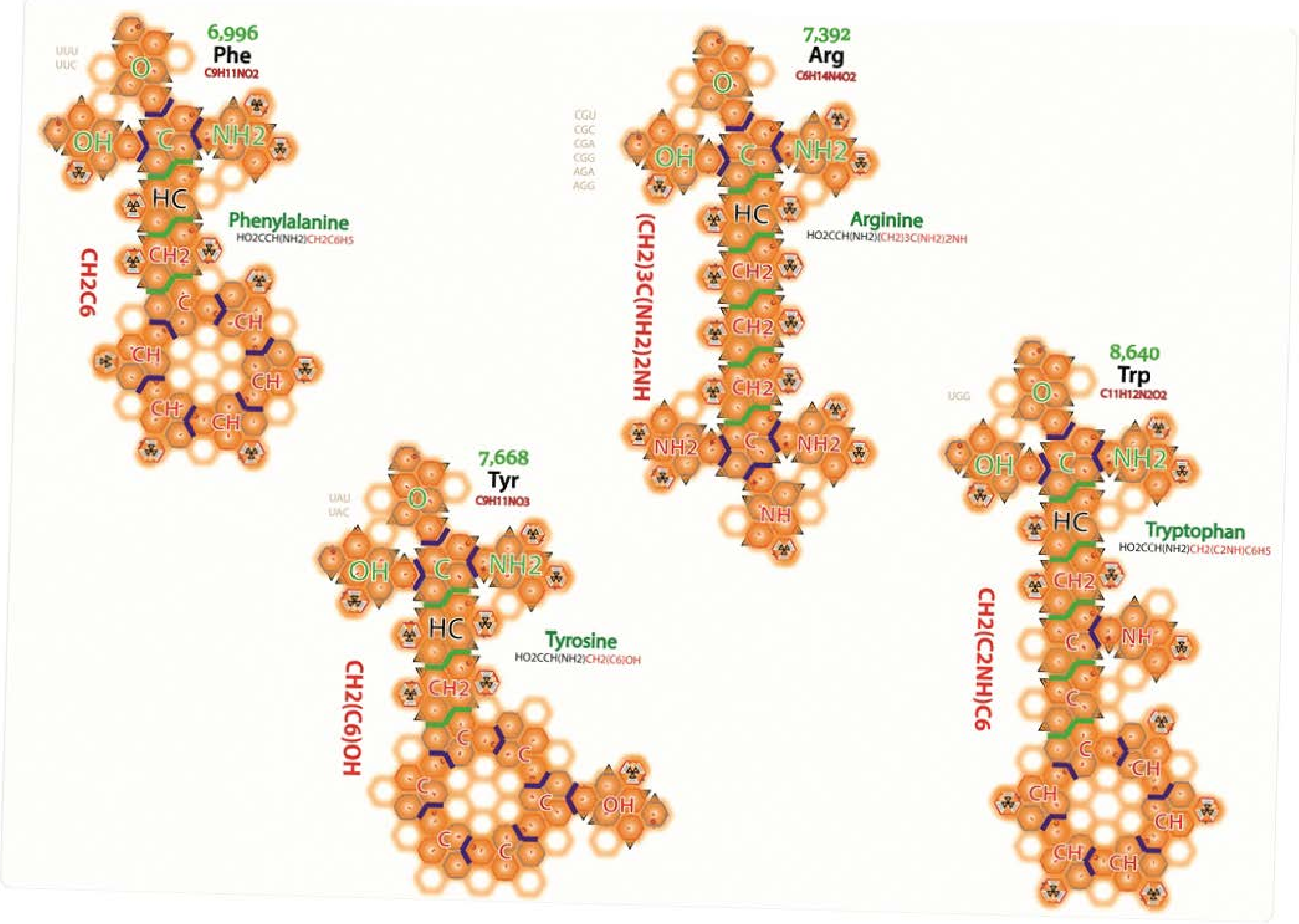
UUU
UUC
6,996
Phe
C9H11NO2
O
OH
C
NH2
HC
Phenylalanine
HO2CCH(NH2)CH2C6H5
CH2
CH2C6
C
CH
CH
CH
CH
CH
UAU
UAC
7,668
Tyr
C9H11NO3
O
OH
C
NH2
HC
Tyrosine
HO2CCH(NH2)CH2(C6)OH
CH2
CH2(C6)OH
C
C
C
C
OH
C
C
CGU
CGC
CGA
CGG
AGA
AGG
7,392
Arg
C6H14N4O2
O
OH
C
NH2
HC
Arginine
HO2CCH(NH2)(CH2)3C(NH2)2NH
(CH2)3C(NH2)2NH
CH2
CH2
CH2
NH2
C
NH2
NH
UGG
8,640
Trp
C11H12N2O2
O
OH
C
NH2
HC
Tryptophan
HO2CCH(NH2)CH2(C2NH)C6H5
CH2
CH2(C2NH)C6
C
NH
C
C
CH
CH
CH
CH
CH

Copyright ABRAHAM [2008] - All rights reserved

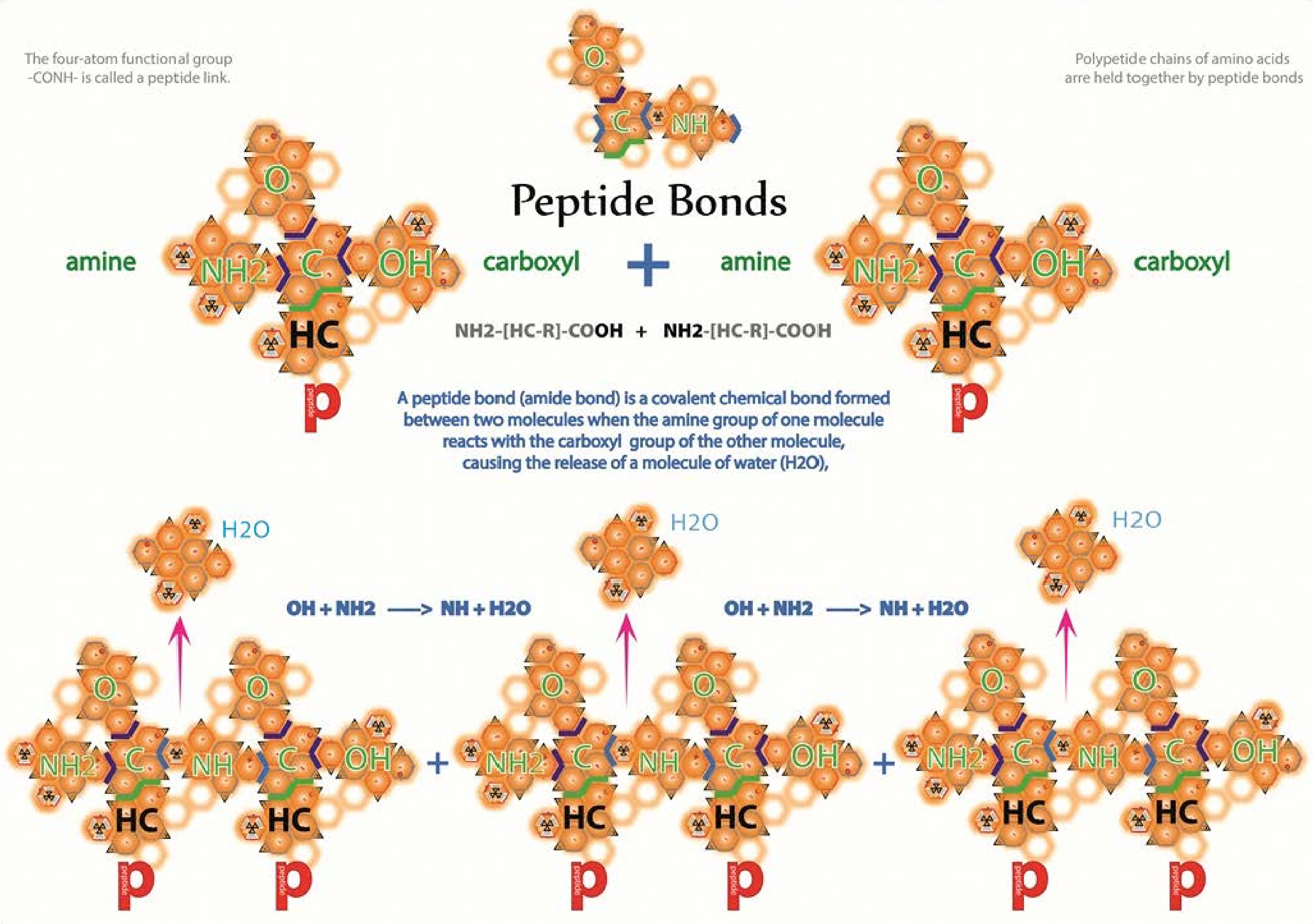

O
C
NH
O
NH2
C
OH
HC
p
peptide
amine
carboxyl
+
amine
carboxyl
H2O
H2O
H2O
NH2
C
NH
C
OH
HC
HC
+
NH2
C
NH
C
OH
HC
HC
+
NH2
C
NH
C
OH
HC
HC

The four-atom functional group
-CONH- is called a peptide link.

Peptide Bonds

Polypeptide chains of amino acids
arre held together by peptide bonds

NH2-[HC-R]-COOH + NH2-[HC-R]-COOH

A peptide bond (amide bond) is a covalent chemical bond formed between two molecules when the amine group of one molecule reacts with the carboxyl group of the other molecule, causing the release of a molecule of water (H2O),

OH + NH2 ——> NH + H2O

OH + NH2 ——> NH + H2O

Tetryonics 59.15 - Peptide bonds

Copyright ABRAHAM [2008] - All rights reserved

amino terminus

The beginning of a polypeptide protein molecule has a free amino group.

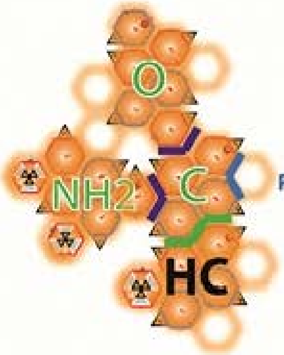

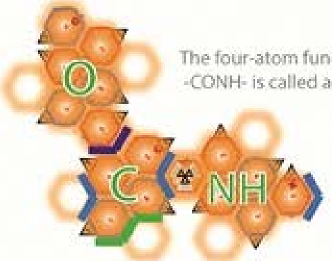

The four-atom functional group -CONH- is called a peptide link.

carboxy terminus

The end of a polypeptide protein molecule has a free carboxyl group.

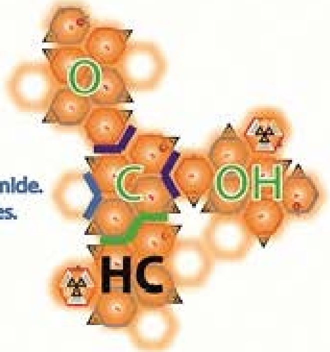

Proteins [poly-peptides] are formed by joining the -CO2H [carboxyl] end of one amino acid with the -NH2 [amine] end of another to form an amide. The -CONH- bond between amino acids is known as a peptide bond because relatively short polymers of amino acids are known as peptides.

polypetide chains

Proteins are formed when amino acids are covalently linked together.

DNA and RNA have a deoxyribose and ribose sugar backbone, respectively, whereas PNA's backbone is composed of repeating (2-aminoethyl)-glycine units linked by peptide bonds.

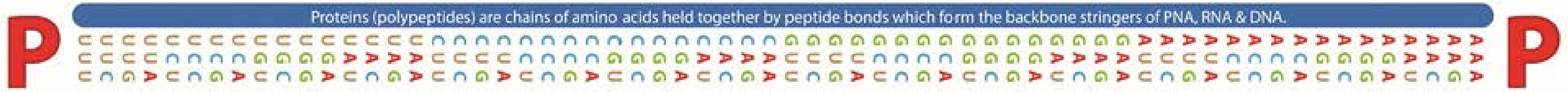

Proteins (polypeptides) are chains of amino acids held together by peptide bonds which form the backbone stringers of PNA, RNA & DNA.

Proteins

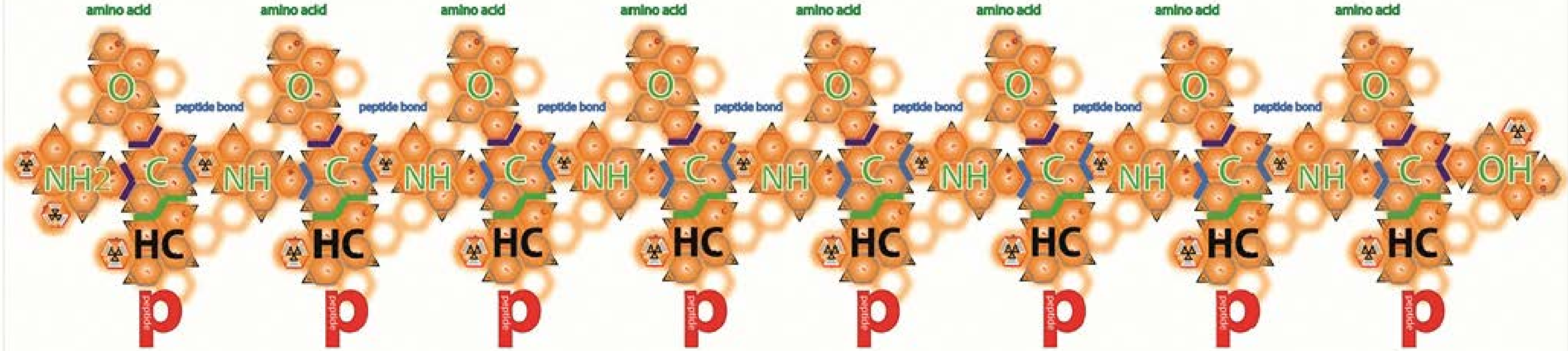

amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
amino acid
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
peptide bond
NH2
O
C
NH
HC
OH
p
peptide

There are 20 different nucleotide side chains present in biological molecules.

Copyright ABRAHAM [2008] - All rights reserved

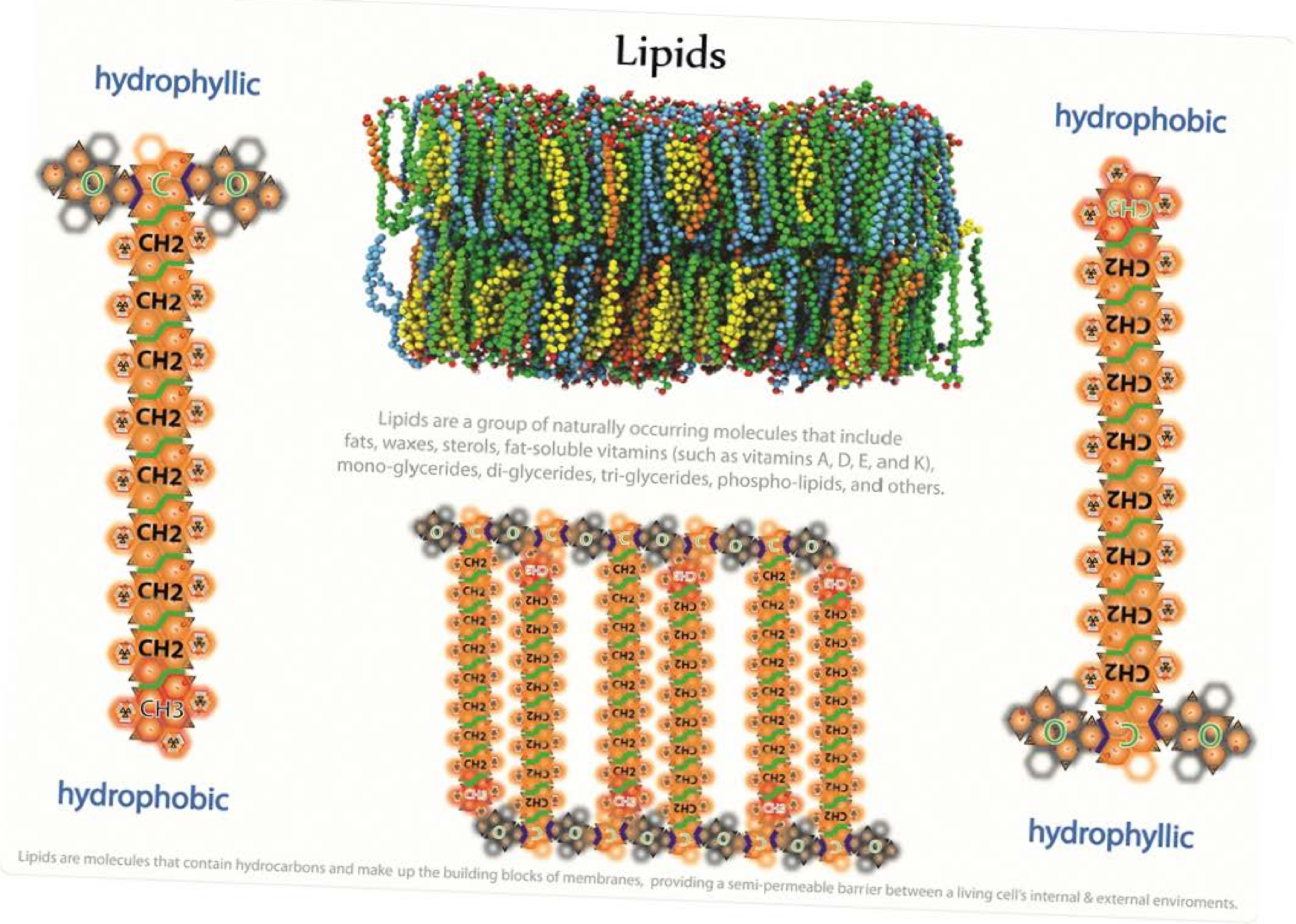

Lipids

hydrophyllic
O
C
O
CH2
CH3
hydrophobic
hydrophobic
hydrophyllic

Lipids are a group of naturally occurring molecules that include fats, waxes, sterols, fat-soluble vitamins (such as vitamins A, D, E, and K), mono-glycerides, di-glycerides, tri-glycerides, phospho-lipids, and others.

Lipids are molecules that contain hydrocarbons and make up the building blocks of membranes, providing a semi-permeable barrier between a living cell's internal & external enviroments.

Tetryonics 59.17 - Lipids

Acetylcholine

is an organic molecule that acts as a neurotransmitter in many organisms, including humans. It is an ester of acetic acid and choline

Ach

C7H16NO2

CH3COO(CH2)2N+(CH3)3

Lewis chemical structure

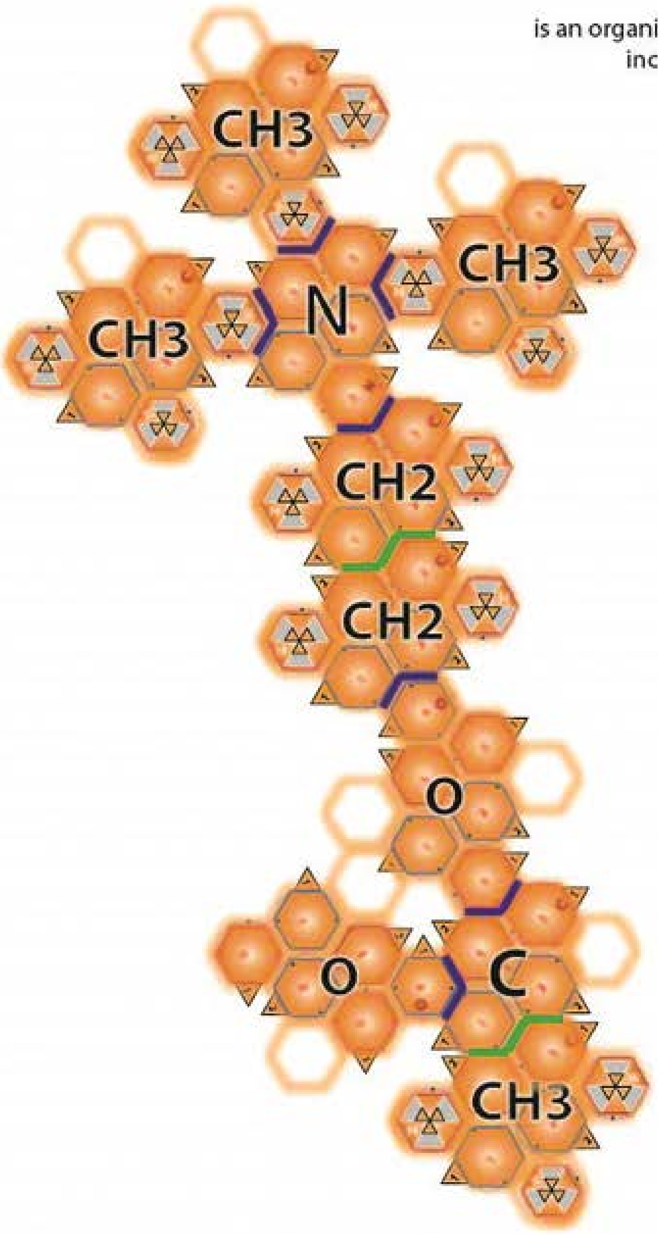
CH3
CH3
N
CH3
CH2
CH2
O
O
C
CH3

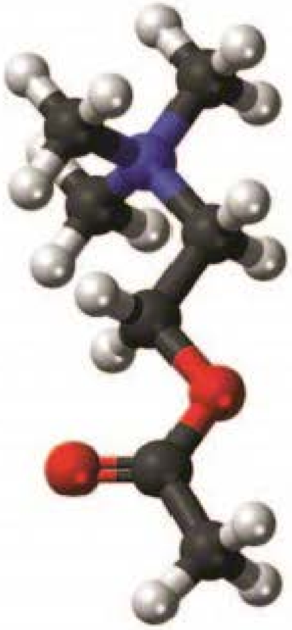

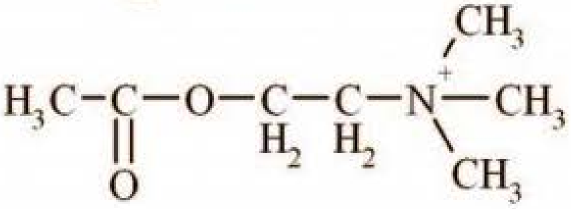

$$H_3C-\underset{\underset{O}{\|}}{C}-O-\underset{H_2}{C}-\underset{H_2}{C}-\overset{CH_3}{\underset{CH_3}{\overset{+}{N}}}-CH_3$$

Tetryonic charge topology

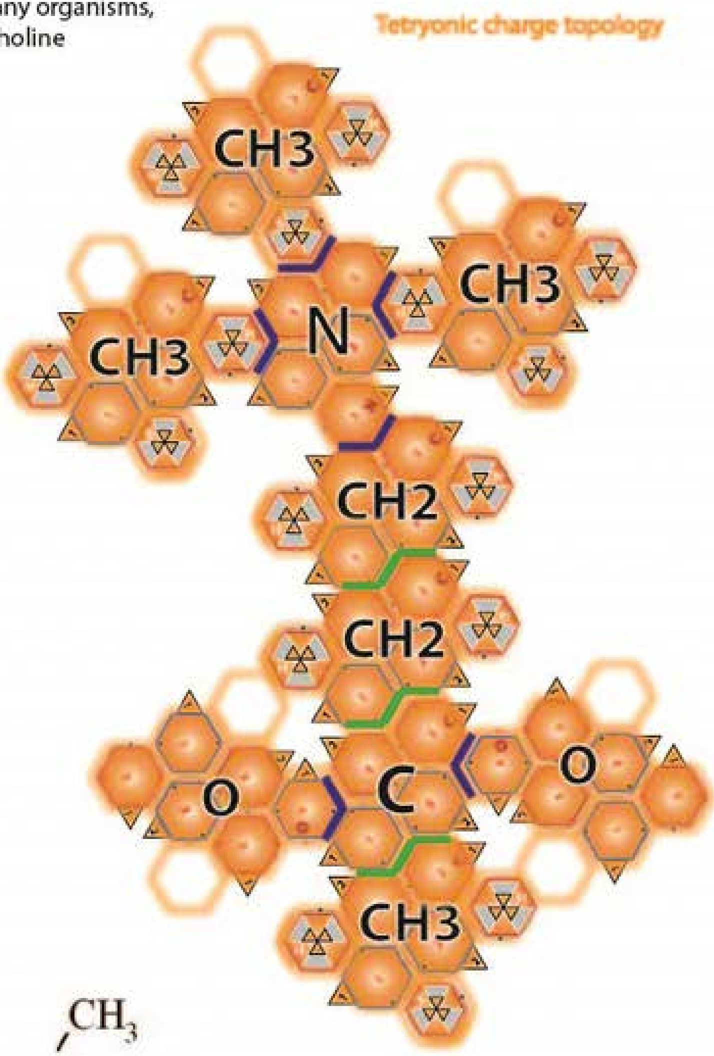
CH3
CH3
N
CH3
CH2
CH2
O
C
O
CH3

6,228

Tetryonics 59.18 - Acetylcholine [Ach]

Copyright ABRAHAM [2008] - All rights reserved

Ribose

6,360

[3,180-3,180]

C5-H10-O5

Carbohydrate compounds can form acyclic sugars

Carbohydrate compounds can also form cyclic sugars

Ribose is an organic compound with the formula C5H10O5; specifically, a monosaccharide (simple sugar) with linear form H–(C=O)–(CHOH)4–H

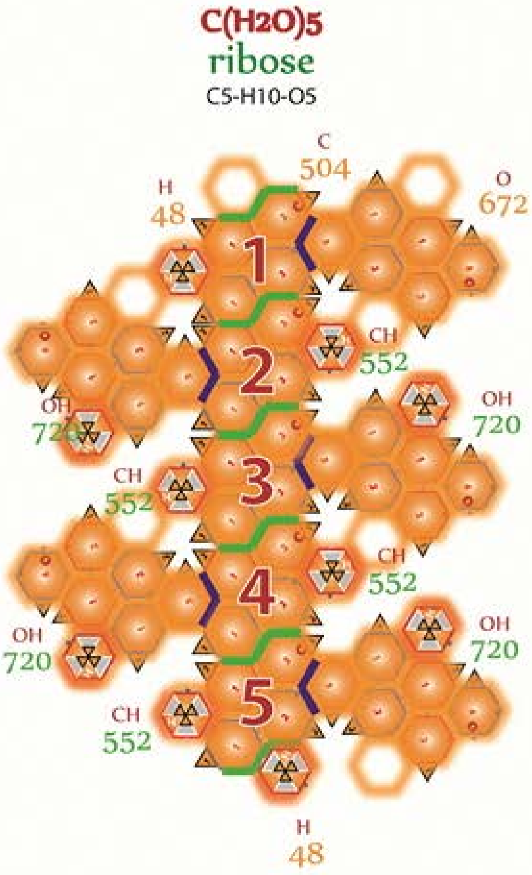
C(H2O)5
ribose
C5-H10-O5
H
48
C
504
O
672
1
CH
552
2
OH
720
OH
720
3
CH
552
CH
552
4
OH
720
OH
720
5
CH
552
H
48

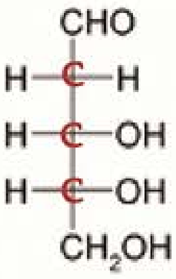
CHO
H
C
H
H
C
OH
H
C
OH
CH2OH

H-(CO)-(CHOH)4-H

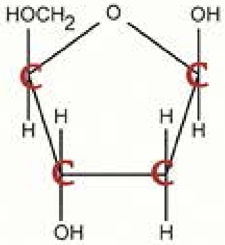
HOCH2
O
OH
C
C
H
H
H
H
C
C
OH
H

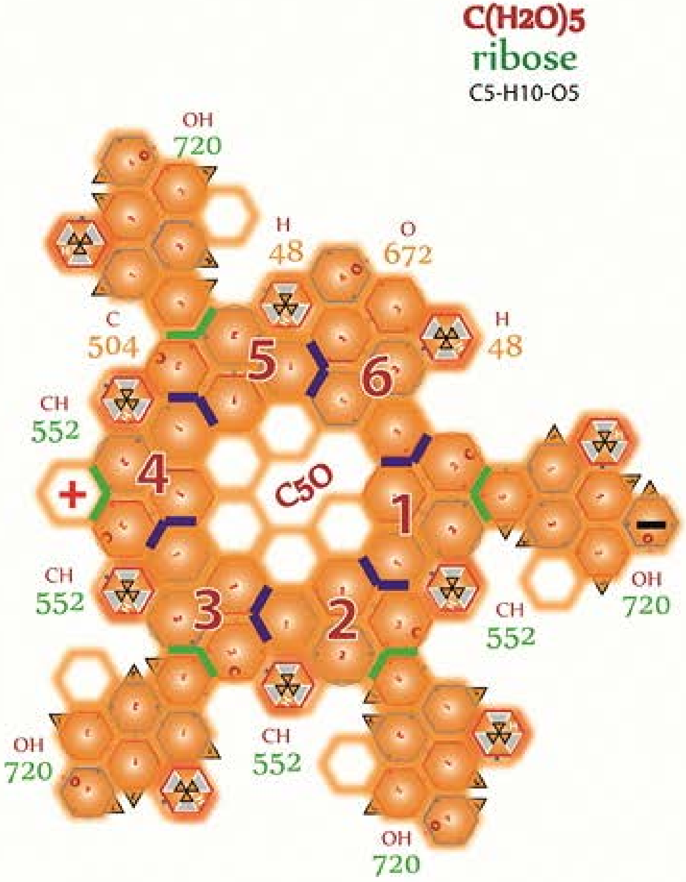
C(H2O)5
ribose
C5-H10-O5
OH
720
H
48
O
672
C
504
H
48
5
6
CH
552
4
+
C5O
1
−
CH
552
3
2
OH
720
CH
552
OH
720
CH
552
OH
720

Ribose forms part of the backbone of RNA.
It is related to deoxyribose, which is found in DNA.

**Ribose is present within every living cell of the body and is used to manufacture ATP (the energy currency of the cell) from scratch.
Whilst the body can manufacture its own ribose from glucose, this requires energy and is a very slow process.**

As its name indicates it is a deoxygenated sugar, meaning that it is derived from the sugar ribose by loss of an oxygen atom

Deoxyribose

5,688

[2,844-2,844]

As a component of DNA, 2-deoxyribose derivatives have an important role in biology

The DNA (deoxyribonucleic acid) molecule, which is the main repository of genetic information in life, consists of a long chain of deoxyribose-containing units called nucleotides, linked via phosphate groups

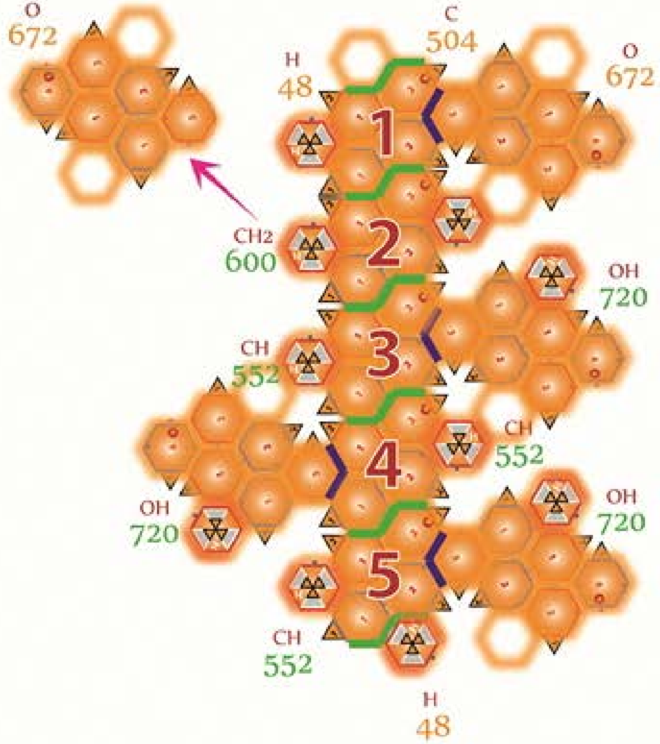
O
672
H
48
C
504
O
672
1
2
CH2
600
OH
720
3
CH
552
4
CH
552
OH
720
OH
720
5
CH
552
H
48

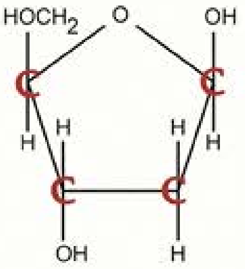
HOCH2
O
OH
C
C
H
H
H
H
C
C
OH
H

Deoxyribose, more, precisely 2-deoxyribose, is a monosaccharide

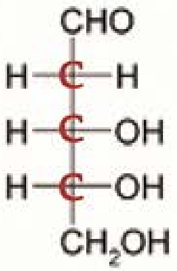
CHO
H
C
H
H
C
OH
H
C
OH
CH2OH

C5-H10-O4
deoxyribose

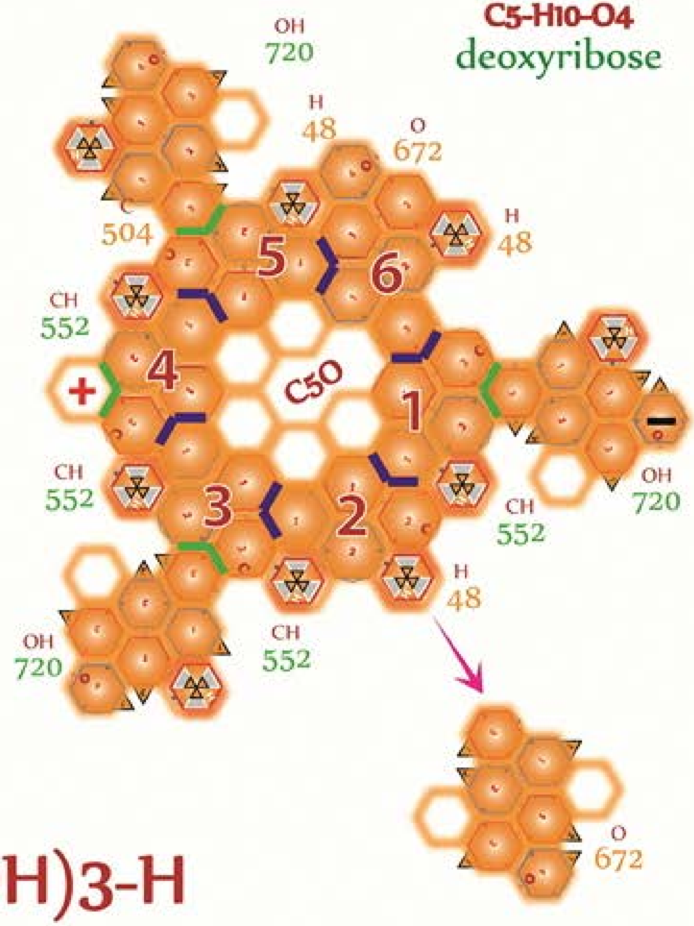
OH
720
H
48
O
672
C
504
H
48
5
6
CH
552
4
+
C5O
1
−
CH
552
3
2
CH
552
OH
720
H
48
OH
720
CH
552
O
672

C5-H10-O4
deoxyribose

H-(CO)-(CH2)-(CHOH)3-H

Hermann Emil Fischer won the Nobel Prize in Chemistry (1902) for his work in determining the structure of the D-aldohexoses.

However, the linear, free-aldehyde structures that Fischer proposed represent a very minor percentage of the forms that hexose sugars adopt in solution.
It was Edmund Hirst and Clifford Purves, in the research group of Walter Haworth, who conclusively determined that the hexose sugars preferentially form a pyranose, or six-membered, ring. Haworth drew the ring as a flat hexagon with groups above and below the plane of the ring – the Haworth projection

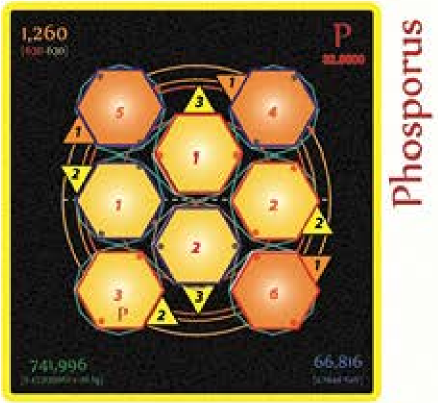
1,260
P
741,996
66,816
Phosporus

3,996

[1,998-1,998]

HPO4

phosphate group

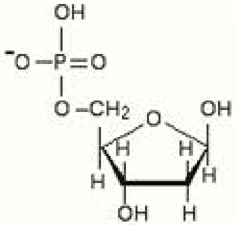
OH
O
P
O
O
CH2
O
OH
H
H
H
H
OH
H

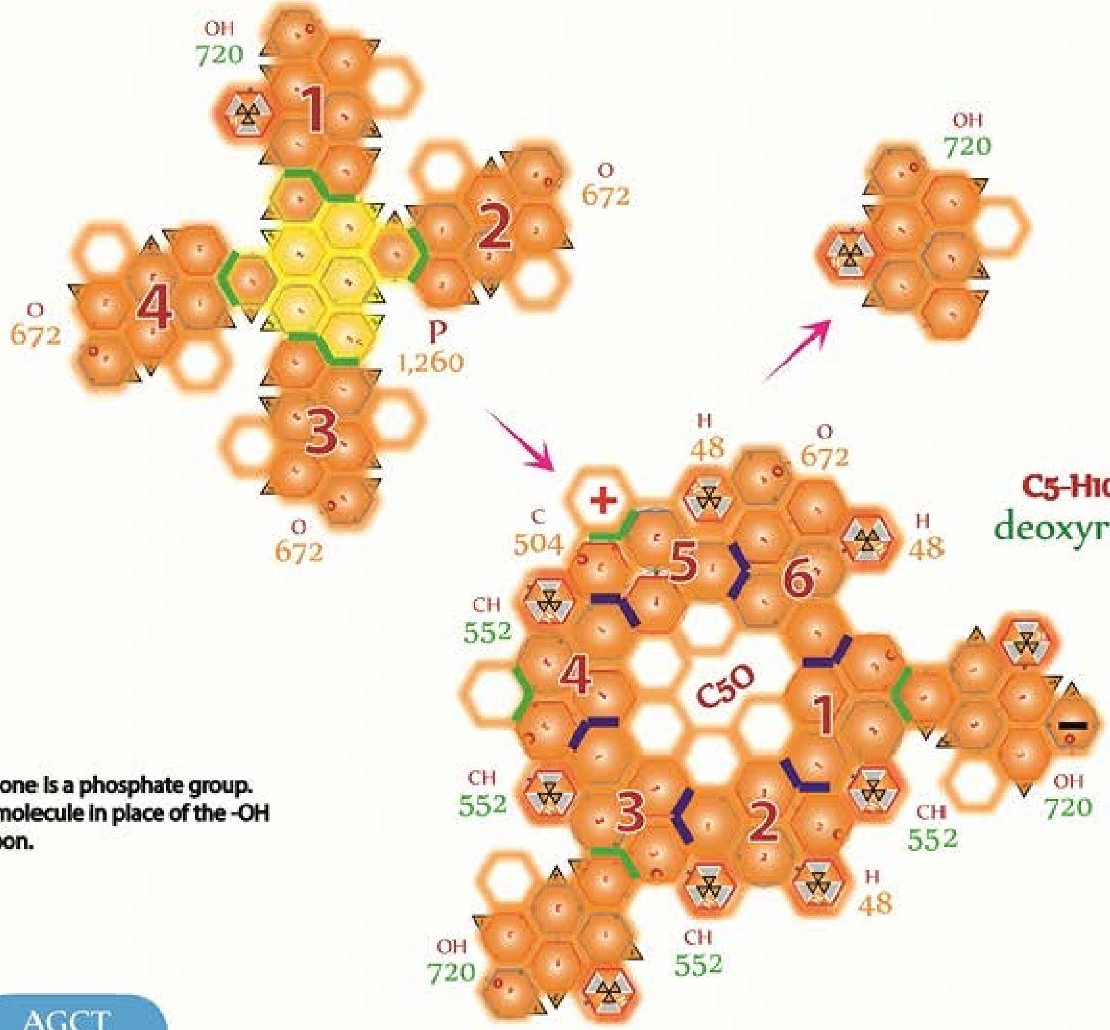
OH
720
1
O
672
2
O
672
4
P
1,260
3
O
672
OH
720
H
48
O
672
+
C
504
5
6
H
48
CH
552
4
C5O
1
−
CH
552
3
2
OH
720
CH
552
H
48
OH
720
CH
552

C5-H10-O4

deoxyribose

The other repeating part of the DNA backbone is a phosphate group. A phosphate group is attached to the sugar molecule in place of the -OH group on the 5' carbon.

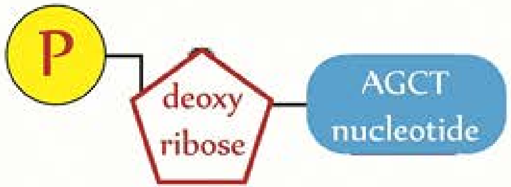
P
deoxy ribose
AGCT nucleotide

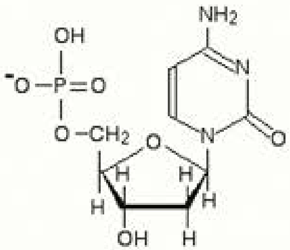
NH2
N
N
O
OH
O
P
O
O
CH2
O
H
H
H
H
OH
H

5,688

[2,844-2,844]

Copyright ABRAHAM [2008] - All rights reserved

Radioactive Decay

The Atom releases excess energies in various forms as it seeks a lower energy, state of equilibria with its surrounding EM energy environment

Energy in all its forms seeks equilibirium

Photons, Heat, Vibration & Kinetics can all raise the energy levels of atomic nuclei

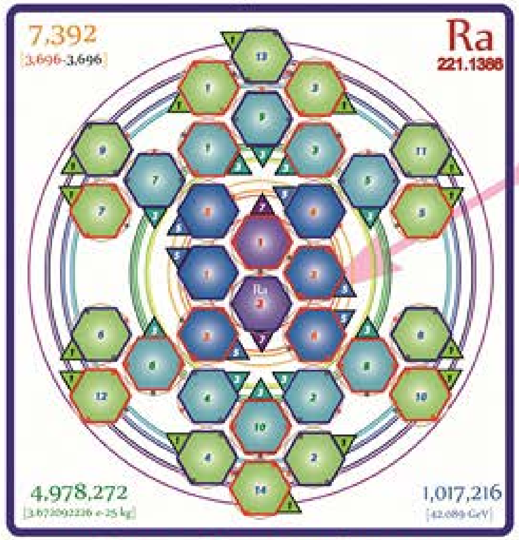
7,392
[3,696-3,696]
Ra
221.1388
4,978,272
[3.673092226 e-25 kg]
1,017,216
[42.089 GeV]

stable core electron configurations

The spectral line emissions of photons from bound electrons in addition to Alpha, Beta and Gamma Particle emissions from nuclei can release KEM energies over long time spans

E

scalar equilateral

mass-ENERGY-Matter

geometries

topologies

mc^2

Mc^4

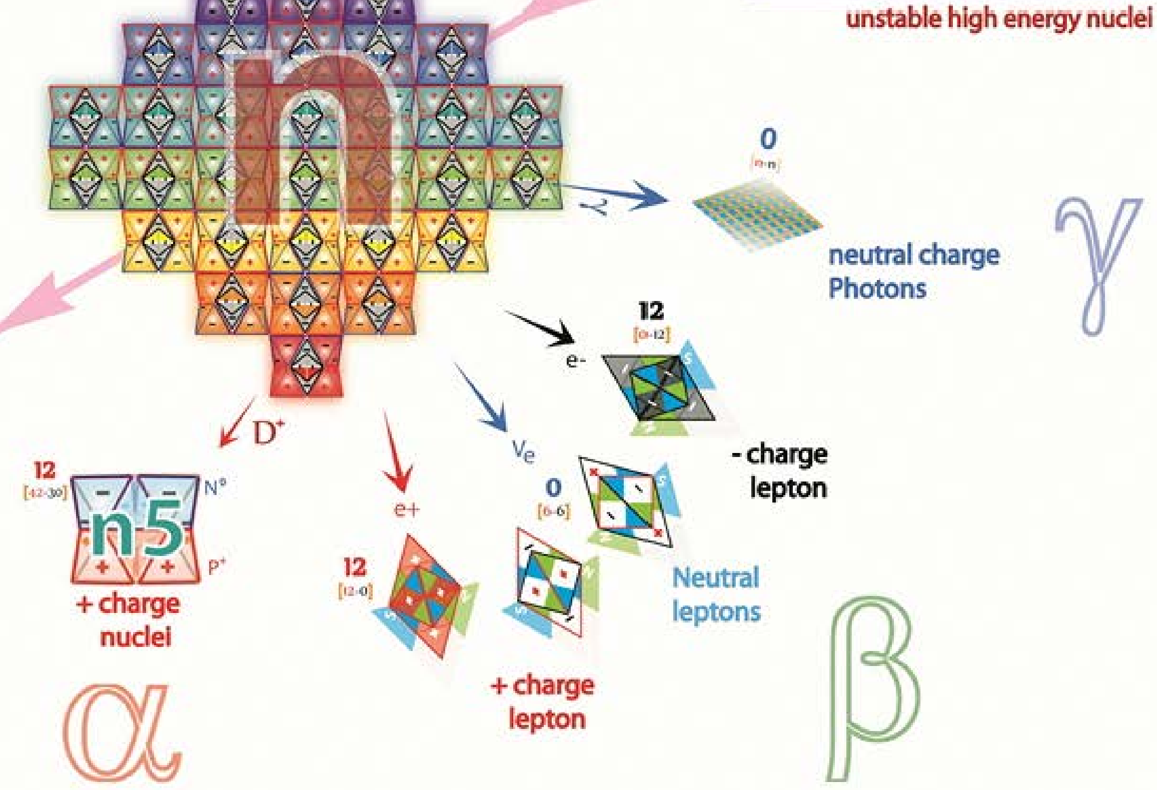
n
0
[n-n]
neutral charge Photons
γ
12
[0-12]
e-
- charge lepton
D+
Ve
12
[12-30]
N°
n5
P+
+ charge nuclei
α
e+
12
[12-0]
0
[6-6]
+ charge lepton
Neutral leptons
β

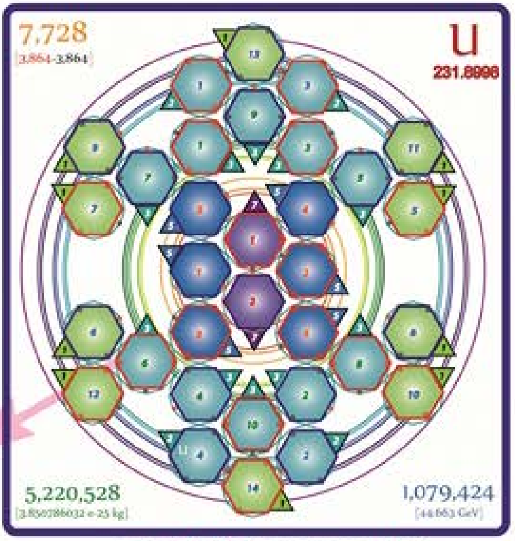
7,728
[3,864-3,864]
U
231.8898
5,220,528
[3.850786032 e-25 kg]
1,079,424
[44.663 GeV]

unstable high energy nuclei

Radio-Carbon isotope dating

Radiocarbon dating (sometimes simply known as carbon dating) is a radiometric dating method that uses the naturally occurring radioisotope carbon-14 (14C) to estimate the age of carbon-bearing materials up to about 58,000 to 62,000 years

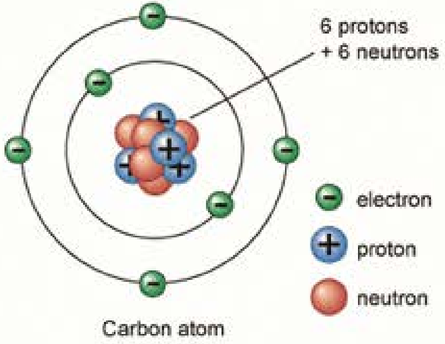
6 protons
+ 6 neutrons
electron
proton
neutron
Carbon atom

Standing-wave KE not 'excess' Neutrons

It is the stored Kinetic Energies [KE] that increase the mass of Carbon12 to create its isotopes, historically these mass-energies have been mistakenly thought to be extra Neutrons in the nuclei

C14 half-life
5730 ± 40

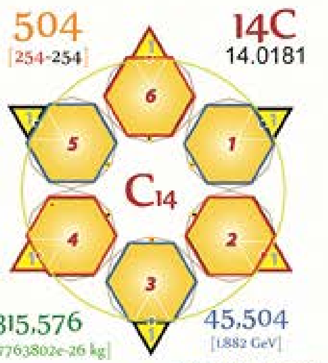
504
[254-254]
14C
14.0181
6
5
1
4
2
3
1
C_{14}
315,576
[2.3277763802e-26 kg]
45,504
[1.882 GeV]

[2.0224 Neutrons]

Carbon-14, 14C, or radiocarbon, is a radioactive isotope of carbon with a nucleus containing 6 protons, 6 neutrons and 6 electrons with an excess energy content of 1.88 GeV

[6 Deuterium nuclei]

ALL elemental Carbon atoms have an atomic configuration of 6 protons, 6 neutrons and 6 electrons

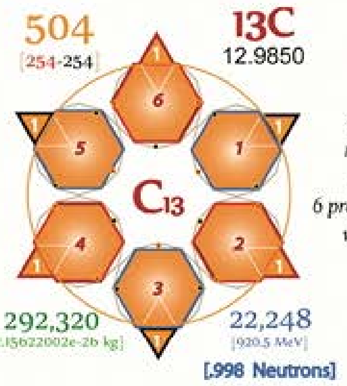
504
[254-254]
13C
12.9850
6
5
1
4
2
3
1
C_{13}
292,320
[2.15622002e-26 kg]
22,248
[920.5 MeV]

[.998 Neutrons]

Carbon-13, 13C, is a weakly radioactive isotope of carbon with a nucleus containing 6 protons, 6 neutrons and 6 electrons with an excess energy content of 920 MeV

Radio-Carbon dating can be used to determine the age of carbonaceous materials up to about 60,000 years old

Isotopic mass-charge quotient determines an element's properties

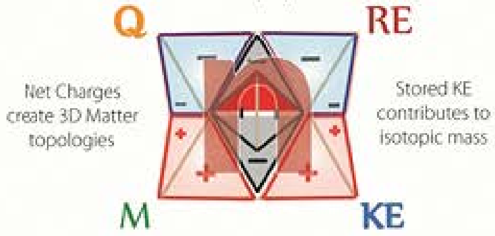

Net Charges create 3D Matter topologies

Stored KE contributes to isotopic mass

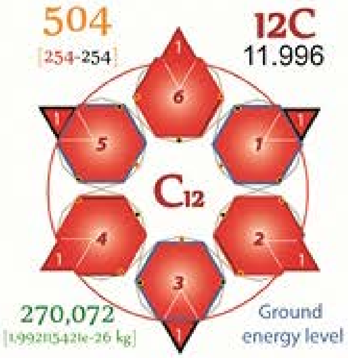
504
[254-254]
12C
11.996
6
5
1
4
2
3
1
C_{12}
270,072
[1.99211542le-26 kg]
Ground
energy level

Tetryonic theory corrects historical errors of atomic geometries and nuclear energies

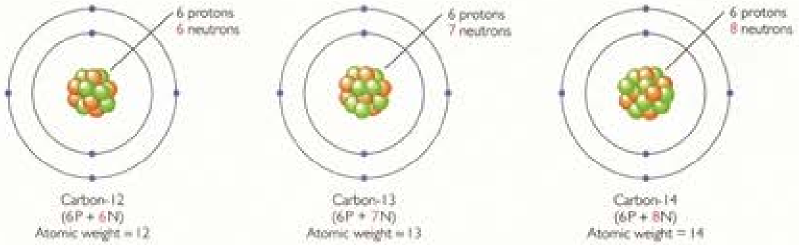
6 protons
6 neutrons
Carbon-12
(6P + 6N)
Atomic weight = 12
6 protons
7 neutrons
Carbon-13
(6P + 7N)
Atomic weight = 13
6 protons
8 neutrons
Carbon-14
(6P + 8N)
Atomic weight = 14

The stored KE [chemical] mass-energies of the Carbon 12+ isotopes comes from various radiological [and biological] processes and its steady, predictable release through the quantum mechanics of the synchronous convertors that make up Carbon nuclei make it useful in dating objects based on their decay rates

Radioactive Isotopes

Where the elementary nuclei's Deuteron energy levels are raised from their ground levels radioactive isotopes are created

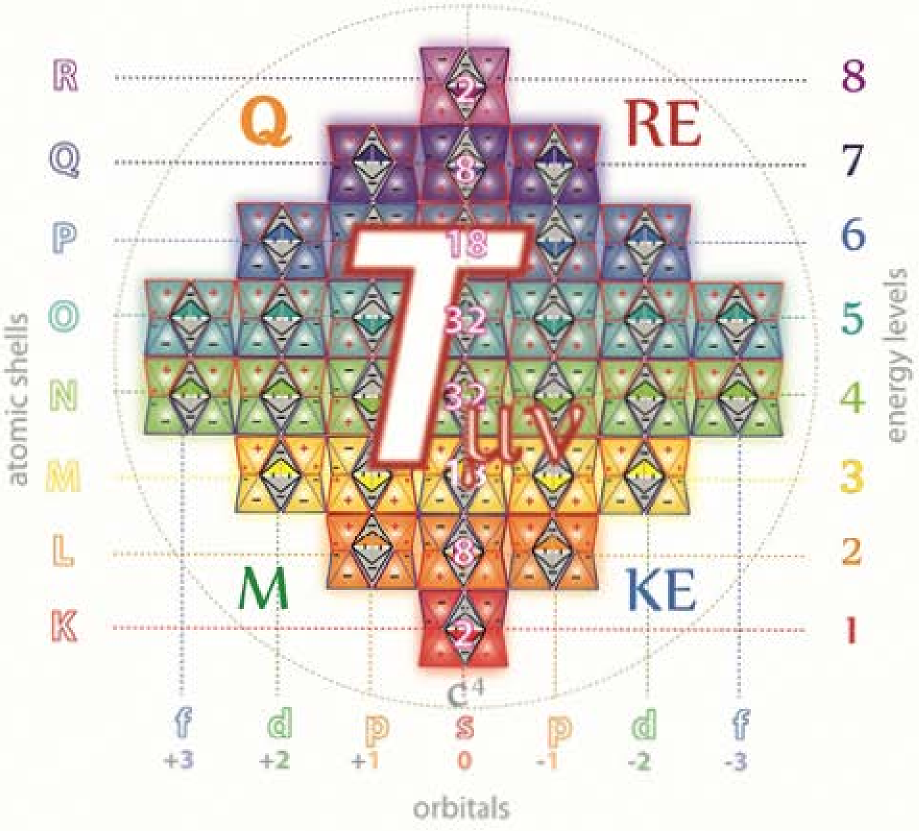
R
Q
P
O
N
M
L
K
atomic shells
8
7
6
5
4
3
2
1
energy levels
Q
RE
M
KE
2
8
18
32
32
18
8
2
$T_{\mu\nu}$
c^4
f
d
p
s
p
d
f
+3
+2
+1
0
-1
-2
-3
orbitals

Each periodic element is comprised of an EQUAL number of Protons, Neutrons & electrons that form each element's unique 2D mass-energy geometries & 3D Matter topology and contribute to its observed properties

The $\frac{\text{mass}}{c^2}$ ~ $\frac{\text{energy}}{c^4}$ content of Matter

Einstein's relativistic stress energy tensor models mass-energy-Matter as a nebulous energy-momenta density-pressure gradient

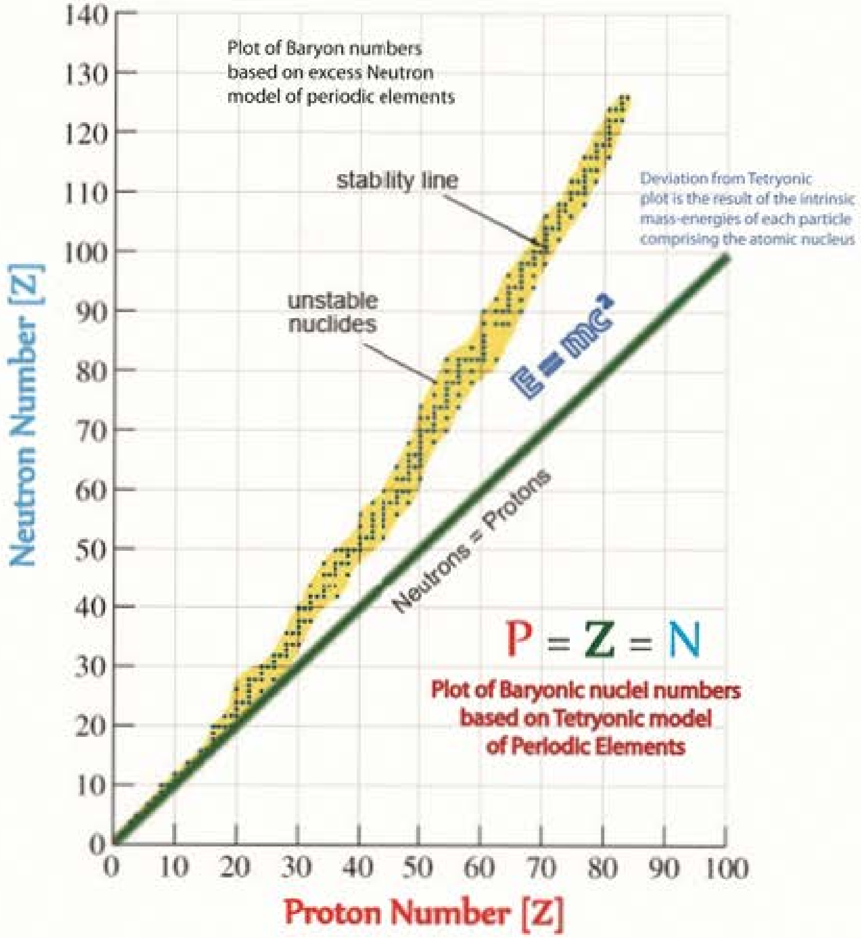

Tetryonic theory redefines the relativistic stress energy tensor [$T_{\mu\nu}$] into a geometric measure of the charged 2D electromagnetic mass-energies & 3D Matter topologies within any spatial co-ordinate system

Copyright ABRAHAM [2008] - All rights reserved

Nuclear Decay processes

is the set of processes by which an unstable atomic nucleus emits subatomic particles

Charge & mass-energy momenta are conservative physical properties

Radioactivity was discovered in 1896 by the French scientist Henri Becquerel, while working on phosphorescent materials. These materials glow in the dark after exposure to light, and he suspected that the glow produced in cathode ray tubes by X-rays might be associated with phosphorescence

Charged Matter equivalence

Neutrogen $\underset{[6\text{-}6]}{ve} + \underset{[18\text{-}18]}{N^0} \overset{[24\text{-}24]}{\sim} \underset{[24\text{-}12]}{P^+} + \underset{[0\text{-}12]}{e^-}$ **Hydrogen**

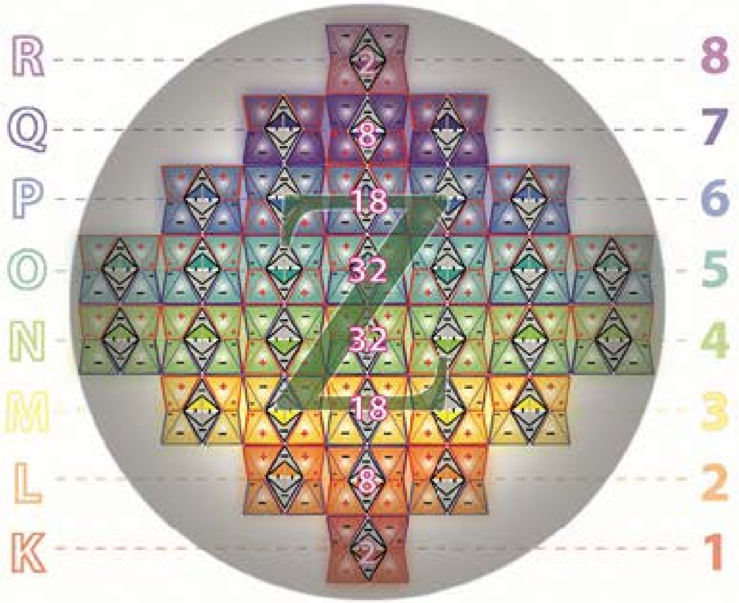

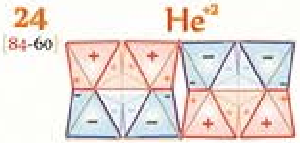

alpha-particles

$$\begin{matrix} D^+ \\ [42\text{-}30] \\ [42\text{-}30] \\ D^+ \end{matrix} \longrightarrow \underset{[84\text{-}60]}{He^{+2}}$$

alpha particle production

E

Energy can be released from atomic nuclei through various processes

alpha-particles
beta-particles
gamma rays
spectral lines
heat & motion

positrons
beta-particles
neutrinos
electrons

Charge Matter particle creation follows chemical equilibrium formulae

$$\underset{[18\text{-}18]}{N^0} \longrightarrow \underset{[12\text{-}0]}{e^+} + \underset{[6\text{-}6]}{ve} + \underset{[0\text{-}12]}{e^-}$$

Neutron decay processes

$$\underset{[18\text{-}18]}{N^0} \longrightarrow \underset{[6\text{-}6]}{ve} + \underset{[6\text{-}6]}{ve} + \underset{[6\text{-}6]}{ve}$$

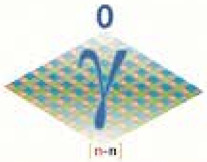

gamma rays

$$\underset{[v\text{-}v]}{KE} \longrightarrow \underset{[n\text{-}n]}{\gamma}$$

gamma ray production

m *All nuclear decay particles are determined by Tetryonic charge topologies* M

3D Matter topologies are not conservative

Alpha radiation

(Deuteron emissions)

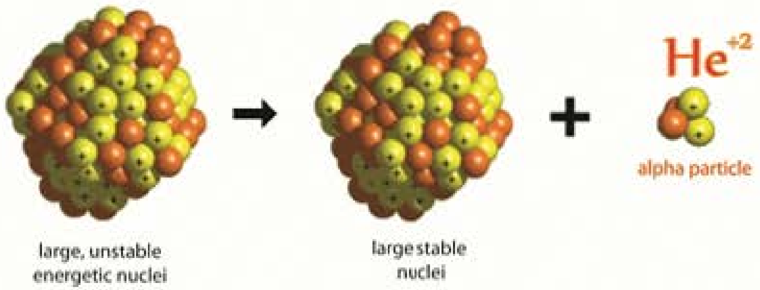
He+2
alpha particle
large, unstable energetic nuclei
large stable nuclei

Alpha particles are released during alpha decay processes in ultra-heavy nuclei like uranium, thorium, actinium, and radium

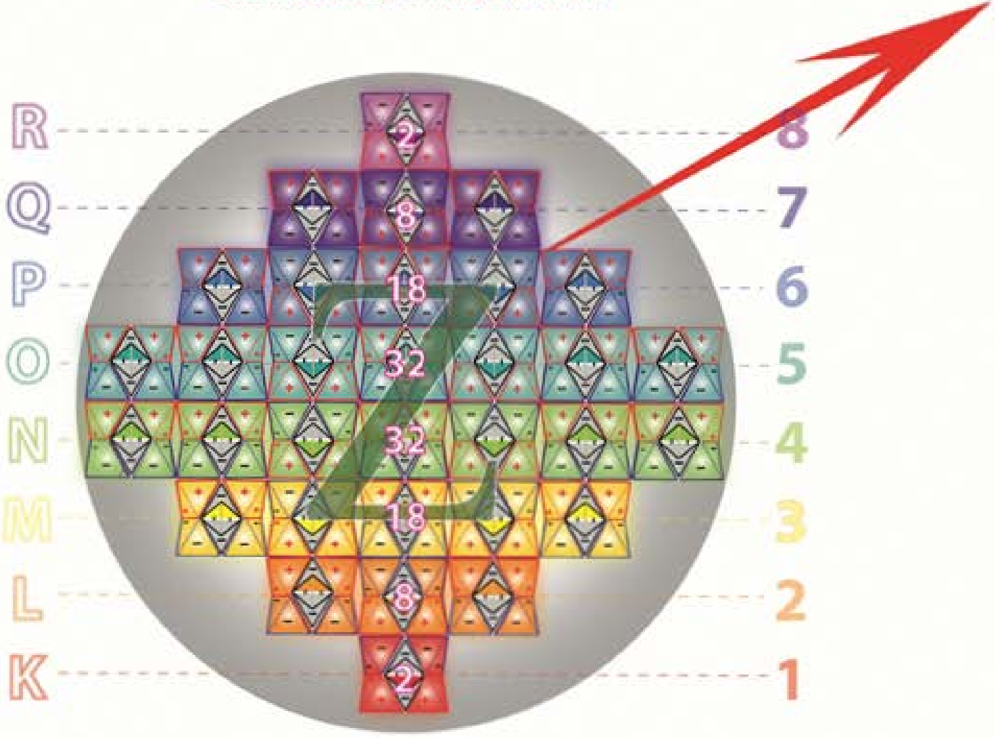
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z

Beta radiation - Negative charge sets (Nuclei)
Gamma radiation - Neutral Charge sets (Photons)

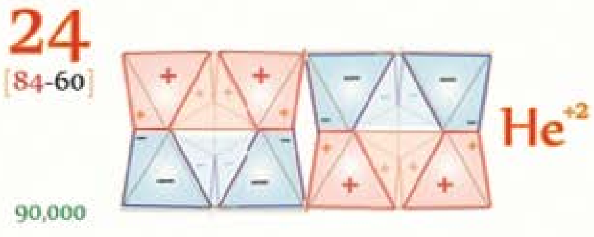
24
[84-60]
90,000
He+2

Helium nuclei creation

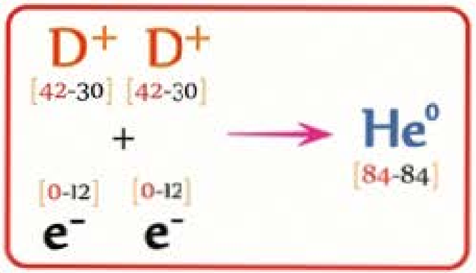
D+ D+
[42-30] [42-30]
+
[0-12] [0-12]
e- e-
He0
[84-84]

charged alpha nuclei bond & capture electrons to form neutral Helium

alpha particles

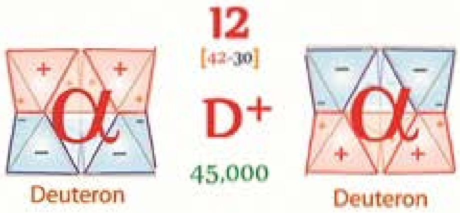
12
[42-30]
D+
45,000
Deuteron
Deuteron

The radioactive isotope Americium 241 emits alpha particles which are used in smoke detectors.

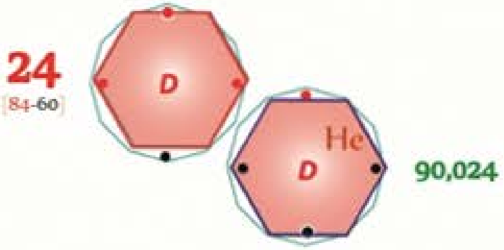
24
[84-60]
D
D
He
90,024

Due to their larger masses compared to beta and gamma particles, alpha particles can be easily stopped by a piece of paper or human skin.

Copyright ABRAHAM [2008] - All rights reserved

Beta particle decay

$$N^0_{[18\text{-}18]} \rightarrow e^+_{[12\text{-}0]} + \nu e_{[6\text{-}6]} + e^-_{[0\text{-}12]}$$

Neutron decay paths

$$N^0_{[18\text{-}18]} \rightarrow \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]}$$

(the emission of leptons from atomic nuclei)

beta-minus decay

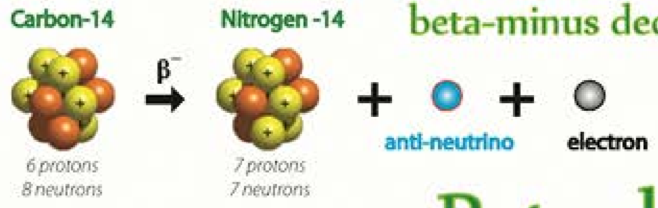
Carbon-14
6 protons
8 neutrons
β⁻
Nitrogen -14
7 protons
7 neutrons
+
anti-neutrino
+
electron

Beta decay

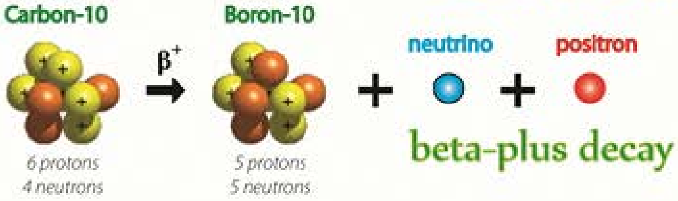
Carbon-10
6 protons
4 neutrons
β⁺
Boron-10
5 protons
5 neutrons
+
neutrino
+
positron

beta-plus decay

Charge is a conserved physical property of all material objects

$$N^0_{[18\text{-}18]} + \nu e_{[6\text{-}6]} \xrightarrow{[24\text{-}24]} P^+_{[24\text{-}12]} + e^-_{[0\text{-}12]}$$

Neutron decay is not a spontaneous process

All beta decay processes are the result of neutrino interaction with Neutrons

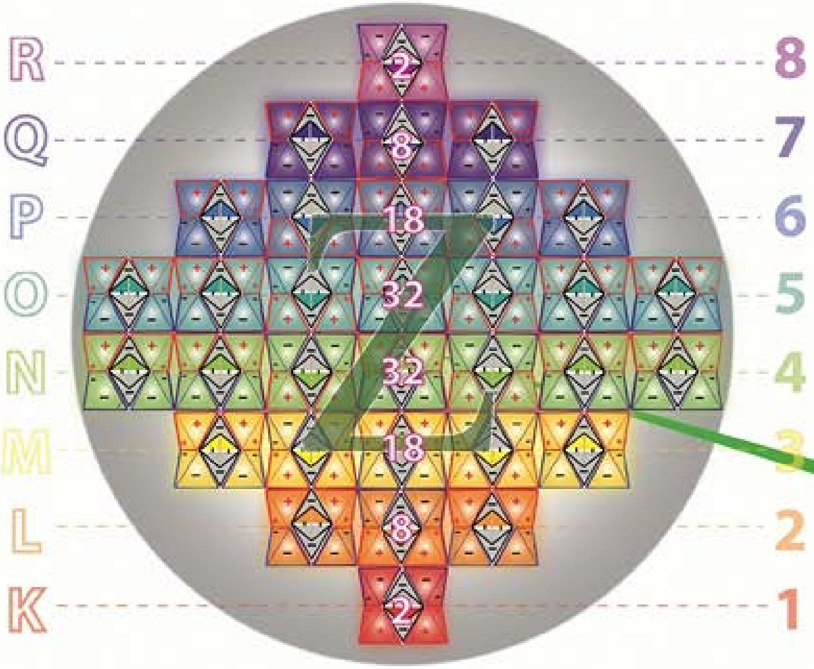
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z

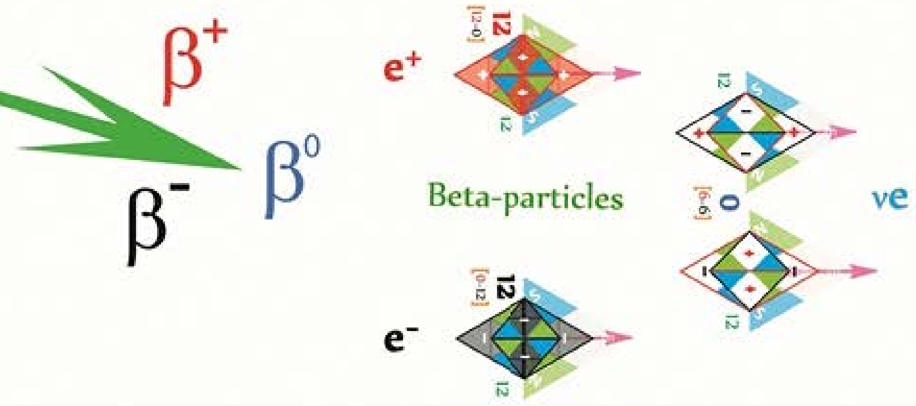
β⁺
β⁰
β⁻
e⁺
[12-0]
12
12
Beta-particles
12
0
[6-6]
12
νe
e⁻
[0-12]
12
12

Alpha radiation - Positive charge particles (Nuclei)
Gamma radiation - Neutral charge quanta (Photons)

Gamma Radiation

(high energy photon emission)

Natural sources of gamma rays on Earth include gamma decay from naturally occurring radioisotopes, and secondary radiation from atmospheric interactions with cosmic ray particles.

λ [f]

All ejected gamma ray photons are neutral energy quanta sets [photons / EM waves]

[f] λ

Gamma rays typically have frequencies above 10 exahertz (or >10^19 Hz), and therefore generally have energies above 100 keV and wavelengths less than 10 picometers (less than the diameter of an atom)

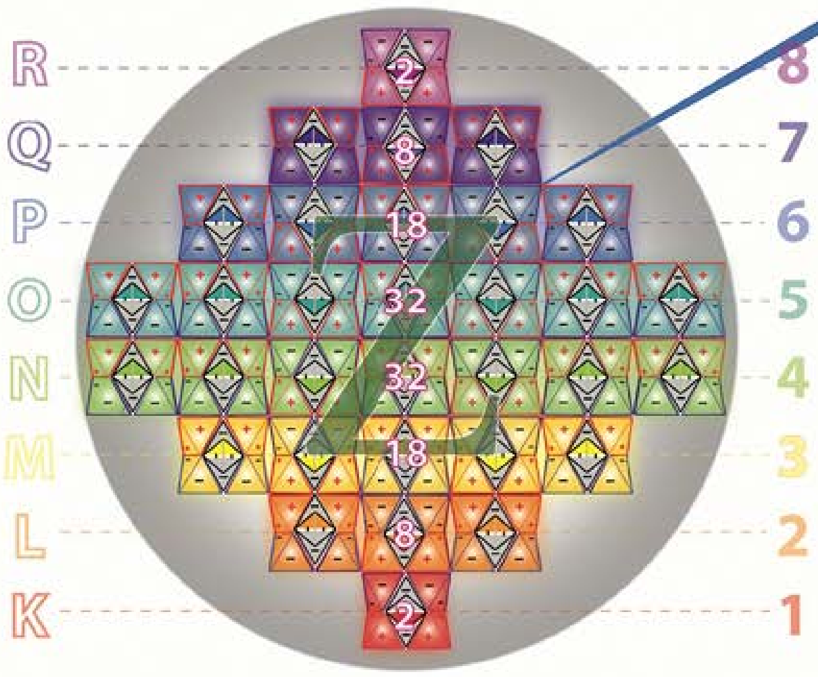

Alpha radiation - Positive charge particles (nuclei)
Beta radiation - Negative charge particles (leptons)

gamma rays

Gamma decay produces rays with energies of only a few hundred keV, and almost always less than 10MeV

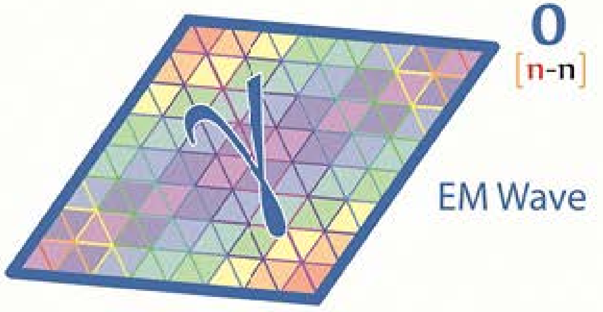

EM Field Planck quanta

$$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Photons ElectroMagnetic mass velocity

Gamma rays are a form of ionizing radiation, and they have very good penetrating power. They result from the release of atomic energies and will cause biological damage to living tissue.

$$2hv = E = hf$$

Care must always be taken to distingush between charged Planck quanta [v] and photons [f]

gamma ray bursts

Baryon mass-Matter topology

N^0 [18-18] → [2.25 e23] [v-v]

$$E = 2mc^2 = hf$$

930 MeV 930 MeV

gamma ray mass-energy geometry

are the result of the stored KEM mass-energies of Matter topologies being released as mass-energy momenta geometries

neutrino-Neutron Interactions

Neutrogen

$\nu e_{[6\text{-}6]} + N^0_{[18\text{-}18]} \overset{[24\text{-}24]}{\sim} P^+_{[24\text{-}12]} + e^-_{[0\text{-}12]}$

Charged Matter equivalence

Hydrogen

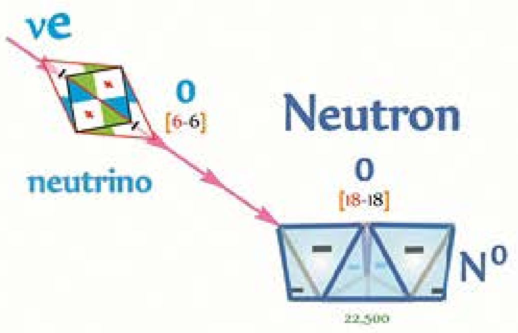

neutrino capture

Neutron decay is not the spontaneous process hypothesied by modern nuclear physicists

The observed decay products are the result of solar neutrinos interacting with Neutrons (both within nuclei and free)

the neutrinos also have an equally likely probability of interacting with Protons in the nuclei

[24-24]

A Neutronium atom has an identical mass-charge quotient to that of Hydrogen

[24-24]

Neutronium decay processes

A Neutron/neutrino interaction can decay into 4 differing particle sets

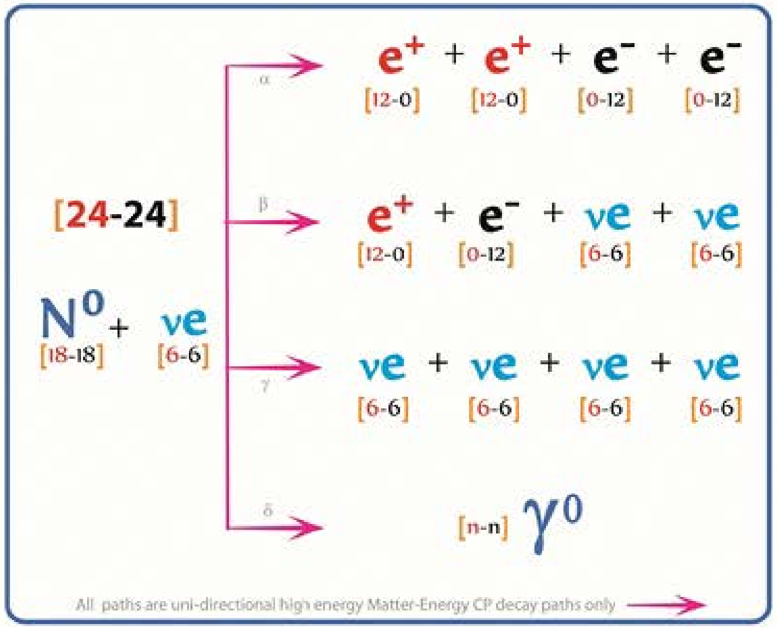

[24-24]

$N^0_{[18\text{-}18]} + \nu e_{[6\text{-}6]}$

α → $e^+_{[12\text{-}0]} + e^+_{[12\text{-}0]} + e^-_{[0\text{-}12]} + e^-_{[0\text{-}12]}$

β → $e^+_{[12\text{-}0]} + e^-_{[0\text{-}12]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]}$

γ → $\nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]} + \nu e_{[6\text{-}6]}$

δ → $_{[n\text{-}n]}\,\gamma^0$

All paths are uni-directional high energy Matter-Energy CP decay paths only →

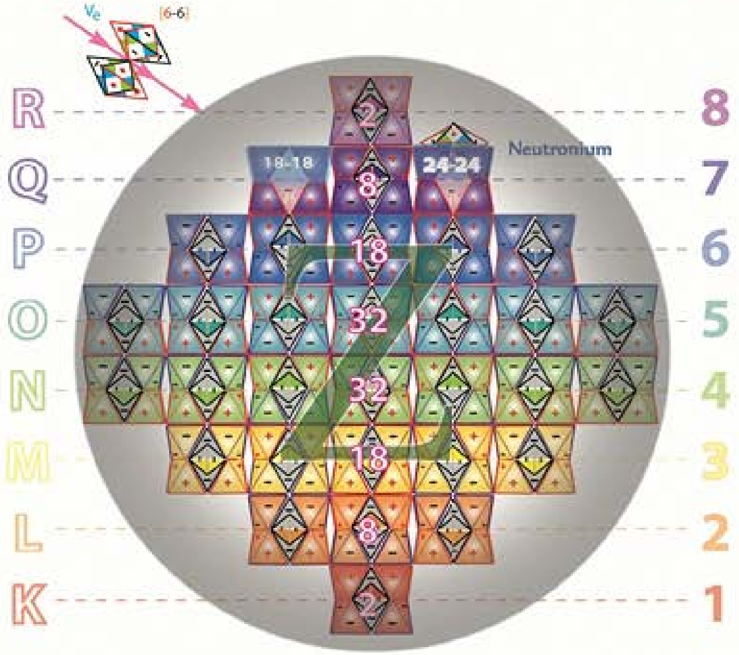

Excluding their nett Charge,
Neutrons have neutral particle geometries [18-18]
identical to that of a Proton [24-12]

$N^0_{[18\text{-}18]} \neq P^+_{[24\text{-}12]} + e^-_{[0\text{-}12]}$

[24-24]

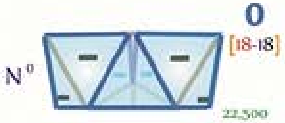

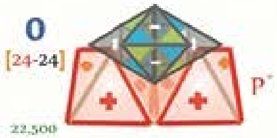

Spontaneous Neutron decay into Proton-electron-neutrinos is not possible without the interaction of Muonic neutrinos (see Tetryonic Charge numbers)

A Neutron is NOT a Proton that has absorbed an electron

Copyright ABRAHAM [2008] - All rights reserved

neutrino-Proton Interactions

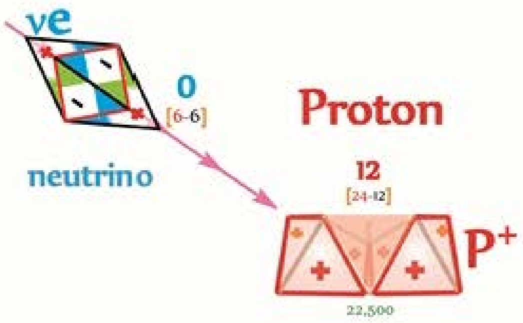

neutrino capture

Neutron decay is not the spontaneous process hypothesied by modern nuclear physicists

The observed decay products are the result of solar neutrinos interacting with Neutrons (both within nuclei and free)

the neutrinos also have an equally likely probability of interacting with Neutrons in the nuclei

[30-18]

A Proton-neutrino atom would have an identical charge quotient to that of a Proton

[24-12]

Although very weakly interacting, neutrinos can still be attracted to and bind to Protons via their negative charge fascia

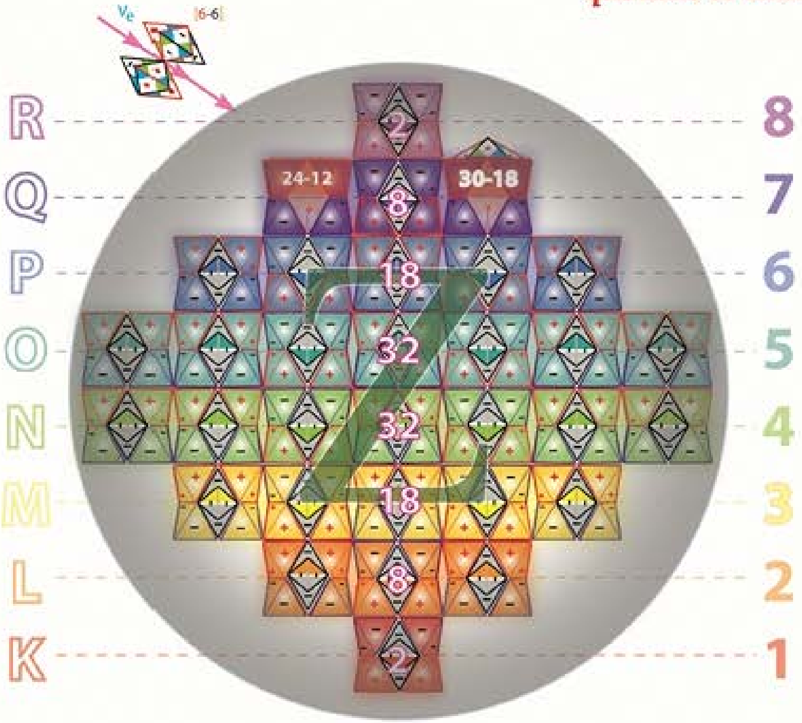

Proton decay paths

A Proton/Neutrino interaction can decay into 4 differing particle sets

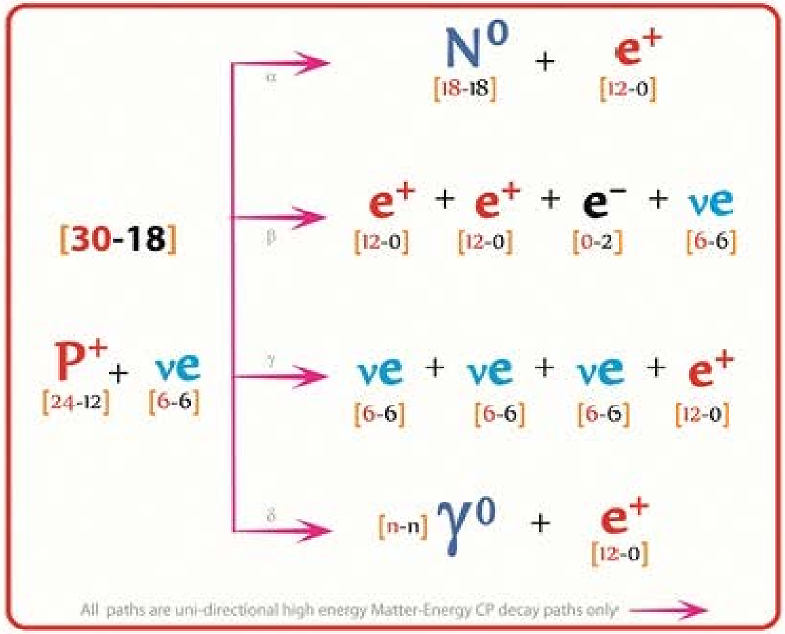

P^+ [24-12] + ve [6-6]

α: N^0 [18-18] + e^+ [12-0]

β: e^+ [12-0] + e^+ [12-0] + e^- [0-2] + ve [6-6]

γ: ve [6-6] + ve [6-6] + ve [6-6] + e^+ [12-0]

δ: [n-n] γ^0 + e^+ [12-0]

All paths are uni-directional high energy Matter-Energy CP decay paths only

If they exist, Proton-neutrino particle couplings would function in a manner identical to that of Proton-electron couplings and could be detected by the anomalous spin measurements that would result

A Proton has a Positive charge geometry [24-12] equivalent to that of a Neutral Neutron [18-18]

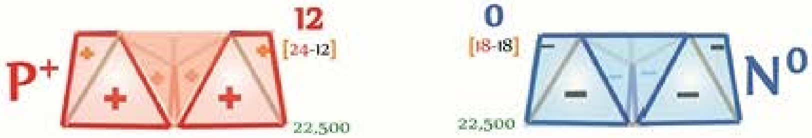

A Proton is topologically identical to a Neutron (differing only in the net charge created)

Quantum Batteries

Atomic nuclei can be easily scaled to non-quantum sizes to offer clean, safe and portable long term Energy storage devices that can store energy indefinitely and release it on demand anywhere in the World

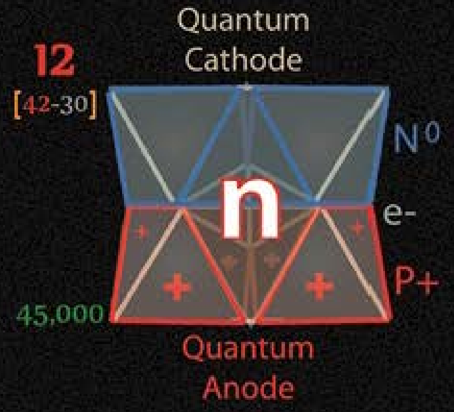
12
[42-30]
Quantum
Cathode
n
N 0
e-
P+
45,000
Quantum
Anode

The quantum battery is unique in that in addition to storing energy indefinitely, when an electron binds to the Deuteron nuclei it has the ability to release specific energies [photons] by way of its quantum-scale synchronous converter topologies

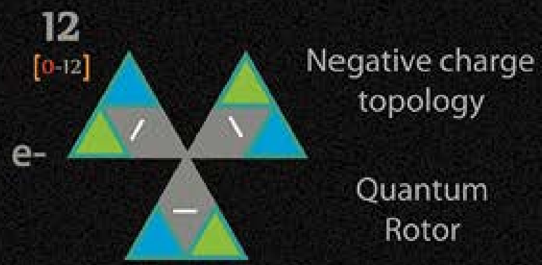
12
[0-12]
e-
Negative charge
topology
Quantum
Rotor

12 loop quantum inductive rotor

(Atomic Nuclei)

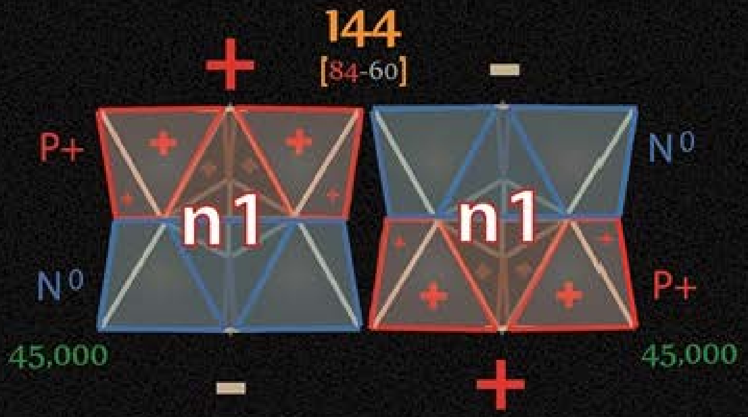
144
[84-60]
+
-
P+
N 0
n1
n1
N 0
P+
45,000
45,000
-
+

anti-Parallel Configuration

Synchronous quantum converter topologies can be connected in parallel or series to meet varying power requirements anywhere in the World and provides for the safe storage of nuclear energy as mass

Series Configuration

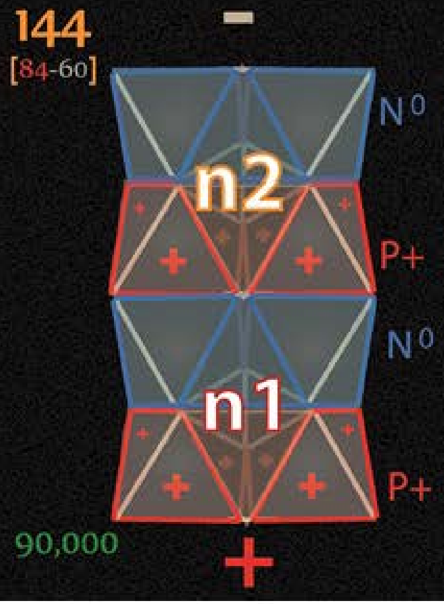
144
[84-60]
-
N 0
n2
P+
N 0
n1
P+
90,000
+

Quantum synchronous Converters

Energy can be stored in macro-scaled quantum converters as mass

The electron has a charged Tetryonic Matter topology that is electrically equivalent to a 6 loop inductive rotor

Energy

$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]^{n8}_{n1}$$

momenta

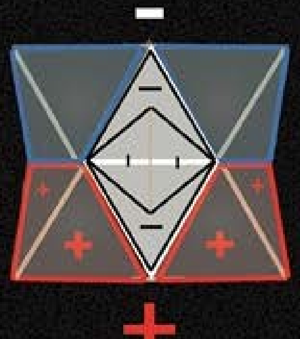

Matter is radiant EM energy in a standing-waveform

Energy

$$\text{Nuclei}\ \frac{72\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]^{n32}_{n25}$$

momenta

Building on the charged topology of Deuterium nuclei, scaled electromechanical quantum converters can be manufactured to provide efficient electrical mass-energy storage & distribution devices

Energy

$$\text{Deuterons}\ \frac{84\pi}{c^4}{}^{n32}_{n25}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]^{n8}_{n1}$$

momenta

These devices can be transported any where demand requires them worldwide with their energies stored in the form of mass

Negating the need for centralised power stations and distribution lines extending vast distances to provide power to remote communities

A quantum synchronous converter can store and release 3 forms of energy

Angular Momentum (motional energy)
square root linear momentum

Radiant Energy (emission/absorption)
of photon/boson mass-energy geometries

EM mass-Matter (stored masses in Matter topologies)
radiant EM mass-energies stored in standing wave Matter topologies

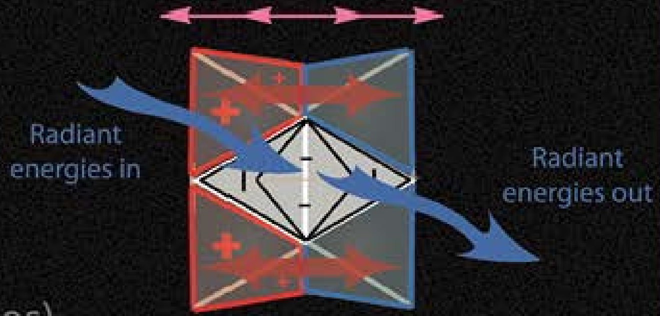

Leptons rotate within the Baryonic EM fields of the Nucleus

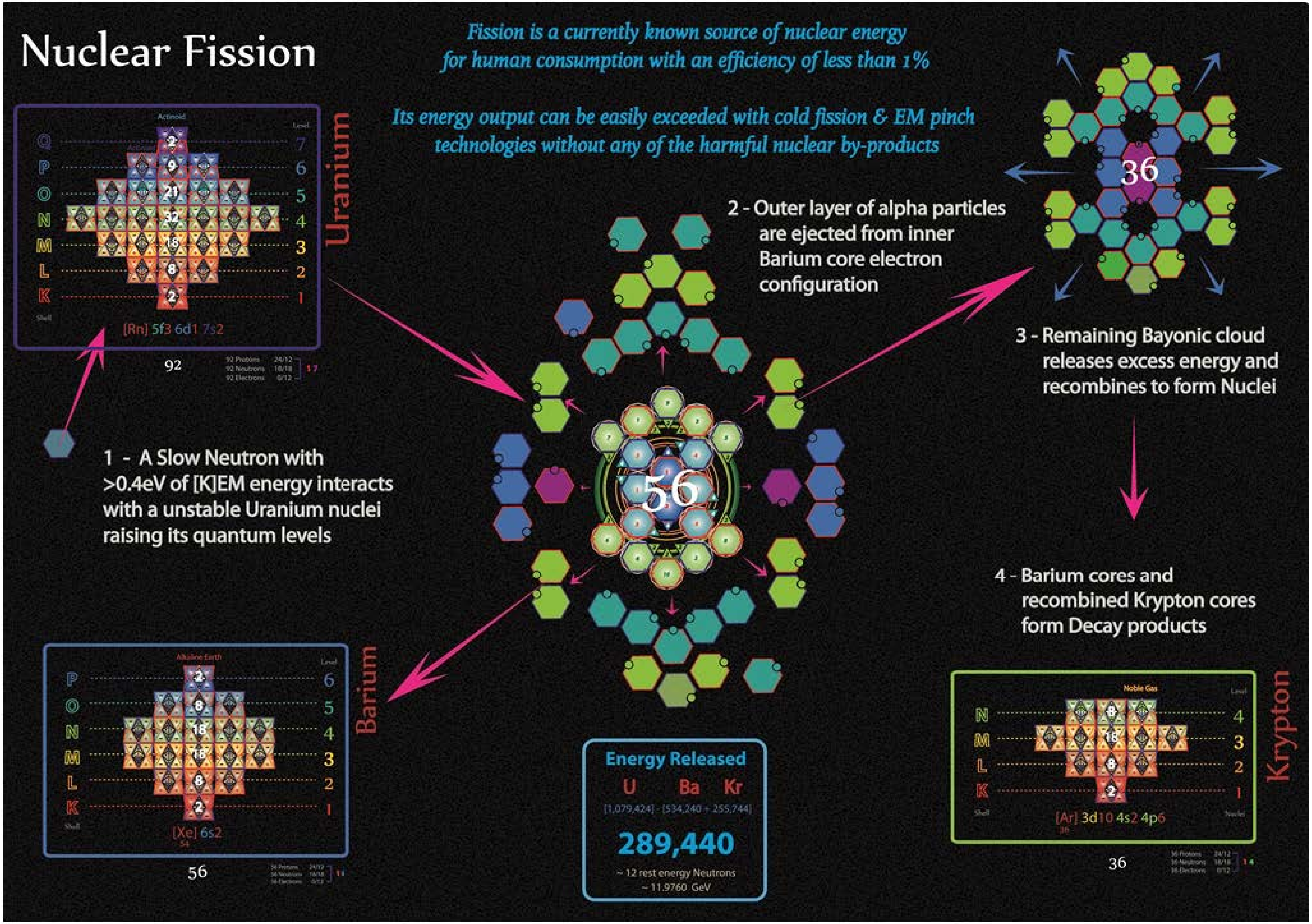
Actinoid
Level
Q
P
O
N
M
L
K
Shell
7
6
5
4
3
2
1
2
9
21
32
18
8
2
Uranium
[Rn] 5f3 6d1 7s2
92
92 Protons 24/12
92 Neutrons 18/18
92 Electrons 0/12
1 7
56
36
Alkaline Earth
Level
P
O
N
M
L
K
Shell
6
5
4
3
2
1
2
8
18
18
8
2
Barium
[Xe] 6s2
56
Noble Gas
Level
N
M
L
K
Shell
4
3
2
1
8
18
8
2
Nuclei
Krypton
[Ar] 3d10 4s2 4p6
36
36 Protons 24/12
36 Neutrons 18/18
36 Electrons 0/12
1 4

Nuclear Fission

Fission is a currently known source of nuclear energy for human consumption with an efficiency of less than 1%

Its energy output can be easily exceeded with cold fission & EM pinch technologies without any of the harmful nuclear by-products

1 - A Slow Neutron with >0.4eV of [K]EM energy interacts with a unstable Uranium nuclei raising its quantum levels

2 - Outer layer of alpha particles are ejected from inner Barium core electron configuration

3 - Remaining Bayonic cloud releases excess energy and recombines to form Nuclei

4 - Barium cores and recombined Krypton cores form Decay products

Energy Released

U Ba Kr

[1,079,424] - [534,240 + 255,744]

289,440

~ 12 rest energy Neutrons
~ 11.9760 GeV

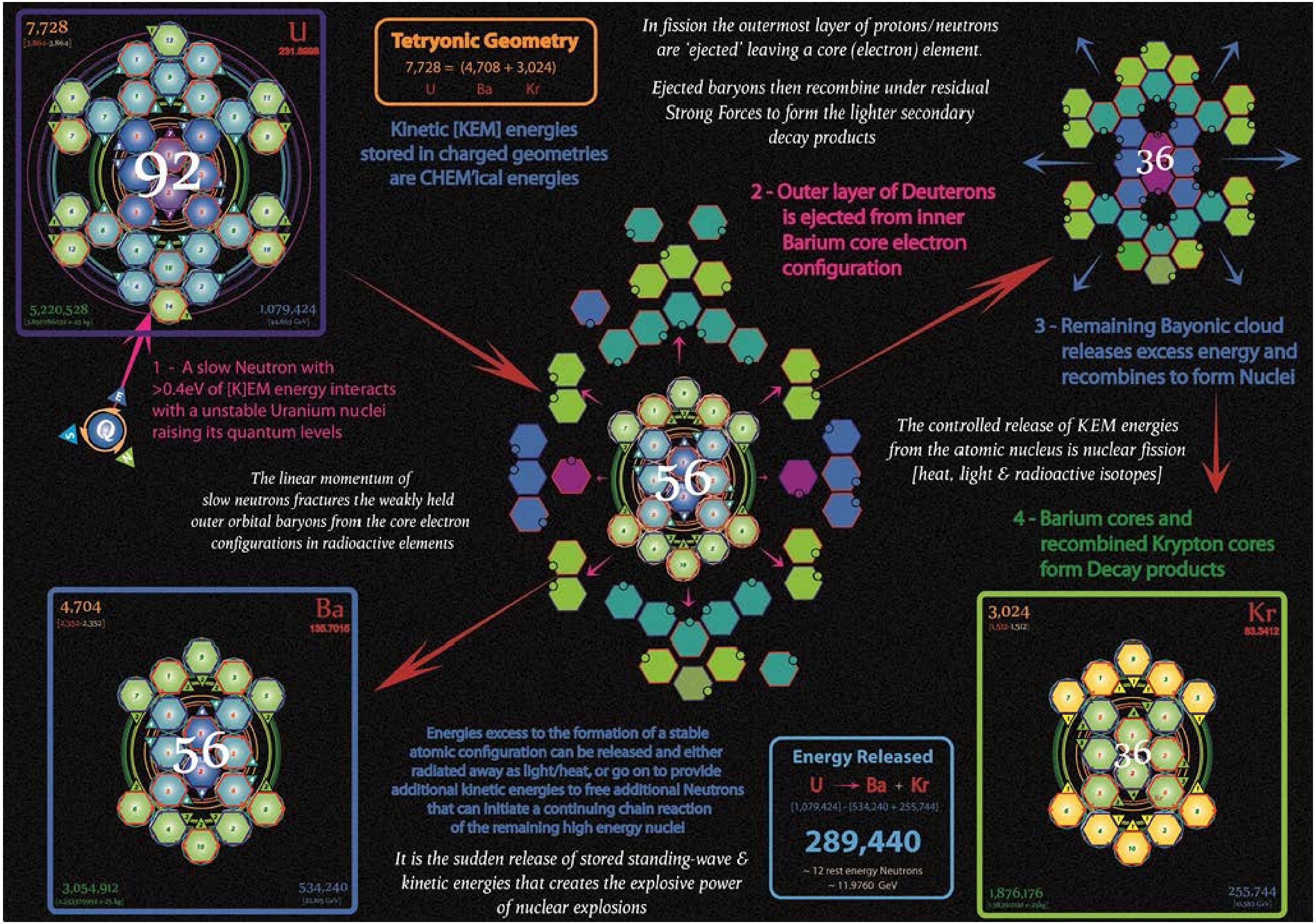

Tetryonic Geometry

7,728 = (4,708 + 3,024)
U Ba Kr

Kinetic [KEM] energies stored in charged geometries are CHEM'ical energies

In fission the outermost layer of protons/neutrons are 'ejected' leaving a core (electron) element.

Ejected baryons then recombine under residual Strong Forces to form the lighter secondary decay products

1 - A slow Neutron with >0.4eV of [K]EM energy interacts with a unstable Uranium nuclei raising its quantum levels

The linear momentum of slow neutrons fractures the weakly held outer orbital baryons from the core electron configurations in radioactive elements

2 - Outer layer of Deuterons is ejected from inner Barium core electron configuration

3 - Remaining Bayonic cloud releases excess energy and recombines to form Nuclei

The controlled release of KEM energies from the atomic nucleus is nuclear fission [heat, light & radioactive isotopes]

4 - Barium cores and recombined Krypton cores form Decay products

Energies excess to the formation of a stable atomic configuration can be released and either radiated away as light/heat, or go on to provide additional kinetic energies to free additional Neutrons that can initiate a continuing chain reaction of the remaining high energy nuclei

It is the sudden release of stored standing-wave & kinetic energies that creates the explosive power of nuclear explosions

Energy Released

U → Ba + Kr

[1,079,424] - [534,240 + 255,744]

289,440

~ 12 rest energy Neutrons
~ 11.9760 GeV

Tetryonics 60.13 - Nuclear Fission process

Sonoluminescence

Sonoluminescence is the first hint at the energies that can be released from Tetryonic collapse and involves the emission of short bursts of light from imploding bubbles in a liquid when excited by sound.

It is a key step on the road to realising technologies that can provide Humanity with clean, safe, efficient Energy production through the conversion of Matter into various forms of EM radiation

The production of light as a result of the passing of sound waves through a liquid medium.

The sound waves cause the formation of bubbles that emit bright flashes of light when they collapse.

Sonoluminesence is the result of energy releases from the collapse of standing-wave Matter topologies within Fluids

Finding an efficient means of inducing and managing a controlled reaction where 3D Matter topologies are converted into 2D radiant energy is a key step on the path to creating a future energy source of all Humanity

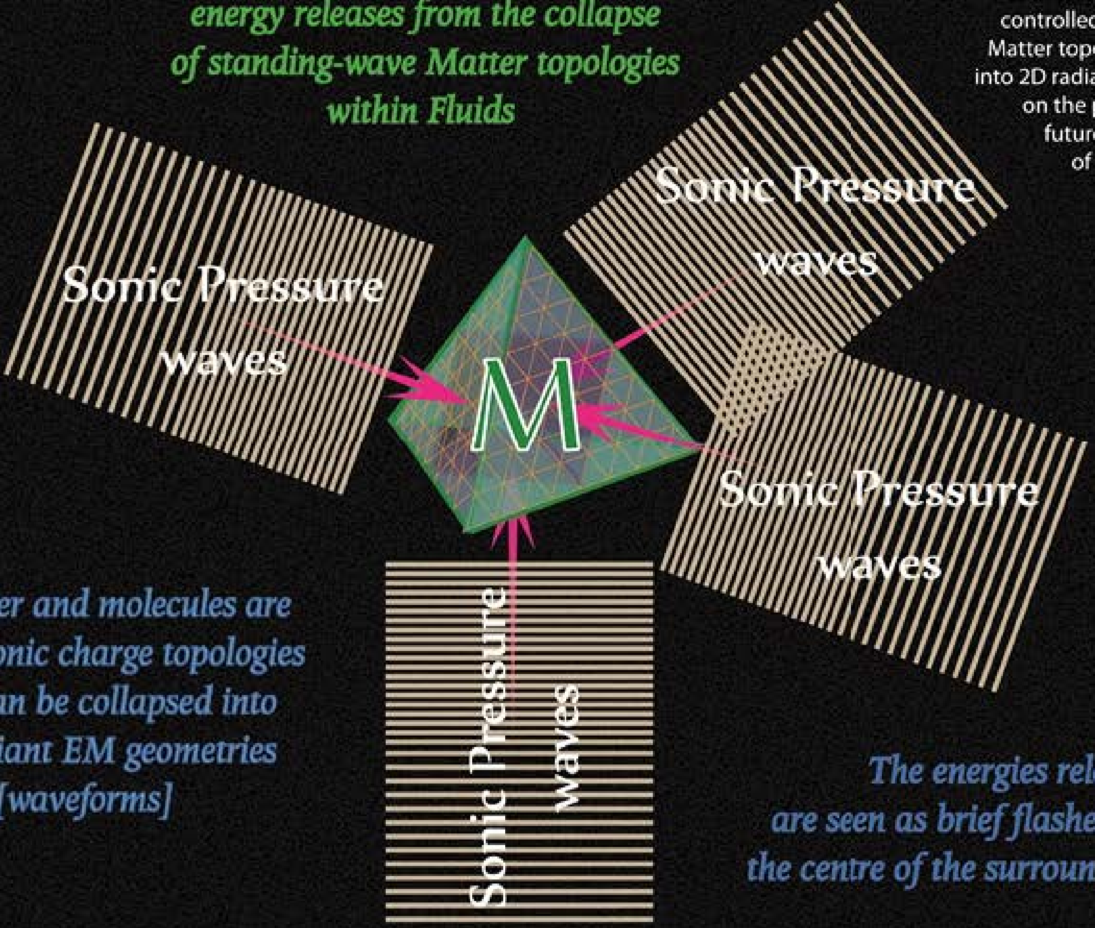

The gauge transformation of Matter into EM radiation is the most efficient form of energy release available to us

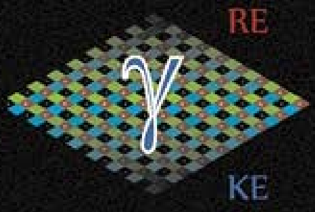

All Matter and molecules are 3D Tetryonic charge topologies which can be collapsed into 2D radiant EM geometries [waveforms]

The energies released are seen as brief flashes of light at the centre of the surrounding medium

Stanley Pons Martin Fleischmann

(23 August 1943) (29 March 1927 – 3 August 2012)

'Nuclear fusion of the type postulated would be inconsistent with current understanding and, if verified, would require theory to be extended in an unexpected way'

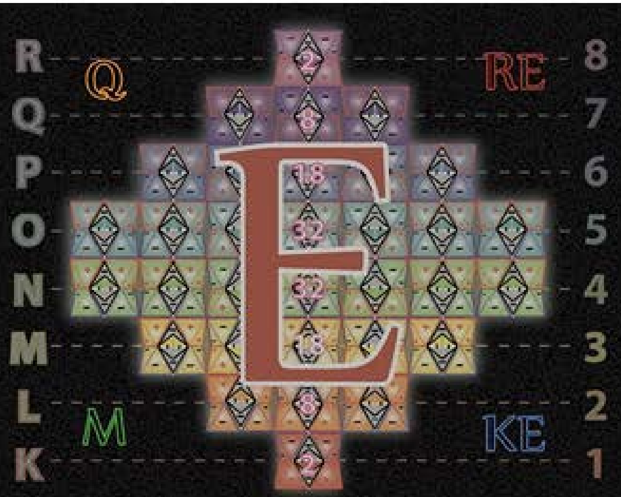
R
Q
P
O
N
M
L
K
Q
RE
M
KE
E
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2

The reported excess energy released is often refused as being impossible as it woud require the release of massive neutron radiation bath.

Cold Fission

Atomic nuclei can easily release the reported energies without Neutron particle emissions when an accurate model of the atomic nuclei is used

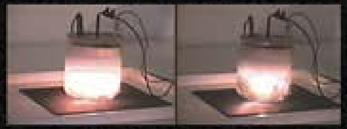

'Fusion' is a fictious nuclear process 'cold fusion' is in fact a form of nuclear fission

Palladium's rest molar mass is 108.7867 with a stored Kinetic energy content of 15.699 Gev [16 neutron equivalence]

Expelling the non-core atomic nuclei results in a release of stored KEM energies as the deuterium nuclei seek to reach a ground energy level state

Krypton's rest molar mass is 83.3412 with a stored Kinetic energy content of 10.582 Gev [11 neutron equivalence]

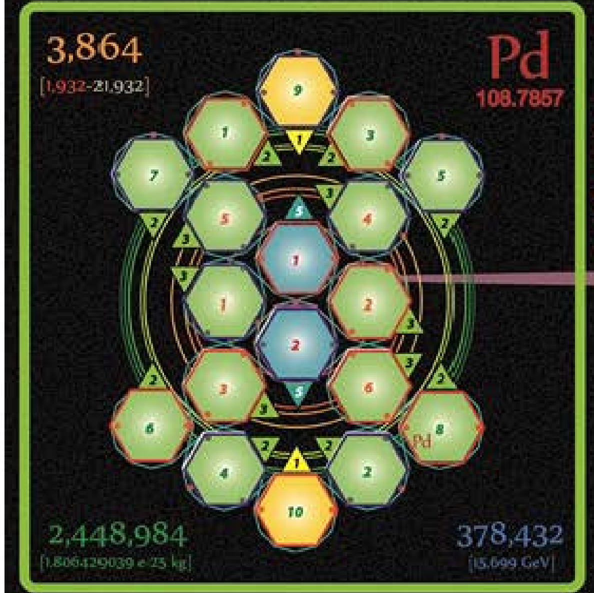
3,864
[1,932-21,932]
Pd
108.7857
9
1
3
7
5
2
1
2
5
4
5
1
1
2
2
3
3
6
8
Pd
4
2
10
2,448,984
[1.806429039 e-25 kg]
378,432
[15.699 GeV]

Palladium

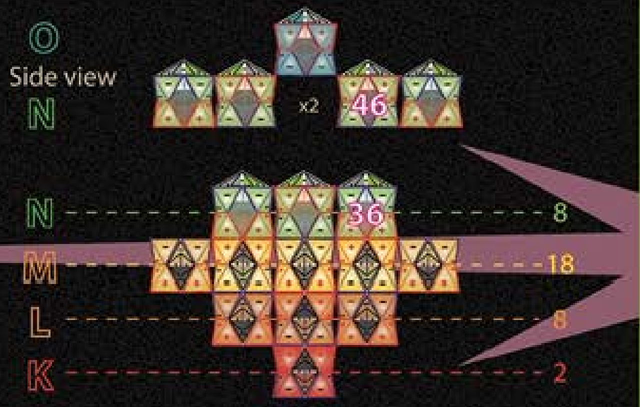
O
Side view
N
x2
46
N
M
L
K
36
8
18
8
2

2 x n5
10 Deuterons
8 x n4

122,688
[5.076 GeV]

Energy equivalent to 5+ Neutrons

NO excess Neutrons are ever released ~ only stored KEM energies and alpha particles

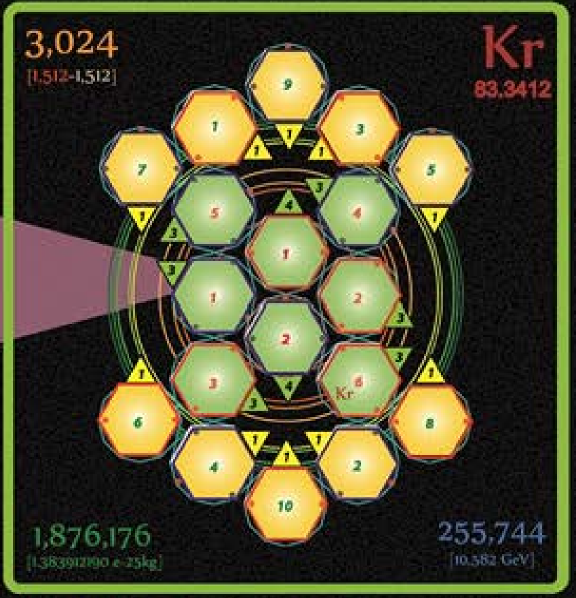
3,024
[1,512-1,512]
Kr
83.3412
9
1
3
7
5
5
4
1
1
2
2
3
6
Kr
6
8
4
2
10
1,876,176
[1.383912190 e-25kg]
255,744
[10.582 GeV]

Krypton

[Kr] 4d8 5s2
36

Transition Metal

46 Protons 24-12
46 Neutrons 18-18
46 Electrons 0-12
1 5

[Ar] 3d10 4s2 4p6
36

Noble Gas

36 Protons 24-12
36 Neutrons 18-18
36 Electrons 0-12
1 4

Copyright ABRAHAM [2008] - All rights reserved

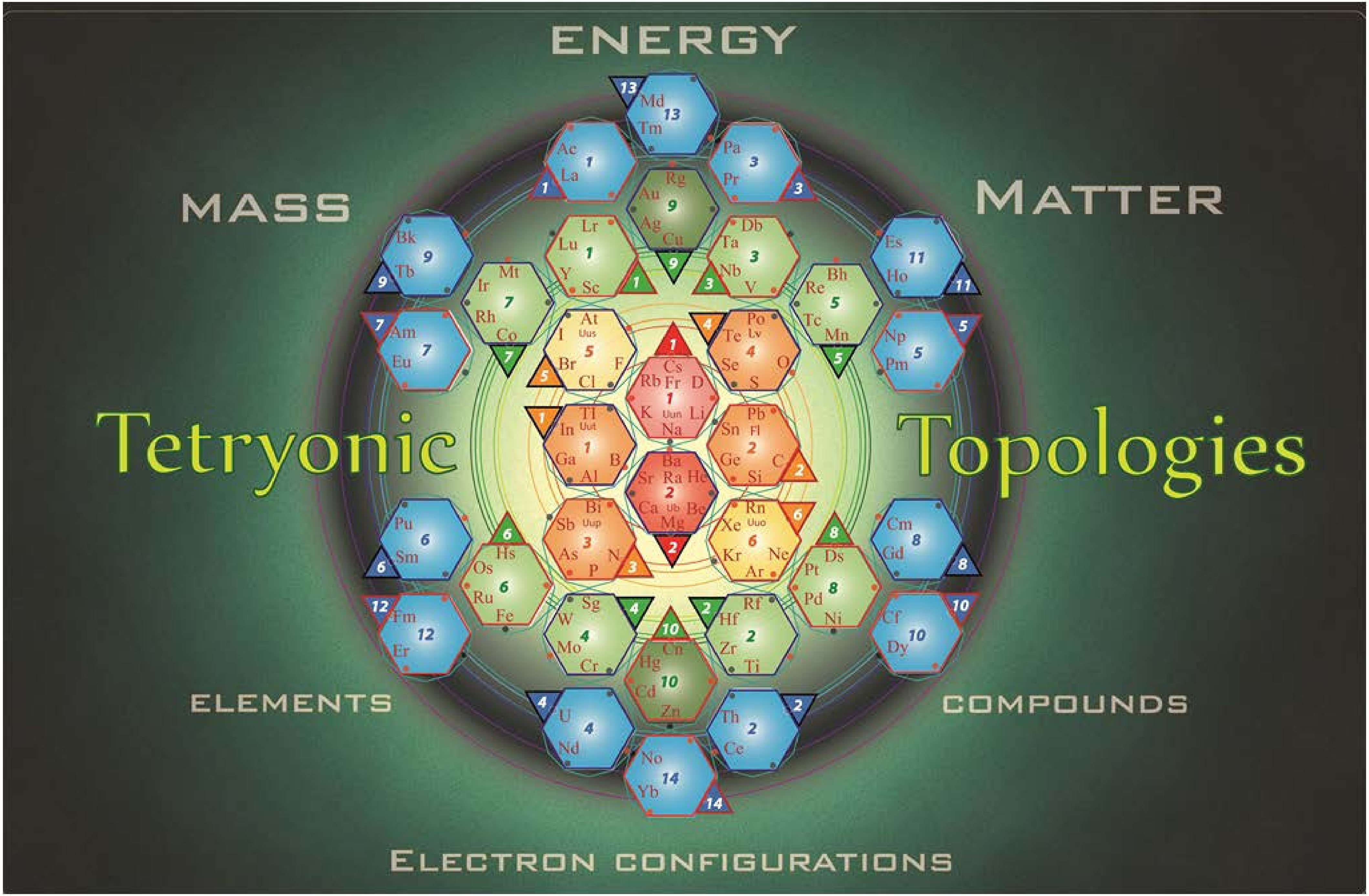
ENERGY
MASS
MATTER
Tetryonic
Topologies
ELEMENTS
COMPOUNDS
ELECTRON CONFIGURATIONS
13 Md 13 Tm
Ac 1 La 1
Pa 3 Pr 3
Rg Au 9 Ag Cu 9
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Bk 9 Tb 9
Es 11 Ho 11
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
7 Am 7 Eu
5 Np 5 Pm
At I Uus 5 Br F Cl 5
4 Po Te Lv 4 Se O S
1 Cs Rb Fr D 1 K Uun Li Na
1 Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Bi Sb Uup 3 As N P 3
Rn Xe Uuo 6 Kr Ne Ar 6
Pu 6 Sm 6
Cm 8 Gd 8
6 Hs Os 6 Ru Fe
8 Ds Pt 8 Pd Ni
12 Fm 12 Er
Cf 10 Dy 10
Sg W 4 Mo Cr 4
10 Cn Hg 10 Cd Zn
2 Rf Hf 2 Zr Ti
4 U 4 Nd
Th 2 Ce 2
No 14 Yb 14

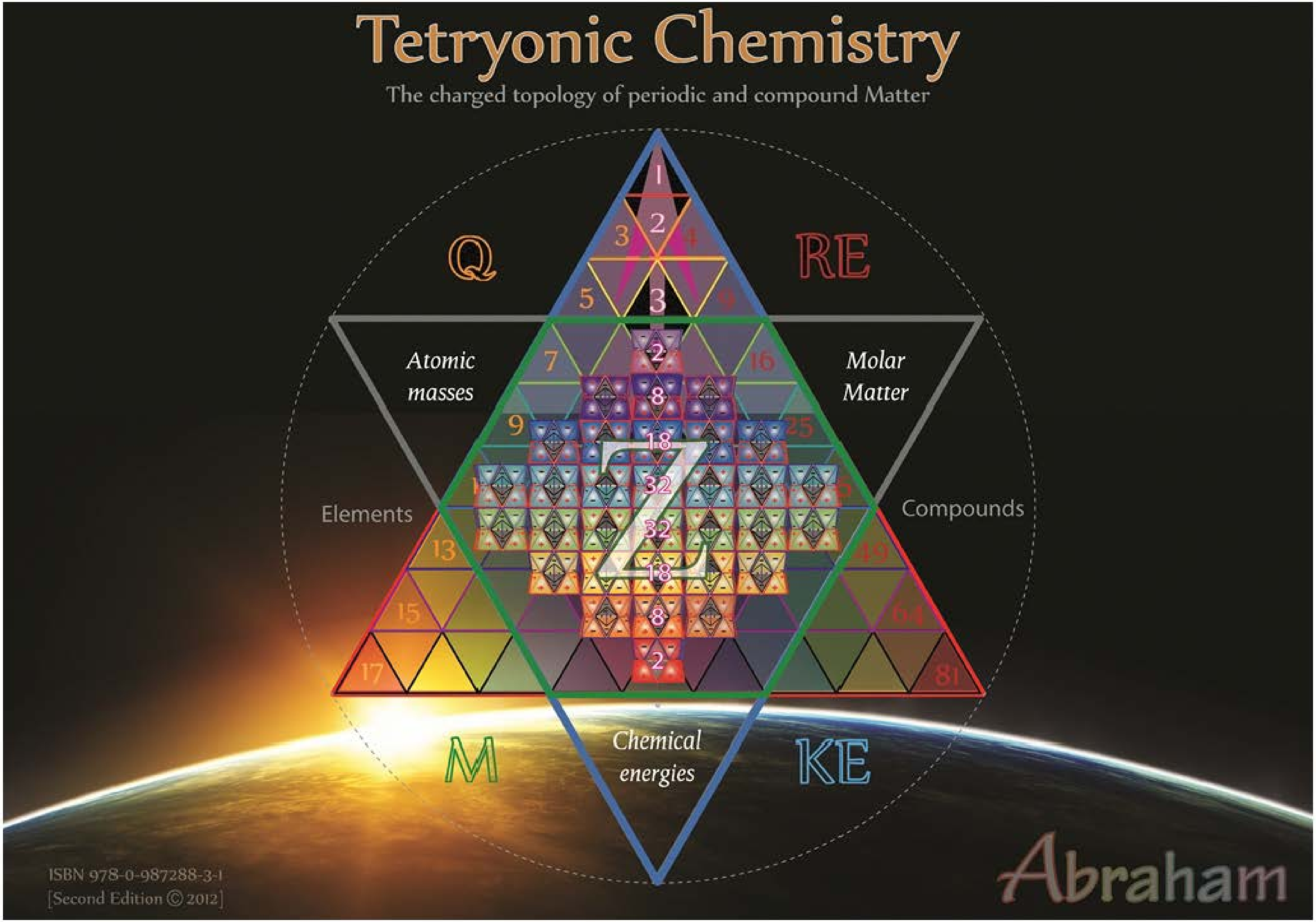

Tetryonic Chemistry

The charged topology of periodic and compound Matter

Q
RE
M
KE
Atomic masses
Molar Matter
Elements
Compounds
Chemical energies
Z
1
2
3
3
4
5
9
7
16
9
25
13
49
15
64
17
81
2
8
18
32
32
18
8
2

ISBN 978-0-987288-3-1
[Second Edition © 2012]

Abraham

TETRYONICS

The charged interactions of universal Gravitation

Q
RE
ε
G
Radiant masses
Standing-wave Matter
Gravitational GEM fields
Interactive KEM fields
μ
k
M
KE
Vacuum Energies
1
3
5
7
9
11
13
15
17
4
5
6
7
8
9
25
36
49

Foundational Quantum Cosmology

Abraham

ISBN 978-0-987288-4-8

[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Copyright ABRAHAM [2008] - All rights reserved

Laws of Force

F = ma

Sir Issac Newton

(25 December 1642 – 20 March 1727)

w = mg

Gravity

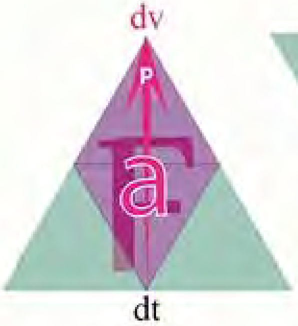

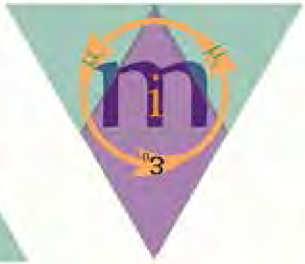

changing Magnetic lines of Force

Michael Faraday

(22 September 1791 – 25 August 1867)

create changing Electric fields of Force

Newton's Principia formulated the laws of motion and of universal gravitation

Faraday discovered the principles of electromagnetic induction, diamagnetism, and the laws of electromagnetism.

Electro-Magnetism

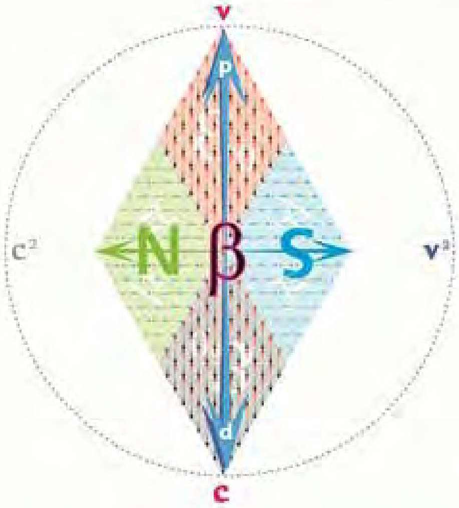

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

Gauss

Light is an electromagnetic wave

Faraday

$$\oint \vec{E}\cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

James Clerk Maxwell

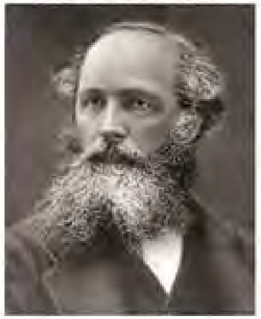

(13 June 1831 – 5 November 1879)

$$\oint \vec{B}\cdot d\vec{A} = 0$$

Gauss

The separate works of Gauss, Ampere & Faraday were unified to create an electromagnetic wave field theory

Ampere

$$\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}$$

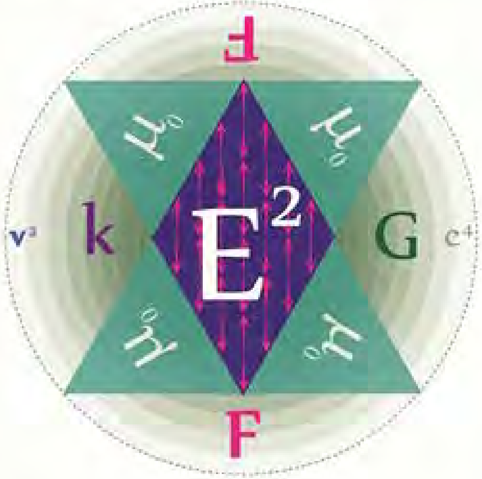

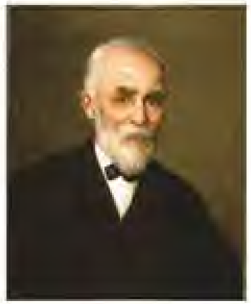

Lorentz Contraction

Lorentz proposed that moving bodies contract in the direction of motion (Lorentz–Fitzgerald contraction)

Hendrik Lorentz

(18 July 1853 – 4 February 1928)

Lorentz Transformations

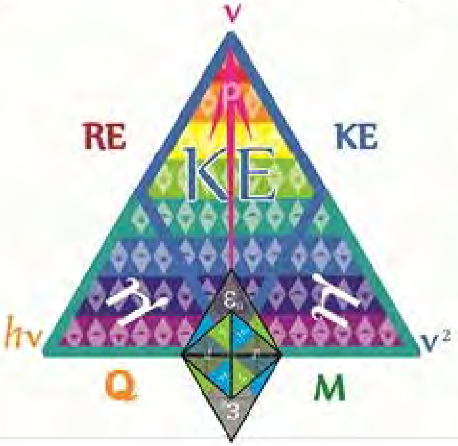

Special Relativity

attempted to explain the constant speed of Light using relative motion & Lorentz contractions

Albert Einstein

(14 March 1879 – 18 April 1955)

&

later attempted to explain gravity as a curvature of spacetime caused by Matter

General Relativity

Copyright ABRAHAM [2008] - All rights reserved

Early physics of Motion

In the 4th century BC, the Greek philosopher Aristotle believed that
there is no effect or motion without a cause

Greek philosopher Aristotle

(384 BC - 322 BC)

Aristotle defined space to be
the boundary geometry of a body

[x,y,z]
geometry, toplogy

Aristotle defined time to be
a measure of movement

the past Δ the future

'Nature abhors a vacuum'

In Book VII of his De Architectura, the Roman engineer and architect Vitruvius
contends that gravity is not dependent on a substance's "weight" but rather on its "nature".

***The cause of the downward motion of heavy bodies, such as the element earth,
was related to their nature, which caused them to move***

In the 7th Century the Indian mathematician Brahmagupta stated
"Bodies fall towards the earth as it is in the nature of the earth to attract bodies,
just as it is in the nature of water to flow."

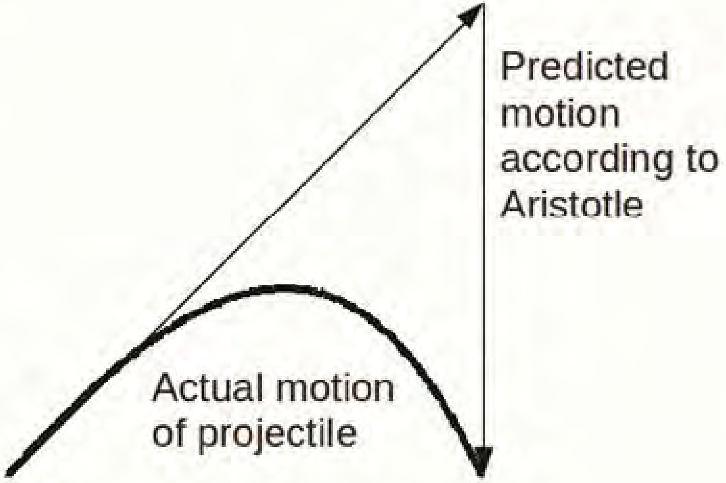
Predicted motion according to Aristotle
Actual motion of projectile

Heavier things fall faster, with speed being proportional to weight.

The speed with which a body falls is inversely proportional
to the density of the medium it is falling through.

Indian Mathematician Brahmagupta

(597–668 AD)

Brahmagupta dedicated a substantial
portion of his work to geometry

height
base

Brahmagupta is considered the first
to formulate the concept of zero

negative numbers 0 positive numbers

Brahmagupta introduced new algebric
methods for solving quadratic equations

Copyright ABRAHAM [2008] - All rights reserved

Galilean Gravity

Galileo Galilei

(15 February 1564 – 8 January 1642)

Galileo claimed to have a Universal Law of Gravitation covering both terrestrial gravity and the motion of planets which he was afraid to discuss.

But this looks more an inspired aspiration than a reality, as he seems not to have considered gravitational force as decreasing with distance from its source.

He did however, definitively show that objects experience a uniform acceleration in a gravitational field (irrespective of their masses) which was central to Newton's later Universal Law of Gravitation

The Motion of Falling Bodies

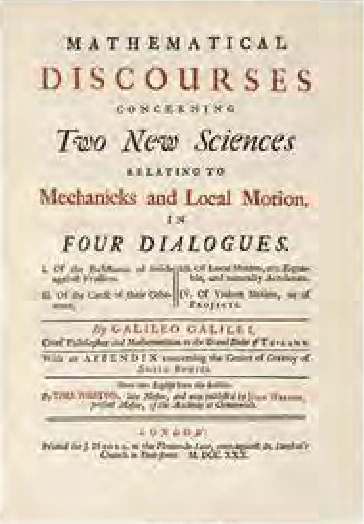

MATHEMATICAL
DISCOURSES
CONCERNING
Two New Sciences
RELATING TO
Mechanicks and Local Motion,
IN
FOUR DIALOGUES.

by GALILEO GALILEI,

LONDON:

(ca. 1590)

Galileo dismissed as a "useless fiction" the idea, held by his contemporary Johannes Kepler, that the moon caused the tides.

Galileo also refused to accept Kepler's elliptical orbits of the planets, considering the circle the "perfect" shape for planetary orbits.

Galileo was the first to counter Aristotle's earlier idea's of motion by teaching that
All objects experience a uniform acceleration in a Gravitational field
(irrespective of their different mass-Matter contents)

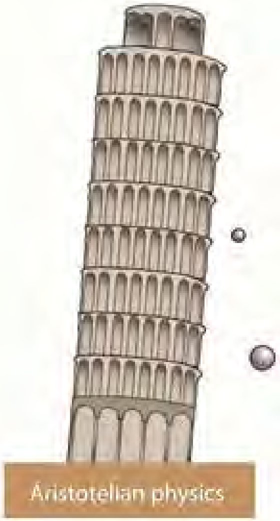

Aristotelian physics

In De motu Galileo proposed that in free fall bodies dropped with a characteristic uniform speed determined not by their weight but by their specific gravity.

He put this theory to the test by dropping bodies from heights and found that his experiments did not confirm his theory.

It is quoted in 'Galileo's Daughter' that the lighter body (i.e. that of the lower specific gravity) will move ahead of the heavier body at the start of the fall, and that the heavier body then overtakes it and arrives at the bottom slightly earlier.

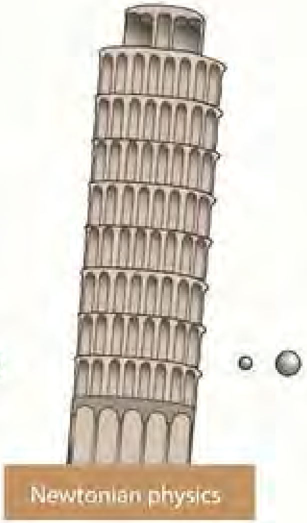

Newtonian physics

Copyright ABRAHAM [2008] - All rights reserved

Sir Issac Newton

(25 December 1642 – 20 March 1727)

In the late 17th century, Robert Hooke's suggested
'that there is a gravitational force which depends on the inverse square of the distance'

Newtonian Gravitation

Newton's monograph Philosophiæ Naturalis Principia Mathematica, published in 1687, lays the foundations for most of classical mechanics. In this work, Newton described Universal gravitation and the three laws of motion, which dominated the scientific view of the physical universe for the next three centuries.

Newton showed that the motions of objects on Earth and of celestial bodies are governed by the same set of natural laws, by demonstrating the consistency between Kepler's laws of planetary motion and his theory of gravitation, thus removing the last doubts about heliocentrism and advancing the Scientific Revolution.

The Principia is generally considered to be one of the most important scientific books ever written.

Principia Mathematica

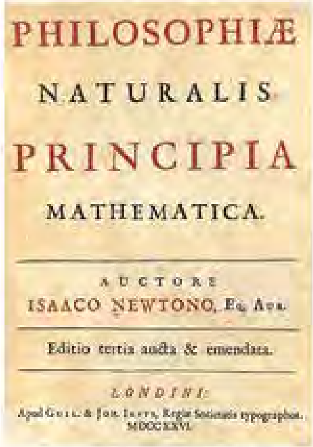

PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

AUCTORE
ISAACO NEWTONO, Eq. Aur.

Editio tertia aucta & emendata.

LONDINI:

first edition (1686/1687)

First Law of Motion

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

*Bodies at rest - remain at rest
unless acted upon by an outside Force*

Second Law of Motion

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

*The nett force on a particle is equal to the time
rate of change of its linear momentum [p]*

Third Law of Motion

$$\sum \mathbf{F}_{a,b} = -\sum \mathbf{F}_{b,a}$$

*For every action there is
an equal and opposite reaction*

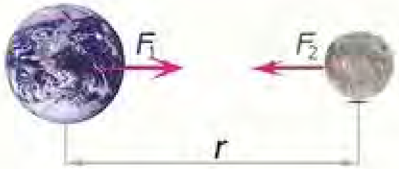

Newton's Law of universal Gravitation

"I deduced that the forces which keep the planets in their orbs must [be] reciprocally as the squares of their distances from the centers about which they revolve: and thereby compared the force requisite to keep the Moon in her Orb with the force of gravity at the surface of the Earth; and found them answer pretty nearly"

$$F = -G \frac{M_1 m_2}{r^2}$$

"hitherto I have not been able to discover the cause of those properties of gravity from the phenomena, and I frame no hypothesis"

Copyright ABRAHAM [2008] - All rights reserved

According to Bošković's simple dynamistic atomism, Matter is not only endowed with forces (dynamic system), but it is composed of forces (dynamistic system). Forces flow out of the atom and permeate empty space.

This idea led to the concept of the field, much later formulated by M. Faraday (1844), who together with J.C. Maxwell introduced the idea of fields of force into science.

Boskovic is famous for his atomic theory and made many important contributions to astronomy, including the first geometric procedure for determining the equator of a rotating planet from three observations of a surface feature and for computing the orbit of a planet from three observations of its position.

In 1753 he also discovered the absence of atmosphere on the Moon.

Rudjer Boskovic

(18 May 1711 – 13 February 1787)

k

Boskovic proposed that attractive Gravitational fields could be explained as diminshed interactive EM fields that result in a nett converent force at great distances

G

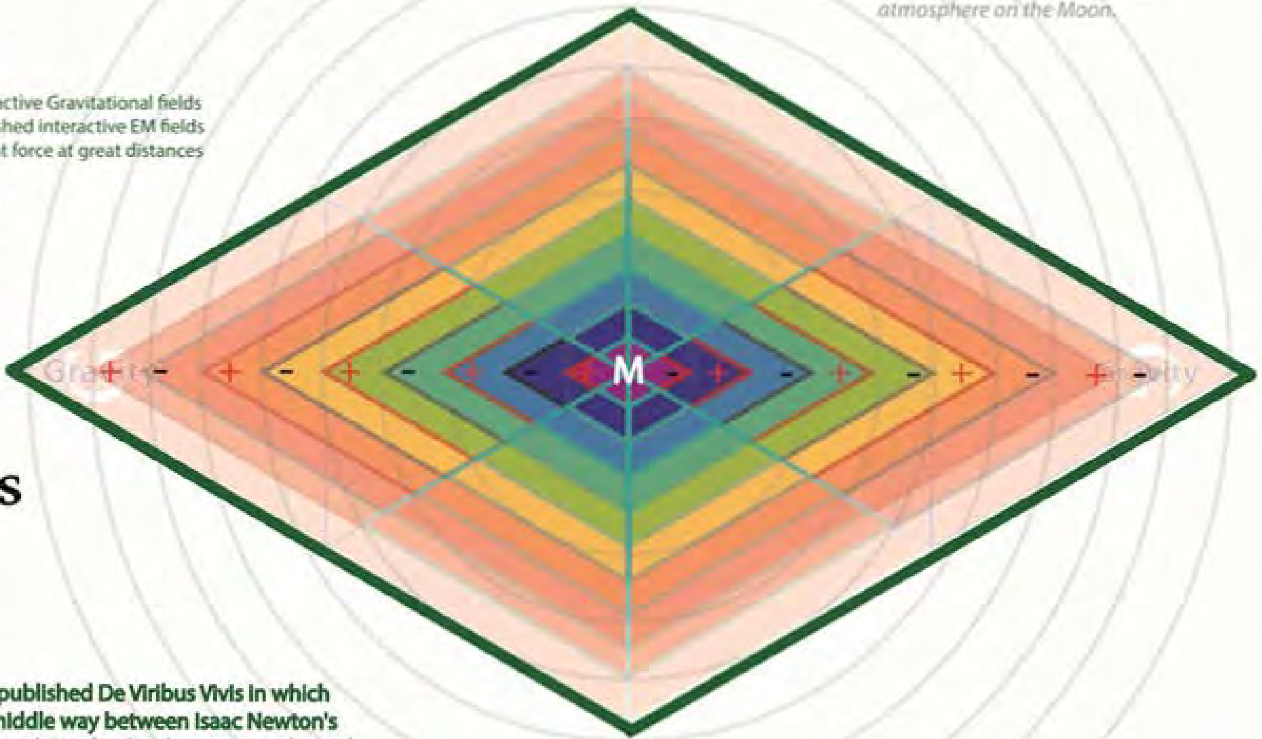

Diminished EM waves

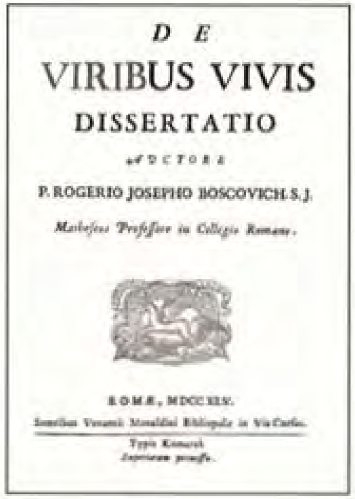

D E

VIRIBUS VIVIS

DISSERTATIO

AUCTORE

P. ROGERIO JOSEPHO BOSCOVICH S.J.

Matheſeos Profeſſore in Collegio Romano.

ROMÆ, MDCCXLV.

Sumtibus Venantii Monaldini Bibliopolæ in Via Curſus.

Typis Komarek

Superiorum permiſſu.

In 1745 Bošković published De Viribus Vivis in which he tried to find a middle way between Isaac Newton's gravitational theory and Gottfried Leibniz's metaphysical theory of monad-points

The force is repulsive when this curve lies above the line AC and attractive when it lies below it.

At very large distances (at and beyond V), it is attractive and approaches Newton's inverse-square law of force produced by gravity

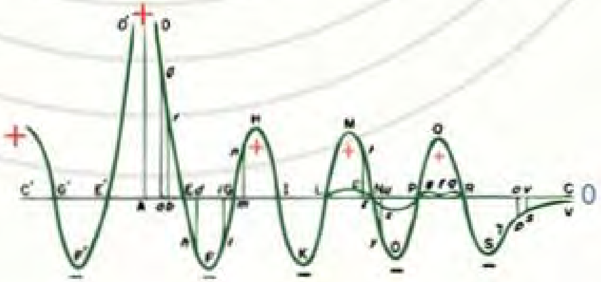

nett attractive force at large distances

Gaussian Gravity

Gauss's law for gravity bears the same mathematical relation to Newton's law that Gauss's law for electricity bears to Coulomb's law.

G

Gauss's law for gravity is mathematically similar in form to Gauss's law for electrostatics.

k

the striking mathematical similarity in the formulation of electrostatics and gravitation has puzzled science for centuries

$$F_{elec}\ k \propto \frac{Q_1 Q_2}{r^2}$$

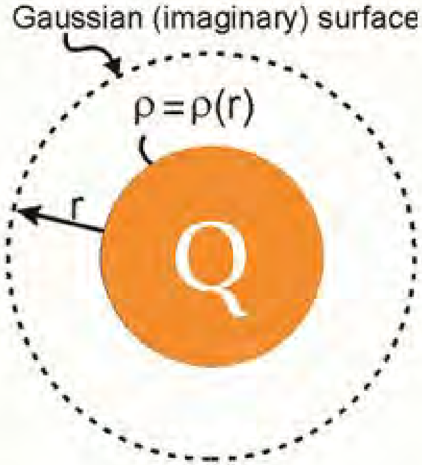

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0} = 4\pi kq$$

The electric flux through any closed surface is proportional to the enclosed electric charge.

$$\nabla \cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k\rho$$

Carl Friedrich Gauss

(30 April 1777 – 23 February 1855)

Gauss's law for gravity, also known as Gauss's flux theorem for gravity, is a law of physics which is essentially equivalent to Newton's law of universal gravitation.

Although Gauss's law for gravity is physically equivalent to Newton's law, there are many situations where Gauss' law for gravity offers a more convenient and simple way to do a calculation than Newton's law.

note: 4π spherical topologies are equivalent to 4π tetrahedral topologies
[see Euler characteristics & Gauss-Bonnet topology theorem]

$$F_{grav}\ G \propto \frac{M_1 m_2}{r^2}$$

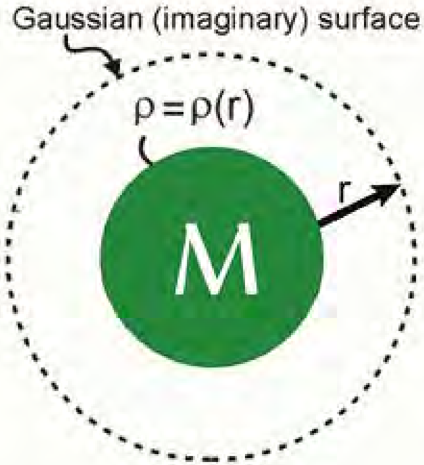

$$\oint_{\partial V} \mathbf{g}\cdot d\mathbf{A} = -4\pi GM$$

The gravitational flux through any closed surface is proportional to the enclosed Matter

$$\nabla \cdot \mathbf{g} = -4\pi G\rho$$

Poisson's Gravity fields

Charges are the sources
and sinks of EM fields

Poisson recognised that charges are the sources and sinks of electrostatic fields:
positive charges emanate electric field lines, and the field lines terminate at negative charges.

Similarly, in Newton's gravitation masses are the sources of the gravity field
so that field lines terminate at objects that have mass.

Matter is the source
of Gravitational fields

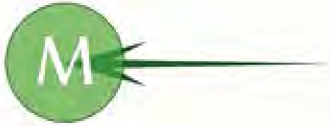

Poisson formalised Gauss' Law for electric fields (using the more general divergence theorem)

$$\iint \vec{E} \cdot d\vec{S} = \frac{q_e}{\epsilon_0} \Rightarrow \vec{\nabla} \cdot \vec{E} = \frac{\rho_e}{\epsilon_0}$$

and then applied the same formulation to spherical Matter for gravitation

$$\iint \vec{g} \cdot d\vec{S} = -4\pi G_c m \Rightarrow \vec{\nabla} \cdot \vec{g} = -4\pi G_c \rho_m$$

Simeon Poisson

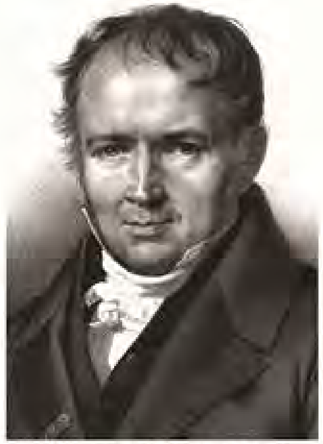

(21 June 1781 – 25 April 1840)

All EM field interactions are the result of superpositioned EM fields (comprising of convergent AND divergent energy momenta) whose quanta create differing field strengths,
The inverse square forces of gravity can be modelled using the same field geometry thus illustrating the reason for the geometric & mathematical similarity between Newton's & Coulomb's constants

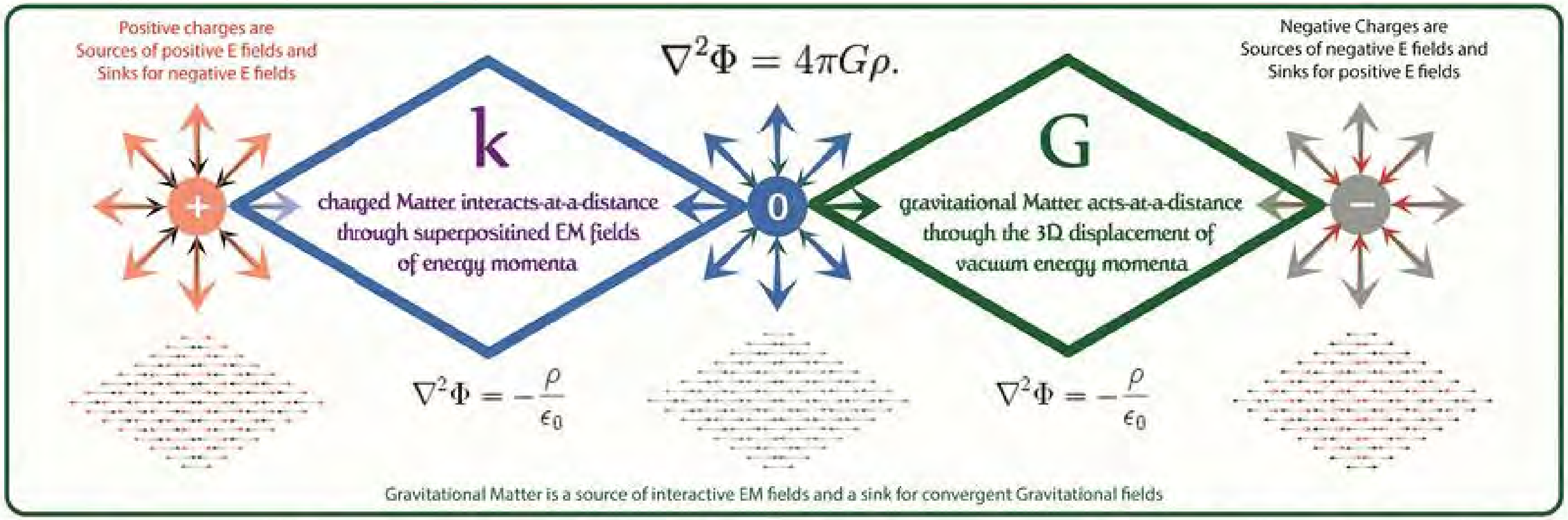

Positive charges are
Sources of positive E fields and
Sinks for negative E fields

$$\nabla^2\Phi = 4\pi G\rho.$$

Negative Charges are
Sources of negative E fields and
Sinks for positive E fields

charged Matter interacts-at-a-distance
through superpositined EM fields
of energy momenta

gravitational Matter acts-at-a-distance
through the 3D displacement of
vacuum energy momenta

$$\nabla^2\Phi = -\frac{\rho}{\epsilon_0}$$

$$\nabla^2\Phi = -\frac{\rho}{\epsilon_0}$$

Gravitational Matter is a source of interactive EM fields and a sink for convergent Gravitational fields

Tetryonic QM & QED clearly shows that both the sink and source fields are the result of bidirectional energy momenta

Poison's field equations shows that weak KEM field geometries can be used to model the convergent forces of Gravitational fields

Copyright ABRAHAM [2008] - All rights reserved

Gravito-Electro-Magnetism

inverse squared law

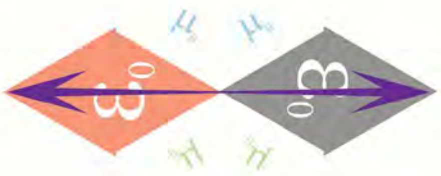

nett divergent
Electric field forces

$1/r^2$

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

k

$$\oint \vec{B}\cdot d\vec{A} = 0$$

Maxwell's 1864 achievements concerning electromagnetism have been called the "second great unification in physics", following on from the first one realised by Isaac Newton.

inverse cubed law

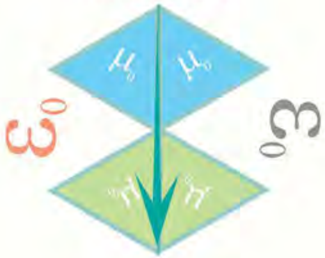

vector
Magnetic forces

$1/r^3$

Maxwell's equations
$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}$
$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
$\nabla \times \mathbf{B} = \frac{1}{\epsilon_0 c^2}\mathbf{J} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$

James Clerk Maxwell

(13 June 1831 – 5 November 1879)

Any opinion as to the form in which the energy of gravitation exists in space is of great importance, and whoever can make his opinion probable will have, made an enormous stride in physical speculation.

The apparent universality of gravitation, and the equality of its effects on matter of all kinds are most remarkable facts, hitherto without exception; but they are purely experimental facts, liable to be corrected by a single observed exception.

We cannot conceive of Matter with negative inertia or mass; but we see no way of accounting for the proportionality of gravitation to mass by any legitimate method of demonstration

GEM equations
$\nabla \cdot \mathbf{E}_g = -4\pi G \rho_g$
$\nabla \cdot \mathbf{B}_g = 0$
$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$
$\nabla \times \mathbf{B}_g = 4\left(-\frac{4\pi G}{c^2}\mathbf{J}_g + \frac{1}{c^2}\frac{\partial \mathbf{E}_g}{\partial t}\right)$

$1/r^2$

strictly convergent vector force

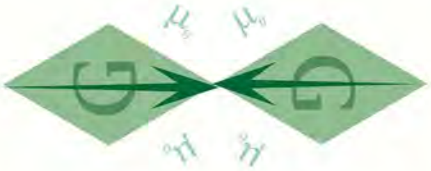

Maxwell proposed a set of equations based on his EM field laws that created a Gravito-Electro-Magnetic field reflective of the similarities between Newton's & Coulomb's Laws

$$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$$

G

$$\nabla \cdot \mathbf{B}_g = 0$$

giving rise to spectulation about the existence of 'gravitons'
a theorised Gravitational 'charge' carrier

$1/r^3$

gravitomagnetic dipole

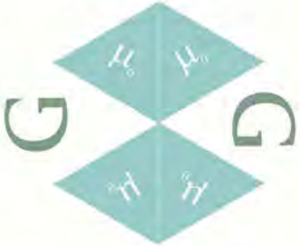

Copyright ABRAHAM [2008] - All rights reserved

General Relativity

The General Theory of Relativity

In 1915, Einstein extended SR and proposed a new theory of gravitation to account for accelerating bodies

Einstein said that when he "was sitting in a chair in the patent office at Bern when all of a sudden a thought occurred to me: 'If a person falls freely he will not feel his own weight.' I was startled. This simple thought ... impelled me toward a theory of gravitation. ..."

$$G_{ab} = 8\pi G T_{\mu\nu}$$

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

The EFE is a tensor equation relating a set of symmetric 4 x 4 tensors.

the scalar curvature

the cosmological constant

Newton's gravitational constant

Ricci curvature tensor

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

the stress-energy tensor

the metric tensor

the speed of light in vacuum

Each tensor has 10 independent components

Spacetime

Matter

$$G_{\mu\nu} + g_{\mu\nu}\Lambda = 8\pi T_{\mu\nu}$$

the cosmological constant

Albert Einstein

(14 March 1879 – 18 April 1955)

$$G_{ab} = \kappa T_{\mu\nu}$$

curved spacetime

mass-energy tensor

where

$$\kappa = -\frac{8\pi G}{c^4}$$

General Relativity describes gravity as a warping of space itself, not as a force.
Einstein pictured space as a three-dimensional version of a thin rubber sheet

Einstein's General Theory of Relativity, which was published in 1916 had as its foundation that the laws of nature in an accelerating frame are equivalent to the laws of a gravitational field.

This is known as the Equivalence Principle.

Copyright ABRAHAM [2008] - All rights reserved

Sir Issac Newton

(25 December 1642 – 20 March 1727)

GEM fields & Gravity

Gravity is one of the four fundamental interactions of nature,
along with electromagnetism and the nuclear strong and weak forces

[for centuries scientific enquiry has been aimed squarely at discerning a physical mechanism for gravitation]

$$\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

Albert Einstein

(14 March 1879 – 18 April 1955)

Tetryonics reveals Gravity to be the result of the displacement of vacuum energies by mass-Matter topologies which is accurately modelled using GEM field geometry

$$F = -G\,\overset{\text{masses}}{\frac{M_1 m_2}{r^2}}$$

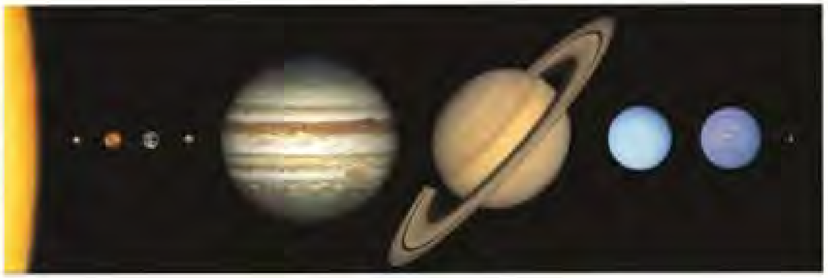

$$G_{ab} = \frac{8\pi G}{c^4}\overset{\text{energies}}{T_{\mu\nu}}$$

Gravity is responsible for keeping the Earth and the other planets in their orbits around the Sun;
for keeping the Moon in its orbit around the Earth; for the formation of tides; for natural convection,
by which fluid flow occurs under the influence of a density gradient and gravity;
for heating the interiors of forming stars and planets to very high temperatures;
and for various other phenomena observed on Earth.

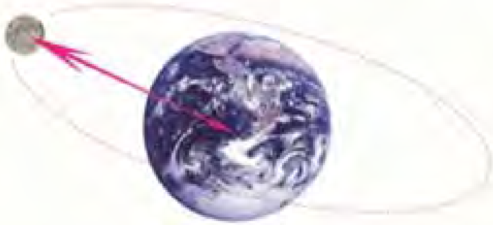

instantaneous
action-at-a-distance

Gravity is a natural phenomenon by which physical bodies attract each other with a force proportional to their mass Matter content.

Gravity is most familiar as the agent that gives weight to objects with mass and causes them to fall to the ground when dropped.

Gravity causes dispersed Matter to coalesce. and coalesced Matter to remain intact. to the creation of the SUN, the Earth, and to source of energy in the stars that light our universe.

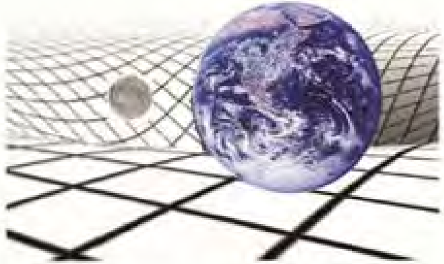

motion in
curved spacetime

Copyright ABRAHAM [2008] - All rights reserved

Separating Space & Time

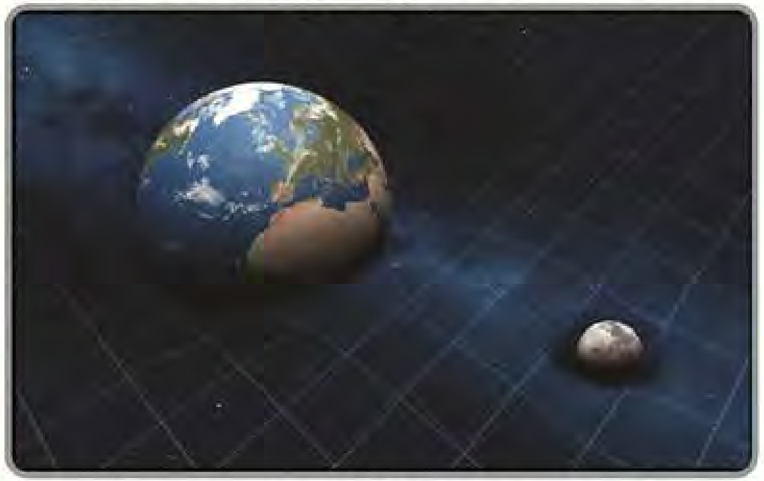

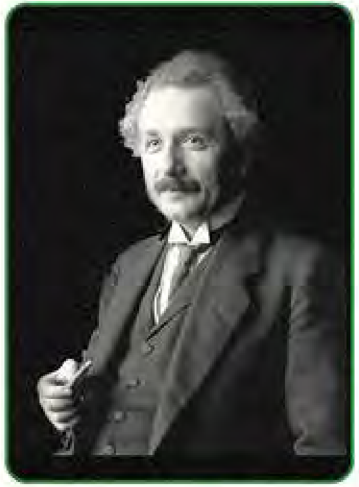

(in the search for Gravity)

Instantaneous
'action-at-a-distance'

Relativistic
spacetime 'curvature'

DOES

The Earth pull on the moon and
the moon pull proportionally back on the Earth

OR

Does the Earth curve spacetime in its vicinity
with the moon following the geodesic path created

Tetryonic theory through its clear definitions of charged mass-ENERGY-Matter provides the long-sought answer to Gravitation

Copyright ABRAHAM [2008] - All rights reserved

Empty Space

is defined in Tetryonic & relativistic theory
by the speed of light

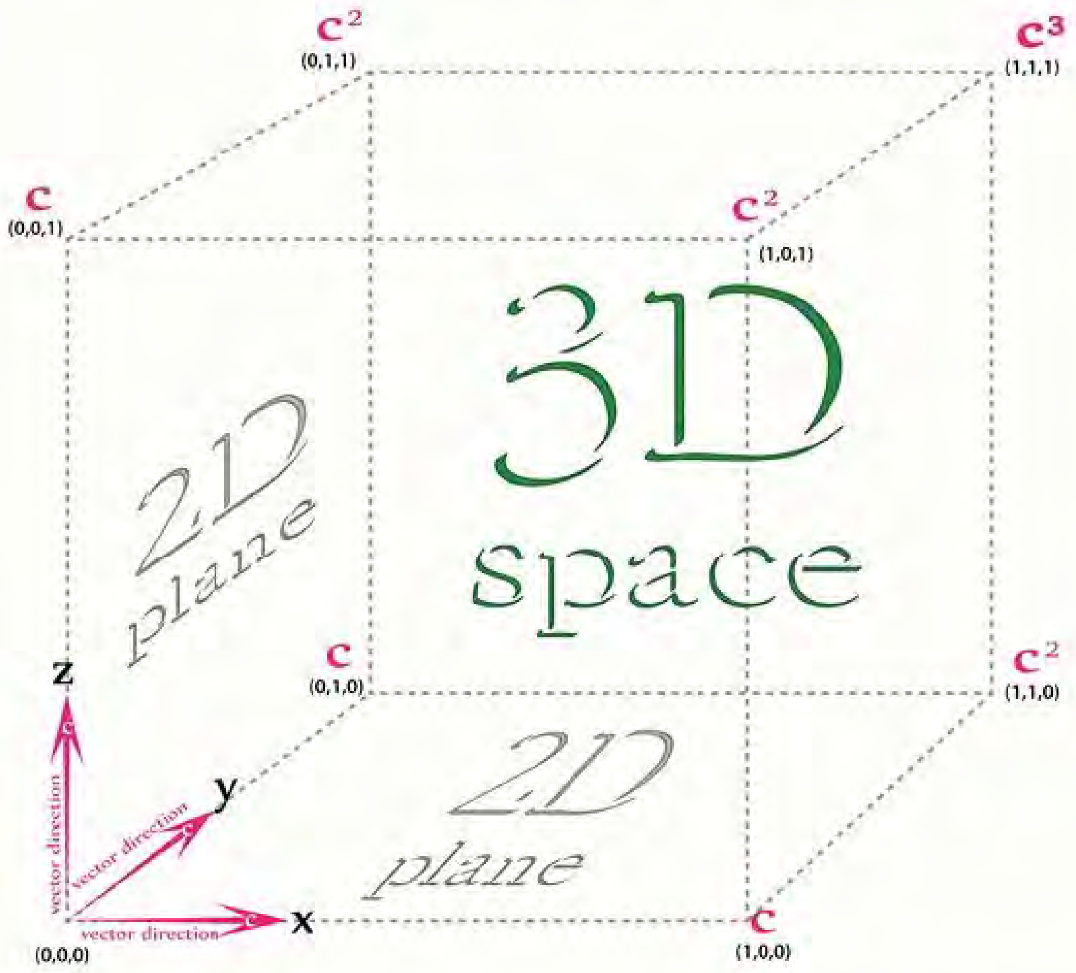
c^2
(0,1,1)
c^3
(1,1,1)
c
(0,0,1)
c^2
(1,0,1)
3D
space
2D
plane
z
c
(0,1,0)
c^2
(1,1,0)
y
2D
plane
vector direction
vector direction
vector direction
x
(0,0,0)
c
(1,0,0)

Energy

content of free space is Zero

A Spatial region is defined
by a co-ordinate system so as to
measure the physics and motion
of mass-ENERGY-Matter
within its confines

mass

is Energy content per unit of time

EM mass-energies can
be found in (and move through)
empty Space in various forms:
radiant energies, Matter etc

Matter

is Energy content per unit of time squared

Space can be any regular geometry
[Cubic, Spherical or polyhedral]
limited by the spatial co-ordinates
used to define the region or volume

Empty Space is defined as a geometric volume devoid of any form of Energy

Spacetime co-ordinates

The linear force vectors of divergent Energy
can form various co-ordinate systems
based on their vector directions
per unit of time

Charge radiates unilaterally
in the direction of its
vector linear momentum

Vector forces

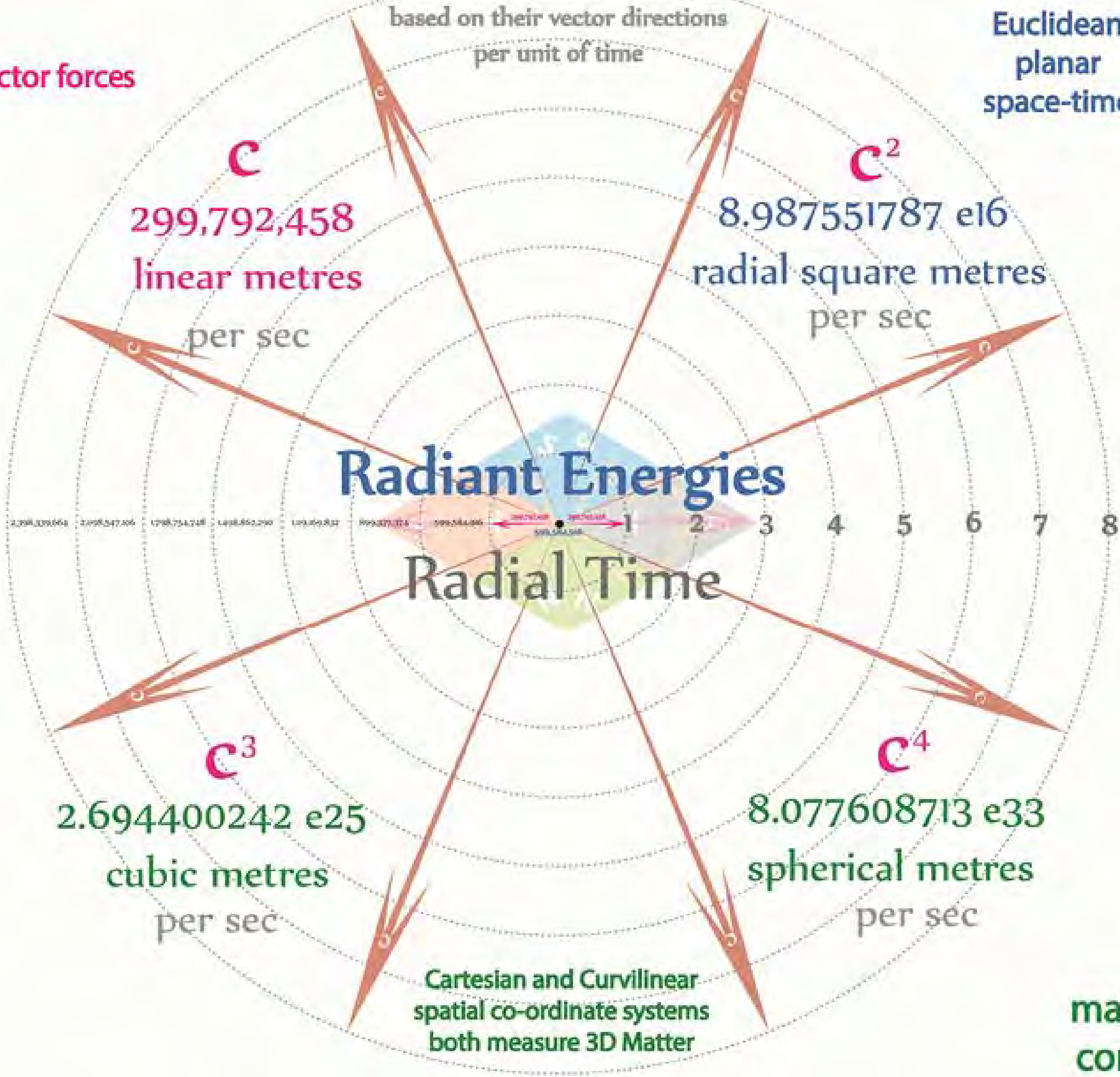

c
299,792,458
linear metres
per sec

c^2
8.987551787 e16
radial square metres
per sec

c^3
2.694400242 e25
cubic metres
per sec

c^4
8.077608713 e33
spherical metres
per sec

Radiant Energies

Radial Time

Cartesian and Curvilinear
spatial co-ordinate systems
both measure 3D Matter

Euclidean
planar
space-time

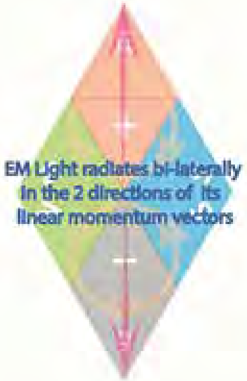

EM Light radiates bi-laterally
in the 2 directions of its
linear momentum vectors

Time in Physics is a measure
of how long it takes for light to
travel 299,792,458 metres
from its source in a vacuum

$$\frac{\text{299,792,458 m}}{\text{second}}$$

$$\frac{\text{metre}}{\text{299,792,458 sec}}$$

A metre in Physics is
the distance light travels in
1/ 299,792,458 of a second
from its source

Energy entropy
of a system

mass-energy-Matter
content of a system

Spatial co-ordinate geometries

based on the vector speed of light
form distinct spatial co-ordinate systems
for the measurement of physics

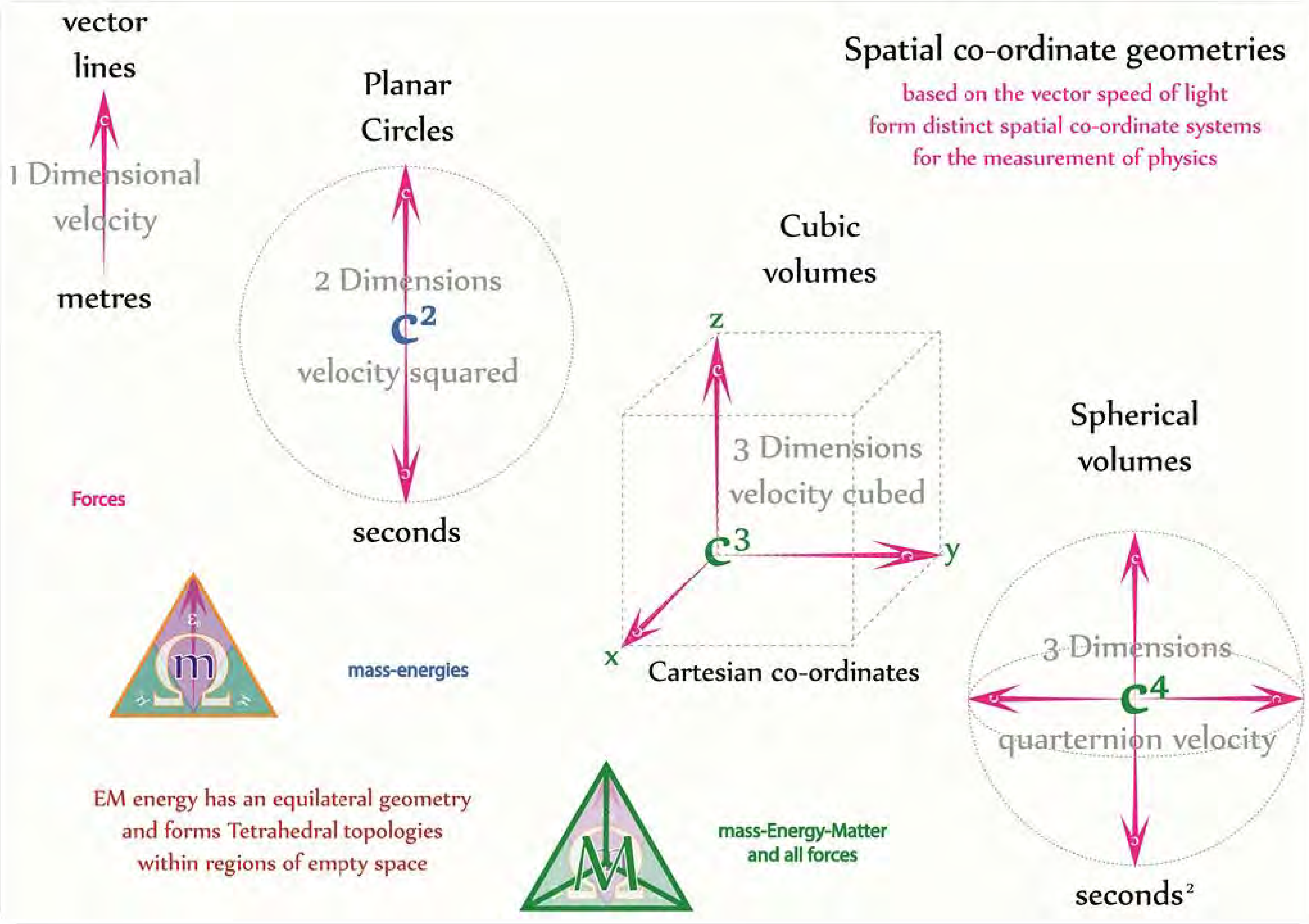

vector lines

1 Dimensional velocity

metres

Planar Circles

2 Dimensions

c^2

velocity squared

seconds

Cubic volumes

3 Dimensions

velocity cubed

c^3

Cartesian co-ordinates

Spherical volumes

3 Dimensions

c^4

quarternion velocity

seconds^2

Forces

mass-energies

mass-Energy-Matter
and all forces

EM energy has an equilateral geometry
and forms Tetrahedral topologies
within regions of empty space

Copyright ABRAHAM [2008] - All rights reserved

Spatial co-ordinate systems

Spacetime (or space-time, or space time) is any mathematical model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional and time playing the role of a fourth dimension that is different from the spatial dimensions.

Tetryonic theory redefines as time to be a measure of the changing quantised angular momenta of Planck scale Charges in any spatial plane or volume

By combining space and time into a single manifold, we can significantly simply a large number of physical theories, and re-formulate them in a more uniform way to explain the mechanics of the Universe at all scales

Tetryonics maps radial spacetime co-ordinates through the vector momentum of Energy

Cubic

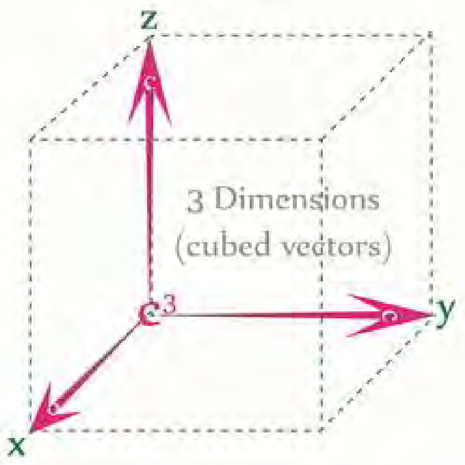

Spherical

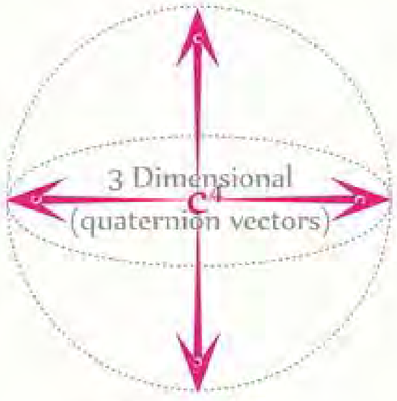

seconds^2

Cartesian Space-Time

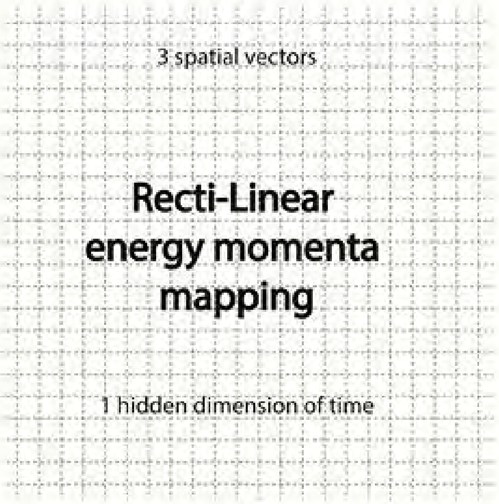

3 spatial vectors

Recti-Linear energy momenta mapping

1 hidden dimension of time

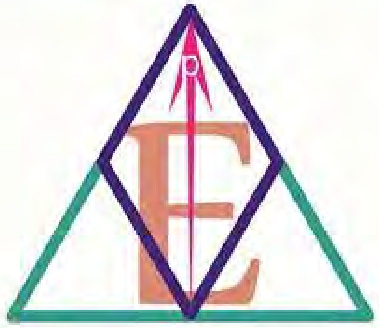

The quantised angular momenta of mass-Energy and Matter are triangular geometries

The current methodologies for mapping energy densities onto Euclidean, Cartesian and Reimannian co-ordinates are inaccurate reflections of the charged mass-ENERGY-Matter geometries, introducing a mathematical complexity to a otherwise simple foundational equilateral geometry

Tetryonic Space-Time

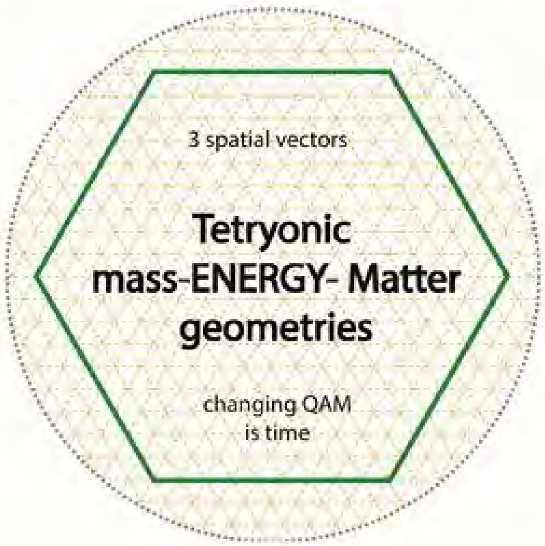

3 spatial vectors

Tetryonic mass-ENERGY- Matter geometries

changing QAM is time

By mapping equilateral mass-energies & Matter into radial spatial co-ordinates Tetryonic theory is able to reveal the hidden 4th dimension of time

Space & time separated

p All equilateral mass-energies & tetrahedral Matter can be mapped with planar spatial geometries within spherical space-time co-ordinates **Ω**

Space is the region of measurement defined by the vector propagation of light
time is a measure of the changing QAM within any spatial region

$T\pi$ standing-waves

c^4 spatial regions contain mass-Matter

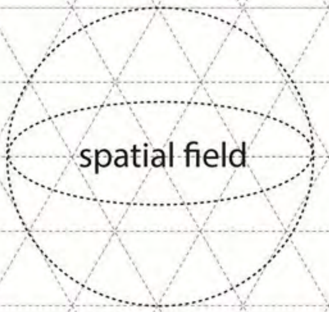
spatial field

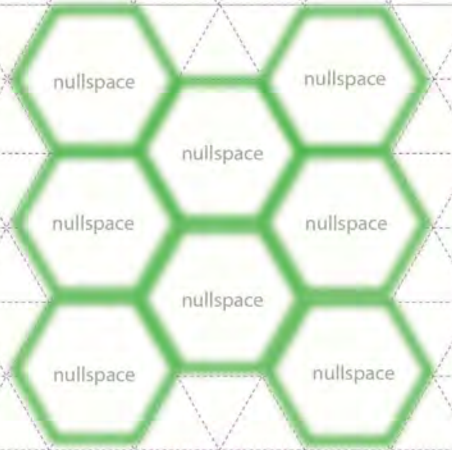
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace

compounds

c^2 spatial regions contain mass-energies

$n\pi$ radiant geometries

weak irradiated mass-energies permeate 'empty' space
[vacuum energies]

γ

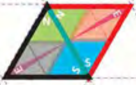

photons

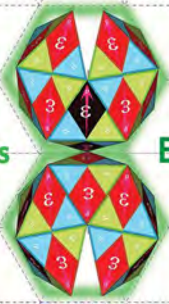

Baryons **Elements**

3D Matter
topologies

Molecules

W+

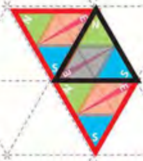

2D mass-energy
geometries

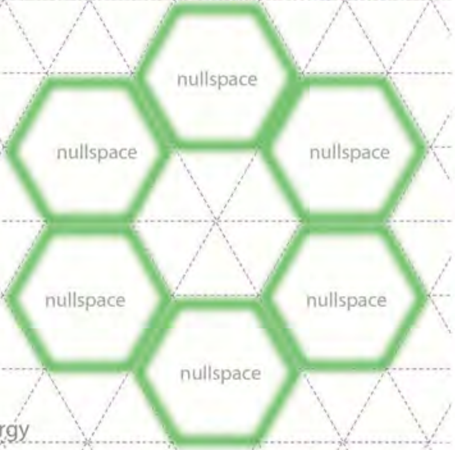
nullspace
nullspace
nullspace
nullspace
nullspace
nullspace

+q

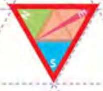

photons

bosons

charges

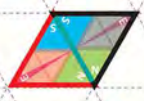

γ

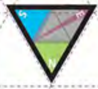

-q

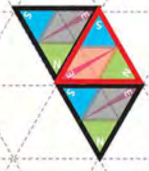

W-

'nullspace' is a spatial region devoid of all forms of energy
[found only within tetrahedral Matter topologies]

Copyright ABRAHAM [2008] - All rights reserved

2D Euclidean spatial geometry

All 2D mass-energy momenta map onto **FLAT** Euclidian co-ordinates

[x,y] ΔΩ
(2D space & time)

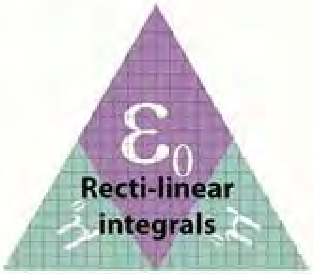

ε_0
Recti-linear
integrals

Cartesian spatial co-ordinates
fail to accurately map
equilateral energy momenta

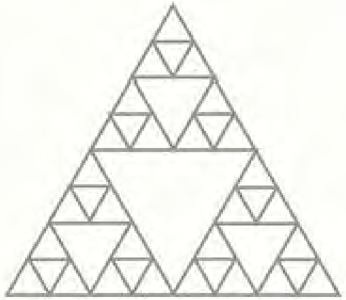

Minkowski spacetime
is a close approximation of
Tetryonic energy momenta

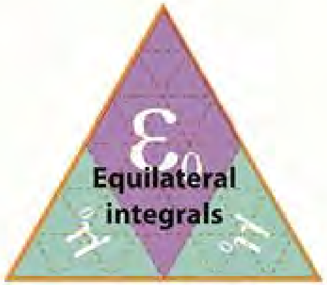

ε_0
Equilateral
integrals

Tetryonic space-time

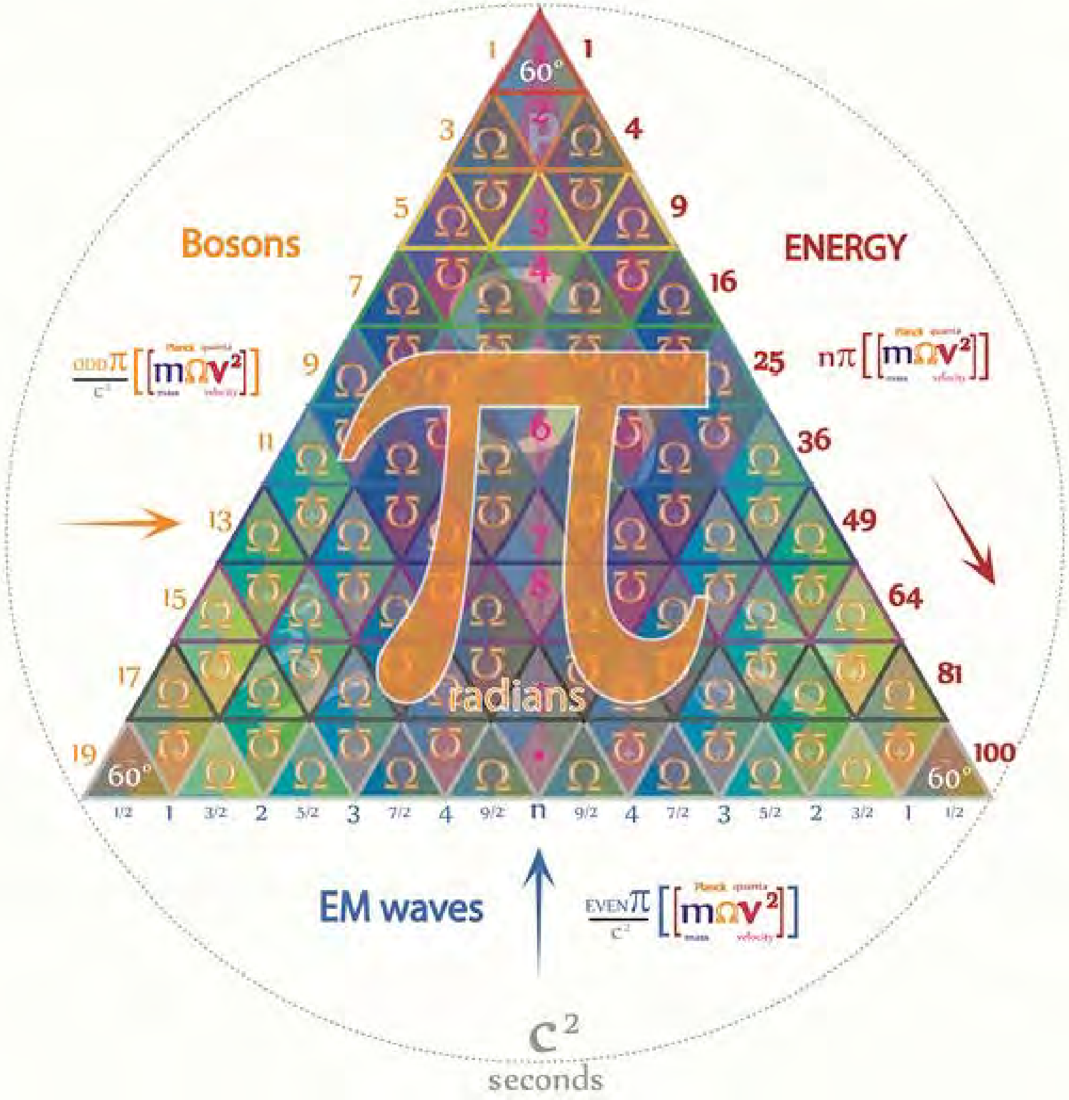

1
60°
1
3
4
5
9
7
16
9
25
36
11
13
49
15
64
17
81
19
100
60°
60°
Ω
π
radians
1/2
1
3/2
2
5/2
3
7/2
4
9/2
n
9/2
4
7/2
3
5/2
2
3/2
1
1/2
Bosons
ENERGY
$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$
$n\pi\left[\left[m\Omega v^2\right]\right]$
$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$
Planck quanta
mass
velocity
EM waves
c^2
seconds

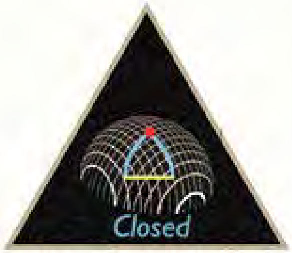

negative curvative
angles add up to > 180 °

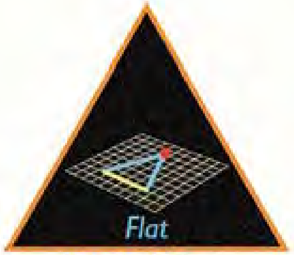

flat Euclidean
angles = 180 °

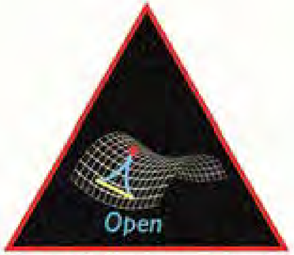

positive curvature
angles add up to < 180 °

Copyright ABRAHAM [2008] - All rights reserved

Light cones space & time

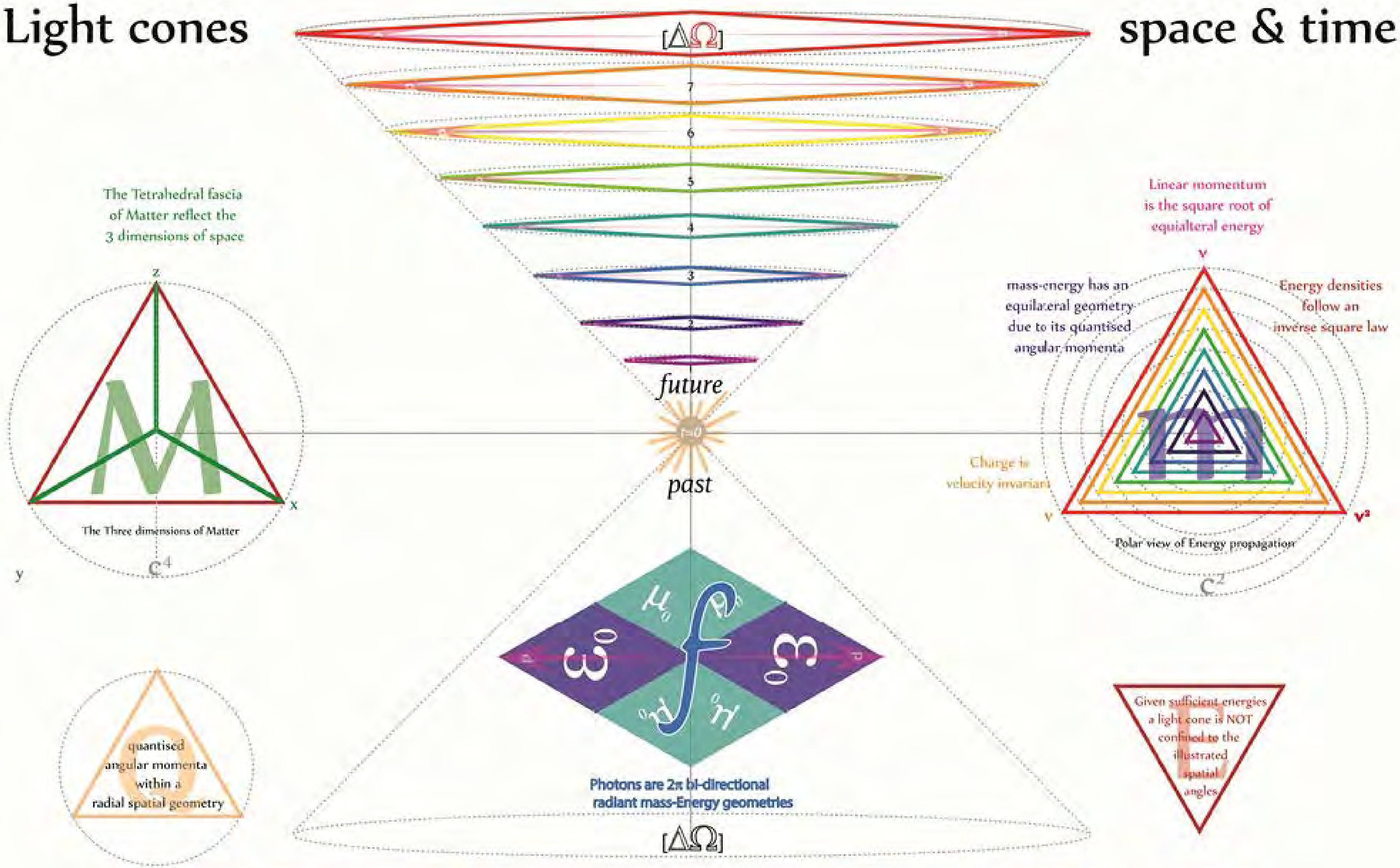

The Tetrahedral fascia
of Matter reflect the
3 dimensions of space

The Three dimensions of Matter

Linear momentum
is the square root of
equialteral energy

mass-energy has an
equilateral geometry
due to its quantised
angular momenta

Energy densities
follow an
inverse square law

Charge is
velocity invariant

Polar view of Energy propagation

quantised
angular momenta
within a
radial spatial geometry

Photons are 2π bi-directional
radiant mass-Energy geometries

Given sufficient energies
a light cone is NOT
confined to the
illustrated
spatial
angles

time is the measurement of the changing quantised angular momenta in any spatial region

A light cone is the path that a flash of light, emanating from a single event
(localized to a single point in space and a single moment in time) and traveling in all directions,
would take through space & time.

Tetryonics 62.08 - Minkowski Light Cones

Charge, energy momenta & time

measurements of the changing quantised angular momenta of scalar energy is the basis of measuring time

Ω

$\boxed{.0012}$

$\frac{m^2}{s}$

quantised angular momentum

$\beta = \left[\frac{v}{c}\right]$

Linear Lorentz factor

metres

$\frac{m^2}{s} \cdot \frac{s}{m}$

$\boxed{\frac{\Omega}{v}}$

λ

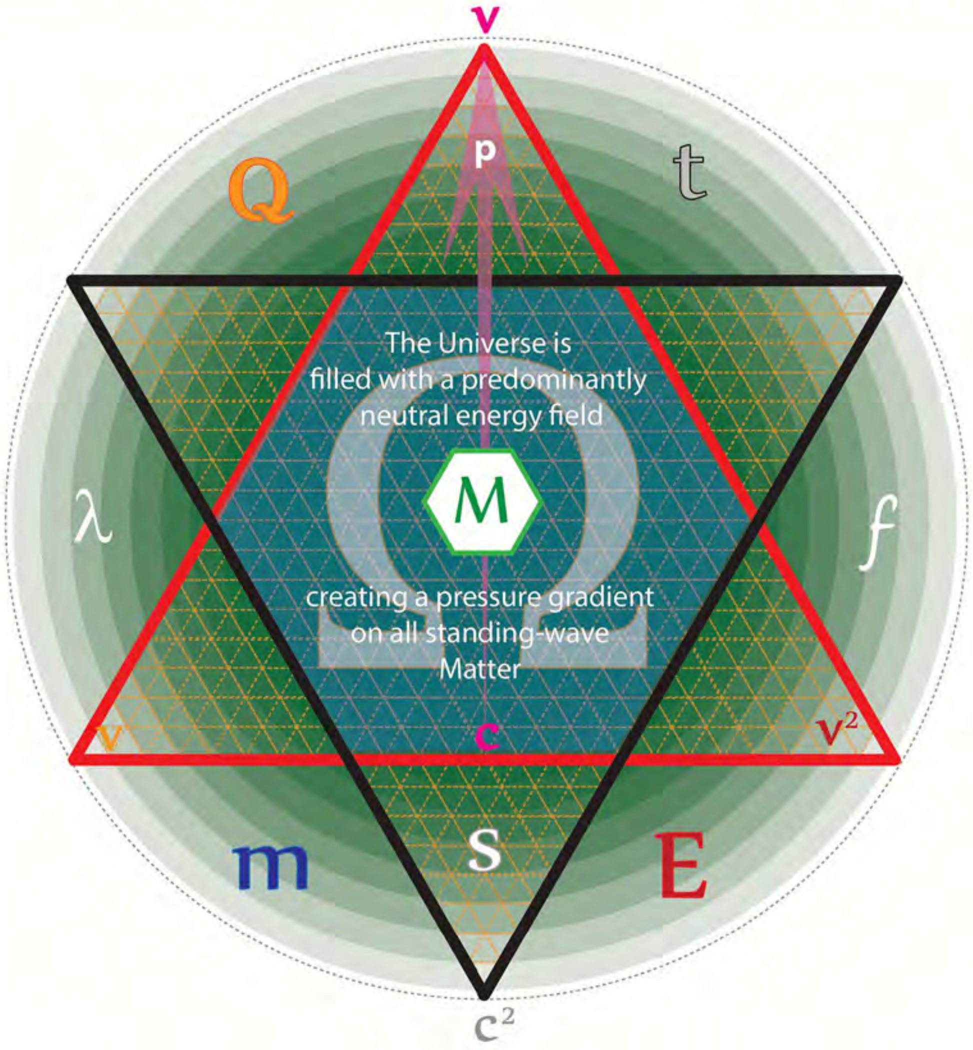

The Universe is filled with a predominantly neutral energy field

creating a pressure gradient on all standing-wave Matter

q

$\boxed{\frac{\Omega}{c^2}}$

$\frac{m^2}{s} \cdot \frac{s^2}{m^2}$

seconds

$\left[\frac{v^2}{c^2}\right] = \beta^2$

Scalar Lorentz factor

per second

$\frac{m^2}{s^2} \cdot \frac{s}{m^2}$

$\boxed{\frac{v^2}{\Omega}}$

f

Velocity of Light

$$c = \left[\underbrace{\overset{\lambda}{\left[\frac{\Omega}{c}\right]}}_{\text{Wavelength}} \cdot \underbrace{\overset{f}{\left[\frac{c^2}{\Omega}\right]}}_{\text{Frequency}} \right]$$

Copyright ABRAHAM [2008] - All rights reserved

$\vec{p}$

Spatial co-ordinate systems are defined by the vectors of their linear energy momentum

Space
[x,y,z]

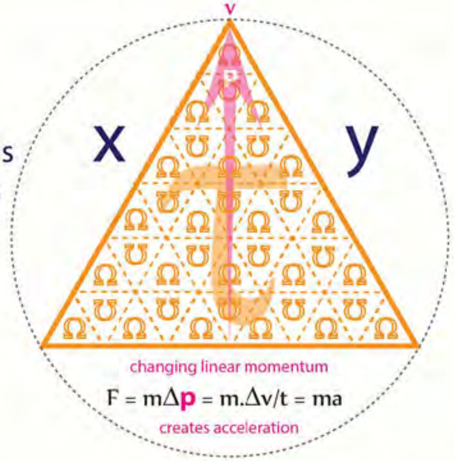

changing linear momentum

$F = m\Delta p = m.\Delta v/t = ma$

creates acceleration

$\Delta\Omega$

Time is a measure of the changing quantised angular momenta in any defined region of space

Time
[seconds]

Spacetime Fields

are spatial geometries that contain changing Quantised Angular energy-Momenta [Planck quanta]

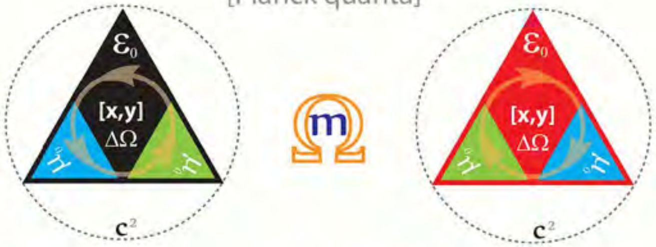

Charge is the nett distribution of Planck quanta mass.[Quantised Angular Momenta] within any spatial co-ordintate system

EM masses are 2D planar energy waveforms

2D mass-energy quanta have linear & angular momenta

Matter has a 3D volumetric energy waveform

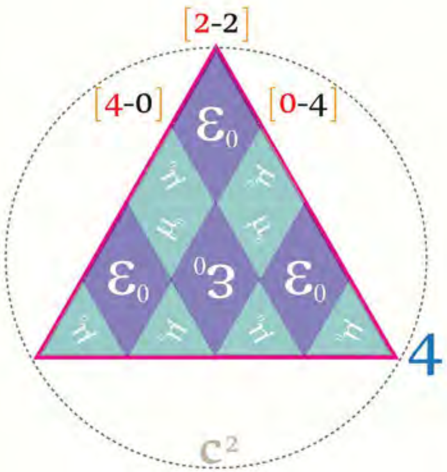

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

Tetryonic geometry restores symmetry to Einstein's mass-energy equivalence and reveals a mass-ENERGY-Matter equivalance

radiant mass-energies can form standing-wave Matter

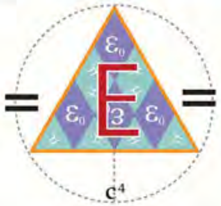

$$\frac{m}{c^2} = E = M$$

standing-wave Matter can release radiant mass-energies

where space and time are separated

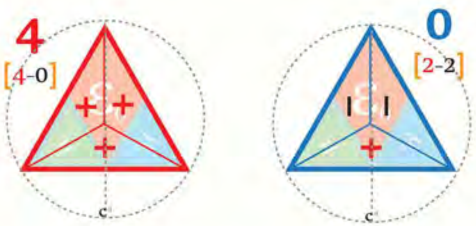

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

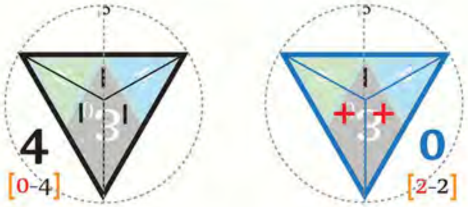

Copyright ABRAHAM [2008] - All rights reserved

The 4th dimension of Time

is charge, the angular momenta associated with any planar mass-energy

Accelerate

$\overrightarrow{p}$

increasing a systems's linear momentum

$F = m\Delta p = m.\Delta v/t = ma$

creates acceleration

Deccelerate

$\overleftarrow{p}$

decreasing a systems's linear momentum

$F = m\Delta p = m.\Delta v/t = ma$

creates decceleration

The square root vector force of Planck mass-energies createss linear momentum

The equilateral geometry of Planck mass-energy is Quantised Angular Momenta

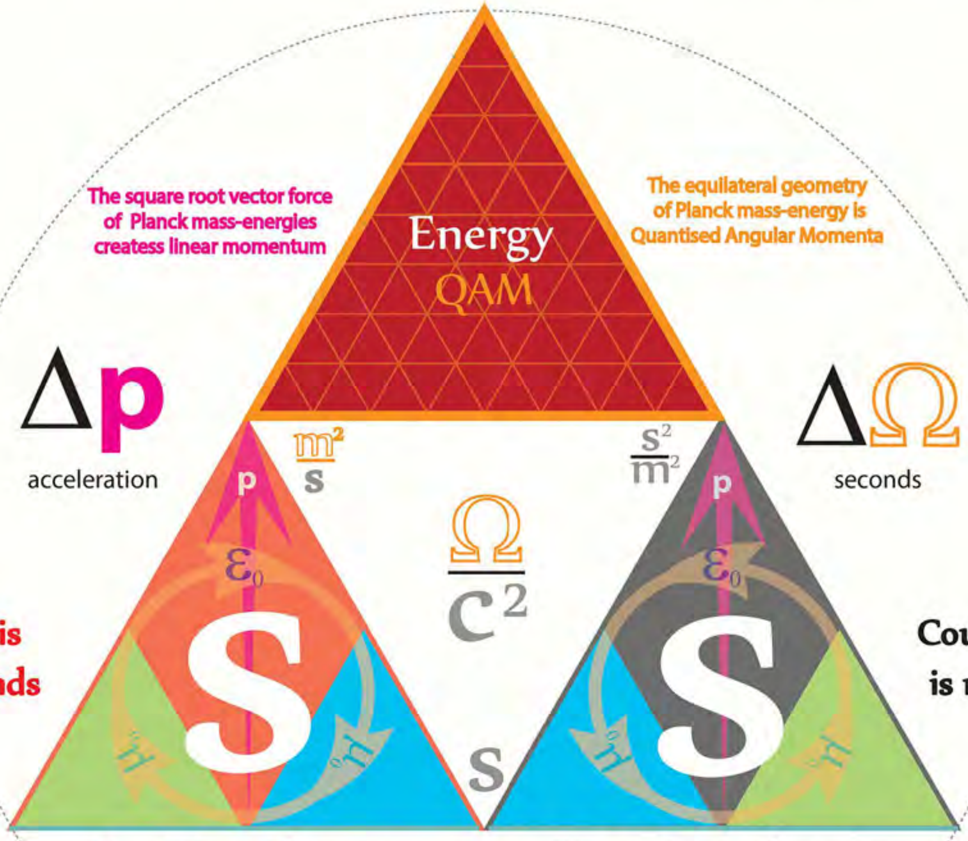

Δp

acceleration

$\Delta\Omega$

seconds

Clockwise charge is positive mass-seconds

Counter-clockwise charge is negative mass-seconds

Positive charge

Negative charge

the two 'arrows' of time in physics are a mathematical mis-interpretation of charge

Positive and Negative charges on opposite sides of Planck energy quanta, the direct result of the equilateral geometry of all energy momenta has been historically mis-represented in mathematics as being two directional arrows of time

$\Delta\Omega$

Positive Time

Time is a measure of the changing angular momenta-energy (Planck geometry) of any given spatial region

$\Delta\Omega$

Negative Time

CHARGE

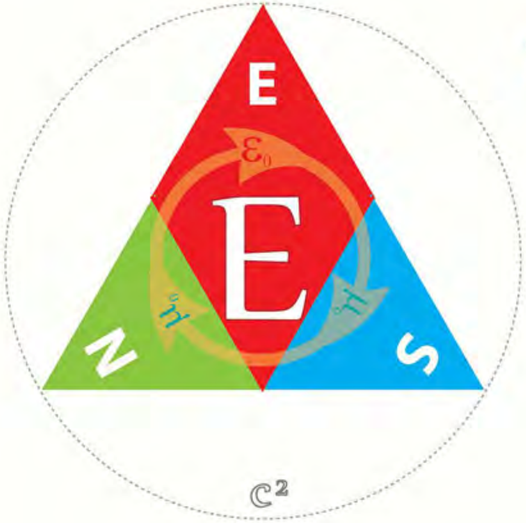

all ideal inductive loops
resist changes to their
Energy levels

INERTIAL mass

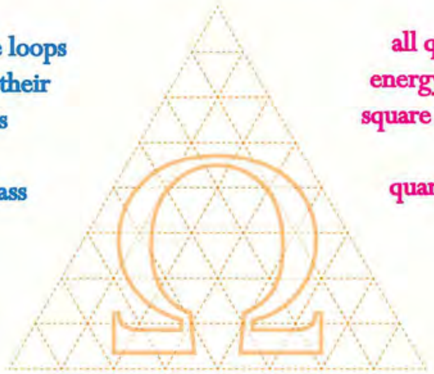

All quantised angular momentum
is an equilateral geometry

all quantised Planck
energy momenta posses
square root vector Forces

quantum of WORK

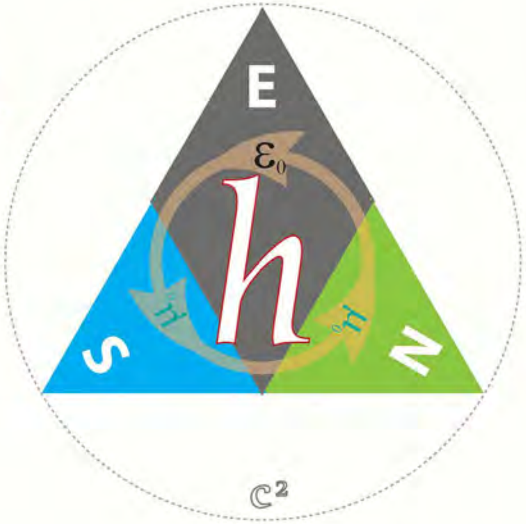

vector velocity results from
linear momentum - the square root
of all scalar energies

$$\mathrm{kg}\left[\frac{m^2}{s^2}\;\frac{s^2}{m^2}\right]$$

Scalar energy
per unit time

mass

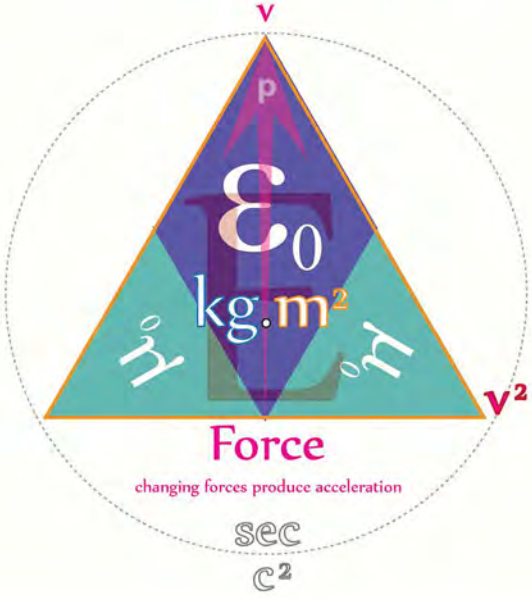

Force

changing forces produce acceleration

$$\mathrm{kg}\left[\frac{m^2}{s}\;\frac{s^2}{m^2}\right]$$

Planck quanta
per unit time

mass.seconds

Copyright ABRAHAM [2008] - All rights reserved

Positive charged mass-energy

Charge is the equilateral geometry of Energy that gives form to all mass & Matter

Negative charged mass-energy

1
[1-0]

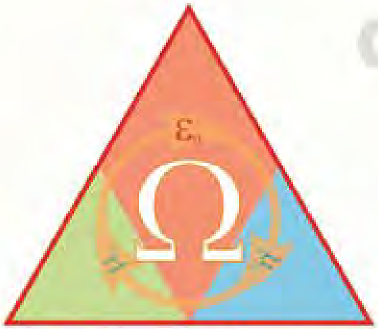

Clockwise inductive energy flux

charge

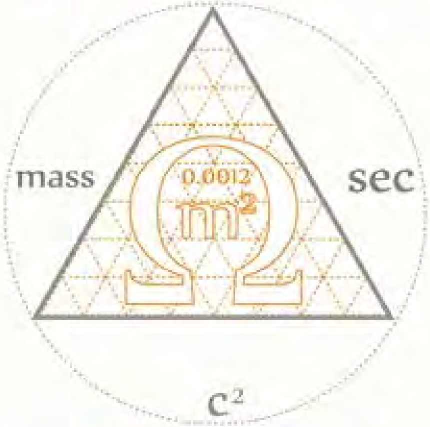

time

1
[0-1]

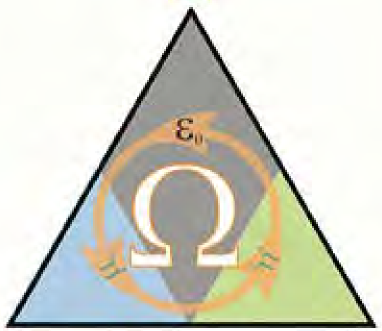

Counter clockwise energy flux

mass.seconds in physics is expressed as Coulombs [charge]

ε_0
m
Ω

ε_0
m
Ω

t

$$\frac{\Delta\Omega}{c^2}$$

$$\frac{m^2}{s}\,\frac{s^2}{m^2}$$

s

seconds

t

ElectroMagnetic Charge is a quantum property resulting from the equilateral QAM geometry of mass-energy

1.33518 e-20 s

Changes to nett quantised angular momenta [charged mass-energies] within any spatial co-ordinate system forms the basis of time

Two ElectroMagnetic charges are possible & both are created through the tesselation of EM energy momenta in ideal inductive loops

Copyright ABRAHAM [2008] - All rights reserved

The Quantum Arrow of Time

in physics is a measurement of changes to quantised angular momenta

Thermodynamic arrow of time

The Entropy of a system always increases

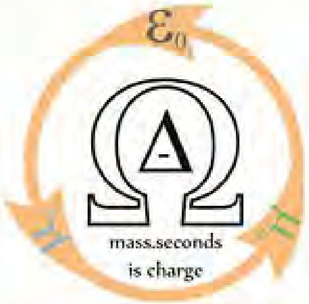

ε_0
Ω
Δ
-
mass.seconds
is charge

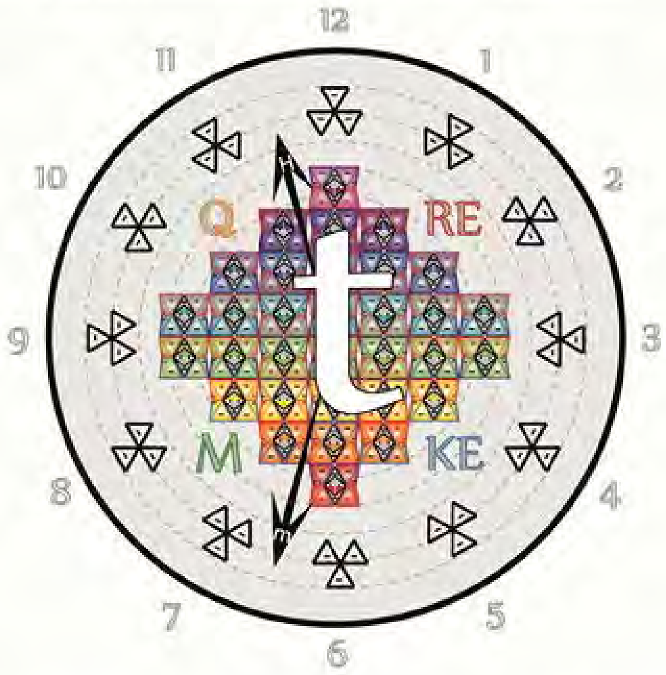

12
1
2
3
4
5
6
7
8
9
10
11
Q
RE
M
KE
t

Cosmological arrow of time

The Universe is evolving from a big bang to ?

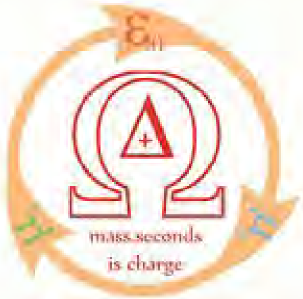

ε_0
Ω
Δ
+
mass.seconds
is charge

Radiative arrow of time

EM mass-energies rative outward from a point source

Causal arrow of time

A cause always precedes its effect

Tetryonic theory maps QAM as equilateral Planck geometries & reveals ΔQAM to be time
Time and the spatial dimensions it creates are mapped with radial & spherical co-ordinates

ΔmΩ

Decceleration

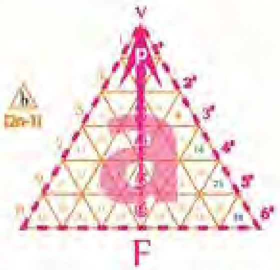

The Arrow of Time

is our perception of changing quantum-scale angular momenta

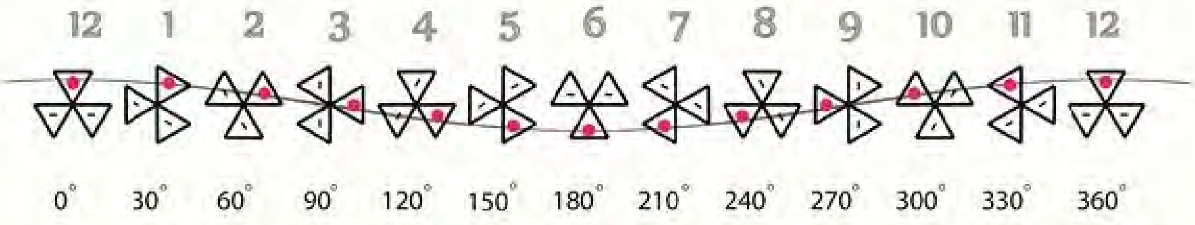

12
1
2
3
4
5
6
7
8
9
10
11
12
0°
30°
60°
90°
120°
150°
180°
210°
240°
270°
300°
330°
360°

ΔmΩ

Acceleration

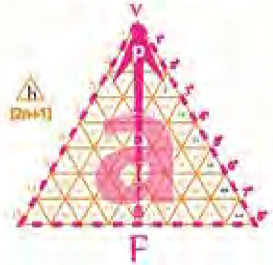

Copyright ABRAHAM [2008] - All rights reserved

Time dilation

An accurate clock at rest with respect to one observer may be measured to tick at a different rate when compared to a second observer's own equally accurate clocks.
This effect arises neither from technical aspects of the clocks nor from the fact that signals need time to propagate, but from the nature of spacetime itself.

The changing quantised angular momenta of
all mass-energy-Matter densities in any spatial region
all impact on our measurement of time [ΔΩ]

increasing the mass of Matter
increases the stress energy tensor
[mass-energy gradient]

Stronger G fields

GR

Vacuum energy-Matter
pressure gradients

Weaker G fields

gravity is proportional to
the material displacement volume of mass-Matter
in a given spatial region of vacuum energy

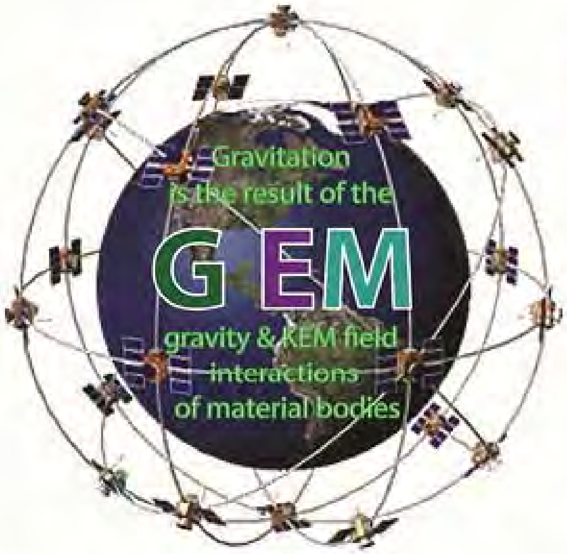

tests of GR using
photons are erroneous
tests of SR

accelerating material objects
increases the Planck quanta
[mass.QAM] in their KEM fields

Stronger KEM fields

SR

Kinetic EM energies
of motion

Weaker KEM fields

decclerating material objects
decreases the Planck quanta
[mass.QAM] in their KEM fields

Any attempt to accurately model the full effections of Gravitation must include
a complete definition and differentiation of all mass-energy & Matter
along with the vacuum energy pressure gradient created by them

In the theory of relativity, time dilation is an actual difference of elapsed time between two events as measured by observers either moving relative to each other or differently situated from gravitational masses.

Entropy in physical systems

is the result of changes to the net organisation of quantised angular momenta as it seeks equilibrium

Second law of thermodynamics

An isolated physical system, if not already in its own internal state of thermodynamic equilibrium, spontaneously evolves towards it.
In an isolated physical system, there is a tendency towards spatial homogeneity.

The Boltzmann equation was
developed to describe
the dynamics of an ideal gas

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + \frac{F}{m}\frac{\partial f}{\partial v} = \frac{\partial f}{\partial t}$$

Ludwig Eduard Boltzmann

(February 20, 1844 – September 5, 1906)

$$S = k{\cdot}[\log(W)]$$

Boltzmann entropy formula

Tetryonic theory utilises
the equilateral geometry of
ideal quantum inductors
to model EM energies

mass-energy radiator

normal statisitcal distribution of energies

Energy in all its forms seeks equilibrium

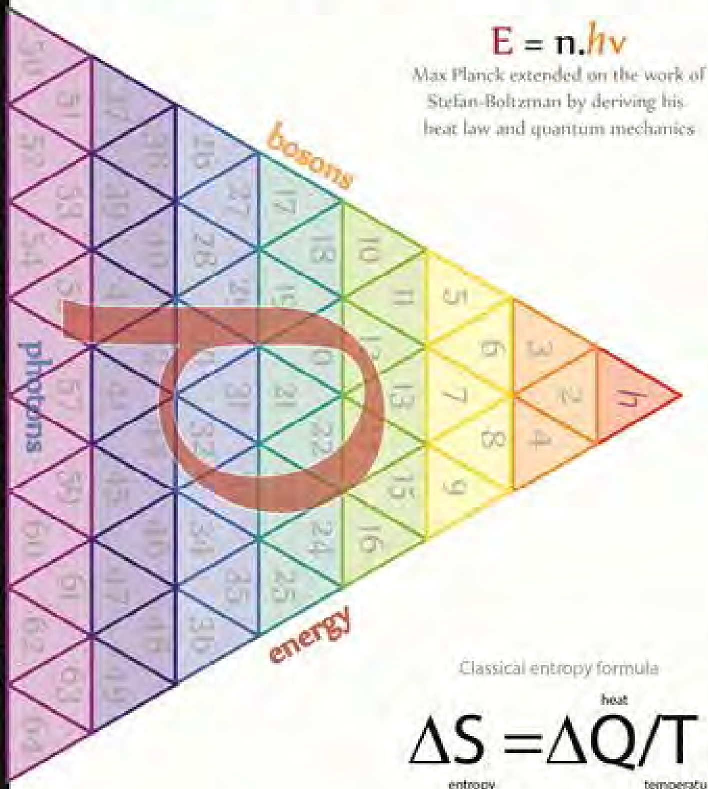

$E = n.hv$

Max Planck extended on the work of
Stefan-Boltzman by deriving his
heat law and quantum mechanics

Classical entropy formula

$$\underset{\text{entropy}}{\Delta S} = \Delta \overset{\text{heat}}{Q} / \underset{\text{temperature}}{T}$$

Energy always moves from higher energy organised systems
to a lower dis-organised system

lower energy blackbody

A black body is an idealized physical body that absorbs all incident electromagnetic radiation

Jacob David Bekenstein

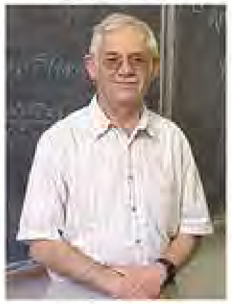

(born May 1, 1947)

Bekenstein-Hawking entropy formula

$$S = \frac{\pi A k c^3}{2hG}$$

still uses 3D [c³]
classical cartesian
co-ordinates

Stephen William Hawking

(born 8 January 1942)

Tetryonic theory unites classical and quantum mechanics with relativity through equilateral QAM

Copyright ABRAHAM [2008] - All rights reserved

The Aether

Aether theories in physics propose the existence of a medium, the aether
a space-filling substance or field, thought to be necessary as a transmission
medium for the propagation of electromagnetic waves.

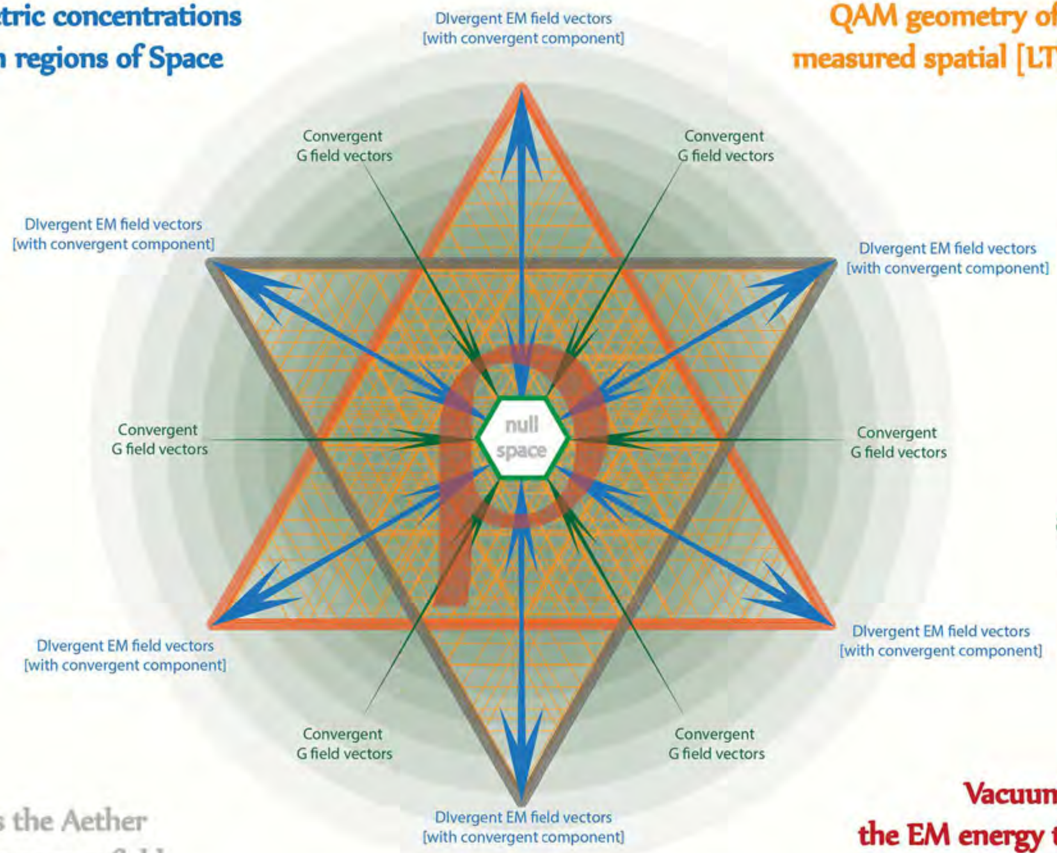
Divergent EM field vectors [with convergent component]
Convergent G field vectors
Convergent G field vectors
Divergent EM field vectors [with convergent component]
Divergent EM field vectors [with convergent component]
Convergent G field vectors
null space
Convergent G field vectors
Divergent EM field vectors [with convergent component]
Divergent EM field vectors [with convergent component]
Convergent G field vectors
Convergent G field vectors
Divergent EM field vectors [with convergent component]

**bosons, photons and Matter
are simply geometric concentrations
of EM energy in regions of Space**

**Spacetime is the
QAM geometry of any
measured spatial [LT] region**

**Vast radiant energy fields
exist throughout all of Space
save inside Matter topologies**

**Matter geometries form
perfect quantum Faraday cages
excluding all energies
thus creating nullspaces**

**The low level EM fields
permeating the space between
Matter forms a background
vacuum energy density**

**The vacuum energy field
acts upon the internal 'null space'
volumes of all Matter creating a
gravitational pressure gradient**

Historically refered to as the Aether
(or the Ether) this vacuum energy field
has been the source of much contention

**Vacuum Energy is
the EM energy that permeates and
fills any spatial region**

**Photons do not require the aether to propagate however
the aether is permeated with 'weak' superpositioned EM fields**

Copyright ABRAHAM [2008] - All rights reserved

The Luminiferous Aether

$$\frac{\rho}{m^2}$$

Photons are radiant 2D mass-energy geometries

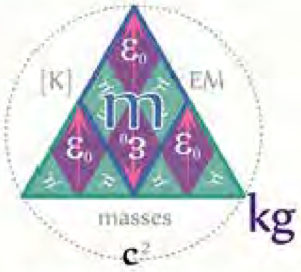

Vacuum energies are very long wavelength photons

Hendrik Lorentz created an electron/aether theory, in which he introduced a strict separation between Matter and the aether

In the late 19th century, luminiferous aether or ether, meaning light-bearing aether, was proposed as a medium for the propagation of light.

Newton's Opticks (1704) postulated an "Aethereal Medium" transmitting vibrations faster than light, by which light, is put into "Fits of easy Reflexion and easy Transmission"

Maxwell proposed a mechanical sea of molecular vortices to explain the transfer of ElectroMagnetic energies

Einstein wrote that one can actually speak about a "new aether", but one may not speak of motion in relation to that aether

In Tetryonics all EM radiation & Matter are revealed to be concentrations of geometric Energies seeking equilibrium in regions of Spacetime

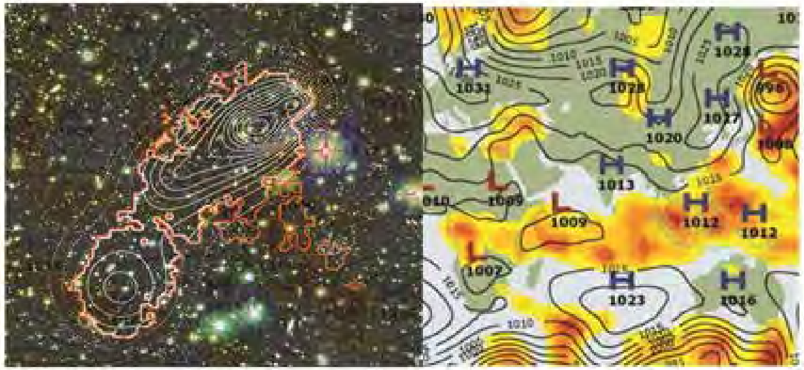

The aether does NOT facilitate the transmission of Energies throughout the Universe, it is the result of radiated [K]EM energies

Vacuum energies can be likened to the mean atmospheric pressure, with 2D mass-energies being HIGHER pressure 'gusts'

and

the nullspaces within 3D Matter creating a LOW pressure gradient [providing a driving mechanism akin to to that of Storms, Cyclones or Tornados]

$$\frac{\rho}{m^3}$$

All Matter are 3D mass-energy topologies containing nullspaces

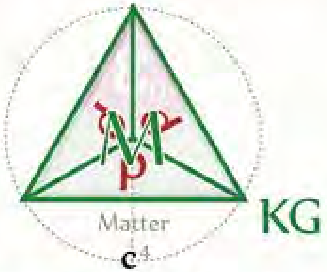

Planets, Stars and Galaxies contain high densities of mass-energies

Michaelson and Morley sought to detect the Relative motion between the Earth and aether by measuring the speed of light

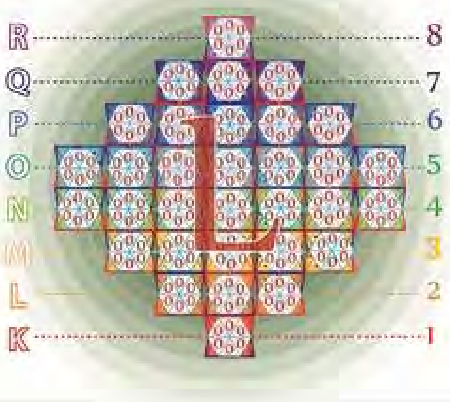

Copyright ABRAHAM [2008] - All rights reserved

Vacuum Energy

Vacuum energy is an underlying background mass-energy that exists
in free space even when the space is devoid of matter

The Universe is Electric in nature
and all Matter in it interacts via
Gravito-Electro-magnetic [GEM] fields

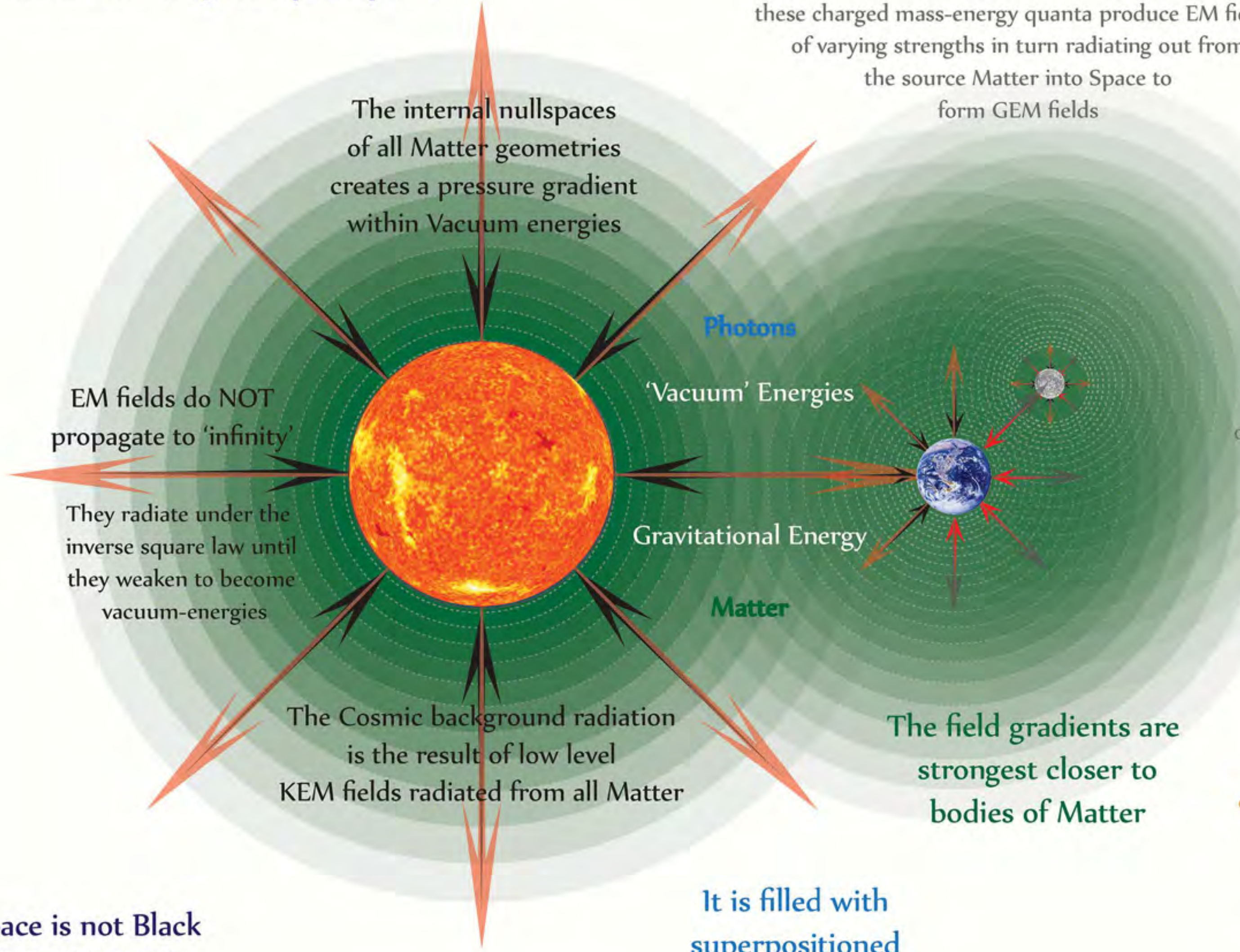

The internal nullspaces
of all Matter geometries
creates a pressure gradient
within Vacuum energies

Tetryonics shows that all Matter is quantised and
these charged mass-energy quanta produce EM fields
of varying strengths in turn radiating out from
the source Matter into Space to
form GEM fields

EM fields do NOT
propagate to 'infinity'

They radiate under the
inverse square law until
they weaken to become
vacuum-energies

The Cosmic background radiation
is the result of low level
KEM fields radiated from all Matter

The field gradients are
strongest closer to
bodies of Matter

v
P
Q
t
λ
M
Ω
f
c
v²
m
s
E
c²

VE is not a 'quantum foam' as
popularly depicted, it is a
triangular Euclidean geometry
of neutral mass-energies

There are no 'virtual particles' in free space
only varying mass-Energy density geometries
[Bosons, Photons & Matter]

Coulombic fields are DIVERGENT
[K]EM fields with interactive forces

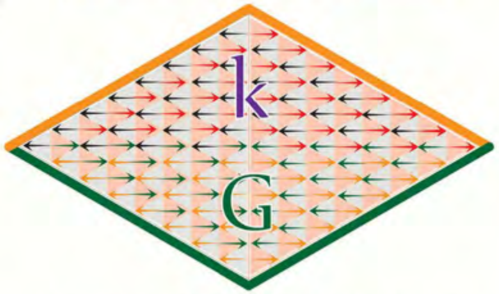

Gravitational fields are CONVERGENT
Gravity is always attractive

Space is not Black
[it is INDIGO]

It is filled with
superpositioned
EM fields

Copyright ABRAHAM [2008] - All rights reserved

The Higgs Field

is another name for the Aether
(or vacuum energy field) that fills
free Space, its interaction with
inductive Tetryonic topologies
creates the inertial properties
observed in mass-Matter

Peter Higgs

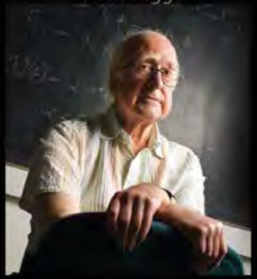

(born 29 May 1929)

Free Space

is any spatial region
free of EM energy quanta
(nullspace within Tetryons)

The 'Higgs Boson' is
just a W boson by
another name

All Higgs vacuum fields
are equilateral mass-
Energy-momenta
fields

Vacuum Energy

is the lowest denisty [longest wavelength]
electromagnetic mass-energy that
having diverged from Matter
radiates outwards to fill space
& create a pressure gradient

'Virtual particles'
do NOT exist

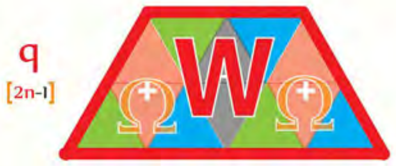

All Bosons have ODDn planar electromagnetic charge geometries

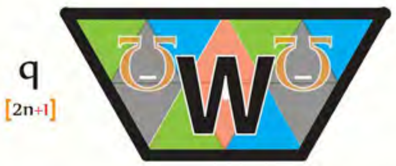

Inductive resistance to changes of the energy content of charged fascia through vacuum energy fields creates inertial mass

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

Electromagnetic mass
is a measure of the Energy
content of a 2D geometry
per unit of Time

$$m$$

mass-energy is a 2D geometric property of 3D Matter

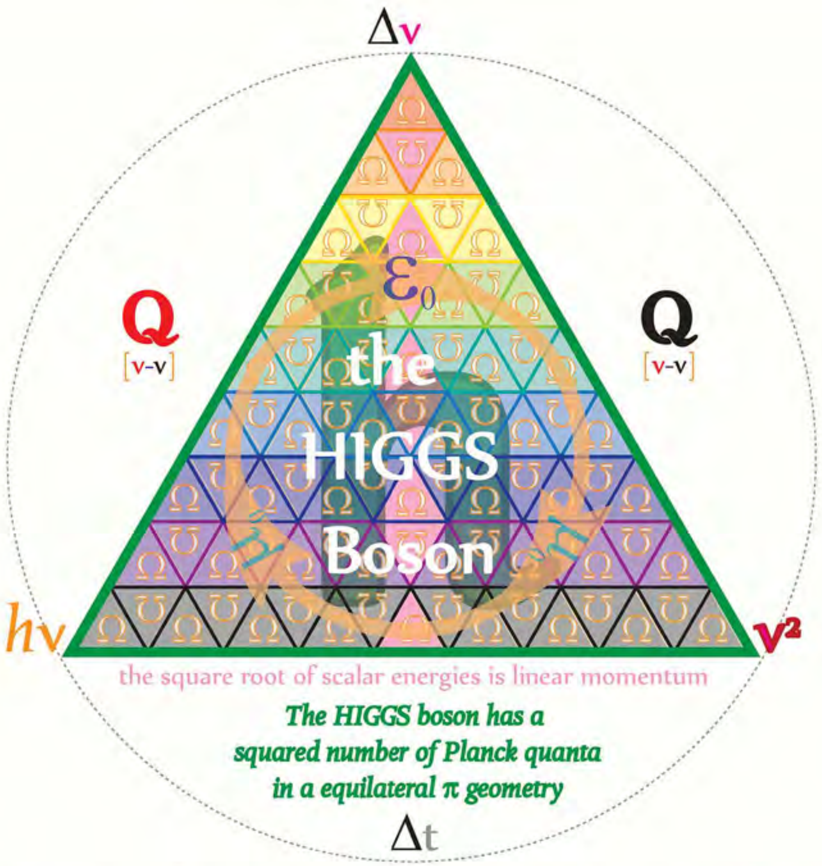

the square root of scalar energies is linear momentum

The HIGGS boson has a squared number of Planck quanta in a equilateral π geometry

The HIGGS boson is an alternative name for the 2D charged mass-energy fascia of all Matter

[whose inductive QAM quanta create the physical property of inertial mass]

$$F = ma$$

inertial mass is a measure of a body's resistance to changes in acceleration due to external forces

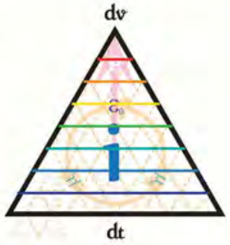

Charged Planck quanta form the geometry of of all mass-ENERGY-Matter
[each charge fascia contains equilateral mass-energy momenta]

TETRYONS

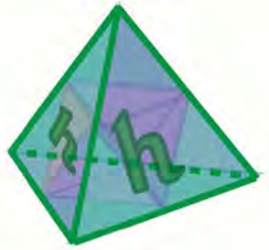

All Matter has 4n enclosed tetrahedral charge topologies

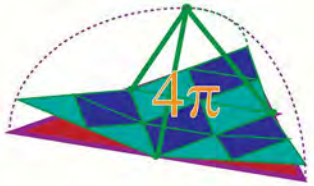

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

3D material Matter is a
EM standing-wave topology
whose mass is a measure of its
geometric Energy content

$$M$$

Matter is the 3D topology created by mass-energies

The energy momenta density of standing wave Matter

determines its inertial mass

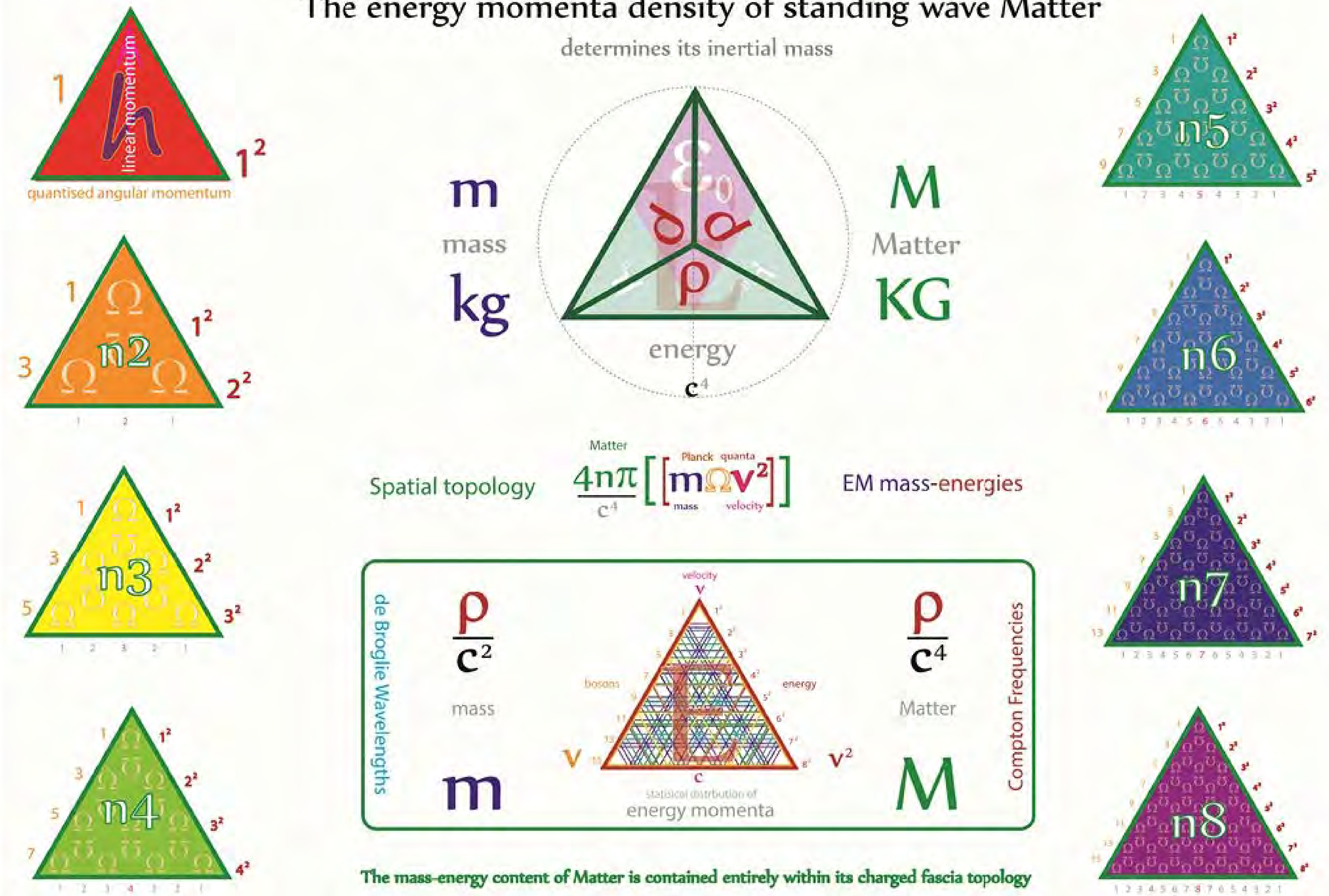
1
1^2
h
linear momentum
quantised angular momentum
n2
n3
n4
n5
n6
n7
n8
ε_0
ρ
v
c
bosons
energy
v
v^2
1
3
5
7
9
11
13
15
2^2
3^2
4^2
5^2
6^2
7^2
8^2

m

mass

kg

M

Matter

KG

energy

c^4

Spatial topology

Matter

$$\frac{4n\pi}{c^4}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

EM mass-energies

de Broglie Wavelengths

$\frac{\rho}{c^2}$

mass

m

velocity

statistical distribution of

energy momenta

$\frac{\rho}{c^4}$

Matter

M

Compton Frequencies

The mass-energy content of Matter is contained entirely within its charged fascia topology

3D Nullspaces

Planck mass-energy quanta clothe
only the topologies of Matter

**3D EM fields with volume
are measurements of
Energy per unit Time squared
[Matter]**

$$T\pi$$

**Nullspace is proportional to the
internal volumes of Tetryons**

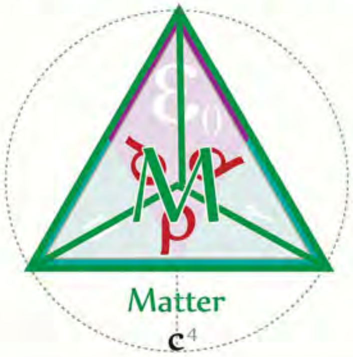

**2D planar EM fields
are measurements of
Energy per unit Time
[EM mass]**

$$[m\Omega v^2]$$

**The mass-energy densities [fascia] of
Matter do NOT influence nullspaces**

**Inside the Tetrahedral topologies
of all Matter exists the only true vacuum**

**Any radiant energy trying to enter
the interior volume of a Tetryon
will be absorbed into the fascia,
increasing its Tetryonic mass
(or increasing its KEM field energy)**

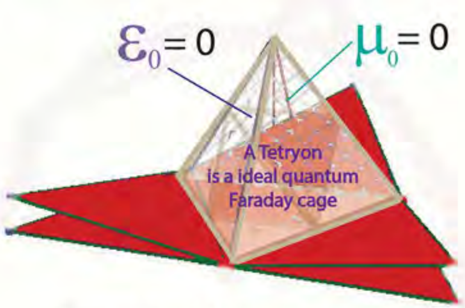

A volume devoid of any possible energy quanta
[ie The True vacuum]

**The interior volume of Tetryons has the
lowest energy density/pressure possible
and is accumulative as Matter forms
increasingly complex states, creating
regions of low energy density within
the nuclei of all Matter**

*Outer space is considered a high-quality vacuum,
with the equivalent of just a few hydrogen atoms
per cubic meter on average. However even it is
permeated with superpositioned EM fields
of vacuum mass-energies*

$$\frac{\rho_0}{T}$$

*A true vacuum is a spatial volume
that is completely devoid of EM mass-energies,
such that its energy density/pressure is much lower
than of the surrounding energy density/pressure
[a nullspace]*

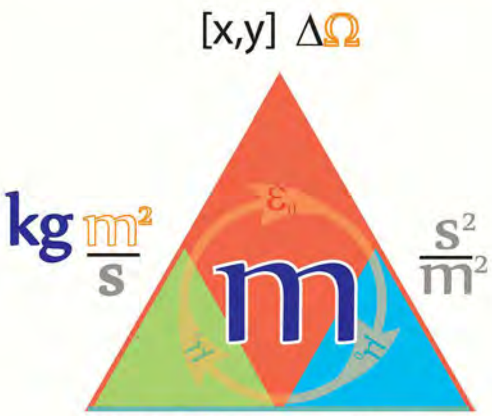

Positive charges in 2D space
are classically modelled as
clockwise energy fluxs

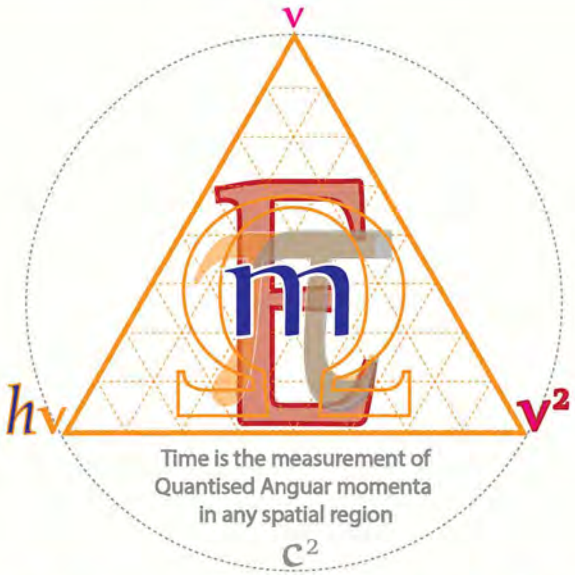

Time is the measurement of
Quantised Anguar momenta
in any spatial region

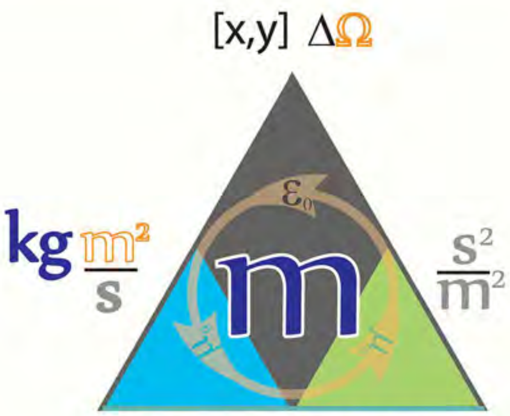

Negative charges in 2D space
are classically modelled as
anti-clockwise energy fluxs

Charged Planck mass-energies provide the scaffolding of 3D Matter

mass

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

mass

[x,y,z] ΔΩ

3D Space and Time

Matter

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

***2D
mass geometries
are a property of
3D Matter***

2 Dimensional mass-energies

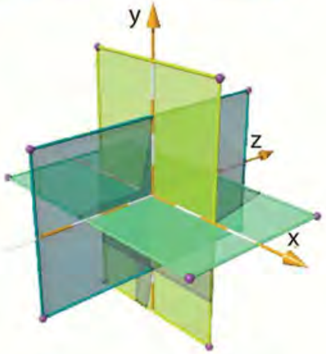

3 Dimensions of Space

kg

**radiant
2D mass-energy momenta**

$$T\pi\left[\underset{\text{impedance}}{\overset{\text{EM Field}}{\left[\varepsilon_0\mu_0\right]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

**creates 3D standing-wave
Matter**

KG

2D charged fascia

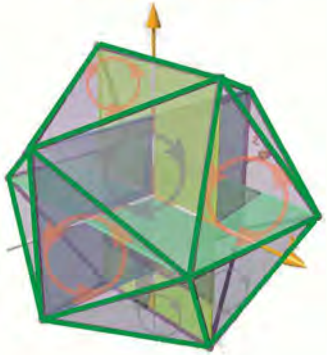

3D Space + Time

***3D
Matter topology
is not a property of
2D mass-energy***

The impedance of any spatial region

is determined by its scalar energy content

Positive Matter

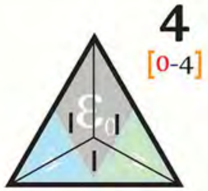

Negative Matter

$$\frac{m}{hv^2}$$

$$\frac{mc^2}{v^2}$$

$$\frac{mc^2}{h}$$

$$\frac{v^2}{c^2}$$

1.112650056 e-17

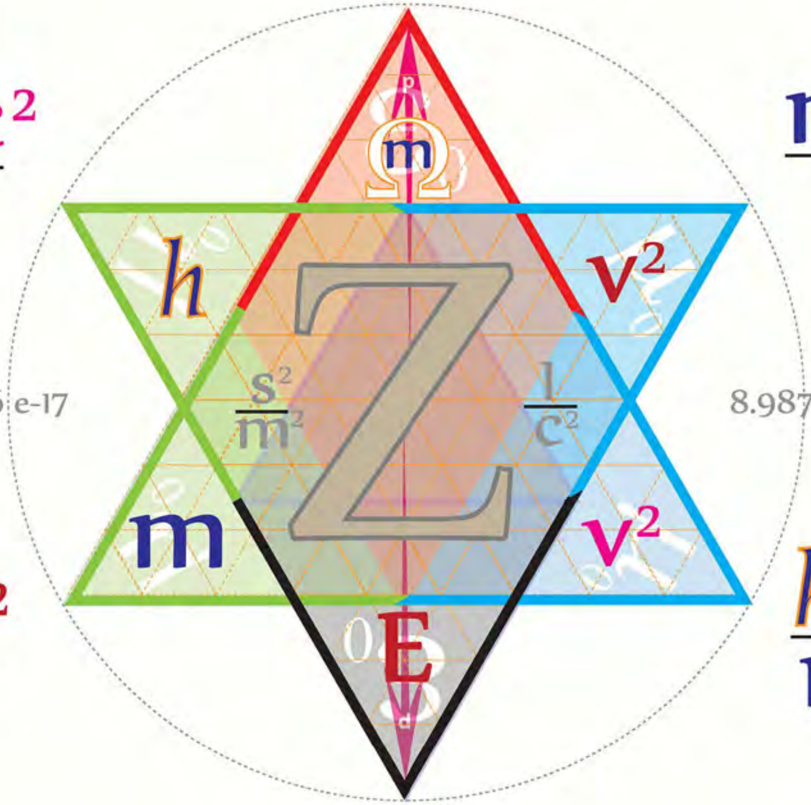

8.987551787 e16 $^{-1}$

$$\frac{h}{m}$$

$$\frac{hv^2}{c^2}$$

$$\frac{hv^2}{m}$$

111.265 ohms

The impedance of any medium determines
the velocity of propagation for mass-energy momenta through it

Neutral Matter

Neutral Matter

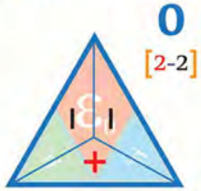

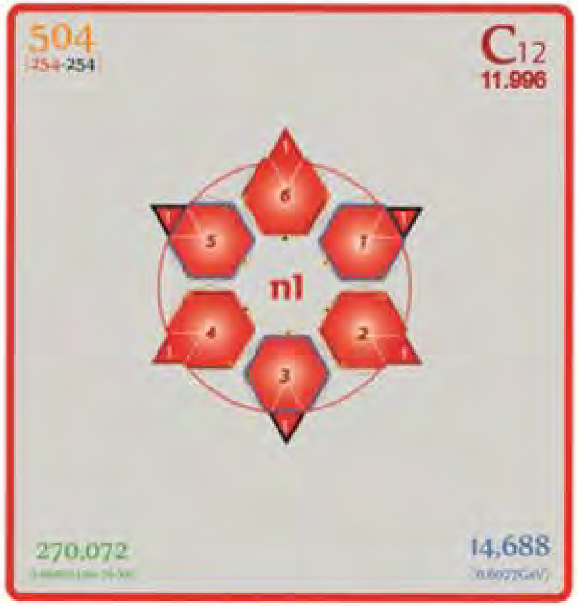
504
[254-254]
C12
11.996
n1
1
2
3
4
5
6
270,072
14,688

EM mass vs. Matter

Often a source of confusion,
Tetryonics provides a rigorous
definition of both EM mass & Matter

6.022141579 e23
[Avagadros number]

1 ton of Carbon & 1 ton of Lead
have the same mass-energies

But they have differing Matter topologies
that create a differing energy densities
for any spational region of measurement

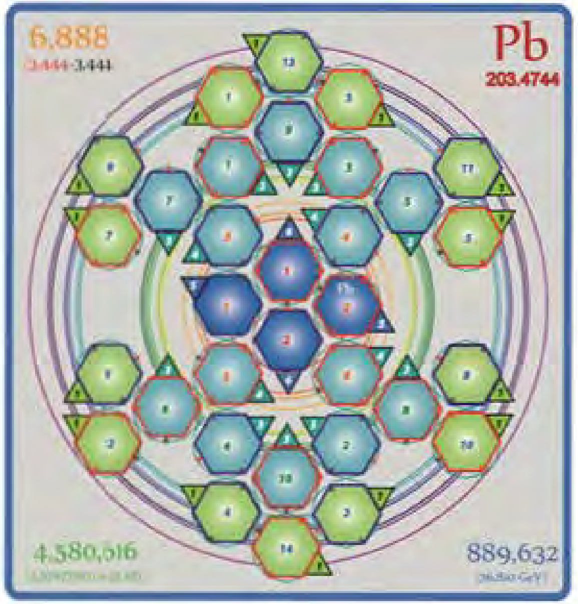
6,888
3444-3444
Pb
203.4744
4,580,516
889,632

1 mole of Carbon atoms
[11.996 grams]

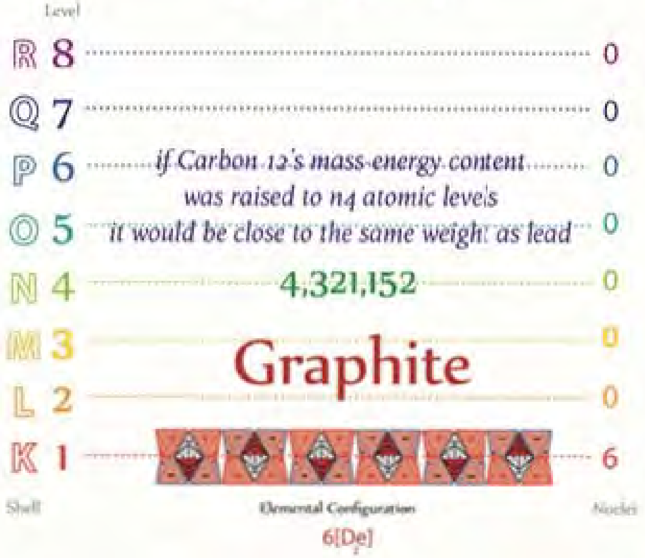
Level
R 8
0
Q 7
0
P 6
0
O 5
0
N 4
4,321,152
0
M 3
0
L 2
Graphite
0
K 1
6
Shell
Elemental Configuration
Nuclei
6[De]

if Carbon 12's mass-energy content
was raised to n4 atomic levels
it would be close to the same weight as lead

In terms of charged Matter topology
Carbon12 displaces 1/14th the volume of Lead

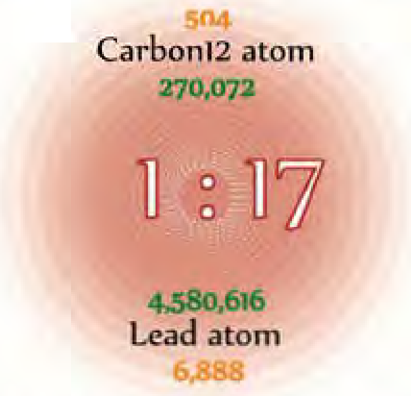
504
Carbon12 atom
270,072
1 : 17
4,580,616
Lead atom
6,888

In terms of mass-ENERGY-Matter densities
Lead is 17 times denser than Carbon
as a result of the mass quanta per charge fascia

1 mole of Lead
[203.4744 grams]

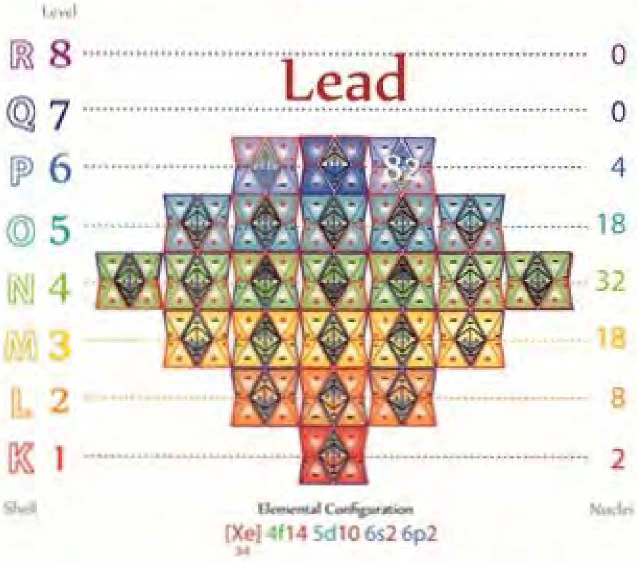
Level
R 8
0
Lead
Q 7
0
P 6
82
4
O 5
18
N 4
32
M 3
18
L 2
8
K 1
2
Shell
Elemental Configuration
Nuclei
[Xe] 4f14 5d10 6s2 6p2
54

Copyright ABRAHAM [2008] - All rights reserved

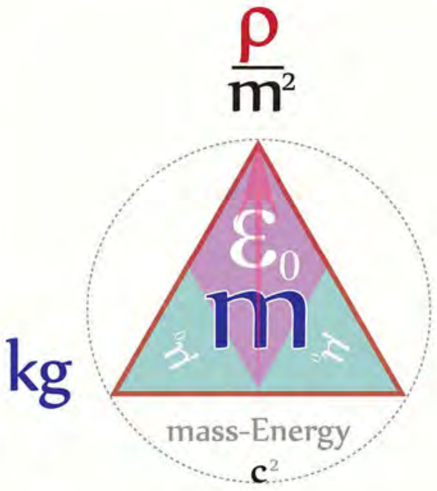

EM mass-energy geometries

are 2D Euclidean EM fields posessing energy momenta

Charge

$$\frac{1\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Photons

$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

mass

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

m

ENERGY

$$n\pi\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

M

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Equilateral mass-energy geometries & Matter Topologies are measured in radial spatial-time co-ordinate systems

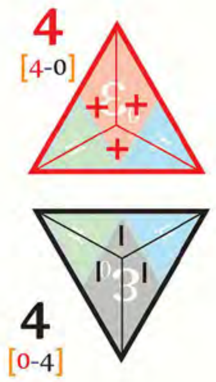

Tetryons

$$\frac{4\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Fermions

$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Baryons

$$\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Deuterons

$$\frac{84\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

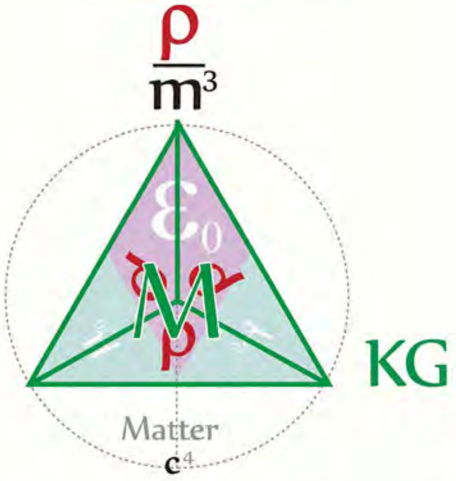

Matter Topologies

are 3D standing-wave [4nπ] mass-energy geometries

mass-energy-Matter quanta

are all defined by charged geometries & topologies

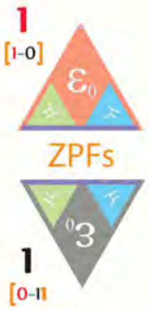
1
[1-0]
ε0
ZPFs
μ0
1
[0-1]

Bosons (v)

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\Big[\big[m\Omega v^2\big]\Big]$$

Planck quanta mass velocity

charges

1π ZPF v

ODDπ boson v

Inductive charges

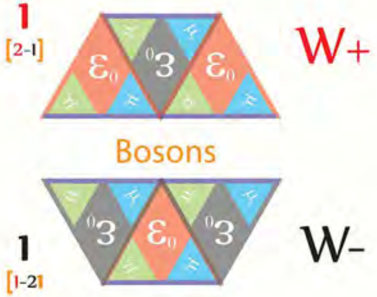
1
[2-1]
W+
Bosons
1
[1-21
W-
ε0
μ0

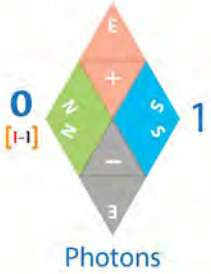
0
[1-1]
1
Photons

Photons(*f*)

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\Big[\big[m\Omega v^2\big]\Big]$$

Planck quanta mass velocity

mass-energies

2π photon $2v^2$

EVENπ EM wave $2v^2$

EM waves

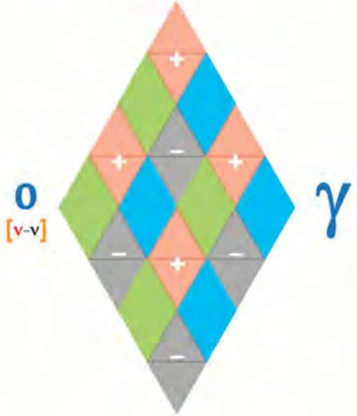

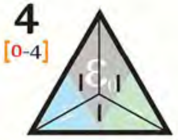

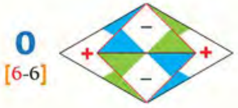

Tetryons (n)

Matter

$$\frac{4n\pi}{c^4}\Big[\big[m\Omega v^2\big]\Big]$$

Planck quanta mass velocity

mass-Matter

4π tetryon n

12π lepton n

24π meson n

36π Baryon n

Matter particles

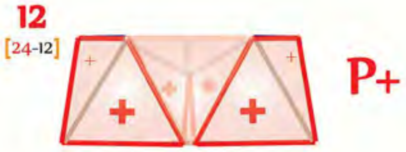

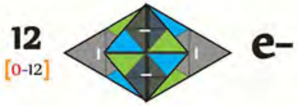

Differentiating mass & Matter

geometries topologies

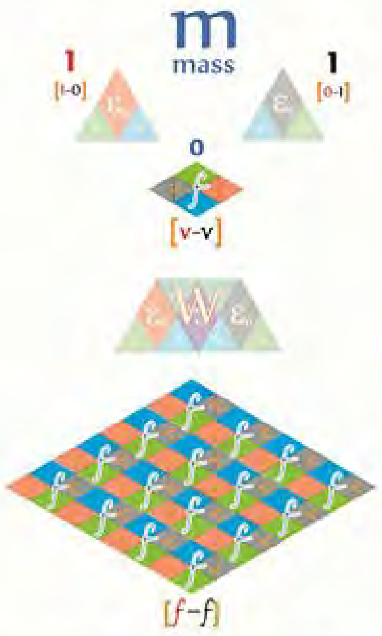

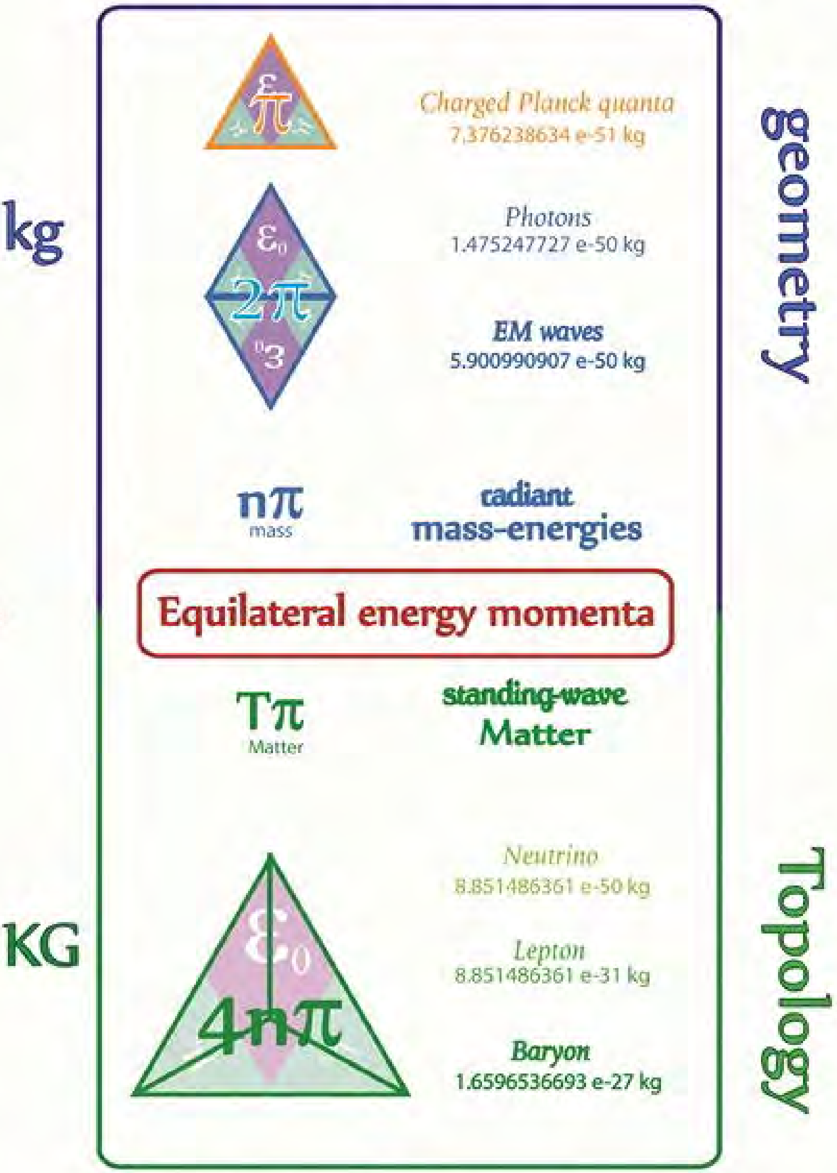

Charged Planck quanta
7.376238634 e-51 kg

Photons
1.475247727 e-50 kg

EM waves
5.900990907 e-50 kg

nπ
mass

radiant mass-energies

Equilateral energy momenta

Tπ
Matter

standing-wave Matter

Neutrino
8.851486361 e-50 kg

Lepton
8.851486361 e-31 kg

Baryon
1.6596536693 e-27 kg

Increasing a particle's speed increases its Kinetic EM mass-energies
(rest Matter's EM mass-energy content is velocity invariant)

All modern physical theories including relativity
fail to define and differentiate between mass & Matter

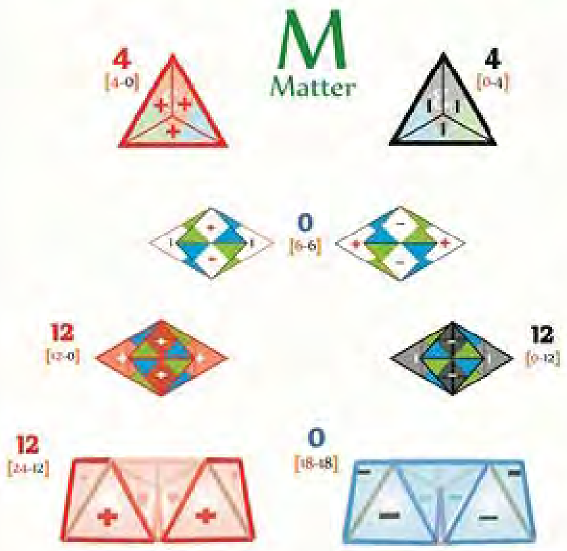

Charge
$$\frac{1\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

Photons
$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

Bosons
$$\frac{\text{ODD}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

EM waves
$$\frac{\text{EVEN}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

All Matter has EM mass-energies
not all EM mass-energies form Matter

Tetryons
$$\frac{4\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

Fermions
$$\frac{12\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

Baryons
$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

Deuterons
$$\frac{84\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$
Planck quanta mass velocity

$$E = mv^2$$

E

Energy

$$kg\,\frac{m^2}{s^2}$$

6.629432351 e-34 J

$$\underset{\text{Energy}}{E} = n\pi\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

ENERGY

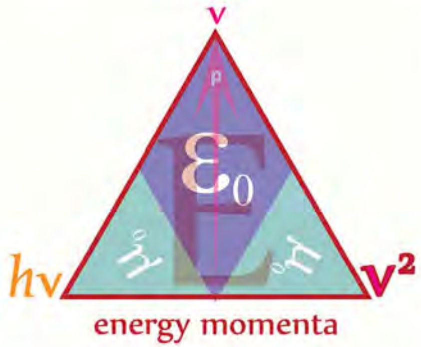
v
p
ε_0
hv
v^2
energy momenta

the proagation velocity of energy is determined
by the impedance of any spatial region

E=hv²

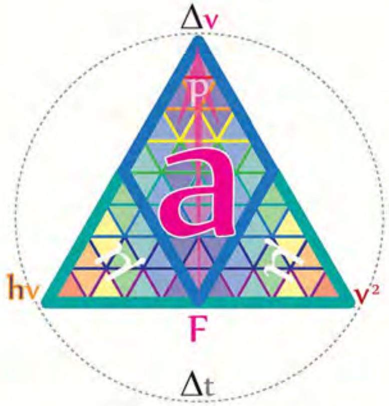
Δv
p
a
hv
v^2
F
Δt

F=ma

the scalar impedance of space

$$\frac{s^2}{m^2}$$

$$\varepsilon_0 \times \mu_0$$

Electric permittivity* & *Magnetic permeability

$$1.11265 \text{ e-}17\ \frac{s^2}{m^2}$$

$$\frac{1}{c^2}$$

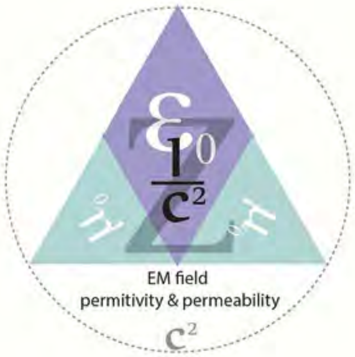
ε_0
$\frac{1}{c^2}$
EM field
permitivity & permeability
c^2

the Forces created by energy momenta in any spatial region
are related through the scalar constant of mass

EM mass

$$m = \frac{E}{c^2}$$

m

mass

kg

$$\underset{\text{mass}}{n\pi}\left[\overset{\text{field impedance}}{\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}}.\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

7.376238376 e-51 kg

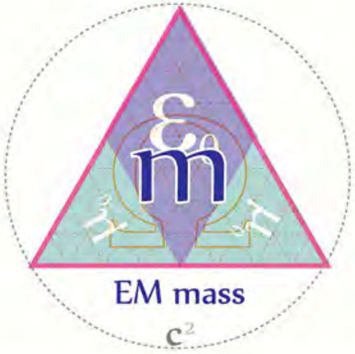
ε
m
EM mass
c^2

Copyright ABRAHAM [2008] - All rights reserved

The energy-momenta of Matter

$$p = mv$$

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

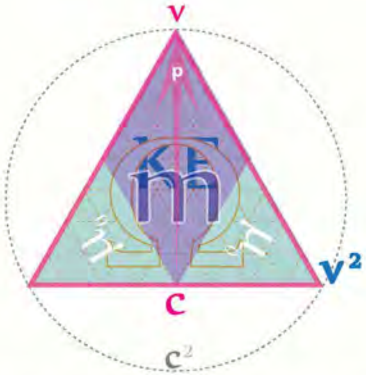

The Energy comprising all KEM fields is proportional to the particle's velocity

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

$$p^2 = E = mv^2$$

$$p = T\pi\left[\left[m\overset{\text{Energy}}{\underset{\text{momenta}}{\Omega}}v\right]\right]$$

As the velocity of a particle increases so does its KEM field's mass-energy momentum

$$mv = p = hv$$

$$kg\,\frac{m}{s}$$

$$\sqrt{6.629432351}\ e\text{-}34\ J$$

$$4n\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v}\right]$$

2.5476847 e-17 kg m/s

The linear momentum of mass-energies in standing-wave rest Matter topologies propagates at c

$$p_M = 4\pi[mv]$$

The rest mass of a particle with NO motion is the total EM mass-energy of the particle

Its linear four-momentum is the square root of its scalar energies

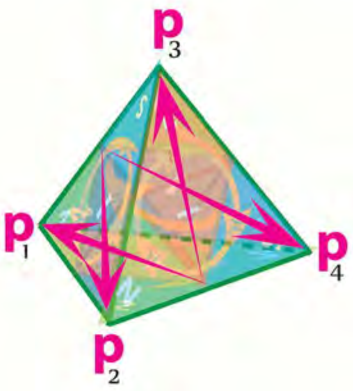

rest mass-Matter

In standing-wave Matter, where there is NO external forces acting on the particle, all internal momenta act in opposition and add to ZERO

$$\vec{p}_0 = \overset{\text{all fascia}}{\sum} \vec{p}$$

Copyright ABRAHAM [2008] - All rights reserved

Gravitational mass-Matter quanta

gravity pulls on things proportional to their mass-Matter content

$$F = -G \frac{M_1 M_2}{r^2}$$

Radiant mass-energy geometries

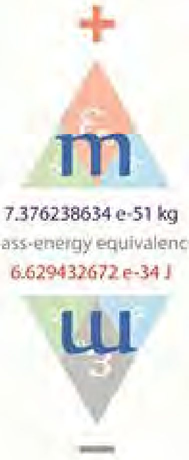

7.376238634 e-51 kg
mass-energy equivalence
6.629432672 e-34 J

All 3D Tetryonic topologies create displacement Matter volumes that in turn generate KEM fields as a result of the motion of their charge fascia

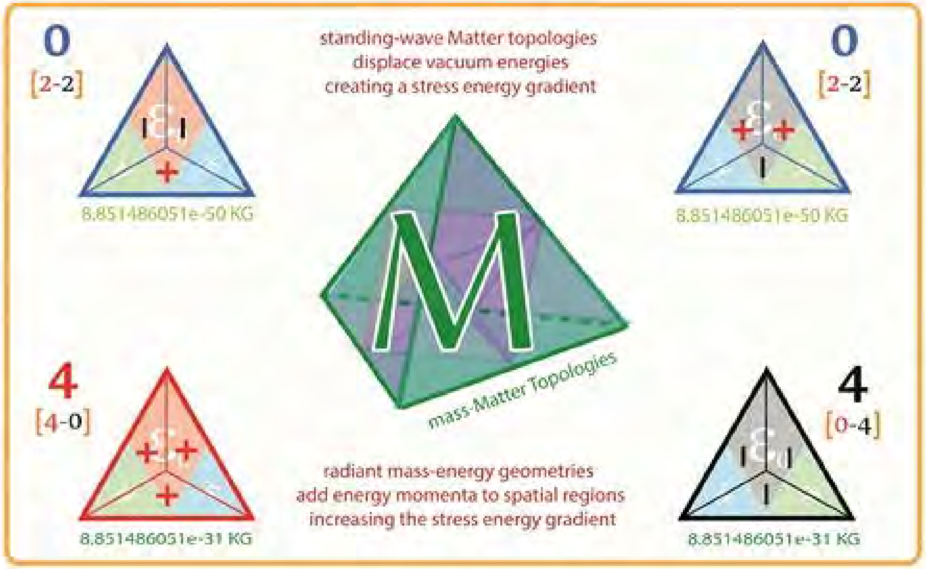

standing-wave Matter topologies
displace vacuum energies
creating a stress energy gradient

radiant mass-energy geometries
add energy momenta to spatial regions
increasing the stress energy gradient

Standing-wave Matter Topologies

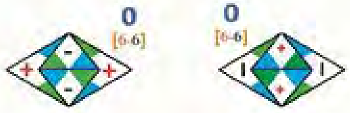

8.851486051e-50 KG

Matter particles can have the same Tetryonic topologies but possess a differing charged mass-energy momenta content

8.851486051e-31 KG

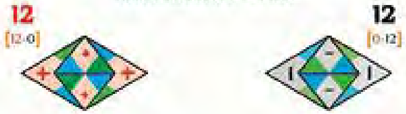

Gravitational Attraction can be modelled as a dipole attraction of opposite charged mass fascia of Matter topologies through its differing energy field interactions

Gravitation attraction is a strictly UNI-DIRECTIONAL CONVERGENT FORCE

$k = 8.99\text{e}9 \frac{N m^2}{C^2}$

$G = 6.67\text{e-}11 \frac{N m^2}{KG^2}$

KEM field energy momenta produce a BI-DIRECTIONAL INTERACTION FIELD

Gravitation attraction is 20 orders of magnitude weaker

KEM field energy momenta produce a BI-DIRECTIONAL INTERACTION FIELD

Stress energy gradients

Energy's highest density form is 3D Matter [comprised of planar 2D EM mass-energies]
whilst its lowest pressure form is the vacuum energies created by radiated mass-energies filling space

ENERGY in all its forms seeks equilibrium

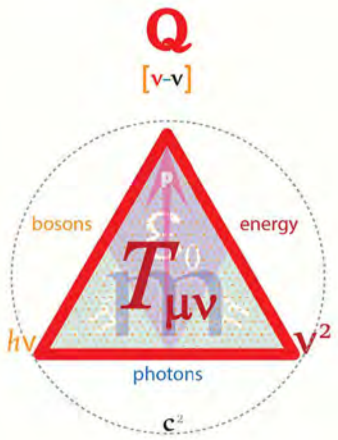

Where Newton's Law of Gravitation is modelled on geometric means

$$F = k\frac{Q_1 Q_2}{r^2}$$

Inverse Square Law

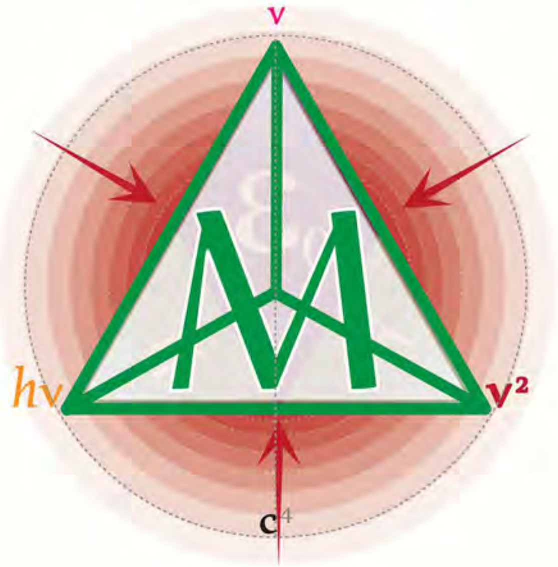

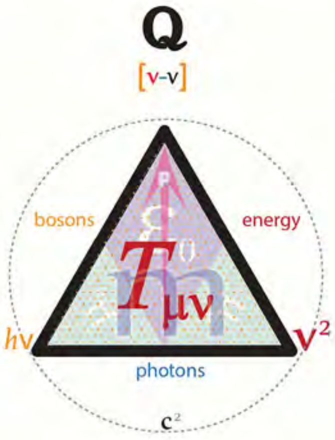

Einstein's General Relativity is modelled on energy pressure

$$PV = nRT$$

Matter tells Spacetime how to bend
Spacetime tells Mater how to move

$$F = -G\frac{M_1 m_2}{r^2}$$

Like any pressure gradient, ENERGY always flows from areas of highest density to regions of low density seeking to reduce 'pressure' differentials and reach a state of equilibrium.

$$R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

mass-Matter topologies create stress energy differentials in the vacuum energies of space

Copyright ABRAHAM [2008] - All rights reserved

Gravitational Matter & weight

is a measure of the amount of mass-energy in Matter

The gravitational mass of a body determined by its response to the force of gravity

mass-Matter

Matter

Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

The gravitational acceleration of a body in any gravity field is uniform and proportional to its mass-Matter

$$w = F = M.g$$

electrons

Planck quanta

$$\frac{12\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy-Matter

8.8514586361 e-31 KG

48

[24-24]

charge topology

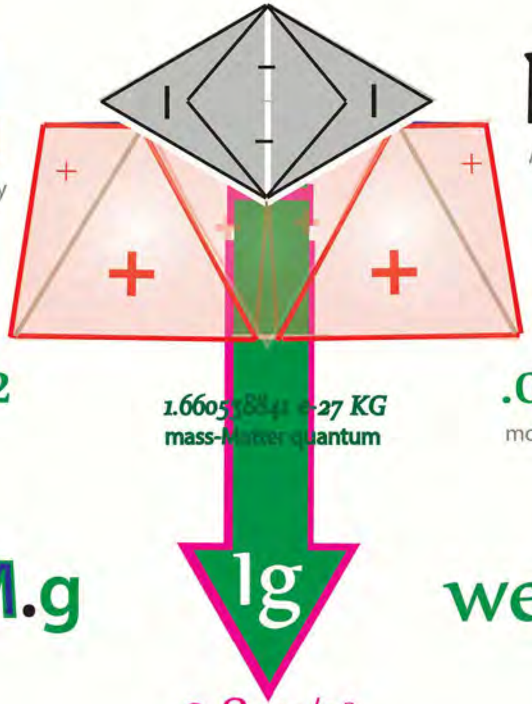

N_A

Avagadro

Baryons

Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy-Matter

1.659653693 e-27 KG

22,512

mass-Matter

1.660538841 e-27 KG

mass-Matter quantum

.001 KG

molar mass

$$w = M.g$$

weight

9.8 m/s²

$$F = Ma_g$$

$= [.001] \times [9.8]$

$= .0098$ N

KG

g

Force

$$\frac{\left[kg\,\frac{m}{s^2}\right]}{KG}$$

Copyright ABRAHAM [2008] - All rights reserved

Zero Point Fields

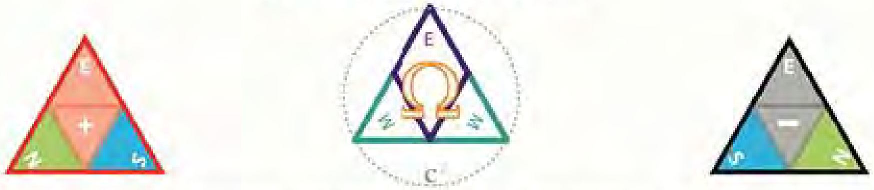
E
+
N
S
E
Ω
W
W
C²
E
−
S
N

ZPF quanta charges are the opposite sides of the same 'quantum coin'

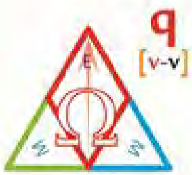

ZPF quanta have unidirectional energy momenta orientations

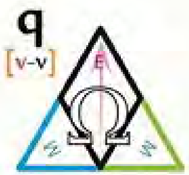

KEM fields

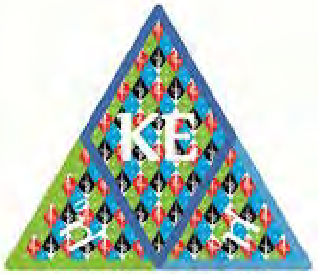

N-S magnetic moment vector

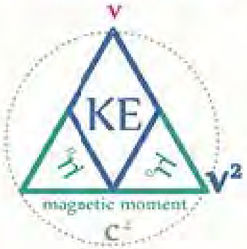
v
KE
v²
magnetic moment
c²

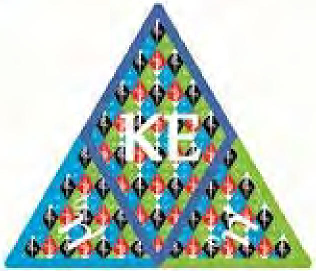

S-N magnetic moment vector

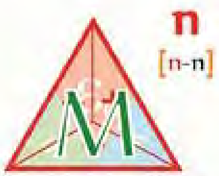

standing-wave mass-Matter

As the velocity of Matter increases its relativistic mass-Energy quanta increase as a result of the associated momentum, kinetic energy and magnetic moment

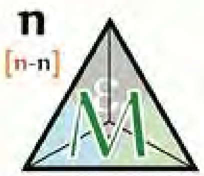

standing-wave mass-Matter

in addition to their standing-wave topologies mass-Matter also posesses

radiant mass-energies

Gravity & EM fields

All energy fields are velocity dependent and combine to form vectorial K-EM fields

All charged Matter posseses bi-directional interactive radiant EM fields

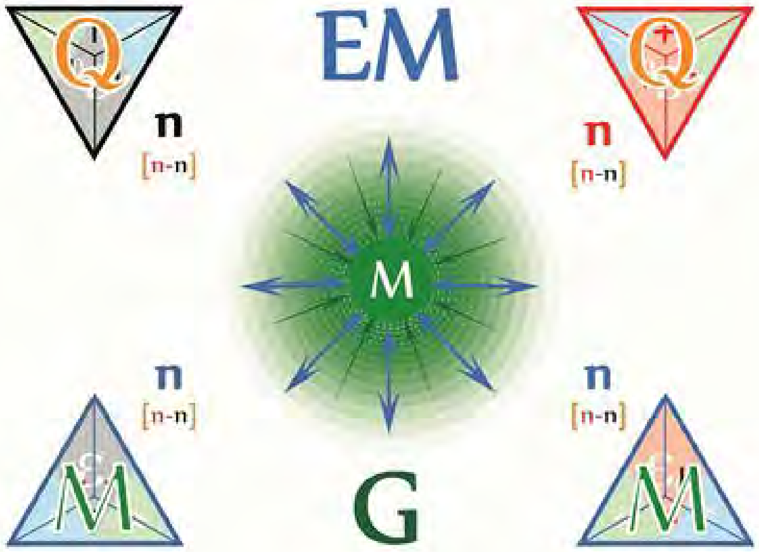
Q
n
[n-n]
EM
Q
n
[n-n]
M
n
[n-n]
M
n
[n-n]
M
G

charged Matter topologies create uni-directional convergent energy gradiants

All mass-Matter is velocity invariant and creates convergent G fields

Copyright ABRAHAM [2008] - All rights reserved

Matterless EM mass-energies

Photons are often termed as 'mass-less' particles in physics which is a complete mis-nomer.
Leading to many erroneous supersitions on the true nature and physics of light and gravity

$$\frac{E}{c^2} = m$$

mass geometry

$$M = \frac{E}{c^4}$$

Matter topology

mass is a measure of planar Energy per second

mass-less infers no energy momenta

Photons are weightless radiant EM energies not massless particles

photons possess 2D planar radiant EM mass-energy momenta

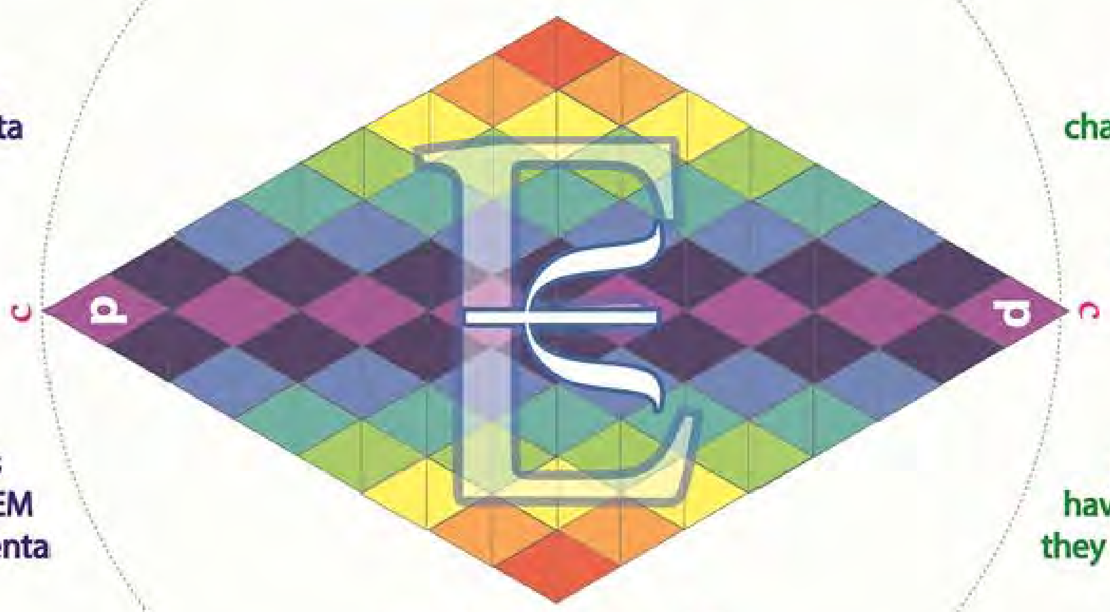

A Particle has a charged Matter topology

Matter is a closed tetrahedral standing-wave energy possessing volume

photons & EM waves have no 3D Matter topology, they are Matter-less geometries

ENERGY

$$\frac{n\pi\left[\left[m\Omega v^2\right]\right]}{c^2}$$

Planck quanta mass velocity

second

mass

$$\frac{n\pi\left[\left[m\Omega v^2\right]\right]}{c^2}$$

Planck quanta mass velocity

Matter

$$\frac{T\pi\left[\left[m\Omega v^2\right]\right]}{c^4}$$

Planck quanta mass velocity

The definition and distinction between mass & Matter
is a major failing of modern physics, corrected for by Tetryonic theory

Copyright ABRAHAM [2008] - All rights reserved

EM fields, vacuum energies and gravity

all result from radiant equilateral mass-energy momenta
and standing-wave Matter in radial space

$$F = k\frac{Q_1 Q_2}{r^2}$$

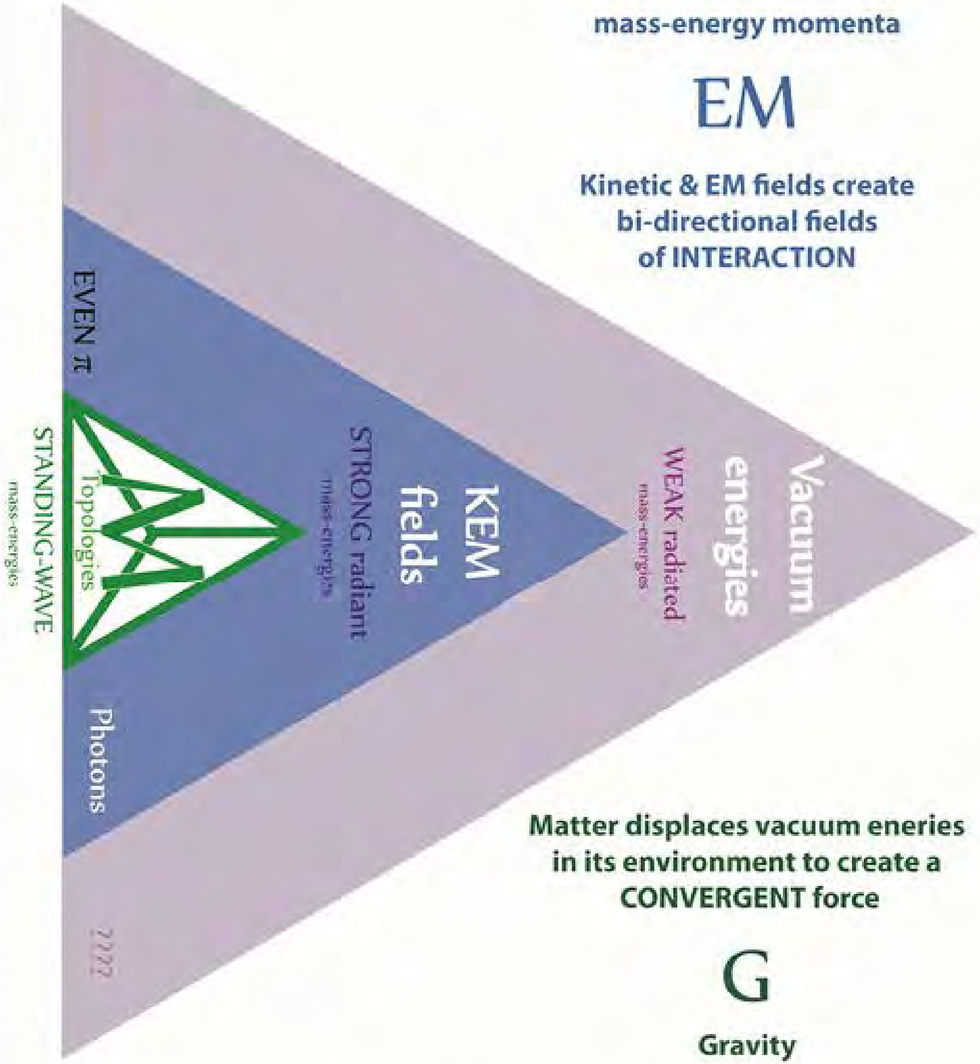
EVEN π
STANDING-WAVE
mass-energies
Topologies
Photons
????
STRONG radiant
mass-energies
KEM
fields
WEAK radiated
mass-energies
Vacuum
energies

mass-energy momenta

EM

**Kinetic & EM fields create
bi-directional fields
of INTERACTION**

**Matter displaces vacuum eneries
in its environment to create a
CONVERGENT force**

G

Gravity

**All GEM fields
have equilateral
quantum geometries**

**All G-EM fields can
be modelled with
Tetryonic geometry**

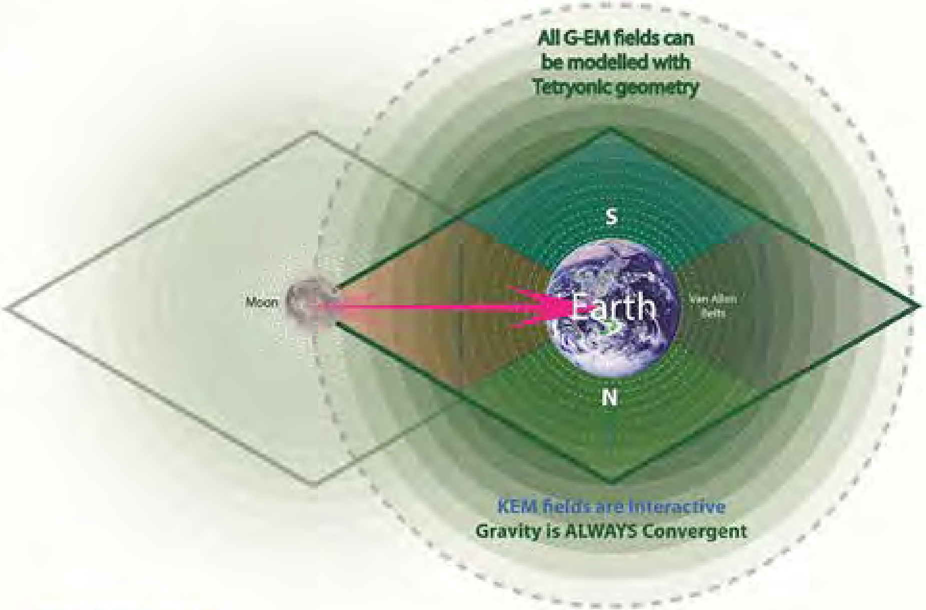
S
Moon
Earth
Van Allen
Belts
N

KEM fields are Interactive
Gravity is ALWAYS Convergent

*the proportional attraction of
material bodies towards each other
can be easily modeled with Tetryonic fields*

$$F = -G\frac{M_1 m_2}{r^2}$$

geometric mean

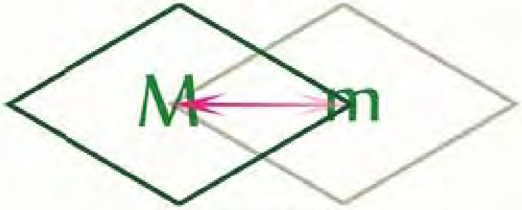

inverse square law

Copyright ABRAHAM [2008] - All rights reserved

Charges vs. Gravity

both can be modelled as interactive or attractive geometric forces

Charged particles are comprised of non-neutralised Planck quanta

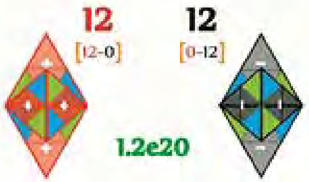
12
[12-0]
12
[0-12]
1.2e20

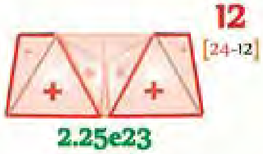
12
[24-12]
2.25e23

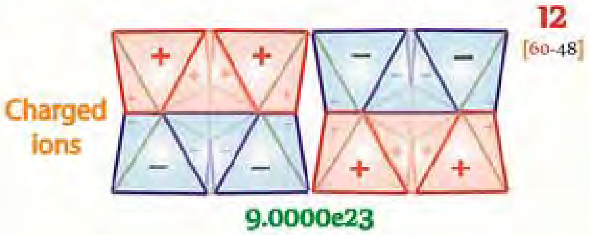
12
[60-48]
Charged ions
9.0000e23

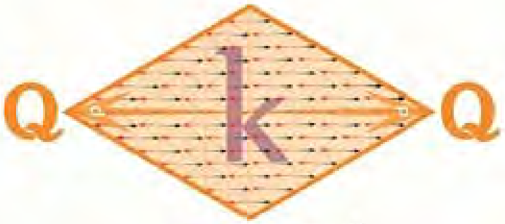

$8.99 \times 10^{9}\ \frac{N\,m^2}{kg^2}$

Opposites Attract
Similars Repel

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

Charges produce bi-lateral interactive forces

mass

Matter

Nuclei

Elements

$$F = -G\,\frac{M_1 M_2}{r^2}$$

Gravity is a convergent attractive force

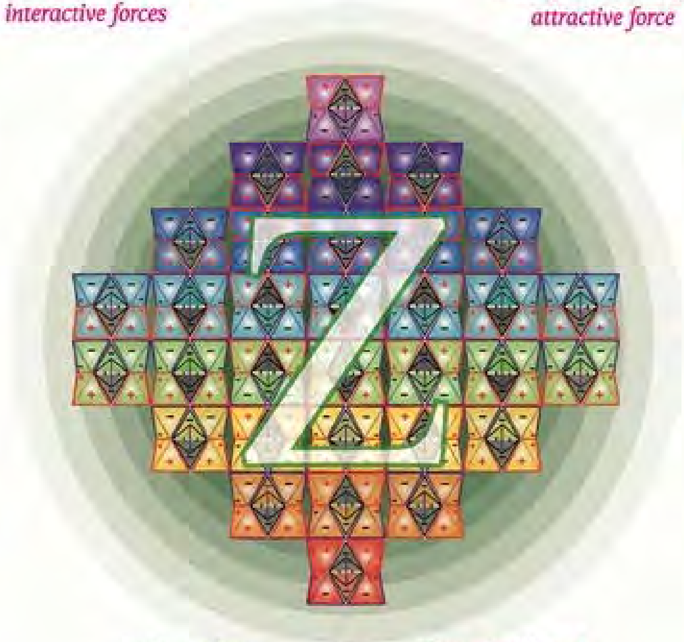

Interactive mass-energies & **Gravitational mass-Matter**

both produce inverse squared forces

Neutral Matter is comprised of neutralised Planck quanta

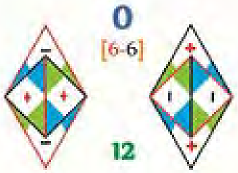

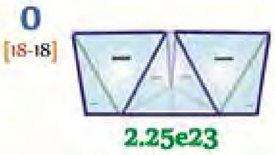
0
[18-18]
2.25e23

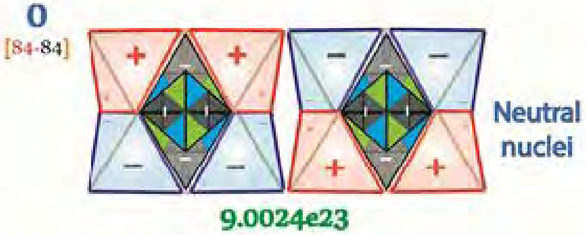
0
[84-84]
Neutral nuclei
9.0024e23

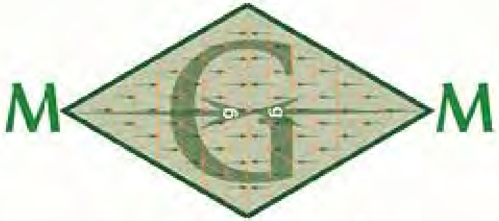

$7.49 \times 10^{-11}\ \frac{N\,m^2}{KG^2}$

Gravitational Matter is always Attractive

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

GravitoElectroMagnetism, (abbreviated GEM), refers to a set of formal analogies between Maxwell's field equations and an approximation, valid under certain conditions, of the Einstein field equations for General Relativity.

$$\nabla \cdot \mathbf{B} = 0$$

$$\mathbf{F} = q\left(\frac{-1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}\right) = q\mathbf{E}$$

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

$$k = 8.99 \times 10^{9}\,\frac{\mathrm{N\,m}^2}{\mathrm{C}^2}$$

nett Charge is a DIVERGENT field

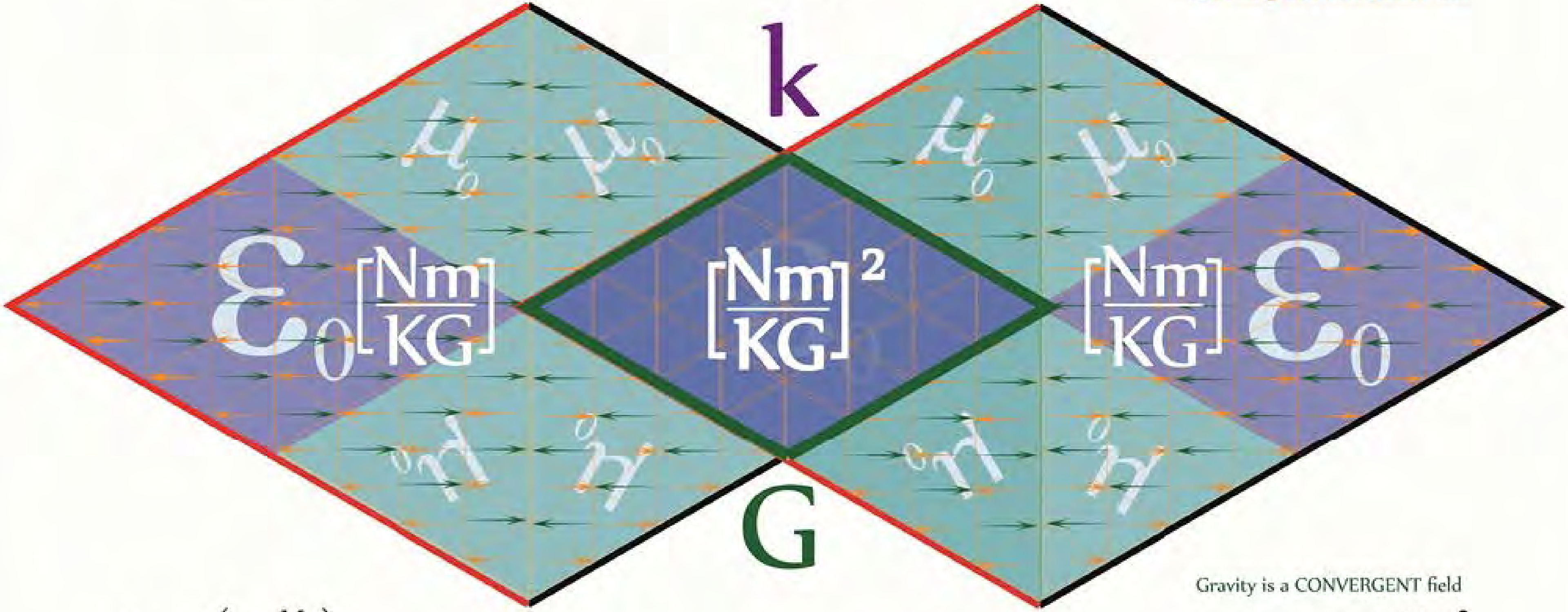
k
μ_0
μ_0
μ_0
μ_0
$\varepsilon_0\left[\frac{\mathrm{Nm}}{\mathrm{KG}}\right]$
$\left[\frac{\mathrm{Nm}}{\mathrm{KG}}\right]^2$
$\left[\frac{\mathrm{Nm}}{\mathrm{KG}}\right]\varepsilon_0$
μ_0
μ_0
μ_0
μ_0
G

Gravity is a CONVERGENT field

$$G = 6.67 \times 10^{-11}\,\frac{\mathrm{N\,m}^2}{\mathrm{kg}^2}$$

$$\mathbf{F} = m\left(-G\frac{M}{r^2}\hat{\mathbf{r}}\right) = m\mathbf{g}.$$

$$F = -G\,\frac{M_1 M_2}{r^2}$$

$$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$$

GEM fields are an accurate way of modelling Gravitational attraction between massive bodies using weak KEM field vectors

$$\nabla \cdot \mathbf{B}_g = 0$$

Gravitational fields can be modelled as weak convergent EM fields

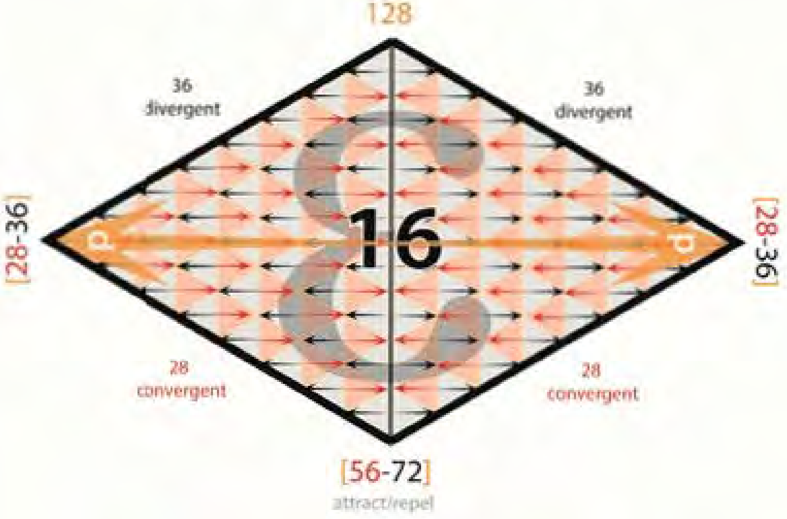
128
36
divergent
36
divergent
[28-36]
p
16
p
[28-36]
28
convergent
28
convergent
[56-72]
attract/repel

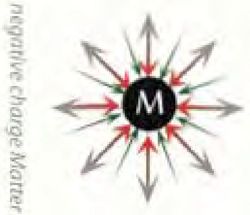
negative charge Matter
M

Electrostatic fields

Charges attract or repel dependent on polarity
Neutral Matter is largely unaffected (equally repel & attract)

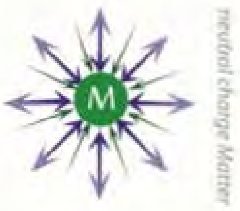
M
neutral charge Matter

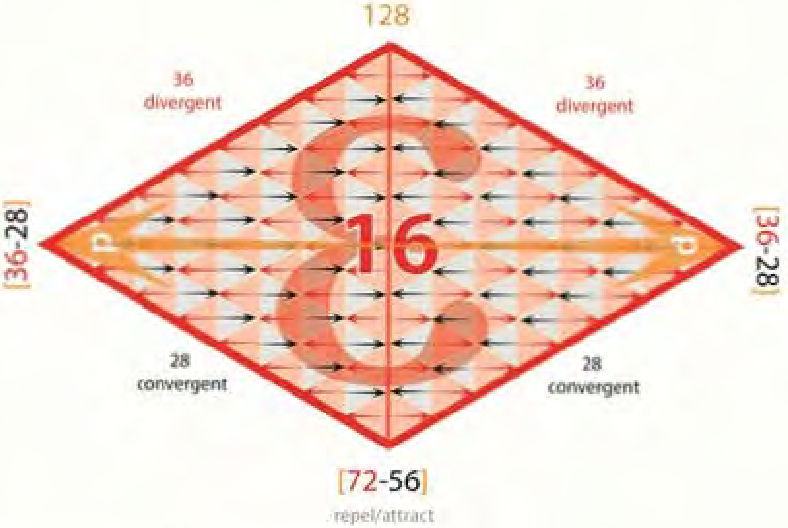
128
36
divergent
36
divergent
[36-28]
p
16
p
[36-28]
28
convergent
28
convergent
[72-56]
repel/attract

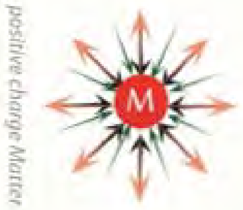
positive charge Matter
M

Charged Matter INTERACTS electrically

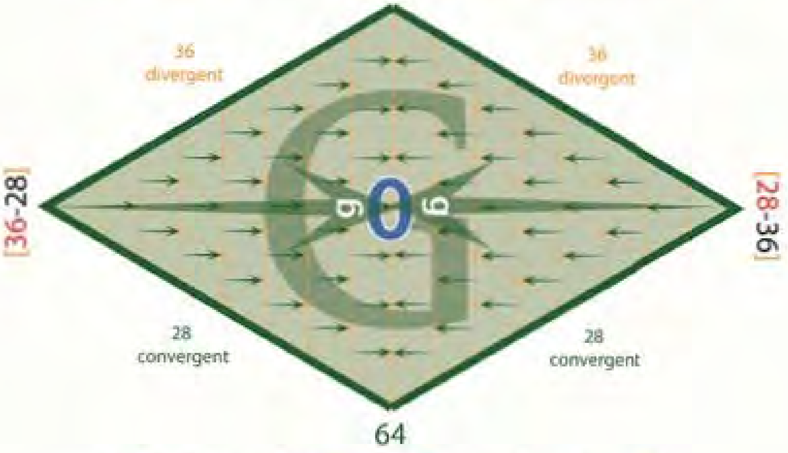
36
divergent
36
divergent
[36-28]
G
0
g
g
[28-36]
28
convergent
28
convergent
64

All Matter ATTRACTS gravitationally

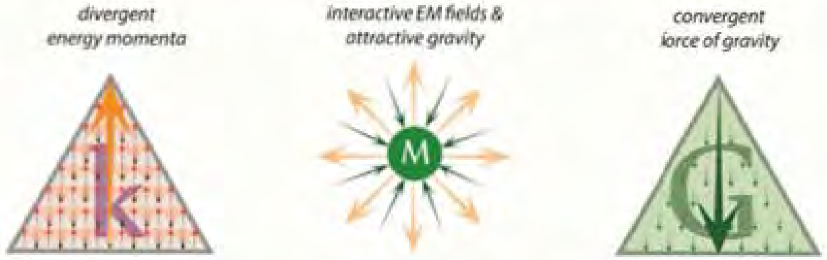
divergent
energy momenta
k
interactive EM fields &
attractive gravity
M
convergent
force of gravity
G

Nett KEM field interaction is divergent from source
Gravitational attraction is the convergent mean force between two sources

Copyright ABRAHAM [2008] - All rights reserved

Gravitational & Electric constants

are scaling constant relating Gravitational gradients to superpositioned EM fields

[1665]

Newton was the first to apply the Inverse Square law to Gravity

but was unaware of ElectroMagnetic forces

GEM

Gravito ElectroMagnetism

[1769]

Coulomb applied the Inverse Square Law to Charge Interactions

and obtained the value for the Electric Constant

Newton's convergent Gravity & Couolomb's Charge interaction
both follow inverse squared laws
(geometric means)

Attractive Gravity has the negative vector of nett Charge Interaction

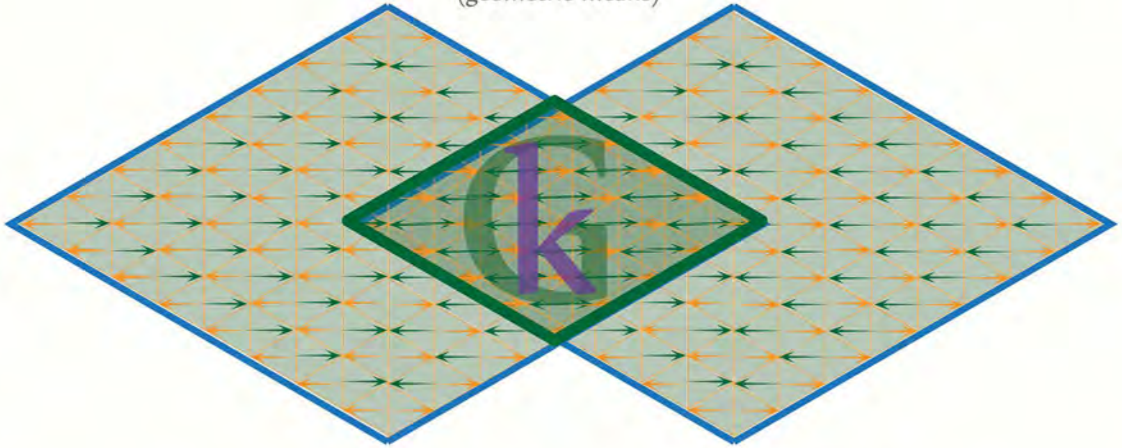

Interactive EM fields are 20 orders of magnitude stronger than Gravity

'Graviton'
Gravitational 'charge' carrier

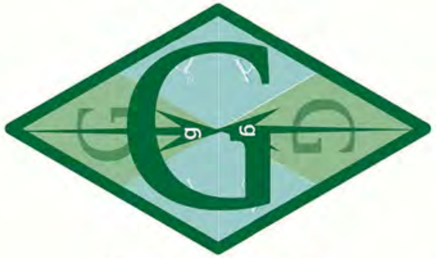

Convergent force vectors

superpositioned EM fields can be used to model the strength and direction of the weaker Graviational force vectors at any point in the fields as they are both inverse squared forces

Producing the ongoing speculation of the existence of a gravitational charge carrier

(an erroneous assumption that stems from using mathematics alone, without a solid understanding of geometric mass-energies & Matter topologies)

Photon
Electric charge carrier

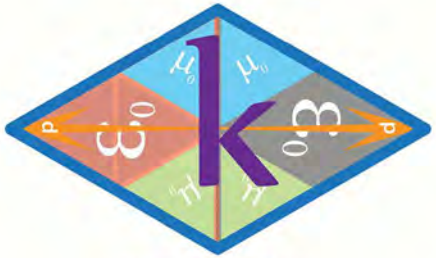

Divergent force vectors

Copyright ABRAHAM [2008] - All rights reserved

Newton's attractive Gravity & Coulomb's Charge interaction
both follow inverse square laws

Instantaneous
action-at-a-distance

The Force of Gravity

Absolute
Space & Time
(x,y,z) (t)

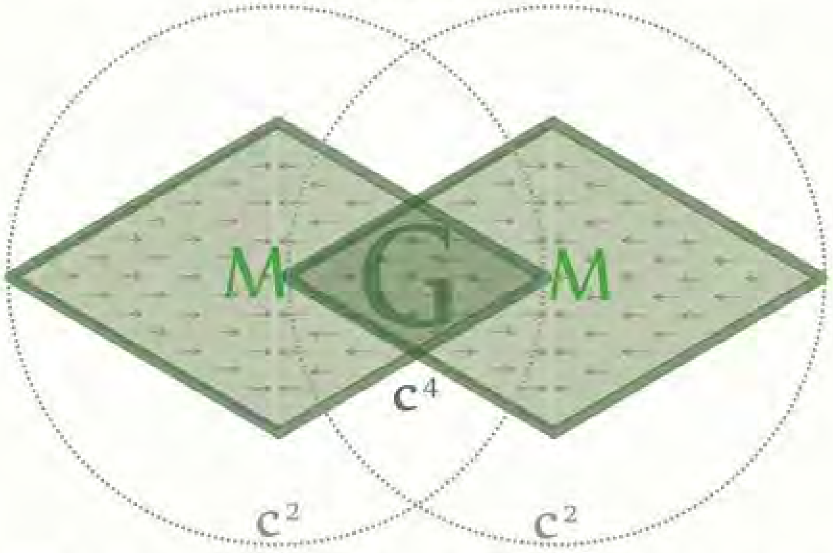

Gravitational action
limited to c

mass-ENERGY-Matter

Curved relativistic
Spacetime
(x,y,z,t)

Modelling Gravity with EM field geometries

while superpositioned EM fields can be used to model the strength and vector direction
of weaker Gravitational fields at any point in time within the field geometry
they DO NOT reflect the actual physicality of the fields involved

$$R_{00} = 4\pi\rho$$

*The speed of Gravity is not limited to c
as changes to the energy density-pressure
gradient are immediately translated through
the aether from one object to the other*

$$R_{\mu\nu} - \tfrac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

*Gravity is a the pressure gradient created
by the displacement of vacuum energies
though the presence of 3D mass-Matter
topologies in any region of space*

Copyright ABRAHAM [2008] - All rights reserved

Fields of Interaction

All topological Matter in motion creates a interactive Kinetic EM field
in addition to its inherent convergent Gravitational field

Convergent Gravity fields produce a uni-directional attraction between two or more Matter topologies

Divergent EM fields produce a bi-directional interactive force between charged objects

interactive EM fields of force are created by geometrically superpositioned equilateral energy momenta vectors

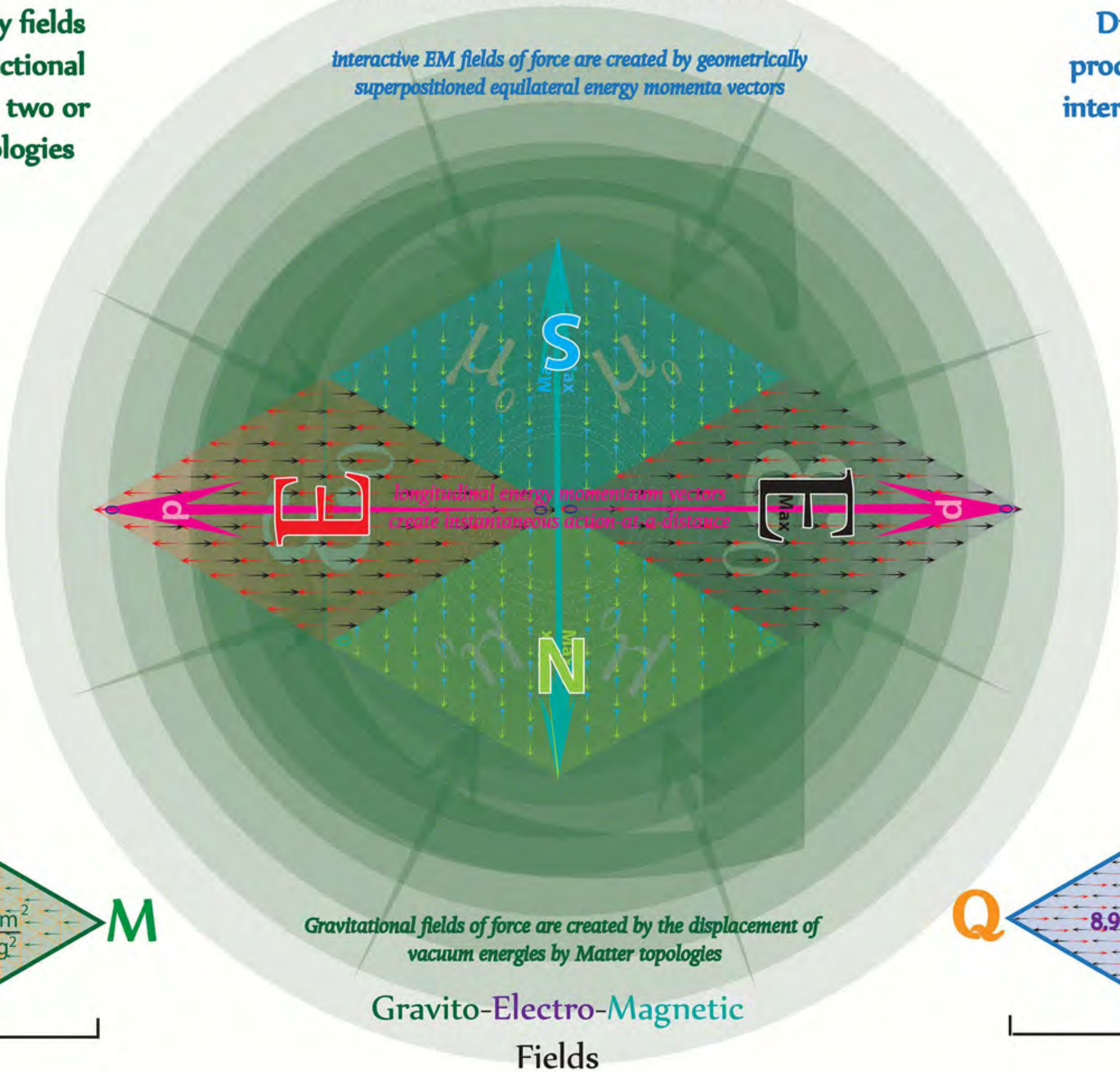
S
N
E
E
Max
Max
p
p
0
0

longitudinal energy momentaum vectors create instantaneous action-at-a-distance

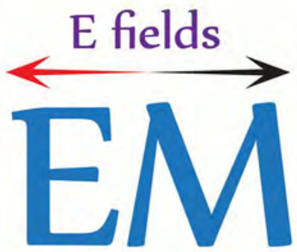

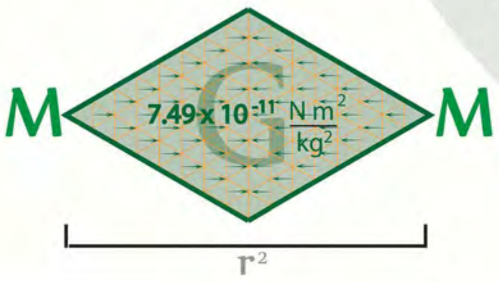
M
M
G
$7.49 \times 10^{-11} \frac{N\,m^2}{kg^2}$
r^2

Gravitational fields of force are created by the displacement of vacuum energies by Matter topologies

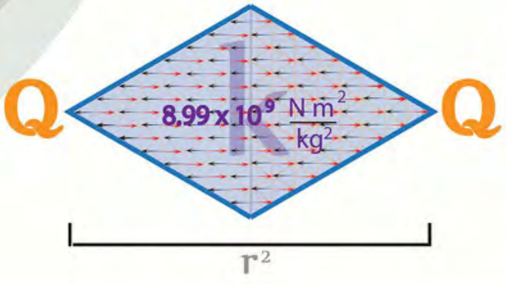
Q
Q
k
$8{,}99 \times 10^{9} \frac{N\,m^2}{kg^2}$
r^2

Gravito-Electro-Magnetic
Fields

Copyright ABRAHAM [2008] - All rights reserved

GEM gradients

The gravitational field of Material objects is the result of zero energy nullspace volumes encompassed by all Matter which in turn creates a displacement energy pressure differential gradient with the surrounding Vacuum Energies

Gravitational attraction can be modelled as a CONVERGENT EM FIELD created by the nullspace differential of Matter

All Matter in motion creates an eternally dynamic system of Gravitational fields and radiative ElectroMagnetic Energies

EM radiation is a bidirectional DIVERGENT-CONVERGENT FIELD resulting from quanta imbalances

radiative EM fields are created by all Matter as it seeks equilibrium with its environment, the impedance of space creates a propagation velocity equal to the speed of light

convergent Gravitational fields

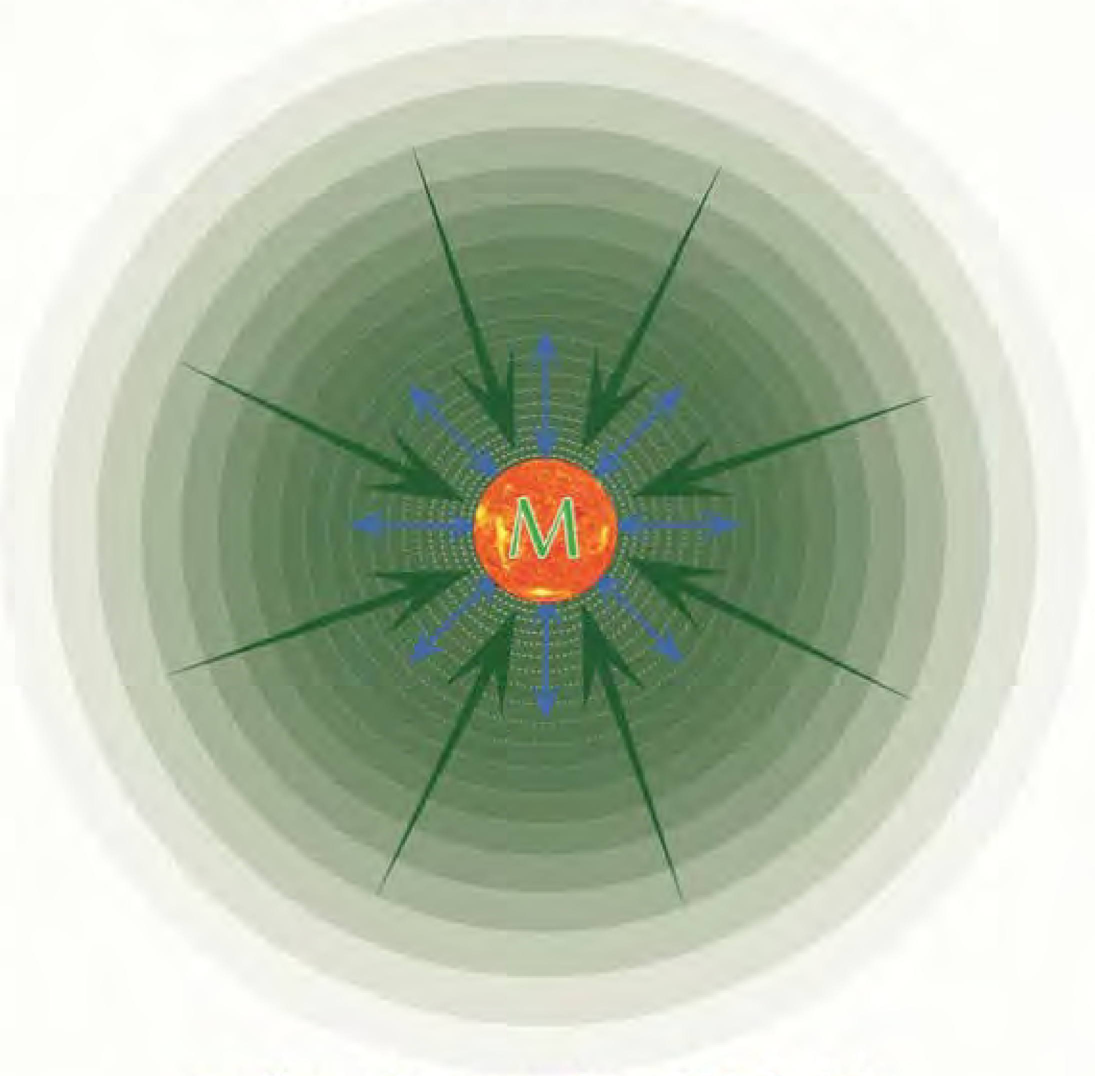

radiant interactive EM fields

Gravity and ElectroMagnetic fields

Newtonian Gravity

Gravity & Electro-Magnetic field strengths
all converge to zero at the centre of the sphere

General Relativity

All EM fields and
Matter are comprised
of charged EM quanta

The total relativisitc
energy of any object
is the sum of its mass-
Matter & Kinetic energies

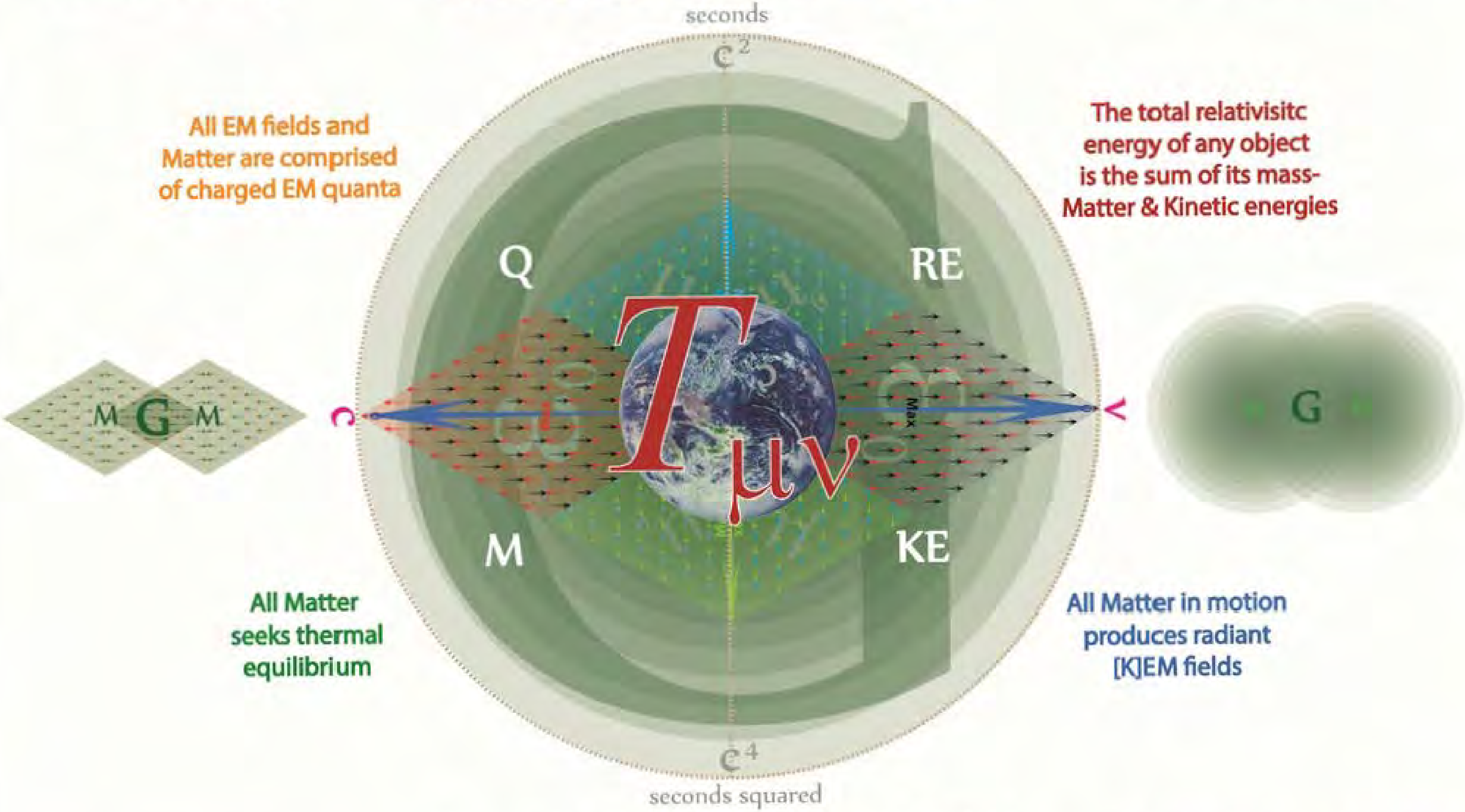

All Matter
seeks thermal
equilibrium

All Matter in motion
produces radiant
[K]EM fields

Charges are Interactive

Gravimetric fields can be modelled as weak EM fields
[space-time curvatures] created by Matter

Gravity is always attractive

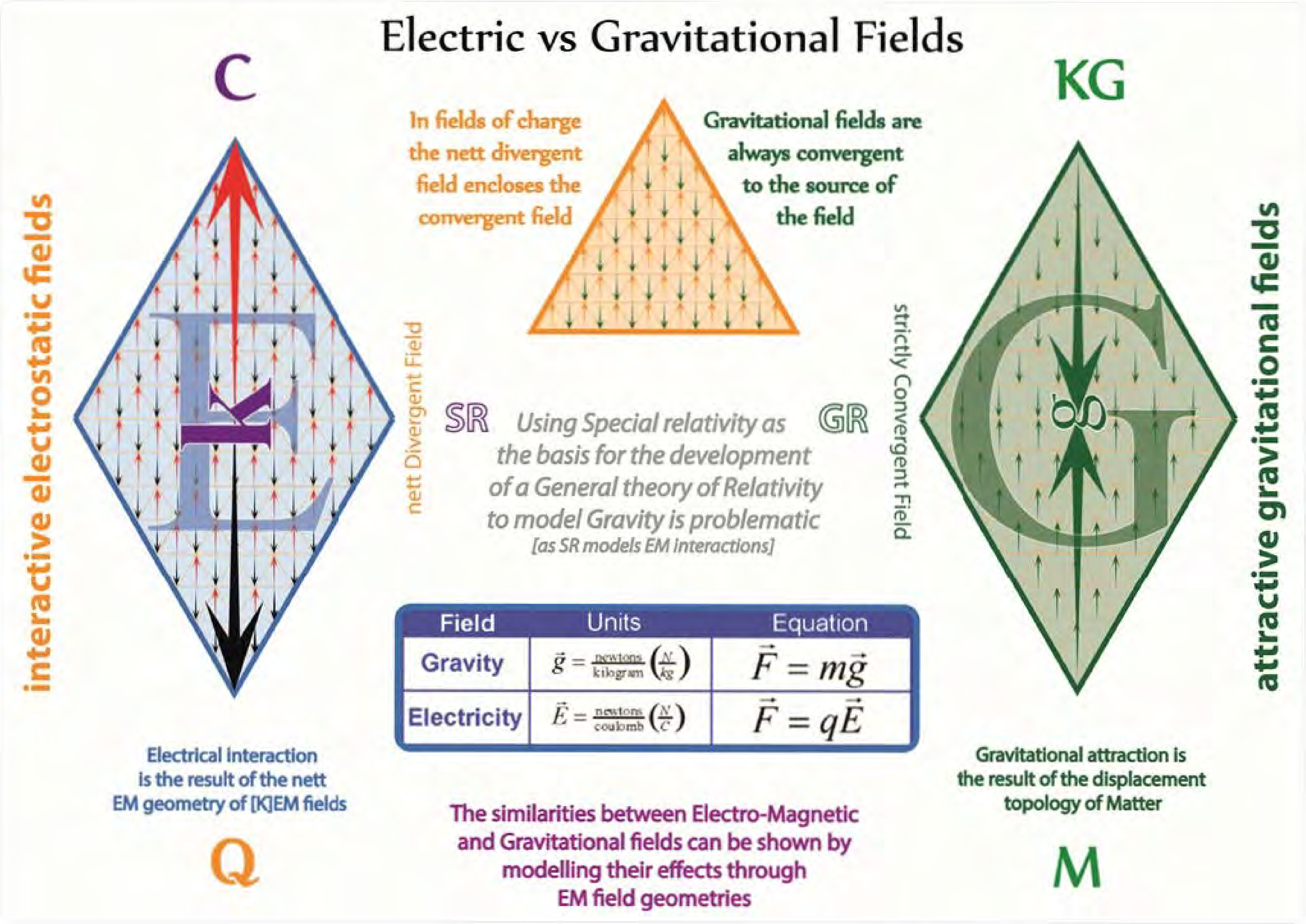

Electric vs Gravitational Fields

C
KG
interactive electrostatic fields
E
K
nett Divergent Field
strictly Convergent Field
G
g
attractive gravitational fields
SR
GR
Q
M

In fields of charge the nett divergent field encloses the convergent field

Gravitational fields are always convergent to the source of the field

Using Special relativity as the basis for the development of a General theory of Relativity to model Gravity is problematic

[as SR models EM interactions]

Field	Units	Equation
Gravity	$\vec{g} = \frac{\text{newtons}}{\text{kilogram}} \left(\frac{N}{kg}\right)$	$\vec{F} = m\vec{g}$
Electricity	$\vec{E} = \frac{\text{newtons}}{\text{coulomb}} \left(\frac{N}{c}\right)$	$\vec{F} = q\vec{E}$

Electrical interaction is the result of the nett EM geometry of [K]EM fields

Gravitational attraction is the result of the displacement topology of Matter

The similarities between Electro-Magnetic and Gravitational fields can be shown by modelling their effects through EM field geometries

Tetryonics 66.01 - Electric vs. Gravitational fields

Copyright ABRAHAM [2008] - All rights reserved

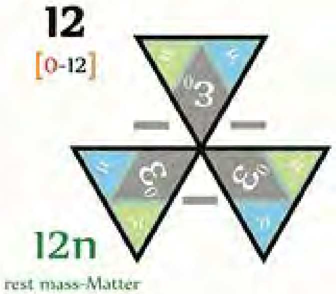

Electric fields and Gravitation

Electrostatic forces are much greater than Gravitational forces (by about 10^20 times).
The force of gravity is attractive for all Matter, whereas electrostatic forces are repulsive for like charged Matter,
There are no negative [repulsive] gravitational charges, while there are both positive and negative electric charges.

Electrostatic forces are interactive
while Gravitational forces are always attractive

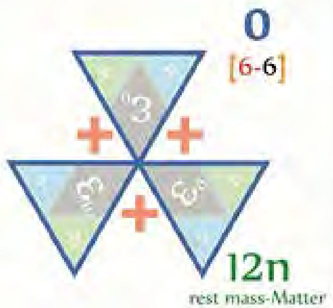

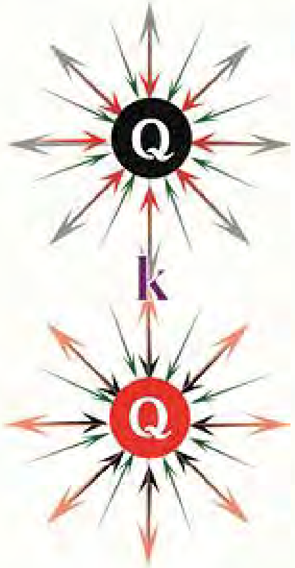

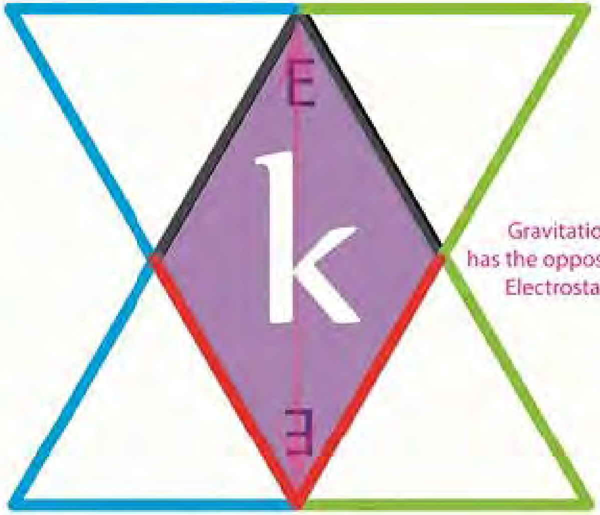

Gravitational attraction
has the opposite force vector of
Electrostatic interaction

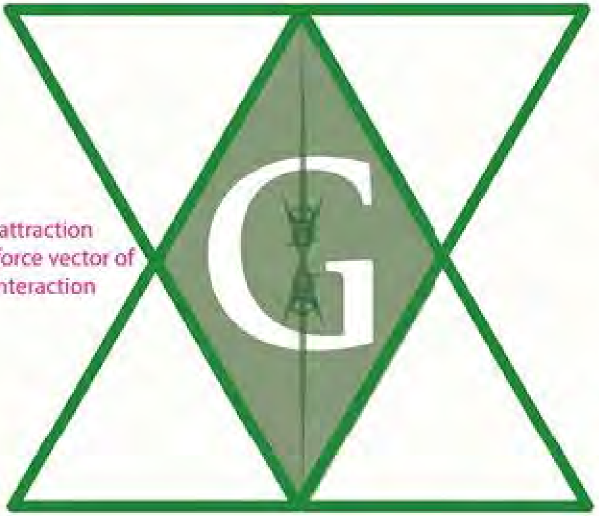

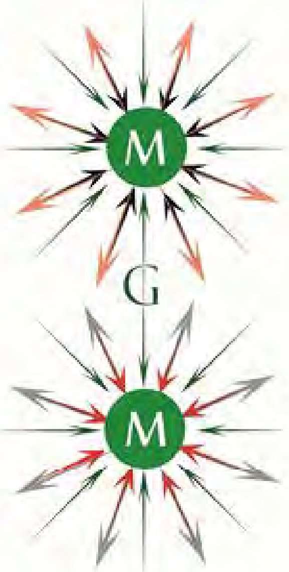

All ElectroMagnetic field quanta are subject to
Lorentz wavelength-frequency corrections due to velocity

Gravitational Matter topologies are NOT affected by
Lorentz corrections due to their velocity of motion

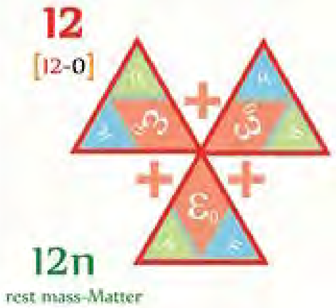

There are similarities between the Electric field and the Gravity field of Matter,
(so sometimes Matter is refered to as "gravitational charge")

Both act in a vacuum and are central and conservative.
Both obey an inverse-square law (both are inversely proportional to square of r).
Both field types propagate at the finite speed of c, the speed of light.
Electric charges and EM mass-energies are conservative

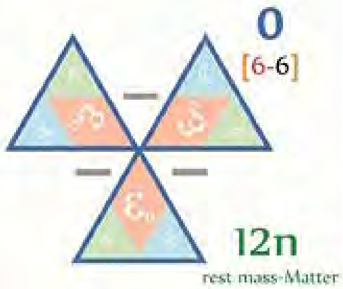

Copyright ABRAHAM [2008] - All rights reserved

Charge vs Gravitation

[empirically determined value]

8.987551787e9 N

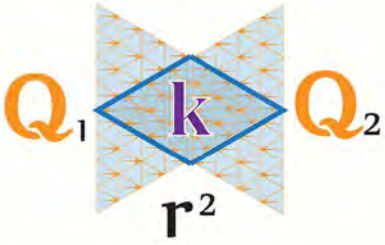

"The magnitude of the electrostatic force between two charges is directly proportional to the product of the magnitudes of each charge and inversely proportional to the square of the distance between the charges."

Charge interaction can be modelled as a DIVERGENT EM FIELD

We have two fundamental laws stating that the forces between two particles are directly proportional to the product of their charges/masses and inversely proportional to the square of the distance between them

Gravitational attraction can be modelled as a CONVERGENT EM FIELD

"The magnitude of the gravitational force between two masses is directly proportional to the product of the magnitudes of each mass and inversely proportional to the square of the distance between the charges."

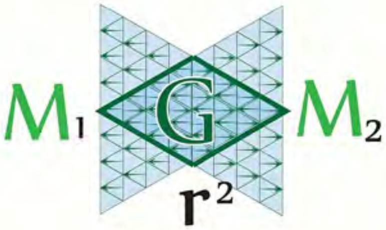

6.67428e-11 N

[empirically determined value]

Coulombic interaction is proportional to Charge (energy geometry imbalances)

$$F = k \frac{Q_1 Q_2}{r^2}$$

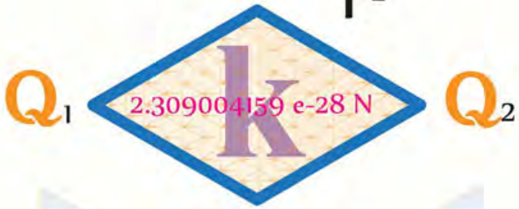

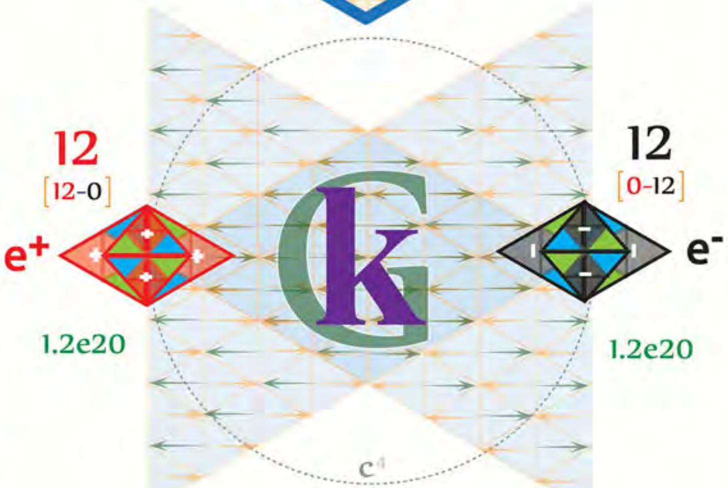

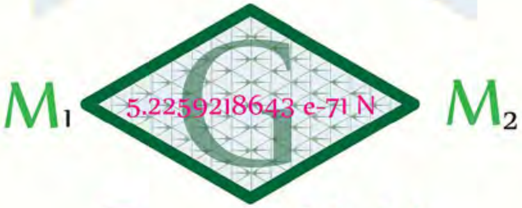

$$F = -G \frac{M_1 M_2}{r^2}$$

Gravitational attraction is proportional to Matter (total rest EM mass-energy content)

*It can now be clearly demonstrated that **Newtonian Gravity** can be modelled as a weak convergent Electro-Magnetic Force*

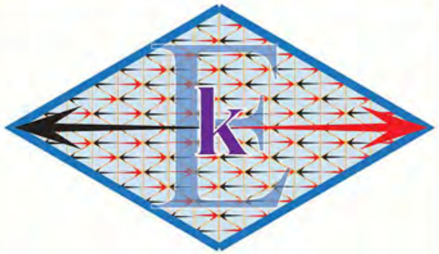

$$\frac{k \quad 8.9875 \text{ e9 N}}{G \quad 6.67384 \text{ e-11 N}}$$

= 1.346 e20

The field strength differential is reflective of the Compton frequency of electron [1.20 e20 quanta]

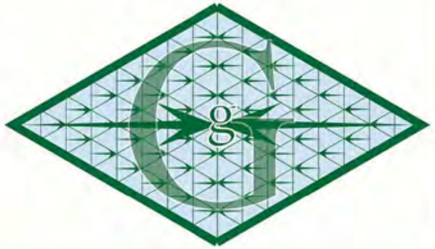

Einstein's general relativity offers a different model of Gravitational attraction however

Gravito-Electro-Magnetic interaction

$$F \propto \frac{Q_1 Q_2}{r^2}$$

The ElectroMagnetic [Coulombic] field strength is proportional to an object's relativistic charge

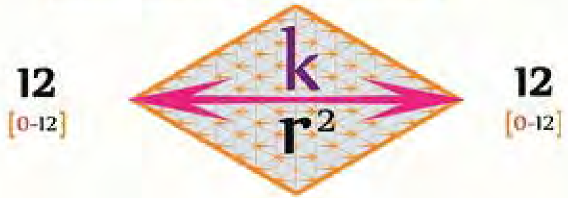
12
[0-12]
k
r^2
12
[0-12]

$$F = k\frac{Q_1 Q_2}{r^2}$$

Charge interactions can be modelled as a superpositional
DIVERGENT FORCE

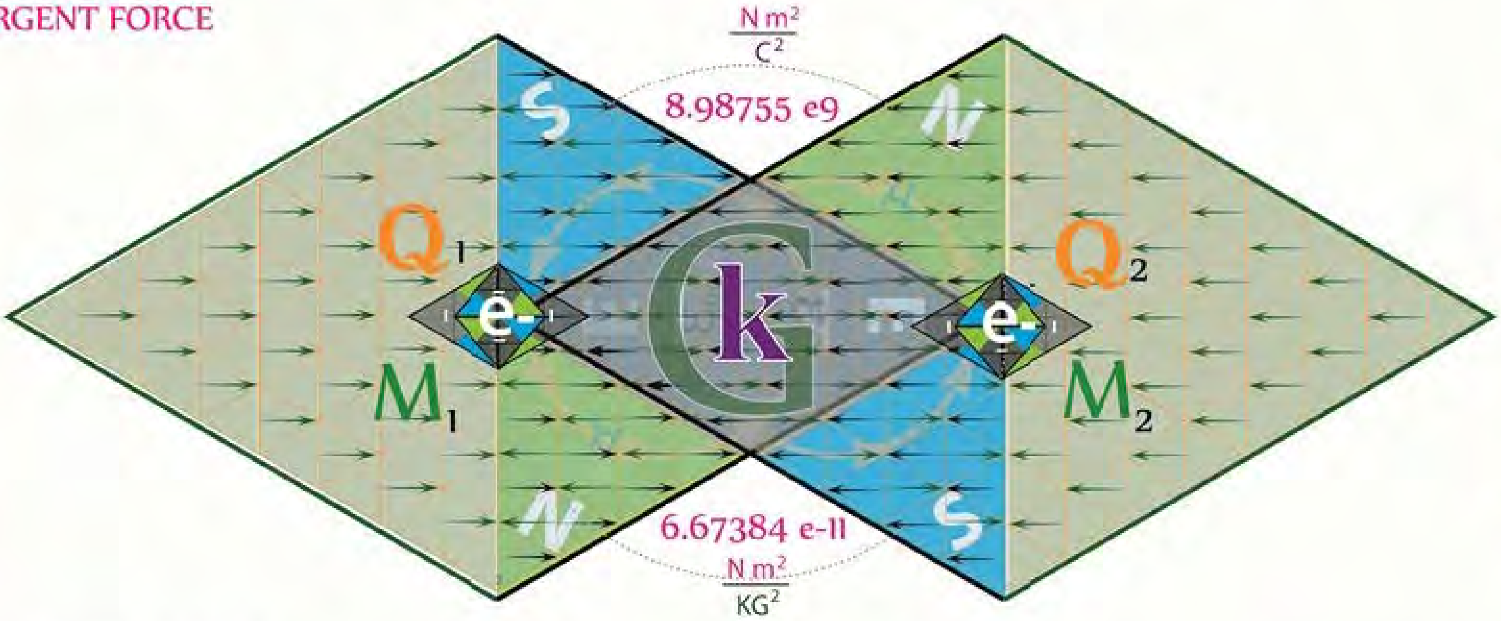
$\frac{N\,m^2}{C^2}$
8.98755 e9
S
N
Q_1
e-
M_1
G
k
Q_2
e-
M_2
N
S
6.67384 e-11
$\frac{N\,m^2}{KG^2}$

equilateral mass-energy momenta quanta
create the interactive fields of Force
between material bodies of Matter

The attractive strength of Gravity
is 20 orders of magnitude weaker than
the strength of Electrostatic interaction

Gravitational attraction can be modelled as a superpositional
CONVERGENT FORCE

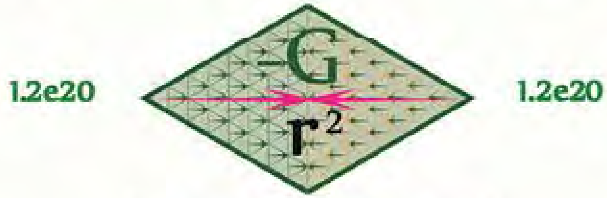
1.2e20
-G
r^2
1.2e20

$$F = -G\frac{M_1 m_2}{r^2}$$

$$F \propto \frac{M_1 m_2}{r^2}$$

Gravitational field strenth is proportional to the rest Matter topology (and can be modelled with weak coulombic fields)

Charged Matter Gravitational interactions

**Electrostatic force fields dominate Gravitational fields
and can be modelled via their superpositioned
energy momenta force fields**

*Following the field symmetries
provided by Tetryonic geometry
we can now model all fields*

$$F = k\frac{Q_1 Q_2}{r^2}$$

2.30900415 e-28 Newtons

*Electromagnetic interactions
as well as the convergent
Gravitational atraction*

**Proton
Matter-charge
9.6577107 e7 C/kg**

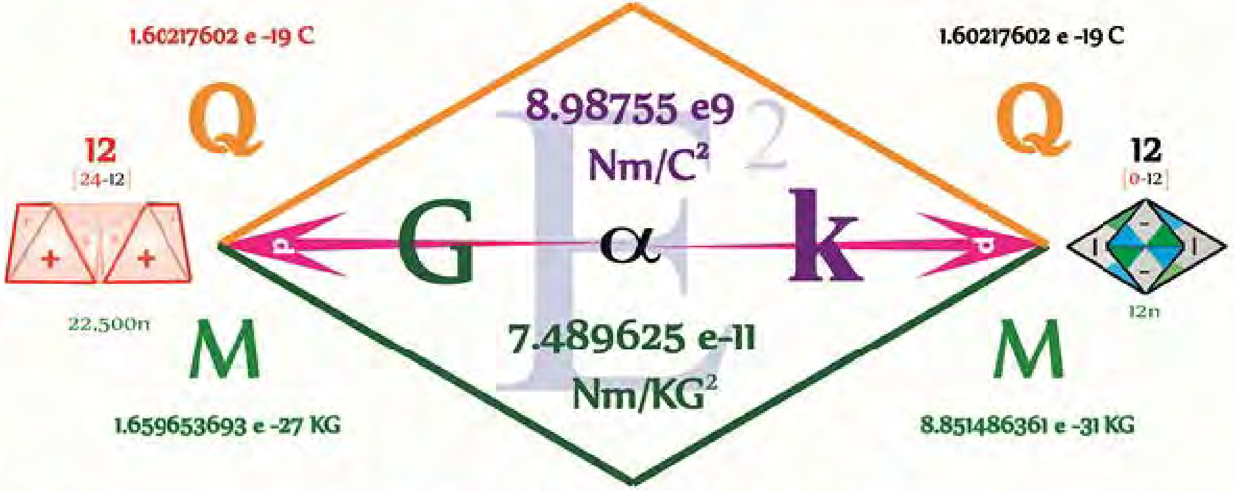
1.60217602 e -19 C
Q
12
[24-12]
+
+
22,500n
M
1.659653693 e -27 KG
8.98755 e9
Nm/C²
E²
G
α
k
p
p
7.489625 e-11
Nm/KG²
1.60217602 e -19 C
Q
12
[0-12]
12n
M
8.851486361 e -31 KG

**electron
Matter-charge
1.810820762 e11 C/kg**

*Reflecting all of the
geometric force interactions in terms of
EM mass-ENERGY & Matter*

1.10025594 e-67 Newtons

$$F = -G\frac{M_1 m_2}{r^2}$$

*Allowing an accurate value
of the graviationl constant
to be derived*

**Gravity [like electrostatic] fields
can be modelled as superpositioned
Energy momenta force fields**

The Gravitational Constant

The gravitational constant is a difficult physical constant to measure to high accuracy, in fact it was not measured until over seventy years after Newton's death by Henry Cavendish with his torsion beam experiment, performed in 1798

G

The gravitational constant is denoted by letter G, it is an empirical physical constant involved in the calculation of the gravitational force between two material bodies.

It usually appears in Sir Isaac Newton's law of universal gravity, and in Albert Einstein's theory of general relativity.

Henry Cavendish

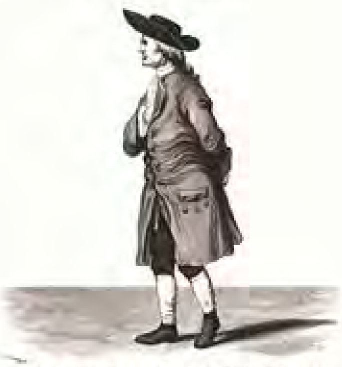

(10 October 1731 – 24 February 1810)

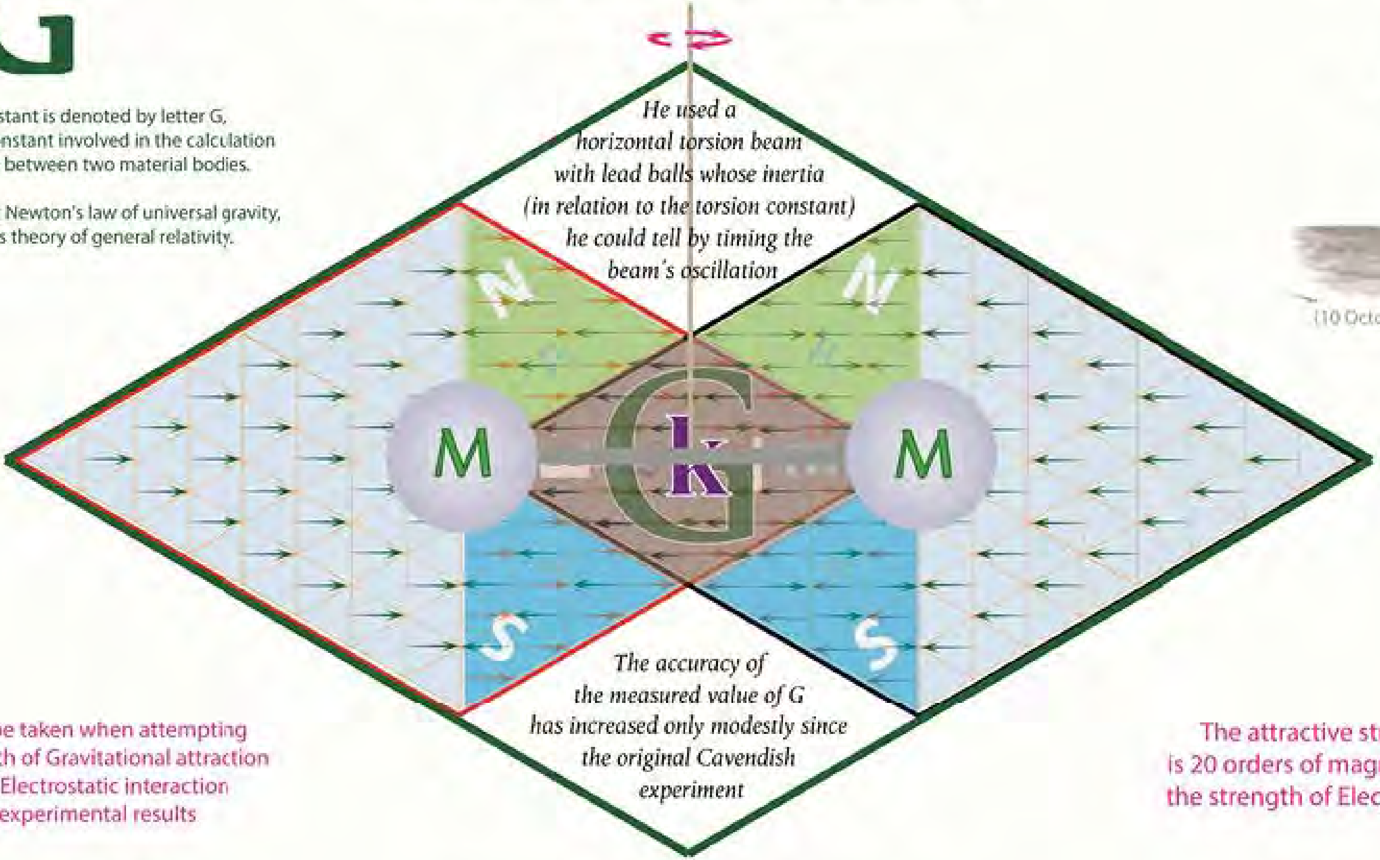

He used a horizontal torsion beam with lead balls whose inertia (in relation to the torsion constant) he could tell by timing the beam's oscillation

The accuracy of the measured value of G has increased only modestly since the original Cavendish experiment

Extreme care must be taken when attempting to measure the strength of Gravitational attraction to ensure that NO Electrostatic interaction influences the experimental results

The attractive strength of Gravity is 20 orders of magnitude weaker than the strength of Electrostatic interaction

8.98755 e9 N

$$F = k \frac{Q_1 Q_2}{r^2}$$

It is also known as the universal gravitational constant, Newton's constant, and colloquially as Big G.

g

It should not be confused with "little g" (g), which is the local gravitational field (equivalent to the free-fall acceleration), especially that at the Earth's surface.

G

6.67384 e-11 N

$$F = -G \frac{M_1 m_2}{r^2}$$

Gravity has no established relation to other fundamental forces, so it does not appear possible to calculate it indirectly from other constants that can be measured more accurately, as is done in some other areas of physics

Fundamental field strengths

are the result of equilateral energy momenta geometries

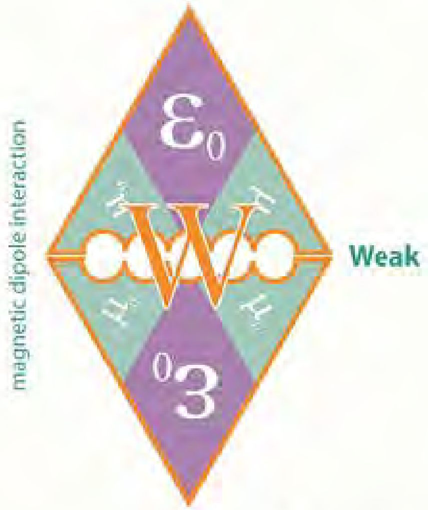
magnetic dipole interaction
ε_0
W
^{0}3
Weak

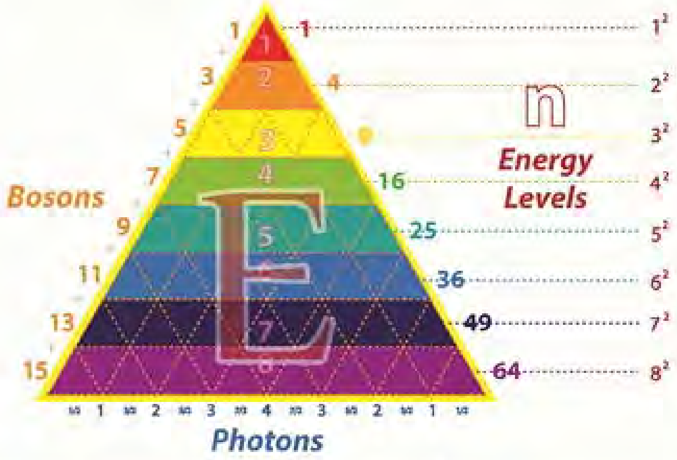
Bosons
1
3
5
7
9
11
13
15
1
4
16
25
36
49
64
1²
2²
3²
4²
5²
6²
7²
8²
n
Energy
Levels
E
Photons

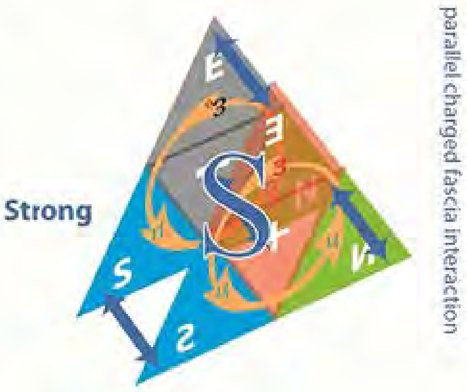
Strong
S
parallel charged fascia interaction

Force	Approximate Relative Strength	Range
Strong Nuclear	10^{38}	10^{-15}
ElectroMagnetic	10^{36}	Infinite
Weak	10^{2}	10^{-18}
Gravitational	1	infinite

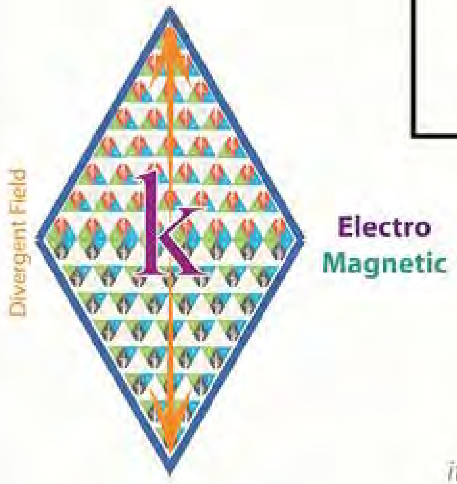
Divergent Field
k
Electro
Magnetic

All the field strengths of the fundamental interactions are a direct result of the interactions of their Tetryonic mass-energy goemetries and Tetryonic Matter topologies

m E M

geometries density topologies

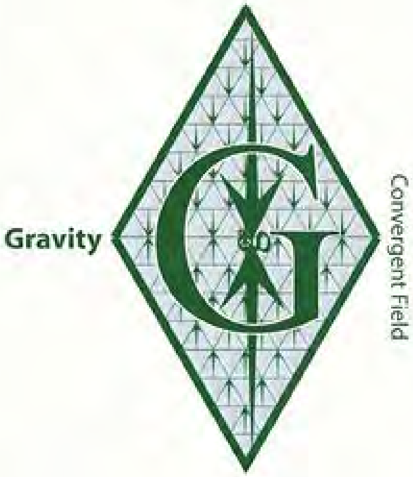
Gravity
G
Convergent Field

(whilst Gravity can be modelled as a weak convergent Coulombic attraction its actual physics are the result of vacuum energy displacement by Matter topologies)

The Strength of the Gravitational Constant

Coulomb's Constant

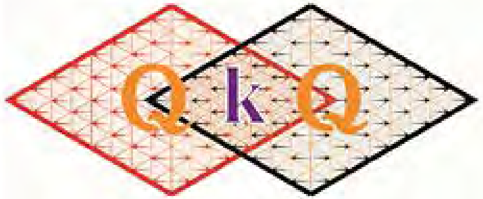

[24-12]

Proton nett charge

1.60217646 e -19 C

equal
(but opposite)
nett Charges

[0-12]

Electron nett charge

1.60217646 e -19 C

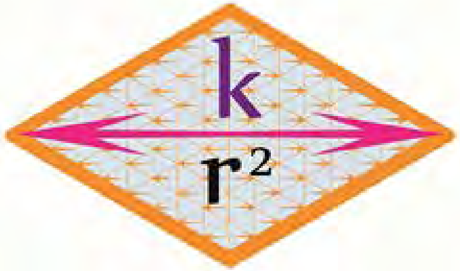

8.987551787e9 N

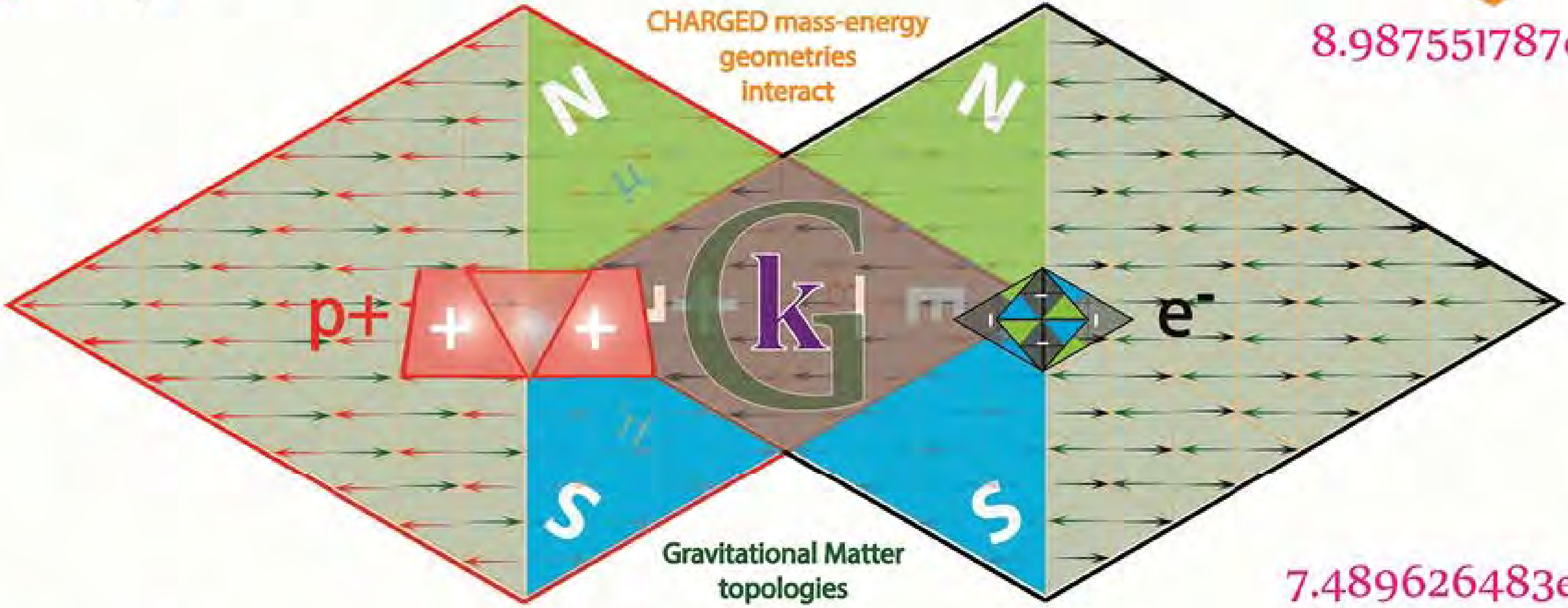

Gravitational Matter
topologies
attract

7.489626483e-11 N

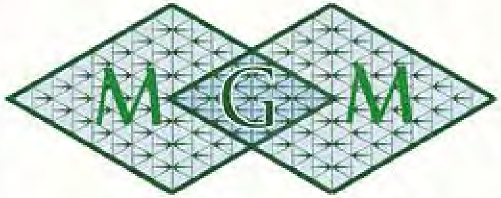

Newton's Constant

1.659653693 e -27 KG

Proton rest mass-Matter

[1.5e23-7.5e22]

1875

8.851486361 e -31 KG

electron rest mass-Matter

[1.2e20]

Although it provides a coherent and accurate model of all the motive forces & interactions present the modelling of Gravity with EM fields does not account for the source, strength & unilateral direction of Gravitational fields

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Matter interaction and Gravitation

Electrostatic convergence can serve as a model for Gravitational attraction

All Matter is comprised of EM mass-Energies

Q

10,080

[5,040-5,040

Charged mass-energy geometries form Matter topologies

$$F = -G\frac{M_1 m_2}{r^2}$$

EM mass & Matter must be differentiated

Matter

$$\frac{4n\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

M

7,036,992

[5.190653044e-25 Kg

All Matter is gravitationally attractive

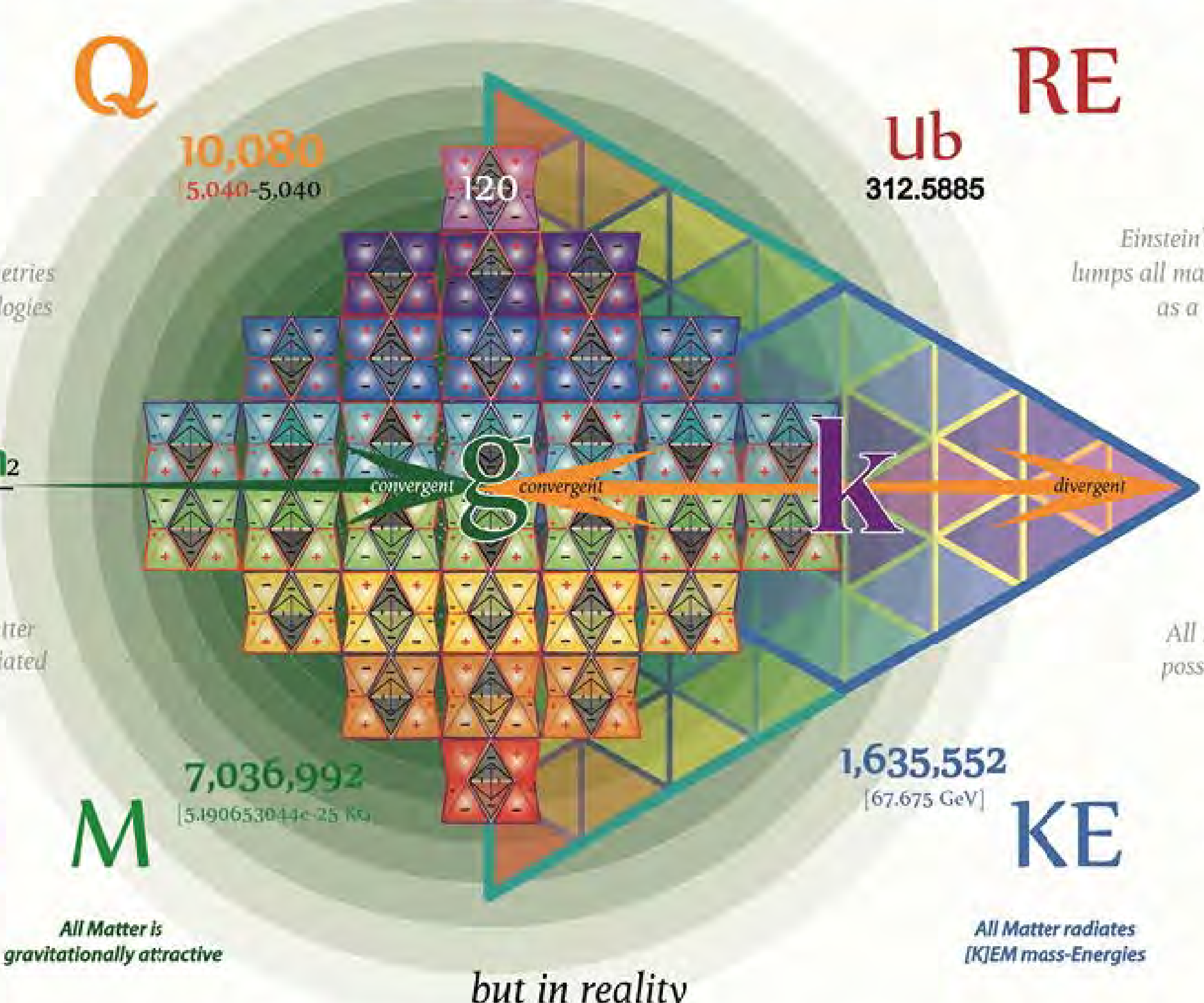

ub

312.5885

RE

All gravitational systems are comprised of Matter & KEM fields

Einstein's stress-energy tensor lumps all mass-Energy-matter together as a pressure gradient

$$F = k\frac{Q_1 Q_2}{r^2}$$

All Matter in motion posseses [K]EM fields

1,635,552

[67.675 GeV]

KE

EM waves

$$\frac{\text{EVEN}\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter radiates [K]EM mass-Energies

but in reality

Gravitational attraction results from the displacement of vacuum energies by Matter topologies

[The effects of stronger [K]EM fields can be excluded from the long distance attractive [gravitational] motion of material bodies to each other]

Copyright ABRAHAM [2008] - All rights reserved

Modelling the force of Gravity

It is in many ways very similar to Coulomb's law of Interaction
between charged particles developed by Coulomb
(after Newton's Gravitation)

Gravity can be modelled as weak convergent EM field

For many people, it is easiest to picture gravity
as a kind of special invisible force,
emanating from ALL the Matter in the universe.
Newton's law of Gravity leads one to sees
it this way and it is probably the view
most of us were given when we learned
about gravity in elementary school.

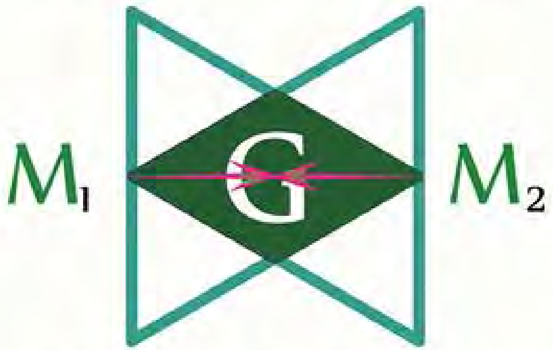

However, there is another way
to model Gravity's effects.
It stems from Albert Einstein's
General theory of Relativity,
a theory that describes gravitation
as being the result of the curvature of
spacetime as a result of energy gradiant

Gravity can also be modelled as convergent pressure gradients

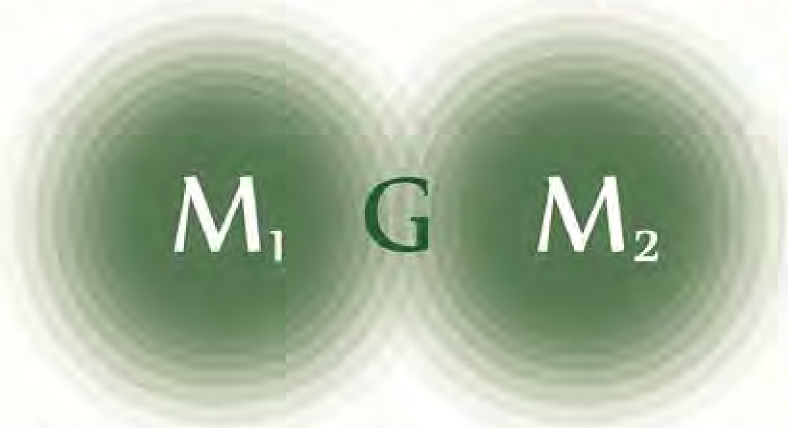

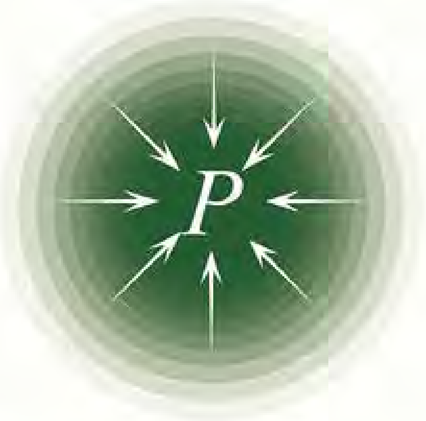

To understand what gravity really is you must have a clear understanding of the distinctions between radiant EM mass and standing-wave Matter and the various quantum interactions they produce.

Tetryonics provides a clear geometric definition and distinction between 2D radiant mass-energy geometries and 3D Matter topologies along with a new unified model of the mechanics of Universal Gravitation at all scales

radiant mass-energy geometries

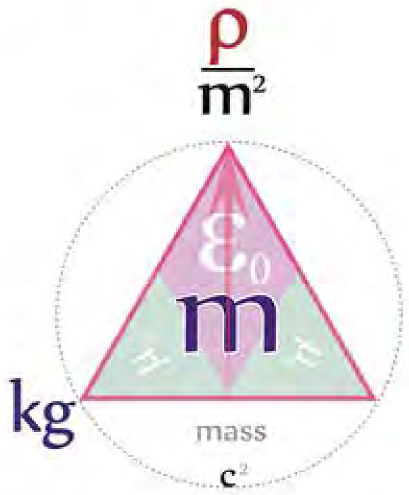

'Action-at-a-Distance'

$$\overset{\text{mass-energies}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

[K]EM mass-Energies with QAM create interactive divergent [and convergent] charge geometries

Newtonian Gravity

Toplogical mass is Matter

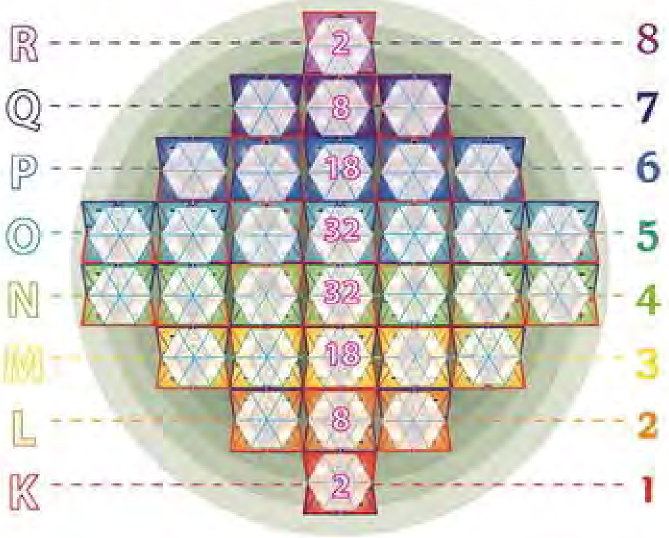

All Matter contains nullspaces and produces motional KEM fields

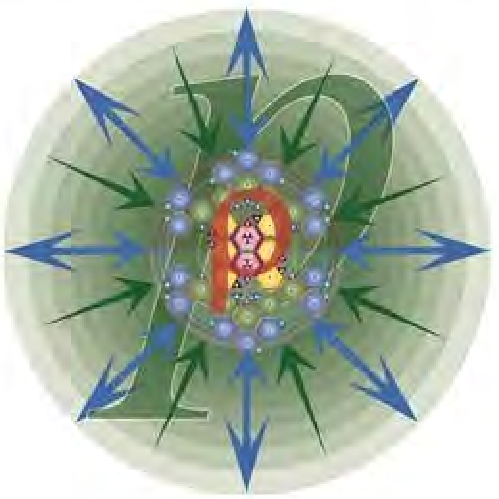

standing-wave Matter toplogies

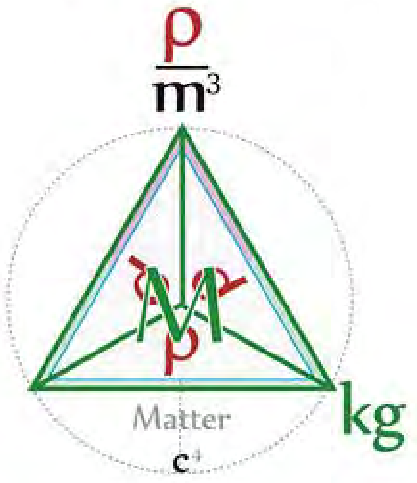

'Folded Space-Time'

$$\overset{\text{Matter}}{\frac{4n\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Tetryonic Matter creates low pressure nullspace resulting in convergent Energy pressure gradients

General Relativity

Instantaneous action-at-a-distance

Tetryonic QED theory shows that where the linear energy momentum of longitudinal waves align and act in a co-linear direction they can facilitate near instantanous action-at-a-distance

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Free particles move with constant vector-velocity

(that is, with zero acceleration, unless acted upon by an outside force).

The Earth orbits the SUN where it is now about a common centre of gravity

'Action-at-a-Distance' and 'Instantaneous' forces were introduced into Physics within Newton's notion of absolute space and time

If the SUN were to vanish the Earth would cease traveling in its current elipical orbit immediately

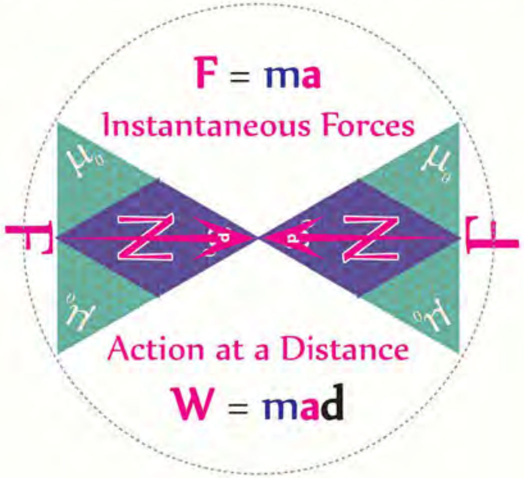
F = ma
Instantaneous Forces
μ_0
μ_0
F
N
p
p
N
F
μ_0
μ_0
Action at a Distance
W = mad

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

The vector-force on a particle equals the product of its mass by its vector acceleration

$$\mathbf{F} = m\mathbf{a} = \frac{\Delta \mathbf{p}}{\Delta t} = \Delta\left[\frac{m\mathbf{v}}{t}\right]$$

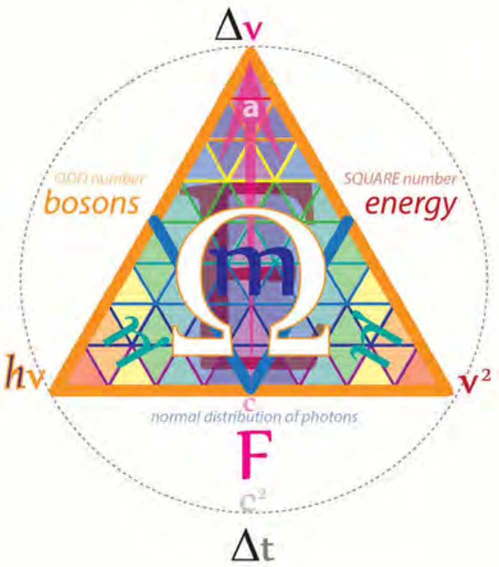
Δv
a
ODD number
bosons
SQUARE number
energy
m
Ω
hv
v^2
c
normal distribution of photons
F
c^2
Δt

The speed of gravity is limited to c

Einstein's GR theory replaces Newton's force of gravity and absolute space & time with a relativistic spacetime gradient curved by mass-energies

If the SUN were to vanish the Earth would continue traveling in its current orbit for 8$^{1/3}$ minutes

The forces of action and reaction are equal and opposite;

(if a particle A exerts a force f on a particle B, then B exerts a force –f on A)

$$\sum \mathbf{F}_{a,b} = -\sum \mathbf{F}_{b,a}$$

Copyright ABRAHAM [2008] - All rights reserved

Newton's Gravitational Constant

$$F = -G\frac{M_1 m_2}{r^2}$$

According to Newton's law of universal gravitation,

the attractive force (F) between the centres of two bodies of mass-Matter, is proportional to the product of their Matter (M1 and m2), and inversely proportional to the square of the distance (r) between them

Long distance force

spacetime curvature

Newton's Gravitional Force models a motive force for gravitational acceleration

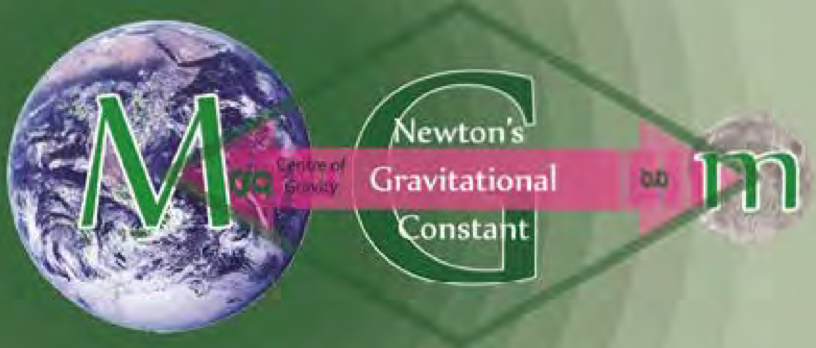

Einstein's General relativity has no motive force for gravitational accelerations

6.67384 e-11

$$\left[\frac{Nm}{kg}\right].\left[\frac{Nm}{kg}\right]$$

The gravitational force between two interacting bodies of Matter is

$$\left[N\frac{m^2}{kg^2}\right] = \left[kg\frac{m}{s^2}\frac{m^2}{kg^2}\right] = \left[\frac{m^3}{kg\,s^2}\right]$$

$4\pi G\rho$

$8\pi G\rho$

Copyright ABRAHAM [2008] - All rights reserved

Action-at-a-Distance

All Matter creates KEM fields of energy momenta

All Matter creates a displacement of vacuum energies

radiant mass-energies INTERACT

Once established longitudinal EM fields between Matter form an 'inelastic rod' of energy momenta that can facilitate an instantaneous 'Action-at-a-Distance'

standing-wave Matter ATTRACTS

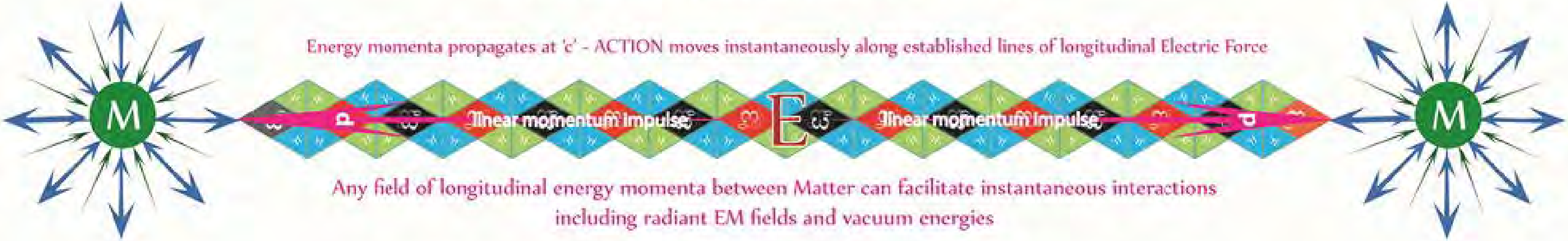

Energy momenta propagates at 'c' - ACTION moves instantaneously along established lines of longitudinal Electric Force

Any field of longitudinal energy momenta between Matter can facilitate instantaneous interactions including radiant EM fields and vacuum energies

Gravity is known as a Universal force of attraction between two objects that is proportional to their mass-Matter content

Newton's theory of gravity offered no prospect of identifying any mediator of gravitational interaction. His theory assumed that gravity acts instantaneously, regardless of distance.

Energy momenta creates universal motive forces

According to Albert Einstein's theory of special relativity, instantaneous action-at-a-distance was seen to violate the relativistic upper limit on speed of propagation of information.

If one of the interacting objects were to suddenly be displaced from its position, the other object would feel its influence instantaneously, meaning information had somehow been transmitted faster than the speed of light.

Gravity fields create a universal attractive field

Copyright ABRAHAM [2008] - All rights reserved

Gravitational Constant

The gravitational constant denoted by letter G, is an empirical physical constant involved in the calculation(s) of gravitational force between two bodies.
It appears in Sir Isaac Newton's law of universal gravitation, and in Albert Einstein's theory of general relativity.
It is also known as the universal gravitational constant, Newton's constant, and colloquially as Big G.

It should not be confused with "little g", which is the local gravitational field (equivalent to the free-fall acceleration), at the Earth's surface

Gravitational Forces
can be modelled

$6.67 \text{ e-11} \left[\frac{Nm}{kg}\right]^2$

linear Newton metres

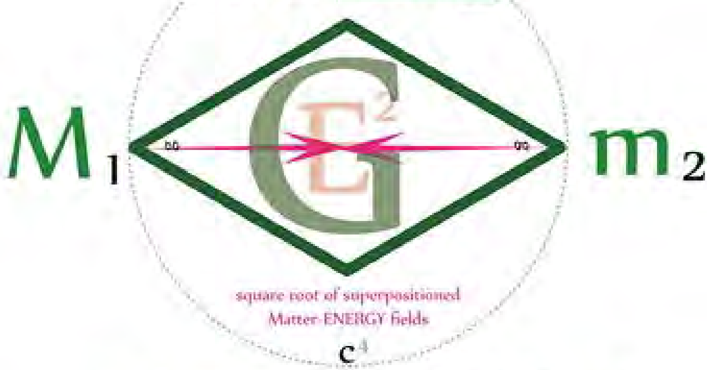

$6.67 \text{ e-11} \left[\frac{E}{kg}\right]^2$

scalar Energy momentum

as the geometric MEAN of
superpositioned fields of gravitational energy

$$F = -G\frac{M_1 m_2}{r^2}$$

The gravitational force between
two neutral interacting bodies is

N.m = Joules

$$\left[\frac{Nm}{kg}\right] \cdot \left[\frac{Nm}{kg}\right]$$

$[N.m/kg]^2 \sim [E/kg]^2$

Gravitational acceleration

Gravitational acceleration can be
viewed as a reflection of the change
in an object's momentum over time

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

$$\frac{\left[\frac{m}{s}\right]}{s}$$

$$\vec{g} = \frac{\vec{F}_g}{m} = -G\frac{mM}{mr^2}\hat{r} = -\frac{GM}{r^2}\hat{r}$$

expressed vectorially the force is always
ATTRACTIVE to the larger masss
and acts in opposition to the
KEM field's divergent momenta

$F=ma$

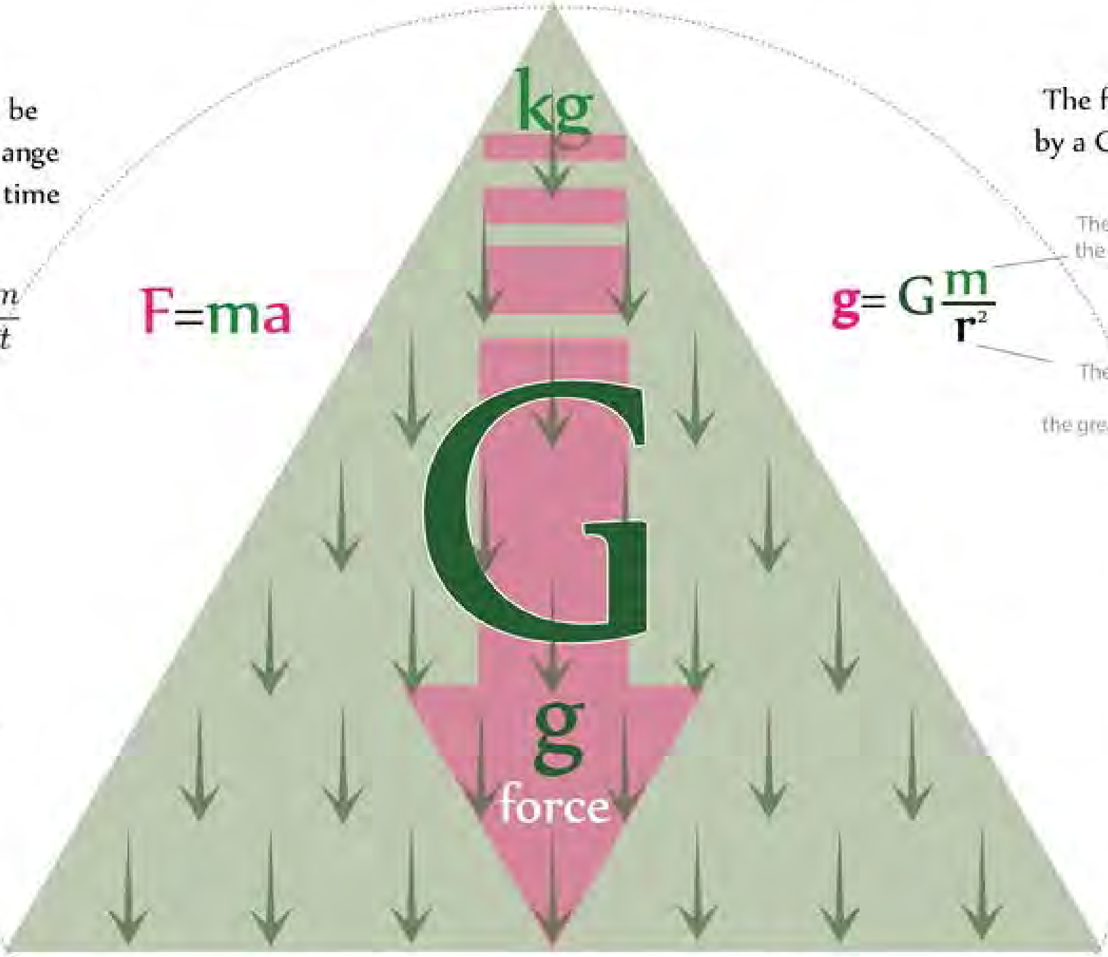

The force on a body acted upon
by a Gravitational field is given by

$$g = G\frac{m}{r^2}$$

The greater the mass of a gravitating object
the greater the acceleration due to the mass

The closer a body is to a the source
of a Gravitational field
the greater the acceleration due to the field

$$\left[\frac{N}{kg}\right] = \left[\frac{m}{s^2}\right] = \frac{\left[kg\frac{m}{s^2}\right]}{kg}$$

gravitational acceleration
can be modelled as the
convergent vector component
of superpositioned EM fields

$$\left[\frac{Nm}{kg}\right]^2$$

The gravitational force between
TWO neutral interacting bodies is

$$\left[\frac{Nm}{kg}\right].\left[\frac{Nm}{kg}\right]$$

Copyright ABRAHAM [2008] - All rights reserved

Modelling Gravitational fields

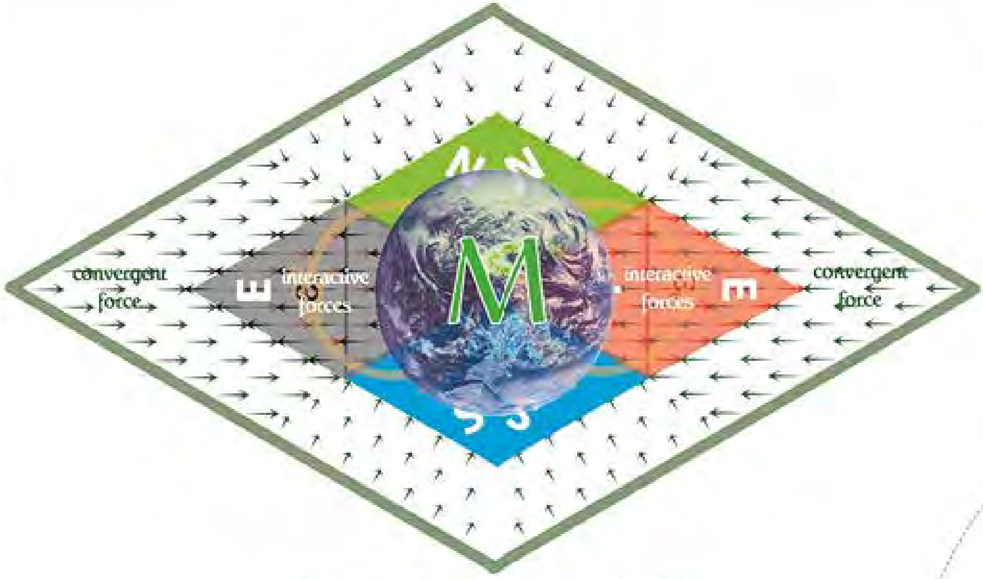
convergent force
E
interactive forces
N
M
S
interactive forces
E
convergent force

The [K]EM fields of planetary bodies are the result of their mass-Matter density and motion in space

Gravity can be modelled as weak convergent force created by and acting between the mass-Matter in planetary bodies

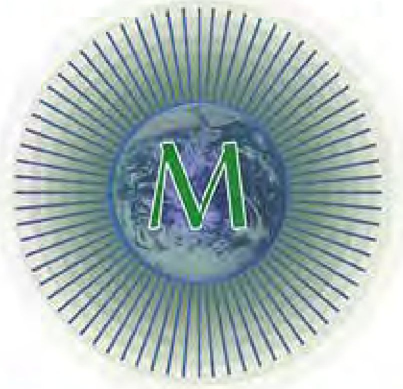

Matter topologies
displace vacuum energies
creating Gravity

Weak KEM fields
can accurately model
Gravitational fields

Gravitational fields are the result of the displacement of vacuum energies by Matter topologies and the interactive energy momenta of their radiant KEM fields

$$F = k\,\frac{Q_1 Q_2}{r^2}$$

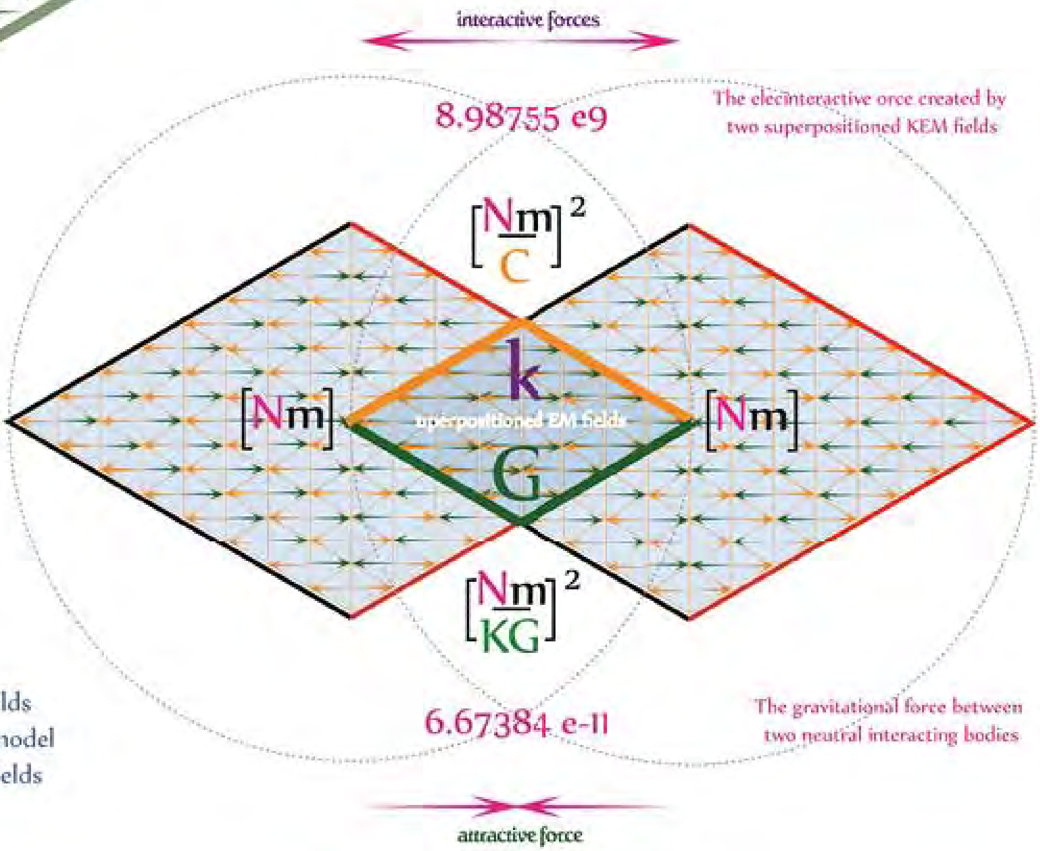
interactive forces
8.98755 e9
The elecinteractive orce created by two superpositioned KEM fields
$\left[\frac{Nm}{C}\right]^2$
[Nm]
k
superpositioned EM fields
G
[Nm]
$\left[\frac{Nm}{KG}\right]^2$
6.67384 e-11
The gravitational force between two neutral interacting bodies
attractive force

$$F = -G\,\frac{M_1 M_2}{r^2}$$

Centres of Gravity

are determined by
the distribution of the combined
gravitational mass-Matter in a system

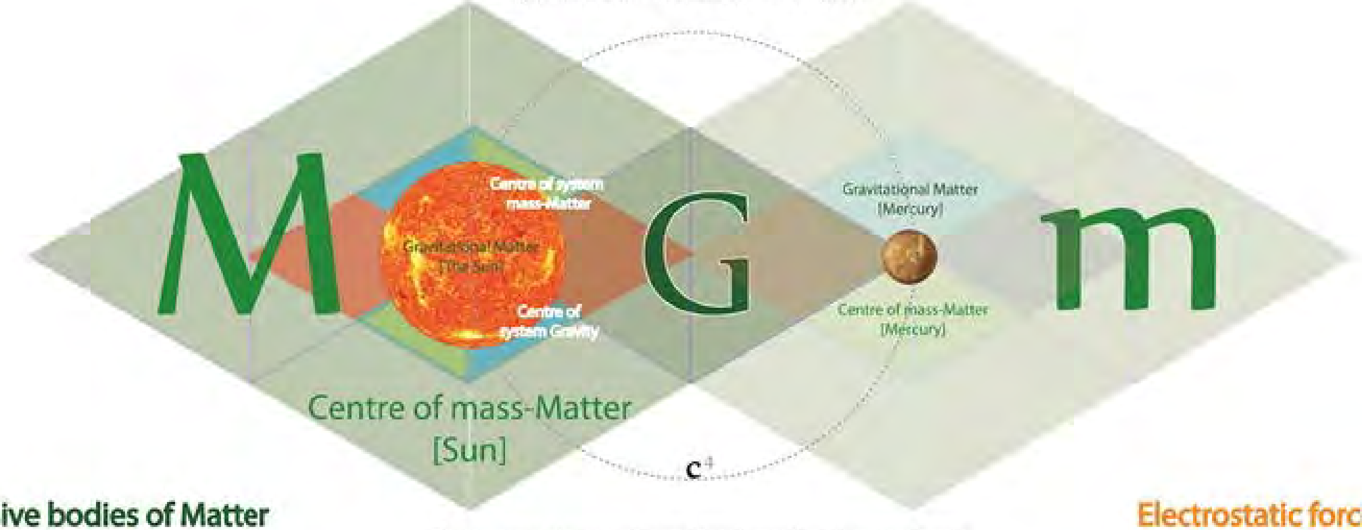
M
G
m
Centre of system mass-Matter
Gravitational Matter [The Sun]
Centre of system Gravity
Gravitational Matter (Mercury)
Centre of mass-Matter (Mercury)
Centre of mass-Matter [Sun]
c^4

Any two massive bodies of Matter (be they Stars or Neutrinos) will orbit about their 'Centre of Gravity'

The term center of mass is often used interchangeably with center of gravity, but they are physically diffenent concepts. They happen to coincide in a uniform gravitational field, but where gravity is not uniform, center of gravity refers to the mean location of the gravitational force acting on a body.

Electrostatic forces are many orders of magnitude greater than gravity & dominate nuclear interactions at the quantum scale

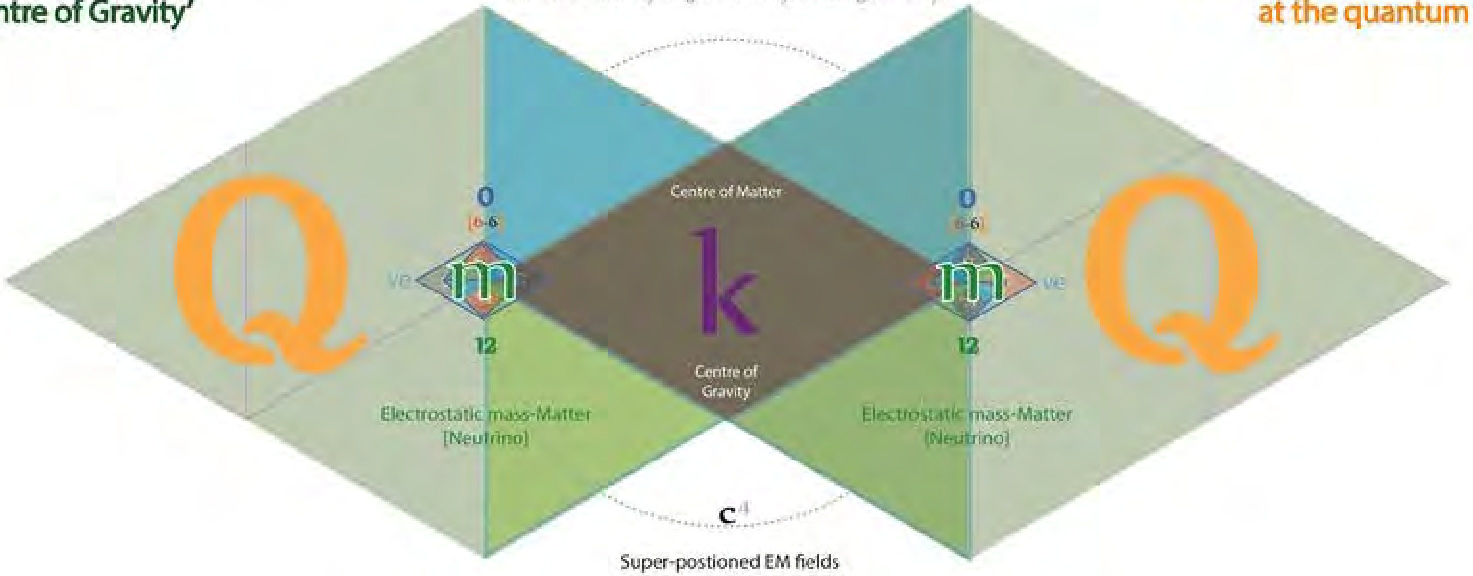
Q
0
(6-6)
m
ve
12
Centre of Matter
k
Centre of Gravity
0
(6-6)
m
ve
12
Q
Electrostatic mass-Matter [Neutrino]
Electrostatic mass-Matter (Neutrino)
c^4

Super-postioned EM fields
have c^4 geometries [c^2x c^2]

$$F=ma$$

There are two quite distinct types of mass in Newton's theory of mechanics and gravitation.

(i) inertial mass, which occurs as the ratio between force and acceleration in Newton's second law
and thus measures a particle's resistance to acceleration, and

(ii) Gravitational mass, which may be regarded as the gravitational analog of electric charge,
and which occurs in Newton's Gravitational equation

$$F_g = -G\,\frac{m}{r^2}$$

Inertial mass
is a measure of the total
energy momenta geometries
of Matter topologies

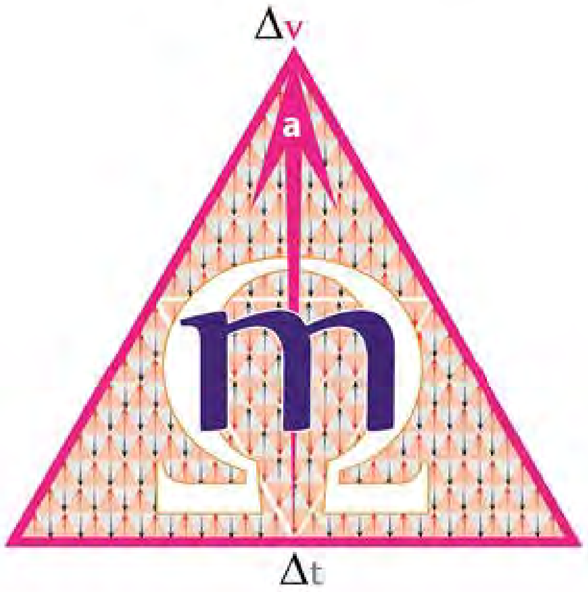

Inertial mass

results from constituent quanta
acting against (or with) any
Force applied to a mass

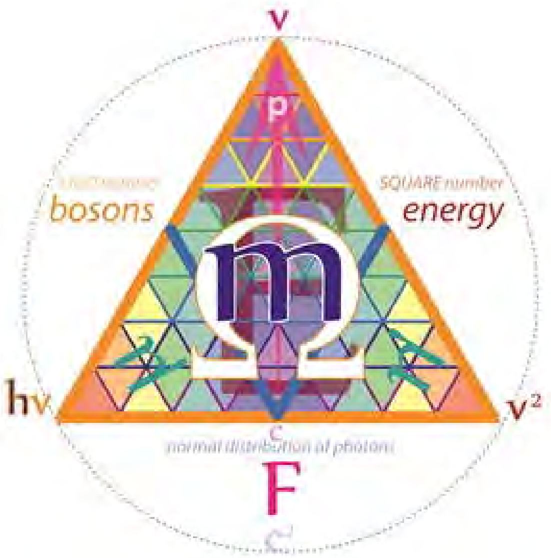

Often stated as the 'Principle of Equivalence'
on which General relativity is founded ,
it can be summarised as follows:

$$a = \frac{m}{m}\,g$$

*"Inertial mass and Gravitational mass
are equivalent"*

***whilst mass-energies contribute equally
to inertial & gravitational mass
it is Matter topologies that creates Gravity***

All Gravitational Matter
has EM mass geometries in
its Matter topology

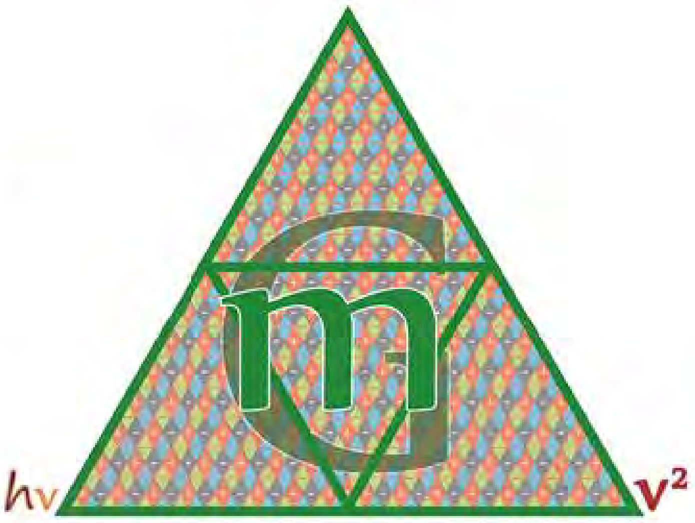

'Gravitational' mass

is seen as being the constituent quanta
of opposite charge polarity
attracting another particle's quanta

Copyright ABRAHAM [2008] - All rights reserved

Discerning between the competing laws of Gravity

$M + KE = RE$

gravitational mass-Matter — *fields of force* — *relativisitic mass-energy densities*

At the heart of resolving the true nature of Gravity is the formal definitions of mass-Energy-Matter and the interplay between the various field interactions they produce and possess

Newton's view of gravity

Sir Isaac Newton described gravity as a force that attracts things possessing Matter

Sir Issac Newton

(25 December 1642 – 20 March 1727)

$$F = -G\frac{M_1 m_2}{r^2}$$

Newton taught that masses attract each other with a force inversly proportional to the distance between the masses.

Newton's idea was that planets are held in their orbits around the Sun by a force proprtional to the mass of the Sun and the planets, and that force is always pointing inwards (i.e. planets get pulled in the direction of the Sun, and the Sun gets pulled in the directions of the planets).

Newton said that a person standing on the Earth is being held to the Earth by a force due to the mass of the Earth and the mass of the person, and that force is always pointing inwards (i.e. a person gets pulled in the direction toward the center of the Earth, and the Earth gets pulled in the direction of that person).

Newton's and Einstein's view of gravity differ markedly

Einstein, on the other hand, taught that all things with mass (including the Sun, Earth, and the rest of the planets) create a energy density gradient that causes spacetime to curve and that this curvature is an alteration the geometry of spacetime.

The curvature casues Matter to move toward the center of the largest nearby mass. The closer to this mass the larger the degree of curvature in the local spacetime.

Additionally, Einstein taught that if there are no external forces acting upon an object (even those without any mass such as photons) then that object will simply follow its natural path (geodesic) through the altered geometry of the local spacetime - regardless of the degree of curvature of that local spacetime. This implies that if an object is somehow kept from following its natural path (geodesic) then a net force does exist and is measurable on that object.

Einstein's view of gravity

Albert Einstein described gravity as a 3D curvature of spacetime caused by Matter

Albert Einstein

(14 March 1879 – 18 April 1955)

Gravity is an instantaneous 'Action-at-a-Distance' mediated by a means unknown

Ricci curvature — *scalar curvature* — *Gravitational constant*

$$R_{\mu\nu} - \tfrac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

stress-energy tensor

spatial geometry

Matter and energy tell spacetime how to curve. Spacetime tells matter how to move

There is no doubt that both Sir Isaac Newton & Albert Einstein were both brilliant, their theories describing Universal Gravitation have remained at the fore for hundreds of years now.

However, a more accurate model of the Universal Gravitation is now at hand and its true nature is revealed via the Tetryonic geometry of EM mass-Energy-Matter and its associated field interactions. Tetryonic Gravitation demonstrates the validity of both approaches and reveals the true geometry behind the field interplays that result in the familiar force of Gravity

Tetryonics 67.11 - The competing Laws of Gravity

The two models of Gravitation

Newtonian gravitation models the observed large-scale convergent force between all material bodies

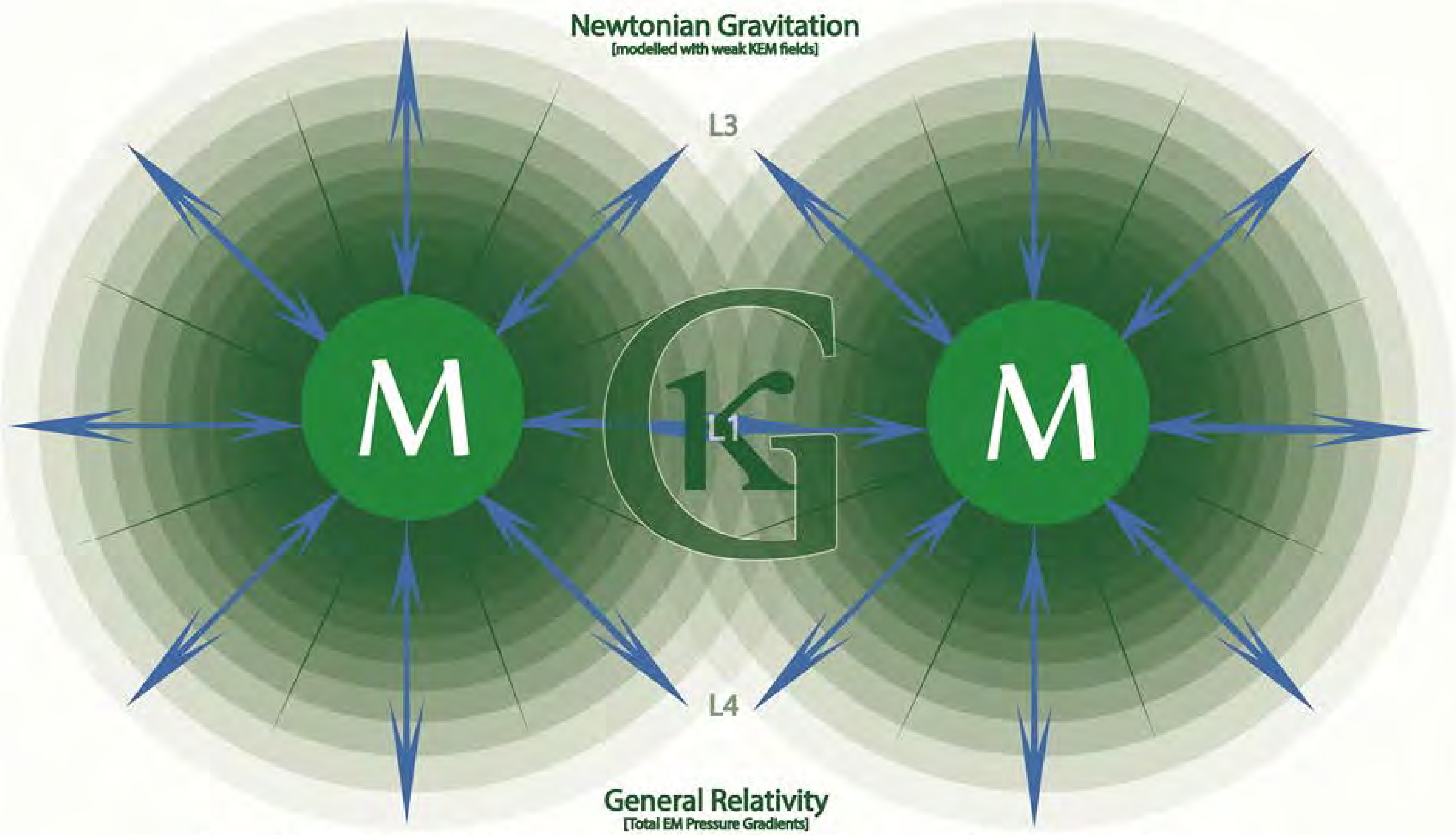
Newtonian Gravitation
[modelled with weak KEM fields]
L3
M
G
K
L1
M
L4
General Relativity
[Total EM Pressure Gradients]

General Relativity models the convergent pressure gradiant created by all forms of energy

While both models are very successful they both fail to reveal the quantum-scale geometries and topologies that give rise to universal gravitation

Theories of gravitation

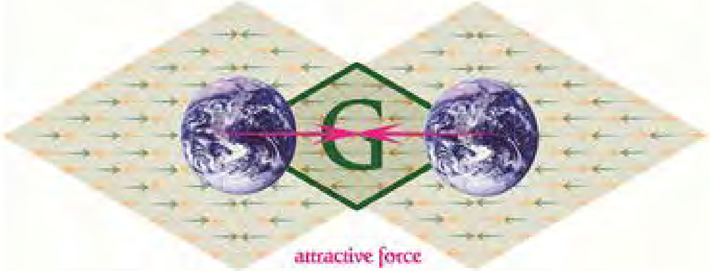
G
attractive force

Newton viewed Gravitation as
an attractive FORCE of Matter
(inverse square field strength)

for the mechanics of gravitation
"I feign no Hypothesis"

"Gravity is the result of
curved spacetime"

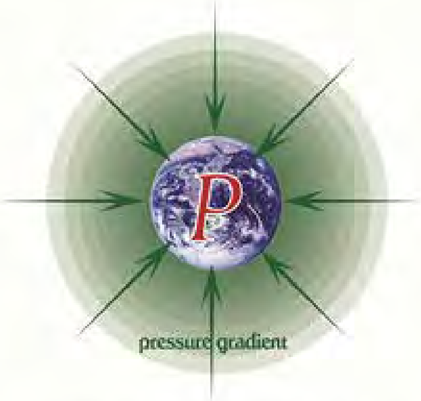
P
pressure gradient

Einstein viewed it as the
curvature of 4D spacetime
(PRESSURE gradient)

$$F = -G\frac{M_1 M_2}{r^2}$$

$$G_{ab} = \kappa T_{\mu\nu}$$

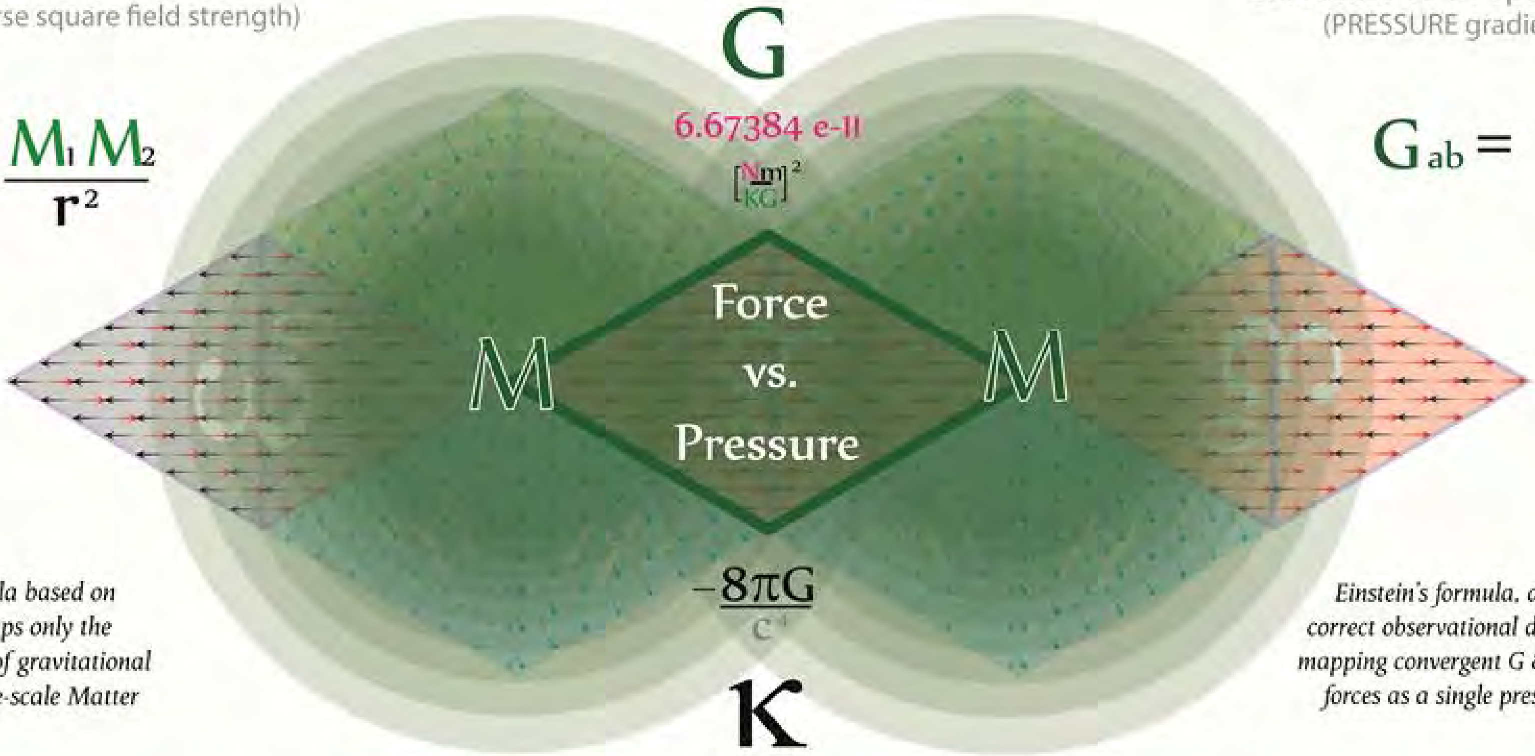
G
6.67384 e-11
$\left[\frac{Nm}{KG}\right]^2$
Force
vs.
Pressure
M
M
$-\frac{8\pi G}{c^4}$
κ

Newton's formula based on observations maps only the observed nett force of gravitational attraction on large-scale Matter

Einstein's formula, attempted to correct observational discrepancies by mapping convergent G & divergent EM forces as a single pressure gradient

Both Newton and Einstein failed to define and distingush between mass-energy geometries & Matter topologies

Copyright ABRAHAM [2008] - All rights reserved

Newton's G vs Einstein GR

Both Newton and Einstein developed their gravitational models
from the observed motions of celestial bodies without any distinction
between mass and Matter or the quantum forces at work

Gravitational Matter

Newton maps Matter's gravitational attraction
[excluding any perturbations caused by [K]EM fields]

All mass-energies

Einstein's GR maps Newton's G & SR effects
[mass-energy momenta [K]EM interactions]

Φ

F

$\nabla^2\Phi$

ρ

$g_{\mu\nu}$

$\Gamma^{\sigma}_{\mu\nu}$

$R_{\mu\nu}$

$T_{\mu\nu}$

Increasing accuracy in the measurements of the motions
of spacecraft interacting with planetary bodies [using EM waves]
has resulted in the need for refinement of both theories

Einstein's doubling of Newton's gravitational constant
was done to fit later observed perturbations in
Mercury's orbital mechanics

Newtonian G field

Any valid tests of the purely gravitational fields of Matter
must exclude KEM fields and not use EM masses to test it

$$[\text{Newton}] = \frac{4\pi GMm}{r^2}$$

$$\nabla^2\Phi = 4\pi G\rho$$

Poisson

$$[\text{Einstein}] = \frac{8\pi GT_{\mu\nu}}{c^4}$$

Pressure gradients and GR

Tensor mathematics is utilised in General relativity to model mass-energy geometries and energy pressure gradients as curved spacetime

$$F = -G\,\frac{M_1}{r^2}$$

Newton's Gravitational fields are modelled on the observed motions of large-scale Matter attraction

Newtonian G field

$$4\pi$$

G

Vector Force between Matter

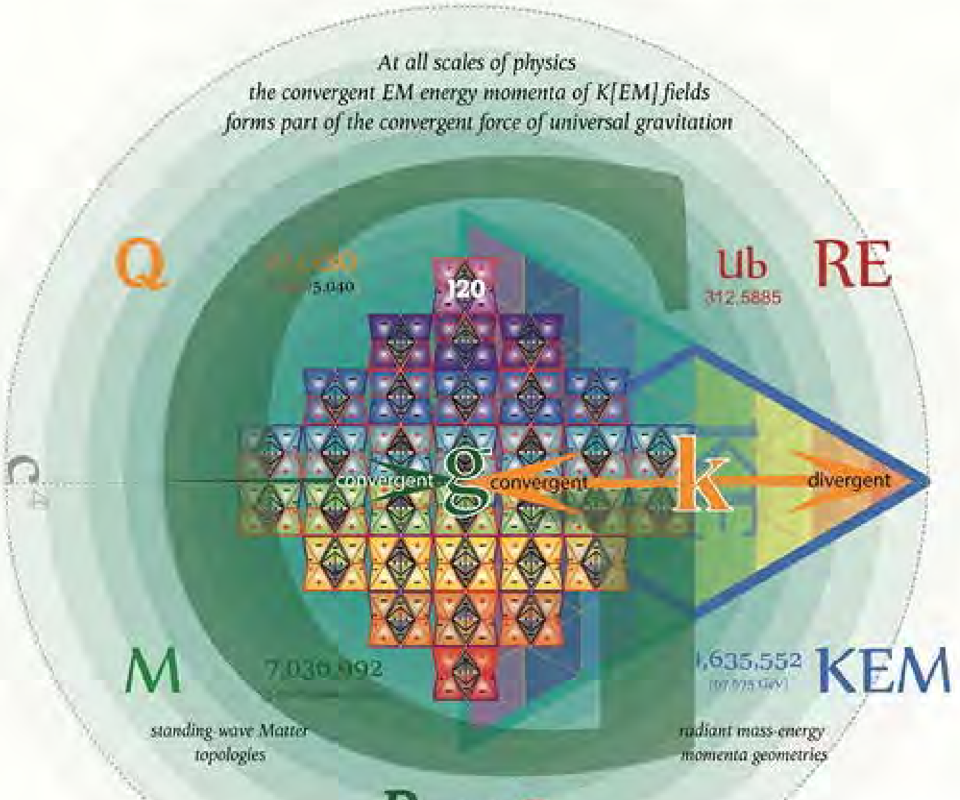
Q
5.040
120
Ub
312.5885
RE
convergent
g
convergent
k
divergent
M
7,030,992
,635,552
KEM
standing-wave Matter topologies
radiant mass-energy momenta geometries

At all scales of physics the convergent EM energy momenta of K[EM] fields forms part of the convergent force of universal gravitation

$$P \ll \rho$$

At the quantum level KEM mass-energy densities overwhelm the vacuum pressure gradients created by Matter topologies

$$n\pi\rho = \underset{\text{geometries}}{n\pi}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right] = T_{\mu\nu}$$

representive of ALL the mass-energy momenta of any spatial volume

$$G_{ab} = \kappa\, T_{\mu\nu}$$

Einstein's GR gradients corrects Newtonian gravity by modelling all forms of EM mass-Energy momenta

Einstein's G field

$$8\pi$$

K

Pressure gradient per unit Volume

Copyright ABRAHAM [2008] - All rights reserved

Stress energy-pressure gradiants

are the result of Tetryonic charge geometries & topologies

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Radiant mass-energies have planar energy geometries

nπ	**Energy quanta**
EVENπ	**photons**
ODDπ	**bosons**
4π	**tetryons**
12π	**quarks**
12π	**leptons**
24π	**mesons**
36π	**Baryons**

All Matter topologies are tetrahedral standing-wave mass-energy geometries

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

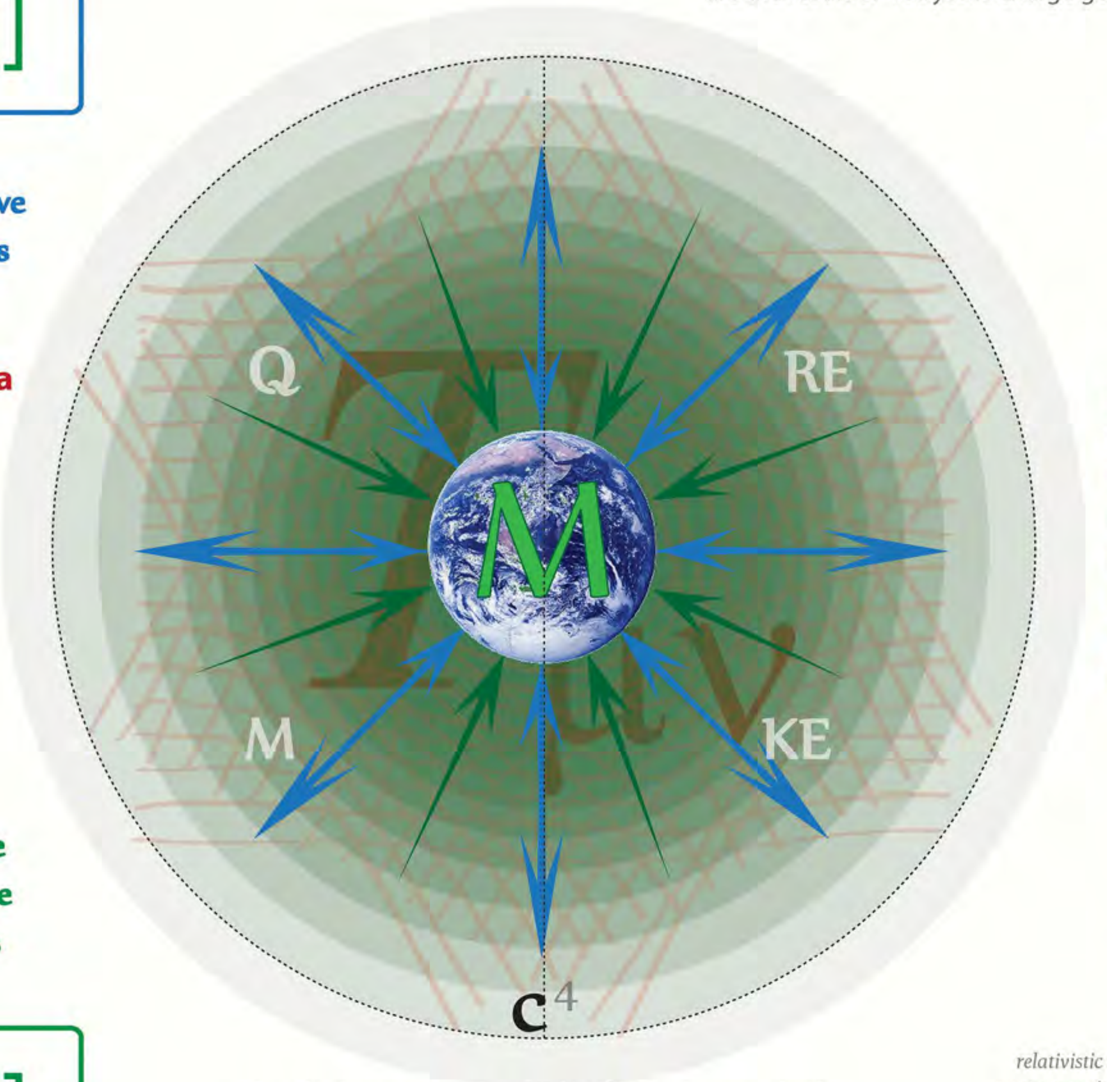

On the non-quantum scale all Matter topologies can be generalised and mathematically modelled as SPHERICAL 4π mass-energy topologies

radiated [K]EM fields have interactive energy momenta [convergent & divergent]

fields of Force

KEM

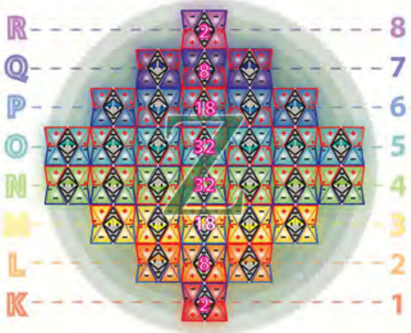

M

standing-wave topologies

displace vacuum energies creating convergent Gravity

relativistic mass-energies *standing-wave Matter* *radiant mass-energies*

RE = M + KEM

stress energies *topologies* *geometries*

Stress Energy tensor

In the Einstein field equations [EFE] of general relativity, the stress-energy tensor is the source of gravitational mass
just as Matter is the source of such a field in Newtonian gravity

$$T_{\mu\nu}$$

In Einstein's general relativity,
the symmetric stress-energy tensor
acts as the source of spacetime curvature

mass *ENERGY* *Matter*

m E M

geometries *momenta* *topologies*

Matter

Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

The Tetryonic unified field equation
models all mass-ENERGY geometries
& Matter topologies

A recognised problem in General Relativity is that
the stress tensor models all forms of mass-energy-Matter
in the form of a generalised stress energy pressure gradient
but fails to differentiate between either mass or Matter
and their resulting interactions within the fields

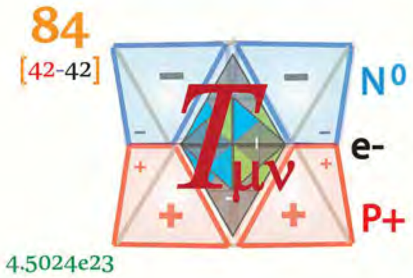

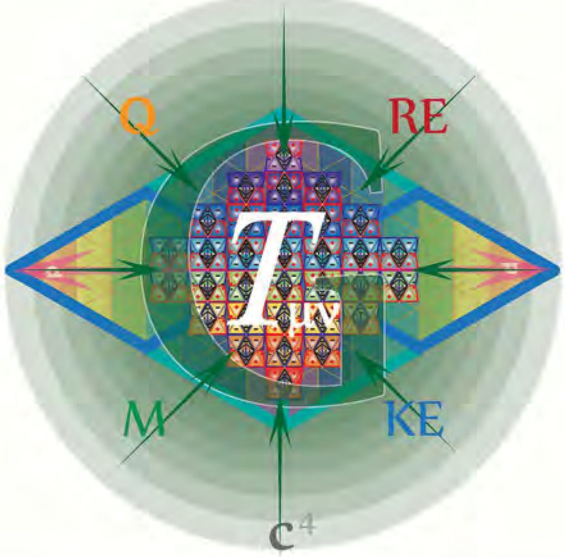

Planck quanta

$$n\pi\left[m\Omega v^2\right]$$

mass velocity

radiant mass-energies
create interactive
fields of force

$$T\pi[\rho]$$

standing-wave Matter
creates vacuum energy
pressure gradients

Copyright ABRAHAM [2008] - All rights reserved

Gravimetric topologies

Charges create radiant mass-energy geometries & standing-wave Matter topologies

kg

2D mass geometries

EM mass is not Matter

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Charged energy geometries are the basis for 2D mass

Photons have mass-energy momenta geometries but no Matter topology

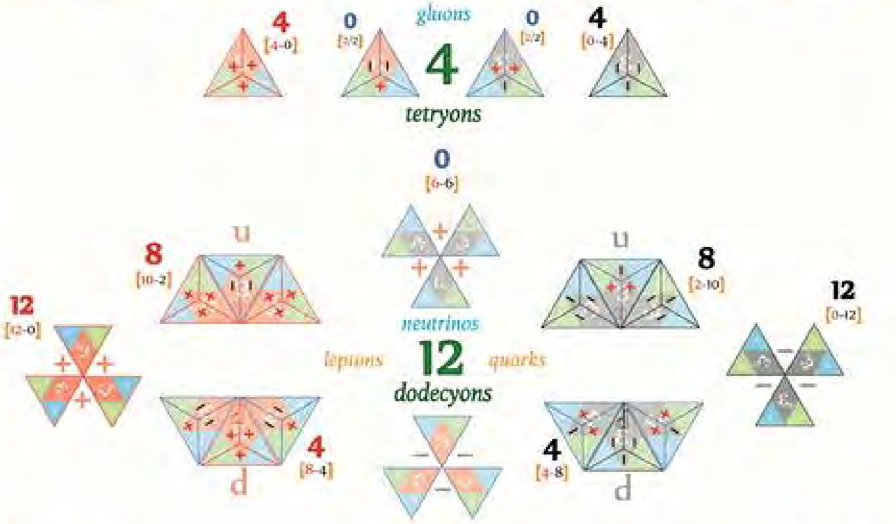
4
[4-0]
0
[2/2]
gluons
4
0
[2/2]
4
[0-4]
tetryons
0
[6-6]
u
8
[10-2]
12
[12-0]
neutrinos
leptons
12
quarks
dodecyons
u
8
[2-10]
12
[0-12]
4
[8-4]
d
4
[4-8]
d

2D mass-energy geometries form surface integral areas of charged Matter topologies

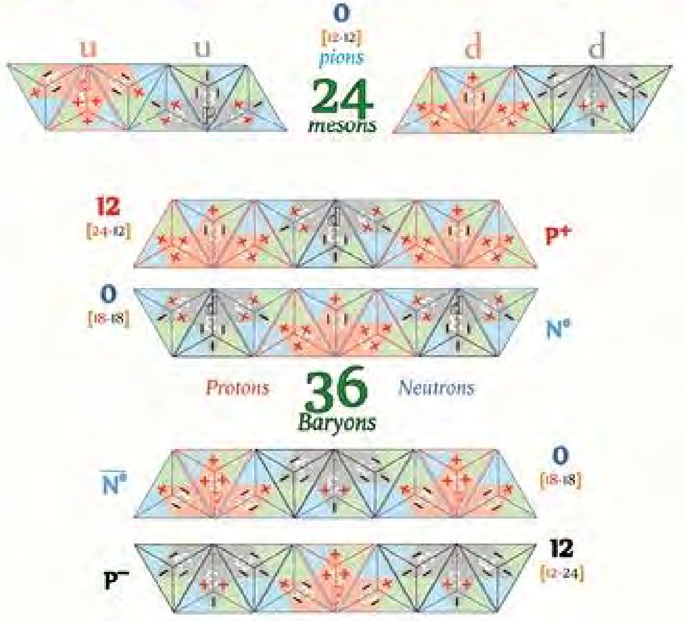
0
[12-12]
pions
u
u
d
d
24
mesons
12
[24-12]
P+
0
[18-18]
N°
Protons
36
Neutrons
Baryons
N̄°
0
[18-18]
P⁻
12
[12-24]

KG

3D Matter topologies

Matter has mass energies

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

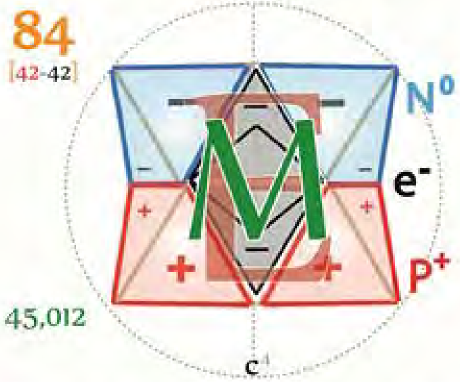
84
[42-42]
N⁰
e⁻
P⁺
M
E
45,012
c⁴

Charged mass-energy topologies are the basis for 3D Matter

Its topology forms a quantum Faraday cage [with a charged surface integral of mass-energies]

Interactive mass-energies vs. gravitational Matter

energy in all its forms seeks equilibrium

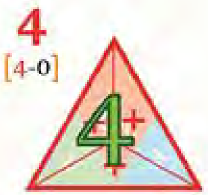

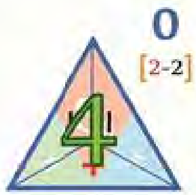

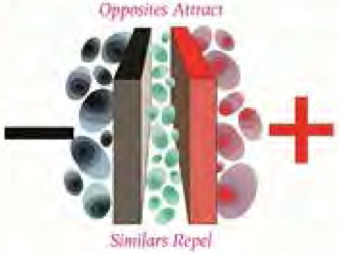
Opposites Attract
–
+
Similars Repel

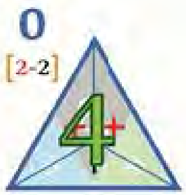

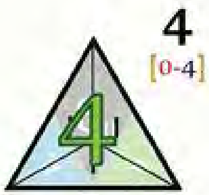

radiant mass-energy geometries

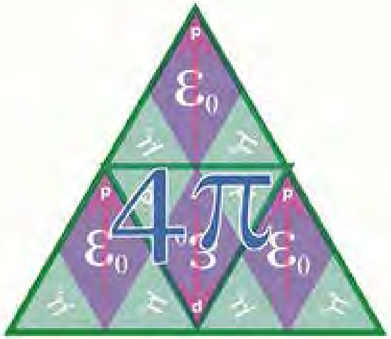
ε0
ε0
4π
ε0
P
P
P

$$\frac{\overset{\text{mass}}{n\pi}}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

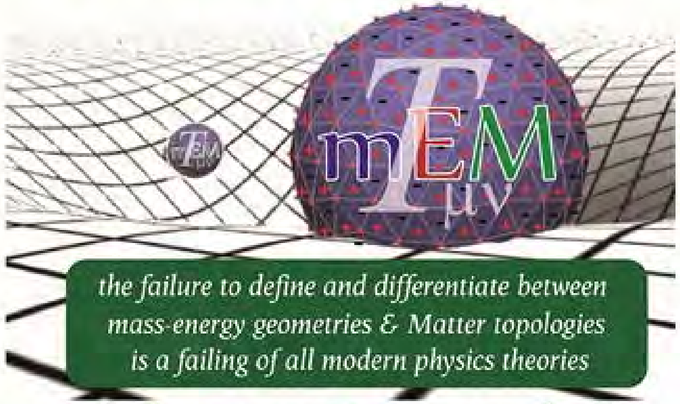

the failure to define and differentiate between mass-energy geometries & Matter topologies is a failing of all modern physics theories

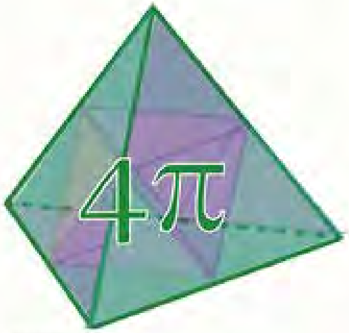

$$\frac{\overset{\text{Matter}}{T\pi}}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

standing-wave Matter topologies

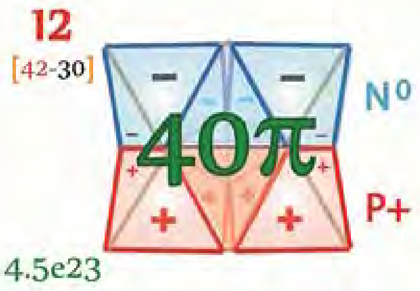
12
[42-30]
N0
40π
P+
4.5e23

Charged Planck quanta of mass-energy momenta that form topologies displace the vacuum energies surrounding them in turn creating a mass-energy-Matter pressure gradient

Compton frequency of rest Matter is what creates inertial mass, its Gravitational Matter results from its charged topology

The integral inductive EM mass-energy momenta in Matter fascia create the physical property of inertial mass

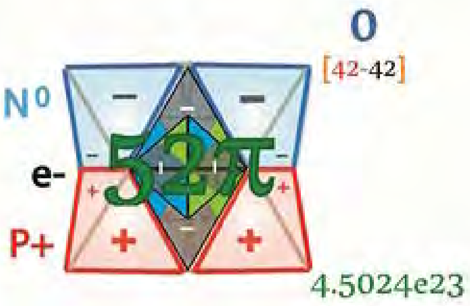
0
[42-42]
N0
e-
52π
P+
4.5024e23

$$F_z = \Delta P . \Delta A$$

Pressure is the force per unit area
applied in a direction perpendicular to the surface of an object

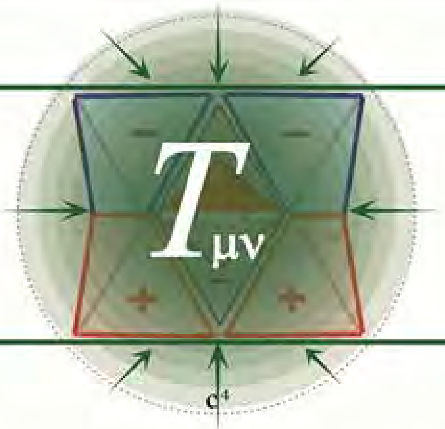

The theory of general relativity uses a stress-energy tensor to reflect to the pressure differentials created by Energy in ALL its forms (making no distinction between EM masses & Matter)

In Tetryonics the stress energy tensor is expanded, through geometric terms, to diferentiate between radiant mass-energy interactions and Matter topologies in radial spatial co-ordinate systems defined by the speed of light

Radiant EM masses

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Photons are often referred to as 'massless particles' are better defined as Matterless EM mass-energy momenta

ε_0
μ_0
μ_0
n
μ_0
μ_0
ε_0

$$\overset{\text{Photons}}{\frac{2\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

interactive mass-energy momenta geometries

Energy Pressure gradients

The vacuum energy pressure gradient created by the charged topology of Matter is distinct from its intrinsic & radiative mass-energy geometries

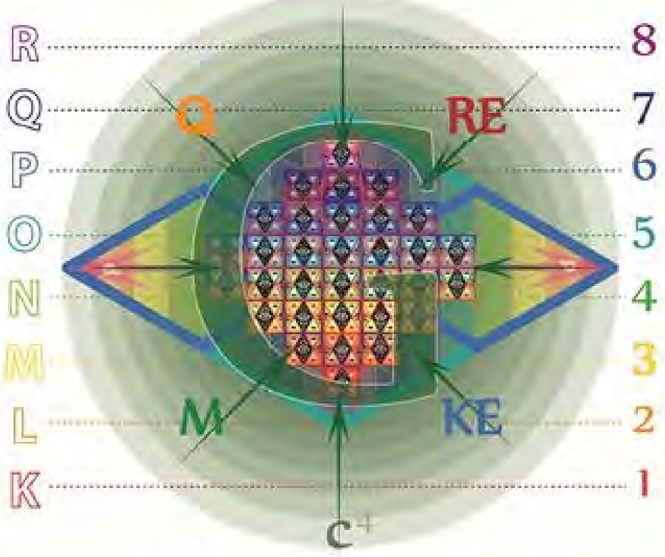

Any theory of gravitation that attempts to model it on any scale must account for energy in all its forms

Standing-wave Matter

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Tetryons, the quantum building blocks of Matter have EM mass-energy topologies that create a vaccum pressure gradient

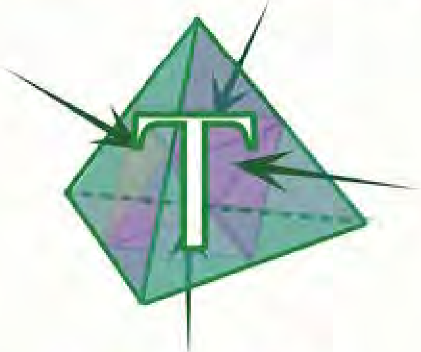

$$\overset{\text{Tetryons}}{\frac{4\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

gravitational Matter topologies

Tetryonics 68.09 - Energy Pressure gradients

Convergent mass-energies & Gravitational Matter

m

$$R_{\mu\nu} - \tfrac{1}{2}R g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Spacetime curvature

Stress-Energy Tensor

M

The concept of mass in general relativity (GR) is more complex than the concept of mass in special relativity. In fact, general relativity does not offer a single definition for the term mass, but offers several different definitions which are applicable under different circumstances.

Komar mass | **Bondi mass** | **ADM mass**

Under some circumstances, the mass or Matter components of a system may not even be defined

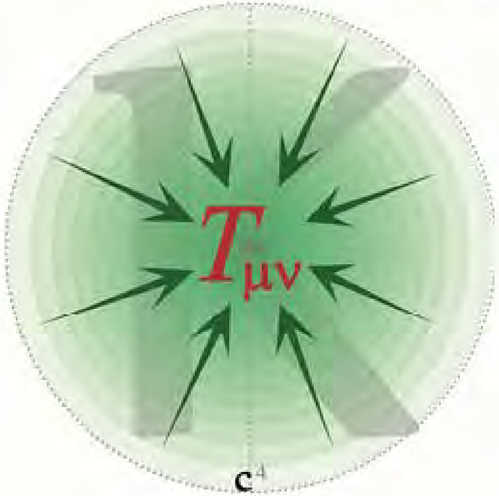

Einstein Field equation

Spacetime curvature

total mass-energy-Matter

$$R_{\mu\nu} - \tfrac{1}{2}R g_{\mu\nu} = \underset{\text{constant}}{\overset{\text{gravitation}}{K}} \frac{n\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

per spatial volume

Tetryonic geometries & topologies

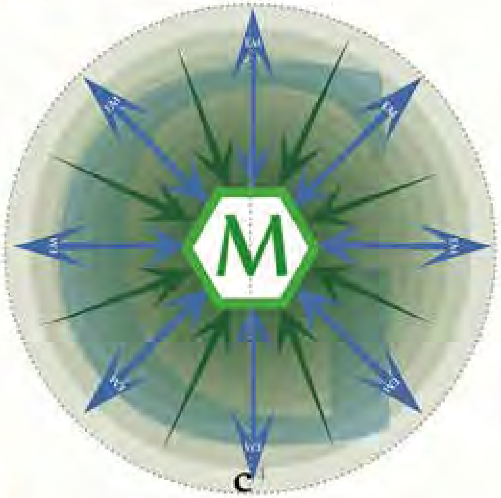

In contrast Tetryonic field equations explicitly differentiate all EM mass-energies and Matter via their charged geometries along with their contributions to the total energy density of any spatial region

GR models the convergent force of gravitation as a convergent energy density pressure gradient created by mass-energies

$$\underset{\text{mass-Matter}}{\overset{\text{gravitational}}{M}} + \underset{\text{mass-energies}}{\overset{\text{interactive}}{kEM}} = \underset{\text{energies}}{\overset{\text{stress}}{U}}$$

Tetryonics models gravitation as the result of mass-energy interactions AND Matter topology displacement of vacuum energies

Interactive EM force fields are subject to Lorentz velocity corrections

Faraday

$$\mathcal{E} = -\frac{d\Phi_B}{dt},$$

$$|\mathcal{E}| = N\left|\frac{d\Phi_B}{dt}\right|$$

$$\oint \vec{E}\cdot d\vec{s} = -\frac{d\Phi_B}{dt}$$

$$\oint \vec{E}\cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

$$\oint \vec{B}\cdot d\vec{A} = 0$$

$$\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}$$

Maxwell

Special Relativity was developed from the foundation of Lorentz corrections to EM fields

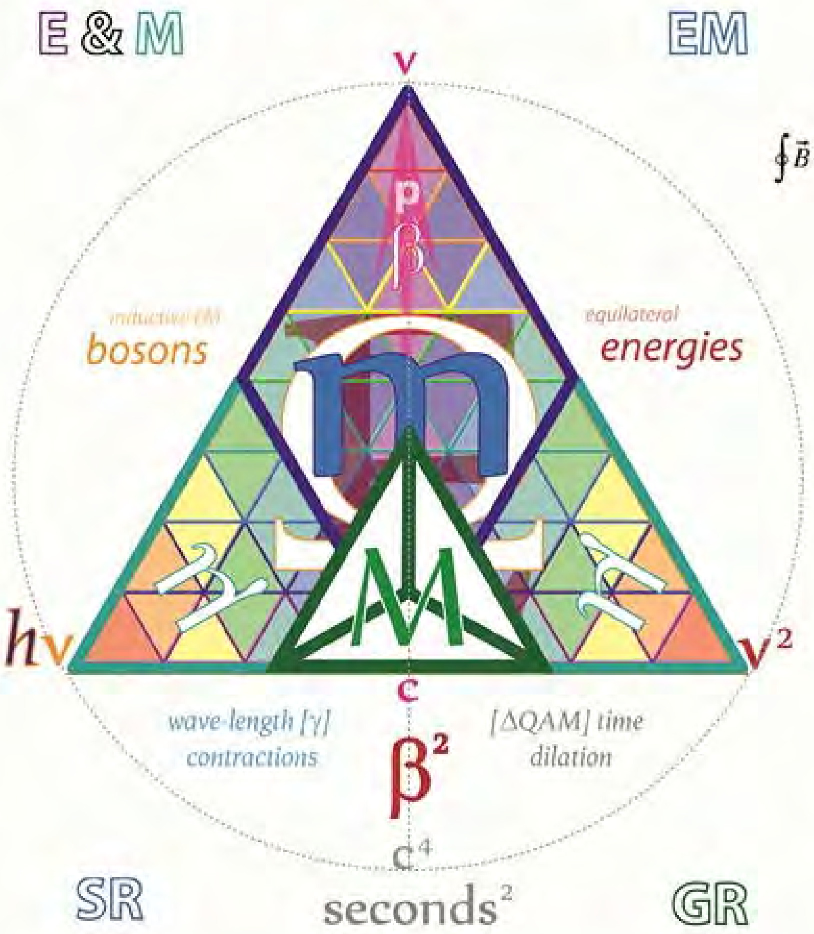
E & M
EM
v
p
β
inductive EM
bosons
equilateral
energies
m
M
γ
γ
hv
v²
c
wave-length [γ] contractions
[ΔQAM] time dilation
β²
c⁴
SR
seconds²
GR

General Relativity was developed from Special Relativity and incorporates EM field interactions

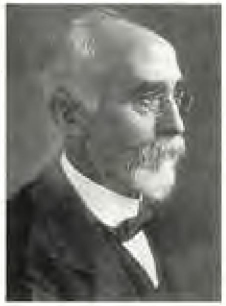

Lorentz

$$L = L'\sqrt{1-\frac{v^2}{c^2}}$$

$$t' = \frac{t}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}$$

$$E = \frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}}$$

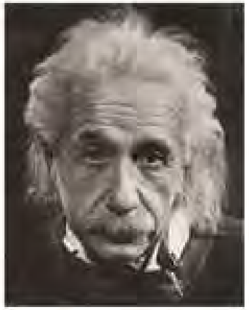

Einstein

Standing-wave Matter topologies are Lorentz invariant

the failure to differentiate between mass & Matter has lead to major misconceptions in physics

Copyright ABRAHAM [2008] - All rights reserved

Tensor Rankings

Rank 0

scalars

A tensor may consist of a single number, in which case it is referred to as a tensor of order zero, or simply a scalar

An example of a scalar would be the EM mass-energy of a particle or field

A second example of a scalar field would be the value of the gravitational potential energy as a function of position

Rank 1

The next most complicated tensor is the tensor of order one, otherwise known as a vector.

vectors

A vector can be defined as a direction between two points, its strength may vary continuously from point-to-point, thereby defining a vector field

Next above a vector are tensors of order 2, which are often referred to as matrices.

Rank 2

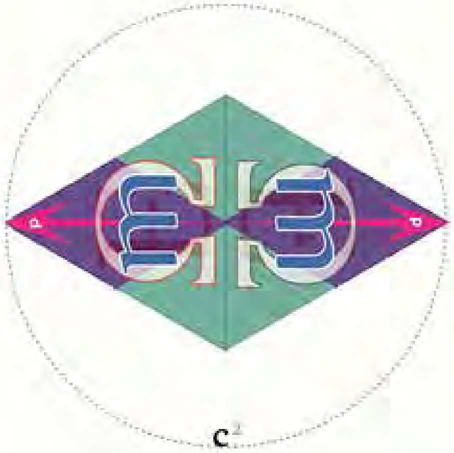

mass-energy momenta

A tensor of rank 2 is defined as a system that has a magnitude and two associated vector directions within the fields that it describes

Rank 3

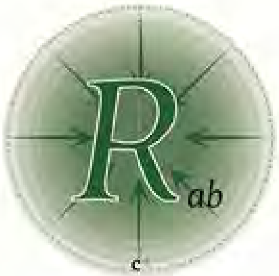

curved spacetime

Rank 4 Tensors as used in Einstein's General Theory of Relativity, describe the curvature of spacetime [the so-called Riemann curvature tensor].

Tetryonic Matrices

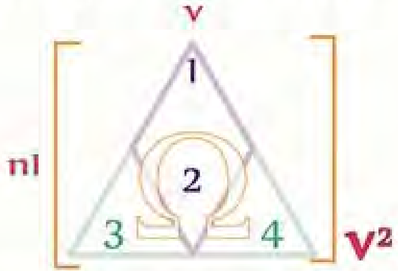

EM fields can be represented by Tetryonic Matrices [Tensors] identical to current square Tensor Matrices with the only difference being the change in the geometry of the matrix to better reflect the true geometry of EM fields

This better facilitates the easy recognition of Electric & Magnetic scalar fields and their associated energy-momenta vectors

Rank 4

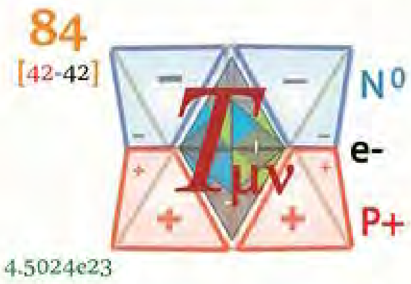

mass-Matter topology

the total mass-energies of Matter & fields thatcontribute to creating gravity.

Tetryonics 69.02 - Tensor rankings

ElectroMagnetic field tensor

The electromagnetic tensor or electromagnetic field tensor (sometimes called the field strength tensor, Faraday tensor or Maxwell bivector) is a mathematical object that describes the electromagnetic field of a physical system in Maxwell's theory of electromagnetism

$$2\pi_{\text{Photons}}\left[\frac{1}{c^2}\cdot\left[m\overset{\text{Energy}}{\Omega}v^2\right]_{\text{momenta}}\right]$$

Photons

$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Visualising EM waves and photons in terms of their Tetryonic geometry the scalar relationships between Electric and Magnetic fields becomes readily apparent.

ε_0
γ
μ_0

$$F^{\mu\nu} = -F^{\nu\mu}$$

photons are their own anti-particles

ε_0
γ
μ_0

A EM field tensor has six degrees of freedom, representing the three components of each of the electric and magnetic fields

$$c = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

Speed of Light

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

EM Fields

EM wave factors such as Lorentz's velocity corrections arise naturally from the geometry providing a clear and concise illustration of the relationships between EM mass-energies, Charge, Electric & Magnetic fields and the speed of Light

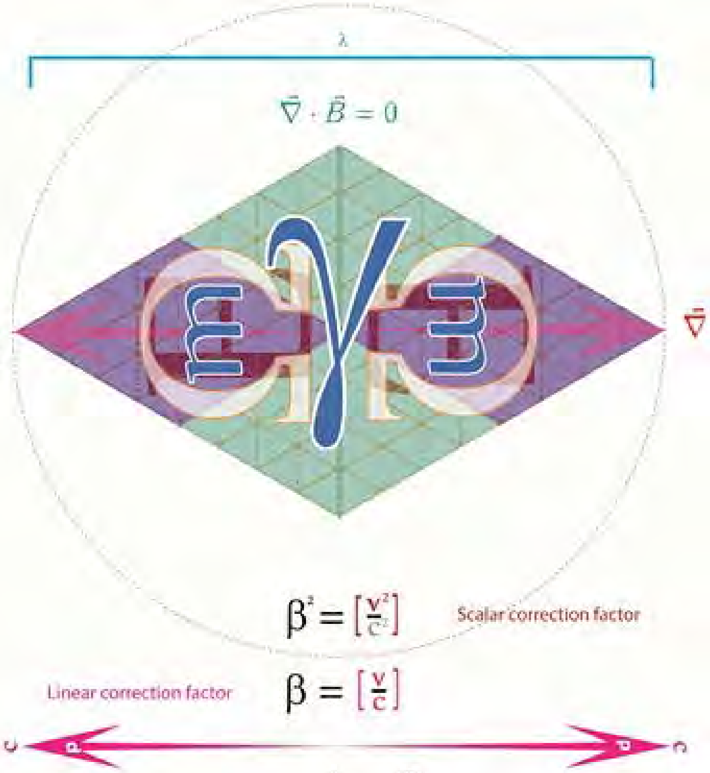

$$\vec{\nabla}\cdot\vec{B} = 0$$

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho}{\epsilon_0}$$

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho}{\epsilon_0}$$

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

$$\frac{\partial\vec{B}}{\partial t} + \vec{\nabla}\times\vec{E} = 0$$

the change of the magnetic field with respect to time plus the curl of the electric field is equal to zero

$$\vec{\nabla}\times\vec{B} - \frac{1}{c^2}\frac{\partial\vec{E}}{\partial t} = \mu_0\vec{J}$$

The curl of the magnetic field vector minus the change of the electric field with respect to time, is equal to the field's current density

Wavelength contraction

$$L = L'\sqrt{1-\frac{v^2}{c^2}}$$

$$p = \frac{h}{\lambda} = \frac{E}{c}$$

Linear Momentum

$$\det(F) = \frac{1}{c^2}\left(\vec{B}\cdot\vec{E}\right)^2$$

the electromagnetic tensor may also be written in terms of the 4-vector potential

speed of Light

$$\underset{\text{velocity of propagation}}{c} = \left[\underset{\text{Frequency}}{\overset{f}{\left[\frac{c^2}{\Omega}\right]}}\cdot\underset{\text{Wavelength}}{\overset{\lambda}{\left[\frac{\Omega}{c}\right]}}\right]$$

Copyright ABRAHAM [2008] - All rights reserved

3D space time

$$[\Delta x, \Delta y, \Delta z] \quad \Delta\tau$$

In Tetryonics, the relativistic Four vector reflects the changing quantised angular momenta [time] of EM mass-energies & Matter in 3D space

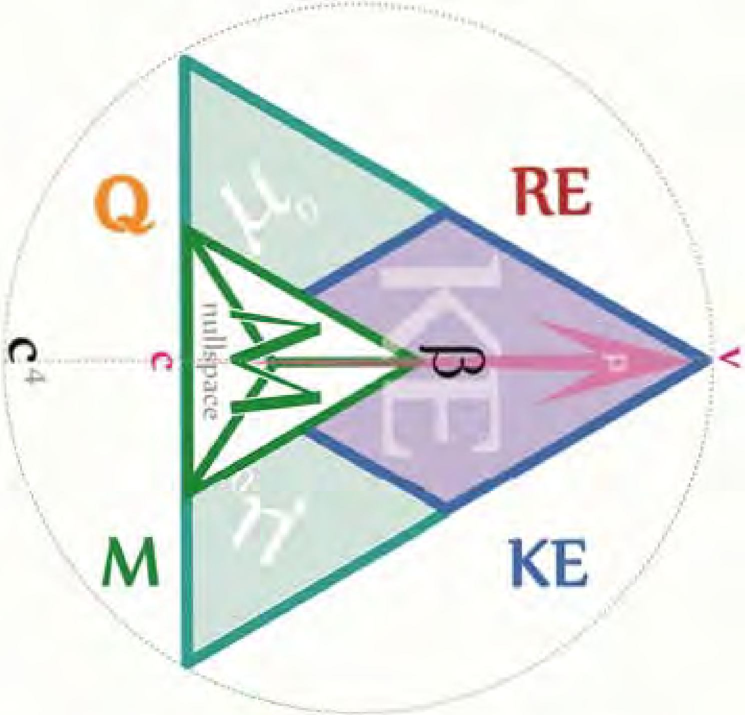

$$M_0 + KE = E$$

Standing-wave four-vector

Lorentz corrected mass-energies

Stress energy momenta tensor

Relativistic Four-vectors

Charged Matter interacts at-a-distance through super-positioned energy momenta fields from which our force constants are derived

The electromagnetic four-potential is a potential from which electromagnetic field can be derived. It combines both the electric scalar potential and the magnetic vector potential into a single space-time four-vector

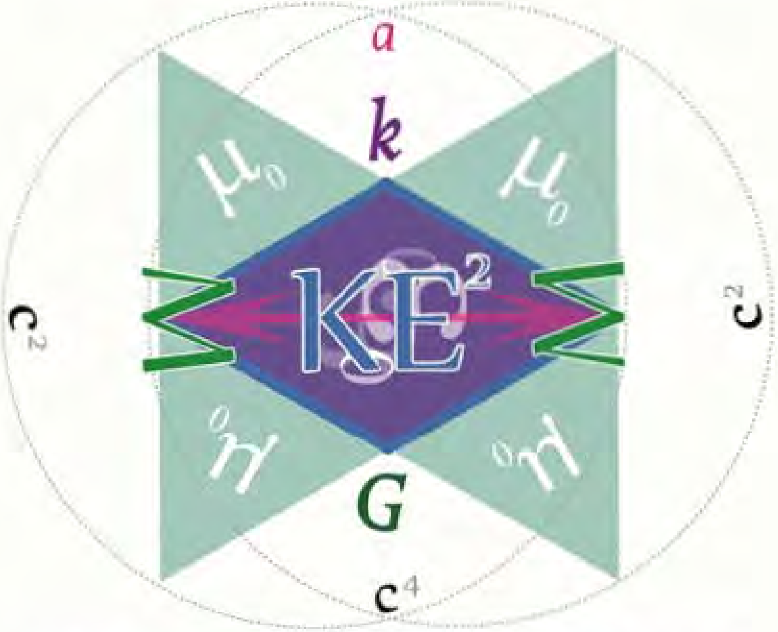

Inverse sqaured Gravity can be mathematically modeled using the same geometry explaining the similar formulations for charge interaction & gravitation

$$m^2v^4 = KE^2 = p^2v^2$$

EM mass

radiant energies

momenta

Four-vectors are reflective of the total EM mass-energy momenta of superpositioned [K]EM fields where the KE is the interactive force component

Special and General relativity

General relativity was developed by Einstein in an attempt to describe the acceleration forces experienced by gravitational Matter (on the foundation of special relativity) which in turn was an extension of Lorentz corrections of EM waves developed to account for the fixed speed of light and established Newtonian mechanics

SR

Special Relativity

(describes inertial frames of reference
in flat Euclidean spacetime)

mass-energy

mass

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta

mass velocity

geometries

GR

General Relativity

(describes accelerating frames of reference
in curved Riemannian spacetime)

Matter

Matter

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta

mass velocity

topologies

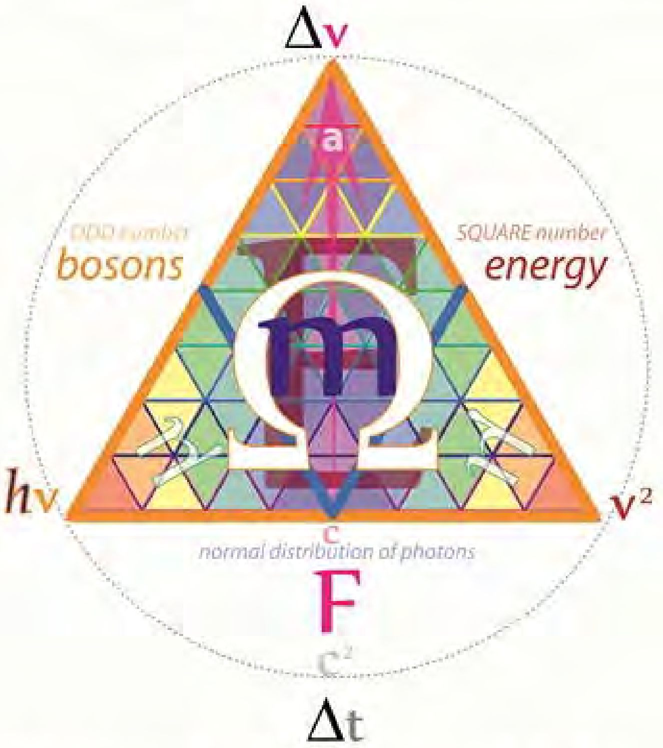
Δv
a
ODD number
bosons
SQUARE number
energy
m
Ω
hv
v²
c
normal distribution of photons
F
c²
Δt

2D

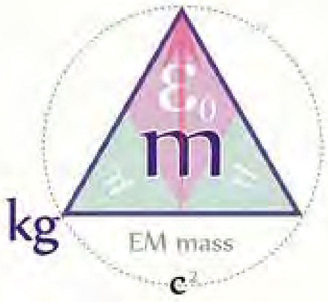
ε₀
m
kg
EM mass
c²

3D

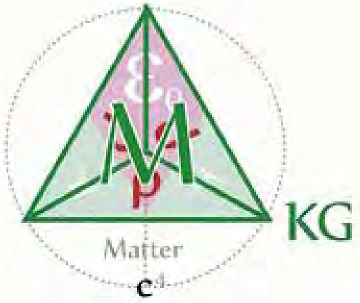
ε₀
M
ρ
KG
Matter
c⁴

all physical theories to date have failed to explain the mechanics of gravitation at all scales of energy and distance

$$\frac{\rho}{m^2}$$

6.629432351 e-34 J

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta

mass velocity

7.376238376 e-51 kg

$$\frac{\rho}{m^3}$$

The Weak Equivalence Principle

$$m_g . a = m_i . g$$

Newton

The uniqueness of free fall states than any two test bodies must fall with the same acceleration in a given external gravitational field

$$m_g = m_i \sum_A \eta^A \frac{E^A}{c^2}$$

Einstein

The uniqueness of free fall trajectories allows one to regard spacetime as filled being with curved energy gradients

$$F = ma$$

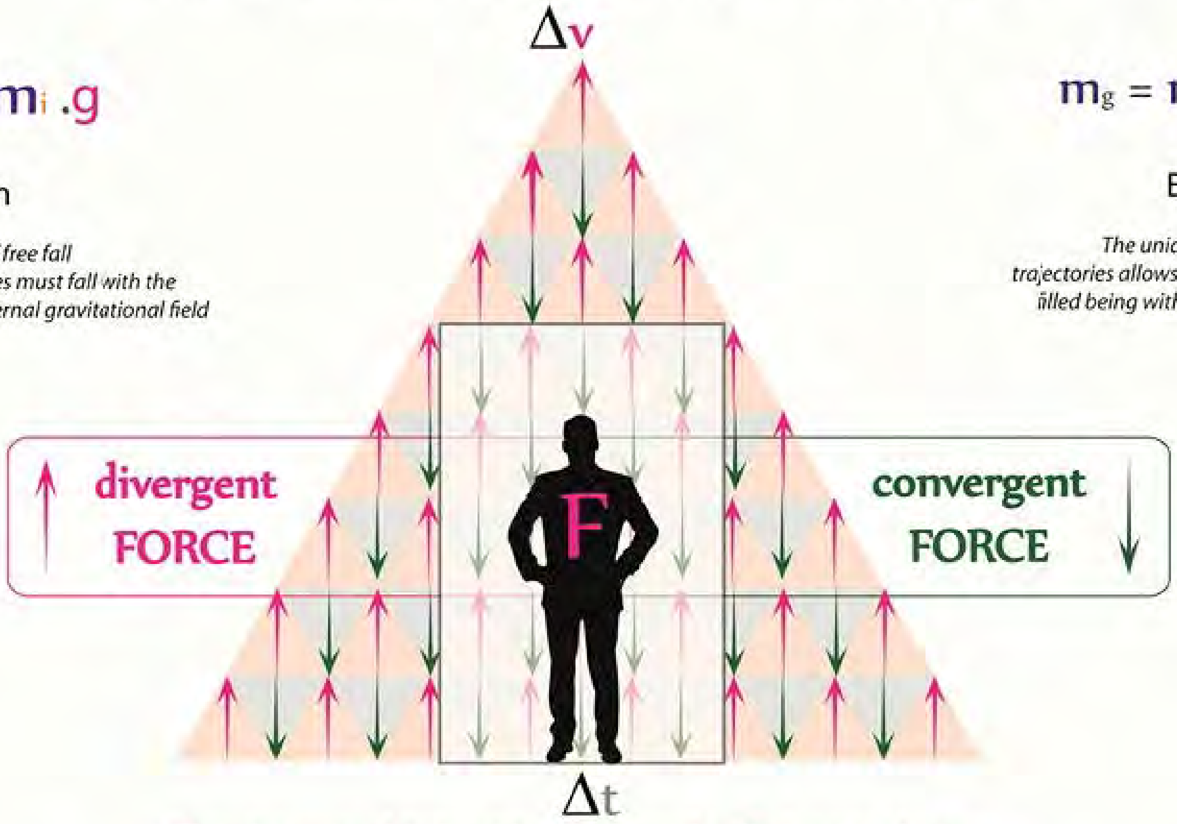

$$F = mg$$

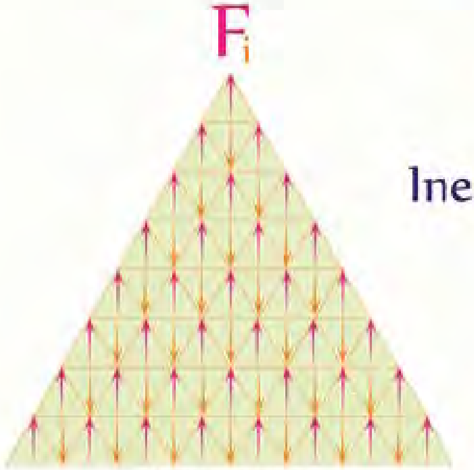

acceleration vs inertial mass

'There is no way of distinguishing between the effects on an observer of a uniform gravitational field and that of constant acceleration'

Inertial and Gravitational masses are equivalent

Of particular note is the fact that the relativistic formulation of the weak equivalence principle hints at the fact that mass-energies in all their forms must be accounted for and correctly modeled for an accurate picture of the mechanics of gravitiation on all-scales to be developed

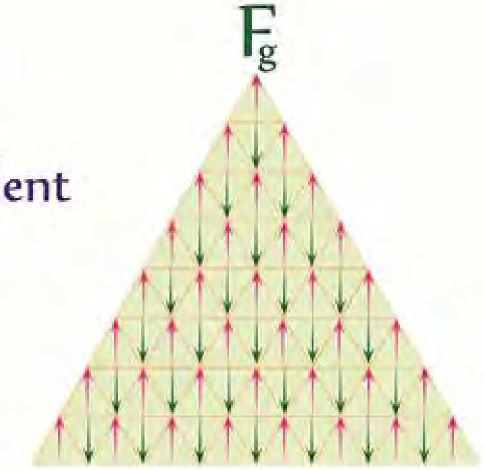

acceleration vs gravitational mass

Einstein's Gravitational Constant

Newton gave a law for the behavior of objects that experienced gravitational forces:
$F = ma$, where $F = -m\nabla\phi$ for a given gravitational field ϕ

And along with Gauss & Poisson also gave a law determining how φ is generated:
$\nabla^2\phi = 4\pi G\rho$.

Einstein subsiquently re-expressed Newtonian Gravitation as a curvation of spacetime brought about by the presence of all mass-Energies

$$G_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

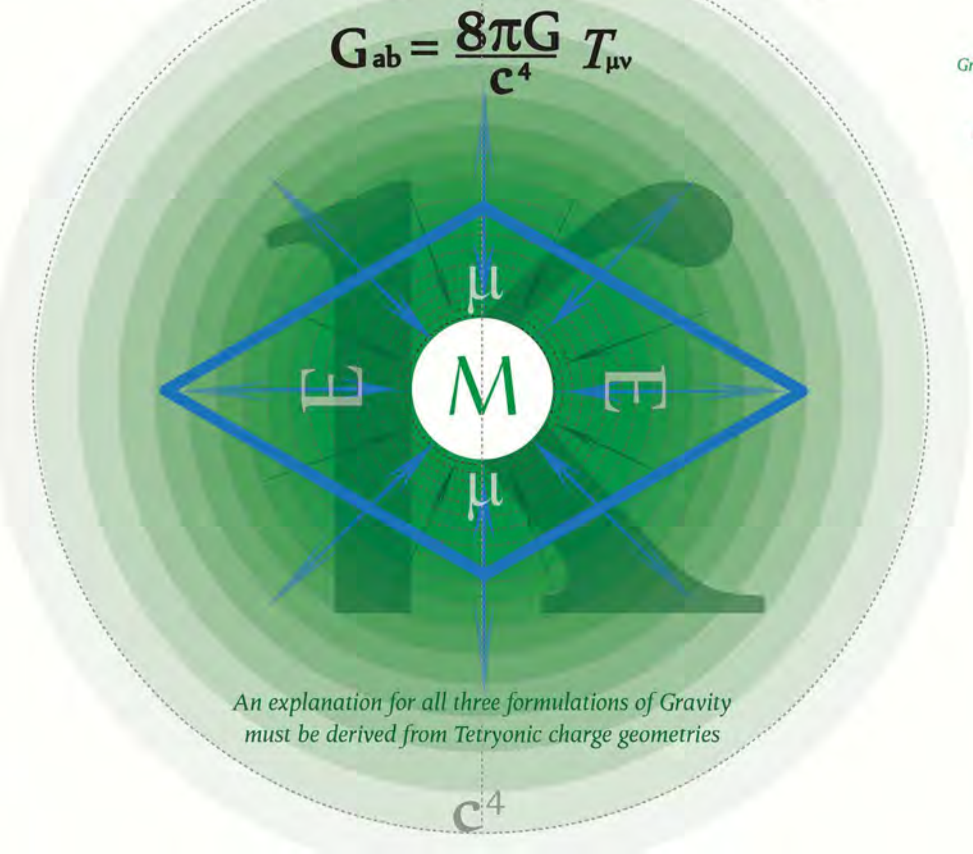

An explanation for all three formulations of Gravity must be derived from Tetryonic charge geometries

$$G_{ab} = \kappa T_{\mu\nu}$$

General Relativity

spactime curvature

Gravitational constant

stress energies

$$G_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

speed of light

Einstein viewed gravitation as the result of a curvature of spacetime created by a mass-energy-pressure gradiant

EM

2D euclidean mass-energy-momenta

Rank 2 tensors

convergent & divergent forces

$$m = \frac{\rho}{c^2}$$

$$n\pi[mEM]$$

Tetryonic charged geometries

Gravitational Matter topologies

Interactive mass-energy geometries

$$T\pi\left[\frac{\rho}{c^4}\right] + n\pi\left[\frac{\rho}{c^2}\right]^2$$

3D spatial region

2D spatial fields

Tetryonics differentiates between Matter & interactive mass-energies via their charged geometries and topologies

M

3D mass-Matter topologies

Rank 4 tensors

convergent pressure gradient

$$M = \frac{\rho}{c^4}$$

By modeling all mass-energy geometries and Matter topologies as a stress energy tensor General Relativity fails to differentiate between the interactive Forces of mass-energy momenta and gravitational Matter

mass-energy-Matter spatial interactions

**No two Fermions [3D Matter topologies]
can occupy the same physical space
at the same time - Pauli exclusion**

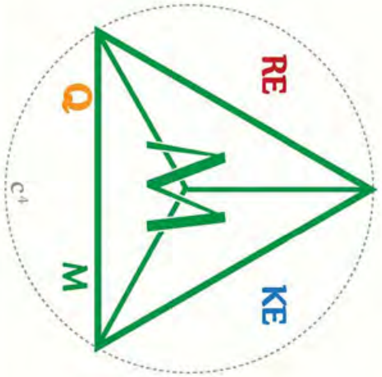

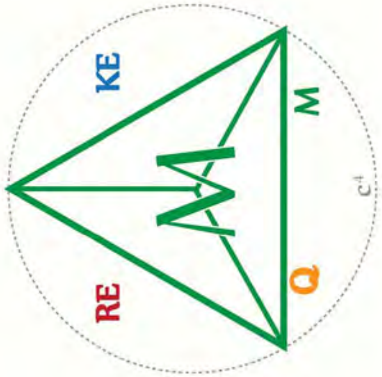

c^4

is a volumetric measure
of the amount of space
taken up by Matter

Fermions · Planck quanta · mass · velocity

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

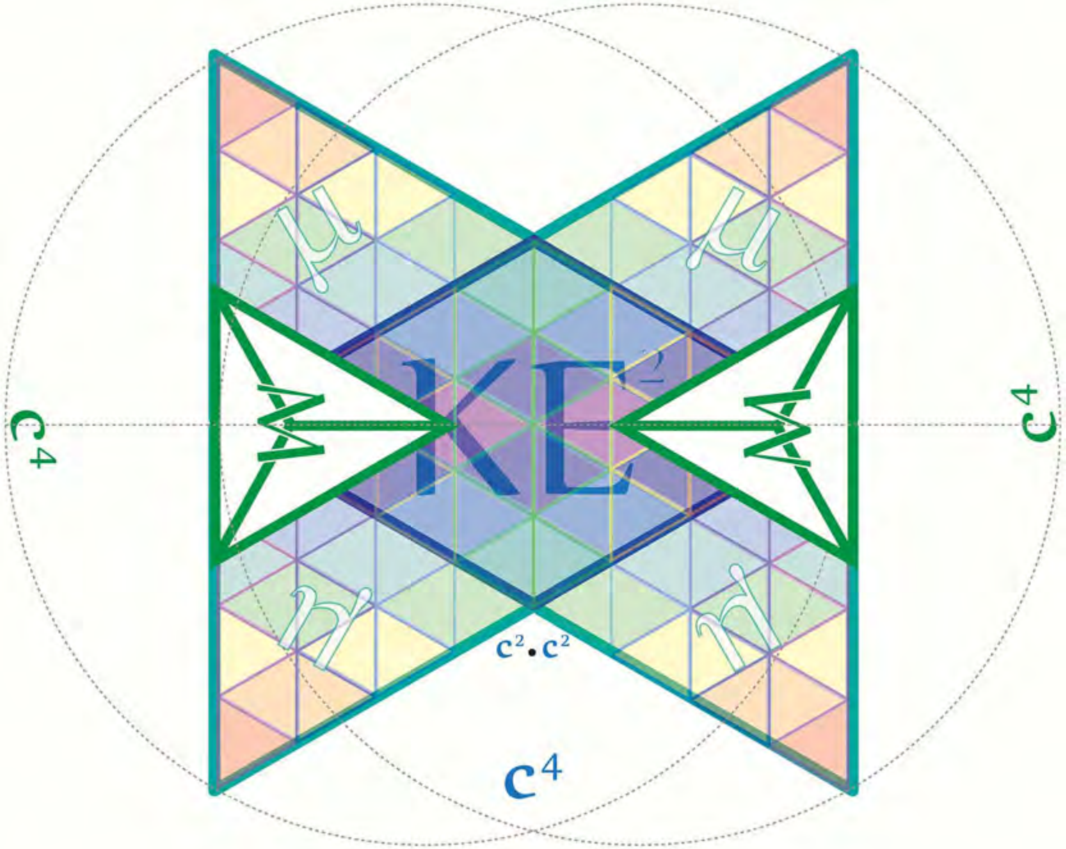
μ
μ
μ
μ
KE²
M
M
c⁴
c⁴
c²•c²
c⁴

c^4

is also a spatial measure
of two superpositioned
interactive KEM fields

Force interaction · Planck quanta · mass · velocity

$$\frac{2\pi}{c^2}\cdot\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

**All separated particles
interact with each other via 2D fields of Force
(superpositioned [K]EM fields)**

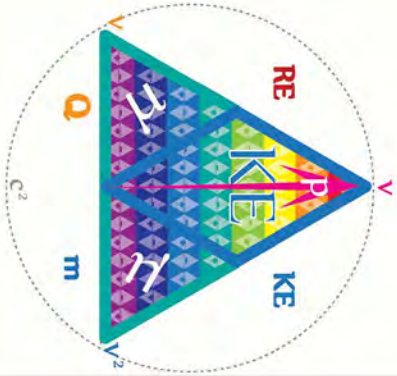
v
Ω
RE
μ
KE
p
v
c²
m
μ
KE
v²

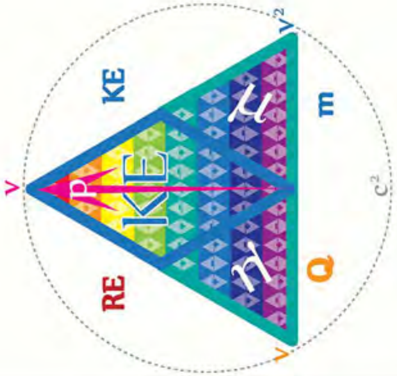
v²
KE
m
μ
v
p
KE
c²
μ
RE
Ω
v

Copyright ABRAHAM [2008] - All rights reserved

$$\frac{F}{m^2}$$

c^2

SR

Special relativity applies
Lorentz velocity corrections to
radiant mass-energy momenta

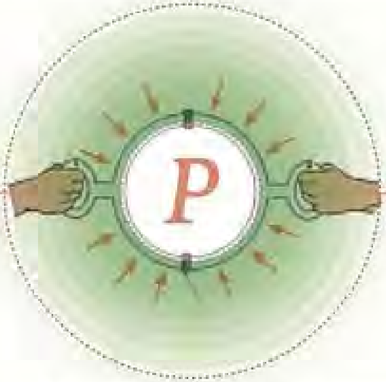

GR

General relativity applies SR
to total energy momenta
density-pressure gradients

$$\frac{E}{m^3}$$

c^3

Otto von Guericke, (1602-1686), a German physicist, performed a famous experiment: the "Magdeburg Hemispheres".

When two halves of a large, hollow metal ball had all the air removed from the volume they created when joined, two teams of horses couldn't pull them apart, because the pressure of the Earth's atmosphere (external to the volume created) exerted a very large force on the ball as it sought to equalise the pressure gradient created.

Nullspace energy pressure gradients

mass-energy geometries

$$P = \frac{\text{Force}}{\text{Area}} = \frac{F}{A} \Rightarrow \frac{F.d}{A.d} \Leftarrow \frac{W}{V} = \frac{\text{Energy}}{\text{Volume}} = U$$

mass-Matter topologies

The scalar mass-energies of all Matter topologies are all contained in their charged fascia

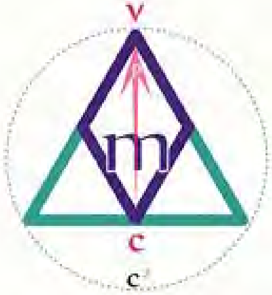

$$n\pi\left[\frac{\rho}{c^2}\right]$$

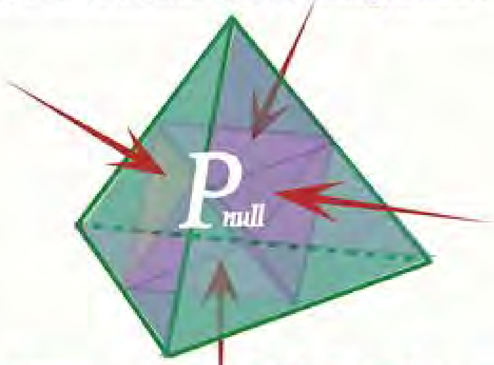

$$T\pi\left[\frac{\rho}{c^4}\right]$$

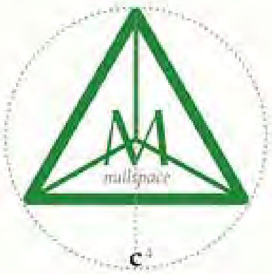

A Faraday cage is a metallic enclosure that prevents the entry or escape of an electromagnetic field (EM field)

All Matter topologies are ideal Faraday cages, quantum standing-waves of EM mass-energies that in turn create regions of nullspace

An ideal quantum Faraday cage consists of an unbroken, perfectly conducting shell. (a Matter topology of scalar mass-energies)

Total stress energies & Matter

The stress energy tensor can be regarded only as a provisional means of representing all of the interactive mass-energies & gravitational Matter in any physical system.

$T_{\mu\nu}$

GR's Stress Energy Tensor does NOT differentiate between EM mass & Matter

Einstein

It is only in present circumstances that we find we have insufficient knowledge of the electromagnetic field of concentrated charges that compels us, provisionally, to leave, undetermined in present theory, the true form of this tensor...

The right hand side the EFE is a formal condensation of all things whose comprehension in the sense of a field theory is still problematic.

Not for a moment... did I doubt that this formulation was merely a makeshift in order to give the general principle of relativity a preliminary closed form expression.

For it was essentially no more than a theory of the gravitational field, which was isolated somewhat artificially from a total field of as yet unknown structure.

Abraham

mass

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

changing quantised angular momenta creates charge~Time

3D MATTER topologies displace vacuum energies creating Gravity

Vacuum energies are formed from weak radiated mass-energies

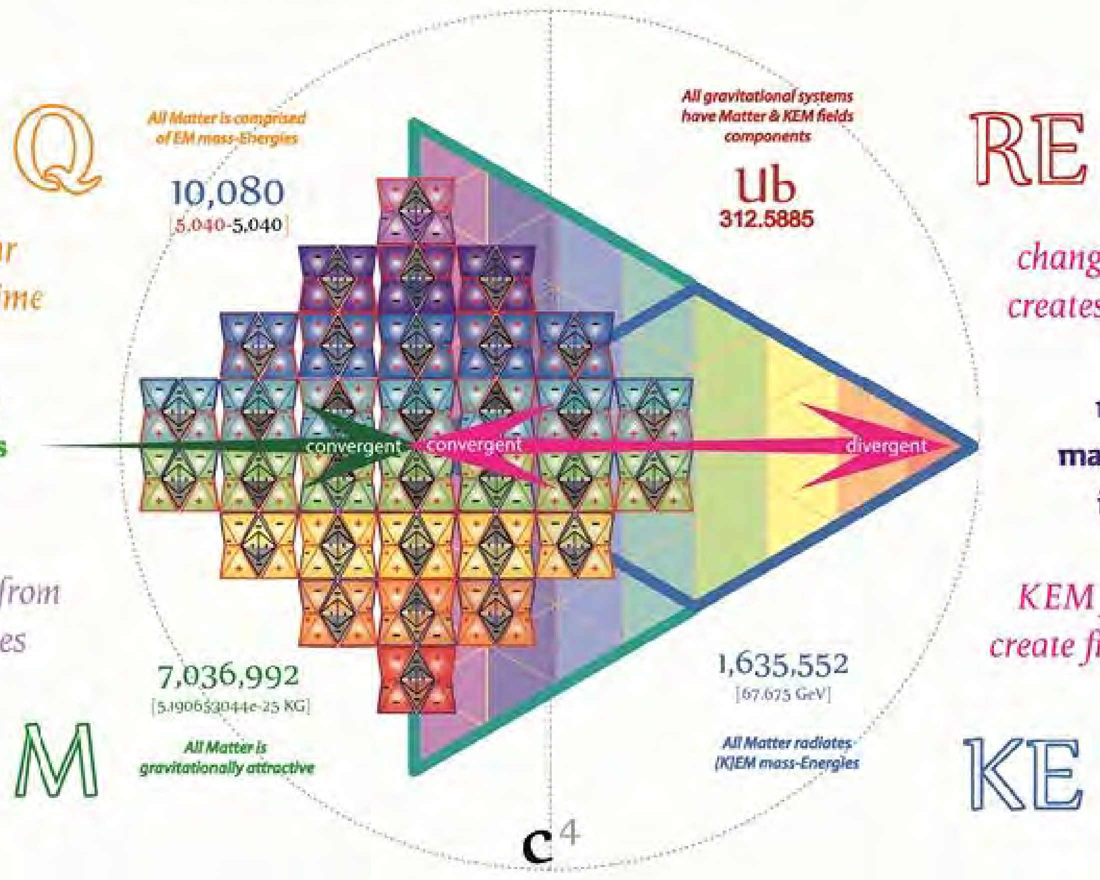
Q
All Matter is comprised of EM mass-Energies
10,080
[5,040-5,040]
All gravitational systems have Matter & KEM fields components
Ub
312.5885
RE
convergent
convergent
divergent
7,036,992
[5.190653044e-25 KG]
All Matter is gravitationally attractive
M
1,635,552
[67.675 GeV]
All Matter radiates (K)EM mass-Energies
KE
c^4

changing linear momentum creates forces of Acceleration

radiant 2D planar mass-ENERGIES create interactive forces

KEM field energy momenta create fields of interactive Force

Tetryonics 69.10 - Total stress energies & Matter

Copyright ABRAHAM [2008] - All rights reserved

Einstein's Field Equation

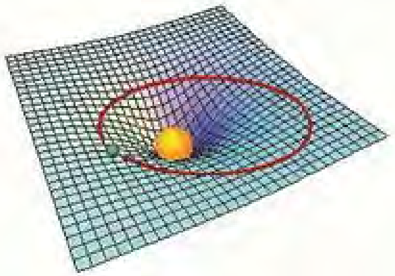

"Spacetime tells Matter how to move: Matter tells spacetime how to curve"
John Archibald Wheeler

$$R_{ab} - \frac{1}{2}R g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

The solutions of the Einstein field equations are metrics of spacetime

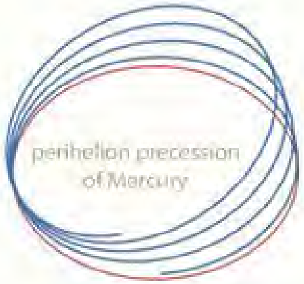

General Relativity was developed to account for the observed perihilion precession of Mercury's orbit around the SUN
[only noticable through detailed observations after Newton developed his law of universal gravitation]

Newtonian G field

$$4\pi$$

the scalar curvature

the cosmological constant

Newton's gravitational constant

Ricci curvature tensor

$$R_{ab} - \frac{1}{2}g_{ab}R + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

the stress-energy tensor

the metric tensor

the spatial region of mass-energy-Matter being measured as determined by the speed of Energy in vacuum

The EFE is a tensor equation relating a set of symmetric 4 x 4 tensors. Each tensor has 10 independent components

Einstein's G field

$$8\pi$$

The EFE (based on SR principles) reduces to Newton's law of gravitation where the radiant mass-energy fields are weak and velocities are much less than the speed of light
[highlighting the inability of GR to distingush between radiant mass-energy geometries & the standing-wave Matter topologies comprising the total energies within the stress Energy tensor]

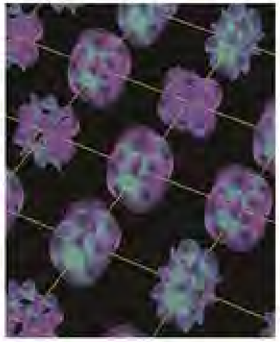

Curved Spacetime

Einstein's gravitational constant

$$G_{ab} + \Lambda g_{ab} = \kappa T_{\mu\nu}$$

the cosmological constant

mass-ENERGY-Matter

Despite the successes of General Relativity, the problems of quantum gravity and spacetime singularities remain open problems in its formulation of curved spacetime geometries

Copyright ABRAHAM [2008] - All rights reserved

ZPF

All Energies are quantised

Quantum theory

m

Gravitation is Universal

Newtonian Mechanics

M

Gravity results from curved spacetime

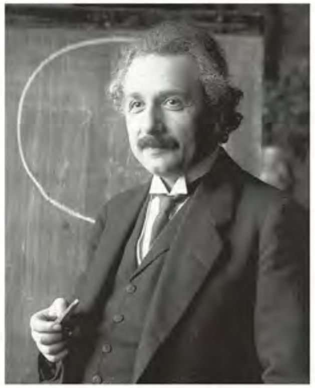

Relativity theory

VE

Quantum gravity is the effort in theoretical physics to create a theory that includes both general relativity and the standard model of particle physics. Currently, these two theories describe different scales of nature and attempts to explore the scale where they overlap yield results that don't quite make sense, like the force of gravity (or curvature of spacetime) becoming infinite.

Euclidean quantum gravity

Scalar theories

Modified Newtonian dynamics (MOND)

Einstein-Cartan Bimetric theories

Tensor-vector-scalar

DGP model

Geometrodynamics

Loop Quantum Gravity

Lovelock theory

Kaluza-Klein theory

Induced gravity

Canonical Quantum Gravity

Jackiw-Teitelboim gravity

Liouville gravity

Quantum gravity

String theory

Composite gravity

Causal Set theory

Brans-Dicke

Massive gravity

Gravity as it is currently understood cannot be reconciled with the laws of quantum mechanics.

Scalar-tensor-vector

Teleparallelism

Causal dynamical triangulation

2+1D topological gravity

Conformal gravity

Nonsymmetric gravitation

Gauss-Bonnet gravity

Supergravity

Gauge theory gravity

f(R) gravity

Unified theory

$h\nu$
1899
Planck's law

Q
1785
Couolomb's law

G
1686
Newton's law of Gravitation

SR
1905
Special relativity

GR
1915
General relativity

Copyright ABRAHAM [2008] - All rights reserved

Without the aid of Tetryonic's charged geometries
Gravitational & EM fields point to mathematical singularities

Inverse squared
Electric fields produce
INTERACTIVE
fields of force

Inverse squared
Magnetic fields produce
PURTURBATIVE
fields of force

There are NO
Black Holes

There are NO
point Particles

Matter produces
CONVERGENT
inverse squared fields
of Gravity

Gravitation is the nett
convergent force
produced by all
G-E-M fields

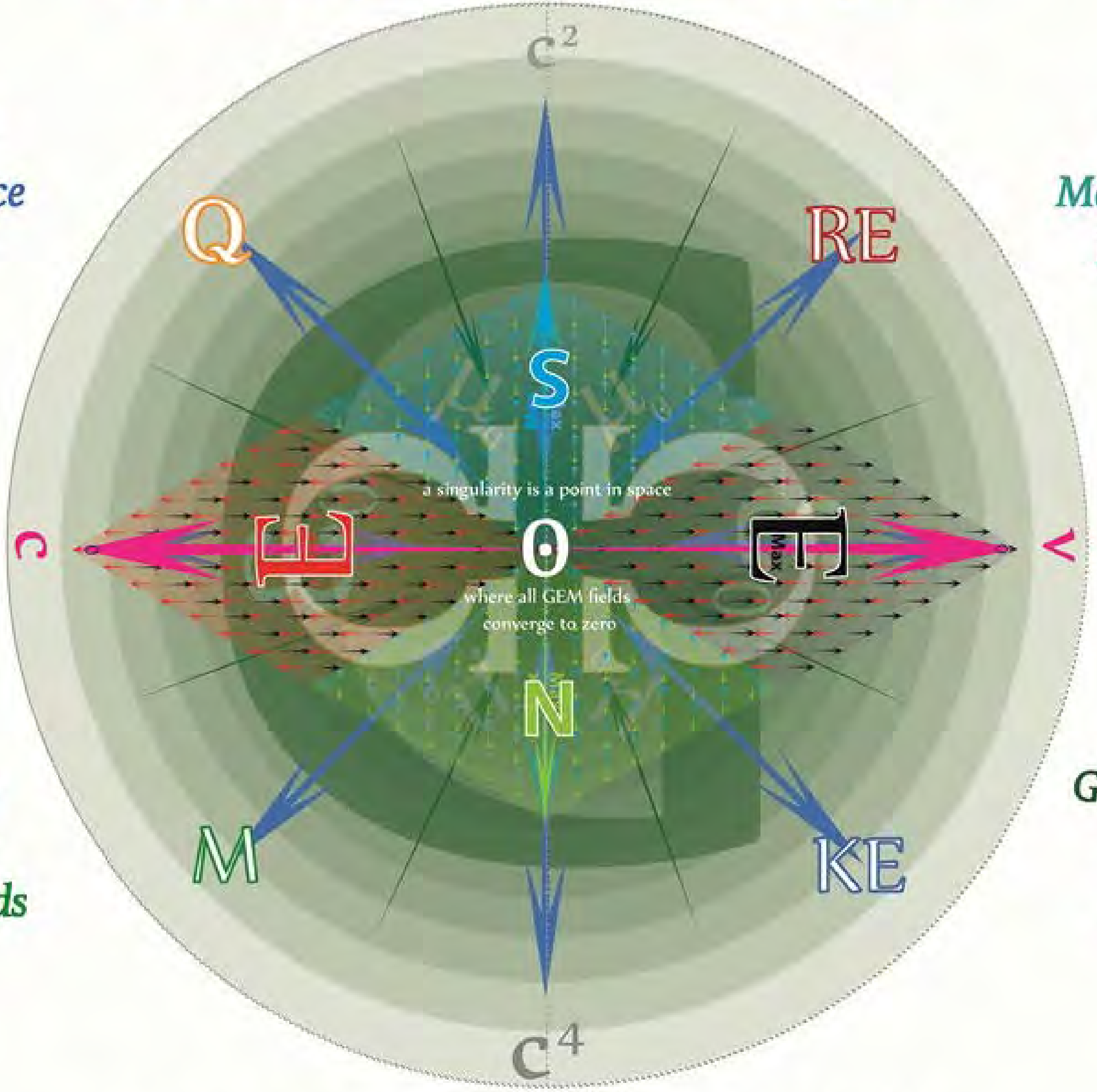
c^2
Q
RE
S
a singularity is a point in space
c
E
0
E
Max
v
where all GEM fields
converge to zero
N
M
KE
c^4

GravitoElectroMagnetic fields

Copyright ABRAHAM [2008] - All rights reserved

Matter topologies and gravity

the positive spatial displacements created by tetrahedral standing-wave mass-energies displace vacuum energies found in those regions to create a energy density pressure gradient

A convergent Gravity field is the result of the higher density-pressure vacuum energies in any spatial region seeking to equalise the lower pressure regions of nullspace within Matter topologies

Matter has a
Tetrahedral mass-energy
quantum topology
[non-Spherical]

M

The mass-energies
of Matter topologies
are stored in the integral
surface area geometries
of their charged fascia

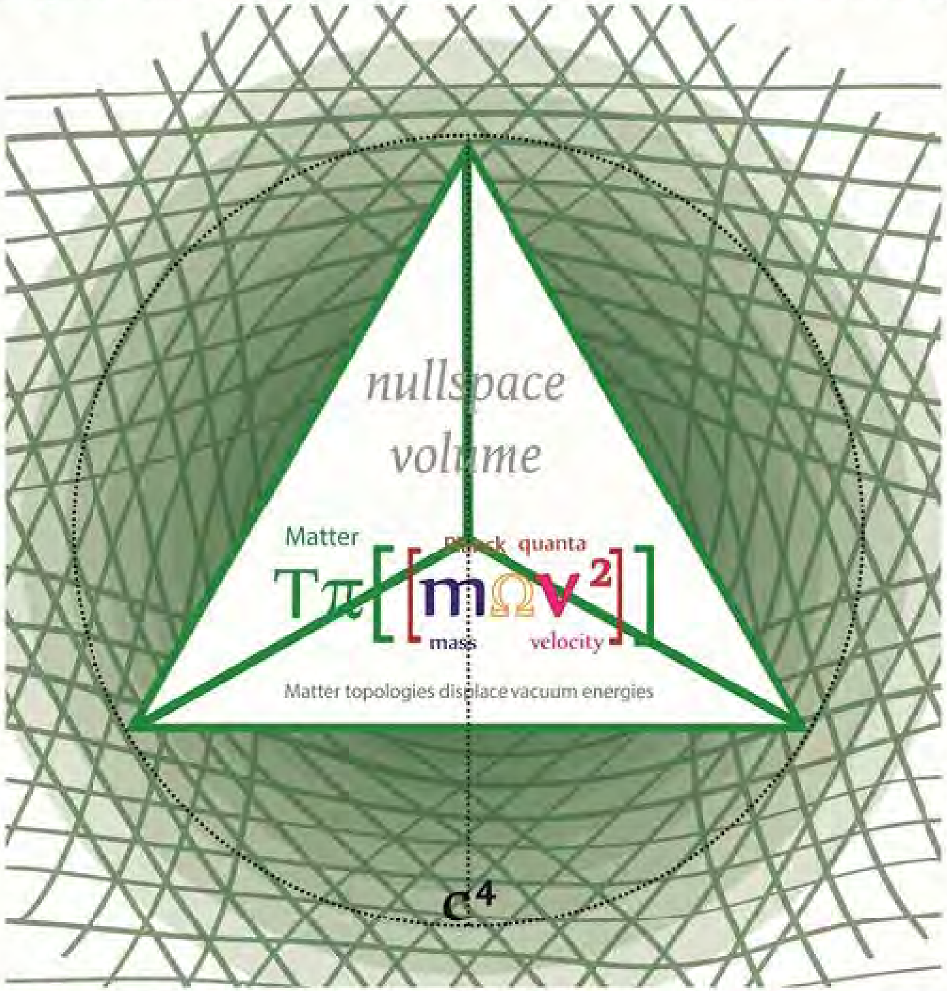

Gravity itself is
the result of vacuum
energy pressure gradients
not Coloumbic fields

G

Gravitation is the
result of the convergent
gravity AND interactive
[k]EM fields that all quantum
Matter topologies create

The resultant density-pressure gradients created by the displacement of vacuum energies by Matter topologies can be modelled using equilateral EM field geometries and a reduced Coloumb constant

Spatial displacement of GEM energies by Matter topologies

The positive 4nπ displacement of vacuum eneriges by Matter topologies creates

Spatial energy density-pressure

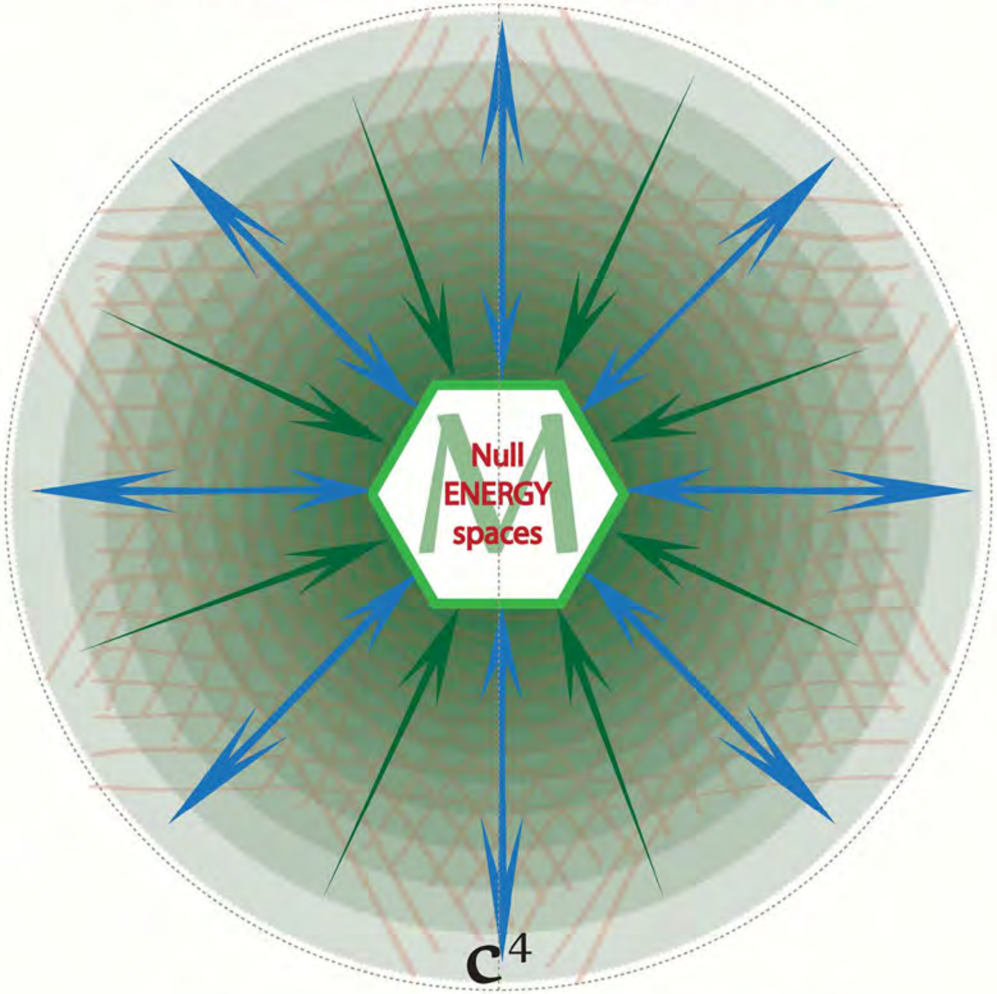

Convergent Gravitational fields

$$F = -G \frac{M_1 m_2}{r^2}$$

Planets orbit around the SUN

Newtonian fields models the resultant attraction as gravity

EM fields accelerate charged Matter

$$F = k \frac{Q_1 Q_2}{r^2}$$

$$G_{ab} = 8\pi G T_{\mu\nu}$$

All EM mass & Matter create stress energies

Einstein models the GEM energy densities as curved spacetime

Orbital precessions must be accounted for

Newton Einstein

4π vs. 8π

G G+SR

Copyright ABRAHAM [2008] - All rights reserved

Newtonian Gravity

model the observed, long range attraction between two bodies of Matter
[exclusive of the close range electromagnetic interactions of each body]

A gravity field is produced by the charged topology of mass-Matter as it displaces vacuum energies and creates intrisic nullspace energy differentials

The strength of a Gravity field produced by Matter is proportional to the total EM mass-energy geometry/topology of any body and its convergent force can be modelled with equilateral Tetryonic geometries

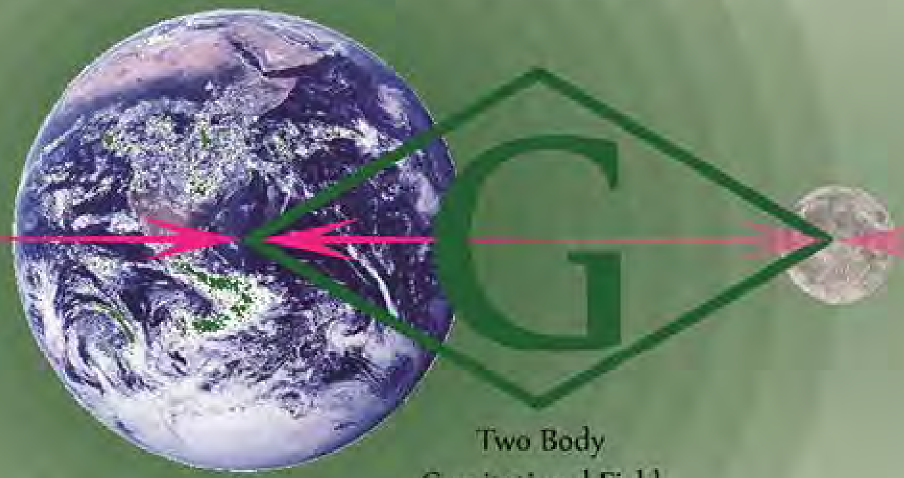

Single Body
Gravitational Field

$$F = -G\frac{M_1}{r^2}$$

Single Body
Gravitational Field

$$F = -G\frac{m_2}{r^2}$$

Two Body
Gravitational Field

$$F = -G\frac{M_1 m_2}{r^2}$$

However, unless it is at absolute zero all Matter topologies produce interactive EM fields that contribute to gravitational forces as evidenced with the perihelion of Mercury where later observations revealed inaccuracies in Newton's formulation for the motion of material bodies subject to convergent G forces & interactive EM fields

Copyright ABRAHAM [2008] - All rights reserved

Gravitational energies as a geometric mean

the geometric mean of two numbers is the square root of their product

EM standing waves [Matter] can be created and destroyed

Sir Issac Newton was the first to realise that gravity could be modelled as an inverse square Force

EM mass-energies can be neither created or destroyed

$$F = -G \frac{M_1 m_2}{r^2}$$

M

E

The Force of Gravity can be accurately modelled

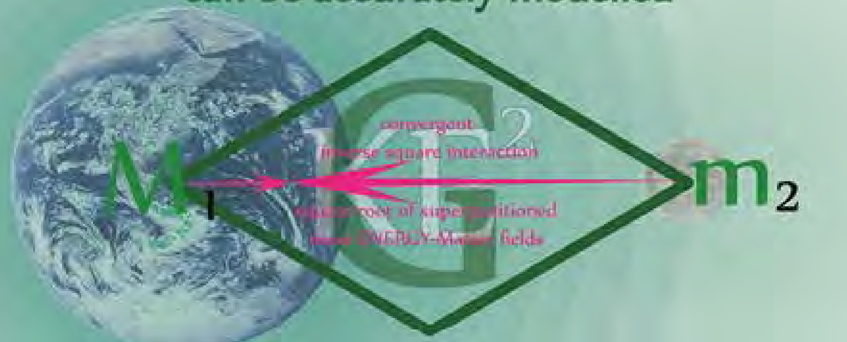

M_1
G^2
convergent
inverse square interaction
square root of superpositioned
mass ENERGY-Matter fields
m_2

as the convergent geometric MEAN of interactive superpositioned E-fields that represent the total mass-energies of the gravitational Matter

Gravitational field of mass-Matter only

$$F_G = 4\pi G \rho$$

[ignoring EM interactions]

Gravitational energy of Matter in motion

$$G_{ab} = 8\pi G T_{\mu\nu}$$

[including KEM field interactions]

$$F_G = \frac{M_1 . m_2}{r^2}$$

$$F_G = \sqrt{M_1 . m_2}$$

Gravity is the result of spatial energy density differentials created by the internal null-space volumes of Matter.
Energy (in all its forms) will seek to fill areas of lower pressure in search of universal equilbrium

Newton modelled the attractive forces between objects of Matter

Matter topologies are non-conservative and as they are destroyed the low pressure nullspace gradient [convergent gravitational Force] they created is replaced with conservative forms of radiative EM energy [interactive vacuum energies-photons]

GR models the energy density created by mass-Matter

Gravitational Tidal Forces

The tidal force is a secondary effect of the force of gravity resulting from the geometry of massive bodies attracting each other.

Its geometric effects allow acceleration fields to be distingushed from pure graviation fields

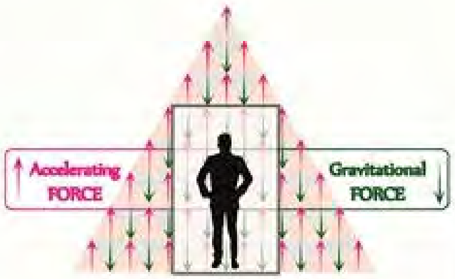

Over a small area on the surface of a Gravitating body, gravitational attraction is often modelled as a uniform field acting downwards across the whole area.

The Gravity of a material body of mass-Matter has historically been defined as a strictly attractive force between bodies

$$F = -G\frac{M_1 m_2}{r^2}$$

Tetryonics reveals Gravitation to be the nett result of 3 distinct interactions:
Convergent G-fields
Interactive E-fields and
Perturbative M-fields

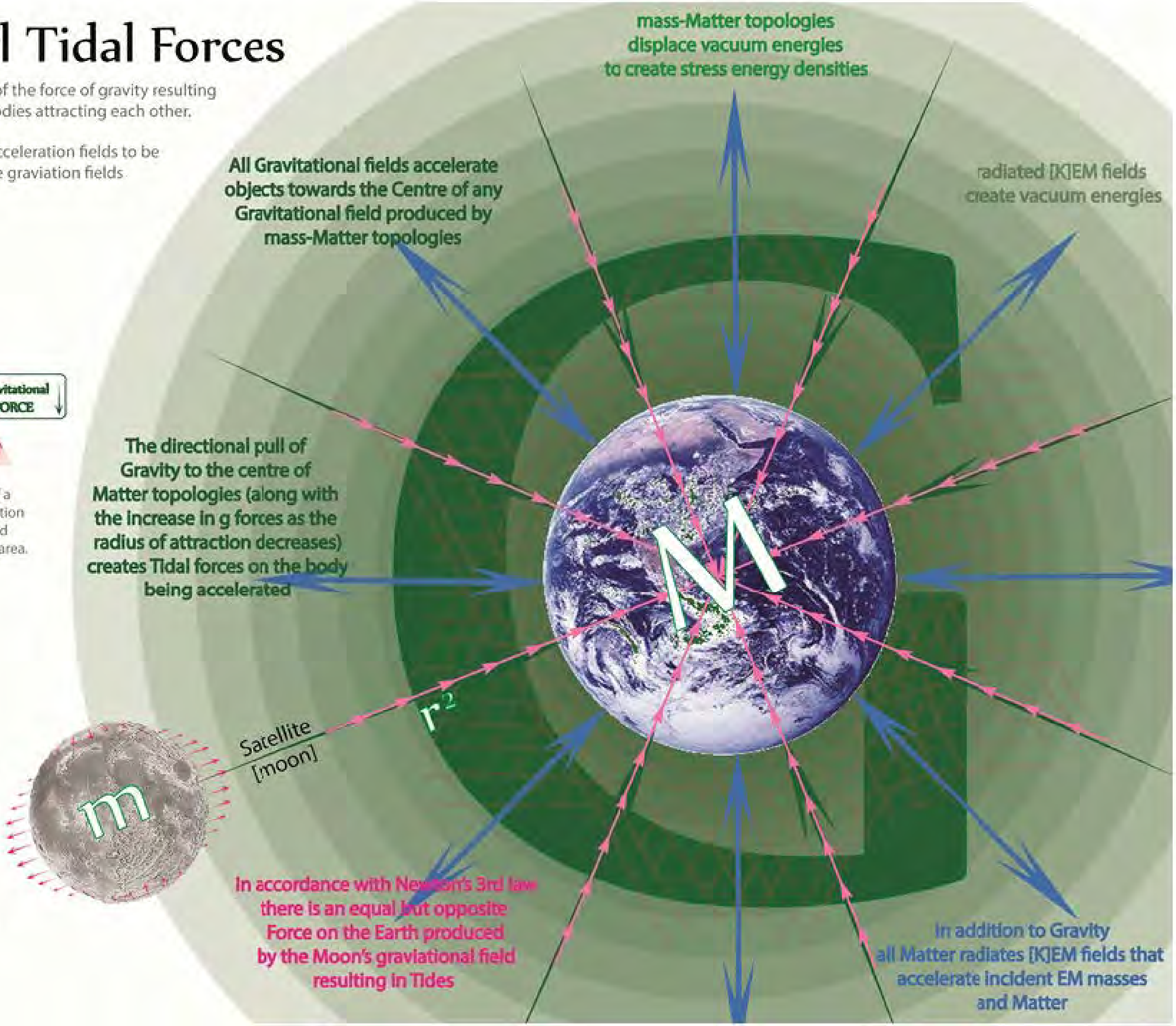

mass-Matter topologies displace vacuum energies to create stress energy densities

All Gravitational fields accelerate objects towards the Centre of any Gravitational field produced by mass-Matter topologies

radiated [K]EM fields create vacuum energies

The directional pull of Gravity to the centre of Matter topologies (along with the increase in g forces as the radius of attraction decreases) creates Tidal forces on the body being accelerated

In accordance with Newton's 3rd law there is an equal but opposite Force on the Earth produced by the Moon's graviational field resulting in Tides

In addition to Gravity all Matter radiates [K]EM fields that accelerate incident EM masses and Matter

Tetryonic Gravitation

is the result of the combined interactions
of radiant EM geometries & Matter topologies
in any spatial region of measurement

[k]EM fields & Vacuum energy gradients
create energy momenta pressure differentials
that accelerate Matter

3D mass-Matter topologies
create nullspace
vacuum energy differentials

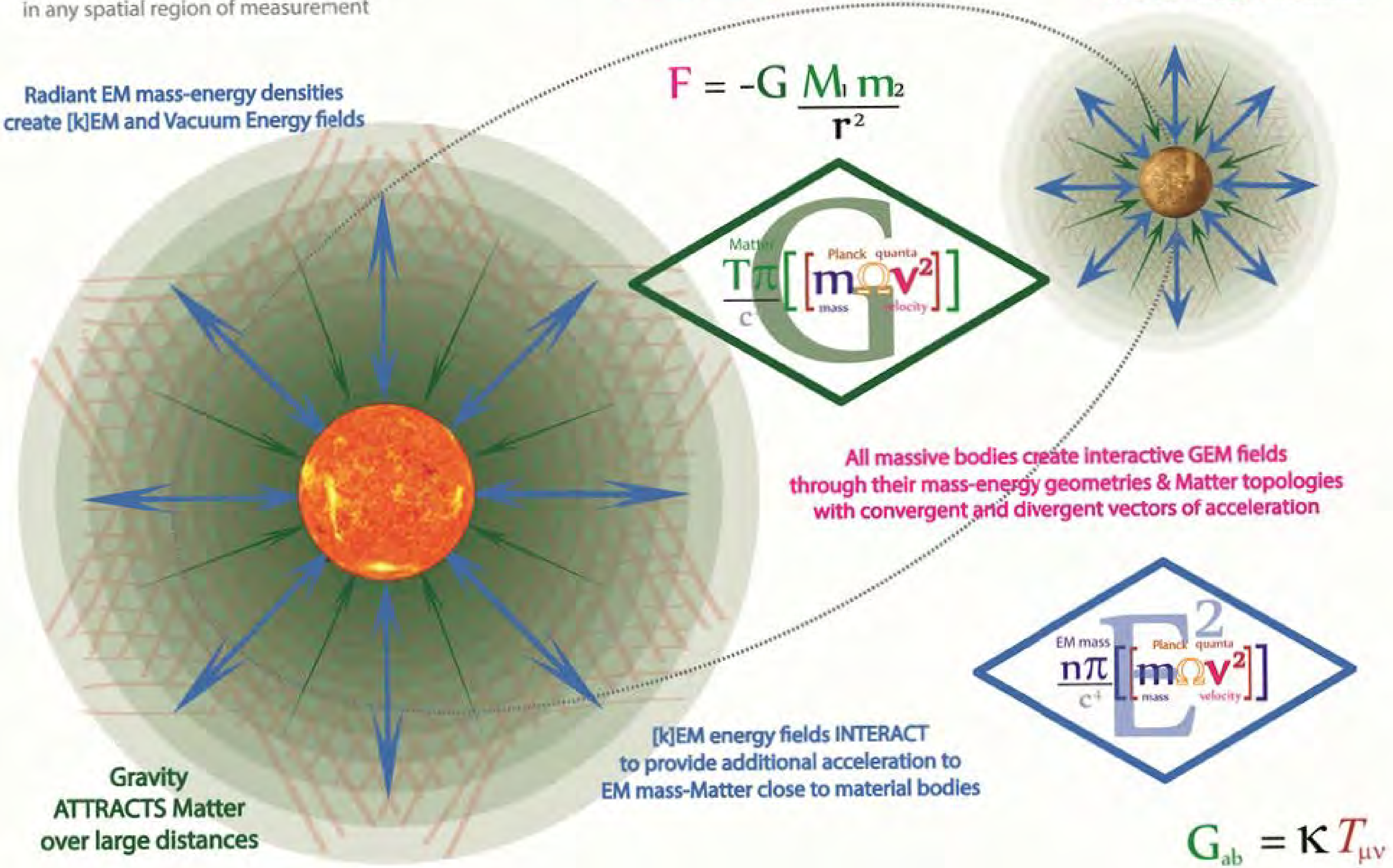
Matter
Tπ
c
Planck quanta
mΩv²
mass
velocity
G
EM mass
nπ
c⁴
Planck quanta
mΩv²
mass
velocity
E²

Radiant EM mass-energy densities
create [k]EM and Vacuum Energy fields

$$F = -G\frac{M_1 m_2}{r^2}$$

All massive bodies create interactive GEM fields
through their mass-energy geometries & Matter topologies
with convergent and divergent vectors of acceleration

[k]EM energy fields INTERACT
to provide additional acceleration to
EM mass-Matter close to material bodies

Gravity
ATTRACTS Matter
over large distances

$$G_{ab} = \kappa T_{\mu\nu}$$

Copyright ABRAHAM [2008] - All rights reserved

EM masses & Gravitational Matter

$$\frac{\rho}{m^2}$$

The core problem preventing the unification of Classical gravitation with QM & Relativity theory
is the imprecise formulation, definition and differentiation between mass & Matter

$$\frac{\rho}{m^3}$$

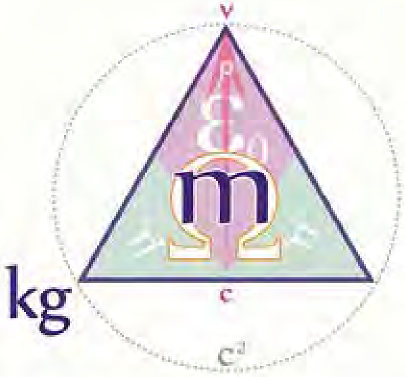

$$F = ma$$

mass is a conserved quantity
Matter is not a conserved quantity

$$F = -G\frac{M_1 m_2}{r^2}$$

inertial mass
active gravitational mass
passive gravitational mass

$$G_{ab} = \kappa T_{\mu\nu}$$

Even GR offers no clear distinction

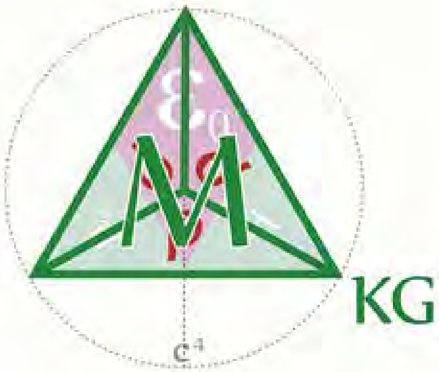

$$E = mc^2$$

radiant mass-energy geometry

$$E = Mc^4$$

standing-wave Matter topology

General relativity has been heralded as a highly successful model of gravitation and cosmology, which has so far passed every unambiguous observational and experimental test.

Even so, there are strong indications the theory is incomplete

Without a clear, concise understanding and accurate models of the relationship between mass & Matter [and their interactions] a Unified Field theory is impossible

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\overset{\text{Energy}}{\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}\,\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]}$$

mass is a geometric measure of a field's
planar energy momenta content per unit Time

2D masses

bosons photons

$$mc^2 = E = Mc^4$$

Note: in both cases the amount of Energy described is identical
It is the changing spatial co-ordinate systems used to measure Energy in all it forms
that allows for the differentiation betweeen mass-energy geometries & Matter-energy topologies

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\overset{\text{Energy}}{\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}\,\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]}$$

Matter is a topological measure of an object's
planar mass-energy content per unit of Time

3D Matter

Fermions, Baryons

Copyright ABRAHAM [2008] - All rights reserved

mass-ENERGY-Matter equivalence

Einstein's mass-energy equivalence can be further refined and differentiated through charge geometries [in turn creating a symetric mass-ENERGY-Matter equivalence formulation]

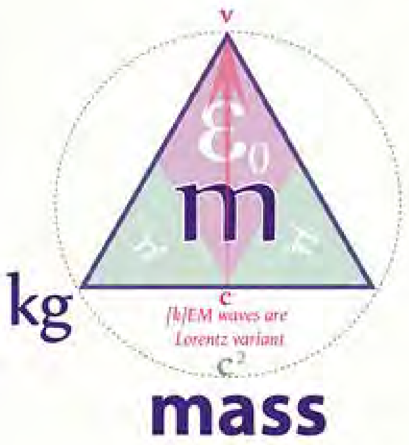
v
ε_0
m
c
[k]EM waves are
Lorentz variant
c^2
kg
mass

standing-wave
Z
topologies

The mass-energies of EM waveforms is contained within its charged geometry

$$E = mc^2$$

EM mass-energy is a property of Matter

radiant EM mass geometry

$$\boxed{\frac{m}{c^2} = M}$$

standing-wave Matter topology

$$m = \frac{E}{c^2}$$

$$\frac{E}{c^4} = M$$

Matter is NOT a property of EM mass-energy

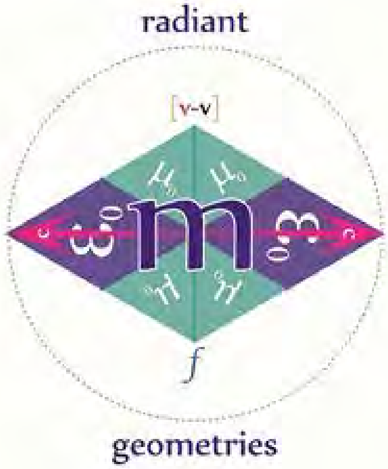
radiant
[v-v]
μ_0
μ_0
ε_0
ε_0
m
c
c
f
geometries

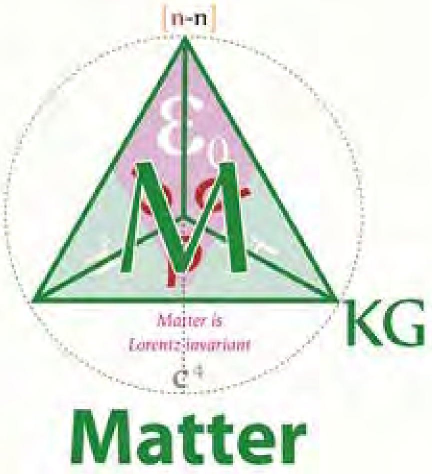
[n-n]
ε_0
M
Matter is
Lorentz invariant
c^4
KG
Matter

$$E = Mc^4$$

The mass-energies of Matter is contained entirely within its charged fascia topology

Tetryonic unified GEM fields

Equilateral Tetryonic geometry/topology provides a unified quantum field equation to model all Gravito-Electro-Magnetic mass-ENERGY-Matter interactions

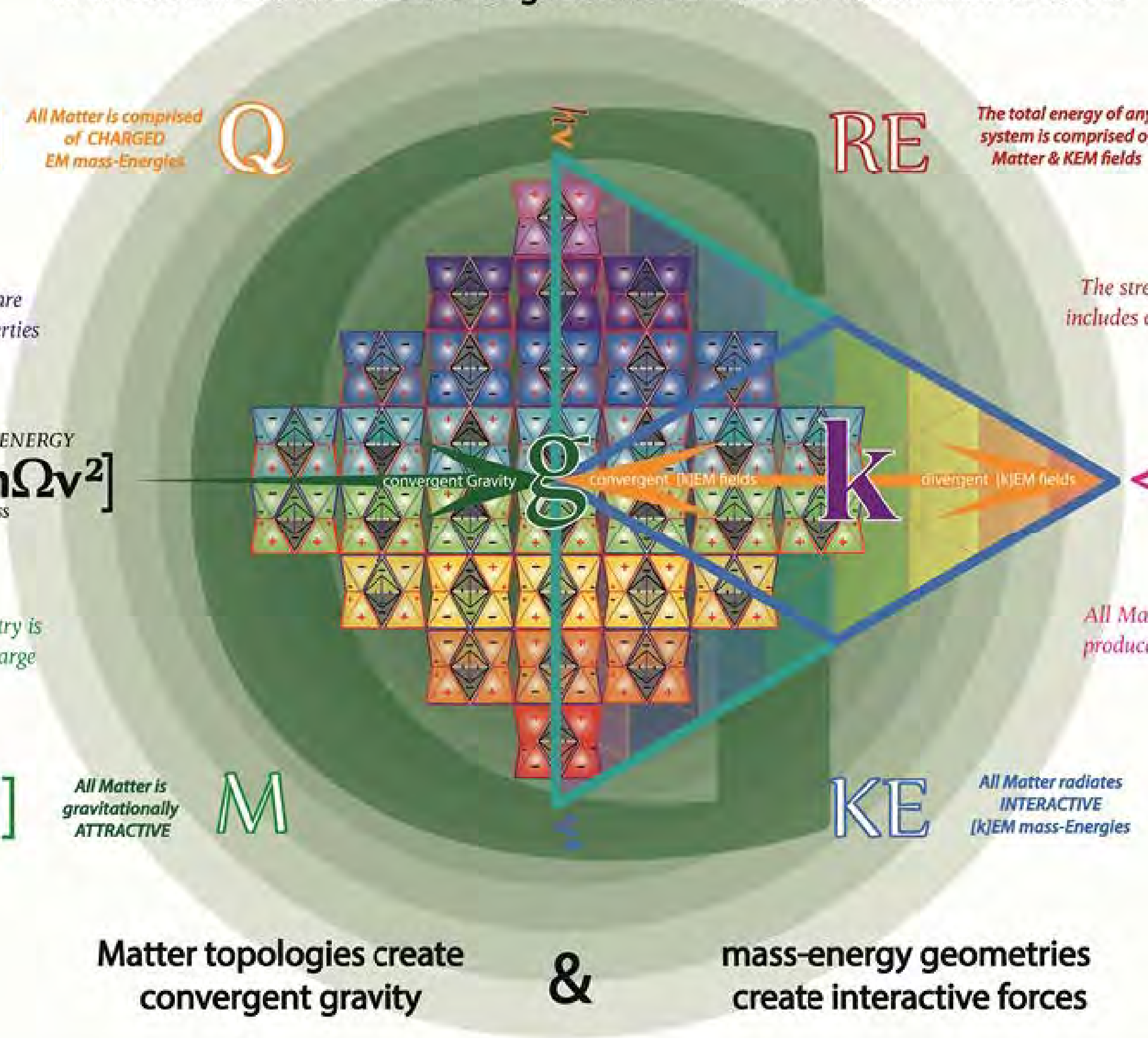

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter is comprised of CHARGED EM mass-Energies

Q

RE

The total energy of any system is comprised of Matter & KEM fields

ENERGY

$$n\pi\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

mass & Matter are distinct EM properties

The stress-Energy tensor includes all forms of energy

ENERGY

$$\underset{\text{Matter}}{T\pi}\left[\underset{\text{mass}}{m}\Omega v^2\right]$$

Equilateral energy momenta are geometrically related to the square of velocity

3D Matter geometry is determined by charge

All Matter in motion produces a [k]EM field

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

All Matter is gravitationally ATTRACTIVE

M

KE

All Matter radiates INTERACTIVE [k]EM mass-Energies

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Matter topologies create convergent gravity

&

mass-energy geometries create interactive forces

Copyright ABRAHAM [2008] - All rights reserved

Quantum Gravity

At present, one of the deepest problems of modern theoretical physics is the unification of General Relativity, which describes gravitation, and applies to large-scale structures (stars, planets, galaxies), with Quantum Mechanics, which describes the other three fundamental forces acting on the atomic scale

M

Newtonian Gravitation models observed motions of large-scale MATTER

G

The large scale Cosmos is ruled by GRAVITY

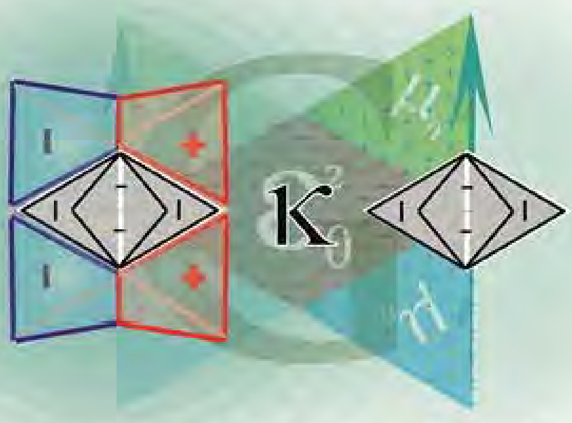

There are major points of tension between Quantum mechanics and General relativity.

Firstly, classical general relativity breaks down at singularities, and Quantum mechanics becomes inconsistent with general relativity in the neighborhood of singularities

Secondly, it is not clear how to determine the gravitational field of a particle, since under the Heisenberg uncertainty principle of Quantum mechanics its location and velocity at any instant of Time cannot be known with certainty.

Thirdly, there is the Problem of Time in Quantum gravity. Time has a different meanings in Quantum mechanics and General relativity and hence there are subtle issues to resolve when trying to formulate a theory which combines the two.

$T_{\mu\nu}$

Einstein's GR models interactions of all mass-energy-Matter

k

The Quantum world is ruled by ELECTROMAGNETISM

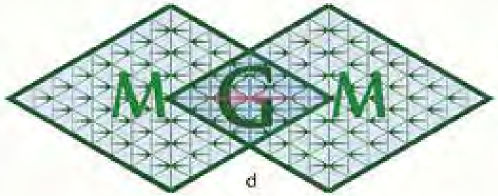

Gravitational attraction is a convergent Force between Matter

Newton

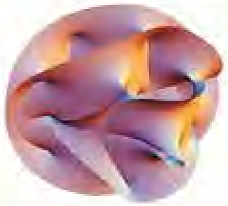

Gravitation results from energy density-pressure gradients

Einstein

Attempts to generalize ordinary quantum field theories, used in elementary particle physics to describe the fundamental interaction of gravity have to date led to serious problems.

Gravitons

Gravitons are not required in Tetryonic theory as gravitational attraction is a convergent force created by the EM energy differential produced by 3D Matter topologies in a vacuum energy field

All standing-wave Matter geometries create Gravity

G

CONVERGENT
Gravitation

Gravitons are the hypothetical carriers of the attractive Gravitational FORCE

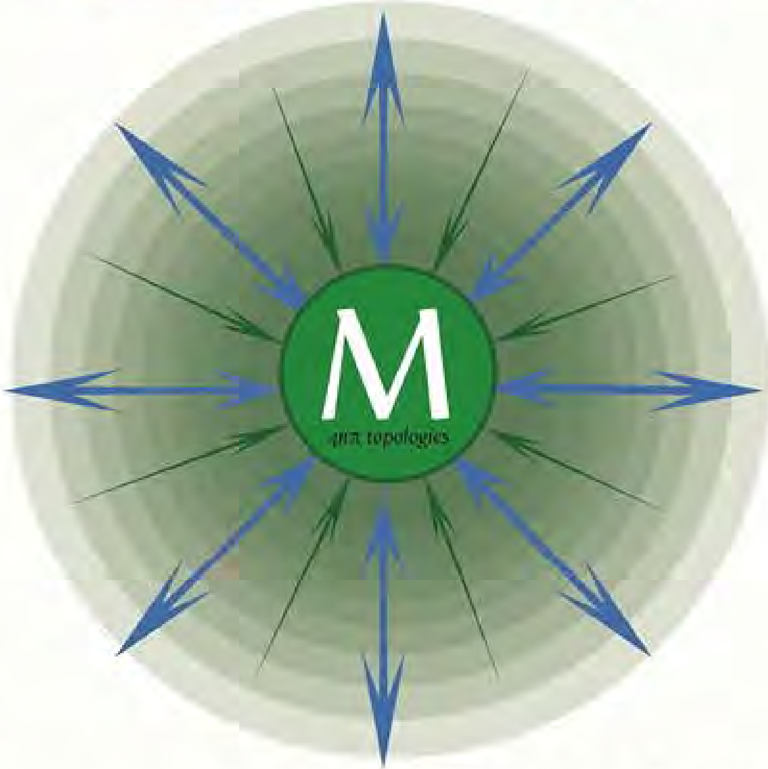

Radiant EM fields are produced by all Matter in motion

EM

DIVERGENT
ElectroMagnetism

Photons are the carriers of the interactive ElectroMagnetic FORCE

Gravitons were proposed as the gravitational equivalent of the Photon in ElectroMagnetic theory but have never been detected

4πG

8πG

They were postulated because of the great success of quantum field theory (in particular, the Standard Model) at modeling the behavior of all other known forces of nature as being mediated by elementary particles: electromagnetism by the photon, the strong interaction by the gluons, and the weak interaction by the W and Z bosons.

The hypothesis was that the gravitational interaction is likewise mediated by a – yet undiscovered – elementary particle, dubbed the graviton. In the classical limit the theory would allow Newton's law of gravitation in the weak-field limit to conform with Einstein's field equations of General relativity

Copyright ABRAHAM [2008] - All rights reserved

Laplacian field interactions

energy density

electric permittivity

impedance of a medium $Z = \frac{\rho}{c^2} = \varepsilon_0 \mu_0$

field geometry

magnetic permeability

The planar two-dimensional Laplacian is given by

$$\Delta = \nabla^2 = \nabla \quad \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

energy density

Einstein's constant

$G_P = \kappa \frac{\rho}{c^4}$

energy density pressure gradient

spatial volume

This fundamental differential operator of Applied Mathematics was introduced by Laplace in his 1782 study of the force of gravitational attraction exerted by spheroids and was named after him by Maxwell in his 1873 treatise on electromagnetism.

The net divergence of Positive charge fields is due to the anti-symmetry of their component Planck quanta

The net divergence of Negative charge fields is due to the anti-symmetry of their component Planck quanta

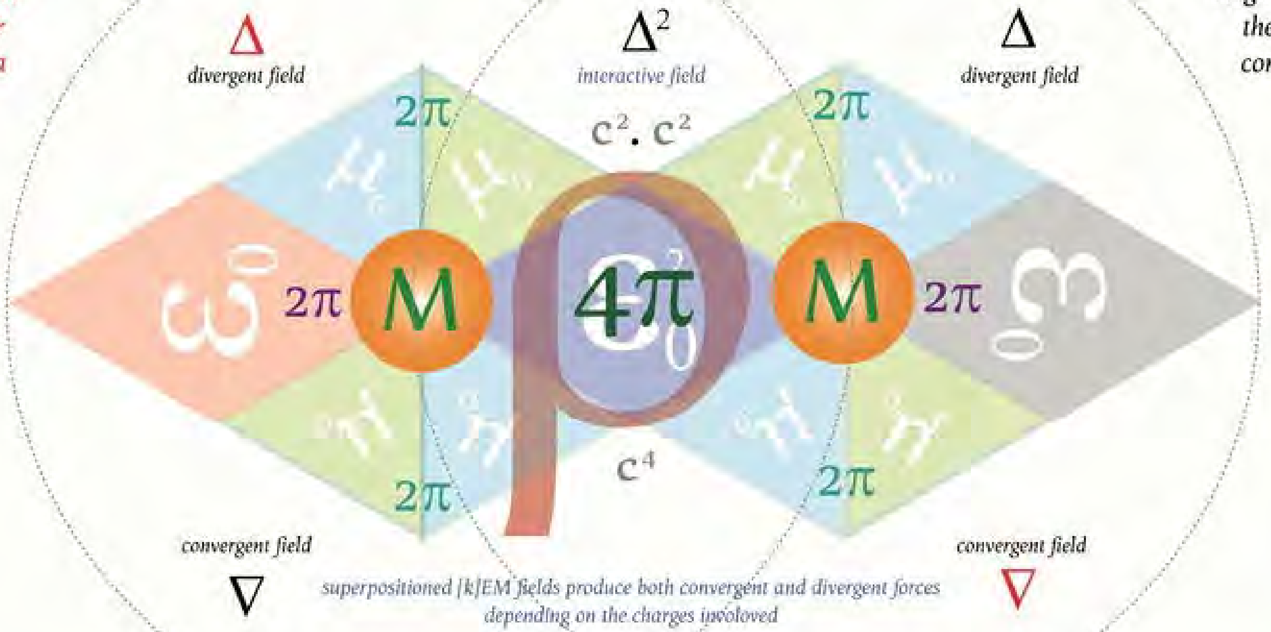
Δ
divergent field
Δ^2
interactive field
Δ
divergent field
2π
$c^2 . c^2$
2π
μ_0
μ_0
μ_0
μ_0
ε_0
2π
M
4π
ε_0
M
2π
ε_0
γ
γ
γ
γ
2π
c^4
2π
convergent field
∇
convergent field
∇

$$E = \frac{\rho}{c^2 \mu_0}$$

$$B = \frac{\rho}{c^2 \varepsilon_0}$$

superpositioned [k]EM fields produce both convergent and divergent forces depending on the charges involoved

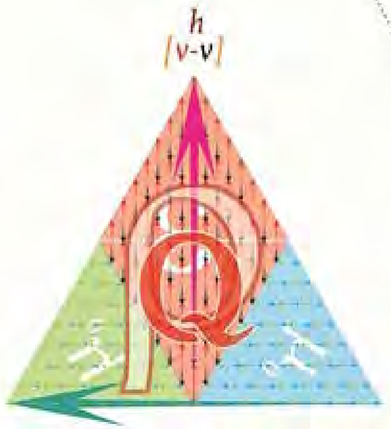
h
[ν-ν]
Q
γ
γ

h
[ν-ν]
Q
γ
γ

The Laplacian of any scalar field is interpreted as being the limit of the net outward flux per unit enclosed area of this field through a closed contour surrounding the point of evaluation

$$\Delta\phi = \lim_{A \to 0} \frac{1}{A} \oint_{\partial D} \frac{\partial \phi}{\partial \nu} \, dl, \quad A = \text{area}(D)$$

co-ordinate-free representation of the Laplacian

Newton's Gravitational Constant
has been shown to be a greatly
reduced version of Coulomb's
Electric interaction Constant

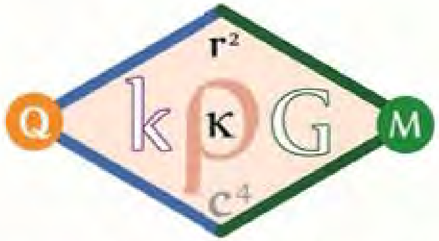

Newtonian Gravitation models
the observed convergent force of
GEM interactions that results in
the motion of large-scale Matter

*Any compete Unified field equation for
Gravitation must take into account both
the Nullspace gravitation resulting from
Matter topologies as well as the Coulombic
2D interaction fields that result from charge
and the motion of Matter through the aether*

Combined Interactive & Gravitational Forces

mass-energy-Matter

$$G_P = \kappa \frac{\rho}{c^4}$$

spatial volume

KE [1-0] KE [0-1]

[24-12] [0-12]

Coulomb's Law $$P+ \left[\frac{2\pi [mv^2]}{c^2} \right] k \left[\frac{2\pi [mv^2]}{c^2} \right] e-$$

Energy, mass, velocity

Q_1 c^4 Q_2

*Coulombic
[k]EM fields*

**Divergent [Interaction]
radiative EM Fields**

[k]EM interactive superpositioned EM fields of Force [k]EM

$$P+ \left[\frac{T\pi}{c^4}[m] + \left[n\pi \frac{[mv^2]}{c^2} \; \boldsymbol{\kappa} \; n\pi \frac{[mv^2]}{c^2} \right] + \frac{T\pi}{c^4}[m] \right] e-$$

Energy, mass, Lorentz, correction, Einstein's constant

Matter *convergent gravitational field of Force* *Matter*

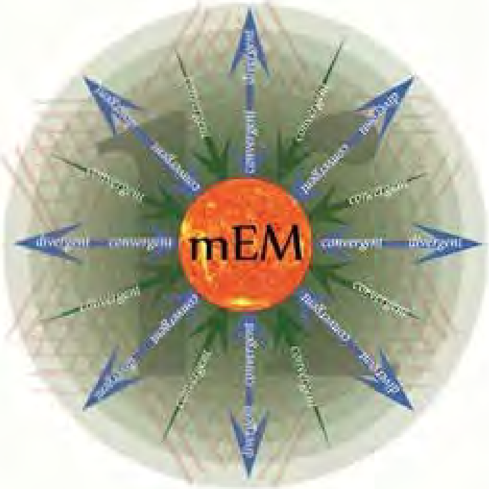

M_1 M_2

Newton's Law $$P+ \left[\frac{4n\pi [mc^2]}{c^2} \right] G \left[\frac{4n\pi [mc^2]}{c^2} \right] e-$$

Energy, mass, velocity

[2,25e23] c^4 [1.2e20]

*Gravitational
Matter*

**Convergent [Attractive]
Gravity Field**

*The observed net convergent force between
objects of mass-Matter are the result of
convergent G & interactive EM fields*

vacuum impedance

$$Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$$

electric permittivity

magnetic permeability

The energy density pressure gradient produced by the positive dispacement of vacuum energies by 3D standing-wave mass-Matter topologies is a strictly convergent force

mass-energy density

$$G_P = \kappa \frac{\rho}{c^4}$$

spatial region

$$P = \frac{kT}{V}$$

mass-Matter-Vacuum energy displacements contributes to STRESS energies

CHARGED mass-energy geometries create Matter topologies

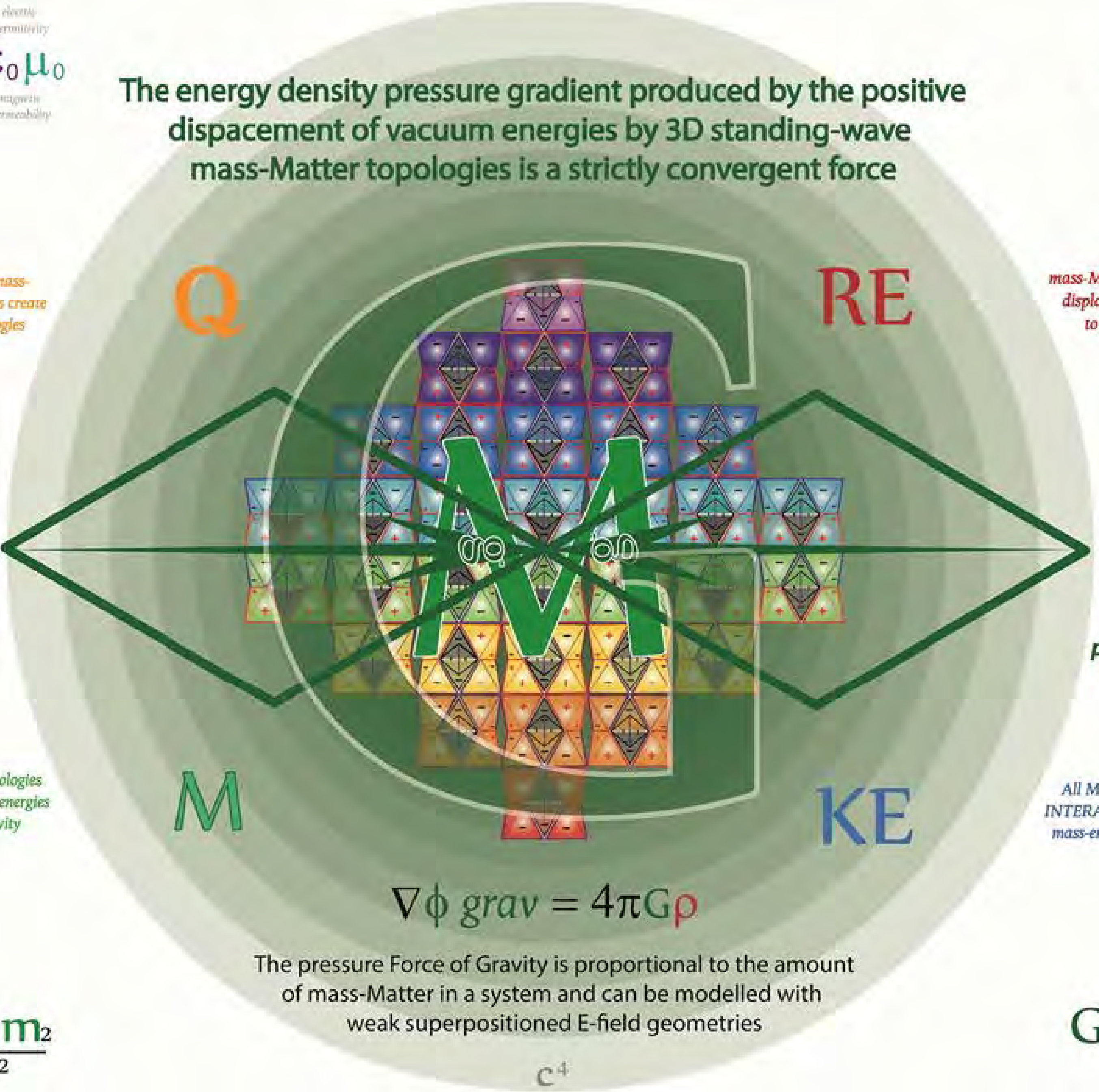

M

All Matter is gravitationally ATTRACTIVE

m

Gravity fields produce a convergent inverse square force

mass-Matter topologies displace vacuum energies to create gravity

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

$$\nabla \phi \; grav = 4\pi G \rho$$

The pressure Force of Gravity is proportional to the amount of mass-Matter in a system and can be modelled with weak superpositioned E-field geometries

c^4

$$F = -G \frac{M_1 m_2}{r^2}$$

$$G = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Inverse squared interactive E-fields

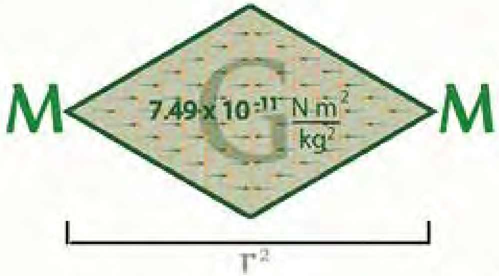

wherever mass-energies
create standing-wave
topologies they also
create Gravity

Although present, the Gravitational Attraction between quantum scale Matter particles is vastly overwhelmed by the Coulombic Interaction field resulting from the nett Charge imbalances of Matter and its fields of motion

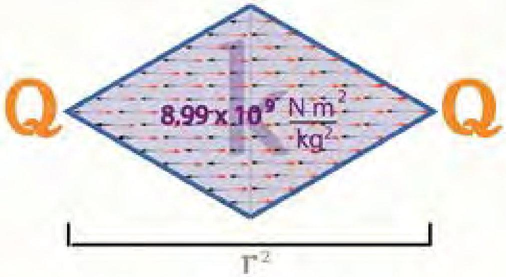

wherever there exists
mass-energy imbalances
energy will seek equilibirum
via radiant EM fields

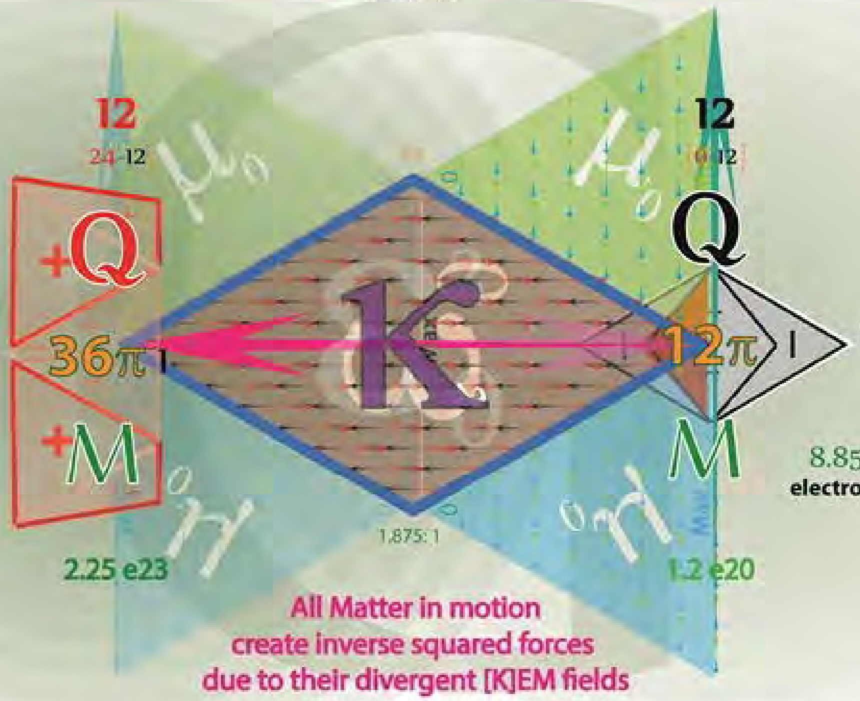

Protons have
Stronger G fields than electrons
but weaker [k]EM fields

P+

1.6596×10^{-27} KG
proton mass-Matter

electrons have
Weaker G fields than Protons
but stronger [k]EM fields

8.8514×10^{-31} KG
electron mass-Matter

All Matter in motion create inverse squared forces due to their divergent [K]EM fields

M

Matter mass-energy
36π * 25n²
topology geometry

KEM
superpositioned kEM fields create forces of attraction & repulsion
Gravity creates a strictly convergent force of attraction
G
c^4

Matter mass-energy
12π * 1n²
topology geometry

[k]EM

mass-Matter topologies displace vacuum energies to form convergent gravity

Equilateral [k]EM mass-energies contribute to the total stress energies of any region

Copyright ABRAHAM [2008] - All rights reserved

vacuum impedance

electric permittivity $\varepsilon_0 = \frac{1}{c^2 \mu_0}$ magnetic permeability

ε_0 = 8.8541878176 e $^{-12}$ F/m

Vacuum Permittivity

The Electric fields produced by Matter produce interactive nett divergent forces

Opposites ATTRACT
Similars REPEL

$P = \frac{kT}{V}$

Electric field permittivity contributes to STRESS energies

CHARGED mass-energy geometries create Matter topologies

Q

RE

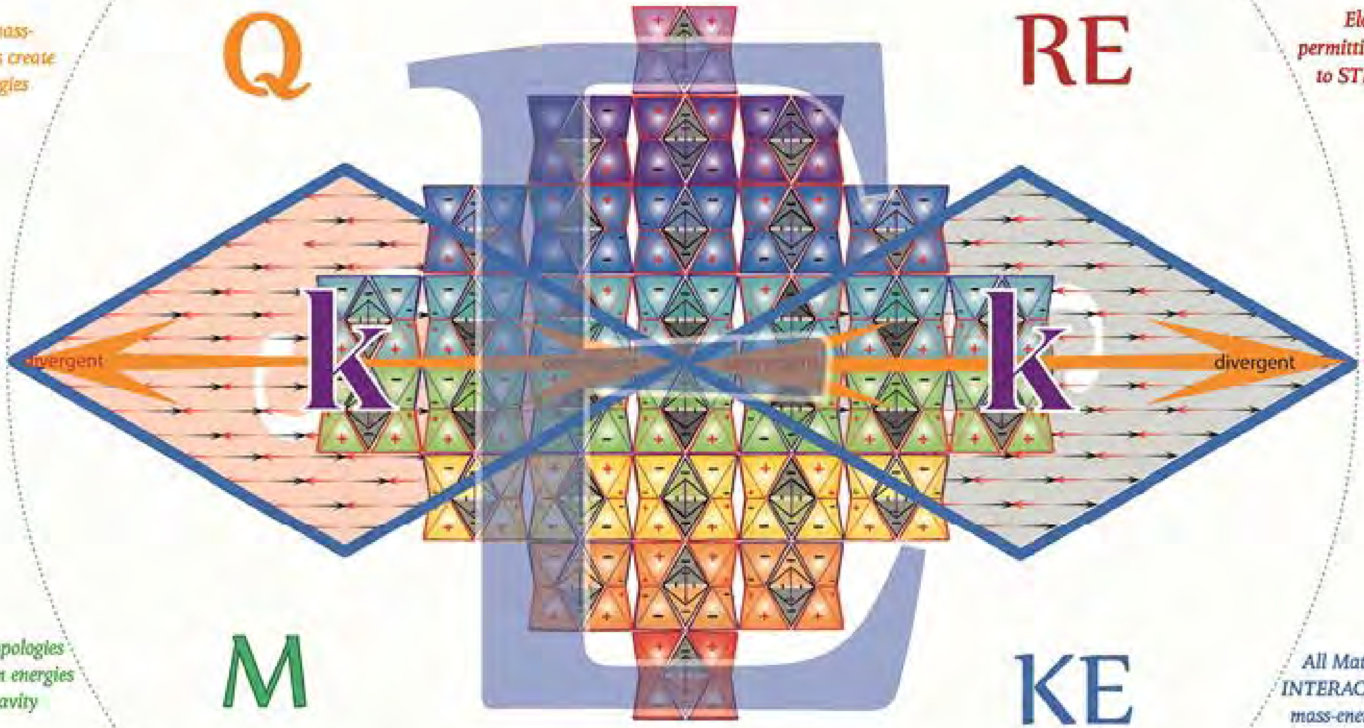

All Electric fields are nett divergent charge geometries >

< ***Electric fields follow the inverse square law***

M

KE

mass-Matter topologies displace vacuum energies to create gravity

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

Equal opposing E-field energy-momenta create electrostatic fields around Matter [that produce Lorentz forces]

vacuum impedance $Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$ electric permittivity, magnetic permeability

c^2

electric permittivity of a spatial region

$4\pi \frac{1}{c^2 \mu_0} \left[\left[m \Omega v^2 \right] \right]$

Planck quanta
mass
velocity

Perturbative M-forces

ElectroMagnetism is a velocity dependent vector interaction that dominates the Gravitational attraction of Matter

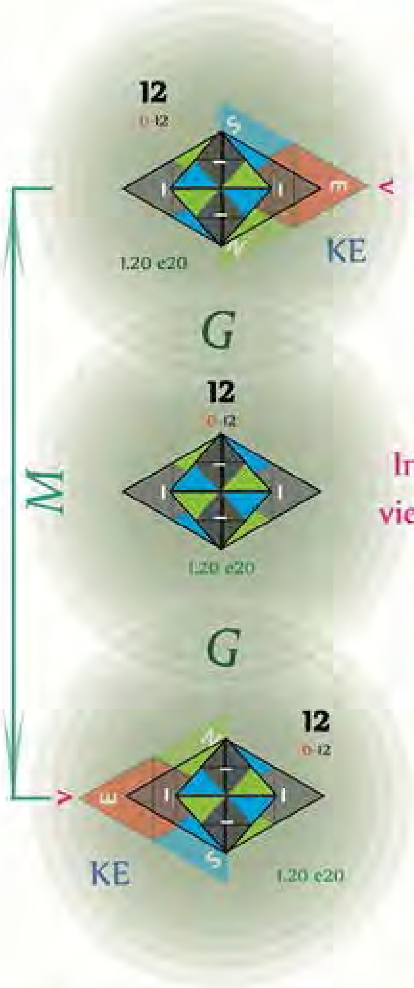

Inertial viewpoint

EM

Maxwells ElectroMagnetism

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}$$

$$\nabla \times \mathbf{B} = \frac{1}{\epsilon_0 c^2}\mathbf{J} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$$

[k]EM fields have electric and magnetic field components

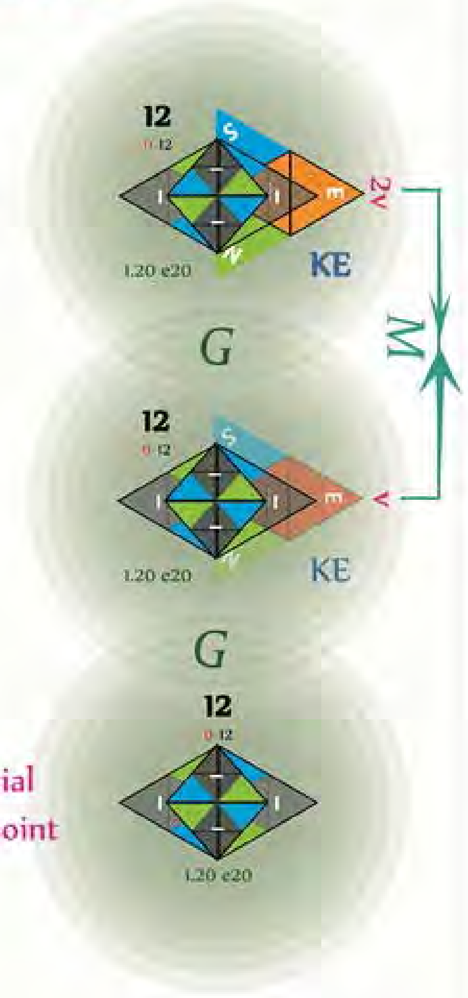

The Tetryonic GEM model reveals the interplay of Convergent [Gravity], interactive [E-fields] & perturbative [M-field] forces

Gravitational Matter is ALWAYS attractive vectorial ElectroMagnetic fields are interactive

Gravity can be modelled with pseudo-electric and pseudo-magnetic field components

GravitoElectroMagnetism

$$\nabla \cdot \mathbf{E}_g = -4\pi G \rho_g$$

$$\nabla \cdot \mathbf{B}_g = 0$$

$$\nabla \times \mathbf{E}_g = -\frac{\partial \mathbf{B}_g}{\partial t}$$

$$\nabla \times \mathbf{B}_g = -\frac{4\pi G}{c^2}\mathbf{J}_g + \frac{1}{c^2}\frac{\partial \mathbf{E}_g}{\partial t}$$

G

Inertial viewpoint

The sign of the effect is different (attraction in one case, repulsion in another), but this is just because in electromagnetism the sign of the electric part of the acceleration is opposite that of gravity: electric charges of the same type repel each other, while in gravity two massive objects attract.

Apart from this sign, there is such a close analogy to magnetism that this velocity-dependent gravitational effect is often termed Gravito-magnetism

The Gravitational forces on the central particle are always equal irrespective of the inertial viewpoints

The ElectroMagnetic forces resulting from inertial motion in a different frame of reference produces different KINETIC [EM] forces

vacuum impedance

magnetic permeability $\mu_0 = \frac{1}{c^2 \varepsilon_0}$ electric permittivity

μ_0 = 1.2566370614 e^{-12} N/A^2

Vacuum Permeability

The Magnetic fields of Matter produce perturbative interactions

$P = \frac{kT}{V}$

Magnetic field permeability contributes to STRESS energies

CHARGED mass-energy geometries create Matter topologies

Q

RE

Magnetic dipoles follow the inverse cubed law

Magnetic fields follow the inverse square law

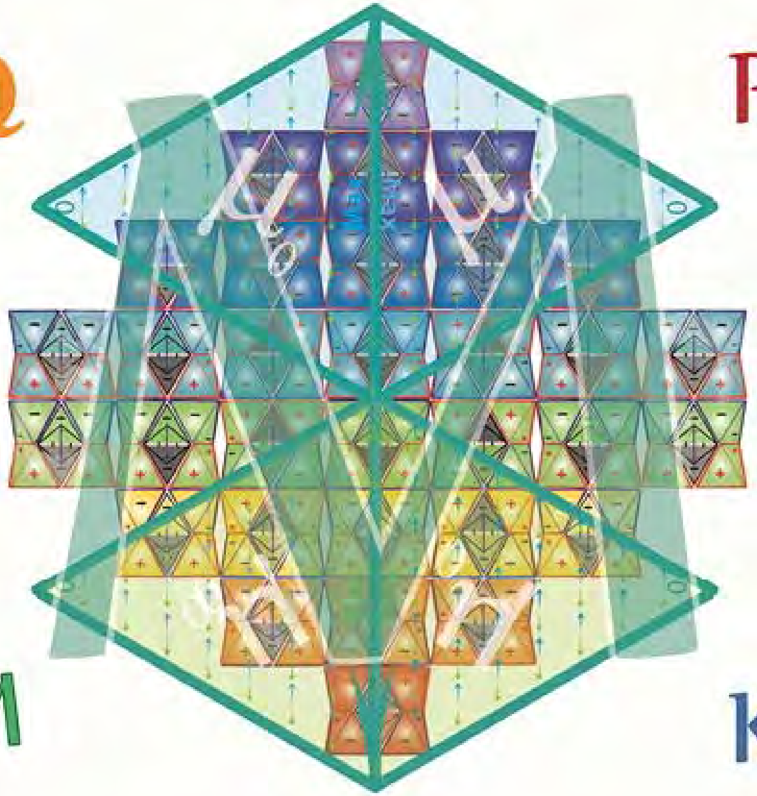

M

KE

mass-Matter topologies displace vacuum energies to create gravity

All Matter radiates INTERACTIVE [k]EM mass-energy momenta

Matter posesses a secondary magnetic moment that is subject to Lorentz forces

vacuum impedance $Z = \frac{1}{c^2} = \varepsilon_0 \mu_0$ (electric permittivity; magnetic permeability)

c^2

magnetic permeability of a spatial region

$4\pi \frac{1}{c^2 \varepsilon_0} \left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}} \overset{\text{quanta}}{\Omega} \underset{\text{velocity}}{v^2} \right] \right]$

Gravitational & EM fields

Any unified Field equation that attempts to accurately model celestial scale gravitational attraction must also take into account KEM field interactions between the bodies resulting from their respective motions

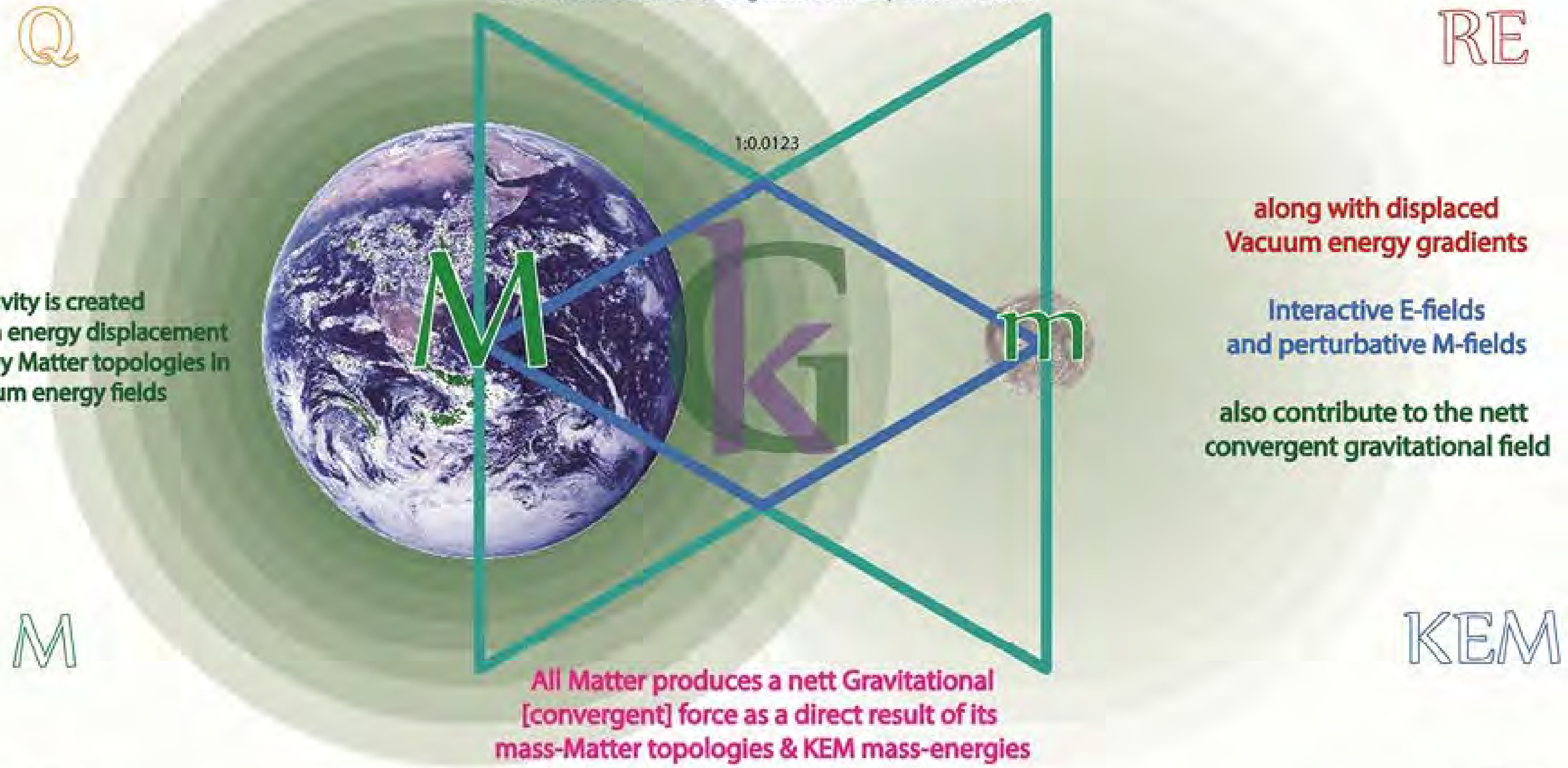

Gravity is created by vacuum energy displacement produced by Matter topologies in vacuum energy fields

along with displaced Vacuum energy gradients

Interactive E-fields and perturbative M-fields

also contribute to the nett convergent gravitational field

All Matter produces a nett Gravitational [convergent] force as a direct result of its mass-Matter topologies & KEM mass-energies

mass-Matter
5.9736×10^{24} KG
Earth
$2.8673e26\ \pi$
displacement topology

interactive KEM field mass-energy momenta

$$\left[\frac{\overset{\text{Matter}}{4n\pi}}{c^4}\left[\overset{\text{Energy}}{\rho}\right]_{\text{mass}} + \left[\frac{2n\pi\left[\overset{\text{EM mass}}{\rho}\right]}{c^2}\;K\;\frac{2n\pi\left[\overset{\text{EM mass}}{\rho}\right]}{c^2}\right] + \frac{\overset{\text{Matter}}{4n\pi}}{c^4}\left[\overset{\text{Energy}}{\rho}\right]_{\text{mass}}\right]$$

$$r^2$$

mass-Matter
7.3477×10^{22} KG
moon
$3.526e24\ \pi$
displacement topology

Gravitational gradients can be modelled as weak KEM fields

Independently and almost simultaneously in1964 three groups of physicists proposed the proposed the Higgs mechanism thorugh which the inertial mass properties of Matter are created:
François Englert and Robert Brout; by Peter Higgs (inspired by ideas of Philip Anderson); and by Gerald Guralnik, C. R. Hagen, and Tom Kibble.

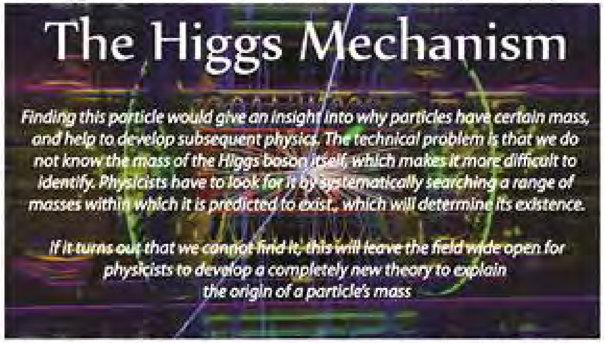

The Higgs Mechanism

Finding this particle would give an insight into why particles have certain mass, and help to develop subsequent physics. The technical problem is that we do not know the mass of the Higgs boson itself, which makes it more difficult to identify. Physicists have to look for it by systematically searching a range of masses within which it is predicted to exist., which will determine its existence.

If it turns out that we cannot find it, this will leave the field wide open for physicists to develop a completely new theory to explain the origin of a particle's mass

The Higgs mechanism was incorporated into modern particle physics by Steven Weinberg and Abdus Salam, and is an essential part of the standard model.

As the Universe cooled and the temperature fell below a critical value, an invisible force field called the 'Higgs field' was formed together with the associated 'Higgs boson'

The Higgs mechanism is a process by which vector bosons can obtain mass

M

TETRYONS / FERMIONS / BARYONS
3D tetrahedral standing-waves
[4nπ EM mass-Matter topologies]
interact with the vacuum energy aether

Matter

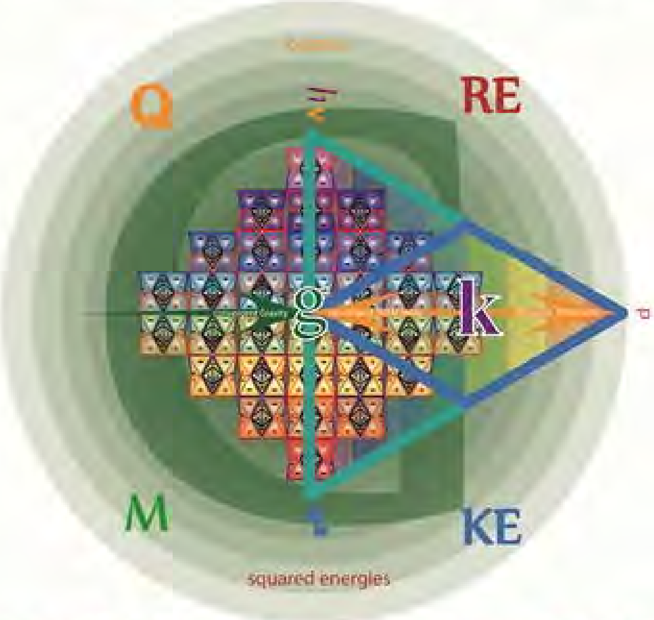

m

BOSONS / PHOTONS
2D planar matterless fields
[nπ EM mass-energy geometries]
'slice' through the vacuum energy aether

EM masses

Tetryonics is a full relativistic unified theory of electromagnetic mass-ENERGY & Matter and motion in which the classical properties of inertial mass and quantum mechanical quantised energy momenta are explained as being direct result of the geometric properties of inductive equilateral energies moving through external electromagnetic fields

Applying the Unified Field Equation

for ElectroMagnetic masses & Matter interactions

All Matter has a
4π standing-wave
mass-energy topology

Matter topologies

$$G_{ab} = 8\pi G T_{\mu\nu}$$

The net gravitational force is the geometric mean of all EM mass-ENERGY-Matter interactions

All kEM waves have
4π superpositioned
mass-energy geometries

mass-ENERGY geometries

$$-G\frac{M_1 m_2}{r^2} = F$$

$$F = k\frac{Q_1 Q_2}{r^2}$$

G

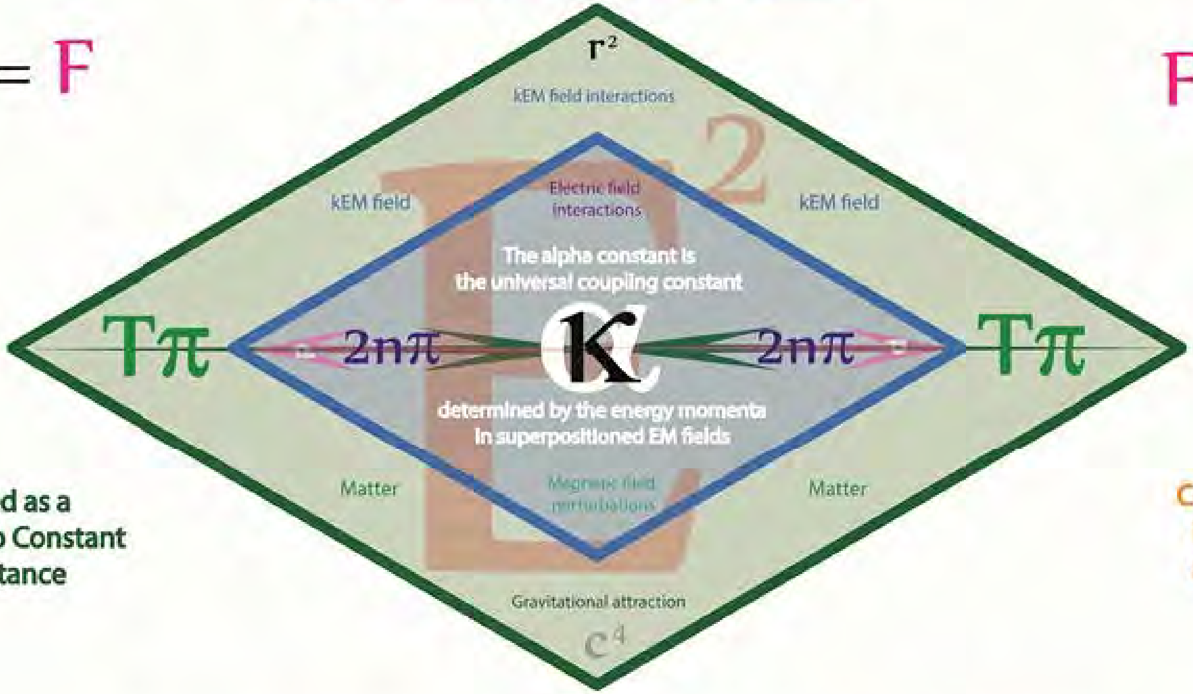
r²
kEM field interactions
kEM field
Electric field interactions
kEM field
The alpha constant is the universal coupling constant
Tπ
2nπ
α
2nπ
Tπ
determined by the energy momenta in superpositioned EM fields
Matter
Magnetic field perturbations
Matter
Gravitational attraction
c⁴
E²

k

Gravity can be modelled as a reduced negative Coloumb Constant acting over a large distance

Coloumb's Constant is a large interactive Electric constant acting over a short distance

Tetryons - 4π
Leptons - 12π
Quarks - 12π
Mesons - 24π
Baryons - 36π

Ignoring interactive [k]EM field interactions reduces GR [k] to Newtonian [G]

Charges - 1π

Photons -2π

EM Fields - 2nπ

$$G_{ab} = 4\pi G T_{\mu\nu}$$

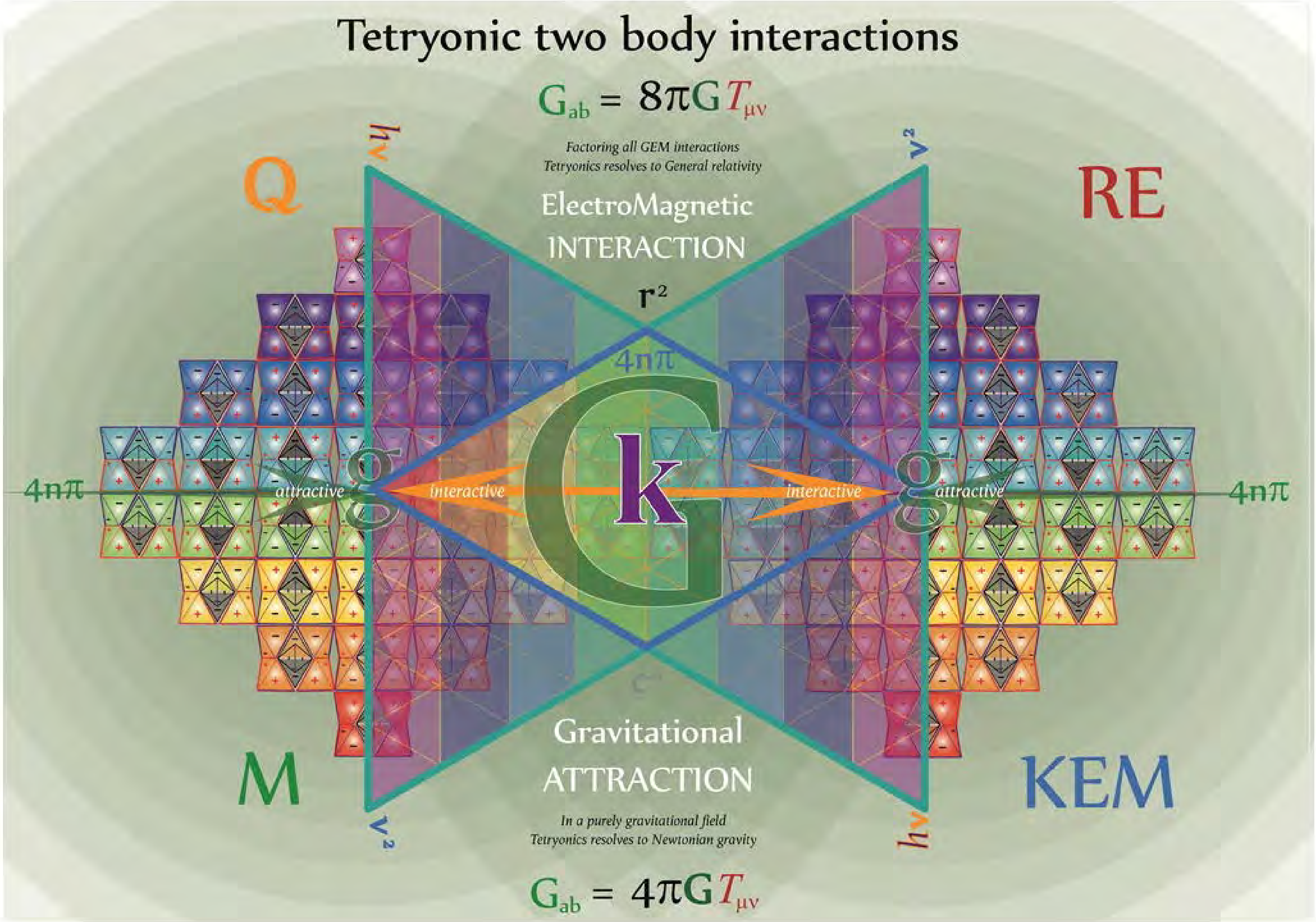

Tetryonic two body interactions

$$G_{ab} = 8\pi G T_{\mu\nu}$$

Factoring all GEM interactions
Tetryonics resolves to General relativity

ElectroMagnetic
INTERACTION

Gravitational
ATTRACTION

In a purely gravitational field
Tetryonics resolves to Newtonian gravity

$$G_{ab} = 4\pi G T_{\mu\nu}$$

Q
RE
M
KEM
hv
v²
v²
hv
r²
4nπ
G
k
g
g
4nπ
4nπ
attractive
interactive
interactive
attractive
c²

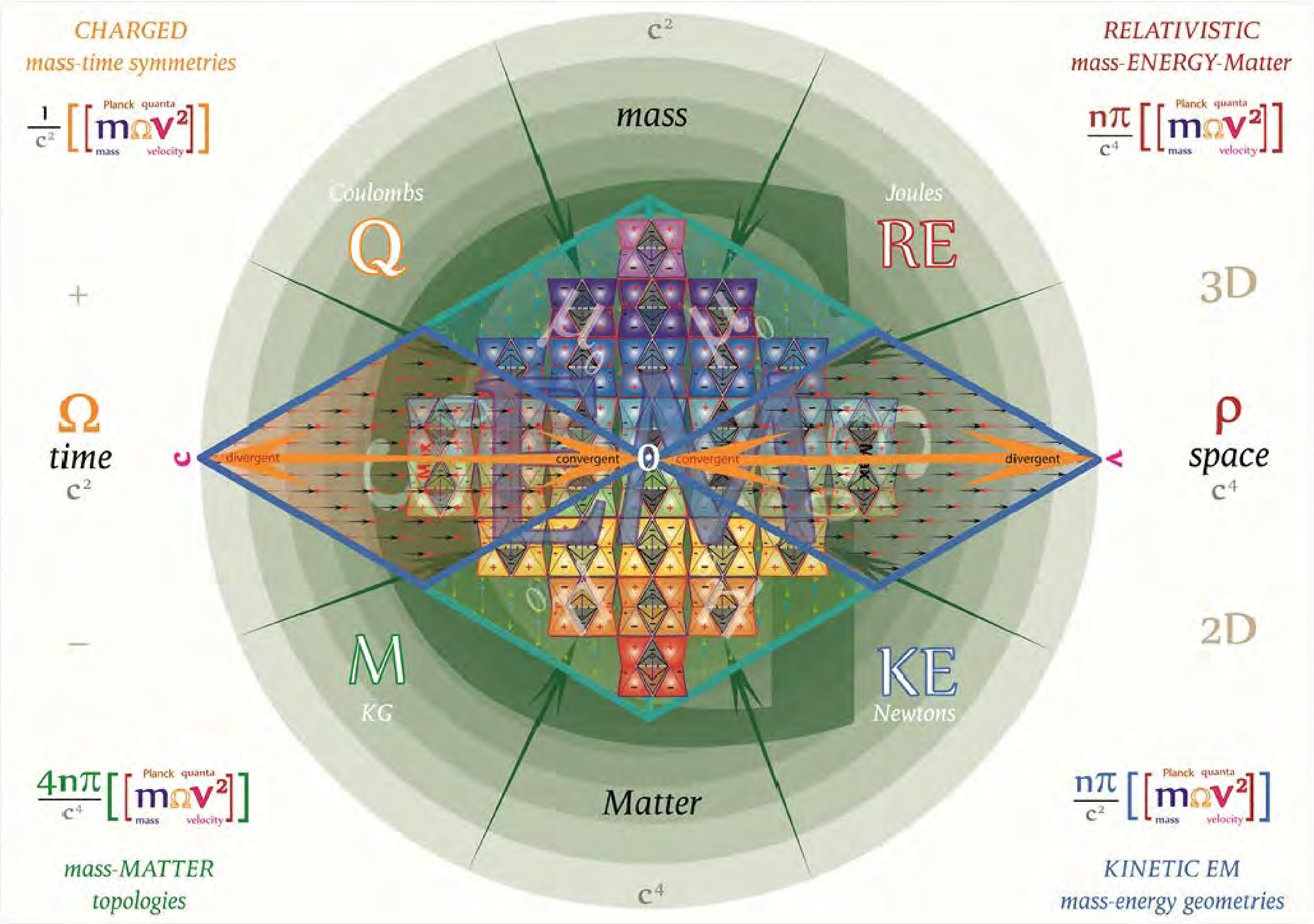
CHARGED
mass-time symmetries
$\frac{1}{c^2}[[m\Omega v^2]]$ Planck quanta mass velocity
RELATIVISTIC
mass-ENERGY-Matter
$\frac{n\pi}{c^4}[[m\Omega v^2]]$ Planck quanta mass velocity
c^2
mass
Coulombs
Q
Joules
RE
3D
+
Ω
time
c^2
c
divergent
convergent
0
convergent
divergent
<
ρ
space
c^4
2D
–
M
KG
KE
Newtons
Matter
c^4
$\frac{4n\pi}{c^4}[[m\Omega v^2]]$ Planck quanta mass velocity
mass-MATTER
topologies
$\frac{n\pi}{c^2}[[m\Omega v^2]]$ Planck quanta mass velocity
KINETIC EM
mass-energy geometries

Tetryonics 71.14 - Quantum GEM space-time fields

Our EM SUN

The SUN (like the Planets) is a CHARGED body moving through an EM environment

G & E fields obey the inverse SQUARED laws

The nett Charge of a body is the result of an excess of one type of energy quanta and is reflective of their charged topologies [ie ionised Matter]

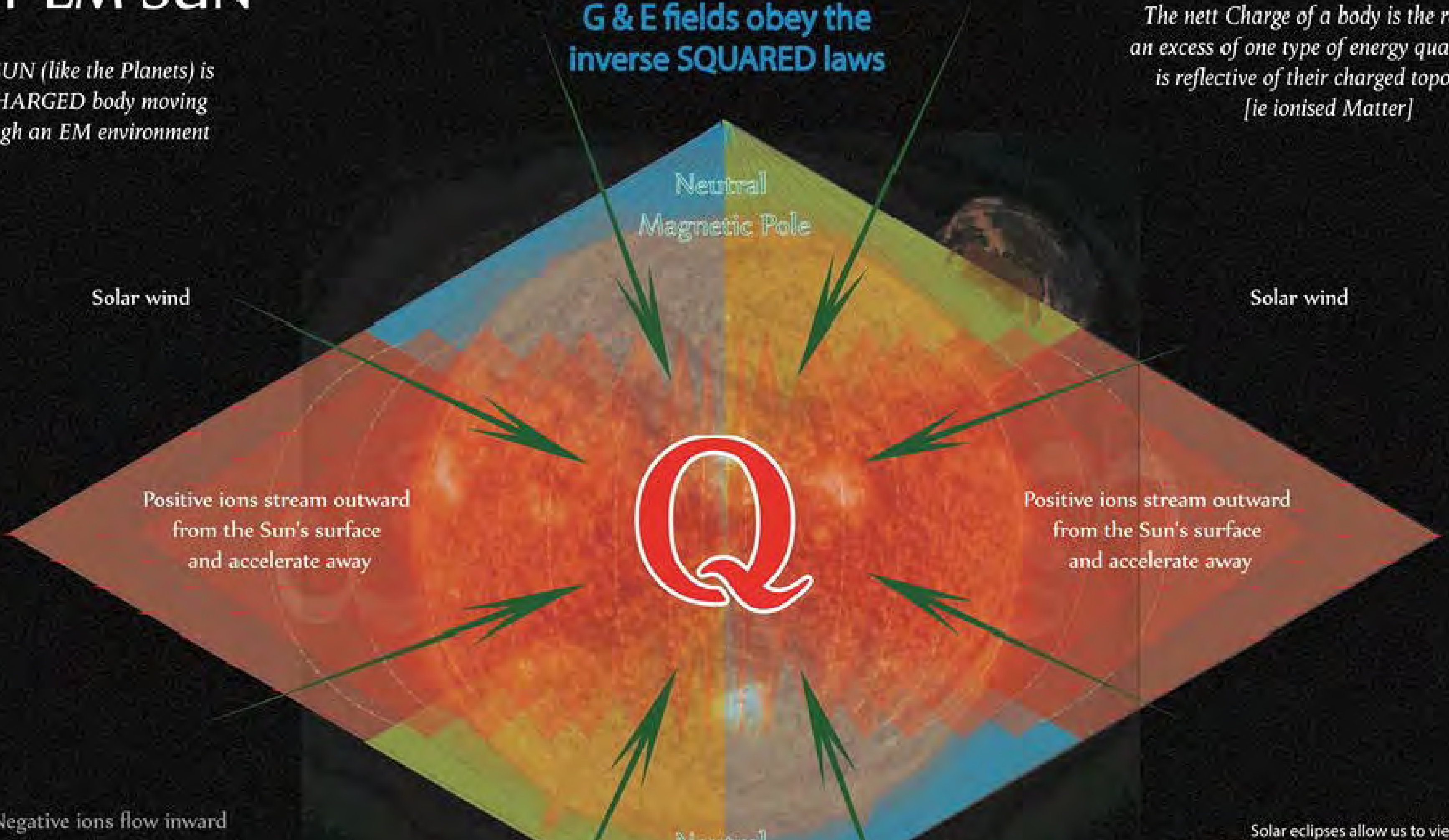

Solar wind

Solar wind

Positive ions stream outward from the Sun's surface and accelerate away

Positive ions stream outward from the Sun's surface and accelerate away

Negative ions flow inward

The charged E field [heliosphere] of our SUN interacts with other charged planetary bodies

Magnetic dipole fields follow the inverse CUBED law

Solar eclipses allow us to view

the SUN's EM corona in detail

Solar Dynamics

Prominences

Flares, Prominences and CMEs are the result of Electro-Magnetic geometries

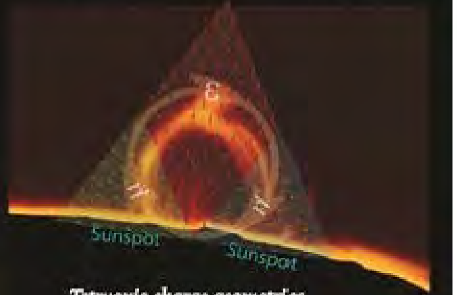

Tetryonic charge geometries explain Solar EM field dynamics

The SUN is a ElectroMagnetohydrodynamo

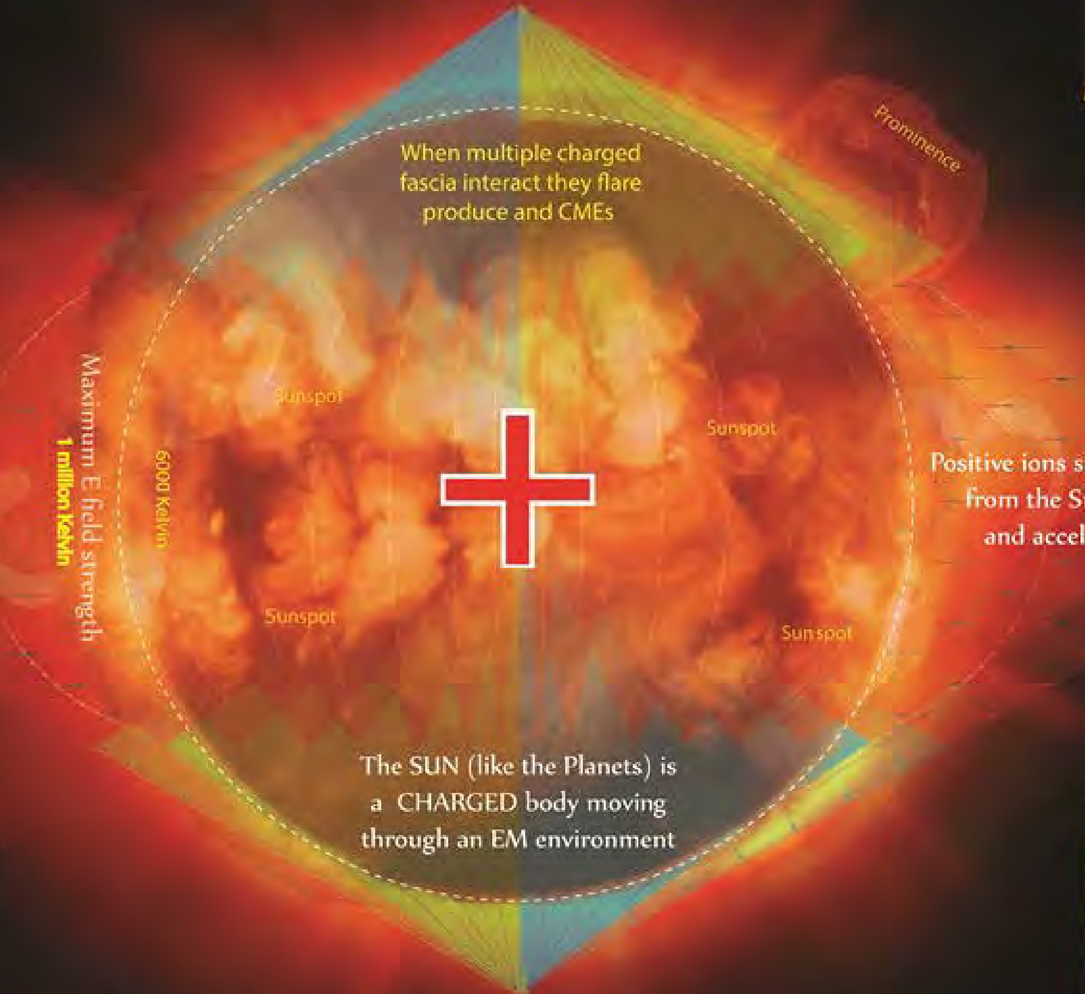

When multiple charged fascia interact they flare produce and CMEs

The SUN (like the Planets) is a CHARGED body moving through an EM environment

Positive ions stream outward from the Sun's surface and accelerate away

Where Solar EM loops breakdown or interact flares result

Solar Flares

The breaking of large inductive EM loops creates CMEs [Coronal Matter Ejections]

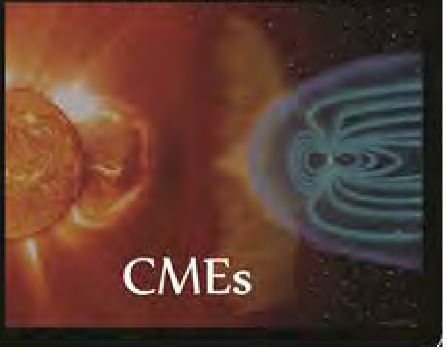

CMEs

The Earth is located at an average distance of about 150 million kilometers [1AU] from the Sun
It takes sunlight travelling at the speed of Light [c] about 8.3 minutes to cross this distance and reach the Earth.

BUT due to the nature of Longitudinal waves emitted by the SUN mechanical momenta impulses [action-at-a-distance] can be transfered from the SUN to the planets nearly instantanteously providing the mechanism for Newton's instataneous pull of Gravity on planetary bodies

Solar Coronal Heating

is a question that has perplexed solar physicists for over half a century.

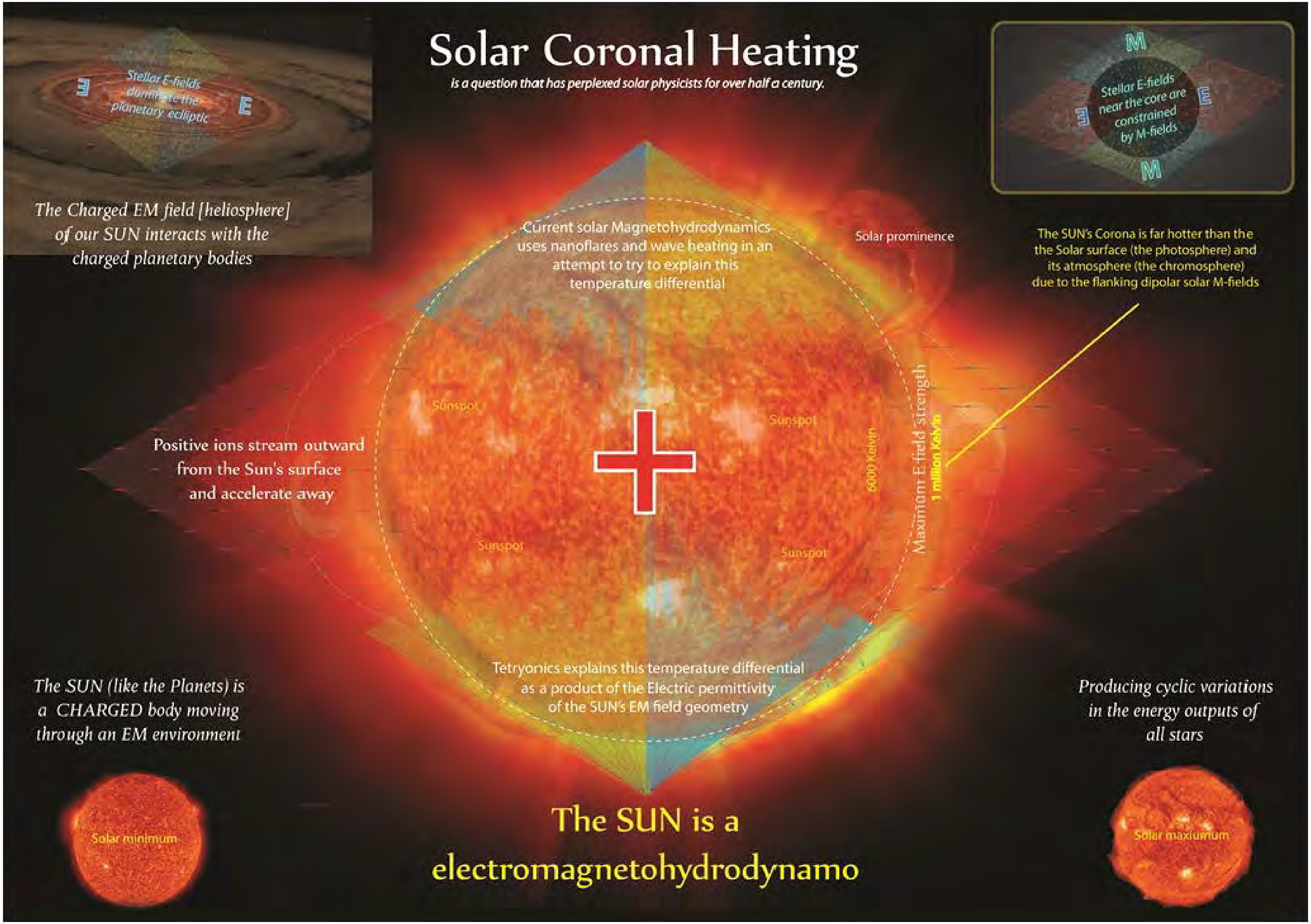
Stellar E-fields dominate the planetary ecliptic
E
E
Stellar E-fields near the core are constrained by M-fields
M
M
E
E
Solar prominence
Sunspot
Sunspot
Sunspot
Sunspot
6000 Kelvin
Maximum E-field strength
1 million Kelvin
Solar minimum
Solar maxiumum

The Charged EM field [heliosphere] of our SUN interacts with the charged planetary bodies

Current solar Magnetohydrodynamics uses nanoflares and wave heating in an attempt to try to explain this temperature differential

The SUN's Corona is far hotter than the the Solar surface (the photosphere) and its atmosphere (the chromosphere) due to the flanking dipolar solar M-fields

Positive ions stream outward from the Sun's surface and accelerate away

Tetryonics explains this temperature differential as a product of the Electric permittivity of the SUN's EM field geometry

The SUN (like the Planets) is a CHARGED body moving through an EM environment

Producing cyclic variations in the energy outputs of all stars

The SUN is a electromagnetohydrodynamo

Solar cycles

On an 11 year cycle, the shape of the corona oscillates from a wide crown about the Sun's equator to a completely closed envelope surrounding the Sun

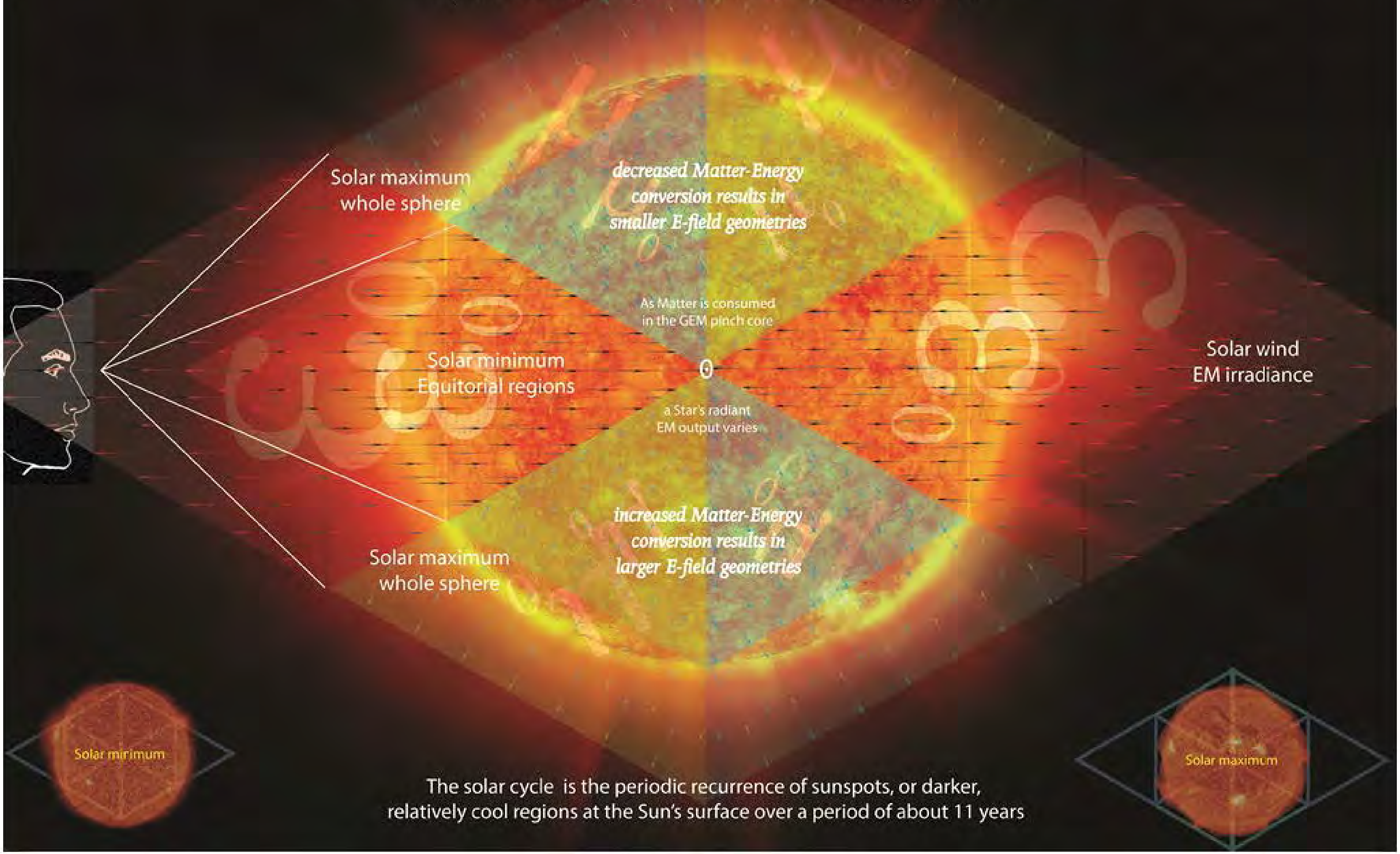
Solar maximum
whole sphere
decreased Matter-Energy conversion results in smaller E-field geometries
As Matter is consumed in the GEM pinch core
Solar minimum
Equitorial regions
0
a Star's radiant EM output varies
Solar wind
EM irradiance
increased Matter-Energy conversion results in larger E-field geometries
Solar maximum
whole sphere
Solar minimum
Solar maximum

The solar cycle is the periodic recurrence of sunspots, or darker, relatively cool regions at the Sun's surface over a period of about 11 years

Tetryonics 72.04 - Solar cycles

Copyright ABRAHAM [2008] - All rights reserved

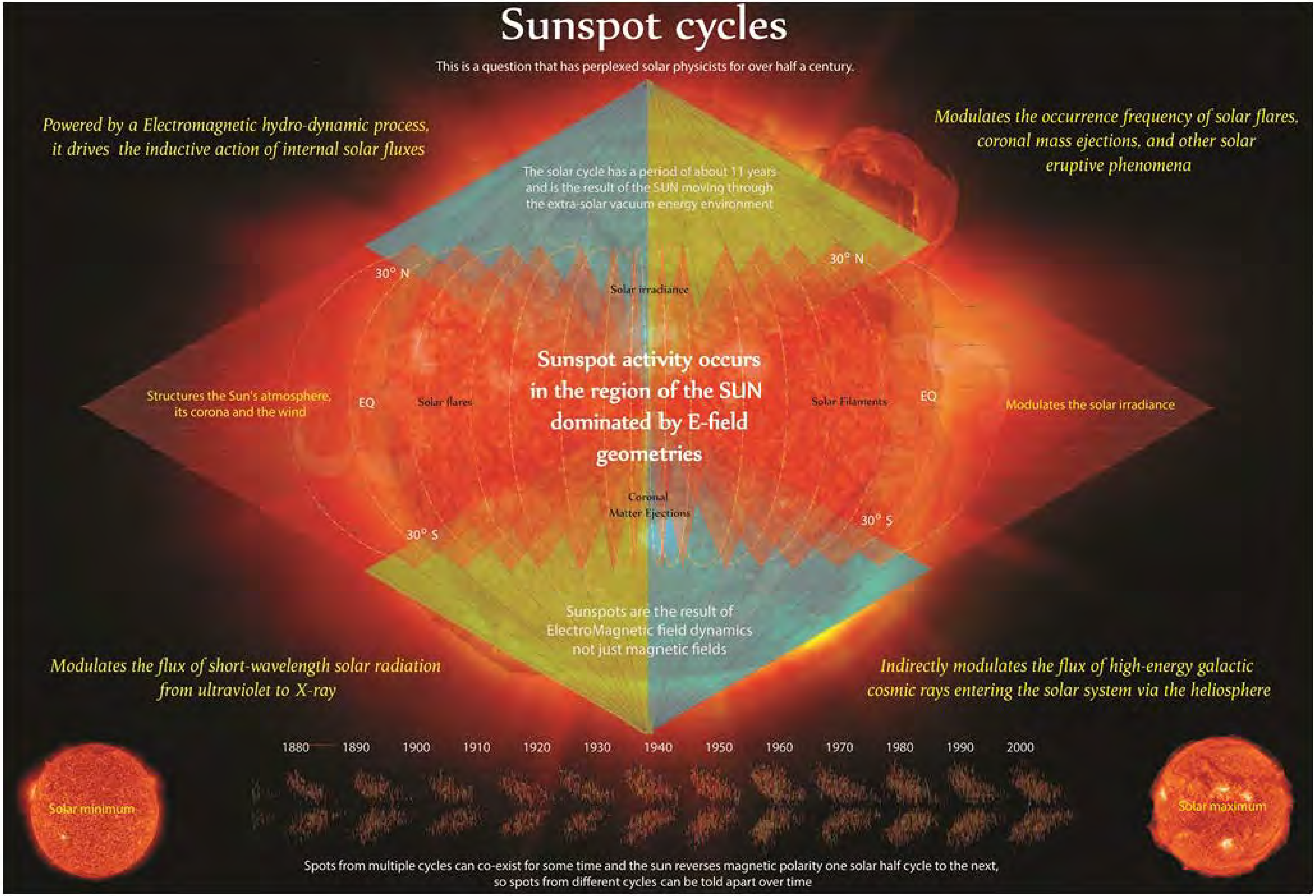

Sunspot cycles

This is a question that has perplexed solar physicists for over half a century.

Powered by a Electromagnetic hydro-dynamic process, it drives the inductive action of internal solar fluxes

Modulates the occurrence frequency of solar flares, coronal mass ejections, and other solar eruptive phenomena

The solar cycle has a period of about 11 years and is the result of the SUN moving through the extra-solar vacuum energy environment

Structures the Sun's atmosphere, its corona and the wind

Sunspot activity occurs in the region of the SUN dominated by E-field geometries

Modulates the solar irradiance

Sunspots are the result of ElectroMagnetic field dynamics not just magnetic fields

Modulates the flux of short-wavelength solar radiation from ultraviolet to X-ray

Indirectly modulates the flux of high-energy galactic cosmic rays entering the solar system via the heliosphere

Spots from multiple cycles can co-exist for some time and the sun reverses magnetic polarity one solar half cycle to the next, so spots from different cycles can be told apart over time

30° N
30° N
Solar irradiance
EQ
Solar flares
Solar Filaments
EQ
Coronal Matter Ejections
30° S
30° S
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
Solar minimum
Solar maximum

Tetryonics 72.05 - Sunspot cycles

Copyright ABRAHAM [2008] - All rights reserved

Solar irradiance

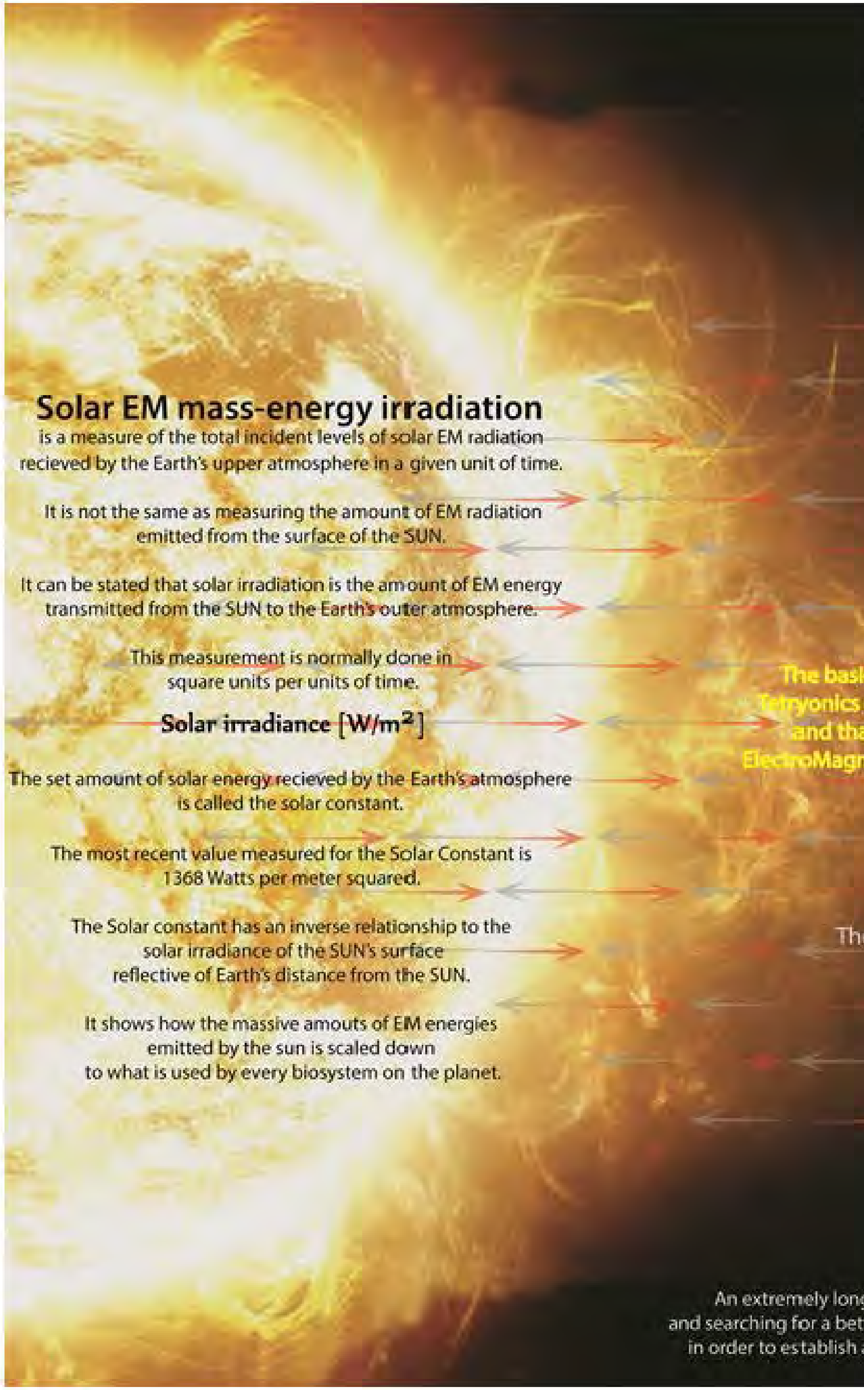

Solar EM mass-energy irradiation

is a measure of the total incident levels of solar EM radiation recieved by the Earth's upper atmosphere in a given unit of time.

It is not the same as measuring the amount of EM radiation emitted from the surface of the SUN.

It can be stated that solar irradiation is the amount of EM energy transmitted from the SUN to the Earth's outer atmosphere.

This measurement is normally done in square units per units of time.

Solar irradiance [W/m^2]

The set amount of solar energy recieved by the Earth's atmosphere is called the solar constant.

The most recent value measured for the Solar Constant is 1368 Watts per meter squared.

The Solar constant has an inverse relationship to the solar irradiance of the SUN's surface reflective of Earth's distance from the SUN.

It shows how the massive amouts of EM energies emitted by the sun is scaled down to what is used by every biosystem on the planet.

The SUN is an ElectroMagnetic pinch reactor

The Sun emits longitudinal EM radiation across most of the electromagnetic spectrum.

The basic causes of the solar variability and solar cycles are still under debate, Tetryonics dictates a EM mechanism for the energies produced at the core of Stars and that the variability of the stellar outputs may be due to changes in the ElectroMagnetic environment that Stars [and their planetary systems] move through

The spectrum of the Sun's solar radiation is close to that of a black body with a temperature of about 5,800 K.

Solar irradiation produces Carbon 14

An extremely long stretch of low activity in recent years has left scientists baffled and searching for a better model of the SUN's interal processes and how to forcast its output in order to establish a better link between the solar cycle and global climate flucuations

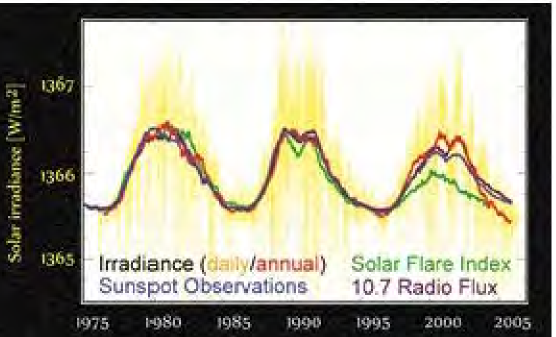

The total solar irradiance (TSI) is the amount of solar radiative energy incident on the Earth's upper atmosphere

Changes in the SUN's EM output can have a direct effect on the weather and electrical activity of the Earth and other planets

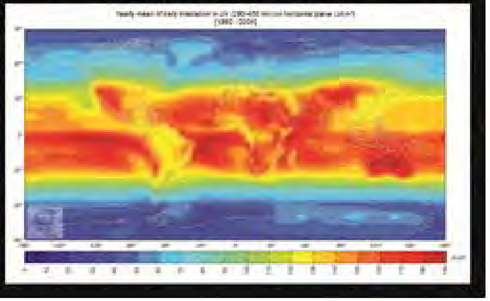

Stellar refraction of EM waves

Any test of General Relativity using photons (or EM waves) will only measure SR interactions
[ie Gravitational red-shifting is a SR effect on EM masses not graviational Matter]

G

The motion of Matter is affected by gravitational attraction and EM interactions

EM

The motion of EM masses is affected by EM interactions and creates refraction

Photons are NOT 'bent' by Gravitational fields they are refracted by propagation velocity changes resulting from their EM interactions with the solar corona

$$\alpha[\text{einstein}] = \frac{8\pi GM}{c^2}$$

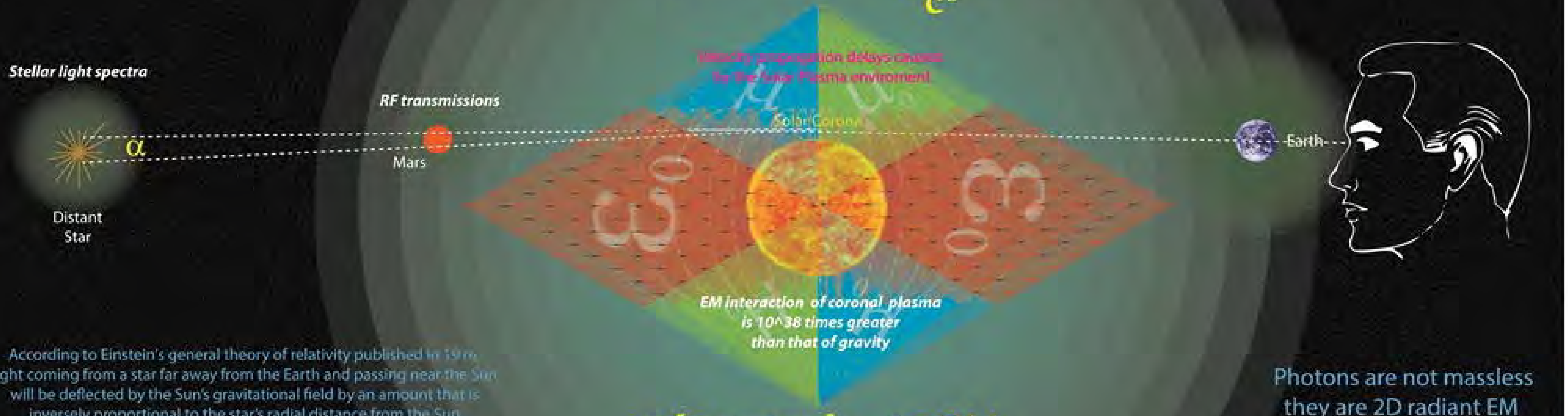

EM interaction of coronal plasma is 10^38 times greater than that of gravity

According to Einstein's general theory of relativity published in 1916, light coming from a star far away from the Earth and passing near the Sun will be deflected by the Sun's gravitational field by an amount that is inversely proportional to the star's radial distance from the Sun

1.745" at the Sun's limb
(this is the result of EM refraction NOT Matter gravitation)

$$\alpha[\text{newton}] = \frac{4\pi GM}{r}$$

Photons are not massless they are 2D radiant EM mass-energy momenta geometries

r

The additional 'bending' of light rays as they pass by a Gravitational body as predicited in Einstein's general relativity is caused by refractive EM interactions not by the force of gravity

4π

4π

c^2

Gravitation in General relativity is Newton's Gravity with SR corrections

Tetryonics 72.07 - Stellar refraction of EM Waves

The Solar Heliosphere

The planetary obits, often hypothesised as being created by the SUN's accreation disk during its formation and 'spun up' to their current orbital momenta can now be visualised as having been formed by the SUN's constant interactive Gravito-Electro-Magnetic fields

The eight planets (plus Pluto and all those other dwarf planets) orbit within a very small volume of the heliosphere (the volume of space dominated by the E field influence of the Sun)

Comets go around the Sun in a highly elliptical orbits.

They can spend hundreds and thousands of years out in the depths of the solar system before they return to the Sun at their perihelion

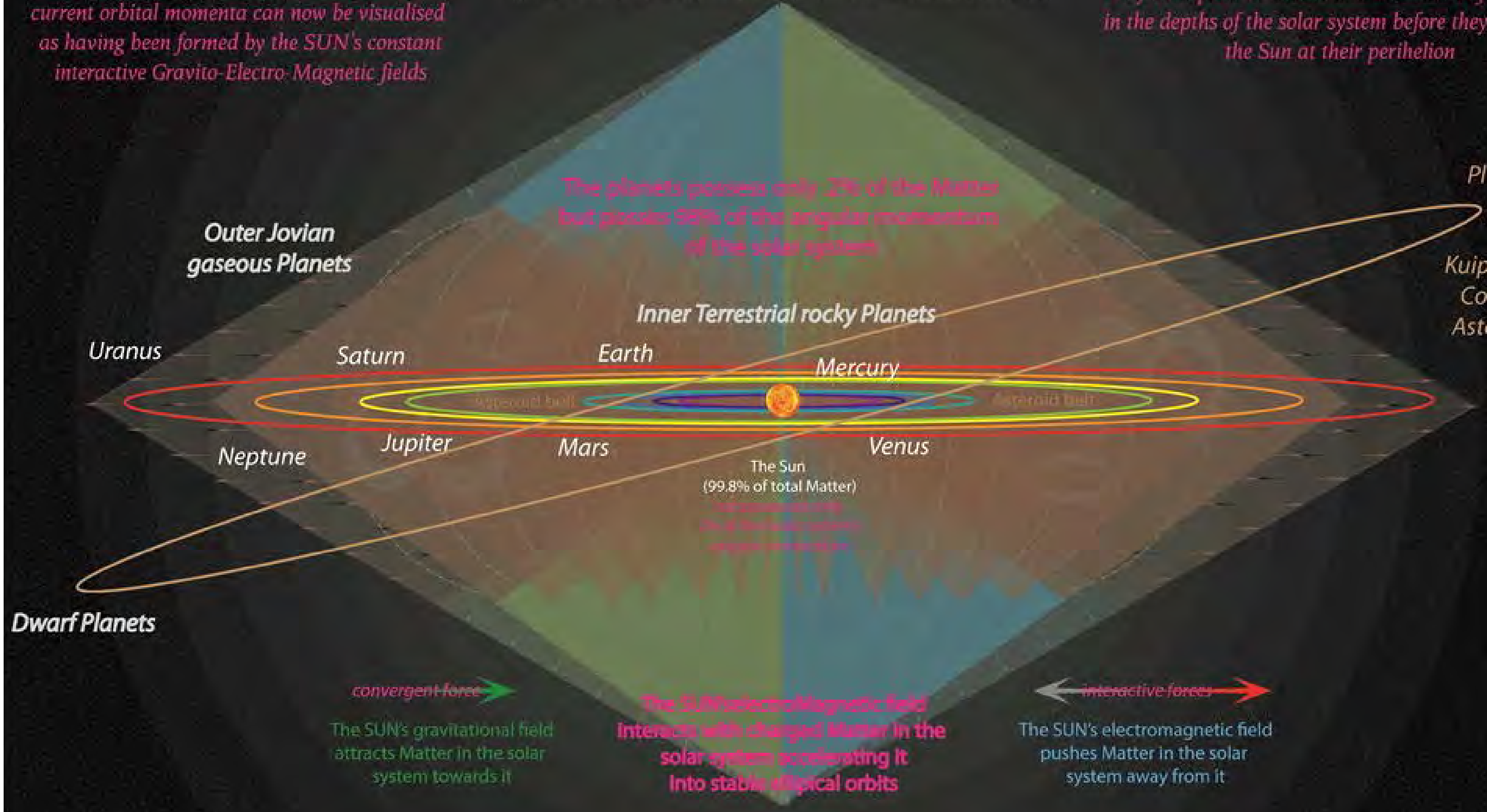
The planets possess only 2% of the Matter
but possess 98% of the angular momentum
of the solar system
Pluto
Outer Jovian gaseous Planets
Kuiper Belt
Comets
Asteroids
Inner Terrestrial rocky Planets
Uranus
Saturn
Earth
Mercury
asteroid belt
asteroid belt
Neptune
Jupiter
Mars
Venus
The Sun
(99.8% of total Matter)
Dwarf Planets
convergent force
interactive forces

The SUN's gravitational field attracts Matter in the solar system towards it

The SUN's electroMagnetic field Interacts with charged Matter in the solar system accelerating it Into stable elliptical orbits

The SUN's electromagnetic field pushes Matter in the solar system away from it

Planetary Orbits are NOT to scale

The SUN is the Solar system's dominant charged electromagnetic mass-Matter body

Planetary solar Orbits, though technically ellipses, are nearly circular

Tetryonics 72.08 - The Solar Heliosphere

Copyright ABRAHAM [2008] - All rights reserved

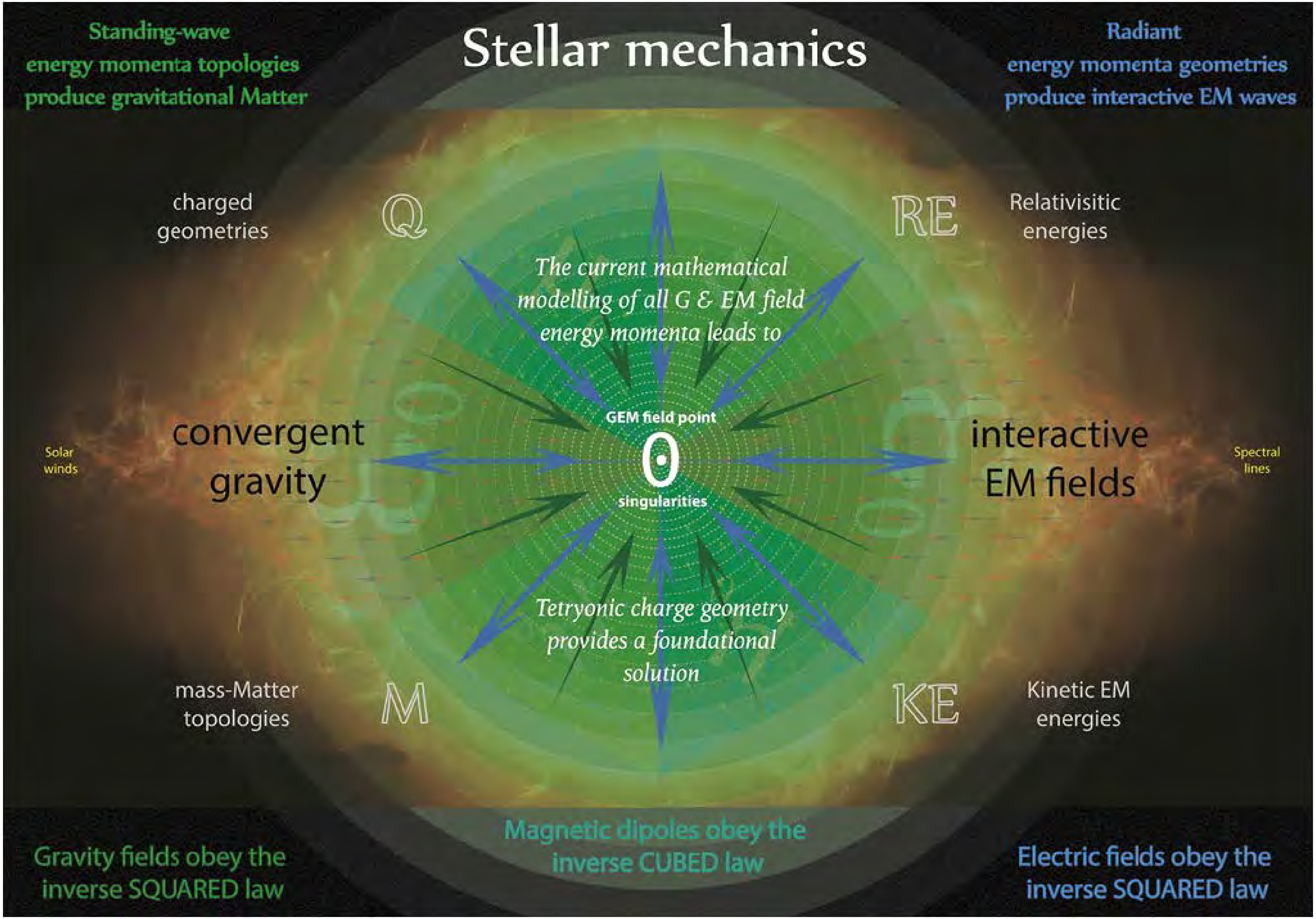

Stellar mechanics

Standing-wave
energy momenta topologies
produce gravitational Matter

Radiant
energy momenta geometries
produce interactive EM waves

charged
geometries

Q

RE

Relativisitic
energies

The current mathematical modelling of all G & EM field energy momenta leads to

Solar
winds

convergent
gravity

GEM field point

0

singularities

interactive
EM fields

Spectral
lines

Tetryonic charge geometry provides a foundational solution

mass-Matter
topologies

M

KE

Kinetic EM
energies

Gravity fields obey the
inverse SQUARED law

Magnetic dipoles obey the
inverse CUBED law

Electric fields obey the
inverse SQUARED law

Tetryonics 73.01 - Stellar mechanics

Anti-Matter Stars

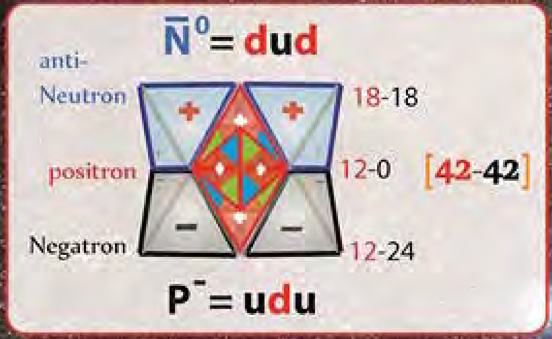
$\bar{N}^0$ = dud
anti-Neutron
18-18
positron
12-0
[42-42]
Negatron
12-24
P^- = udu

anti-Matter stars have
the opposite charge topology
to Matter stars

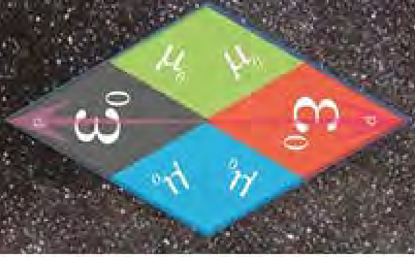
μ_0
μ_0
ε_0
ε_0
μ_0
μ_0

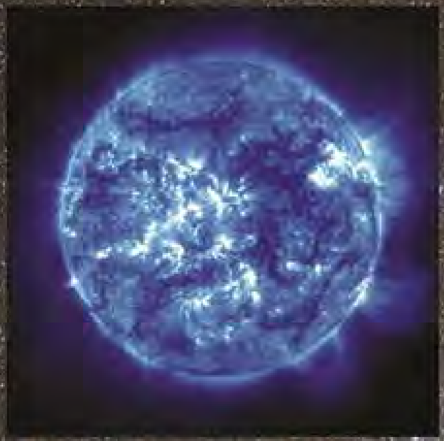

Neutron Stars

Colours used are illustrative of Matter types only

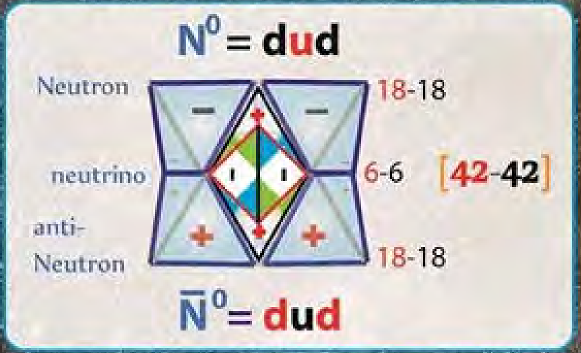
N^0 = dud
Neutron
18-18
neutrino
6-6
[42-42]
anti-Neutron
18-18
$\bar{N}^0$ = dud

All stars types emit Neutral Photons

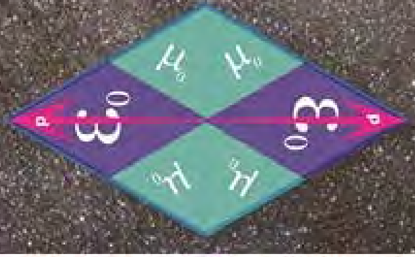
μ_0
μ_0
ε_0
ε_0
μ_0
μ_0

Matter Stars

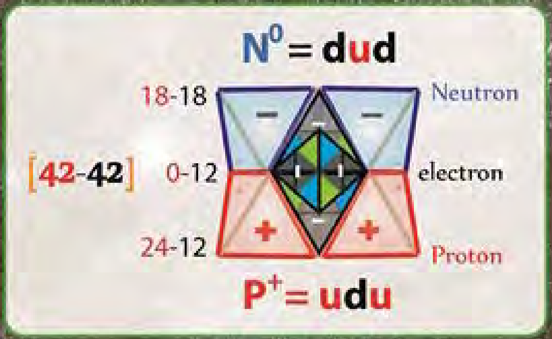
N^0 = dud
18-18
Neutron
[42-42]
0-12
electron
24-12
Proton
P^+ = udu

Matter stars have
the opposite charge topology
to anti-Matter stars

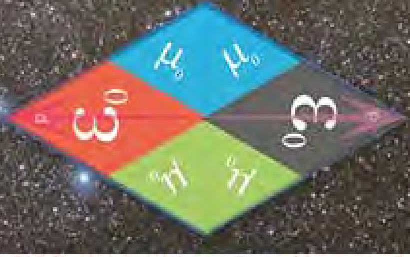
μ_0
μ_0
ε_0
ε_0
μ_0
μ_0

Star classifications

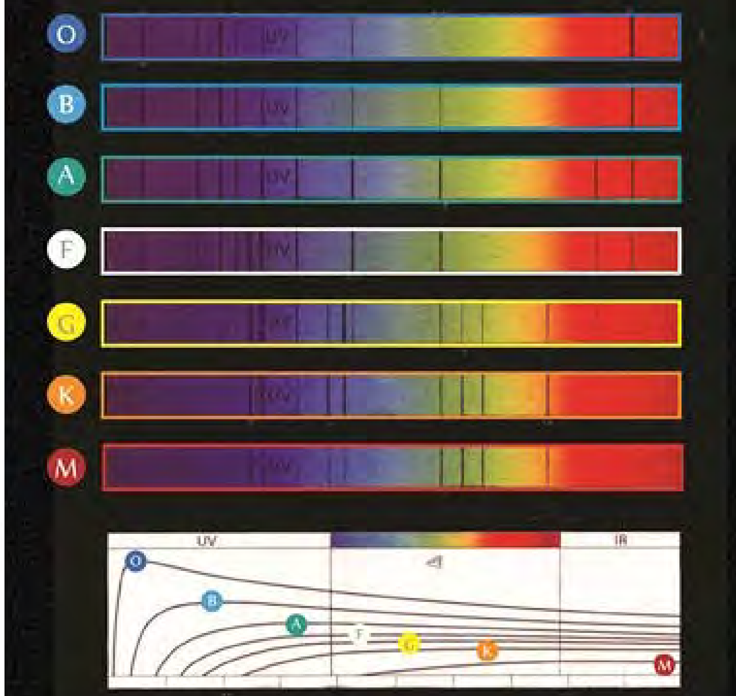

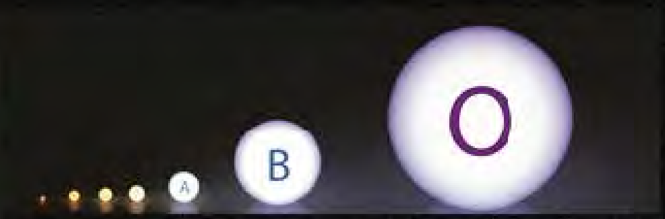

The greater the EM mass-energies radiated in a given volume of space the higher the Temperatures

White dwarves

Average Matter in Small volumes

The greater the voume for a given amount of radiated EM mass-energy the 'cooler' the Frequency of light produced from stellar processes

O

≥ 16 M

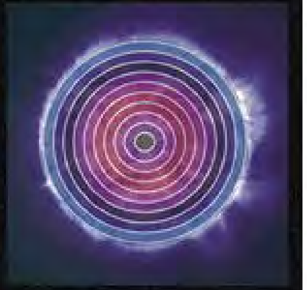

Blue giants

Greater mass-Matter in Larger volumes

B

2.1–16 M

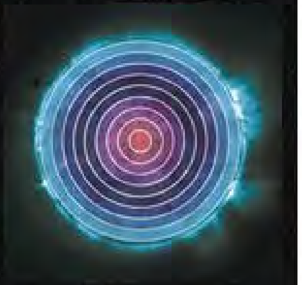

A

1.4–2.1 M

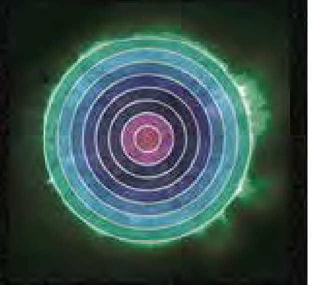

Red giants

Less mass-Matter in Larger volumes

M

≤ 0.45 M

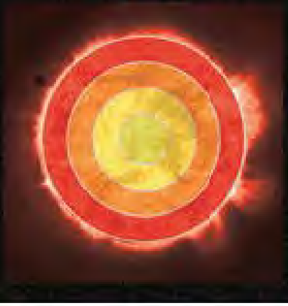

K

0.45–0.8 M

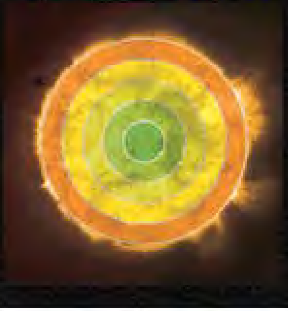

G

0.8–1.04 M

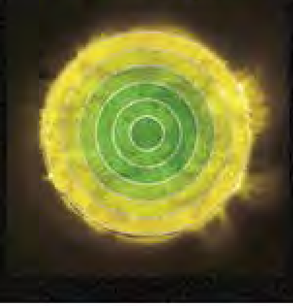

F

1.04–1.4 M

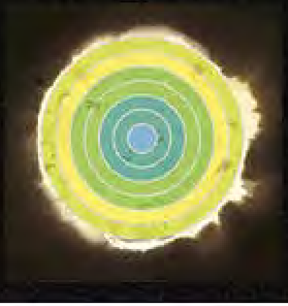

Main sequence

Average mass-Matter in medium volumes

Copyright ABRAHAM [2008] - All rights reserved

Nucleosynthesis

The goal of nucleosynthesis is to understand the vastly differing abundances of the chemical elements and their several isotopes as being a result of natural history

Nucleosynthesis is the process of creating new atomic nuclei from pre-existing nucleons (protons and neutrons)

All Stars have a dominant spectral colour and emit a multitude of spectral lines

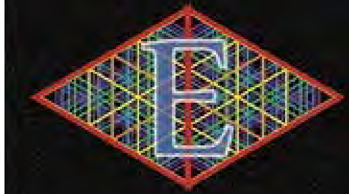

All Stars emit various colour spectra that combine to form spectral lines reflective of the star's energy output

The CNO cycle describing element creation inside Stars is erroneous ....

All spectral lines can be produced by the excitation of stellar nuclei by photons released by GEM pinch Matter-Energy conversion

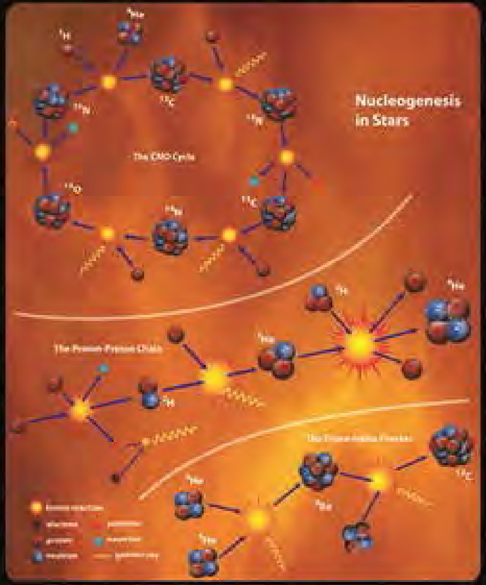

Elements in the Sun

Element	% of total atoms	% of total mass
Hydrogen	91.2	71.0
Helium	8.7	27.1
Oxygen	0.078	0.97
Carbon	0.043	0.40
Nitrogen	0.0088	0.096
Silicon	0.0045	0.099
Magnesium	0.0038	0.076
Neon	0.0035	0.058
Iron	0.030	0.014
Sulfur	0.015	0.040

In 1920, Arthur Eddington was the first to suggest that stars obtained their energy from nuclear fusion of hydrogen to form helium.

The P-P Chain describing the FUSION of elements to release energy is erroneous ....

All Matter-Energy conversion inside Stars is the result of ElectroMagnetic Pinches

Nucleosynthesis problems such as the abundances of observed elements & low mass element synthesis remain open problems for the current theory

Despite intense efforts to create suitable conditions, FUSION as described in nuclear theory, remains an ellusive source of Energy

The CNO cycle and P-P reactions are incorrect theorical models, developed to explain the SUN's GEM pinch dynamics and broad spectral line emissions

Solar Arc lamp spectra

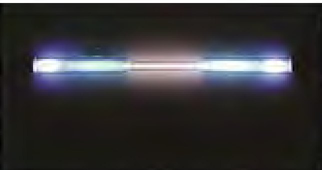

spectral lines are produced by accelerating electrons

Energy-momentum relationship

$$E = mv^2$$

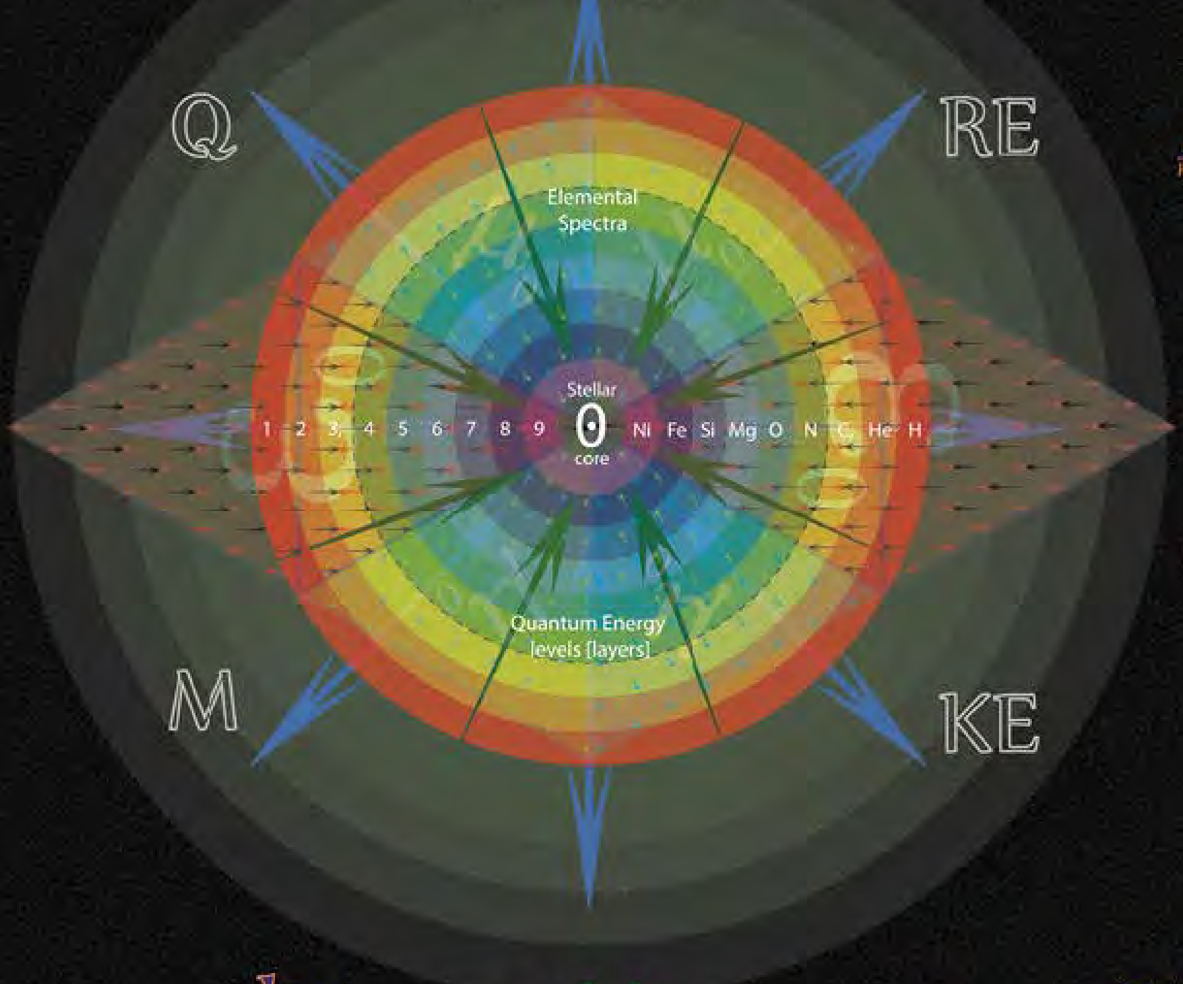
Q
RE
M
KE
Elemental Spectra
Stellar
0
core
1 2 3 4 5 6 7 8 9
Ni Fe Si Mg O N C He H
Quantum Energy levels [layers]

$$\Delta hf = \Delta Mv = \Delta p$$

Spectral line transitions

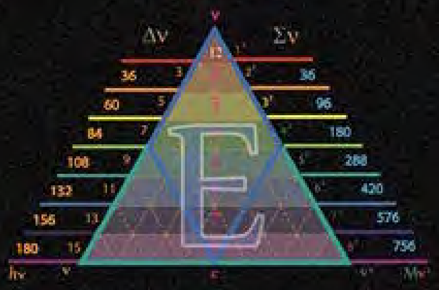
v
Δv
Σv
E
36
60
84
108
132
156
180
3
5
7
9
11
13
15
12
36
96
180
288
420
576
756
hv
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

Spectral lines

KEM
48
108
192
300
432
588
768

Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named

classical physics and relativity united through equilateral geometry

Astrophysicists and astronomers use arc lamps as spectral line reference sources for many elemental emission and absorption lines

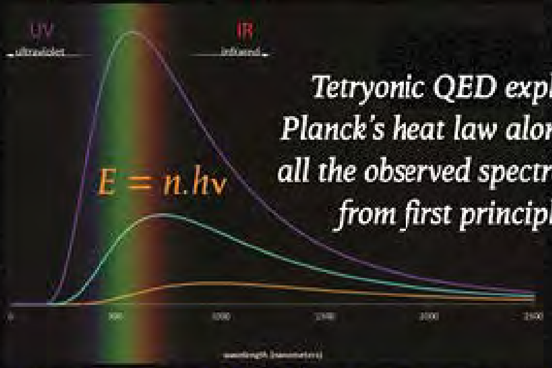

Tetryonic QED explains Planck's heat law along with all the observed spectral lines from first principles

Planck's quantum radiation Law gives a energy distribution that peaks at a certain wavelength, Wien's Law explains the peak shifts to shorter wavelengths for higher temperatures, and the Stefan-Boltzman law explains the growth in the height [the area under] the curve with increasing temperatures

Tetryonics 73.05 - Stellar Arc Lamps

Copyright ABRAHAM [2008] - All rights reserved

Stellar spectral lines

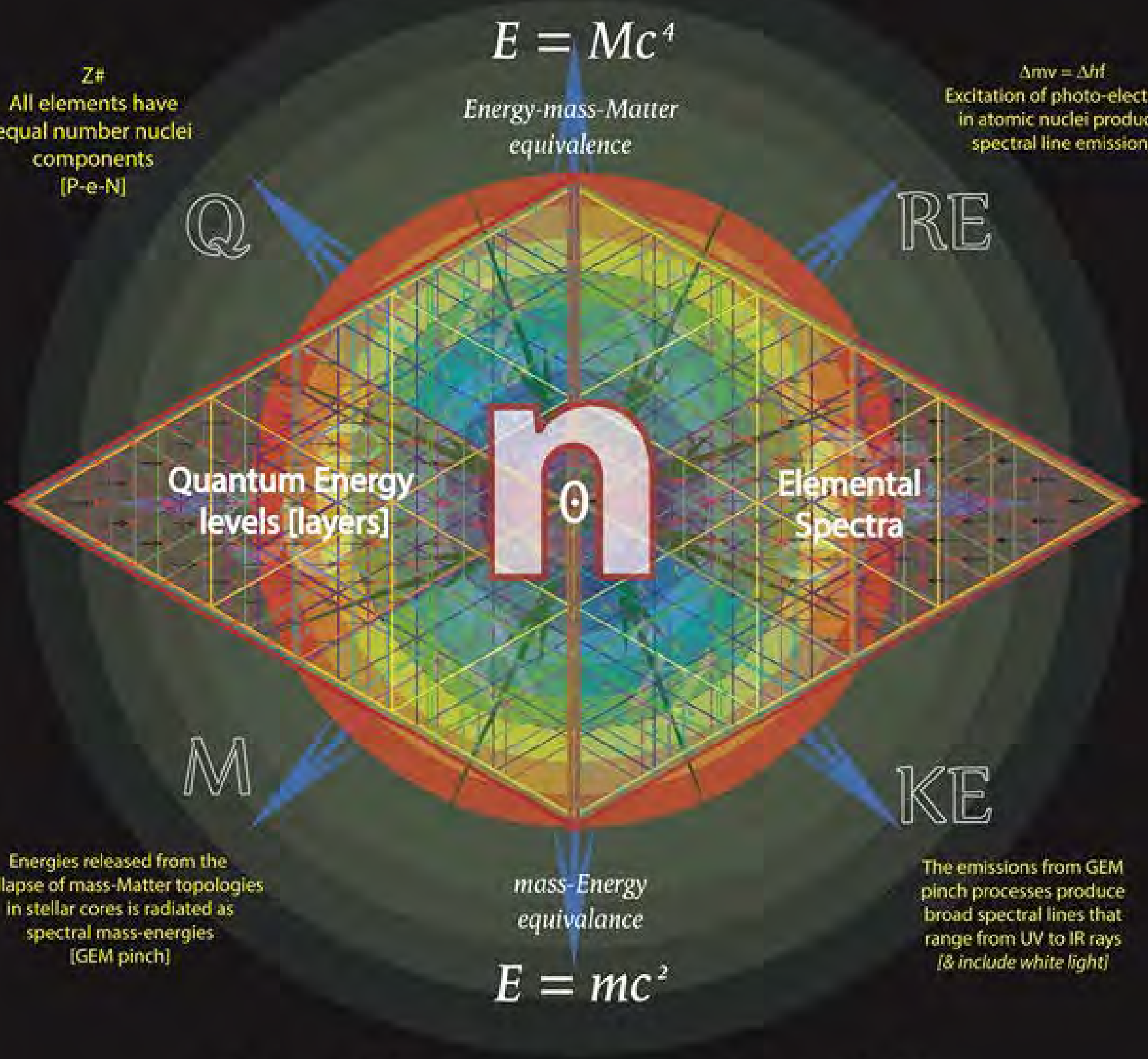
Q
RE
M
KE
Quantum Energy levels [layers]
n
0
Elemental Spectra

$E = Mc^4$

Energy-mass-Matter equivalence

Z#
All elements have equal number nuclei components
[P-e-N]

$\Delta mv = \Delta hf$
Excitation of photo-electrons in atomic nuclei produces spectral line emissions

Energies released from the collapse of mass-Matter topologies in stellar cores is radiated as spectral mass-energies
[GEM pinch]

mass-Energy equivalance

$E = mc^2$

The emissions from GEM pinch processes produce broad spectral lines that range from UV to IR rays
[& include white light]

The spectral lines of all elements can be mimicked by Deuterium nuclei with rasied energy levels undergoing energy transitions

Tetryonic Electrodynamics

Δhv
Quantum differentials

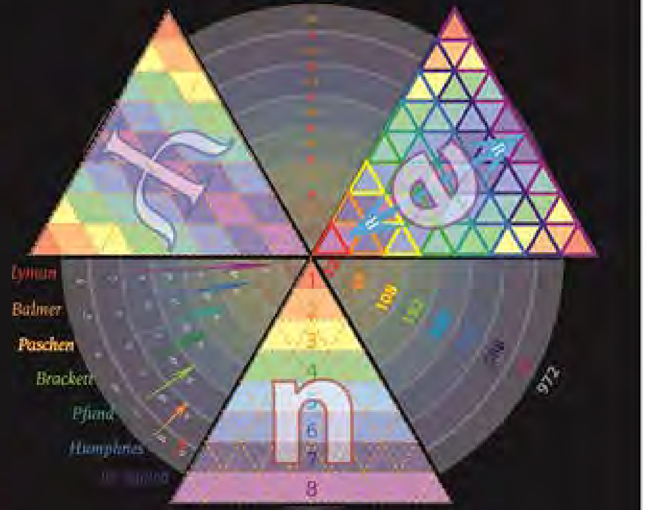
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
1
2
3
4
5
6
7
8
n
972

$hf = \Delta Mv = \Delta p$

spectral lines are produced by accelerating electrons

As energy is 'released' by the collapse of mass-Matter topologies in the stellar core they radiate outward raising the nuclear energies of the surrounding Matter which in turn radiate spectral series photons

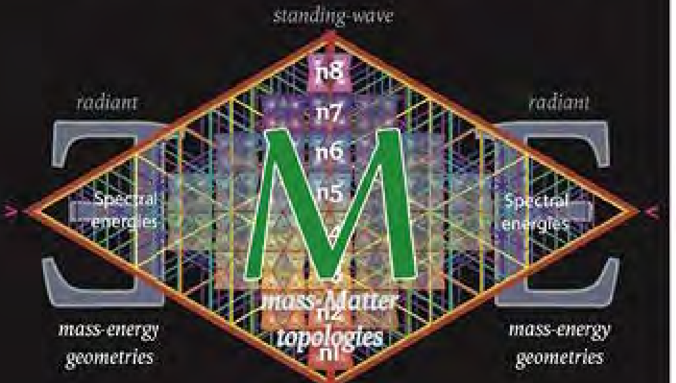
standing-wave
radiant
radiant
n8
n7
n6
n5
M
Spectral energies
Spectral energies
mass-Matter topologies
n2
n1
mass-energy geometries
mass-energy geometries

Joseph von Fraunhofer

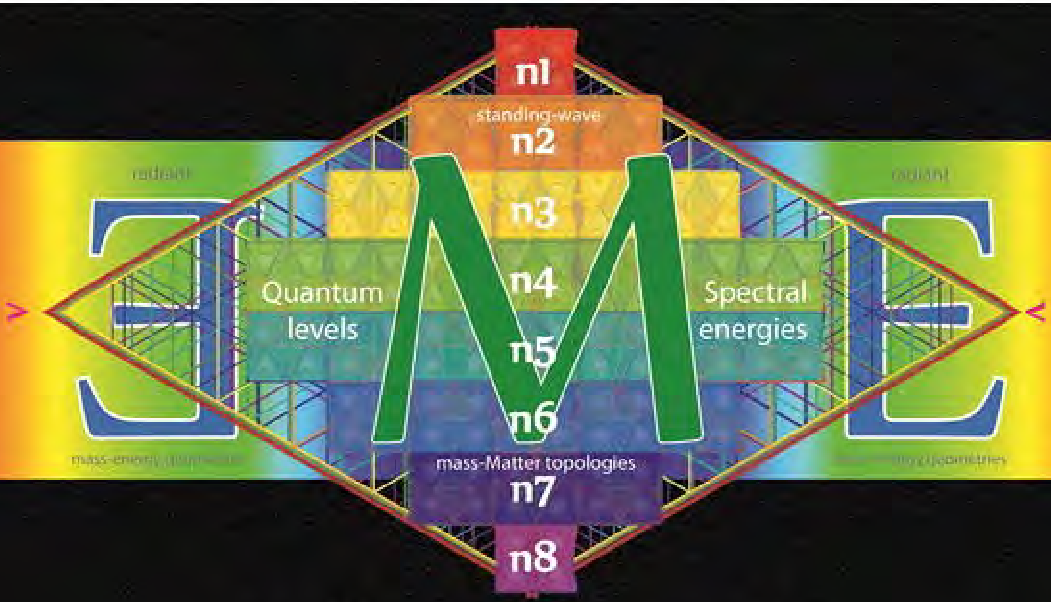
n1
standing-wave
n2
n3
n4
n5
n6
mass-Matter topologies
n7
n8
radiant
radiant
Quantum levels
Spectral energies
mass-energy geometries
mass-energy geometries
M
E
E
>
<

Annie Jump Cannon

(6 March 1787 – 7 June 1826)

(December 11, 1863 – April 13, 1941)

SPECTRAL SERIES

Lyman
Balmer
Paschen
Brackett
Pfund
Humphries

Stellar spectroscopy

Astronomical spectroscopy is the technique of spectroscopy used in astronomy.
The object of study is the spectrum of electromagnetic radiation, including visible light,
which radiates from stars and other celestial objects.

ATOMIC SHELLS

K
L
M
N
O
P

Spectral absorption lines
are often referred to
as Fraunhofer lines

QUANTUM LEVELS

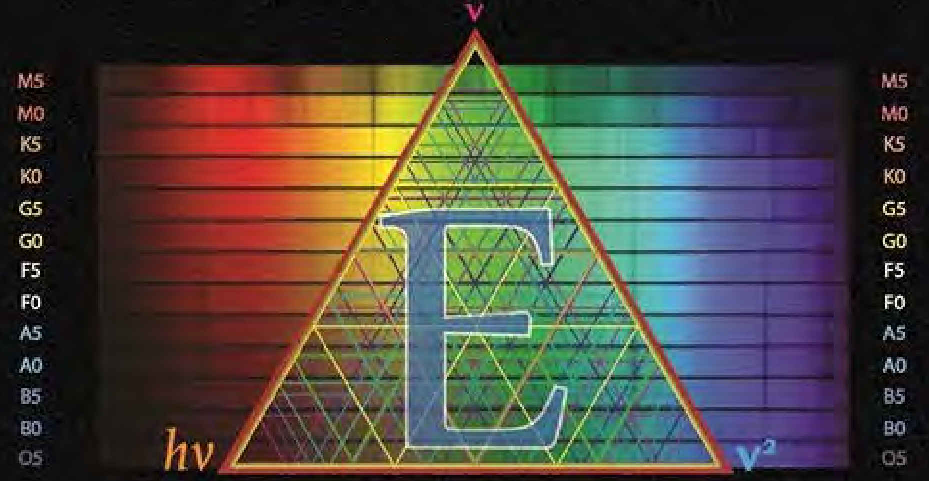
v
M5
M0
K5
K0
G5
G0
F5
F0
A5
A0
B5
B0
O5
hv
E
v²
M5
M0
K5
K0
G5
G0
F5
F0
A5
A0
B5
B0
O5

In 1901 Annie Cannon improved
upon the 1897 work of Harvard
associate Antonia Maury to produce
the Harvard spectral class

STELLAR CLASSIFICATIONS

Spectroscopy can be used to derive many properties of distant stars and galaxies,
such as their chemical composition, and motion through Doppler shift measurements.

Radiant electromagnetic waves propagate

at the speed of light in a vacuum [c]
taking 8.3 minutes to reach Earth

$$E = hf$$

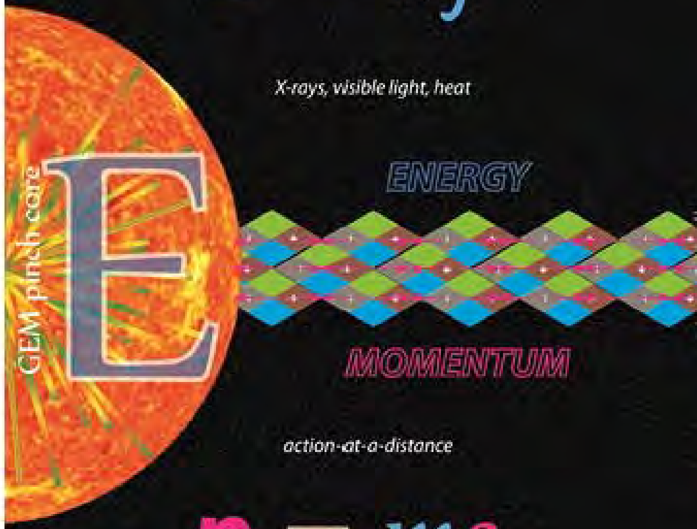
X-rays, visible light, heat
GEM pinch core
E
ENERGY
MOMENTUM
action-at-a-distance

$$p = mc$$

Square root linear momenta present in
all longitudinal EM waveforms from rigid rods

and once established propagate impulses of momenta
near-instantaneously along the longitudinal waves
arriving 8.3 minutes before visible light

mass-ENERGY-Matter

Energy from the SUN

EM energies released by the collapse of standing-wave
Matter topologies in the SUN radiate through space
as mass-energy momenta and interact with Matter

All atoms are comprised of Deuterium nuclei

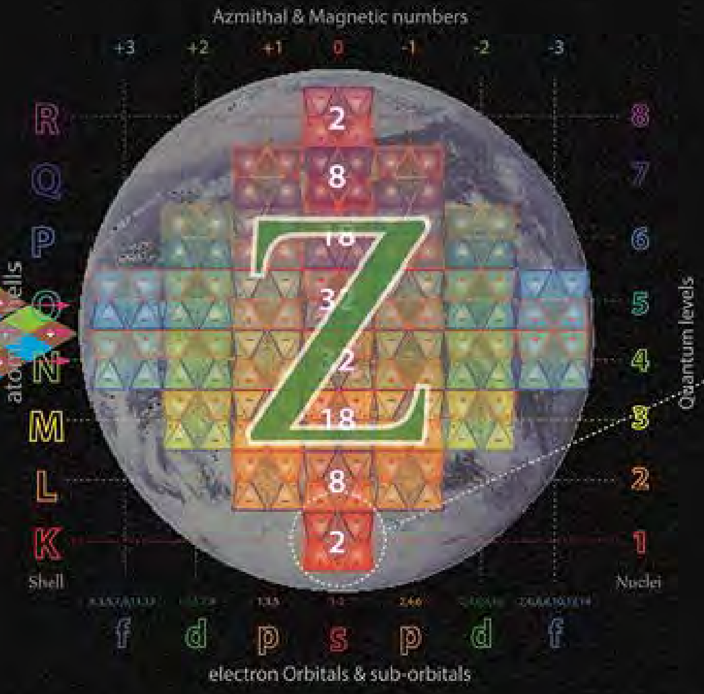
Azmithal & Magnetic numbers
+3
+2
+1
0
-1
-2
-3
R
Q
P
O
N
M
L
K
Shell
atomic shells
8
7
6
5
4
3
2
1
Nuclei
Quantum levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron Orbitals & sub-orbitals

The Earth receives a broad spectrum of energies from the SUN
[with visible light falling into the Balmer spectral series]

$$hf = \underset{366.859\ \text{nm}}{kEM} = \frac{hcR}{n^2}$$

By inductively coupling to the raised quantum
KEM fields of the bound electrons their
energies can be extracted and utilised

Deuterium nuclei are
quantum synchronous convertors
of electromagnetic energies

$$kEM = Mv^2$$

As energy momenta is recieved the energy levels
of the Baryons increases in turn affecting and
raising the KEM field energies of bound electrons

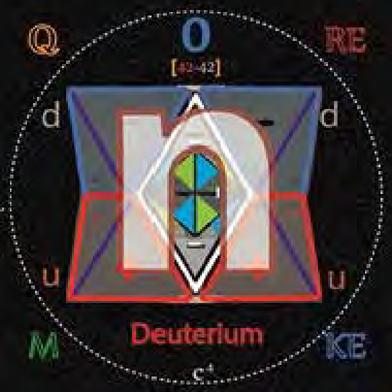
Q
O
RE
[42-42]
d
d
n
u
u
Deuterium
M
c⁴
KE

spectral lines, atomic transitions,
photo-electric effect

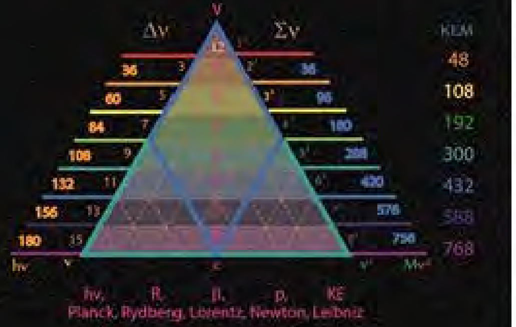
v
Δv
Σv
KEM
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
48
108
192
300
432
588
768
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

changing spectral emissions $\Delta hf = \Delta Mv = \Delta p$ accelerate photo-electrons

Copyright ABRAHAM [2008] - All rights reserved

Stellar Energies

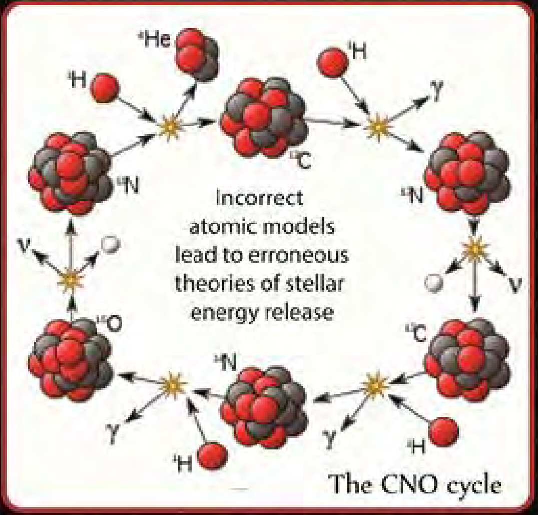
⁴He
¹H
¹H
γ
¹²C
¹⁵N
¹³N
ν
ν
¹⁵O
¹³C
¹⁴N
γ
γ
¹H
¹H
Incorrect atomic models lead to erroneous theories of stellar energy release

The CNO cycle

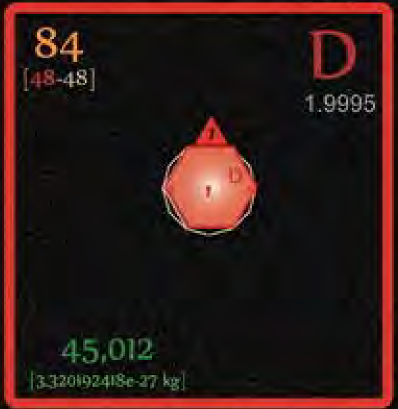
84
[48-48]
D
1.9995
45,012
[3.320192418e-27 kg]

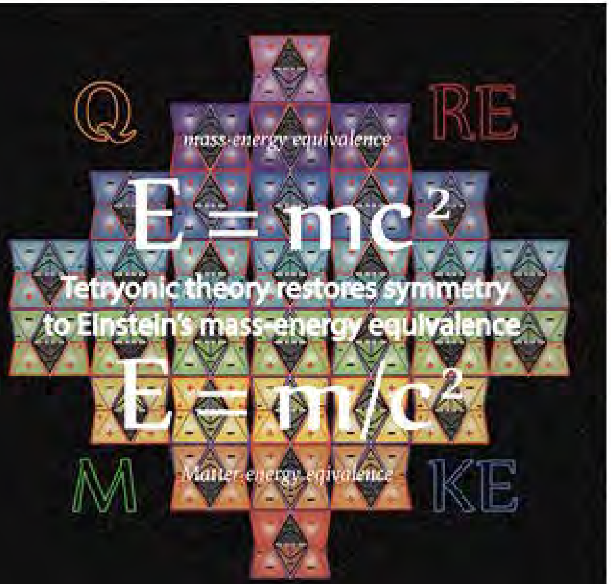
Q
RE
mass-energy equivalence
M
KE
Matter-energy equivalence

$$E = mc^2$$

Tetryonic theory restores symmetry to Einstein's mass-energy equivalence

$$E = m/c^2$$

ENERGY

$$mc^2 = E = Mc^4$$

mass Matter

mass-ENERGY-Matter equivalence shows that a tiny amount of mass-Matter can be converted into a tremendous amount of radiant EM mass-energy

The energy released hypothesised by 'Fusion' is 1/200th that of the Tetryonic collapse of Matter into radiant EM mass-energies

All mass-ENERGY-Matter topologies are modelled in Tetryonic theory and Baryon decay is instigated by neutrinos

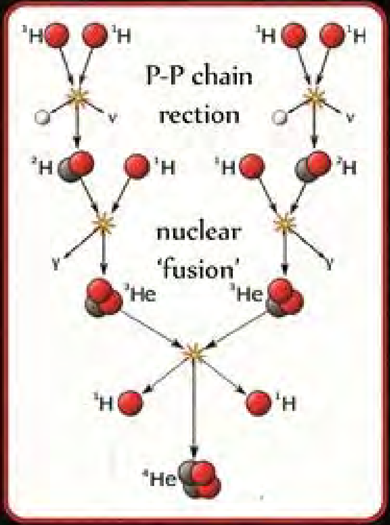
¹H
¹H
¹H
¹H
P-P chain rection
ν
ν
²H
¹H
¹H
²H
nuclear 'fusion'
γ
γ
³He
³He
¹H
¹H
⁴He

Neutrons are NOT Protons with an absorbed electron inside their charge topology

The P-P fusion chain of Hydrogen->Helium is completely erroneous

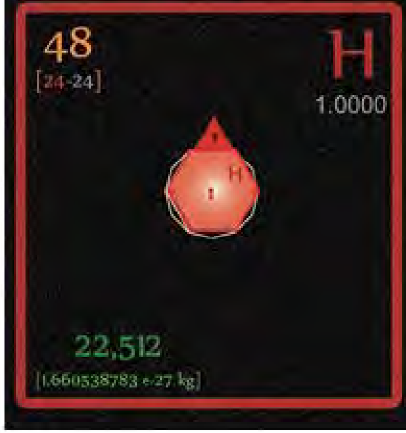
48
[24-24]
H
1.0000
22,512
[1.660538783 e-27 kg]

High pressure tunneling through the Coulomb barrier does NOT occur

The neutrino production of Proton-Proton chain reactions is 1/3 of what is detected

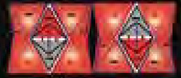

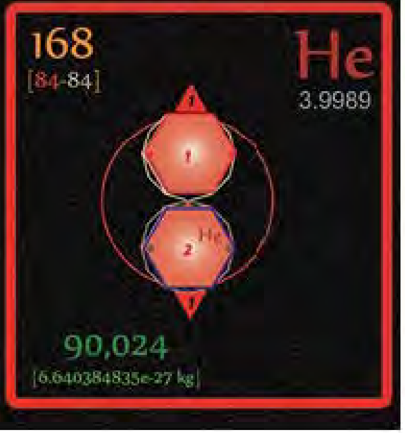
168
[84-84]
He
3.9989
90,024
[6.640384835e-27 kg]

The SUN's energy comes from Tetryonic Matter collapse NOT the fusion of atomic nuclei

Copyright ABRAHAM [2008] - All rights reserved

Stellar Nuclear Fusion

is the process of fusing of light atomic nuclei in order to facilitate energy releases

'COLD fusion'

can be realised through the fragmenting of large atomic nuclei thus releasing stored nuclear [K]EM waveform energies

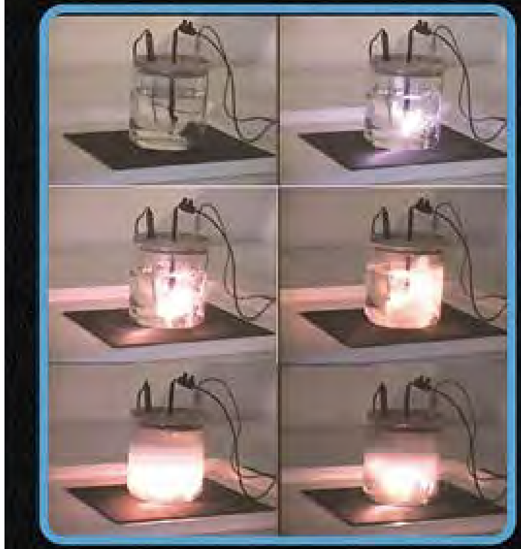

The much-maligned 'COLD fusion' of Palladium is in fact a form of atomic fission that releases the stored chemical [kEM] energies of atomic nuclei and is 12-13% efficient

Long sought after, but never realised, Fusion energies remain an ellusive source of clean limitless energy.

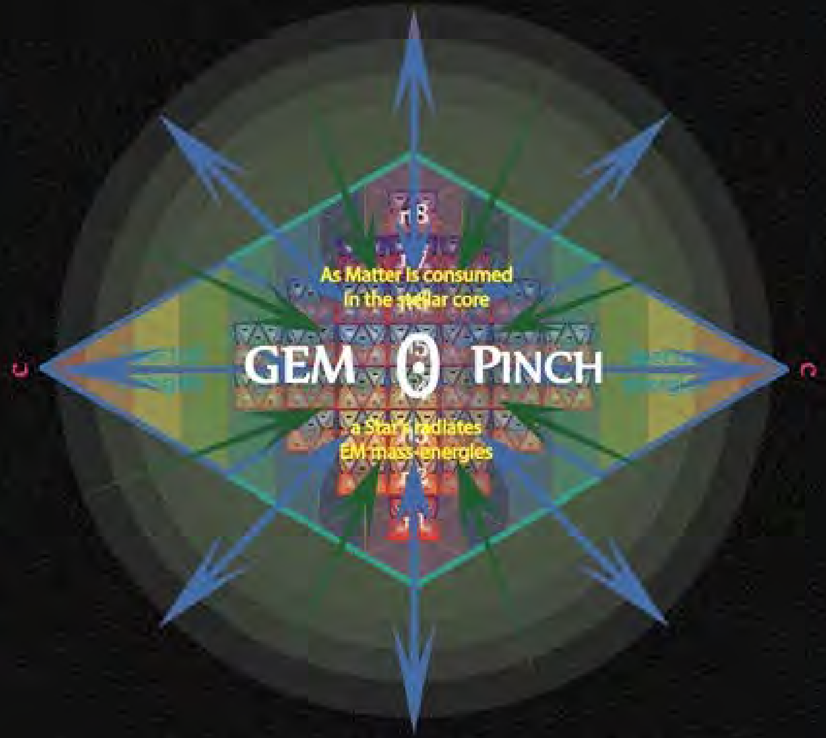
As Matter is consumed in the stellar core
c
GEM
PINCH
c
a Star radiates EM mass-energies

Tetryonics reveals the true mechanics behind the cold 'fusion' observed by Drs. Pons & Fleischmann and corrects the model of stellar processes to reveal new energy sources for the future

'HOT fusion'

is a fallacy, Stars release energy by collapsing standing-wave Matter into radiant EM mass-energy geometries in electromagnetic pinches

The hypothesised 'HOT fusion' of elements in stars is the Tetryonic conversion of mass-Matter within the star into radiant mass-energies requiring an electromagnetic pinch reaction but is 100% efficient

Where Fusion went Wrong

The conversion of standing-wave Matter into radiant mass-energy constitutes a sustainable, virtually unlimited environmentally compatible long-term source of energy that can meet all of Humanity's needs.

The toriodal design of Tokamak reactors does not allow for the formation of plasma pinches at their cores

Tokamak design

Among a small number of approaches, the concept of toroidal magnetic confinement of fusion plasmas has achieved the most impressive scientific and technical progress towards energy release by thermonuclear burn of deuterium-tritium fuels.

The plasma consists of charged nuclei and electrons, which move in tight spirals around the lines of force of strong magnetic fields.

Of the two designs only inertial confinement offered the best chance of creating the conditions required to effectively collapse Matter topologies in order to release the EM energies it contains.

However a far superior approach would be the electromagnetic pinching of Matter (as present in all stellar cores)

The basic problems in attaining useful nuclear fusion conditions are

(1) to heat the gas to these very high temperatures and

(2) to confine a sufficient quantity of the reacting nuclei for a long enough time to permit the release of more energy than is needed to heat and confine the gas.

A subsequent major problem is the capture of this energy and its conversion to electrical energy for work

Inertial confinement fusion produces brief shockwaves, excessive heating and no sustained EM field

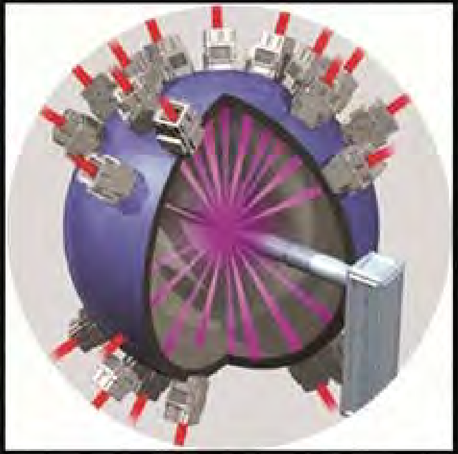

Inertial confinement

In inertia confinement the fuel—tritium or deuterium is contained within a tiny glass sphere that is then bombarded on several sides by a pulsed laser or heavy ion beam.

This causes an implosion of the glass sphere, setting off a thermonuclear reaction that ignites the fuel but offer no self-regualation and relase of radiant mass-energies.

E_{in}

In any useful fusion device, the energy output must exceed the energy required to confine and heat the plasma

E_{out}

ElectroMagnetic Pinch technology

can serve as the gateway to realising new forms of clean, safe, efficient Energy production through the conversion of any Matter into various forms of EM radiation utilising electromagnetic pinch geometries developed from Tetryonic theory

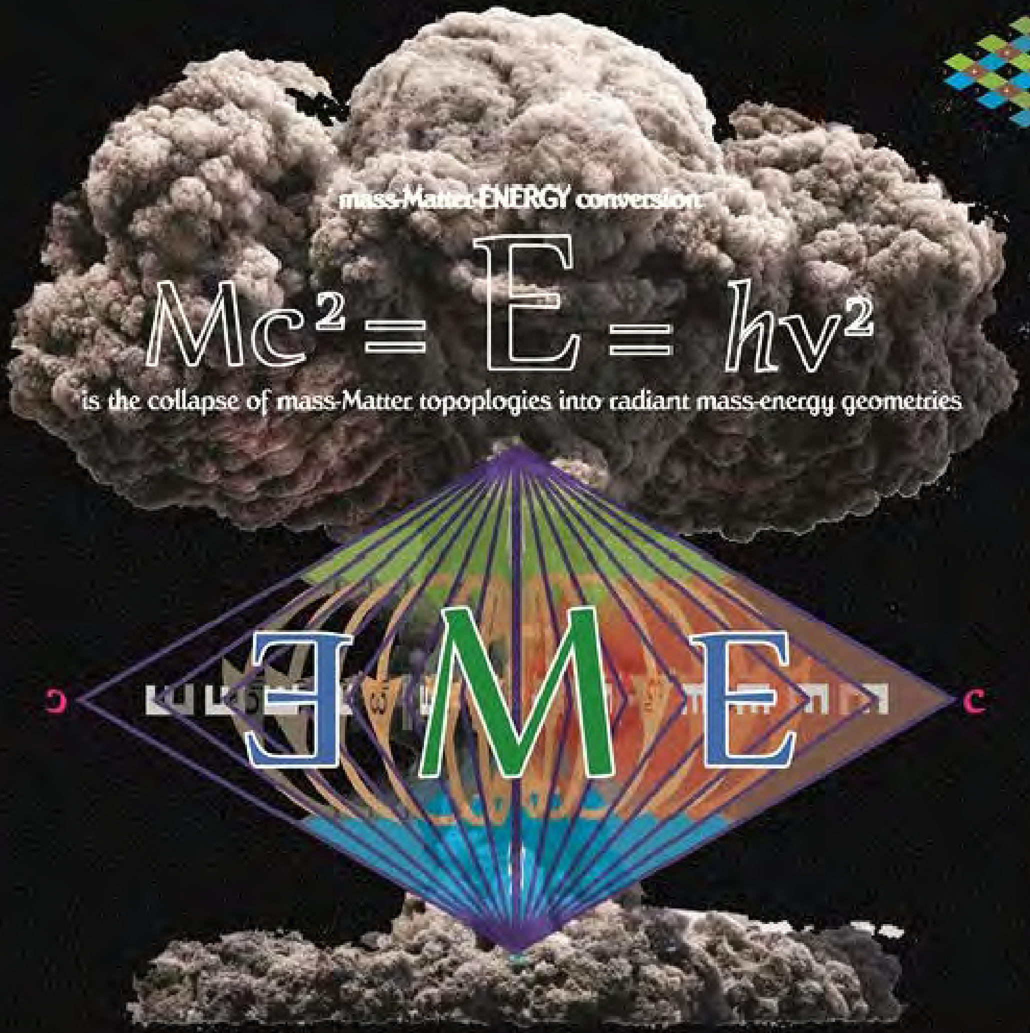

mass-Matter-ENERGY conversion

$$Mc^2 = E = hv^2$$

is the collapse of mass-Matter topoplogies into radiant mass-energy geometries

Charged Matter topologies are standing wave EM mass-energies

Radiant mass-energies form kEM wave geometries

An electromagnetic pulse (sometimes abbreviated EMP) is a burst of electromagnetic radiation that results from an explosion (usually from the detonation of a nuclear weapon) and/or a suddenly fluctuating magnetic field. The resulting rapidly changing electric fields or magnetic fields may couple with electrical/electronic systems to produce damaging current and voltage surges.

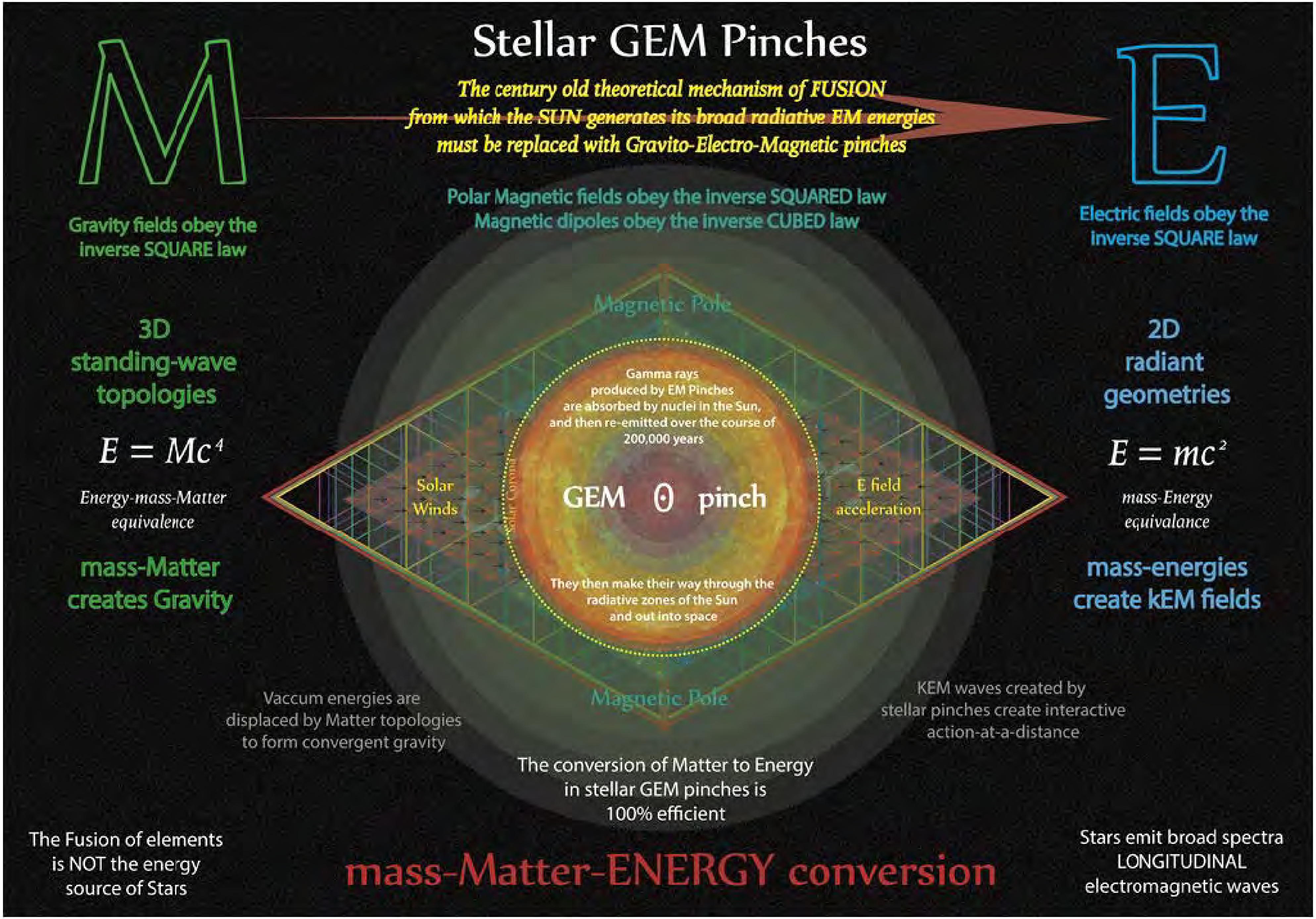

Stellar GEM Pinches

The century old theoretical mechanism of FUSION from which the SUN generates its broad radiative EM energies must be replaced with Gravito-Electro-Magnetic pinches

Gravity fields obey the inverse SQUARE law

Polar Magnetic fields obey the inverse SQUARED law
Magnetic dipoles obey the inverse CUBED law

Electric fields obey the inverse SQUARE law

3D standing-wave topologies

$E = Mc^4$

Energy-mass-Matter equivalence

mass-Matter creates Gravity

2D radiant geometries

$E = mc^2$

mass-Energy equivalance

mass-energies create kEM fields

Gamma rays produced by EM Pinches are absorbed by nuclei in the Sun, and then re-emitted over the course of 200,000 years

They then make their way through the radiative zones of the Sun and out into space

Vaccum energies are displaced by Matter topologies to form convergent gravity

KEM waves created by stellar pinches create interactive action-at-a-distance

The conversion of Matter to Energy in stellar GEM pinches is 100% efficient

The Fusion of elements is NOT the energy source of Stars

mass-Matter-ENERGY conversion

Stars emit broad spectra LONGITUDINAL electromagnetic waves

Tetryonics 73.13 - GEM Pinches

Copyright ABRAHAM [2008] - All rights reserved

GEM Pinch efficiency

GEM Matter-Energy conversion is the 100% efficient dimensional transformation of 3D Matter into 2D mass-energy

M

Theorised Proton-Proton chain reactions of Stellar fusion release <1% of the total Matter as EM energies

3D
mass-Matter
topologies

c^4

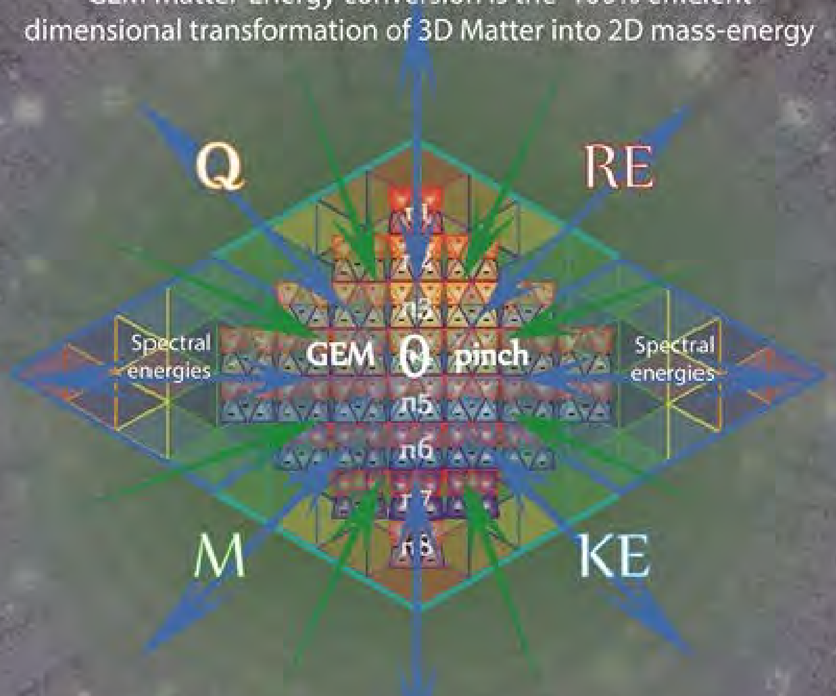
Q
RE
M
KE
Spectral energies
GEM
0
pinch
Spectral energies
n1
n2
n3
n5
n6
n7
n8

The collapse and conversion of 3D mass-Matter topologies into radiant mass-energies releases as much energy as

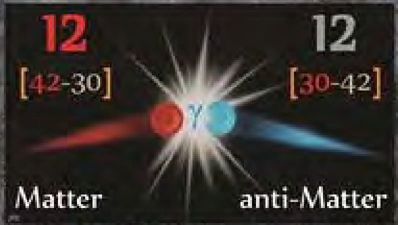
12
[42-30]
12
[30-42]
γ
Matter
anti-Matter

annihilation

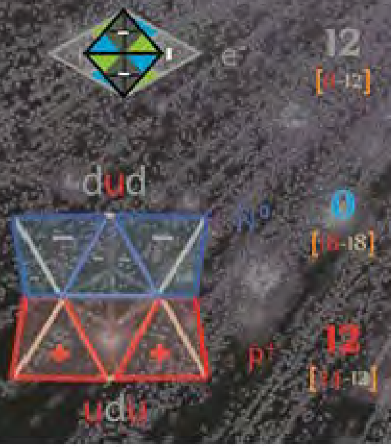
e⁻
12
[0-12]
dud
N⁰
0
[18-18]
p⁺
12
[24-12]
udu

m

Matter to energy conversion by EM pinches releases 100% of the total Tetryonic Matter as radiant EM mass-energies

2D
mass-energy
geometries

c^2

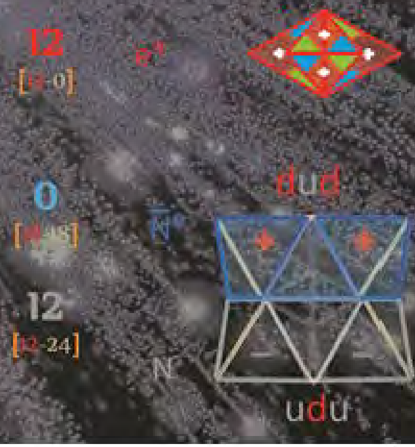
12
[12-0]
e⁺
0
[18-18]
N̄⁰
12
[12-24]
N⁻
dud
udu

Tetryonics 73.14 - Matter-Energy collapse

Copyright ABRAHAM [2008] - All rights reserved

EM Pinch reactors

The technology required to create clean, limitless energy and meet all of humanity's future energy requirements has existed since the 1900's.

The basic steps to build a EM Pinch reactor are

(1) design and build a EM pinch field into which a fuel pellet can be injected

(2) the fuel can be any form of Matter that exhibits both Electric & Magnetic dipoles to facilitate the automatic positioning of the material fuel into the pinch core [the EM fields interact untill they equalise]

(3) enclose the EM Pinch core in a containment casing that can both capture and absorb the radiant EM energies produced by the collapse of Matter at the core and convert these energies into electricity

Th efficiency of the containment vessel along with energy storage and distribution can all be improved through the use of various Tetryonic field geometries to capture and re-circulate all EM energies produced

There are no moving parts to wear out or require extensive maintenance regimes

Upon any loss of Power to the system the EM pinch will be unsustainable and shutdown automatically preventing any possibility of runaway reactions

Photons

$$\frac{2\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

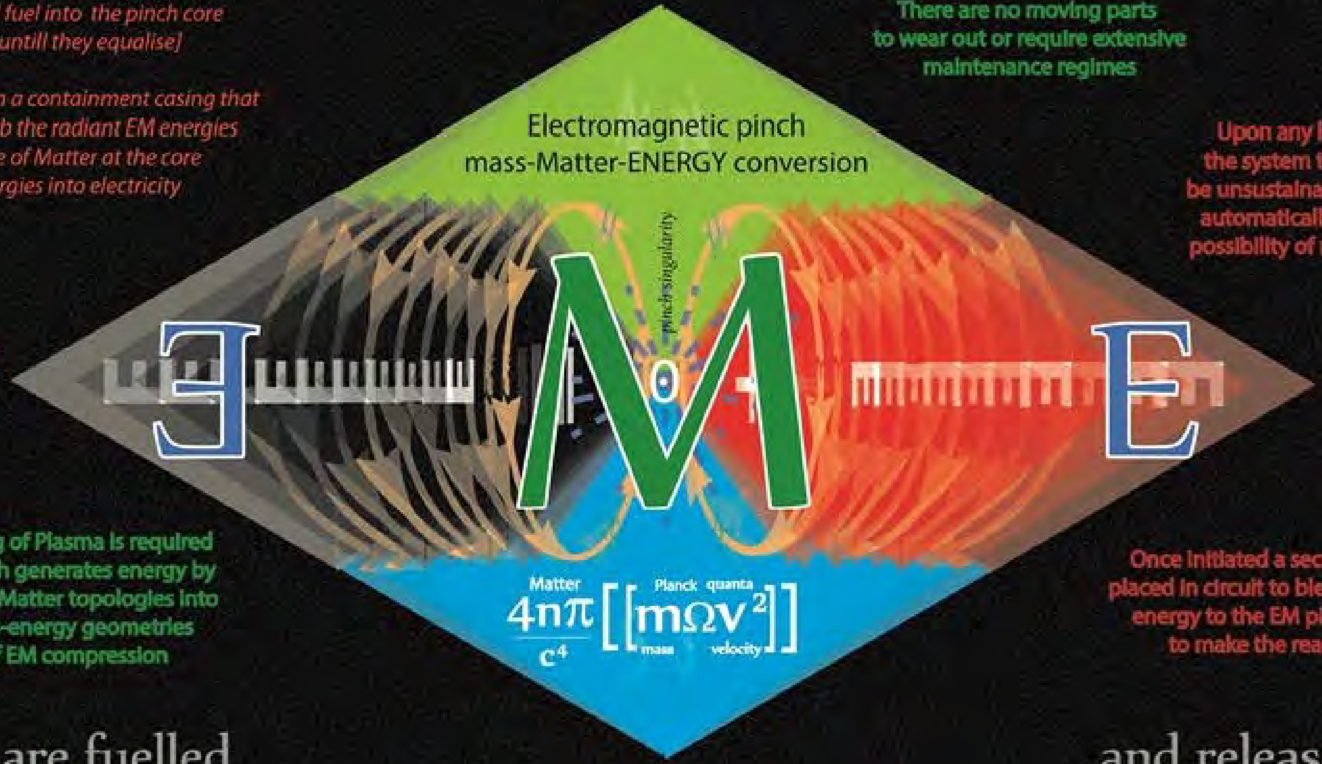

EM waves

$$\frac{2n\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

No preheating of Plasma is required as the EM pinch generates energy by collapsing' 3D Matter topologies into 2D EM mass-energy geometries by way of EM compression

Once initiated a secondary 'circuit' could be placed in circuit to bleed off and return suficient energy to the EM pinch field circuit in order to make the reaction self-sustaining

EM pinches are fuelled

by any form of Matter

The electromagnetic pinching of Matter (as utilised in Nature in all stellar cores) produces only EM radiation [light & heat] as a by-product - there are no harmful wastes - as the Matter is converted with 100% efficiency

and release mass-energy

as radiative KEM fields

Tetryonics 73.15 - The EM Pinch reactor

EM Pinch reactor containment vessel

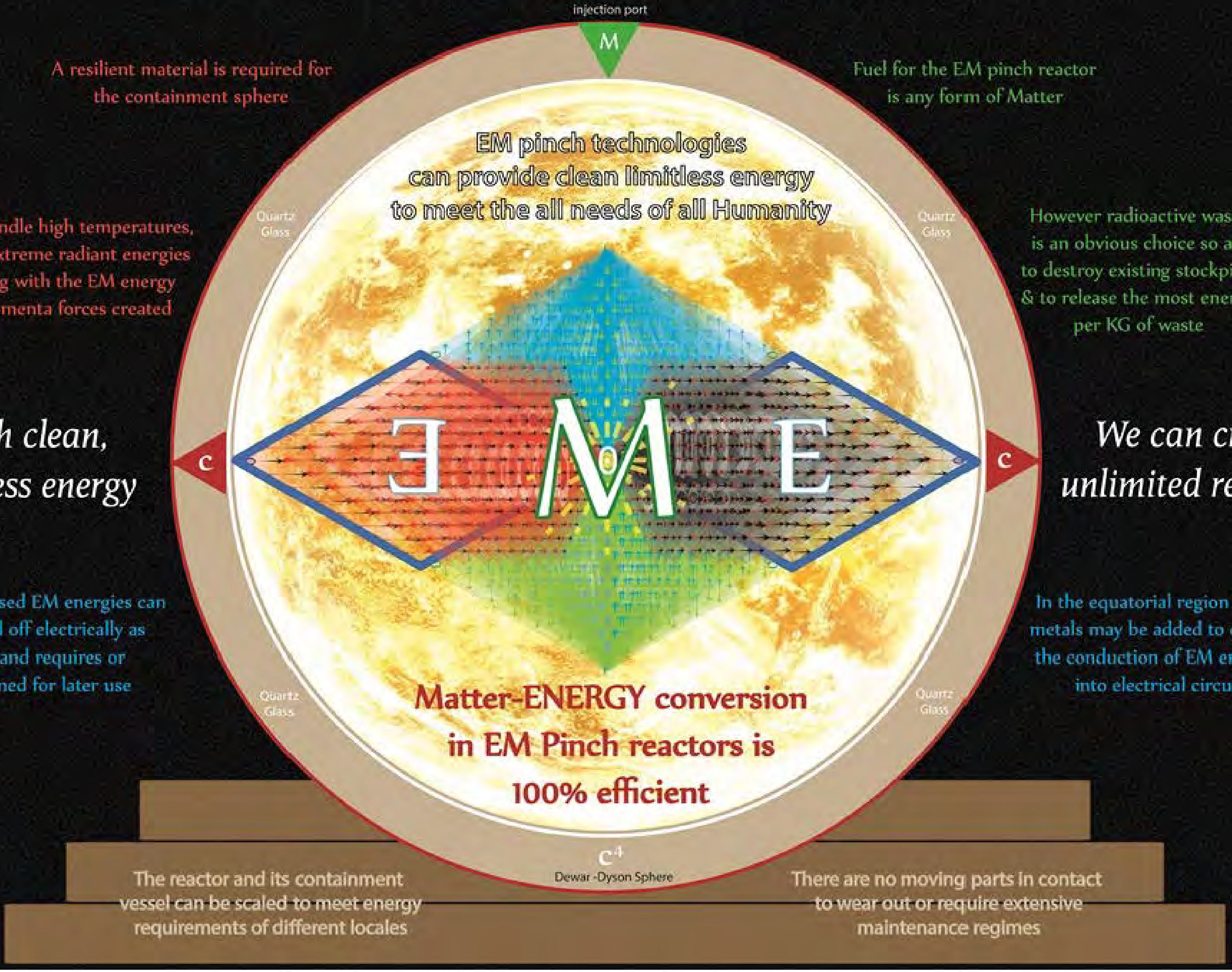

A resilient material is required for the containment sphere

Fuel for the EM pinch reactor is any form of Matter

EM pinch technologies can provide clean limitless energy to meet the all needs of all Humanity

To handle high temperatures, the extreme radiant energies along with the EM energy momenta forces created

However radioactive waste is an obvious choice so as to destroy existing stockpiles & to release the most energy per KG of waste

With clean, limitless energy

We can create unlimited resources

The released EM energies can be bled off electrically as demand requires or confined for later use

In the equatorial region doping metals may be added to enhance the conduction of EM energies into electrical circuits

Matter-ENERGY conversion in EM Pinch reactors is 100% efficient

The reactor and its containment vessel can be scaled to meet energy requirements of different locales

There are no moving parts in contact to wear out or require extensive maintenance regimes

Tetryonics 73.16 - EM Pinch reactor containment

Copyright ABRAHAM [2008] - All rights reserved

The Earth has finite resources and an ever increasing human population puts pressure on its ability to supply resources to meet our needs

Unlimited material Resources

With clean, limitless energy sources

Tetryonics offers a solution to human problems that have plagued civilization from its beginnings leading to wars, disease and famines

Unlimited resources can be created

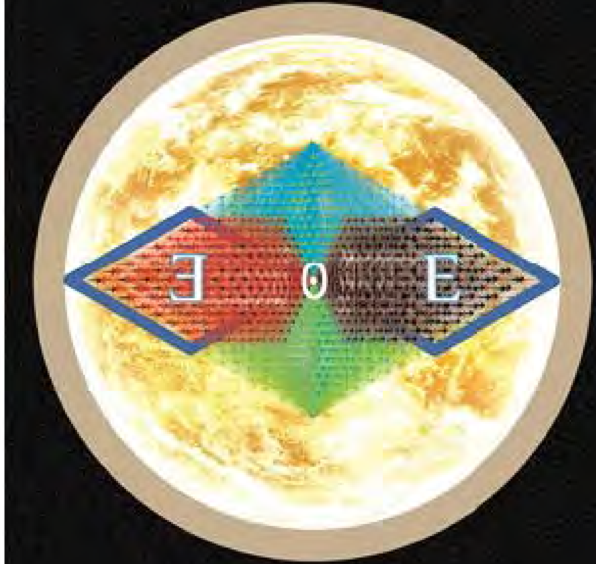

mass-energy geometries
cannot be created or destroyed

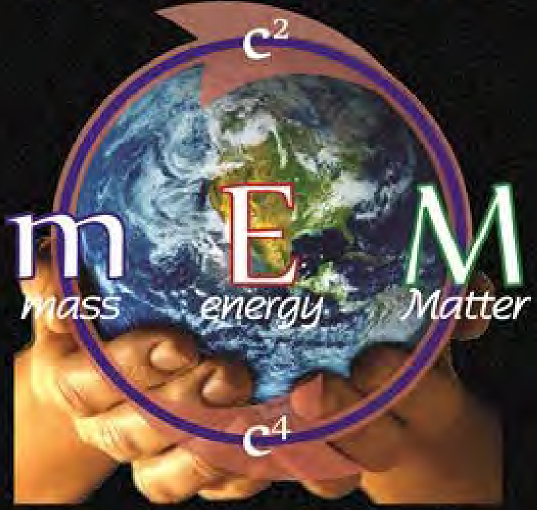

mass-Matter topologies
can be created and destroyed

Resources

Health

With this new understanding of the charged mass-energy geometries of Matter excess energies generated from the destruction of Matter topologies can be reformed into any desired element or compound, and all harmful waste can be removed from our environment

Food

Water

Tetryonics 73.17 - Unlimited resources

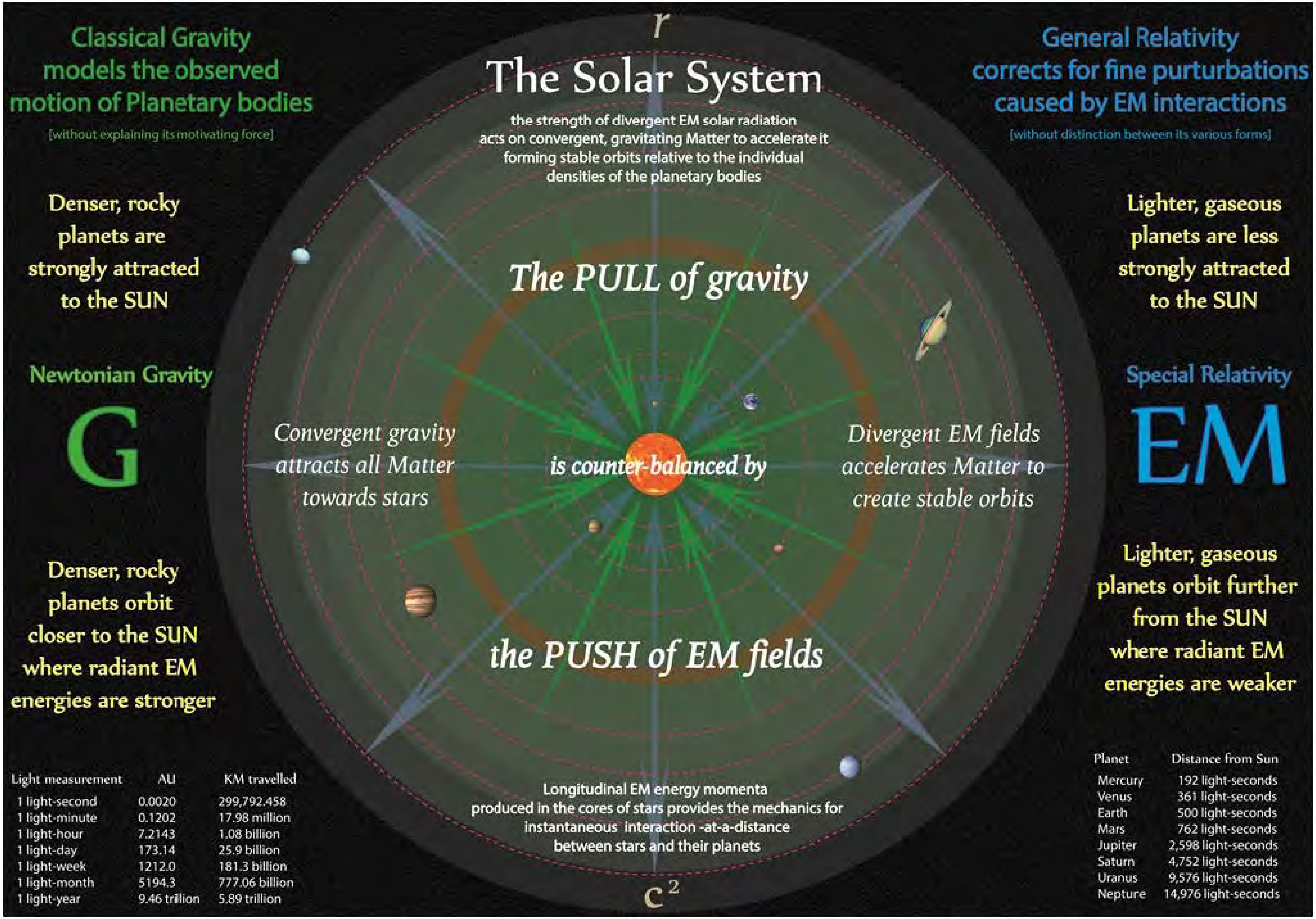

The Solar System

the strength of divergent EM solar radiation acts on convergent, gravitating Matter to accelerate it forming stable orbits relative to the individual densities of the planetary bodies

Classical Gravity models the observed motion of Planetary bodies

[without explaining its motivating force]

General Relativity corrects for fine purturbations caused by EM interactions

[without distinction between its various forms]

Denser, rocky planets are strongly attracted to the SUN

Lighter, gaseous planets are less strongly attracted to the SUN

Newtonian Gravity

G

Special Relativity

EM

The PULL of gravity

Convergent gravity attracts all Matter towards stars

is counter-balanced by

Divergent EM fields accelerates Matter to create stable orbits

the PUSH of EM fields

Denser, rocky planets orbit closer to the SUN where radiant EM energies are stronger

Lighter, gaseous planets orbit further from the SUN where radiant EM energies are weaker

Longitudinal EM energy momenta produced in the cores of stars provides the mechanics for instantaneous interaction -at-a-distance between stars and their planets

Light measurement	AU	KM travelled
1 light-second	0.0020	299,792,458
1 light-minute	0.1202	17.98 million
1 light-hour	7.2143	1.08 billion
1 light-day	173.14	25.9 billion
1 light-week	1212.0	181.3 billion
1 light-month	5194.3	777.06 billion
1 light-year	9.46 trillion	5.89 trillion

Planet	Distance from Sun
Mercury	192 light-seconds
Venus	361 light-seconds
Earth	500 light-seconds
Mars	762 light-seconds
Jupiter	2,598 light-seconds
Saturn	4,752 light-seconds
Uranus	9,576 light-seconds
Nepture	14,976 light-seconds

Tetryonics 74.01 - The Solar System

Kepler's Laws

Kepler's three laws of planetary motion can be described as follows:

The path of the planets about the sun is elliptical in shape,
with the center of the sun being located at one focus.
(The Law of Ellipses)

An imaginary line drawn from the center of the sun to the center of
the planet will sweep out equal areas in equal intervals of time.
(The Law of Equal Areas)

The ratio of the squares of the periods of any two planets is equal to
the ratio of the cubes of their average distances from the sun.
(The Law of Harmonies)

$$T^2 = k\,a^3$$

The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

Gravity attracts Matter towards stars

The divergent energy momenta of kEM fields create an acceleration FORCE on material bodies

The orbit of every planet is an ellipse
with the Sun at one of the two foci.

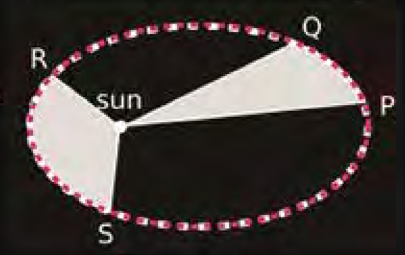

A line joining a planet and the Sun sweeps
out equal areas during equal intervals of time.

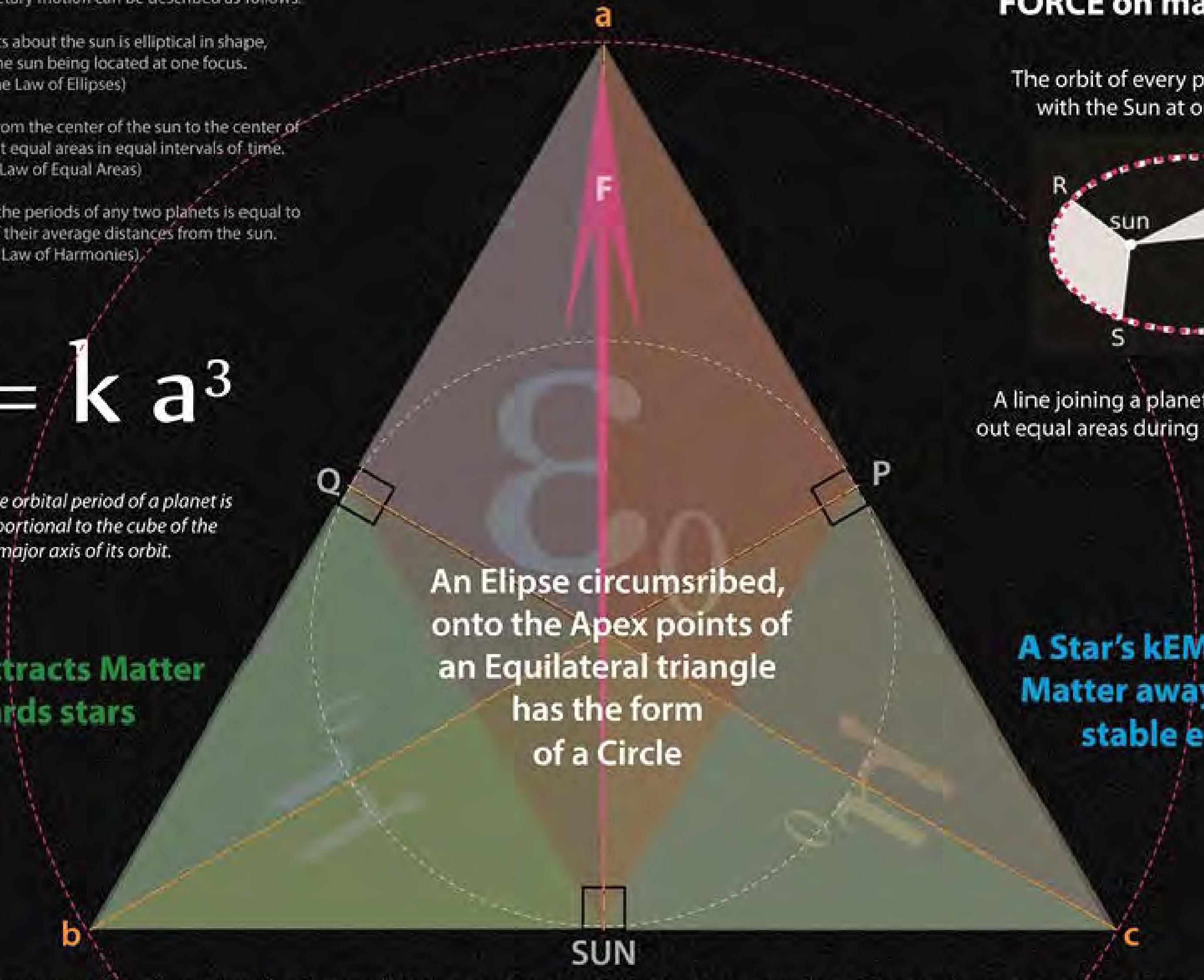

An Elipse circumsribed,
onto the Apex points of
an Equilateral triangle
has the form
of a Circle

A Star's kEM field accelerates Matter away from it creating stable ellipical orbits

*With an equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles.
Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another.
Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal.
This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral*

Copyright ABRAHAM [2008] - All rights reserved

The Major celestial bodies

All planetary orbits are the result of a balance between gravitation Matter and interactive kEM fields

The sun is a rotating charged mass-Matter object, and it creates a solar system-wide kEM field.

The solar mass is large enough to produce a EM field with lines of Force reaching the orbit of Pluto.

Planet	*Object Axial tilt (°)*	*Axial tilt (radians)*
Mercury	~0.01	0.00
Venus	177.4	3.10
Earth	23.44	0.41
Moon	1.5424	0.03
Mars	25.19	0.44
Jupiter	3.13	0.06
Saturn	26.73	0.47
Uranus	97.77	1.71
Neptune	28.32	0.49
Pluto	119.61	2.09

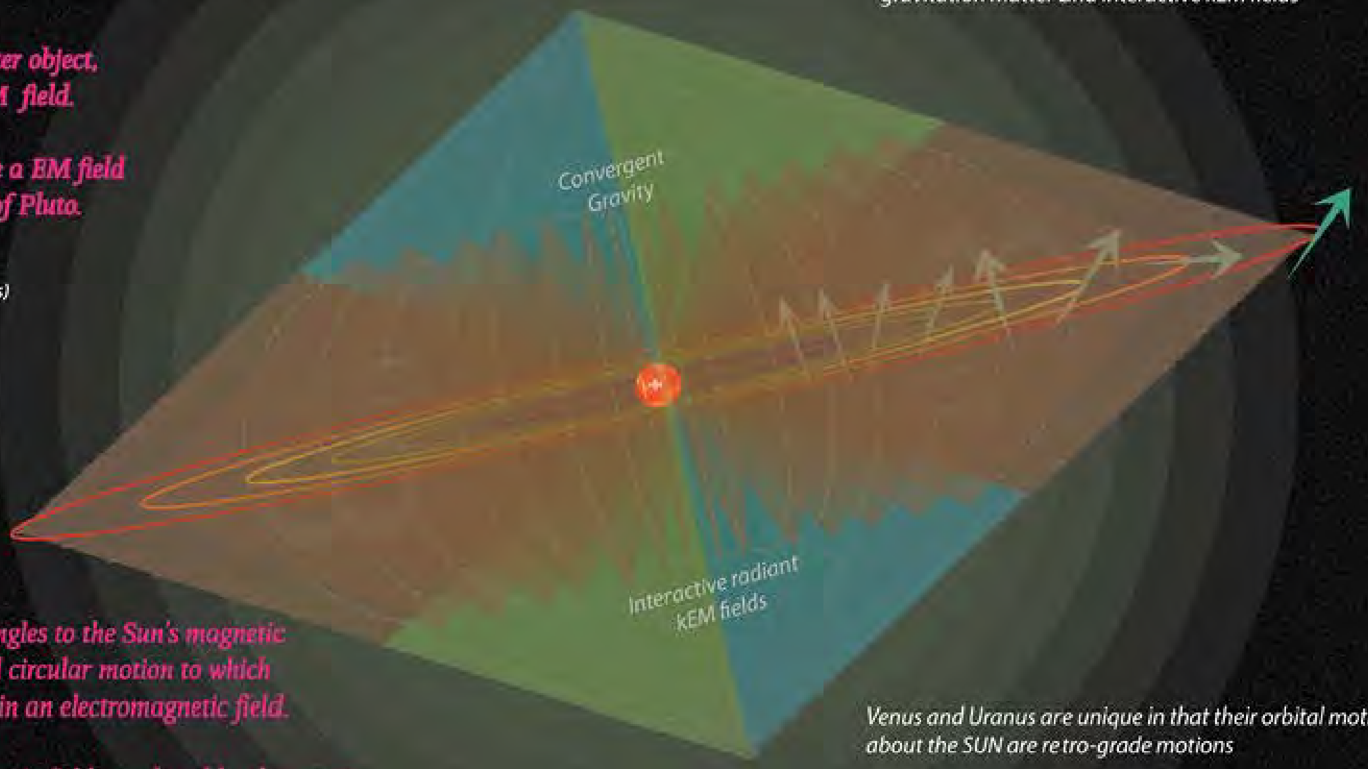

The charged planets move at right angles to the Sun's magnetic lines of force and describe the usual circular motion to which moving charged bodies are subject to in an electromagnetic field.

Venus and Uranus are unique in that their orbital motions about the SUN are retro-grade motions

Satellites, in turn, revolve in the smaller magnetic fields produced by the rotation of the planets. Mercury and Venus have no satellites as they produce extremely weak electromagnetic fields in addition to their Gravitational fields.

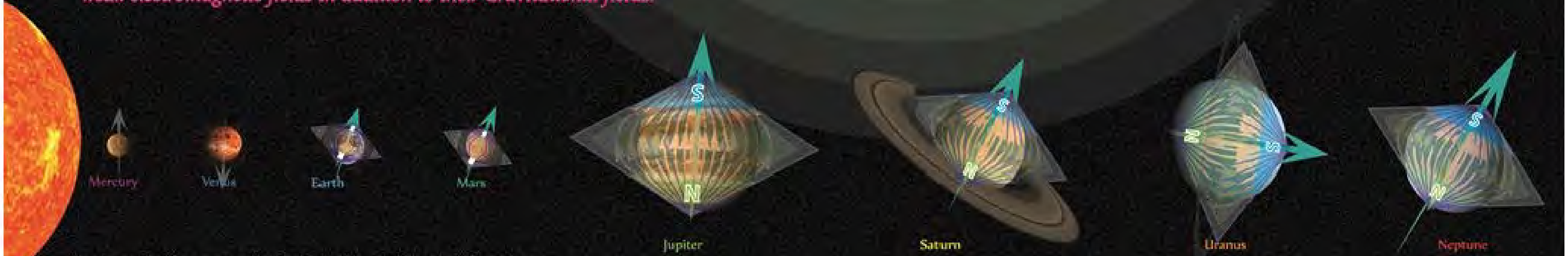

"In comparing the masses of the Earth, Jupiter and Saturn to their volumes, one remarks that the densities of these planets are, to some degree, inverse to their mean distances from the Sun. This rule is not true for Venus and Uranus"

Urbain Le Verrier

Solar orbit Inclinations

Inclination to Sun's equator (°)

Rocky Planets	
Mercury	3.38
Venus	3.86
Earth	7.15
Mars	5.65

Gas giants	
Jupiter	6.09
Saturn	5.51
Uranus	6.48
Neptune	6.43

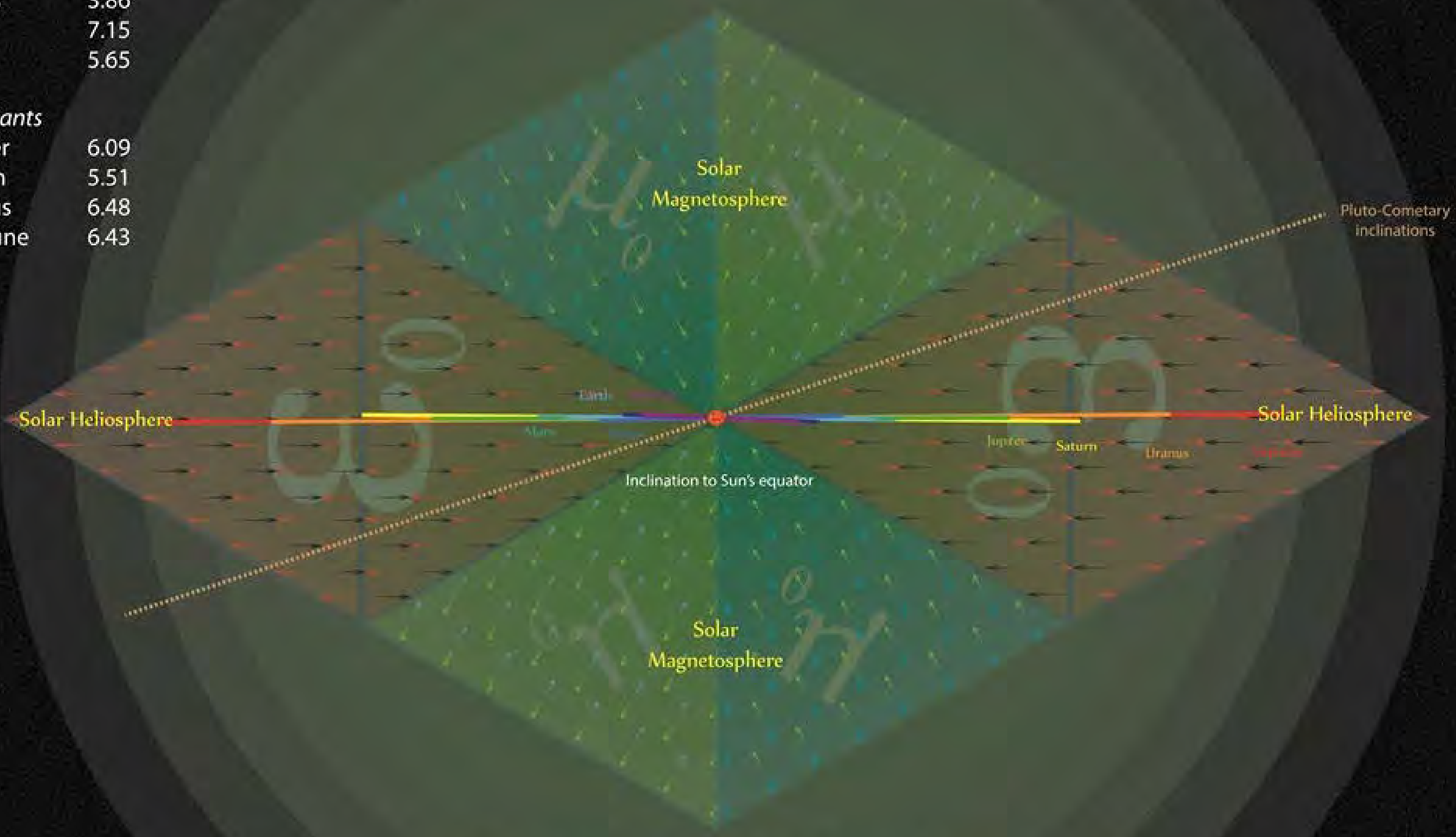
Solar Magnetosphere
μ_0
μ_0
Pluto-Cometary inclinations
ε_0
ε_0
Earth
Solar Heliosphere
Mars
Jupiter
Saturn
Uranus
Solar Heliosphere
Inclination to Sun's equator
μ_0
μ_0
Solar Magnetosphere

Orbital inclination is one of the six orbital parameters describing the shape and orientation of a celestial orbit.

It is the angular distance of the orbital plane from the plane of reference (usually the primary's equator or the ecliptic), normally stated in degrees.

Gravity is a strictly convergent Force

← ← ↔

GR = G + EM

Einstein Newton Lorentz

General Relativity models Newtonian Gravity with SR interactions

The Lagrange Points

are the five positions in an orbital configuration where a small object affected only by gravity can be stationary relative to two larger material bodies of mass-Matter

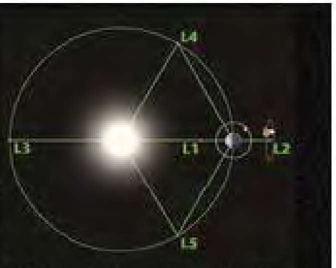

GEM forces between two bodies can be modelled with super-positioned equilateral energy geometries

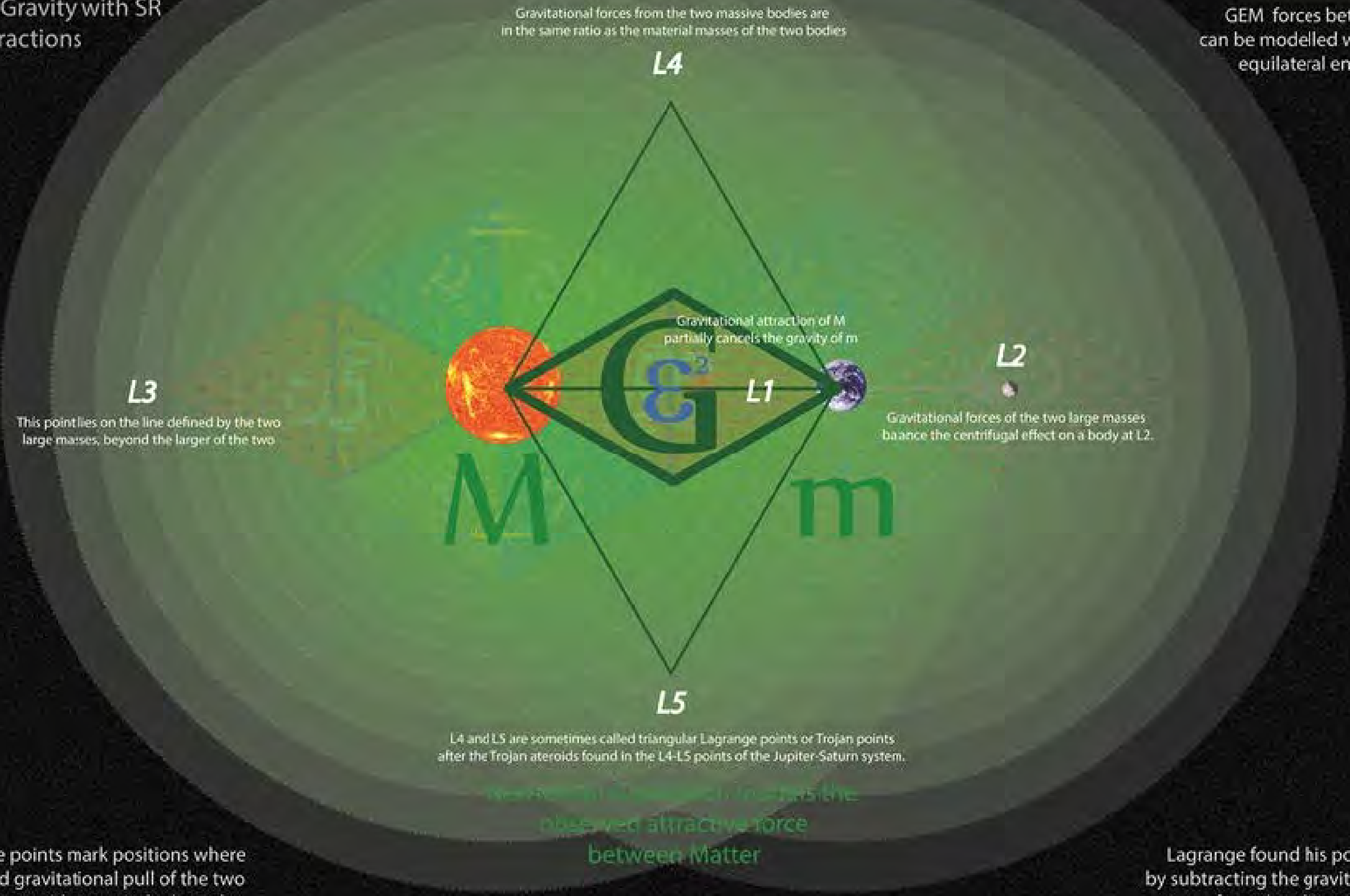

Gravitational forces from the two massive bodies are in the same ratio as the material masses of the two bodies

This point lies on the line defined by the two large masses, beyond the larger of the two

Gravitational attraction of M partially cancels the gravity of m

Gravitational forces of the two large masses baance the centrifugal effect on a body at L2.

L4 and L5 are sometimes called triangular Lagrange points or Trojan points after the Trojan ateroids found in the L4-L5 points of the Jupiter-Saturn system.

Newtonian Gravitation [illegible] the observed attractive force between Matter

The Lagrange points mark positions where the combined gravitational pull of the two large masses provides precisely the right centripetal force required to rotate with them

Lagrange found his points of stabilty by subtracting the gravitational potential energy from the object's kinetic energy.

Copyright ABRAHAM [2008] - All rights reserved

Earth-Moon Gravito-Electro-Magnetic fields

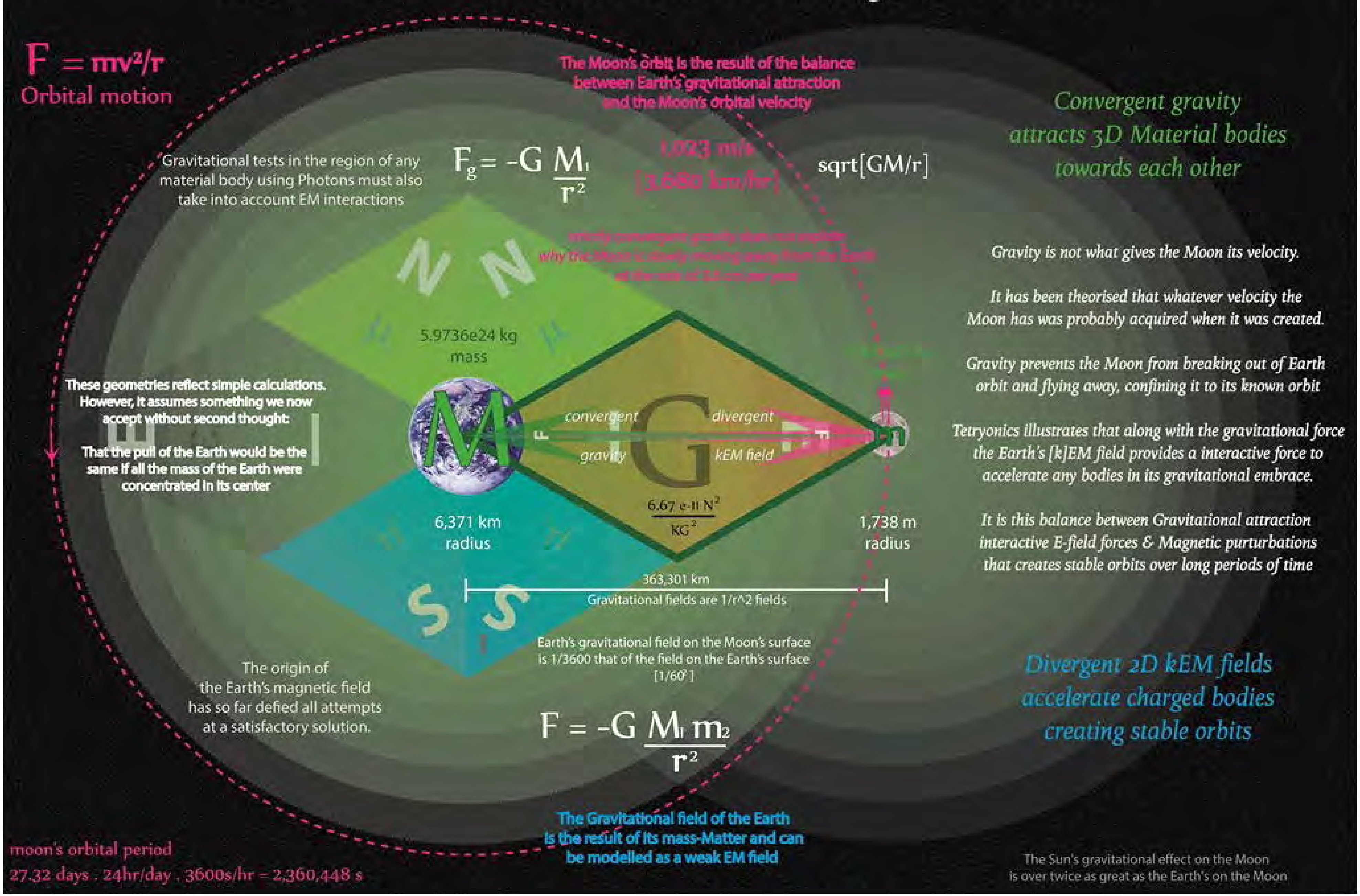

$F = mv^2/r$
Orbital motion

The Moon's orbit is the result of the balance
between Earth's gravitational attraction
and the Moon's orbital velocity

strictly convergent gravity does not explain
why the Moon is slowly moving away from the Earth
at the rate of 3.8 cm per year

Gravitational tests in the region of any material body using Photons must also take into account EM interactions

$$F_g = -G \frac{M_1}{r^2}$$

These geometries reflect simple calculations. However, it assumes something we now accept without second thought:

That the pull of the Earth would be the same if all the mass of the Earth were concentrated in its center

Earth's gravitational field on the Moon's surface
is 1/3600 that of the field on the Earth's surface
$[1/60^2]$

The origin of
the Earth's magnetic field
has so far defied all attempts
at a satisfactory solution.

$$F = -G \frac{M_1 m_2}{r^2}$$

The Gravitational field of the Earth is the result of its mass-Matter and can be modelled as a weak EM field

moon's orbital period
27.32 days . 24hr/day . 3600s/hr = 2,360,448 s

*Convergent gravity
attracts 3D Material bodies
towards each other*

Gravity is not what gives the Moon its velocity.

It has been theorised that whatever velocity the Moon has was probably acquired when it was created.

Gravity prevents the Moon from breaking out of Earth orbit and flying away, confining it to its known orbit

Tetryonics illustrates that along with the gravitational force the Earth's [k]EM field provides a interactive force to accelerate any bodies in its gravitational embrace.

It is this balance between Gravitational attraction interactive E-field forces & Magnetic perturbations that creates stable orbits over long periods of time

*Divergent 2D kEM fields
accelerate charged bodies
creating stable orbits*

The Sun's gravitational effect on the Moon
is over twice as great as the Earth's on the Moon

Tetryonics 74.06 - Earth-Moon EM interaction field

Copyright ABRAHAM [2008] - All rights reserved

The Earth's ElectroMagnetic field

Earth's magnetic field (also known as the geomagnetic field) is the magnetic field that extends from the Earth's inner core to where it meets the solar wind,

It is the magnetic dipole field tilted at an angle of 11 degrees with respect to the rotational axis

The Earth's magnetic field is really generated by the motion of the Earth through solar & vacuum energies and it is enhanced by the molten alloys in the Earth's core

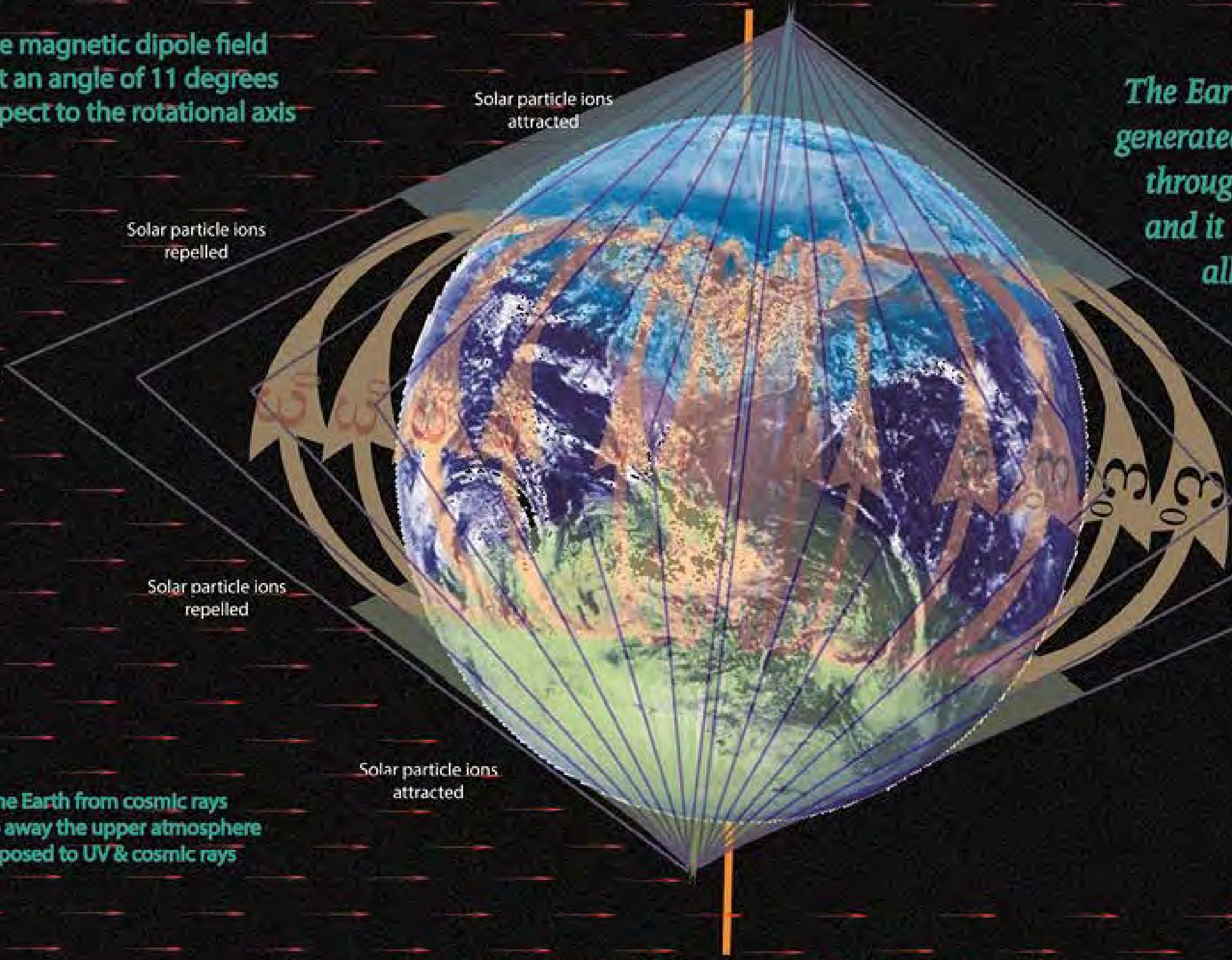
Solar particle ions attracted
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions attracted

It protects the Earth from cosmic rays that would strip away the upper atmosphere leaving us exposed to UV & cosmic rays

The magnetic poles precess and even flip over time

At random intervals (several hundred thousand years) the Earth's field reverses (the north and south geomagnetic poles change places with each other)

Tetryonics 74.07 - The Earth's Magnetic field

Copyright ABRAHAM [2008] - All rights reserved

The Earth's molten core

The Earth has an outer silicate solid crust, a highly viscous mantle, a liquid outer core that is much less viscous than the mantle, and a solid inner core.
Scientific understanding of Earth's internal structure is based on observations of topography and bathymetry.

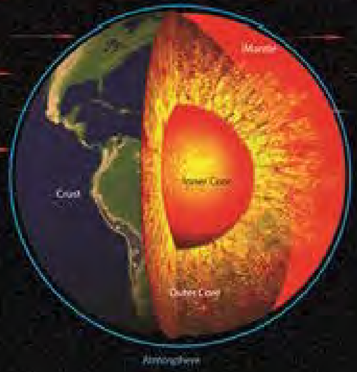

Q

The SUN and planetary bodies are all charged Matter moving in, displacing and interacting with solar and vacuum energies

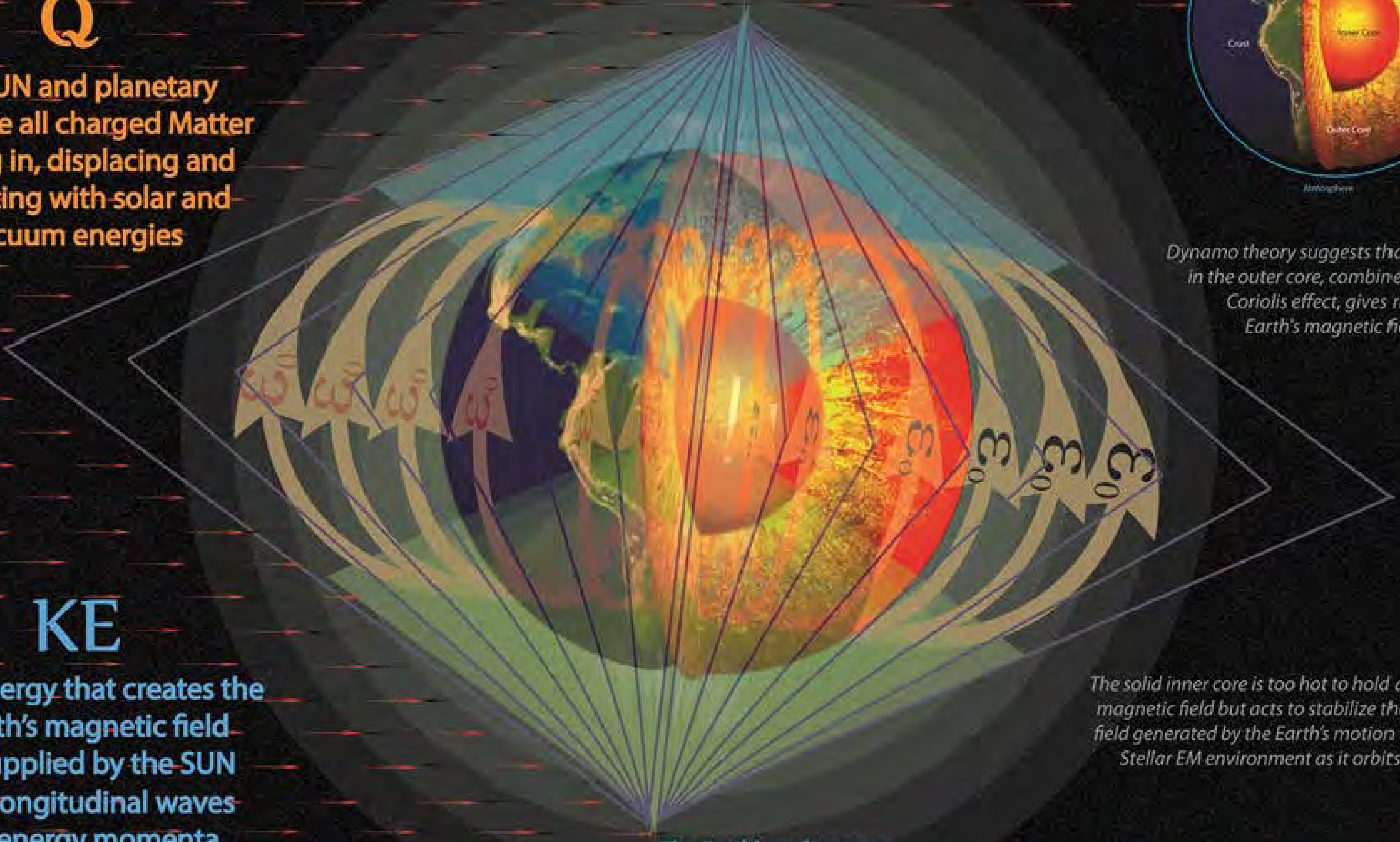

Dynamo theory suggests that convection in the outer core, combined with the Coriolis effect, gives rise to Earth's magnetic field.

KE

The energy that creates the Earth's magnetic field is supplied by the SUN via longitudinal waves of energy momenta

The solid inner core is too hot to hold a permanent magnetic field but acts to stabilize the magnetic field generated by the Earth's motion through the Stellar EM environment as it orbits the SUN

The Earth's molten core does not produce the Earth's magnetic field it enhances it

The force exerted by Earth's gravity can be used to calculate its mass, and by estimating the volume of the planet, its average density can be calculated.

Astronomers can also calculate Earth's mass from its orbit and effects on nearby planetary bodies.

Terrestrial Aurora

result from the interaction of the Earth's magnetic field and the energetic solar wind, which provides energised ions and photons to the polar regions of Planets

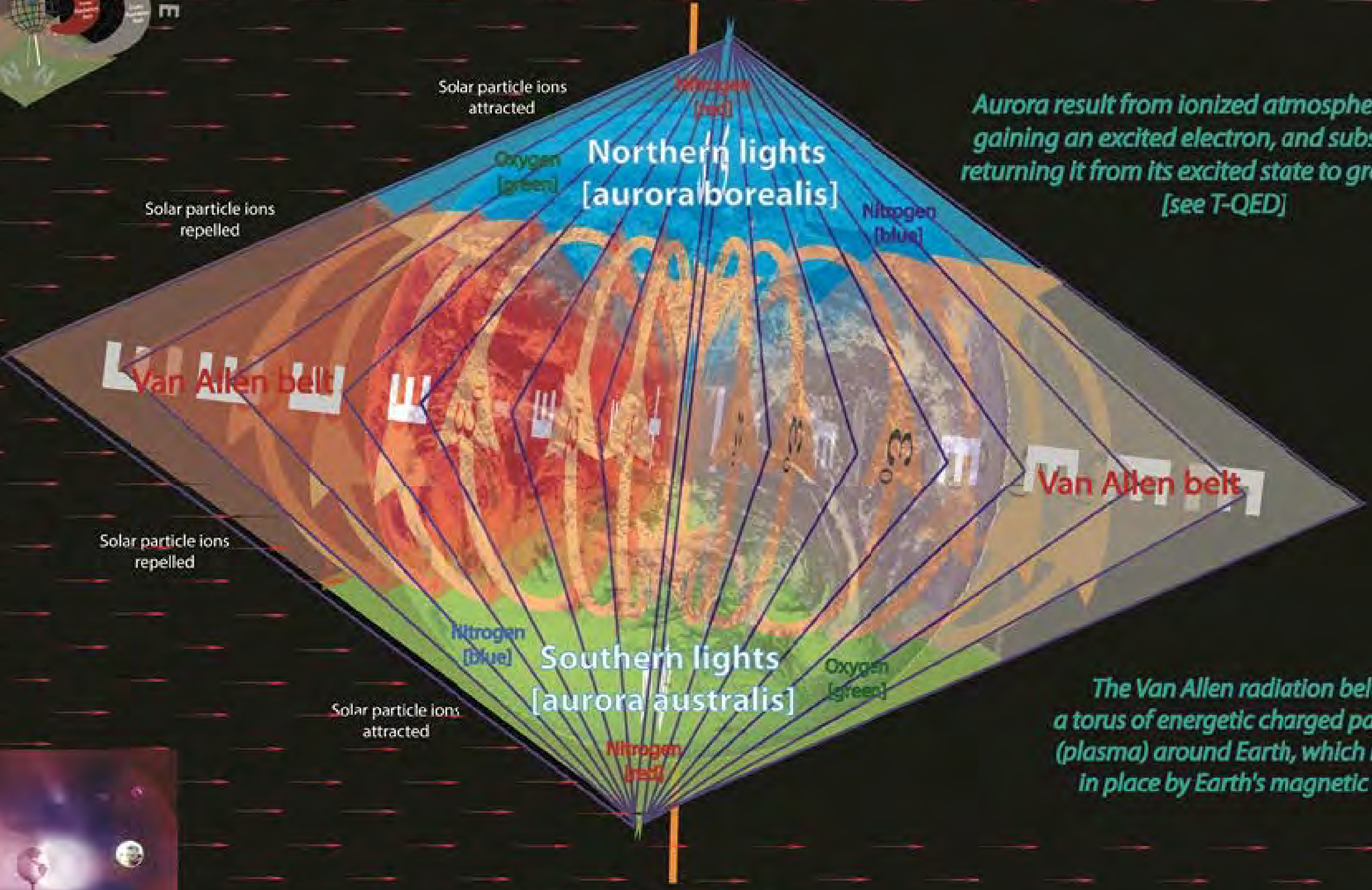
Solar particle ions attracted
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions repelled
Solar particle ions attracted
Nitrogen [red]
Oxygen [green]
Northern lights [aurora borealis]
Nitrogen [blue]
Van Allen belt
Van Allen belt
Nitrogen [blue]
Southern lights [aurora australis]
Oxygen [green]
Nitrogen [red]

Aurora result from ionized atmospheric atoms gaining an excited electron, and subsequently returning it from its excited state to ground state [see T-QED]

The Van Allen radiation belt is a torus of energetic charged particles (plasma) around Earth, which is held in place by Earth's magnetic field

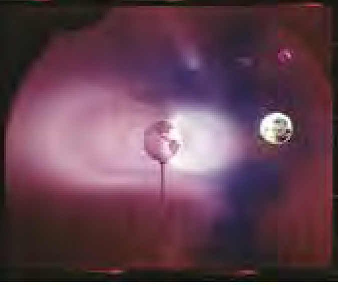

Aurora occur on many planets throughout the Solar system

Copyright ABRAHAM [2008] - All rights reserved

Absolute Rest Frames

Absolute Rest is intrinsically linked to the concepts of absolute time and space originally introduced by Sir Isaac Newton in the Philosophiæ Naturalis Principia Mathematica, which provided a theoretical foundation that facilitated Newtonian mechanics.

According to Newton, absolute time and space respectively are independent aspects of objective reality.

Rest in physics refers to an object being stationary relative to a particular frame of reference or another object

In Tetryonics an Inertial Rest frame can be determined by the absence of a Kinetic EM field & its Magnetic moment

The concepts of space and time were separate in physical theory prior to the advent of special relativity theory, which connected the two and showed both to be dependent upon the observer's inertial state of motion relative to other inertial frames.

In Einstein's theories, the ideas of absolute time and space were superseded by the notion of spacetime in special relativity, and by dynamically curved spacetime in general relativity.

$$p^2 = E = mv^2$$

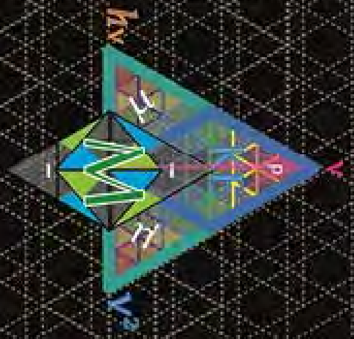

$$\frac{p^2}{2} = KE = \frac{1}{2}Mv^2$$

In some of his later paper (especially in 1920 and 1924), Einstein gave a new definition of the aether by identifying it with "properties of space", and this aether can be called absolute, as long as its state cannot be influenced by Matter.

So he argued that Newton's absolute space can be considered as an "absolute aether". Also the four-dimensional spacetime of special relativity (which replaces the absolute space of Newton) would be some sort of "absolute aether", as its states cannot be influenced by Matter.

Tetryonics clearly establishes free space as a 3D volume filled with a interactive vacuum [aether] and EM energies against which any relative motion can be deduced by the presence (or absence) of divergent [K]EM fields produced as a direct result of Matter in motion additionally, 3D mass-Matter topologies displace the aether [vacuum energy field] producing convergent differential energy gradients [gravity fields]

Copyright ABRAHAM [2008] - All rights reserved

Red-shifted Photonic Energies

Energy is ALWAYS conserved... longer wavelength [red-shifted] photon energies are a reflection of the time specific measurement of photonic energies at a particular distance from the source of the light

Photons are neutral, coupled charge equilateral mass-energy momenta geometries that radiate bilaterally at the speed of light in a vacuum

Photons

$$\frac{2\pi}{c^2}\left[\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

mass velocity

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[\overset{\text{Planck quanta}}{m\Omega v^2}\right]\right]$$

mass velocity

As photons radiate from a point source at 'c'

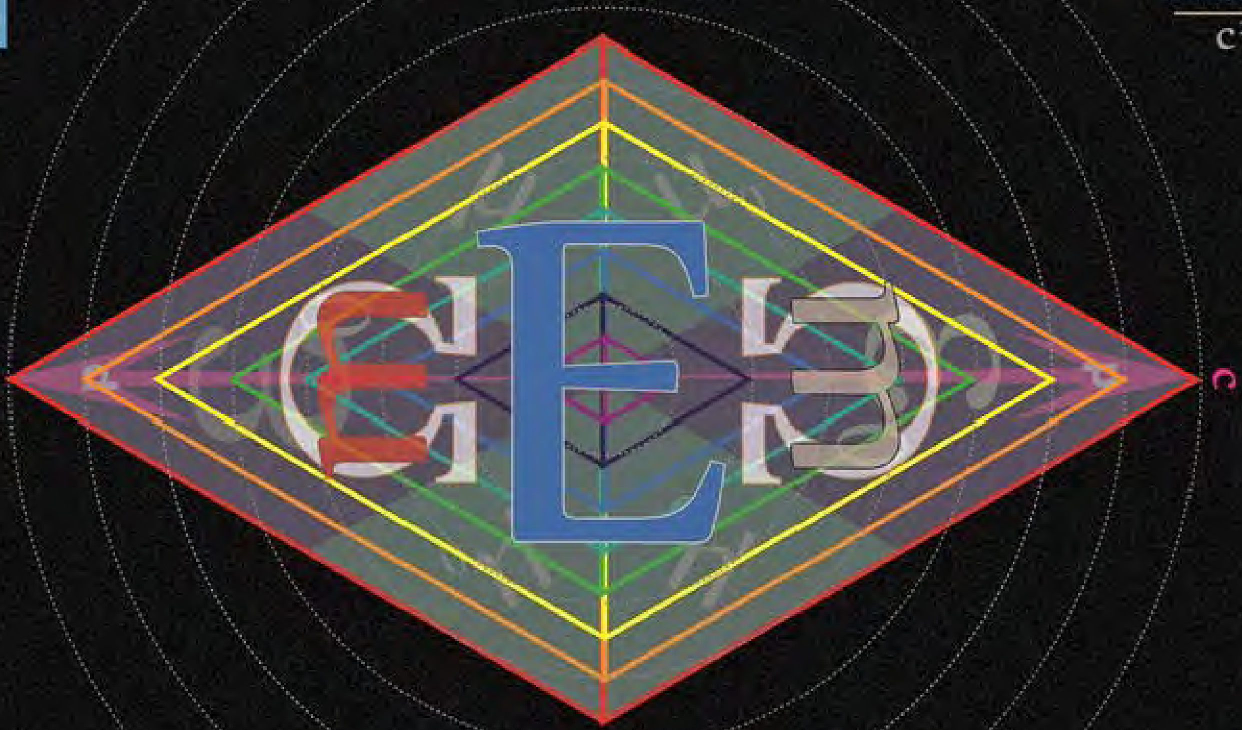

their EM energies diminsh as per the inverse square law

producing the familiar red-shifting of spectral lines

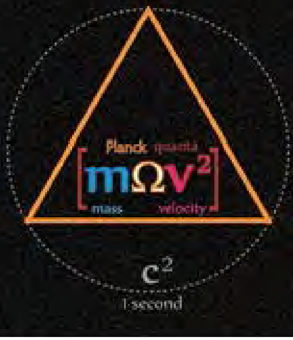

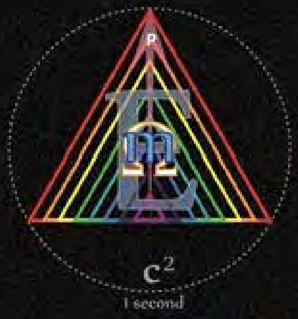

Tetryonics 75.02 - Red-shifted Photonic energies

Copyright ABRAHAM [2008] - All rights reserved

The red shift of Photons

dλ

df

According to special relativity, as something approaches the speed of light, the passage of time slows.
A twin moving at the speed of light would not age or change relative to the stationary twin.

General relativity additionally states that the cosmological red shift is the result of a photon traveling through an expanding space time field which causes the photon to loose energy and it's wavelength correspondingly increases.

A photon travels at the speed of light; therefore it will not experience the passage of time, therefore it will not change; yet it does.

The divergent spreading of EM energy over time from a source creates the spacetime aether

The Quantised Angular Momentum and Energy content are directly related through the geometry of Photons

All EM Energy waveforms (save Matter) radiate outward following the inverse square law & reducing their energy content over time

Blue-shifted photon travelling towards observer

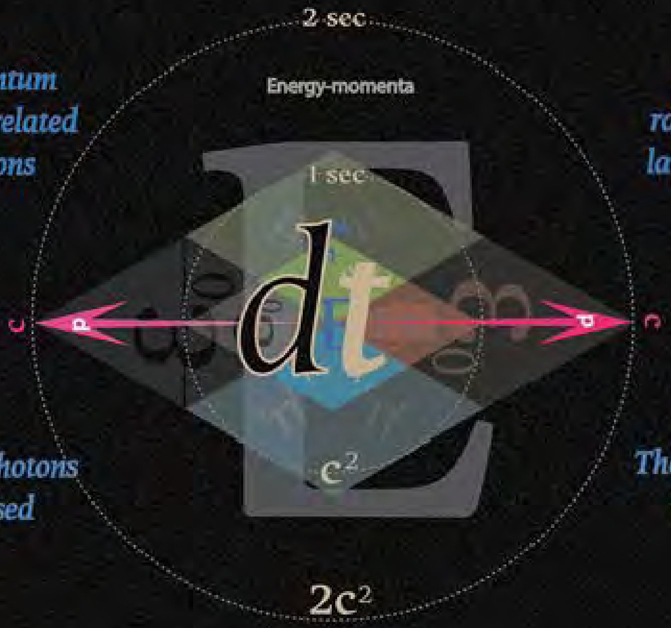

Red-shifted photon moving away from observer

The EM mass-energy content of photons reveals photons to be a quantised electromotive field

The wavelength of a photon is related directly to its quantised mass-energy momenta geometry per unit of time

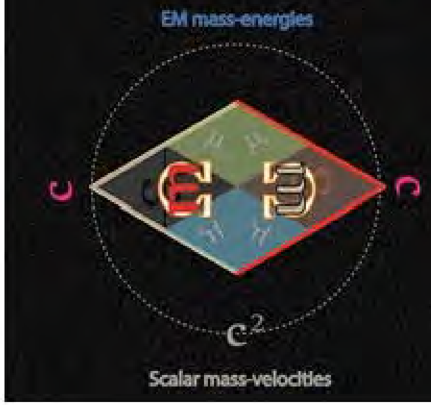

Tetryonic geometry relates photon mass-energy momenta to the constant velocity of light in a vacuum and its intrinsic quantised angular momenta

$$\frac{mc^2}{c^2} = KE = \frac{hv^2}{c^2}$$

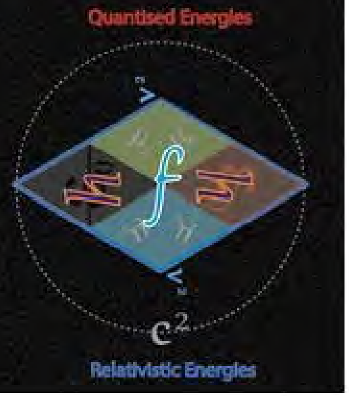

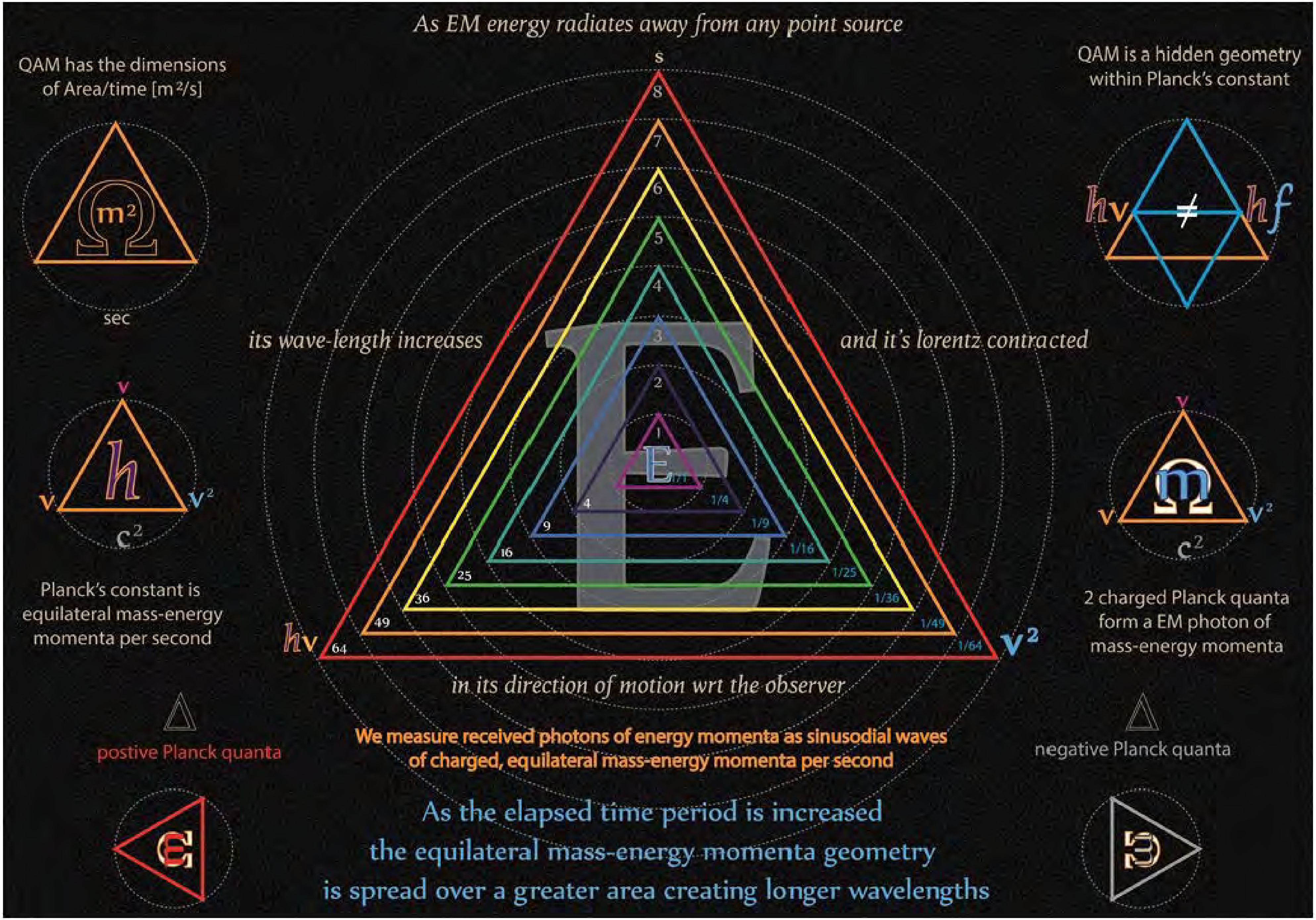

As EM energy radiates away from any point source

its wave-length increases

and it's lorentz contracted

in its direction of motion wrt the observer

QAM has the dimensions of Area/time [m²/s]

QAM is a hidden geometry within Planck's constant

Planck's constant is equilateral mass-energy momenta per second

2 charged Planck quanta form a EM photon of mass-energy momenta

postive Planck quanta

negative Planck quanta

We measure received photons of energy momenta as sinusodial waves of charged, equilateral mass-energy momenta per second

As the elapsed time period is increased
the equilateral mass-energy momenta geometry
is spread over a greater area creating longer wavelengths

Tetryonics 75.04 - Doppler wavelengths

Copyright ABRAHAM [2008] - All rights reserved

The Doppler shifting of EM waves

Photons

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta

mass velocity

EM waves

$$\frac{\text{EVEN}\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta

mass velocity

General relativity incorrectly holds that the cosmological red shift is the result of a photon traveling through an expanding space time field which causes the photon to loose energy and it's wavelength correspondingly increases

Photonic energies form
inifitie energy loops
[quantum transformers]

It is the Photon's intrinsic geometry that increases the further it travels from the source due to the inverse square law resulting in less energy (per unit of time) being recieved at increasing distances from the source as the Photon's specific QAM [m²/s] has increased over time

All EM waves are comprised
of Photons of specific energies
[Light is superpositioned EM waves]

All Photons are quantum harmonic oscillators
and create sinusoidal EM waveforms
are they radiate and are detected

Radiating geometric EM mass-energies
from a source (over Distance-Time)
creates the spatial vacuum aether

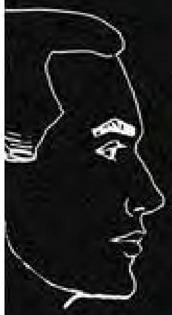

Blue-shifted

observer moves towards source *or* source moves towards observer

Any relative motion between the source & observer that reduces the distance between them produces a Blue-shift in the measured quanta-frequency/ energy of the EM wave

observer moves away from source *or* source moves away from observer

Red-shifted

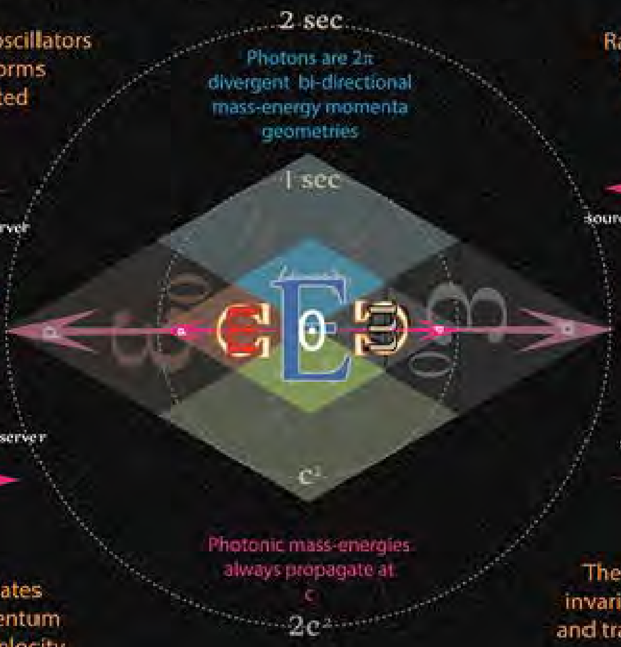

Photons are 2π
divergent bi-directional
mass-energy momenta
geometries

Photonic mass-energies
always propagate at
c

Red-shifted

source moves away from observer *or* observer moves away from source

Any relative motion between the source and observer that increases the distance bewteen them produces a Red-shift in the measured quanta-frequency/ energy of the EM wave

source moves towards observer *or* observer moves towards source

Blue-shifted

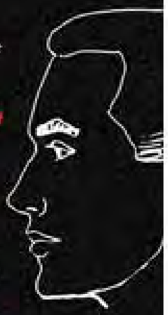

Tetryonics clearly shows and relates
Time to quantised angular momentum
and Planck energy momenta to velocity

The energy of indiviudal photons is
invariant until they interact with Matter
and transfer their mass-energy momenta

motion toward the source increases the number of quanta per sec measured

[v+v]

Cosmological doppler-shifting of Light is the direct result of the motion of Matter in the Universe

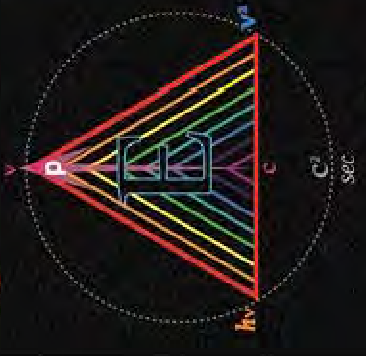

Poor definintions of Energy per unit of Time
along with the interchange of quanta and frequency
have led to the mistaken concept of photons losing energy

$$mv^2 = E = hv^2$$

Photons DO NOT lose energy as they travel.
the group energy [per unit of time] of the EM wave
they form spreads out over time as a direct result
of the inverse square law

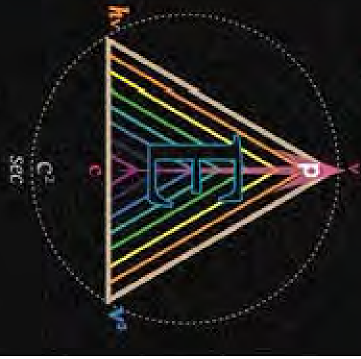

motion away from the source decreases the number of quanta per sec measured

[v-v]

Gravitational Red-shifting is an EM effect not gravitational

Tetryonics 75.05 - The Doppler shifting of EM waves

Relativistic Corrections

are applied to only [kinetic]EM mass-energy geometries

$$d\lambda \qquad \frac{Mv^2}{c^2} = KEM = \frac{hv^2}{c^2} \qquad dt$$

Wave-length contraction

linear
Lorentz corrections

$$\beta = \left[\frac{v}{c}\right]$$

SR - measuring rods in differing interial frames have different contracted lengths

GR - lengths measured in curved spacetime are different to straight lines in flat spacetime

$$\frac{\Delta\Omega}{c}$$

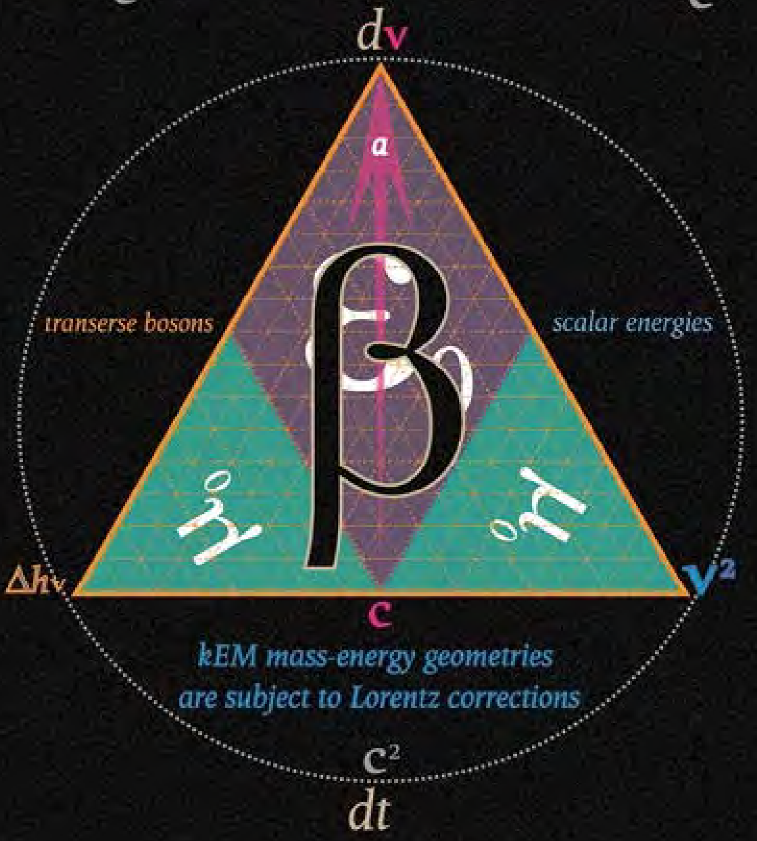

kEM mass-energy geometries
are subject to Lorentz corrections

scalar energy momenta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v\right]^2\right]$$

Matter, Planck quanta, mass, velocity

linear momentum

Time Dilation

scalar
Lorentz corrections

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

SR - clocks in differing inertial frames experience time dilation

GR - clocks in differing Gravitational fields experience time dilation

$$\frac{\Delta\Omega}{c^2}$$

The mass-energies of standing-wave Matter topologies are Lorentz invariant

GPS relativistic corrections

Claims that the Global Positioning System (GPS) provides a valid test of General relativity are erroneous
[GR provides Lorentzian [SR] correcttions under the guise of Gravitational effects]

Photons are mass-energy momenta quanta
[they are Matter-less NOT mass-less]
and are subect to Lorentz Forces and corrections

Newton G + SR = GR *Einstein*
Lorentz

Graviational Matter

M

All Matter topologies
produce
a Gravitational field

Newtonian Gravity
is universal

The gravitational effects
caused by Matter can only
be accurately modelled [or
tested] through the exclusion
of all EM fields interactions

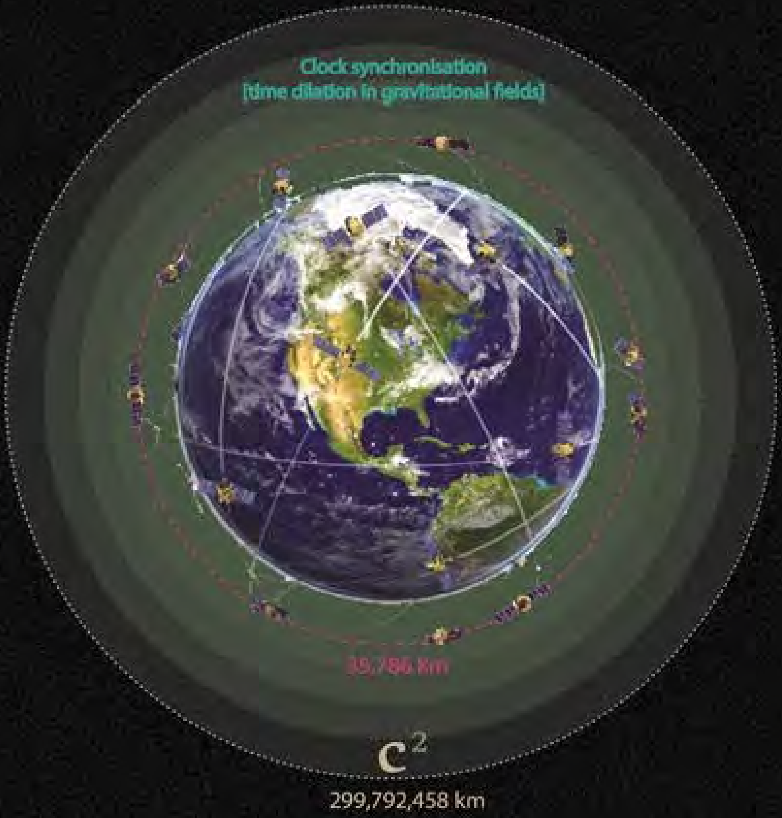

Interactive masses

m

All Matter in motion
produces
interactive mass-energies

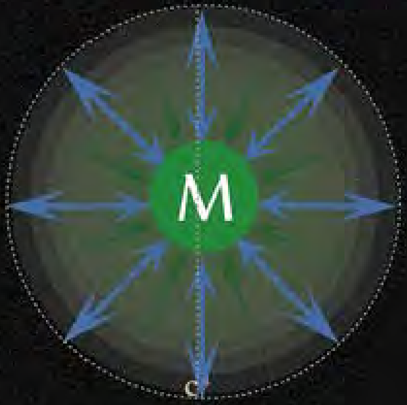

General Relativity
is limted to c

Any test of GR using Photons
as a testing mechanism
by definition is affected by
and subject to Special relativity
[ElectroMagnetic effects]

The 'gravitational' frequency shifting [Red-Blue shifting of EM waves] is in fact an electromagnetic effect

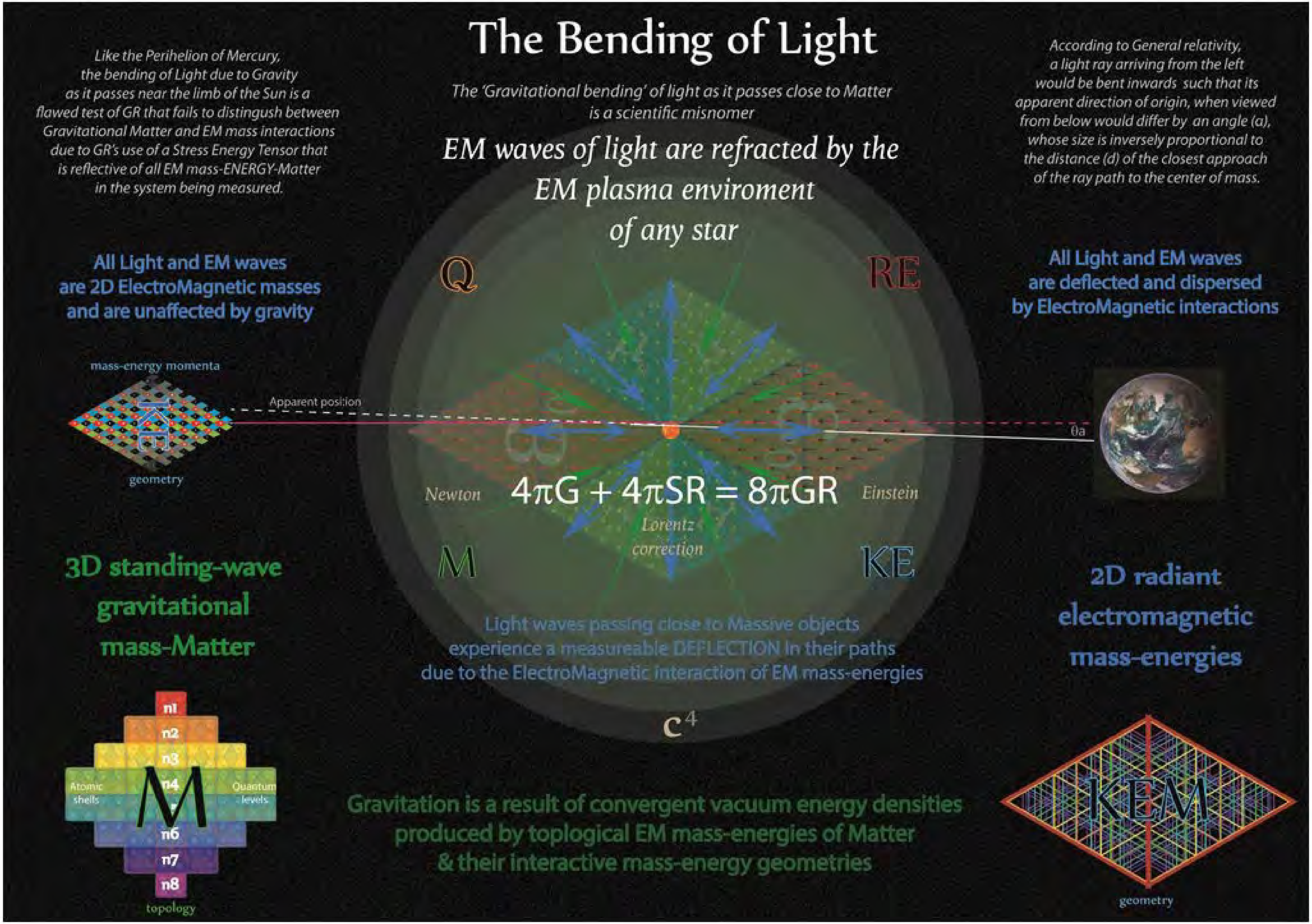

The Bending of Light

Like the Perihelion of Mercury, the bending of Light due to Gravity as it passes near the limb of the Sun is a flawed test of GR that fails to distingush between Gravitational Matter and EM mass interactions due to GR's use of a Stress Energy Tensor that is reflective of all EM mass-ENERGY-Matter in the system being measured.

The 'Gravitational bending' of light as it passes close to Matter is a scientific misnomer

EM waves of light are refracted by the EM plasma enviroment of any star

According to General relativity, a light ray arriving from the left would be bent inwards such that its apparent direction of origin, when viewed from below would differ by an angle (a), whose size is inversely proportional to the distance (d) of the closest approach of the ray path to the center of mass.

All Light and EM waves are 2D ElectroMagnetic masses and are unaffected by gravity

All Light and EM waves are deflected and dispersed by ElectroMagnetic interactions

$$4\pi G + 4\pi SR = 8\pi GR$$

3D standing-wave gravitational mass-Matter

Light waves passing close to Massive objects experience a measureable DEFLECTION in their paths due to the ElectroMagnetic interaction of EM mass-energies

2D radiant electromagnetic mass-energies

Gravitation is a result of convergent vacuum energy densities produced by toplogical EM mass-energies of Matter & their interactive mass-energy geometries

Q
RE
mass-energy momenta
KE
Apparent position
geometry
θa
Newton
Einstein
Lorentz correction
M
KE
c^4
n1
n2
n3
n4
n6
n7
n8
Atomic shells
Quantum levels
M
topology
KEM
geometry

Tetryonics 75.08 - The Bending of Light

Copyright ABRAHAM [2008] - All rights reserved

The Perihelion of Mercury

is the result of EM interactions in addition to gravitational attraction

The perihelion of Mercury's orbit precesses at a measurable rate, but even after accounting for gravitational perturbations caused all other planets in the solar system, Newton's theory (assuming a precise inverse-square relationship for distance) predicts a rate of precession that differs from the measured rate by approximately 43 arcseconds per century.

Both Newton and Einstein saw Gravity as a singular force of attraction between objects and developed their math accordingly from the observed motions without any physical distinction between mass & Matter

$4\pi G$

Newton modelled only the observed nett attractive force of Matter in the system

$8\pi G$

Einstein modelled the nett convergent force of all energies without differentiation

$$F = -G\frac{M_1 m_2}{r^2}$$

$$G_{ab} = K T_{ab}$$

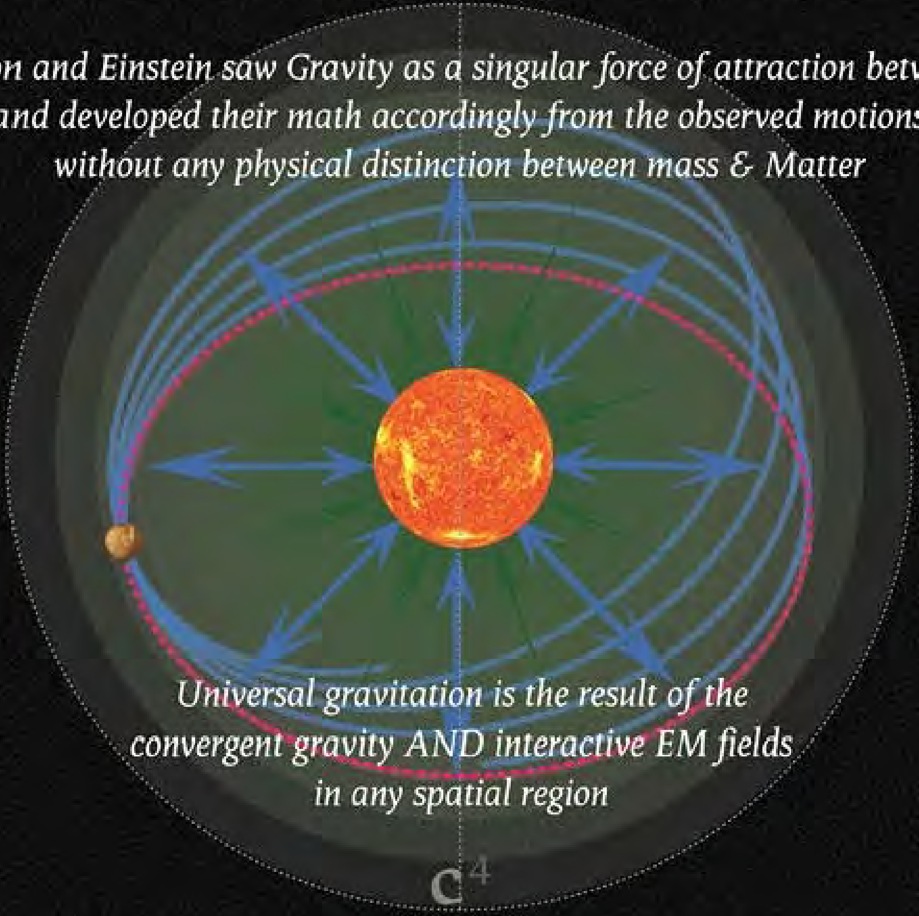

Universal gravitation is the result of the convergent gravity AND interactive EM fields in any spatial region

c^4

Newton

G

Einstein

$G+EM$

Newton Lorentz

General relativity was developed in part to provide an estimate for this rate of precession that better matches observations.

Newton's theory can also be 'incorrectly modified' to explain this perihelion by factoring in the gravitational pull due to other planets or making tiny adjustments to parameters in the gravitational equation to reflect the effect of ElectroMagnetic interactions between the Sun and Mercury.

By modelling the quantum interactions of ALL GEM mass-ENERGIES and their distinct geometries Tetryonic theory explains recently measured discrepancies in both Newtonian & GR formulations of Gravity

Copyright ABRAHAM [2008] - All rights reserved

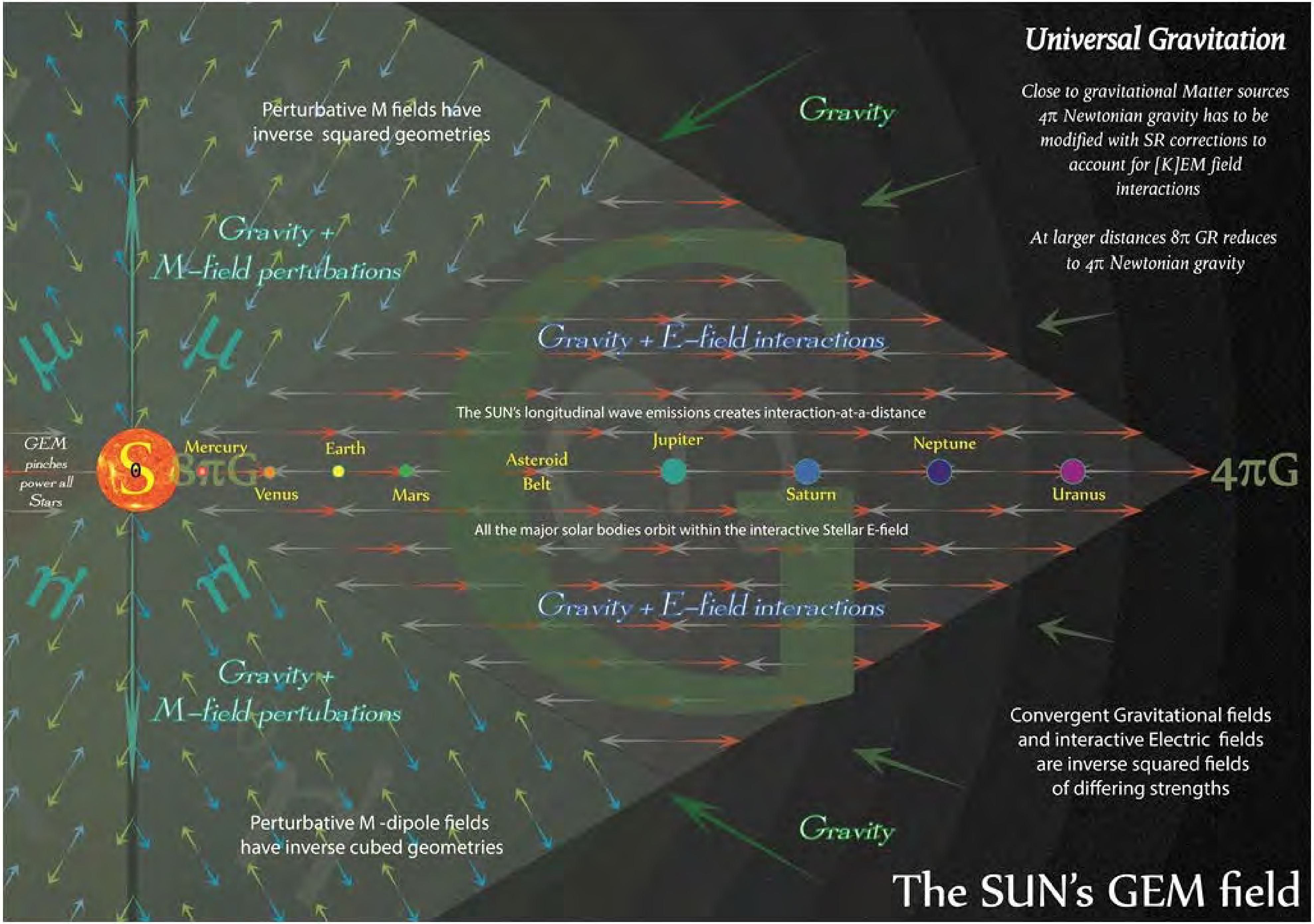

Universal Gravitation

Close to gravitational Matter sources 4π Newtonian gravity has to be modified with SR corrections to account for [K]EM field interactions

At larger distances 8π GR reduces to 4π Newtonian gravity

Perturbative M fields have inverse squared geometries

The SUN's longitudinal wave emissions creates interaction-at-a-distance

All the major solar bodies orbit within the interactive Stellar E-field

Perturbative M -dipole fields have inverse cubed geometries

Convergent Gravitational fields and interactive Electric fields are inverse squared fields of differing strengths

GEM pinches power all Stars

The SUN's GEM field

Gravity
Gravity + M–field pertubations
μ
μ
Gravity + E–field interactions
S
8πG
Mercury
Venus
Earth
Mars
Asteroid Belt
Jupiter
Saturn
Neptune
Uranus
4πG
η
η
Gravity + E–field interactions
Gravity + M–field pertubations
Gravity

Tetryonics 76.01 - The SUN's GEM field

Copyright ABRAHAM [2008] - All rights reserved

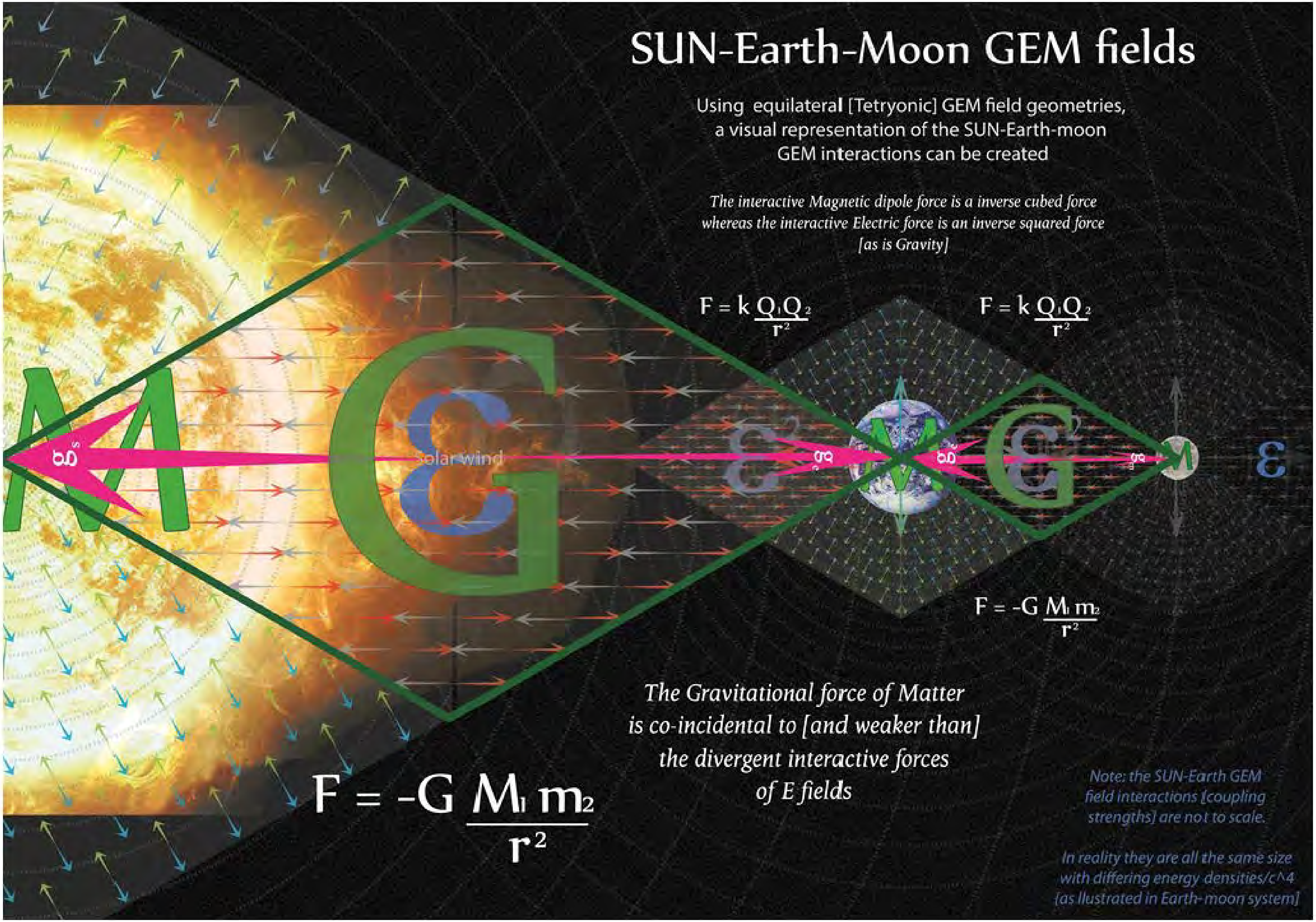

SUN-Earth-Moon GEM fields

Using equilateral [Tetryonic] GEM field geometries, a visual representation of the SUN-Earth-moon GEM interactions can be created

The interactive Magnetic dipole force is a inverse cubed force whereas the interactive Electric force is an inverse squared force [as is Gravity]

$$F = k\frac{Q_1 Q_2}{r^2}$$

$$F = k\frac{Q_1 Q_2}{r^2}$$

$$F = -G\frac{M_1 m_2}{r^2}$$

The Gravitational force of Matter is co-incidental to [and weaker than] the divergent interactive forces of E fields

$$F = -G\frac{M_1 m_2}{r^2}$$

Note: the SUN-Earth GEM field interactions [coupling strengths] are not to scale.

In reality they are all the same size with differing energy densities/c^4 [as illustrated in Earth-moon system]

Tetryonics 76.02 - SUN-Earth-moon GEM fields

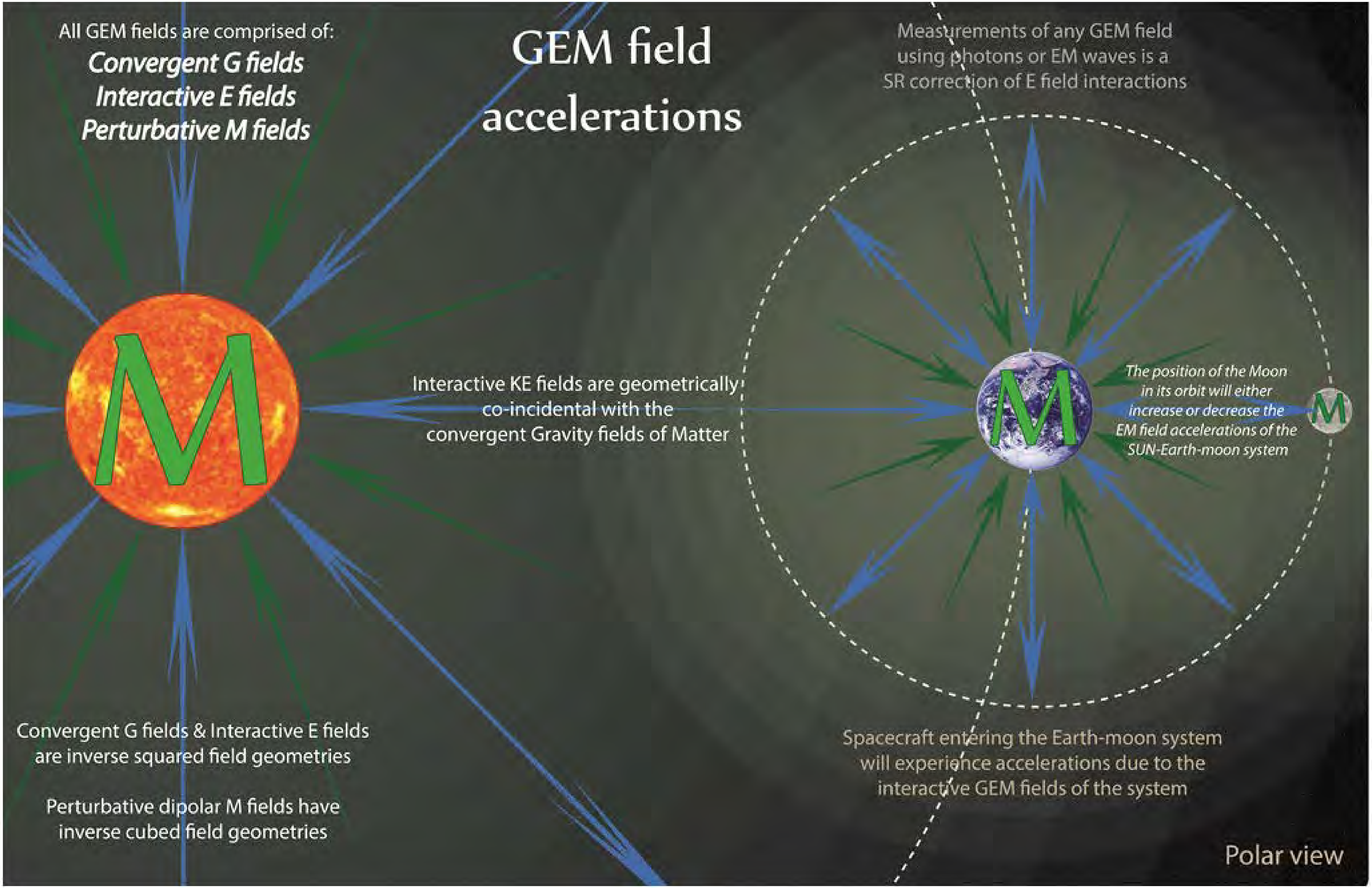

GEM field accelerations

All GEM fields are comprised of:

Convergent G fields
Interactive E fields
Perturbative M fields

Measurements of any GEM field using photons or EM waves is a SR correction of E field interactions

Interactive KE fields are geometrically co-incidental with the convergent Gravity fields of Matter

The position of the Moon in its orbit will either increase or decrease the EM field accelerations of the SUN-Earth-moon system

Convergent G fields & Interactive E fields are inverse squared field geometries

Perturbative dipolar M fields have inverse cubed field geometries

Spacecraft entering the Earth-moon system will experience accelerations due to the interactive GEM fields of the system

Polar view

Tetryonics 76.03 - GEM fields [Polar]

GEM field accelerations

All GEM fields are comprised of:
Convergent G fields
Interactive E fields
Perturbative M fields

Interactive KE field vectors are geometrically co-incidental with the strictly convergent vector of Gravity

The Electric field forces of GEM fields are produced by Longitudinal EM waveforms

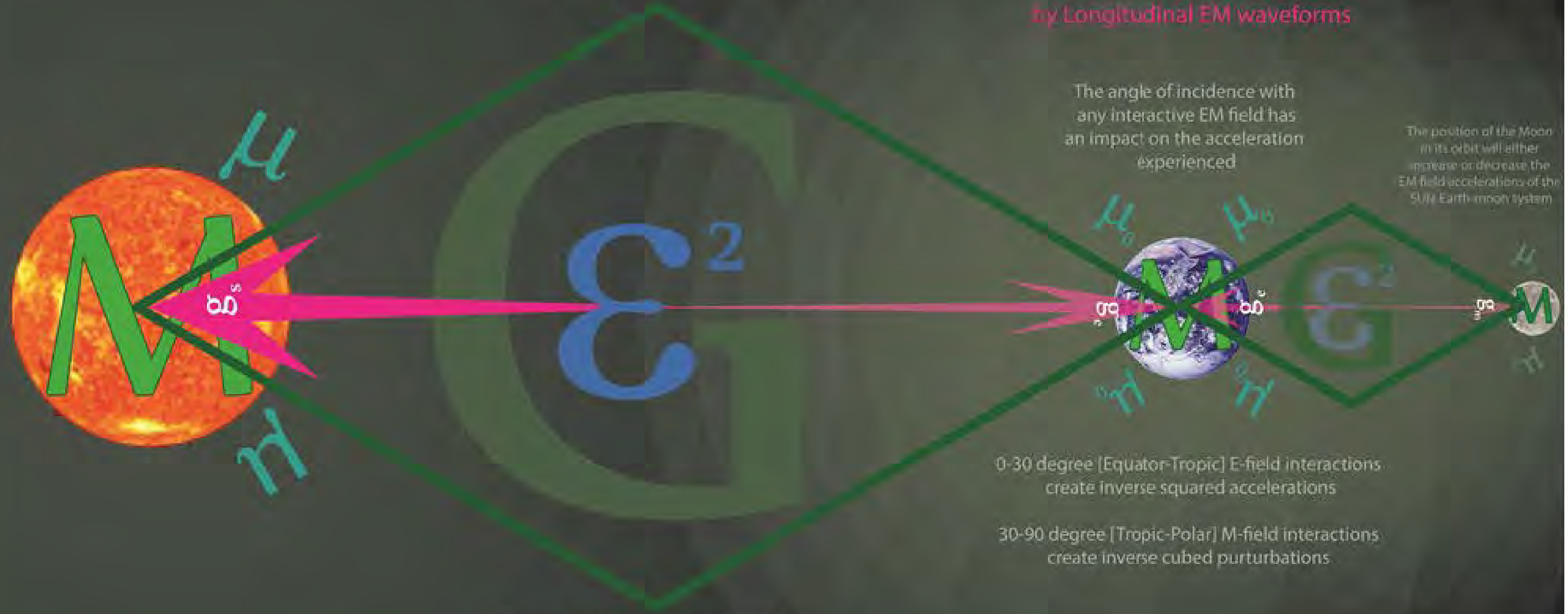

The angle of incidence with any interactive EM field has an impact on the acceleration experienced

The position of the Moon in its orbit will either increase or decrease the EM field accelerations of the SUN Earth-moon system

0-30 degree [Equator-Tropic] E-field interactions create inverse squared accelerations

30-90 degree [Tropic-Polar] M-field interactions create inverse cubed purturbations

The combined interactions of GEM fields create the nett convergent motional force we have historically observed and modelled as gravity

Strictly Newtonian Gravity fields are the result of the spatial energy differential created by Matter as it topologically displaces the surrounding Vacuum Energies

Convergent G fields & interactive E fields are inverse squared geometries

Perturbative dipolar M fields have inverse cubed geometries

Side elevation

Tetryonics 76.04 - GEM fields [Side]

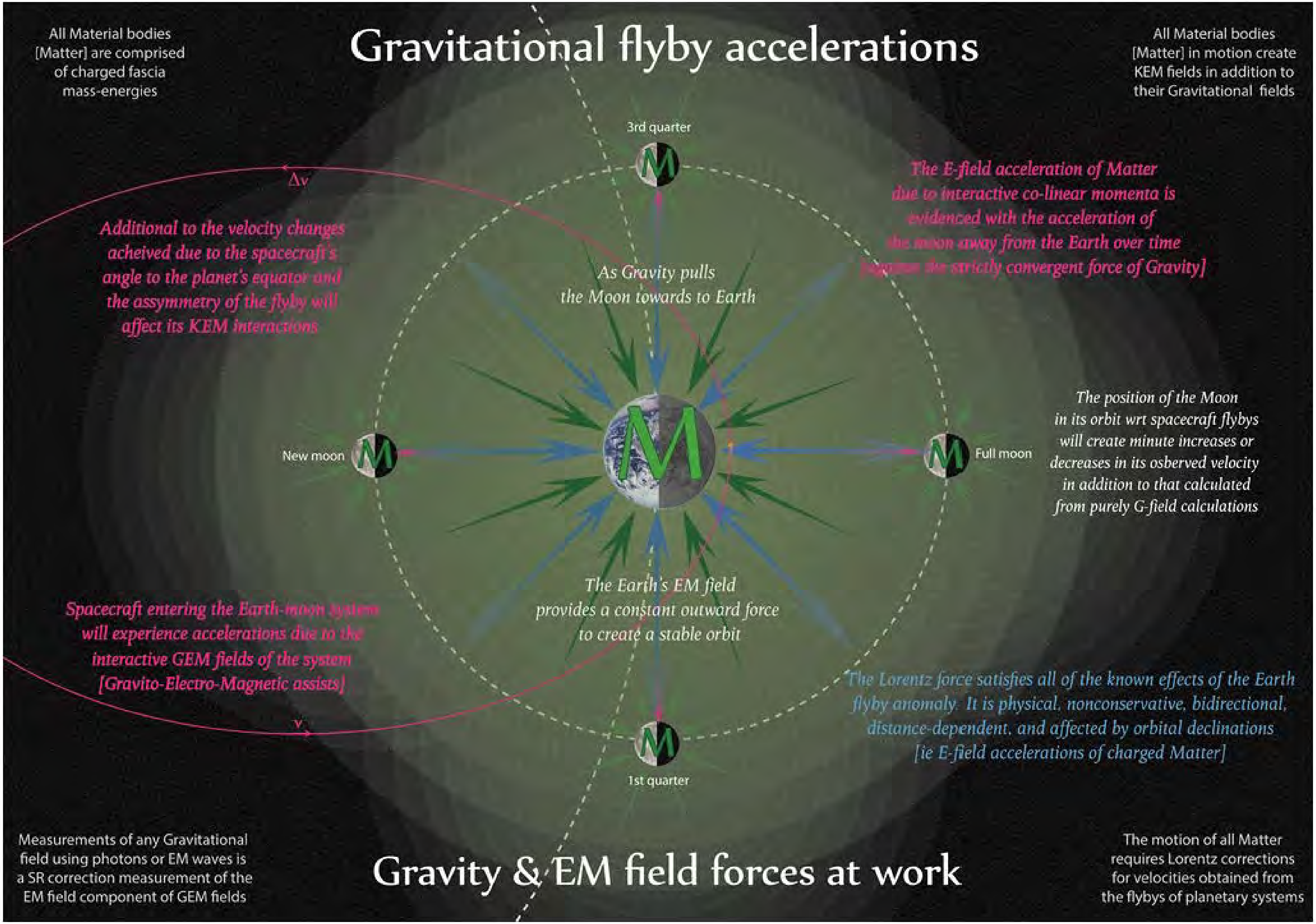

Gravitational flyby accelerations

All Material bodies
[Matter] are comprised
of charged fascia
mass-energies

All Material bodies
[Matter] in motion create
KEM fields in addition to
their Gravitational fields

*The E-field acceleration of Matter
due to interactive co-linear momenta is
evidenced with the acceleration of
the moon away from the Earth over time
[against the strictly convergent force of Gravity]*

*Additional to the velocity changes
acheived due to the spacecraft's
angle to the planet's equator and
the assymmetry of the flyby will
affect its KEM interactions*

*As Gravity pulls
the Moon towards to Earth*

*The position of the Moon
in its orbit wrt spacecraft flybys
will create minute increases or
decreases in its osberved velocity
in addition to that calculated
from purely G-field calculations*

*The Earth's EM field
provides a constant outward force
to create a stable orbit*

*Spacecraft entering the Earth-moon system
will experience accelerations due to the
interactive GEM fields of the system
[Gravito-Electro-Magnetic assists]*

*The Lorentz force satisfies all of the known effects of the Earth
flyby anomaly. It is physical, nonconservative, bidirectional,
distance-dependent, and affected by orbital declinations
[ie E-field accelerations of charged Matter]*

Measurements of any Gravitational
field using photons or EM waves is
a SR correction measurement of the
EM field component of GEM fields

Gravity & EM field forces at work

The motion of all Matter
requires Lorentz corrections
for velocities obtained from
the flybys of planetary systems

Tetryonics 76.05 - Gravitational flyby accelerations

Flyby anomalies

ALL Matter creates GravitoElectoMagnetic fields

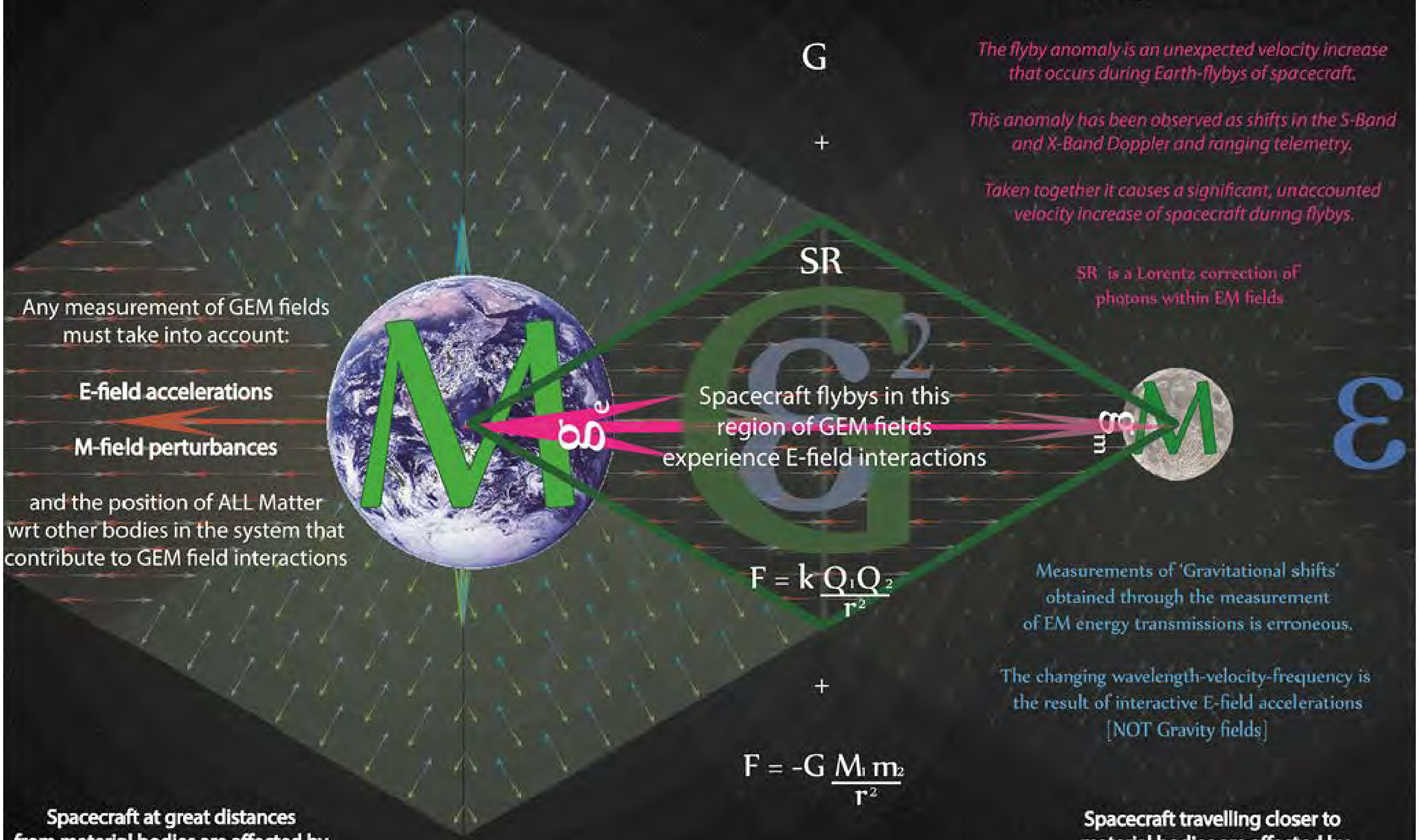

The flyby anomaly is an unexpected velocity increase that occurs during Earth-flybys of spacecraft.

This anomaly has been observed as shifts in the S-Band and X-Band Doppler and ranging telemetry.

Taken together it causes a significant, unaccounted velocity increase of spacecraft during flybys.

SR is a Lorentz correction of photons within EM fields

Any measurement of GEM fields must take into account:

E-field accelerations

M-field perturbances

and the position of ALL Matter wrt other bodies in the system that contribute to GEM field interactions

Spacecraft flybys in this region of GEM fields experience E-field interactions

$$F = k\frac{Q_1Q_2}{r^2}$$

$$F = -G\frac{M_1m_2}{r^2}$$

Measurements of 'Gravitational shifts' obtained through the measurement of EM energy transmissions is erroneous.

The changing wavelength-velocity-frequency is the result of interactive E-field accelerations [NOT Gravity fields]

Spacecraft at great distances from material bodies are affected by convergent Gravitational fields [Newtonian gravity]

Spacecraft travelling closer to material bodies are affected by interactive E-fields as well as G-fields [General Relativity]

Pioneer anomaly

Despite many proposed solutions,
there is not yet an accepted explanation
for the cause of the Pioneer velocity anomaly

Newtonian Gravitation

$$F = -G \frac{M_1 m_2}{r^2}$$

measures the nett gravitational attraction
of Matter towards other Matter
[using weak [K]EM fields]

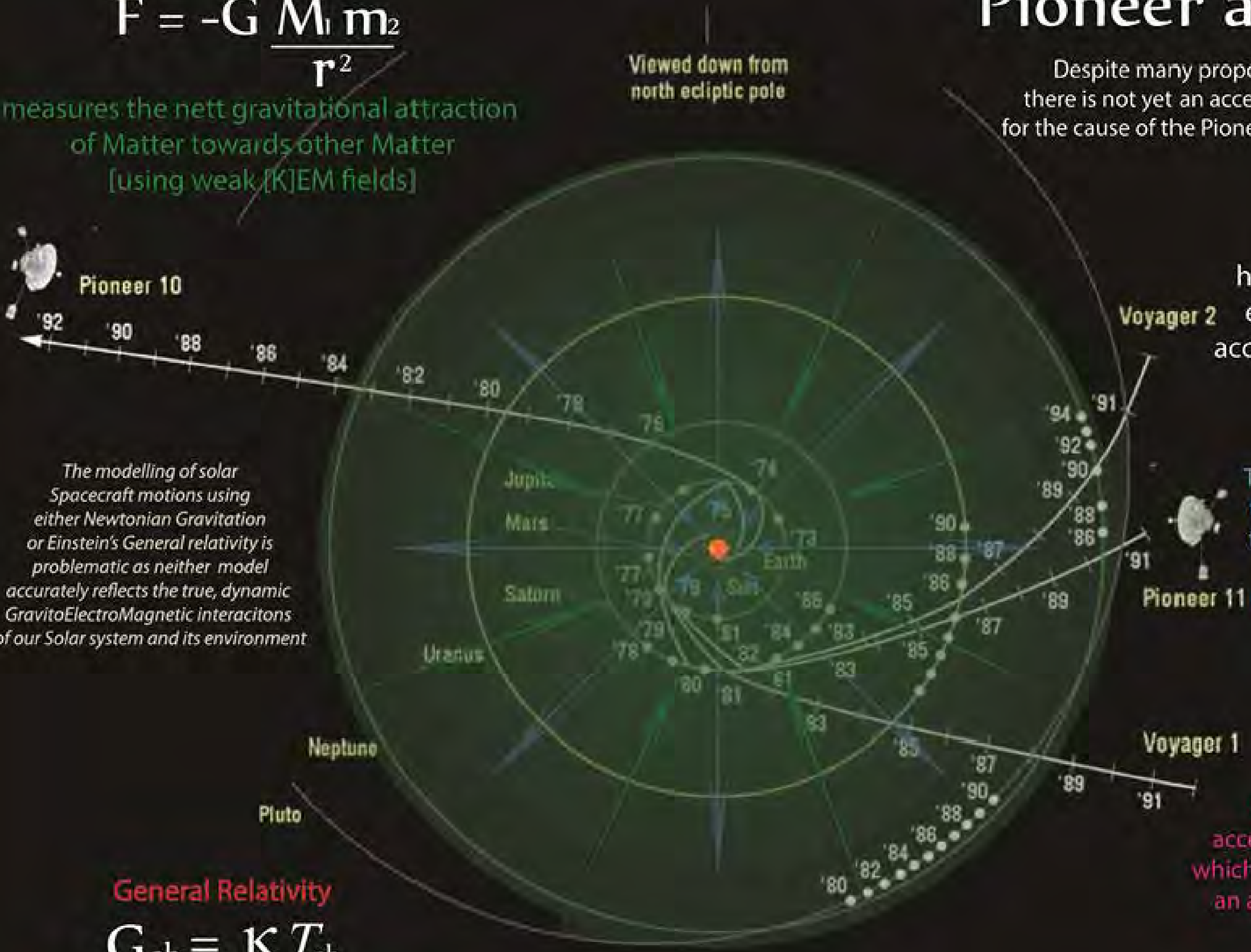

The modelling of solar
Spacecraft motions using
either Newtonian Gravitation
or Einstein's General relativity is
problematic as neither model
accurately reflects the true, dynamic
GravitoElectroMagnetic interacitons
of our Solar system and its environment

Many proposals
have been put forward to
explain this unexpected
acceleration towards the Sun

Tetryonic GEM interactions shows
that the measured acceleration is
the result of a weakening, inverse
squared stellar EM interaction
throughout the Solar heliopause

This would result in
a decrease in divergent
acceleration away from the Sun
which has been mis-interpretted as
an acceleration toward the Sun

General Relativity

$$G_{ab} = \kappa T_{ab}$$

calculates for the total Gravitational and
EM interactions for energy in all its forms
[using Energy pressure gradients]

"Both Pioneer spacecraft are escaping the Solar System, and are slowing under the influence of the Sun's nett gravitational [GEM] field.

Upon very close examination of navigational data, the spacecraft were found to be slowing slightly more than expected.
The effect is an extremely small but unexplained acceleration towards the Sun, of 8.74±1.33×10−10 m/s2"

Tetryonics 76.07 - Pioneer anomaly

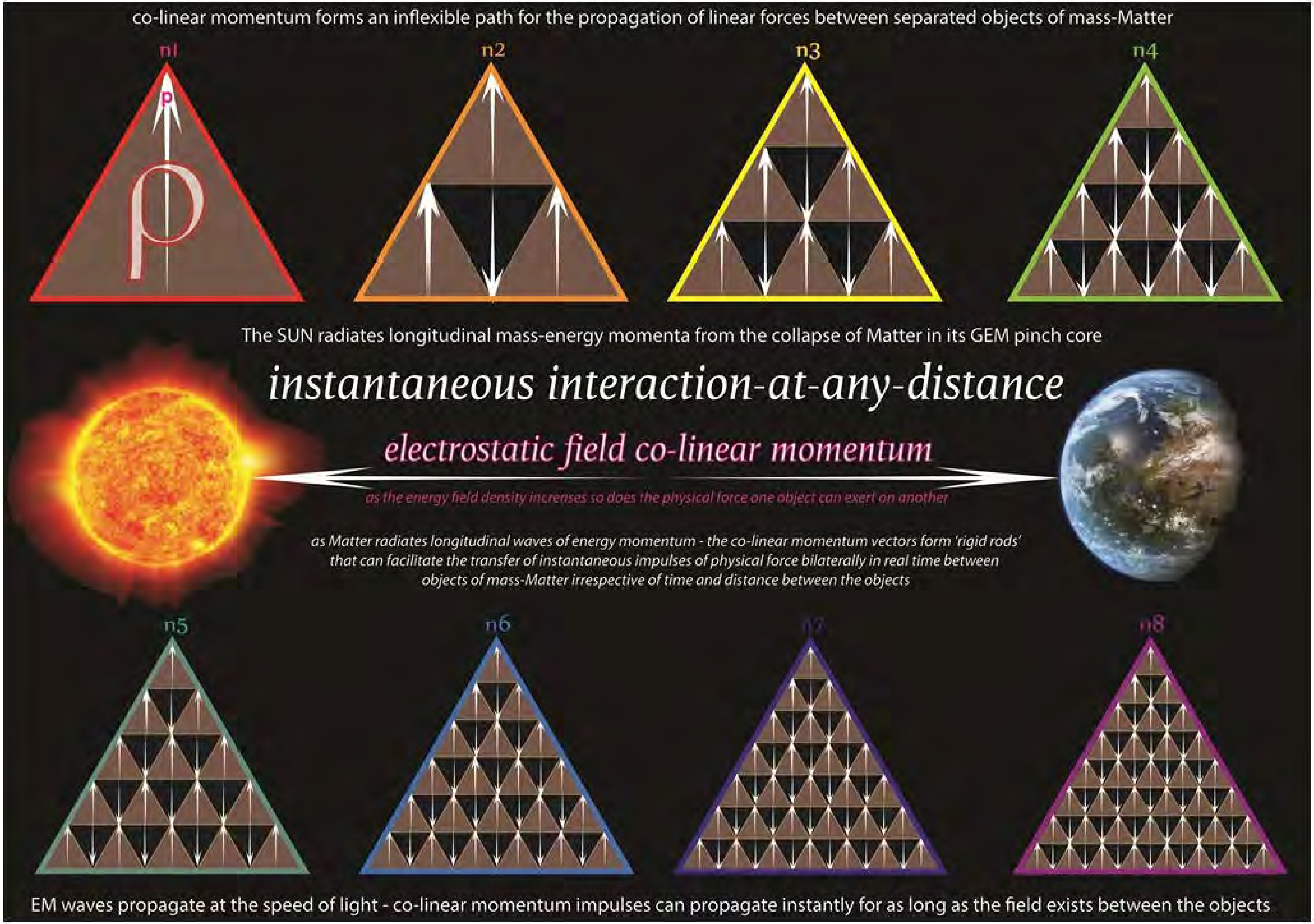

co-linear momentum forms an inflexible path for the propagation of linear forces between separated objects of mass-Matter

The SUN radiates longitudinal mass-energy momenta from the collapse of Matter in its GEM pinch core

instantaneous interaction-at-any-distance

electrostatic field co-linear momentum

as the energy field density increases so does the physical force one object can exert on another

as Matter radiates longitudinal waves of energy momentum - the co-linear momentum vectors form 'rigid rods' that can facilitate the transfer of instantaneous impulses of physical force bilaterally in real time between objects of mass-Matter irrespective of time and distance between the objects

EM waves propagate at the speed of light - co-linear momentum impulses can propagate instantly for as long as the field exists between the objects

Tetryonics 76.08 - Co-linear momentum [interaction-at-a-distance]

Instantaneous interaction-at-a-distance

One of the major points of disagreement between Newtonian Graviation and General Relativity concerns Einstein's limiting of information and energy momenta propagation to the speed of light whereas the Newtonian physics for Gravitation acts instantly between Matter irrespective of distance and time

partilces interact with others in real time irrespective of distance

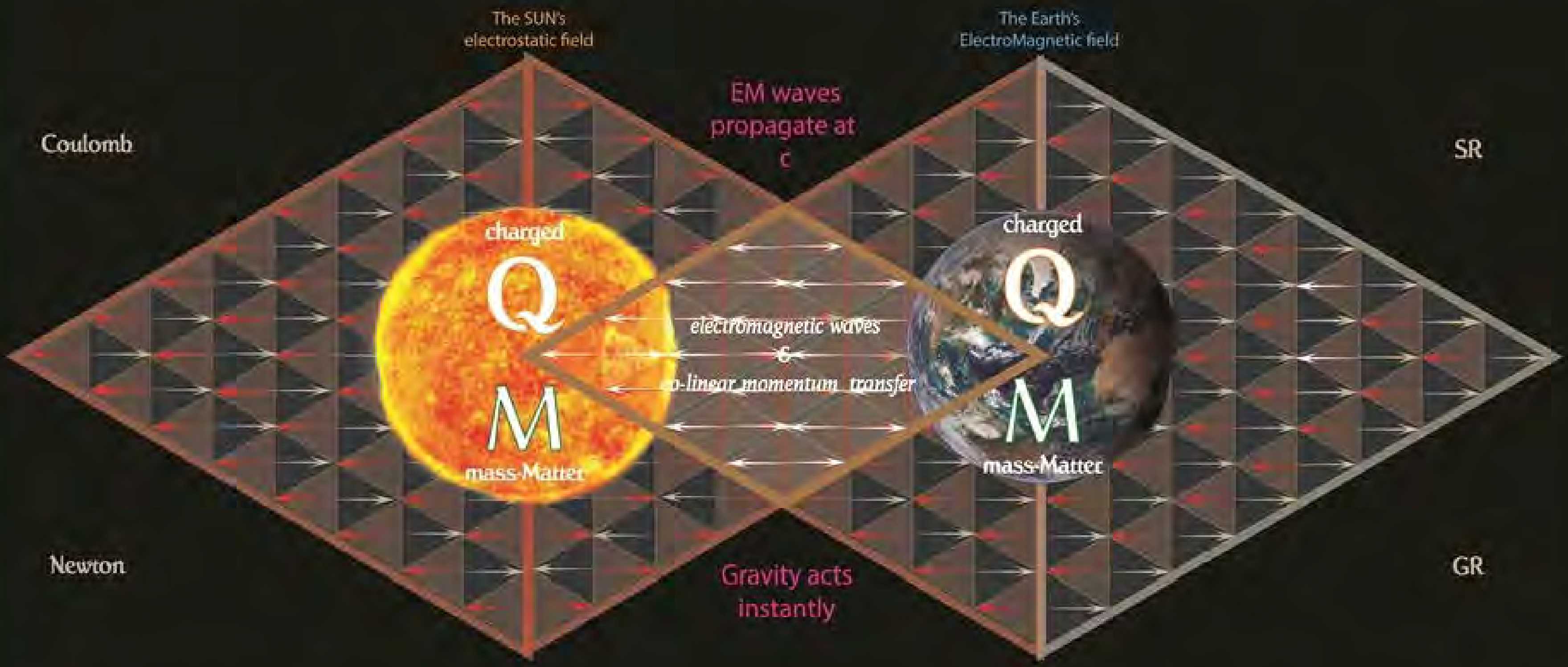
The SUN's electrostatic field
The Earth's ElectroMagnetic field
Coulomb
EM waves propagate at c
SR
charged
Q
M
mass-Matter
electromagnetic waves & co-linear momentum transfer
charged
Q
M
mass-Matter
Newton
Gravity acts instantly
GR

the flow of energy and information is limited to the speed of light

Where electrostatic and/or electromagnetic fields overlap they form fields of interaction between bodies of Matter

the 'quantum rods' of co-linear momentum in their EM fields facilitate 'FTL spooky interaction-at-a-distance' facilitating near instantaneous [real time] impulses of linear momentum to convey physical forces over vast distances without loss for as long as the fields exist between the objects

The efficient wireless transmission of electrical energy requires the creation of a large electrostatic field of longitudinal energy momenta

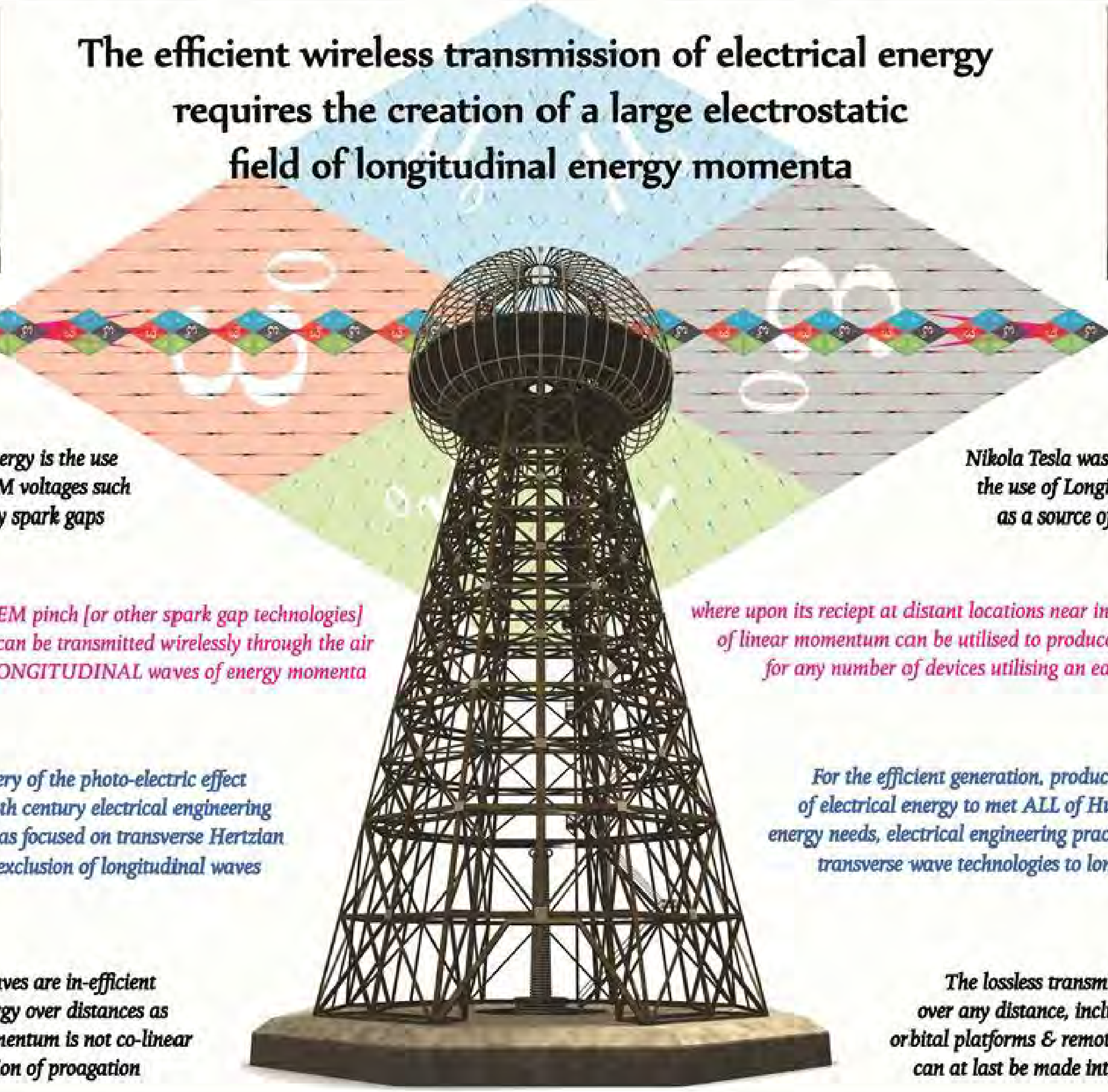

The key to wireless energy is the use of LONGITUDINAL EM voltages such as those produced by spark gaps

Nikola Tesla was the first to promote the use of Longitudinal EM waves as a source of wireless energy

Once generated by EM pinch [or other spark gap technologies] real electrical power can be transmitted wirelessly through the air without loss using LONGITUDINAL waves of energy momenta

where upon its reciept at distant locations near instanteous impulses of linear momentum can be utilised to produce electrical work for any number of devices utilising an earth return

Since the discovery of the photo-electric effect at the turn of the 19th century electrical engineering and physics theory has focused on transverse Hertzian waveforms to the exclusion of longitudinal waves

For the efficient generation, production and storage of electrical energy to met ALL of Humanities growing energy needs, electrical engineering practices must move from transverse wave technologies to longitudinal waves

Transverse EM waves are in-efficient for transmitting energy over distances as their linear energy momentum is not co-linear with their direction of proagation

The lossless transmission of power over any distance, including to and from orbital platforms & remote locations on Earth, can at last be made into a physical reality

Super-luminal communications

Creating a near instantaneous real-time communications network

Once established the linear energy momenta of a scalar electric field

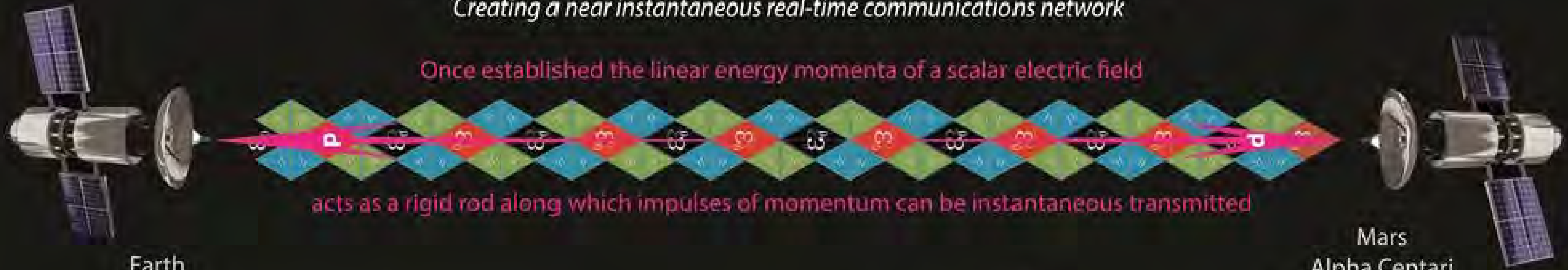

acts as a rigid rod along which impulses of momentum can be instantaneous transmitted

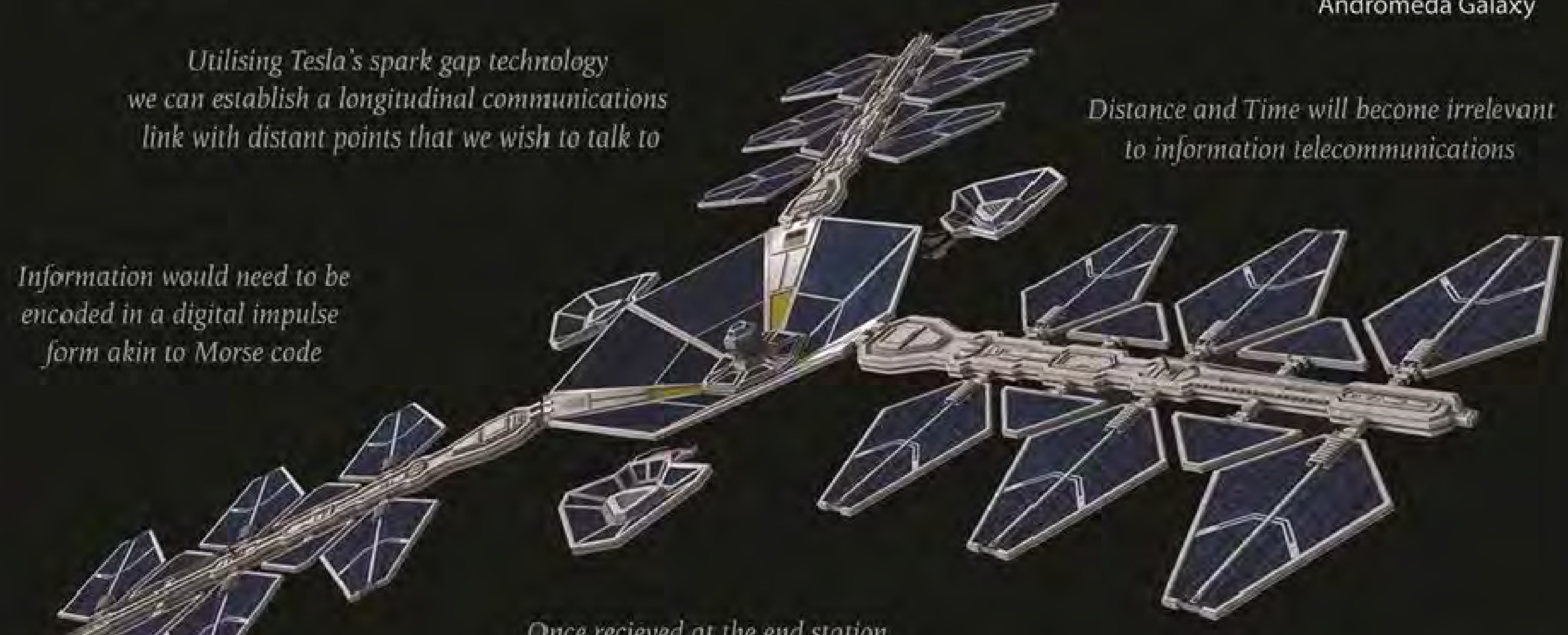

Utilising Tesla's spark gap technology we can establish a longitudinal communications link with distant points that we wish to talk to

Distance and Time will become irrelevant to information telecommunications

Information would need to be encoded in a digital impulse form akin to Morse code

Once recieved at the end station the information could be decoded and rebroadcast as transverse waves to the local stations

Orthagonal electromagnetic waves can also propagate along the rigid rod of momentum but only at the speed of Light

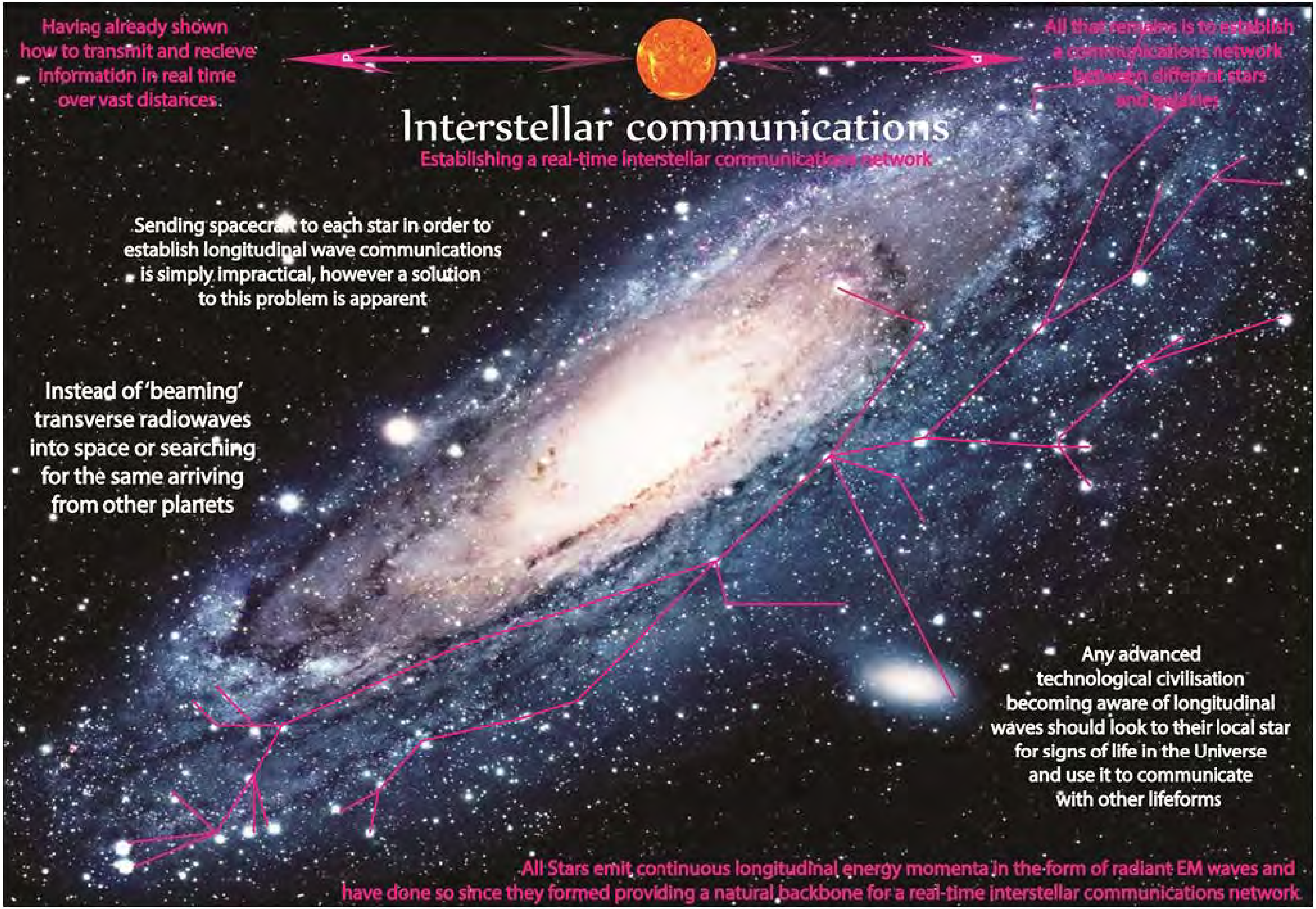

Having already shown
how to transmit and recieve
information in real time
over vast distances

Interstellar communications

Establishing a real-time interstellar communications network

All that remains is to establish
a communications network
between different stars
and galaxies

Sending spacecraft to each star in order to
establish longitudinal wave communications
is simply impractical, however a solution
to this problem is apparent

Instead of 'beaming'
transverse radiowaves
into space or searching
for the same arriving
from other planets

Any advanced
technological civilisation
becoming aware of longitudinal
waves should look to their local star
for signs of life in the Universe
and use it to communicate
with other lifeforms

All Stars emit continuous longitudinal energy momenta in the form of radiant EM waves and
have done so since they formed providing a natural backbone for a real-time interstellar communications network

Tetryonics 76.12 - Interstellar communication

Copyright ABRAHAM [2008] - All rights reserved

Mercurial anomalies

Mercury has a weak magnetic field

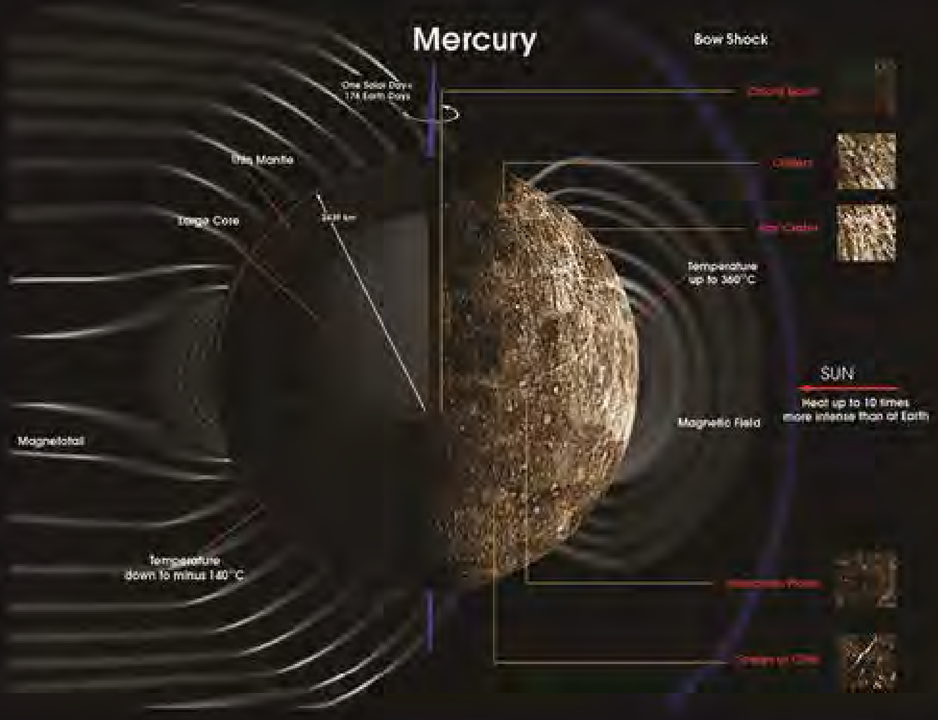
Mercury
Bow Shock
One Solar Day = 176 Earth Days
Thin Mantle
Large Core
Temperature up to 360°C
SUN
Heat up to 10 times more intense than at Earth
Magnetic Field
Magnetotail
Temperature down to minus 140°C

Mercury has heat of its own, not just reflected heat of the Sun;

Mercury has still an atmosphere of Hydrogen, possibly the last vestiges of a more extensive halo and trail (caduceus)

Mercury rotates because it is in its present orbit it is strongly affected by the EM field of the SUN,

Because the ElectroMagnetic forces near the Sun need to be accounted for in the motion of any satellite about a more massive body [ie Satellite-Planet or Planet- Sun], the stronger E & perturbative M forces are responsible for the observed precession of the perihelion of Mercury, and Leverrier's discovery of this precession can be modelled with Einsteins' General relativity through Tetryonic geometries

Laplace showed that should a celestial body attracted by its primary as inverse square of distance be subject to another attraction that changes as the inverse cube of distance, a precession by that body would result.

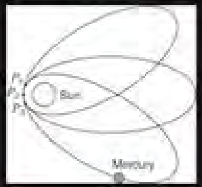
Sun
Mercury

Einstein corrected Newton's gravitational results by calculating for ALL the energies in a system but failed to realise that his 'graviational correction' was in fact a Lorentz correction for EM interactions

Copyright ABRAHAM [2008] - All rights reserved

Venusian anomalies

is sometimes referrd to as Earth's "twin" owing to their similar size, gravity, and bulk composition.

Venus has the densest atmosphere of all the terrestrial planets
in the Solar System, consisting of mostly carbon dioxide.
The atmospheric pressure at the planet's surface is 92 times that of the Earth.

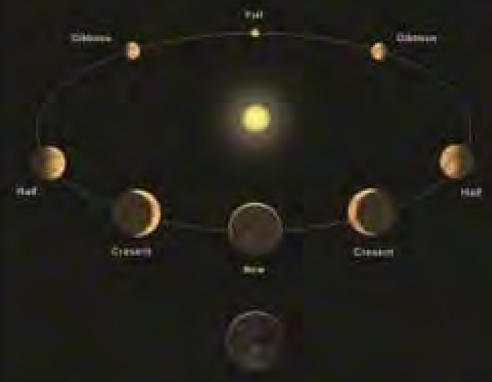

Galileo's discovery that Venus showed phases
(while being never very far from the Sun in our sky)
proved that it orbits the Sun and not the Earth

The Venusian year is about 225 Earth days long.

Venus takes 243 Earth days to rotate on its axis,
which means that days on Venus would be longer
than years. However, because of Venus' curious
retrograde rotation, the time from one sunrise
to the next is only about 117 Earth days

The very top layer of Venus' clouds zip around the planet
every four Earth days, propelled by hurricane-force winds
traveling roughly 224 miles (360 kilometers) per hour.

This super-rotation of the planet's atmosphere, some 60 times faster
than Venus itself rotates, may be one of Venus' biggest mysteries.

**Venus is the hottest world in the solar system.
Although Venus is not the planet closest to the sun,
its dense atmosphere traps heat in a runaway version
of the greenhouse effect that warms up the Earth.**

As a result, temperatures on Venus reach 465 degrees C

The lack of an intrinsic magnetic field at Venus was surprising
given it is similar to Earth in size, and was expected also to
contain a dynamo at its core

Venus has no moons or rings

Copyright ABRAHAM [2008] - All rights reserved

Earth-Moon anomalies

The Moon's Origin

Most scientists think the moon was born from a gargantuan collision, when a young, 30-million-year-old Earth was sideswiped by an embryonic planet the size of Mars some 4.5 billion years ago, with debris from our planet and this impactor eventually coalescing from the ejecta to form our Moon.

The Lunar a lunar day is a month long, the same as the Moon's orbital period

The Moon is slowly accelerating away from the Earth

Athough it is attracted gravitationally to the Earth the Moon is moving away from the Earth at about 4cm per year as a result of the Earth's EM field interactions with Matter in its vicinity.

The Earth's EM field accelerates the moon against the pull of gravity stabilising its orbit

its current field strength forces the moon to slowly accelerate away from the Earth over time

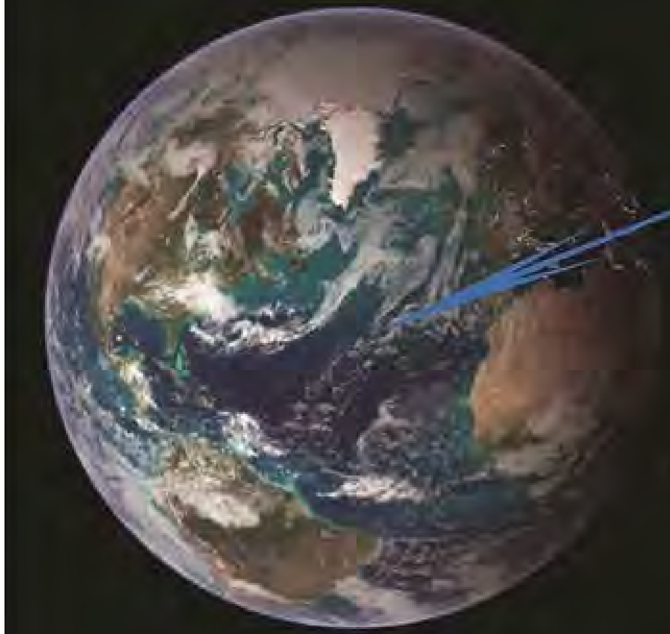

The Moon only shows one face to the Earth

it is gravitational locked in its orbit with one side always facing the Earth when viewed from the Earthat night. There is no such thing as the "dark side of the moon". The sun shines on all sides

The Moon is 400 times smaller than the SUN
The SUN is 400 times more distant

The Moon looks exactly the same size as the Sun

"There is no astronomical reason why the moon and the sun should fit so well. It is the sheerest of coincidences, and only the Earth among all the planets is blessed in this fashion." ....Isaac Asimov

A striking coincidence that has had astronomers thinking since ancient times is why the Sun and the Moon appear to be the same size during an eclipse. It's known that the Sun's diameter is 400 times that of the Moon, but, then again, the Moon is 400 times further away from the Sun.

This creates the optical illusion that the two are actually the same size, allowing us to observe the Solar corona and its EM field lines

Our moon is the only moon in the solar system that has a stationary, near-perfect circular orbit. Stranger still, the moon's center of mass is about 6000 feet closer to the Earth than its geometric center (which should cause wobbling), but the moon's bulge is on the far side of the moon, away from the Earth. "Something" had to put the moon in orbit with its precise altitude, course, and speed.

Copyright ABRAHAM [2008] - All rights reserved

Martian anomalies

There is no completely satisfactory theory as to how Deimos and Phobos came to be in orbit around Mars

Phobos

With a dametre of 6 km Deimos
is one of the smallest known moons
in the solar system

Deimos

Although Mars is much smaller than Earth, [approx 1/4 the size]
its surface area is about the same as the land surface area of Earth

Deimos and Phobos are probably asteroids perturbed
by Jupiter into orbits that allowed them to be captured by Mars

The southern hemisphere of Mars
is predominantly ancient
cratered highlands somewhat
similar to the Moon.

In contrast, the northern hemisphere
consists of plains which are much younger,
lower in elevation and have a much
more complex history.

Mars orbits about
1.5au from the SUN

An abrupt elevation change of several kilometers seems to occur at the boundary

Jupiters' anomalies

Jupiter has twelve satellites.

The five inner moons revolve around their planet in orbits only slightly inclined to the planet's equator at distances from about 110,000 miles for the innermost to about 640,000 miles for the outermost.

Jupiter's mass is 2.5 times that of all the other planets in our Solar System combined

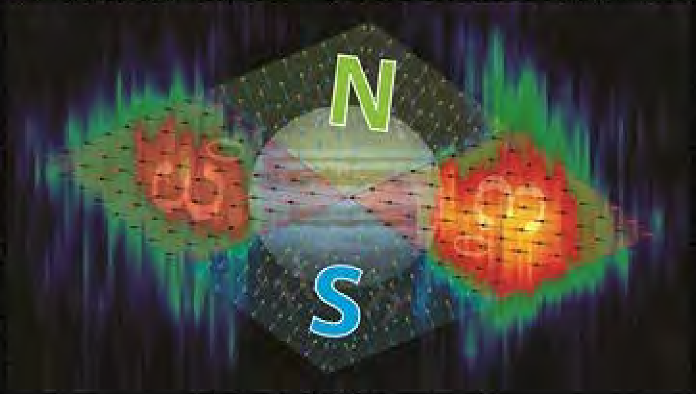

Tetryonic EM geometry
offers answers to many of the dynamic
mysteries of Jupiter and its moons

Then there is a group of three moons whose orbits are inclined to the planet's equator by almost 30° at distances of about 7 million miles from Jupiter.

These three moons also revolve around the planet in the predicted direction.

The four outer moons, however, move around the planet in retrograde motion, or opposite to that of the other eight satellites, at distances from about 12 to 13 million miles from the planet

Jupiter has an intense and far-reaching magnetic field.
Exactly what drives the magnetic powerhouse and
how deep inside the planet it originates, however, is unknown

Jupiter's diameter is approx
1/10 that of the SUN and
10 times the size of Earth

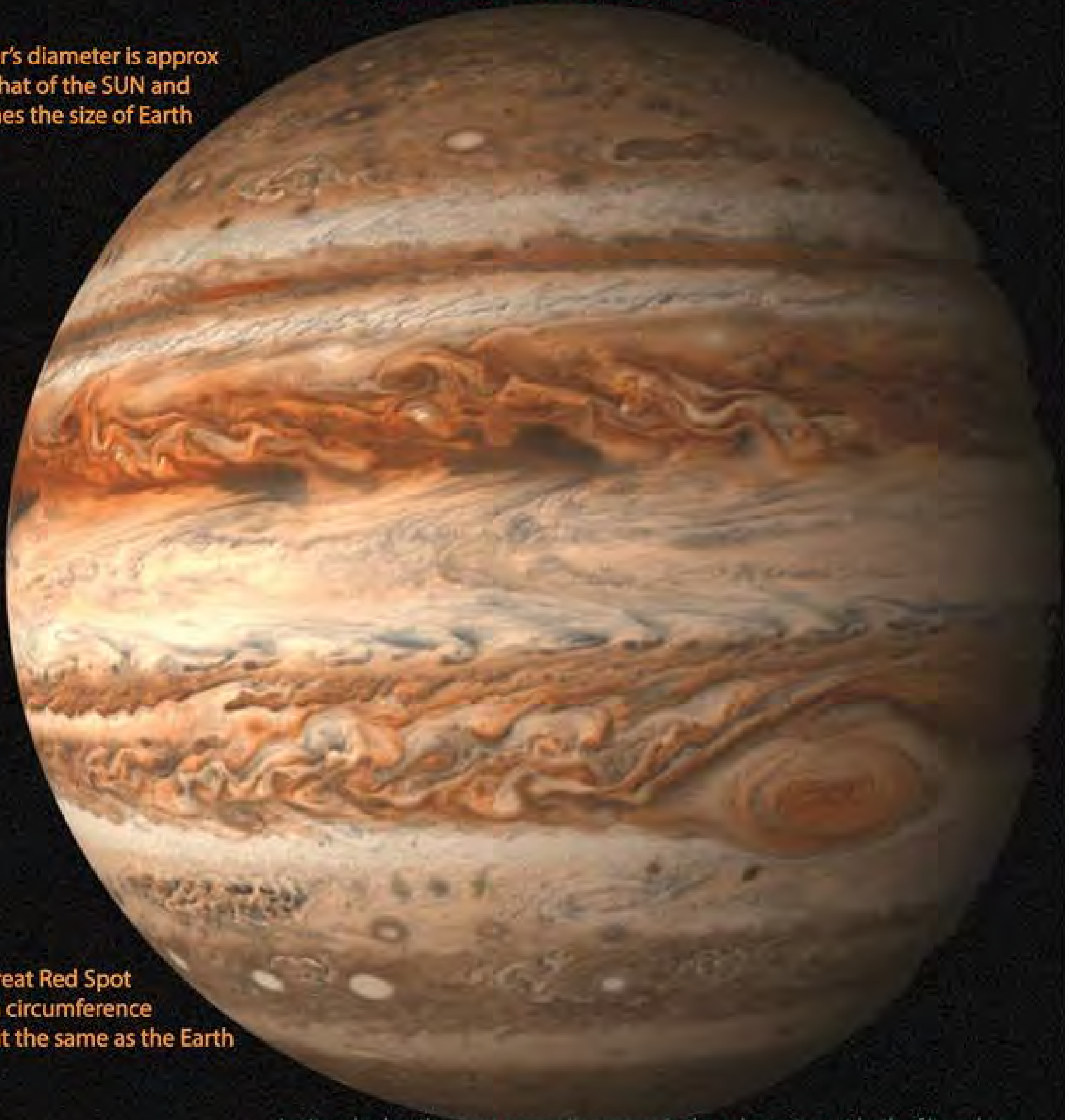

Its Great Red Spot
has a circumference
about the same as the Earth

Jupiter also has the most energetic auroras in the solar system at both of its poles

Saturnian anomalies

In addition to the planet's giant rings, Saturn has nine major moons and more than 60 known satellites.

The motion of the outermost, Phoebe, is retrograde, moving in a direction opposite to the other eight moons and opposite to that predicted from an evolutionary origin.

Tetryonic theory suggests that this retrograde motion is the result of it being a gravitationally captured body whose motion over time will be slowed, eventually forcing it into the same motion as the others

TITAN
is the only moon that
possesses a thick atmosphere, and
it's the only place beyond Earth to have
stable bodies of liquid Methane on its surface

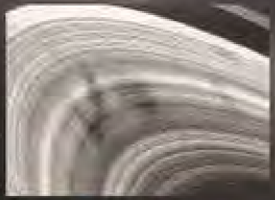

The rings of Saturn occasionally exhibit ghostly radial spokes that seemingly defy the laws of gravity

The Saturnian rings, like many others, are formed by the electric field of the planet and subject to seasonal fluctuations of the SUN's heliosheric E-field

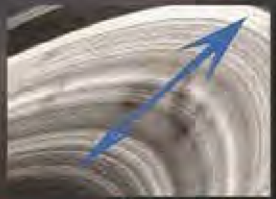

Tetryonic GEM interactions shows that these features are the result of electric forces additional to gravity

Strange features abound within the rings of Saturn. These include hundreds of "record grooves" or narrow fluctuations in the B ring, the most massive ring.

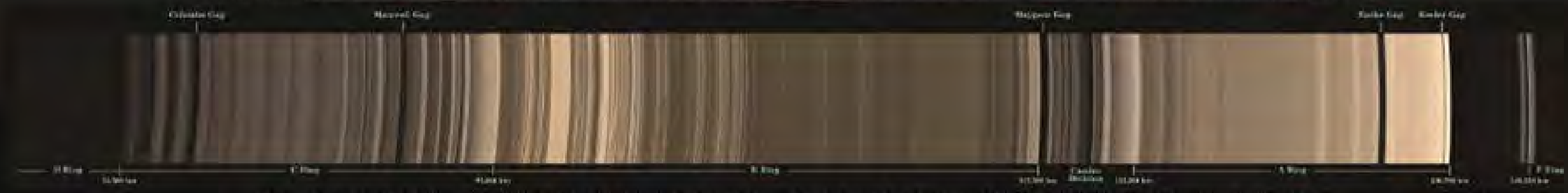

These also include "plateaus" in the C ring, the innermost main ring—broad, dark, sharp-edged features thicker than elsewhere in the ring.

Copyright ABRAHAM [2008] - All rights reserved

Neptunes' anomalies

Neptune was the first planet found by mathematical prediction
rather than by empirical observation

Neptune has two satellites.

Nereid, a small moon, moves around Neptune in the predicted direction, but Triton, one of the larger satellites in the solar system with a mass almost twice that of the earth's moon.

Of Neptune's 13 moons, Triton is by far the biggest and the only one massive enough to be spheroidal.

Weirdly, Triton has a "retrograde" orbit, revolving in the opposite direction of the planet and other moons.

Plus, the orbit is at an angle rather than in the plane around the equator like typical satellites

The magnetic field of Neptune, like that of Uranus,
is highly tilted at 47 degrees from the rotation axis and offset
at least 0.55 radii (about 13,500 kilometers or 8,500 miles) from the physical center

Copyright ABRAHAM [2008] - All rights reserved

Uranus' anomalies

Uranus is remarkable, even though it rotates around the sun in the same direction as the other planets in an orbit inclined less than a degree (46') from the ecliptic (the plane of the earth's orbit around the sun), the axis of rotation of Uranus is nearly in the plane of its orbit.

Accordingly, the inclination of the equator of Uranus to the plane of its orbit is 98°, and its axial rotation is retrograde.

The five moons or satellites of Uranus move exactly in the equatorial plane of the planet and they revolve in the same direction as the planet rotates.

Their motion, with respect to the remainder of the solar system, is, therefore, also retrograde.

Tetryonic theory dictates that Uranus' peculiar planetary mechanics may be the result of the EM interactions of Uranus itself with its satellites rather than the current evolutionary explanation

The direction of the axial rotation of Uranus and the motion of its satellites is opposite to that predicted on the basis of an evolutionary origin.

Pluto's anomalies

Pluto is much smaller than any of the official planets
and due to its size and orbital inclination
is now classified as a "dwarf planet"

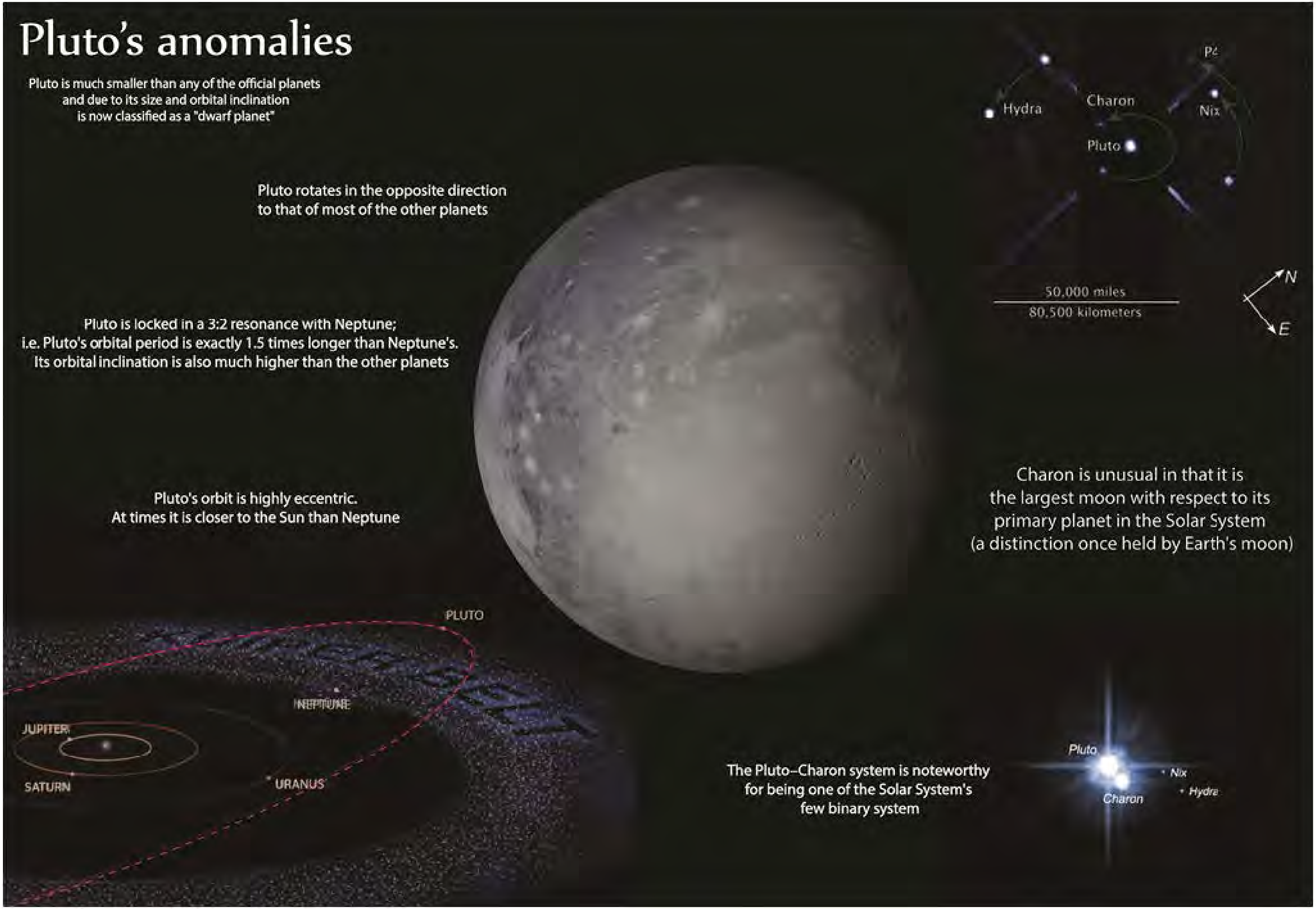

Hydra
Charon
P4
Nix
Pluto
50,000 miles
80,500 kilometers
N
E
PLUTO
NEPTUNE
JUPITER
SATURN
URANUS
Pluto
Nix
Hydra
Charon

Pluto rotates in the opposite direction
to that of most of the other planets

Pluto is locked in a 3:2 resonance with Neptune;
i.e. Pluto's orbital period is exactly 1.5 times longer than Neptune's.
Its orbital inclination is also much higher than the other planets

Pluto's orbit is highly eccentric.
At times it is closer to the Sun than Neptune

Charon is unusual in that it is
the largest moon with respect to its
primary planet in the Solar System
(a distinction once held by Earth's moon)

The Pluto–Charon system is noteworthy
for being one of the Solar System's
few binary system

Copyright ABRAHAM [2008] - All rights reserved

Cometary anomalies

Comets provide some of the clearest evidence of Electromagnetic interactions between the SUN and other planetary bodies

While the solid nucleus of comets is generally less than 50 km across, the coma may be larger than the Sun, and Ion tails have been observed to extend 3-4 astronomical units (5-600 million km)

Cometary discharging may also occur due to any disturbances of its electrical plasma sheath as it passes through regions of varying electric potential.

This seems to have occurred in the recent "totally surprising" outburst of Comet Holmes 17P as it moved away from the Sun's domain.

In 2000 the explosive break up of Comet Linear provoked great amazement as the event occurred well over a hundred million kilometers from the Sun

Other anomalies that can be explained by a CHARGED model of cometary nuclei are: Unexpectedly high temperatures and X-ray emissions from cometary comas

The sharply carved relief of comets [the exact opposite of what is expected under the "dirty snowball" model]

The unexplained ability of a relatively minuscule comet nucleus to hold in place a highly spherical coma, up to millions of miles in diamater, against the force of the solar wind;

Ejection of larger particles and "gravel" that was never anticipated under the idea that comets accreted from primordial clouds of ice, gas, and dust;

A short supply or complete absence of water and other volatiles on comets" nuclei;

The Electric Universe's model of charged solar bodies predicted the occurrence of an electrical flash in advance to the impact of a projectile into the nucleus of Comet Tempel 1 (Deep Impact.)

Comets have two tails:
A Dust tail and
A Ion tail
most recently observed
in comet Hale-Bopp

The Dust tail is the result of reflective ejected Matter following the comet's path throught solar system while the Ion tail is the result of ejected charged particles of Matter and is heavily influenced by the SUN's radiated E fields

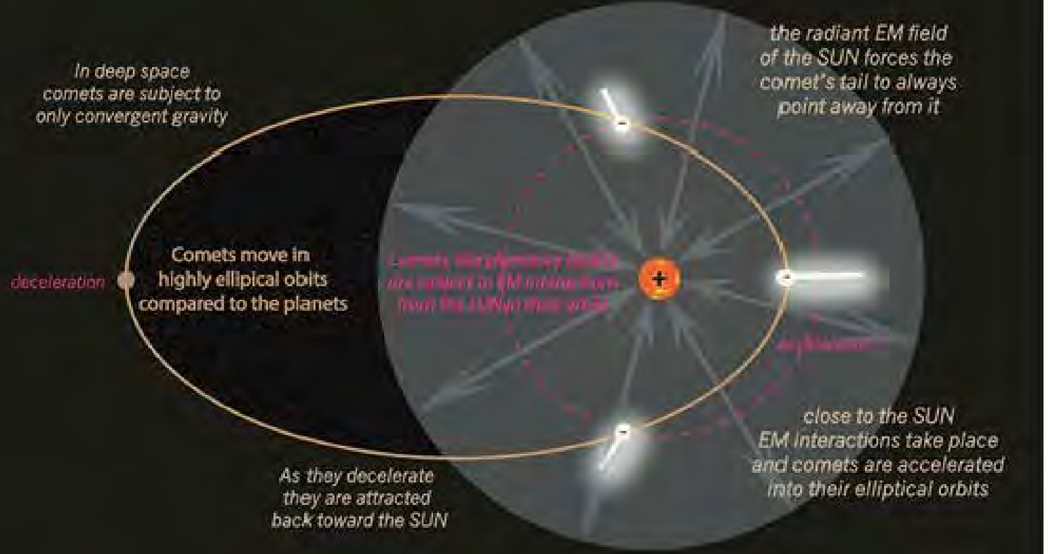

In deep space
comets are subject to
only convergent gravity
the radiant EM field
of the SUN forces the
comet's tail to always
point away from it
deceleration
Comets move in
highly ellipical obits
compared to the planets
Comets like planetary bodies
are subject to EM interactions
from the SUN in their orbit
+
acceleration
close to the SUN
EM interactions take place
and comets are accelerated
into their elliptical orbits
As they decelerate
they are attracted
back toward the SUN

The tails of comets always point away from the Sun, so after a comet has passed the Sun it actually travels tail first

Copyright ABRAHAM [2008] - All rights reserved

The Plasma universe

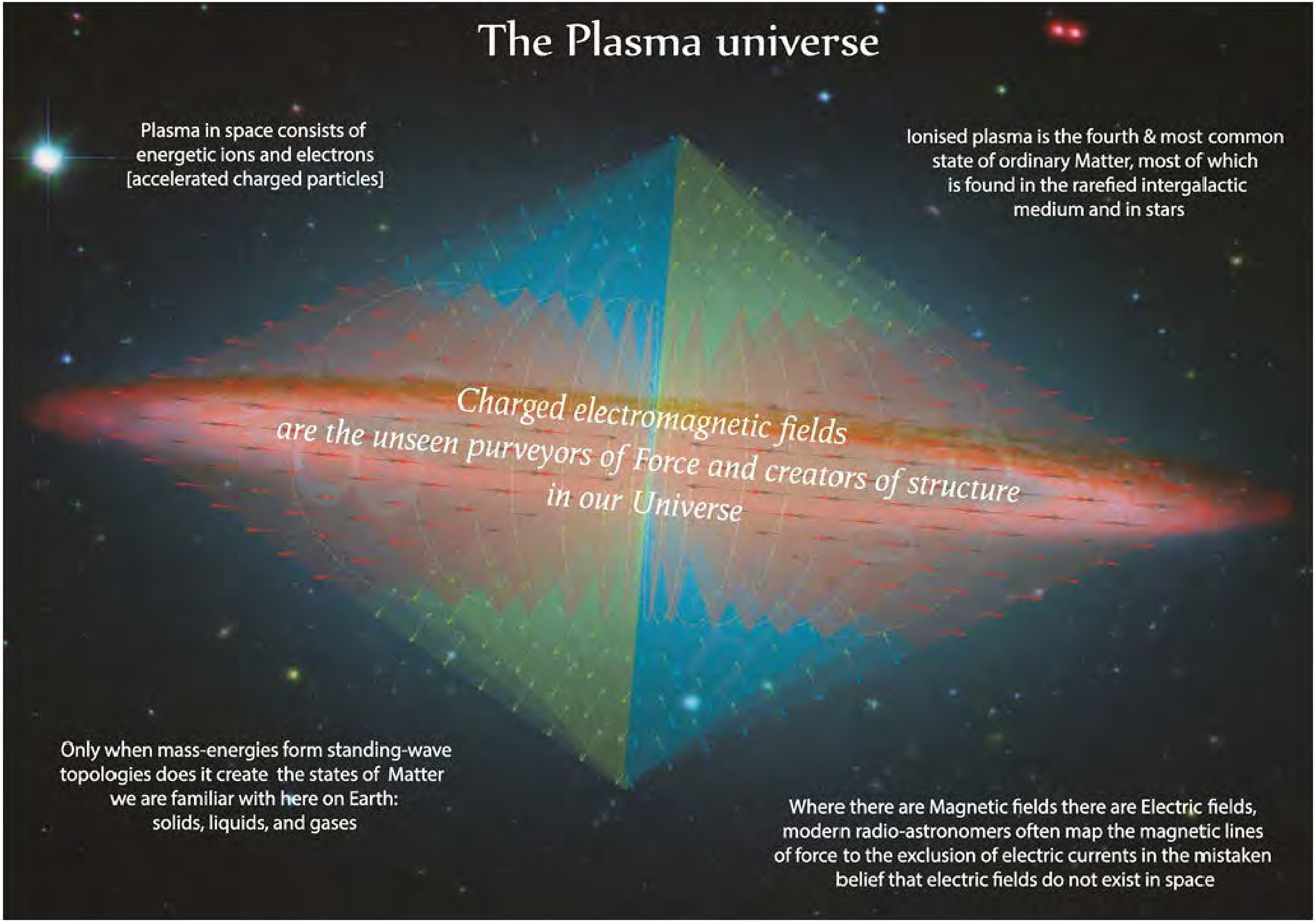

Plasma in space consists of
energetic ions and electrons
[accelerated charged particles]

Ionised plasma is the fourth & most common
state of ordinary Matter, most of which
is found in the rarefied intergalactic
medium and in stars

*Charged electromagnetic fields
are the unseen purveyors of Force and creators of structure
in our Universe*

Only when mass-energies form standing-wave
topologies does it create the states of Matter
we are familiar with here on Earth:
solids, liquids, and gases

Where there are Magnetic fields there are Electric fields,
modern radio-astronomers often map the magnetic lines
of force to the exclusion of electric currents in the mistaken
belief that electric fields do not exist in space

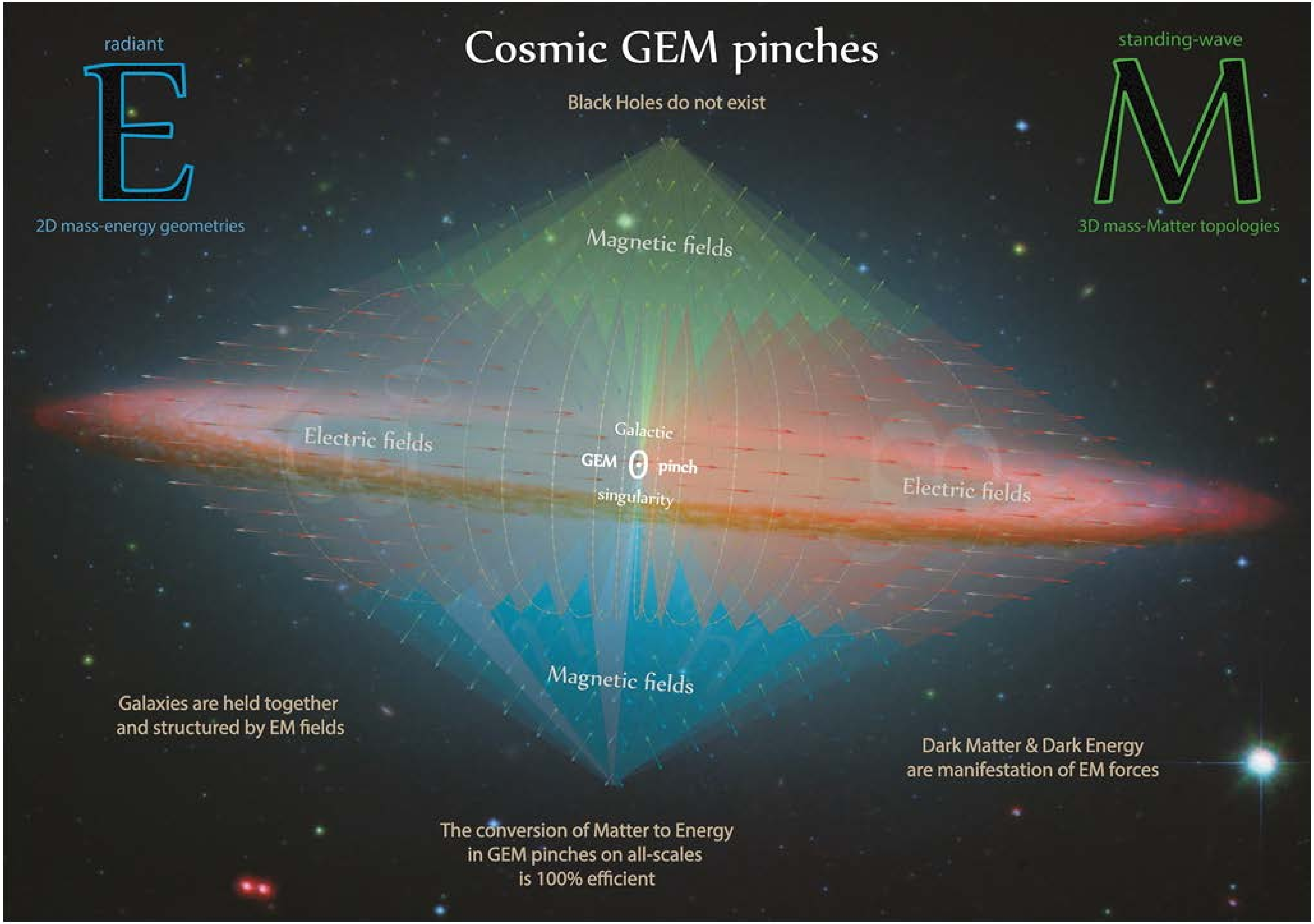

Cosmic GEM pinches

Black Holes do not exist

radiant
E
2D mass-energy geometries
standing-wave
M
3D mass-Matter topologies
Magnetic fields
Electric fields
Galactic
GEM
pinch
singularity
Electric fields
Magnetic fields

Galaxies are held together and structured by EM fields

Dark Matter & Dark Energy are manifestation of EM forces

The conversion of Matter to Energy in GEM pinches on all-scales is 100% efficient

Tetryonics 78.02 - GEM pinches

Copyright ABRAHAM [2008] - All rights reserved

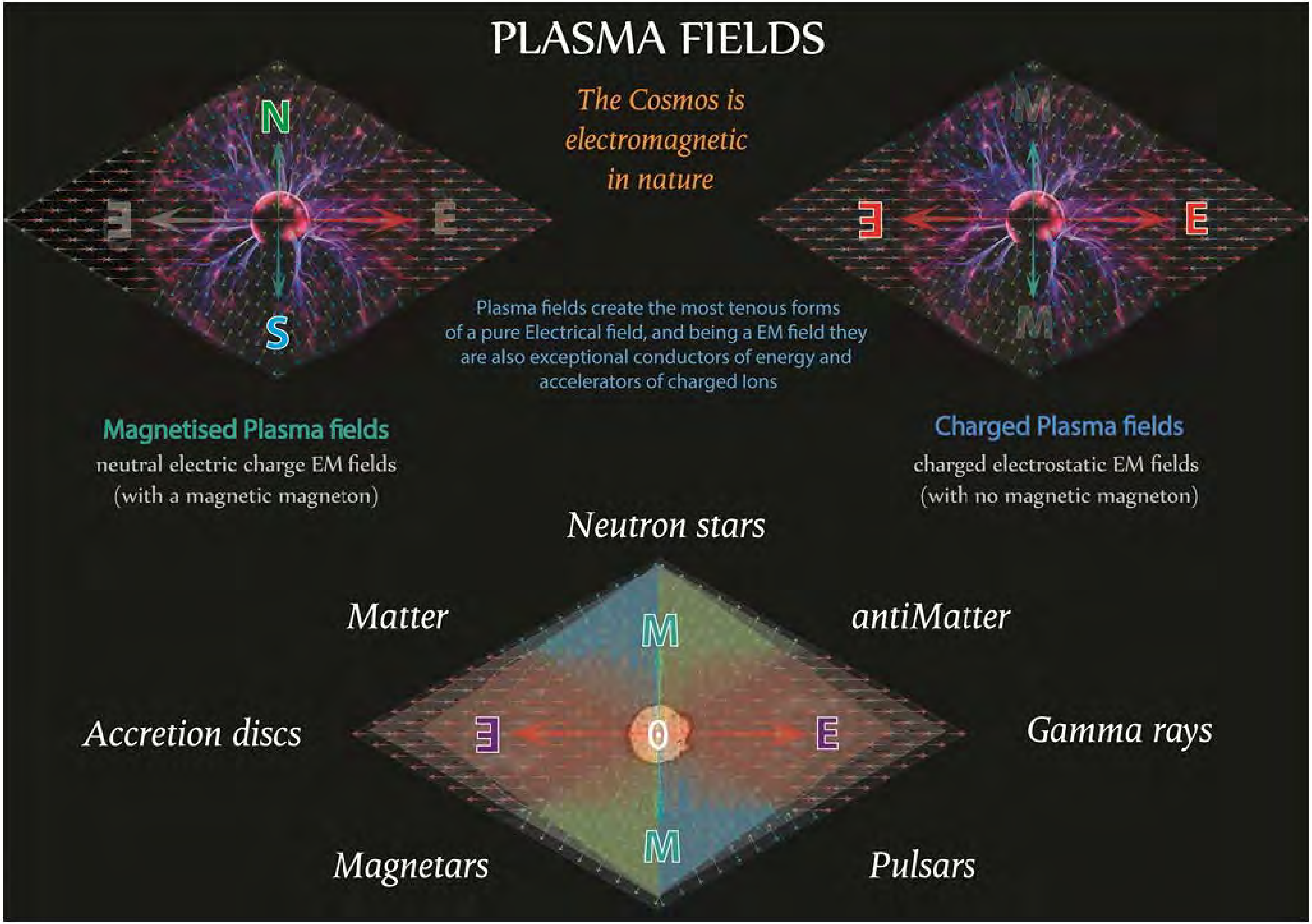

PLASMA FIELDS

The Cosmos is electromagnetic in nature

Plasma fields create the most tenous forms of a pure Electrical field, and being a EM field they are also exceptional conductors of energy and accelerators of charged Ions

Magnetised Plasma fields

neutral electric charge EM fields (with a magnetic magneton)

Charged Plasma fields

charged electrostatic EM fields (with no magnetic magneton)

Neutron stars

Matter

antiMatter

Accretion discs

Gamma rays

Magnetars

Pulsars

Tetryonics 78.03 - Plasma Fields

Copyright ABRAHAM [2008] - All rights reserved

Neutron Stars

All Gravitational fields are comprised of quantum GEM fields

In current astrophysical theories in order to create a

Neutron star

the central region of the star collapses under gravity.

It collapses so much that protons and electrons combine to form neutrons. and neutrogenic Matter

Hence the name "neutron star".

Typical Neutron stars have very strong M-fields

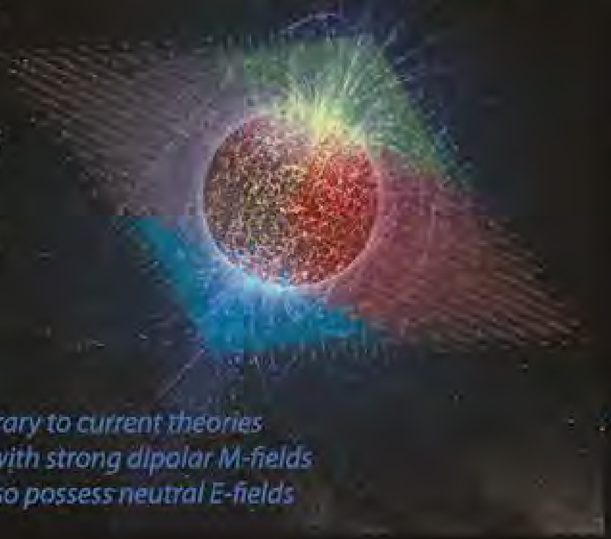

Contrary to current theories any star with strong dipolar M-fields must also possess neutral E-fields

"A neutron star is a type of remnant that can result from the gravitational collapse of a massive star during a Type II, Type Ib or Type Ic supernova event.

Such stars are composed almost entirely of neutrogenic Matter, [Neutrons, antiNeutrons and neutrinos] which against current theory act in exactly the same manner as ordinary Matter in producing spectral lines.

"Neutron stars are very hot and are supported against further collapse because of the Pauli exclusion principle.

This principle states that no two neutrons (or any other fermionic particle) can occupy the same place and quantum state simultaneously'"

As shown throughout Tetryonic theory Neutron stars can be formed comprised completely of 'Neutral Matter'

Neutral Matter is comprised of Neutron - anti-Neutron - Neutrino atomic nuclei

Neutral matter atoms function identically to normal matter-antimatter particle topologies and releases energy in the form of photons

The gravitational collapse of any stellar material must overcome the radiant EM energies that it releases

Pulsars are non-neutral stellar objects that have weak M-fields but emit extremely strong E-field radiation as they collapse

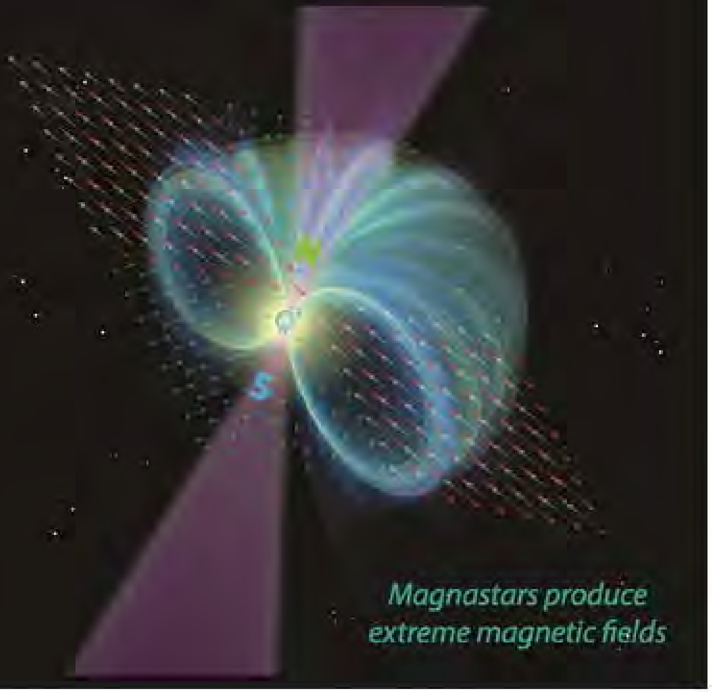

Magnastars produce extreme magnetic fields

Neutral matter in Neutron stars experience the same ElectroMagnetic forces as normal matter and anti-matter stars but have differing charge arrangements

Copyright ABRAHAM [2008] - All rights reserved

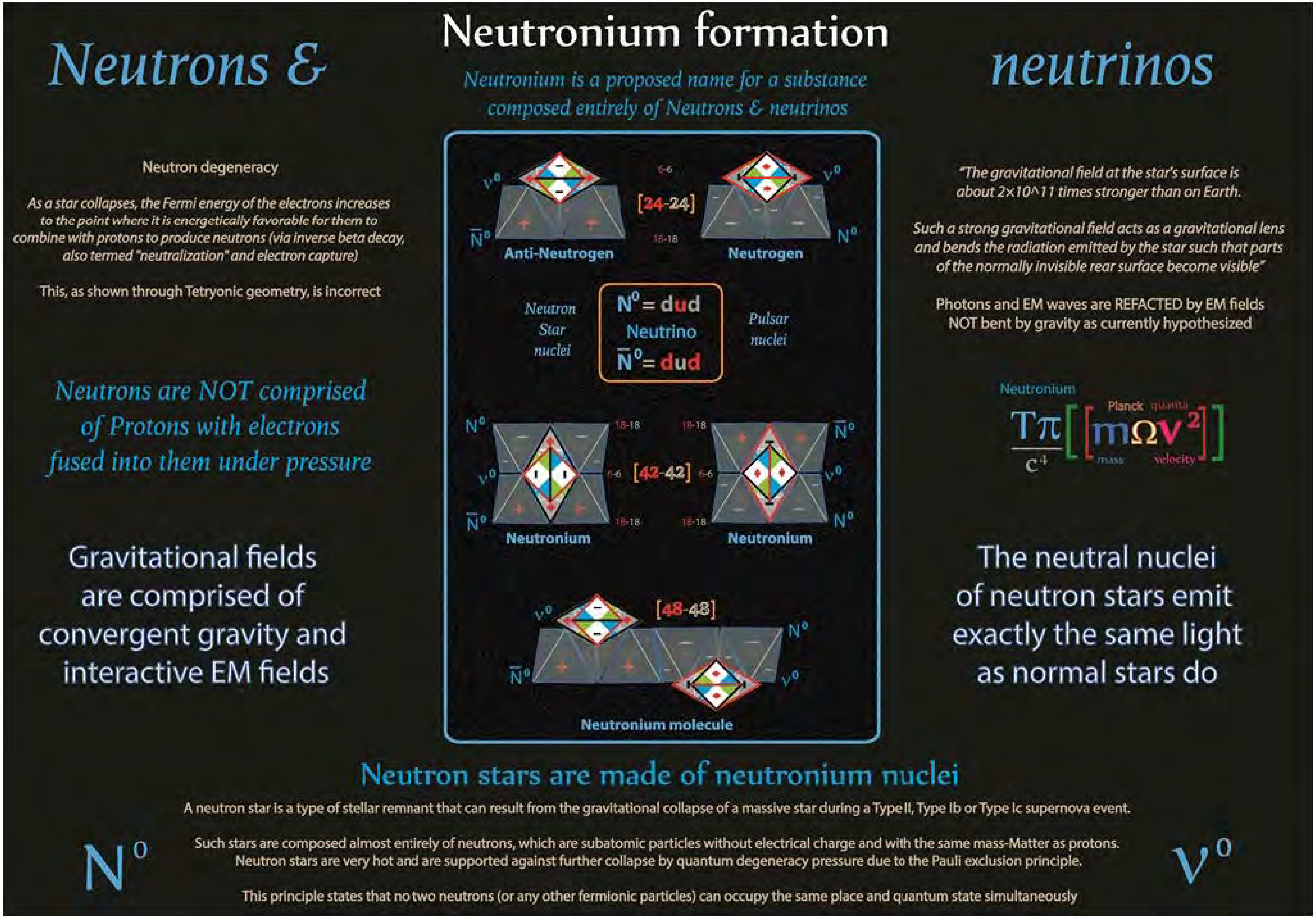

Neutrons & Neutronium formation *neutrinos*

Neutronium is a proposed name for a substance composed entirely of Neutrons & neutrinos

Neutron degeneracy

As a star collapses, the Fermi energy of the electrons increases to the point where it is energetically favorable for them to combine with protons to produce neutrons (via inverse beta decay, also termed "neutralization" and electron capture)

This, as shown through Tetryonic geometry, is incorrect

Neutrons are NOT comprised of Protons with electrons fused into them under pressure

Gravitational fields are comprised of convergent gravity and interactive EM fields

Neutron Star nuclei

$N^0 = dud$

Neutrino

$\bar{N}^0 = dud$

Pulsar nuclei

"The gravitational field at the star's surface is about 2×10^11 times stronger than on Earth.

Such a strong gravitational field acts as a gravitational lens and bends the radiation emitted by the star such that parts of the normally invisible rear surface become visible"

Photons and EM waves are REFACTED by EM fields NOT bent by gravity as currently hypothesized

Neutronium

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

The neutral nuclei of neutron stars emit exactly the same light as normal stars do

Neutron stars are made of neutronium nuclei

A neutron star is a type of stellar remnant that can result from the gravitational collapse of a massive star during a Type II, Type Ib or Type Ic supernova event.

Such stars are composed almost entirely of neutrons, which are subatomic particles without electrical charge and with the same mass-Matter as protons. Neutron stars are very hot and are supported against further collapse by quantum degeneracy pressure due to the Pauli exclusion principle.

This principle states that no two neutrons (or any other fermionic particles) can occupy the same place and quantum state simultaneously

N^0 ν^0

Copyright ABRAHAM [2008] - All rights reserved

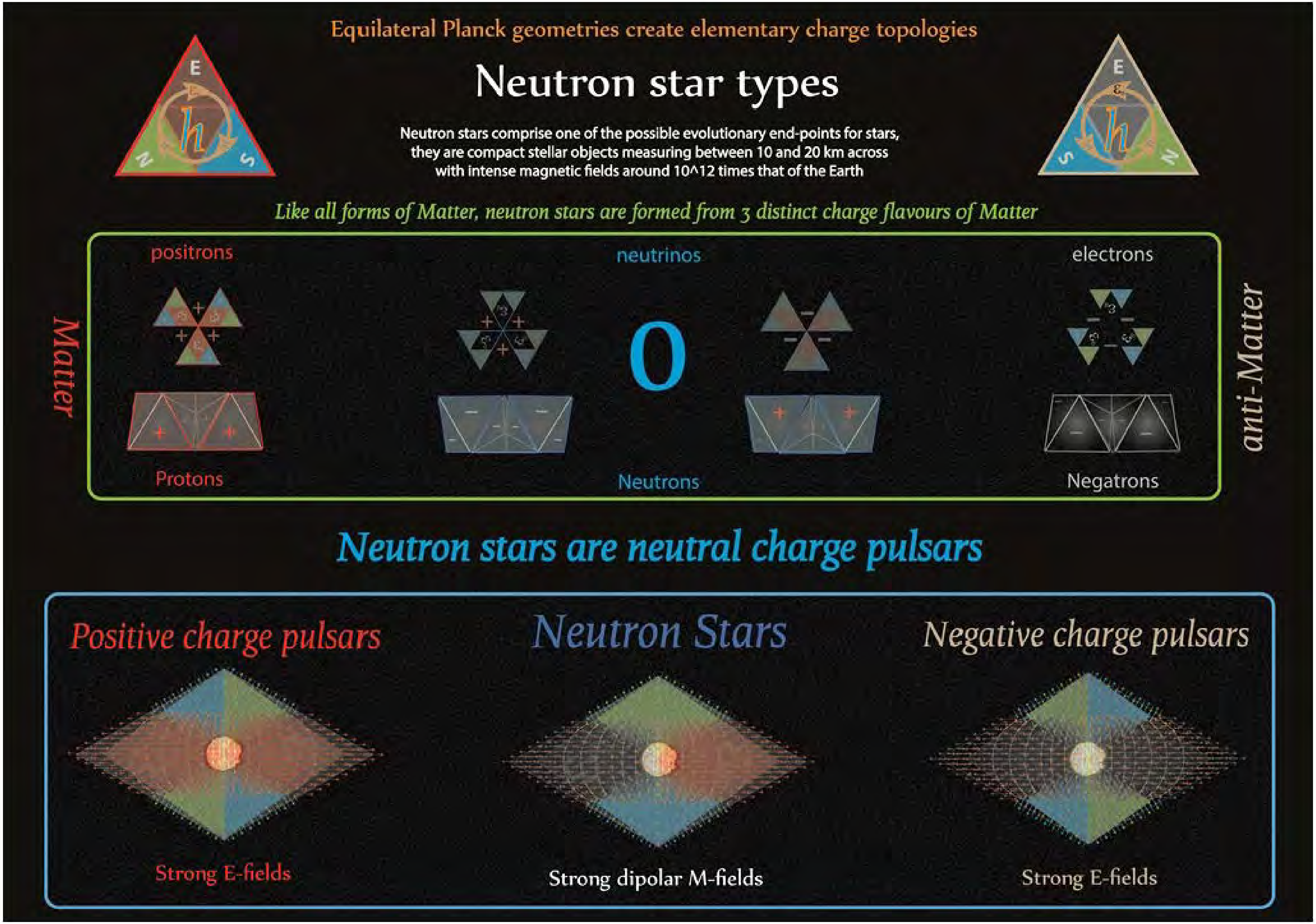

Equilateral Planck geometries create elementary charge topologies

Neutron star types

Neutron stars comprise one of the possible evolutionary end-points for stars, they are compact stellar objects measuring between 10 and 20 km across with intense magnetic fields around 10^12 times that of the Earth

Like all forms of Matter, neutron stars are formed from 3 distinct charge flavours of Matter

Matter

anti-Matter

positrons

neutrinos

electrons

Protons

Neutrons

Negatrons

Neutron stars are neutral charge pulsars

Positive charge pulsars

Strong E-fields

Neutron Stars

Strong dipolar M-fields

Negative charge pulsars

Strong E-fields

Tetryonics 78.06 - Neutron star types

Pulsars

A pulsar is a rotating neutron star that emits a beam of electromagnetic radiation.

This radiation can only be observed when the beam of emission is pointing towards the Earth,
much the way a lighthouse can only be seen when the light is pointed in the direction of an observer,
and is responsible for the pulsed appearance of emission.

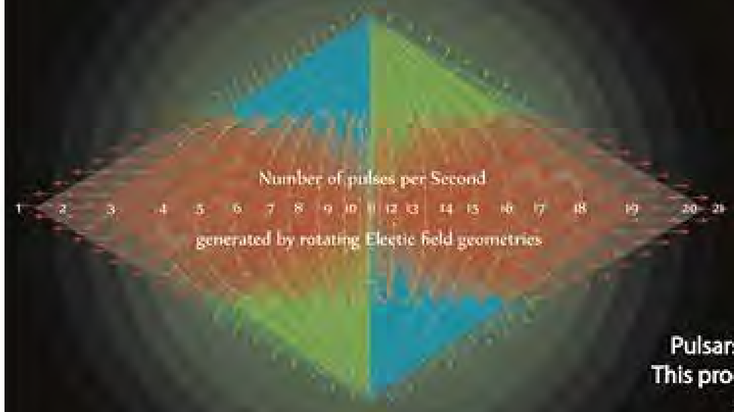
Number of pulses per Second
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
generated by rotating Electic field geometries

Pulsars are electrially charged versions of neutron stars that have weak M-fields in comparision to Neutron stars themselves

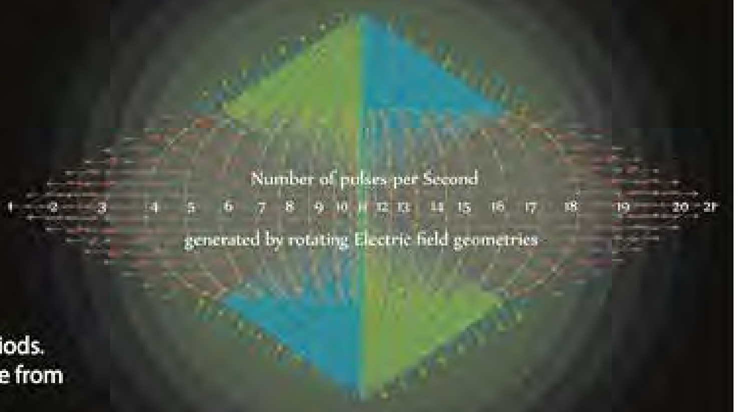
Number of pulses per Second
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
generated by rotating Electric field geometries

Pulsars are very dense, and have short, regular rotational periods.
This produces a very precise interval between pulses that range from
milliseconds to several seconds for individual pulsars

E-field waves accelerate charged particles M-fields do no work

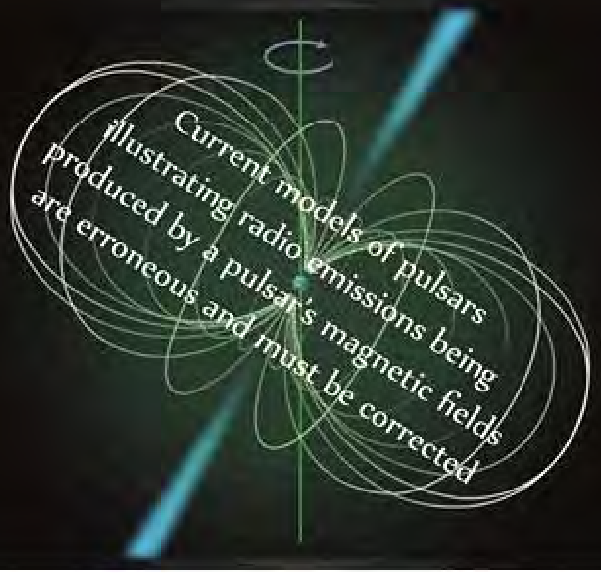

Current models of pulsars illustrating radio emissions being produced by a pulsar's magnetic fields are erroneous and must be corrected

Radio wave astomony detects E-field waves of energy momenta not magnetic waves

Ampere

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

Electric force

Magnetic field

Lorentz

Maxwell

Because a magnetic force is always perpendicular to the motion, the magnetic field can do no work on an isolated charge.

It can only do work indirectly, via the electric field generated by a changing magnetic field

Faraday

Pulsars

Are compact Matter bodies with strong GEM fields

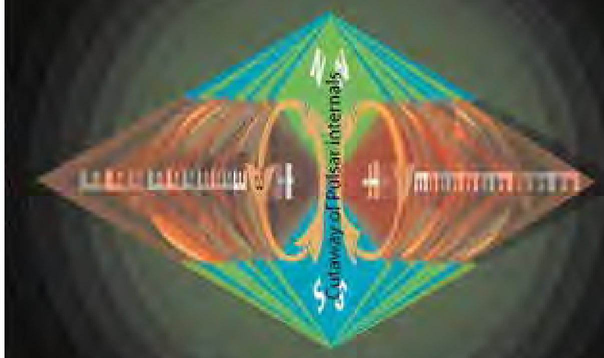

Like all Stellar bodies their GEM fields resolve to a point at their core

The gravitational collapse of Matter produces radiative electromagnetic waves

A millisecond pulsar (MSP) is a pulsar with a rotational period in the range of about 1-10 milliseconds.

Millisecond pulsars have been detected in the radio, X-ray, and gamma ray portions of the electromagnetic spectrum.

The origin of millisecond pulsars is still unknown.

The leading theory is that they begin life as longer period pulsars but are spun up or "recycled" through accretion. For this reason, millisecond pulsars are often called recycled pulsars.

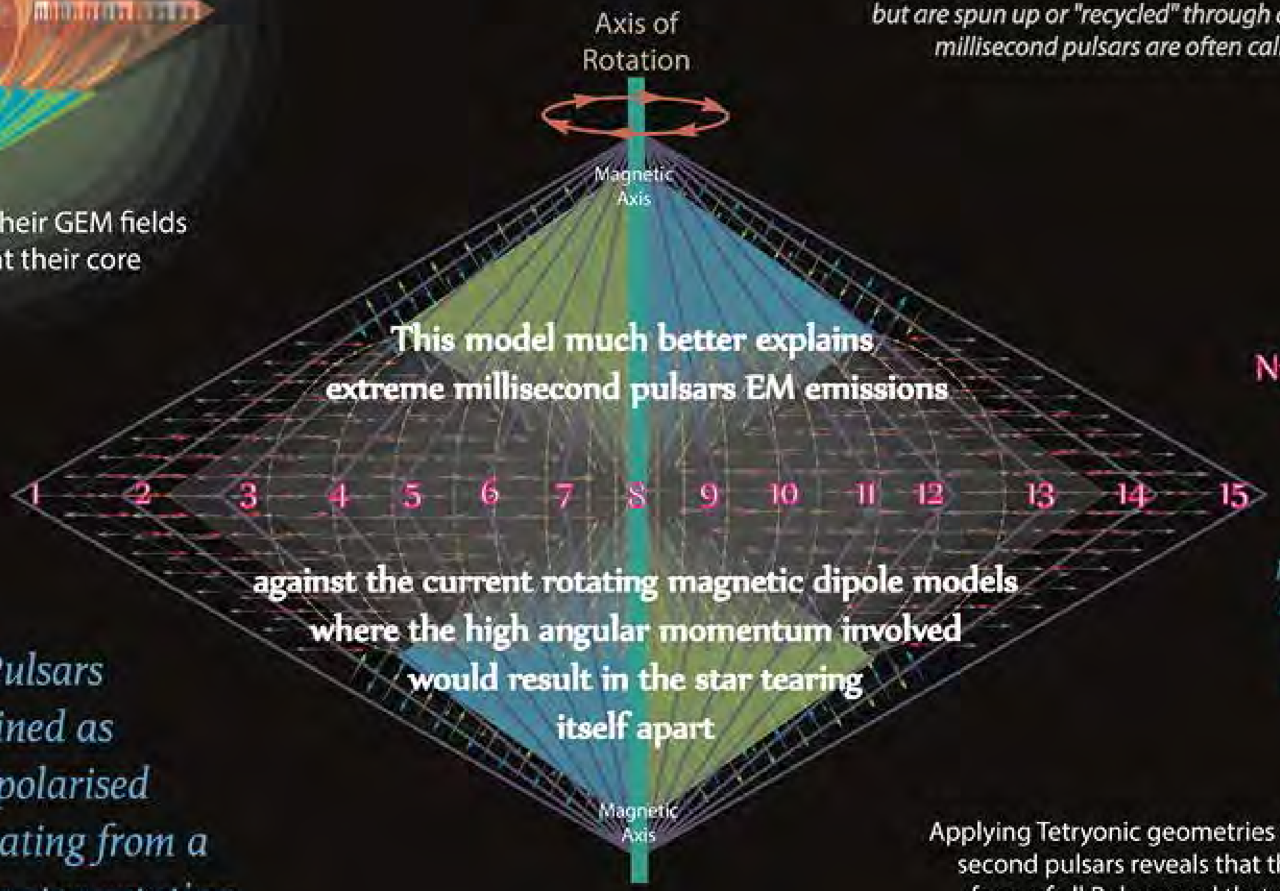

This model much better explains extreme millisecond pulsars EM emissions against the current rotating magnetic dipole models where the high angular momentum involved would result in the star tearing itself apart

Number of Pulses per Second

Radio wave astomony detects E-field energy momenta impulses not magnetic waves

Millisecond Pulsars are best explained as the detection of polarised Electric fields originating from a non-charged Neutron star rotating with high angular momentum

Applying Tetryonic geometries to the question of millisecond pulsars reveals that they are the formative form of all Pulsars and that over time they slow down to become the observed longer period pulsars [obeying the conservation of Angular momentum law]

Tetryonics 78.08 - Pulsar rotations

Copyright ABRAHAM [2008] - All rights reserved

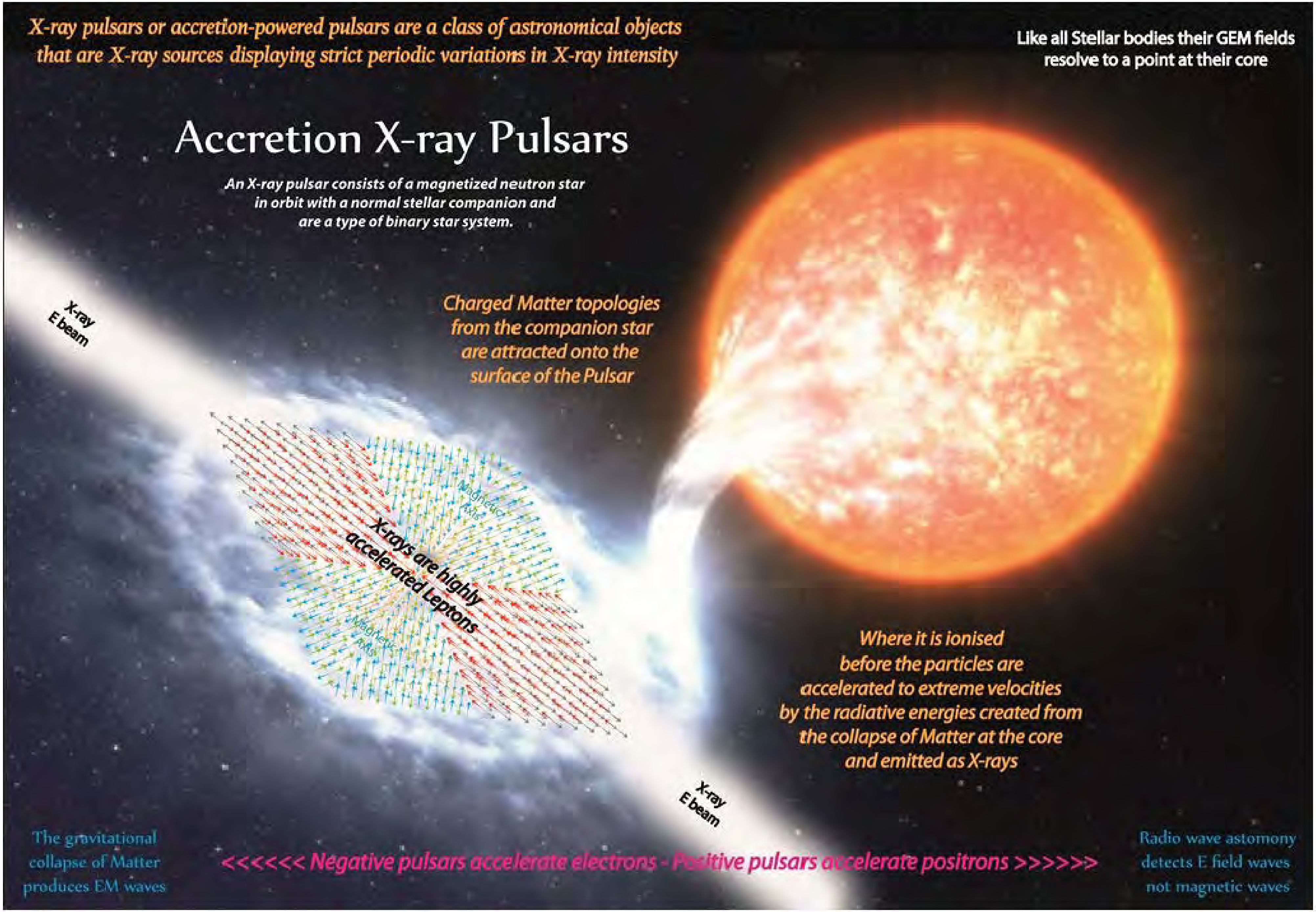

X-ray pulsars or accretion-powered pulsars are a class of astronomical objects that are X-ray sources displaying strict periodic variations in X-ray intensity

Like all Stellar bodies their GEM fields resolve to a point at their core

Accretion X-ray Pulsars

An X-ray pulsar consists of a magnetized neutron star in orbit with a normal stellar companion and are a type of binary star system.

Charged Matter topologies from the companion star are attracted onto the surface of the Pulsar

Where it is ionised before the particles are accelerated to extreme velocities by the radiative energies created from the collapse of Matter at the core and emitted as X-rays

The gravitational collapse of Matter produces EM waves

<<<<<< Negative pulsars accelerate electrons - Positive pulsars accelerate positrons >>>>>>

Radio wave astomony detects E field waves not magnetic waves

Tetryonics 78.09 - Accretion X-ray Pulsars

Stellar Novae

Novae are cataclysmic nuclear explosions in a white dwarf star.

It is caused by the accretion of hydrogen on to the surface of the star, which ignites and starts a runaway nuclear 'fusion' process

All stellar 'fusion' processes are in fact GEM pinches that convert standing wave Matter into radiant mass-energy

Novae are cataclysmic stellar phenomena that take place in binary systems consisting of a compact stellar object and a low-mass star.

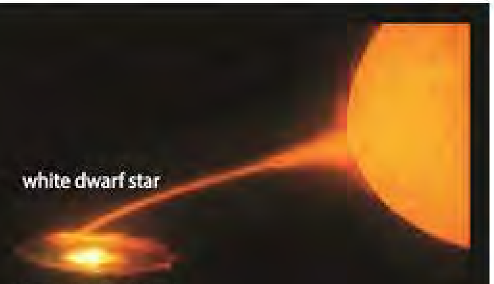

white dwarf star

the size of a planet
but with a mass of up to
1.4 times that of the Sun

The stars must be close enough for the intense gravitational field of the white dwarf to tear material away from the outer layers of its companion.

photons of energy momenta
form kEM fields that accelerate
ions and electrons

spectral
line transitions

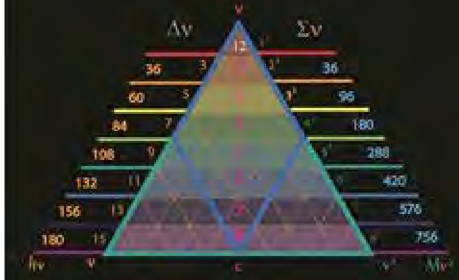

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

all elemental spectral lines
are the result of accelerating
photo-electrons

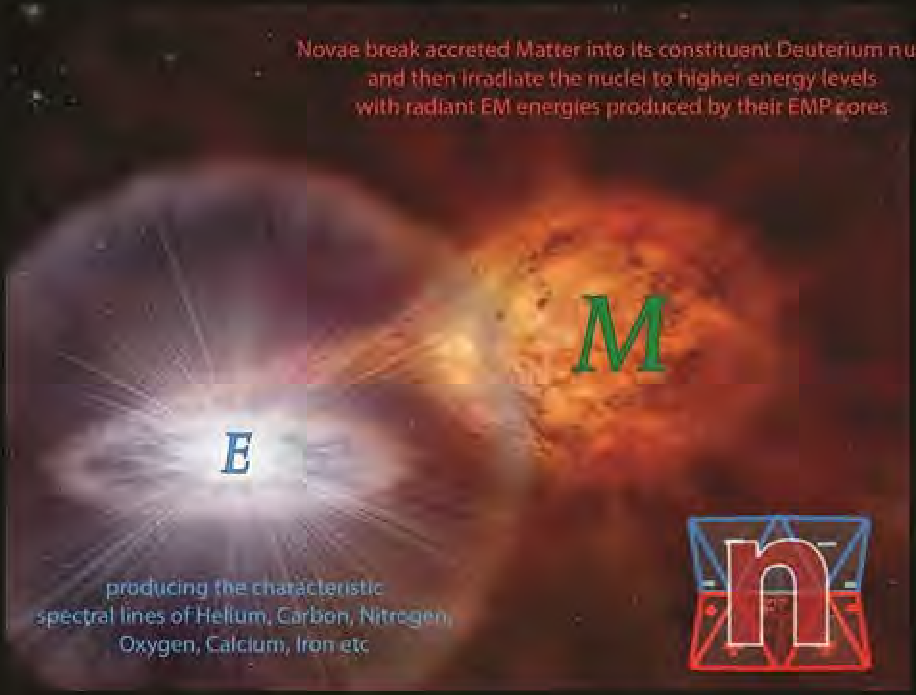

Novae break accreted Matter into its constituent Deuterium nuclei
and then irradiate the nuclei to higher energy levels
with radiant EM energies produced by their EMP cores

producing the characteristic
spectral lines of Helium, Carbon, Nitrogen,
Oxygen, Calcium, Iron etc

runaway
Matter-energy conversion
within the supernova

12	*Lyman*
48	*Balmer*
108	*Paschen*
192	*Brackett*
300	*Pfund*
432	*Humphries*
588	*un-named*
768	*ionised leptons*

accelerates ions & photo-electrons
producing spectral line emissions
and creates gamma ray bursts

An enigma that has puzzled experts for over 50 years is the origin of the irregular, inhomogeneous distribution of nova ejecta

The material transferred by the companion star is often of solar composition (i.e., close to 98% hydrogen and helium by mass).

Elements, in the range between carbon (C) and calcium (Ca) on the periodic table, can account for 30% to 50% of the material ejected during a nova explosion.

Copyright ABRAHAM [2008] - All rights reserved

Supernovae explosions

Supernovae are extremely luminous and cause a burst of radiation that often briefly outshines an entire galaxy,

The GEM pinching [collapse of Matter waveforms]
always results in divergent neutral charge
KEM mass-energies

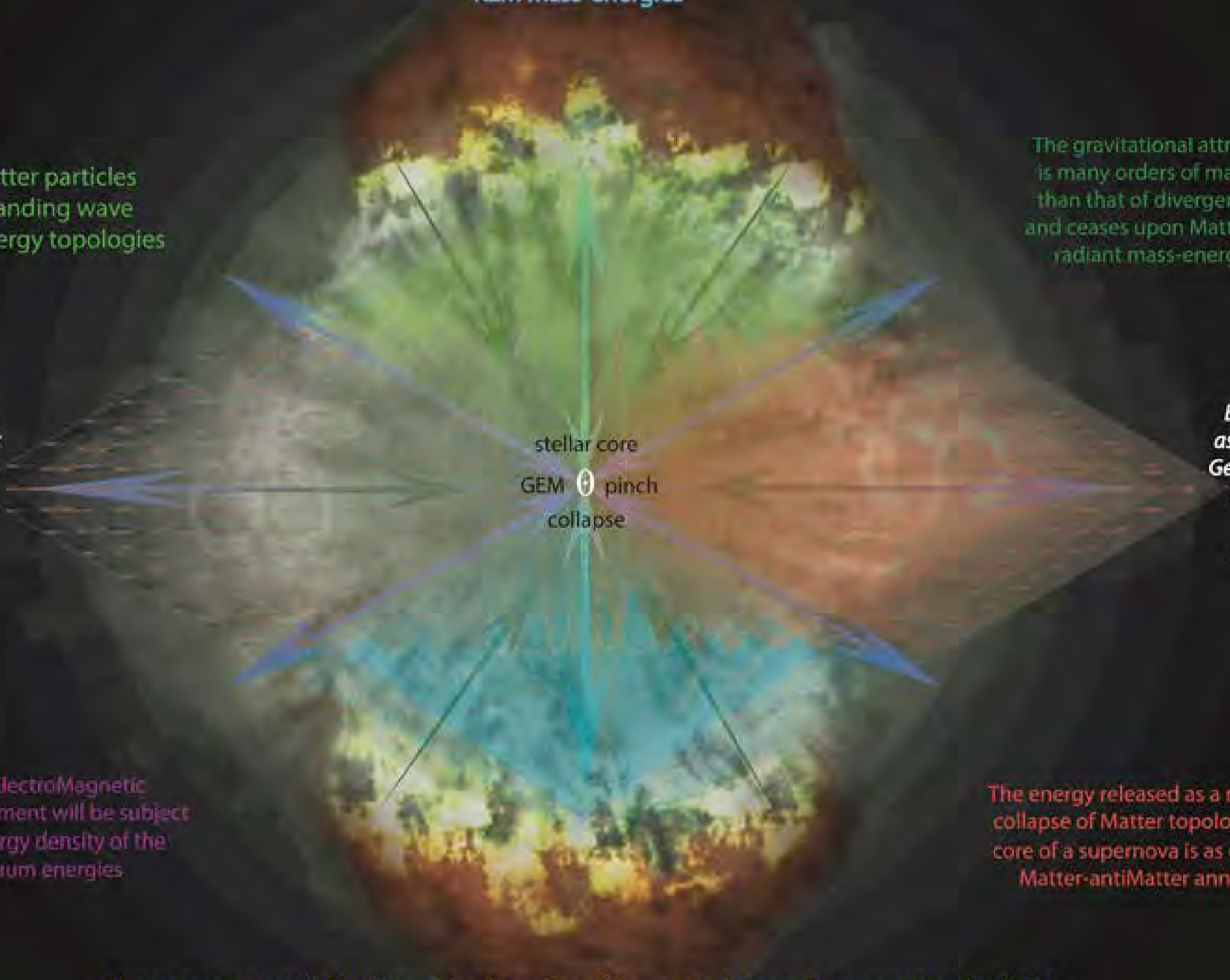

All Matter particles
are standing wave
mass-energy topologies

The gravitational attraction of Matter
is many orders of magnitude weaker
than that of divergent KEM energies
and ceases upon Matter's collapse into
radiant mass-energy geometries

If the EM pinch at the core of a Star
rapidly increases its efficiency or
loses its abili'ty to confine the
resultant KEM energies the
star will be blasted apart

Black holes, hypothesied in physics
as a solution to the math of Einstein's
General relativity since it was was first
formulated in 1915, do NOT exist

All their physics are provided for
by GEM pinch field mechanics

A star moving in a ElectroMagnetic
vacuum energy environment will be subject
to changes in the energy density of the
surrounding vacuum energies

The energy released as a result of the
collapse of Matter topologies in the
core of a supernova is as efficient as
Matter-antiMatter annihilation

Supernovae are created by the sudden disruption of the counter balanced convergent and radiative
mass-ENERGY-Matter output of GEM pinch singularities at the core of compact stellar objects
which results in the sudden release of massive amounts of divergent mass-energy
all at once from the stellar core blasting away the remaining stellar Matter

Tetryonics 79.02 - Supernovae explosions

Copyright ABRAHAM [2008] - All rights reserved

Singularities

A gravitational singularity or spacetime singularity is a location where the quantities that are used to measure the gravitational electromagnetic fields become infinite in a way that does not depend on the spatial co-ordinate system

Just as the radiative KEM fields of Matter in motion and charged EM fields gives rise to the modern physics misconceptions of spherical point particles and String theories

Tetryonic theory precludes point particles and black holes through analytic mass-energy geometries & Gaussian Matter topologies that result from equilateral Planck energy momenta interactions

The collapse of 3D Matter releases its standing-wave energies as radiant 2D mass-energies foms the basis for all Stellar energy releases

Current theories based on mathematical modelling of General relativity without accurate quantum mechanical models predicts that at densities greater than those supported by fermion degeneracy, gravity overwhelms all other forces collapsing the 3D material body of a large star to create a singularity and the eventual formation of a black hole.

Equilateral Planck charge geometries are the key to understanding stellar processes

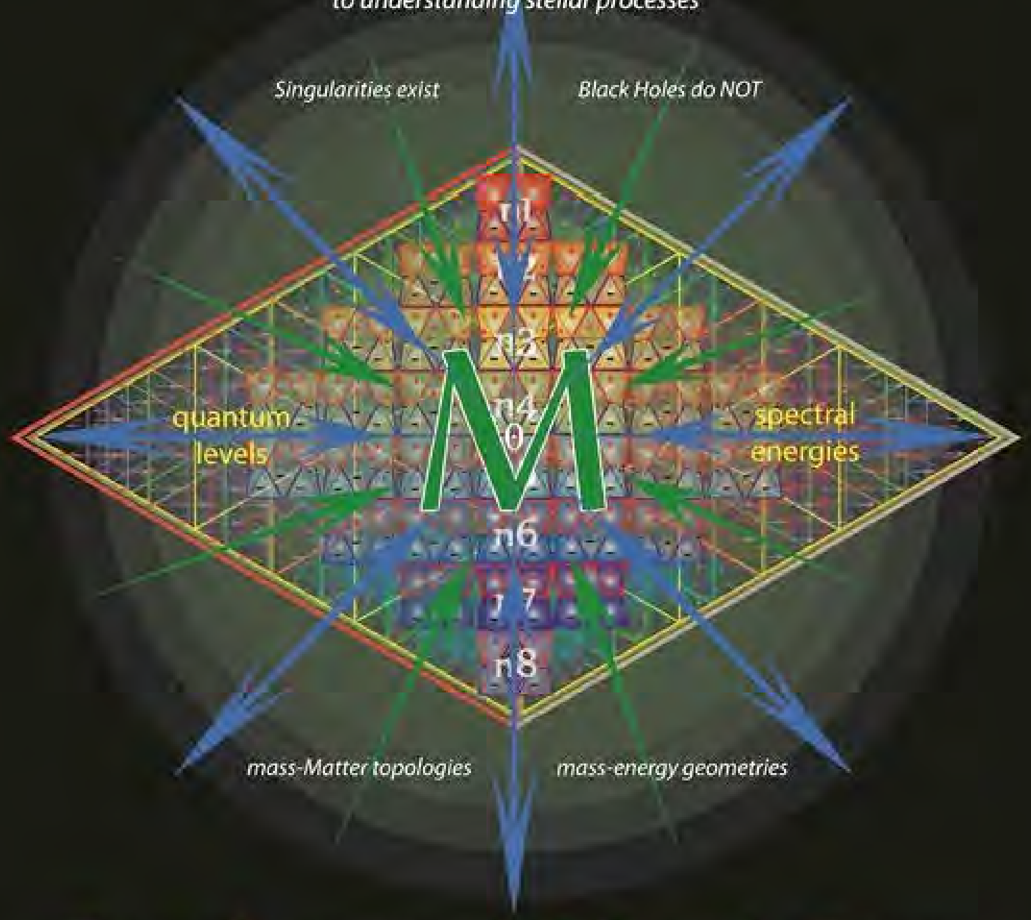

The smallest quanta of 3D Matter is a Tetryon

(any attempt to reduce its 3 spatial dimensions results in its collapse into 2D divergent mass-energy waveforms)

Tracing the GravitoElectroMagnetic fields of Matter back to their origin point can give rise to the misconception of a EM field singularity leading to a black hole

Black holes do not exist

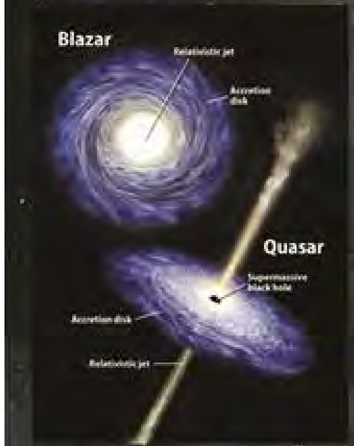
Blazar
Relativistic jet
Accretion disk
Quasar
Supermassive black hole
Accretion disk
Relativistic jet

Quasi-stellar objects

As standing-wave Matter is converted to radiant EM masses and pure energy the outputs of GEM pinches can increase dramatically over short periods

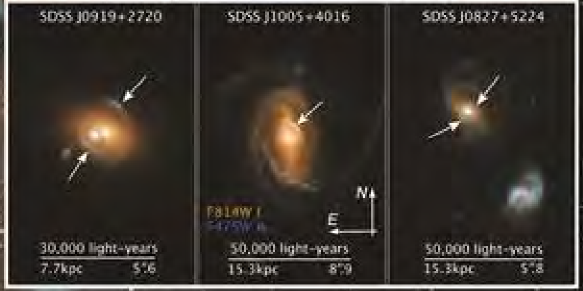
SDSS J0919+2720
30,000 light-years
7.7kpc
5"6
SDSS J1005+4016
F814W I
F475W B
N
E
50,000 light-years
15.3kpc
8"9
SDSS J0827+5224
50,000 light-years
15.3kpc
5"8

A quasi-stellar radio source ("quasar") is a very energetic and distant active galactic nucleus.

Quasars and Blazars are extremely luminous and were first identified as being high redshift sources of electromagnetic energy, including radio waves and visible light, that were point-like, similar to stars, rather than extended sources similar to galaxies

Long hypothesised as having supermassive Blackholes at their cores Tetryonics reveals Matter~Energy GEM pinches to be the source of the observed, distant large-scale stellar dynamics

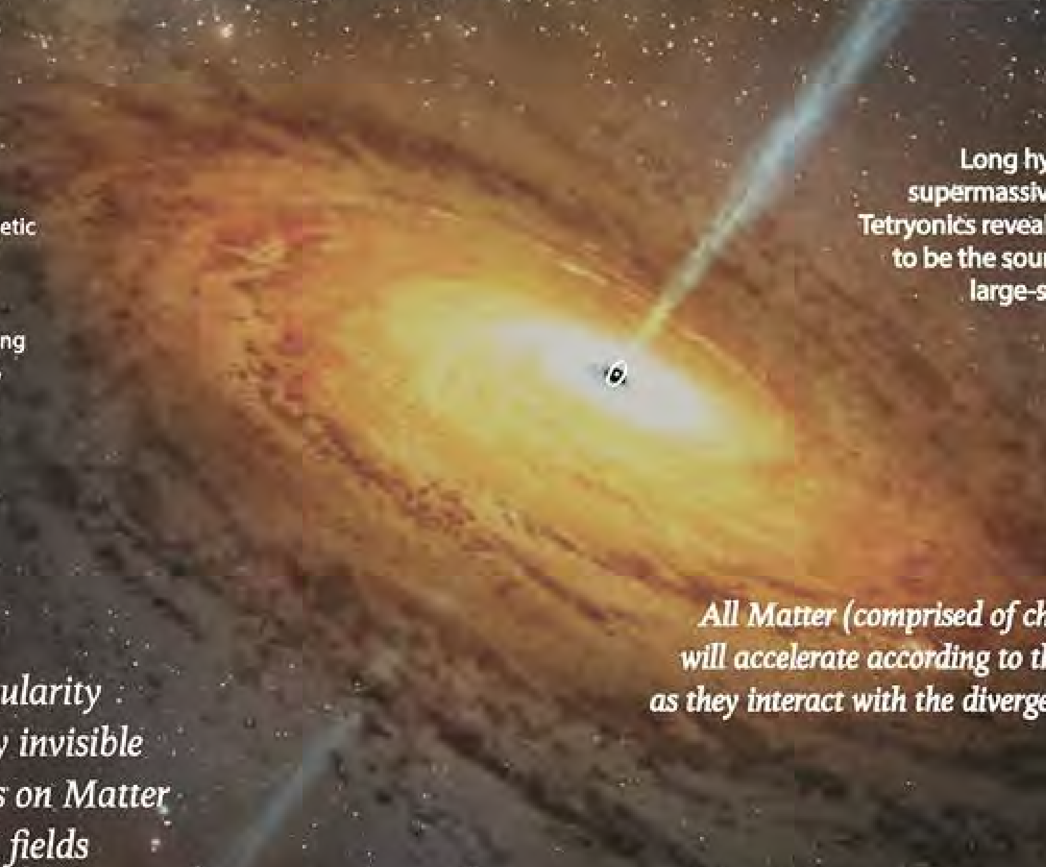

All Matter (comprised of charged Planck quanta) will accelerate according to their charge distributions as they interact with the divergent energies of EM pinches

The central 'core' point singularity of any GEM pinch is completely invisible and is revealed only by its effects on Matter through its interactive EM fields

$E = mc^2$

mass-energy equivalence

mass geometries are energy per c^2

Stellar GEM pinches convert Matter topologies into radiant mass-energy geometries with 100% efficiency

Matter topologies are energy per c^4

Matter-energy equivalence

$E = \frac{m}{c^2}$

Copyright ABRAHAM [2008] - All rights reserved

Active Galaxies

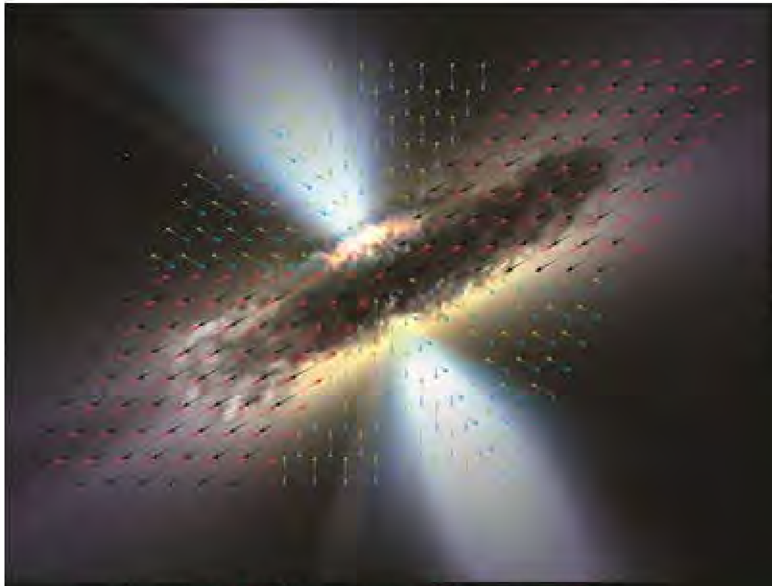

In addition to the E-field acceleration of charged particles GEM pinches can also release energies along their Magnetic axis

An active galactic nucleus (AGN) is a compact region at the centre of a galaxy that has a much higher than normal luminosity over at least some portion, and possibly all, of the electromagnetic spectrum.

The motion of stars about a central point in a galaxy's core is the result of a GEM pinch's radiated interactive forces

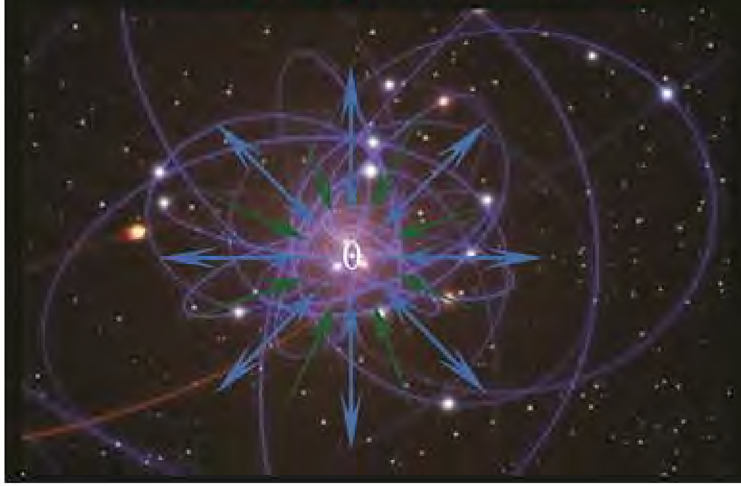

GR uses stress tensors to model the gravitational effects of ALL energy in a system in its various forms without distinction or differentiation between mass-energy geometries and Matter topologies

EM PINCHES ARE 100% EFFICIENT CONVERTERS OF MATTER INTO ENERGY

Using Tetryonic geometry the Universe is revealed as being completely dominated by ElectroMagnetic forces, with Gravity aiding at the larger scales to help shape the differing forms of Galaxies that we see

EM PINCHES REDUCE TO SINGULARITIES AT THE FOCUS OF THEIR DIVERGENT FIELDS

Tetryonic geometry reveals divergent interactive EM masses & convegent gravitational Matter all contribute to the nett force of Gravitational fields

The current astrophysical models of active galaxies propose supermassive Blackholes as the only mechanism possible to produce the observed large-scale dynamics

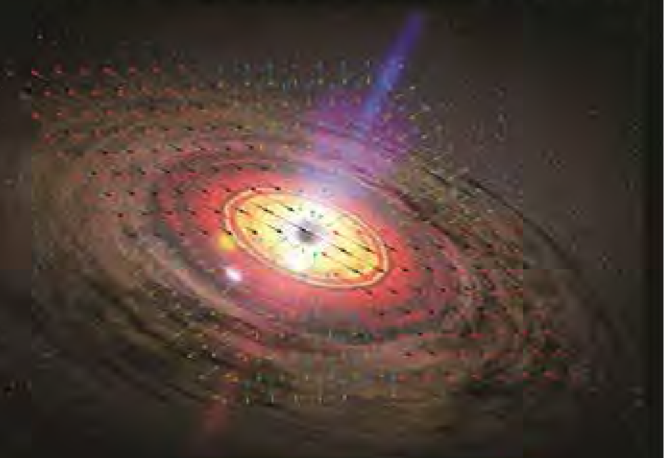

E-fields hold Matter in a disc and M-fields release highly polarised energies

Tetryonics reveals the Galactic EM Pinches are also capable of fitting the observed data and explaining many of the current mysteries surrounding how active galaxies generate their power.

The energies released by the conversion of Matter into energy in GEM pinches is the most energetic process in the Universe

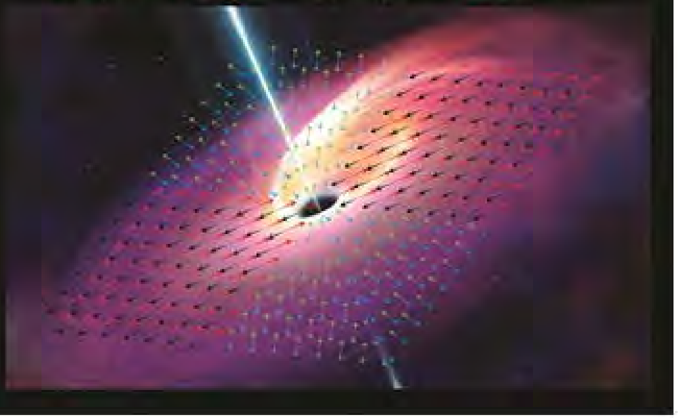

Black holes

[Garbage IN - Garbage OUT]

Computer-based modelling of GR
mathematics without charge geometries
leads to paradoxial & erroneous outcomes

Tetryonics allows for Supermassive Matter to form but does NOT allow for the classically defined Black holes

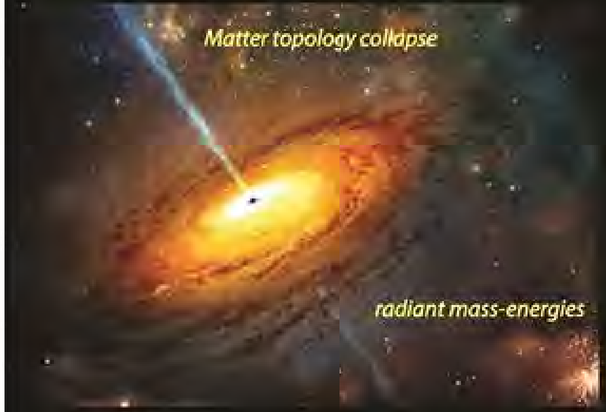

The centres of Galaxies & Quasars hypothesised as being powered by supermassive blackholes are in fact the foci of ElectroMagnetic pinches

In General relativity, a black hole is a region of space in which the gravitational field is so powerful that nothing, including light, can escape its gravitational pull.

The black hole has a one-way surface, called an event horizon, into which objects can fall, but out of which nothing can come. It is called "black" because it absorbs all the light that hits it, reflecting nothing, just like a perfect blackbody in thermodynamics.

'Standard Model' quantum analysis of black holes shows them to possess a temperature and emit Hawking radiation.

In Standard Model Quantum Mechanics
a problem arises in that Charge and Gravity equations
break down and produce infinities and
other impossible results

Very large non-radiating 'Dark Matter'
would be the closest thing possible to
the currently theorised black-holes.

However the 3D volume of such an object
would be limited to the space taken up by its
constituent Tetryonic charge Matter topologies.
(ie Tetryons face-to face without Space
present between any facsias)

This is similar to the current models of
Neutron Stars , WIMPs and MACHOs

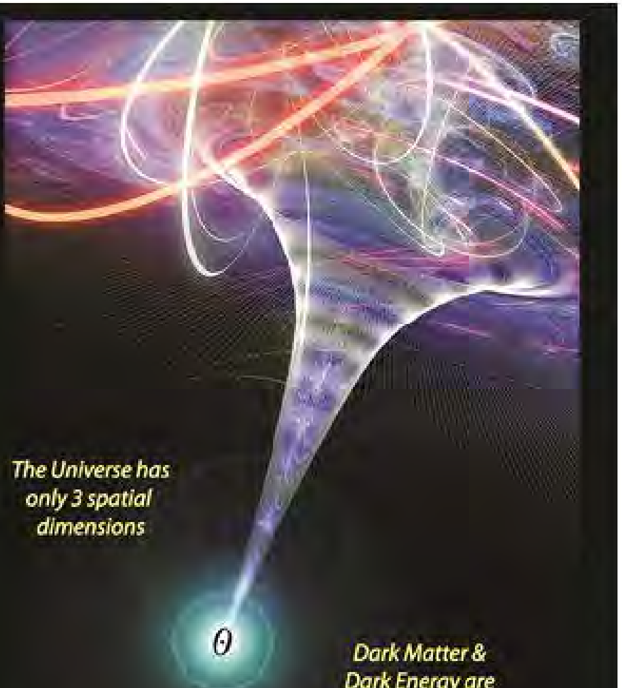

The Universe has
only 3 spatial
dimensions

Dark Matter &
Dark Energy are
EM vector forces

Black holes, Worm holes &
other passable singularities
do NOT exist

The collapse of standing-wave Matter topologies results in the release of divergent radiative mass-energy geometries

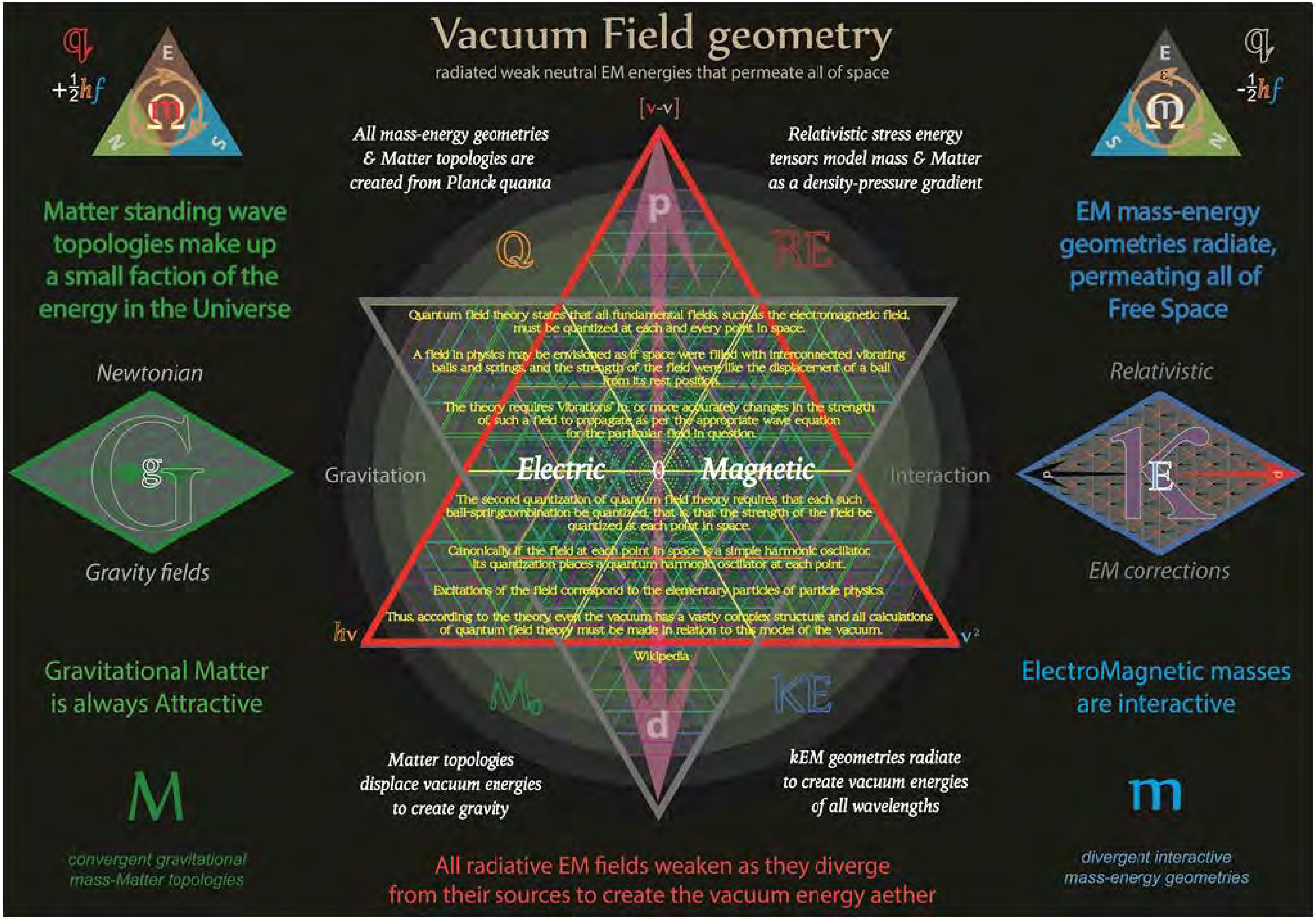

Vacuum Field geometry

radiated weak neutral EM energies that permeate all of space

Matter standing wave topologies make up a small faction of the energy in the Universe

Newtonian

Gravity fields

Gravitational Matter is always Attractive

M

convergent gravitational mass-Matter topologies

All mass-energy geometries & Matter topologies are created from Planck quanta

Relativistic stress energy tensors model mass & Matter as a density-pressure gradient

Quantum field theory states that all fundamental fields, such as the electromagnetic field, must be quantized at each and every point in space.

A field in physics may be envisioned as if space were filled with interconnected vibrating balls and springs, and the strength of the field were like the displacement of a ball from its rest position.

The theory requires 'vibrations' in, or more accurately changes in the strength of such a field to propagate as per the appropriate wave equation for the particular field in question.

Gravitation **Electric** **Magnetic** Interaction

The second quantization of quantum field theory requires that each such ball-springcombination be quantized, that is, that the strength of the field be quantized at each point in space.

Canonically, if the field at each point in space is a simple harmonic oscillator, its quantization places a quantum harmonic oscillator at each point.

Excitations of the field correspond to the elementary particles of particle physics.

Thus, according to the theory, even the vacuum has a vastly complex structure and all calculations of quantum field theory must be made in relation to this model of the vacuum.

Wikipedia

Matter topologies displace vacuum energies to create gravity

kEM geometries radiate to create vacuum energies of all wavelengths

All radiative EM fields weaken as they diverge from their sources to create the vacuum energy aether

EM mass-energy geometries radiate, permeating all of Free Space

Relativistic

EM corrections

ElectroMagnetic masses are interactive

m

divergent interactive mass-energy geometries

Tetryonics 79.07 - Vacuum Field geometry

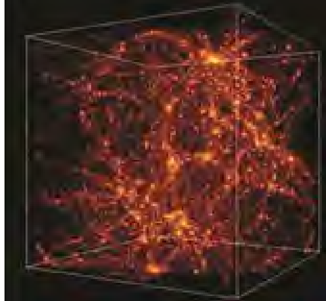

The Big Bang theory

The Big Bang is a considered to be a well-tested scientific theory which is widely accepted within the scientific community because it is the most accurate and comprehensive explanation for the full range of phenomena astronomers observe.

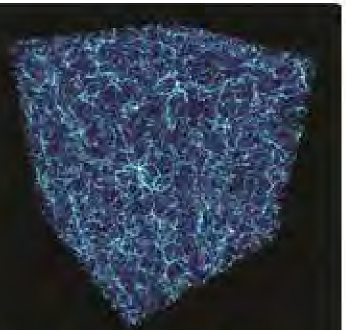

The framework for the Big Bang model relies on Albert Einstein's general relativity and on simplifying assumptions such as homogeneity and isotropy of space.

Primodial Atom

According to the Big Bang theory, the Universe was once in an extremely hot and dense state which expanded rapidly

The Big Crunch

If the mass density of the Universe were greater than the critical density, then the Universe would reach a maximum size and then begin to collapse.

It would become denser and hotter again, ending with a state similar to that in which it started

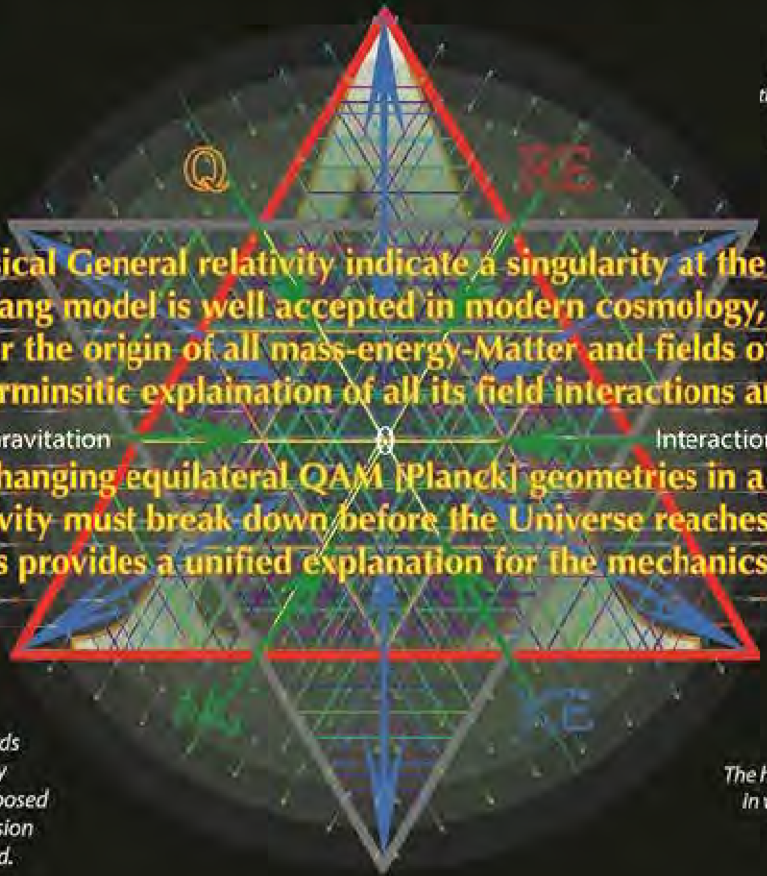

The equations of classical General relativity indicate a singularity at the origin of cosmic time, and while the Big Bang model is well accepted in modern cosmology, Tetryonics provides a clear causal model for the origin of all mass-energy-Matter and fields of force in our Universe, offering a unified, determinsitic explaination of all its field interactions and ultimate energy fate.

Tetryonics relates Time to changing equilateral QAM [Planck] geometries in a dynamic evolving Universe. Moreover, general relativity must break down before the Universe reaches the Planck temperature while Tetryonics provides a unified explanation for the mechanics of the Universe

Big Bang

Extrapolation of the expansion of the Universe backwards in time using general relativity yields an infinite density and temperature at a finite time in the past, which it is proposed underwent a explosive birth and a period of rapid expansion in order to produce the homognetity currently observed.

The Big Chill

The heat death of the universe is a suggested ultimate fate of the universe, in which the universe has diminished to a state of no thermodynamic free energy and therefore can no longer sustain motion or life.

Inflation

The Big Bang theory is the prevailing cosmological model that attempts to explain the early development of the Universe

Energy death

Tetryonics 79.08 - The Big Bang

Copyright ABRAHAM [2008] - All rights reserved

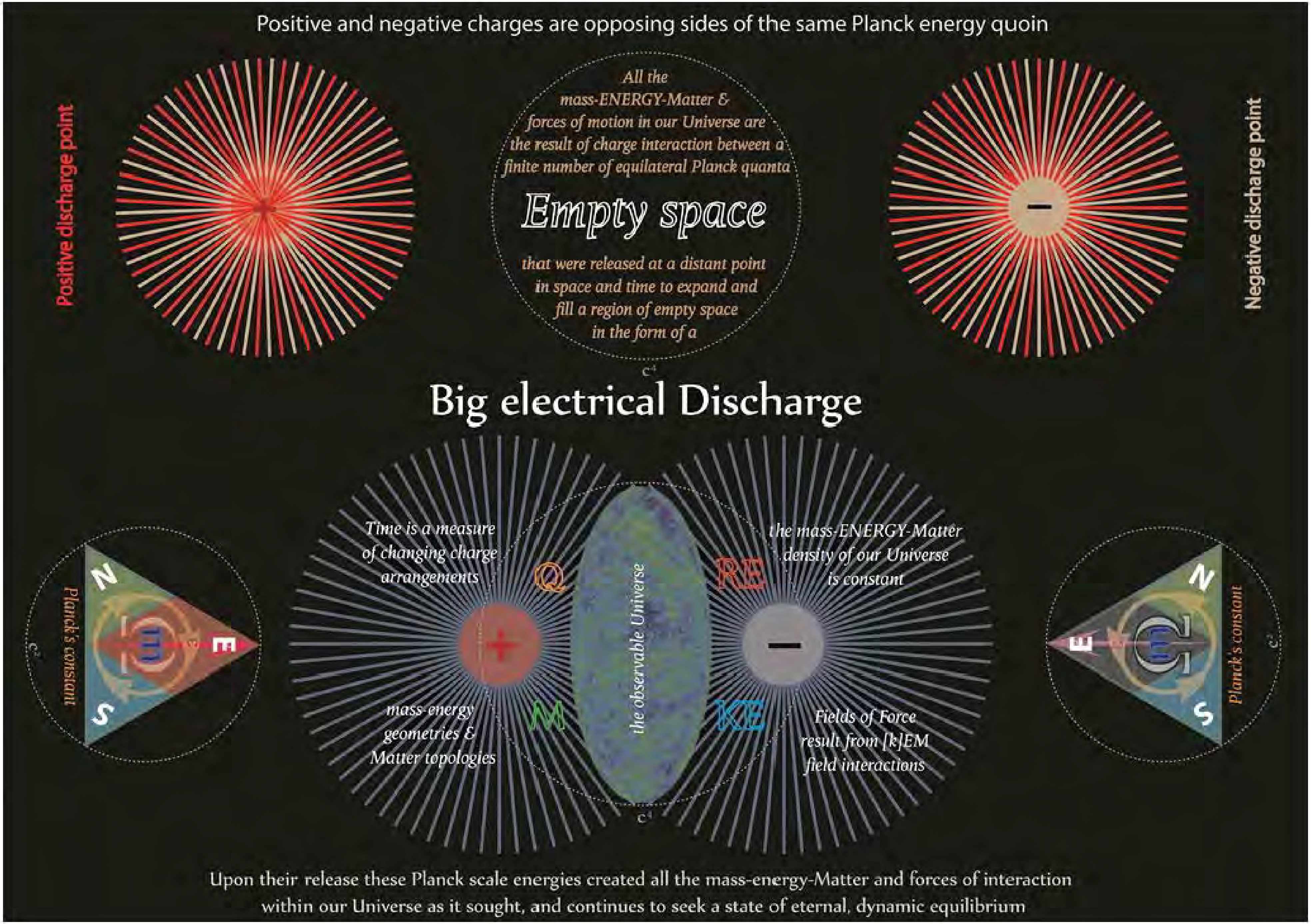

Positive and negative charges are opposing sides of the same Planck energy quoin

All the mass-ENERGY-Matter & forces of motion in our Universe are the result of charge interaction between a finite number of equilateral Planck quanta

Empty space

that were released at a distant point in space and time to expand and fill a region of empty space in the form of a

Big electrical Discharge

Time is a measure of changing charge arrangements

the mass-ENERGY-Matter density of our Universe is constant

mass-energy geometries & Matter topologies

Fields of Force result from [k]EM field interactions

Upon their release these Planck scale energies created all the mass-energy-Matter and forces of interaction within our Universe as it sought, and continues to seek a state of eternal, dynamic equilibrium

Positive discharge point
Negative discharge point
−
c^4
Q
M
RE
KE
+
−
the observable Universe
c^4
N
E
S
Planck's constant
c^2
m

The Dynamic Universe

Like a living breathing Organism, the Universe can now be seen as an eternally dynamic system that is constantly evolving through Matter creation, EM interactions & Matter destruction

Gravitational contraction

Singularities

arise from the vector modelling of convergent EM fields of force without an understanding of the EM energy momenta geometries

The formation of Matter creates regions of attractive low vacuum energy pressure

[Convergent Gravity]

Dark Matter

results from modern modelling of convergent EM interactions without an understanding of the EM energy momenta geometries

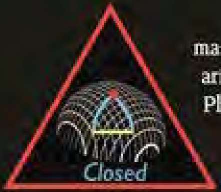

All mass-ENERGY -Matter arises from equilateral Planck energy quanta

Flat Euclidean geometries and Platonic topologies create everything

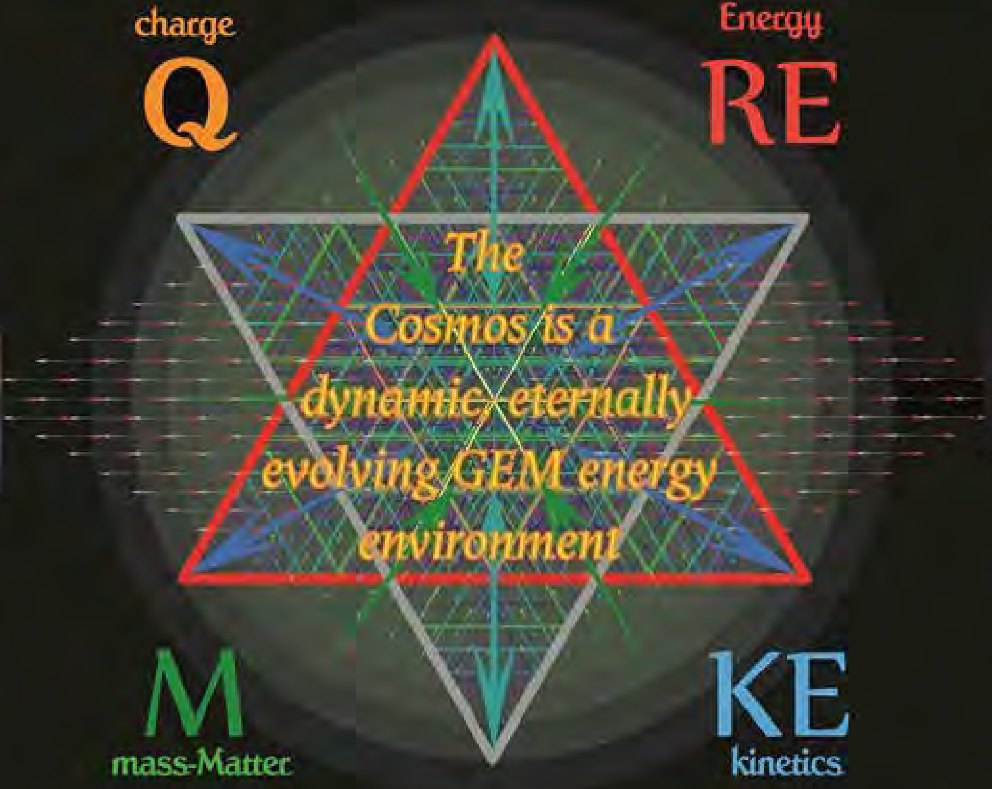
charge
Q
Energy
RE
The Cosmos is a dynamic eternally evolving GEM energy environment
M
mass-Matter
KE
kinetics

Radiant expansion

Inflation

arises from an attempt to explain the observed homogenity of the Universe without an understanding of kEM energy momenta geometries

The collapse of standing-wave Matter releases interactive EM mass-energies

[Divergent EM Radiation]

Dark Energy

results from modern modelling of divergent EM interactions without an understanding of the EM energy momenta geometries

The energetic 'space-time fabric' of our Universe is not curved, it is created from flat Euclidean Planck quanta in a finely tuned state of dynamic equilibrium between gravitational Matter & radiant EM mass-energies

Tetryonics 79.10 - The Dynamic Universe

Copyright ABRAHAM [2008] - All rights reserved

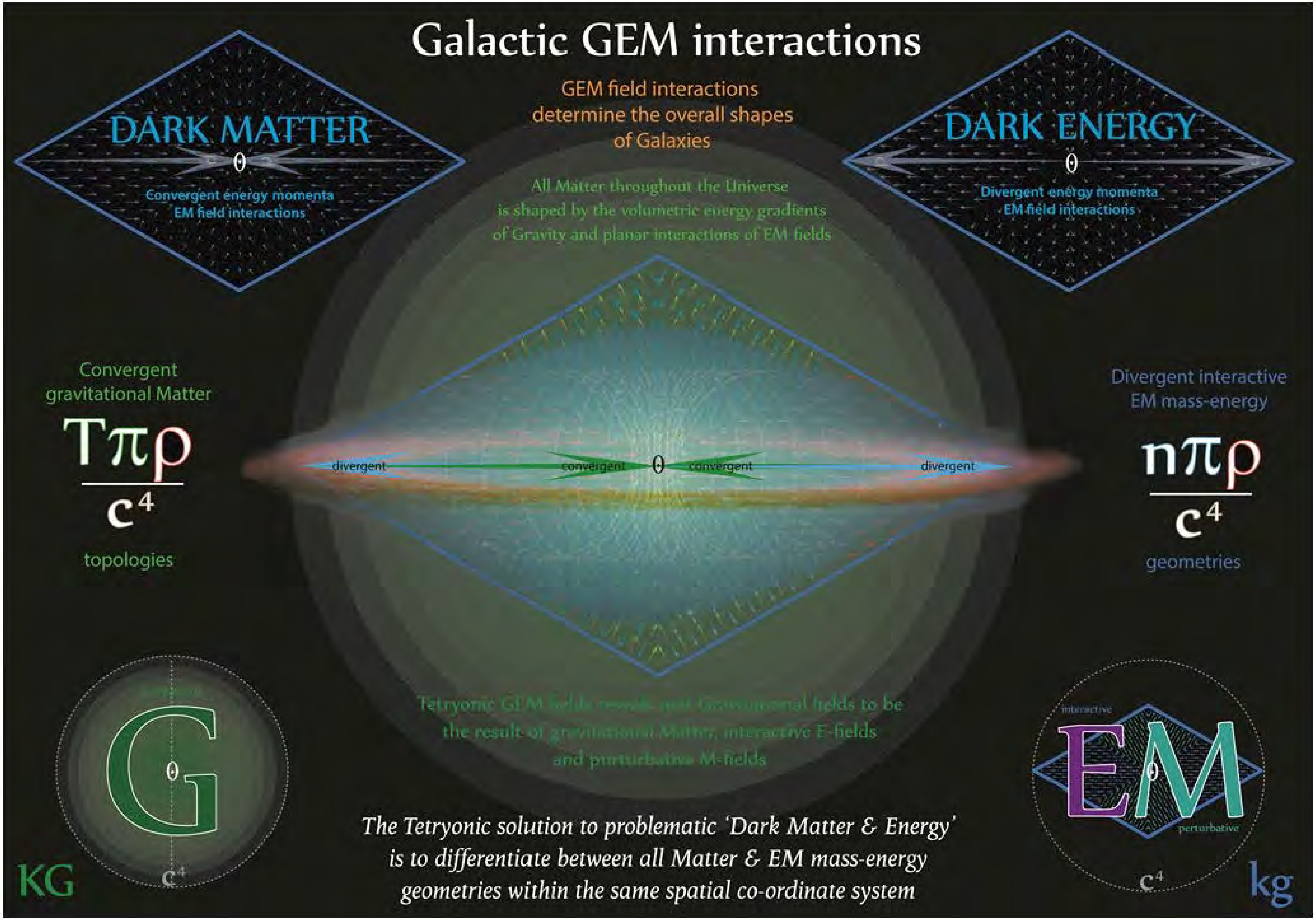

Galactic GEM interactions

GEM field interactions
determine the overall shapes
of Galaxies

All Matter throughout the Universe
is shaped by the volumetric energy gradients
of Gravity and planar interactions of EM fields

DARK MATTER

Convergent energy momenta
EM field interactions

DARK ENERGY

Divergent energy momenta
EM field interactions

Convergent
gravitational Matter

$$\frac{T\pi\rho}{c^4}$$

topologies

Divergent interactive
EM mass-energy

$$\frac{n\pi\rho}{c^4}$$

geometries

Tetryonic GEM fields reveal all Gravitational fields to be
the result of gravitational Matter, interactive E-fields
and perturbative M-fields

The Tetryonic solution to problematic 'Dark Matter & Energy'
is to differentiate between all Matter & EM mass-energy
geometries within the same spatial co-ordinate system

Tetryonics 79.11 - Universal GEM interactions

Copyright ABRAHAM [2008] - All rights reserved

The Dark Forces

Gravitational attraction is the result of the nullspace of Matter.
It is many orders of magnitude weaker than ElectroMagnetic forces
(Gravity does not hold the whole galaxy together)

charged energy geometries

energy momenta quanta

$$n\pi\underbrace{[m\Omega v]}_{\text{mass} \quad \text{velocity}}^2$$

$$G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Newtonian Gravity
models the motion resulting
from the attractive force
between Matter

General Relativity
models perturbative force
resulting from the total energy
of any system

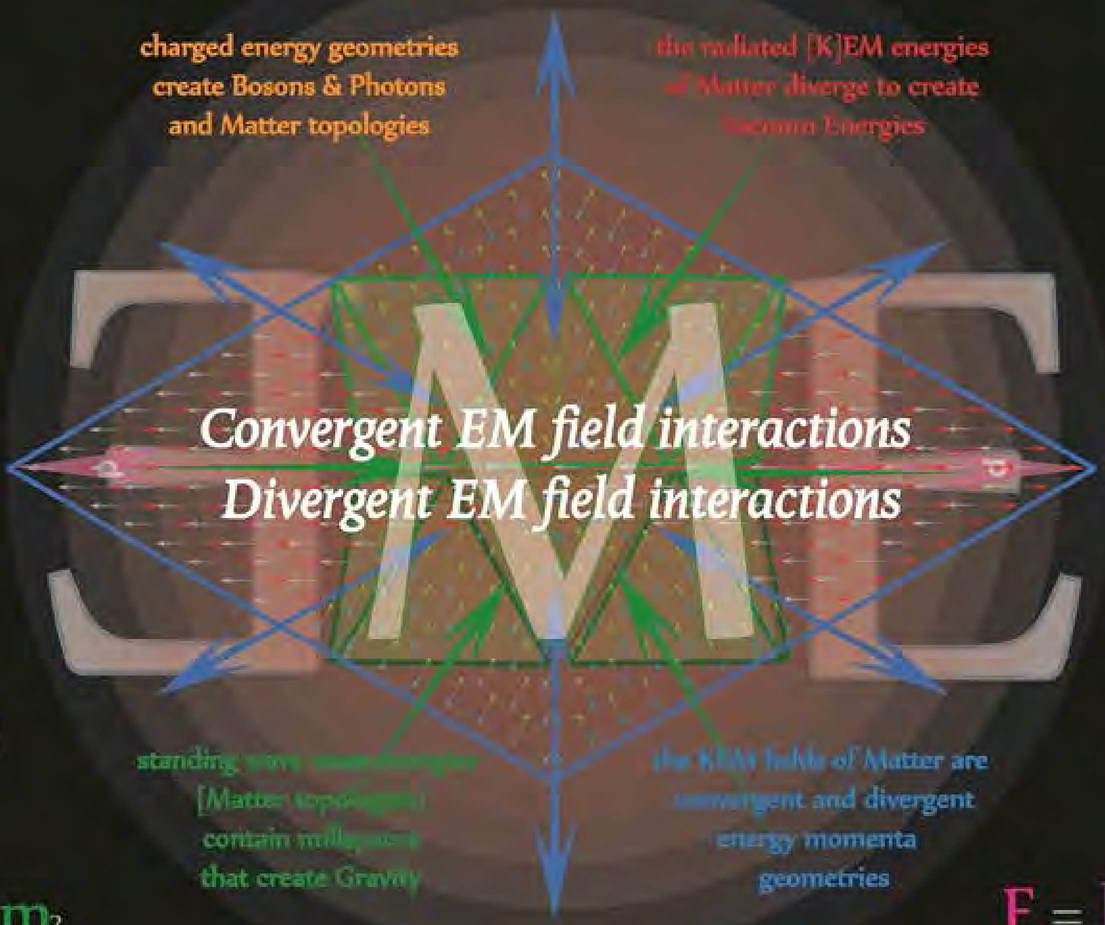

charged energy geometries
create Bosons & Photons
and Matter topologies

the radiated [K]EM energies
of Matter diverge to create
Vacuum Energies

Convergent EM field interactions
Divergent EM field interactions

standing wave mass-energies
[Matter topologies]
contain nullspaces
that create Gravity

the [K]EM fields of Matter are
convergent and divergent
energy momenta
geometries

All [K]EM radiation has
a CONVERGENT component in its
interactive field that increases
when Matter is in motion

DARK MATTER

or when its standing-wave Matter
topology is destroyed

All [K]EM radiation has
a DIVERGENT omponent in its
interactive field that increases
when Matter is in motion

DARK ENERGY

or when its standing-wave Matter
topology is destroyed

$$E \xrightarrow[\text{energy momenta}]{\text{convergent}} M$$

$$M \xrightarrow[\text{energy momenta}]{\text{divergent}} E$$

$$F = -G\frac{M_1 m_2}{r^2}$$

$$F = k\frac{Q_1 Q_2}{r^2}$$

Gravitational
Matter
is attractive

G

Tetryonics dictates that in order to correct for measured Gravitational accelerations
we must distingush between EM mass and Matter and take into account
all the EM forces as well as that of Gravity produced by Matter

k

ElectroMagnetic
mass-energies
are interactive

Tetryonics 80.01 - The Dark Forces

Copyright ABRAHAM [2008] - All rights reserved

Vacuum Energies

Vacuum energy is an underlying background energy that exists in space even when it is devoid of Matter (free space) and results from the introduction of the Cosmological constant [Λ] by Albert Einstein as a modification of his original theory of General relativity in order to achieve a static universe.

Dark Matter is a measure of CONVERGENT KEM interactions additional to that of Gravity

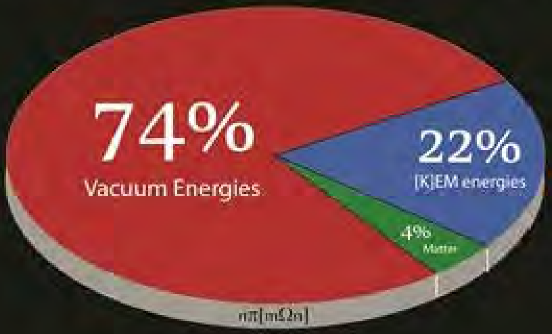

Dark Energy is a measure of DIVERGENT KEM interactions opposing that of Gravity

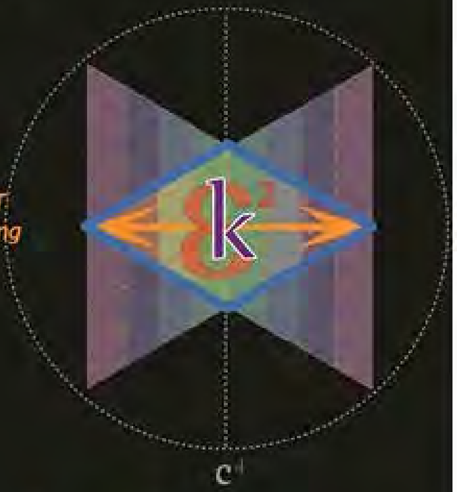

He later abandoned this concept after observations of the Hubble redshift indicated that the universe might not be stationary

Gravitational Matter is always attractive

$$G_{ab} + \Lambda g_{ab} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

total system energy

ElectroMagnetic masses are always interactive

Matter

M

topologies

Unlike Einstein's Special and General theories of Relativity that utilizes a scalar energy density tensor, Tetryonics utilizes analytical geometry & topology to differentiate between EM mass & Matter and model all energy interactions in any spatial region

mass

m

geometries

Gravito-Electro-magnetic interactions

Total energy geometries *quantised energy momenta*

$$GEM = \frac{n\pi[m\Omega v]^2}{c^4}$$

of all EM mass-ENERGY-Matter in the Universe can be defined as

spatial volume of mass-energy-Matter interactions

Gravity results from standing-wave Matter

EM interactions result from superpositioned EM masses

Tetryonics 80.02 - Vacuum Energies

Copyright ABRAHAM [2008] - All rights reserved

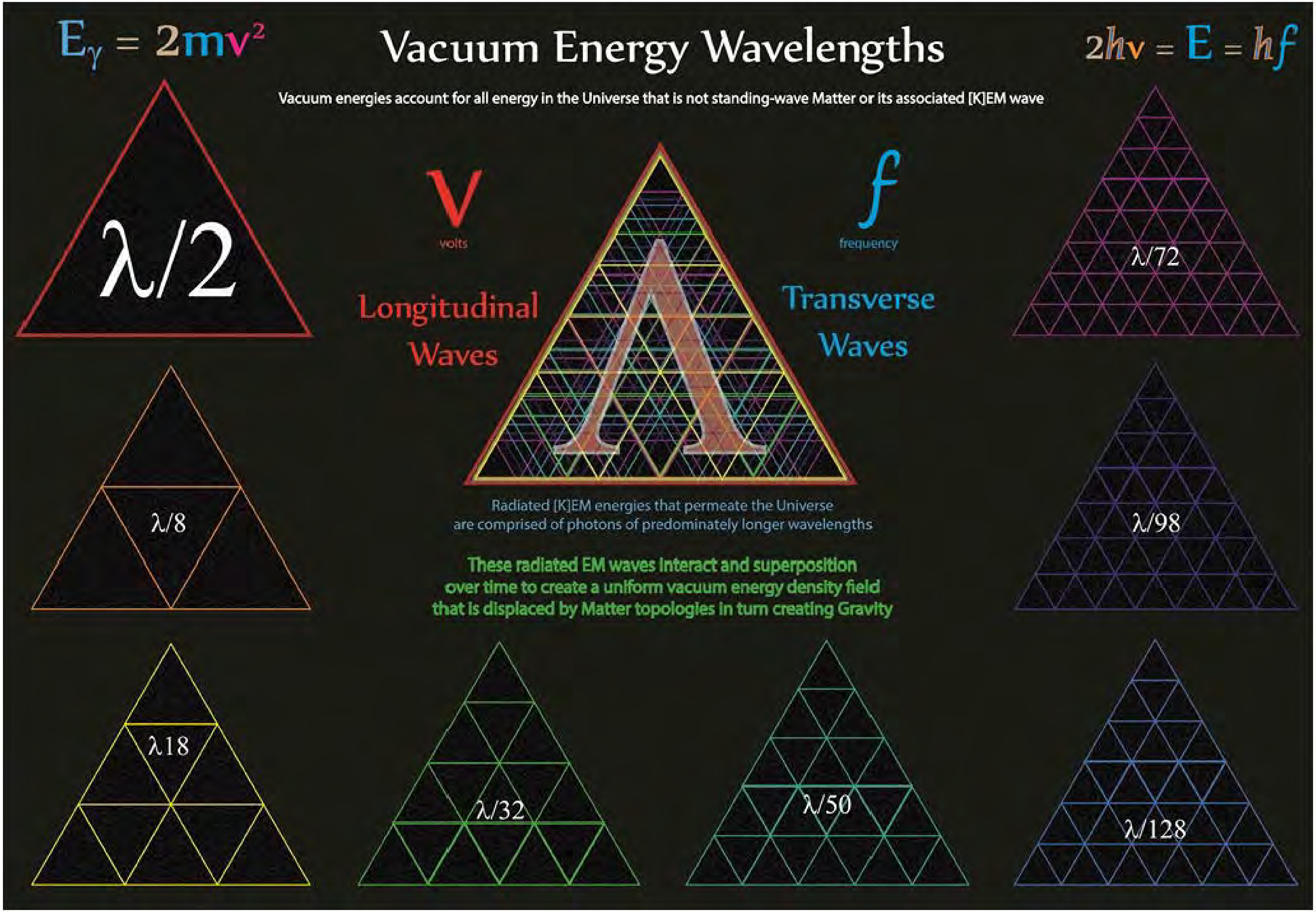

Vacuum Energy Wavelengths

$E_\gamma = 2mv^2$

$2hv = E = hf$

Vacuum energies account for all energy in the Universe that is not standing-wave Matter or its associated [K]EM wave

Radiated [K]EM energies that permeate the Universe
are comprised of photons of predominately longer wavelengths

**These radiated EM waves interact and superposition
over time to create a uniform vacuum energy density field
that is displaced by Matter topologies in turn creating Gravity**

λ/2
V
volts
Longitudinal Waves
Λ
f
frequency
Transverse Waves
λ/72
λ/8
λ/98
λ18
λ/32
λ/50
λ/128

Tetryonics 80.03 - Vacuum Energy Wavelengths

Copyright ABRAHAM [2008] - All rights reserved

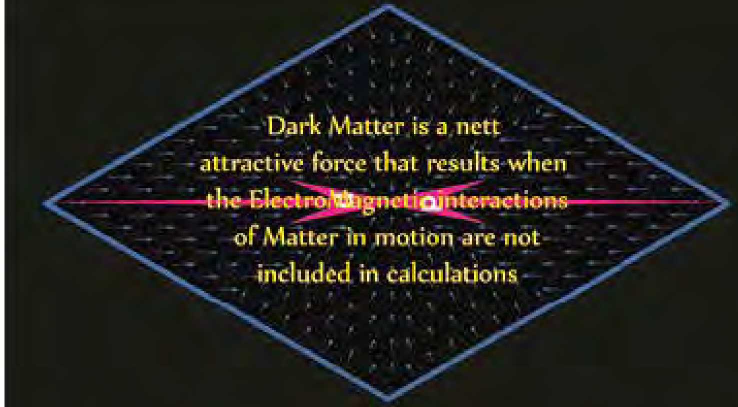

Dark Matter is a nett attractive force that results when the ElectroMagnetic interactions of Matter in motion are not included in calculations

In addition to Gravity the [K]EM waves of all Matter in motion possess a convergent component which in turn creates a unidirectional 'force'

The convergent EM force component can be modelled as being the result of an 'invisible' form of Matter or additional gravitational force

DARK MATTER

Convergent EM field vector momenta

All Matter emits divergent EM radiation photons comprised of bidirectional vector momenta

Matter produces Gravity
EM waves are interactive

Dark Matter is matter that neither emits nor scatters light or other electromagnetic radiation, and so cannot be directly detected via optical or radio astronomy.

Gravity is the result of the displacement of vacuum energies by 3D Matter topologies

All [K]EM radiation has a convergent component in its interactive field that increases when Matter is in motion or created in the form of EM standing-wave topologies

DARK MATTER

As Matter is destroyed and energy is released the attractive force of Gravity decreases but the convergent acceleration due to Dark Matter increases

Dark Matter

Dark matter is estimated to constitute 84% of the Matter in the universe and 23% of the total mass-energies

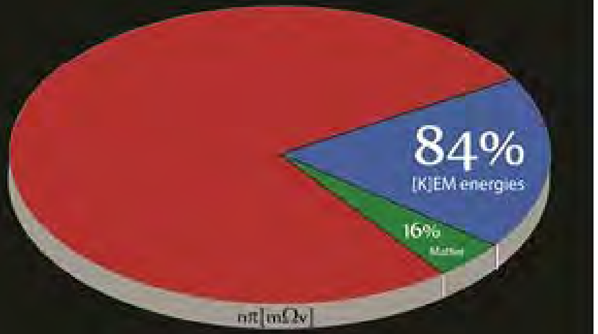

Dark Matter created by KEM field convergence contributes to the total energies of the cyclic Universe

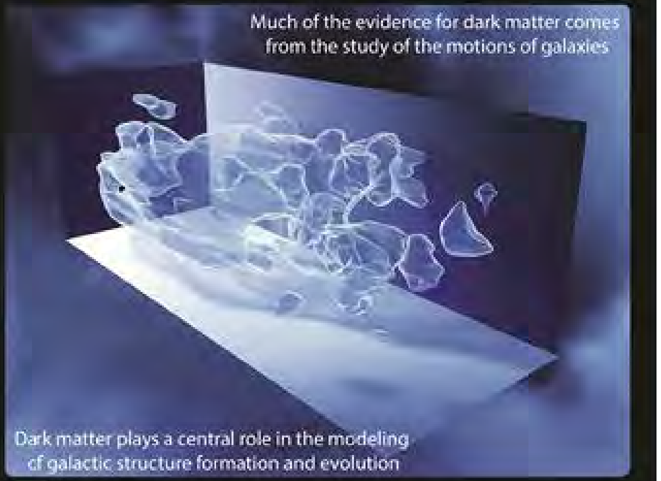

Much of the evidence for dark matter comes from the study of the motions of galaxies

Dark matter plays a central role in the modeling of galactic structure formation and evolution

Dark Matter's existence is inferred from its gravitational effects on visible Matter and the convergent lensing of background radiation

Copyright ABRAHAM [2008] - All rights reserved

In 1986, Nobel laureate Hannes Alfven
postulated at an electrical model
for the solar system

Galaxy Rotations

The rotation curve of a galaxy (also called a velocity curve)
can be represented by a graph that plots the orbital speed (in km/s) of the stars or gas
in the galaxy on the y-axis against the distance from the center of the galaxy on the x-axis

The dynamic modelling
of large scale Matter must take
into account all forces present

Based on Newtonian mechanics and assuming, as was originally thought, that most of the Matter of the galaxy had to be in the galactic bulge near the center, Matter (such as stars and gas) in the disk portion of a spiral should orbit the center of the galaxy similar to the way in which planets in the solar system orbit the sun, i.e. where the average orbital speed of an object any distance away from the majority of the Matter would decrease inversely with the square root of the radius of the orbit

G

**Gravitational
Matter
is attractive**

Newtonian gravity
models convergent
Gravity only

$$F = -G\frac{M_1}{r^2}$$

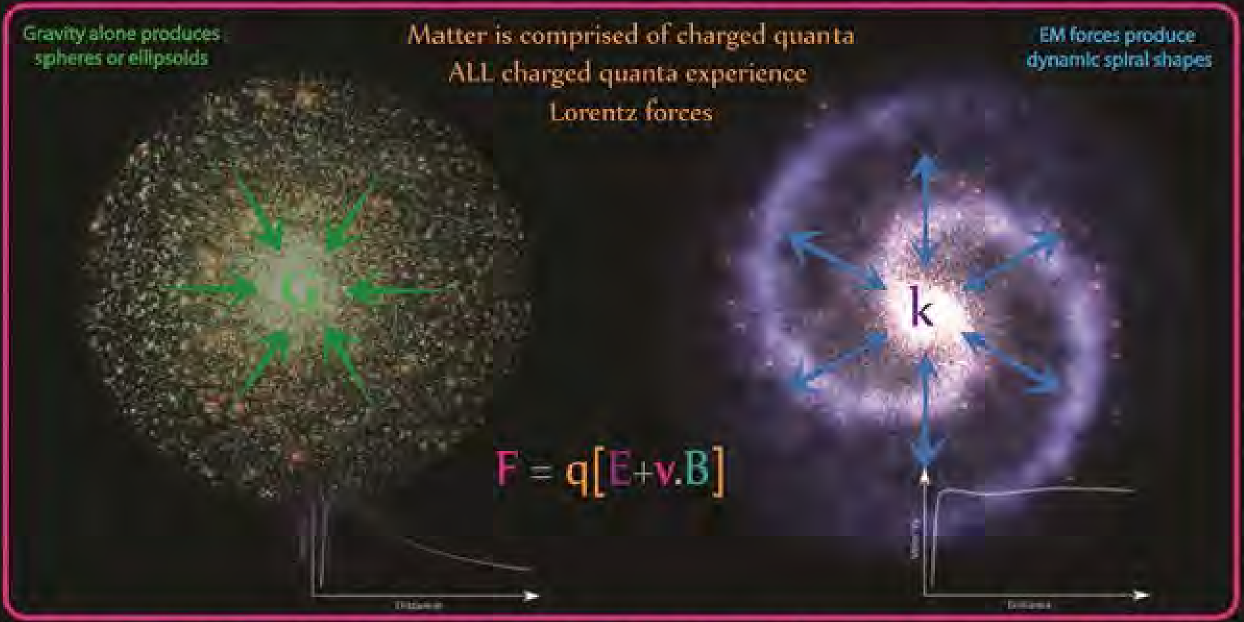

Gravity alone produces
spheres or ellipsoids

Matter is comprised of charged quanta
ALL charged quanta experience
Lorentz forces

EM forces produce
dynamic spiral shapes

$$F = q[E+v.B]$$

*The curves of the rotational velocities of stars within Galaxies do not decrease in the expected inverse square root relationship
but are "flat", i.e. outside of the central bulge their speed is nearly constant from the centre out ot the edge of the galaxy
[identical to the forces produced by homopolar motors]*

SR

**ElectroMagnetic
mass-Energy
is interactive**

General relativity
models interactive
GEM energies

$$F = k\frac{Q_1}{r^2}$$

*Tetryonics dictates that we must take into account EM forces as well as Gravity
in order to account for any perturbations that convergent Gravitation alone cannot account for*

Observations have shown there is not sufficient visible Matter in the Universe to account for the apparent strength of gravitational forces within and between galaxies leading to the hypothesis that up to 90% of the Matter in the Universe is

Dark Matter

***Gravity alone does not hold the whole galaxy together
It is many orders of magnitude weaker than ElectroMagnetic forces***

Gravitational attraction is the nett convergent force resulting from three distinct quantum forces

To account for the redshift-magnitude relation of type Ia supernovae General relativity requires that much of the energy in the Universe consists of a component with large negative pressure, this negative-gravity has been dubbed

Dark Energy

Dark Energy

Dark energy is a hypothetical form of energy that permeates all of space and tends to increase the rate of expansion of the universe.

EM mass-energies are conservative

Dark Energy is a nett repulsive force that results when Gravitational Matter topologies are destroyed by stellar GEM pinch processes

Matter is not conservative

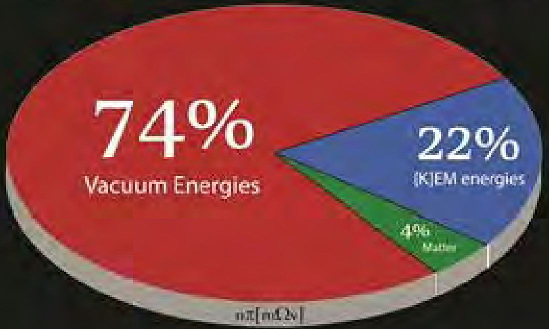

Independent from its actual nature, dark energy would need to have a strong negative pressure (repulsive force or divergent energy opposite that of gravitational attraction) in order to explain the observed acceleration in the expansion rate of the universe.

Dark Energy is the divergent scalar field created by [K]EM field energies

2D [K]EM fields create interactive forces between Matter

The cosmological constant is physically equivalent to vacuum energy

All [K]EM radiation has a divergent component in its interactive field that increases when Matter is in motion

DARK ENERGY

or its standing-wave topology is destroyed

As Matter is destroyed and energy is released the attractive force of Gravity decreases and the divergent acceleration due to Dark Energy increases

The [K]EM fields of Matter
are composed of
[illegible]
DARK ENERGY
Divergent EM field interactions
The [K]EM energy released by divergent
KEM fields contributes to the total energies
of the cyclic Universe

DARK MATTER
Convergent EM field interactions

A major outstanding problem is that most quantum field theories predict a huge energy value for the quantum vacuum, more than 100 orders of magnitude too large.

This would need to be cancelled almost, but not exactly, by an equally large term of the opposite sign.

DARK ENERGY
Divergent EM field interactions

In fact a negative pressure does not influence the gravitational interaction between Matter [which remains attractive] but rather alters the overall evolution of the universe at the cosmological scale, typically resulting in the localised accelerating expansion of the universe despite the attraction between Matter present throughout the Universe.

Tetryonics 80.06 - Dark Energy

Energy-momenta geometry

A major outstanding problem is that most quantum field theories predict a huge cosmological constant from the energy of the quantum vacuum, more than 100 orders of magnitude too large.

If the Higgs field does exist, another field with an unknown negative energy is needed to reduce the total energy in the vacuum to a slightly positive value to correspond with the current accelerated rate of expansion

Summing over all possible Energy quanta at all points in space gives a large but NOT infinite quantity of energy-momenta

The super-positioned EM fields of free space is comprised of Planck energy momenta

Dark Energy is Divergent Vacuum Energy momenta in free Space

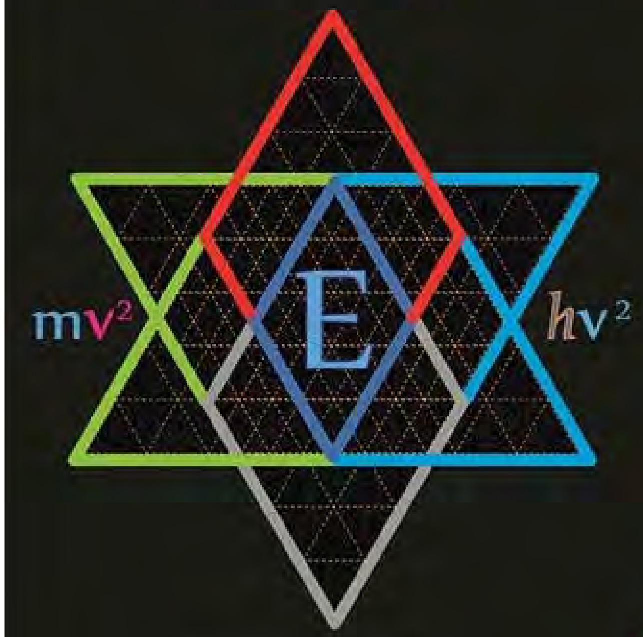

Convergent Vacuum Energy momenta in free Space is Dark Matter

The total Vacuum Energy in any spatial region contains fields of Planck energy-momenta seeking an equalised, homgenous state

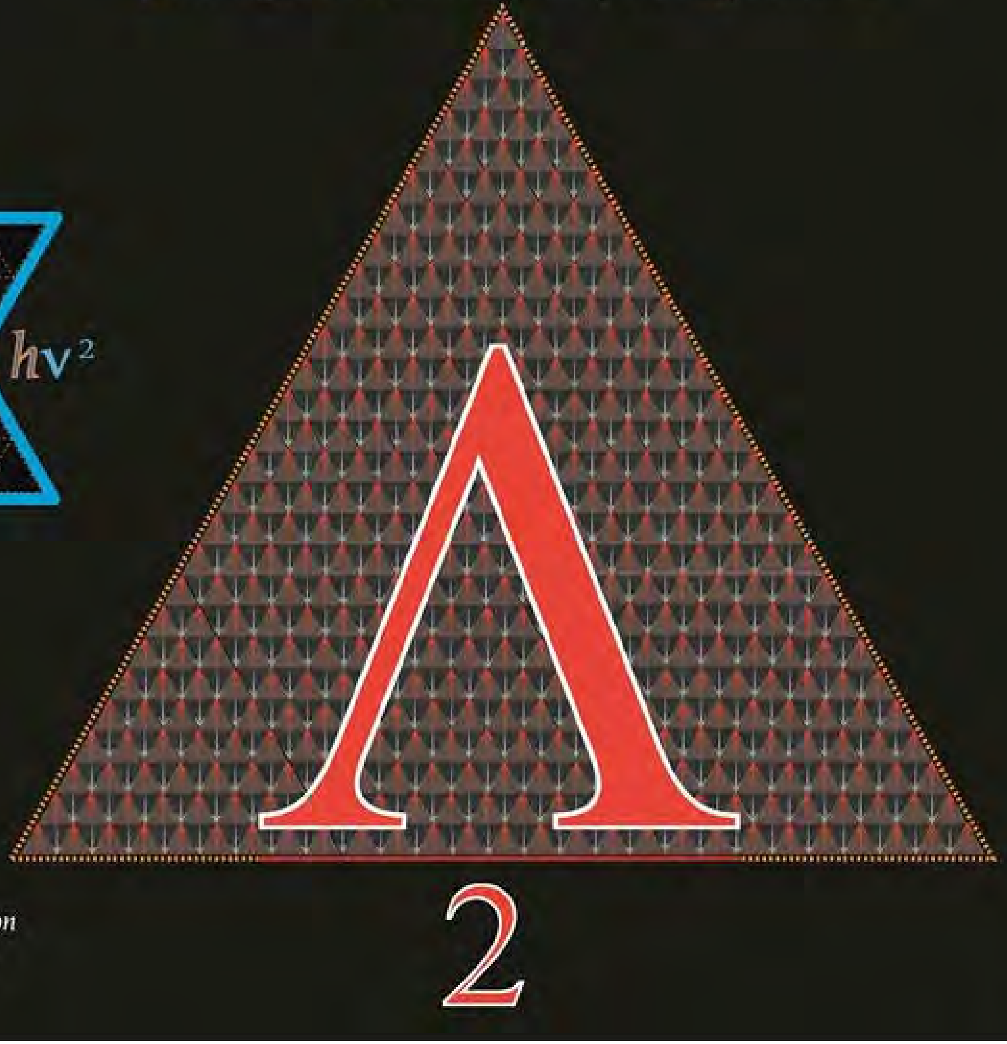

Due to the equilateral asymmetry of all the energy momenta quanta in charged electromagnetic fields

E Field Coulombic forces

convergent momenta ↓

divergent momenta ↑

Vacuum charge momentum

1	1		1
2	3		0.666667
3	5		0.6
4	7		0.571429
5	9		0.555556
6	11		0.545455
7	13		0.538462
8	15		0.533333
9	17		0.529412
10	19		0.526316
11	21		0.52381
12	23		0.521739
13	25		0.52
14	27		0.518519
15	29		0.517241
16	31		0.516129
17	33		0.515152
18	35		0.514286
19	37		0.513514
20	39		0.512821
21	41		0.512195
22	43		0.511628
23	45		0.511111
24	47		0.510638
25	49		0.510204
26	51		0.509804
27	53		0.509434
28	55		0.509091
29	57		0.508772
30	59		0.508475

Divergent energy momenta always dominates over Convergent energy momenta

Tetryonics 80.07 - Vacuum Energy-momenta

Copyright ABRAHAM [2008] - All rights reserved

Charged electrostatic fields

E Field Coulombic forces

convergent momenta

divergent momenta

Charge interaction

1	1		1
2	3		0.666667
3	5		0.6
4	7		0.571429
5	9		0.555556
6	11		0.545455
7	13		0.538462
8	15		0.533333
9	17		0.529412
10	19		0.526316
11	21		0.52381
12	23		0.521739
13	25		0.52
14	27		0.518519
15	29		0.517241
16	31		0.516129
17	33		0.515152
18	35		0.514286
19	37		0.513514
20	39		0.512821
21	41		0.512195
22	43		0.511628
23	45		0.511111
24	47		0.510638
25	49		0.510204
26	51		0.509804
27	53		0.509434
28	55		0.509091
29	57		0.508772
30	59		0.508475

7/15

DIVERGENT CHARGE FIELDS
are external Energy momenta

momenta

CONVERGENT CHARGE FIELDS
are internal Energy momenta

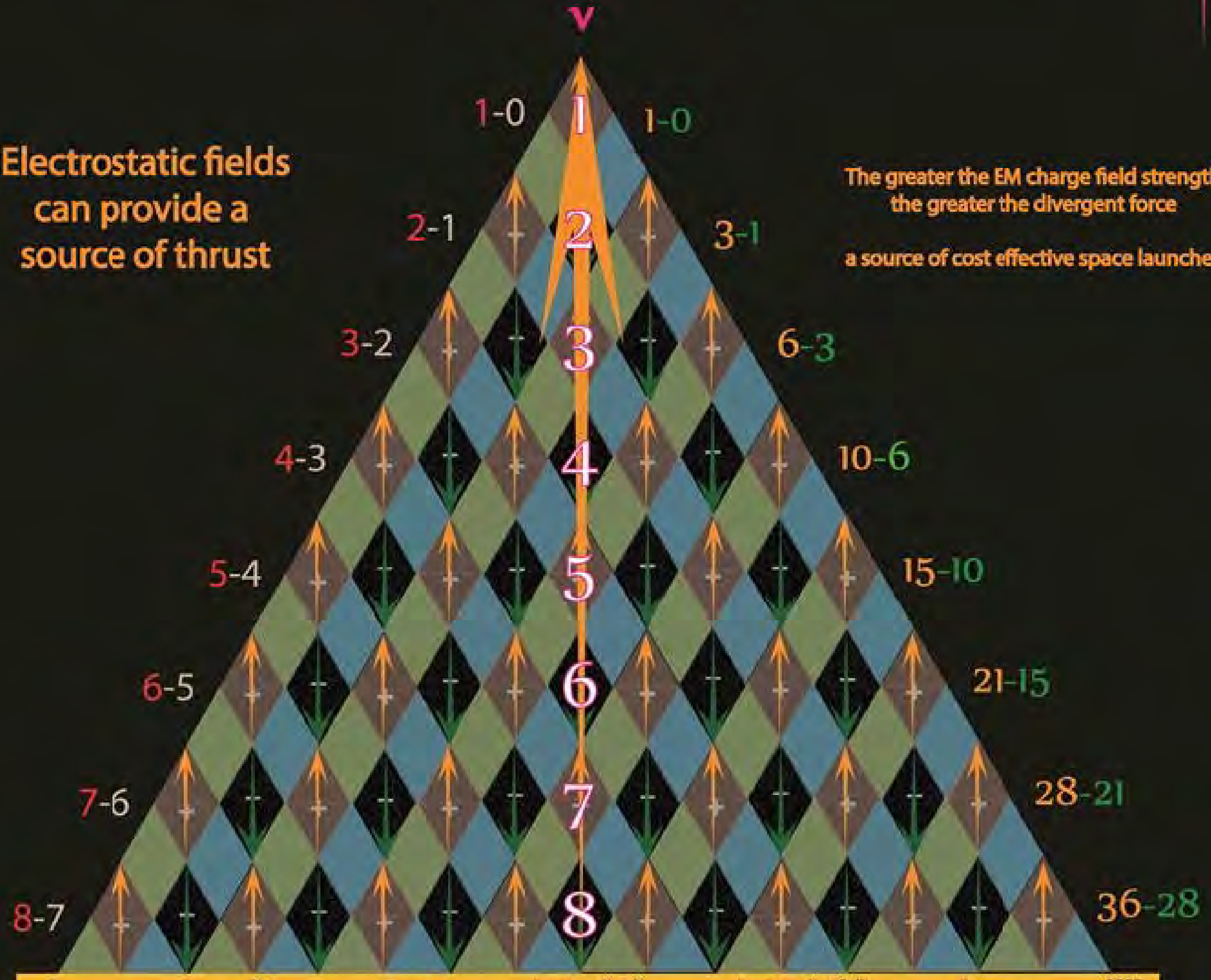

Electrostatic fields can provide a source of thrust

The greater the EM charge field strength the greater the divergent force

a source of cost effective space launches

Energy-momenta forces

Forward velocity

downward Thrust

Momenta acceleration

1	0
3	.25
2	.333
1.666	.375
1.5	.4
1.4	.4166
1.333	.4375
1.2857	.4375

#		Pos	Neg		Charge		Pos %	Neg %		Pos Split	Neg Split		Power		SQRT
1000		1000	999		1		0.50025	0.49975		500500	499500		1000000		1000

Launch Pad

Momentum splits are reversed for Negative charges

Electric Charge divergence is always superior to Charge convergence

Momentum trends to 50/50 split (but never reaches it)

Launch Pad

Tetryonics 80.08 - Electro-Gravitic Charge lauches

Electrostatic lifters

Electrostatic levitation is the process of using an electric field to levitate a charged object and counteract the effects of gravity

Tetryonic geometries and Coulombic forces offers an alternative means of propulsion for space lauches and frictionless transports

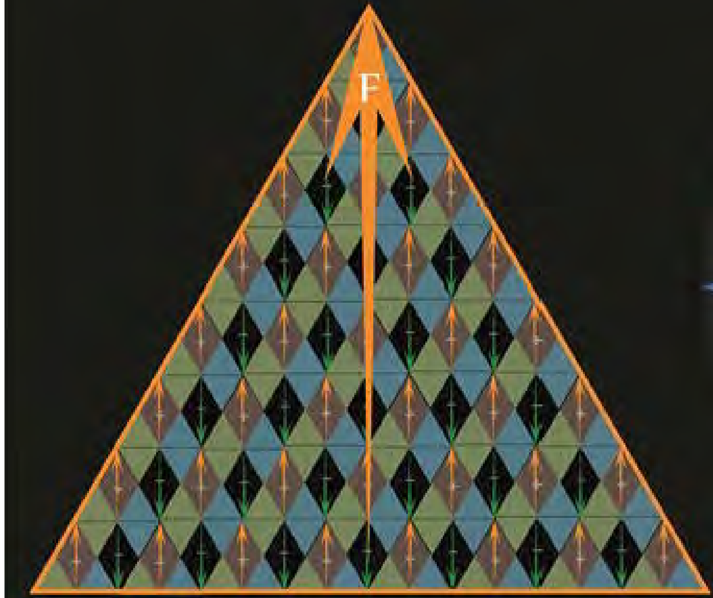

Charged voltage fields contain nett divergent energy momenta [Planck quanta] that can do work over distances

Electrostatic charges are always divergent from their sources

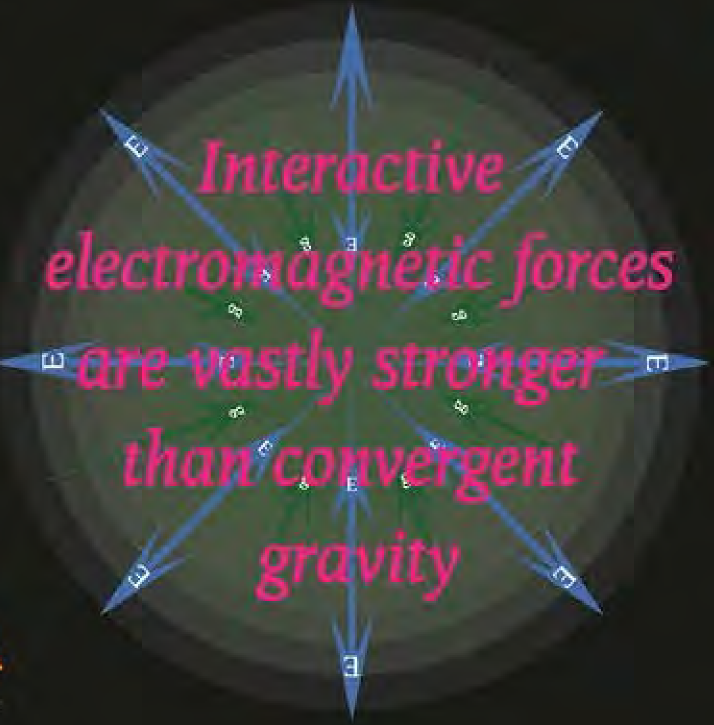

Interactive electromagnetic forces are vastly stronger than convergent gravity

Static electricity is the result of electrostatic charges

Electrostatic charges are comprised of divergent energy momenta

Electrostatic levitation

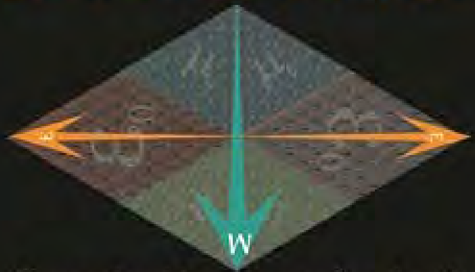

forces should not be confused with

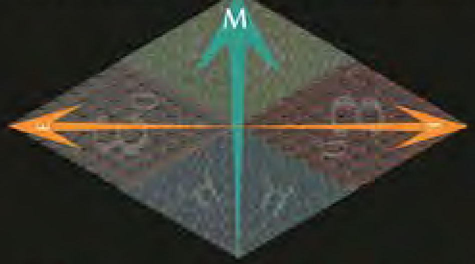

Magnetic levitation

Copyright ABRAHAM [2008] - All rights reserved

The Light speed barrier

BOSONS/PHOTONS
2D Planar massless particles
[EM mass-Energies]
'slice' throught the
vacuum energy aether

CHARGED
particle or massive body

ENERGY-MOMENTA
of populsion field

The Compton frequency of quanta
in any EM field is directly proportional
to the field's velocity of propagation

**The 'speed of Light'
is the limiting speed of
transverse EM accelerations
[it is not THE limit]**

$$E = Mv^2$$

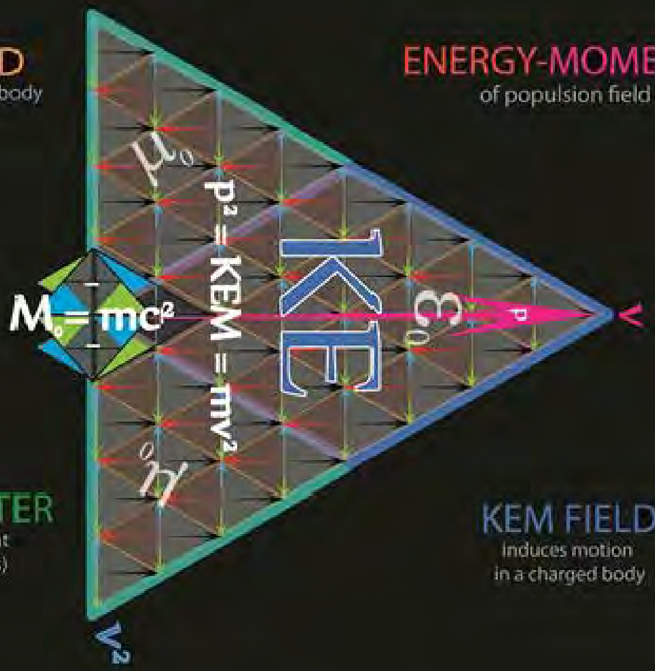

$$KE = \frac{1}{2}Mv^2$$

MOTION VECTOR
direction of propulsion
of a charged body

REST MATTER
remains invariant
(but accelerates)

KEM FIELD
induces motion
in a charged body

TETRYONS/BARYONS
3D Tetrahedral EM Matter
[4nπ EM mass-Energies]
interact at various angles with the
vacuum energy aether

**Tetryonic geometry clearly shows that here is NO limit to the
energy content that can be released into any spatial region,
negating the long held belief of a Light speed barrier**

V

According to special relativity, c is the maximum speed at which all EM mass-Energy and information in the universe can travel.
It is predicted by General relativity to be the absolute velocity of all 'massless particles' and associated fields
- Including ElectroMagnetic radiation such as Light in a vacuum.

Such particles and waves travel at c regardless of the motion of the source or the inertial frame of reference of the observer.
In the theory of relativity, c interrelates space and time, and appears in the famous equation of mass–energy equivalence E = mc2.

C

The speed at which light waves propagate in vacuum
is independent both of the motion of the wave source
and of the inertial frame of reference of the observer

$$E = n\pi\left[\overset{\text{Planck quanta}}{\left[\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}\right]}\right]$$

The speed of light in vacuum, usually denoted by c,
is a physical constant important in many areas of physics
and is determined by the impedance of the medium

Inflation

While the detailed particle physics mechanisms responsible for inflation are not known,
the basic picture makes a number of predictions that have been confirmed by observation.

Inflation is thus now considered part of the standard hot Big Bang cosmology

$$\underset{\text{momentum}}{\overset{\text{ENERGY}}{n\pi}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]^2\right]$$

Tetryonics now provides a detailed physical model of ALL ponderable Matter and its kinetics through the charged geometry of mass-ENERGY-Matter

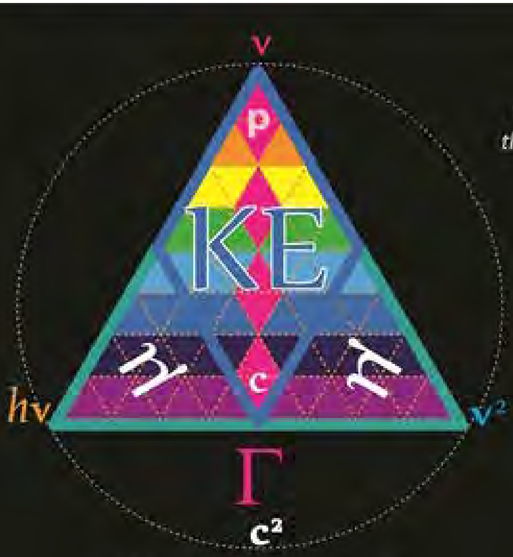

All mass-energy propagates
at a velocity directly related to
the impedance of the medium

The observable Universe
could have originated from a
region of pure energy in it's past.

This point would have contained all
the Energy that ultimately formed all the
mass-Matter of the Universe and Forces
that we now observe and measure
in two 2D planar fields

GIVEN SUFFICIENT ENERGIES ANY REGION OF SPACE-TIME
WILL UNDERGO SUPERLUMINAL EXPANSION [INFLATION]

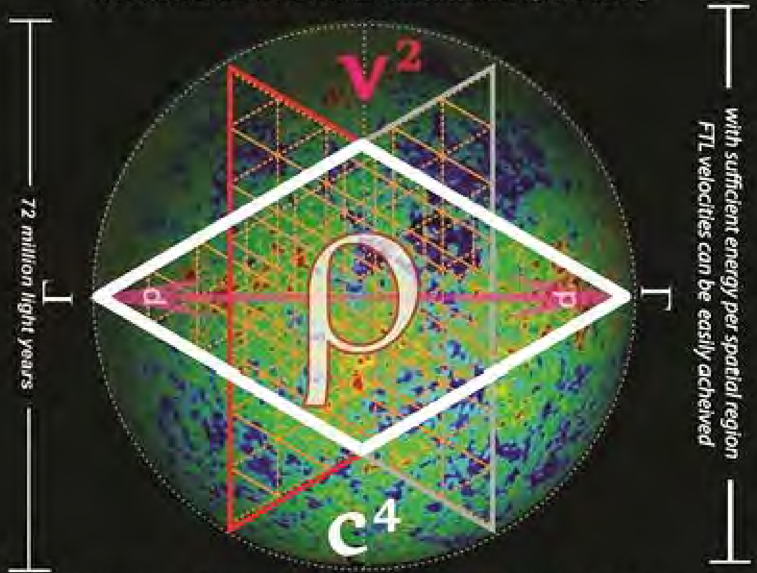
v²
ρ
p
p
c⁴
72 million light years
with sufficient energy per spatial region
FTL velocities can be easily achieved

UNTIL IT REACHES A CRITICAL ENERGY DENSITY THAT RESULTS IN
THE OBSERVED SPEED OF LIGHT THAT WE NOW MEASURE

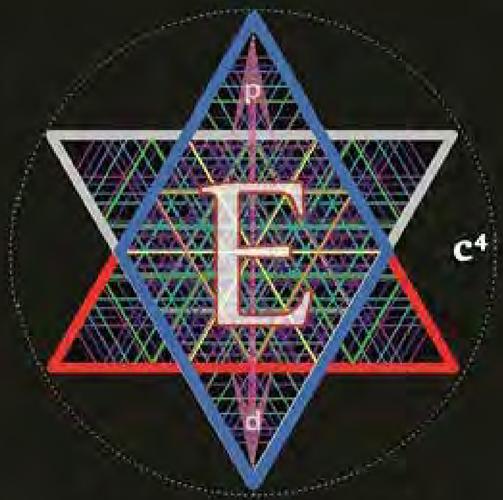

The energy in any spatial region
possesses a related mass-velocity
[linear momentum] component

"And God said,
Let there be light:
and there was light"

The planar EM Energies released
at that point would have resulted
in faster-than-light divergence from
that point as dictated by the
geometric energy-momenta
velocity relationship

$$\frac{\overset{\text{mass}}{n\pi}}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

It prescribes a dynamical EM Universe evolving over time
into a state of equilibrium between gravitational mass-Matter topologies
and radiant, interactive mass-energy momenta geometries

$$\frac{\overset{\text{Matter}}{T\pi}}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Superluminal velocities

Any velocity is obtainable by Matter providing
sufficient energy is released in very short durations
to a region of space containing a charged body

$$\Sigma F = \frac{\Delta p}{dt} = m.\frac{dv}{dt}$$

REQUIRED VELOCITY

$$\beta = \left[\frac{v}{c}\right]$$

Linear correction factor

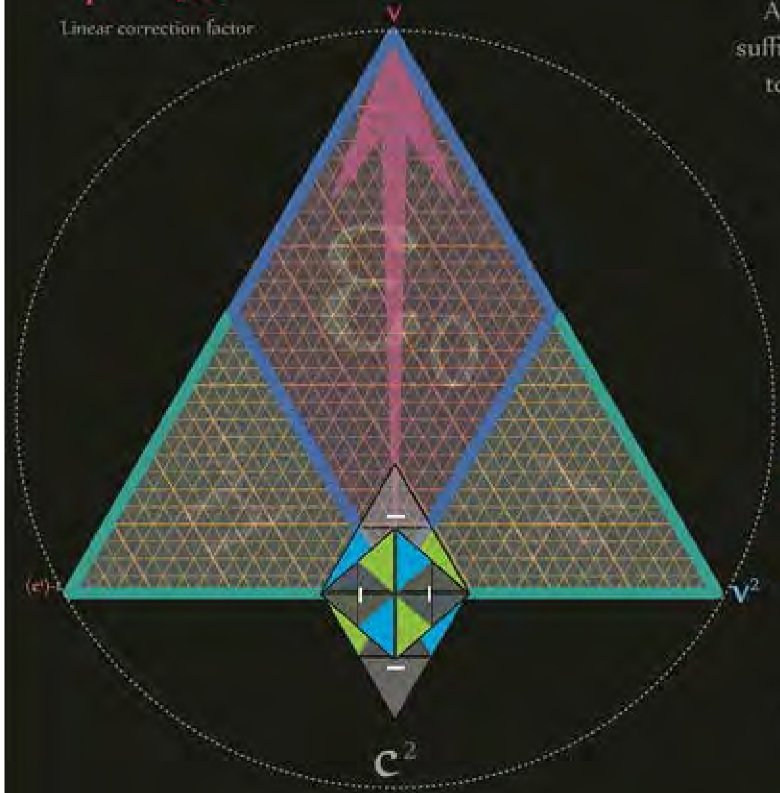

W bosons — W bosons

$$[hv]- \quad E \quad +[hv]$$

decelerate — *accelerate*

Every quantum energy level change requires a
related change in the number of quanta compared to
the previous energy level (ad infinitum)

Note: Inertial considerations still apply
Instanteous release of energy to acheive
superluminal velocities is not advised

The speed of light is a limiting velocity
for accelerations due to transverse wave
electromagnetic energies

Providing the KEM geometry contains enough
energy per c² geometry the long held belief of
'the speed of light' being an ultimate speed barrier
is shown to be invalid

KEM fields contain
energy-momenta

$$p^2 = KEM = mv^2$$

Any mass-energy field moving
at superluminal velocities is still
subject to Lorentz contractions

Any Matter held within, and moving along
with, a superluminal EM field remains
Lorentz invariant

$$\frac{mv^2}{c^2} = \frac{KEM}{c^2} = \frac{hv^2}{c^2}$$

The mass-energy of Kinetic EM
fields is velocity dependent and
subject to Lorentz corrections

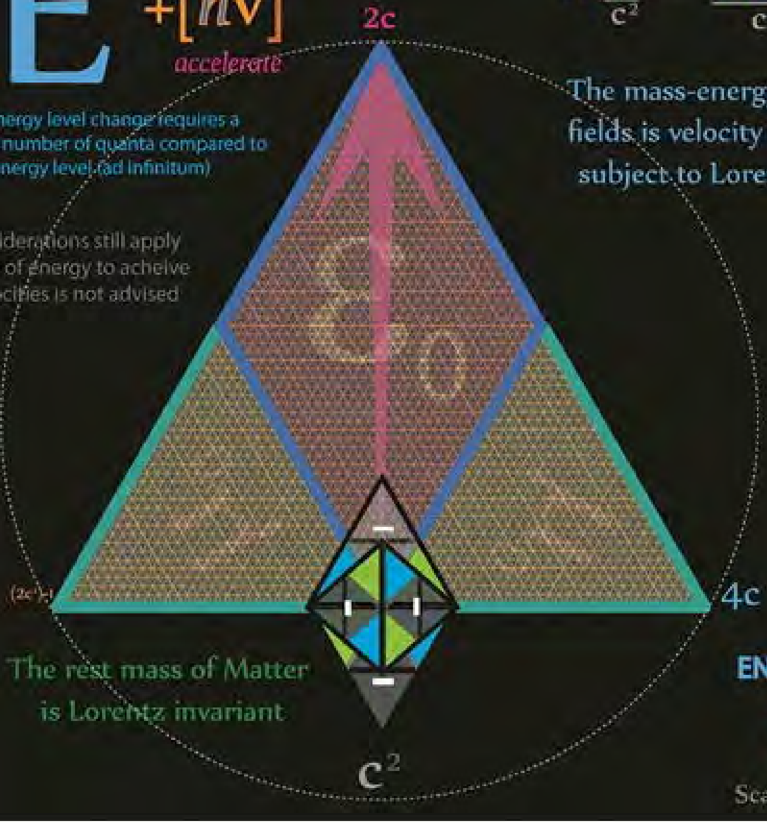

The rest mass of Matter
is Lorentz invariant

ENERGY REQUIRED

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Scalar correction factor

Copyright ABRAHAM [2008] - All rights reserved

Faster than Light Travel

It has been well established by Tetryonic geometry that the speed of light 'c' is not a insurmountable barrier as long expounded by Relativity theory

The natural speed of Light is easily surpassed given the right technologies (and sources of energy)

Any charged object in an Electric field will undergo acceleration

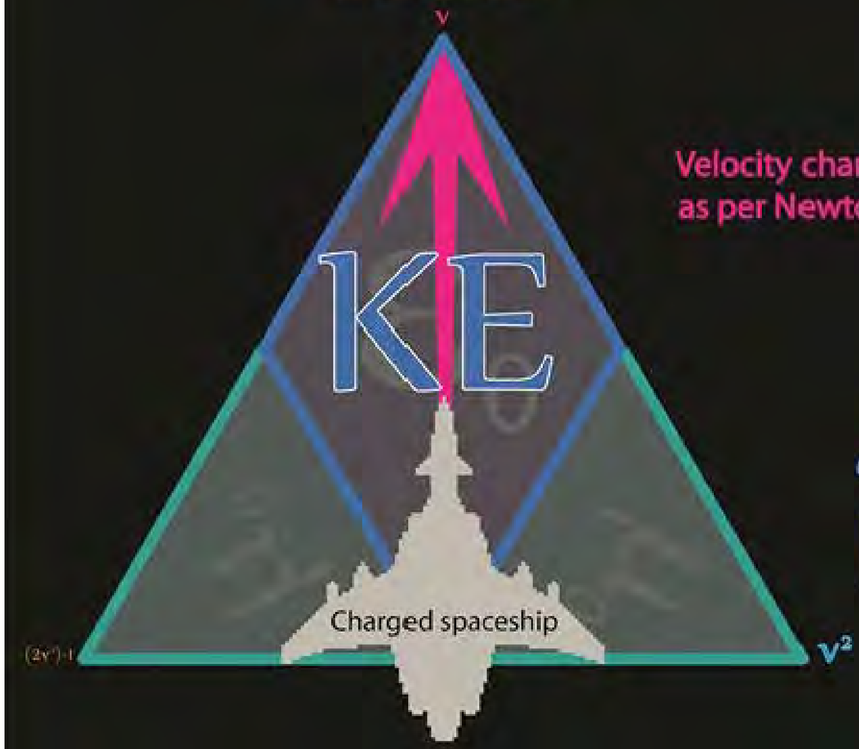

Velocity changes are additive as per Newtonian mechanics

The energy requirements to accelerate Matter to increasing velocities are exponental and follow equilateral geometries

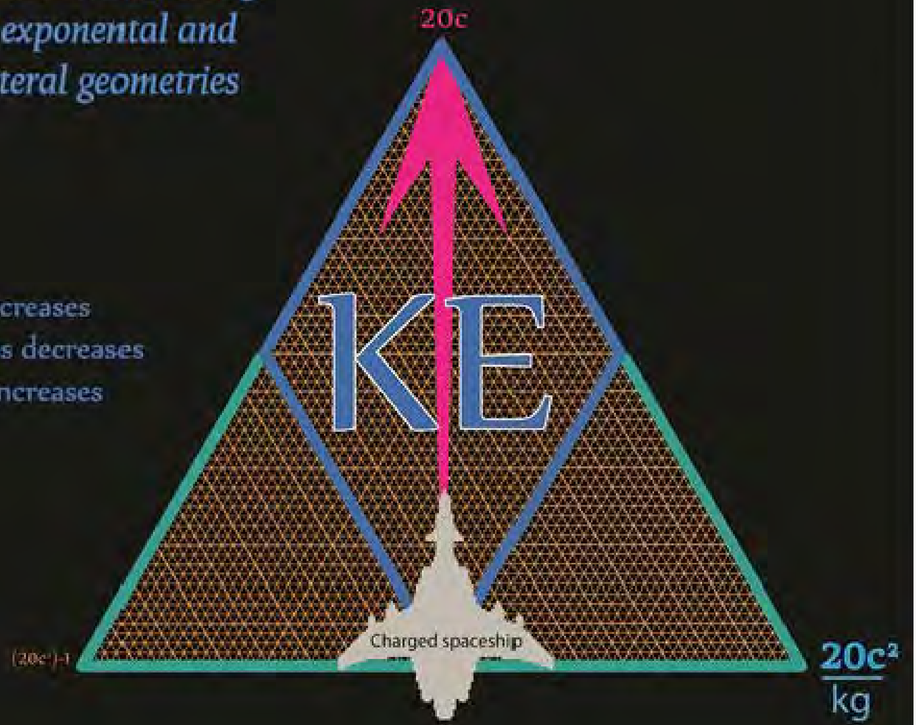

As the Spaceship's Velocity increases the wavelength of Kinetic Energies decreases and the Compton Frequency increases

Warp Speed Scotty

Proposed explanation to GAMMA RAY BURSTS [GRB]

Any massive body using KEM fields as a motive force for faster-than-Light travel would produce an extremely bright release of Gamma rays at their point of acceleration (as only 1/2 of the photonic energy would be used to accelerate the craft in any one direction). additionally anybody noticing the burst of required Energy would not find a 'source' for the Gamma ray burst when they looked as the craft would have accelerated away from that point in free space

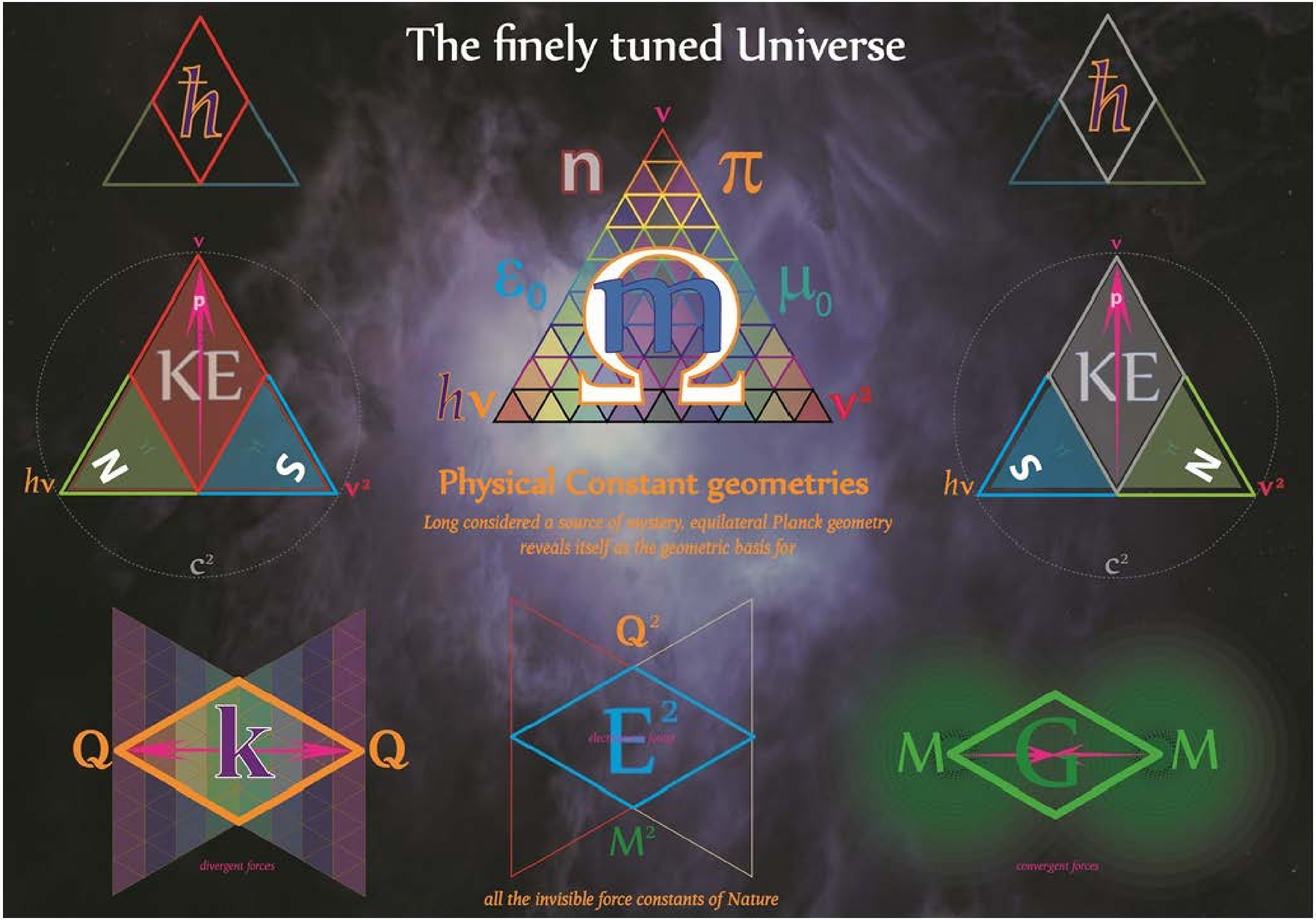

The finely tuned Universe

Physical Constant geometries

Long considered a source of mystery, equilateral Planck geometry reveals itself as the geometric basis for

all the invisible force constants of Nature

ħ
ħ
v
n
π
ε0
μ0
m
Ω
hv
v²
v
p
KE
N
S
hv
v²
c²
v
p
KE
S
N
hv
v²
c²
Q
k
Q
divergent forces
Q²
E²
M²
M
G
M
convergent forces

Tetryonics 80.14 - Physical Constant geometries

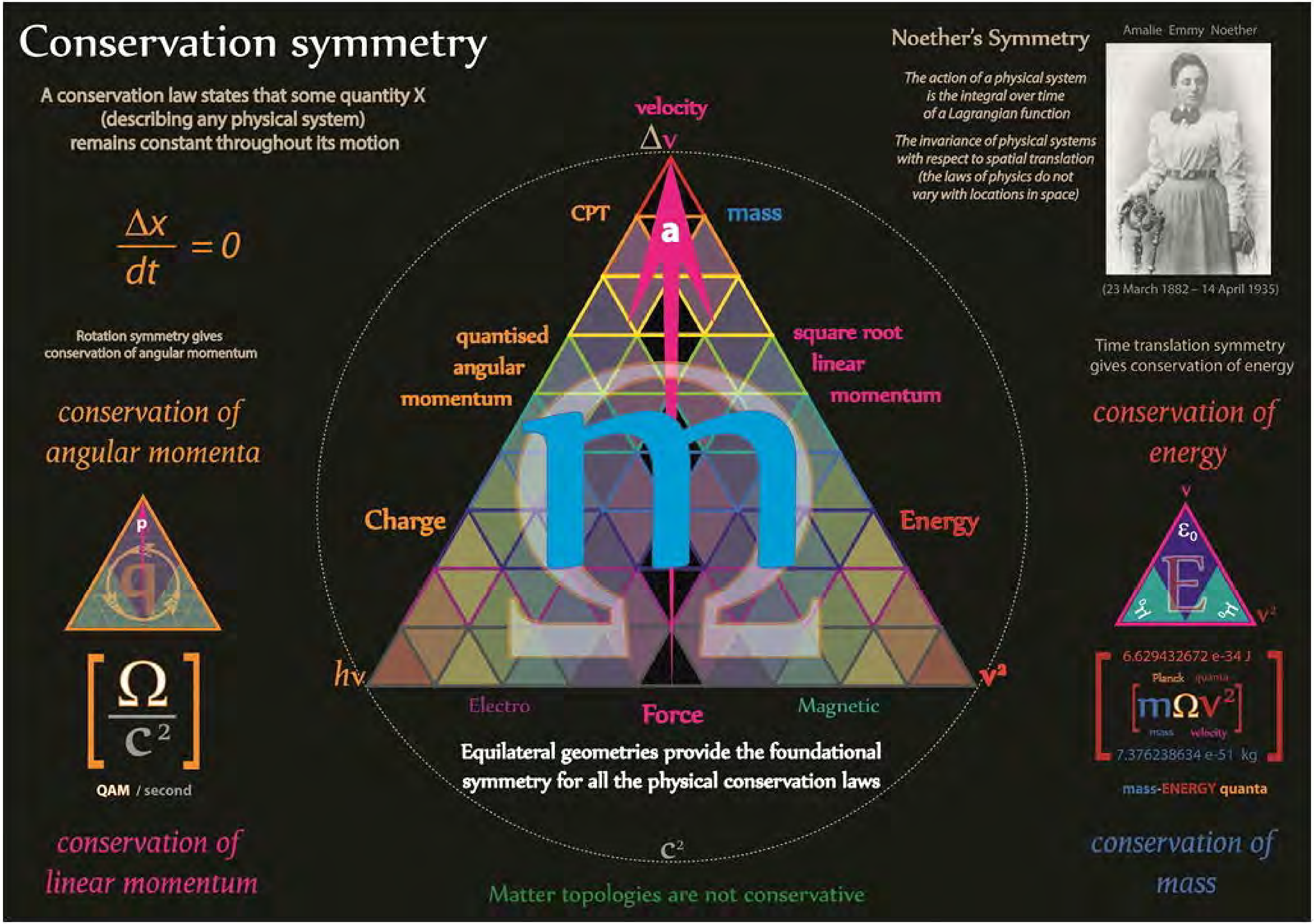
velocity
Δv
CPT
mass
a
quantised
angular
momentum
square root
linear
momentum
m
Ω
Charge
Energy
hv
v^2
Electro
Force
Magnetic
c^2
p
q
v
ε_0
E
v^2

Conservation symmetry

A conservation law states that some quantity X
(describing any physical system)
remains constant throughout its motion

$$\frac{\Delta x}{dt} = 0$$

Rotation symmetry gives
conservation of angular momentum

conservation of
angular momenta

$$\left[\frac{\Omega}{c^2}\right]$$

QAM / second

conservation of
linear momentum

Equilateral geometries provide the foundational
symmetry for all the physical conservation laws

Matter topologies are not conservative

Noether's Symmetry

The action of a physical system
is the integral over time
of a Lagrangian function

The invariance of physical systems
with respect to spatial translation
(the laws of physics do not
vary with locations in space)

Amalie Emmy Noether

(23 March 1882 – 14 April 1935)

Time translation symmetry
gives conservation of energy

conservation of
energy

6.629432672 e-34 J

Planck quanta

$$\left[m\Omega v^2\right]$$

mass velocity

7.376238634 e-51 kg

mass-ENERGY quanta

conservation of
mass

Tetryonics 80.15 - Conservation symmetry

Virtual Particles

In physics, a virtual particle is a particle that exists for a limited time and space.
The energy and momentum of a virtual particle are uncertain according to the uncertainty principle.
The degree of uncertainty of each is inversely proportional to
time duration (for energy) or to position span (for momentum).

In Tetryonics the Uncertainty principle has been negated with deterministic, re-normalisable charged mass-ENERGY geometries & Matter topologies

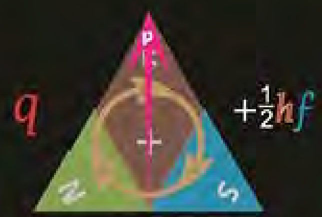
q
$+\frac{1}{2}hf$
P
+
N
S

Positive Zero Point Field
energy momenta

Longitudinal waves facilitate
action-at-a-distance

*Convergent Gravitation
is attractive*

Material Gravitation

M

Charged Matter topologies
in vacuum energy fields
create Gravity

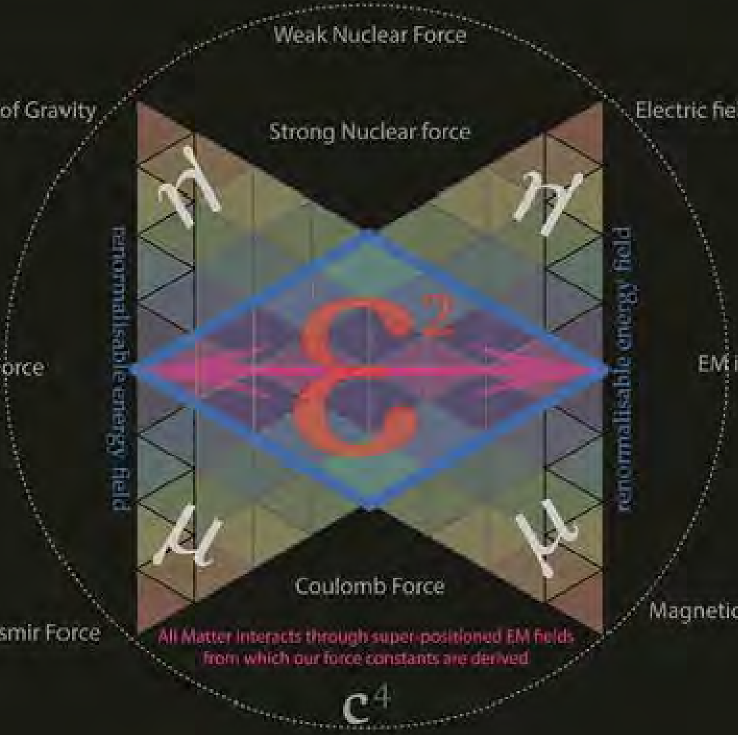
Weak Nuclear Force
Force of Gravity
Electric fields
Strong Nuclear force
η
η
renormalisable energy field
renormalisable energy field
Van der Waals Force
EM induction
E^2
μ
μ
Coulomb Force
Casmir Force
Magnetic fields
c^4

All Matter interacts through super-positioned EM fields
from which our force constants are derived

$-\frac{1}{2}hf$
q
P
-
S
N

Negative Zero Point Field
energy momenta

Energy propagates at the
speed of light

*Divergent Coulomb forces
are interactive*

EM Interactions

m

Superpositioned EM fields
create interactive
EM forces

Vacuum field Energies

$$m^2c^4 = E^2 = p^2c^2$$

EM mass momenta

virtual Particles are not a required part of tetryonic theory

Tetryonics 80.16 - Virtual Particles

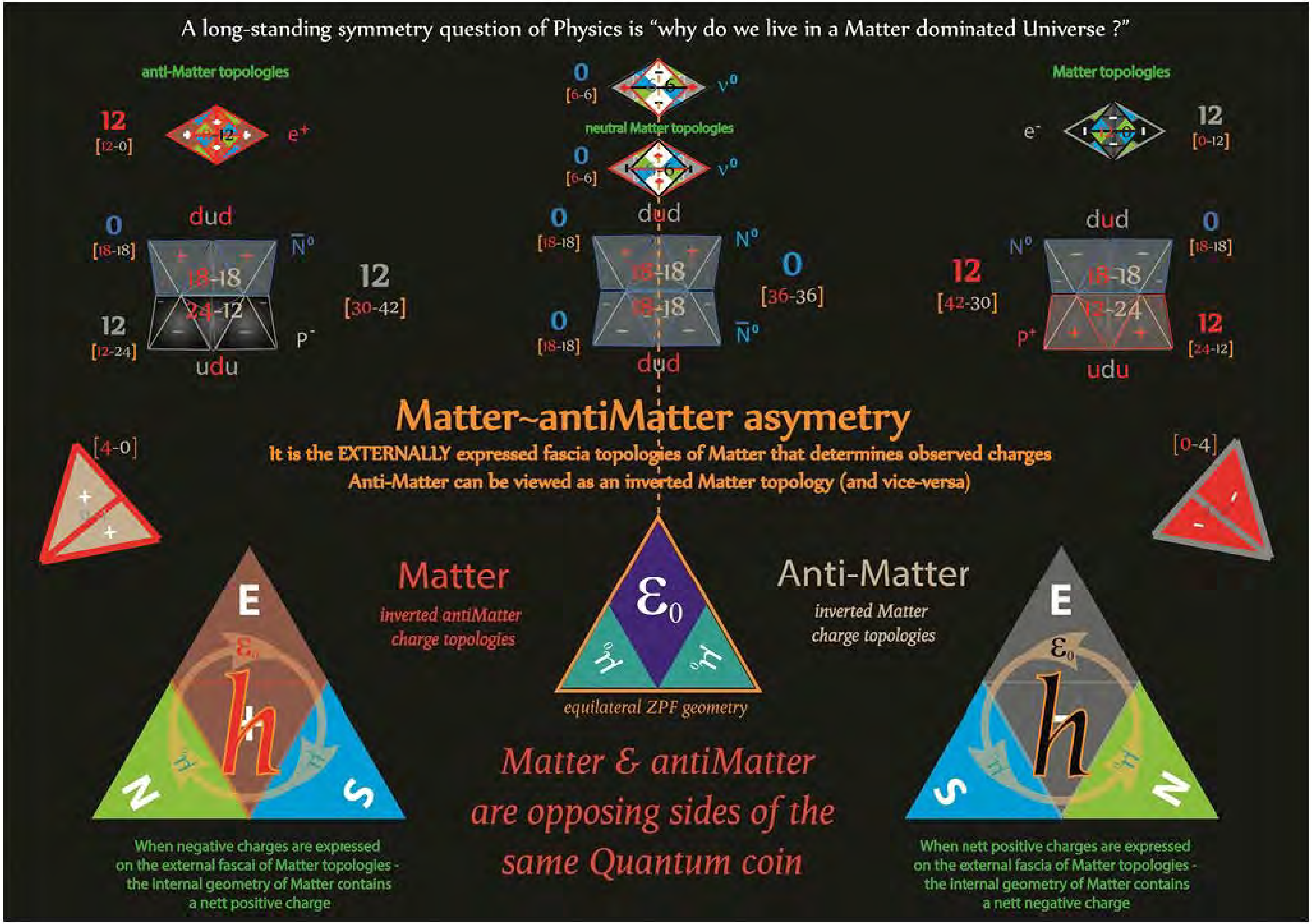
anti-Matter topologies
12
[12-0]
e^+
0
[6-6]
ν^0
neutral Matter topologies
0
[6-6]
ν^0
Matter topologies
e^-
12
[0-12]
0
[18-18]
dud
$\bar{N}^0$
18-18
24-12
12
[12-24]
P^-
udu
12
[30-42]
0
[18-18]
dud
N^0
18-18
18-18
0
[18-18]
$\bar{N}^0$
dud
0
[36-36]
12
[42-30]
N^0
dud
18-18
12-24
P^+
udu
0
[18-18]
12
[24-12]
[4-0]
[0-4]
E
ε_0
h
N
S
ε_0
μ_0
μ_0
E
ε_0
h
S
N

A long-standing symmetry question of Physics is "why do we live in a Matter dominated Universe ?"

Matter~antiMatter asymetry

It is the EXTERNALLY expressed fascia topologies of Matter that determines observed charges

Anti-Matter can be viewed as an inverted Matter topology (and vice-versa)

Matter

inverted antiMatter charge topologies

Anti-Matter

inverted Matter charge topologies

equilateral ZPF geometry

Matter & antiMatter are opposing sides of the same Quantum coin

When negative charges are expressed on the external fascai of Matter topologies - the internal geometry of Matter contains a nett positive charge

When nett positive charges are expressed on the external fascia of Matter topologies - the internal geometry of Matter contains a nett negative charge

Tetryonics 80.17 - Charged fascia chirality

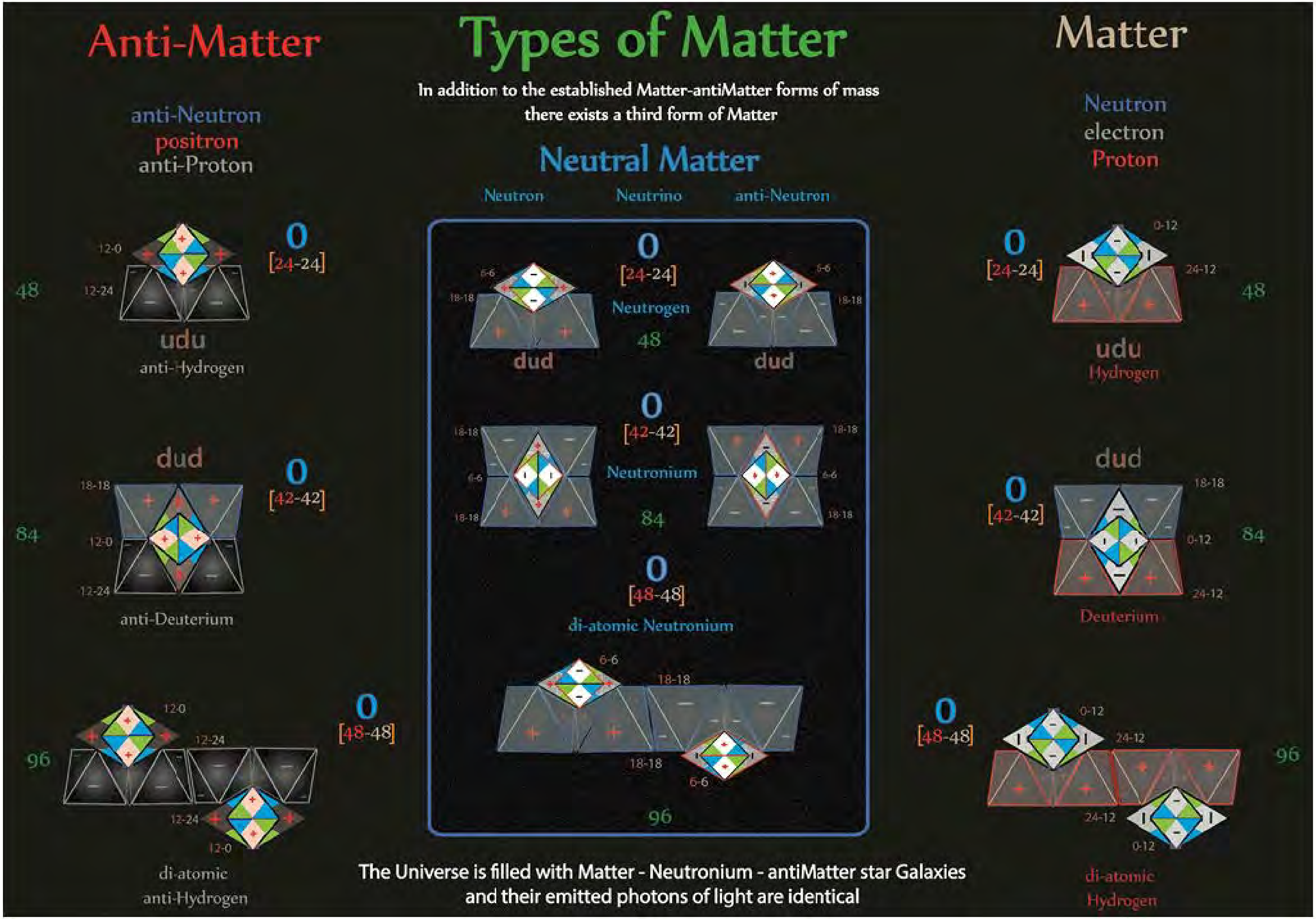

Types of Matter

In addition to the established Matter-antiMatter forms of mass
there exists a third form of Matter

Neutral Matter

Anti-Matter
anti-Neutron
positron
anti-Proton
12-0
12-24
48
0
[24-24]
udu
anti-Hydrogen
dud
18-18
12-0
12-24
84
0
[42-42]
anti-Deuterium
12-0
12-24
96
12-24
12-0
0
[48-48]
di-atomic
anti-Hydrogen
Neutron
Neutrino
anti-Neutron
6-6
18-18
0
[24-24]
Neutrogen
48
dud
dud
6-6
18-18
18-18
6-6
18-18
0
[42-42]
Neutronium
84
18-18
6-6
18-18
0
[48-48]
di-atomic Neutronium
6-6
18-18
18-18
6-6
96
Matter
Neutron
electron
Proton
0
[24-24]
0-12
24-12
48
udu
Hydrogen
dud
0
[42-42]
18-18
0-12
24-12
84
Deuterium
0
[48-48]
0-12
24-12
24-12
0-12
96
di-atomic
Hydrogen

The Universe is filled with Matter - Neutronium - antiMatter star Galaxies
and their emitted photons of light are identical

Copyright ABRAHAM [2008] - All rights reserved

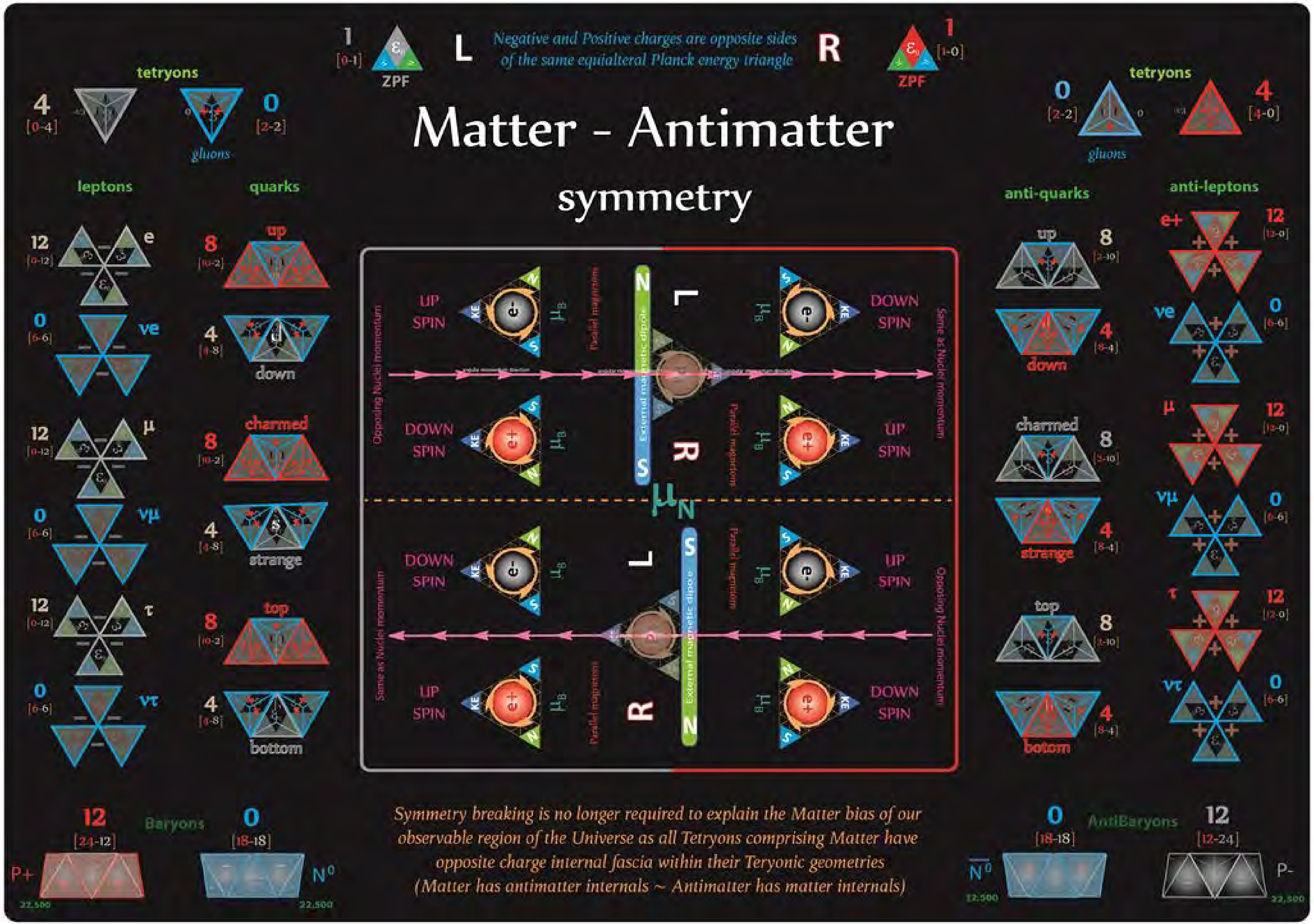

Negative and Positive charges are opposite sides of the same equialteral Planck energy triangle

Symmetry breaking is no longer required to explain the Matter bias of our observable region of the Universe as all Tetryons comprising Matter have opposite charge internal fascia within their Teryonic geometries (Matter has antimatter internals ~ Antimatter has matter internals)

Matter - Antimatter
symmetry
1
[0-1]
ZPF
L
R
1
[1-0]
ZPF
tetryons
4
[0-4]
0
[2-2]
gluons
tetryons
0
[2-2]
gluons
4
[4-0]
leptons
quarks
anti-quarks
anti-leptons
12
[0-12]
e
0
[6-6]
ve
12
[0-12]
μ
0
[6-6]
vμ
12
[0-12]
τ
0
[6-6]
vτ
8
[10-2]
up
4
[4-8]
down
8
[10-2]
charmed
4
[4-8]
strange
8
[10-2]
top
4
[4-8]
bottom
up
8
[2-10]
4
[8-4]
down
charmed
8
[2-10]
4
[8-4]
strange
top
8
[2-10]
4
[8-4]
botom
e+
12
[12-0]
ve
0
[6-6]
μ
12
[12-0]
vμ
0
[6-6]
τ
12
[12-0]
vτ
0
[6-6]
UP SPIN
DOWN SPIN
DOWN SPIN
UP SPIN
DOWN SPIN
UP SPIN
UP SPIN
DOWN SPIN
e-
e+
KE
μ_B
N
S
L
R
External magnetic dipole
Parallel magnetons
Opposing Nuclei momentum
Same as Nuclei momentum
μ_N
12
Baryons
[24-12]
P+
22,500
0
[18-18]
N⁰
22,500
0
AntiBaryons
[18-18]
$\overline{N}^0$
22,500
12
[12-24]
P-
22,500

Tetryonics 80.19 - Matter - Antimatter symmetry

Baryon asymmetry

Anti-Matter

Matter

The Big Bang should have produced equal amounts of Matter and antiMatter, as such, there should have been total cancellation of both.

It is not yet understood why the Universe has more matter than antimatter.
It is generally assumed that when the Universe was young and very hot,
it was in equilibrium and contained equal numbers of baryons and antibaryons.

Protons should have cancelled with antiprotons, electrons with antielectrons (positrons),
neutrons with antineutrons, and so on for all elementary particles.
This would have resulted in a sea of photons in the universe with no matter.

However, observations suggest that the Universe, including its most distant parts,
is made almost entirely of normal matter.

dud

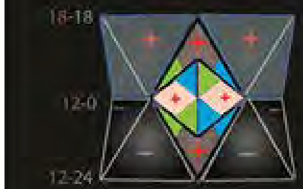

0
[42-42]
84

anti-Deuterium

dud

0
[42-42]
84

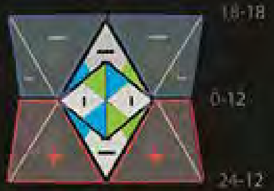

Deuterium

anti-Neutron
positron
anti-Proton

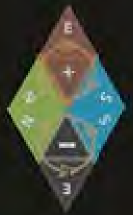

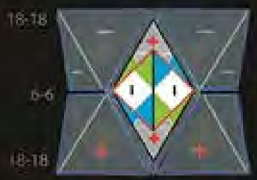

0
[42-42]
Neutronium
84

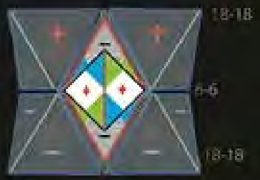

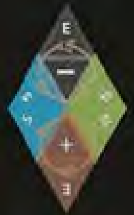

Neutron
electron
Proton

antiMatter dominated Galaxy

Matter dominated Galaxy

Neutral Matter

Neutron Neutrino anti-Neutron

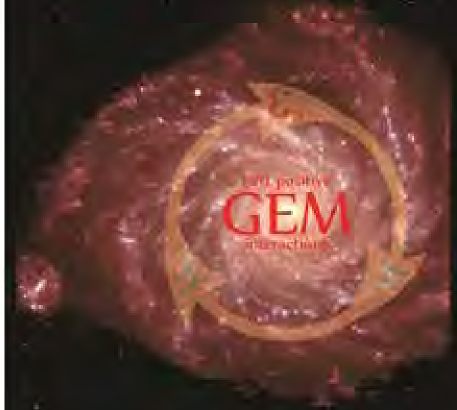

Clockwise rotation

Tetryonics dictates that the Universe is filled with Matter - Neutronium - Antimatter star Galaxies and their emitted light photons are identical

Tetryonics also demonstrates that the charge topologies of each Matter type prevalent in the formation of each Galaxy determines the overall resultant EM field geometries that shapes their large-scale evolution

It is hypothesised that the rotation direction of galaxies is also influenced by how each galaxy was formed
in particular the predominance of Matter over antiMatter interactions and
the electrodynamic fields of each Galaxy's GEM pinch core

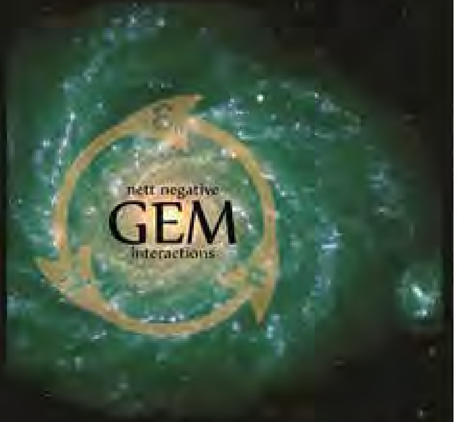

Counter-clockwise rotation

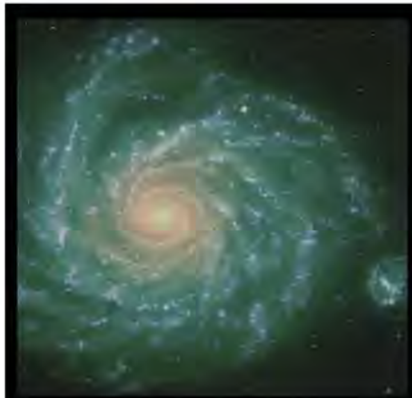

Opposites attract - Similars repel
even on the galactic-scale

Galactic Matter distributions

Tetryonics dictates that the Universe is filled with 3 forms of
Matter - Neutronium - Antimatter stars and Galaxies
their light is identical and cannot be used to distingush them

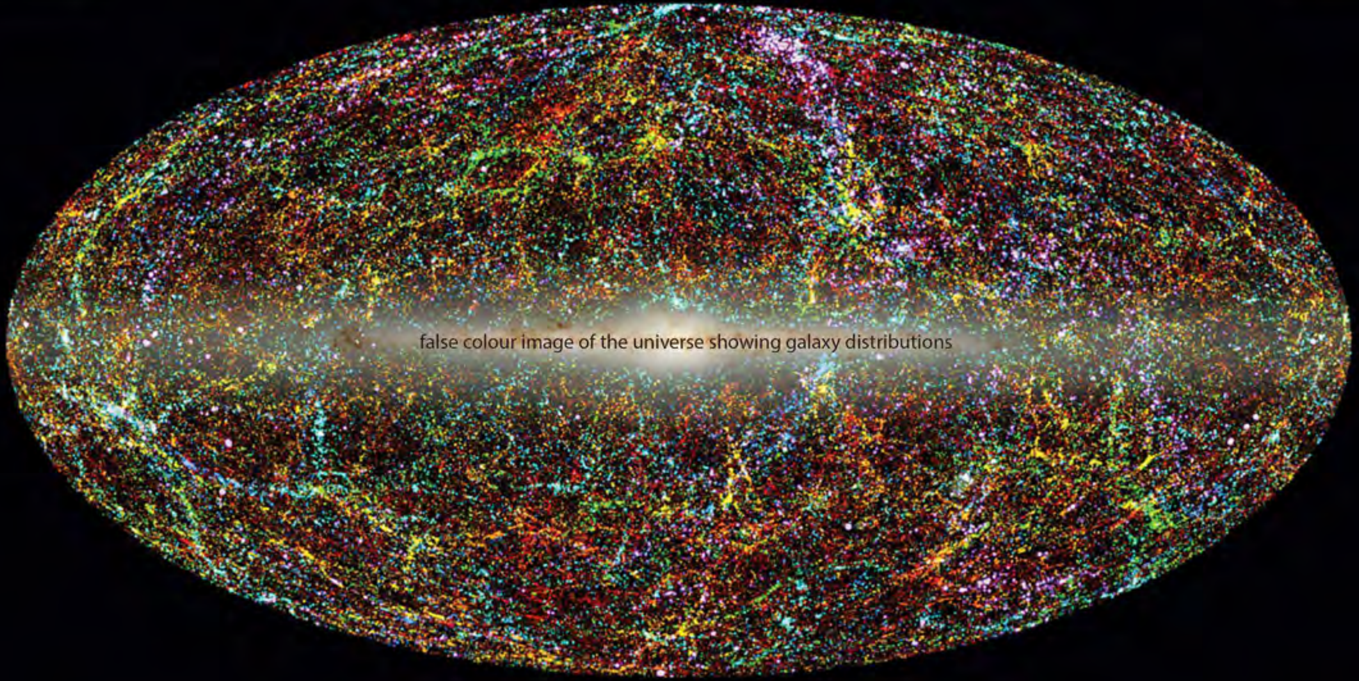

false colour image of the universe showing galaxy distributions

Matter

Neutron
Electron
Proton

Neutronium

Neutron Neutrino anti-Neutron

Neutronium is found predominantly in Neutron stars and Pulsars
spread throughout Matter and antiMatter galaxies
however it is conceivable that neutral galaxies may exist

Anti-Matter

anti-Neutron
Positron
anti-Proton

Tetryonics 80.21 - Galactic Matter distribution

Copyright ABRAHAM [2008] - All rights reserved

Planck

$$h = m.\Omega$$

quantised angular momentum is equilateral geometry

$$E = hv^2$$

E

scalar Energy

$$kg\,\frac{m^2}{s^2}$$

h

Planck's constant

$$kg\,\frac{m^2}{s}$$

Leibniz-Newton

linear momentum

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

Lorentz-Einstein

scalar energies

$$E = \frac{mv^2}{c^2}$$

mass-energy equivalence

Energy per second is mass

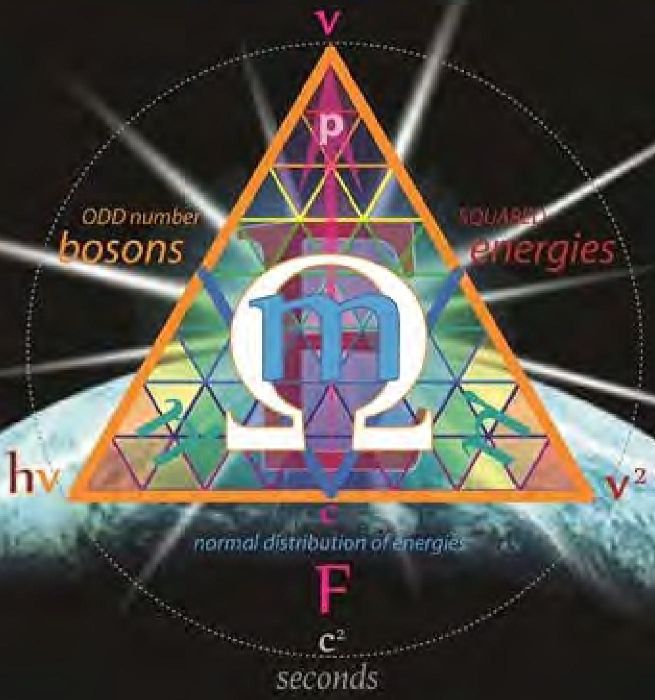
v
p
ODD number
bosons
SQUARED
energies
hv
v²
c
normal distribution of energies
F
c²
seconds

p

linear momentum

$$kg\,\frac{m}{s}$$

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

m

scalar mass

kg

Tetryonic Theory

The charged geometry of ElectroMagnetic mass-ENERGY
& Matter in motion

Charged geometries

Bosons

$$\frac{ODD\pi\left[\left[m\Omega v^2\right]\right]}{c^2}$$

Planck quanta mass velocity

Total Relativistic Energies

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta mass velocity

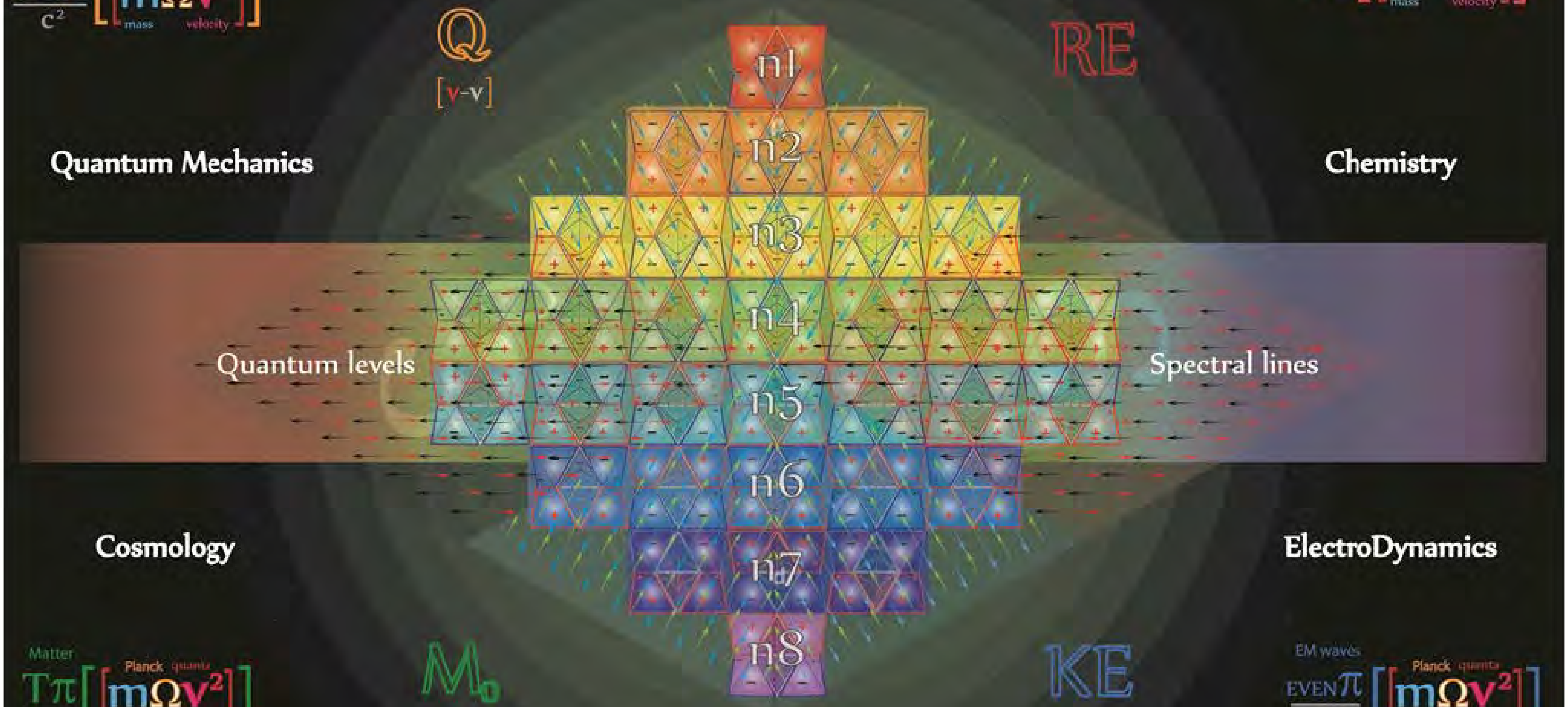
Q
[v-v]
RE
Quantum Mechanics
Chemistry
n1
n2
n3
n4
n5
n6
n7
n8
Quantum levels
Spectral lines
Cosmology
ElectroDynamics
M_0
KE

Matter

$$\frac{T\pi\left[\left[m\Omega v^2\right]\right]}{c^4}$$

Planck quanta mass velocity

EM waves

$$\frac{EVEN\pi\left[\left[m\Omega v^2\right]\right]}{c^2}$$

Planck quanta mass velocity

mass/Matter topologies

Unifying QM, QED, Chemistry and Cosmology
through charged equilateral Planck mass-energy momenta

Kinetic Energies of motion

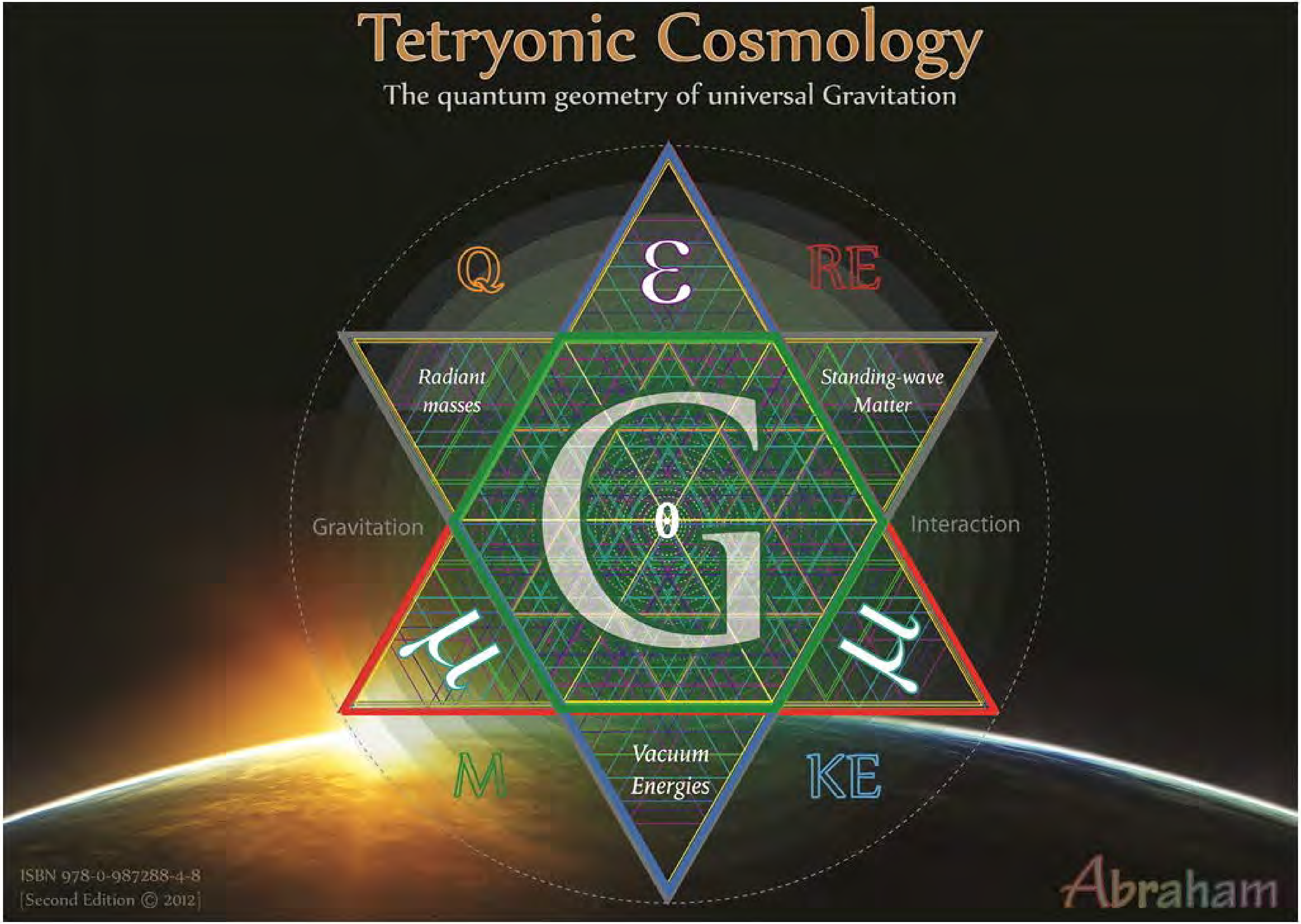

Tetryonic Cosmology

The quantum geometry of universal Gravitation

Q
ε
RE
Radiant masses
Standing-wave Matter
G
0
Gravitation
Interaction
μ
γ
M
Vacuum Energies
KE
Abraham

ISBN 978-0-987288-4-8
[Second Edition © 2012]

Tetryonics 80.24 - Quantum Cosmology

Copyright ABRAHAM [2008] - All rights reserved

TETRYONICS

The equilateral geometry underpinning the mathematics of Physics

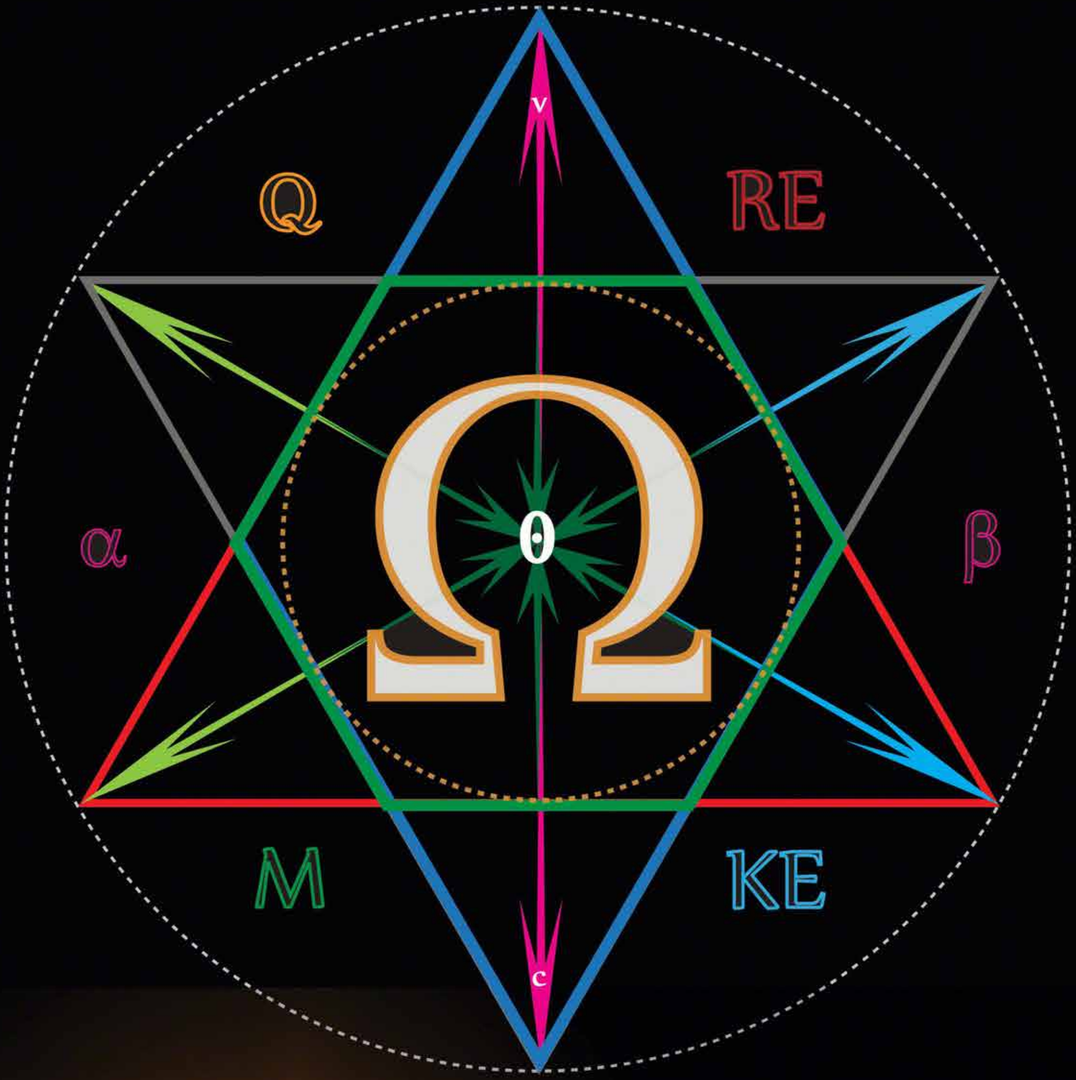

Foundational physics Mathematics

Abraham

ISBN 978-0-987288-5-5

[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Geometry and the Theory of Everything

Plato

(c.428-348 BC)

DIFFERENT TYPE OF GEOMETRIES

2D PLANE | **3D** SPHERE | **3D** SADDLE

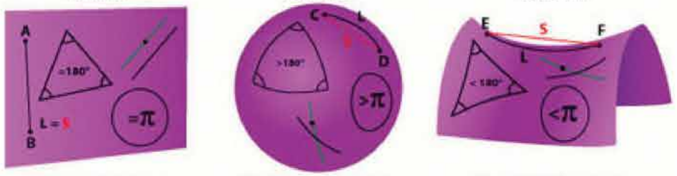

Zero Curvature *Euclidian geometry* | **Positive Curvature** *Elliptic geometry* | **Negative Curvature** *Hyperbolic geometry*

(studied by Omar Khayyam, Girolamo Saccheri, Bernhard Riemann, ...)

Euclid

(c.330-275 BC, fl. c.300 BC)

The Socratic tradition was not particularly congenial to mathematics, as may be gathered from Socrates' inability to convince himself that 1 plus 1 equals 2, but it seems that his student Plato gained an appreciation for mathematics after a series of conversations with his friend Archytas in 388 BC.

One of the things that most caught Plato's imagination was the existence and uniqueness of what are now called the five "Platonic solids".

It's uncertain who first described all five of these shapes - it may have been the early Pythagoreans - but some sources (including Euclid) indicate that Theaetetus (another friend of Plato's) wrote the first complete account of the five regular solids.

Presumably this formed the basis of the constructions of the Platonic solids that constitute the concluding Book XIII of Euclid's Elements.

In any case, Plato was mightily impressed by these five definite shapes that constitute the only perfectly symmetrical arrangements of a set of (non-planar) points in space, and late in life he expounded a complete "theory of everything", in the treatise called Timaeus, based explicitly on these five solids.

Interestingly, almost 2000 years later, Johannes Kepler was similarly fascinated by these five shapes, and developed his own cosmology from them

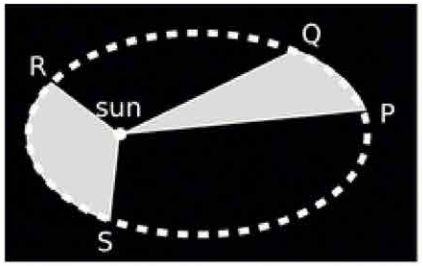

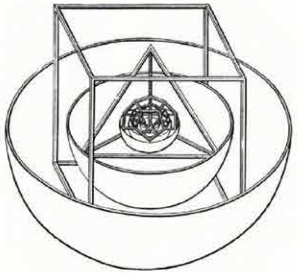

Tetractys

The Greek Tetractys is a
triangular figure consisting of ten points
arranged in four rows:

one, two, three, and four points in each row,
which is the geometrical representation of the
fourth triangular number.

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

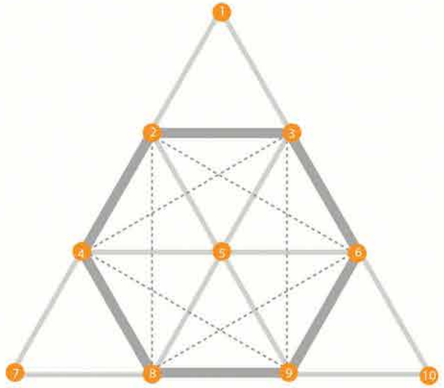

Sacred numbers

A

The Tetractys historically symbolized the four elements
[Earth, Air, Fire, and Water]
and the relationship between Humanity
and the cosmos created by GOD

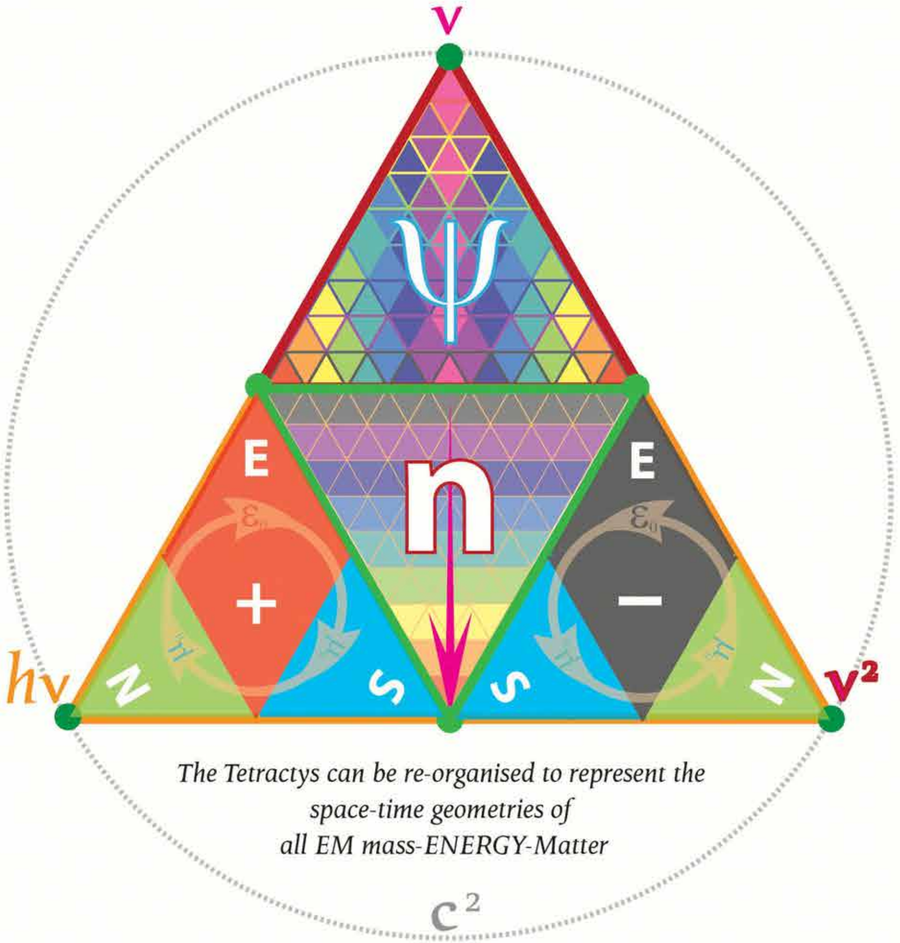

The Tetractys can be re-organised to represent the
space-time geometries of
all EM mass-ENERGY-Matter

The tetrad was the name given to
the number four -
in Pythagorean philosophy
there were four seasons
and four elements,
and the number was also associated
with planetary motions and music

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

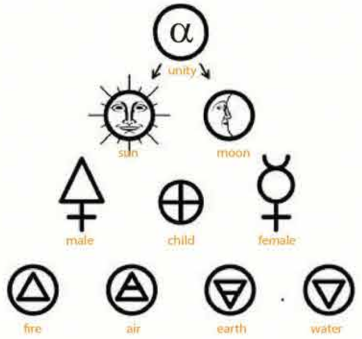

The Cosmos

Ω

The single triangle in the first row represents zero-dimensions (a point)
A vector direction in one-dimension can be represented as a line between any two points
The second row represents a Boson (two-dimensions in a plane defined by a rhombus of three triangles)
The whole figure folded represents three-dimensions (a tetrahedron defined by four apex points)
Photons of ElectroMagnetic mass-Energy quanta are represented by two opposing triangles

The Greek Zodiac

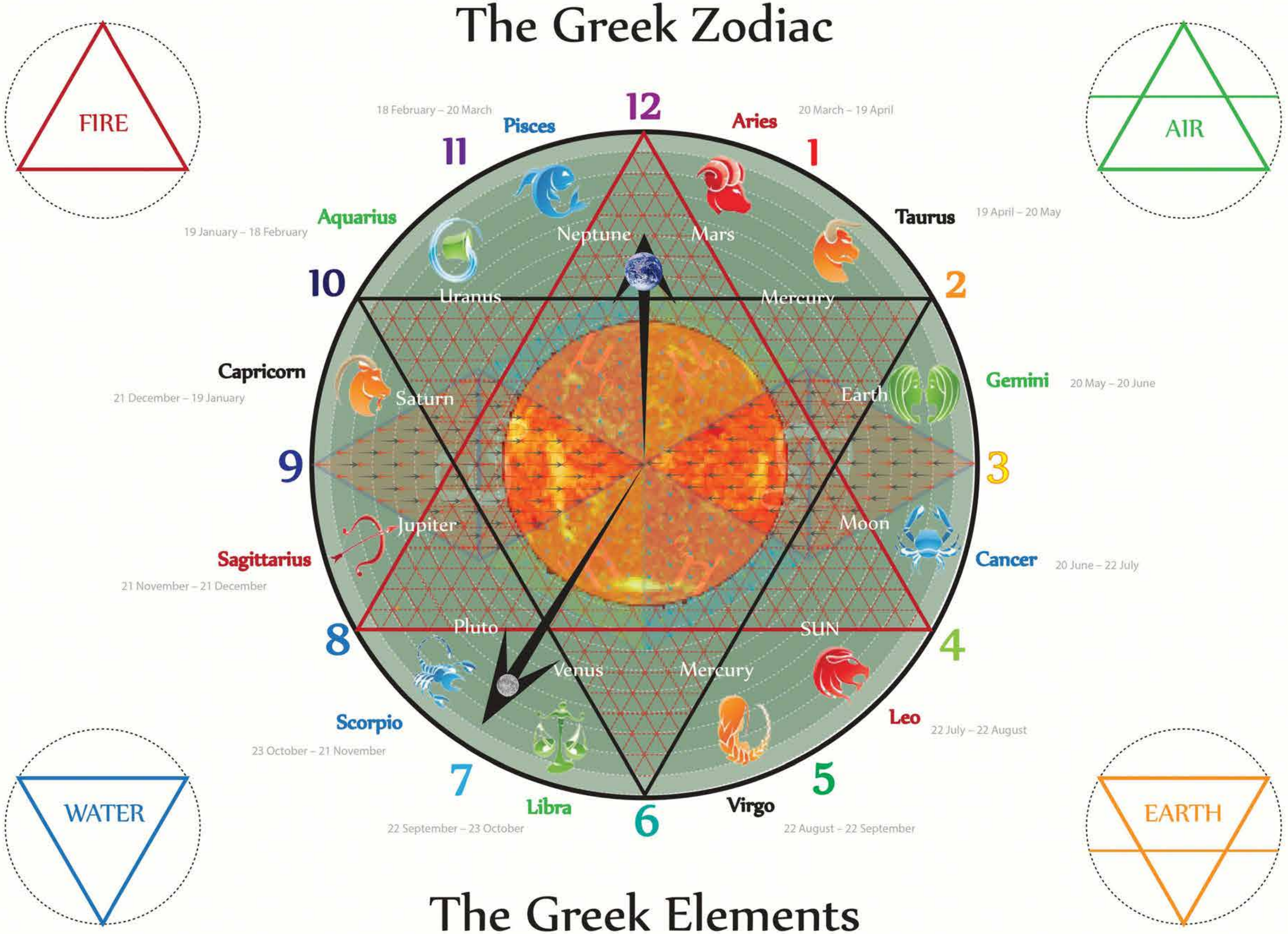
FIRE
AIR
12
Pisces
18 February – 20 March
11
Aries
20 March – 19 April
1
Taurus
19 April – 20 May
2
Aquarius
19 January – 18 February
10
Neptune
Mars
Uranus
Mercury
Capricorn
21 December – 19 January
Saturn
Earth
Gemini
20 May – 20 June
9
3
Jupiter
Moon
Sagittarius
21 November – 21 December
Cancer
20 June – 22 July
8
Pluto
SUN
4
Venus
Mercury
Scorpio
23 October – 21 November
Leo
22 July – 22 August
7
Libra
22 September – 23 October
6
Virgo
22 August – 22 September
5
WATER
EARTH

The Greek Elements

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Triangles

An equilateral triangle is a triangle in which all three sides are equal

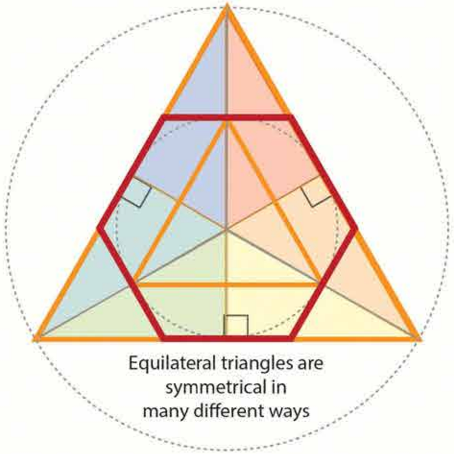

Equilateral triangles are
symmetrical in
many different ways

It is unique in that it is the only polygon
that can be tiled [or divided] and produce
only identical geometries and squares numbers

The equilateral triangle
is eminently suited for
the construction of fractals

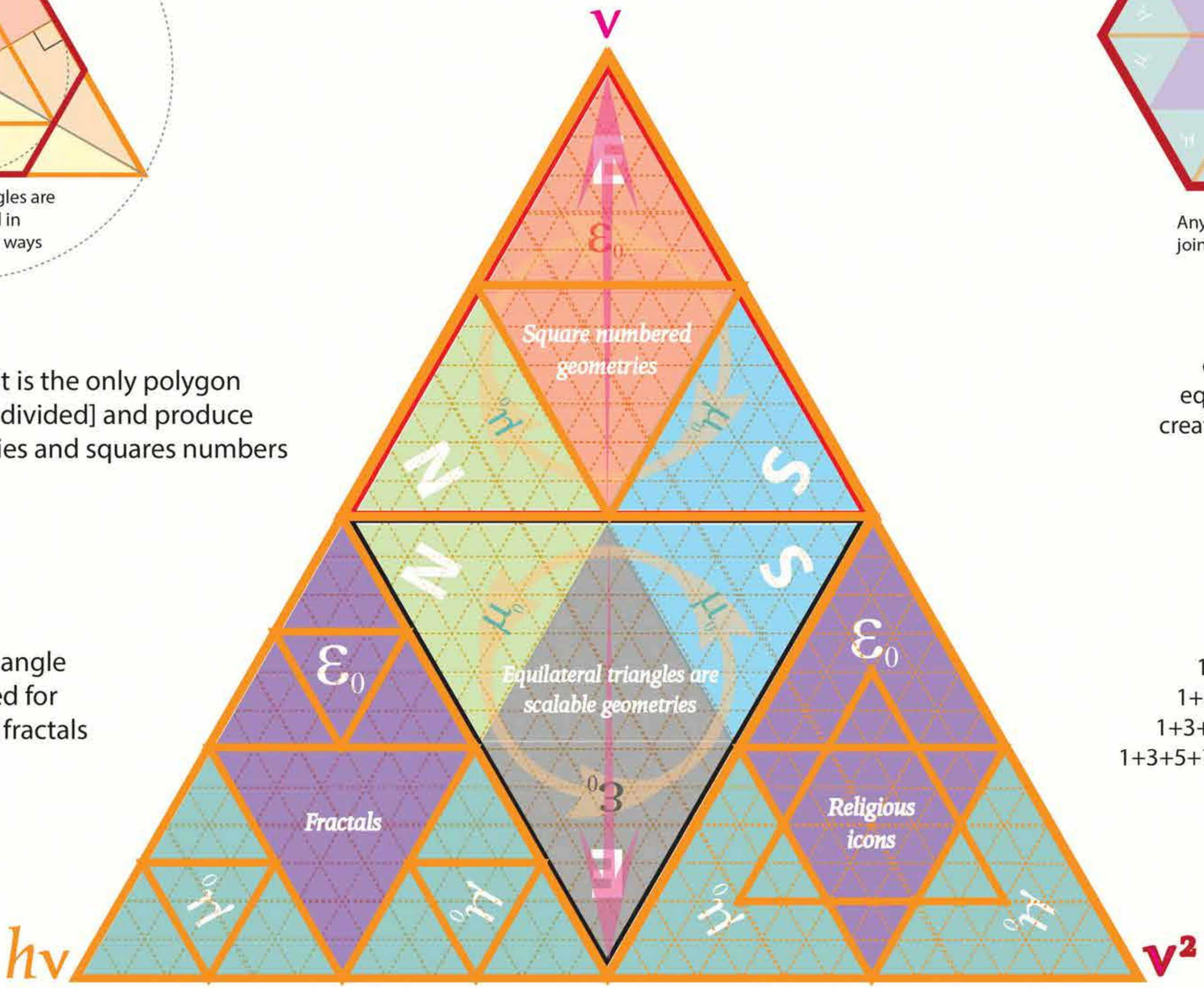
v
Square numbered geometries
N
S
ε_0
μ_0
Equilateral triangles are scalable geometries
Fractals
Religious icons
hv
v²

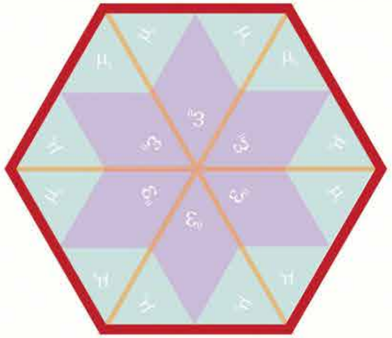
ε_0
μ_0

Any six equilateral triangles
joined can make a hexagon.

The
tesselation of
odd numbered
equilateral triangles
creates square numbers

That is,

1
1+3 = 4
1+3+5 = 9
1+3+5+7 = 16
1+3+5+7+9 = 25
1+3+5+7+9+11 = 36
1+3+5+7+9+11+13 = 49
1+3+5+7+7+9+11+13+15 = 64
etc

An equilateral triangle is simply a specific case of a regular polygon with 3 sides

The Pythagorean Theorem

Though attributed to Pythagoras, it is not certain that he was the first person to prove it.

The first clear proof came from Euclid, and it is possible the concept was known
1000 years before Pythoragas by the Babylonians

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

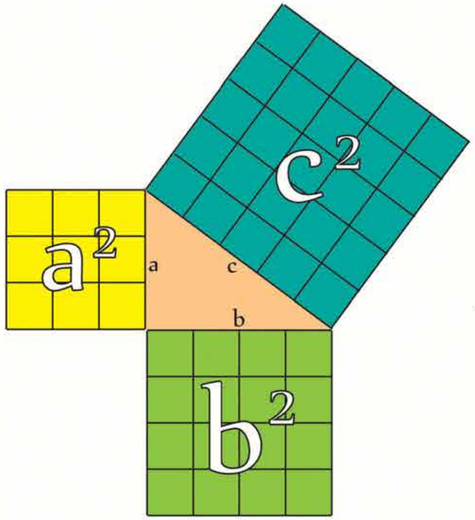

In any right triangle,
the area of the square whose side is the hypotenuse
(the side opposite the right angle) is equal to
the sum of the areas of the squares
whose sides are the two legs
(the two sides that meet at a right angle)

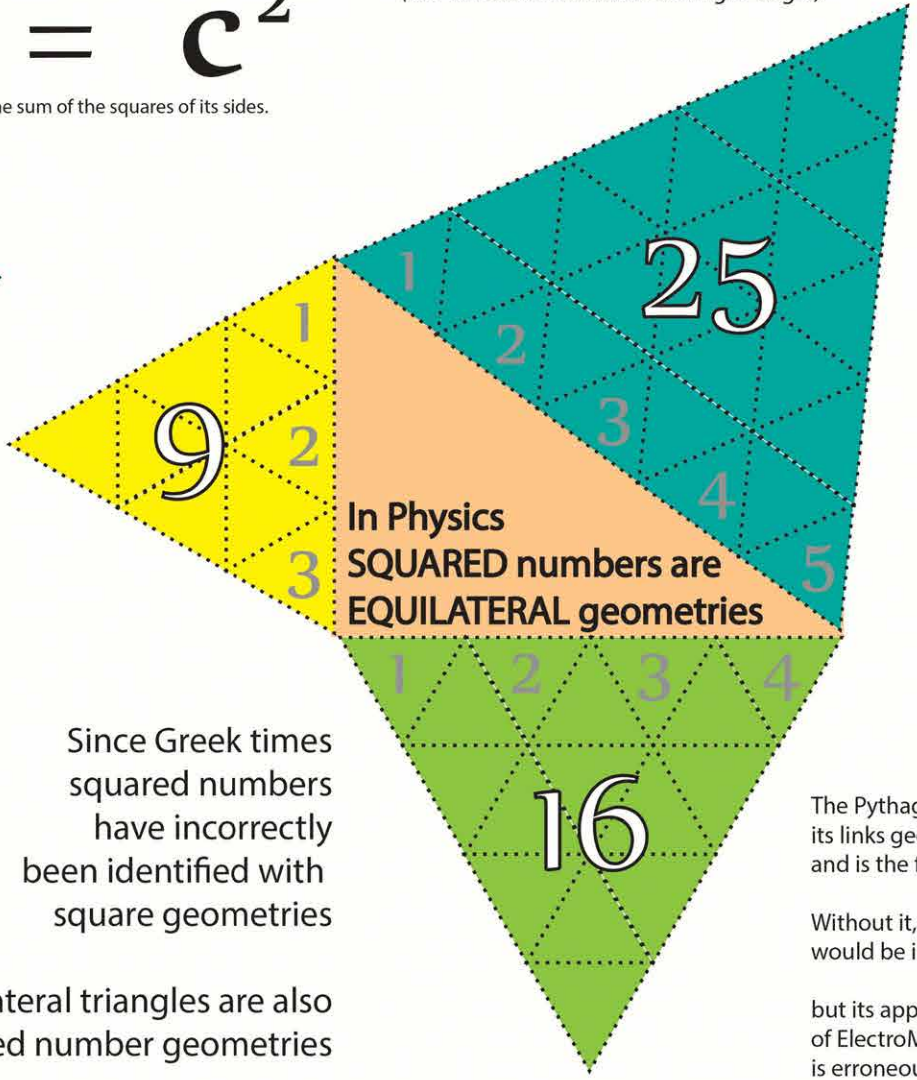

In Physics
SQUARED numbers are
EQUILATERAL geometries

Pythagoras of Samos

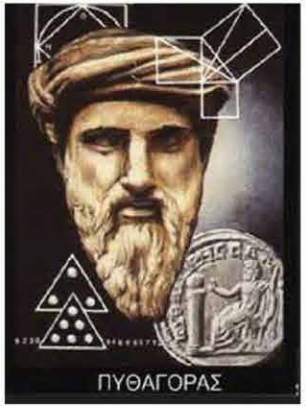

about (570 – 495 BC)

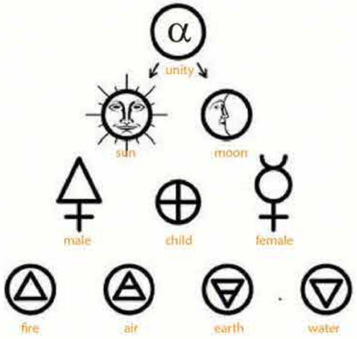

Pythagorean Tetractys

Since Greek times
squared numbers
have incorrectly
been identified with
square geometries

Equilateral triangles are also
squared number geometries

The Pythagorean equation is at the core of much of geometry,
its links geometry with algebra,
and is the foundation of trigonometry.

Without it, accurate surveying, mapmaking, and navigation
would be impossible,

but its application to the energy-momenta geometries
of ElectroMagnetic fields and Matter in motion in Physics
is erroneous and must be corrected for science to advance

Copyright ABRAHAM [2008] - All rights reserved

Hexagons

A regular hexagon can be subdivided into six equilateral triangles

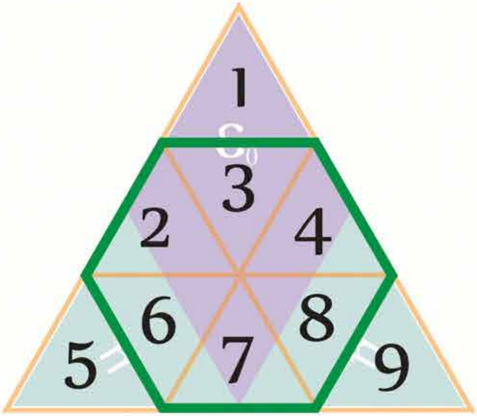

Energy geometries

Atomic nuclei geometries

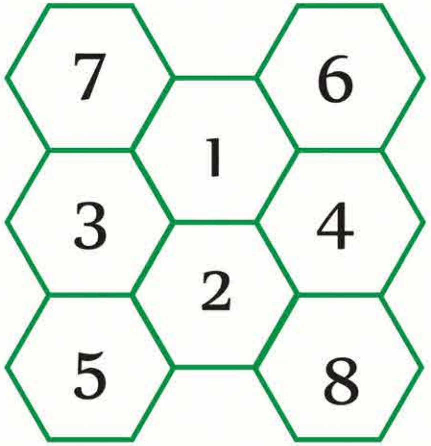

Hexagons can be tiled or tessellated in a regular pattern on a flat two-dimensional plane

Hexagons are the only regular polygon that can be subdivided into another regular polygon

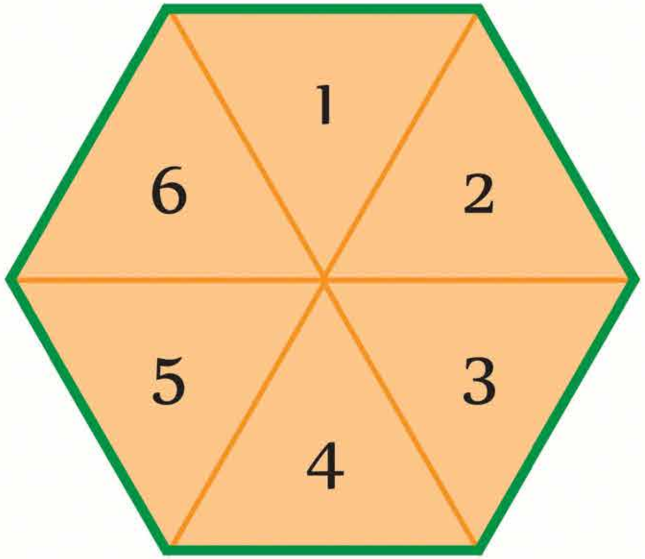

Hexagons are the unique regular polygon such that the distance between the center and each vertex is equal to the length of each side

Six is a highly composite number, the second-smallest composite number, and the first perfect number.

That is, 1*2*3=1+2+3=6

An interesting relationship between circular and hexagonal geometry is that hexagonal patterns often appear spontaneously when natural forces are trying to approximate circles

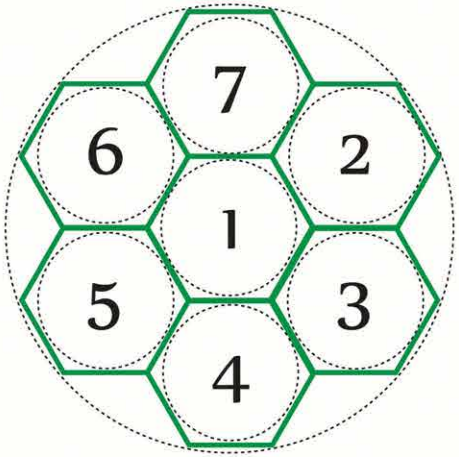

Hexagonal tessellation is topologically identical to the close packing of circles on a plane

Copyright ABRAHAM [2008] - All rights reserved

Platonic Solids

The Pythagoreans knew that there were only five regular convex solids, the tetrahedon, cube, octahedron, icosohedron and dodecahedron and each one could be accurately circumscribed by a sphere.

FIRE

4 triangles meet to form a Tetrahedron

EARTH

6 squares meet to form a Cube

AIR

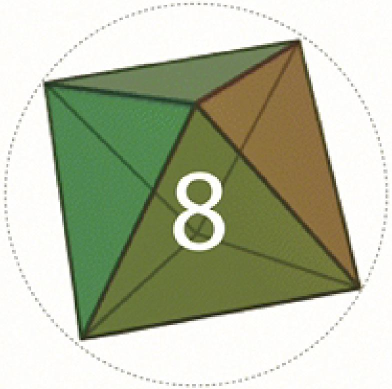

8 triangles meet to form a Octahedron

tetryons **4**

regular deltahedrons

leptons **12**

F - E + V = 2

faces edges vertexs

8 ***quarks***

regular deltahedrons

20 ***Baryons***

12 pentagons meet to form a Dodecahedron

THE HEAVENS

Plato

(c.428-348 BC)

The philosopher Plato concluded that they must be the fundamental building blocks – the atoms – of nature, and assigned to them what he believed to be the essential elements of the universe.

20 triangles meet to form a Icosahedron

WATER

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic Solids

Despite their unique topologies Tetryonic solids are not unlike Platonic solids save that their toplogies are comprised entirely from complex hitherto undescribed **4nπ equilateral Planck mass-energy momenta geometries** that also match the Euler numbers of Platonic solids

+ tetryon

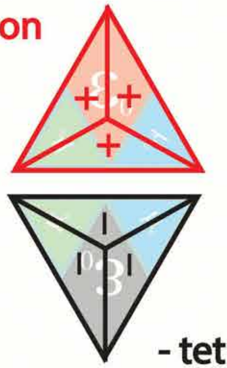

4 faces
6 edges
4 vertices

2

- tetryon

up quark

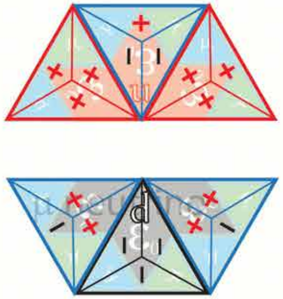

2

8 faces
12 edges
6 vertices

down quark

mass geometries

4π *tetryons* **4π**

regular deltahedrons

12π *leptons* **12π**

Matter topologies

Matter topologies

8π *quarks* **12π**

regular deltahedrons

20π *Baryons* **36π**

mass geometries

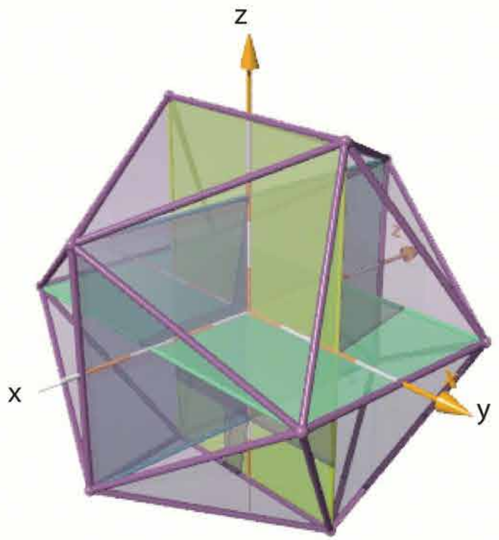

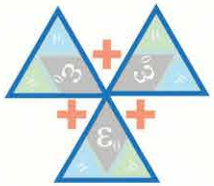

neutrino

12 faces
18 edges
8 vertices

2

electron

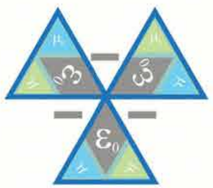

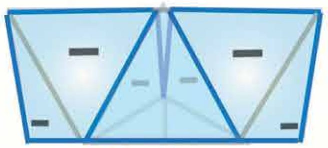

Neutron

2

20 faces
30 edges
12 vertices

Proton

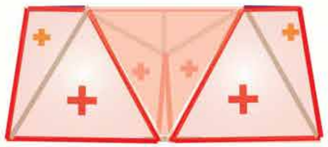

Their equilateral topologies are best described as regular topologic-deltahedrons:

tetra-delta-hedrals	**tetryons**	4π external charge fascia
octa-delta-hedrals	**quarks**	8π external charge fascia
dodeca-delta-hedrals	**leptons**	12π external charge fascia
icoso-delta-hedrals	**Baryons**	20π external charge fascia

note:
Charged mass-energy fascia geometries and edges become "hidden" upon the meshing of delta-hedra to form Matter topologies

Copyright ABRAHAM [2008] - All rights reserved

Euclidean geometry

Euclid

(c.330-275 BC, fl. c.300 BC)

Arguably the most influential Mathematics book ever written is Euclid's 'The Elements'

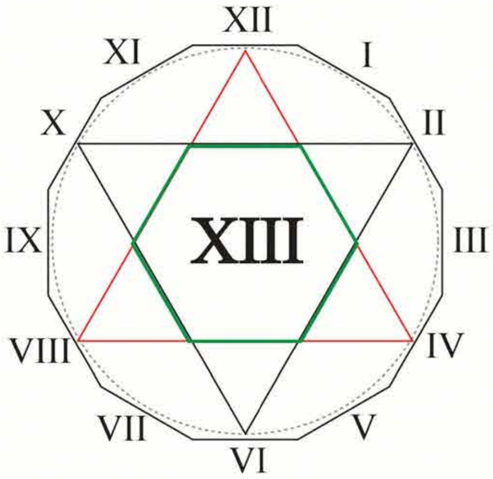

In all, it contains 465 theorems and proofs, described in a clear, logical and elegant style, and using only a compass and a straight edge.

The Elements - Book 1 - Definition 20

Of the trilateral figures, an equilateral triangle is that which has its three EQUAL sides

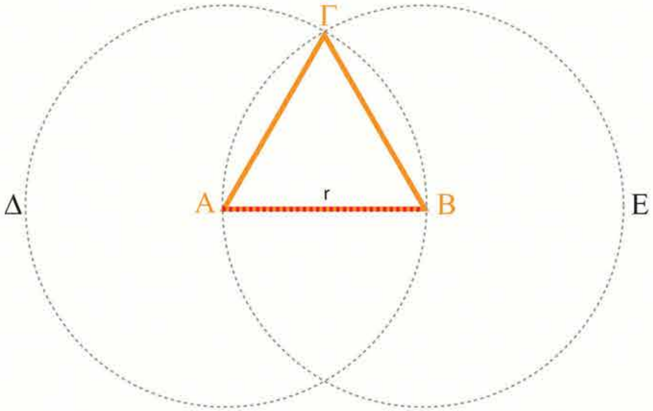

Euclid's Elements - Book 1 - Proposition 1
Method of constructing an Equialteral triangle

Euclid's five general axioms were:

Things which are equal to the same thing are equal to each other.

If equals are added to equals, the wholes (sums) are equal.

If equals are subtracted from equals, the remainders (differences) are equal.

Things that coincide with one another are equal to one another.

The whole is greater than the part

Copyright ABRAHAM [2008] - All rights reserved

Archimedes

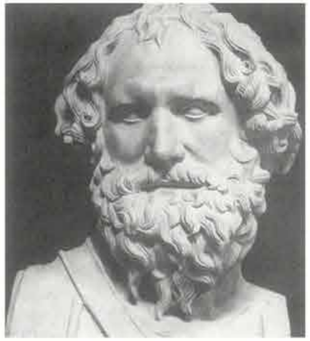

c. (287 BC – c. 212 BC)

As its definition relates to the circle,
π is found in many formulae in trigonometry and geometry,
especially those concerning circles, ellipses, or spheres.

It is also found in formulae from other branches of science,
such as cosmology, number theory, statistics, fractals,
thermodynamics, mechanics, and electromagnetism

Incorrect identification of Pi [c/d] as opposed to Pi radians
in Physics has led to the inappropriate association
of spherical particles to the physical sciences
whereas equilateral triangles & tetrahedra
form its true geometry

The number π is a mathematical constant that is
the ratio of a circle's circumference to its diameter.

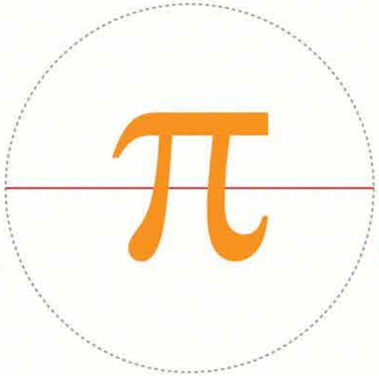

Pi

π is an irrational number
its decimal representation
never ends and never repeats.

$\pi = c/d$

The ratio C/d is constant,
regardless of the circle's size

3.141592654.......

Proof of the fact that C=2πr and how Archimedes proved it

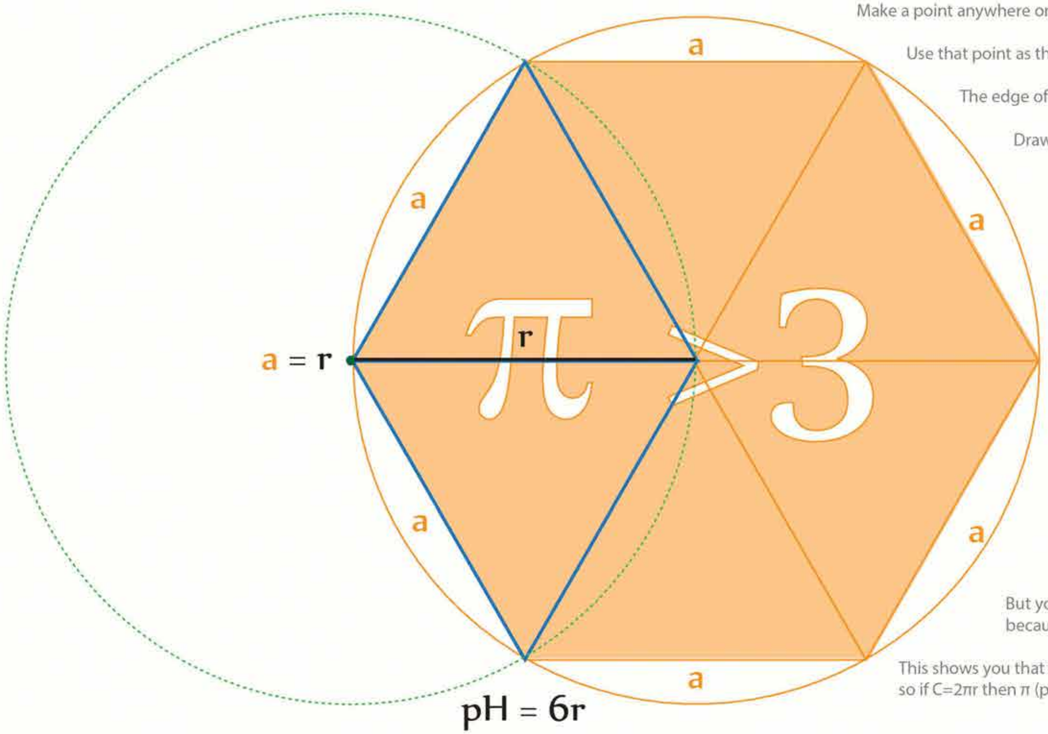

Draw any circle.

Make a point anywhere on the circumference of the [green] circle.

Use that point as the center of a second [blue] circle with the same radius as the green circle.

The edge of the blue circle should touch the center of the green circle.

Draw the line segment connecting the centers of the two circles.

That forms the radius of both of the circles.

Now draw the line connecting the center of the blue circle to where it crosses the green circle
on both sides, and complete the triangles.

You should have two equilateral triangles whose sides are equal to the radius of the green circle

Now extend all of the radius lines so they become diameter lines,
all the way across the circle, and finish drawing all of the triangles to connect them.

You've got six equilateral triangles now, that make an orange hexagon.

So the perimeter of your hexagon is the same as six times the radius of your circle.

But your circumference is a little bigger than the perimeter of your hexagon,
because the shortest distance between two points is always a straight line.

This shows you that the circumference of the blue circle has to be more than 6r,
so if C=2πr then π (pi) has to be a little bigger than 3, which it is.

The more sides we draw on our polygon, the closer we will get to the real value of pi (3.14159 etc.).
Using a polygon with 96 sides, Archimedes was able to calculate that π was a little bigger than 3.1408

Copyright ABRAHAM [2008] - All rights reserved

The Golden Ratio

Two quantities are in the golden ratio (φ) if the ratio of the sum of the quantities to the larger quantity is equal to the ratio of the larger quantity to the smaller one.

The figure to the right illustrates the geometric relationship

[1.61803]

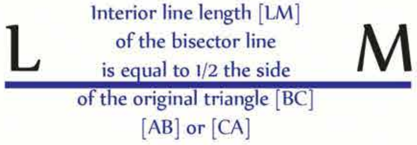

Interior line length [LM] of the bisector line is equal to 1/2 the side of the original triangle [BC] [AB] or [CA]

Φ

Golden ratio phi

$$\hbar = \frac{h}{2\pi}$$

Tetryonic geometry reveals the maximum E-field amplitude of the reduced Planck constant to be an example of the Golden ratio in physics

velocity vectors

The height of the triangle [ALM] produced by the bisecting line is 1/2 the height of the height of [ABC]

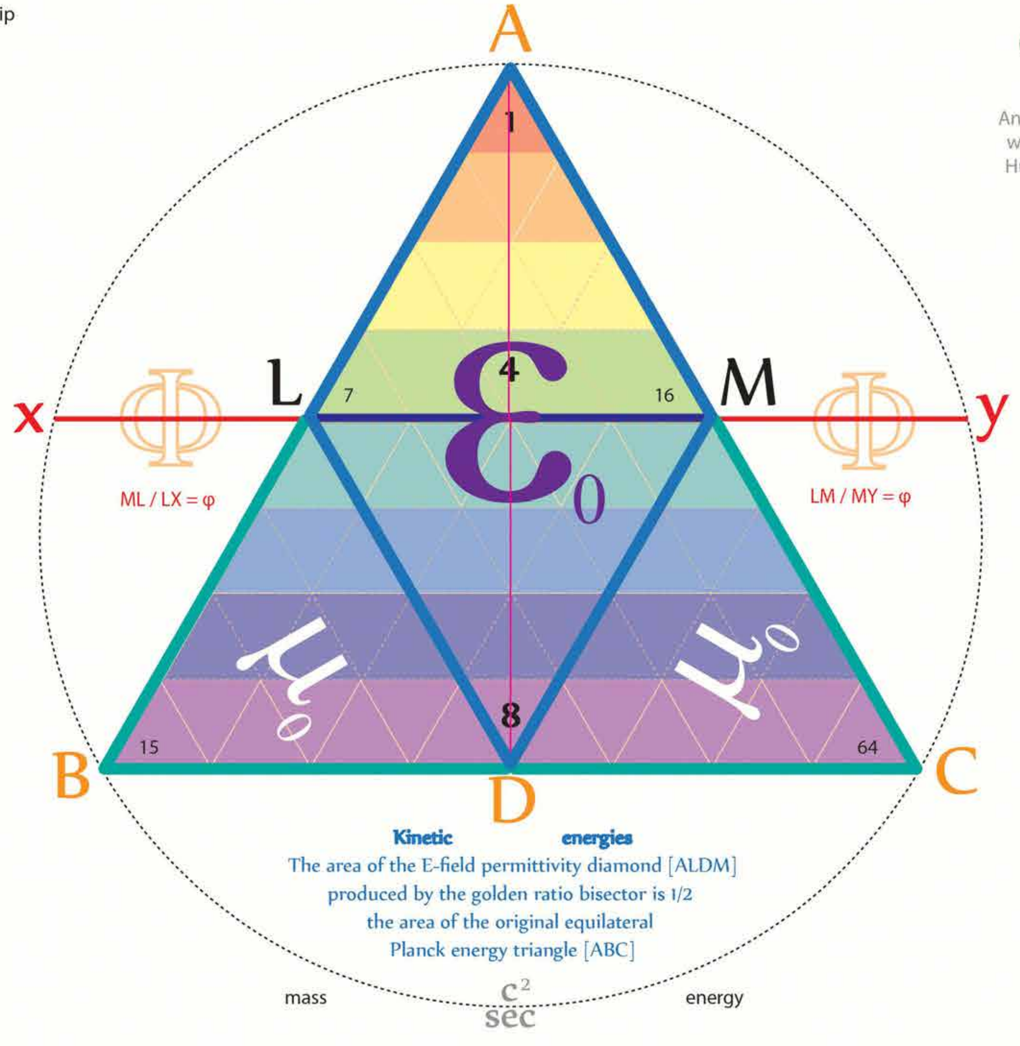

Kinetic **energies**

The area of the E-field permittivity diamond [ALDM] produced by the golden ratio bisector is 1/2 the area of the original equilateral Planck energy triangle [ABC]

Planck's constant [h]

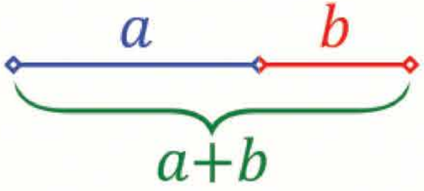

$a+b$ is to a as a is to b

An intriguing showing of φ in an equilateral triangle was observed by George Odom, a resident of the Hudson River Psychiatric Center, in the early 1980s

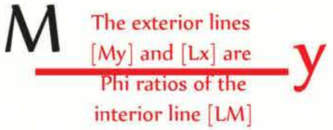

The exterior lines [My] and [Lx] are Phi ratios of the interior line [LM]

Let L and M be the midpoints of the sides AB and AC of an equilateral triangle ABC

By measurement and the Intersecting Chords Theorem

MX·MY = AM·MC.

which is of the form

(a + b)·b = a·a.

Denoting a/b = x, we see that

$1 + x = x^2$,

George Odom

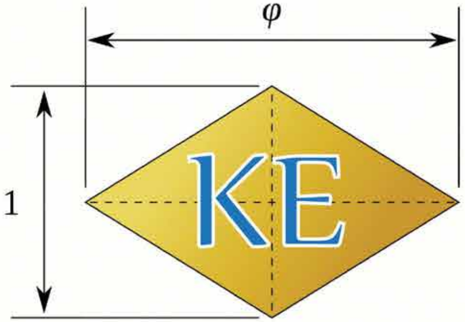

The Golden Rhombus

Applying the golden ratio (φ) to quantum scale electrodynamic geometry we can quickly determine that the linear momentum and magnetic moment vectors of photons & EM waves can also be expressed as a golden ratio

vector velocities
vs
magnetic vector

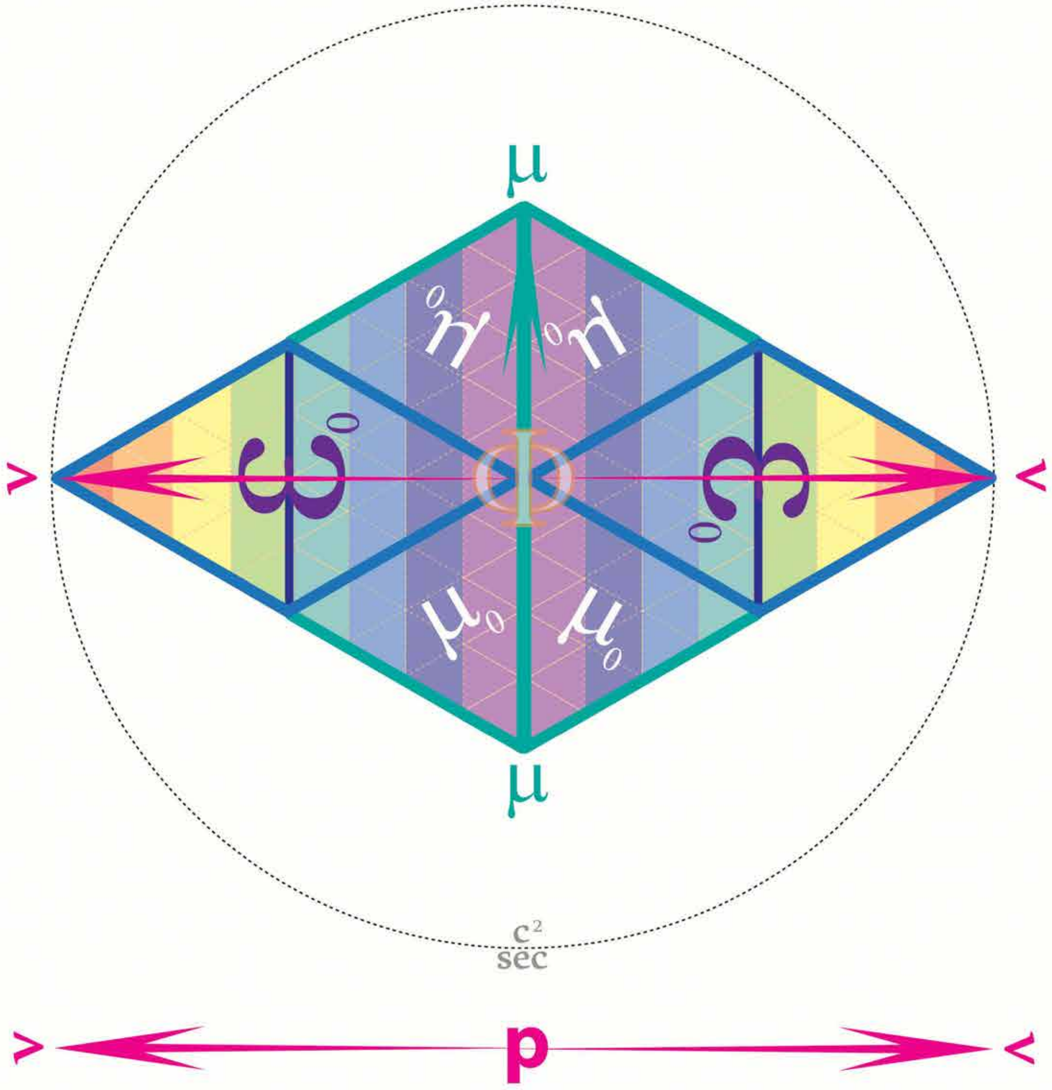

Φ

Golden ratio phi

[1.61803]

magnetic Force
vs
electric Force

photons
of
EM energy

Copyright ABRAHAM [2008] - All rights reserved

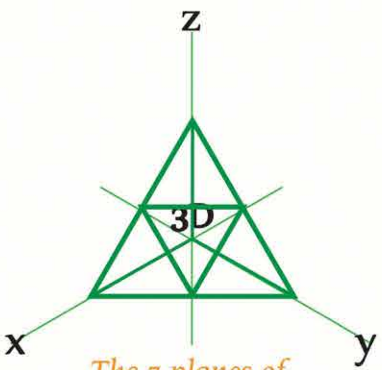

The 3 planes of Cartesian co-ordinates

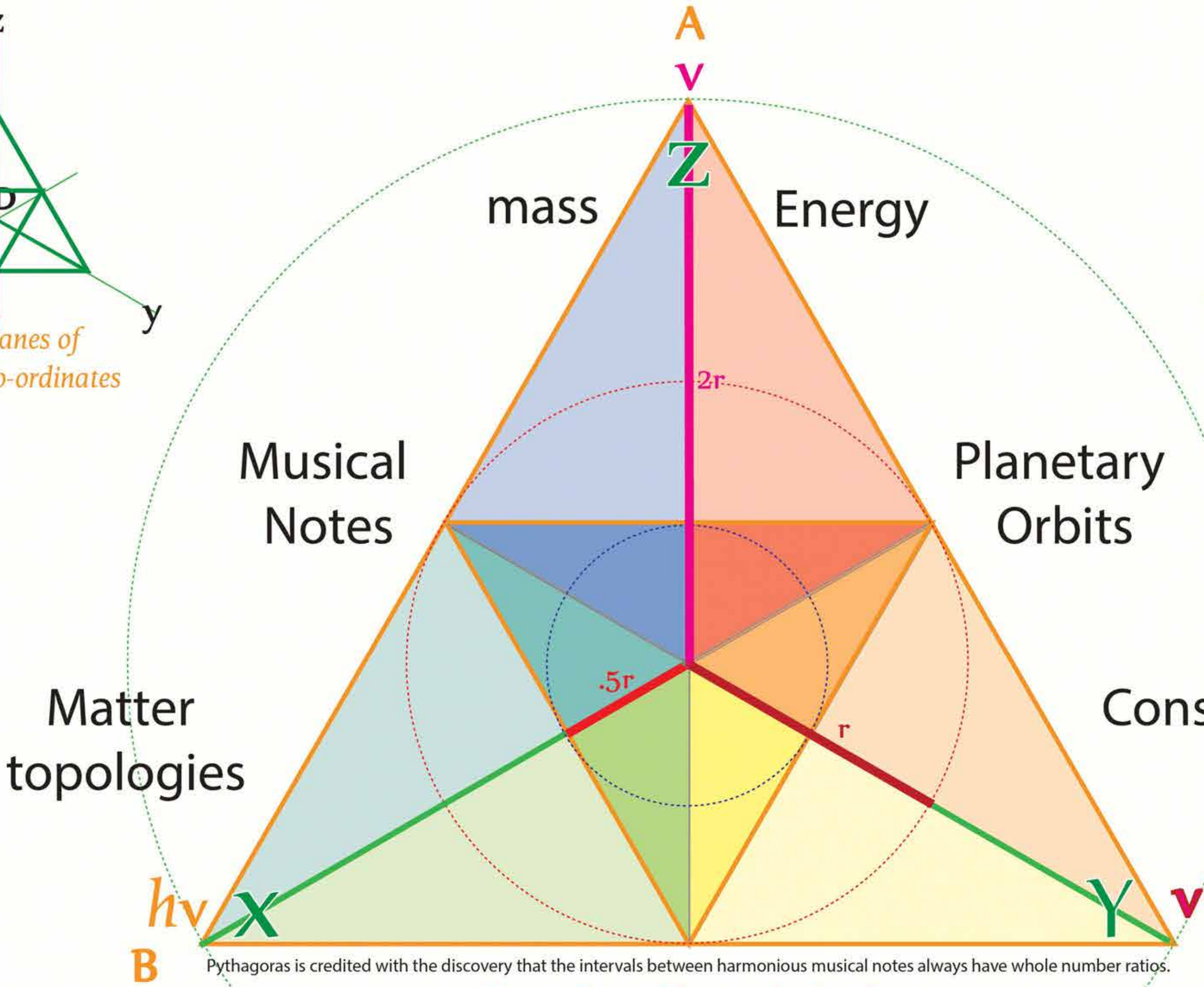
A
v
Z
mass
Energy
2r
Musical Notes
Planetary Orbits
.5r
r
Matter topologies
Conservation Laws
hv
X
Y
v^2
B
C

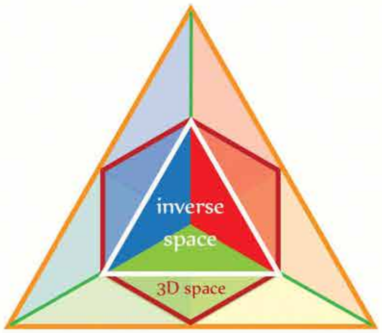
inverse space
3D space

An equilateral [square] triangle divided into 12 equal semitones

Pythagoras is credited with the discovery that the intervals between harmonious musical notes always have whole number ratios.

An equilateral [square] triangle
divided into 6 equal semitones
and 6 equal quarter tones

With an Equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles. Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another. Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal. This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Fifths

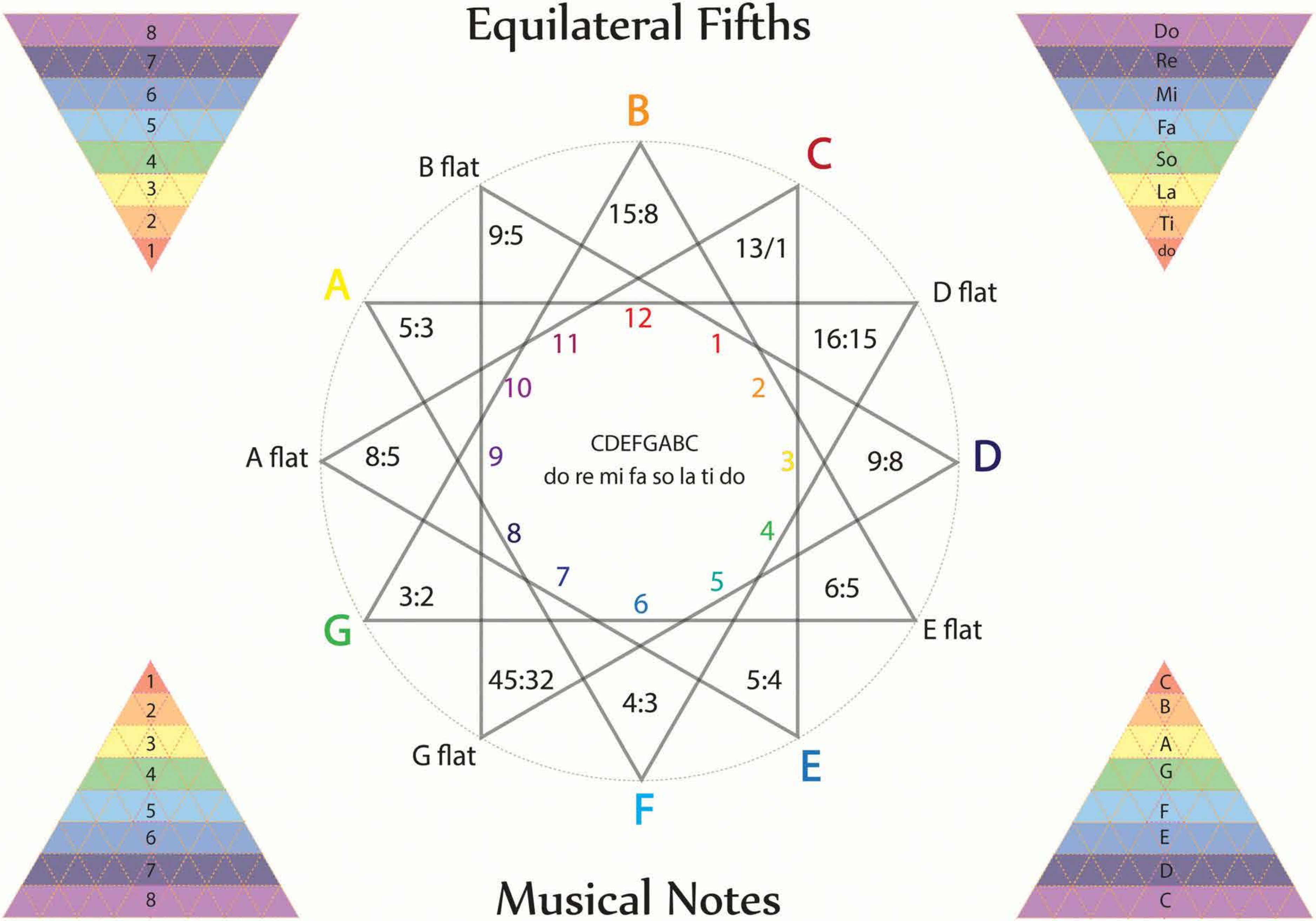
8
7
6
5
4
3
2
1
Do
Re
Mi
Fa
So
La
Ti
do
B
C
B flat
D flat
A
D
A flat
G
E flat
G flat
E
F
15:8
9:5
13/1
5:3
16:15
8:5
9:8
3:2
6:5
45:32
4:3
5:4
12
1
2
3
4
5
6
7
8
9
10
11
CDEFGABC
do re mi fa so la ti do
1
2
3
4
5
6
7
8
C
B
A
G
F
E
D
C

Musical Notes

2D radiated mass-energies

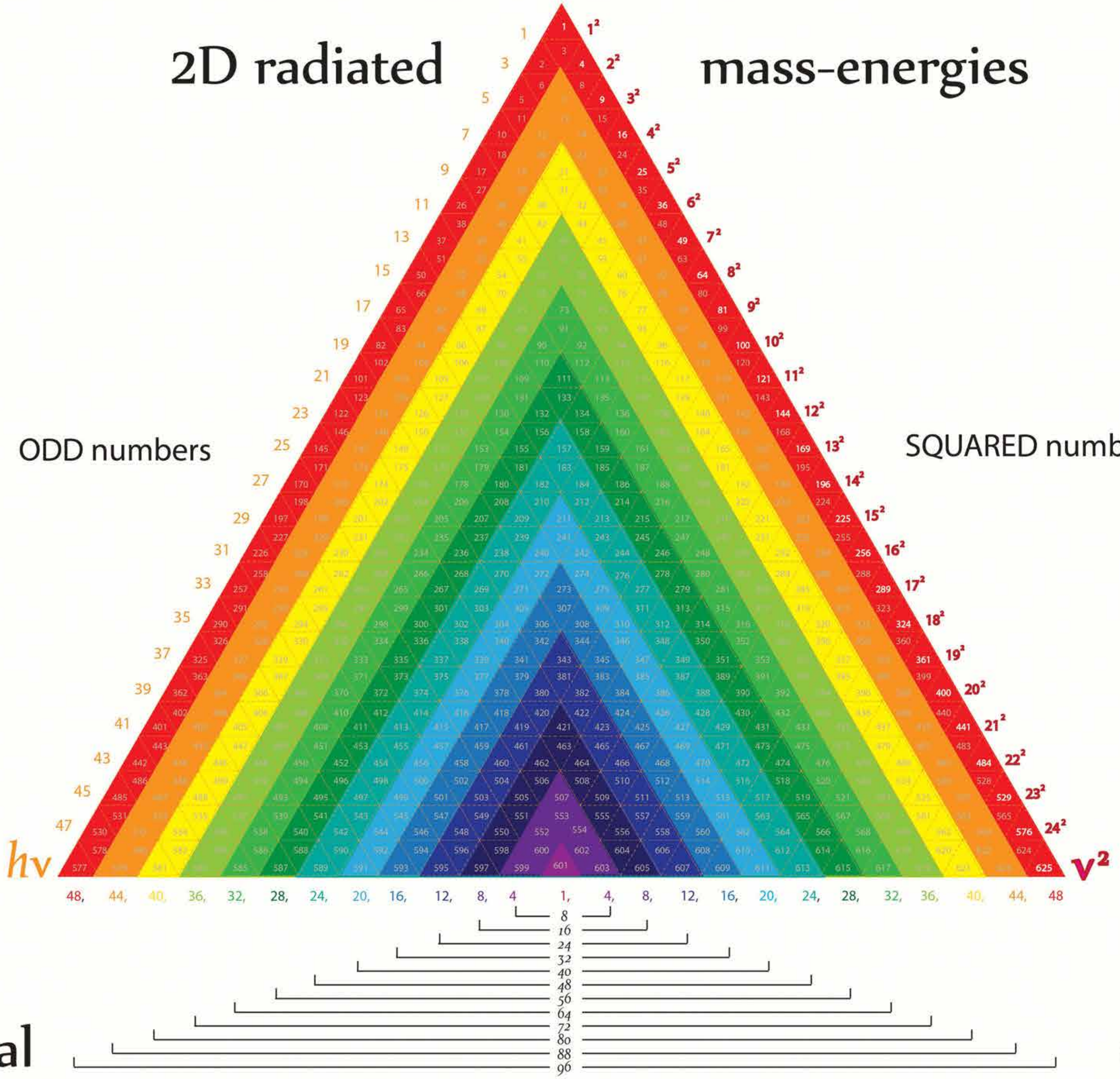
ODD numbers
SQUARED numbers
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
hv
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
9^2
10^2
11^2
12^2
13^2
14^2
15^2
16^2
17^2
18^2
19^2
20^2
21^2
22^2
23^2
24^2
v^2
48, 44, 40, 36, 32, 28, 24, 20, 16, 12, 8, 4
1,
4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48
8
16
24
32
40
48
56
64
72
80
88
96

Equilateral Octaves

Ternary Diagrams

Viviani's Theorem implies that lines parallel to the sides of an equilateral triangle
provide (homogeneous/barycentric/areal/trilinear) coordinates for ternary diagrams
for representing three quantities A,B,C whose sum is a constant
(which can be normalized to unity).

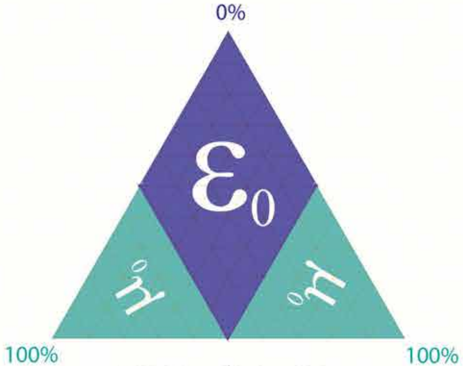

Ternary diagrams
should NOT be used
to model EM field
strengths

In a ternary plot, of EM energy
the Electric field [E] and the
Magnetic dipole [N-S]
must renormalise
to 1

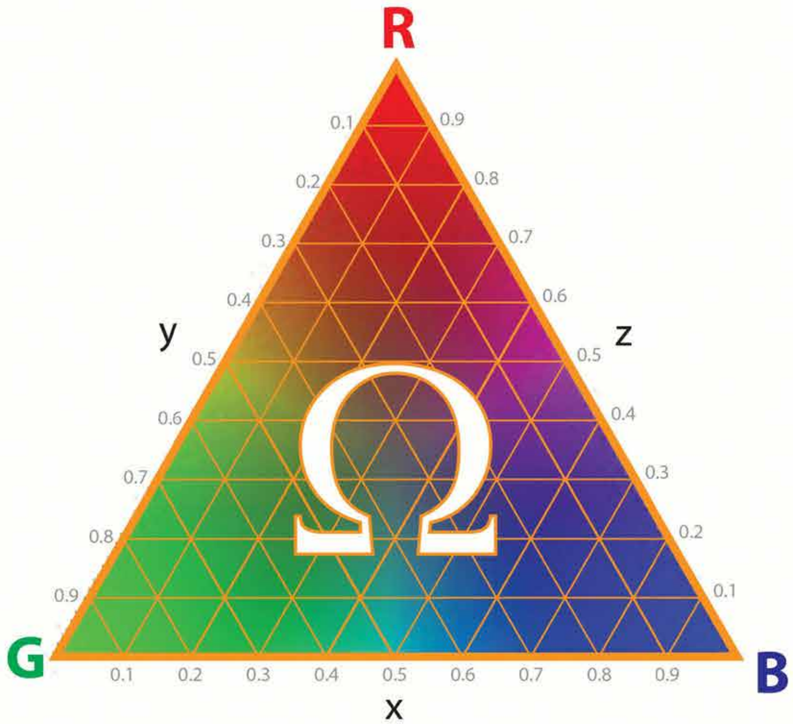

A ternary diagram is simply a triangular coordinate system in which the 3 edges correspond to the axes.

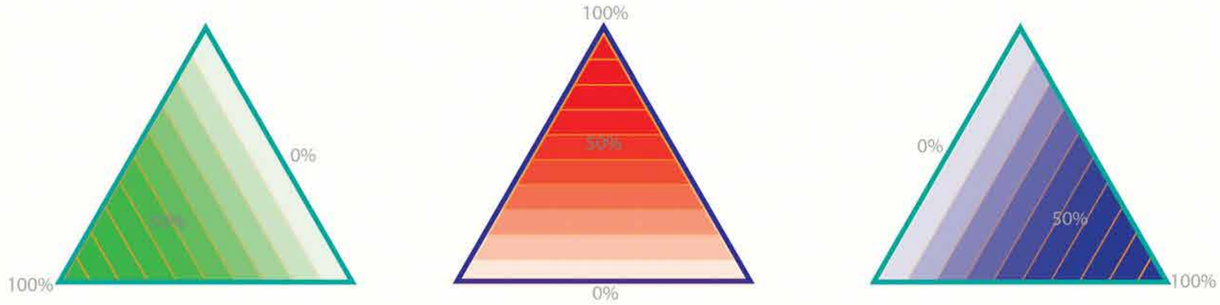

Trilinear charts are commonly used for finding the result of mixing three components (such as gases, chemical compounds, soil, color, etc.) that add to 100% of a quantity.

Whilst the Pythagorean Theorem boasts a slightly greater economy of terms than the Eutrigon Theorem (Wayne Roberts 2003), the latter contains an important area not included in the former:

the area enclosed or swept out by the three points of the triangle in question

$$ab = a^2 + b^2 - c^2$$

the area of any eutrigon is equal to the combined areas of the equilateral triangles on legs 'a' and 'b', minus the area of the equilateral triangle on its hypotenuse 'c'.

The Pythagorean theorem

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

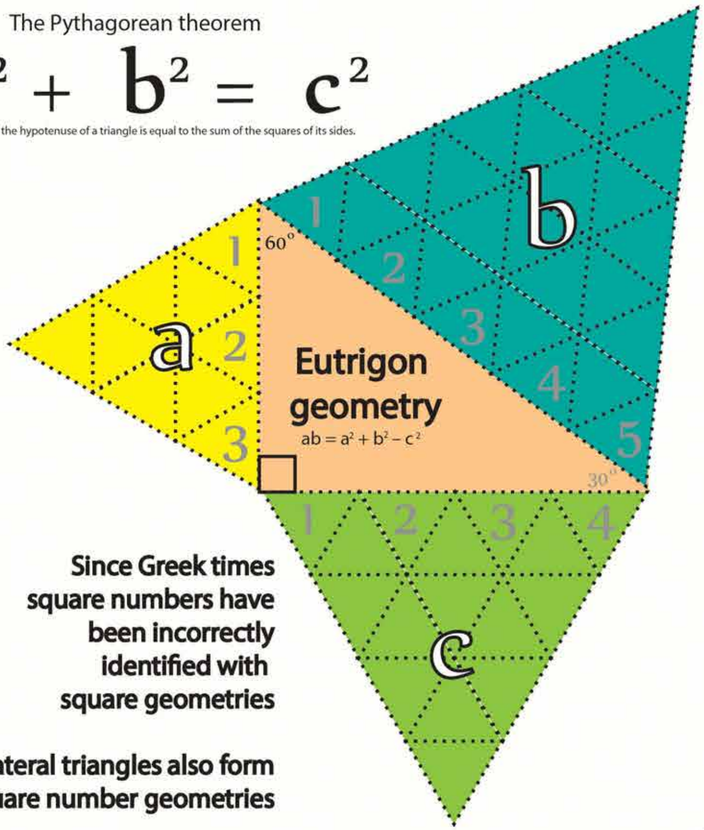

Since Greek times square numbers have been incorrectly identified with square geometries

Equilateral triangles also form square number geometries

Eutrigons

are an important new class of triangle (mathematically defined by Wayne Roberts), as the analogue of the right-triangle in orthogonal (Cartesian) coordinate geometry

Kepler's Second Law of planetary motion

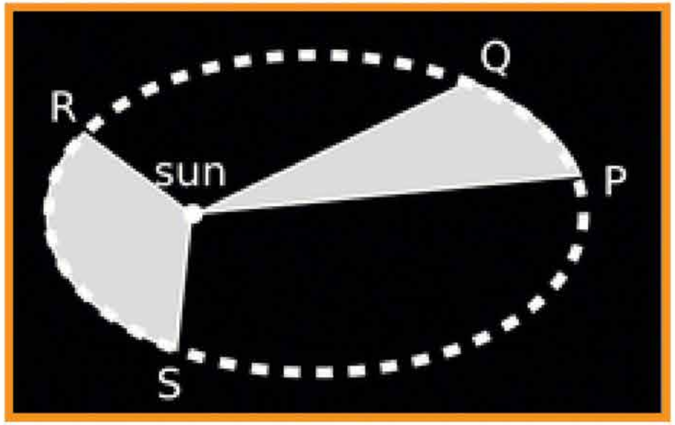

The orbit of every planet is an ellipse with the Sun at one of the two foci.

A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

The algebraic form of the Eutrigon Theorem, (like the algebraic form of Pythagoras' Theorem), is proven to be special case of the Cosine Rule...

Tetryonic theory reveals the equilateral [square] energy geometry
that reveals the 'harmonics at play' in physical laws
such as the second law of Kepler, and in many other phenomena in
physics, chemistry, cosmology, biochemistry and number theory
thus providing the foundation for the mathematics of quantum mechanics

Copyright ABRAHAM [2008] - All rights reserved

Number theory

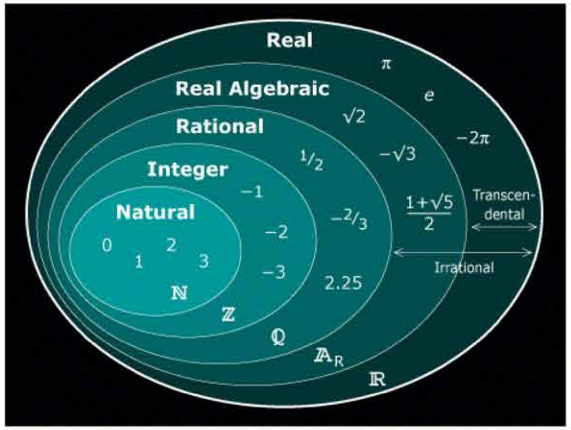
Real
π
e
Real Algebraic
√2
−2π
Rational
1/2
−√3
Integer
−1
1+√5 / 2
Transcen-dental
Natural
−2
−2/3
0
1
2
3
−3
2.25
Irrational
ℕ
ℤ
ℚ
𝔸ℝ
ℝ

The Pythagoreans also established the foundations of number theory, with their investigations of triangular, square and also perfect numbers (numbers that are the sum of their divisors).

They discovered several new properties of square numbers, such as that the square of a number n is equal to the sum of the first n odd numbers (e.g. 1 + 3 + 5 + 7 = 16).

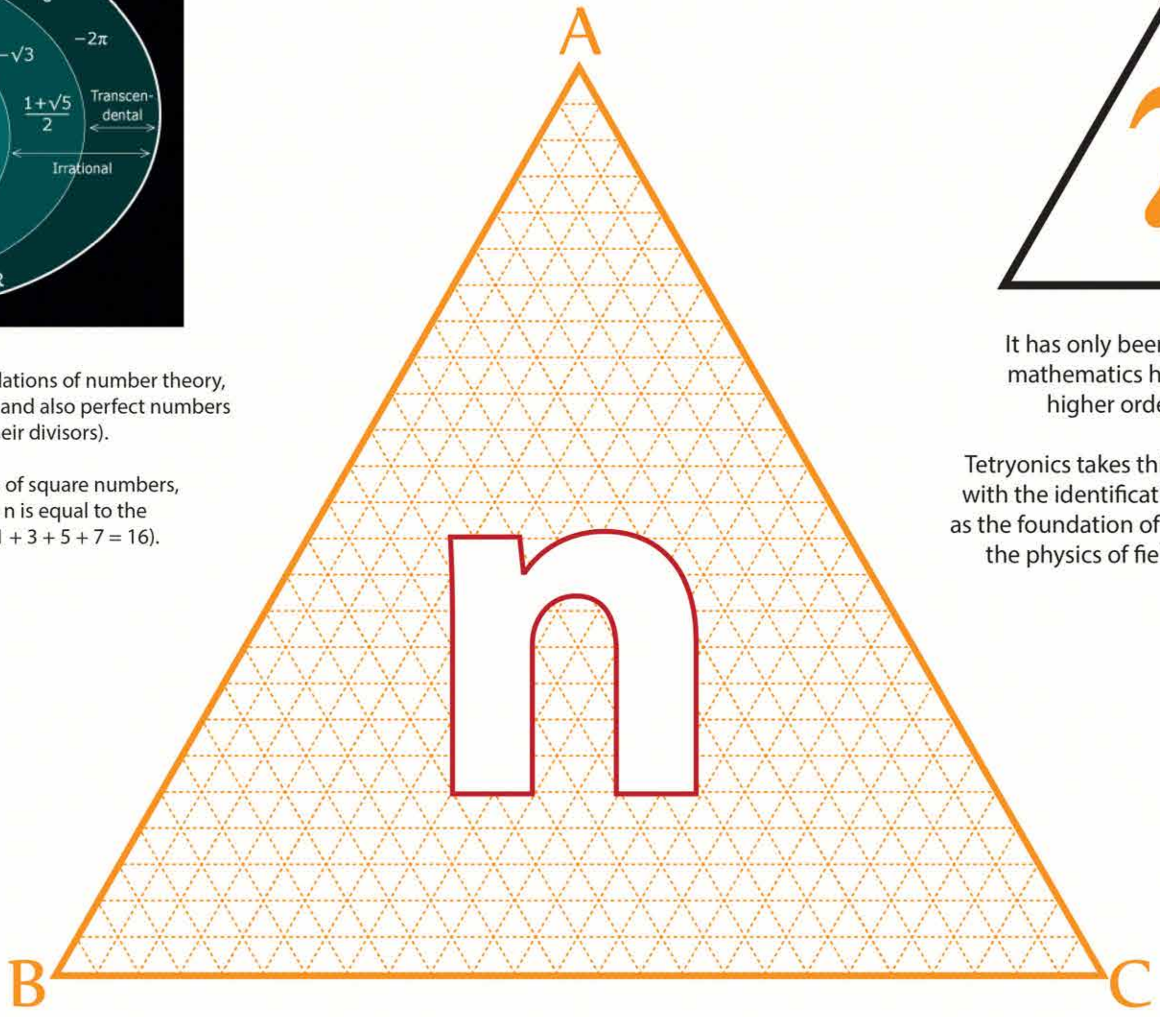

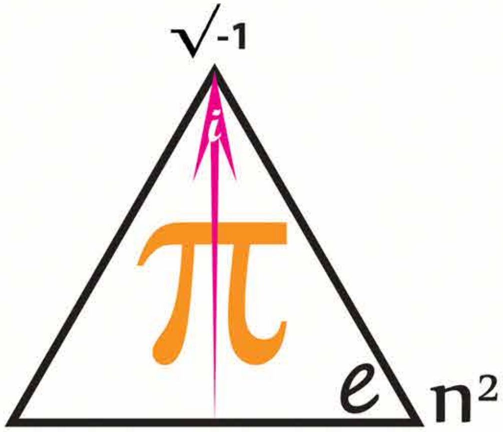

It has only been in recent centuries that mathematics has begun to explore the higher order irrational numbers

Tetryonics takes this investigation to new levels with the identification of equilateral geometries as the foundation of transcendental numbers and the physics of fields and particles in motion

What mathematics has failed to appreciate is the significance application of equilateral geometries to the 'square' numbers of physics and science in general

Copyright ABRAHAM [2008] - All rights reserved

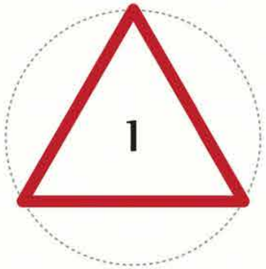

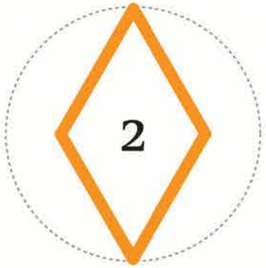

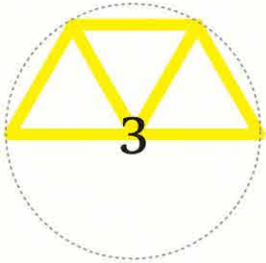

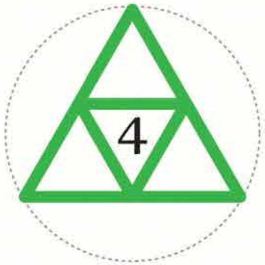

TETRYONICS

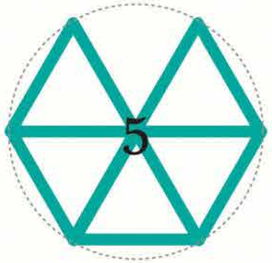

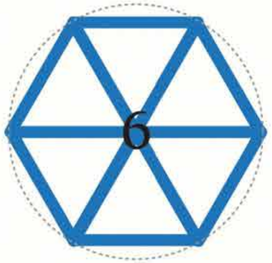

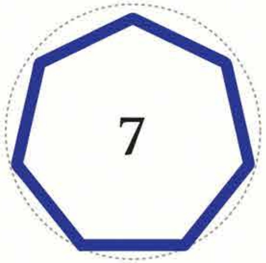

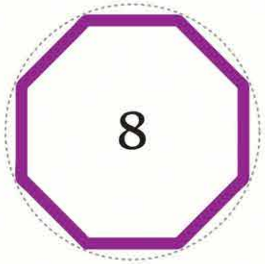

COUNTING POLYGONS

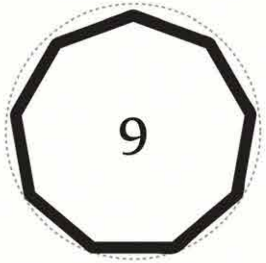

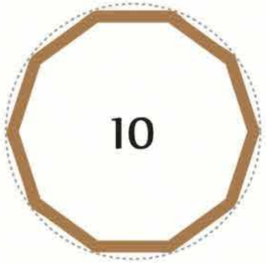

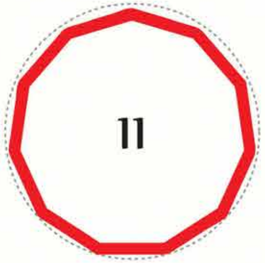

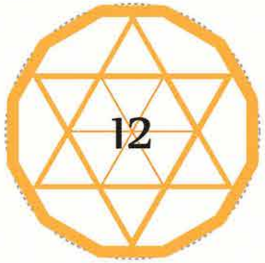

SQUARED energies in quantum mechanics are EQUILATERAL geometries

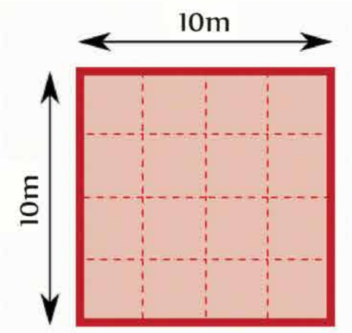

Square

area = s^2 = [100]

Circles

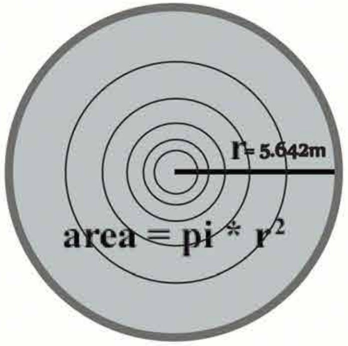
r= 5.642m
area = pi * r²

$= pi *[5.642]^2$
= 100

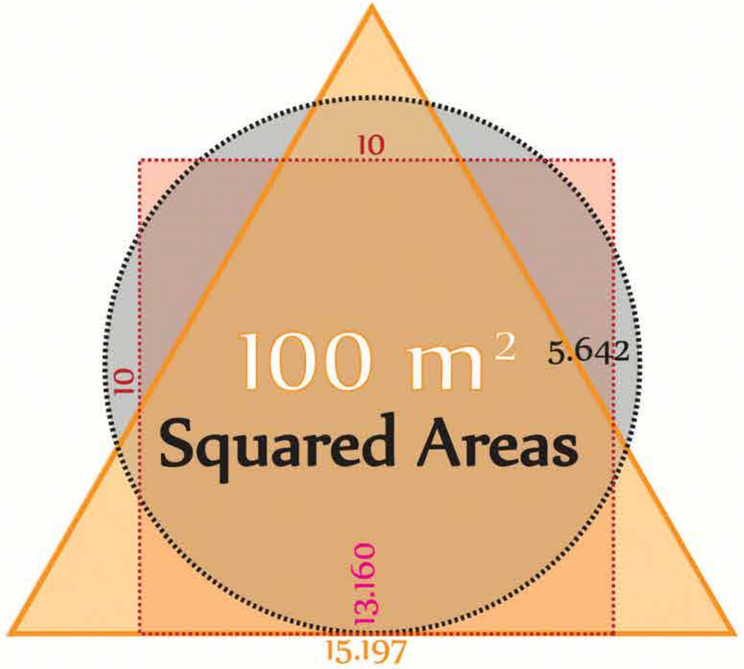
10
10
100 m²
Squared Areas
5.642
13.160
15.197

can be created by a number of planar geometries

For a long time it has been assumed by scientists (and mathematicians)
that circular [and squared] geometries are the geometric foundation of all physics,
leading to a serioulsy flawed model of particles and forces in quantum mechanics

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

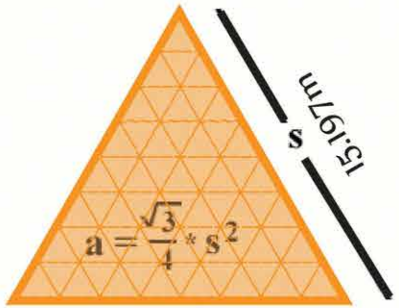
15.197m
s
$a = \frac{\sqrt{3}}{4} * s^2$

Equilateral

area = $(\frac{1}{2}*b)*h$

Triangles

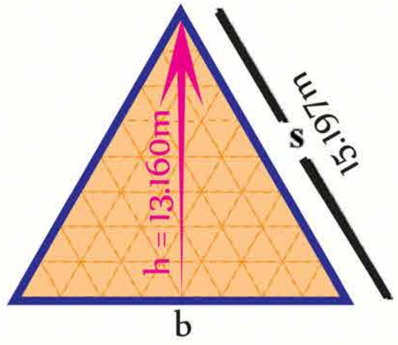
15.197m
s
h = 13.160m
b

$[\overset{b}{.5\text{x}15.197}] \times \overset{h}{13.160}$
= 100

Copyright ABRAHAM [2008] - All rights reserved

0.987654321

SQUARE ROOTS

80/81

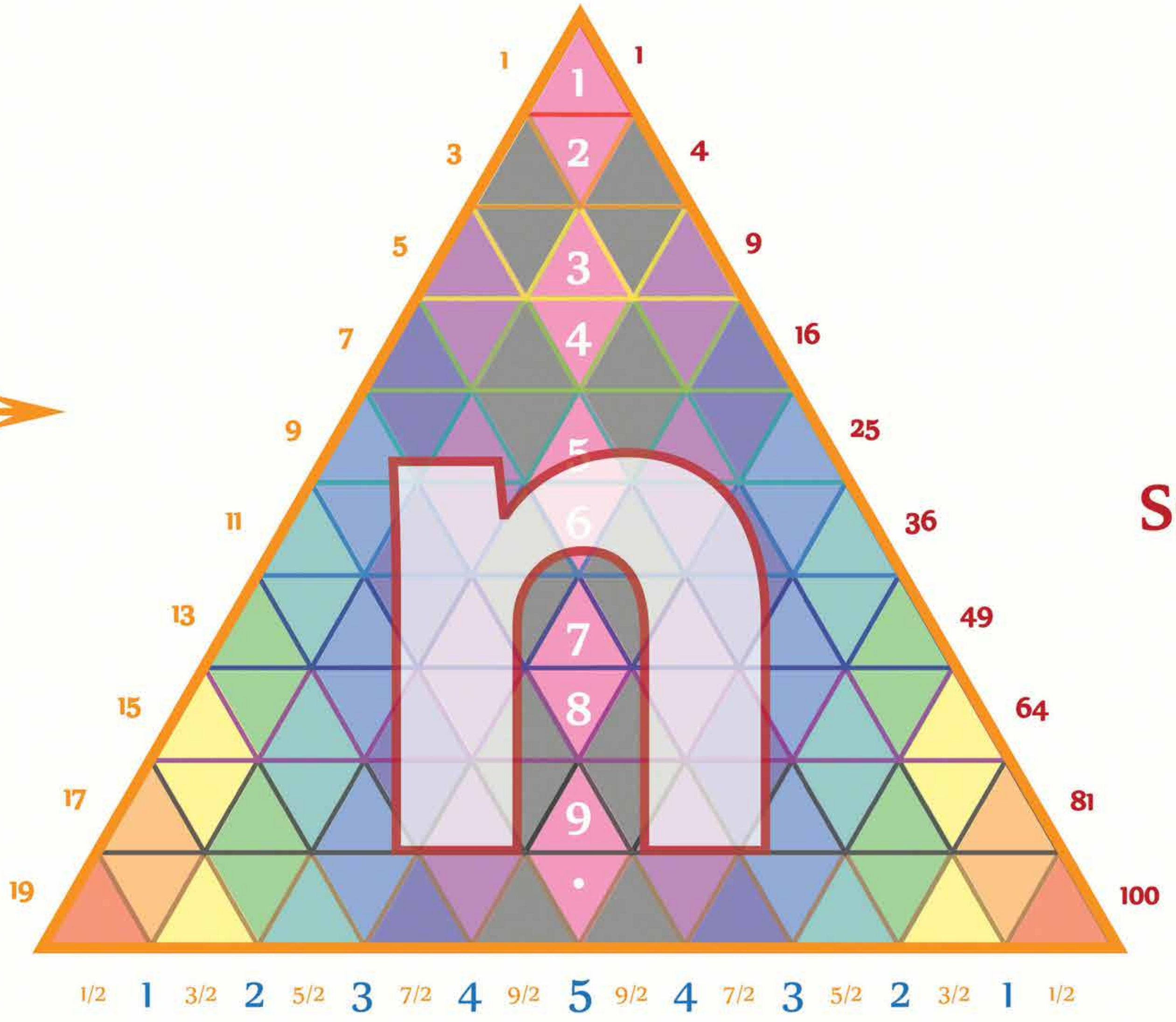
1
2
3
4
5
6
7
8
9
•
n
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2

ODDS

SQUARES

EVEN DISTRIBUTIONS

Copyright ABRAHAM [2008] - All rights reserved

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the Planck charge quantum

$$2n-1$$

$$[[n]+[n-1]]$$

Triangular numbered geometries are NOT equialteral geometries

0, 1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

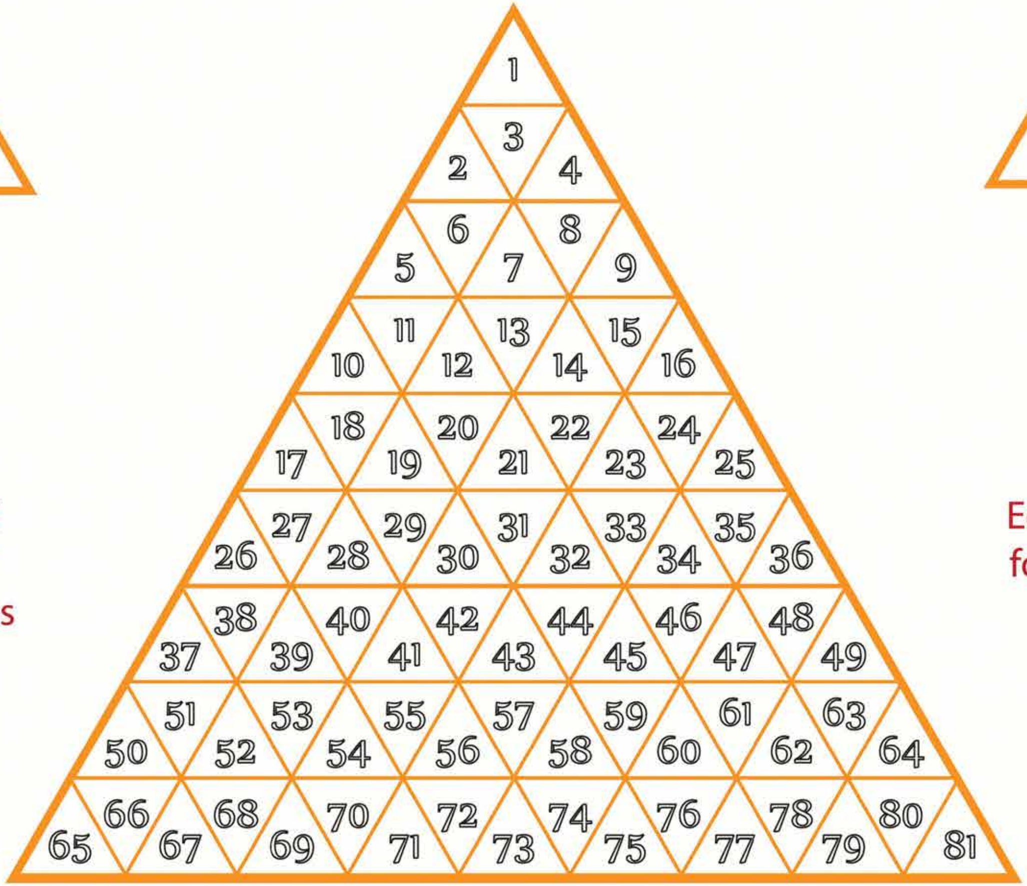
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1

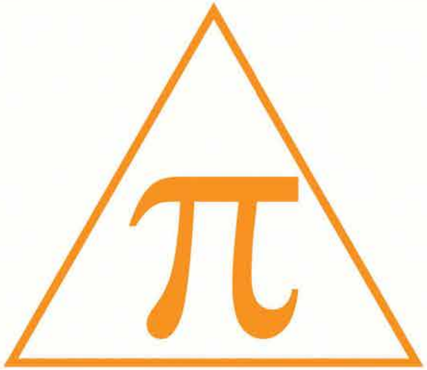

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form **SQUARE** number geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Equilateral energy quanta form a normal longitudinal distribution

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

ε_0
π
μ_0
μ_0

Bosons are a transverse measure of scalar energy momenta

ODD numbers

An odd number is an integer which is not a multiple of two.

E
μ_0
n
μ_0
E
ε_0
ε_0
+
3
–
N
μ_0
μ_0
S
S
μ_0
μ_0
N

Quantum

Bosons have ODD number quanta

Quantum Levels

hv

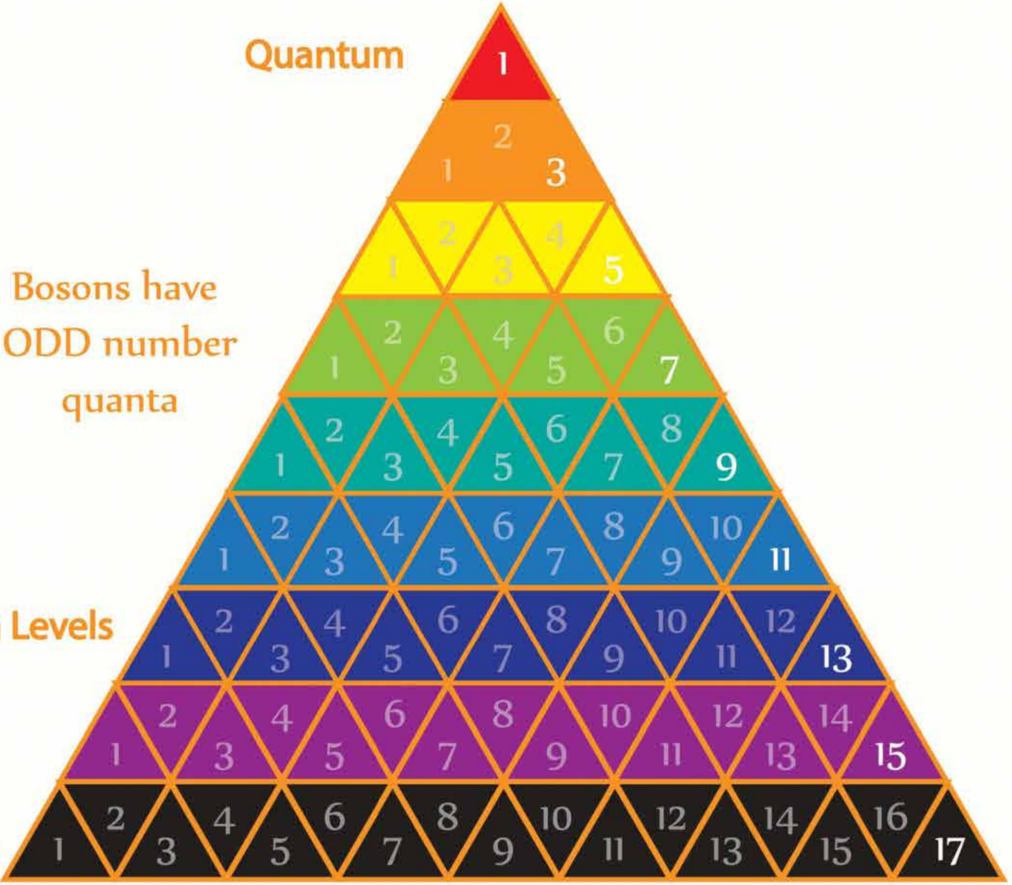

1
1 2 3
1 2 3 4 5
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

1

3

5

7

9

11

13

15

17

An odd number, when divided by two, will result in a fraction

Copyright ABRAHAM [2008] - All rights reserved

Tau radians

Around the whole outside of a circle,
there are about 6.283 radians - or, Tau radians

3.141592654

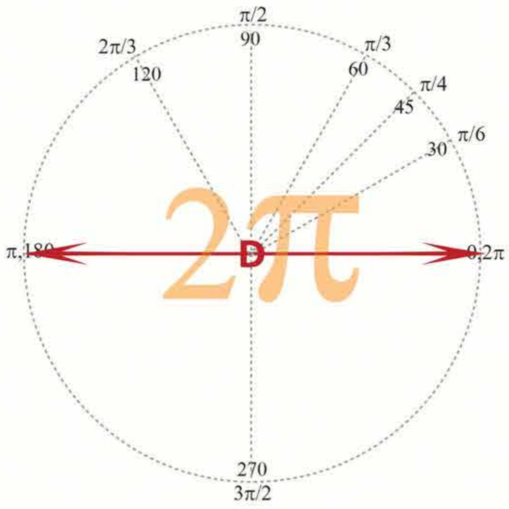
π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
0,2π
2π
D
270
3π/2

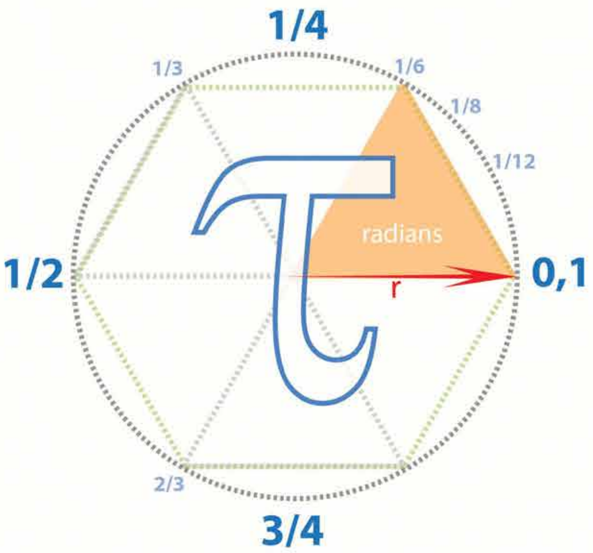
1/4
1/3
1/6
1/8
1/12
1/2
0,1
τ
radians
r
2/3
3/4

6.283185307

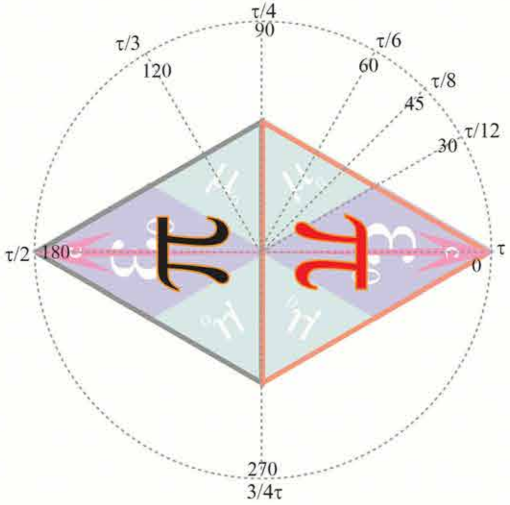
τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
τ
0
π
π
270
3/4τ

τ is a more 'natural' radian system for geometric physics than π
Tau = 2π = 360 degree rotation about a point

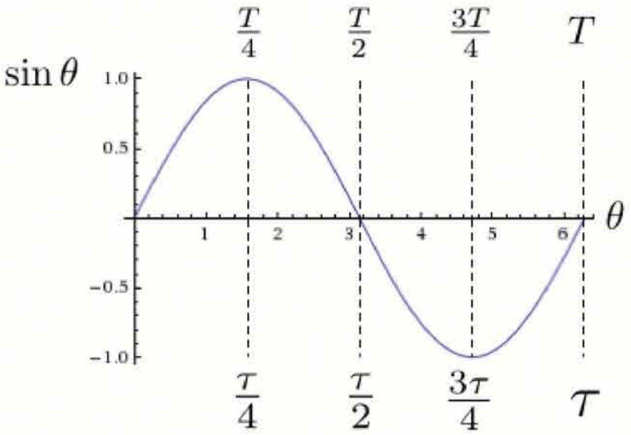

π

historically defined as the
ratio of a circle's circumference to its DIAMETER
should be redefined in physics to

τ

the ratio of its circumference to its RADIUS

in doing so many of the π/2 terms common to physics
will re automatically rationalised and will better reflect
the Tetryonic geometry of mass-ENERGY-Matter in motion

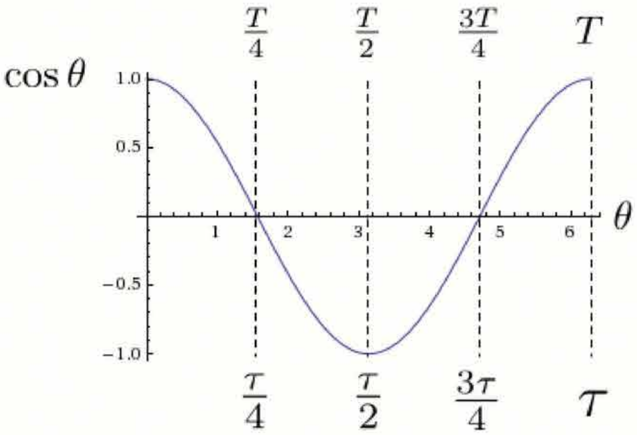

Photons are a longitudinal measure of scalar energy momenta

E
ε_0
+
2
N
N
S
S
−
μ_0
E

Photons are
EVEN number
quantums

EVEN numbers

An integer that is not an odd number is an even number

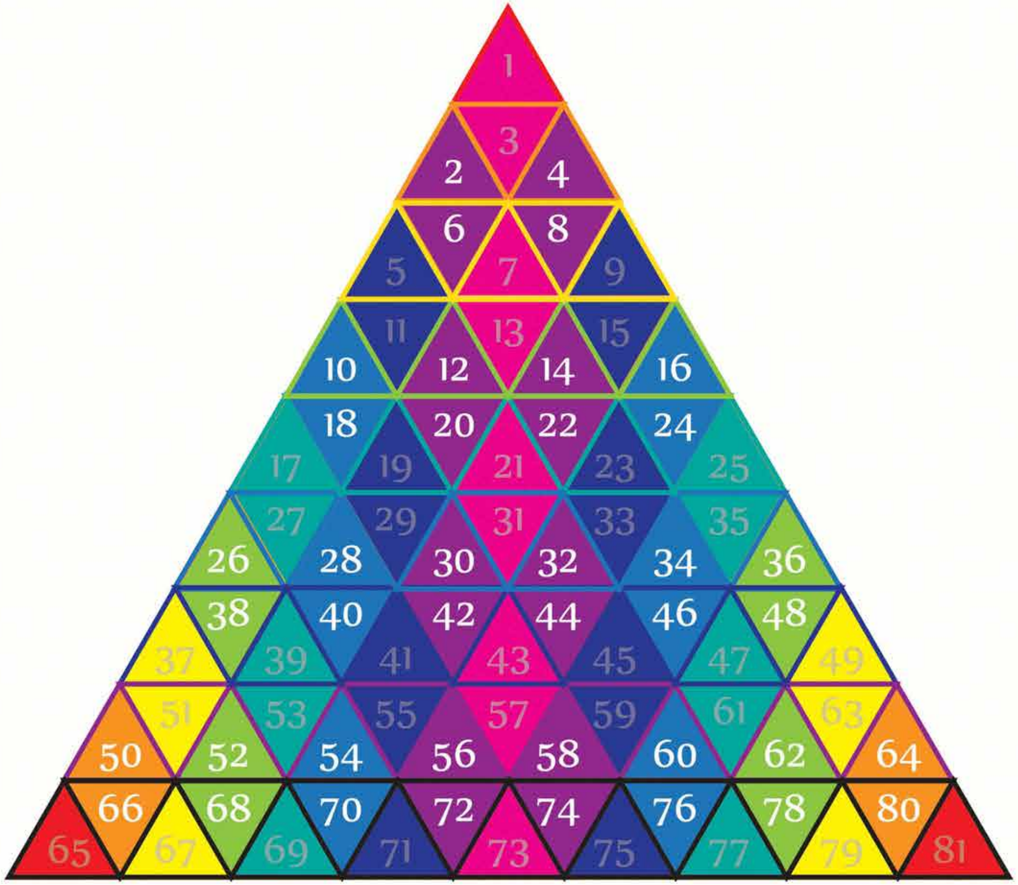
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

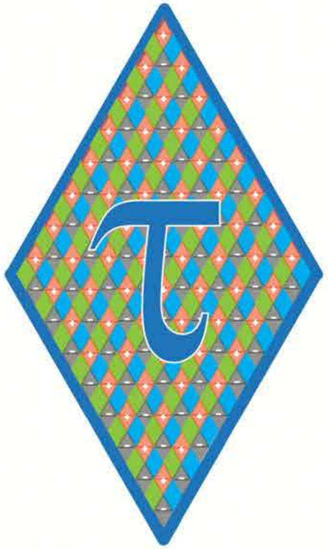

EM waves are
comprised of EVEN
numbered quanta

If an even number is divided by two, the result is another whole number

1/2 **1** 3/2 **2** 5/2 **3** 7/2 **4** 9/2 **4** 7/2 **3** 5/2 **2** 3/2 **1** 1/2

If an odd number is divided by two, the result is a fractional number

Triangular numbers

Historically, a triangular number counts quanta that can pack together to form an equilateral triangle

1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

this form of geometric counting of same charges over-complicated the simpler physical reality

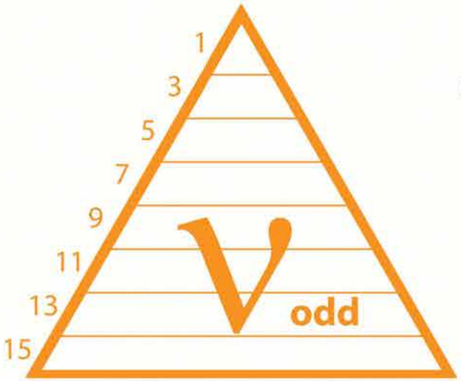
1
3
5
7
9
11
13
15
ν
odd

$$2\nu-1$$

$$[[\nu] + [\nu-1]]$$

Equilateral chords or quantum levels are ODD numbers

1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,....

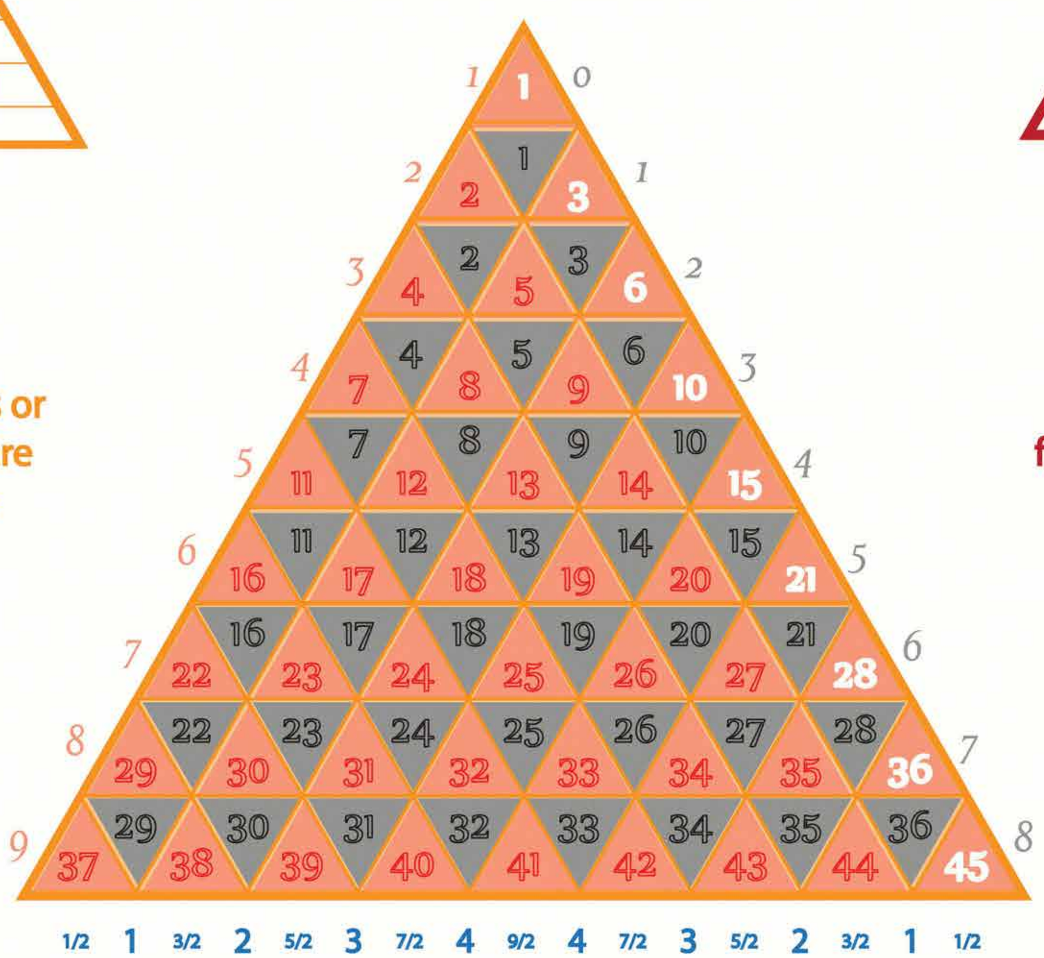
1
2
3
4
5
6
7
8
9
0
1
2
3
4
5
6
7
8
1
2 1 3
4 2 5 3 6
7 4 8 5 9 6 10
11 7 12 8 13 9 14 10 15
16 11 17 12 18 13 19 14 20 15 21
22 16 23 17 24 18 25 19 26 20 27 21 28
29 22 30 23 31 24 32 25 33 26 34 27 35 28 36
37 29 38 30 39 31 40 32 41 33 42 34 43 35 44 36 45
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

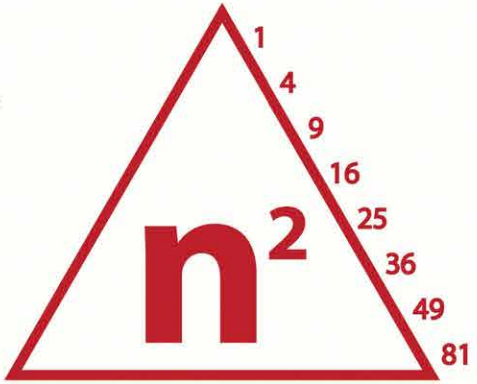
n²
1
4
9
16
25
36
49
81

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form SQUARE numbered geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Triangular energy quanta form normal distributions

Copyright ABRAHAM [2008] - All rights reserved

Squared numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

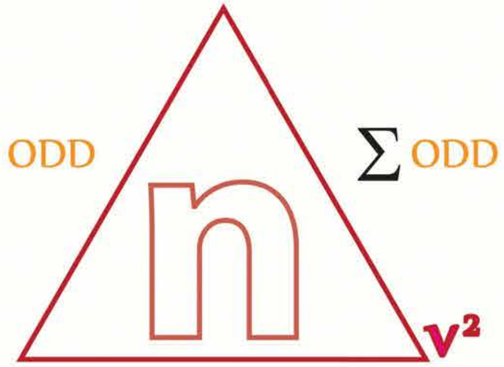
ODD
$\sum$ ODD
n
v^2

1
+
3
+
5
+
7
+
9
+
11
+
13
+
15
+
17

Square numbers result from
the summing of consecutive
ODD numbers

ODDS

2n-1

[[n] + [n-1]]

Quantum Energies

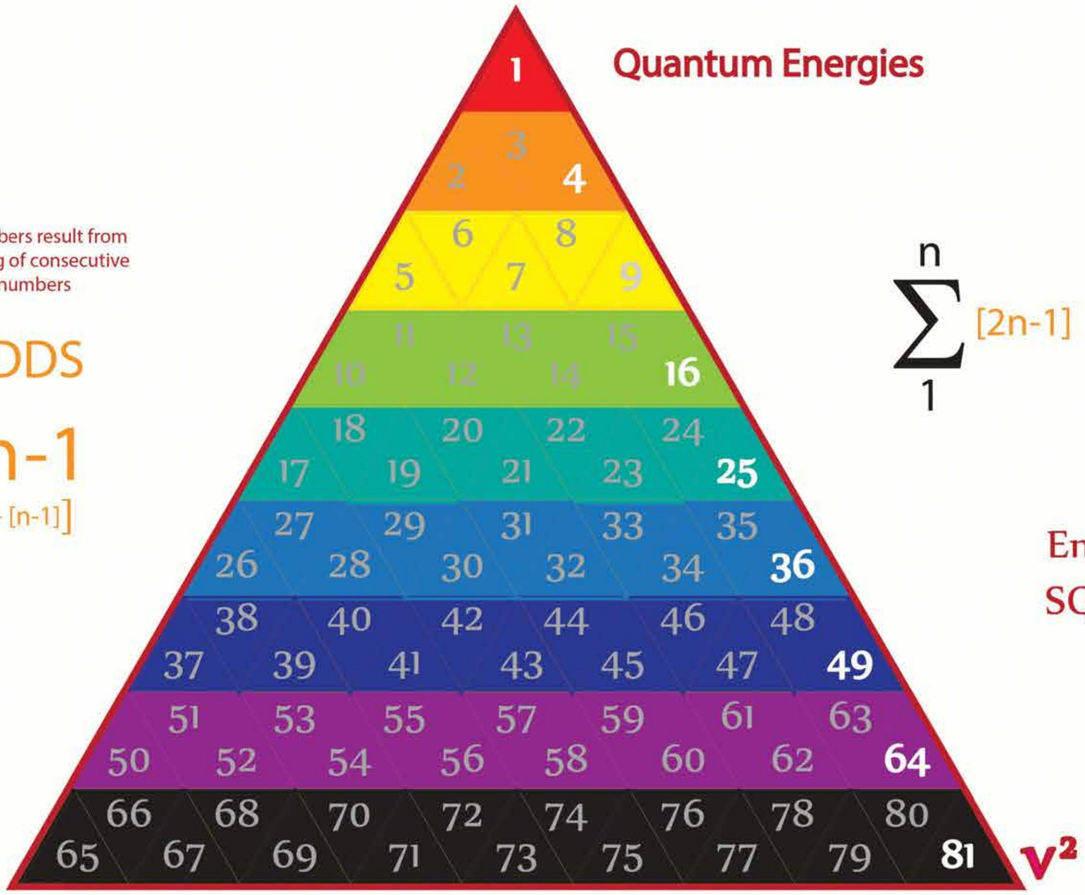
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
v^2

$$\sum_{1}^{n}[2n-1]$$

Energy levels have
SQUARE number
quanta

Compton Frequency

In Tetryonics Square numbers produce equilateral geometries

Copyright ABRAHAM [2008] - All rights reserved

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

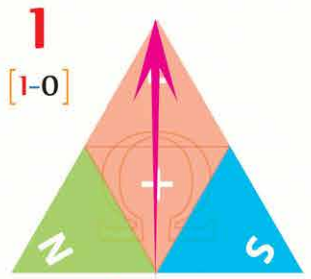

Root of positive one

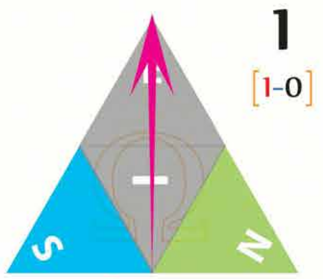

Root of negative one

-*i* and +*i*

Square roots of negative numbers have a basis in physical reality

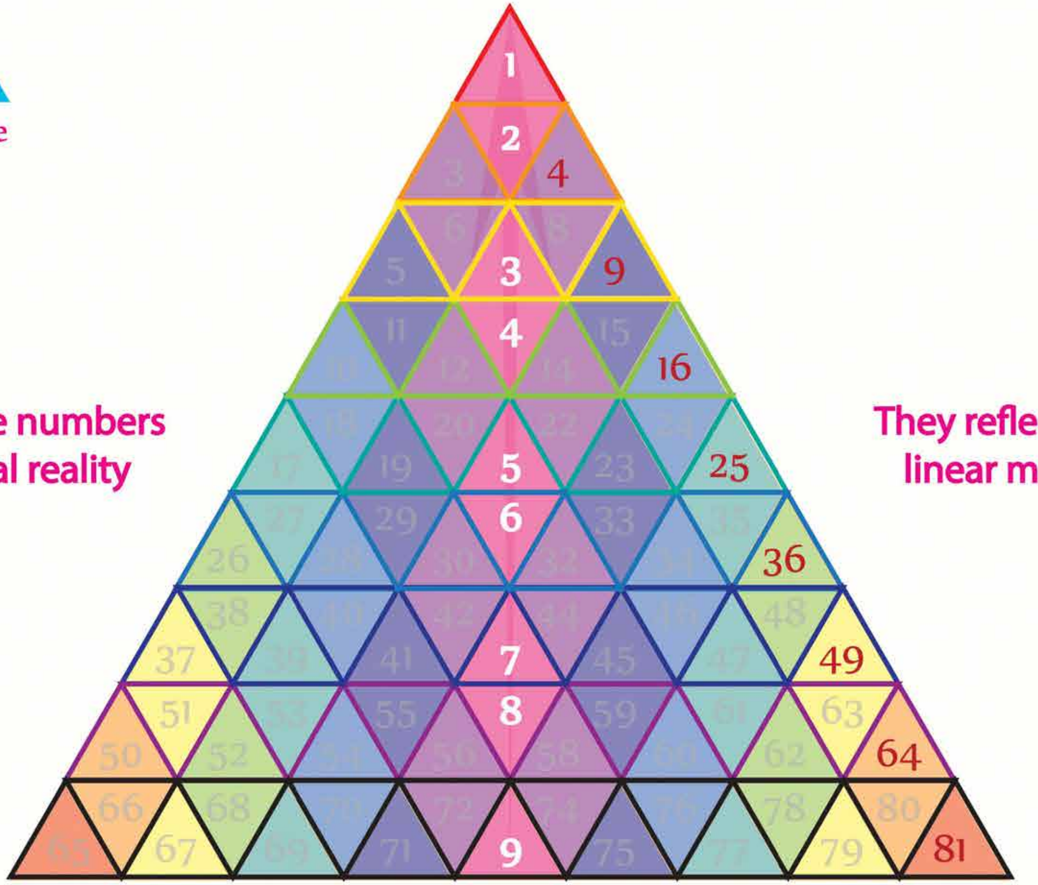

They reflect the real non-negative linear momentum of a system

A whole number with a square root that is also a whole number is called a perfect square in Tetryonic theory they are actually equilateral geometries

Real Numbers

In mathematics, a real number is a value that represents a quantity along a continuous line. The real numbers include all the rational numbers,

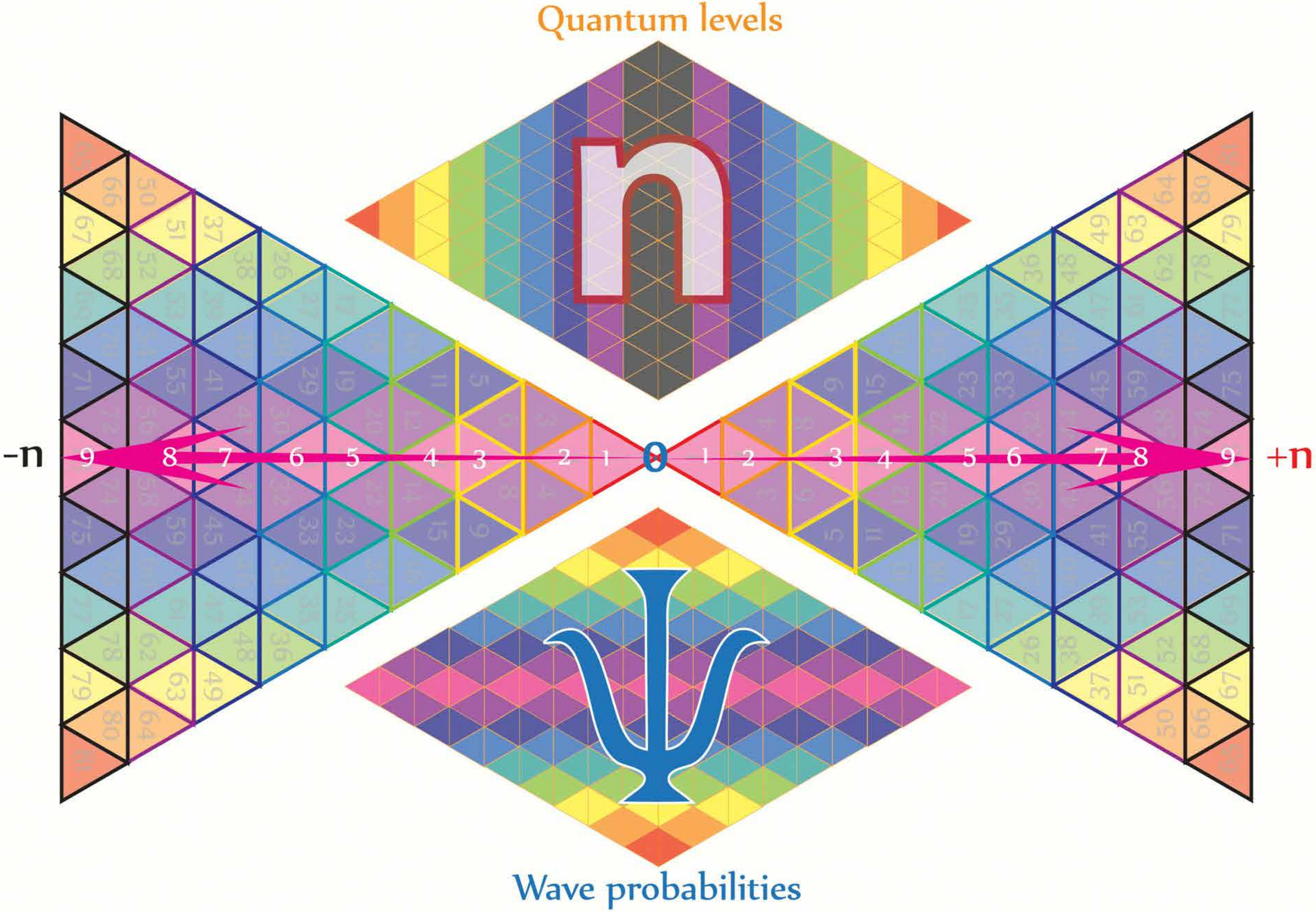
Quantum levels
n
-n
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
+n
Ψ
Wave probabilities

Copyright ABRAHAM [2008] - All rights reserved

ONE to ∞^2

Quantum levels

ODD numbers

[2n-1]

1, 3, 5, 7, 9......

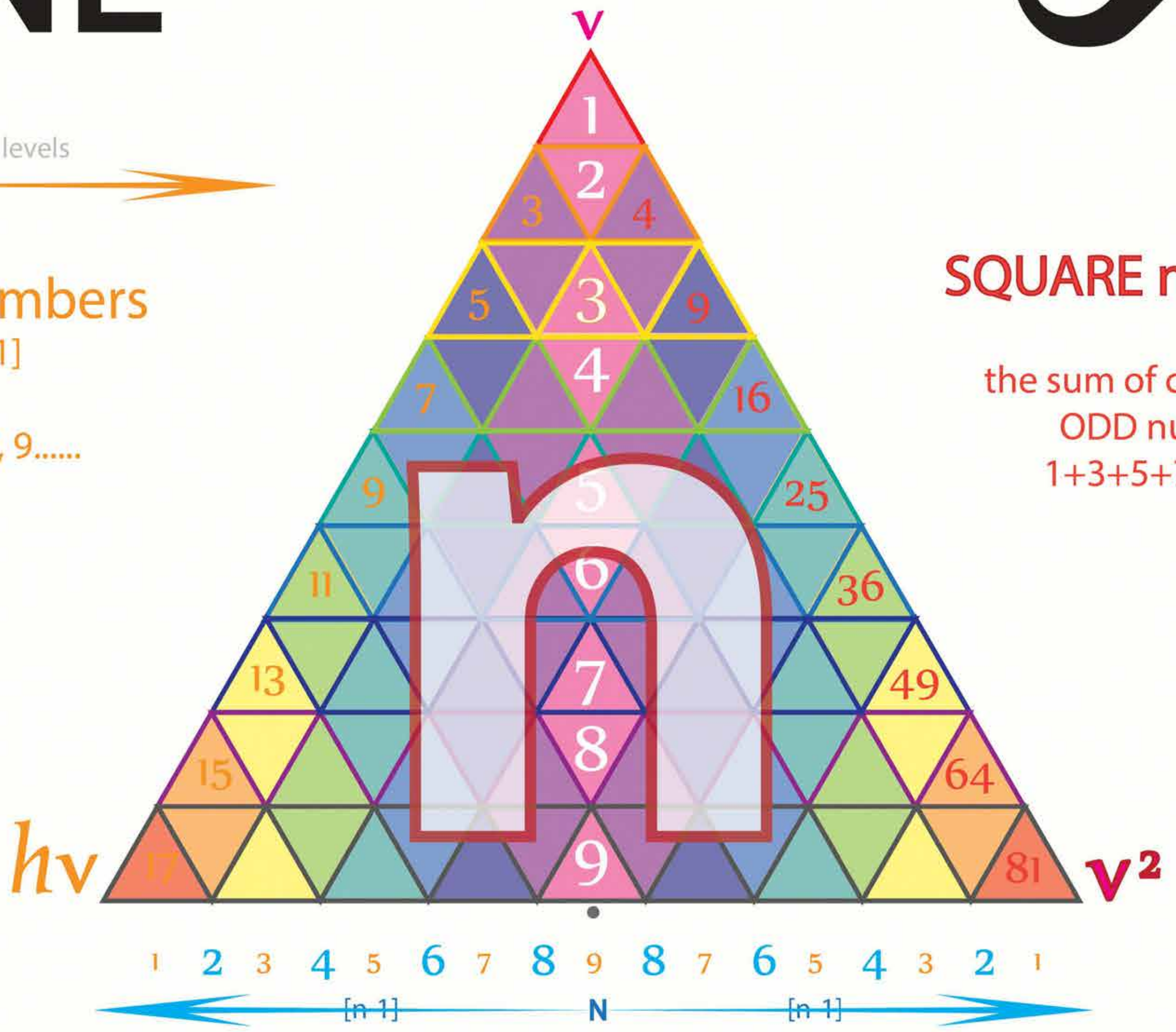
v
1
2
3
4
5
3
9
7
4
16
9
5
25
11
6
36
13
7
49
15
8
64
17
9
81
n
hv
v²
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
N
[n-1]

SQUARE numbers

the sum of consecutive
ODD numbers
1+3+5+7+9+......

Scalar Energies

Linear energy momentum

NORMAL DISTRIBUTION

Tetryonics 82.13 - One to Infinity squared

Copyright ABRAHAM [2008] - All rights reserved

Basic Properties of nested scribed Equilateral Triangles

Given an equilateral triangle of side s

altitude

$\sqrt{3/2}\,S$

$3S$

perimeter

$\sqrt{3/4}\,S^2$

area

$\sqrt{3/6}\,S$

in-radius

$\sqrt{\pi/12}\,S^2$

in-circle area

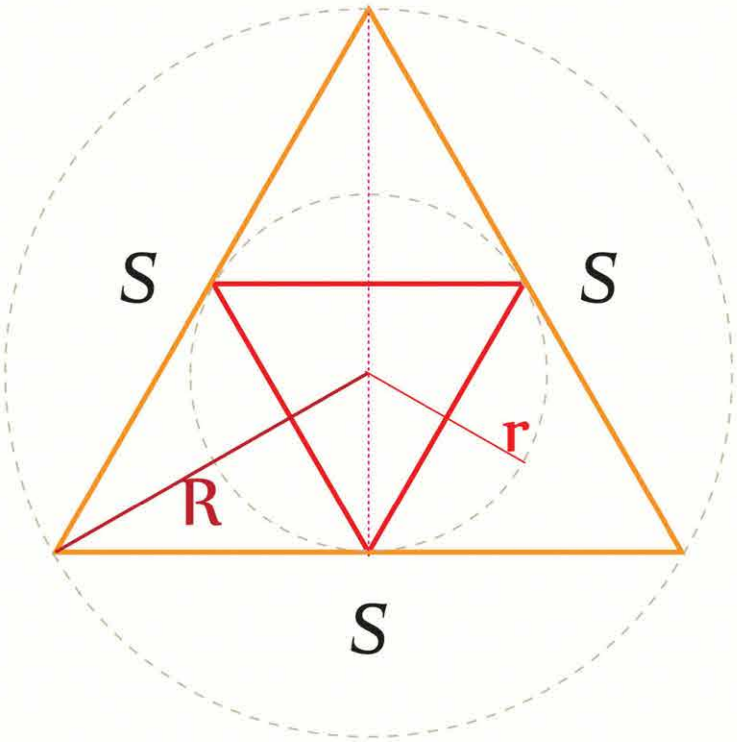

circum-radius

$\sqrt{3/3}\,S$

circum-circle area

$\sqrt{\pi/3}\,S^2$

Tetryonic [equilateral] geometry

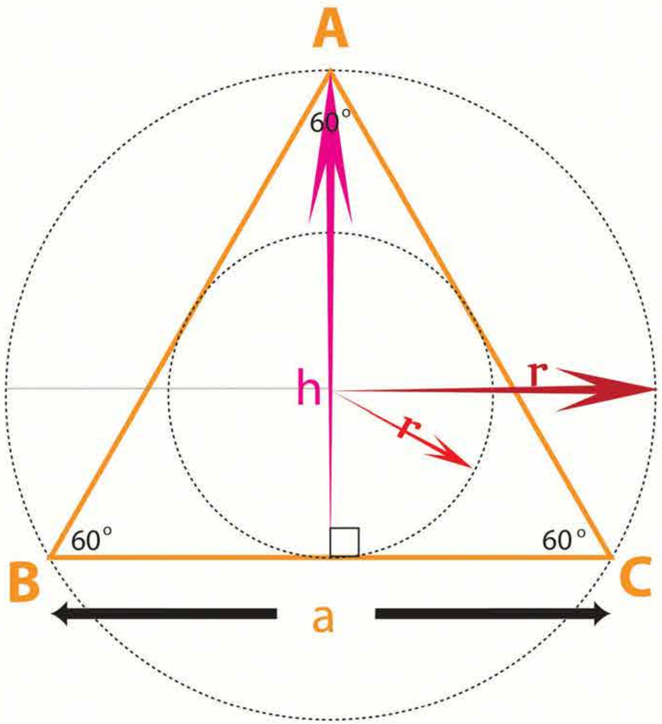

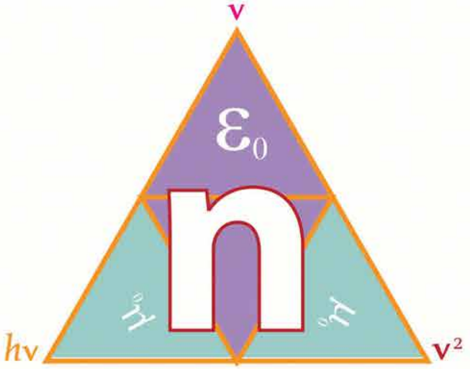

The equilateral triangle exhibits 'square symmetry'
it can always be divided into n^2 [number]
of smaller self-similar parts

**All triangles are
flat euclidean
π radian
geometries
[180°]**

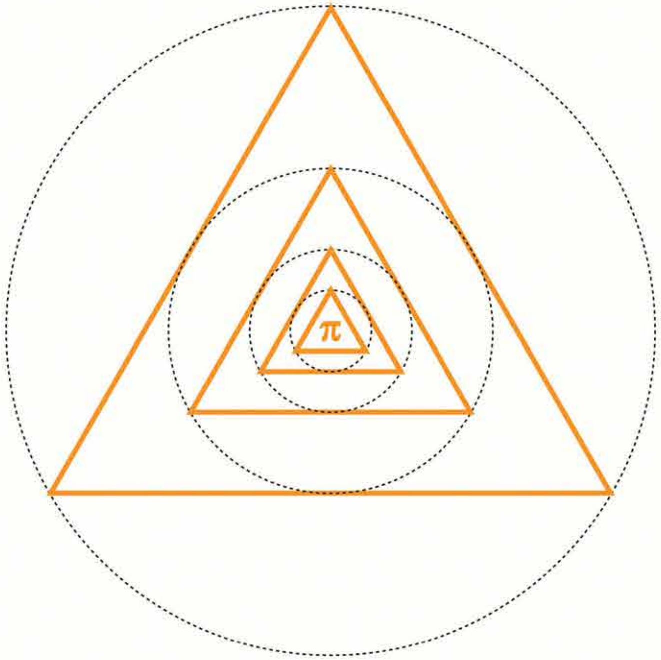

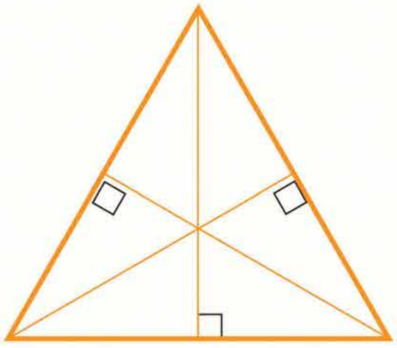

An equilateral triangle is the most symmetrical triangle,
having 3 lines of reflection and rotational symmetry of order 3 about its center

Copyright ABRAHAM [2008] - All rights reserved

CHARGE

Scribed equilateral geometries

reflect space-time's geometric relationship with charged mass-energy

Energy

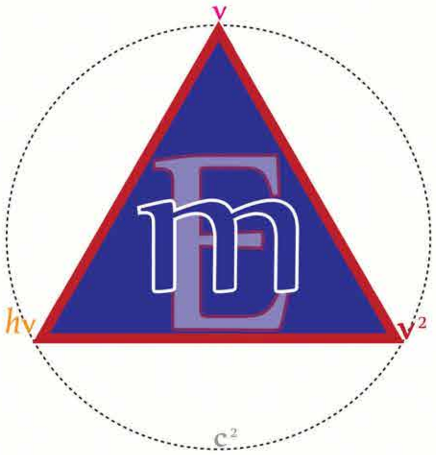

Energy per
spatial unit

mass

$$m = \frac{E}{c^2}$$

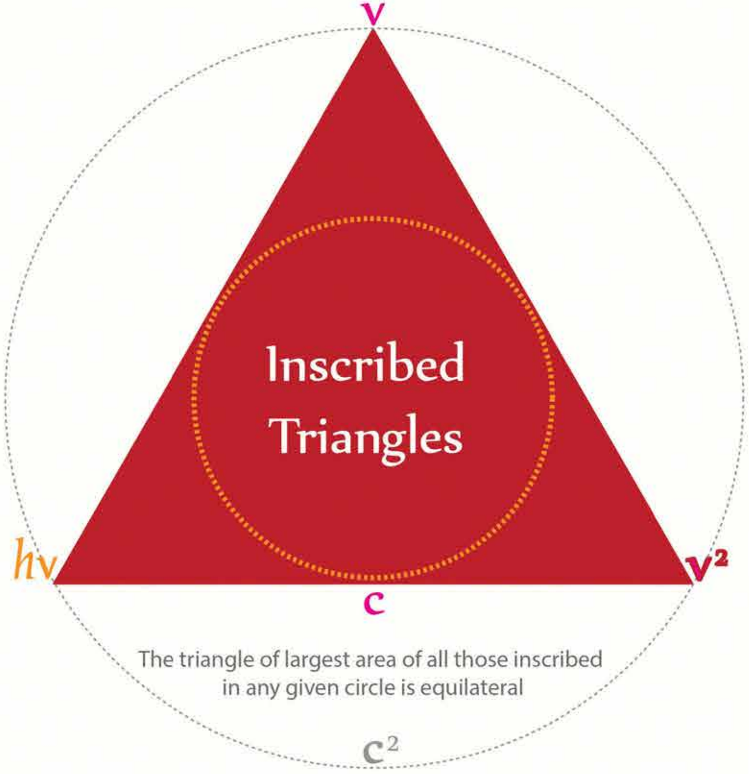

v
Inscribed
Triangles
hv
v²
c
c²

The triangle of largest area of all those inscribed
in any given circle is equilateral

QAM

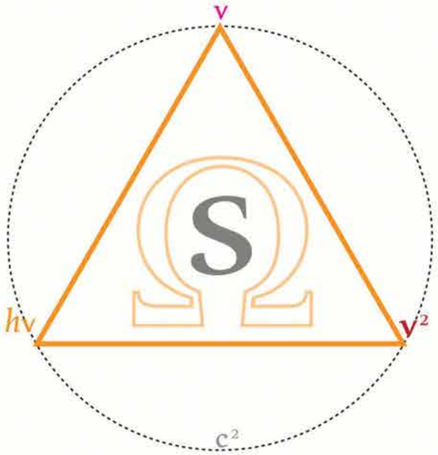

QAM per
spatial unit

seconds

$$\frac{\Omega}{c^2} = s$$

Copyright ABRAHAM [2008] - All rights reserved

Circumscribed Triangles

reflect Energy's relationship with Time

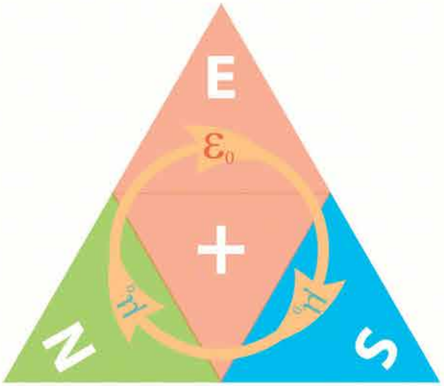
E
ε_0
+
μ_0
μ_0
N
S

Positive Planck Charge

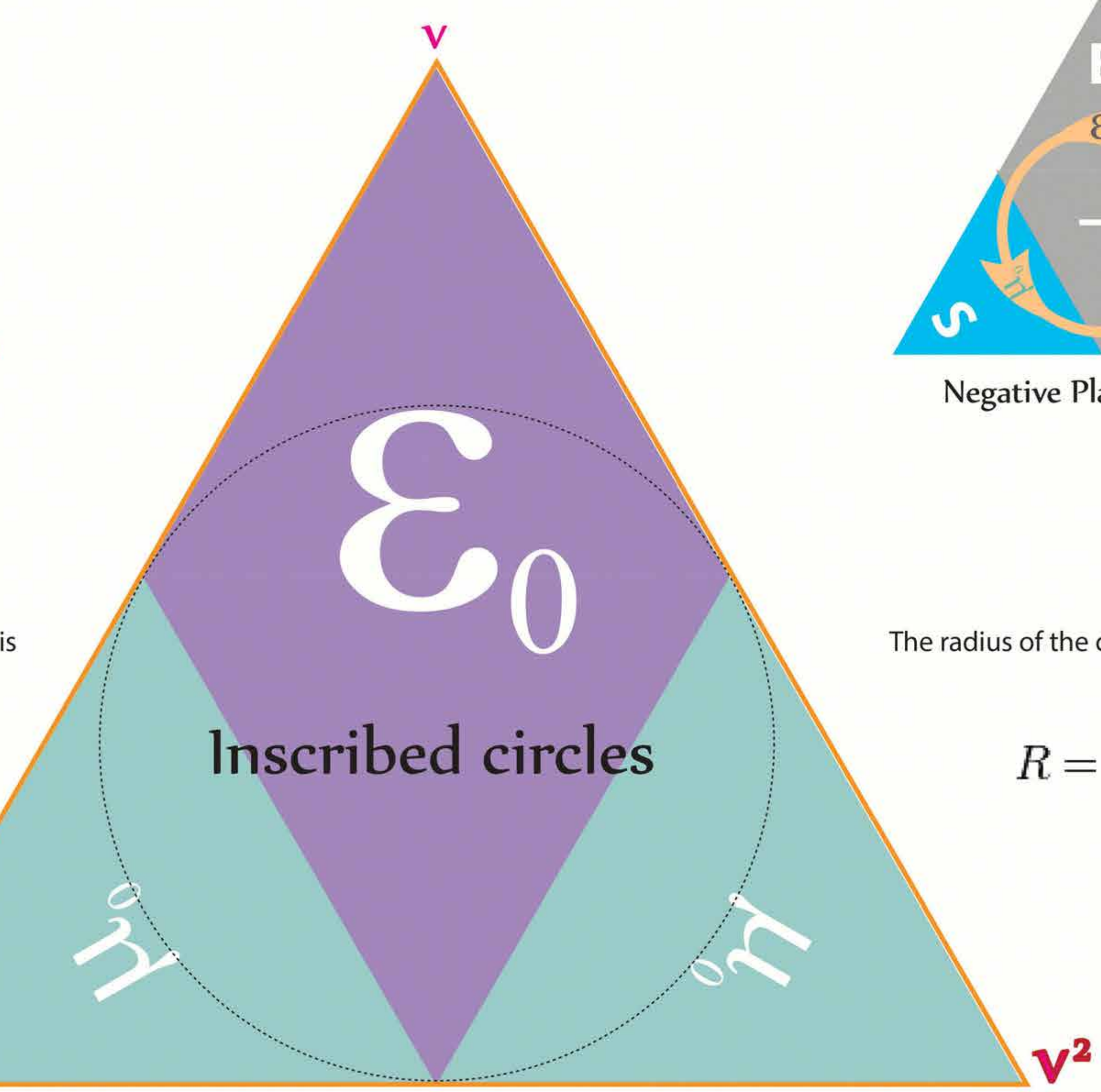
v
ε_0
Inscribed circles
μ_0
μ_0
hv
v^2

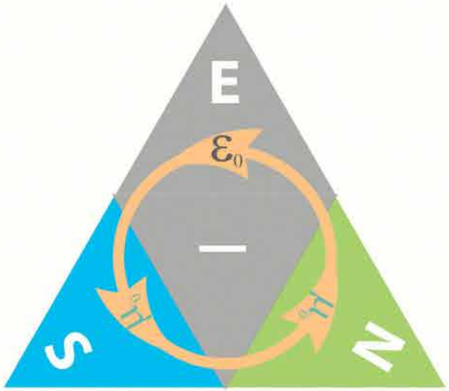
E
ε_0
−
μ_0
μ_0
S
N

Negative Planck Charge

The perimiter of an equilateral triangle is

$$p = 3a$$

The radius of the circumscribed circle is

$$R = \frac{\sqrt{3}}{3}a$$

The equilateral triangle has the smallest area of all those circumscribed around a given circle

Circumscribed circles

By Euler's inequality, the equilateral triangle has the smallest ratio R/r of the circumradius to the inradius of any triangle: specifically, R/r = 2

The ratio of the area to the square of the perimeter of an equilateral triangle

$$\frac{1}{12\sqrt{3}}$$

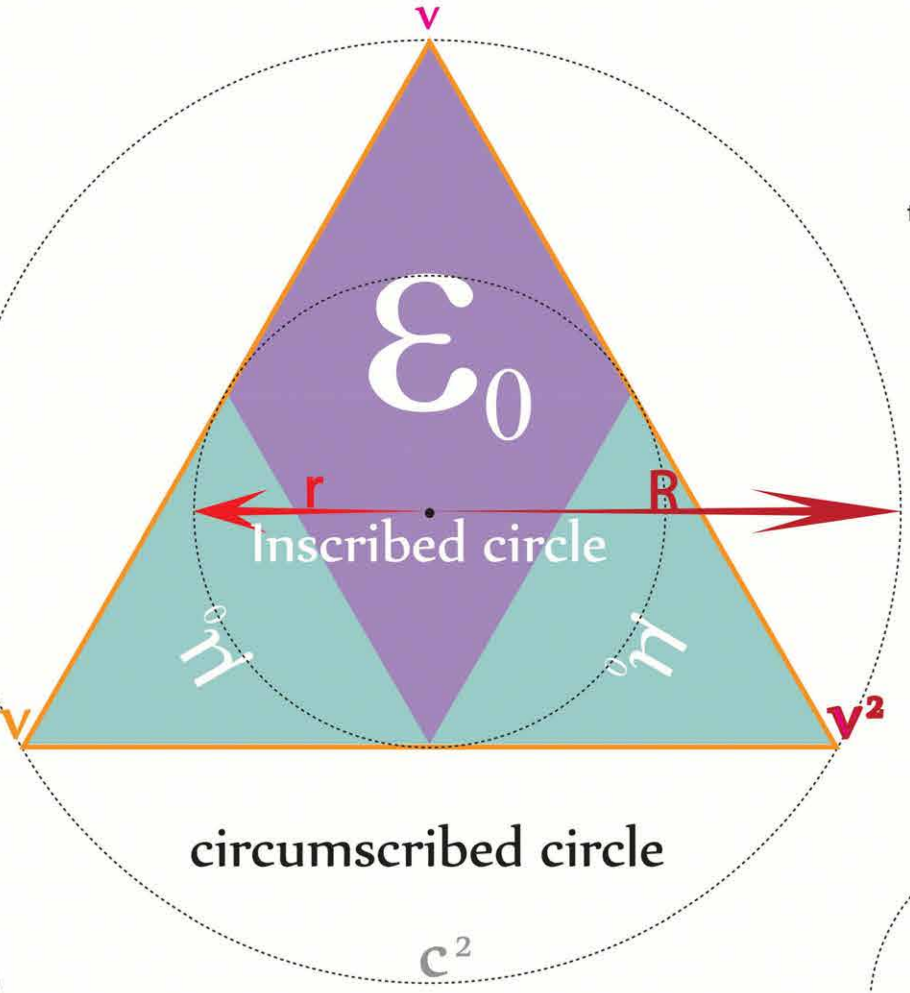
v
ε_0
r
R
Inscribed circle
μ_0
μ_0
hv
v^2
circumscribed circle
c^2

The ratio of the area of the incircle to the area of an equilateral triangle

$$\frac{\pi}{3\sqrt{3}}$$

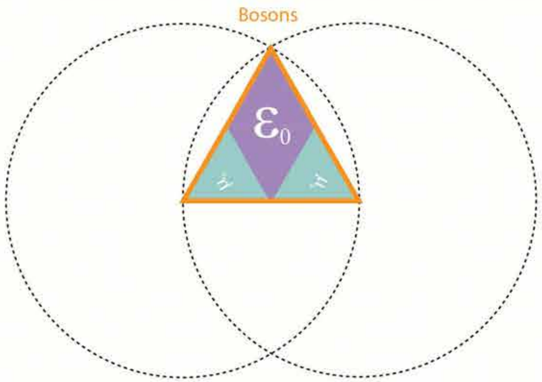
Bosons
ε_0

Equilateral triangles and Tetrahedrons will scale at exactly the same proportion as Circles and Spheres scribing them

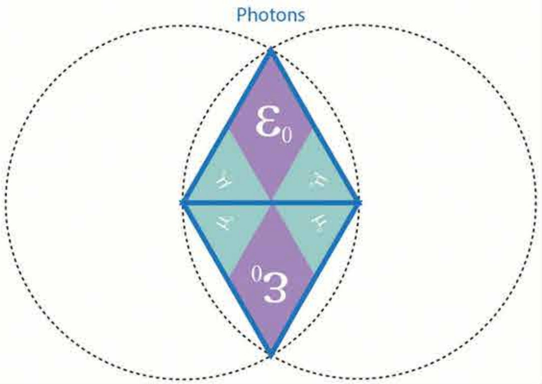
Photons
ε_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

Charged mass-ENERGIES

Charge is the result of quantised angular momenta
[the inscribed circular flux of energy in equilateral geometries]

Time is a measure of changing quantised angular momenta
[the circumscribed spatial co-ordinate of equilateral energy geometries]

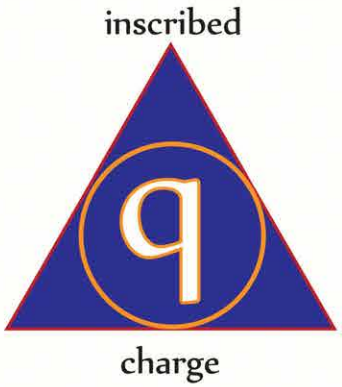

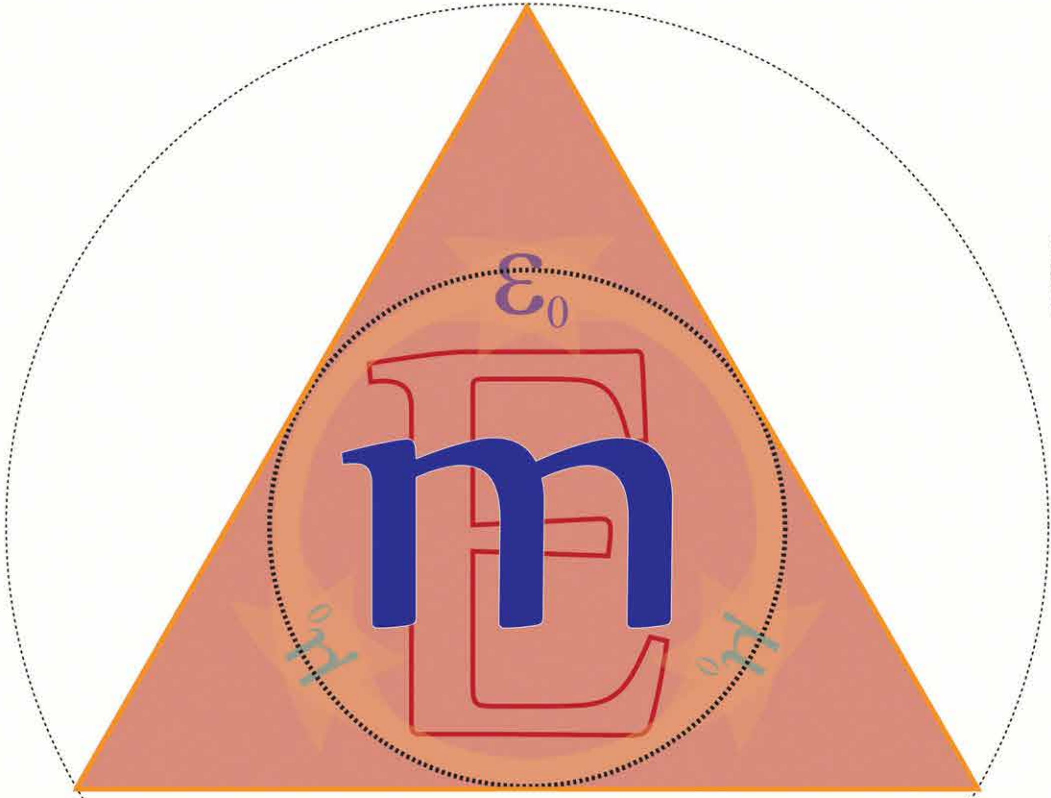

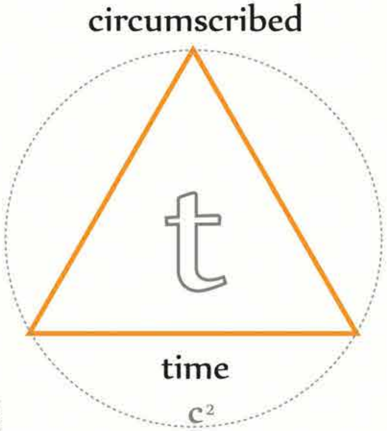

It is the equilateral geometry of energy not a classical vector rotation that creates QAM

Scalar EM mass is a measure of equilateral Planck energy per spatial co-ordinate system

m

Charged mass-energies can be described as
a vector measure of scalar Energy momenta per unit of time
[inscribed circular fluxes of equilateral energy per circumscribed temporal geometry]

E

Copyright ABRAHAM [2008] - All rights reserved

Trigometric functions

$$\sin\,\theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos\,\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan\,\theta = \frac{\text{opposite}}{\text{adjacent}}$$

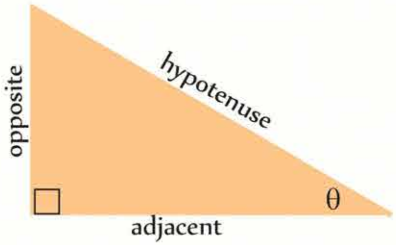
hypotenuse
opposite
θ
adjacent

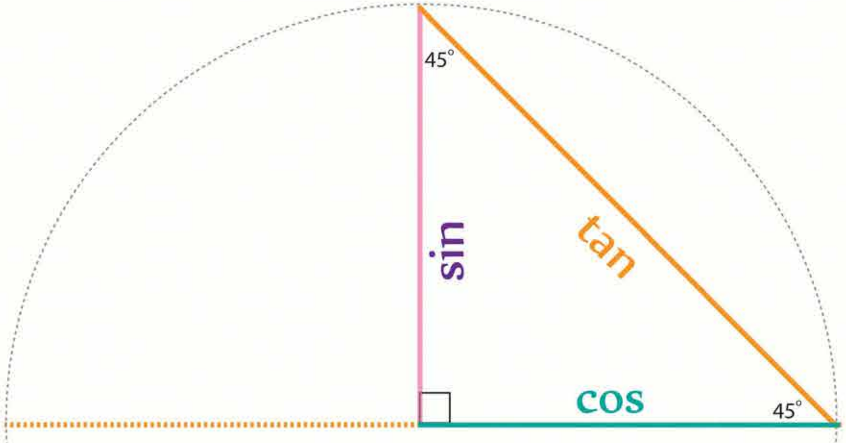
45°
sin
tan
cos
45°

The most familiar trigonometric functions are the sine, cosine, and tangent.

In the context of the standard unit circle with radius 1,
where a triangle is formed by a ray originating at the origin and
making some angle with the x-axis,

the SINE of the angle gives the length of the y-component (rise) of the triangle,
the COSINE gives the length of the x-component (run), and
the TANGENT function gives the slope
(y-component divided by the x-component)

Standard trigometric functions must be carefully applied to
measurements of equilateral Planck mass-energy geometries
in scribed circular space-time co-ordinate systems

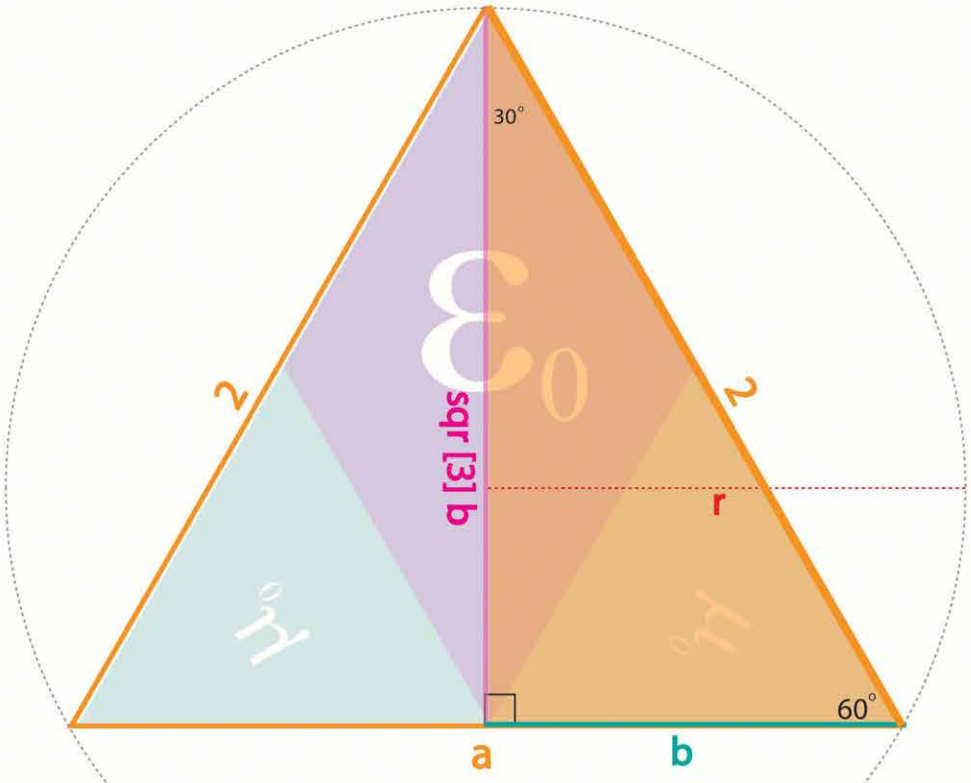
30°
ε0
sqr [3] b
2
2
r
μ0
μ0
60°
a
b

In most cases it is in fact easier to
ignore the trigonomteric functions and math
and simpy draw their related equilateral geometries

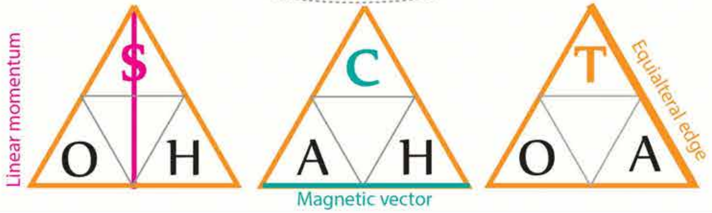
Linear momentum
S
O
H
C
A
H
Magnetic vector
T
O
A
Equilateral edge

Copyright ABRAHAM [2008] - All rights reserved

The roots of scribed equilateral triangles

Scalar equilateral energies map directly onto circular space-time co-ordinates
through their square root linear momentum

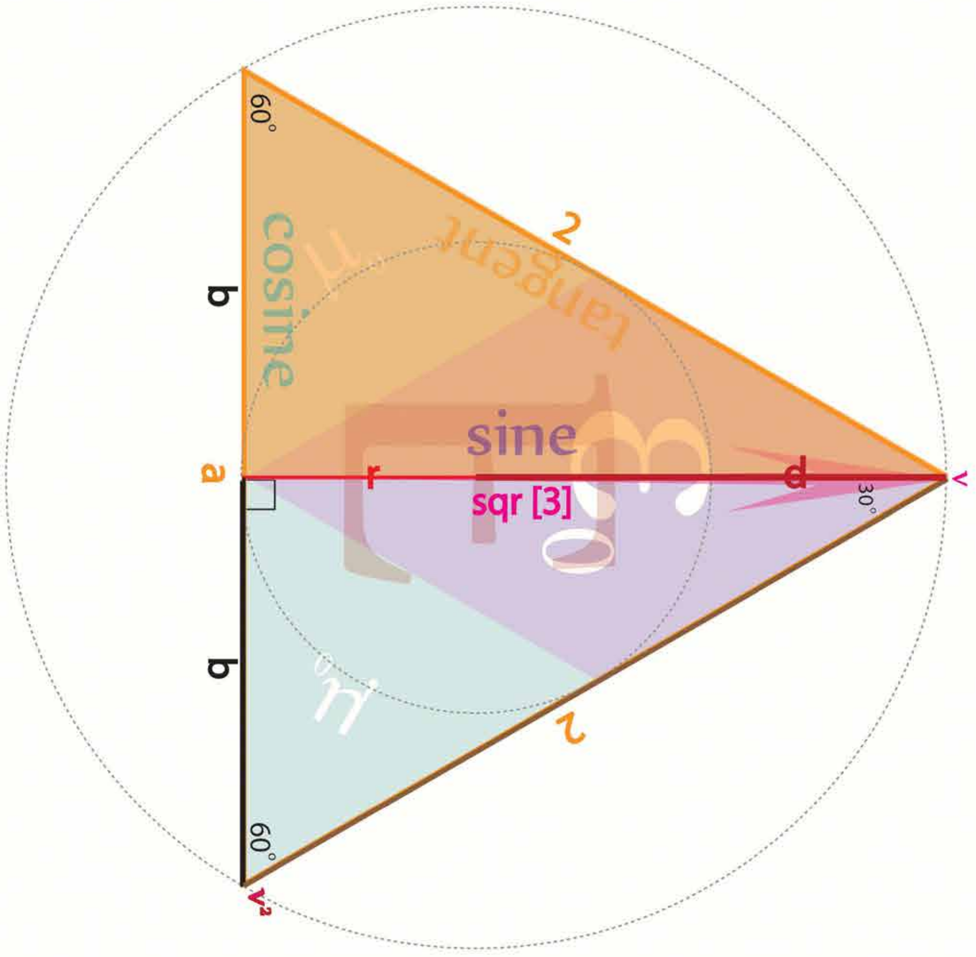
60°
cosine
tangent
2
b
a
r
sine
sqr [3]
d
v
30°
b
2
60°
v²

r

The ratio of the
circumscribed circle of
an equilateral triangle
to its inscribed circle
is 2:1

d

0.866 r

The ratio of the
side of an equilateral triangle
to the radius of
its incribed circle
is sqr [3]/2

a

The ratio of the
side of an equilateral triangle
to the radius of
its circumscribed circle
is sqr [3]

1.732 d

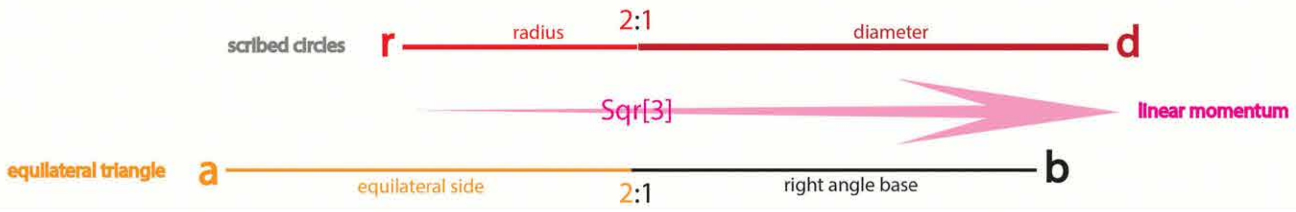
scribed circles
r
radius
2:1
diameter
d
Sqr[3]
linear momentum
equilateral triangle
a
equilateral side
2:1
right angle base
b

Equilateral triangles and scribed circles

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1-\beta^2}} = \frac{dt}{d\tau}$$

$$\gamma(\beta) = 1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \frac{35}{128}\beta^8 + \ldots$$

$$L = L'\sqrt{1 - \frac{v^2}{c^2}}$$

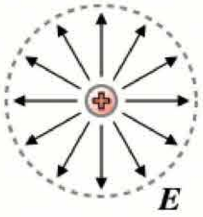

$$mc^2\left\{\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right\}$$

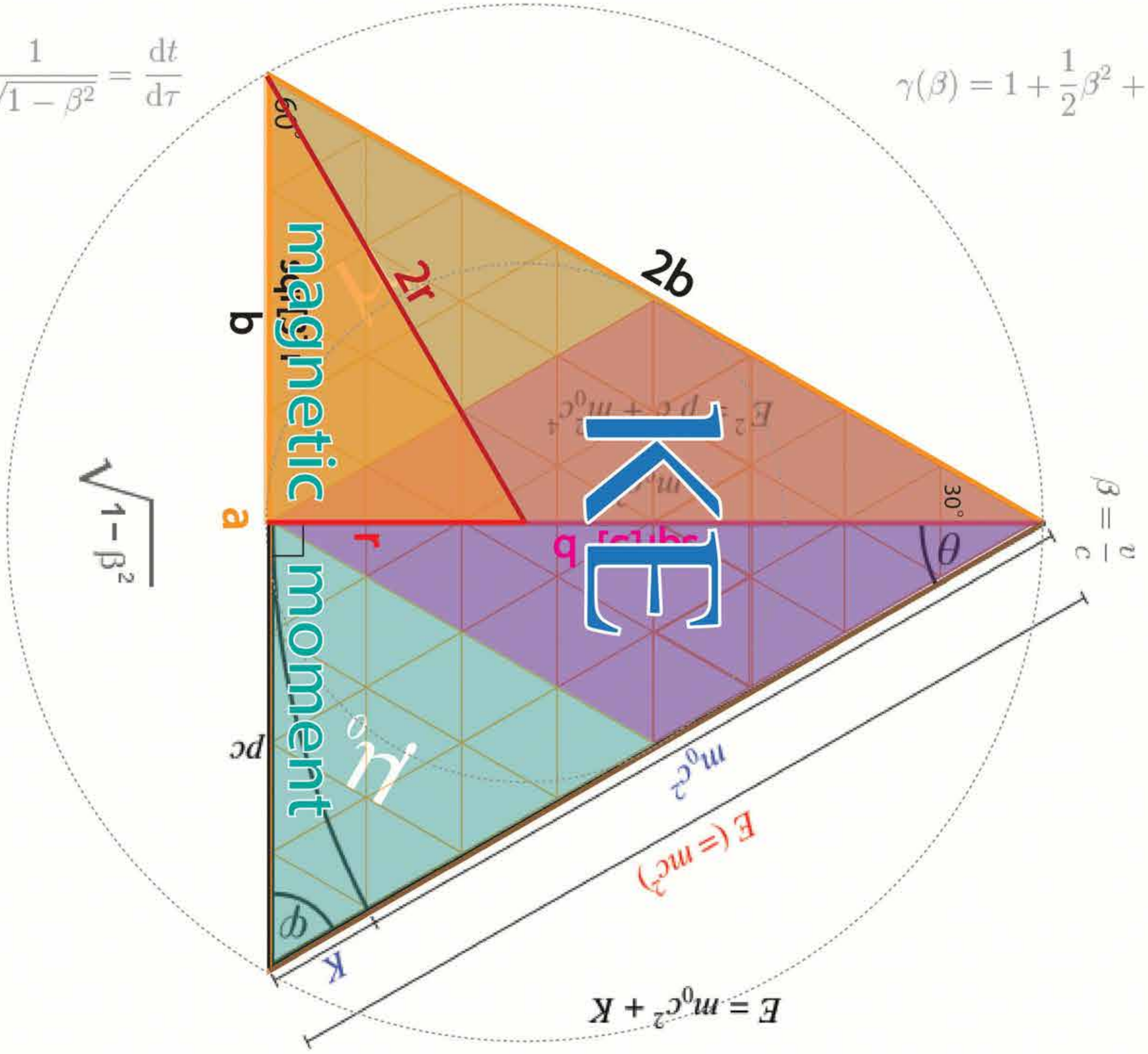
60°
2r
2b
b
magnetic
moment
KE
a
r
q
30°
θ
φ
pc
μ₀
m_0c^2
E (= mc²)
K
$E = m_0c^2 + K$
$E^2 = p^2c^2 + m_0^2c^4$
$\beta = \frac{v}{c}$
$\sqrt{1-\beta^2}$

$$E_k = \frac{mc^2}{\sqrt{1 - (v/c)^2}} - mc^2$$

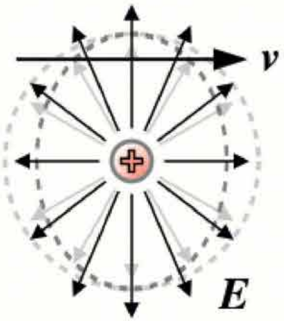

$$p = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma m_0 v$$

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\sqrt{1 - \left(\frac{v}{c}\right)^2}$$

All the relativistic relationships historically attributed to circularised energies and modelled through the use of the math of right angled triangles are in fact the result of equilateral, scalar geometries

Copyright ABRAHAM [2008] - All rights reserved

Finite sequences and series have defined first and last terms,

$$\sum a_n = a_0 + a_1 + a_2$$

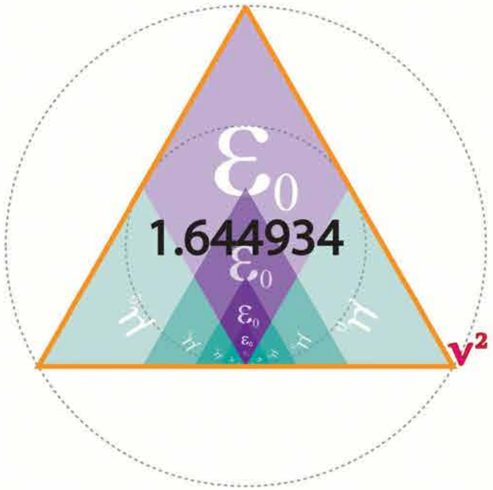
ε_0
1.644934
ε_0
ε_0
ε_0
π
π
v^2

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \to +\infty} \left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \right).$$

whereas infinite sequences and series continue indefinitely

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots = \frac{\pi^2}{6}$$

Tetryonics now provides a geometric solution to visualising and solving the Basel problem

$$\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6} \approx 1.645;$$

Tetryonic Infinite Series

is a infinite sequence of square numbers,
the result of adding all those terms together
[or their geometric inverse]

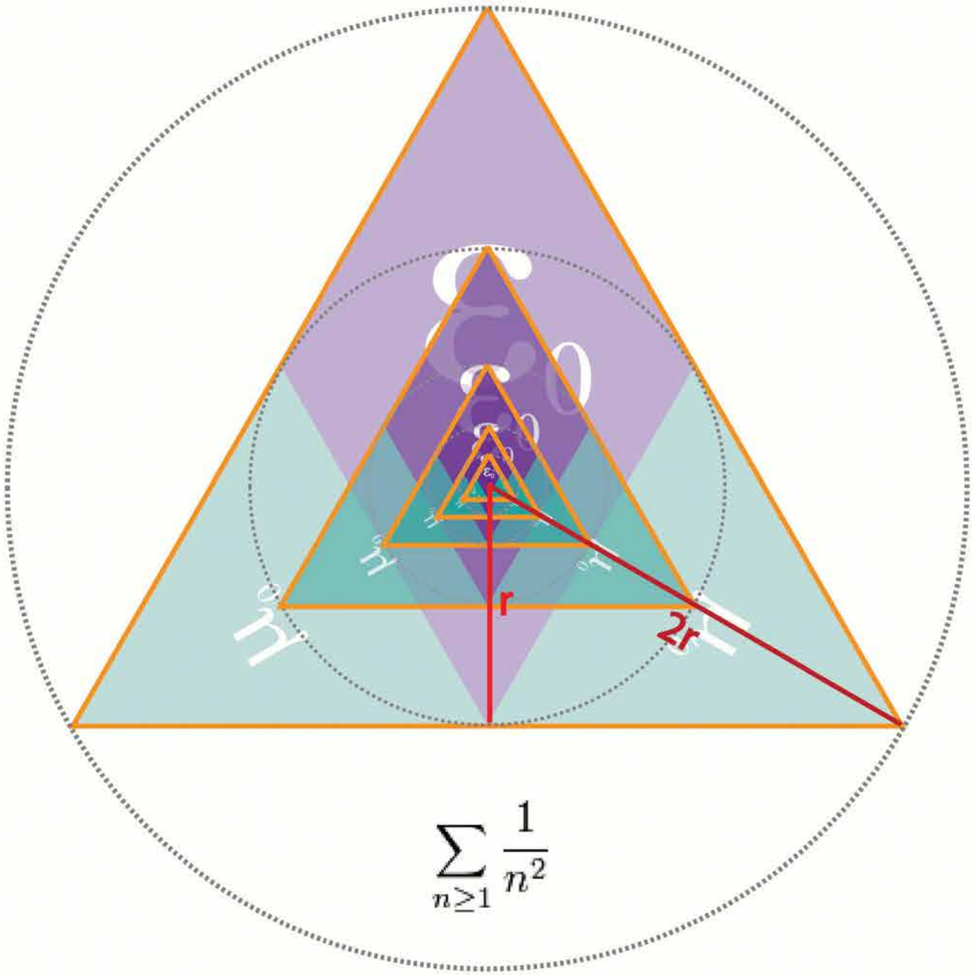
ε_0
ε_0
ε_0
π
π
π
π
r
2r

$$\sum_{n \geq 1} \frac{1}{n^2}$$

The entire sum of the series is equal to
twice the size of the radius of the largest inscribed circle
which is equal to the largest circle circumscribing the triangular series.

Copyright ABRAHAM [2008] - All rights reserved

Inverting the Circle

Electric
Permittivity

ε_o

Magnetic
Permeability

μ_o

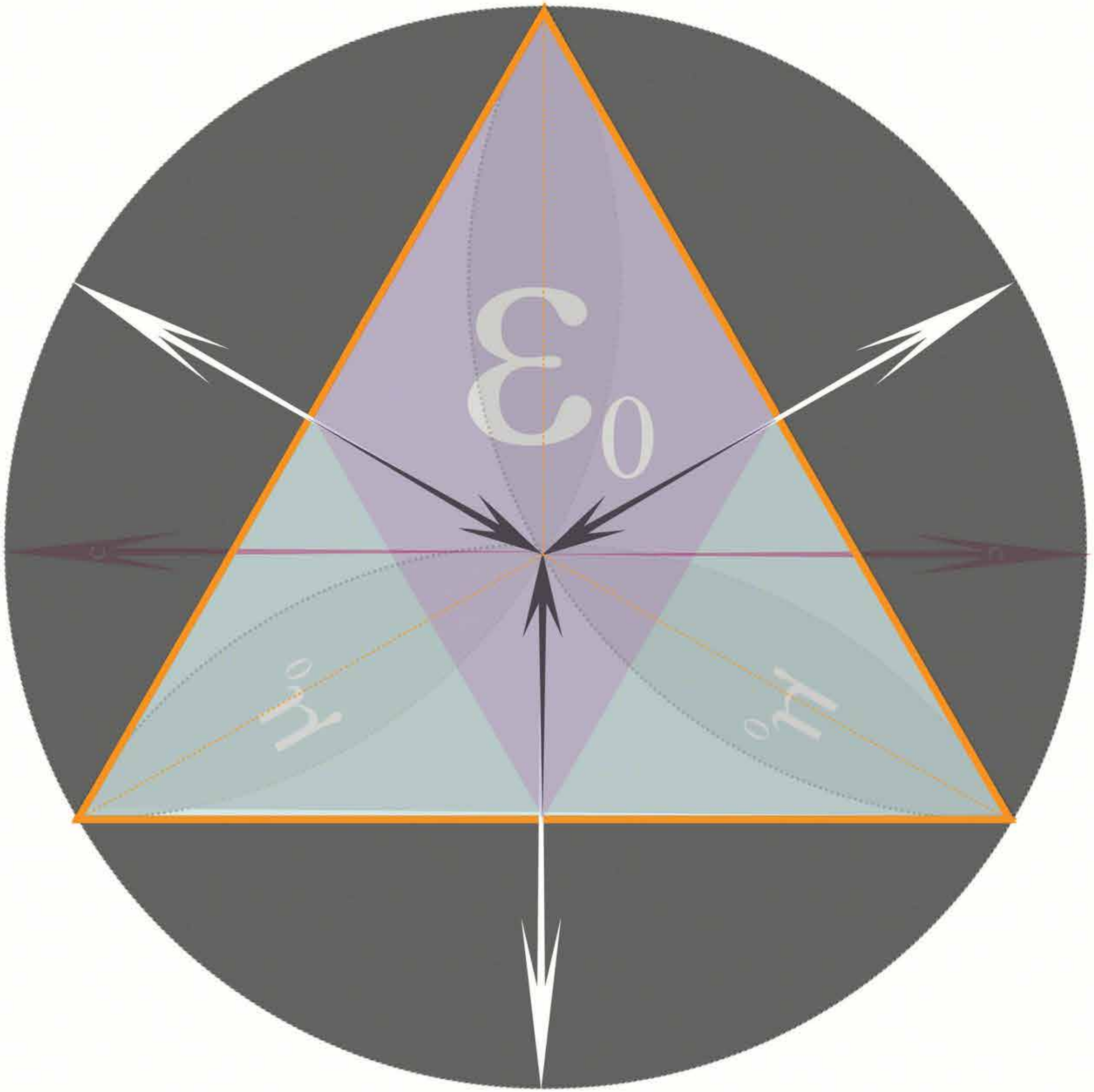

c^2

2D spatial
co-ordinates

$\frac{1}{c^2}$

ElectroMagnetic
fields

Copyright ABRAHAM [2008] - All rights reserved

Irrational numbers

an irrational number cannot be represented as a simple fraction.
Irrational numbers are those real numbers that cannot be represented as terminating or repeating decimals

sin(x)=-square root of (3)/2

0.866025403.......

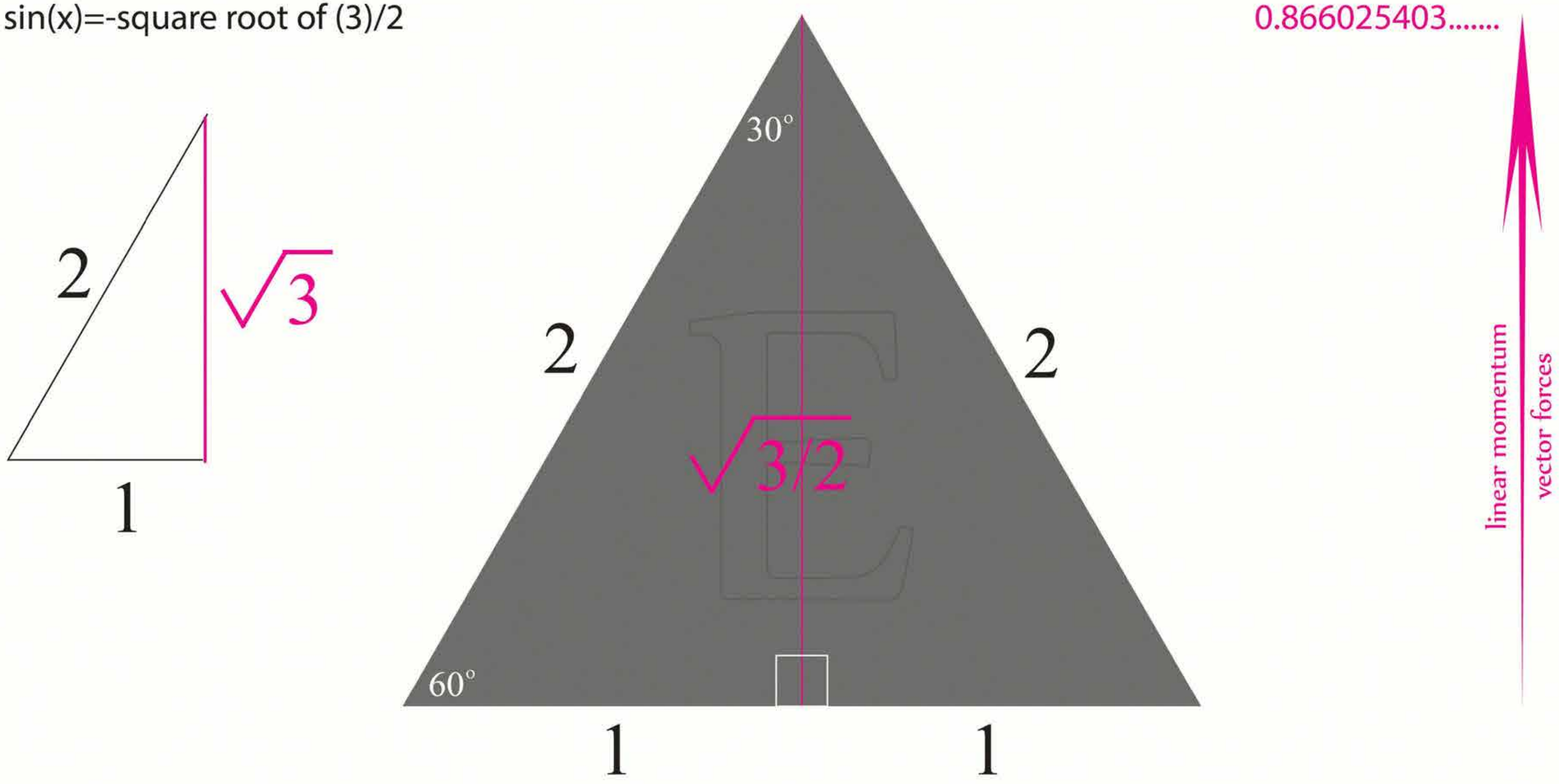
2
$\sqrt{3}$
1
30°
2
2
$\sqrt{3/2}$
60°
1
1
linear momentum
vector forces

Pythagφras' theoreom
and irratiφnal numbers expressed in terms of right-angled triangles in Physics
offer a 'half truth' regarding the equilateral geometry of Energy

Copyright ABRAHAM [2008] - All rights reserved

Leibniz

Leibniz introduced the concept of vis viva (living force) which attempted to quantify life force

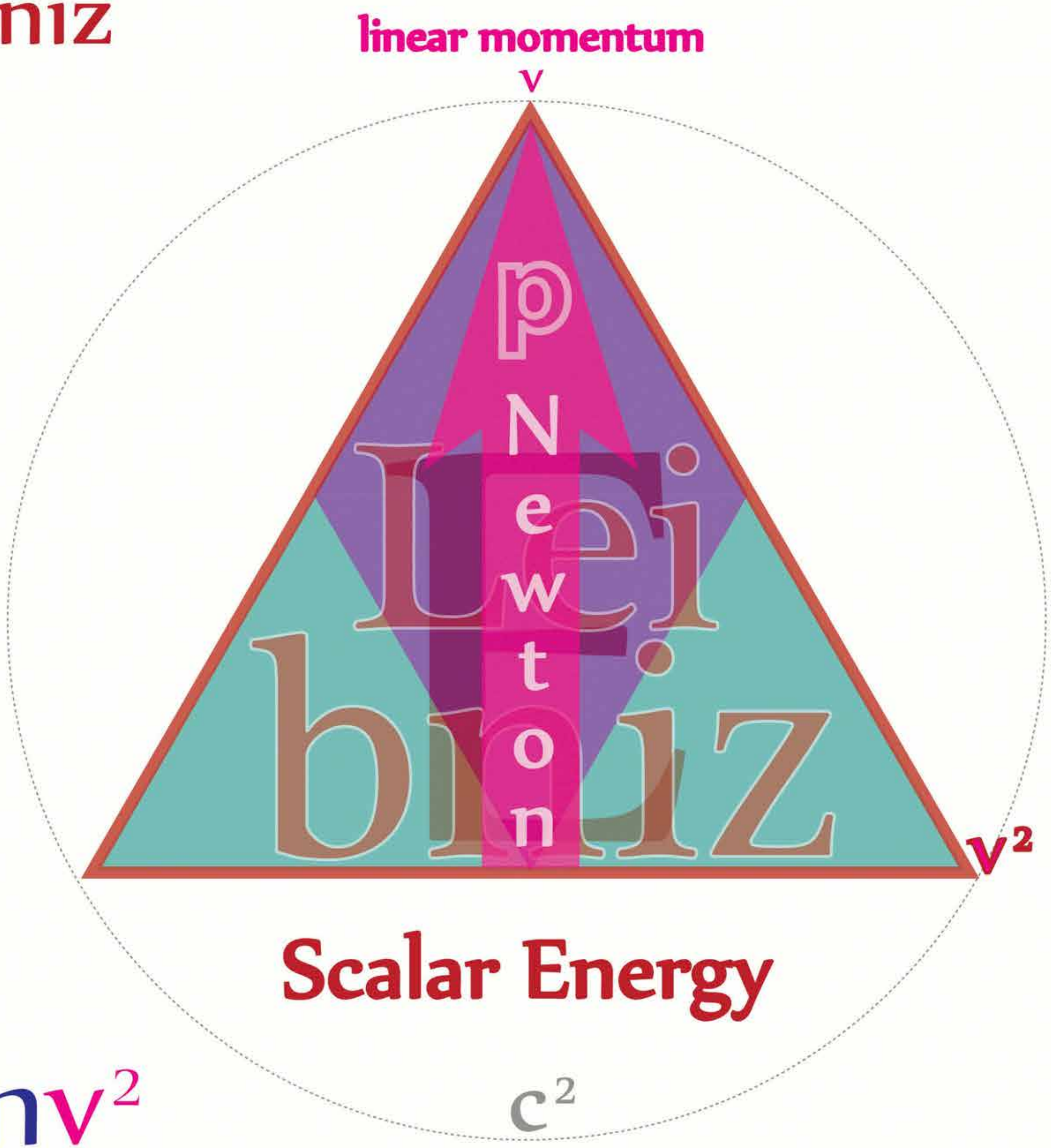

Newton

Newton focused his work on linear momentum which he developed into his famous laws of motion

Newton and Leibniz disagreed about what the world is made of and how its physics shaped our scientific concepts of force, energy, and momentum

Tetryonics reveals the physical relationships they both described mathematically as geometric properties of equilateral Planck energy momenta

$E=mv^2$

$p=mv$

Copyright ABRAHAM [2008] - All rights reserved

Geometric Square Roots

In geometrical terms, the square root function maps the area of a square to its side length.

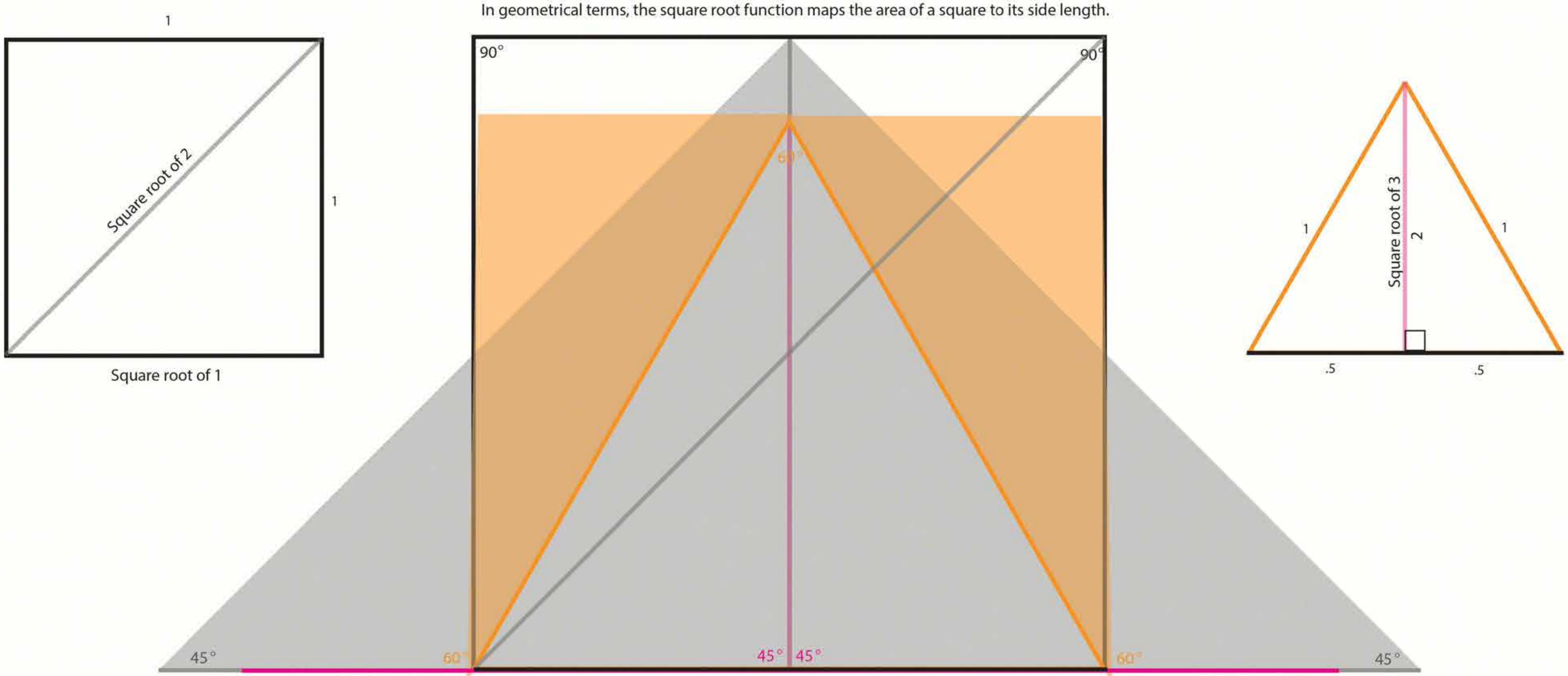

"In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward abstraction seeks to crystallize the logical relations inherent in the maze of material that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward intuitive understanding fosters a more immediate grasp of the objects one studies, a live rapport with them, so to speak, which stresses the concrete meaning of their relations"

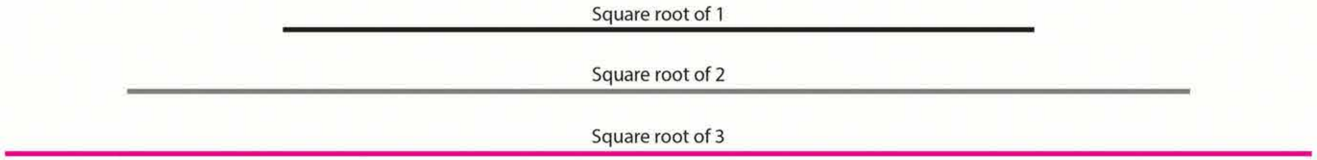

"As to geometry, in particular, the abstract tendency has here led to the magnificent systematic theories of Algebraic Geometry, of Riemannian Geometry, and of Topology; these theories make extensive use of abstract reasoning and symbolic calculation in the sense of algebra.

Notwithstanding this, it is still as true today as it ever was that intuitive understanding plays a major role in geometry. And such concrete intuition is of great value not only for the research worker, but also for anyone who wishes to study and appreciate the results of research in geometry"

David Hilbert [Geometry and the Imagination]

Copyright ABRAHAM [2008] - All rights reserved

Square Roots in Physics

In mathematics, a square root of a number a is a number [n] such that [n]2 = x,
or, in other words, a number [n] whose square (the result of multiplying the number by itself, or [n × n]) is x.

Modern calculators use the
Square Root Algorithm
to calculate the value

It is an approximate numerical value

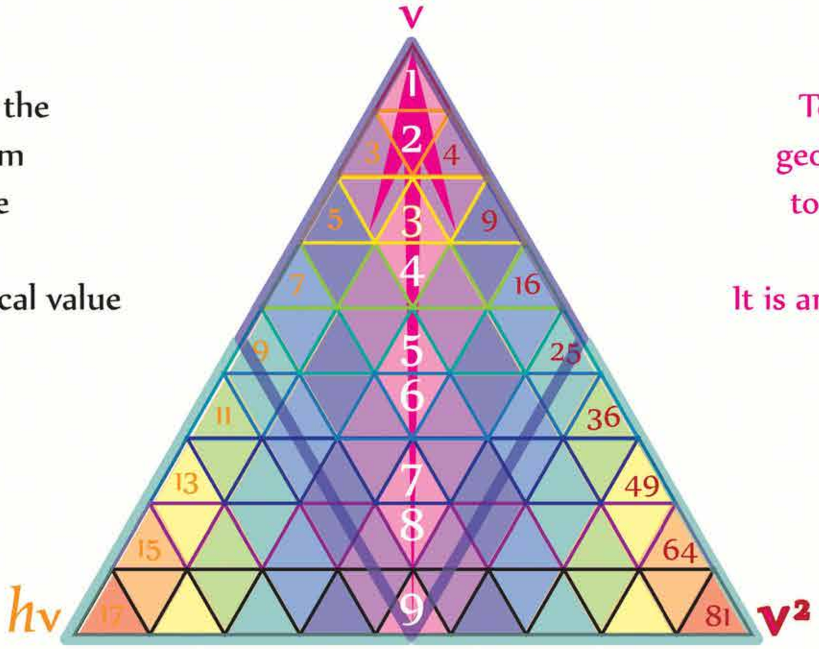

Tetryonics uses the
geometric Square Root
to calculate the value

It is an exact geometric value

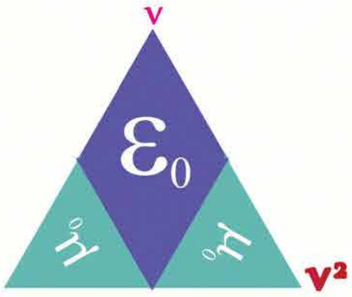

$$E = mv^2$$

In classical geometry,
the square root function maps the
area of a square to its side length.

$$p = \sqrt{E}$$

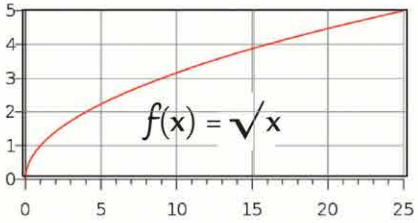

Geometry can easily map irrational numbers

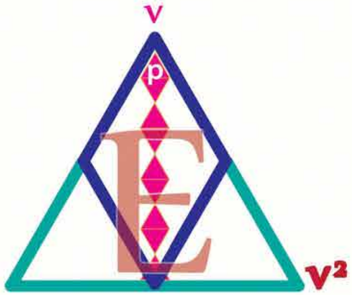

$$p^2 = E$$

In physics,
the square root function maps
ENERGY [E] to momentum [mv]

The Square roots of n

Historically, any number raised to the power of 2 has been modeled using a polygon--the square
That's why we call raising a number to the second power "squaring the number."

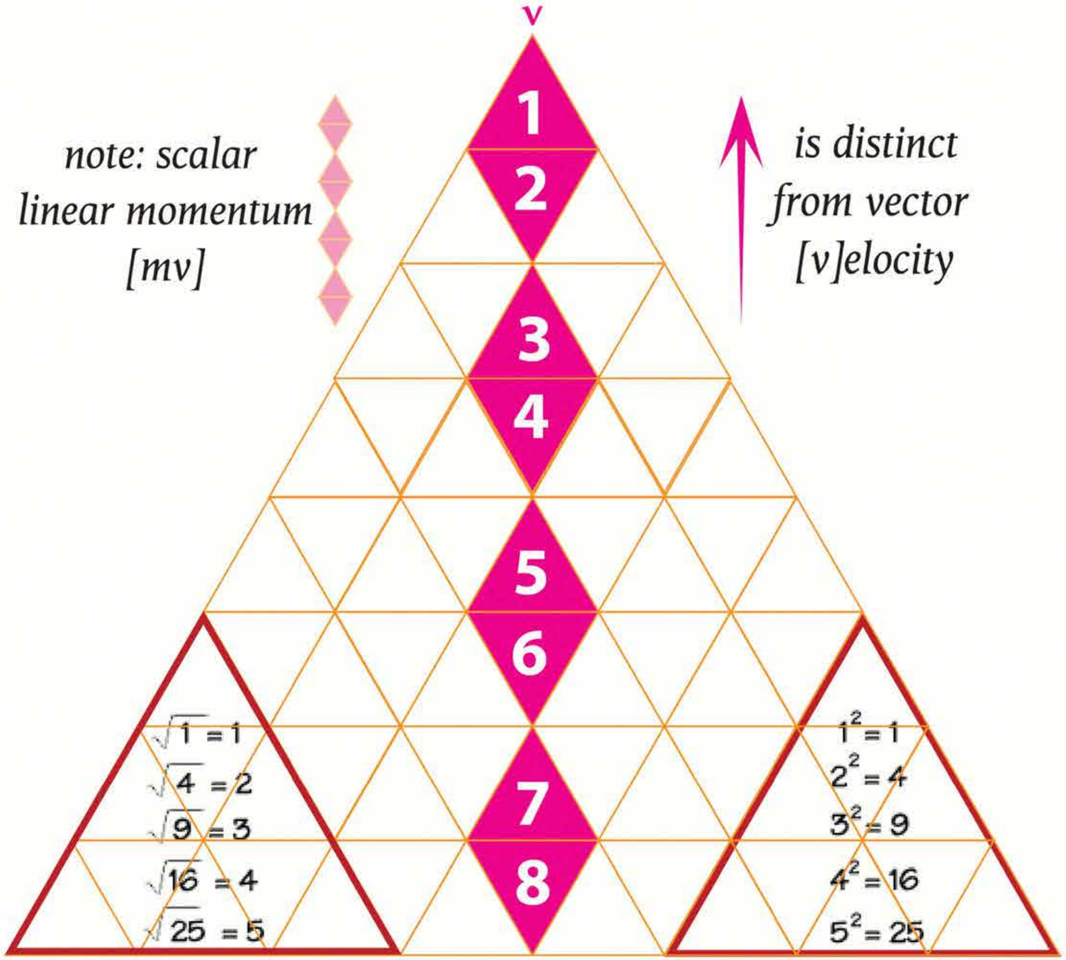

note: scalar linear momentum [mv]

is distinct from vector [v]elocity

$\sqrt{1}=1$
$\sqrt{4}=2$
$\sqrt{9}=3$
$\sqrt{16}=4$
$\sqrt{25}=5$

$1^2=1$
$2^2=4$
$3^2=9$
$4^2=16$
$5^2=25$

[In physics square numbers are in fact equilateral geometries]
The perfect squares are squares of whole numbers.
Here are the first eight perfect squares

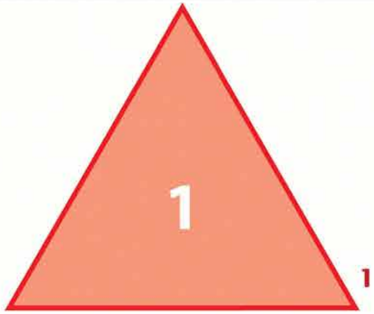

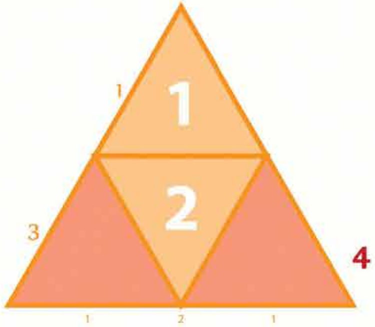

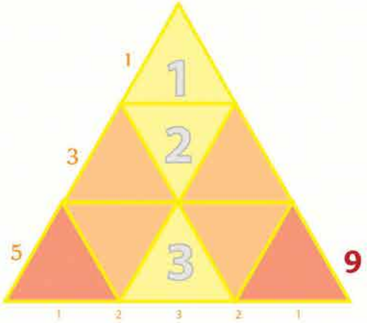
1
2
3
1
3
5
9
1 2 3 2 1

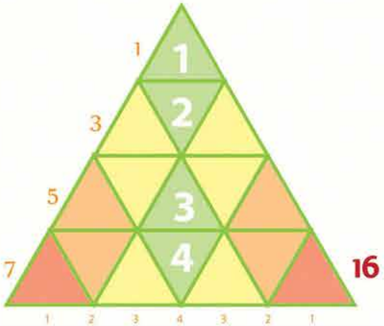
1
2
3
4
1
3
5
7
16
1 2 3 4 3 2 1

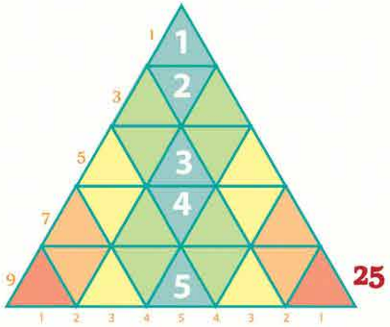
1
2
3
4
5
1
3
5
7
9
25
1 2 3 4 5 4 3 2 1

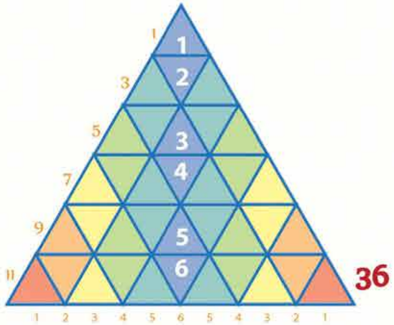
1
2
3
4
5
6
1
3
5
7
9
11
36
1 2 3 4 5 6 5 4 3 2 1

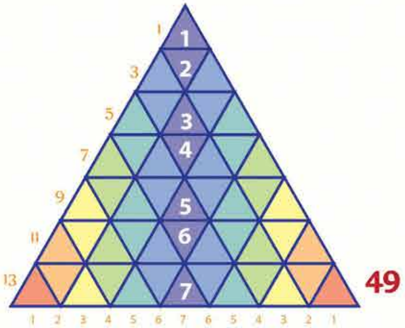
1
2
3
4
5
6
7
1
3
5
7
9
11
13
49
1 2 3 4 5 6 7 6 5 4 3 2 1

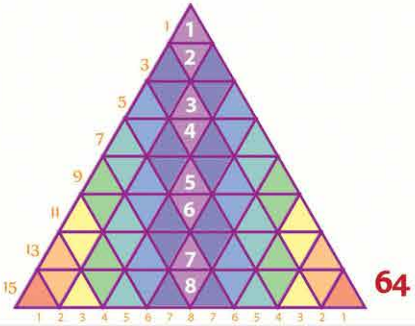
1
2
3
4
5
6
7
8
1
3
5
7
9
11
13
15
64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

The Square root of Negative 1

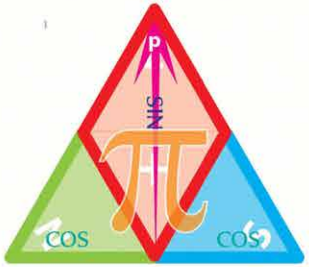

Positive fields are out of phase with Negative fields

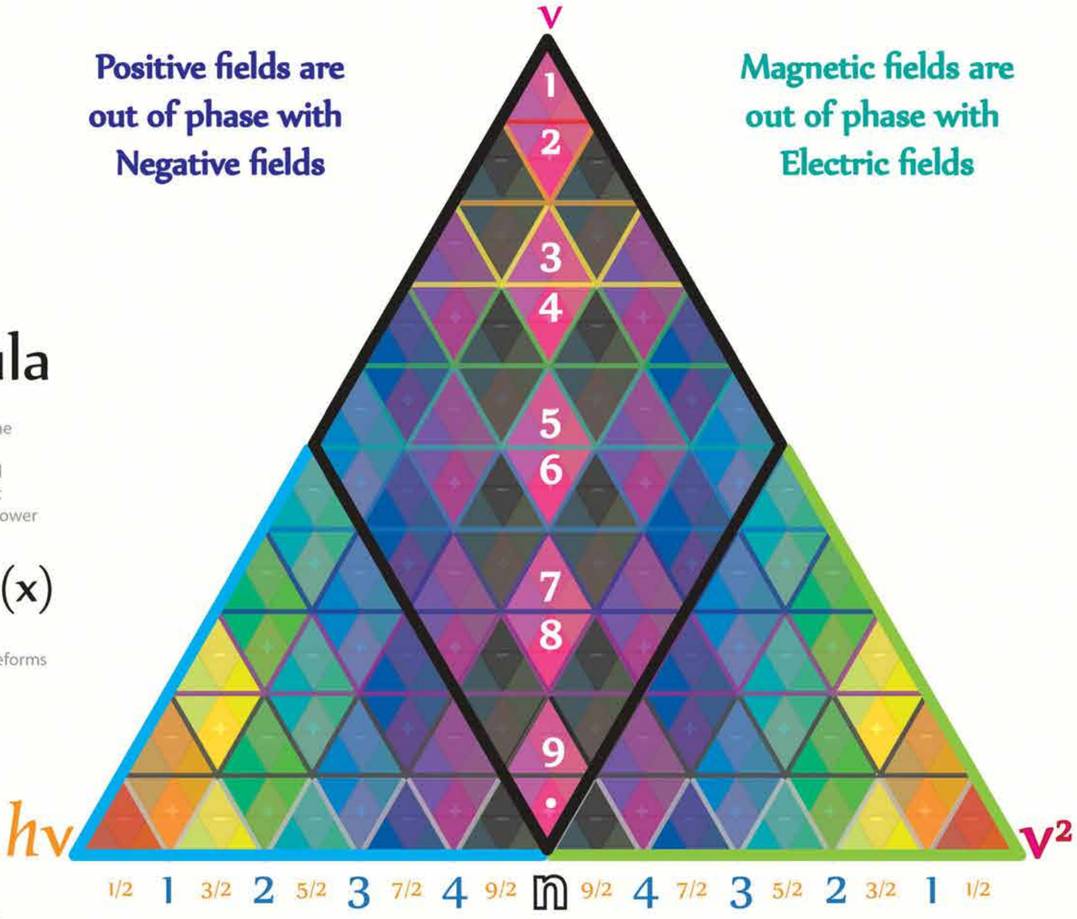

Magnetic fields are out of phase with Electric fields

p
COS
SIN
SIN

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system.
In deriving this formula, Euler established a relationship between the trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

Leonhard Euler

(15 April 1707 – 18 September 1783)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

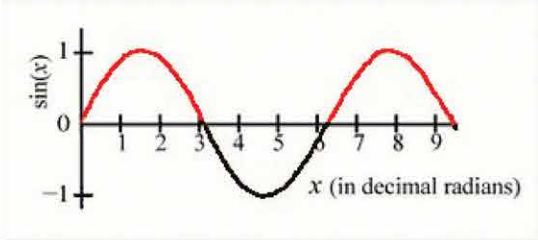

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

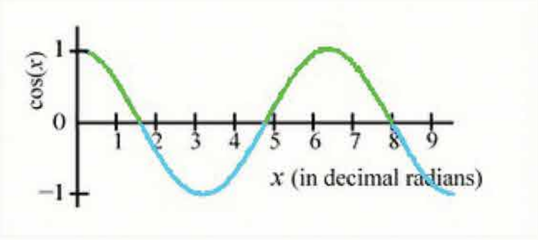

Geometric means

The geometric mean of two numbers, is the square root of their product

geometric square root
of positive one

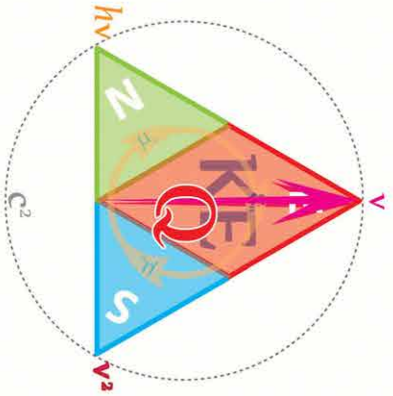

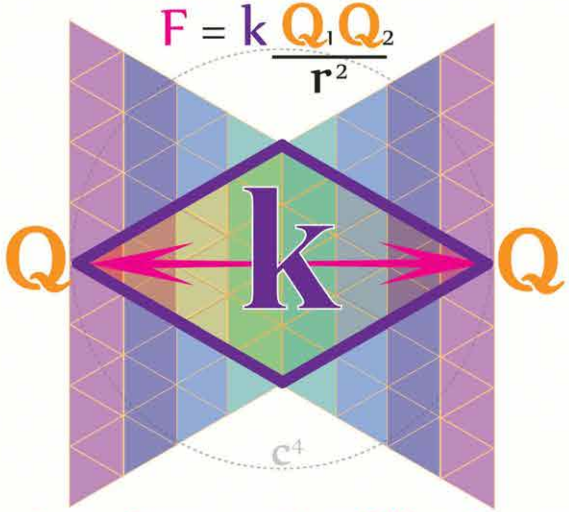

$$F = k\frac{Q_1 Q_2}{r^2}$$

geometric square root
of negative one

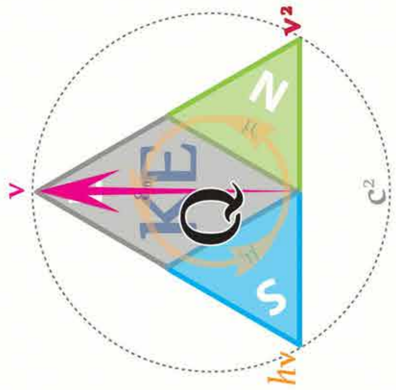

In physics, the geometric mean of two superpositioned fields produces a vector square root Force

It is generally stated that
the geometric mean applies
only to positive numbers.

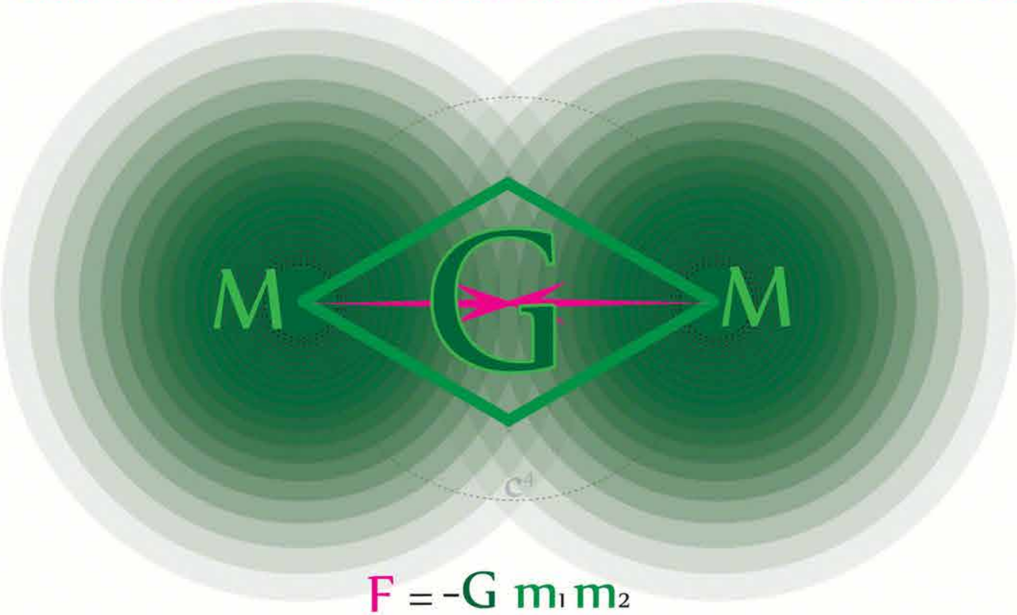

In Tetryonic geometry
the geometric mean applies
to positive & negative fields.

$$F = -G\frac{m_1 m_2}{r^2}$$

Superpositioning

When two or more waves traverse the same space,
the net amplitude at each point is the
sum of the amplitudes of the individual waves.

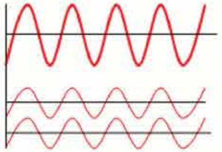

constructive interference

In phase

Additive

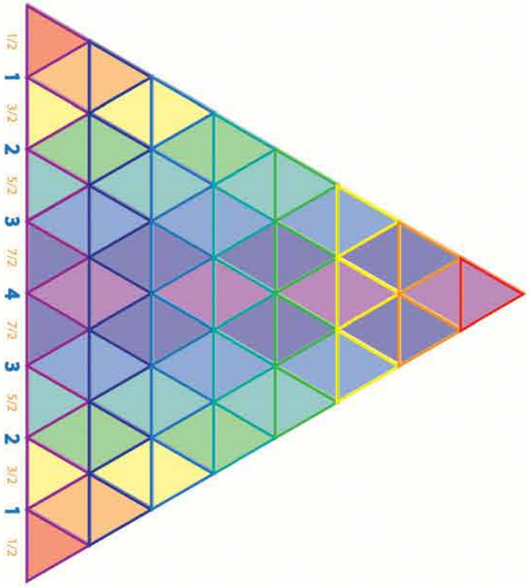
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

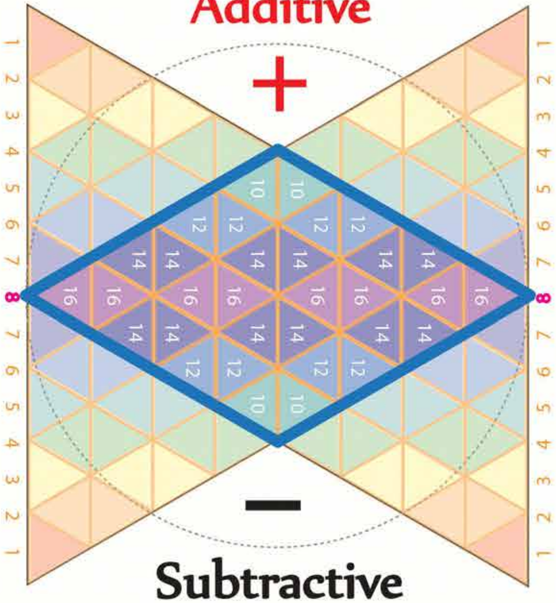
+
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
10
12
14
16
−

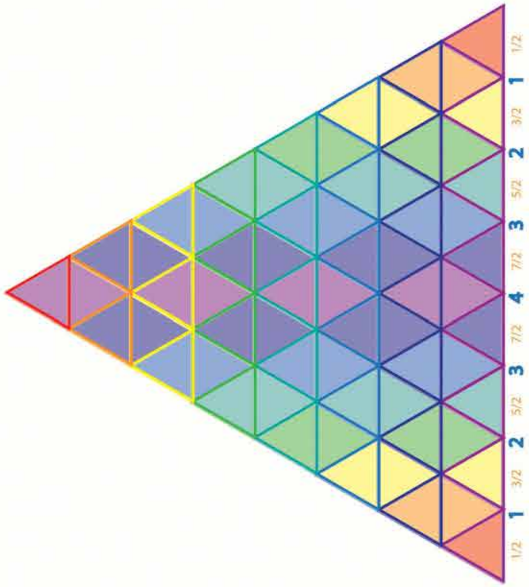
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Subtractive

Out of phase

destructive interference

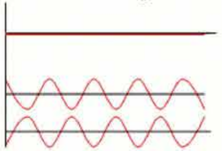

When two or more waves traverse the same space,
if the summed variation has a smaller amplitude
than the constiuent component variations.

The lines of Force

Tetryonic Multiplication table

A multiplication table is a mathematical grid used to define a multiplication operation and its results

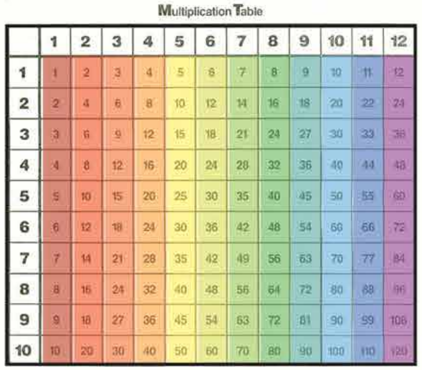

Multiplication Table

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120

Historically Multiplication tables
have been based on Square geometries

In Tetryonics multiplication tables can
also be based on EQUILATERAL geometries

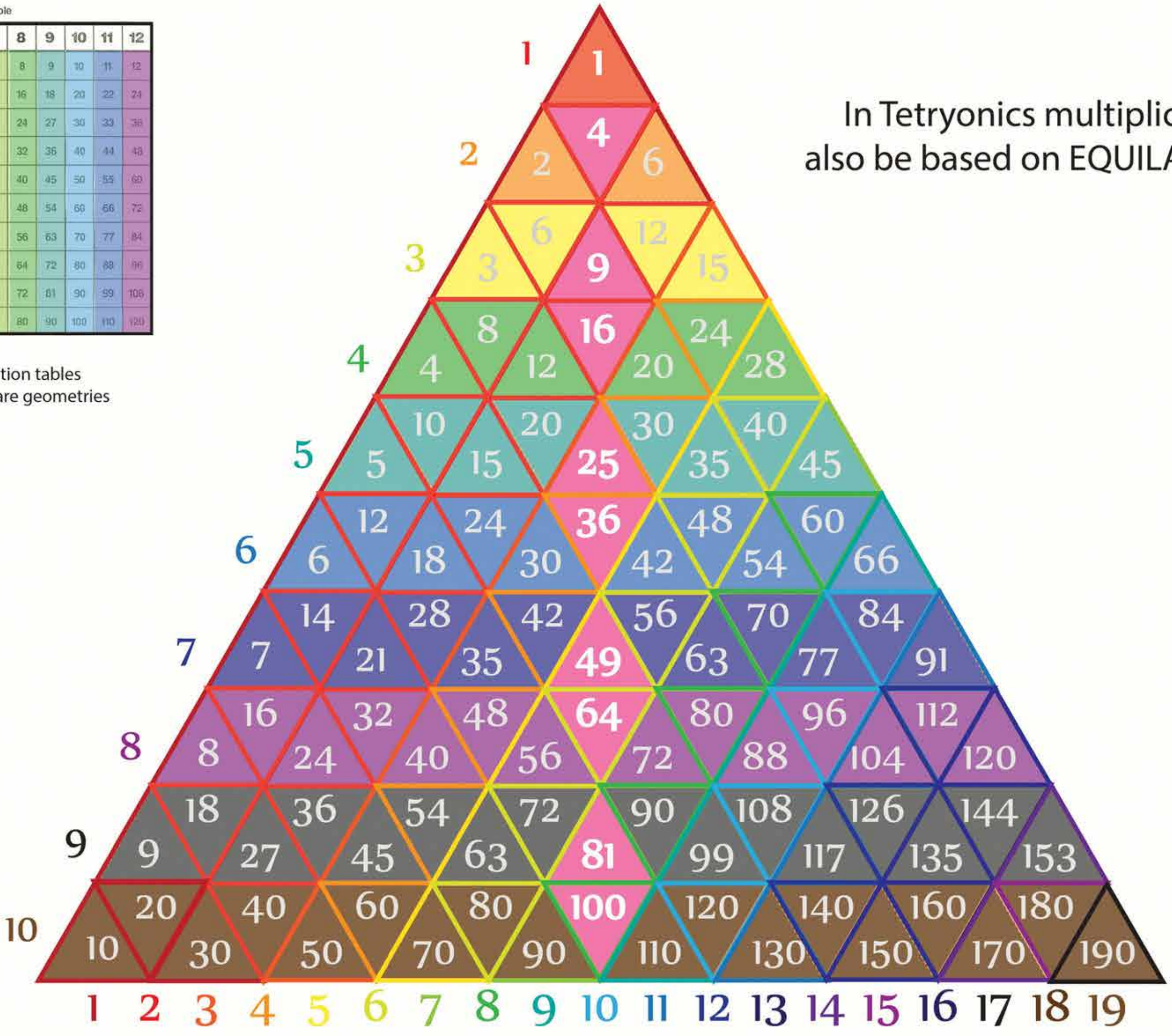

1
2 4 6
3 6 9 12 15
4 8 12 16 20 24 28
5 10 15 20 25 30 35 40 45
6 12 18 24 30 36 42 48 54 60 66
7 14 21 28 35 42 49 56 63 70 77 84 91
8 16 24 32 40 48 56 64 72 80 88 96 104 112 120
9 18 27 36 45 54 63 72 81 90 99 108 117 126 135 144 153
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190

The integer multiplicitors are colour coded

Copyright ABRAHAM [2008] - All rights reserved

Rhombic Multiplication Tables

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

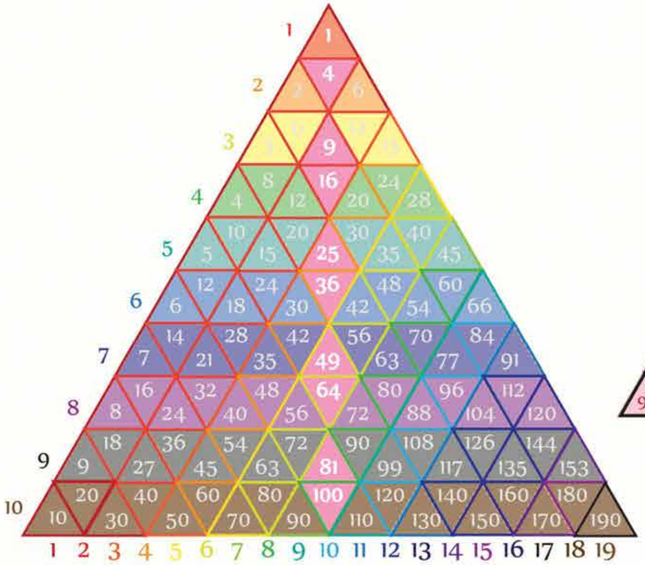

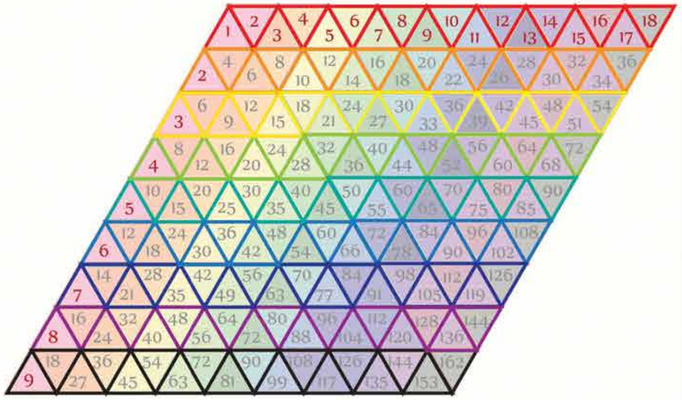

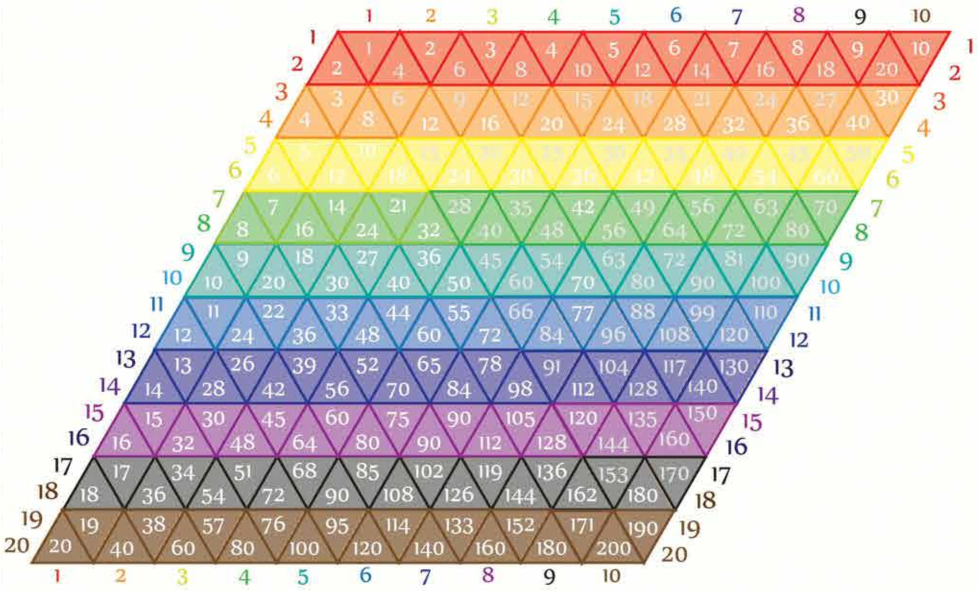

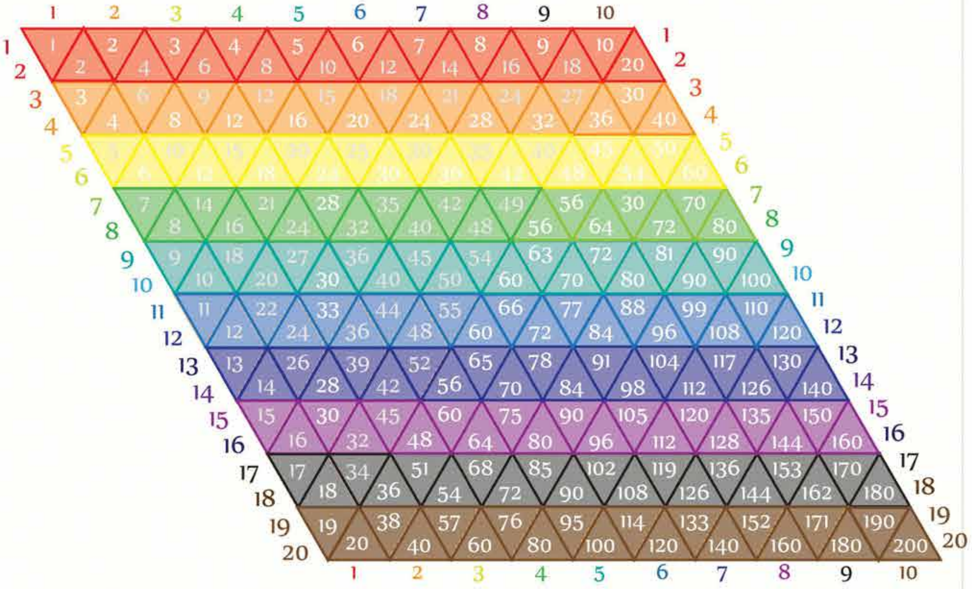

Tetryonics 85.03 - Rhombic multiplication tables

Copyright ABRAHAM [2008] - All rights reserved

Photonic Root Tables

Tetryonic multiplication table can take a number of geometric forms

Square root median

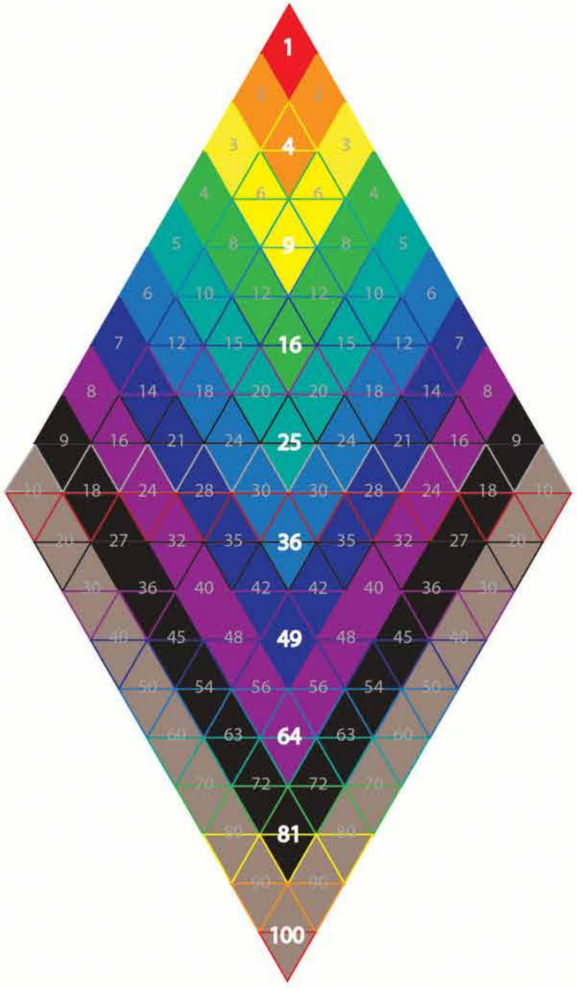
1
4
9
16
25
36
49
64
81
100

Table read diagonally

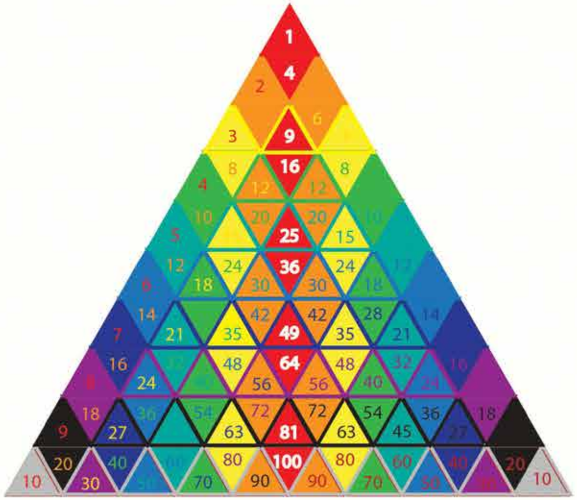
1
4
9
16
25
36
49
64
81
100

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

Integer median

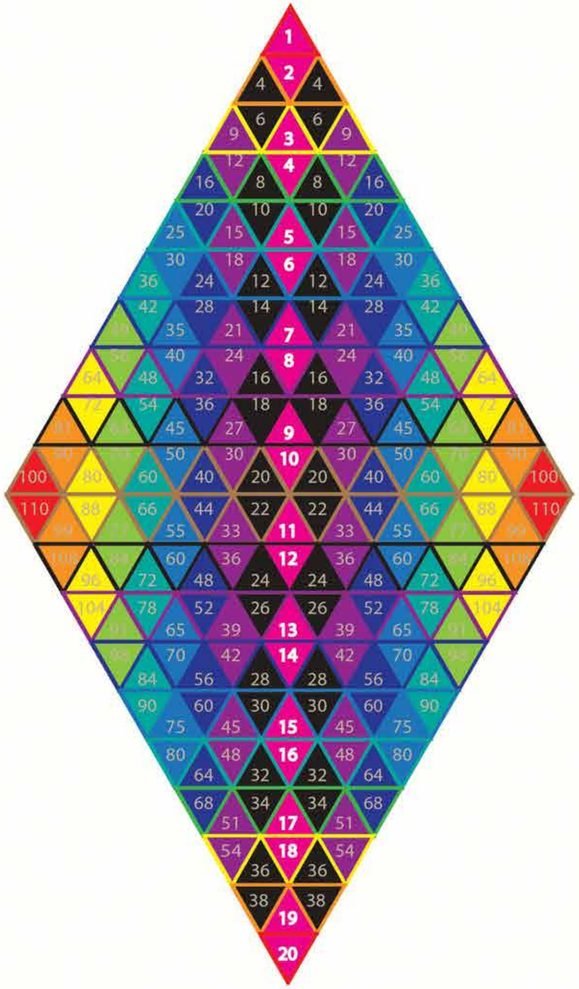
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Table read from centre
to outside edge then down

Copyright ABRAHAM [2008] - All rights reserved

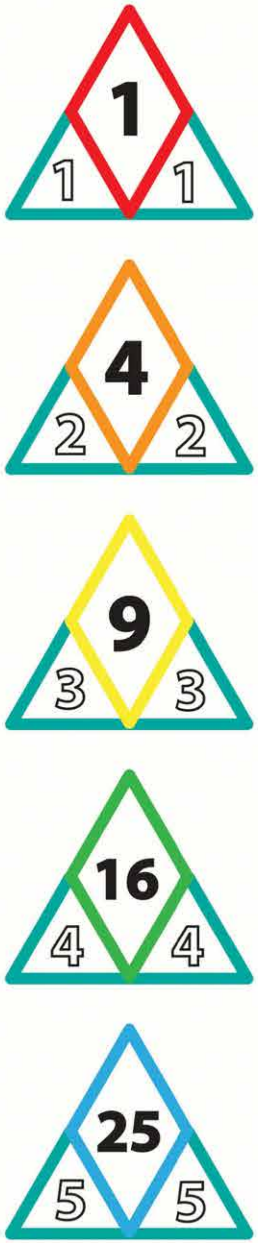
1
1 1
4
2 2
9
3 3
16
4 4
25
5 5

Divide [division] tables

Minus [subtraction] tables

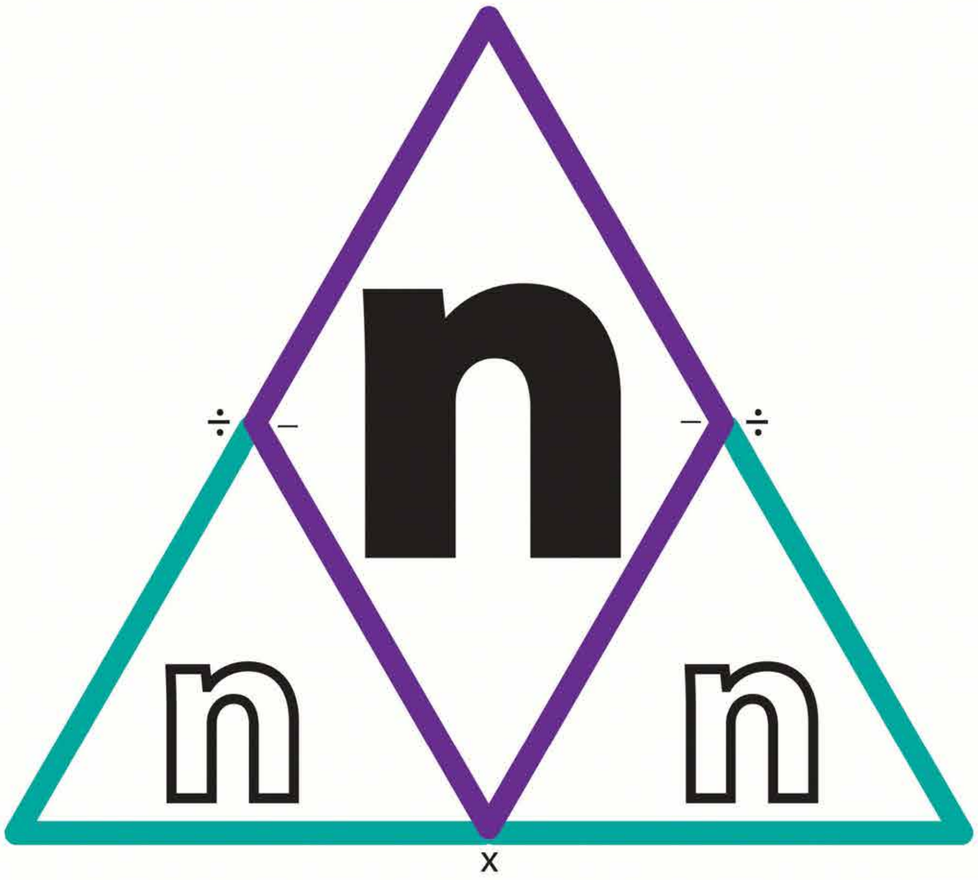

Times [multiplication] tables

Plus [addition] tables

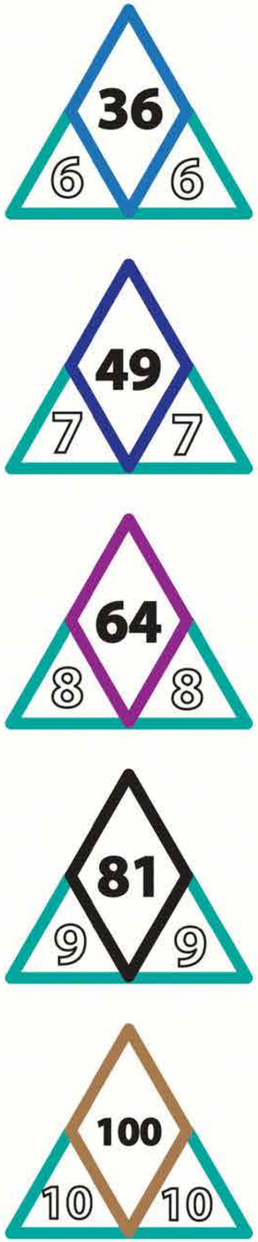
36
6 6
49
7 7
64
8 8
81
9 9
100
10 10

Copyright ABRAHAM [2008] - All rights reserved

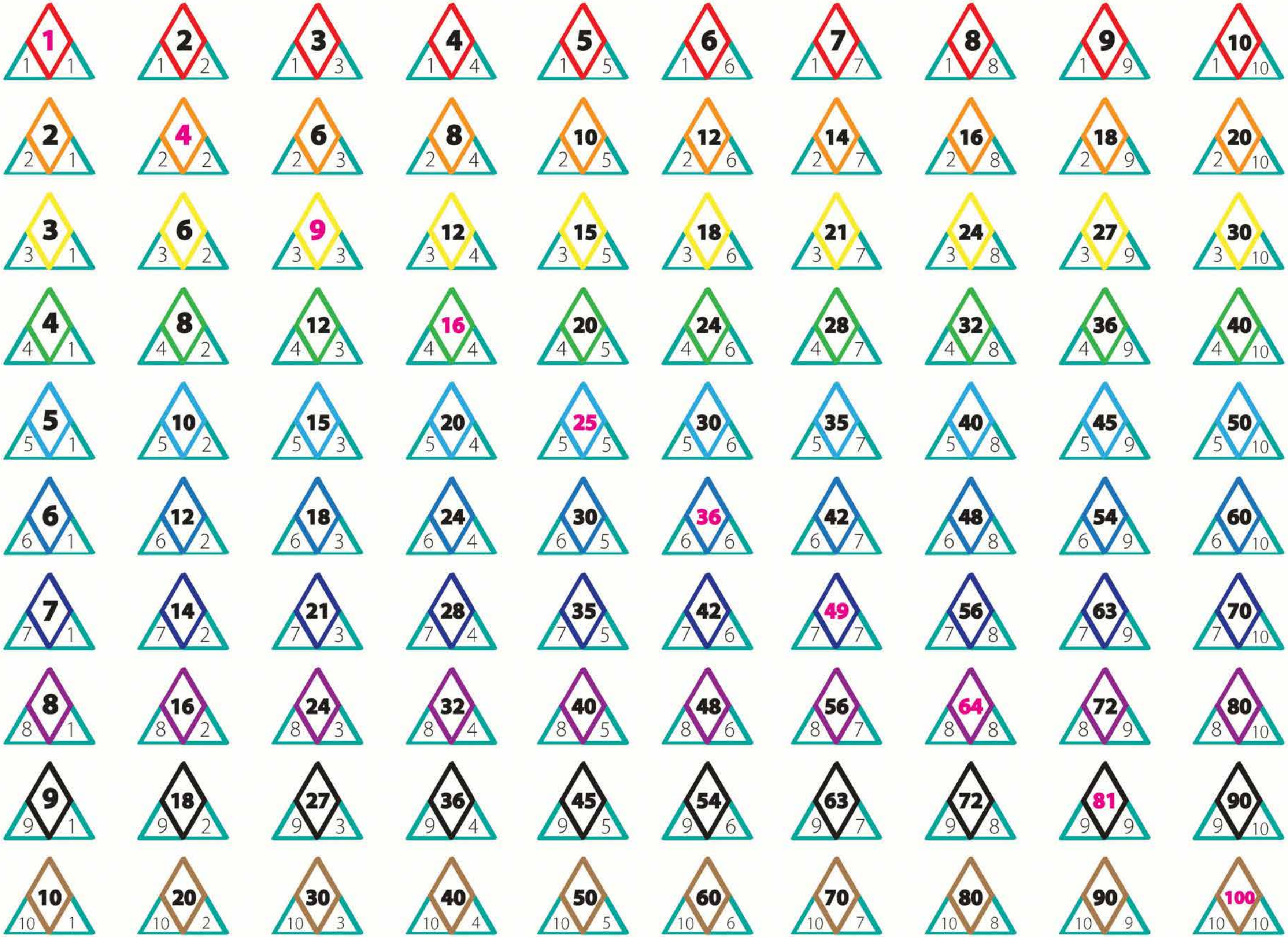

1 1 1	2 1 2	3 1 3	4 1 4	5 1 5	6 1 6	7 1 7	8 1 8	9 1 9	10 1 10
2 2 1	4 2 2	6 2 3	8 2 4	10 2 5	12 2 6	14 2 7	16 2 8	18 2 9	20 2 10
3 3 1	6 3 2	9 3 3	12 3 4	15 3 5	18 3 6	21 3 7	24 3 8	27 3 9	30 3 10
4 4 1	8 4 2	12 4 3	16 4 4	20 4 5	24 4 6	28 4 7	32 4 8	36 4 9	40 4 10
5 5 1	10 5 2	15 5 3	20 5 4	25 5 5	30 5 6	35 5 7	40 5 8	45 5 9	50 5 10
6 6 1	12 6 2	18 6 3	24 6 4	30 6 5	36 6 6	42 6 7	48 6 8	54 6 9	60 6 10
7 7 1	14 7 2	21 7 3	28 7 4	35 7 5	42 7 6	49 7 7	56 7 8	63 7 9	70 7 10
8 8 1	16 8 2	24 8 3	32 8 4	40 8 5	48 8 6	56 8 7	64 8 8	72 8 9	80 8 10
9 9 1	18 9 2	27 9 3	36 9 4	45 9 5	54 9 6	63 9 7	72 9 8	81 9 9	90 9 10
10 10 1	20 10 2	30 10 3	40 10 4	50 10 5	60 10 6	70 10 7	80 10 8	90 10 9	100 10 10

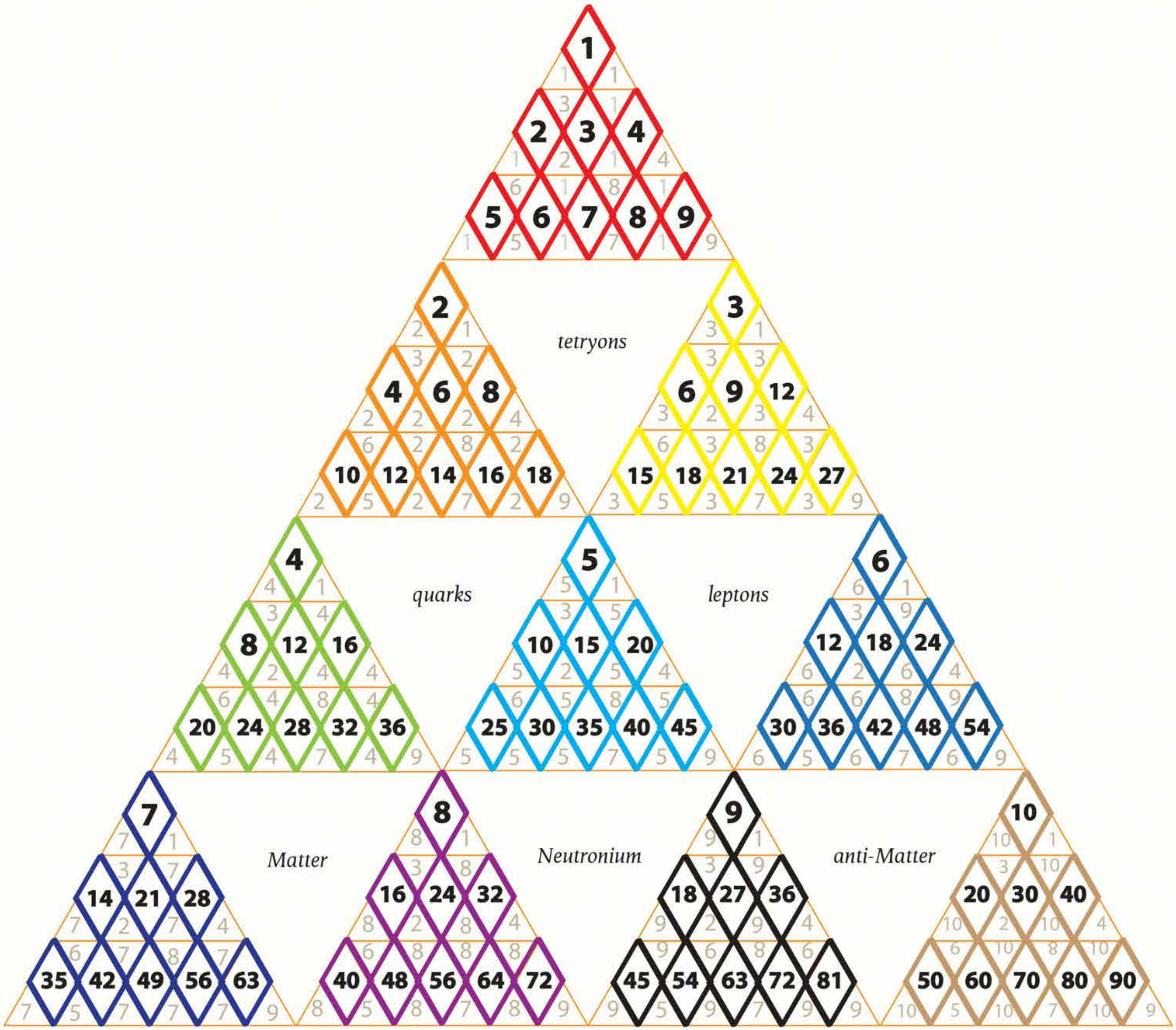
tetryons
quarks
leptons
Matter
Neutronium
anti-Matter

Copyright ABRAHAM [2008] - All rights reserved

Irrational Numbers

An irrational number is defined to be any number that is the part of the real number system that cannot be written as a complete ratio of two integers

One well known irrational number is π

Irrational numbers can be easily represented geometrically

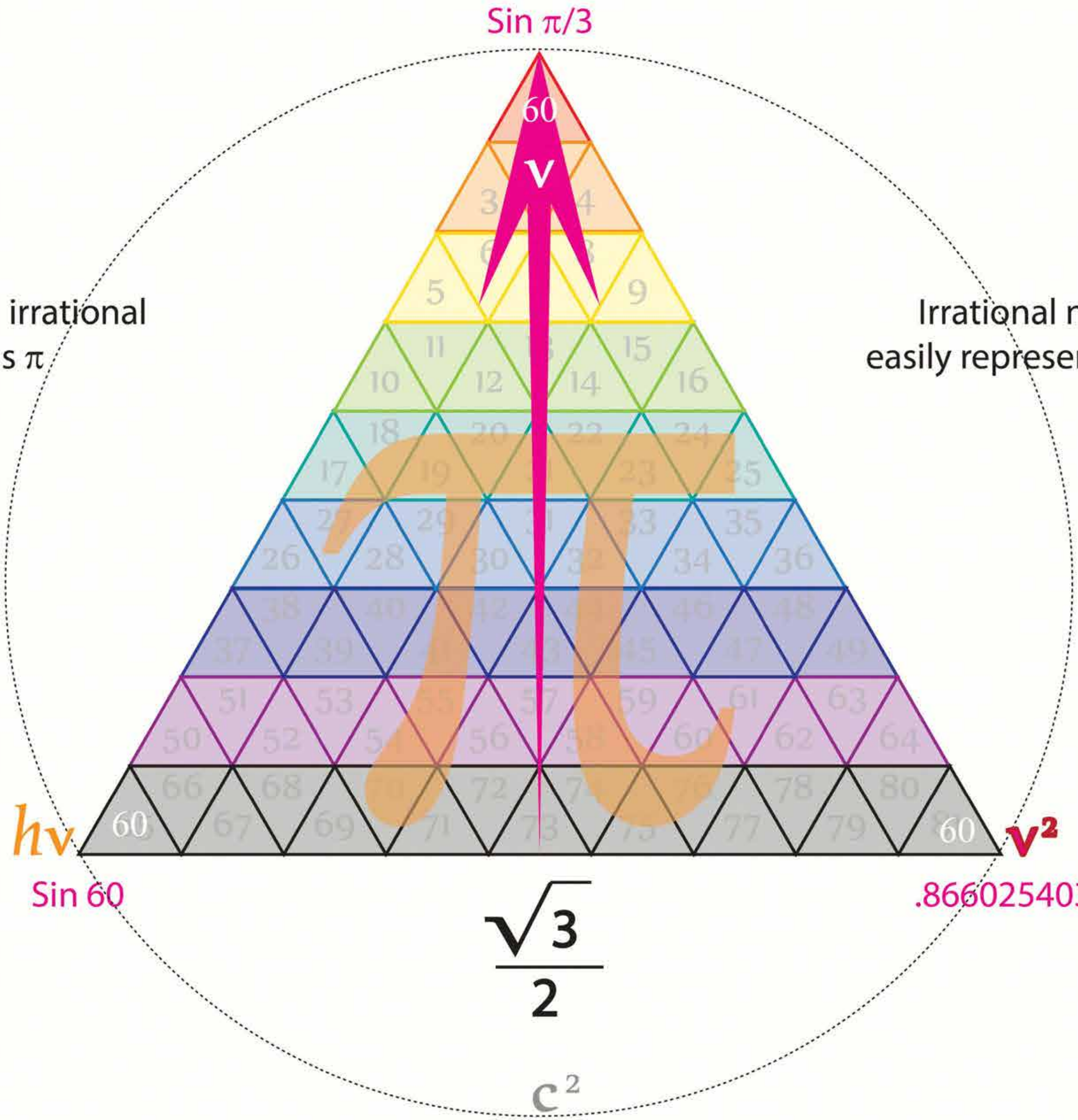
Sin π/3
60
v
3
4
5
9
10
11
12
14
15
16
17
18
20
22
24
25
26
28
29
30
33
34
35
36
38
40
42
46
48
50
51
52
53
56
57
59
60
61
62
63
64
66
67
68
69
71
72
73
77
78
79
80
π
hv
60
Sin 60
60
v²
.866025403
√3
2
c²

Exponentials & Logarithms

e and the Natural Log are inverse functions of each other:

e^x is the amount of continuous GROWTH after a certain amount of time.
Natural Log (ln) is the amount of TIME of continuous growth to reach a certain level

1	0
10	1
100	2
1,000	3

$10{,}000 = 10^4$

$\log(10{,}000) = 4$

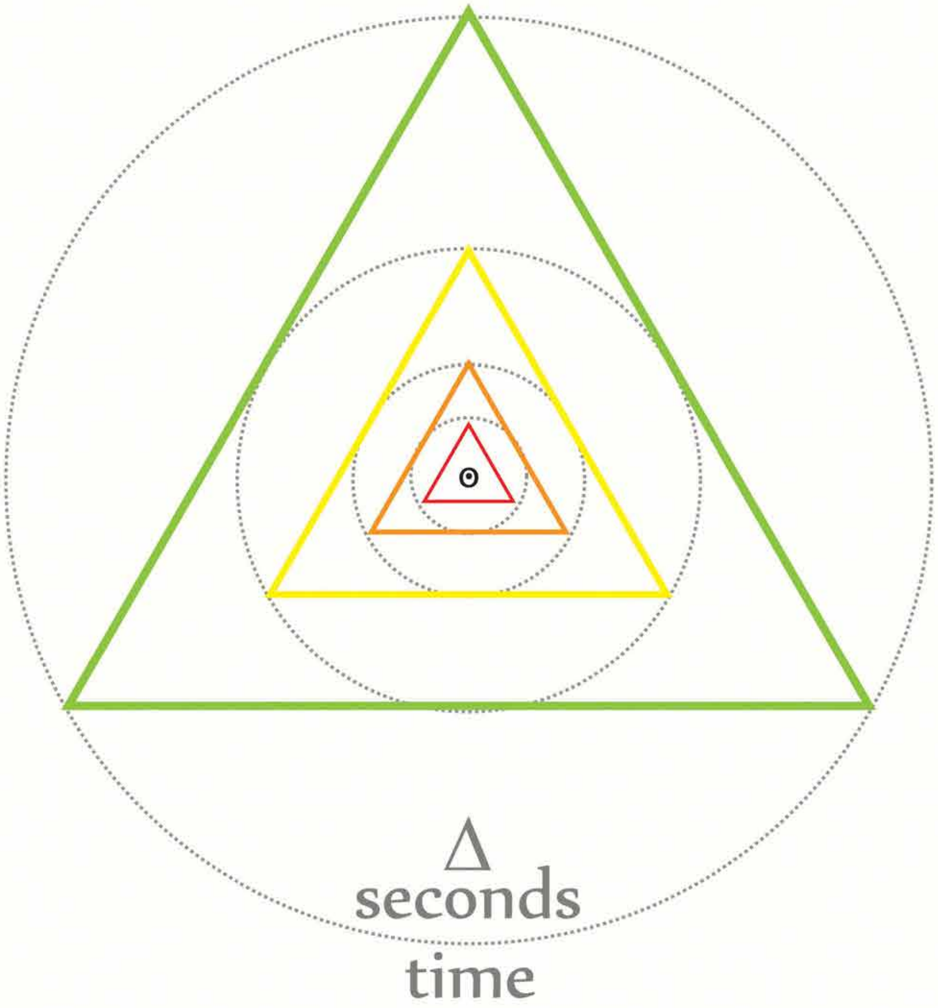

e

How much growth after x units of time
(and 100% continuous growth)

$$e^x = \lim_{n\to\infty} (1 + x/n)^n,$$

GROWTH

ln

ln(x) lets us plug in continuous growth
and get the time it would take.

$$\ln(x) = \lim_{n\to\infty} n(x^{1/n} - 1).$$

PERIOD

Copyright ABRAHAM [2008] - All rights reserved

Exponential growth

$$\text{growth} = e = \lim_{n\to\infty}\left(1 + \frac{r}{n}\right)^{t}$$

e represents the idea that all continually growing systems
are scaled versions of a common rate

GEOMETRIC

3.141592654

π

Pi is the ratio between
circumference and diameter
shared by all circles.

It is a fundamental ratio inherent in all circles
and therefore impacts any calculation of
circumference, area, volume, and surface areas

Pi radians are equally important and show
all quantised equilateral energy geometries
are related to their scribed circles

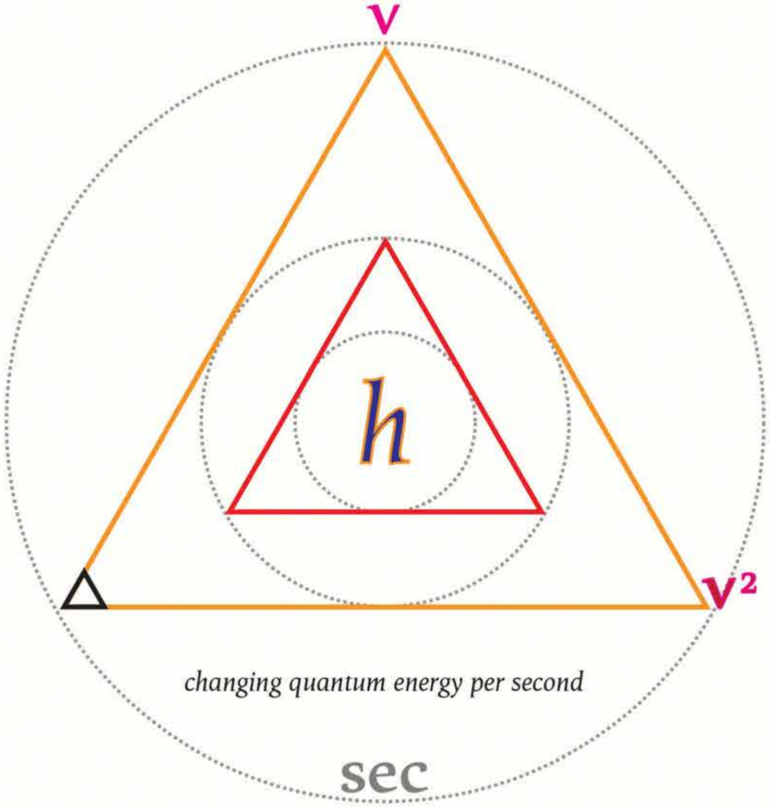

GROWTH

2.718281828

e

e is the base rate of growth
shared by all continually growing processes.

e lets you take a simple growth rate
(where all the change happens all at once
at the end of a period of time - ie quantised growth)

e shows up whenever systems grow
exponentially and continuously.....
radioactive decay, interest calculations
and populations

tetryons

EM fields

quarks

leptons

Baryons

e can be applied to the equilateral energy geometries of physical systems
only where the rate of increase is a integer factor of a squared number

Nuclear Energy levels

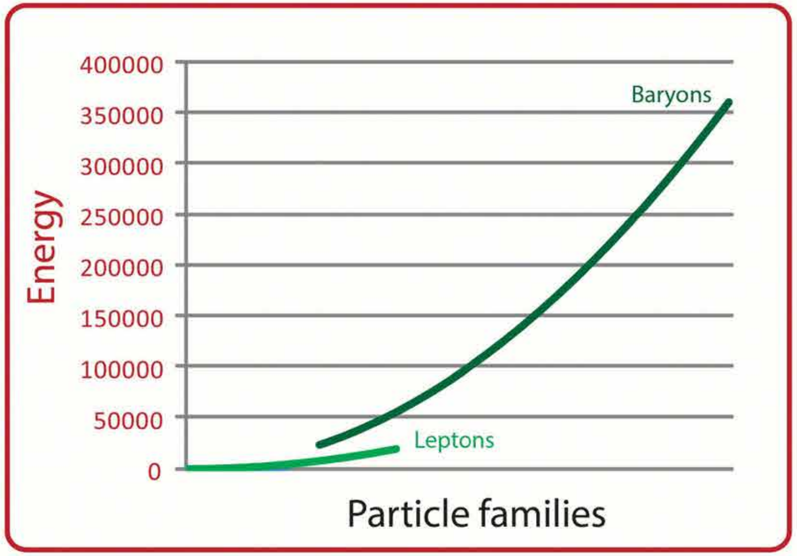

The emission and absorption of bosons and Photons within sub-atomic nuclei

Increase and decrease in integer amounts according to the charged Tetryonic topologies of the particles involved

Exponential energy levels

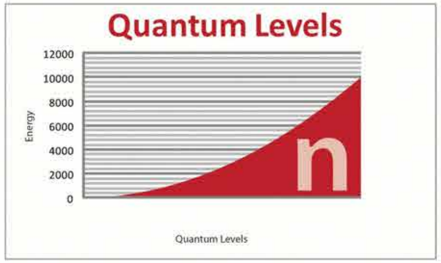

Radioactive decays follow exponential curves determined by the Tetryonic topology of the sub-atomic particle families

Proton
Neutron

antiNeutron
antiProton

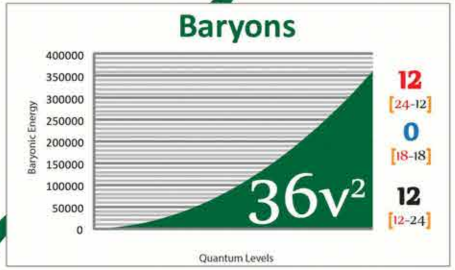

12 [24-12]
0 [18-18]
12 [12-24]

EM Field · Planck · quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons · ElectroMagnetic · mass · velocity

12 [12-0]
12 [0-12]
0 [6-6]
8 [10-2]
4 [4-8]

Electron
Positron

Neutrino

Up
Down
Strange
Charmed
Top
Bottom

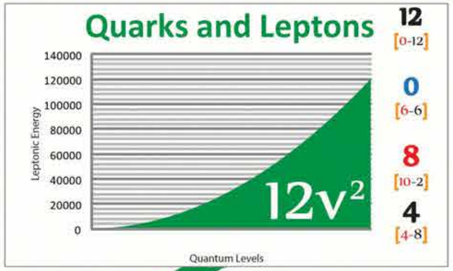

EM Field · Planck · quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

quarks / leptons · ElectroMagnetic · mass · velocity

Positive
Negative
Neutral

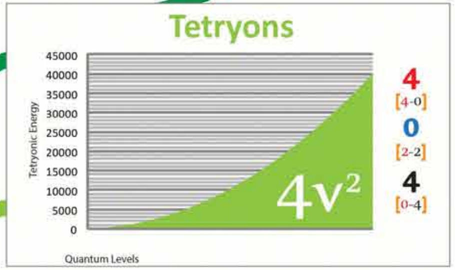

4 [4-0]
0 [2-2]
4 [0-4]

EM Field · Planck · quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

tetryons · ElectroMagnetic · mass · velocity

Copyright ABRAHAM [2008] - All rights reserved

Series addition & the Riemann Zeta Function

The second series addition of the Reimann Zeta function is where x=2: (pi^2)/6=1+1/2^2+1/3^2+1/4^2+...
(the sum of the reciprocals of the squares)

$$\zeta(s) = \sum_{n} \frac{1}{n^s}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^x}$$

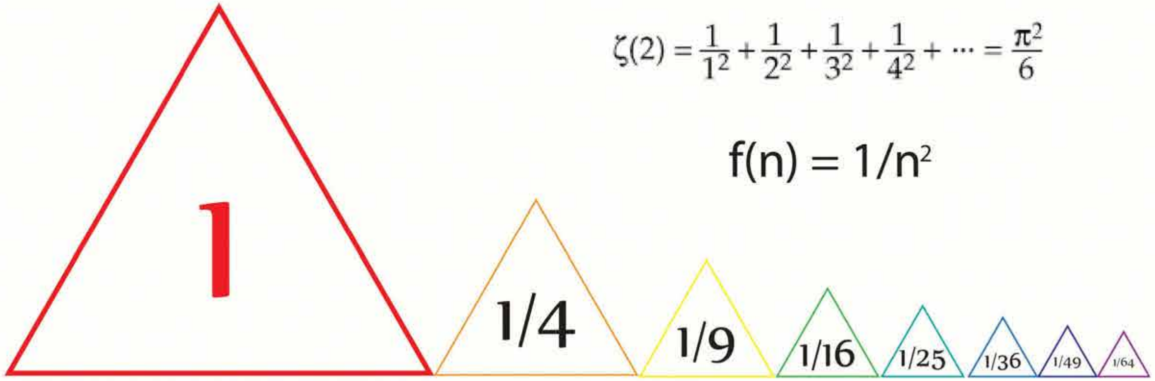

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}$$

$$f(n) = 1/n^2$$

In mathematics, the Riemann zeta function, is a prominent function of great significance in number theory. It is a named after German mathematician Bernhard Riemann. It is so important because of its relation to the distribution of prime numbers. It also has applications in other areas such as physics, probability theory, and applied statistics

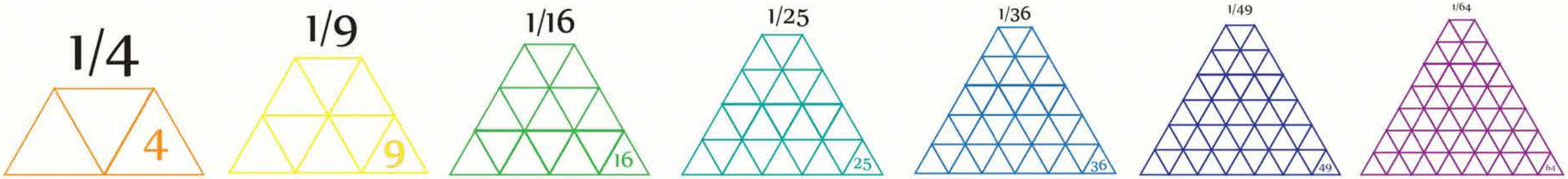

The mystery of prime numbers

Question: which natural numbers are prime? how are they distributed among natural numbers?

Primes are basic building blocks for natural numbers:

- any natural number is a product of prime numbers
- a prime number is only divisible by itself and by 1:
  (it cannot be further simplified)

We don't know how to predict where the prime numbers are:

“Prime numbers grow like weeds among the natural numbers, seeming to obey no other law than that of chance but also exhibit stunning regularity" (Don Zagier, number theorist)

Adding the odd numbers in order produces the square numbers

Apart from 2, all primes are odd numbers;
the difference between two consecutive squares being odd,
every prime can be expressed as the difference between two squares

SPECTRAL LINES

Primes can be modelled as Bosons [ODD number energy momenta geometries]

[2n-1]

KEM ENERGIES

Primes can be expressed as Bosons [the difference of two squared energies]

$n^2 - [n-1]^2$

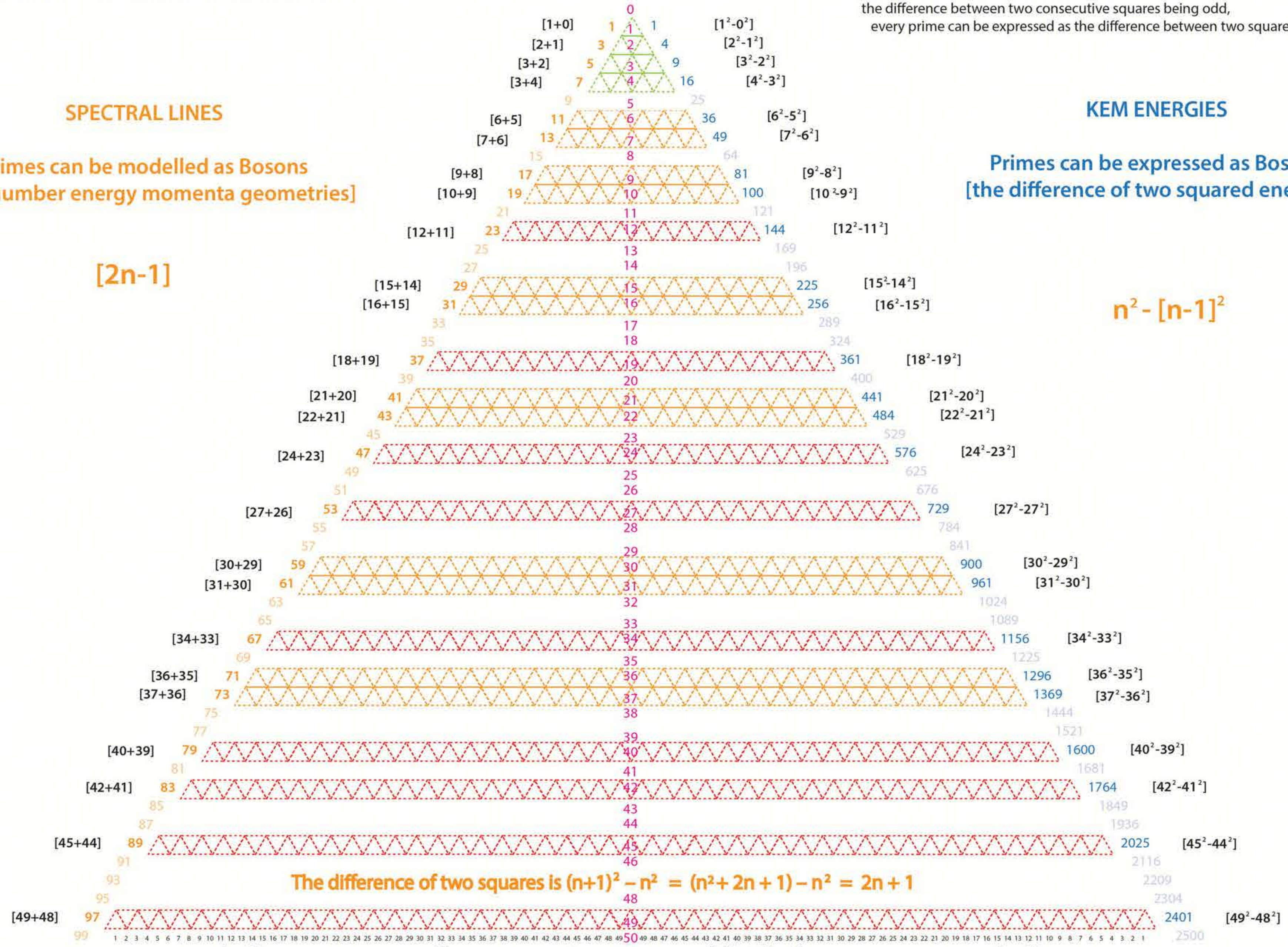
[1+0]
[2+1]
[3+2]
[3+4]
[6+5]
[7+6]
[9+8]
[10+9]
[12+11]
[15+14]
[16+15]
[18+19]
[21+20]
[22+21]
[24+23]
[27+26]
[30+29]
[31+30]
[34+33]
[36+35]
[37+36]
[40+39]
[42+41]
[45+44]
[49+48]
[1²-0²]
[2²-1²]
[3²-2²]
[4²-3²]
[6²-5²]
[7²-6²]
[9²-8²]
[10²-9²]
[12²-11²]
[15²-14²]
[16²-15²]
[18²-19²]
[21²-20²]
[22²-21²]
[24²-23²]
[27²-27²]
[30²-29²]
[31²-30²]
[34²-33²]
[36²-35²]
[37²-36²]
[40²-39²]
[42²-41²]
[45²-44²]
[49²-48²]

The difference of two squares is $(n+1)^2 - n^2 = (n^2 + 2n + 1) - n^2 = 2n + 1$

[n-1] **all energy quanta create normal distributions** **[n-1]**

Copyright ABRAHAM [2008] - All rights reserved

Prime number Distributions

A twin prime is a prime number that differs from another prime number by two, for example the twin prime pair (41, 43). Sometimes the term twin prime is used for a pair of twin primes; an alternative name for this is prime twin or prime pair.

Twin primes appear despite the general tendency of gaps between adjacent primes to become larger as the numbers themselves get larger due to the prime number theorem (the "average gap" between primes less than n is log(n)).

$$p_2 - p_1 = 2$$

Every twin prime pair except (3, 5) is of the form (6n – 1, 6n + 1) for some natural number n, and with the exception of n = 1, n must end in 0, 2, 3, 5, 7, or 8

What Yitang Zhang just proved is that there are infinitely many pairs of primes that differ by at most 70,000,000. In other words, that the gap between one prime and the next is bounded by 70,000,000 infinitely often—thus, the "bounded gaps" conjecture.

On first glance, this might seem a miraculous phenomenon. If the primes are tending to be farther and farther apart, what's causing there to be so many pairs that are close together?

It is the geometry of Numbers themselves that causes a natural geometric convergence

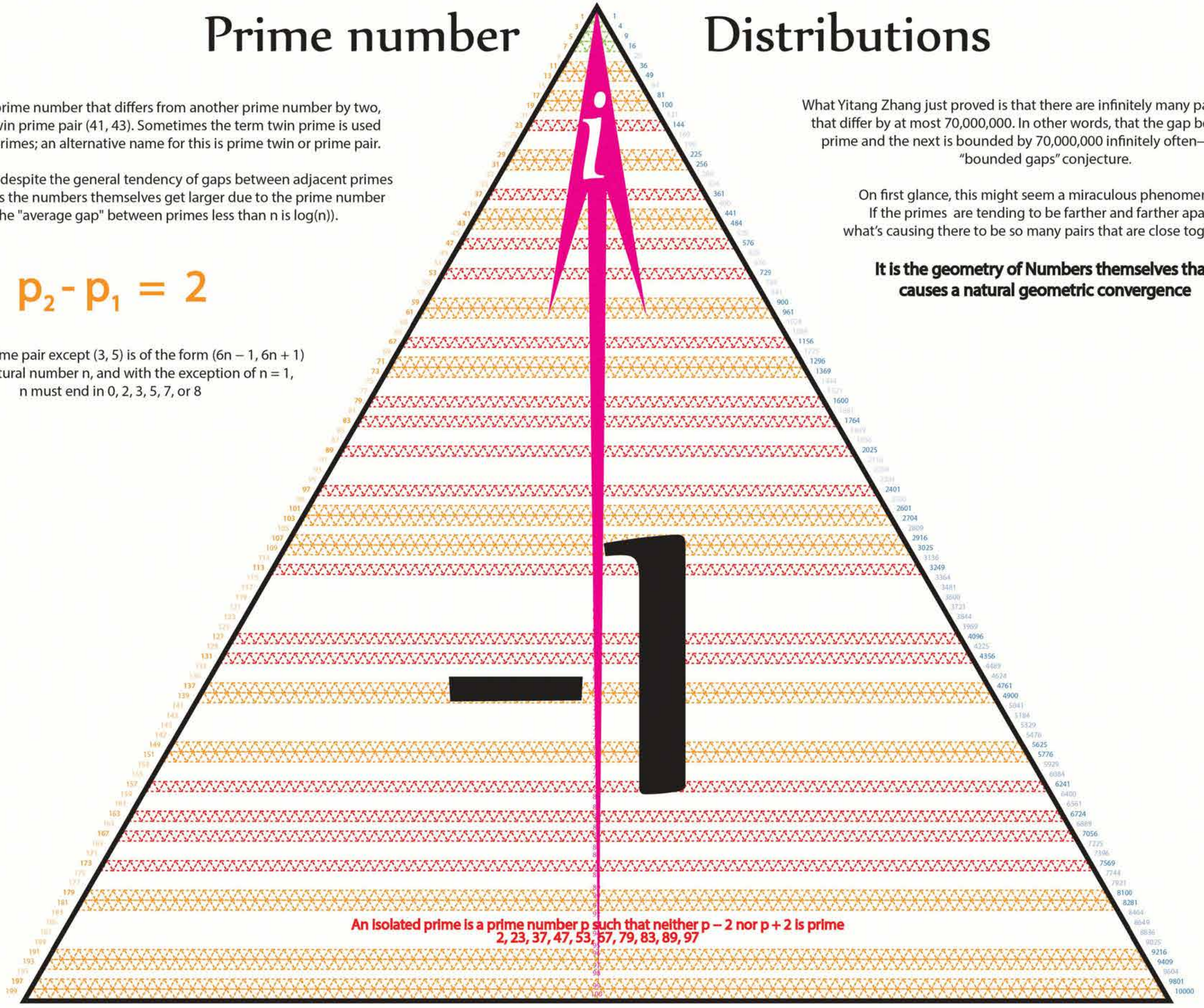

An isolated prime is a prime number p such that neither p – 2 nor p + 2 is prime
2, 23, 37, 47, 53, 67, 79, 83, 89, 97

Copyright ABRAHAM [2008] - All rights reserved

The Digital roots of Prime numbers

The digital root (also repeated digital sum) of a number is the (single digit) value obtained by an iterative process of summing digits, on each iteration using the result from the previous iteration to compute a digit sum.

The process continues until a single-digit number is reached.

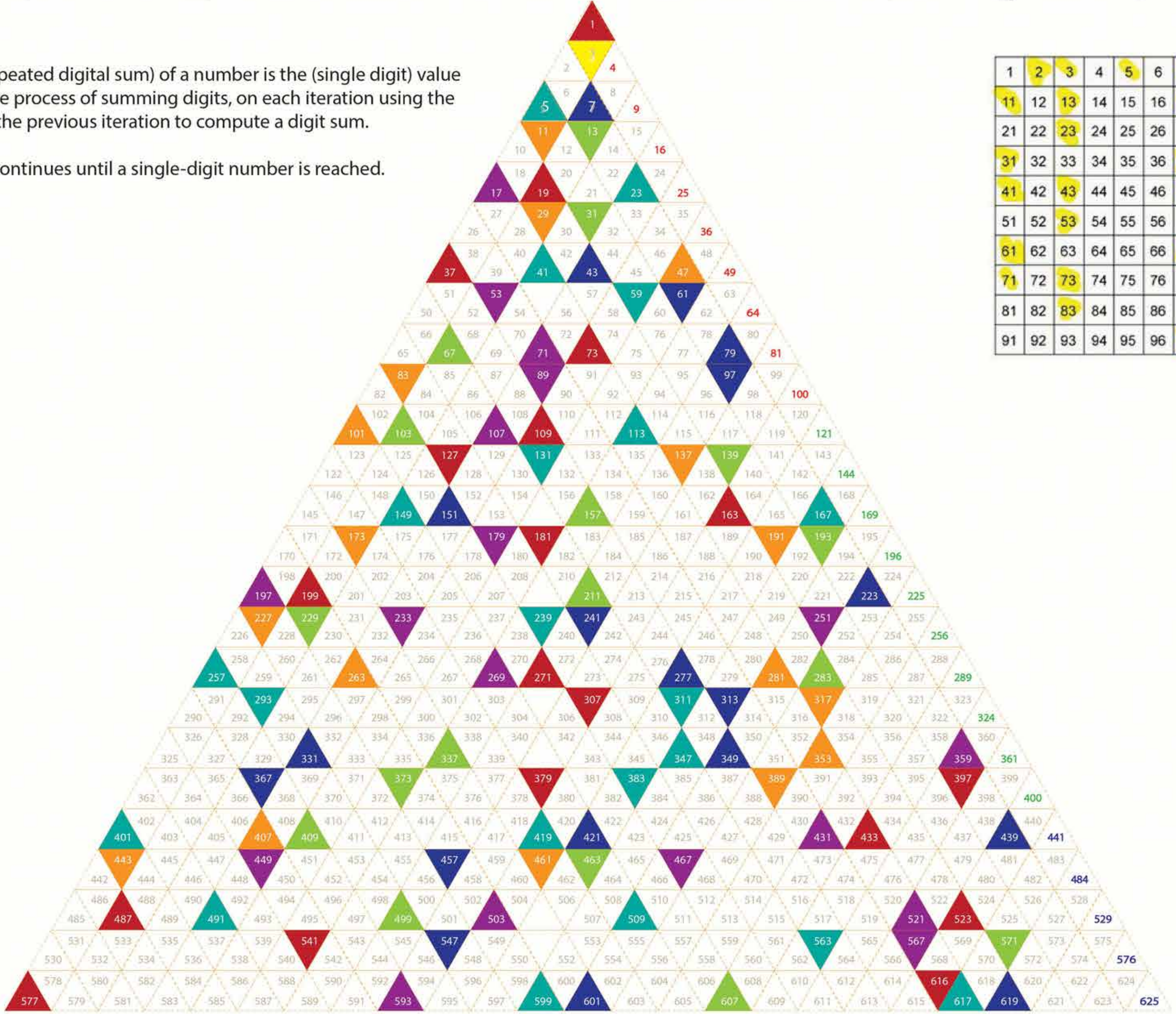
1
4
9
16
25
36
49
64
81
100
121
144
169
196
225
256
289
324
361
400
441
484
529
576
625

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Copyright ABRAHAM [2008] - All rights reserved

Archimedes is given credit for the first calculus.

Archimedes infinite series

Today's calculus was published by Newton.

1 + 3/4 + 3/16 + 3/64 + 3/256 + 3/1024 + ...

1 + .75 + .1875 + .046875 + .01171875 + ...

$$3\left(\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots\right)=1.$$

$$\frac{3}{4}+\frac{3}{4^2}+\frac{3}{4^3}+\frac{3}{4^4}+\cdots=1.$$

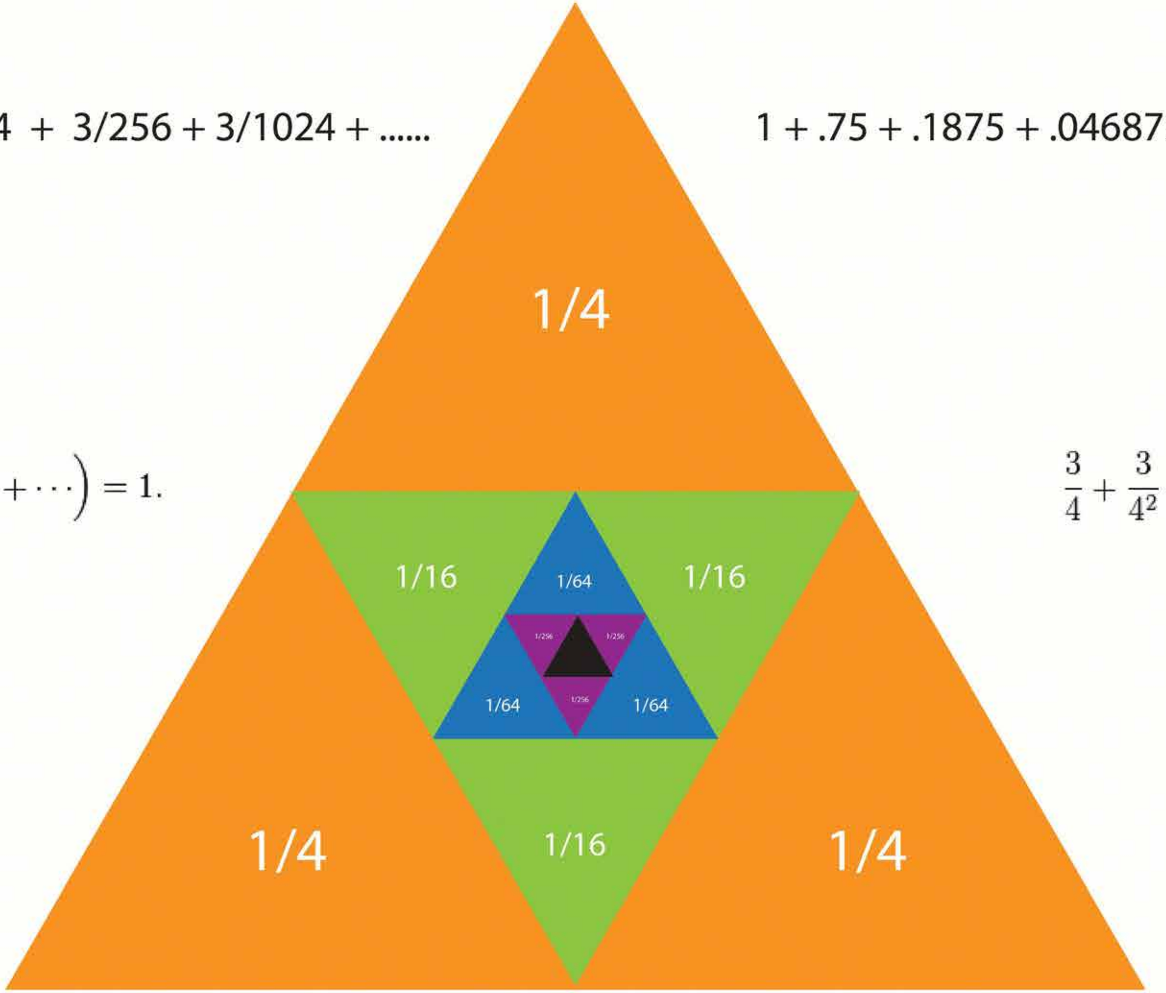
1/4
1/16
1/64
1/16
1/256
1/256
1/64
1/256
1/64
1/4
1/16
1/4

Nested convergent infinite series

Summing a convergent infinite series

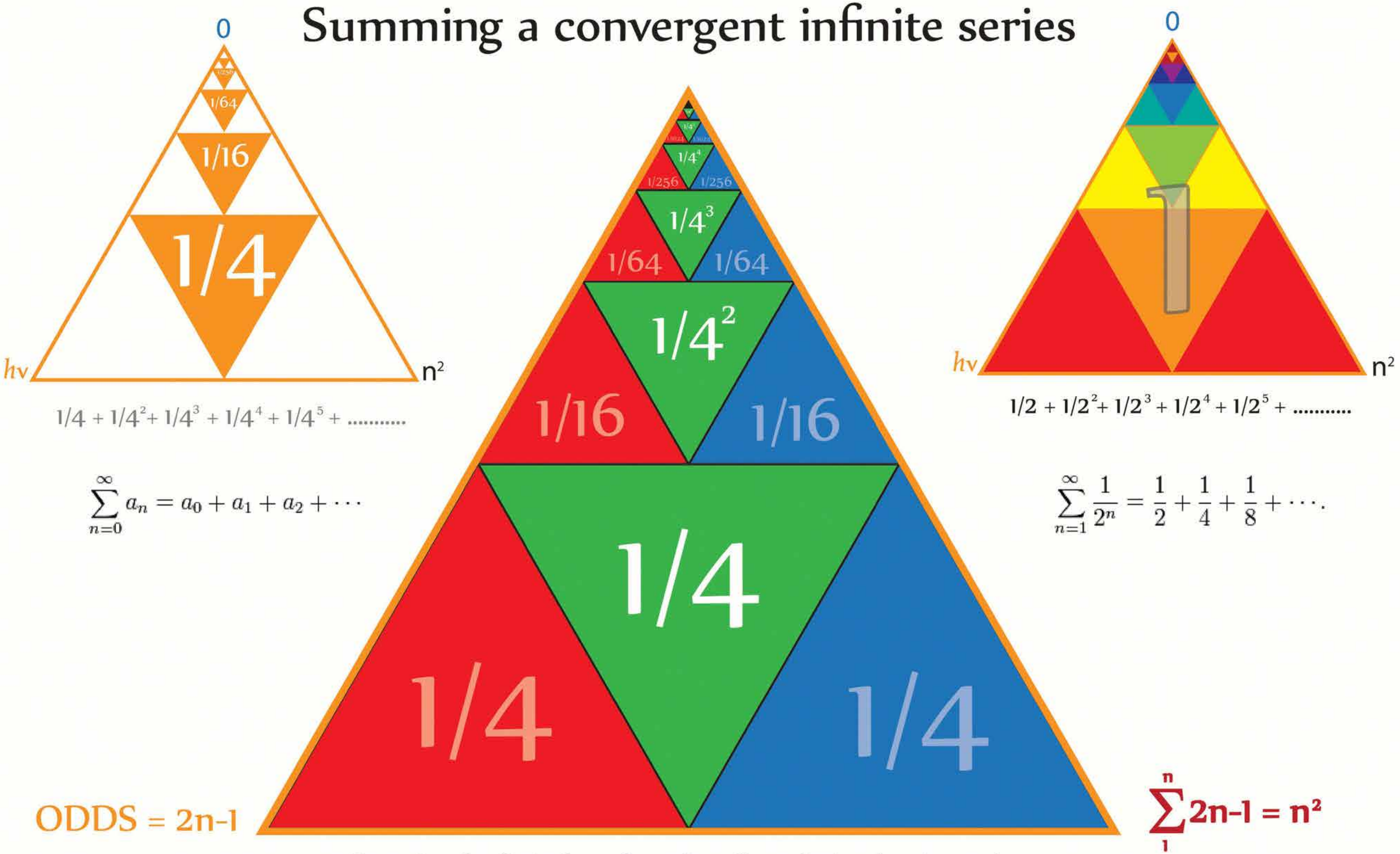
0
1/256
1/64
1/16
1/4
hv
n²
1/4⁴
1/256
1/256
1/4³
1/64
1/64
1/4²
1/16
1/16
1/4
1/4
1/4
0
1
hv
n²

1/4 + $1/4^2$ + $1/4^3$ + $1/4^4$ + $1/4^5$ + ...

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

1/2 + $1/2^2$ + $1/2^3$ + $1/2^4$ + $1/2^5$ + ...

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

ODDS = 2n-1

$$\sum_{1}^{n} 2n\text{-}1 = n^2$$

Summing the dissimilar coloured equilateral triangles gives unity

1/3 1/3 1/3

1/3 1/3 1/3

Summing the dissimilar coloured equilateral triangles gives unity

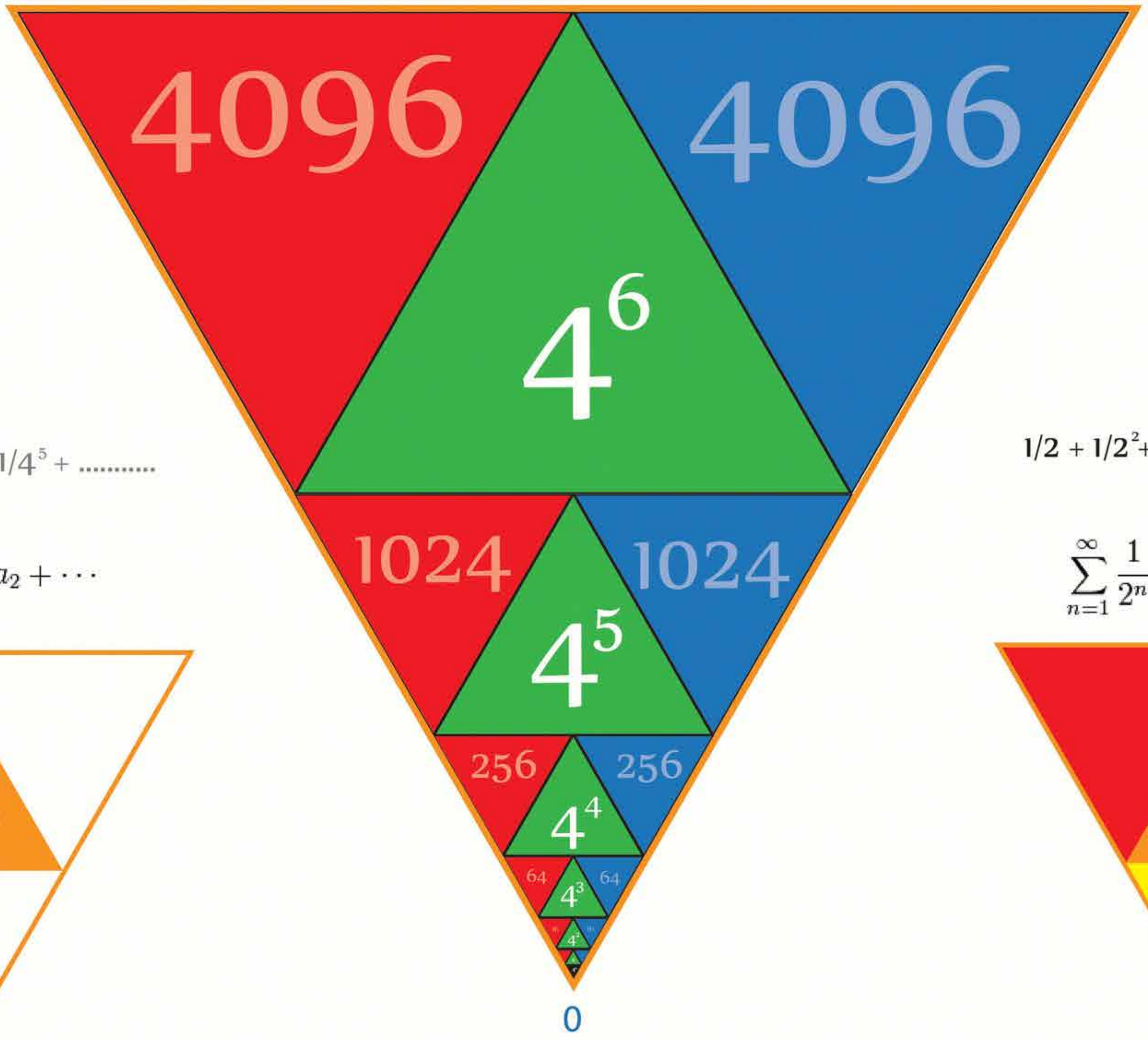
4096
4096
4^6
1024
1024
4^5
256
256
4^4
64
64
4^3
4^2
0

$1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 +$ ....

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

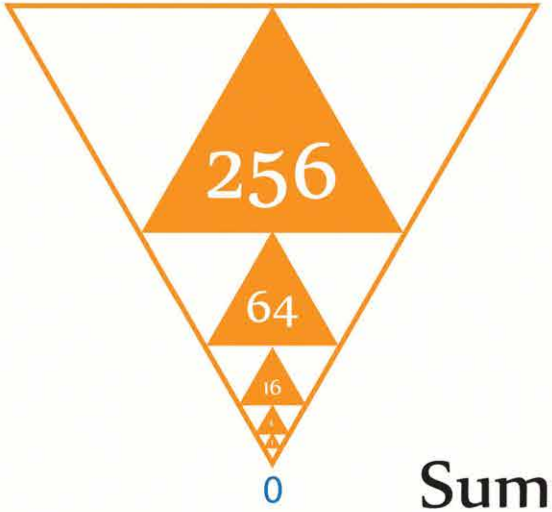

$1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 +$ ....

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

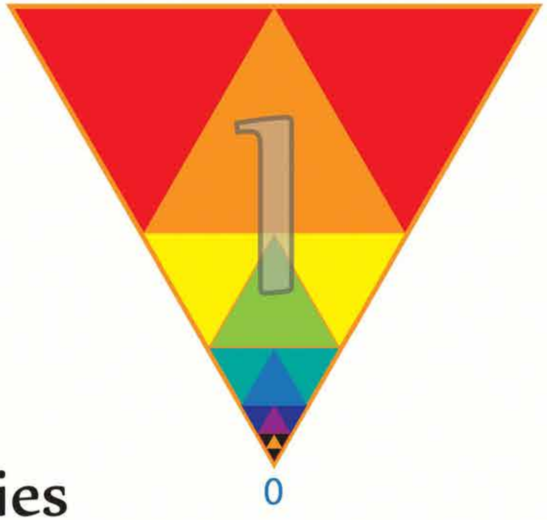

Summing a divergent infinite series

Pietro Mengoli

(1626 – June 7, 1686)

The Basel Problem

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6} \approx 1.644934$$

Leonhard Euler

(15 April 1707 – 18 Sept 1783)

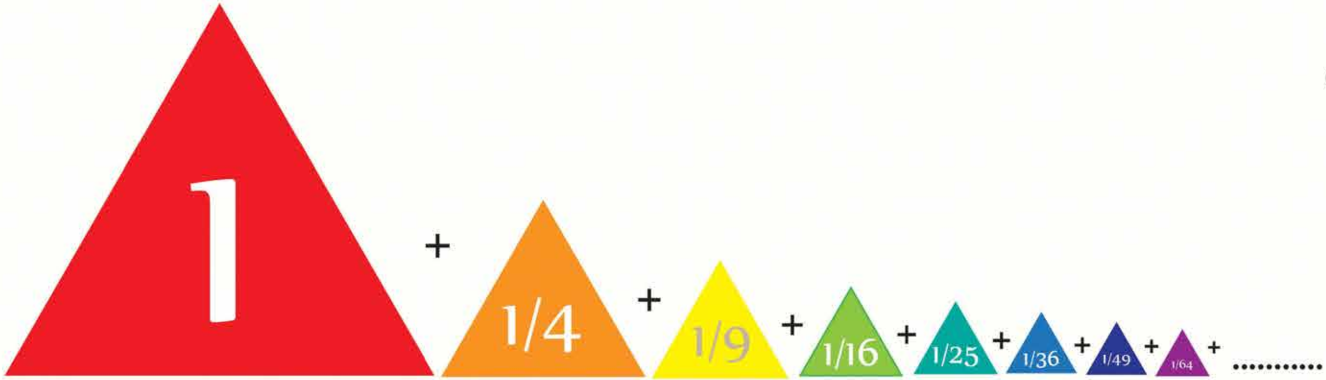

The Basel problem asks for the precise summation of the reciprocals of the squares of the natural numbers, i.e. the precise sum of the infinite series

$$f(n) = 1/n^2$$

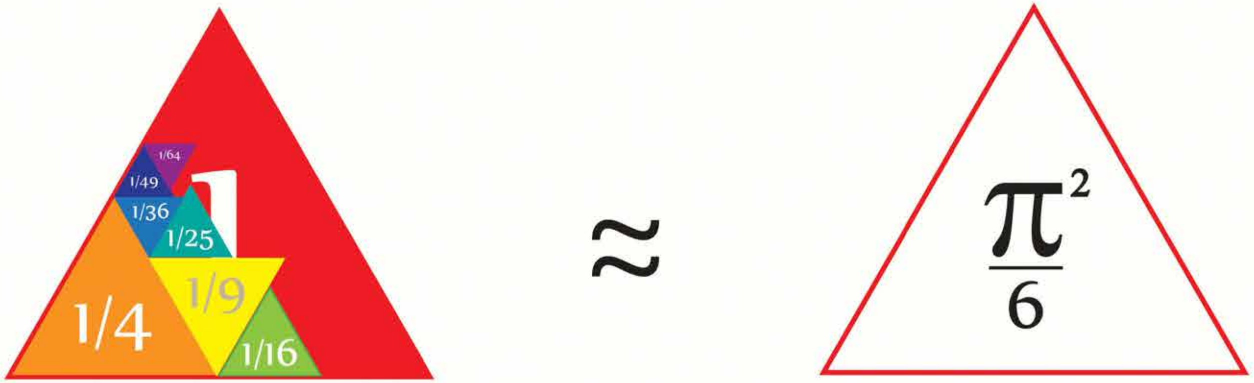

Integrals of mass-energy

is a means of finding scalar areas using summation and limits.

Integration is a micro adding of ***CONTINUOUS*** quantities.

$$\int_{n}^{\infty} f(x)$$

Integration is a special case of summation.

Integration is defined as the limit of a summation as the number of elements approches infinity while a part of their respective value approaches zero.

Summation is the finite sum of multiple, fixed values.

$$\sum_{n}^{\infty} x$$

Summation is a macro adding of ***DISCRETE*** quantities.

The summation of equilateral energy momenta quanta with respect to their linear vector components

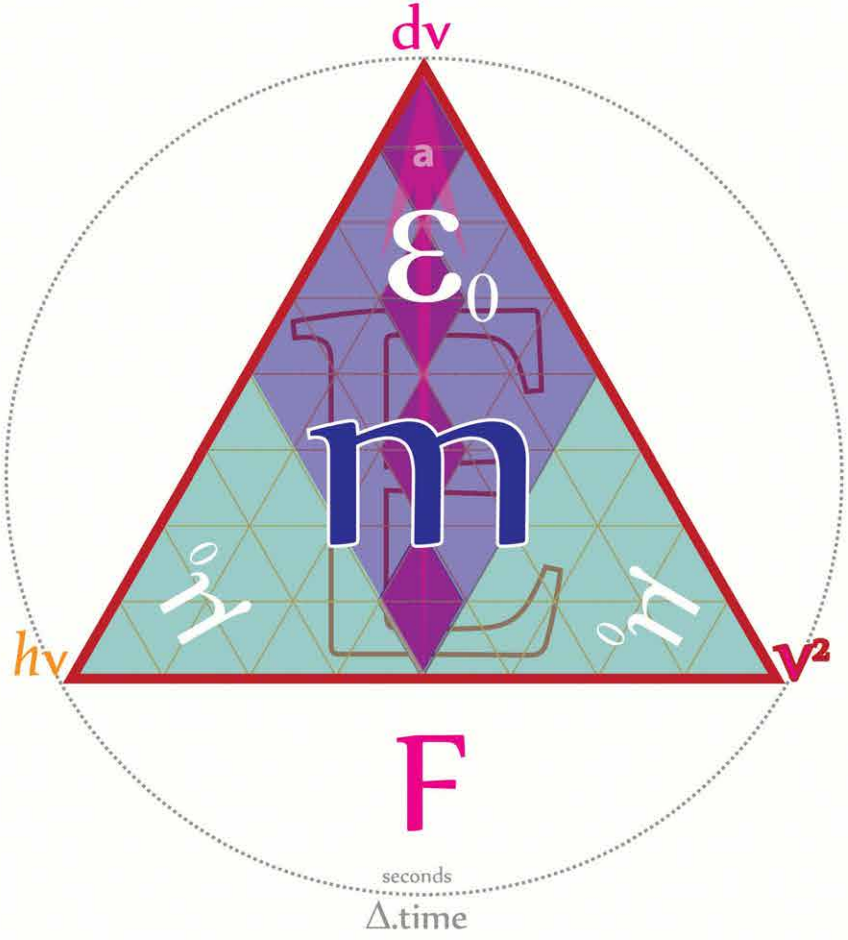

Integrating the energy quanta contained within equilateral charge geometry gives the variable Force required to acheive changes in motion [Energy, work, acceleration]

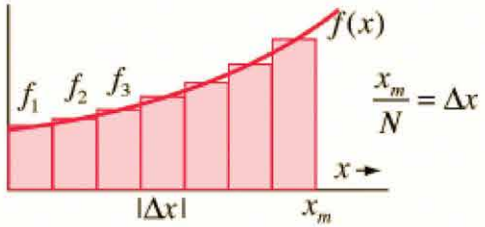

The Integral of the continuous area under the curve is the summation of an infinte number of disctrete rectangular measurements made to a specified limit

An integration isn't a simple summation, but the limit of a sequence of summations

All Planck energies are discrete equilateral geometries

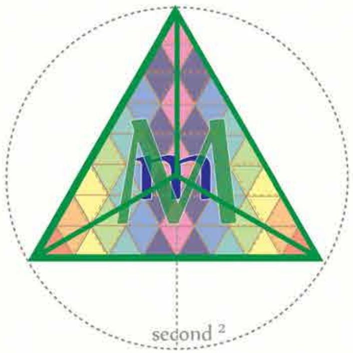

mass is the surface integral of EM energy geometries per unit of time

"The calculus of infinitesimals"

The fundamental theorem of calculus
simply states that the sum of infinitesimal changes in a quantity over time
adds up to the net change in the quantity.

Scalar Energy

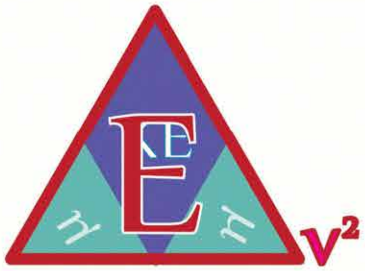

**Leibniz's vis viva
(Latin for living force)
is mv2,**

[2 x kinetic energy]

$$E = mv^2$$

Gottfried Wilhelm von Leibniz

(July 1, 1646 – November 14, 1716)

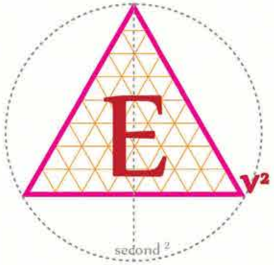

integral calculus

The founders of calculus thought
of the integral as an infinite sum of
rectangles of infinitesimal width

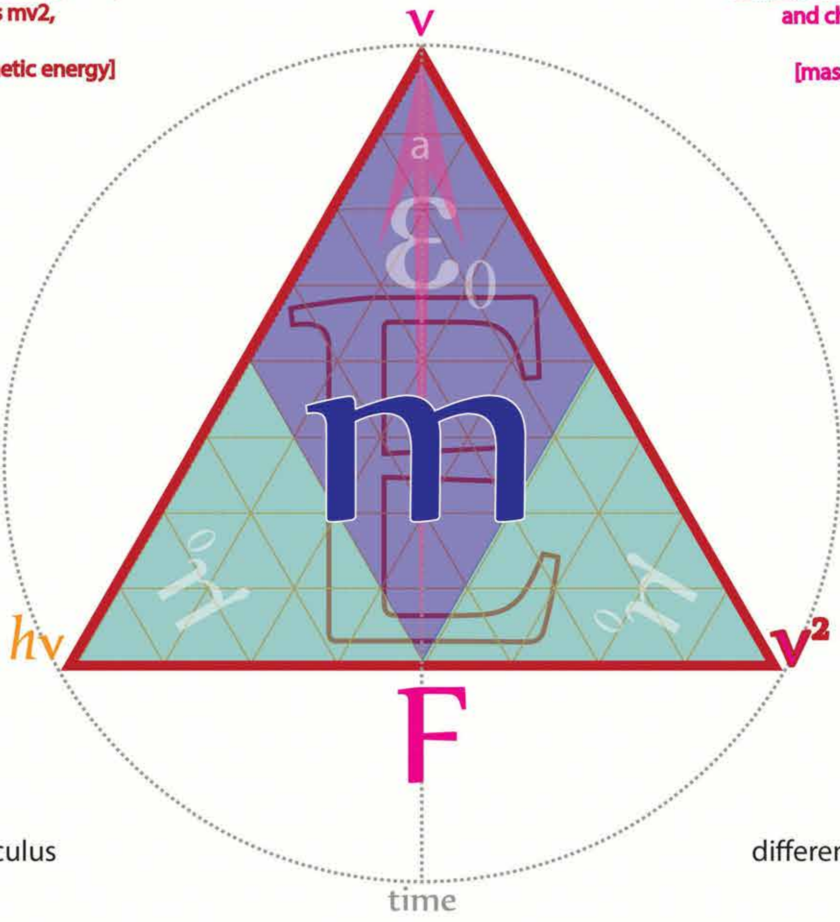

*The fundamental theorem of calculus is a theorem that links
the concept of the integral with the derivative of a function.*

vector Force

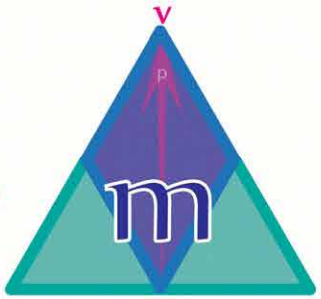

**Much of Newton's work
centred around momentum
and changes to it**

[mass.velocity]

$$F = m[dv/dt]$$

Sir Isaac Newton

(1643-1727)

differential calculus

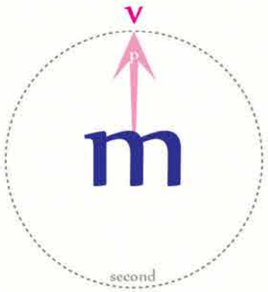

In calculus, a branch of mathematics,
the derivative is a measure of how
a function changes as its input changes

Energy momentum geometries

Leibniz

scalar energies

linear momentum is the square root vector force of the scalar energy required to do a set amount of work

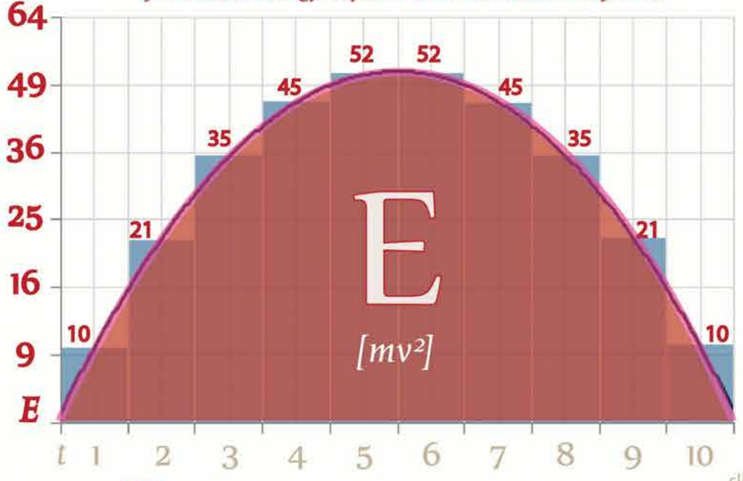

Newton

linear momentum

$d[mv]/dt = dp/dt = ma$

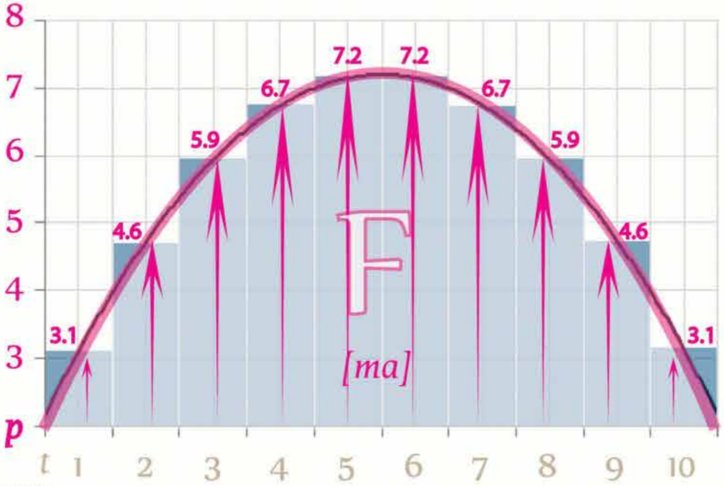

$$E = p^2$$

Energy momentum

$$kg\left[\frac{m}{s}\right]^2$$

print out this page

then cut out the 10 Planck energy momenta triangles shown

slice fine lines along the pink linear momentum arrows with a razor blade or similar as highlighted under the Newtonian acceleration curve

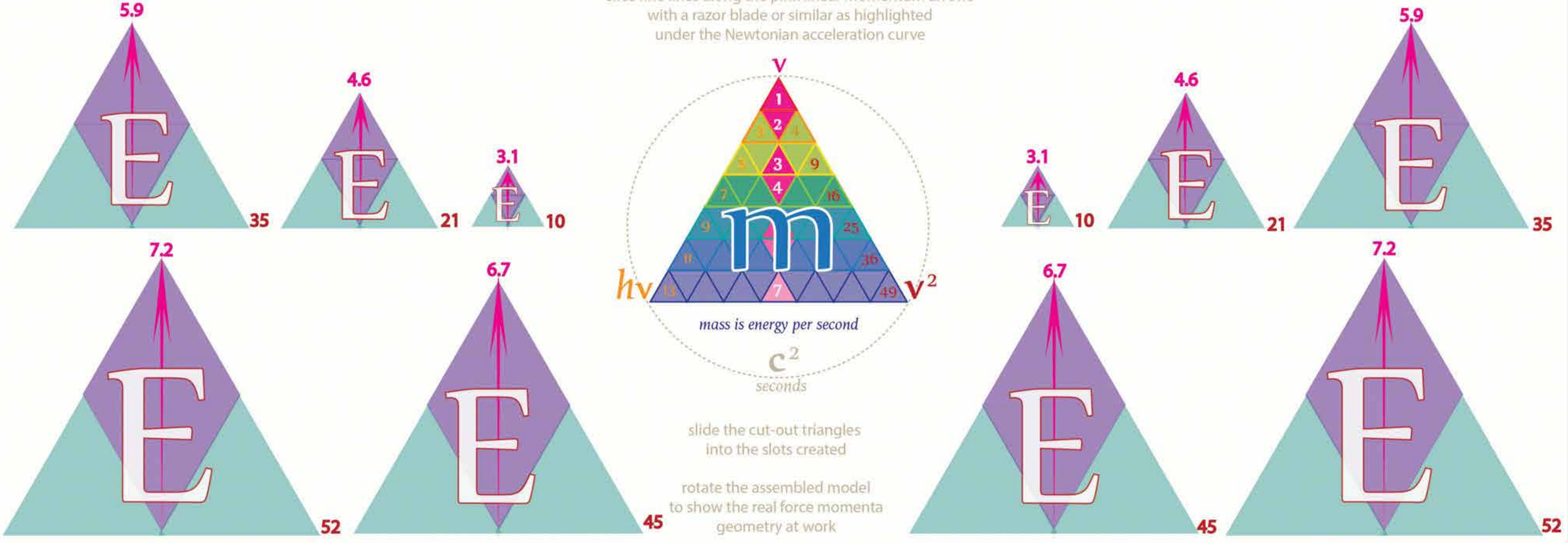
5.9
E
35
4.6
E
21
3.1
E
10
7.2
E
52
6.7
E
45
v
1
2
3
4
5
6
7
8
9
11
13
16
25
36
49
7
m
hv
v²
mass is energy per second
c²
seconds
3.1
E
10
4.6
E
21
5.9
E
35
6.7
E
45
7.2
E
52

slide the cut-out triangles into the slots created

rotate the assembled model to show the real force momenta geometry at work

Copyright ABRAHAM [2008] - All rights reserved

linear momentum is the square root vector force
of the scalar energy required to do a set amount of work

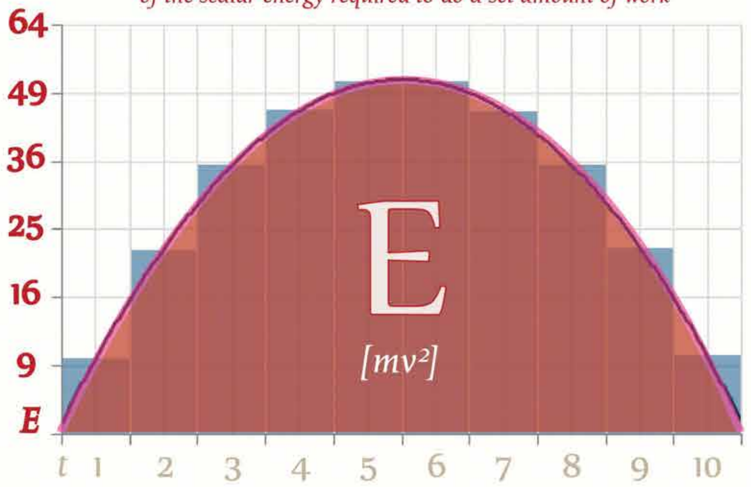

p

linear momentum

$\mathrm{kg}\,\frac{\mathrm{m}}{\mathrm{s}}$

print out this page

d[mv]/dt = dp/dt = ma

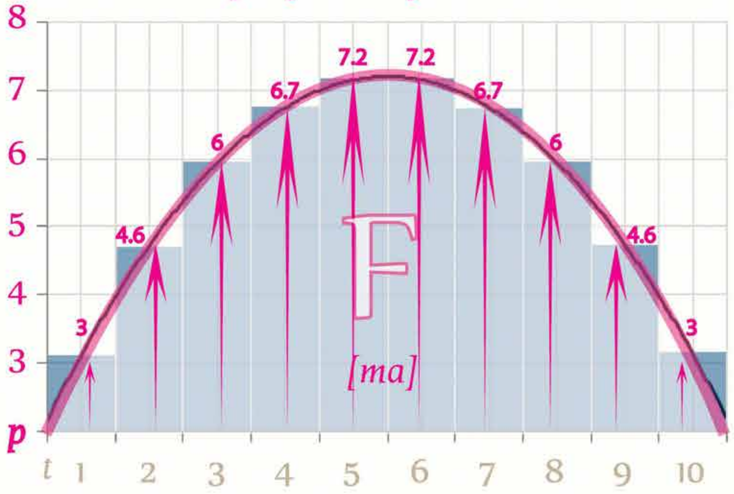

Leibniz

scalar energies

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

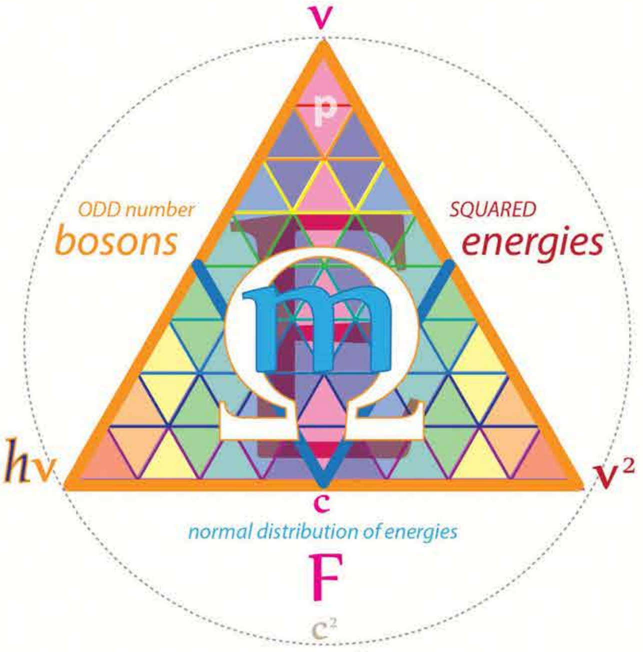
v
p
ODD number
bosons
SQUARED
energies
m
Ω
hv
v²
c
normal distribution of energies
F
c²

seconds

Newton

linear momentum

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

Planck quanta and their vector linear momentum
lie at 90 degrees to the angle use in the graphs of motion in calculus

Differentiation

Differentiation is concerned with things like speeds and accelerations, slopes and curves etc.
These are Rates of Change, they are things that are defined locally.

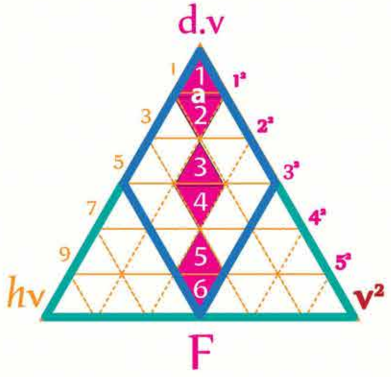
d.v
1
a
2
3
4
5
6
1
3
5
7
9
1²
2²
3²
4²
5²
hv
v²
F

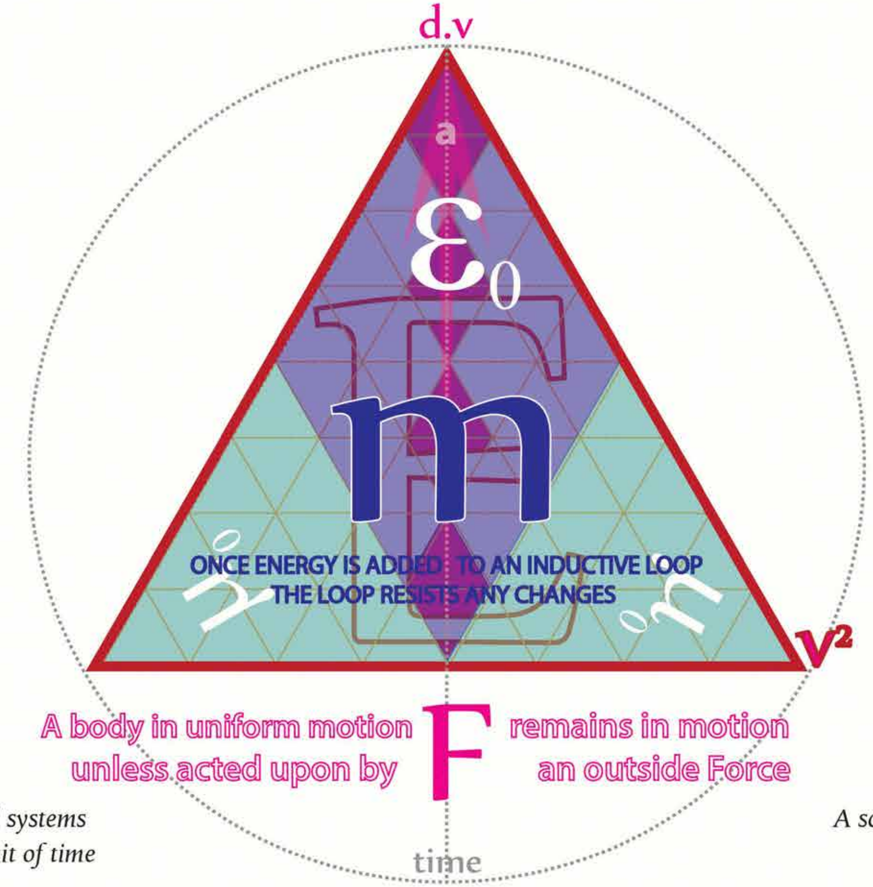
d.v
a
ε0
E
m
ONCE ENERGY IS ADDED TO AN INDUCTIVE LOOP
THE LOOP RESISTS ANY CHANGES
v²
time

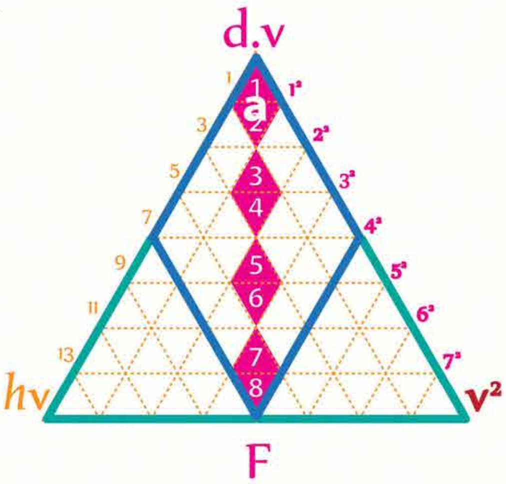
d.v
1
a
2
3
4
5
6
7
8
1
3
5
7
9
11
13
1²
2²
3²
4²
5²
6²
7²
hv
v²
F

An increase in a force
opposing an object's
vector velocity results in
DECCELERATION

An increase in a force
in line with an object's
vector velocity results in
ACCELERATION

A body in uniform motion remains in motion
unless acted upon by **F** an outside Force

A linear measure of Forces acting on physical systems resulting in changes to distance covered per unit of time

A scalar measure of Forces acting on physical systems resulting in changes to their rate of motion

$$v = \frac{\Delta d}{\Delta t}$$

$$= m/s$$

$$F = ma$$

In physics, the derivative of the displacement of a moving body with respect to time is the velocity of the body, and the derivative of velocity with respect to time is acceleration.

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

Newton's second law of motion states that the derivative of the momentum of a body equals the force applied to the body.

$$a = \frac{\Delta v}{\Delta t}$$

$$= m/s^2$$

Visualising the geometric half-truths of relativistic physics

The source of all the physical relationships of mass-energy momenta and the constants in Physics is the Equilateral Triangle (and all texts must be corrected)

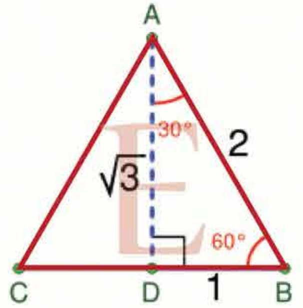

Energy geometries within Physics including Special Relativity with its Lorentz corrections have historically been incorrectly illustrated through the geometry of right angled triangles

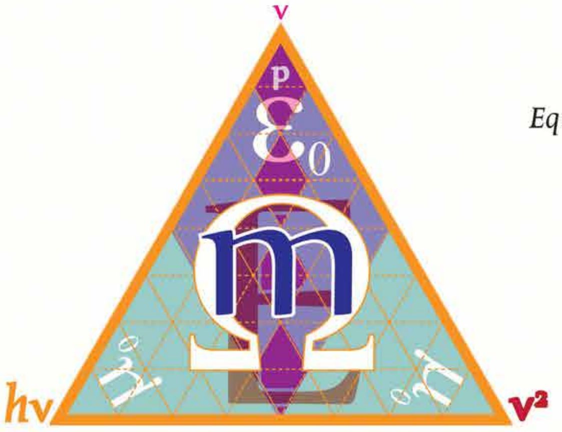

F

$$\sum \mathbf{F} = \frac{d\mathbf{P}}{dt} = m\frac{d\mathbf{v}}{dt} = m\mathbf{a}$$

$$\mathbf{F} = m\mathbf{a}\,.$$

Physics is geometry, one cannot be separated from the other

Equilateral geomtries lead to a intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all the other apsects of Nature.

$$mv^2 = E = hv^2$$

ENERGY

6.629432673 e-34 J

$$\left[m\Omega v^2 \right]$$

Planck quanta
mass velocity

7.376238634 e-51 kg

momentum

$$p^2 = mv^2$$

$$E = m_0c^2 + K$$

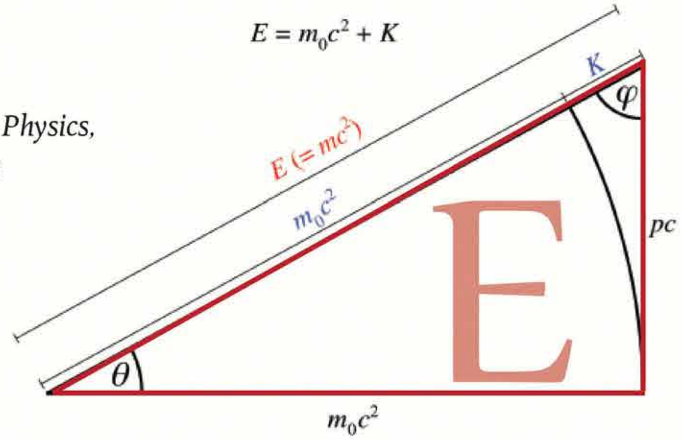

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total energy, mass, or distance in spacetime is the sum of the components squared.

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related tot the Energy content of a body by the velocity of Energy

$$E = mc^2$$

Velocity

In physics, velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The scalar absolute value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms–1) when using the SI (metric) system.

$$\frac{m}{s}$$

Speed is the scalar value of the Distance traveled per unit of TIme

$$\bar{\mathbf{v}} = \frac{\Delta \mathbf{x}}{\Delta t}.$$

Velocity is the vector value of the Distance traveled per unit of Time

$$\frac{\vec{m}}{s}$$

Velocity squared is the scalar value of the Distance traveled per unit of Time squared (Energy of a given spatial volume)

$$\frac{m^2}{s^2}$$

v Velocity $\frac{m}{s}$

sec

is a 2D
radial Space-time
MEASUREMENT

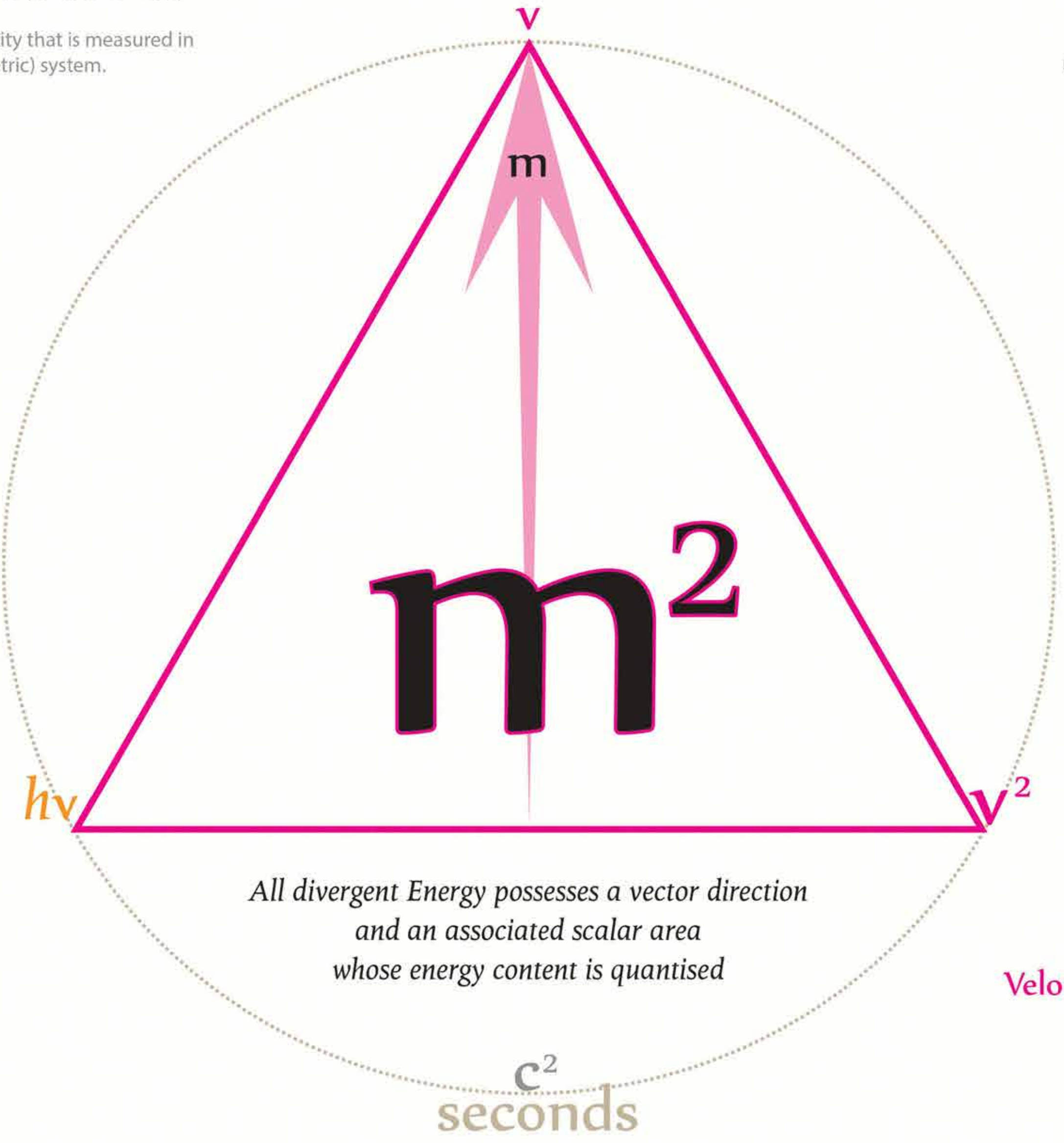

All divergent Energy possesses a vector direction and an associated scalar area whose energy content is quantised

Velocity squared

$$\frac{m^2}{s^2}$$

Copyright ABRAHAM [2008] - All rights reserved

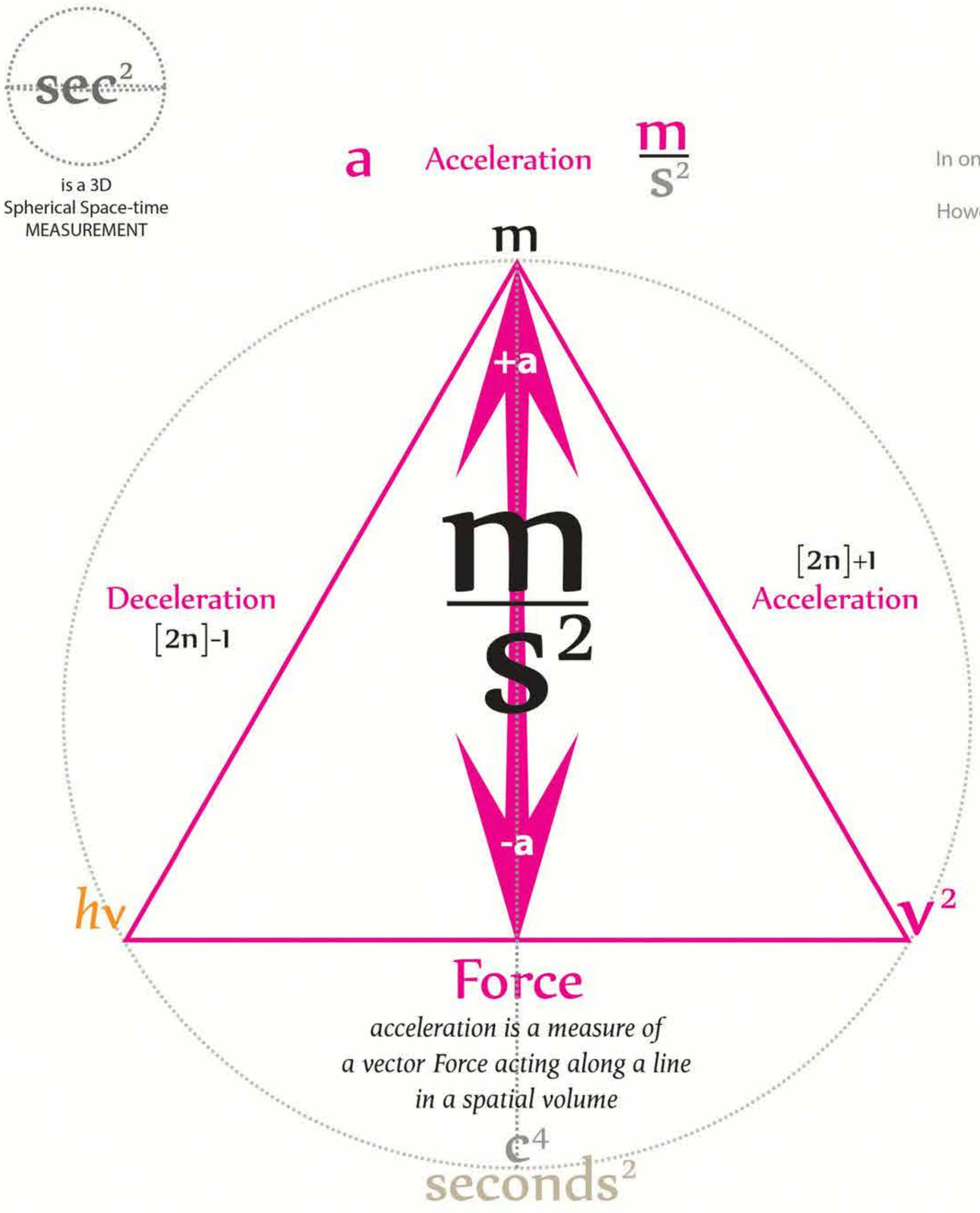
sec²
is a 3D
Spherical Space-time
MEASUREMENT
a
Acceleration
m/s²
m
+a
m/s²
-a
Deceleration
[2n]-1
[2n]+1
Acceleration
hv
v²
Force
acceleration is a measure of
a vector Force acting along a line
in a spatial volume
c⁴
seconds²

Acceleration

In physics, acceleration is the rate of change of velocity over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

Force

$$kg\,\frac{m}{s^2}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

Copyright ABRAHAM [2008] - All rights reserved

$\mathrm{kg}\frac{\mathrm{m}}{\mathrm{s}}$

linear
momentum

p

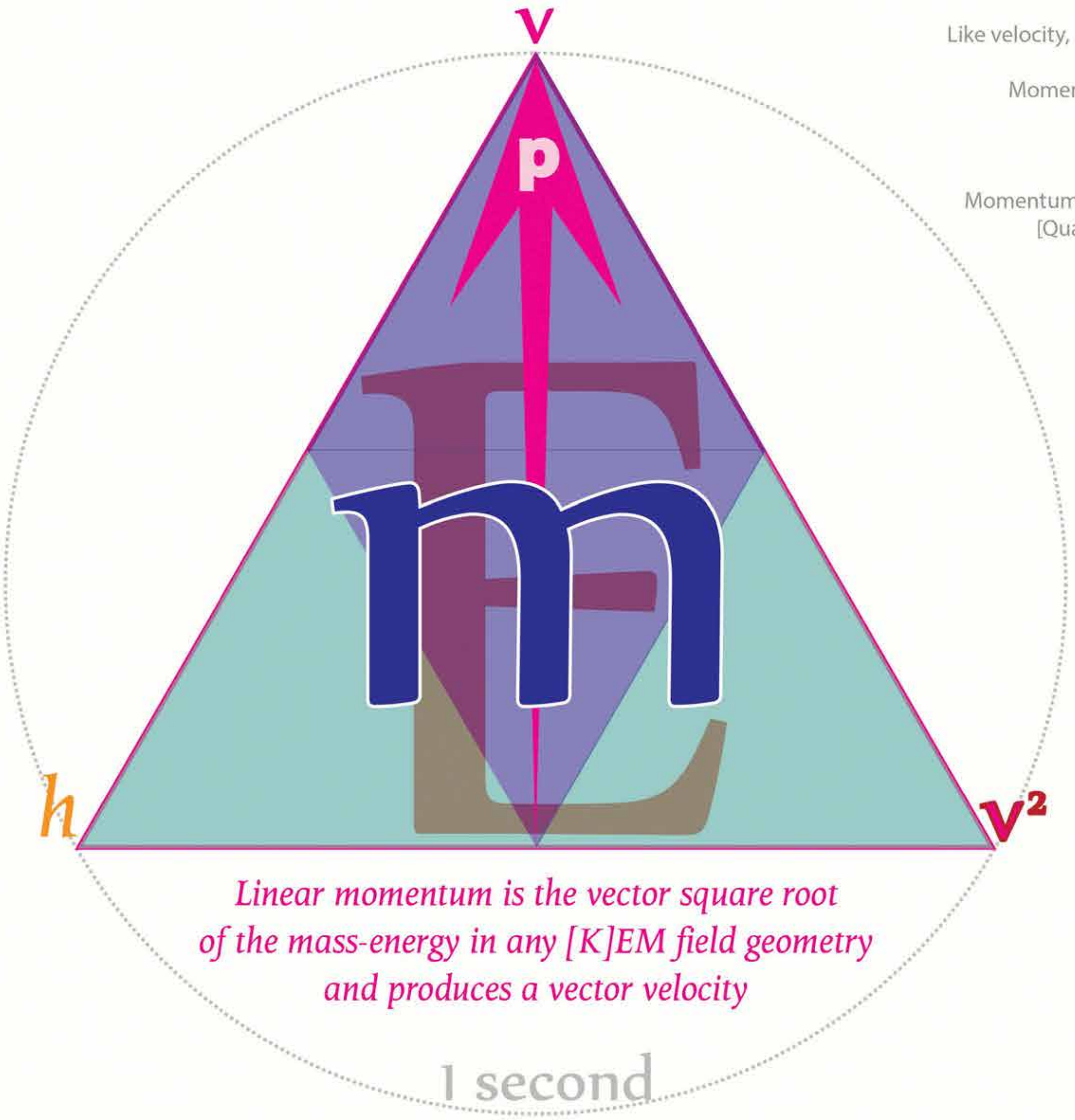

Linear momentum is the vector square root
of the mass-energy in any [K]EM field geometry
and produces a vector velocity

Momentum

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s)
is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum),
meaning that if a closed system is not affected by external forces,
its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms
[Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{h v^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as
the square of linear momentum

$$\mathrm{kg}\frac{\mathrm{m}^2}{\mathrm{s}^2}$$

equilateral Planck energy momenta

$$m\Omega v^2$$

Copyright ABRAHAM [2008] - All rights reserved

Energy-momentum relationship

The total number of equilateral Planck quanta [quantised mass-energy momenta]
is directly related to the square of its linear momentum [mass-velocity]

Quantum Mechanics

$$h = \frac{E}{v^2}$$

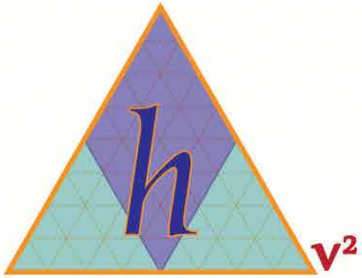

Quantised energy equilateral momenta

$$hv^2$$

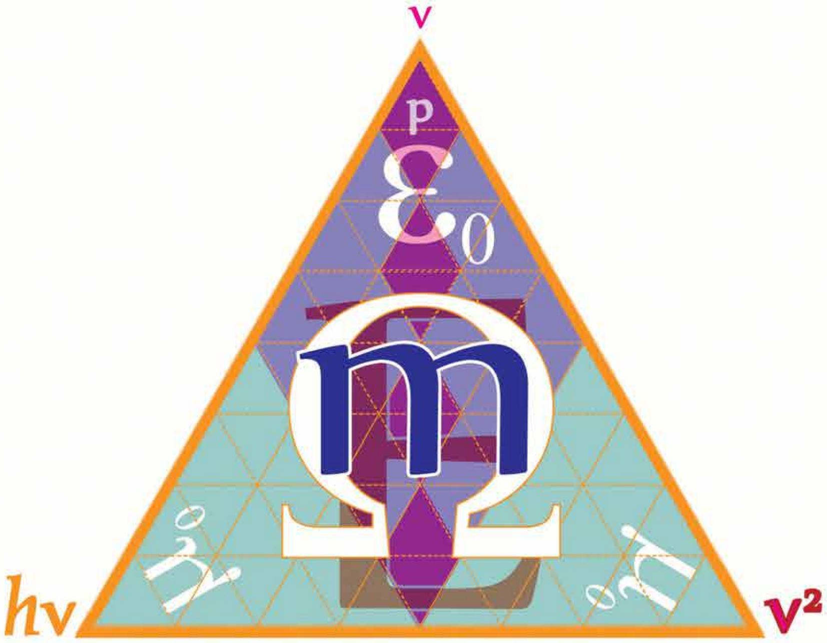

Quantised Energy momenta is related to
Scalar mass energy momenta through
the equilateral geometry of Planck's constant

Newtonian physics

$$m = \frac{E}{v^2}$$

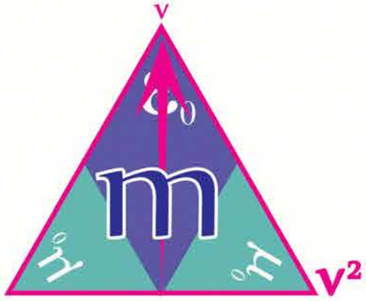

Scalar energy linear momentum

$$pv$$

$$m\Omega v^2 = E = mv^2$$

Copyright ABRAHAM [2008] - All rights reserved

Inertial resistance to Force

Inertia is the resistance of any physical object
to a change in its state of motion.

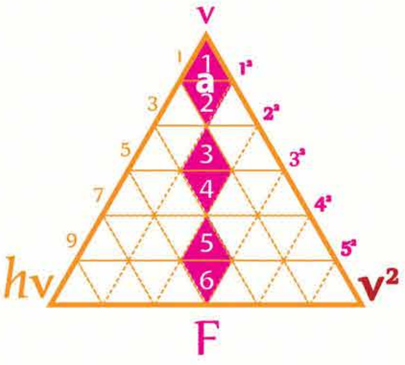
v
1
a
2
3
4
5
6
1
3
5
7
9
1^2
2^2
3^2
4^2
5^2
hv
v^2
F

Changes to mass-velocity require a corresponding change to boson quanta

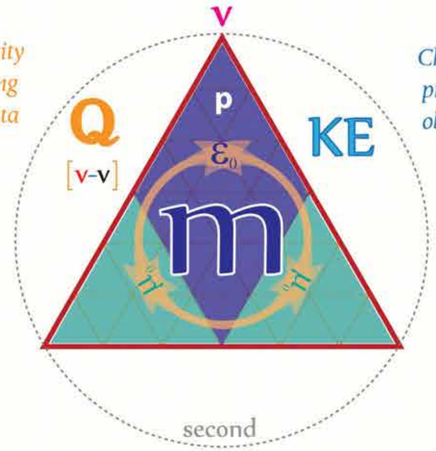
v
Q
[v-v]
p
KE
ε_0
m
μ
μ
second

Changes to mass-velocity produces a change in an object's Kinetic Energies

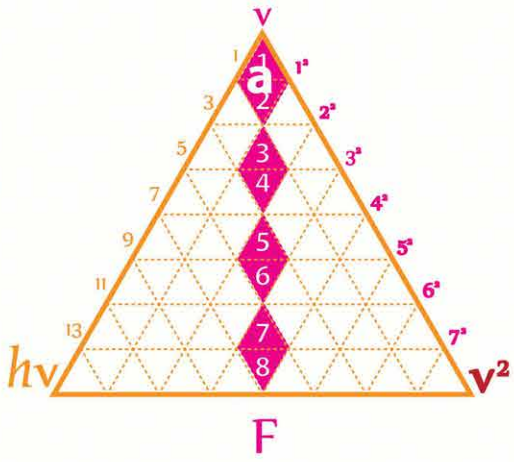
v
1
a
2
3
4
5
6
7
8
1
3
5
7
9
11
13
1^2
2^2
3^2
4^2
5^2
6^2
7^2
hv
v^2
F

$$E = mc^2$$

The total intrinsic momenta of all energy waveforms is the sum of their constituent Quantised Angular momenta (mass-energy momenta)

$$M_0 = m/c^2$$

Any change in motion results in changes to the Charge geometries creating in turn proportional changes to KEM mass and momenta components

$$F = ma$$

The 'inductive resistance' of Charge quanta fields to changes in their mass-energy momenta content is what we term Inertia

$$KEM = Mv^2$$

Matter in motion has Kinetic Energies in addition to invariant rest mass-Energy

KE = RE - rest Matter

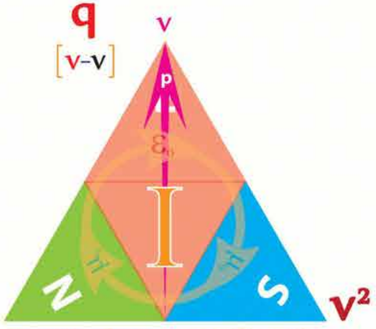
q
[v-v]
v
p
ε_0
I
μ
μ
N
S
v^2

quantised moment of inertia

Any change to an object's velocity results in a corresponding change to its mass-Energy momenta which is reflected by its inertia

$$p^2 = E = mv^2$$

Inertial mass can be related to Charge through inductive Planck quanta

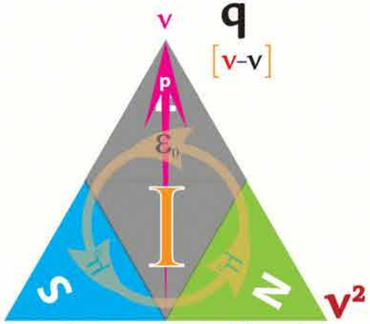
v
q
[v-v]
p
ε_0
I
μ
μ
S
N
v^2

quantised moment of inertia

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

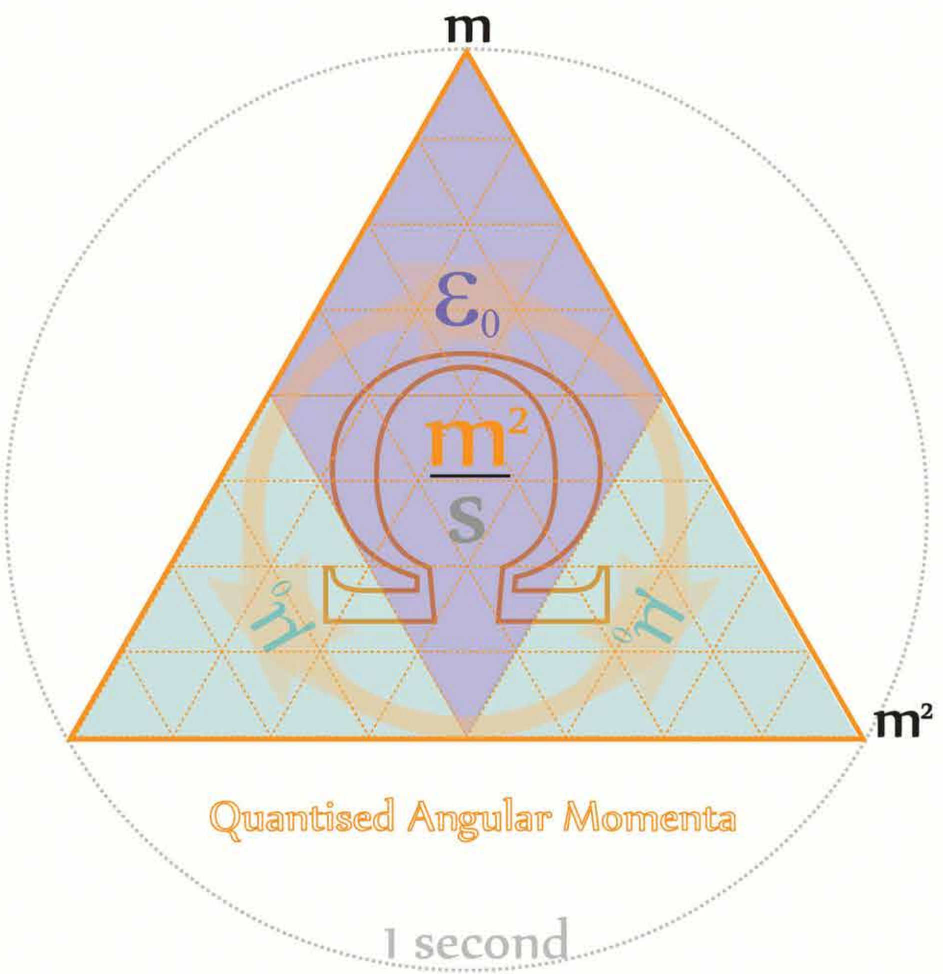
m
ε_0
$\frac{m^2}{s}$
Ω
m^2
Quantised Angular Momenta
1 second

Angular momentum is sometimes described as the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of quantised mass-energy momenta within any defined space-time co-ordinate system

A major re-definition of quantised angular momentum in physics is revealed

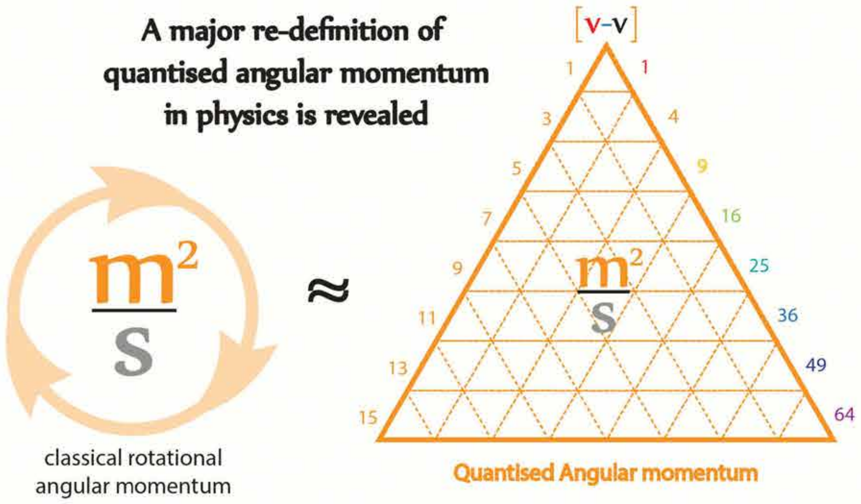
$\frac{m^2}{s}$
classical rotational angular momentum
≈
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
$\frac{m^2}{s}$
Quantised Angular momentum

In quantum mechanics, angular momentum is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum, or rotational momentum is a conserved vector quantity that can be used to describe the overall state of a physical system.

When applied to specific mass-Energy-Matter systems QAM reveals the true quantum geometry and nature of Energy in our universe

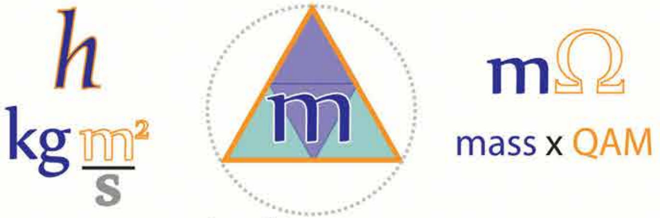
h
kg $\frac{m^2}{s}$
m
mΩ
mass x QAM

Planck's Constant

Normally viewed as an expression of rotational momentum Quantised Angular Momentum [QAM] is in fact a result of the equilateral geometric quantization of mass-energy

Copyright ABRAHAM [2008] - All rights reserved

Charged geometries

All charge geometries are nett divergent

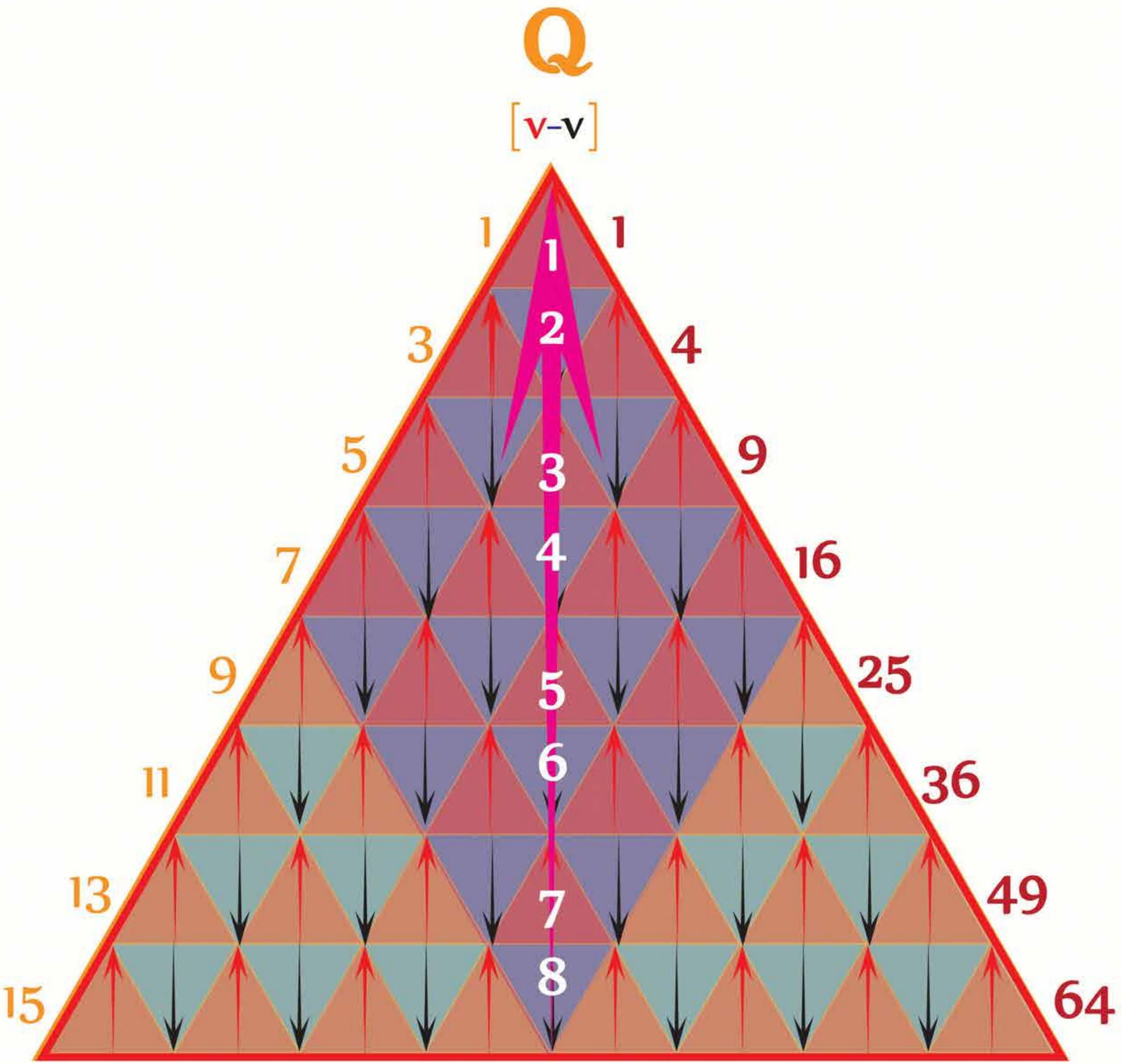
Q
[v-v]
1
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
1
2
3
4
5
6
7
8

Divergent energy momenta	Convergent energy momenta
1	0
3	1
6	3
10	6
15	10
21	15
28	21
36	28

All charge geometries are comprised of finite equilateral energy momenta quanta

Copyright ABRAHAM [2008] - All rights reserved

Renormalisation

Renormalization was first developed in quantum electrodynamics (QED) to make sense of infinite integrals in perturbation theory. The problem of infinities first arose in the classical electrodynamics of point particles in the19th and later in the calculation of Gravitational fields in General Relativity in the early 20th century.

In QED Infinities must be cancelled

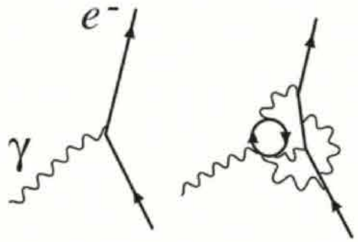

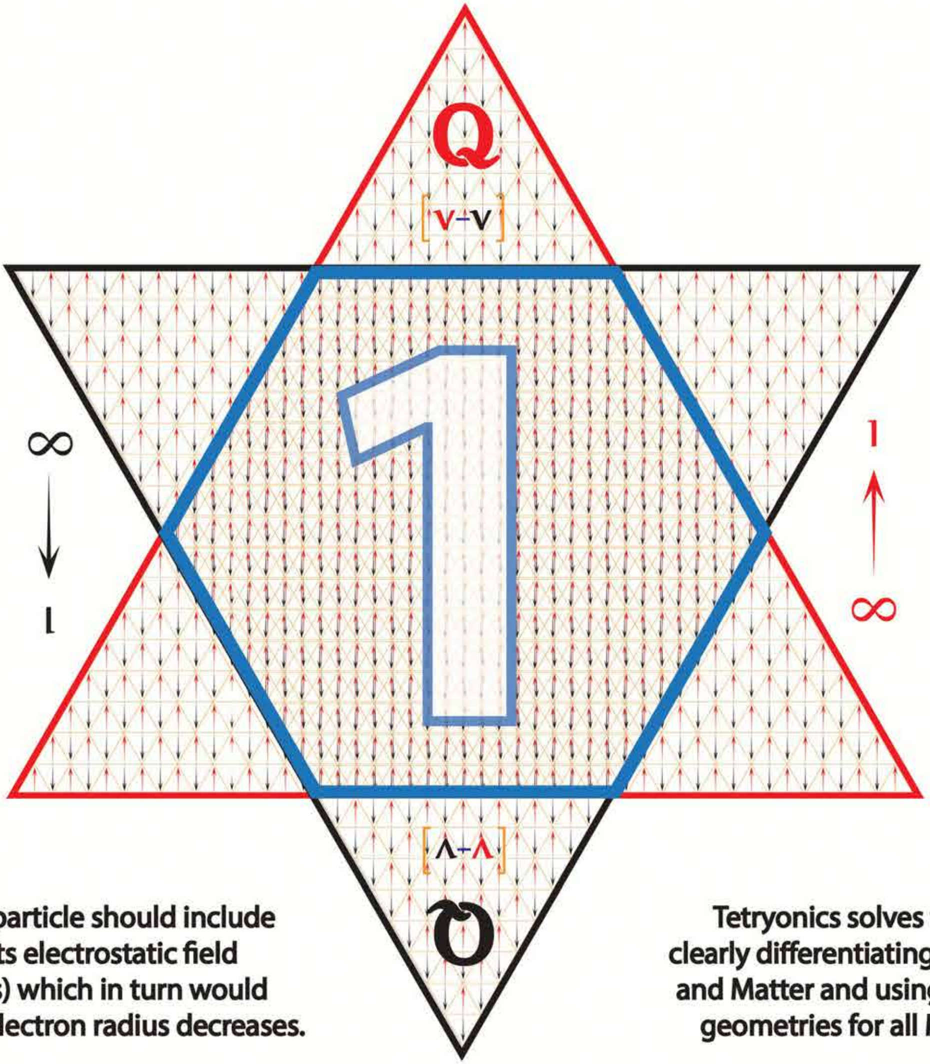

In Tetryonics Infinities do not exist

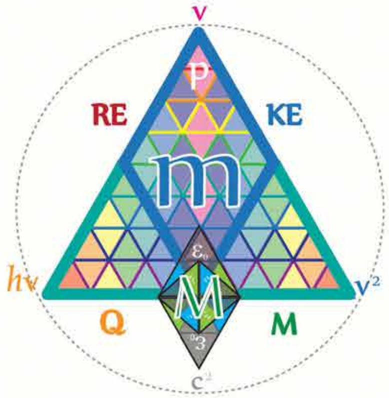

The mass of a charged particle should include the mass-energy in its electrostatic field (Electromagnetic mass) which in turn would approach infinity as the electron radius decreases.

Tetryonics solves the problem by clearly differentiating between EM mass and Matter and using finite equilateral geometries for all Matter in motion

Initially viewed as a suspicious provisional procedure by some of its originators, renormalization was eventually embraced as an important and self-consistent tool in several fields of physics and mathematics.

2D space [c^2]

The adjective Cartesian refers to the French mathematician and philosopher René Descartes who developed the coordinate system in 1637

Since then many other coordinate systems have been developed such as the polar coordinates for the plane, and the spherical and cylindrical coordinates for three-dimensional space.

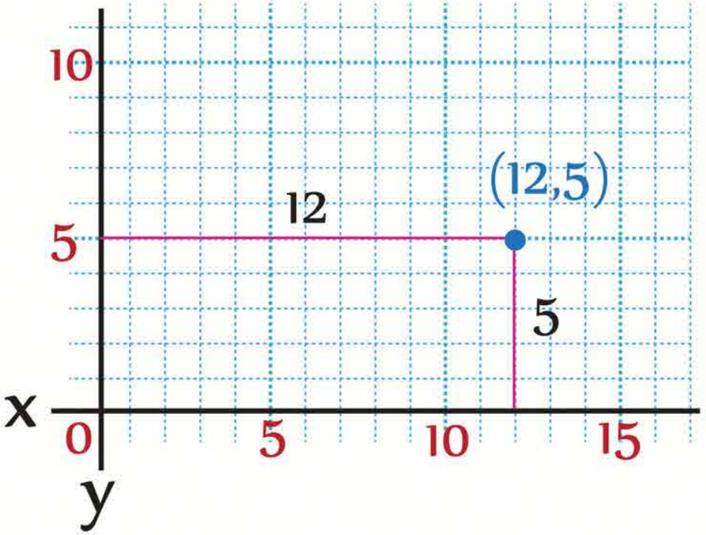

Cartesian coordinates are the foundation of analytic geometry, and provide enlightening geometric interpretations for many other branches of mathematics, such as linear algebra, complex analysis, differential geometry, multivariate calculus, group theory, and more

Mapping 3D spaces using Recti-linear co-ordinates

Cartesian coordinates can be defined as the positions of the perpendicular projections of a point onto the two or more axes, expressed as signed distances from the origin.

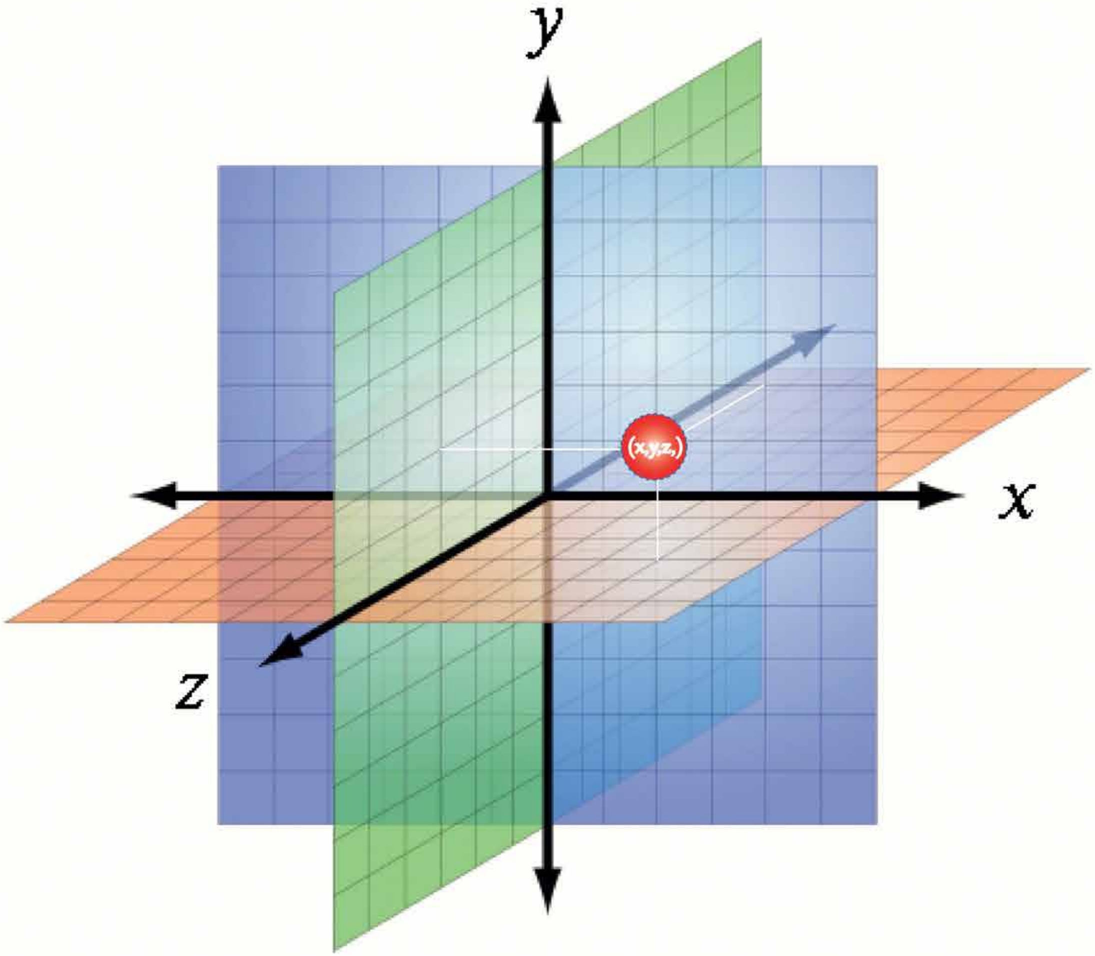

3D Cartesian co-ordinate [c^3] systems are distinct from spherical co-ordinate [c^4] systems

Polar co-ordinates

In mathematics, the polar coordinate system is a two-dimensional co-ordinate system in which each point on a plane is determined by a distance from a fixed point and an angle from a fixed direction.

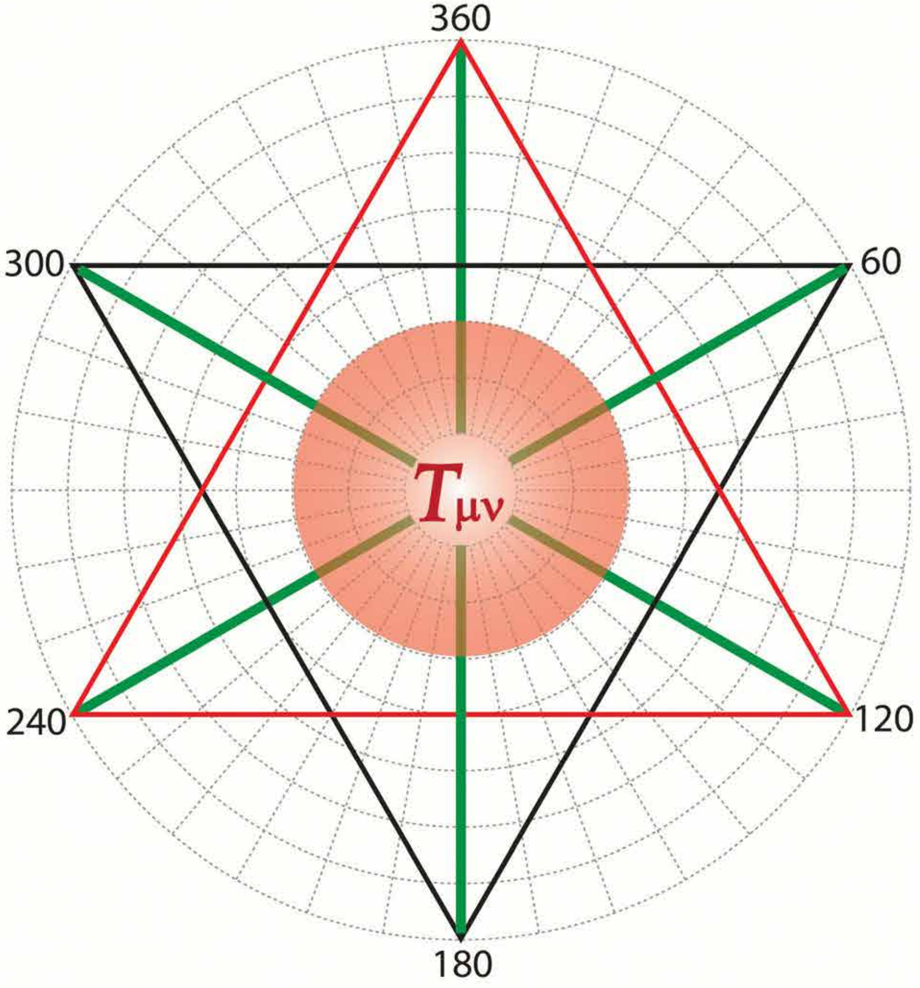

In geometry, curvilinear coordinates are a coordinate system for Euclidean space in which the co-ordinate lines may be curved.

Action Dynamics

Curvilinear co-ordinates may be derived from a set of rectilinear Cartesian coordinates by using a locally invertible transformation that maps one point to another in both systems

Metric Tensors

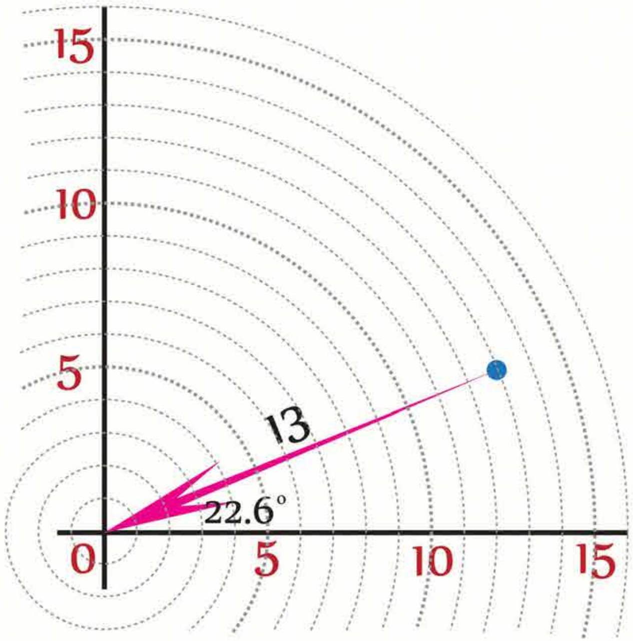

Gravitational acceleration

Polar or curvilinear co-ordinate systems are used extensively by Einstein in his theory of General Relavtivity

Reimannian curved space-time

Copyright ABRAHAM [2008] - All rights reserved

Co-ordinate transformations

There are many different possible coordinate systems for describing geometrical figures and they can all be related to one another.

Such relations are described by coordinate transformations which give formulas for the coordinates in one system in terms of the coordinates in another system

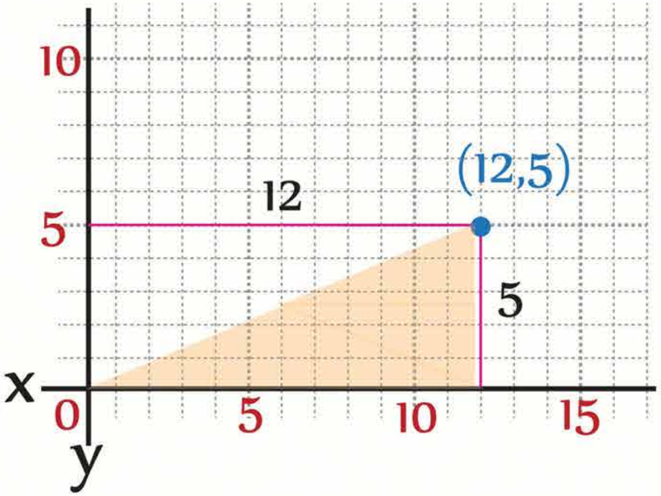

Converting between Polar and Cartesian coordinates

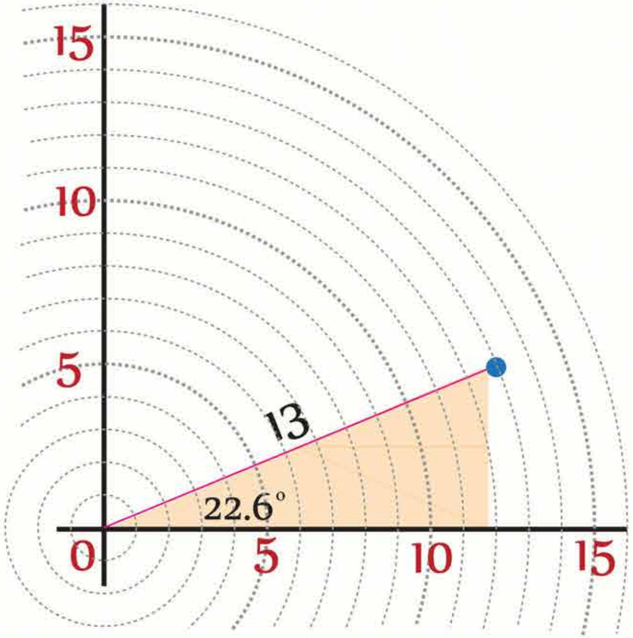

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$$

$$= \tan^{-1}(5/12)$$

$$= 22.619$$

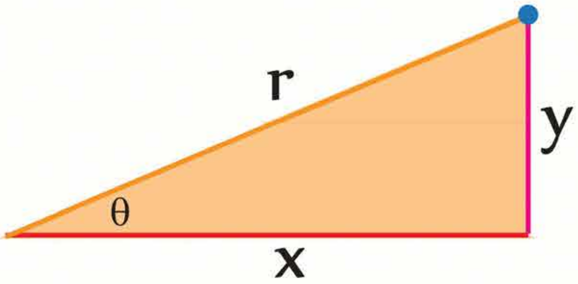

$$r^2 = 12^2 + 5^2$$

$$= \sqrt{169}$$

$$= 13$$

solving for the triangle using trigometric functions

Copyright ABRAHAM [2008] - All rights reserved

Spatial co-ordinate systems

Cubic

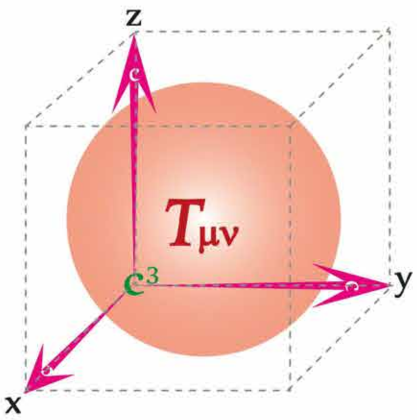

Spacetime is any mathematical co-ordinate system or model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional with time playing the role of a fourth dimension that is different from the spatial dimensions.

From a Euclidean space perspective, the universe has three spatial dimensions and one dimension of time [reflected by quantised angular momentum].

Spherical

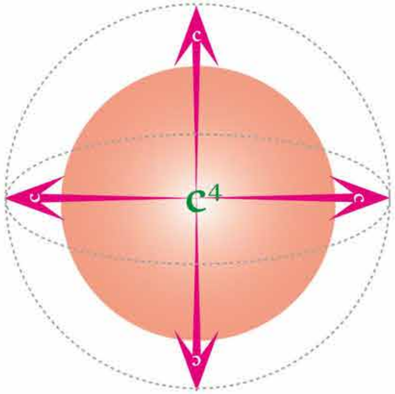

Tetryonics maps spatial co-ordinates through the momenta vectors of equilateral Energy

Cartesian Space-Time

Tetryonic Space-Time

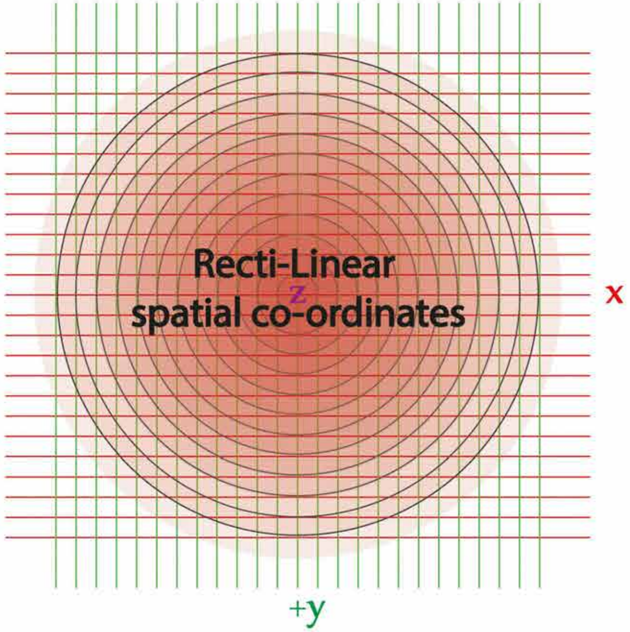

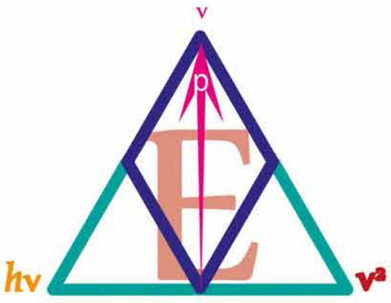

EM mass-ENERGY momenta are equilateral geometries

Mapping equilateral Energy geometries onto recti or curvi-linear spatial co-ordinate systems introduces mathematical complexity to a otherwise simplistic geometry for all EM mass-Energy-Matter interactions

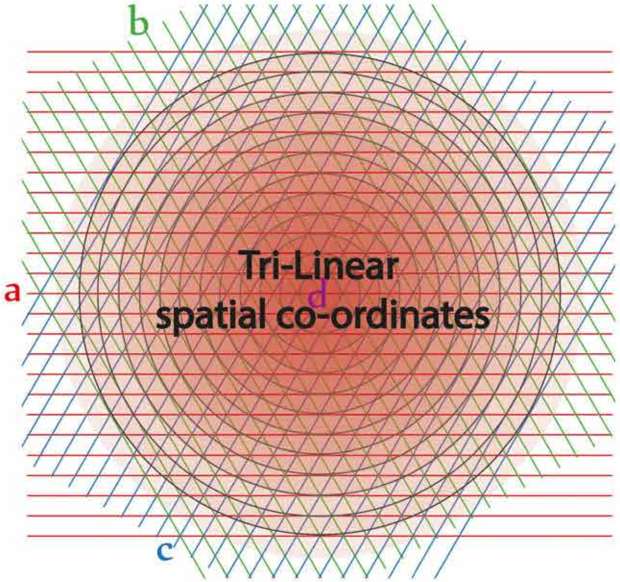

In physics spatial co-ordinates to date have been based on Cartesian co-ordinates when in fact Energy momenta follow a Tri-Linear co-ordinate geometry

Space-time co-ordinates

The propagation of Energy momenta forms distinct spatial co-ordinate systems

Where angles are typically measured in degrees (°)
or radians (rad), where 360° = 2π rad = τ radians.

vectors

1 Dimensional velocity

Euclidean

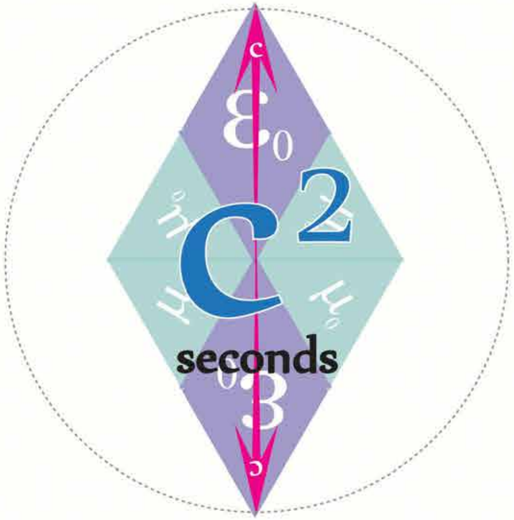
c
ε_0
c^2
μ_0
seconds
c

2 Dimensions velocity squared

$$m = \frac{E}{c^2}$$

mass-energy equivalence

mass-Matter

$$\frac{m}{c^2} = M$$

equivalence

$$T_{\mu\nu}$$

energy-Matter equivalence

$$M = \frac{E}{c^4}$$

GR fails differentiate betweem EM mass-ENERGY-Matter unlike Tetryonic theory

Cartesian co-ordinates

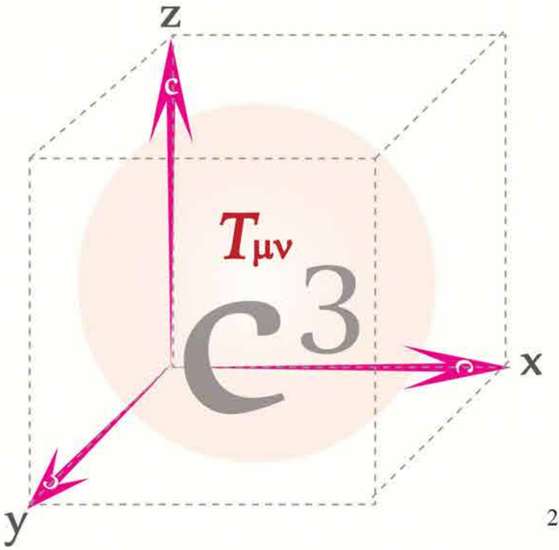
z
c
$T_{\mu\nu}$
c^3
c
x
c
y

3 Dimensions velocity cubed

Spherical

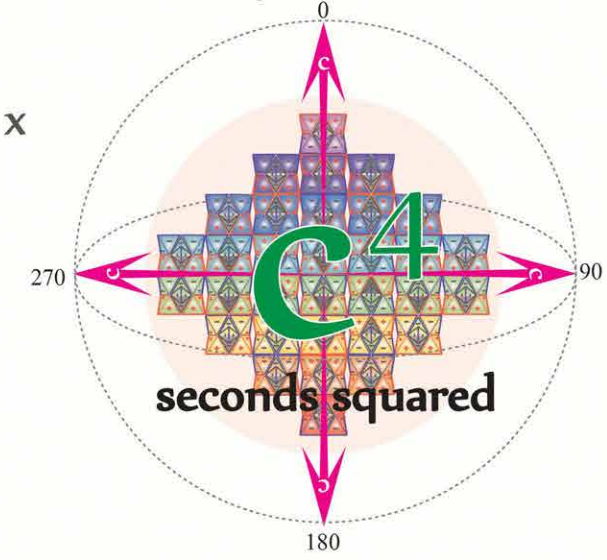
0
c
270
c
c^4
c
90
seconds squared
c
180

3 Dimensions quadrature velocity

Tetryonic co-ordinate systems

Energy per second

Energy per second squared

m

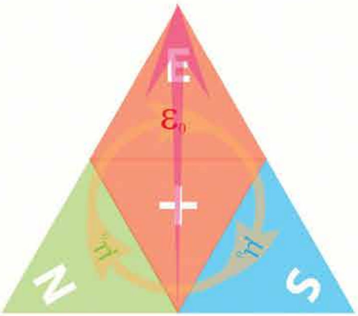

2D mass-energy geometries

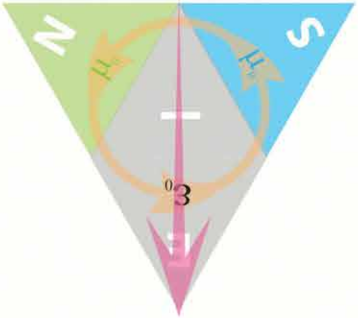

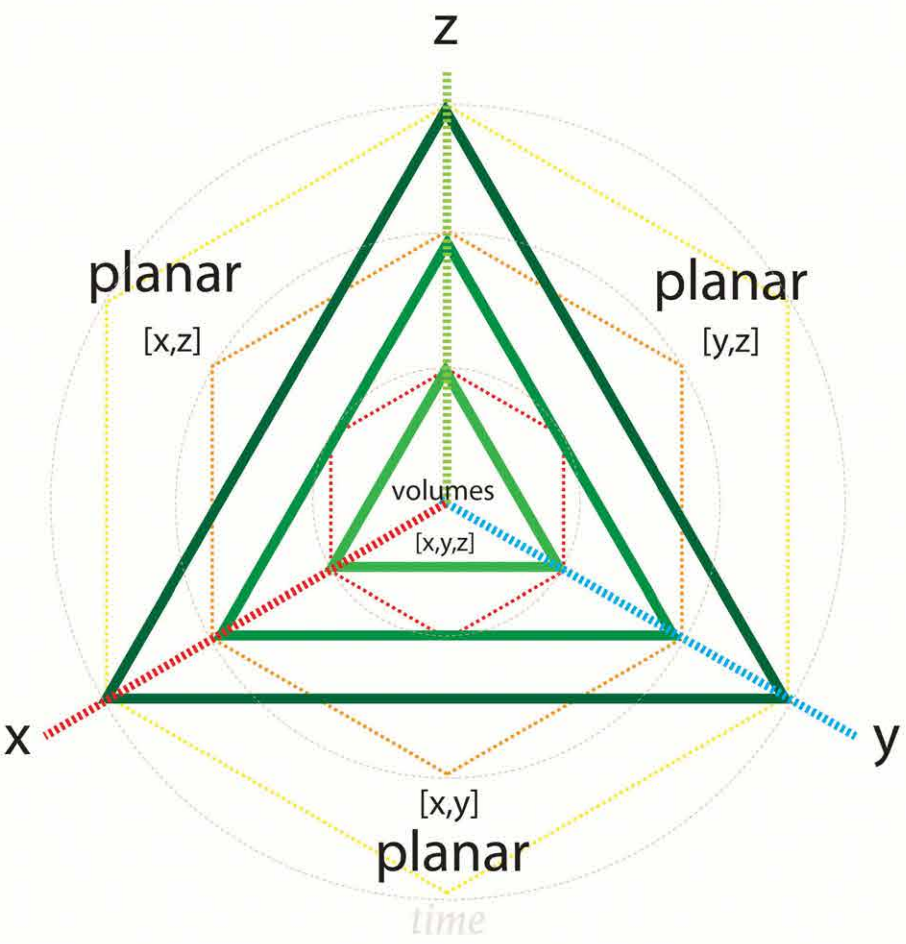
z
planar
[x,z]
planar
[y,z]
volumes
[x,y,z]
x
y
[x,y]
planar
time

M

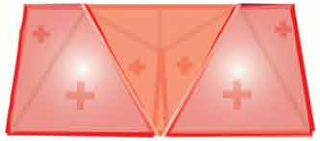

Proton

3D Matter topologies

Neutron

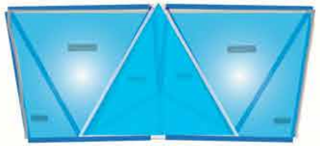

Planar mass-energy geometries have no z-components

Differentiation between 2D mass-energy & 3D Matter is key to extending our understanding of physics

Matter topologies have z-components

Copyright ABRAHAM [2008] - All rights reserved

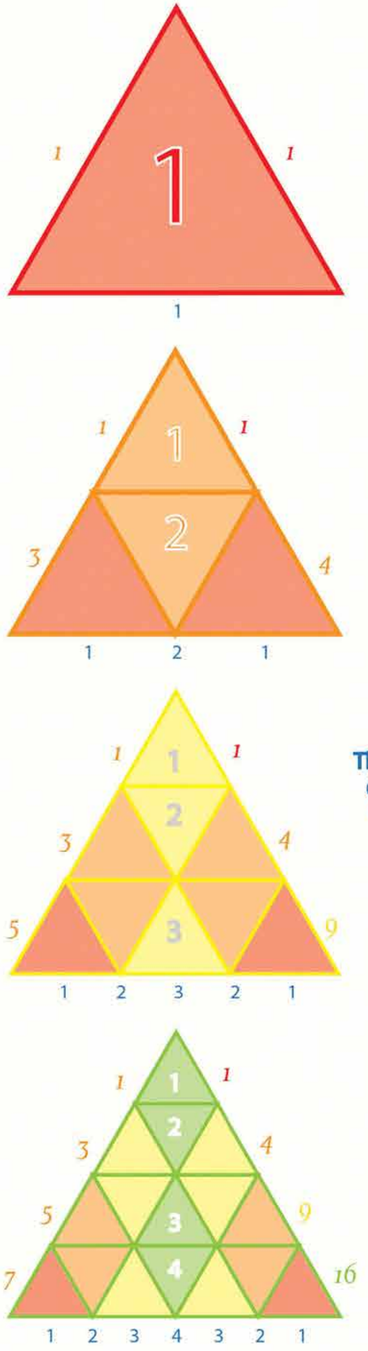
1
1 1
1
1 2
1 1
3 4
1 2 1
1 2 3
1 1
3 4
5 9
1 2 3 2 1
1 2 3 4
1 1
3 4
5 9
7 16
1 2 3 4 3 2 1

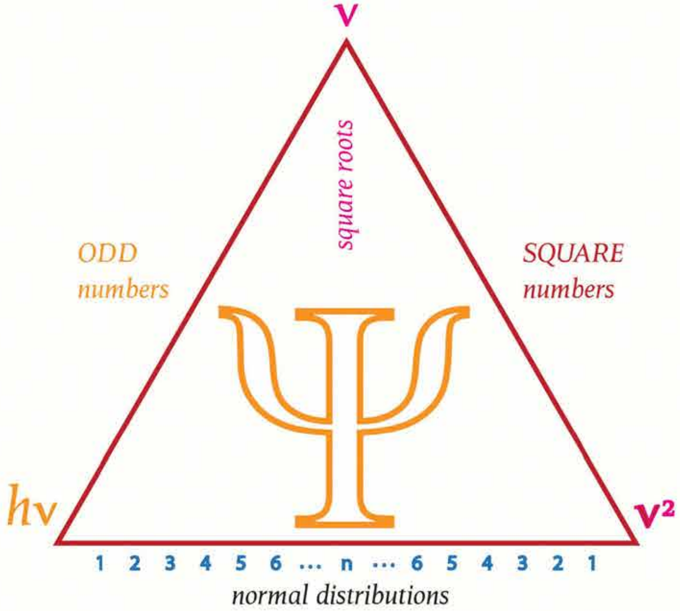
v
square roots
ODD numbers
SQUARE numbers
hv
v²
1 2 3 4 5 6 … n … 6 5 4 3 2 1
normal distributions

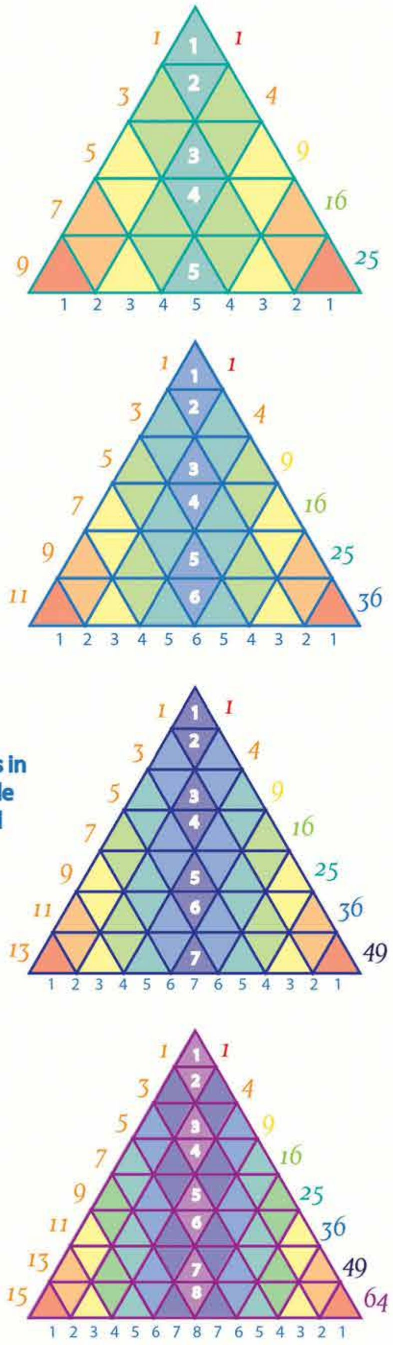
1 2 3 4 5
1 1
3 4
5 9
7 16
9 25
1 2 3 4 5 4 3 2 1
1 2 3 4 5 6
1 1
3 4
5 9
7 16
9 25
11 36
1 2 3 4 5 6 5 4 3 2 1
1 2 3 4 5 6 7
1 1
3 4
5 9
7 16
9 25
11 36
13 49
1 2 3 4 5 6 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8
1 1
3 4
5 9
7 16
9 25
11 36
13 49
15 64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Quantum probability distributions

The equilateral geometry and distribution of quantised Energy momenta provides the basis for all statistical probabilities in Quantum mechanics, thermodynamic & information entropy - including a solution to Heisenberg's Uncertainty Principle thus paving the way forward for a new understanding, and manipulation of physical phenomena at the quantum level

1

1 2 1

1 2 3 2 1

1 2 3 4 3 2 1

1 2 3 4 5 4 3 2 1

1 2 3 4 5 6 5 4 3 2 1

1 2 3 4 5 6 7 6 5 4 3 2 1

1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Normal distributions are extremely important in statistics, and are often used in the natural and social sciences for real-valued random variables whose distributions are not known

Copyright ABRAHAM [2008] - All rights reserved

Quantum Probability Distributions

The normal distribution is a probability distribution.
It is also called Gaussian distribution because it was discovered by Carl Friedrich Gauss.

x/n^2

Probabilities are the square of the Amplitude

n

$\Sigma\ x/n^2$

All probabilities are re-normalisable and sum to Unity

1

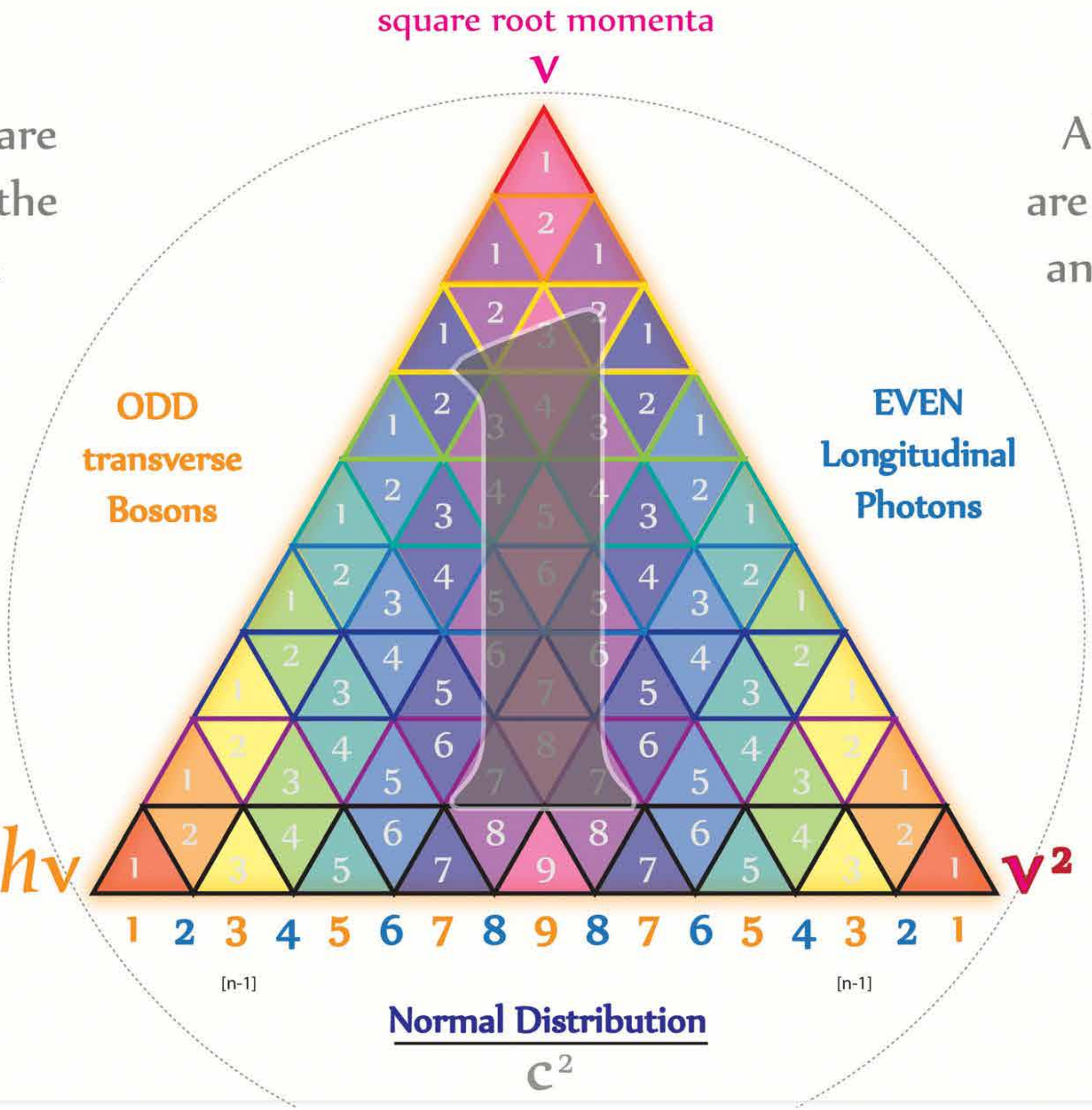
square root momenta
v
1
2
1 1
hv
v^2
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
[n-1]
Normal Distribution
c^2

ODD transverse Bosons

EVEN Longitudinal Photons

Planck quanta

Square energies

Quanta Distributions

$$2h\nu = E = hf$$

Quantum Levels

transverse bosons

Quantum Probabilities

longitudinal photons

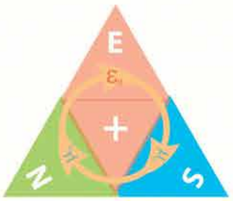

W+

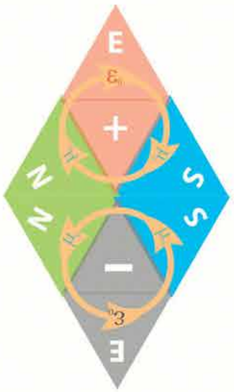
E
ε₀
+
N
S
μ₀
N
S
−
ε₀
E

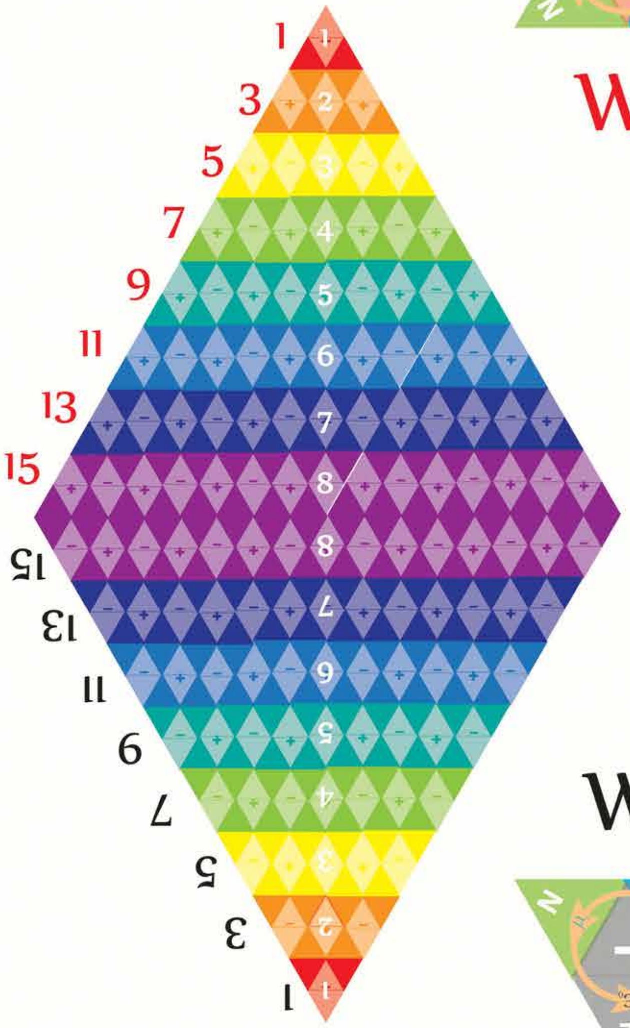
1
3
5
7
9
11
13
15
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1
15
13
11
9
7
5
3
1

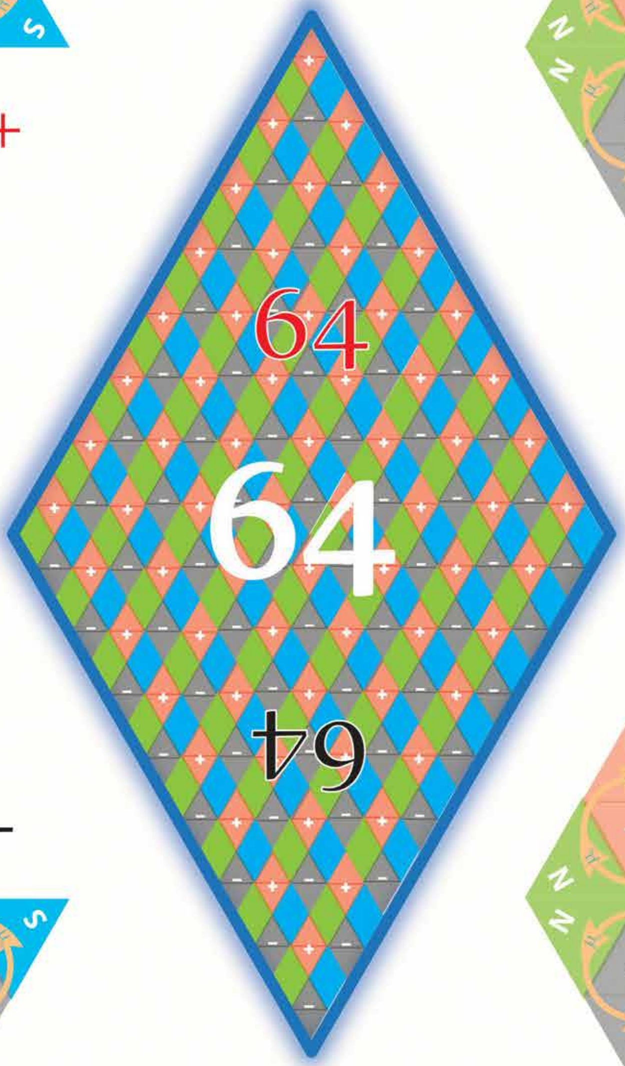

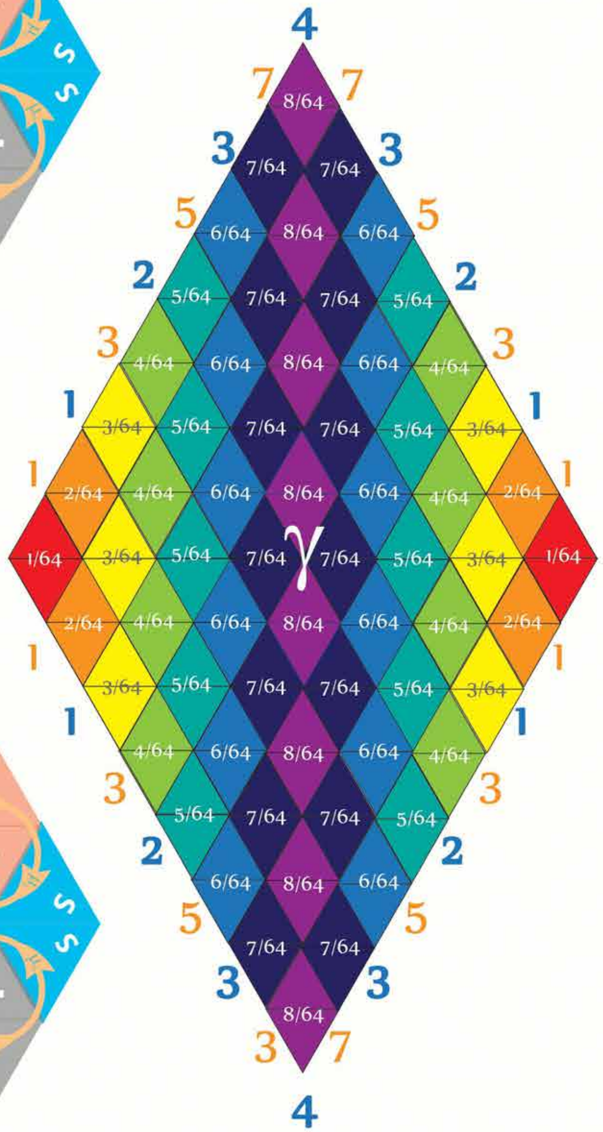
4
7
7
8/64
3
3
7/64
7/64
5
5
6/64
8/64
6/64
2
2
5/64
7/64
7/64
5/64
3
3
4/64
6/64
8/64
6/64
4/64
1
1
3/64
5/64
7/64
7/64
5/64
3/64
1
1
2/64
4/64
6/64
8/64
6/64
4/64
2/64
1/64
3/64
5/64
7/64
γ
7/64
5/64
3/64
1/64
2/64
4/64
6/64
8/64
6/64
4/64
2/64
1
1
3/64
5/64
7/64
7/64
5/64
3/64
1
1
4/64
6/64
8/64
6/64
4/64
3
3
5/64
7/64
7/64
5/64
2
2
6/64
8/64
6/64
5
5
7/64
7/64
3
3
8/64
3
7
4

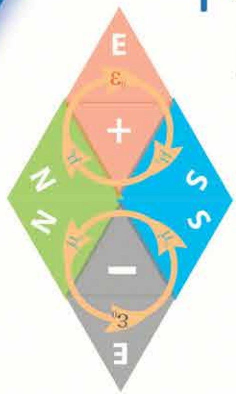
E
ε₀
+
N
S
μ₀
N
S
−
ε₀
E

W-

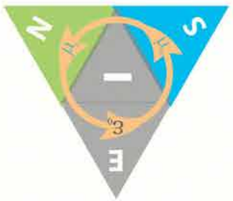

Wavefunctions

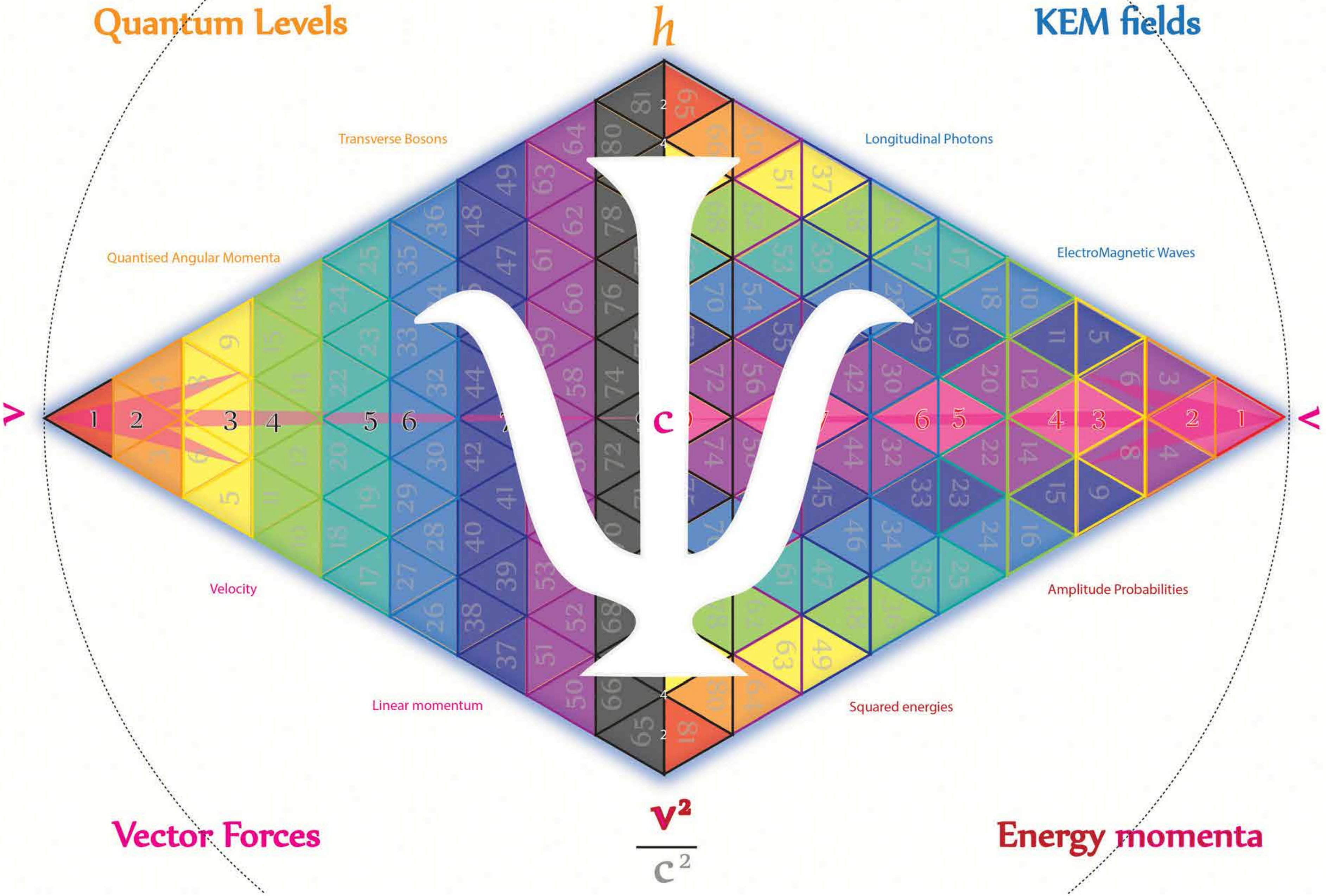
Quantum Levels
h
KEM fields
Transverse Bosons
Longitudinal Photons
Quantised Angular Momenta
ElectroMagnetic Waves
v
1 2 3 4 5 6 7
c
7 6 5 4 3 2 1
v
Velocity
Amplitude Probabilities
Linear momentum
Squared energies
Vector Forces
$\frac{v^2}{c^2}$
Energy momenta

EM waveforms

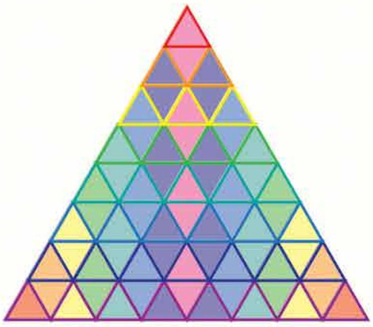

All EM waveforms can be measured by either their Transverse EM masses [Bosons] or their Longitudinal EM masses [Photons]

BOSONS

Bosons Planck quanta

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

vs.

PHOTONS

EM waves Planck quanta

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

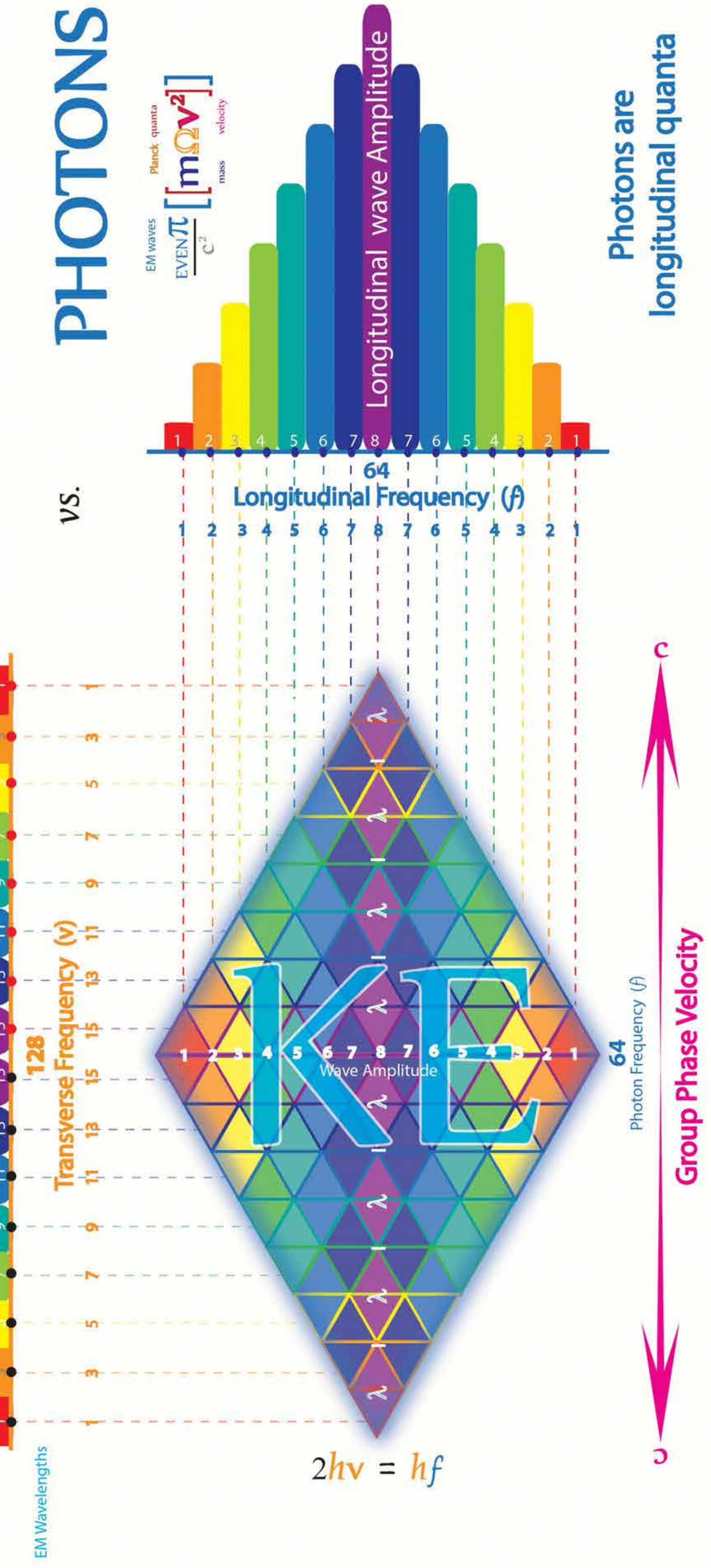
Longitudinal wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Longitudinal Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
KE
Wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
λ
64
Photon Frequency (f)
Group Phase Velocity
c
c
128
Transverse Frequency (v)
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
EM Wavelengths
2hv = hf

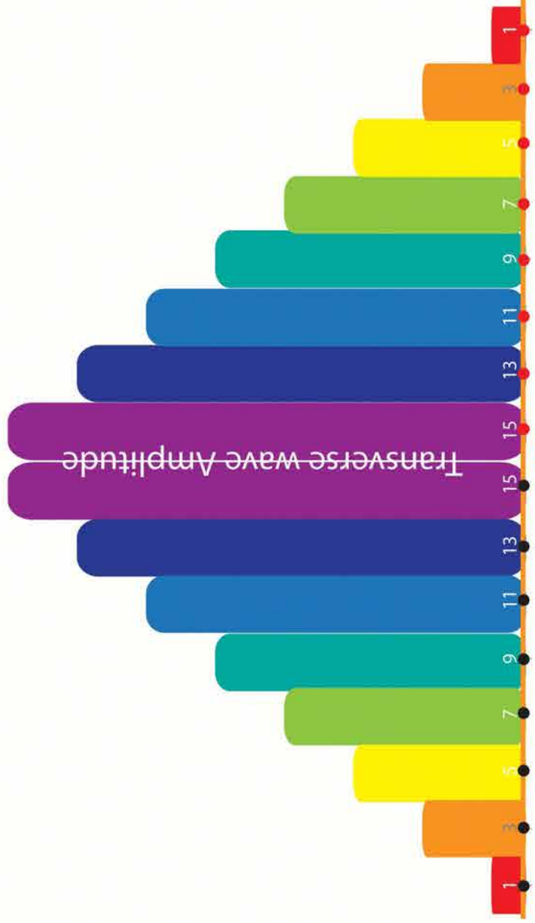
Transverse wave Amplitude
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1

Photons are longitudinal quanta

Bosons are transverse quanta

Copyright ABRAHAM [2008] - All rights reserved

Quantum Energy distributions

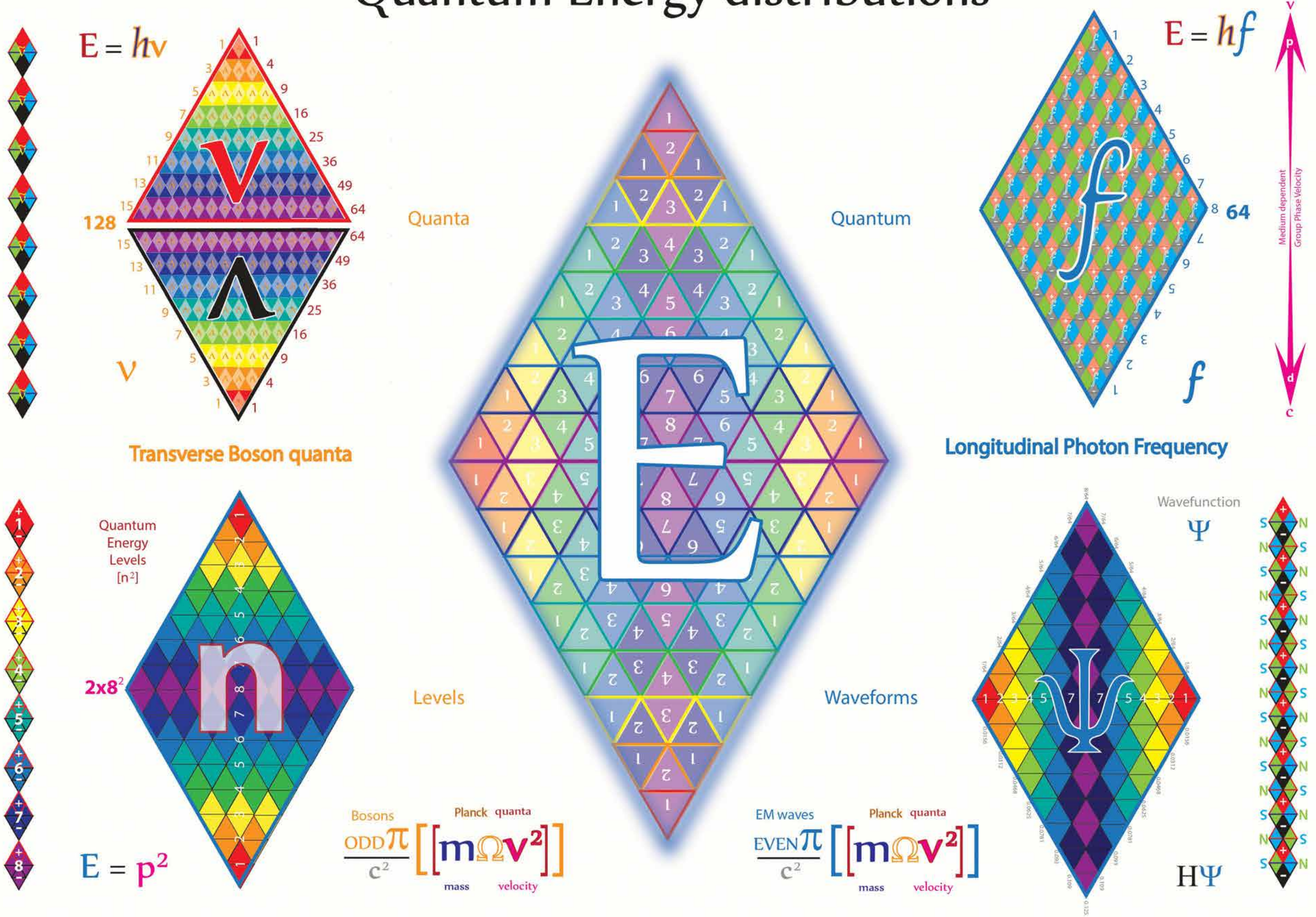
E = hν
ν
λ
128
ν
Quanta
Transverse Boson quanta
E
Quantum
E = hf
f
64
f
Medium dependent
Group Phase Velocity
v
p
d
c
Longitudinal Photon Frequency
Quantum
Energy
Levels
[n²]
2x8²
n
E = p²
Levels
Waveforms
Wavefunction
Ψ
HΨ
S
N

Bosons Planck quanta

$$\frac{\mathrm{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

EM waves Planck quanta

$$\frac{\mathrm{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Normal Distributions

The Gaussian distribution sometimes informally called the bell curve.

Pierre de Fermat

Pierre de Fermat is given credit
for early developments that led to
infinitesimal calculus.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

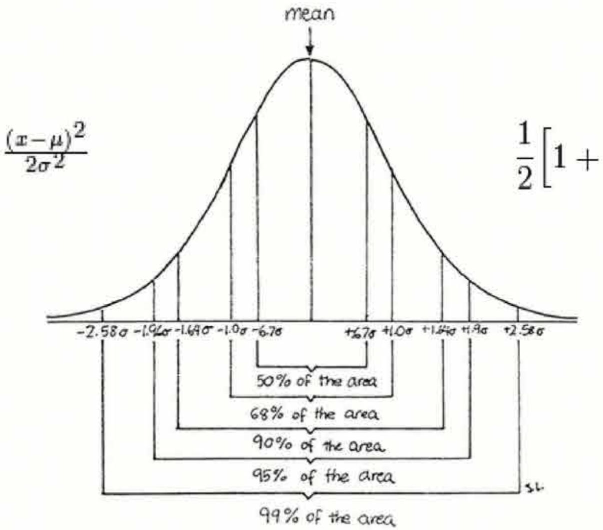
mean
-2.58σ
-1.96σ
-1.69σ
-1.0σ
-6.7σ
+6.7σ
+1.0σ
+1.69σ
+1.9σ
+2.58σ
50% of the area
68% of the area
90% of the area
95% of the area
99% of the area
s.l.

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

Leonhard Euler

Leonhard Euler developed a formula
which links complex exponentiation
with trigonometric functions

**A bell shaped curve defines the standard normal distribution,
in which the probability of observing a point is greatest near the average,
and declines rapidly as one moves away from the mean.**

all ODD numbers are

$$a^2 - b^2 = [a-b].[a+b]$$

the difference of two squares

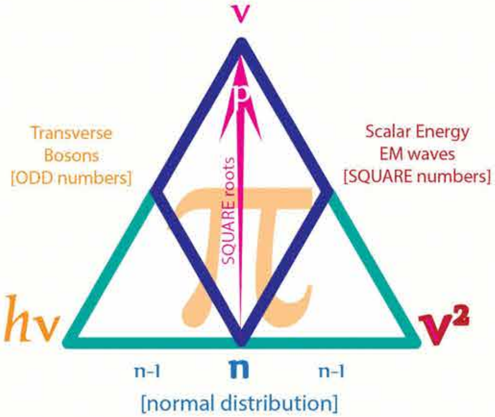
v
p
SQUARE roots
Transverse
Bosons
[ODD numbers]
Scalar Energy
EM waves
[SQUARE numbers]
hv
v²
n-1
n
n-1
[normal distribution]

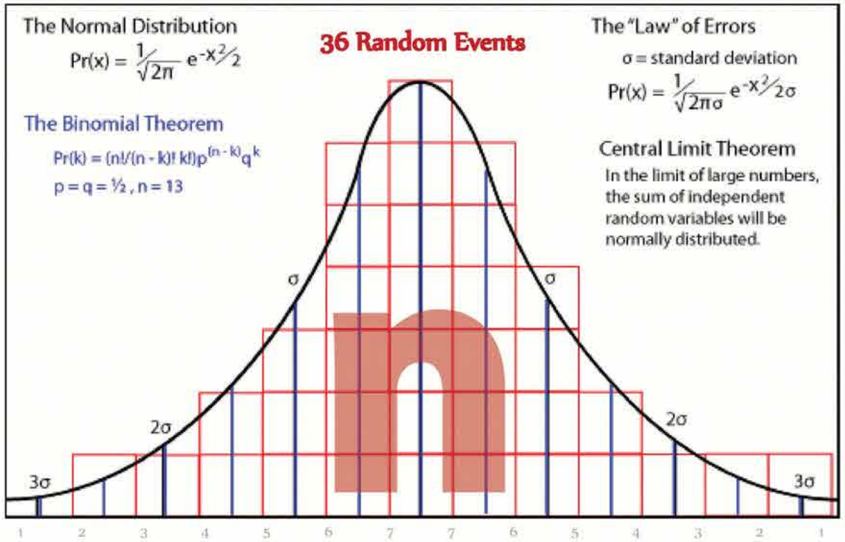

The Normal Distribution

$$\Pr(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

The Binomial Theorem

$\Pr(k) = (n!/(n-k)!\,k!)p^{(n-k)}q^k$

$p = q = ½, n = 13$

The "Law" of Errors

σ = standard deviation

$$\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-x^2/2\sigma}$$

Central Limit Theorem

In the limit of large numbers, the sum of independent random variables will be normally distributed.

$$e^{i\pi} = -1$$

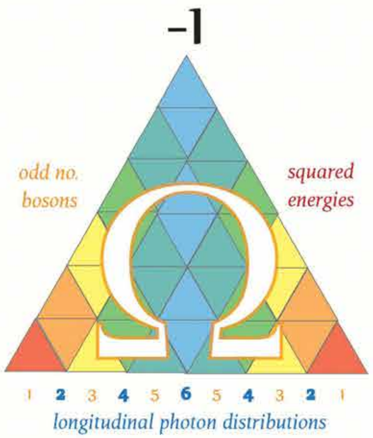
-1
odd no.
bosons
squared
energies
Ω
1
2
3
4
5
6
5
4
3
2
1
longitudinal photon distributions

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function

Copyright ABRAHAM [2008] - All rights reserved

Fundamental theorem of Energy momenta

A nth level scalar energy momenta waveform has exactly n linear momentum in unit circle co-ordinate systems
(with Longitudinal and Transverse equilateral Planck waveforms being orthogonal to each other)

scalar energies

E

mv^2

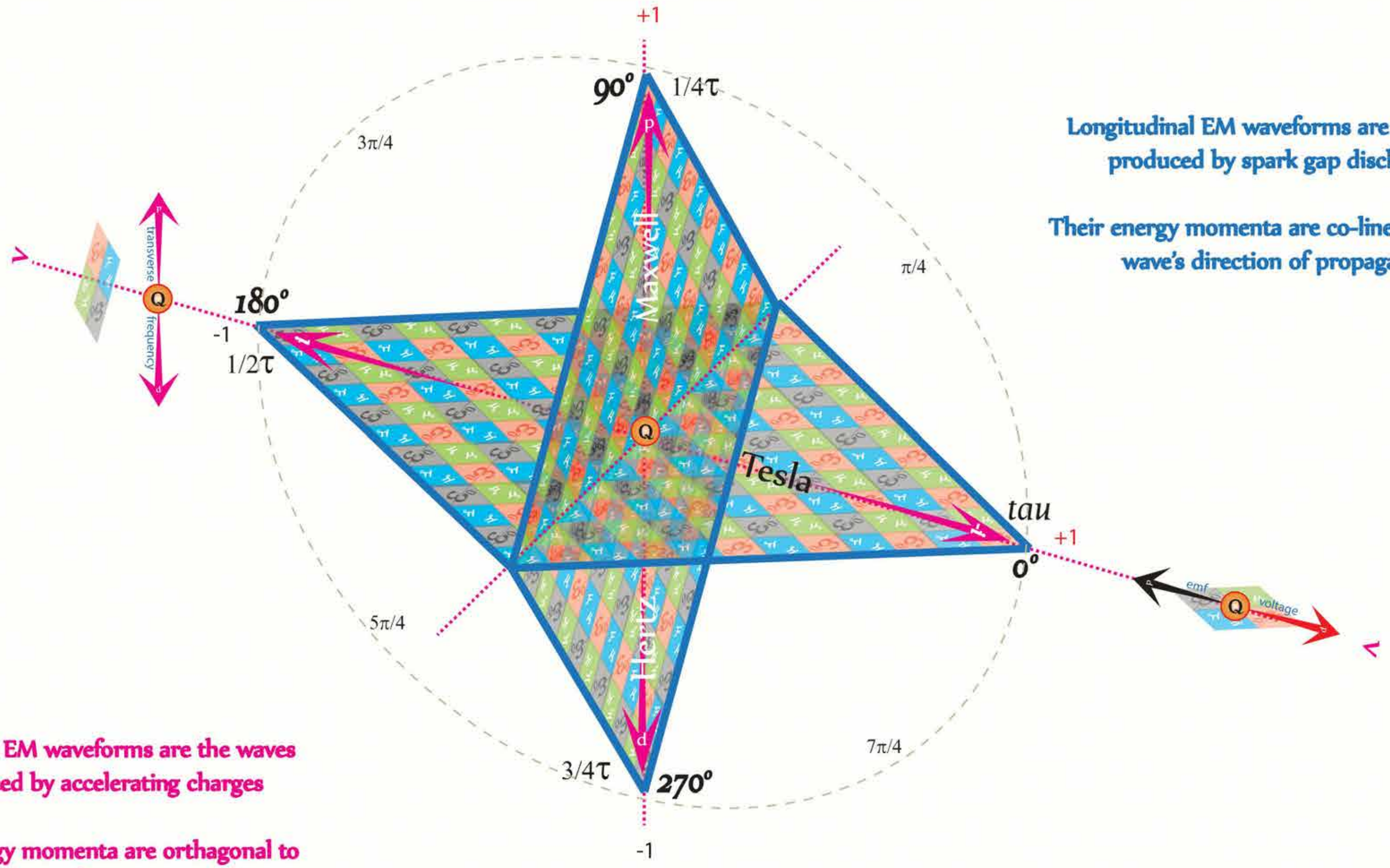
+1
90°
1/4τ
3π/4
π/4
transverse
frequency
Q
180°
-1
1/2τ
Maxwell
Q
Tesla
tau
+1
0°
emf
voltage
Q
5π/4
Hertz
7π/4
3/4τ
270°
-1

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthogonal to the direction of wave propagation

[Mathematics] *i or j* *[Electrical theory]*

square root of negative one

p

mv

linear momentum

Velocity and Time dependent EM fields

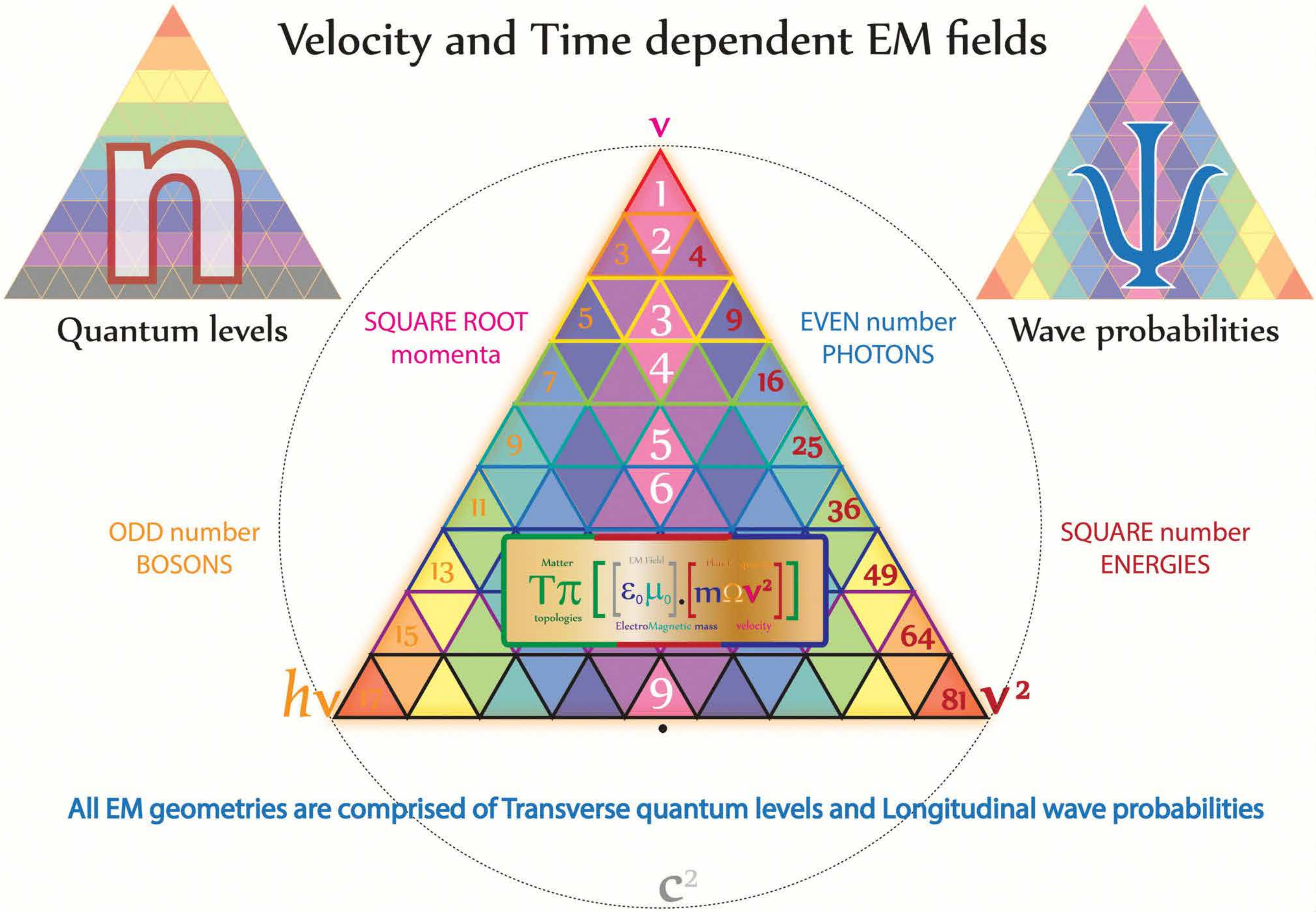
n
Ψ
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
13
49
15
64
17
9
81
hv
v²
c²
Matter
Tπ [[ε₀μ₀].[mΩv²]]
topologies
EM Field
Planck quanta
ElectroMagnetic mass
velocity

Quantum levels

Wave probabilities

SQUARE ROOT
momenta

EVEN number
PHOTONS

ODD number
BOSONS

SQUARE number
ENERGIES

All EM geometries are comprised of Transverse quantum levels and Longitudinal wave probabilities

Matrices

Matrices are a key tool in linear algebra.
One use of matrices is to represent linear transformations,
which are higher-dimensional analogs of linear functions
of the form f(x) = cx, where c is a constant; matrix multiplication
corresponds to composition of linear transformations.

Further developing equilateral Matrices and tensor mathematics to
reflect the 2D geometry of EM , KEM and GEM quantum fields,
along with the geometric quantisation of mass-energiy momenta
and their energy distributions allows for field interactions
to be accurately visualised and modelled

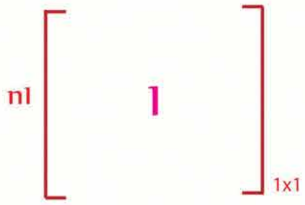

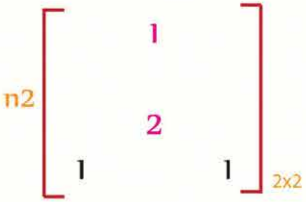

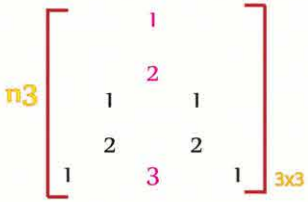
n3
1
2
1 1
2 2
1 3 1
3x3

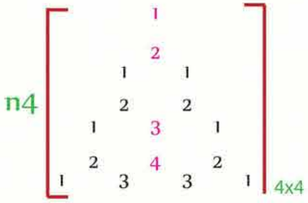
n4
1
2
1 1
2 2
1 3 1
2 4 2
1 3 3 1
4x4

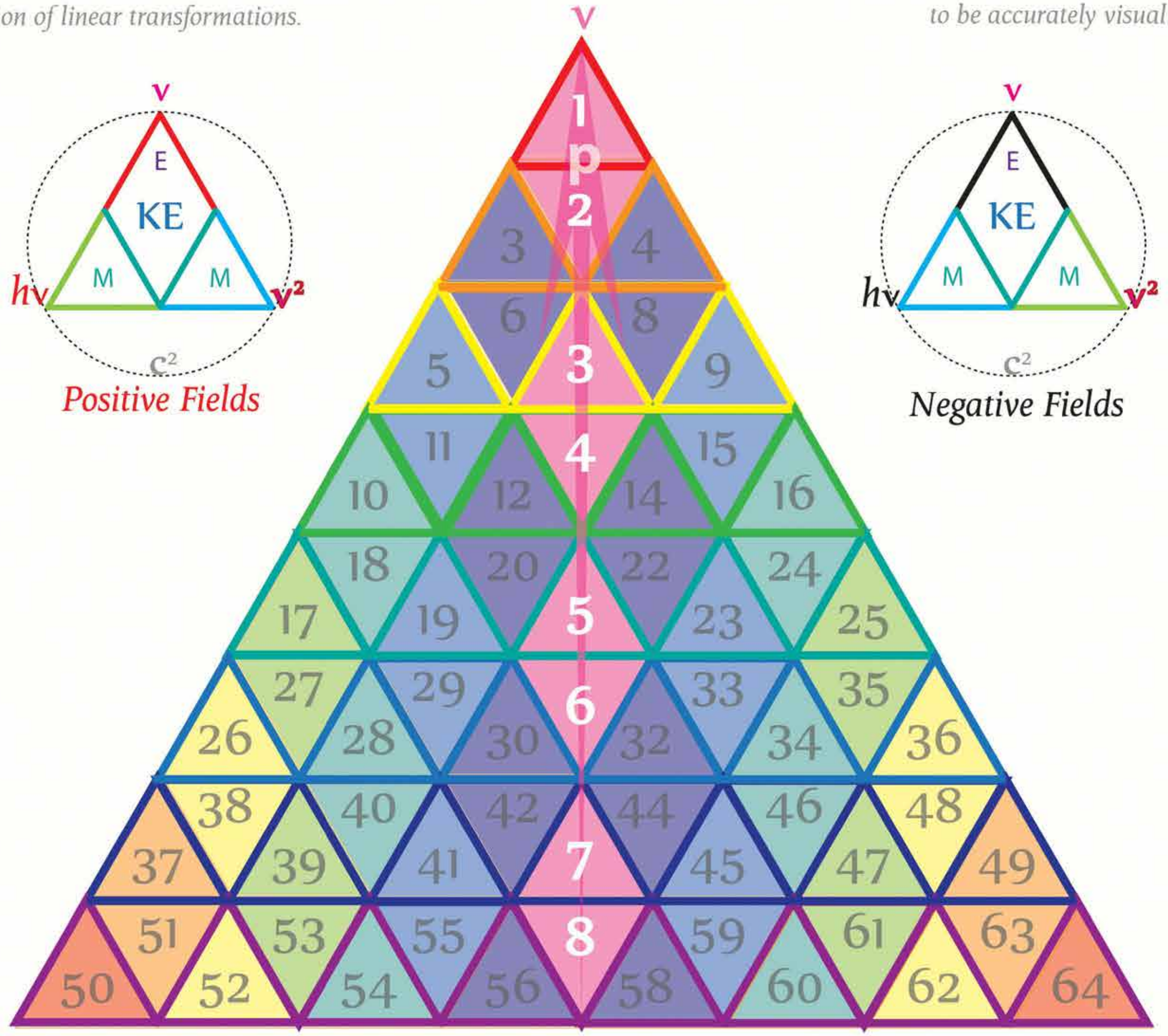
v
E
KE
M
M
hv
v²
c²
Positive Fields
v
E
KE
M
M
hv
v²
c²
Negative Fields
1
p
2
3
4
5
6
7
8
3 4
5 6 8 9
10 11 12 14 15 16
17 18 19 20 22 23 24 25
26 27 28 29 30 32 33 34 35 36
37 38 39 40 41 42 44 45 46 47 48 49
50 51 52 53 54 55 56 58 59 60 61 62 63 64

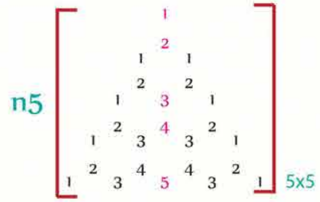

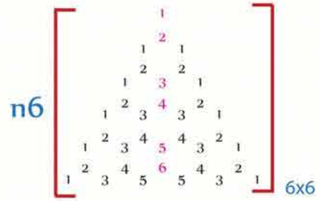

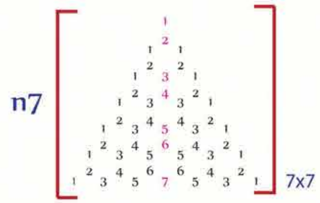

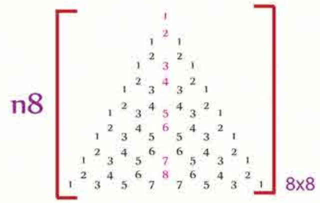

Tensors are geometric entities introduced into mathematics and physics to extend the notion of scalars, geometric vectors, and matrices to increasingly higher orders.

Modifying Square matrices to reflect the equialteral geometries of Tetryonic fields allows for the accurate geometric modelling of all Scalar & Vector fields along with their varied intrinsic quantum energies and physical properties

Copyright ABRAHAM [2008] - All rights reserved

Energy momenta Tensors

momenta
(a property of Energy)
is converative

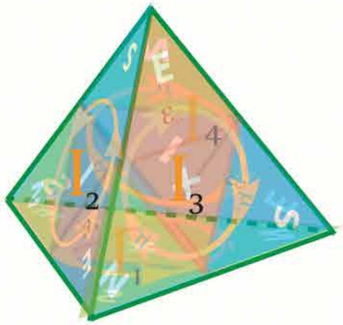

Matter
(a geometric property)
is NOT conservative

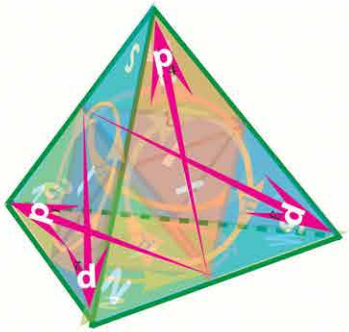

$$p_M \longrightarrow (E, p1, p2, p3, p4).$$

All standing-wave Matter topologies can be modelled using its Tetryonic charge energy momenta Tensors

with an additional Kinetic EM energy-momenta tensor required for Matter in motion

$$p_{KEM} \longrightarrow (E, p5).$$

2D+1 [SR] mass-energy momenta can be folded into 3D+1 [GR] Matter
that can be modelled using 4 Energy-momenta tensors

Positive standing-wave
Matter

$$E = \sum_{\text{all fascia}} E$$

Energy of a massive particle
is the total of all Planck quanta
(compton frequency) in a 3D geometry

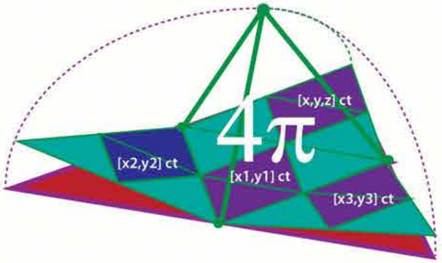

$$\vec{p} = \sum_{\text{all fascia}} \vec{p}$$

Total Momentum
is the total of all quanta
linear Momenta in a 3D particle

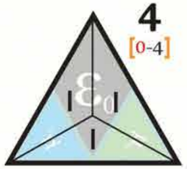

Negative standing-wave
Matter

Relativistic Matter
in motion

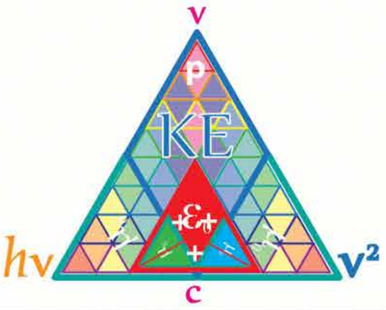

$$I = \sum_{\text{all fascia}} I$$

Inertial mass
is the total of all inertia
in a 3D particle

Photons have
null
energy-momentum
tensors

Note that both
2D mass-energies [Special Relativity]
and 3D Matter [General Relativity]
have distinct Energy momenta

It is the 3D Tetrahedral topologies that
provides a definative basis for Matter

Relativistic Matter
in motion

v
p
KE
ε
hv
v²
c

Tetryonic Energy momenta tensors should not be confused with Four vector tensors
which map energy-momenta vectors in 3D spatial[cartesian] co-ordinate systems

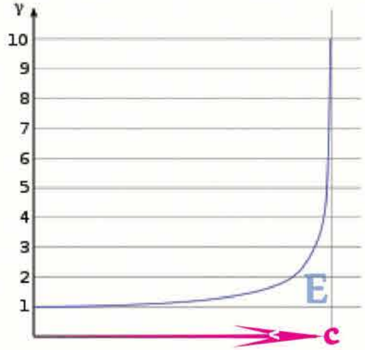

Relativistic Lorentz correction factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity
It is often incorrectly graphed [as pictured to the right] as an increasing function of velocity approaching but never reaching c

**'c' is the maximum velocity acheivable through the electrical acceleration of particles
is NOT the maximum velocity possible for Matter in motion**

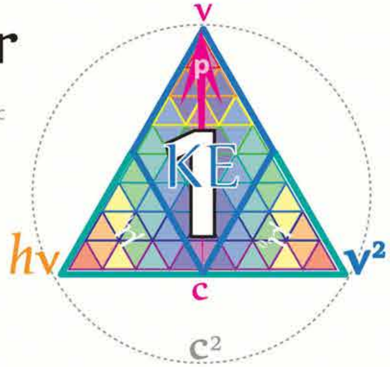

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

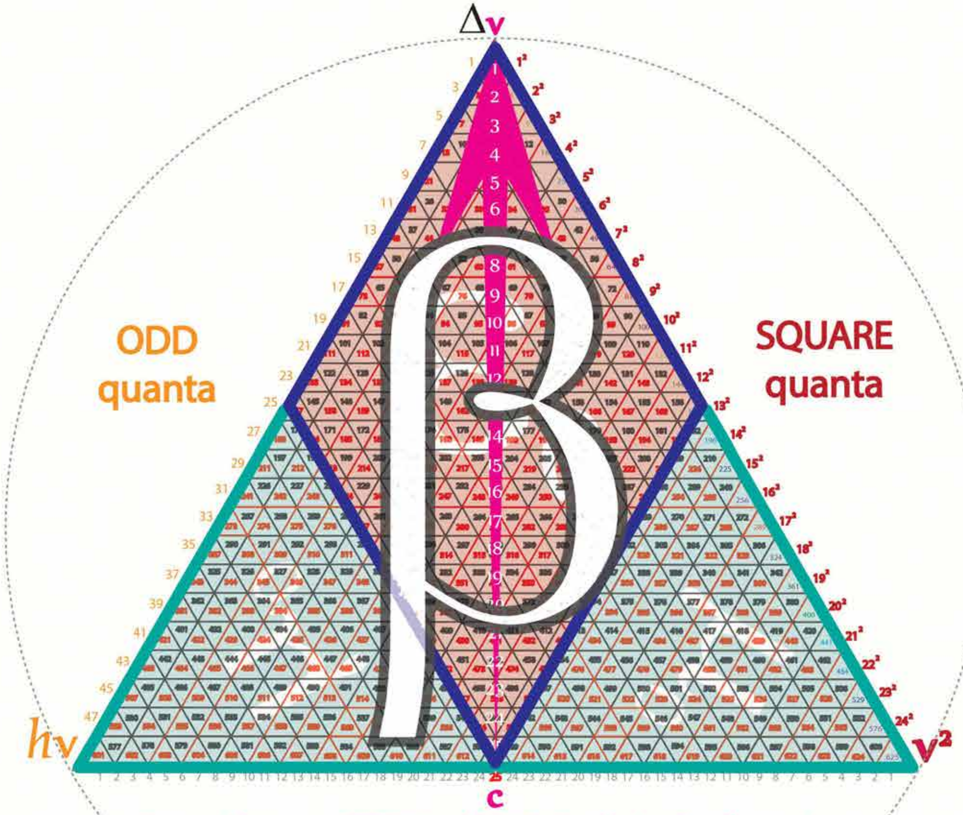

Lorentz factors are LINEAR and SCALAR velocity related corrections to the relativistic mass-energy momenta content of any physical system accelerated by EM Forces

$$\frac{\Gamma}{c^2}$$

Lorentz corrections

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

$$\Delta\frac{\gamma}{c^2}$$ Wavelength contraction

$$\Delta\frac{mv}{c^2}$$ Relativistic momentum

Time Dilation $$\Delta\frac{\Omega}{c^2}$$

Relativistic mass-energy $$\Delta\frac{hv}{c^2}$$

Equilateral energies

Any measurements of mass without a Space-Time co-ordinate system are measurements of Energy

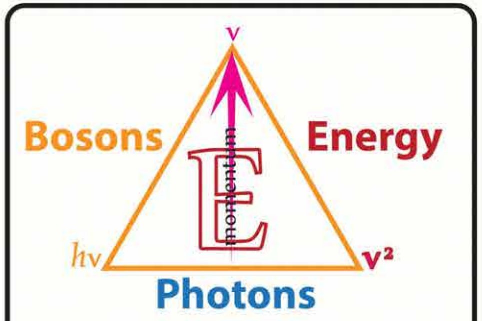
v
Bosons
Energy
E
momentum
hv
v²
Photons

form the basis for all charged energy geometries and their associated quanta distributions

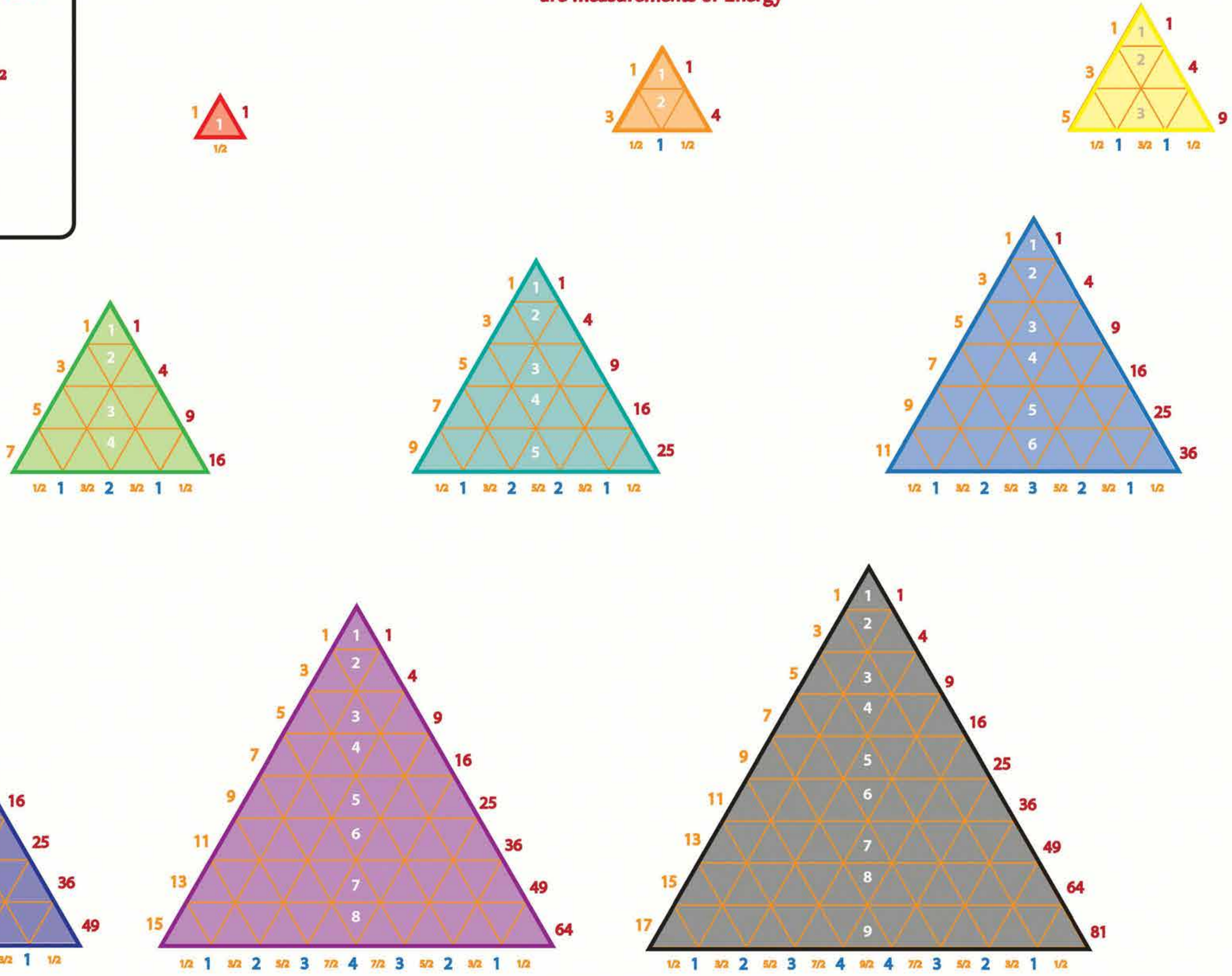
1 1 1
1/2
1 1 1 2
3 4
1/2 1 1/2
1 1 1 2 3
3 4
5 9
1/2 1 3/2 1 1/2
1 1 3 4 5 9 7 16
1 2 3 4
1/2 1 3/2 2 3/2 1 1/2
1 1 3 4 5 9 7 16 9 25
1 2 3 4 5
1/2 1 3/2 2 5/2 2 3/2 1 1/2
1 1 3 4 5 9 7 16 9 25 11 36
1 2 3 4 5 6
1/2 1 3/2 2 5/2 3 5/2 2 3/2 1 1/2
1 1 3 4 5 9 7 16 9 25 11 36 13 49
1 2 3 4 5 6 7
1/2 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 1/2
1 1 3 4 5 9 7 16 9 25 11 36 13 49 15 64
1 2 3 4 5 6 7 8
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2
1 1 3 4 5 9 7 16 9 25 11 36 13 49 15 64 17 81
1 2 3 4 5 6 7 8 9
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

Copyright ABRAHAM [2008] - All rights reserved

Electromagnetic mass

2D mass-energy is the surface integral of 3D Matter

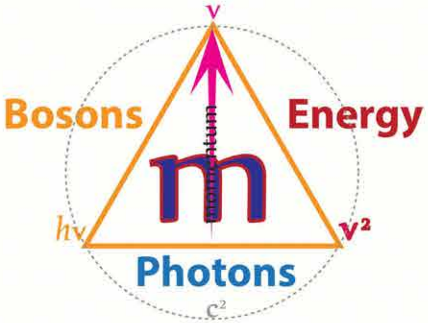
v
Bosons
Energy
m
hv
v²
Photons
c²

Time specific measurements of equilateral ENERGY momenta geometries forms the basis for EM mass

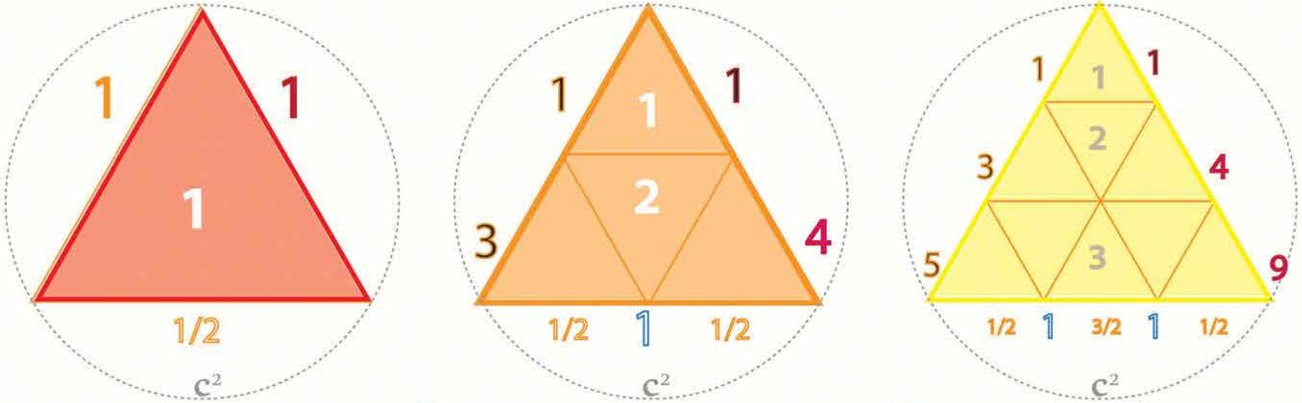
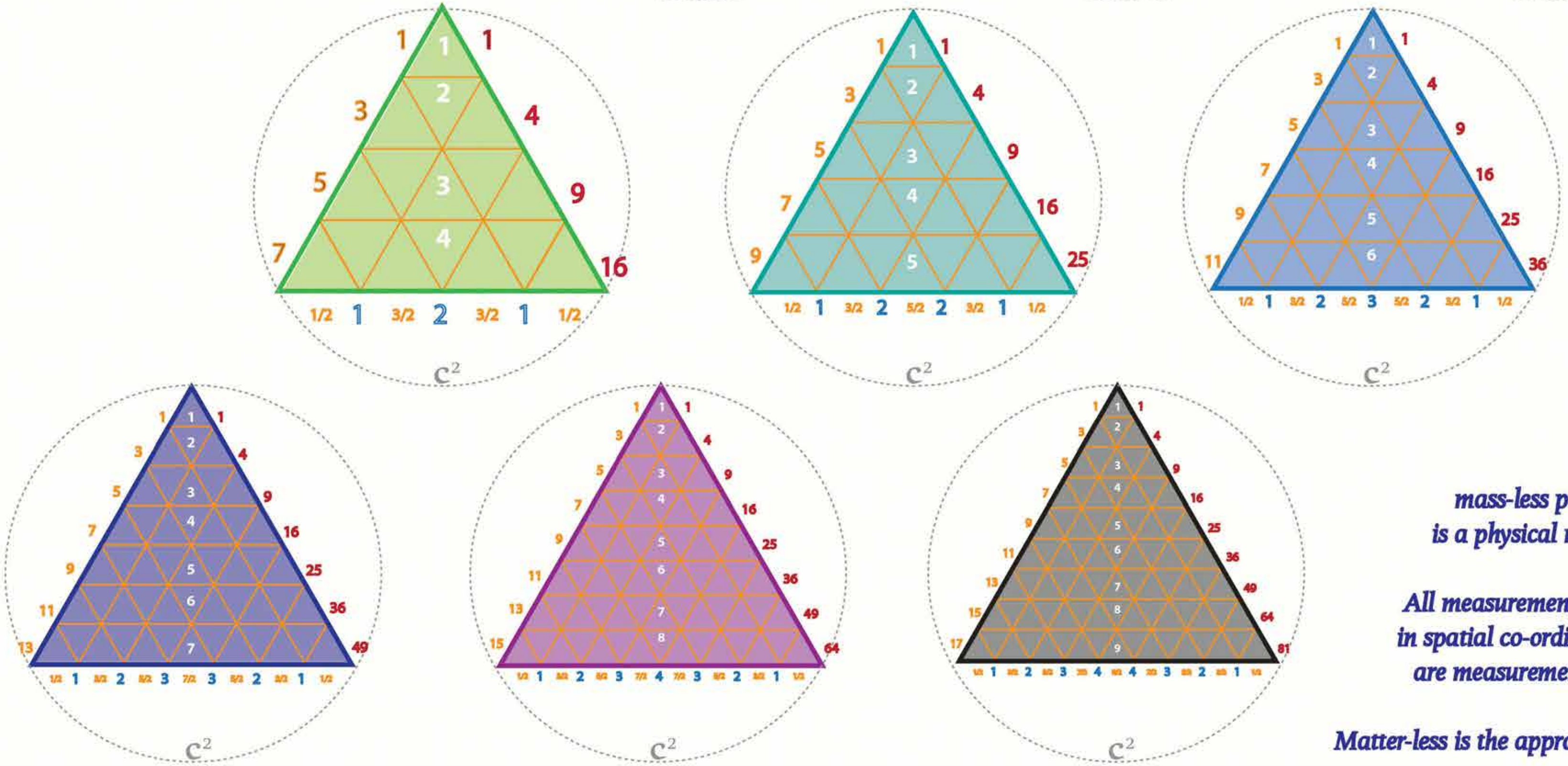
1
1
1
1/2
c²
1
1
1
2
3
4
1/2
1
1/2
c²
1
1
1
2
3
3
4
5
9
1/2
1
3/2
1
1/2
c²
1/2
1
3/2
2
3/2
1
1/2
1
4
9
16
3
5
7
c²

mass-less particles is a physical misnomer

All measurements of energy in spatial co-ordinate systems are measurements of mass

Matter-less is the appropriate terminology

Standing-wave Matter

Matter is a higher order 3D topology created by standing wave mass-energies

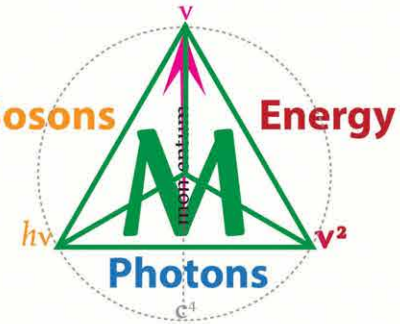
v
Bosons
Energy
momentum
M
hν
v²
Photons
c⁴

The equilateral mass-ENERGY content of 4nπ Tetrahedral standing-wave topologies forms the basis of Matter

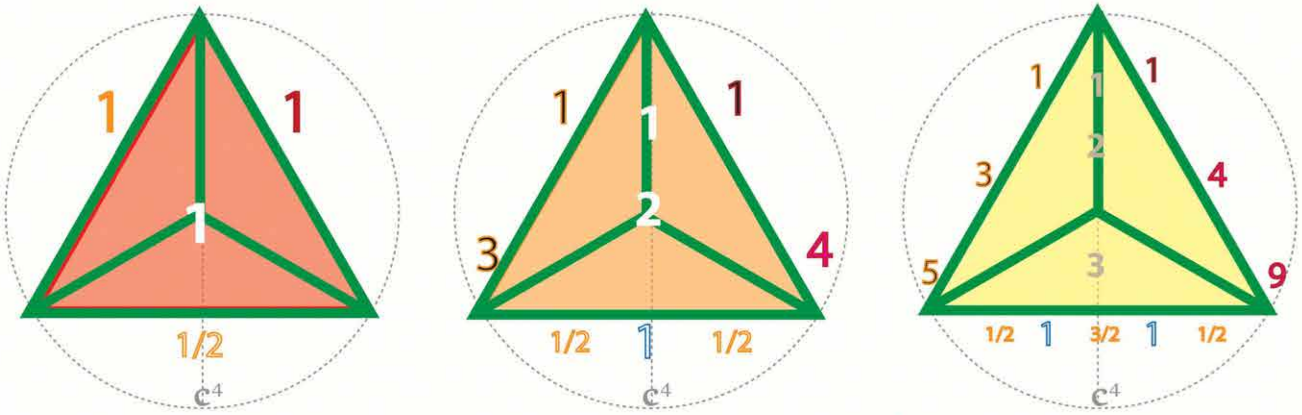
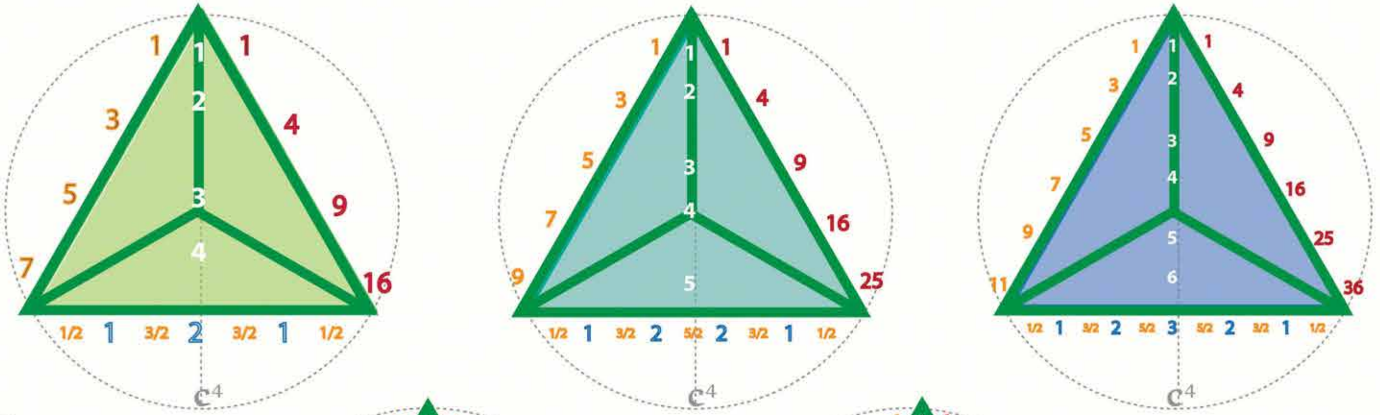
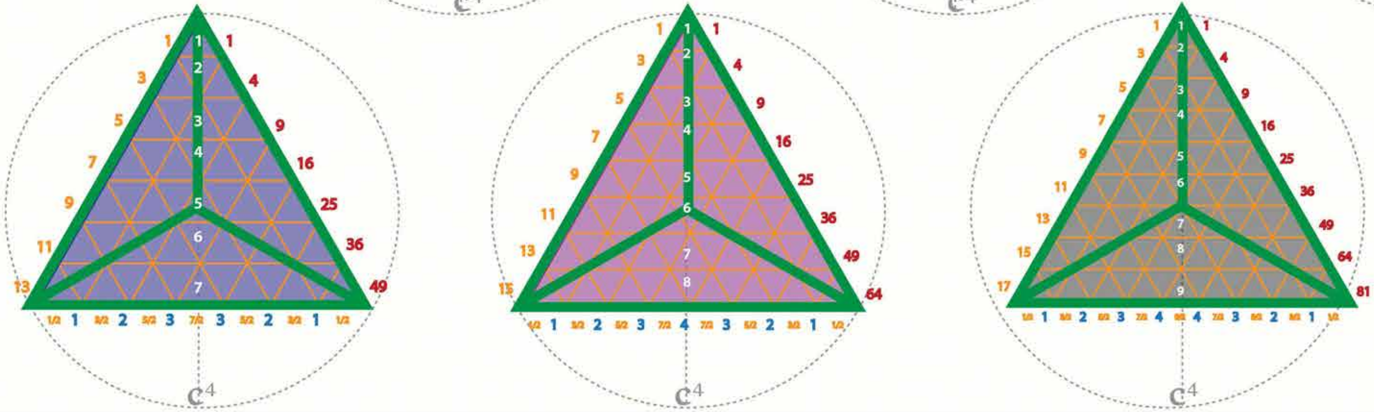
1
1
1
1/2
c⁴
1
1
1
2
3
4
1/2
1
1/2
c⁴
1
1
1
2
3
3
5
4
9
1/2
1
3/2
1
1/2
c⁴
c⁴
c⁴
c⁴
c⁴
c⁴
c⁴

Only mass-ENERGIES contained in the tetryonic fascia of standing-wave topologies contribute to weight

Copyright ABRAHAM [2008] - All rights reserved

Periodic element nuclei

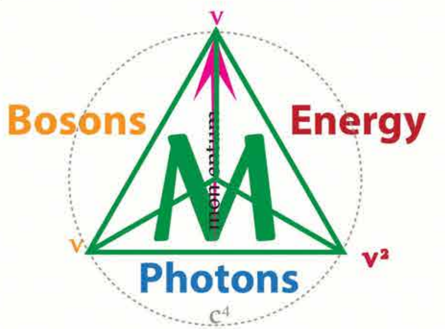
v
Bosons
Energy
momentum
M
v
v²
Photons
c⁴

The mass-ENERGY content of
4nπ Tetrahedral standing-wave topologies
forms the basis of all periodic elements

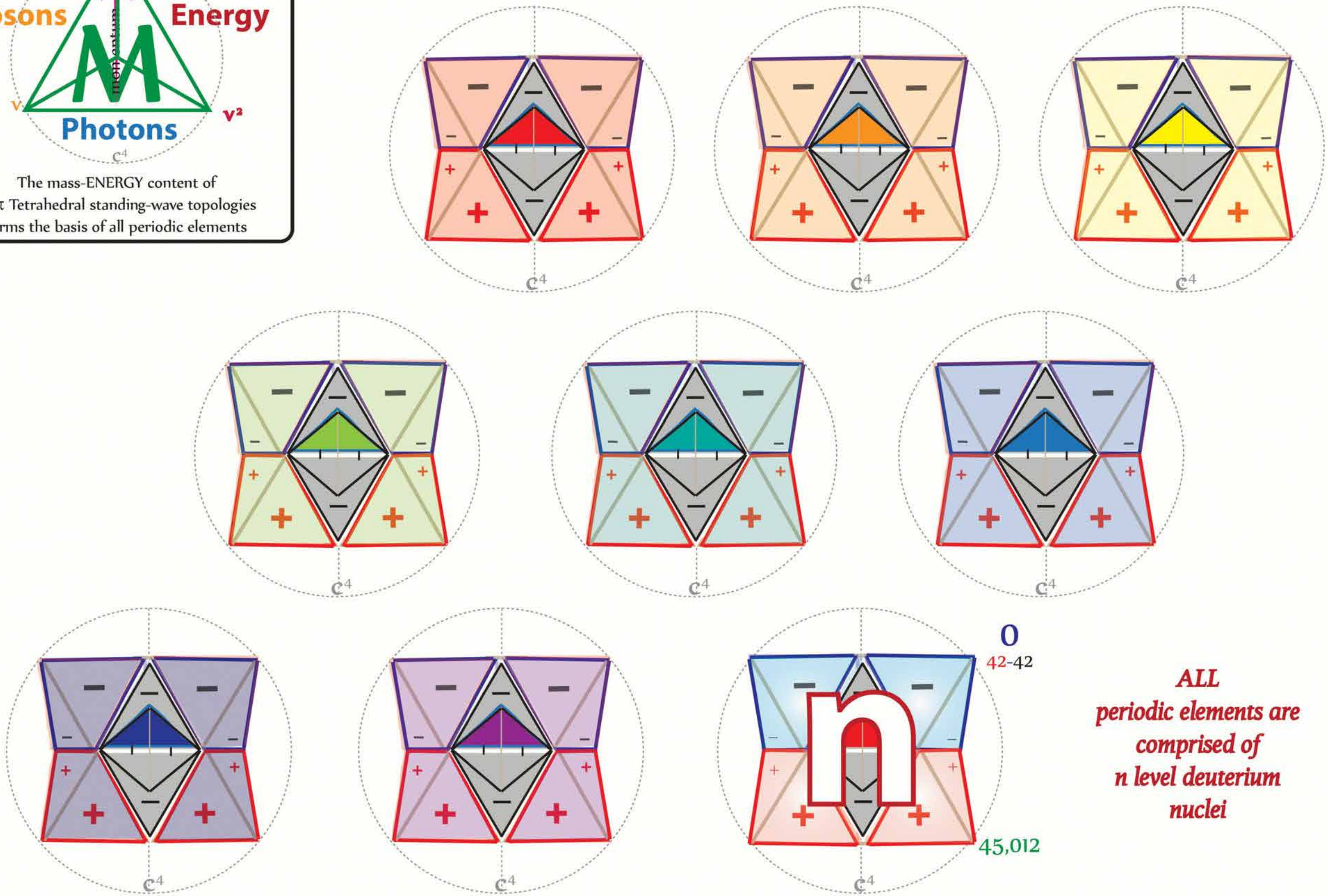
c⁴
n
0
42-42
45,012

ALL
periodic elements are
comprised of
n level deuterium
nuclei

Copyright ABRAHAM [2008] - All rights reserved

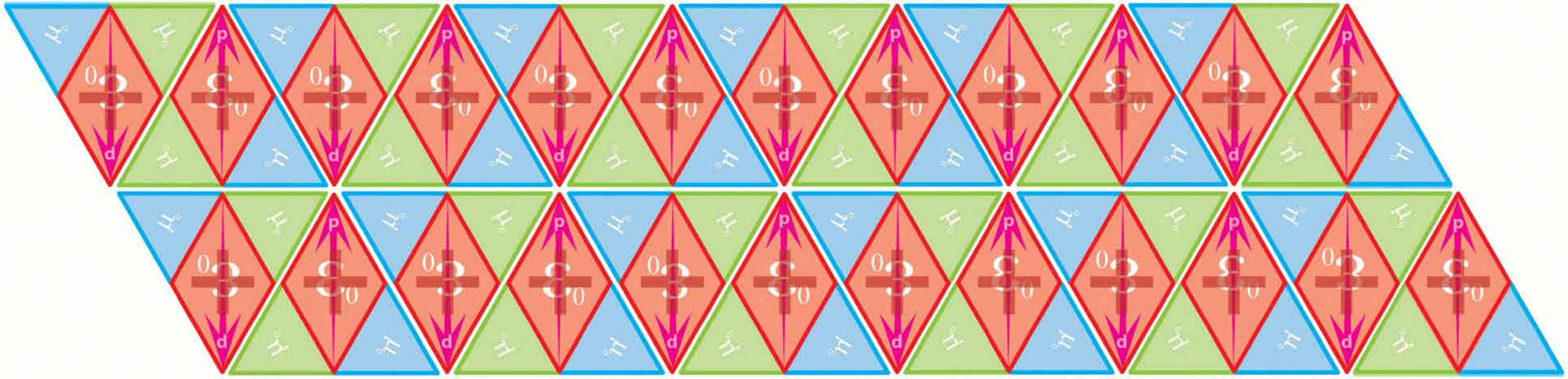

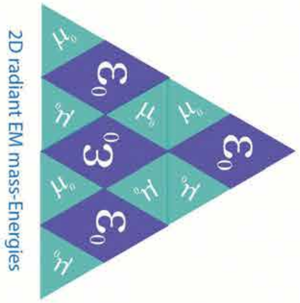
2D radiant EM mass-Energies
μ0
ε0

n1 - Charged Tetryon Templates

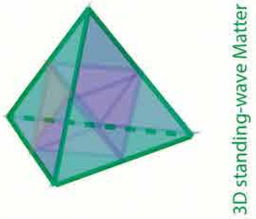
3D standing-wave Matter

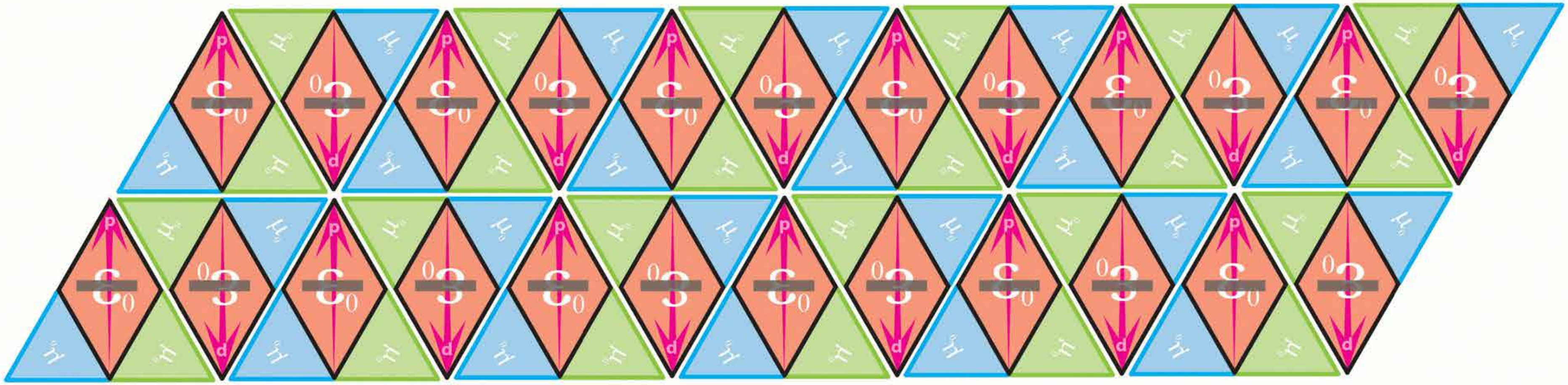

Copyright ABRAHAM [2008] - All rights reserved

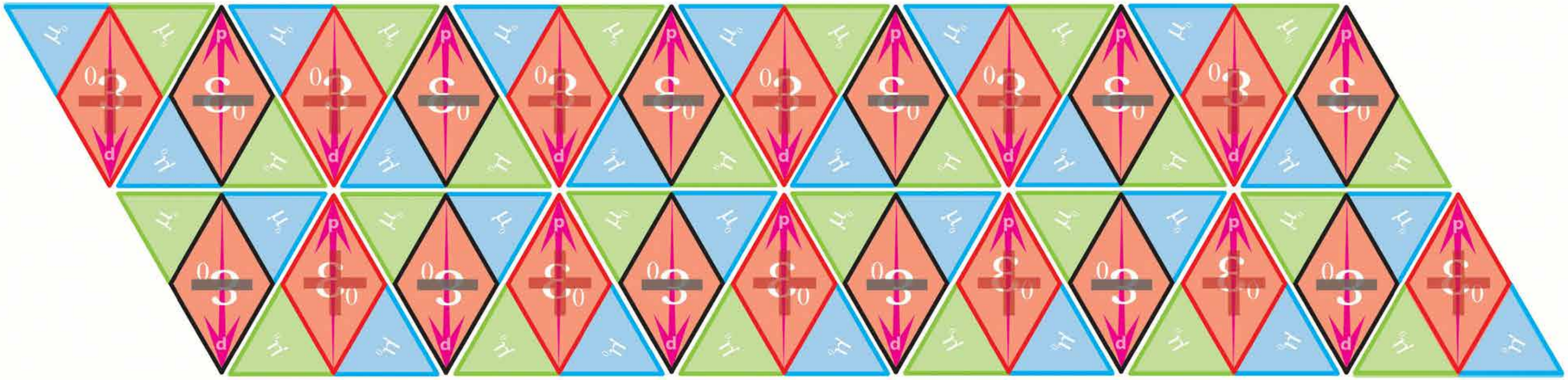

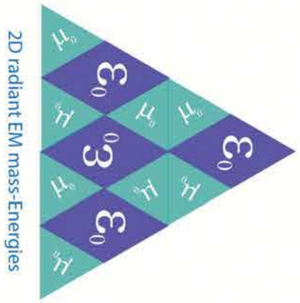
2D radiant EM mass-Energies
μ0
ε0

n1 - Neutral Tetryon Templates

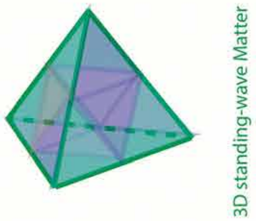
3D standing-wave Matter

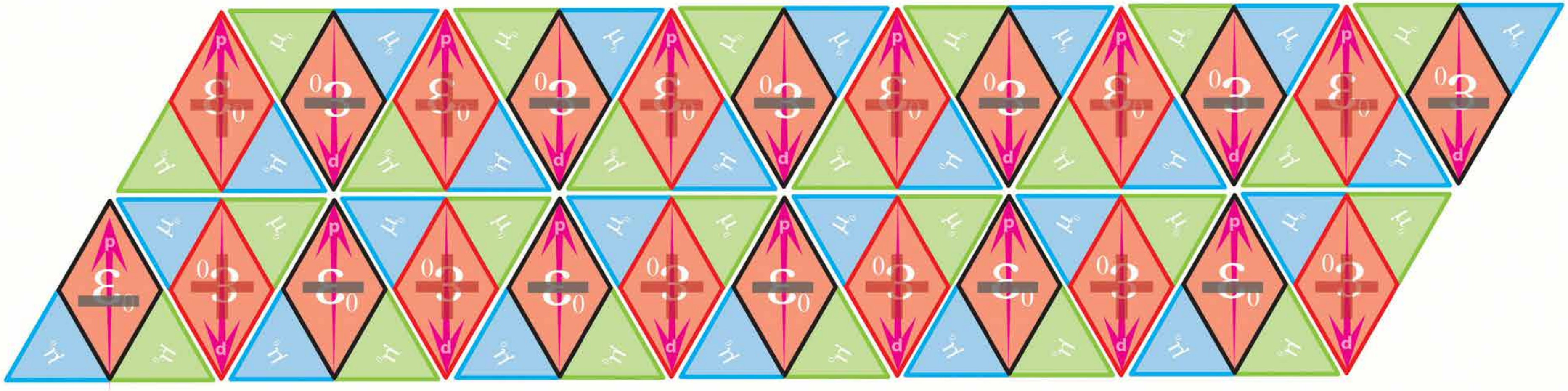

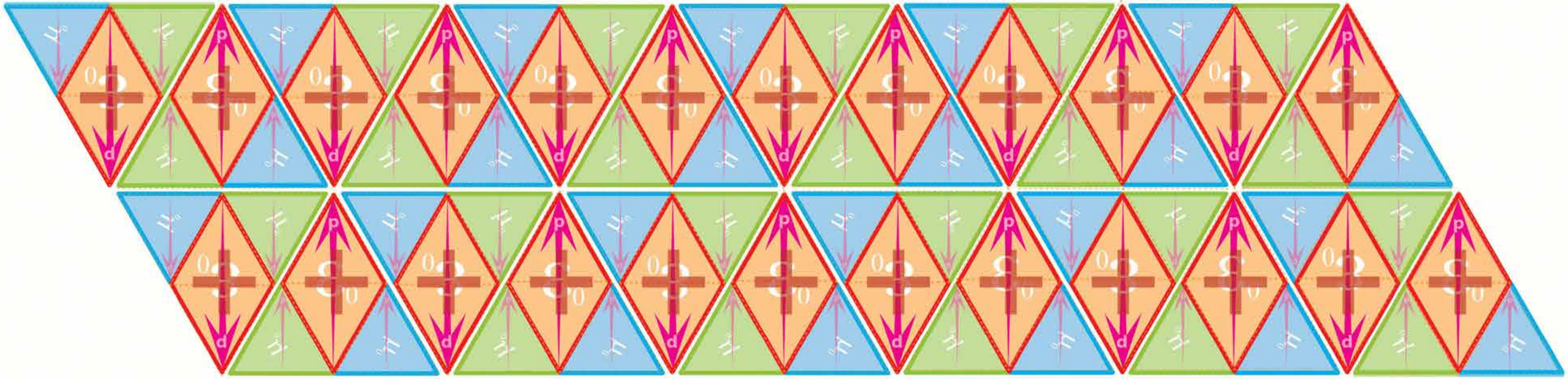
p
d
ε_0
μ_0

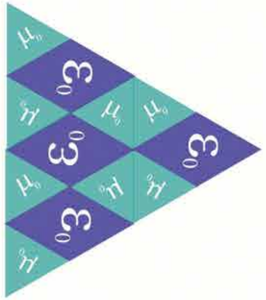
ε_0
μ_0

2D radiant EM mass-Energies

n2 - Charged Tetryon Templates

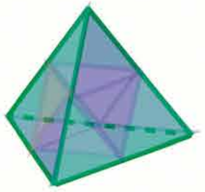

3D standing-wave Matter

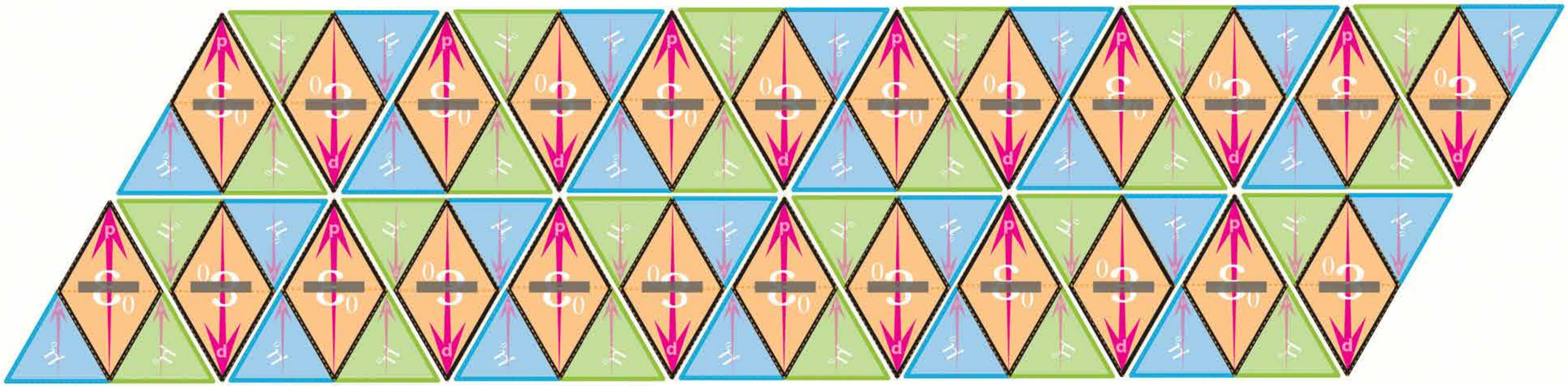
p
d
ε_0
μ_0

Copyright ABRAHAM [2008] - All rights reserved

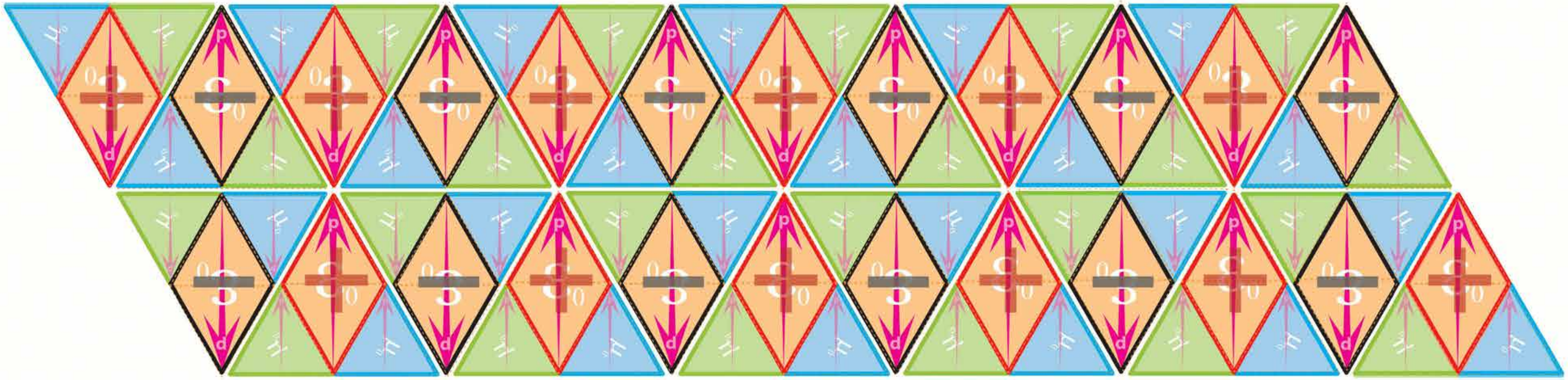

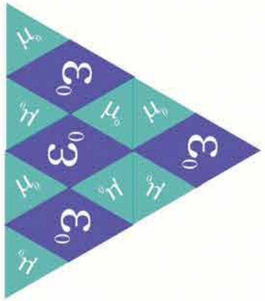

2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

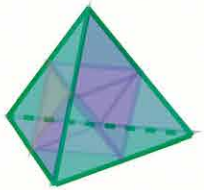

3D standing-wave Matter

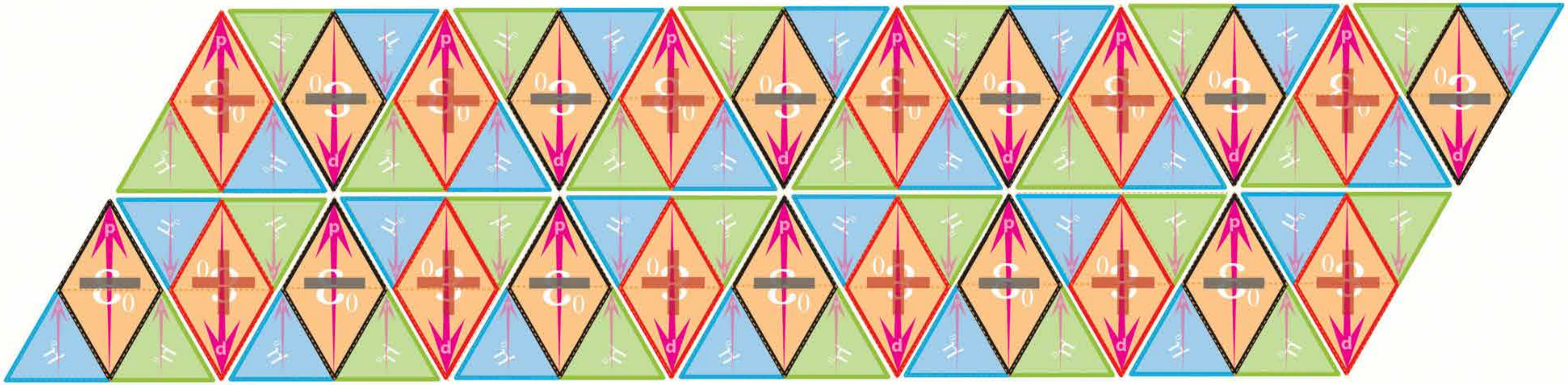

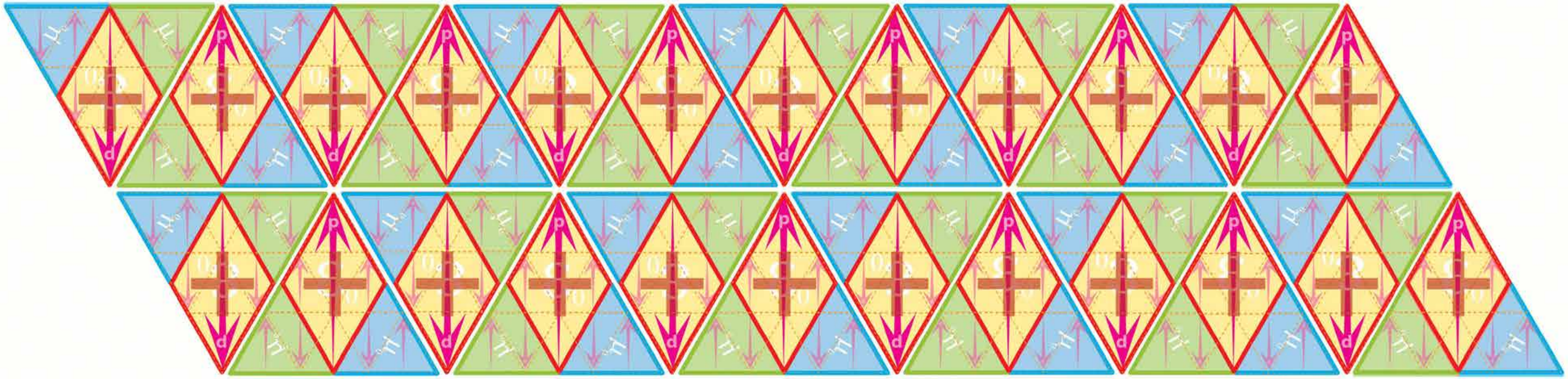

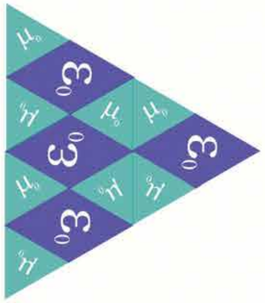

2D radiant EM mass-Energies

n3 - Charged Tetryon Templates

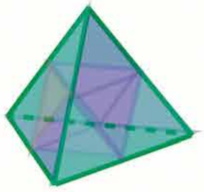

3D standing-wave Matter

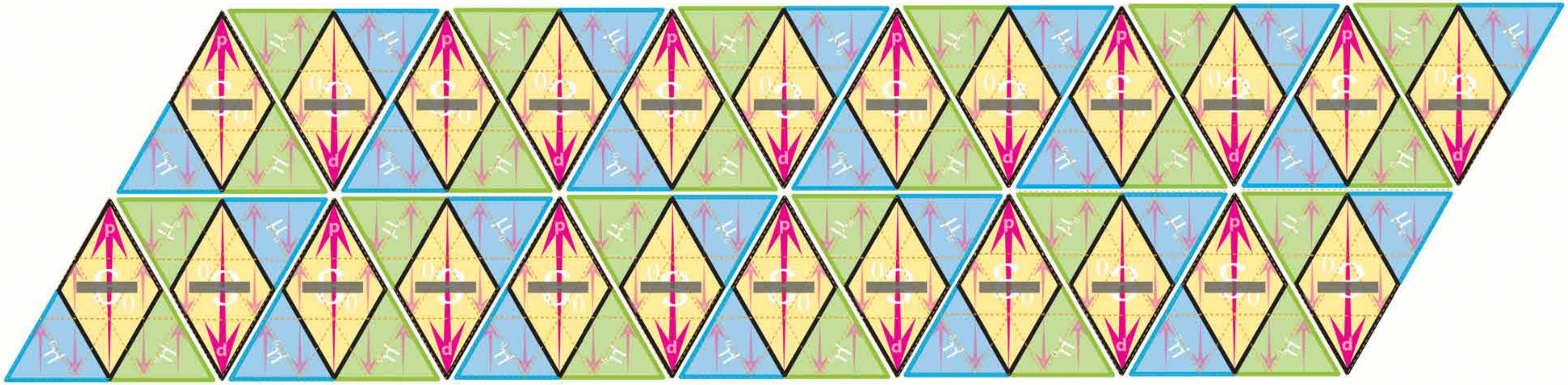

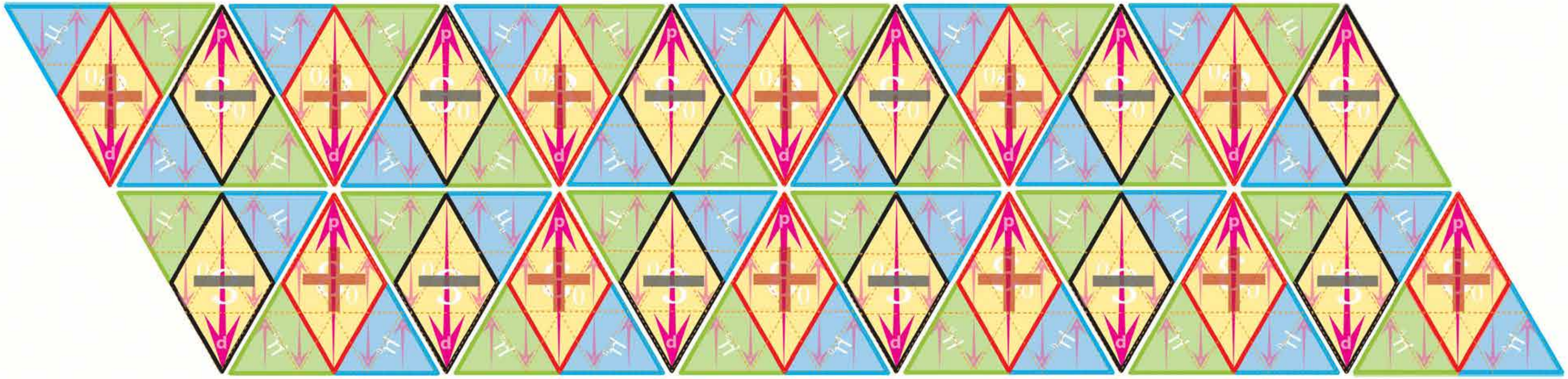

n3 - Neutral Tetryon Templates

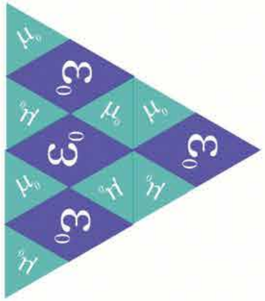

2D radiant EM mass-Energies

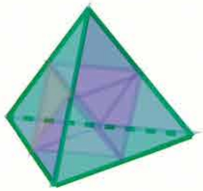

3D standing-wave Matter

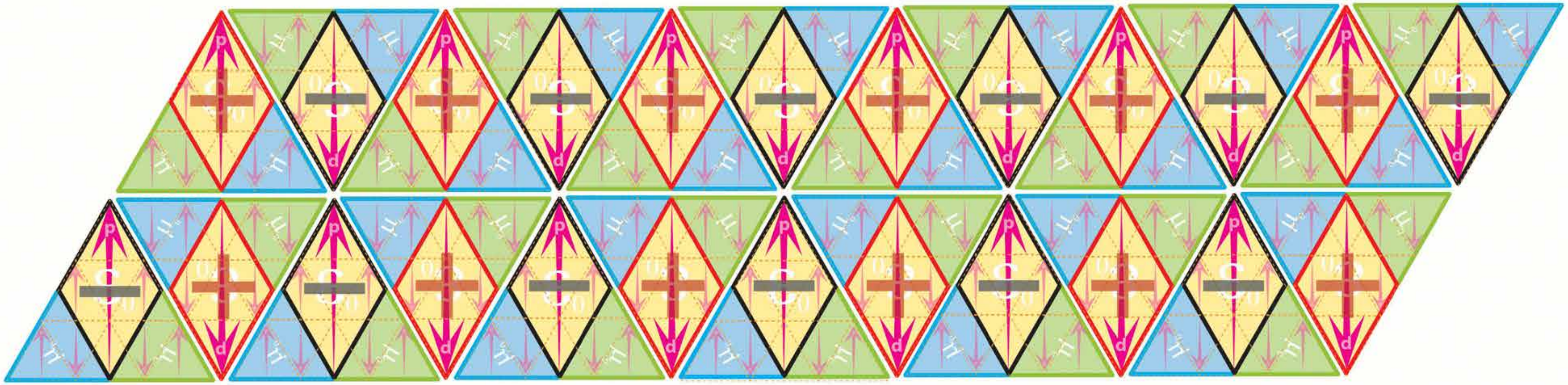

Copyright ABRAHAM [2008] - All rights reserved

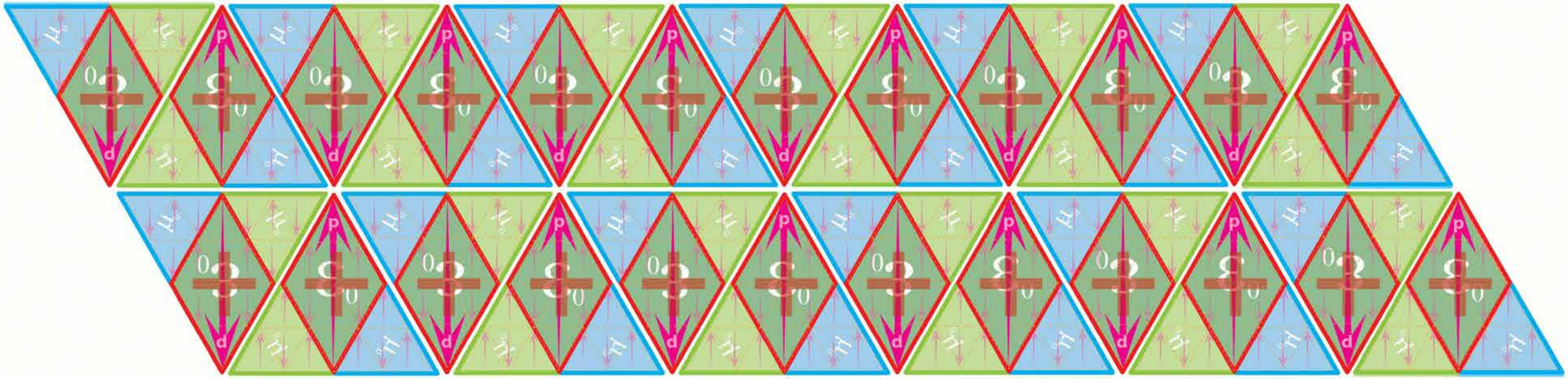

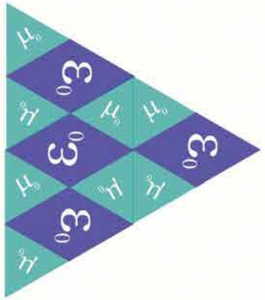

2D radiant EM mass-Energies

n4 - Charged Tetryon Templates

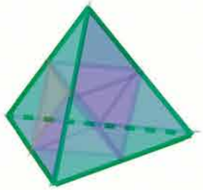

3D standing-wave Matter

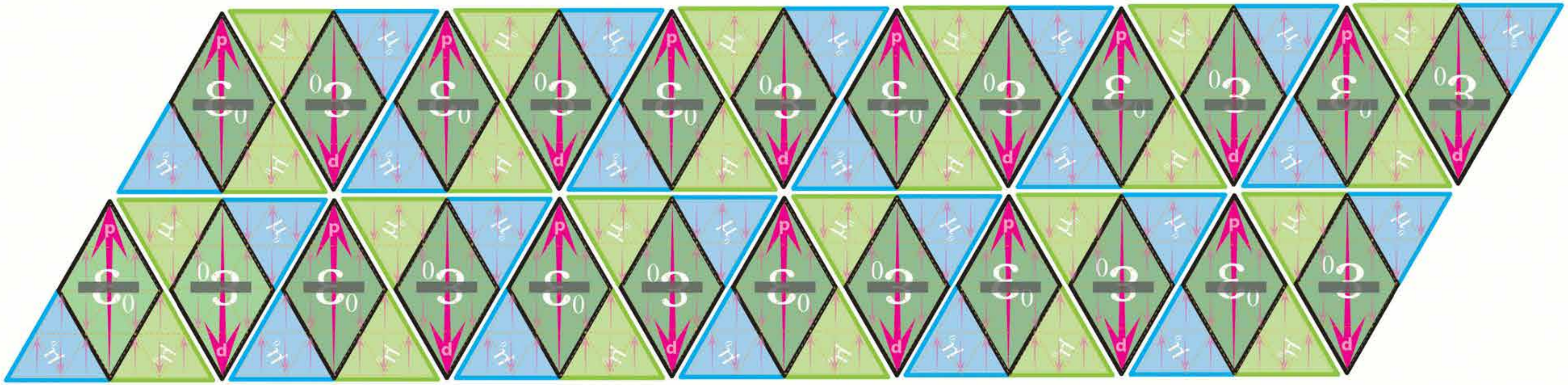

Copyright ABRAHAM [2008] - All rights reserved

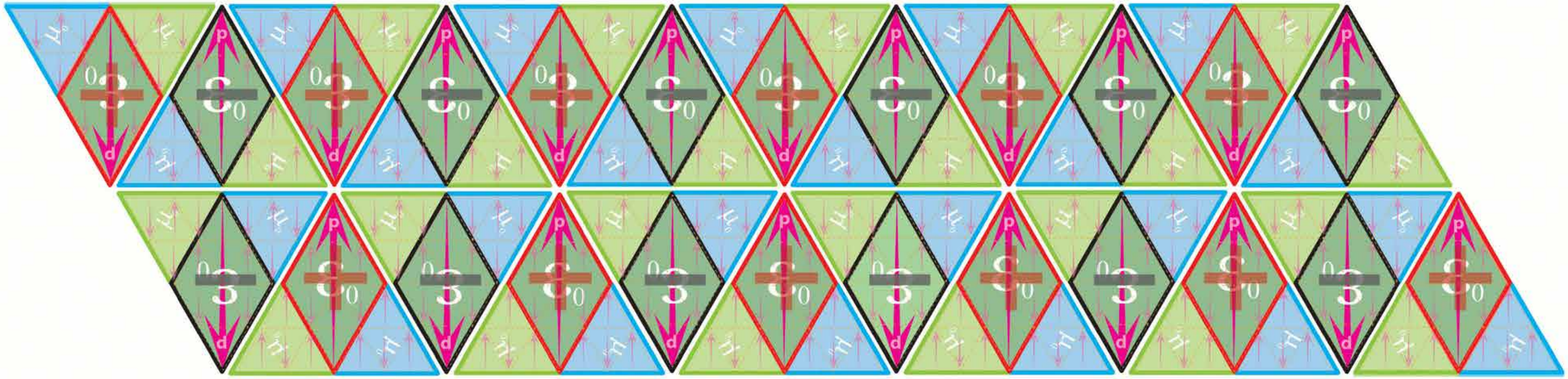

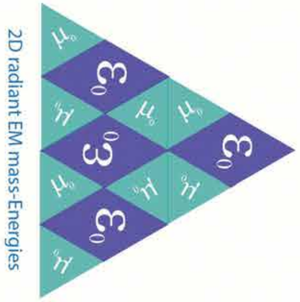
2D radiant EM mass-Energies
μ0
ε0

n4 - Neutral Tetryon Templates

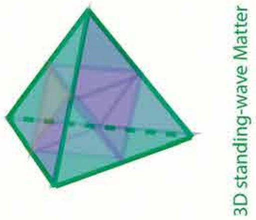
3D standing-wave Matter

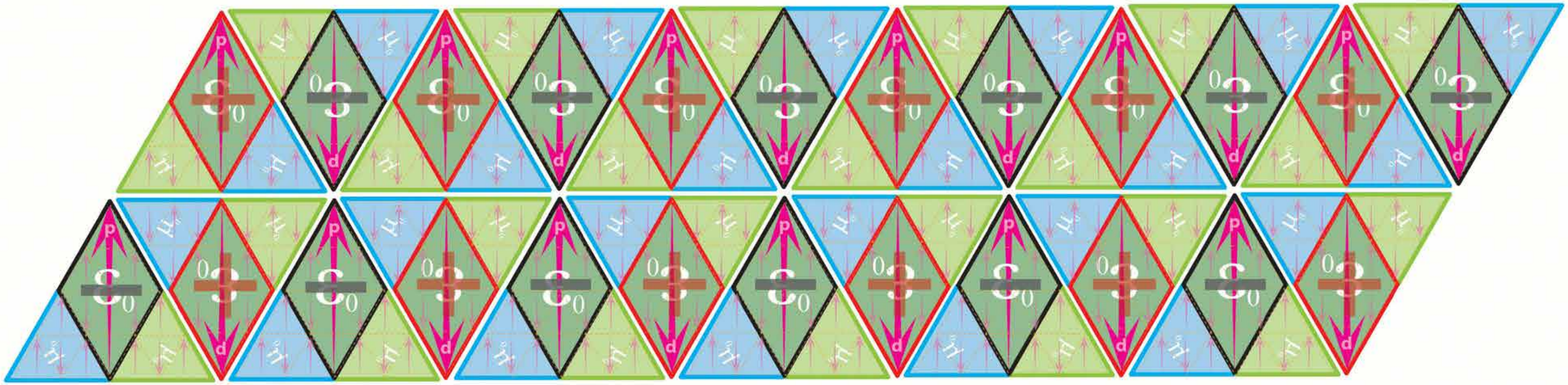

Copyright ABRAHAM [2008] - All rights reserved

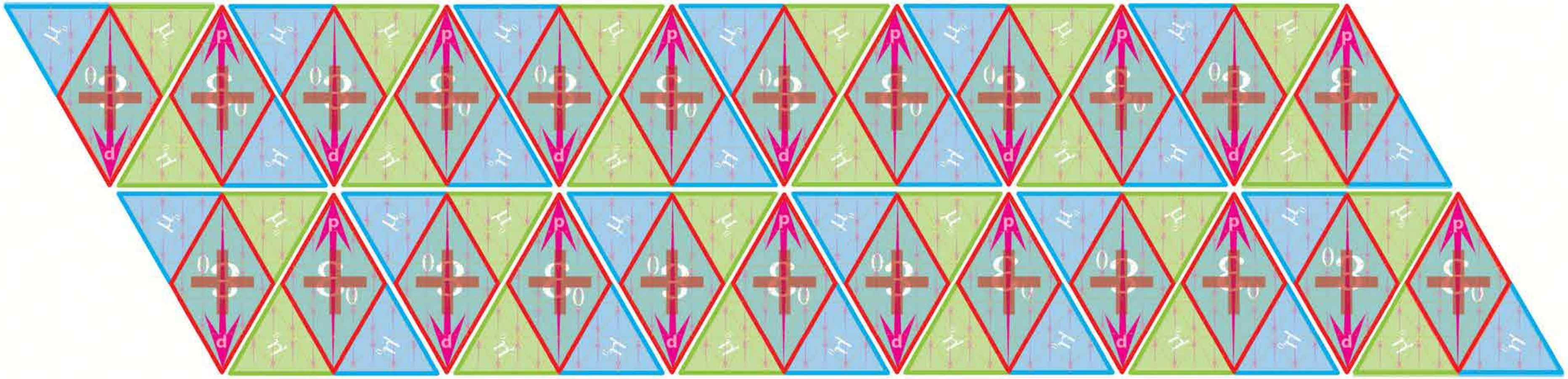
p
d
ε_0
μ_0

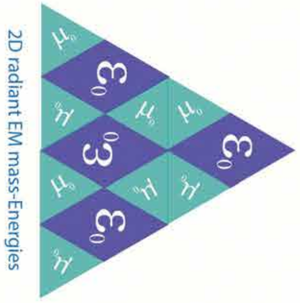
2D radiant EM mass-Energies
ε_0
μ_0

n5 - Charged Tetryon Templates

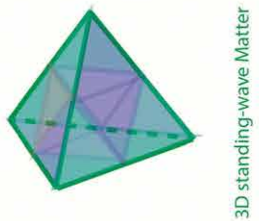
3D standing-wave Matter

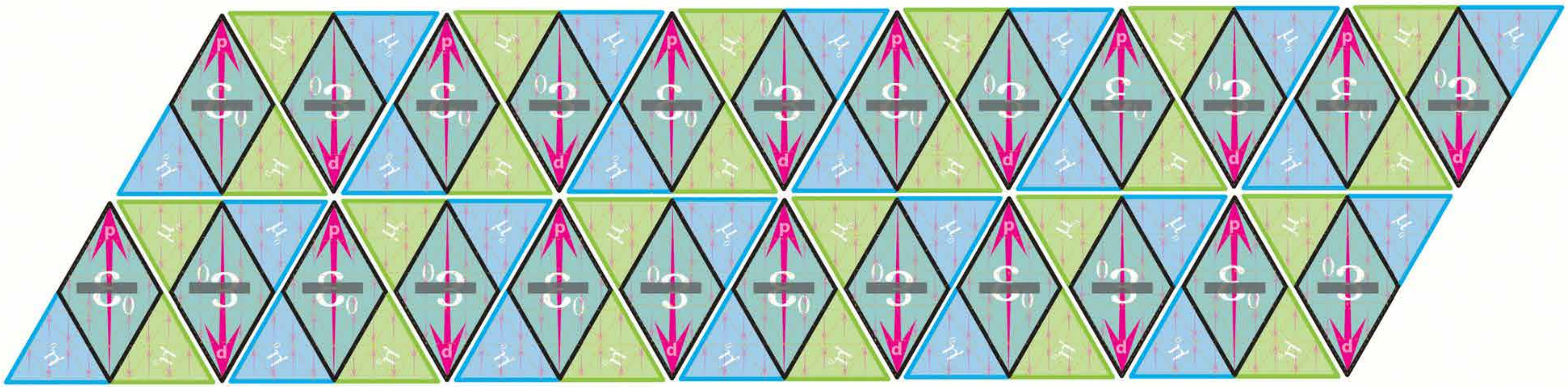
p
d
ε_0
μ_0

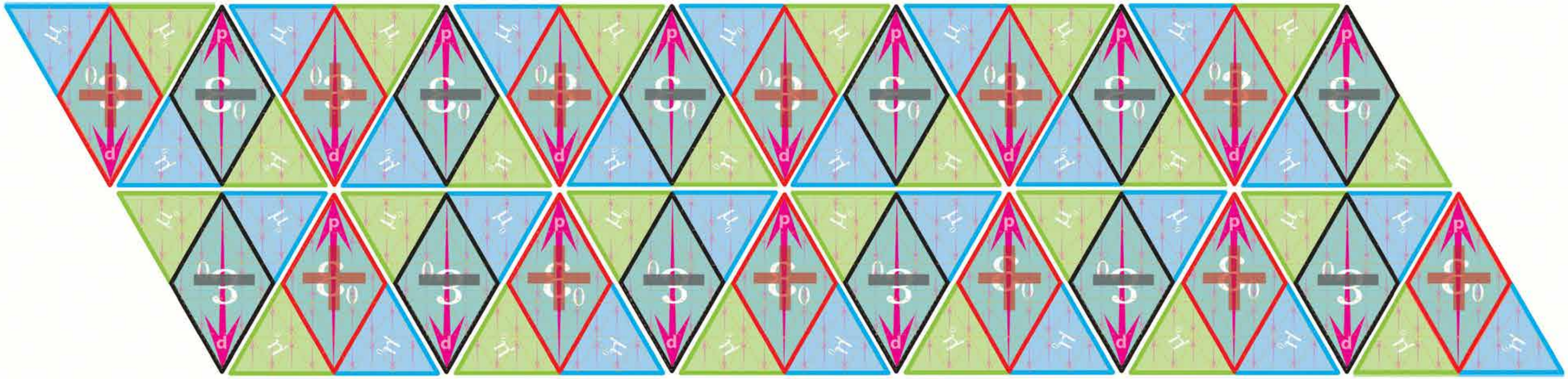
p
d
μ_0
ε_0

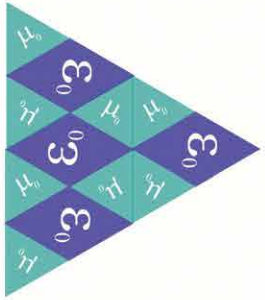
μ_0
ε_0

2D radiant EM mass-Energies

n5 - Neutral Tetryon Templates

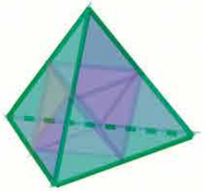

3D standing-wave Matter

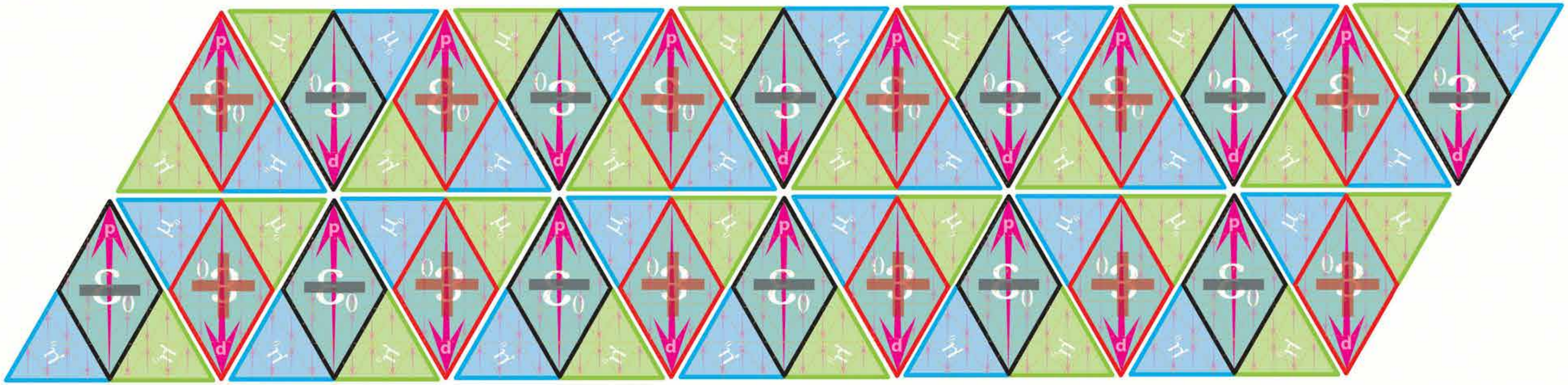
p
d
μ_0
ε_0

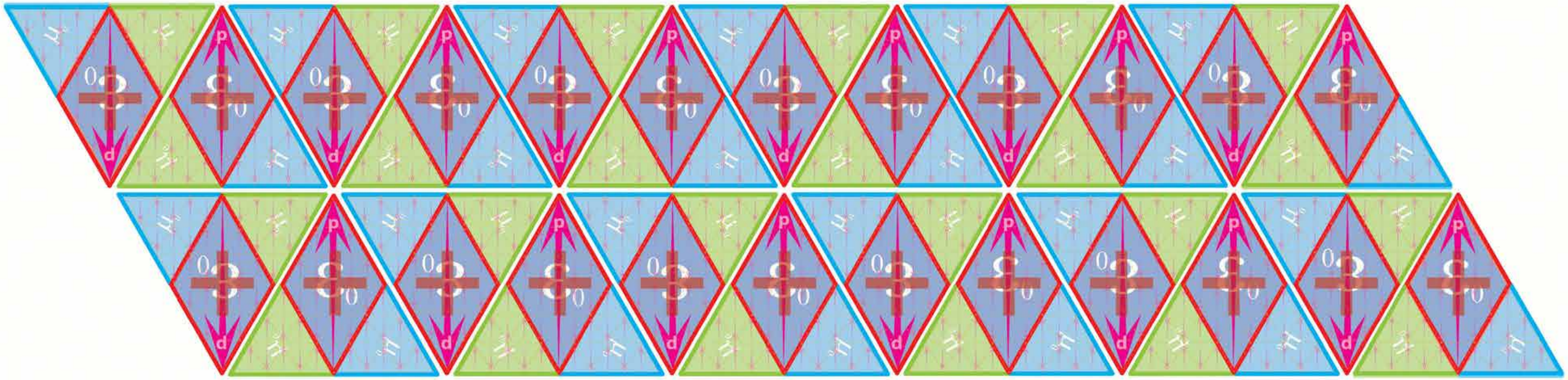

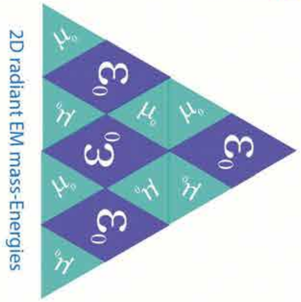
2D radiant EM mass-Energies
ε0
μ0

n6 - Charged Tetryon Templates

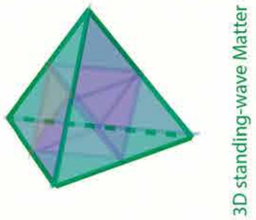
3D standing-wave Matter

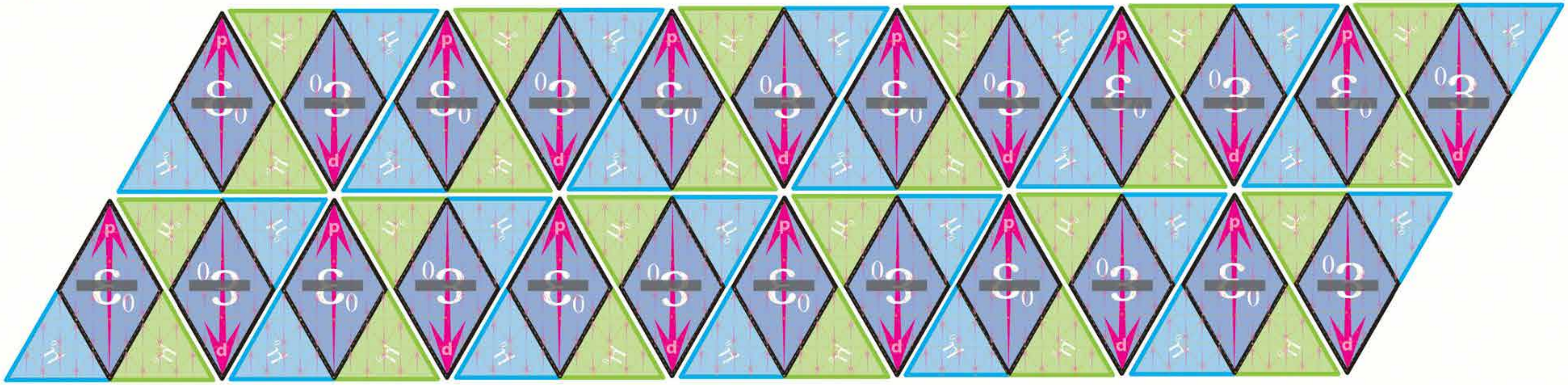

Copyright ABRAHAM [2008] - All rights reserved

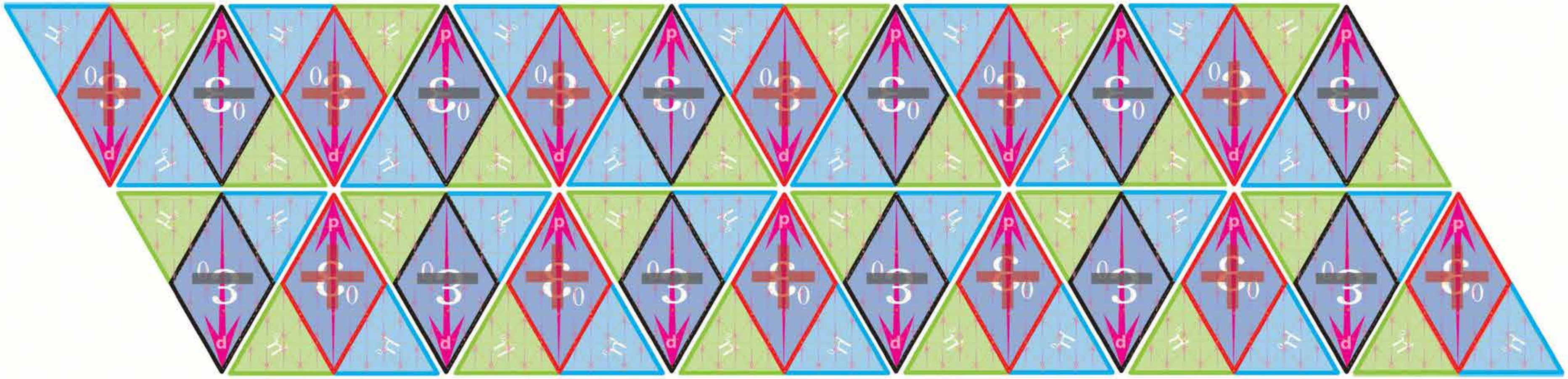

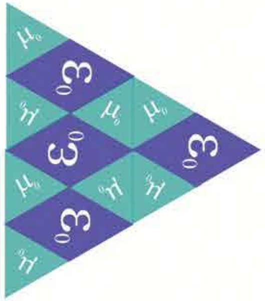

2D radiant EM mass-Energies

n6 - Neutral Tetryon Templates

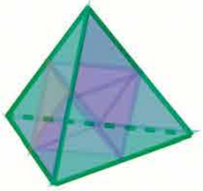

3D standing-wave Matter

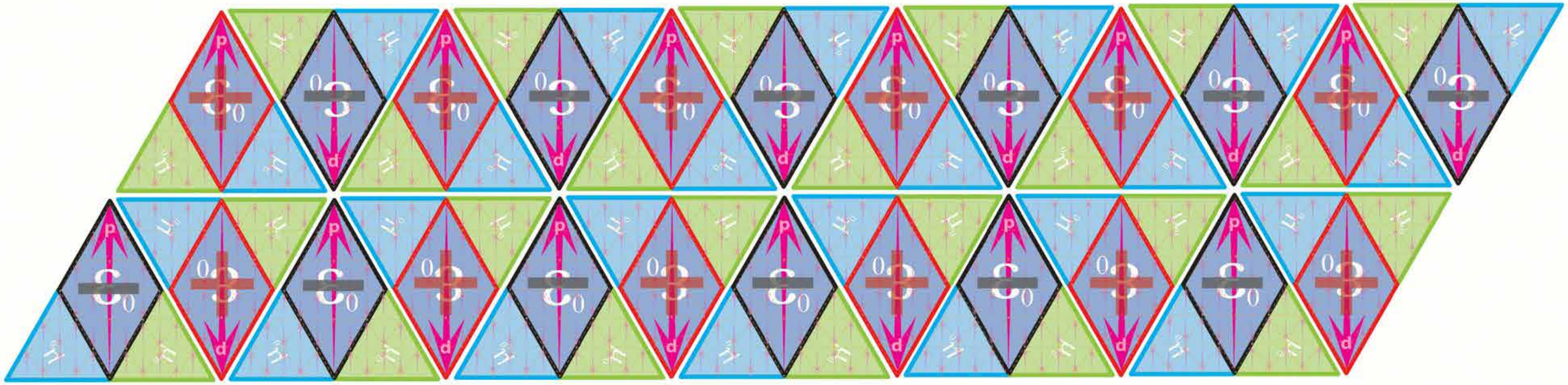

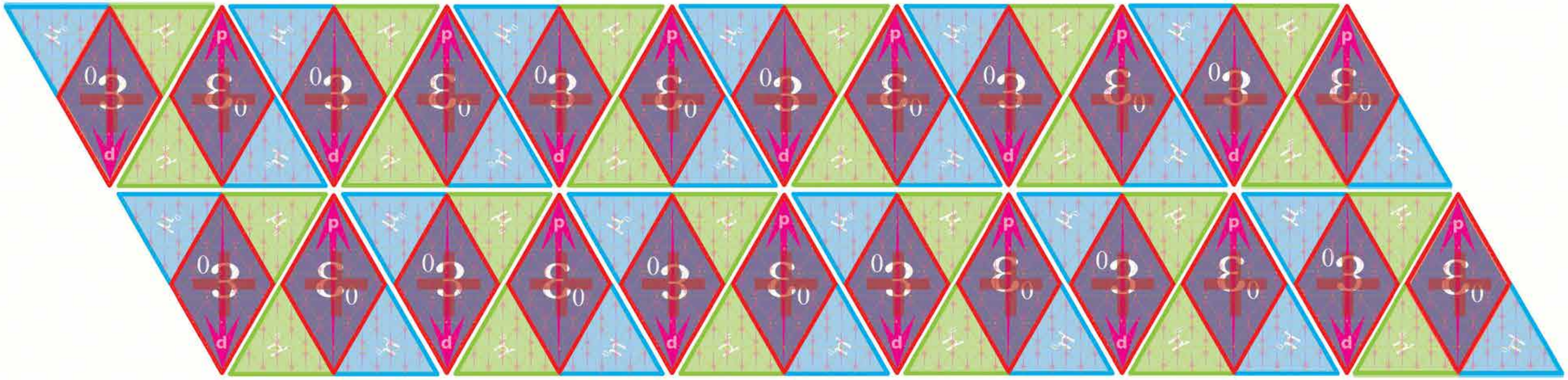
p
d
ε_0
μ_0

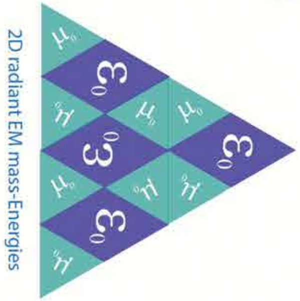
2D radiant EM mass-Energies
ε_0
μ_0

n7 - Charged Tetryon Templates

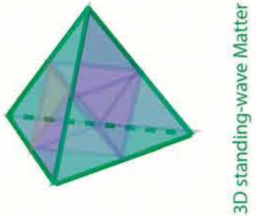
3D standing-wave Matter

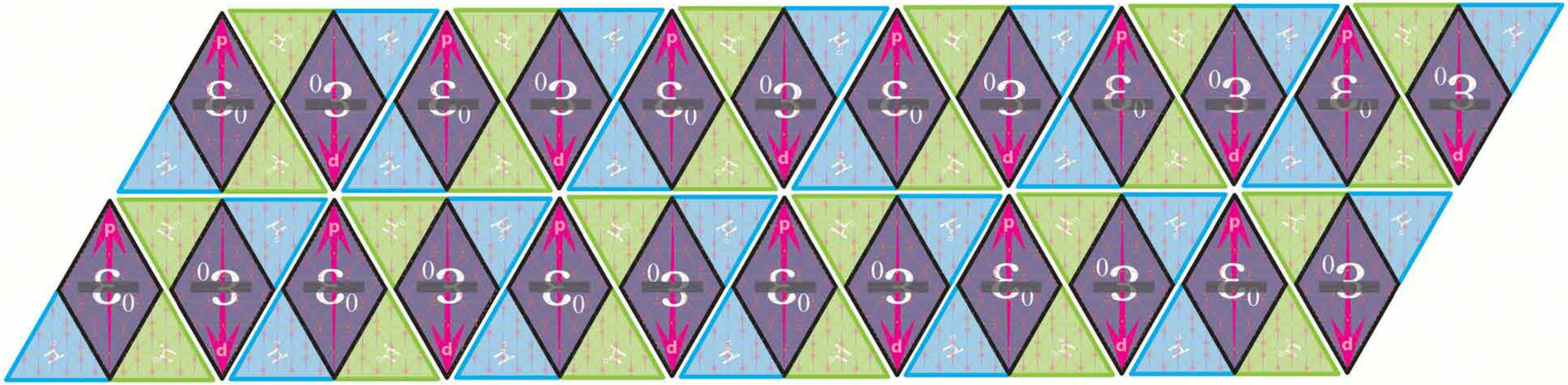
p
d
ε_0
μ_0

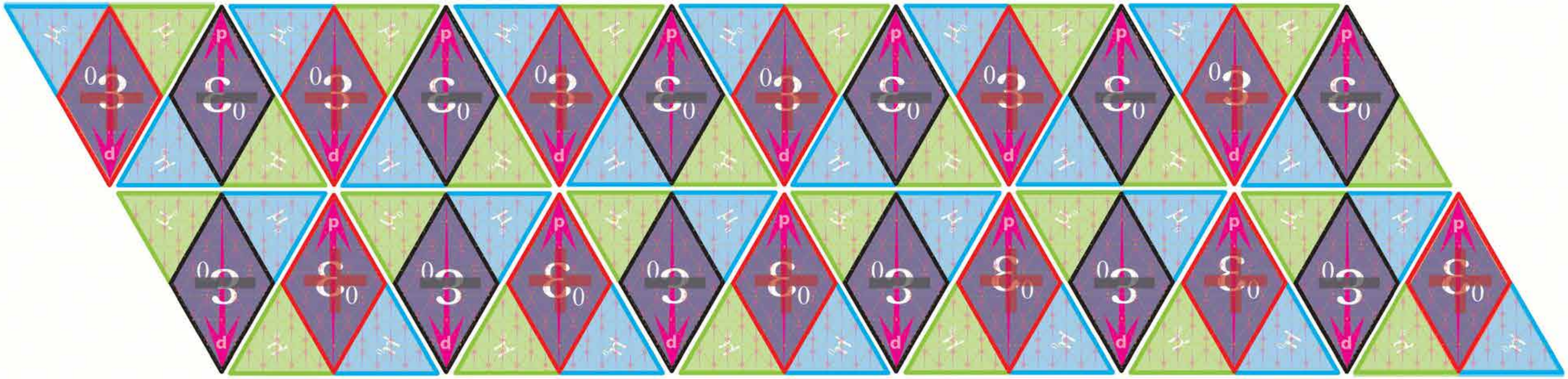

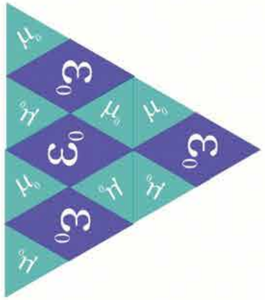

2D radiant EM mass-Energies

n7 - Neutral Tetryon Templates

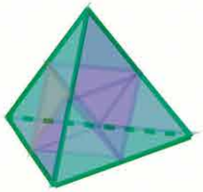

3D standing-wave Matter

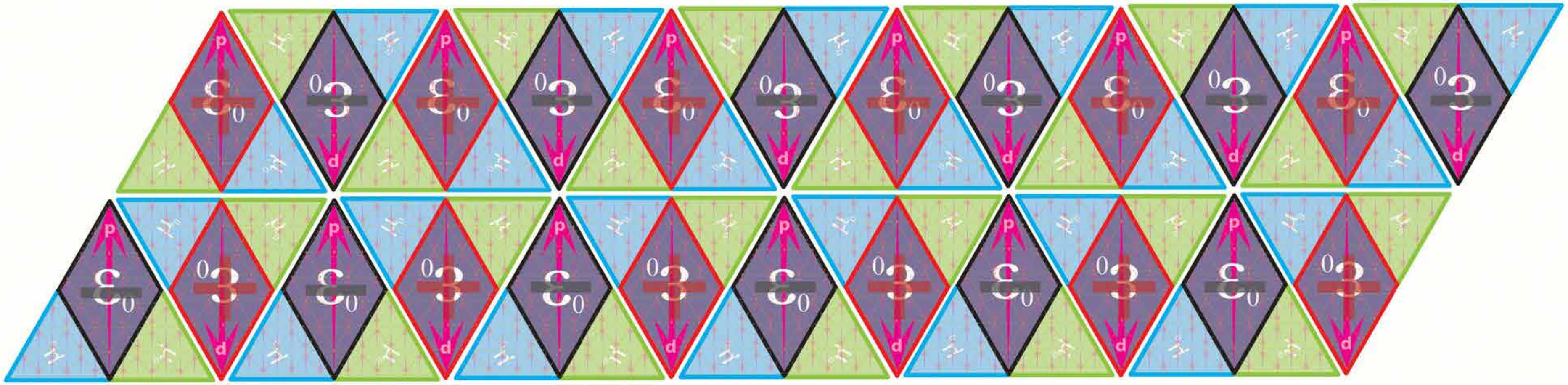

Copyright ABRAHAM [2008] - All rights reserved

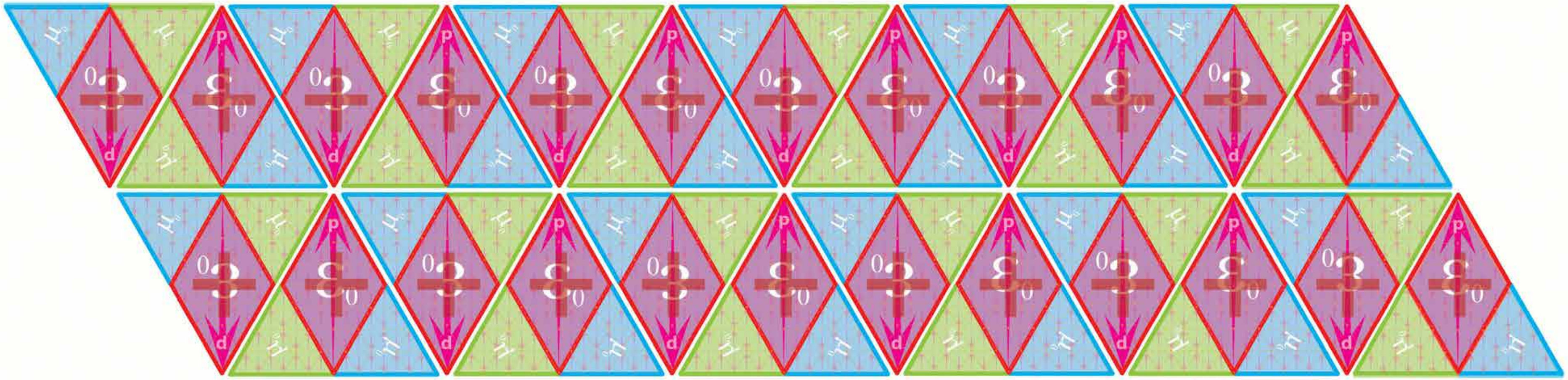
p
d
μ_0
ε_0

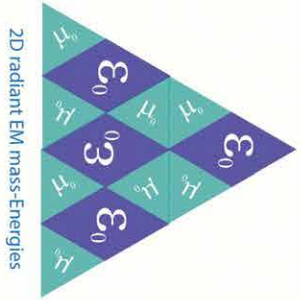
μ_0
ε_0

2D radiant EM mass-Energies

n8 - Charged Tetryon Templates

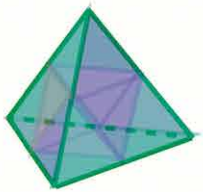

3D standing-wave Matter

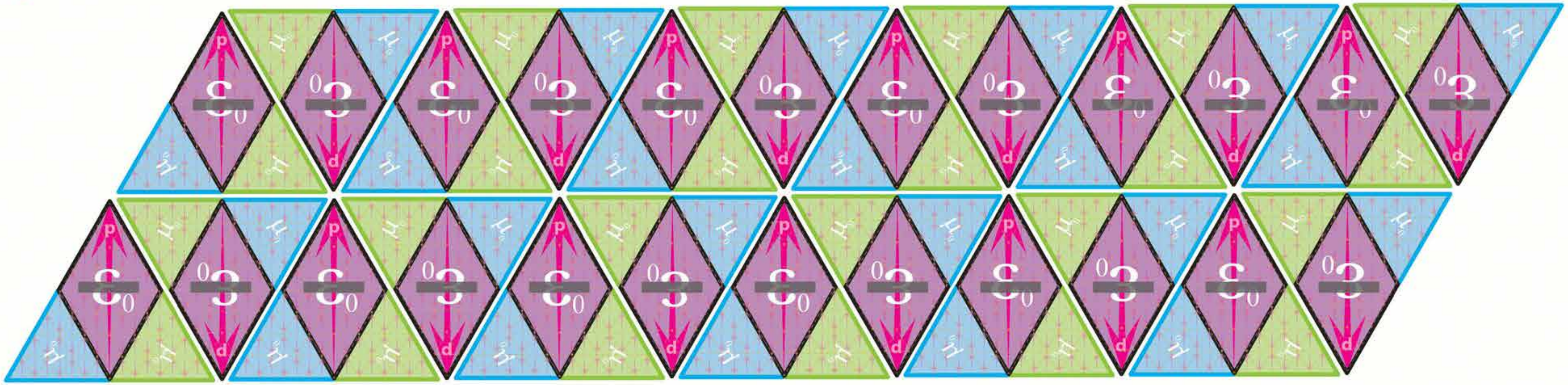
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

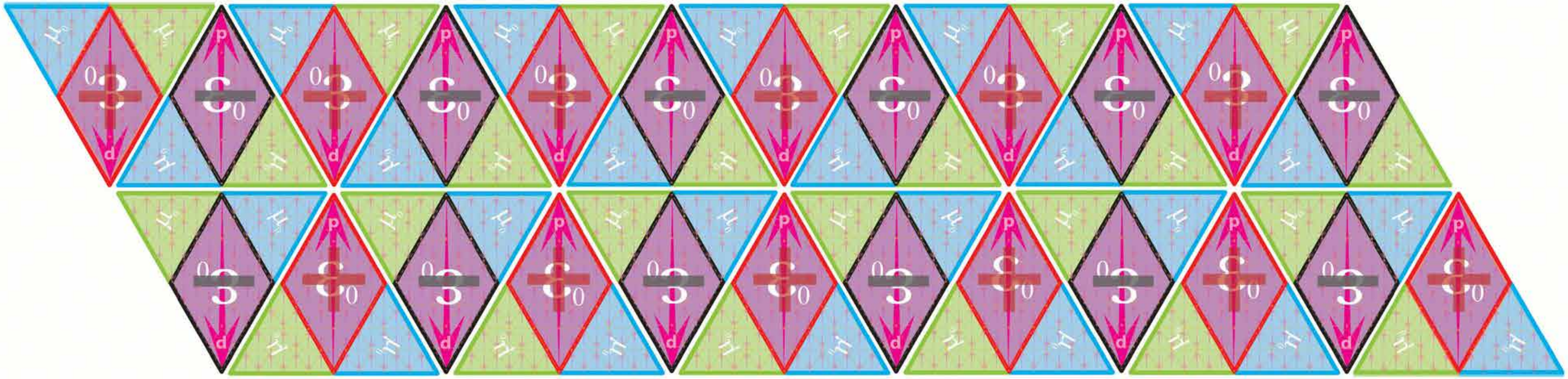
p
d
ε_0
μ_0

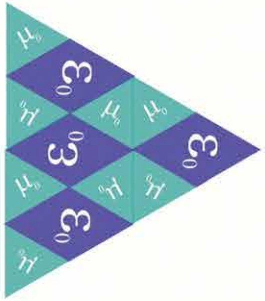
ε_0
μ_0

2D radiant EM mass-Energies

n8 - Neutral Tetryon Templates

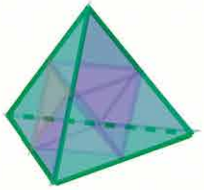

3D standing-wave Matter

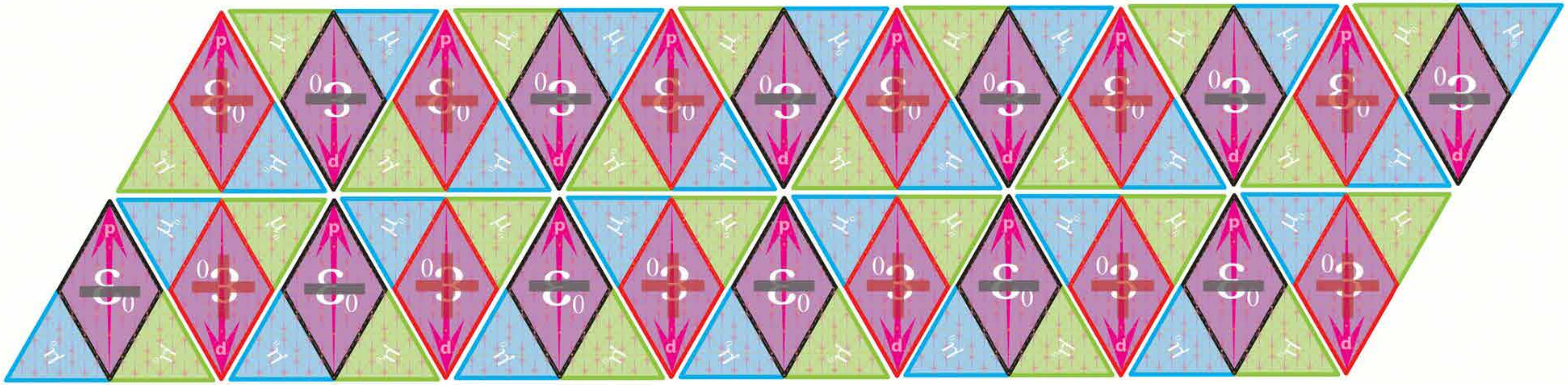
p
d
ε_0
μ_0

Copyright ABRAHAM [2008] - All rights reserved

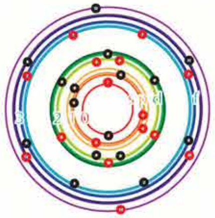

Each shell consists of one or more subshells, and
each subshell consists of one or more atomic orbitals.

Periodic element geometries

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus.
The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the
"2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus.
The shell letters K, L, M, ... are alphabetical

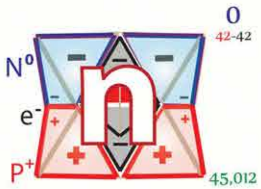
0
42-42
N⁰
e⁻
P⁺
45,012

Each shell can contain only an integer
number of whole deuterium nuclei
[Proton, Neutron & electron]

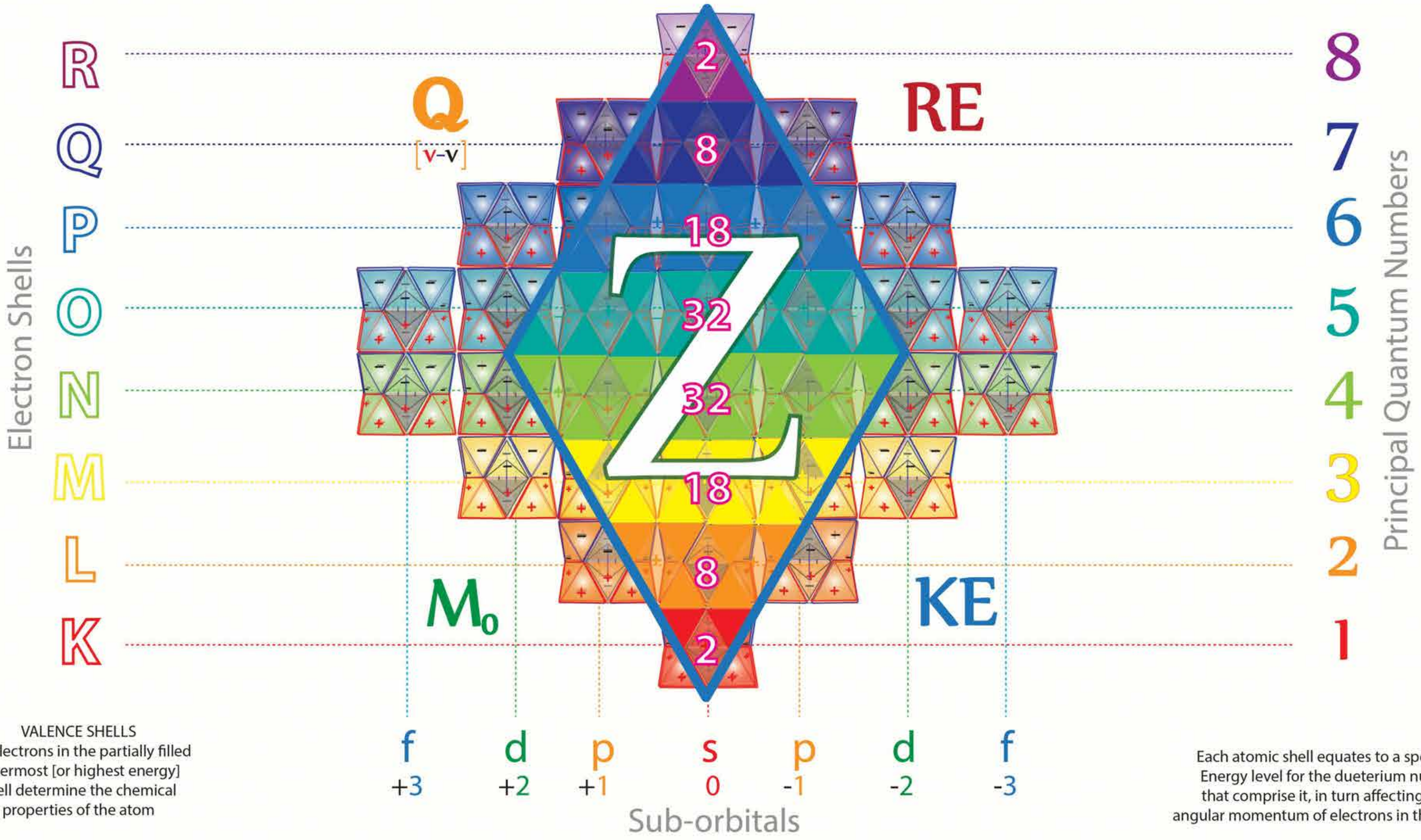
Electron Shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
M₀
KE
Z
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
Principal Quantum Numbers
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
Sub-orbitals

VALENCE SHELLS
The electrons in the partially filled
outermost [or highest energy]
shell determine the chemical
properties of the atom

Each atomic shell equates to a specific
Energy level for the dueterium nuclei
that comprise it, in turn affecting the
angular momentum of electrons in that shell

Copyright ABRAHAM [2008] - All rights reserved

Periodic Harmonic motions

$x = A\cos(\omega t + \varphi)$

Circular motion

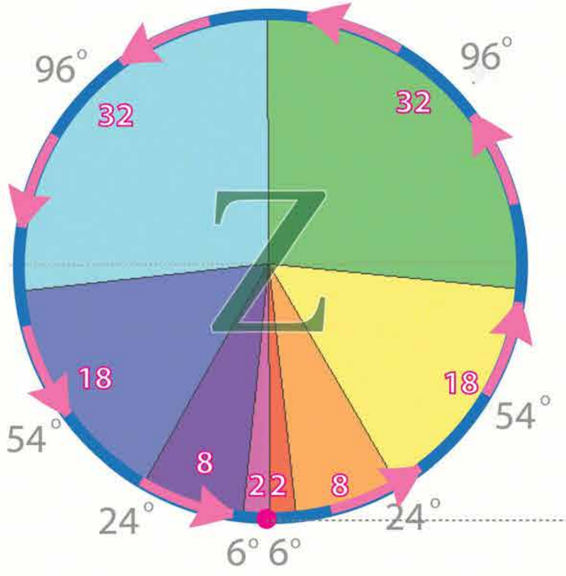

96°
96°
32
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

circular harmonic motion

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

Much of the math in of modern physics is predicated on the assumption that π [where it appears] is related to the properties of a circle

$$P\sum 2(n^2)$$

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

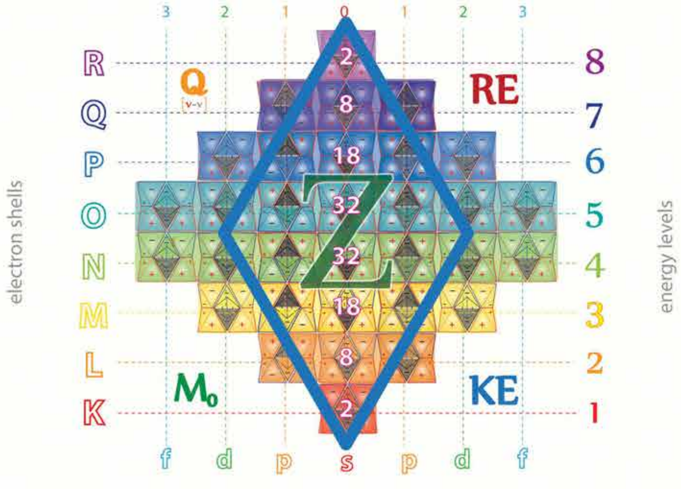

Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M₀
KE
2
8
18
Z
32
32
18
8
2
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

$F = -kx$

Linear motion

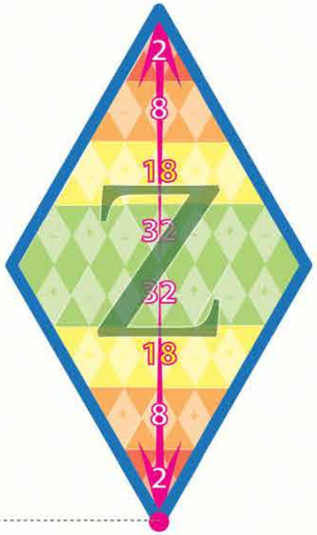

2
8
18
Z
32
32
18
8
2

simple harmonic motion

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

Copyright ABRAHAM [2008] - All rights reserved

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

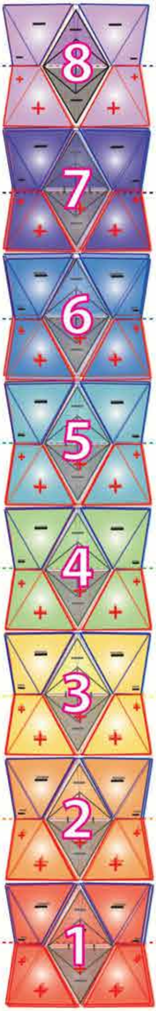

Baryon mass-energies	n		electron KEM field hv^2
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13

Compton frequencies
930 MeV

1.2 e20
electron rest Matter is velocity invariant
496 keV

spectral frequencies
13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

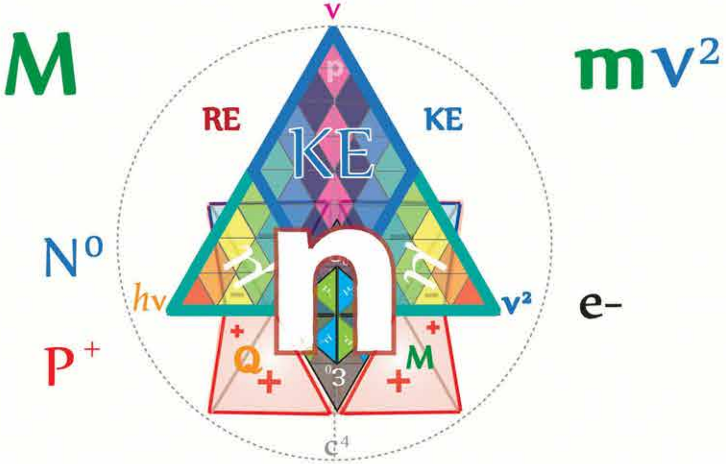

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses — lepton rest mass — KEM

$$\overset{8}{\underset{1}{n}}\Big[\,[72(n)^2]+[12e19]+[m_e v^2]\,\Big]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Baryons **930.947 MeV** + **KEM fields** **13.525 ev** + **electrons** **496.519 keV**

Mapping Planck mass-energy contributions to elementary Matter and isotopes

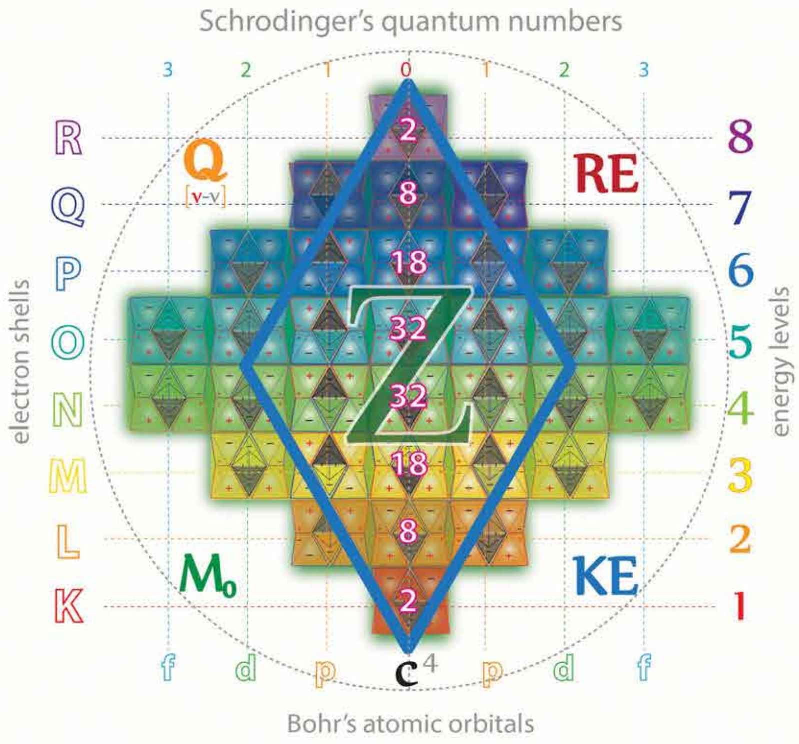
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M_0
KE
2
8
18
32
Z
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

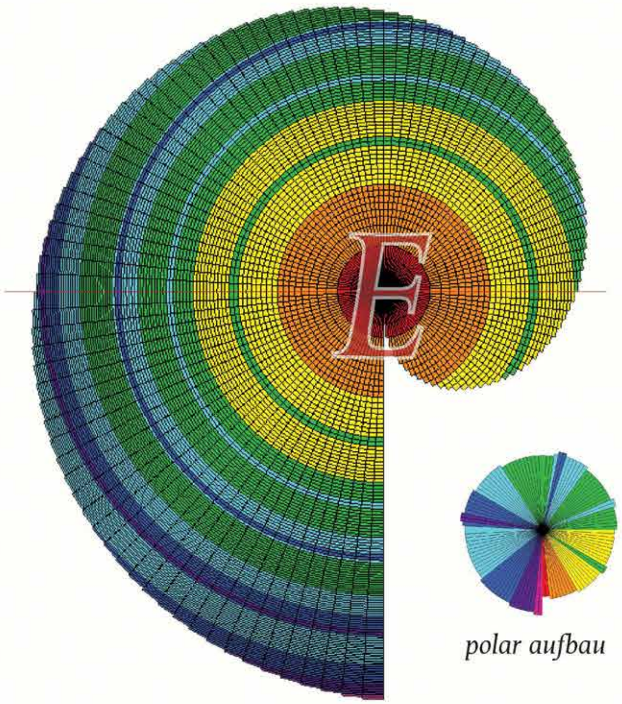
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[\underset{\text{Deuteron rest mass}}{72[v^2]} + \underset{\text{Spectral lines}}{v} + \underset{\text{electron rest mass}}{1.20\ \text{e}20\ v}\right]$$

reveal a quadratic formulation for all Z numbers

$$\sum_{K}^{R}$$

nuclei per shell

STEP ONE

Periodic summation follows the atomic shell electron config

$$_{P}\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{1}^{120} Z$$

element number

STEP TWO

Periodic elements build up following the aufbau sequence

$$_{P}\sum_{K}^{R} 2(x^2) =$$

1 R
2
3
4
4
3
2
1 K

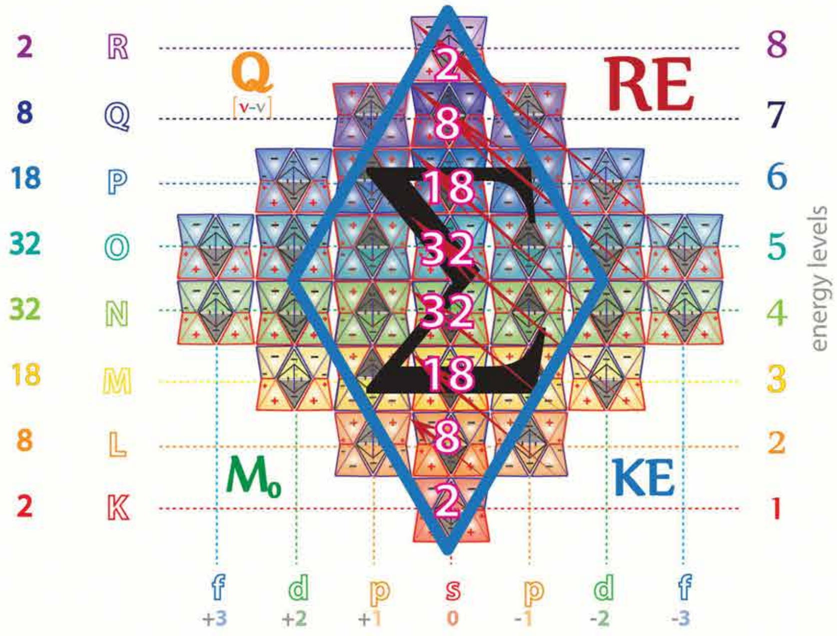

Each atomic shell can hold only a fixed number of deuterium nuclei

${}_{P}\sum R = 2$	2 nuclei [74,496 ea] +	120	Unbinilium
${}_{P}\sum Q = 8$	8 nuclei [69,780 ea] +	118	Ununoctium
${}_{P}\sum P = 18$	18 nuclei [65,232 ea] +	110	Darmstadtium
${}_{P}\sum O = 32$	32 nuclei [60,852 ea] +	92	Uranium
${}_{P}\sum N = 32$	32 nuclei [56,640 ea] +	60	Neodymuim
${}_{P}\sum M = 18$	18 nuclei [52,596 ea] +	28	Argon
${}_{P}\sum L = 8$	8 nuclei [48,720 ea] +	10	Neon
${}_{P}\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

$$Z\#\left[\begin{array}{ll} z\ \text{Protons} & [24\text{-}12] \\ z\ \text{Neutrons} & [18\text{-}18] \\ z\ \text{electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Copyright ABRAHAM [2008] - All rights reserved

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

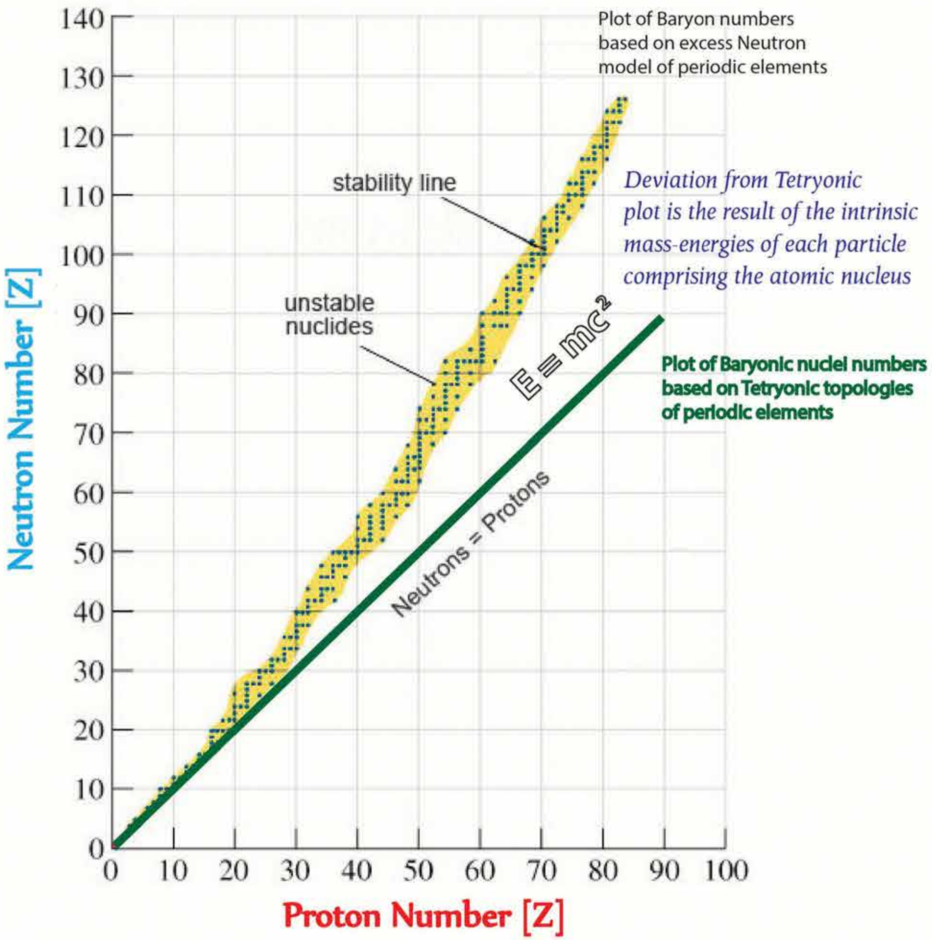

Plot of Baryon numbers based on excess Neutron model of periodic elements

Deviation from Tetryonic plot is the result of the intrinsic mass-energies of each particle comprising the atomic nucleus

Plot of Baryonic nuclei numbers based on Tetryonic topologies of periodic elements

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

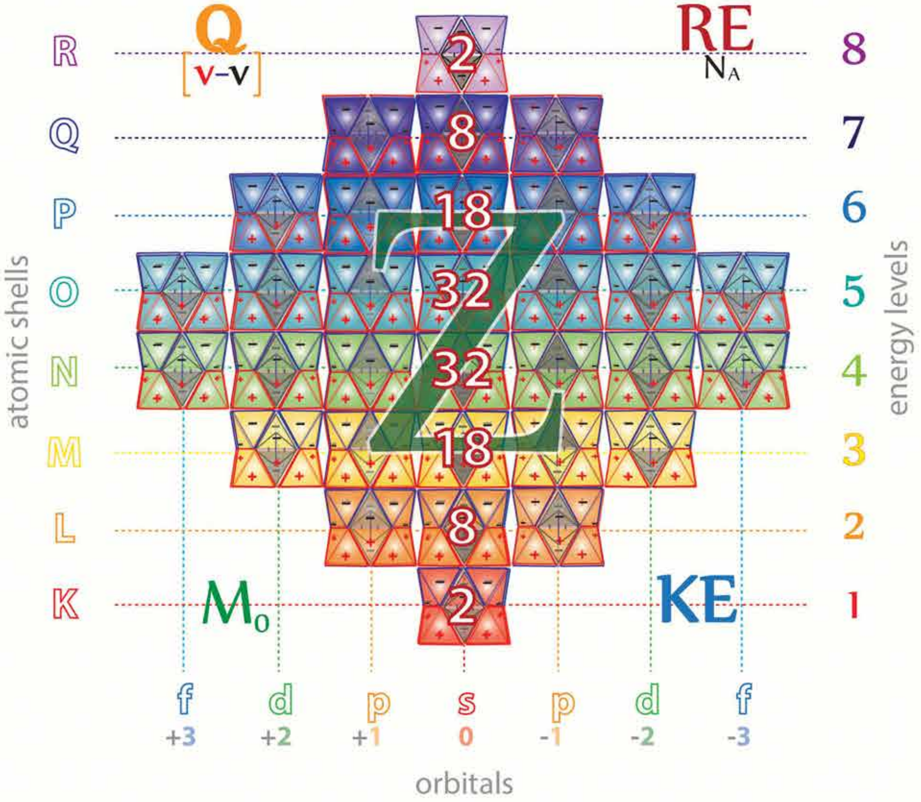

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

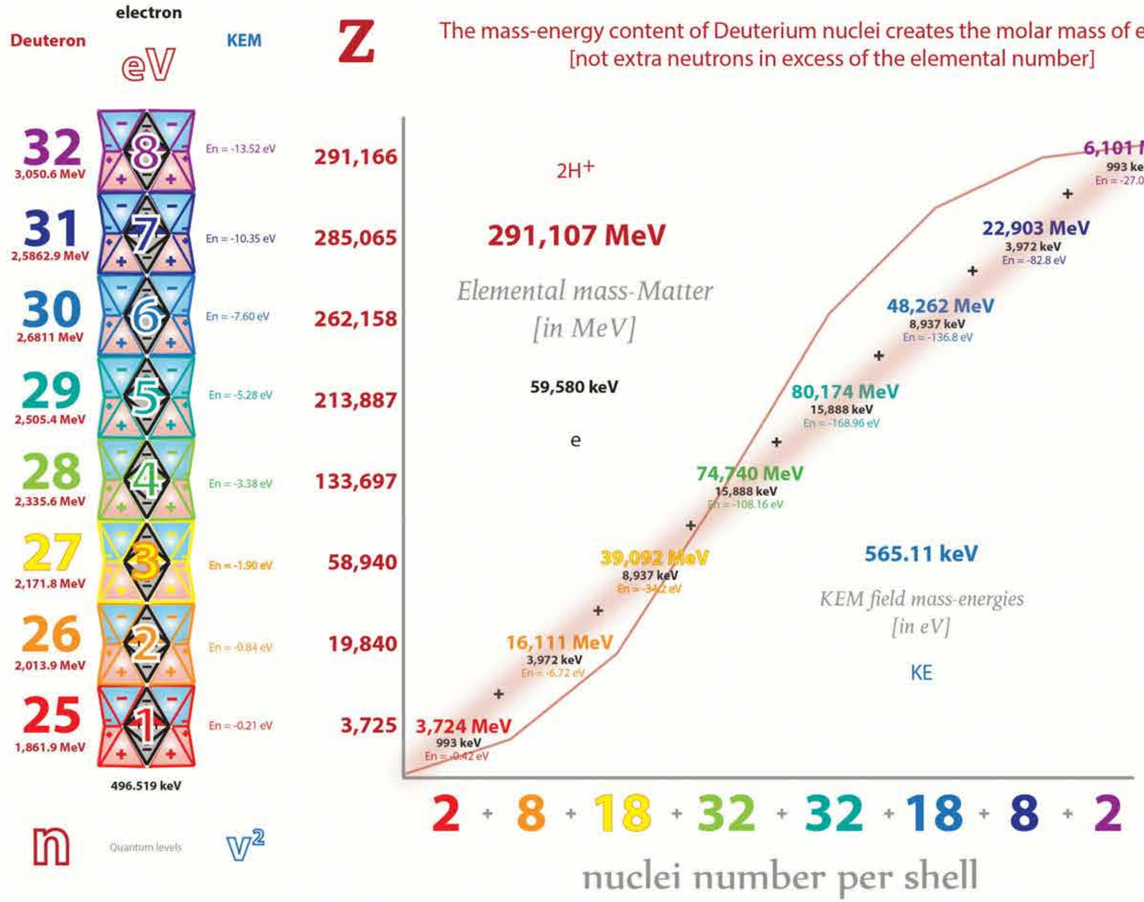

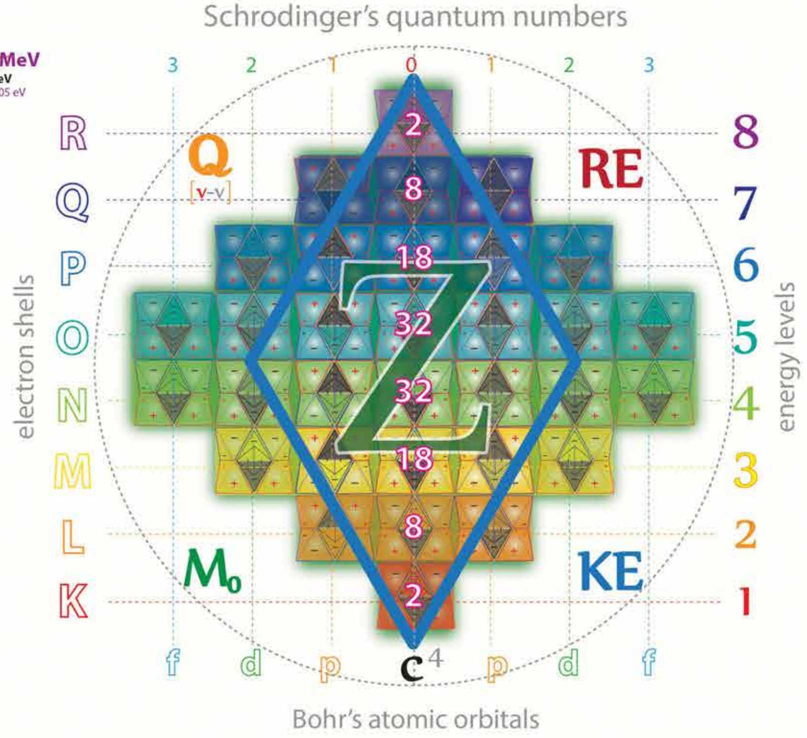
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M_0
KE
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons 930.947 MeV + **electrons** 496.519 keV + **KEM fields** 13.525 ev

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

All elements are comprised of n level Duetrium nuclei

The atomic shell energy levels of Deuterium nuclei in elements

Deuteron	electron eV	KEM
32 3,050.6 MeV	8	En = -13.52 eV
31 2,5862.9 MeV	7	En = -10.35 eV
30 2,6811 MeV	6	En = -7.60 eV
29 2,505.4 MeV	5	En = -5.28 eV
28 2,335.6 MeV	4	En = -3.38 eV
27 2,171.8 MeV	3	En = -1.90 eV
26 2,013.9 MeV	2	En = -0.84 eV
25 1,861.9 MeV	1	En = -0.21 eV

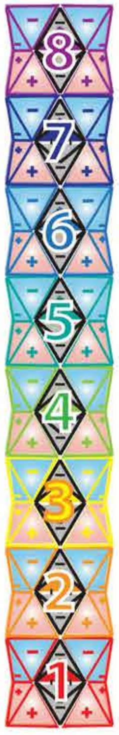

496.519 keV

n v^2

Determines the spectral line [KEM field energies] of electrons bound to them

Baryons	**KEM fields**	**electrons**
Z [[72n²]	+ [12v²]	+ [1.2e20]]
1,861,949 MeV	13.525 eV	496,519 keV

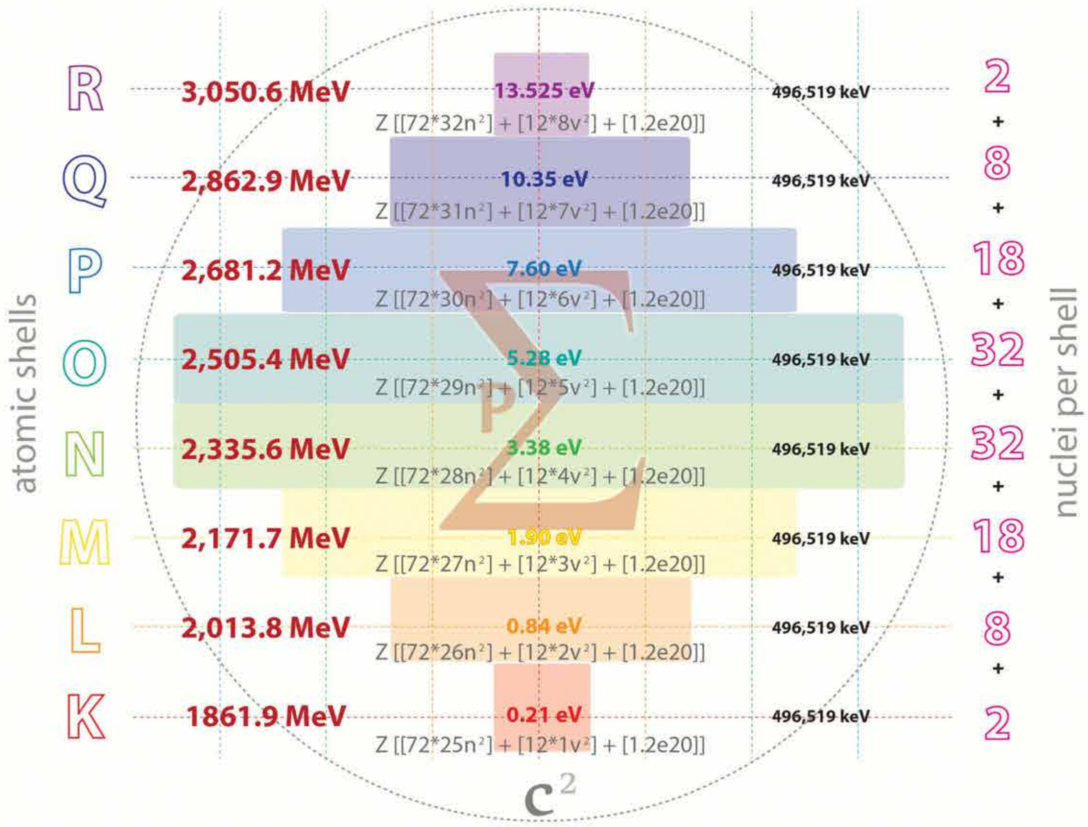

atomic shells	Baryons	KEM fields		electrons	nuclei per shell
R	3,050.6 MeV	13.525 eV	Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2
Q	2,862.9 MeV	10.35 eV	Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	+ 8
P	2,681.2 MeV	7.60 eV	Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	+ 18
O	2,505.4 MeV	5.28 eV	Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	+ 32
N	2,335.6 MeV	3.38 eV	Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	+ 32
M	2,171.7 MeV	1.90 eV	Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	+ 18
L	2,013.8 MeV	0.84 eV	Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	+ 8
K	1861.9 MeV	0.21 eV	Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	+ 2

$2H^+$ [72n²] **Z** γ [$M_e v^2$]

Elemental mass-Matter [in MeV]

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

the rest mass-Matter of bound photo-electrons is velocity invariant

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

Ionisation energies

Fractional quantum differentials

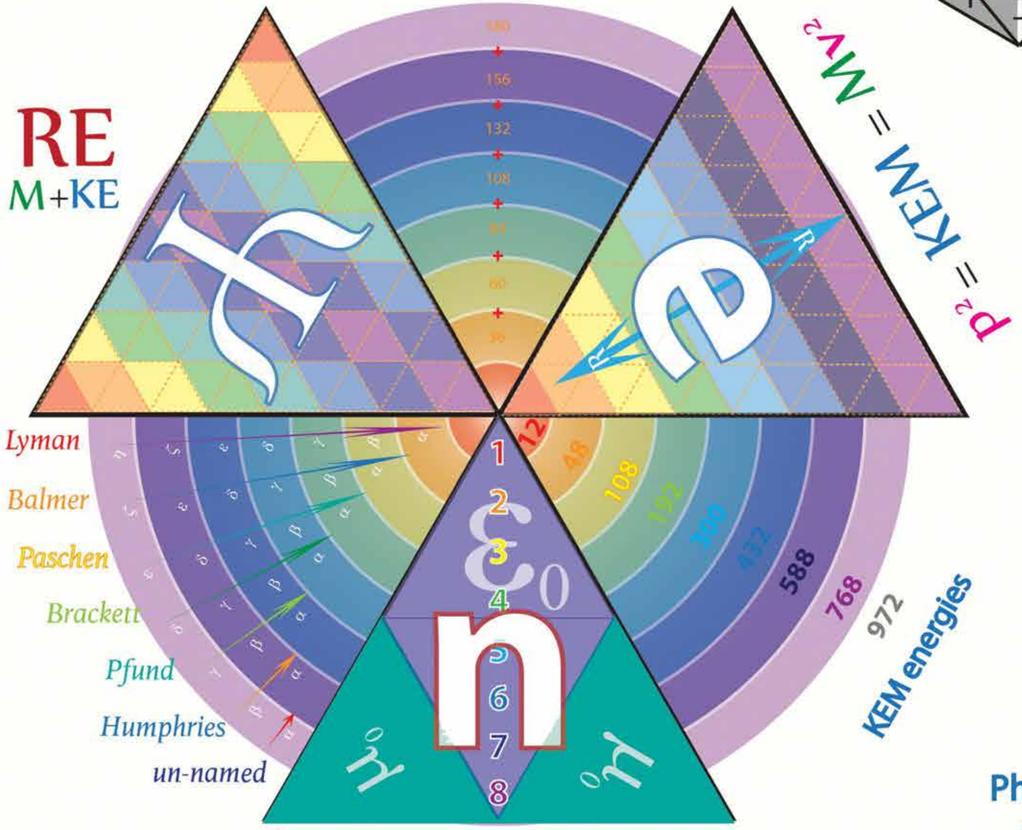
RE
M+KE
h
e
R
R
p^2 = KEM = Mv^2
36
60
84
108
132
156
180
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
12
48
108
192
300
432
588
768
972
KEM energies
1
2
3
4
5
6
7
8
ε_0
n

$$hf = \Delta Mv = \Delta p$$

spectral lines are produced by accelerating electrons

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

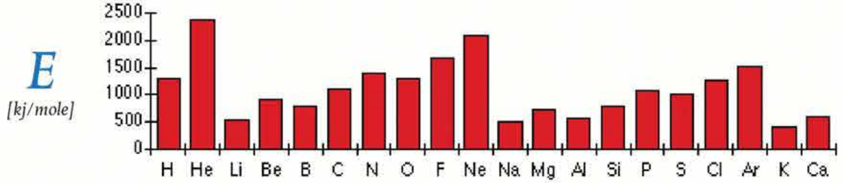

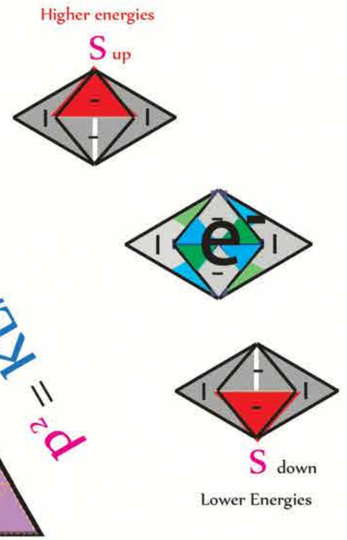
Higher energies
S up
e⁻
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

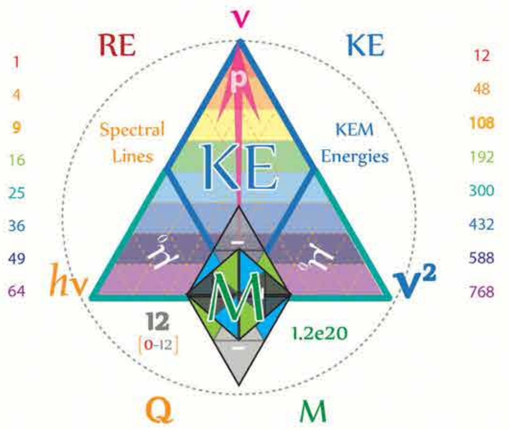
v
RE
KE
p
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768
Spectral Lines
KEM Energies
KE
hv
v^2
M
12
[0-12]
1.2e20
Q
M

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons that results from their transitions to specific energy nuclei in elements results in differing QAM quanta and produces spectral lines and fine line splitting

Spectral line transitions

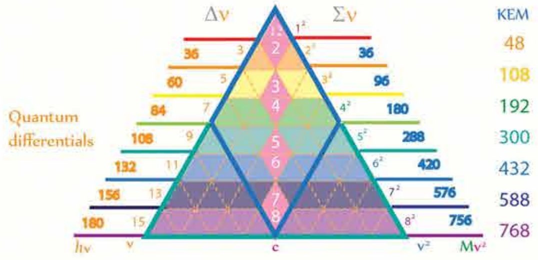
Δv
Σv
KEM
Quantum differentials
36
60
84
108
132
156
180
3
5
7
9
11
13
15
1
2
3
4
5
6
7
8
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
36
96
180
288
420
576
756
48
108
192
300
432
588
768
hv
v
c
v^2
Mv^2

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

1 R
2
3
4
4
3
2
1 K

$$P\sum 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	
Hydrogen	0	

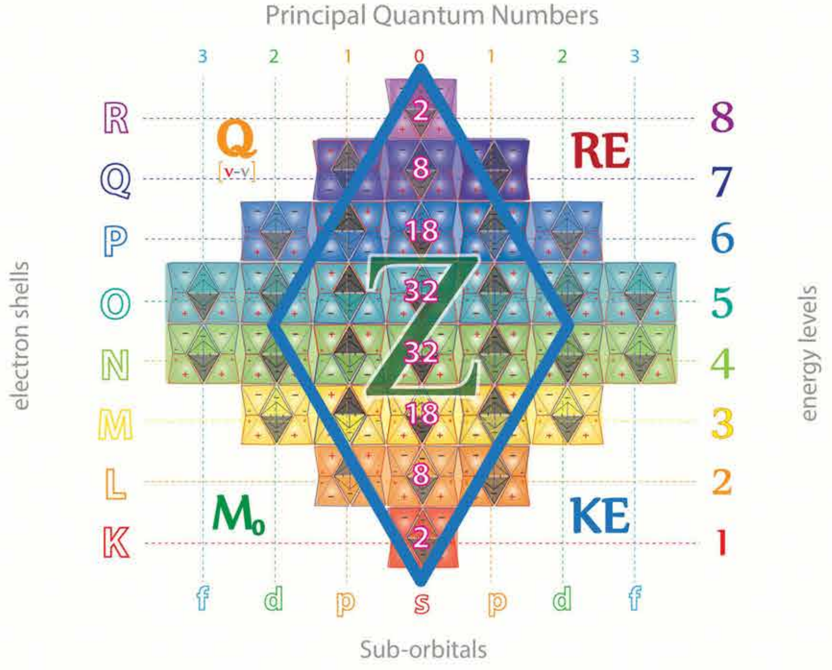
Principal Quantum Numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
[v-v]
RE
M0
KE
Z
2
8
18
32
32
18
8
2
electron shells
energy levels
f d p s p d f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\sideset{^R_K}{}{P}\sum_{25}^{32}\Big[\underbrace{[72(n)^2]}_{}+[12e19]+[m_e v^2]\Big]_1^8$$

Baryon rest masses — lepton rest mass — KEM

Deuterium mass-energy per shell

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

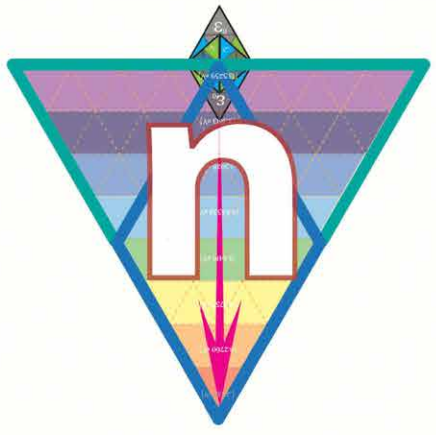

γ

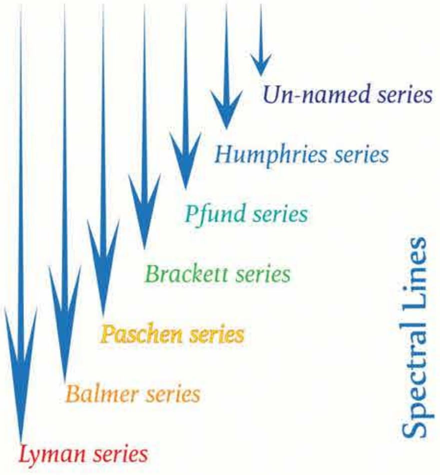
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$Mv^2 = KEM = hcR_H$$

Photon emission/absorption

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

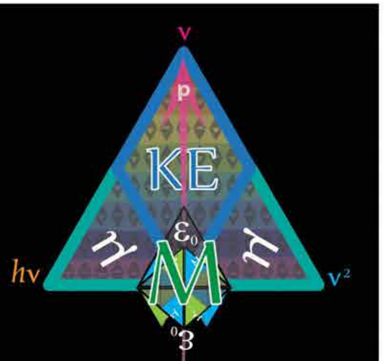
v
p
KE
ε_0
M
hv
v²
3

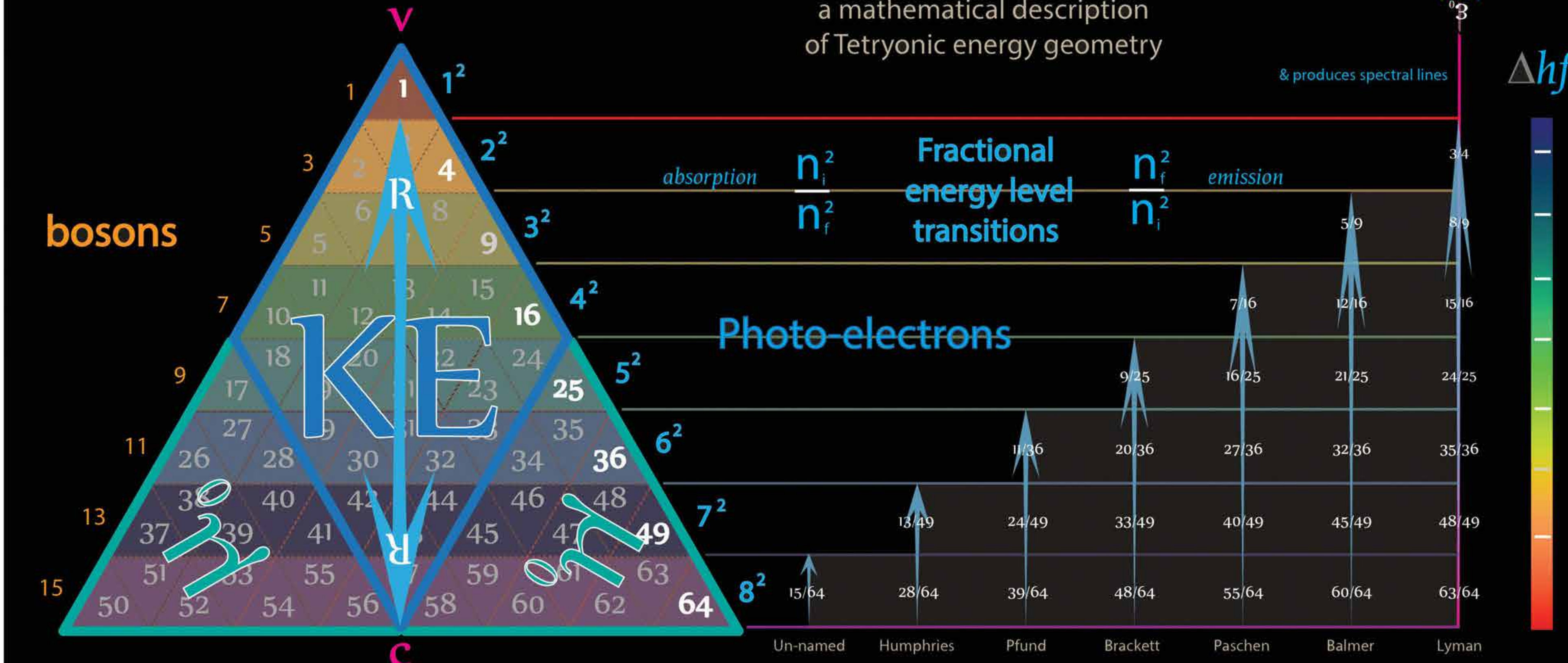
v
c
bosons
1 3 5 7 9 11 13 15
1² 2² 3² 4² 5² 6² 7² 8²
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
KE
R
absorption
$\frac{n_i^2}{n_f^2}$
Fractional energy level transitions
$\frac{n_f^2}{n_i^2}$
emission
Photo-electrons
& produces spectral lines
Δhf
3/4 8/9 15/16 24/25 35/36 48/49 63/64
5/9 12/16 21/25 32/36 45/49 60/64
7/16 16/25 27/36 40/49 55/64
9/25 20/36 33/49 48/64
11/36 24/49 39/64
13/49 28/64
15/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Ryberg's constant reflects the changing energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

All of the transitions of photo-electrons bound to Hydrogen atoms can now be revealed in the fractional geometry of KEM field energies

KEM field energies

Fermat's method of Factoring

also known as 'the difference of two squares' is used to factorise large numbers

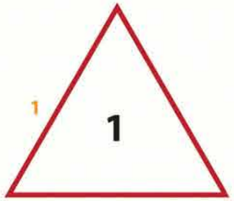

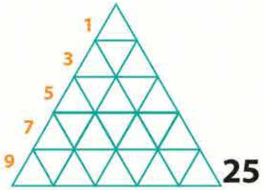

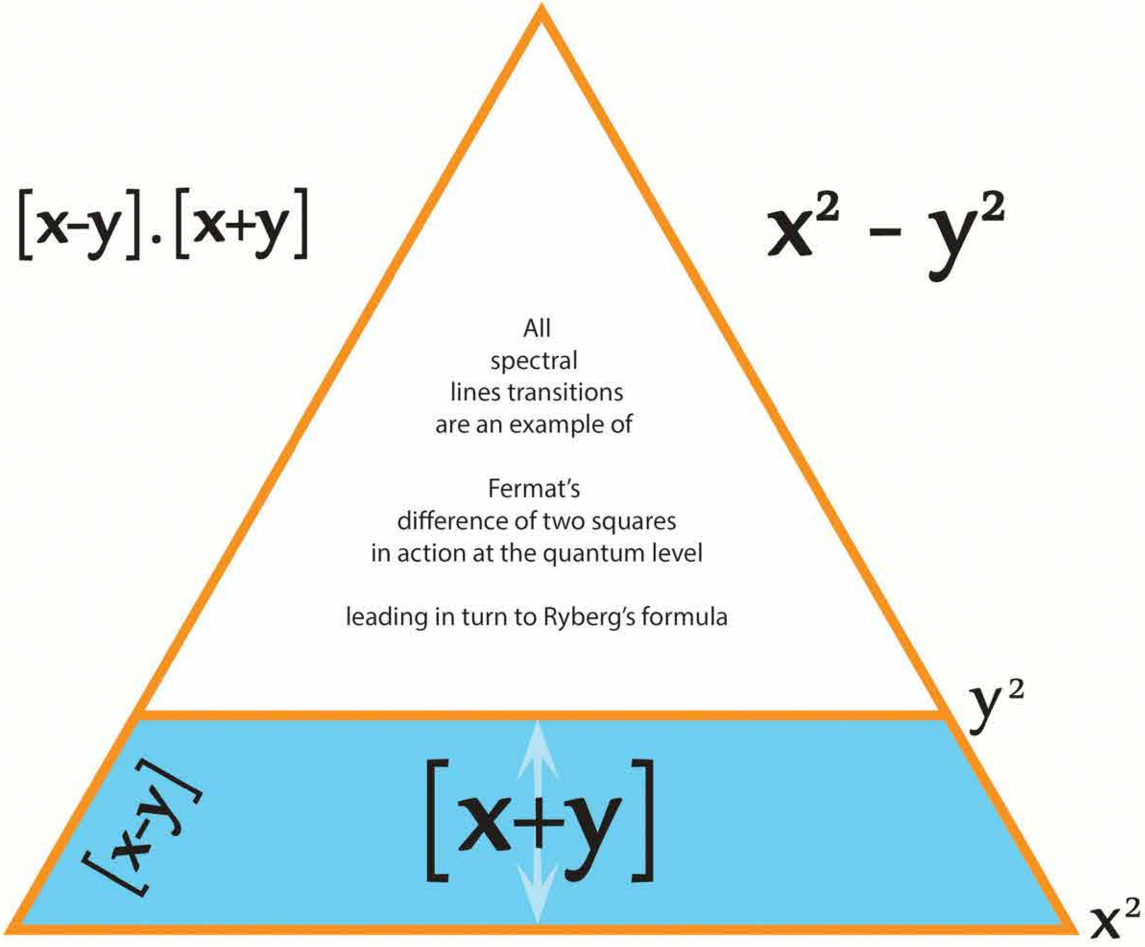

$[x-y].[x+y]$

$x^2 - y^2$

All
spectral
lines transitions
are an example of

Fermat's
difference of two squares
in action at the quantum level

leading in turn to Ryberg's formula

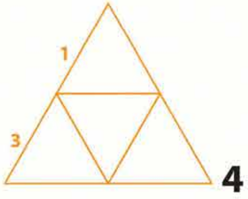

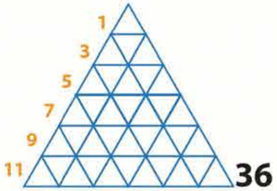

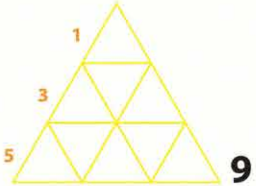

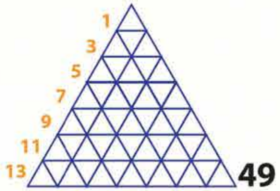
1
3
5
7
9
11
13
49

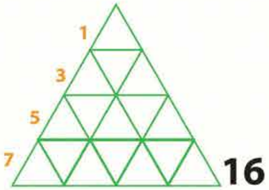

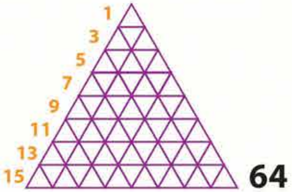
1
3
5
7
9
11
13
15
64

Fermat knew that every odd number could be written as the difference of two squares
or as revealed geometrically through Tetryonic theory's equilateral geometry,
every 'SQUARE' number is the sequential sum of ODD numbers

Copyright ABRAHAM [2008] - All rights reserved

Fractions

A fraction is a number that shows how many equal parts there are

In quantum mechanics fractions appear in quantum steps
as a result of the equilateral geometry of Planck energy momenta

3/4

24/25

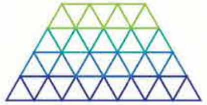

40/49

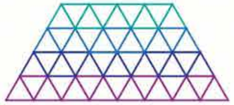

48/64

80/81

.987654321

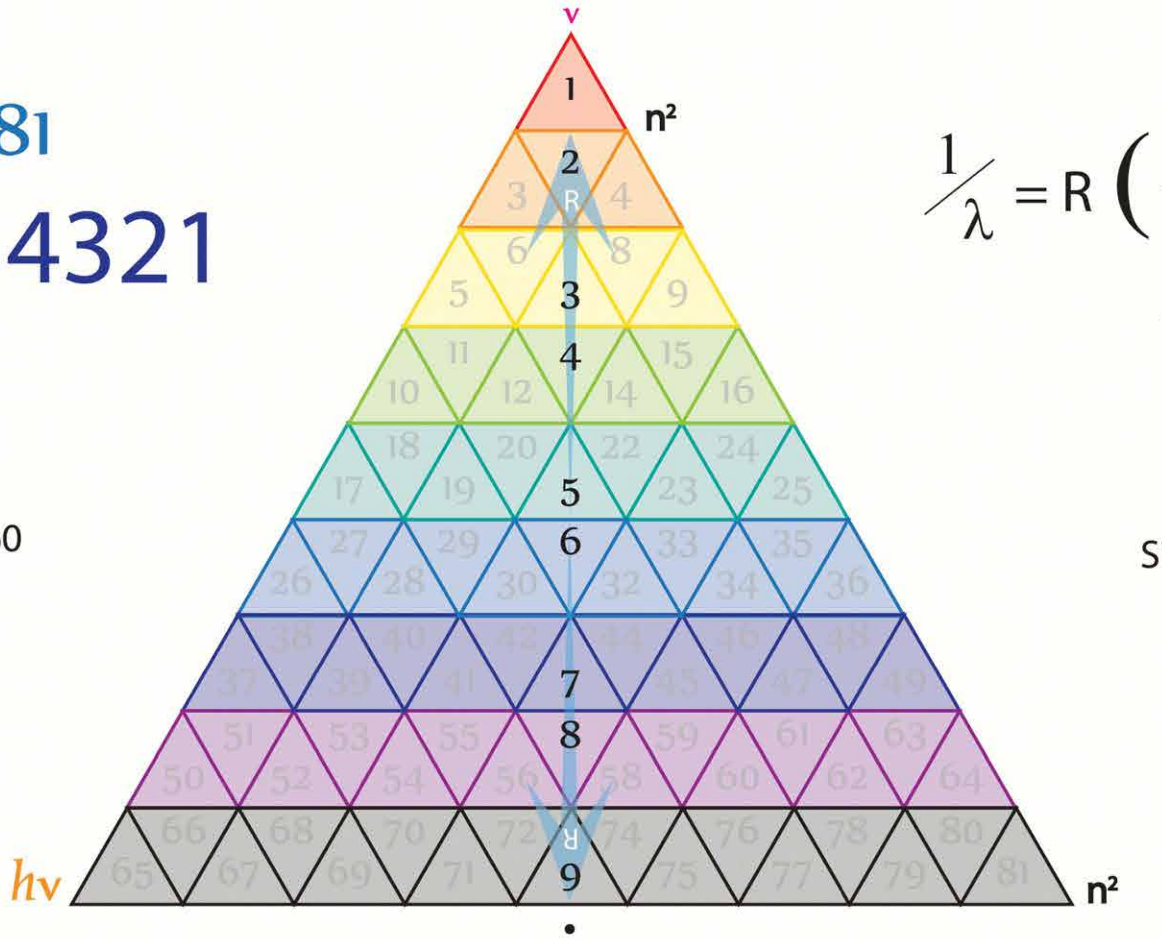
v
n²
1
2
3
4
5
6
7
8
9
R
R
3 4
6 8
5 9
11 15
10 12 14 16
18 20 22 24
17 19 23 25
27 29 33 35
26 28 30 32 34 36
38 40 42 44 46 48
37 39 41 45 47 49
51 53 55 59 61 63
50 52 54 56 58 60 62 64
66 68 70 72 74 76 78 80
65 67 69 71 75 77 79 81
hv
n²

$$\frac{1}{\lambda} = R\left(\frac{1}{1} - \frac{1}{81}\right)$$

.987654321

Cos 60

.5

1/2

Sin 60

.866025403

$\sqrt{3}/2$

Tetryonic geometry explains the fractional mathematics of Rydberg's formula

Copyright ABRAHAM [2008] - All rights reserved

Lyman spectral transitions

Δhv = KEM = hcR

Planck — 13.525 eV — Rydberg

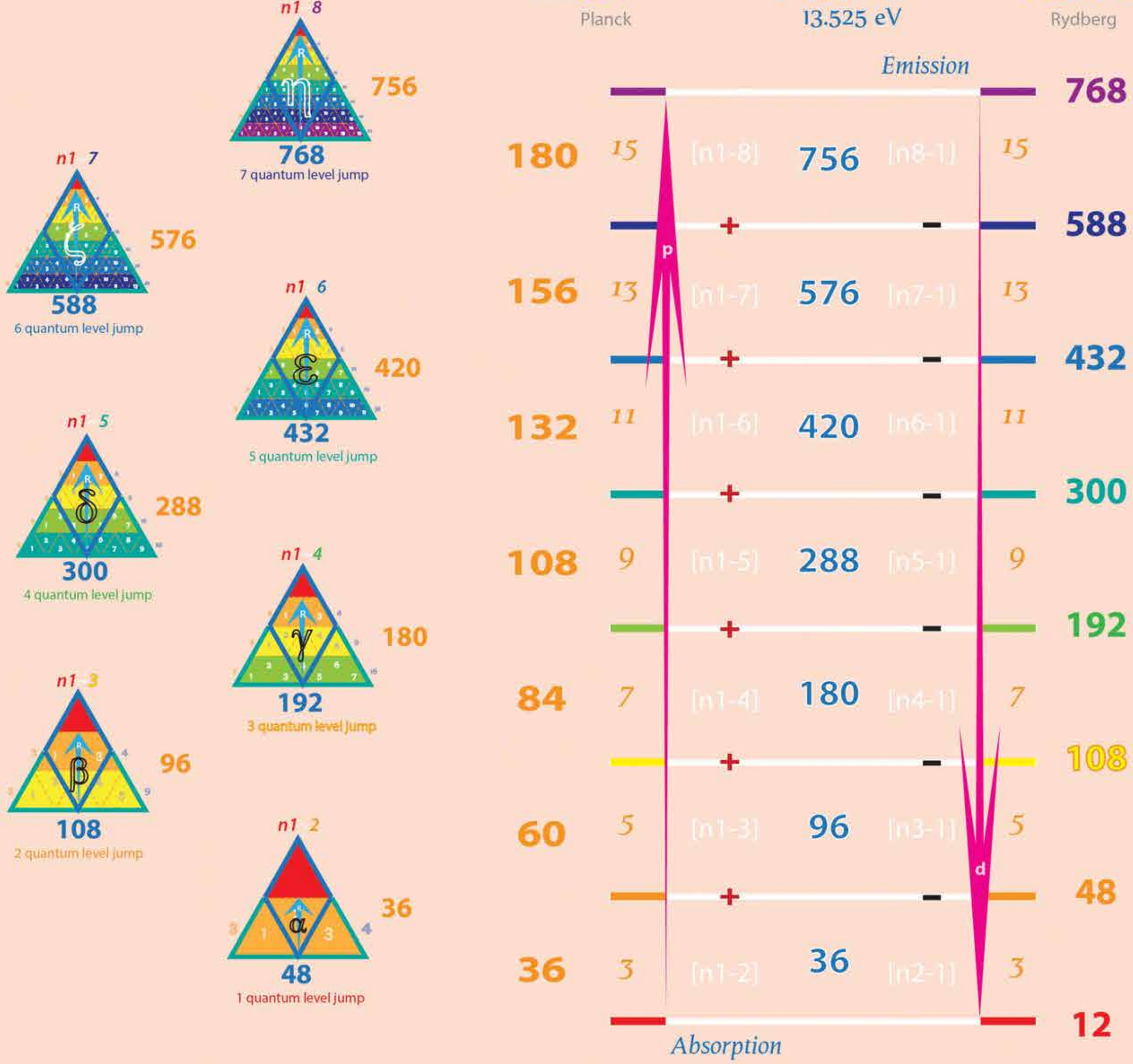
n1 8
756
768
7 quantum level jump
n1 7
576
588
6 quantum level jump
n1 6
420
432
5 quantum level jump
n1 5
288
300
4 quantum level jump
n1 4
180
192
3 quantum level jump
n1 3
96
108
2 quantum level jump
n1 2
36
48
1 quantum level jump
Emission
Absorption
p
d
180 15 [n1-8] 756 [n8-1] 15
156 13 [n1-7] 576 [n7-1] 13
132 11 [n1-6] 420 [n6-1] 11
108 9 [n1-5] 288 [n5-1] 9
84 7 [n1-4] 180 [n4-1] 7
60 5 [n1-3] 96 [n3-1] 5
36 3 [n1-2] 36 [n2-1] 3
+
−
768
588
432
300
192
108
48
12

		λ (nm)	$\tilde{\nu}$ (m-1)	f (Hz)	E (eV)	Ω (m²/s)	
63/64	.9843	93.17075598	10,732,981.5	$3.217666905 \times 10^{15}$	13.31395504	27.93188995	$R\left(\frac{1}{1}-\frac{1}{64}\right)$ .984375
48/49	.9795	93.62569132	10,680,829.01	$3.202031983 \times 10^{15}$	13.24926138	28.06827613	$R\left(\frac{1}{1}-\frac{1}{49}\right)$ .9795
35/36	.9722	94.33539043	10,600,475.55	$3.177942622 \times 10^{15}$	13.14958523	28.28103857	$R\left(\frac{1}{1}-\frac{1}{36}\right)$ .9722
24/25	.9600	95.53641971	10,467,212.43	$3.137991343 \times 10^{15}$	12.98427616	28.64109809	$R\left(\frac{1}{1}-\frac{1}{25}\right)$ .96
15/16	.9375	97.82929378	10,221,5587.14	$3.064444671 \times 10^{15}$	12.67995718	29.32848443	$R\left(\frac{1}{1}-\frac{1}{16}\right)$ .9375
8/9	.8888	103.1793333	9,691,863.362	$2.90554754 \times 10^{15}$	12.02247792	30.93238594	$R\left(\frac{1}{1}-\frac{1}{9}\right)$ .888
3/4	.7500	122.2866172	8,177,509.712	$2.451555737 \times 10^{15}$	10.14396575	36.66060556	$R\left(\frac{1}{1}-\frac{1}{4}\right)$ .75

ΔMv^2

Balmer spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck 3.381 eV Rydberg

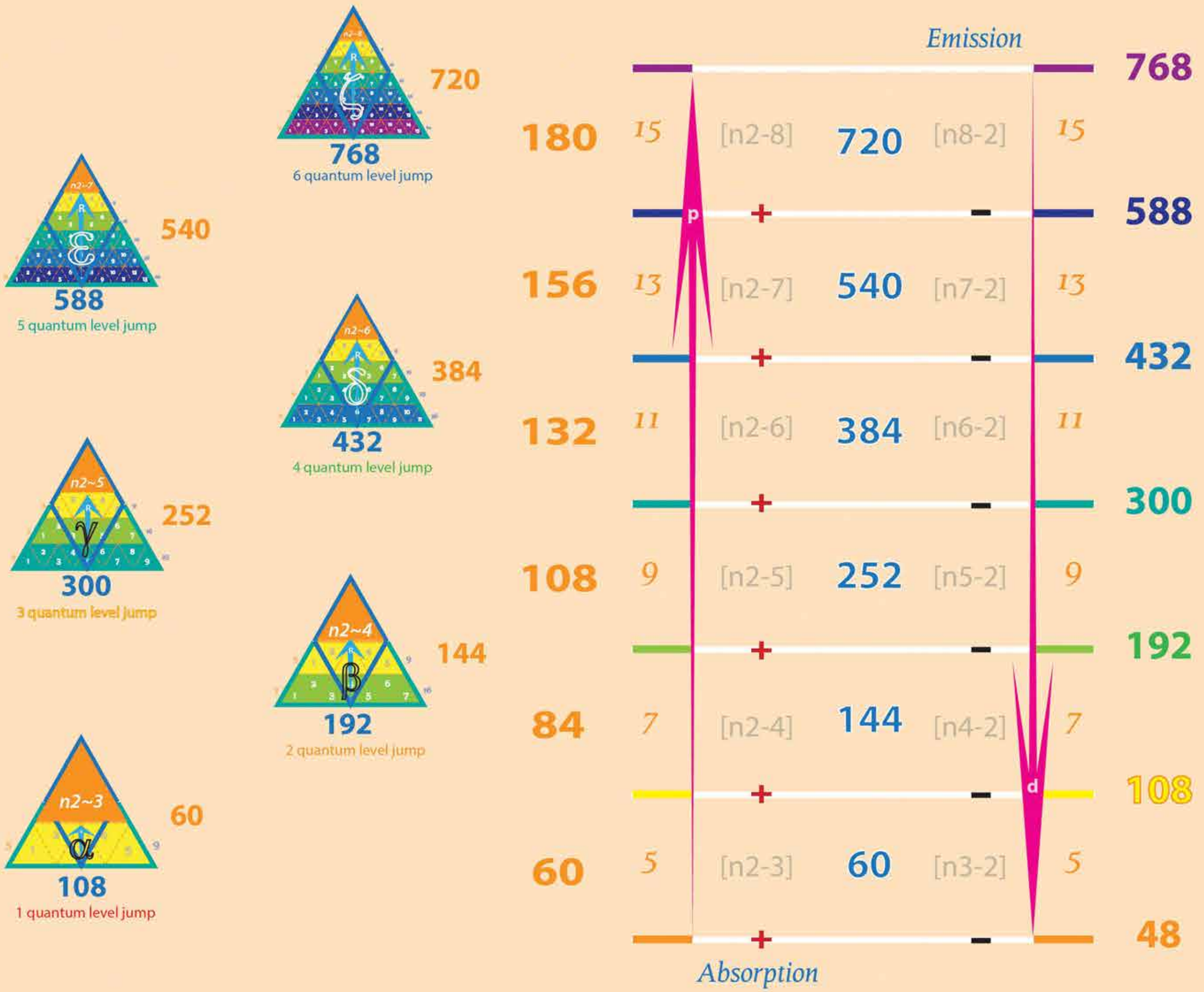
n2~8
R
ζ
720
768
6 quantum level jump
n2~7
R
ε
540
588
5 quantum level jump
n2~6
R
δ
384
432
4 quantum level jump
n2~5
R
γ
252
300
3 quantum level jump
n2~4
R
β
144
192
2 quantum level jump
n2~3
α
60
108
1 quantum level jump
Emission
768
180
15
[n2-8]
720
[n8-2]
15
p
+
-
588
156
13
[n2-7]
540
[n7-2]
13
+
-
432
132
11
[n2-6]
384
[n6-2]
11
+
-
300
108
9
[n2-5]
252
[n5-2]
9
+
-
192
84
7
[n2-4]
144
[n4-2]
7
d
+
-
108
60
5
[n2-3]
60
[n3-2]
5
+
-
48
Absorption

60/64
.2343

λ	391.3171751	nm
ṽ	2,555,471.1785	m-1
f	$7.661111677 \times 10^{14}$	Hz
E	**3.169989296**	**eV**
Ω	103.1079531	m^2/s

$R\left(\frac{1}{4}-\frac{1}{64}\right)$ .2343

45/49
.2295

λ	399.4696163	nm
ṽ	2,5033,319.299	m-1
f	$7.50476246 \times 10^{14}$	Hz
E	**3.105295637**	**eV**
Ω	101.0037092	m^2/s

$R\left(\frac{1}{4}-\frac{1}{49}\right)$ .2295

32/36
.2222

λ	412.7173331	nm
ṽ	2,422,965.84	m-1
f	$7.26386885 \times 10^{14}$	Hz
E	**3.00561948**	**eV**
Ω	97.76161483	m^2/s

$R\left(\frac{1}{4}-\frac{1}{36}\right)$ .2222

21/25
.2100

λ	436.7379187	nm
ṽ	2,289,702.719	m-1
f	$6.864356063 \times 10^{14}$	Hz
E	**2.840310409**	**eV**
Ω	92.38472601	m^2/s

$R\left(\frac{1}{4}-\frac{1}{25}\right)$ .21

12/16
.1875

λ	489.1464689	nm
ṽ	2,044,377.428	m-1
f	$6.128889342 \times 10^{14}$	Hz
E	**2.535991437**	**eV**
Ω	82.48636251	m^2/s

$R\left(\frac{1}{4}-\frac{1}{16}\right)$ .1875

5/9
.138

λ	660.347733	nm
ṽ	1,514,353.65	m-1
f	$4.539918031 \times 10^{14}$	Hz
E	**1.878512175**	**eV**
Ω	61.10100927	m^2/s

$R\left(\frac{1}{4}-\frac{1}{9}\right)$ .138

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Paschen spectral transitions

Δhv = KEM = hcR

Planck — 1.502 eV — Rydberg

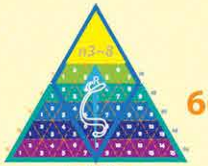

660

768

5 quantum level jump

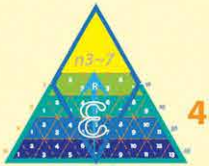

480

588

4 quantum level jump

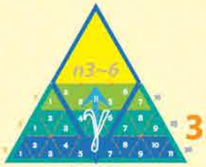

324

432

3 quantum level jump

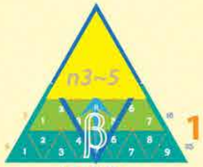

192

300

2 quantum level jump

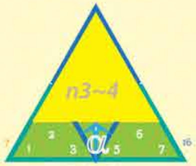

84

192

1 quantum level jump

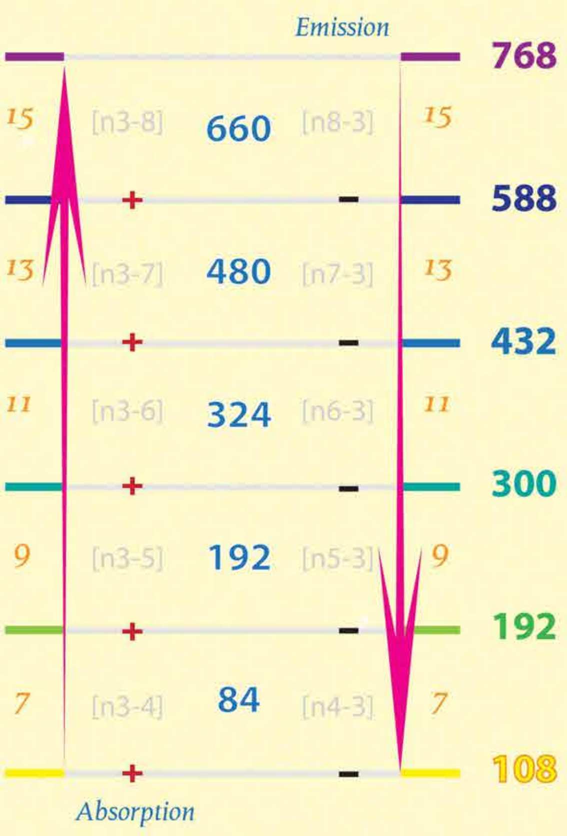

Emission

768

180 *15* [n3-8] 660 [n8-3] *15*

+ – 588

156 *13* [n3-7] 480 [n7-3] *13*

+ – 432

132 *11* [n3-6] 324 [n6-3] *11*

+ – 300

108 *9* [n3-5] 192 [n5-3] *9*

+ – 192

84 *7* [n3-4] 84 [n4-3] *7*

+ – 108

Absorption

55/64

.0954

λ	960.5057935	nm
$\tilde{v}$	1,041,118.135	m-1
f	3.121193646 x 10^{14}	Hz
E	**1.29147712**	**eV**
Ω	287.9523928	m^2/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$

.09548

40/49

.0907

λ	1,011.157466	nm
$\tilde{v}$	988,965.6492	m-1
f	2.9464844428 x 10^{14}	Hz
E	**1.226783461**	**eV**
Ω	303.1373822	m^2/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$

.0907

27/36

.8033

λ	1,100.579555	nm
$\tilde{v}$	908,612.1902	m-1
f	2.723950819 x 10^{14}	Hz
E	**1.127107305**	**eV**
Ω	329.94545	m^2/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$

.0833

16/25

.0711

λ	1,289.741666	nm
$\tilde{v}$	775,349.069	m-1
f	2.324438032 x 10^{14}	Hz
E	**0.961798233**	**eV**
Ω	386.6548243	m^2/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$

.0711

7/16

.0486

λ	1,886.707809	nm
$\tilde{v}$	530,023.7776	m-1
f	1.588971311 x 10^{14}	Hz
E	**0.657479261**	**eV**
Ω	365.6207715	m^2/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$

.0486

ΔMv^2

Brackett spectral transitions

$$\Delta h\nu \;=\; \text{KEM} \;=\; hcR$$

Planck

0.8451 eV

Rydberg

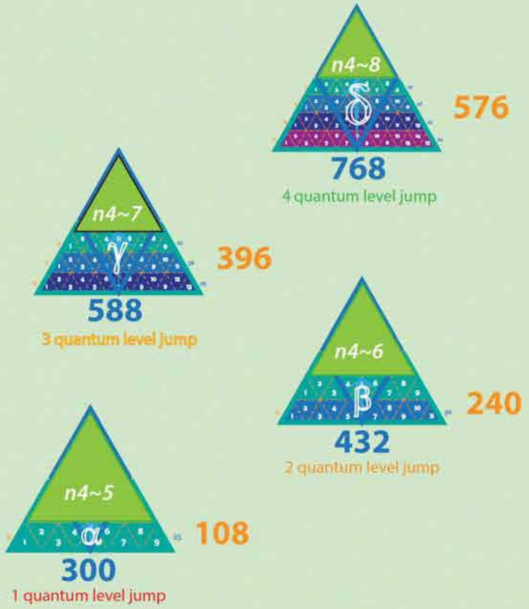
n4~8
δ
576
768
4 quantum level jump
n4~7
γ
396
588
3 quantum level jump
n4~6
β
240
432
2 quantum level jump
n4~5
α
108
300
1 quantum level jump

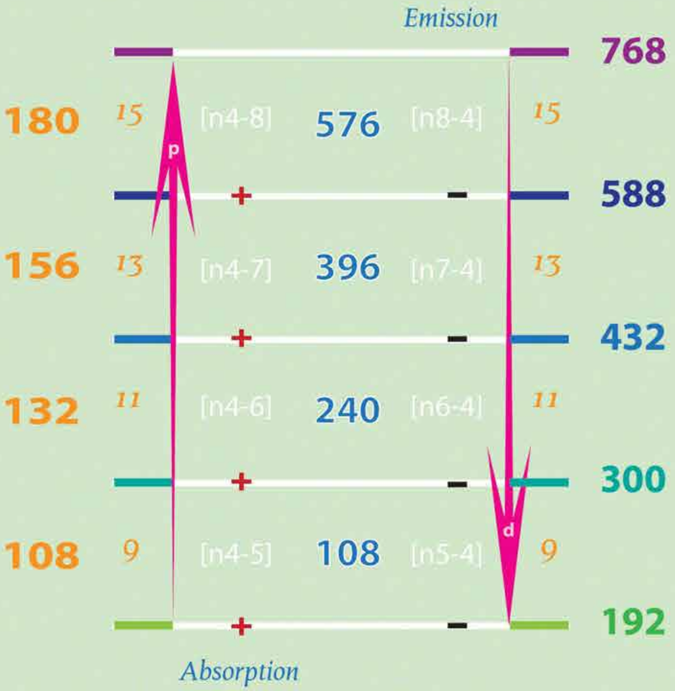
Emission
768
180
15
[n4-8]
576
[n8-4]
15
p
+
−
588
156
13
[n4-7]
396
[n7-4]
13
+
−
432
132
11
[n4-6]
240
[n6-4]
11
+
−
300
108
9
[n4-5]
108
[n5-4]
9
d
+
−
192
Absorption

48/64

.0468

λ	1,956.585876	nm
ṽ	511,094.357	m-1
ƒ	$1.532222335 \times 10^{14}$	Hz
E	**0.634001085**	**eV**
Ω	586.569689	m²/s

$R\left(\frac{1}{16}-\frac{1}{64}\right)$

.0468

33/49

.04209

λ	2,178.92518	nm
ṽ	458,941.8716	m-1
ƒ	$1.375873118 \times 10^{14}$	Hz
E	**0.569307096**	**eV**
Ω	653.2253354	m²/s

$R\left(\frac{1}{16}-\frac{1}{49}\right)$

.04209

20/36

.0347

λ	2,641.390932	nm
ṽ	378,588.4126	m-1
ƒ	$1.134979508 \times 10^{14}$	Hz
E	**0.469630433**	**eV**
Ω	791.8690801	m²/s

$R\left(\frac{1}{16}-\frac{1}{36}\right)$

.0347

9/25

.0225

λ	4,076.220574	nm
ṽ	245,325.2914	m-1
ƒ	$7.35466721 \times 10^{13}$	Hz
E	**0.30432052**	**eV**
Ω	1,222.020185	m²/s

$R\left(\frac{1}{16}-\frac{1}{25}\right)$

.0225

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Pfund spectral transitions

$$\Delta h\nu = \mathrm{KEM} = hc\mathrm{R}$$

Planck

0.541 eV

Rydberg

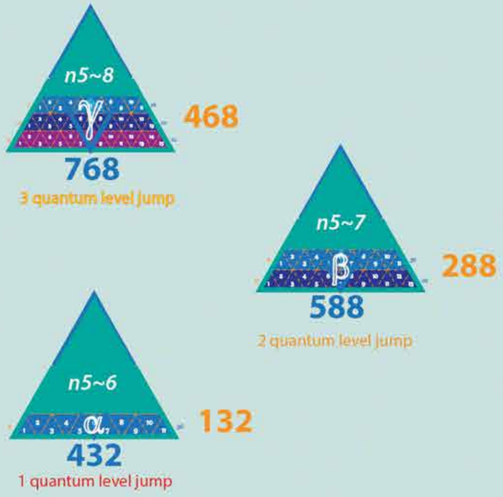
n5~8
γ
468
768
3 quantum level jump
n5~7
β
288
588
2 quantum level jump
n5~6
α
132
432
1 quantum level jump

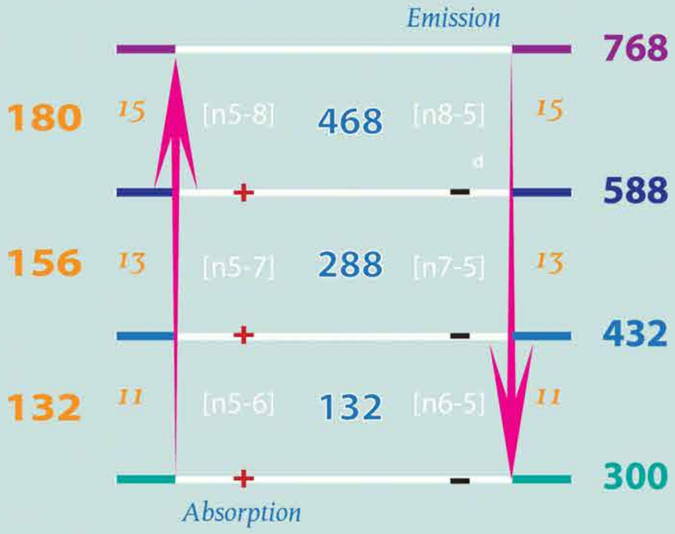
Emission
768
180
15
[n5-8]
468
[n8-5]
15
+
-
588
156
13
[n5-7]
288
[n7-5]
13
+
-
432
132
11
[n5-6]
132
[n6-5]
11
+
-
300
Absorption

39/64

.0243

λ	3,762.665146	nm
ṽ	265,769.0656	m-1
f	$7.967556145 \times 10^{13}$	Hz
E	0.329678886	eV
Ω	1,128.018633	m²/s

$R\left(\frac{1}{25}-\frac{1}{64}\right)$

.0243

24/49

.0195

λ	4,681.284566	nm
ṽ	213,616.5802	m-1
f	$6.404063965 \times 10^{13}$	Hz
E	0.264985227	eV
Ω	1,403.413807	m²/s

$R\left(\frac{1}{25}-\frac{1}{49}\right)$

.01959

11/36

.0122

λ	7,503.951512	nm
ṽ	133,263.1212	m-1
f	$3.995127867 \times 10^{13}$	Hz
E	0.165309071	eV
Ω	2,249.628068	m²/s

$R\left(\frac{1}{25}-\frac{1}{36}\right)$

.0122

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Humphreys spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck — 0.375 eV — Rydberg

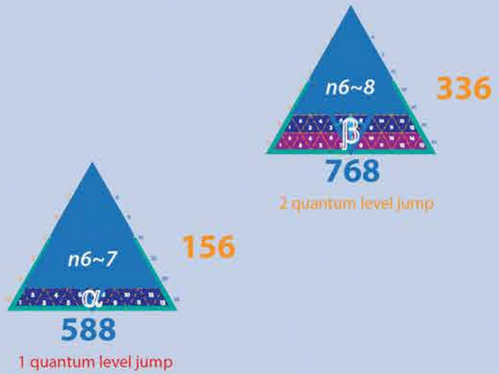
n6~8
336
β
768
2 quantum level jump
n6~7
156
α
588
1 quantum level jump

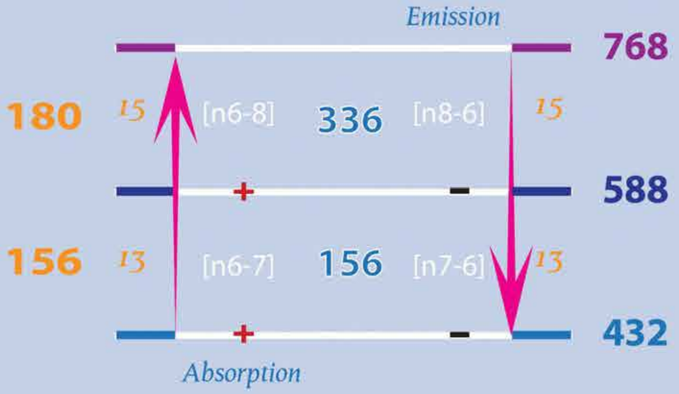
Emission
768
180
15
[n6-8]
336
[n8-6]
15
+
−
588
156
13
[n6-7]
156
[n7-6]
13
+
−
432
Absorption

28/64

.01215

λ	7,546.831235	nm
ṽ	132,505.9444	m-1
ƒ	3.972428277 x 10^{13}	Hz
E	**0.164369815**	**eV**
Ω	1,128.018633	m^2/s

$R\left(\frac{1}{36} - \frac{1}{64}\right)$

.01215

13/49

.00733

λ	12,445.01497	nm
ṽ	80,353.459	m-1
ƒ	2.4089936098x 10^{13}	Hz
E	**0.099676156**	**eV**
Ω	1,403.413807	m^2/s

$R\left(\frac{1}{36} - \frac{1}{49}\right)$

.0073

$\Delta M v^2$

Un-named spectral transition

$\Delta h\nu$ = KEM = hcR

Planck 0.276 eV Rydberg

$\Delta\nu$ $\Sigma\nu$ $M\nu^2$

180

768

1 quantum level jump

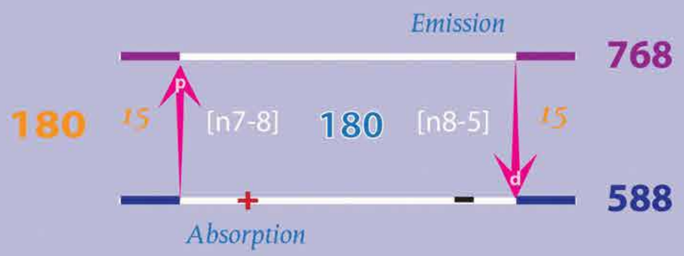
Emission
768
180
15
p
[n7-8]
180
[n8-5]
15
d
+
−
588
Absorption

15/64

.0047

λ	19,174.54158	nm
$\tilde{\nu}$	52,152.48541	m-1
f	1.563492179 x 10^{13}	Hz
E	0.064693658	eV
Ω	5,748.382952	m^2/s

$R\left(\frac{1}{49}-\frac{1}{64}\right)$

.0047

$\Delta M\nu^2$

Copyright ABRAHAM [2008] - All rights reserved

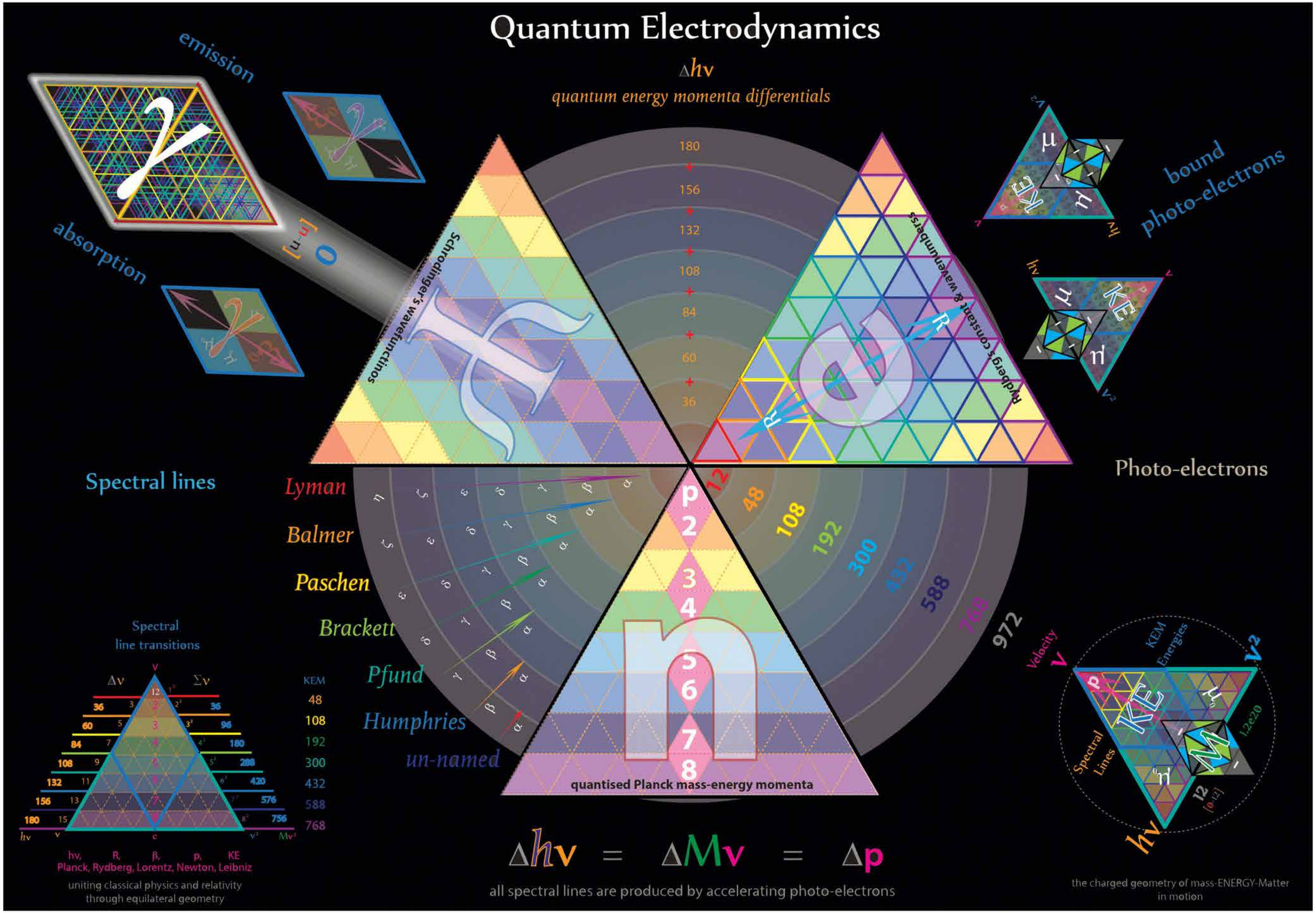
Quantum Electrodynamics
Δhv
quantum energy momenta differentials
emission
absorption
[n-u]
0
Schrodinger's wavefunctions
Rydberg's constant & wavenumbers
R
R
180
156
132
108
84
60
36
bound
photo-electrons
μ
KE
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
η ζ ε δ γ β α
12
48
108
192
300
432
588
768
972
p
2
3
4
5
6
7
8
n
quantised Planck mass-energy momenta
Spectral
line transitions
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity
through equilateral geometry
Velocity
v
v²
KEM
Energies
1.2e20
Spectral
Lines
12
hv
M
the charged geometry of mass-ENERGY-Matter
in motion
Δhv = ΔMv = Δp
all spectral lines are produced by accelerating photo-electrons

Tetryonics 96.11 - Atomic spectral series transitions

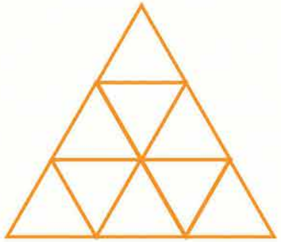

ninths

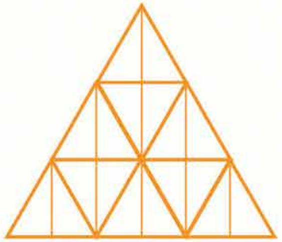

eigteenths

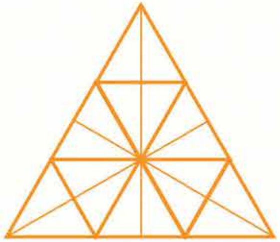

Eighteenths

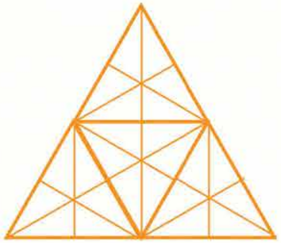

Twenty-fourths

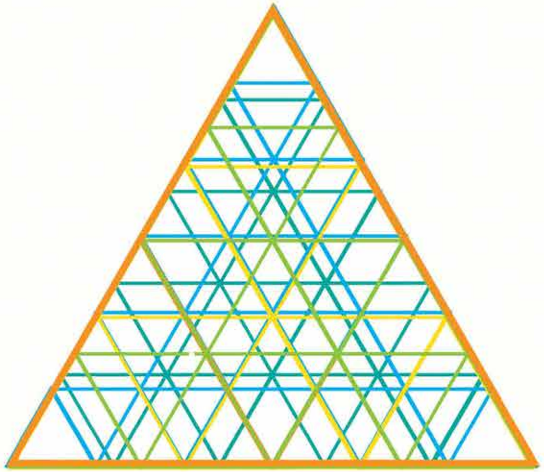

Fractionals and fractals

White light
[EM radiation]
is comprised of many
superpositioned
frequencies

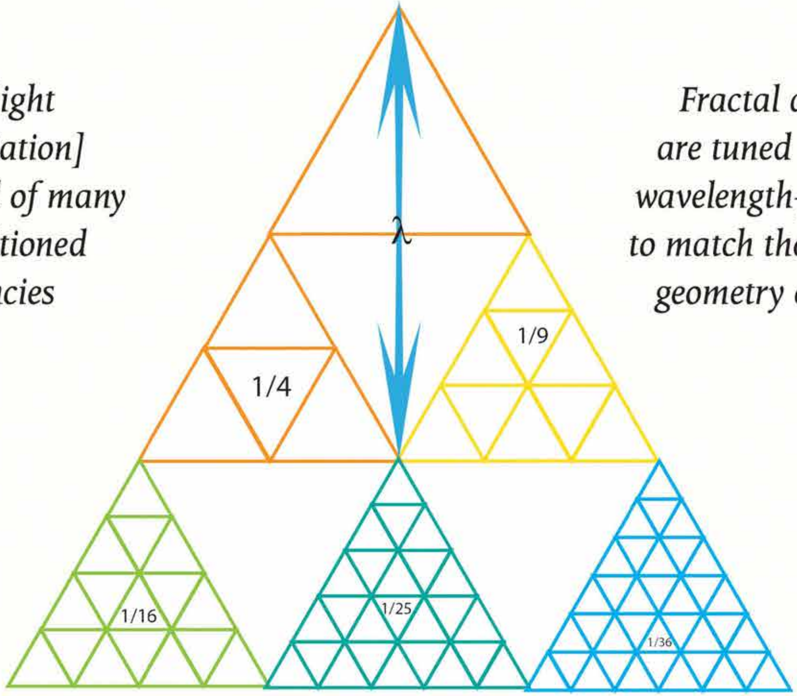

Fractal antennas
are tuned to specific
wavelength-frequencies
to match the equilateral
geometry of photons

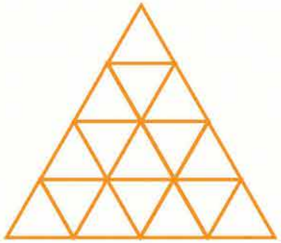

Sizteenths

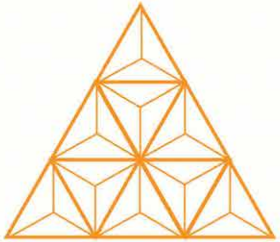

Twenty-sevenths

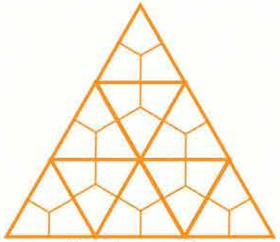

Twenty-sevenths

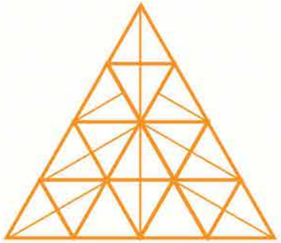

Thirty-seconds

Copyright ABRAHAM [2008] - All rights reserved

Polygons and Angles

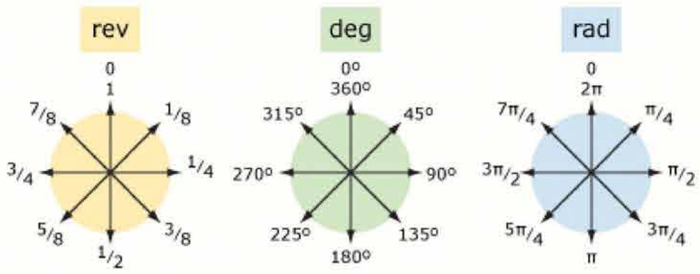
rev
0
1
1/8
1/4
3/8
1/2
5/8
3/4
7/8
deg
0°
360°
45°
90°
135°
180°
225°
270°
315°
rad
0
2π
π/4
π/2
3π/4
π
5π/4
3π/2
7π/4

polygon exterior and interior angles

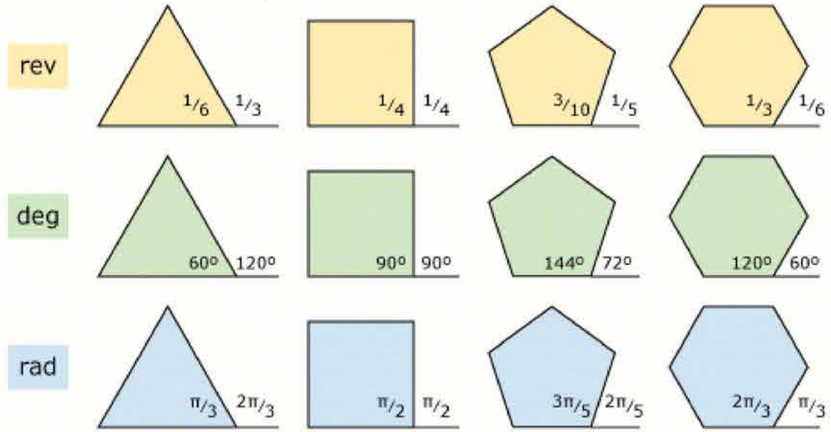
rev
1/6
1/3
1/4
1/4
3/10
1/5
1/3
1/6
deg
60°
120°
90°
90°
144°
72°
120°
60°
rad
π/3
2π/3
π/2
π/2
3π/5
2π/5
2π/3
π/3

dividing the circle

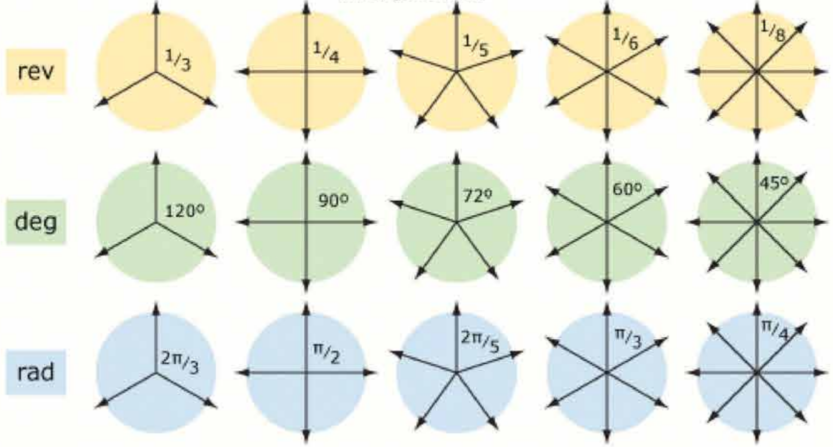
rev
1/3
1/4
1/5
1/6
1/8
deg
120°
90°
72°
60°
45°
rad
2π/3
π/2
2π/5
π/3
π/4

supplementary angles

complementary angles

triangles

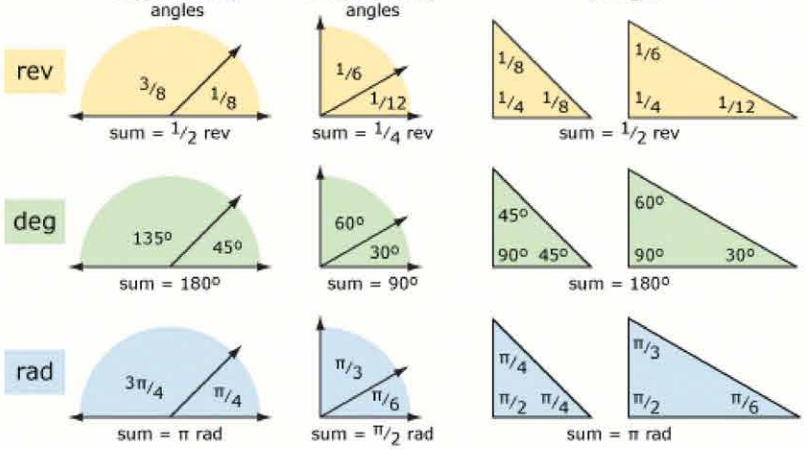
rev
3/8
1/8
sum = 1/2 rev
1/6
1/12
sum = 1/4 rev
1/8
1/4
1/8
1/6
1/4
1/12
sum = 1/2 rev
deg
135°
45°
sum = 180°
60°
30°
sum = 90°
45°
90°
45°
60°
90°
30°
sum = 180°
rad
3π/4
π/4
sum = π rad
π/3
π/6
sum = π/2 rad
π/4
π/2
π/4
π/3
π/2
π/6
sum = π rad

Copyright ABRAHAM [2008] - All rights reserved

Pi vs Tau

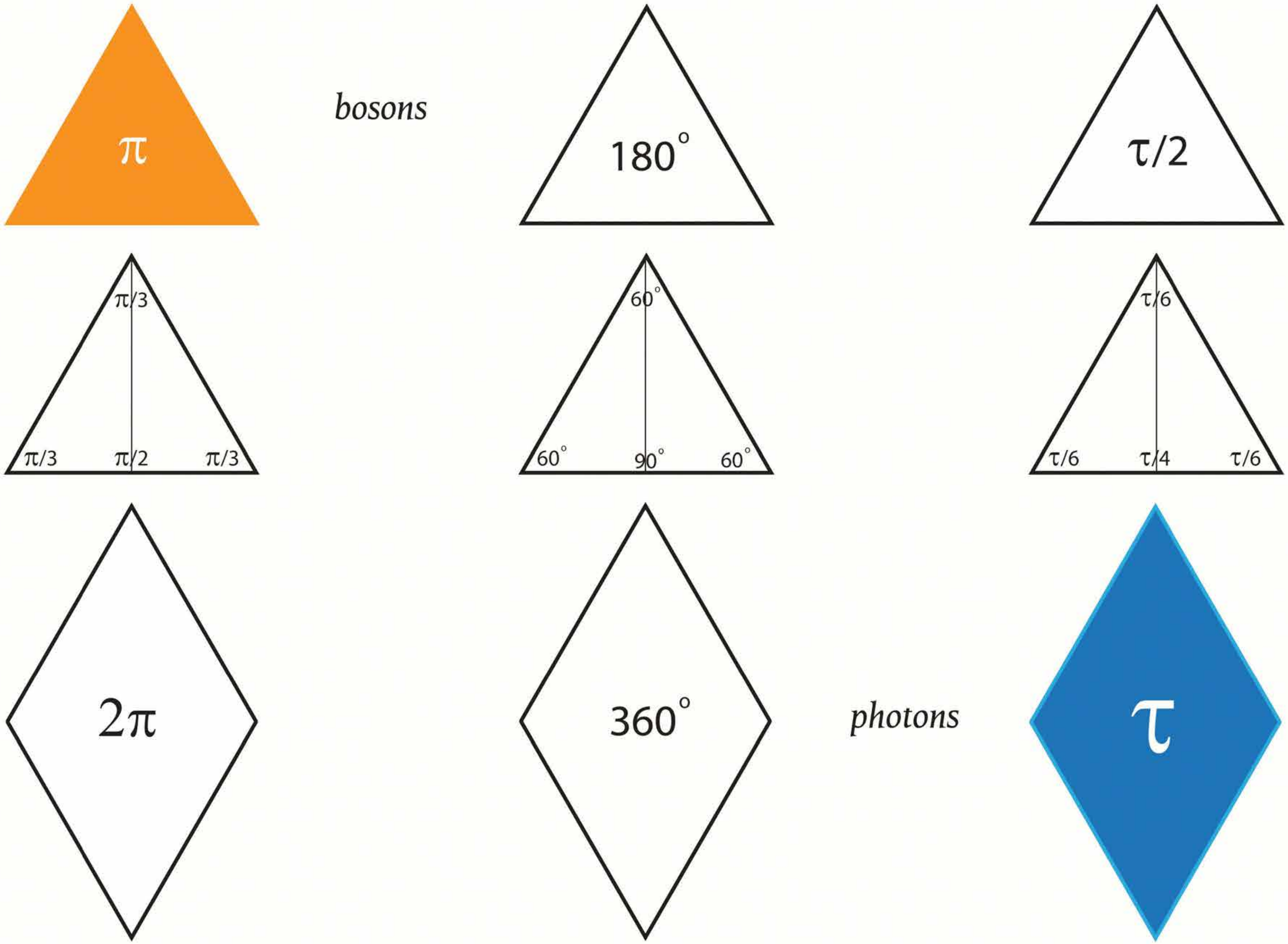
π
bosons
180°
τ/2
π/3
π/3
π/2
π/3
60°
60°
90°
60°
τ/6
τ/6
τ/4
τ/6
2π
360°
photons
τ

Triangular dissection of an equilateral triangle

is a way of dividing up a original triangle into smaller equilateral triangles,
such that none of the smaller triangles overlap

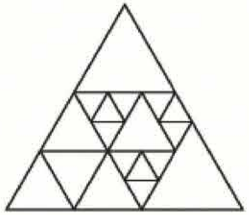

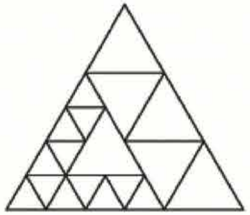

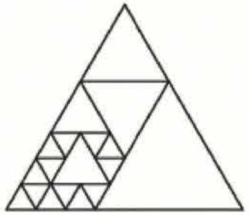

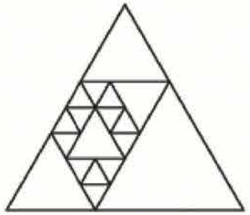

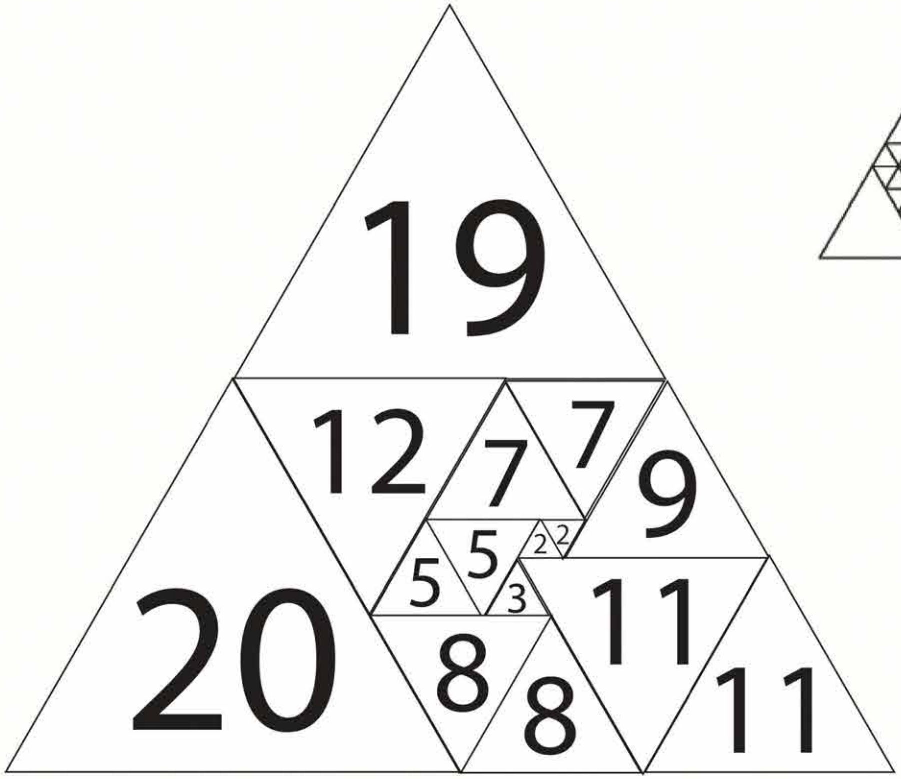

lowest order perfect equilateral triangle dissected by equilateral triangles

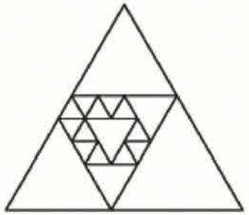

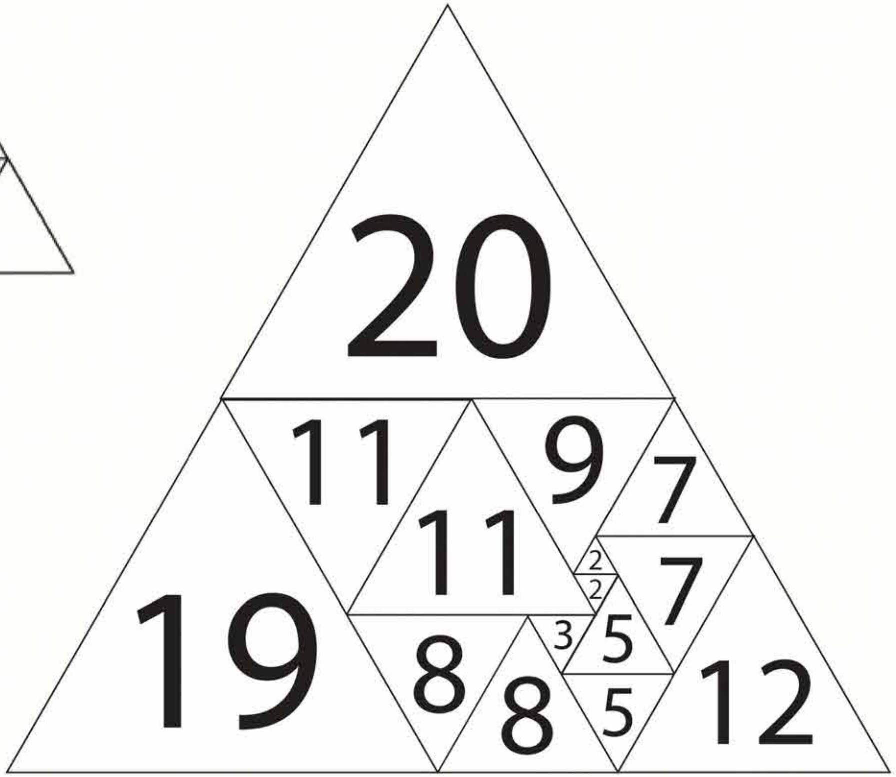

lowest order perfect dissected equilateral triangle, an isomer of the first

Golden mean Spirals

Golden Mean Spiral – This spiral is derived via the golden rectangle, a unique rectangle which has the golden ratio.
This form is found everywhere in nature: the Nautilus Shell, the face of a Sunflower, fingerprints, our DNA, and the shape of the Milky Way

Golden spiral

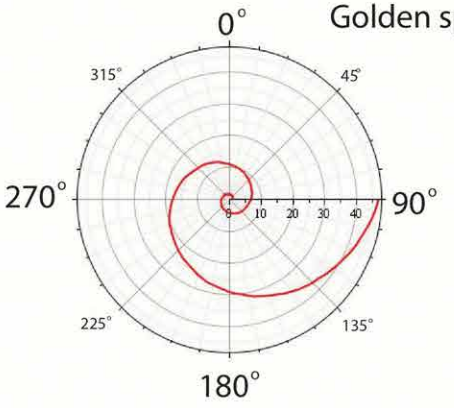

$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, \ldots$

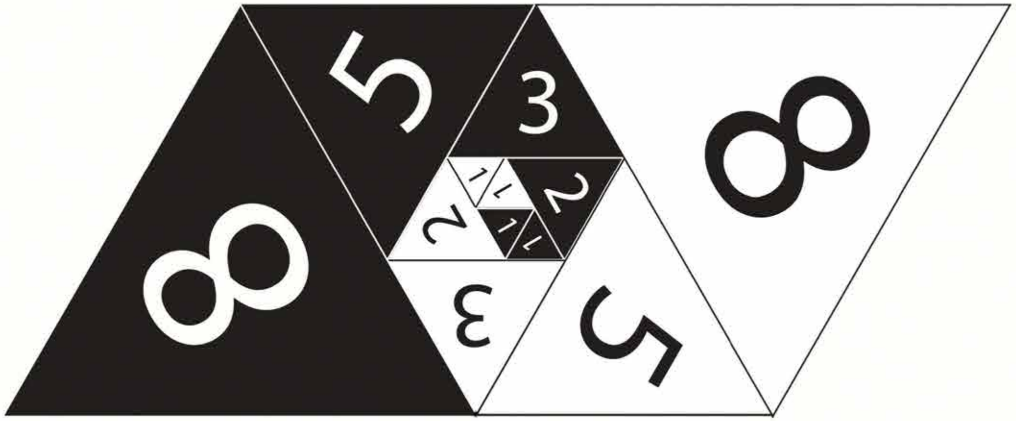

Golden spirals

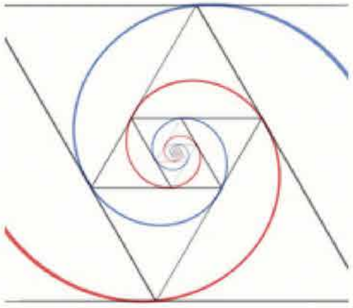

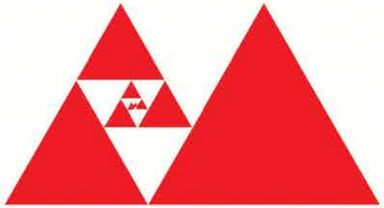

Continued fractions and the Fibonacci sequence

The convergence of the continued fractions

$$1+\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\,\cdots$$

are:

$$\frac{2}{1},\ \frac{3}{2},\ \frac{5}{3},\ \frac{8}{5},\ \frac{13}{8},\ \frac{21}{13},\ \frac{34}{21},\ \frac{55}{34},\ \frac{89}{55},\ \frac{144}{89},\ \frac{233}{144},\ \frac{377}{233},\ \frac{610}{377},$$

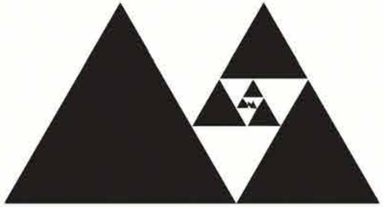

The Golden Ratio (Golden Mean, Golden Section)
is defined mathematically as:

$$\varphi = \frac{1+\sqrt{5}}{2} = 1.6180339887\ldots.$$

Copyright ABRAHAM [2008] - All rights reserved

Koch fractal Curve

Niels Fabian Helge von Koch (January 25, 1870 – March 11, 1924) was a Swedish mathematician
who gave his name to one of the earliest fractal curves ever known

He described the Koch curve, or Koch snowflakes as it popularly known, in a 1904 paper entitled
“On a continuous curve without tangents constructible from elementary geometry”

Von Koch wrote several papers on number theory .
One of his results was a 1901 theorem proving that
the Riemann hypothesis is equivalent to a strengthened
form of the prime number theorem.

Three Koch curves form the snowflake.

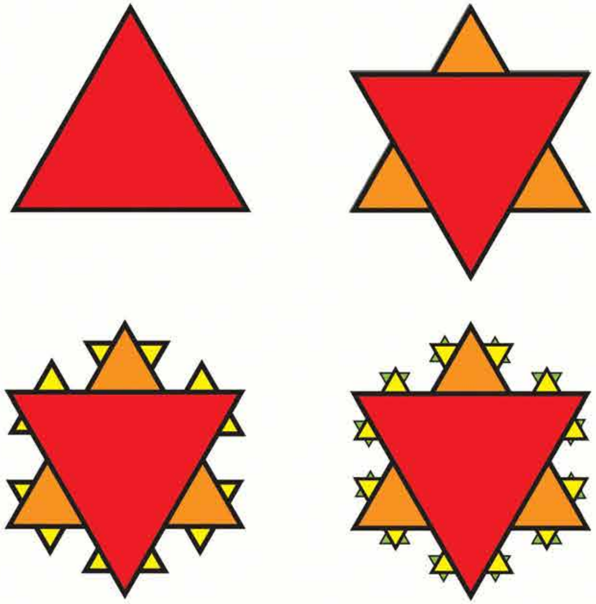

The Koch snowflake (or Koch star) is a mathematical curve
and one of the earliest fractal curves to have been described.

Actually Koch described what is now known as the Koch curve,
which is the same as the now popular snowflake, except it starts
with a line segment instead of an equilateral triangle.

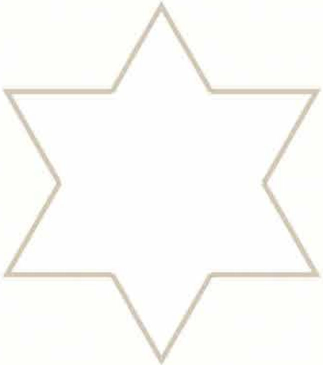

$$a = \frac{1}{2} + \frac{i}{\sqrt{12}}$$

The Koch curve is a special case of the Cesaro curve where:

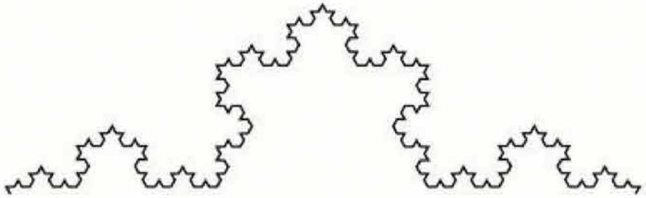

which is in turn a special case of the de Rham curve.

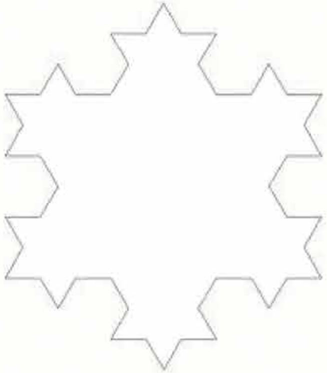

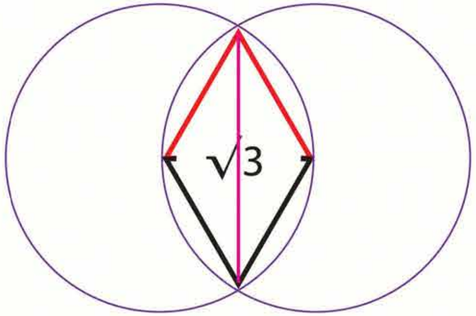

vesica piscis

$$\frac{1351}{780} > \sqrt{3} > \frac{265}{153}\,.$$

Koch snowflake

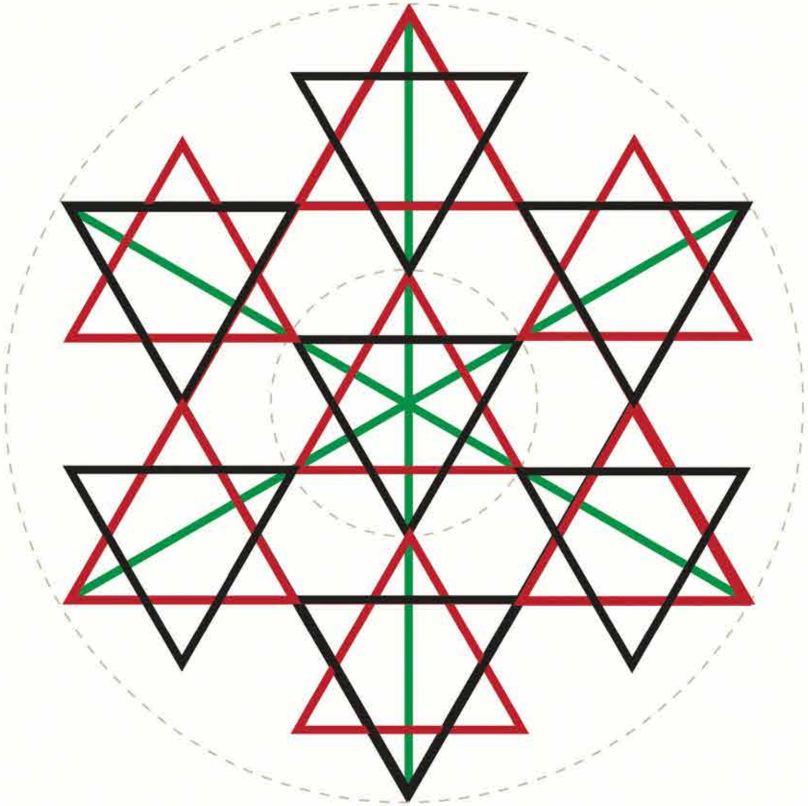

space-time

star tetrahedron

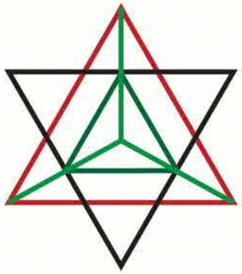

Flower of Life

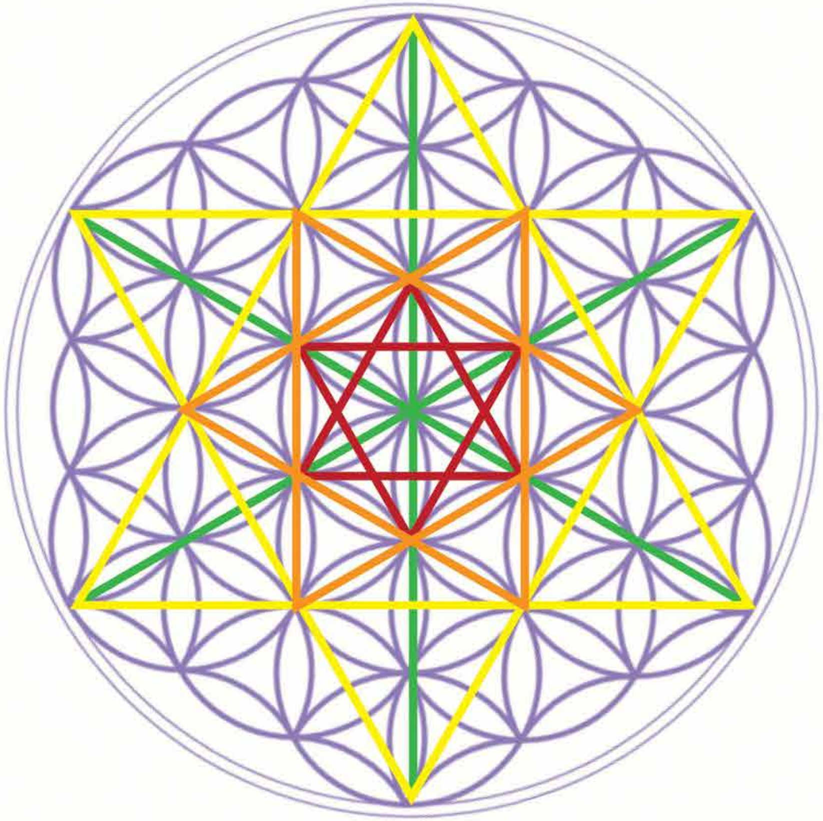

The Flower of Life is a name for a geometrical figure composed of multiple evenly-spaced, overlapping circles.

Unit circles - SINE WAVES - Photons

Classic model of a photon

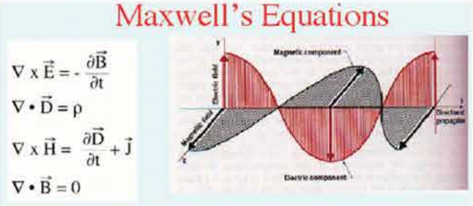

Maxwell's Equations

$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$

$\nabla \cdot \vec{D} = \rho$

$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}$

$\nabla \cdot \vec{B} = 0$

Magnetic component
Electric component

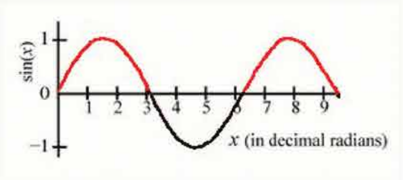

electric fields

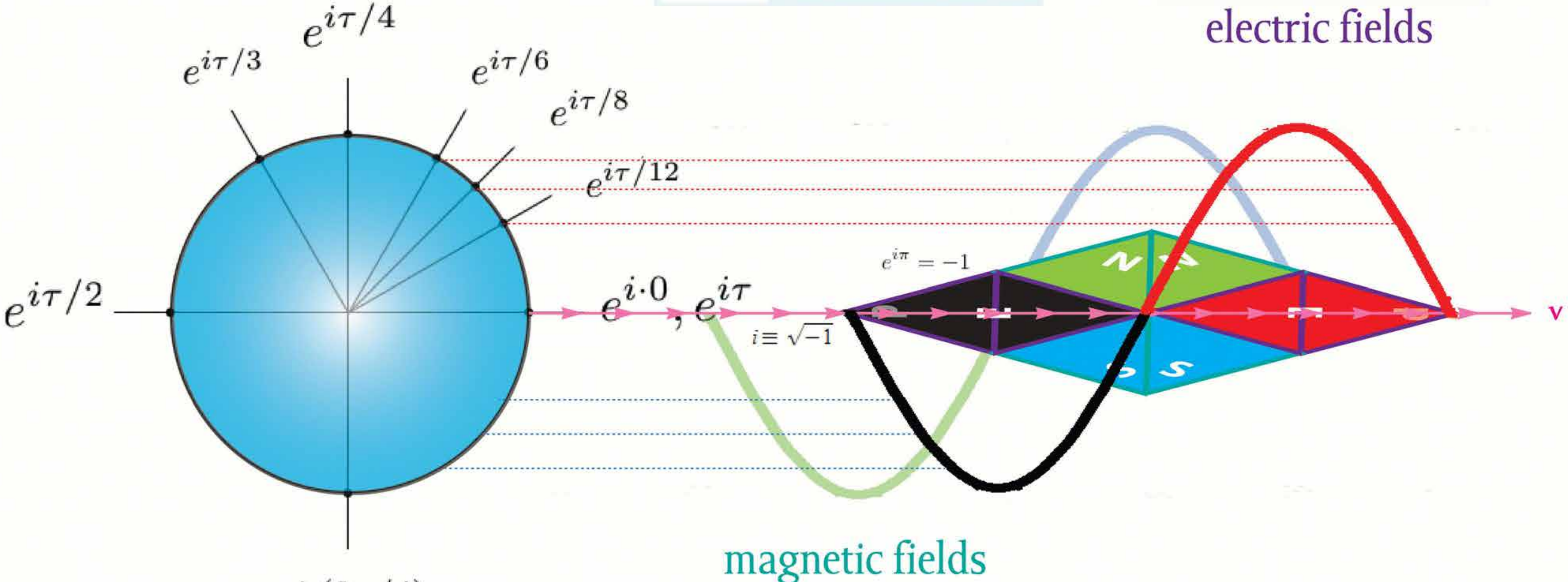
$e^{i\tau/4}$
$e^{i\tau/3}$
$e^{i\tau/6}$
$e^{i\tau/8}$
$e^{i\tau/12}$
$e^{i\tau/2}$
$e^{i\cdot 0}, e^{i\tau}$
$e^{i\pi} = -1$
$i \equiv \sqrt{-1}$
N
N
S
S
v
$e^{i\cdot(3\tau/4)}$

magnetic fields

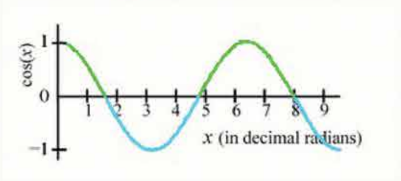

Tetryonic model of a photon

Magnetic waveforms are 90 degrees out of phase with Electric waveforms

Boson distributions in monochromatic EM waves

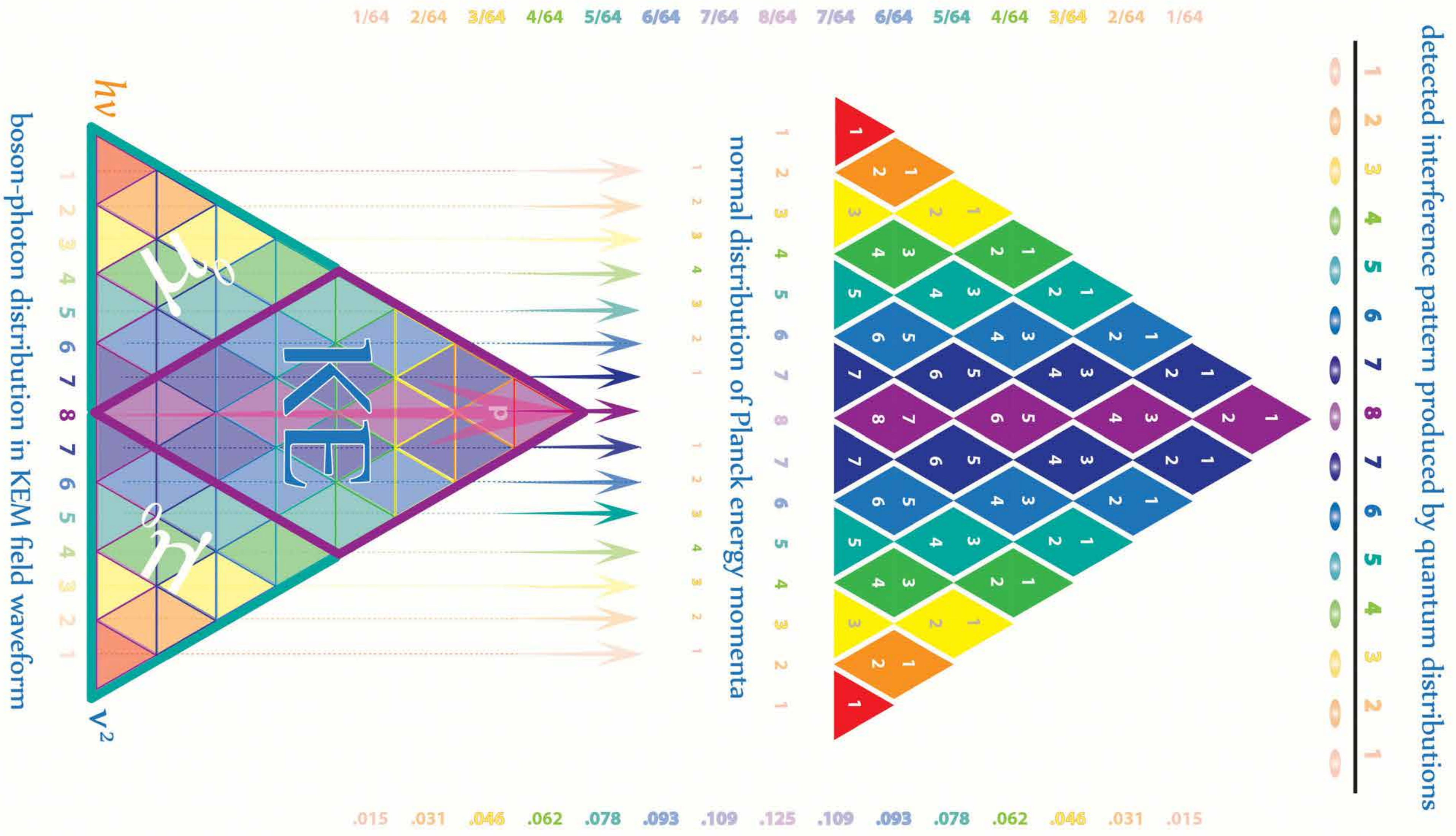
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
hν
v²
μ₀
ε₀
KE
p
boson-photon distribution in KEM field waveform
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
normal distribution of Planck energy momenta
detected interference pattern produced by quantum distributions
.015 .031 .046 .062 .078 .093 .109 .125 .109 .093 .078 .062 .046 .031 .015

Probability distributions of monochromatic EM waves

Tetryonics 98.02 - Boson distributions in monochromatic EM waves

Wave interference patterns

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

The double-slit experiment, sometimes called Young's experiment, is a demonstration that matter and energy can display characteristics of both waves and particles, and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and Establishes the quantum interference principle known as wave–particle duality.

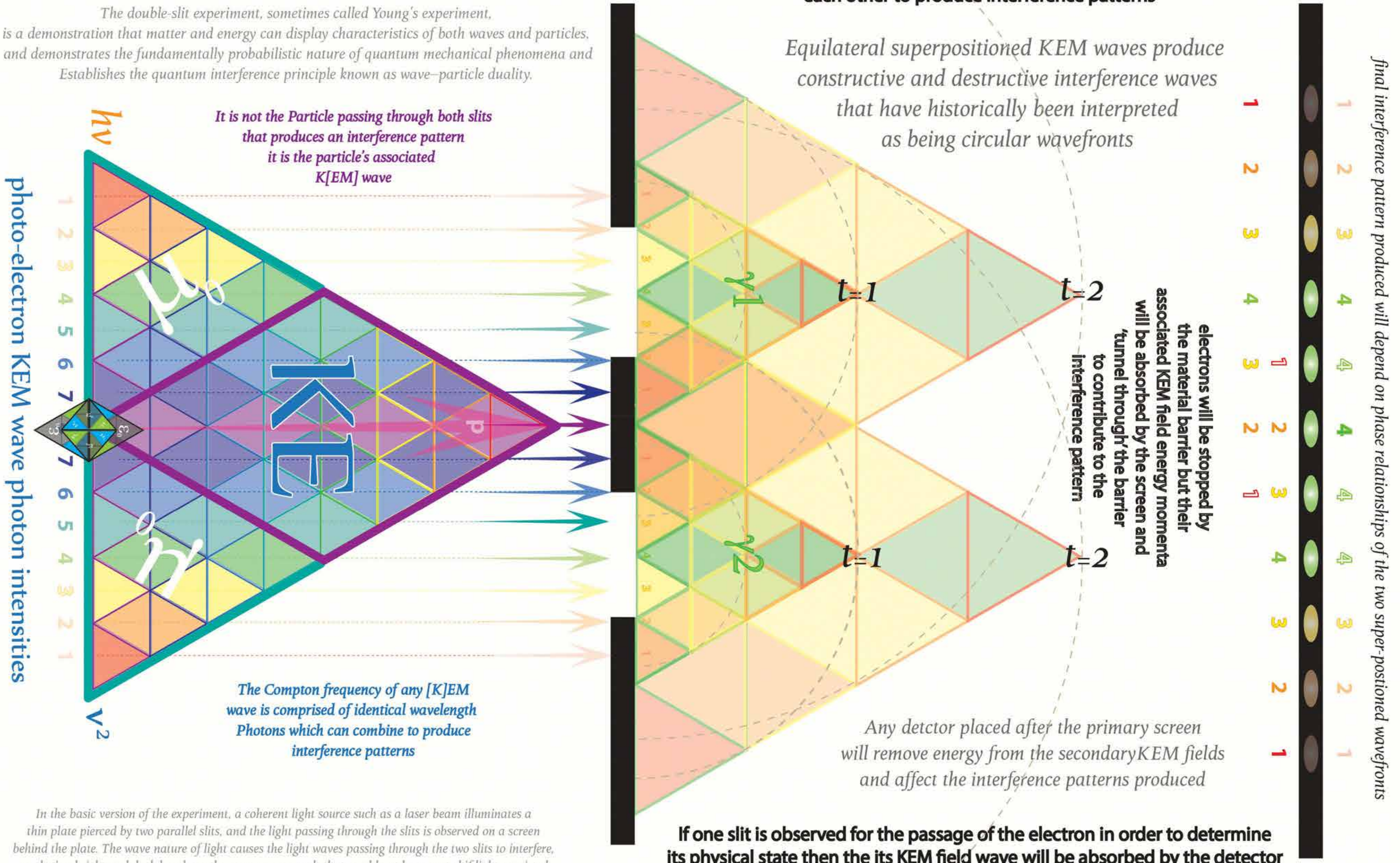

It is not the Particle passing through both slits that produces an interference pattern it is the particle's associated K[EM] wave

Equilateral superpositioned KEM waves produce constructive and destructive interference waves that have historically been interpreted as being circular wavefronts

electrons will be stopped by the material barrier but their associated KEM field energy momenta will be absorbed by the screen and 'tunnel through' the barrier to contribute to the interference pattern

final interference pattern produced will depend on phase relationships of the two super-postioned wavefronts

The Compton frequency of any [K]EM wave is comprised of identical wavelength Photons which can combine to produce interference patterns

Any detctor placed after the primary screen will remove energy from the secondaryKEM fields and affect the interference patterns produced

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

Quantum computing via EM wave super-positioning

BY superpositioning two beams
of EM radiation the resultant 'colours'
will perform quantum level computations
that can be read via the resultant
interference patterns produced

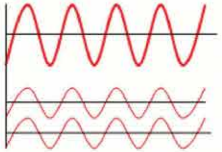

In phase

constructive interference

additive in-phase EM waves

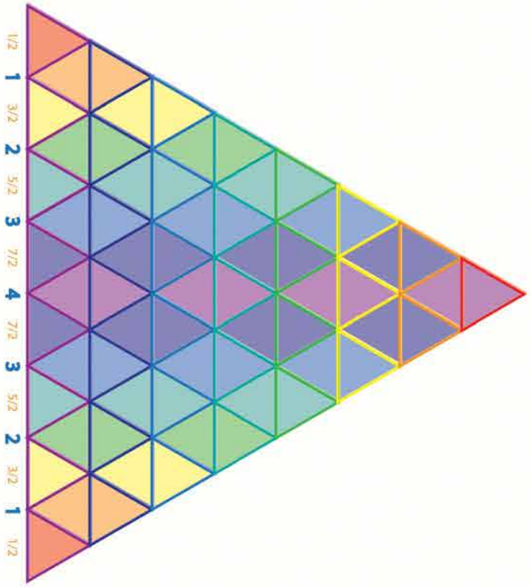
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

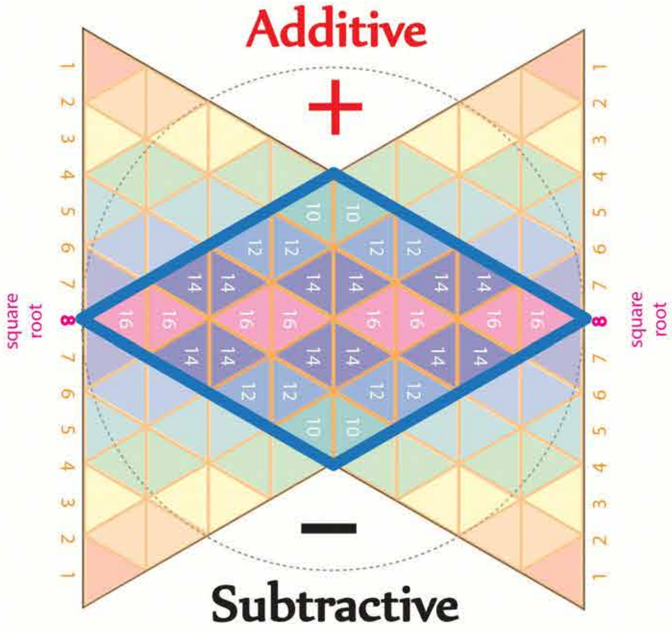
Additive
+
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
square root
10
12
14
16
−
Subtractive

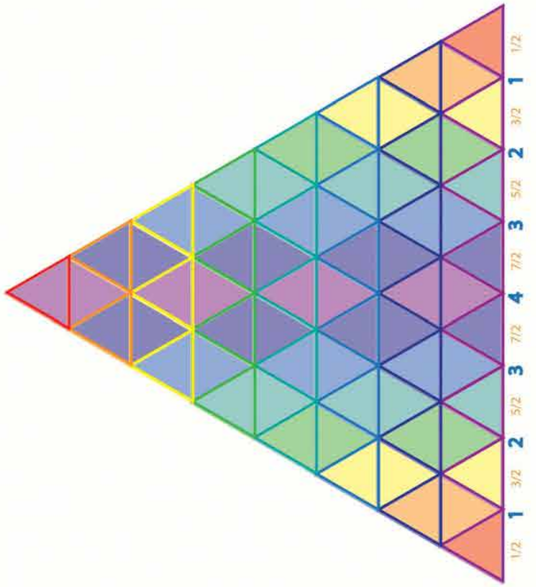
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Out of phase

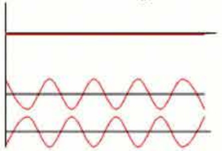

destructive interference

subtractive out-of-phase EM waves

Various basic operations, such as
ADDITION, SUBTRACTION and
SQUARE ROOTS etc
are all easily computed using
EM wave super-positioning

The lines of Force

By utilising the statiscal distribution of equilateral Planck energy momenta quanta in EM waves
Tetryonic theory provides a practical geometric solution to quantum computing problems

q

bosons

P vs. NP

The set of all decision problems for which an algorithm exists which can be carried out by a deterministic Turing machine in polynomial time

the set of all decision problems for which an algorithm exists which can be carried out by a non-deterministic Turing machine in polynomial time

E

photons

An algorithm of time complexity $O(n)$ is one which increases in time linearly as the "size of the problem" (whatever n stands for) increases.

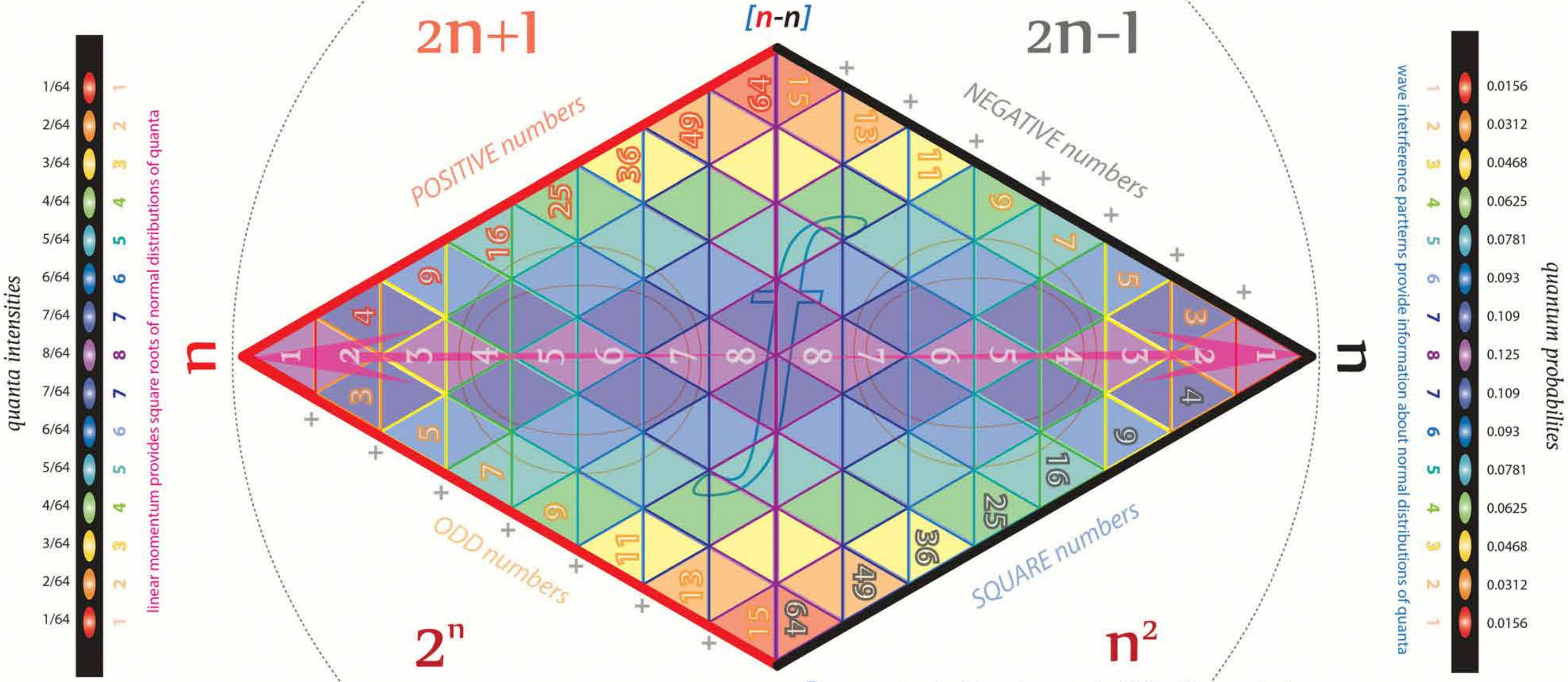
[n-n]
2n+1
2n-1
POSITIVE numbers
NEGATIVE numbers
ODD numbers
SQUARE numbers
n
n
2ⁿ
n²
64 49 36 25 16 9 4
15 13 11 9 7 5 3
3 5 7 9 11 13 15
4 9 16 25 36 49 64
1 2 3 4 5 6 7 8 8 7 6 5 4 3 2 1
quanta intensities
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
linear momentum provides square roots of normal distributions of quanta
wave interference parterns provide information about normal distributions of quanta
quantum probabilities
0.0156 0.0312 0.0468 0.0625 0.0781 0.093 0.109 0.125 0.109 0.093 0.0781 0.0625 0.0468 0.0312 0.0156

An algorithm of complexity $O(2^n)$ utlises exponential quanta; increasing n by 1 will double the quanta required

$2n^2$

An algorithm of complexity $O(n2)$ utilises quadratic quanta, meaning that if you double n it will use four times as many quanta.

n^o

Unlike Math treatise on P vs Np that require exponential polynomial time $O(n^k)$ Tetryonic geometry of EM fields utilise exponential energies per second

2hv

1 second

hf

By relating p to the number of equilateral wave fields in any EM wave very large [pn²] data sets can be modelled and processed every second

Quantum Cryptography

If the numerical sequences were applied to amplitude modulation
their non-repeating numerical sequences would appear
to be purely random noise

**Using Tetryonic geometry
advanced non-repeating cyphers
of any complexity
can be easily developed**

The level of encryption can
be easily increased (without limit)
by increasing the dimensions
of the cypher geometry

eg: number of quanta
2D vs 3D geometries

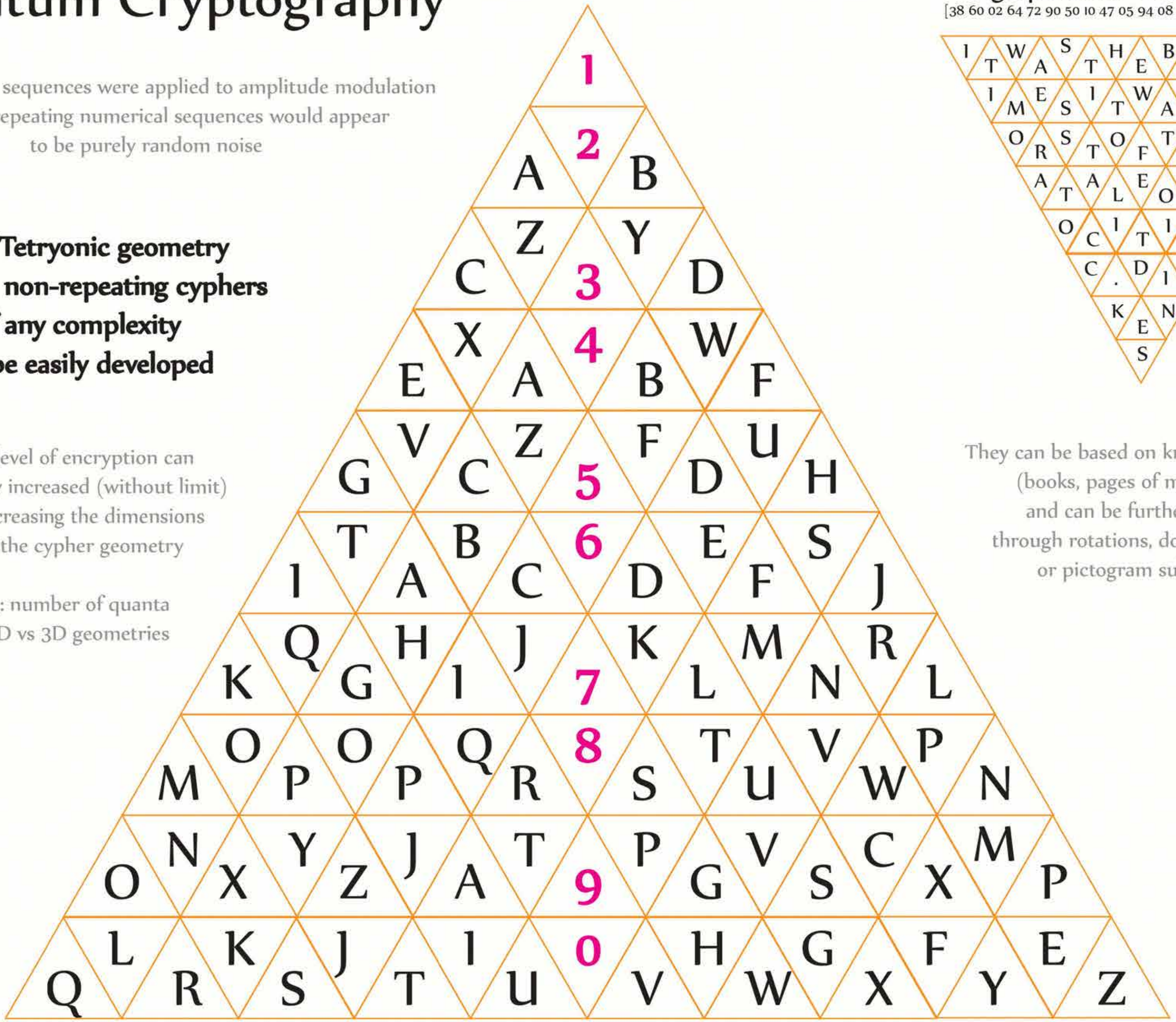
1
A 2 B
C Z 3 Y D
E X A 4 B W F
G V C Z 5 F D U H
I T A B C 6 D E F S J
K Q G H I J 7 K L M N R L
M O P O P Q R 8 S T U V W P N
O N X Y Z J A T 9 P G V S C X M P
Q L R K S J T I U 0 V H W G X F Y E Z

eg: quantum encryption 1024
[38 60 02 64 72 90 50 10 47 05 94 08 48 52 27 26 51 66 01 91 03 20]

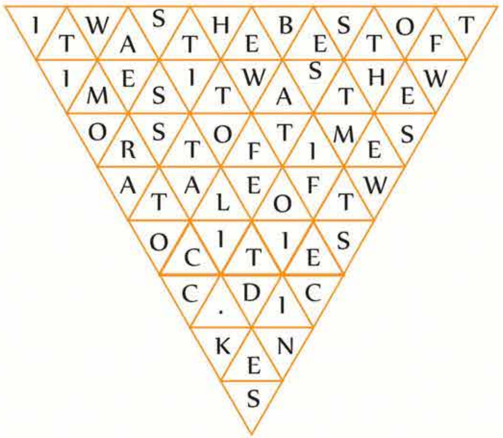
I T W A S T H E B E S T O F T
I M E S I T W A S T H E W
O R S T O F T I M E S
A T A L E O F T W
O C I T I E S
C . D I C
K E N
S

They can be based on known letter sources
(books, pages of magazines etc)
and can be further encrypted
through rotations, double encrypting
or pictogram substitutions

Copyright ABRAHAM [2008] - All rights reserved

Quantum Computing

The Proton/Neutron geometries of atomic nuclei can be built at the quantum scale
to create an atomic nuclei that can operate as a Opto-memory-transistive computing element,
many elements can then be combined in lattices to create super computers no larger than bacterium

Spin UP

Energy can be gated through individual nuclei
using the centre Baryon as the base transistor element,
in turn effecting the energies of bound photo-electrons

Spin DOWN

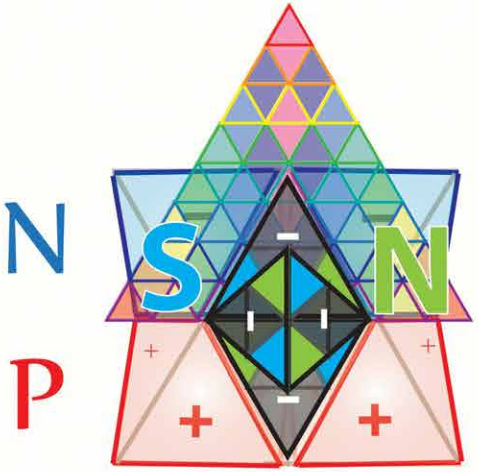

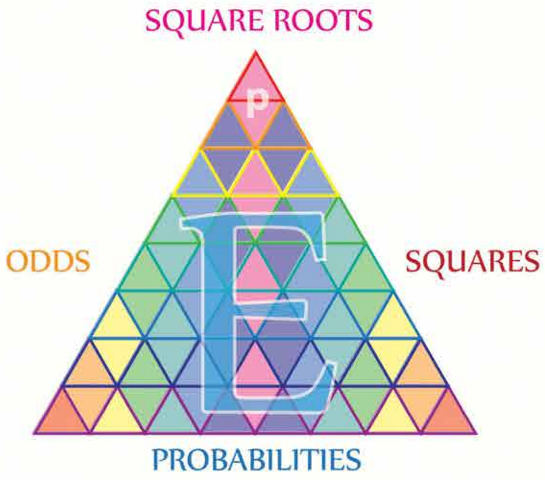

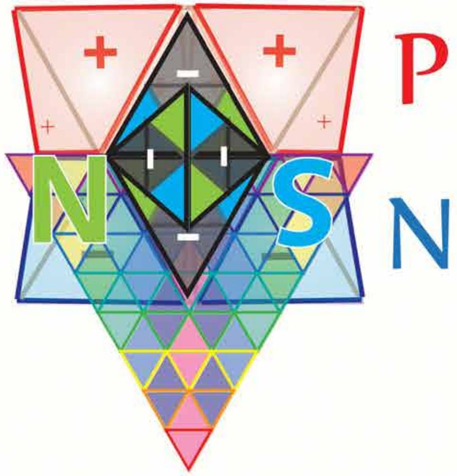

Q-bits

1

photo-electronic transitions can be used to directly recieve
or emit memory states through the absorption and emission
of spectral photons of specific energy momenta

0

Scalar mass-energies

Quantum distributions

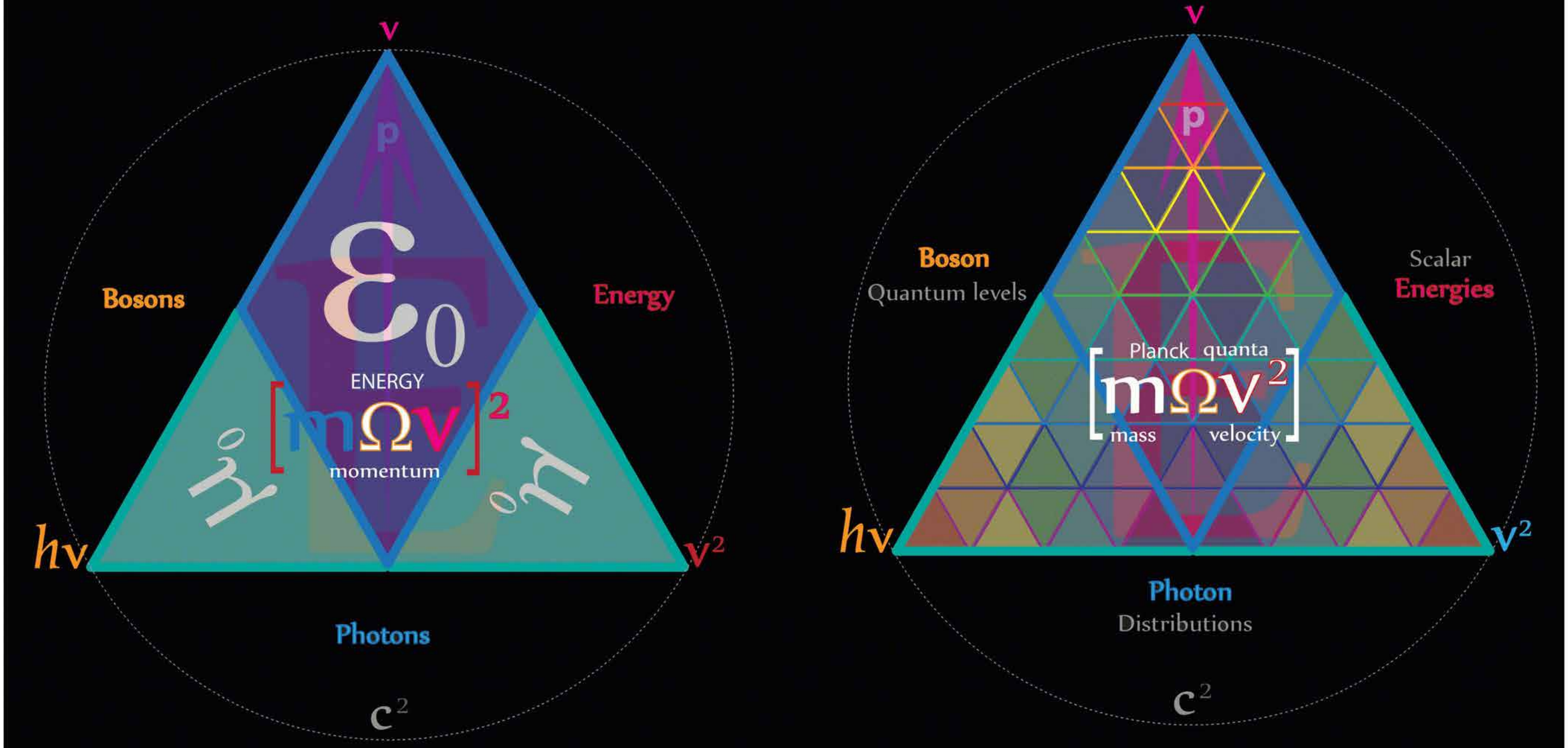
v
p
$\mathcal{E}_0$
ENERGY
$[m\Omega v]^2$
momentum
Bosons
Energy
Photons
hv
v²
c²
v
p
Boson
Quantum levels
Scalar
Energies
Planck quanta
$[m\Omega v^2]$
mass velocity
Photon
Distributions
hv
v²
c²

EM mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Copyright ABRAHAM [2008] - All rights reserved

ENERGY

2D mass

EM mass

Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

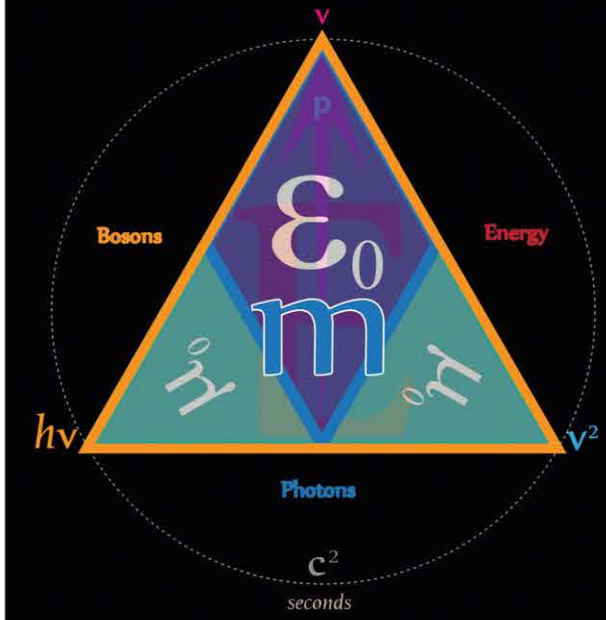
v
p
ε_0
m
Bosons
Energy
hv
v^2
Photons
c^2
seconds

Measure of planar energies
in a nπ radiant wave
geometry
[per second]

mass quantum

7.376238634 e-51 kg

ENERGY

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

momenta

Equilateral energy momenta
has a mass equivalence
per unit of Time

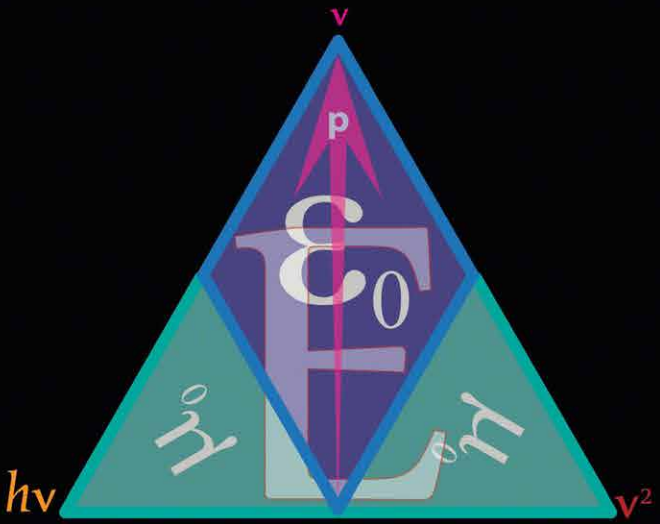
v
p
ε_0
E
hv
v^2

c^2 is the radial geometry
created by radiant energy
in 1 second

Planck Constant

6.629432672 e-34 J.s

3D Matter

Matter

Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Measure of mass-energy
in a 4nπ standing wave
topology
[per second squared]

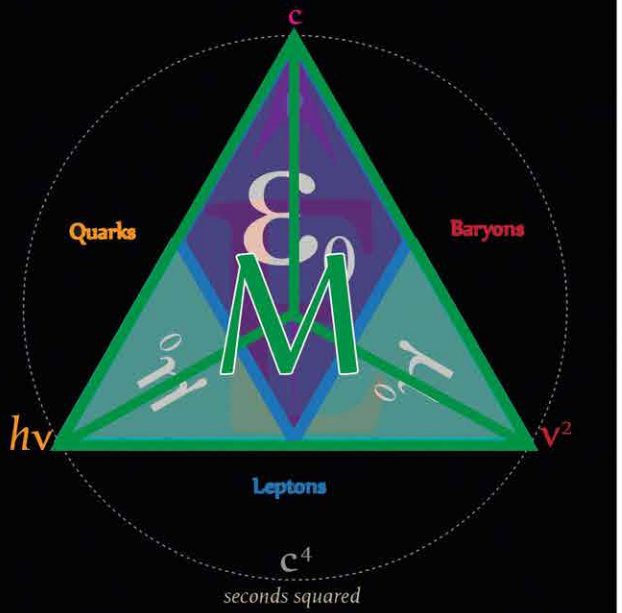
c
ε_0
M
Quarks
Baryons
hv
v^2
Leptons
c^4
seconds squared

Matter quantum

2.9504955454 e-50 kg

Copyright ABRAHAM [2008] - All rights reserved

Planck's constant - the quantum of Action

all ideal quantum inductive loops
resist changes to their
energy levels

INERTIAL mass

all equilateral energy momenta
produce square root vector
Forces

quantum of WORK

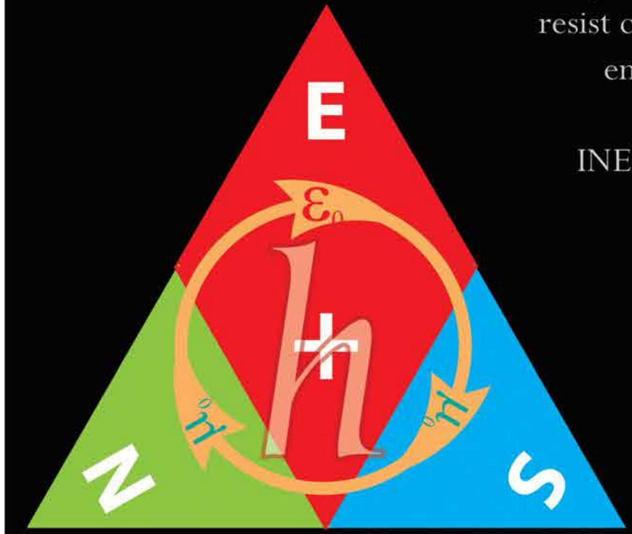
E
ε_0
$\hbar$
μ_0
μ_0
N
S

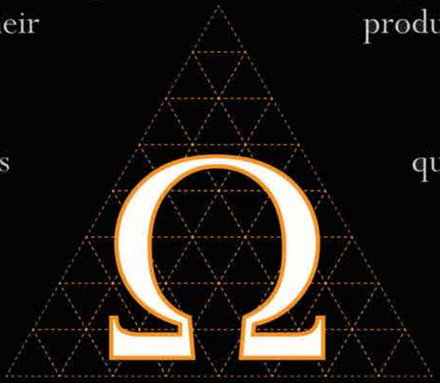

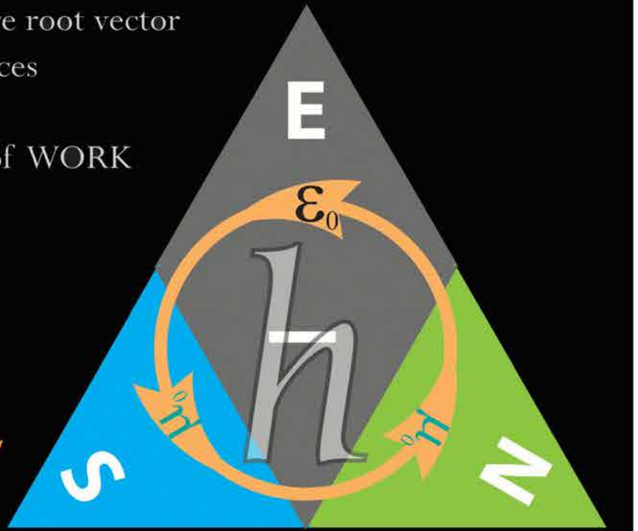
E
ε_0
$\hbar$
μ_0
μ_0
S
N

Quantised angular momentum is equilateral geometry

The two sides of Planck's Constant

The square root of all Energy is linear momentum

Equilateral
energy quanta
per second

mass

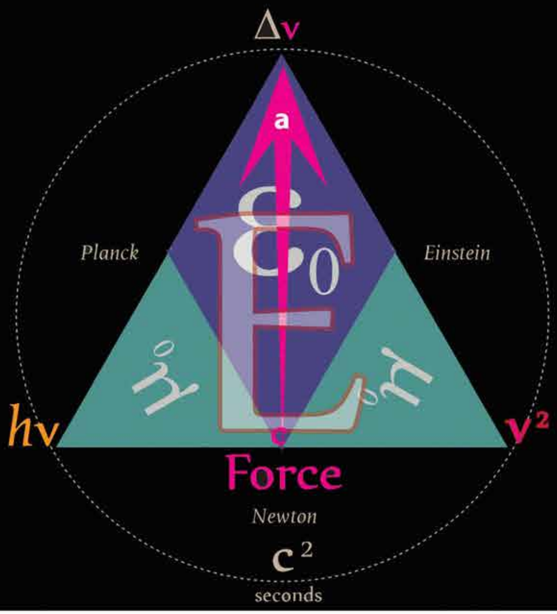
Δv
a
ε_0
E
Planck
Einstein
μ_0
μ_0
hv
v²
Force
Newton
c²
seconds

Quantised
angular momenta
per second

time

Copyright ABRAHAM [2008] - All rights reserved

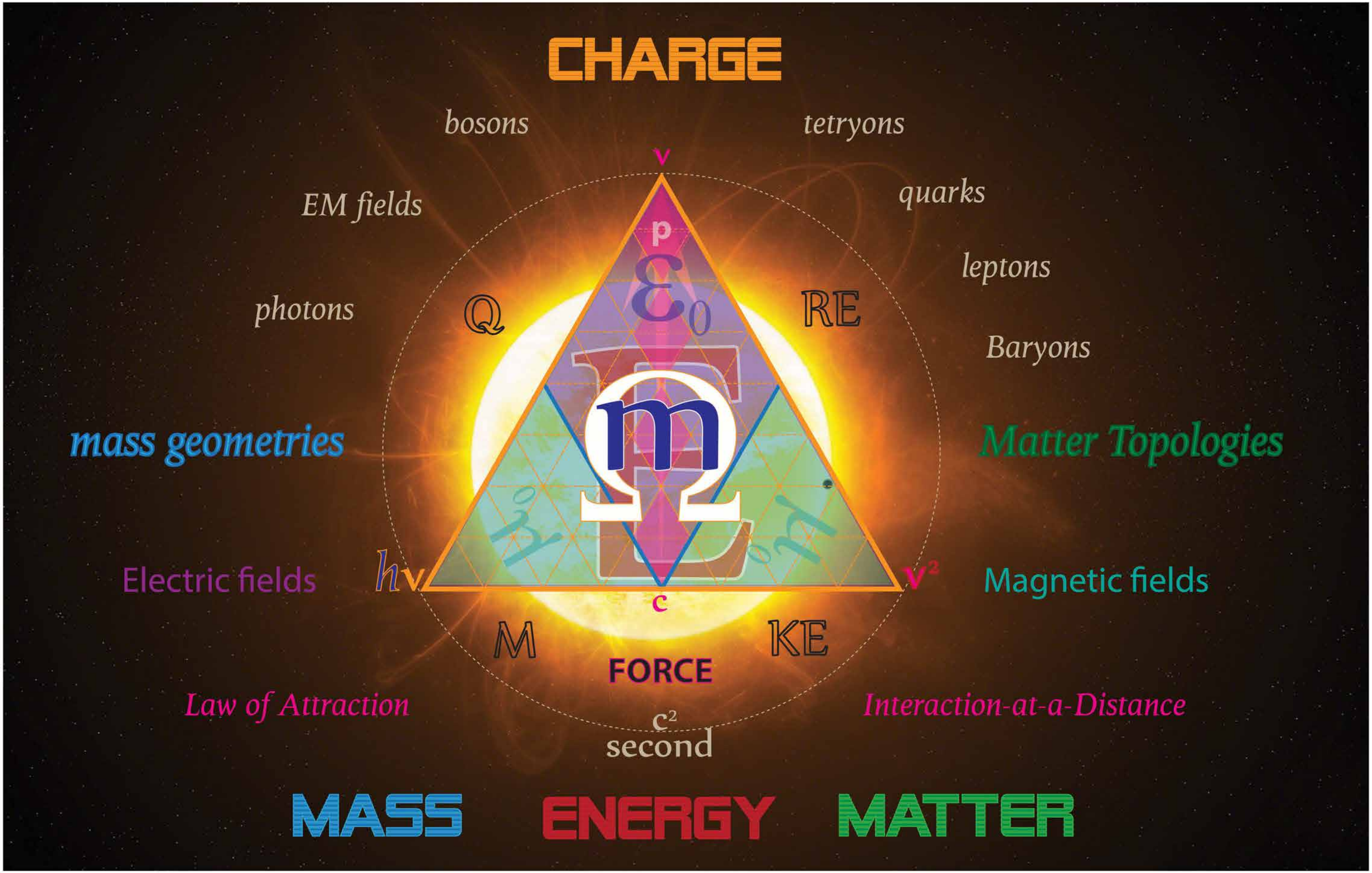
CHARGE
bosons
tetryons
v
EM fields
quarks
p
leptons
ε0
photons
Q
RE
Baryons
E
m
Ω
mass geometries
Matter Topologies
μ0
Electric fields
hv
v²
Magnetic fields
c
M
KE
FORCE
Law of Attraction
Interaction-at-a-Distance
c²
second
MASS
ENERGY
MATTER

Tetryonics 99.04 - Charged Planck mass-ENERGY-Matter

All squared energies have equilateral geometries

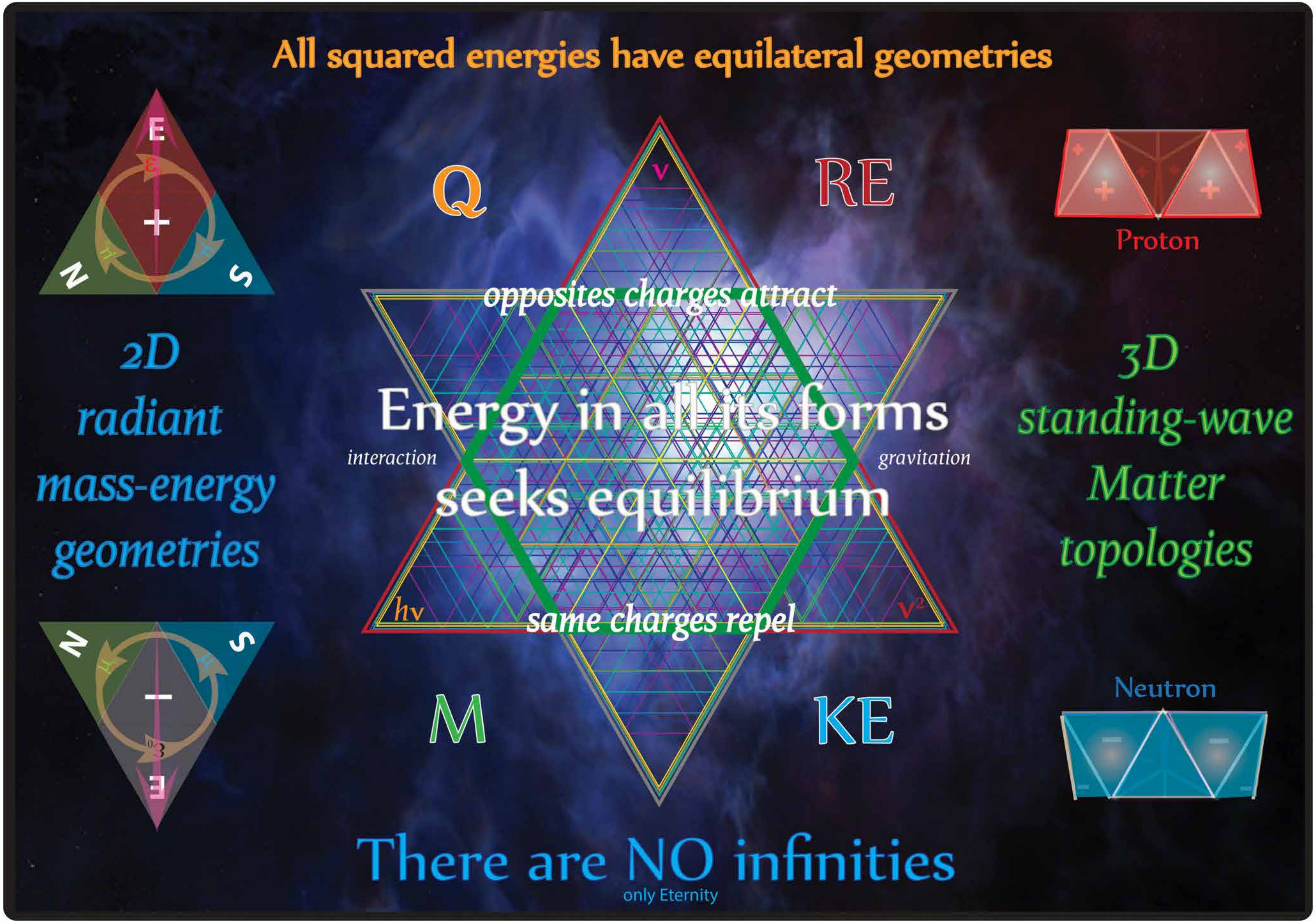
E
+
N
S
Q
RE
ν
Proton
interaction
gravitation
$h\nu$
ν^2
N
S
−
E
M
KE
Neutron

2D
radiant
mass-energy
geometries

opposites charges attract

Energy in all its forms
seeks equilibrium

same charges repel

3D
standing-wave
Matter
topologies

There are NO infinities
only Eternity

Copyright ABRAHAM [2008] - All rights reserved

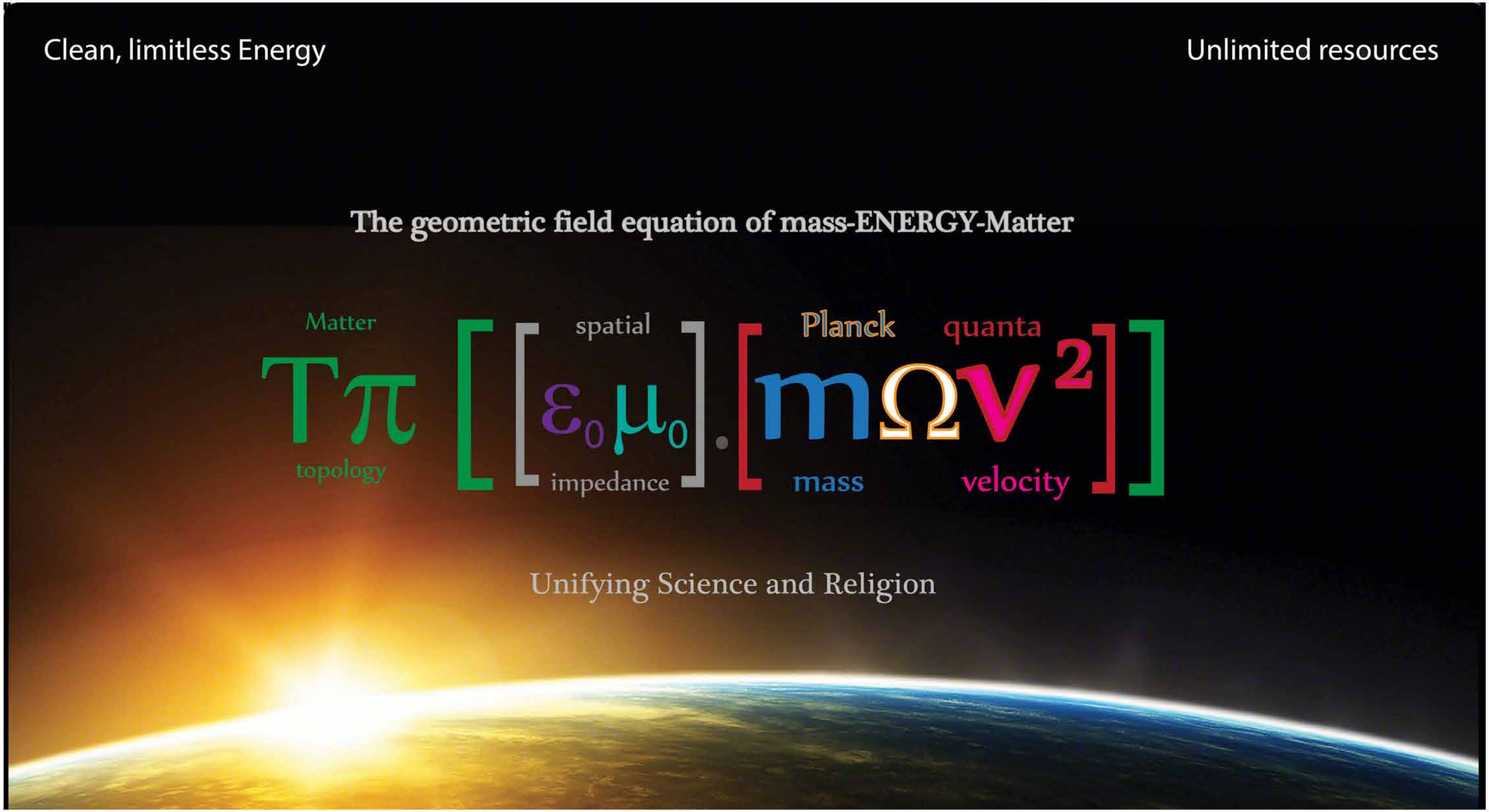

Clean, limitless Energy

Unlimited resources

The geometric field equation of mass-ENERGY-Matter

$$T\pi\left[\left[\varepsilon_0\mu_0\right]\cdot\left[m\Omega v^2\right]\right]$$

Matter
topology
spatial
impedance
Planck
quanta
mass
velocity

Unifying Science and Religion

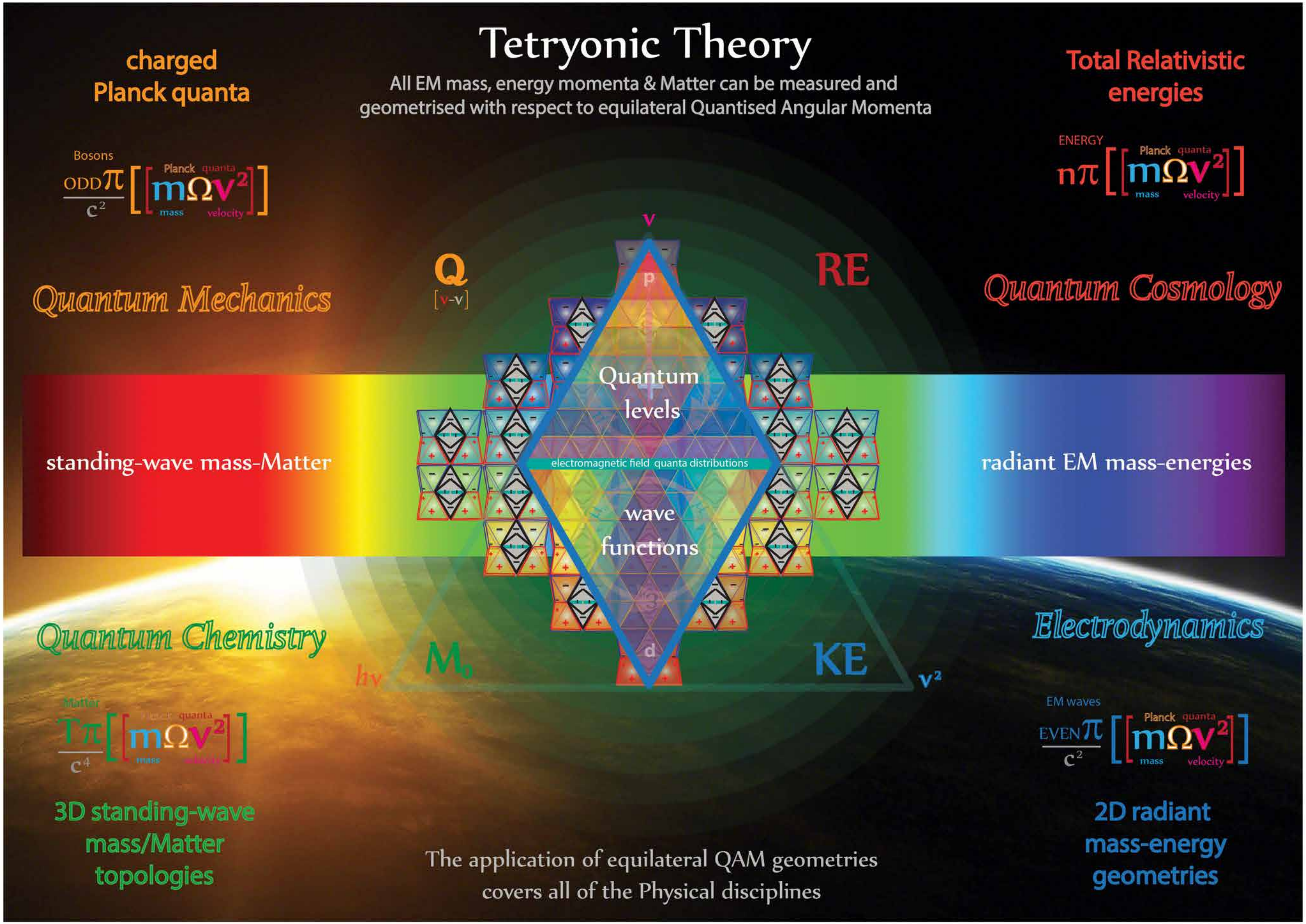

Tetryonic Theory

All EM mass, energy momenta & Matter can be measured and geometrised with respect to equilateral Quantised Angular Momenta

charged Planck quanta

Bosons

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

Total Relativistic energies

ENERGY

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

Quantum Mechanics

Quantum Cosmology

standing-wave mass-Matter

radiant EM mass-energies

Quantum Chemistry

Matter

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

Electrodynamics

EM waves

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

Planck quanta
mass velocity

3D standing-wave mass/Matter topologies

2D radiant mass-energy geometries

The application of equilateral QAM geometries
covers all of the Physical disciplines

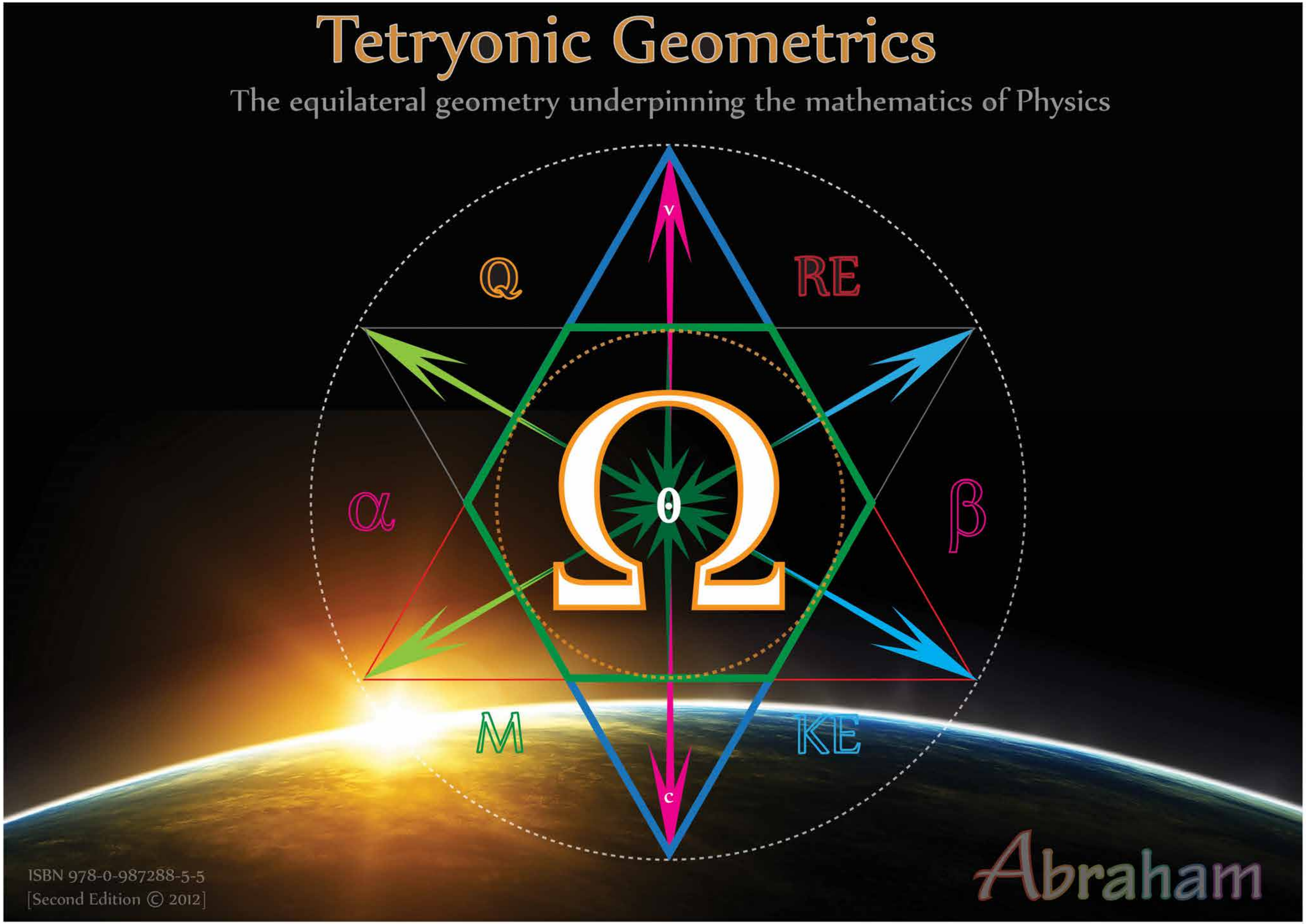

Tetryonic Geometrics

The equilateral geometry underpinning the mathematics of Physics

ISBN 978-0-987288-5-5
[Second Edition © 2012]

Abraham

TETRYONICS

The equilateral geometry underpinning the mathematics of Physics

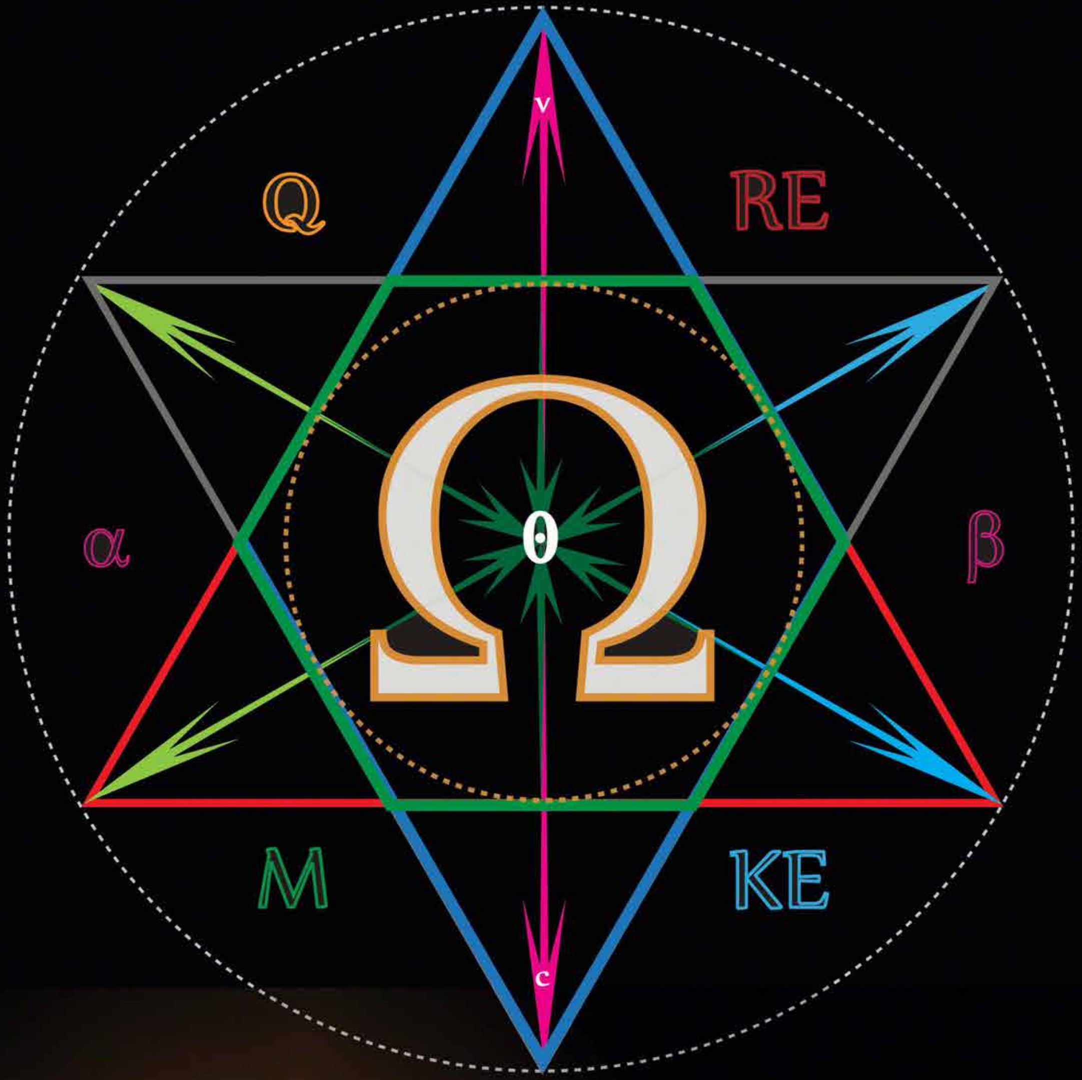

Foundational physics Mathematics

Abraham

ISBN 978-0-987288-5-5
[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Geometry and the Theory of Everything

Plato

(c.428-348 BC)

DIFFERENT TYPE OF GEOMETRIES

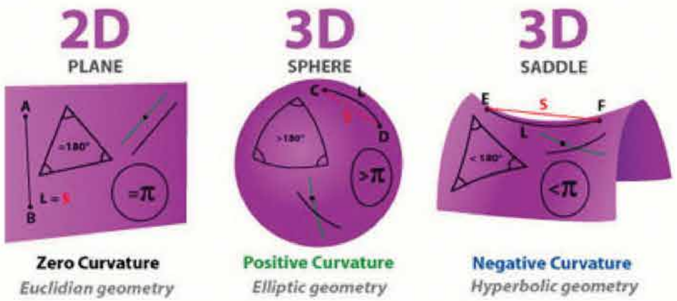
2D
PLANE
A
B
L = S
=180°
=π
3D
SPHERE
C
L
S
D
>180°
>π
3D
SADDLE
E
S
F
L
<180°
<π
Zero Curvature
Euclidian geometry
Positive Curvature
Elliptic geometry
Negative Curvature
Hyperbolic geometry

(studied by Omar Khayyam, Girolamo Saccheri, Bernhard Riemann, ...)

Euclid

(c.330-275 BC, fl. c.300 BC)

The Socratic tradition was not particularly congenial to mathematics, as may be gathered from Socrates' inability to convince himself that 1 plus 1 equals 2, but it seems that his student Plato gained an appreciation for mathematics after a series of conversations with his friend Archytas in 388 BC.

One of the things that most caught Plato's imagination was the existence and uniqueness of what are now called the five "Platonic solids".

It's uncertain who first described all five of these shapes - it may have been the early Pythagoreans - but some sources (including Euclid) indicate that Theaetetus (another friend of Plato's) wrote the first complete account of the five regular solids.

Presumably this formed the basis of the constructions of the Platonic solids that constitute the concluding Book XIII of Euclid's Elements.

In any case, Plato was mightily impressed by these five definite shapes that constitute the only perfectly symmetrical arrangements of a set of (non-planar) points in space, and late in life he expounded a complete "theory of everything", in the treatise called Timaeus, based explicitly on these five solids.

Interestingly, almost 2000 years later, Johannes Kepler was similarly fascinated by these five shapes, and developed his own cosmology from them

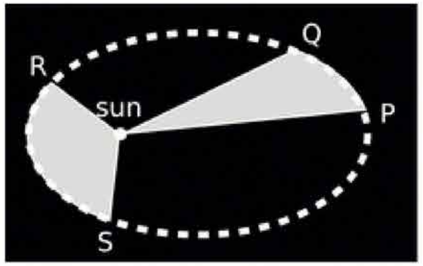

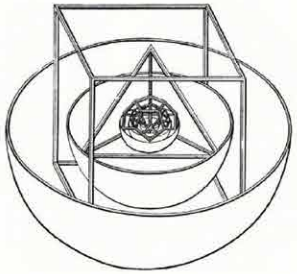

Copyright ABRAHAM [2008] - All rights reserved

Tetractys

The Tetractys historically symbolized the four elements
[Earth, Air, Fire, and Water]
and the relationship between Humanity
and the cosmos created by GOD

The Greek Tetractys is a
triangular figure consisting of ten points
arranged in four rows:

one, two, three, and four points in each row,
which is the geometrical representation of the
fourth triangular number.

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

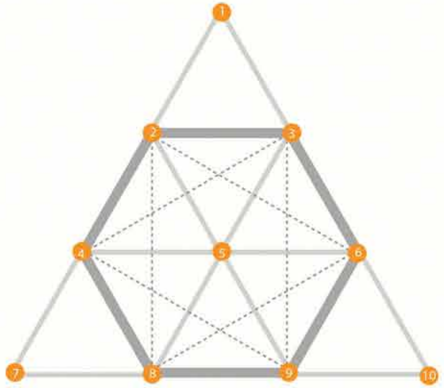

Sacred numbers

A

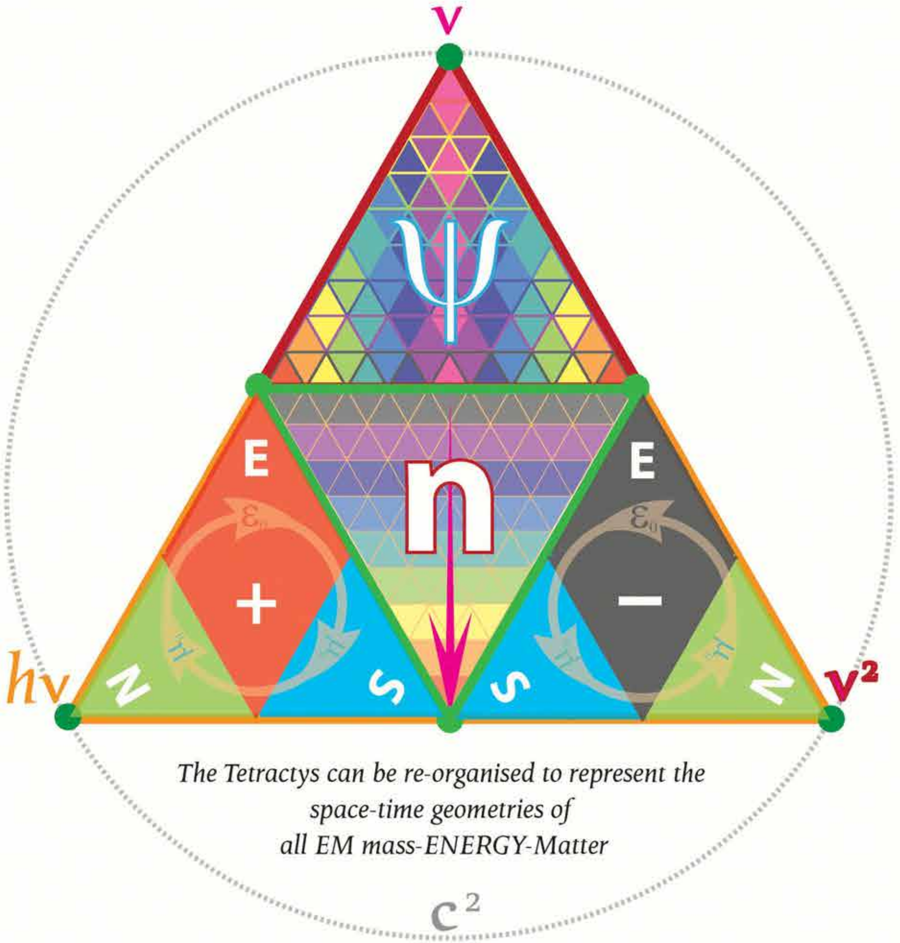

The Tetractys can be re-organised to represent the space-time geometries of all EM mass-ENERGY-Matter

The tetrad was the name given to
the number four -
in Pythagorean philosophy
there were four seasons
and four elements,
and the number was also associated
with planetary motions and music

As a mystical symbol, it was very important to
the secret worship of the Pythagoreans.

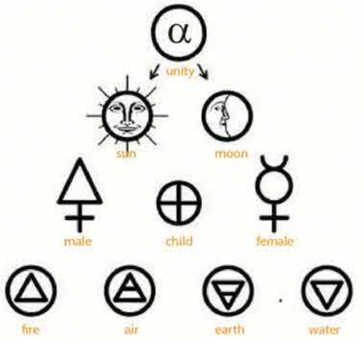

The Cosmos

Ω

The single triangle in the first row represents zero-dimensions (a point)
A vector direction in one-dimension can be represented as a line between any two points
The second row represents a Boson (two-dimensions in a plane defined by a rhombus of three triangles)
The whole figure folded represents three-dimensions (a tetrahedron defined by four apex points)
Photons of ElectroMagnetic mass-Energy quanta are represented by two opposing triangles

The Greek Zodiac

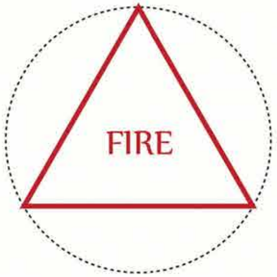

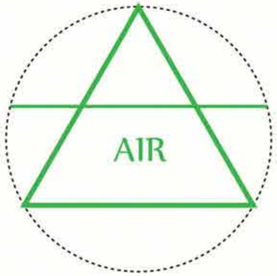

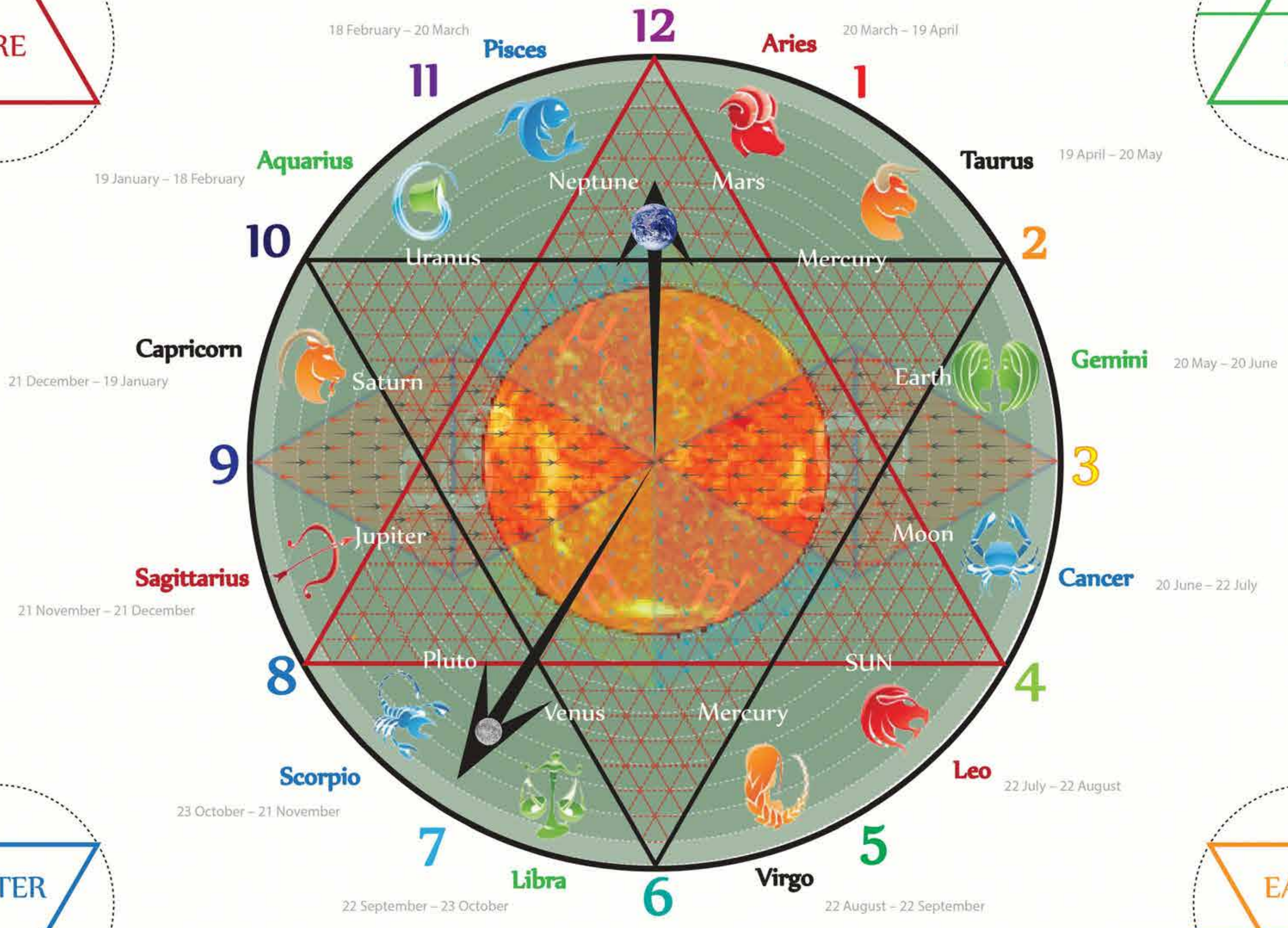
12
Aries
1
20 March – 19 April
Taurus
19 April – 20 May
2
Gemini
20 May – 20 June
3
Cancer
20 June – 22 July
4
Leo
22 July – 22 August
5
Virgo
22 August – 22 September
6
Libra
22 September – 23 October
7
Scorpio
23 October – 21 November
8
Sagittarius
21 November – 21 December
9
Capricorn
21 December – 19 January
10
Aquarius
19 January – 18 February
11
Pisces
18 February – 20 March
Neptune
Mars
Uranus
Mercury
Saturn
Earth
Jupiter
Moon
Pluto
SUN
Venus
Mercury

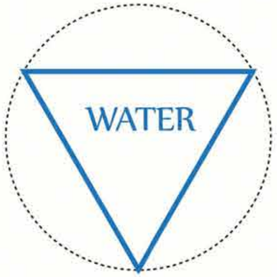

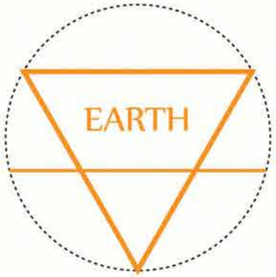

The Greek Elements

Copyright ABRAHAM [2008] - All rights reserved

Equilateral Triangles

An equilateral triangle is a triangle in which all three sides are equal

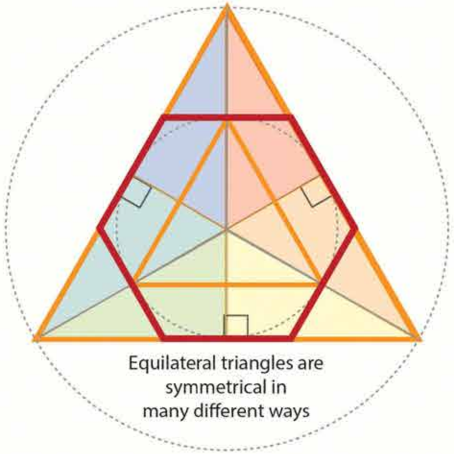

Equilateral triangles are
symmetrical in
many different ways

It is unique in that it is the only polygon
that can be tiled [or divided] and produce
only identical geometries and squares numbers

The equilateral triangle
is eminently suited for
the construction of fractals

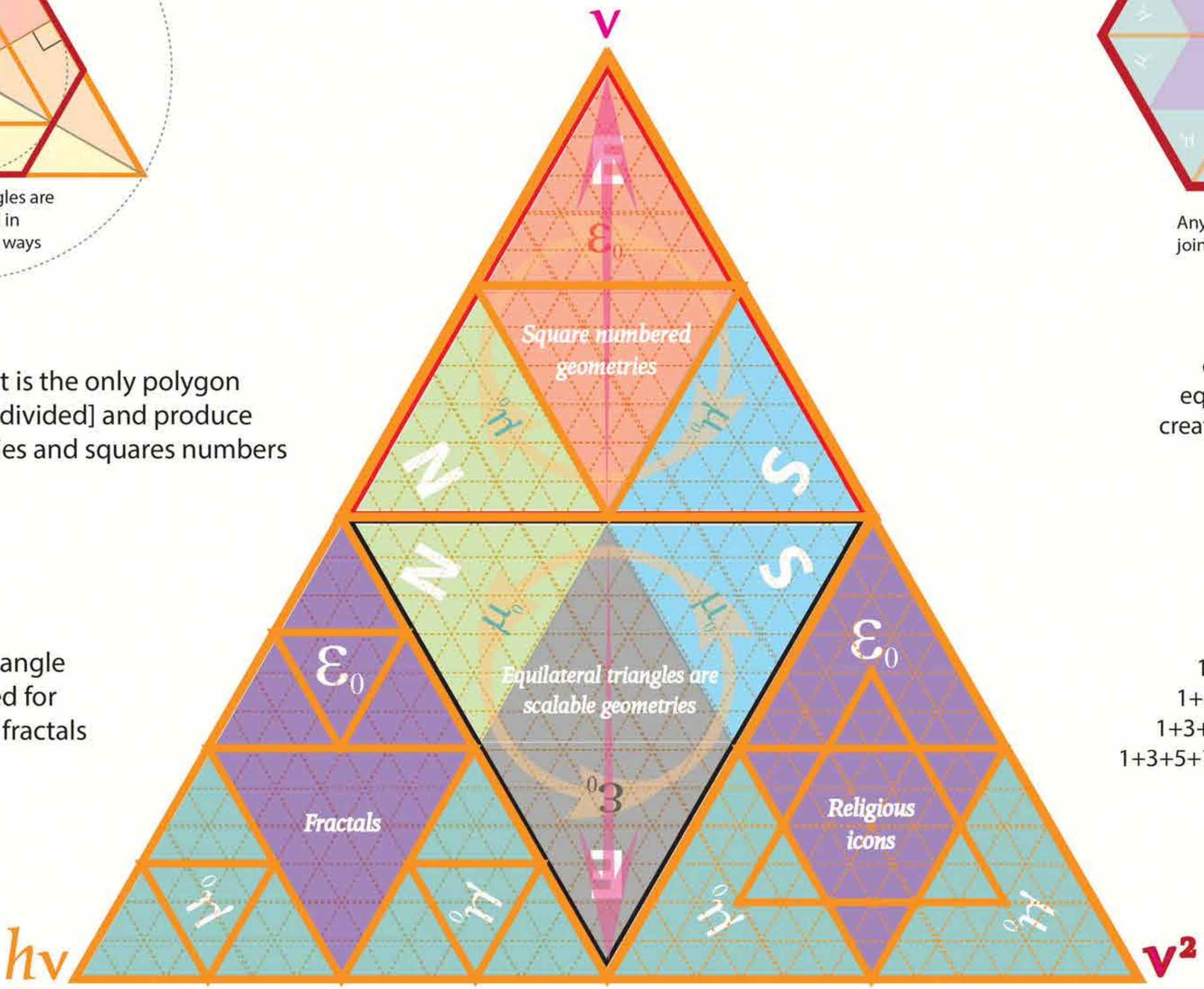
v
ε_0
Square numbered geometries
μ_0
N
S
μ_0
ε_0
Equilateral triangles are scalable geometries
ε_0
Fractals
Religious icons
ε_0
μ_0
hv
v²

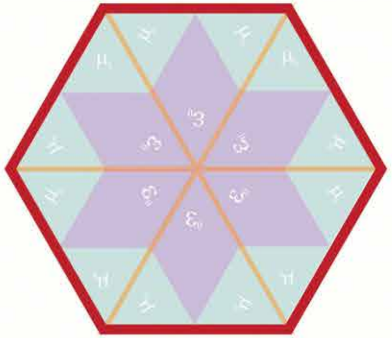

Any six equilateral triangles
joined can make a hexagon.

The
tesselation of
odd numbered
equilateral triangles
creates square numbers

That is,

1
1+3 = 4
1+3+5 = 9
1+3+5+7 = 16
1+3+5+7+9 = 25
1+3+5+7+9+11 = 36
1+3+5+7+9+11+13 = 49
1+3+5+7+7+9+11+13+15 = 64
etc

An equilateral triangle is simply a specific case of a regular polygon with 3 sides

The Pythagorean Theorem

Though attributed to Pythagoras, it is not certain that he was the first person to prove it.

The first clear proof came from Euclid, and it is possible the concept was known
1000 years before Pythoragas by the Babylonians

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

*In any right triangle,
the area of the square whose side is the hypotenuse
(the side opposite the right angle) is equal to
the sum of the areas of the squares
whose sides are the two legs
(the two sides that meet at a right angle)*

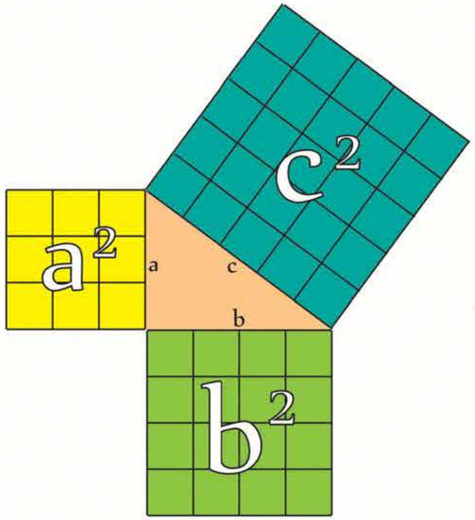

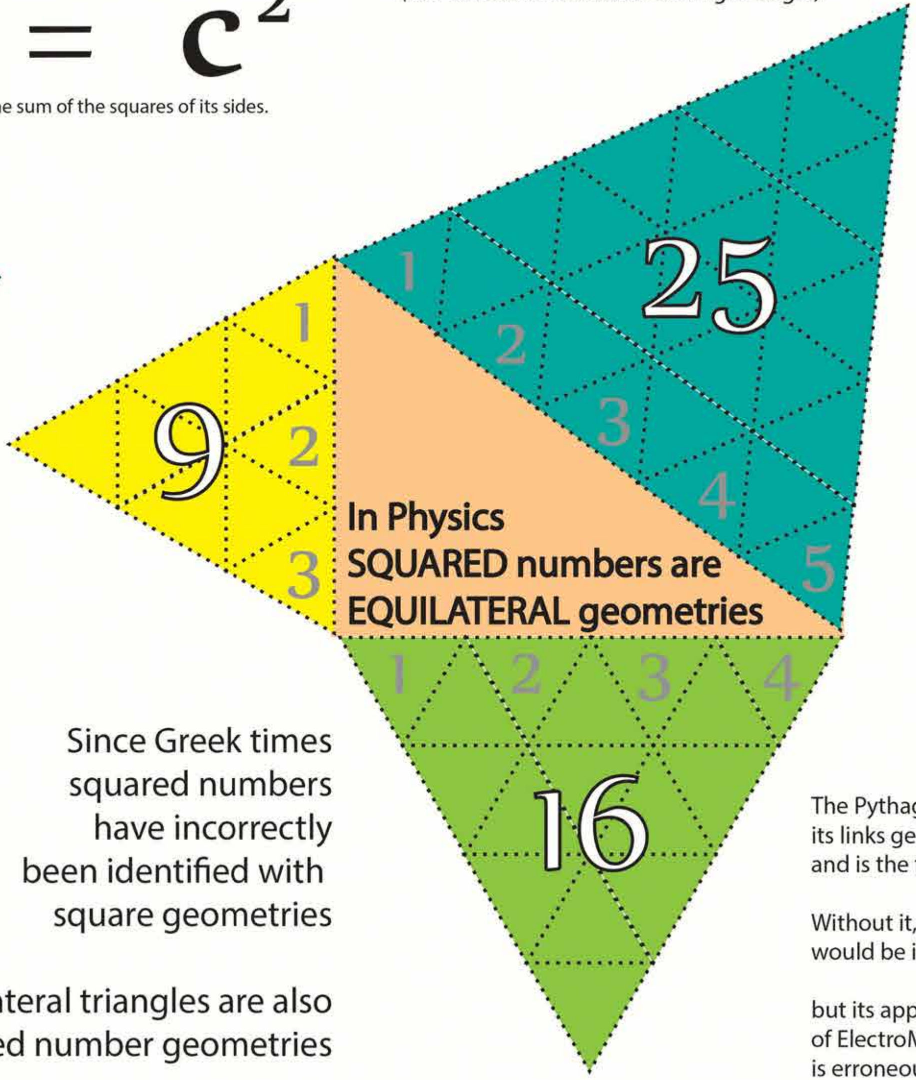

**In Physics
SQUARED numbers are
EQUILATERAL geometries**

Pythagoras of Samos

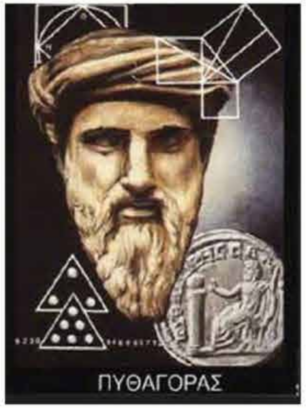

about (570 – 495 BC)

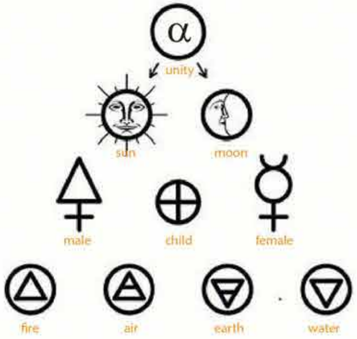

Pythagorean Tetractys

Since Greek times
squared numbers
have incorrectly
been identified with
square geometries

Equilateral triangles are also
squared number geometries

The Pythagorean equation is at the core of much of geometry,
its links geometry with algebra,
and is the foundation of trigonometry.

Without it, accurate surveying, mapmaking, and navigation
would be impossible,

but its application to the energy-momenta geometries
of ElectroMagnetic fields and Matter in motion in Physics
is erroneous and must be corrected for science to advance

Hexagons

A regular hexagon can be subdivided into six equilateral triangles

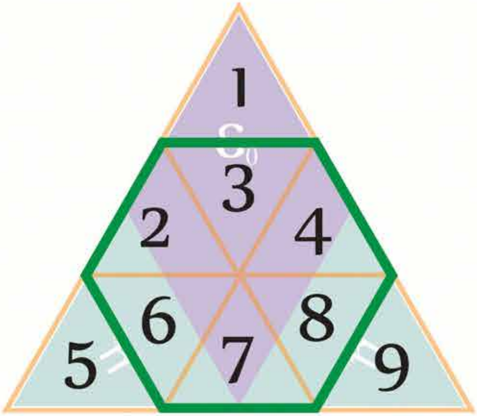

Energy geometries

Atomic nuclei geometries

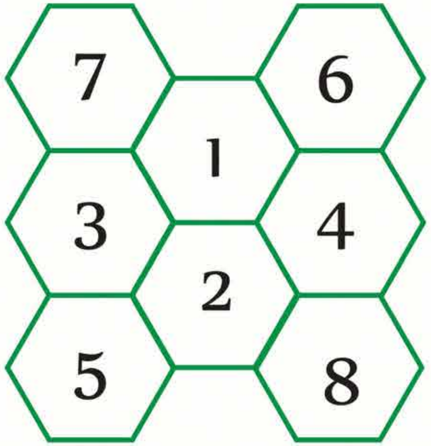

Hexagons can be tiled or tessellated in a regular pattern on a flat two-dimensional plane

Hexagons are the only regular polygon that can be subdivided into another regular polygon

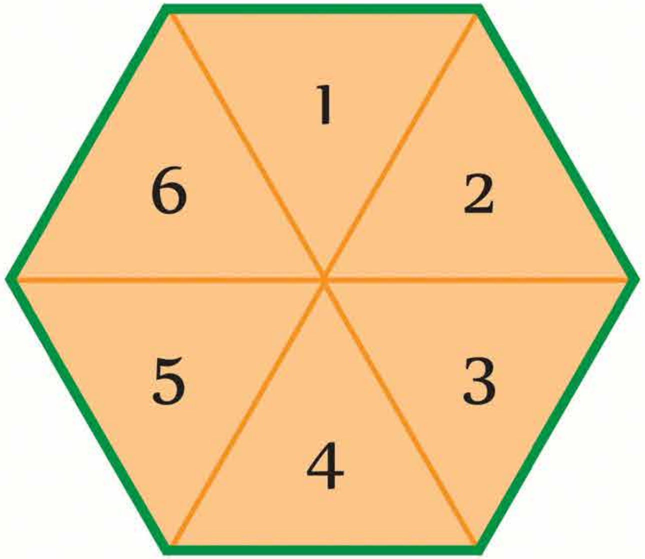

Hexagons are the unique regular polygon such that the distance between the center and each vertex is equal to the length of each side

Six is a highly composite number,
the second-smallest composite number,
and the first perfect number.

That is, 1*2*3=1+2+3=6

An interesting relationship between circular and hexagonal geometry is that hexagonal patterns often appear spontaneously when natural forces are trying to approximate circles

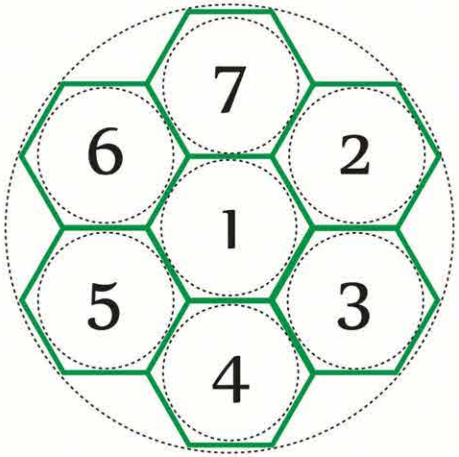

Hexagonal tessellation is topologically identical to the close packing of circles on a plane

Copyright ABRAHAM [2008] - All rights reserved

Platonic Solids

The Pythagoreans knew that there were only five regular convex solids,
the tetrahedon, cube, octahedron, icosohedron and dodecahedron
and each one could be accurately circumscribed by a sphere.

FIRE

4 triangles meet to form
a Tetrahedron

EARTH

6 squares meet to form
a Cube

AIR

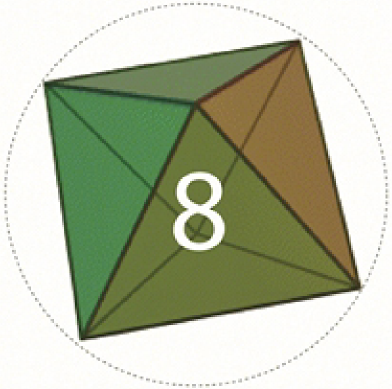

8 triangles meet to form
a Octahedron

tetryons **4**

regular deltahedrons

leptons **12**

$$F - E + V = 2$$

faces edges vertexs

8 *quarks*

regular deltahedrons

20 *Baryons*

12 pentagons meet to form
a Dodecahedron

THE HEAVENS

Plato

(c.428-348 BC)

The philosopher Plato concluded that they must be the fundamental building blocks
– the atoms – of nature, and assigned to them what he believed to be
the essential elements of the universe.

20 triangles meet to form
a Icosahedron

WATER

Tetryonic Solids

Despite their unique topologies Tetryonic solids are not unlike Platonic solids save that their toplogies are comprised entirely from complex hitherto undescribed **4nπ equilateral Planck mass-energy momenta geometries** that also match the Euler numbers of Platonic solids

+ tetryon

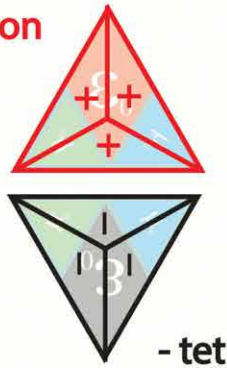

- tetryon

4 faces
6 edges
4 vertices

2

up quark

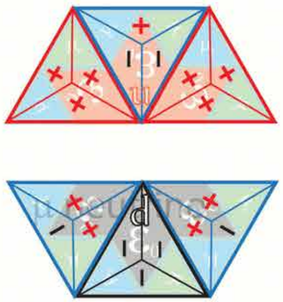

down quark

8 faces
12 edges
6 vertices

2

mass geometries

4π	*tetryons*	**4π**
	regular deltahedrons	
12π	*leptons*	**12π**

Matter topologies

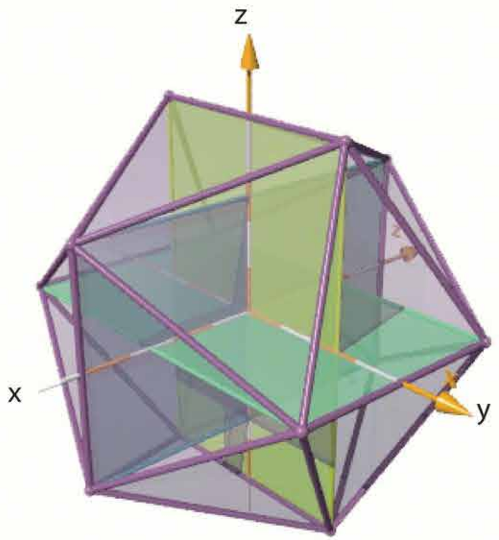

Matter topologies

8π	*quarks*	**12π**
	regular deltahedrons	
20π	*Baryons*	**36π**

mass geometries

neutrino

electron

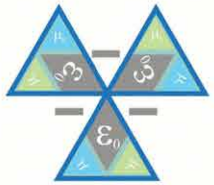

12 faces
18 edges
8 vertices

2

Their equilateral topologies are best described as regular topologic-deltahedrons:

tetra-delta-hedrals	**tetryons**	4π external charge fascia
octa-delta-hedrals	**quarks**	8π external charge fascia
dodeca-delta-hedrals	**leptons**	12π external charge fascia
icoso-delta-hedrals	**Baryons**	20π external charge fascia

note:
Charged mass-energy fascia geometries and edges become "hidden" upon the meshing of delta-hedra to form Matter topologies

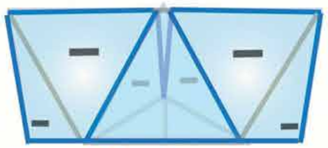

Neutron

Proton

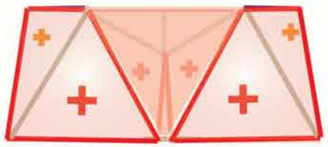

2

20 faces
30 edges
12 vertices

Copyright ABRAHAM [2008] - All rights reserved

Euclidean geometry

Euclid

(c.330-275 BC, fl. c.300 BC)

Arguably the most influential Mathematics book ever written is Euclid's 'The Elements'

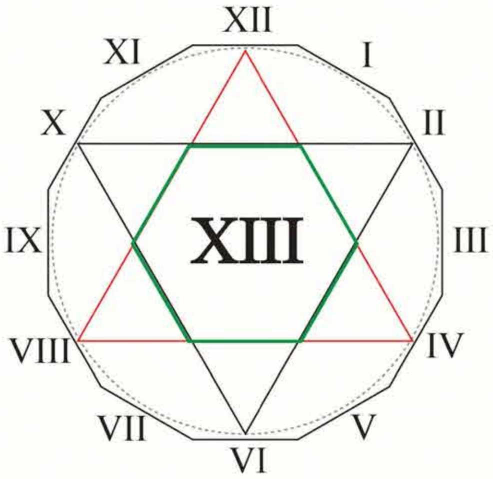

In all, it contains 465 theorems and proofs, described in a clear, logical and elegant style, and using only a compass and a straight edge.

The Elements - Book 1 - Definition 20

Of the trilateral figures, an equilateral triangle is that which has its three EQUAL sides

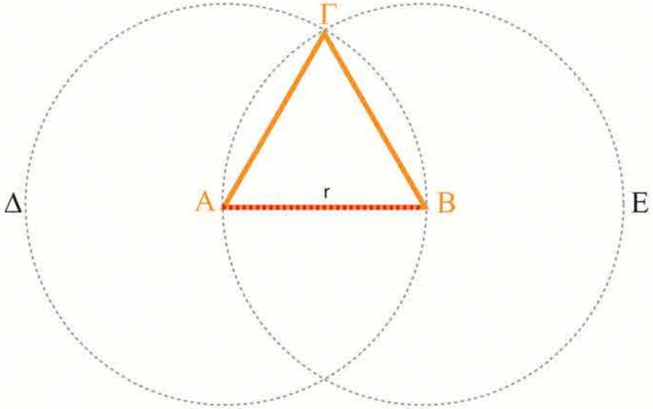

Euclid's Elements - Book 1 - Proposition 1
Method of constructing an Equialteral triangle

Euclid's five general axioms were:

Things which are equal to the same thing are equal to each other.

If equals are added to equals, the wholes (sums) are equal.

If equals are subtracted from equals, the remainders (differences) are equal.

Things that coincide with one another are equal to one another.

The whole is greater than the part

Copyright ABRAHAM [2008] - All rights reserved

Archimedes

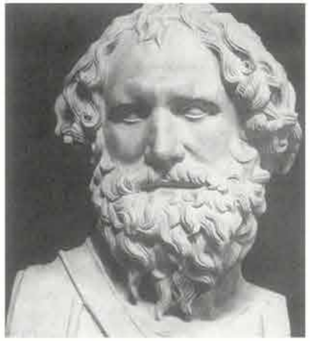

c. (287 BC – c. 212 BC)

As its definition relates to the circle,
π is found in many formulae in trigonometry and geometry,
especially those concerning circles, ellipses, or spheres.

It is also found in formulae from other branches of science,
such as cosmology, number theory, statistics, fractals,
thermodynamics, mechanics, and electromagnetism

Incorrect identification of Pi [c/d] as opposed to Pi radians
in Physics has led to the inappropriate association
of spherical particles to the physical sciences
whereas equilateral triangles & tetrahedra
form its true geometry

The number π is a mathematical constant that is
the ratio of a circle's circumference to its diameter.

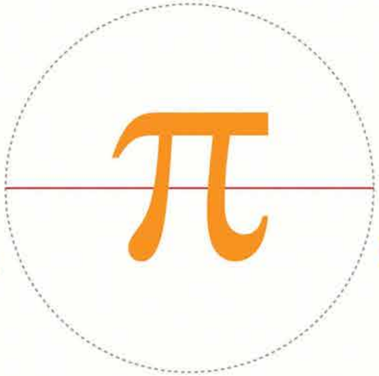

Pi

π is an irrational number
its decimal representation
never ends and never repeats.

$\pi = c/d$

The ratio C/d is constant,
regardless of the circle's size

3.141592654.......

Proof of the fact that C=2πr and how Archimedes proved it

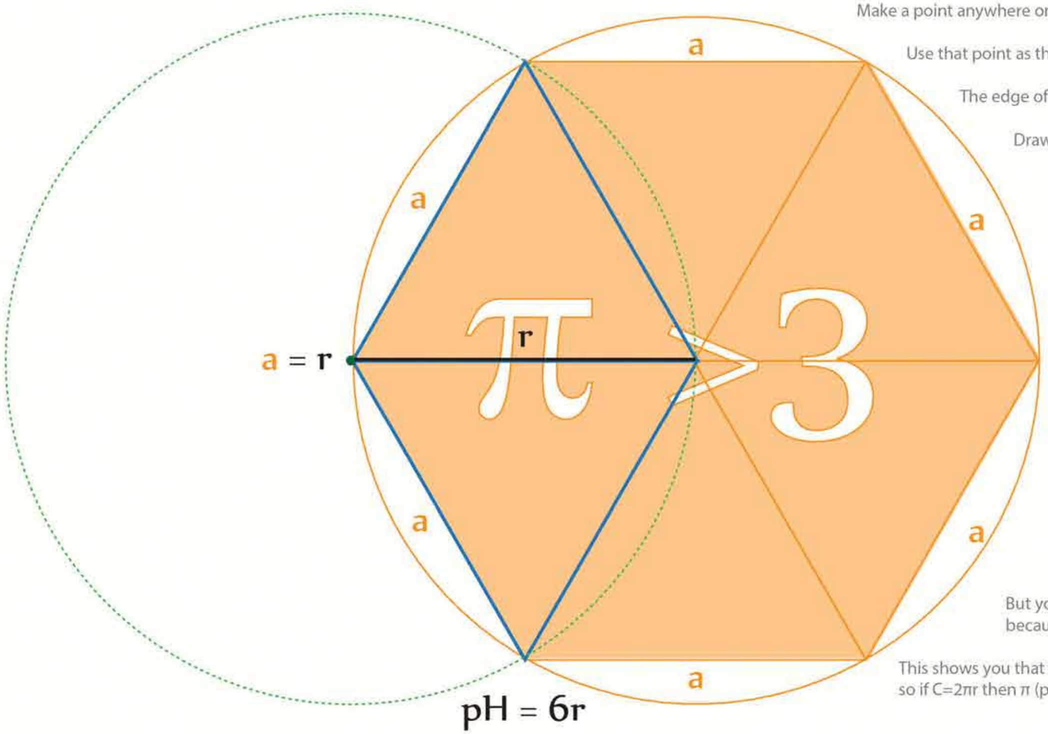

Draw any circle.

Make a point anywhere on the circumference of the [green] circle.

Use that point as the center of a second [blue] circle with the same radius as the green circle.

The edge of the blue circle should touch the center of the green circle.

Draw the line segment connecting the centers of the two circles.

That forms the radius of both of the circles.

Now draw the line connecting the center of the blue circle to where it crosses the green circle
on both sides, and complete the triangles.

You should have two equilateral triangles whose sides are equal to the radius of the green circle

Now extend all of the radius lines so they become diameter lines,
all the way across the circle, and finish drawing all of the triangles to connect them.

You've got six equilateral triangles now, that make an orange hexagon.

So the perimeter of your hexagon is the same as six times the radius of your circle.

But your circumference is a little bigger than the perimeter of your hexagon,
because the shortest distance between two points is always a straight line.

This shows you that the circumference of the blue circle has to be more than 6r,
so if C=2πr then π (pi) has to be a little bigger than 3, which it is.

The more sides we draw on our polygon, the closer we will get to the real value of pi (3.14159 etc.).
Using a polygon with 96 sides, Archimedes was able to calculate that π was a little bigger than 3.1408

The Golden Ratio

Two quantities are in the golden ratio (φ) if the ratio of the sum of the quantities to the larger quantity is equal to the ratio of the larger quantity to the smaller one.

The figure to the right illustrates the geometric relationship

[1.61803]

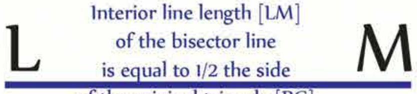

Interior line length [LM]
of the bisector line
is equal to 1/2 the side
of the original triangle [BC]
[AB] or [CA]

Φ

Golden ratio phi

$$\hbar = \frac{h}{2\pi}$$

Tetryonic geometry reveals
the maximum E-field amplitude
of the reduced Planck constant
to be an example of the
Golden ratio in physics

velocity vectors

The height of the triangle [ALM] produced by
the bisecting line is 1/2 the height of the height of [ABC]

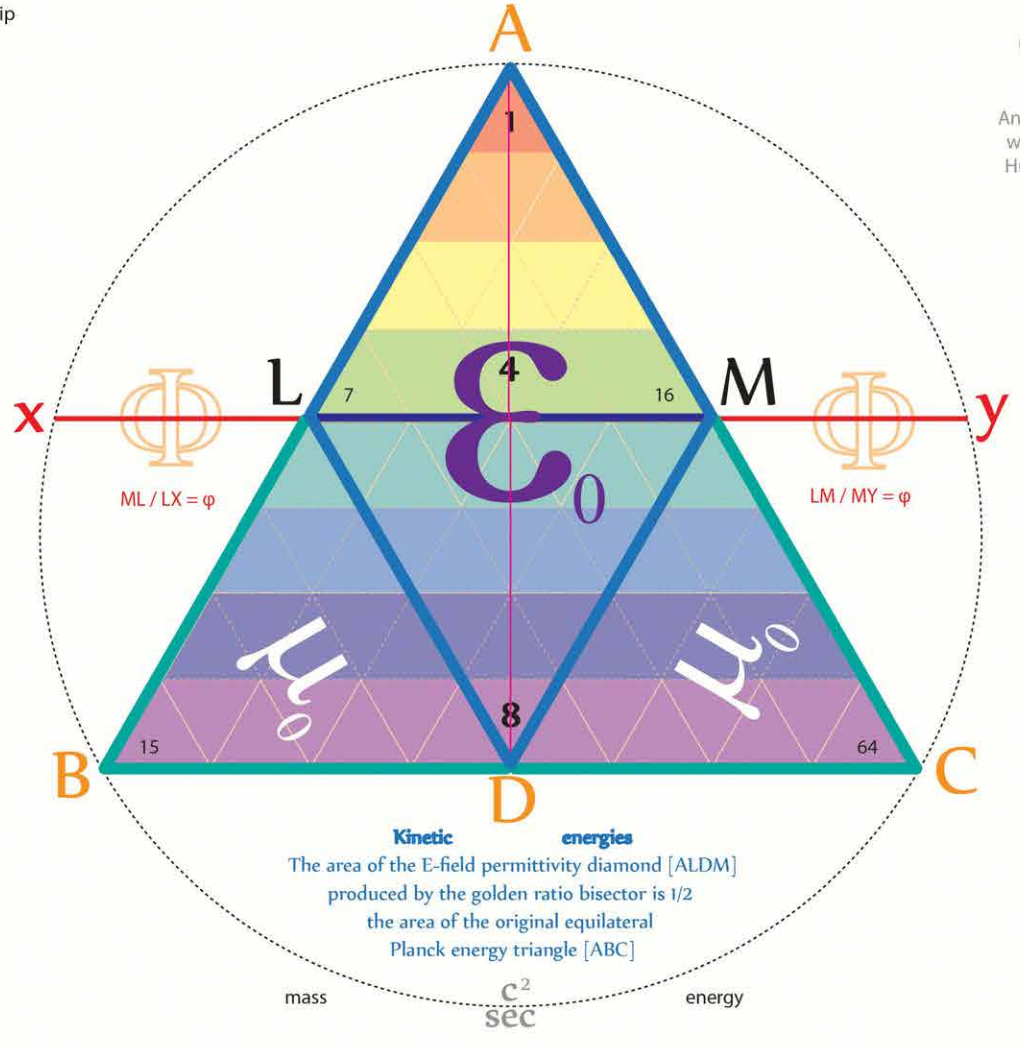

Kinetic energies

The area of the E-field permittivity diamond [ALDM]
produced by the golden ratio bisector is 1/2
the area of the original equilateral
Planck energy triangle [ABC]

Planck's constant [h]

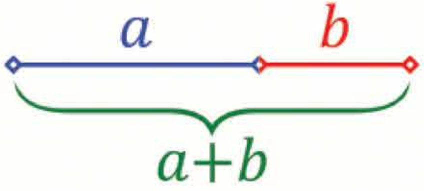

a+b is to a as a is to b

An intriguing showing of φ in an equilateral triangle was observed by George Odom, a resident of the Hudson River Psychiatric Center, in the early 1980s

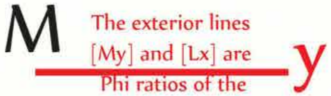

The exterior lines
[My] and [Lx] are
Phi ratios of the
interior line [LM]

Let L and M be the midpoints of the sides AB and AC of an equilateral triangle ABC

By measurement and
the Intersecting Chords Theorem

MX·MY = AM·MC.

which is of the form

(a + b)·b = a·a.

Denoting a/b = x, we see that

$1 + x = x^2$,

George Odom

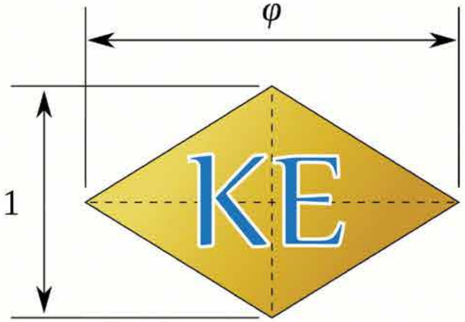

The Golden Rhombus

Applying the golden ratio (φ) to quantum scale
electrodynamic geometry we can quickly determine that
the linear momentum and magnetic moment vectors of photons
& EM waves can also be expressed as a golden ratio

vector velocities
vs
magnetic vector

Φ
Golden ratio phi
[1.61803]

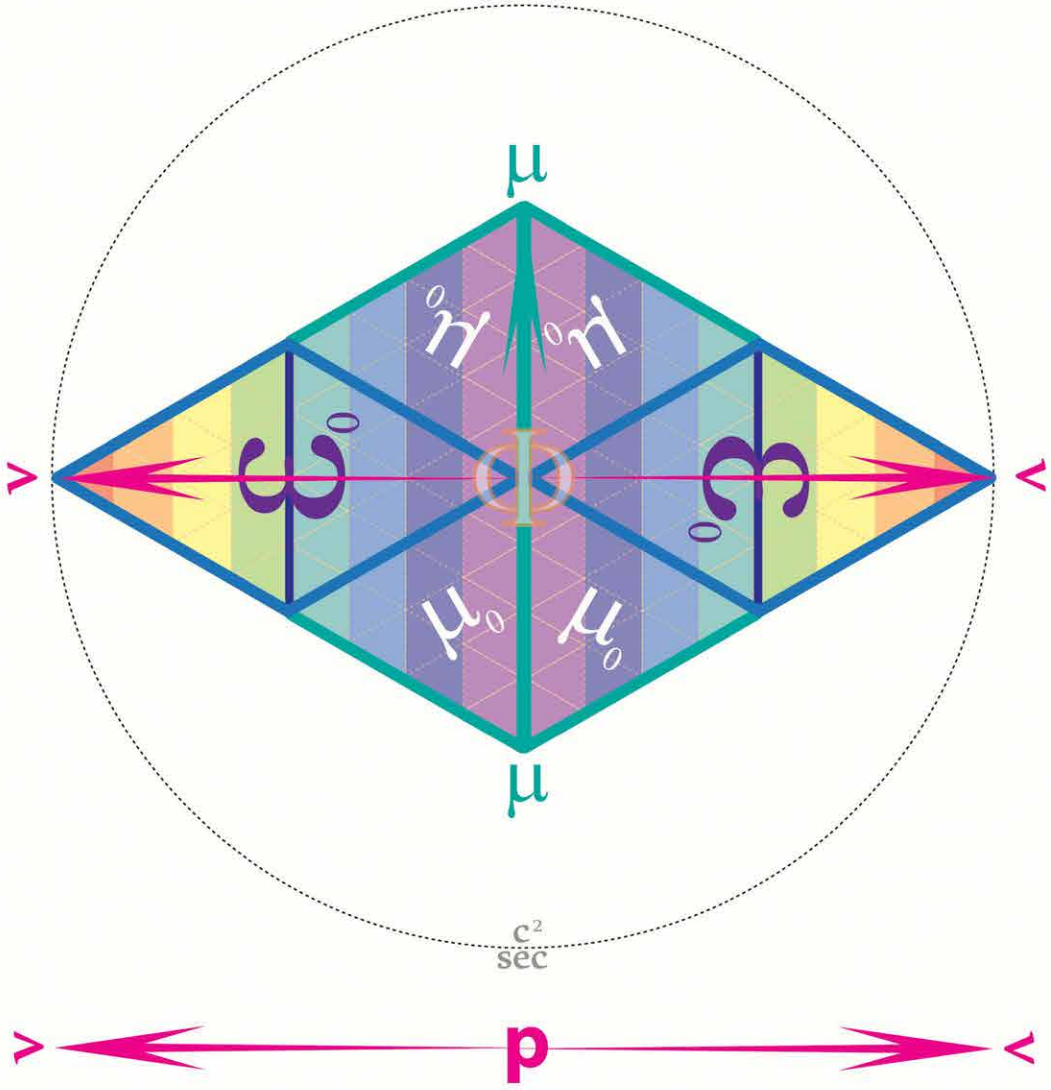

magnetic Force
vs
electric Force

**photons
of
EM energy**

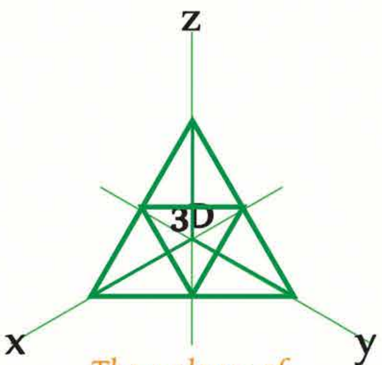

The 3 planes of Cartesian co-ordinates

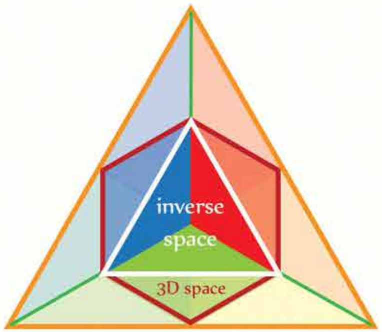
inverse space
3D space

An equilateral [square] triangle divided into 12 equal semitones

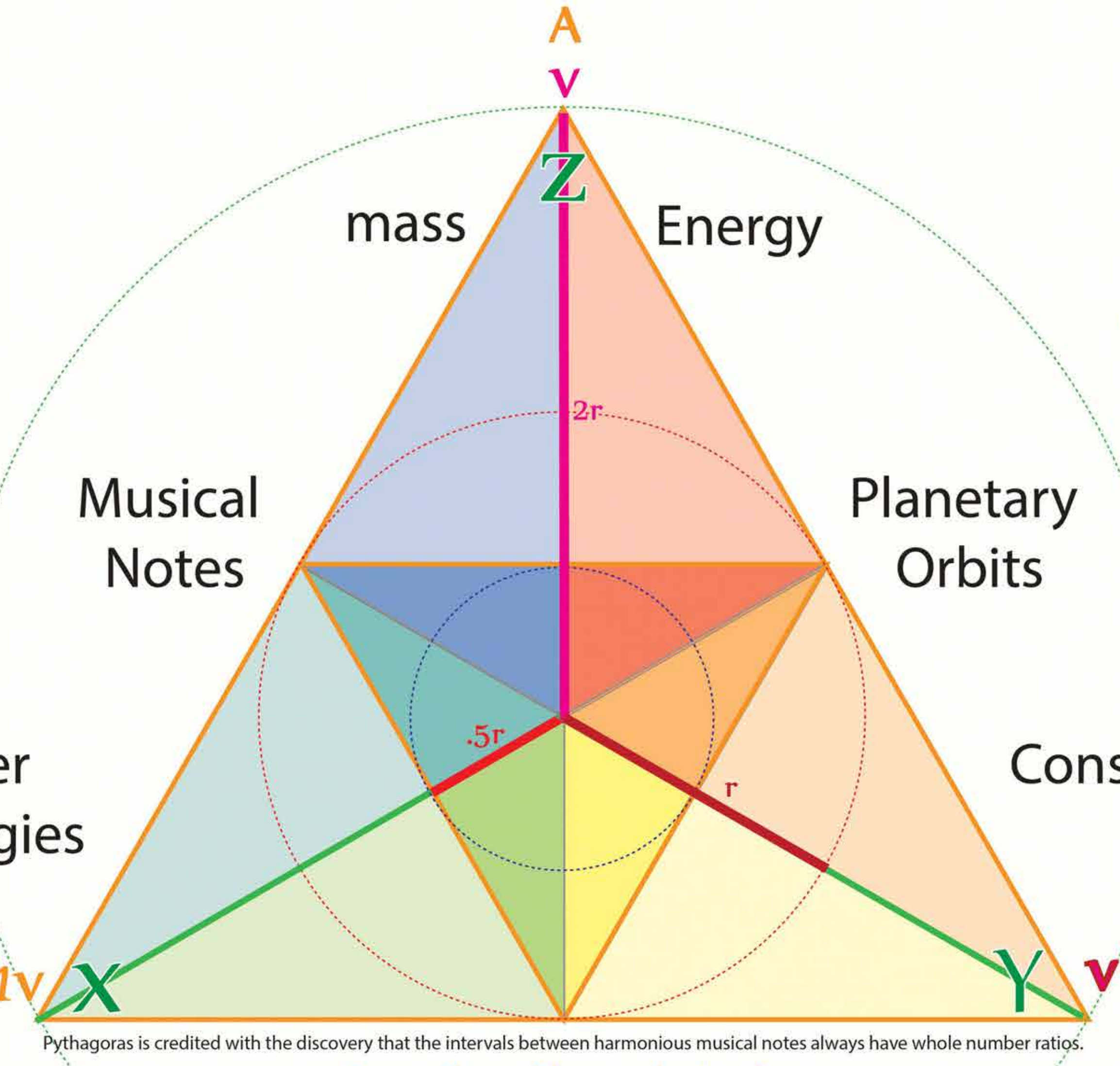
A
v
Z
mass
Energy
2r
Musical Notes
Planetary Orbits
.5r
r
Matter topologies
Conservation Laws
hv
X
Y
v²
B
C

Pythagoras is credited with the discovery that the intervals between harmonious musical notes always have whole number ratios.

An equilateral [square] triangle divided into 6 equal semitones and 6 equal quarter tones

With an Equilateral triangle, draw lines from the center of the circle to each vertex and each midpoint, creating six right triangles, as shown with six different colored triangles.
Each right triangle has a radius of the circle for one leg, and half of a side of the original triangle for another.
Any two right triangles sharing one of the radial lines must therefore be congruent, and that implies that the hypotenuses of the triangles are all equal.
This in turn shows that the six right triangles are all congruent, and so the large triangle they combine to form is equilateral

Equilateral Fifths

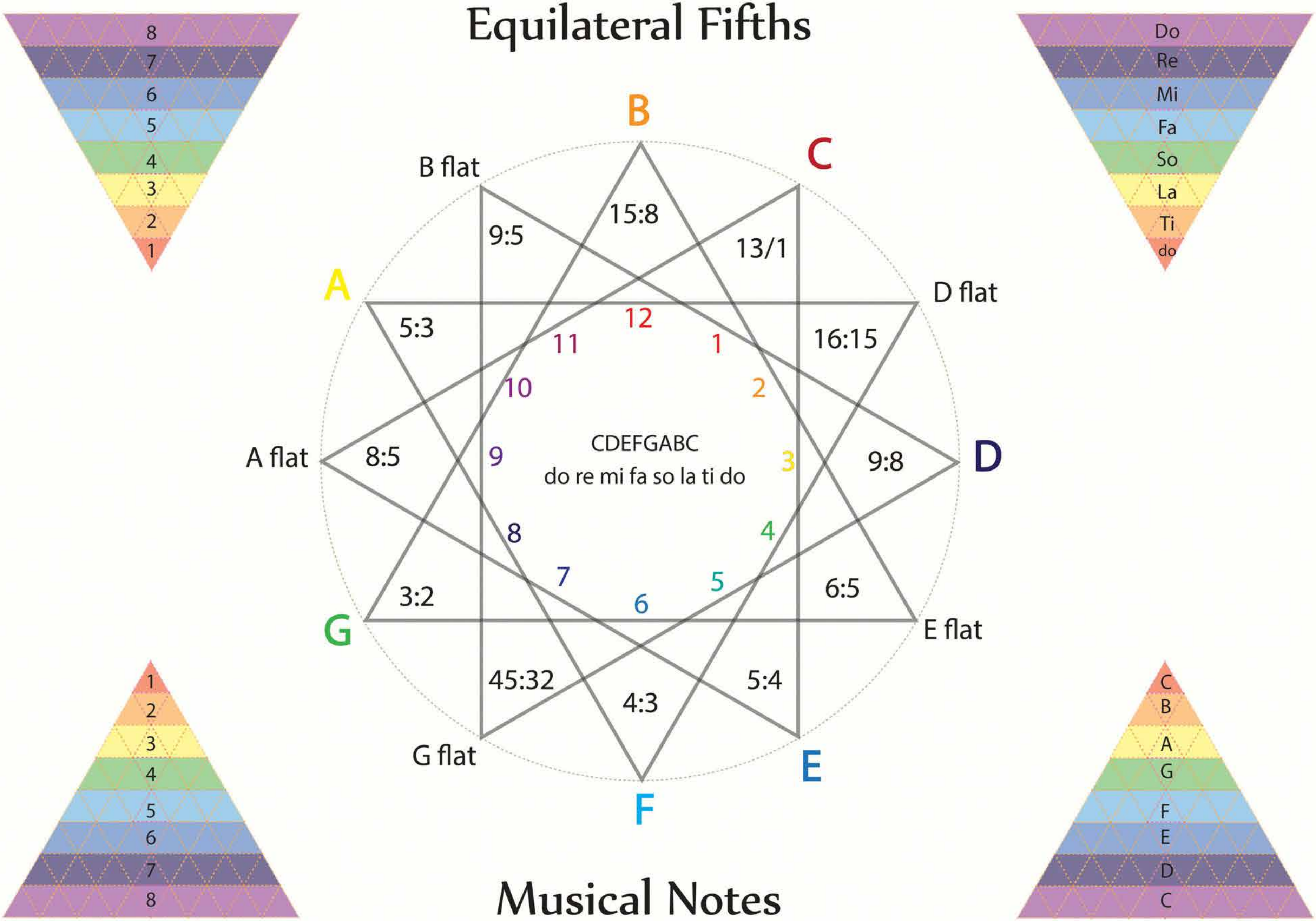
8
7
6
5
4
3
2
1
Do
Re
Mi
Fa
So
La
Ti
do
B
C
B flat
D flat
A
D
A flat
G
E flat
G flat
E
F
15:8
13/1
9:5
5:3
16:15
8:5
9:8
3:2
6:5
45:32
4:3
5:4
12
1
2
3
4
5
6
7
8
9
10
11
CDEFGABC
do re mi fa so la ti do
1
2
3
4
5
6
7
8
C
B
A
G
F
E
D
C

Musical Notes

Copyright ABRAHAM [2008] - All rights reserved

2D radiated mass-energies

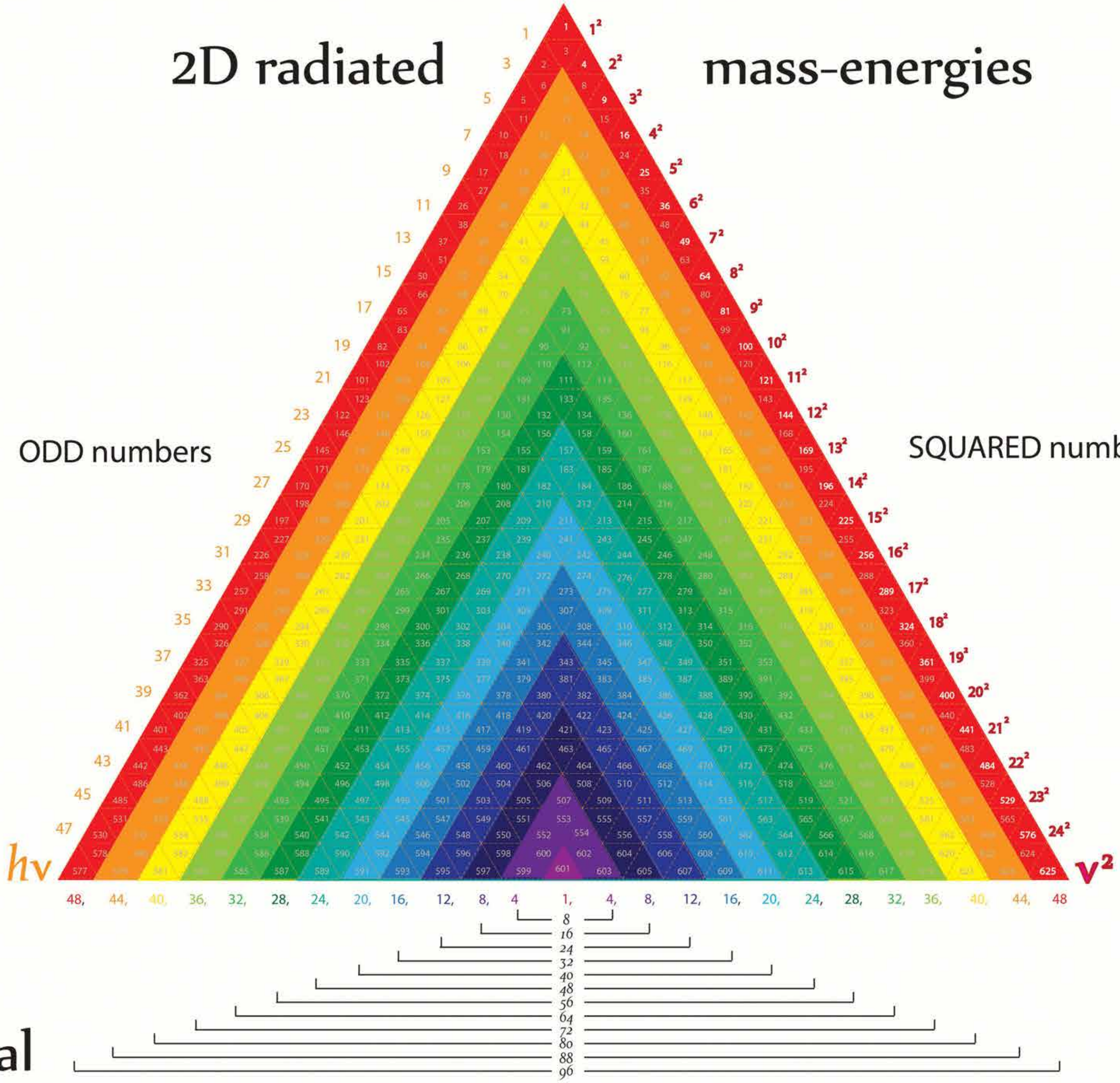
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
9^2
10^2
11^2
12^2
13^2
14^2
15^2
16^2
17^2
18^2
19^2
20^2
21^2
22^2
23^2
24^2
hv
v^2
48, 44, 40, 36, 32, 28, 24, 20, 16, 12, 8, 4
1,
4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48
8
16
24
32
40
48
56
64
72
80
88
96

ODD numbers

SQUARED numbers

Equilateral Octaves

Ternary Diagrams

Viviani's Theorem implies that lines parallel to the sides of an equilateral triangle provide (homogeneous/barycentric/areal/trilinear) coordinates for ternary diagrams for representing three quantities A,B,C whose sum is a constant (which can be normalized to unity).

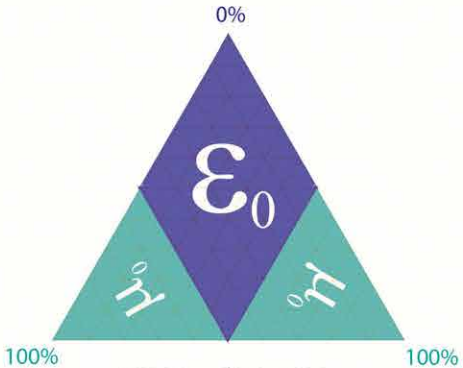
0%
ε_0
μ_0
μ_0
100%
100%

Ternary diagrams should NOT be used to model EM field strengths

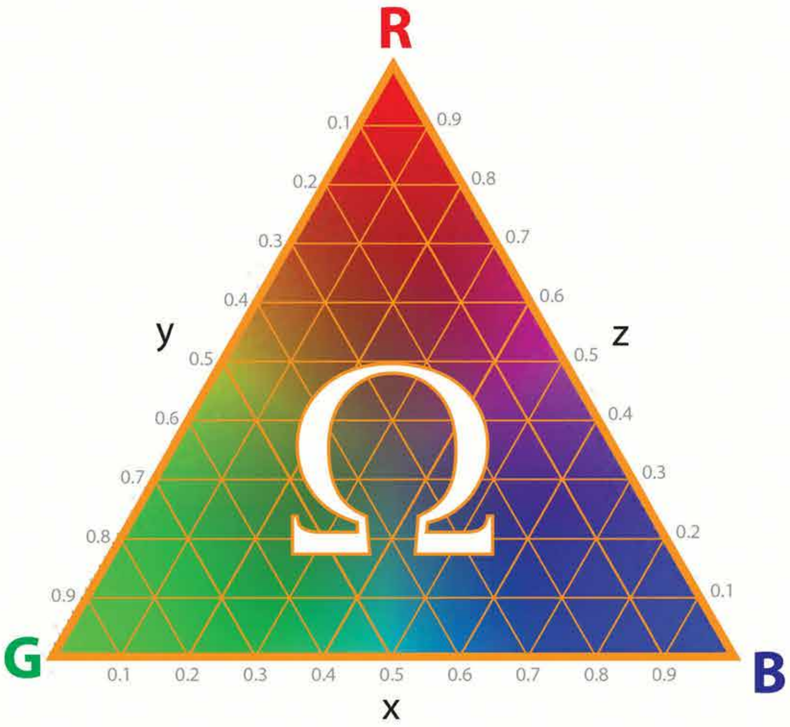
R
G
B
Ω
y
z
x
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

In a ternary plot, of EM energy the Electric field [E] and the Magnetic dipole [N-S] must renormalise to 1

A ternary diagram is simply a triangular coordinate system in which the 3 edges correspond to the axes.

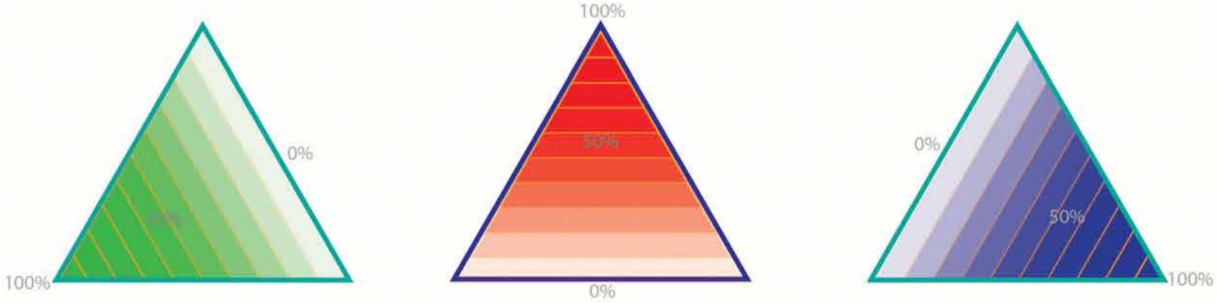
0%
50%
100%
100%
50%
0%
0%
50%
100%

Trilinear charts are commonly used for finding the result of mixing three components (such as gases, chemical compounds, soil, color, etc.) that add to 100% of a quantity.

Whilst the Pythagorean Theorem boasts a slightly greater economy of terms than the Eutrigon Theorem (Wayne Roberts 2003), the latter contains an important area not included in the former:

the area enclosed or swept out by the three points of the triangle in question

$$ab = a^2 + b^2 - c^2$$

the area of any eutrigon is equal to the combined areas of the equilateral triangles on legs 'a' and 'b', minus the area of the equilateral triangle on its hypotenuse 'c'.

The Pythagorean theorem

$$a^2 + b^2 = c^2$$

The square of the hypotenuse of a triangle is equal to the sum of the squares of its sides.

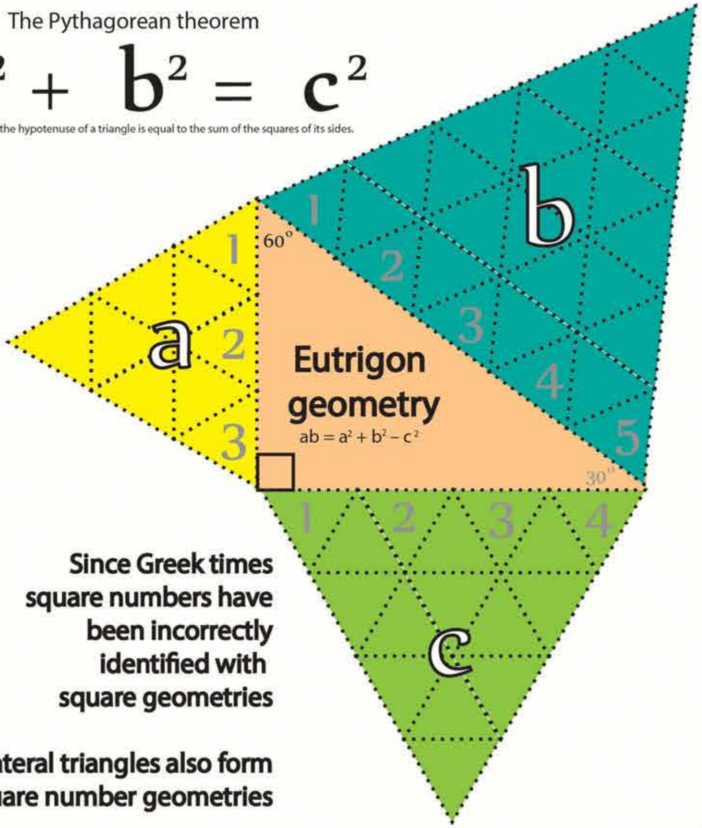

Since Greek times square numbers have been incorrectly identified with square geometries

Equilateral triangles also form square number geometries

Eutrigons

are an important new class of triangle (mathematically defined by Wayne Roberts), as the analogue of the right-triangle in orthogonal (Cartesian) coordinate geometry

Kepler's Second Law of planetary motion

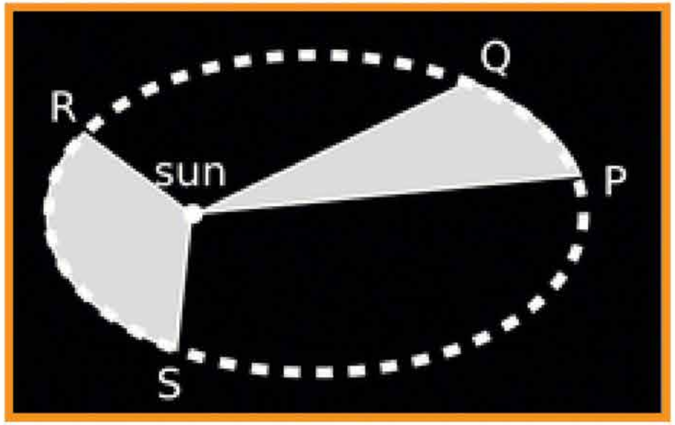

The orbit of every planet is an ellipse with the Sun at one of the two foci.

A line joining a planet and the Sun sweeps out equal areas during equal intervals of time.

The algebraic form of the Eutrigon Theorem, (like the algebraic form of Pythagoras' Theorem), is proven to be special case of the Cosine Rule...

Tetryonic theory reveals the equilateral [square] energy geometry that reveals the 'harmonics at play' in physical laws such as the second law of Kepler, and in many other phenomena in physics, chemistry, cosmology, biochemistry and number theory thus providing the foundation for the mathematics of quantum mechanics

Number theory

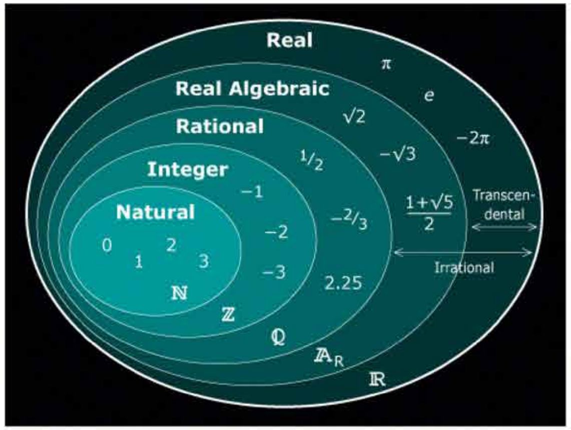
Real
Real Algebraic
Rational
Integer
Natural
0
1
2
3
ℕ
−1
−2
−3
ℤ
1/2
−2/3
2.25
ℚ
√2
−√3
(1+√5)/2
𝔸_R
π
e
−2π
Transcen-dental
Irrational
ℝ

The Pythagoreans also established the foundations of number theory, with their investigations of triangular, square and also perfect numbers (numbers that are the sum of their divisors).

They discovered several new properties of square numbers, such as that the square of a number n is equal to the sum of the first n odd numbers (e.g. 1 + 3 + 5 + 7 = 16).

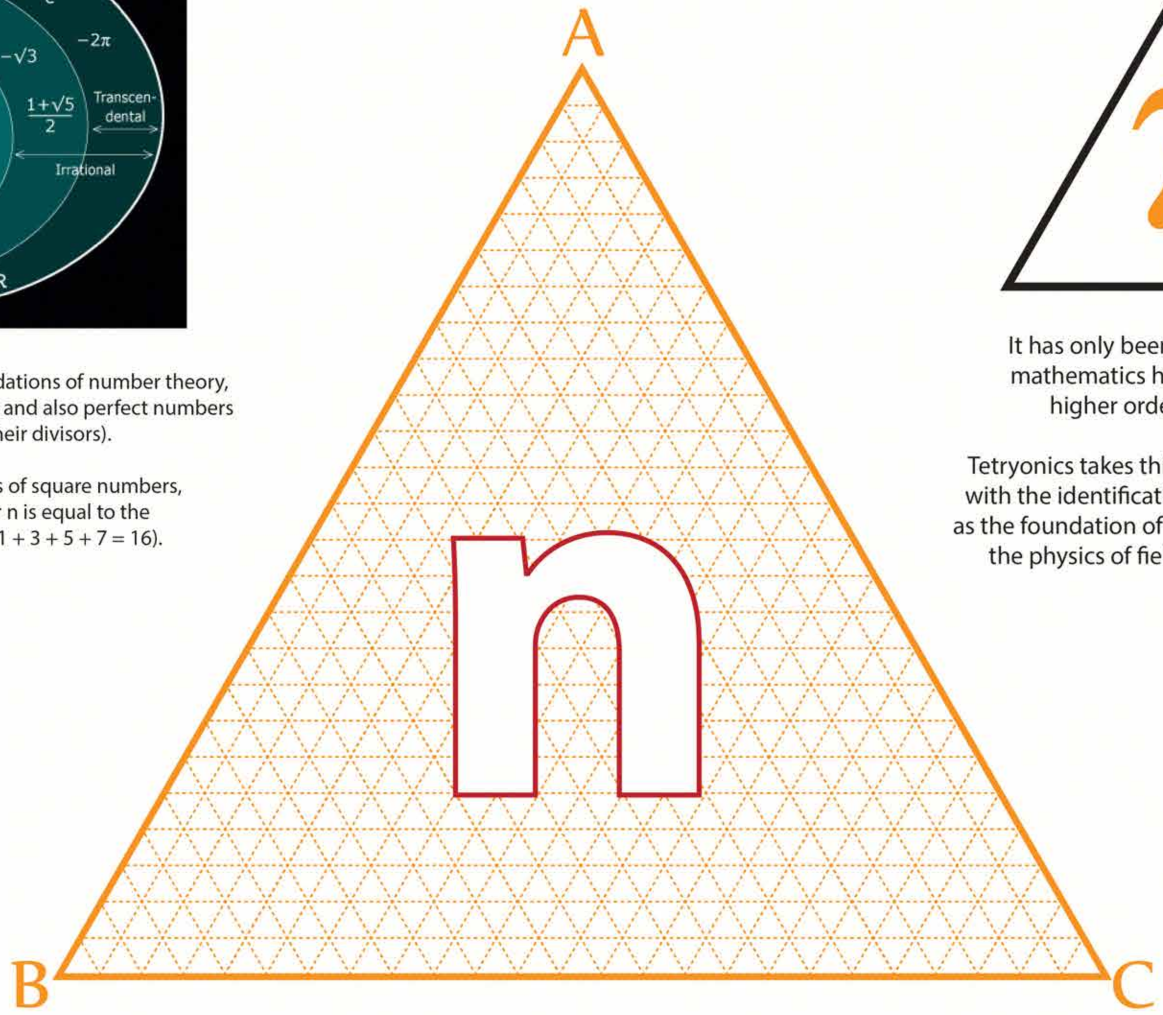

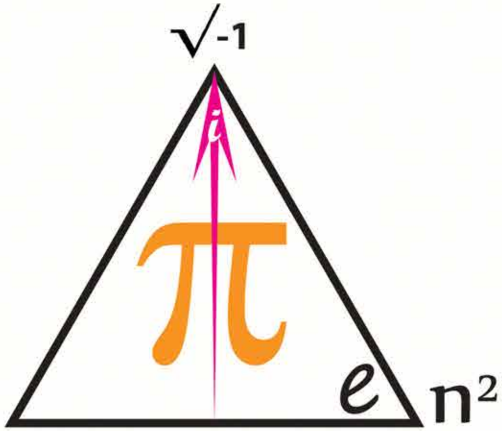

It has only been in recent centuries that mathematics has begun to explore the higher order irrational numbers

Tetryonics takes this investigation to new levels with the identification of equilateral geometries as the foundation of transcendental numbers and the physics of fields and particles in motion

What mathematics has failed to appreciate is the significance application of equilateral geometries to the 'square' numbers of physics and science in general

Copyright ABRAHAM [2008] - All rights reserved

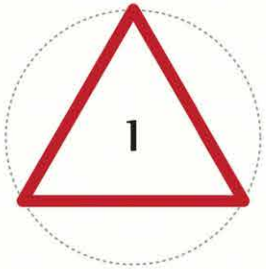

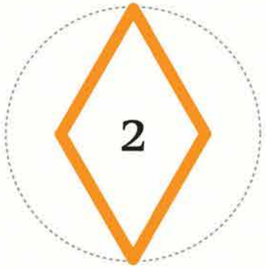

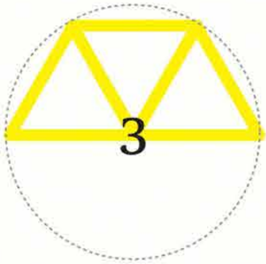

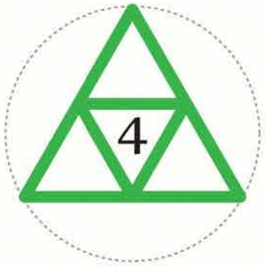

TETRYONICS

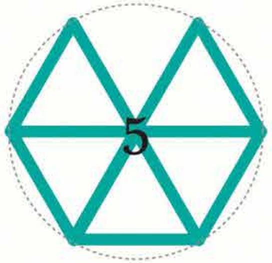

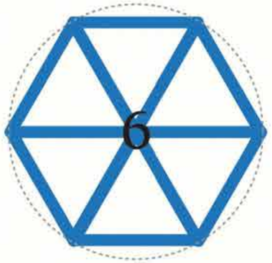

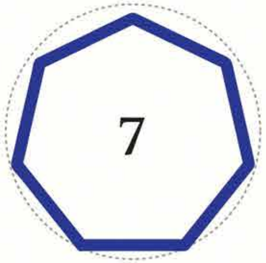

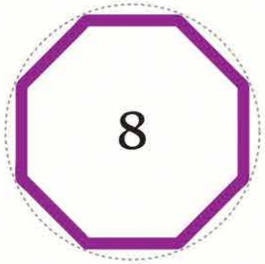

COUNTING POLYGONS

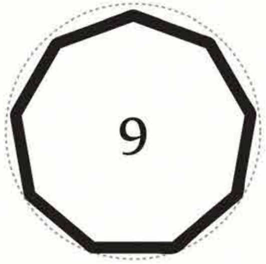

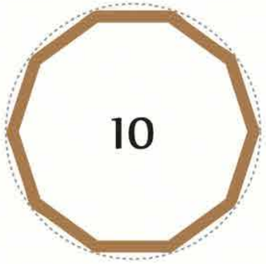

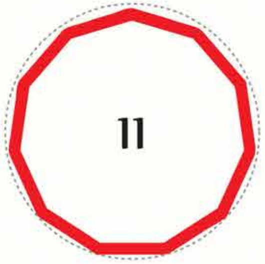

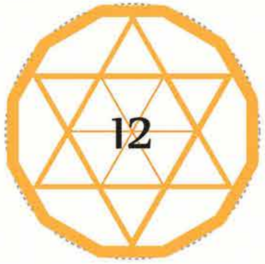

Tetryonics 82.02 - Counting Polygons

SQUARED energies in quantum mechanics are EQUILATERAL geometries

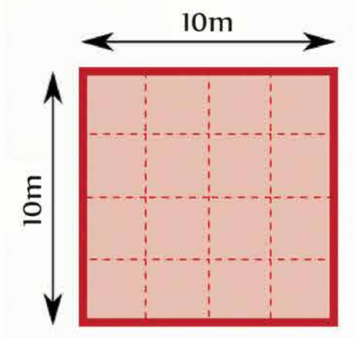

Square

area = s^2 = [100]

Circles

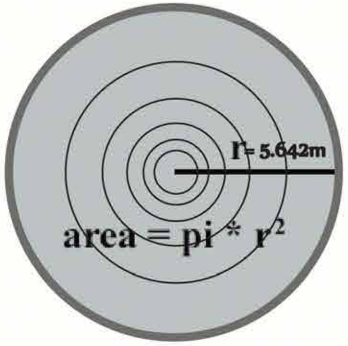
r= 5.642m
area = pi * r²

$= \text{pi} * [5.642]^2$
= 100

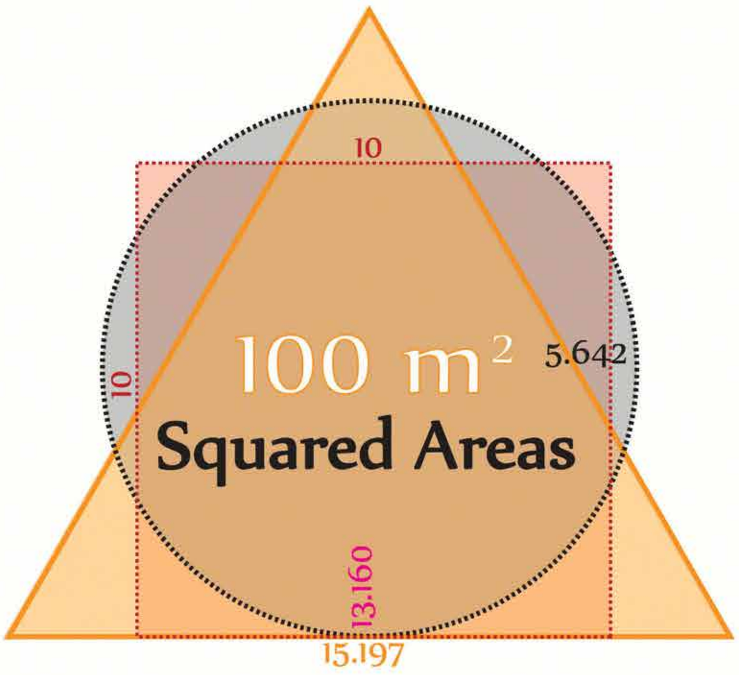
10
100 m²
Squared Areas
5.642
10
13.160
15.197

can be created by a number of planar geometries

For a long time it has been assumed by scientists (and mathematicians) that circular [and squared] geometries are the geometric foundation of all physics, leading to a serioulsy flawed model of particles and forces in quantum mechanics

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

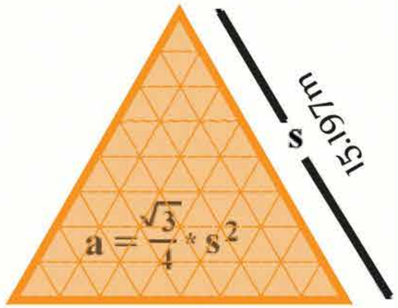
a = √3/4 * s²
s
15.197m

Equilateral

area = $(\frac{1}{2} * b) * h$

Triangles

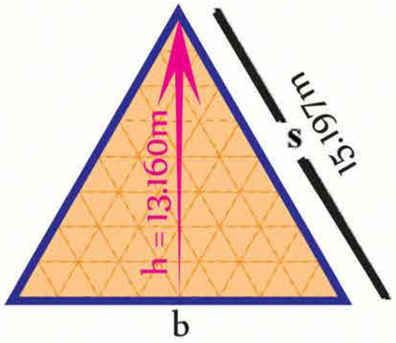
h = 13.160m
s
15.197m
b

$\overset{b}{[.5\text{x}15.197]} \times \overset{h}{13.160}$
= 100

Copyright ABRAHAM [2008] - All rights reserved

0.987654321

SQUARE ROOTS

80/81

ODDS

SQUARES

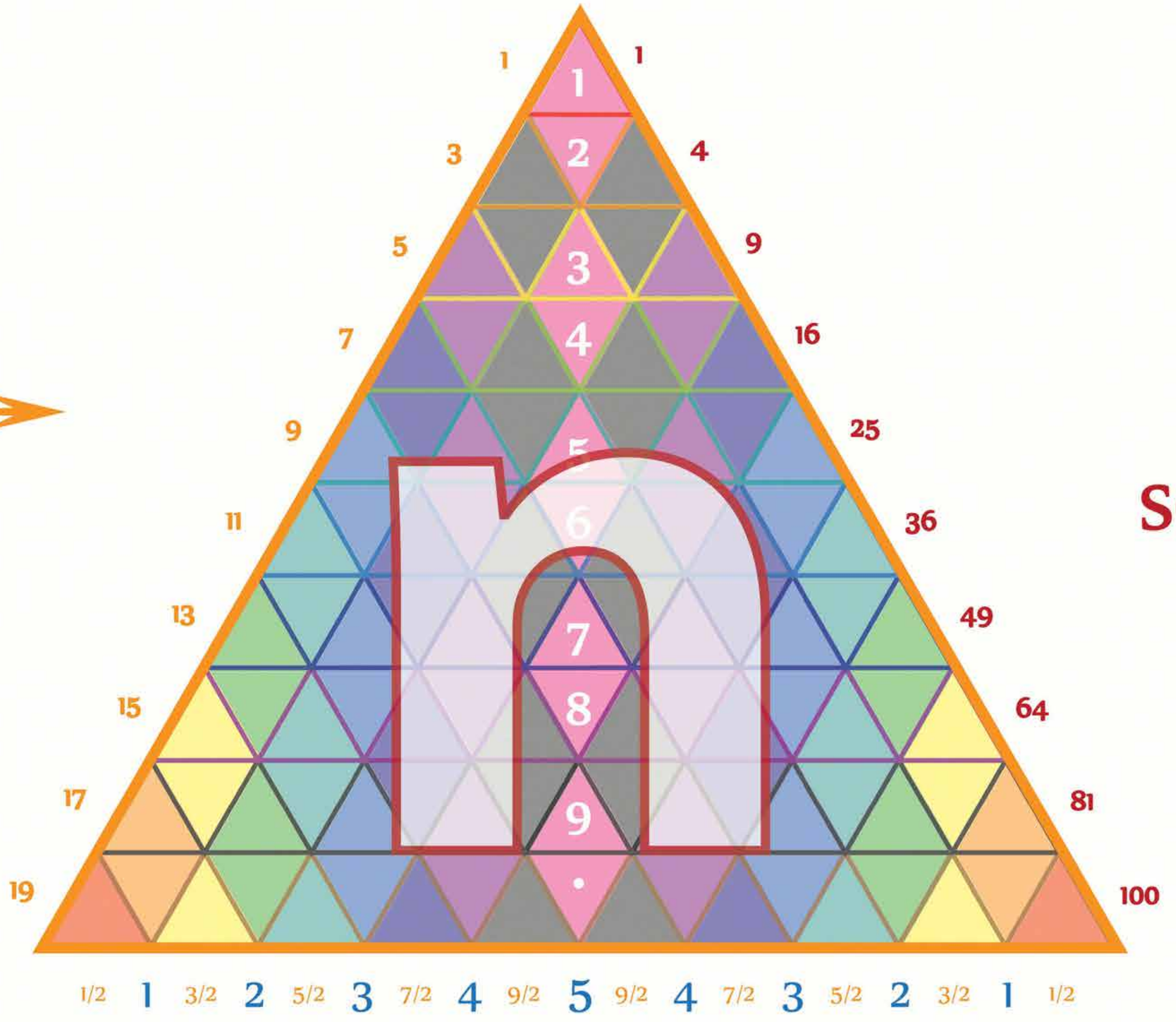
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
1
2
3
4
5
6
7
8
9
•
n
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2

EVEN DISTRIBUTIONS

Copyright ABRAHAM [2008] - All rights reserved

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the Planck charge quantum

$$2n-1$$

$$[[n] + [n-1]]$$

Triangular numbered geometries are NOT equialteral geometries

0, 1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

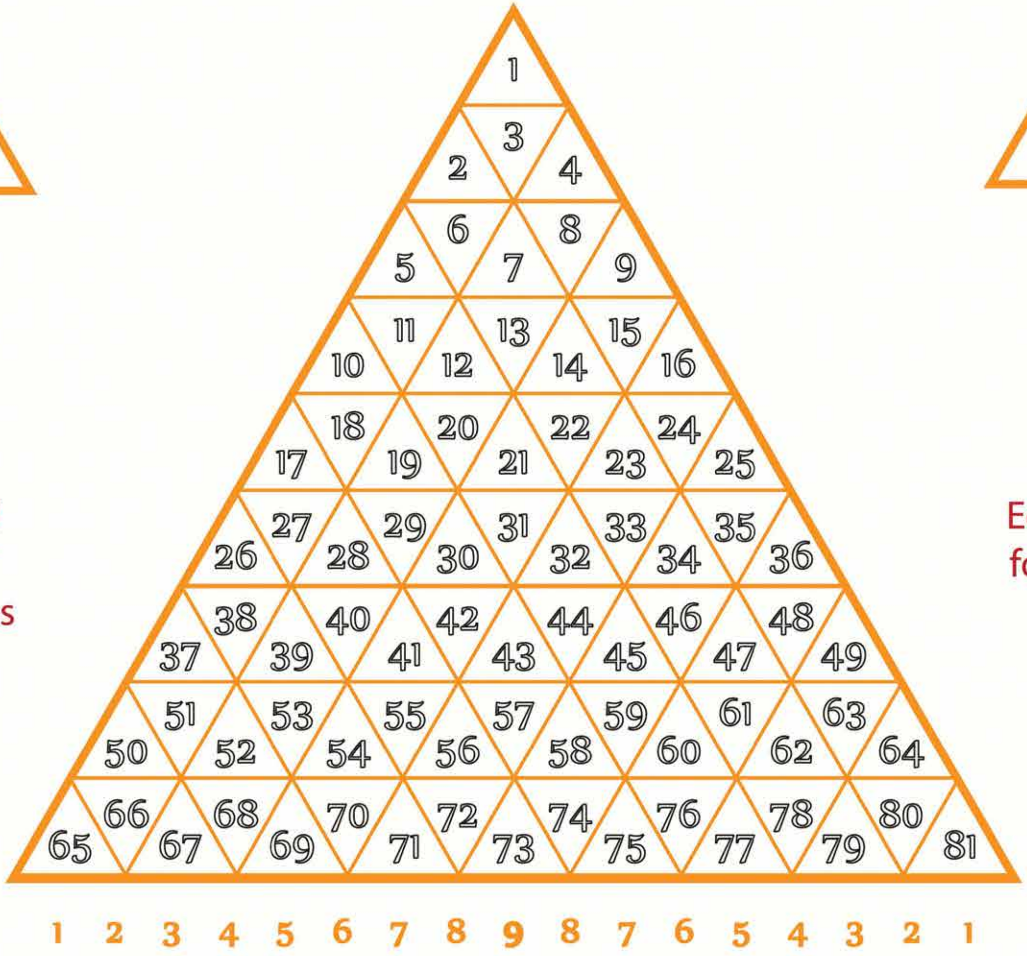
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1

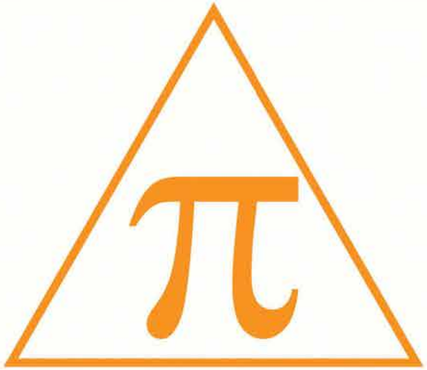

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form **SQUARE** number geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Equilateral energy quanta form a normal longitudinal distribution

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

Bosons are a transverse measure of scalar energy momenta

ODD numbers

An odd number is an integer which is not a multiple of two.

E
μ0
n
3
μ0
E
ε0
ε0
+
−
N
S
S
N

Quantum

Bosons have
ODD number
quanta

Quantum Levels

hv

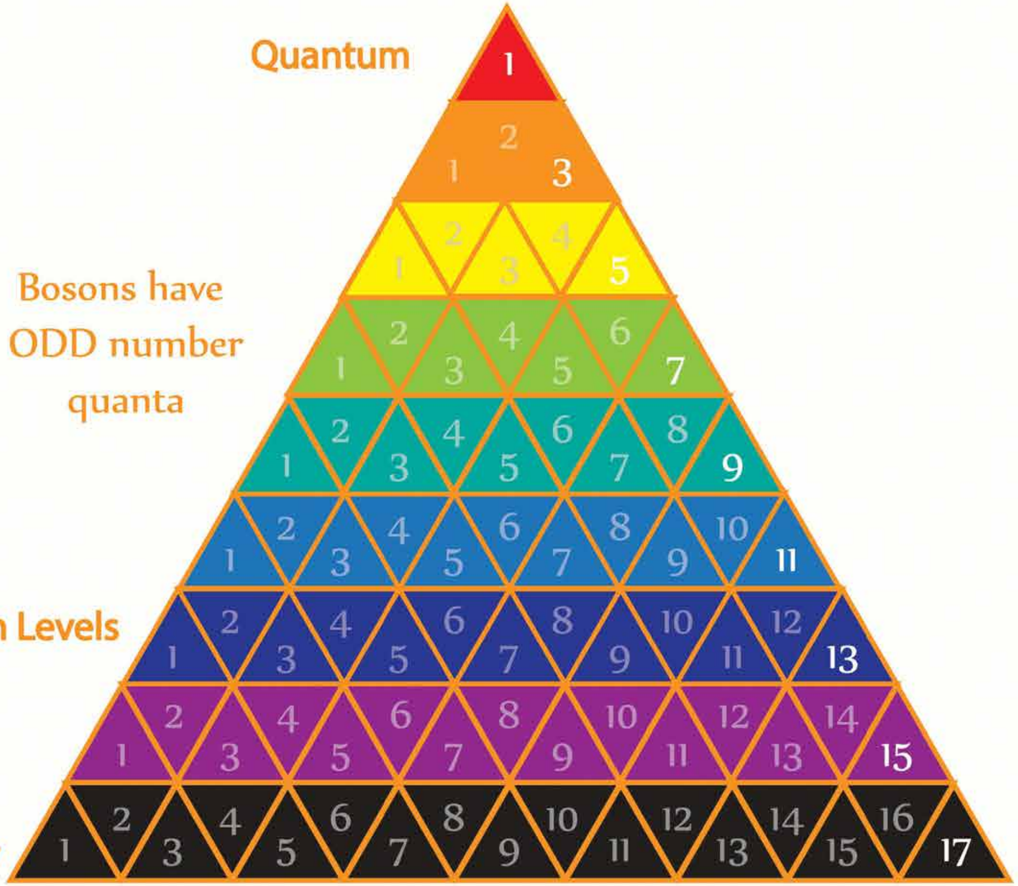
1
1 2 3
1 2 3 4 5
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11
1 2 3 4 5 6 7 8 9 10 11 12 13
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

1

3

5

7

9

11

13

15

17

An odd number, when divided by two, will result in a fraction

Tau radians

Around the whole outside of a circle,
there are about 6.283 radians - or, Tau radians

3.141592654

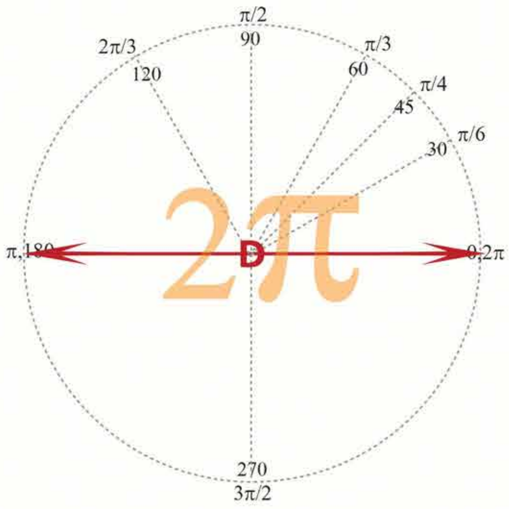
π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
0,2π
2π
D
270
3π/2

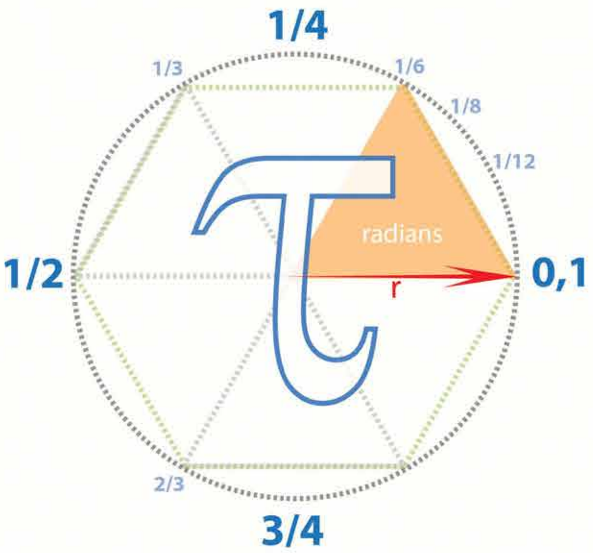
1/4
1/3
1/6
1/8
1/12
1/2
0,1
τ
radians
r
2/3
3/4

6.283185307

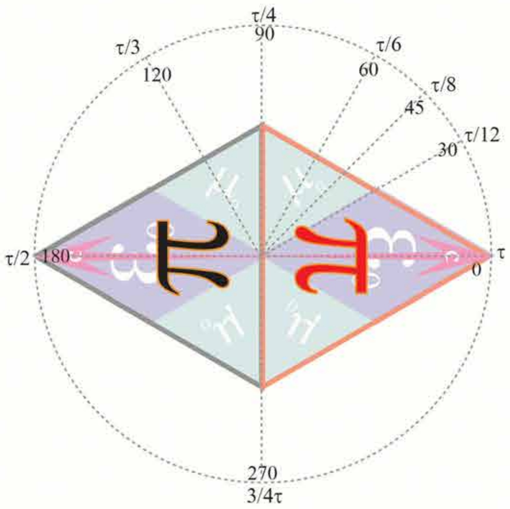
τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
τ
0
π
π
270
3/4τ

τ is a more 'natural' radian system for geometric physics than π
Tau = 2π = 360 degree rotation about a point

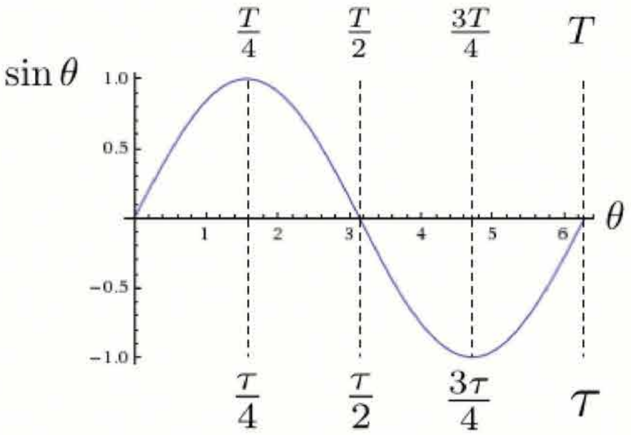

π
historically defined as the
ratio of a circle's circumference to its DIAMETER
should be redefined in physics to

τ
the ratio of its circumference to its RADIUS

in doing so many of the π/2 terms common to physics
will re automatically rationalised and will better reflect
the Tetryonic geometry of mass-ENERGY-Matter in motion

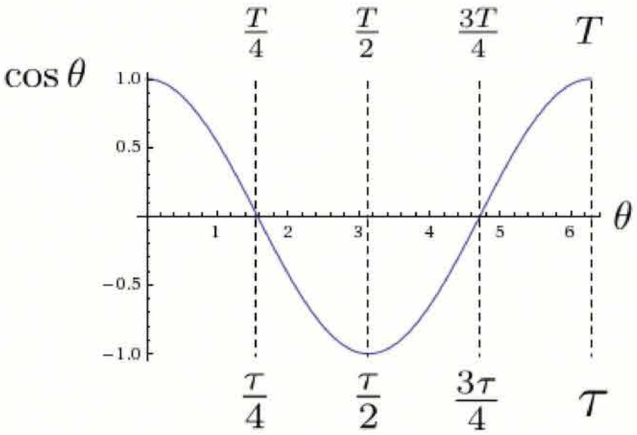

Copyright ABRAHAM [2008] - All rights reserved

Photons are a longitudinal measure of scalar energy momenta

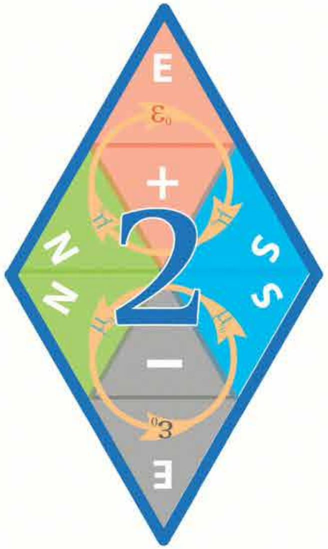

Photons are
EVEN number
quantums

EVEN numbers

An integer that is not an odd number is an even number

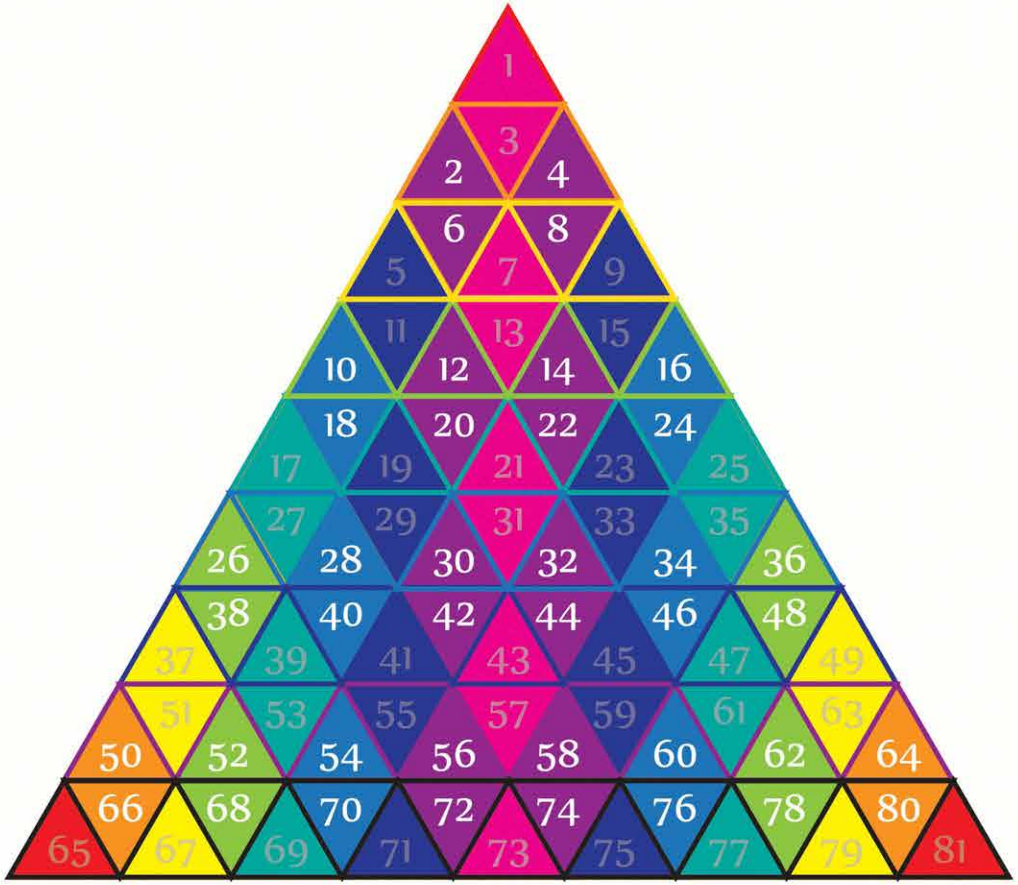
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

If an even number is divided by two, the result is another whole number

1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

If an odd number is divided by two, the result is a fractional number

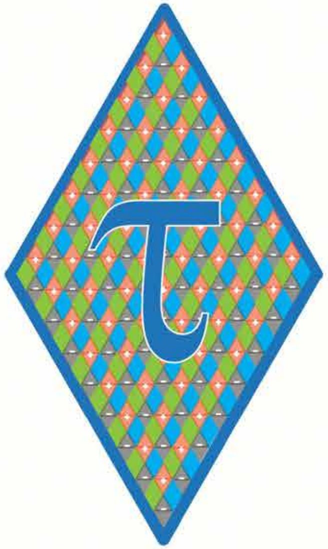

EM waves are
comprised of EVEN
numbered quanta

Triangular numbers

Historically, a triangular number counts quanta that can pack together to form an equilateral triangle

1, 3, 6, 10, 15, 21, 28, 36, 45, 55, ...

this form of geometric counting of same charges over-complicated the simpler physical reality

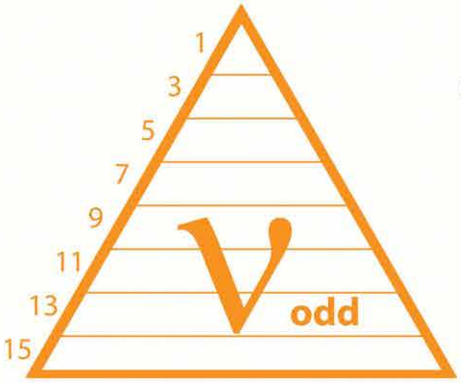
1
3
5
7
9
11
13
15
ν odd

$$2\nu - 1$$

$$[\nu] + [\nu - 1]$$

Equilateral chords or quantum levels are ODD numbers

1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,....

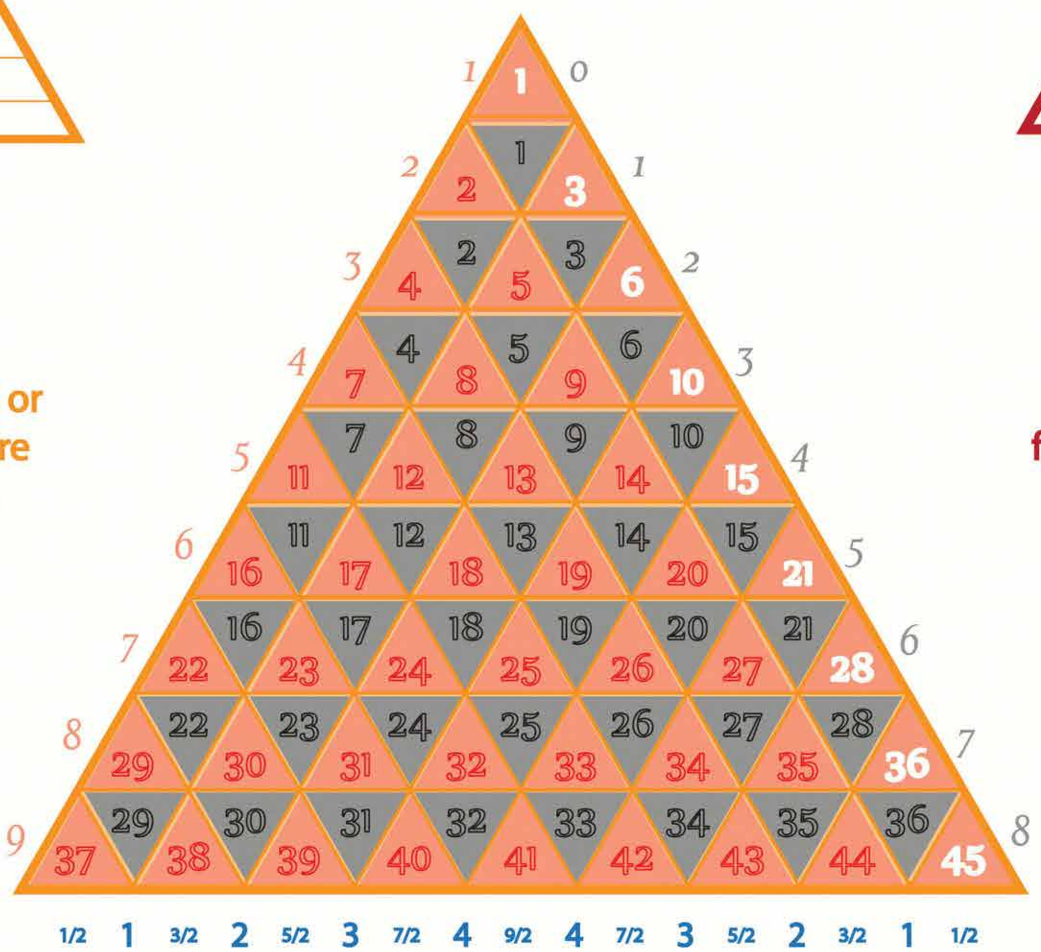
1
0
2
1
3
2
4
3
5
4
6
5
7
6
8
7
9
8
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

Triangular energy quanta form normal distributions

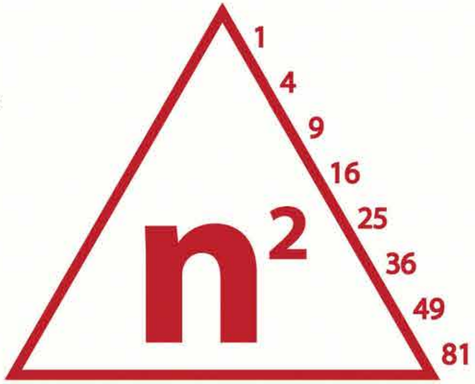
n²
1
4
9
16
25
36
49
81

$$\sum_{1}^{n} [2n-1]$$

Equilateral geometries form SQUARE numbered geometries

1, 4, 9, 16, 25, 49, 64, 81, 100, 121, 144, ...

Copyright ABRAHAM [2008] - All rights reserved

Squared numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

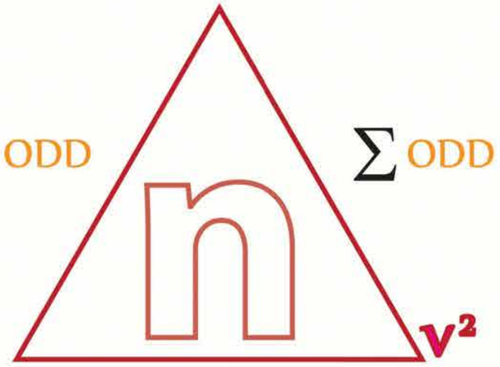
ODD
n
$\sum$ ODD
V^2

1
+
3
+
5
+
7
+
9
+
11
+
13
+
15
+
17

Square numbers result from
the summing of consecutive
ODD numbers

ODDS

2n-1

[[n] + [n-1]]

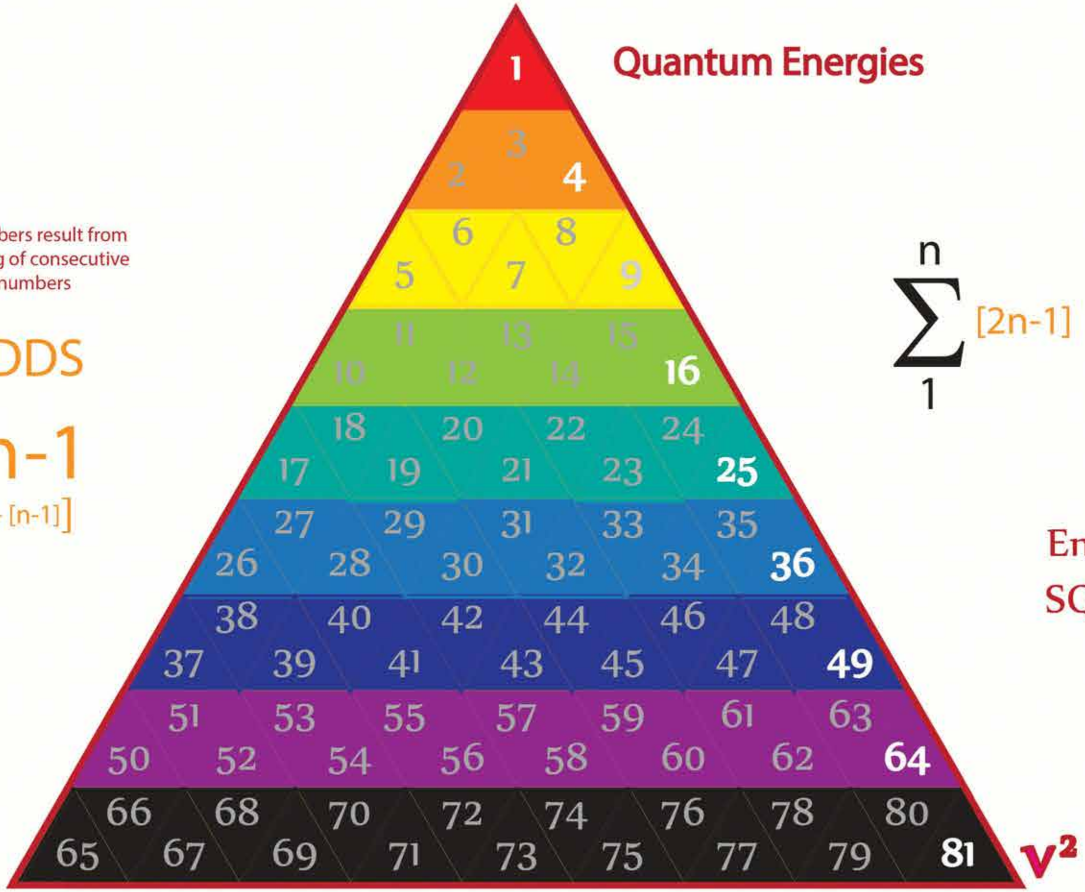
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
V^2

Quantum Energies

$$\sum_{1}^{n}[2n-1]$$

Energy levels have
SQUARE number
quanta

Compton Frequency

In Tetryonics Square numbers produce equilateral geometries

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

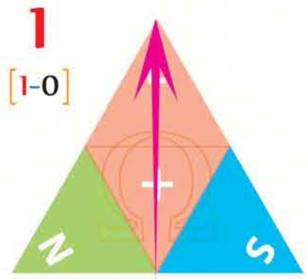

Root of positive one

-*i* and +*i*

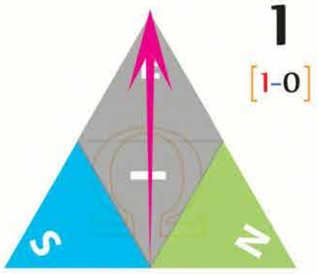

Root of negative one

Square roots of negative numbers have a basis in physical reality

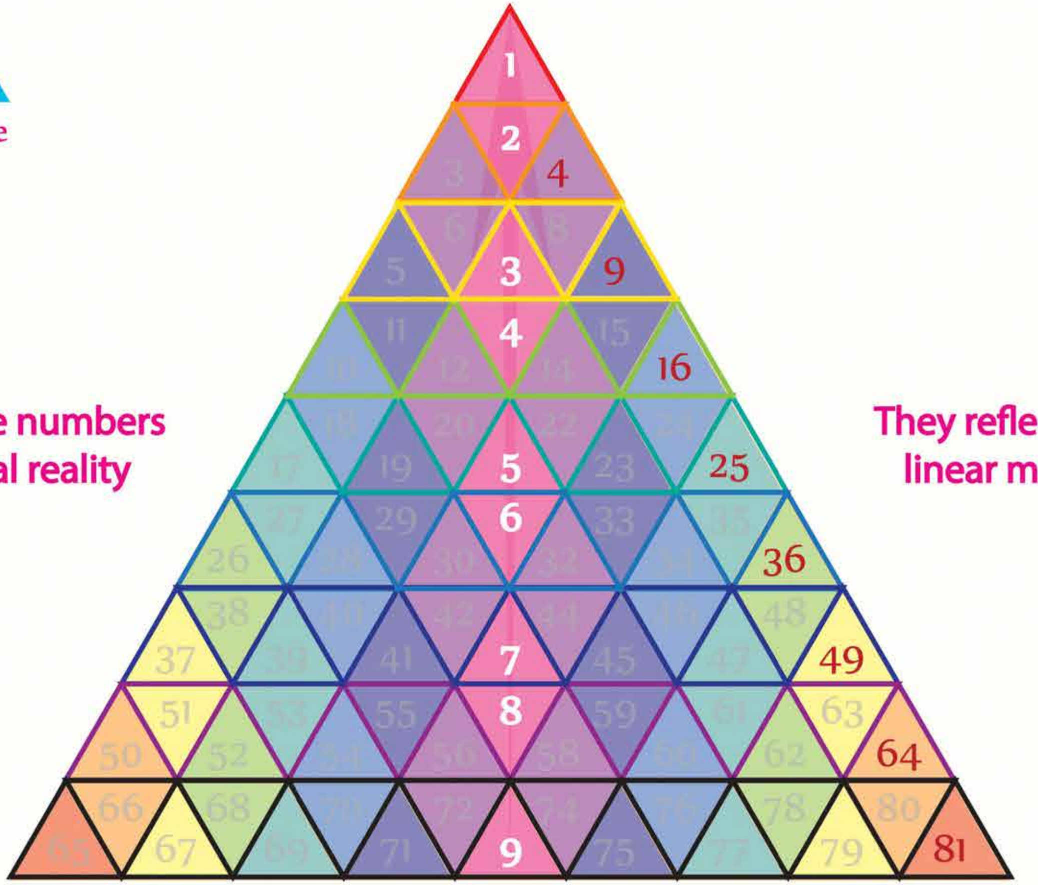

They reflect the real non-negative linear momentum of a system

A whole number with a square root that is also a whole number is called a perfect square in Tetryonic theory they are actually equilateral geometries

Copyright ABRAHAM [2008] - All rights reserved

Real Numbers

In mathematics, a real number is a value that represents a quantity along a continuous line. The real numbers include all the rational numbers,

Quantum levels

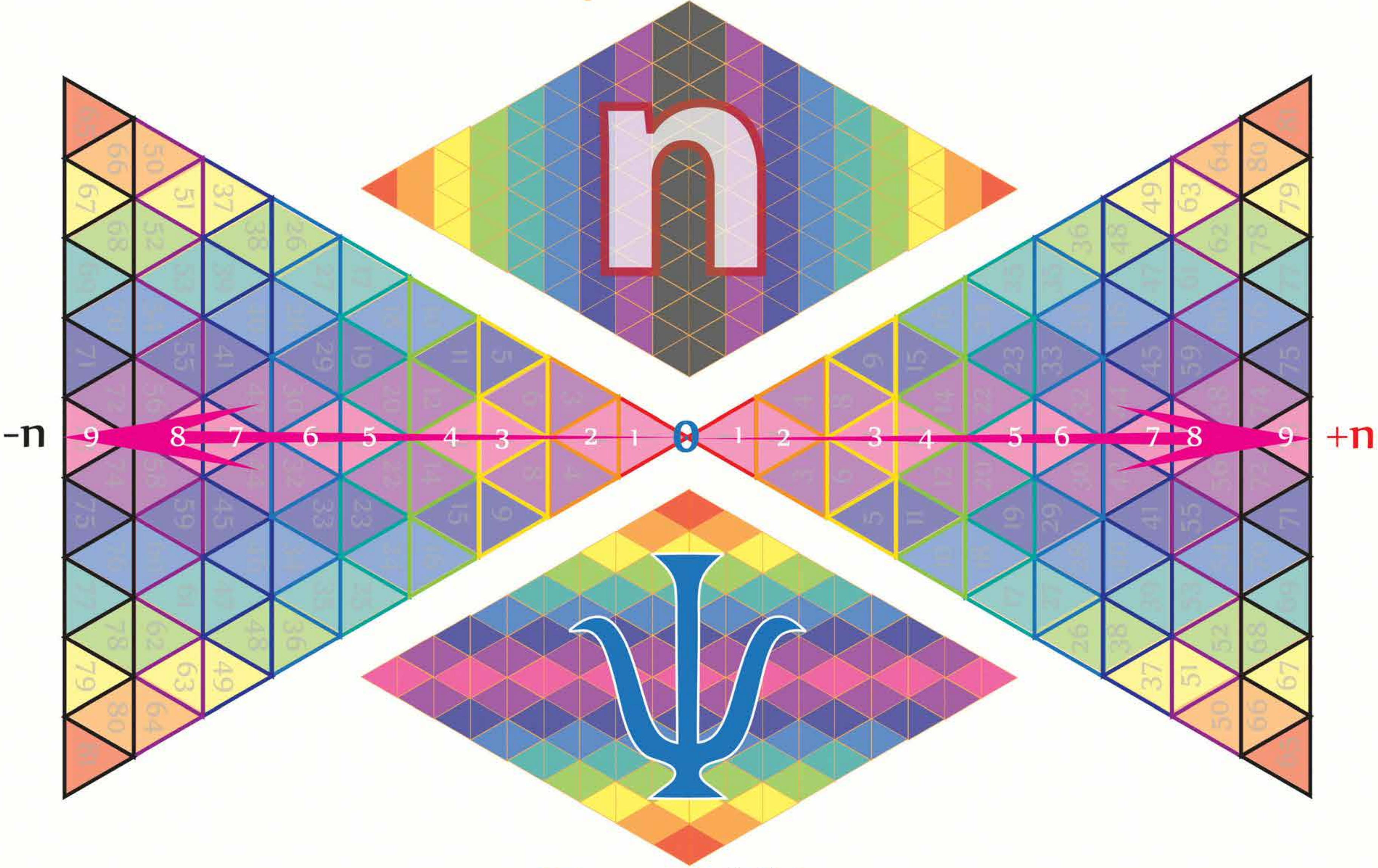

Wave probabilities

Copyright ABRAHAM [2008] - All rights reserved

ONE to ∞^2

Quantum levels →

ODD numbers

[2n-1]

1, 3, 5, 7, 9......

SQUARE numbers

the sum of consecutive
ODD numbers
1+3+5+7+9+......

Scalar Energies ↓

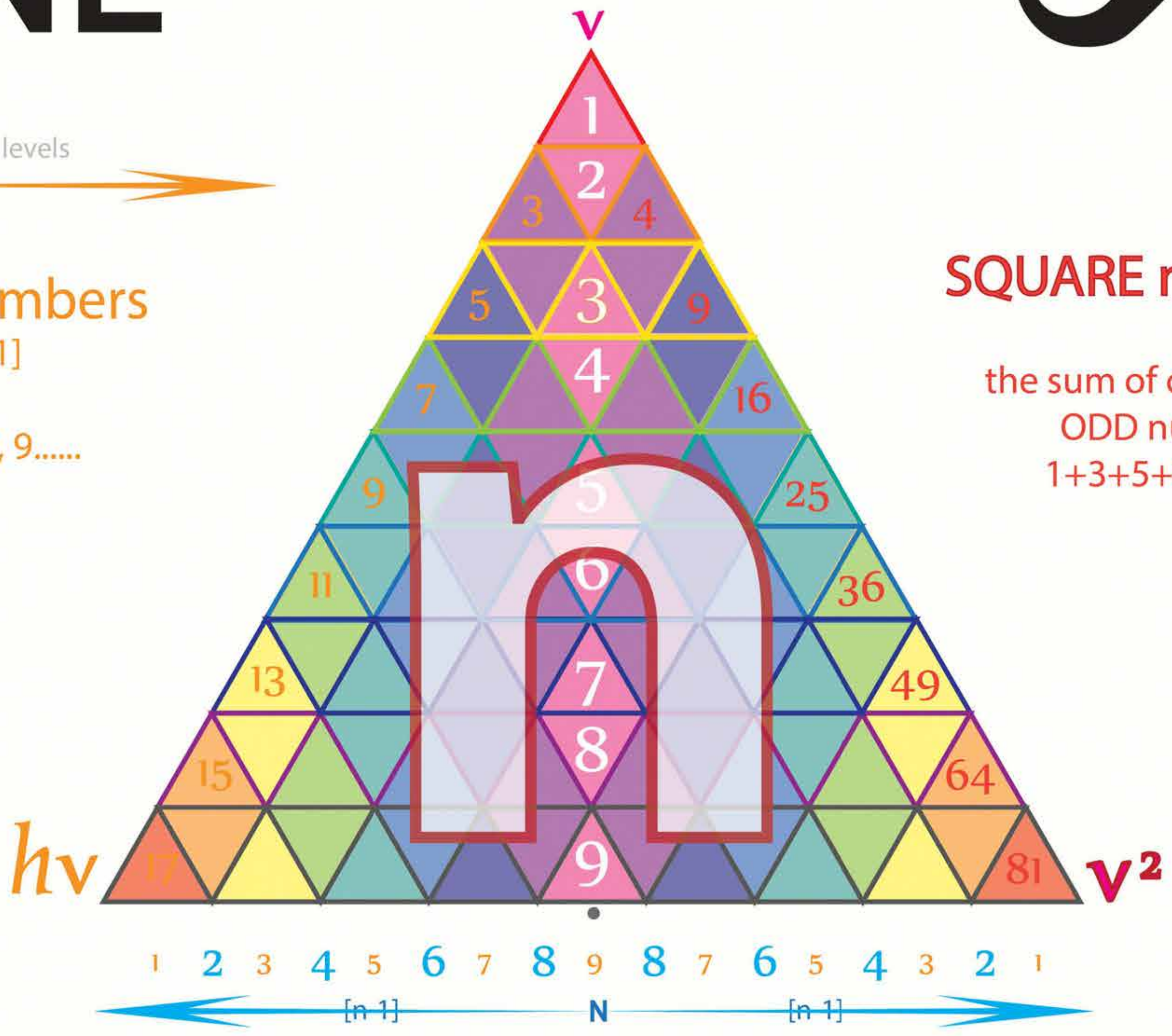
v
1
2
3
4
5
3
9
7
4
16
9
5
25
11
6
36
13
7
49
15
8
64
17
9
81
n
hv
v²
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
N
[n-1]

Linear energy momentum

NORMAL DISTRIBUTION

Tetryonics 82.13 - One to Infinity squared

Copyright ABRAHAM [2008] - All rights reserved

Basic Properties of nested scribed Equilateral Triangles

Given an equilateral triangle of side s

altitude

$\sqrt{3/2}\,S$

$3S$

perimeter

$\sqrt{3/4}\,S^2$

area

$\sqrt{3/6}\,S$

in-radius

$\sqrt{\pi/12}\,S^2$

in-circle area

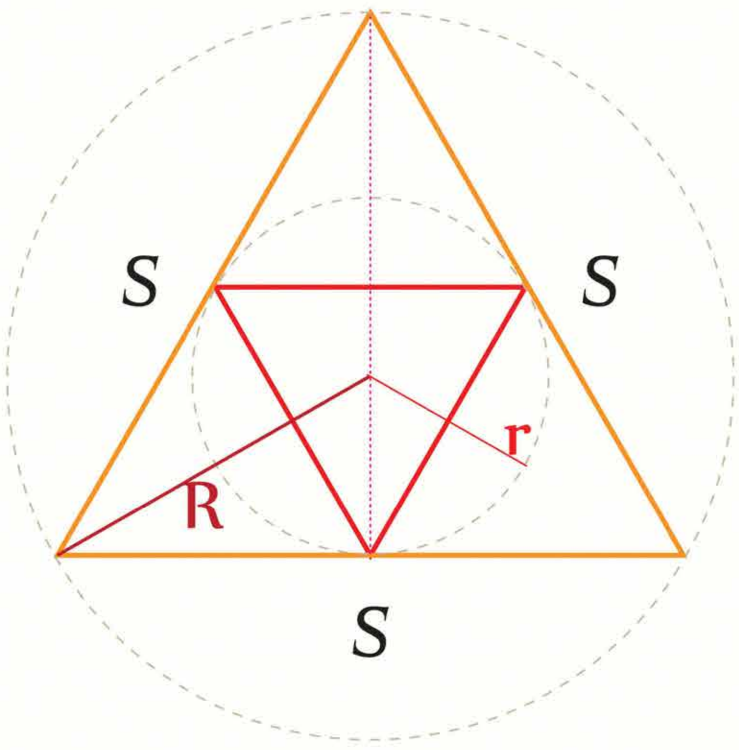

circum-radius

$\sqrt{3/3}\,S$

circum-circle area

$\sqrt{\pi/3}\,S^2$

Copyright ABRAHAM [2008] - All rights reserved

Tetryonic [equilateral] geometry

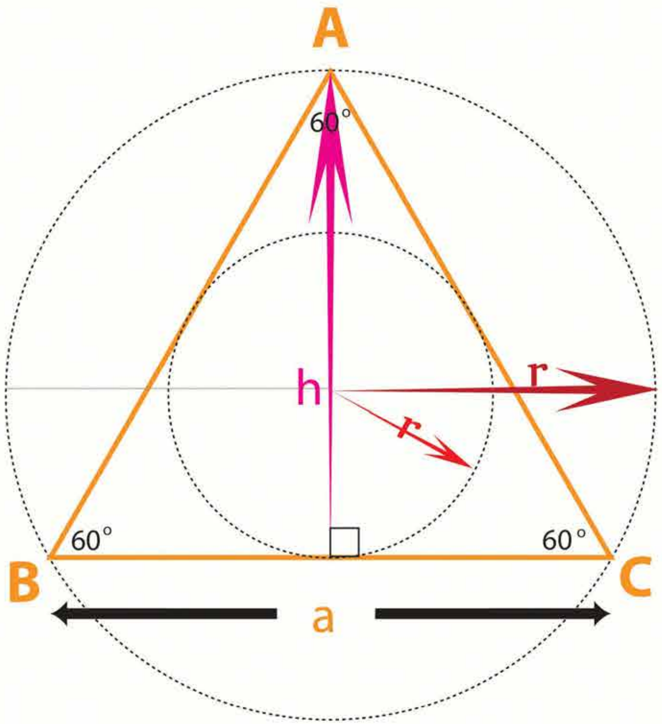

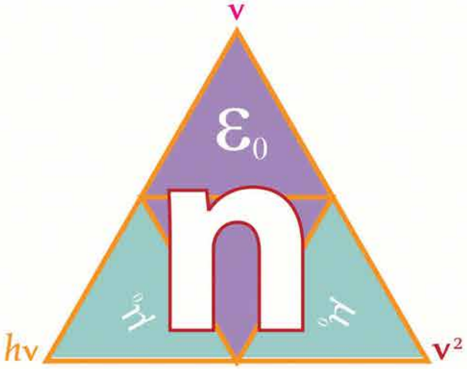

The equilateral triangle exhibits 'square symmetry'
it can always be divided into n^2 [number]
of smaller self-similar parts

**All triangles are
flat euclidean
π radian
geometries
[180°]**

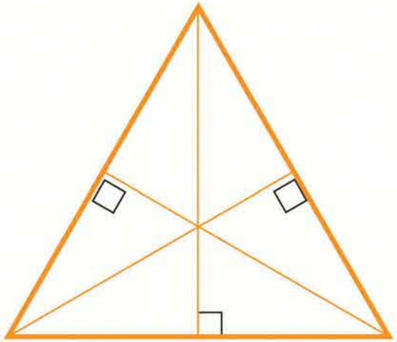

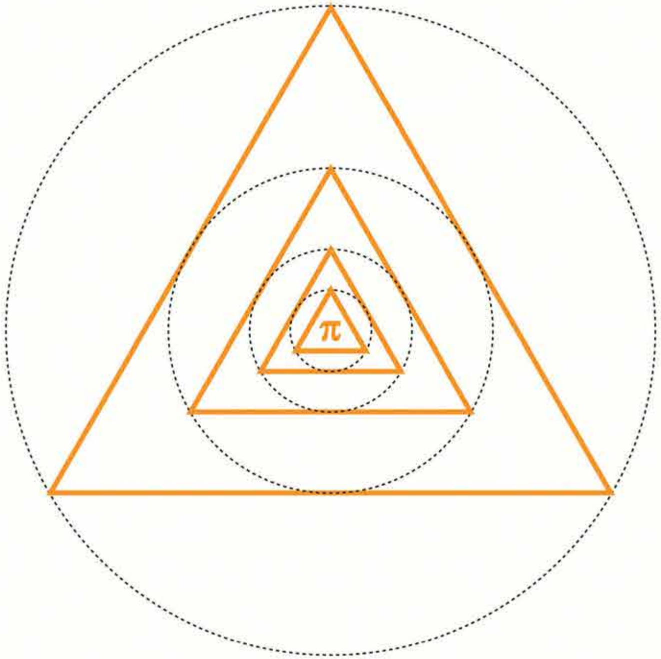

An equilateral triangle is the most symmetrical triangle,
having 3 lines of reflection and rotational symmetry of order 3 about its center

Copyright ABRAHAM [2008] - All rights reserved

CHARGE

Scribed equilateral geometries

reflect space-time's geometric relationship with charged mass-energy

Energy

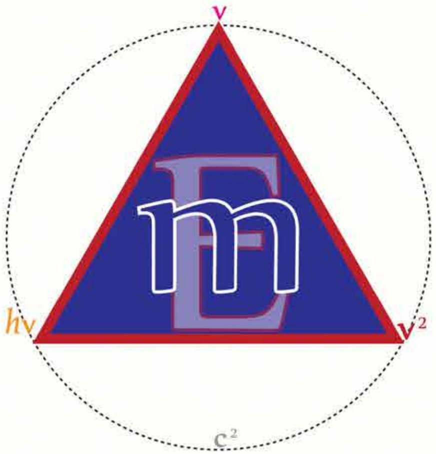

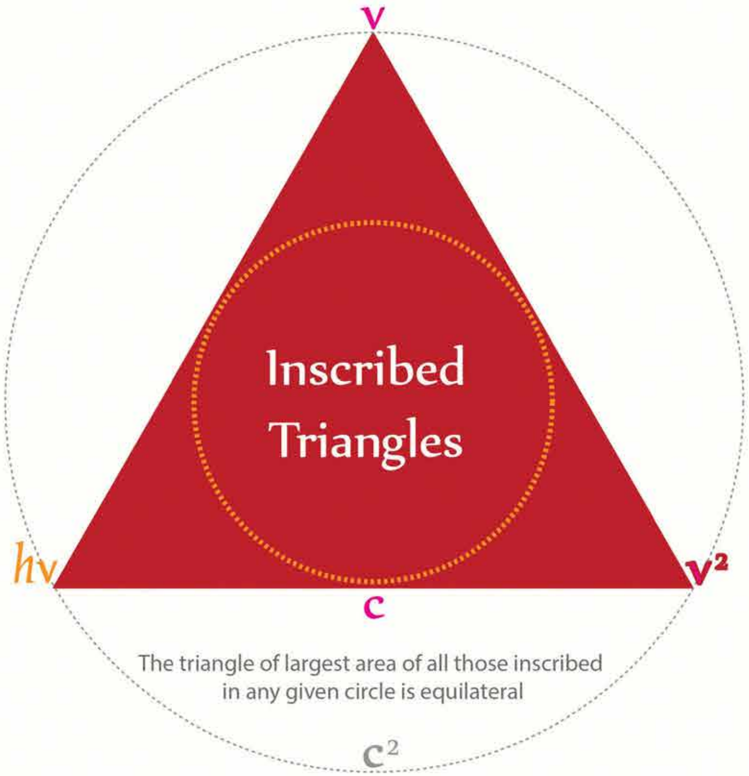
v
Inscribed
Triangles
hv
v²
c
c²

The triangle of largest area of all those inscribed in any given circle is equilateral

QAM

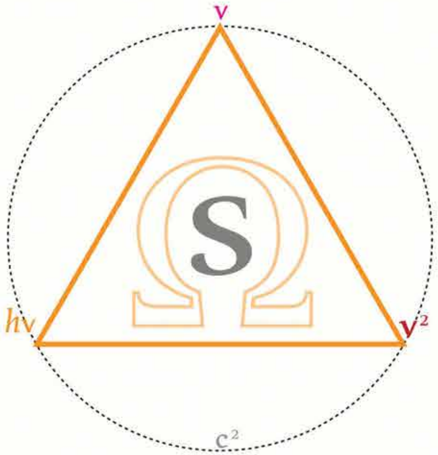

Energy per spatial unit

mass

QAM per spatial unit

seconds

$$m = \frac{E}{c^2}$$

$$\frac{\Omega}{c^2} = s$$

Circumscribed Triangles

reflect Energy's relationship with Time

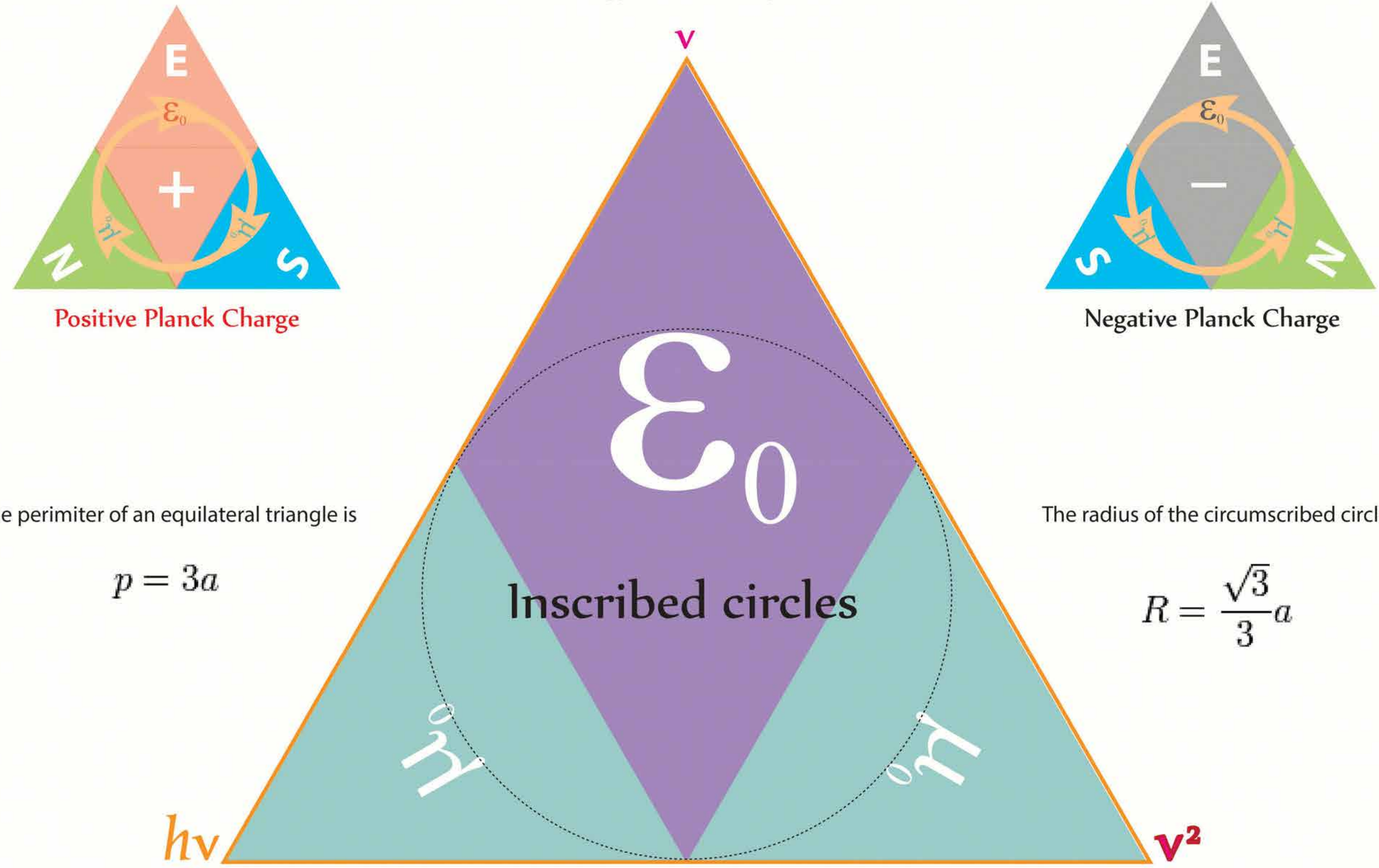
E
ε_0
+
μ_0
μ_0
N
S

Positive Planck Charge

v
ε_0
Inscribed circles
μ_0
μ_0
hv
v²

E
ε_0
–
μ_0
μ_0
S
N

Negative Planck Charge

The perimiter of an equilateral triangle is

$$p = 3a$$

The radius of the circumscribed circle is

$$R = \frac{\sqrt{3}}{3}a$$

The equilateral triangle has the smallest area of all those circumscribed around a given circle

Copyright ABRAHAM [2008] - All rights reserved

Circumscribed circles

By Euler's inequality, the equilateral triangle has the smallest ratio R/r of the circumradius to the inradius of any triangle: specifically, R/r = 2

The ratio of the area to the square of the perimeter of an equilateral triangle

$$\frac{1}{12\sqrt{3}}$$

The ratio of the area of the incircle to the area of an equilateral triangle

$$\frac{\pi}{3\sqrt{3}}$$

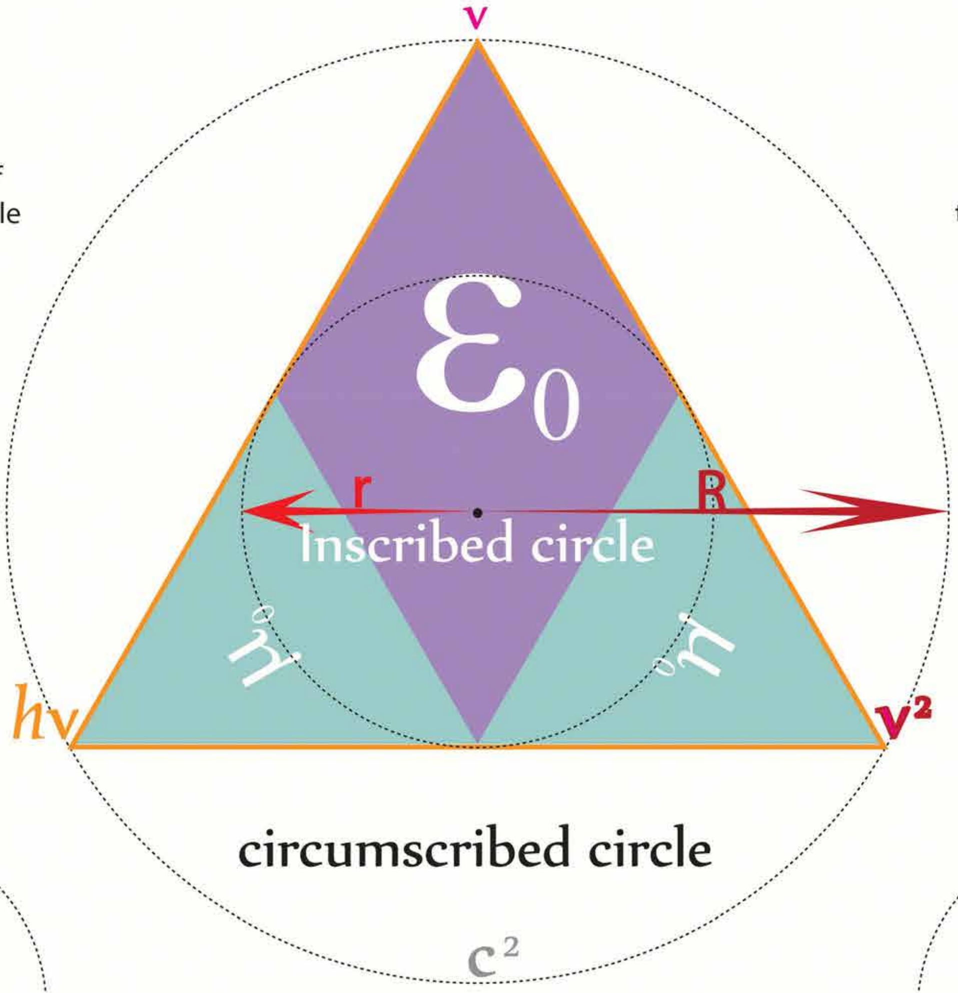
v
ε_0
r
R
Inscribed circle
μ_0
μ_0
hv
v²
circumscribed circle
c^2

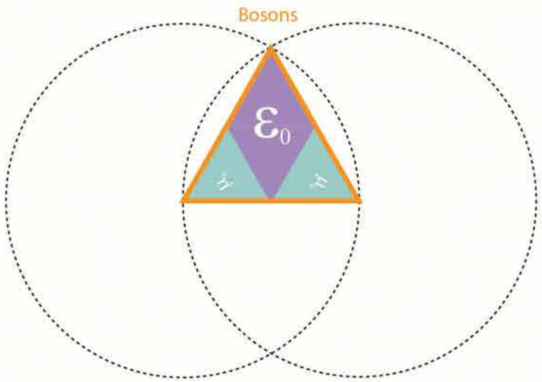
Bosons
ε_0

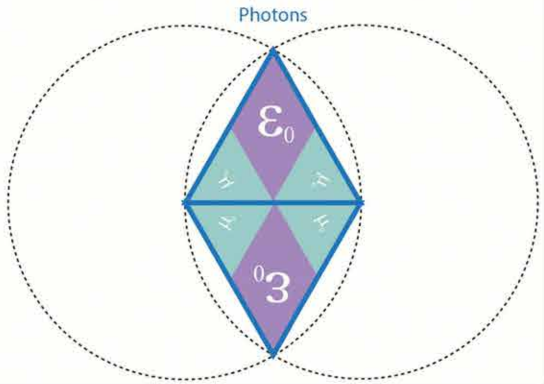
Photons
ε_0

Equilateral triangles and Tetrahedrons will scale at exactly the same proportion as Circles and Spheres scribing them

Charged mass-ENERGIES

Charge is the result of quantised angular momenta

[the inscribed circular flux of energy in equilateral geometries]

Time is a measure of changing quantised angular momenta

[the circumscribed spatial co-ordinate of equilateral energy geometries]

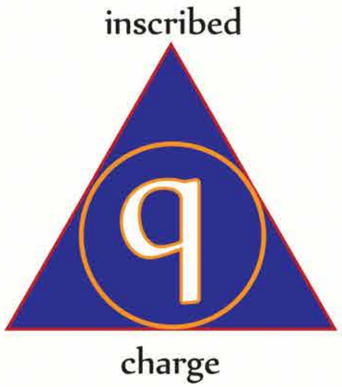

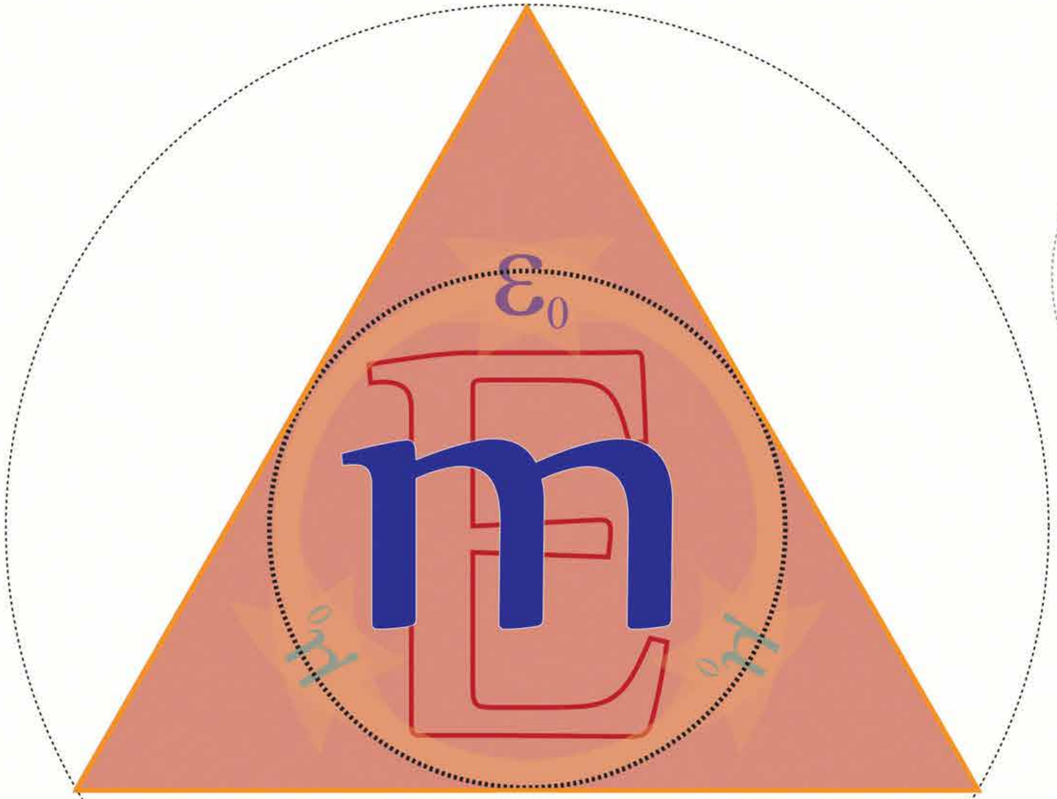

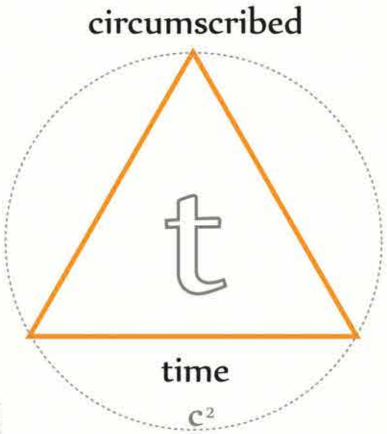

It is the equilateral geometry of energy not a classical vector rotation that creates QAM

Scalar EM mass is a measure of equilateral Planck energy per spatial co-ordinate system

m

Charged mass-energies can be described as a vector measure of scalar Energy momenta per unit of time

[inscribed circular fluxes of equilateral energy per circumscribed temporal geometry]

E

Trigometric functions

$$\sin\,\theta = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos\,\theta = \frac{\text{adjacent}}{\text{hypotenuse}}$$

$$\tan\,\theta = \frac{\text{opposite}}{\text{adjacent}}$$

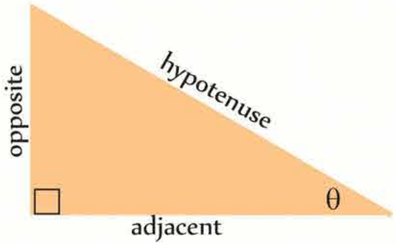
opposite
hypotenuse
θ
adjacent

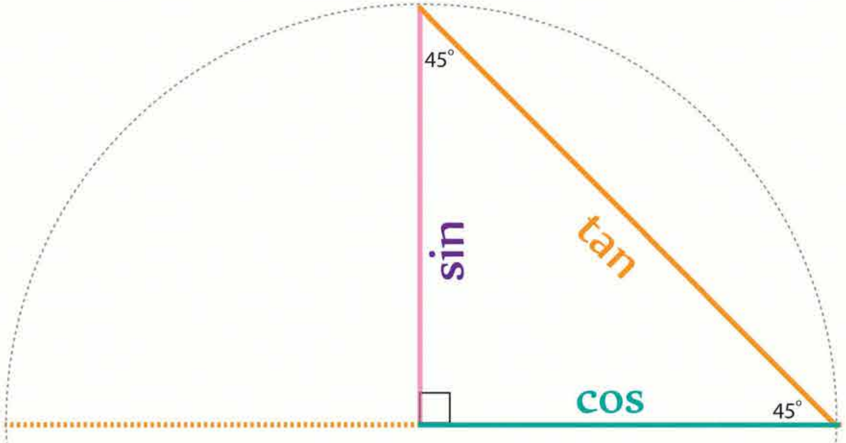
45°
sin
tan
cos
45°

The most familiar trigonometric functions are the sine, cosine, and tangent.

In the context of the standard unit circle with radius 1,
where a triangle is formed by a ray originating at the origin and
making some angle with the x-axis,

the SINE of the angle gives the length of the y-component (rise) of the triangle,
the COSINE gives the length of the x-component (run), and
the TANGENT function gives the slope
(y-component divided by the x-component)

Standard trigometric functions must be carefully applied to
measurements of equilateral Planck mass-energy geometries
in scribed circular space-time co-ordinate systems

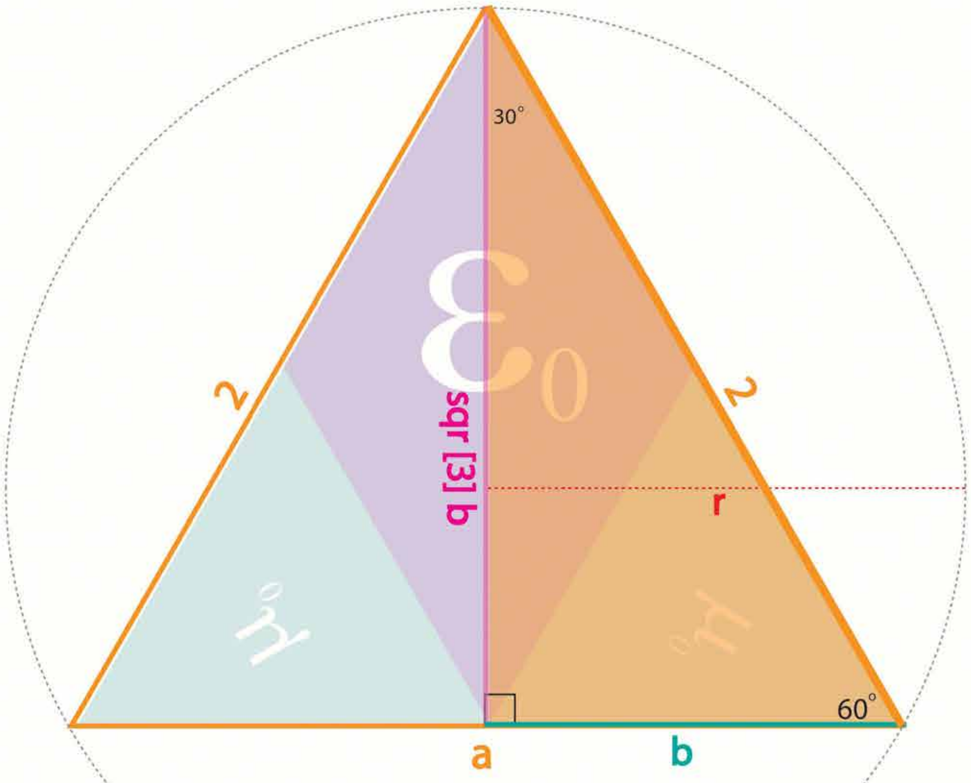
30°
ε0
sqr [3] b
2
2
r
μ0
μ0
60°
a
b

In most cases it is in fact easier to
ignore the trigonomteric functions and math
and simpy draw their related equilateral geometries

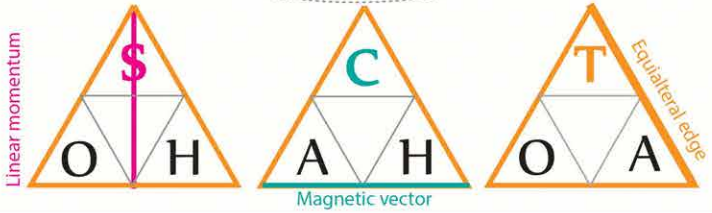
Linear momentum
S
O
H
C
A
H
Magnetic vector
T
O
A
Equilateral edge

Copyright ABRAHAM [2008] - All rights reserved

The roots of scribed equilateral triangles

Scalar equilateral energies map directly onto circular space-time co-ordinates
through their square root linear momentum

r

The ratio of the
circumscribed circle of
an equilateral triangle
to its inscribed circle
is 2:1

d

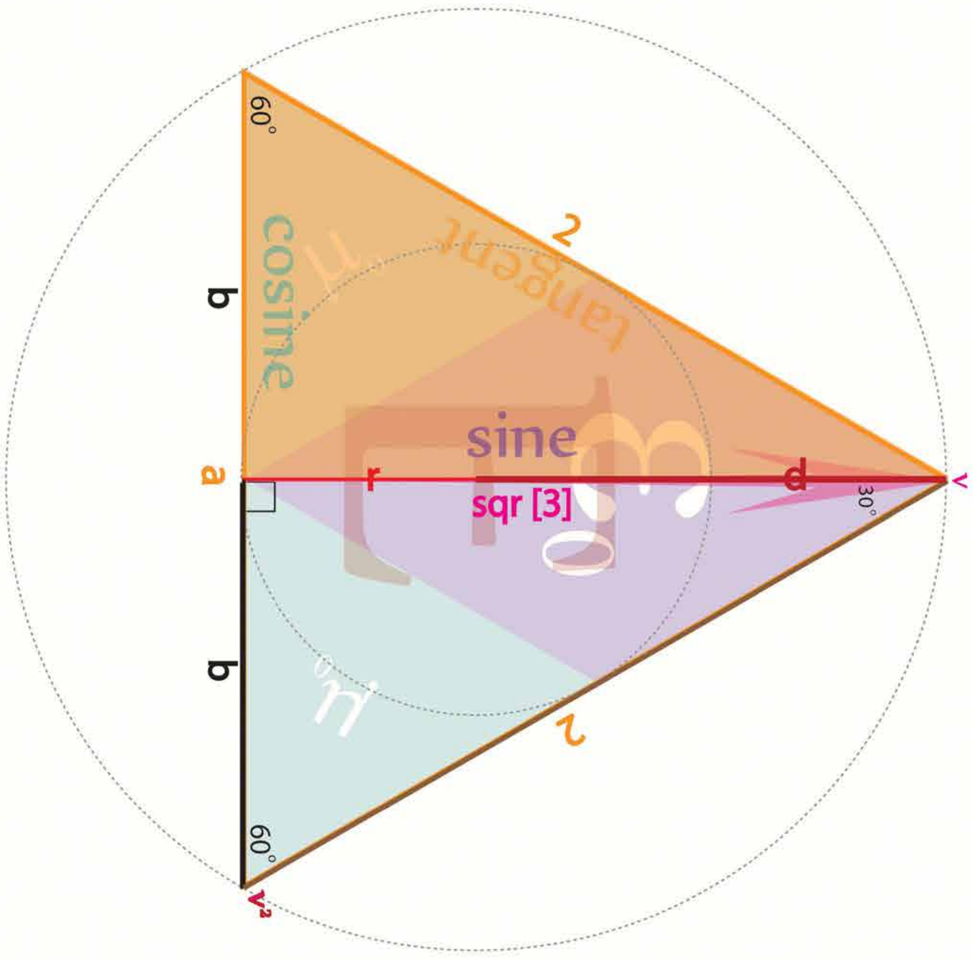

60°
cosine
tangent
2
b
sine
a
r
d
sqr [3]
30°
v
b
2
60°
v²

0.866 r

The ratio of the
side of an equilateral triangle
to the radius of
its incribed circle
is sqr [3]/2

a

The ratio of the
side of an equilateral triangle
to the radius of
its circumscribed circle
is sqr [3]

1.732 d

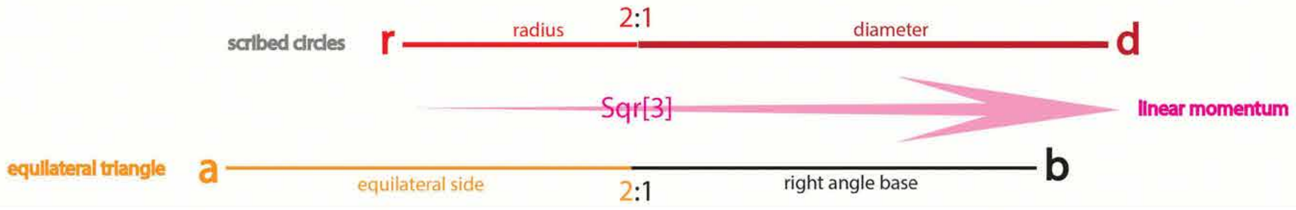

scribed circles
r
radius
2:1
diameter
d
Sqr[3]
linear momentum
equilateral triangle
a
equilateral side
2:1
right angle base
b

Equilateral triangles and scribed circles

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = \frac{1}{\sqrt{1 - \beta^2}} = \frac{dt}{d\tau}$$

$$\gamma(\beta) = 1 + \frac{1}{2}\beta^2 + \frac{3}{8}\beta^4 + \frac{5}{16}\beta^6 + \frac{35}{128}\beta^8 + \ldots$$

$$L = L'\sqrt{1 - \frac{v^2}{c^2}}$$

$$E_k = \frac{mc^2}{\sqrt{1 - (v/c)^2}} - mc^2$$

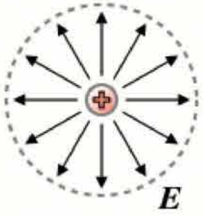

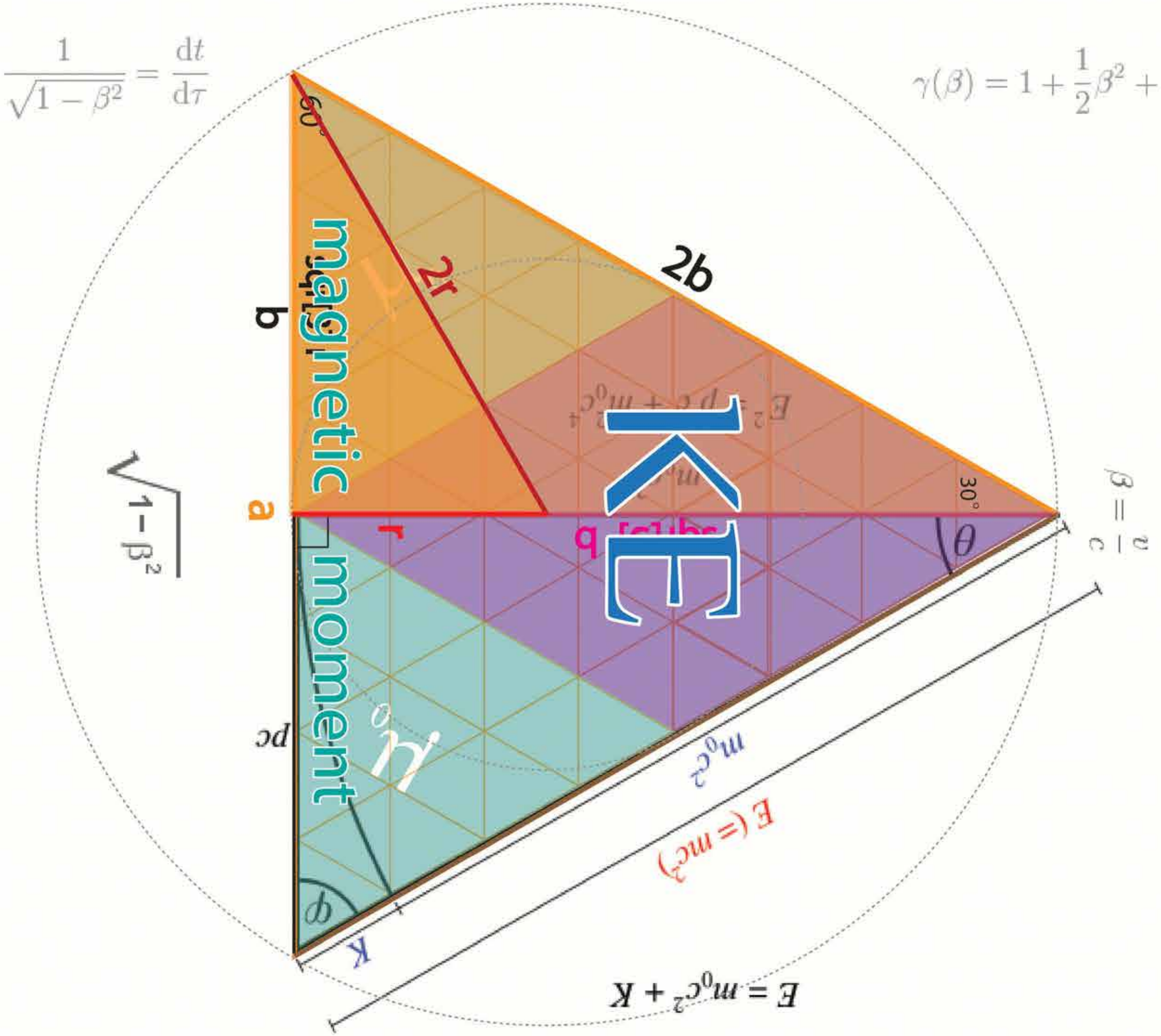
60°
2r
2b
b
magnetic moment
μ
a
r
q
KE
30°
θ
$\sqrt{1-\beta^2}$
$\beta = \frac{v}{c}$
$E^2 = p^2c^2 + m_0^2c^4$
pc
μ_0
m_0c^2
$E\ (= mc^2)$
K
φ
$E = m_0c^2 + K$

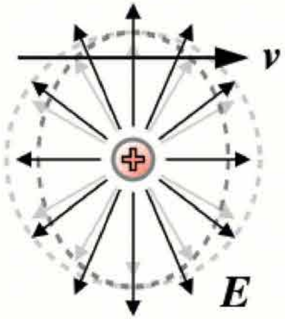

$$mc^2\left\{\frac{1}{\sqrt{1 - v^2/c^2}} - 1\right\}$$

$$p = \frac{m_0 v}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma m_0 v$$

$$E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$$

All the relativistic relationships historically attributed to circularised energies and modelled through the use of the math of right angled triangles are in fact the result of equilateral, scalar geometries

$$\sqrt{1 - \left(\frac{v}{c}\right)^2}$$

Finite sequences and series have defined first and last terms,

$$\sum a_n = a_0 + a_1 + a_2$$

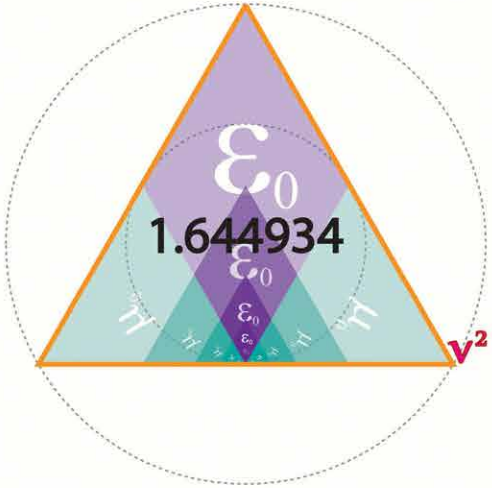

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \lim_{n \to +\infty} \left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \right).$$

whereas infinite sequences and series continue indefinitely

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots = \frac{\pi^2}{6}$$

Tetryonics now provides a geometric solution to visualising and solving the Basel problem

$$\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6} \approx 1.645;$$

Tetryonic Infinite Series

is a infinite sequence of square numbers,
the result of adding all those terms together
[or their geometric inverse]

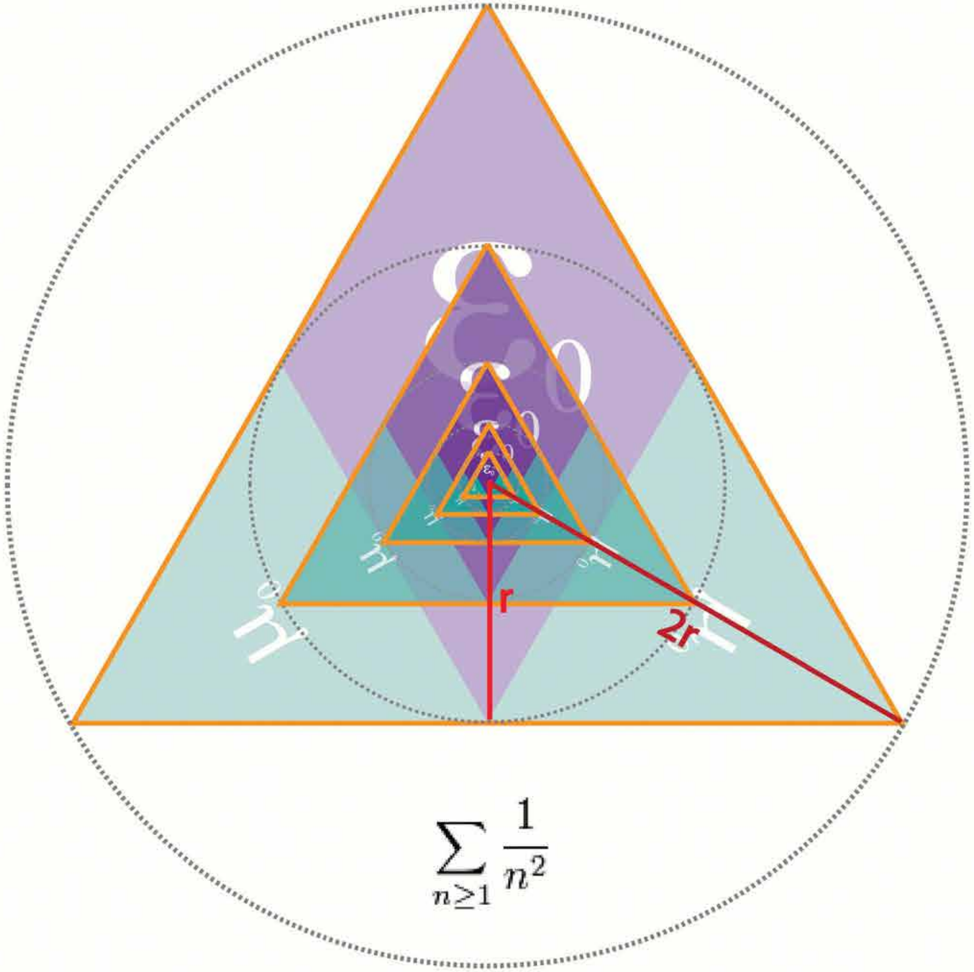

$$\sum_{n \geq 1} \frac{1}{n^2}$$

The entire sum of the series is equal to
twice the size of the radius of the largest inscribed circle
which is equal to the largest circle circumscribing the triangular series.

Copyright ABRAHAM [2008] - All rights reserved

Inverting the Circle

Electric Permittivity

ε_o

Magnetic Permeability

μ_o

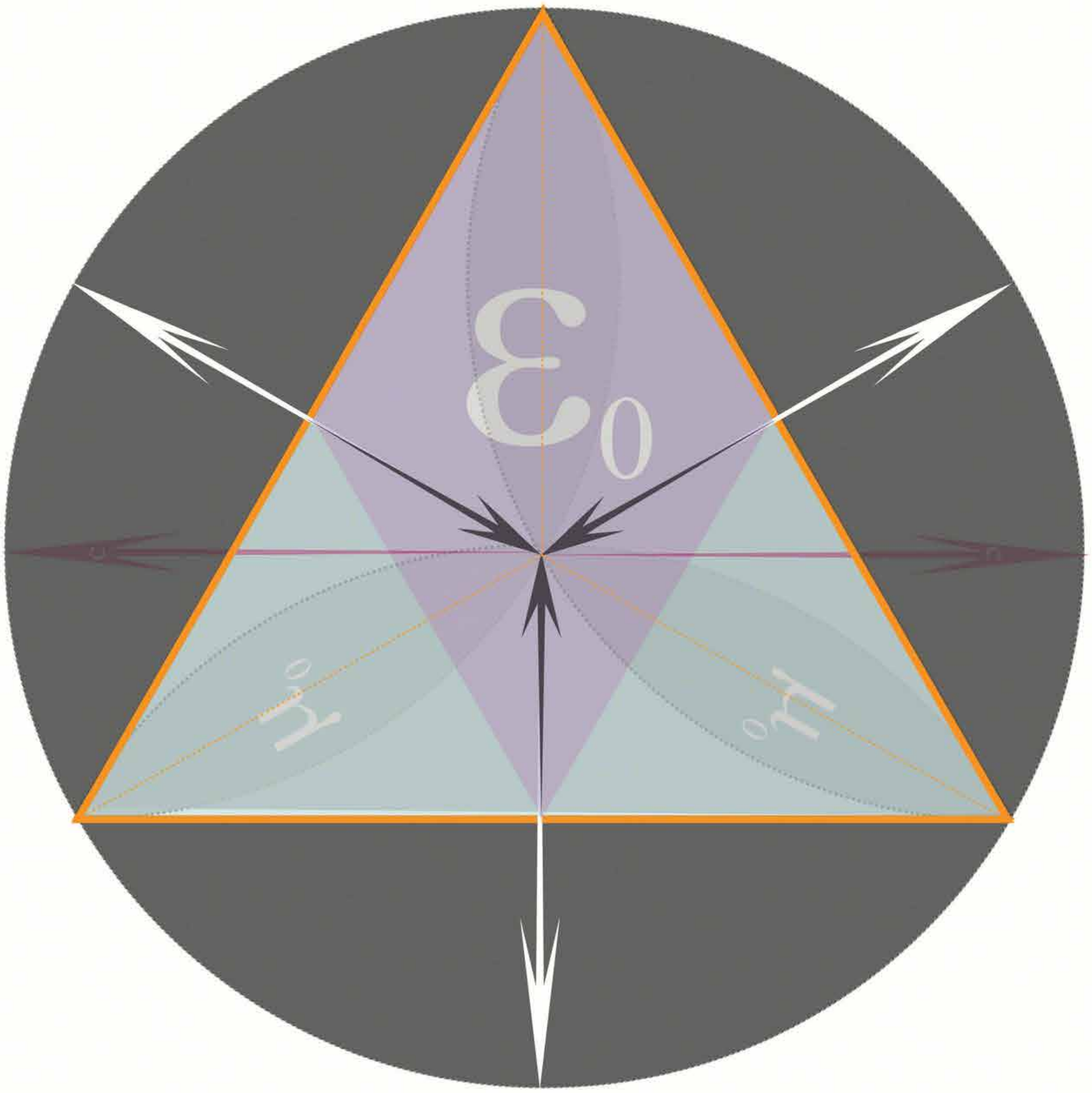

c^2

2D spatial co-ordinates

$\frac{1}{c^2}$

ElectroMagnetic fields

Copyright ABRAHAM [2008] - All rights reserved

Irrational numbers

an irrational number cannot be represented as a simple fraction.
Irrational numbers are those real numbers that cannot be represented as terminating or repeating decimals

sin(x)=-square root of (3)/2

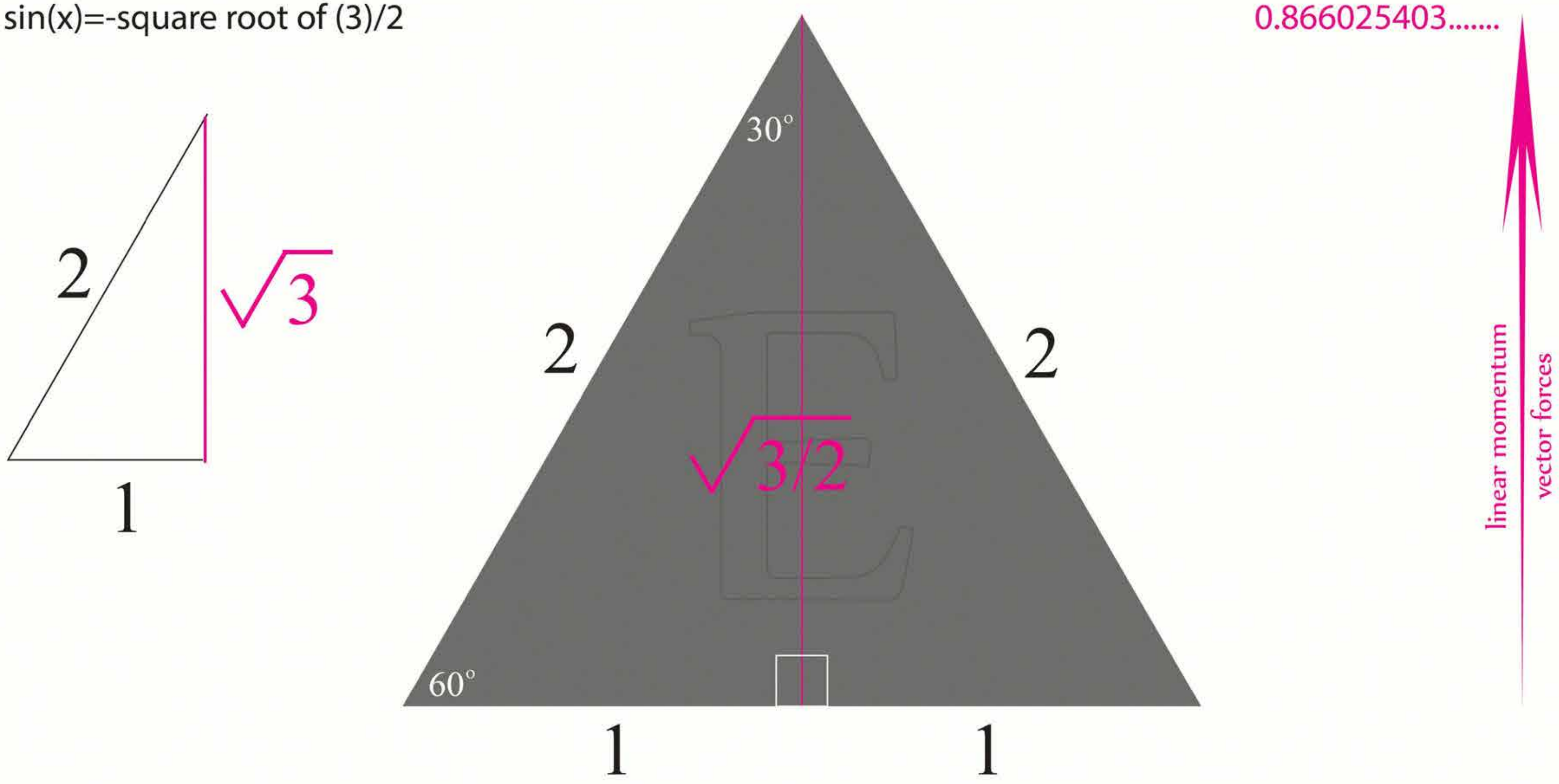
0.866025403.......
2
$\sqrt{3}$
1
30°
2
2
$\sqrt{3/2}$
60°
1
1
linear momentum
vector forces

Pythagφras' theoreom
and irratiφnal numbers expressed in terms of right-angled triangles in Physics
offer a 'half truth' regarding the equilateral geometry of Energy

Leibniz

Leibniz introduced the concept of vis viva (living force) which attempted to quantify life force

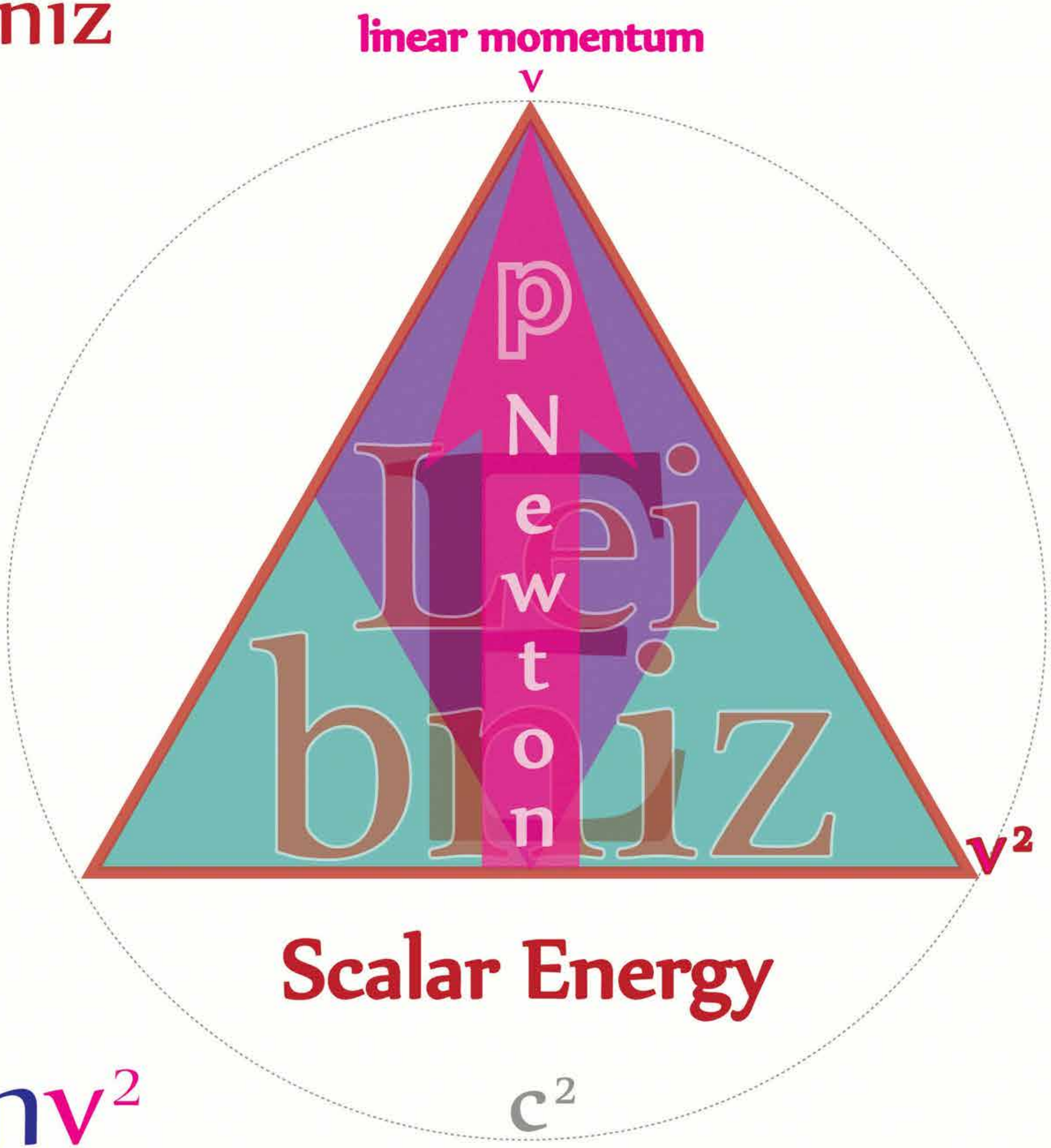

Newton

Newton focused his work on linear momentum which he developed into his famous laws of motion

Newton and Leibniz disagreed about what the world is made of and how its physics shaped our scientific concepts of force, energy, and momentum

Tetryonics reveals the physical relationships they both described mathematically as geometric properties of equilateral Planck energy momenta

$$E=mv^2$$

$$p=mv$$

Copyright ABRAHAM [2008] - All rights reserved

Geometric Square Roots

In geometrical terms, the square root function maps the area of a square to its side length.

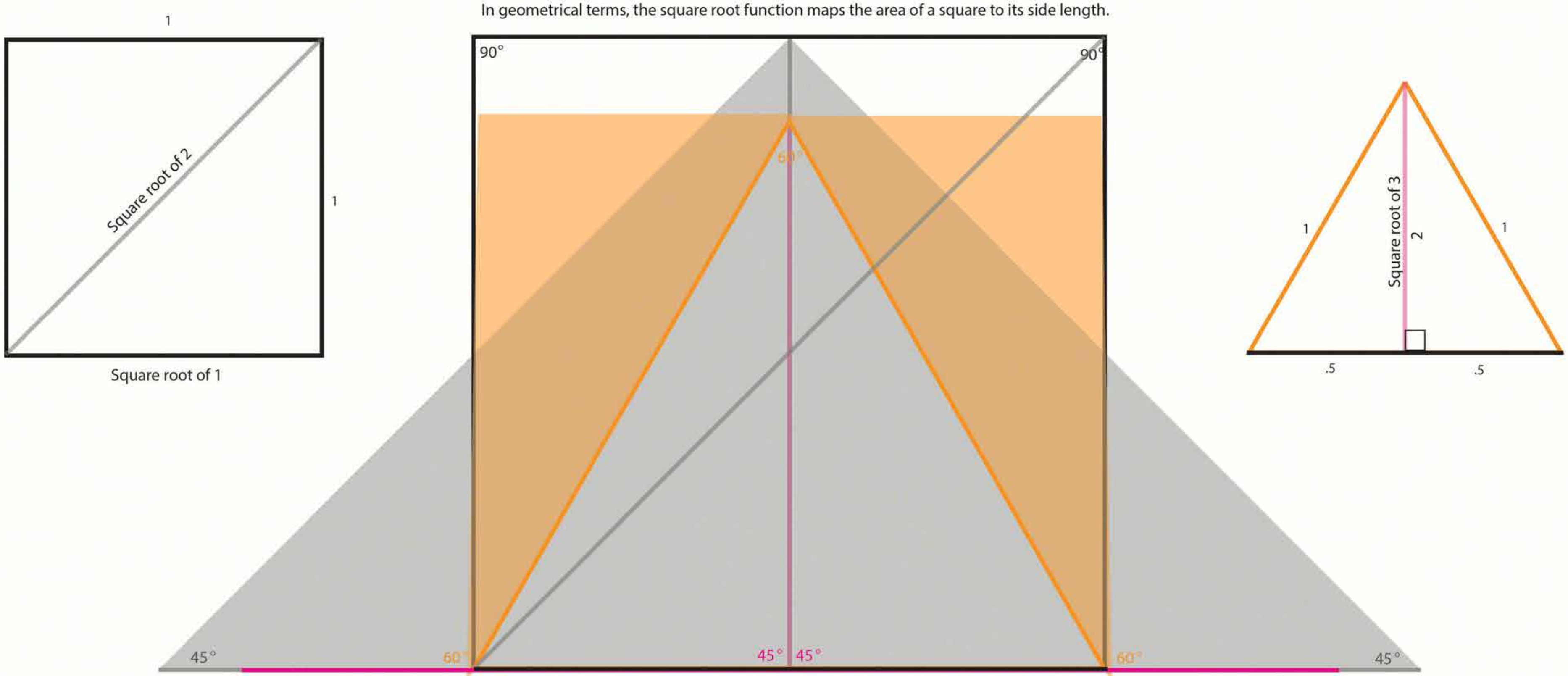

Square root of 1

"In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward abstraction seeks to crystallize the logical relations inherent in the maze of material that is being studied, and to correlate the material in a systematic and orderly manner. On the other hand, the tendency toward intuitive understanding fosters a more immediate grasp of the objects one studies, a live rapport with them, so to speak, which stresses the concrete meaning of their relations"

Square root of 1

Square root of 2

Square root of 3

"As to geometry, in particular, the abstract tendency has here led to the magnificent systematic theories of Algebraic Geometry, of Riemannian Geometry, and of Topology; these theories make extensive use of abstract reasoning and symbolic calculation in the sense of algebra.

Notwithstanding this, it is still as true today as it ever was that intuitive understanding plays a major role in geometry. And such concrete intuition is of great value not only for the research worker, but also for anyone who wishes to study and appreciate the results of research in geometry"

David Hilbert [Geometry and the Imagination]

Copyright ABRAHAM [2008] - All rights reserved

Square Roots in Physics

In mathematics, a square root of a number a is a number [n] such that [n]2 = x,
or, in other words, a number [n] whose square (the result of multiplying the number by itself, or [n × n]) is x.

Modern calculators use the
Square Root Algorithm
to calculate the value

It is an approximate numerical value

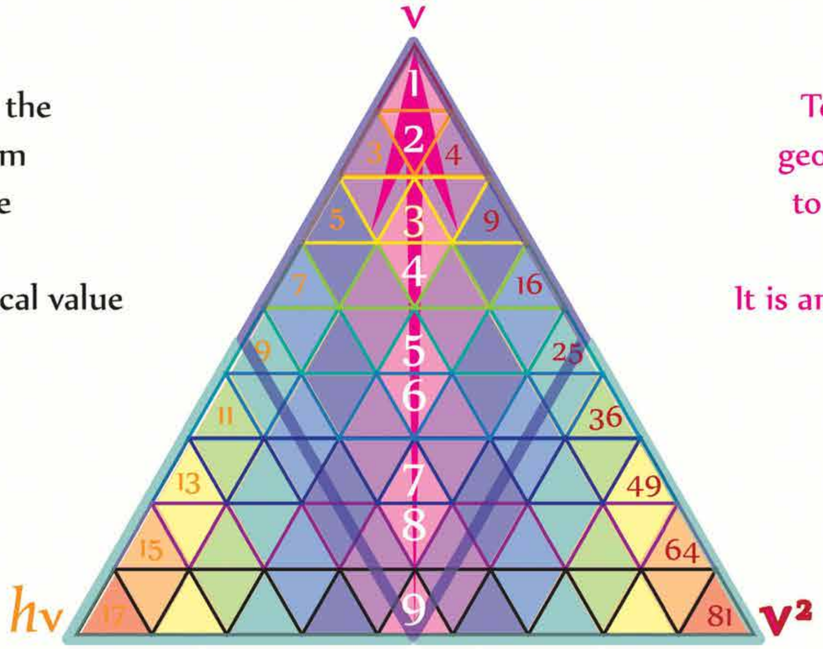

Tetryonics uses the
geometric Square Root
to calculate the value

It is an exact geometric value

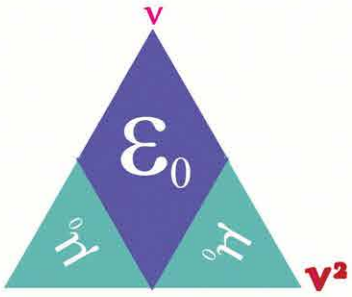

$$E = mv^2$$

*In classical geometry,
the square root function maps the
area of a square to its side length.*

$$p = \sqrt{E}$$

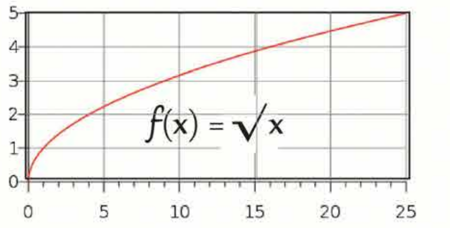

Geometry can easily map irrational numbers

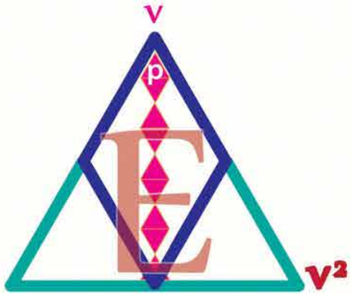

$$p^2 = E$$

*In physics,
the square root function maps
ENERGY [E] to momentum [mv]*

The Square roots of n

Historically, any number raised to the power of 2 has been modeled using a polygon--the square
That's why we call raising a number to the second power "squaring the number."

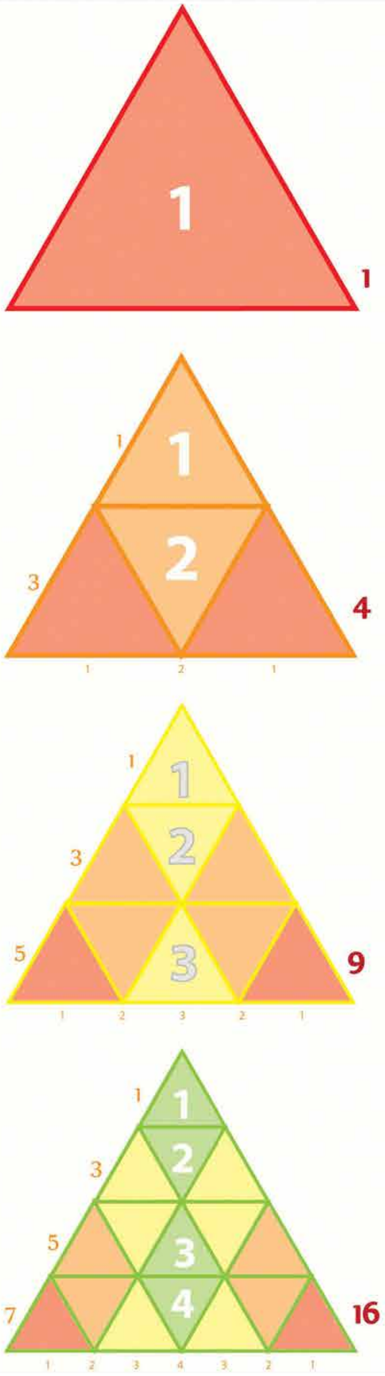

note: scalar linear momentum [mv]

is distinct from vector [v]elocity

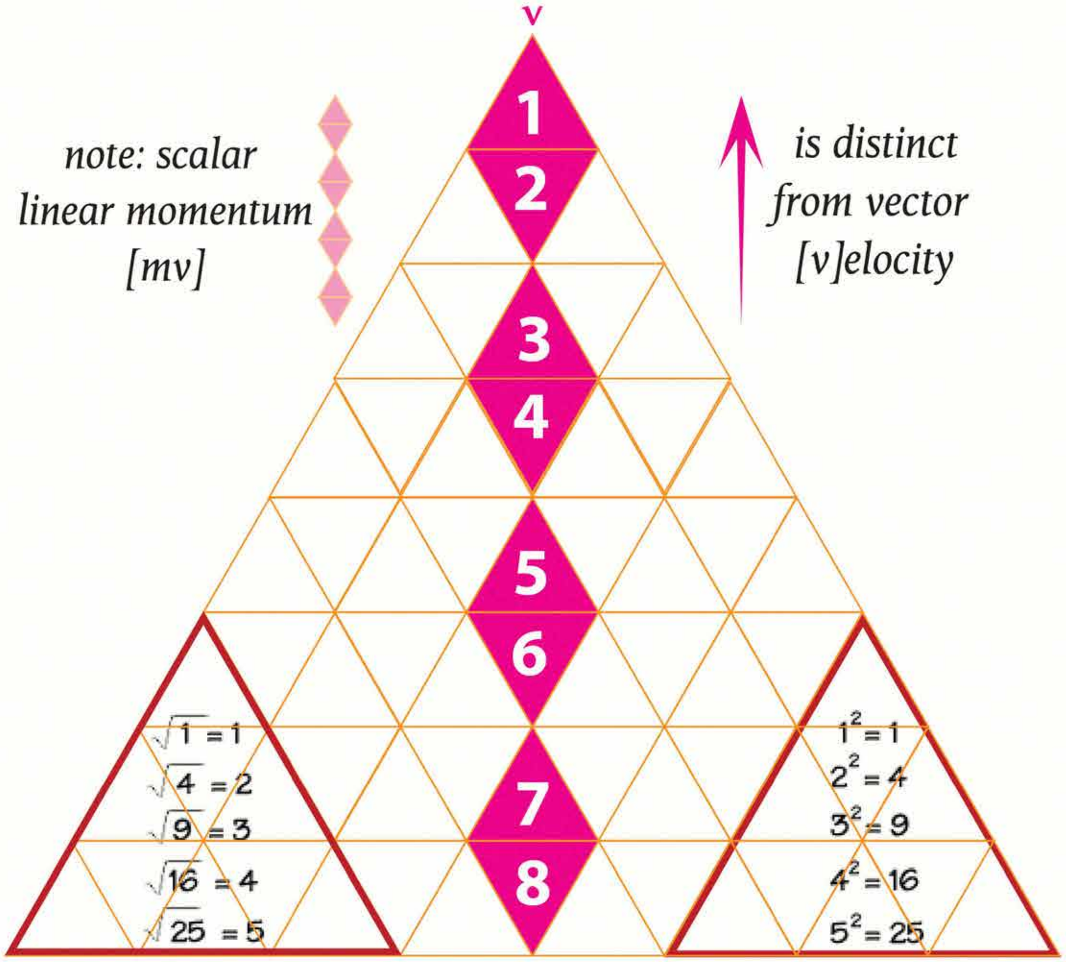

$\sqrt{1} = 1$
$\sqrt{4} = 2$
$\sqrt{9} = 3$
$\sqrt{16} = 4$
$\sqrt{25} = 5$

$1^2 = 1$
$2^2 = 4$
$3^2 = 9$
$4^2 = 16$
$5^2 = 25$

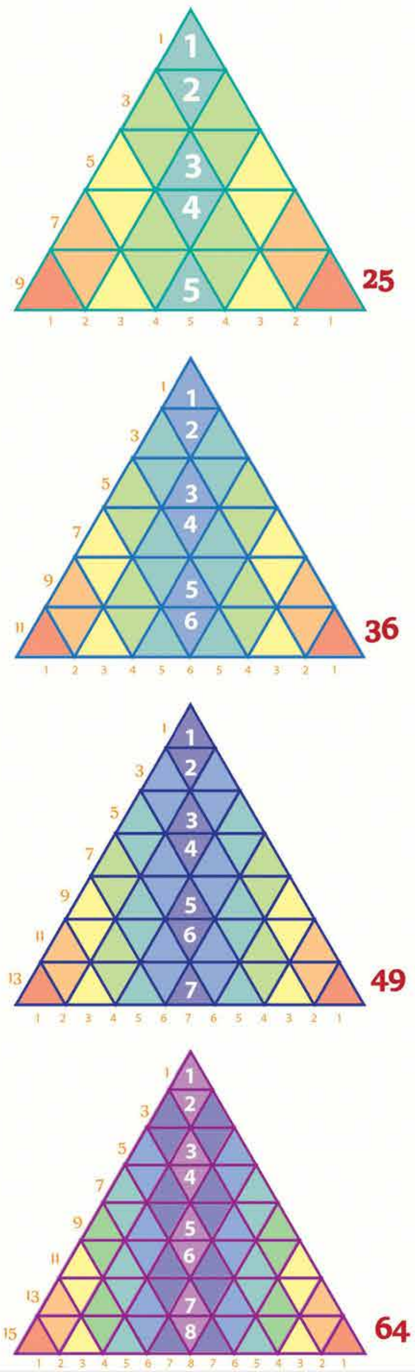

[In physics square numbers are in fact equilateral geometries]
The perfect squares are squares of whole numbers.
Here are the first eight perfect squares

The Square root of Negative 1

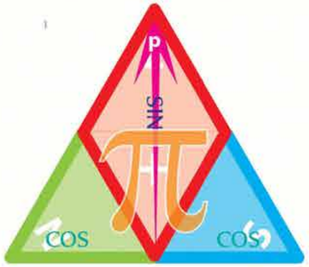

Positive fields are out of phase with Negative fields

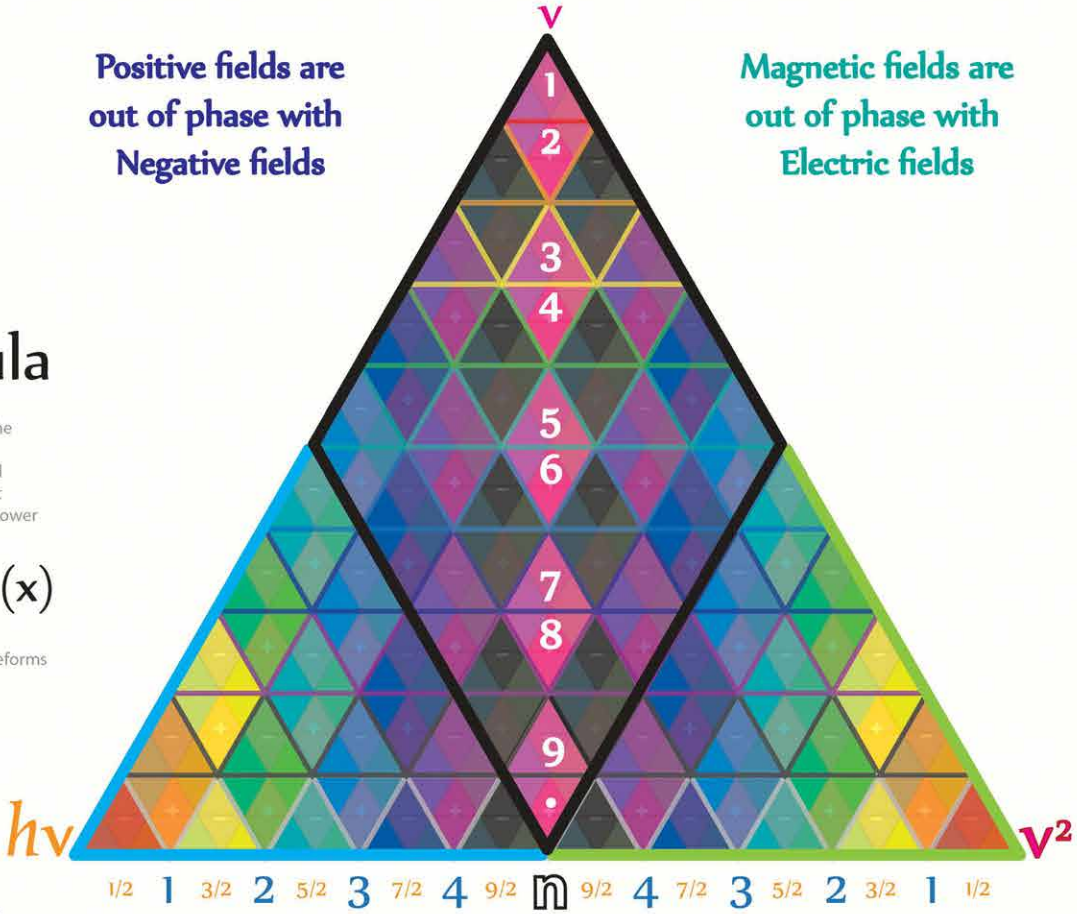

Magnetic fields are out of phase with Electric fields

P
COS
π
SIN
SIN

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system.
In deriving this formula, Euler established a relationship between the trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

Leonhard Euler

(15 April 1707 – 18 September 1783)

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$$

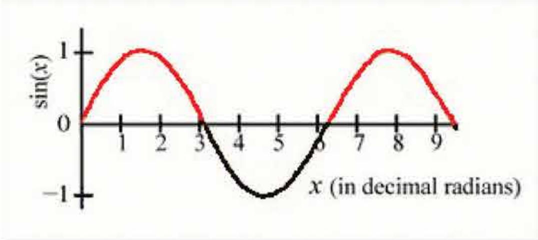

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

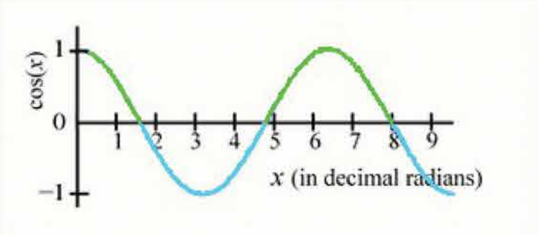

Copyright ABRAHAM [2008] - All rights reserved

Geometric means

The geometric mean of two numbers, is the square root of their product

geometric square root of positive one

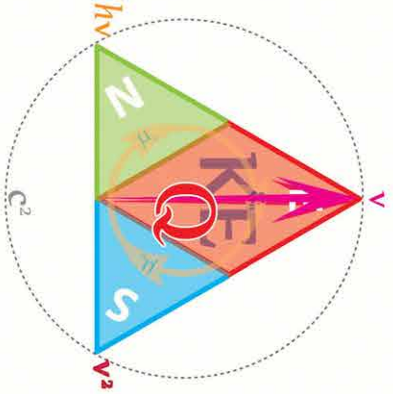

$$F = k\frac{Q_1 Q_2}{r^2}$$

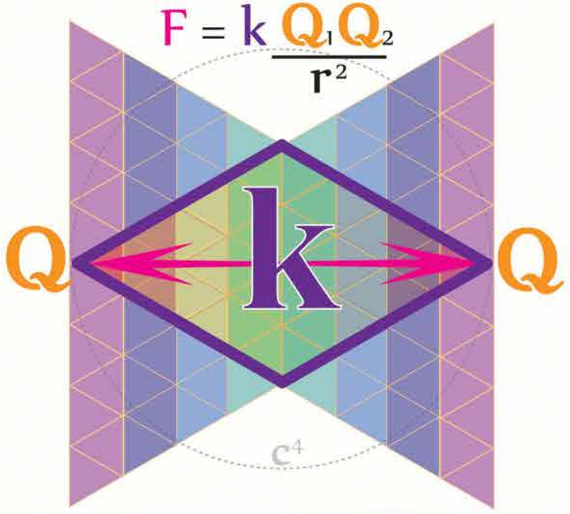

geometric square root of negative one

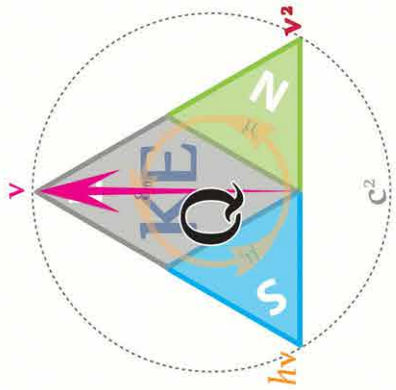

In physics, the geometric mean of two superpositioned fields produces a vector square root Force

It is generally stated that
the geometric mean applies
only to positive numbers.

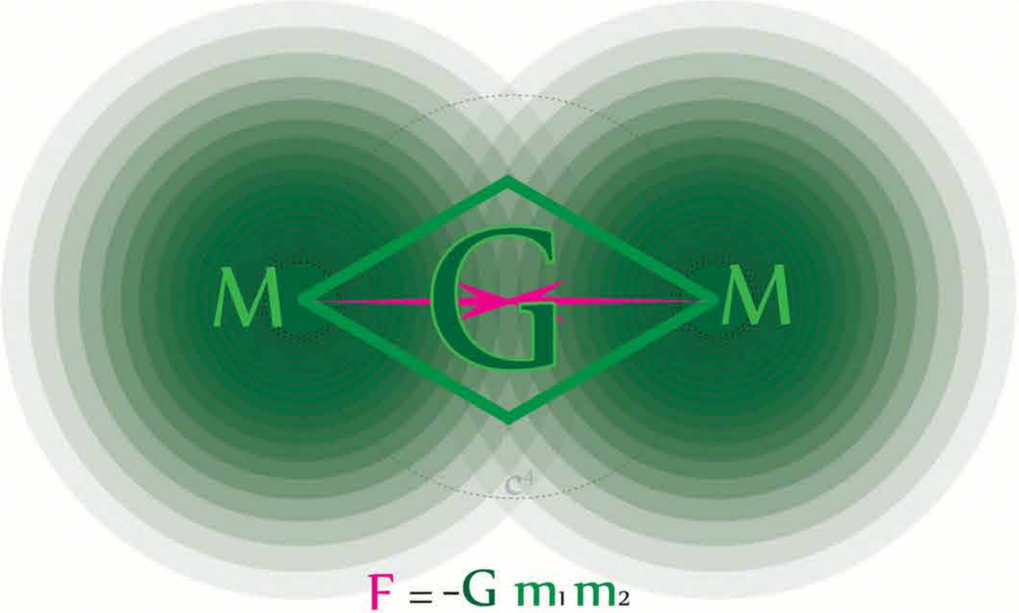

In Tetryonic geometry
the geometric mean applies
to positive & negative fields.

$$F = -G\frac{m_1 m_2}{r^2}$$

Superpositioning

When two or more waves traverse the same space,
the net amplitude at each point is the
sum of the amplitudes of the individual waves.

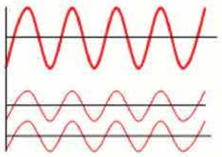

constructive interference

In phase

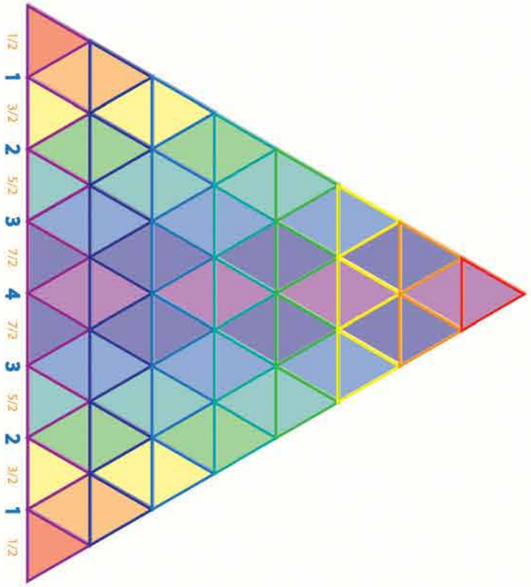
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Additive

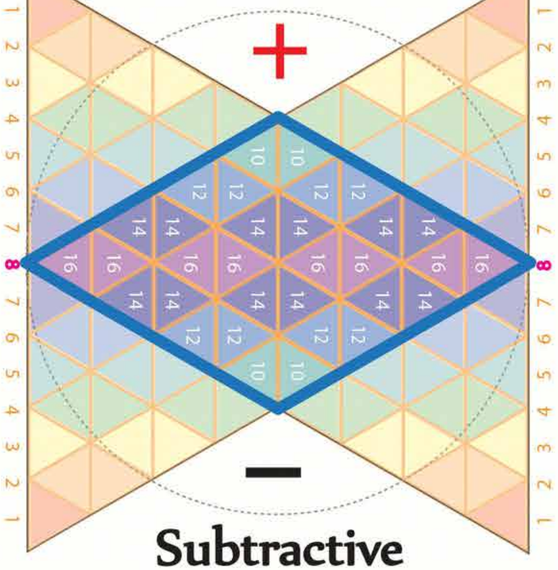
+
−
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
10
12
14
16

Subtractive

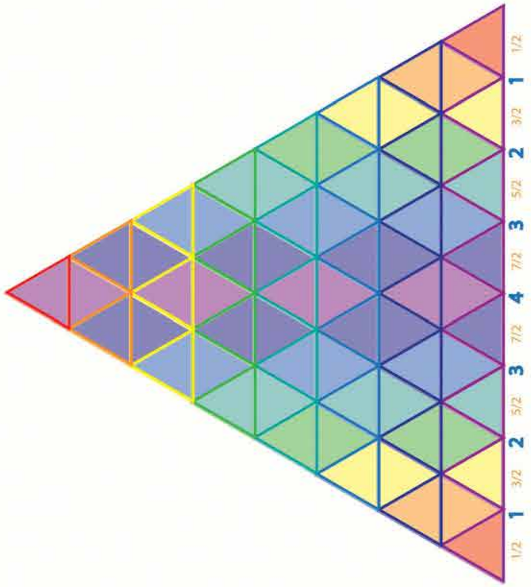
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Out of phase

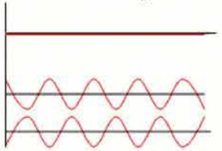

destructive interference

When two or more waves traverse the same space,
if the summed variation has a smaller amplitude
than the constiuent component variations.

The lines of Force

Tetryonic Multiplication table

A multiplication table is a mathematical grid used to define a multiplication operation and its results

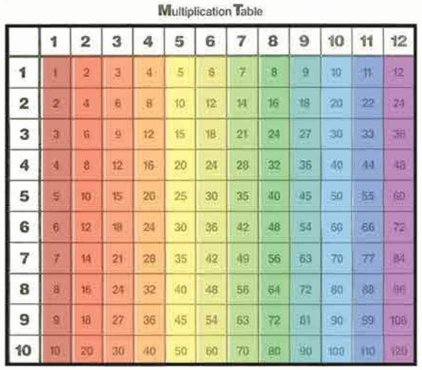

Multiplication Table

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120

Historically Multiplication tables
have been based on Square geometries

In Tetryonics multiplication tables can
also be based on EQUILATERAL geometries

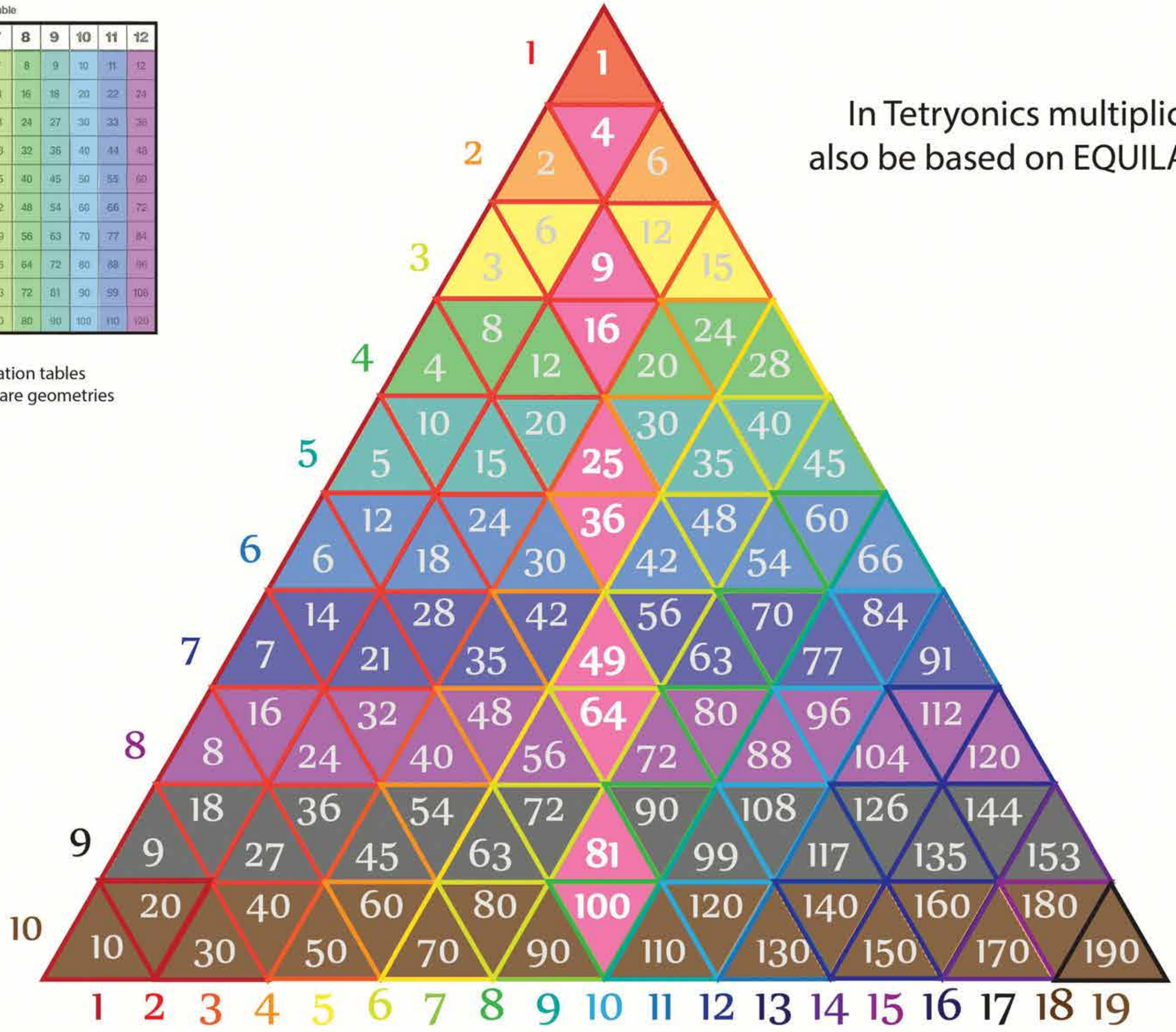

1: 1
2: 2 4 6
3: 3 6 9 12 15
4: 4 8 12 16 20 24 28
5: 5 10 15 20 25 30 35 40 45
6: 6 12 18 24 30 36 42 48 54 60 66
7: 7 14 21 28 35 42 49 56 63 70 77 84 91
8: 8 16 24 32 40 48 56 64 72 80 88 96 104 112 120
9: 9 18 27 36 45 54 63 72 81 90 99 108 117 126 135 144 153
10: 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

The integer multiplicitors are colour coded

Rhombic Multiplication Tables

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

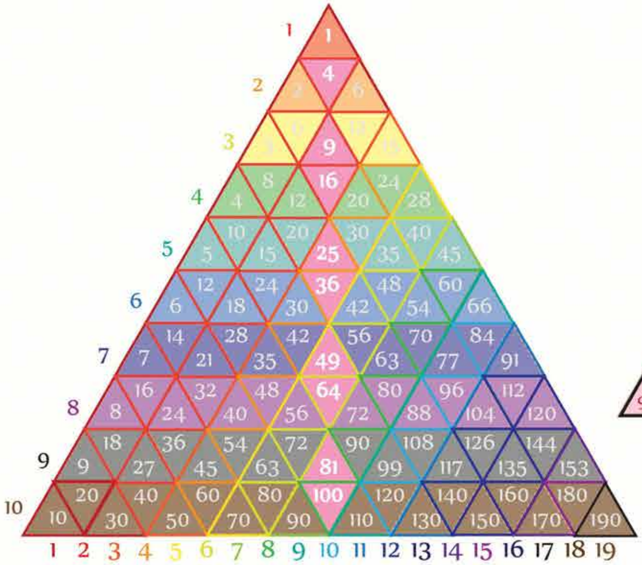
1
2
3
4
5
6
7
8
9
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

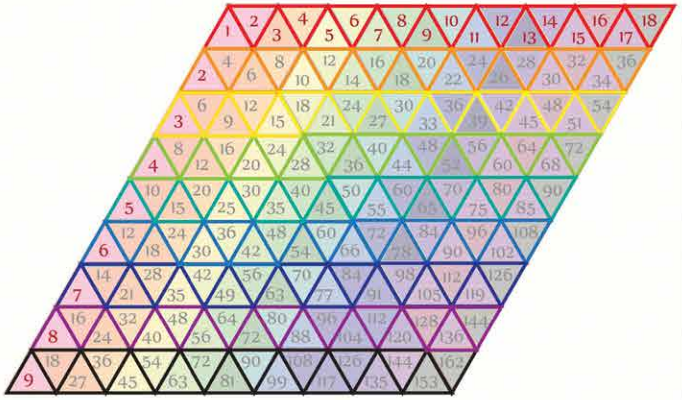

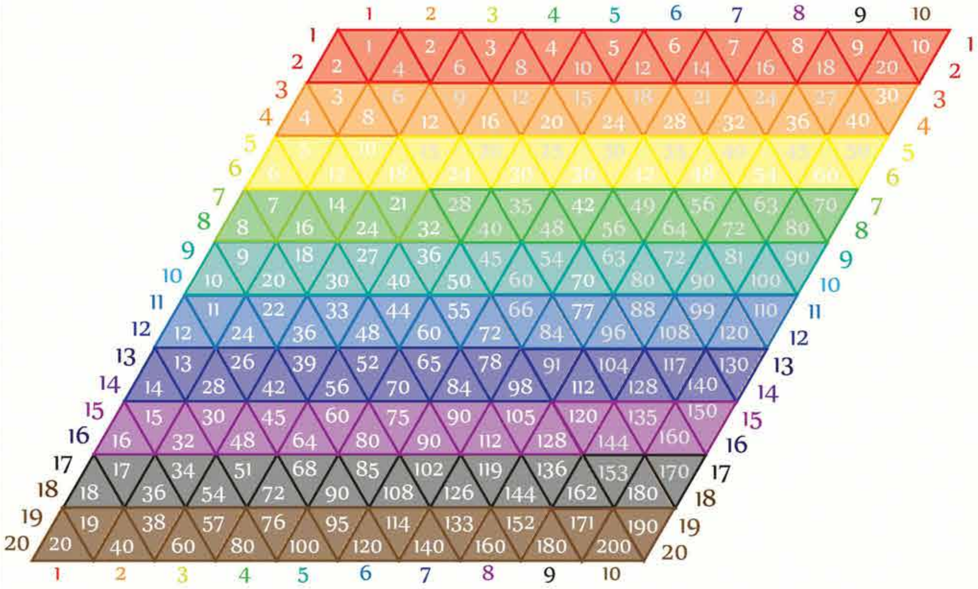
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1 2 3 4 5 6 7 8 9 10

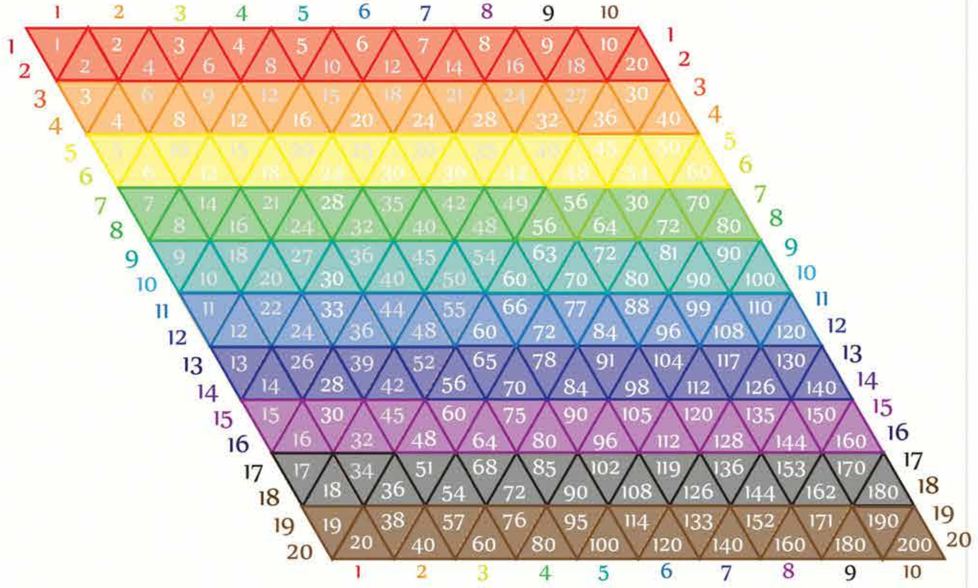
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1 2 3 4 5 6 7 8 9 10

Tetryonics 85.03 - Rhombic multiplication tables

Copyright ABRAHAM [2008] - All rights reserved

Photonic Root Tables

Tetryonic multiplication table can take a number of geometric forms

Square root median

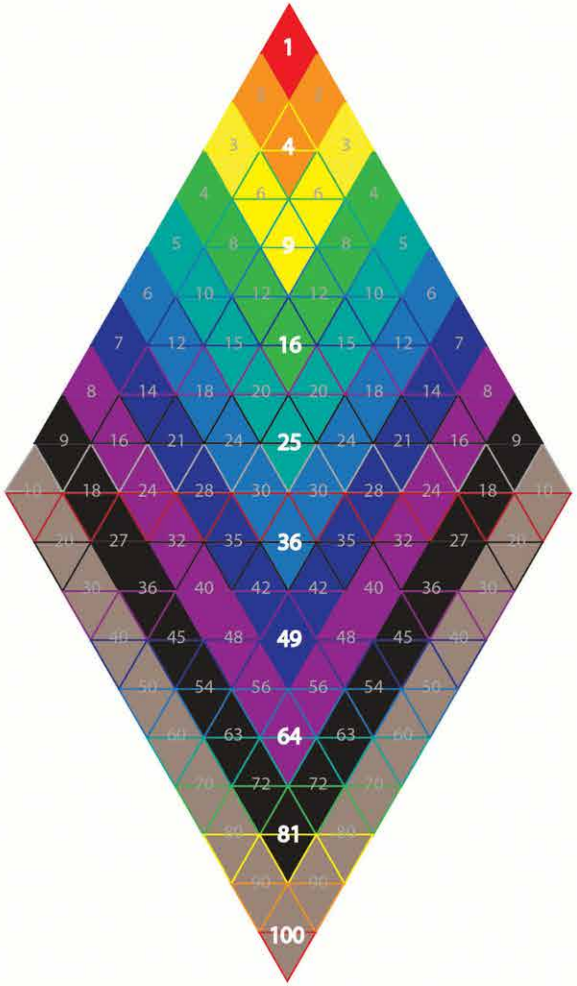

Table read diagonally

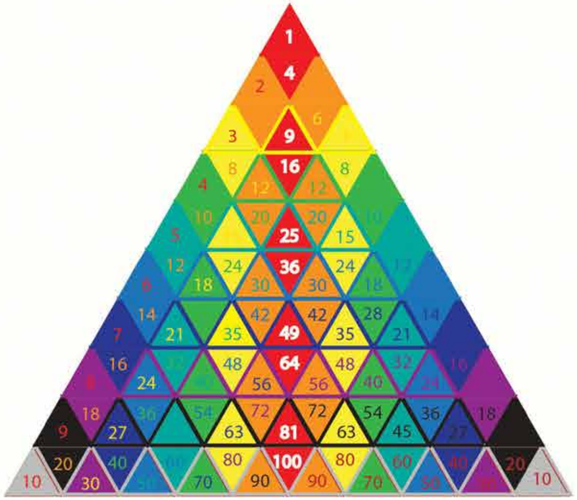

1	2	3	4	5	6	7	8	9	10
2	**4**	6	8	10	12	14	16	18	20
3	6	**9**	12	15	18	21	24	27	30
4	8	12	**16**	20	24	28	32	36	40
5	10	15	20	**25**	30	35	40	45	50
6	12	18	24	30	**36**	42	48	54	60
7	14	21	28	35	42	**49**	56	63	70
8	16	24	32	40	48	56	**64**	72	80
9	18	27	36	45	54	63	72	**81**	90
10	20	30	40	50	60	70	80	90	**100**

Integer median

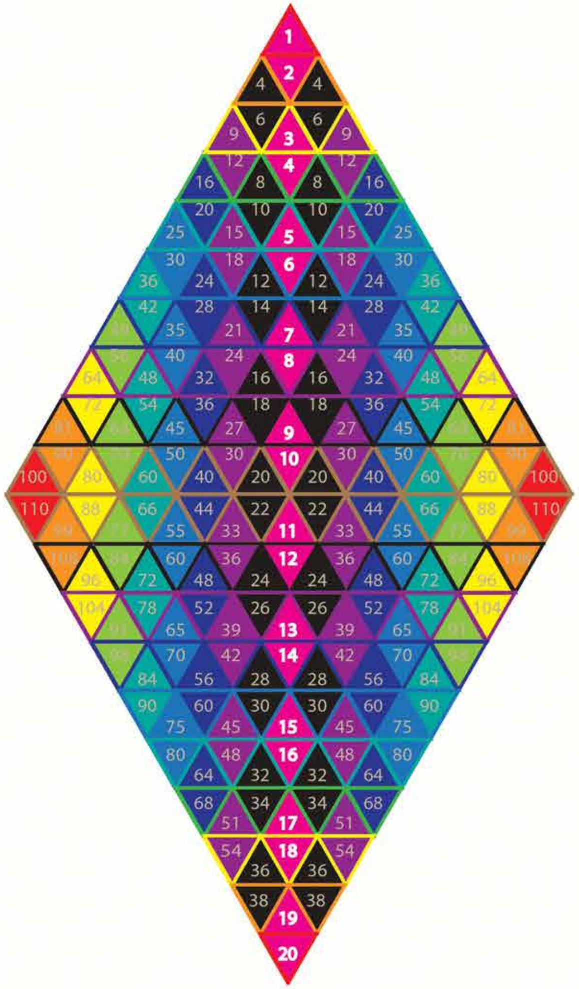

Table read from centre
to outside edge then down

Copyright ABRAHAM [2008] - All rights reserved

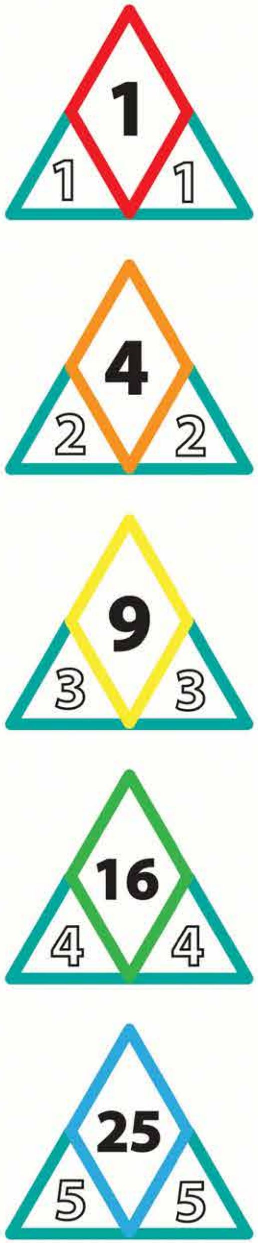
1
1
1
4
2
2
9
3
3
16
4
4
25
5
5

Divide [division] tables

Minus [subtraction] tables

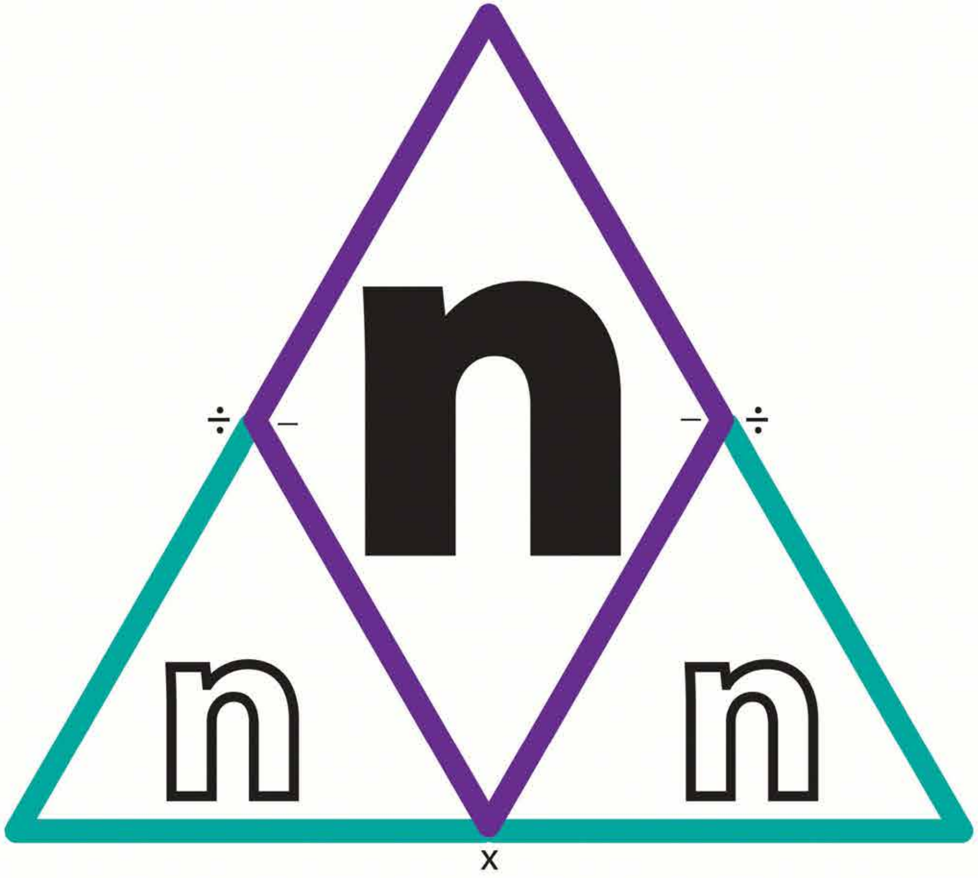

Times [multiplication] tables

Plus [addition] tables

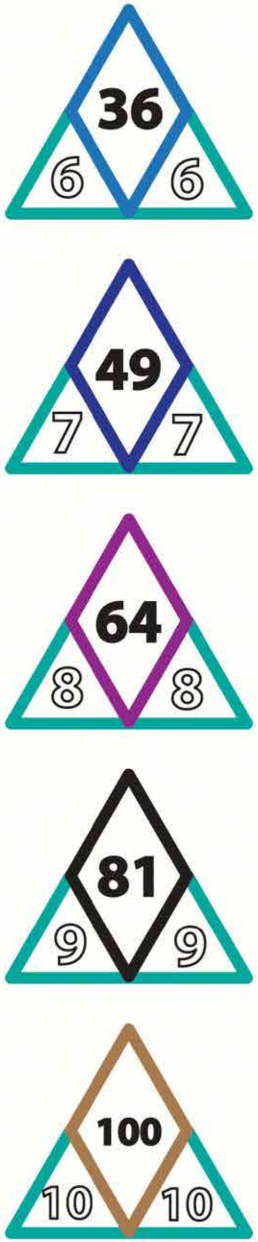
36
6
6
49
7
7
64
8
8
81
9
9
100
10
10

Copyright ABRAHAM [2008] - All rights reserved

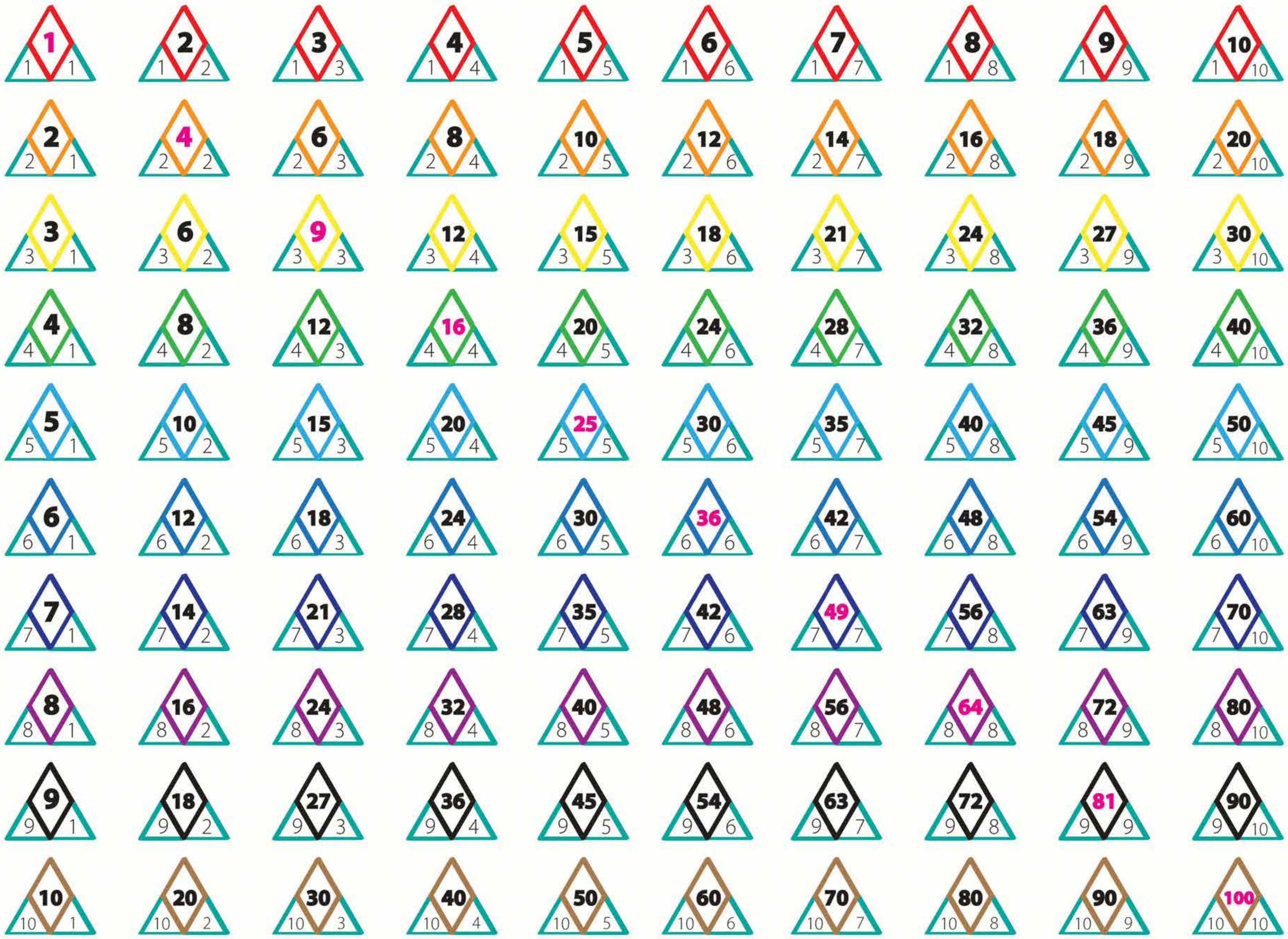

1 1 1	2 1 2	3 1 3	4 1 4	5 1 5	6 1 6	7 1 7	8 1 8	9 1 9	10 1 10
2 2 1	4 2 2	6 2 3	8 2 4	10 2 5	12 2 6	14 2 7	16 2 8	18 2 9	20 2 10
3 3 1	6 3 2	9 3 3	12 3 4	15 3 5	18 3 6	21 3 7	24 3 8	27 3 9	30 3 10
4 4 1	8 4 2	12 4 3	16 4 4	20 4 5	24 4 6	28 4 7	32 4 8	36 4 9	40 4 10
5 5 1	10 5 2	15 5 3	20 5 4	25 5 5	30 5 6	35 5 7	40 5 8	45 5 9	50 5 10
6 6 1	12 6 2	18 6 3	24 6 4	30 6 5	36 6 6	42 6 7	48 6 8	54 6 9	60 6 10
7 7 1	14 7 2	21 7 3	28 7 4	35 7 5	42 7 6	49 7 7	56 7 8	63 7 9	70 7 10
8 8 1	16 8 2	24 8 3	32 8 4	40 8 5	48 8 6	56 8 7	64 8 8	72 8 9	80 8 10
9 9 1	18 9 2	27 9 3	36 9 4	45 9 5	54 9 6	63 9 7	72 9 8	81 9 9	90 9 10
10 10 1	20 10 2	30 10 3	40 10 4	50 10 5	60 10 6	70 10 7	80 10 8	90 10 9	100 10 10

Copyright ABRAHAM [2008] - All rights reserved

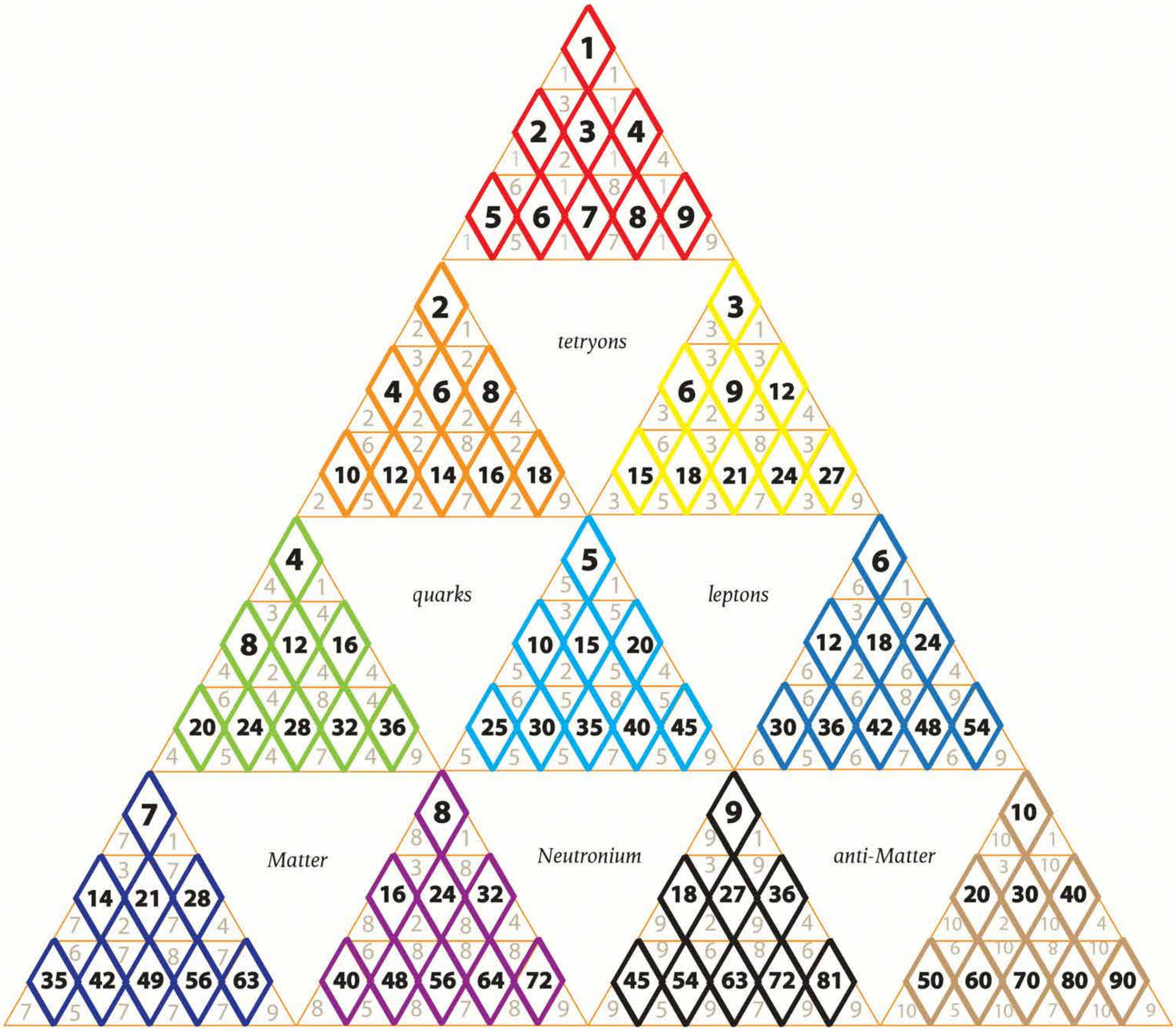
1
2 3 4
5 6 7 8 9
tetryons
2
4 6 8
10 12 14 16 18
3
6 9 12
15 18 21 24 27
quarks
4
8 12 16
20 24 28 32 36
5
10 15 20
25 30 35 40 45
leptons
6
12 18 24
30 36 42 48 54
Matter
7
14 21 28
35 42 49 56 63
8
16 24 32
40 48 56 64 72
Neutronium
9
18 27 36
45 54 63 72 81
anti-Matter
10
20 30 40
50 60 70 80 90

Copyright ABRAHAM [2008] - All rights reserved

Irrational Numbers

An irrational number is defined to be any number that is the part of the real number system that cannot be written as a complete ratio of two integers

One well known irrational number is π

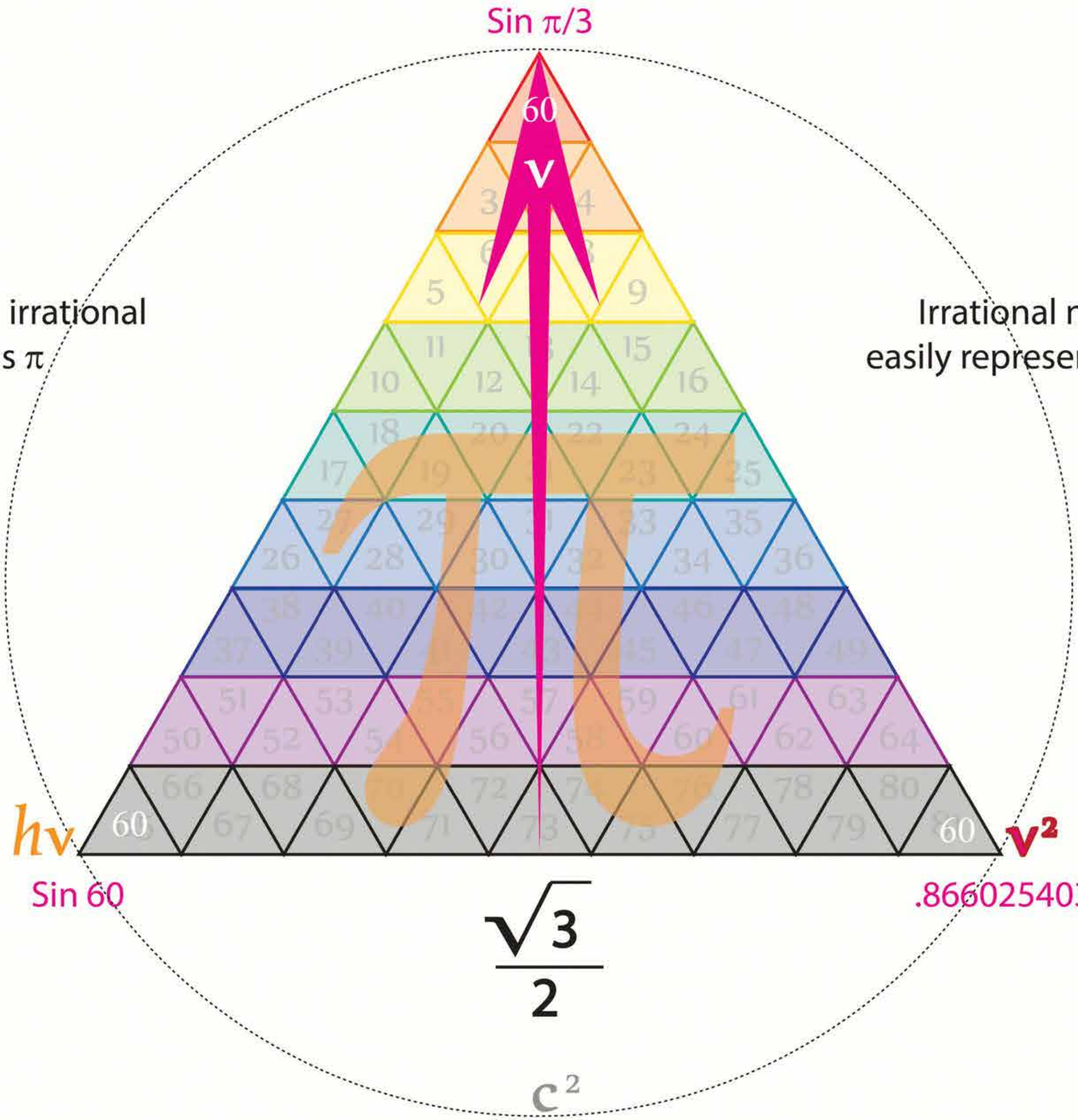
Sin π/3
60
v
hv
60
Sin 60
60
v²
.866025403
√3 / 2
c²

Irrational numbers can be easily represented geometrically

Copyright ABRAHAM [2008] - All rights reserved

Exponentials & Logarithms

e and the Natural Log are inverse functions of each other:

e^x is the amount of continuous GROWTH after a certain amount of time.
Natural Log (ln) is the amount of TIME of continuous growth to reach a certain level

1	0
10	1
100	2
1,000	3

$10{,}000 = 10^4$

$\log(10{,}000) = 4$

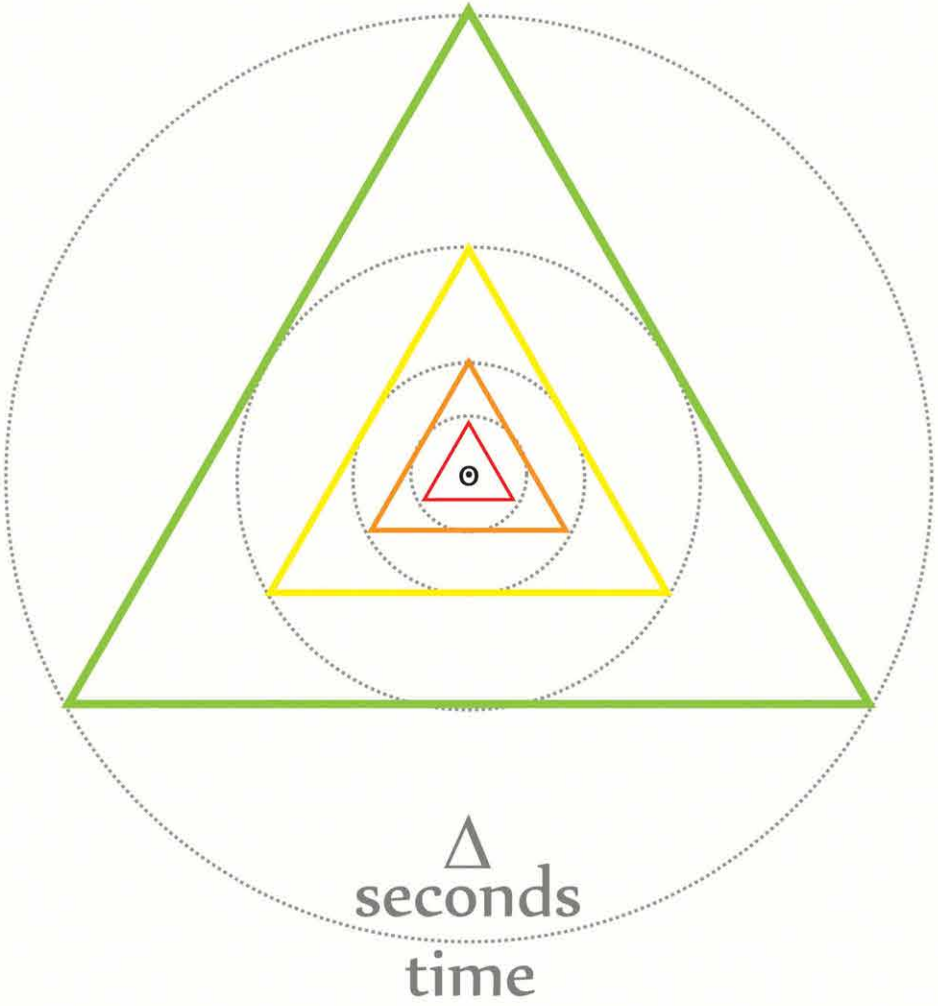

e

How much growth after x units of time
(and 100% continuous growth)

$$e^x = \lim_{n\to\infty}(1 + x/n)^n,$$

GROWTH

$\ln$

ln(x) lets us plug in continuous growth
and get the time it would take.

$$\ln(x) = \lim_{n\to\infty} n(x^{1/n} - 1).$$

PERIOD

Copyright ABRAHAM [2008] - All rights reserved

Exponential growth

$$\text{growth} = e = \lim_{n \to \infty}\left(1 + \frac{r}{n}\right)^{t}$$

e represents the idea that all continually growing systems are scaled versions of a common rate

GEOMETRIC

3.141592654

π

Pi is the ratio between circumference and diameter shared by all circles.

It is a fundamental ratio inherent in all circles and therefore impacts any calculation of circumference, area, volume, and surface areas

Pi radians are equally important and show all quantised equilateral energy geometries are related to their scribed circles

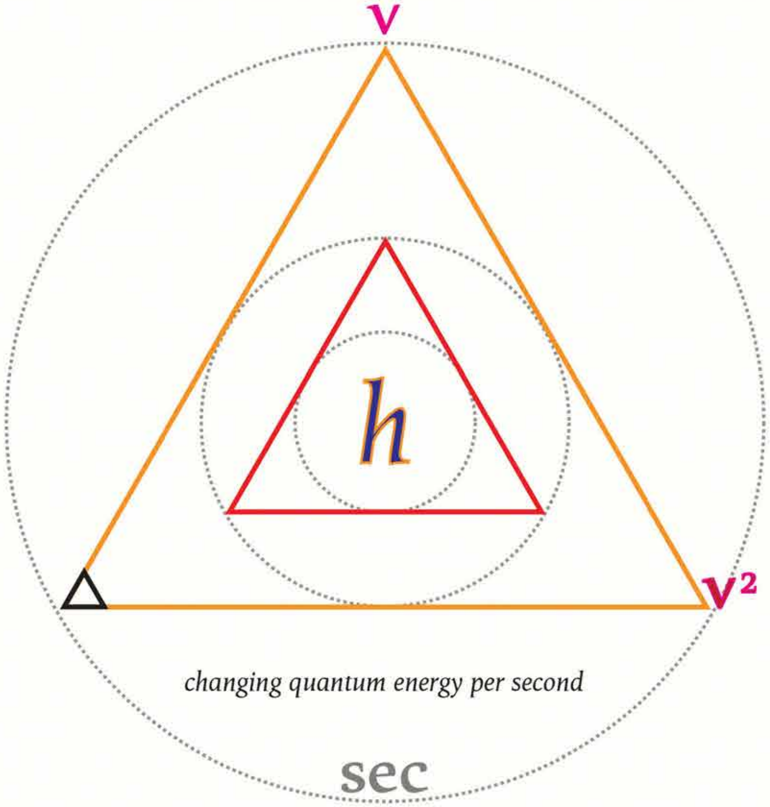

GROWTH

2.718281828

e

e is the base rate of growth shared by all continually growing processes.

e lets you take a simple growth rate (where all the change happens all at once at the end of a period of time - ie quantised growth)

e shows up whenever systems grow exponentially and continuously..... radioactive decay, interest calculations and populations

tetryons

EM fields

quarks

e can be applied to the equilateral energy geometries of physical systems only where the rate of increase is a integer factor of a squared number

leptons

Baryons

Nuclear Energy levels

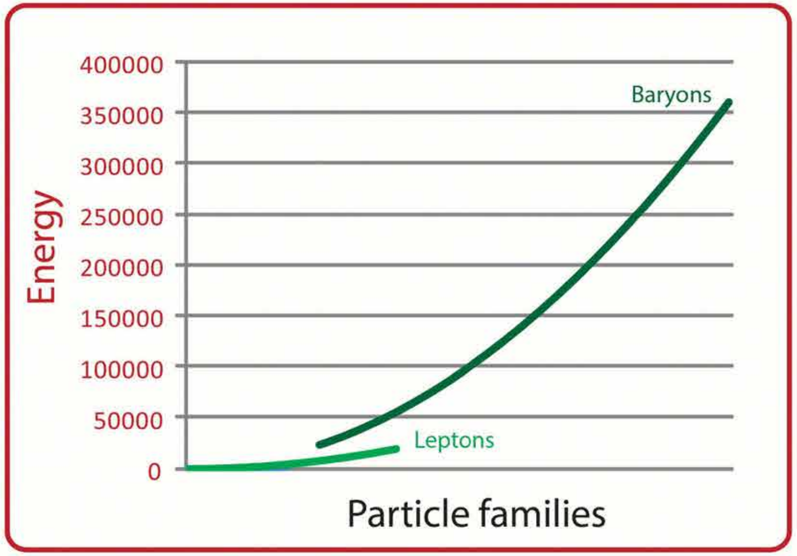

The emission and absorption of bosons and Photons within sub-atomic nuclei

Increase and decrease in integer amounts according to the charged Tetryonic topologies of the particles involved

Exponential energy levels

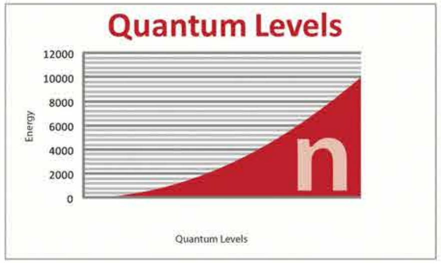

Radioactive decays follow exponential curves determined by the Tetryonic topology of the sub-atomic particle families

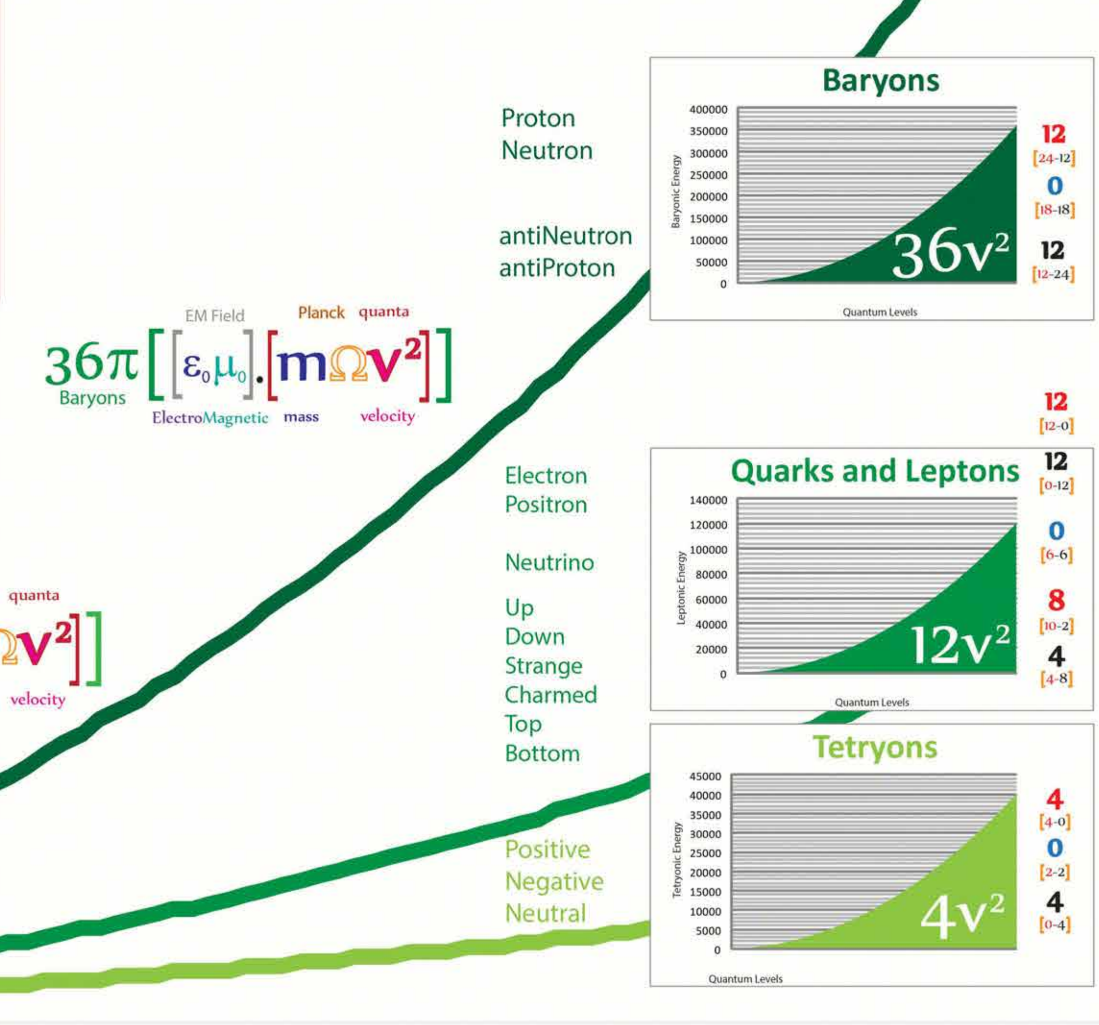

Proton
Neutron

antiNeutron
antiProton

EM Field Planck quanta

$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

Baryons

ElectroMagnetic mass velocity

Electron
Positron

Neutrino

Up
Down
Strange
Charmed
Top
Bottom

quarks EM Field Planck quanta

$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

leptons

ElectroMagnetic mass velocity

Positive
Negative
Neutral

EM Field Planck quanta

$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

tetryons

ElectroMagnetic mass velocity

Copyright ABRAHAM [2008] - All rights reserved

Series addition & the Riemann Zeta Function

The second series addition of the Reimann Zeta function is where x=2: (pi^2)/6=1+1/2^2+1/3^2+1/4^2+...
(the sum of the reciprocals of the squares)

$$\zeta(s) = \sum_n \frac{1}{n^s}$$

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}$$

$$f(n) = 1/n^2$$

$$\sum_{n=1}^{\infty} \frac{1}{n^x}$$

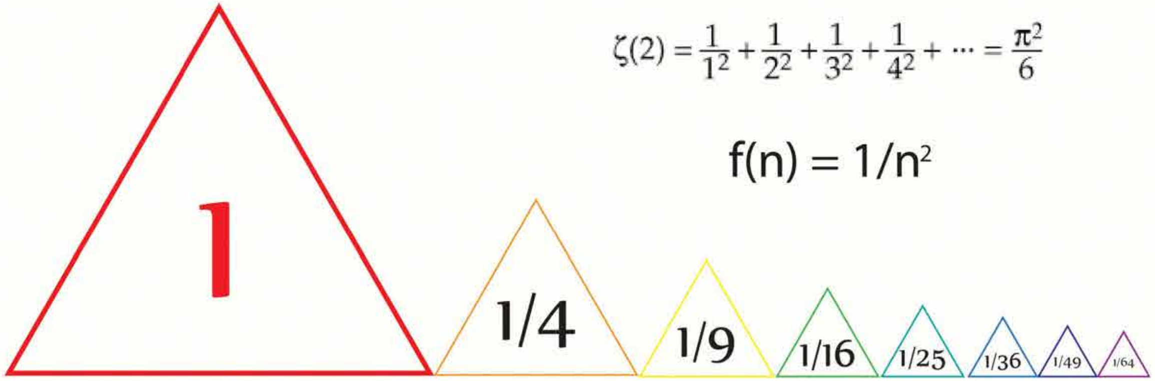

In mathematics, the Riemann zeta function, is a prominent function of great significance in number theory. It is a named after German mathematician Bernhard Riemann. It is so important because of its relation to the distribution of prime numbers. It also has applications in other areas such as physics, probability theory, and applied statistics

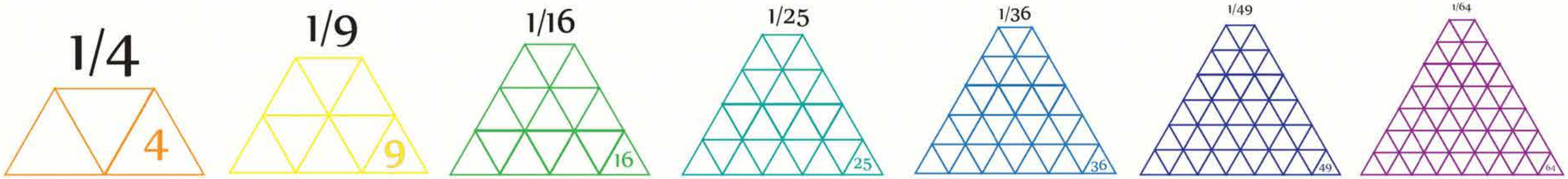

The mystery of prime numbers
Question: which natural numbers are prime? how are they distributed among natural numbers?

Primes are basic building blocks for natural numbers:

- any natural number is a product of prime numbers
- a prime number is only divisible by itself and by 1:
  (it cannot be further simplified)

We don't know how to predict where the prime numbers are:

"Prime numbers grow like weeds among the natural numbers, seeming to obey no other law than that of chance but also exhibit stunning regularity" (Don Zagier, number theorist)

Copyright ABRAHAM [2008] - All rights reserved

Adding the odd numbers in order produces the square numbers

Apart from 2, all primes are odd numbers;
the difference between two consecutive squares being odd,
every prime can be expressed as the difference between two squares

SPECTRAL LINES

Primes can be modelled as Bosons
[ODD number energy momenta geometries]

[2n-1]

KEM ENERGIES

Primes can be expressed as Bosons
[the difference of two squared energies]

$n^2 - [n-1]^2$

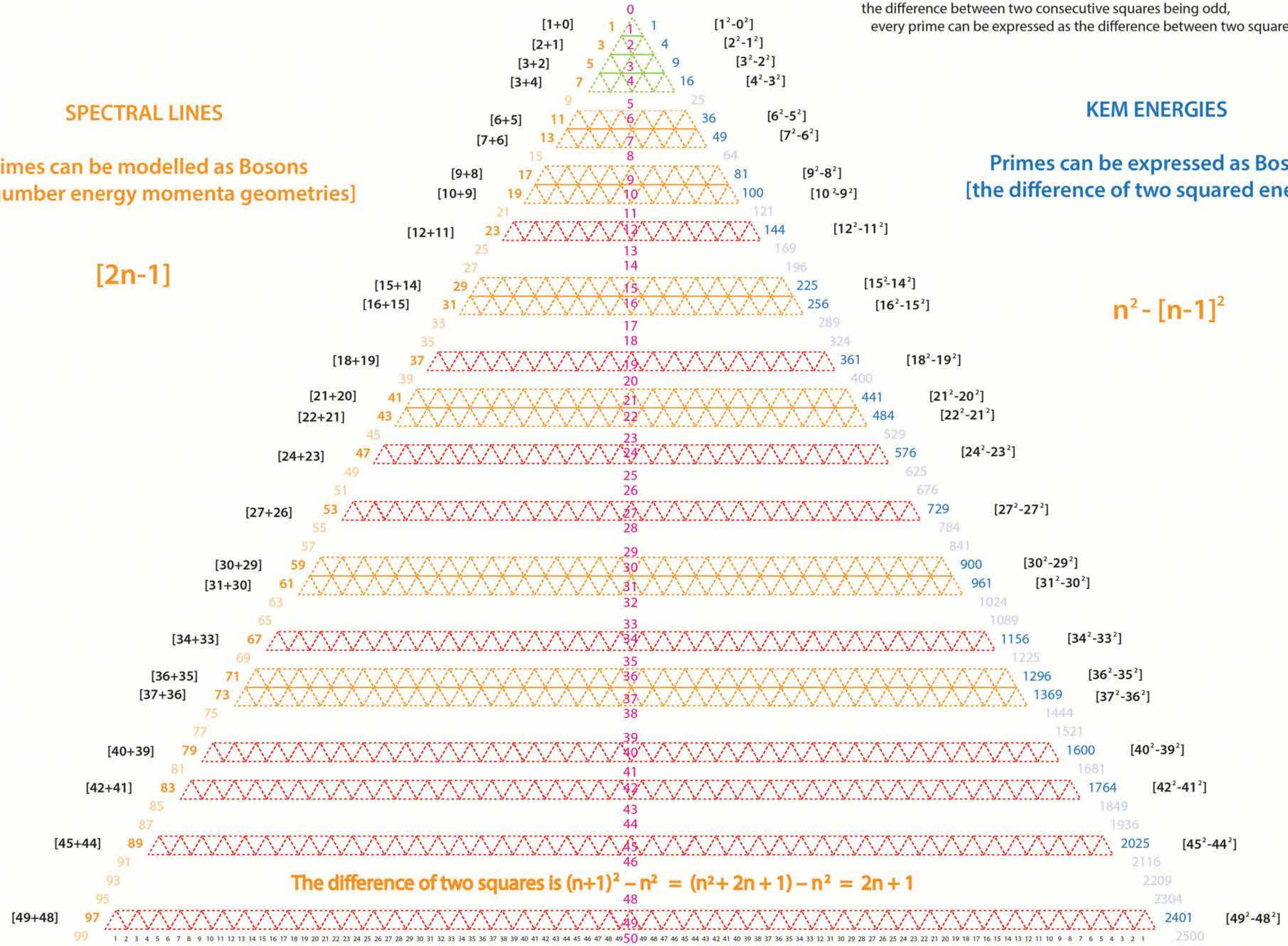
0
[1+0] 1 1 1 $[1^2-0^2]$
[2+1] 3 2 4 $[2^2-1^2]$
[3+2] 5 3 9 $[3^2-2^2]$
[3+4] 7 4 16 $[4^2-3^2]$
9 5 25
[6+5] 11 6 36 $[6^2-5^2]$
[7+6] 13 7 49 $[7^2-6^2]$
15 8 64
[9+8] 17 9 81 $[9^2-8^2]$
[10+9] 19 10 100 $[10^2-9^2]$
21 11 121
[12+11] 23 12 144 $[12^2-11^2]$
25 13 169
27 14 196
[15+14] 29 15 225 $[15^2-14^2]$
[16+15] 31 16 256 $[16^2-15^2]$
33 17 289
35 18 324
[18+19] 37 19 361 $[18^2-19^2]$
39 20 400
[21+20] 41 21 441 $[21^2-20^2]$
[22+21] 43 22 484 $[22^2-21^2]$
45 23 529
[24+23] 47 24 576 $[24^2-23^2]$
49 25 625
51 26 676
[27+26] 53 27 729 $[27^2-27^2]$
55 28 784
57 29 841
[30+29] 59 30 900 $[30^2-29^2]$
[31+30] 61 31 961 $[31^2-30^2]$
63 32 1024
65 33 1089
[34+33] 67 34 1156 $[34^2-33^2]$
69 35 1225
[36+35] 71 36 1296 $[36^2-35^2]$
[37+36] 73 37 1369 $[37^2-36^2]$
75 38 1444
77 39 1521
[40+39] 79 40 1600 $[40^2-39^2]$
81 41 1681
[42+41] 83 42 1764 $[42^2-41^2]$
85 43 1849
87 44 1936
[45+44] 89 45 2025 $[45^2-44^2]$
91 46 2116
93 47 2209
95 48 2304
[49+48] 97 49 2401 $[49^2-48^2]$
99 50 2500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

The difference of two squares is $(n+1)^2 - n^2 = (n^2 + 2n + 1) - n^2 = 2n + 1$

[n-1] **all energy quanta create normal distributions** **[n-1]**

Copyright ABRAHAM [2008] - All rights reserved

Prime number Distributions

A twin prime is a prime number that differs from another prime number by two, for example the twin prime pair (41, 43). Sometimes the term twin prime is used for a pair of twin primes; an alternative name for this is prime twin or prime pair.

Twin primes appear despite the general tendency of gaps between adjacent primes to become larger as the numbers themselves get larger due to the prime number theorem (the "average gap" between primes less than n is log(n)).

$$p_2 - p_1 = 2$$

Every twin prime pair except (3, 5) is of the form (6n – 1, 6n + 1) for some natural number n, and with the exception of n = 1, n must end in 0, 2, 3, 5, 7, or 8

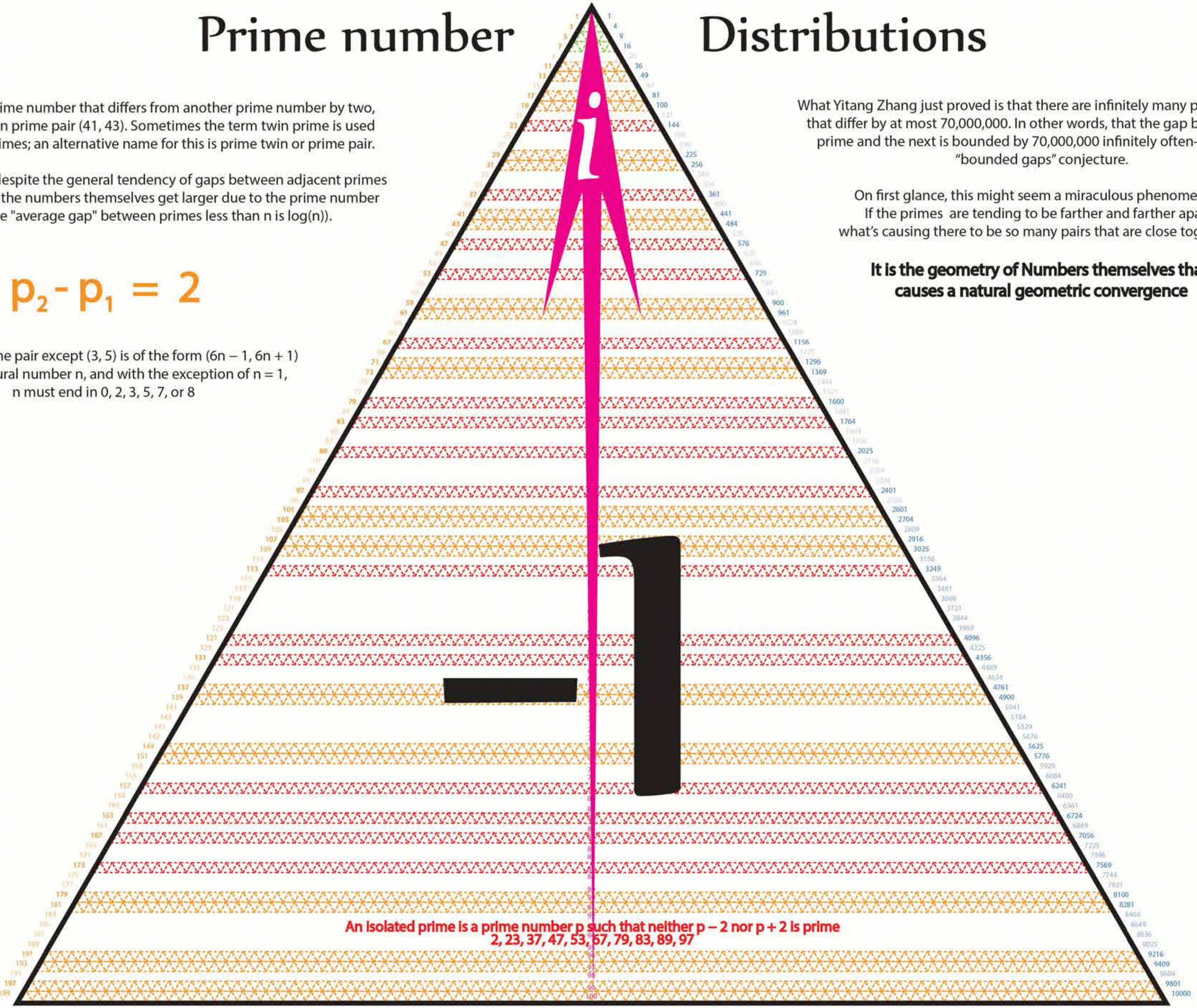

An isolated prime is a prime number p such that neither p – 2 nor p + 2 is prime
2, 23, 37, 47, 53, 67, 79, 83, 89, 97

What Yitang Zhang just proved is that there are infinitely many pairs of primes that differ by at most 70,000,000. In other words, that the gap between one prime and the next is bounded by 70,000,000 infinitely often—thus, the "bounded gaps" conjecture.

On first glance, this might seem a miraculous phenomenon. If the primes are tending to be farther and farther apart, what's causing there to be so many pairs that are close together?

It is the geometry of Numbers themselves that causes a natural geometric convergence

The Digital roots of Prime numbers

The digital root (also repeated digital sum) of a number is the (single digit) value obtained by an iterative process of summing digits, on each iteration using the result from the previous iteration to compute a digit sum.

The process continues until a single-digit number is reached.

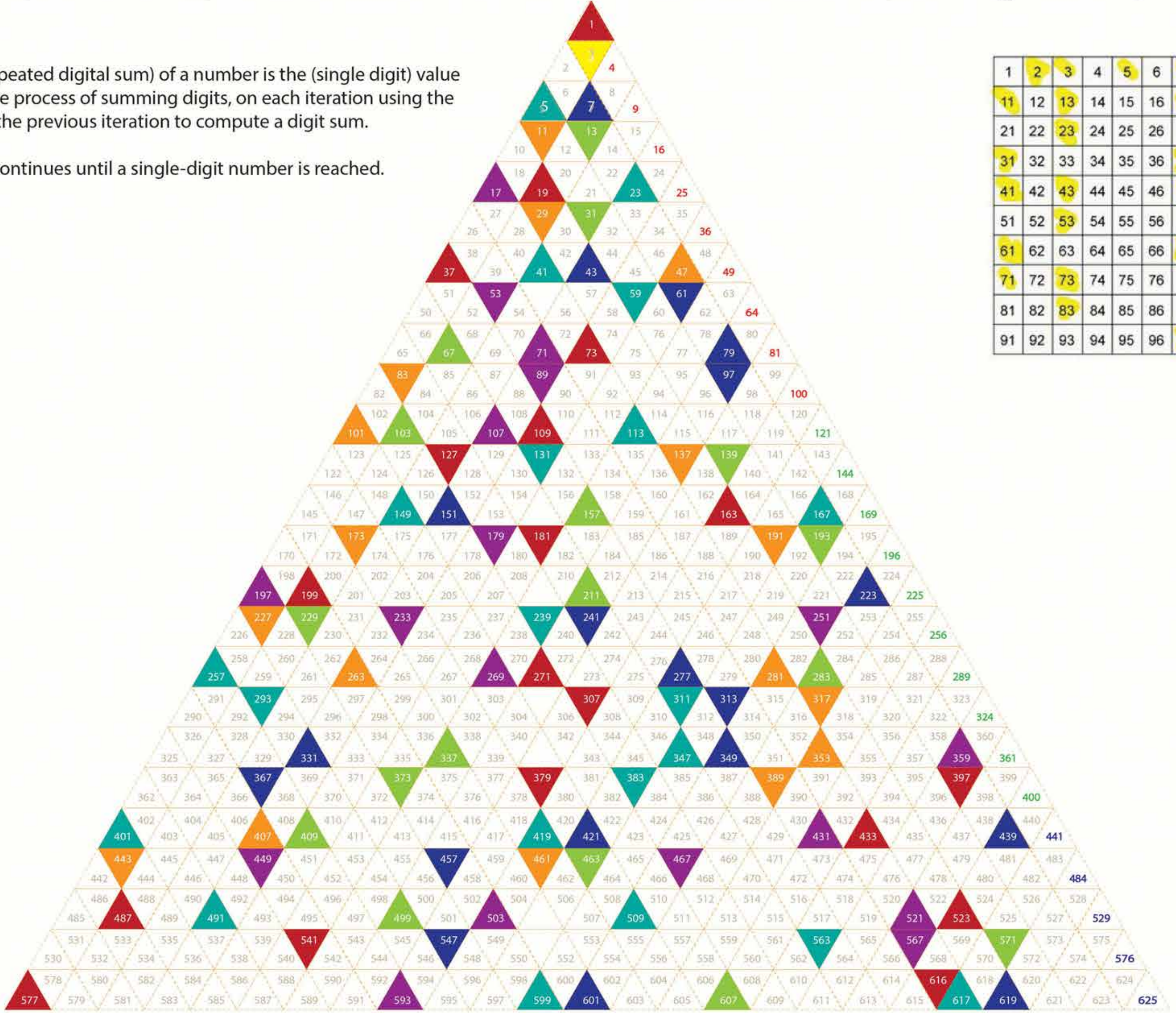

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Copyright ABRAHAM [2008] - All rights reserved

Archimedes is given credit for the first calculus.

Archimedes infinite series

Today's calculus was published by Newton.

1 + 3/4 + 3/16 + 3/64 + 3/256 + 3/1024 + ...

1 + .75 + .1875 + .046875 + .01171875 + ...

$$3\left(\frac{1}{4}+\frac{1}{4^2}+\frac{1}{4^3}+\frac{1}{4^4}+\cdots\right)=1.$$

$$\frac{3}{4}+\frac{3}{4^2}+\frac{3}{4^3}+\frac{3}{4^4}+\cdots=1.$$

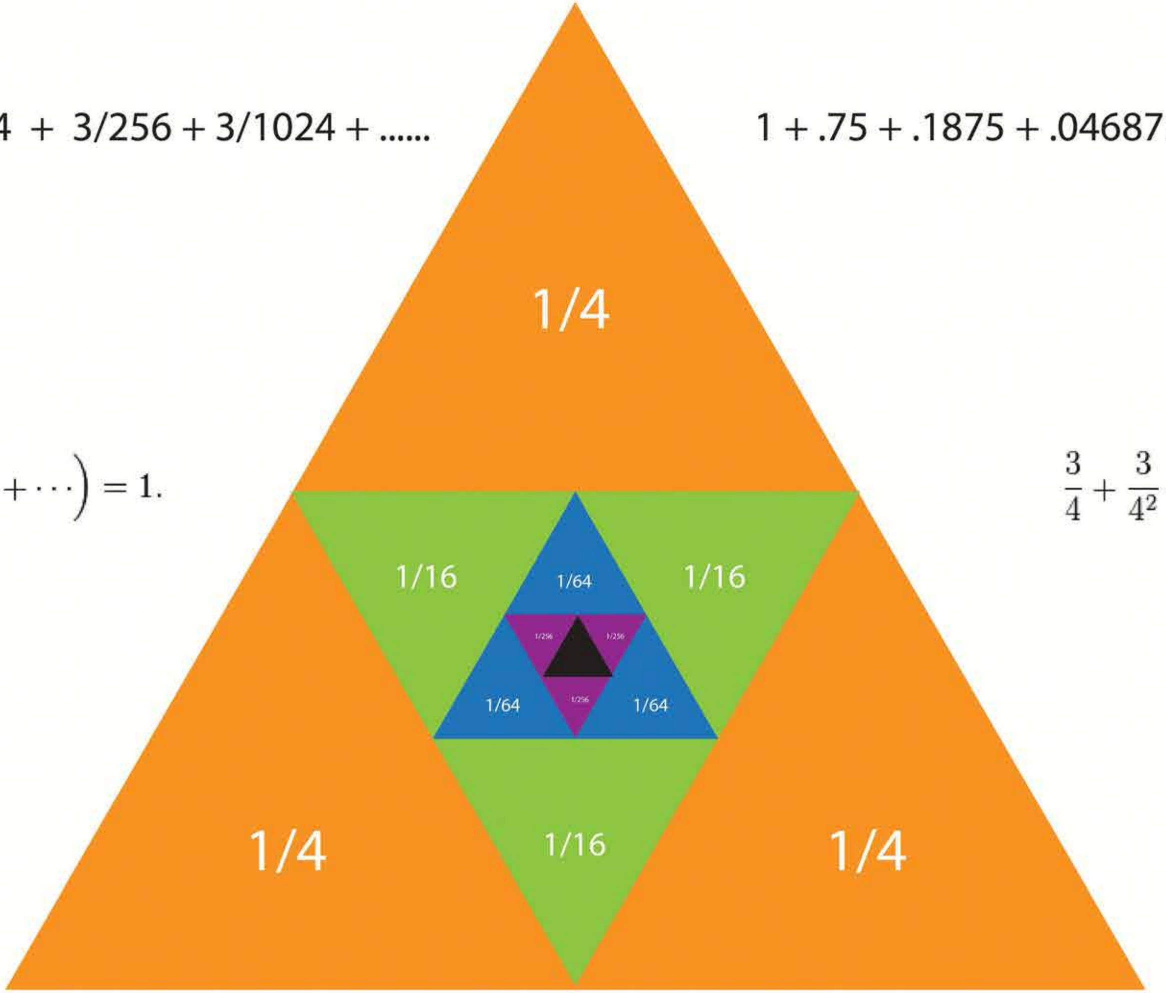
1/4
1/16
1/64
1/16
1/256
1/256
1/64
1/256
1/64
1/4
1/16
1/4

Nested convergent infinite series

Summing a convergent infinite series

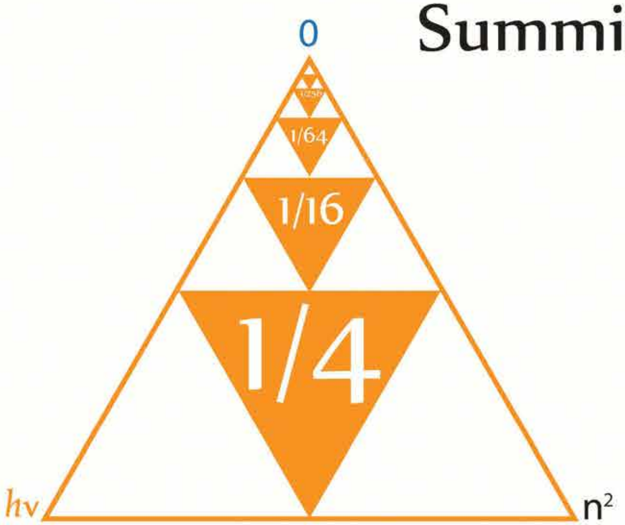
0
1/256
1/64
1/16
1/4
hv
n²

$1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 + \ldots\ldots\ldots\ldots$

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

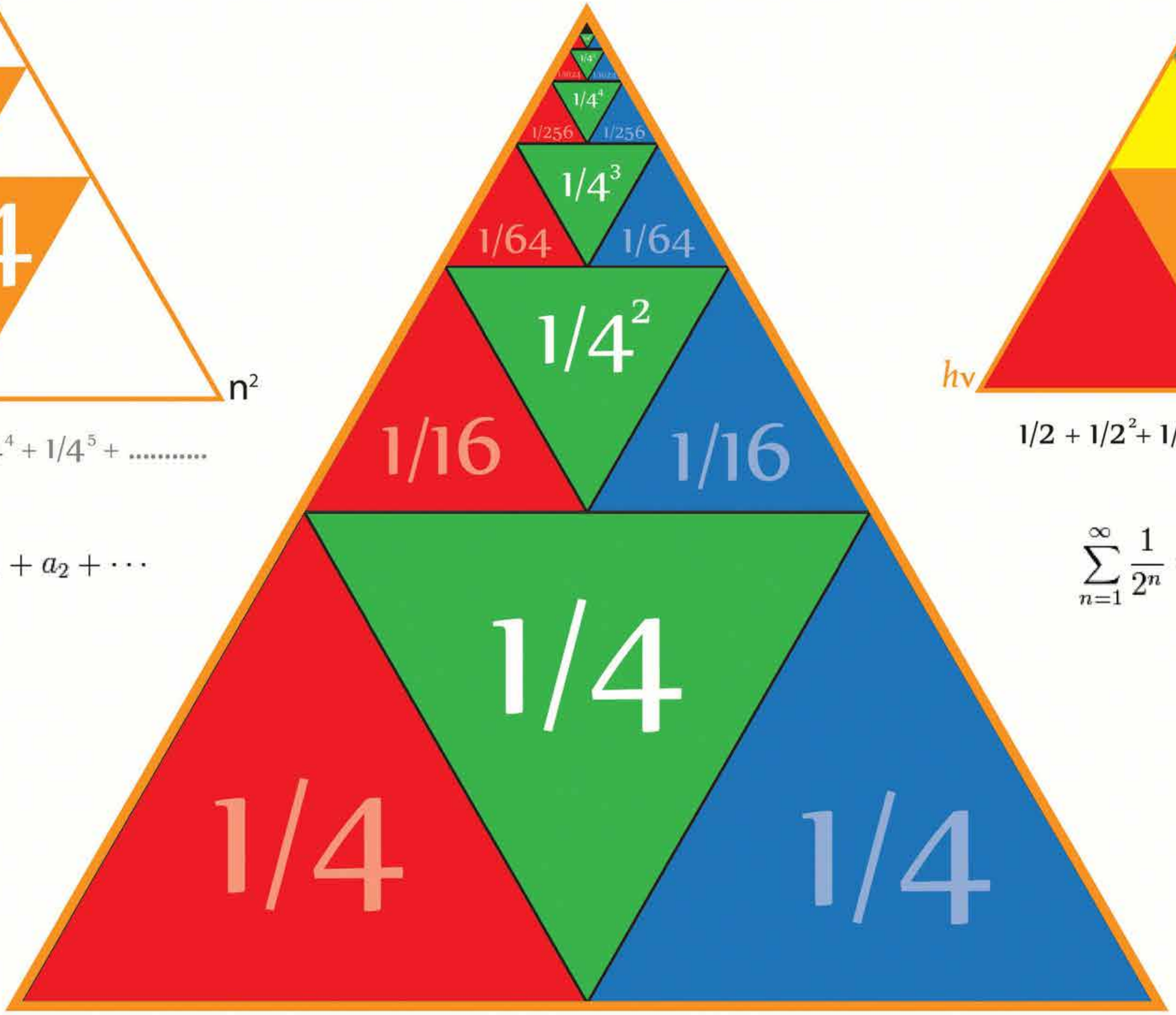
1/4⁵
1/1024
1/1024
1/4⁴
1/256
1/256
1/4³
1/64
1/64
1/4²
1/16
1/16
1/4
1/4
1/4

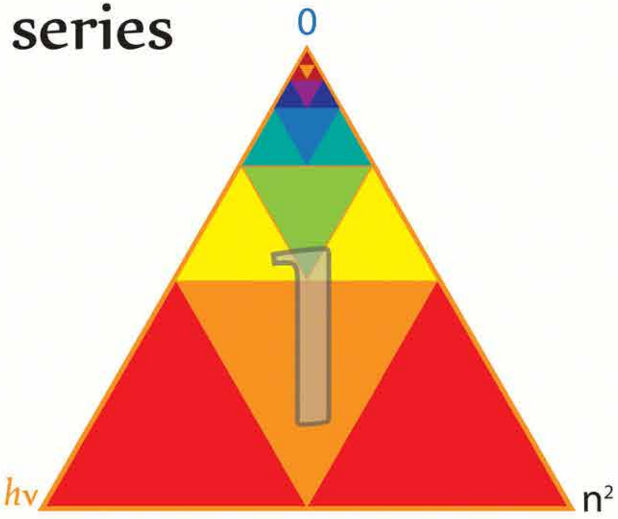

$1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 + \ldots\ldots\ldots\ldots$

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

ODDS = 2n-1

$$\sum_{1}^{n} 2n-1 = n^2$$

Summing the dissimilar coloured equilateral triangles gives unity

1/3 1/3 1/3

1/3 1/3 1/3

Summing the dissimilar coloured equilateral triangles gives unity

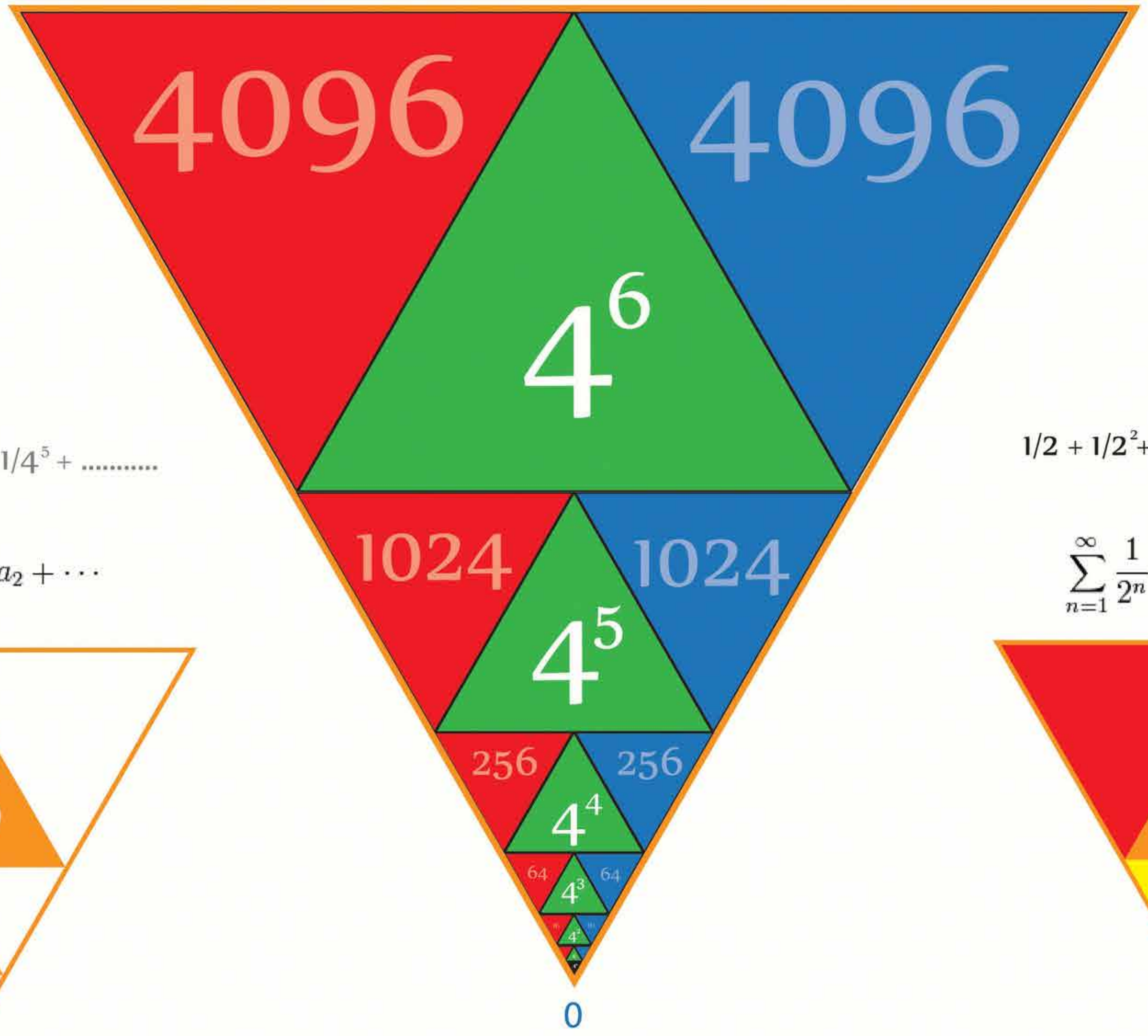
4096
4096
4^6
1024
1024
4^5
256
256
4^4
64
64
4^3
0

$1/4 + 1/4^2 + 1/4^3 + 1/4^4 + 1/4^5 + \ldots\ldots\ldots\ldots$

$$\sum_{n=0}^{\infty} a_n = a_0 + a_1 + a_2 + \cdots$$

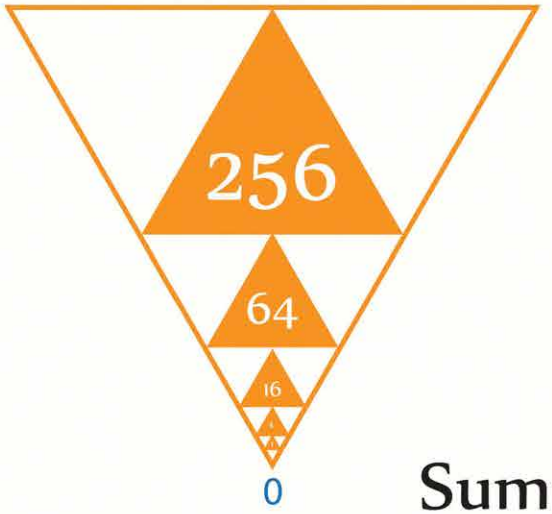

$1/2 + 1/2^2 + 1/2^3 + 1/2^4 + 1/2^5 + \ldots\ldots\ldots\ldots$

$$\sum_{n=1}^{\infty} \frac{1}{2^n} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots.$$

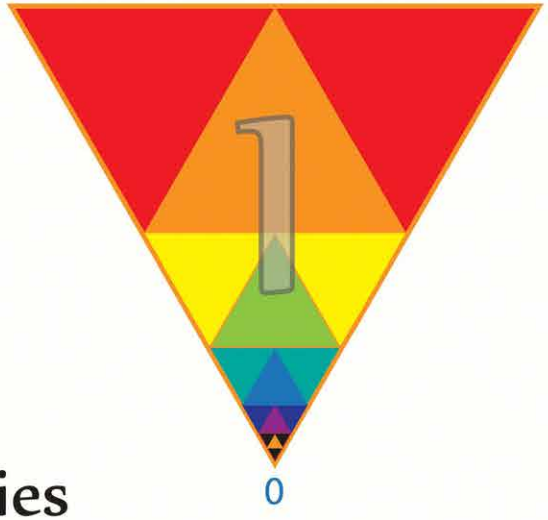

Summing a divergent infinite series

Copyright ABRAHAM [2008] - All rights reserved

Pietro Mengoli

(1626 – June 7, 1686)

The Basel Problem

The Basel problem is a famous problem in mathematical analysis with relevance to number theory, first posed by Pietro Mengoli in 1644 and solved by Leonhard Euler in 1735.

Leonhard Euler

(15 April 1707 – 18 Sept 1783)

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6} \approx 1.644934$$

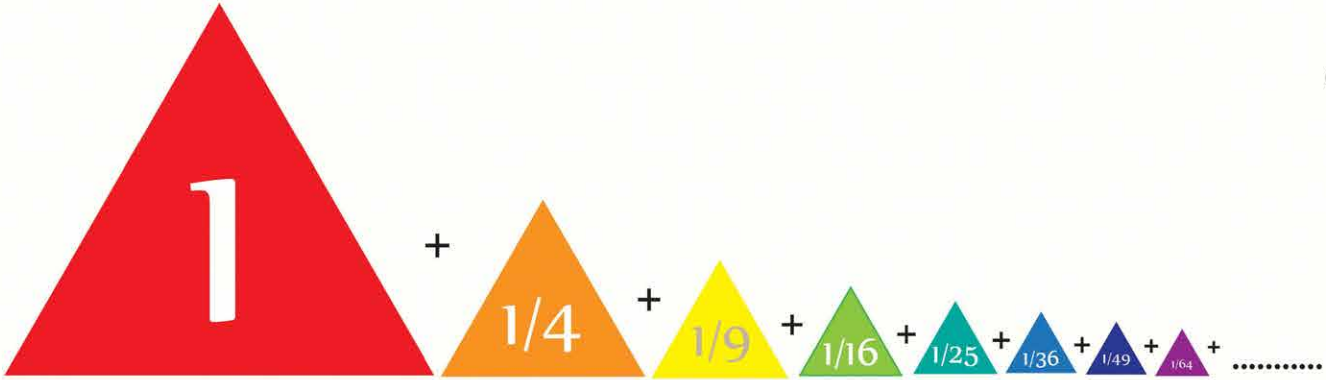

The Basel problem asks for the precise summation of the reciprocals of the squares of the natural numbers, i.e. the precise sum of the infinite series

$$f(n) = 1/n^2$$

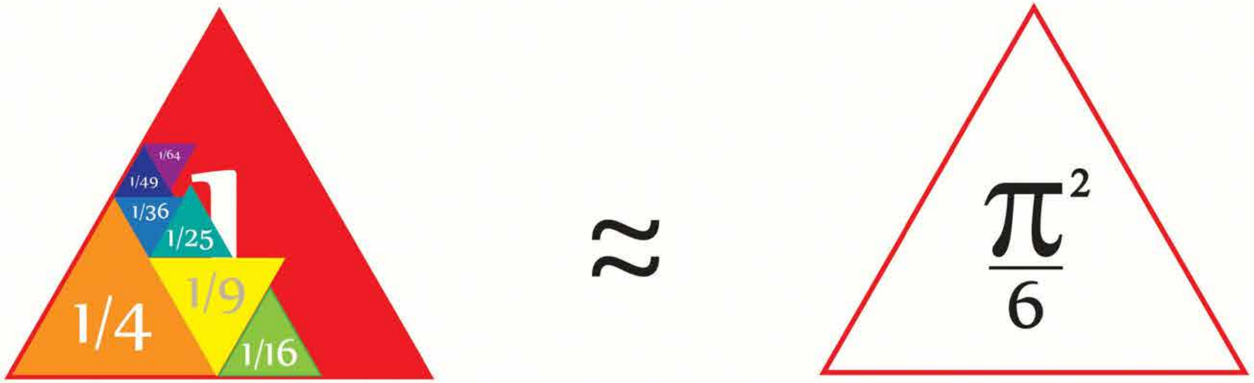

Integrals of mass-energy

is a means of finding scalar areas using summation and limits.

Integration is a micro adding of ***CONTINUOUS*** quantities.

$$\int_{n}^{\infty} f(x)$$

Integration is a special case of summation.

Integration is defined as the limit of a summation as the number of elements approches infinity while a part of their respective value approaches zero.

Summation is the finite sum of multiple, fixed values.

$$\sum_{n}^{\infty} x$$

Summation is a macro adding of ***DISCRETE*** quantities.

The summation of equilateral energy momenta quanta with respect to their linear vector components

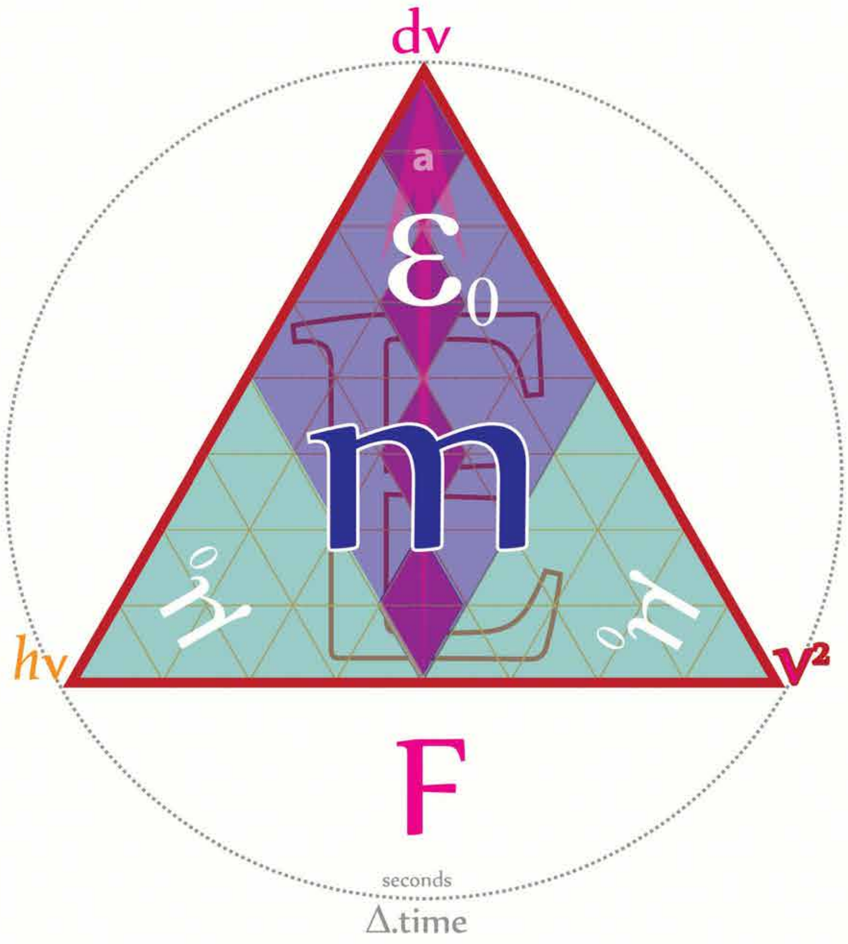

Integrating the energy quanta contained within equilateral charge geometry gives the variable Force required to acheive changes in motion [Energy, work, acceleration]

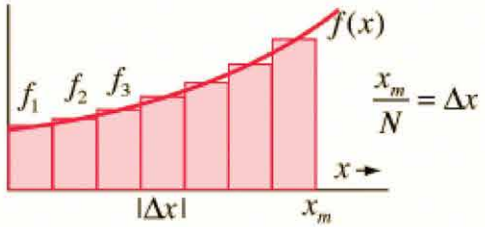

The Integral of the continuous area under the curve is the summation of an infinte number of disctrete rectangular measurements made to a specified limit

An integration isn't a simple summation, but the limit of a sequence of summations

All Planck energies are discrete equilateral geometries

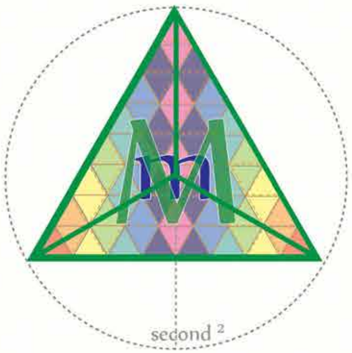

mass is the surface integral of EM energy geometries per unit of time

"The calculus of infinitesimals"

The fundamental theorem of calculus
simply states that the sum of infinitesimal changes in a quantity over time
adds up to the net change in the quantity.

Scalar Energy

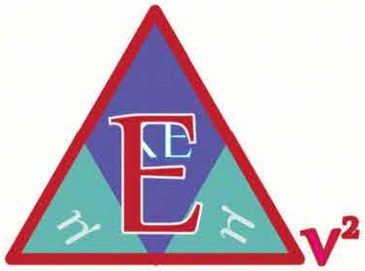

$E = mv^2$

Leibniz's vis viva
(Latin for living force)
is mv2,

[2 x kinetic energy]

vector Force

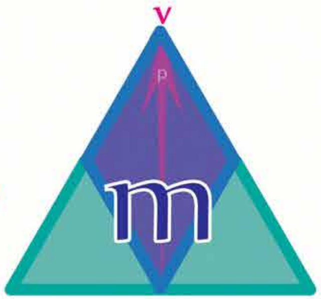

Much of Newton's work
centred around momentum
and changes to it

[mass.velocity]

$F = m[dv/dt]$

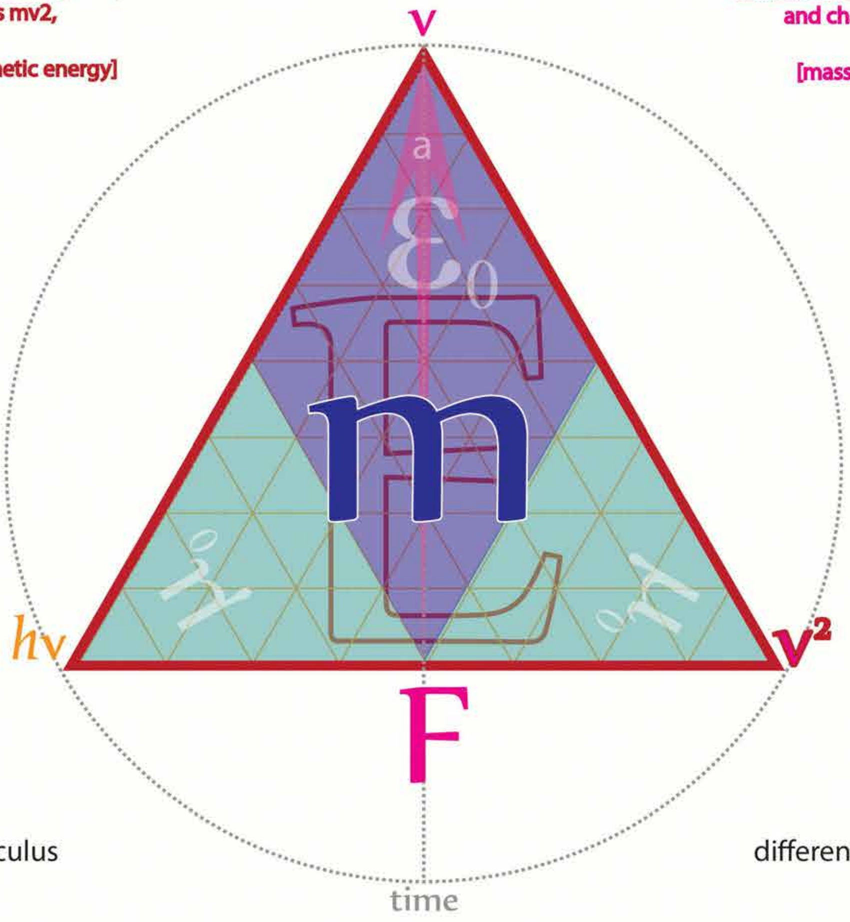

Gottfried Wilhelm von Leibniz

(July 1, 1646 – November 14, 1716)

Sir Isaac Newton

(1643-1727)

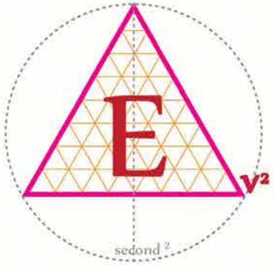

integral calculus

differential calculus

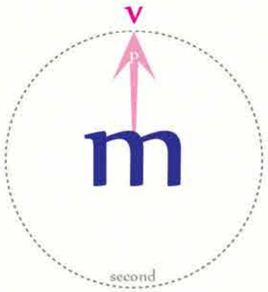

The founders of calculus thought
of the integral as an infinite sum of
rectangles of infinitesimal width

The fundamental theorem of calculus is a theorem that links
the concept of the integral with the derivative of a function.

In calculus, a branch of mathematics,
the derivative is a measure of how
a function changes as its input changes

Copyright ABRAHAM [2008] - All rights reserved

Energy momentum geometries

Leibniz

scalar energies

linear momentum is the square root vector force of the scalar energy required to do a set amount of work

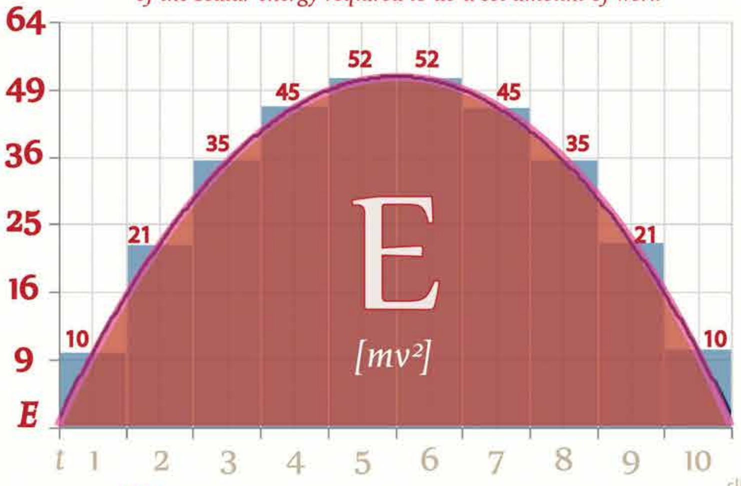

$$E = p^2$$

Energy momentum

$$kg\left[\frac{m}{s}\right]^2$$

Newton

linear momentum

$d[mv]/dt = dp/dt = ma$

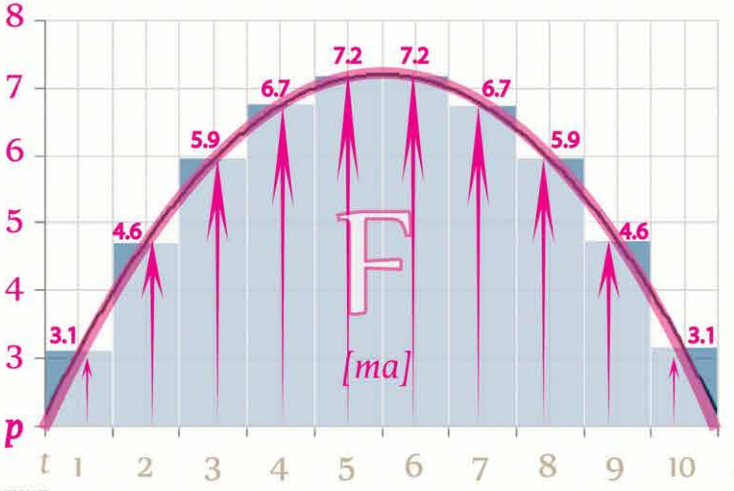

print out this page

then cut out the 10 Planck energy momenta triangles shown

slice fine lines along the pink linear momentum arrows with a razor blade or similar as highlighted under the Newtonian acceleration curve

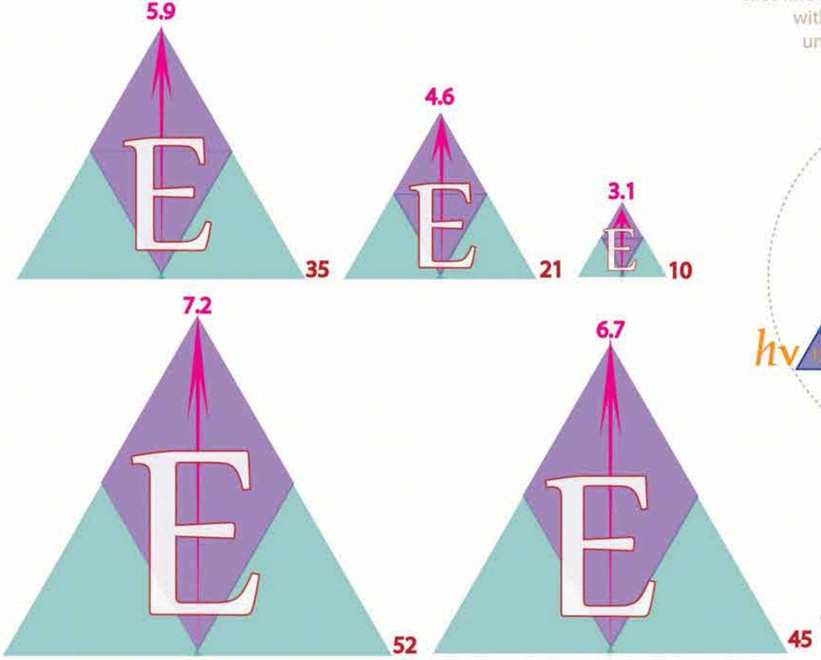
5.9
E
35
4.6
E
21
3.1
E
10
7.2
E
52
6.7
E
45

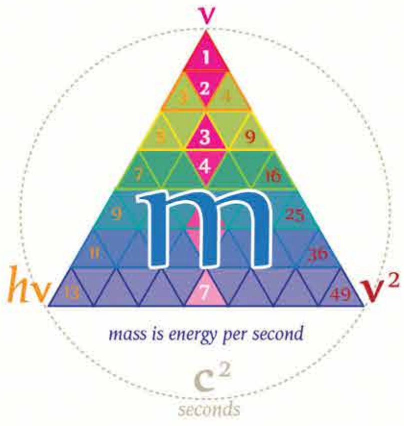
v
1
2
3
4
5
3
9
7
4
16
9
25
11
36
13
7
49
m
hv
v²
c²
seconds

mass is energy per second

slide the cut-out triangles into the slots created

rotate the assembled model to show the real force momenta geometry at work

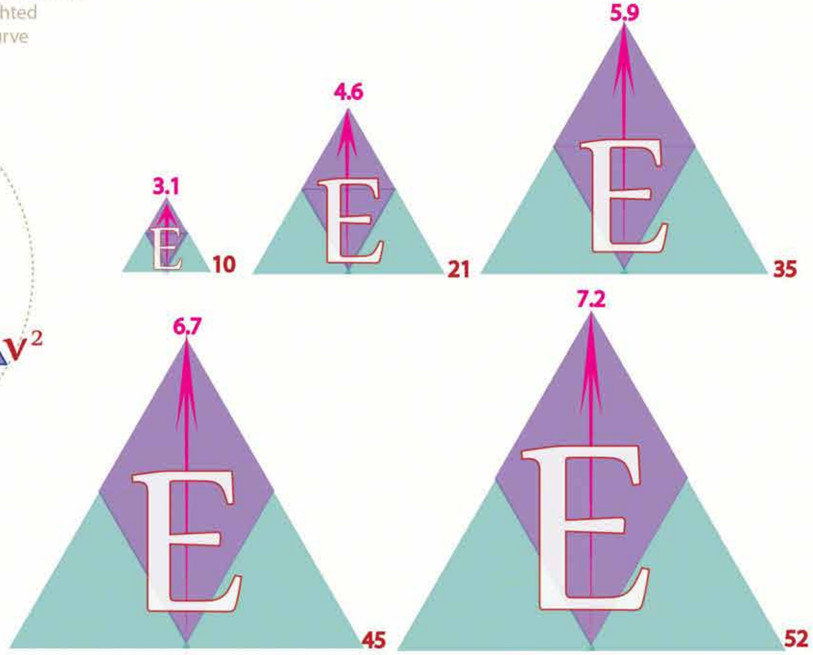
3.1
E
10
4.6
E
21
5.9
E
35
6.7
E
45
7.2
E
52

Tetryonics 87.07 - Physically modeling the geometric forces of acceleration in calculus *linear momentum is the square root vector force of the scalar energy required to do a set amount of work*

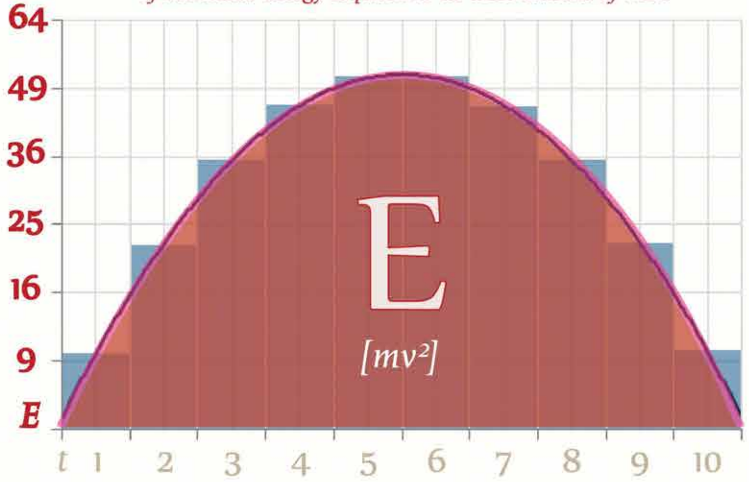

p

linear momentum

$$kg\,\frac{m}{s}$$

print out this page

$d[mv]/dt = dp/dt = ma$

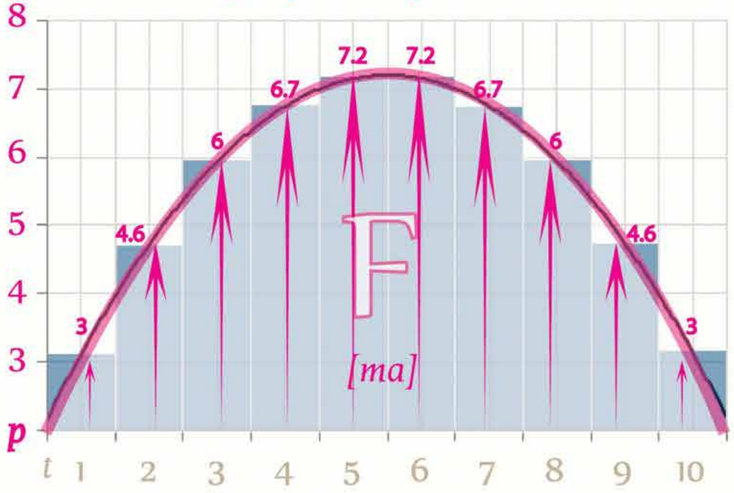

Leibniz

scalar energies

$$E = mv^2$$

mass-energy momenta relationship

square root equilateral energy is linear momentum

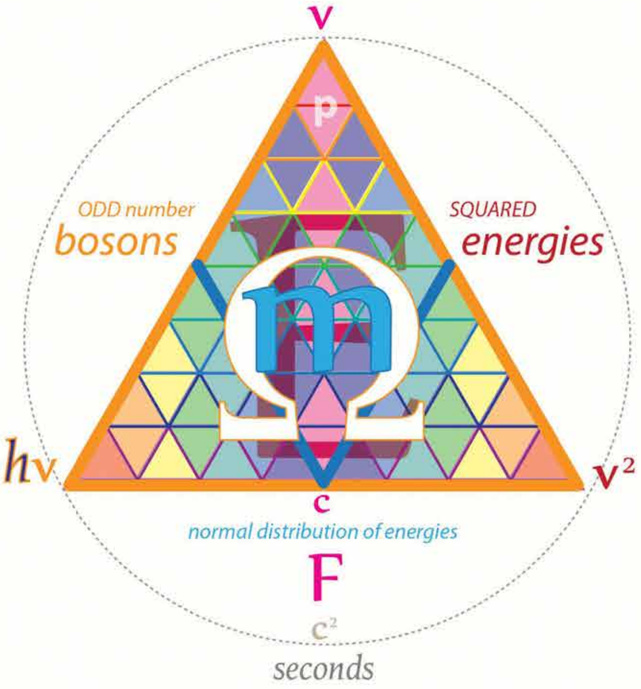
v
p
ODD number
bosons
SQUARED
energies
m
Ω
hv
v²
c
normal distribution of energies
F
c²
seconds

Newton

linear momentum

$$F = ma$$

mass is a scalar constant relating Force to acceleration

$$F = m\frac{\Delta v}{\Delta t}$$

Planck quanta and their vector linear momentum lie at 90 degrees to the angle use in the graphs of motion in calculus

Differentiation

Differentiation is concerned with things like speeds and accelerations, slopes and curves etc.
These are Rates of Change, they are things that are defined locally.

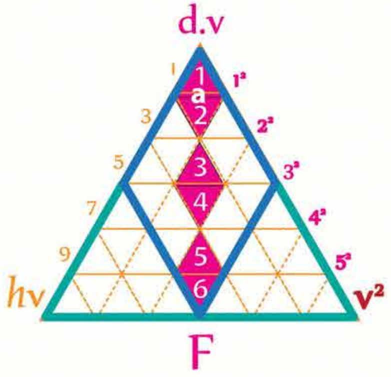

An increase in a force
opposing an object's
vector velocity results in
DECCELERATION

A linear measure of Forces acting on physical systems
resulting in changes to distance covered per unit of time

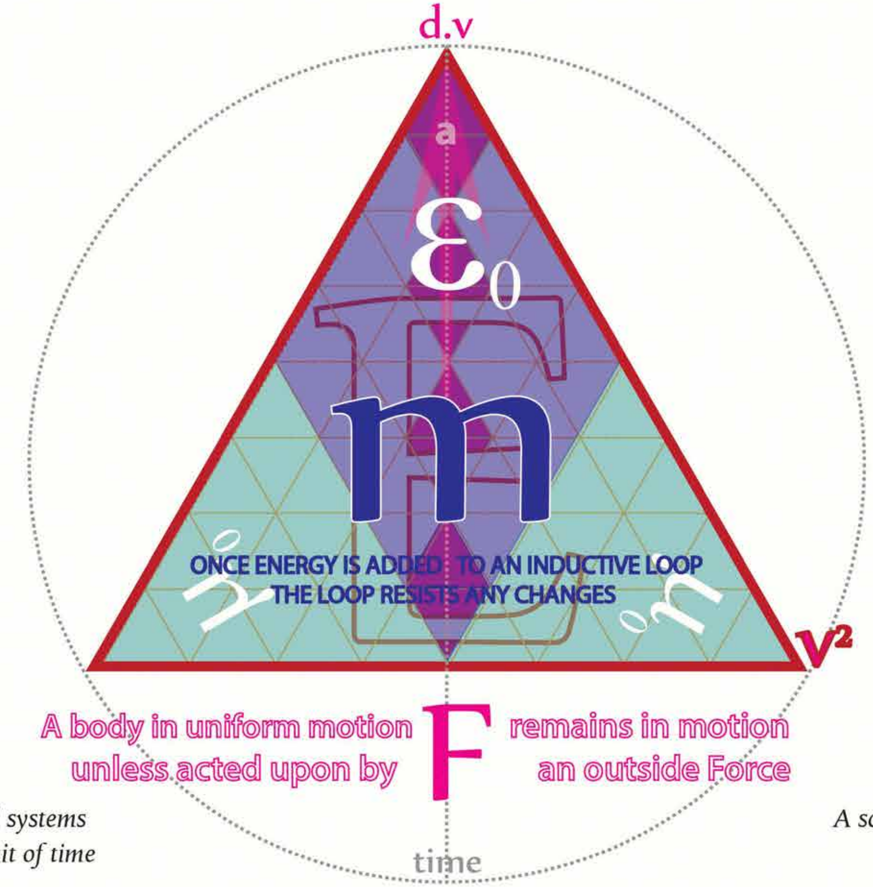

ONCE ENERGY IS ADDED TO AN INDUCTIVE LOOP
THE LOOP RESISTS ANY CHANGES

A body in uniform motion remains in motion
unless acted upon by **F** an outside Force

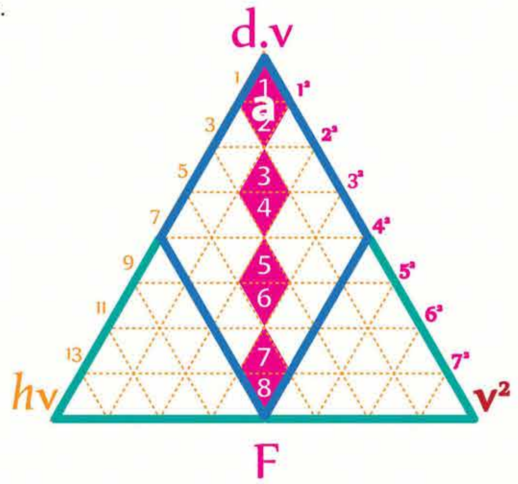

An increase in a force
in line with an object's
vector velocity results in
ACCELERATION

A scalar measure of Forces acting on physical systems
resulting in changes to their rate of motion

$$v = \frac{\Delta d}{\Delta t}$$

$$= m/s$$

$$F = ma$$

In physics, the derivative of the displacement of a moving body with respect to time is the velocity of the body, and the derivative of velocity with respect to time is acceleration.

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

Newton's second law of motion states that the derivative of the momentum of a body equals the force applied to the body.

$$a = \frac{\Delta v}{\Delta t}$$

$$= m/s^2$$

Visualising the geometric half-truths of relativistic physics

The source of all the physical relationships of mass-energy momenta and the constants in Physics is the Equilateral Triangle (and all texts must be corrected)

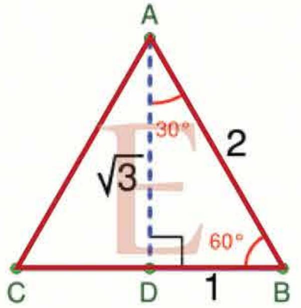

Energy geometries within Physics including Special Relativity with its Lorentz corrections have historically been incorrectly illustrated through the geometry of right angled triangles

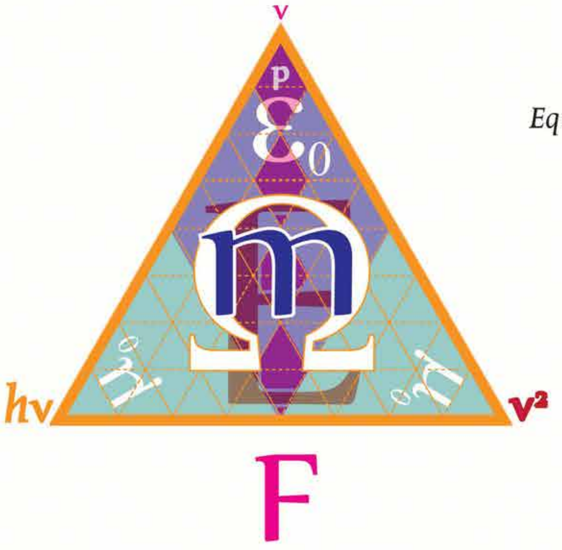

$$\sum \mathbf{F} = \frac{d\mathbf{P}}{dt} = m\frac{d\mathbf{v}}{dt} = m\mathbf{a}$$

$$\mathbf{F} = m\mathbf{a}\,.$$

**Physics is geometry,
one cannot be separated from the other**

Equilateral geomtries lead to a intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all the other apsects of Nature.

$$mv^2 = E = hv^2$$

ENERGY

6.629432673 e-34 J

Planck
quanta
mΩv²
mass
velocity

7.376238634 e-51 kg

momentum

$$p^2 = mv^2$$

$$E = m_0c^2 + K$$

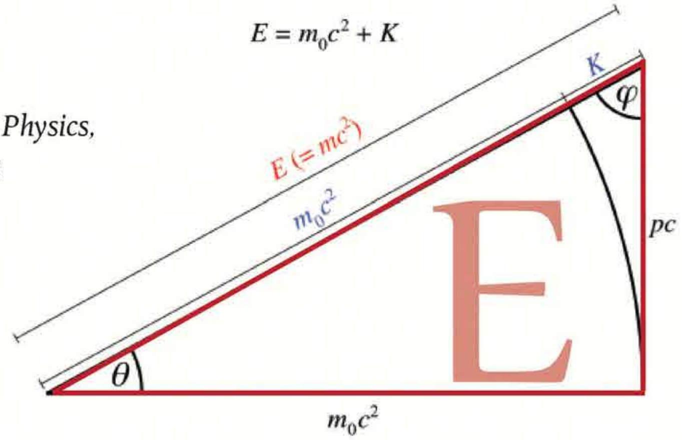

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total energy, mass, or distance in spacetime is the sum of the components squared.

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related tot the Energy content of a body by the velocity of Energy

$$E = mc^2$$

Velocity

In physics, velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The scalar absolute value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms−1) when using the SI (metric) system.

$$\frac{m}{s}$$

Speed is the scalar value of the
Distance traveled per unit of TIme

$$\bar{\mathbf{v}} = \frac{\Delta \mathbf{x}}{\Delta t}.$$

Velocity is the vector value of the
Distance traveled per unit of Time

$$\frac{\vec{m}}{s}$$

Velocity squared is the scalar value of the
Distance traveled per unit of Time squared
(Energy of a given spatial volume)

$$\frac{m^2}{s^2}$$

v Velocity $\frac{m}{s}$

sec

is a 2D
radial Space-time
MEASUREMENT

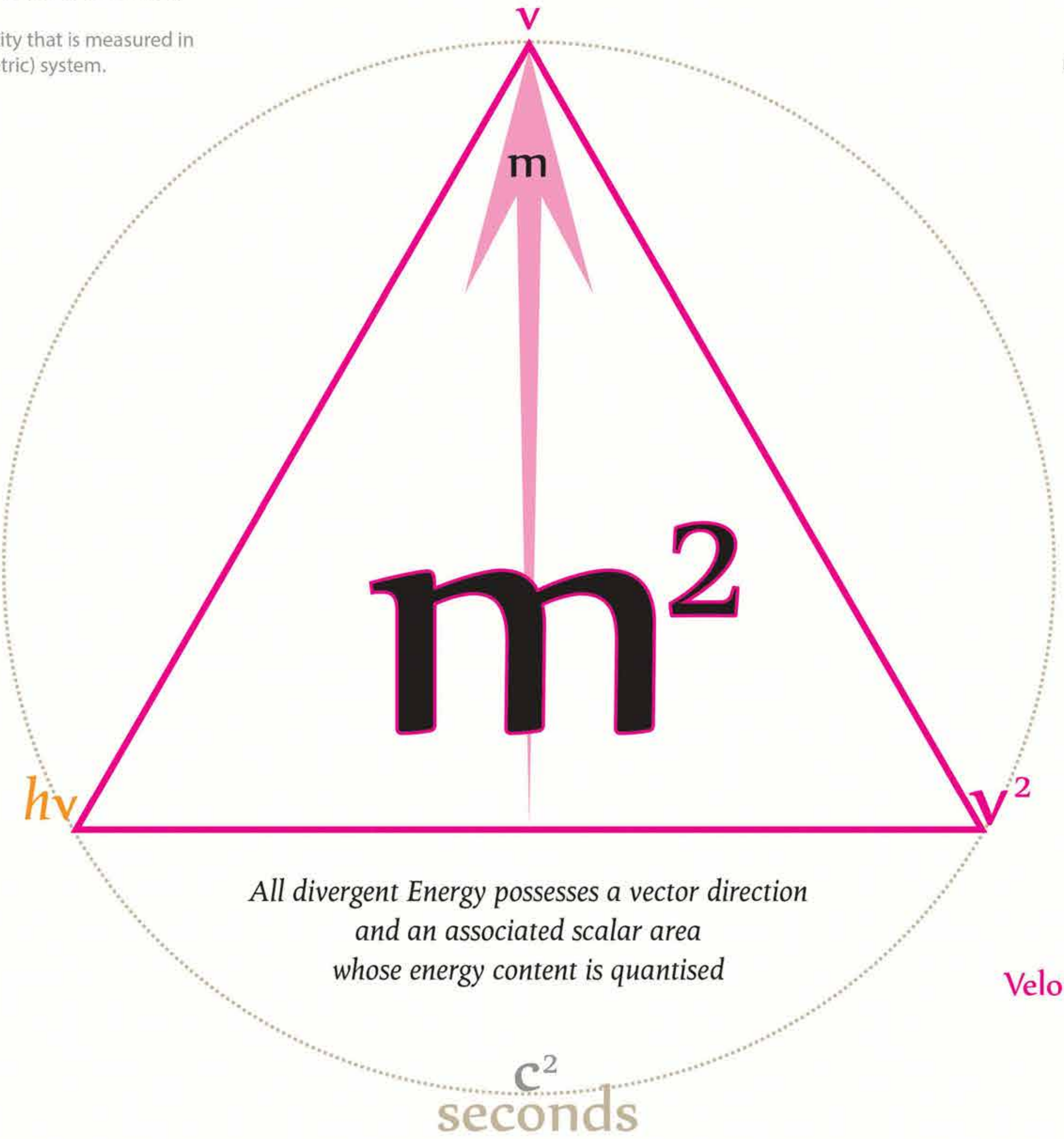

*All divergent Energy possesses a vector direction
and an associated scalar area
whose energy content is quantised*

Velocity squared

$$\frac{m^2}{s^2}$$

Copyright ABRAHAM [2008] - All rights reserved

Acceleration

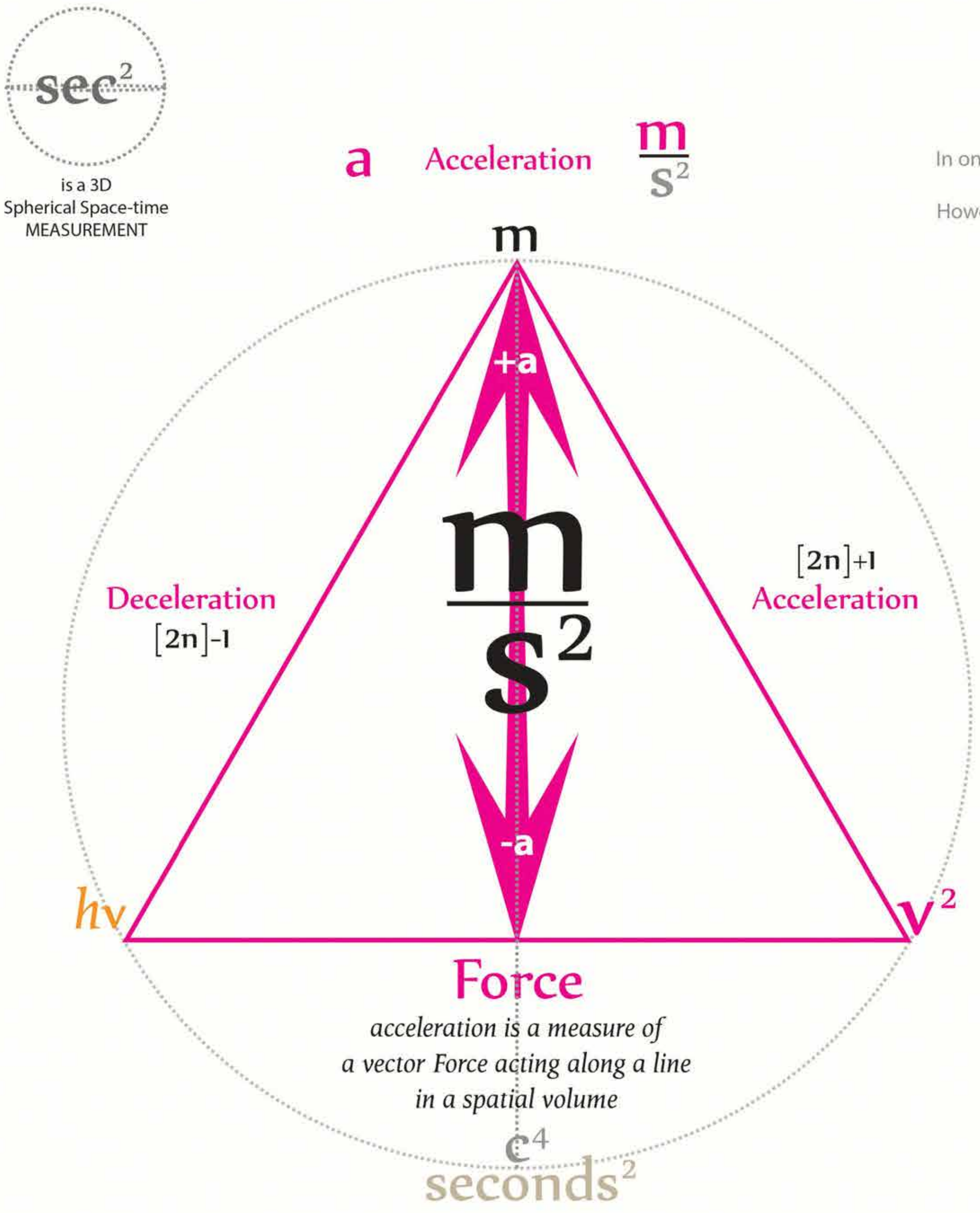
sec^2
is a 3D
Spherical Space-time
MEASUREMENT
a
Acceleration
$\frac{m}{s^2}$
m
+a
$\frac{m}{s^2}$
-a
Deceleration
[2n]-1
[2n]+1
Acceleration
hν
v^2
Force
c^4
seconds^2

acceleration is a measure of
a vector Force acting along a line
in a spatial volume

In physics, acceleration is the rate of change of velocity over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

Force

$$kg\,\frac{m}{s^2}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

$kg\frac{m}{s}$

linear momentum

p

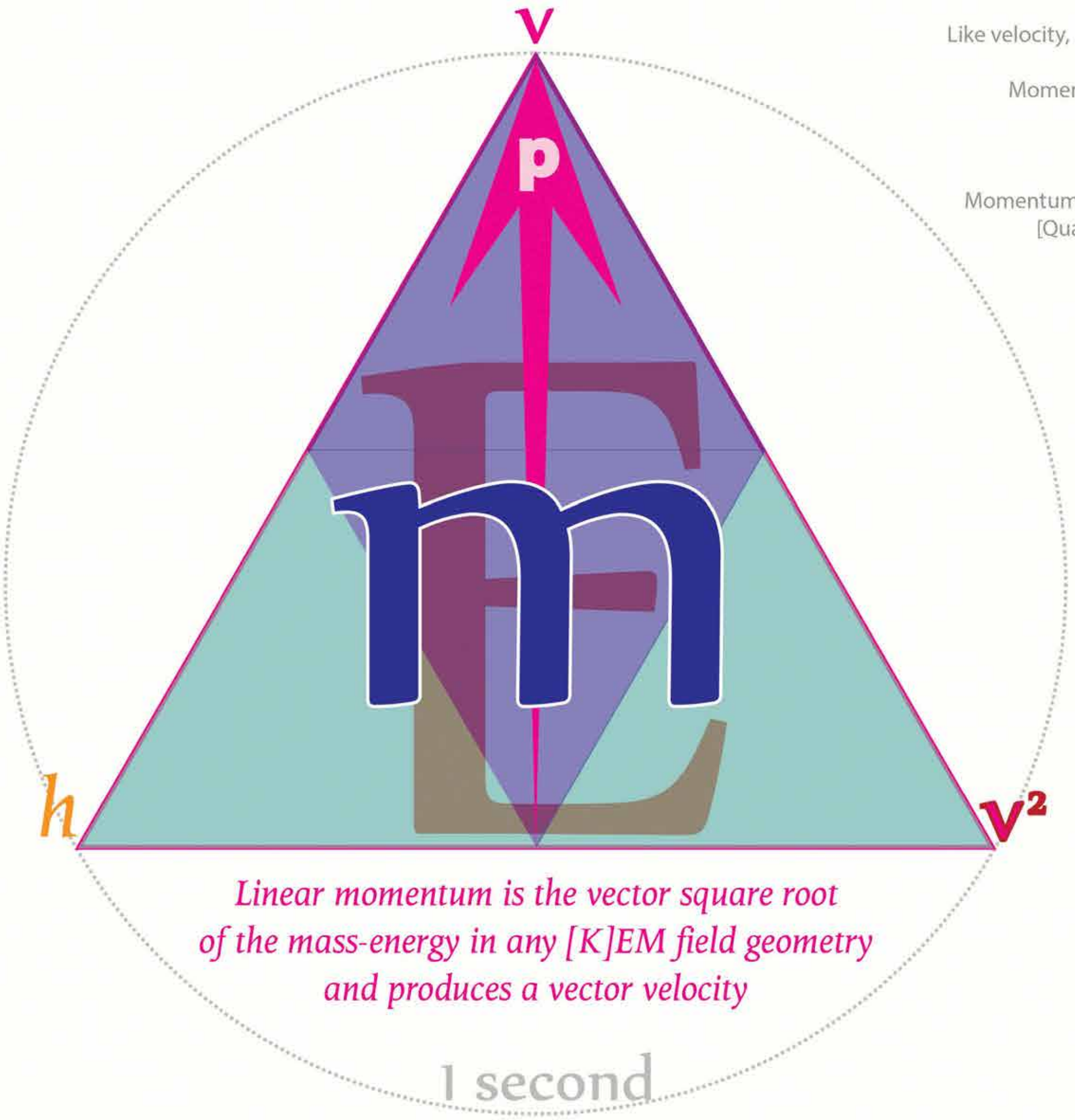

Linear momentum is the vector square root of the mass-energy in any [K]EM field geometry and produces a vector velocity

Momentum

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s) is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum), meaning that if a closed system is not affected by external forces, its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms [Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{hv^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as the square of linear momentum

$$kg\frac{m^2}{s^2}$$

equilateral Planck energy momenta

$$m\Omega v^2$$

Energy-momentum relationship

The total number of equilateral Planck quanta [quantised mass-energy momenta]
is directly related to the square of its linear momentum [mass-velocity]

Quantum Mechanics

$$h = \frac{E}{v^2}$$

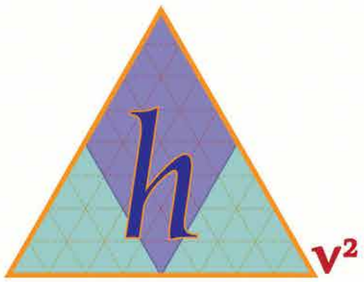

Quantised energy
equilateral momenta

$$hv^2$$

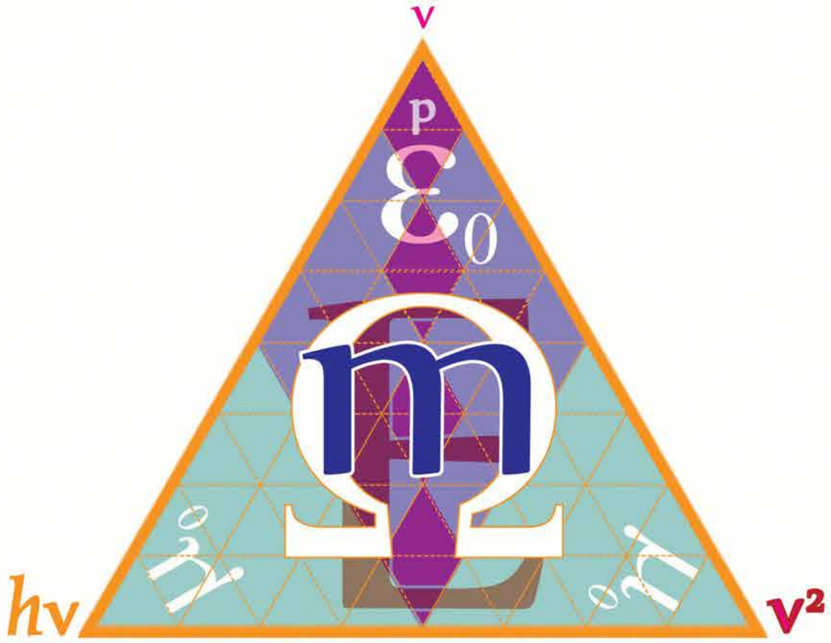

Quantised Energy momenta is related to
Scalar mass energy momenta through
the equilateral geometry of Planck's constant

Newtonian physics

$$m = \frac{E}{v^2}$$

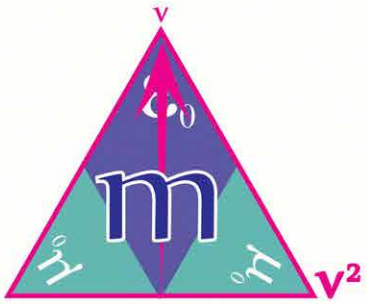

Scalar energy
linear momentum

$$pv$$

$$m\Omega v^2 = E = mv^2$$

Inertial resistance to Force

Inertia is the resistance of any physical object
to a change in its state of motion.

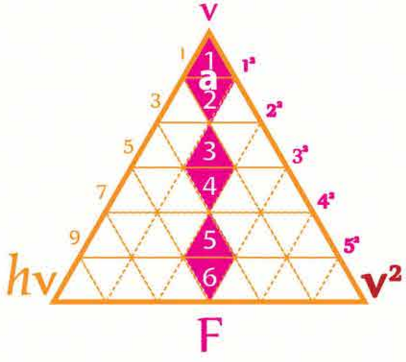
v
a
1
2
3
4
5
6
1
3
5
7
9
1²
2²
3²
4²
5²
hv
v²
F

Changes to mass-velocity
require a corresponding
change to boson quanta

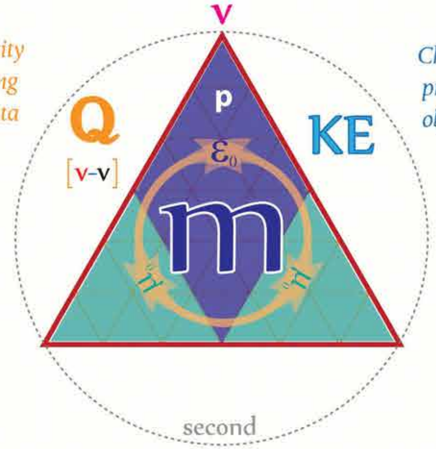
v
Q
[v-v]
p
ε0
KE
m
second

Changes to mass-velocity
produces a change in an
object's Kinetic Energies

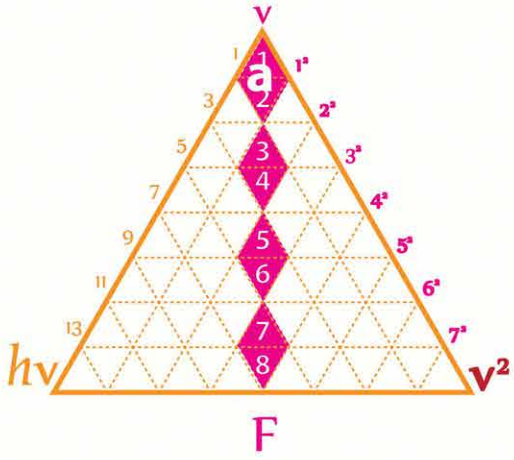
v
a
1
2
3
4
5
6
7
8
1
3
5
7
9
11
13
1²
2²
3²
4²
5²
6²
7²
hv
v²
F

$$E = mc^2$$

The total intrinsic momenta of
all energy waveforms is the
sum of their constituent
Quantised Angular momenta
(mass-energy momenta)

$$M_0 = m/c^2$$

Any change in motion results in changes to
the Charge geometries creating in turn proportional
changes to KEM mass and momenta components

$$F = ma$$

The 'inductive resistance' of Charge quanta fields to
changes in their mass-energy momenta content
is what we term Inertia

$$KEM = Mv^2$$

Matter in motion
has Kinetic Energies
in addition to invariant
rest mass-Energy

KE = RE - rest Matter

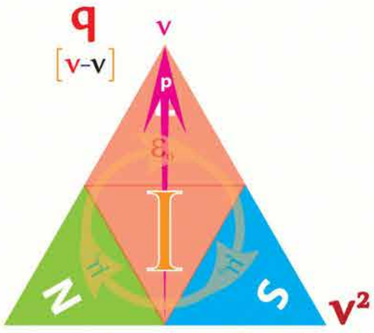
q
[v-v]
v
p
ε0
I
N
S
v²

quantised moment of inertia

Any change to an object's velocity results in a corresponding
change to its mass-Energy momenta which is reflected by its inertia

$$p^2 = E = mv^2$$

Inertial mass can be related to Charge through inductive Planck quanta

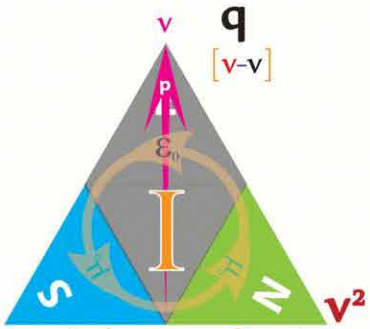
v
q
[v-v]
p
ε0
I
S
N
v²

quantised moment of inertia

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

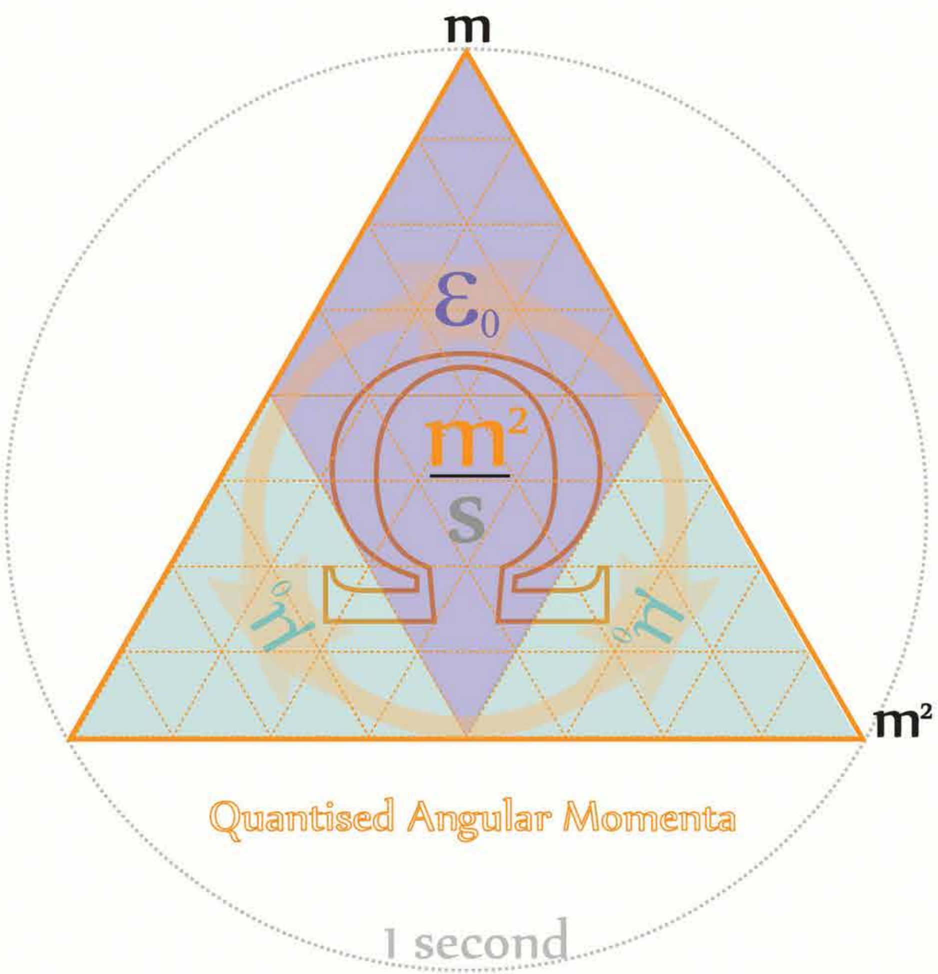

Angular momentum is sometimes described as the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of quantised mass-energy momenta within any defined space-time co-ordinate system

A major re-definition of quantised angular momentum in physics is revealed

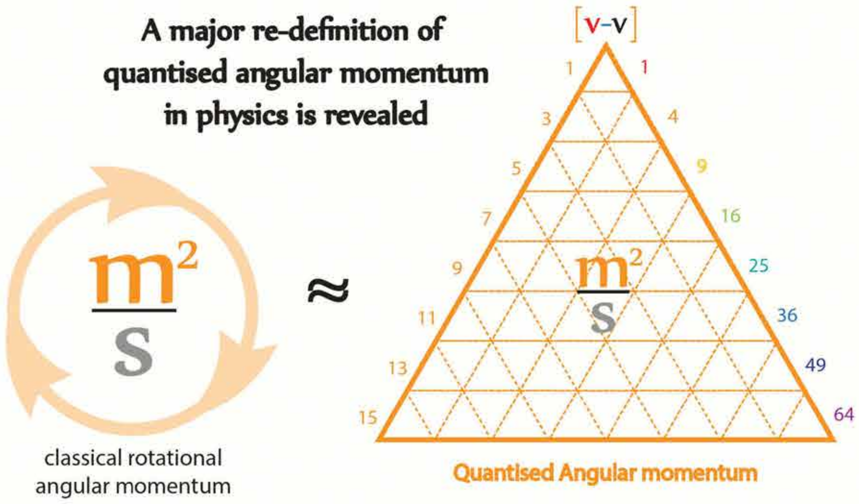

classical rotational angular momentum

Quantised Angular momentum

In quantum mechanics, angular momentum is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum, or rotational momentum is a conserved vector quantity that can be used to describe the overall state of a physical system.

When applied to specific mass-Energy-Matter systems QAM reveals the true quantum geometry and nature of Energy in our universe

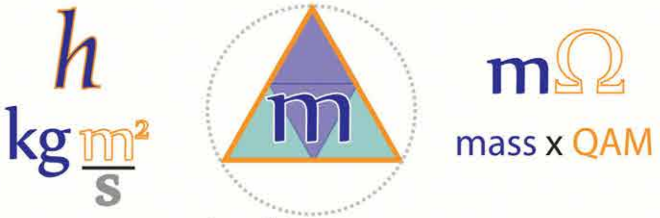

Planck's Constant

Normally viewed as an expression of rotational momentum Quantised Angular Momentum [QAM] is in fact a result of the equilateral geometric quantization of mass-energy

Charged geometries

All charge geometries are nett divergent

Divergent energy momenta	Convergent energy momenta
1	0
3	1
6	3
10	6
15	10
21	15
28	21
36	28

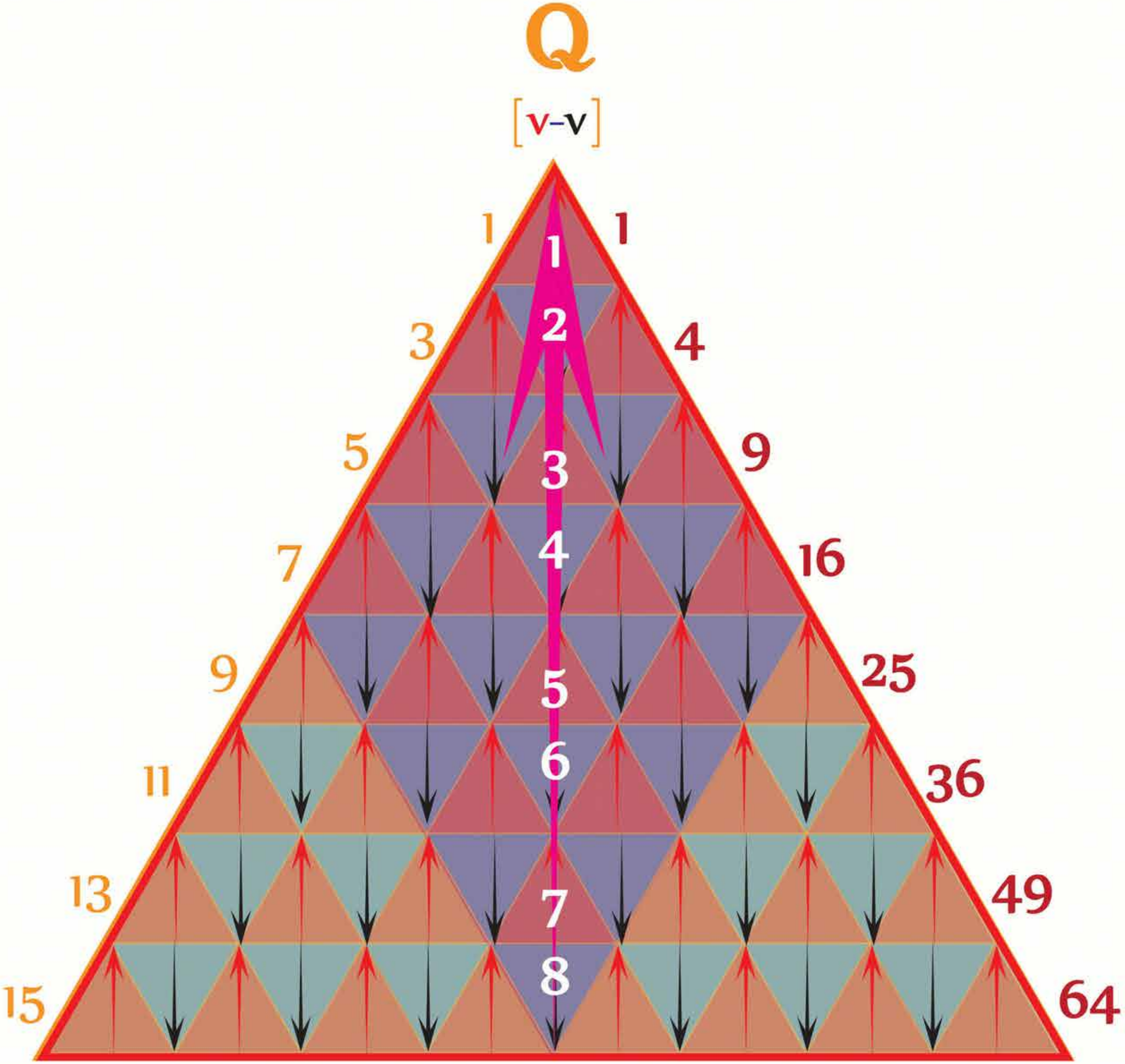
Q
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
1
2
3
4
5
6
7
8

All charge geometries are comprised of finite equilateral energy momenta quanta

Renormalisation

Renormalization was first developed in quantum electrodynamics (QED) to make sense of infinite integrals in perturbation theory.
The problem of infinities first arose in the classical electrodynamics of point particles in the19th and
later in the calculation of Gravitational fields in General Relativity in the early 20th century.

**In QED
Infinities must
be cancelled**

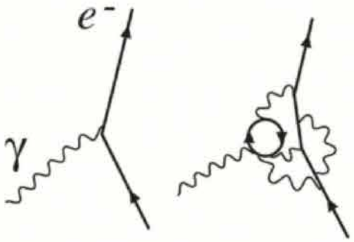

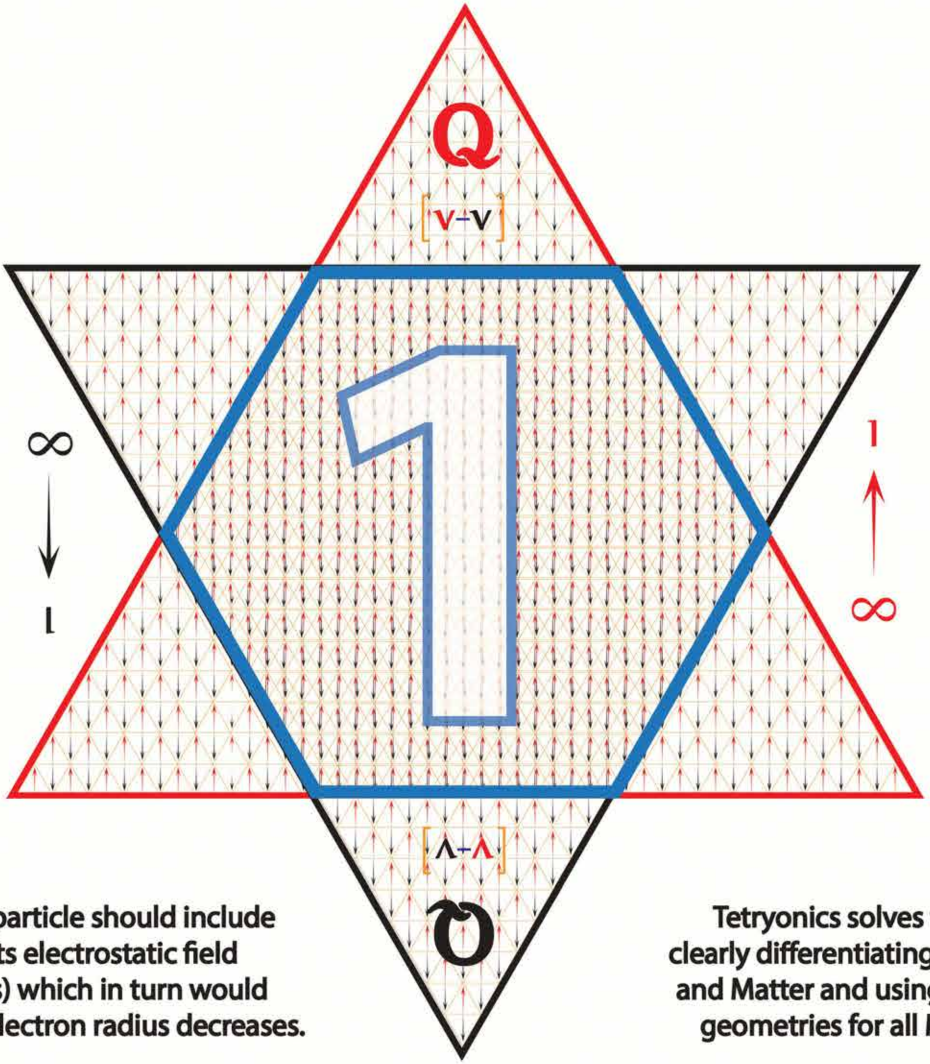

**In Tetryonics
Infinities do not
exist**

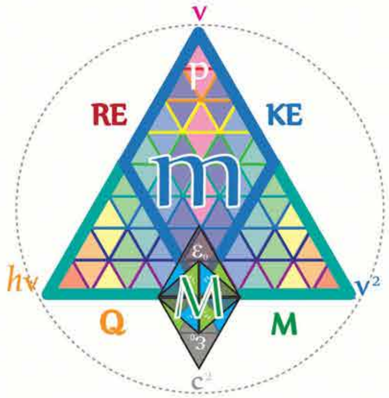

**The mass of a charged particle should include
the mass-energy in its electrostatic field
(Electromagnetic mass) which in turn would
approach infinity as the electron radius decreases.**

**Tetryonics solves the problem by
clearly differentiating between EM mass
and Matter and using finite equilateral
geometries for all Matter in motion**

Initially viewed as a suspicious provisional procedure by some of its originators, renormalization was eventually embraced as an important and self-consistent tool
in several fields of physics and mathematics.

2D space [c^2]

The adjective Cartesian refers to the French mathematician and philosopher René Descartes who developed the coordinate system in 1637

Since then many other coordinate systems have been developed such as the polar coordinates for the plane, and the spherical and cylindrical coordinates for three-dimensional space.

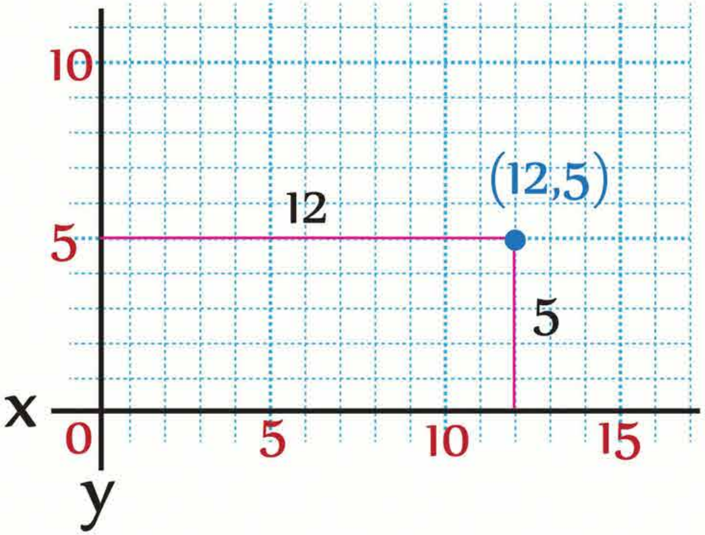

Cartesian coordinates are the foundation of analytic geometry, and provide enlightening geometric interpretations for many other branches of mathematics, such as linear algebra, complex analysis, differential geometry, multivariate calculus, group theory, and more

Mapping 3D spaces using Recti-linear co-ordinates

Cartesian coordinates can be defined as the positions of the perpendicular projections of a point onto the two or more axes, expressed as signed distances from the origin.

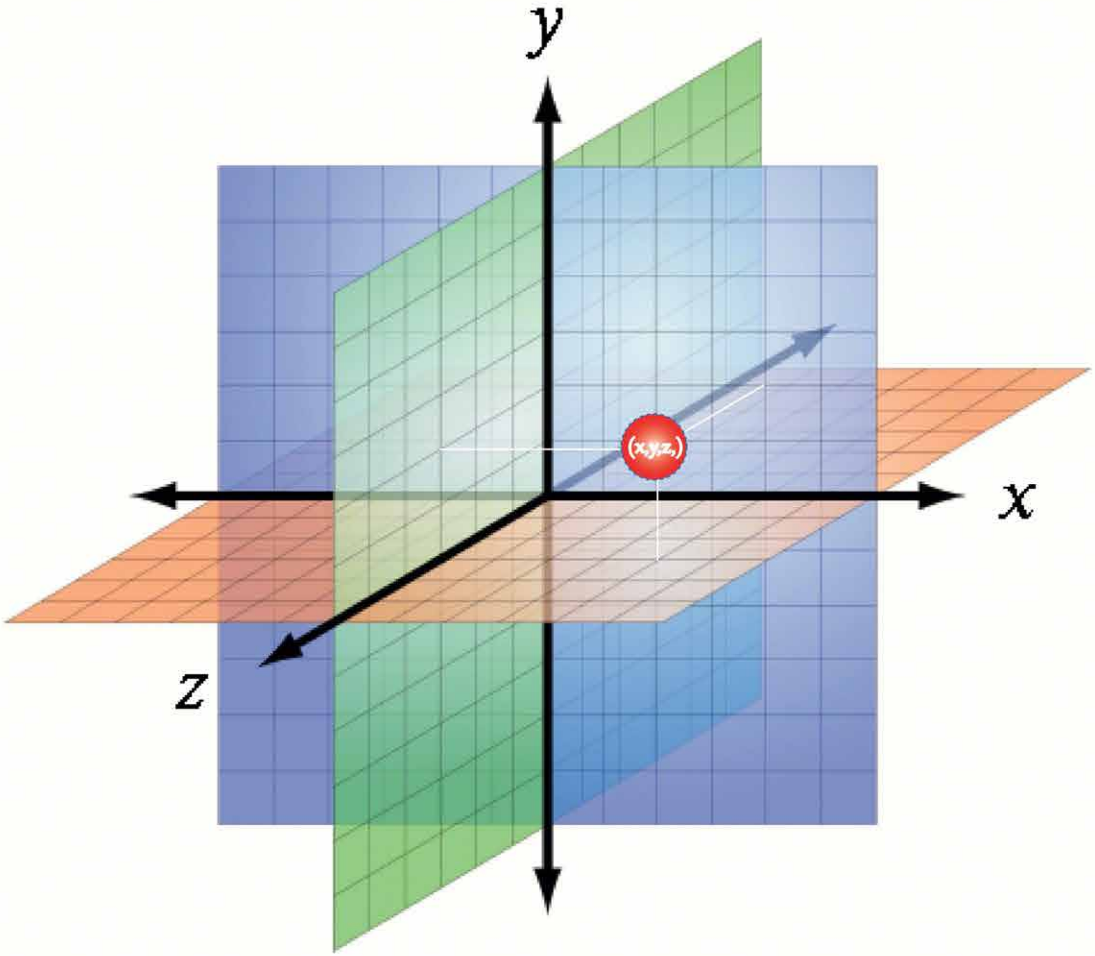

3D Cartesian co-ordinate [c^3] systems are distinct from spherical co-ordinate [c^4] systems

Copyright ABRAHAM [2008] - All rights reserved

Polar co-ordinates

In mathematics, the polar coordinate system is a two-dimensional co-ordinate system in which each point on a plane is determined by a distance from a fixed point and an angle from a fixed direction.

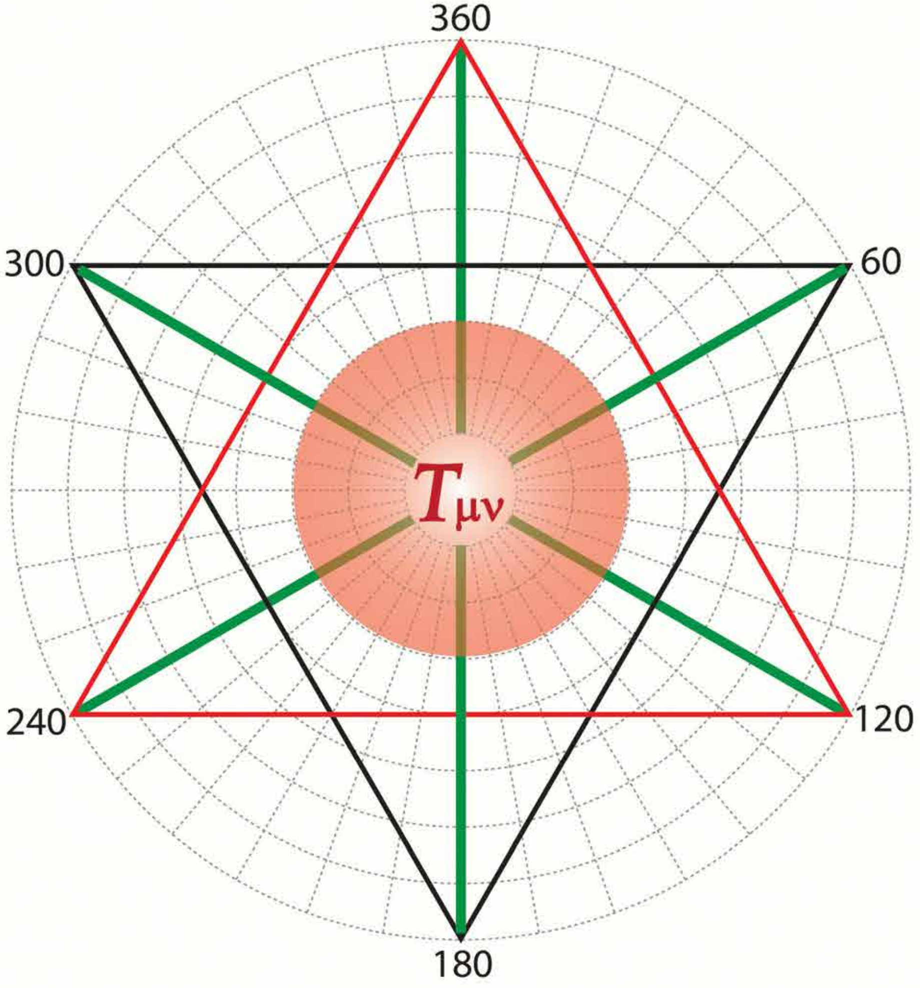

In geometry, curvilinear coordinates are a coordinate system for Euclidean space in which the co-ordinate lines may be curved.

Action Dynamics

Curvilinear co-ordinates may be derived from a set of rectilinear Cartesian coordinates by using a locally invertible transformation that maps one point to another in both systems

Metric Tensors

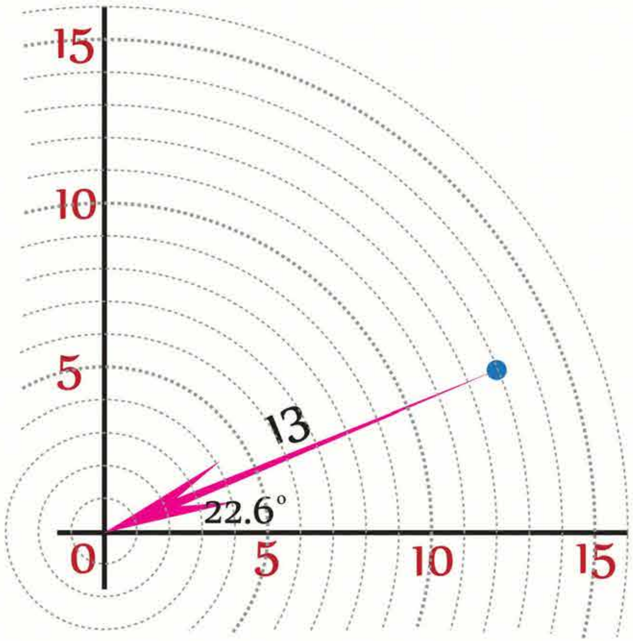

Gravitational acceleration

Polar or curvilinear co-ordinate systems are used extensively by Einstein in his theory of General Relavtivity

Reimannian curved space-time

Copyright ABRAHAM [2008] - All rights reserved

Co-ordinate transformations

There are many different possible coordinate systems for describing geometrical figures and they can all be related to one another.

Such relations are described by coordinate transformations which give formulas for the coordinates in one system in terms of the coordinates in another system

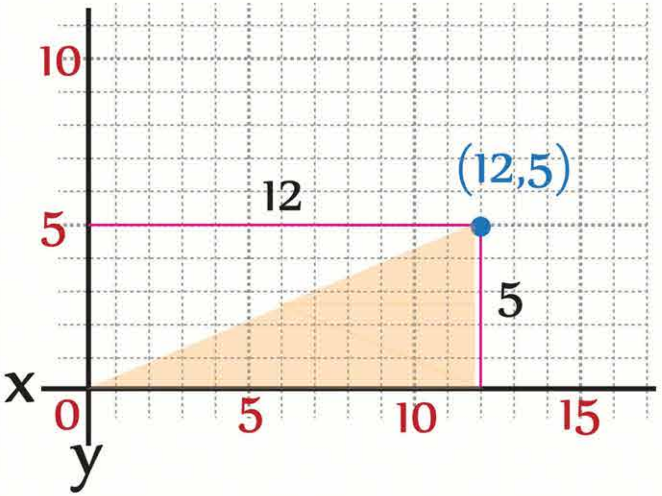

Converting between
Polar and Cartesian
coordinates

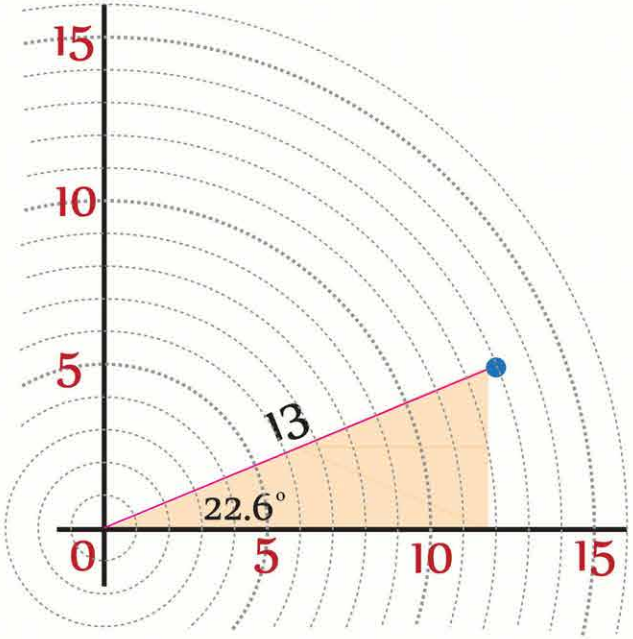

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}}$$

$$= \tan^{-1}(5/12)$$

$$= 22.619$$

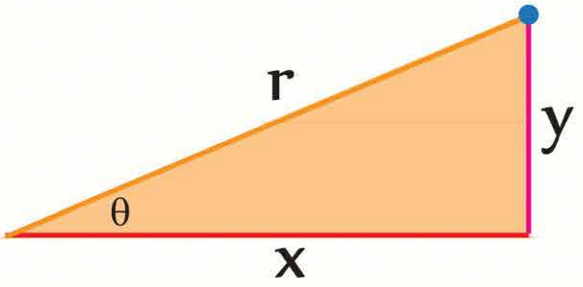

$$r^2 = 12^2 + 5^2$$

$$= \sqrt{169}$$

$$= 13$$

solving for the triangle
using trigometric functions

Spatial co-ordinate systems

Cubic

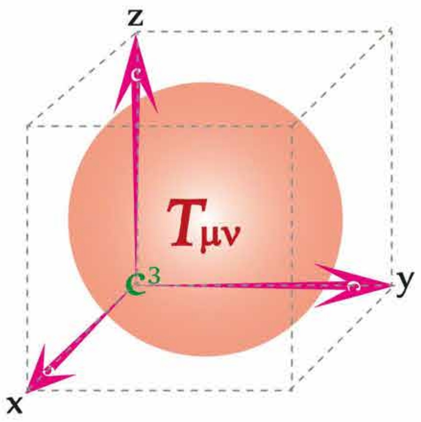

Spacetime is any mathematical co-ordinate system or model that combines space and time into a single continuum.

Spacetime is usually interpreted with space as being three-dimensional with time playing the role of a fourth dimension that is different from the spatial dimensions.

From a Euclidean space perspective, the universe has three spatial dimensions and one dimension of time [reflected by quantised angular momentum].

Spherical

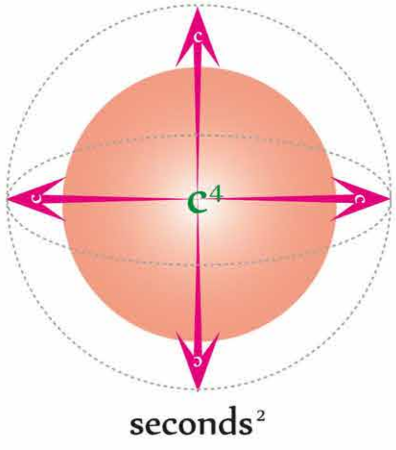

Tetryonics maps spatial co-ordinates through the momenta vectors of equilateral Energy

Cartesian Space-Time

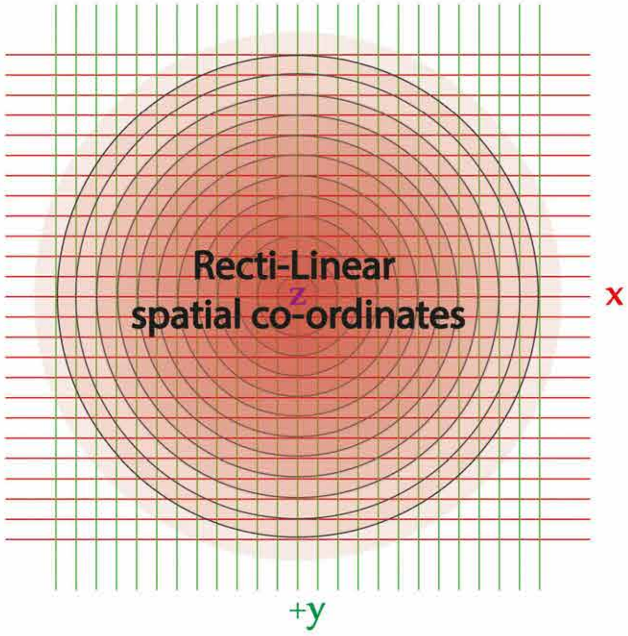

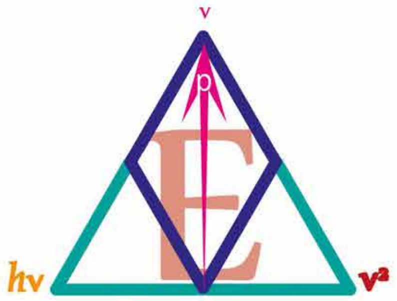

EM mass-ENERGY momenta are equilateral geometries

Tetryonic Space-Time

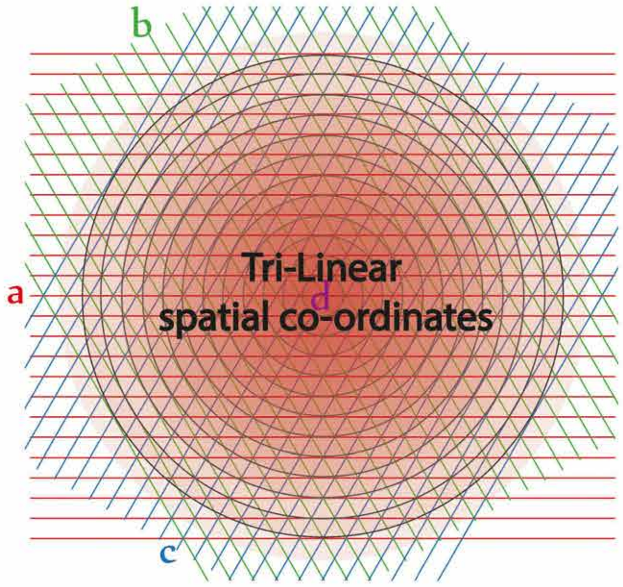

Mapping equilateral Energy geometries onto recti or curvi-linear spatial co-ordinate systems introduces mathematical complexity to a otherwise simplistic geometry for all EM mass-Energy-Matter interactions

In physics spatial co-ordinates to date have been based on Cartesian co-ordinates when in fact Energy momenta follow a Tri-Linear co-ordinate geometry

vectors

1 Dimensional velocity

Euclidean

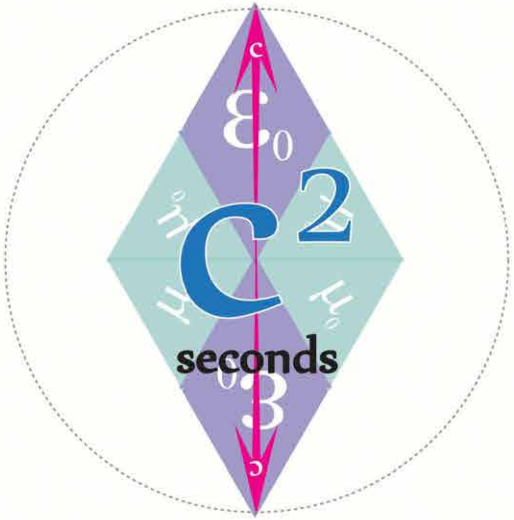

2 Dimensions velocity squared

$$m = \frac{E}{c^2}$$

mass-energy equivalence

mass-Matter

$$\frac{m}{c^2} = M$$

equivalence

$$T_{\mu\nu}$$

energy-Matter equivalence

$$M = \frac{E}{c^4}$$

GR fails differentiate betweem EM mass-ENERGY-Matter unlike Tetryonic theory

Cartesian co-ordinates

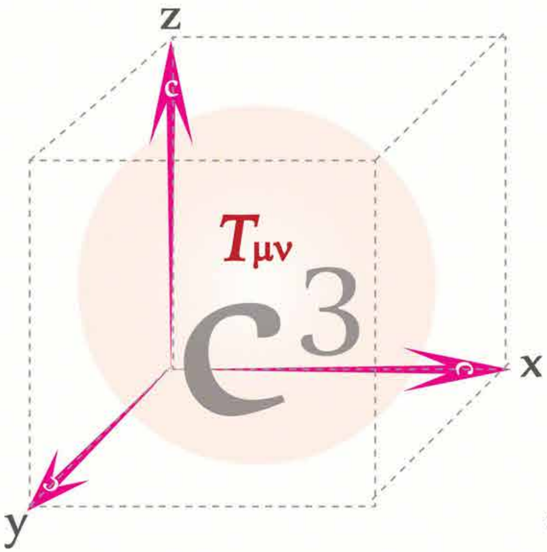

3 Dimensions velocity cubed

Spherical

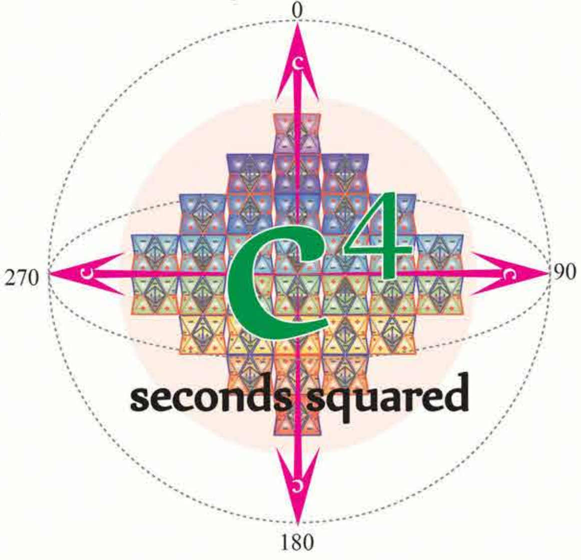

3 Dimensions quadrature velocity

Space-time co-ordinates

The propagation of Energy momenta forms distinct spatial co-ordinate systems

Where angles are typically measured in degrees (°) or radians (rad), where 360° = 2π rad = τ radians.

Tetryonic co-ordinate systems

Energy per second

m

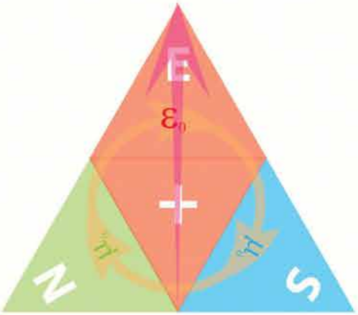

2D mass-energy geometries

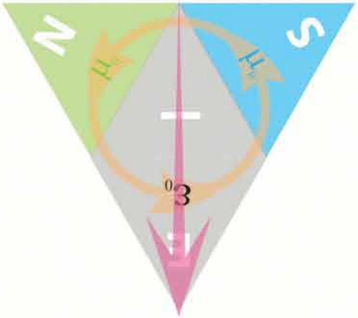

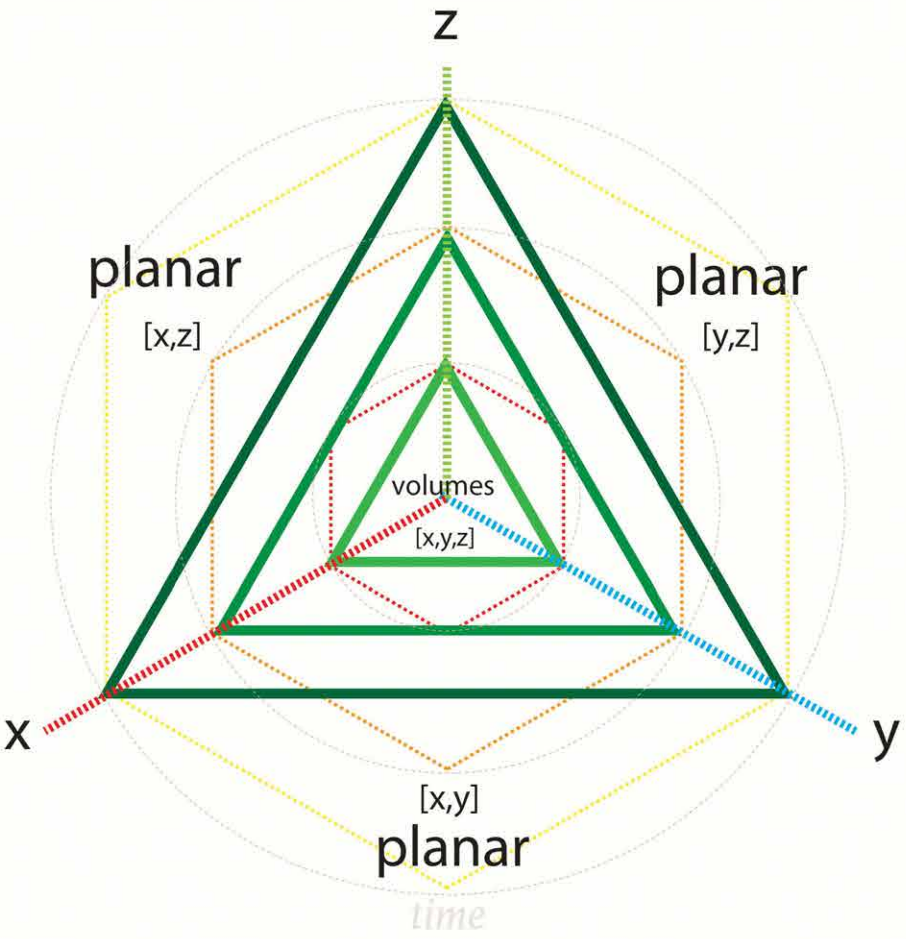

z
planar
[x,z]
planar
[y,z]
volumes
[x,y,z]
x
y
[x,y]
planar
time

Planar mass-energy geometries have no z-components

Differentiation between 2D mass-energy & 3D Matter is key to extending our understanding of physics

Matter topologies have z-components

Energy per second squared

M

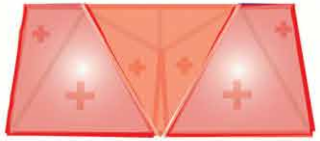

Proton

3D Matter topologies

Neutron

Copyright ABRAHAM [2008] - All rights reserved

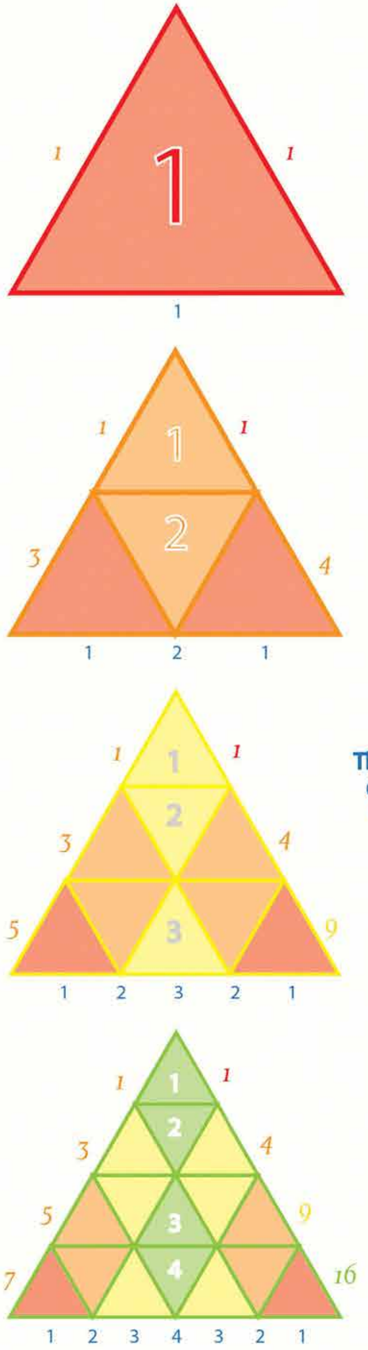
1
1 1
1
1
2
1 1
3 4
1 2 1
1
2
3
1 1
3 4
5 9
1 2 3 2 1
1
2
3
4
1 1
3 4
5 9
7 16
1 2 3 4 3 2 1

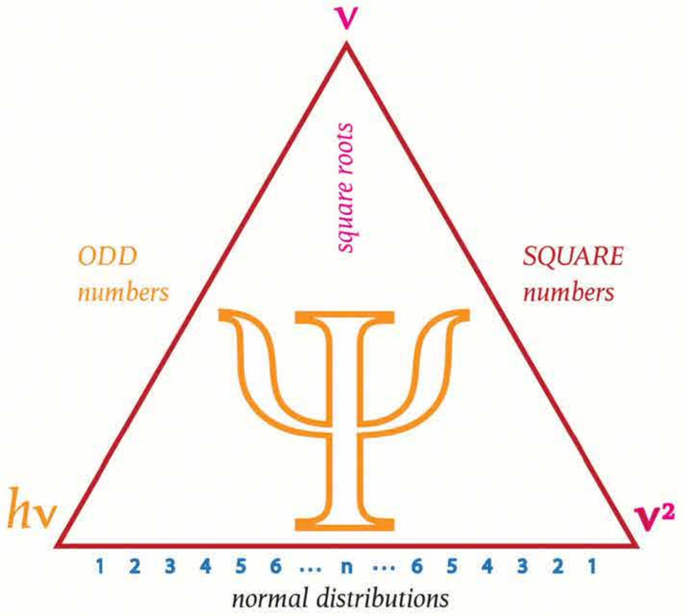
v
square roots
ODD numbers
SQUARE numbers
hv
v²
1 2 3 4 5 6 … n … 6 5 4 3 2 1
normal distributions

Quantum probability distributions

The equilateral geometry and distribution of quantised Energy momenta provides the basis for all statistical probabilities in Quantum mechanics, thermodynamic & information entropy - including a solution to Heisenberg's Uncertainty Principle thus paving the way forward for a new understanding, and manipulation of physical phenomena at the quantum level

1
1 2 1
1 2 3 2 1
1 2 3 4 3 2 1
1 2 3 4 5 4 3 2 1
1 2 3 4 5 6 5 4 3 2 1
1 2 3 4 5 6 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Normal distributions are extremely important in statistics, and are often used in the natural and social sciences for real-valued random variables whose distributions are not known

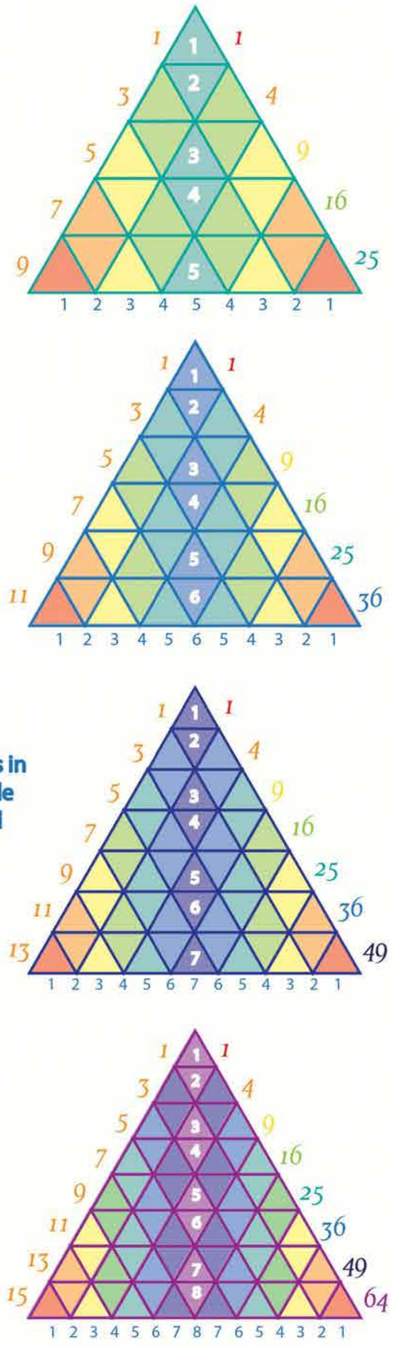
1 1
3 4
5 9
7 16
9 25
1 2 3 4 5 4 3 2 1
1 1
3 4
5 9
7 16
9 25
11 36
1 2 3 4 5 6 5 4 3 2 1
1 1
3 4
5 9
7 16
9 25
11 36
13 49
1 2 3 4 5 6 7 6 5 4 3 2 1
1 1
3 4
5 9
7 16
9 25
11 36
13 49
15 64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Quantum Probability Distributions

The normal distribution is a probability distribution.
It is also called Gaussian distribution because it was discovered by Carl Friedrich Gauss.

x/n^2

Probabilities are the square of the Amplitude

n

square root momenta

v

$\Sigma\ x/n^2$

All probabilities are re-normalisable and sum to Unity

1

ODD transverse Bosons

EVEN Longitudinal Photons

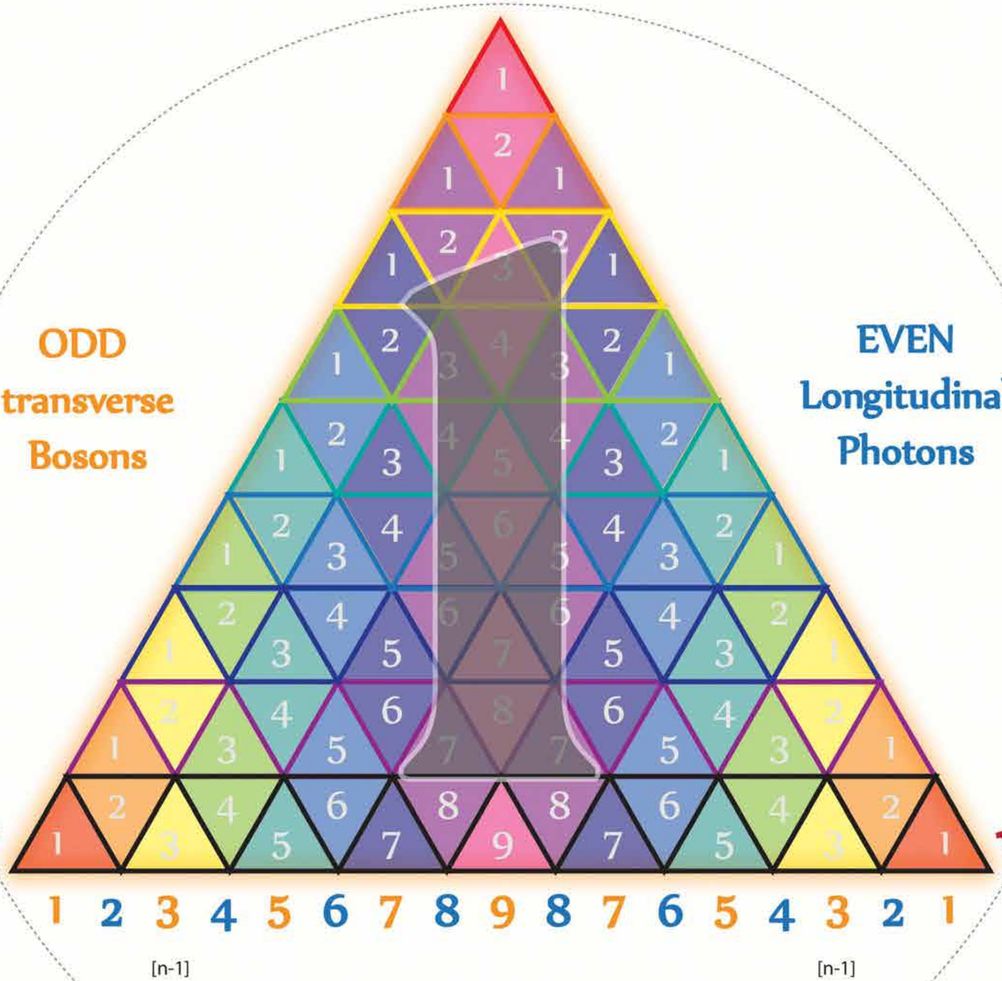
1
2
1 1
3
1 2 2 1
4
1 2 3 3 2 1
5
1 2 3 4 4 3 2 1
6
1 2 3 4 5 5 4 3 2 1
7
1 2 3 4 5 6 6 5 4 3 2 1
8
1 2 3 4 5 6 7 7 6 5 4 3 2 1
9
1 2 3 4 5 6 7 8 8 7 6 5 4 3 2 1
1 2 3 4 5 6 7 8 9 8 7 6 5 4 3 2 1
[n-1]
[n-1]

Planck quanta

hv

v^2

Square energies

Normal Distribution

c^2

Copyright ABRAHAM [2008] - All rights reserved

Quanta Distributions

Quantum Levels

transverse bosons

$$2h\nu = E = hf$$

Quantum Probabilities

longitudinal photons

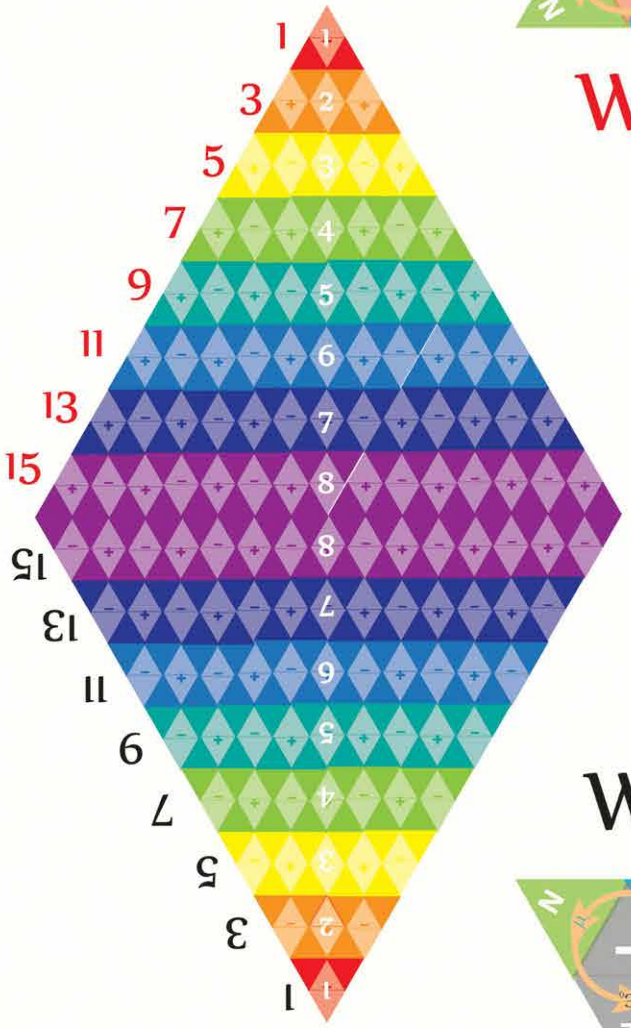
1
3
5
7
9
11
13
15
15
13
11
9
7
5
3
1
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1

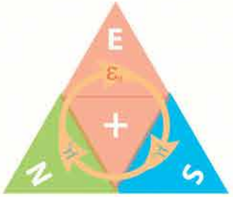

W+

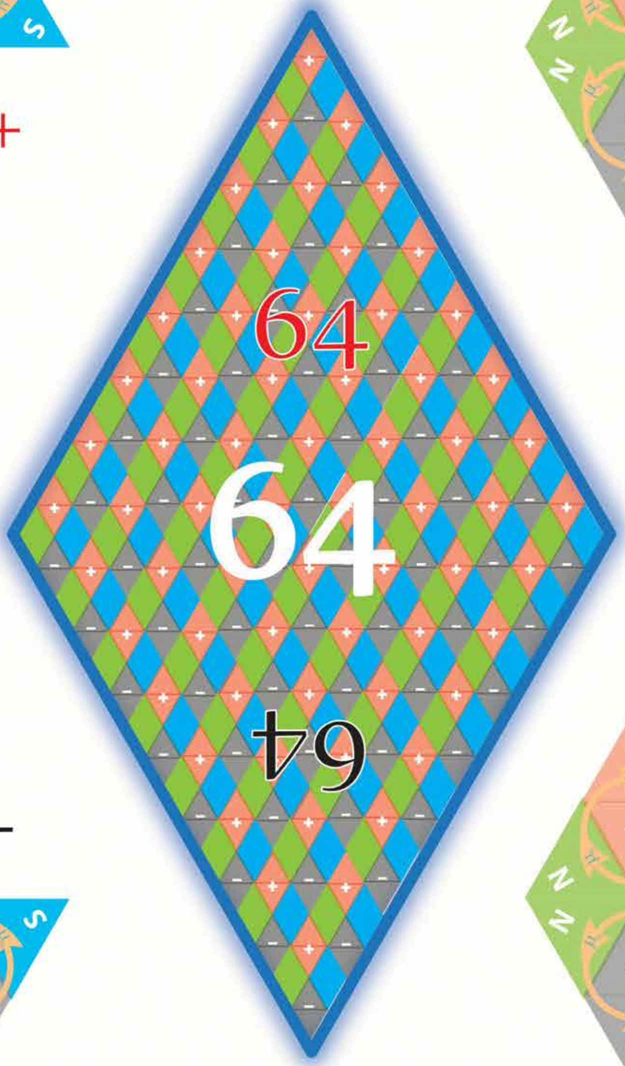

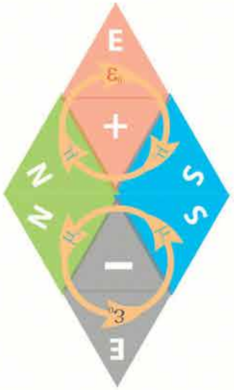

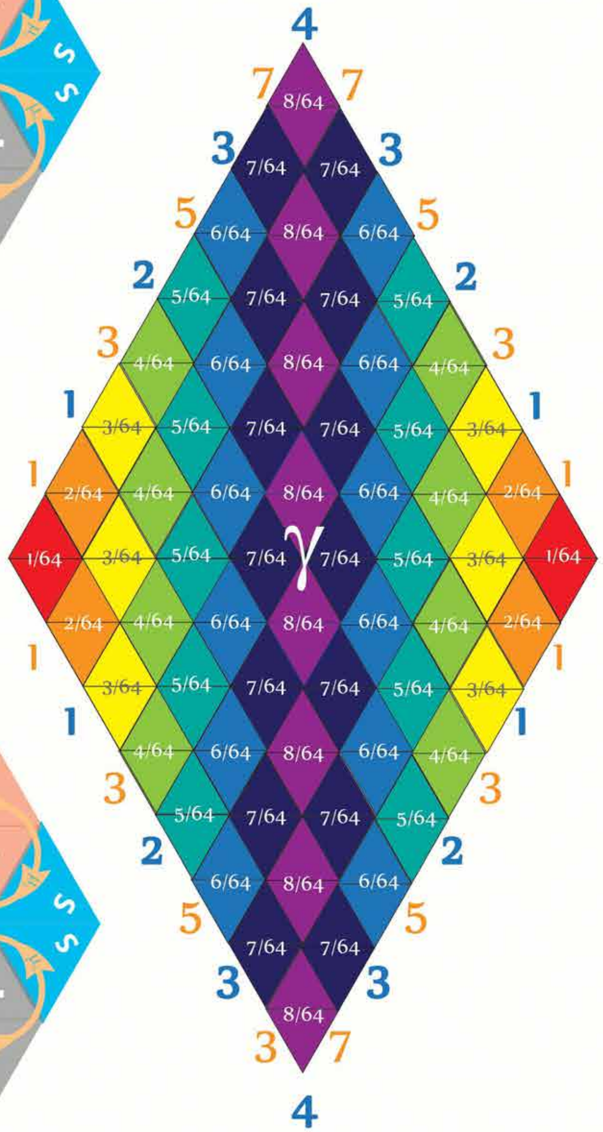
4
7
7
8/64
3
3
7/64
7/64
5
5
6/64
8/64
6/64
2
2
5/64
7/64
7/64
5/64
3
3
4/64
6/64
8/64
6/64
4/64
1
1
3/64
5/64
7/64
7/64
5/64
3/64
1
1
2/64
4/64
6/64
8/64
6/64
4/64
2/64
1/64
3/64
5/64
7/64
γ
7/64
5/64
3/64
1/64
1
1
2/64
4/64
6/64
8/64
6/64
4/64
2/64
1
1
3/64
5/64
7/64
7/64
5/64
3/64
3
3
4/64
6/64
8/64
6/64
4/64
2
2
5/64
7/64
7/64
5/64
5
5
6/64
8/64
6/64
3
3
7/64
7/64
3
7
8/64
4

W-

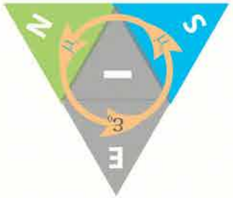

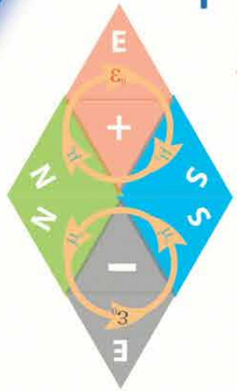

Copyright ABRAHAM [2008] - All rights reserved

Wavefunctions

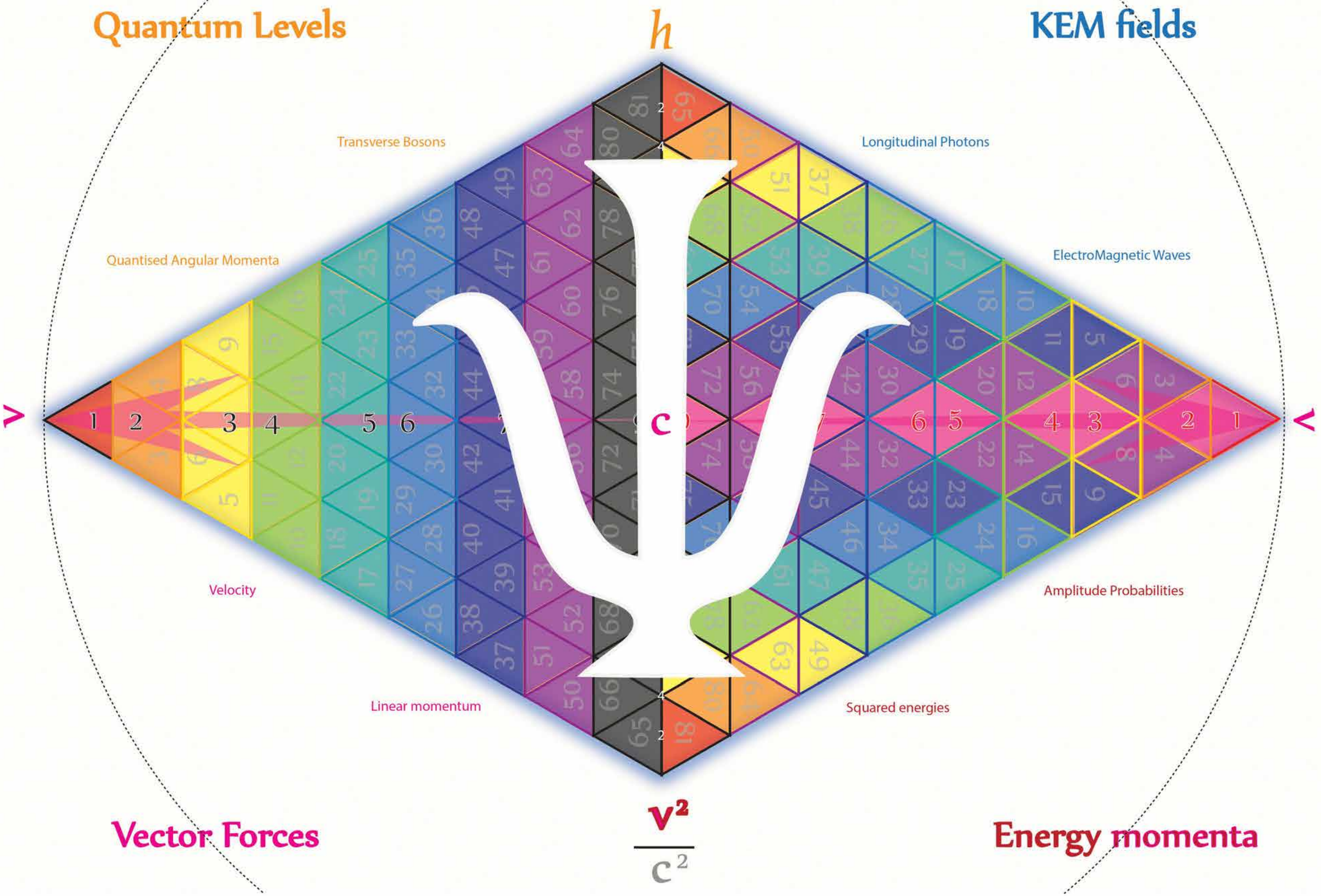
Quantum Levels
h
KEM fields
Transverse Bosons
Longitudinal Photons
Quantised Angular Momenta
ElectroMagnetic Waves
v
c
v
Velocity
Amplitude Probabilities
Linear momentum
Squared energies
Vector Forces
v²
c²
Energy momenta

Copyright ABRAHAM [2008] - All rights reserved

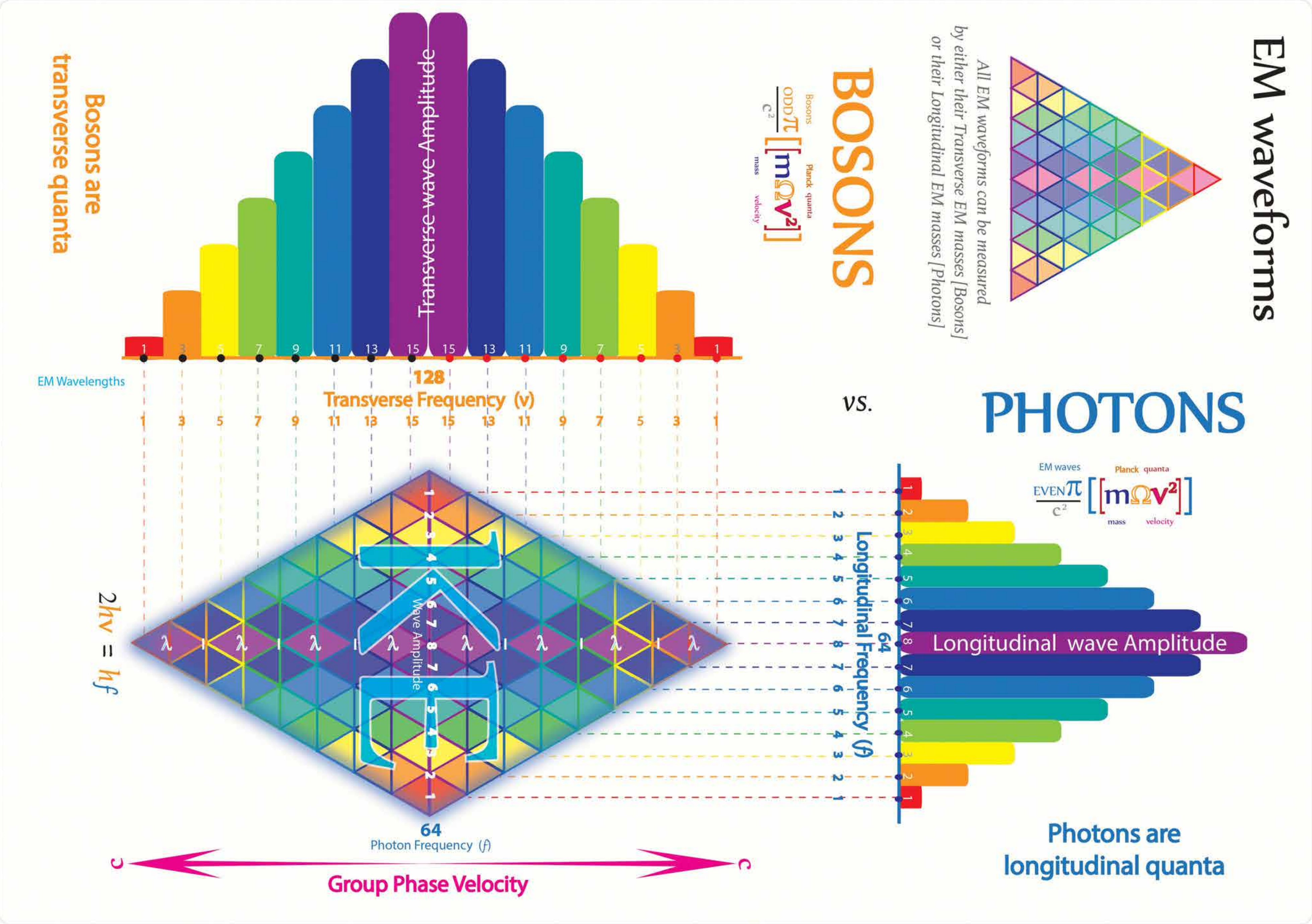
EM waveforms
All EM waveforms can be measured
by either their Transverse EM masses [Bosons]
or their Longitudinal EM masses [Photons]
BOSONS
Bosons
Planck quanta
$\frac{\text{ODD}\pi}{c^2}[[m\Omega v^2]]$
mass
velocity
vs.
PHOTONS
EM waves
Planck quanta
$\frac{\text{EVEN}\pi}{c^2}[[m\Omega v^2]]$
mass
velocity
Bosons are transverse quanta
Transverse wave Amplitude
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
EM Wavelengths
128
Transverse Frequency (v)
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
$2hv = hf$
λ λ λ λ λ λ λ λ
KE
Wave Amplitude
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Photon Frequency (f)
c
c
Group Phase Velocity
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
64
Longitudinal Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Longitudinal wave Amplitude
Photons are longitudinal quanta

Quantum Energy distributions

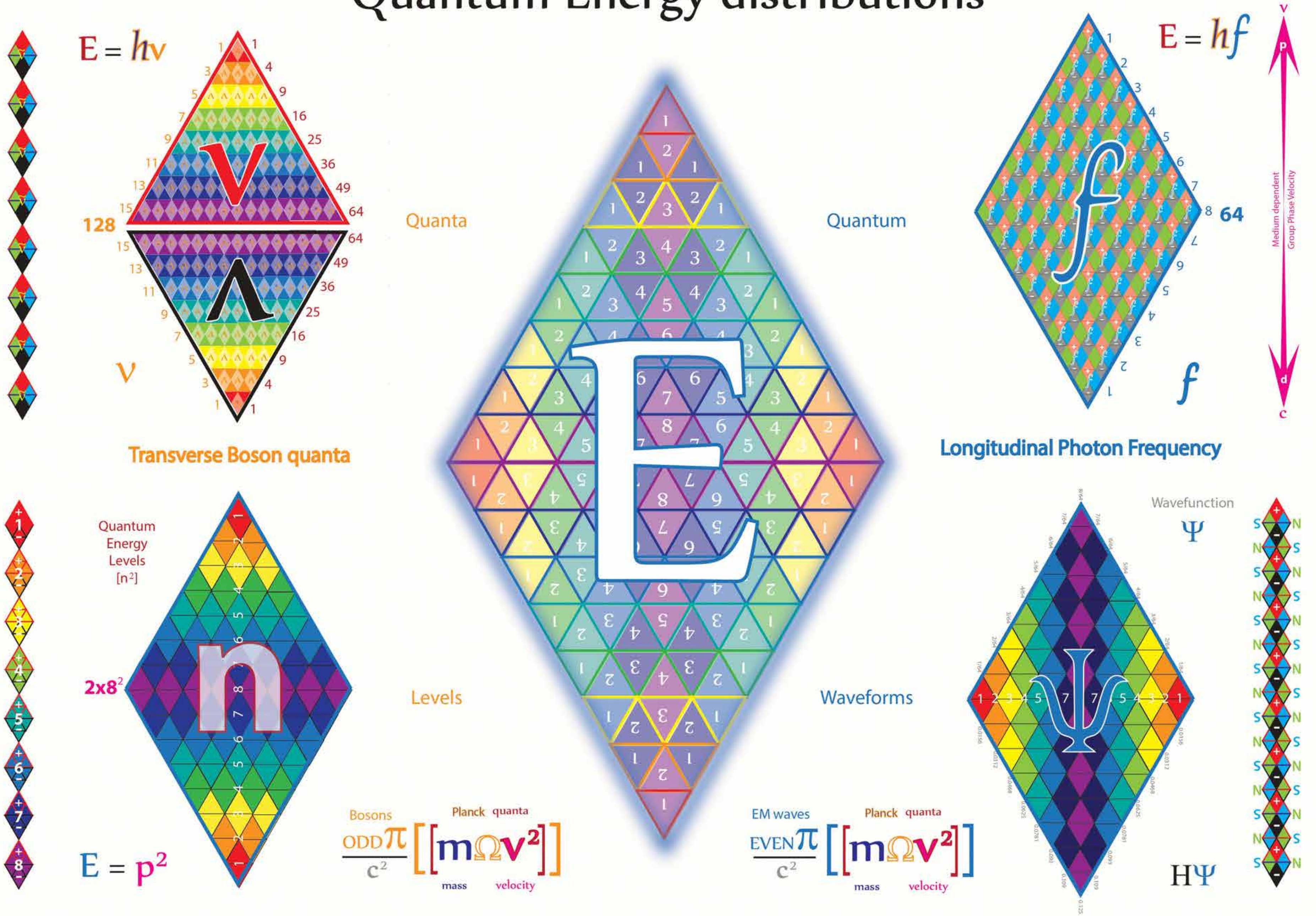

$E = h\nu$

Quanta

Transverse Boson quanta

Quantum

$E = hf$

f

Longitudinal Photon Frequency

Medium dependent
Group Phase Velocity

Quantum
Energy
Levels
$[n^2]$

2×8^2

Levels

$E = p^2$

Bosons Planck quanta

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Wavefunction

Ψ

Waveforms

EM waves Planck quanta

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

$H\Psi$

Copyright ABRAHAM [2008] - All rights reserved

Normal Distributions

The Gaussian distribution sometimes informally called the bell curve.

Pierre de Fermat

Pierre de Fermat is given credit for early developments that led to infinitesimal calculus.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

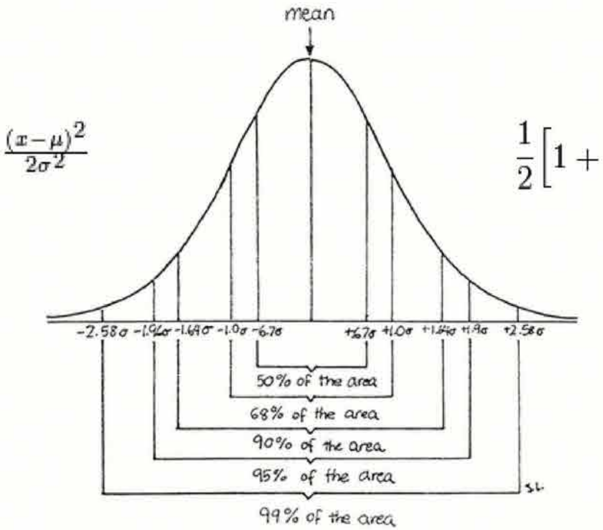

$$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$$

Leonhard Euler

Leonhard Euler developed a formula which links complex exponentiation with trigonometric functions

A bell shaped curve defines the standard normal distribution, in which the probability of observing a point is greatest near the average, and declines rapidly as one moves away from the mean.

all ODD numbers are

$$a^2 - b^2 = [a-b].[a+b]$$

the difference of two squares

$$e^{i\pi} = -1$$

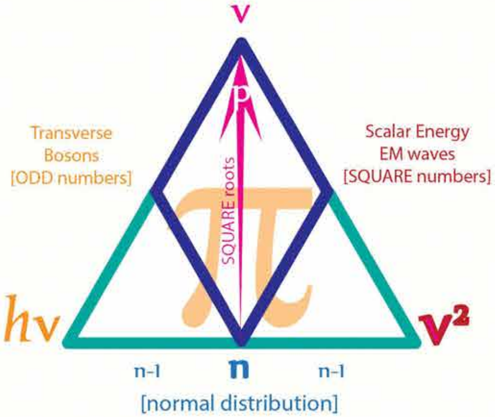
v
p
SQUARE roots
π
Transverse
Bosons
[ODD numbers]
Scalar Energy
EM waves
[SQUARE numbers]
hv
v²
n-1
n
n-1

[normal distribution]

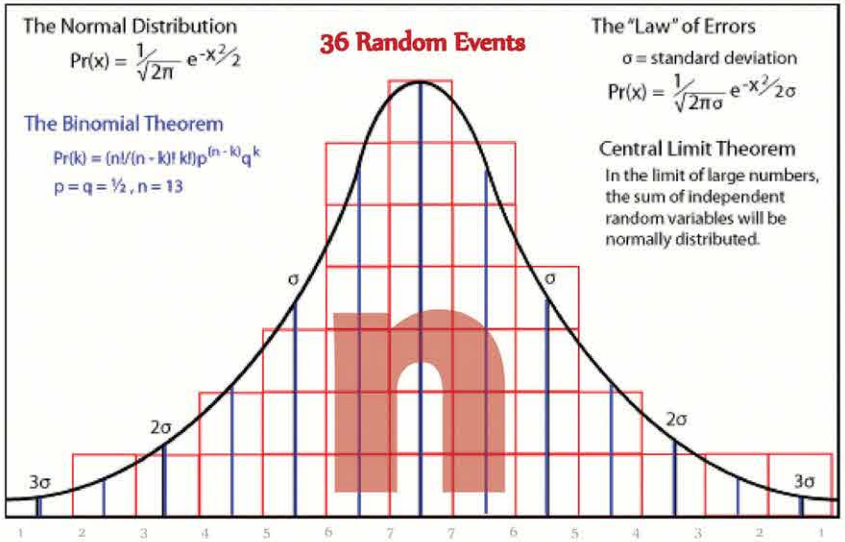

The Normal Distribution

$$\Pr(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

The Binomial Theorem

$\Pr(k) = (n!/(n-k)!\,k!)p^{(n-k)}q^k$

$p = q = ½ , n = 13$

The "Law" of Errors

σ = standard deviation

$$\Pr(x) = \frac{1}{\sqrt{2\pi\sigma}}\, e^{-x^2/2\sigma}$$

Central Limit Theorem

In the limit of large numbers, the sum of independent random variables will be normally distributed.

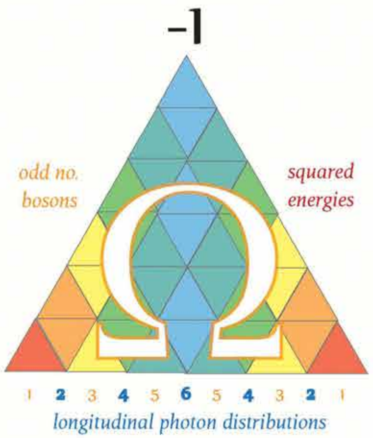
-1
odd no.
bosons
squared
energies
Ω
1
2
3
4
5
6
5
4
3
2
1

longitudinal photon distributions

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution that has a bell-shaped probability density function, known as the Gaussian function

Copyright ABRAHAM [2008] - All rights reserved

Fundamental theorem of Energy momenta

A nth level scalar energy momenta waveform has exactly n linear momentum in unit circle co-ordinate systems
(with Longitudinal and Transverse equilateral Planck waveforms being orthogonal to each other)

scalar energies

E

mv^2

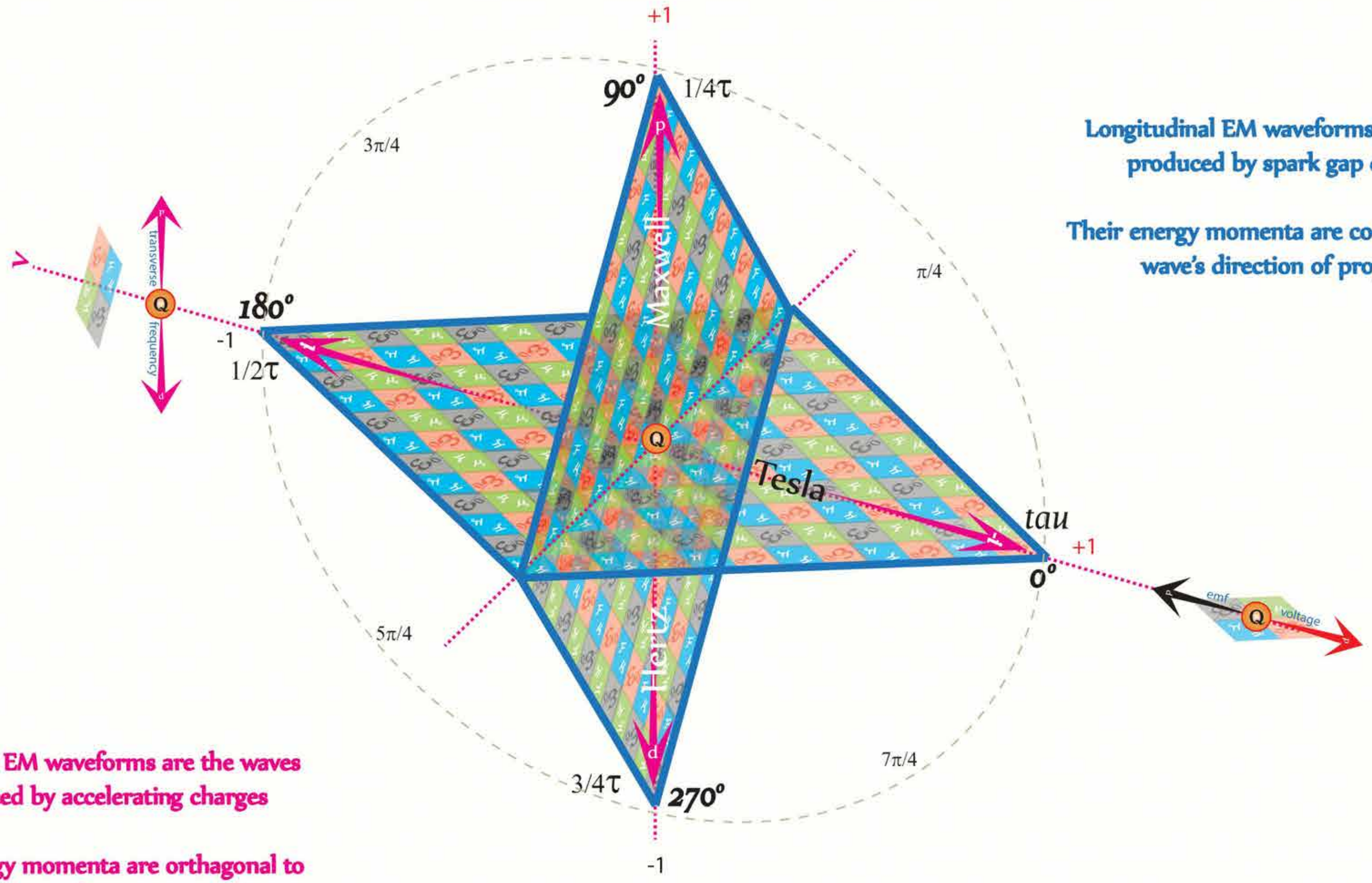
+1
90°
1/4τ
3π/4
π/4
transverse
Q
frequency
180°
-1
1/2τ
Maxwell
Q
Tesla
tau
+1
0°
emf
Q
voltage
5π/4
Hertz
7π/4
3/4τ
270°
-1

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthogonal to the direction of wave propagation

[Mathematics] ***i or j*** *[Electrical theory]*

square root of negative one

p

mv

linear momentum

Velocity and Time dependent EM fields

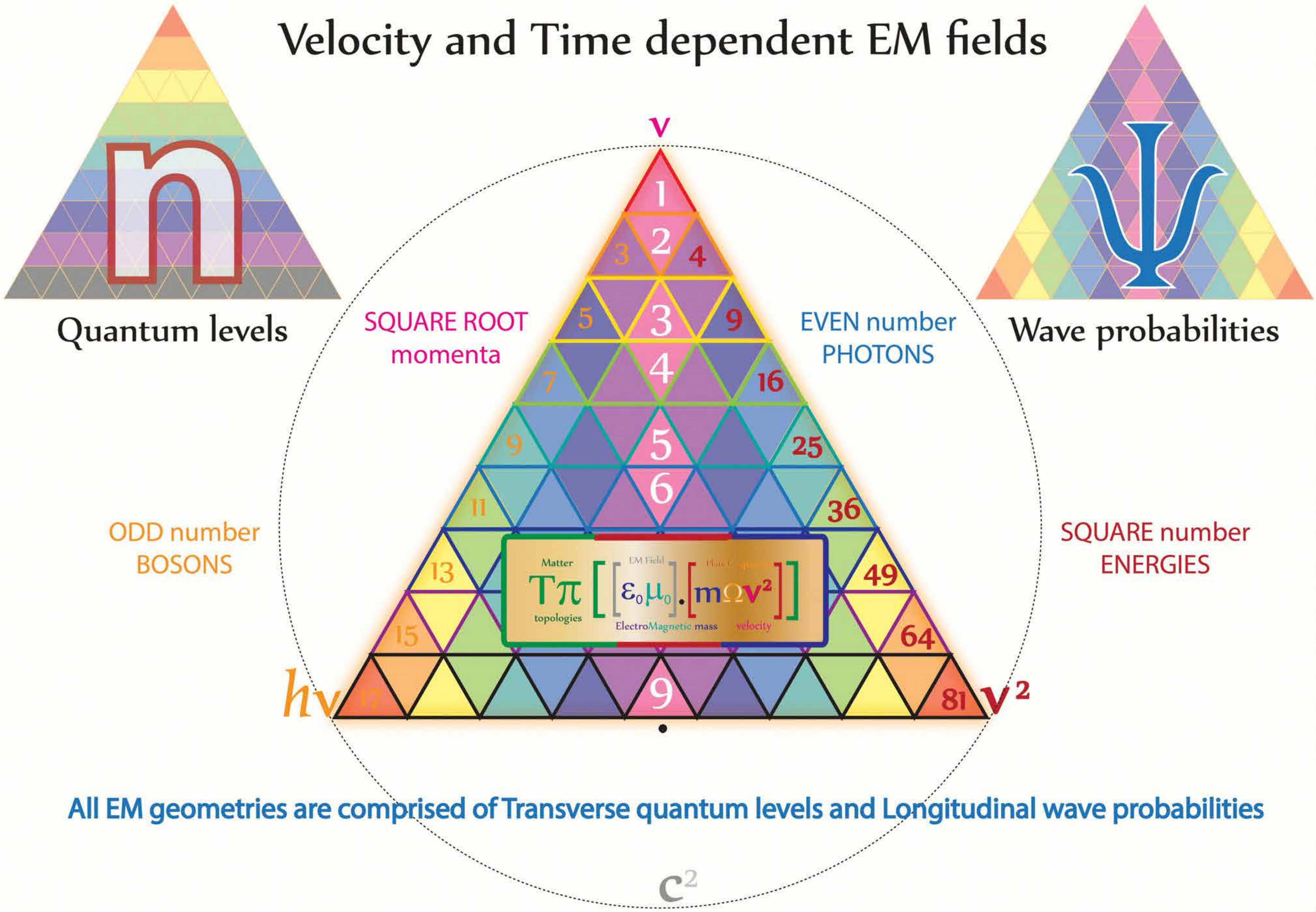
n
Quantum levels
Ψ
Wave probabilities
v
1
2
3
4
3
5
9
4
7
16
5
9
25
6
11
36
13
49
15
64
9
17
81
hv
v²
c²
SQUARE ROOT momenta
EVEN number PHOTONS
ODD number BOSONS
SQUARE number ENERGIES
Matter
Tπ
topologies
EM Field
Planck quanta
ElectroMagnetic mass
velocity

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

All EM geometries are comprised of Transverse quantum levels and Longitudinal wave probabilities

Matrices

*Matrices are a key tool in linear algebra.
One use of matrices is to represent linear transformations,
which are higher-dimensional analogs of linear functions
of the form f(x) = cx, where c is a constant; matrix multiplication
corresponds to composition of linear transformations.*

*Further developing equilateral Matrices and tensor mathematics to
reflect the 2D geometry of EM , KEM and GEM quantum fields,
along with the geometric quantisation of mass-energiy momenta
and their energy distributions allows for field interactions
to be accurately visualised and modelled*

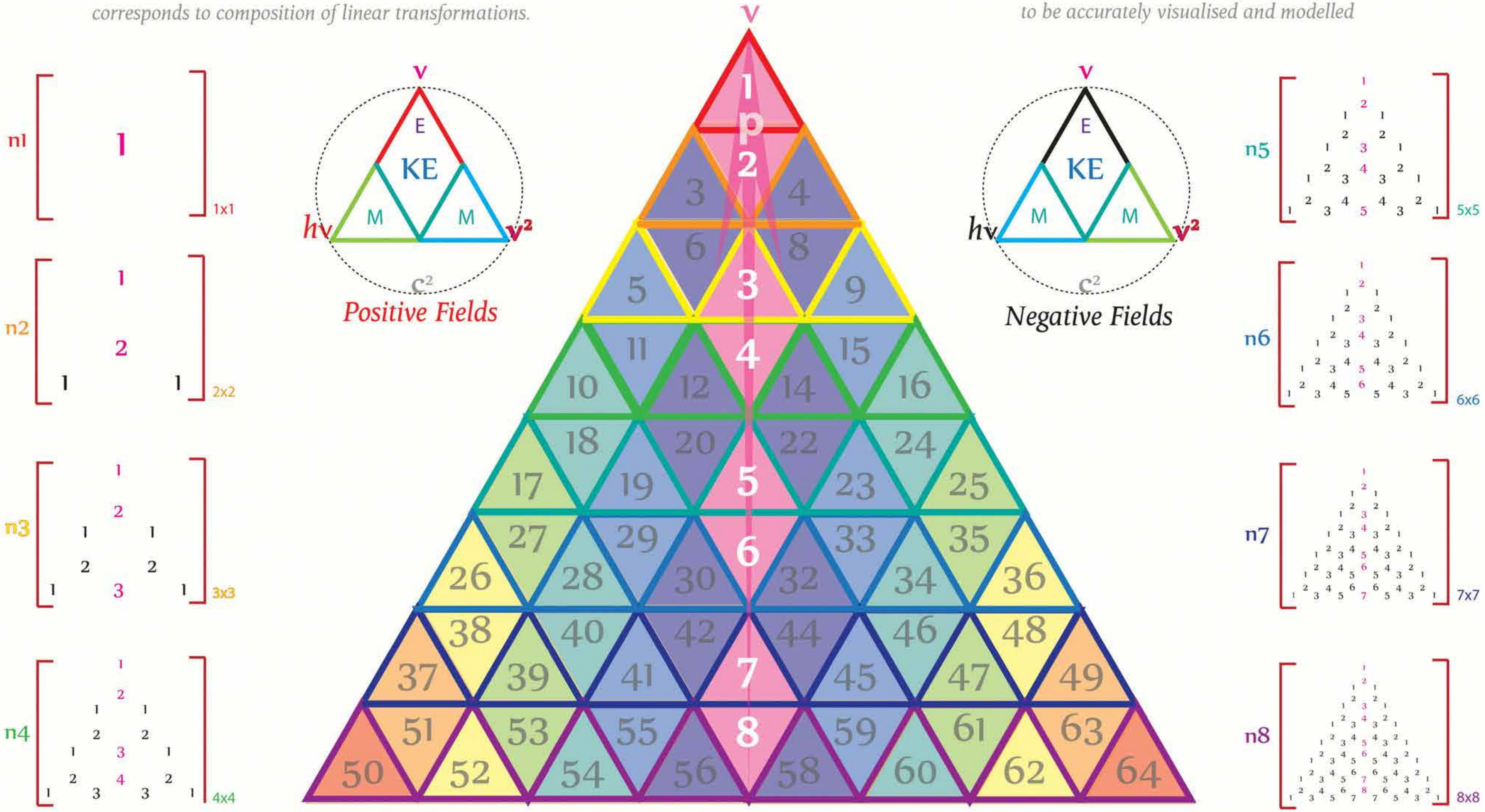

Tensors are geometric entities introduced into mathematics and physics to extend the notion of scalars, geometric vectors, and matrices to increasingly higher orders.

Modifying Square matrices to reflect the equialteral geometries of Tetryonic fields allows for the accurate geometric modelling of all Scalar & Vector fields along with their varied intrinsic quantum energies and physical properties

Copyright ABRAHAM [2008] - All rights reserved

Energy momenta Tensors

momenta
(a property of Energy)
is converative

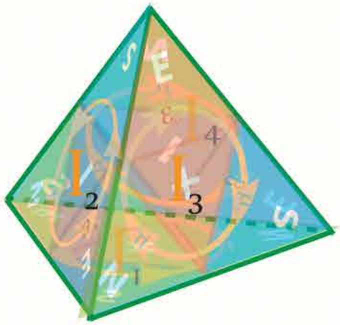

$$p_M \longrightarrow (E, p1, p2, p3, p4).$$

All standing-wave Matter topologies can be modelled using its Tetryonic charge energy momenta Tensors

with an additional Kinetic EM energy-momenta tensor required for Matter in motion

$$p_{KEM} \longrightarrow (E, p5).$$

Matter
(a geometric property)
is NOT conservative

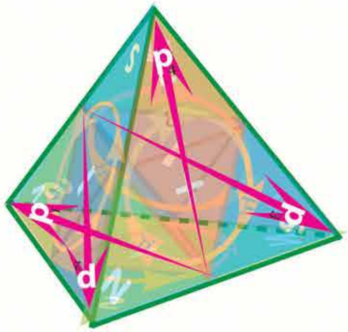

2D+1 [SR] mass-energy momenta can be folded into 3D+1 [GR] Matter that can be modelled using 4 Energy-momenta tensors

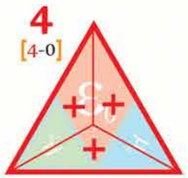

Positive standing-wave Matter

$$E = \sum_{\text{all fascia}} E$$

Energy of a massive particle is the total of all Planck quanta (compton frequency) in a 3D geometry

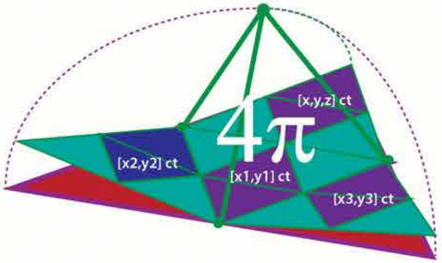

$$\vec{p} = \sum_{\text{all fascia}} \vec{p}$$

Total Momentum is the total of all quanta linear Momenta in a 3D particle

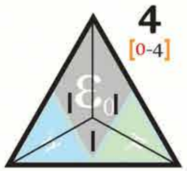

Negative standing-wave Matter

Relativistic Matter in motion

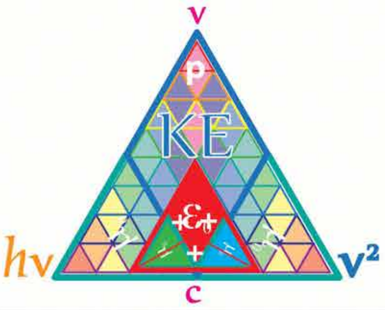

$$I = \sum_{\text{all fascia}} I$$

Inertial mass is the total of all inertia in a 3D particle

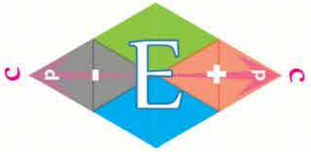

Photons have null energy-momentum tensors

Note that both 2D mass-energies [Special Relativity] and 3D Matter [General Relativity] have distinct Energy momenta

It is the 3D Tetrahedral topologies that provides a definative basis for Matter

Relativistic Matter in motion

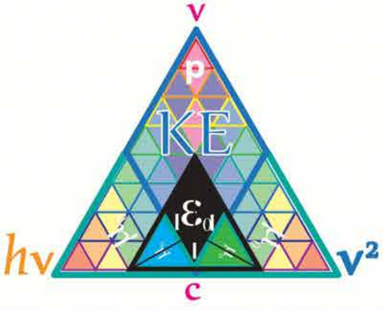

Tetryonic Energy momenta tensors should not be confused with Four vector tensors which map energy-momenta vectors in 3D spatial[cartesian] co-ordinate systems

Relativistic Lorentz correction factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity
It is often incorrectly graphed [as pictured to the right] as an increasing function of velocity approaching but never reaching c

'c' is the maximum velocity acheivable through the electrical acceleration of particles
is NOT the maximum velocity possible for Matter in motion

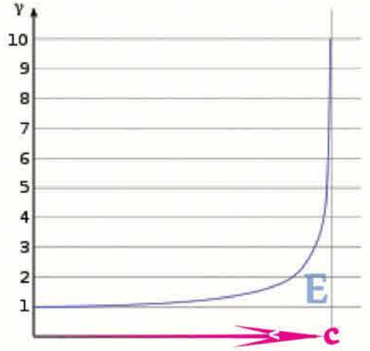

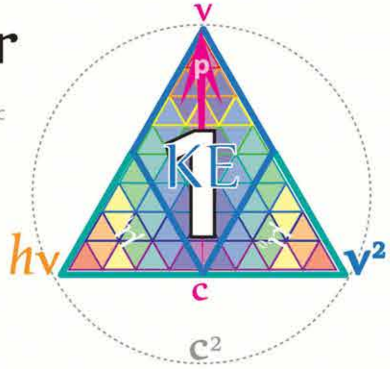

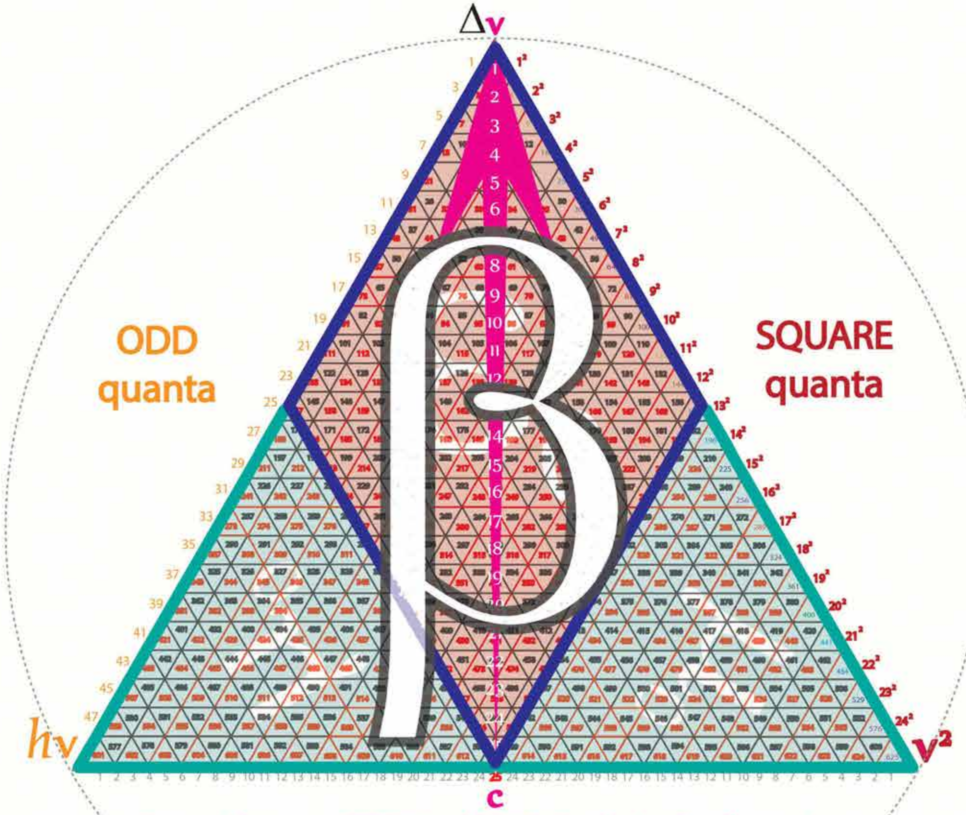

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Lorentz factors are LINEAR and SCALAR velocity related corrections
to the relativistic mass-energy momenta content of any
physical system accelerated by EM Forces

$$\frac{\Delta\gamma}{c^2}$$ Wavelength contraction

Time Dilation $$\frac{\Delta\Omega}{c^2}$$

$$\frac{\Gamma}{c^2}$$ Lorentz corrections

$$\frac{\Delta mv}{c^2}$$ Relativistic momentum

Relativistic mass-energy $$\frac{\Delta hv}{c^2}$$

Equilateral energies

Any measurements of mass without a Space-Time co-ordinate system are measurements of Energy

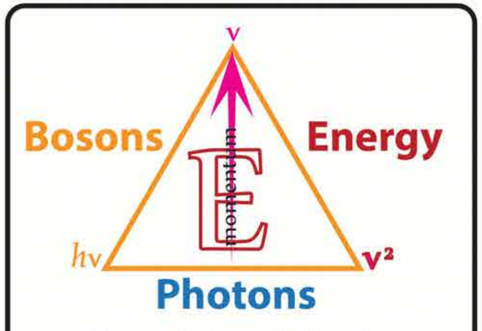
v
Bosons
Energy
E
momentum
hv
v²
Photons

form the basis for all charged energy geometries and their associated quanta distributions

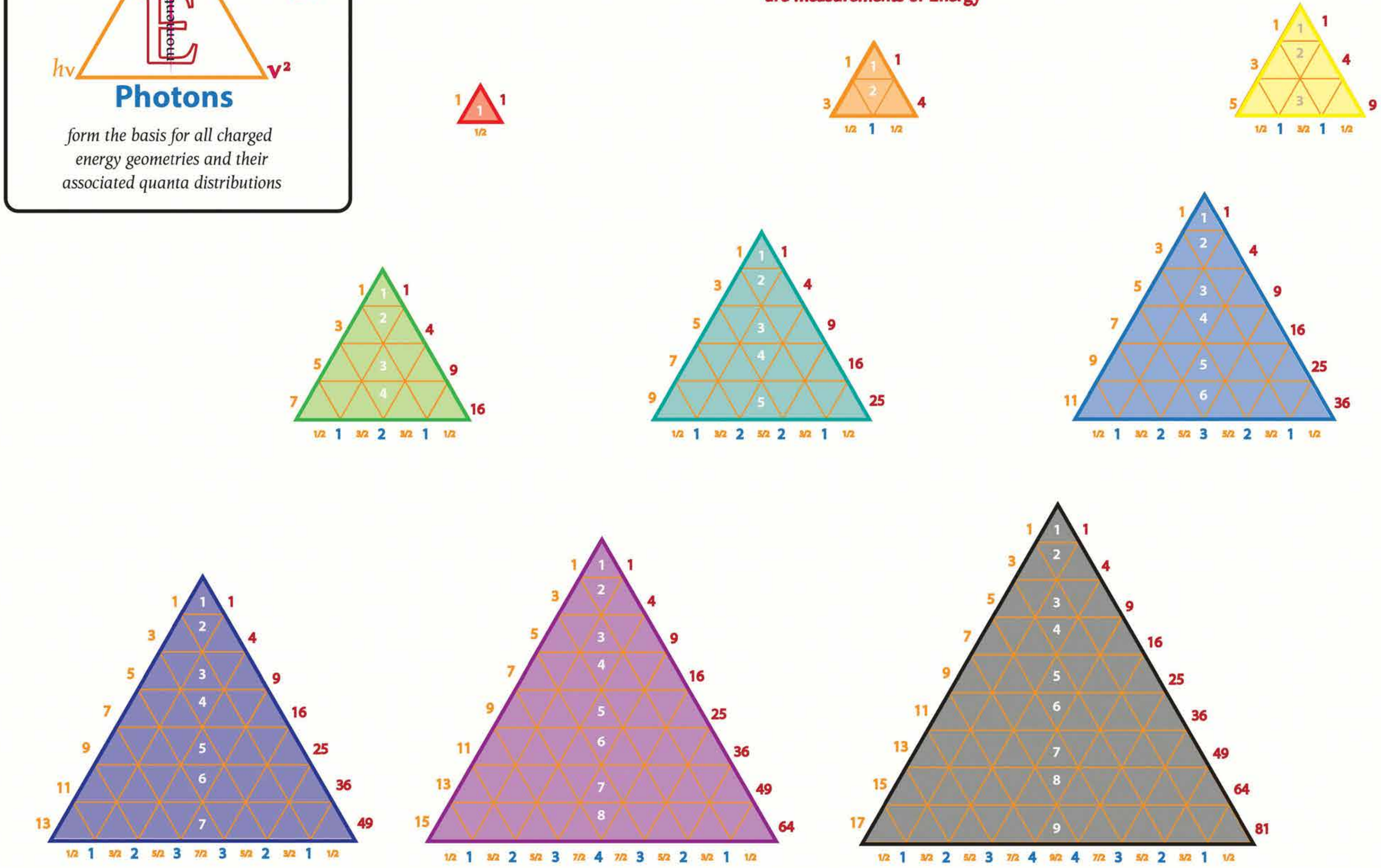
1 1 1
1/2
1 1 1 2 3 4
1/2 1 1/2
1 3 5 1 4 9 1 2 3
1/2 1 3/2 1 1/2
1 3 5 7 1 4 9 16 1 2 3 4
1/2 1 3/2 2 3/2 1 1/2
1 3 5 7 9 1 4 9 16 25 1 2 3 4 5
1/2 1 3/2 2 5/2 2 3/2 1 1/2
1 3 5 7 9 11 1 4 9 16 25 36 1 2 3 4 5 6
1/2 1 3/2 2 5/2 3 5/2 2 3/2 1 1/2
1 3 5 7 9 11 13 1 4 9 16 25 36 49 1 2 3 4 5 6 7
1/2 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 1/2
1 3 5 7 9 11 13 15 1 4 9 16 25 36 49 64 1 2 3 4 5 6 7 8
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2
1 3 5 7 9 11 13 15 17 1 4 9 16 25 36 49 64 81 1 2 3 4 5 6 7 8 9
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2

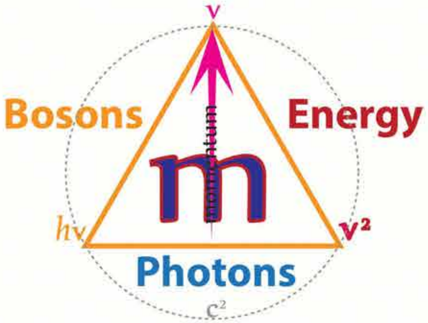
v
Bosons
Energy
m
hv
v²
Photons
c²

Time specific measurements of equilateral ENERGY momenta geometries forms the basis for EM mass

Electromagnetic mass

2D mass-energy is the surface integral of 3D Matter

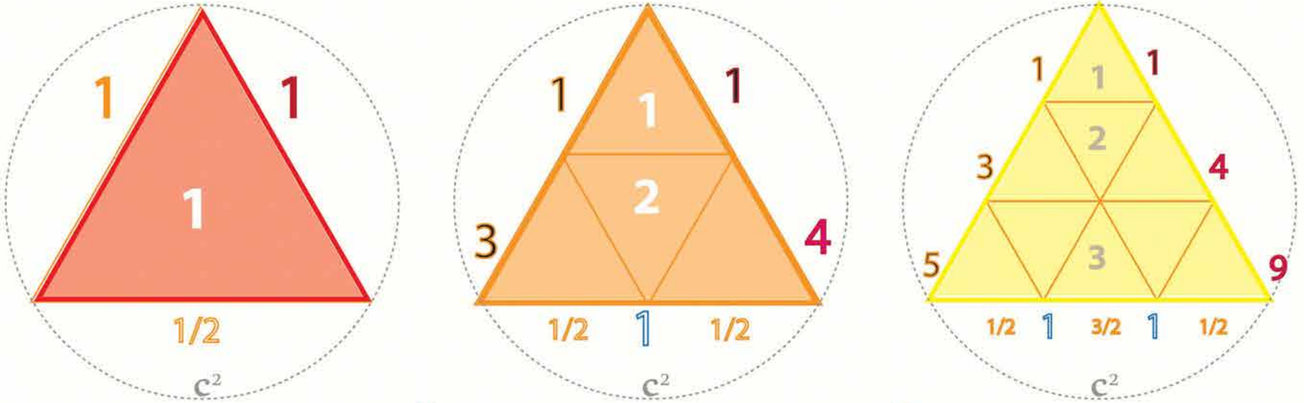

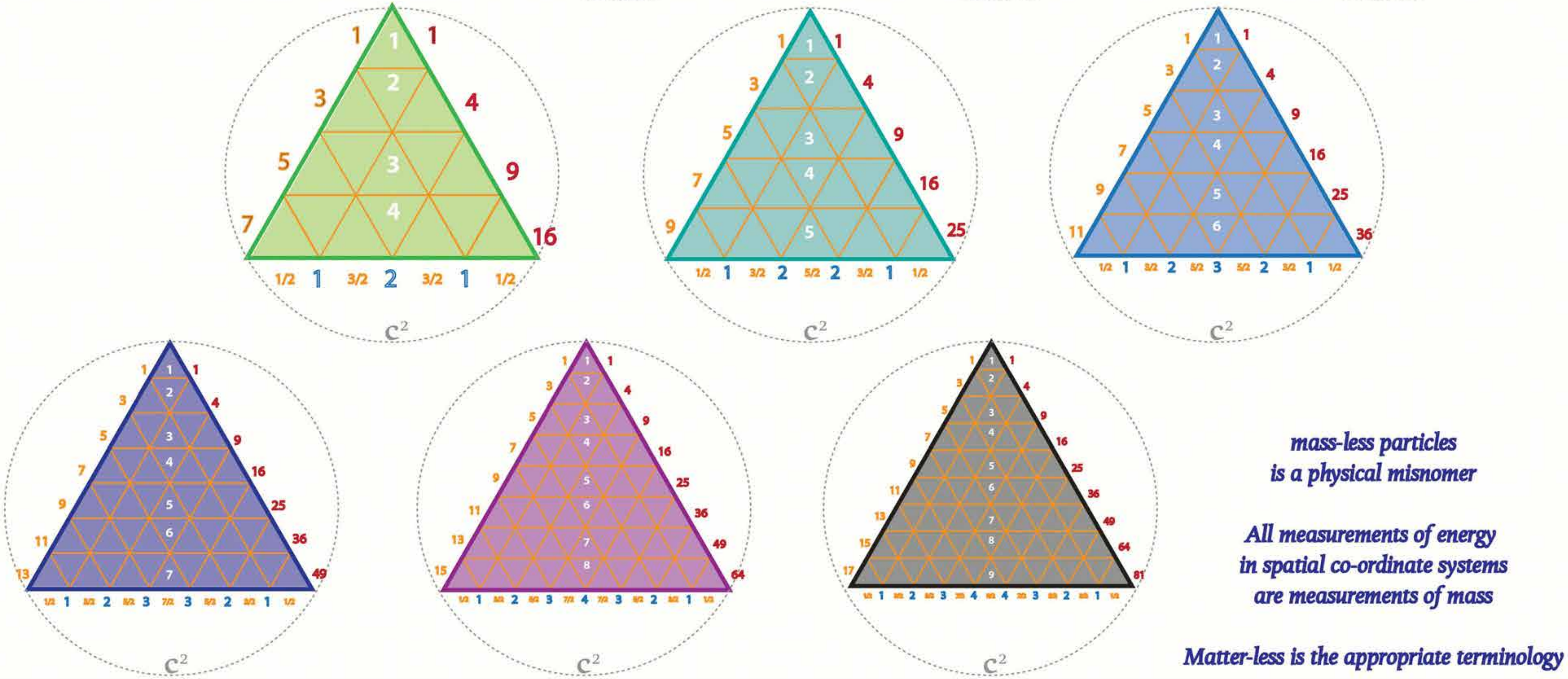
1
1
1
1/2
c²
1 1 3
1 4
1 2
1/2 1 1/2
c²
1 3 5
1 4 9
1 2 3
1/2 1 3/2 1 1/2
c²
1 3 5 7
1 4 9 16
1 2 3 4
1/2 1 3/2 2 3/2 1 1/2
c²
1 3 5 7 9
1 4 9 16 25
1 2 3 4 5
1/2 1 3/2 2 5/2 2 3/2 1 1/2
c²
1 3 5 7 9 11
1 4 9 16 25 36
1 2 3 4 5 6
1/2 1 3/2 2 5/2 3 5/2 2 3/2 1 1/2
c²
1 3 5 7 9 11 13
1 4 9 16 25 36 49
1 2 3 4 5 6 7
1/2 1 3/2 2 5/2 3 7/2 3 5/2 2 3/2 1 1/2
c²
1 3 5 7 9 11 13 15
1 4 9 16 25 36 49 64
1 2 3 4 5 6 7 8
1/2 1 3/2 2 5/2 3 7/2 4 7/2 3 5/2 2 3/2 1 1/2
c²
1 3 5 7 9 11 13 15 17
1 4 9 16 25 36 49 64 81
1 2 3 4 5 6 7 8 9
1/2 1 3/2 2 5/2 3 7/2 4 9/2 4 7/2 3 5/2 2 3/2 1 1/2
c²

mass-less particles is a physical misnomer

All measurements of energy in spatial co-ordinate systems are measurements of mass

Matter-less is the appropriate terminology

Standing-wave Matter

Matter is a higher order 3D topology created by standing wave mass-energies

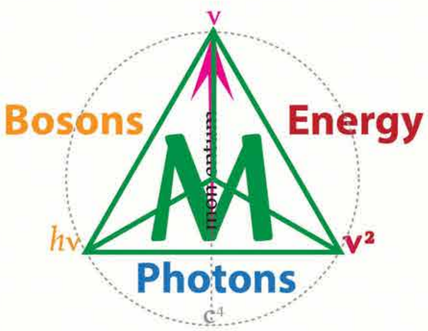
v
Bosons
Energy
M
momentum
hv
v²
Photons
c⁴

The equilateral mass-ENERGY content of 4nπ Tetrahedral standing-wave topologies forms the basis of Matter

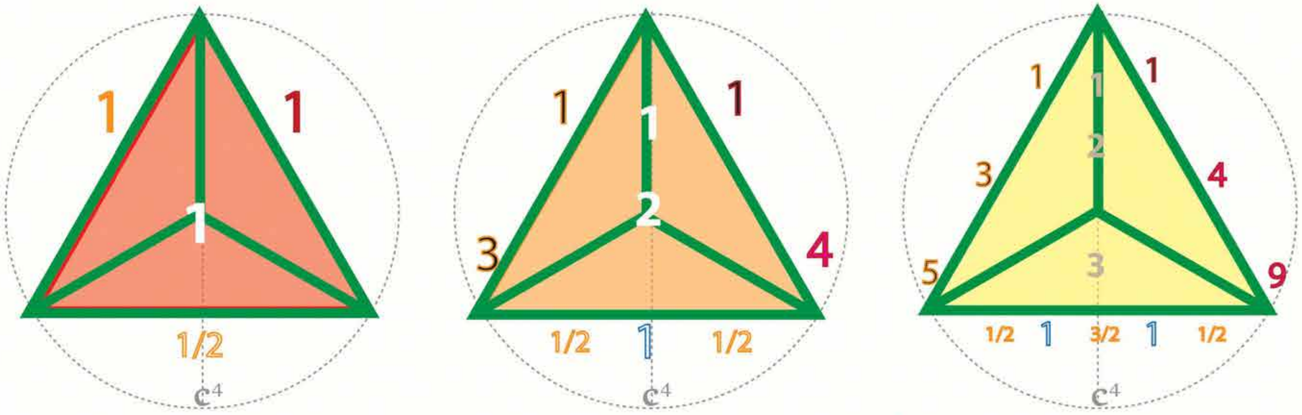
1
1
1
1/2
c⁴
1
1
1
2
3
4
1/2
1
1/2
c⁴
1
1
1
2
3
4
3
5
9
1/2
1
3/2
1
1/2
c⁴

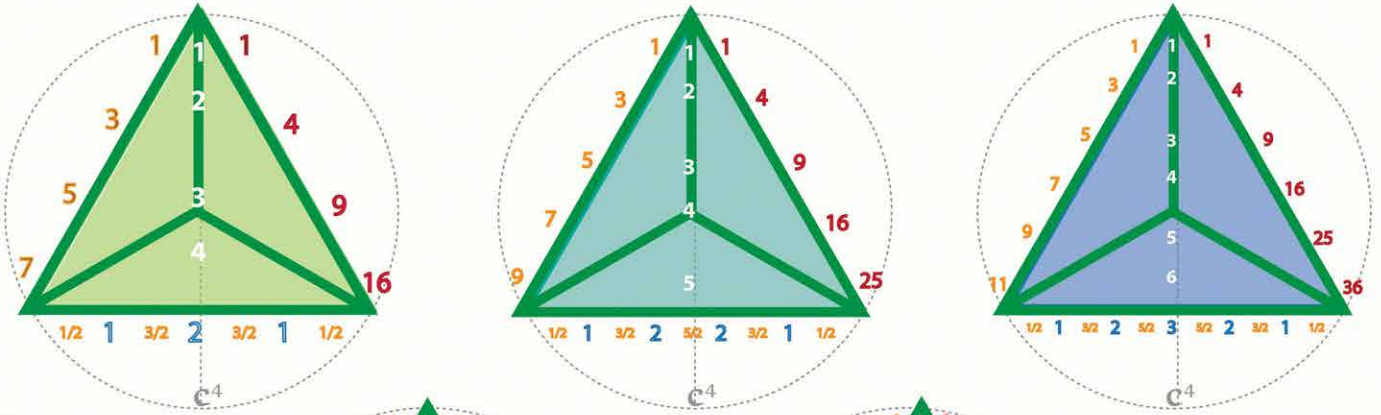
1
1
3
4
5
9
7
16
1
2
3
4
1/2
1
3/2
2
3/2
1
1/2
c⁴
1
1
3
4
5
9
7
16
9
25
1
2
3
4
5
1/2
1
3/2
2
5/2
2
3/2
1
1/2
c⁴
1
1
3
4
5
9
7
16
9
25
11
36
1
2
3
4
5
6
1/2
1
3/2
2
5/2
3
5/2
2
3/2
1
1/2
c⁴

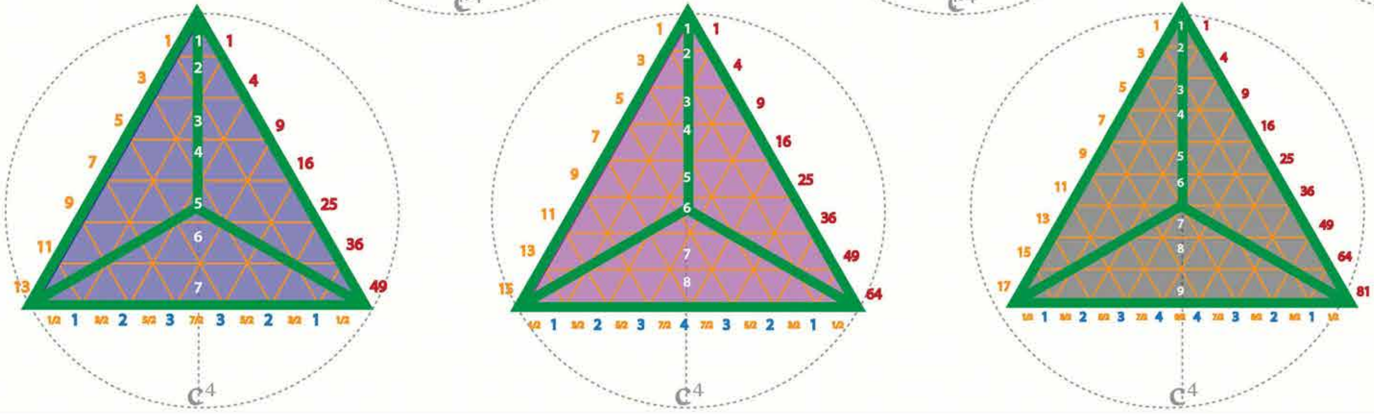
1
1
3
4
5
9
7
16
9
25
11
36
13
49
1
2
3
4
5
6
7
1/2
1
3/2
2
5/2
3
7/2
3
5/2
2
3/2
1
1/2
c⁴
1
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
1
2
3
4
5
6
7
8
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2
c⁴
1
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
17
81
1
2
3
4
5
6
7
8
9
1/2
1
3/2
2
5/2
3
7/2
4
9/2
4
7/2
3
5/2
2
3/2
1
1/2
c⁴

Only mass-ENERGIES contained in the tetryonic fascia of standing-wave topologies contribute to weight

Periodic element nuclei

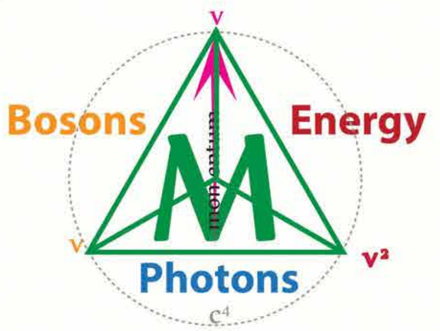
v
Bosons
Energy
momentum
M
v
v²
Photons
c4

The mass-ENERGY content of
4nπ Tetrahedral standing-wave topologies
forms the basis of all periodic elements

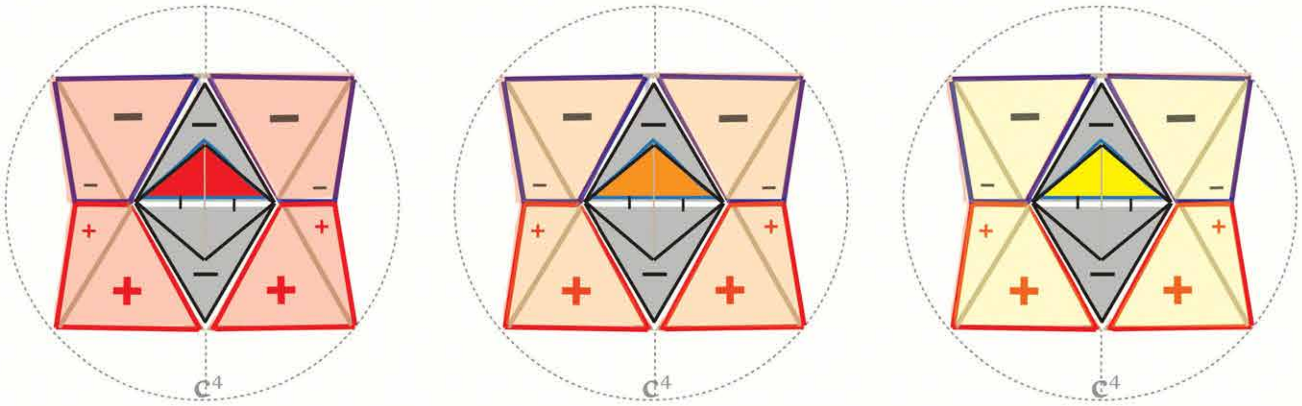
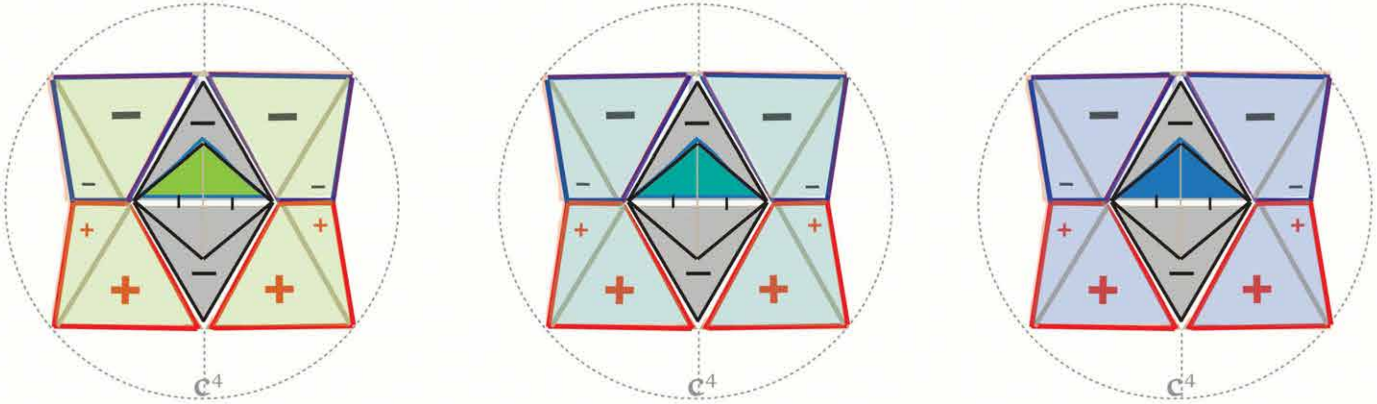
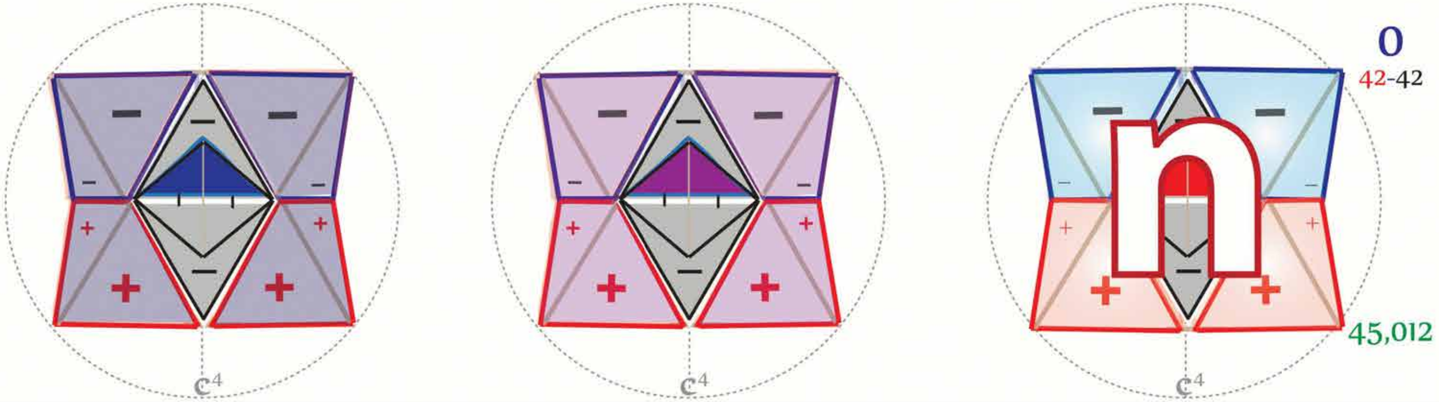
−
+
c4
n
0
42-42
45,012

ALL
periodic elements are
comprised of
n level deuterium
nuclei

Copyright ABRAHAM [2008] - All rights reserved

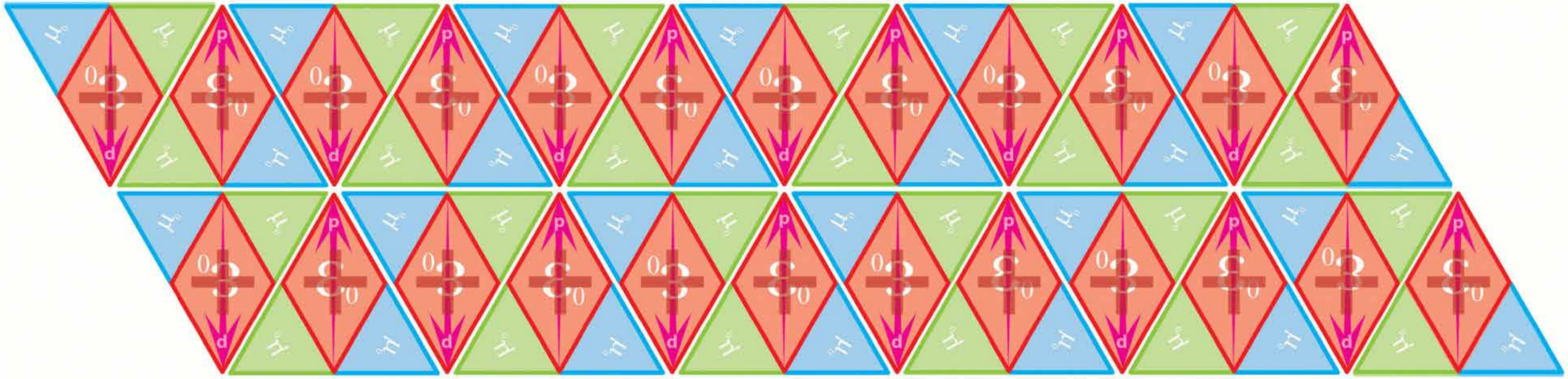

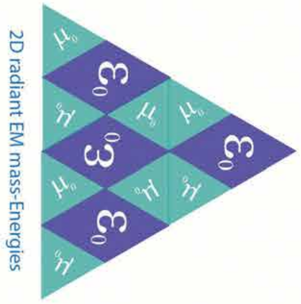
2D radiant EM mass-Energies
ε0
μ0

n1 - Charged Tetryon Templates

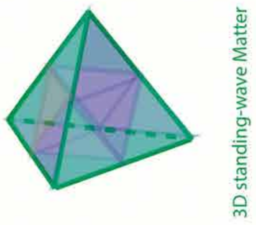
3D standing-wave Matter

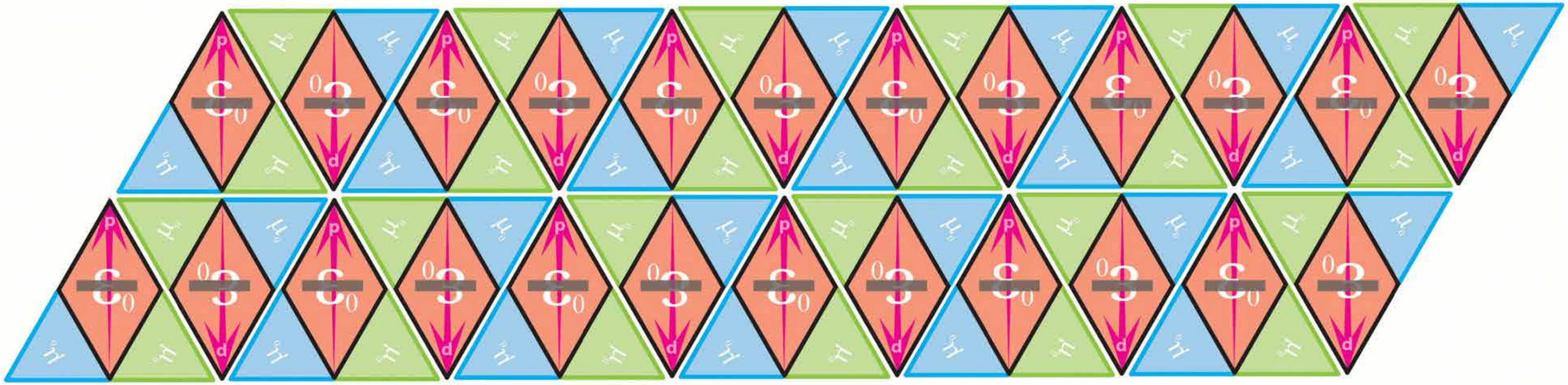

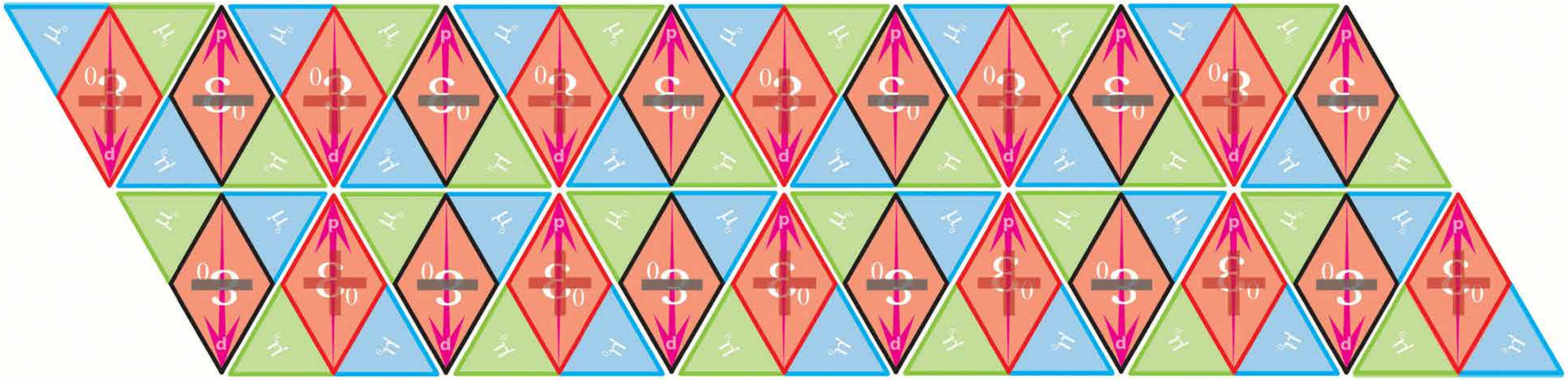

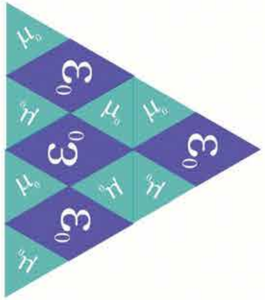

2D radiant EM mass-Energies

n1 - Neutral Tetryon Templates

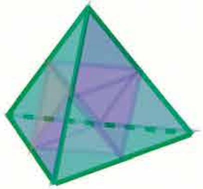

3D standing-wave Matter

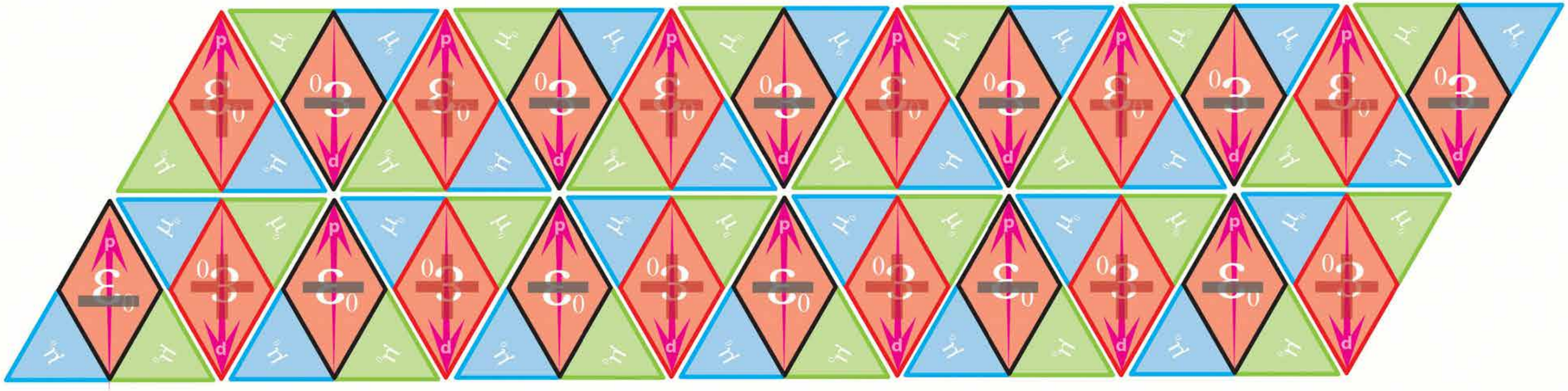

Copyright ABRAHAM [2008] - All rights reserved

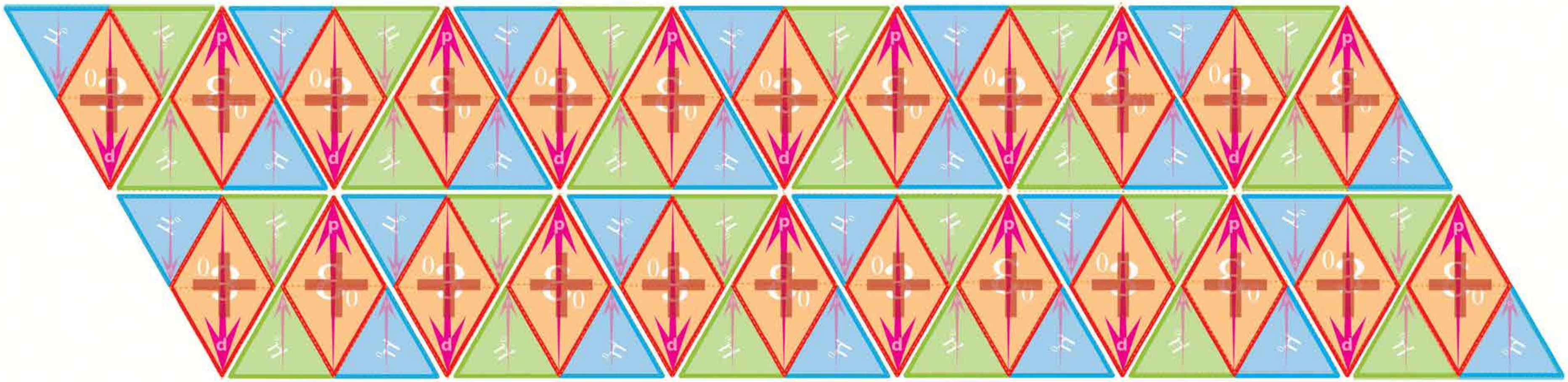

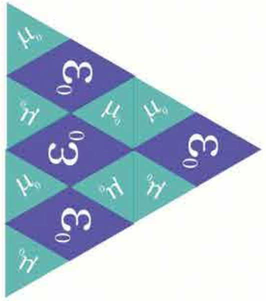

2D radiant EM mass-Energies

n2 - Charged Tetryon Templates

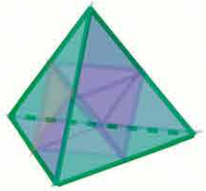

3D standing-wave Matter

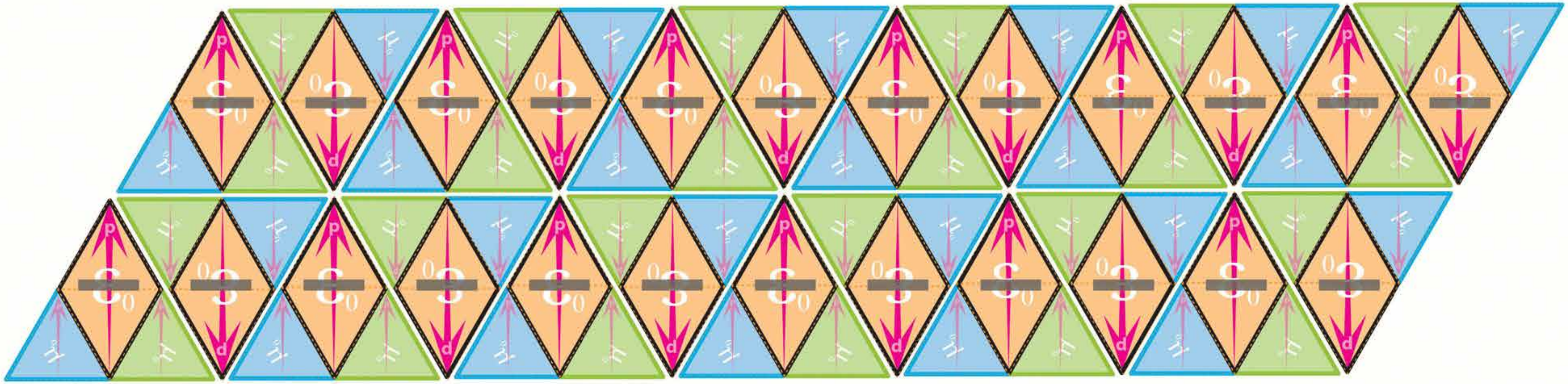

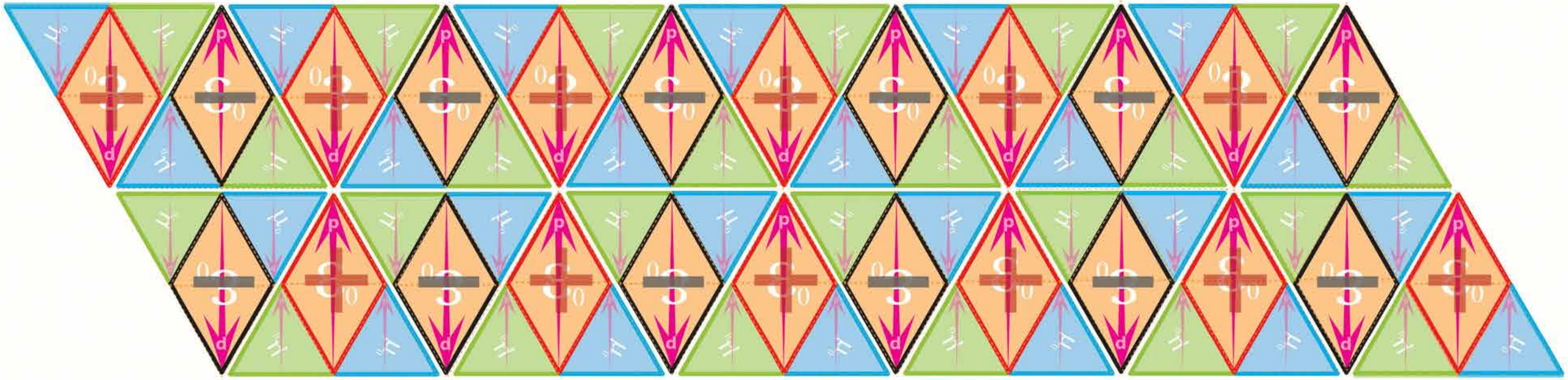

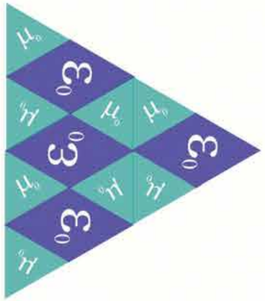

2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

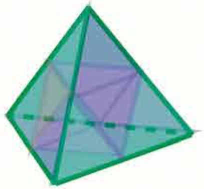

3D standing-wave Matter

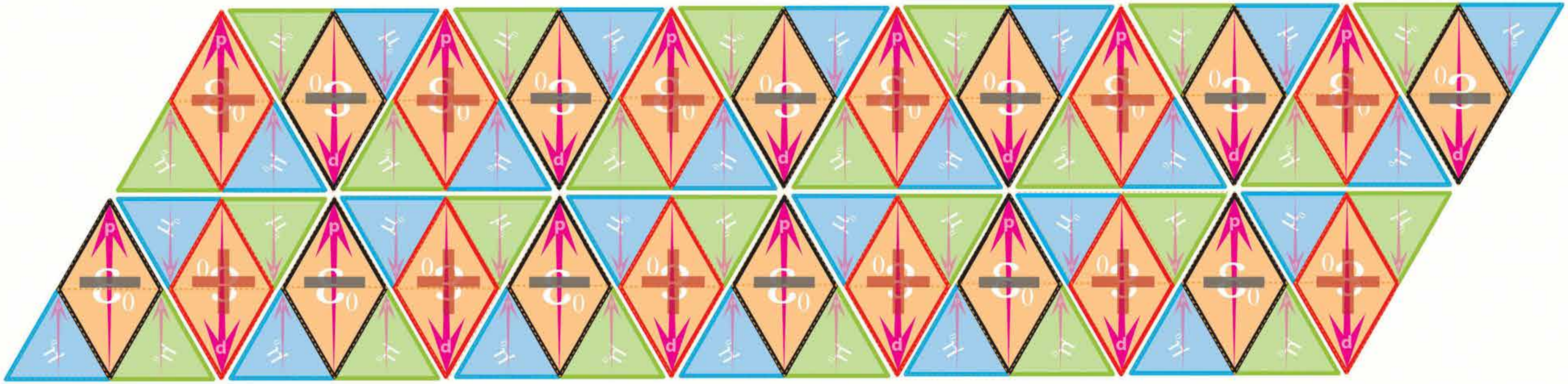

Copyright ABRAHAM [2008] - All rights reserved

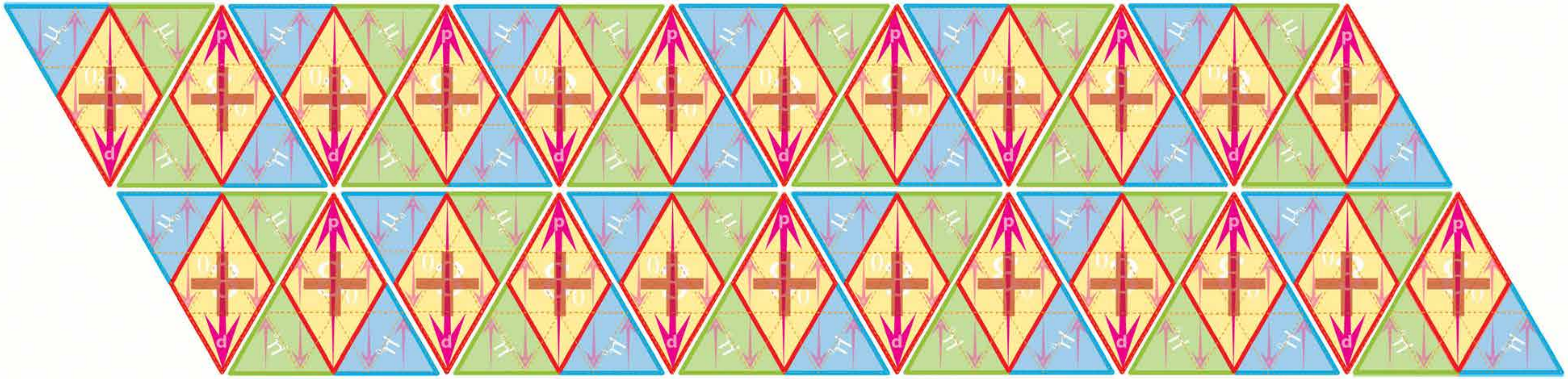
p
d
μ_0
ε_0

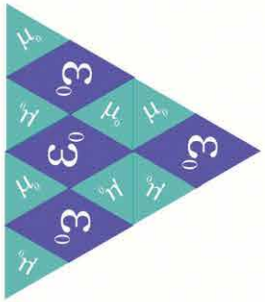
μ_0
ε_0

2D radiant EM mass-Energies

n3 - Charged Tetryon Templates

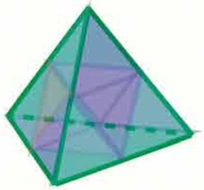

3D standing-wave Matter

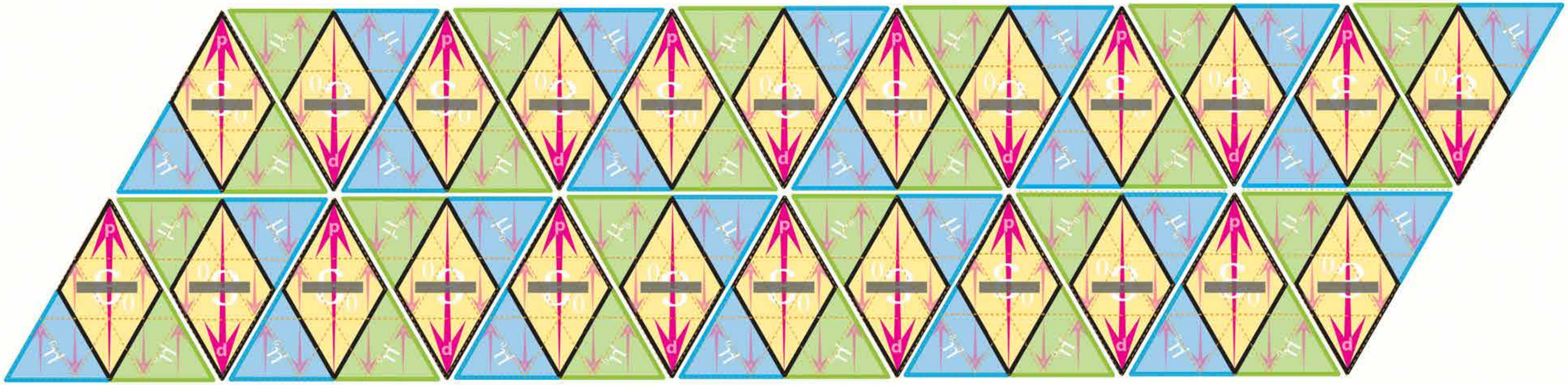
p
d
μ_0
ε_0

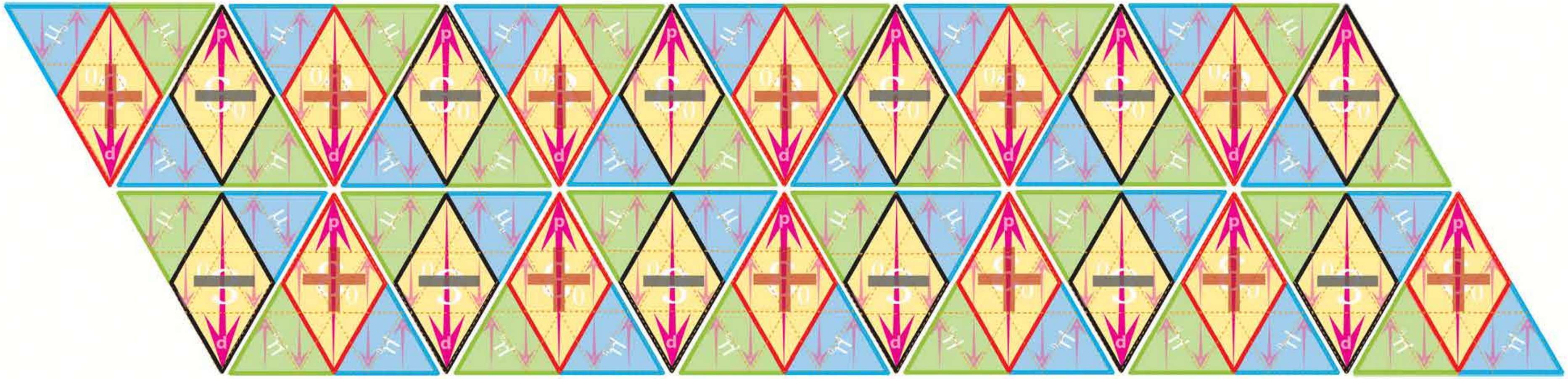

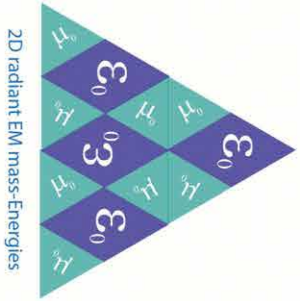
2D radiant EM mass-Energies

n3 - Neutral Tetryon Templates

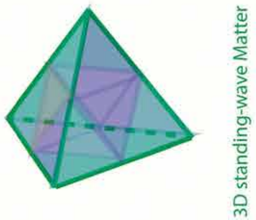
3D standing-wave Matter

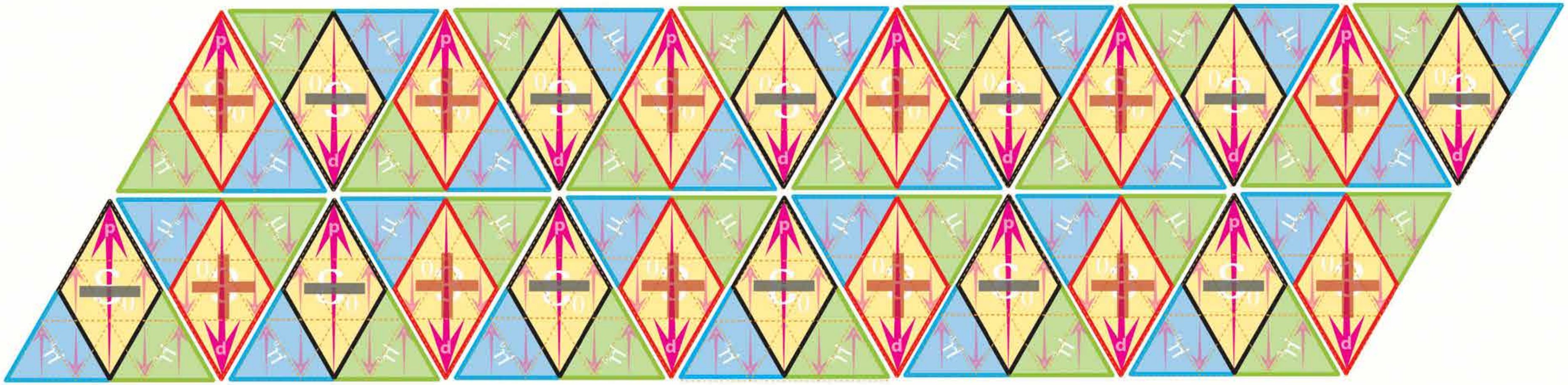

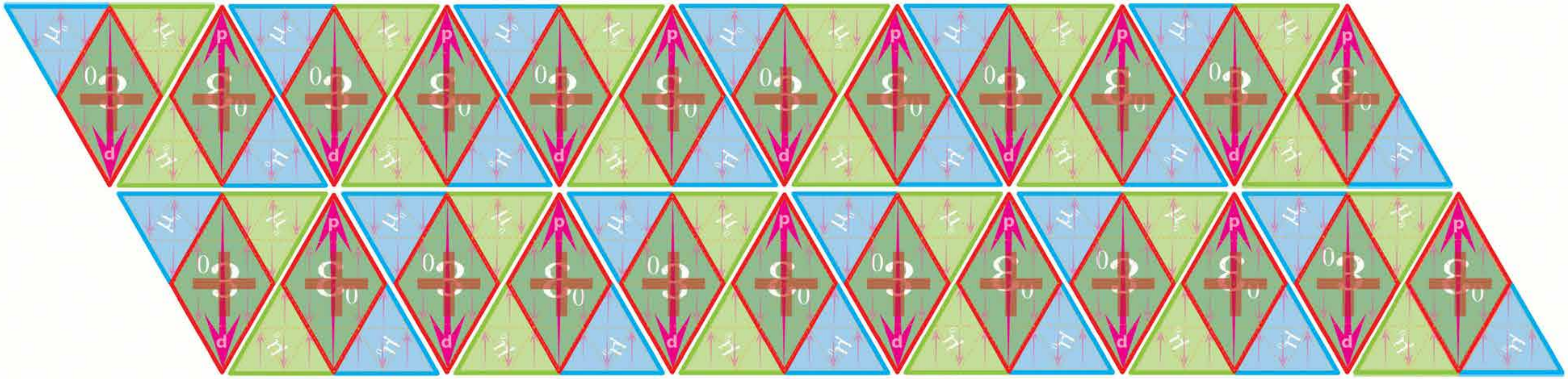
p
d
ε_0
μ_0

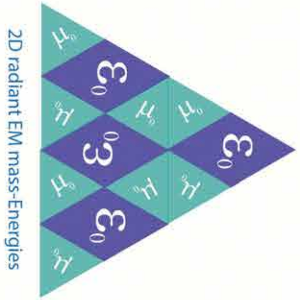
ε_0
μ_0

2D radiant EM mass-Energies

n4 - Charged Tetryon Templates

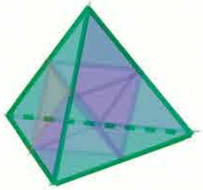

3D standing-wave Matter

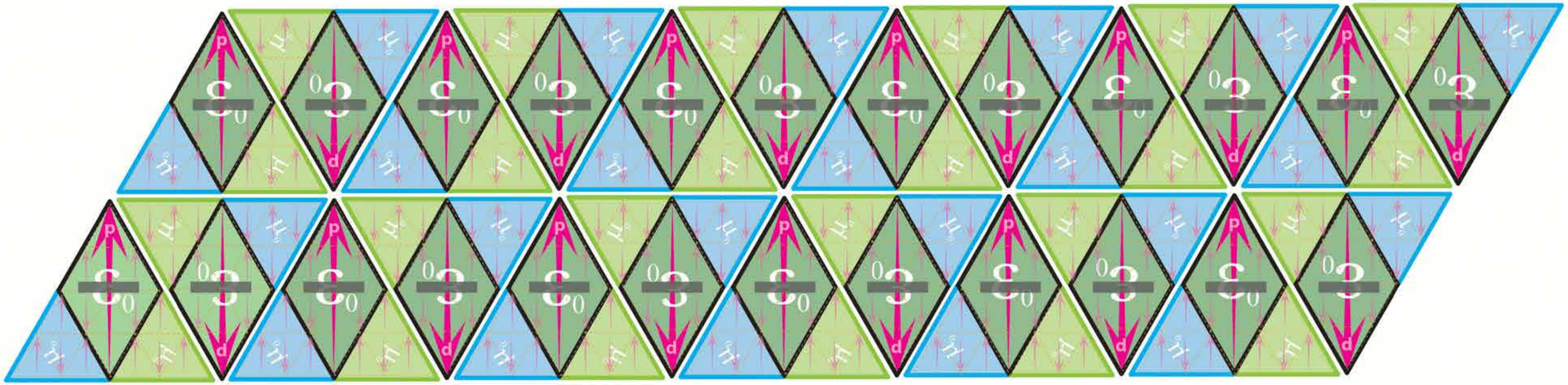
p
d
ε_0
μ_0

Copyright ABRAHAM [2008] - All rights reserved

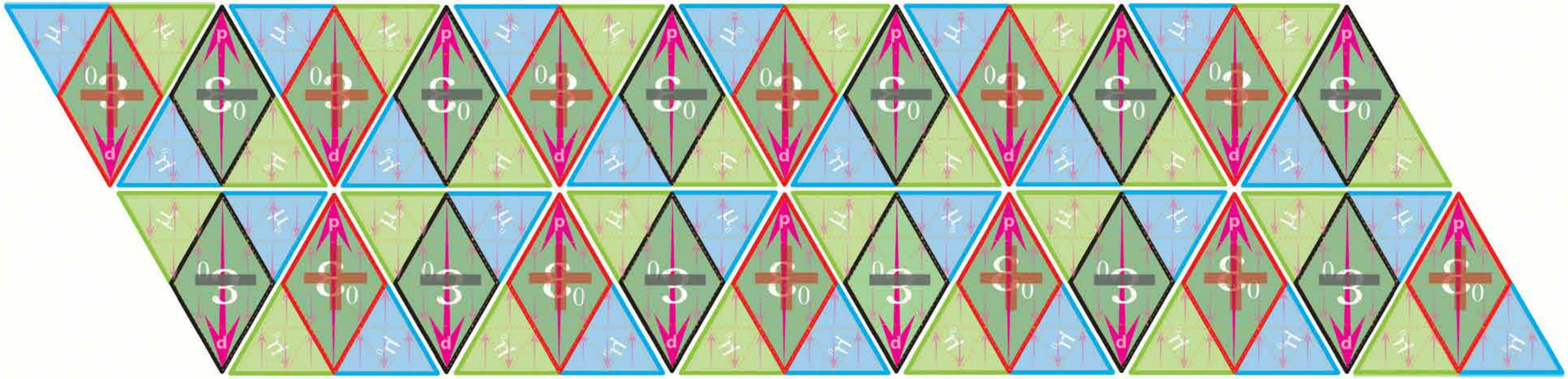
p
d
ε_0
μ_0

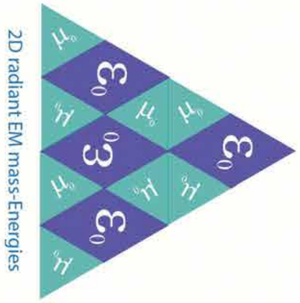
2D radiant EM mass-Energies
ε_0
μ_0

n4 - Neutral Tetryon Templates

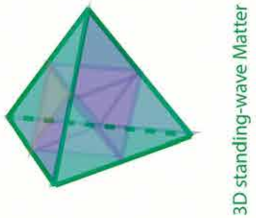
3D standing-wave Matter

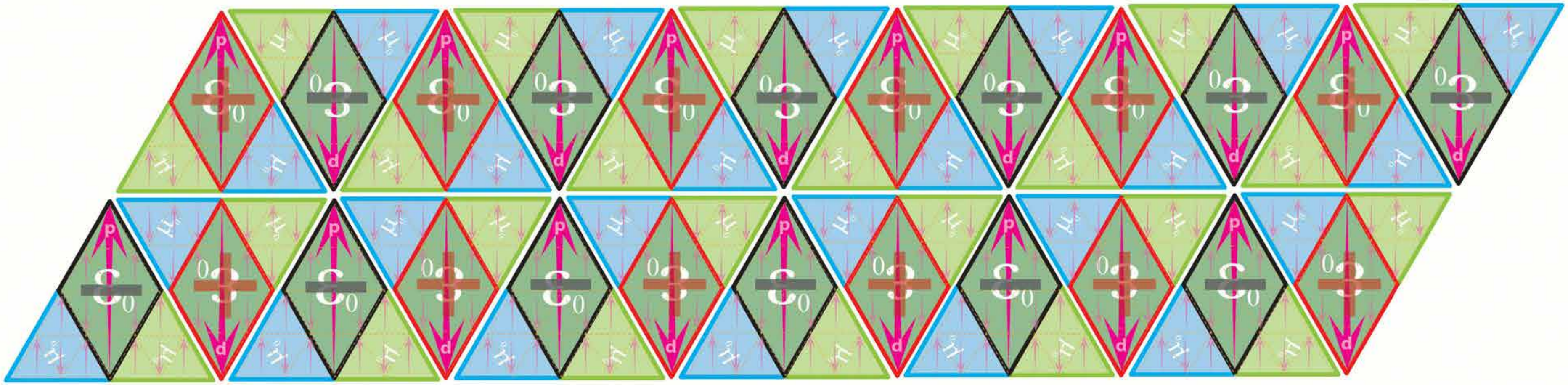
p
d
ε_0
μ_0

Copyright ABRAHAM [2008] - All rights reserved

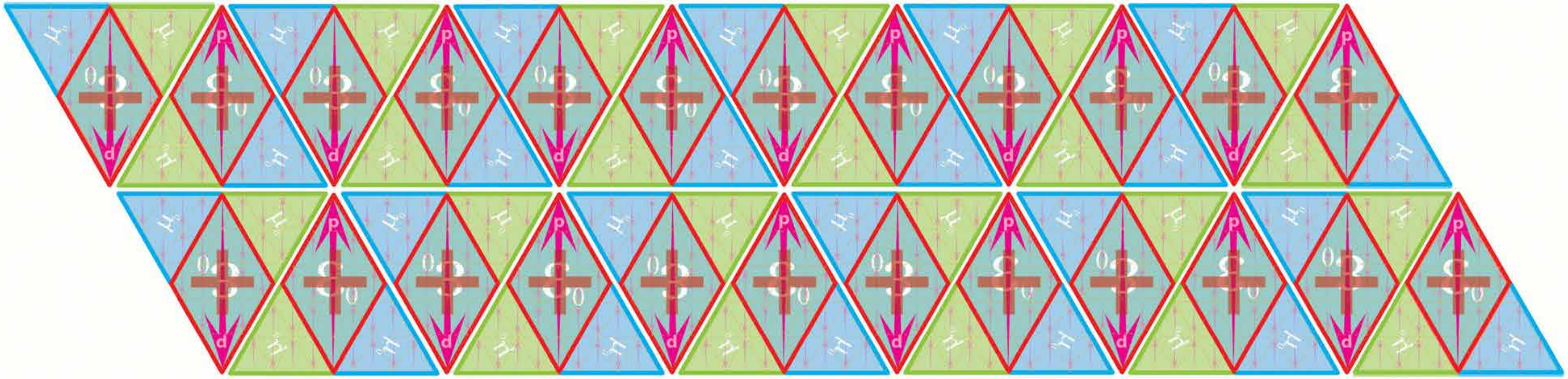

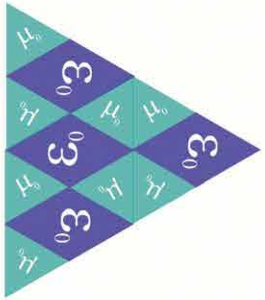
2D radiant EM mass-Energies

n5 - Charged Tetryon Templates

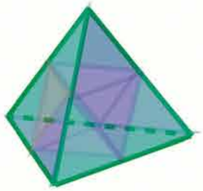
3D standing-wave Matter

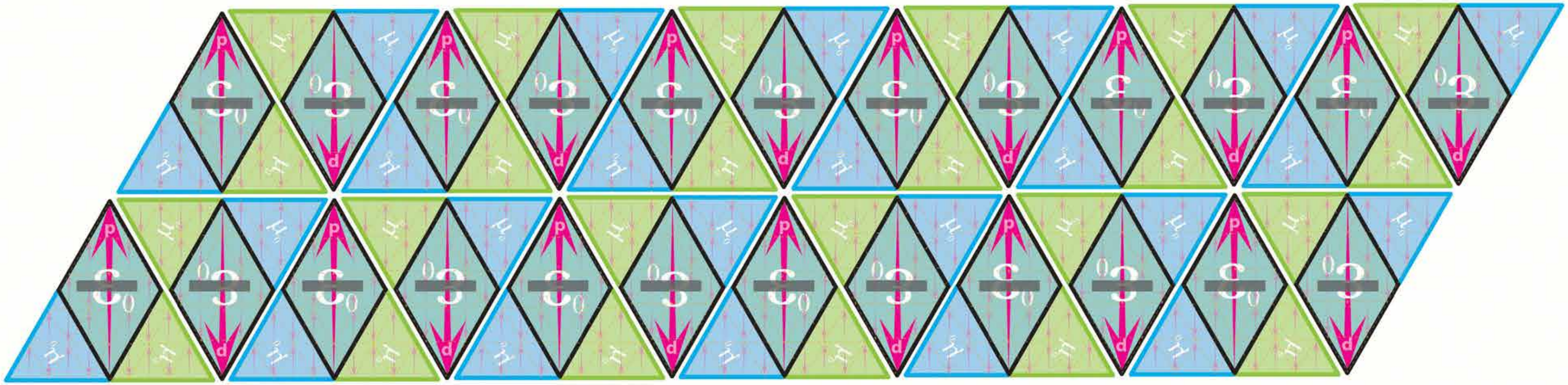

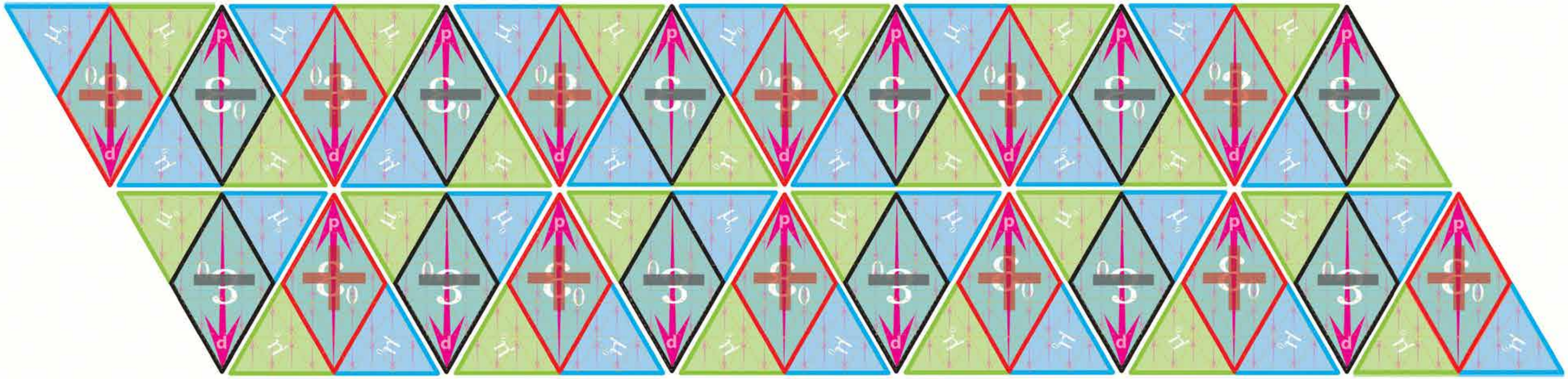

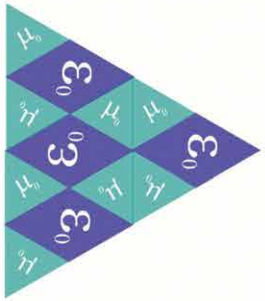

2D radiant EM mass-Energies

n5 - Neutral Tetryon Templates

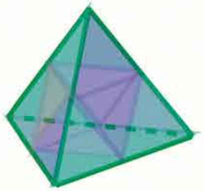

3D standing-wave Matter

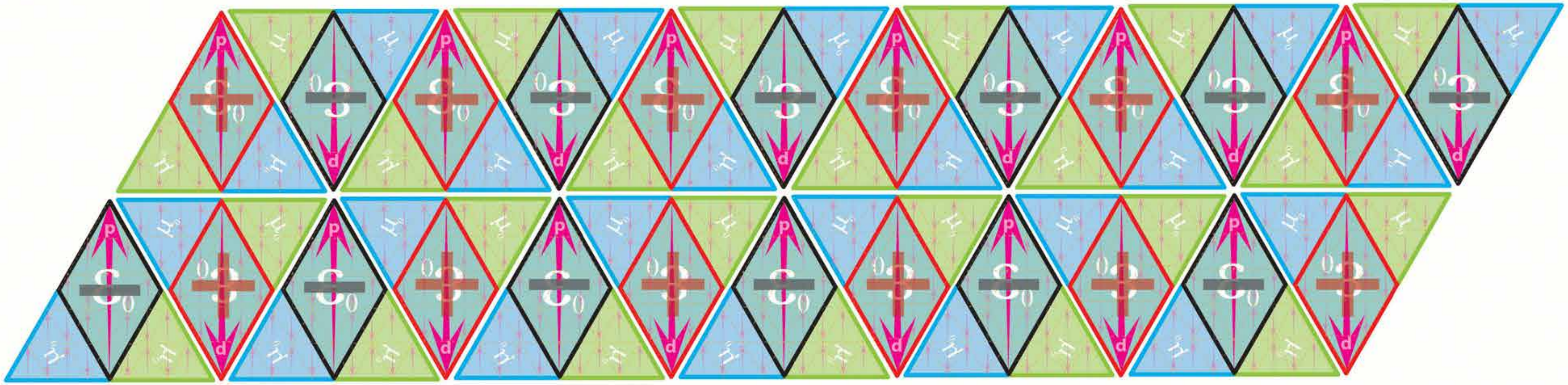

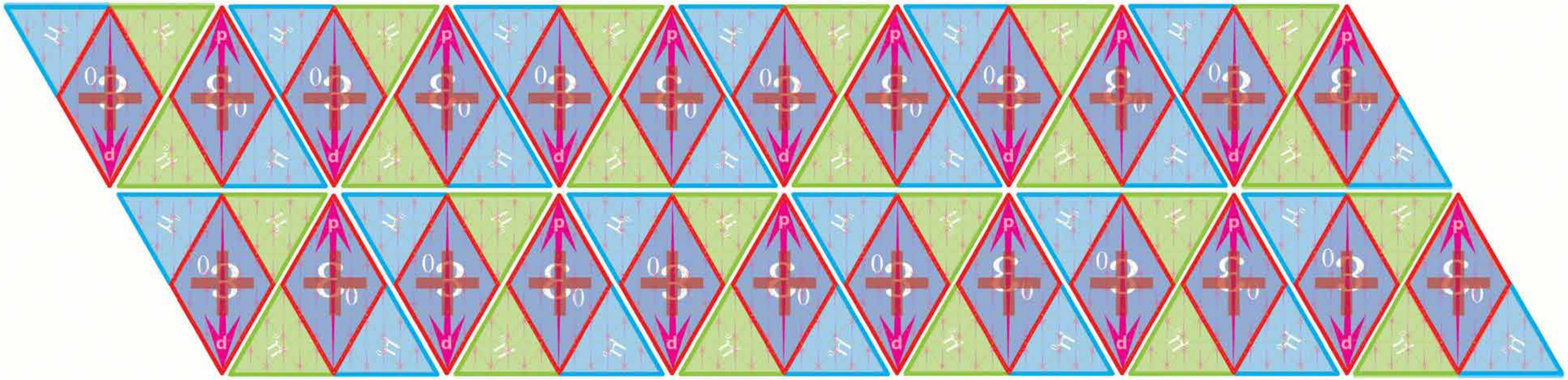

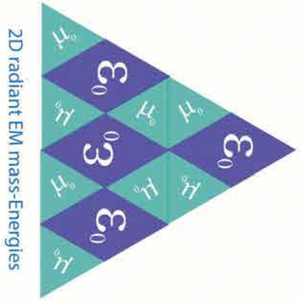
2D radiant EM mass-Energies
ε0
μ0

n6 - Charged Tetryon Templates

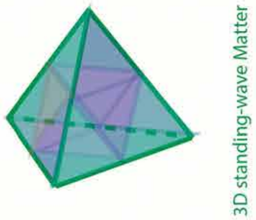
3D standing-wave Matter

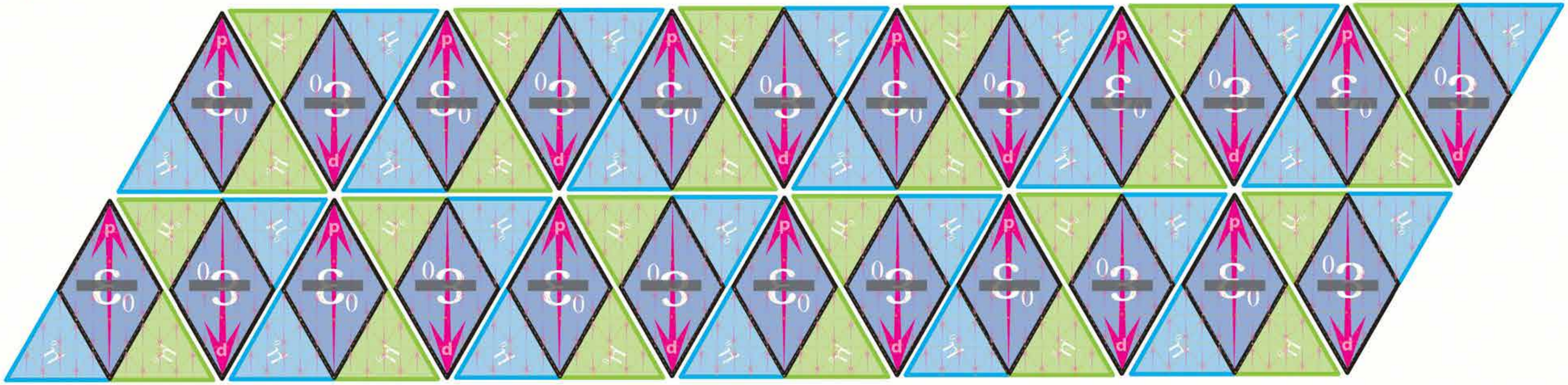

Copyright ABRAHAM [2008] - All rights reserved

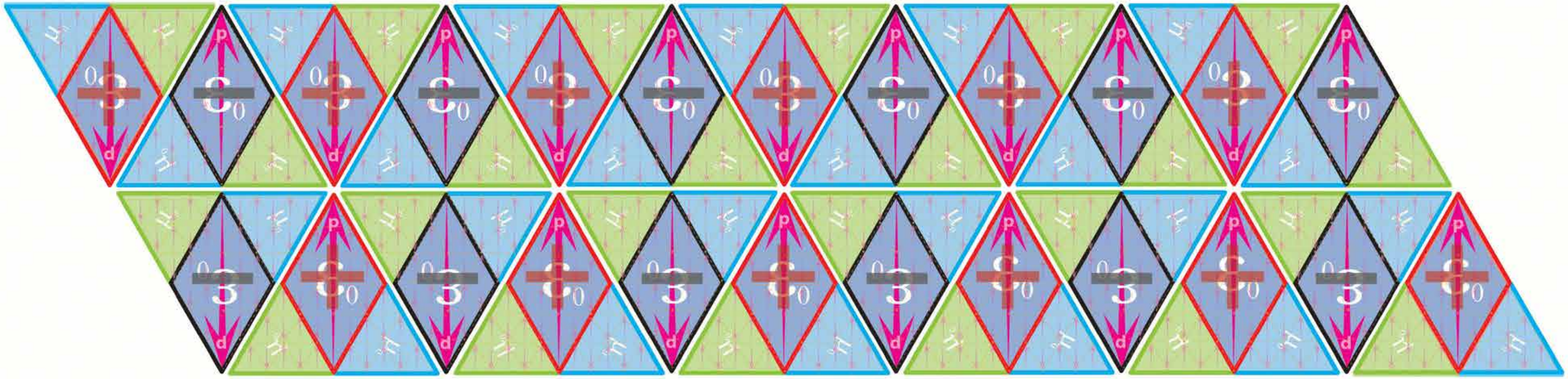

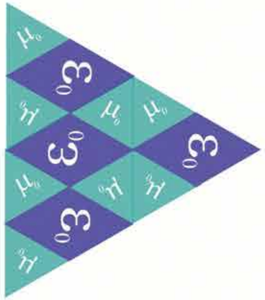

2D radiant EM mass-Energies

n6 - Neutral Tetryon Templates

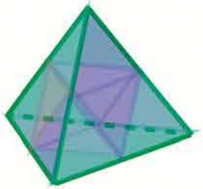

3D standing-wave Matter

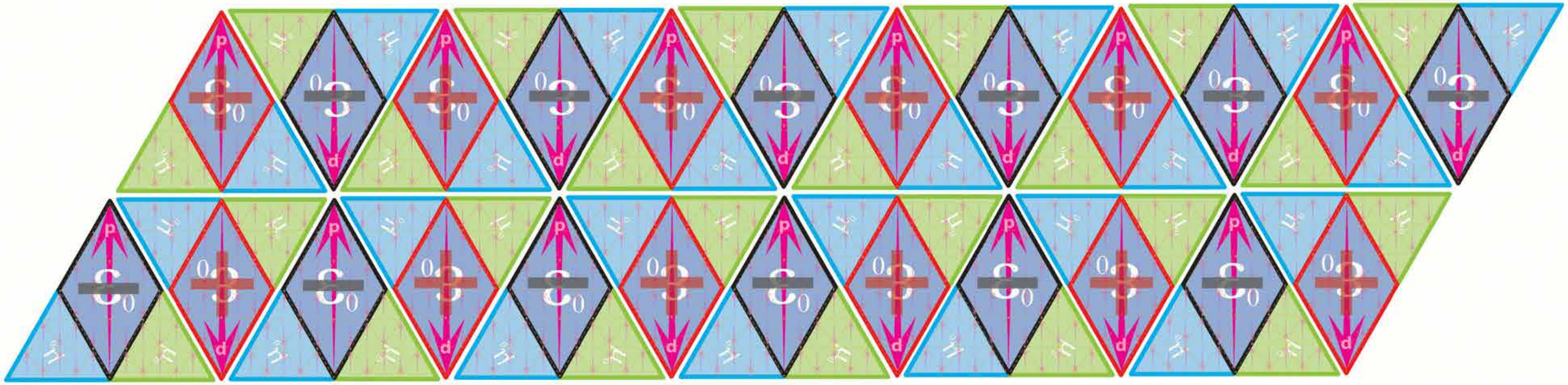

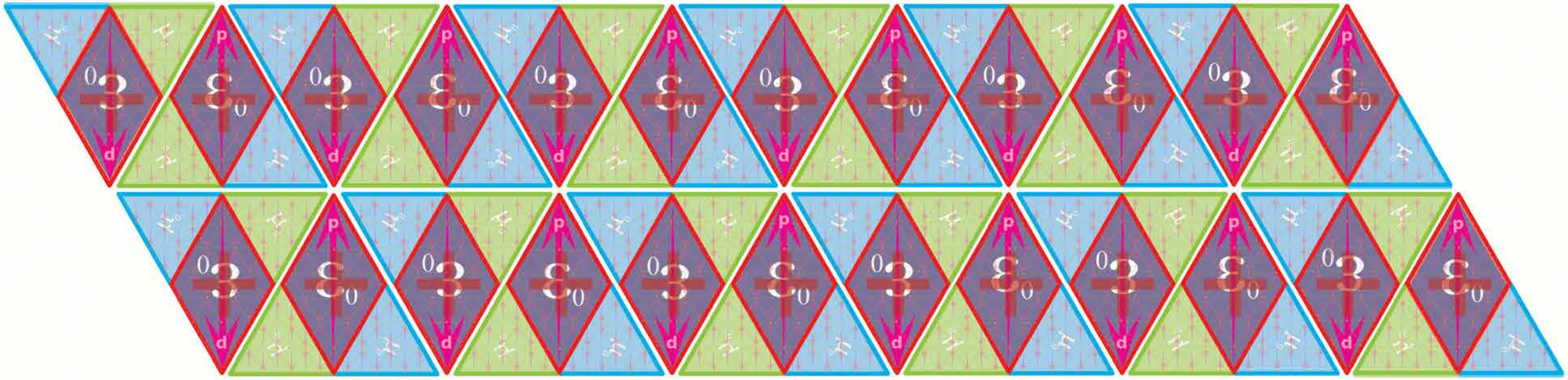

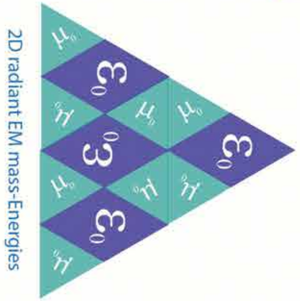
2D radiant EM mass-Energies
ε0
μ0

n7 - Charged Tetryon Templates

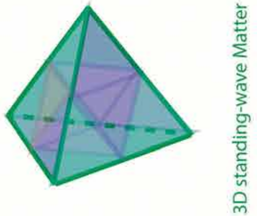
3D standing-wave Matter

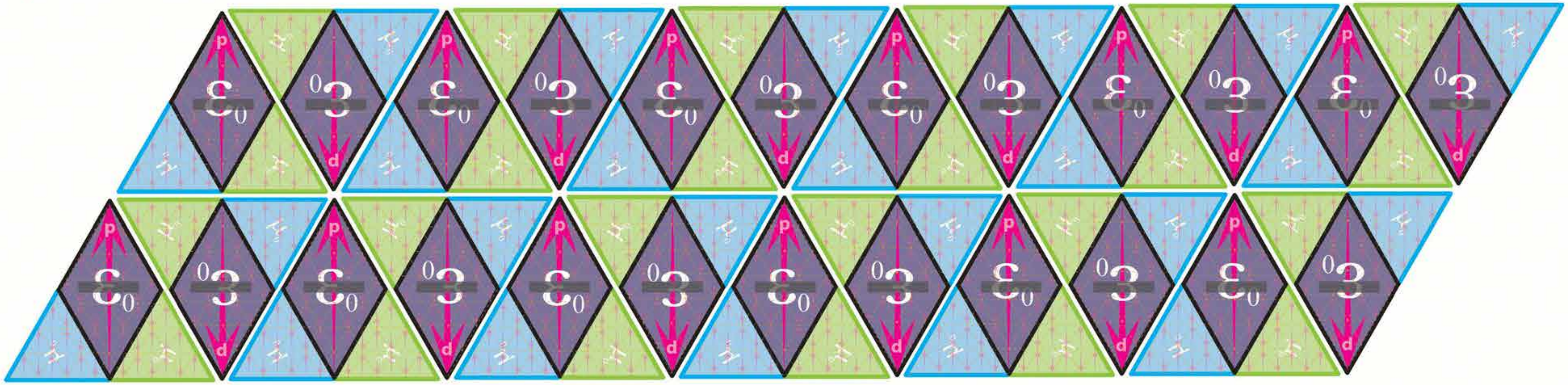

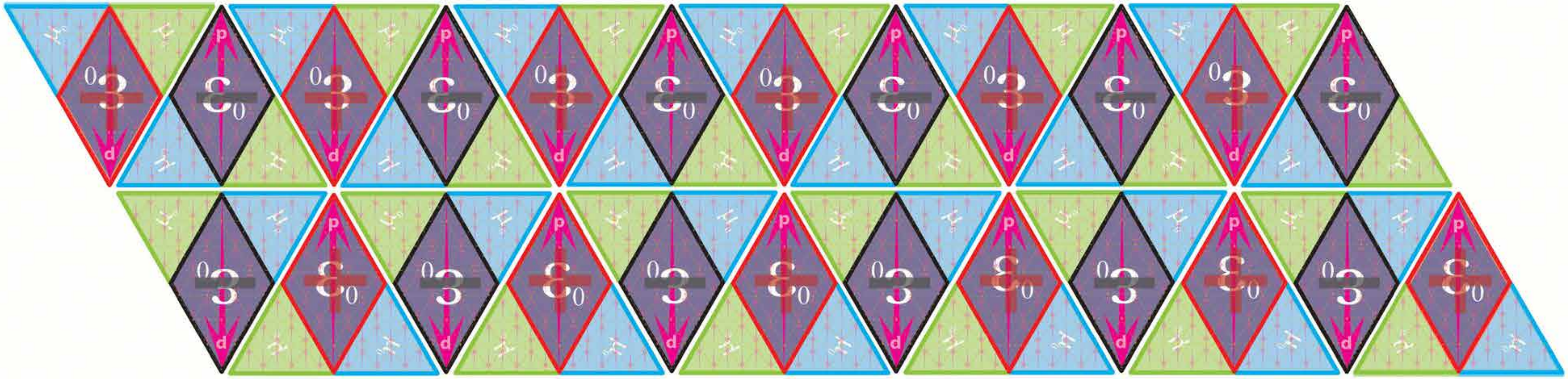

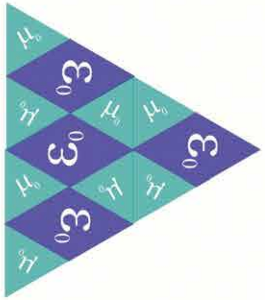

2D radiant EM mass-Energies

n7 - Neutral Tetryon Templates

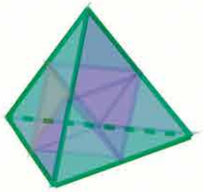

3D standing-wave Matter

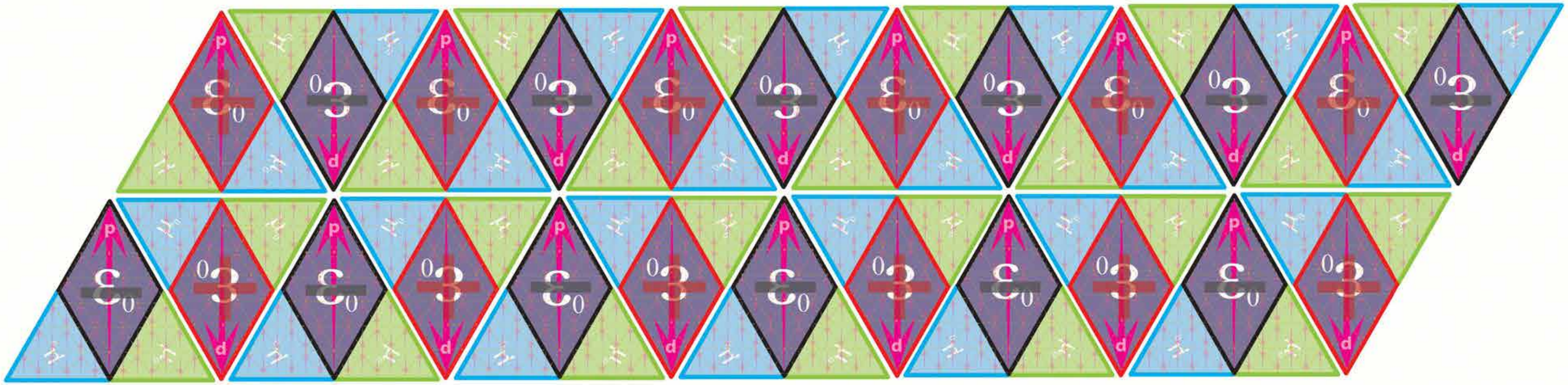

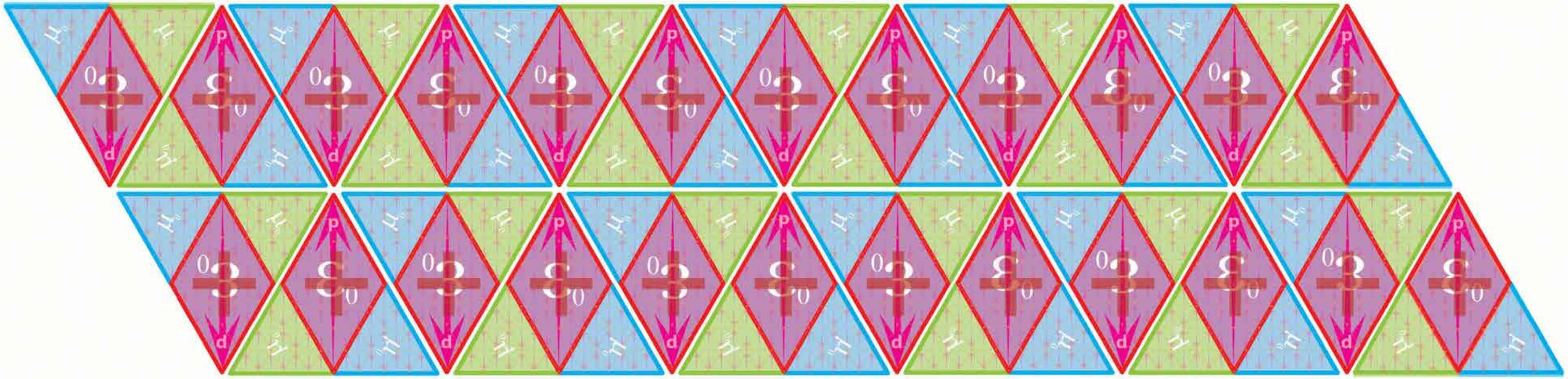
p
d
μ_0
ε_0

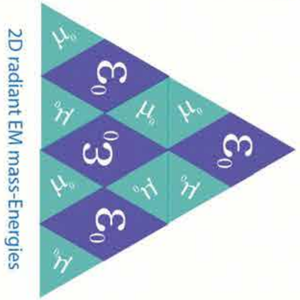
2D radiant EM mass-Energies
μ_0
ε_0

n8 - Charged Tetryon Templates

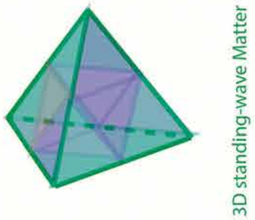
3D standing-wave Matter

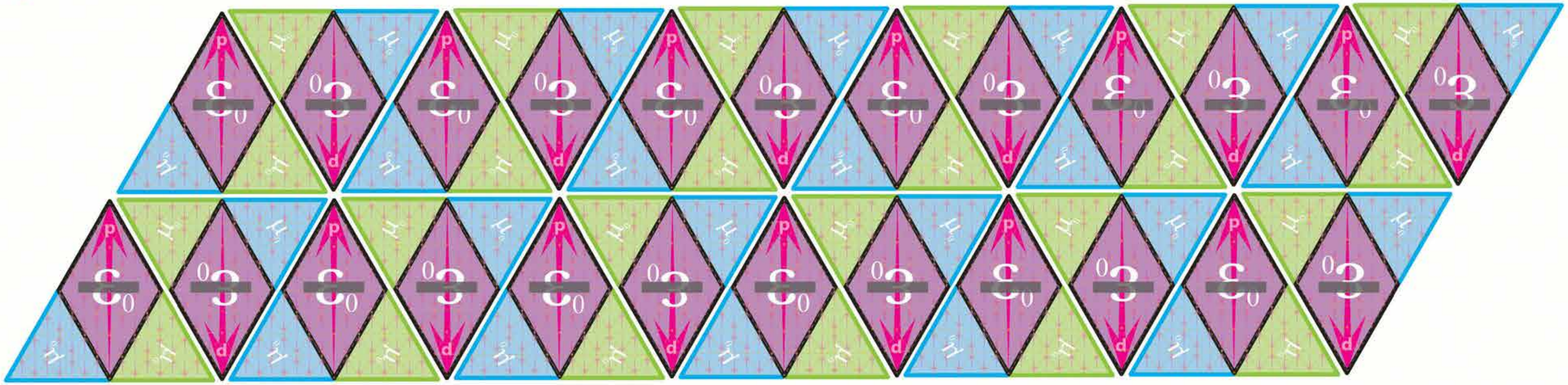
p
d
μ_0
ε_0

Copyright ABRAHAM [2008] - All rights reserved

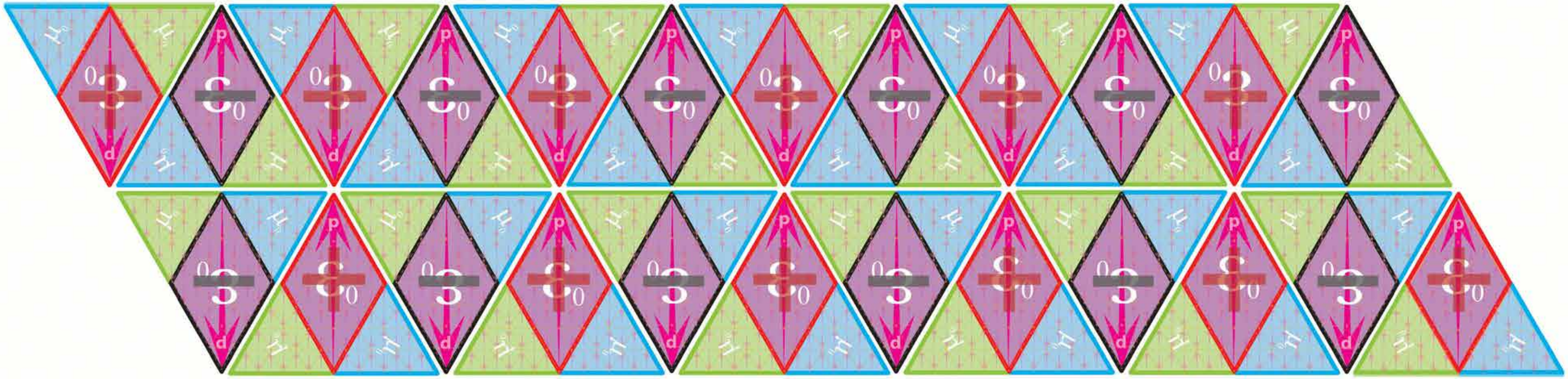

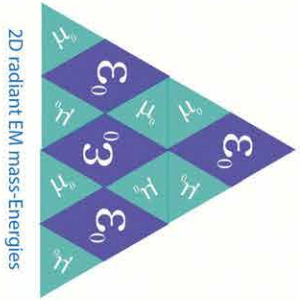
2D radiant EM mass-Energies
μ_0
ε_0

n8 - Neutral Tetryon Templates

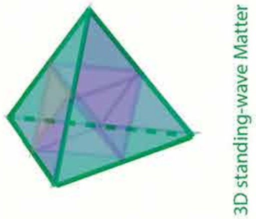
3D standing-wave Matter

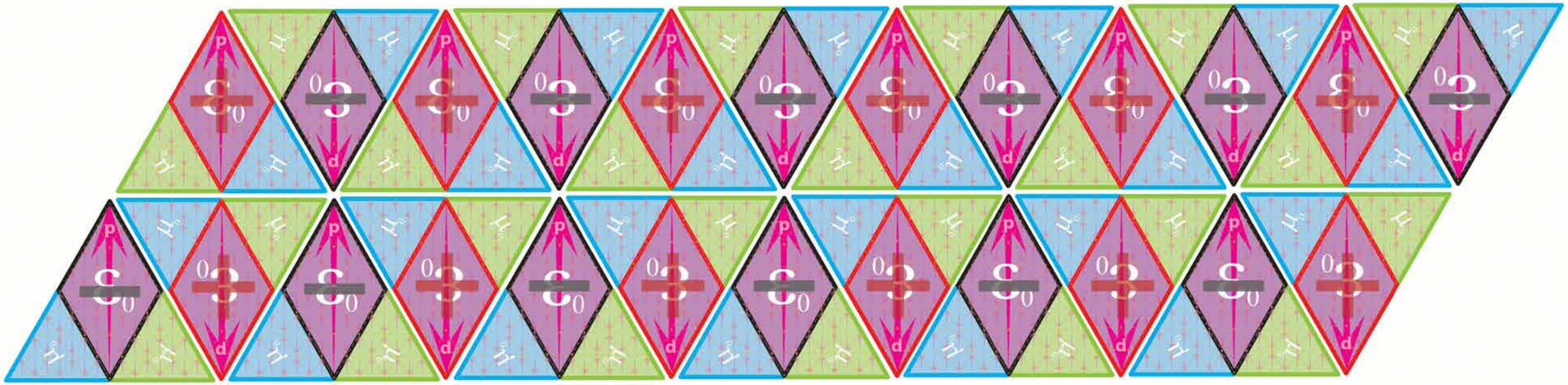

Periodic element geometries

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus.
The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the "2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus.
The shell letters K, L, M, ... are alphabetical

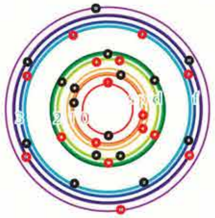

Each shell consists of one or more subshells, and each subshell consists of one or more atomic orbitals.

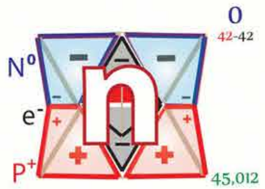
0
42-42
N⁰
e⁻
P⁺
n
45,012

Each shell can contain only an integer number of whole deuterium nuclei [Proton, Neutron & electron]

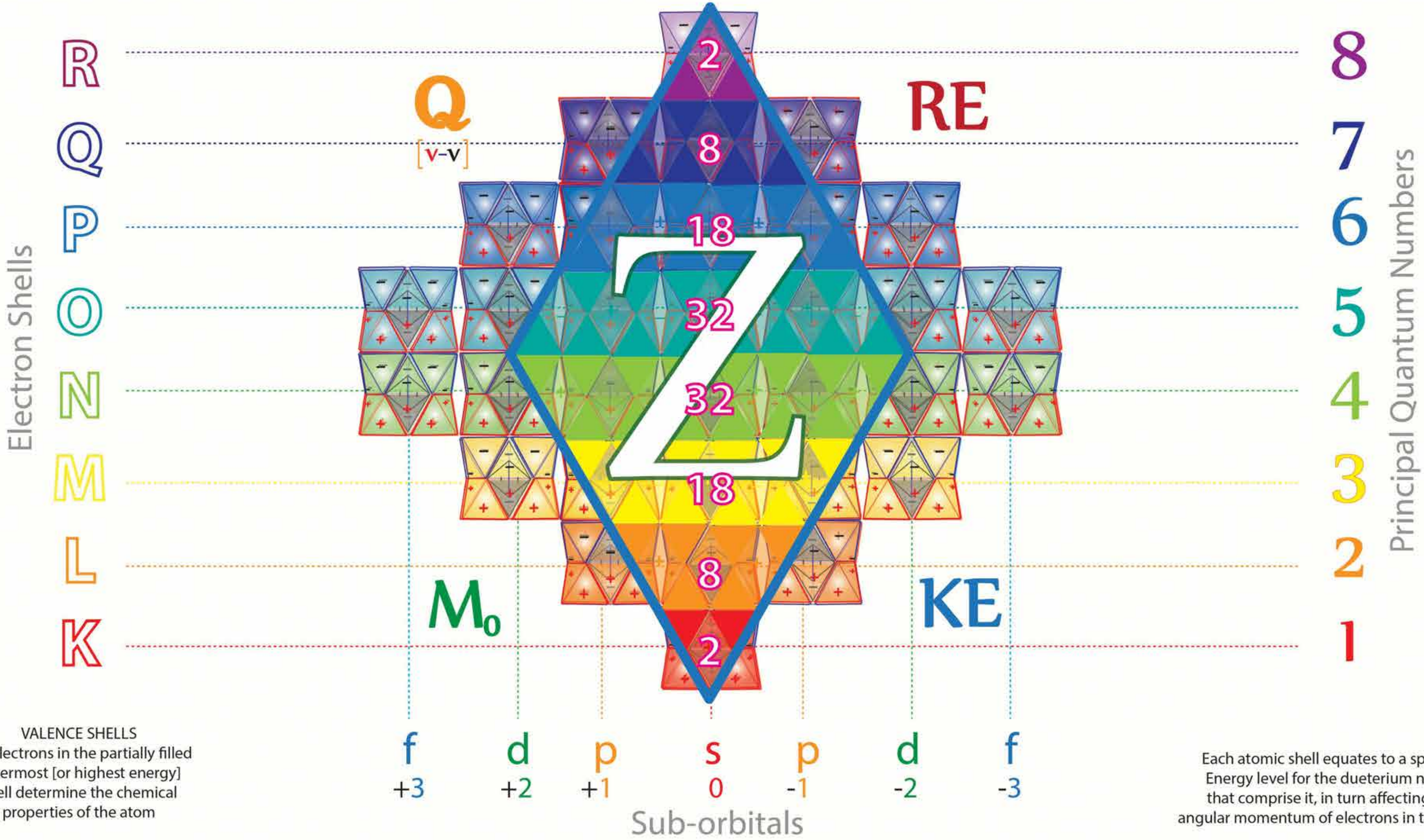
Electron Shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
M₀
KE
Z
2
8
18
32
32
18
8
2
Principal Quantum Numbers
8
7
6
5
4
3
2
1
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
Sub-orbitals

VALENCE SHELLS
The electrons in the partially filled outermost [or highest energy] shell determine the chemical properties of the atom

Each atomic shell equates to a specific Energy level for the dueterium nuclei that comprise it, in turn affecting the angular momentum of electrons in that shell

Periodic Harmonic motions

$x = A \cos(\omega t + \varphi)$

Circular motion

Much of the math in of modern physics is predicated on the assumption that π [where it appears] is related to the properties of a circle

$F = -kx$

Linear motion

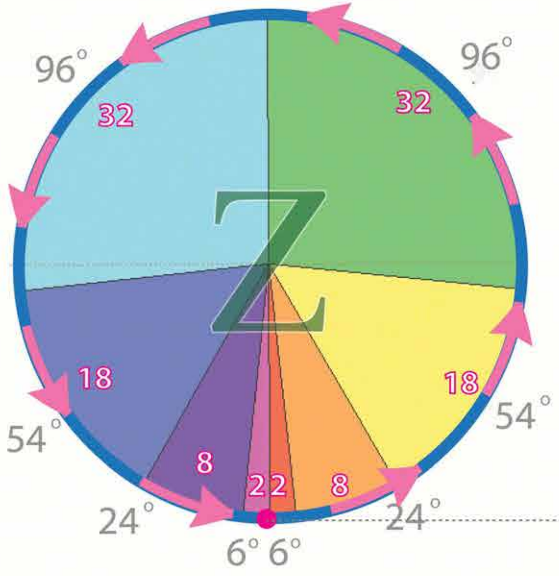
96°
32
32
96°
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

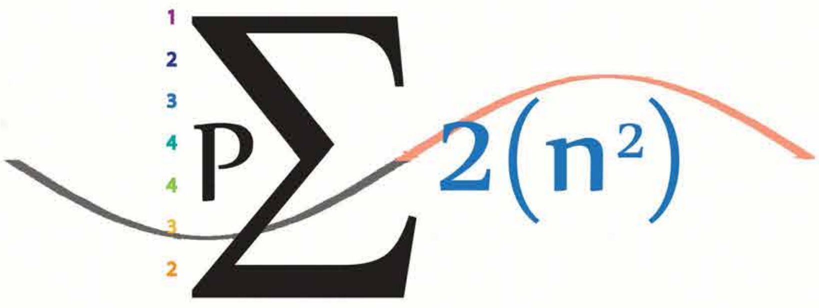
1
2
3
4
4
3
2
PΣ2(n²)

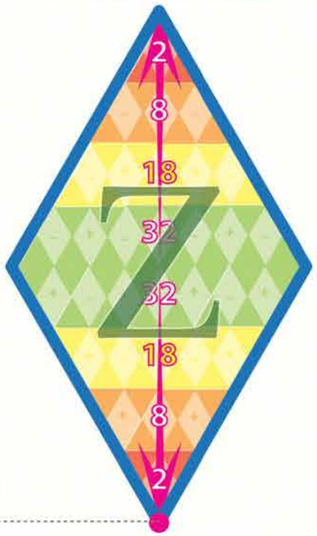
2
8
18
32
32
18
8
2
Z

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

circular harmonic motion

simple harmonic motion

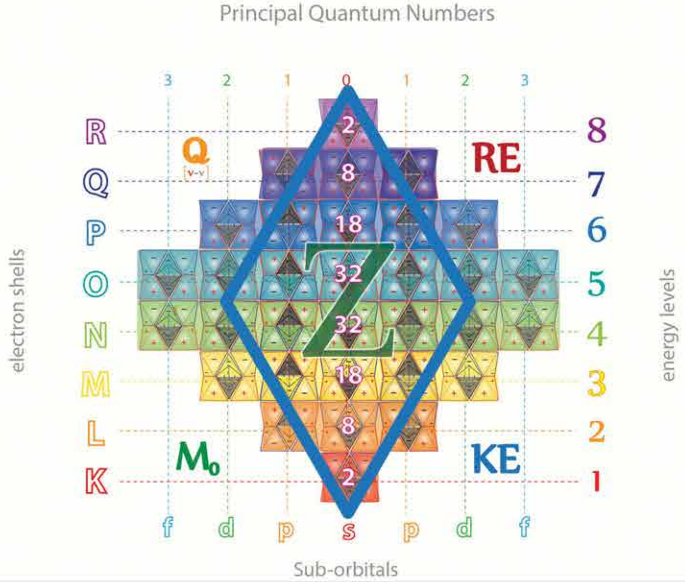
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
2
8
18
Z
32
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

Copyright ABRAHAM [2008] - All rights reserved

Baryon mass-energies

n

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

electron KEM field

hv^2

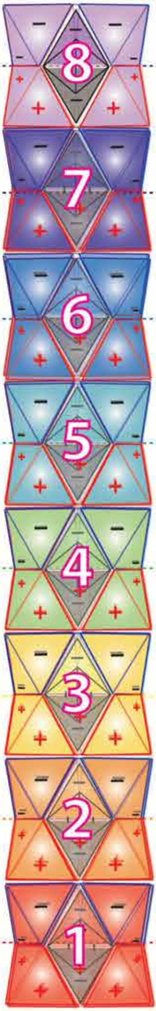

Baryon mass-energies	n	v	electron KEM field
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13

Compton frequencies

930 MeV

1.2 e20
electron rest Matter is velocity invariant

496 keV

spectral frequencies

13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

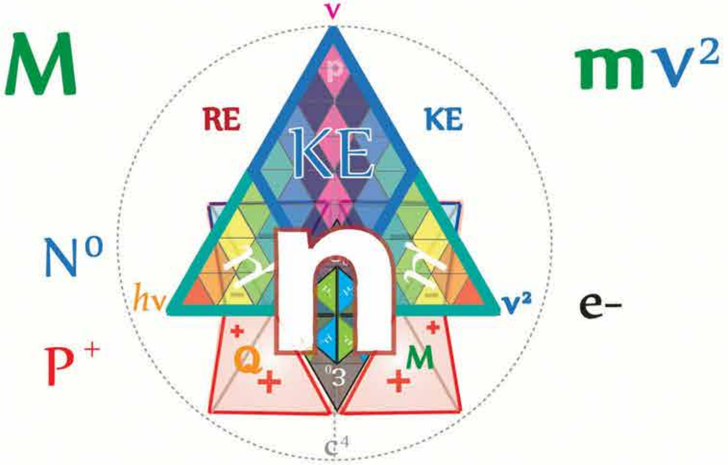

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses   lepton rest mass   KEM

$$\overset{8}{\underset{1}{n}}\left[\,[72(n)^2]+[12e19]+[m_e v^2]\,\right]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Baryons **930.947 MeV** + **KEM fields** **13.525 ev** + **electrons** **496.519 keV**

Mapping Planck mass-energy contributions to elementary Matter and isotopes

Schrodinger's quantum numbers

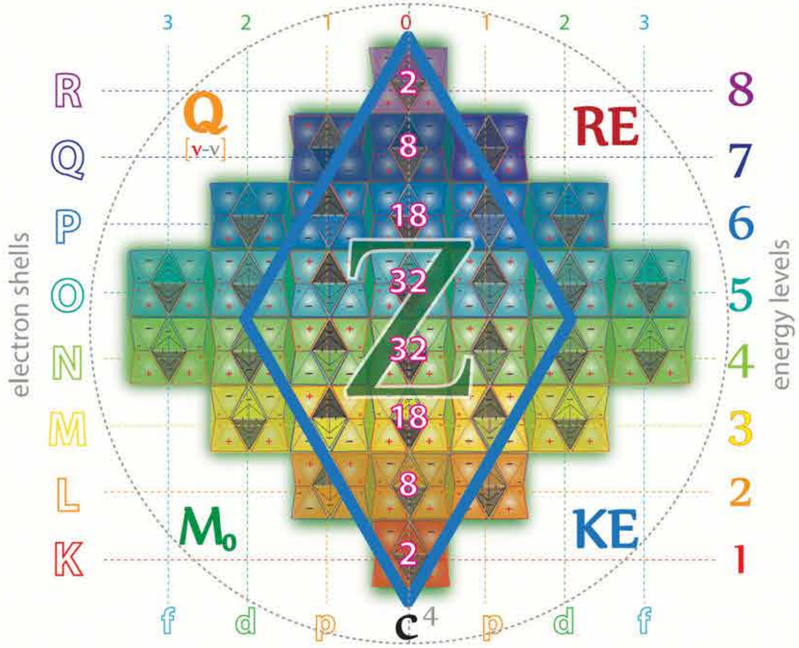
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
8
7
6
5
4
3
2
1
energy levels
Q
[v-v]
RE
M₀
KE
2
8
18
32
32
18
8
2
Z
f
d
p
c⁴
p
d
f

Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

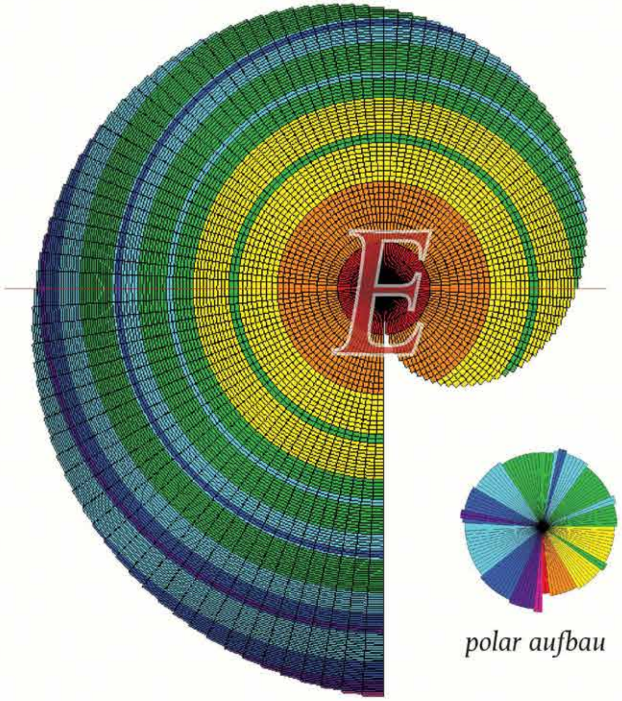

polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\left[\underbrace{72[v^2]}_{\text{Deuteron rest mass}} + \underbrace{v}_{\text{Spectral lines}} + \underbrace{1.20\text{ e}20\text{ v}}_{\text{electron rest mass}}\right]$$

reveal a quadratic formulation for all Z numbers

R
nuclei per shell $\sum$
K

STEP ONE

Periodic summation follows the atomic shell electron config

$$_{P}\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{1}^{120} Z$$ element number

STEP TWO

Periodic elements build up following the aufbau sequence

1 R
2
3
4
4
3
2
1 K

$$P\sum_{K}^{R} 2(x^2) =$$

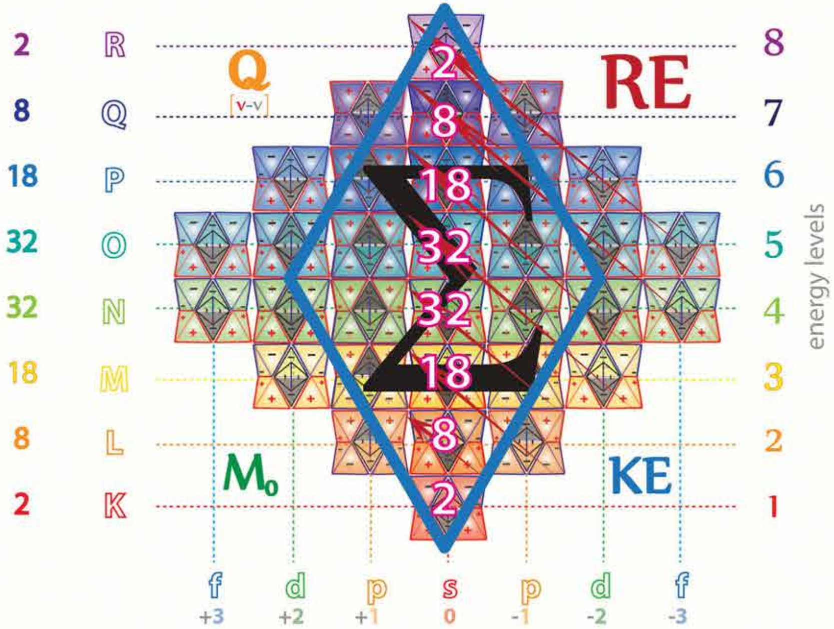

Each atomic shell can hold only a fixed number of deuterium nuclei

$_{P}\sum R = 2$	2 nuclei [74,496 ea] +	120	Unbinilium
$_{P}\sum Q = 8$	8 nuclei [69,780 ea] +	118	Ununoctium
$_{P}\sum P = 18$	18 nuclei [65,232 ea] +	110	Darmstadtium
$_{P}\sum O = 32$	32 nuclei [60,852 ea] +	92	Uranium
$_{P}\sum N = 32$	32 nuclei [56,640 ea] +	60	Neodymuim
$_{P}\sum M = 18$	18 nuclei [52,596 ea] +	28	Argon
$_{P}\sum L = 8$	8 nuclei [48,720 ea] +	10	Neon
$_{P}\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

Z# [z Protons [24-12]
z Neutrons [18-18]
z electrons [0-12]] n1-8

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number.
It is used as rule to determine which nuclei are stable or unstable.

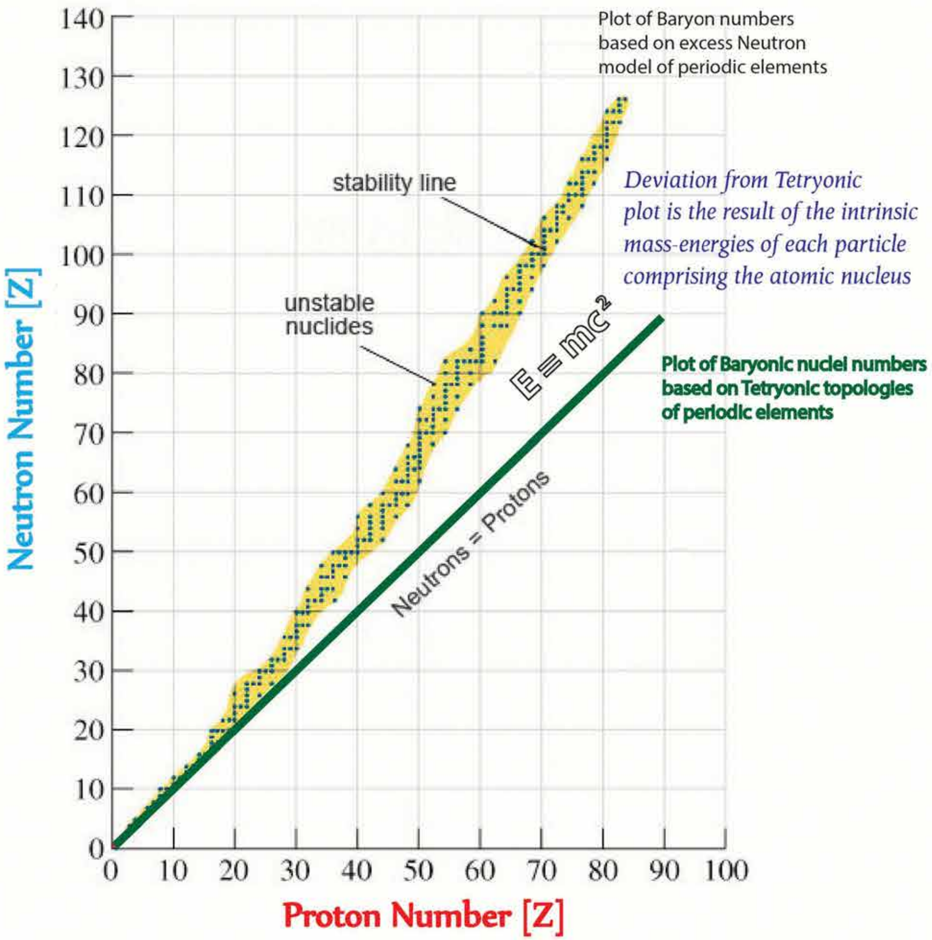

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

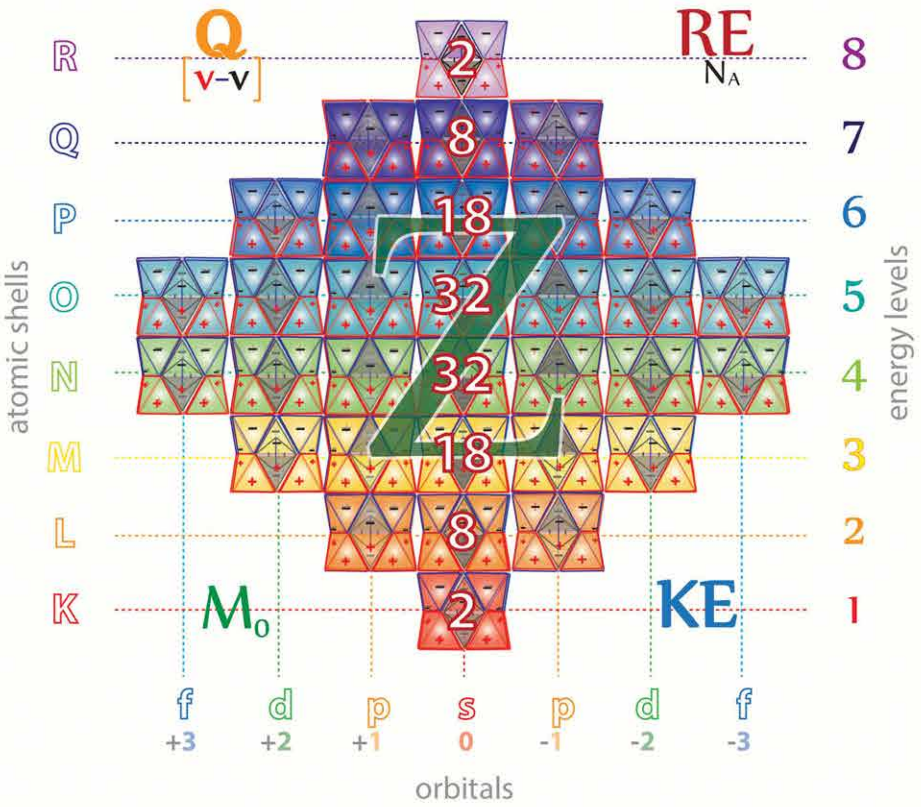
Q
[v-v]
RE
N_A
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
energy levels
M_0
KE
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
orbitals

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

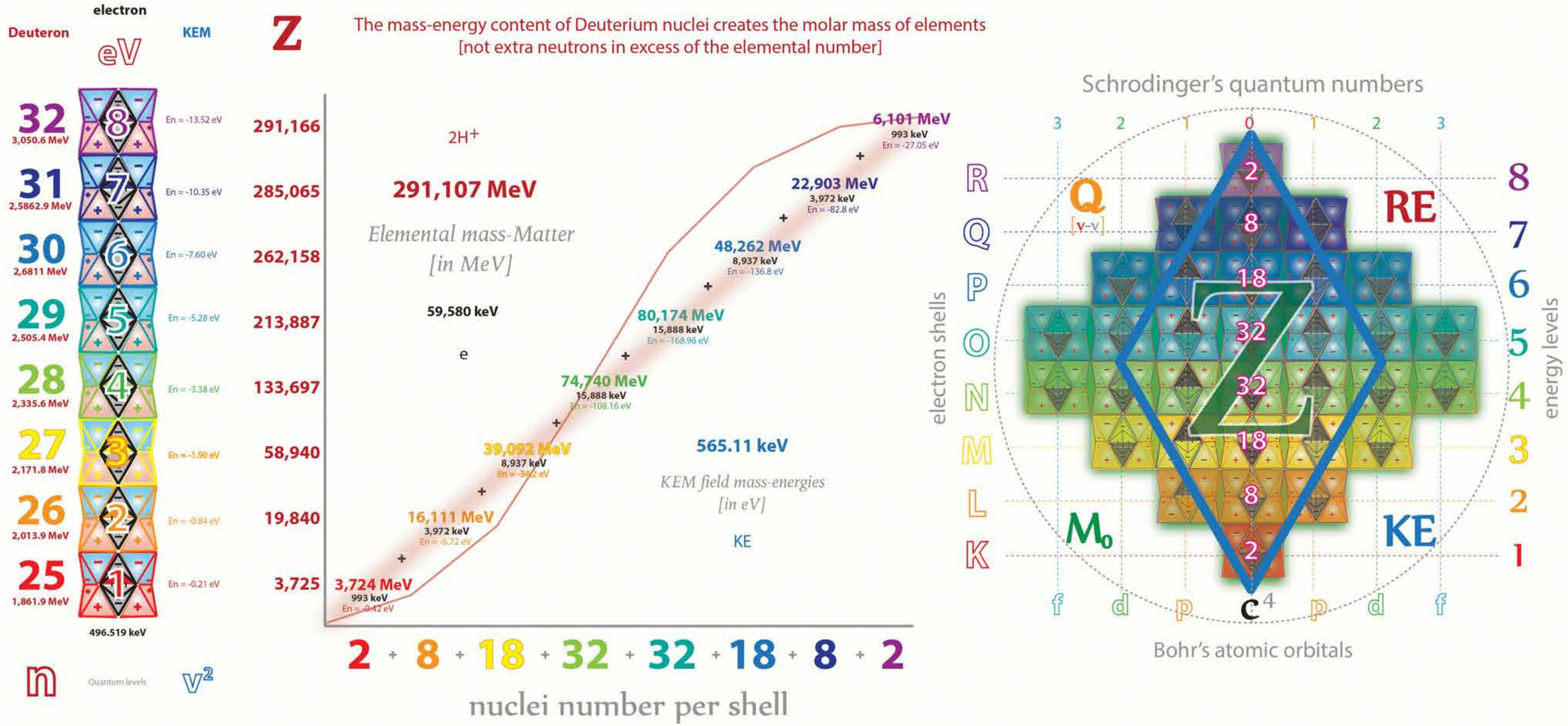
Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
8
En = -13.52 eV
291,166
31
2,5862.9 MeV
7
En = -10.35 eV
285,065
30
2,6811 MeV
6
En = -7.60 eV
262,158
29
2,505.4 MeV
5
En = -5.28 eV
213,887
28
2,335.6 MeV
4
En = -3.38 eV
133,697
27
2,171.8 MeV
3
En = -1.90 eV
58,940
26
2,013.9 MeV
2
En = -0.84 eV
19,840
25
1,861.9 MeV
1
En = -0.21 eV
3,725
496.519 keV
n
Quantum levels
v²
2H+
291,107 MeV
Elemental mass-Matter
[in MeV]
59,580 keV
e
6,101 MeV
993 keV
En = -27.05 eV
22,903 MeV
3,972 keV
En = -82.8 eV
48,262 MeV
8,937 keV
En = -136.8 eV
80,174 MeV
15,888 keV
En = -168.96 eV
74,740 MeV
15,888 keV
En = -108.16 eV
39,092 MeV
8,937 keV
En = -34.2 eV
16,111 MeV
3,972 keV
En = -6.72 eV
3,724 MeV
993 keV
En = -0.42 eV
565.11 keV
KEM field mass-energies
[in eV]
KE
2 + 8 + 18 + 32 + 32 + 18 + 8 + 2
nuclei number per shell
electron shells
R
Q
P
O
N
M
L
K
Schrodinger's quantum numbers
3
2
1
0
1
2
3
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M₀
KE
c⁴
8
7
6
5
4
3
2
1
energy levels
f
d
p
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	**496.519 keV**	+	**13.525 ev**

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

All elements are comprised of n level Duetrium nuclei

Baryons **KEM fields** **electrons**

Z [[72n²] + [12v²] + [1.2e20]]

1,861,949 MeV **13.525 eV** **496,519 keV**

The atomic shell energy levels of Deuterium nuclei in elements

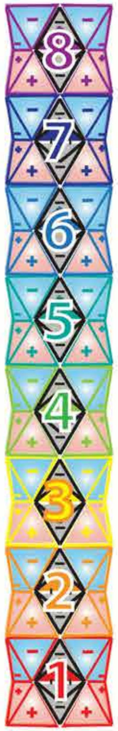

Deuteron	electron eV	KEM
32 3,050.6 MeV		En = -13.52 eV
31 2,5862.9 MeV		En = -10.35 eV
30 2,6811 MeV		En = -7.60 eV
29 2,505.4 MeV		En = -5.28 eV
28 2,335.6 MeV		En = -3.38 eV
27 2,171.8 MeV		En = -1.90 eV
26 2,013.9 MeV		En = -0.84 eV
25 1,861.9 MeV		En = -0.21 eV

496.519 keV

n v^2

Determines the spectral line [KEM field energies] of electrons bound to them

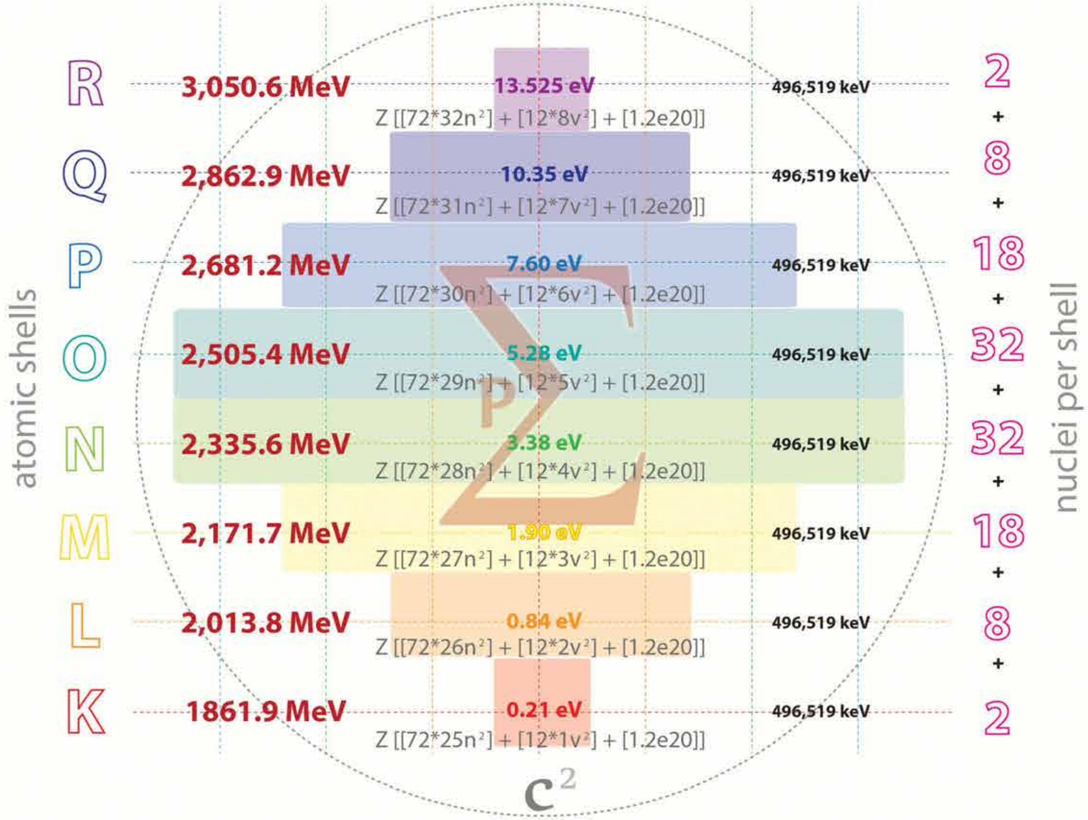

R	**3,050.6 MeV**	13.525 eV Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2	+
Q	**2,862.9 MeV**	10.35 eV Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	8	+
P	**2,681.2 MeV**	7.60 eV Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	18	+
O	**2,505.4 MeV**	5.28 eV Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	32	+
N	**2,335.6 MeV**	3.38 eV Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	32	+
M	**2,171.7 MeV**	1.90 eV Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	18	+
L	**2,013.8 MeV**	0.84 eV Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	8	+
K	**1861.9 MeV**	0.21 eV Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	2	

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

$2H^+$ **Z** γ

[72n²] [$M_e v^2$]

Elemental mass-Matter [in MeV]

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

the rest mass-Matter of bound photo-electrons is velocity invariant

Ionisation energies

Fractional quantum differentials

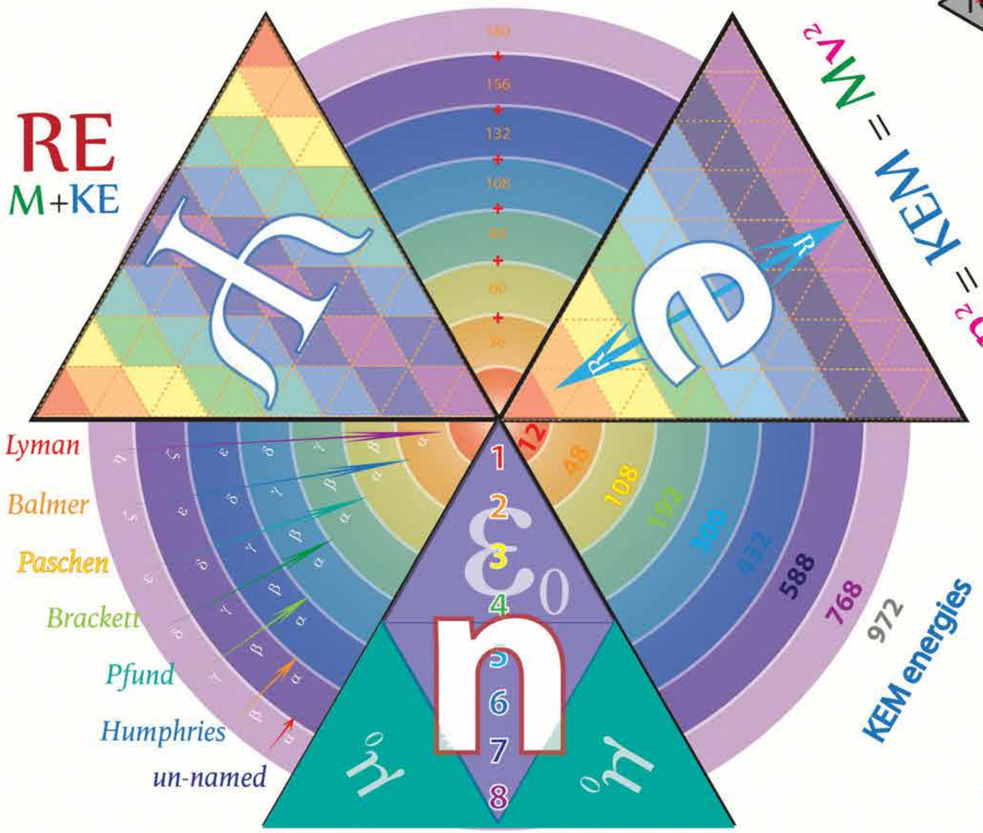
RE
M+KE
h
e
R
R
p² = KEM = Mv²
180
156
132
108
84
60
36
12
48
108
192
300
432
588
768
972
KEM energies
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
1
2
3
4
5
6
7
8
ε0
n

$$hf = \Delta M v = \Delta p$$

spectral lines are produced by accelerating electrons

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

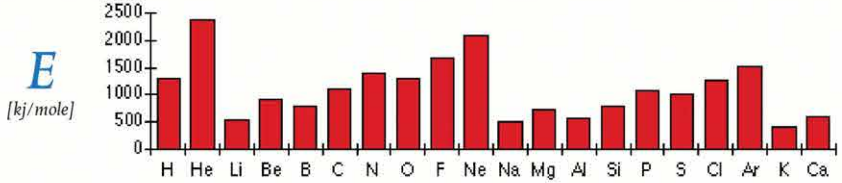

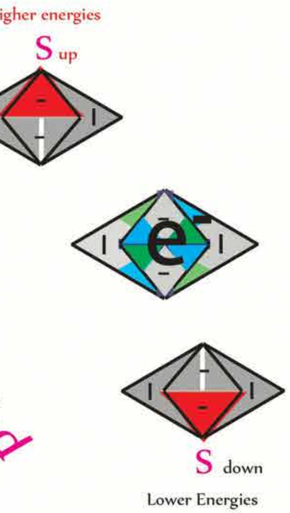
Higher energies
S up
e⁻
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

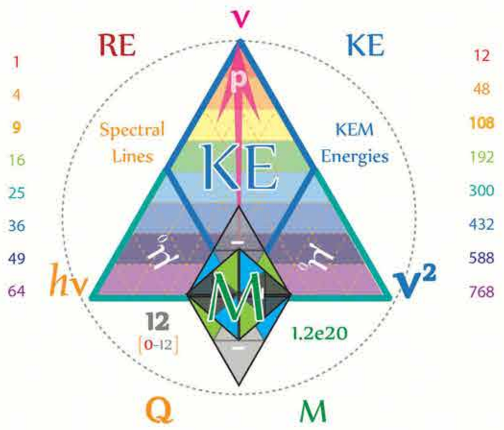
v
RE
KE
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768
p
Spectral Lines
KE
KEM Energies
hv
v²
M
12
[0-12]
1.2e20
Q
M

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

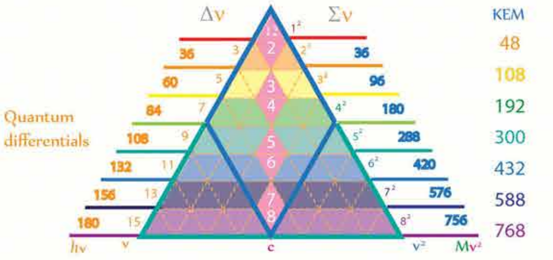
Δv
Σv
KEM
48
108
192
300
432
588
768
Quantum differentials
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
v
c
v²
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

$$\begin{matrix} 1 & R \\ 2 & \\ 3 & \\ 4 & \\ 4 & \\ 3 & \\ 2 & \\ 1 & K \end{matrix} \quad P\sum 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	
Hydrogen	0	

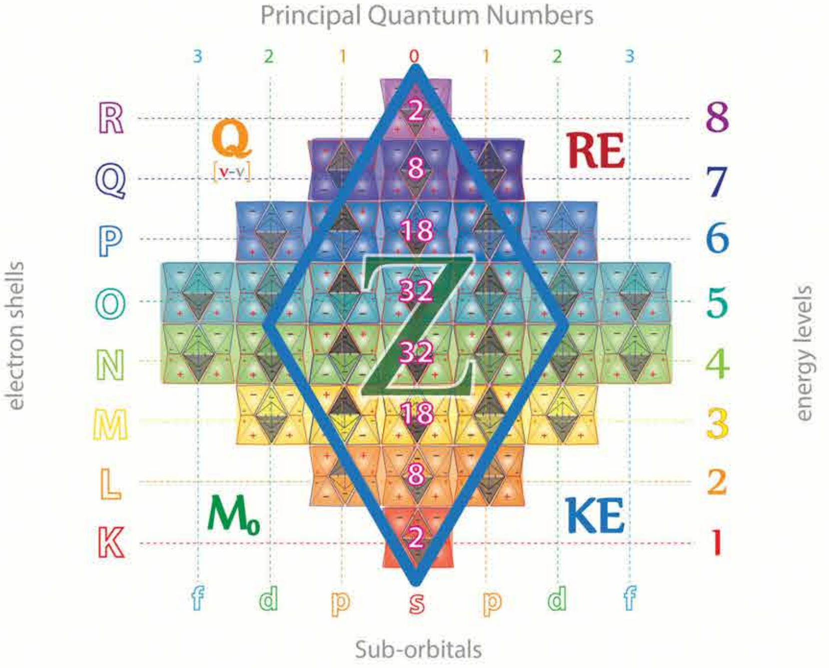
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M₀
KE
2
8
18
32
Z
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\boxed{{}_{K}^{R}\,P\sum_{25}^{32}\Bigg[\overset{\text{Baryon rest masses}}{\left[72(n)^2\right]}+\overset{\text{lepton rest mass}}{\left[12e19\right]}+\overset{\text{KEM}}{\left[m_e v^2\right]}\Bigg]_{1}^{8}}$$

Deuterium mass-energy per shell

The measured weight of Matter in gravitational fields is the result of planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

γ

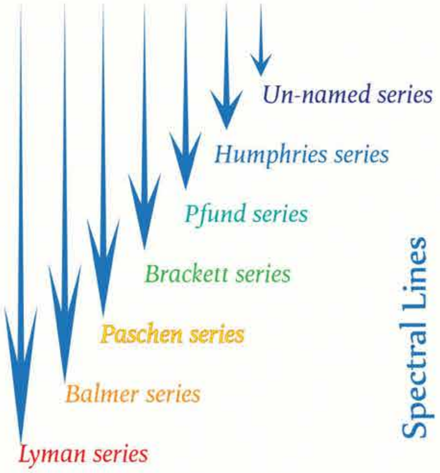
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$\boxed{Mv^2 = KEM = hcR_H}$$

Photon emission/absorption

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

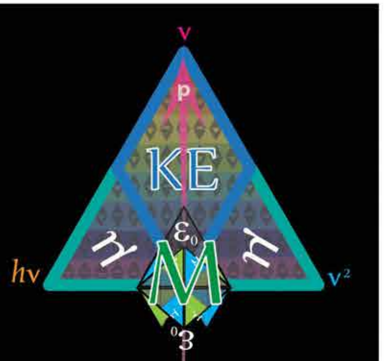

v
p
KE
ε₀
M
hv
v²
3

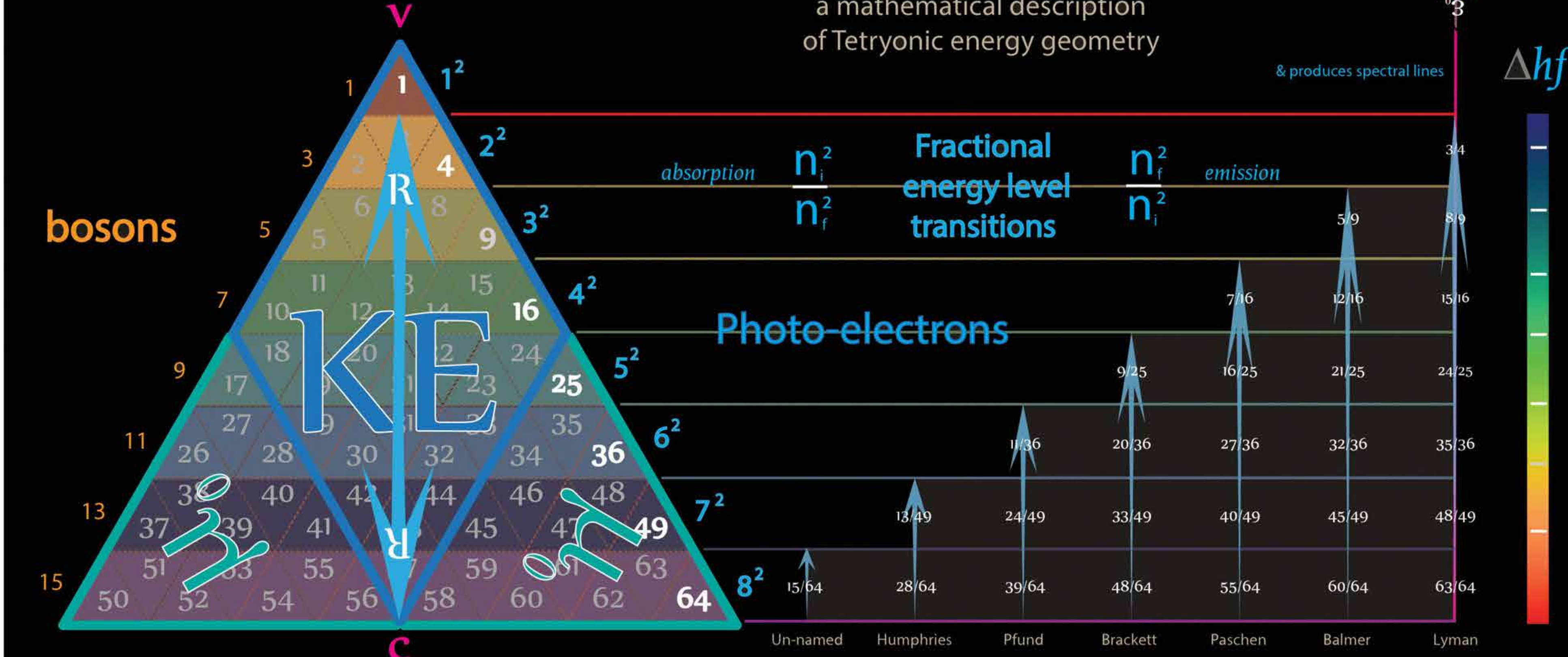

& produces spectral lines
Δhf
v
c
bosons
KE
R
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
1
4
9
16
25
36
49
64
absorption
$\frac{n_i^2}{n_f^2}$
Fractional energy level transitions
$\frac{n_f^2}{n_i^2}$
emission
Photo-electrons
3/4
5/9
8/9
7/16
12/16
15/16
9/25
16/25
21/25
24/25
11/36
20/36
27/36
32/36
35/36
13/49
24/49
33/49
40/49
45/49
48/49
15/64
28/64
39/64
48/64
55/64
60/64
63/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Ryberg's constant reflects the changing
energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

All of the transitions of photo-electrons bound to Hydrogen atoms can now be revealed in the fractional geometry of KEM field energies

KEM field energies

Copyright ABRAHAM [2008] - All rights reserved

Fermat's method of Factoring

also known as 'the difference of two squares' is used to factorise large numbers

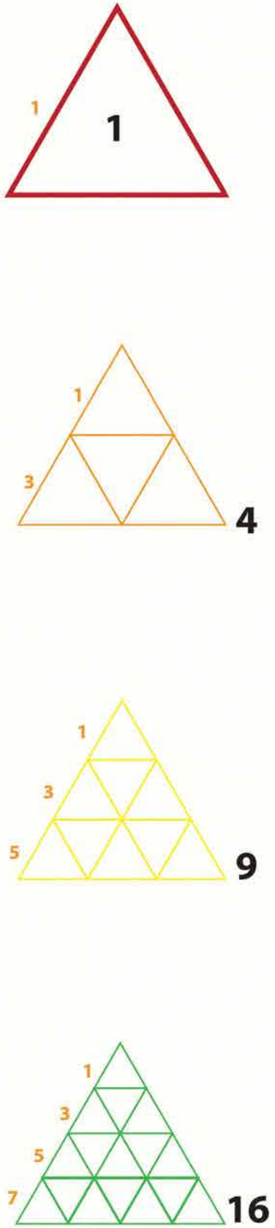
1
1
1
3
4
1
3
5
9
1
3
5
7
16

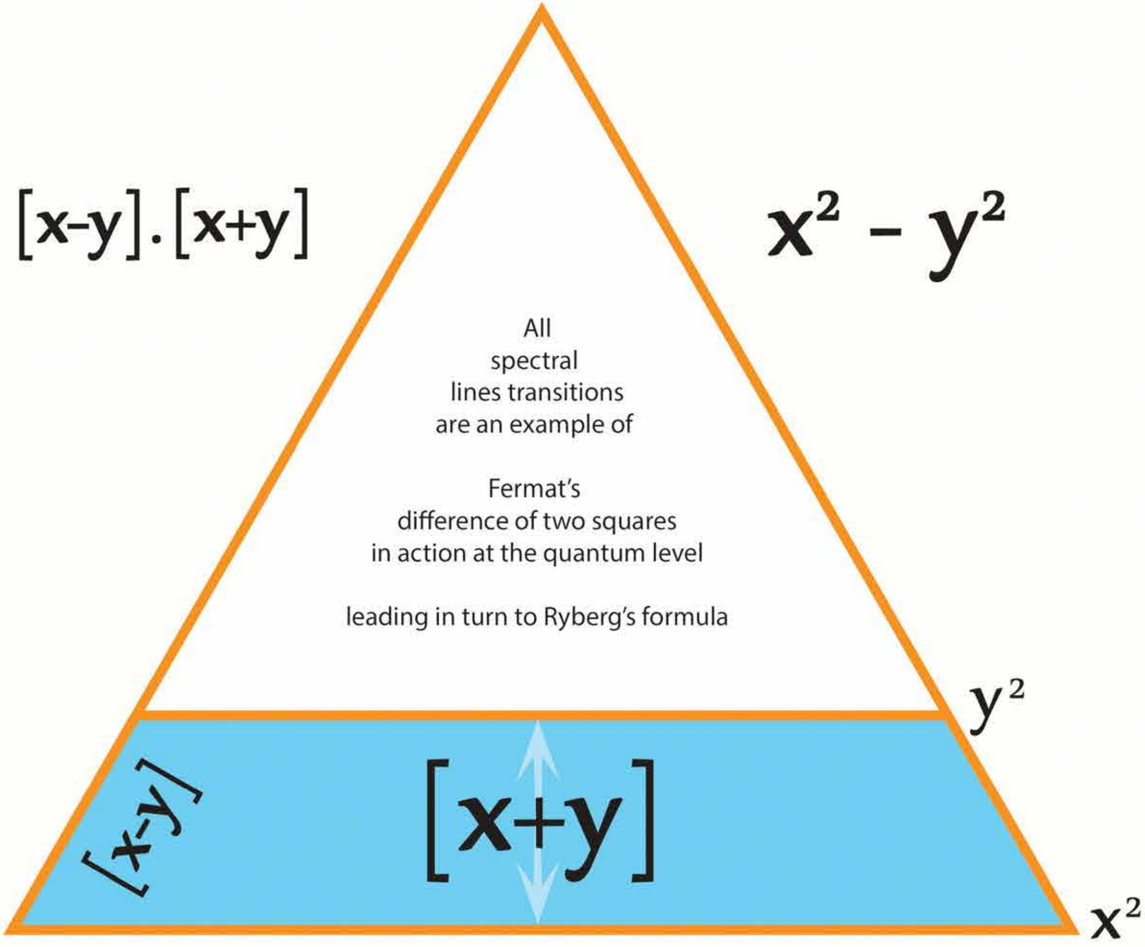
[x-y].[x+y]
x² - y²
y²
[x-y]
[x+y]
x²

All
spectral
lines transitions
are an example of

Fermat's
difference of two squares
in action at the quantum level

leading in turn to Ryberg's formula

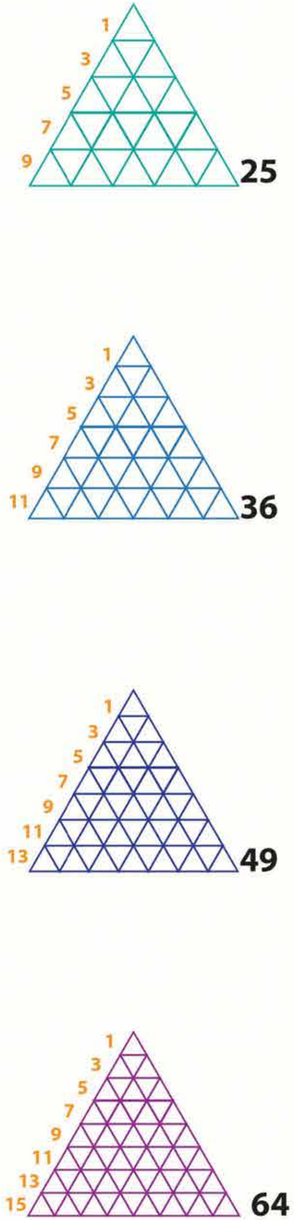
1
3
5
7
9
25
1
3
5
7
9
11
36
1
3
5
7
9
11
13
49
1
3
5
7
9
11
13
15
64

Fermat knew that every odd number could be written as the difference of two squares
or as revealed geometrically through Tetryonic theory's equilateral geometry,
every 'SQUARE' number is the sequential sum of ODD numbers

Fractions

A fraction is a number that shows how many equal parts there are

In quantum mechanics fractions appear in quantum steps
as a result of the equilateral geometry of Planck energy momenta

3/4

24/25

40/49

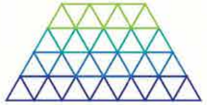

48/64

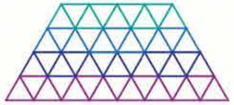

80/81

.987654321

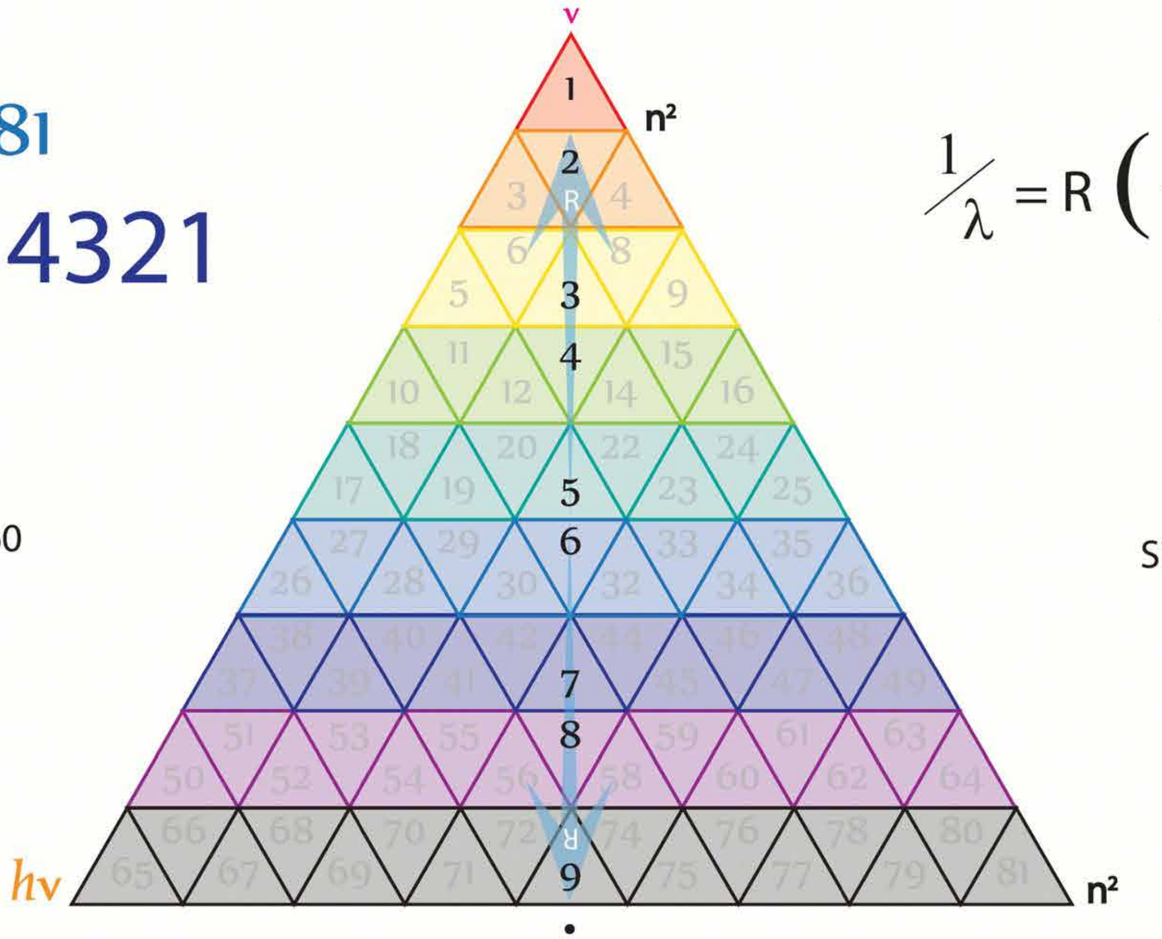
v
1
n²
2
3
R
4
6
8
5
3
9
11
4
15
10
12
14
16
18
20
22
24
17
19
5
23
25
27
29
6
33
35
26
28
30
32
34
36
38
40
42
44
46
48
37
39
41
7
45
47
49
51
53
55
8
59
61
63
50
52
54
56
58
60
62
64
66
68
70
72
74
76
78
80
65
67
69
71
R
9
75
77
79
81
hv
n²

$$\frac{1}{\lambda} = R\left(\frac{1}{1} - \frac{1}{81}\right)$$

.987654321

Cos 60

.5

1/2

Sin 60

.866025403

$\sqrt{3}/2$

Tetryonic geometry explains the fractional mathematics of Rydberg's formula

Copyright ABRAHAM [2008] - All rights reserved

Lyman spectral transitions

$$\Delta h\nu = \text{KEM} = hcR$$

Planck

13.525 eV

Rydberg

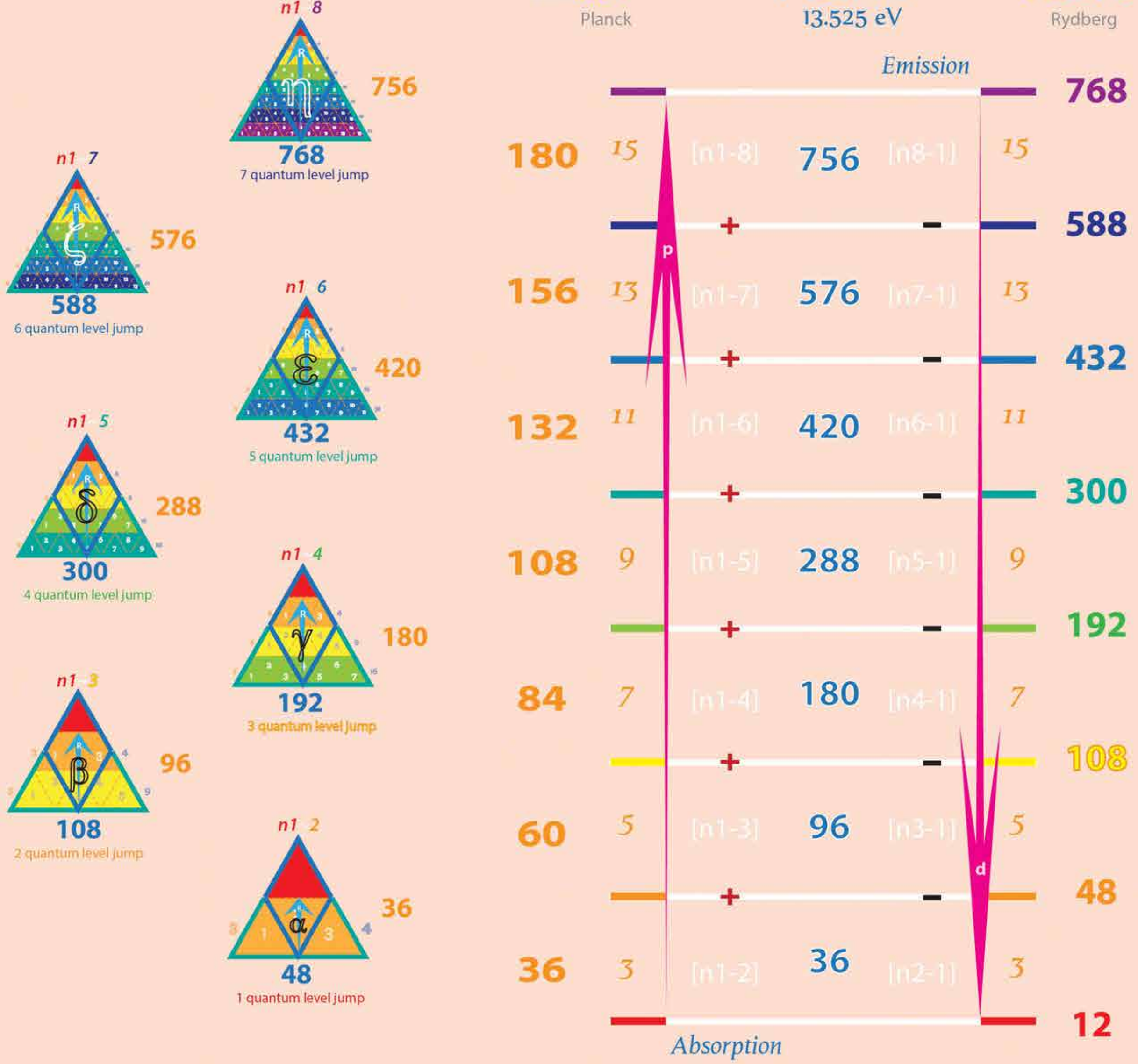

n1 8
756
768
7 quantum level jump

n1 7
576
588
6 quantum level jump

n1 6
420
432
5 quantum level jump

n1 5
288
300
4 quantum level jump

n1 4
180
192
3 quantum level jump

n1 3
96
108
2 quantum level jump

n1 2
36
48
1 quantum level jump

Emission

						768
180	15	[n1-8]	756	[n8-1]	15	
		+		−		588
156	13	[n1-7]	576	[n7-1]	13	
		+		−		432
132	11	[n1-6]	420	[n6-1]	11	
		+		−		300
108	9	[n1-5]	288	[n5-1]	9	
		+		−		192
84	7	[n1-4]	180	[n4-1]	7	
		+		−		108
60	5	[n1-3]	96	[n3-1]	5	
		+		−		48
36	3	[n1-2]	36	[n2-1]	3	
						12

p

d

Absorption

		λ (nm)	$\tilde{\nu}$ (m-1)	f (Hz)	E (eV)	Ω (m²/s)	
63/64	.9843	93.17075598	10,732,981.5	3.217666905 x 10^{15}	13.31395504	27.93188995	$R\left(\frac{1}{1}-\frac{1}{64}\right)$ .984375
48/49	.9795	93.62569132	10,680,829.01	3.202031983 x 10^{15}	13.24926138	28.06827613	$R\left(\frac{1}{1}-\frac{1}{49}\right)$ .9795
35/36	.9722	94.33539043	10,600,475.55	3.177942622 x 10^{15}	13.14958523	28.28103857	$R\left(\frac{1}{1}-\frac{1}{36}\right)$ .9722
24/25	.9600	95.53641971	10,467,212.43	3.137991343 x 10^{15}	12.98427616	28.64109809	$R\left(\frac{1}{1}-\frac{1}{25}\right)$ .96
15/16	.9375	97.82929378	10,221,5587.14	3.064444671 x 10^{15}	12.67995718	29.32848443	$R\left(\frac{1}{1}-\frac{1}{16}\right)$ .9375
8/9	.8888	103.1793333	9,691,863.362	2.90554754 x 10^{15}	12.02247792	30.93238594	$R\left(\frac{1}{1}-\frac{1}{9}\right)$ .888
3/4	.7500	122.2866172	8,177,509.712	2.451555737 x 10^{15}	10.14396575	36.66660556	$R\left(\frac{1}{1}-\frac{1}{4}\right)$ .75

$$\Delta M v^2$$

Copyright ABRAHAM [2008] - All rights reserved

Balmer spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck | 3.381 eV | Rydberg

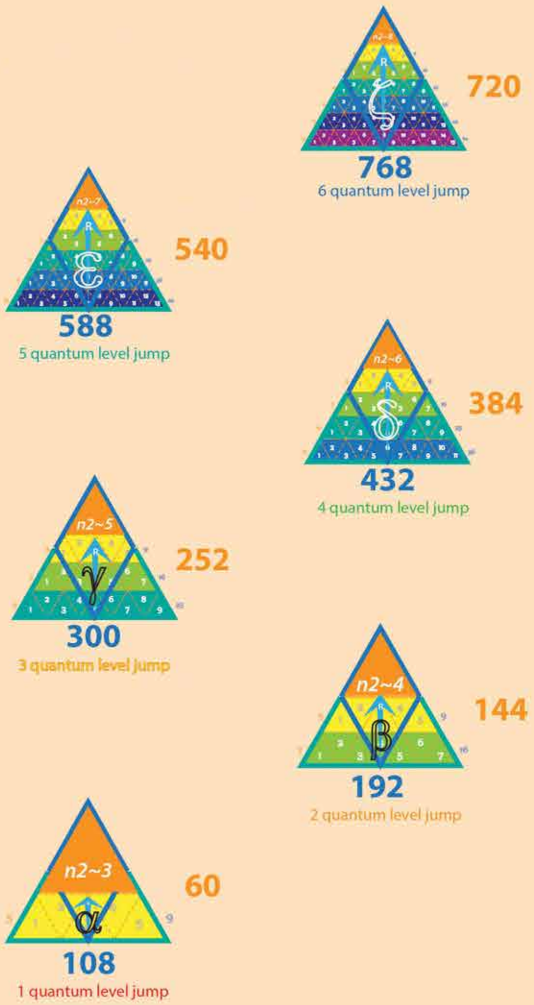

720
768
6 quantum level jump

540
588
5 quantum level jump

384
432
4 quantum level jump

252
300
3 quantum level jump

144
192
2 quantum level jump

60
108
1 quantum level jump

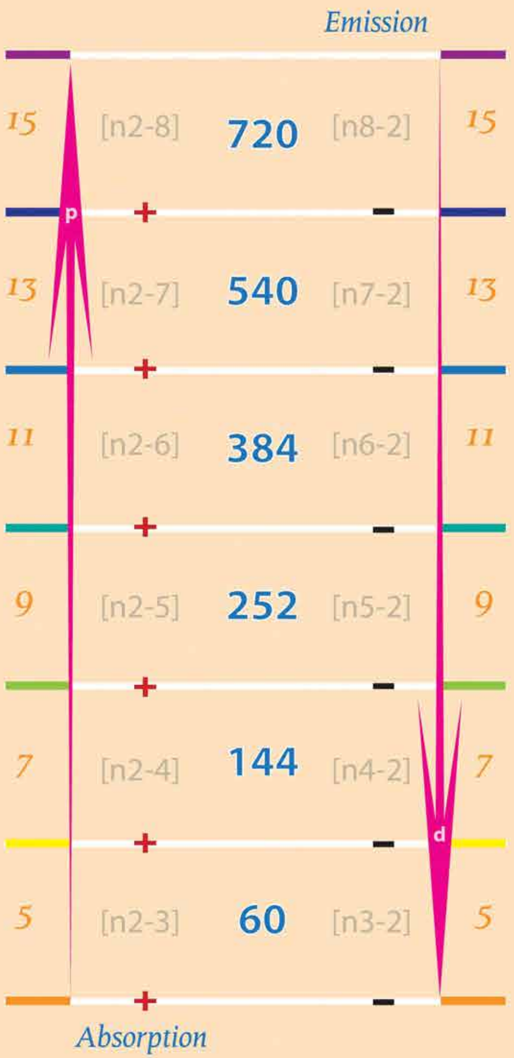

						768
180	15	[n2-8]	720	[n8-2]	15	
						588
156	13	[n2-7]	540	[n7-2]	13	
						432
132	11	[n2-6]	384	[n6-2]	11	
						300
108	9	[n2-5]	252	[n5-2]	9	
						192
84	7	[n2-4]	144	[n4-2]	7	
						108
60	5	[n2-3]	60	[n3-2]	5	
						48

Level	Ratio		λ (nm)	$\tilde{\nu}$ (m-1)	f (Hz)	E (eV)	Ω (m²/s)	
768	60/64	.2343	391.3171751	2,555,471.1785	7.661111677 x 10¹⁴	3.169989296	103.1079531	$R\left(\frac{1}{4}-\frac{1}{64}\right)$ .2343
588	45/49	.2295	399.4696163	2,5033,319.299	7.50476246 x 10¹⁴	3.105295637	101.0037092	$R\left(\frac{1}{4}-\frac{1}{49}\right)$ .2295
432	32/36	.2222	412.7173331	2,422,965.84	7.26386885 x 10¹⁴	3.00561948	97.76161483	$R\left(\frac{1}{4}-\frac{1}{36}\right)$ .2222
300	21/25	.2100	436.7379187	2,289,702.719	6.864356063 x 10¹⁴	2.840310409	92.38472601	$R\left(\frac{1}{4}-\frac{1}{25}\right)$ .21
192	12/16	.1875	489.1464689	2,044,377.428	6.128889342 x 10¹⁴	2.535991437	82.48636251	$R\left(\frac{1}{4}-\frac{1}{16}\right)$ .1875
108	5/9	.138	660.347733	1,514,353.65	4.539918031 x 10¹⁴	1.878512175	61.10100927	$R\left(\frac{1}{4}-\frac{1}{9}\right)$ .138

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Paschen spectral transitions

$$\underset{\text{Planck}}{\Delta h\nu} = \underset{1.502\ \text{eV}}{\text{KEM}} = \underset{\text{Rydberg}}{hcR}$$

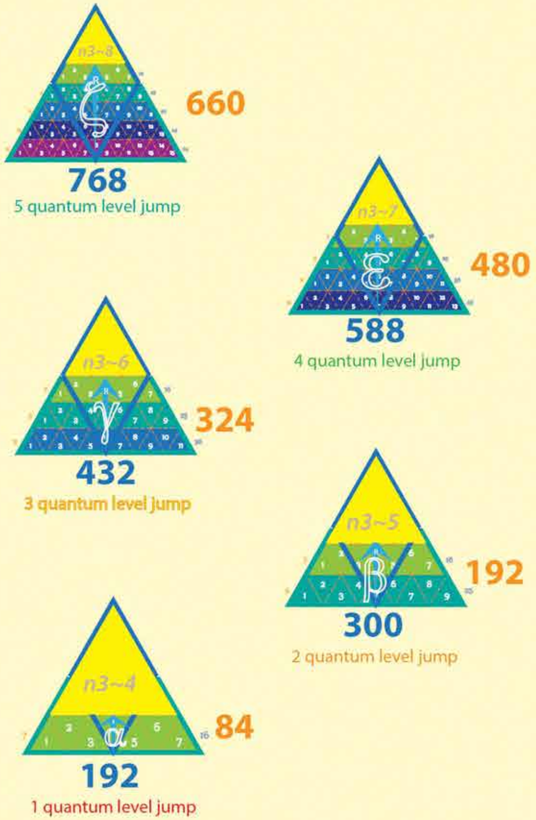
n3-8
660
768
5 quantum level jump
n3-7
480
588
4 quantum level jump
n3-6
324
432
3 quantum level jump
n3-5
192
300
2 quantum level jump
n3-4
84
192
1 quantum level jump

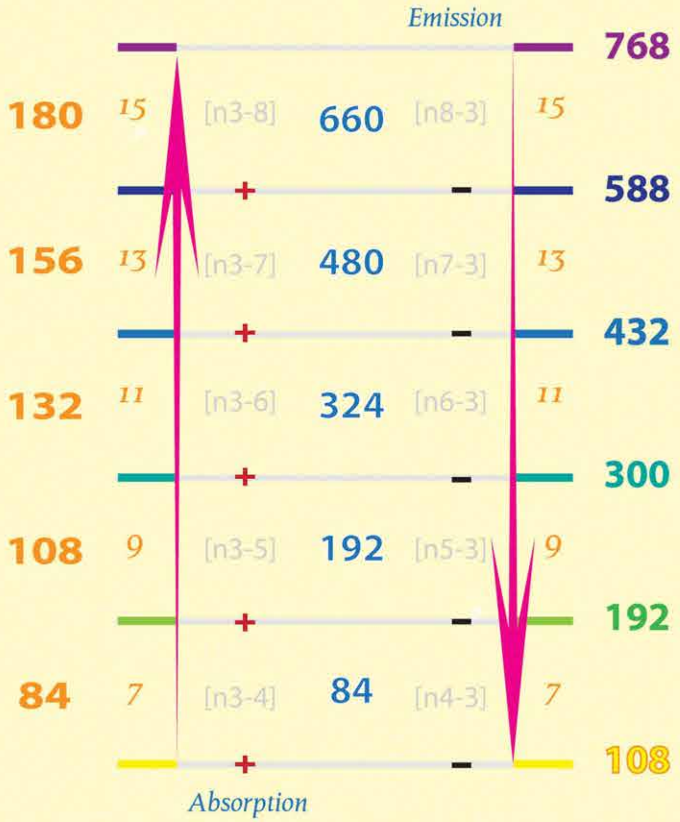
Emission
768
180
15
[n3-8]
660
[n8-3]
15
588
156
13
[n3-7]
480
[n7-3]
13
432
132
11
[n3-6]
324
[n6-3]
11
300
108
9
[n3-5]
192
[n5-3]
9
192
84
7
[n3-4]
84
[n4-3]
7
108
+
−
Absorption

55/64
.0954

λ	960.5057935	nm
ṽ	1,041,118.135	m-1
ƒ	3.121193646 x 10^{14}	Hz
E	**1.29147712**	**eV**
Ω	287.9523928	m^2/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$
.09548

40/49
.0907

λ	1,011.157466	nm
ṽ	988,965.6492	m-1
ƒ	2.9464844428 x 10^{14}	Hz
E	**1.226783461**	**eV**
Ω	303.1373822	m^2/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$
.0907

27/36
.8033

λ	1,100.579555	nm
ṽ	908,612.1902	m-1
ƒ	2.723950819 x 10^{14}	Hz
E	**1.127107305**	**eV**
Ω	329.94545	m^2/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$
.0833

16/25
.0711

λ	1,289.741666	nm
ṽ	775,349.069	m-1
ƒ	2.324438032 x 10^{14}	Hz
E	**0.961798233**	**eV**
Ω	386.6548243	m^2/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$
.0711

7/16
.0486

λ	1,886.707809	nm
ṽ	530,023.7776	m-1
ƒ	1.588971311 x 10^{14}	Hz
E	**0.657479261**	**eV**
Ω	565.6207715	m^2/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$
.0486

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Brackett spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck — 0.8451 eV — Rydberg

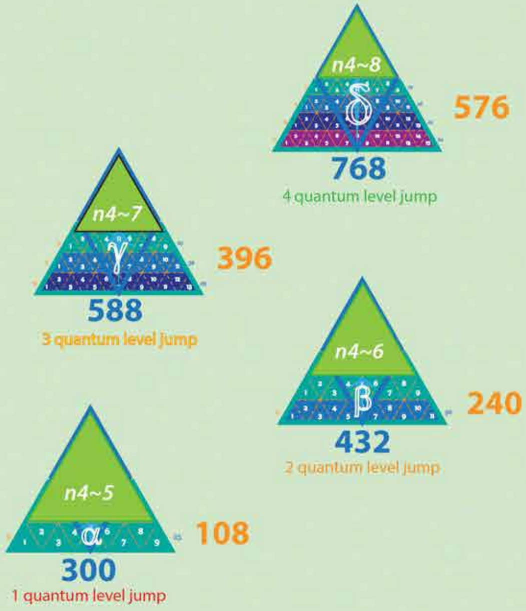

n4~8
δ
576
768
4 quantum level jump
n4~7
γ
396
588
3 quantum level jump
n4~6
β
240
432
2 quantum level jump
n4~5
α
108
300
1 quantum level jump

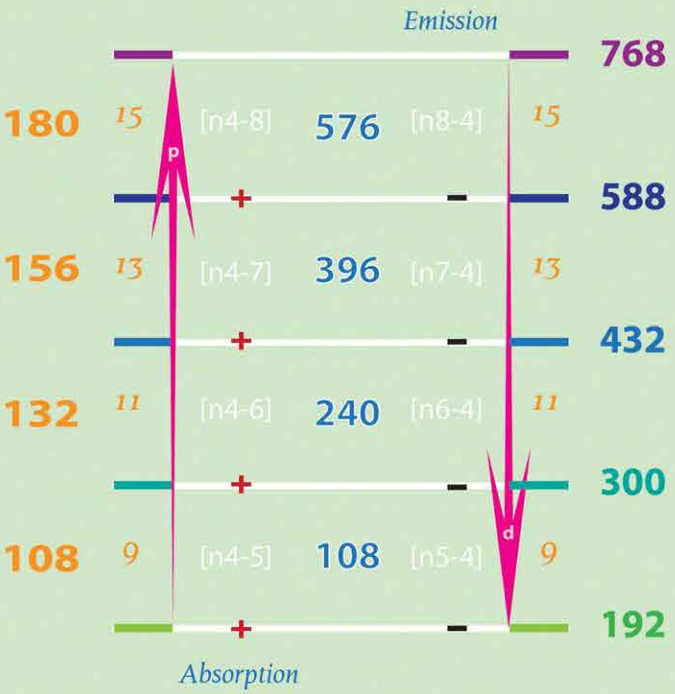

Emission
768
180
15
[n4-8]
576
[n8-4]
15
p
588
156
13
[n4-7]
396
[n7-4]
13
432
132
11
[n4-6]
240
[n6-4]
11
300
108
9
[n4-5]
108
[n5-4]
9
d
192
+
−
Absorption

48/64
.0468

λ	1,956.585876	nm
ṽ	511,094.357	m-1
f	1.532222335 x 10^{14}	Hz
E	0.634001085	eV
Ω	586.569689	m^2/s

$R\left(\frac{1}{16}-\frac{1}{64}\right)$ .0468

33/49
.04209

λ	2,178.92518	nm
ṽ	458,941.8716	m-1
f	1.375873118 x 10^{14}	Hz
E	0.569307096	eV
Ω	653.2253354	m^2/s

$R\left(\frac{1}{16}-\frac{1}{49}\right)$ .04209

20/36
.0347

λ	2,641.390932	nm
ṽ	378,588.4126	m-1
f	1.134979508 x 10^{14}	Hz
E	0.469630433	eV
Ω	791.8690801	m^2/s

$R\left(\frac{1}{16}-\frac{1}{36}\right)$ .0347

9/25
.0225

λ	4,076.220574	nm
ṽ	245,325.2914	m-1
f	7.35466721x 10^{13}	Hz
E	0.30432052	eV
Ω	1,222.020185	m^2/s

$R\left(\frac{1}{16}-\frac{1}{25}\right)$ .0225

ΔMv^2

Pfund spectral transitions

$\Delta h\nu$ = KEM = hcR

Planck P — 0.541 eV — Rydberg

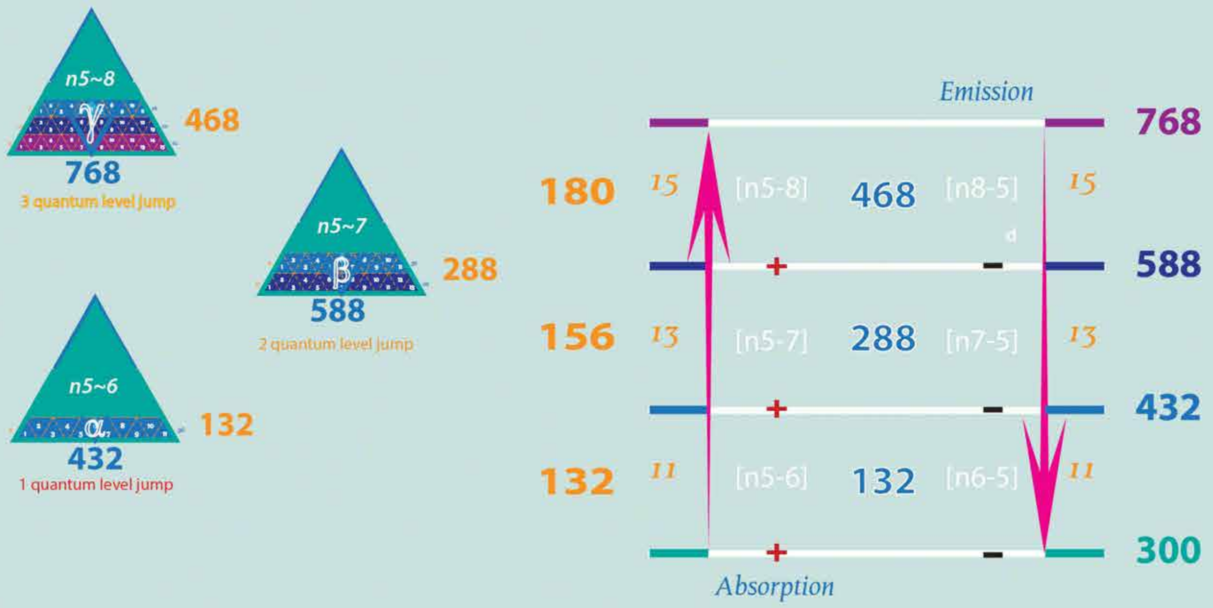
n5~8
γ
468
768
3 quantum level jump
n5~7
β
288
588
2 quantum level jump
n5~6
α
132
432
1 quantum level jump
Emission
768
588
432
300
180
156
132
15
13
11
[n5-8]
468
[n8-5]
[n5-7]
288
[n7-5]
[n5-6]
132
[n6-5]
+
-
d
Absorption

39/64

.0243

λ	3,762.665146	nm
$\tilde{\nu}$	265,769.0656	m-1
f	$7.967556145 \times 10^{13}$	Hz
E	**0.329678886**	**eV**
Ω	1,128.018633	m^2/s

$R\left(\frac{1}{25}-\frac{1}{64}\right)$

.0243

24/49

.0195

λ	4,681.284566	nm
$\tilde{\nu}$	213,616.5802	m-1
f	$6.404063965 \times 10^{13}$	Hz
E	**0.264985227**	**eV**
Ω	1,403.413807	m^2/s

$R\left(\frac{1}{25}-\frac{1}{49}\right)$

.01959

11/36

.0122

λ	7,503.951512	nm
$\tilde{\nu}$	133,263.1212	m-1
f	$3.995127867 \times 10^{13}$	Hz
E	**0.165309071**	**eV**
Ω	2,249.628068	m^2/s

$R\left(\frac{1}{25}-\frac{1}{36}\right)$

.0122

$\Delta M v^2$

Copyright ABRAHAM [2008] - All rights reserved

Humphreys spectral transitions

$$\Delta h\nu \quad = \quad \text{KEM} \quad = \quad hcR$$

Planck

0.375 eV

Rydberg

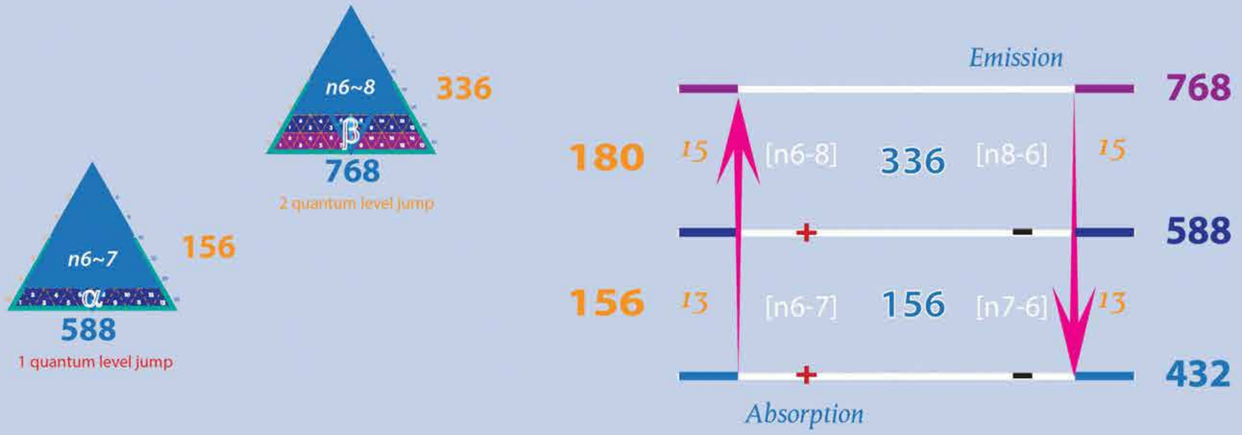
n6~8
β
336
768
2 quantum level jump
n6~7
α
156
588
1 quantum level jump
Emission
768
180
15
[n6-8]
336
[n8-6]
15
588
+
-
156
13
[n6-7]
156
[n7-6]
13
432
+
-
Absorption

28/64

.01215

λ	7,546.831235	nm
ṽ	132,505.9444	m-1
ƒ	$3.972428277 \times 10^{13}$	Hz
E	**0.164369815**	**eV**
Ω	1,128.018633	m²/s

$$R\left(\frac{1}{36} - \frac{1}{64}\right)$$

.01215

13/49

.00733

λ	12,445.01497	nm
ṽ	80,353.459	m-1
ƒ	$2.4089936098 \times 10^{13}$	Hz
E	**0.099676156**	**eV**
Ω	1,403.413807	m²/s

$$R\left(\frac{1}{36} - \frac{1}{49}\right)$$

.0073

$$\Delta M v^2$$

Copyright ABRAHAM [2008] - All rights reserved

Un-named spectral transition

$$\Delta h\nu \;=\; \text{KEM} \;=\; hcR$$

Planck   0.276 eV   Rydberg

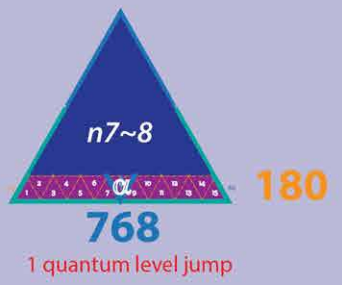
n7~8
1 2 3 4 5 6 7 α 9 10 11 12 13 14 15
180
768
1 quantum level jump

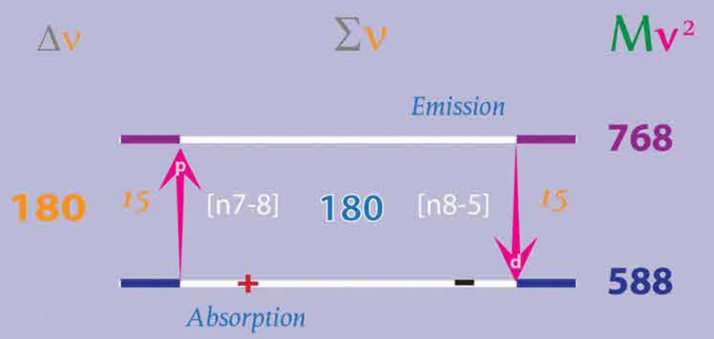
Δν
Σν
Mv²
Emission
768
180
15
p
[n7-8]
180
[n8-5]
15
d
+
−
588
Absorption

15/64

.0047

λ	19,174.54158	nm
$\tilde{\nu}$	52,152.48541	m-1
ƒ	$1.563492179 \times 10^{13}$	Hz
E	0.064693658	eV
Ω	5,748.382952	m²/s

$$R\left(\frac{1}{49}-\frac{1}{64}\right)$$

.0047

ΔMv^2

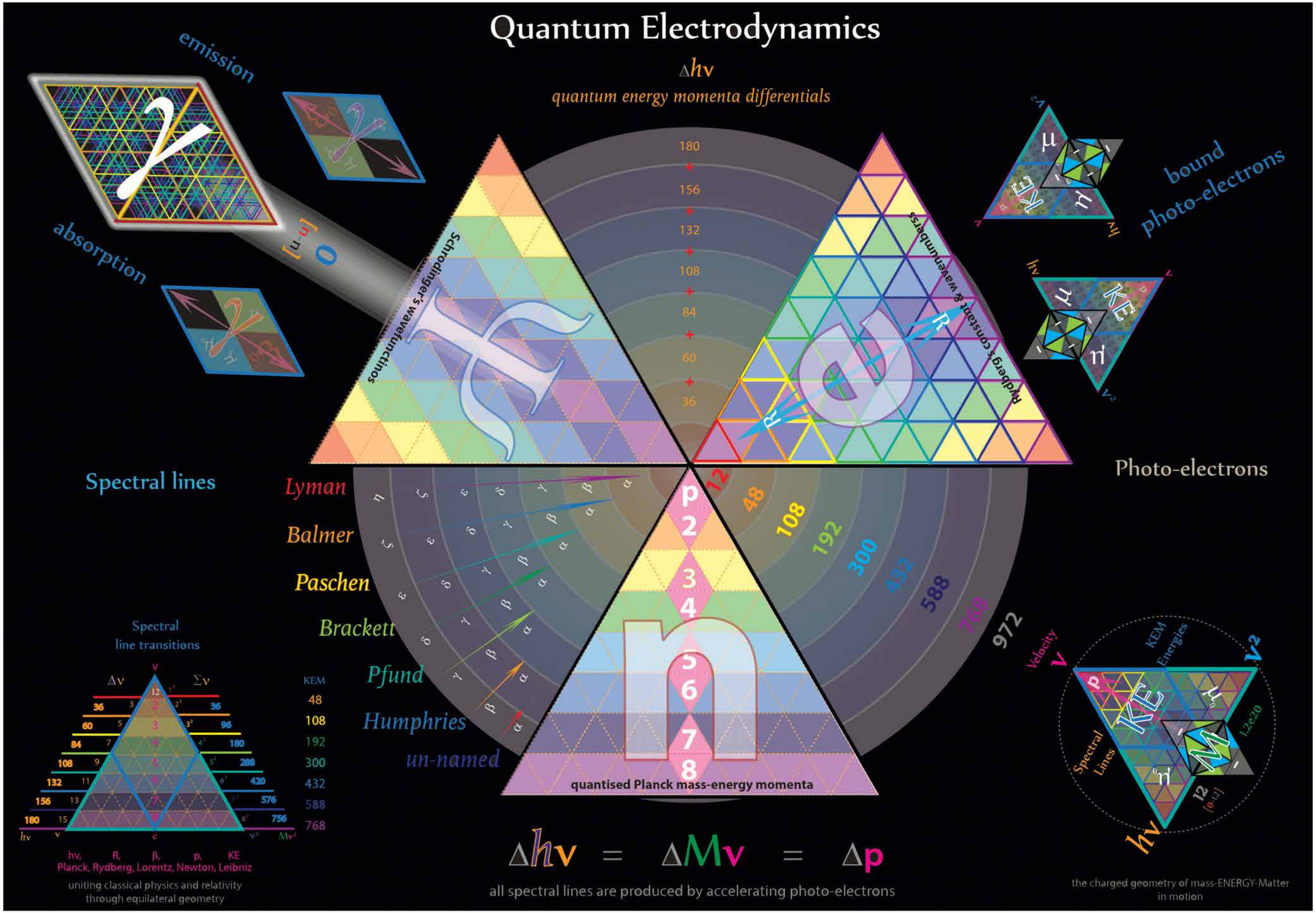

Quantum Electrodynamics

$\Delta h\nu$

quantum energy momenta differentials

emission
absorption
[n-n]
0
Schrodinger's wavefunctions
180
156
132
108
84
60
36
Rydberg's constant & wavenumbers
R
R
bound
photo-electrons
μ
KE
KE
μ
Photo-electrons
Spectral lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
α β γ δ ε ζ η
p
2
3
4
5
6
7
8
n
12
48
108
192
300
432
588
768
972
quantised Planck mass-energy momenta
Spectral
line transitions
Δv
Σv
v
12
36
60
84
108
132
156
180
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
v
c
v²
Mv²
KEM
48
108
192
300
432
588
768
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
Velocity
v
KEM
Energies
v²
p
KE
M
1.2e20
Spectral
Lines
12
[0-12]
hv

$$\Delta h\nu = \Delta M\nu = \Delta p$$

all spectral lines are produced by accelerating photo-electrons

uniting classical physics and relativity through equilateral geometry

the charged geometry of mass-ENERGY-Matter in motion

Tetryonics 96.11 - Atomic spectral series transitions

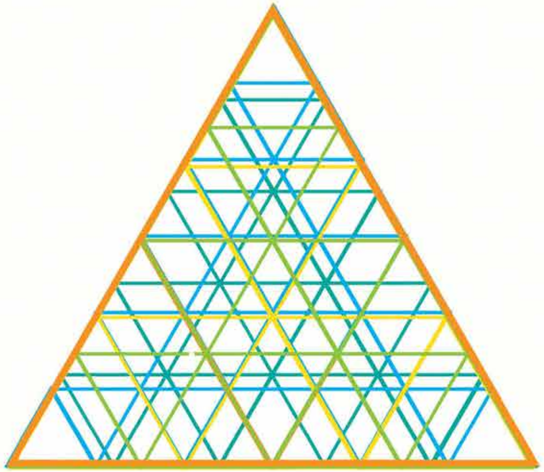

Fractionals and fractals

White light
[EM radiation]
is comprised of many
superpositioned
frequencies

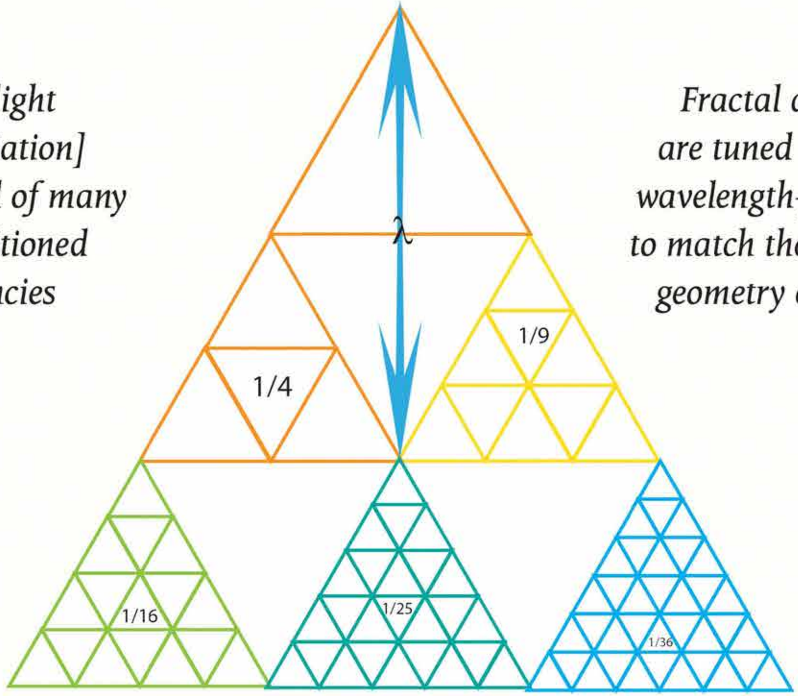

Fractal antennas
are tuned to specific
wavelength-frequencies
to match the equilateral
geometry of photons

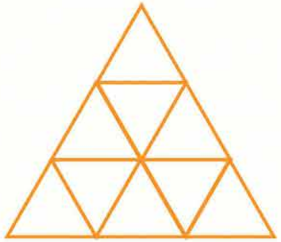

ninths

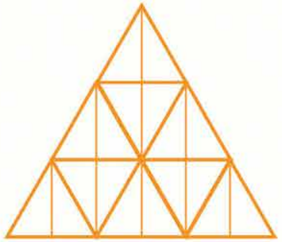

eigteenths

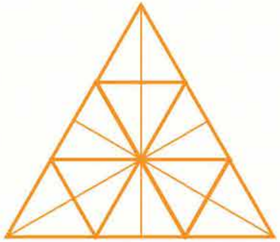

Eighteenths

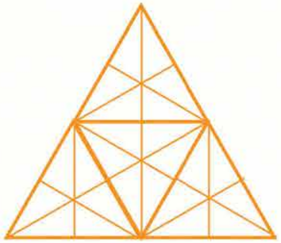

Twenty-fourths

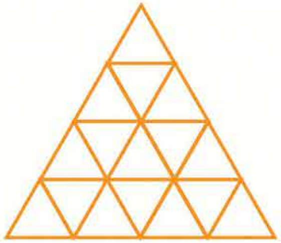

Sizteenths

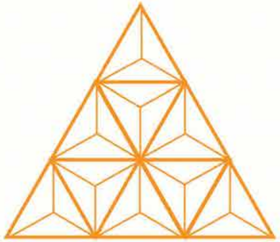

Twenty-sevenths

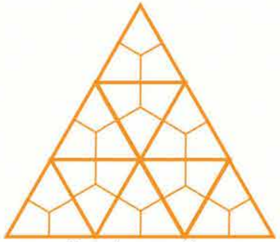

Twenty-sevenths

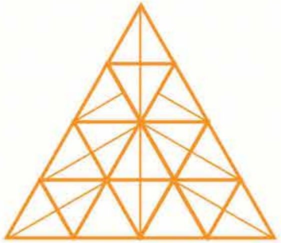

Thirty-seconds

Polygons and Angles

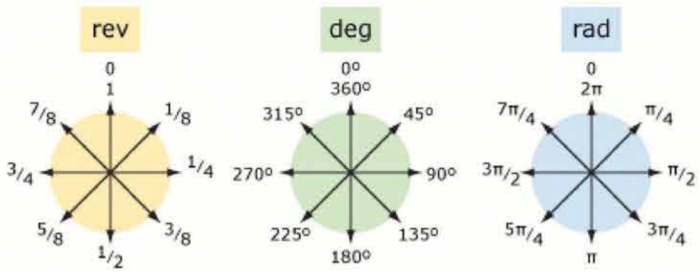
rev
0
1
1/8
1/4
3/8
1/2
5/8
3/4
7/8
deg
0°
360°
45°
90°
135°
180°
225°
270°
315°
rad
0
2π
π/4
π/2
3π/4
π
5π/4
3π/2
7π/4

polygon exterior and interior angles

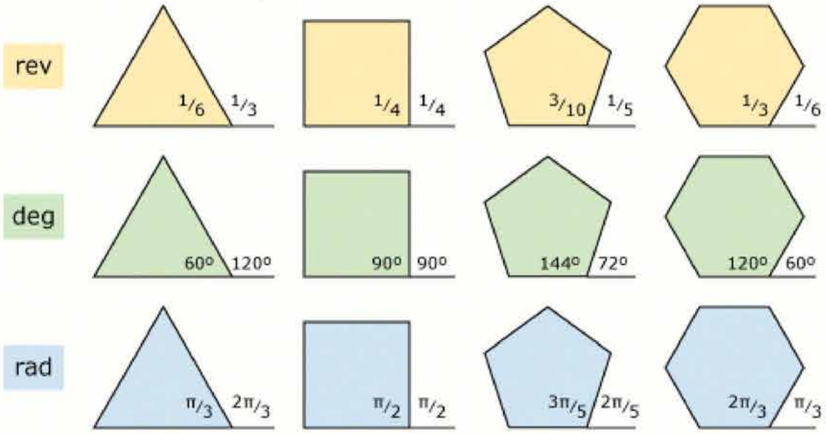
rev
1/6
1/3
1/4
1/4
3/10
1/5
1/3
1/6
deg
60°
120°
90°
90°
144°
72°
120°
60°
rad
π/3
2π/3
π/2
π/2
3π/5
2π/5
2π/3
π/3

dividing the circle

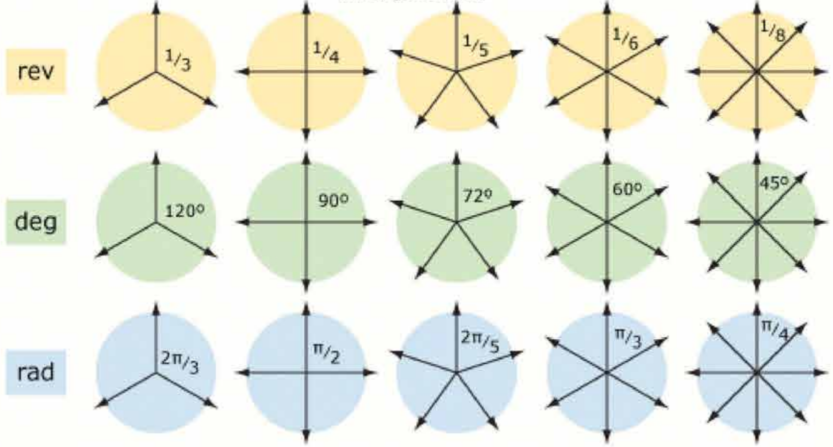
rev
1/3
1/4
1/5
1/6
1/8
deg
120°
90°
72°
60°
45°
rad
2π/3
π/2
2π/5
π/3
π/4

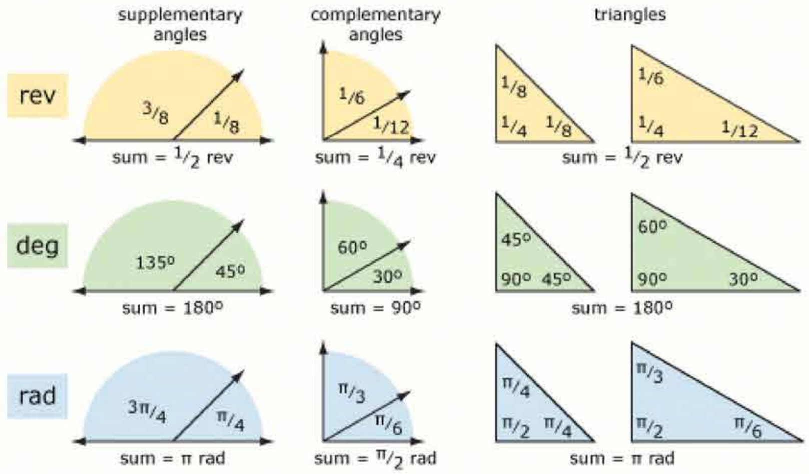
supplementary angles
complementary angles
triangles
rev
3/8
1/8
sum = 1/2 rev
1/6
1/12
sum = 1/4 rev
1/8
1/4
1/8
1/6
1/4
1/12
sum = 1/2 rev
deg
135°
45°
sum = 180°
60°
30°
sum = 90°
45°
90°
45°
60°
90°
30°
sum = 180°
rad
3π/4
π/4
sum = π rad
π/3
π/6
sum = π/2 rad
π/4
π/2
π/4
π/3
π/2
π/6
sum = π rad

Copyright ABRAHAM [2008] - All rights reserved

Pi vs Tau

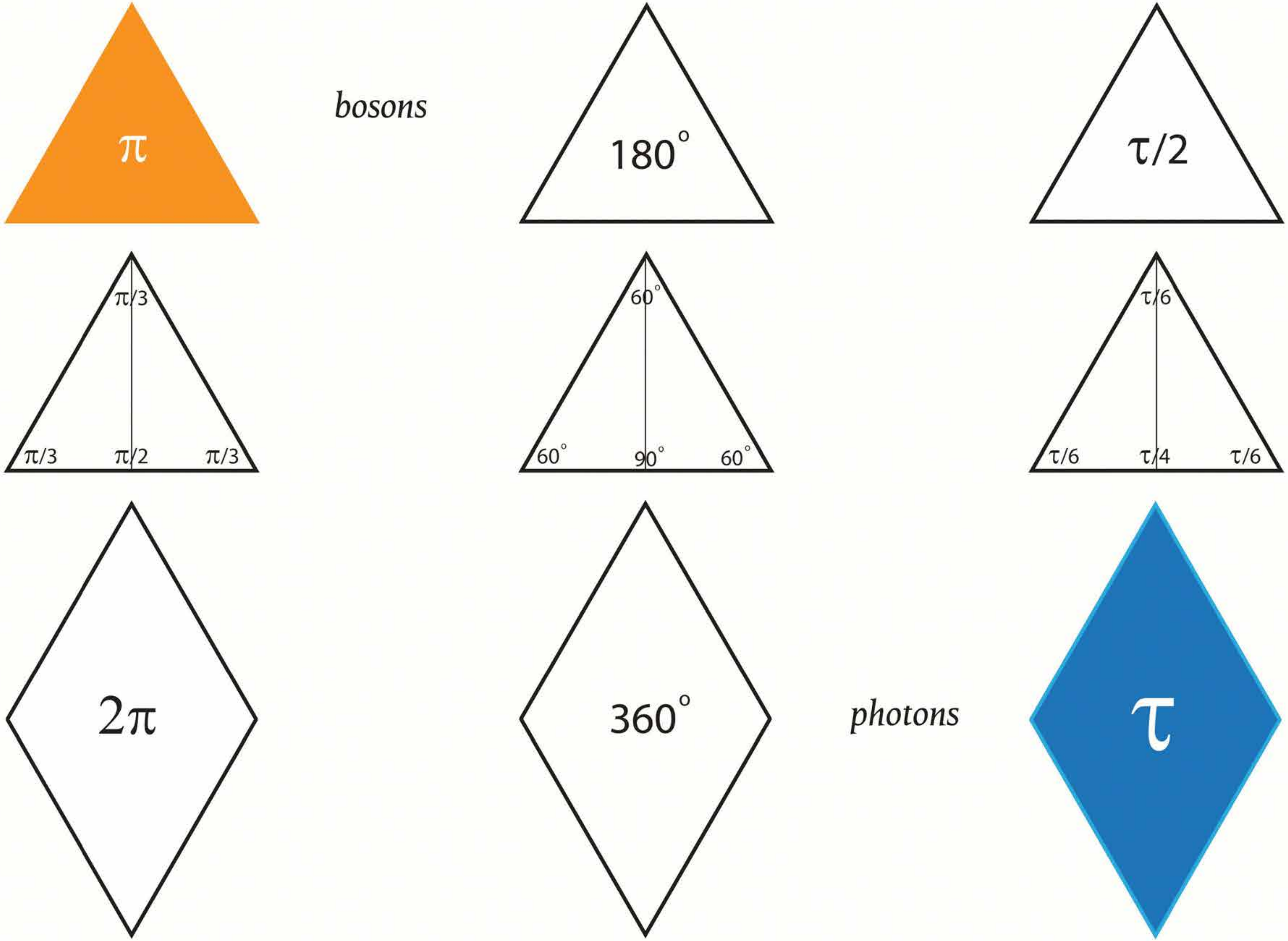
π
bosons
180°
τ/2
π/3
π/3
π/2
π/3
60°
60°
90°
60°
τ/6
τ/6
τ/4
τ/6
2π
360°
photons
τ

Triangular dissection of an equilateral triangle

is a way of dividing up a original triangle into smaller equilateral triangles,
such that none of the smaller triangles overlap

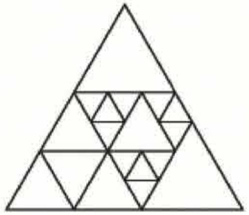

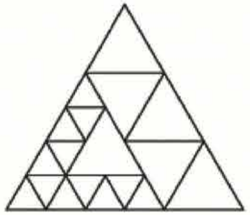

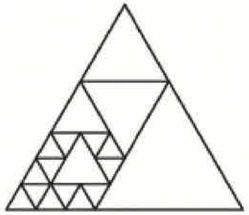

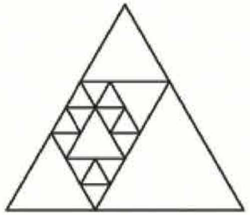

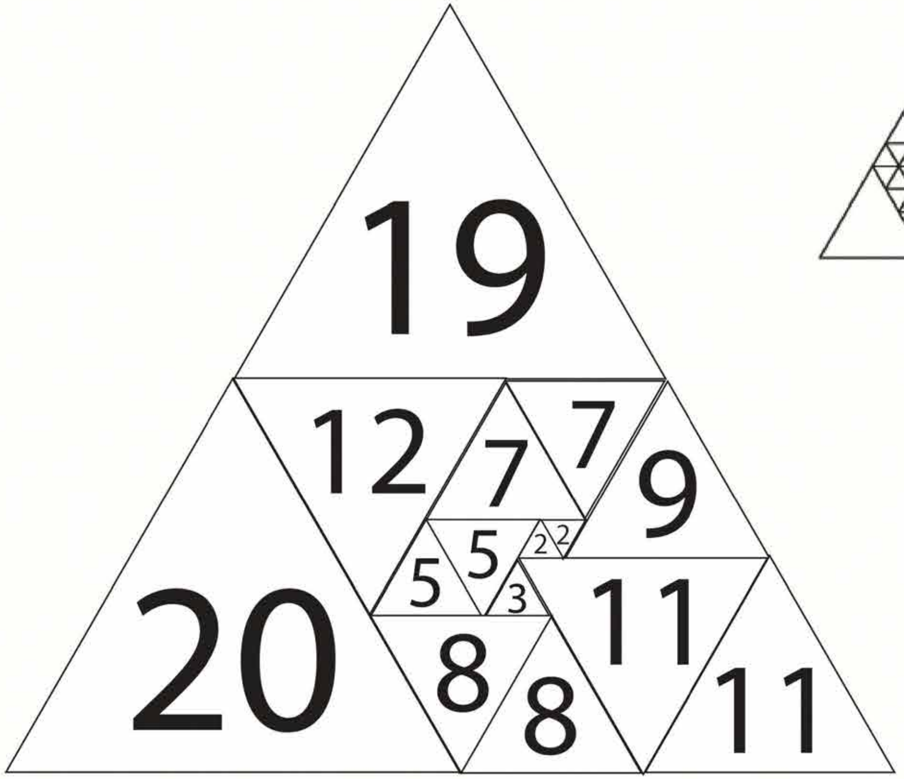

lowest order perfect equilateral triangle dissected by equilateral triangles

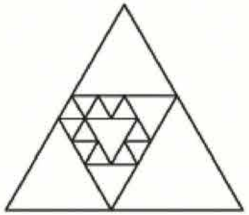

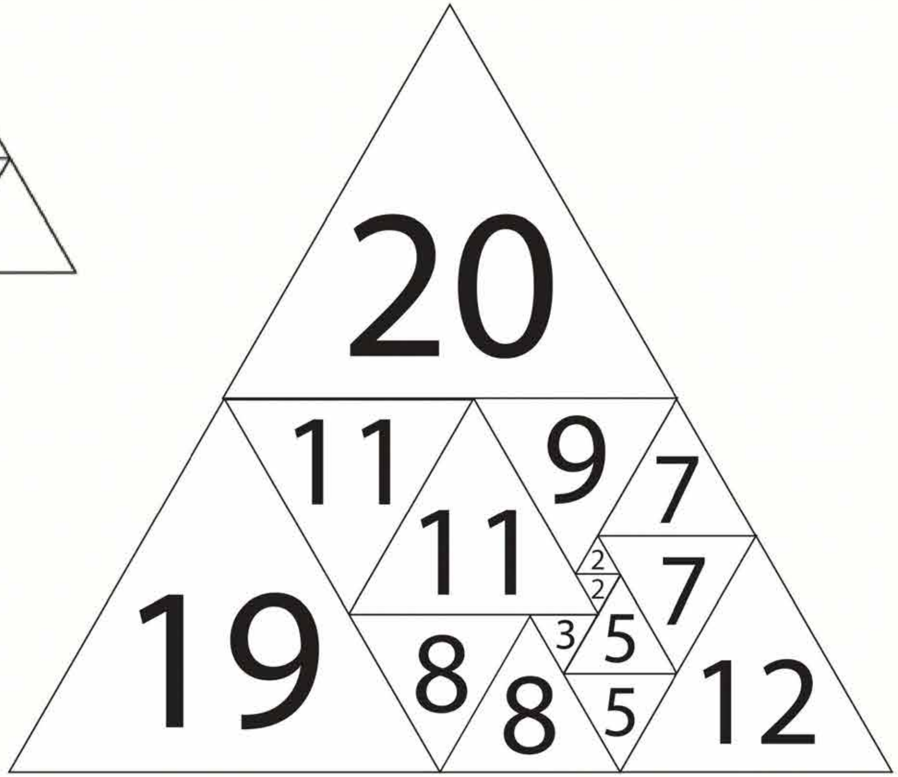

lowest order perfect dissected equilateral triangle, an isomer of the first

Golden mean Spirals

Golden Mean Spiral – This spiral is derived via the golden rectangle, a unique rectangle which has the golden ratio.
This form is found everywhere in nature: the Nautilus Shell, the face of a Sunflower, fingerprints, our DNA, and the shape of the Milky Way

Golden spiral

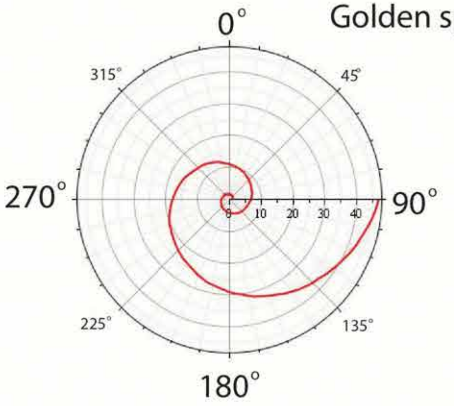

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ...

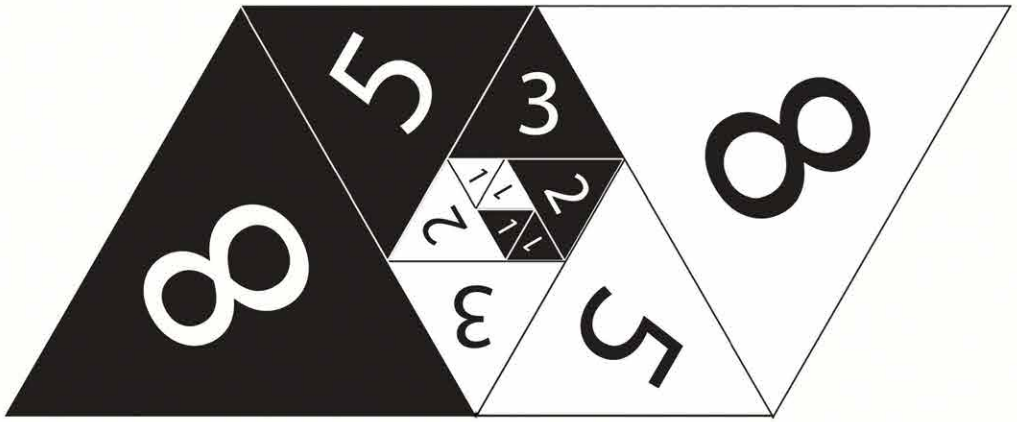

Golden spirals

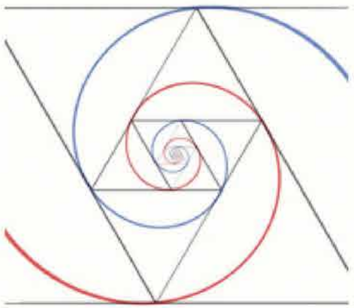

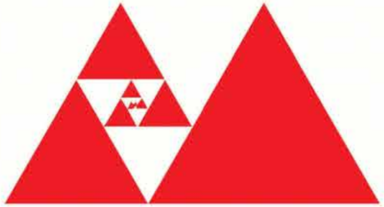

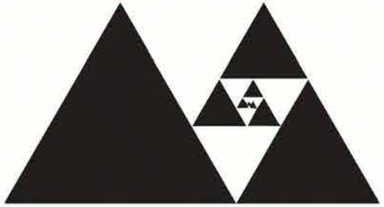

Continued fractions and the Fibonacci sequence

The convergence of the continued fractions

$$1+\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\frac{1}{1+}\cdots$$

are:

$$\frac{2}{1},\ \frac{3}{2},\ \frac{5}{3},\ \frac{8}{5},\ \frac{13}{8},\ \frac{21}{13},\ \frac{34}{21},\ \frac{55}{34},\ \frac{89}{55},\ \frac{144}{89},\ \frac{233}{144},\ \frac{377}{233},\ \frac{610}{377},$$

The Golden Ratio (Golden Mean, Golden Section)
is defined mathematically as:

$$\varphi=\frac{1+\sqrt{5}}{2}=1.6180339887\ldots.$$

Copyright ABRAHAM [2008] - All rights reserved

Koch fractal Curve

Niels Fabian Helge von Koch (January 25, 1870 – March 11, 1924) was a Swedish mathematician
who gave his name to one of the earliest fractal curves ever known

He described the Koch curve, or Koch snowflakes as it popularly known, in a 1904 paper entitled
"On a continuous curve without tangents constructible from elementary geometry"

Von Koch wrote several papers on number theory . One of his results was a 1901 theorem proving that the Riemann hypothesis is equivalent to a strengthened form of the prime number theorem.

Three Koch curves form the snowflake.

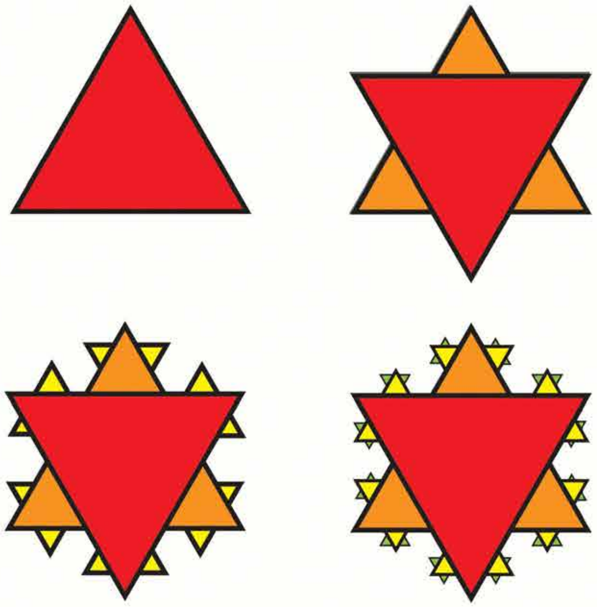

The Koch snowflake (or Koch star) is a mathematical curve and one of the earliest fractal curves to have been described.

Actually Koch described what is now known as the Koch curve, which is the same as the now popular snowflake, except it starts with a line segment instead of an equilateral triangle.

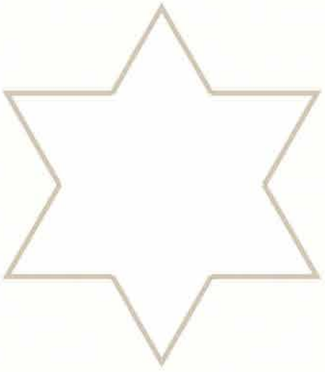

$$a = \frac{1}{2} + \frac{i}{\sqrt{12}}$$

The Koch curve is a special case of the Cesaro curve where:

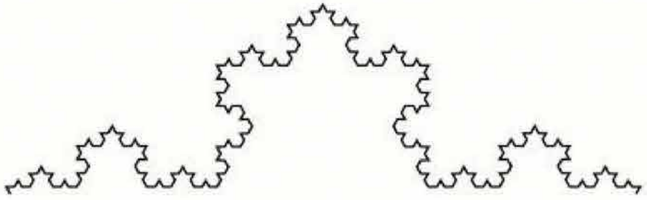

which is in turn a special case of the de Rham curve.

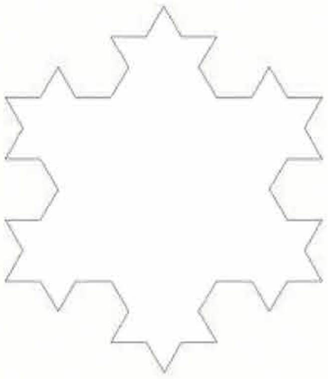

Copyright ABRAHAM [2008] - All rights reserved

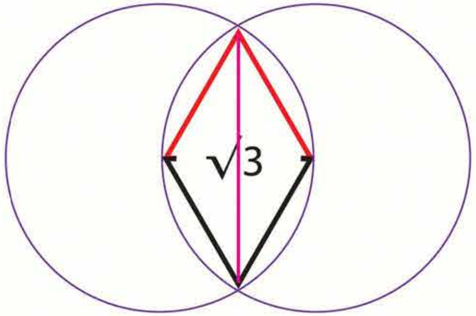

vesica piscis

$$\frac{1351}{780} > \sqrt{3} > \frac{265}{153}.$$

Koch snowflake

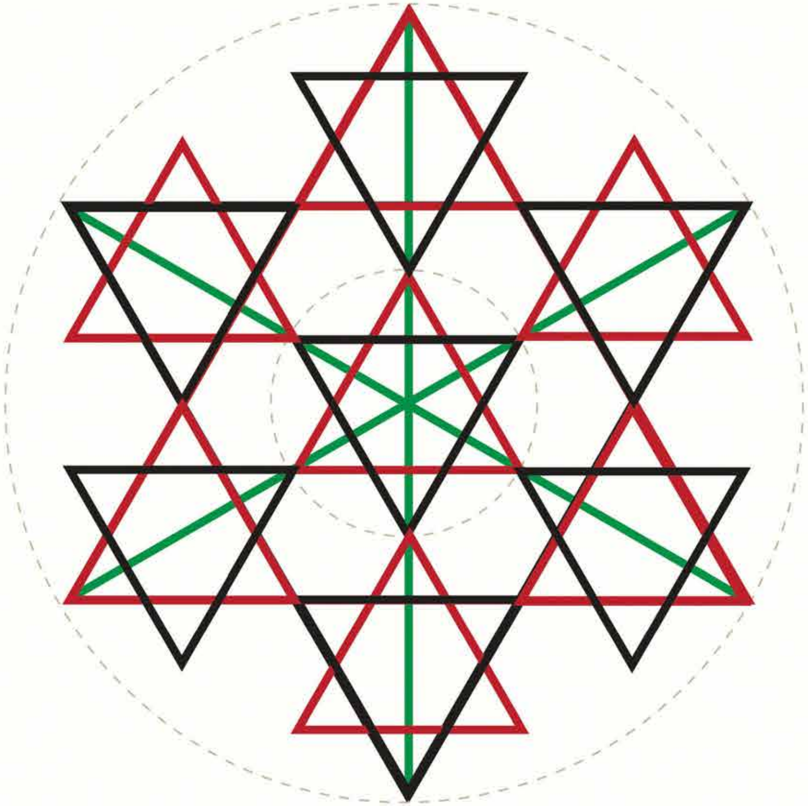

space-time

star tetrahedron

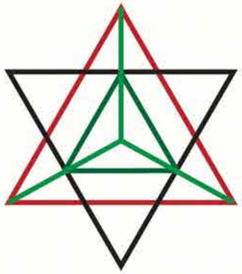

Flower of Life

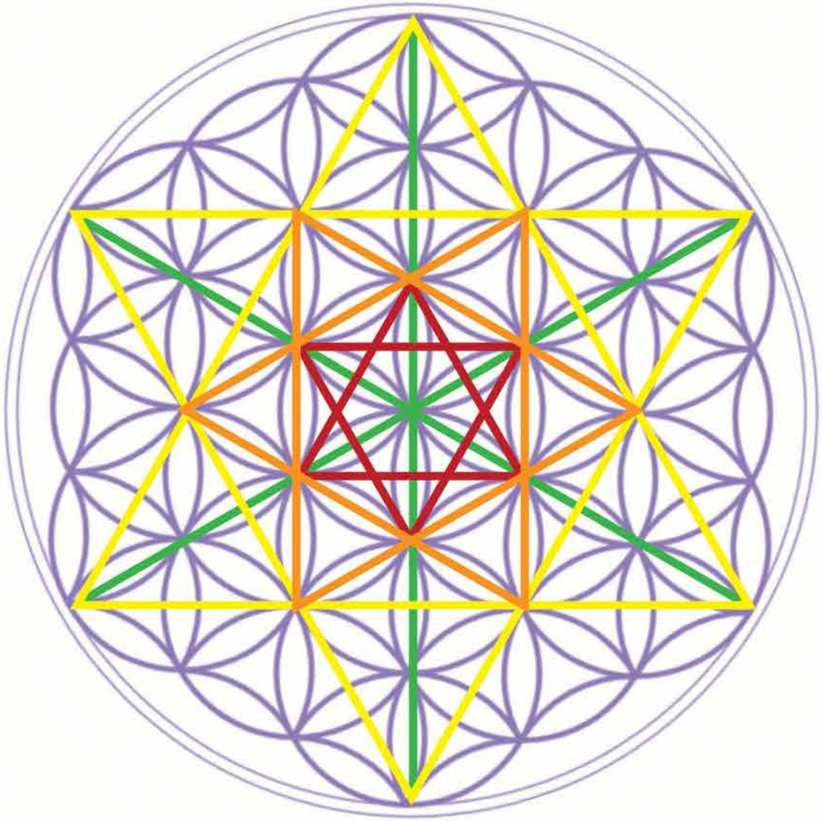

The Flower of Life is a name for a geometrical figure composed of multiple evenly-spaced, overlapping circles.

Unit circles - SINE WAVES - Photons

Classic model of a photon

Maxwell's Equations

$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$

$\nabla \cdot \vec{D} = \rho$

$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}$

$\nabla \cdot \vec{B} = 0$

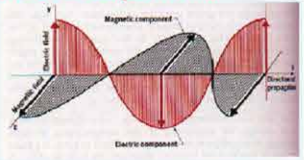
Magnetic component
Electric field
Magnetic field
Direction of propagation
Electric component

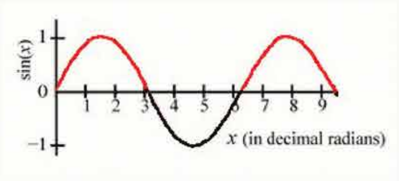

electric fields

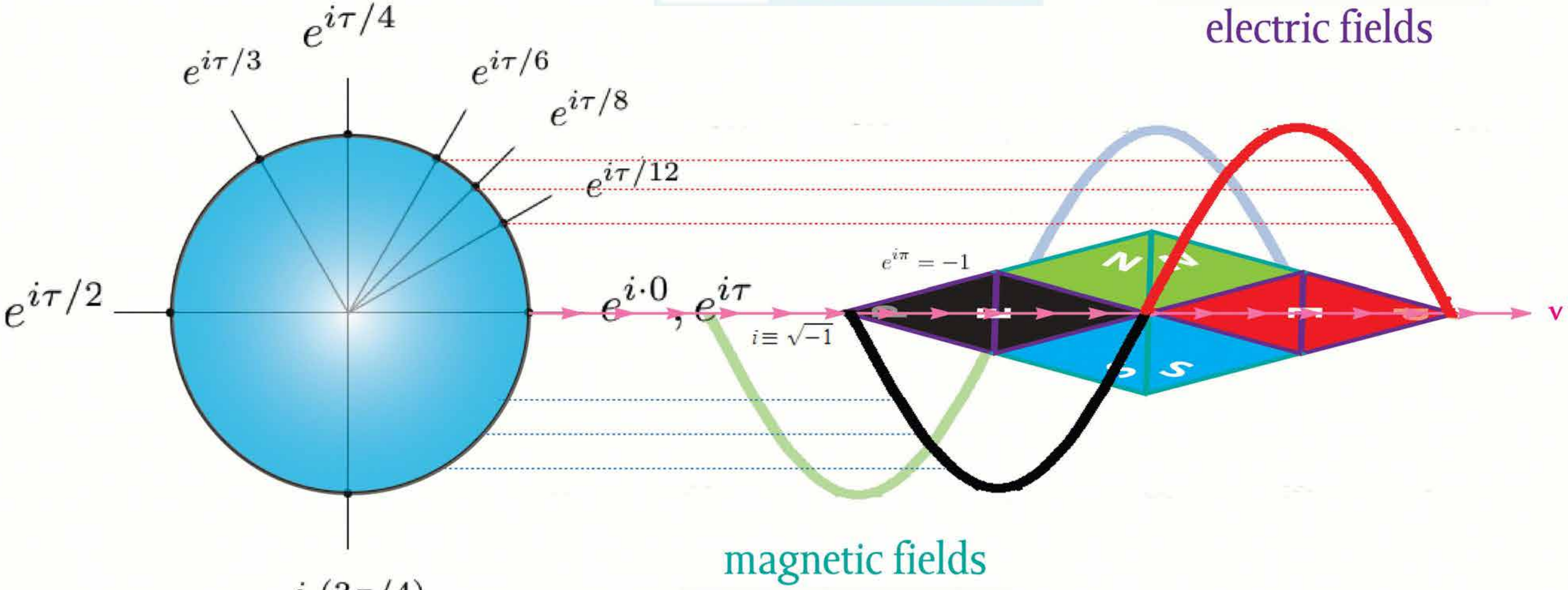
$e^{i\tau/4}$
$e^{i\tau/3}$
$e^{i\tau/6}$
$e^{i\tau/8}$
$e^{i\tau/12}$
$e^{i\tau/2}$
$e^{i\cdot 0}, e^{i\tau}$
$e^{i\pi} = -1$
$i \equiv \sqrt{-1}$
N
N
S
S
v
$e^{i\cdot(3\tau/4)}$

magnetic fields

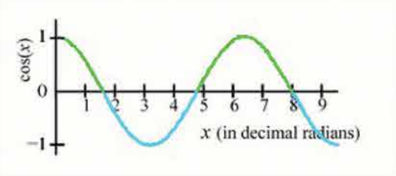

Tetryonic model of a photon

Magnetic waveforms are 90 degrees out of phase with Electric waveforms

Copyright ABRAHAM [2008] - All rights reserved

Boson distributions in monochromatic EM waves

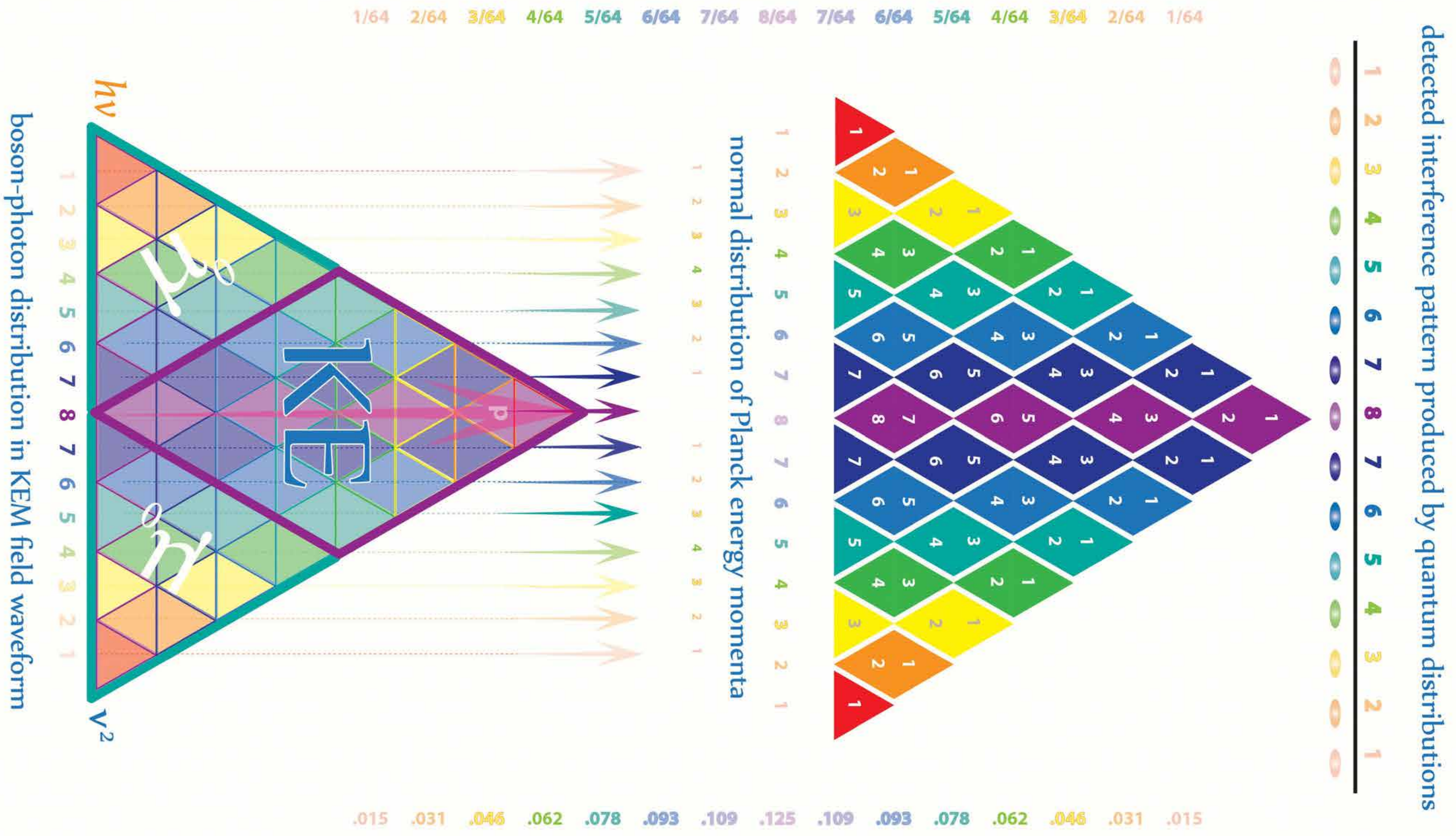
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64 7/64 6/64 5/64 4/64 3/64 2/64 1/64
hν
v²
μ0
ε0
KE
p
boson-photon distribution in KEM field waveform
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
normal distribution of Planck energy momenta
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
detected interference pattern produced by quantum distributions
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
.015 .031 .046 .062 .078 .093 .109 .125 .109 .093 .078 .062 .046 .031 .015

Probability distributions of monochromatic EM waves

Tetryonics 98.02 - Boson distributions in monochromatic EM waves

Wave interference patterns

The double-slit experiment, sometimes called Young's experiment, is a demonstration that matter and energy can display characteristics of both waves and particles, and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and Establishes the quantum interference principle known as wave–particle duality.

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

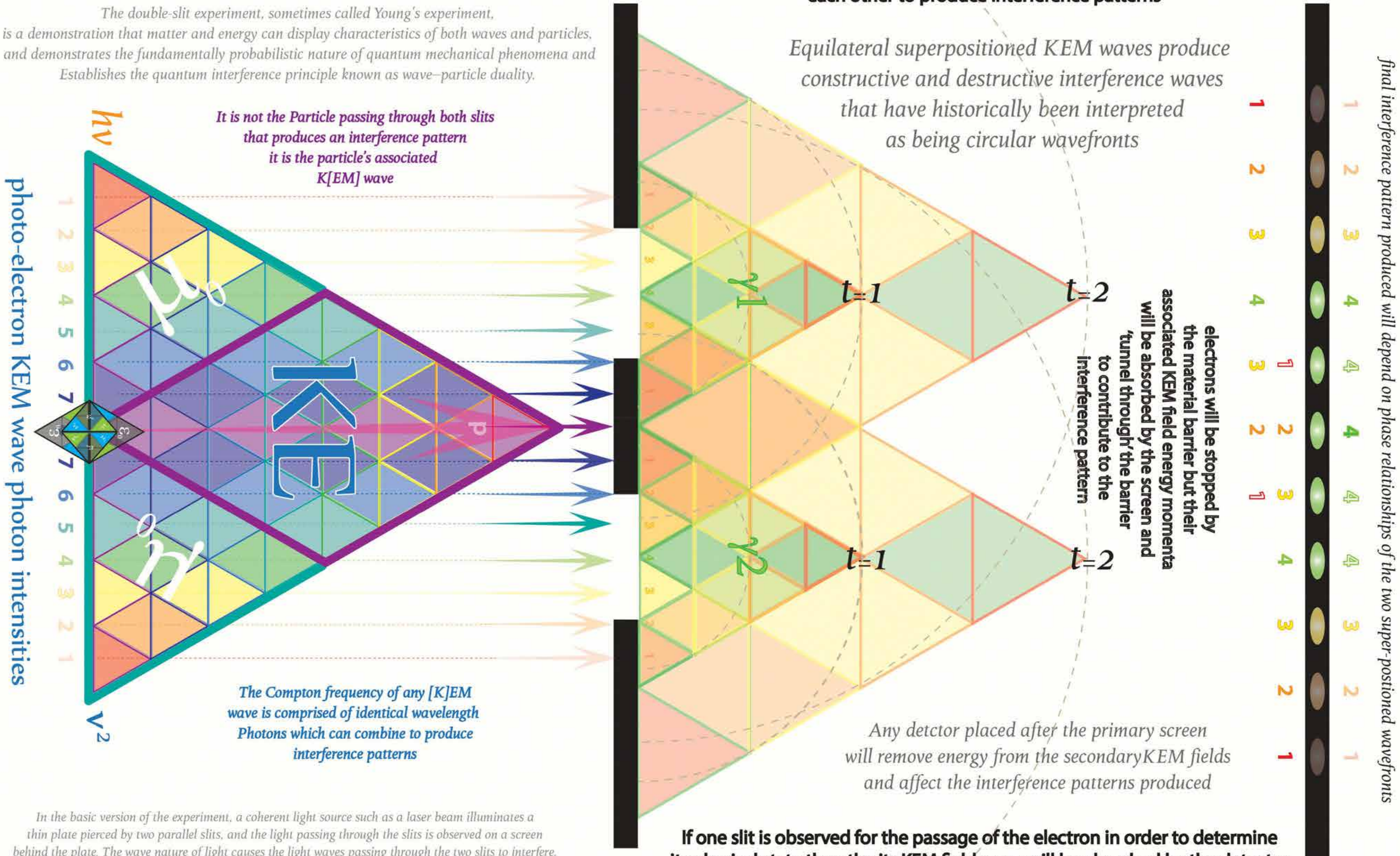

photo-electron KEM wave photon intensities

It is not the Particle passing through both slits that produces an interference pattern it is the particle's associated K[EM] wave

Equilateral superpositioned KEM waves produce constructive and destructive interference waves that have historically been interpreted as being circular wavefronts

electrons will be stopped by the material barrier but their associated KEM field energy momenta will be absorbed by the screen and 'tunnel through' the barrier to contribute to the interference pattern

final interference pattern produced will depend on phase relationships of the two super-postioned wavefronts

The Compton frequency of any [K]EM wave is comprised of identical wavelength Photons which can combine to produce interference patterns

Any detctor placed after the primary screen will remove energy from the secondaryKEM fields and affect the interference patterns produced

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

Quantum computing via EM wave super-positioning

BY superpositioning two beams
of EM radiation the resultant 'colours'
will perform quantum level computations
that can be read via the resultant
interference patterns produced

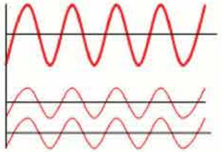

In phase

constructive interference

additive in-phase EM waves

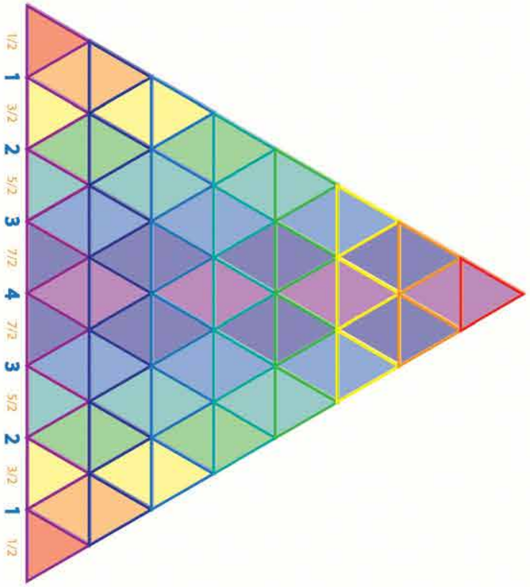
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Additive

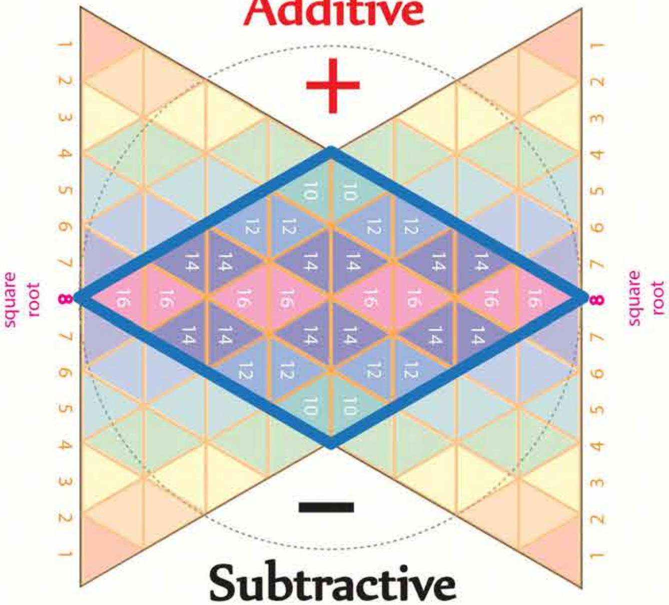
+
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
square root
10
12
14
16
−

Subtractive

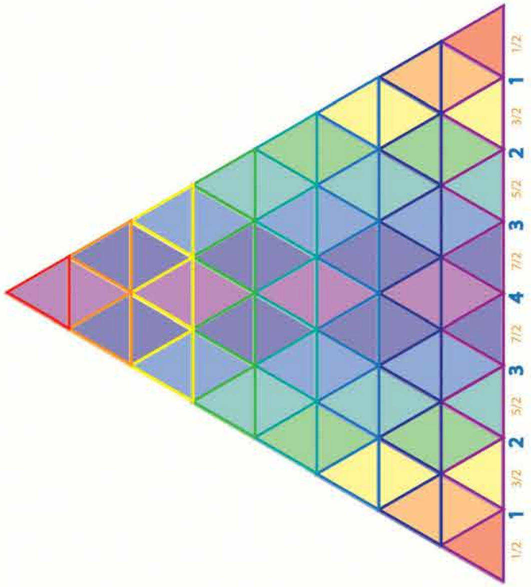
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Out of phase

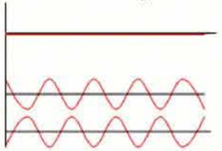

destructive interference

subtractive out-of-phase EM waves

Various basic operations, such as
ADDITION, SUBTRACTION and
SQUARE ROOTS etc
are all easily computed using
EM wave super-positioning

The lines of Force

By utilising the statiscal distribution of equilateral Planck energy momenta quanta in EM waves Tetryonic theory provides a practical geometric solution to quantum computing problems

q

bosons

P vs. NP

The set of all decision problems for which an algorithm exists which can be carried out by a deterministic Turing machine in polynomial *time*

the set of all decision problems for which an algorithm exists which can be carried out by a non-deterministic *Turing machine in* polynomial *time*

E

photons

An algorithm of time complexity ***O***(*n*) is one which increases in time linearly as the "size of the problem" (whatever n stands for) increases.

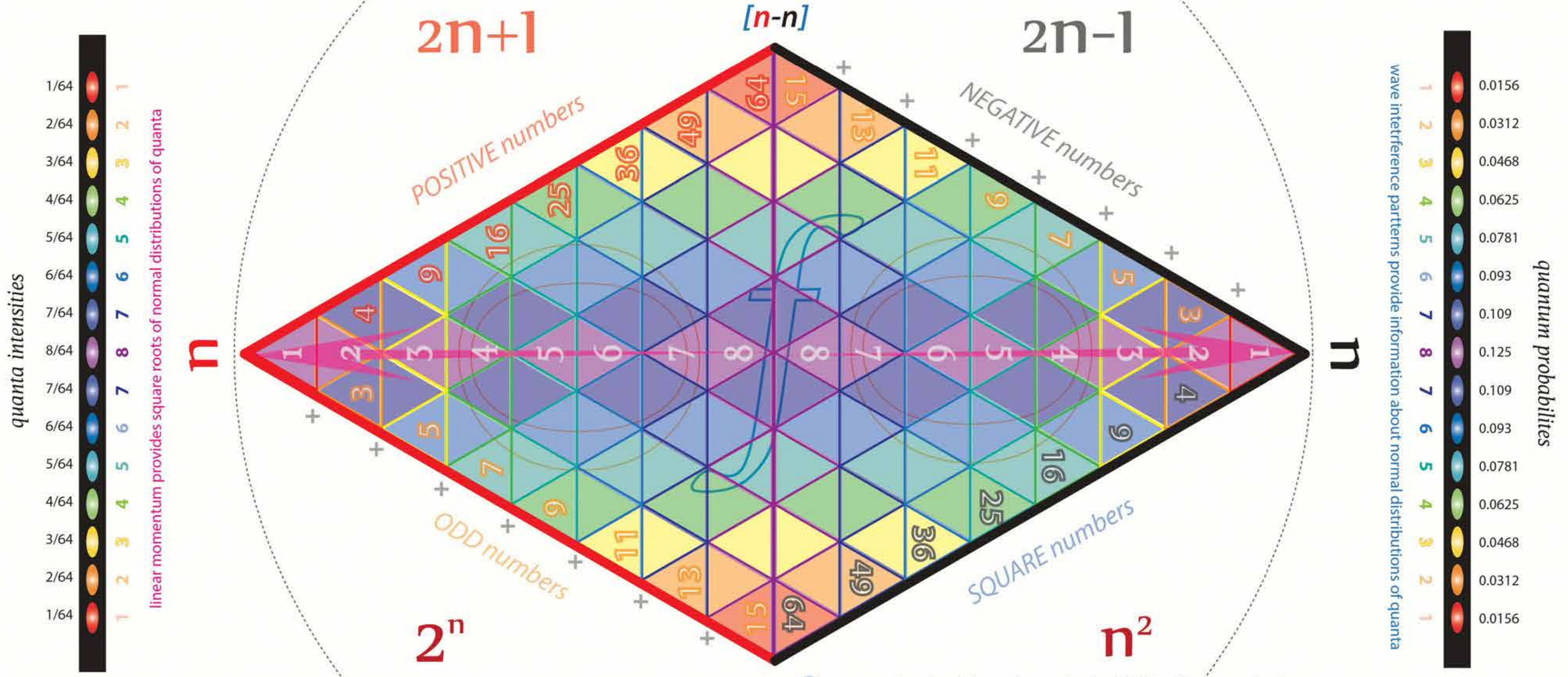
quanta intensities
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
2n+1
[n-n]
2n-1
POSITIVE numbers
NEGATIVE numbers
ODD numbers
SQUARE numbers
n
n
1 2 3 4 5 6 7 8 8 7 6 5 4 3 2 1
4 9 16 25 36 49 64
3 5 7 9 11 13 15
2^n
n^2
quantum probabilites
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156

linear momentum provides square roots of normal distributions of quanta

wave interference parterns provide information about normal distributions of quanta

An algorithm of complexity ***O***(*2ⁿ*) utlises exponential quanta; increasing n by 1 will double the quanta required

$2n^2$

An algorithm of complexity ***O***(*n*2) utilises quadratic quanta, meaning that if you double n it will use four times as many quanta.

n^o

Unlike Math treatise on P vs Np that require exponential polynomial time ***O***(*n*k) Tetryonic geometry of EM fields utilise exponential energies per second

2hv

1 second

hf

By relating p to the number of equilateral wave fields in any EM wave very large [pn²] data sets can be modelled and processed every second

Copyright ABRAHAM [2008] - All rights reserved

Quantum Cryptography

If the numerical sequences were applied to amplitude modulation
their non-repeating numerical sequences would appear
to be purely random noise

**Using Tetryonic geometry
advanced non-repeating cyphers
of any complexity
can be easily developed**

The level of encryption can
be easily increased (without limit)
by increasing the dimensions
of the cypher geometry

eg: number of quanta
2D vs 3D geometries

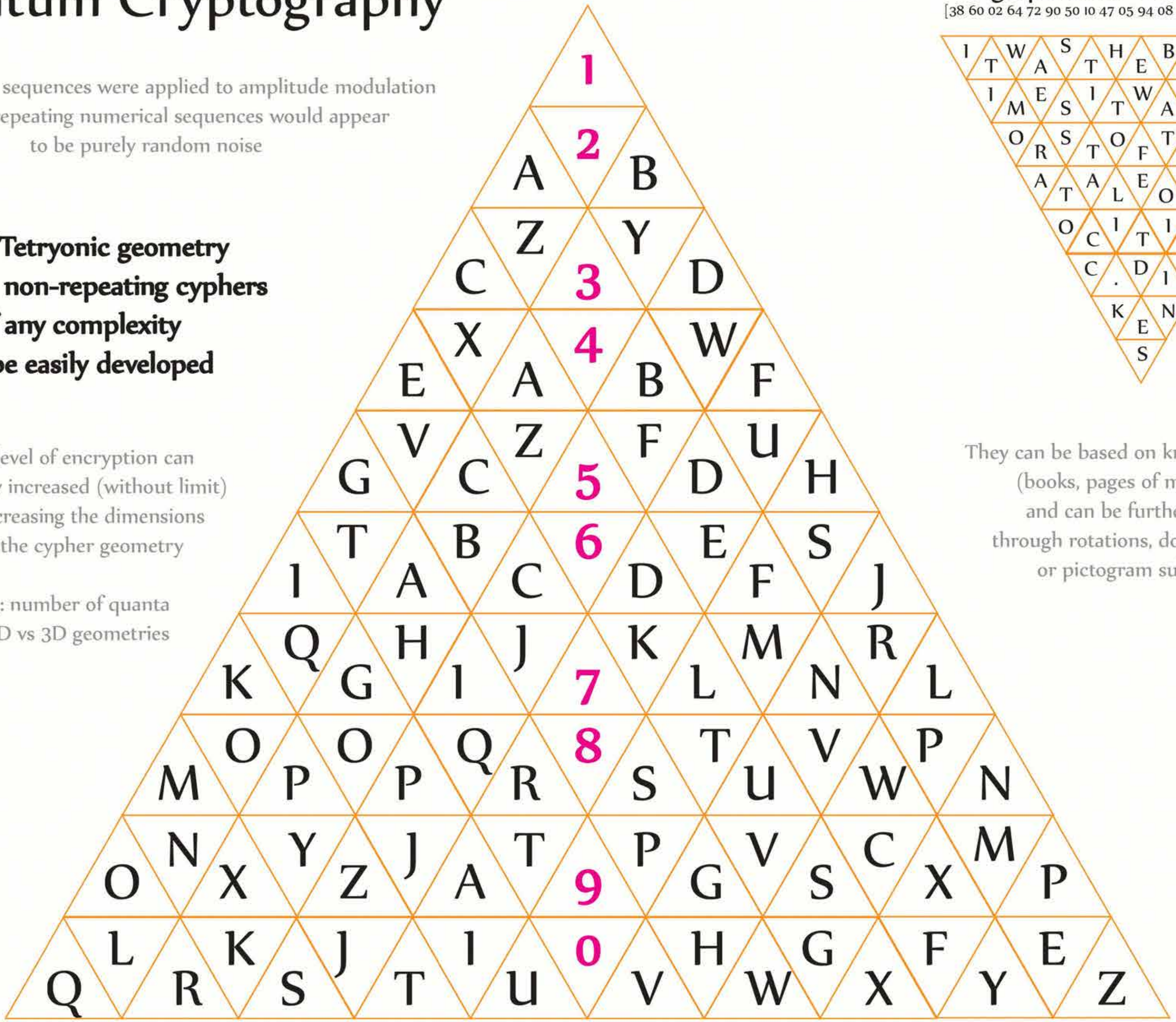
1
A 2 B
C Z 3 Y D
E X A 4 B W F
G V C Z 5 F D U H
I T A B C 6 D E F S J
K Q G H I J 7 K L M N R L
M O P O P Q R 8 S T U V W P N
O N X Y Z J A T 9 P G V S C X M P
Q L R K S J T I U 0 V H W G X F Y E Z

eg: quantum encryption 1024
[38 60 02 64 72 90 50 10 47 05 94 08 48 52 27 26 51 66 01 91 03 20]

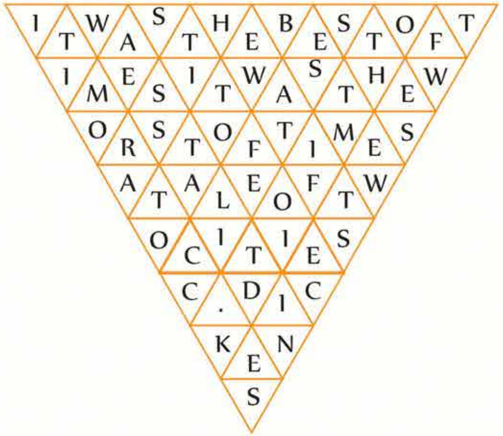
I T W A S T H E B E S T O F T
I M E S I T W A S T H E W
O R S T O F T I M E S
A T A L E O F T W
O C I T I E S
C . D I C
K E N
S

They can be based on known letter sources
(books, pages of magazines etc)
and can be further encrypted
through rotations, double encrypting
or pictogram substitutions

Copyright ABRAHAM [2008] - All rights reserved

Quantum Computing

The Proton/Neutron geometries of atomic nuclei can be built at the quantum scale
to create an atomic nuclei that can operate as a Opto-memory-transistive computing element,
many elements can then be combined in lattices to create super computers no larger than bacterium

Spin UP

Energy can be gated through individual nuclei
using the centre Baryon as the base transistor element,
in turn effecting the energies of bound photo-electrons

Spin DOWN

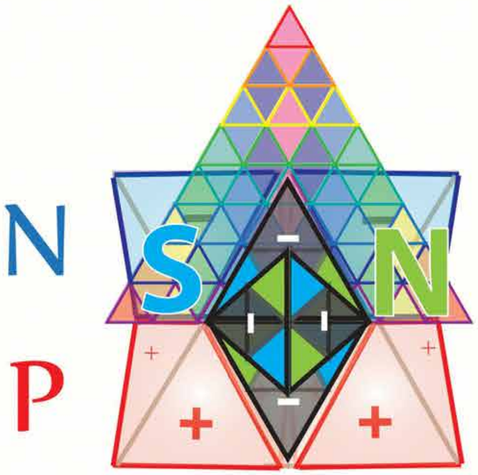

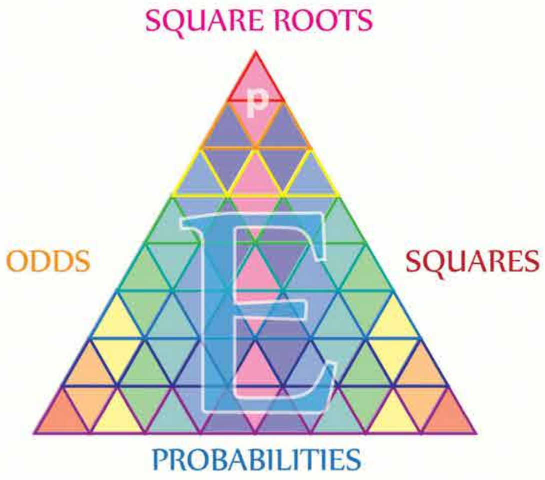

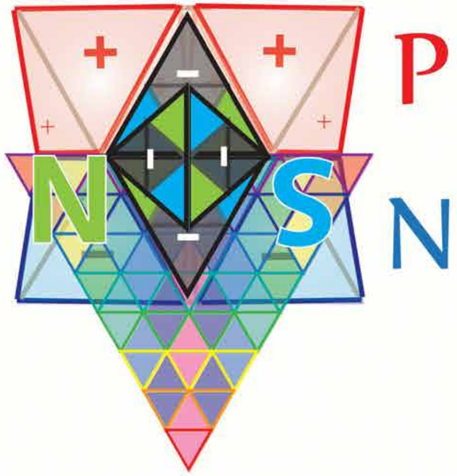

Q-bits

1

photo-electronic transitions can be used to directly recieve
or emit memory states through the absorption and emission
of spectral photons of specific energy momenta

0

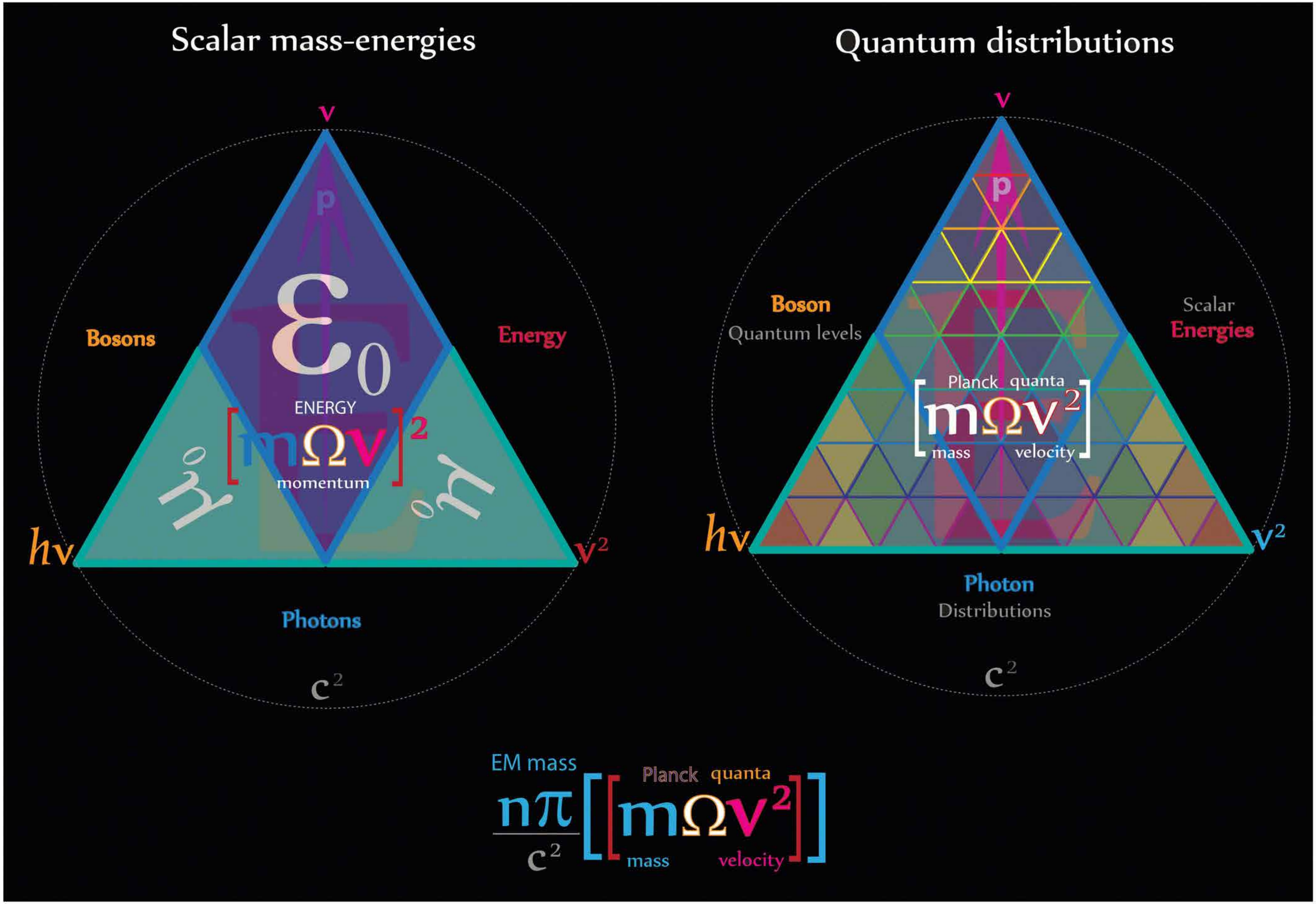
Scalar mass-energies
v
p
ε_0
ENERGY
$[m\Omega v]^2$
momentum
Bosons
Energy
hν
v^2
Photons
c^2
Quantum distributions
v
p
Boson
Quantum levels
Scalar
Energies
Planck quanta
$[m\Omega v^2]$
mass velocity
hν
v^2
Photon
Distributions
c^2
EM mass
$\frac{n\pi}{c^2}[[m\Omega v^2]]$
Planck quanta
mass velocity

2D mass

EM mass

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

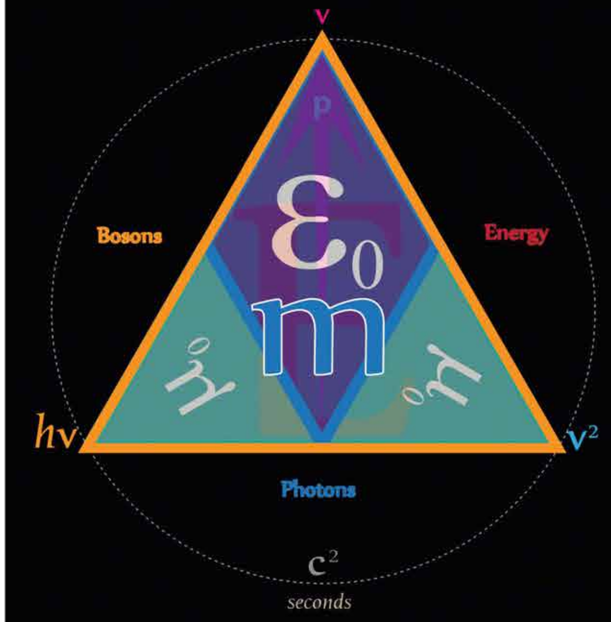
v
p
ε0
m
hν
v²
Bosons
Energy
Photons
c²
seconds

Measure of planar energies
in a nπ radiant wave
geometry
[per second]

mass quantum
7.376238634 e-51 kg

ENERGY

ENERGY

$$n\pi\left[\left[\underset{\text{momenta}}{m\Omega v^2}\right]\right]$$

Equilateral energy momenta
has a mass equivalence
per unit of Time

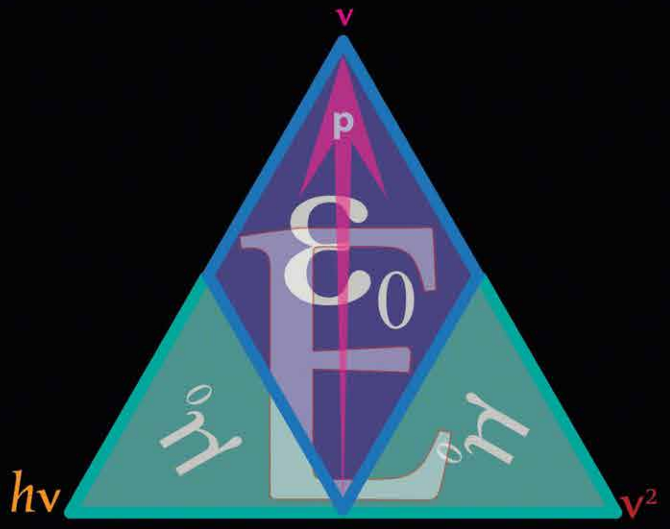

c² is the radial geometry
created by radiant energy
in 1 second

Planck Constant
6.629432672 e-34 J.s

3D Matter

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Measure of mass-energy
in a 4nπ standing wave
topology
[per second squared]

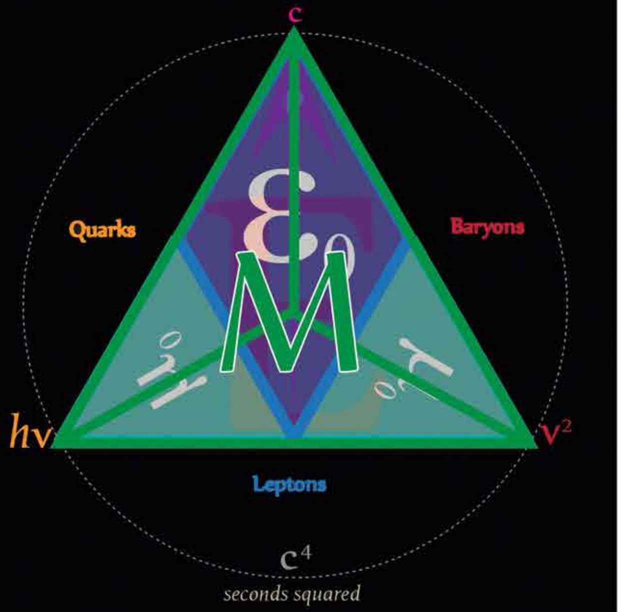
c
ε0
M
Quarks
Baryons
Leptons
hν
v²
c⁴
seconds squared

Matter quantum
2.9504955454 e-50 kg

Planck's constant - the quantum of Action

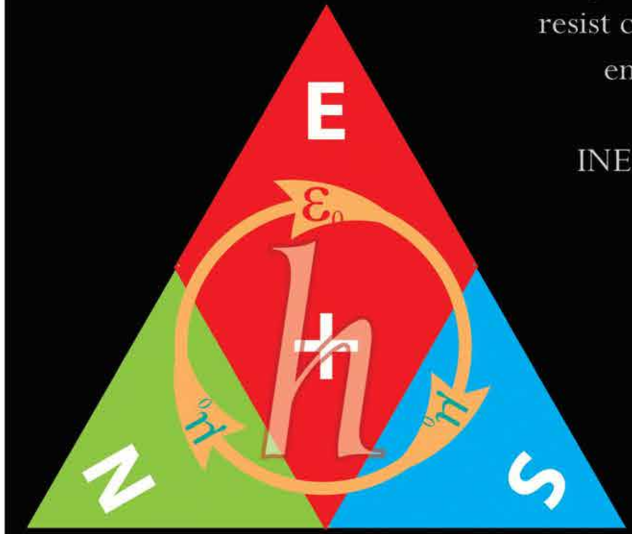
E
ε_0
h
μ_0
μ_0
N
S

all ideal quantum inductive loops
resist changes to their
energy levels

INERTIAL mass

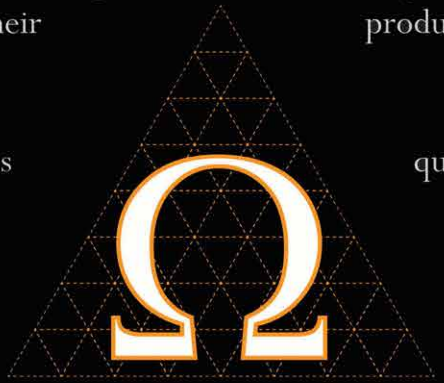

all equilateral energy momenta
produce square root vector
Forces

quantum of WORK

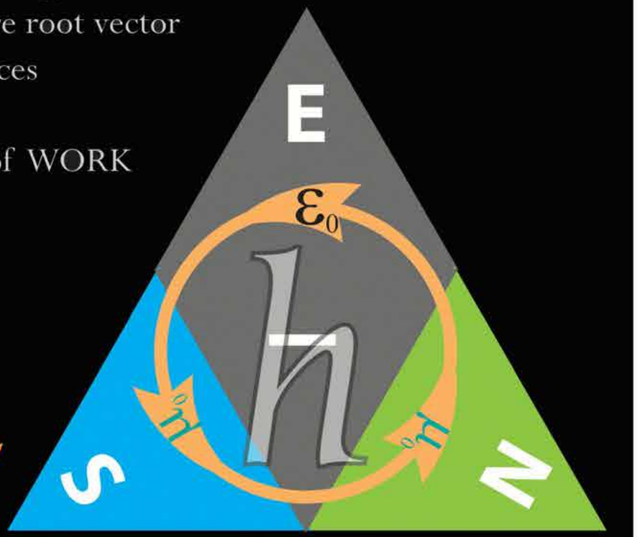
E
ε_0
h
μ_0
μ_0
S
N

Quantised angular momentum is equilateral geometry

The two sides of Planck's Constant

The square root of all Energy is linear momentum

Equilateral
energy quanta
per second

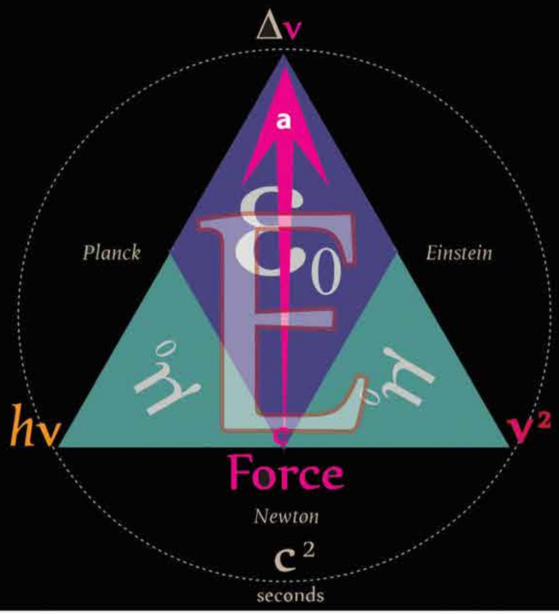
Δv
a
ε_0
E
Planck
Einstein
μ_0
μ_0
hv
v²
Force
Newton
c²
seconds

Quantised
angular momenta
per second

mass

time

Copyright ABRAHAM [2008] - All rights reserved

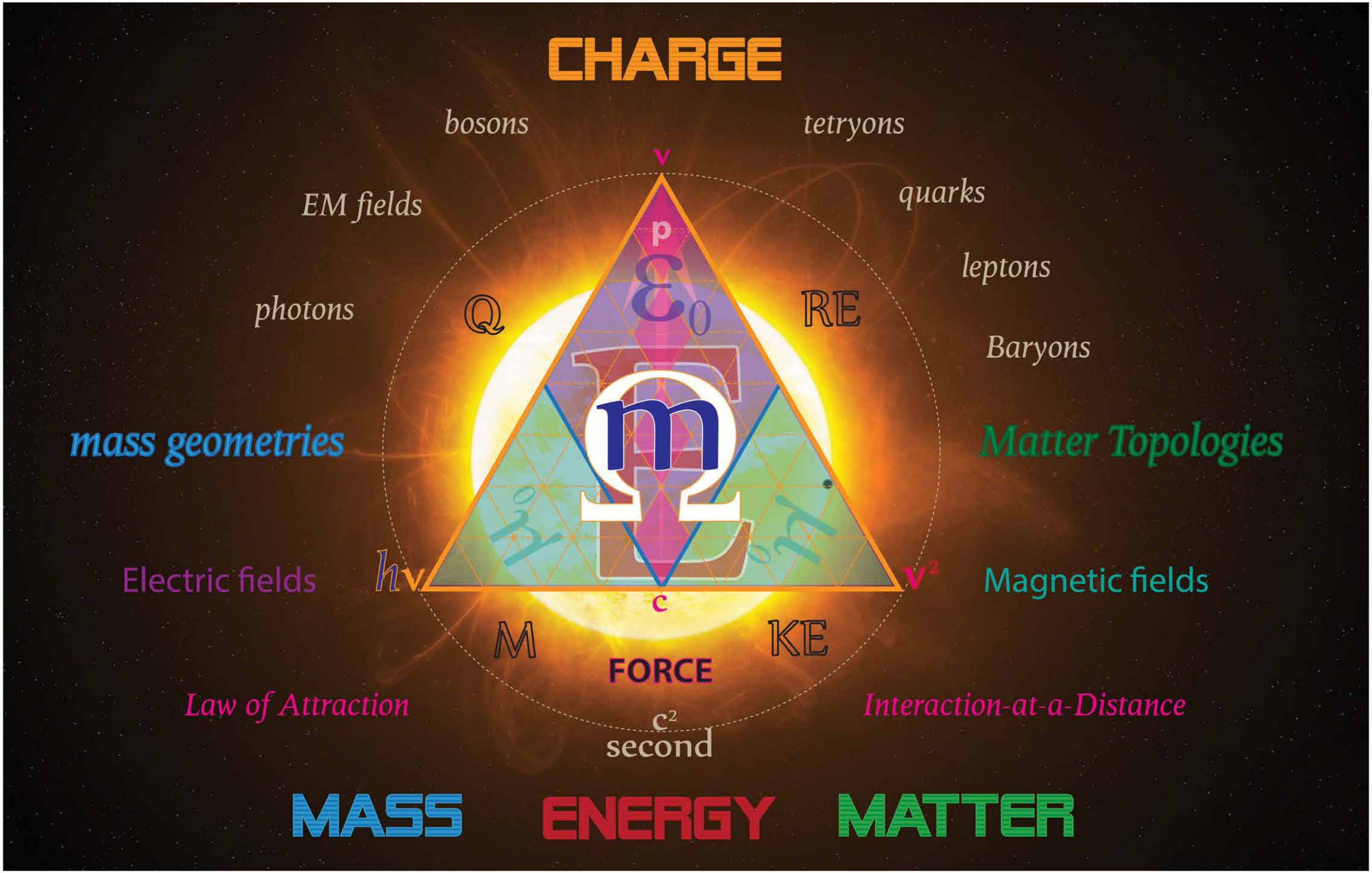

CHARGE

bosons

tetryons

EM fields

quarks

leptons

photons

Baryons

mass geometries

Matter Topologies

Electric fields

Magnetic fields

Law of Attraction

Interaction-at-a-Distance

FORCE

c^2

second

MASS ENERGY MATTER

v
p
ε_0
Q
RE
E
m
Ω
μ_0
hv
v^2
c
M
KE

Tetryonics 99.04 - Charged Planck mass-ENERGY-Matter

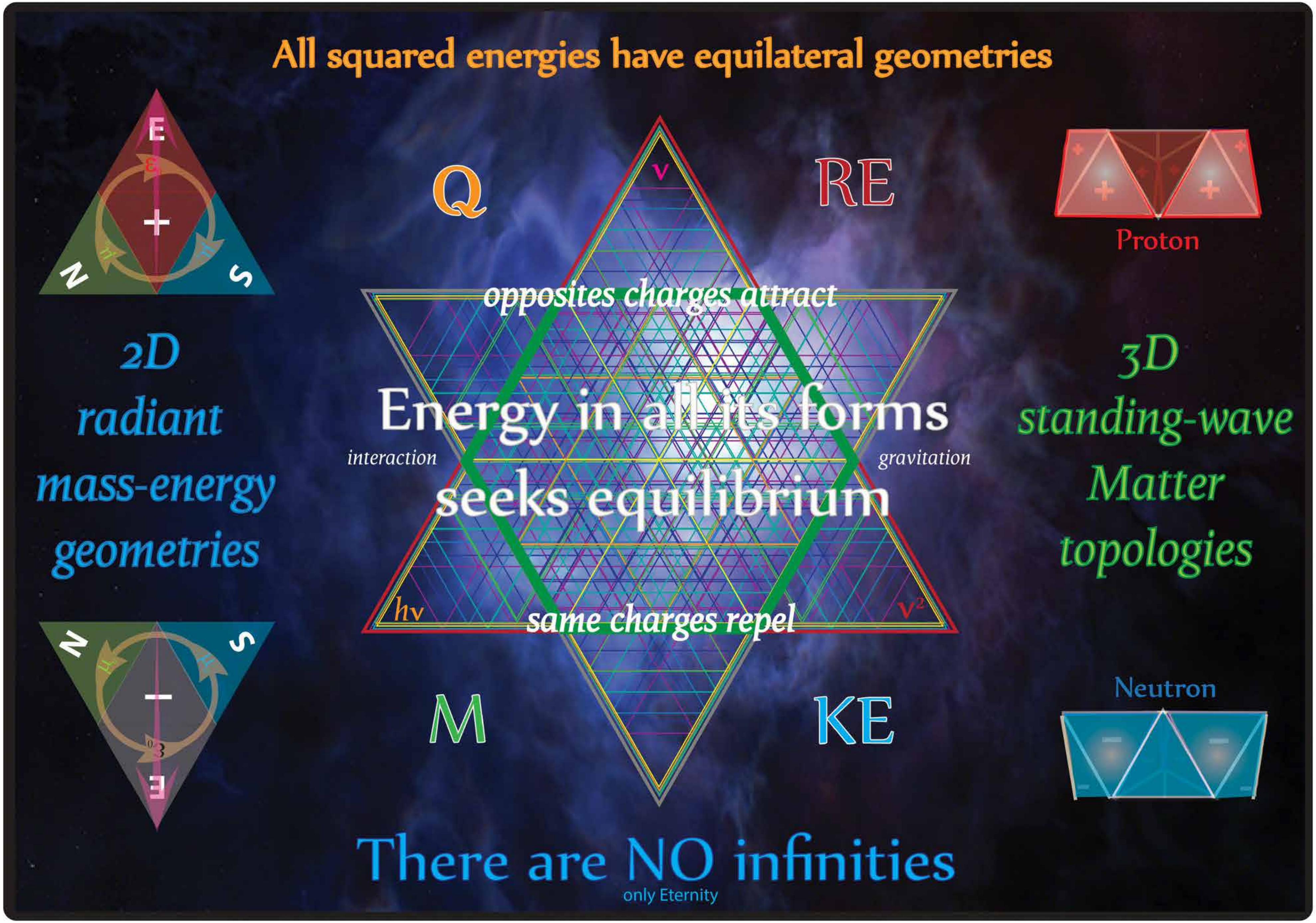

All squared energies have equilateral geometries

2D
radiant
mass-energy
geometries

opposites charges attract

interaction

Energy in all its forms
seeks equilibrium

gravitation

same charges repel

3D
standing-wave
Matter
topologies

There are NO infinities
only Eternity

Copyright ABRAHAM [2008] - All rights reserved

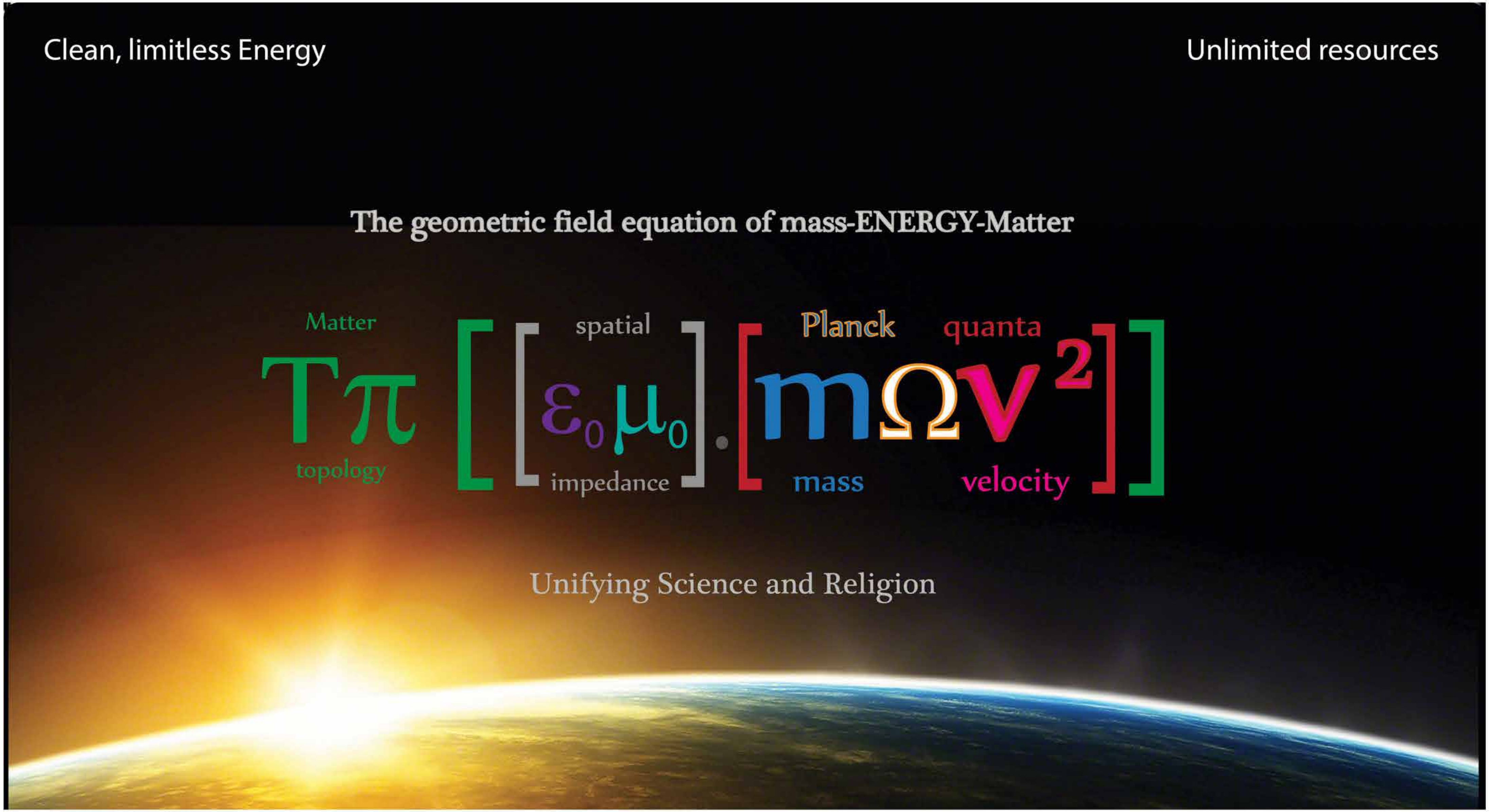

Clean, limitless Energy

Unlimited resources

The geometric field equation of mass-ENERGY-Matter

$$\underset{\text{topology}}{\overset{\text{Matter}}{T\pi}}\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{\varepsilon_0\mu_0}}\right]\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

Unifying Science and Religion

Tetryonics 99.06 - Tetryonic Unified field equation

Tetryonic Theory

All EM mass, energy momenta & Matter can be measured and geometrised with respect to equilateral Quantised Angular Momenta

charged Planck quanta

Bosons

$$\frac{ODD\,\pi}{c^2}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Total Relativistic energies

ENERGY

$$n\pi\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Quantum Mechanics

Quantum Cosmology

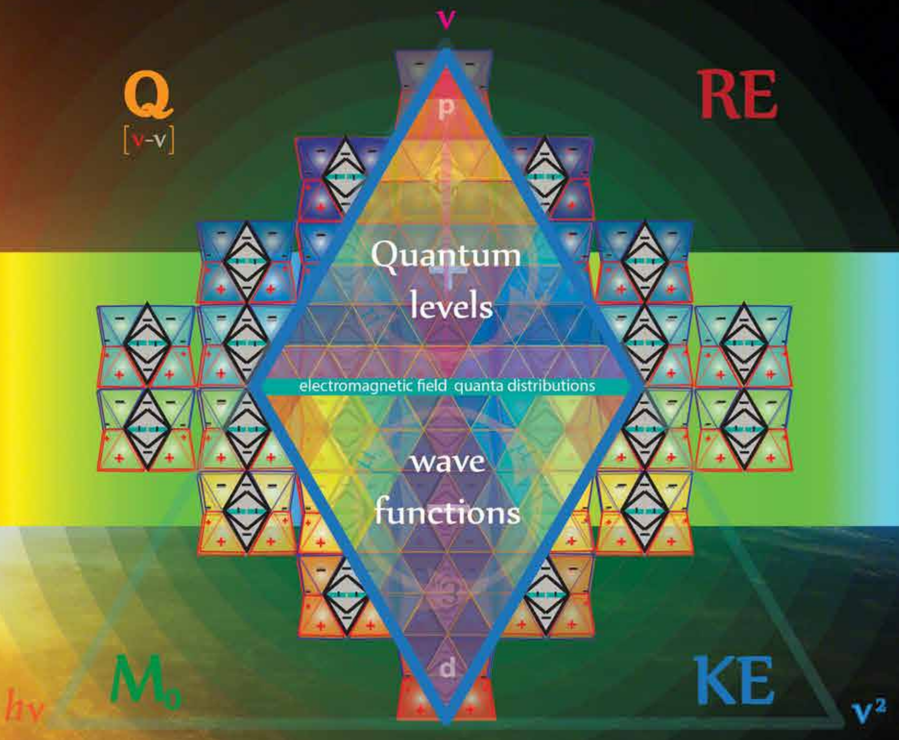

standing-wave mass-Matter

radiant EM mass-energies

Quantum Chemistry

Electrodynamics

Matter

$$\frac{T\pi}{c^4}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

EM waves

$$\frac{EVEN\,\pi}{c^2}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\overset{\text{quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

3D standing-wave mass/Matter topologies

2D radiant mass-energy geometries

The application of equilateral QAM geometries
covers all of the Physical disciplines

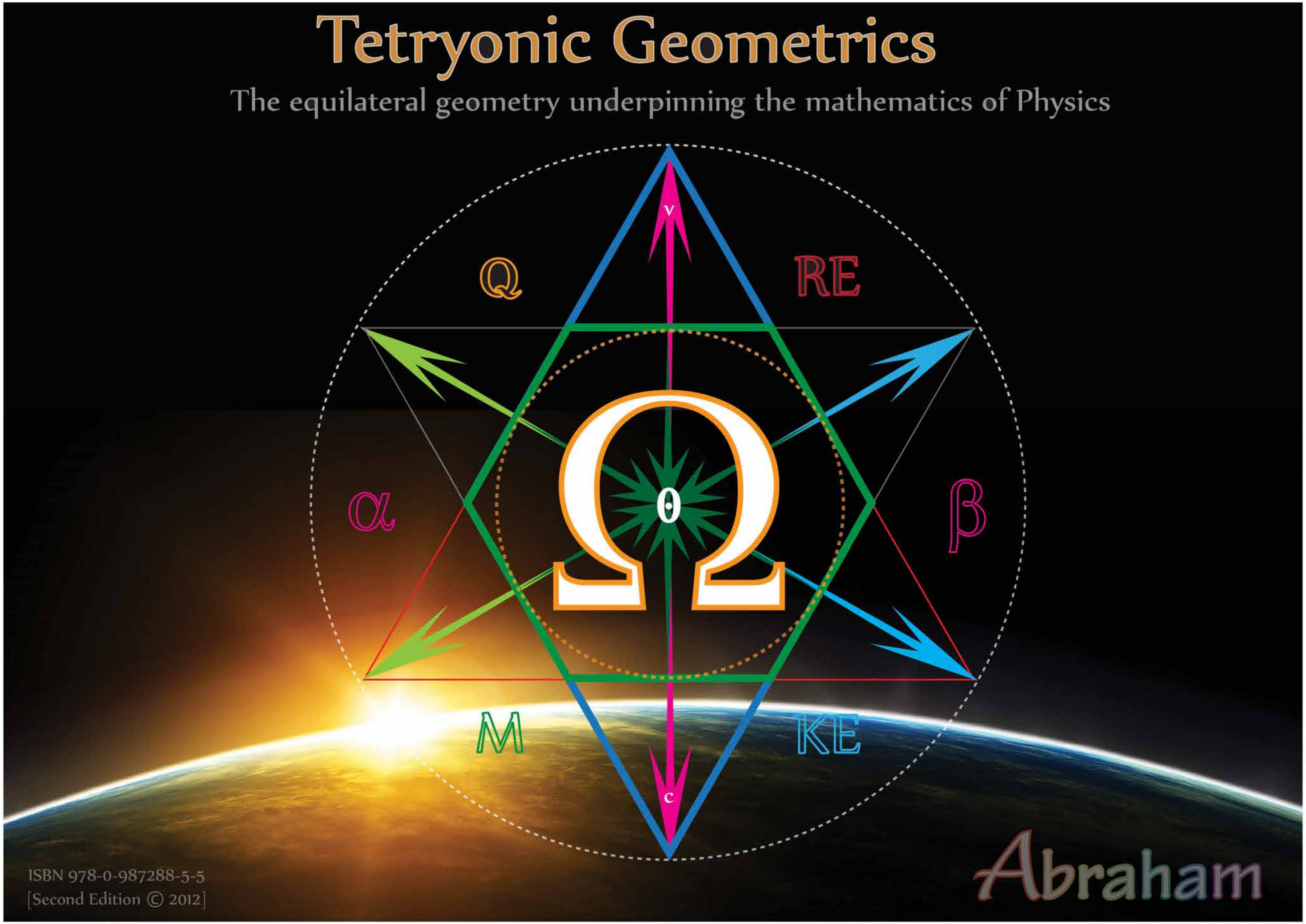

Tetryonic Geometrics

The equilateral geometry underpinning the mathematics of Physics

ISBN 978-0-987288-5-5
[Second Edition © 2012]

Abraham

Copyright ABRAHAM [2008] - All rights reserved

TETRYONICS

The charged topology of periodic & compound Matter

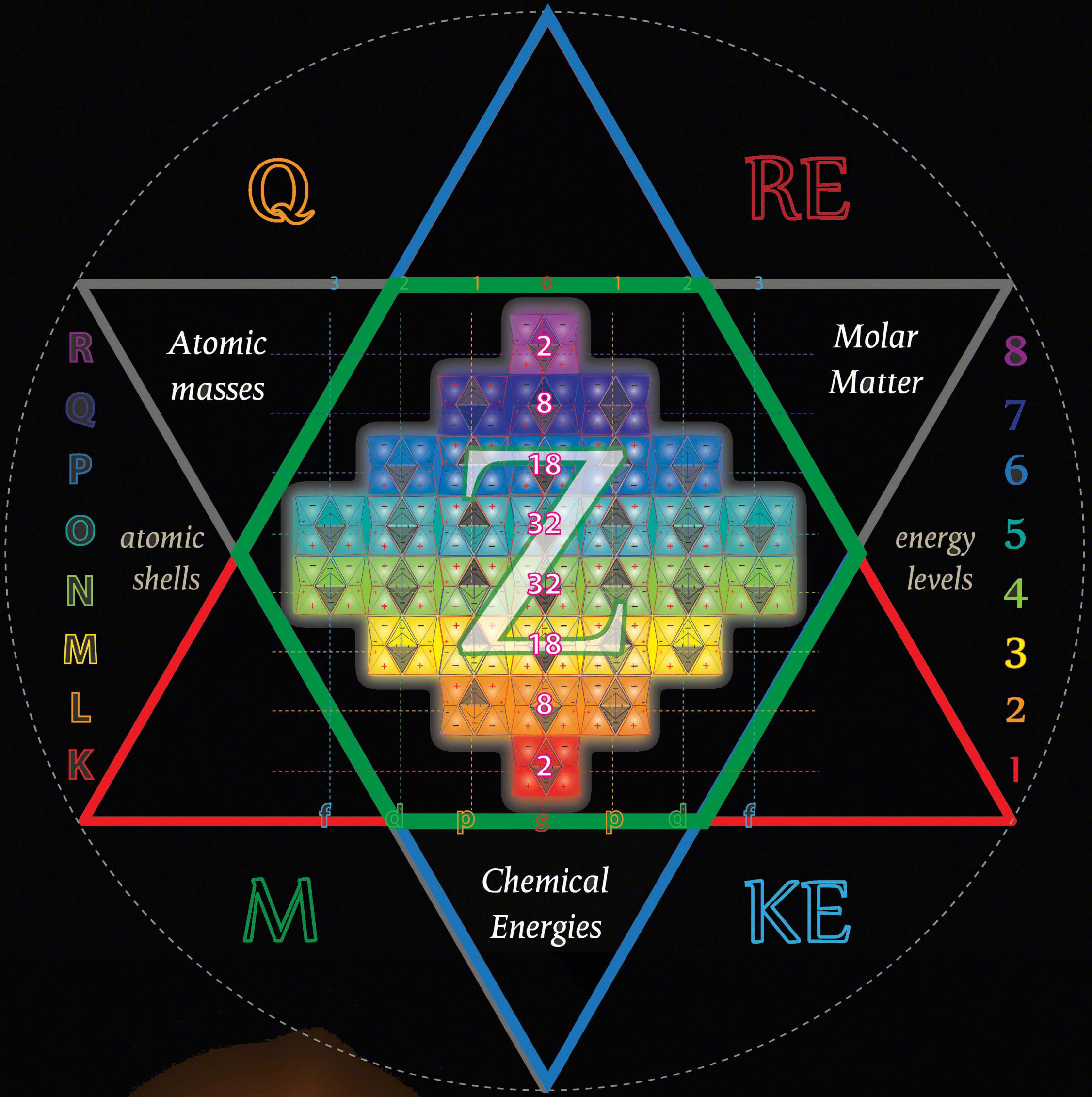
Q
RE
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic masses
atomic shells
Molar Matter
energy levels
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
M
KE
Chemical Energies

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1
[Second Edition © 2012]

Residual Electro-Magnetic Forces

allow Neutrons and Protons to attract via the opposite Electric charge points
created by their constituent Quarks in order to create Elementary Nuclei

Protons **+** *Neutrons*

u+d+u d+u+d

All electrons, quarks and Baryons are made up of 4n+ (Tetryonic) standing-wave EM fields.

As well as having nett Tetryonic charged topologies ranging between [+24] ~ [-12] they all posses distinct ELECTRIC FIELDS that are concentrated in 3 apex points as indicated in the illustrations

These points result from the orientation of Electric apexes and orthagonal Magnetic dipole field edges that make up each particle's externalised EM fields.

The Positive and Negative electric apex points, obey the Law of Interaction forcing separated nuceli to combine due to their individual nett Tetryonic charges and provide a means of orienting nuclei to each other to create larger particles [elements, allotropes and compounds]

External Magnetic (H) fields can interact with the integral magnetic (B) dipoles of Tetryonic particles forcing them to orientate in specific directions to facilite chemical bonding [nuclear forces]

Additionally, external Electric fields can interact with the integral electric fields attracting or repelling them depending on the polarity of the external electric field [Electrostatics]

External energies can be induced into these integral EM fields via inductive coupling or the absorption of spectral photons in turn leading to an increase in the strengthes of the integral EM apexes in turn increasing the Strong Nuclear Force.

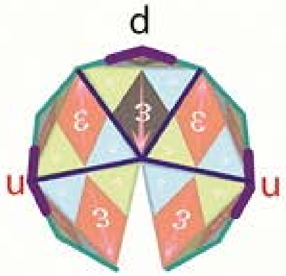

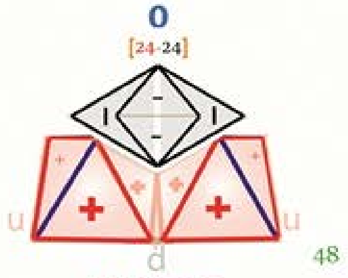

Hydrogen

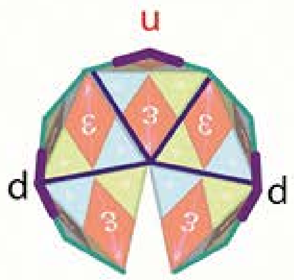

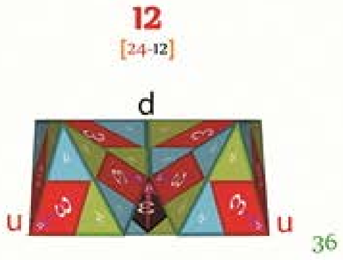

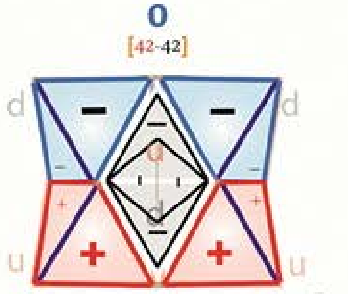

Deuterium

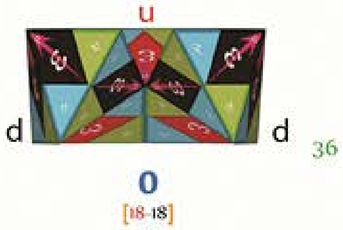

The Strong Nuclear force binds Matter together

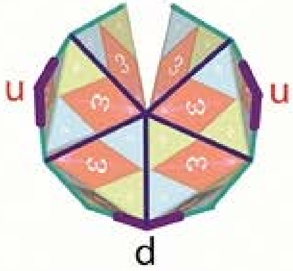

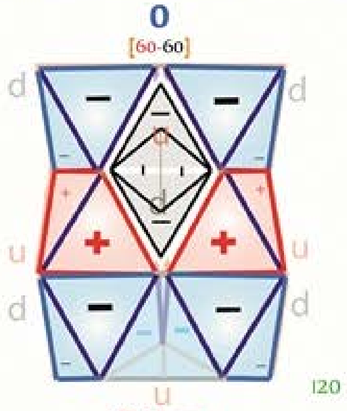

Tritium

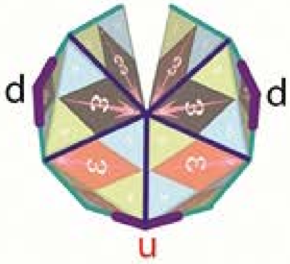

The orientation of the component Electric fields within 3D Matter creates macroscopic force apexes via externalised 'E-points'

The orientation of the component Magnetic fields within 3D Matter creates macroscopic force apexes via externalised 'M-dipoles'

Copyright ABRAHAM [2008] - All rights reserved

Nucleonic residual EM force

Neutrons

Protons

The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei

The residual e-field apexes and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei

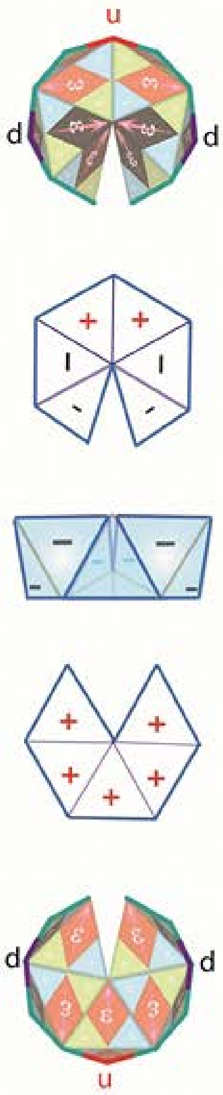

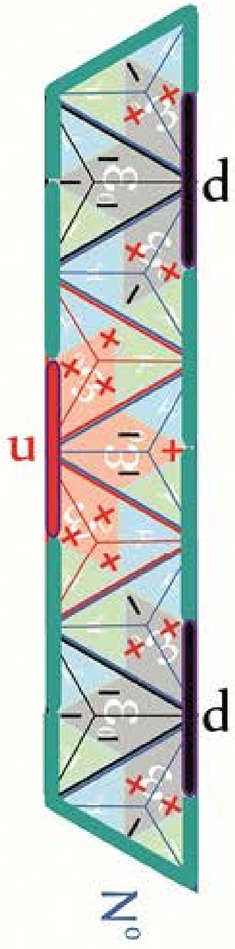

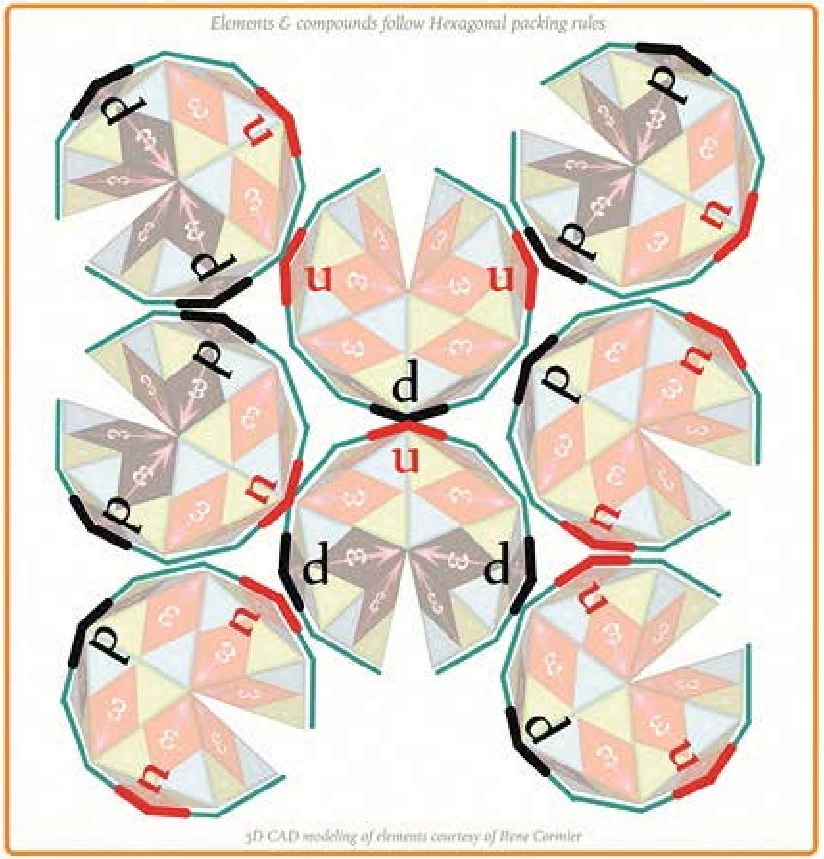

Elements & compounds follow Hexagonal packing rules

3D CAD modelling of elements courtesy of Ilene Cormier

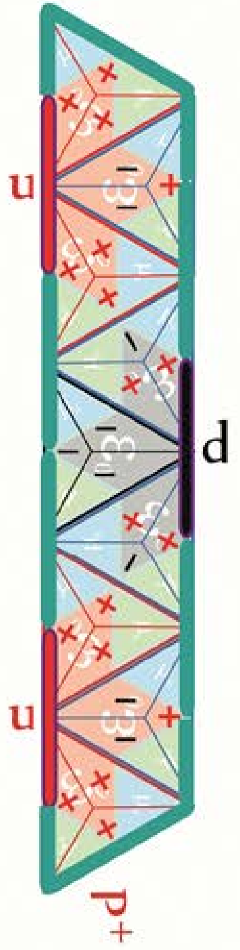

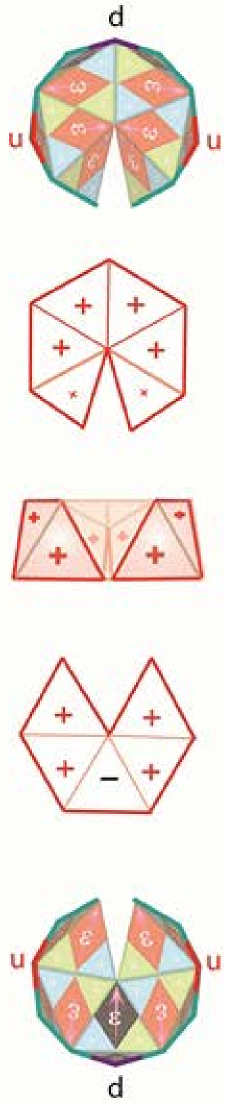

E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements

UP Quark
Positive Electric field apex

DOWN Quark
Negative Electric field apex

Copyright ABRAHAM [2008] - All rights reserved

Insulators and Conductors

The position of electrons in Nuclei within Atomic Elements
results in the properties of Insulators or Conductors

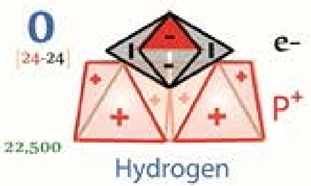
0
[24-24]
e-
P+
22,500
Hydrogen

Coloumbic forces

Electrons are attracted to the residual EM net (+12) positive charge of Protons or n[+12] unbalanced Ionic charges of nuclei

Conductor

Electrical energies move around the material via boson exchanges and electron movement

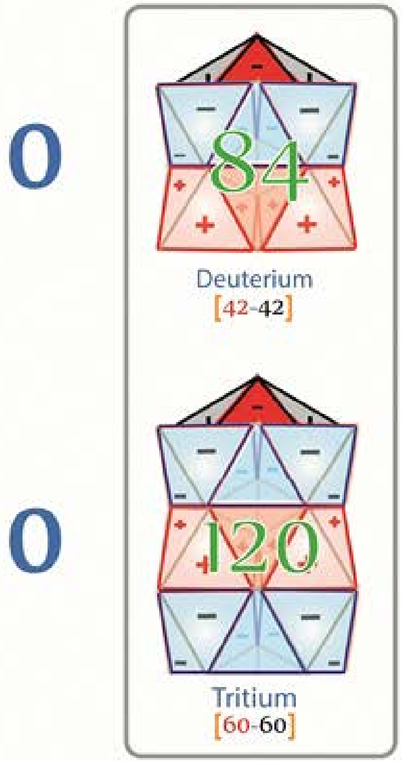
0
84
Deuterium
[42-42]
0
120
Tritium
[60-60]

≈

Insulator

Electrical energy is fixed within the nucleus as electrostatic charges & released upon demand via electron rotation/motion within the nucleus

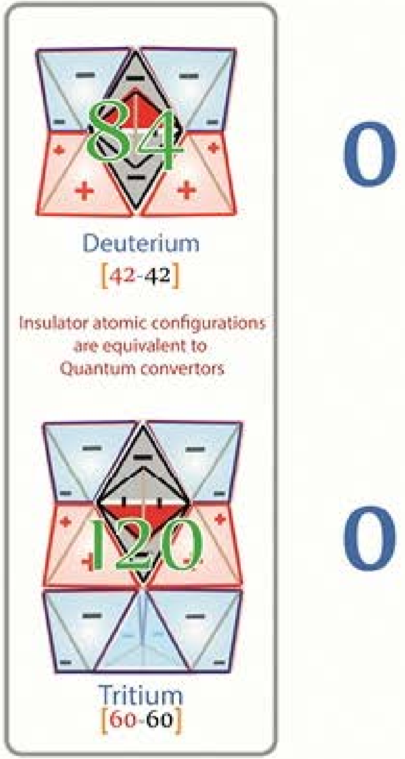
84
Deuterium
[42-42]
0
120
Tritium
[60-60]
0

Insulator atomic configurations are equivalent to Quantum convertors

Conductive materials contain 'free' electrons that can be readily or easily moved within the material

Insulator materials have electrons that are 'bound' tightly to the atoms and store charges locally where they are applied

Ions

Charge (energy) is moved around material via electron movement

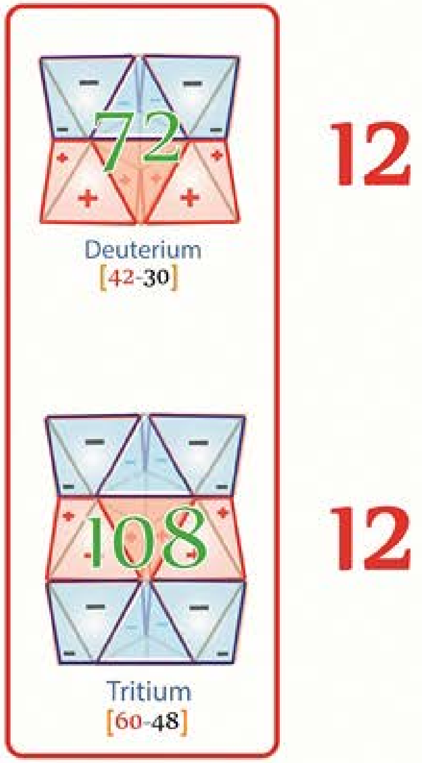
72
Deuterium
[42-30]
12
108
Tritium
[60-48]
12

Materials that have been ionised are more likely to become Conductors as they easily attract and bind free electrons to them

Copyright ABRAHAM [2008] - All rights reserved

Nucleon Quark Arrangement

The nuclei arrangement of each atomic shell [quantum level] is the result of quark EM field interactions

Quantum levels

Atomic shells

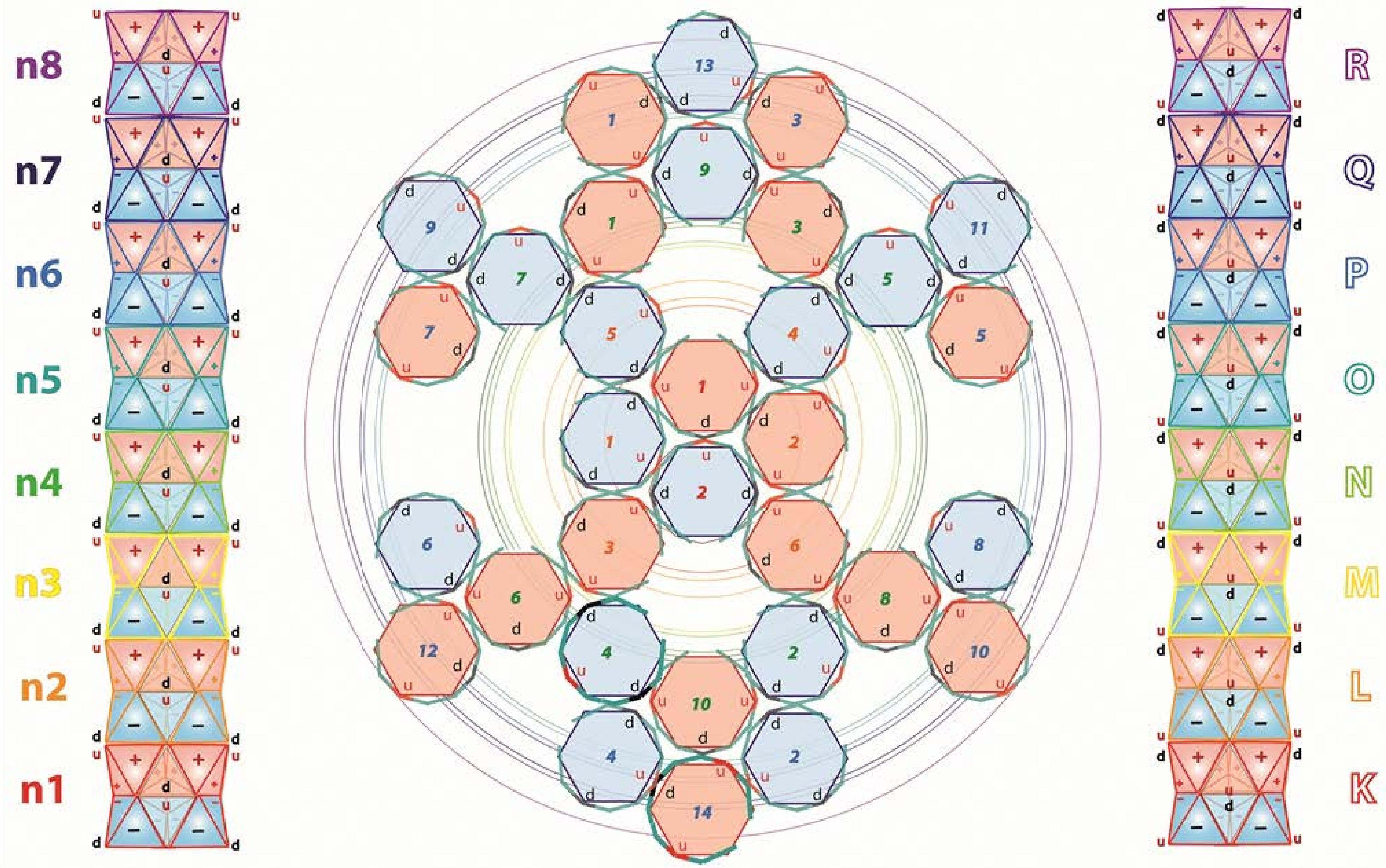
n8
n7
n6
n5
n4
n3
n2
n1
R
Q
P
O
N
M
L
K
u
d
+
−
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Nucleon Charges and Bonding

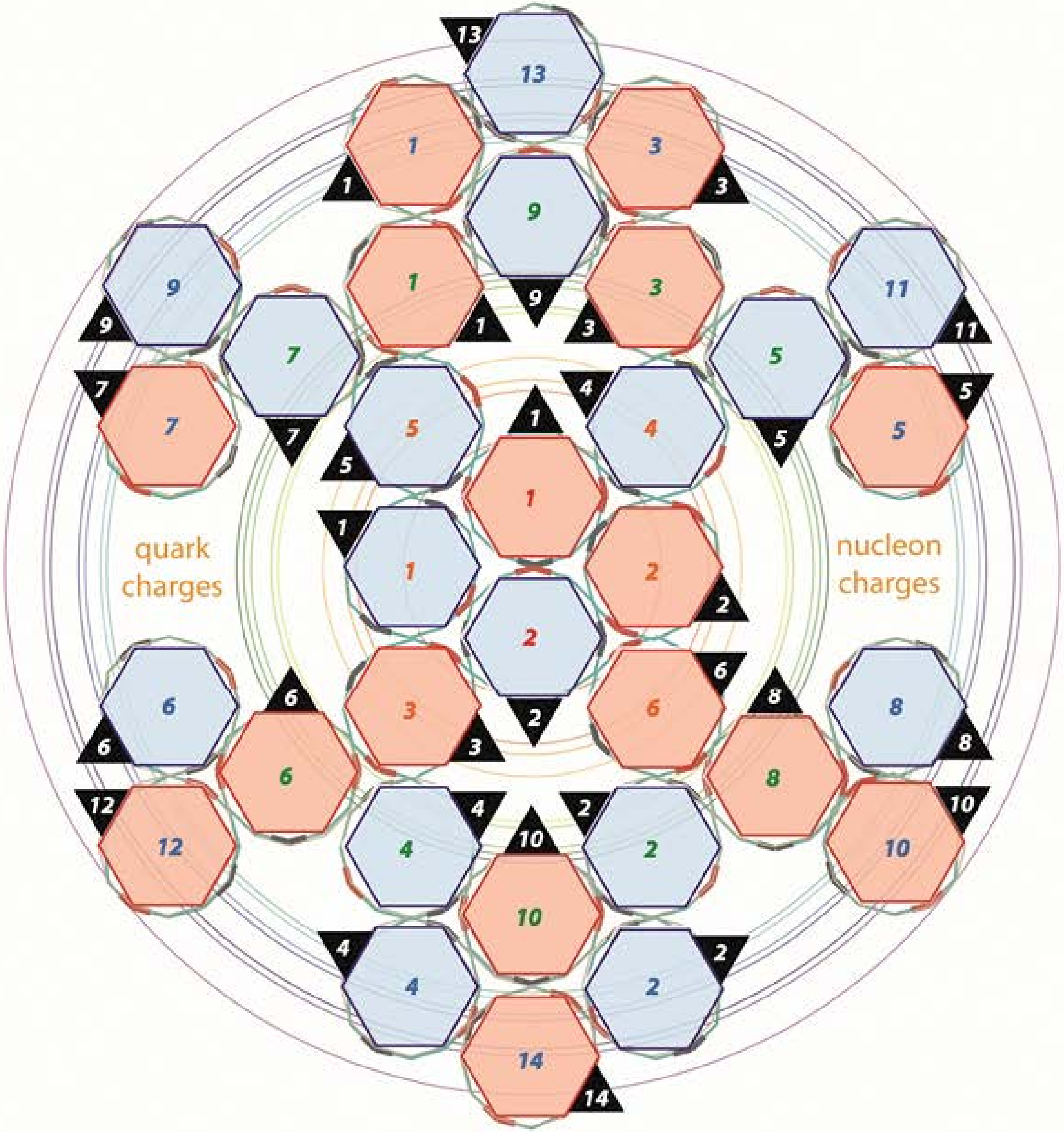
quark charges
nucleon charges

(Strong force - topological Electric Points)

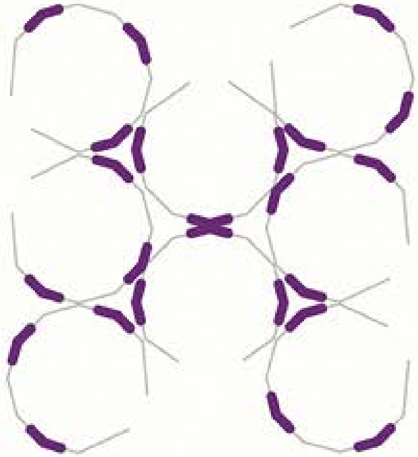

residual EM forces

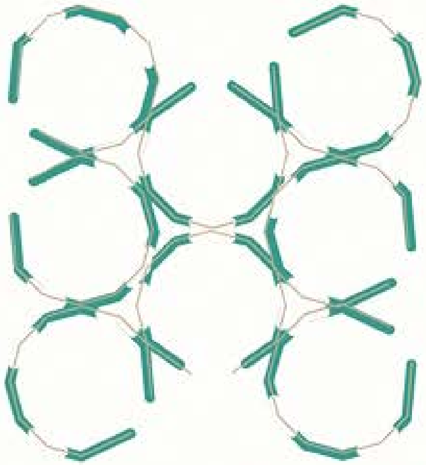

(Strong force - topological Magnetic dipoles)

Bound electron arrangements

Externally bound electrons produce sub-orbital patterns
different to the electron orbitals of internally bound electrons

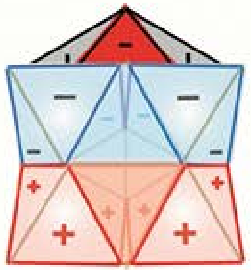

Deuterium

electrons are externally bound
to the Deuteron nuclei

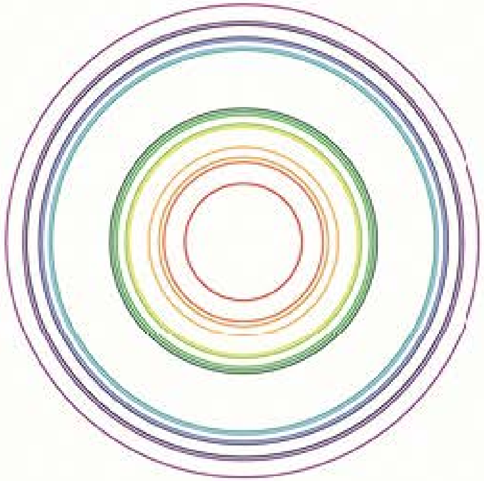

Conductors

Charges are free to move and equalise

electrons require less energy
to 'break free' from Nuclei

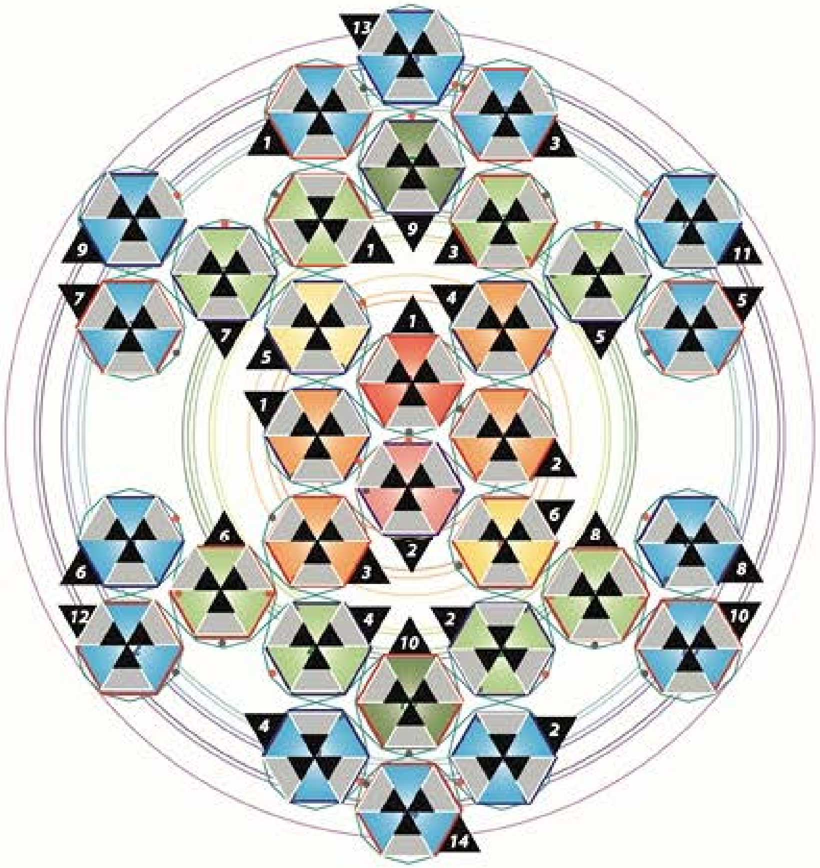

The electron orbitals of
conductors are lower energies
than those of insulators

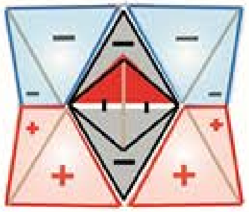

Deuterium

electrons are internally bound
in the Deuteron nuclei

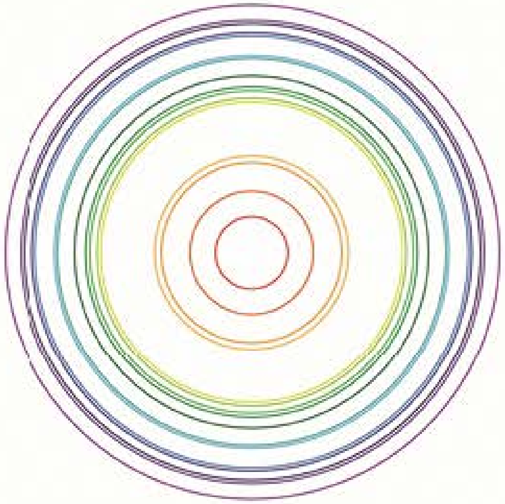

Insulators

Charges are bound to specific locations

electrons require more energy
to 'break free' from Nuclei

2D mass-energy geometries form the fabric of 3D Matter topologies

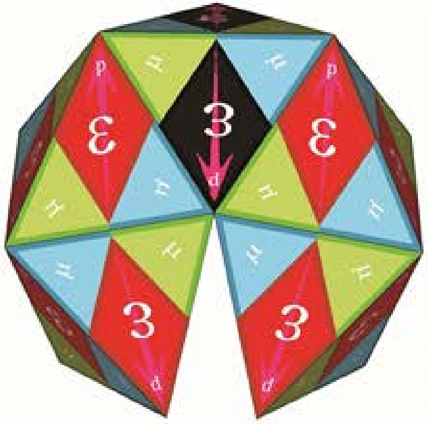

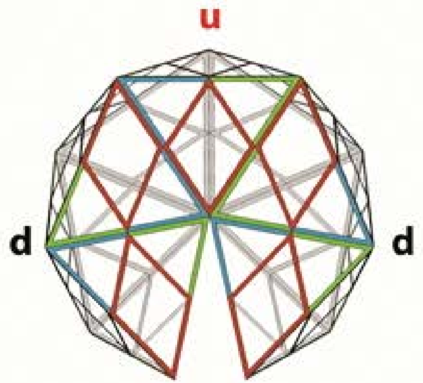

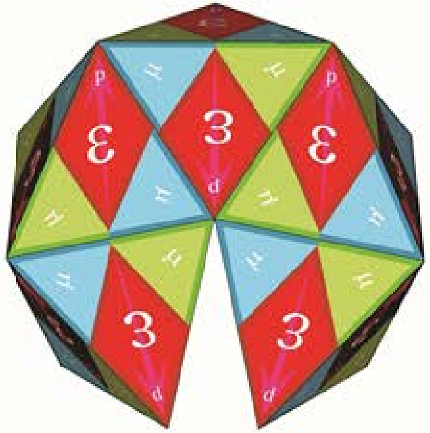

Proton

12
[24-12]

nett Charge
component charges

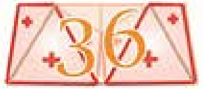

charged mass-energy geometry

20π

Matter topology

12
[42-30]

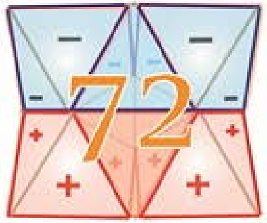

40π

Deuteron

nett Charge
component charges

0
[18-18]

charged mass-energy geometry

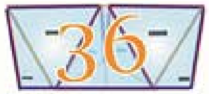

Matter topology

20π

Neutron

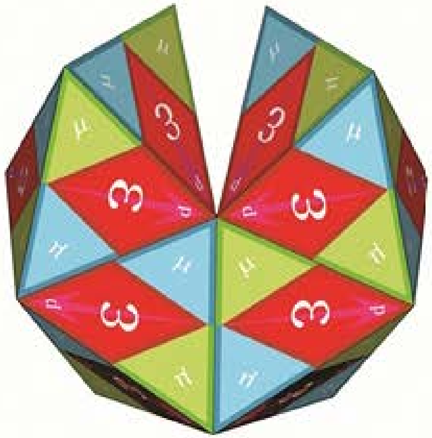

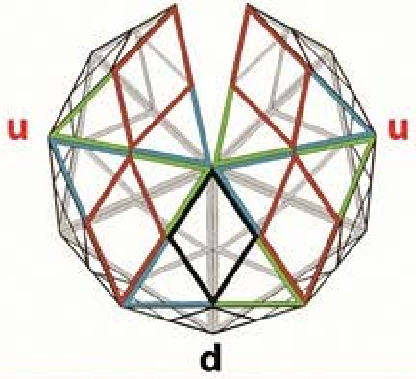

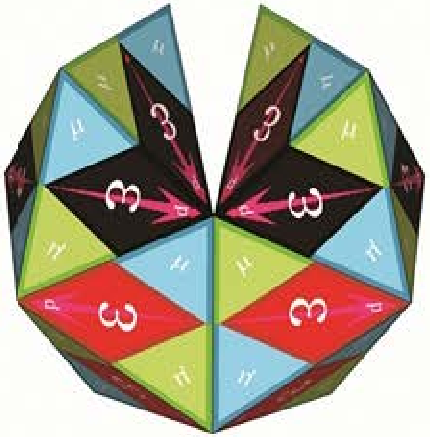

Charge provides the framework for the mass-energy geometry of Matter

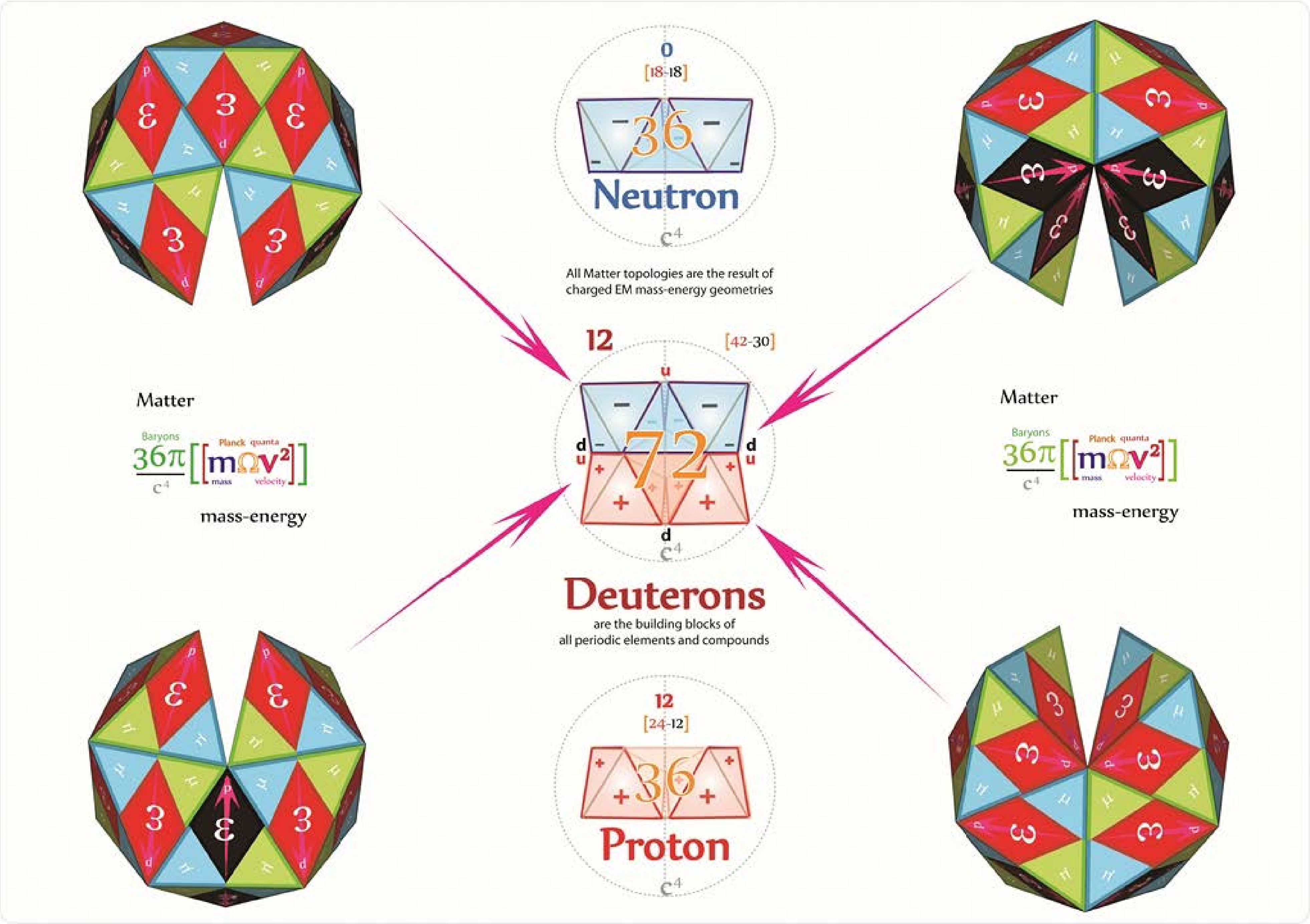
0
[18-18]
36
Neutron
c^4
12
[42-30]
u
d
u
d
u
72
d
c^4
12
[24-12]
36
Proton
c^4
ε
p
d
μ

All Matter topologies are the result of
charged EM mass-energy geometries

Matter

Baryons Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy

Matter

Baryons Planck quanta

$$\frac{36\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

mass-energy

Deuterons
are the building blocks of
all periodic elements and compounds

Copyright ABRAHAM [2008] - All rights reserved

Strong Forces and Nuclear Bonding

How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?

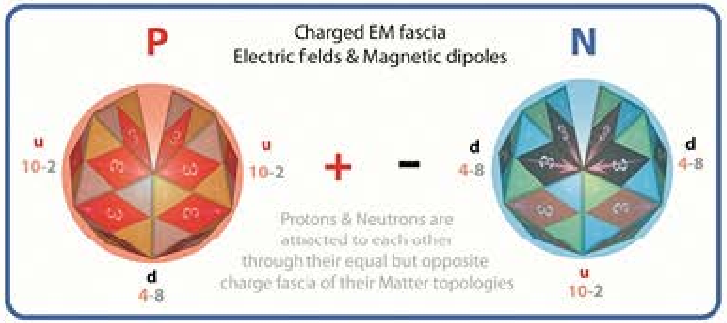
P
Charged EM fascia
Electric fields & Magnetic dipoles
N
u
10-2
u
10-2
+
-
d
4-8
d
4-8
d
4-8
u
10-2

Protons & Neutrons are attracted to each other through their equal but opposite charge fascia of their Matter topologies

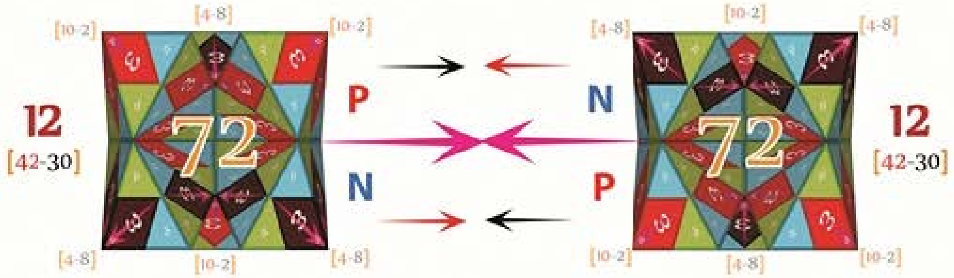
[4-8]
[10-2]
[10-2]
12
[42-30]
72
P
N
N
P
72
12
[42-30]
[4-8]
[10-2]
[4-8]
[4-8]
[10-2]
[4-8]
[10-2]
[4-8]
[10-2]

Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements

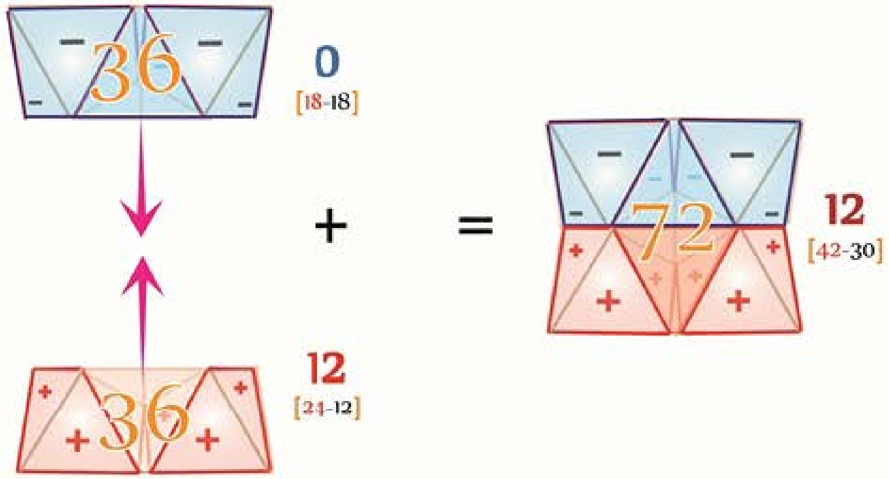
36
0
[18-18]
+
=
72
12
[42-30]
36
12
[24-12]

The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges

The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

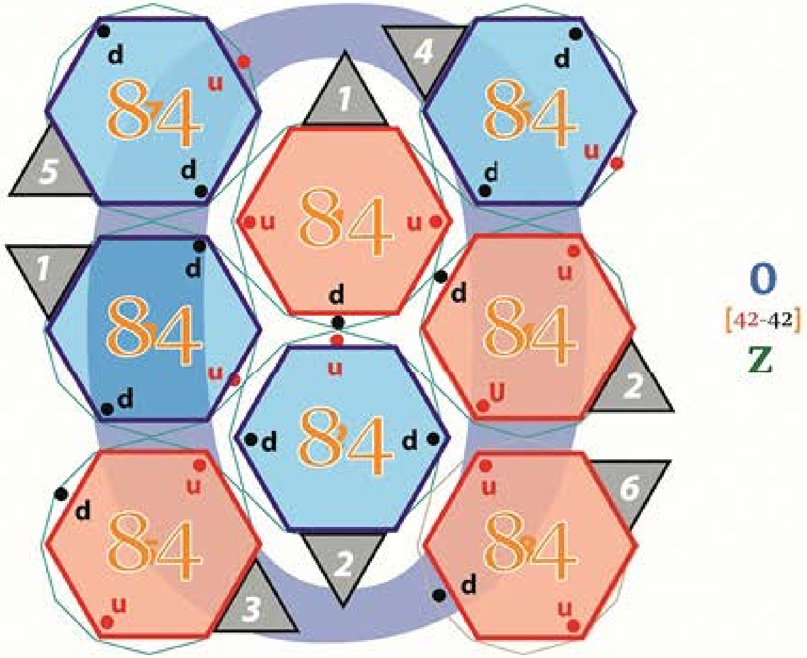
84
84
84
84
84
84
84
84
d
u
1
2
3
4
5
6
0
[42-42]
Z

Copyright ABRAHAM [2008] - All rights reserved

Strong Nucleonic Bonding

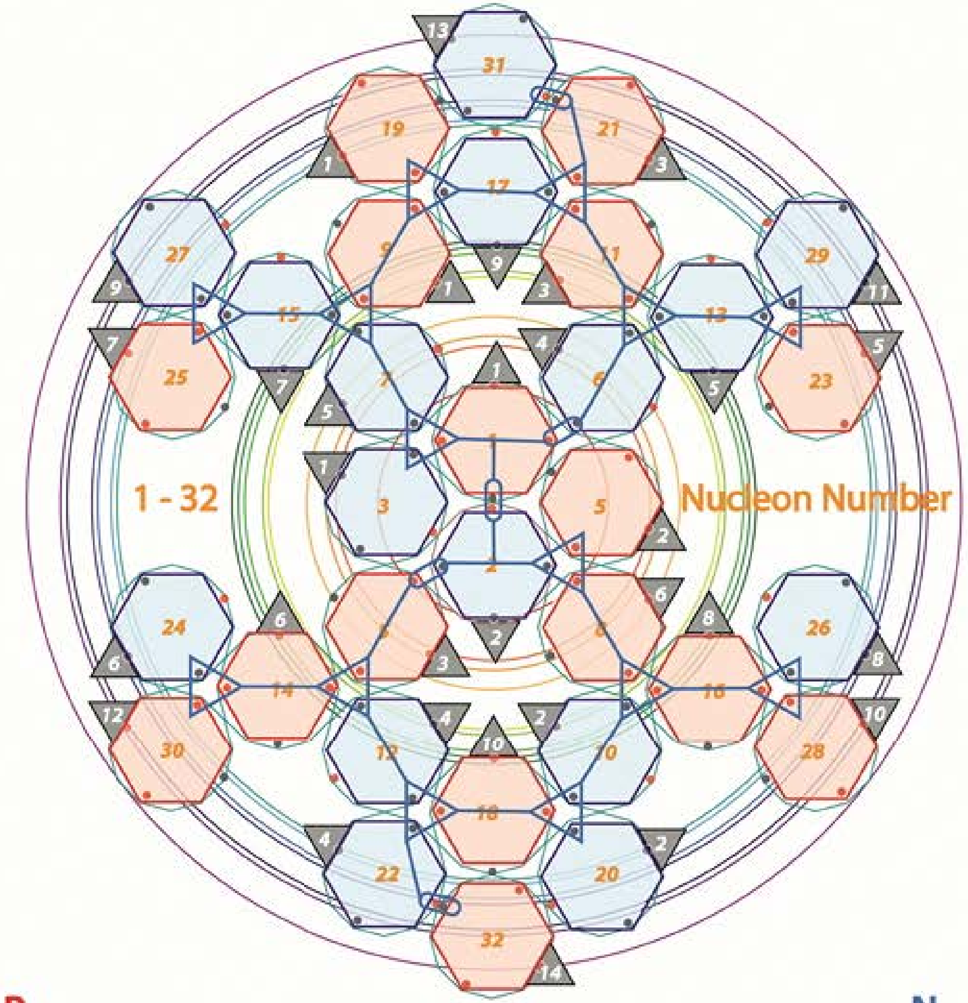
13
31
19
21
1
3
17
27
29
9
1
3
11
9
15
13
7
7
4
5
25
23
7
5
1
5
1 - 32
7
3
5
Nucleon Number
2
24
26
6
6
3
8
8
2
6
14
16
12
4
10
2
10
30
28
12
10
4
18
2
22
20
32
14

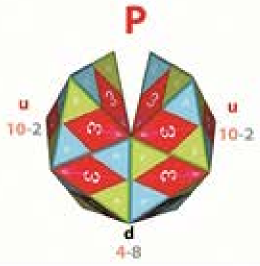
P
u
10-2
u
10-2
d
4-8

+

In addition to the Strong colour force
a separate residual binding force arises
from the external apexes formed by
Positive and Negative Electric points of quarks
in each Baryon [nuclear-chemical bonds]

-

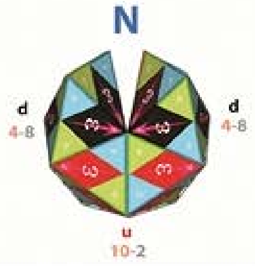
N
d
4-8
d
4-8
u
10-2

All energy seeks equilibrium

Strong E-force
nuclear linkages

External Electric field points
bind via Charge fascia interactions
with Plus and Minus Electric points
combining and sharing energy
throughout the resultant nuclei

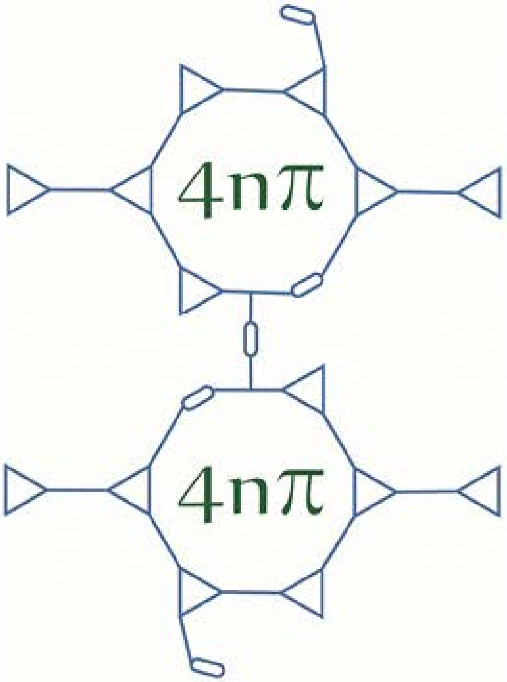

● *Positive Electric field apex*

Negative Electric field apex ●

Copyright ABRAHAM [2008] - All rights reserved

Hydrogenic vs Nucleonic electron binding

If a unbound Proton attracts an Electron
the Electron can be bound to the nuclei
in a number of differing orientations
[each with differing spin energies]

All atomic nuclei (and elements) are Deuteron'ic nuclei with a mixture of *orthagonal*, parallel and anti-paralllel spin orientations

(this is why Rydberg is less accurate for elemental nuclei compared to Hydrogenic atoms - see QM spin)

Electrons can be bound to deuteron nuclei
in four distinct orientations [2 horizontal & 2 vertical]
with each spin coupling orientation producing
differing energy electron orbitals
[wrt to the nuclear magnetic moments]

Ejecting electrons from atomic nuclei
by adding energies to their KEM fields
[the Photoelectric effect]
creates Positive Ions

Vertically orientated electrons
within Proton-Neutron Nuclei [Deuterons]
create quantum synchronous convertor geometries

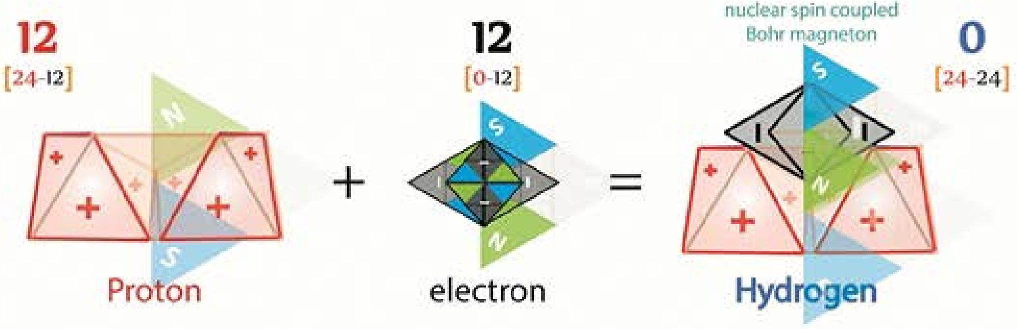
12
[24-12]
Proton
+
12
[0-12]
electron
=
nuclear spin coupled
Bohr magneton
0
[24-24]
Hydrogen

The energy levels of Baryons determines the KEM field energy of bound electrons

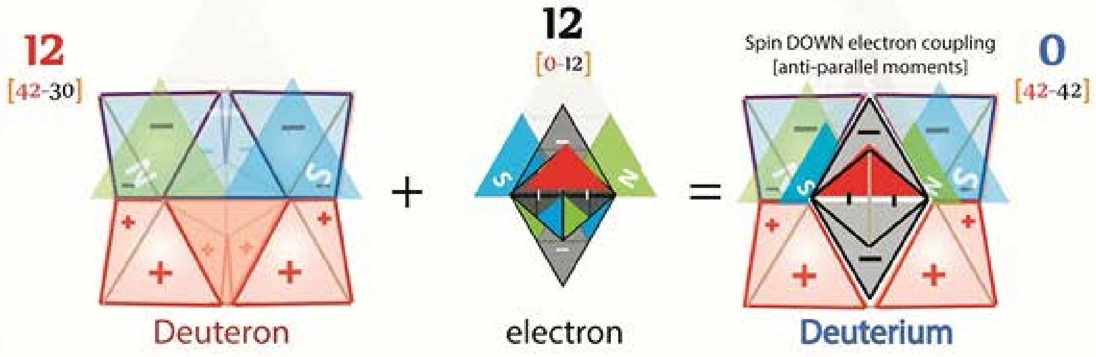
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin DOWN electron coupling
[anti-parallel moments]
0
[42-42]
Deuterium

Bohr magnetons are always referenced wrt the Nuclear magneton

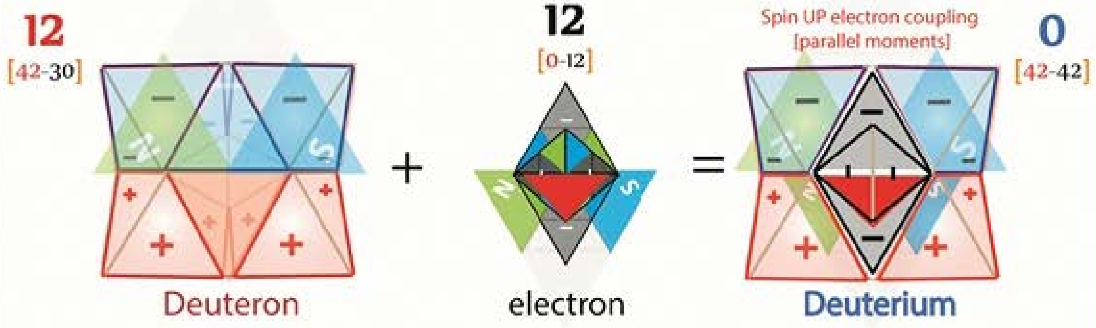
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin UP electron coupling
[parallel moments]
0
[42-42]
Deuterium

electrons produce stronger magnetic moments due to their mass-charge quotient

Atomic Nucleus Master Template

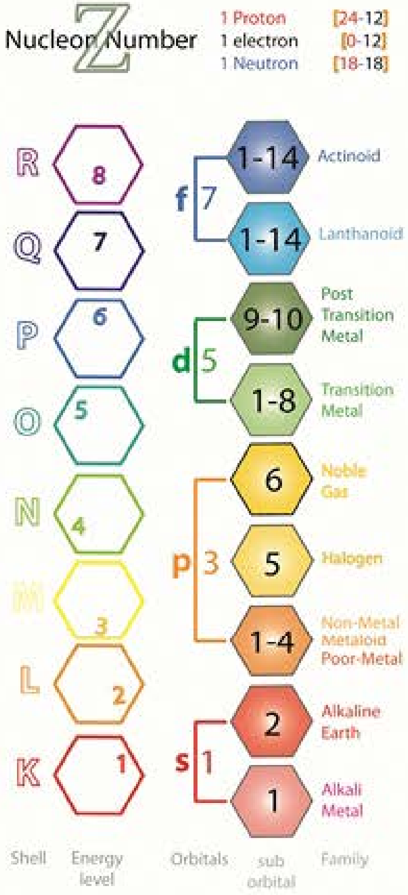
Nucleon Number
Z
1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

○ **s** *1 sub-Orbital (2 electrons max)*
○ **p** *3 sub-Orbitals (6 electrons max)*
○ **d** *5 sub-Orbitals (10 electrons max)*
○ **f** *7 sub-Orbitals (14 electrons max)*

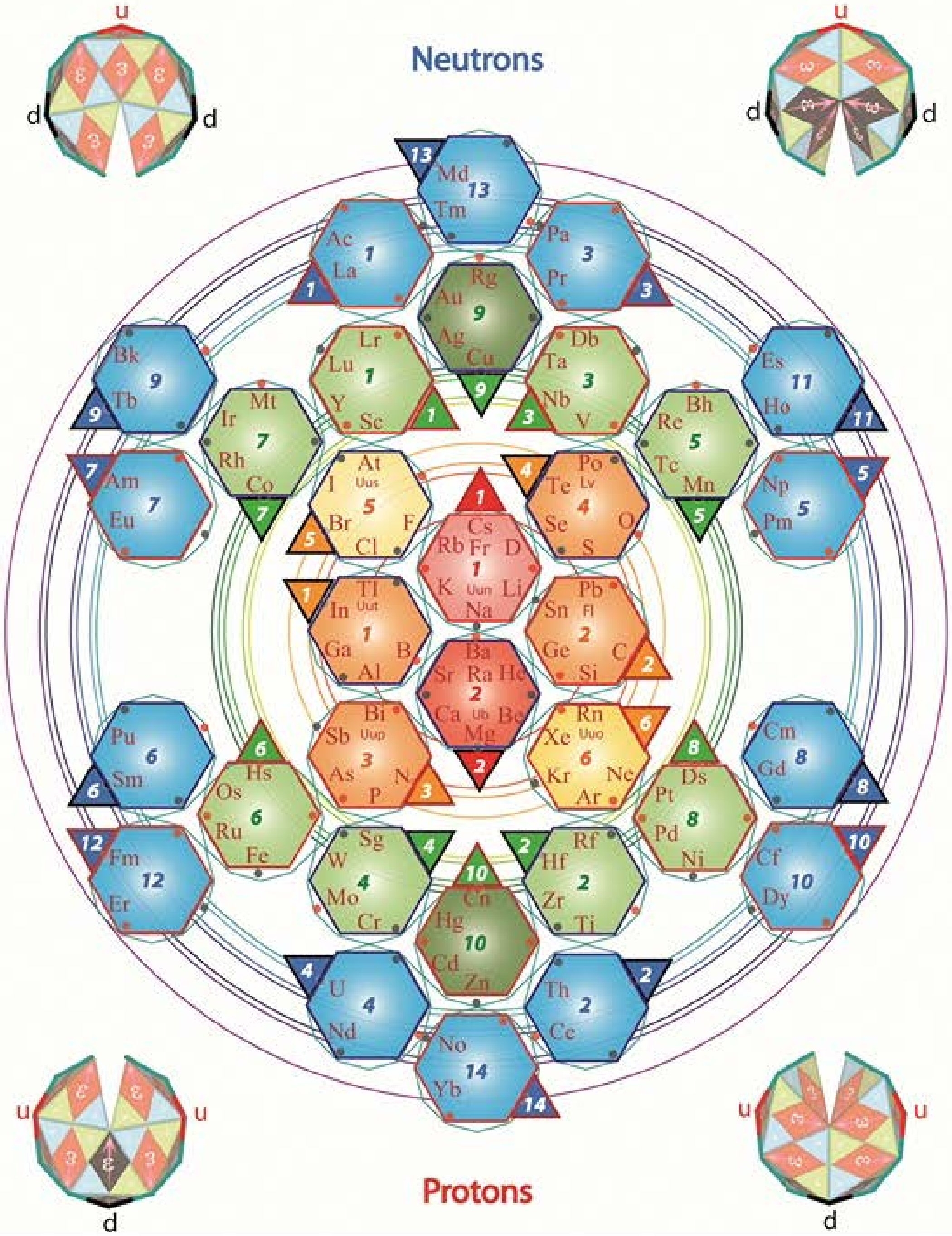
Neutrons
u
d
d
u
d
d
13 Md Tm
1 Ac La
3 Pa Pr
9 Rg Au Ag Cu
9 Bk Tb
1 Lr Lu Y Sc
3 Db Ta Nb V
11 Es Ho
7 Mt Ir Rh Co
5 Bh Re Tc Mn
7 Am Eu
5 At I Uus Br Cl F
4 Po Te Lv Se S O
5 Np Pm
1 Cs Rb Fr D K Uun Li Na
1 Tl In Uut Ga B Al
2 Pb Sn Fl Ge C Si
2 Ba Sr Ra He Ca Ub Be Mg
6 Pu Sm
6 Hs Os Ru Fe
3 Bi Sb Uup As N P
6 Rn Xe Uuo Kr Ne Ar
8 Ds Pt Pd Ni
8 Cm Gd
12 Fm Er
4 Sg W Mo Cr
10 Cn Hg Cd Zn
2 Rf Hf Zr Ti
10 Cf Dy
4 U Nd
2 Th Ce
14 No Yb
u
u
d
u
u
d
Protons

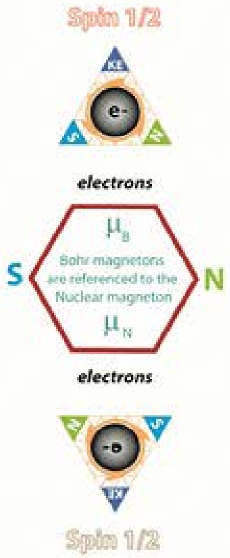
Spin 1/2
e-
electrons
S
μB
Bohr magnetons are referenced to the Nuclear magneton
μN
N
electrons
-e
Spin 1/2

The Chemical elements

Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms

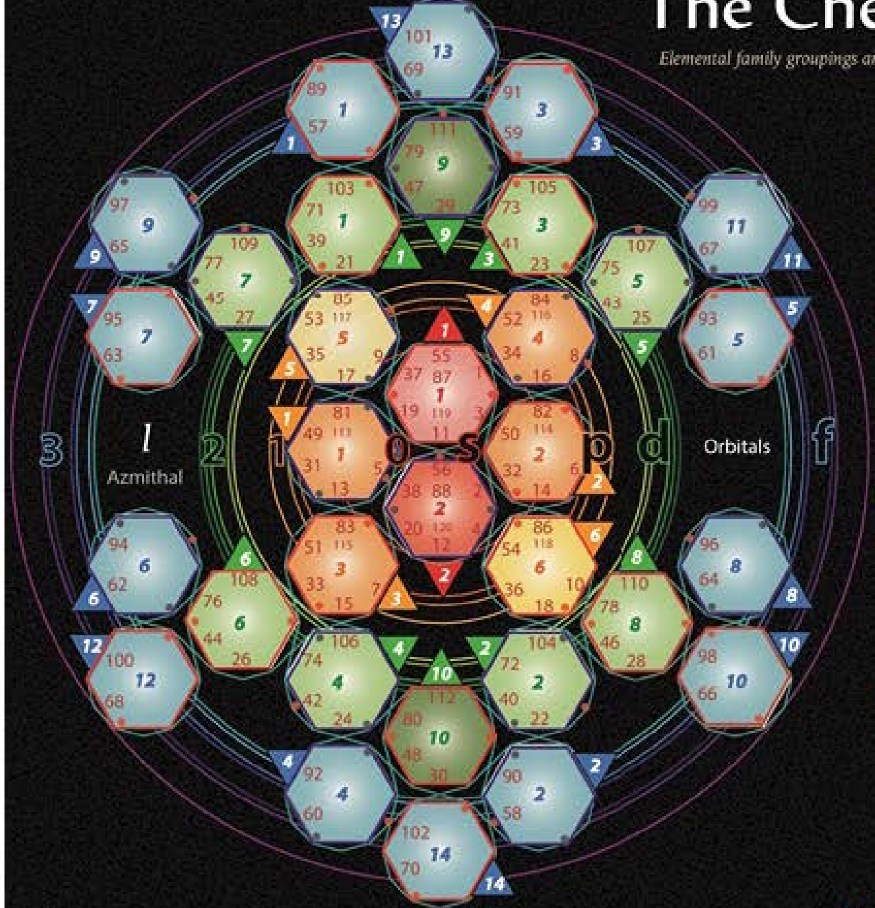

Orbitals	sub orbital	Family
f 7	1-14	Actinoid
	1-14	Lanthanoid
d 5	9-10	Post Transition Metal
	1-8	Transition Metal
p 3	6	Noble Gas
	5	Halogen
	1-4	Non-Metal Metaloid Poor-Metal
s 1	2	Alkaline Earth
	1	Alkali Metal

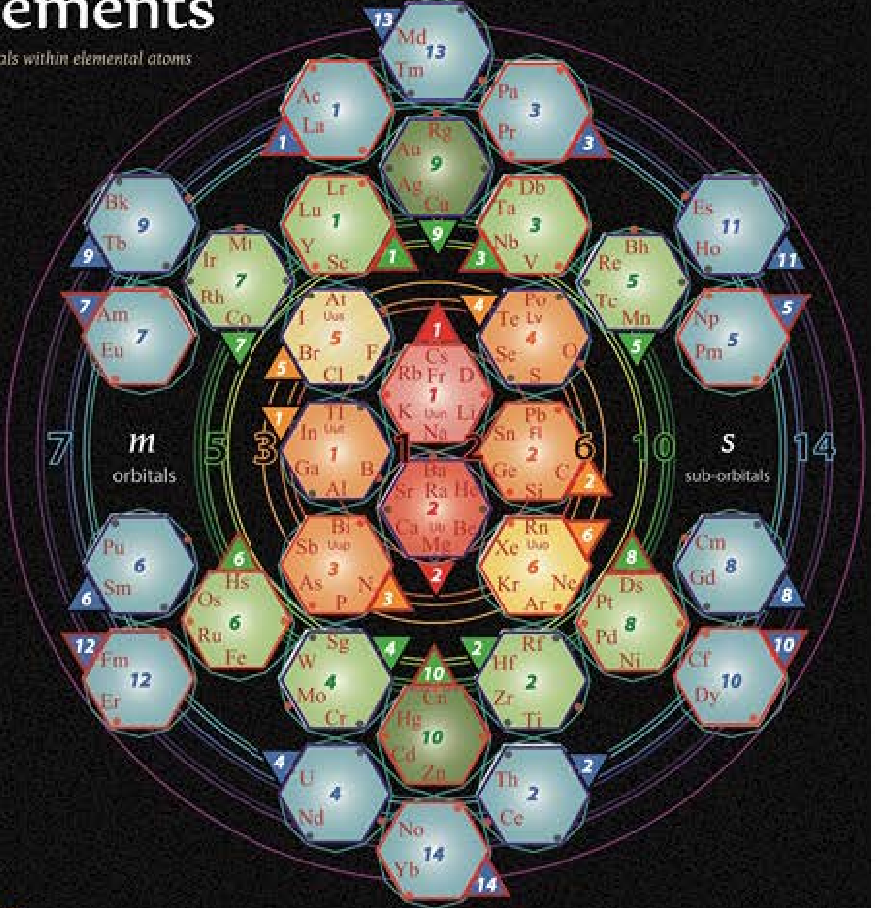

Quantum numbers

	3							2					1			0		1			2					3							
	+3	+2	+1	0	-1	-2	-3	+2	+1	0	-1	-2	+1	0	-1			-1	0	+1	-2	-1	0	+1	+2	-3	-2	-1	0	+1	+2	+3	
R																119	120																8
Q													117	115	113	87	88	114	116	118													7
P								111	109	107	105	103	85	83	81	55	56	82	84	86	104	106	108	110	112								6
O	101	99	97	95	93	91	89	79	77	75	73	71	53	51	49	37	38	50	52	54	72	74	76	78	80	90	92	94	96	98	100	102	5
N	69	67	65	63	61	59	57	47	45	43	41	39	35	33	31	19	20	32	34	36	40	42	44	46	48	58	60	62	64	66	68	70	4
M	13	11	9	7	5	3	1	29	27	25	23	21	17	15	13	11	12	14	16	18	22	24	26	28	30	2	4	6	8	10	12	14	3
L								9	7	5	3	1	9	7	5	3	4	6	8	10	2	4	6	8	10								2
K													5	3	1	1	2	2	4	6													1
				f						d				p		s			p				d						f				

Atomic Shells

Energy levels

Orbitals & sub-Orbitals

Element Numbers

Z *
1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]

Each elemental nuclei is made from Deuterium

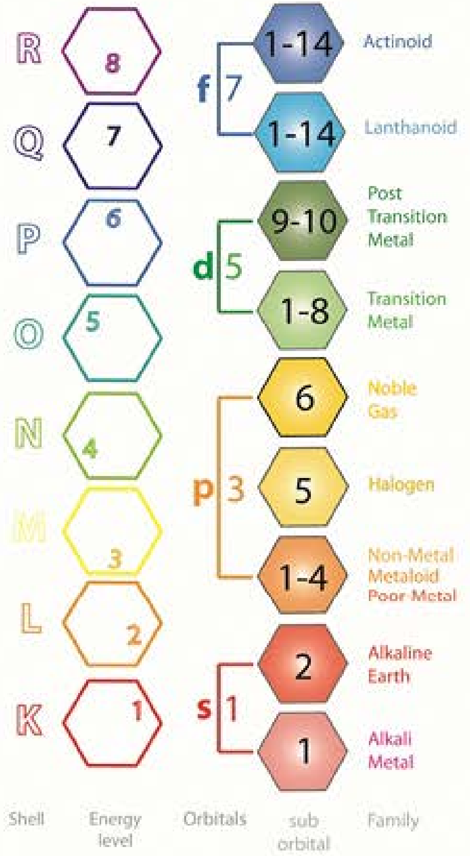
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor Metal
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

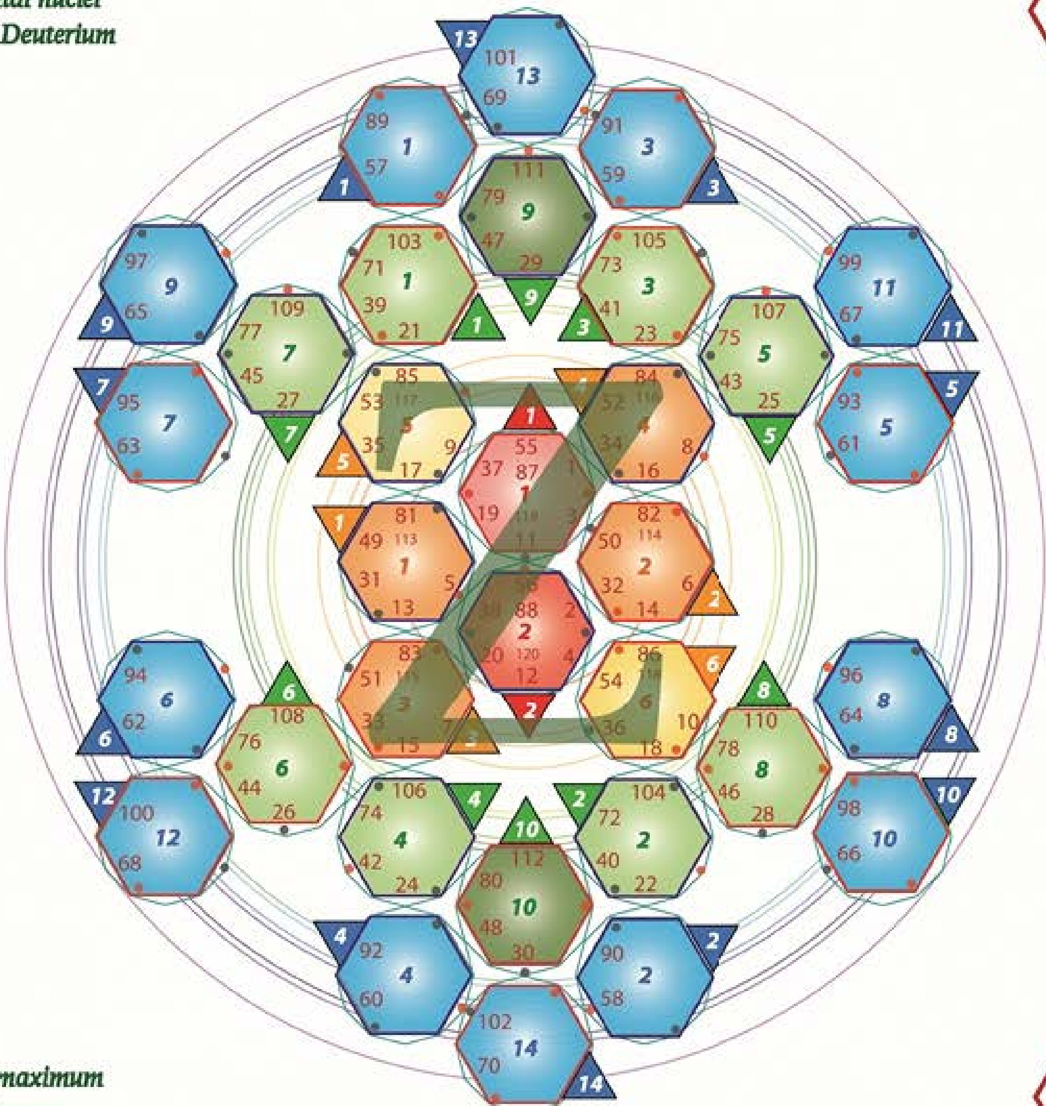
101 13 69
89 1 57
91 3 59
111 79 9 47 29
97 9 65
99 11 67
109 77 7 45 27
103 71 1 39 21
105 73 3 41 23
107 75 5 43 25
95 7 63
93 5 61
85 53 117 5 35 17 9
84 52 116 4 34 16 8
55 37 87 1 19 119 11 3
81 49 113 1 31 13 5
82 50 114 2 32 14 6
56 38 88 2 20 120 12 4
83 51 115 3 33 15 7
86 54 118 6 36 18 10
Z
94 6 62
96 8 64
108 76 6 44 26
110 78 8 46 28
100 12 68
98 10 66
106 74 4 42 24
104 72 2 40 22
112 80 10 48 30
92 4 60
90 2 58
102 14 70

Atomic shells

○ **s** *1 sub-Orbital (2 electrons max)*

○ **p** *3 sub-Orbitals (6 electrons max)*

○ **d** *5 sub-Orbitals (10 electrons max)*

○ **f** *7 sub-Orbitals (14 electrons max)*

There are a maximum of 120 elements possible

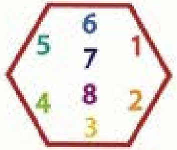

Energy levels

Element Names

Z *
1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]

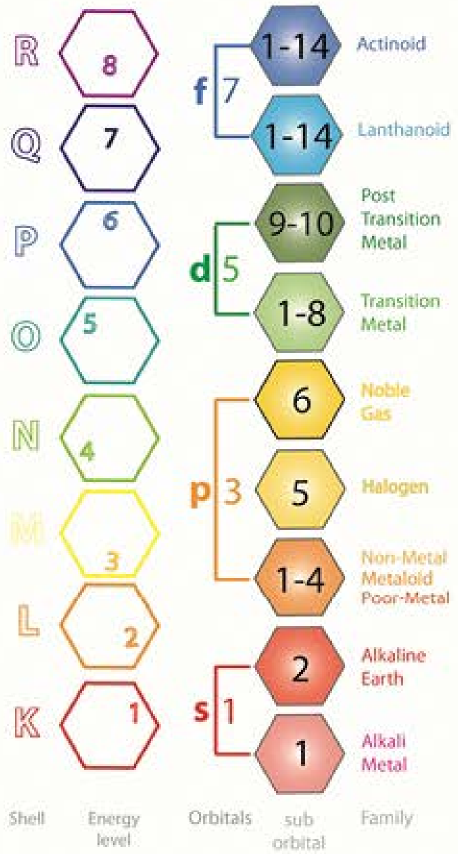
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

o s 1 sub-Orbital (2 electrons max)
O p 3 sub-Orbitals (6 electrons max)
O d 5 sub-Orbitals (10 electrons max)
O f 7 sub-Orbitals (14 electrons max)

There are a maximum of 120 elements possible

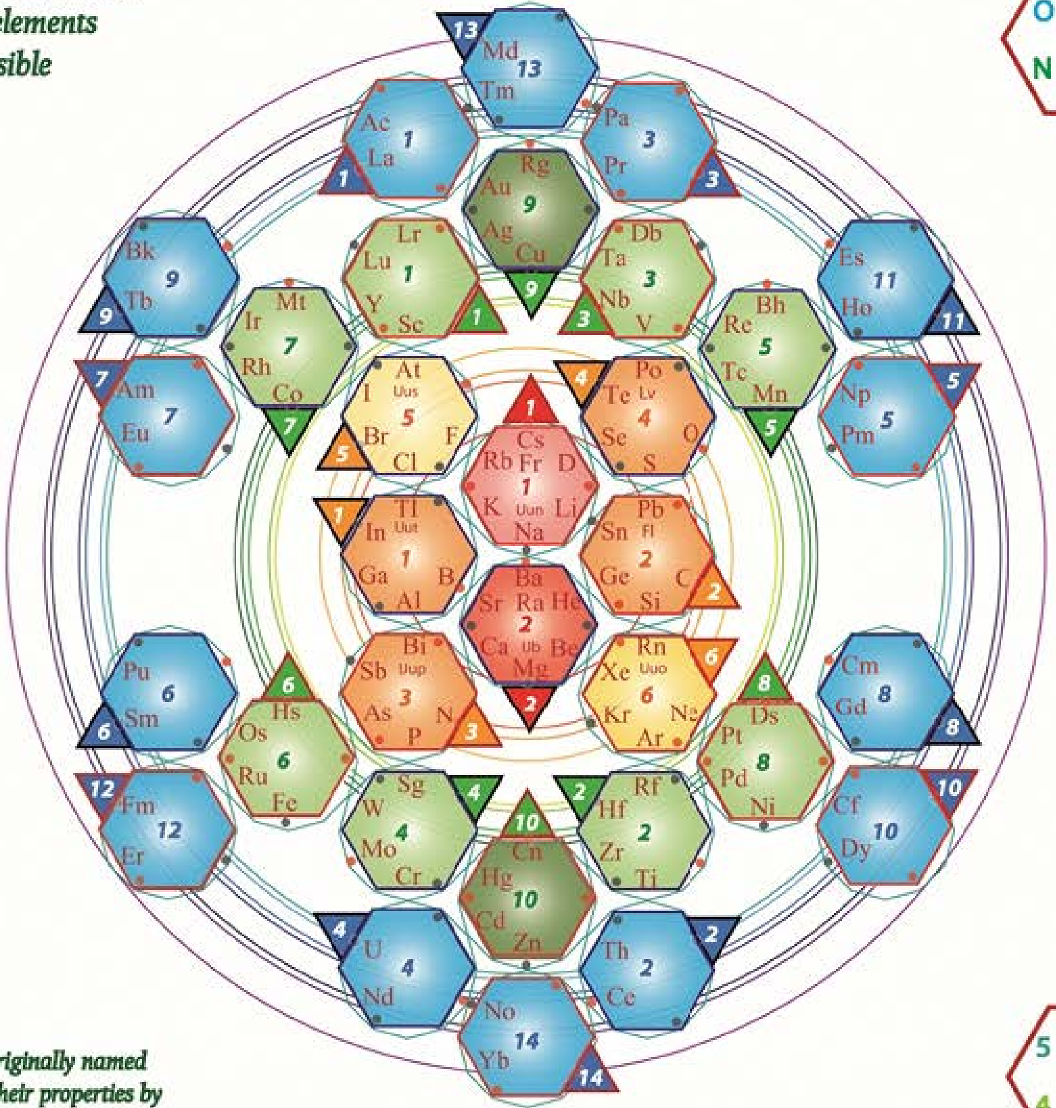
13 Md 13 Tm
Ac 1 La 1
Pa 3 Pr 3
Rg Au 9 Ag Cu 9
Bk 9 Tb 9
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Es 11 Ho 11
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
Am 7 Eu 7
At I Uus 5 Br Cl F 5
Po Te Lv 4 Se S O 4
Np 5 Pm 5
1 Cs Rb Fr D 1 K Uun Li Na
Tl In Uut 1 Ga Al B 1
Pb Sn Fl 2 Ge Si C 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Bi Sb Uup 3 As P N 3
Rn Xe Uuo 6 Kr Ne Ar 6
Pu 6 Sm 6
Hs Os 6 Ru Fe 6
Ds Pt 8 Pd Ni 8
Cm Gd 8 8
12 Fm 12 Er
Sg W 4 Mo Cr 4
2 Rf Hf 2 Zr Ti
Cf 10 Dy 10
10 Cn Hg 10 Cd Zn
4 U 4 Nd
2 Th 2 Ce
No 14 Yb 14

Atomic shells

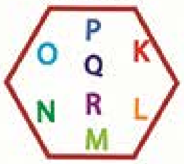

They were originally named according to their properties by their discoverer but have recently been named after famous scientists

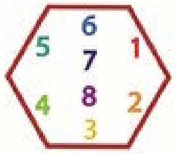

Energy levels

Copyright ABRAHAM [2008] - All rights reserved

Electron orbital configurations

A huge number of differing d and f orbital configurations are possible given the number of nuclei and bond points created by elemental topologies

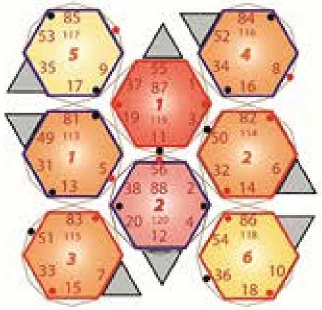
85
53 117
5
35 9
17
84
52 116
4
34 8
16
55
37 87 3
1
19 119
11
81
49 113
1
31 5
13
82
50 114
2
32 6
14
56
38 88 2
2
20 120 4
12
83
51 115
3
33 7
15
86
54 118
6
36 10
18

As the number of nucleons increases so does the complexity of the electron orbitals possible

quantum snowflakes

However all have a stable 'core' grouping of nuclei comprised of s and p electron orbitals

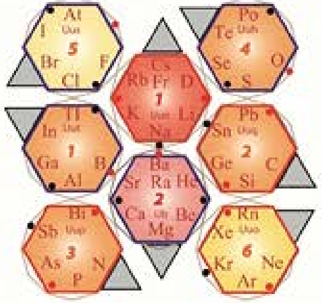
At
I Uus
5
Br F
Cl
Po
Te Uuh
4
Se O
S
Cs
Rb Fr D
1
K Uue Li
Na
Tl
In Uut
1
Ga B
Al
Pb
Sn Uuq
2
Ge C
Si
Ba
Sr Ra He
2
Ca Ubn Be
Mg
Bi
Sb Uup
3
As N
P
Rn
Xe Uuo
6
Kr Ne
Ar

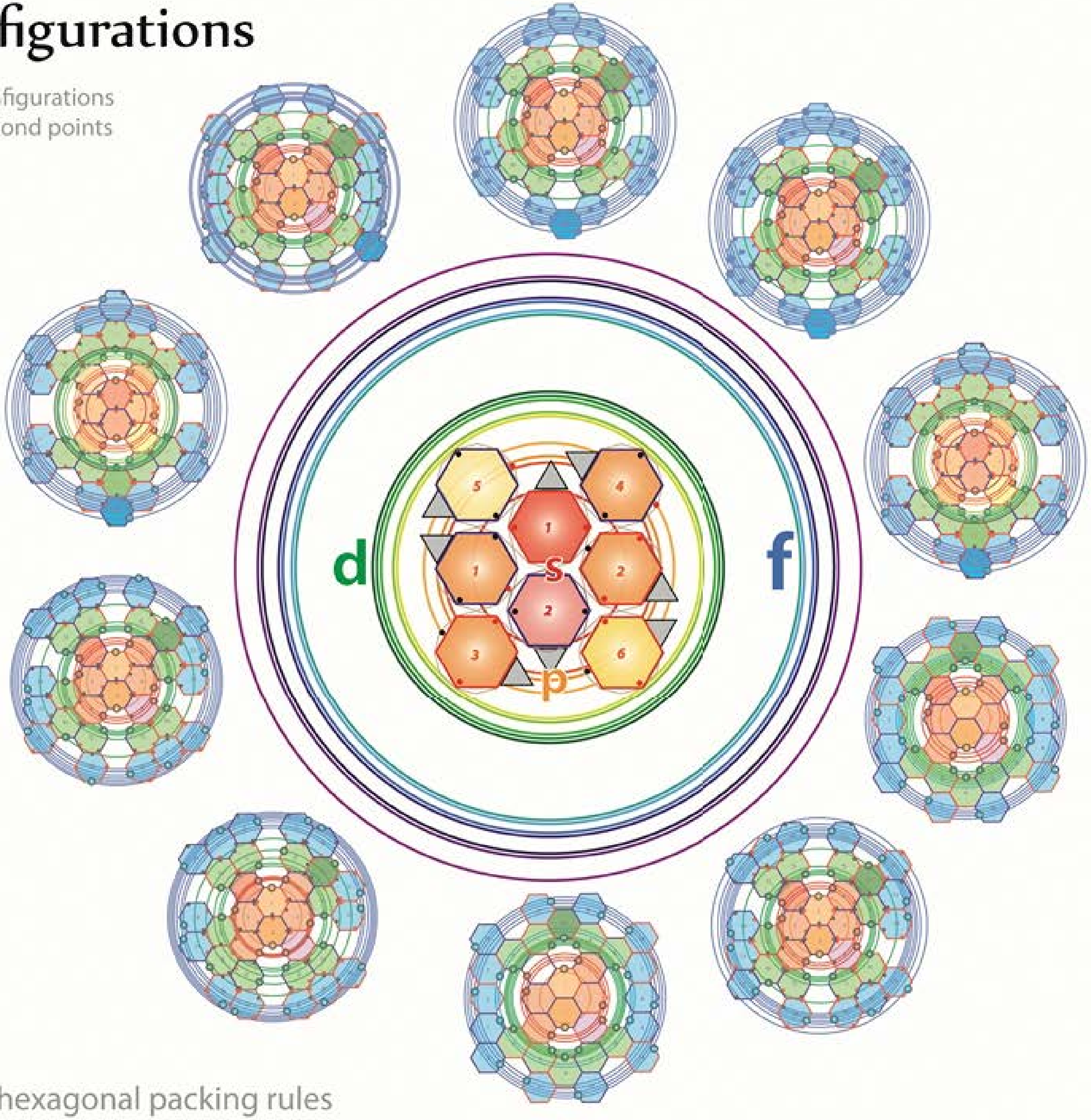
d
f
s
p
5
4
1
1
2
2
3
6

All nuclei bonding closely follows hexagonal packing rules

Atomic Nucleus Exploded view

1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]

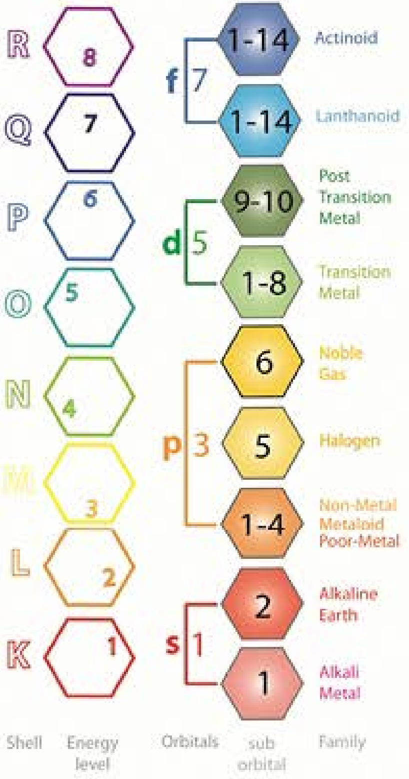
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

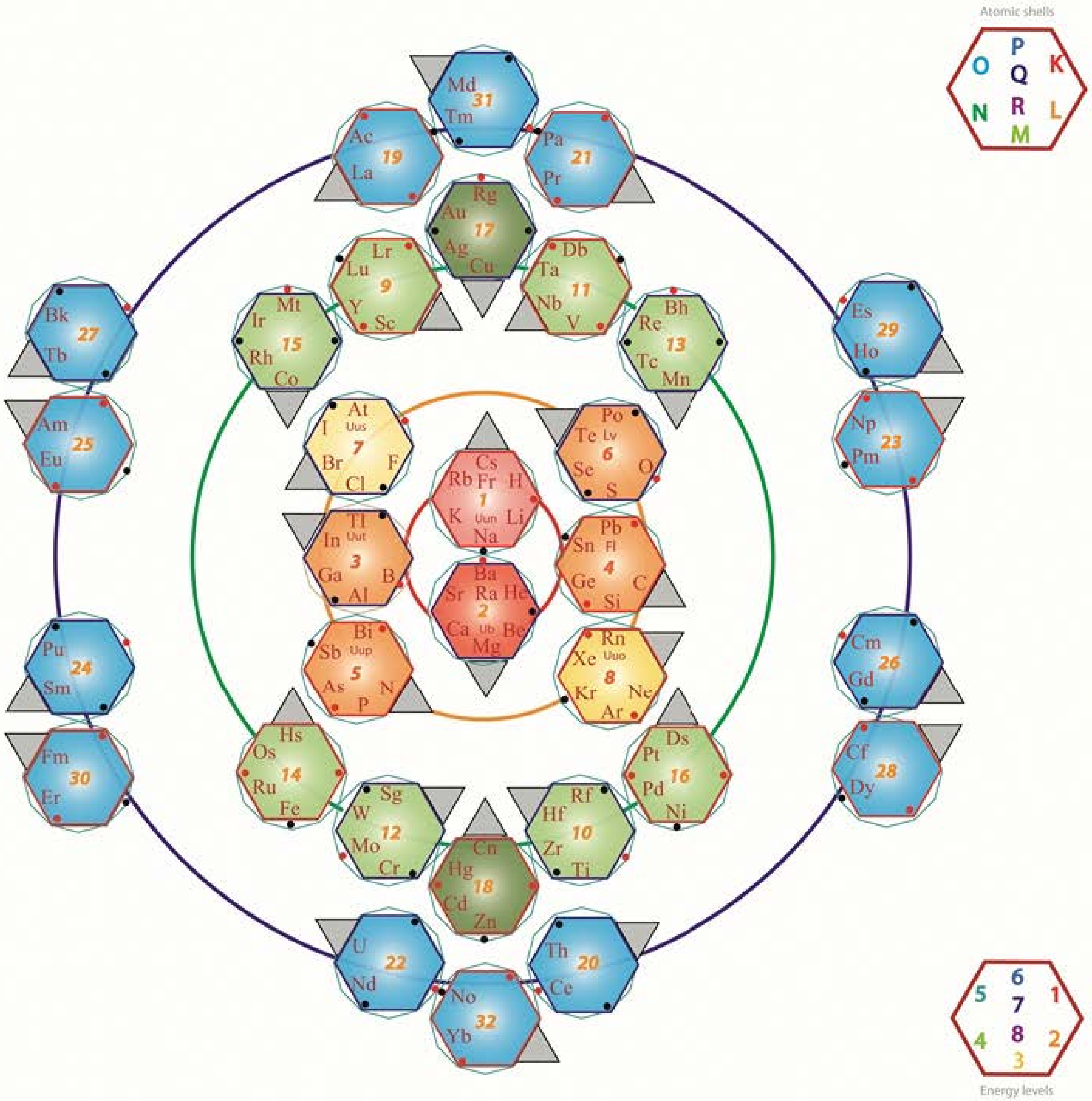
Atomic shells
P
O
Q
K
N
R
L
M
Md Tm 31
Ac La 19
Pa Pr 21
Rg Au Ag Cu 17
Lr Lu Y Sc 9
Db Ta Nb V 11
Mt Ir Rh Co 15
Bh Re Tc Mn 13
Bk Tb 27
Es Ho 29
Am Eu 25
Np Pm 23
At I Uus Br Cl F 7
Cs Rb Fr H K Uun Li Na 1
Po Te Lv Se S O 6
Tl In Uut Ga Al B 3
Ba Sr Ra He Ca Ub Be Mg 2
Pb Sn Fl Ge Si C 4
Bi Sb Uup As P N 5
Rn Xe Uuo Kr Ar Ne 8
Pu Sm 24
Cm Gd 26
Fm Er 30
Cf Dy 28
Hs Os Ru Fe 14
Ds Pt Pd Ni 16
Sg W Mo Cr 12
Rf Hf Zr Ti 10
Cn Hg Cd Zn 18
U Nd 22
Th Ce 20
No Yb 32
Energy levels
6
5
1
7
8
4
2
3

s 1 sub-Orbital (2 electrons max)

p 3 sub-Orbitals (6 electrons max)

d 5 sub-Orbitals (10 electrons max)

f 7 sub-Orbitals (14 electrons max)

Copyright ABRAHAM [2008] - All rights reserved

Atomic radii of elements

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

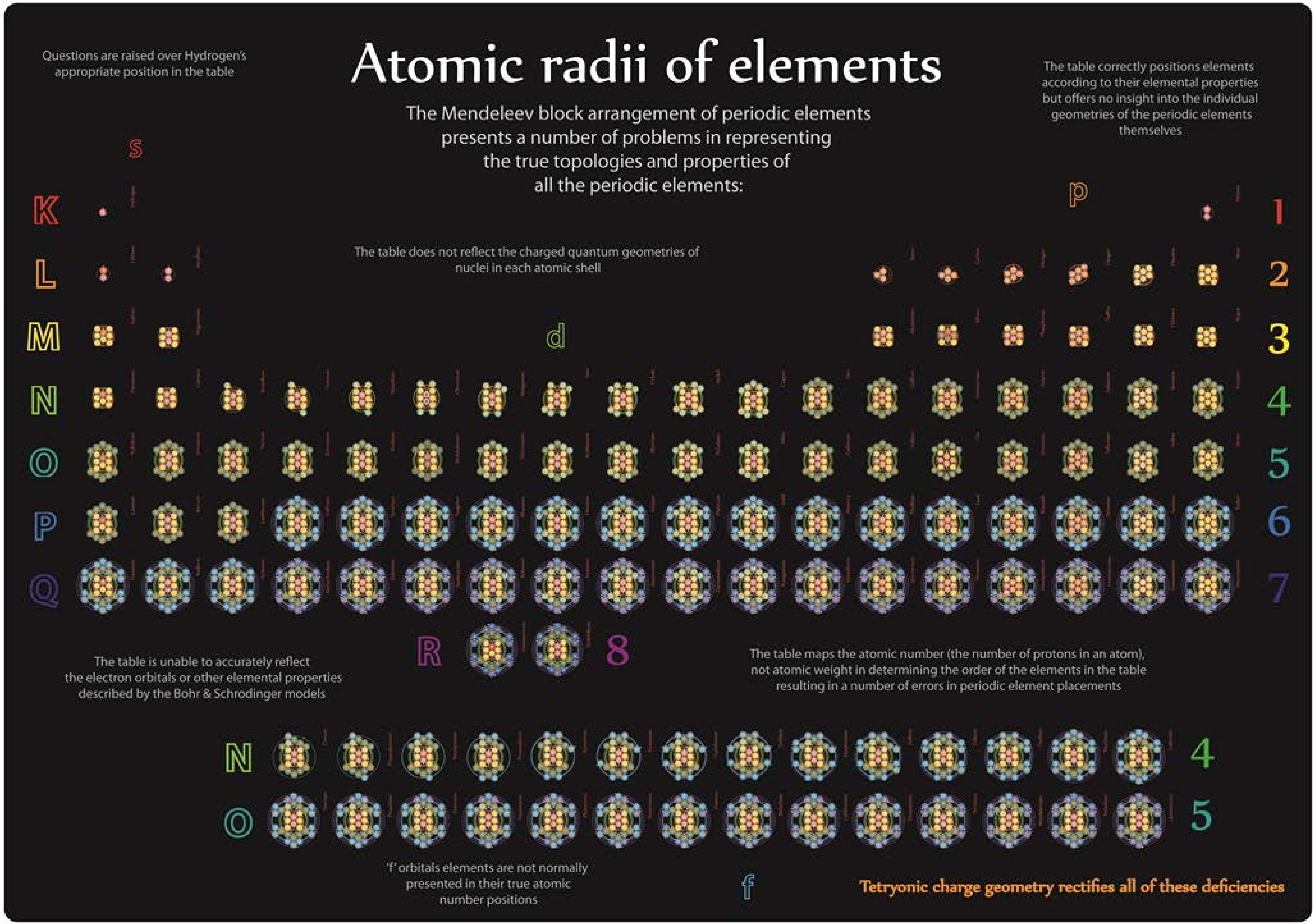

Questions are raised over Hydrogen's appropriate position in the table

The table correctly positions elements according to their elemental properties but offers no insight into the individual geometries of the periodic elements themselves

The table does not reflect the charged quantum geometries of nuclei in each atomic shell

The table is unable to accurately reflect the electron orbitals or other elemental properties described by the Bohr & Schrodinger models

The table maps the atomic number (the number of protons in an atom), not atomic weight in determining the order of the elements in the table resulting in a number of errors in periodic element placements

'f' orbitals elements are not normally presented in their true atomic number positions

Tetryonic charge geometry rectifies all of these deficiencies

Tetryonics 43.07 - Atomic radii of elements

PERIODIC TABLE OF ELEMENTS

modelled on the polar view of the charge topologies of periodic elements

Deuterium [not Hydrogen] is the building block of all elements

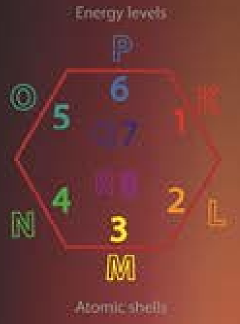
Energy levels
P
6
O
5
Q 7
R 8
1
K
4
2
N
L
3
M
Atomic shells

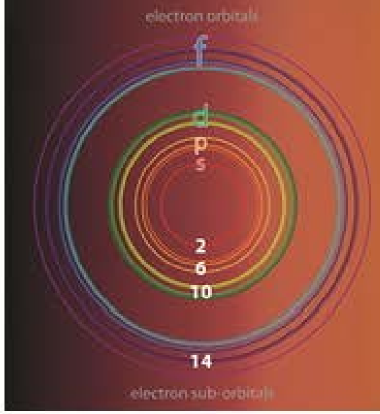
electron orbitals
f
d
p
s
2
6
10
14
electron sub-orbitals

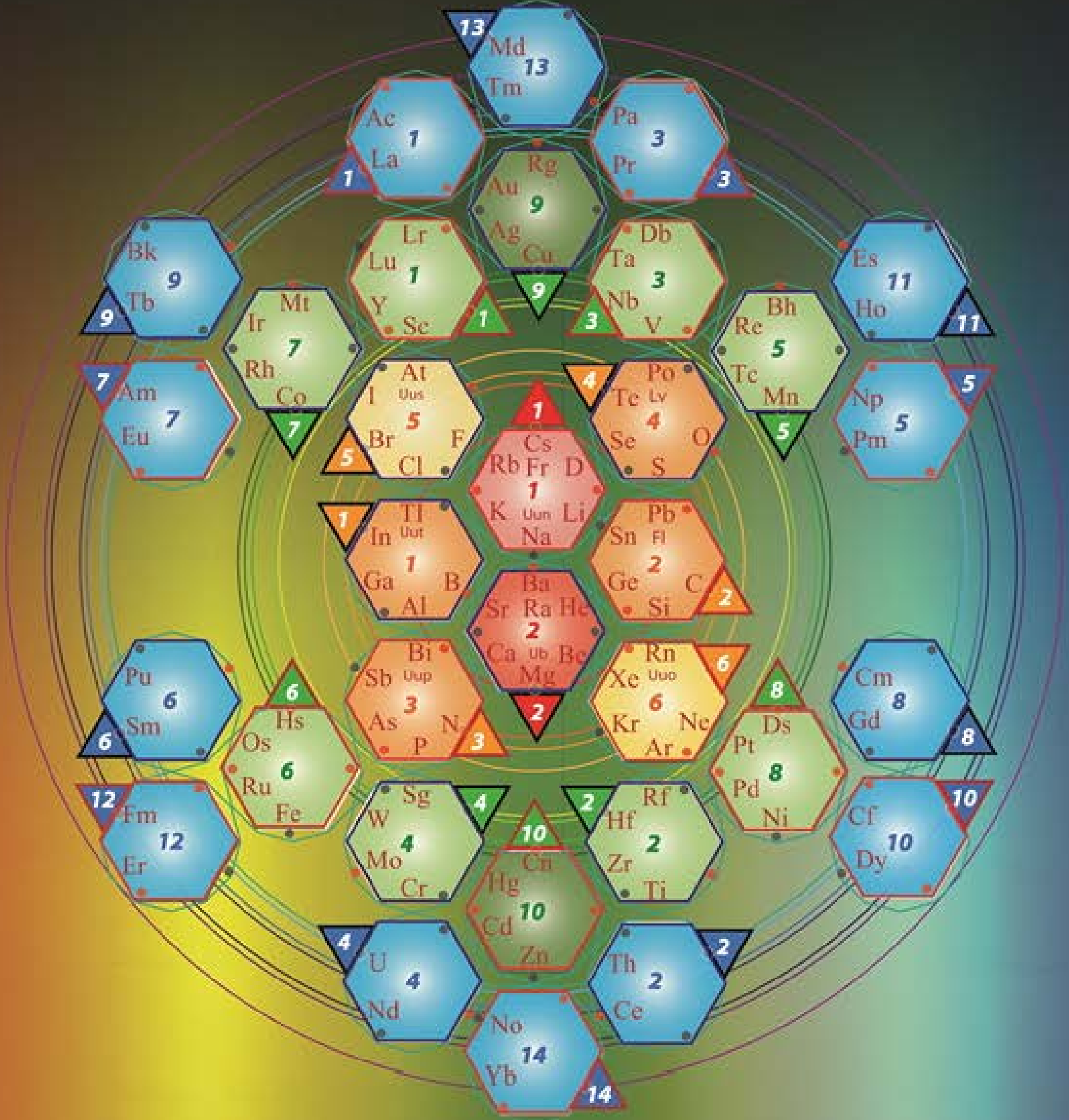
13 Md 13 Tm
Ac 1 La 1
Pa 3 Pr 3
Rg Au 9 Ag Cu 9
Bk 9 Tb 9
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Es 11 Ho 11
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
7 Am 7 Eu
Np 5 Pm 5
At I Uus 5 Br F Cl 5
1 Cs Rb Fr D 1 K Uun Li Na
4 Po Te Lv 4 Se O S
1 Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Pu 6 Sm 6
6 Hs Os 6 Ru Fe
Bi Sb Uup 3 As N P 3
6 Rn Xe Uuo 6 Kr Ne Ar
8 Ds Pt 8 Pd Ni
Cm 8 Gd 8
12 Fm 12 Er
Sg W 4 Mo Cr 4
10 Cn Hg 10 Cd Zn
2 Rf Hf 2 Zr Ti
Cf 10 Dy 10
4 U 4 Nd
No 14 Yb 14
Th 2 Ce 2

Periodic element families

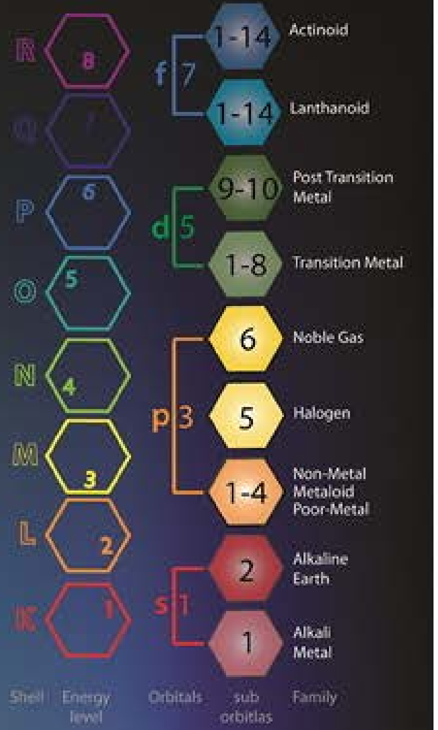
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbitlas
Family

3D quantum charge geometry of periodic element nuclei

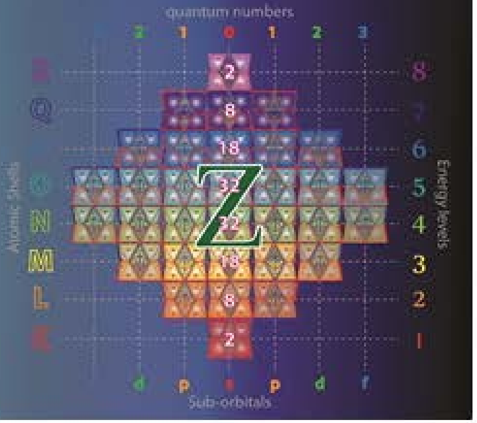
quantum numbers
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
Z
Atomic shells
8
7
6
5
4
3
2
1
Energy levels
d p s p d f
Sub-orbitals

Copyright ABRAHAM [2008] - All rights reserved

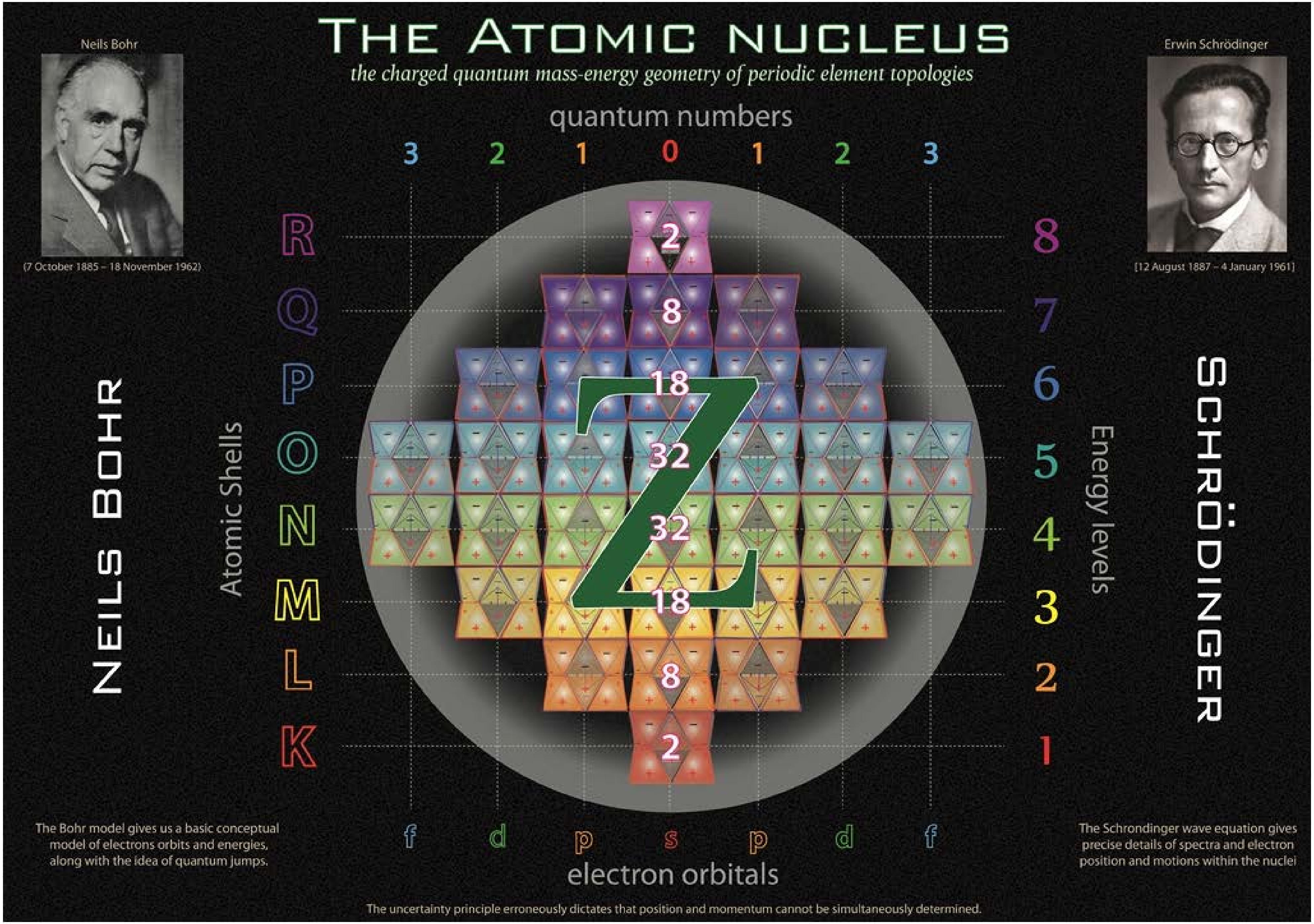

THE ATOMIC NUCLEUS

the charged quantum mass-energy geometry of periodic element topologies

Neils Bohr
(7 October 1885 – 18 November 1962)
Erwin Schrödinger
[12 August 1887 – 4 January 1961]
quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
NEILS BOHR
Atomic Shells
Energy levels
SCHRÖDINGER
f
d
p
s
p
d
f
electron orbitals

The Bohr model gives us a basic conceptual model of electrons orbits and energies, along with the idea of quantum jumps.

The Schrondinger wave equation gives precise details of spectra and electron position and motions within the nuclei

The uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined.

Periodicity of atomic elements

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus. The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the "2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus. The shell letters K, L, M, ... are alphabetical

Each shell can contain only a fixed number of deuterium nuclei [Protons, Neutrons & electrons]

Each shell consists of one or more electron orbitals, and each orbital consists of one or more sub-orbitals.

8 atomic energy shells with electrons

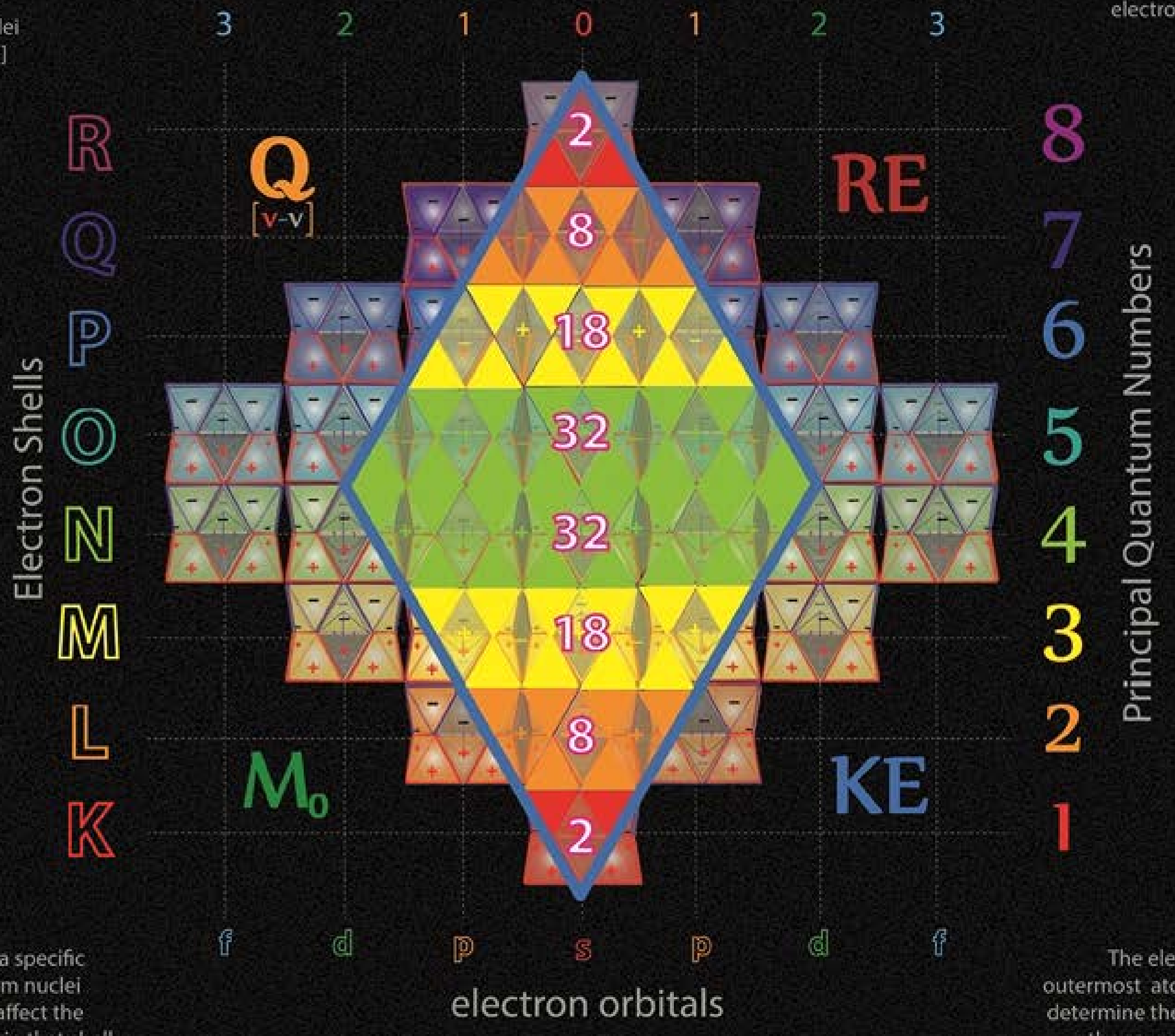

120 periodic elements in total

The number of deuterium nuclei that form each atomic shell in an elemental atom number is double that of Photon energy levels

Each atomic shell equates to a specific energy level for the Deuterium nuclei that comprise it with in turn affect the angular momentum of electrons in that shell

The electrons in the partially filled outermost atomic shell [or highest energy level] determine the chemical properties of the atom; they are called valence shell electrons.

Copyright ABRAHAM [2008] - All rights reserved

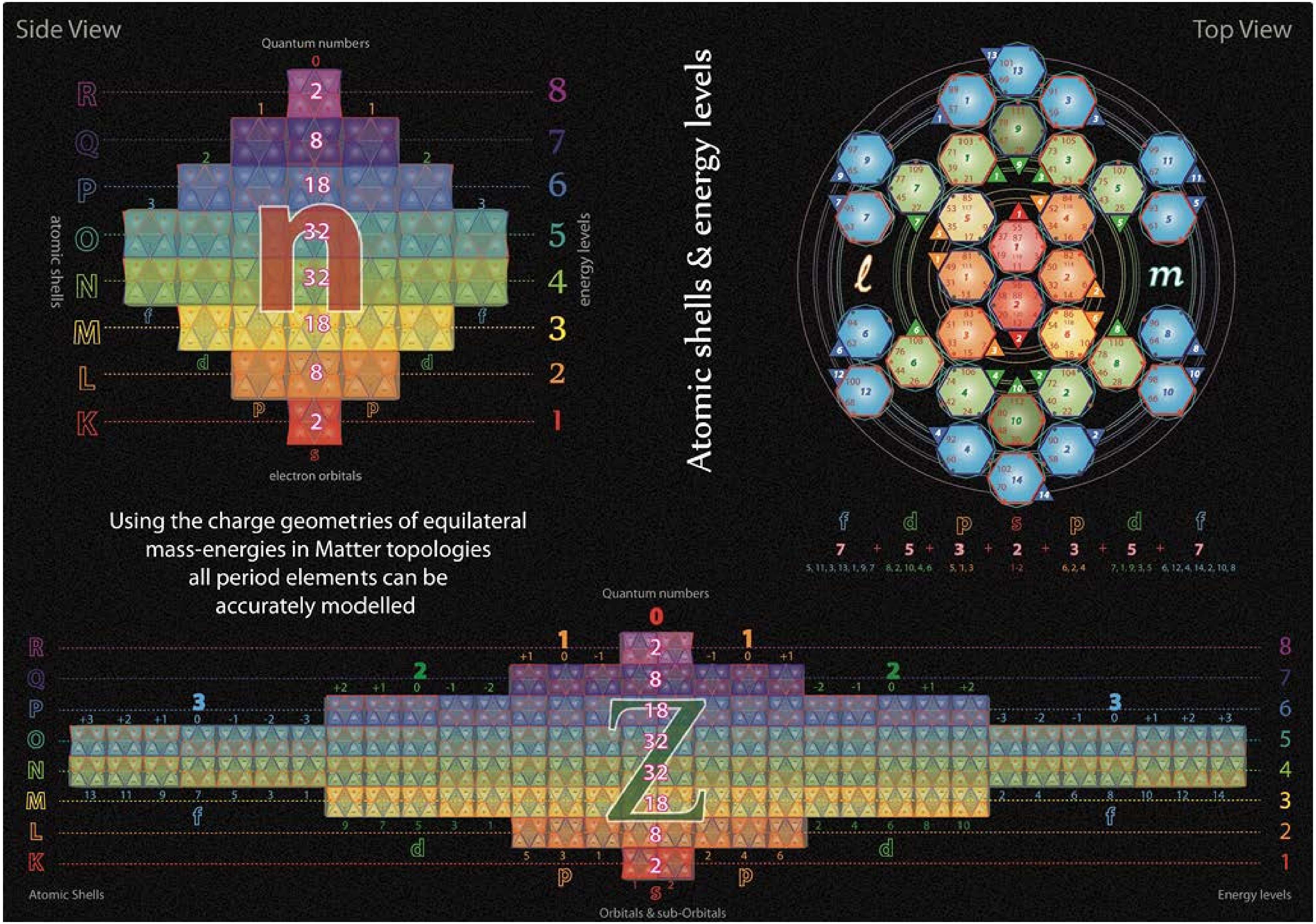

Atomic shells & energy levels

Using the charge geometries of equilateral mass-energies in Matter topologies all period elements can be accurately modelled

Side View
Quantum numbers
0
1
1
2
2
3
3
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
n
8
7
6
5
4
3
2
1
energy levels
f
f
d
d
p
p
s
electron orbitals
Top View
l
m
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7
5, 11, 3, 13, 1, 9, 7
8, 2, 10, 4, 6
5, 1, 3
1-2
6, 2, 4
7, 1, 9, 3, 5
6, 12, 4, 14, 2, 10, 8
Quantum numbers
0
1
+1 0 -1
1
-1 0 +1
2
+2 +1 0 -1 -2
2
-2 -1 0 +1 +2
3
+3 +2 +1 0 -1 -2 -3
3
-3 -2 -1 0 +1 +2 +3
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
f
2 4 6 8 10 12 14
f
9 7 5 3 1
d
2 4 6 8 10
d
5 3 1
p
2 4 6
p
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 44.03 - Shells & energy levels

Copyright ABRAHAM [2008] - All rights reserved

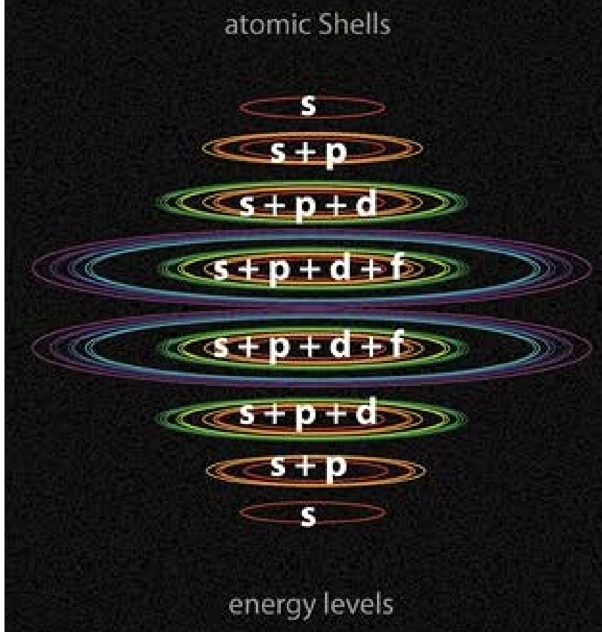
atomic Shells
s
s + p
s + p + d
s + p + d + f
s + p + d + f
s + p + d
s + p
s
energy levels

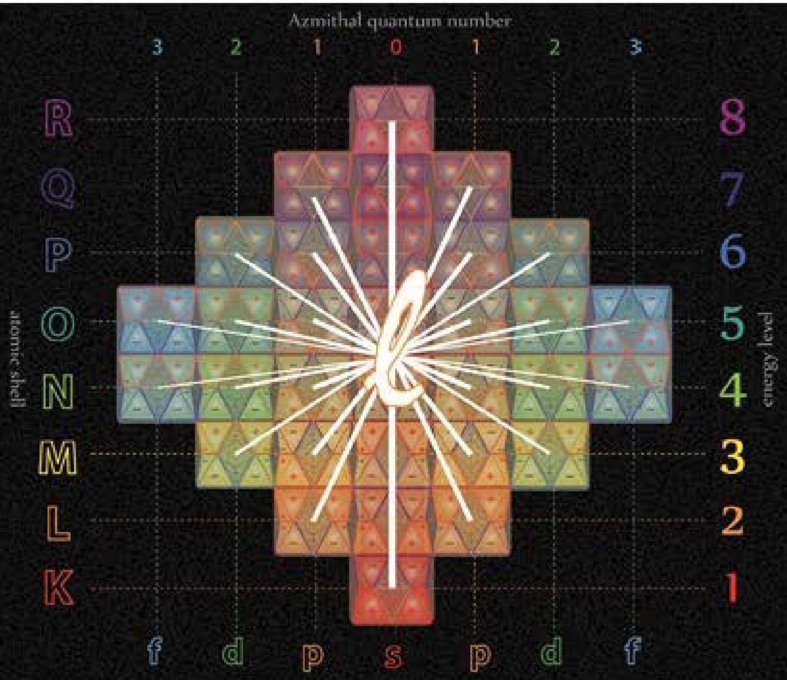
Azmithal quantum number
3 2 1 0 1 2 3
R Q P O N M L K
atomic shell
8 7 6 5 4 3 2 1
energy level
ℓ
f d p s p d f

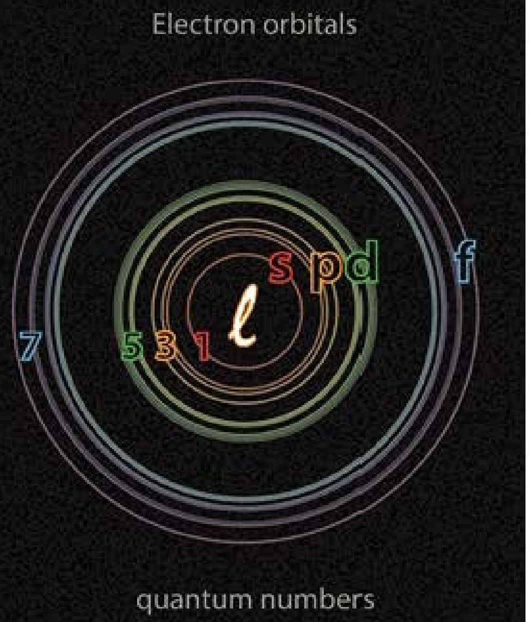
Electron orbitals
s p d f
7 5 3 1
ℓ
quantum numbers

Each electron's location in an elemental nuclei is defined by its quantum numbers

All atoms have shells comprised of concentric s, p, d & f electron orbitals

electron orbitals

each electron orbital can hold a maximum of two spin opposed electrons

a core s-orbital
surrounded by
3 p-orbitals
5 d-orbitals and
7 f-orbitals

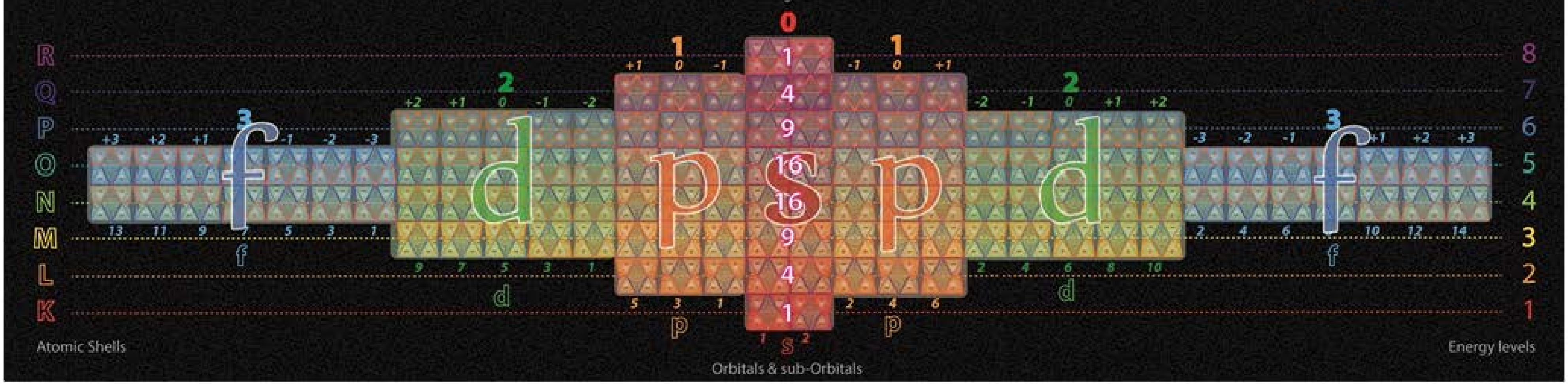
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 -1 -2 -3 -3 -2 -1 +1 +2 +3
R Q P O N M L K
8 7 6 5 4 3 2 1
f d p s p d f
1 4 9 16 16 9 4 1
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
f d p p d f
1 s 2
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Tetryonics 44.04 - Electron Orbitals

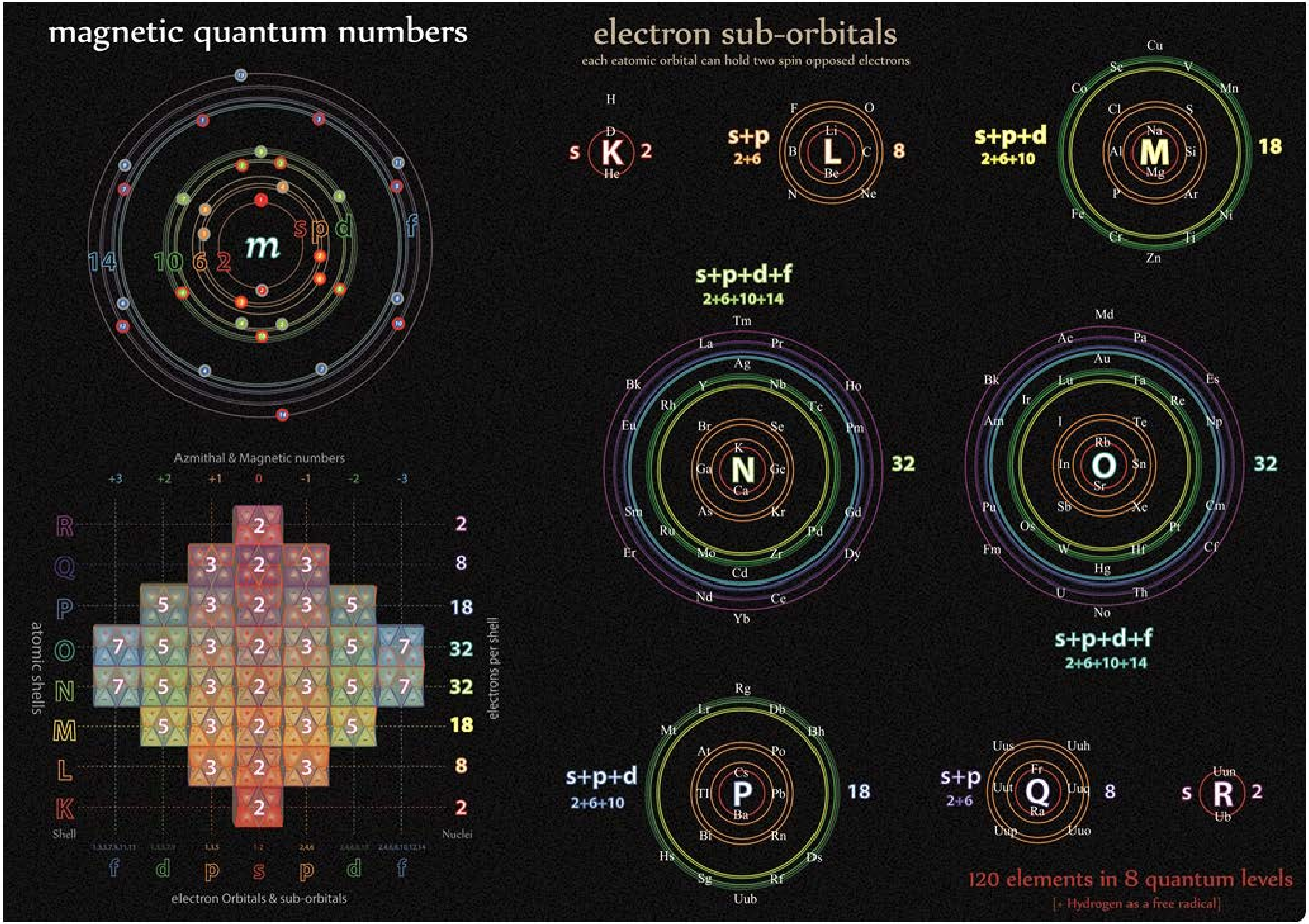
magnetic quantum numbers
m
s p d f
14 10 6 2
Azmithal & Magnetic numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
atomic shells
Shell
2 8 18 32 32 18 8 2
electrons per shell
Nuclei
f d p s p d f
electron Orbitals & sub-orbitals
electron sub-orbitals
each eatomic orbital can hold two spin opposed electrons
s K 2
H D He
s+p 2+6 L 8
s+p+d 2+6+10 M 18
s+p+d+f 2+6+10+14 N 32
s+p+d+f 2+6+10+14 O 32
s+p+d 2+6+10 P 18
s+p 2+6 Q 8
s R 2
120 elements in 8 quantum levels
[+ Hydrogen as a free radical]

Tetryonics 44.05 - Electron sub-Obitals

Copyright ABRAHAM [2008] - All rights reserved

Electron SPIN in atoms

1 sub-Orbit	s	2 electrons
3 sub-Orbits	p	6 electrons
5 sub-Orbits	d	10 electrons
7 sub-Orbits	f	14 electrons

Hund's Rule
Each electron sub-orbital (Azimuthal number)
can not have electrons with the same spin direction

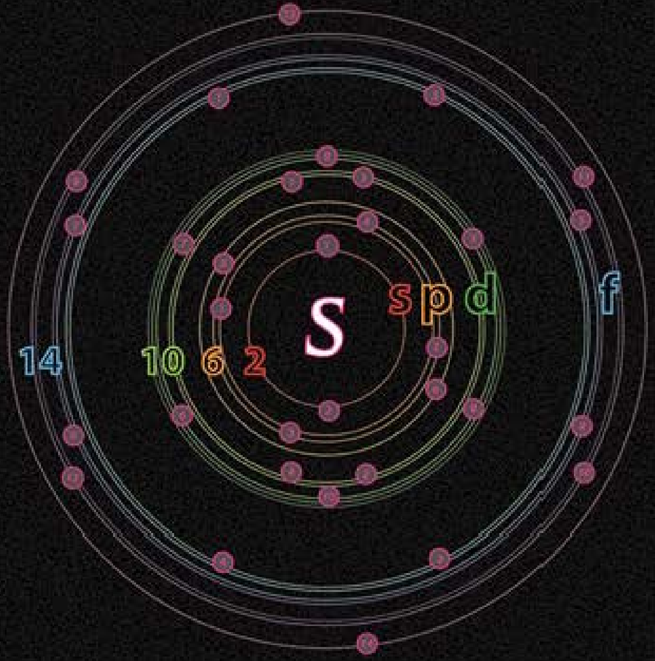
S
s p d f
14 10 6 2

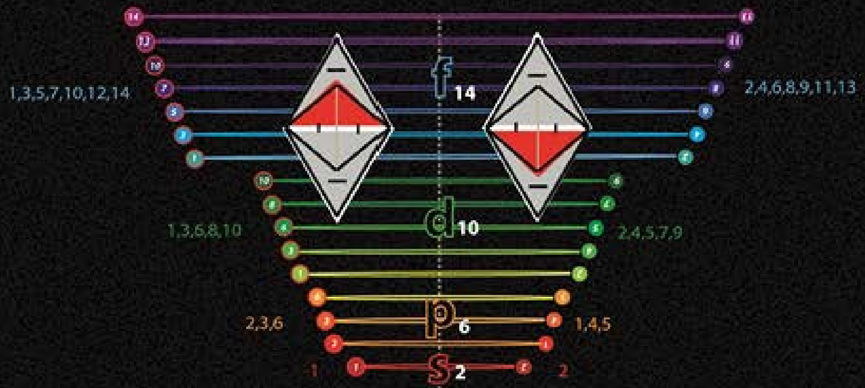
f 14
1,3,5,7,10,12,14
2,4,6,8,9,11,13
d 10
1,3,6,8,10
2,4,5,7,9
p 6
2,3,6
1,4,5
s 2
1
2

electron spin is always referenced to the Nuclear magneton

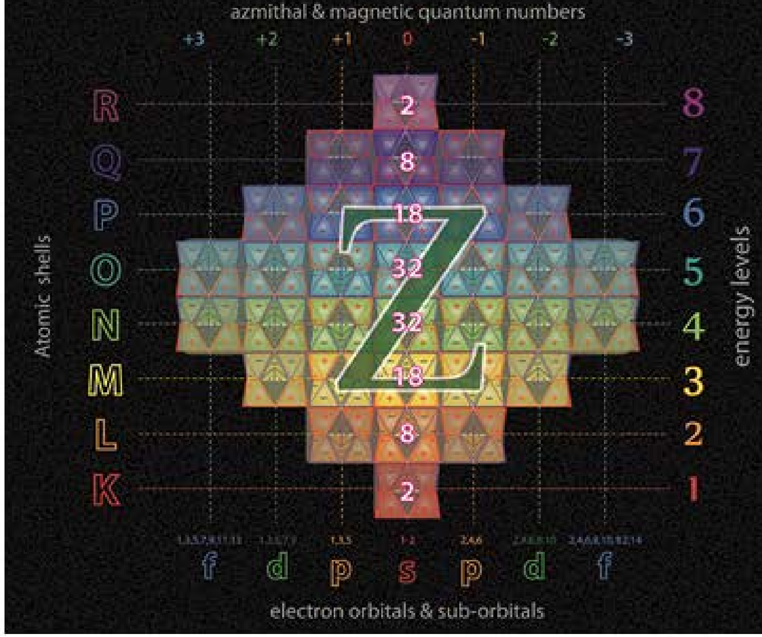
azmithal & magnetic quantum numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
Atomic shells
2 8 18 32 32 18 8 2
Z
8 7 6 5 4 3 2 1
energy levels
1,3,5
1-2
2,4,6
f d p s p d f
electron orbitals & sub-orbitals

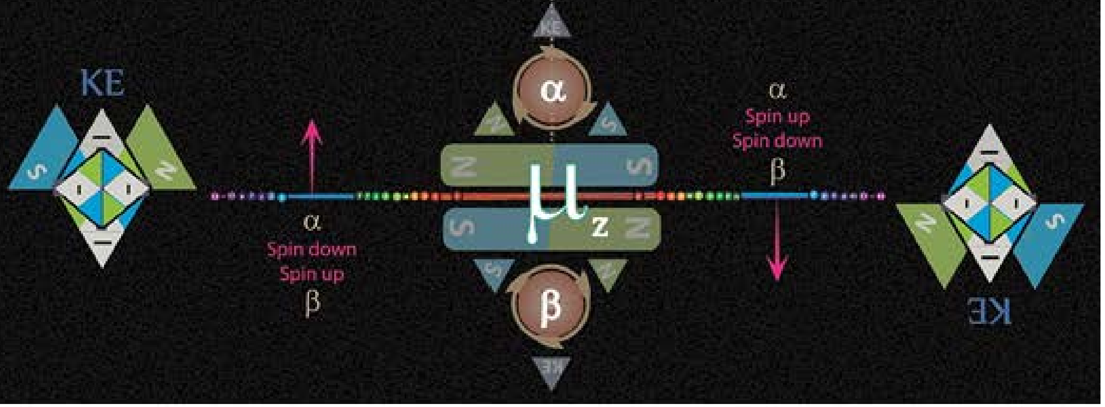
KE
S N
α
Spin down
Spin up
β
KE
α
β
μ z
N S
S N
α
Spin up
Spin down
β
KE
N S

Copyright ABRAHAM [2008] - All rights reserved

Quantum numbers

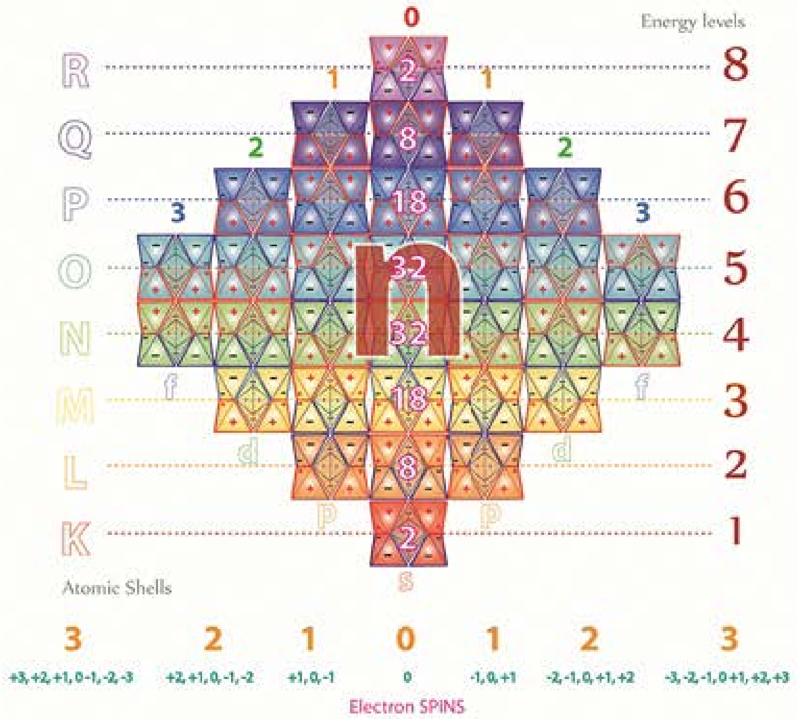
Energy levels
0
R
8
1
1
Q
2
8
2
7
P
3
18
3
6
O
32
n
5
N
32
4
f
f
M
18
3
d
d
L
8
2
p
p
K
2
1
s
Atomic Shells
3
2
1
0
1
2
3
+3, +2, +1, 0 -1, -2, -3
+2, +1, 0, -1, -2
+1, 0, -1
0
-1, 0, +1
-2, -1, 0, +1, +2
-3, -2, -1, 0 +1, +2, +3
Electron SPINS

These four numbers, n, ℓ, m and s can be used to describe any electron in a stable atom [and can be mapped back to the classical terminology of Shells and electron orbitals].

The properties of every atom's specific electron configuration can be described by four quantum numbers:

n (1-8)
Principal
(n = 1, 2, 3, 4 ...)

ℓ (0-3)
Azimuthal
(l = 0, 1 ... n−1)

m (2ℓ+1)
Magnetic
(ml = −l, −l+1 ... 0 ... l−1, l)

s ± ½
electron Spin
(ms = −1/2 or +1/2),

The Bohr model was a one-dimensional model that used one quantum number to describe the distribution of electrons

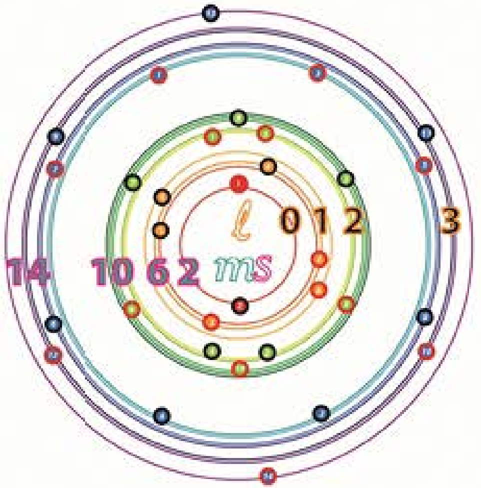
ℓ
0
1
2
3
14
10
6
2
ms

The three coordinates that come from Schrödinger's wave equations are the principal (n), angular (l), and magnetic (m) quantum numbers.

These quantum numbers describe the size, shape, and orientation in space of the orbitals of any particular atom mathematically.

Each electron's quantum numbers are unique and cannot be shared by another electron in that atom.

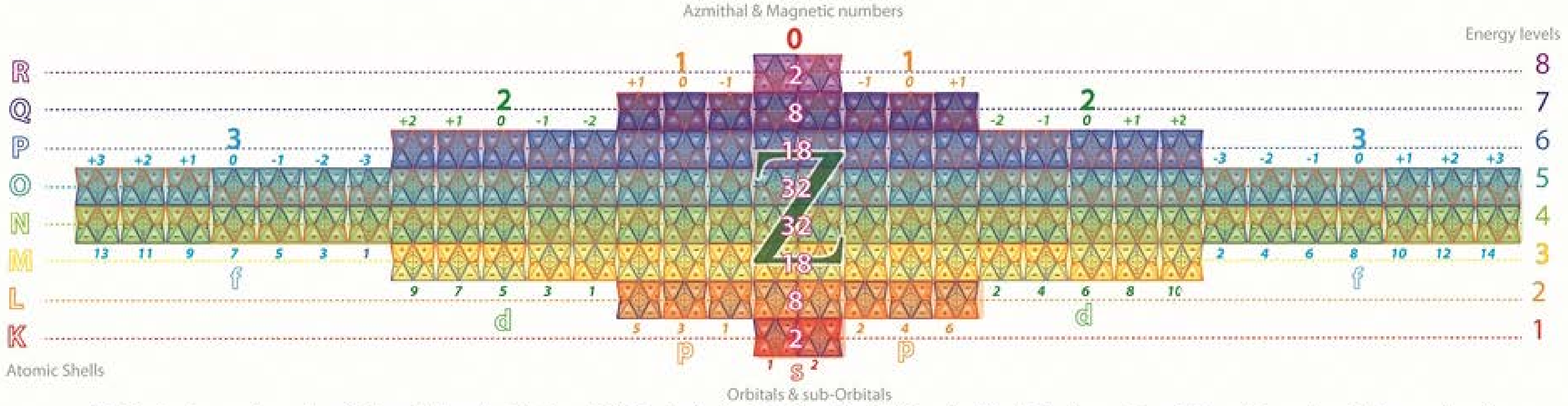
Azmithal & Magnetic numbers
Energy levels
0
1
+1
0
-1
1
-1
0
+1
2
+2
+1
0
-1
-2
2
-2
-1
0
+1
+2
3
+3
+2
+1
0
-1
-2
-3
3
-3
-2
-1
0
+1
+2
+3
R
8
Q
7
P
6
O
5
N
4
M
3
L
2
K
1
2
8
18
32
Z
32
18
8
2
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
f
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Orbitals & sub-Orbitals

A stable atom has equal numebrs of Protons, electrons [and Neutrons], all following the Pauli Exclusion Principle thus orientating their spins so that each element has a unique electron configuration

Copyright ABRAHAM [2008] - All rights reserved

PRINCIPAL quantum number

The first describes the electron shell, or energy level, of an atom.

This is the only quantum number
introduced by the Bohr model

n(1-8)

(n = 1, 2, 3, 4 ...)

The principal quantum number
can only have positive integer values

atomic shells

energy levels

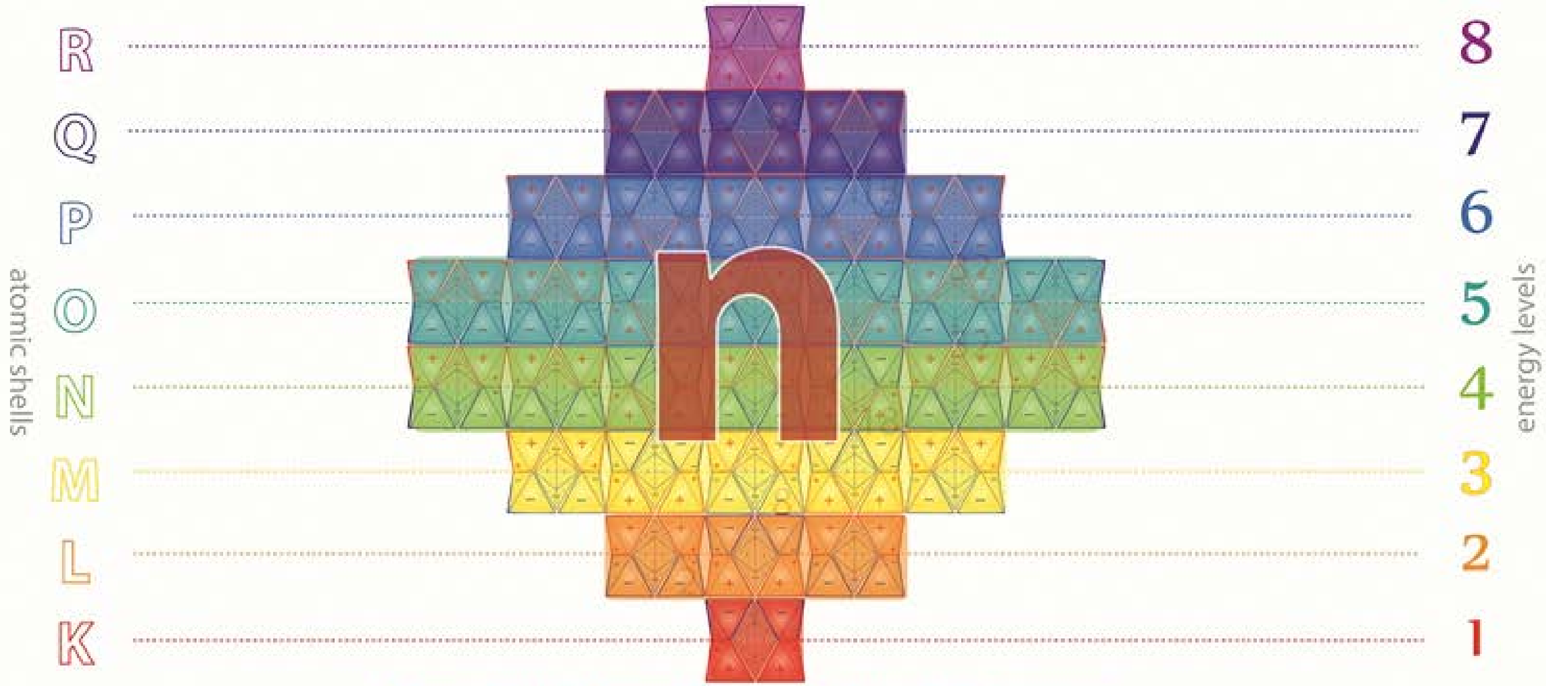

As energies of the Baryons comprising the atomic nuclei increases,
the electron bound to each nuclei also possesses more KEM field energies
and is therefore less tightly bound to the nucleus

Atomic Shells

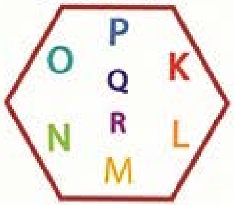

(K, L, M, N, O, P, Q, R)

Atomic shells relate directly to Principal quantum numbers

(1, 2, 3, 4, 5, 6, 7, 8)

Energy levels

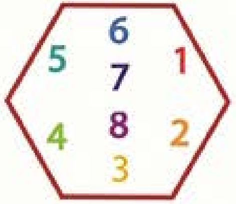

AZMITHAL quantum number

The azimuthal quantum number is a quantum number assigned to any atomic orbital that describes its orbital angular momentum and determines the shape of the electron orbital

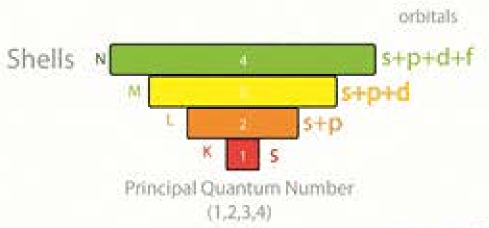
orbitals
Shells
N
4
s+p+d+f
M
s+p+d
L
2
s+p
K
1
s
Principal Quantum Number
(1,2,3,4)

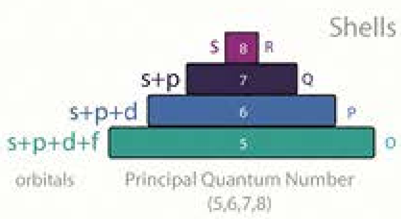
Shells
s
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron orbitals

(l = 0, 1 ... n−1)

ℓ (0-7)

Azimuthal

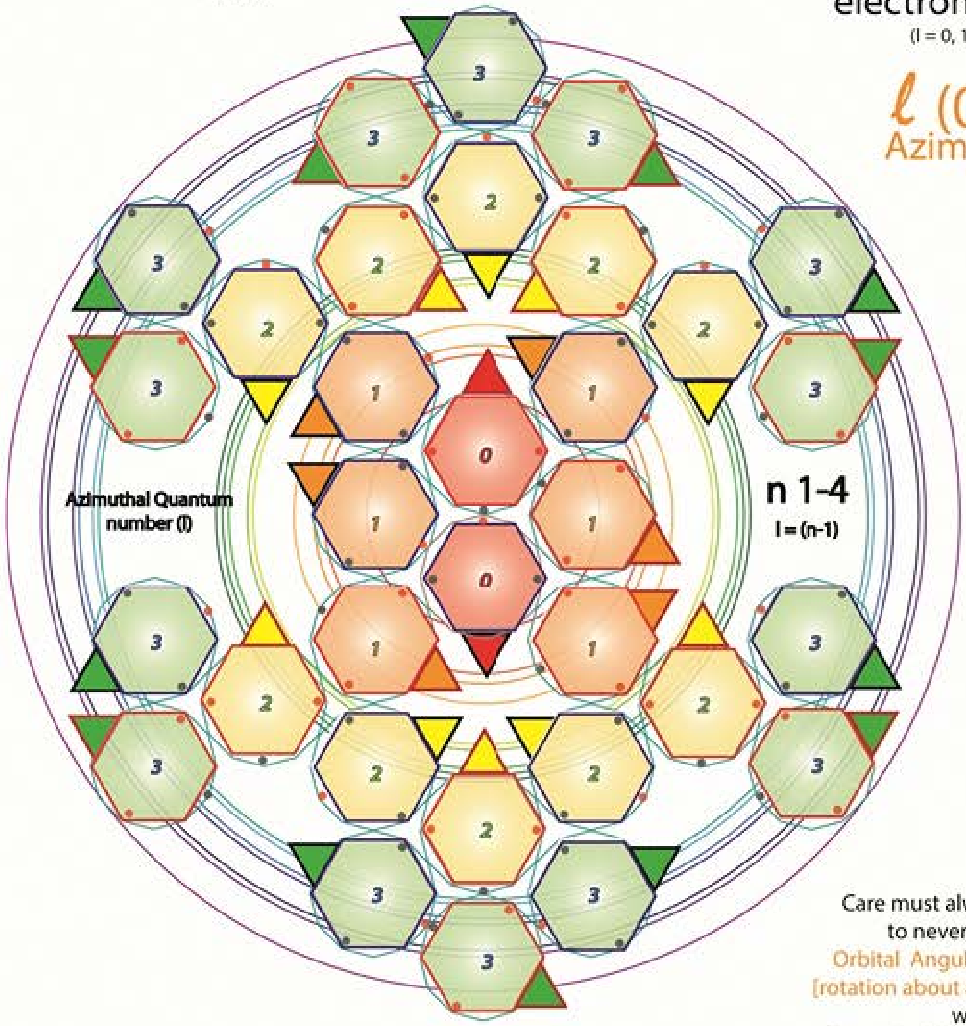
Azimuthal Quantum number (l)
n 1-4
l = (n-1)

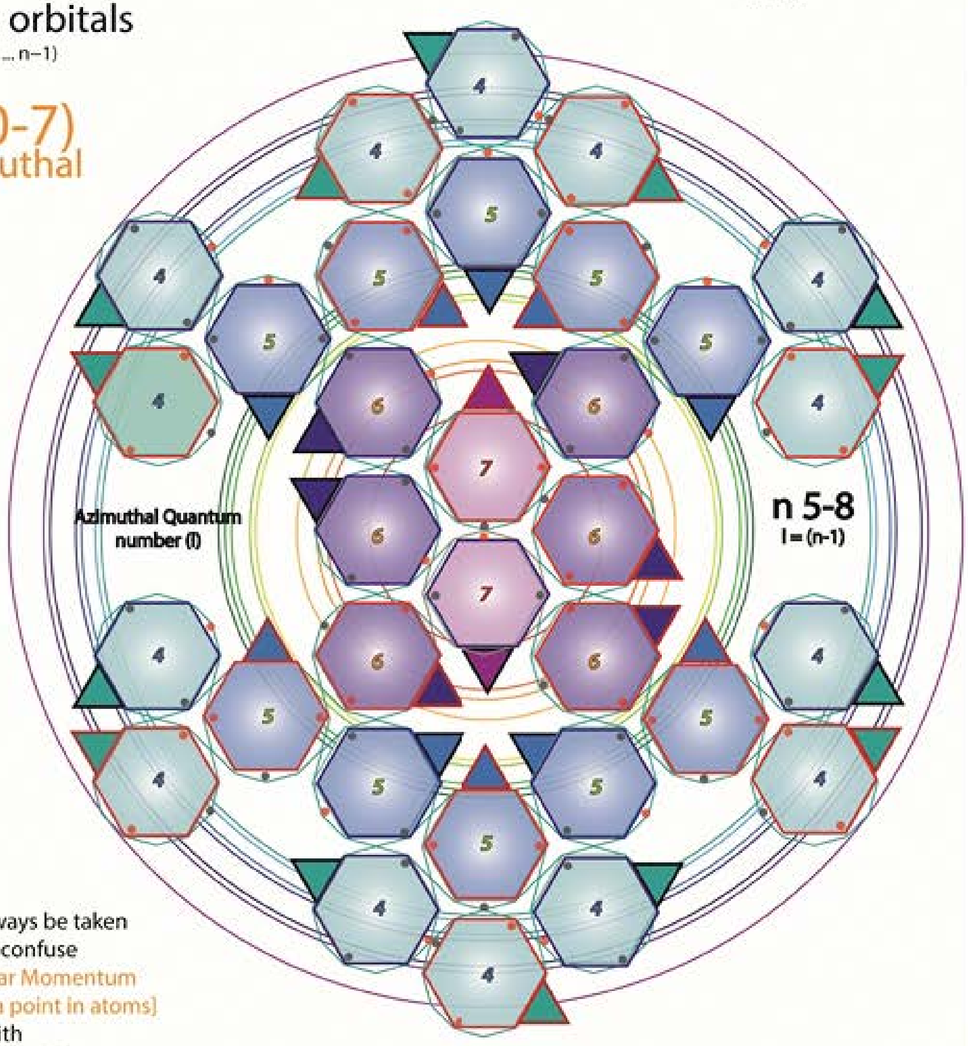
Azimuthal Quantum number (l)
n 5-8
l = (n-1)

Care must always be taken
to never confuse
Orbital Angular Momentum
[rotation about a point in atoms]
with
Quantised Angular Momenta
[equilateral Planck energy geometries]

MAGNETIC quantum number

The magnetic quantum number denotes the energy levels available within any subshell

Magnetic numbers do not continue to increase as the Principal numbers increase
instead they reverse after n4 to reflect the charged quanutm geometry of Elements
and do not follow the current computer models in popular use

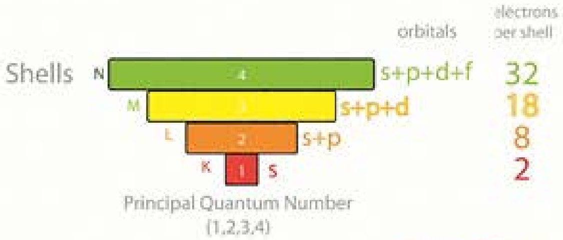
Shells
orbitals
electrons per shell
N
M
L
K
4
3
2
1
s+p+d+f
s+p+d
s+p
s
32
18
8
2
Principal Quantum Number
(1,2,3,4)

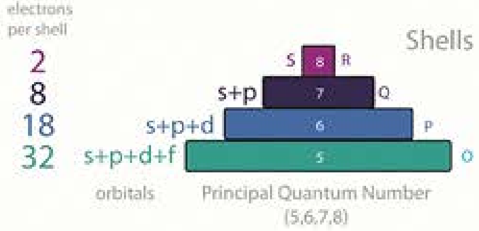
electrons per shell
Shells
2
8
18
32
S
R
Q
P
O
8
7
6
5
s+p
s+p+d
s+p+d+f
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron sub-orbitals

$(m_l = -l, -l+1 \ldots 0 \ldots l-1, l)$

$m\ell(2\ell+1)$

Magnetic

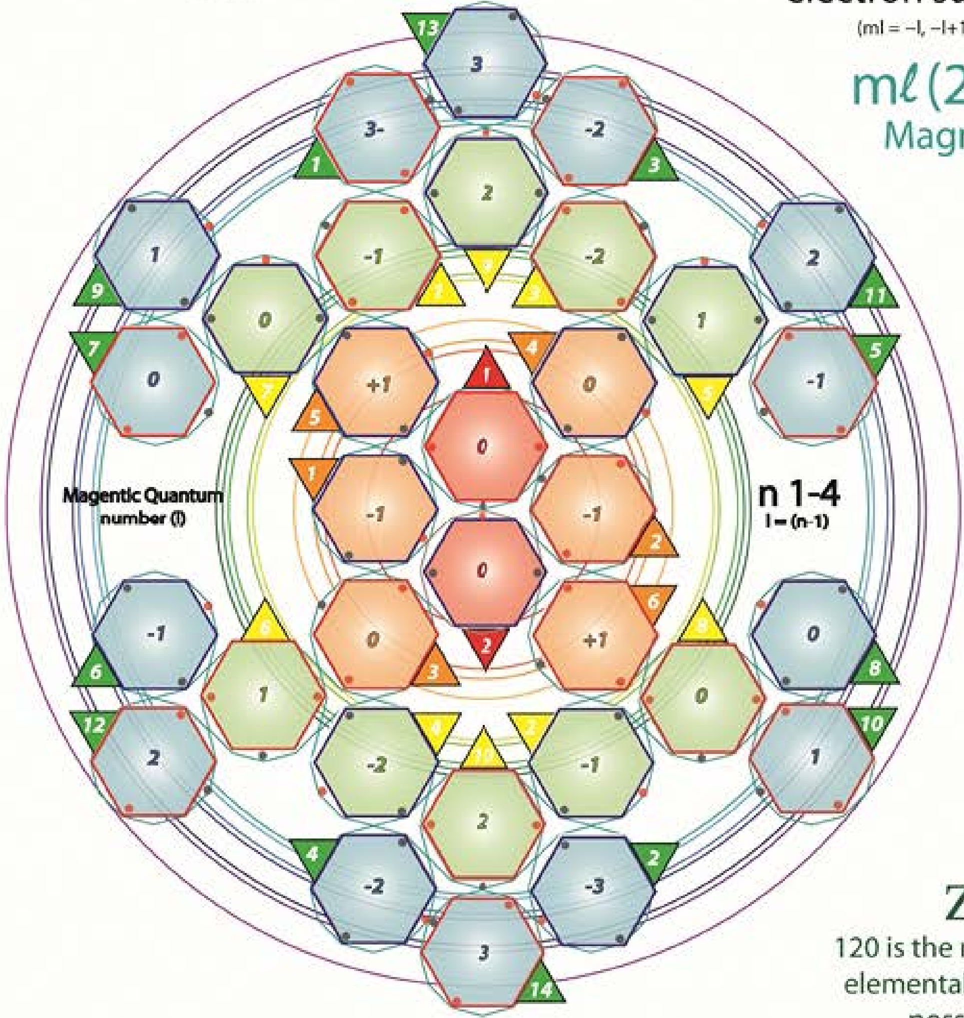
Magentic Quantum number (l)
n 1-4
l = (n-1)

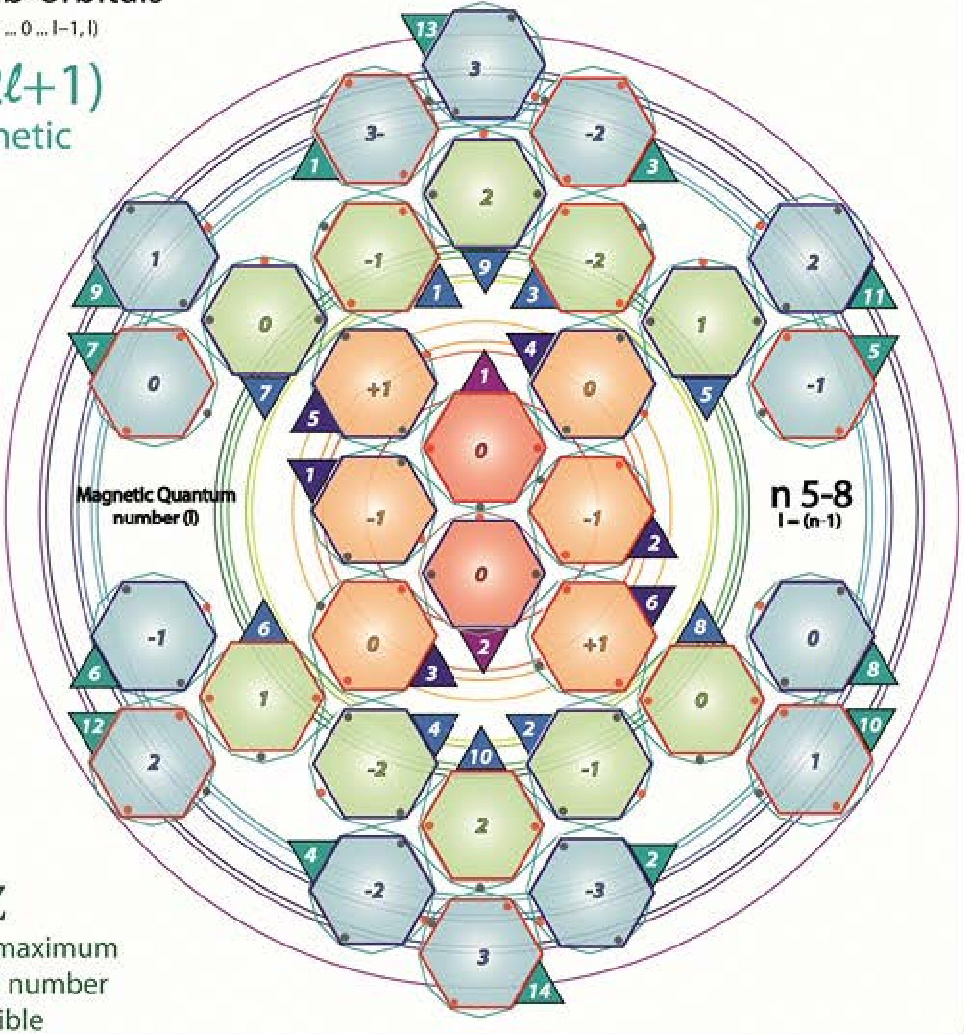
Magnetic Quantum number (l)
n 5-8
l = (n-1)

Z

120 is the maximum
elemental number
possible

Copyright ABRAHAM [2008] - All rights reserved

SPIN quantum number

The spin quantum number is a quantum number that parameterizes the intrinsic angular momentum (or spin angular momentum), of any given electron anywhere in an atomic nucleus

parallel magnetic moments

Spin UP

higher coupling energies

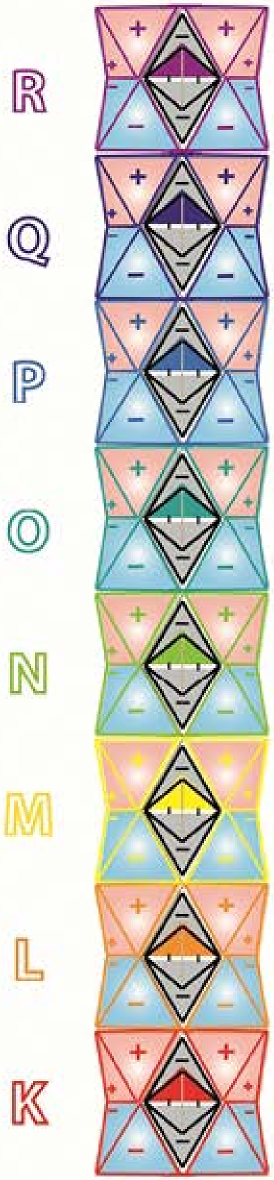

anti-parallel magnetic moments

Spin DOWN

lower coupling energies

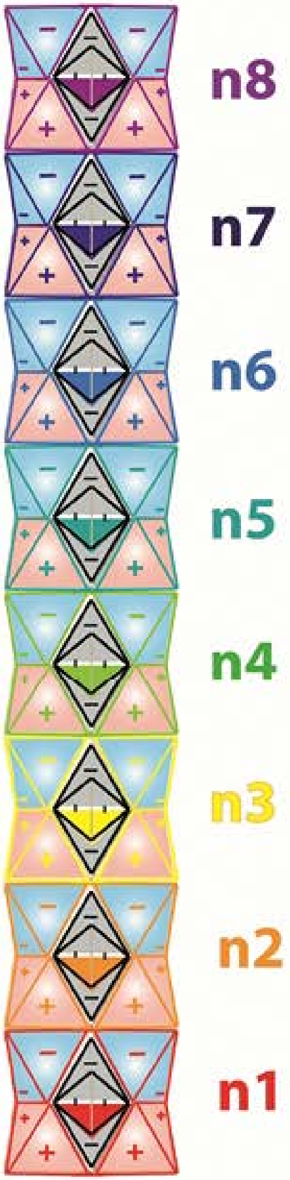
n8
n7
n6
n5
n4
n3
n2
n1
+
−

Electron spin can orientate in either direction within Nuclei, providing the nett spins follow the Hund rule and Pauli exclusion principle

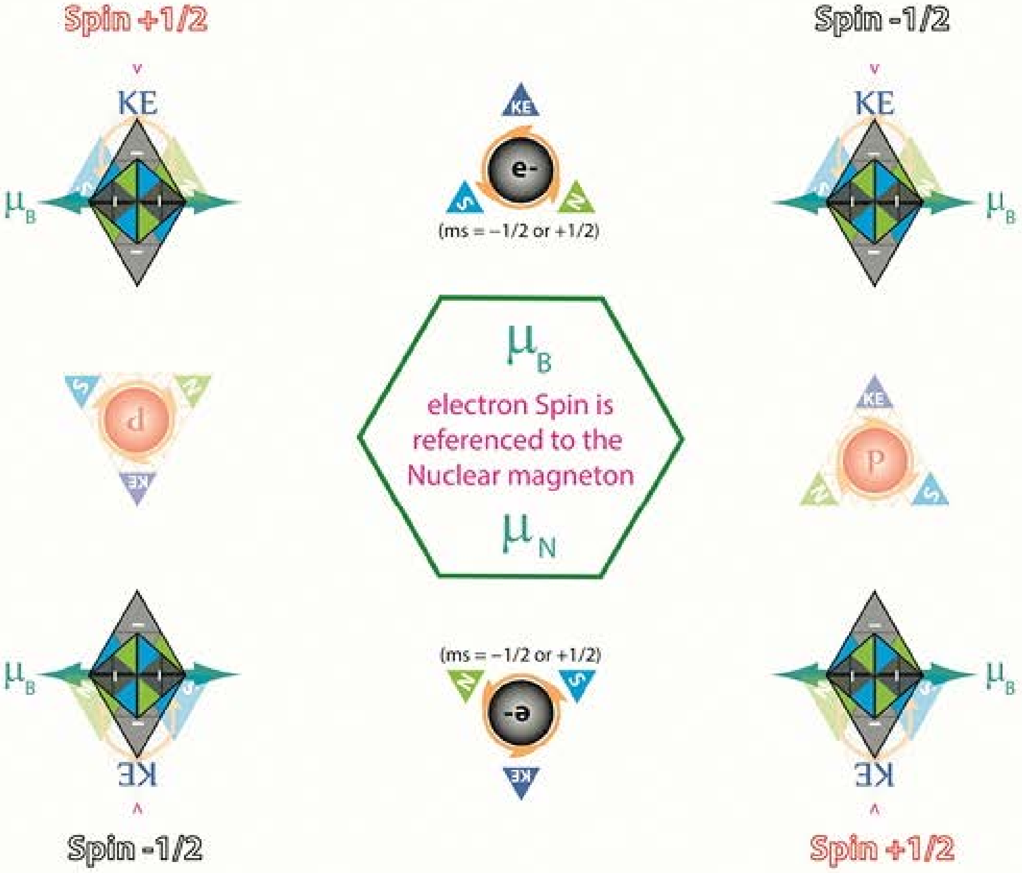
Spin +1/2
Spin -1/2
KE
μ_B
e-
(ms = −1/2 or +1/2)
d
p
μ_B
electron Spin is referenced to the Nuclear magneton
μ_N
Spin -1/2
Spin +1/2

The nuclear energy levels of the Baryons comprising Elemental nuclei determine the energy-momenta of electrons bound to them

Modifying Hund's Rule

Electrons fill orbitals in an alternating sequential numbering pattern due to nucleon placement creating opposed direction electron spins

The increased stability of the atom, most commonly manifested in a lower energy state, arises because the high-spin state forces the unpaired electrons to reside in different spatial orbitals.

A commonly given reason for the increased stability of high multiplicity states is that the different occupied spatial orbitals create a larger average distance between electrons, reducing electron-electron repulsion energy. In reality, it has been shown that the actual reason behind the increased stability is a decrease in the screening of electron-nuclear attractions[1].

The total spin state is calculated as the total number of unpaired electrons + 1, or twice the total spin + 1 written as 2s+1.

As a result of Hund's rule, constraints are placed on the way atomic orbitals are filled using the Aufbau principle.

Before any two electrons occupy an orbital in a subshell, other orbitals in the same subshell must first each contain one electron. Also, the electrons filling a subshell will have parallel spin before the shell starts filling up with the opposite spin electrons (after the first orbital gains a second electron).

As a result, when filling up atomic orbitals, the maximum number of unpaired electrons (and hence maximum total spin state) is assured

s	p	d	f
▲1 1 2▼	▼1 1 2▲	▲1 1 2▼	▲ 1 1 2 ▼
	▲3 2 4▼	▲3 2 4▼	▲ 3 2 4 ▼
	▼5 3 6▲	▼5 3 6▲	▲ 5 3 6 ▼
		▼7 4 8▲	▲ 7 4 8 ▼
		▼9 5 10▲	▼ 9 5 10 ▲
			▼ 11 6 12 ▲
			▼ 13 7 14 ▲

▲

Hund's rule of orbital filling must be modified to reflect the true orbital filling order

▼

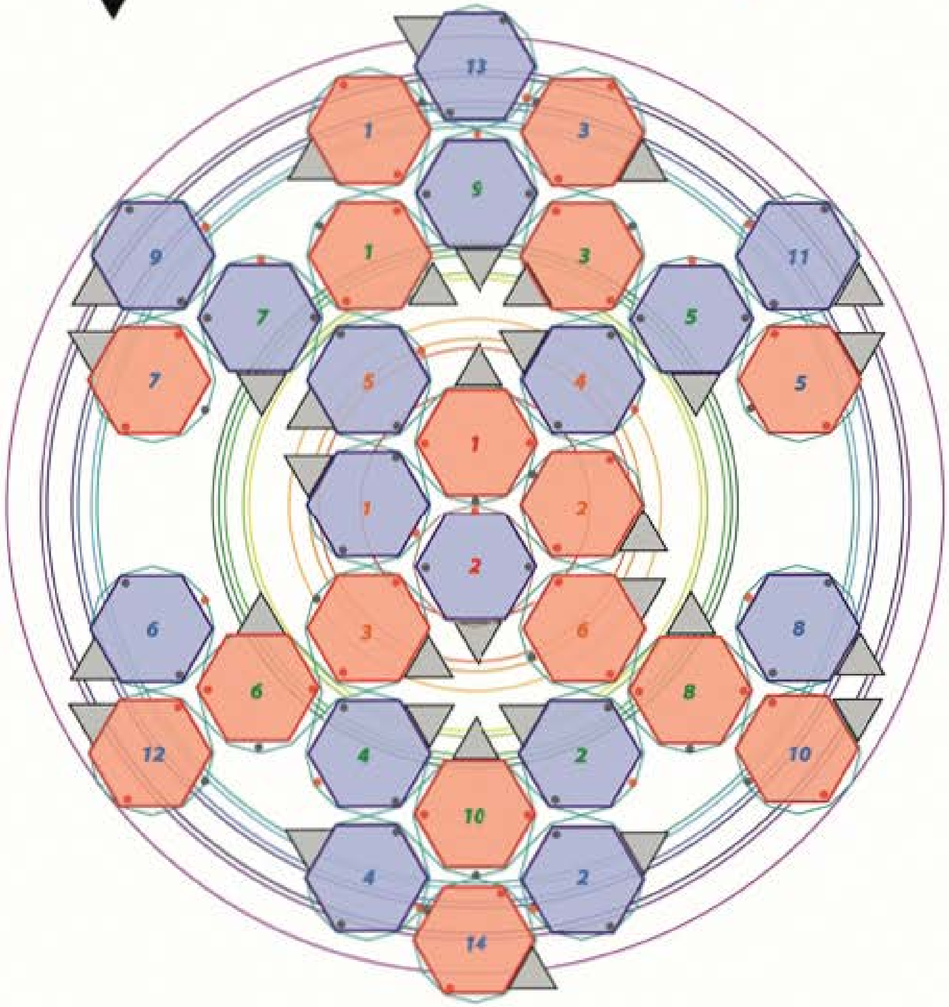

Sub-orbitals fill in order of numbering

Electrons spins pair
before next orbital
is filled

ie.
p1 [DOWN] and p2 [UP] fill
before
p3 [UP] and p4 [DOWN]
before
p5 [DOWN] and p6 [UP] etc

Copyright ABRAHAM [2008] - All rights reserved

Principle quantum Energies

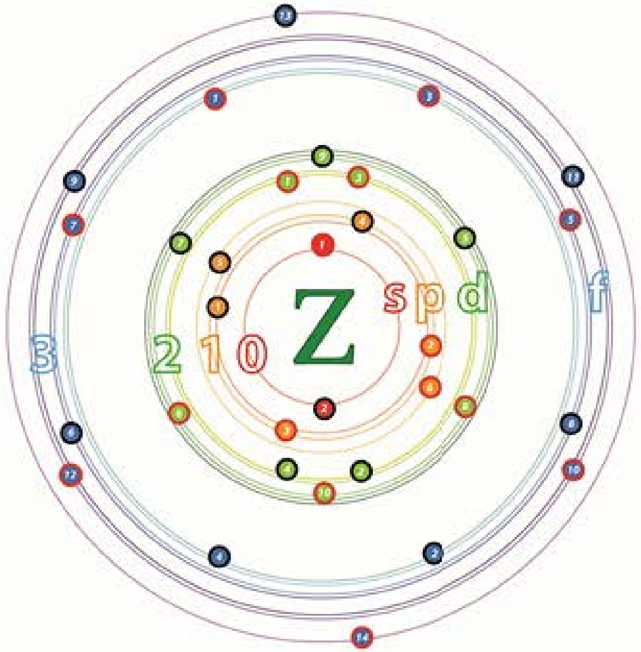

In an atom - electron energies are proportional to their intrinsic Kinetic Energies - which in turn are directly proportional to the quantum energy level of the nuclei which the electron binds to in their respective atomic shells

In a nucleus, lower energy orbits have less 'paired' nuclei supplying energy. The more energy you give a nuclei the faster it casuses the bound electron to rotate. If you give the nuclei enough energy, it will impart eneough energy to its electron for it to leave the system entirely.

The same is true for an electron orbital. Higher values of n mean more energy for the electron and the corresponding KEM field enegies of the electron is larger, resulting in increased angular momentum.

Values of n start at 1 and go up by integer amounts.

If enough energy is added to the system by incident Photons a electron will leave the atom creating a positively charged nuclei [ionisation].

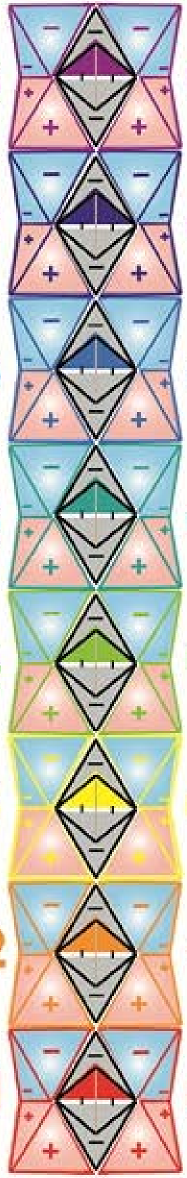

n8 En = -0.211 eV
n7 En = -0.276 eV
n6 En = -0.375 eV
n5 En = -0.541 eV
n4 En = -0.845 eV
n3 En = -1.502 eV
n2 En = -3.381 eV
n1 En = -13.525 eV

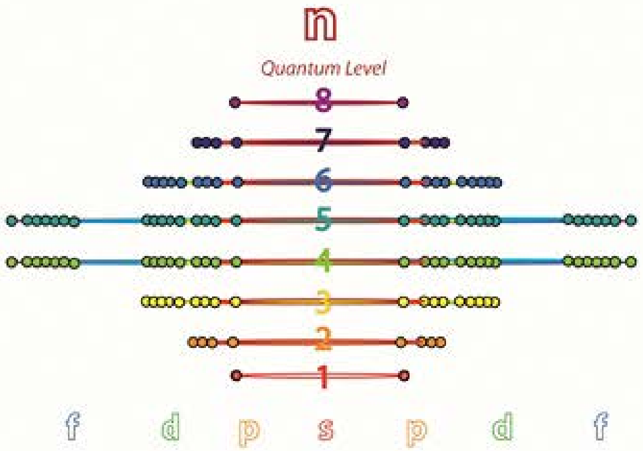

s orbits	n1-8
p orbits	n2-7
d orbits	n3-6
f orbits	n4-5

Eigenstate value

KEM field energy [per n] required to exceed 13.525 eV at which point the photo-electron has sufficient KE to break free of the Nucleus

$$E_n = \frac{E_1}{n^2} = \frac{-13.6eV}{n^2}, n = 1, 2, 3...$$

The possible Kinetic Energies (quantum levels) of an electron are directly related to the energy level of the Nuclei in each Quantum Level

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 1

K shell

n1

Ground State electron

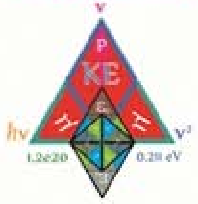
v
P
KE
hv
v^2
1.2e20
0.28 eV

$iE = -13.313\ eV$

The energy levels of bound electrons is determined by Baryons

Z #	Name		Protons	electrons	Neutrons
1	1s1	Deuterium	1	1	1
2	1s2	Helium	2	2	2

2

Deuterium [not Hydrogen] is the building block of elements

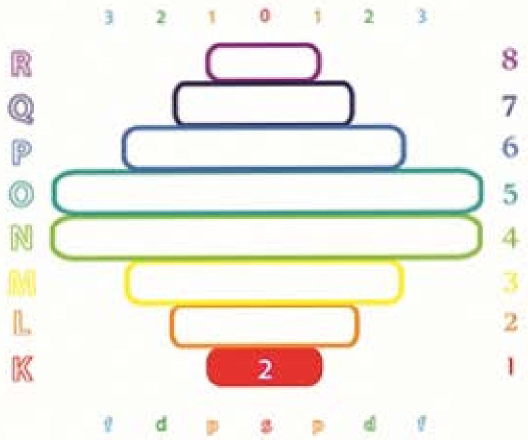
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
f
d
p
s
p
d
f

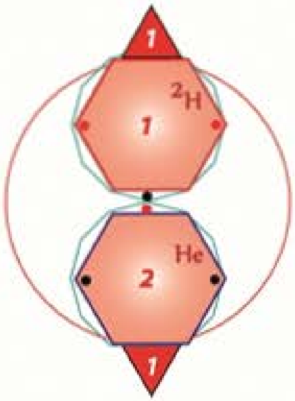

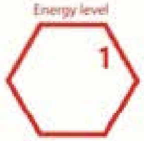
Energy level
1

Z #	Name		Protons	electrons	Neutrons
1	1s1	Hydrogen	1	1	0

Hydrogen is a free radical element

n1

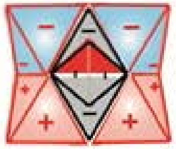

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 2

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
3	2s1	Lithium	3	3	3
4	2s2	Beryllium	4	4	4
5	1p1	Boron	5	5	5
6	1p2	Carbon	6	6	6
7	1p3	Nitrogen	7	7	7
8	1p4	Oxygen	8	8	8
9	1p5	Fluorine	9	9	9
10	1p6	Neon	10	10	10

8

Deuterium [not Hydrogen] is the building block of elements

L shell

n2

Ground State electron

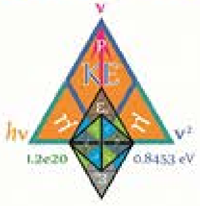
v
KE
hv
v²
1.2e20
0.8453 eV

$iE = -12.679\ eV$

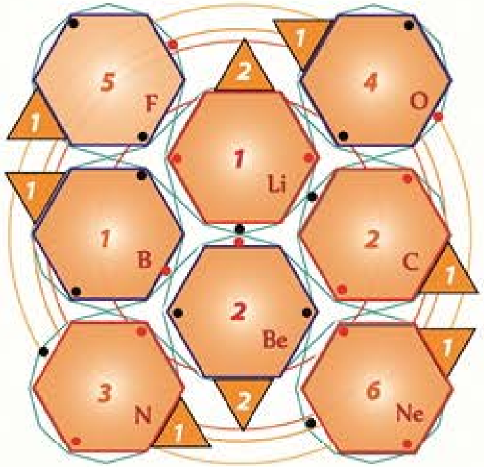
5
F
2
1
4
O
1
1
Li
1
1
B
2
C
1
2
Be
3
N
2
1
1
6
Ne

Energy level

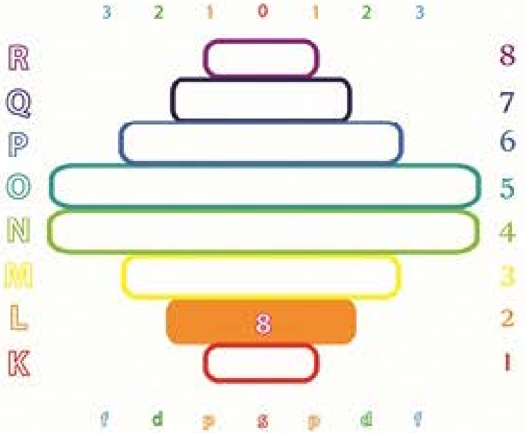
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
f
d
p
s
p
d
f

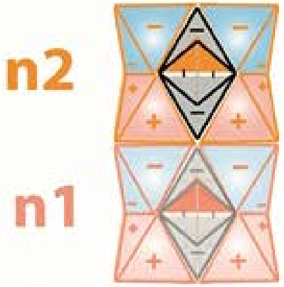

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 3

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
11	3s1	Sodium	11	11	11
12	3s2	Magnesium	12	12	12
13	2p1	Aluminium	13	13	13
14	2p2	Silicon	14	14	14
15	2p3	Phosphorus	15	15	15
16	2p4	Sulfur	16	16	16
17	2p5	Chlorine	17	17	17
18	2p6	Argon	18	18	18
21	3d1	Scandium	21	21	21
22	3d2	Titanium	22	22	22
23	3d3	Vanadium	23	23	23
24	3d4	Chromium	24	24	24
25	3d5	Manganese	25	25	25
26	3d6	Iron	26	26	26
27	3d7	Cobalt	27	27	27
28	3d8	Nickel	28	28	28
29	3d9	Copper	29	29	29
30	3d10	Zinc	30	30	30

18

Deuterium [not Hydrogen] is the building block of elements

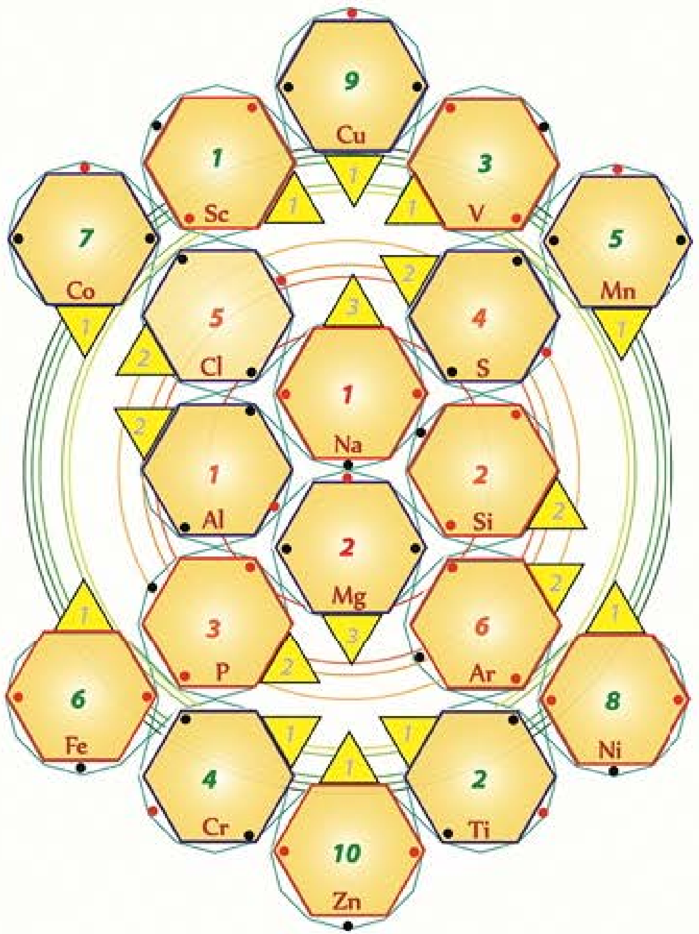
9
Cu
1
Sc
3
V
7
Co
5
Mn
5
Cl
4
S
1
Na
1
Al
2
Si
2
Mg
3
P
6
Ar
6
Fe
8
Ni
4
Cr
2
Ti
10
Zn
1
2
3

M shell

n3

Ground State electron

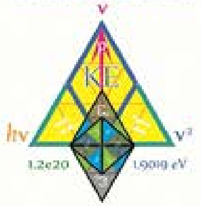
v
KE
hv
v²
1.2e20
1.9019 eV

$iE = -11.623\ eV$

Energy level

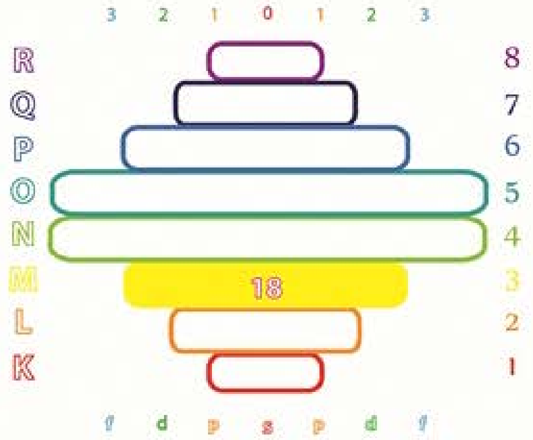
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
18
f
d
p
s
p
d
f

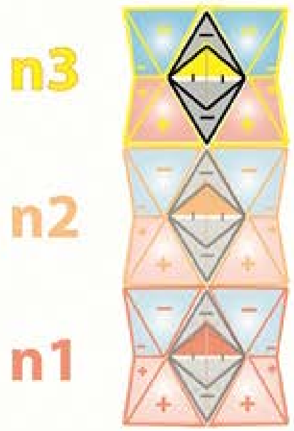

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 4

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
19	4s1	Potassium	19	19	19
20	4s2	Calcium	20	20	20
31	4p1	Gallium	31	31	31
32	4p2	Germanium	32	32	32
33	4p3	Arsenic	33	33	33
34	4p4	Selenium	34	34	34
35	4p5	Bromine	35	35	35
36	4p6	Krypton	36	36	36
39	4d1	Yttrium	39	39	39
40	4d2	Zirconium	40	40	40
41	4d3	Niobium	41	41	41
42	4d4	Molybdenum	42	42	42
43	4d5	Technetium	43	43	43
44	4d6	Ruthenium	44	44	44
45	4d7	Rhodium	45	45	45
46	4d8	Palladium	46	46	46
47	4d9	Silver	47	47	47
48	4d10	Cadmium	48	48	48
57	4f1	Lanthanum	57	57	57
58	4f2	Cerium	58	58	58
59	4f3	Praseodymium	59	59	59
60	4f4	Neodymium	60	60	60
61	4f5	Promethium	61	61	61
62	4f6	Samarium	62	62	62
63	4f7	Europium	63	63	63
64	4f8	Gadolinium	64	64	64
65	4f9	Terbium	65	65	65
66	4f10	Dysprosium	66	66	66
67	4f11	Holium	67	67	67
68	4f12	Erbium	68	68	68
69	4f13	Thulium	69	69	69
70	4f14	Ytterbium	70	70	70

32

Deuterium [not Hydrogen] is the building block of elements

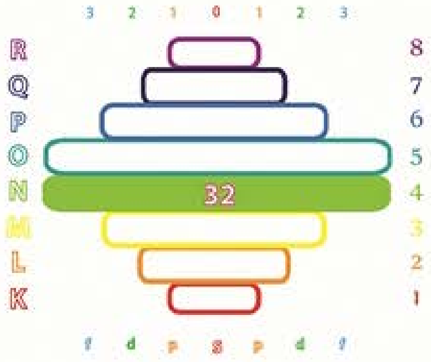
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
32
f d p s p d f

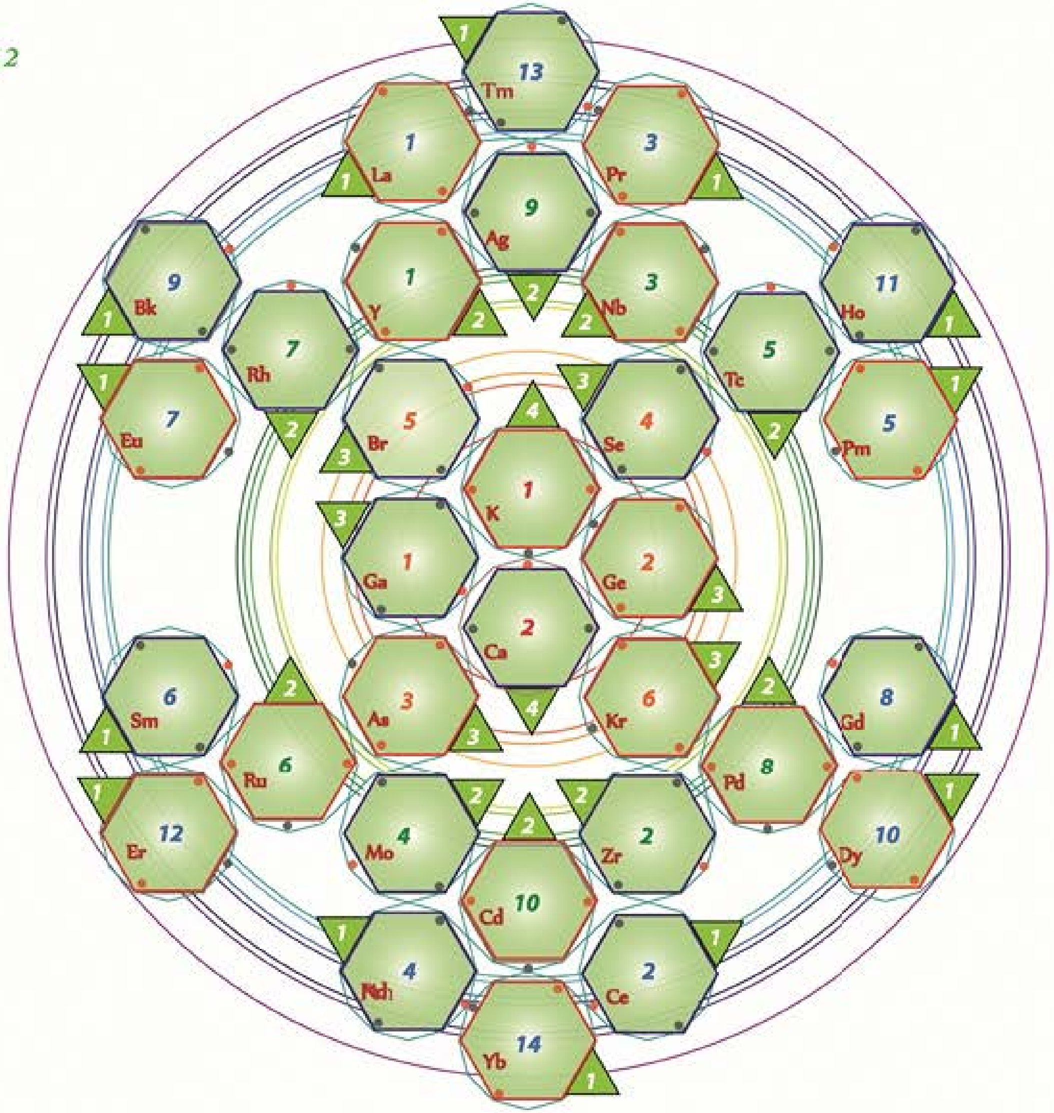
13 Tm
1 La
3 Pr
9 Ag
9 Bk
11 Ho
1 Y
3 Nb
7 Rh
5 Tc
7 Eu
5 Br
4 Se
5 Pm
1 K
1 Ga
2 Ge
2 Ca
6 Sm
3 As
6 Kr
8 Gd
6 Ru
8 Pd
12 Er
4 Mo
2 Zr
10 Dy
10 Cd
4 Nd
2 Ce
14 Yb

N shell

n4

Ground State electron

v
hv
v²
1.2e20
3.3812 eV

$iE = -10.143\ eV$

Energy level
4

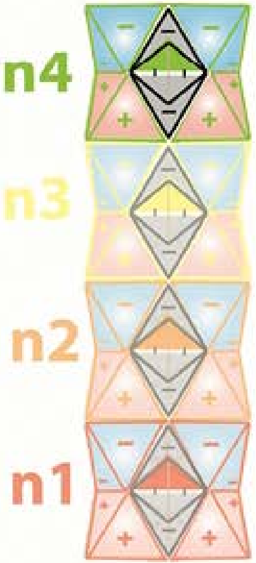

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 5

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
37	5s1	Rubidium	37	37	37
38	5s2	Strontium	38	38	38
49	5p1	Indium	49	49	49
50	5p2	Tin	50	50	50
51	5p3	Antimony	51	51	51
52	5p4	Tellurium	52	52	52
53	5p5	Iodine	53	53	53
54	5p6	Xenon	54	54	54
71	5d1	Lutetium	71	71	71
72	5d2	Hafnium	72	72	72
73	5d3	Tantalum	73	73	73
74	5d4	Tungsten	74	74	74
75	5d5	Rhenium	75	75	75
76	5d6	Osmium	76	76	76
77	5d7	Iridium	77	77	77
78	5d8	Platinum	78	78	78
79	5d9	Gold	79	79	79
80	5d10	Mercury	80	80	80
89	5f1	Actinium	89	89	89
90	5f2	Thorium	90	90	90
91	5f3	Protactium	91	91	91
92	5f4	Uranium	92	92	92
93	5f5	Neptunium	93	93	93
94	5f6	Plutonium	94	94	94
95	5f7	Americium	95	95	95
96	5f8	Curium	96	96	96
97	5f9	Berkelium	97	97	97
98	5f10	Californium	98	98	98
99	5f11	Einsteinium	99	99	99
100	5f12	Fermium	100	100	100
101	5f13	Mendelevium	101	101	101
102	5f14	Nobelium	102	102	102

32

Deuterium [not Hydrogen] is the building block of elements

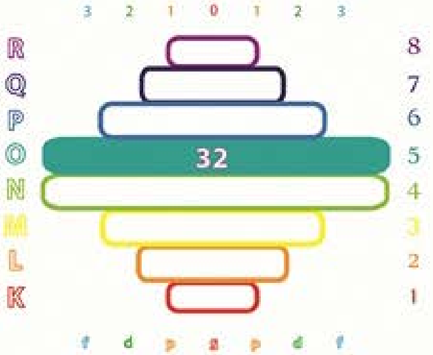
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
32
8
7
6
5
4
3
2
1
f d p s p d f

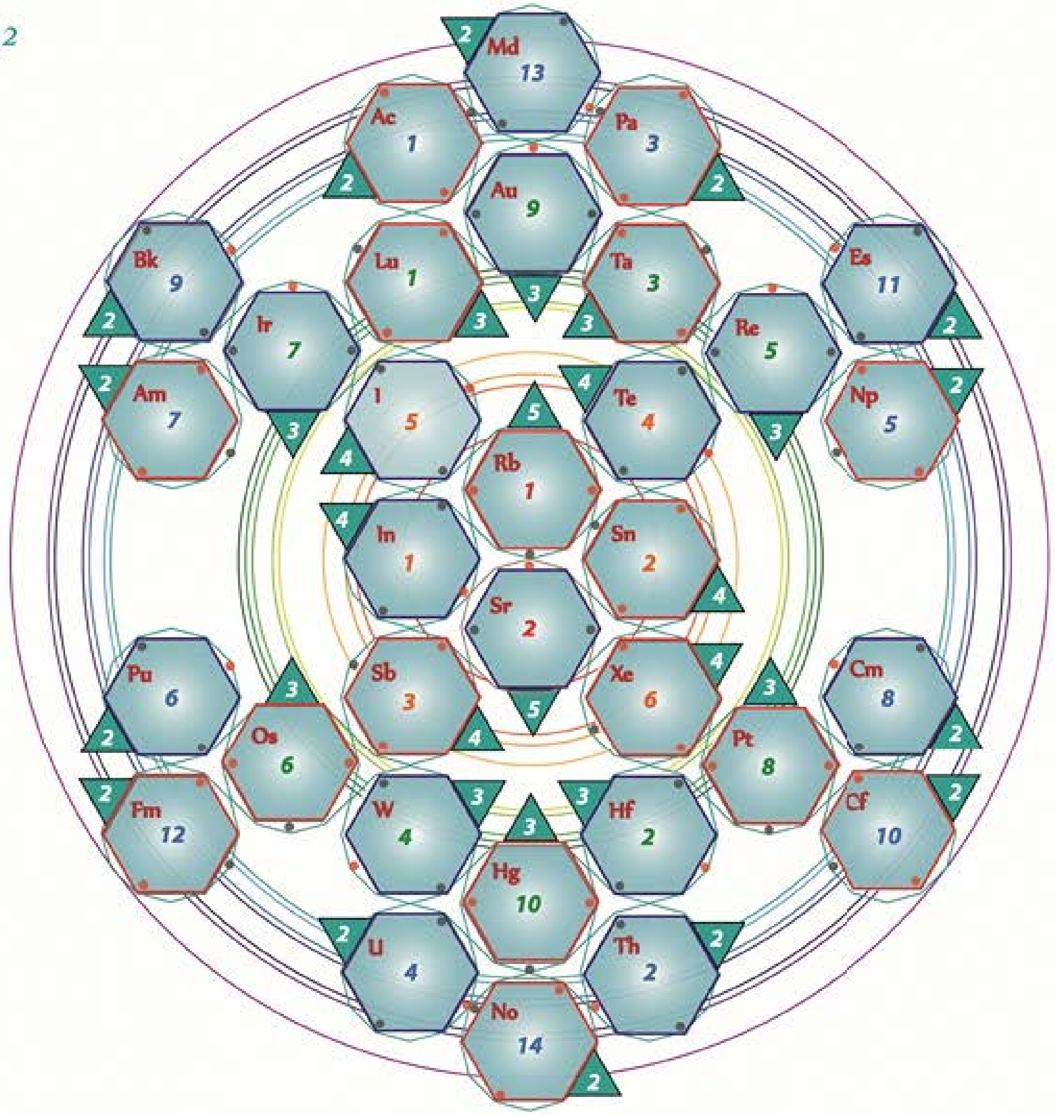
Md 13
Ac 1
Pa 3
Au 9
Bk 9
Lu 1
Ta 3
Es 11
Ir 7
Re 5
Am 7
I 5
Te 4
Np 5
Rb 1
In 1
Sn 2
Sr 2
Pu 6
Sb 3
Xe 6
Cm 8
Os 6
Pt 8
Fm 12
W 4
Hf 2
Cf 10
Hg 10
U 4
Th 2
No 14

O shell

n5

Ground State electron

hv
v
v²
KE
1.2e20
5.5282 eV

$iE = -8.241\ eV$

Energy level
5

n5
n4
n3
n2
n1

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 6

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
55	6s1	Caesium	55	55	55
56	6s2	Barium	56	56	56
81	6p1	Thallium	81	81	81
82	6p2	Lead	82	82	82
83	6p3	Bismuth	83	83	83
84	6p4	Polonium	84	84	84
85	6p5	Astatine	85	85	85
86	6p6	Radon	86	86	86
103	6d1	Lawrencium	103	103	103
104	6d2	Rutherfordium	104	104	104
105	6d3	Dubnium	105	105	105
106	6d4	Seaborgium	106	106	106
107	6d5	Bohrium	107	107	107
108	6d6	Hassium	108	108	108
109	6d7	Meitnerium	109	109	109
110	6d8	Darmstadtium	110	110	110
111	6d9	Roetgenium	111	111	111
112	6d10	Copernicium	112	112	112

18

Deuterium [not Hydrogen] is the building block of elements

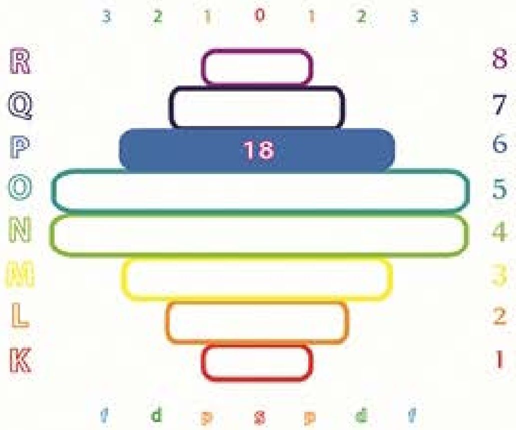
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
18
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level
6

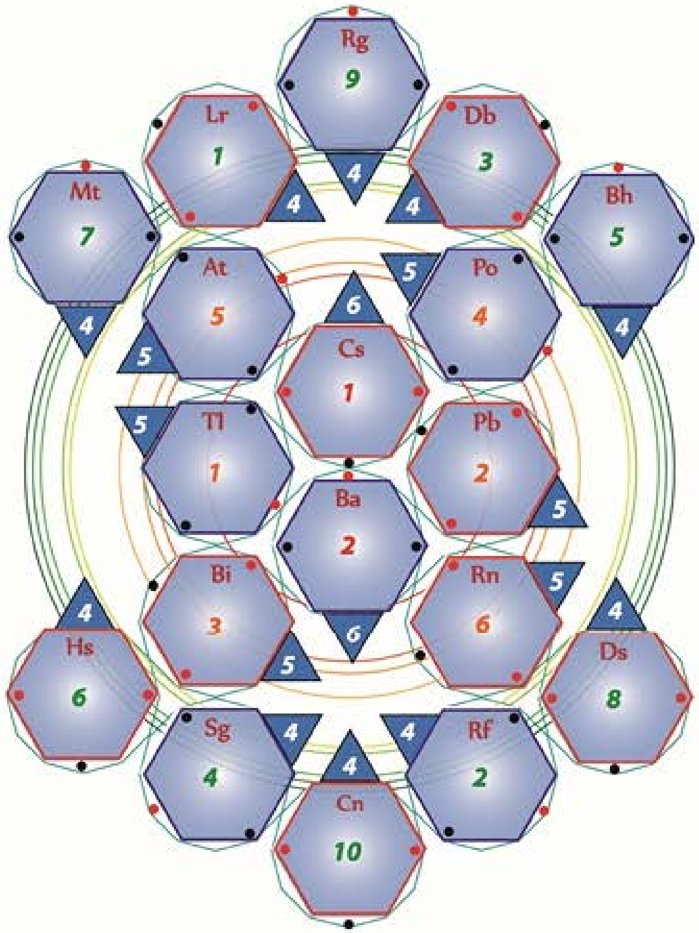
Rg
9
Lr
1
Db
3
Mt
7
Bh
5
At
5
Po
4
Cs
1
Tl
1
Pb
2
Ba
2
Bi
3
Rn
6
Hs
6
Ds
8
Sg
4
Rf
2
Cn
10
4
5
6

P shell

n6

Ground State electron

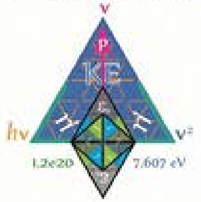
v
P
KE
hv
v²
1.2e20
7.607 eV

iE = -5.917 eV

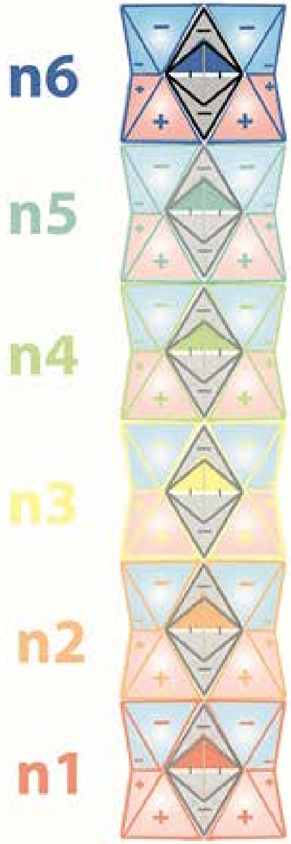
n6
n5
n4
n3
n2
n1
−
+

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 7

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
87	7s1	Francium	87	87	87
88	7s2	Radium	88	88	88
113	7p1	Ununtrium	113	113	113
114	7p2	Flerovium	114	114	114
115	7p3	Ununpentium	115	115	115
116	7p4	Livermorium	116	116	116
117	7p5	Ununseptium	117	117	117
118	7p6	Ununoctium	118	118	118

8

Deuterium [not Hydrogen] is the building block of elements

Q shell

n7

Ground State electron

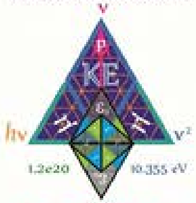
v
p
KE
hv
v²
1.2e20
10.355 eV

iE = -3.169 eV

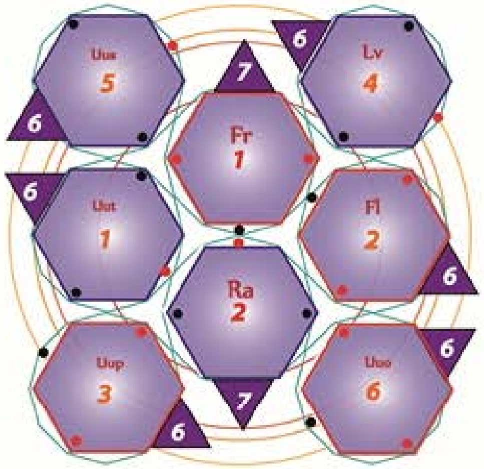
Uus
5
Lv
4
7
6
6
Fr
1
6
Uut
1
Fl
2
Ra
2
6
6
Uup
3
Uuo
6
7
6

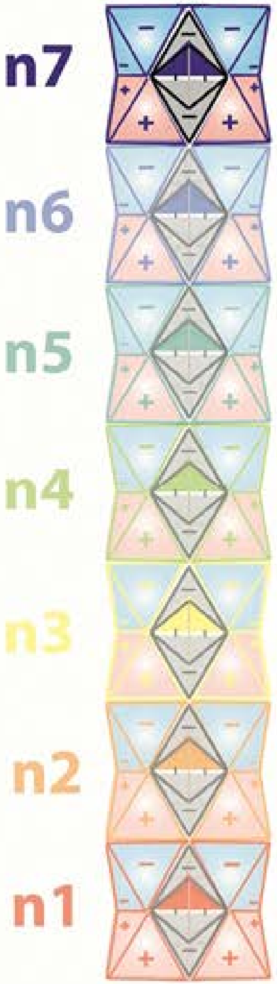
n7
n6
n5
n4
n3
n2
n1

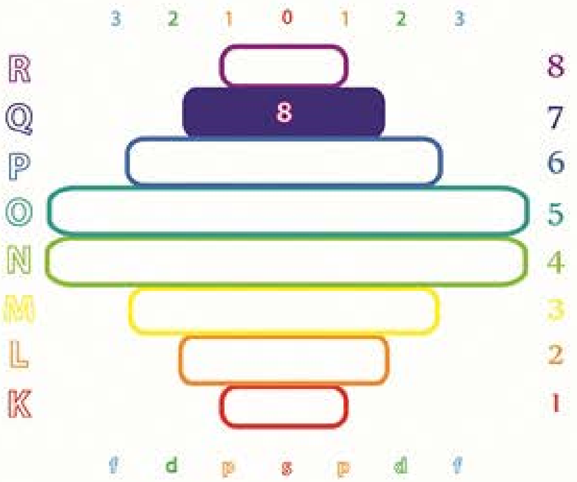
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level

Quantum Level 8

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
119	8s1	Ununnonium	119	119	119
120	8s2	Unbinilium	120	120	120

2

Deuterium [not Hydrogen] is the building block of elements

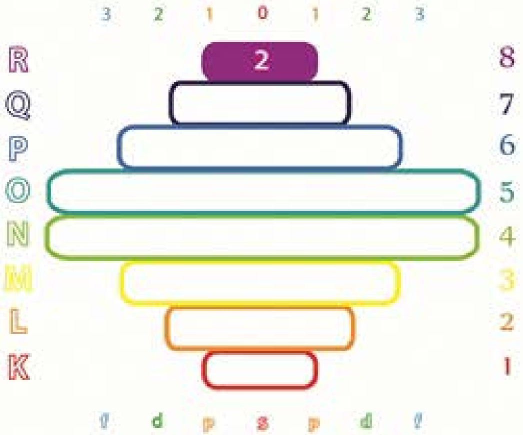
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

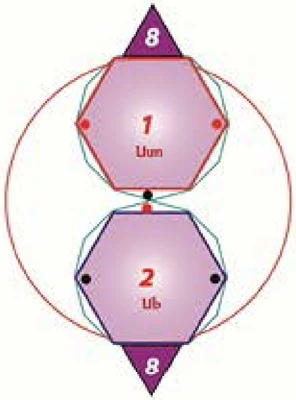

Energy level
8

R shell

n8

Ground State electron

v
P
KE
hv
v²
1.2e20
13.525 eV

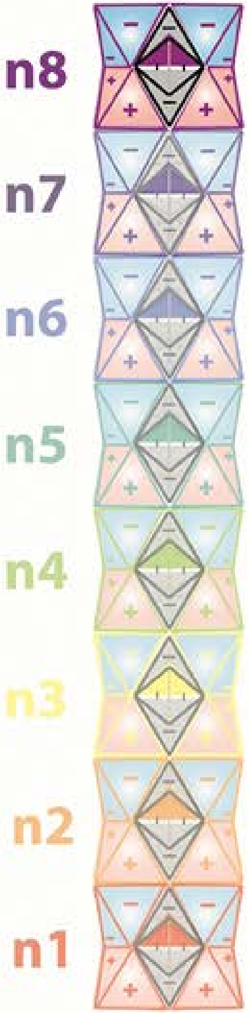
n8
n7
n6
n5
n4
n3
n2
n1

Copyright ABRAHAM [2008] - All rights reserved

Spectral line transitions

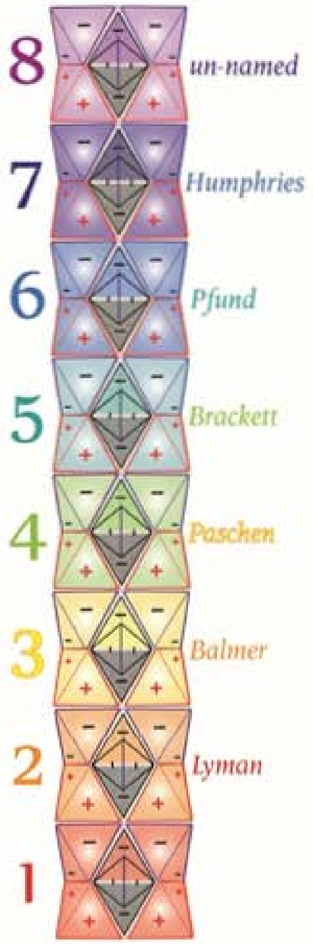
8
un-named
7
Humphries
6
Pfund
5
Brackett
4
Paschen
3
Balmer
2
Lyman
1

Quantum level jumps

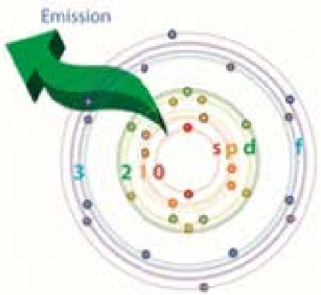
Emission
3
2
1
0
s
p
d
f

Quantum Level Jumps

Photon Absorption and Release

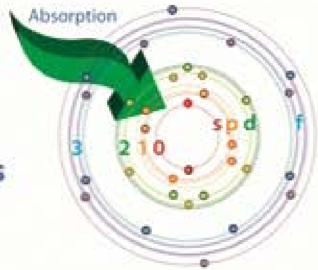
Absorption
3
2
1
0
s
p
d
f

Photo-electrons can only transition between principal energy Baryons in the atomic nuclei in discrete steps [or quantum jumps] because Baryons determine the KEM energy levels of electrons in nuclei

[see Tetryonic QED for full details on spectral line mechanics]

Initial Quantum level

8	768	756	720	660	576	468	336	180	0
7	588	576	540	480	396	288	156	0	
6	432	420	384	324	240	132	0		
5	300	288	252	192	108	0			
4	192	180	144	84	0				
3	108	96	60	0					
2	48	36	0						
1	12	0							
		12	48	108	192	300	432	588	768
		1	2	3	4	5	6	7	8

Final Quantum level

$$\Delta p = \Delta M v = hf$$

accelerating photo-electrons produce spectral lines

Spectral line transitions

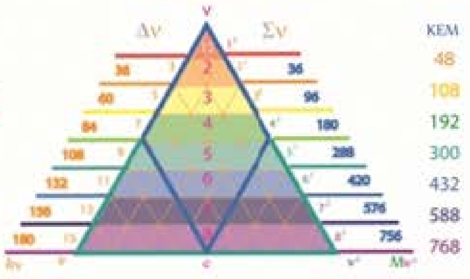
v
Δv
Σv
36
60
84
108
132
156
180
36
96
180
288
420
576
756
2
3
4
5
6
KEM
48
108
192
300
432
588
768
hv
c
Mv²

Quantum differentials

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity through equilateral geometry

Nuclear energy emission~absorption

If atoms are left undisturbed, their electrons usually fill the lowest available energy levels and stay there, in their "ground state."

Occasionally, however, they may also be pushed up to some higher energy ("become excited") e.g. by a collision with a fast atom or electron, one which got extra speed from an electric voltage or from some source of heat.

An atom/electron elevated to one of its higher "excited levels" soon falls back to a lower level ("undergoes a quantum jump"), emitting a photon whose energy corresponds to the difference between the levels.

That need not be the ground state: the atom/electron might descend to that state in several steps, emitting a photon at each step on the way.

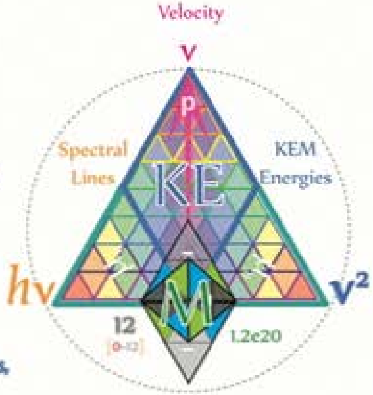
Velocity
v
P
Spectral Lines
KEM Energies
KE
hv
v²
M
12
[0-12]
1.2e20

Copyright ABRAHAM [2008] - All rights reserved

Quantum transitions

(Orbital Shells - Bound energy states)

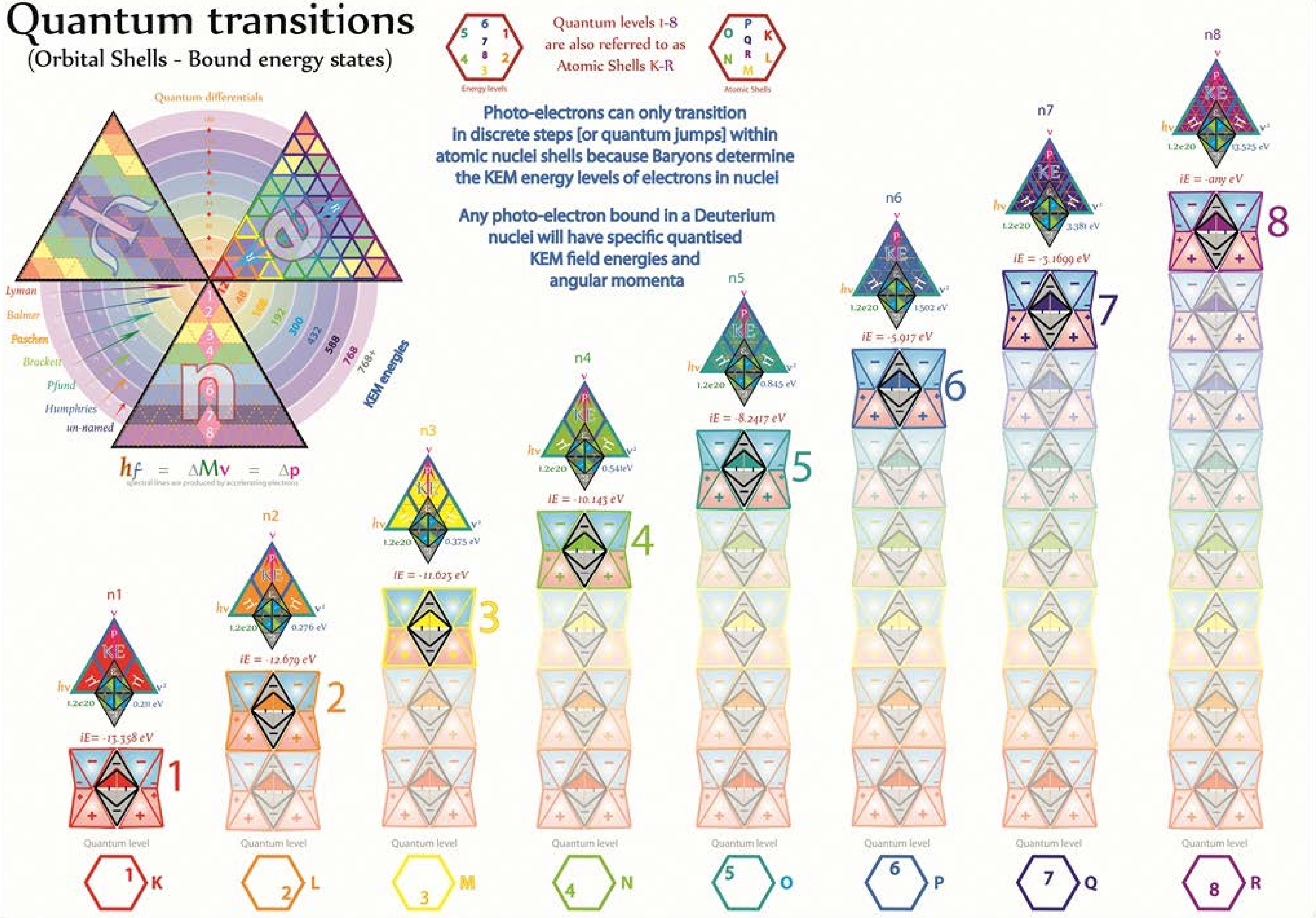

6
5
7
1
4
8
2
3
Energy levels
P
O
Q
K
R
N
L
M
Atomic Shells
Quantum differentials
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
1
2
3
4
6
7
8
n
12
48
108
192
300
432
588
768
768+
KEM energies
n1
p
KE
hv
v²
1.2e20
0.211 eV
iE= -13.358 eV
1
n2
1.2e20
0.276 eV
iE = -12.679 eV
2
n3
1.2e20
0.375 eV
iE = -11.623 eV
3
n4
1.2e20
0.541eV
iE = -10.143 eV
4
n5
1.2e20
0.845 eV
iE = -8.2417 eV
5
n6
1.2e20
1.502 eV
iE = -5.917 eV
6
n7
1.2e20
3.381 eV
iE = -3.1699 eV
7
n8
1.2e20
13.525 eV
iE = -any eV
8
Quantum level
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

Quantum levels 1-8
are also referred to as
Atomic Shells K-R

Photo-electrons can only transition in discrete steps [or quantum jumps] within atomic nuclei shells because Baryons determine the KEM energy levels of electrons in nuclei

Any photo-electron bound in a Deuterium nuclei will have specific quantised KEM field energies and angular momenta

$$hf = \Delta Mv = \Delta p$$

spectral lines are produced by accelerating electrons

Copyright ABRAHAM [2008] - All rights reserved

AZMITHAL MAGNETIC

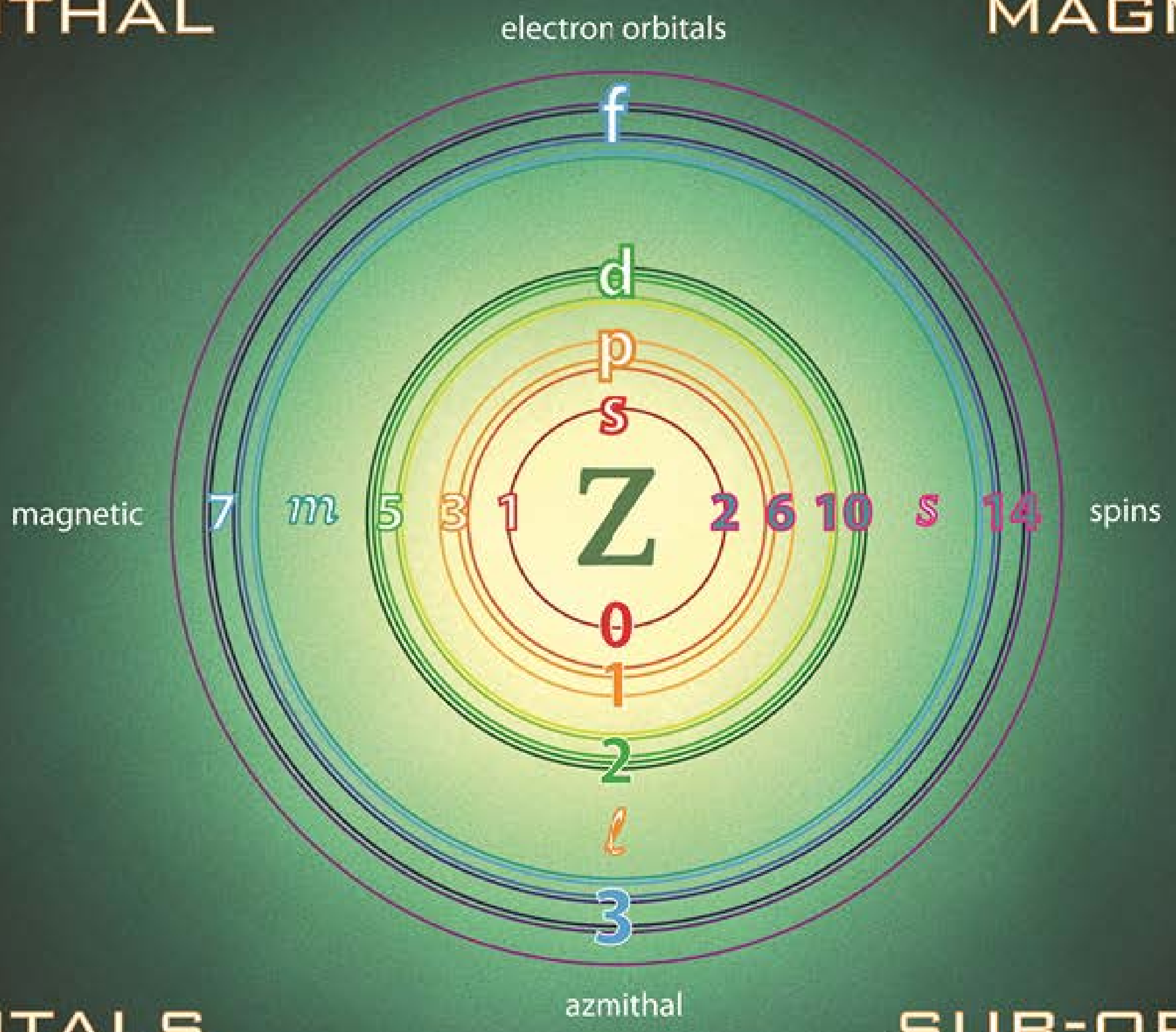
electron orbitals
f
d
p
s
magnetic
7
m
5
3
1
Z
2
6
10
s
14
spins
0
1
2
ℓ
3
azmithal

ORBITALS SUB-ORBITALS

Atomic Orbitals

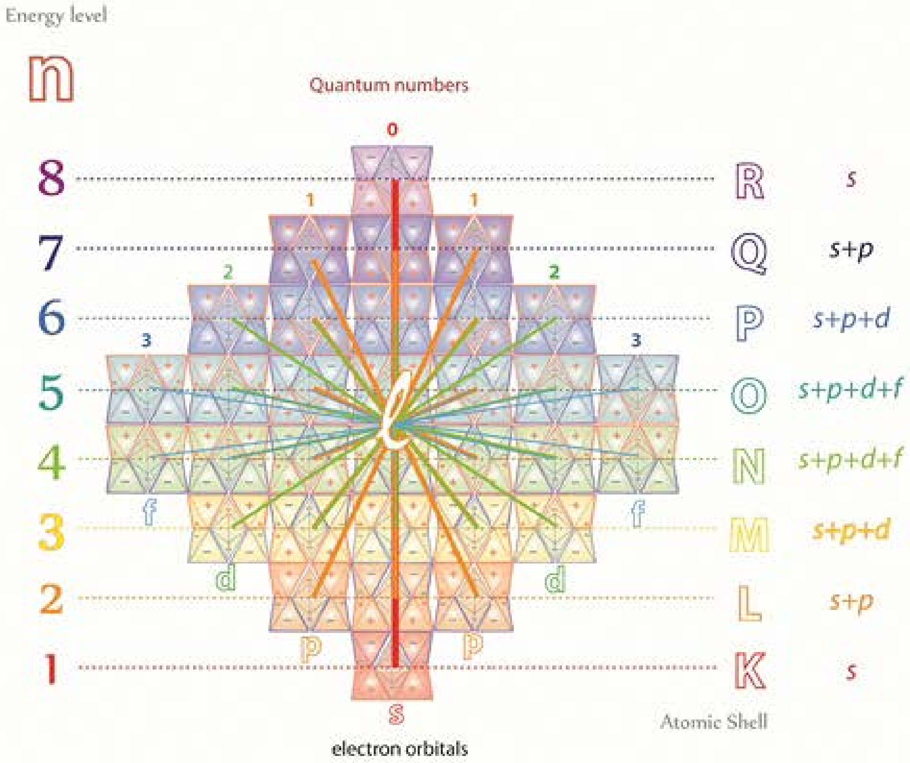
Energy level
n
Quantum numbers
0
1
1
2
2
3
3
8
7
6
5
4
3
2
1
R
Q
P
O
N
M
L
K
s
s+p
s+p+d
s+p+d+f
s+p+d+f
s+p+d
s+p
s
l
f
f
d
d
p
p
s
Atomic Shell
electron orbitals

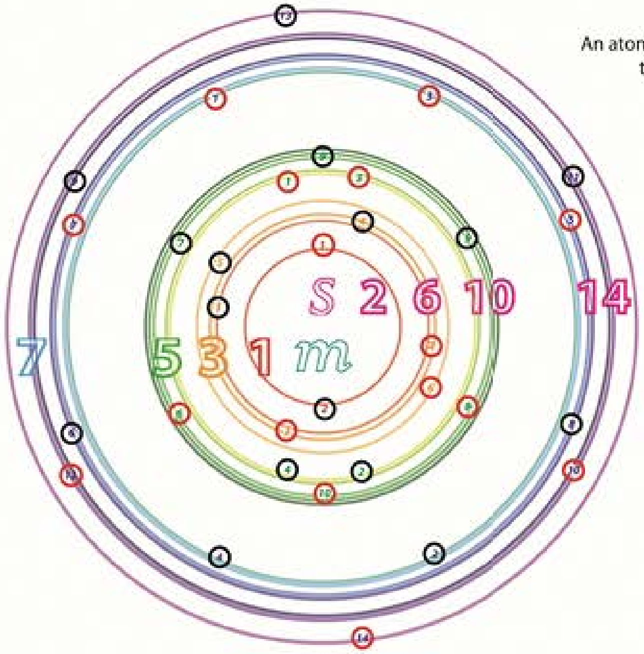
S
2
6
10
14
7
5
3
1
m

An atomic orbital is a mathematical function that describes the wave-like behavior of either one electron or a pair of electrons in an atom

Atomic orbitals are typically categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and an angular momentum vector component, respectively.

Historically used to define the pedagogical electron cloud model of an atom Tetryonics reveals the true geometry of atomic nuclei

Each orbital is defined by a different set of quantum numbers and contains a maximum of two spin opposed electrons.

Aufbau

Azmithal & Magnetic numbers

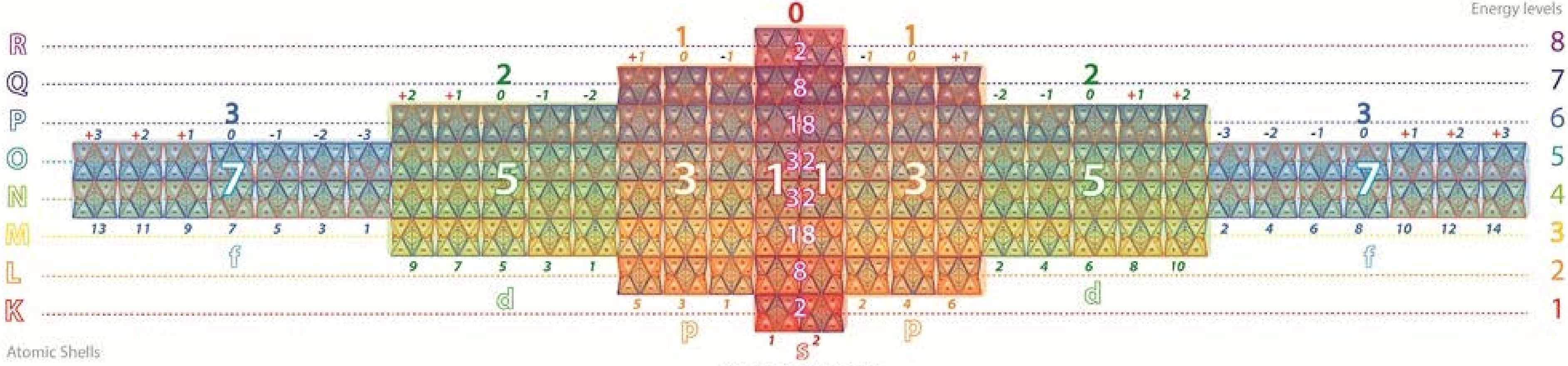
Energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
2
8
18
32
32
18
8
2
7
5
3
1
1
3
5
7
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
f
f
d
d
p
p
1 s 2
Atomic Shells

Orbitals & sub-Orbitals

Electron Spins can be either up or down providing they obey the Pauli exclusion principle

s Orbital

1 Orbit (2 electrons max)

Alkali Metals & Alkali Earths

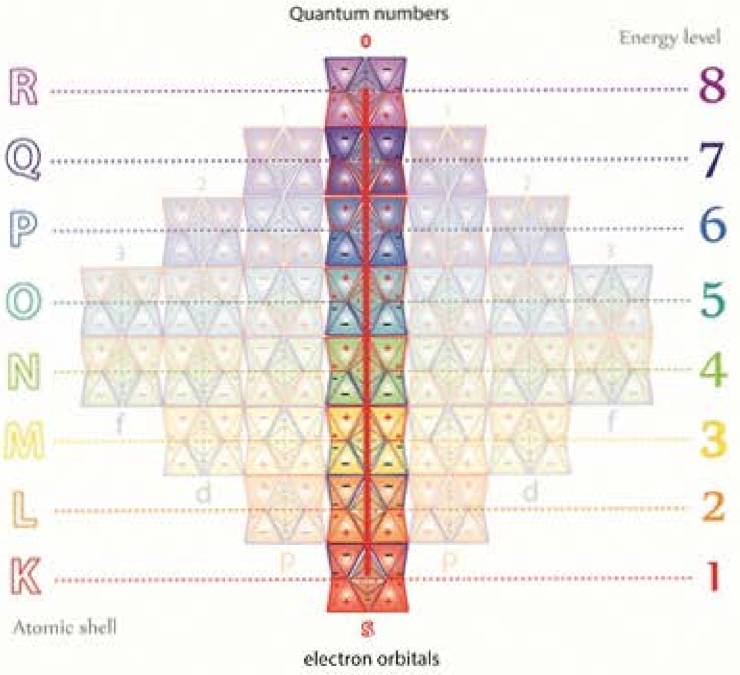
Quantum numbers
0
Energy level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f
d
p
Atomic shell
s
electron orbitals

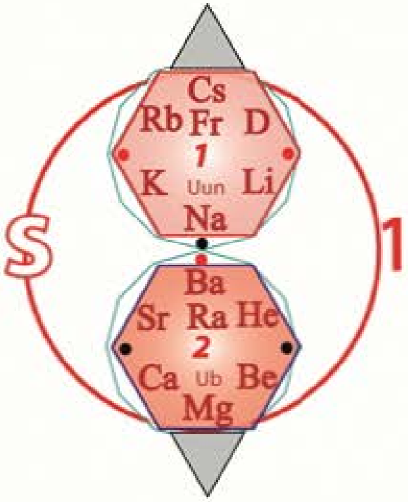
Cs
Rb
Fr
D
1
K
Uun
Li
Na
s
1
Ba
Sr
Ra
He
2
Ca
Ub
Be
Mg

		n
azmithal number	0	l
magnetic numbers	0	m
spins	-1/2 +1/2	s

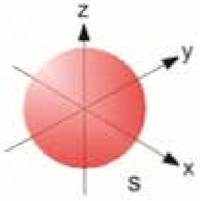

erroneous computer model of 's' electron orbitals

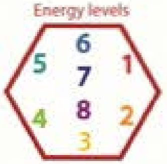
Energy levels
6
5
1
7
4
8
2
3

p Orbital

3 Orbits (6 electrons max)

Non-Metals, Halogens & Nobel gases

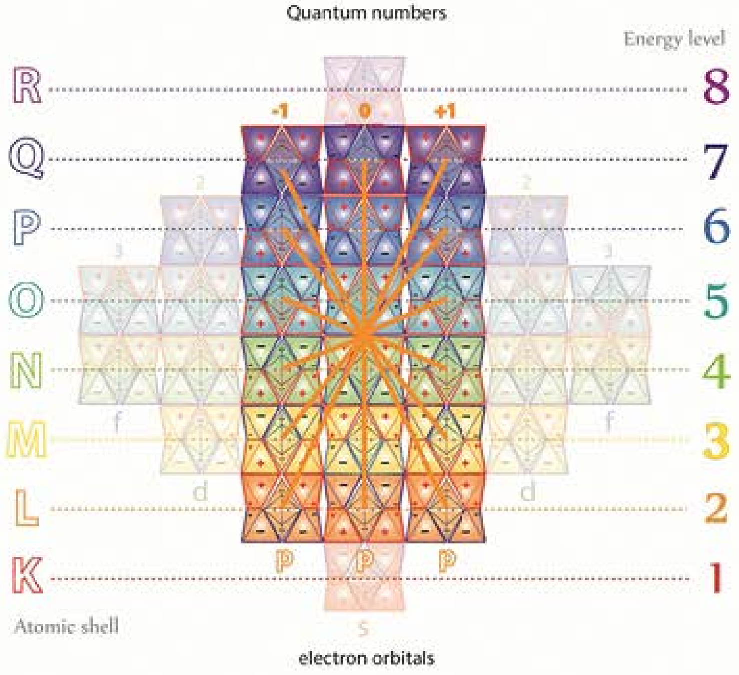
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-1
0
+1
p
p
p
d
d
f
f
s
Atomic shell
electron orbitals

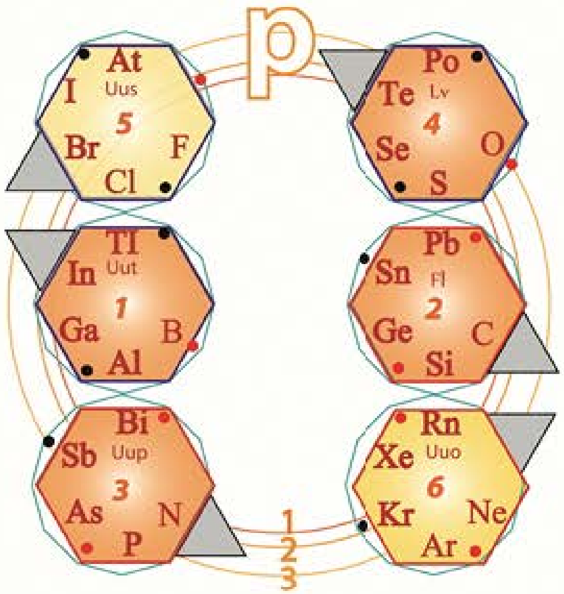
p
At
I
Uus
5
Br
F
Cl
Po
Te
Lv
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar
1
2
3

n

ℓ

m

s

azmithal number		1		
magnetic numbers	-1	0	+1	
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	

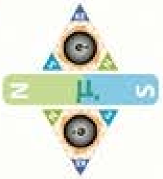

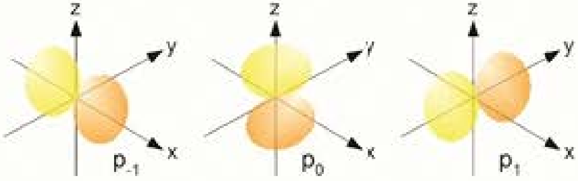
z
y
x
p_{-1}
z
y
x
p_0
z
y
x
p_1

erroneous computer model of 'p' electron orbitals

Quantum levels
6
5
7
4
2
3

d Orbital

5 Orbits (10 electrons max)

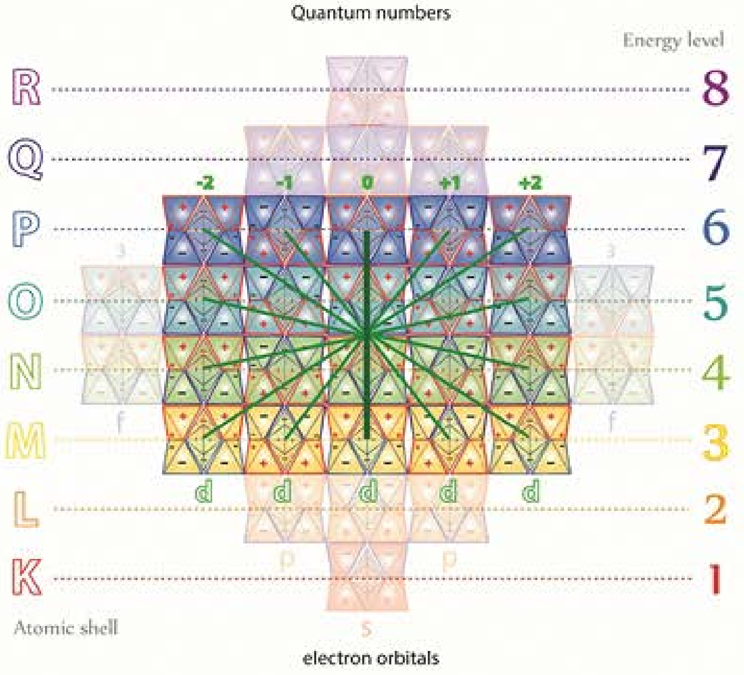
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-2
-1
0
+1
+2
d
d
d
d
d
f
f
p
p
s
Atomic shell
electron orbitals

						n
azimthal number			2			ℓ
magnetic numbers	-2	-1	0	+1	+2	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

Transition & post-Transition Metals

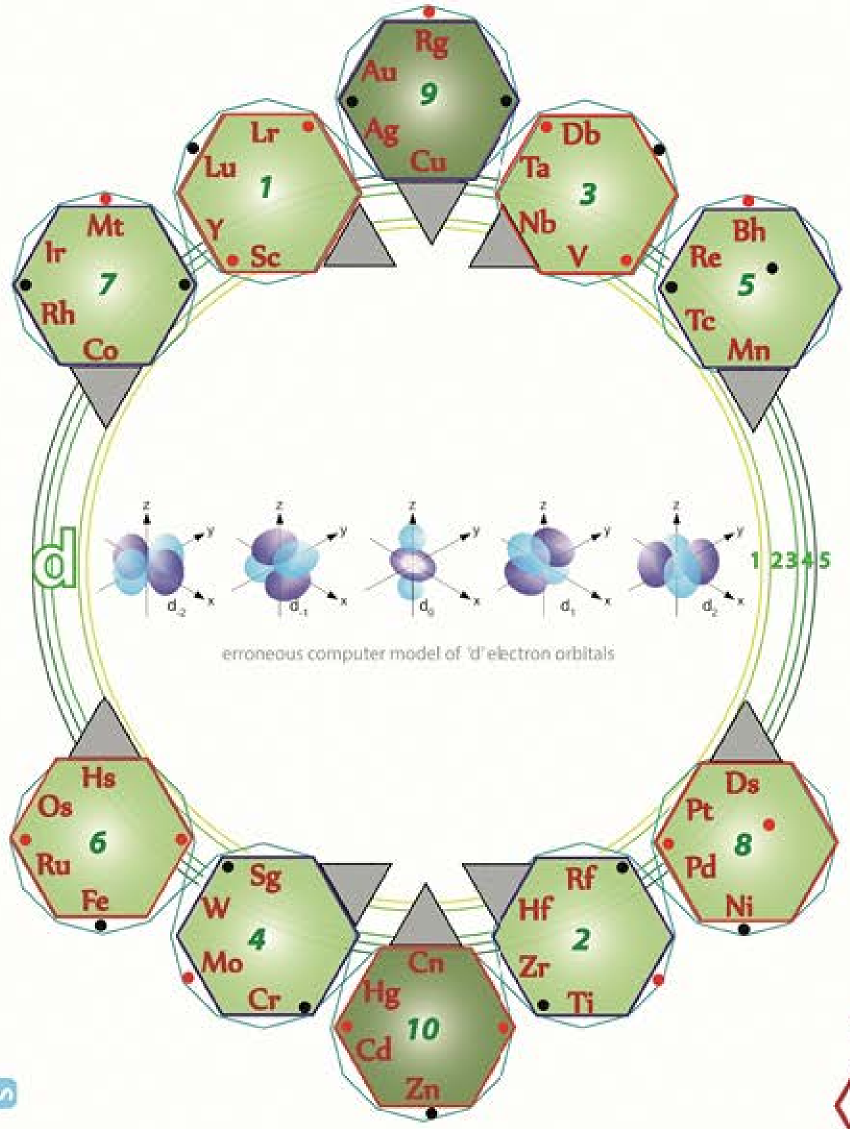
Rg
Au
9
Ag
Cu
Lr
Lu
1
Y
Sc
Db
Ta
3
Nb
V
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
d
z
y
x
d-2
d-1
d0
d1
d2
1 2 3 4 5
erroneous computer model of 'd' electron orbitals
Hs
Os
6
Ru
Fe
Sg
W
4
Mo
Cr
Cn
Hg
10
Cd
Zn
Rf
Hf
2
Zr
Ti
Ds
Pt
8
Pd
Ni

Quantum levels
6
5
4
3

Copyright ABRAHAM [2008] - All rights reserved

f Orbital

7 Orbits (14 electrons max)

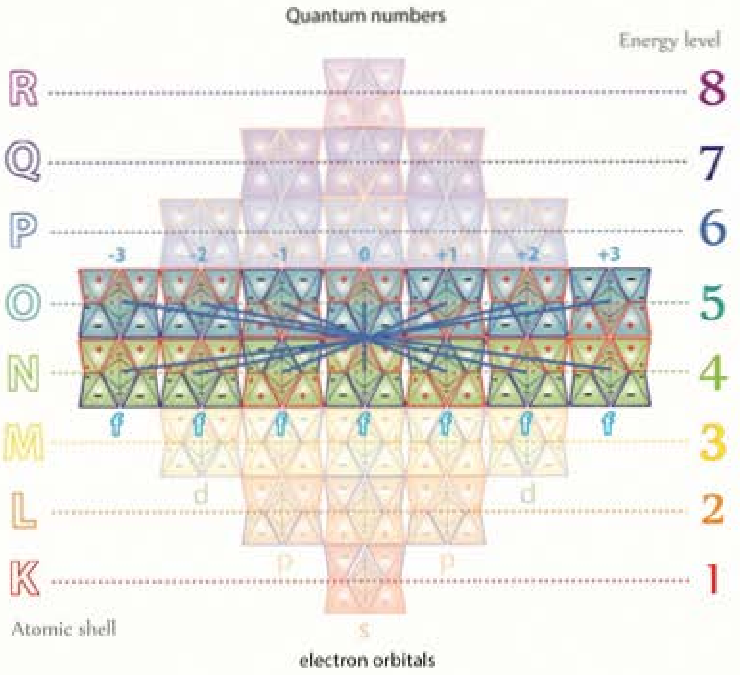
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-3
-2
-1
0
+1
+2
+3
f
d
p
s
Atomic shell
electron orbitals

								n
azmithal number				3				ℓ
magnetic numbers	-3	-2	-1	0	+1	+2	+3	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

Lanthanoids & Actinoids

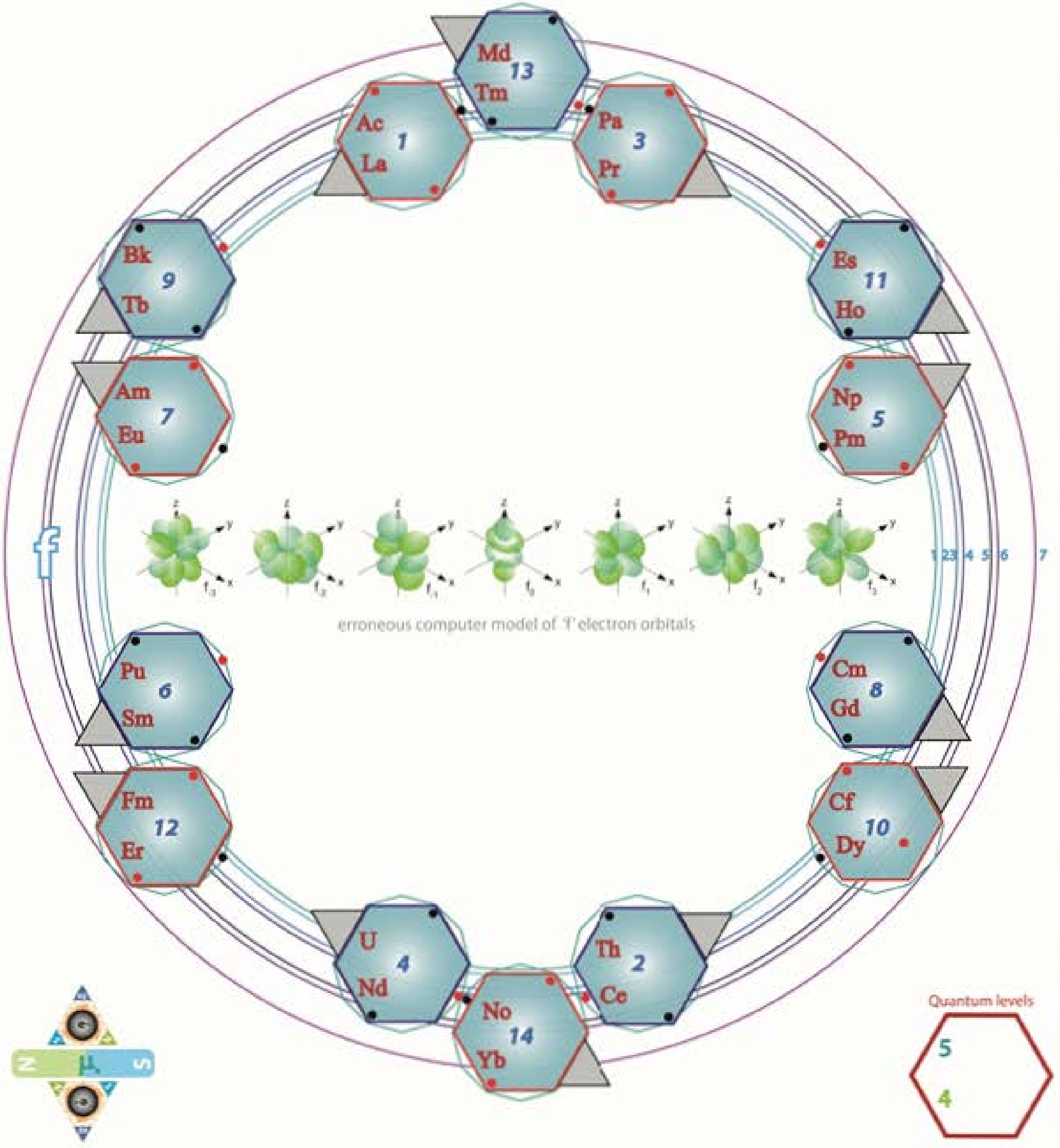
Md 13 Tm
Ac 1 La
Pa 3 Pr
Bk 9 Tb
Es 11 Ho
Am 7 Eu
Np 5 Pm
f
1 2 3 4 5 6 7
Pu 6 Sm
Cm 8 Gd
Fm 12 Er
Cf 10 Dy
U 4 Nd
Th 2 Ce
No 14 Yb
N
μ
S
Quantum levels
5
4

erroneous computer model of 'f' electron orbitals

Copyright ABRAHAM [2008] - All rights reserved

Orbital energy variations

All Elements have stable core electron configurations of s & p orbitals for each energy level as revealed through diffraction studies

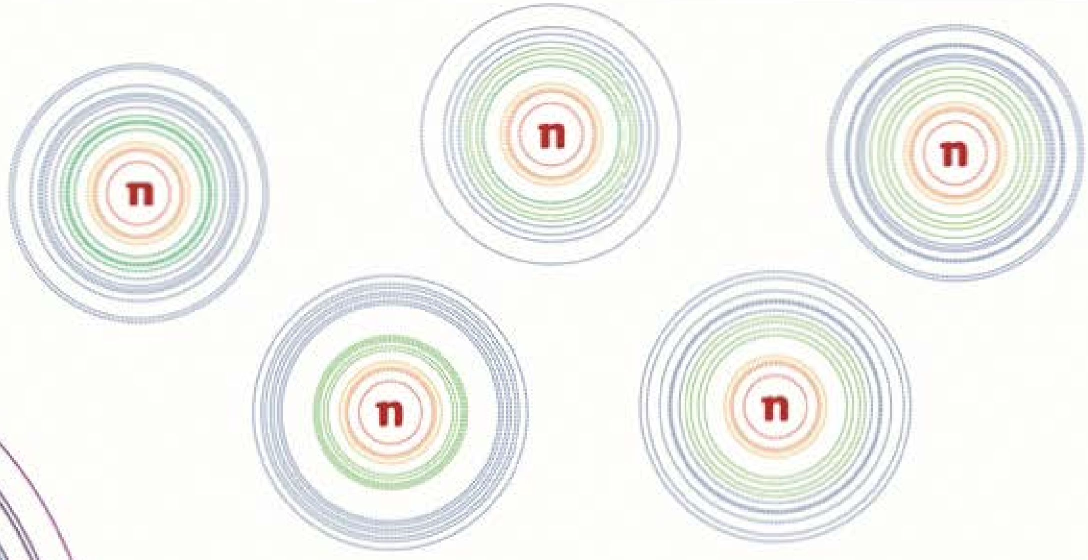

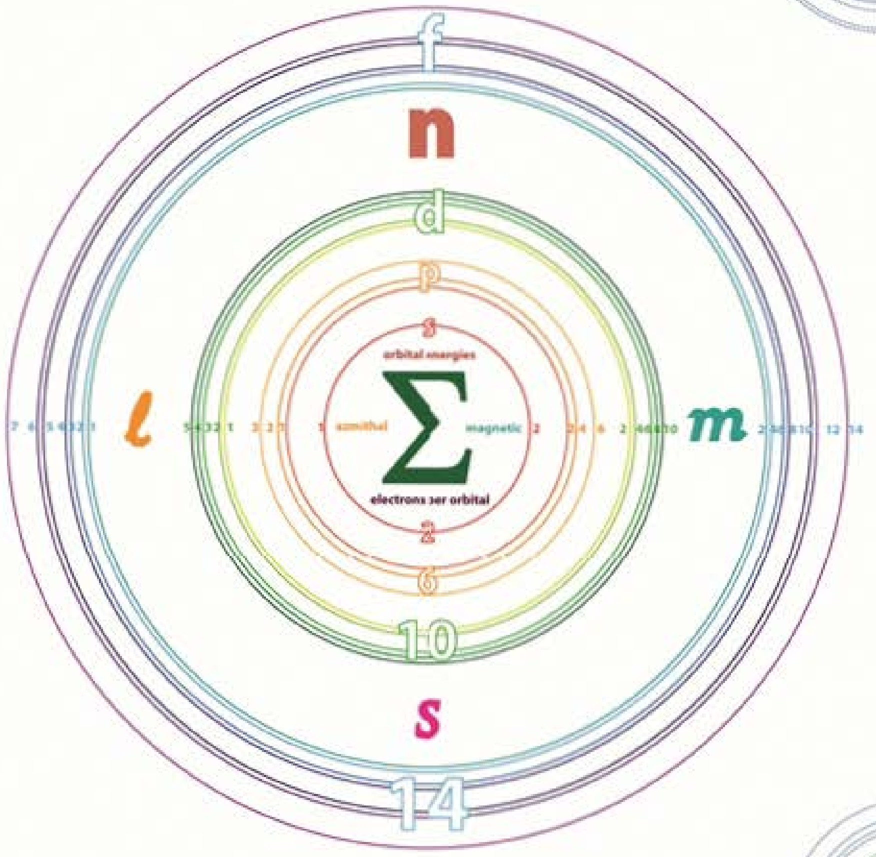
f
n
d
p
s
orbital energies
l
azimuthal
Σ
magnetic
m
electrons per orbital
2
6
10
s
14

As additional nuclei bind to form d & f orbitals they can do so by bonding in many positions, with each location producing different orbital energies for each electron that binds to nuclei in that position

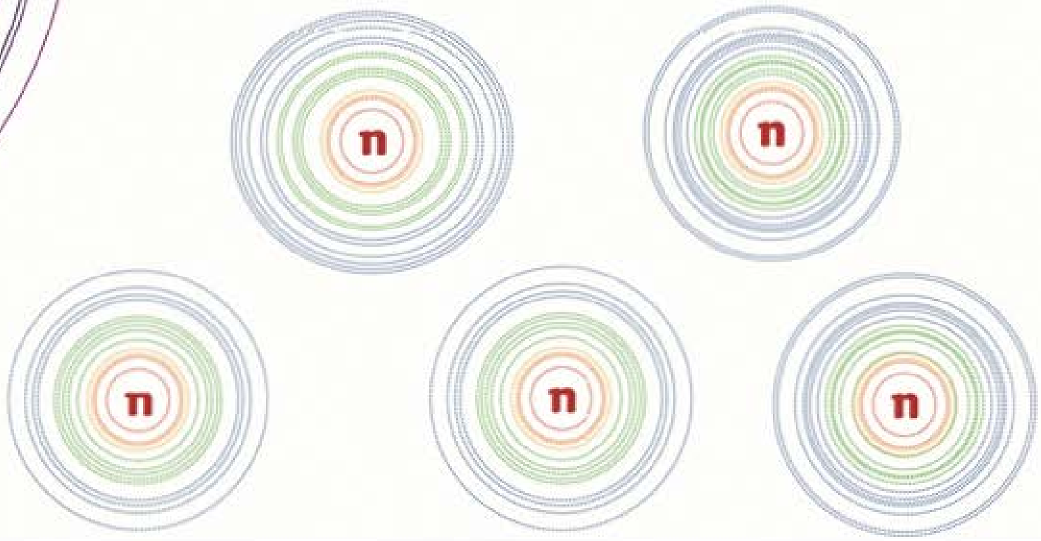

The final energy levels of each orbital is the result of the energy of the Baryons in the nuclei and the spin coupling energies of the photo-electrons bound to them

Copyright ABRAHAM [2008] - All rights reserved

Electrons per shell

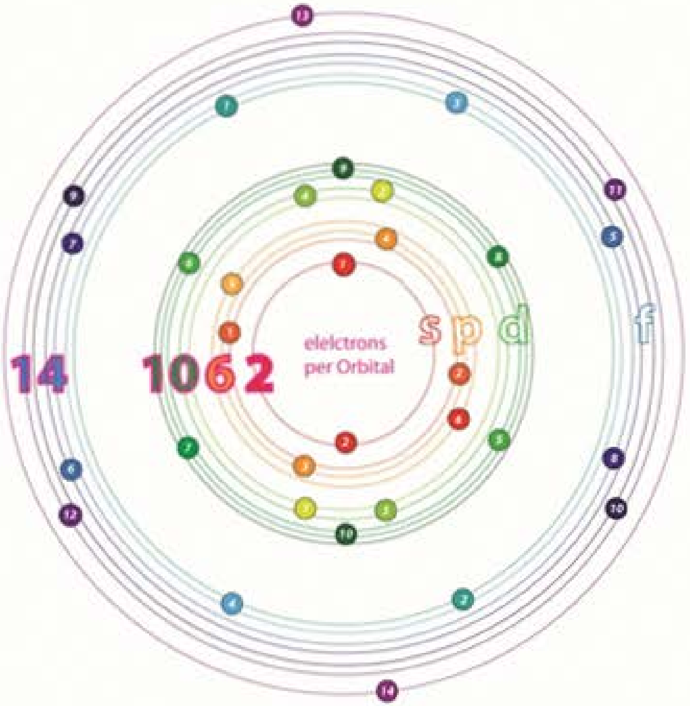

Each energy shell of a periodic element can hold only a fixed number of electrons

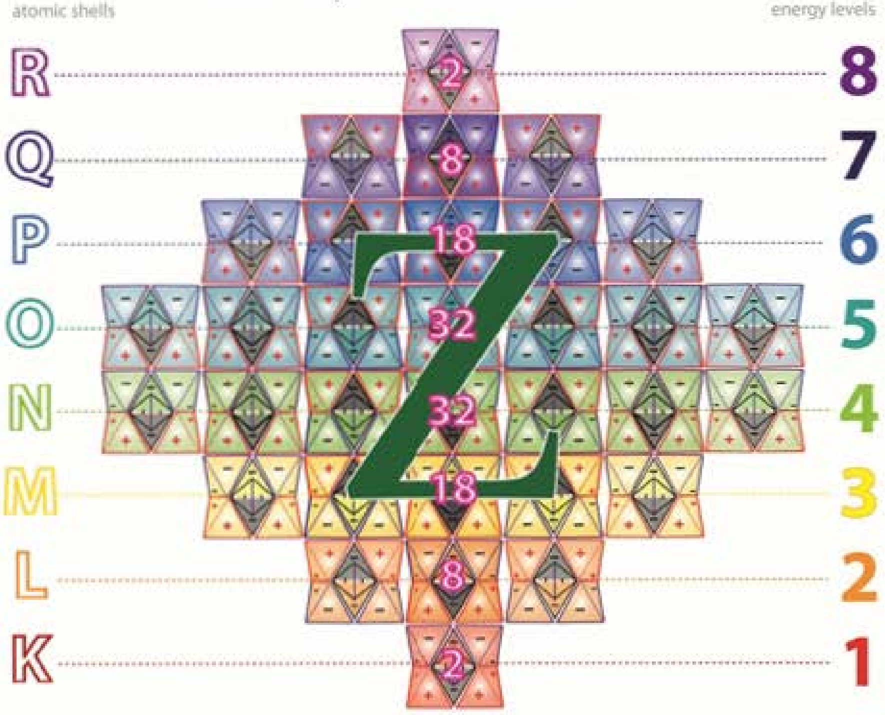

Atomic Orbitals

f	d	p	s	p	d	f
5, 11, 3, 13, 1, 9, 7	8, 2, 10, 4, 6	5, 1, 3	1-2	6, 2, 4	7, 1, 9, 3, 5	6, 12, 4, 14, 2, 10, 8

7 + 5 + 3 + 2 + 3 + 5 + 7

Rules governing the allowed combinations of Quantum Numbers

The three quantum numbers (n, l, and m)
that describe an orbital are integers: 0, 1, 2, 3, and so on

n (1-8) Principal (n = 1, 2, 3, 4 ...)	The principal quantum number (n) cannot be zero.
ℓ (0-3) Azimuthal (l = 0, 1 ... n-1)	The angular quantum number (l) can be any integer between 0 and n - 1.
$m\ell\,(2\ell+1)$ Magnetic (ml = -l, -l+1 ... 0 ... l-1, l)	The magnetic quantum number (m) can be any integer between -l and +l.
m_s : $\frac{1}{2}$ Spin Projection (ms = -1/2 or +1/2)	The Spin of electrons in any nuclear sub-orbital can only be +1/2 (Spin UP) or -1.2 (Spin DOWN).

Z

1, 2, 3, 4, 5, 6, 7, 8

s, p, d, f

1	2
1, 3, 5	2, 4, 6
1, 3, 5, 7, 9	2, 4, 6, 8, 10
1, 3, 5, 7, 9, 11, 13	2, 4, 6, 8, 10, 12, 14
down up	up down

Quantum Numbers

n	4-5	3-6	2-7	1-8	2-7	3-6	4-5
ℓ	3	2	1	0	1	2	3
m	-3	-2	-1	1	+1	+2	+3
s	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2

Tetryonics 47.08 - Schrodinger's Quantum numbers

Copyright ABRAHAM [2008] - All rights reserved

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

Using Tetryonic charged geometries for mass-ENERGY-Matter,
an electron's position and velocity CAN be modelled simultaneously
(but any attempt to measure or interact with it, will affect its component energy-momenta)

Electron Position Uncertainty

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over any spatial region of measurement, categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and a vector momentum component, respectively

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

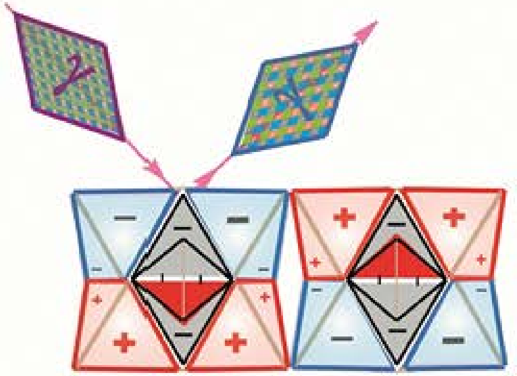

The energies of photo-electrons are determined by the Baryons they bind to & incident photons

Quantum Mechanics is a mathematical representation of equilateral energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

Leptons are 12 loop quantum rotors

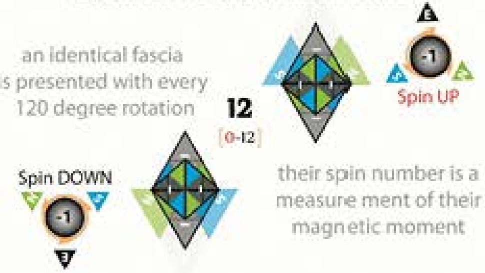

an identical fascia is presented with every 120 degree rotation

their spin number is a measure ment of their magnetic moment

The unique 12 faceted topology of leptons results in an identical EM geometry being oberved for every 120° rotation of the particle

Leading to the interpretation that the Lepton disappears and re-appears when being 'observed' or measured

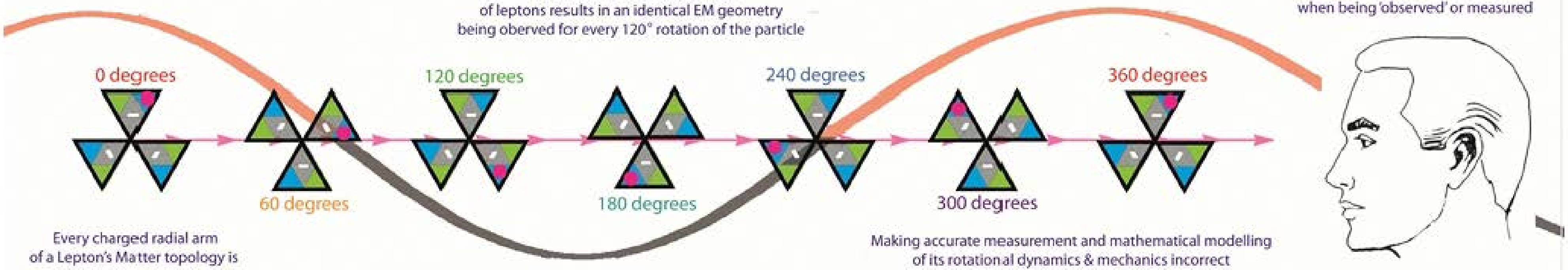

Every charged radial arm of a Lepton's Matter topology is identical to every other

Making accurate measurement and mathematical modelling of its rotational dynamics & mechanics incorrect without the correct physical topologies

Electron modelling & probability calculations

12
[0-12]
RE
S
KE

Electron positional measurements have proven to be historically difficult to accurately model due to the charged 12π rotating topologies of leptons

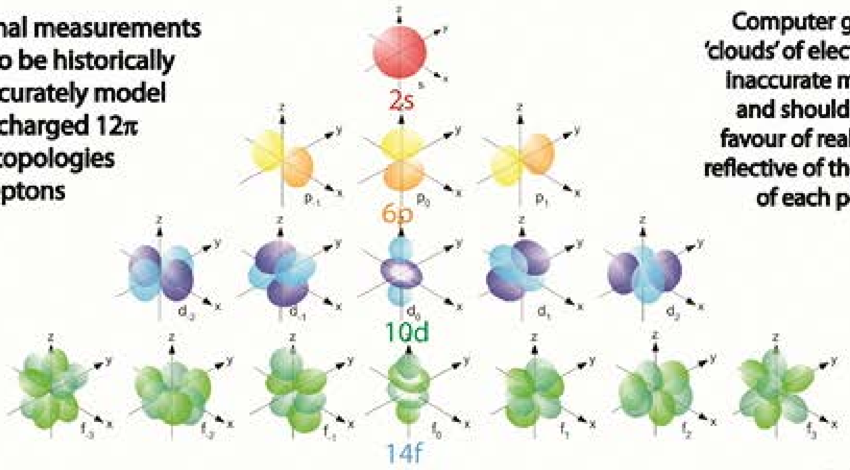

Computer generated plots of 'clouds' of electron probabilites are inaccurate mis-representations and should be abandoned in favour of realistic atomic models reflective of the charged geometry of each periodic element

v
RE
KE
P
KE
hv
Q
M
M
v²
c²

Every elemental atom can be viewed as a quantum carousel with a unique number of oscillating fans positioned around it.

Each fan has 3 blades and a fixed speed n[1-8] related to its height above ground level, AND the carousel is turning around on its axis

All periodic elements are made form Deuterium nuclei

0
[42-42]
N⁰
n
e⁻
M
P⁺

The baryonic energies of nuclei determines the energies of bound photo-electrons

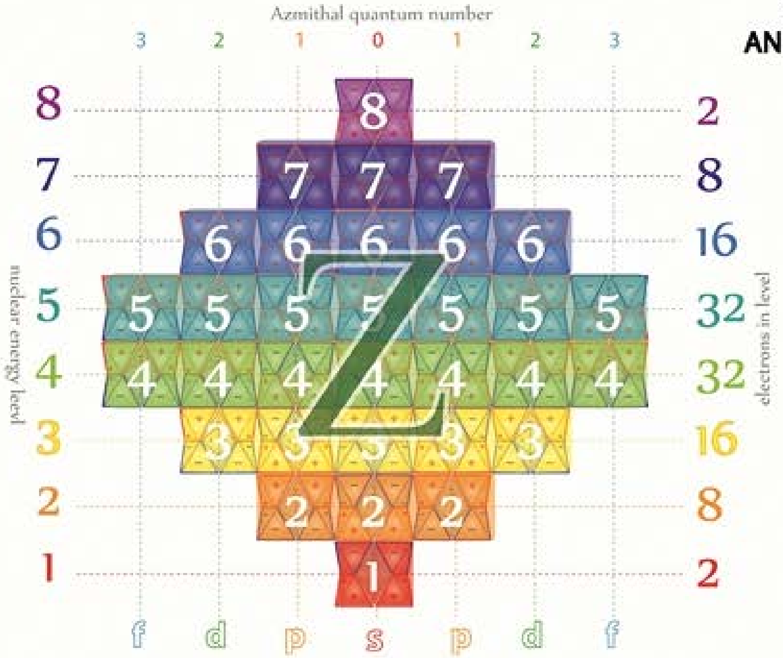

The current computer generated electron probability diagrams in popular use at present can now be show to be a misrepresentative model of mathematical modelling of electron sub-orbital energies

The nuclear quantum levels [n], intrinsic quantised angular momentum [h] and orbital angular momentum [l] of each electron bound within atomic nuclei are all the direct result of the Baryonic energies of the nuclei they are bound to

Each level of the quantum carousel can contain only a limited number of fans each running at a specific speed

Imagine trying to measure (or model) the motion of any 1 quantum scale blade while the carousel rotates

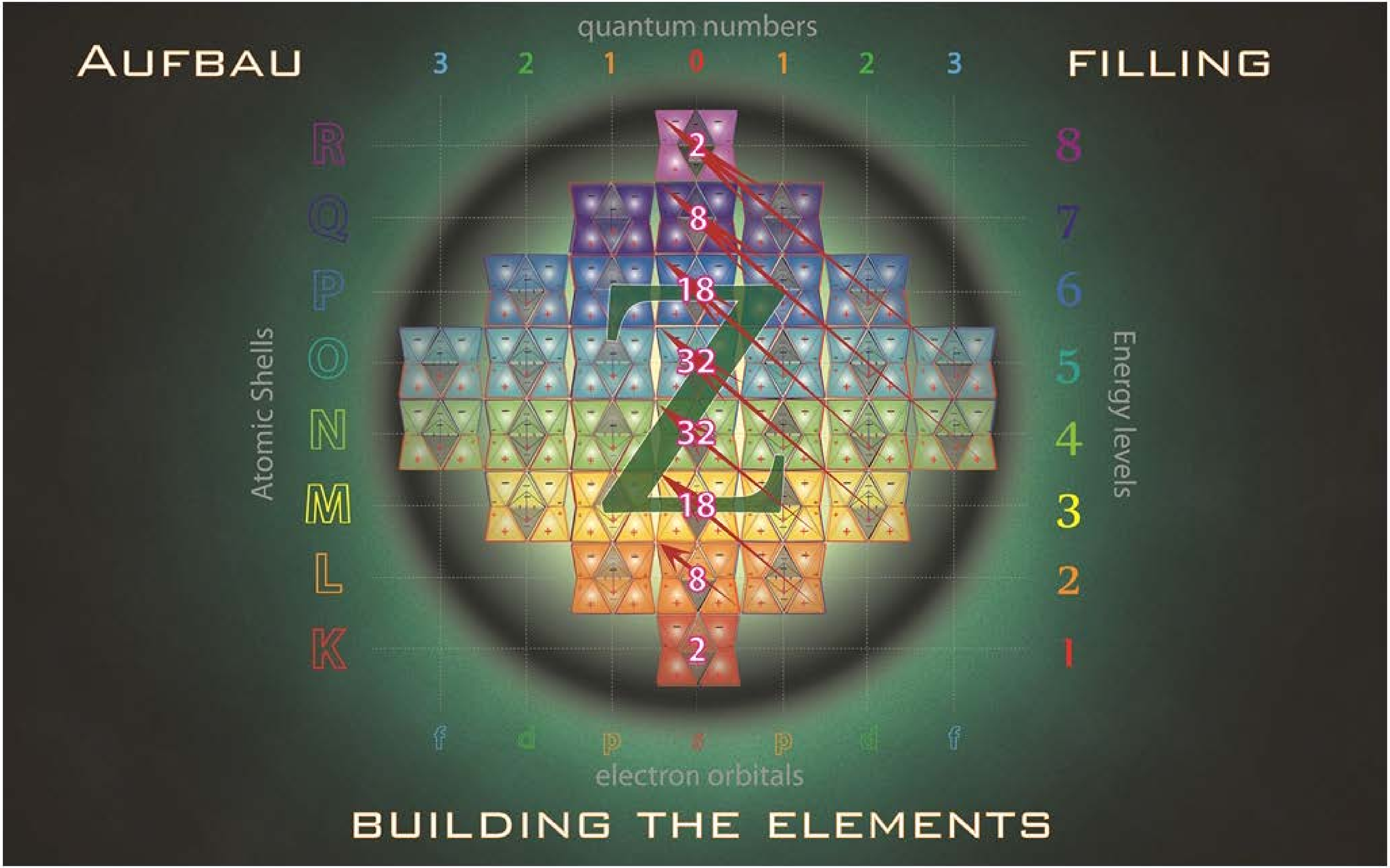
AUFBAU
FILLING
quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic Shells
8
7
6
5
4
3
2
1
Energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals
BUILDING THE ELEMENTS

Tetryonics 48.01 - aufbau geometry

Quantum Topologies

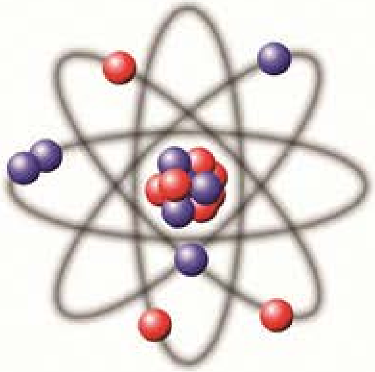

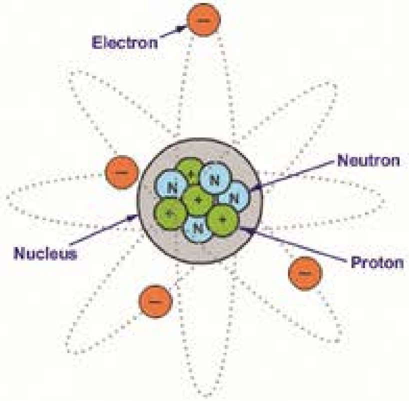
Electron
Neutron
Nucleus
Proton

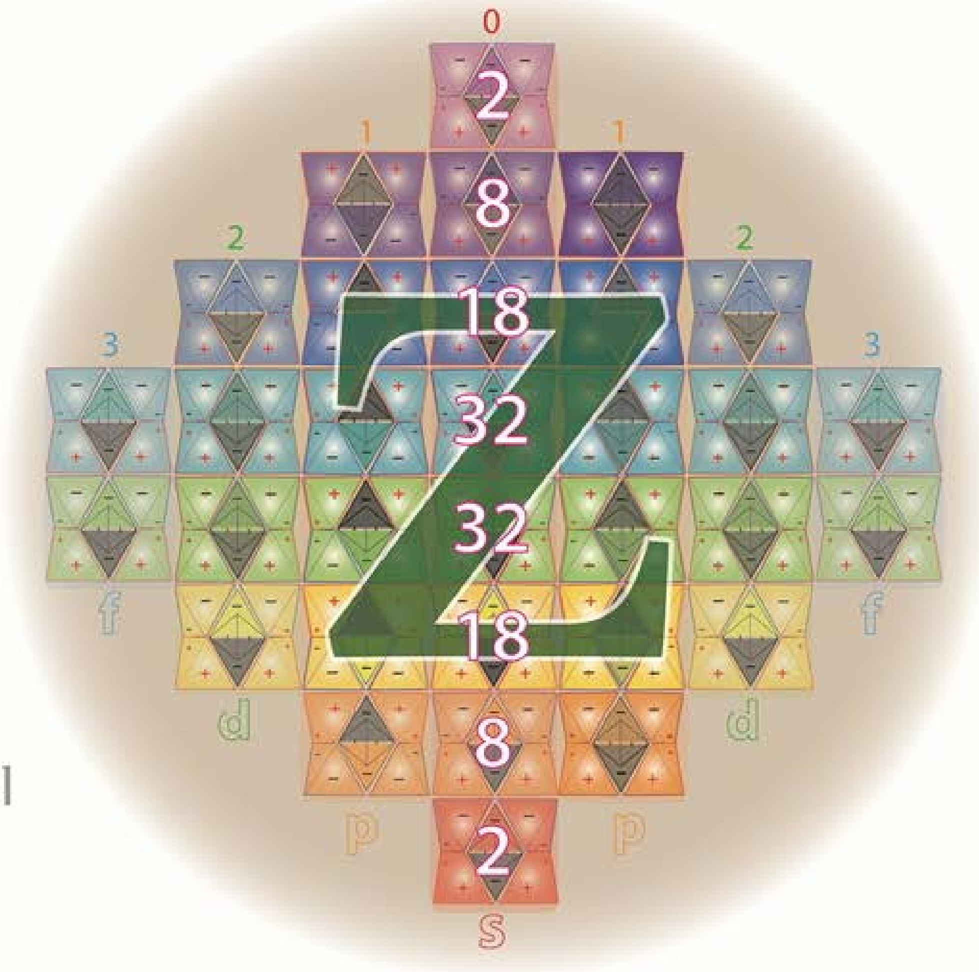
0
1
1
2
2
3
3
2
8
18
32
32
18
8
2
Z
f
f
d
d
p
p
s

Dalton Model

Quantum Model

Thomson Model

Rutherford Model

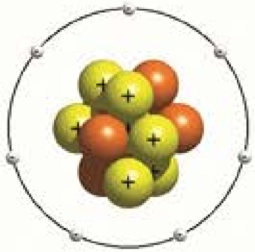

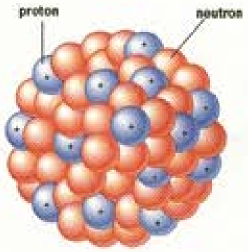
proton
neutron

Historically viewed as a spherical object
Tetryonic charge geometry has finally revealed the
true quantum topology of all atoms

Element numbers

The rule dictating how many nuclei form each Atomic shell is known as the Aufbau principle.

The physical and chemical properties of elements is determined by the atomic structure.
The atomic structure is, in turn, determined by the electrons and
which shells, subshells and orbitals they reside in.

The maximum periodic elemental number is 120

$$Z = \sum_{1,2,3,4,4,3,2,1} 2(n^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

Number of nuclei per level	Element #
2	120 119
8	118 111
18	110 93
32	92 61
32	60 29
18	28 11
8	10 3
2	2 1

The number of nuclei per quantum level is reflective of photonic energy levels and provides the foundational geometry for all of the periodic elements

The number of possible nuclei in each Quantum level follows aufbau principle 'numbers' which can be determined using the following summation formula

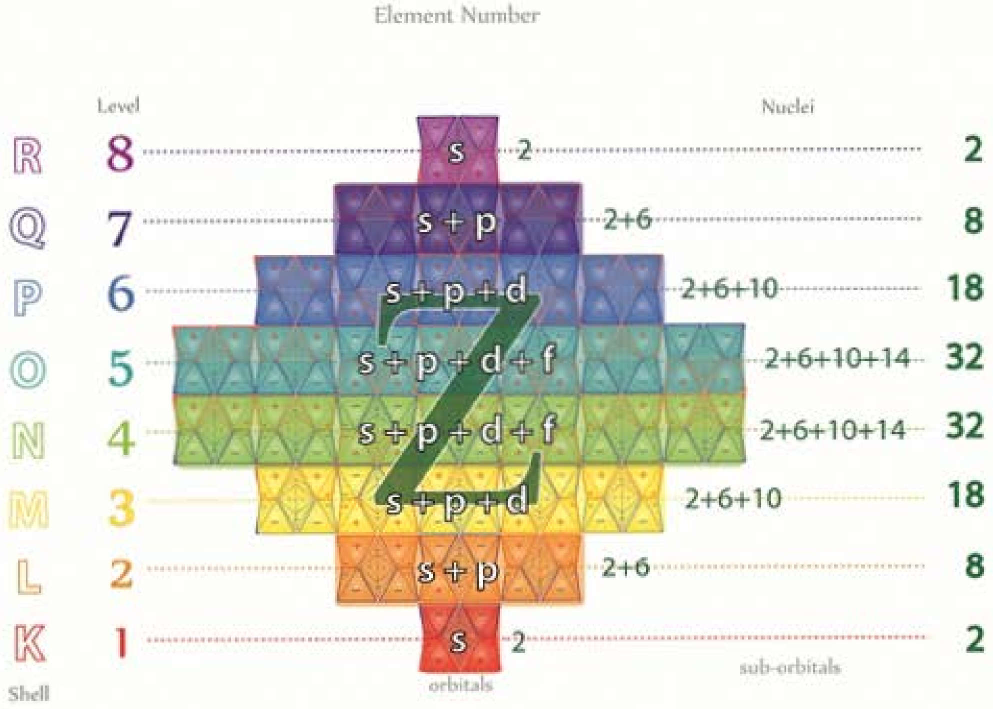
Element Number
Level
Nuclei
R 8 s 2 2
Q 7 s + p 2+6 8
P 6 s + p + d 2+6+10 18
O 5 s + p + d + f 2+6+10+14 32
N 4 s + p + d + f 2+6+10+14 32
M 3 s + p + d 2+6+10 18
L 2 s + p 2+6 8
K 1 s 2 2
Z
Shell
orbitals
sub-orbitals

Deuterium is the building block of all elements

Each element has equal numbers of Protons, electron & Neutrons with their stored mass-energies making up the molar masses of elements not excess neutron as currently modelled

Aufbau Principle

(Nuclei number and position)

1 Sub-Orbit (2 electrons max) ○ s

3 Sub-Orbits (6 electrons max) ○ p

○ d 5 Sub-Orbits (10 electrons max)

○ f 7 Sub-Orbits (14 electrons max)

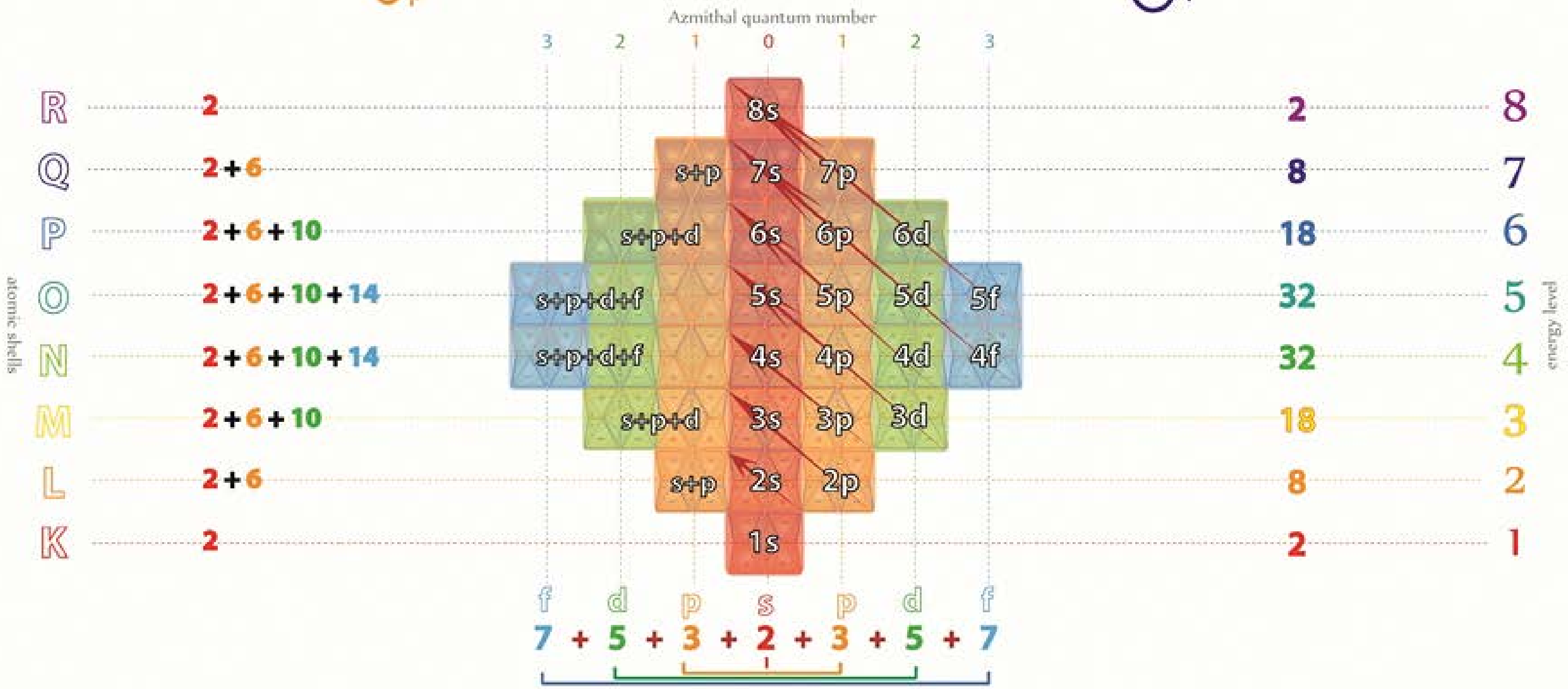
Azmithal quantum number
3 2 1 0 1 2 3
atomic shells
R 2
Q 2+6
P 2+6+10
O 2+6+10+14
N 2+6+10+14
M 2+6+10
L 2+6
K 2
8s
s+p 7s 7p
s+p+d 6s 6p 6d
s+p+d+f 5s 5p 5d 5f
s+p+d+f 4s 4p 4d 4f
s+p+d 3s 3p 3d
s+p 2s 2p
1s
2 8
8 7
18 6
32 5
32 4
18 3
8 2
2 1
energy level
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7

	2		6						10										14															
Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32		per level
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	Quantum number	

sub-orbitals

A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.

The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements

Tetryonics 48.04 - Aufbau Principle

Wolfgang Pauli

(25 April 1900 – 15 December 1958)

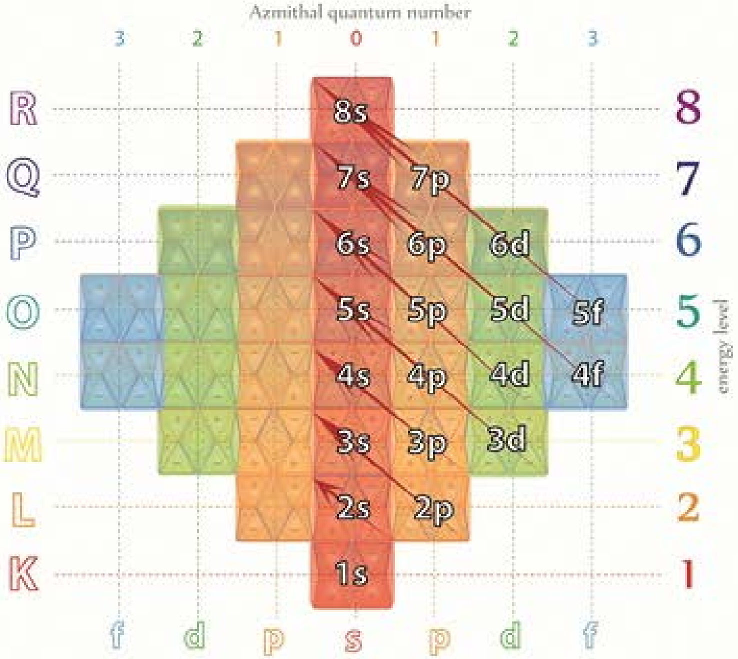
Azmithal quantum number
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
energy level
8s
7s 7p
6s 6p 6d
5s 5p 5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
f d p s p d f

Friedrich Hermann Hund

(4 February 1896 - 31 March 1997)

The orbitals of lower energy are filled in first with the electrons and only then are the higher energy orbitals filled.

aufbau electron orbital filling

Orbitals of equal energy are each occupied by one electron before any orbital is occupied by a second electron, and all electrons in singly occupied orbitals must have the same spin state

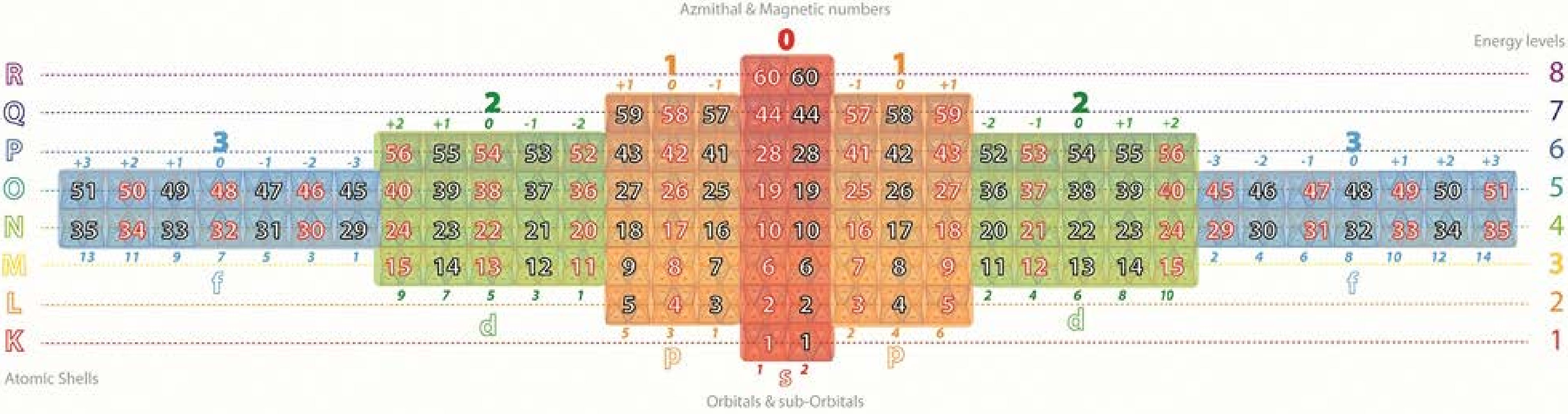
Azmithal & Magnetic numbers
Energy levels
R Q P O N M L K
8 7 6 5 4 3 2 1
0 1 1 2 2 3 3
f d p s p d f
Atomic Shells
Orbitals & sub-Orbitals

Aufbau construction

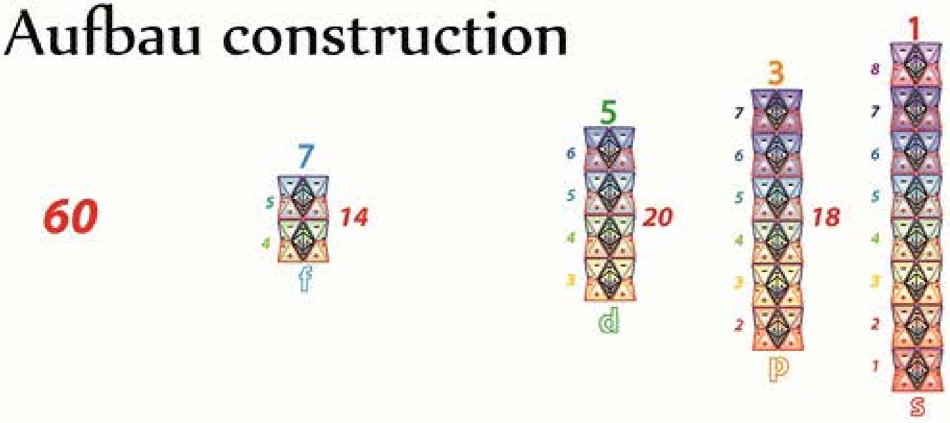
60
7
14
5
4
f
5
20
6
5
4
3
d
3
18
7
6
5
4
3
2
p
1
8
8
7
6
5
4
3
2
1
s

Proton base Hemisphere

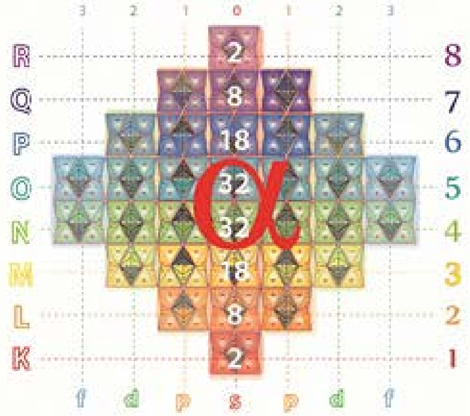
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
2
8
18
32
32
18
8
2
α
8
7
6
5
4
3
2
1
f d p s p d f

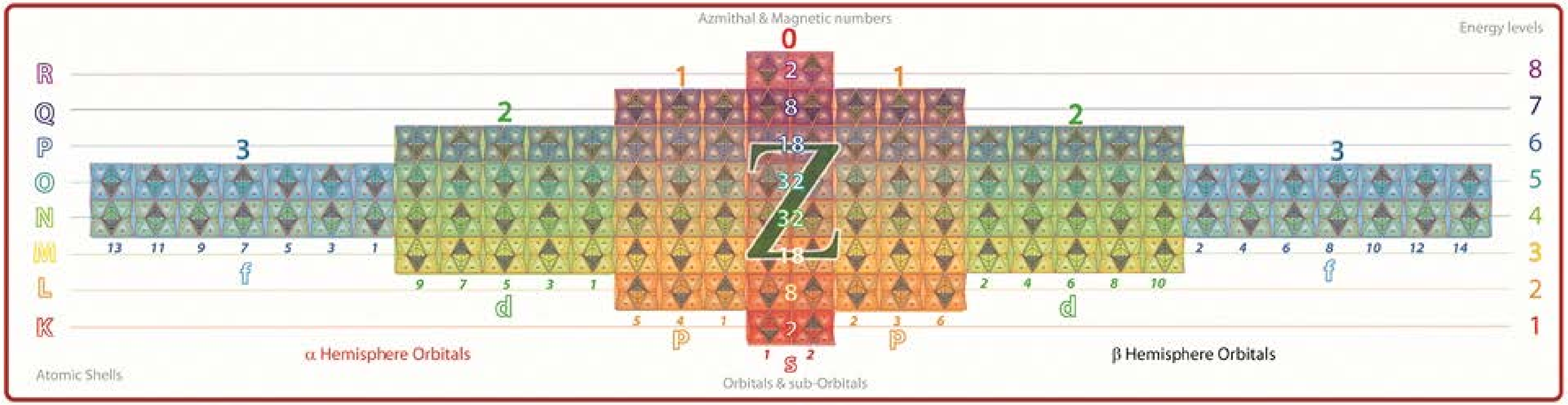
Azmithal & Magnetic numbers
Energy levels
R
Q
P
O
N
M
L
K
0
1
1
2
2
3
3
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 4 1
p
1 s 2
2 3 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
8
7
6
5
4
3
2
1
α Hemisphere Orbitals
β Hemisphere Orbitals
Atomic Shells
Orbitals & sub-Orbitals

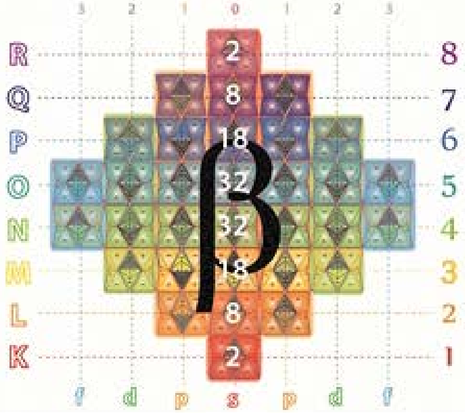
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
β
8
7
6
5
4
3
2
1
f d p s p d f

Neutron base Hemisphere

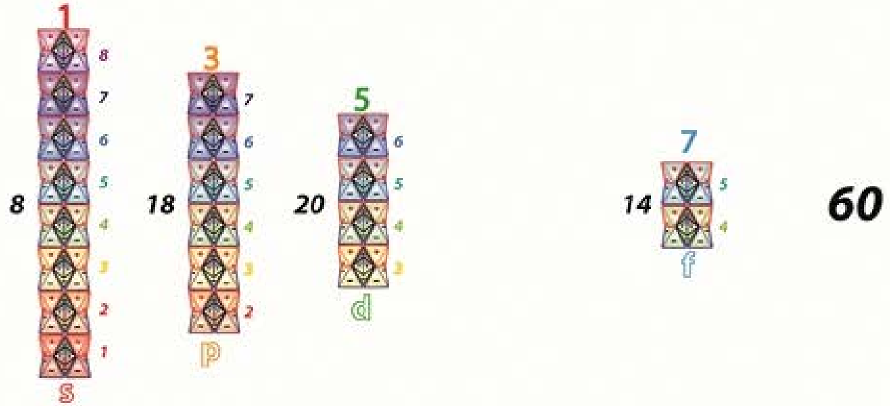
1
8
8
7
6
5
4
3
2
1
s
3
18
7
6
5
4
3
2
p
5
20
6
5
4
3
d
7
14
5
4
f
60

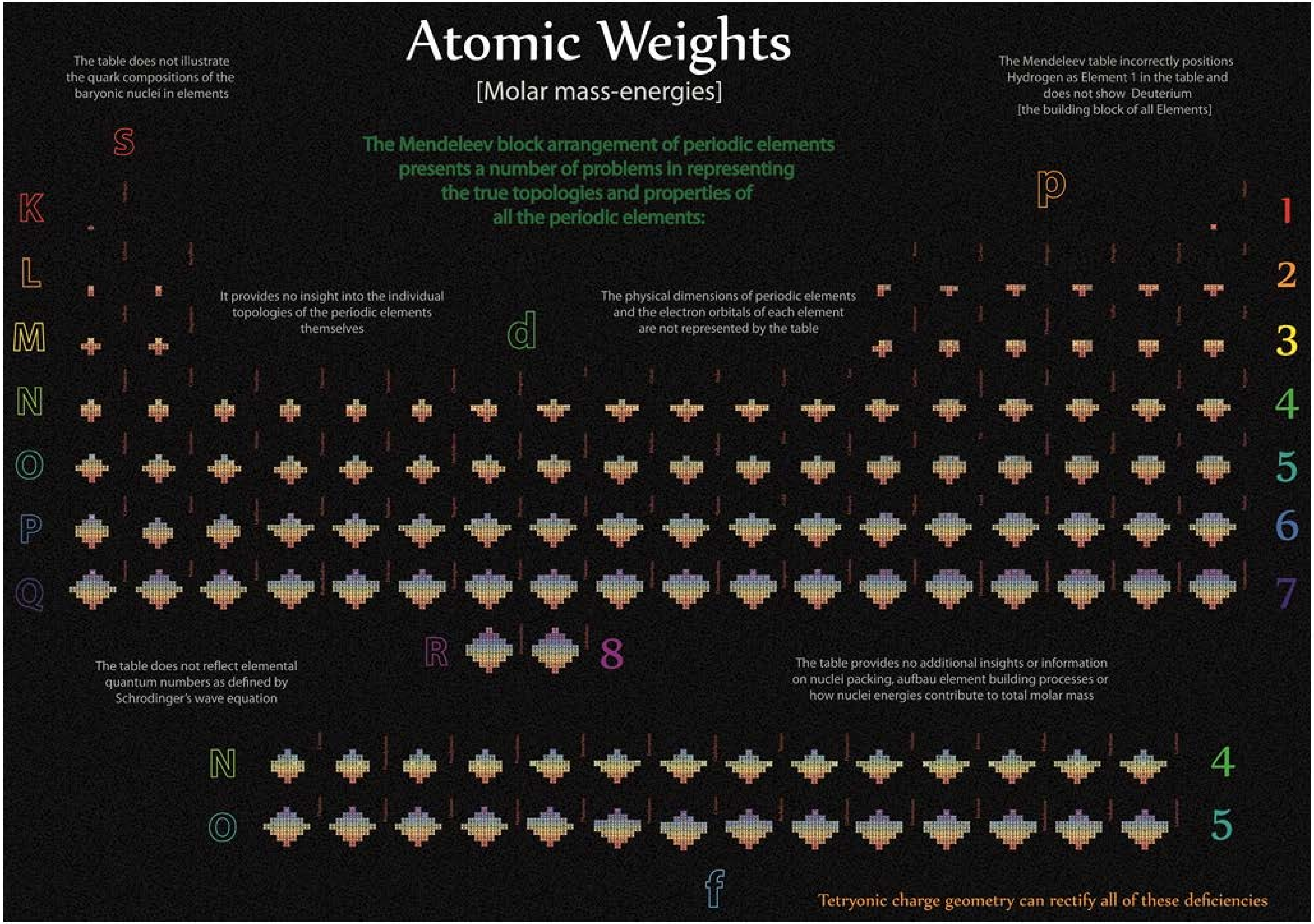

Atomic Weights

[Molar mass-energies]

The table does not illustrate the quark compositions of the baryonic nuclei in elements

The Mendeleev table incorrectly positions Hydrogen as Element 1 in the table and does not show Deuterium [the building block of all Elements]

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

It provides no insight into the individual topologies of the periodic elements themselves

The physical dimensions of periodic elements and the electron orbitals of each element are not represented by the table

The table does not reflect elemental quantum numbers as defined by Schrodinger's wave equation

The table provides no additional insights or information on nuclei packing, aufbau element building processes or how nuclei energies contribute to total molar mass

Tetryonic charge geometry can rectify all of these deficiencies

Copyright ABRAHAM [2008] - All rights reserved

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

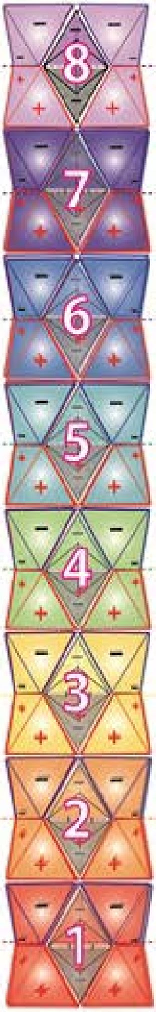

Baryon mass-energies n			electron KEM field hv^2
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13
Compton frequencies 930 MeV		**1.2e20** **electron rest Matter is velocity invariant** 496 keV	**spectral frequencies** 13.6 eV

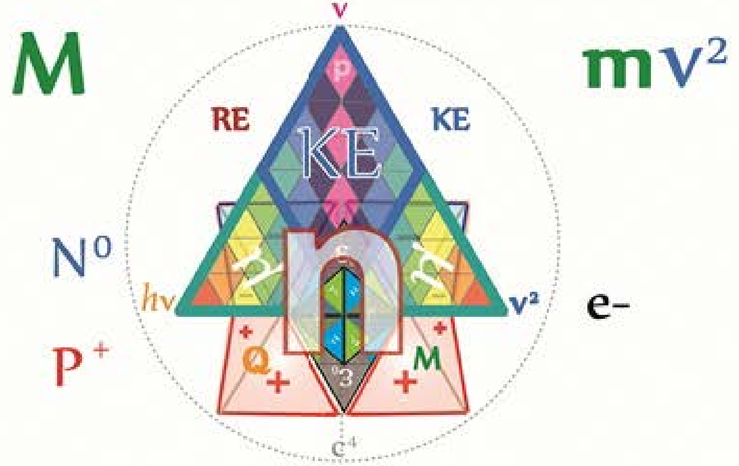

The nuclei forming each atomic shell have specific mass-energy quanta

Baryon rest masses lepton rest mass KEM

$$\overset{8}{\underset{1}{n}}\left[[72(n)^2]+[12e19]+[m_e v^2]\right]$$

Deuterium mass-energy per shell

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Tetryonics 49.02 - Quantum level nuclei masses

Hydrogen Ionisation Energies [/n²]

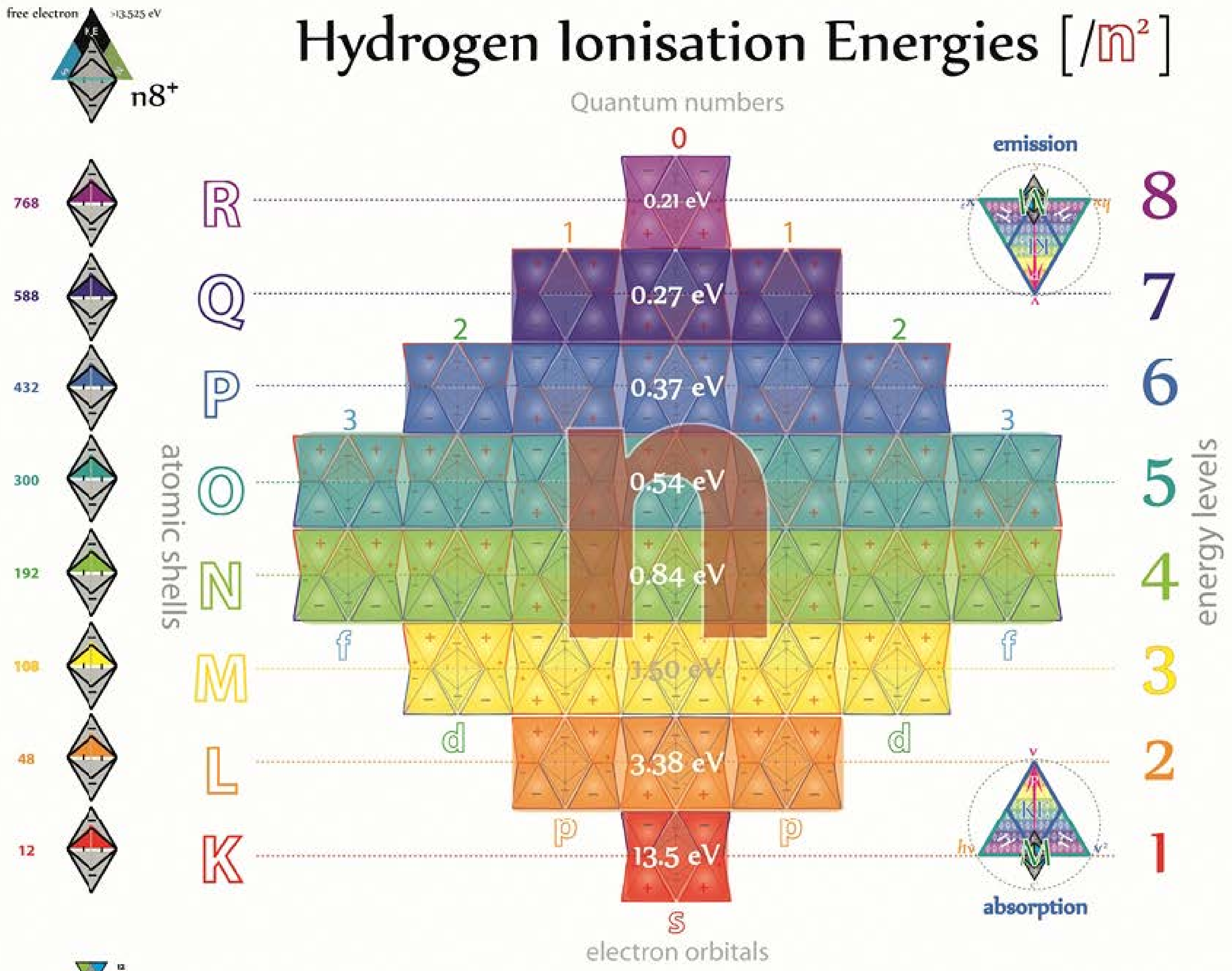
free electron
>13.525 eV
K.E
n8+
Quantum numbers
0
1
1
2
2
3
3
f
f
d
d
p
p
s
electron orbitals
768
588
432
300
192
108
48
12
R
Q
P
O
N
M
L
K
atomic shells
0.21 eV
0.27 eV
0.37 eV
0.54 eV
0.84 eV
1.50 eV
3.38 eV
13.5 eV
n
emission
absorption
hv
8
7
6
5
4
3
2
1
energy levels
n0
12

13.6eV
Free electron

8	5.107765145 e13
7	6.671366720 e13
6	9.078047137 e13
5	1.307587877 e14
4	2.043106058 e14
3	3.632188548 e14
2	8.172424234 e14
1	3.268969693 e15

Planck quanta

Spectral line series frequencies
[Hydrogen]

accelerating electrons $Mv^2 = KEM = hv^2$ produce spectral lines

Copyright ABRAHAM [2008] - All rights reserved

Redefining Atomic weights

Atomic weight (symbol: Ar) is a dimensionless physical quantity,
the ratio of the average mass of atoms of an element (from a given source) to
1/12 of the mass of an atom of carbon-12 (known as the unified atomic mass unit)

Carbon 12

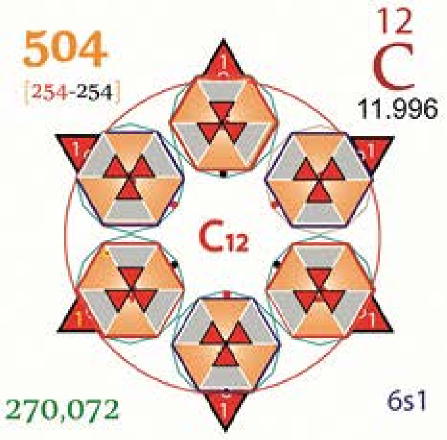

6 Protons [24-12]
6 Neutrons [18-18] n1
6 electrons [0-12]

1/12 C_{12}

$A_r = 22{,}506$

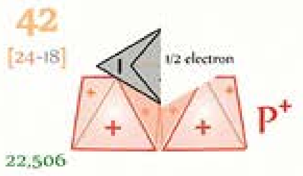

1.660096209e-27 kg

1 Proton [24-12]
.5 electron 0-6 n1

The 'unified atomic mass unit' currently in use
is known to be inaccurate and must be corrected
in order to bring clarity & increased accuracy
to the atomic weights of all elements

$$\boxed{A_r = 22{,}512 \text{ Hydrogen}}$$

Defining Hydrogen as having an exact
atomic Planck mass of 22512_n *quanta*
provides uniformity with Tetryonics

Deuterium is the building block of all elements in the period table

Unified atomic Matter unit

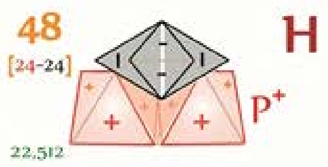

1.660538783e-27 kg

1 Proton [24-12]
1 electron [0-12] n1

$$\boxed{A_r = 45{,}012 \text{ Deuterium}}$$

Defining Deuterium as having an exact
atomic Planck mass of 45012_n *quanta*
reflects the true charged geometries
of all Elements & their topologies

D

1/6 C

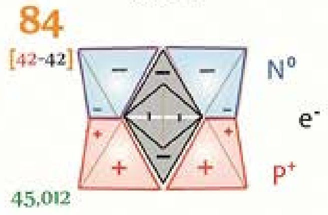

3.320192418e-27 kg

1 Proton [24-12]
1 Neutron [18-18] n1
1 electron [0-12]

Planck mass-energy units

NAu = 0.001 kg

48
[24-24]
22,512

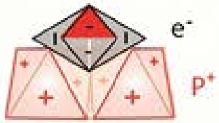

^1H

1.660538841 e-27 kg

One Da is approximately equal to the mass of one proton or one neutron

36
[18-18]
22,500

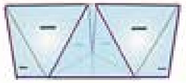

N^0

0.499866702 H

1.659653693 e-27 kg

36
[24-12]
22,500

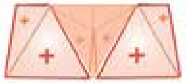

P^+

0.499866702 H

Deuterium is the building block of all elements

84
[42-42]
45,012

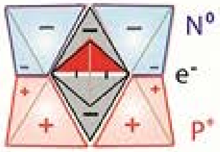

^2H

3.320192534 e-27 kg

The unified atomic mass unit (symbol: u) or Dalton (symbol: Da)
is a unit that is used for indicating mass on an atomic or molecular scale

270,072

1/12 the mass of a C12 graphene atom
at rest in its electronic ground state

1.660538782(83)x10–27 kg

22,506

is an inaccurate means of determining the
exact rest mass of a Hydrogen atom

22,512

Carbon 12 has 270,072n planck quanta
(270,072 / 12 = 22,506)

Hydrogen has a mass of 22,512n (22,500+12)
requiring all mass to be calculated directly using the
Planck mass-energy quantum (.001kg / N_A / 22,512)
& Tetryonic charge geometries

Using Tetryonic theory to define
n Planck mass = 7.376238634 x 10^{-32} Kg
(see Tetryonics QM 15.04)

exact atomic rest masses for all
particles, elements and compounds
can be determined directly from atomic theory

N_A = 6.02214179 e23

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is "mol".

NA(12u) = 0.012 kg

504
[254-254]

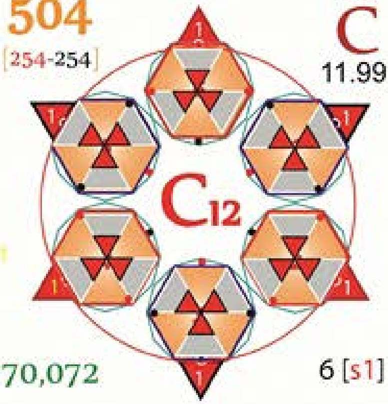

^{12}C
11.996

270,072

6 [s1]

6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1

Carbon has a number of differing atomic configurations (allotropes)

6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1-2

504
[252-252]

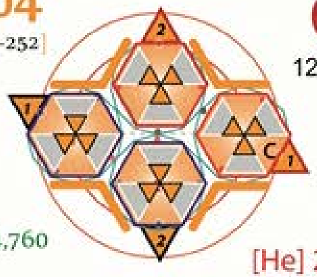

^6C
12.6493

284,760

[He] 2s2 2p2

Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies

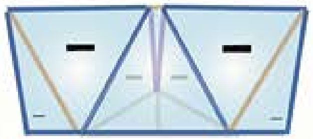

1.659653693 e-27 kg

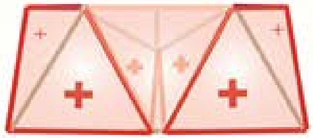

8.851486361 e-31 kg

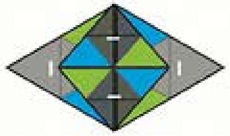

2.41109611 e-35 kg

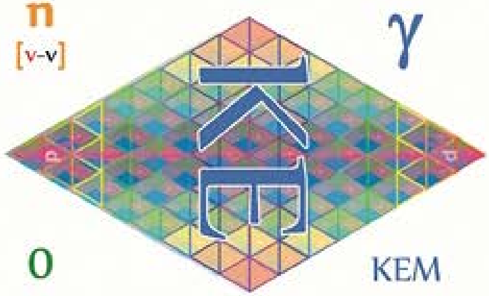

Baryons have 2,25e23 Planck quanta comprising their rest Matter topologies

[930.974 MeV]

Δ 1875 x

[496.5 keV]

Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies

[496.5 keV]

Δ 36,711 x

[13.6 eV]

Photons are planar geometries [Matter-less]
(purely Kinetic mass-Energy and momenta)

The Lyman alpha spectral line mass-energy contribution to the mass of a Deuterium nucleus is negligible

Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements

D^+

[1.862 GeV]

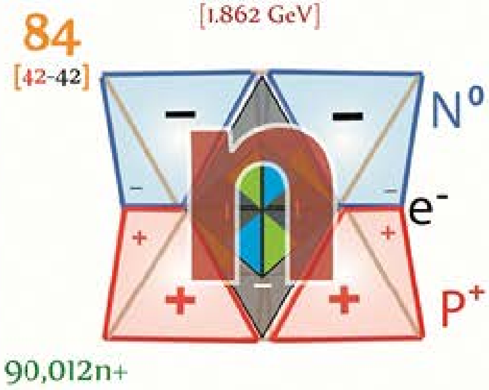

Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]

Photons are 2π charge mass-energy geometries

Ionisation energies

Fractional quantum differentials

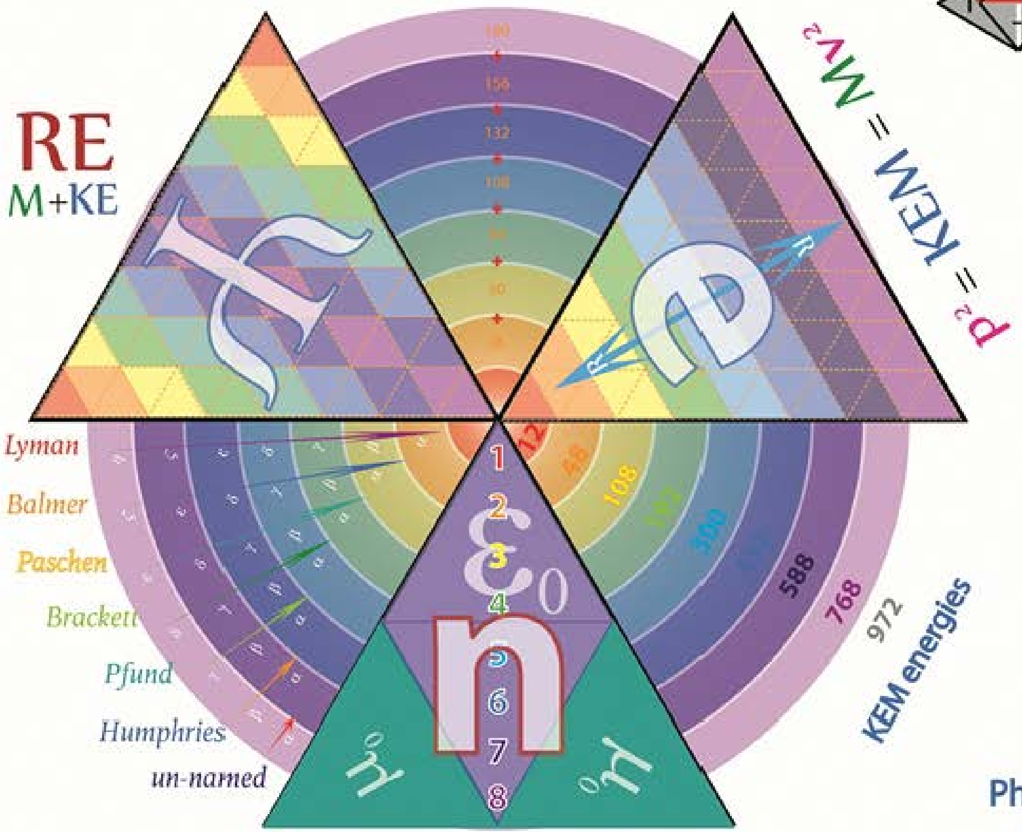
RE
M+KE
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
p² = KEM = Mv²
KEM energies
12
48
108
192
300
432
588
768
972
1
2
3
4
5
6
7
8
ε0
n
h
e
R
μ0

$$hf = \Delta M v = \Delta p$$

spectral lines are produced by accelerating electrons

c^2

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^4

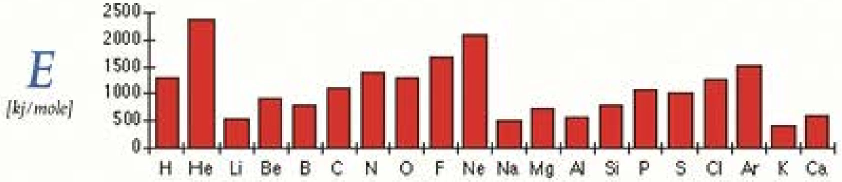

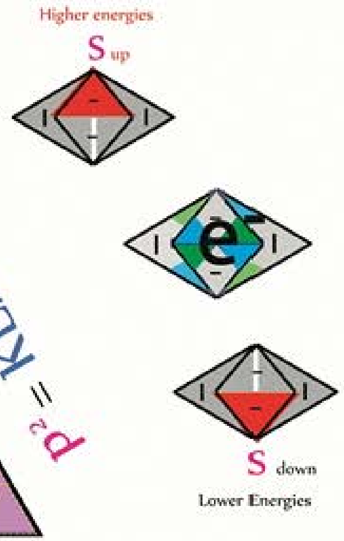
Higher energies
S up
e⁻
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

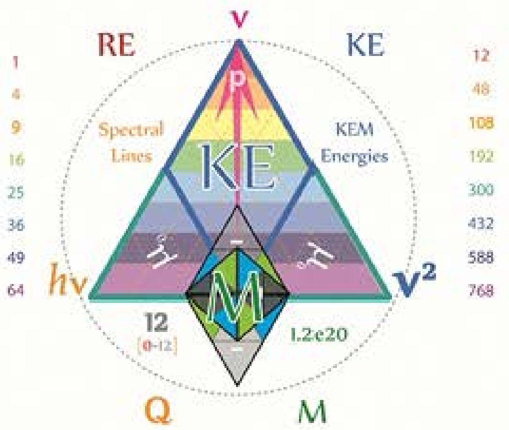
v
p
RE
KE
Spectral Lines
KEM Energies
KE
M
hv
v²
12
[0-12]
1.2e20
Q
M
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

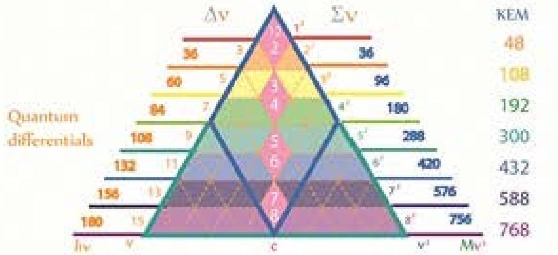
Δv
Σv
KEM
48
108
192
300
432
588
768
Quantum differentials
36
60
84
108
132
156
180
36
96
180
288
420
576
756
hv
v
c
v²
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Elementary ionisation energies

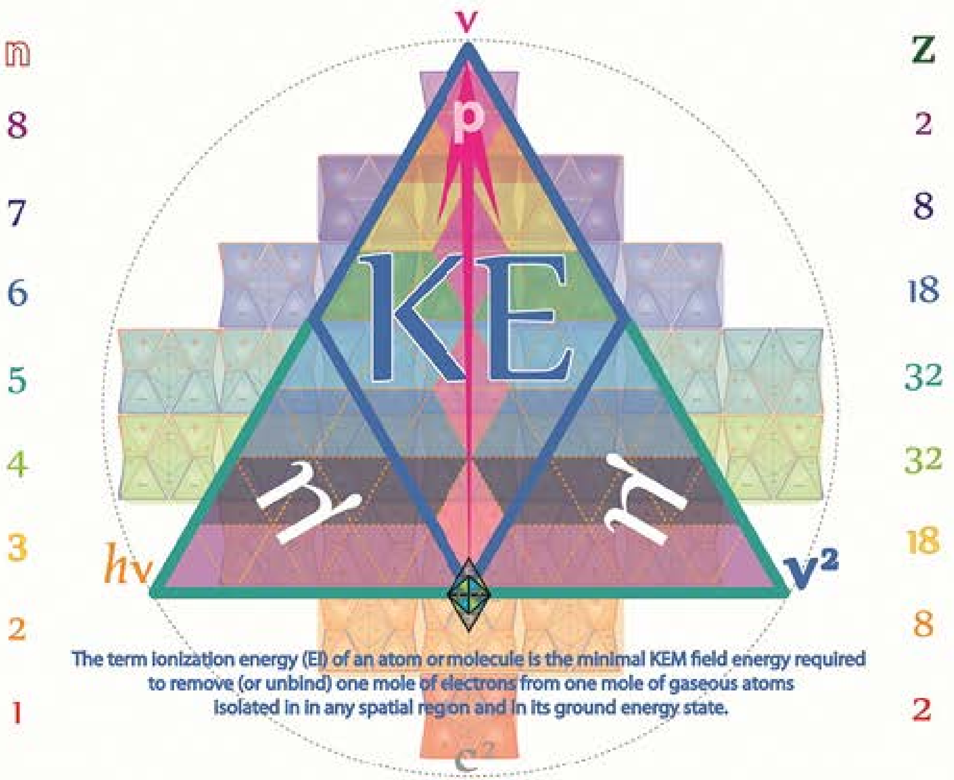

The term ionization energy (EI) of an atom or molecule is the minimal KEM field energy required to remove (or unbind) one mole of electrons from one mole of gaseous atoms isolated in in any spatial region and in its ground energy state.

The term "ionization energy" is sometimes used as a name for the work needed to remove (or un-bind) the highest energy photoelectron from an atom or molecule.

However, due to interactions with surfaces, this value differs from the ionization energy of the atom or molecule in question when it is located by itself in free space.

So, in the case of surface-adsorbed atoms and molecules, it may be better to use the more general term "electron binding energy", in order to avoid confusion.

Both these names are also sometimes used to describe the work needed to remove an electron from a "lower" orbital (i.e., not the topmost orbital) for both free and adsorbed atoms; in such cases it is necessary to specify the orbital from which the electron has been removed

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

Every electron in each elementary orbit has a unique ionisation energy

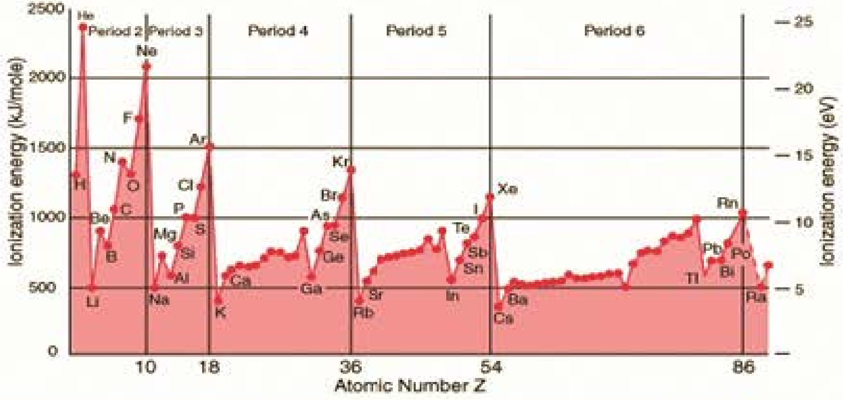

Element number

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	per shell	Energy level
R	119	120																															2	8
Q	87	88	113	114	115	116	117	118																									8	7
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															18	6
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	32	5
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	32	4
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															18	3
L	3	4	5	6	7	8	9	10																									8	2
K	1	2																															2	1
Shells	s				p								d													f								
	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14		

Orbitals

sub-orbitals

Copyright ABRAHAM [2008] - All rights reserved

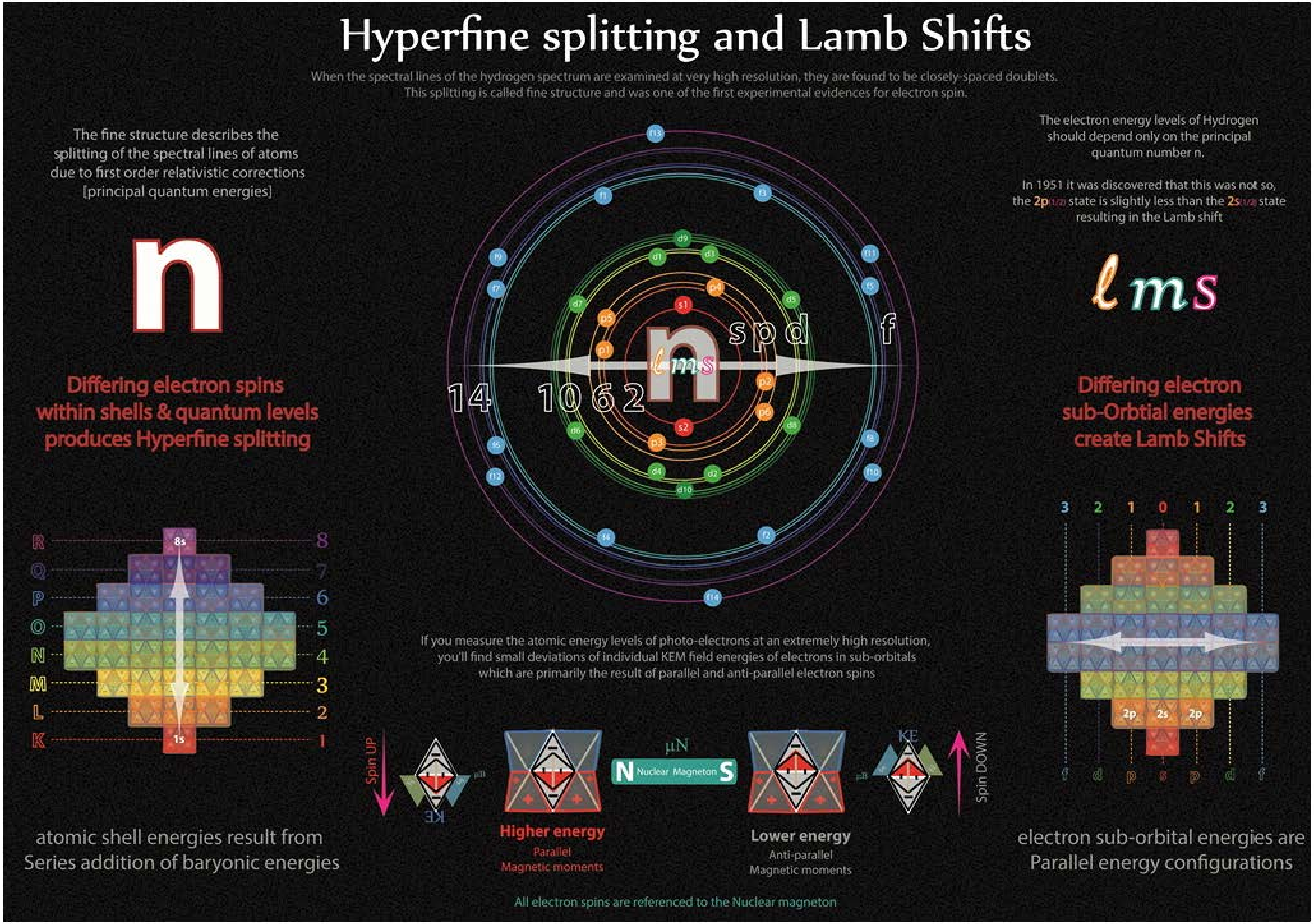

Hyperfine splitting and Lamb Shifts

When the spectral lines of the hydrogen spectrum are examined at very high resolution, they are found to be closely-spaced doublets.
This splitting is called fine structure and was one of the first experimental evidences for electron spin.

The fine structure describes the splitting of the spectral lines of atoms due to first order relativistic corrections [principal quantum energies]

The electron energy levels of Hydrogen should depend only on the principal quantum number n.

In 1951 it was discovered that this was not so, the 2p$_{(1/2)}$ state is slightly less than the 2s$_{(1/2)}$ state resulting in the Lamb shift

n
ℓ ms
f13
f1
f3
f9
f11
f7
f5
d9
d1
d3
d7
d5
p4
p5
s1
p1
s p d
f
n
ℓ ms
14
10 6 2
p2
p6
s2
d6
d8
p3
f6
f8
d4
d2
d10
f12
f10
f4
f2
f14

Differing electron spins within shells & quantum levels produces Hyperfine splitting

Differing electron sub-Orbtial energies create Lamb Shifts

R
Q
P
O
N
M
L
K
8s
1s
8
7
6
5
4
3
2
1
3
2
1
0
1
2
3
2p
2s
2p
f
d
p
s
p
d
f

If you measure the atomic energy levels of photo-electrons at an extremely high resolution, you'll find small deviations of individual KEM field energies of electrons in sub-orbitals which are primarily the result of parallel and anti-parallel electron spins

Spin UP
KE
μB
μN
N Nuclear Magneton S
μB
KE
Spin DOWN

Higher energy
Parallel
Magnetic moments

Lower energy
Anti-parallel
Magnetic moments

atomic shell energies result from Series addition of baryonic energies

electron sub-orbital energies are Parallel energy configurations

All electron spins are referenced to the Nuclear magneton

Tetryonics 49.09 - Hyperfine splitting and Lamb Shifts

Copyright ABRAHAM [2008] - All rights reserved

M_0

3D Matter topologies are comprised of charged 2D mass-energies

Energy per second²

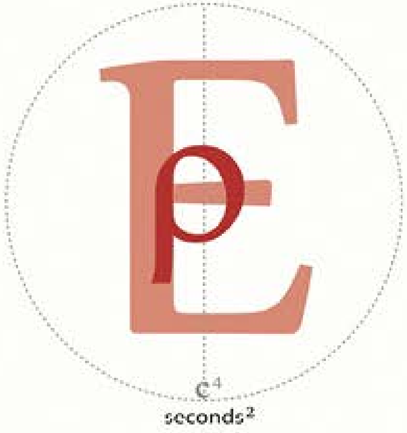

c^4
seconds²

RE

Relativistic mass-ENERGY-Matter

Relativity fails at the foundational level to explain and differentiate between mass-ENERGY and Matter in physical systems

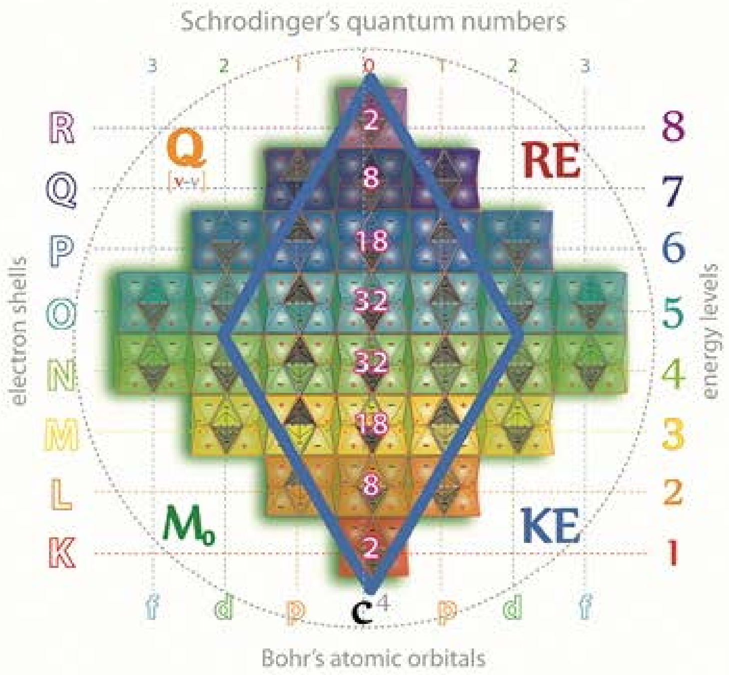

Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M_0
KE
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

Einstein's relativistic [Lorentz corrected] stress energy tensor aggregates all forms of energy into a single energy density gradient

KE

2D equilateral mass-energies are euclidean geometries

Energy per second

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

mass-Matter

standing wave mass-energiescreate
the material substance of all chemical elements

atomic energies

$$\frac{Tm\mathfrak{n}}{c^4} + \frac{mv^2}{c^2}$$

electron spins

3D rest Matter + Lorentz corrected 2D Kinetic Energies = total Relativistic Energies

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

mass-energies

radiant planar mass-energies create
EM fields, spectral lines & chemical interactions

Deuterium is the building block of all Elements (save Hydrogen)

The geometry of any Element is determined by its Charge

The Molar Weight of any Element is a measure of its standing wave mass-energies

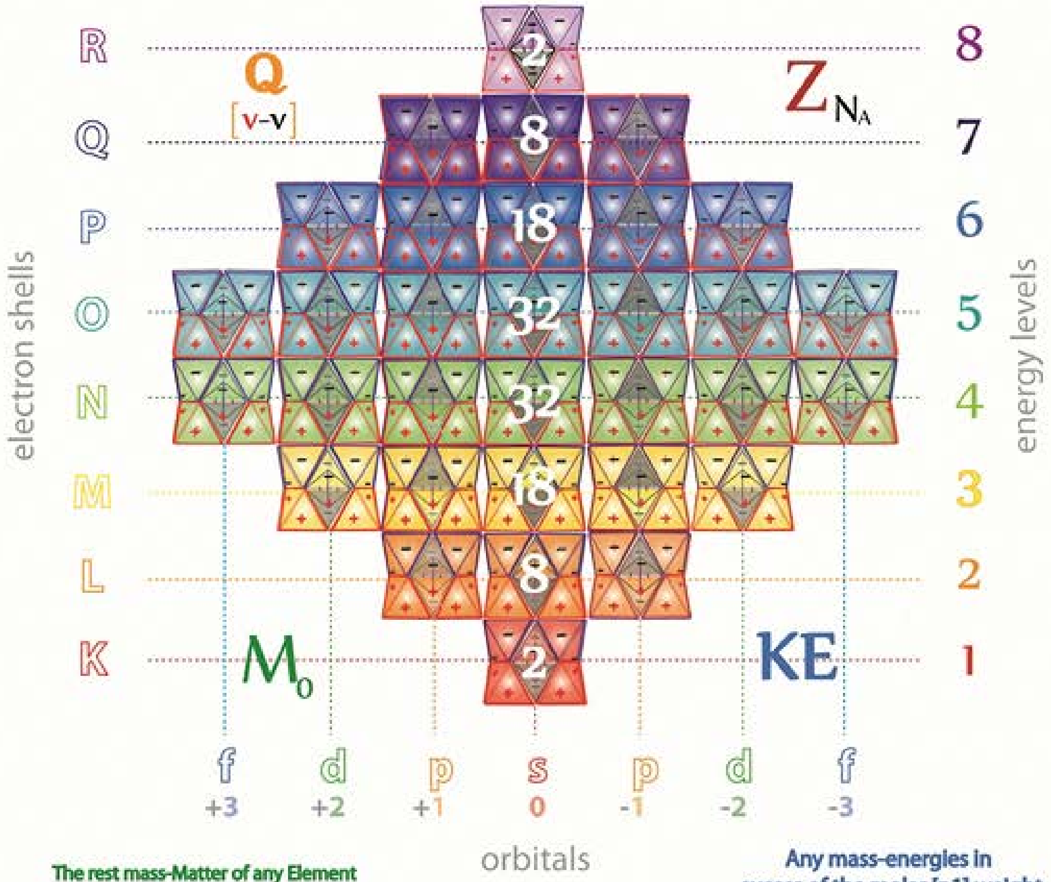

The rest mass-Matter of any Element is determined by the total number of quanta making up the Protons, Neutrons and electrons that comprise them (in their respective energy levels)

Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies

Important point to note:

The Kinetic Energy difference between any Element's total [n1] Deuteron mass-energies and its Molar mass has historically been incorrectly explained as resulting from an excess number of Neutrons in the atom it is not. Z# = (number of Protons = number of electrons = number of Neutrons)

Elementary mass-Matter

$\sum_{K}^{R}$

	n per nuclei 1e19v = n	
2	2 nuclei [74,496 ea]	120 Unbinilium 119 Ununennium
	+	
8	8 nuclei [69,780 ea]	118 Ununoctium 87 Francium
	+	
18	18 nuclei [65,232 ea]	112 Copernicium 55 Caesium
	+	
32	32 nuclei [60,852 ea]	102 Nobelium 37 Rubidium
	+	
32	32 nuclei [56,640 ea]	70 Ytterbuim 19 Potassium
	+	
18	18 nuclei [52,596 ea]	30 Zinc 11 Sodium
	+	
8	8 nuclei [48,720 ea]	10 Neon 3 Lithium
	+	
2	2 nuclei [45,012 ea]	2 Helium 1 Deuterium

The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei

Aufbau

$$Z\#\left[\begin{array}{ll} z\ \text{Protons} & [24\text{-}12] \\ z\ \text{Neutrons} & [18\text{-}18] \\ z\ \text{electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

(ie Calcium [20] = 2+8+10 n level Deuterium nuclei)

Periodic Harmonic motions

$x = A\cos(\omega t + \varphi)$

Much of the math in of modern physics is predicated on the assumption that π [where it appears] is related to the properties of a circle

$F = -kx$

Circular motion

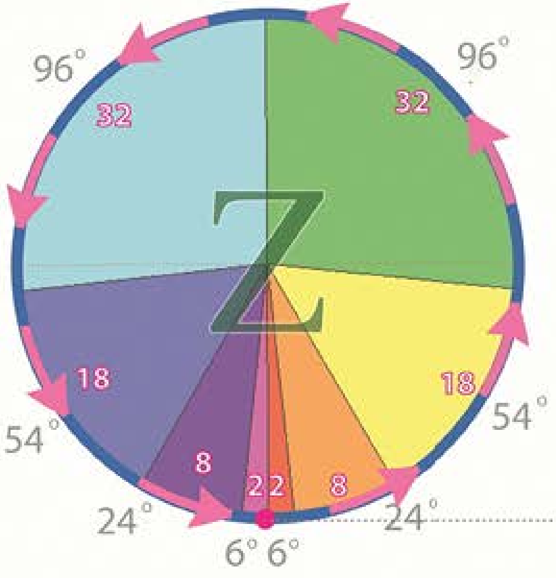
96°
96°
32
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

circular harmonic motion

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

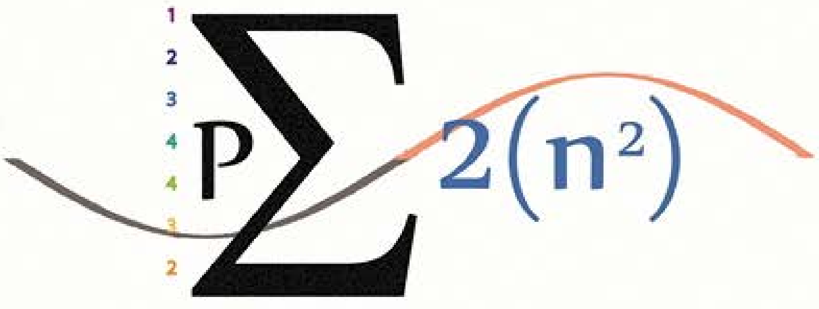

$$P\sum 2(n^2)$$

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

Linear motion

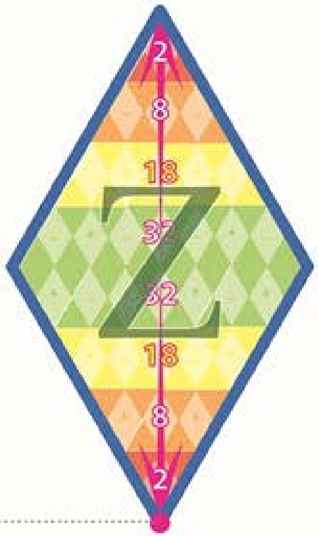
2
8
18
32
Z
32
18
8
2

simple harmonic motion

Principal Quantum Numbers

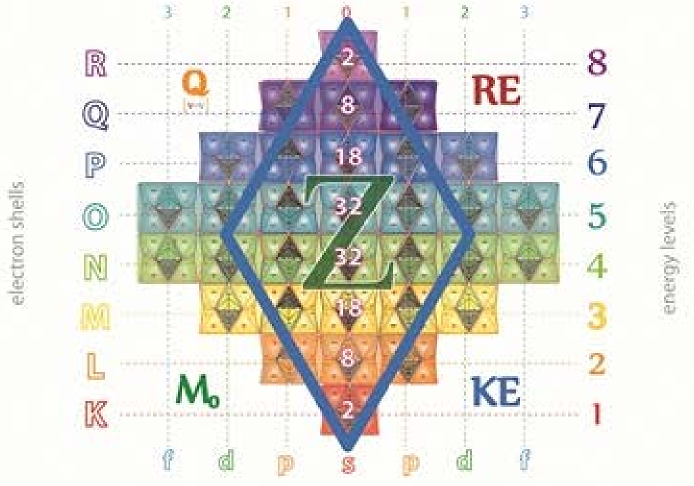
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M₀
KE
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f

Sub-orbitals

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

Copyright ABRAHAM [2008] - All rights reserved

$$P\sum_{K}^{R} Z$$

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of Z[n²] energy level Deuterium nuclei that form the periodic elements

$$\sum_{K}^{R}$$

nuclei per shell

STEP ONE

Periodic summation follows the atomic shell electron config

$$\sum_{1}^{120} Z$$

element number

STEP TWO

Periodic elements build up following the aufbau sequence

$$P\sum_{K}^{R} 2(x^2) =$$

1 2 3 4 4 3 2 1

Each atomic shell can hold only a fixed number of deuterium nuclei

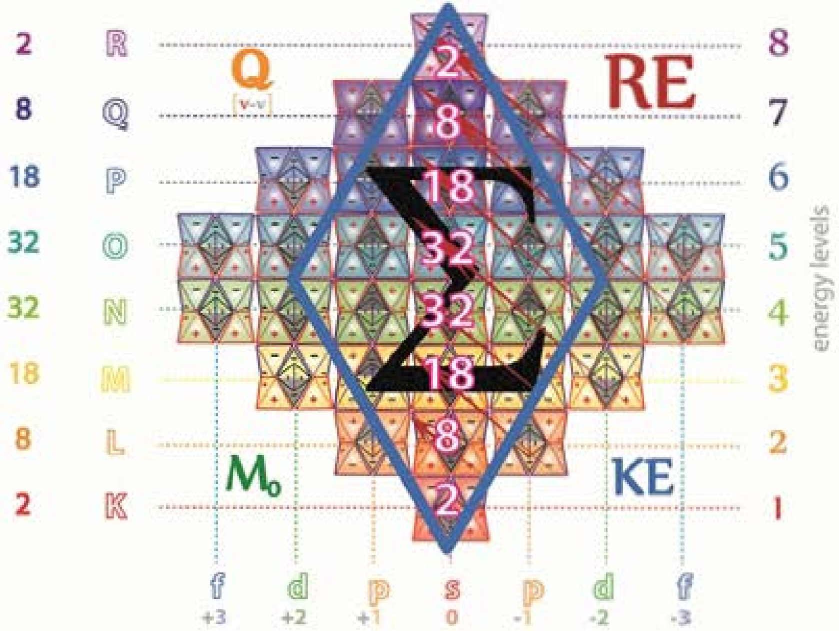

$P\sum R = 2$	2 nuclei [74,496 ea]	120	Unbinilium
+			
$P\sum Q = 8$	8 nuclei [69,780 ea]	118	Ununoctium
+			
$P\sum P = 18$	18 nuclei [65,232 ea]	110	Darmstadtium
+			
$P\sum O = 32$	32 nuclei [60,852 ea]	92	Uranium
+			
$P\sum N = 32$	32 nuclei [56,640 ea]	60	Neodymuim
+			
$P\sum M = 18$	18 nuclei [52,596 ea]	28	Argon
+			
$P\sum L = 8$	8 nuclei [48,720 ea]	10	Neon
+			
$P\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n² energy] deuterium nuclei

Z# [z Protons [24-12]
z Neutrons [18-18]
z electrons [0-12]] n1-8

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

1 R
2
3
4
4
3
2
1 K

$${}^{R}_{K}\text{P}\sum 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	
Hydrogen	0	

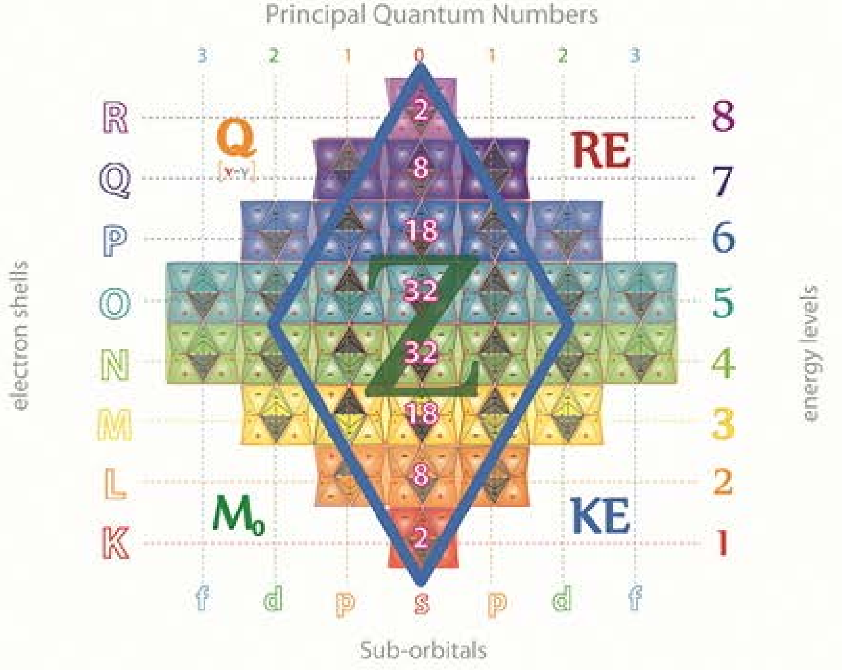
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-γ]
RE
2
8
18
Z
32
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\boxed{{}^{R}_{K}\text{P}\sum_{25}^{32}\underset{\text{Deuterium mass-energy per shell}}{\left[\overset{\text{Baryon rest masses}}{[72(n)^2]}+\overset{\text{lepton rest mass}}{[12e19]}+\overset{\text{KEM}}{[m_e v^2]}\right]_1^8}}$$

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

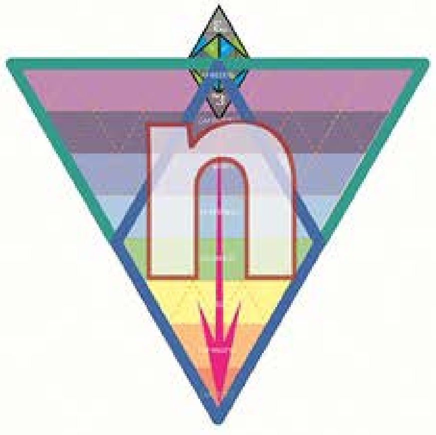

γ

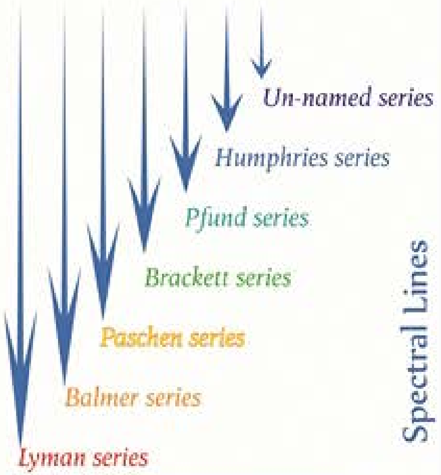
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$\boxed{Mv^2 = \text{KEM} = hcR_H}$$

Photon emission/absorption

Photo-electron ionisation energies

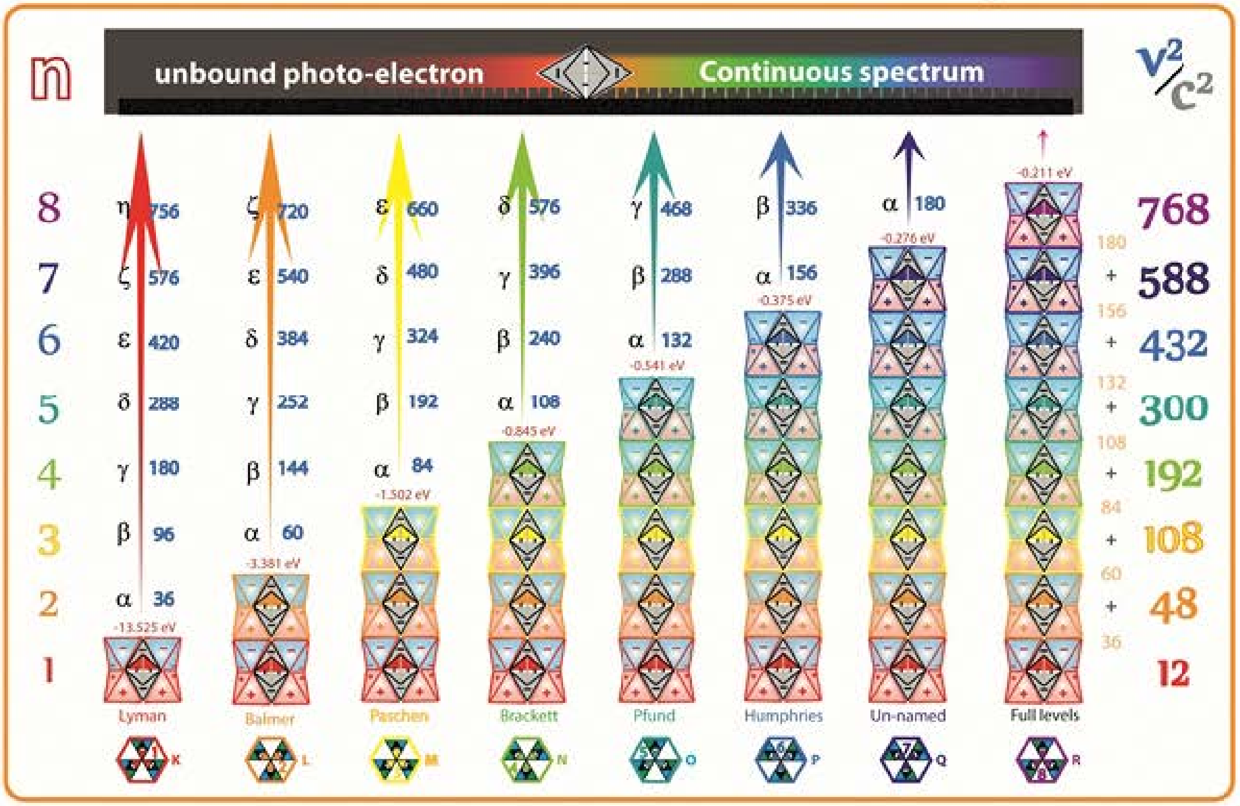
n
unbound photo-electron
Continuous spectrum
v²/c²
8
7
6
5
4
3
2
1
η 756
ζ 576
ε 420
δ 288
γ 180
β 96
α 36
-13.525 eV
Lyman
K
ζ 720
ε 540
δ 384
γ 252
β 144
α 60
-3.381 eV
Balmer
L
ε 660
δ 480
γ 324
β 192
α 84
-1.502 eV
Paschen
M
δ 576
γ 396
β 240
α 108
-0.845 eV
Brackett
N
γ 468
β 288
α 132
-0.541 eV
Pfund
O
β 336
α 156
-0.375 eV
Humphries
P
α 180
-0.276 eV
Un-named
Q
-0.211 eV
Full levels
R
768
180
+
588
156
+
432
132
+
300
108
+
192
84
+
108
60
+
48
36
12

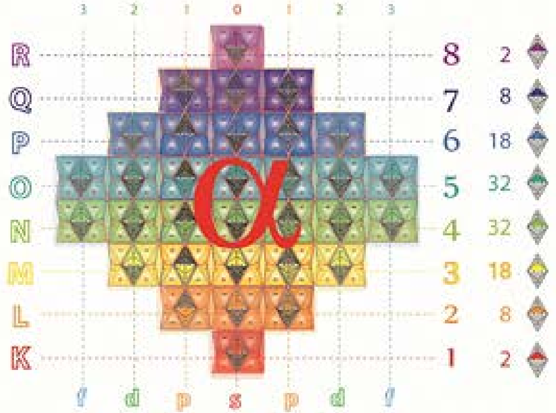
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
α
8 2
7 8
6 18
5 32
4 32
3 18
2 8
1 2
f d p s p d f

Opposed electron spins results in Hyperfine splitting & Lamb shifts

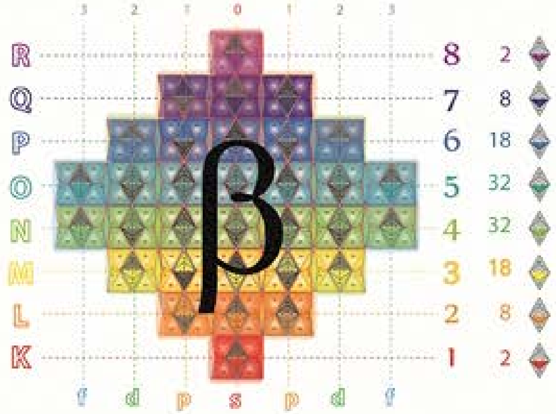
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
β
8 2
7 8
6 18
5 32
4 32
3 18
2 8
1 2
f d p s p d f

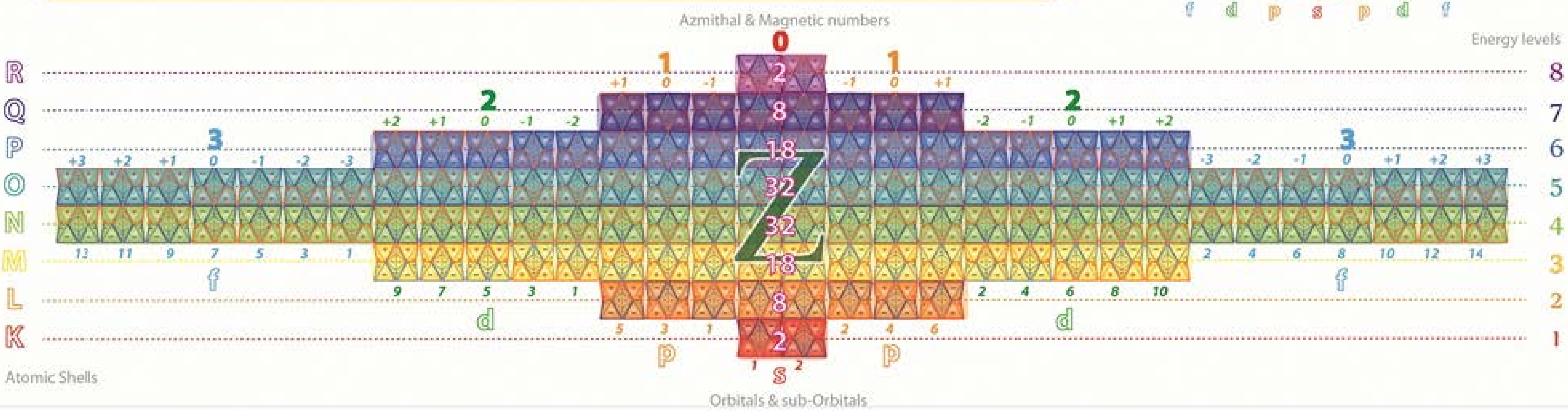
Azmithal & Magnetic numbers
Energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1 +1 0 -1
1 -1 0 +1
2 +2 +1 0 -1 -2
2 -2 -1 0 +1 +2
3 +3 +2 +1 0 -1 -2 -3
3 -3 -2 -1 0 +1 +2 +3
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f
f
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals

Eigenstate - Ionisation energies

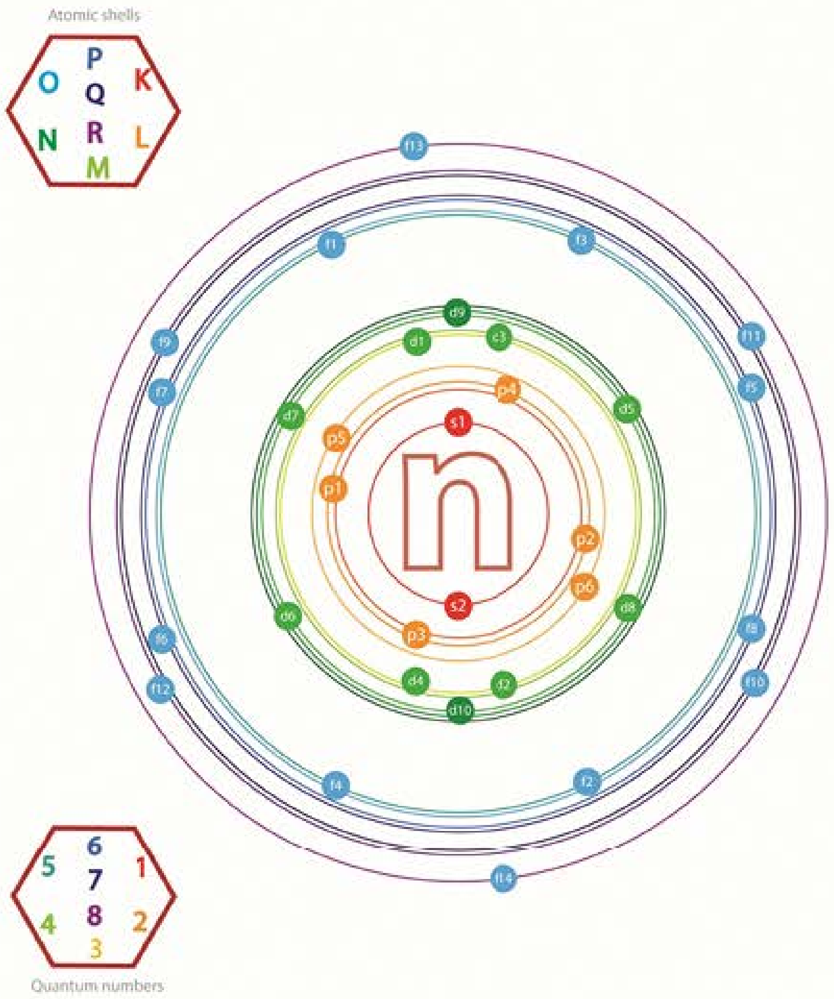
Atomic shells
P
O
Q
K
N
R
L
M
n
s1
s2
p1
p2
p3
p4
p5
p6
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
f14
6
5
7
1
8
4
2
3
Quantum numbers

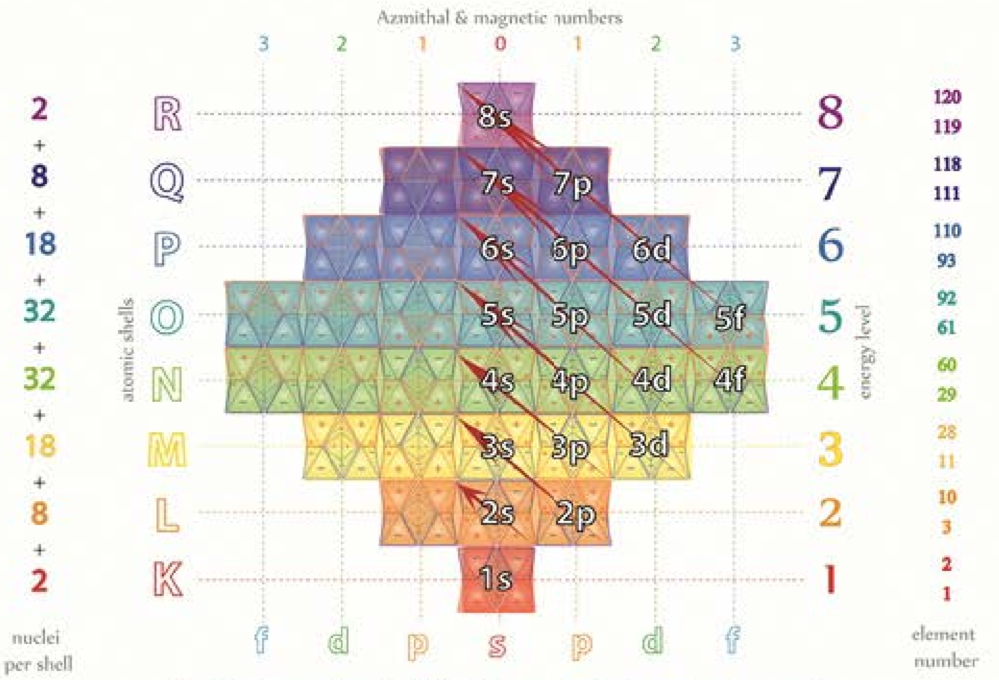
Azmithal & magnetic numbers
3
2
1
0
1
2
3
2
+
8
+
18
+
32
+
32
+
18
+
8
+
2
nuclei per shell
atomic shells
R
Q
P
O
N
M
L
K
8s
7s
7p
6s
6p
6d
5s
5p
5d
5f
4s
4p
4d
4f
3s
3p
3d
2s
2p
1s
f
d
p
s
p
d
f
energy level
8
7
6
5
4
3
2
1
120
119
118
111
110
93
92
61
60
29
28
11
10
3
2
1
element number

The ionisation energies of individual atoms varies due to many factors, namely:
electron spin-orbital coupling with Baryons of specific energies,
the relativistic energies of photo-electrons bound in nuclei
and Zitterbewegung effects on bound electrons

First ionisation energies for all periodic elements

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

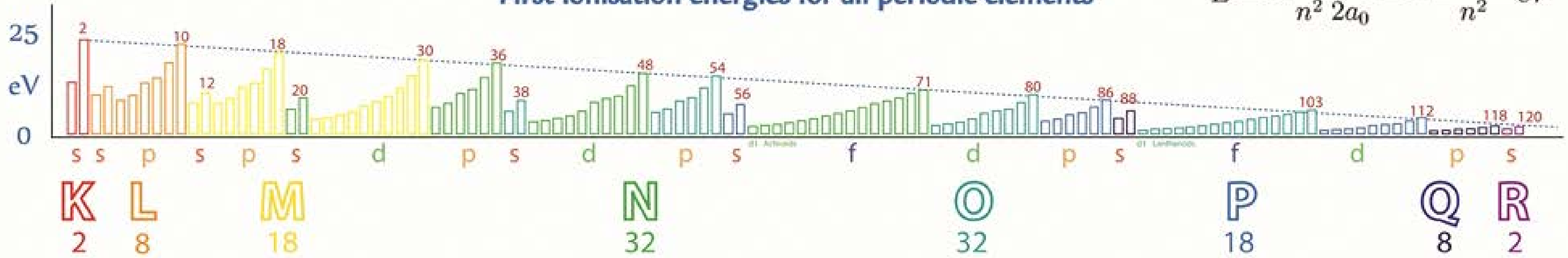

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number. It is used as rule to determine which nuclei are stable or unstable.

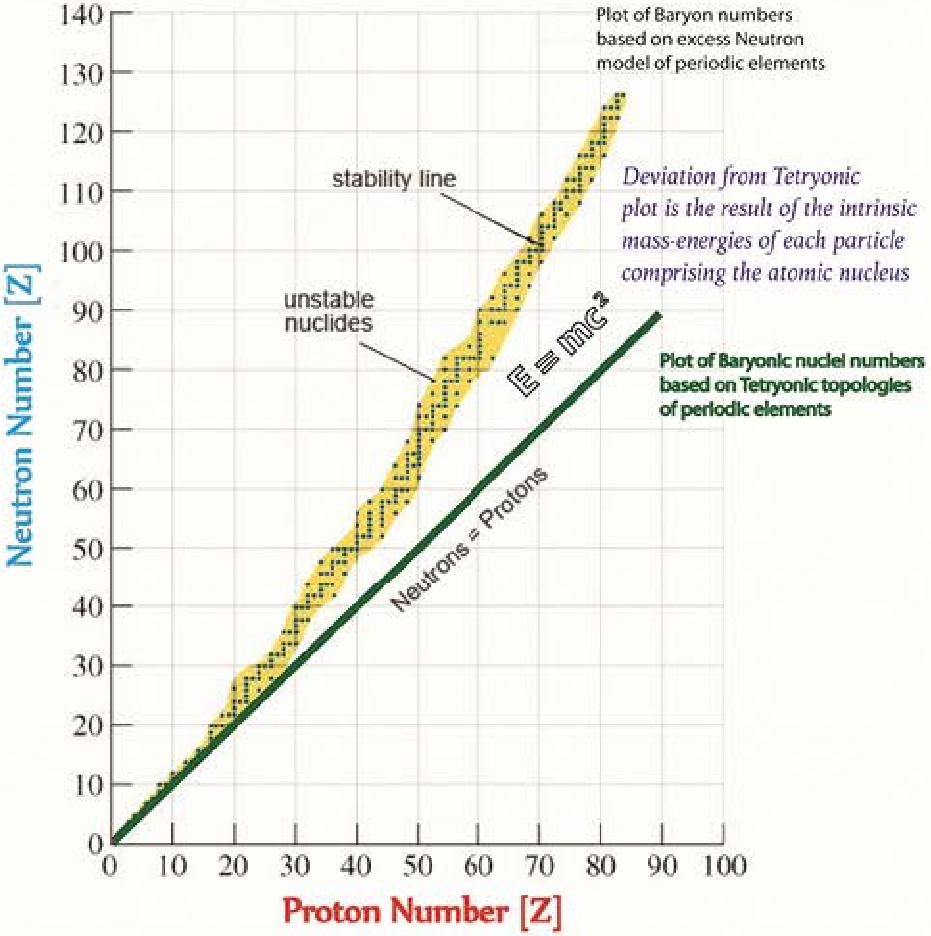

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

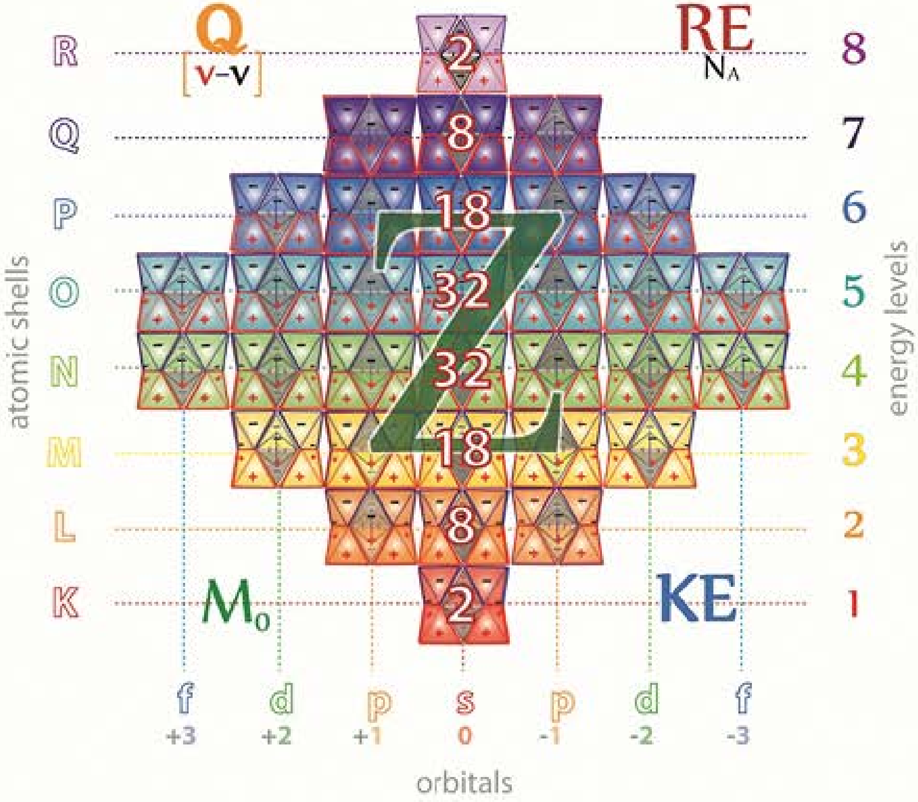

R
Q
P
O
N
M
L
K
atomic shells
Q
[v-v]
RE
N_A
2
8
18
32
32
18
8
2
Z
M0
KE
8
7
6
5
4
3
2
1
energy levels
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
orbitals

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

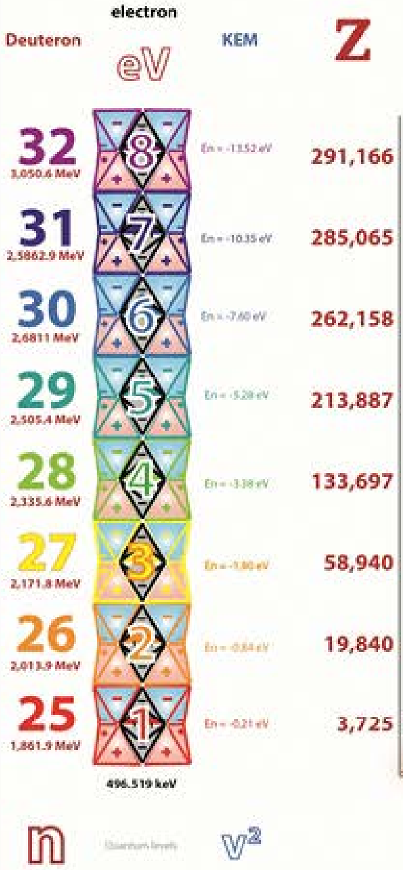
Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
8
En = -13.52 eV
291,166
31
2,5862.9 MeV
7
En = -10.35 eV
285,065
30
2,6811 MeV
6
En = -7.60 eV
262,158
29
2,505.4 MeV
5
En = -5.28 eV
213,887
28
2,335.6 MeV
4
En = -3.38 eV
133,697
27
2,171.8 MeV
3
En = -1.90 eV
58,940
26
2,013.9 MeV
2
En = -0.84 eV
19,840
25
1,861.9 MeV
1
En = -0.21 eV
3,725
496.519 keV
n
Quantum levels
v²

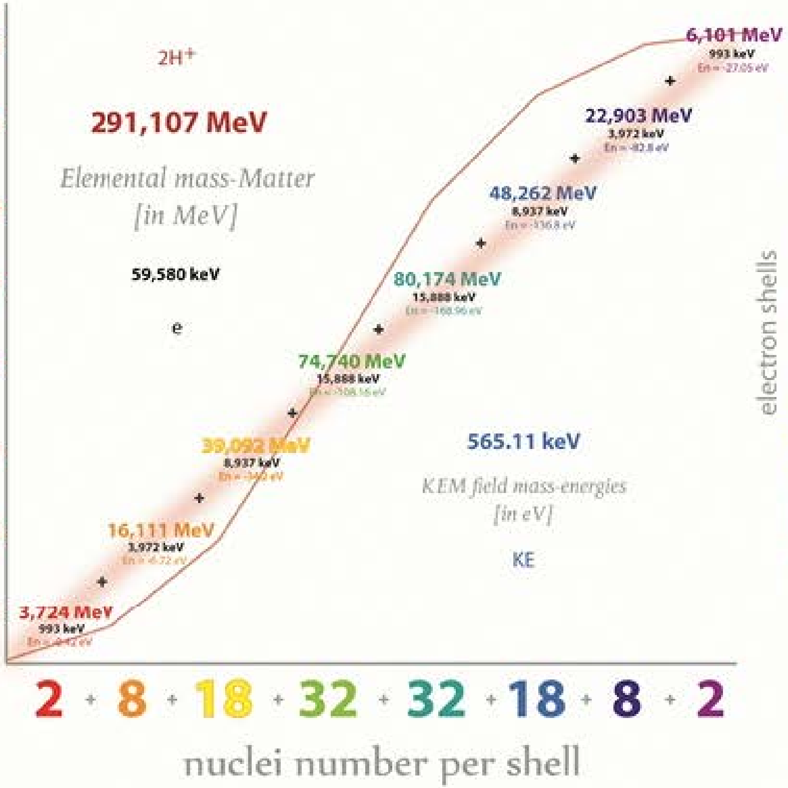

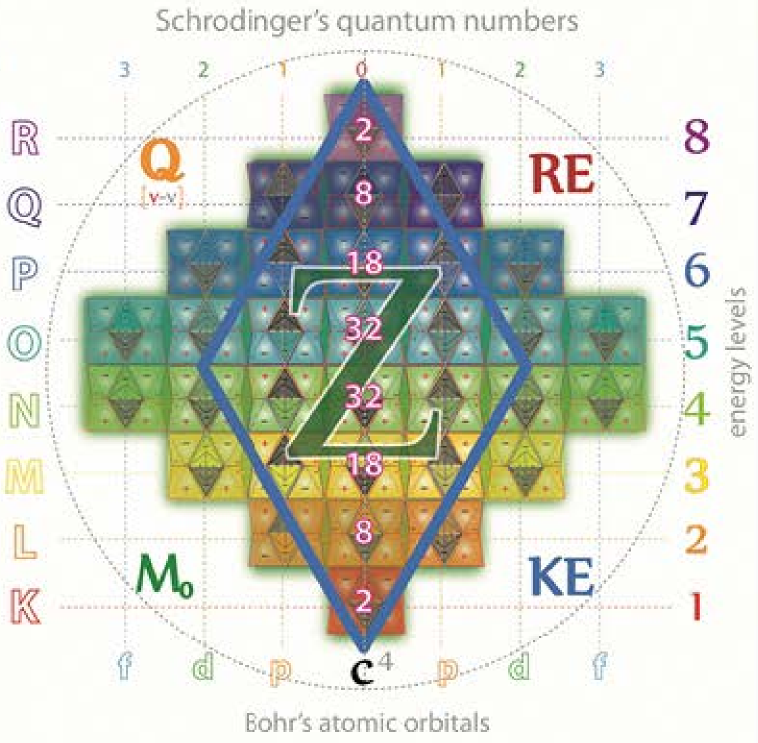
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
2
8
18
32
32
18
8
2
Z
M₀
KE
c⁴
8
7
6
5
4
3
2
1
energy levels
f
d
p
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	496.519 keV	+	13.525 ev

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Baryons 930.947 MeV + **KEM fields** 13.525 ev + **electrons** 496.519 keV

Mapping Planck mass-energy contributions to elementary Matter and isotopes

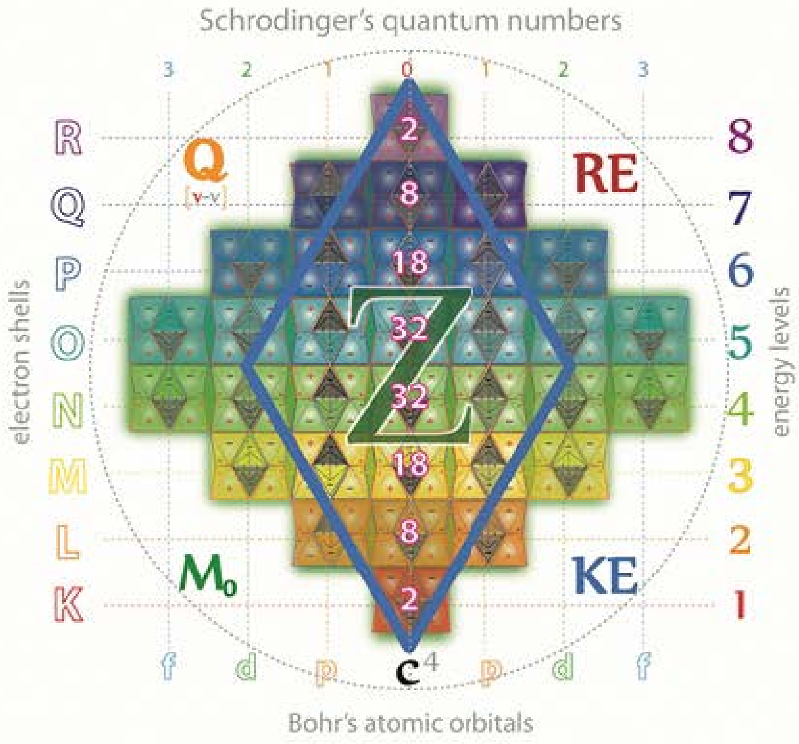
Schrodinger's quantum numbers
3
2
1
0
1
2
3
electron shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
2
8
18
32
Z
32
18
8
2
M₀
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

$$E = hv^2$$

$$n^2 + v^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nhv$$

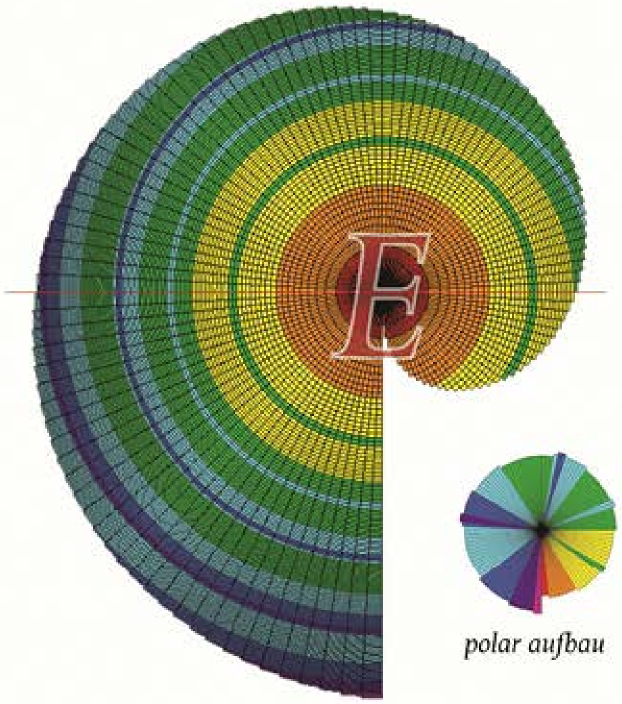
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\Big[\underset{\text{Deuteron rest mass}}{72[v^2]} + \underset{\text{Spectral lines}}{v} + \underset{\text{electron rest mass}}{1.20\text{ e}20\ v}\Big]$$

reveal a quadratic formulation for all Z numbers

All elements are comprised of n level Duetrium nuclei

Baryons	**KEM fields**	**electrons**
1,861,949 MeV	13.525 eV	496,519 keV

$$Z\,[[72n^2] + [12v^2] + [1.2e20]]$$

2H+ $[72n^2]$ **Z** γ $[M_e v^2]$

Elemental mass-Matter [in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

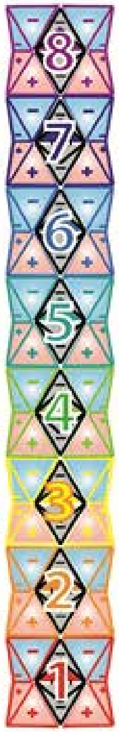

Deuteron	electron eV	KEM
32 3,050.6 MeV	8	En = -13.52 eV
31 2,5862.9 MeV	7	En = -10.35 eV
30 2,6811 MeV	6	En = -7.60 eV
29 2,505.4 MeV	5	En = -5.28 eV
28 2,335.6 MeV	4	En = -3.38 eV
27 2,171.8 MeV	3	En = -1.90 eV
26 2,013.9 MeV	2	En = -0.84 eV
25 1,861.9 MeV	1	En = -0.21 eV

496,519 keV

m v²

Determines the spectral line [KEM field energies] of electrons bound to them

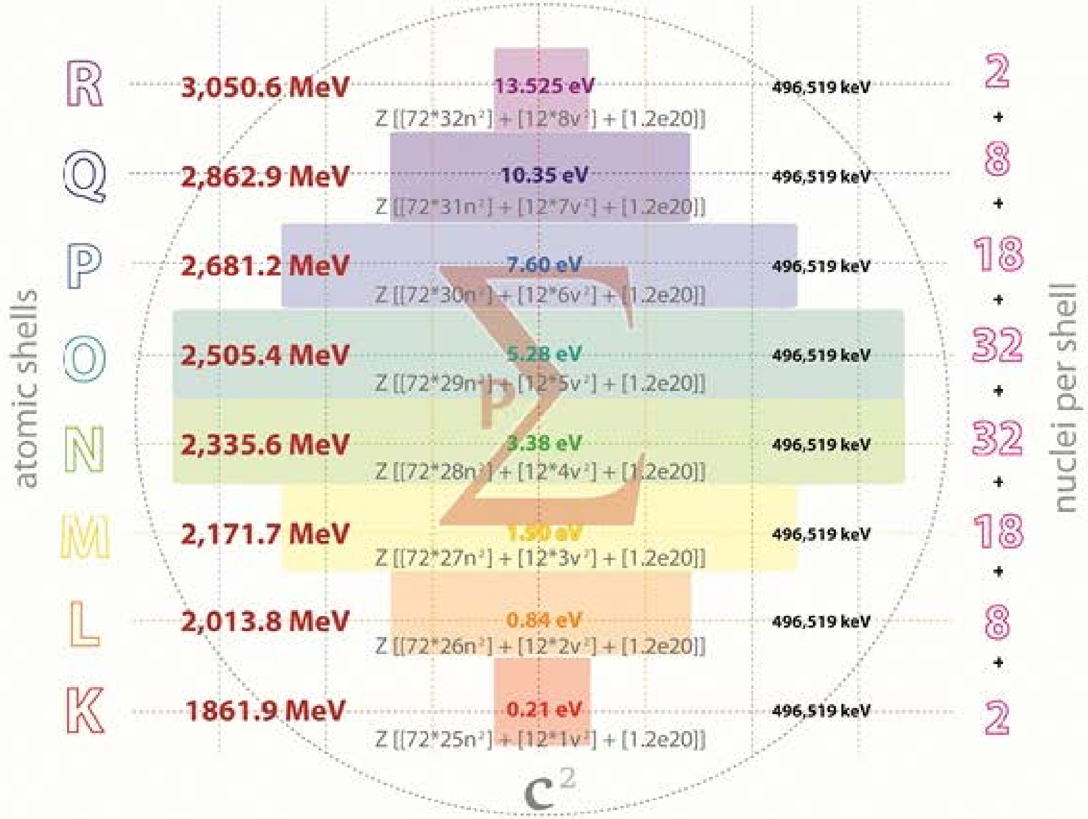

atomic shells					nuclei per shell
R	3,050.6 MeV	13.525 eV	Z [[72*32n²] + [12*8v²] + [1.2e20]]	496,519 keV	2
					+
Q	2,862.9 MeV	10.35 eV	Z [[72*31n²] + [12*7v²] + [1.2e20]]	496,519 keV	8
					+
P	2,681.2 MeV	7.60 eV	Z [[72*30n²] + [12*6v²] + [1.2e20]]	496,519 keV	18
					+
O	2,505.4 MeV	5.28 eV	Z [[72*29n²] + [12*5v²] + [1.2e20]]	496,519 keV	32
					+
N	2,335.6 MeV	3.38 eV	Z [[72*28n²] + [12*4v²] + [1.2e20]]	496,519 keV	32
					+
M	2,171.7 MeV	1.90 eV	Z [[72*27n²] + [12*3v²] + [1.2e20]]	496,519 keV	18
					+
L	2,013.8 MeV	0.84 eV	Z [[72*26n²] + [12*2v²] + [1.2e20]]	496,519 keV	8
					+
K	1861.9 MeV	0.21 eV	Z [[72*25n²] + [12*1v²] + [1.2e20]]	496,519 keV	2

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

the rest mass-Matter of bound photo-electrons is velocity invariant

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

48
H
1.0000
22,512

Avagadro's number

6.022141579 e26
atoms in 1KG of Matter

1 KG mass
[of Matter]

1.660538841 e-27 kg
atomic rest mass-Matter

Hydrogen

01

using SI units Avagadro's number can be expressed exactly as the inverse rest mass of Hydrogen

Weighted atomic mass

Carbon

12

5.019789213 e25
atoms in 1KG of Matter

6.02214078 e 23

1.99211552 e-26 kg
atomic rest mass-Matter

504
[254-254]
C
11.996
C12
270,072

Tetryonic charge geometries make weighted atomic mass measurements and calculations obsolete

22,506 1/12 of Carbon 12 [Graphene] is not equal to 1 Hydrogen atom 22,512
(Deuterium is the building block of all atomic elements)

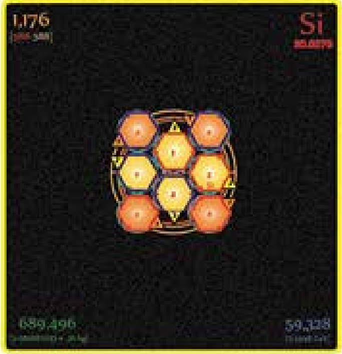

1.966225348 e25
atoms in 1KG of Matter

International Avagadro project

5.085887033 e-26 kg
atomic rest mass-Matter

Silicon

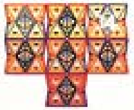

28

The gram was originally defined in 1795 as the mass of one cubic centimeter of water at 4°C,
making the kilogram equal to the mass of one liter of water.

The prototype kilogram, manufactured in 1799 and from which the current kilogram is based
has a mass equal to the mass of 1.000025 liters of water

In recent years two major experiments, namely the Watt balance & Avagadro projects, have been attempting to
measure and define 1KG of mass-Matter in terms of electrical force and the number of atoms respectively
in order to better define 1KG of mass-Matter precisely for all future physical references

Platinum

191

2.817950081 e24
atoms in 1KG of Matter
3.181804449 e23

La Grande K

90% Pt
10% Ir

The La grande K is an alloy of 90% Platinum & 10% Iridium
that has been slowly losing mass since its manufacture

3.1893811012 e-25 kg
atomic rest mass-Matter
3.142870708 e-25 kg

6,552
Pt
4,329,864
818,028

All atomic rest masses are for atoms at absolute zero and any deviation is a measure of the topological Matter's Kinetic energy content [chemical energy, KEM fields and/or spectral lines]

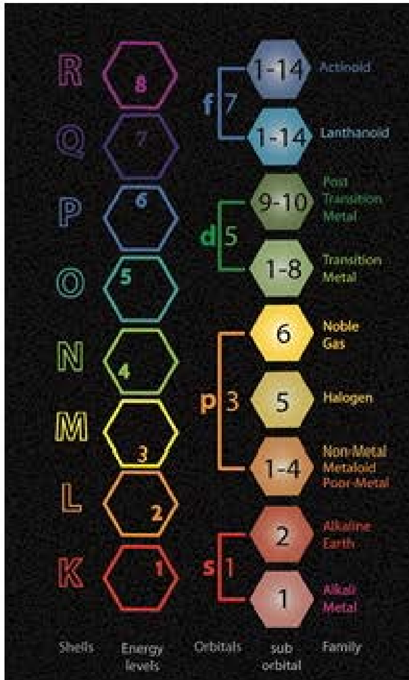
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
d 5
p 3
s 1
1-14 Actinoid
1-14 Lanthanoid
9-10 Post Transition Metal
1-8 Transition Metal
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
2 Alkaline Earth
1 Alkali Metal
Shells
Energy levels
Orbitals
sub orbital
Family

Incorporating

- *Quark configurations*
- *Nuclei geometries*
- *Chemical bond points*
- *Atomic Orbitals*
- *Hund Rule*
- *Pauli exclusion*
- *Aufbau filling principle*
- *Element names, numbers and*
- *Elementary families*

Element placements that reflects the true quantum geometries of periodic elements

● UP Quark

● DOWN Quark

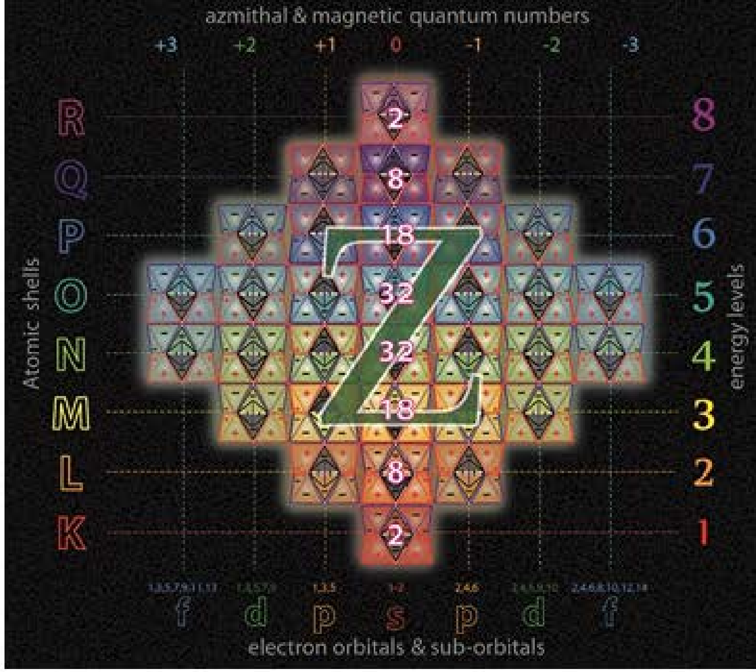
azmithal & magnetic quantum numbers
+3
+2
+1
0
-1
-2
-3
Atomic shells
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals & sub-orbitals

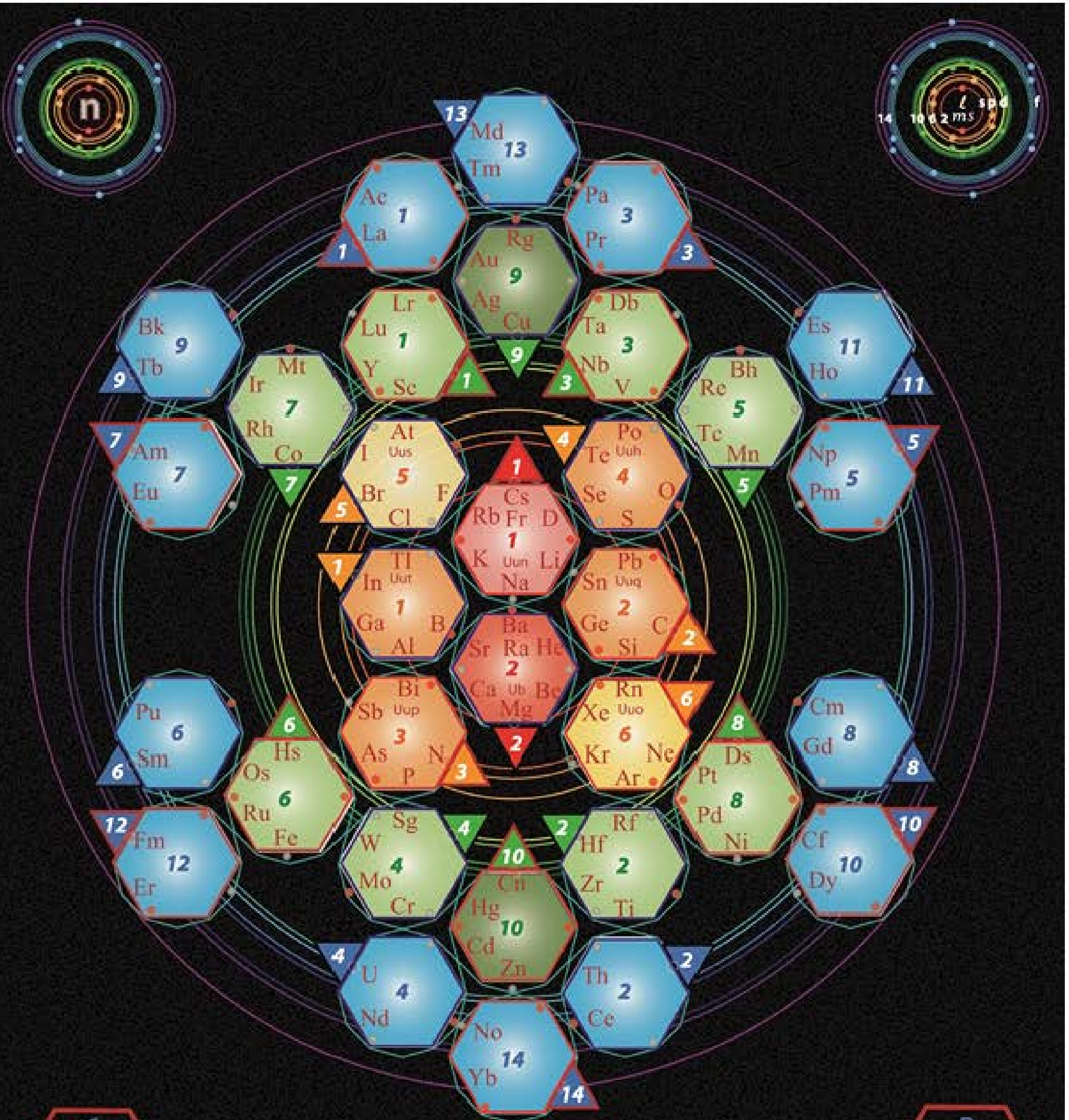
n
ℓ
s p d
f
m s
14
10 6 2
13 Md 13 Tm
Ac 1 La
Pa 3 Pr
Rg Au 9 Ag Cu
Lr Lu 1 Y Sc
Db Ta 3 Nb V
Bk 9 Tb
Mt Ir 7 Rh Co
Bh Re 5 Tc Mn
Es 11 Ho
Am 7 Eu
At I Uus 5 Br F Cl
Cs Rb Fr D 1 K Uun Li Na
Po Te Uuh 4 Se O S
Np 5 Pm
Tl In Uut 1 Ga B Al
Ba Sr Ra He 2 Ca Ub Be Mg
Pb Sn Uuq 2 Ge C Si
Pu 6 Sm
Bi Sb Uup 3 As N P
Rn Xe Uuo 6 Kr Ne Ar
Cm 8 Gd
Hs Os 6 Ru Fe
Ds Pt 8 Pd Ni
Fm 12 Er
Sg W 4 Mo Cr
Rf Hf 2 Zr Ti
Cf 10 Dy
Cn Hg 10 Cd Zn
U 4 Nd
Th 2 Ce
No 14 Yb
1
2
3
4
5
6
7
8
9
10
11
12
13
14

6
5
7
1
4
8
2
3
Energy levels

Periodic Table 2.0

(the quantum geometry of periodic elements)

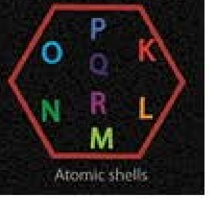
P
O
Q
K
N
R
L
M
Atomic shells

Copyright ABRAHAM [2008] - All rights reserved

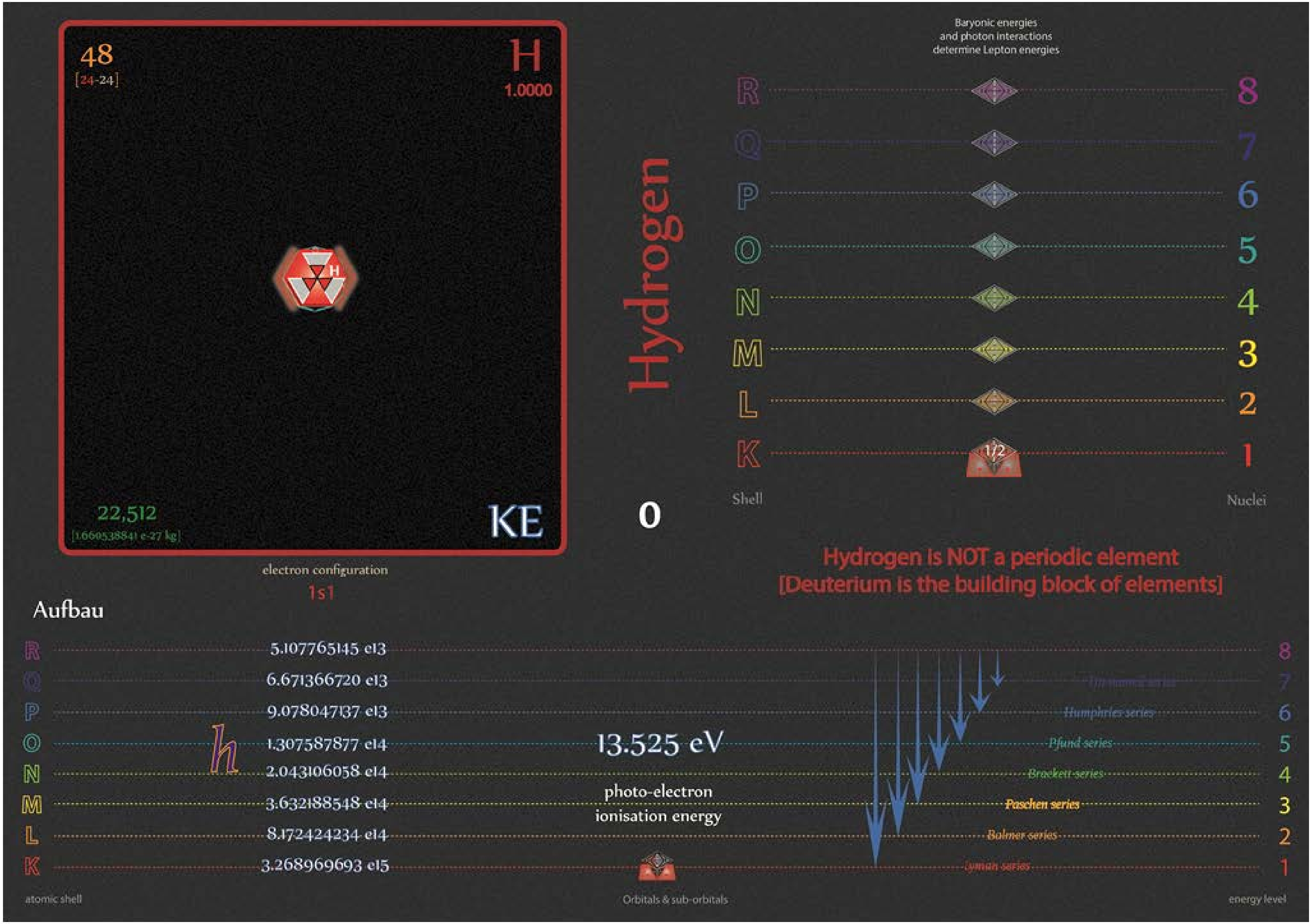

48
[24-24]

H
1.0000

22,512
[1.660538841 e-27 kg]

KE

electron configuration
1s1

Hydrogen

0

Baryonic energies
and photon interactions
determine Lepton energies

Shell		Nuclei
R		8
Q		7
P		6
O		5
N		4
M		3
L		2
K	1/2	1

Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]

Aufbau

atomic shell			energy level
R	5.107765145 e13		8
Q	6.671366720 e13	[illegible]	7
P	9.078047137 e13	Humphries series	6
O	1.307587877 e14	Pfund series	5
N	2.043106058 e14	Brackett series	4
M	3.632188548 e14	Paschen series	3
L	8.172424234 e14	Balmer series	2
K	3.268969693 e15	Lyman series	1

h

13.525 eV

photo-electron
ionisation energy

Orbitals & sub-orbitals

Tetryonics 51.00 - Hydrogen atom

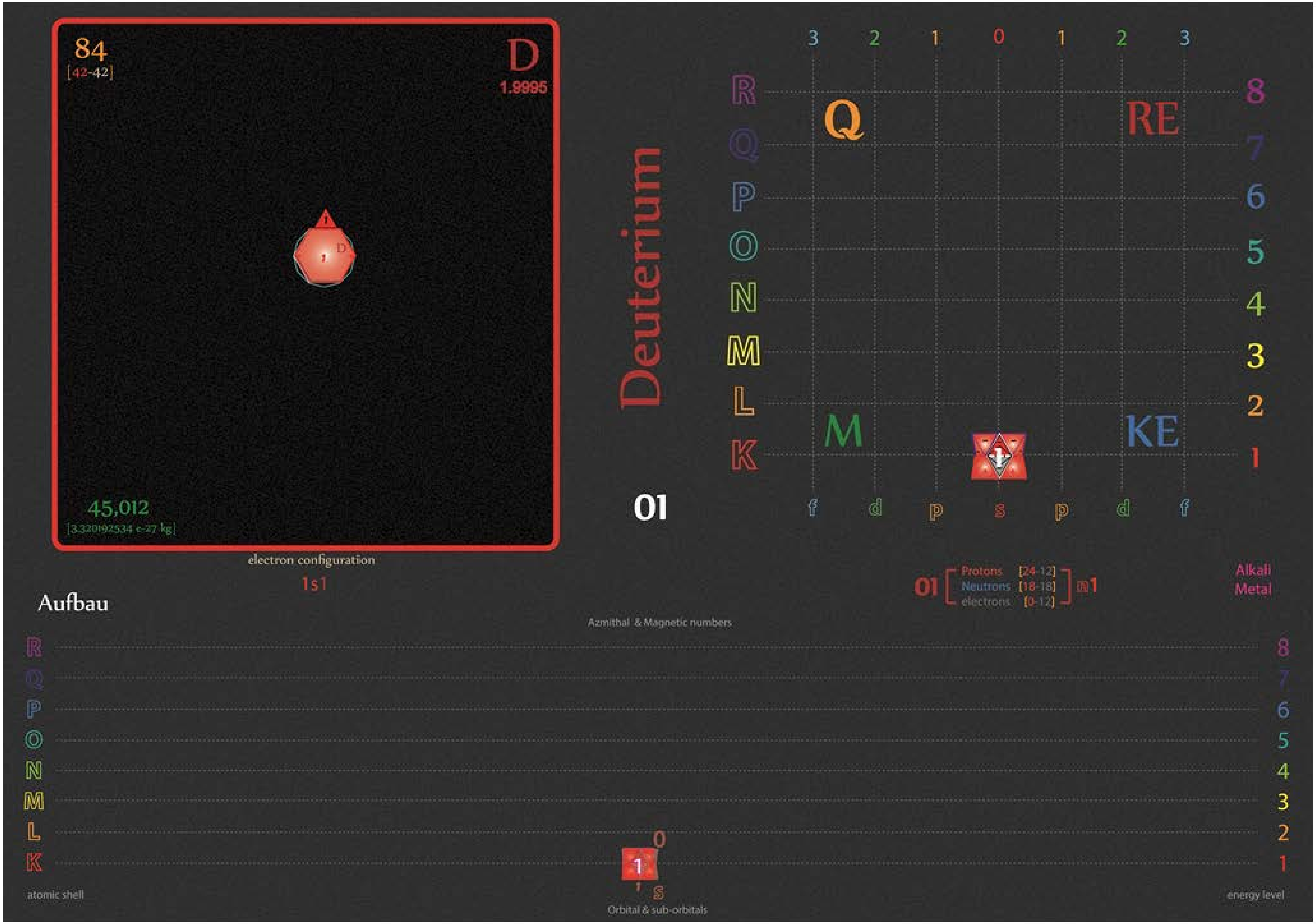
84
[42-42]
D
1.9995
45,012
[3.320192534 e-27 kg]
electron configuration
1s1
Deuterium
01
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
1
f d p s p d f
01
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1
Alkali
Metal
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
s
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.01 - Deuterium atom

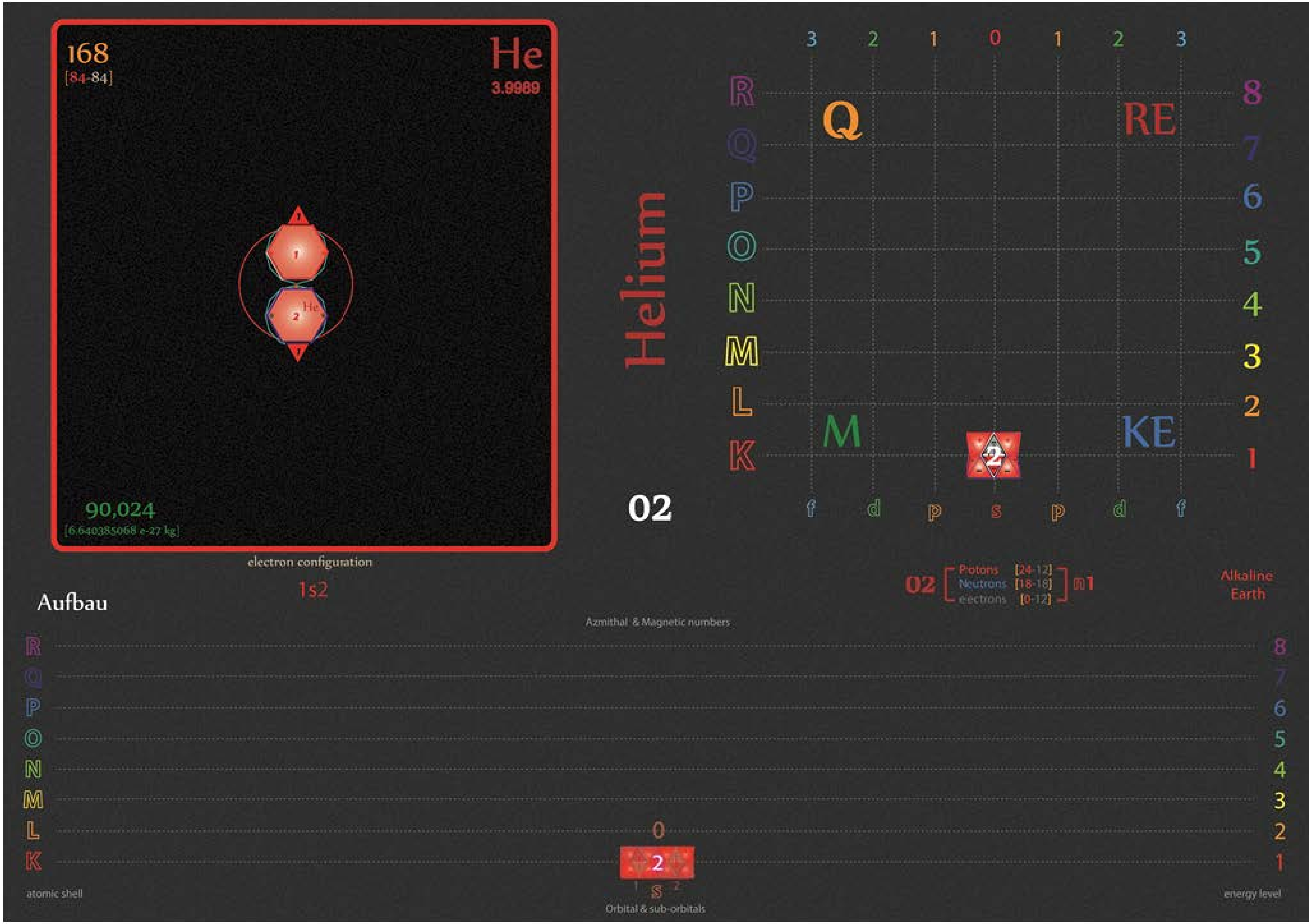
168
[84-84]
He
3.9989
He
90,024
[6.640385068 e-27 kg]
electron configuration
1s2
Aufbau
Helium
02
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
2
f d p s p d f
02
Potons [24-12]
Neutrons [18-18]
electrons [0-12]
m1
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
1 s 2
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.02 - Helium atom

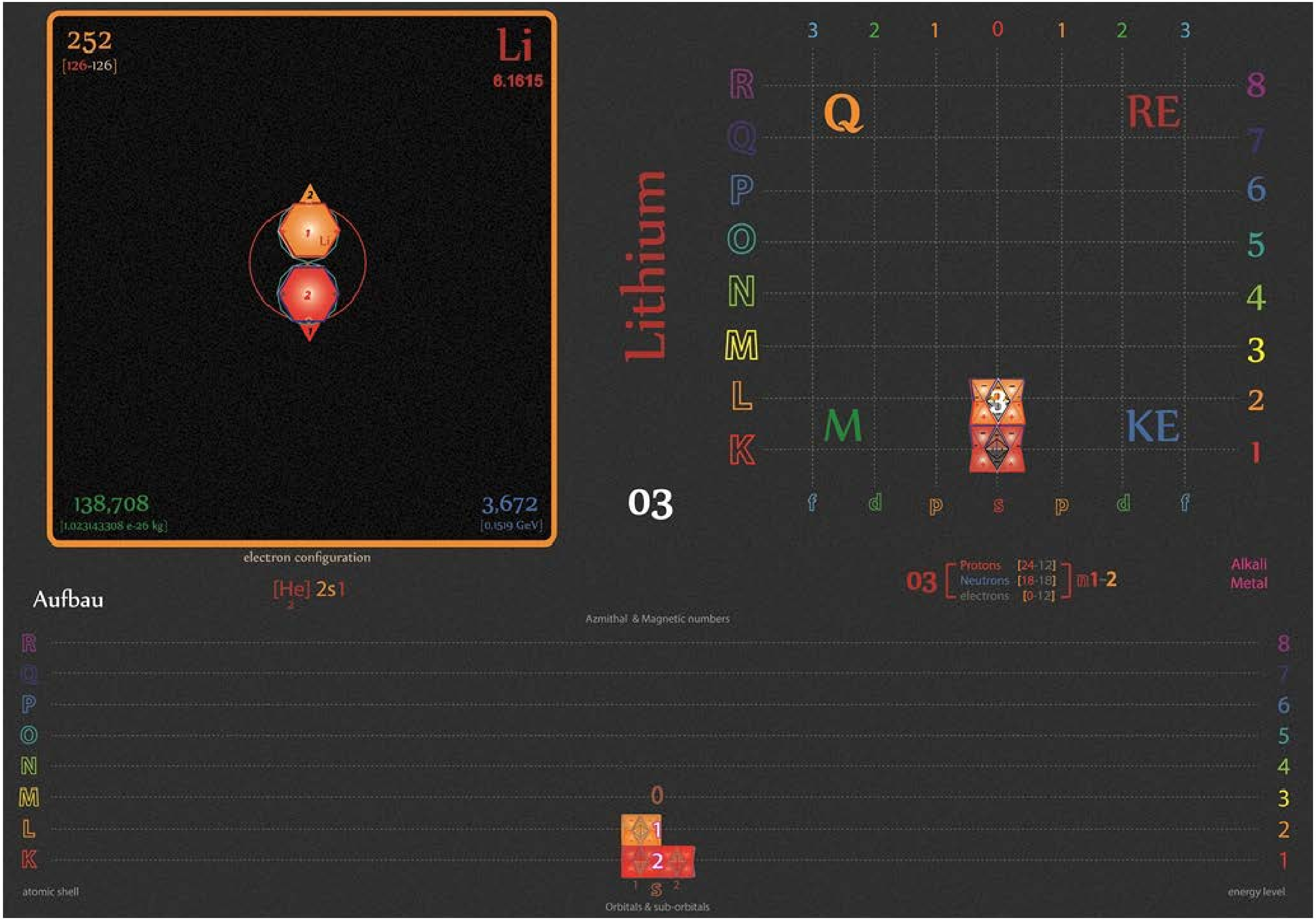
252
[126-126]
Li
6.1615
2
1
Li
2
1
138,708
[1.023143308 e-26 kg]
3,672
[0.1519 GeV]
electron configuration
[He] 2s1
2
Aufbau
Lithium
03
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
3
f d p s p d f
03
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkali
Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
2
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.03 - Lithium atom

Copyright ABRAHAM [2008] - All rights reserved

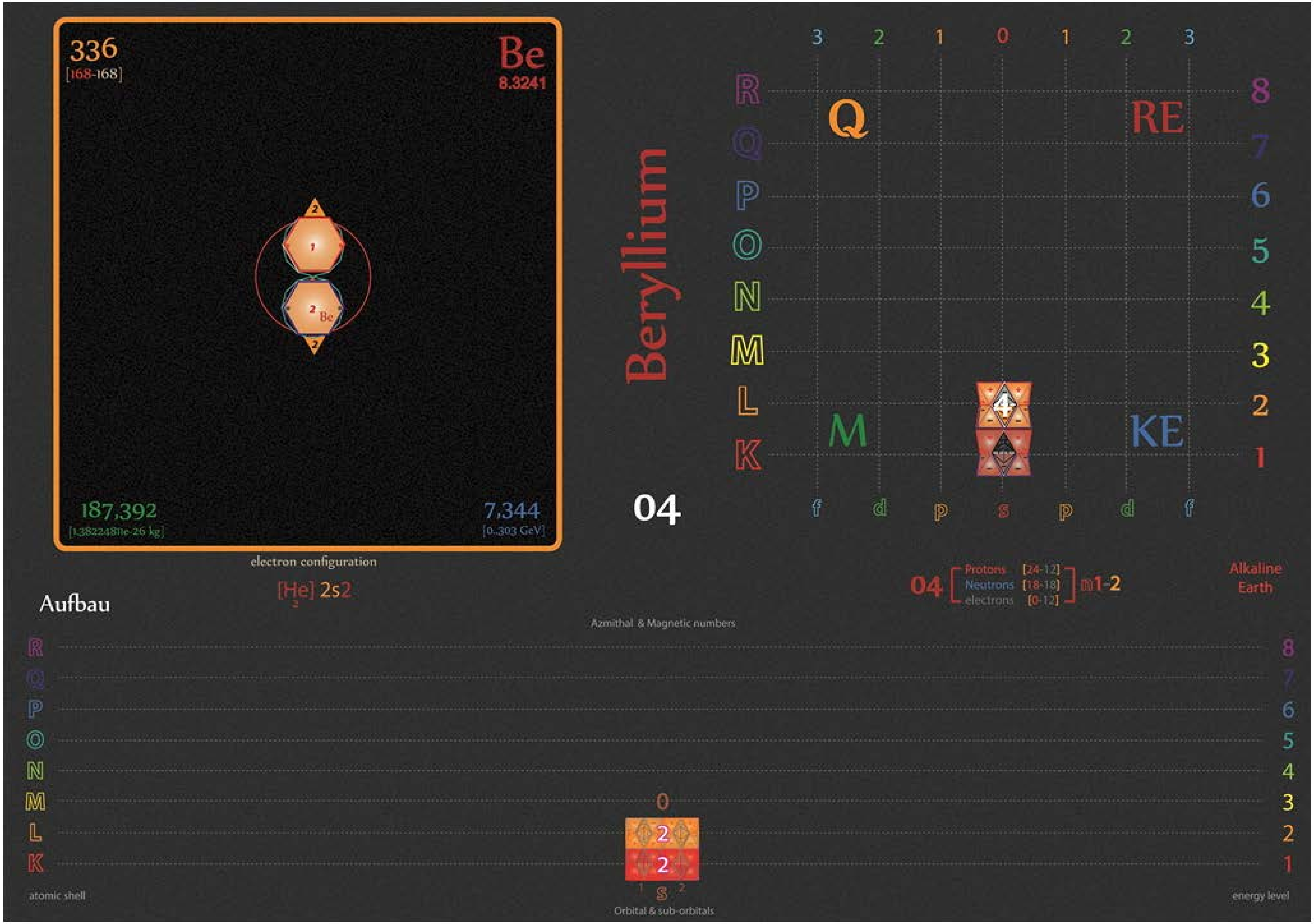
336
[168-168]
Be
8.3241
2
1
2
Be
2
187,392
[1.3822481e-26 kg]
7,344
[0.303 GeV]
electron configuration
[He] 2s2
Aufbau
Beryllium
04
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
4
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
04
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkaline Earth
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
2
2
1
s
2
8
7
6
5
4
3
2
1
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.04 - Beryllium atom

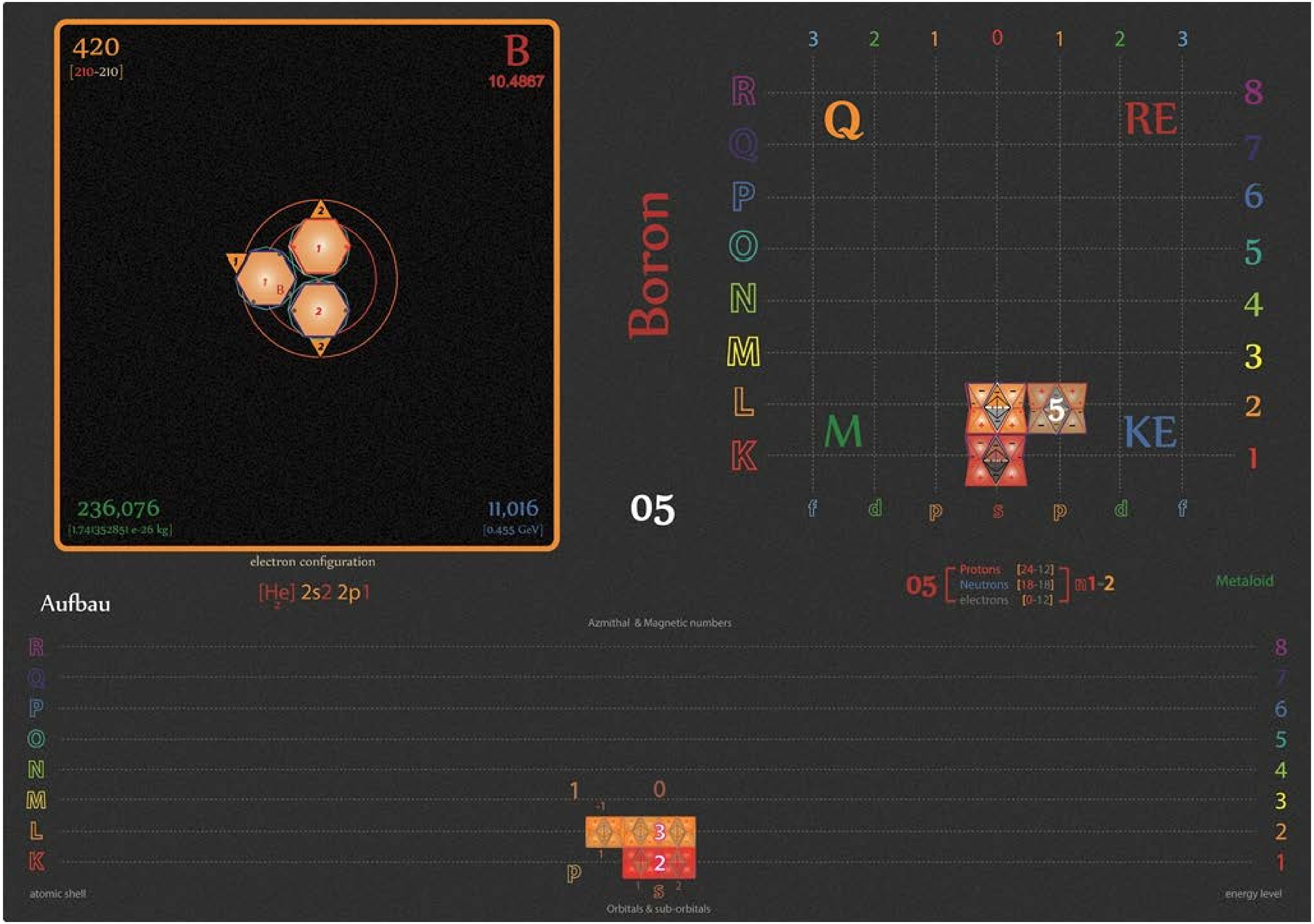
420
[210-210]
B
10.4867
1
2
1
B
2
2
236,076
[1.741352851 e-26 kg]
11,016
[0.455 GeV]
electron configuration
[He] 2s2 2p1
2
Aufbau
Boron
05
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
5
f d p s p d f
05 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-2
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0
-1
1
3
2
p
1 s 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.05 - Boron atom

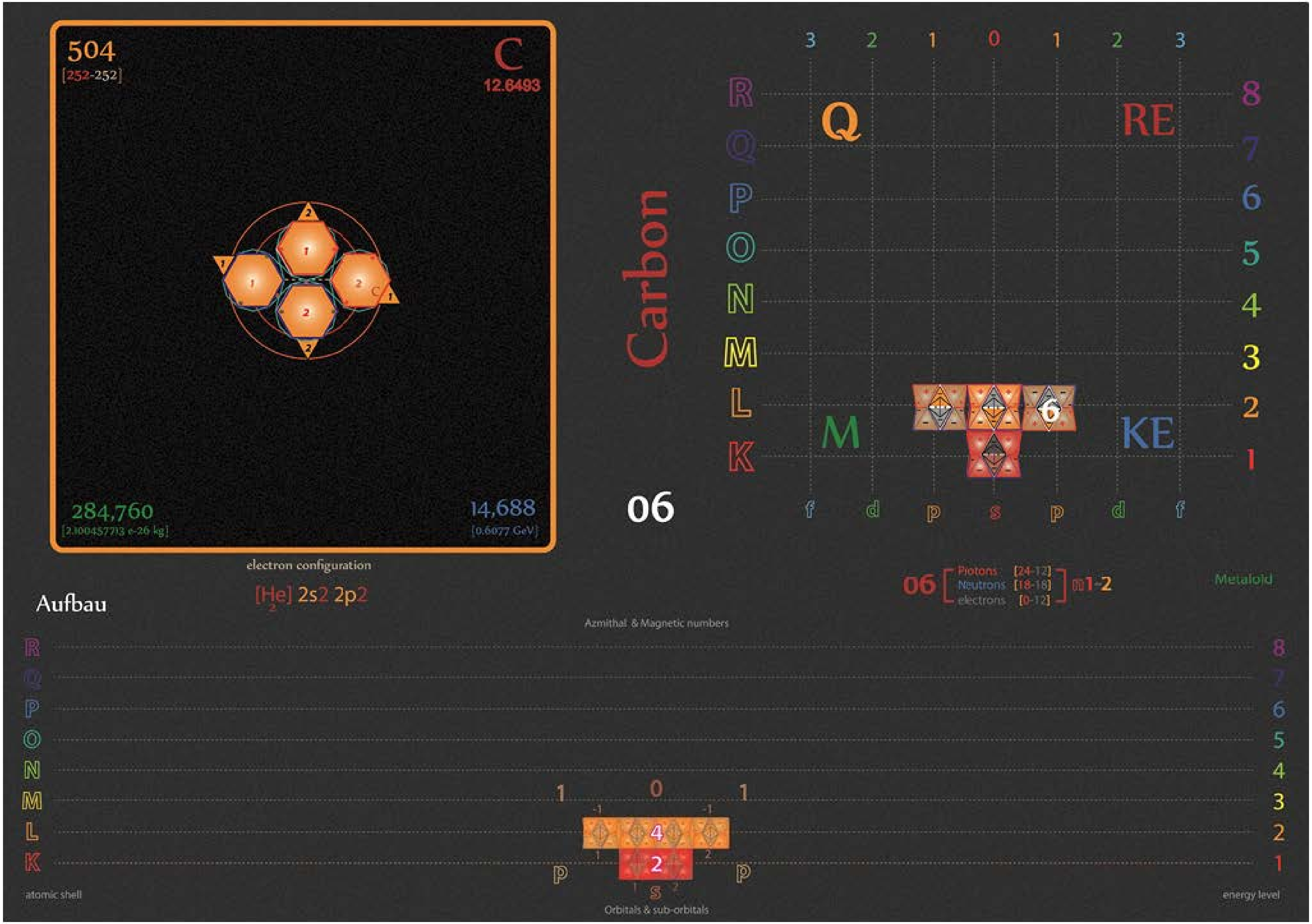
504
[252-252]
C
12.6493
284,760
[2.100457713 e-26 kg]
14,688
[0.6077 GeV]
electron configuration
[He] 2s2 2p2
2
Aufbau
Carbon
06
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
6
f d p s p d f
06
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
-1 -1
4
1 2
2
p p
1 2
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.06 - Carbon atom

Copyright ABRAHAM [2008] - All rights reserved

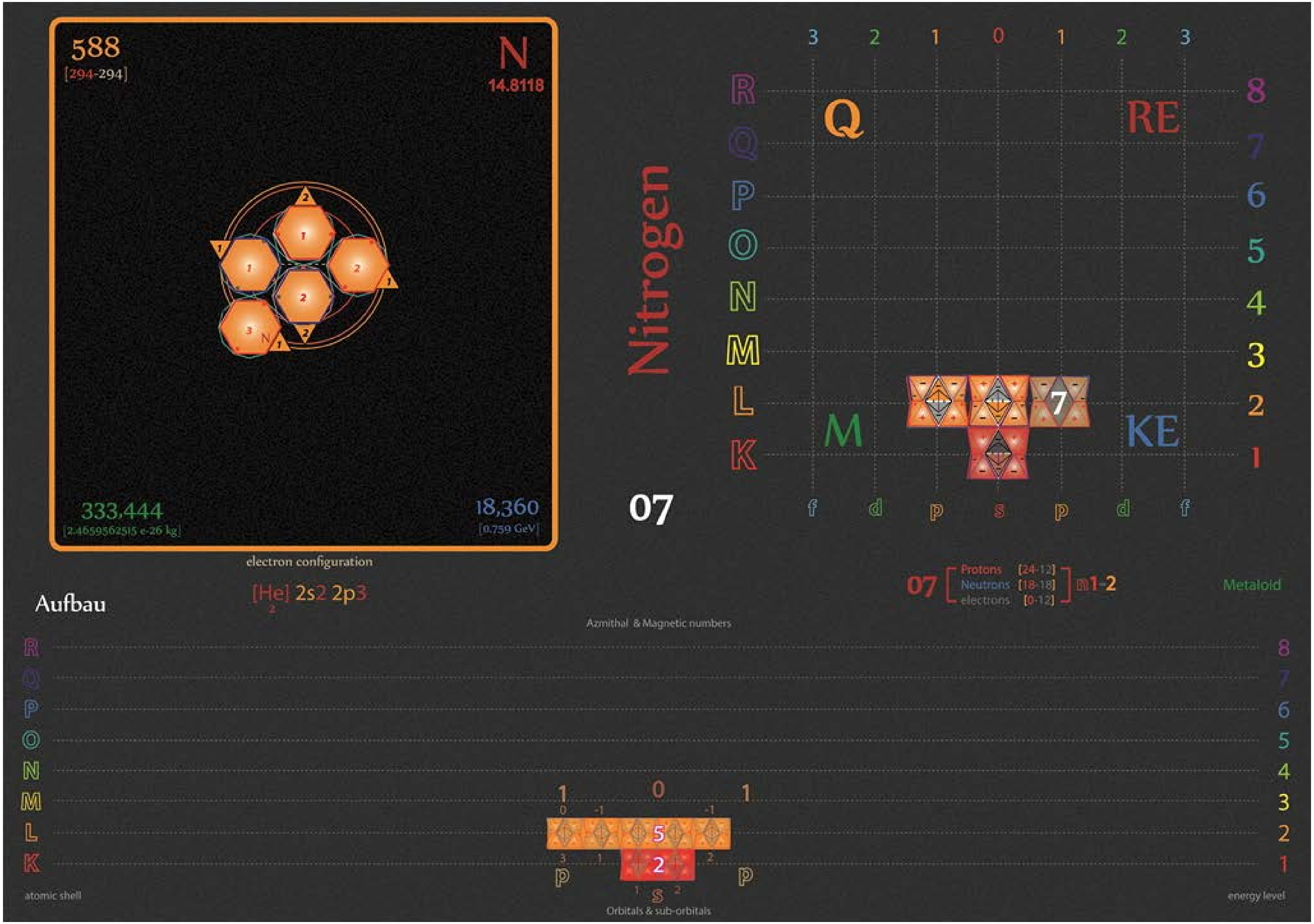
588
[294-294]
N
14.8118
333,444
[2.4659562515 e-26 kg]
18,360
[0.759 GeV]
Nitrogen
07
electron configuration
[He] 2s2 2p3
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
7
f d p s p d f
07
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1
5
2
3 1 2
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.07 - Nitrogen atom

Copyright ABRAHAM [2008] - All rights reserved

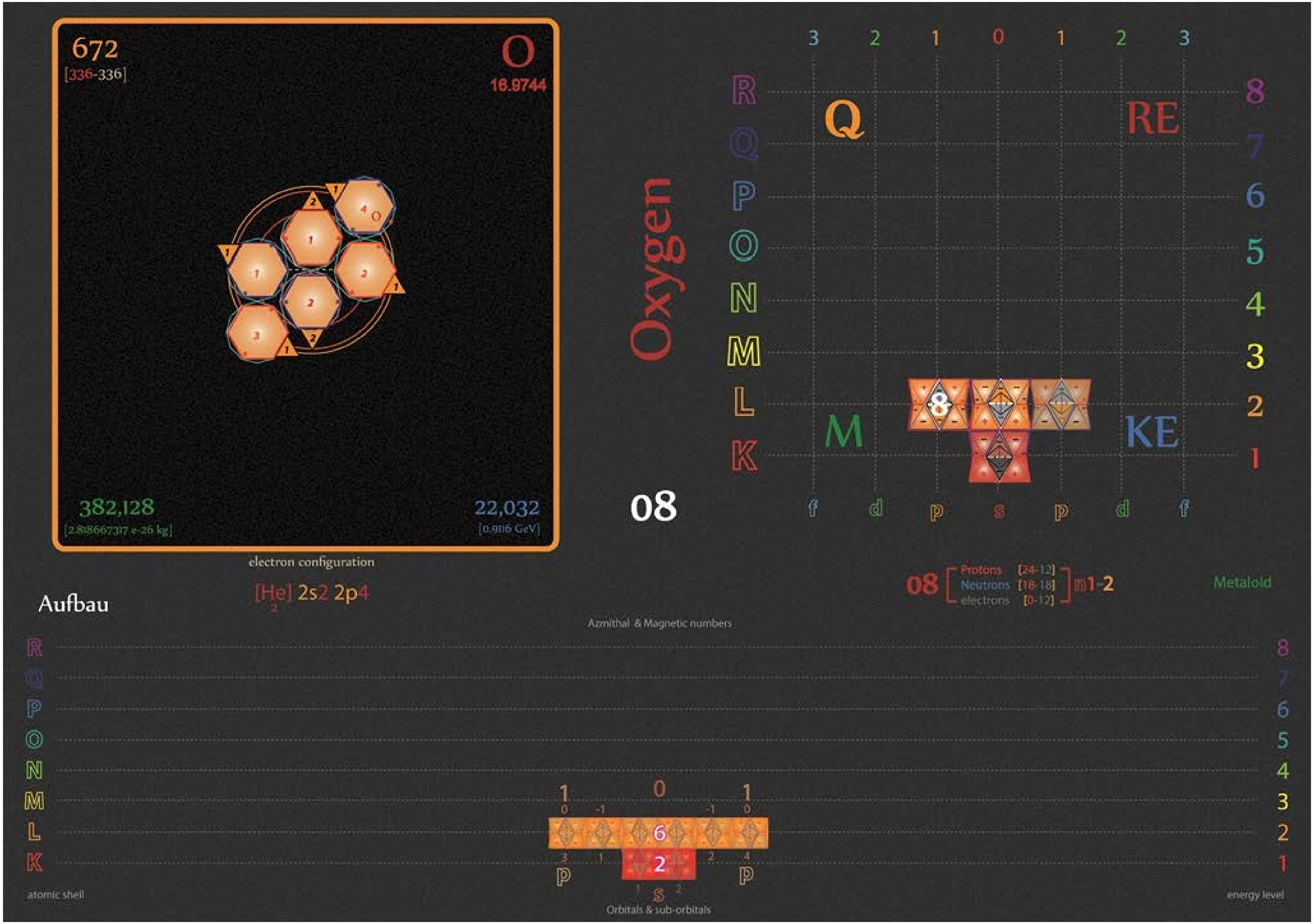
672
[336-336]
O
16.9744
382,128
[2.818667317 e-26 kg]
22,032
[0.9116 GeV]
Oxygen
08
electron configuration
[He] 2s2 2p4
2
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
8
f d p s p d f
08
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
0 -1 -1 0
6
2
3 1 2 4
p p
1 s 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.08 - Oxygen atom

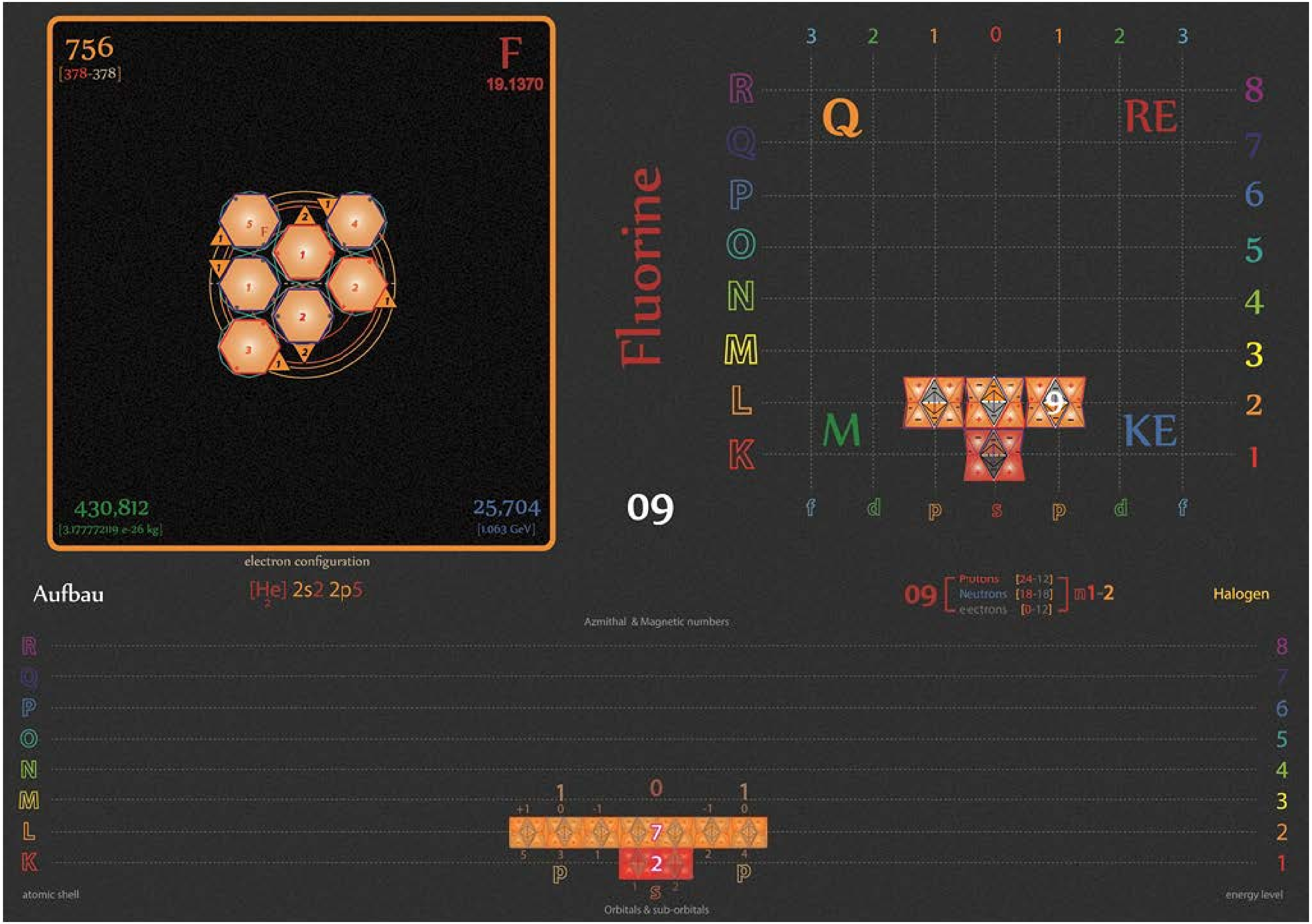
756
[378-378]
F
19.1370
430,812
[3.17772119 e-26 kg]
25,704
[1.063 GeV]
Fluorine
09
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
9
f d p s p d f
electron configuration
Aufbau
[He] 2s2 2p5
Azmithal & Magnetic numbers
09
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Halogen
1 0 1
+1 0 -1 -1 0
7
2
5 3 1 2 4
p s p
1 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.09 - Fluorine atom

Copyright ABRAHAM [2008] - All rights reserved

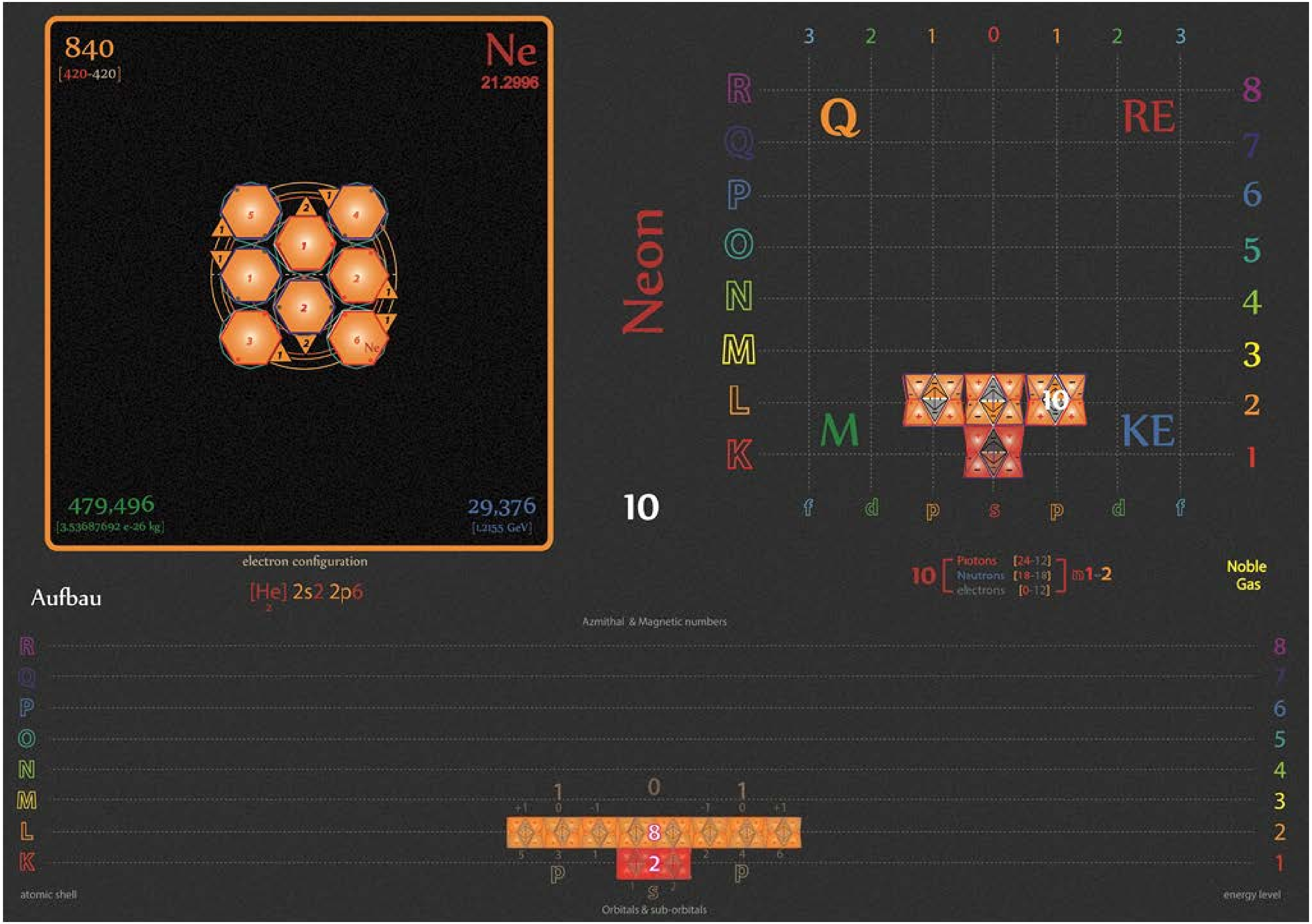
840
[420-420]
Ne
21.2996
479,496
[3.53687692 e-26 kg]
29,376
[1.2155 GeV]
electron configuration
[He] 2s2 2p6
2
Aufbau
Neon
10
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
10
f d p s p d f
10 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Noble Gas
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.10 - Neon atom

Copyright ABRAHAM [2008] - All rights reserved

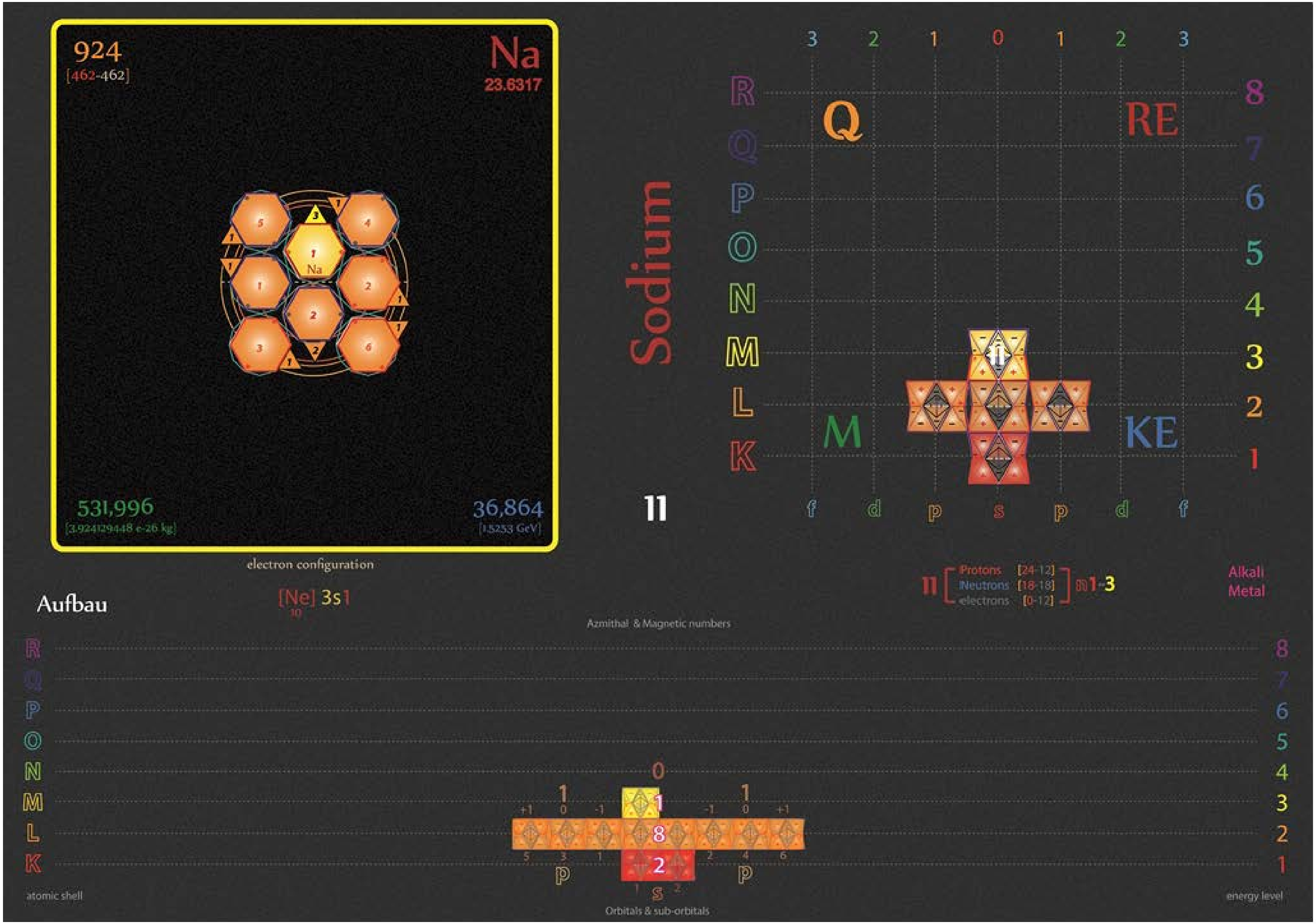
924
[462-462]
Na
23.6317
Na
531,996
[3.924129448 e-26 kg]
36,864
[1.5253 GeV]
electron configuration
Aufbau
[Ne] 3s1
10
Sodium
11
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
11 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Alkali Metal
Azmithal & Magnetic numbers
0
1 1
+1 0 -1 -1 0 +1
1
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.11 - Sodium atom

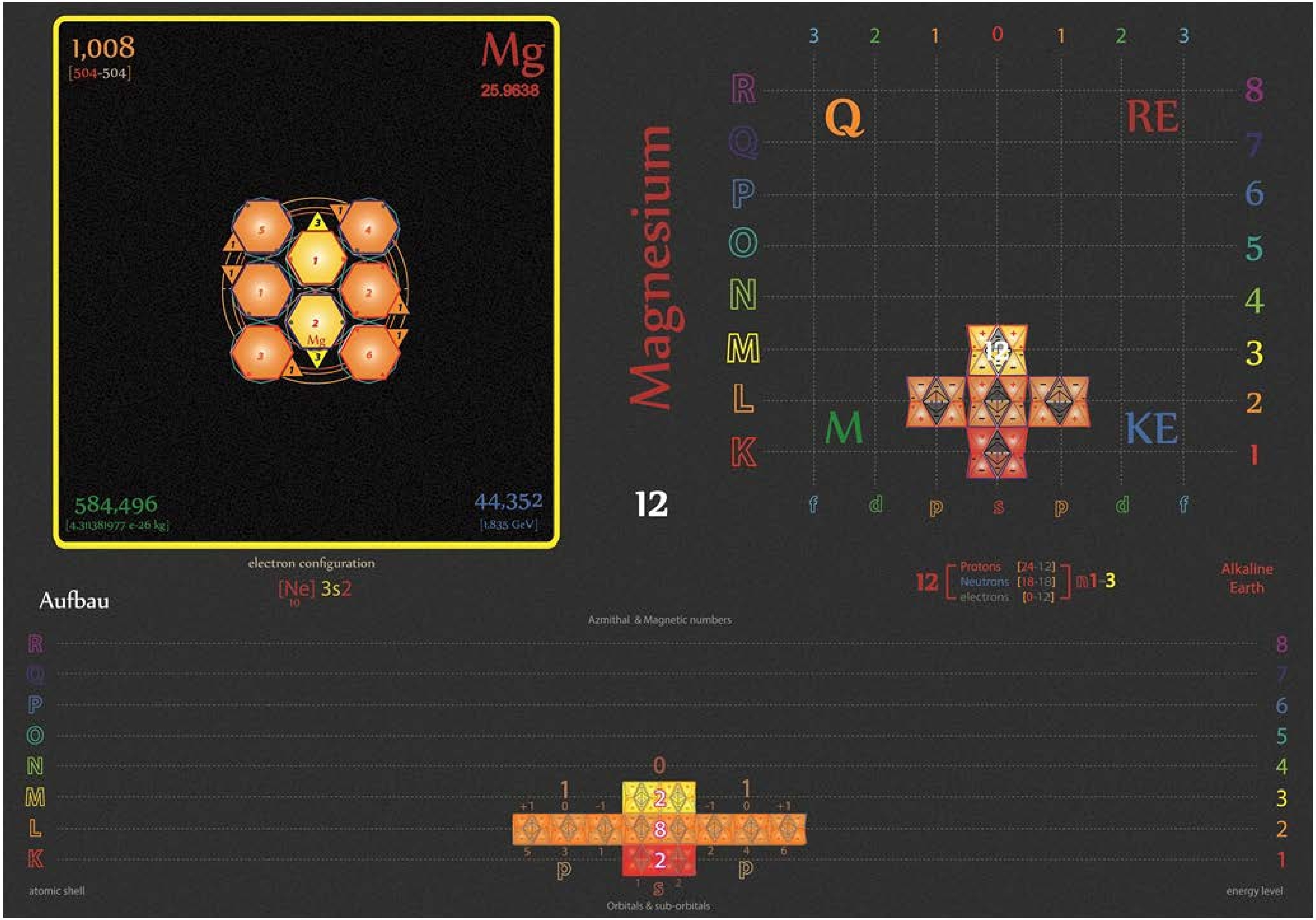
1,008
[504-504]
Mg
25.9638
5
4
3
1
1
2
2
Mg
3
6
584,496
[4.31138977 e-26 kg]
44,352
[1.835 GeV]
electron configuration
[Ne] 3s2
10
Aufbau
Magnesium
12
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
12
8
7
6
5
4
3
2
1
f d p s p d f
12
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Alkaline
Earth
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
+1 0 -1
-1 0 +1
2
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.12 - Magnesium atom

Copyright ABRAHAM [2008] - All rights reserved

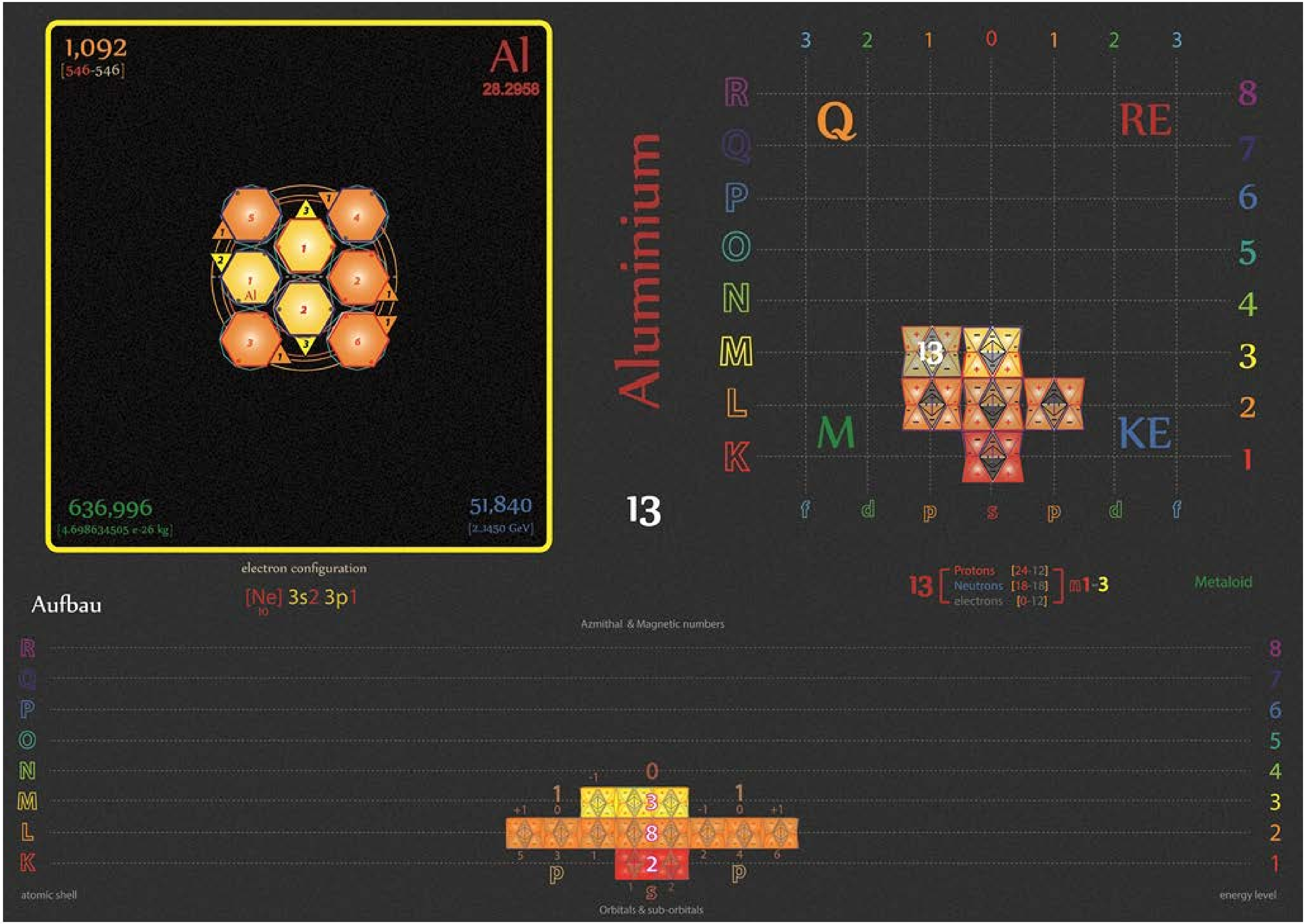
1,092
[546-546]
Al
28.2958
636,996
[4.698634505 e-26 kg]
51,840
[2.1450 GeV]
Aluminium
13
electron configuration
[Ne] 3s2 3p1
10
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
13
13
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Metaloid
Azmithal & Magnetic numbers
0
1
1
-1
+1 0 -1 0 +1
3
8
2
5 3 1 2 4 6
p
s
p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.13 - Aluminium atom

Copyright ABRAHAM [2008] - All rights reserved

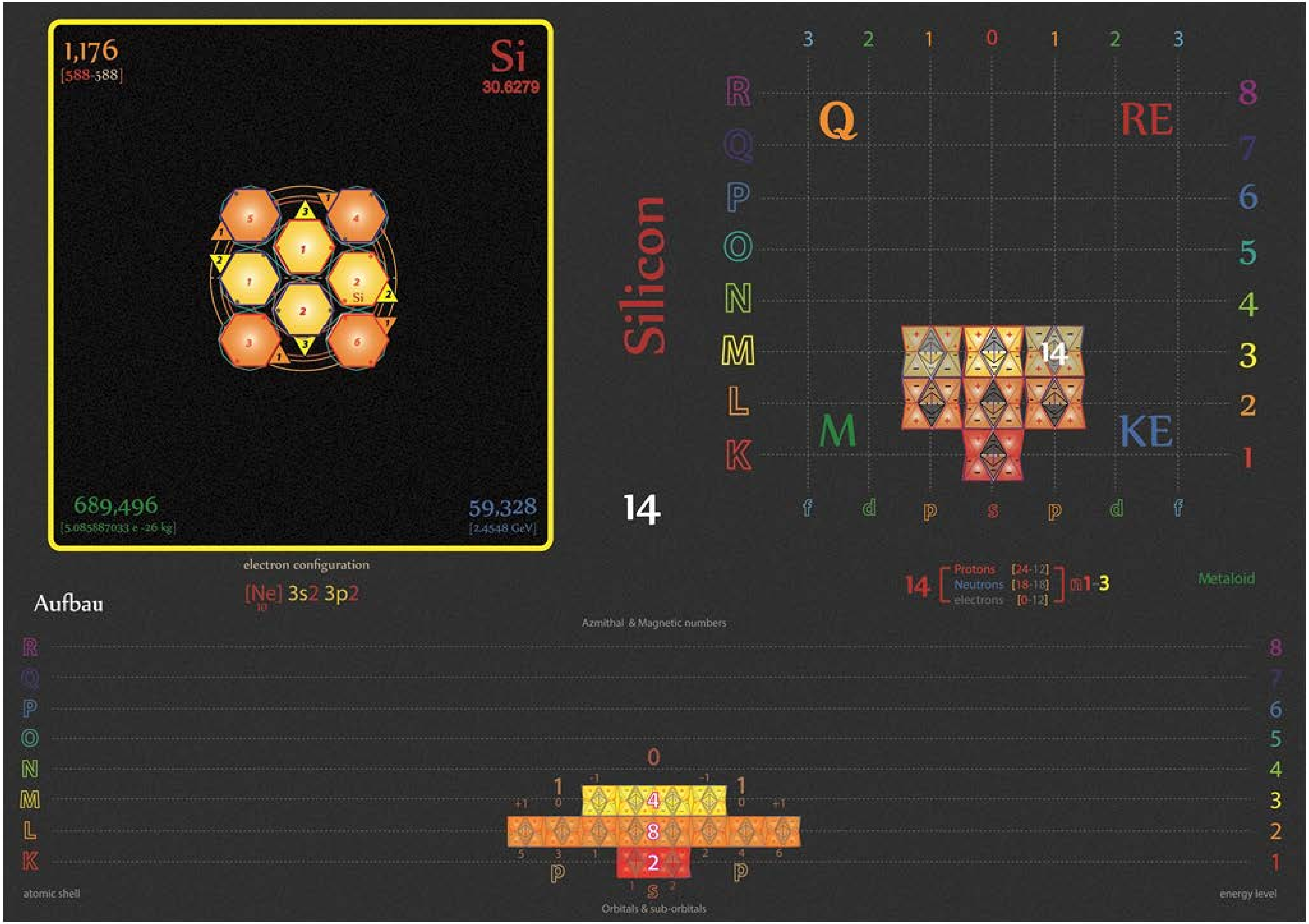
1,176
[588-588]
Si
30.6279
Si
689,496
[5.085887033 e -26 kg]
59,328
[2.4548 GeV]
Silicon
14
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
14
f d p s p d f
14 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
electron configuration
[Ne]10 3s2 3p2
Aufbau
Azmithal & Magnetic numbers
0
1 1
-1 -1
+1 0 0 +1
4
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.14 - Silicon atom

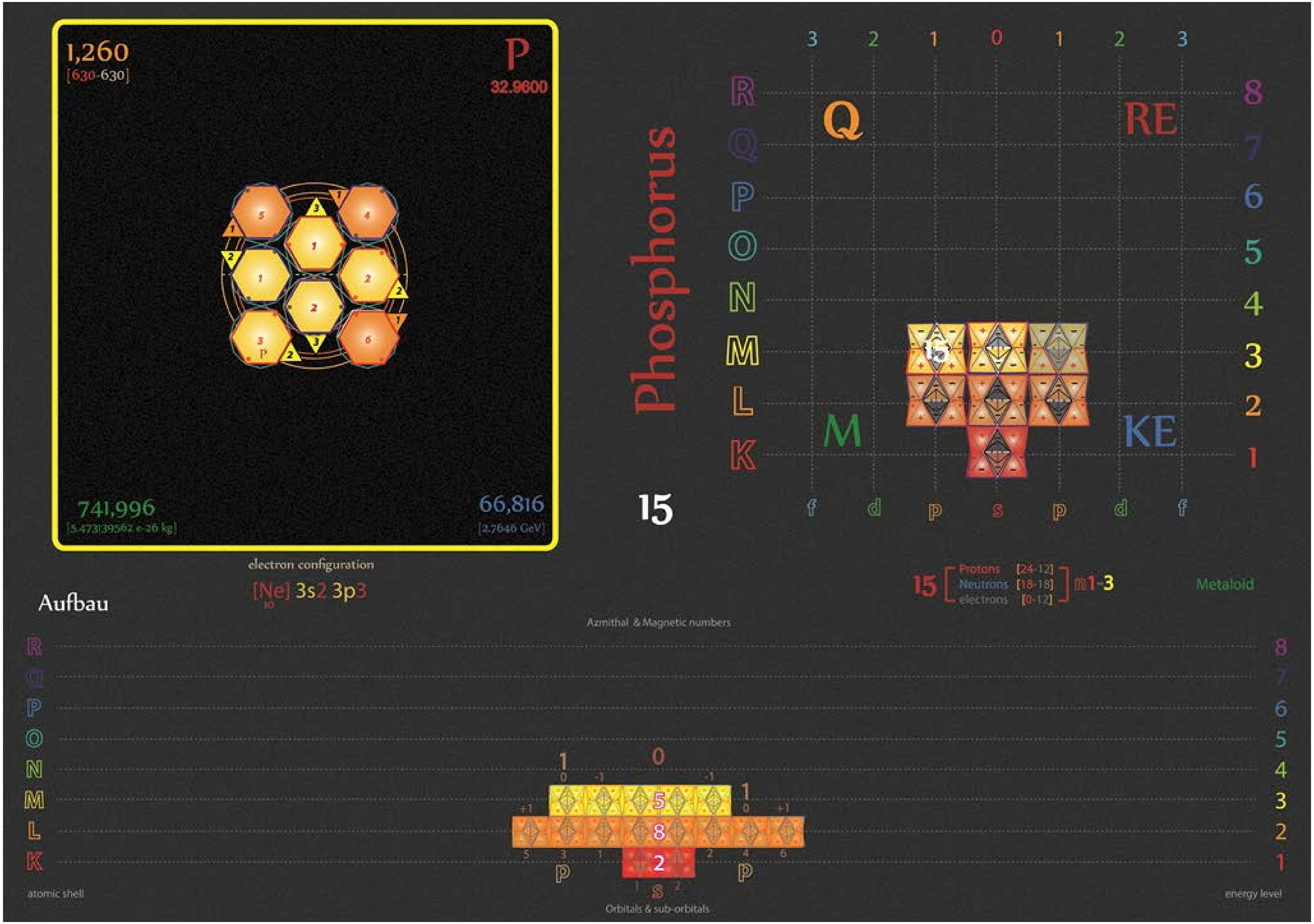
1,260
[630-630]
P
32.9600
741,996
[5.473139562 e-26 kg]
66,816
[2.7646 GeV]
electron configuration
[Ne] 3s2 3p3
10
Phosphorus
15
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
15
f d p s p d f
15 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0
0 -1 -1
+1 1 0 +1
5
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.15 - Phosphorus atom

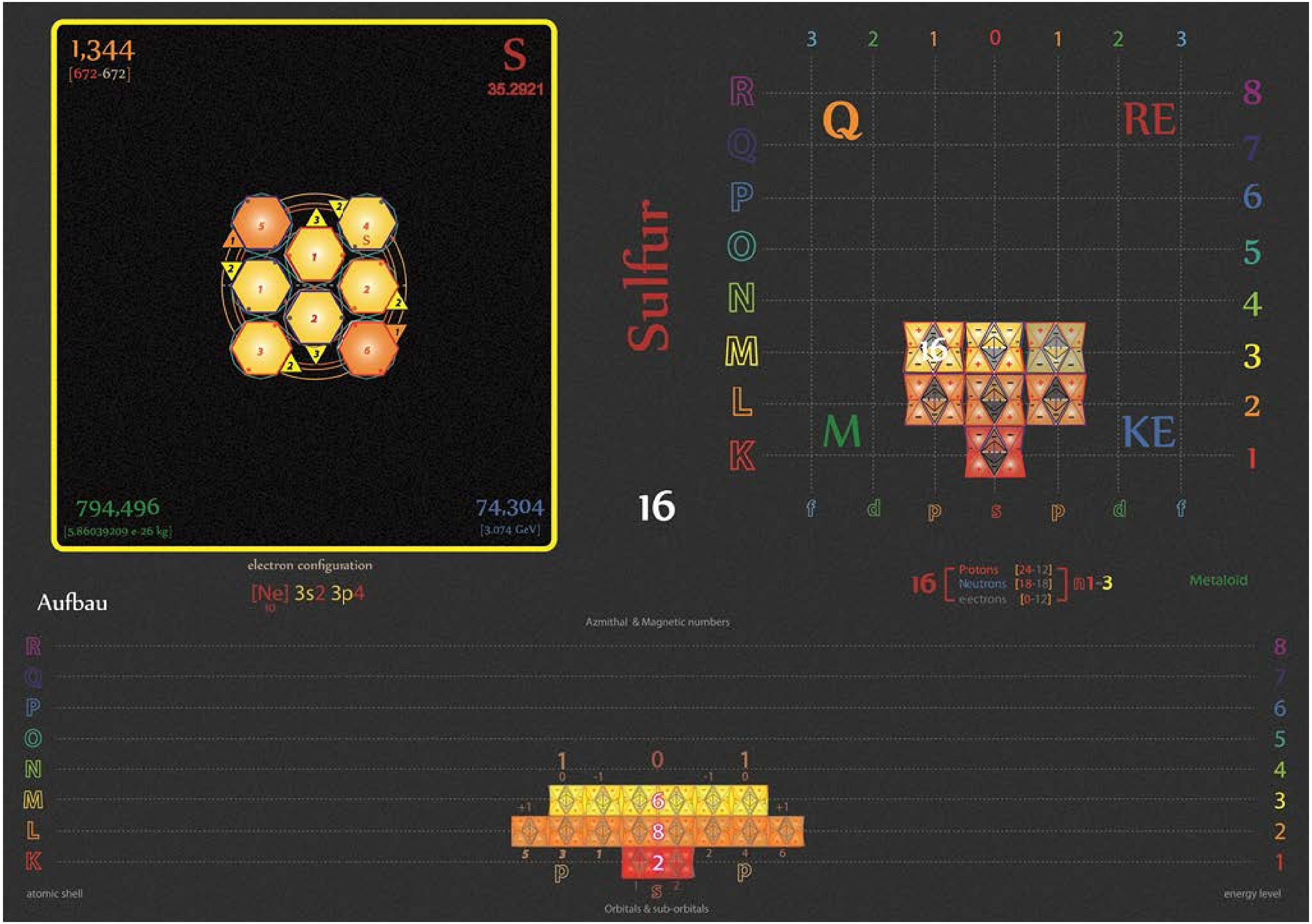
1,344
[672-672]
S
35.2921
794,496
[5.86039209 e-26 kg]
74,304
[3.074 GeV]
Sulfur
16
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
16 [Potons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
electron configuration
[Ne] 3s2 3p4
10
Aufbau
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
+1 +1
6
8
2
5 3 1 2 4 6
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.16 - Sulfur atom

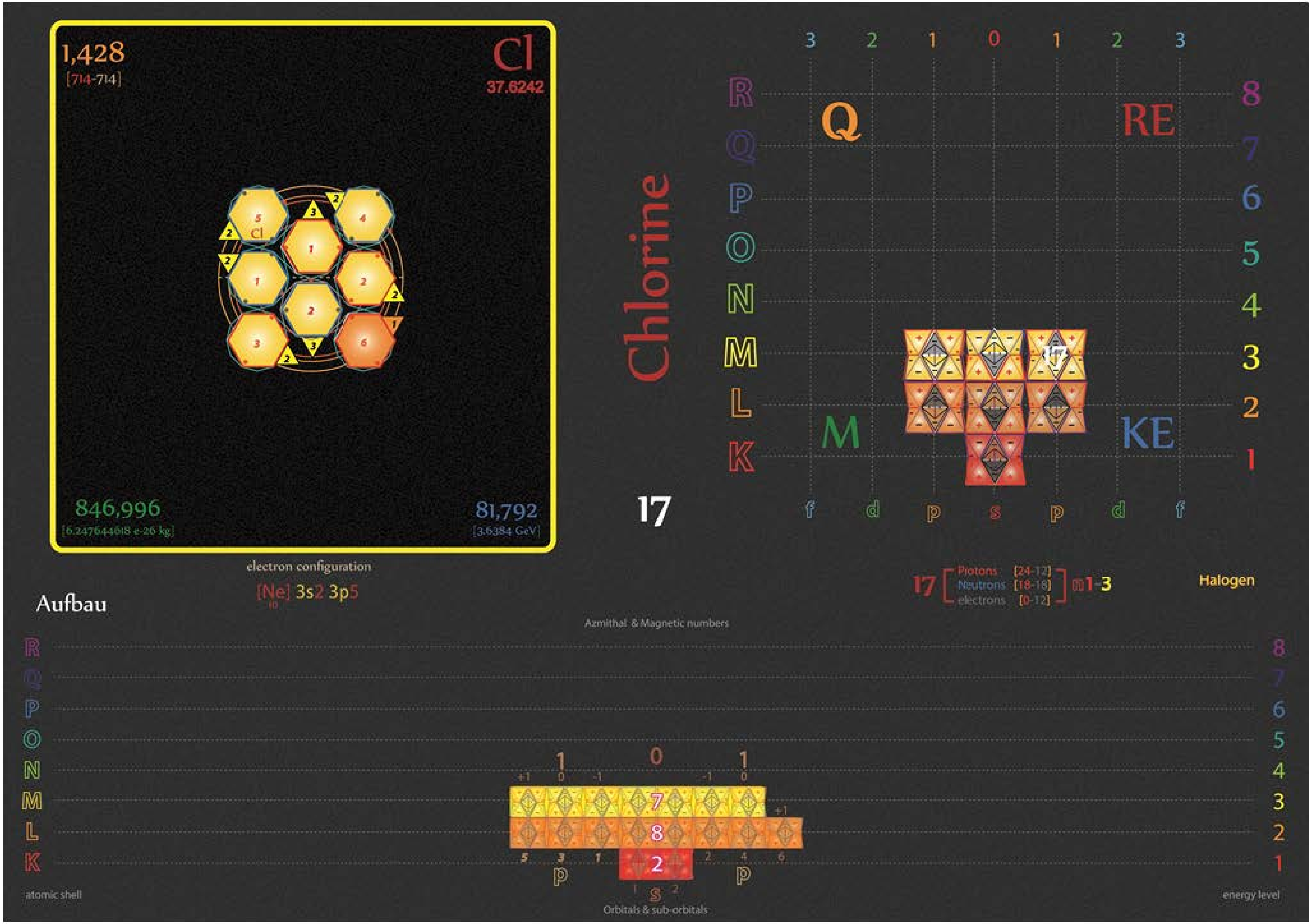
1,428
[714-714]
Cl
37.6242
5
Cl
4
1
1
2
2
3
6
846,996
[6.247644618 e-26 kg]
81,792
[3.6384 GeV]
electron configuration
[Ne] 3s2 3p5
10
Aufbau
Chlorine
17
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
17
f d p s p d f
17 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Halogen
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
7
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.17 - Chlorine atom

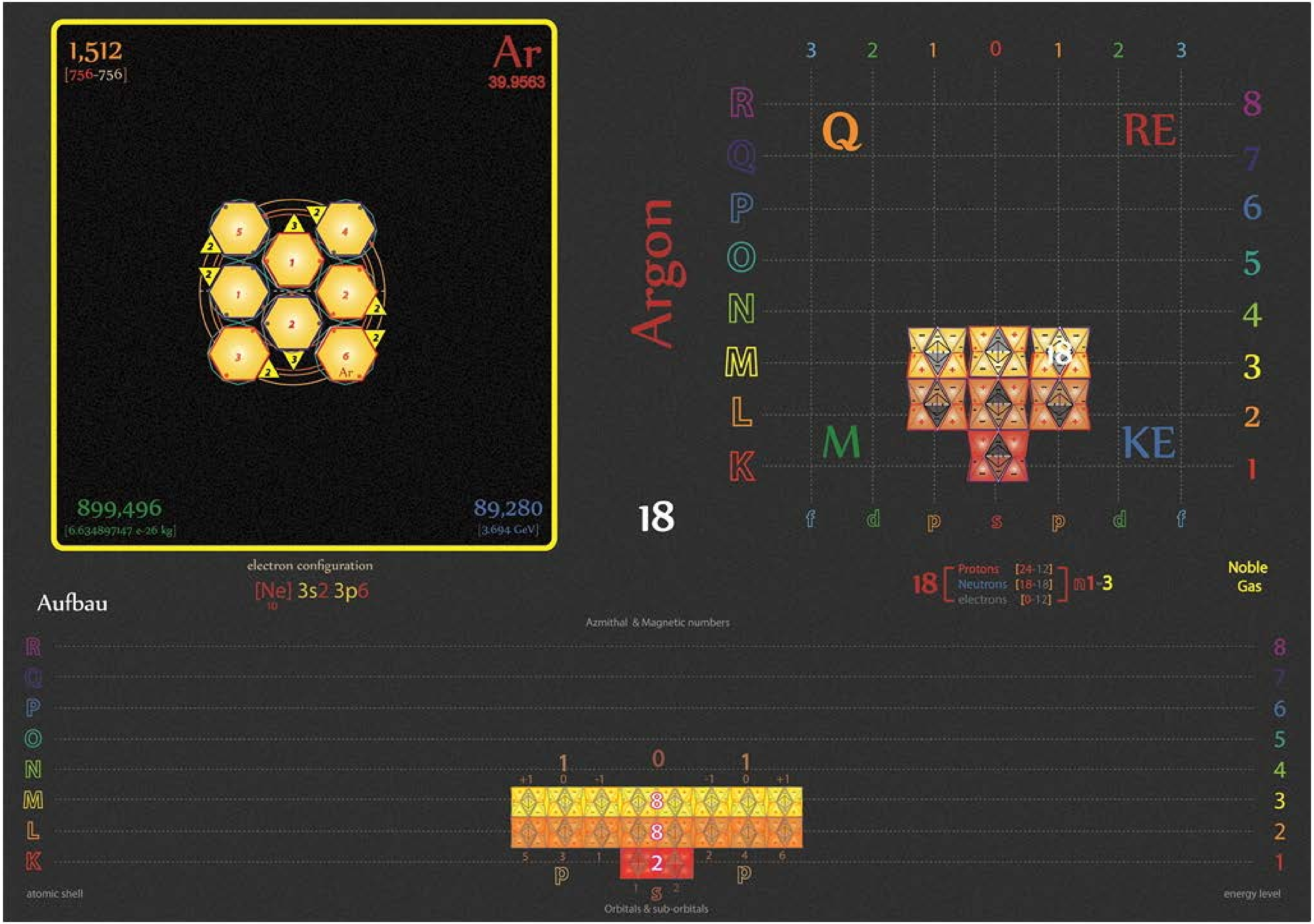
1,512
[756-756]
Ar
39.9563
1 2 3 4 5 6
1 2 3
Ar
899,496
[6.634897147 e-26 kg]
89,280
[3.694 GeV]
Argon
18
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
18
f d p s p d f
electron configuration
[Ne] 3s2 3p6
10
Aufbau
18 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Noble Gas
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0 +1
8
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.18 - Argon atom

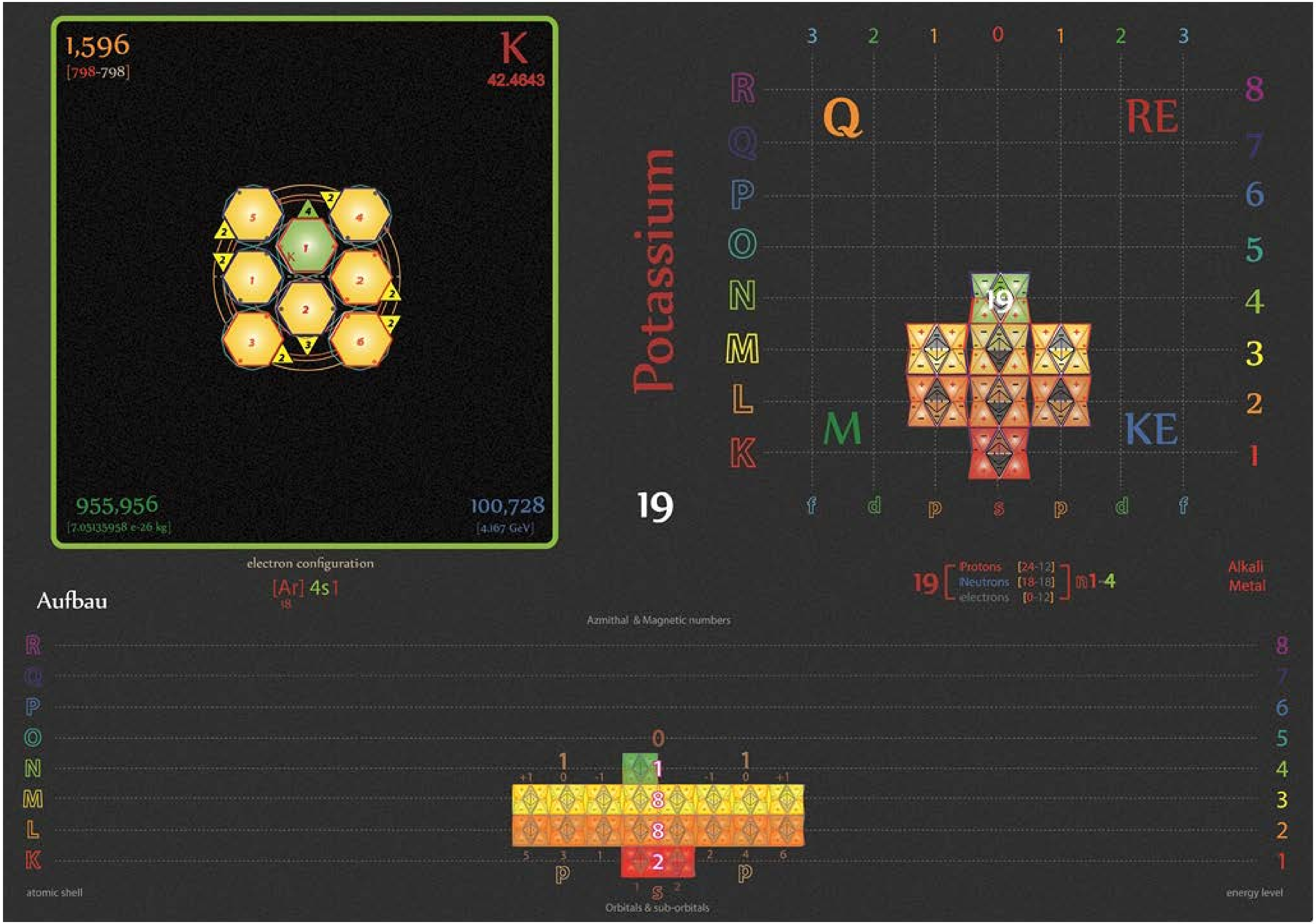
1,596
[798-798]
K
42.4643
5
4
1
2
3
6
955,956
[7.05135958 e-26 kg]
100,728
[4.167 GeV]
Potassium
19
electron configuration
[Ar] 4s1
18
Aufbau
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
19
f
d
p
s
p
d
f
19
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-4
Alkali
Metal
Azmithal & Magnetic numbers
0
1
1
+1
0
-1
-1
0
+1
1
8
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.19 - Potassium atom

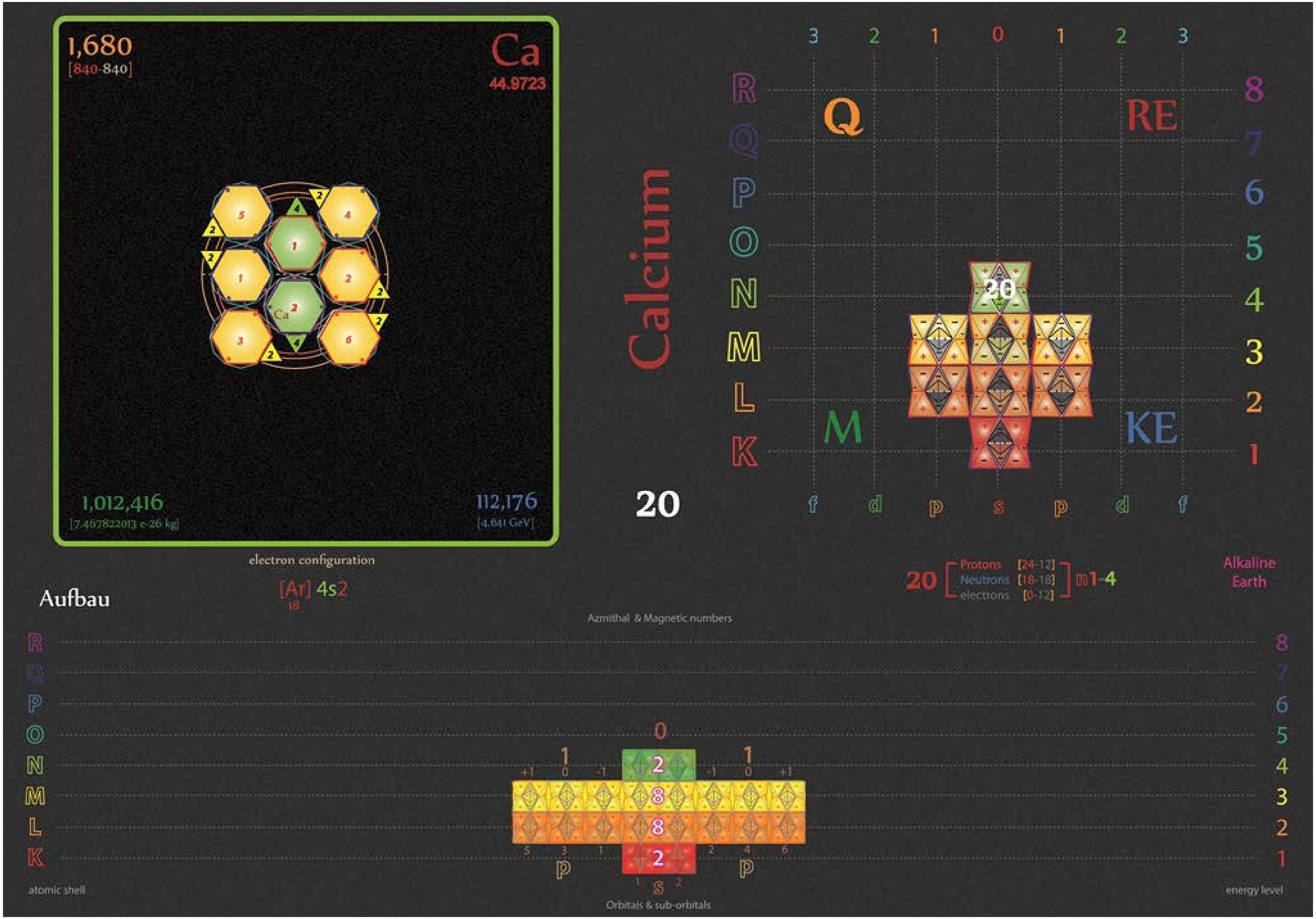
1,680
[840-840]
Ca
44.9723
1,012,416
[7.467822013 e-26 kg]
112,176
[4.641 GeV]
Calcium
20
electron configuration
[Ar] 4s2
18
Aufbau
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
20
20 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
+1 0 -1
-1 0 +1
2
8
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.20 - Calcium atom

Copyright ABRAHAM [2008] - All rights reserved

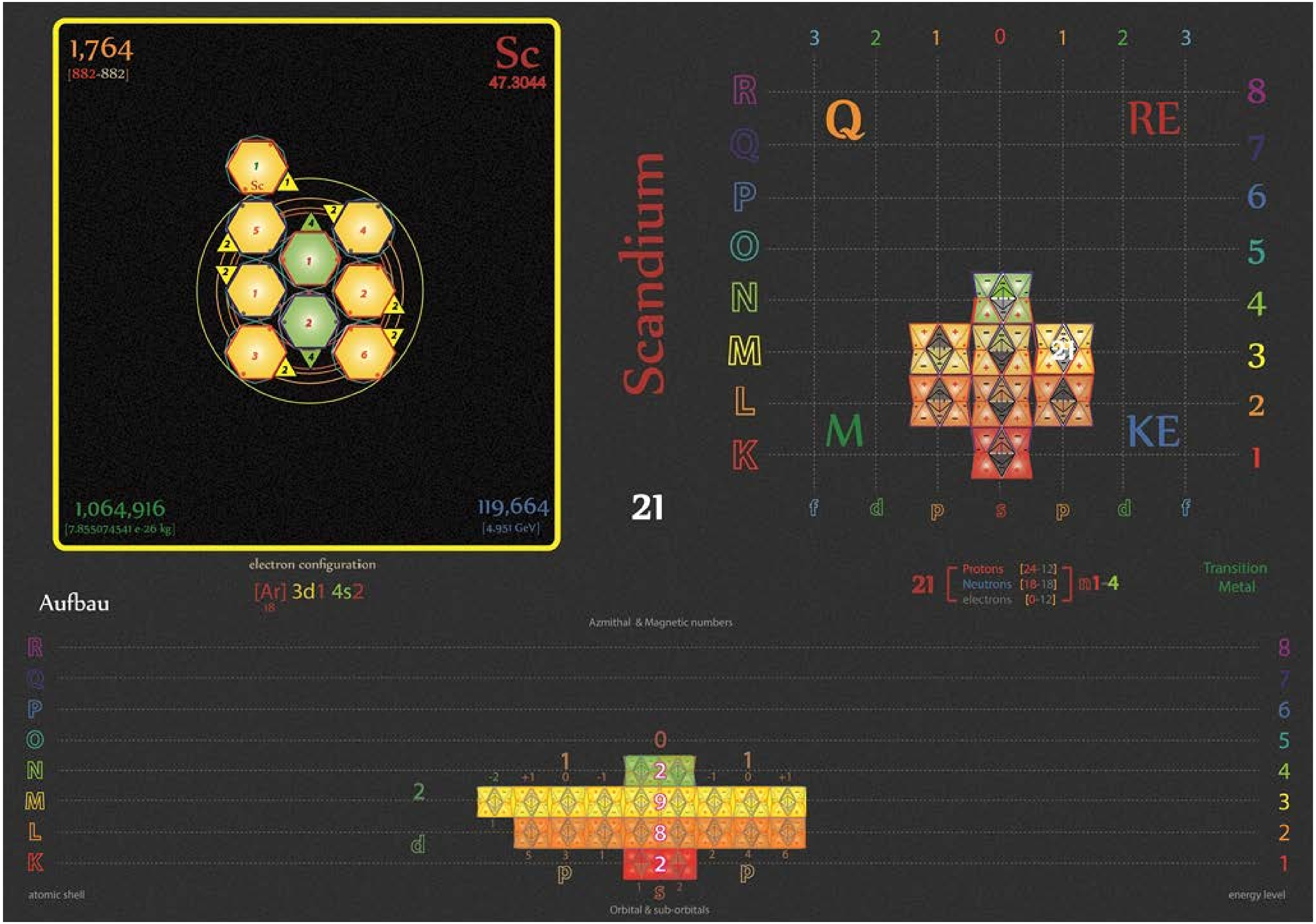
1,764
[882-882]
Sc
47.3044
1,064,916
[7.855074541 e-26 kg]
119,664
[4.951 GeV]
Scandium
21
electron configuration
[Ar] 3d1 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
21
f d p s p d f
21
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
d
-2 +1 0 -1 -1 0 +1
2
9
8
2
5 3 1 2 4 6
p
p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.21 - Scandium atom

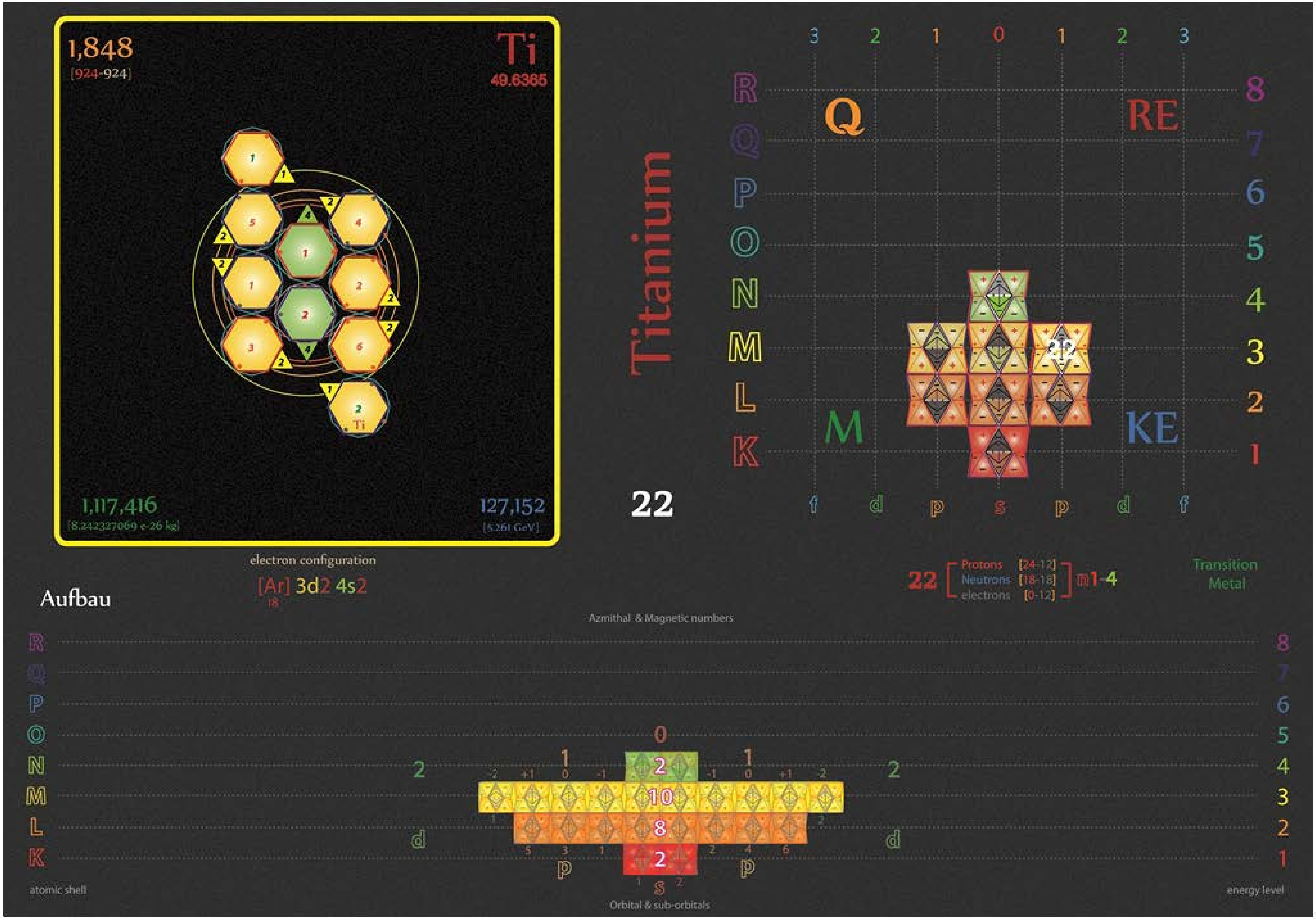
1,848
[924-924]
Ti
49.6365
1,117,416
[8.242327069 e-26 kg]
127,152
[5.261 GeV]
Ti
Titanium
22
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
22
electron configuration
[Ar] 3d2 4s2
18
Aufbau
22
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
10
8
2
-2 +1 0 -1 -1 0 +1 -2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.22 - Titanium atom

Copyright ABRAHAM [2008] - All rights reserved

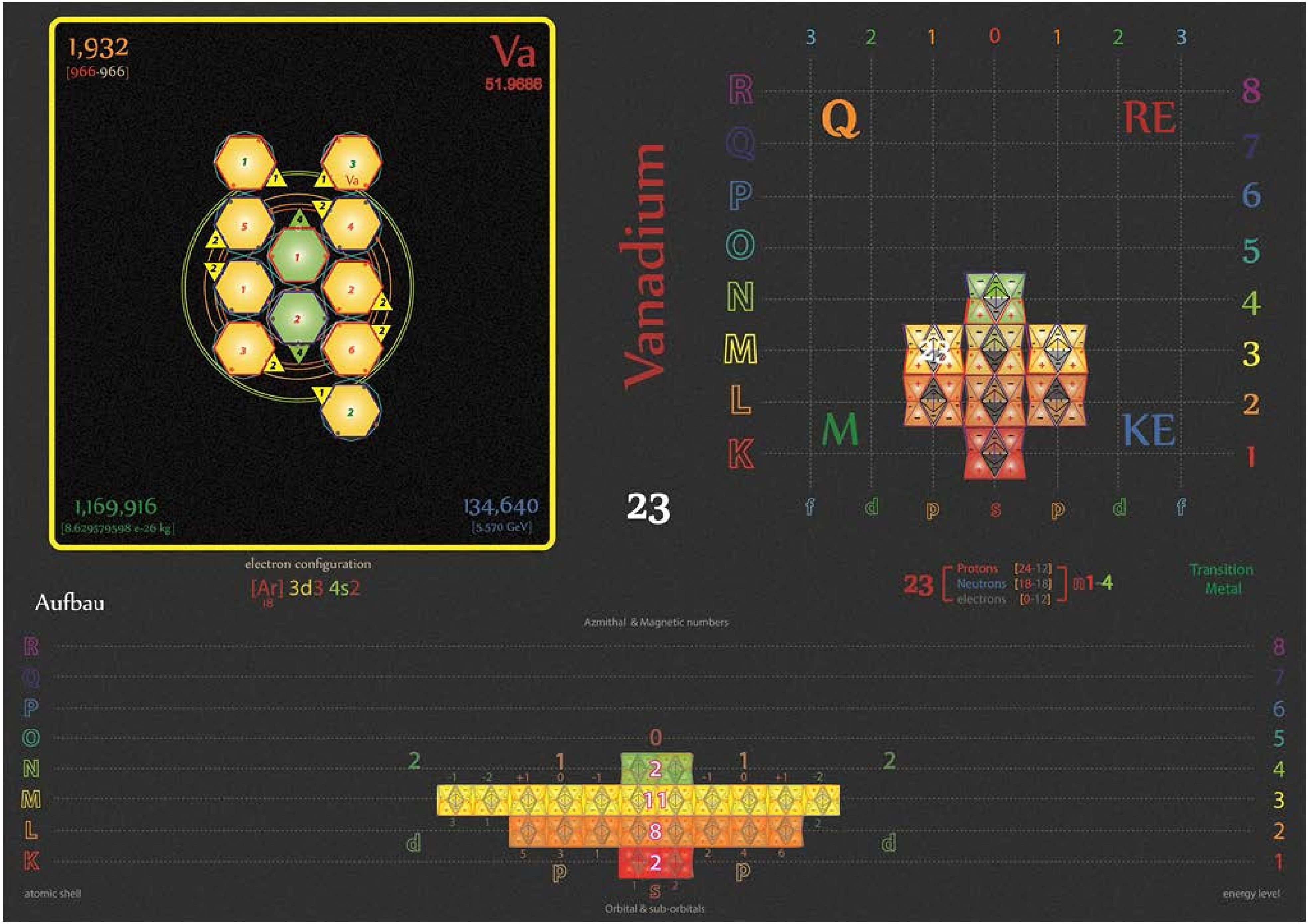
1,932
[966-966]
Va
51.9888
1,169,916
[8.629579598 e-26 kg]
134,640
[5.570 GeV]
Vanadium
23
electron configuration
[Ar] 3d3 4s2
18
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
23 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2
11
8
2
d
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.23 - Vanadium atom

Copyright ABRAHAM [2008] - All rights reserved

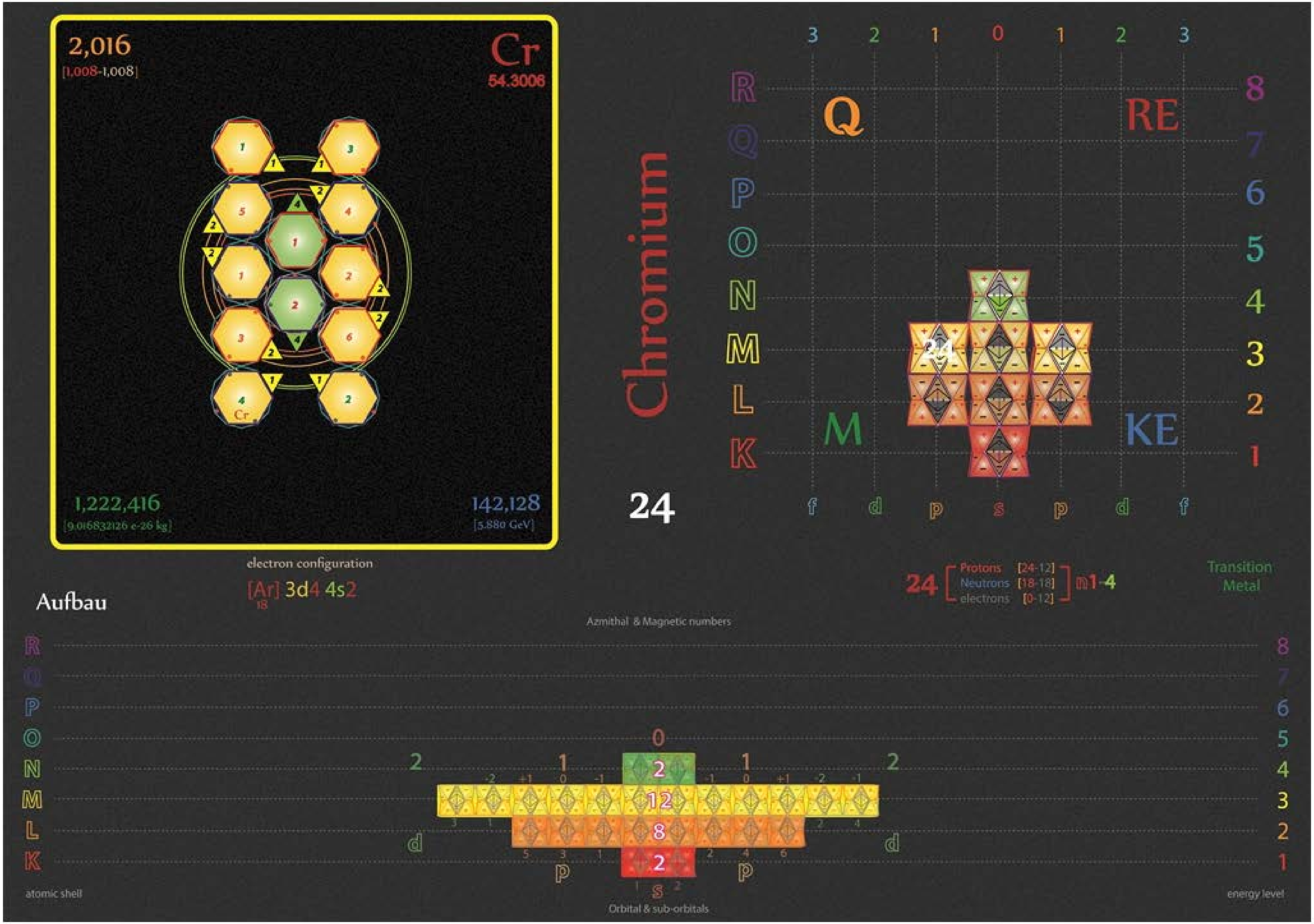
2,016
[1,008-1,008]
Cr
54.3006
1,222,416
[9.016832126 e-26 kg]
142,128
[5.880 GeV]
Chromium
24
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ar] 3d4 4s2
18
Aufbau
24 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
-2 +1 0 -1 -1 0 +1 -2 -1
2
12
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.24 - Chromium atom

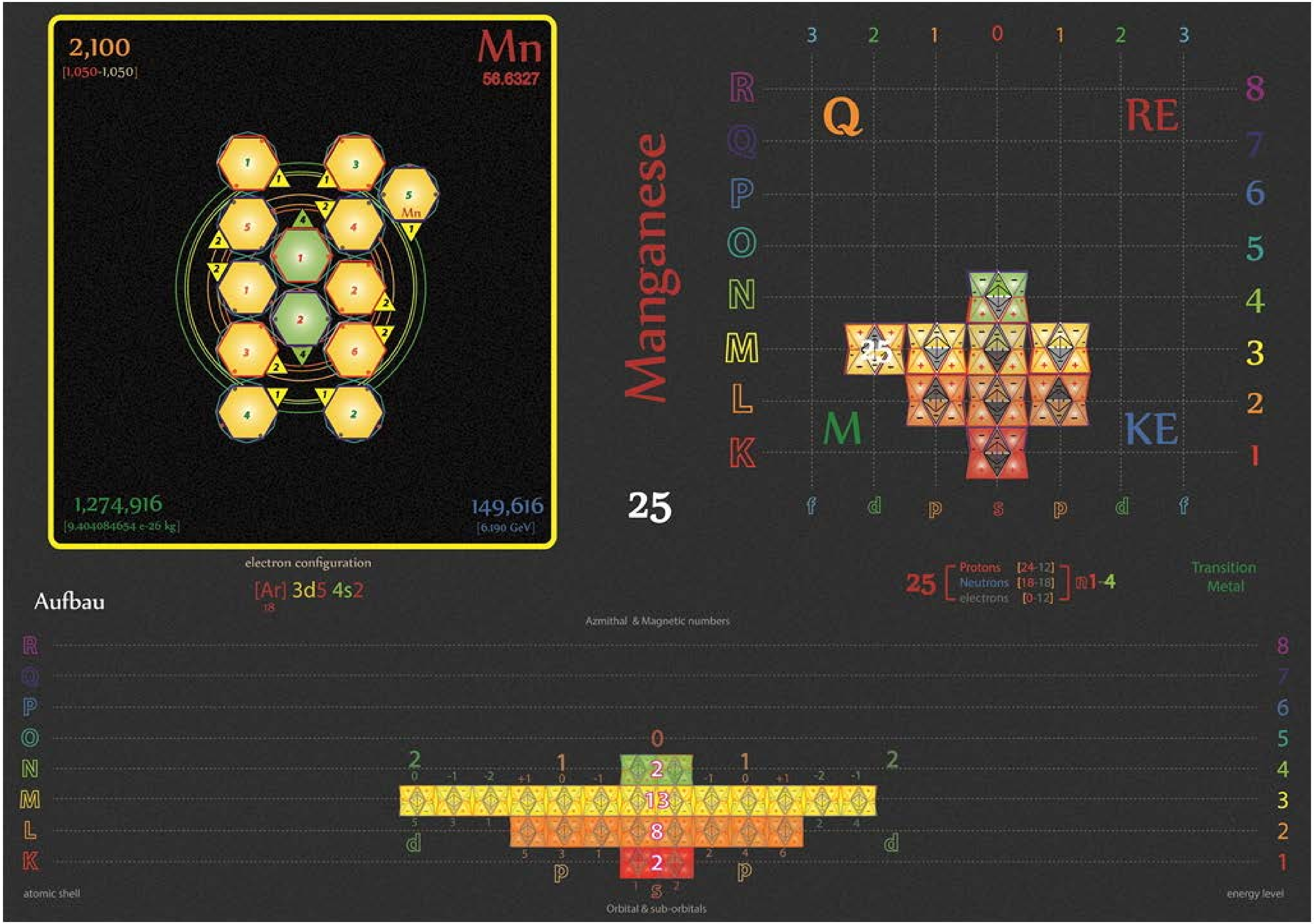
2,100
[1,050-1,050]
Mn
56.6327
1,274,916
[9.404084654 e-26 kg]
149,616
[6.190 GeV]
Manganese
25
electron configuration
[Ar] 3d5 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
25 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
13
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.25 - Manganese atom

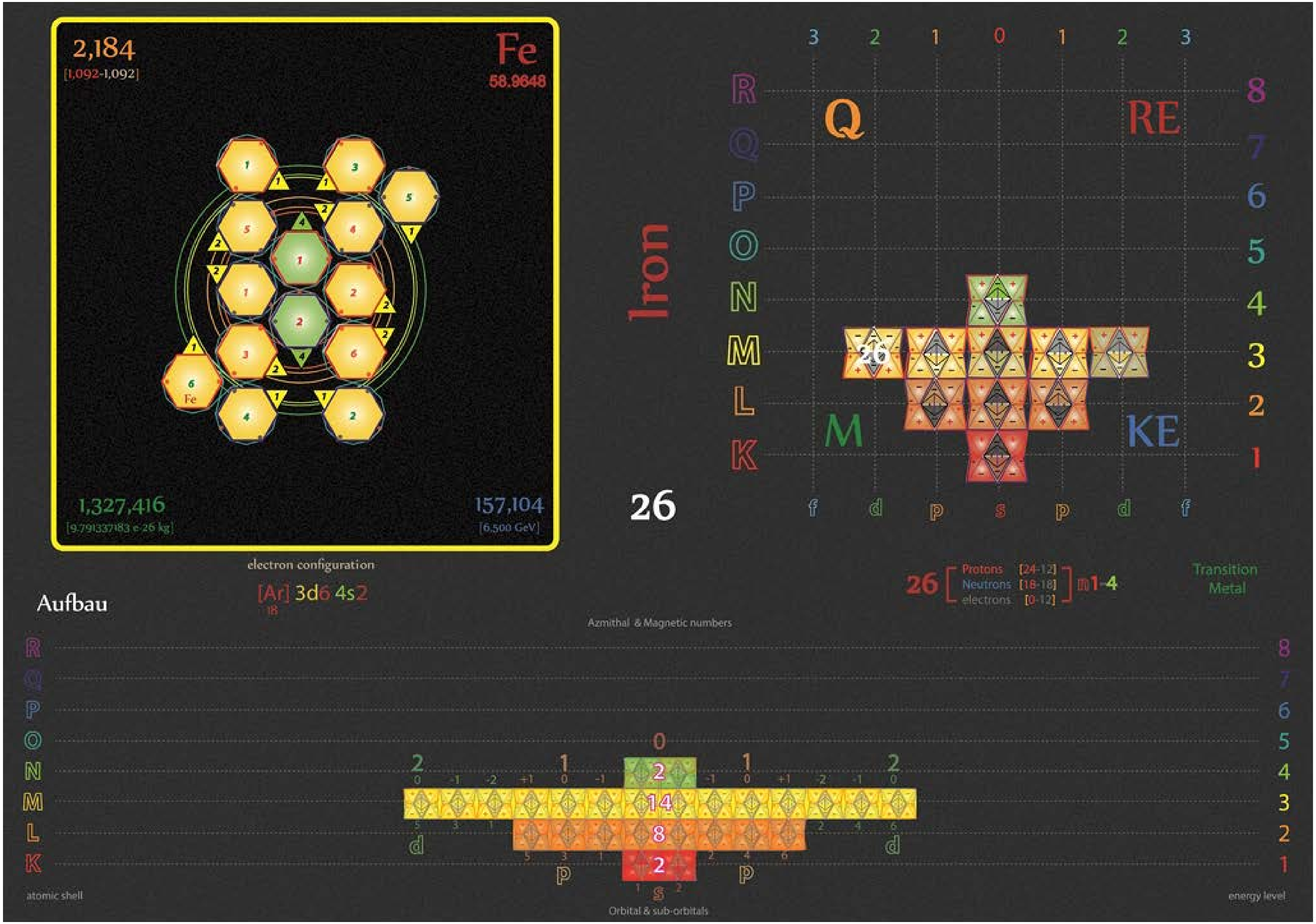
2,184
[1,092-1,092]
Fe
58.9648
1,327,416
[9.791337183 e-26 kg]
157,104
[6,500 GeV]
Iron
26
electron configuration
[Ar] 3d6 4s2
18
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
26 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
2 1 0 1 2
2
14
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.26 - Iron atom

Copyright ABRAHAM [2008] - All rights reserved

2,268
[1,134-1,134]

Co
61.2969

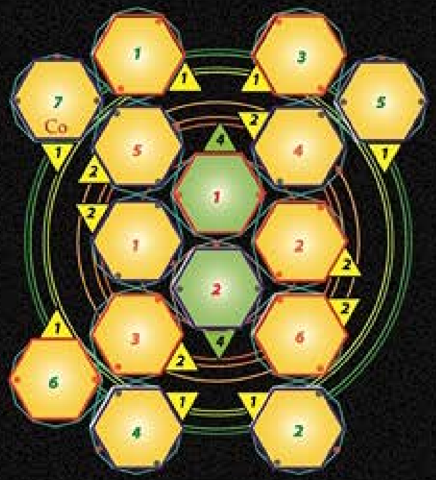
1
3
7
Co
5
5
4
4
1
1
2
2
3
6
6
4
2
1
1
1
1
1
1
2
2
2
2
2
2
1
1
1
1

1,379,916
[1.017858971 e-25 kg]

164,592
[6.810 GeV]

electron configuration

[Ar] 3d7 4s2
18

Cobalt

27

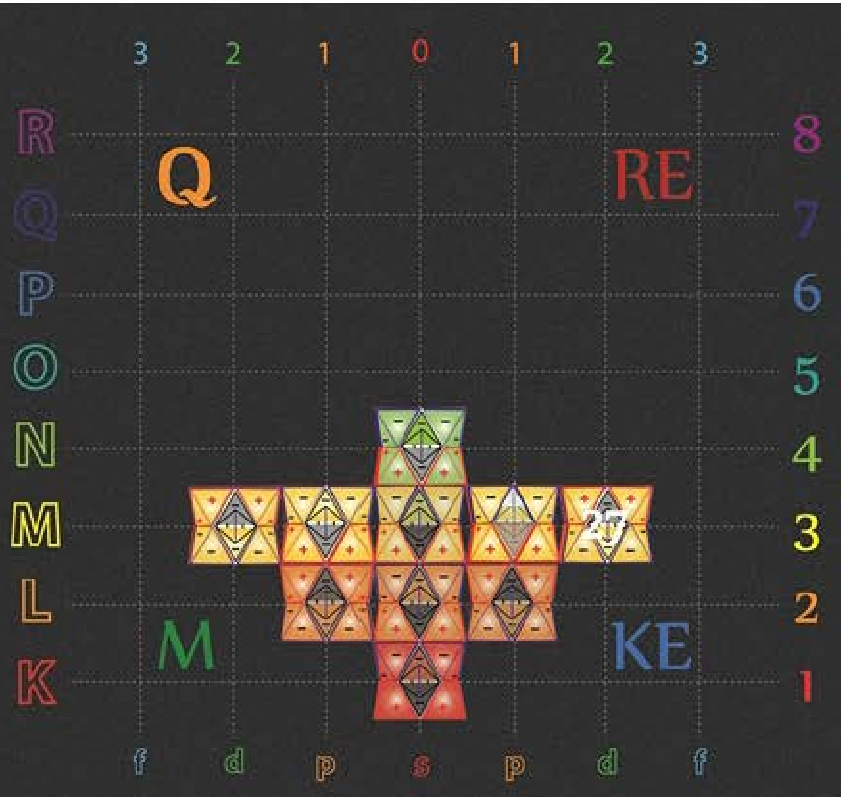
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
27
f
d
p
s
p
d
f

27 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Transition Metal

Aufbau

Azmithal & Magnetic numbers

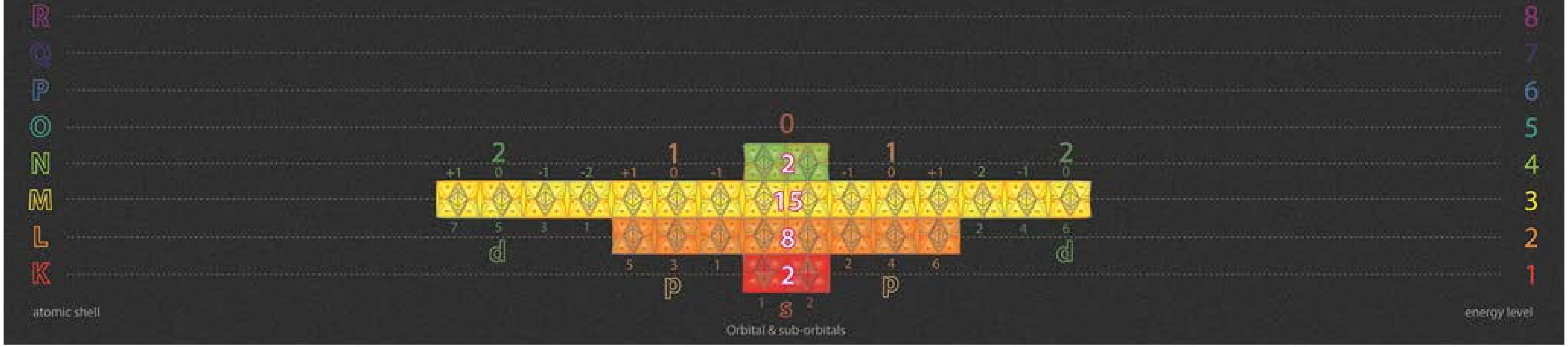
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
2
15
8
2
7
5
3
1
2
4
6
d
d
5
3
1
2
4
6
p
p
1
2
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.27 - Cobalt atom

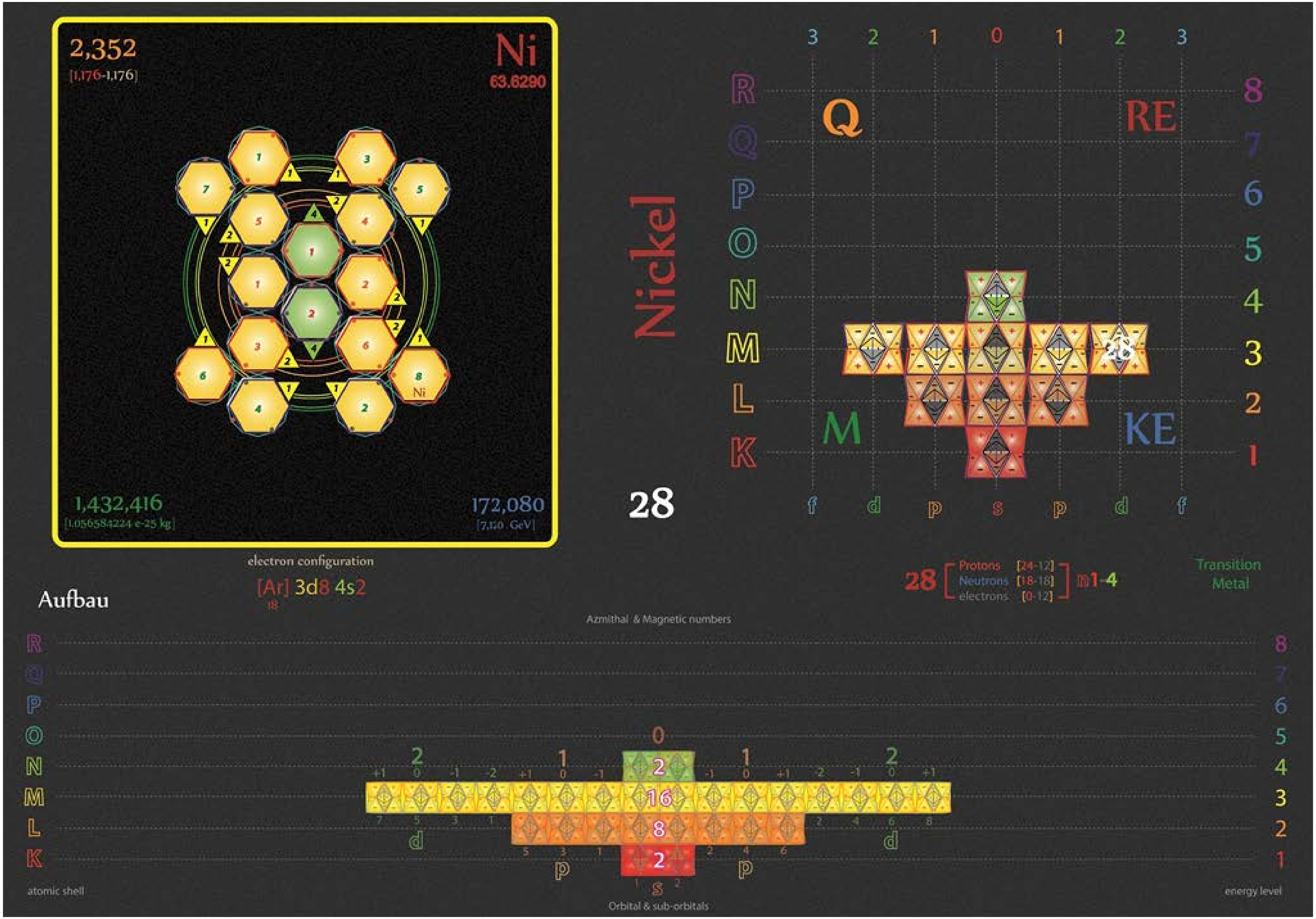
2,352
[1,176-1,176]
Ni
63.6290
1,432,416
[1.056584224 e-25 kg]
172,080
[7,150 GeV]
Nickel
28
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ar] 3d8 4s2
18
Aufbau
28
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
2
16
8
2
7 5 3 1 2 4 6 8
5 3 1 2 4 6
1 2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.28 - Nickel atom

Copyright ABRAHAM [2008] - All rights reserved

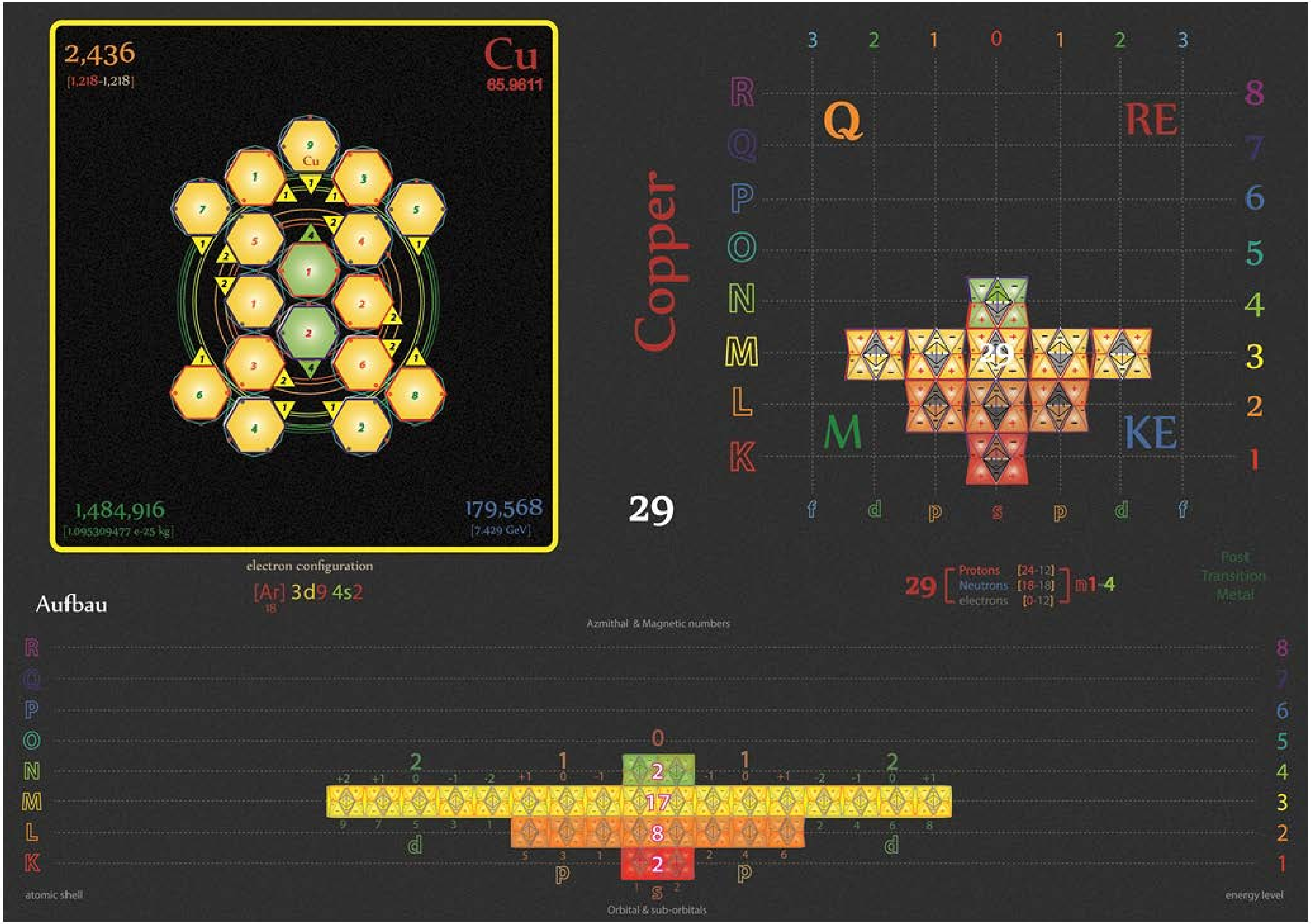
2,436
[1,218-1,218]
Cu
65.9811
Cu
1,484,916
[1.095309477 e-25 kg]
179,568
[7.429 GeV]
Copper
29
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Ar] 3d9 4s2
18
Aufbau
29 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Post Transition Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
2
17
8
2
9 7 5 3 1 2 4 6 8
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.29 - Copper atom

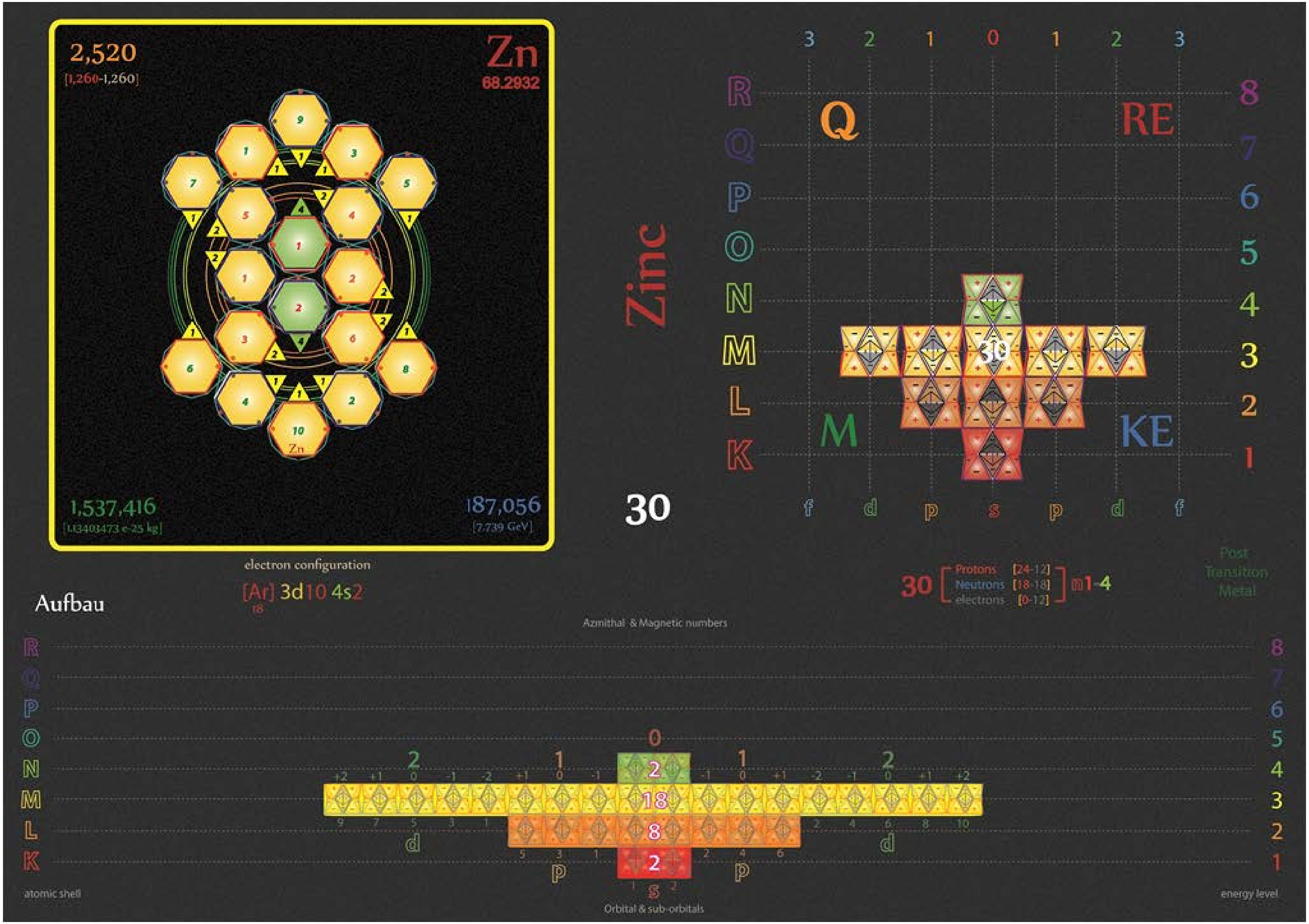
2,520
[1,260-1,260]
Zn
68.2932
Zn
1,537,416
[1.13403473 e-25 kg]
187,056
[7.739 GeV]
Zinc
30
electron configuration
[Ar] 3d10 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
30
f d p s p d f
30
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-4
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.30 - Zinc atom

Copyright ABRAHAM [2008] - All rights reserved

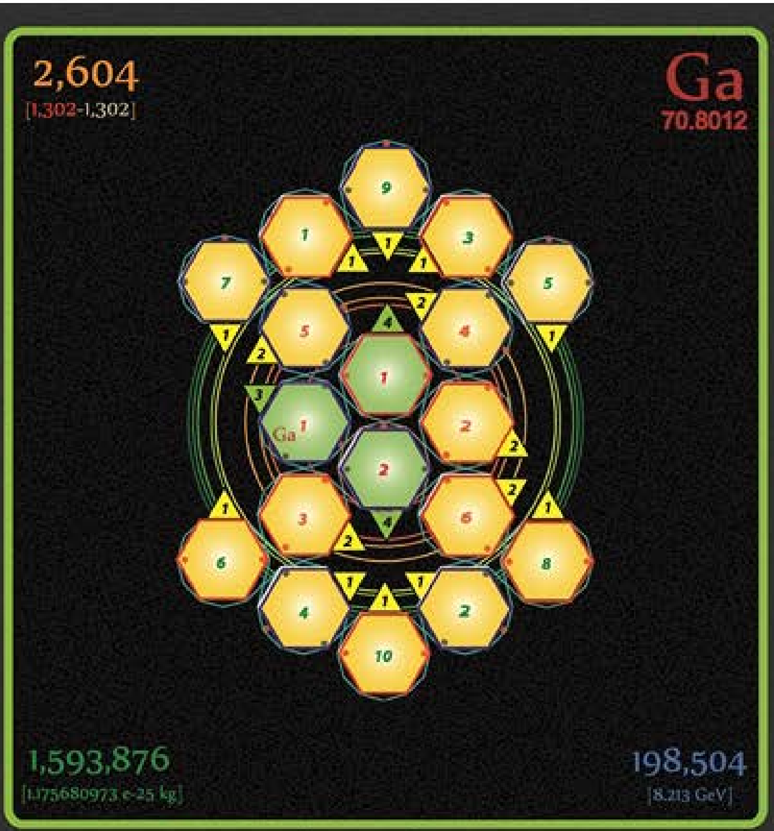
Ga
9
1
3
7
5
1
1
2
5
4
4
1
1
Ga
2
2
3
2
3
6
2
1
1
6
8
4
1
1
2
10

2,604
[1,302-1,302]

Ga
70.8012

1,593,876
[1.175680973 e-25 kg]

198,504
[8.213 GeV]

electron configuration

[Ar] 3d10 4s2 4p1
18

Gallium

31

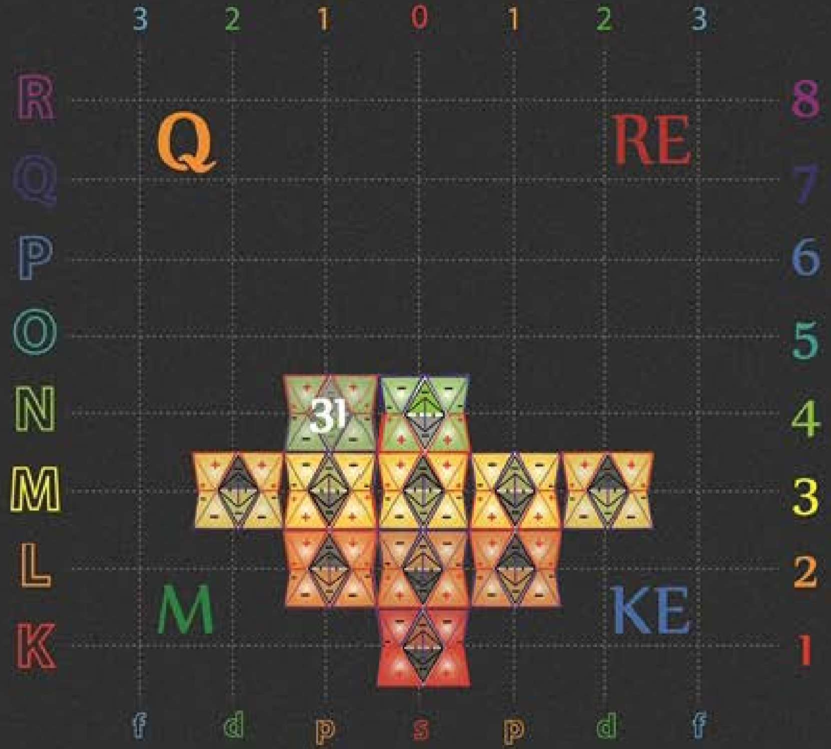
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
31
f
d
p
s
p
d
f

31 [Protons [24-12] / Neutrons [18-18] / electrons [0-12]] m1-4

Metaloid

Aufbau

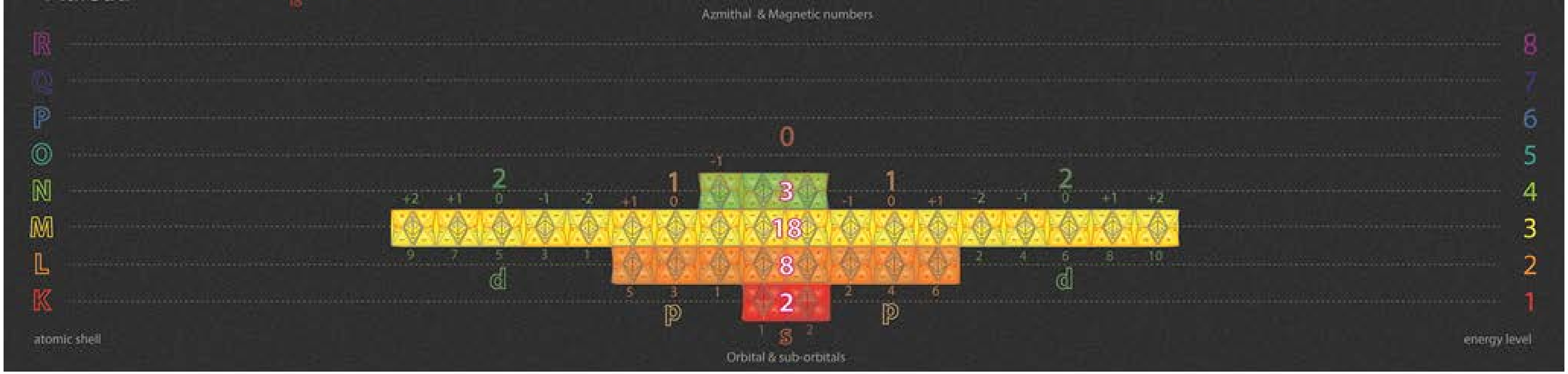
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
-1
2
1
1
2
+2
+1
0
-1
-2
+1
0
-1
0
+1
-2
-1
0
+1
+2
3
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.31 - Gallium atom

Copyright ABRAHAM [2008] - All rights reserved

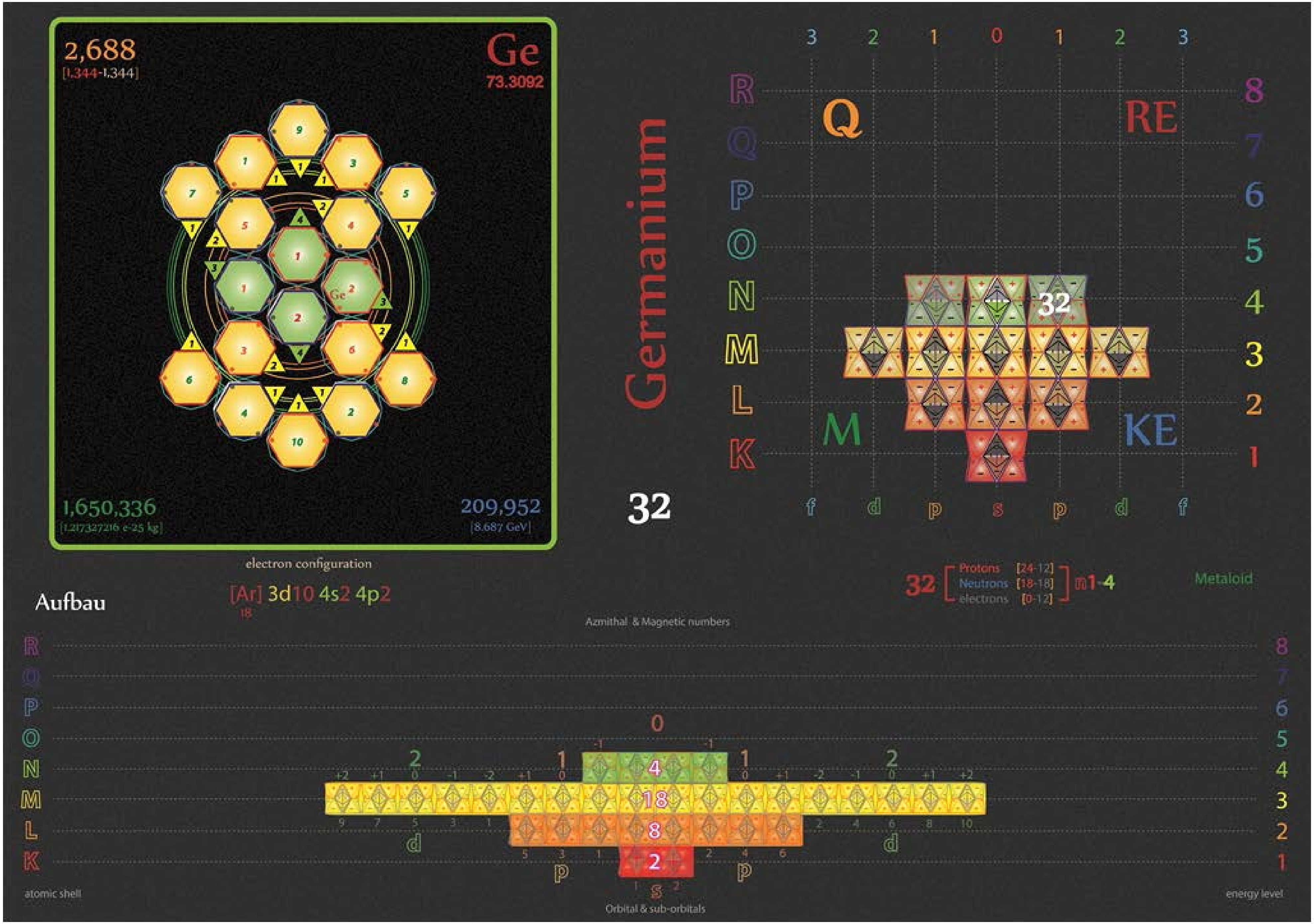
2,688
[1,344-1,344]
Ge
73.3092
Ge
1,650,336
[1.217327216 e-25 kg]
209,952
[8.687 GeV]
Germanium
32
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
32
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[Ar] 3d10 4s2 4p2
18
32 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
0
2 1 1 2
-1 -1
+2 +1 0 -1 -2 +1 0 0 +1 -2 -1 0 +1 +2
4
18
8
2
9 7 5 3 1 2 4 6 8 10
5 3 1 2 4 6
d p s p d
1 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.32 - Germanium atom

Copyright ABRAHAM [2008] - All rights reserved

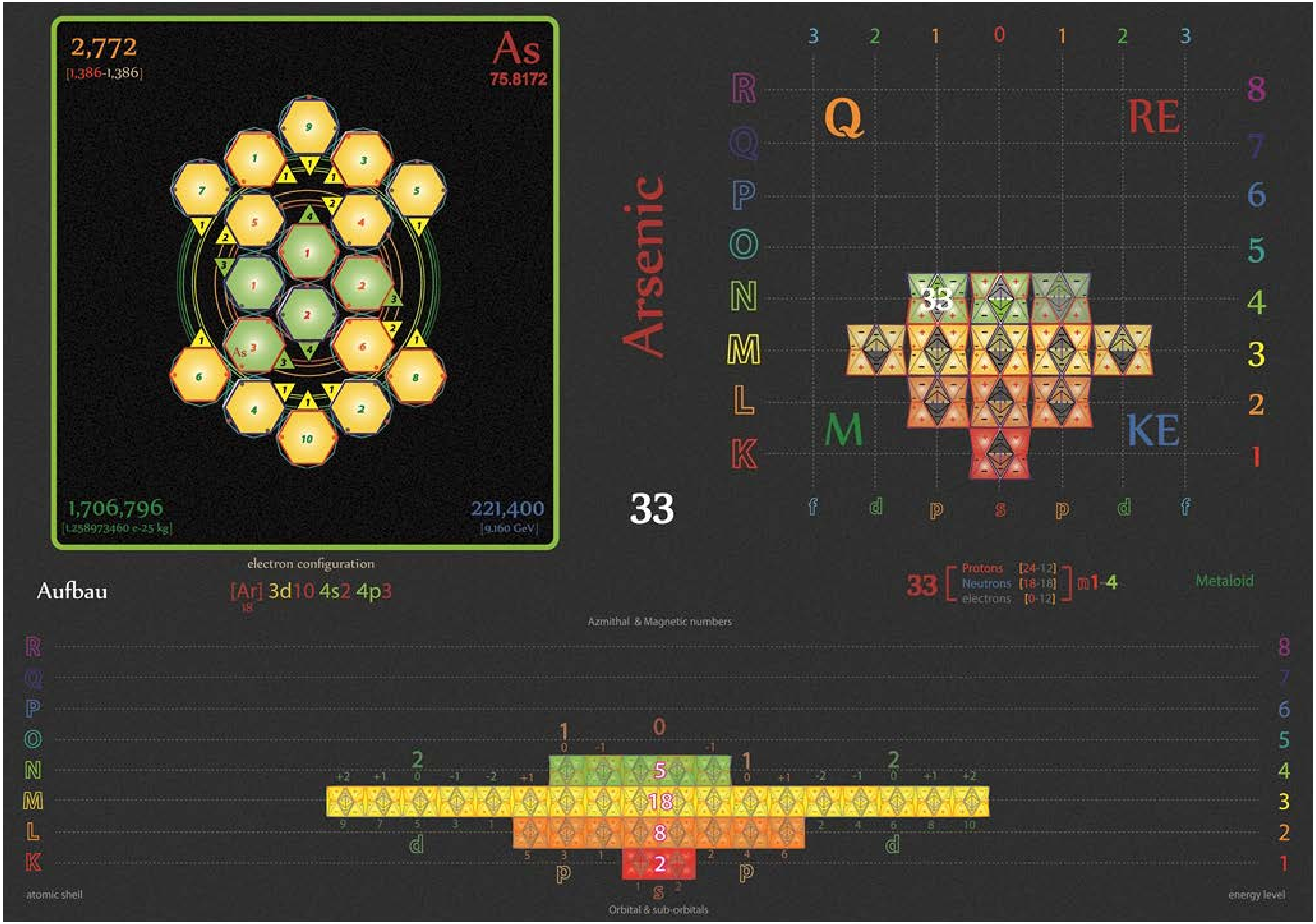
2,772
[1,386-1,386]
As
75.8172
1,706,796
[1.258973460 e-25 kg]
221,400
[9.160 GeV]
Arsenic
33
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
33
f d p s p d f
electron configuration
Aufbau
[Ar] 3d10 4s2 4p3
18
33 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
2 2
5
18
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.33 - Arsenic atom

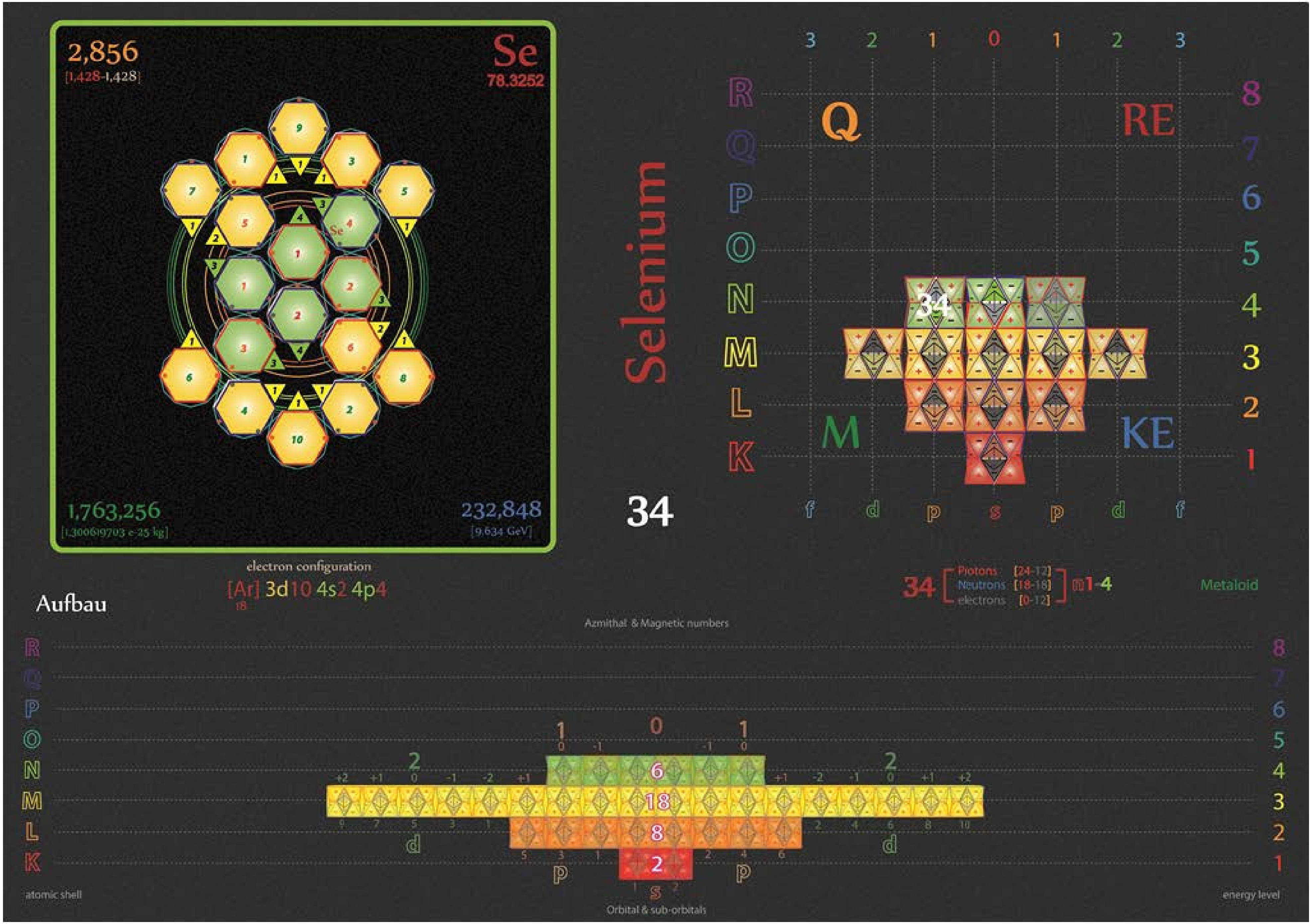
2,856
[1,428-1,428]
Se
78.3252
1,763,256
[1,300619703 e-25 kg]
232,848
[9,634 GeV]
Selenium
34
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
34 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
electron configuration
[Ar] 3d10 4s2 4p4
18
Aufbau
Azmithal & Magnetic numbers
6
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.34 - Selenium atom

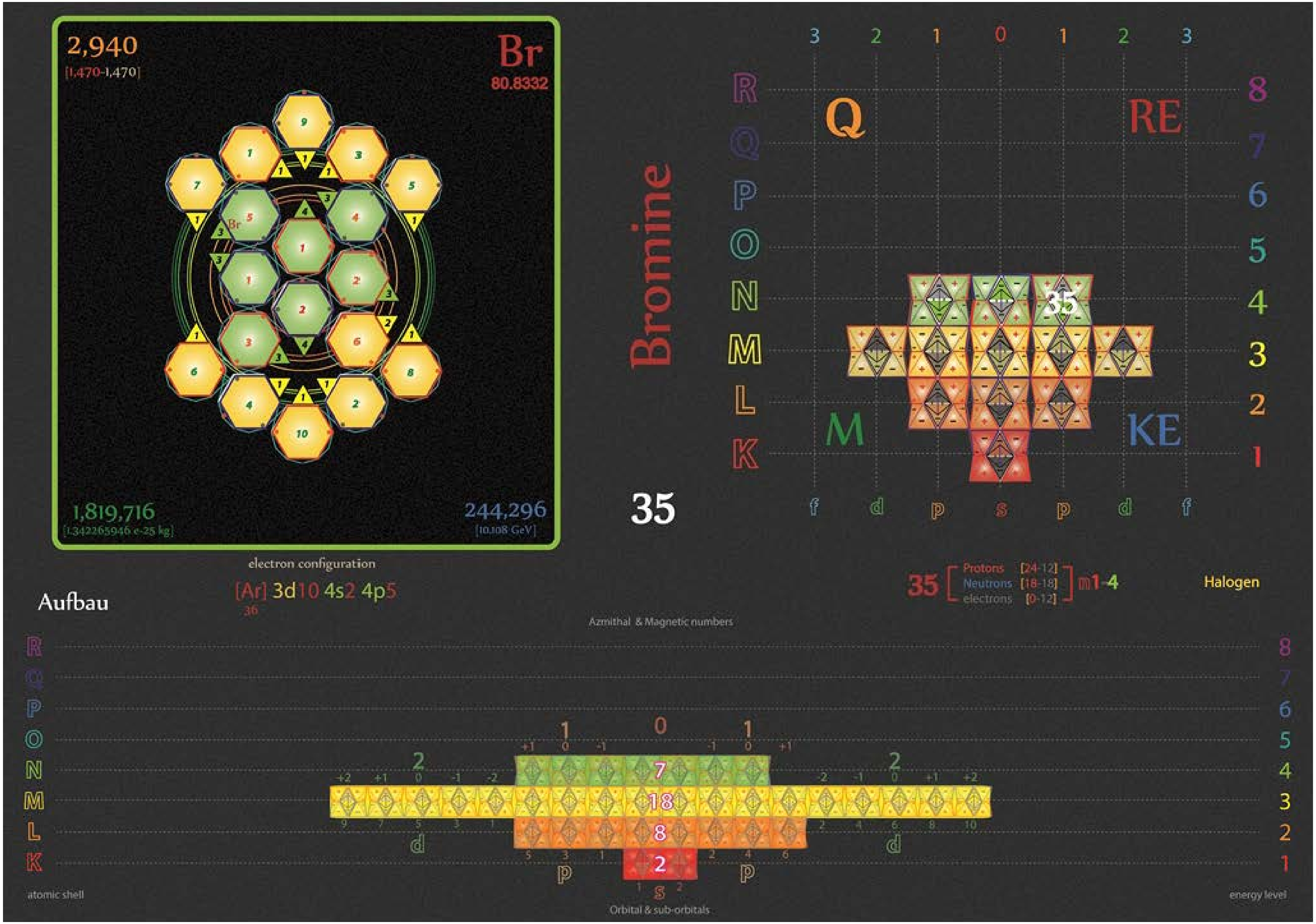
2,940
[1,470-1,470]
Br
80.8332
1,819,716
[1.342265946 e-25 kg]
244,296
[10.108 GeV]
Bromine
35
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
35
f d p s p d f
electron configuration
[Ar] 3d10 4s2 4p5
36
Aufbau
35
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Halogen
Azmithal & Magnetic numbers
7
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.35 - Bromine atom

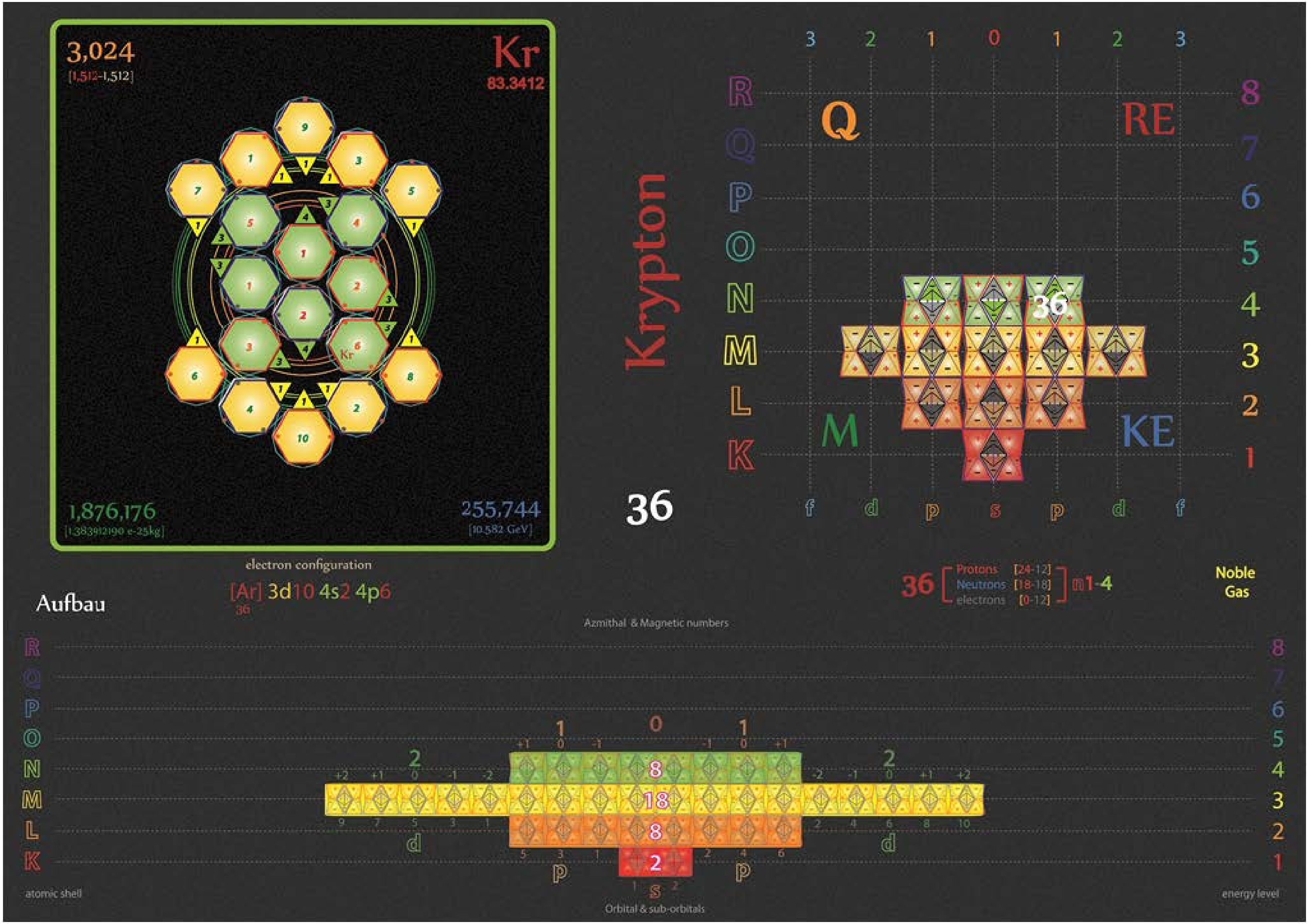
3,024
[1,512-1,512]
Kr
83.3412
1,876,176
[1.383912190 e-25kg]
255,744
[10.582 GeV]
Krypton
36
electron configuration
[Ar] 3d10 4s2 4p6
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
36 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Noble Gas
Azimthal & Magnetic numbers
8
18
8
2
d
p
s
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.36 - Krypton atom

Copyright ABRAHAM [2008] - All rights reserved

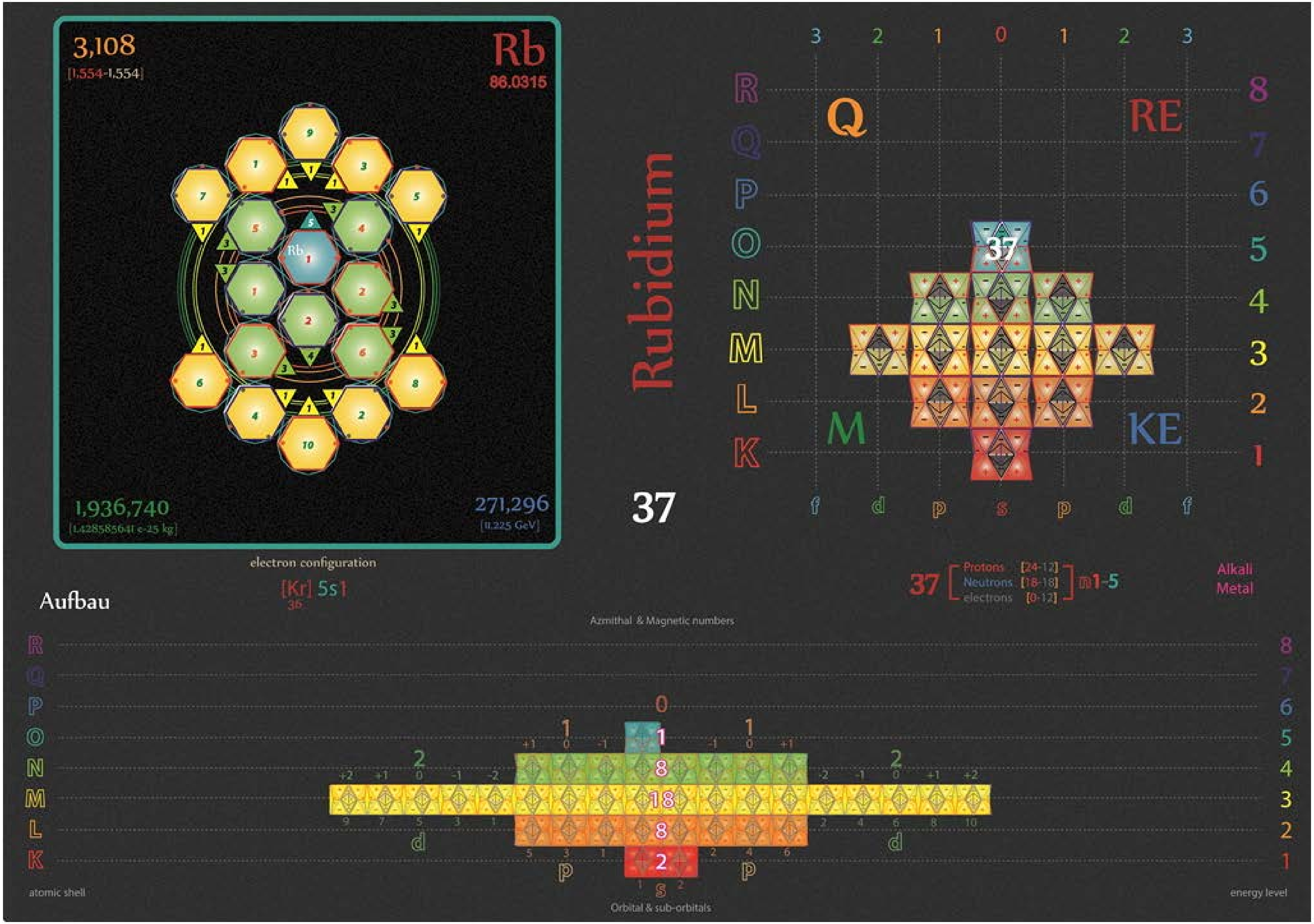
3,108
[1.554-1.554]
Rb
86.0315
1,936,740
[1.428585641 e-25 kg]
271,296
[11.225 GeV]
Rubidium
37
electron configuration
[Kr] 5s1
36
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
37
37 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-5
Alkali Metal
Azmithal & Magnetic numbers
1
8
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.37 - Rubidium atom

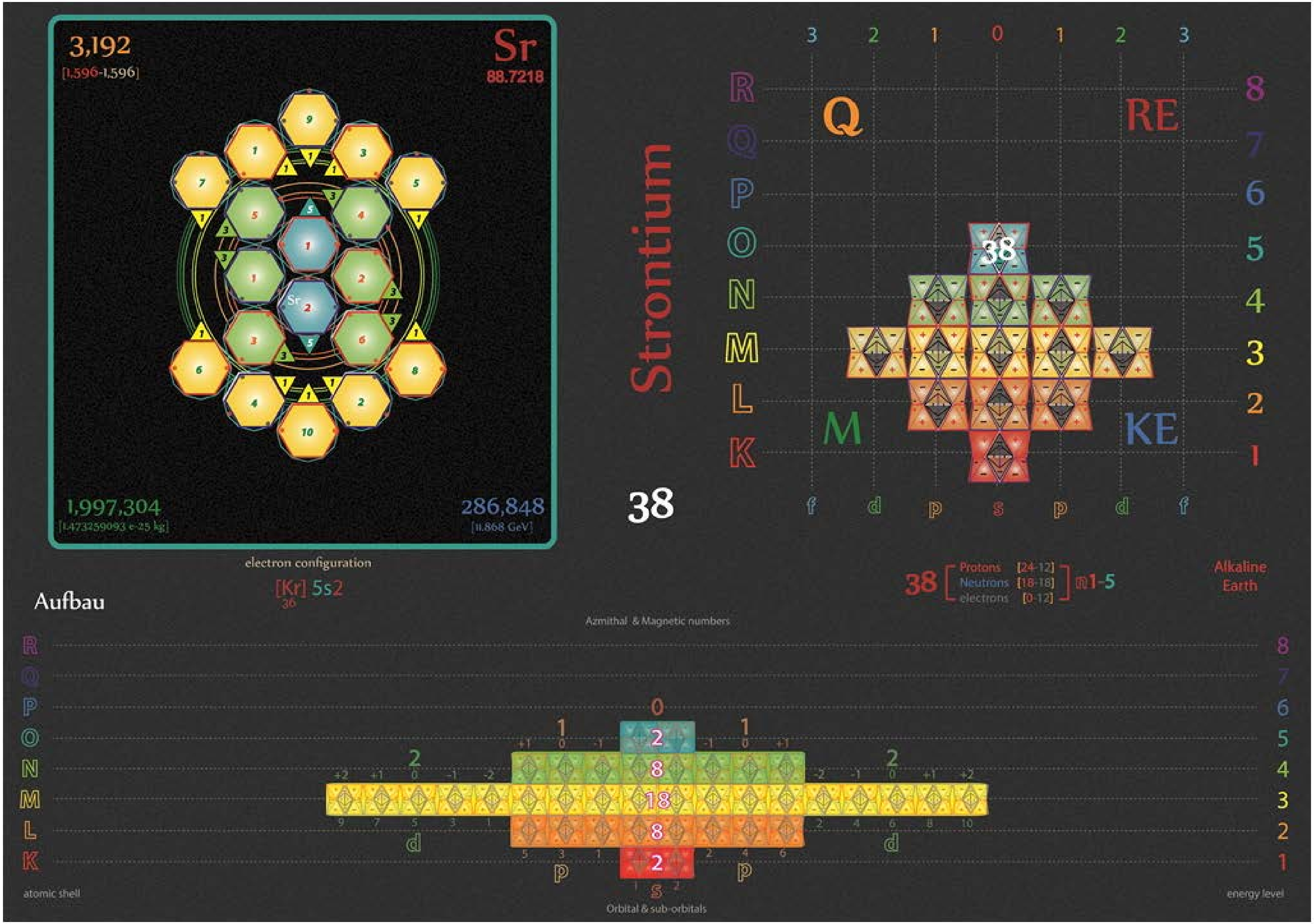
3,192
[1,596-1,596]
Sr
88.7218
1,997,304
[1.473259093 e-25 kg]
286,848
[11.868 GeV]
Strontium
38
electron configuration
[Kr] 5s2
36
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
38 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
2
2
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
2
8
18
8
2
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.38 - Strontium atom

Copyright ABRAHAM [2008] - All rights reserved

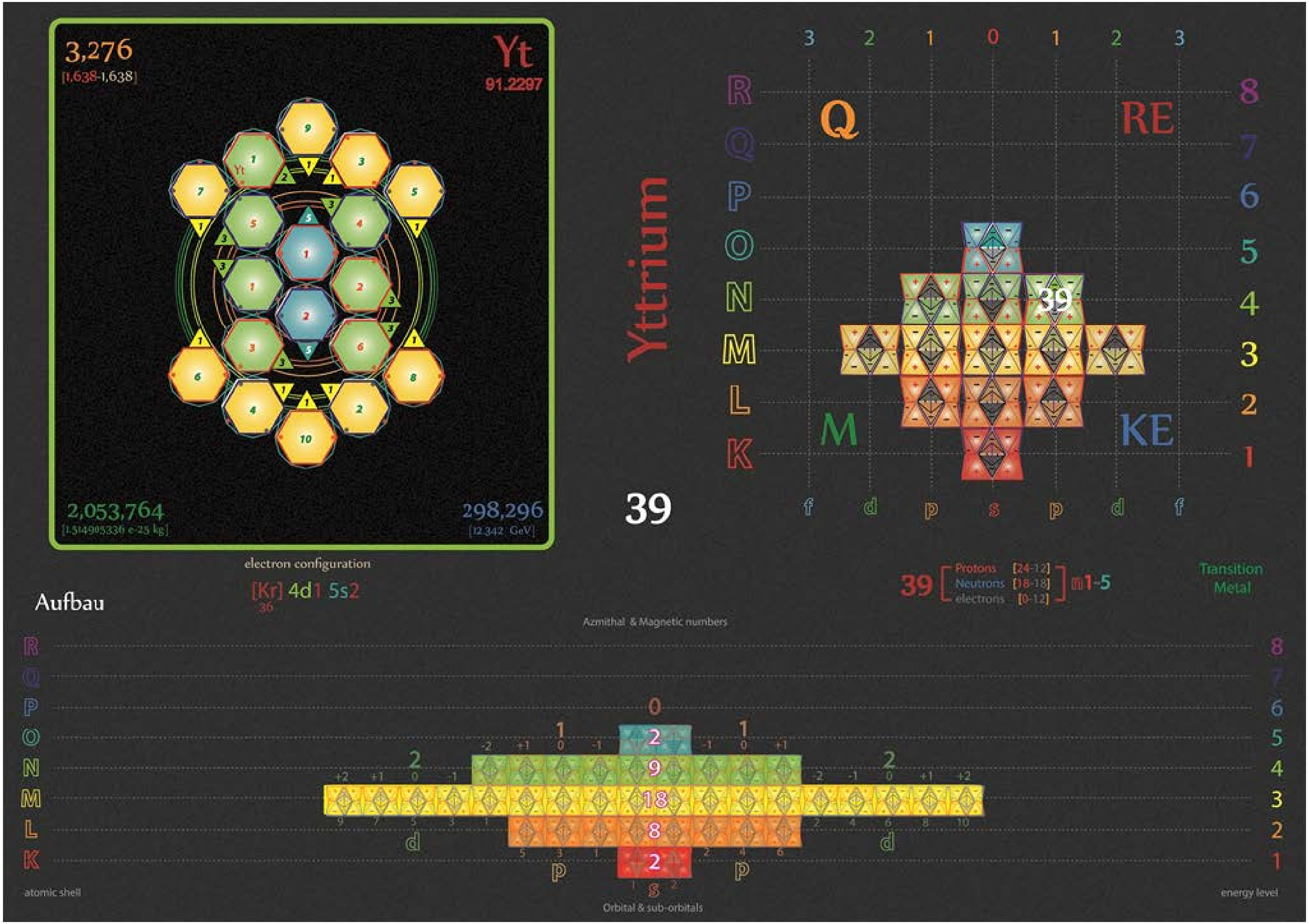
3,276
[1,638-1,638]
Yt
91.2297
2,053,764
[1.514905336 e-25 kg]
298,296
[12.342 GeV]
Yttrium
39
electron configuration
[Kr] 4d1 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
39
f d p s p d f
39 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
9
18
8
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.39 - Yttrium atom

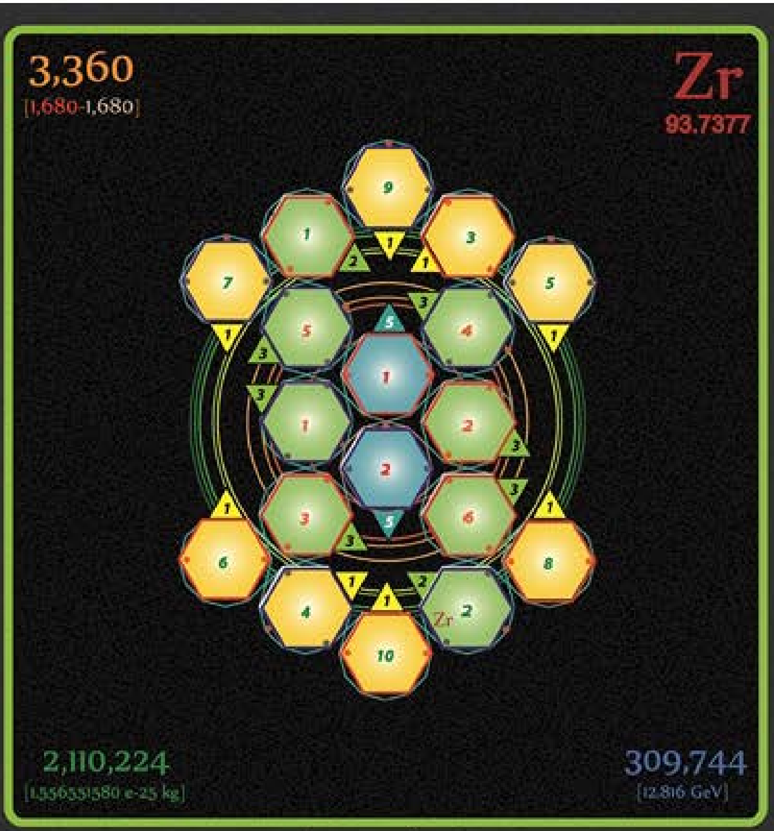
3,360
[1,680-1,680]
Zr
93.7377
9
1
3
7
5
2
1
1
3
5
4
5
1
3
1
2
2
3
3
3
6
5
3
1
6
8
1
1
1
2
4
2
Zr
10
2,110,224
[1.556531580 e-25 kg]
309,744
[12.816 GeV]

Zirconium

40

electron configuration

[Kr] 4d2 5s2
36

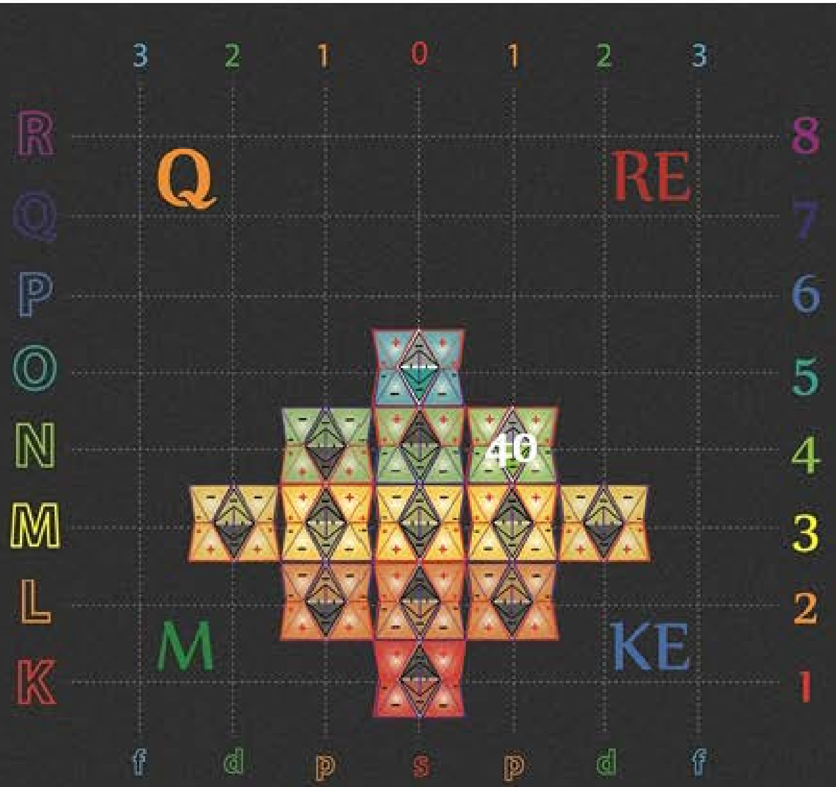
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
40
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

40 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Transition
Metal

Aufbau

Azmithal & Magnetic numbers

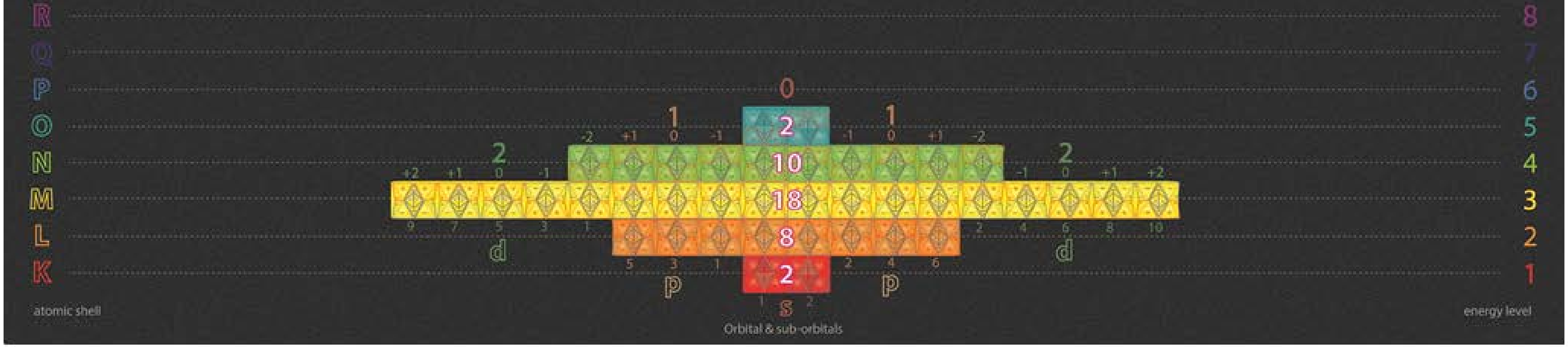
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
-2
+1
0
-1
-1
0
+1
-2
+2
+1
0
-1
-1
0
+1
+2
2
10
18
8
2
9
7
5
3
1
2
4
6
8
10
5
3
1
2
4
6
1
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.40 - Zirconium atom

Copyright ABRAHAM [2008] - All rights reserved

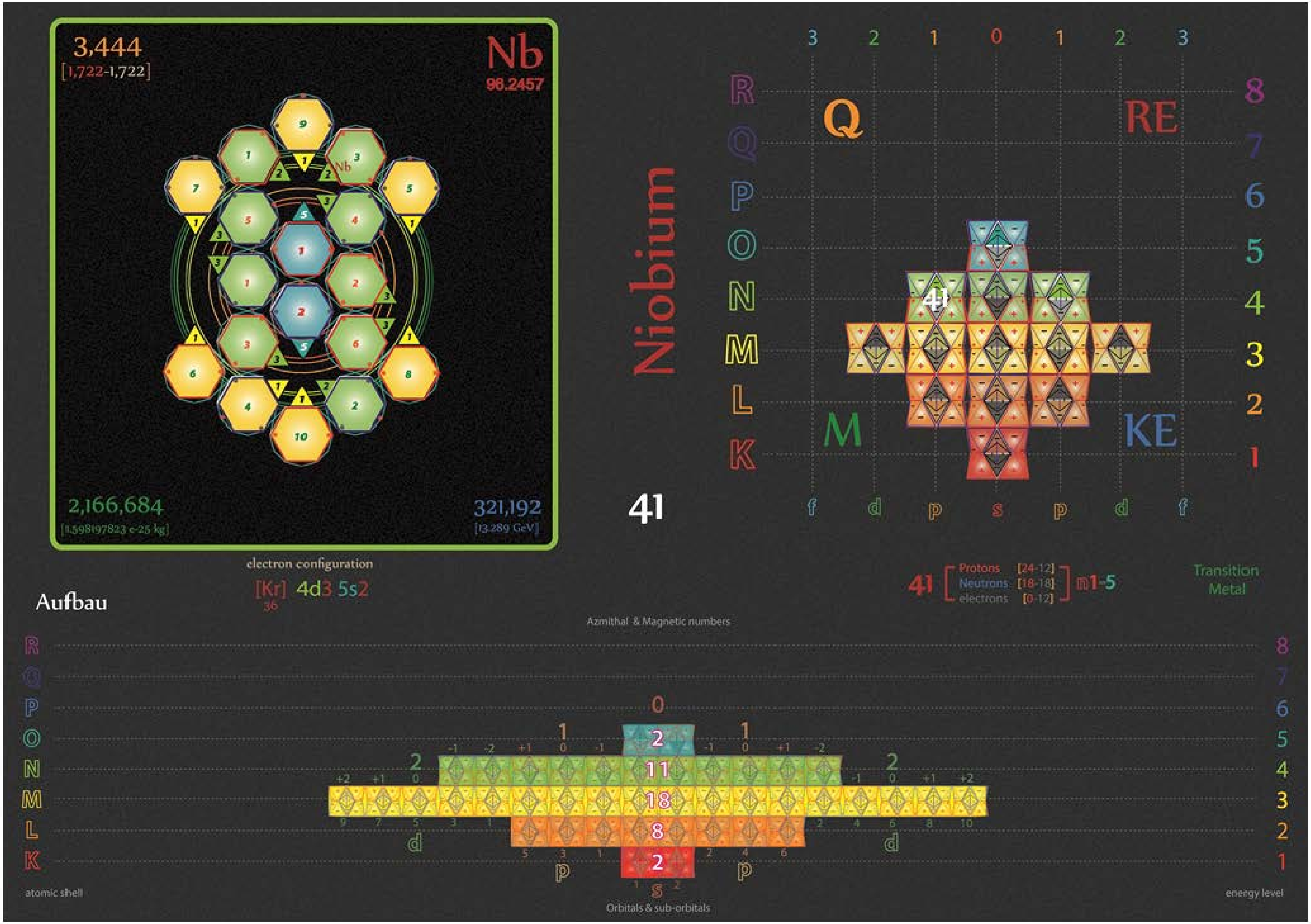
3,444
[1,722-1,722]
Nb
92.9064
2,166,684
[1.598197823 e-25 kg]
321,192
[13.289 GeV]
Niobium
41
electron configuration
[Kr] 4d3 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
41 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
2
11
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.41 - Niobium atom

Copyright ABRAHAM [2008] - All rights reserved

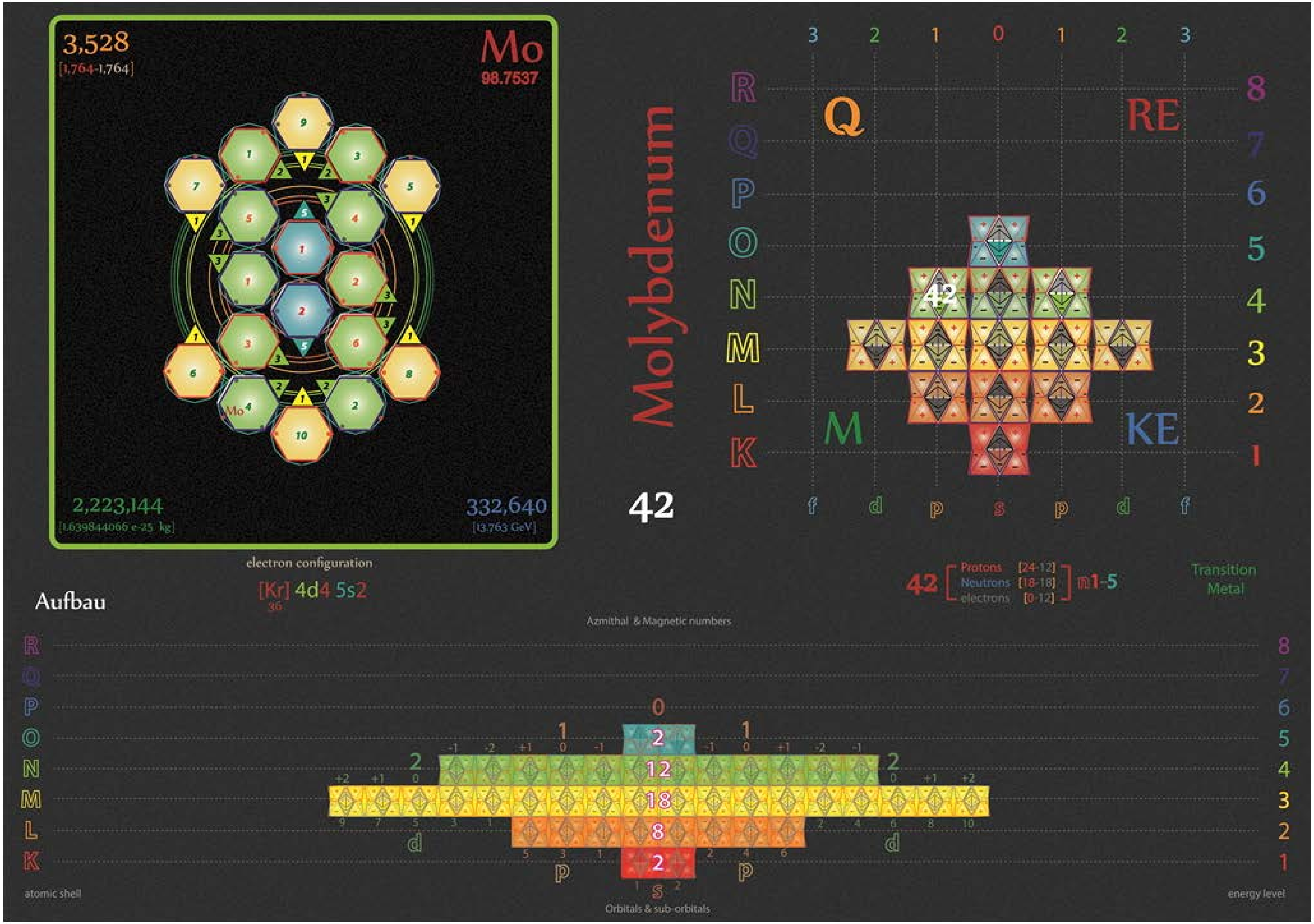
3,528
[1,764-1,764]
Mo
98.7537
2,223,144
[1.639844066 e-25 kg]
332,640
[13.763 GeV]
Molybdenum
42
electron configuration
[Kr] 4d4 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
42 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
2
2
12
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.42 - Molybdenum atom

Copyright ABRAHAM [2008] - All rights reserved

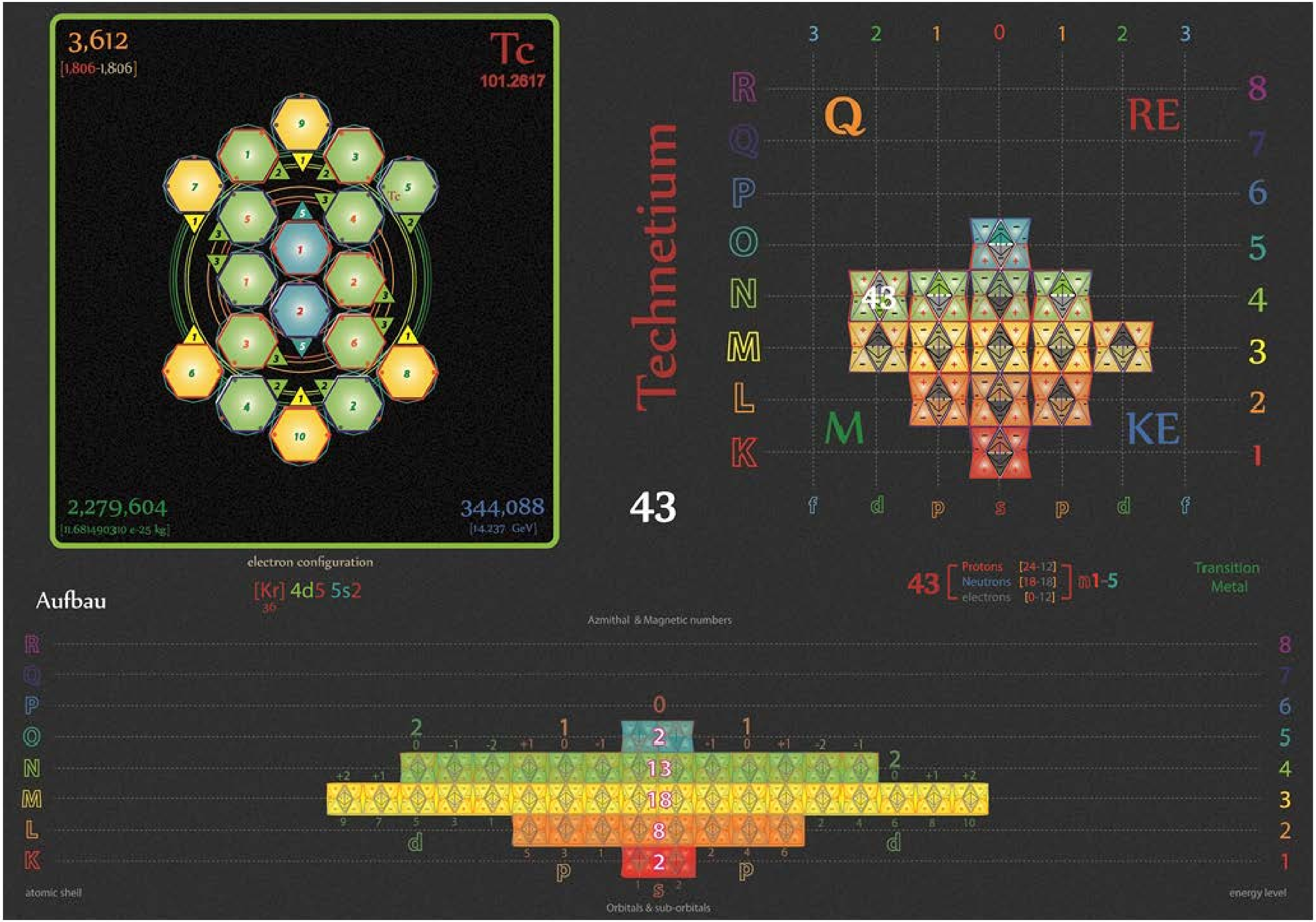
3,612
[1,806-1,806]
Tc
101.2617
2,279,604
[11.6814903110 e-25 kg]
344,088
[14.237 GeV]
Technetium
43
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d5 5s2
36
Aufbau
43 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
2
13
18
8
2
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.43 - Technetium atom

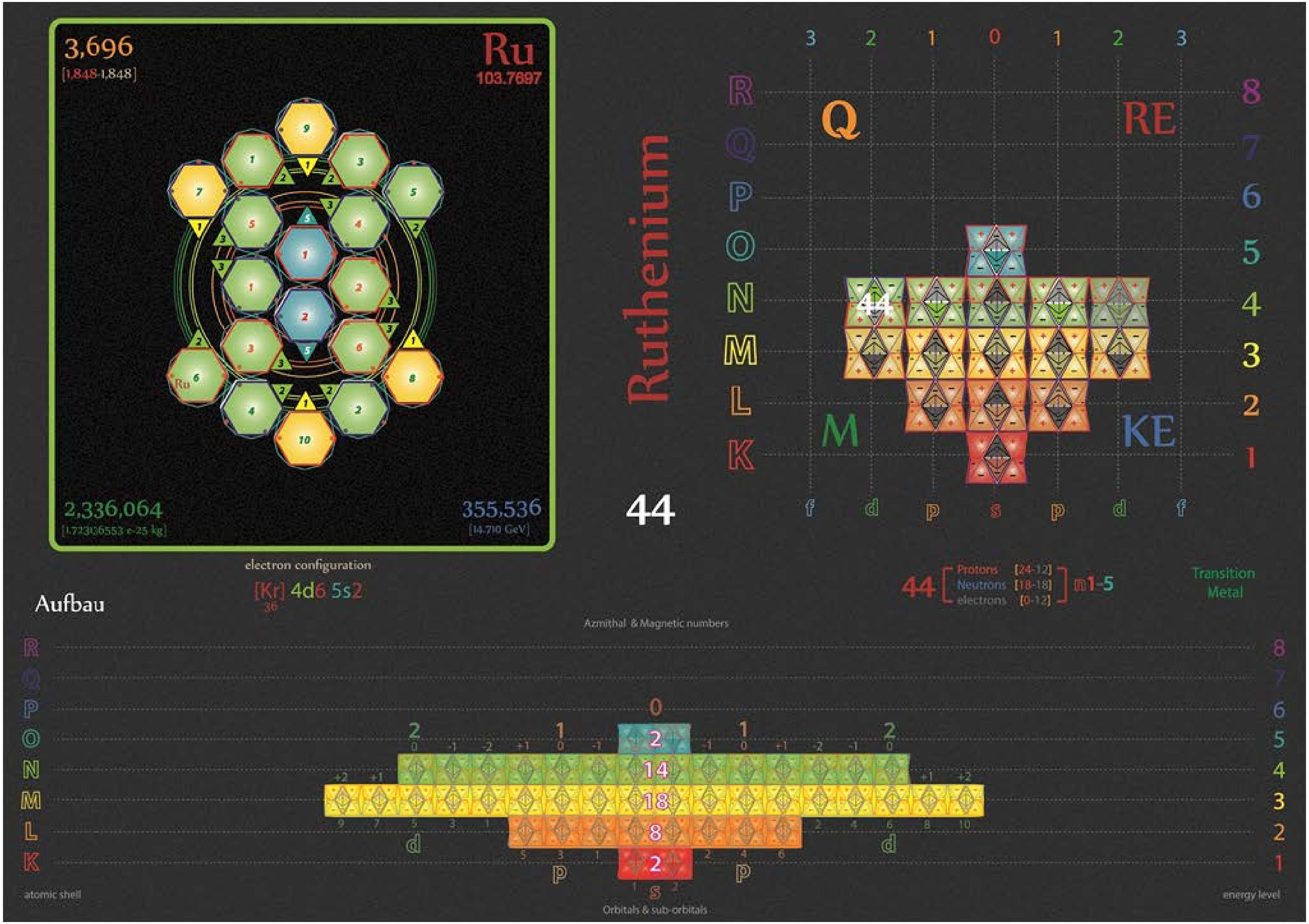
3,696
[1,848-1,848]
Ru
103.7697
2,336,064
[1.72316553 e-25 kg]
355,536
[14.710 GeV]
Ruthenium
44
electron configuration
[Kr] 4d6 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
44
f d p s p d f
44 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2
14
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.44 - Ruthenium atom

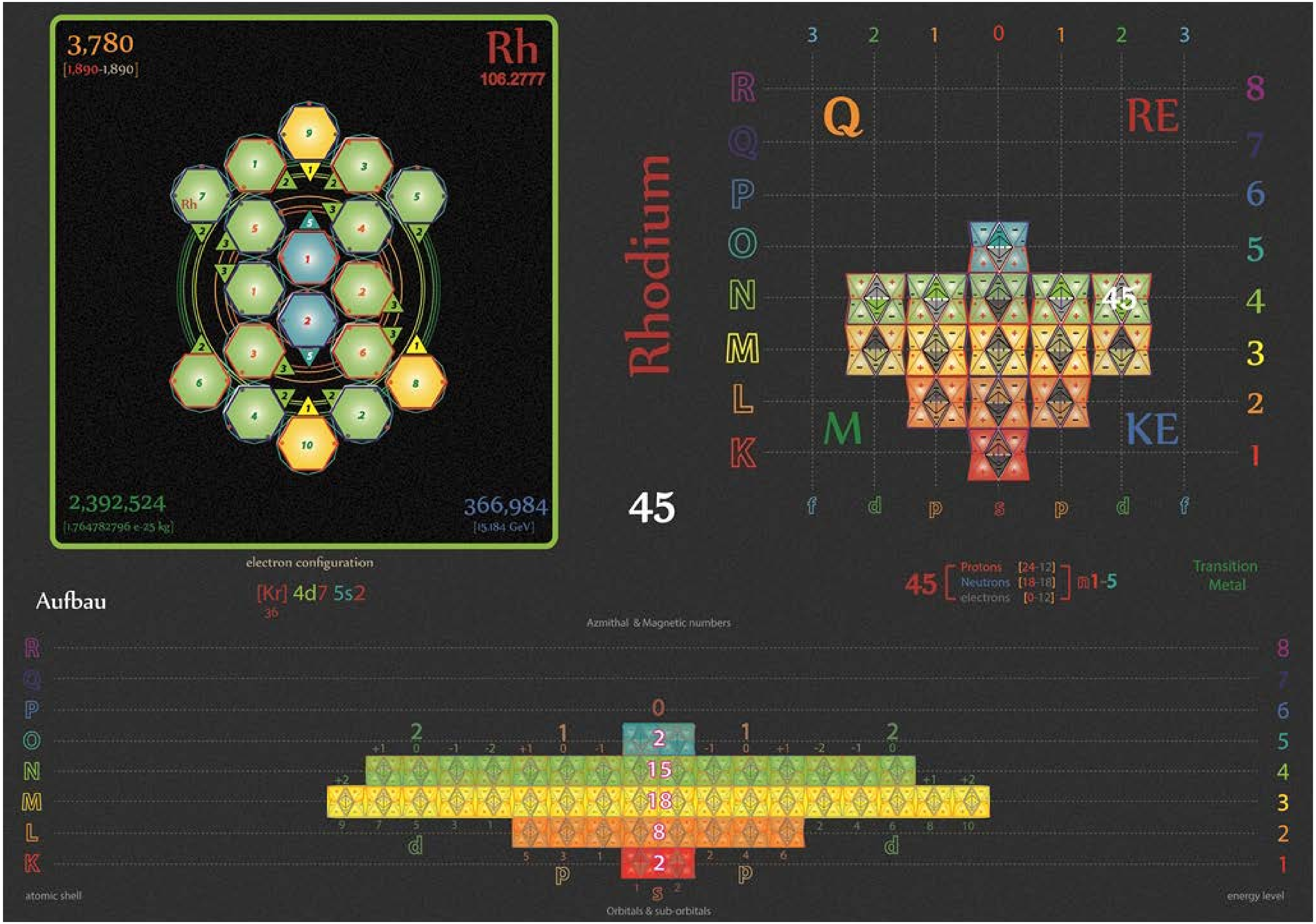
3,780
[1,890-1,890]
Rh
106.2777
Rh
2,392,524
[1.764782796 e-25 kg]
366,984
[15.184 GeV]
Rhodium
45
electron configuration
[Kr] 4d7 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
45
f d p s p d f
45 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
15
18
8
2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
+2 +1 +2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.45 - Rhodium atom

Copyright ABRAHAM [2008] - All rights reserved

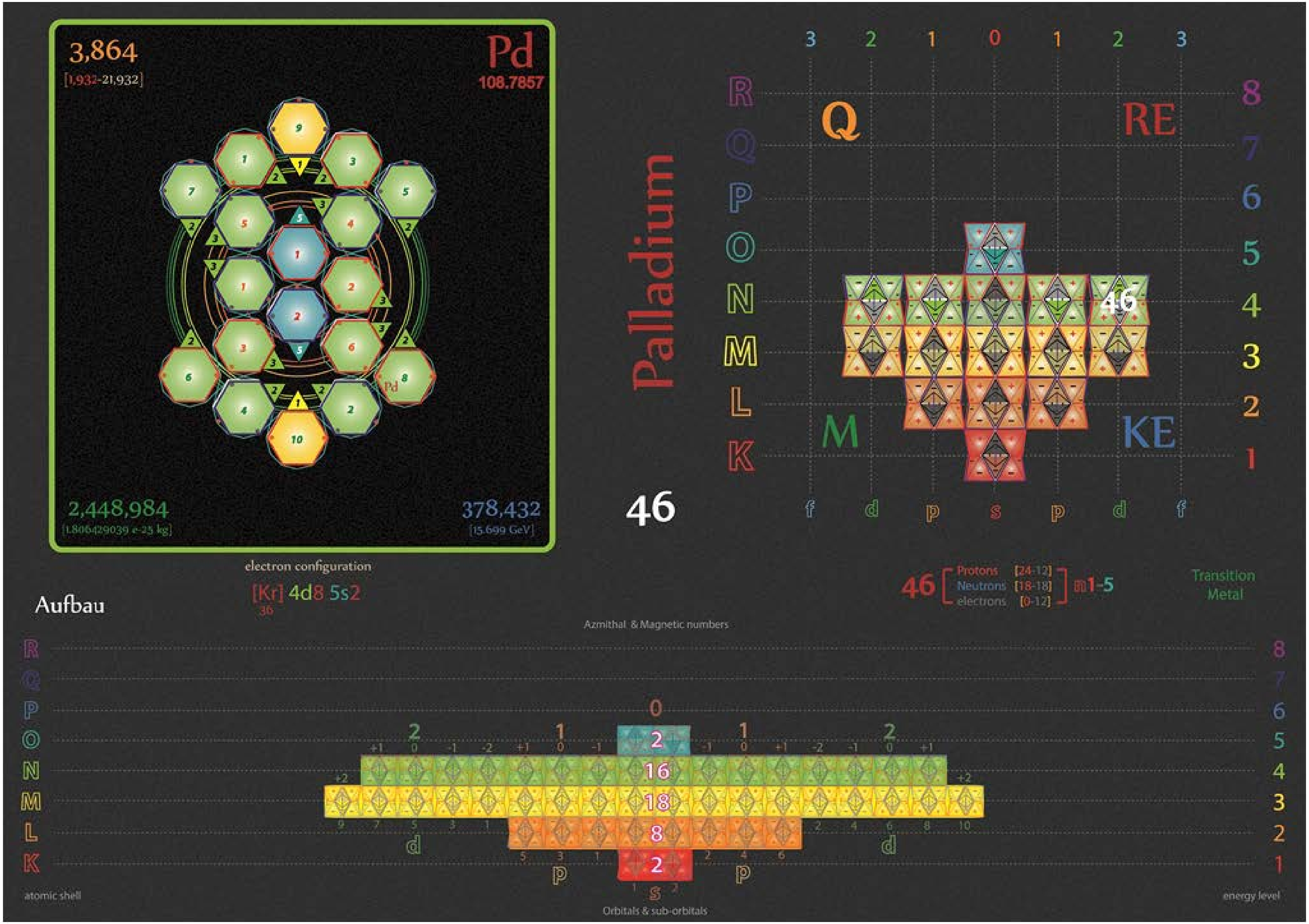
3,864
[1,932-21,932]
Pd
108.7857
2,448,984
[1.806429039 e-25 kg]
378,432
[15.699 GeV]
Palladium
46
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
46
f d p s p d f
electron configuration
[Kr] 4d8 5s2
36
Aufbau
46 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
16
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.46 - Palladium atom

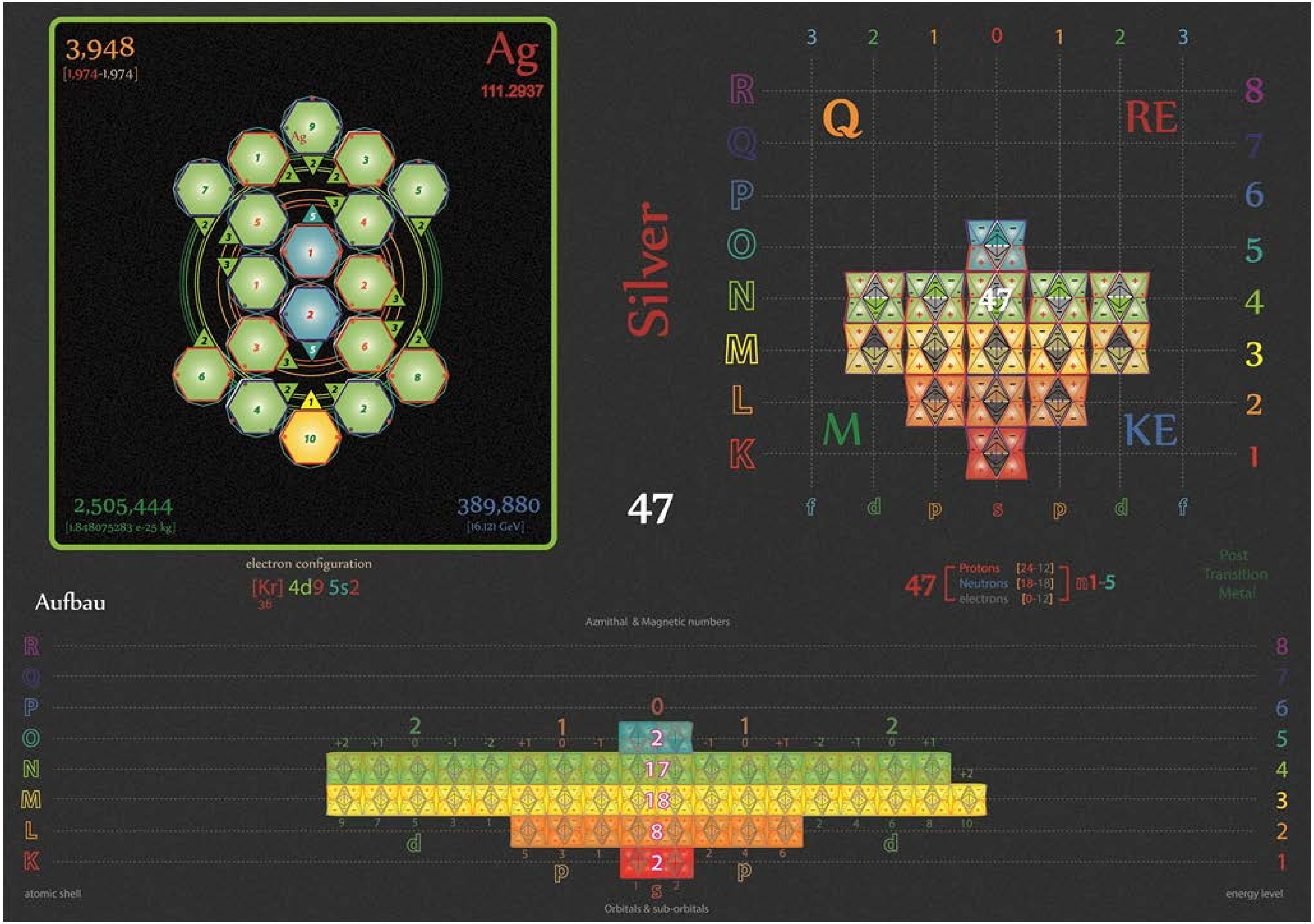
3,948
[1,974-1,974]
Ag
111.2937
2,505,444
[1.848075283 e-25 kg]
389,880
[16.121 GeV]
Silver
47
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
47 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Post Transition Metal
electron configuration
[Kr] 4d9 5s2
36
Aufbau
Azmithal & Magnetic numbers
0
2 1 1 2
2
17
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.47 - Silver atom

Copyright ABRAHAM [2008] - All rights reserved

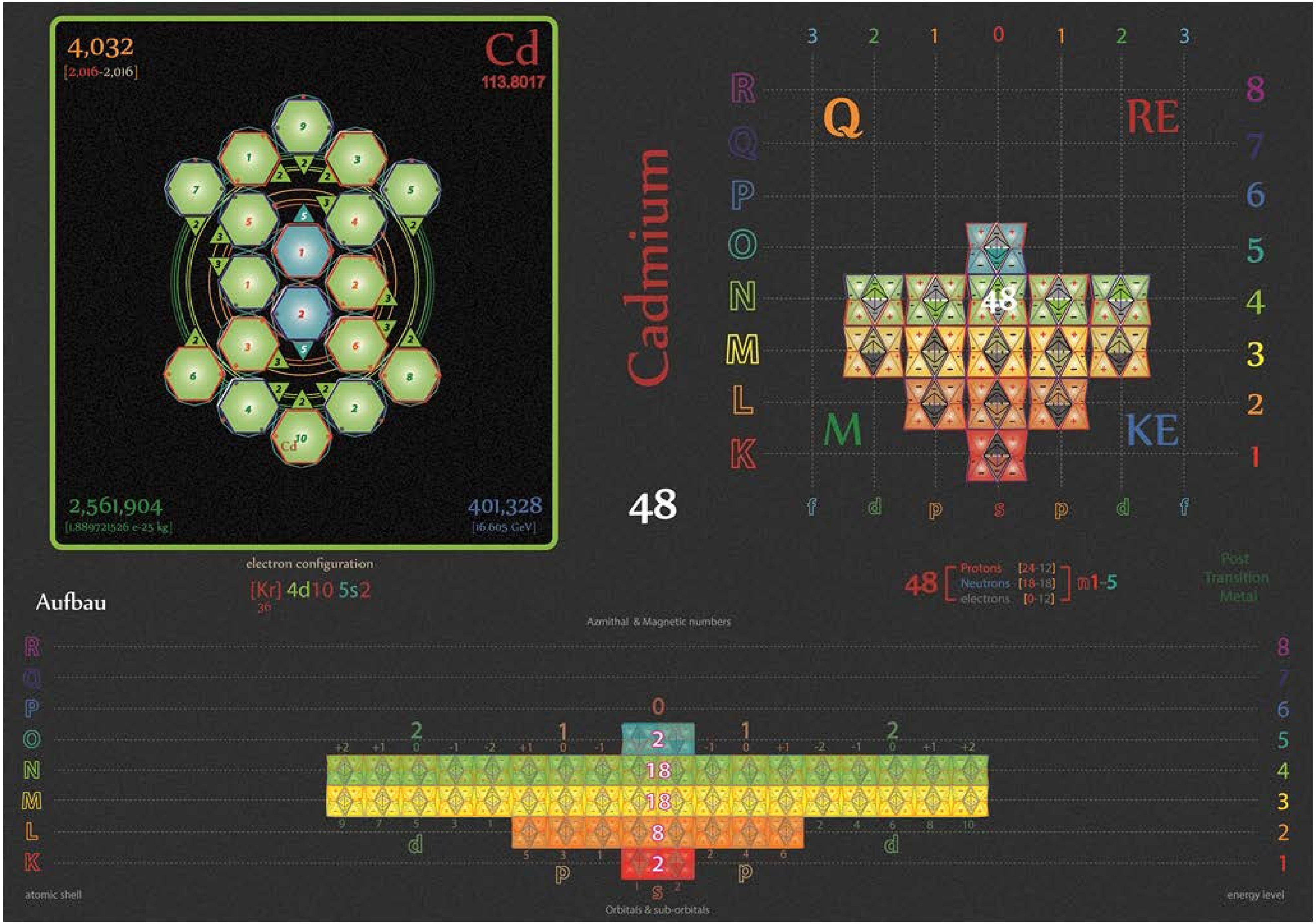
4,032
[2,016-2,016]
Cd
113.8017
2,561,904
[1.889721526 e-25 kg]
401,328
[16.605 GeV]
Cadmium
48
electron configuration
[Kr] 4d10 5s2
36
Aufbau
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
48
48 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 2
s
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.48 - Cadmium atom

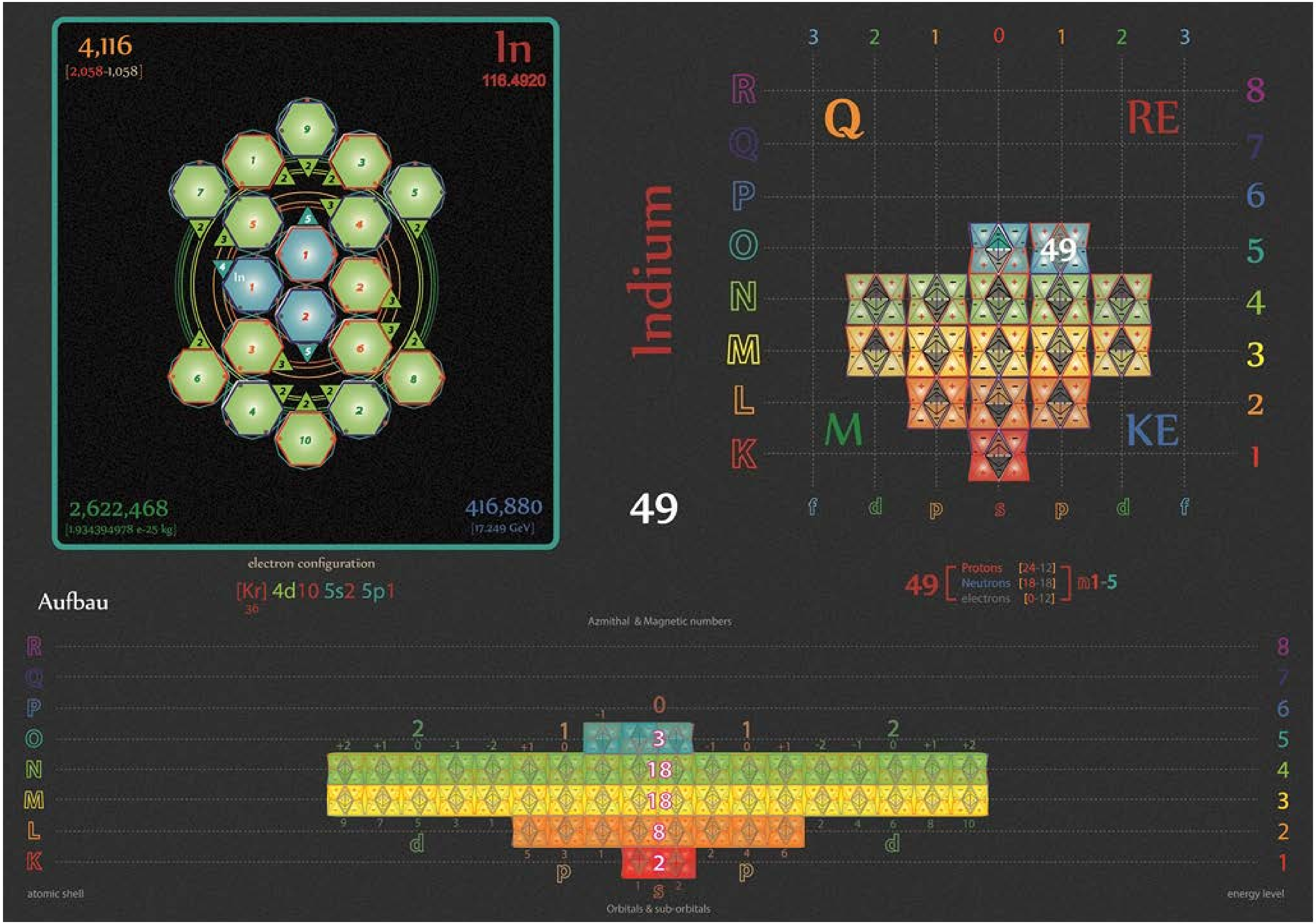
4,116
[2,058-1,058]
In
116.4920
2,622,468
[1.934394978 e-25 kg]
416,880
[17.249 GeV]
Indium
49
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
49
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p1
36
Aufbau
49 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Azmithal & Magnetic numbers
0
2 1 1 2
-1
+2 +1 0 -1 -2 +1 0 -1 0 +1 -2 -1 0 +1 +2
3
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.49 - Indium atom

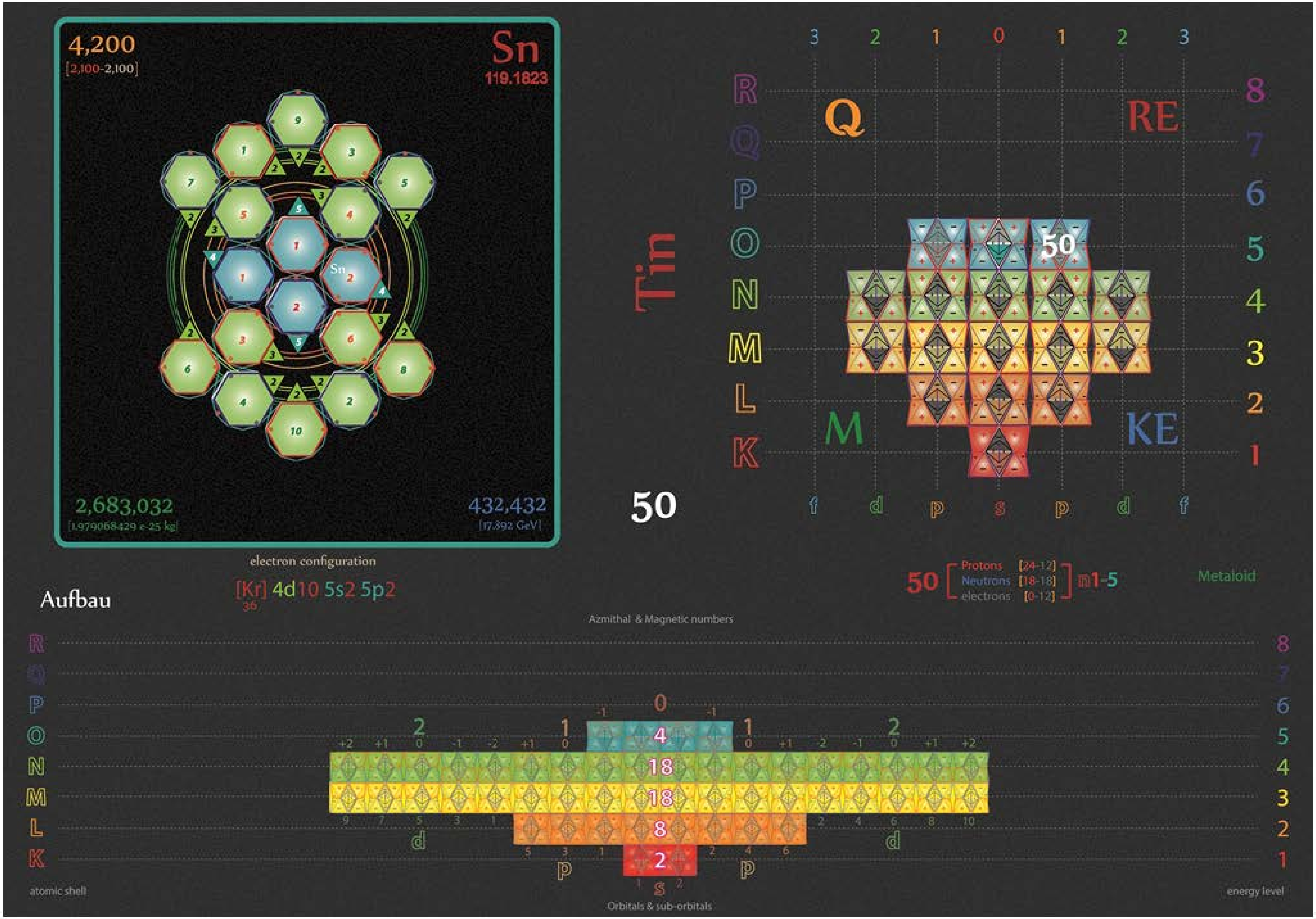
4,200
[2,100-2,100]
Sn
119.1823
2,683,032
[1.979068429 e-25 kg]
432,432
[17.392 GeV]
electron configuration
[Kr] 4d10 5s2 5p2
36
Aufbau
Tin
50
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
50 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
0
-1
-1
2
1
1
2
+2 +1 0 -1 -2 +1 0 0 +1 -2 -1 0 +1 +2
4
18
18
8
2
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.50 - Tin atom

Copyright ABRAHAM [2008] - All rights reserved

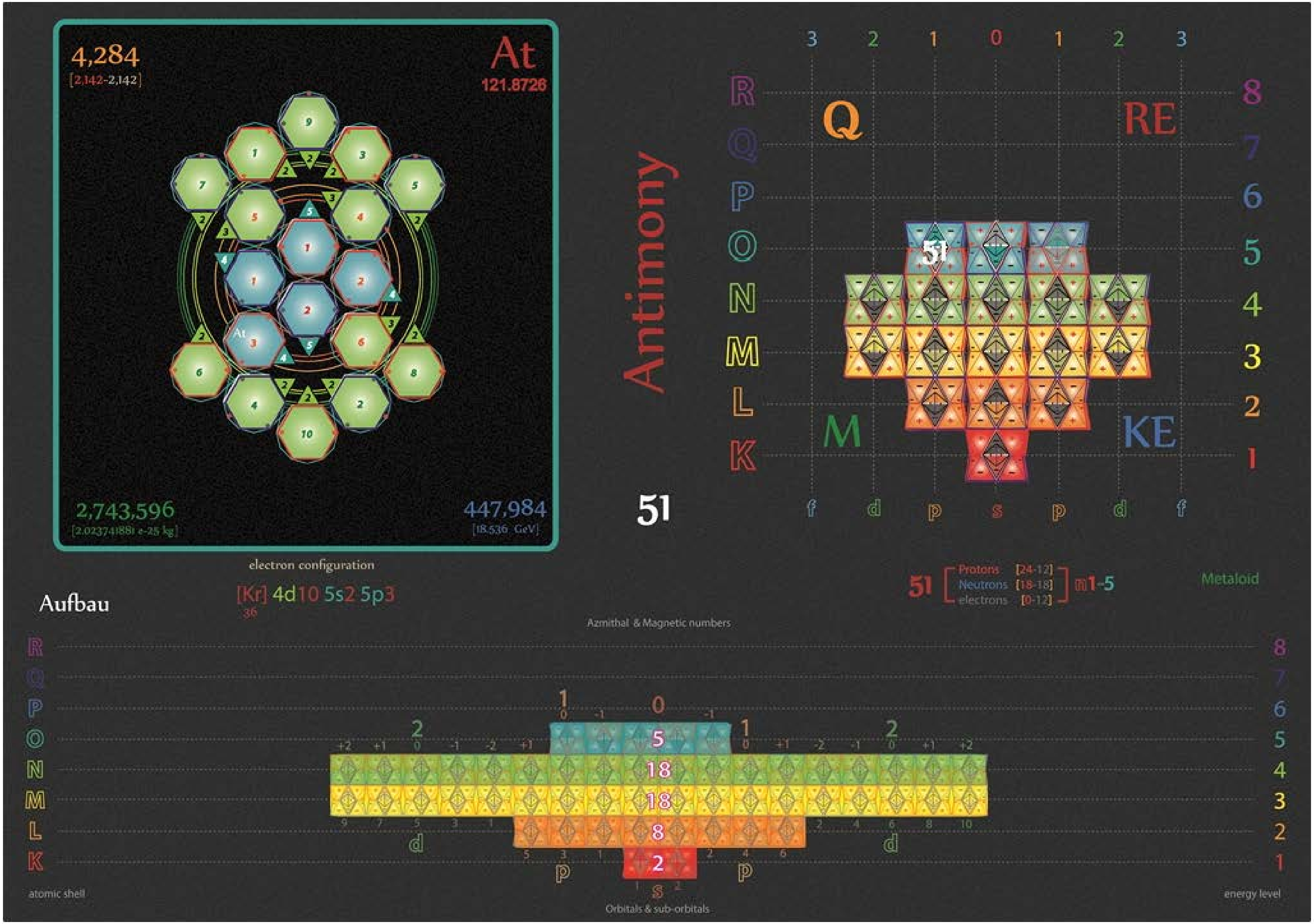
4,284
[2,142-2,142]
At
121.8726
2,743,596
[2.02374188i e-25 kg]
447,984
[18.536 GeV]
Antimony
51
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
51
electron configuration
[Kr] 4d10 5s2 5p3
36
Aufbau
51
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Metaloid
Azmithal & Magnetic numbers
5
18
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.51 - Antimony atom

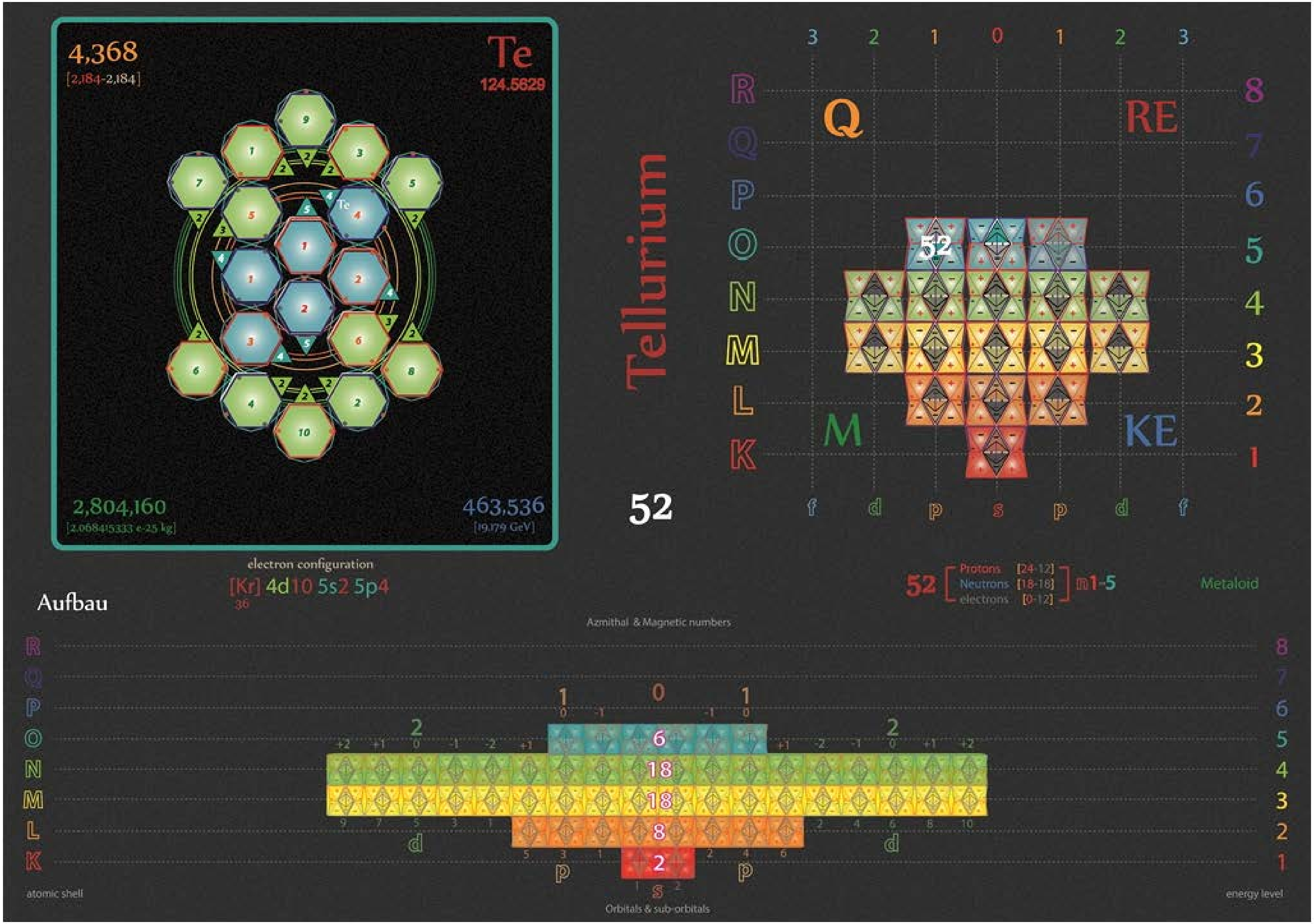
4,368
[2,184-2,184]
Te
124.5629
2,804,160
[2.068415333 e-25 kg]
463,536
[19.179 GeV]
Tellurium
52
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p4
36
Aufbau
52 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
6
18
18
8
2
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.52 - Tellurium atom

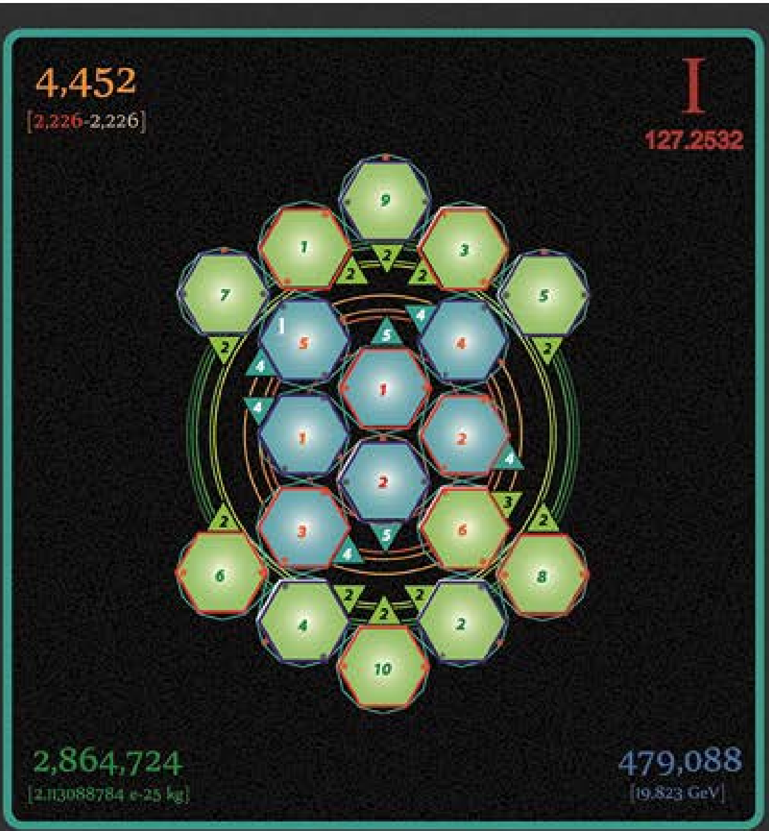
4,452
[2,226-2,226]
I
127.2532
9
1
3
7
5
1
5
4
1
1
2
3
6
2
6
8
4
2
10
2,864,724
[2.113088784 e-25 kg]
479,088
[19.823 GeV]

Iodine

53

electron configuration

[Kr] 4d10 5s2 5p5
36

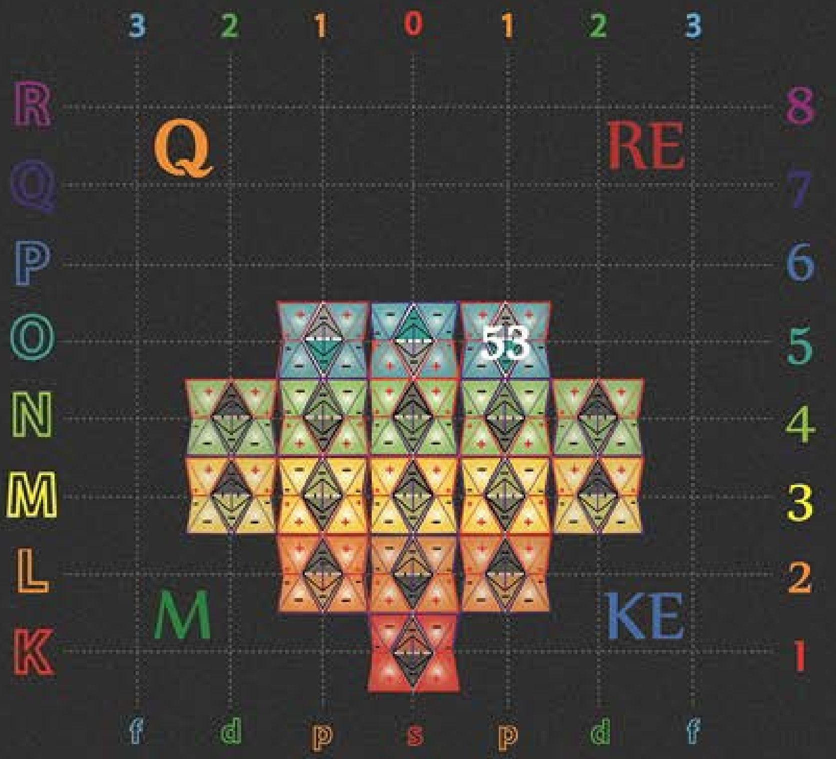
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
53
f
d
p
s
p
d
f

53 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Halogen

Aufbau

Azmithal & Magnetic numbers

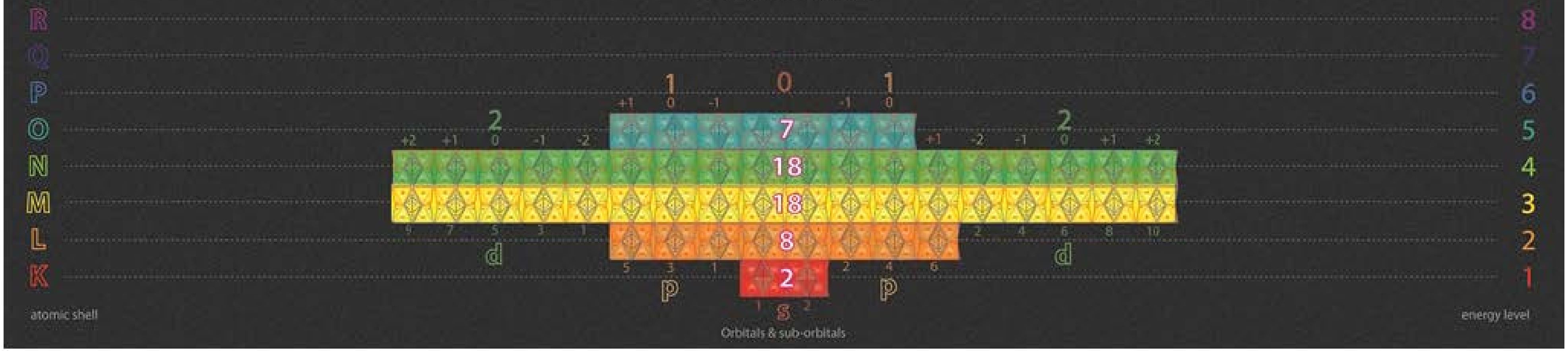
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
0
1
+1
0
-1
-1
0
2
2
+2
+1
0
-1
-2
+1
-2
-1
0
+1
+2
7
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.53 - Iodine atom

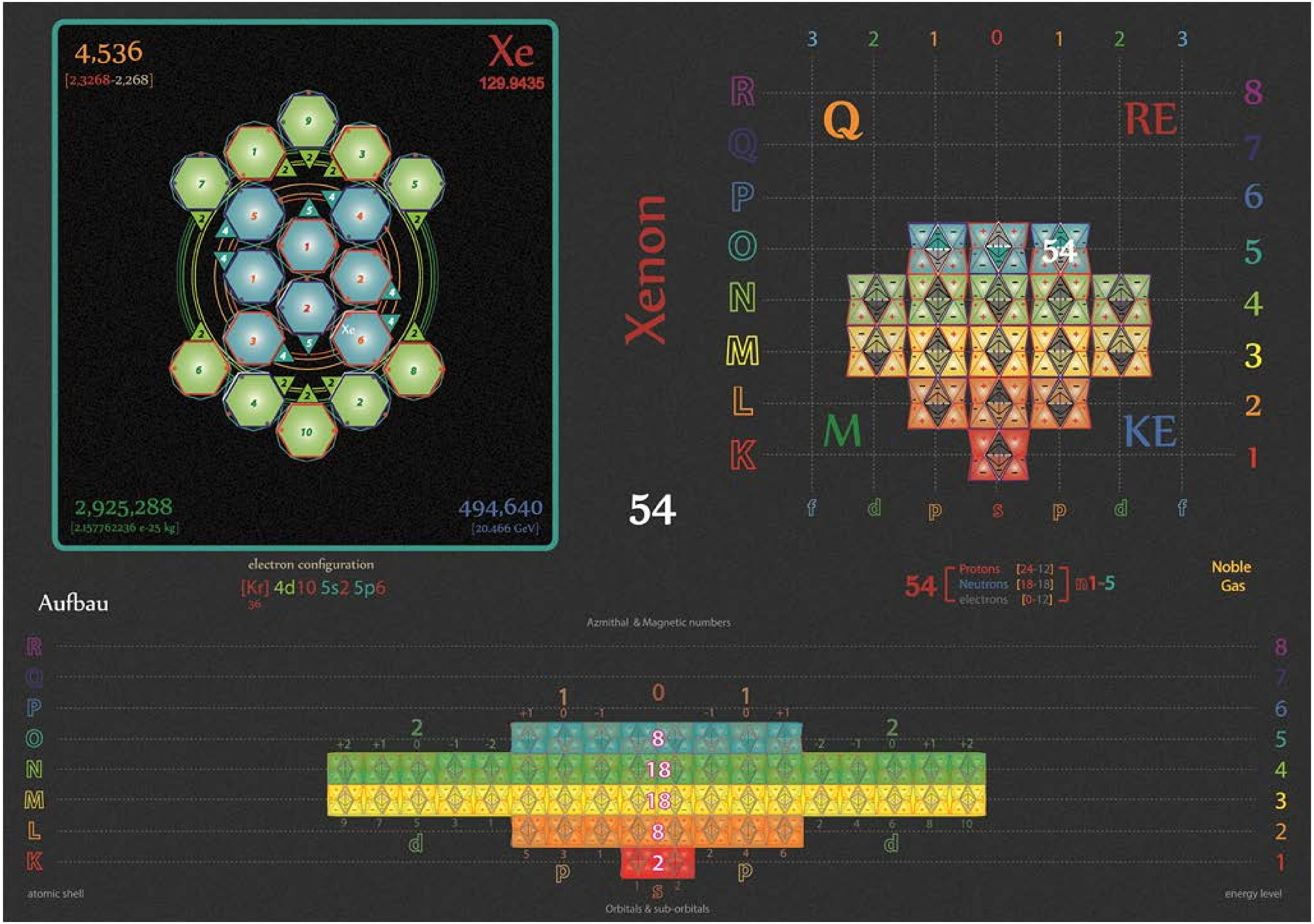
4,536
[2,3268-2,268]
Xe
129.9435
2,925,288
[2.157762236 e-25 kg]
494,640
[20.466 GeV]
Xenon
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p6
36
Aufbau
54 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Noble Gas
Azmithal & Magnetic numbers
1 0 1
2 2
8
18
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.54 - Xenon atom

Copyright ABRAHAM [2008] - All rights reserved

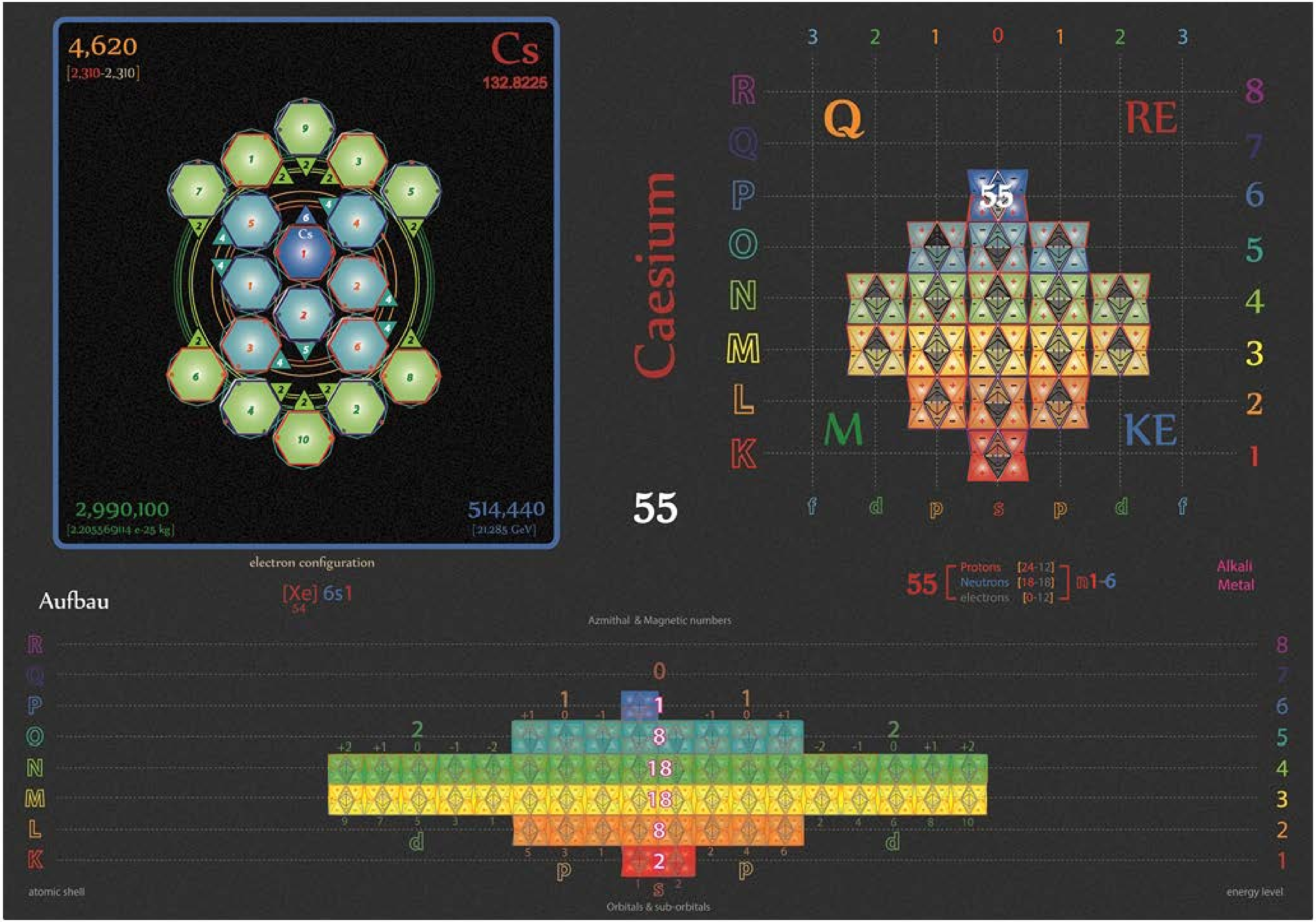
4,620
[2,310-2,310]
Cs
132.8225
Cs
2,990,100
[2.205569114 e-25 kg]
514,440
[21.285 GeV]
electron configuration
[Xe] 6s1
54
Aufbau
Caesium
55
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
55
f d p s p d f
55
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Alkali
Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
1
8
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.55 - Caesium atom

Copyright ABRAHAM [2008] - All rights reserved

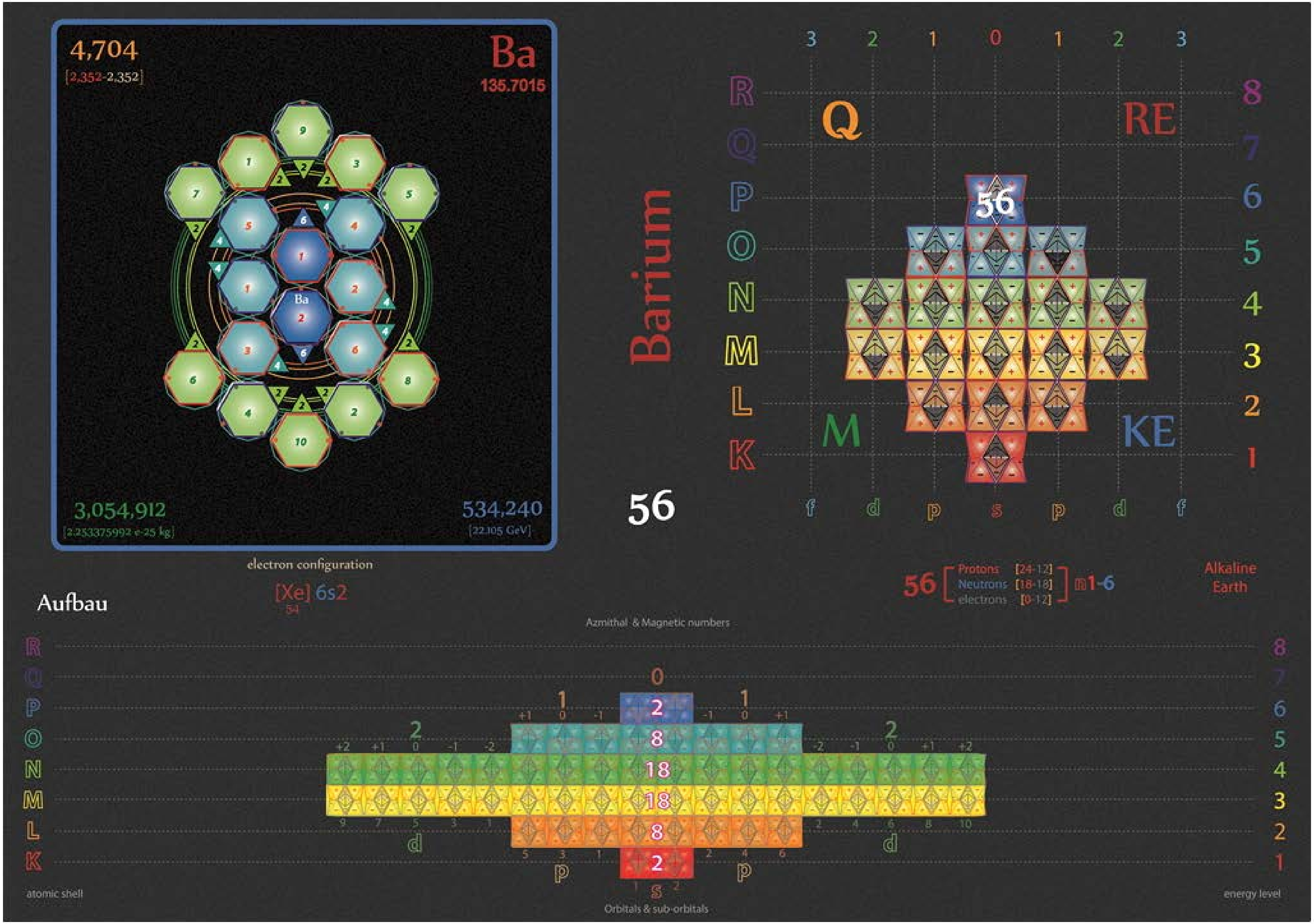
4,704
[2,352-2,352]
Ba
135.7015
3,054,912
[2.253375992 e-25 kg]
534,240
[22.105 GeV]
electron configuration
[Xe] 6s2
54
Aufbau
Barium
56
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
56
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Alkaline Earth
Azmithal & Magnetic numbers
0
1
+1 0 -1
2
+2 +1 0 -1 -2
2
8
18
18
8
2
-1 0 +1
-2 -1 0 +1 +2
9 7 5 3 1
2 4 6 8 10
d
5 3 1
2 4 6
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.56 - Barium atom

Copyright ABRAHAM [2008] - All rights reserved

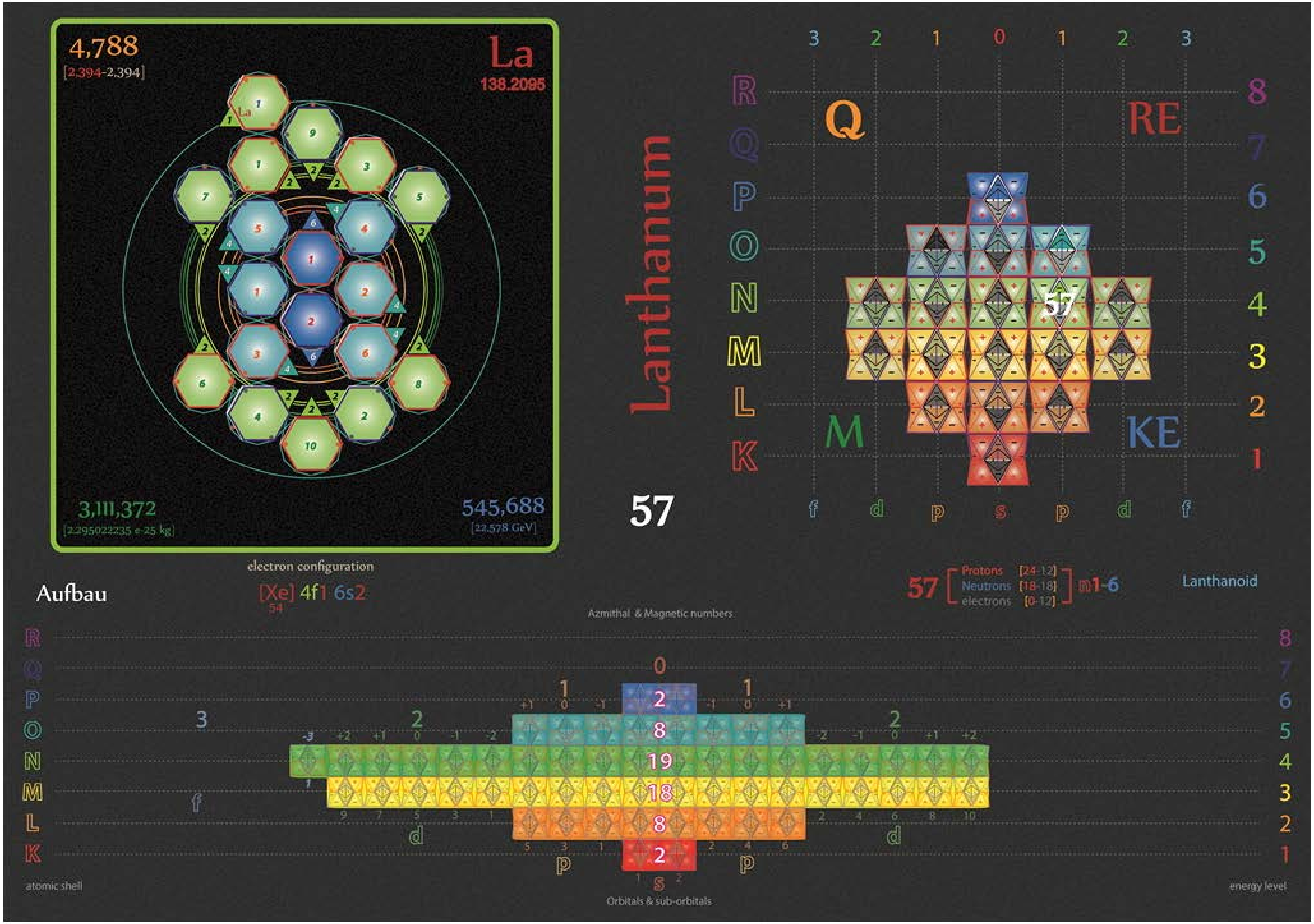
4,788
[2,394-2,394]
La
138.2095
3,111,372
[2.295022235 e-25 kg]
545,688
[22.578 GeV]
Lanthanum
57
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
57
f d p s p d f
Aufbau
electron configuration
[Xe] 4f1 6s2
54
57 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals
2
8
19
18
8
2

Tetryonics 51.57 - Lanthanum atom

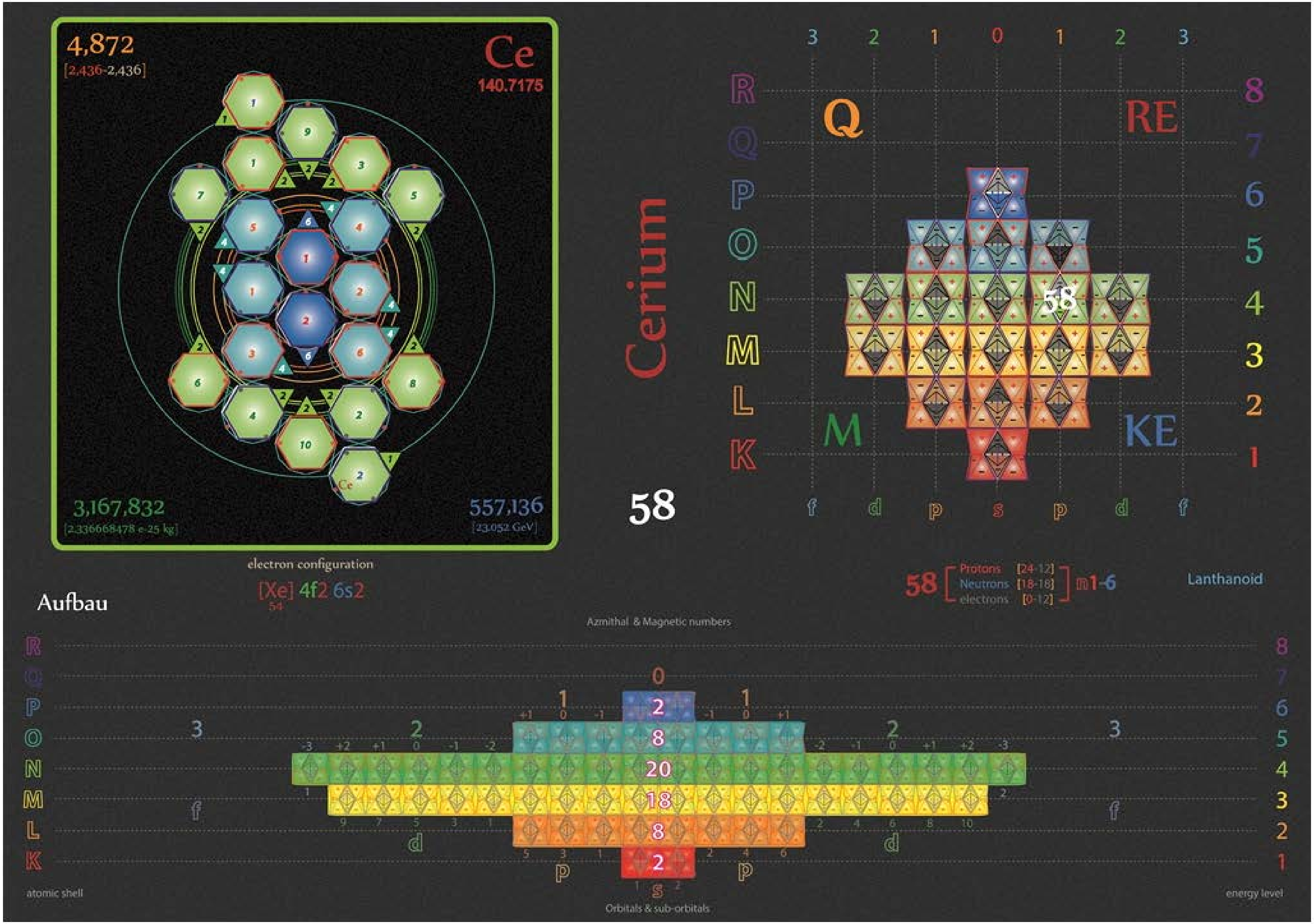
4,872
[2,436-2,436]
Ce
140.7175
3,167,832
[2.336668478 e-25 kg]
557,136
[23,052 GeV]
Cerium
58
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
58 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
electron configuration
[Xe] 4f2 6s2
54
Aufbau
Azmithal & Magnetic numbers
0
2
8
20
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.58 - Cerium atom

Copyright ABRAHAM [2008] - All rights reserved

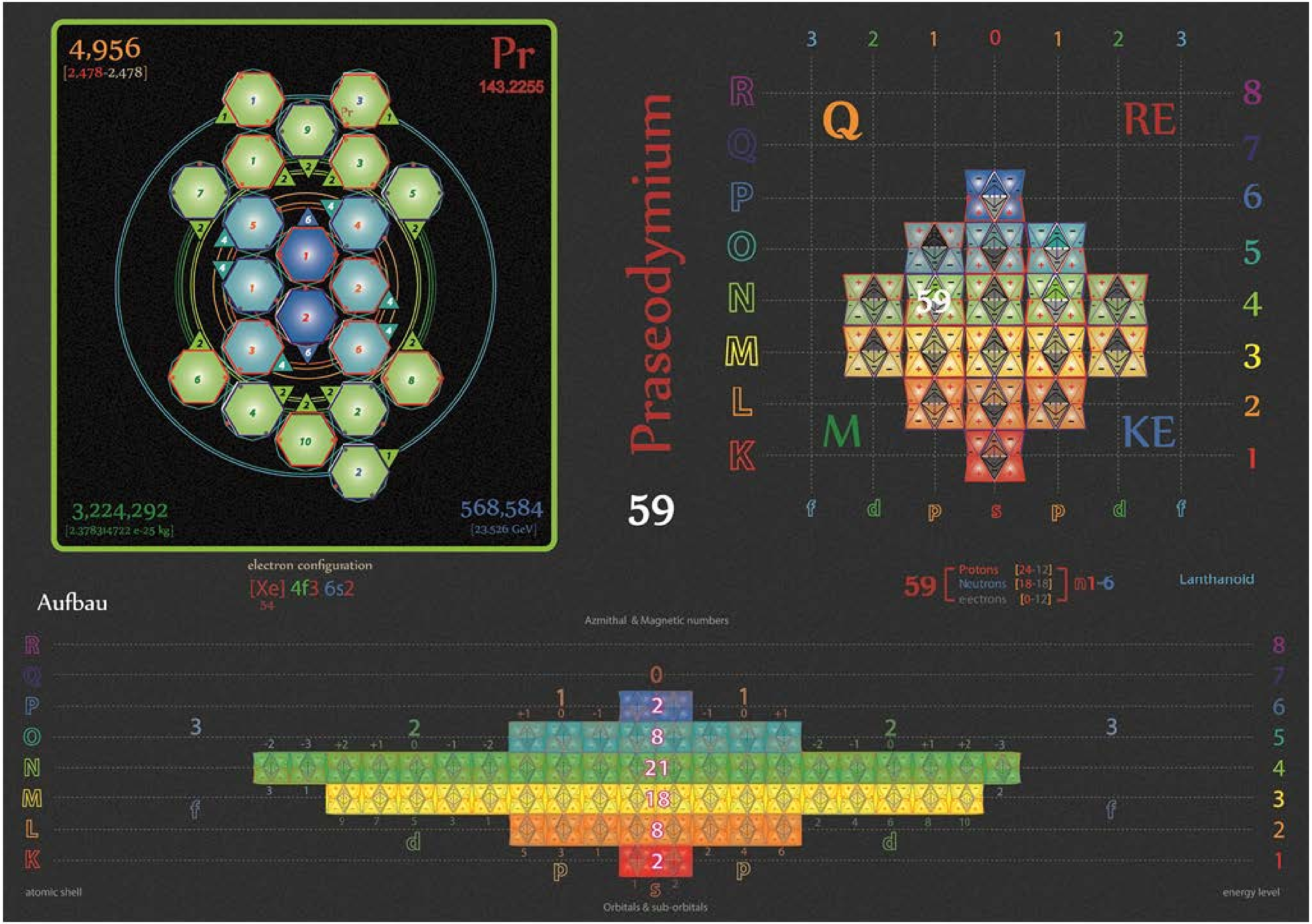
4,956
[2,478-2,478]
Pr
143.2255
3,224,292
[2.378314722 e-25 kg]
568,584
[23.526 GeV]
Praseodymium
59
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
59
electron configuration
[Xe] 4f3 6s2
54
Aufbau
59 [Potons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
8
21
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals
s p d f

Tetryonics 51.59 - Praseodymium atom

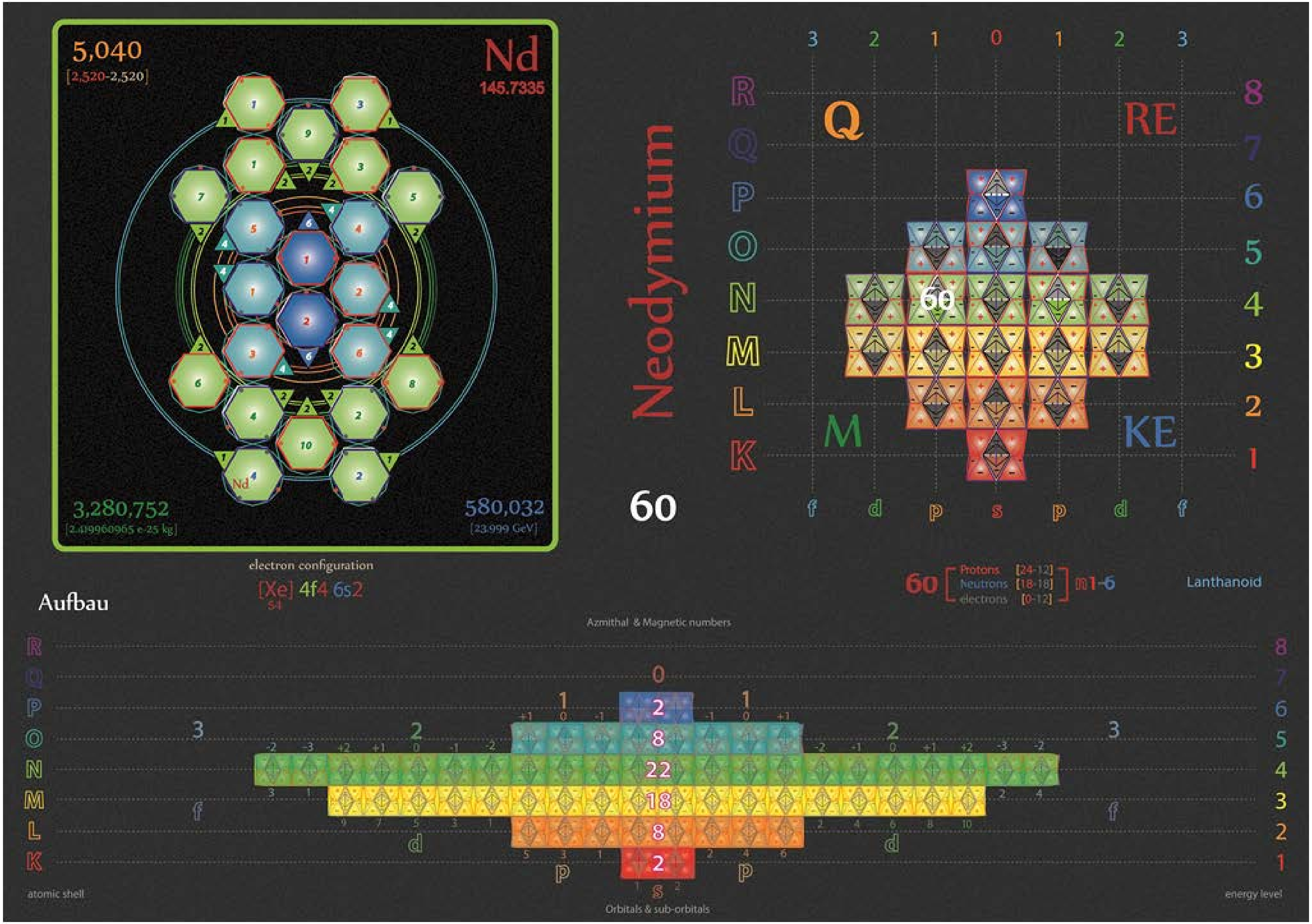
5,040
[2,520-2,520]
Nd
145.7335
3,280,752
[2.49960965 e-25 kg]
580,032
[23.999 GeV]
Neodymium
60
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f4 6s2
54
Aufbau
60 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
22
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.60 - Neodymium atom

Copyright ABRAHAM [2008] - All rights reserved

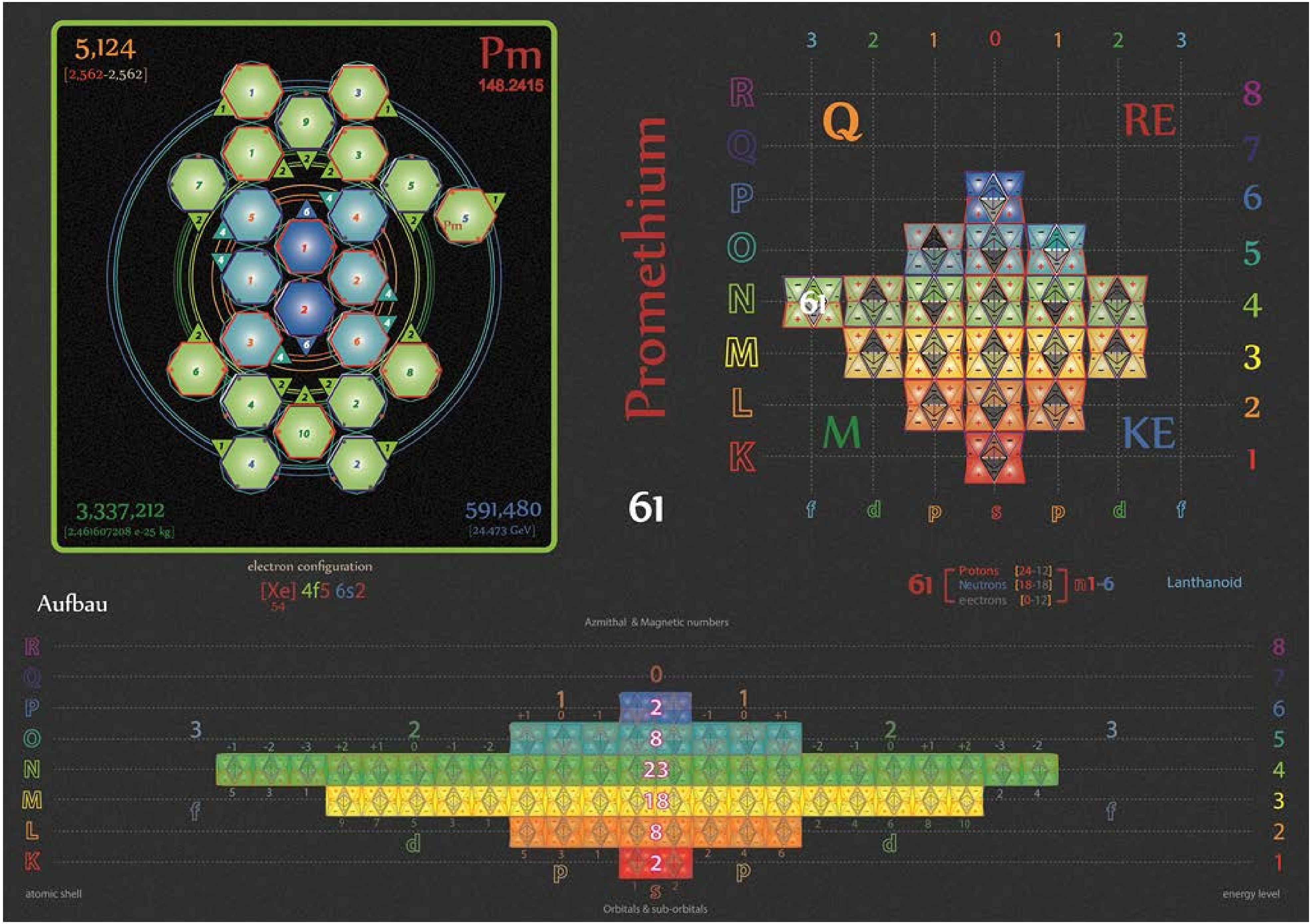
5,124
[2,562-2,562]
Pm
148.2415
3,337,212
[2.461607208 e-25 kg]
591,480
[24.473 GeV]
Promethium
61
electron configuration
[Xe] 4f5 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
61 [Potons [24-12] Neutrons [18-18] eectrons [0-12]] n1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
23
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.61 - Promethium atom

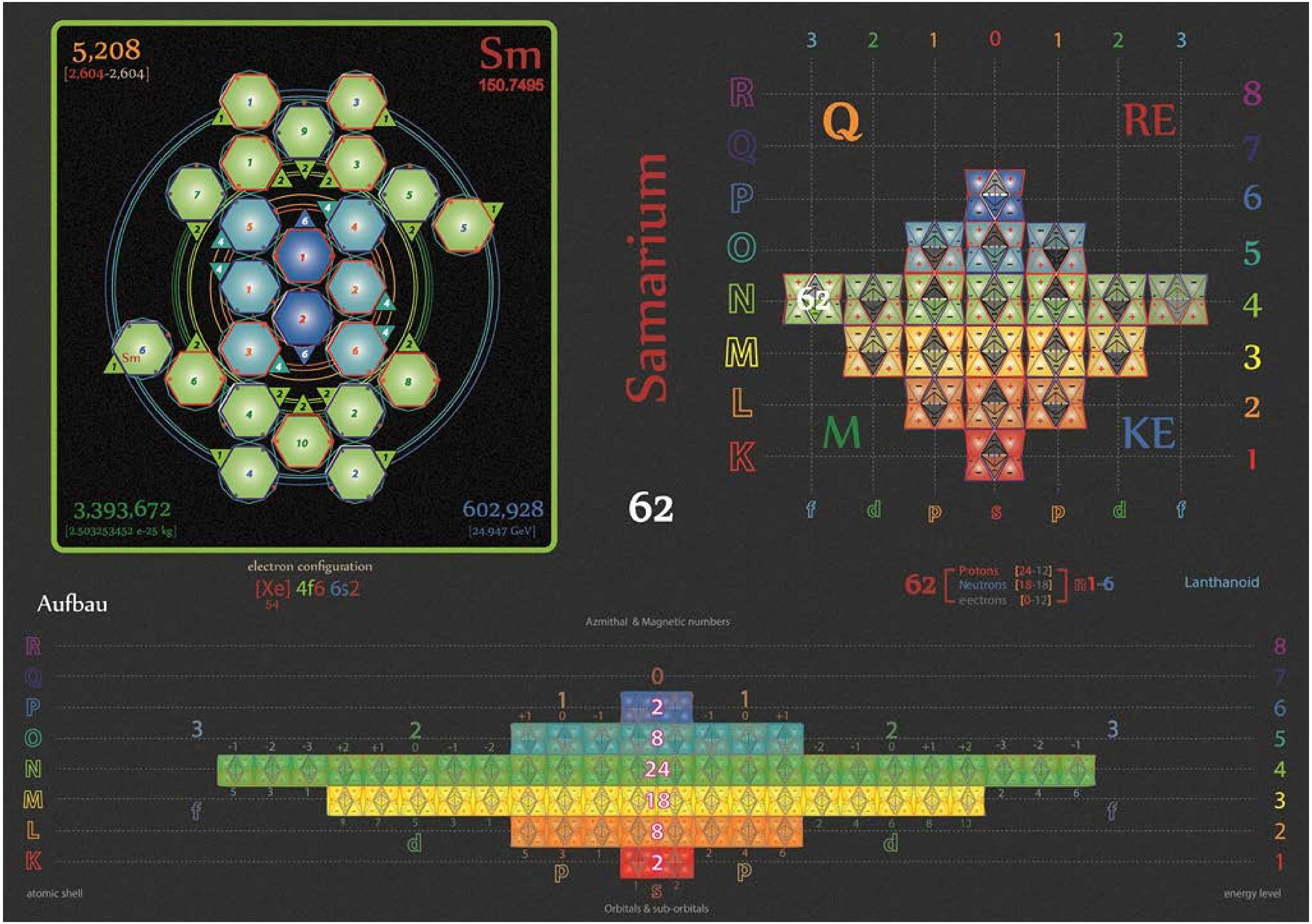
5,208
[2,604-2,604]
Sm
150.7495
3,393,672
[2.503253452 e-25 kg]
602,928
[24.947 GeV]
Samarium
62
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f6 6s2
54
Aufbau
62 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
24
18
8
2
f f
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.62 - Samarium atom

Copyright ABRAHAM [2008] - All rights reserved

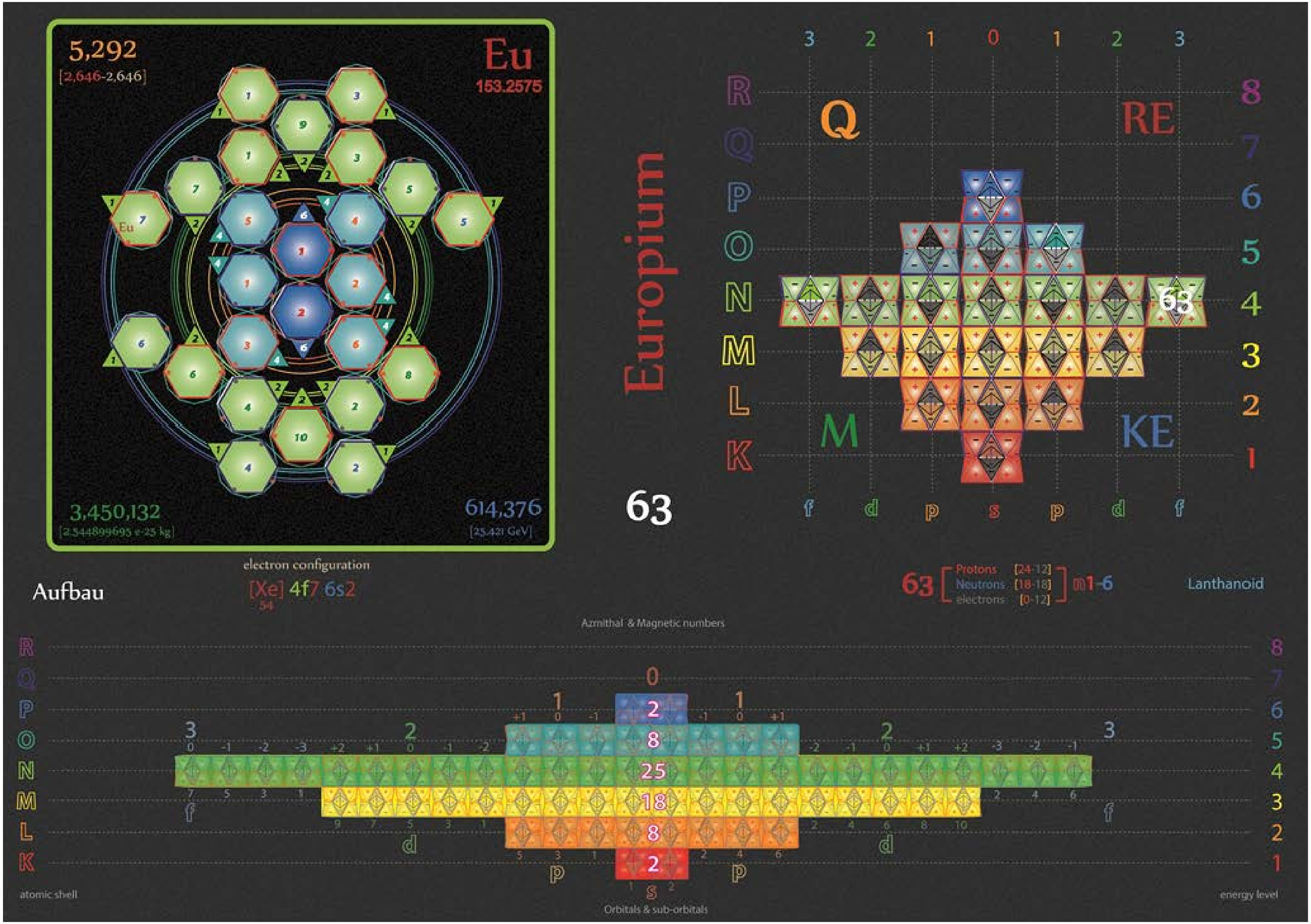
5,292
[2,646-2,646]
Eu
153.2575
3,450,132
[2.544899695 e-25 kg]
614,376
[25.421 GeV]
Europium
63
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Xe] 4f7 6s2
54
63 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
25
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.63 - Europium atom

Copyright ABRAHAM [2008] - All rights reserved

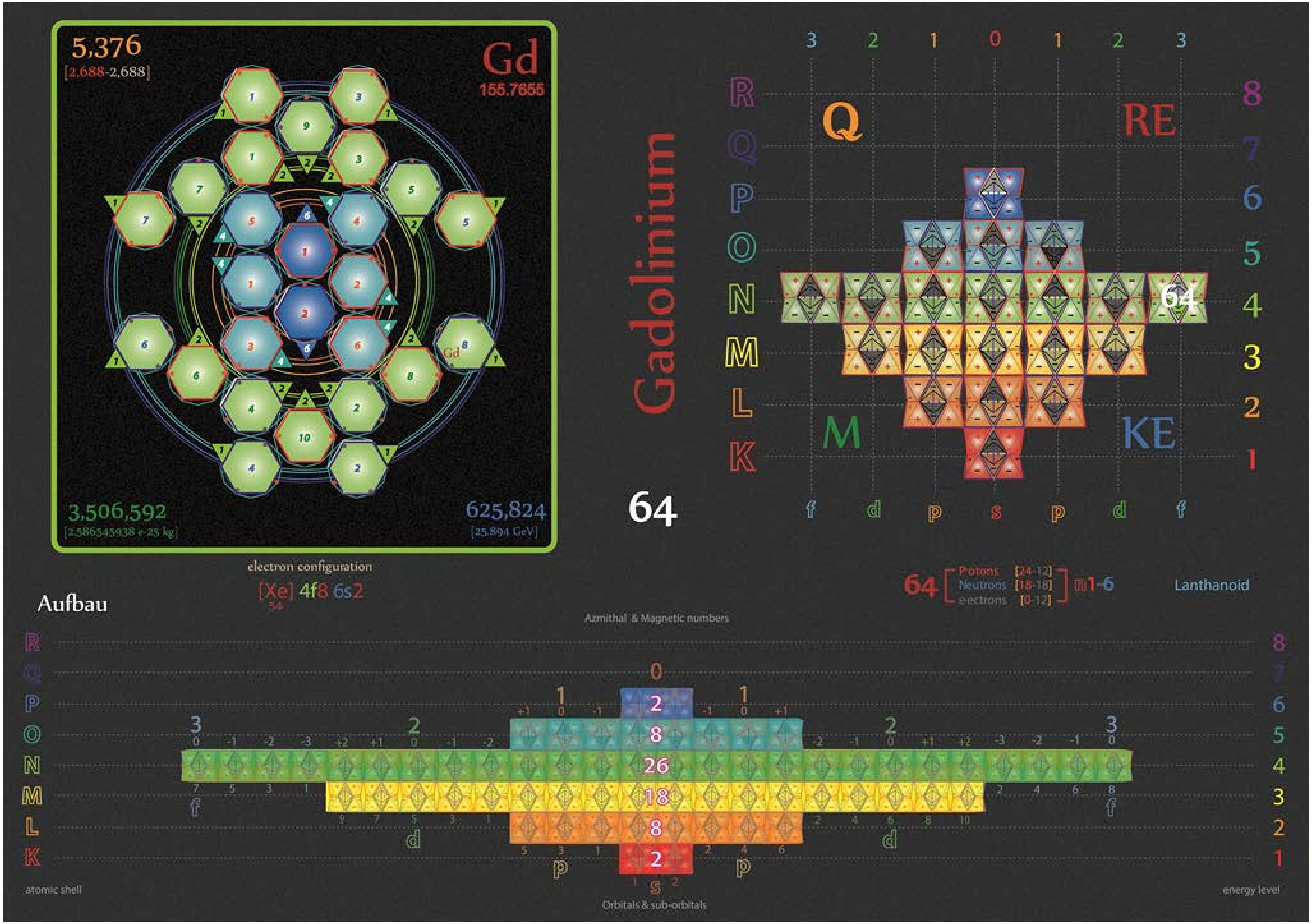
5,376
[2,688-2,688]
Gd
155.7655
3,506,592
[2.586545938 e-25 kg]
625,824
[25.894 GeV]
Gadolinium
64
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
64
f d p s p d f
electron configuration
[Xe] 4f8 6s2
54
Aufbau
64
Potons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
8
26
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.64 - Gadolinium atom

Copyright ABRAHAM [2008] - All rights reserved

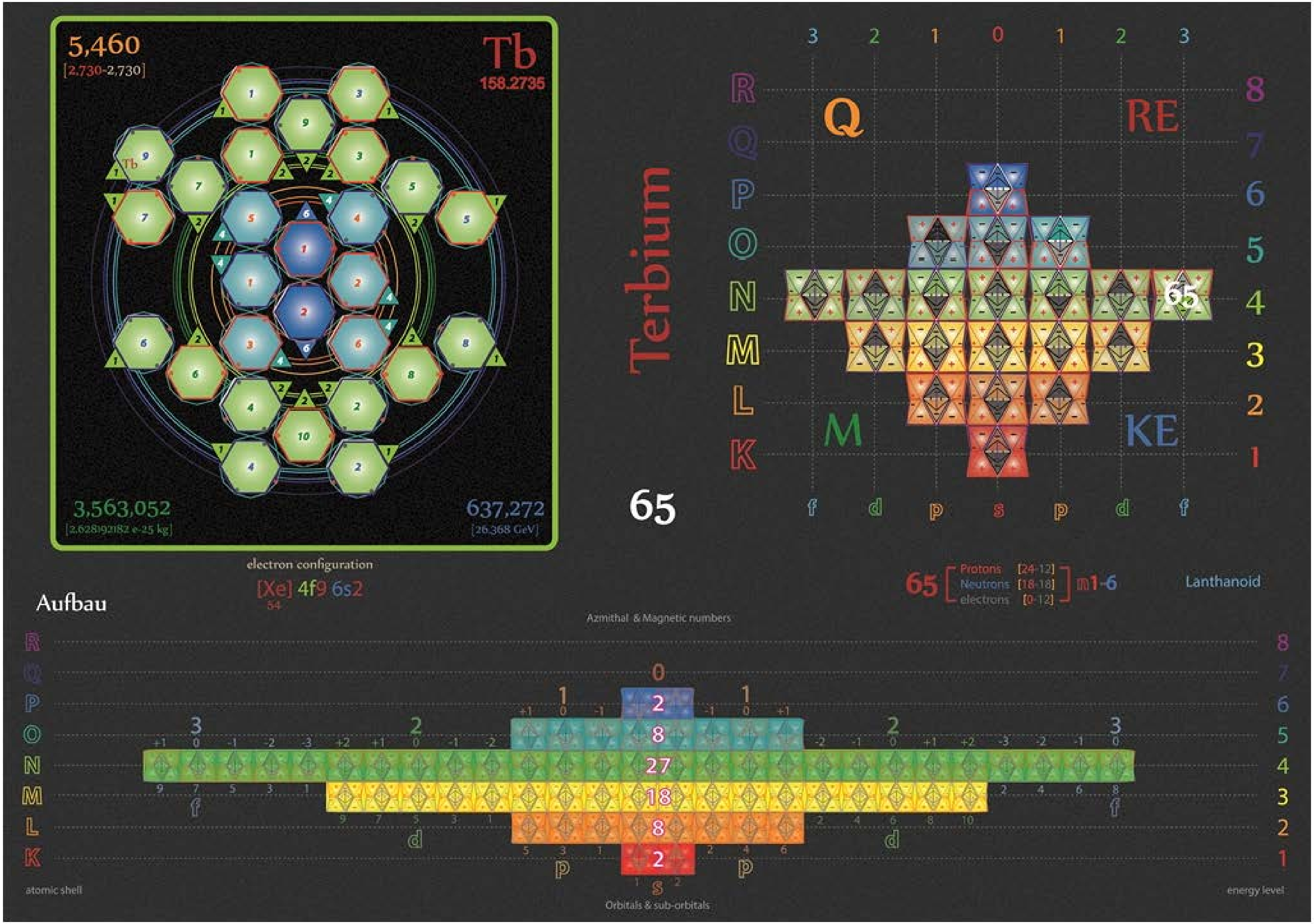
5,460
[2,730-2,730]
Tb
158.2735
3,563,052
[2,628192182 e-25 kg]
637,272
[26,368 GeV]
Terbium
65
electron configuration
[Xe] 4f9 6s2
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
65 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
2
8
27
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.65 -Terbium atom

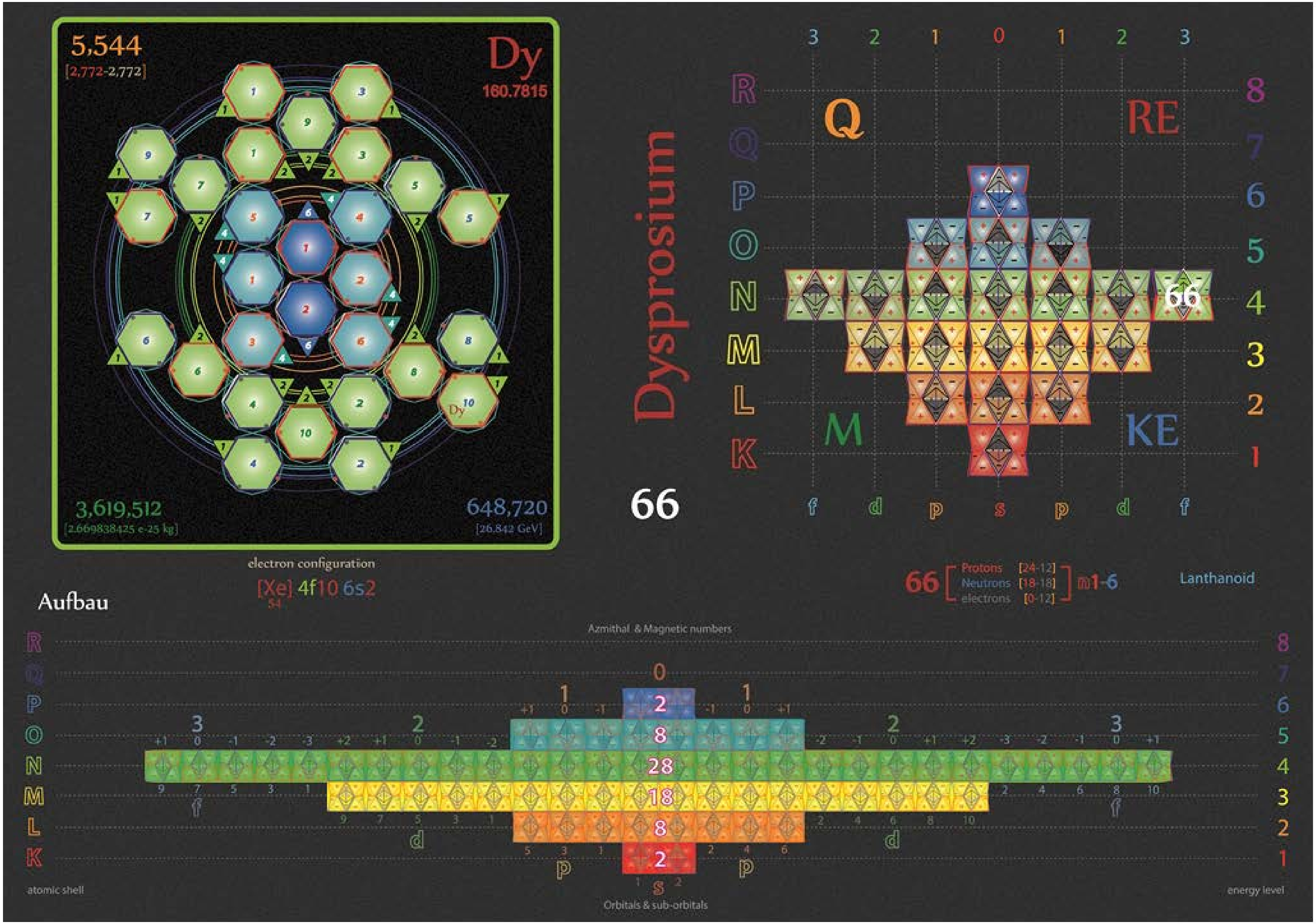
5,544
[2,772-2,772]
Dy
160.7815
3,619,512
[2.669838425 e-25 kg]
648,720
[26.842 GeV]
Dysprosium
66
electron configuration
[Xe] 4f10 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
66
f d p s p d f
66 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
28
18
8
2
f
d
p
s
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.66 - Dysprosium atom

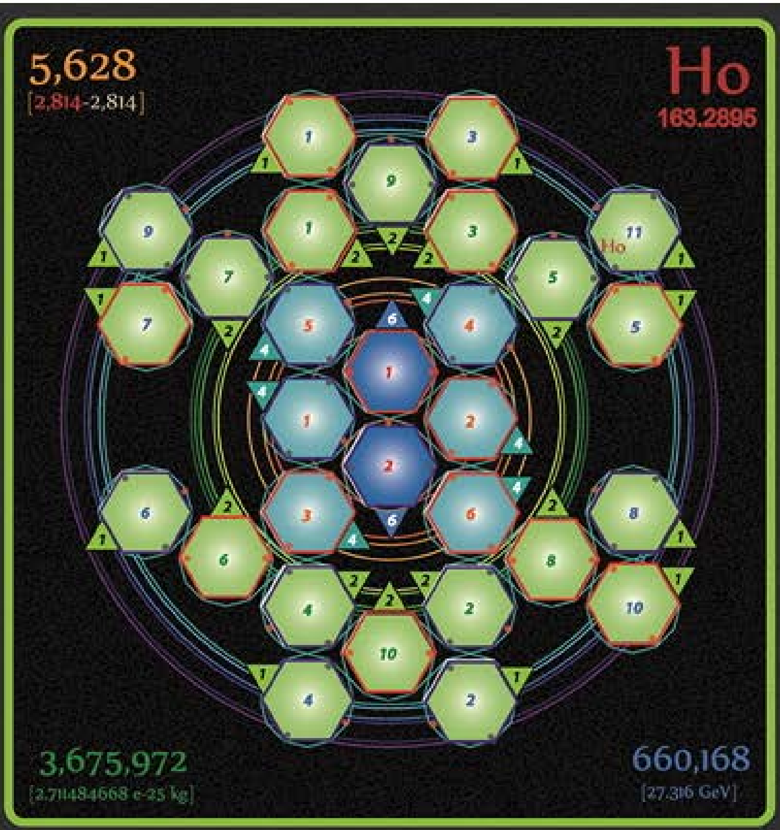

5,628
[2,814-2,814]

Ho
163.2895

3,675,972
[2.711484668 e-25 kg]

660,168
[27.316 GeV]

Holmium

67

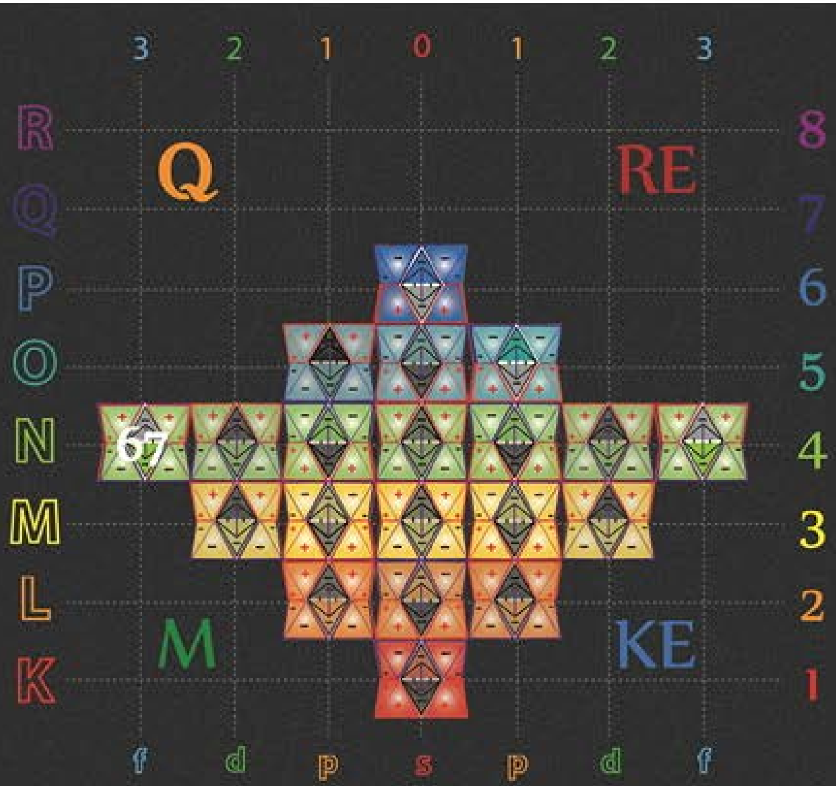
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
67
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Aufbau

electron configuration
[Xe] 4f11 6s2
54

67 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Lanthanoid

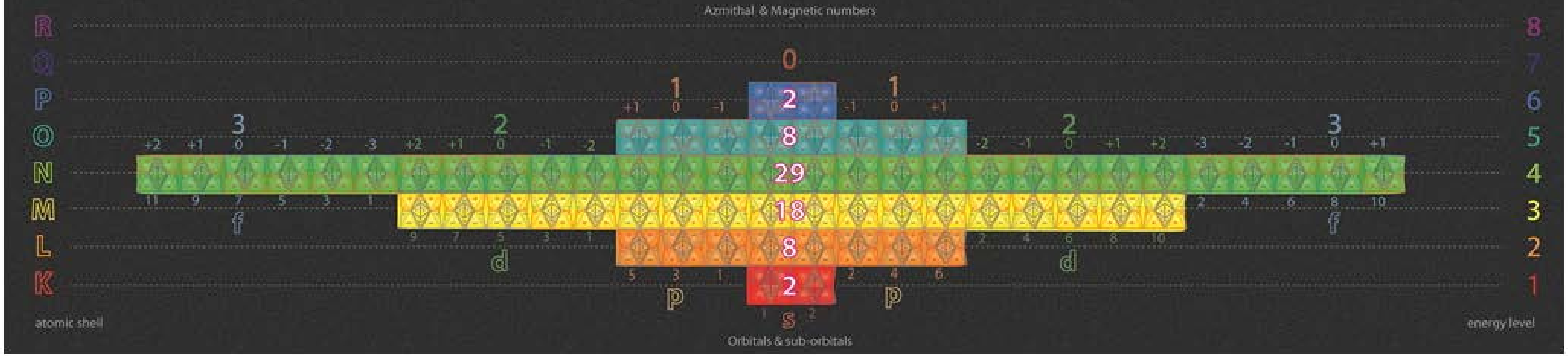
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
8
29
18
8
2
1
1
+1 0 -1
-1 0 +1
3
2
2
3
+2 +1 0 -1 -2 -3 +2 +1 0 -1 -2
-2 -1 0 +1 +2 -3 -2 -1 0 +1
11 9 7 5 3 1
2 4 6 8 10
f
f
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.67 - Holmium atom

Copyright ABRAHAM [2008] - All rights reserved

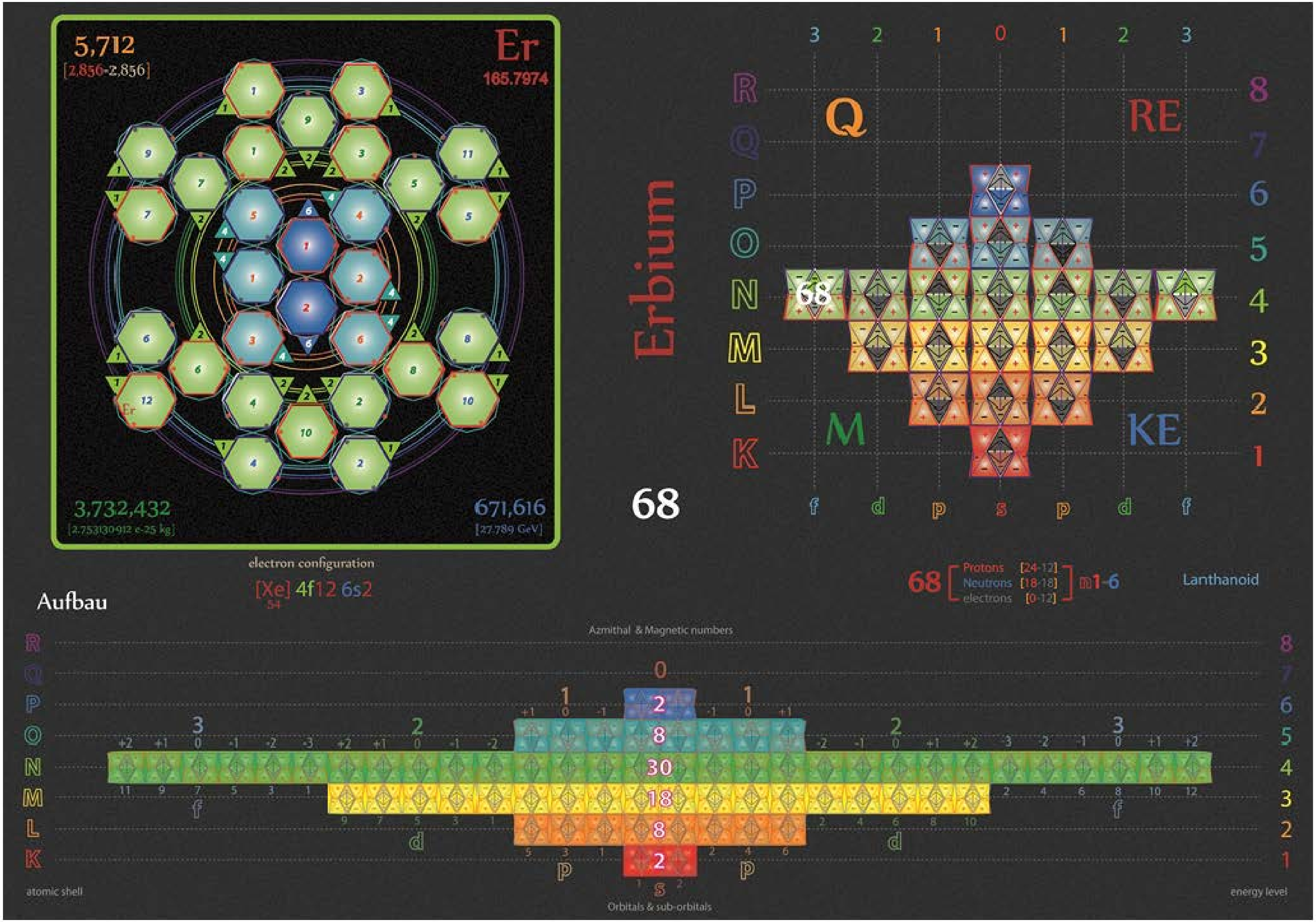
5,712
[2,856-2,856]
Er
165.7974
3,732,432
[2.753130912 e-25 kg]
671,616
[27.789 GeV]
Erbium
68
electron configuration
[Xe] 4f12 6s2
54
Aufbau
68
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Azmithal & Magnetic numbers
0
2
8
30
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.68 - Erbium atom

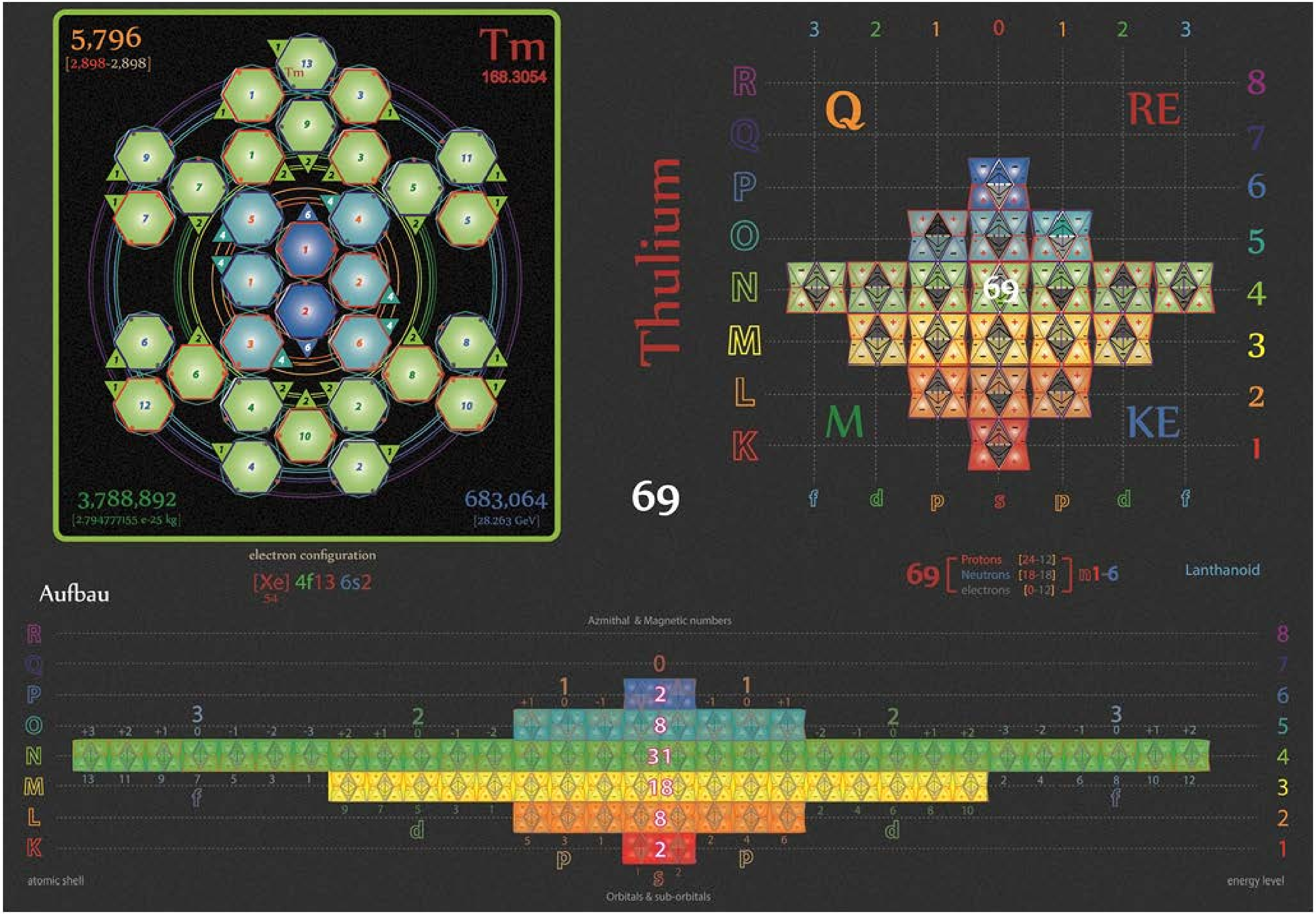
5,796
[2,898-2,898]
Tm
168.3054
3,788,892
[2.794777155 e-25 kg]
683,064
[28.263 GeV]
Thulium
69
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f13 6s2
54
Aufbau
69 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2
8
31
18
8
2
3 2 2 3
atomic shell
energy level
f d p s p d f
Orbitals & sub-orbitals

Tetryonics 51.69 - Thulium atom

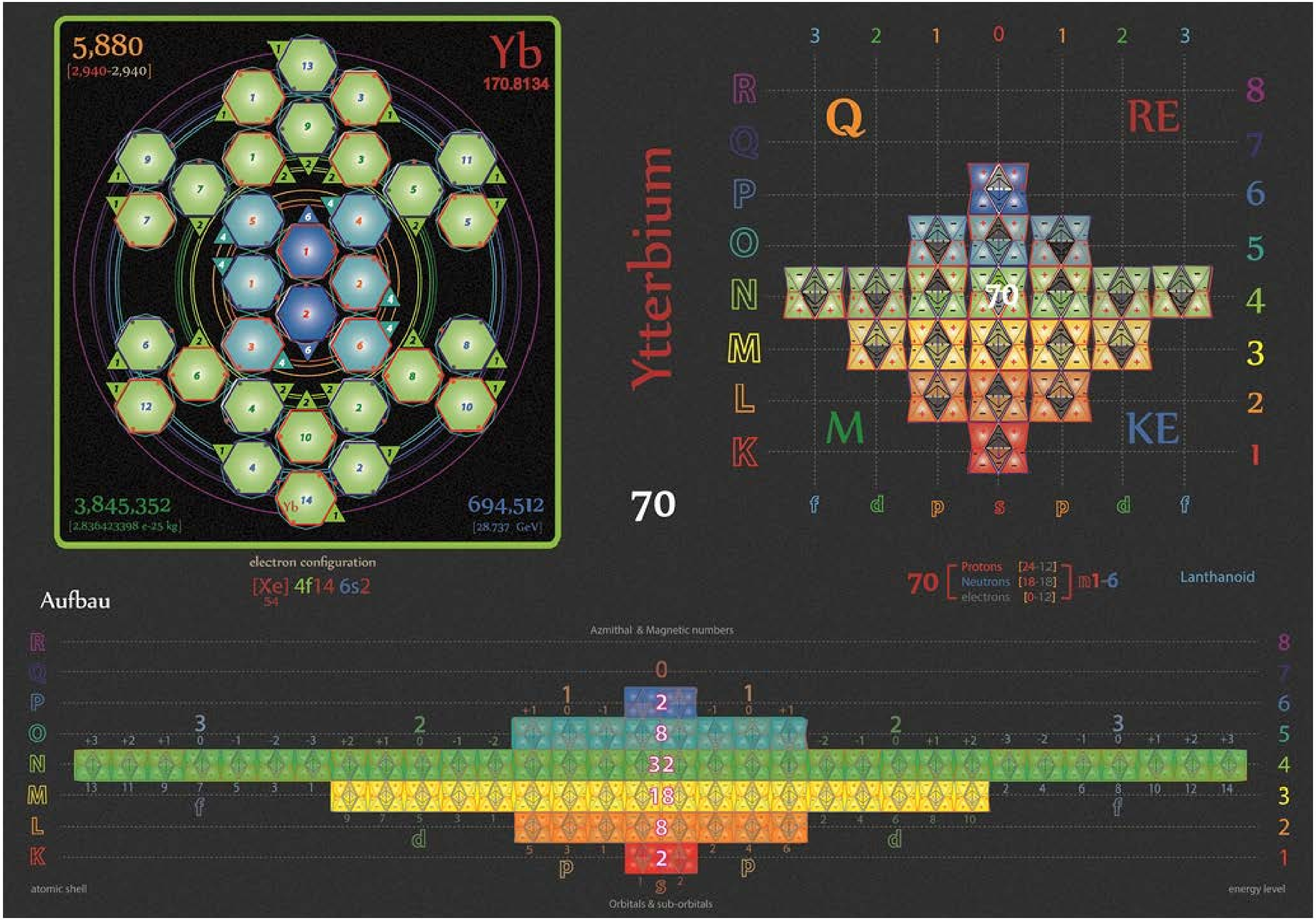
5,880
[2,940-2,940]
Yb
170.8134
3,845,352
[2.836423398 e-25 kg]
694,512
[28.737 GeV]
Ytterbium
70
electron configuration
[Xe] 4f14 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
70
f d p s p d f
70 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
+3 +2 +1 0 -1 -2 -3
2
8
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.70 - Ytterbium atom

Copyright ABRAHAM [2008] - All rights reserved

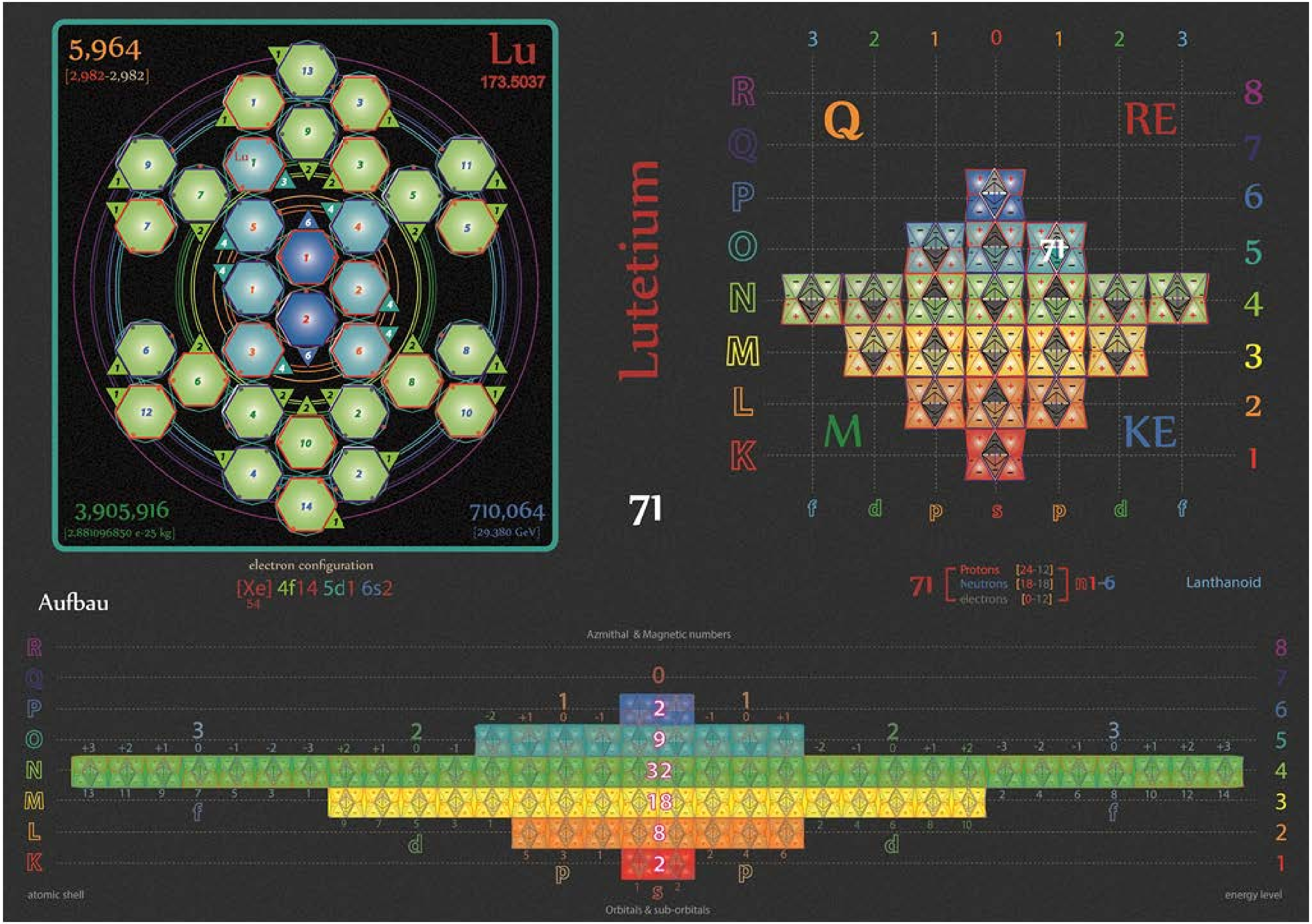
5,964
[2,982-2,982]
Lu
173.5037
Lu
3,905,916
[2.881096850 e-25 kg]
710,064
[29.380 GeV]
Lutetium
71
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
71
f d p s p d f
electron configuration
[Xe] 4f14 5d1 6s2
54
71 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
2
9
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.71 - Lutetium atom

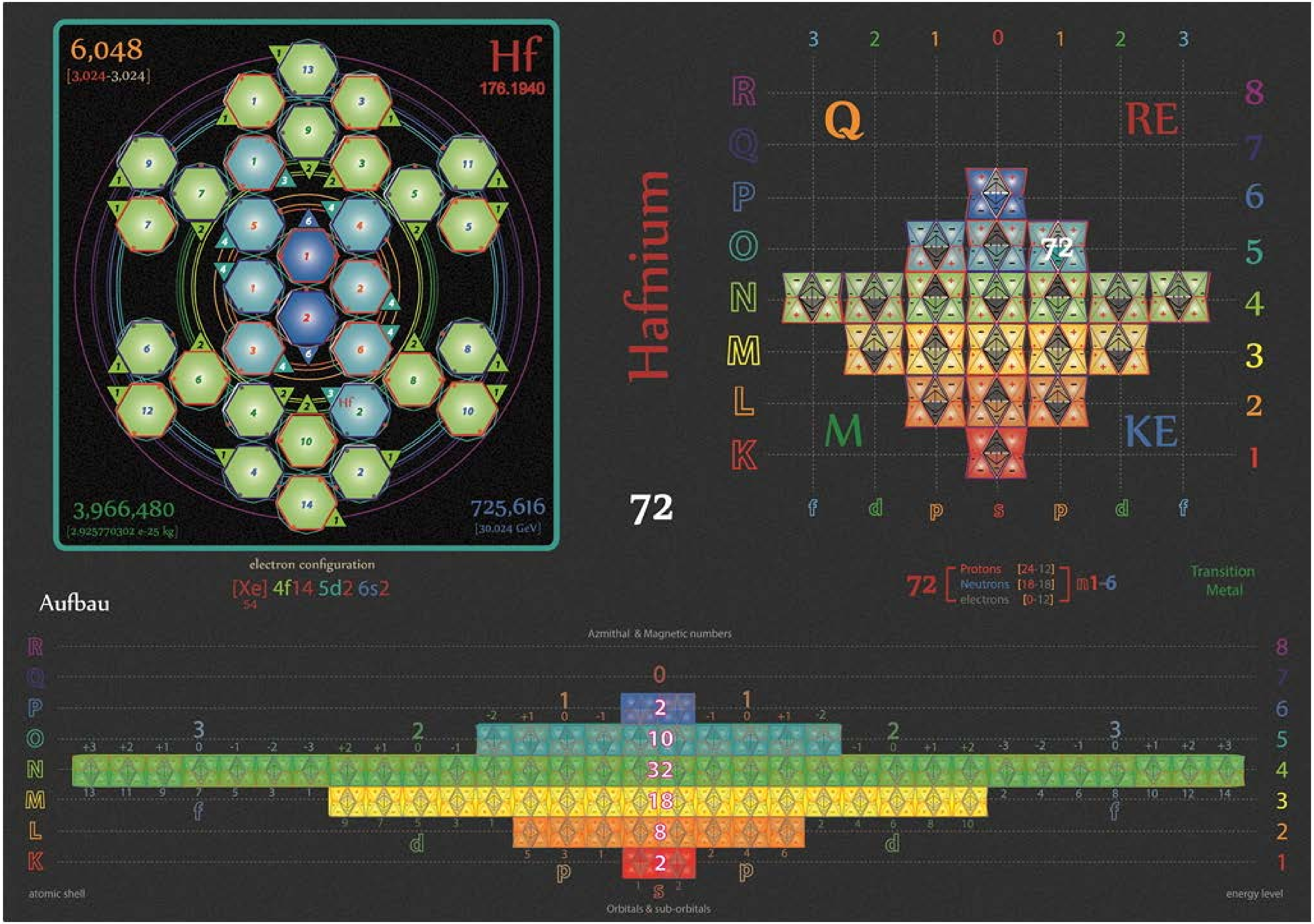
6,048
[3,024-3,024]
Hf
176.1940
3,966,480
[2.925770302 e-25 kg]
725,616
[30.024 GeV]
Hafnium
72
electron configuration
[Xe] 4f14 5d2 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
72
72 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
2
10
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.72 - Hafnium atom

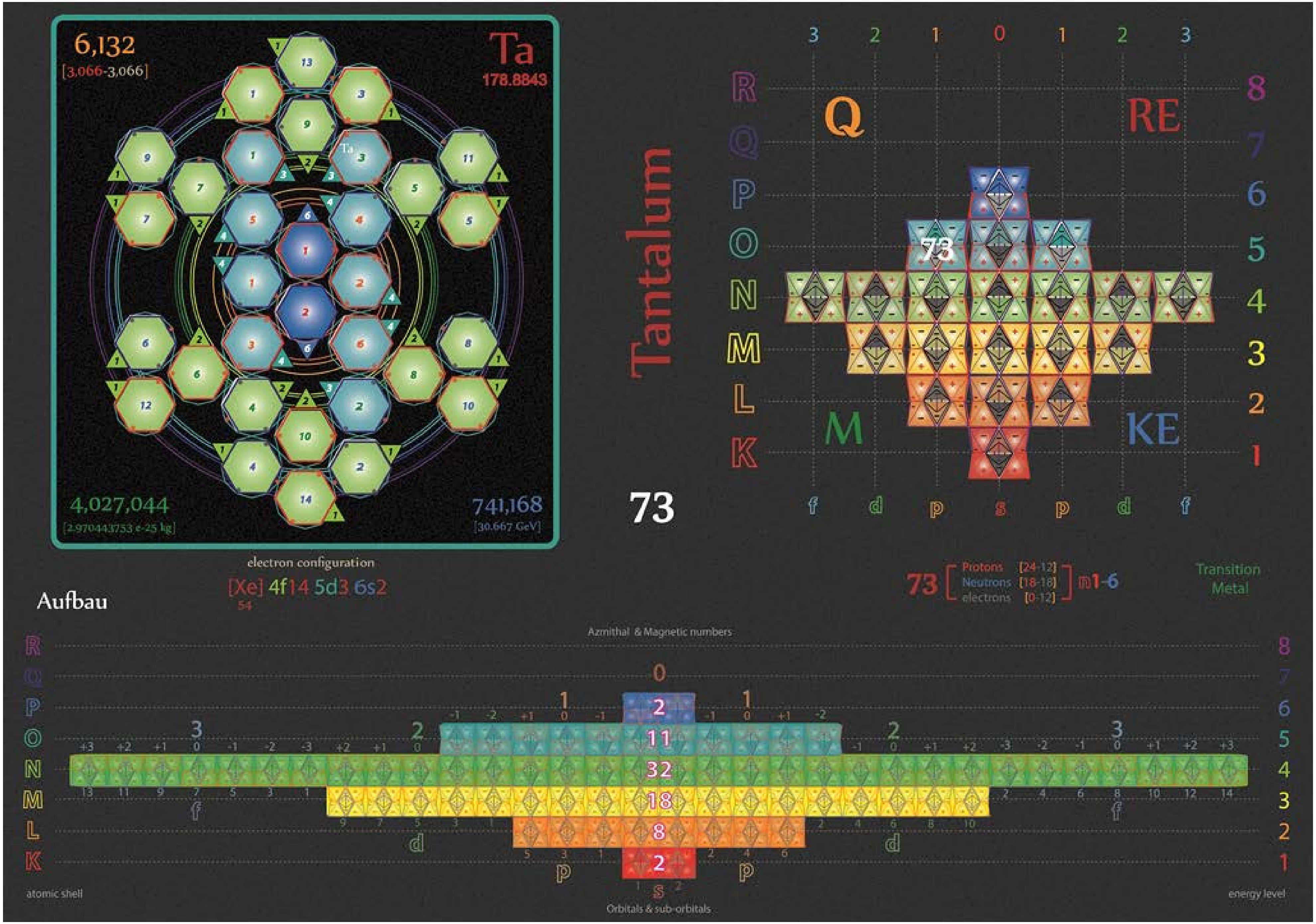
6,132
[3,066-3,066]
Ta
178.8843
4,027,044
[2.970443753 e-25 kg]
741,168
[30.667 GeV]
Tantalum
73
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
73
73 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
electron configuration
[Xe] 4f14 5d3 6s2
54
Aufbau
Azmithal & Magnetic numbers
2
11
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.73 - Tantalum atom

Copyright ABRAHAM [2008] - All rights reserved

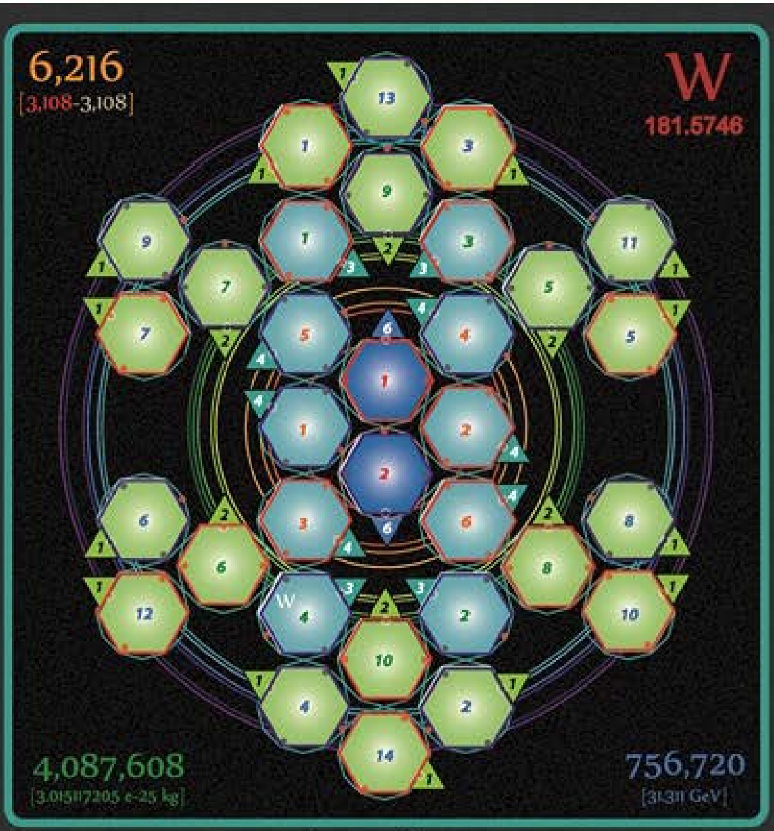
6,216
[3,108-3,108]
W
181.5746
4,087,608
[3.015117205 e-25 kg]
756,720
[31.311 GeV]

electron configuration

[Xe] 4f14 5d4 6s2
54

Tungsten

74

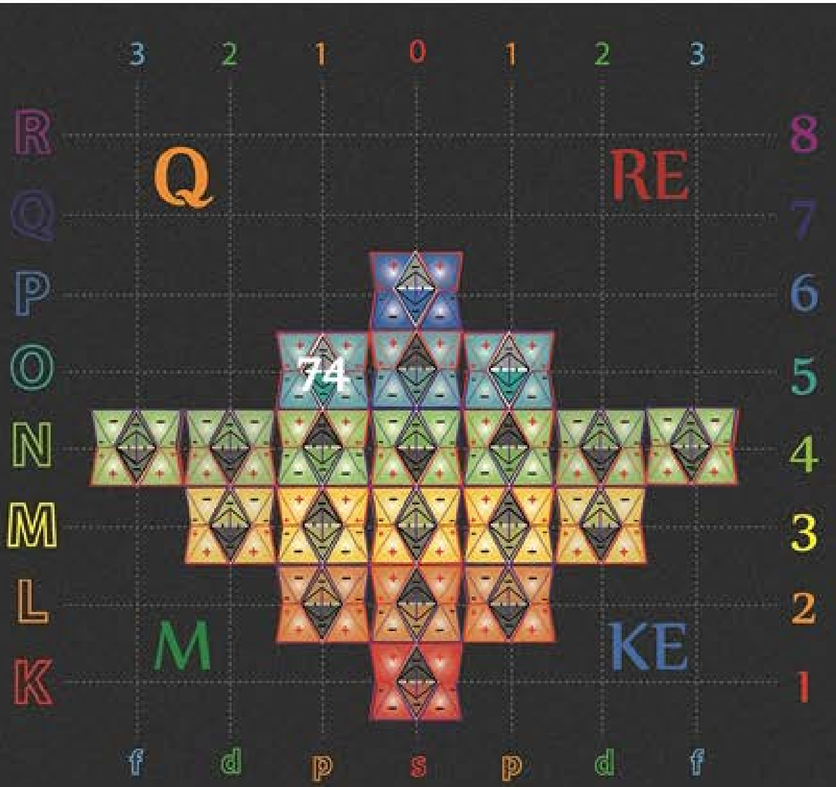
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
74
M
KE
f
d
p
s
p
d
f

74 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Transition Metal

Aufbau

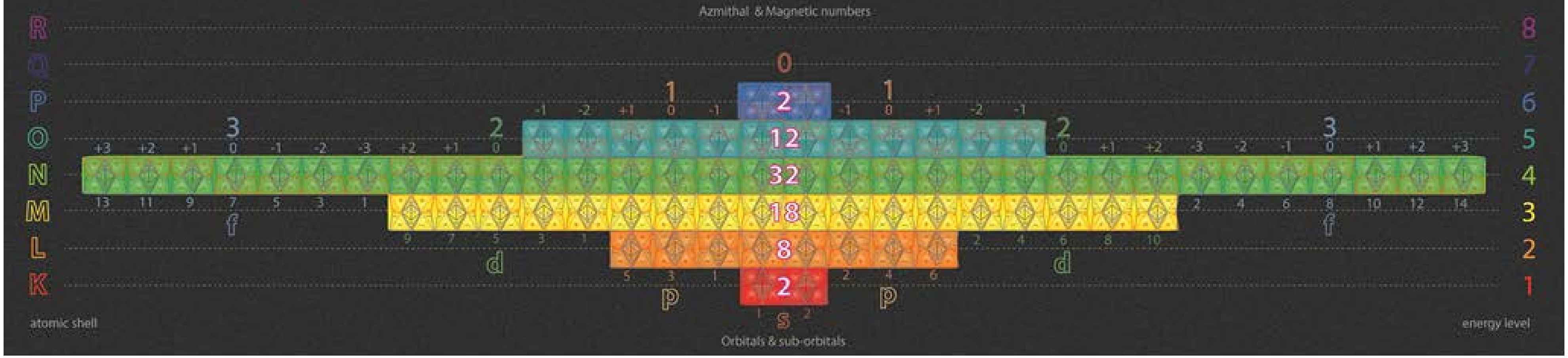
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
12
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.74 - Tungsten atom

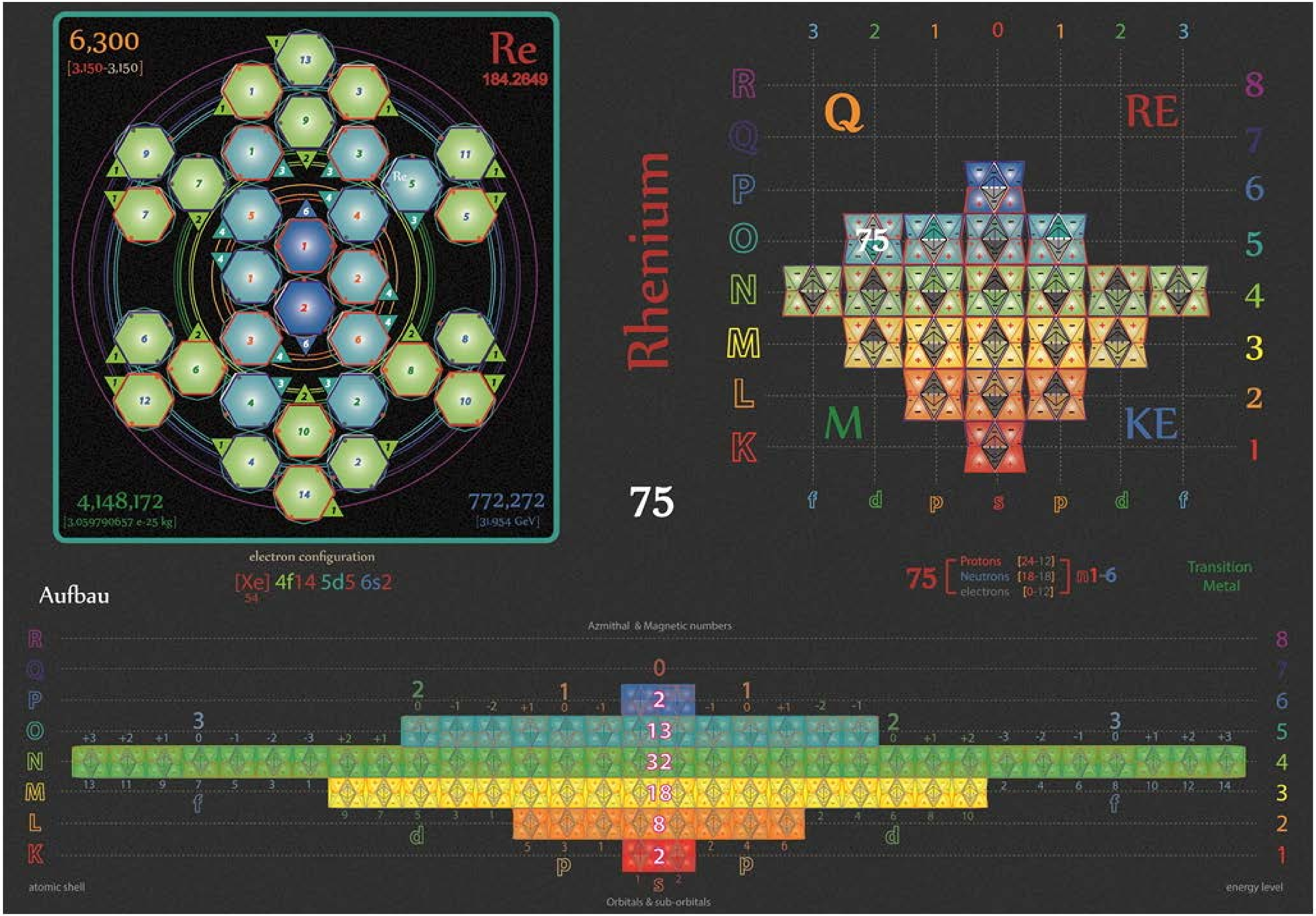
6,300
[3,150-3,150]
Re
184.2849
4,148,172
[3.059790657 e-25 kg]
772,272
[31.954 GeV]
Rhenium
75
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Xe] 4f14 5d5 6s2
54
75 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Transition Metal
Aufbau
Azmithal & Magnetic numbers
2
13
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.75 - Rhenium atom

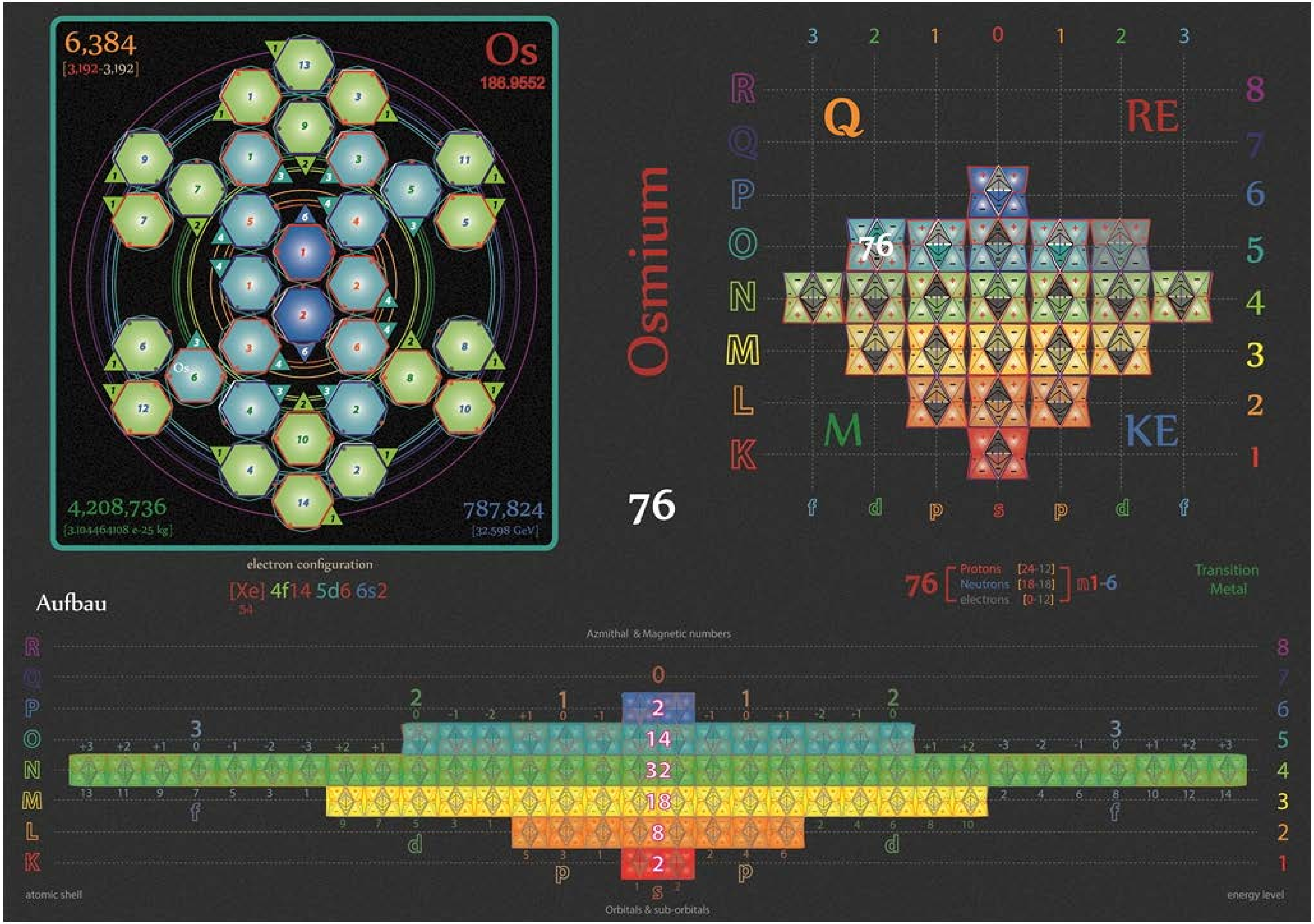
6,384
[3,192-3,192]
Os
186.9552
4,208,736
[3.104464108 e-25 kg]
787,824
[32.598 GeV]
electron configuration
[Xe] 4f14 5d6 6s2
54
Osmium
76
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
76 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2 14 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.76 - Osmium atom

Copyright ABRAHAM [2008] - All rights reserved

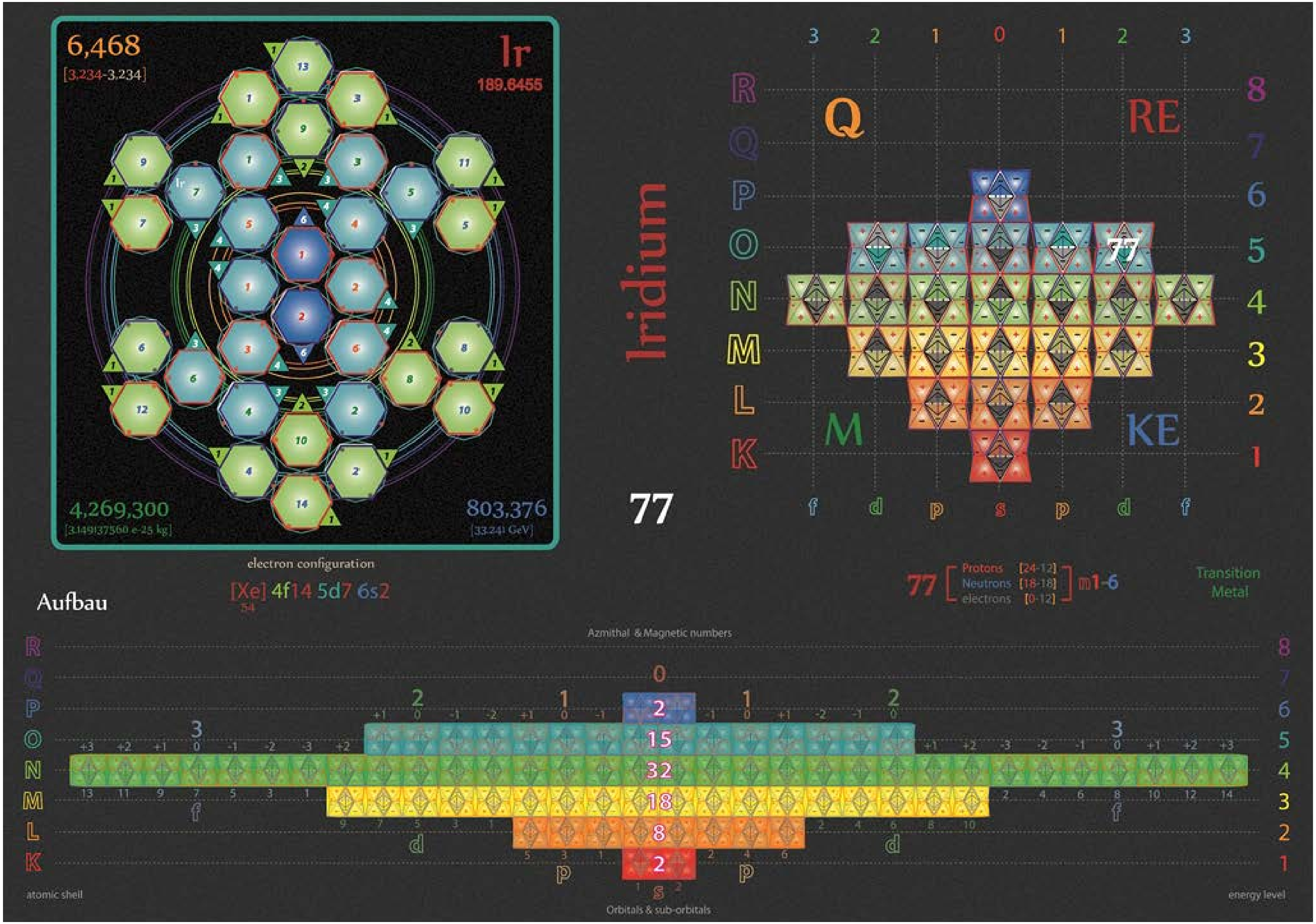
6,468
[3.234-3.234]
Ir
189.6455
4,269,300
[3.149137560 e-25 kg]
803,376
[33.241 GeV]
Iridium
77
electron configuration
[Xe] 4f14 5d7 6s2
54
Aufbau
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
77
77 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
2
15
32
18
8
2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.77 - Iridium atom

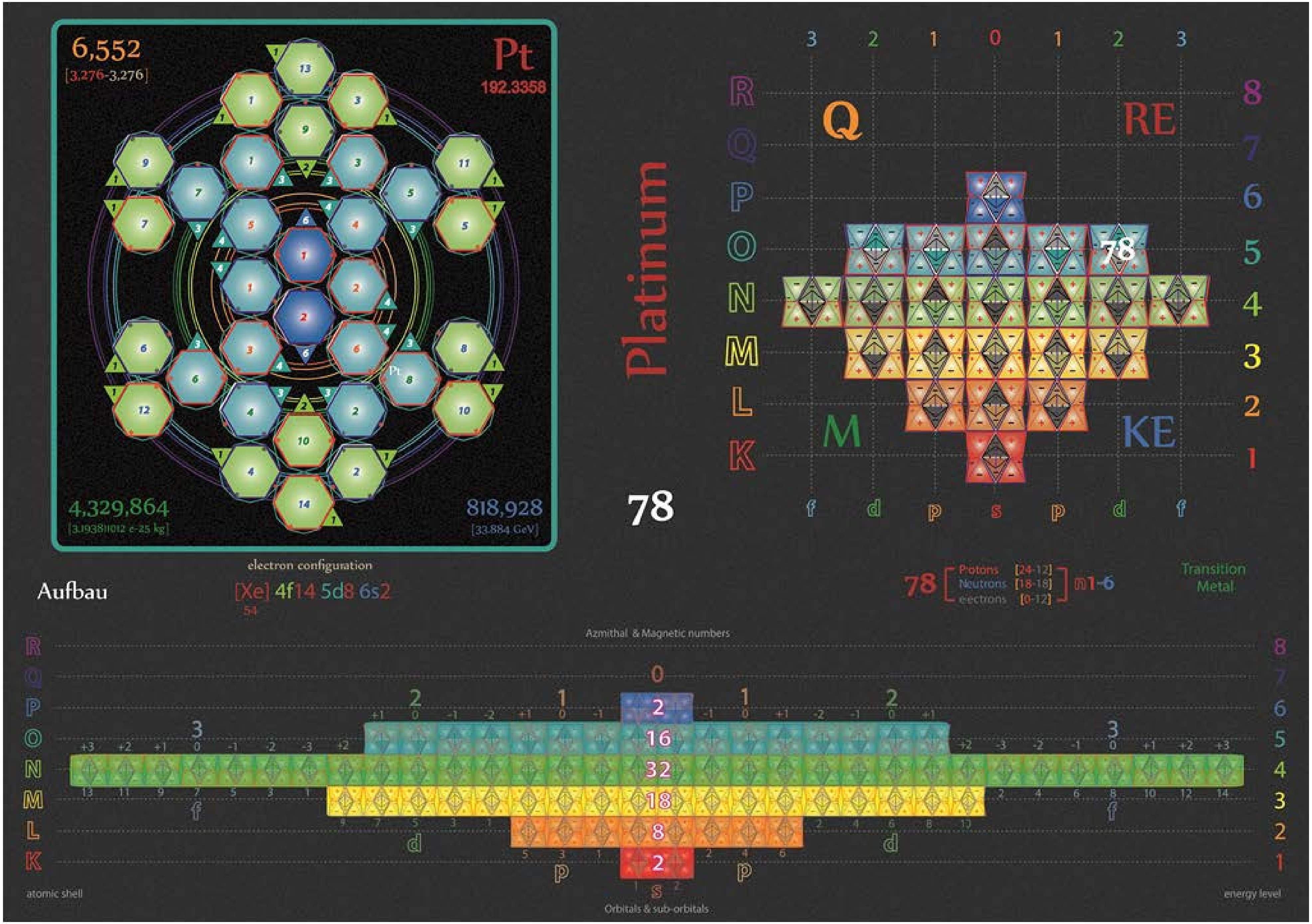
6,552
[3,276-3,276]
Pt
192.3358
4,329,864
[3.19381012 e-25 kg]
818,928
[33.884 GeV]
Platinum
78
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d8 6s2
54
78 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
16
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.78 - Platinum atom

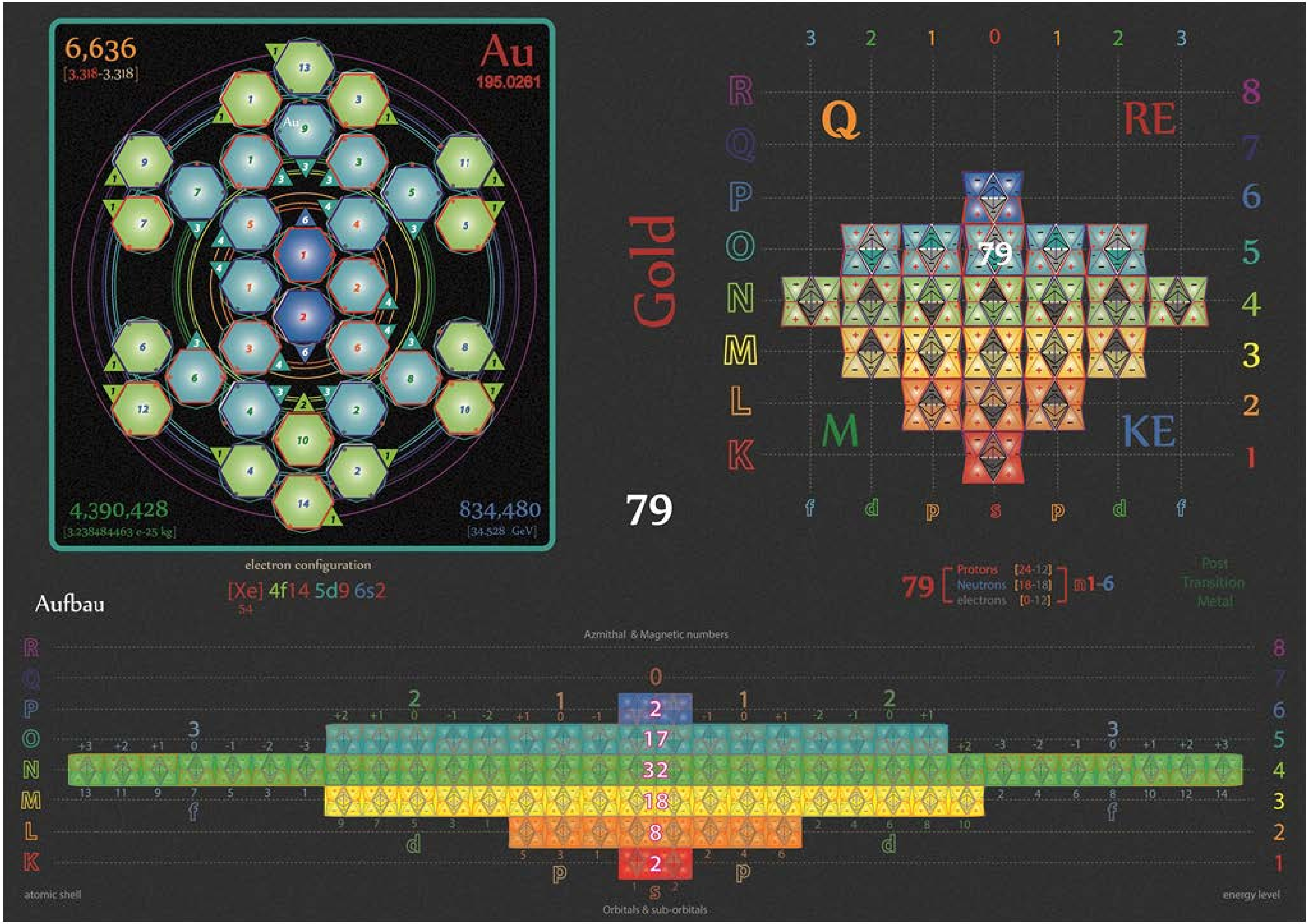
6,636
[3.318-3.318]
Au
195.0261
4,390,428
[3.238484463 e-25 kg]
834,480
[34.528 GeV]
Gold
79
electron configuration
[Xe] 4f14 5d9 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
79
79 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Post Transition Metal
Azmithal & Magnetic numbers
2
17
32
18
8
2
f
d
p
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.79 - Gold atom

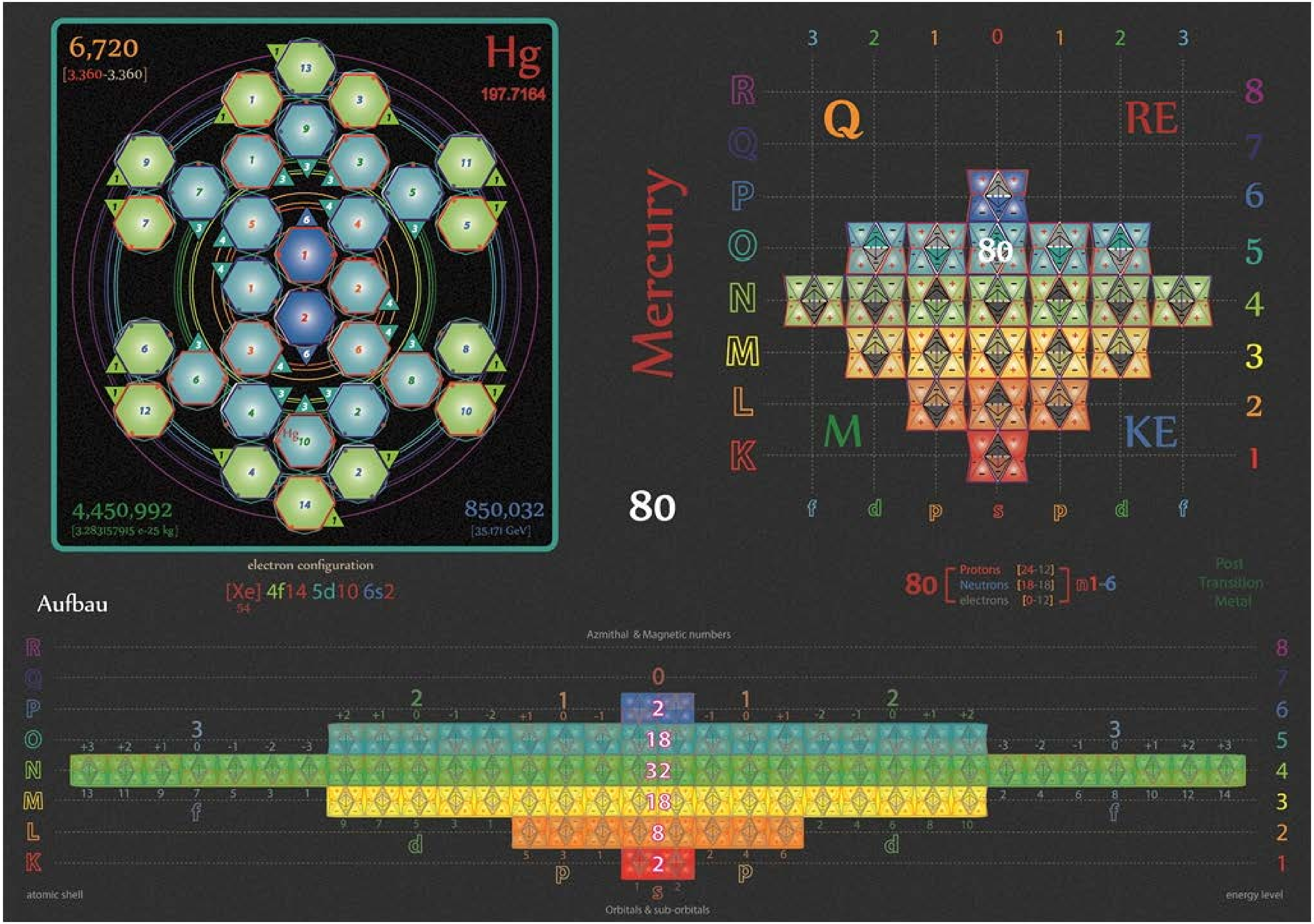
6,720
[3,360-3,360]
Hg
197.7164
4,450,992
[3.283157915 e-25 kg]
850,032
[35.171 GeV]
Hg 10
electron configuration
[Xe] 4f14 5d10 6s2
54
Aufbau
Mercury
80
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
80
f d p s p d f
80 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Post Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
2
18
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.80 - Mercury atom

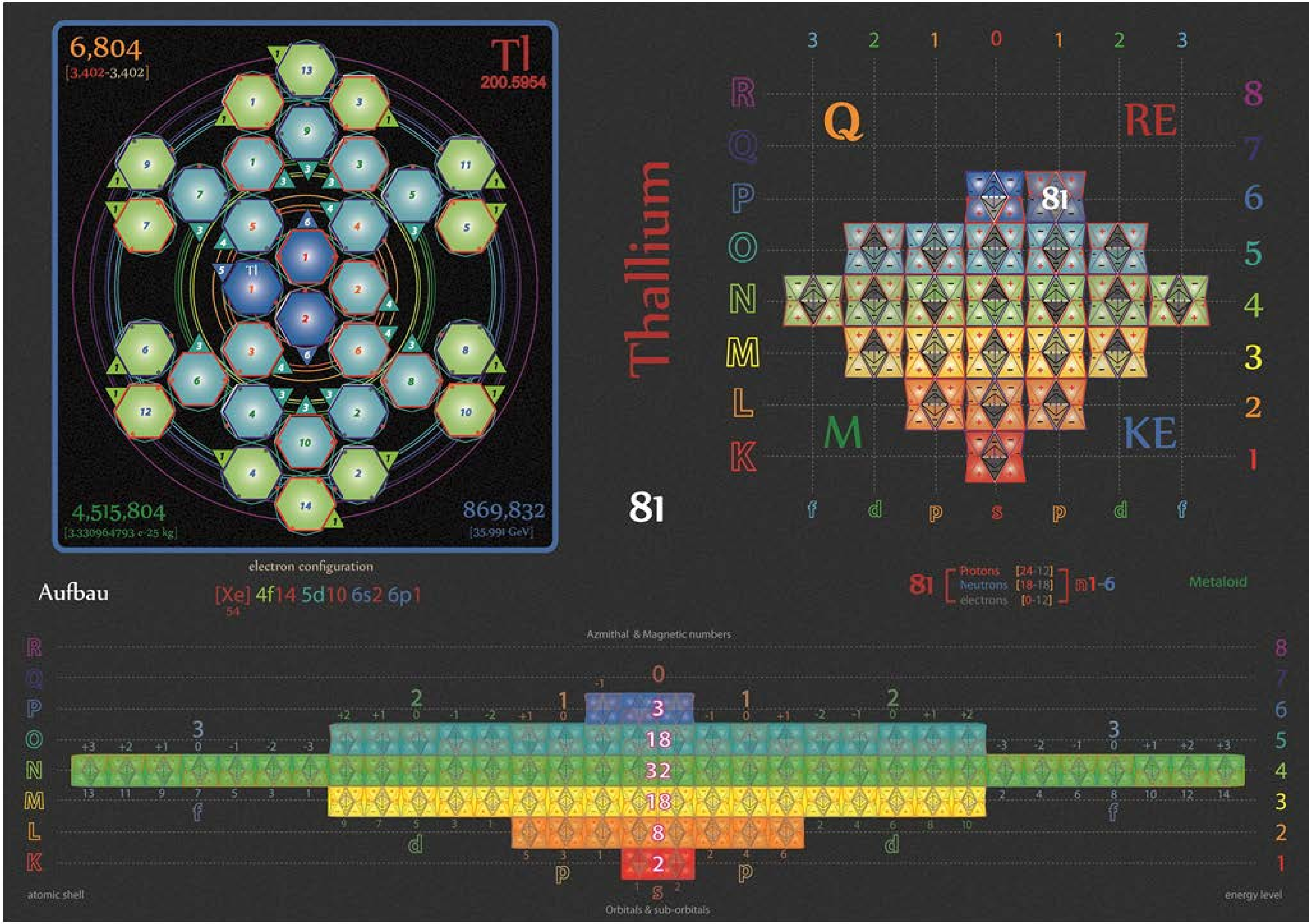
6,804
[3,402-3,402]
Tl
200.5954
4,515,804
[3.330964793 e-25 kg]
869,832
[35.991 GeV]
Thallium
81
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8
7
6
5
4
3
2
1
Q
RE
M
KE
81
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p1
54
81 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Metaloid
Azmithal & Magnetic numbers
0
3
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.81 - Thallium atom

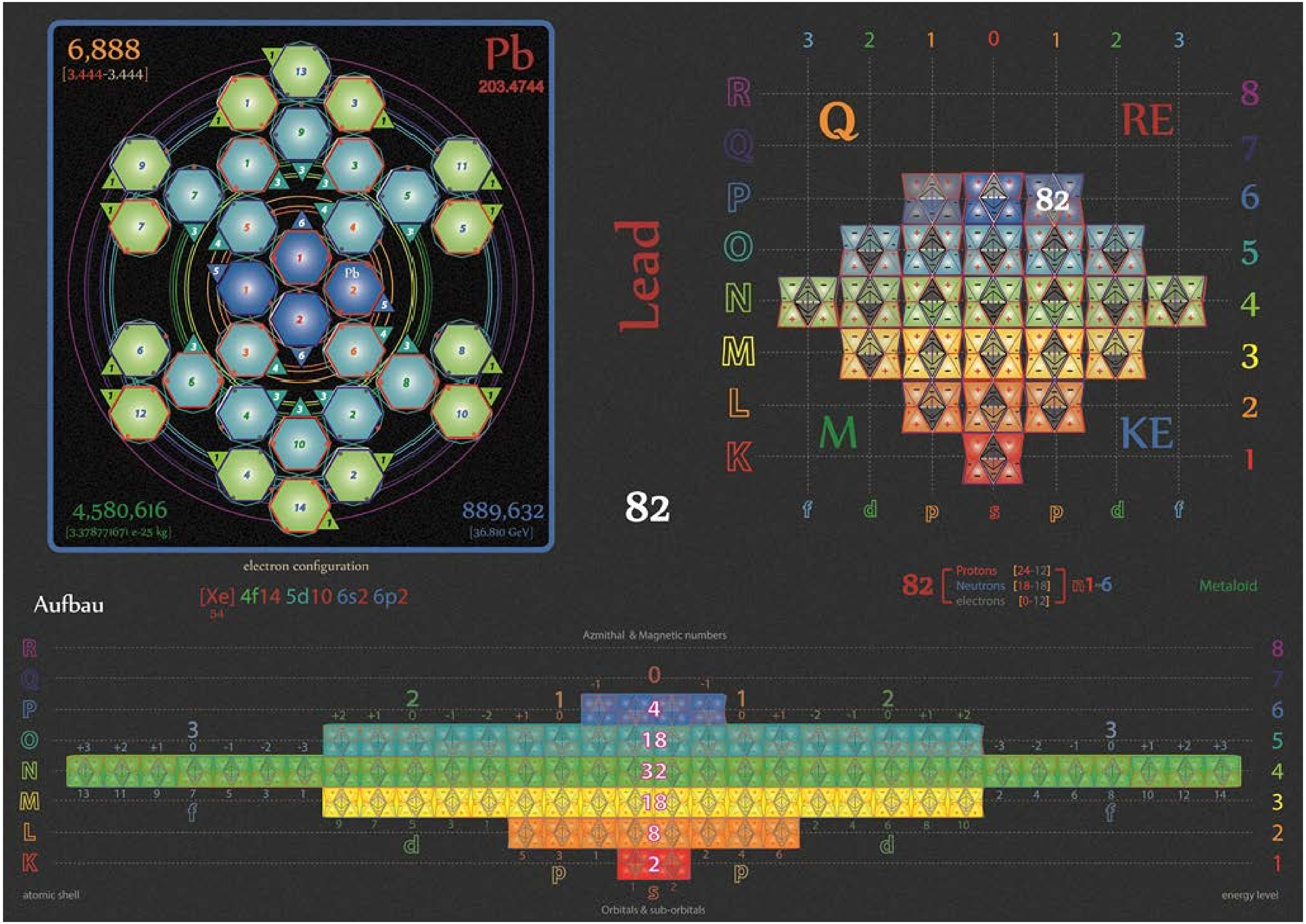
6,888
[3.444-3.444]
Pb
203.4744
4,580,616
[3.37877167 e-25 kg]
889,632
[36.810 GeV]
Lead
82
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p2
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
82
f d p s p d f
82
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.82 - Lead atom

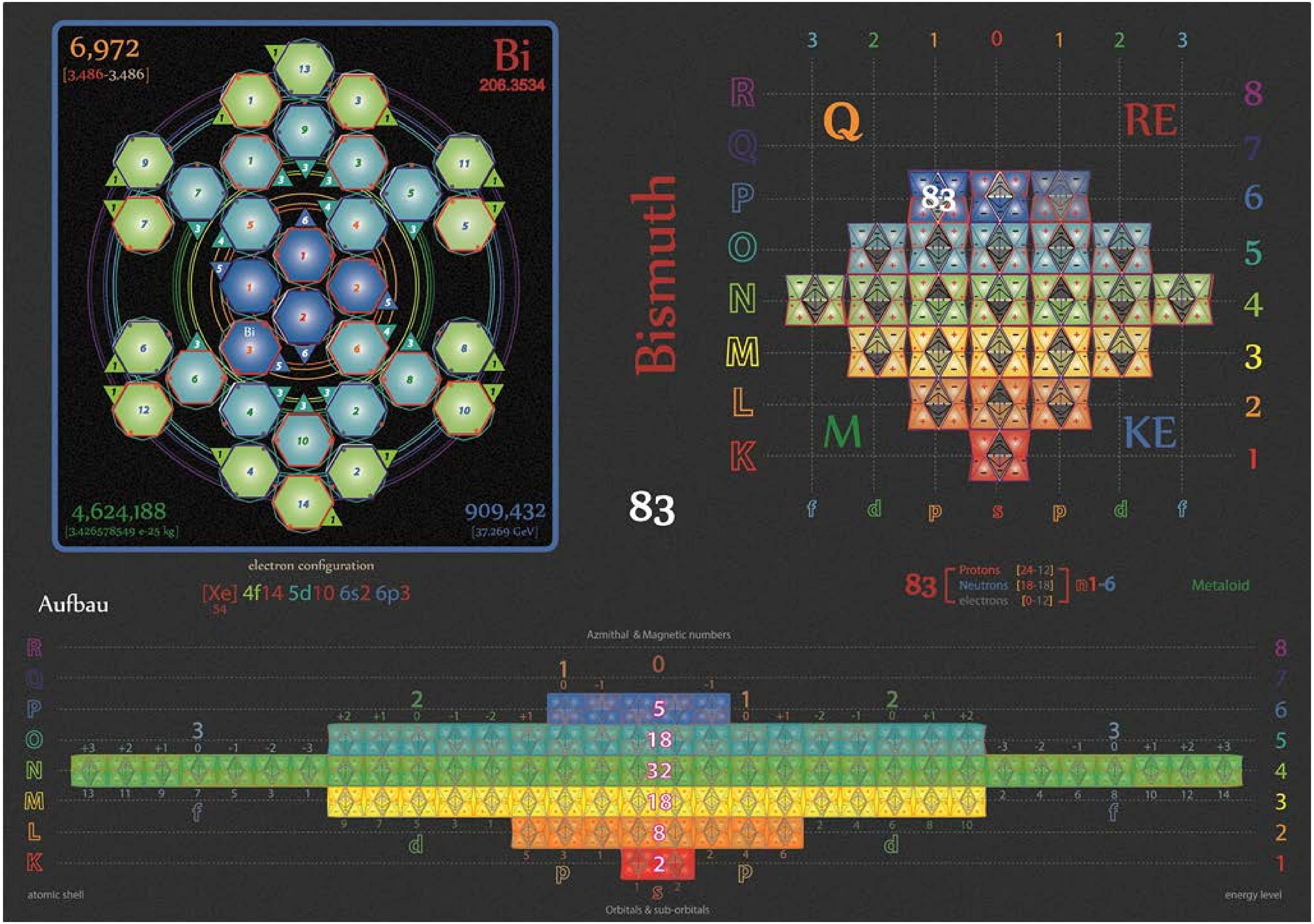
6,972
[3,486-3,486]
Bi
208.3534
4,624,188
[3.426578549 e-25 kg]
909,432
[37.269 GeV]
Bismuth
83
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p3
54
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
f d p s p d f
83 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Metaloid
Azmithal & Magnetic numbers
5
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.83 - Bismuth atom

Copyright ABRAHAM [2008] - All rights reserved

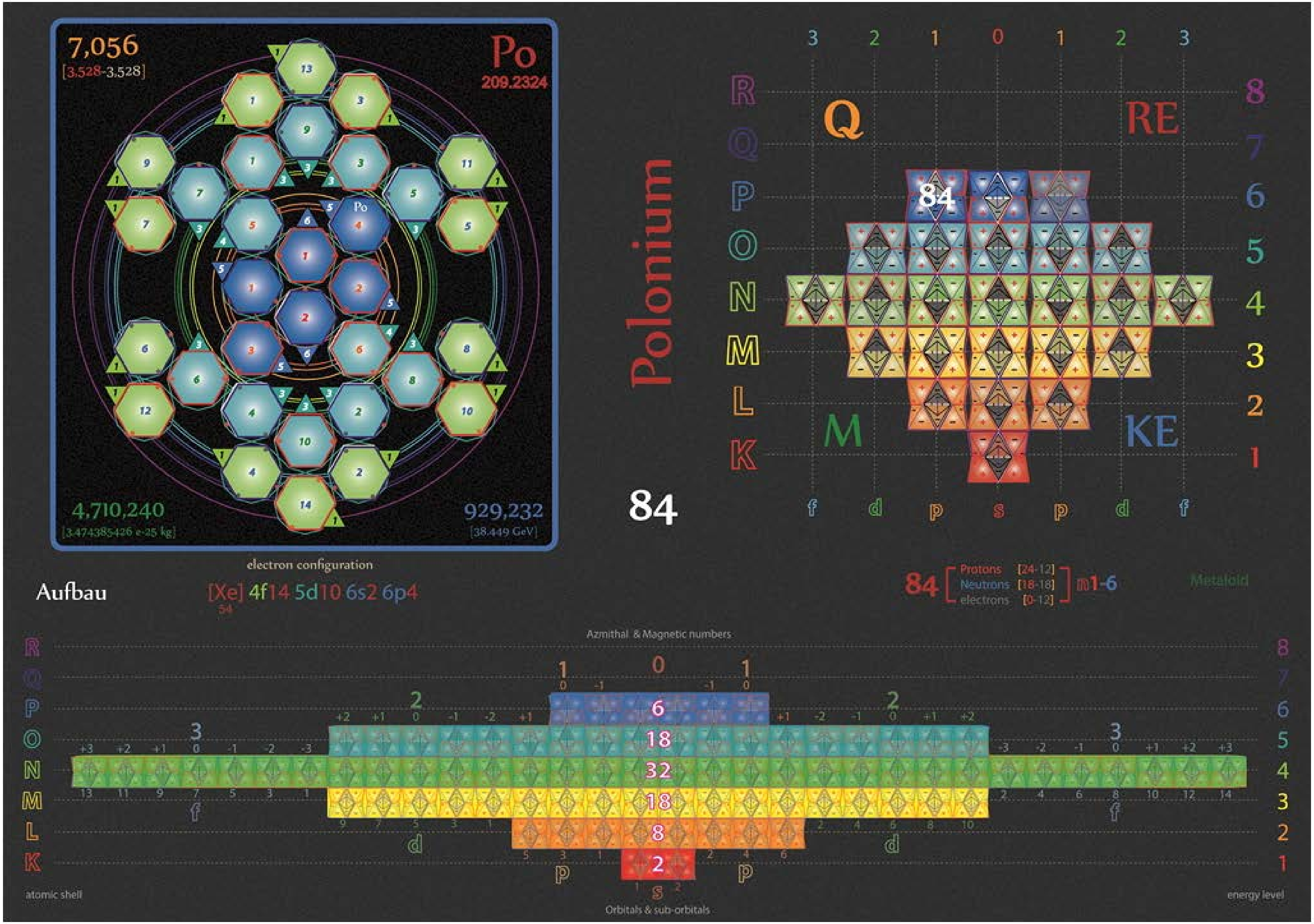
7,056
[3.528-3.528]
Po
209.2324
4,710,240
[3.474385426 e-25 kg]
929,232
[38.449 GeV]
Polonium
84
R
Q
P
O
N
M
L
K
3
2
1
0
1
2
3
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p4
54
84
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Metalloid
Azmithal & Magnetic numbers
6
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.84 - Polonium atom

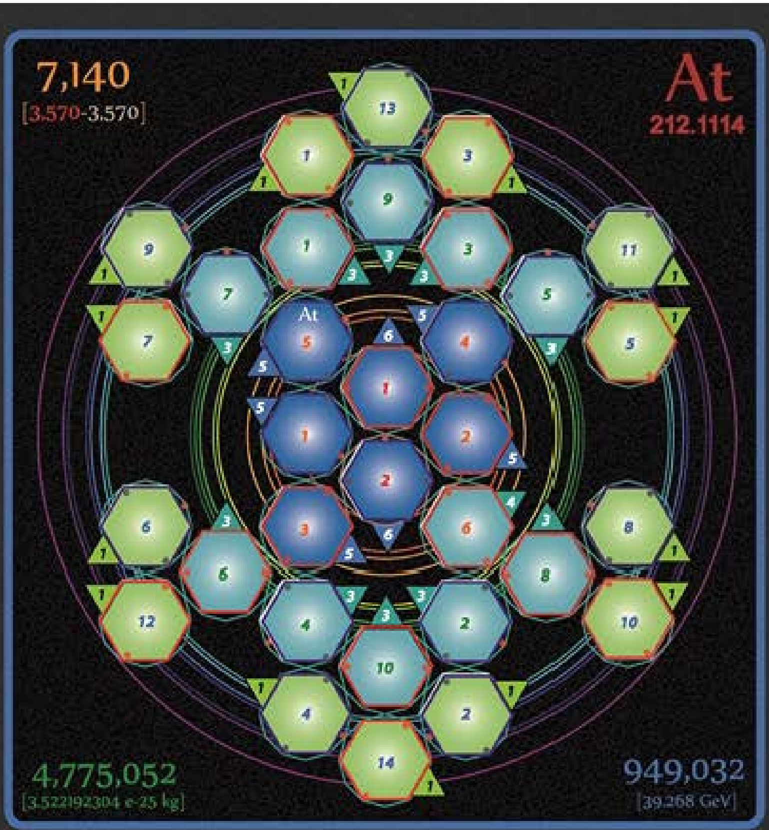
7,140
[3.570-3.570]
At
212.1114
4,775,052
[3.522192304 e-25 kg]
949,032
[39.268 GeV]

Astatine

85

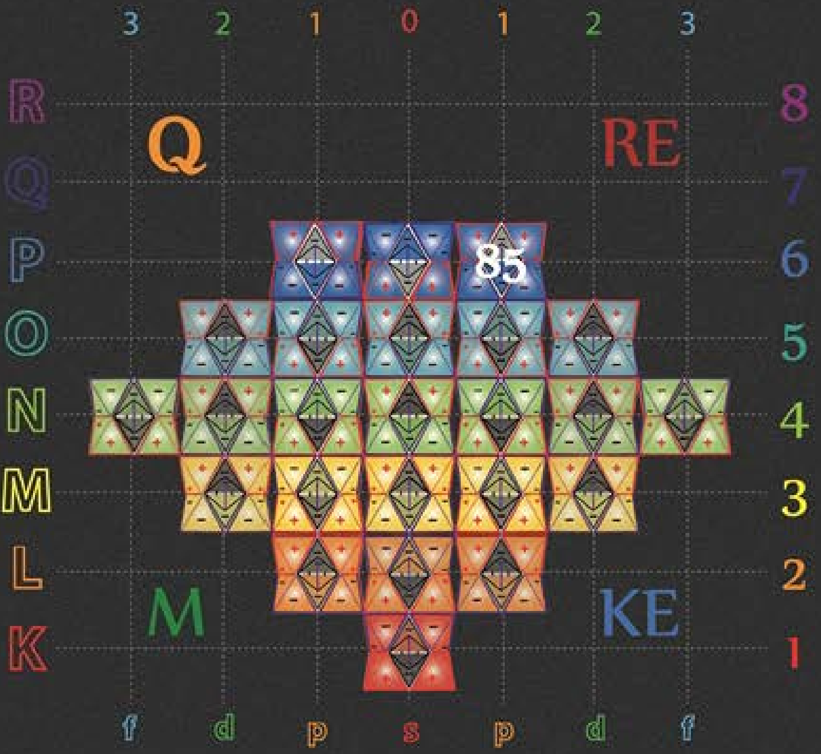
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
85
8 7 6 5 4 3 2 1
f d p s p d f

electron configuration

Aufbau

[Xe] 4f14 5d10 6s2 6p5
54

85 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Halogen

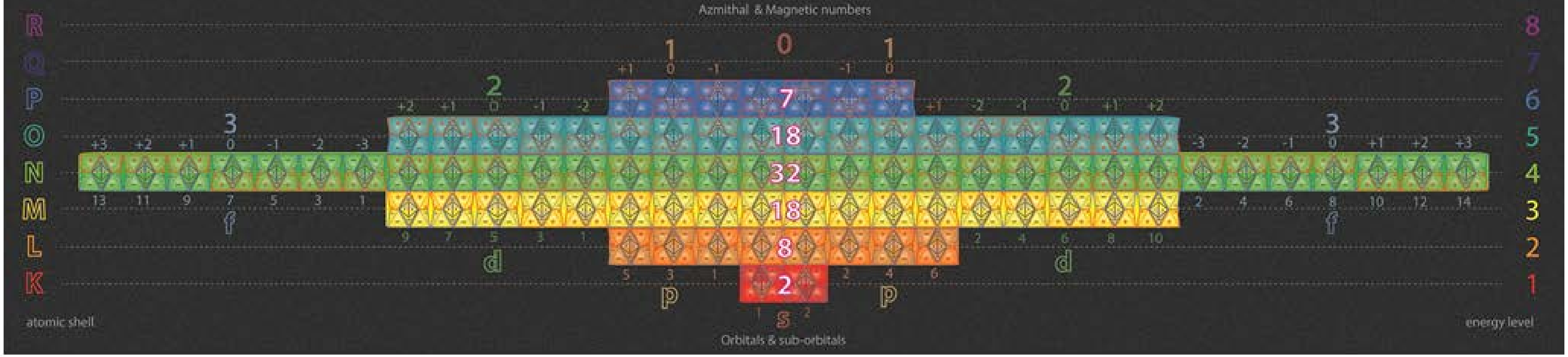
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
7
18
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.85 - Astatine atom

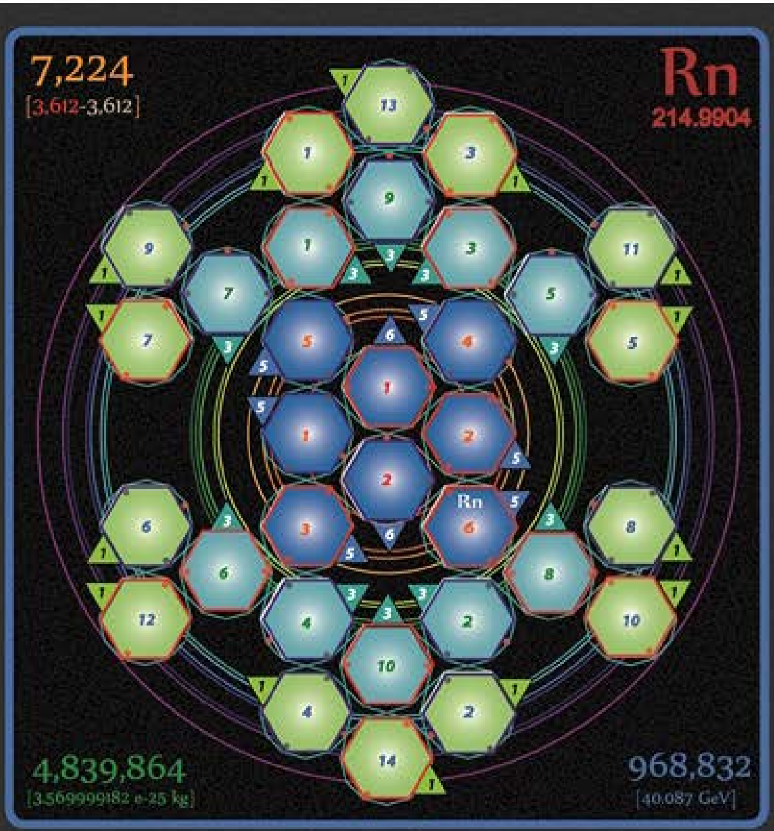
7,224
[3,612-3,612]
Rn
214.9904
Rn
4,839,864
[3.569999182 e-25 kg]
968,832
[40.087 GeV]

Radon

86

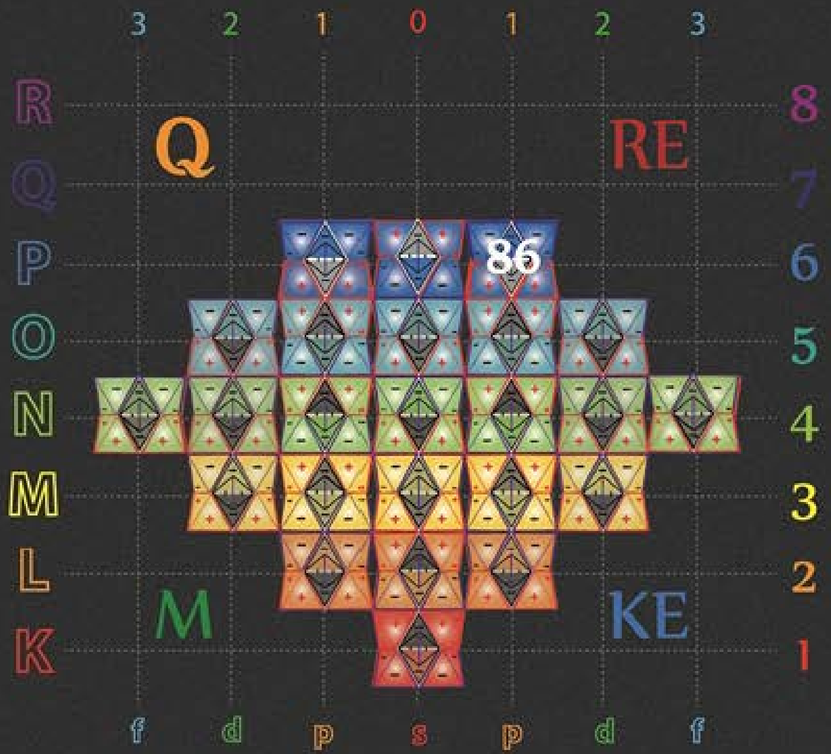
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
86
f d p s p d f

electron configuration

Aufbau

[Xe] 4f14 5d10 6s2 6p6
54

86 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Noble Gas

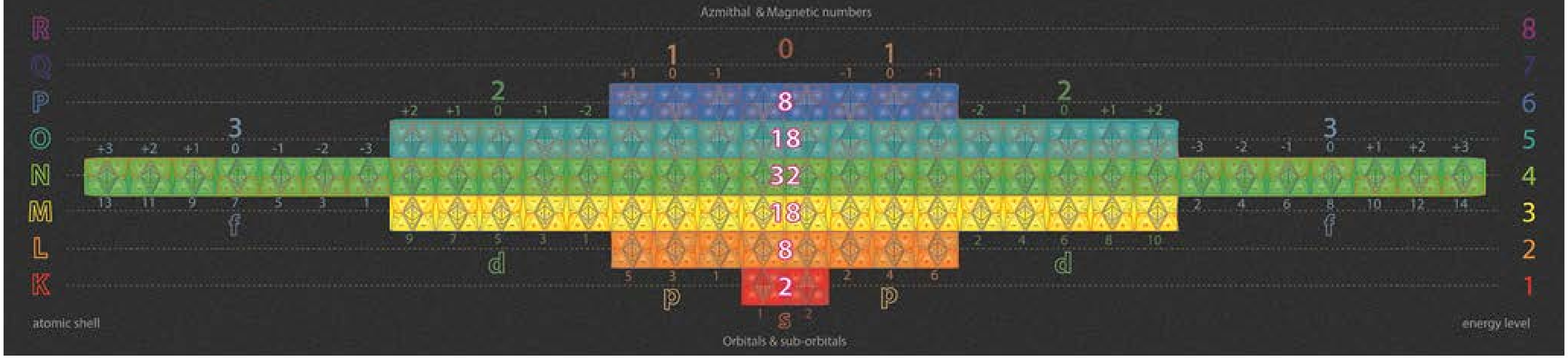
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1 0 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2
-2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
8
18
32
18
8
2
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f f
9 7 5 3 1
2 4 6 8 10
d d
5 3 1
2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.86 - Radon atom

Copyright ABRAHAM [2008] - All rights reserved

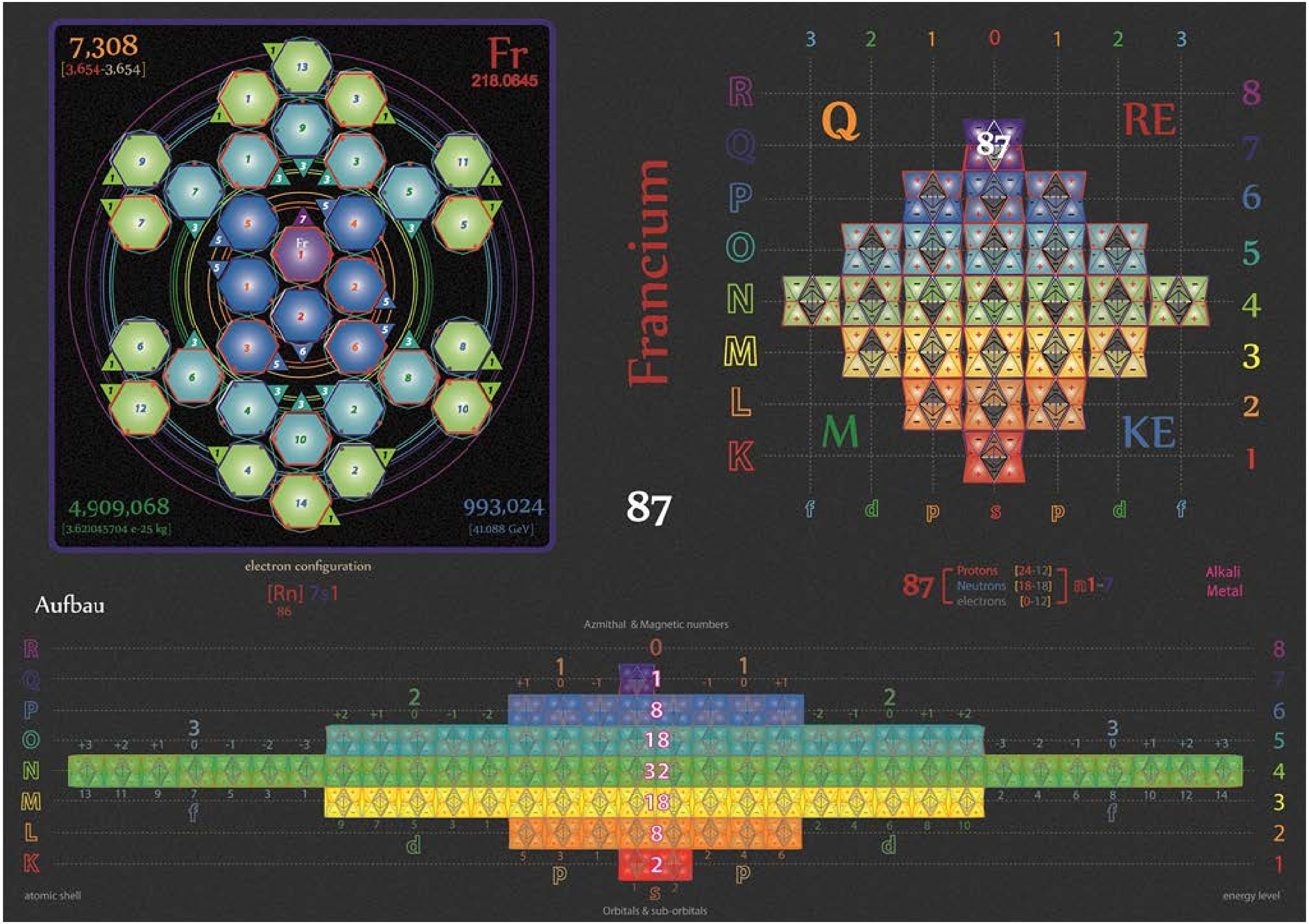
7,308
[3,654-3,654]
Fr
218.0845
4,909,068
[3.62045704 e-25 kg]
993,024
[41.088 GeV]
Francium
87
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Rn] 7s1
86
87 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Alkali Metal
Aufbau
Azmithal & Magnetic numbers
2
8
18
32
18
8
2
1
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.87 - Francium atom

Copyright ABRAHAM [2008] - All rights reserved

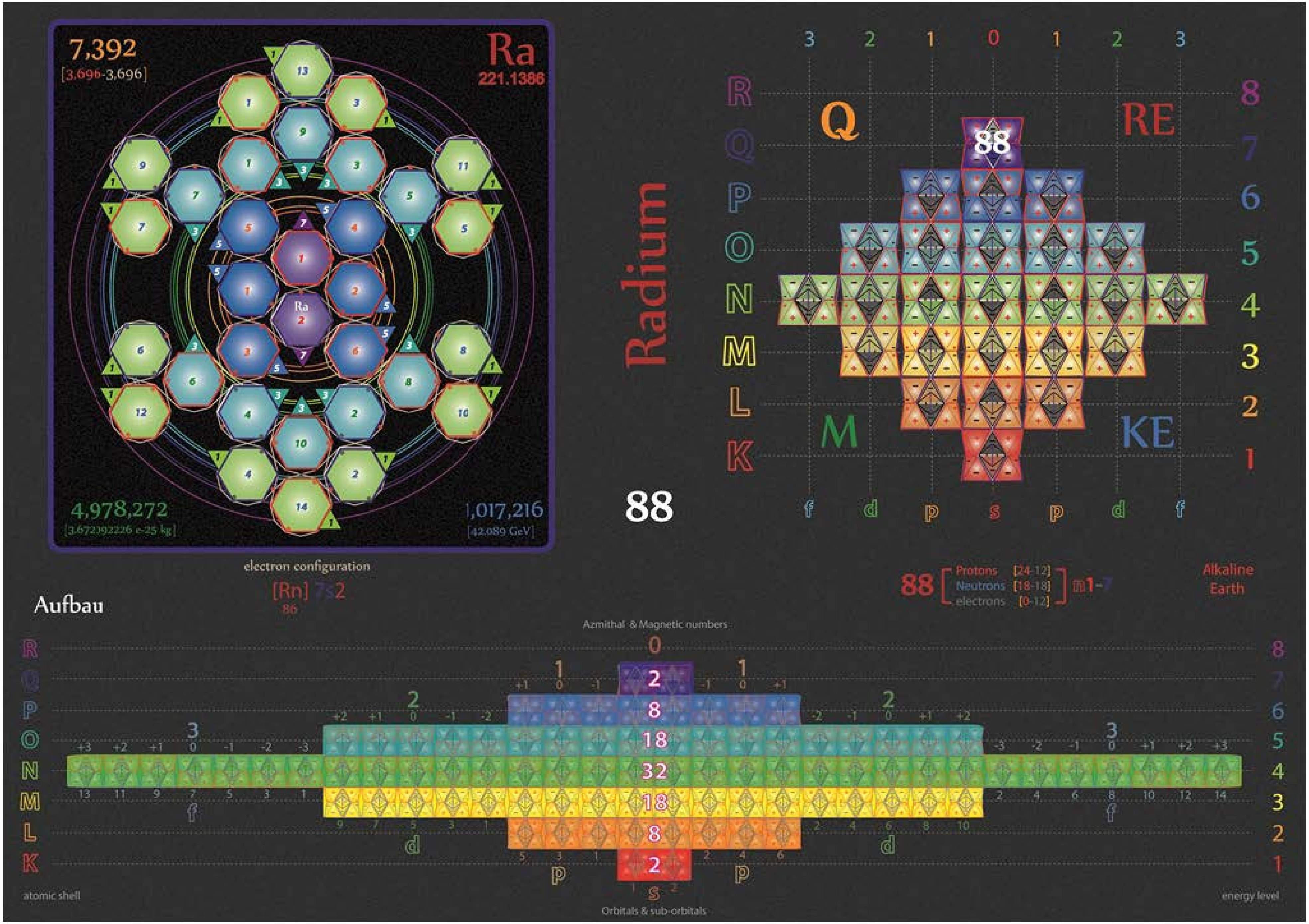
7,392
[3,696-3,696]
Ra
221.1386
4,978,272
[3.67292226 e-25 kg]
1,017,216
[42,089 GeV]
Radium
88
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Rn] 7s2
86
Aufbau
88 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Alkaline Earth
Azmithal & Magnetic numbers
0
2
8
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.88 - Radium atom

Copyright ABRAHAM [2008] - All rights reserved

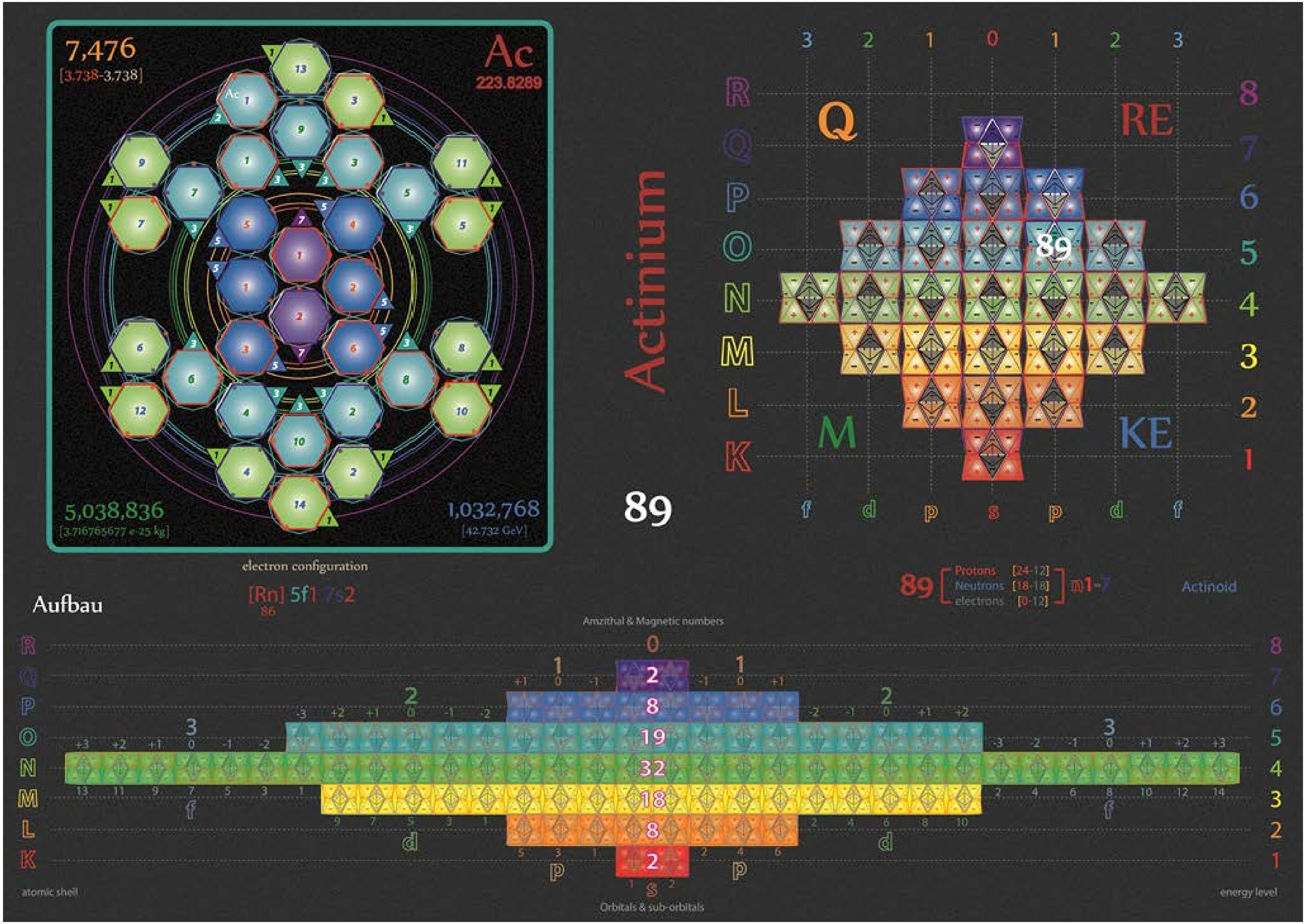
7,476
[3.738-3.738]
Ac
223.8289
5,038,836
[3.716765677 e-25 kg]
1,032,768
[42.732 GeV]
Actinium
89
electron configuration
[Rn] 5f1 7s2
86
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
3 2 1 0 1 2 3
f d p s p d f
89
89
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Amzithal & Magnetic numbers
0
2
8
19
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.89 - Actinium atom

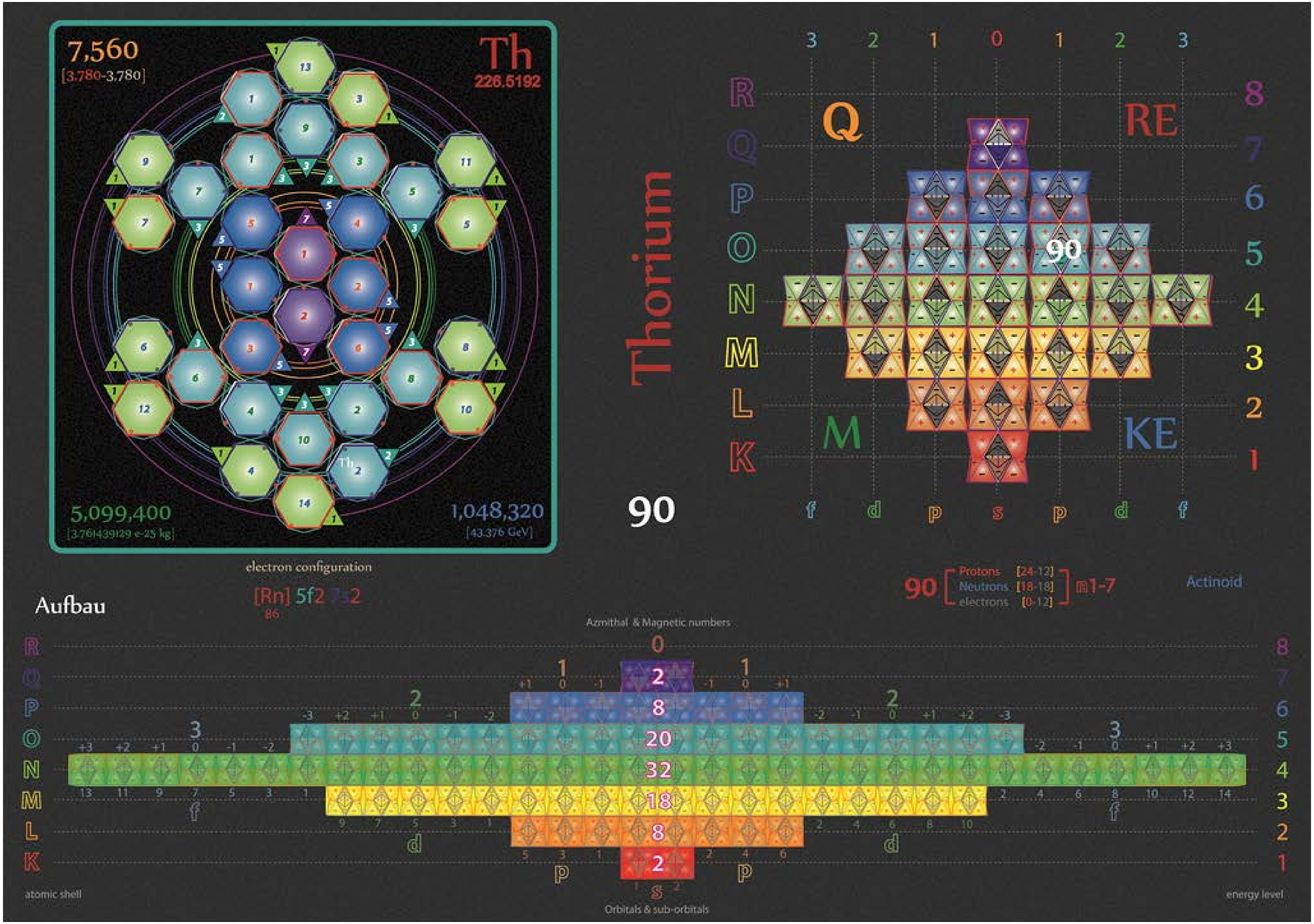
7,560
[3,780-3,780]
Th
226.5192
5,099,400
[3.761439129 e-25 kg]
1,048,320
[43.376 GeV]
Thorium
90
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Rn] 5f2 7s2
86
90 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
Azmithal & Magnetic numbers
2
8
20
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.90 - Thorium atom

Copyright ABRAHAM [2008] - All rights reserved

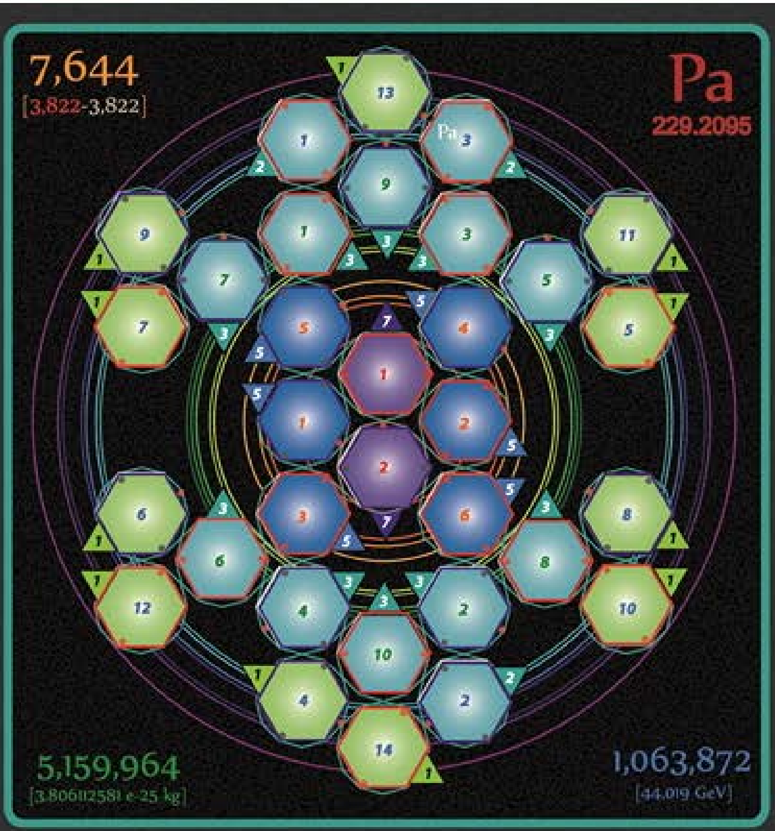
7,644
[3,822-3,822]
Pa
229.2095
5,159,964
[3.806112581 e-25 kg]
1,063,872
[44.019 GeV]

Protactinium

91

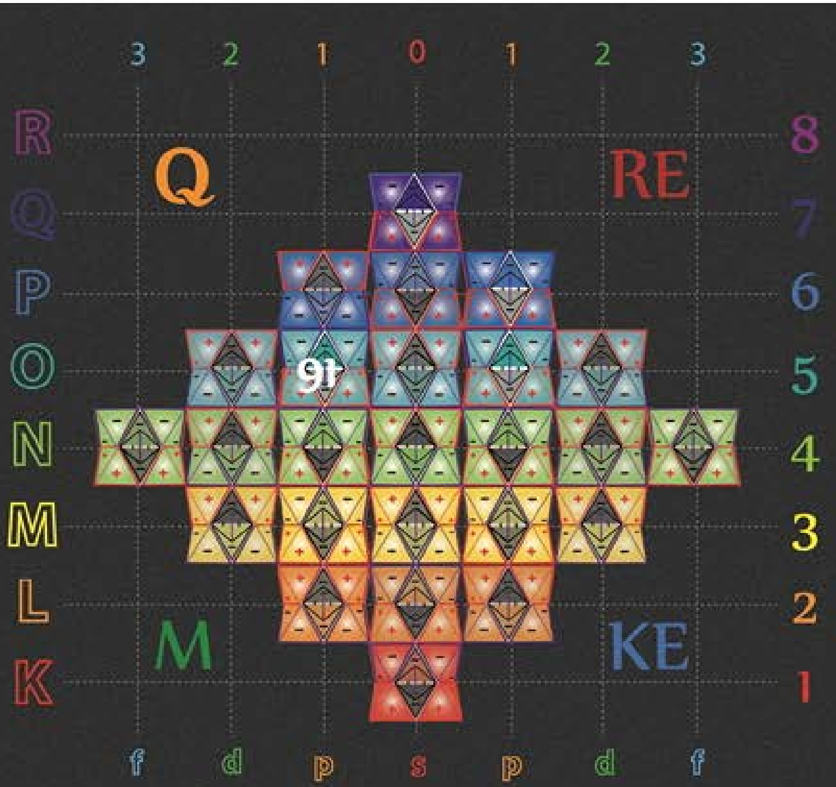
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
91
f
d
p
s
p
d
f

91 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-7

Actinoid

Aufbau

electron configuration

[Rn] 5f3 7s2
86

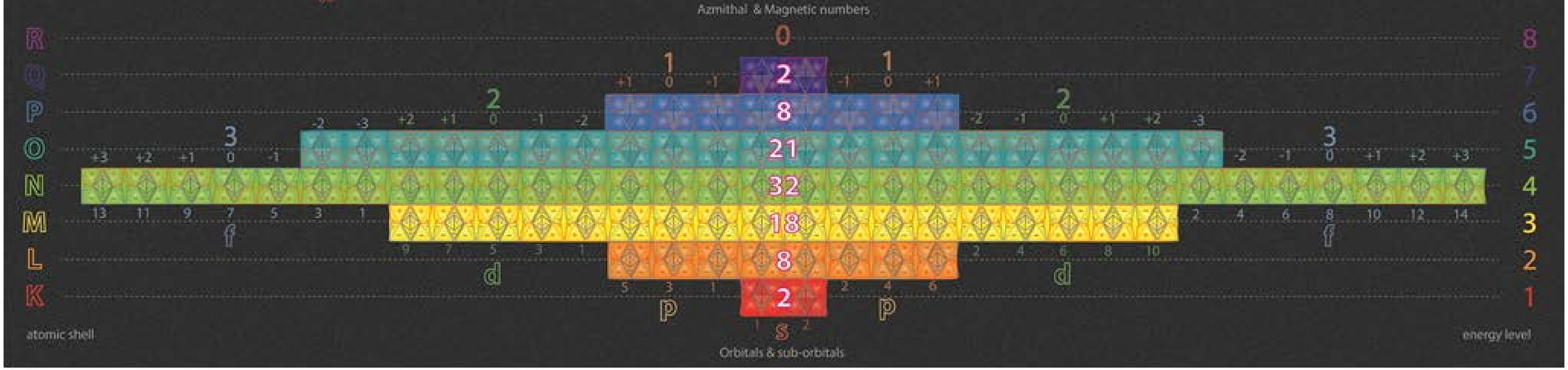
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
2
8
21
32
18
8
2
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.91 - Protactinium atom

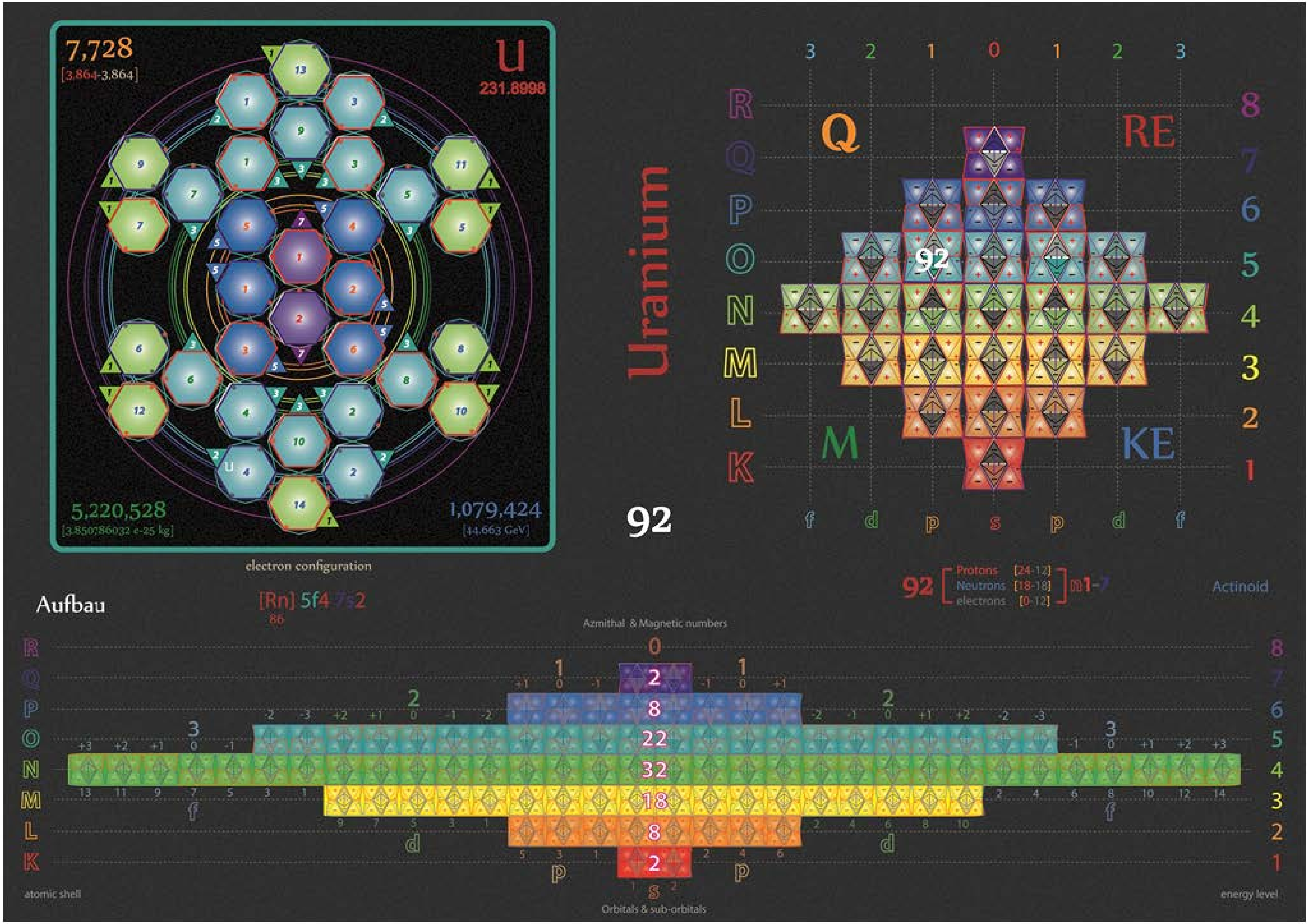
7,728
[3,864-3,864]
U
231.8998
5,220,528
[3.850786032 e-25 kg]
1,079,424
[14.663 GeV]
electron configuration
Uranium
92
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
92 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Aufbau
[Rn] 5f4 7s2
86
Azimuthal & Magnetic numbers
0
1
2
3
2
8
22
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.92 - Uranium atom

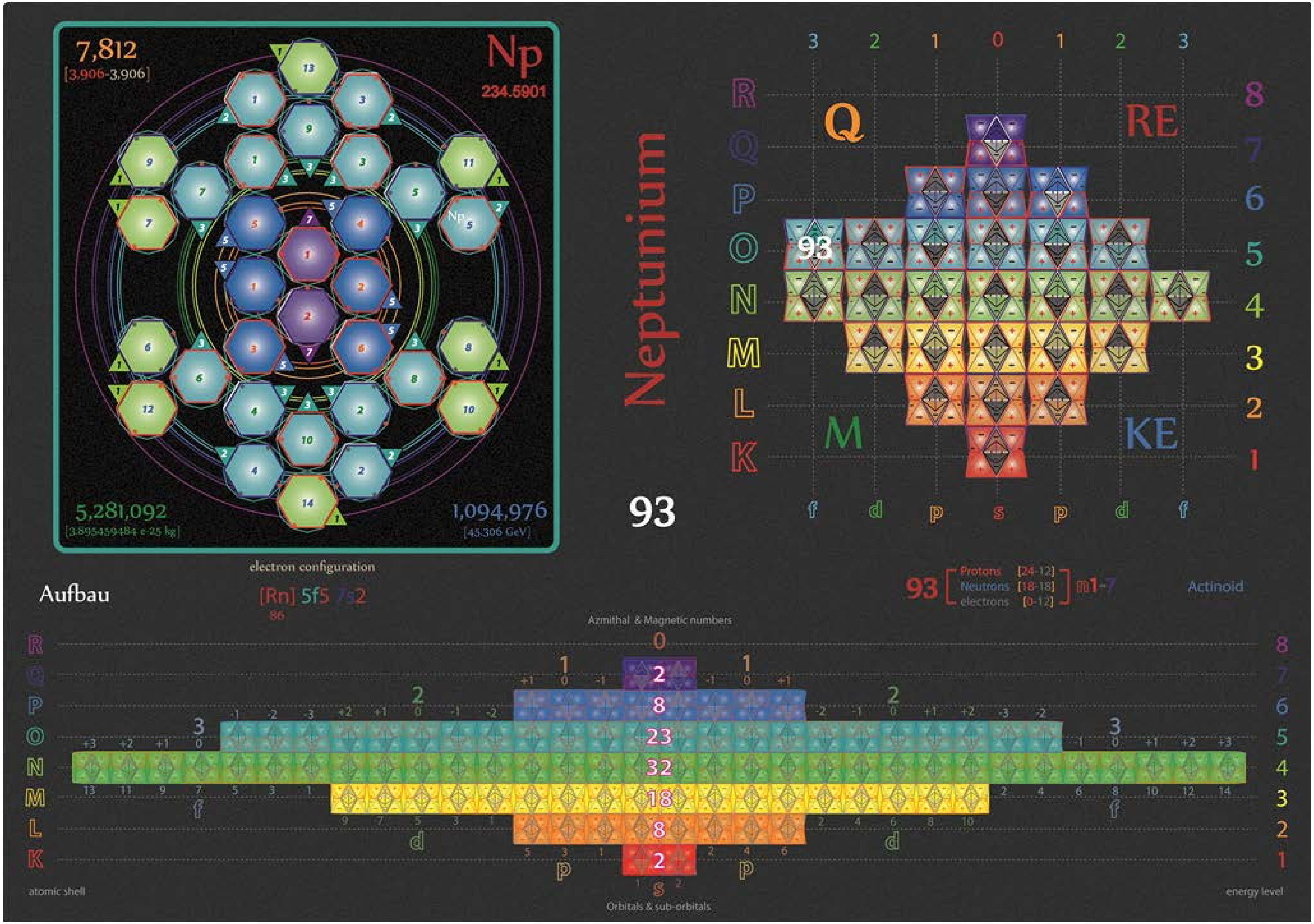
7,812
[3,906-3,906]
Np
234.5901
5,281,092
[3.895459484 e-25 kg]
1,094,976
[45.306 GeV]
Neptunium
93
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f5 7s2
86
93 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
0
2
8
23
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.93 - Neptunium atom

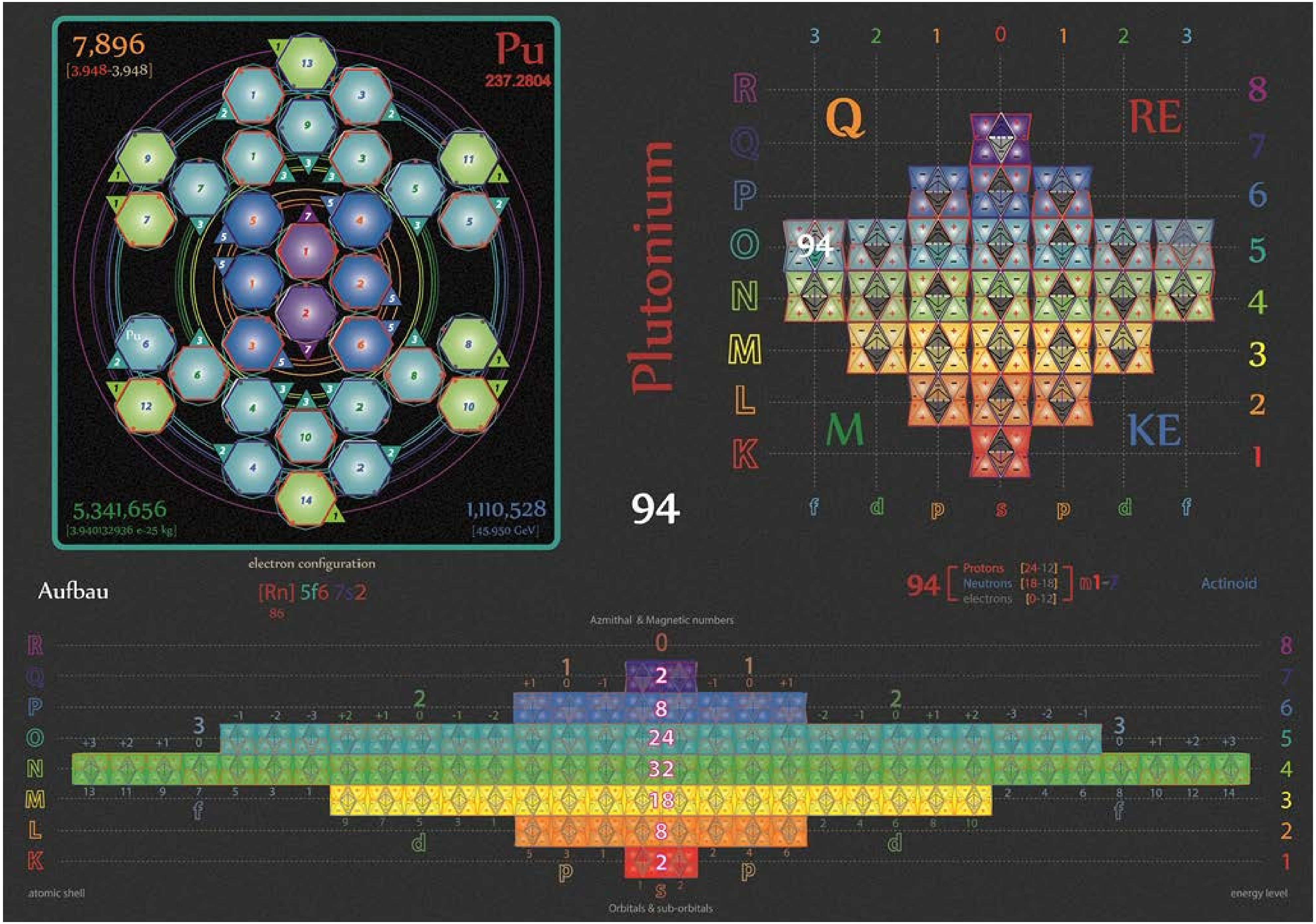
7,896
[3,948-3,948]
Pu
237.2804
5,341,656
[3.940132936 e-25 kg]
1,110,528
[45.950 GeV]
Plutonium
94
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
[Rn] 5f6 7s2
86
Aufbau
94 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
24
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.94 - Plutonium atom

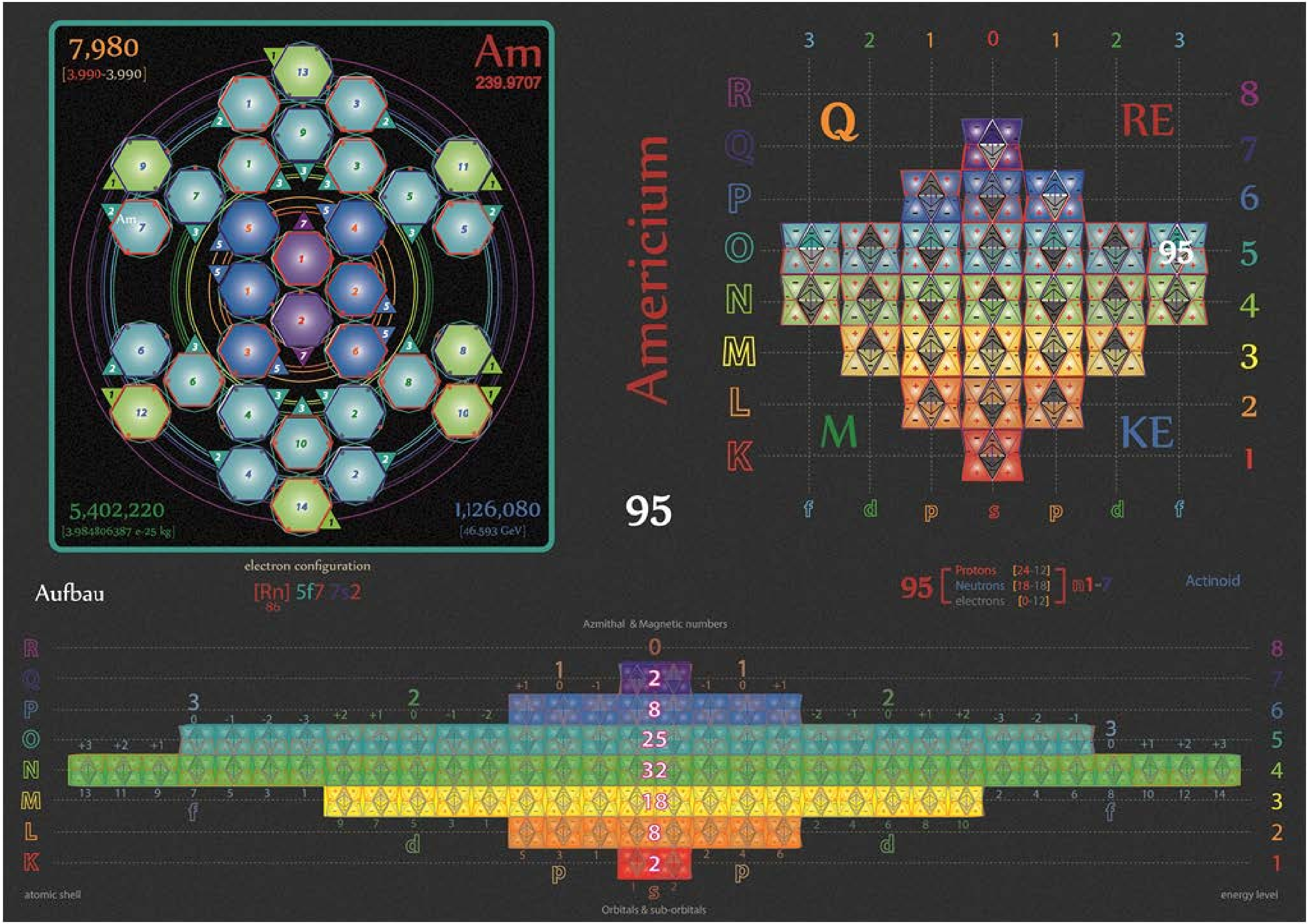
7,980
[3,990-3,990]
Am
239.9707
5,402,220
[3.984806387 e-25 kg]
1,126,080
[46.593 GeV]
Americium
95
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
95
f d p s p d f
electron configuration
[Rn] 5f7 7s2
86
Aufbau
95 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
8
25
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.95 - Americium atom

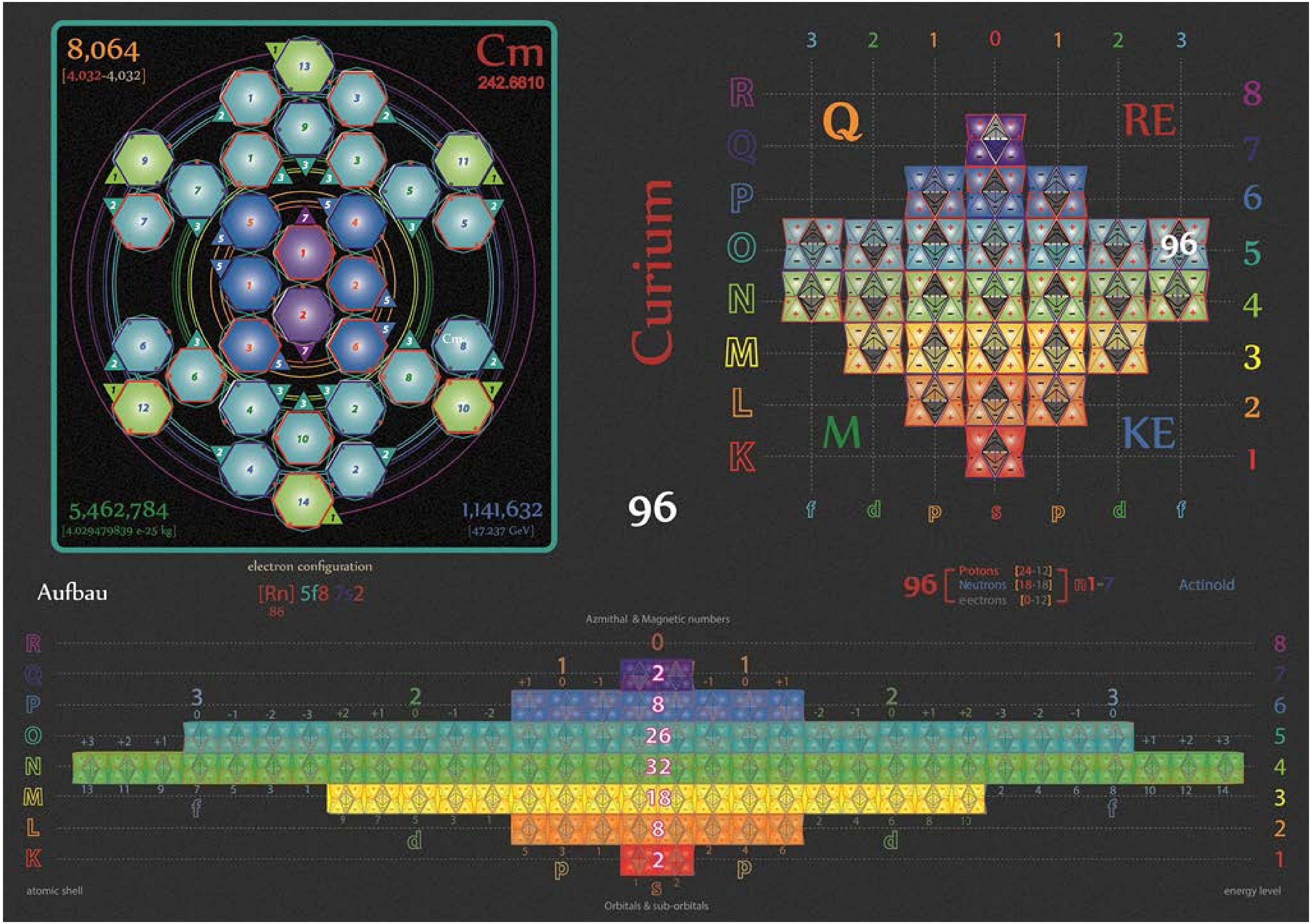
8,064
[4,032-4,032]
Cm
242.6810
5,462,784
[4.029479839 e-25 kg]
1,141,632
[47.237 GeV]
Curium
96
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f8 7s2
86
96 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
26
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.96 - Curium atom

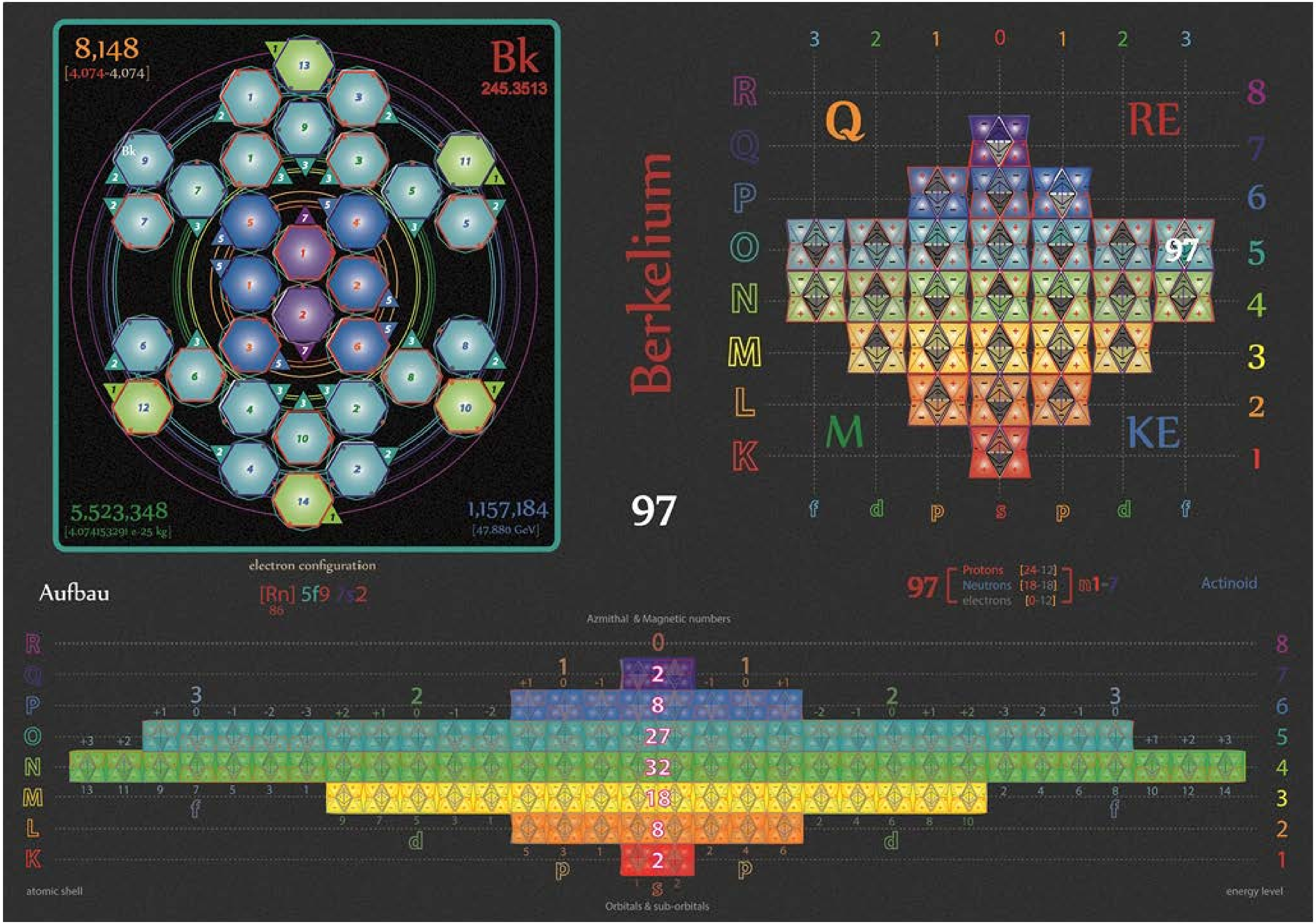
8,148
[4,074-4,074]
Bk
245.3513
5,523,348
[4.07415329l e-25 kg]
1,157,184
[47,880 GeV]
Berkelium
97
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f9 7s2
86
97 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
27
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.97 - Berkelium atom

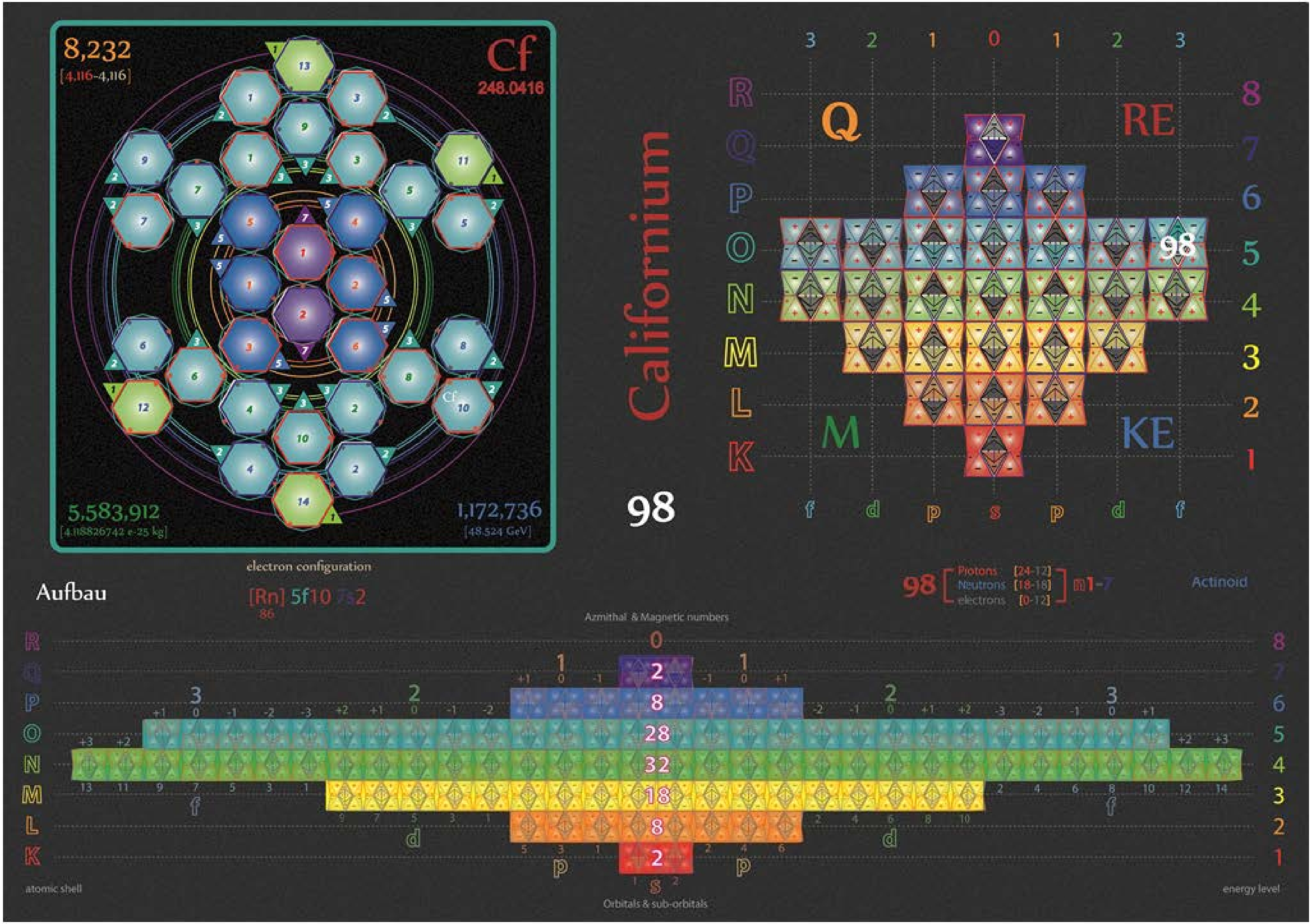
8,232
[4,116-4,116]
Cf
248.0416
5,583,912
[4.118826742 e-25 kg]
1,172,736
[48.524 GeV]
Californium
98
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f10 7s2
86
98 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
28
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.98 - Californium atom

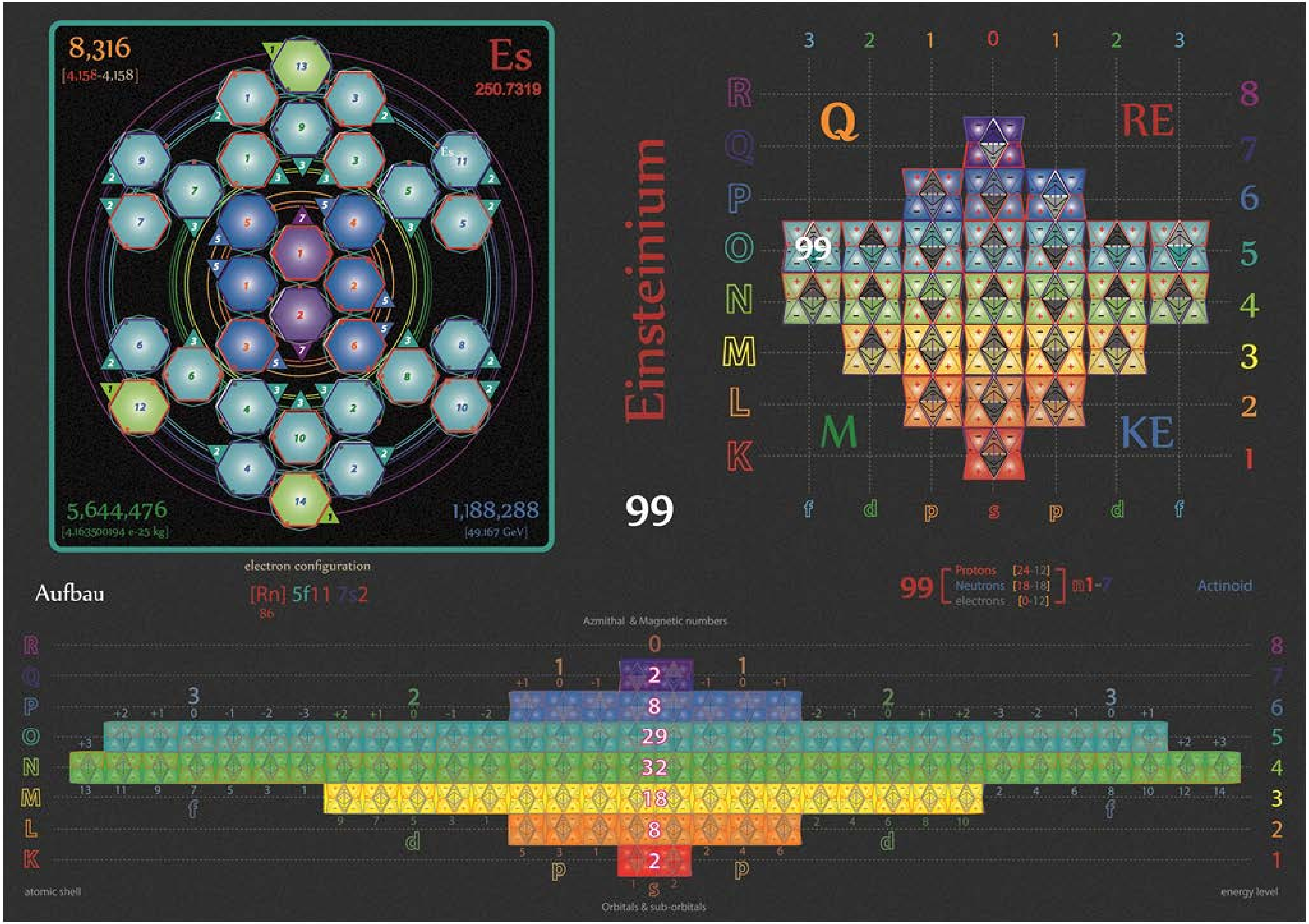
8,316
[4,158-4,158]
Es
250.7319
5,644,476
[4.163500194 e-25 kg]
1,188,288
[49.167 GeV]
Einsteinium
99
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f11 7s2
86
99 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
0
2
8
29
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.99 - Einsteinium atom

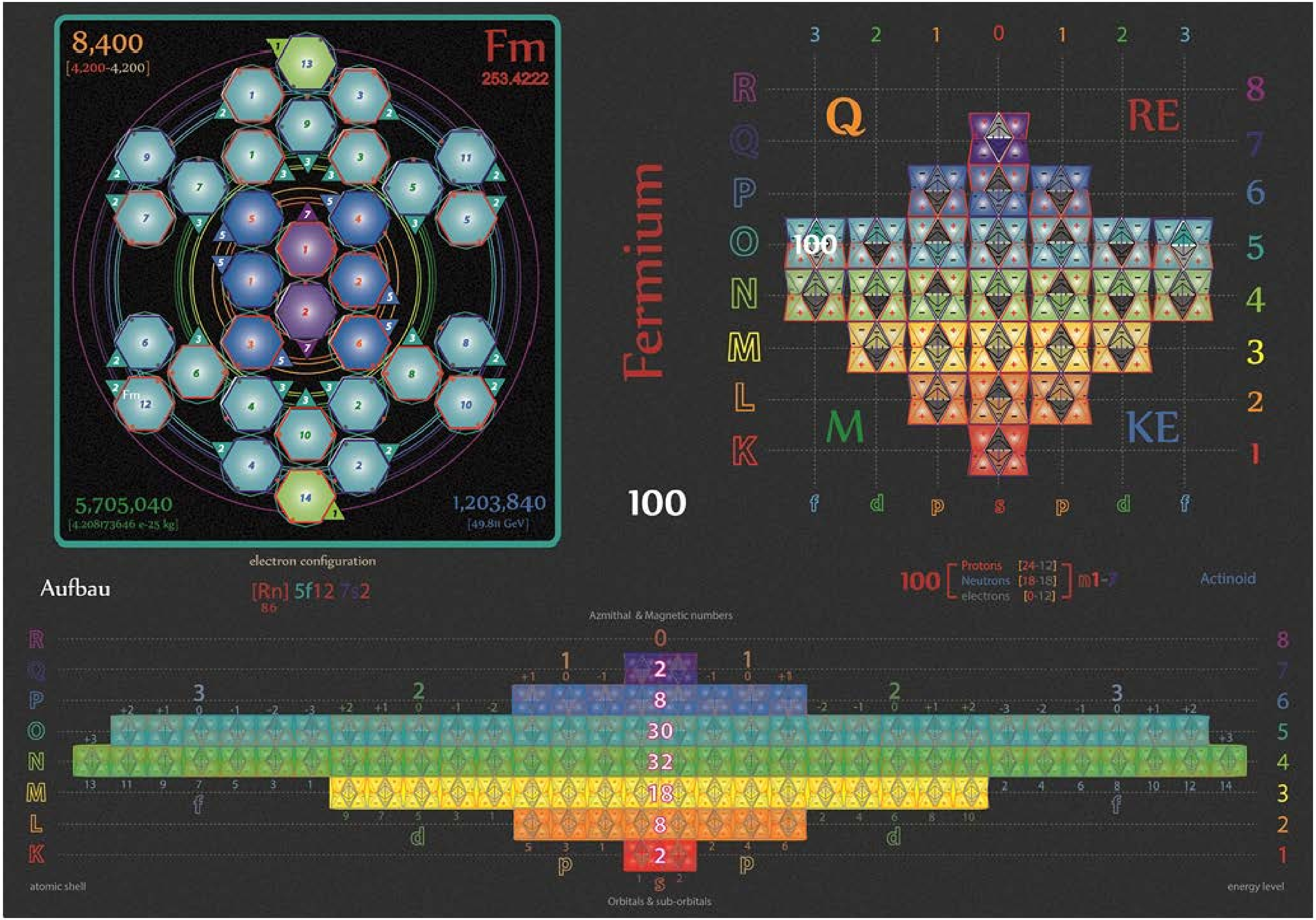
8,400
[4,200-4,200]
Fm
253.4222
5,705,040
[4.208173646 e-25 kg]
1,203,840
[49.811 GeV]
Fermium
100
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f12 7s2
86
100
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
30
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.100 - Fermium atom

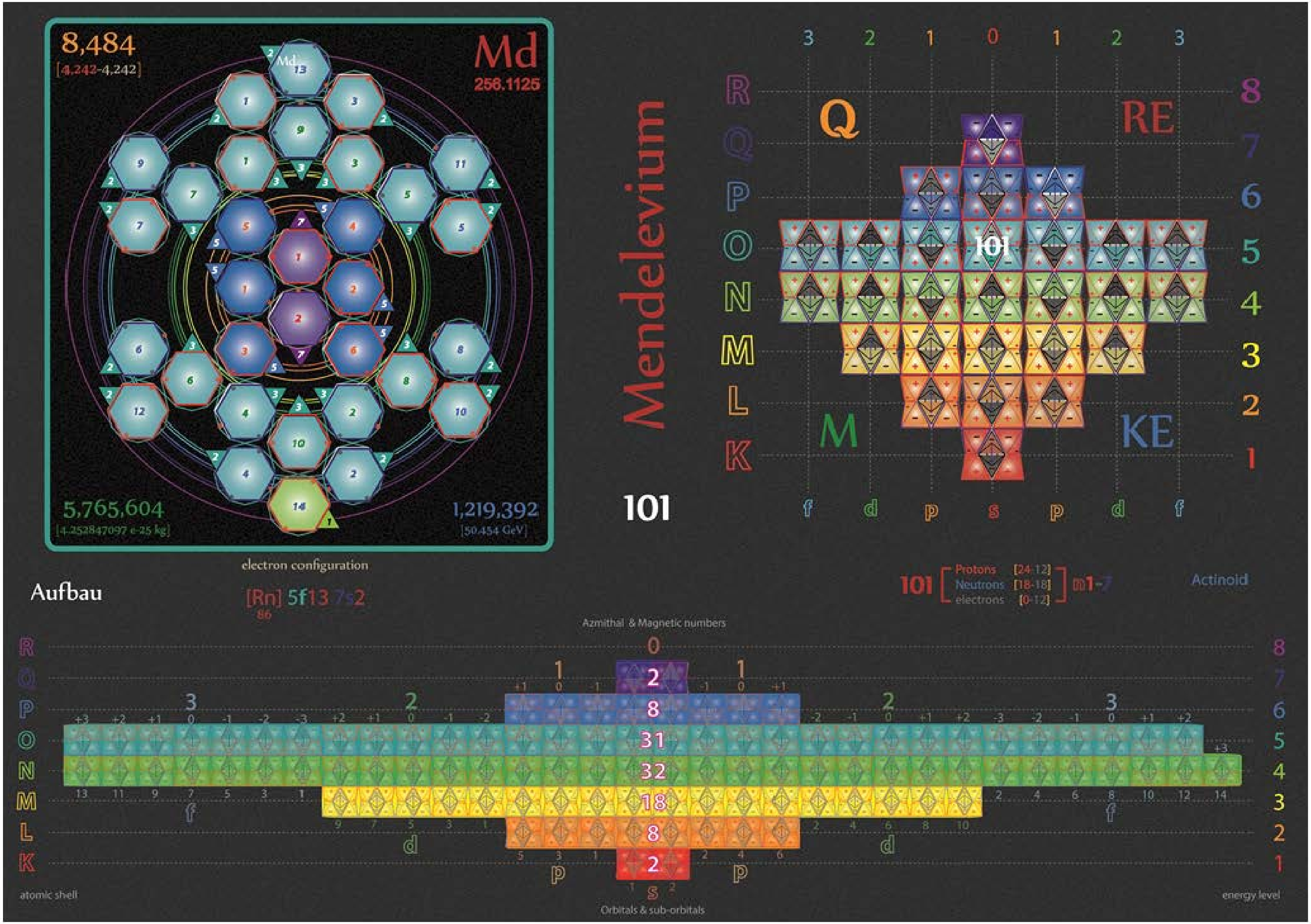
8,484
[4,242-4,242]
Md
256.1125
5,765,604
[4.252847097 e-25 kg]
1,219,392
[50.454 GeV]
Mendelevium
101
electron configuration
Aufbau
[Rn] 5f13 7s2
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
101 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-7
Actinoid
Azmithal & Magnetic numbers
0
2
8
31
32
18
8
2
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.101 - Mendelevium atom

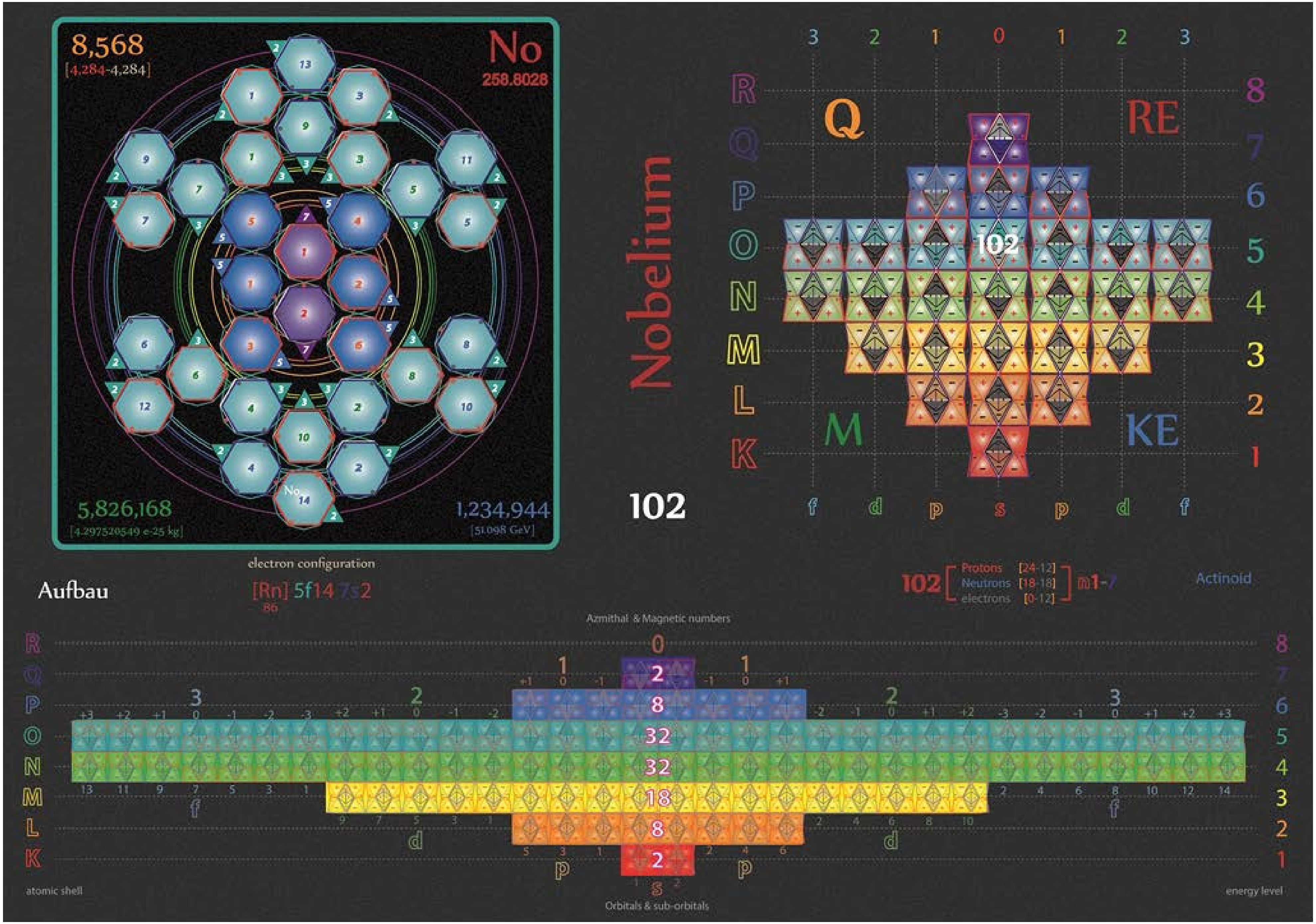
8,568
[4,284-4,284]
No
258.8028
5,826,168
[4.297520549 e-25 kg]
1,234,944
[51.098 GeV]
Nobelium
102
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 7s2
86
102 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.102 - Nobelium atom

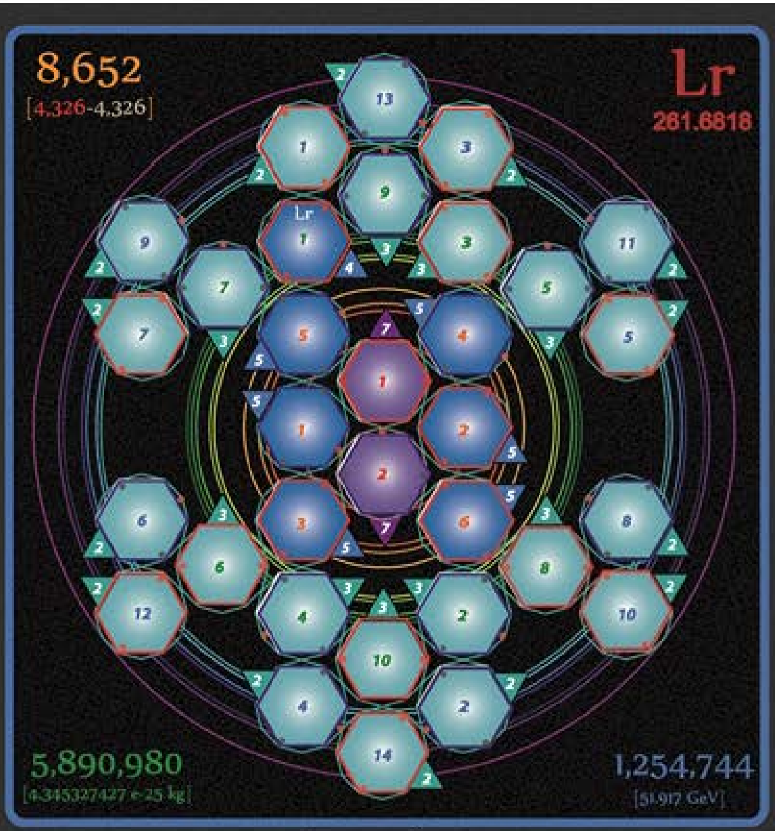
8,652
[4,326-4,326]
Lr
281.6818
5,890,980
[4.345327427 e-25 kg]
1,254,744
[51.917 GeV]

Lawrencium

103

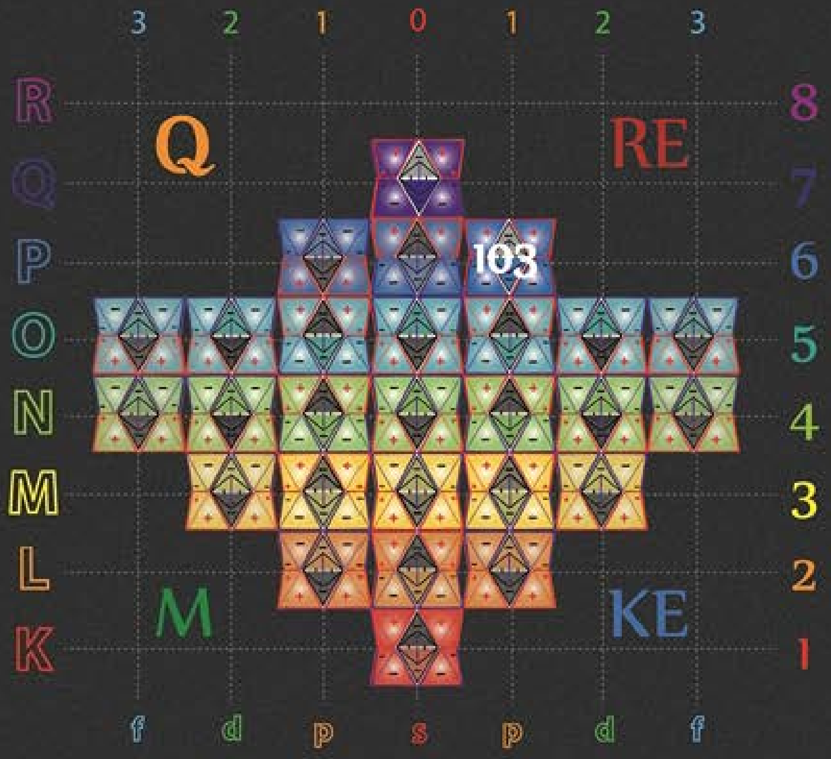
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
103
M
KE
f
d
p
s
p
d
f

electron configuration

Aufbau

[Rn] 5f14 6d1 7s2
86

103 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Actinoid

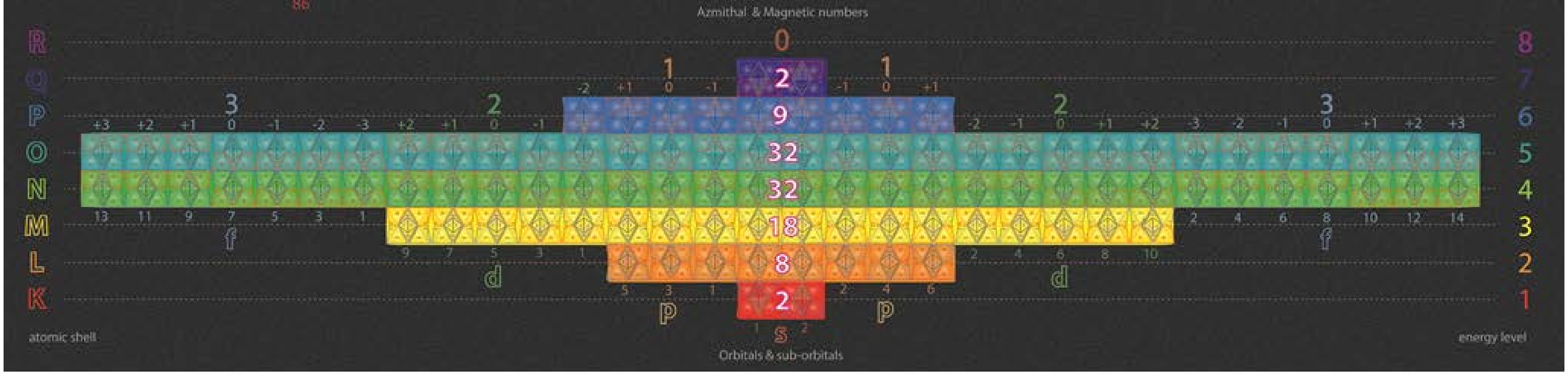
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
9
32
32
18
8
2
1
1
2
2
3
3
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

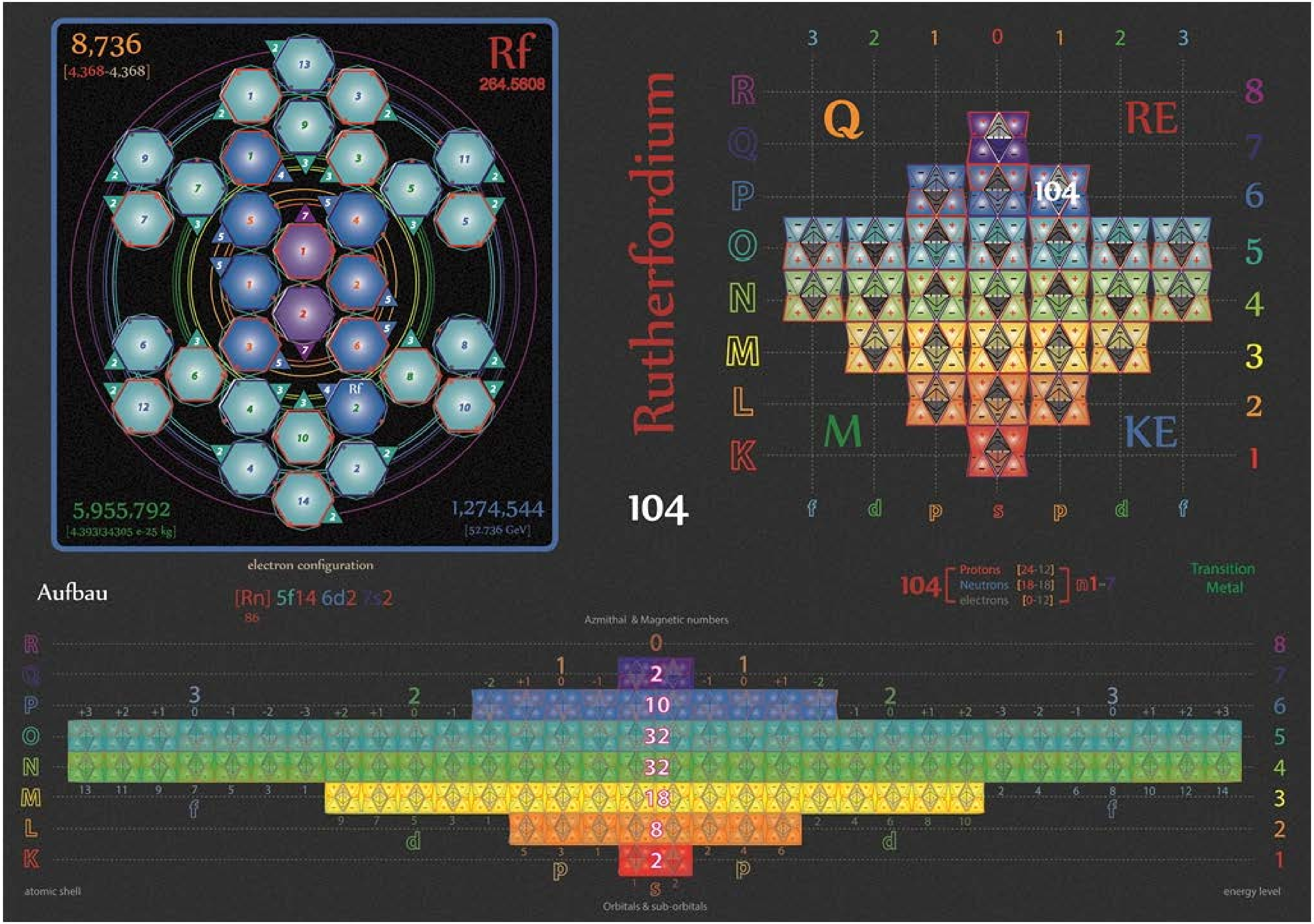
8,736
[4,368-4,368]
Rf
264.5608
5,955,792
[4.393134305 e-25 kg]
1,274,544
[52.736 GeV]
Rutherfordium
104
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d2 7s2
86
104 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Transition Metal
Azmithal & Magnetic numbers
2
10
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.104 - Rutherfordium atom

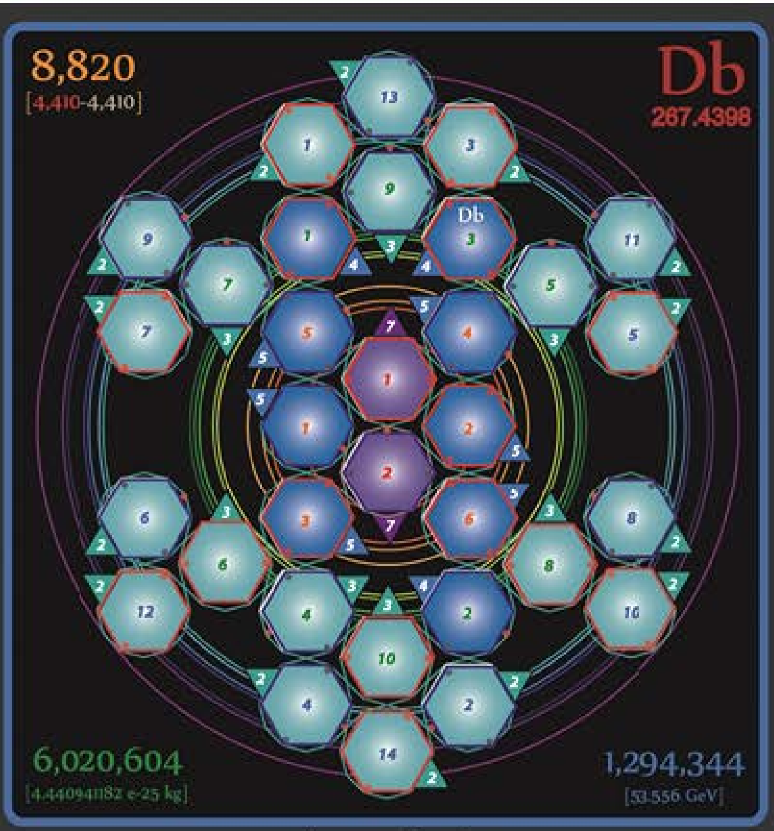
8,820
[4,410-4,410]
Db
287.4398
Db
6,020,604
[4.44094182 e-25 kg]
1,294,344
[53.556 GeV]

Dubnium

105

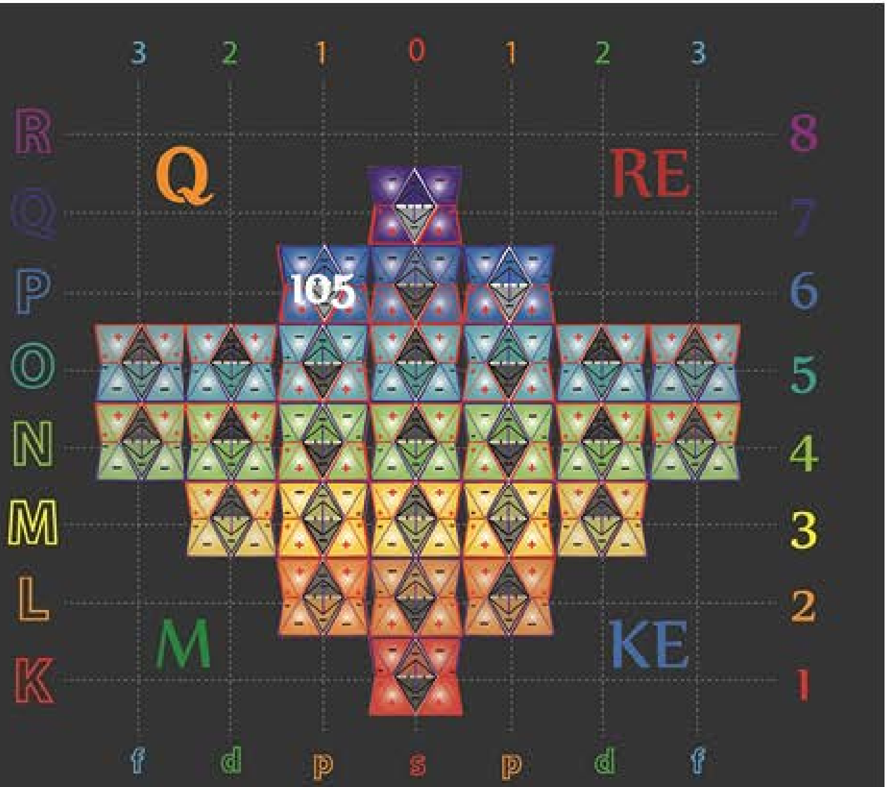
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
105
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Aufbau

electron configuration

[Rn] 5f14 6d3 7s2
86

105 [Protons [24-12] Neutrons [18-16] electrons [0-12]] m1-7

Transition Metal

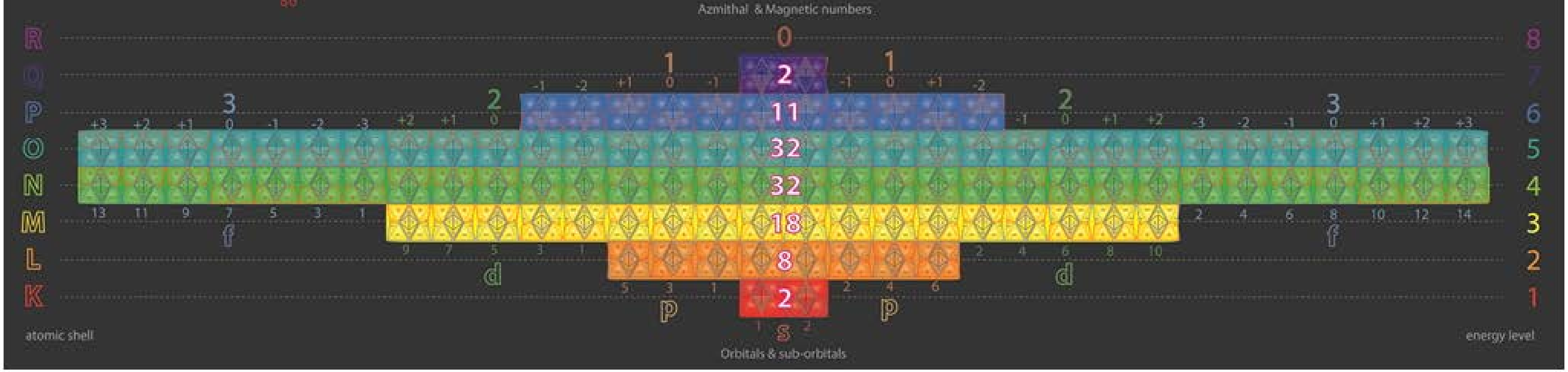
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
3
3
2
11
32
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Copyright ABRAHAM [2008] - All rights reserved

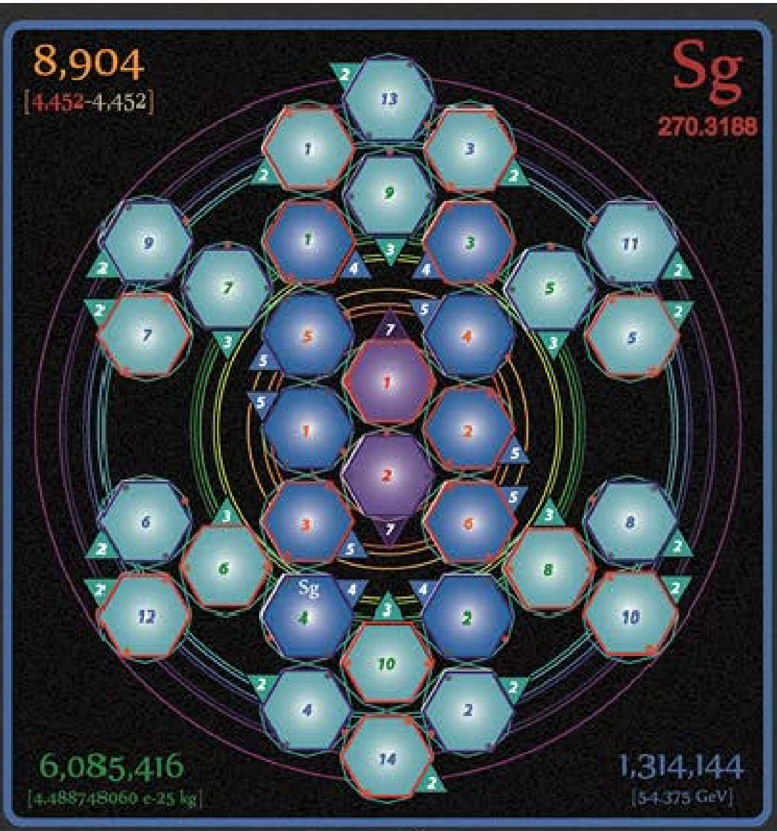
8,904
[4,452-4,452]
Sg
270.3188
6,085,416
[4.488748060 e-25 kg]
1,314,144
[54.375 GeV]

Seaborgium

106

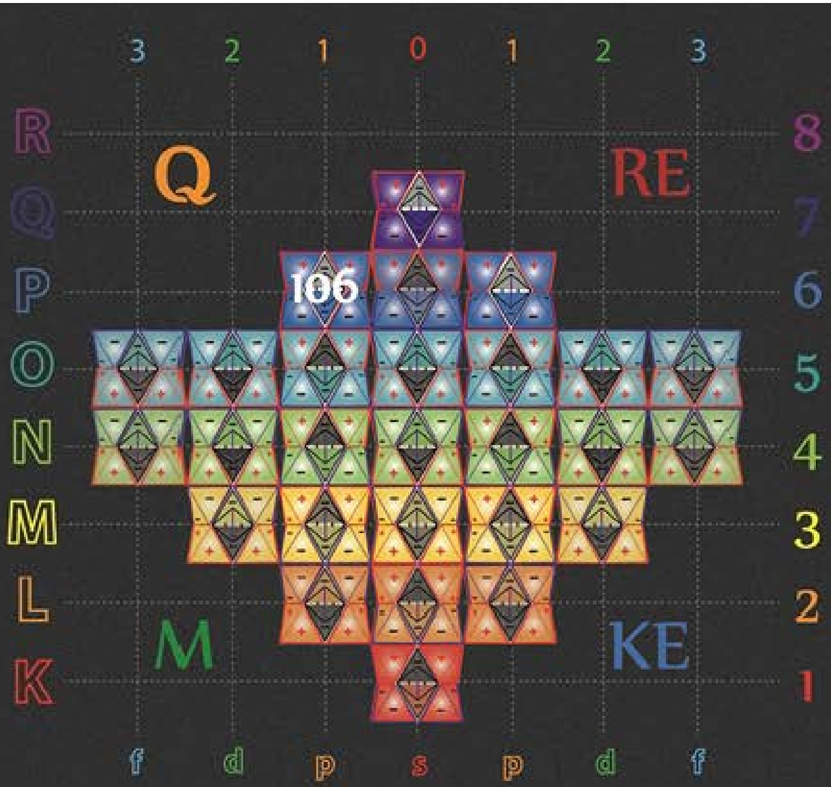
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
106
f d p s p d f

electron configuration

[Rn] 5f14 6d4 7s2
86

106 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Transition Metal

Aufbau

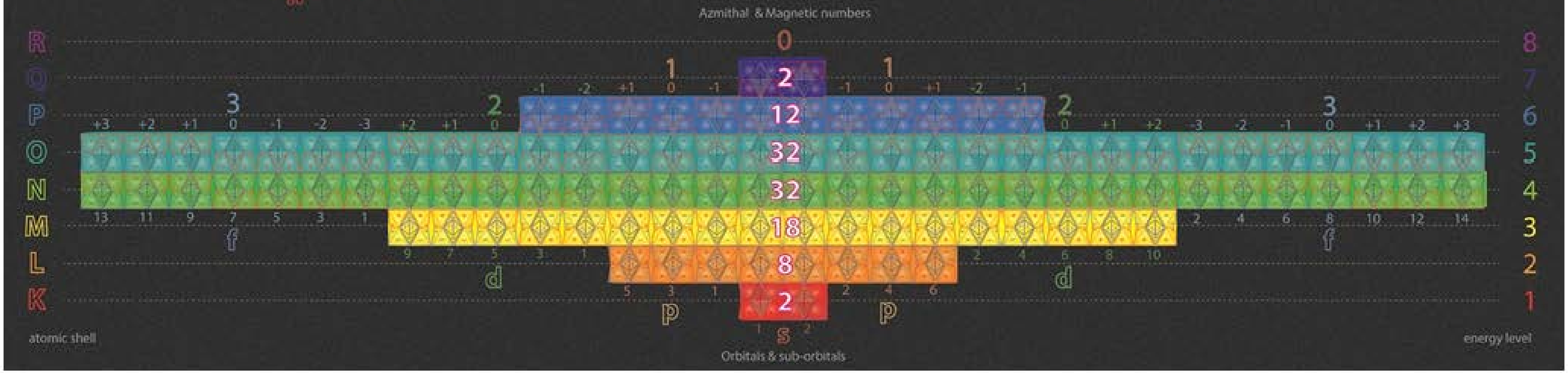
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
12
32
32
18
8
2
R Q P O N M L K
8 7 6 5 4 3 2 1
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.106 - Seaborgium atom

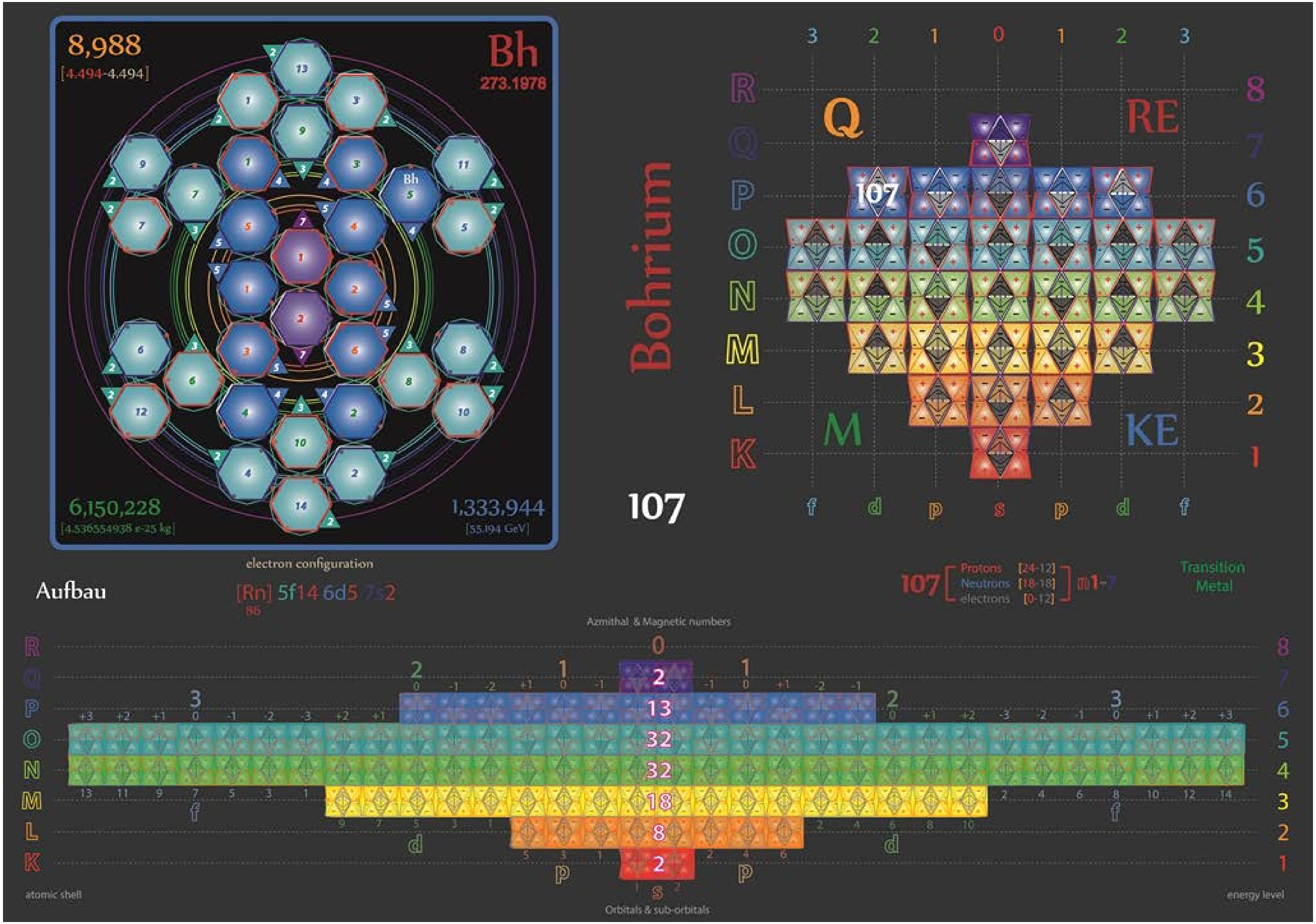
8,988
[4.494-4.494]
Bh
273.1978
6,150,228
[4.536554938 e-25 kg]
1,333,944
[55.194 GeV]
Bohrium
107
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d5 7s2
86
107 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
2 13 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.107 - Bohrium atom

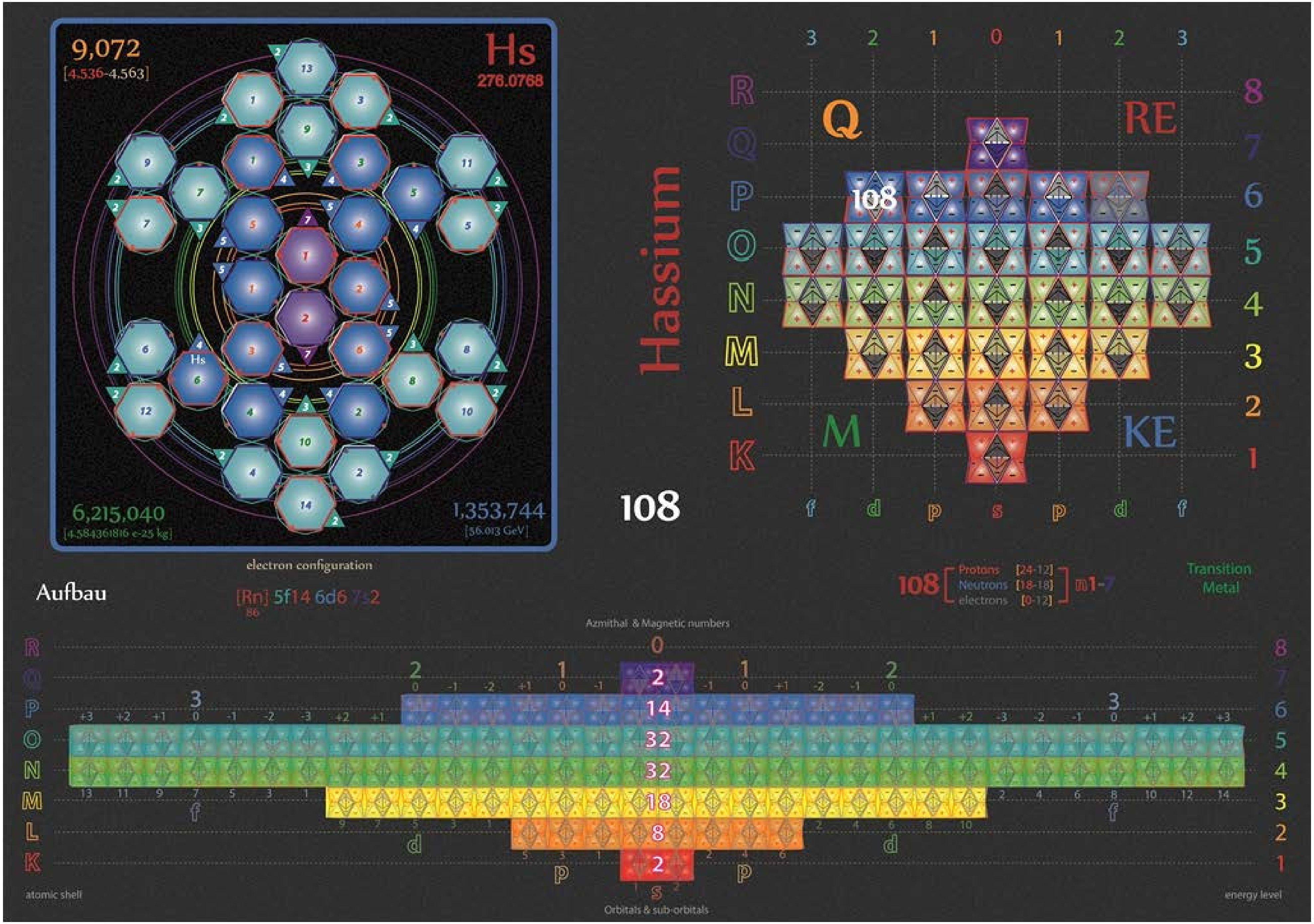
9,072
[4.536-4.563]
Hs
276.0768
6,215,040
[4.584361816 e-25 kg]
1,353,744
[56.013 GeV]
Hassium
108
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d6 7s2
86
108 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2 14 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.108 - Hassium atom

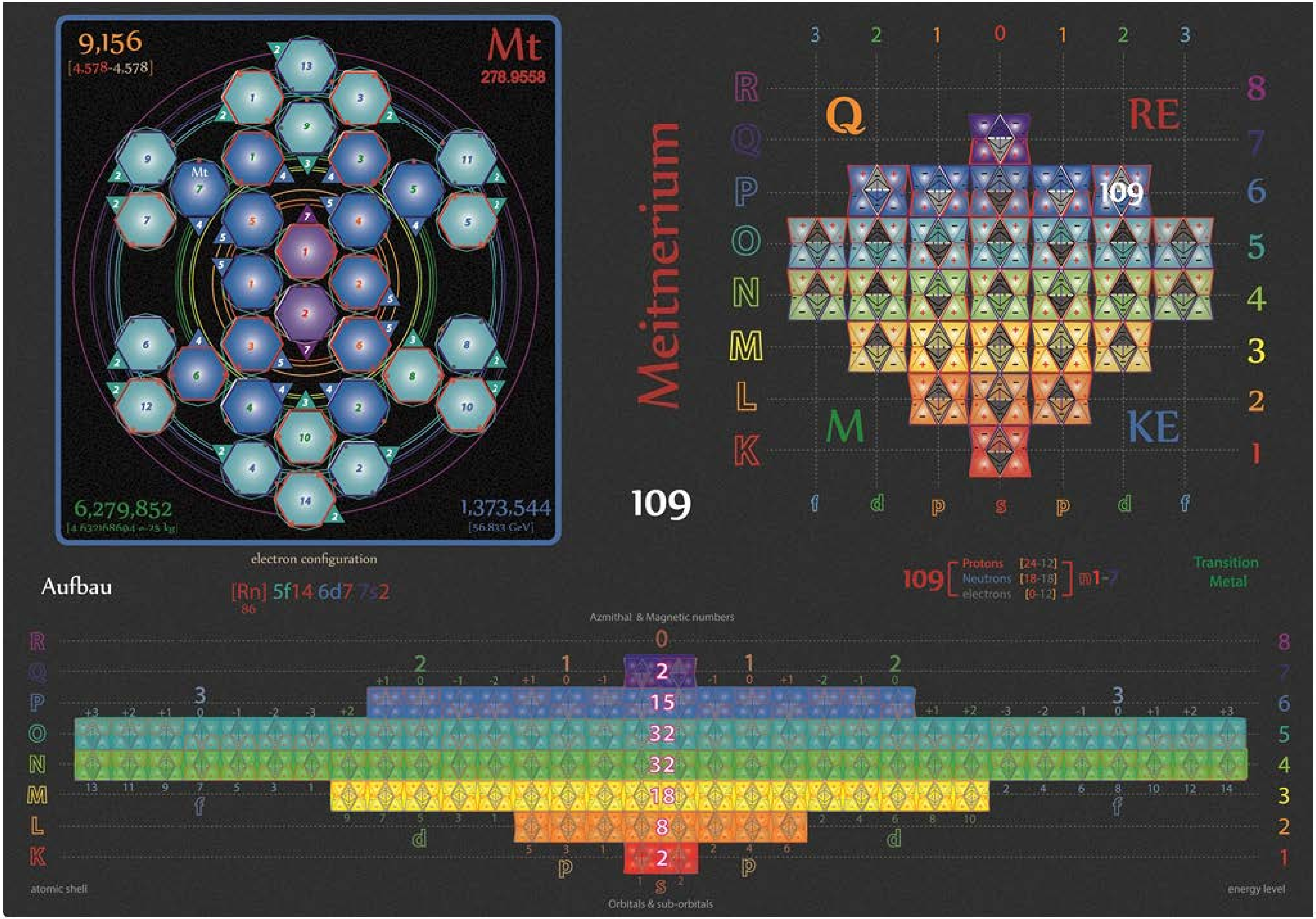
9,156
[4.578-4.578]
Mt
278.9558
6,279,852
[4.63216860 e-25 kg]
1,373,544
[56.833 GeV]
Meitnerium
109
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
109
electron configuration
Aufbau
[Rn] 5f14 6d7 7s2
86
109 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Transition Metal
Azmithal & Magnetic numbers
2 15 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.109 - Meitnerium atom

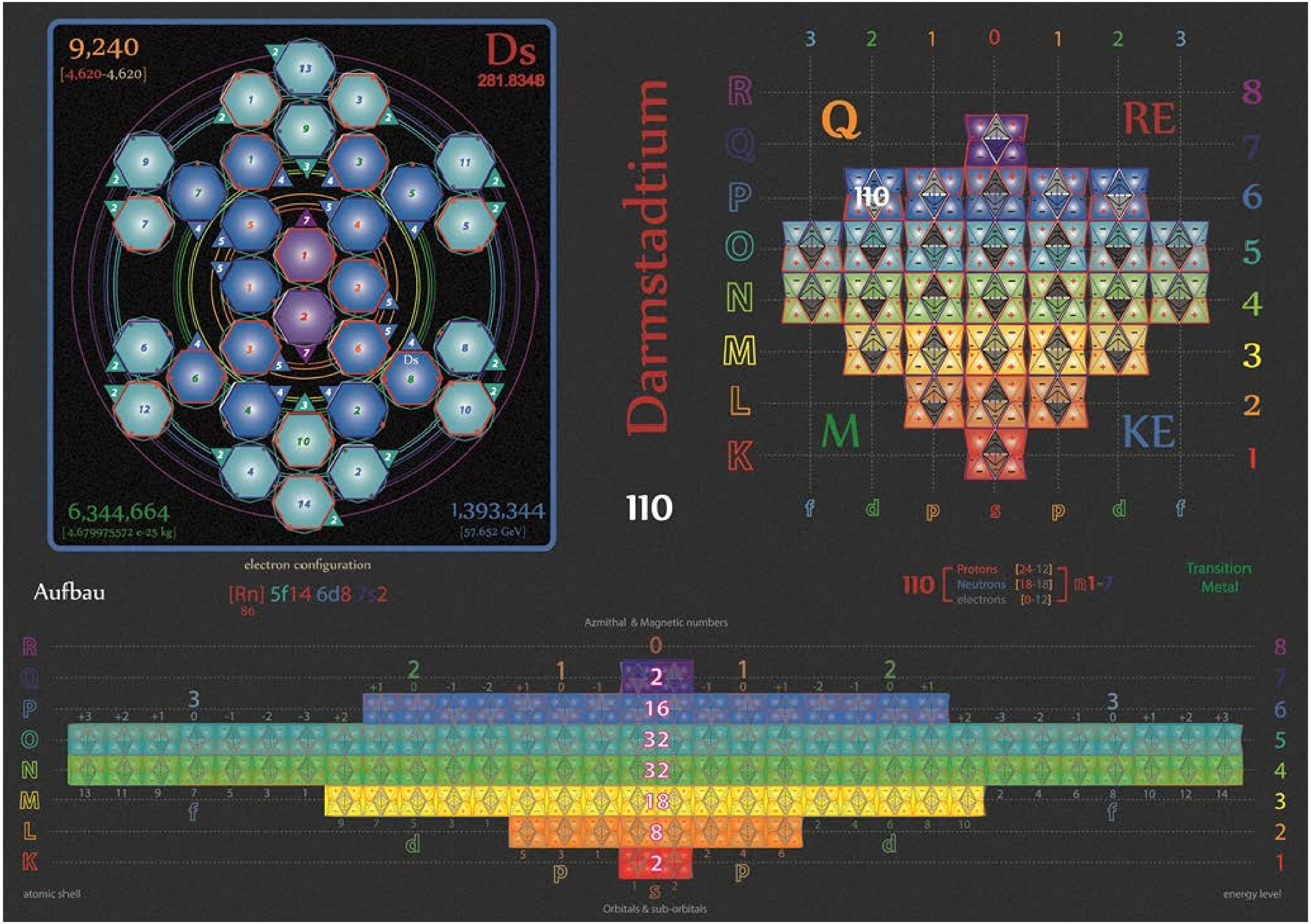
9,240
[4,620-4,620]
Ds
281.8348
6,344,664
[4.679975572 e-25 kg]
1,393,344
[57.652 GeV]
Darmstadtium
110
electron configuration
Aufbau
[Rn] 5f14 6d8 7s2
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
110 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
2
16
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.110 - Darmstadtium atom

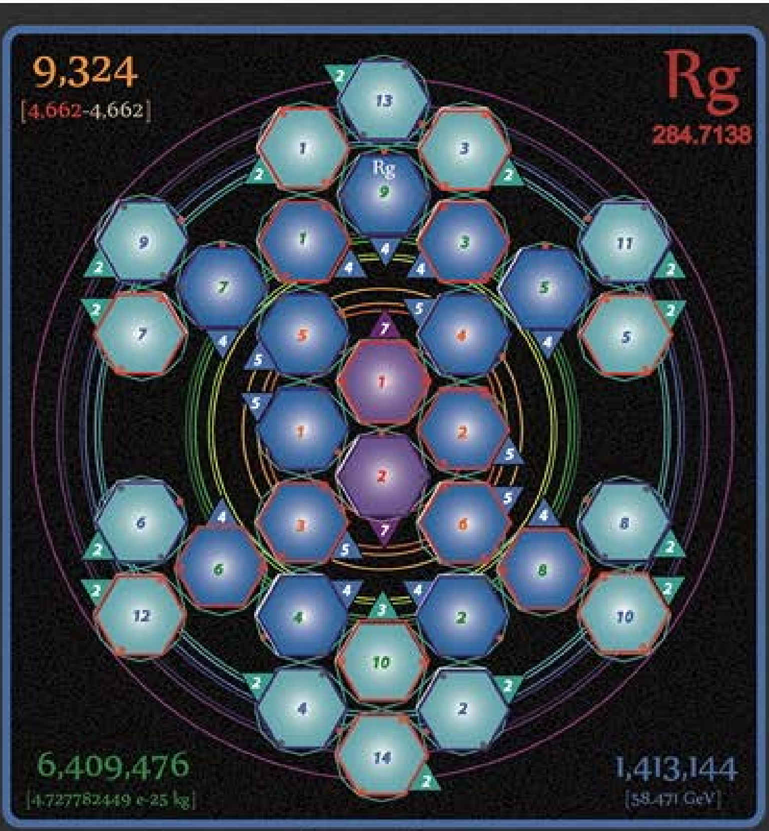
9,324
[4,662-4,662]
Rg
284.7138
Rg
9
6,409,476
[4.727782449 e-25 kg]
1,413,144
[58.471 GeV]

electron configuration

Aufbau

[Rn] 5f14 6d9 7s2
86

Roetgenium

111

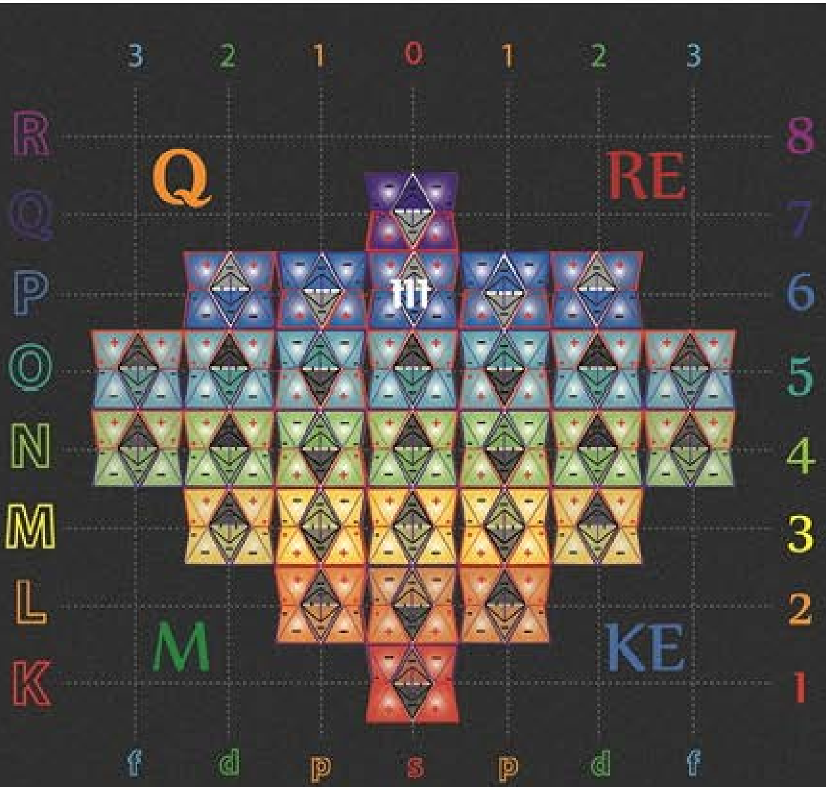
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
111
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

111 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Post Transition Metal

Azmithal & Magnetic numbers

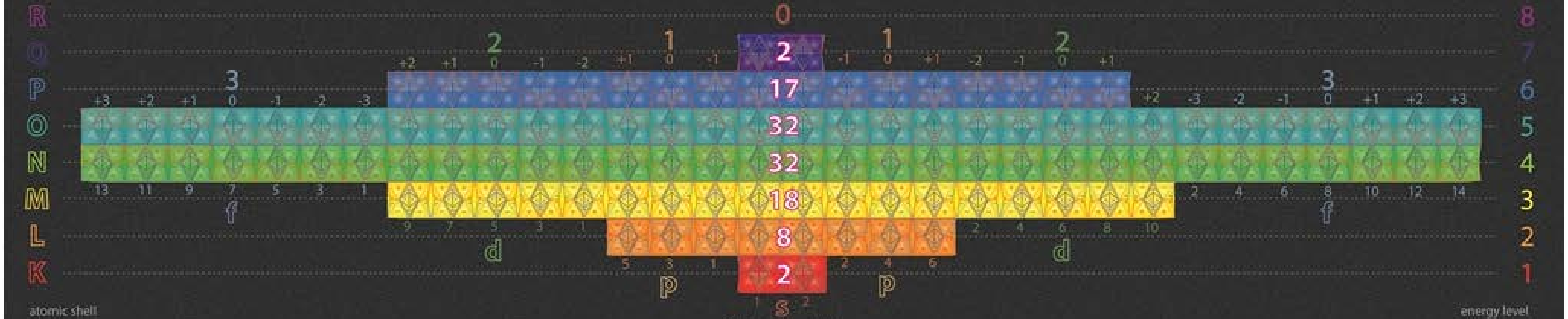
R
Q
P
O
N
M
L
K
0
2
1
1
2
3
3
2
17
32
32
18
8
2
f
d
p
s
p
d
f
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.111 - Roetgenium atom

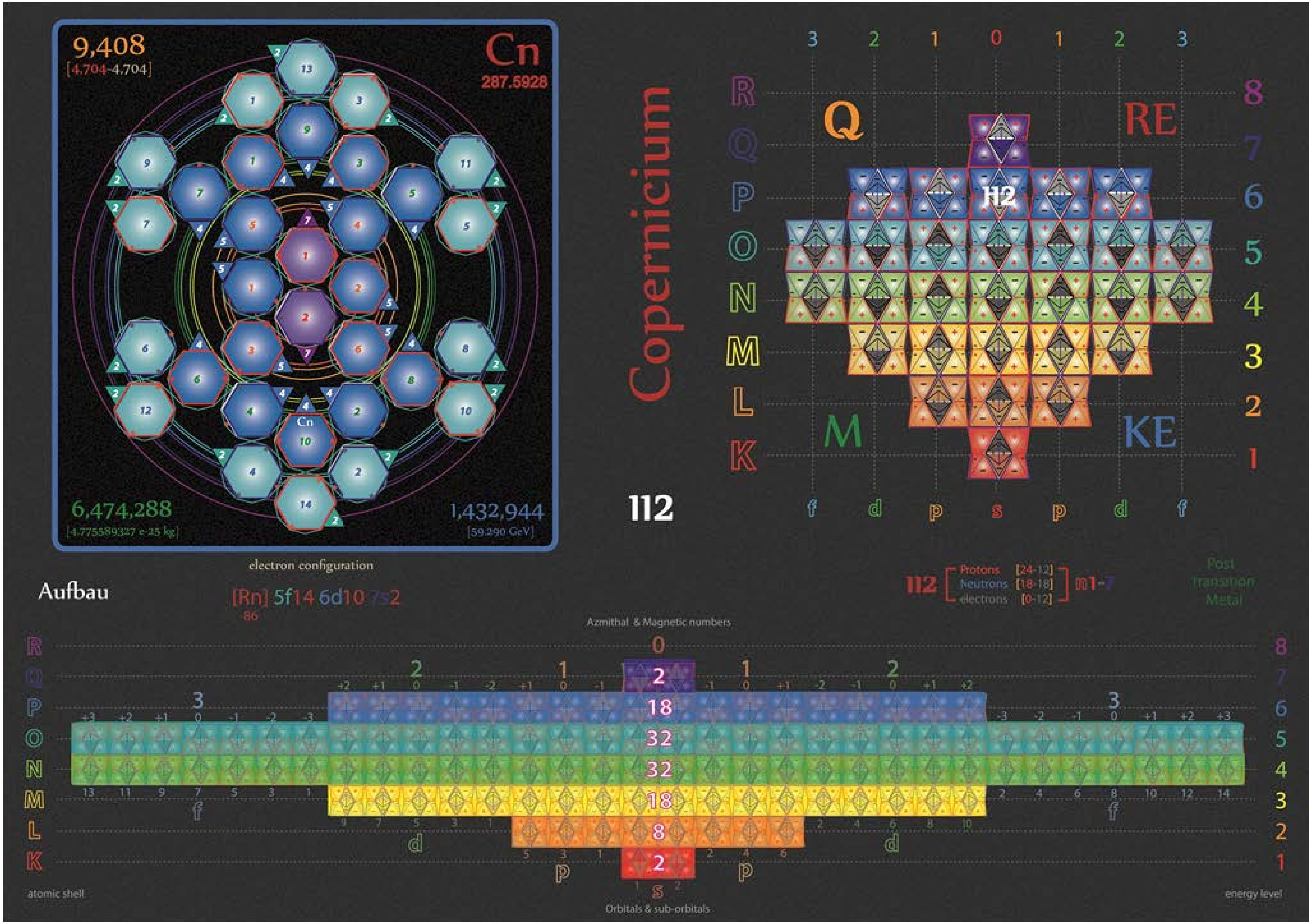
9,408
[4,704-4,704]
Cn
287.5928
6,474,288
[4.775589327 e-25 kg]
1,432,944
[59.290 GeV]
Copernicium
112
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2
86
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
112 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Post transition Metal
Azmithal & Magnetic numbers
2 18 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.112 - Copernicium atom

Copyright ABRAHAM [2008] - All rights reserved

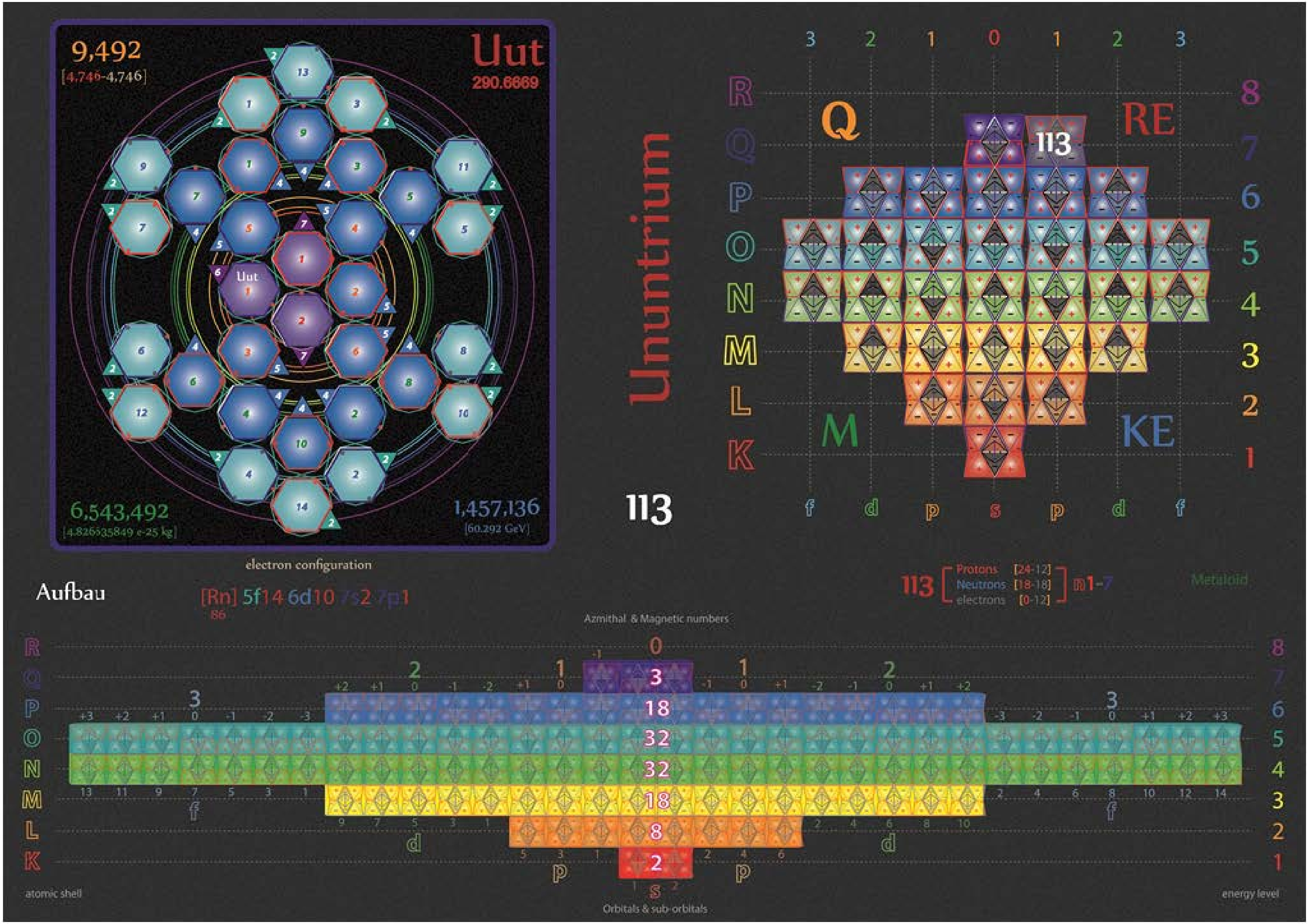
9,492
[4,746-4,746]
Uut
290.6669
Uut
6,543,492
[4.826535849 e-25 kg]
1,457,136
[60.292 GeV]
electron configuration
Ununtrium
113
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
113
f d p s p d f
113
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Metaloid
Aufbau
[Rn] 5f14 6d10 7s2 7p1
86
Azmithal & Magnetic numbers
0
3
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals
f
d
p
s
p
d
f

Tetryonics 51.113 - Ununtrium atom

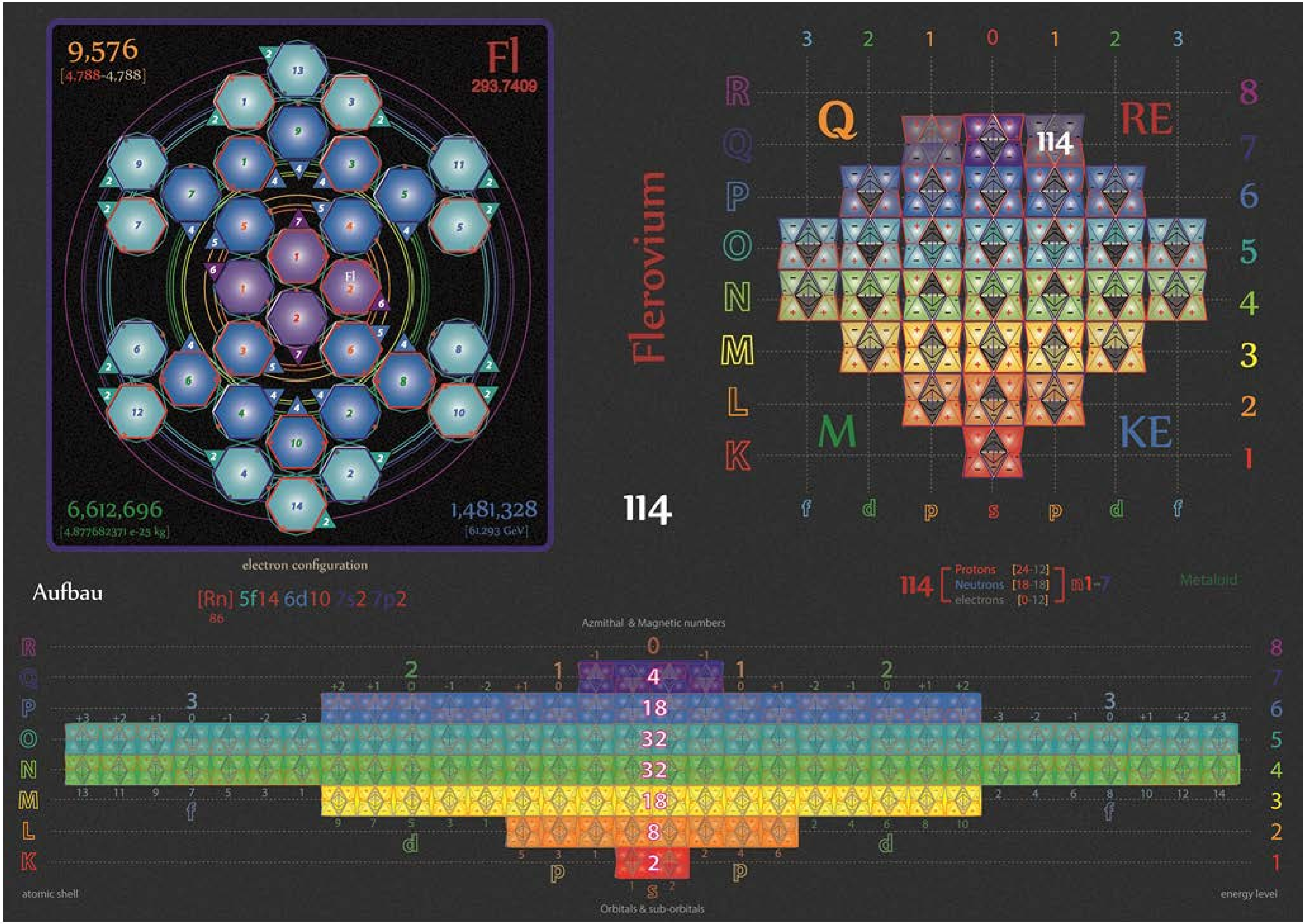
9,576
[4,788-4,788]
Fl
293.7409
6,612,696
[4.877682371 e-25 kg]
1,481,328
[61.293 GeV]
Flerovium
114
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p2
86
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
114 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.114 - Flerovium atom

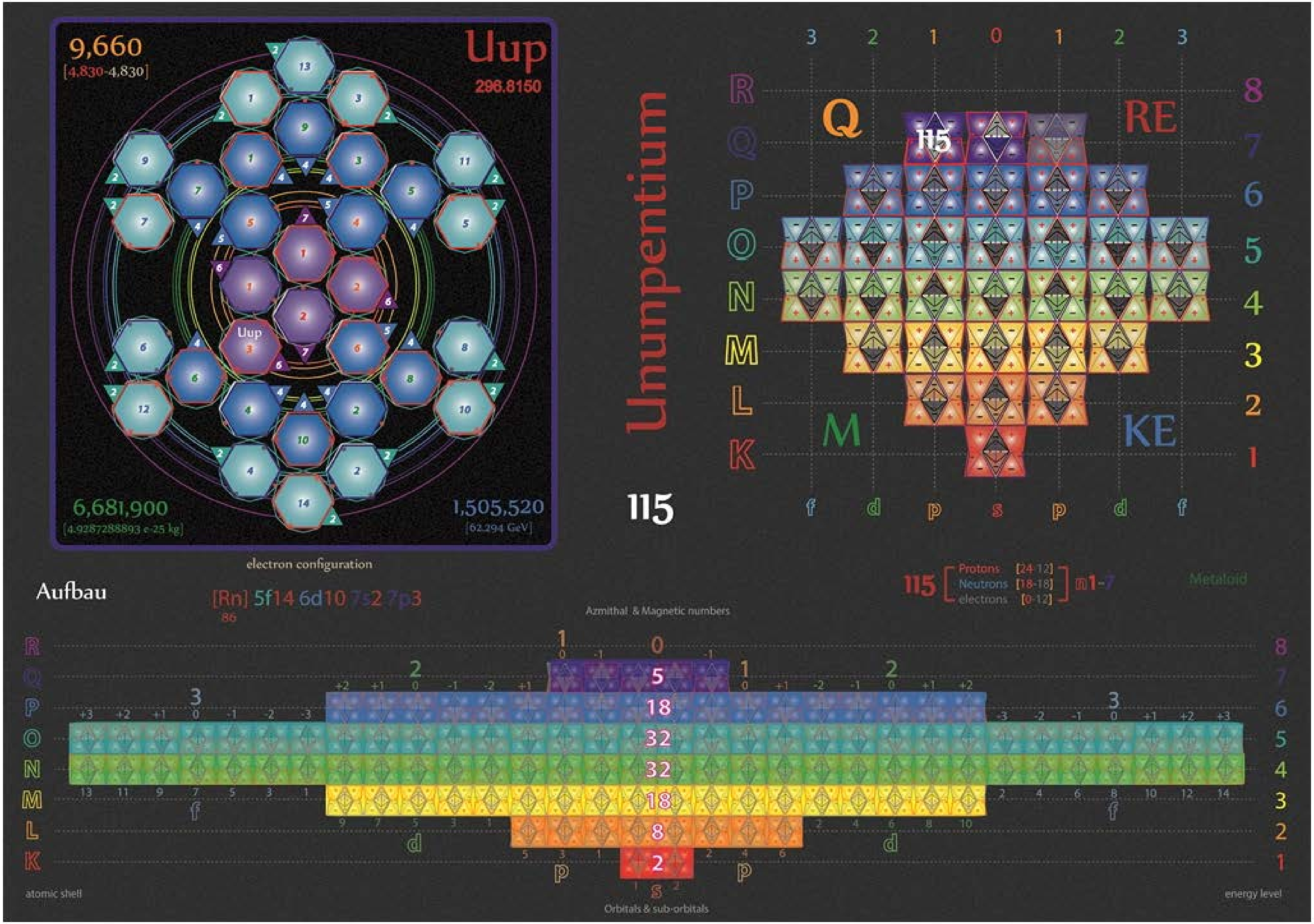
9,660
[4,830-4,830]
Uup
296.8150
6,681,900
[4.9287288893 e-25 kg]
1,505,520
[62.294 GeV]
Ununpentium
115
Q
RE
M
KE
R Q P O N M L K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p3
86
115 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Metaloid
Azmithal & Magnetic numbers
5
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.115 - Ununpentium atom

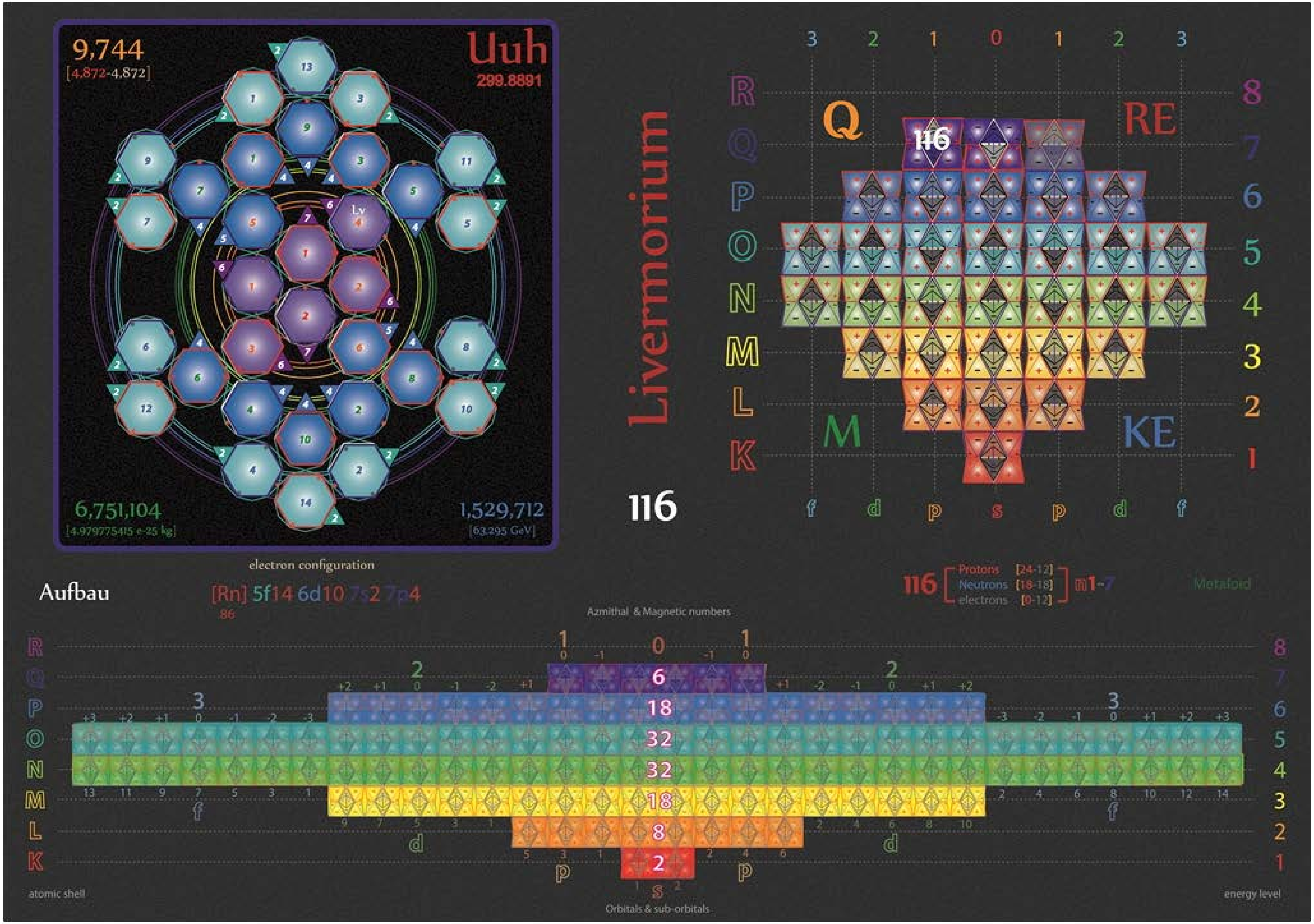
9,744
[4,872-4,872]
Uuh
299.8891
Lv
6,751,104
[4.979775415 e-25 kg]
1,529,712
[63.295 GeV]
Livermorium
116
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p4
86
116 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Metaloid
Azmithal & Magnetic numbers
6
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.116 - Livermorium atom

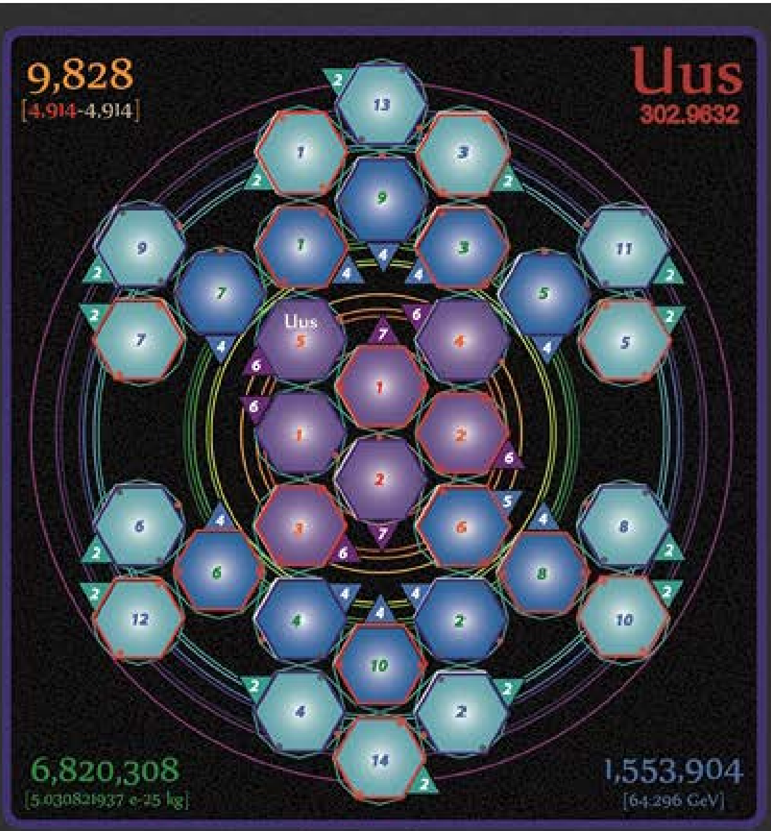
9,828
[4.914-4.914]
Uus
302.9832
Uus
5
6,820,308
[5.030821937 e-25 kg]
1,553,904
[64.296 GeV]

electron configuration

Ununseptium

117

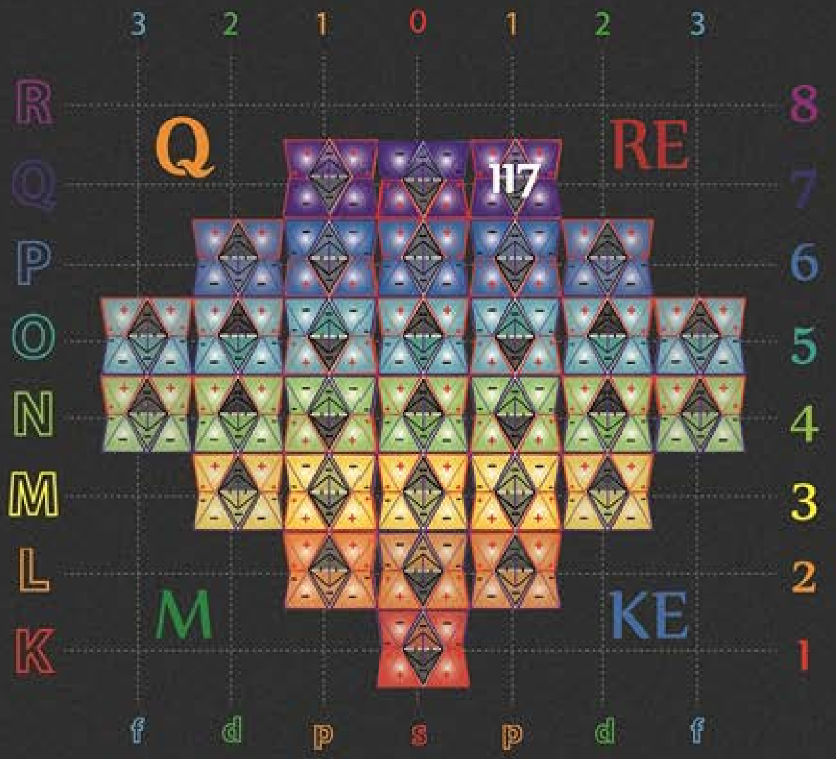
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
117
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

117 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Halogen

Aufbau

[Rn] 5f14 6d10 7s2 7p5
86

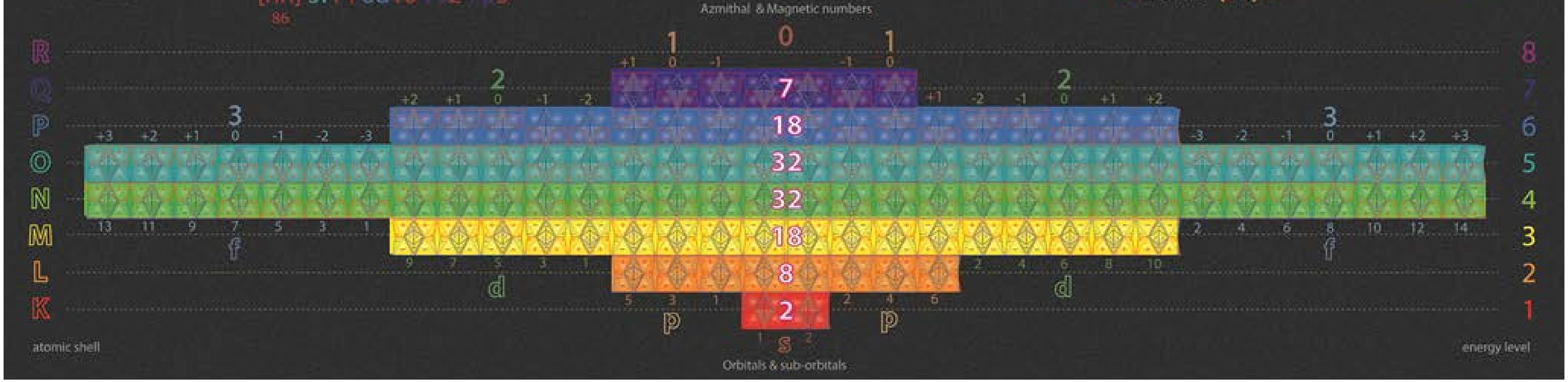
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
1
1
2
2
3
3
7
18
32
32
18
8
2
f
f
d
d
p
p
s
8
7
6
5
4
3
2
1
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.117 - Ununseptium atom

Copyright ABRAHAM [2008] - All rights reserved

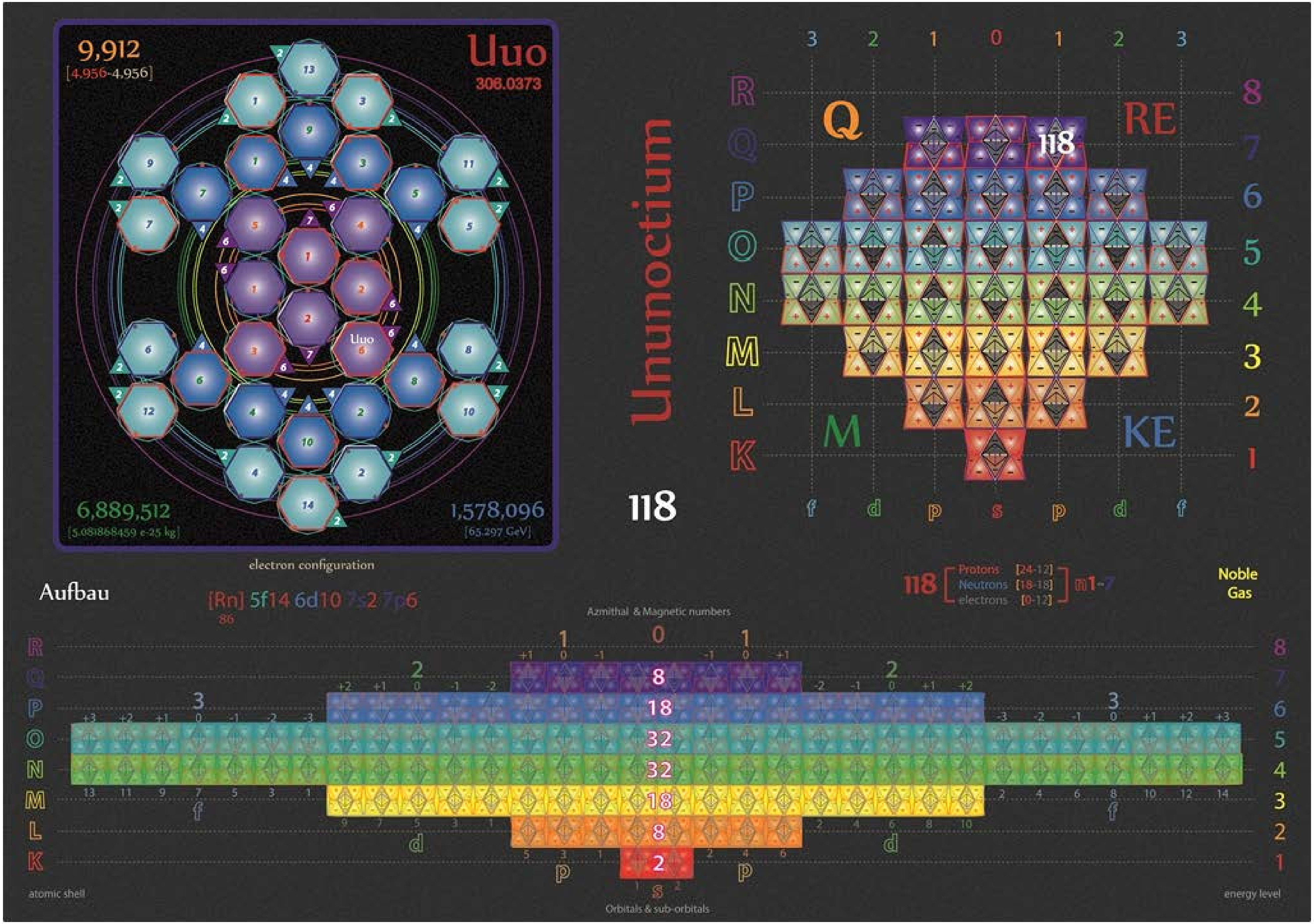
9,912
[4.956-4.956]
Uuo
308.0373
Uuo
6,889,512
[5.081868459 e-25 kg]
1,578,096
[65.297 GeV]
electron configuration
Unuoctium
118
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
118 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Noble Gas
Aufbau
[Rn] 5f14 6d10 7s2 7p6
86
Azmithal & Magnetic numbers
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.118 - Ununoctium atom

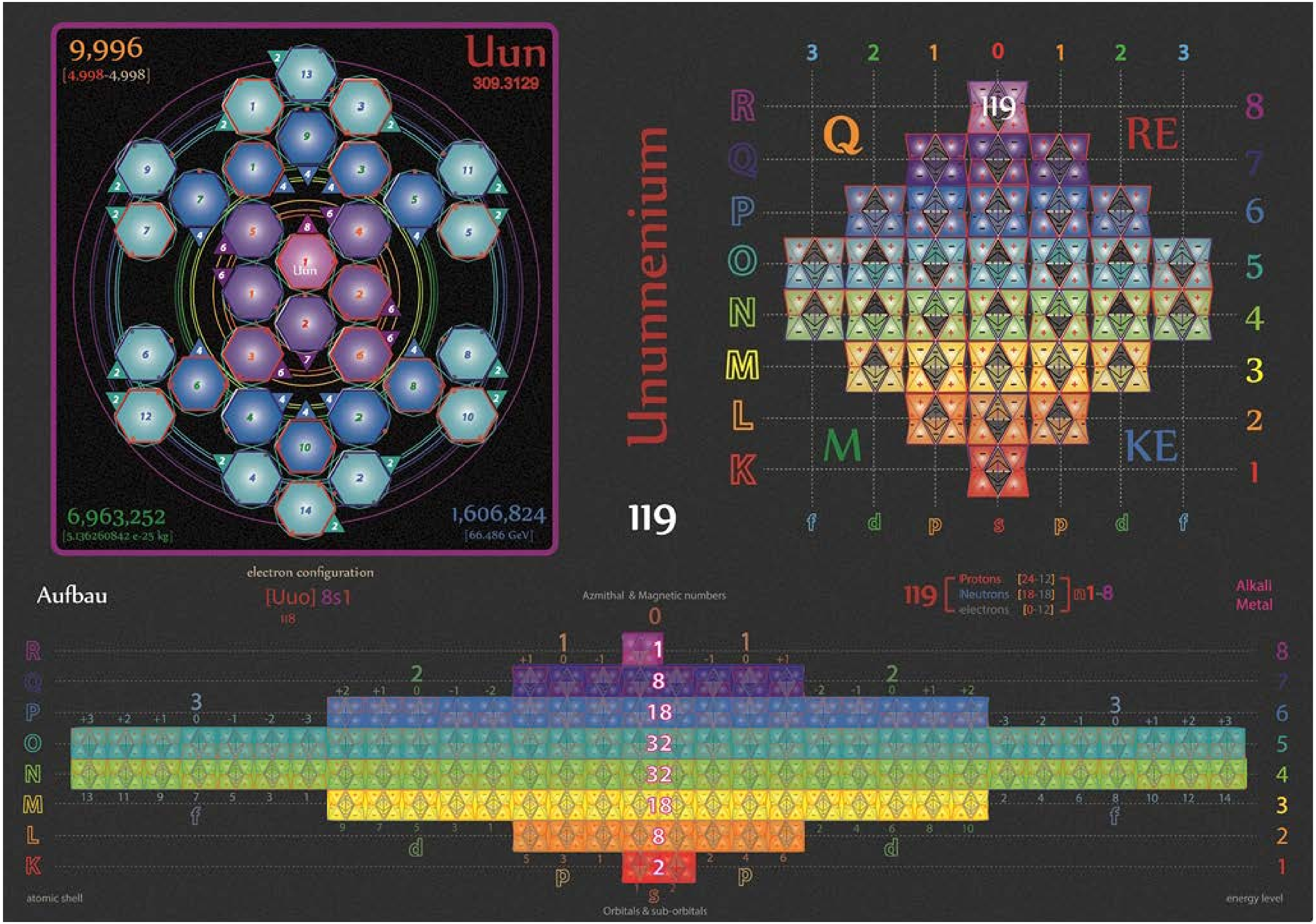
9,996
[4,998-4,998]
Uun
309.3129
Uun
6,963,252
[5.136260842 e-25 kg]
1,606,824
[66.486 GeV]
Ununnenium
119
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
119
f d p s p d f
electron configuration
Aufbau
[Uuo] 8s1
118
Azmithal & Magnetic numbers
119 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-8
Alkali Metal
1
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.119 - Ununnenium atom

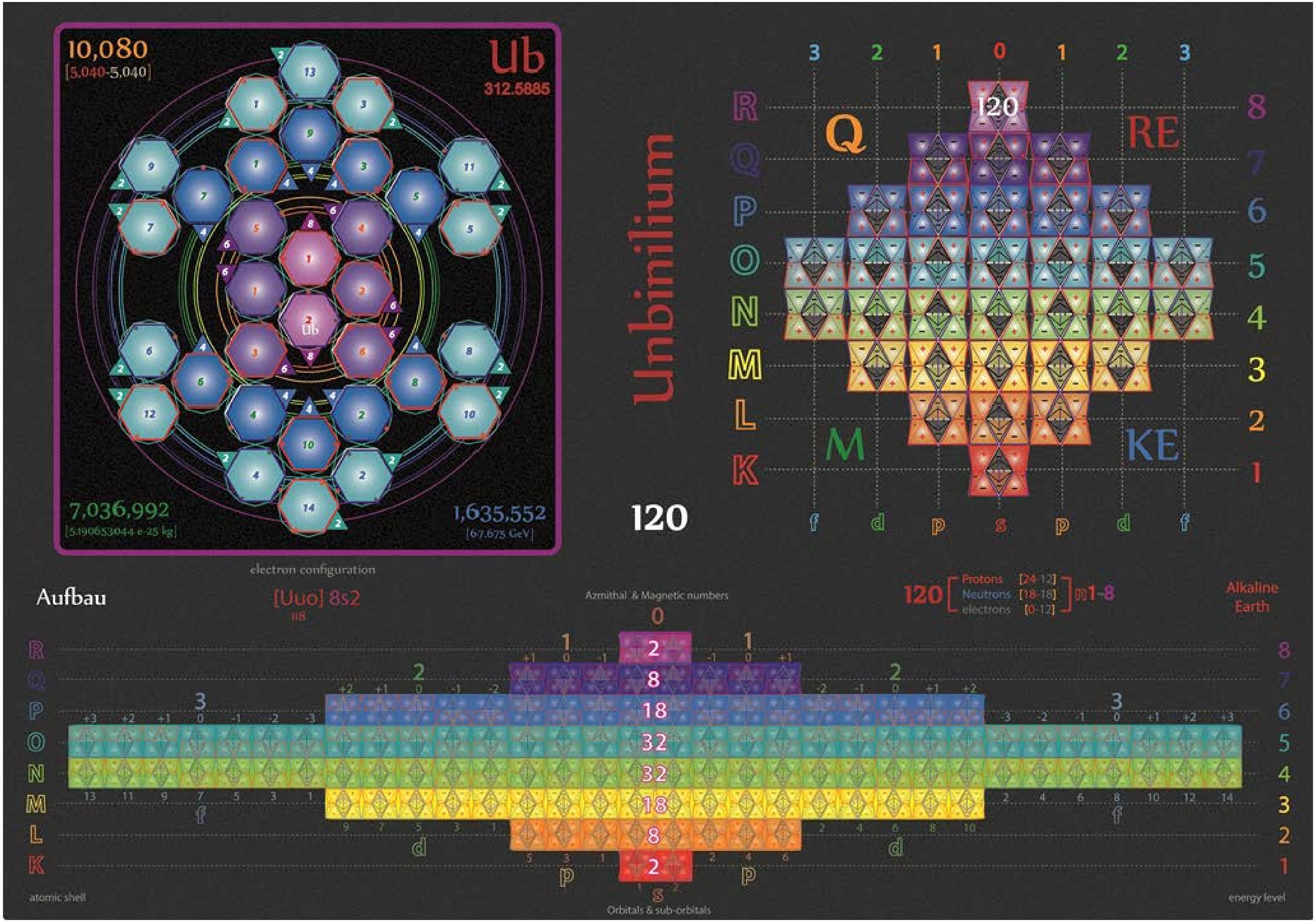
10,080
[5.040-5.040]
Ub
312.5885
7,036,992
[5.190653044 e-25 kg]
1,635,552
[67.675 GeV]
Unbinilium
120
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
120
f d p s p d f
electron configuration
Aufbau
[Uuo] 8s2
118
Azmithal & Magnetic numbers
120 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-8
Alkaline Earth
2
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.120 - Unbinilium atom

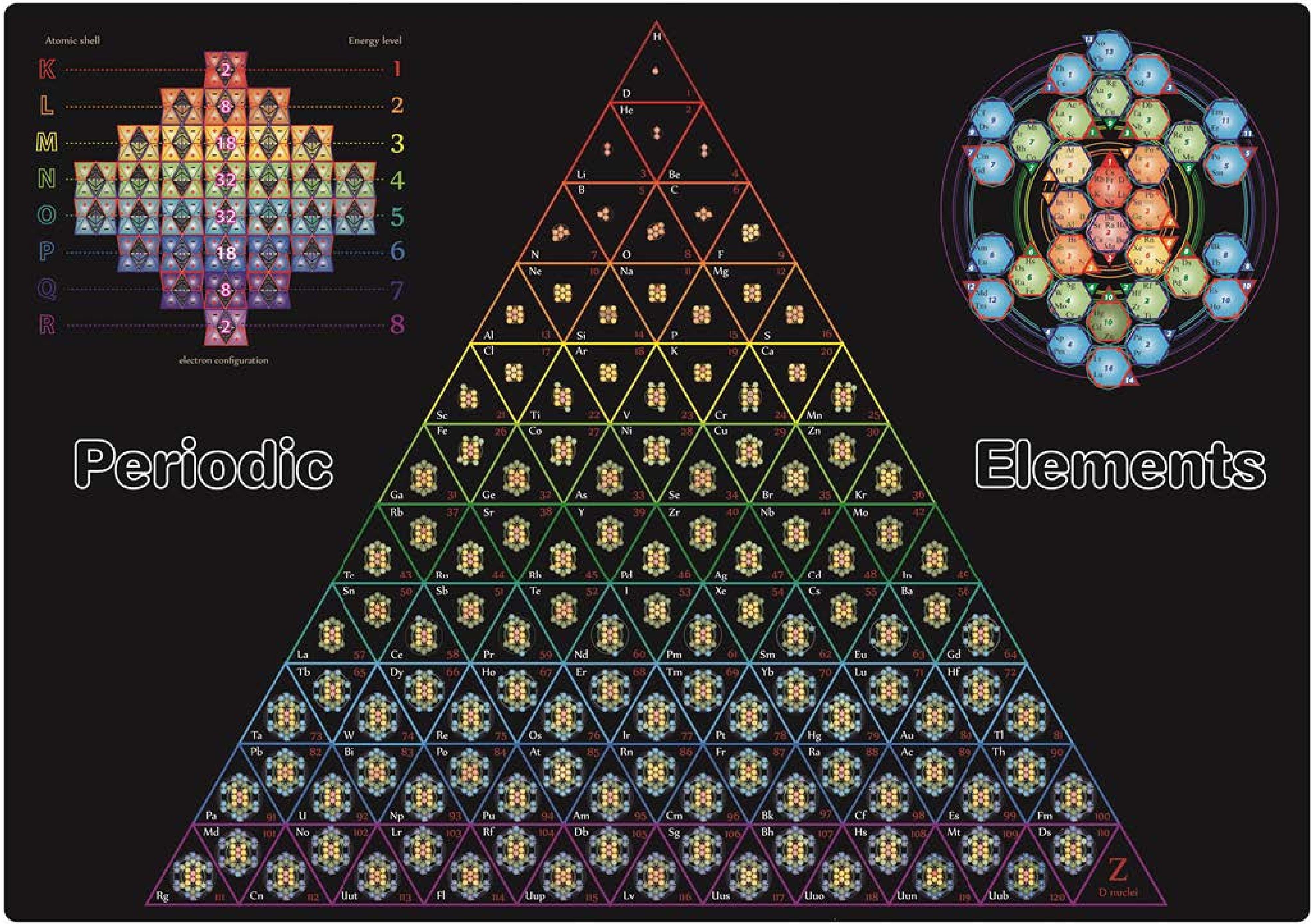
Atomic shell
Energy level
K
L
M
N
O
P
Q
R
1
2
3
4
5
6
7
8
2
8
18
32
32
18
8
2
electron configuration
Periodic
Elements
H
D 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
La 57
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Hg 79
Au 80
Tl 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Ac 89
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103
Rf 104
Db 105
Sg 106
Bh 107
Hs 108
Mt 109
Ds 110
Rg 111
Cn 112
Uut 113
Fl 114
Uup 115
Lv 116
Uus 117
Uuo 118
Uun 119
Uub 120
Z
D nuclei

Tetryonics 52.01 - Tetryonic Element Table

Copyright ABRAHAM [2008] - All rights reserved

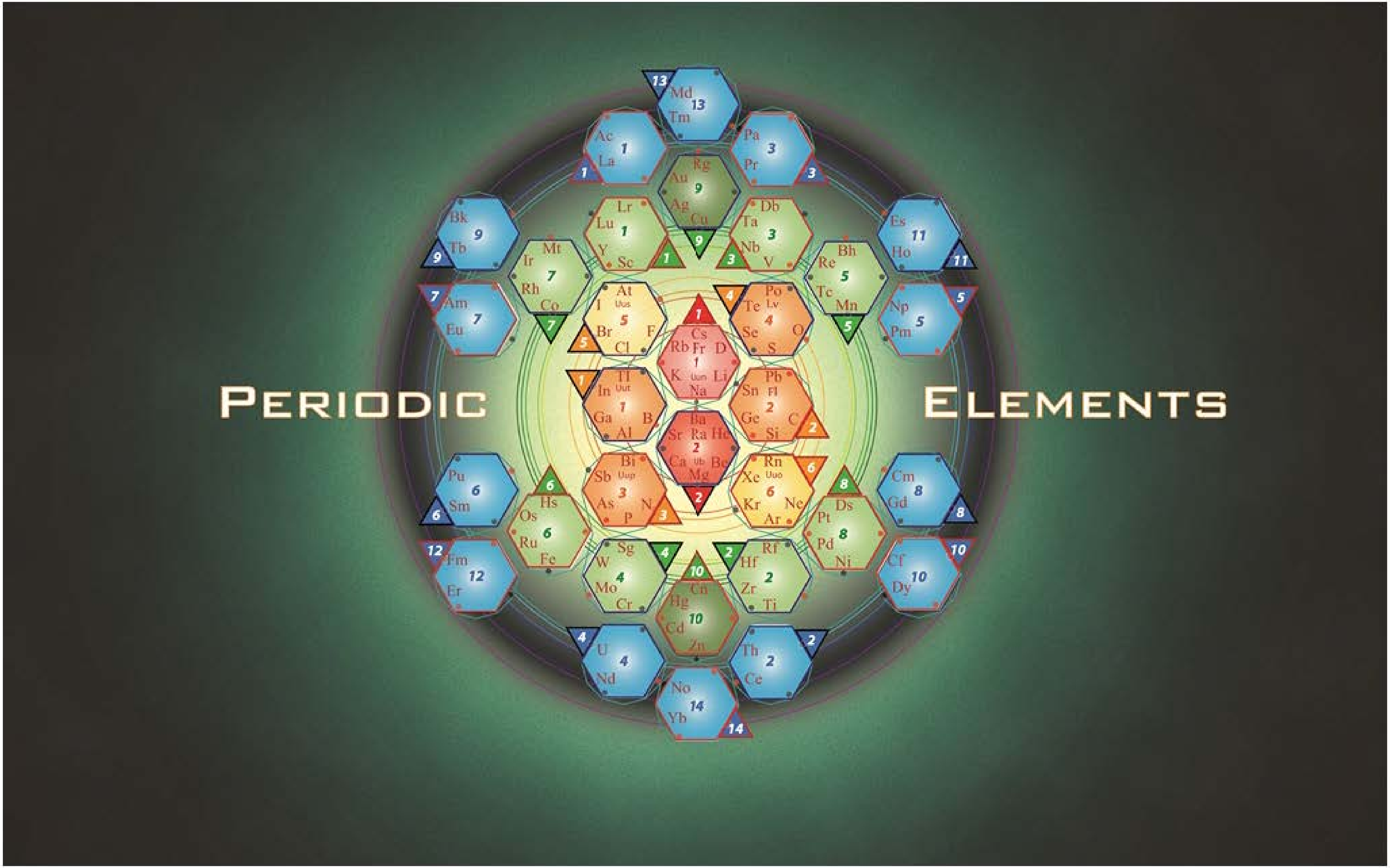
PERIODIC
ELEMENTS
13 Md 13 Tm
Ac 1 La 1
Pa 3 Pr 3
Au Rg 9 Ag Cu 9
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Bk 9 Tb 9
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
Es 11 Ho 11
7 Am 7 Eu
5 Np 5 Pm
At I Uus 5 Br F Cl 5
1 Cs Rb Fr D 1 K Uun Li Na
4 Po Te Lv 4 Se O S
1 Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Bi Sb Uup 3 As N P 3
Rn Xe Uuo 6 Kr Ne Ar 6
Pu 6 Sm 6
6 Hs Os 6 Ru Fe
8 Ds Pt 8 Pd Ni
Cm 8 Gd 8
12 Fm 12 Er
Sg W 4 Mo Cr 4
10 Cn Hg 10 Cd Zn
2 Rf Hf 2 Zr Ti
Cf 10 10 Dy
4 U 4 Nd
Th 2 2 Ce
No 14 Yb 14

Tetryonics 52.02 - Periodic Table [Elements]

Copyright ABRAHAM [2008] - all rights reserved

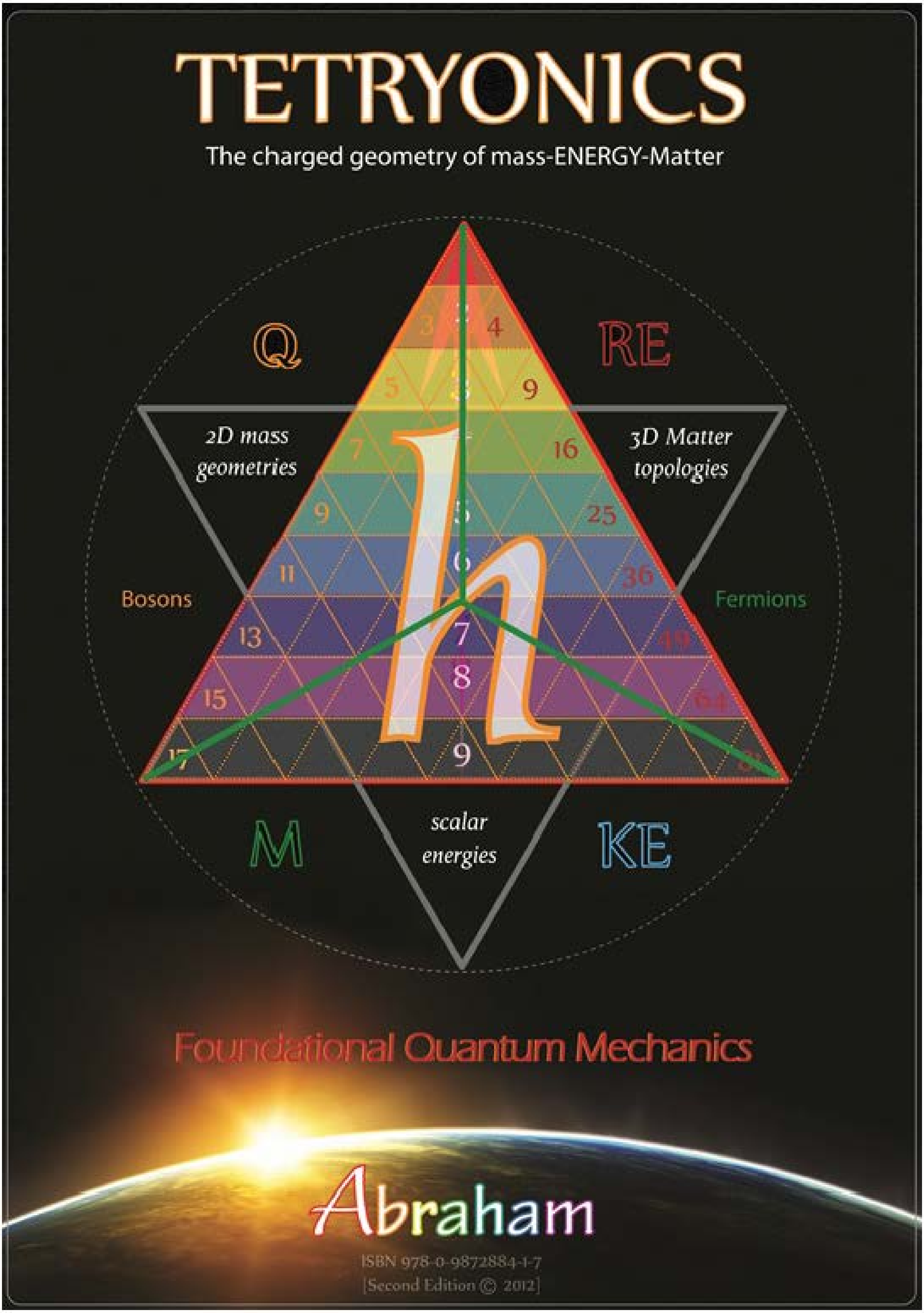
Q
RE
2D mass
geometries
3D Matter
topologies
Bosons
Fermions
h
3
4
5
9
7
16
9
25
11
36
13
49
15
64
17
6
7
8
9
M
KE
scalar
energies

TETRYONICS

The charged geometry of mass-ENERGY-Matter

Foundational Quantum Mechanics

Abraham

ISBN 978-0-9872884-1-7
[Second Edition © 2012]

Copyright ABRAHAM [2008] - all rights reserved

FUNDAMENTUM QUANTUM MECHANICA

questions **Q**

A answers

[TETRYONICS]

A fundamental re-interpretation of the geometry of quantised angular momentum is required to complete the physics of 'The Standard model'

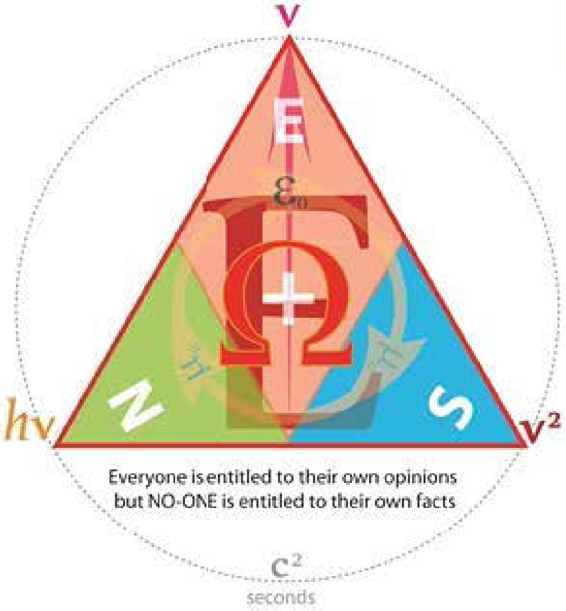

Everyone is entitled to their own opinions
but NO-ONE is entitled to their own facts

Having removed the impossible,
anything that remains, however improbable,
must be the truth

Mathematics is the language of Physics,
and Geometry is its grammar

Philosophiae Naturalis Principia geometricae

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

[Albert Einstein]

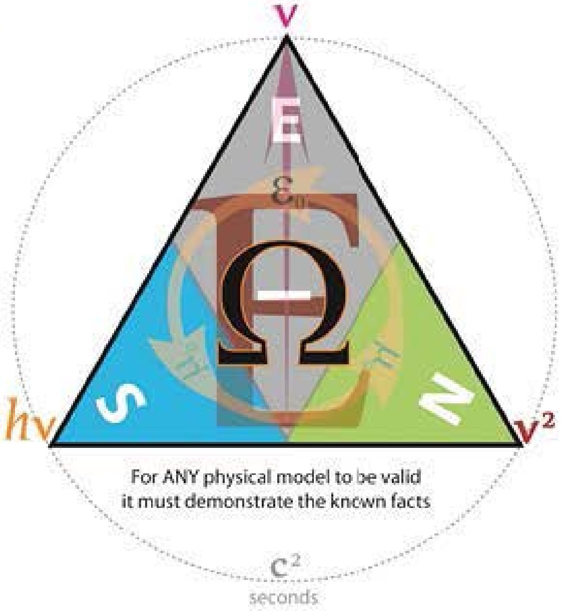

For ANY physical model to be valid
it must demonstrate the known facts

Science is born from observation,
and the reasoning of known facts
in search of underlying truths

In the following pages the true geometry of quantum mechanics is revealed, leading scientific endeavour into new realms of understanding

Copyright ABRAHAM [2008] - all rights reserved

mass-ENERGY-Matter

The a-priori revelation of Tetryonic theory is that all square mass-energies possess equilateral momenta geometries

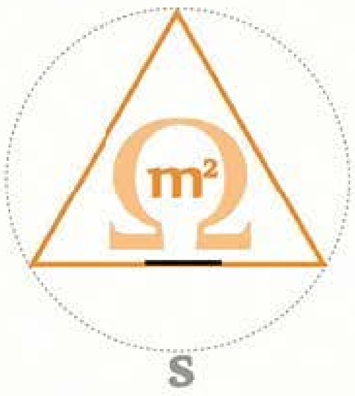

The equilateral Quantised Angular Momentum intrinsic to Planck mass-energy momenta produces charged geometries

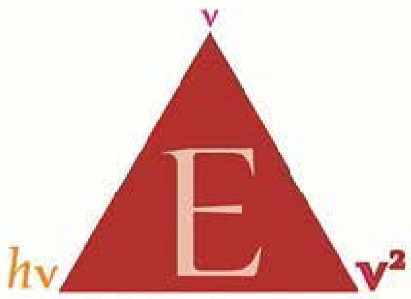

The quantum mechanics of velocity, quanta, EM fields and mass-Energy-Matter can be fully revealed through their equilateral geometries

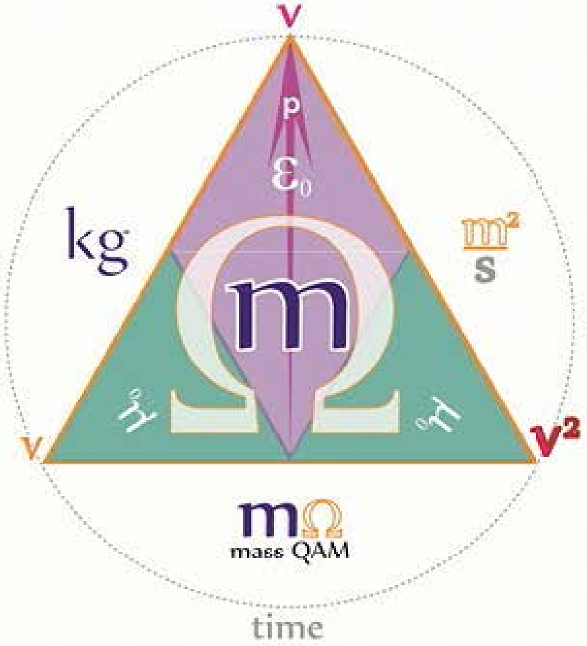

v
p
ε₀
kg
m²
S
m
Ω
v
v²
mΩ
mass QAM
time

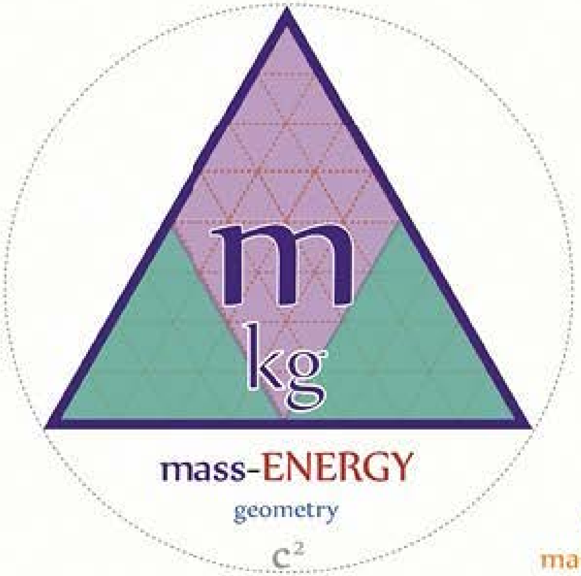

m
kg
mass-ENERGY
geometry
c²

A long hidden topology is revealed

Equilateral triangles are the foundational geometry for all mass-ENERGY-Matter topologies and physical Force interactions

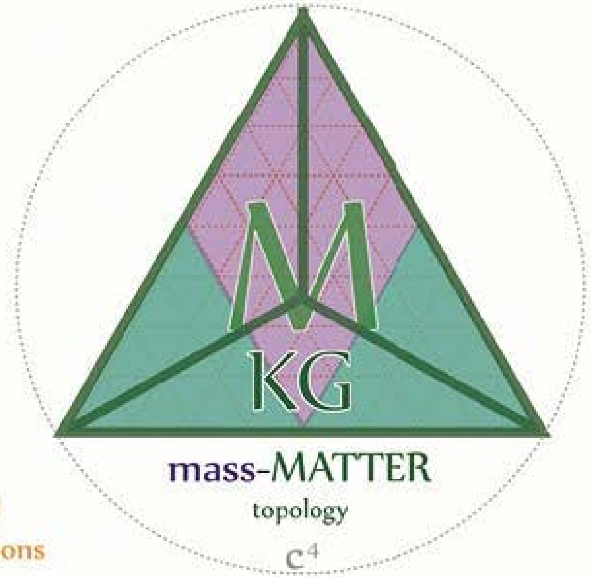

M
KG
mass-MATTER
topology
c⁴

Copyright ABRAHAM [2008] - all rights reserved

SQUARED energies in quantum mechanics are EQUILATERAL geometries

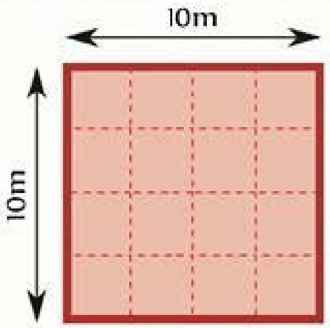

Square

area = s^2 = [100]

Circles

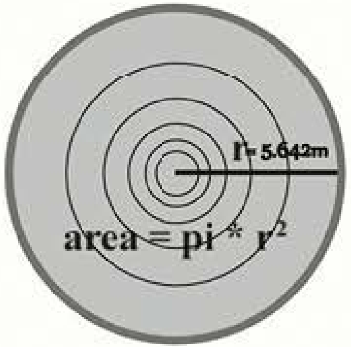
r=5.642m
area = pi * r²

= pi $*[5.642]^2$
= 100

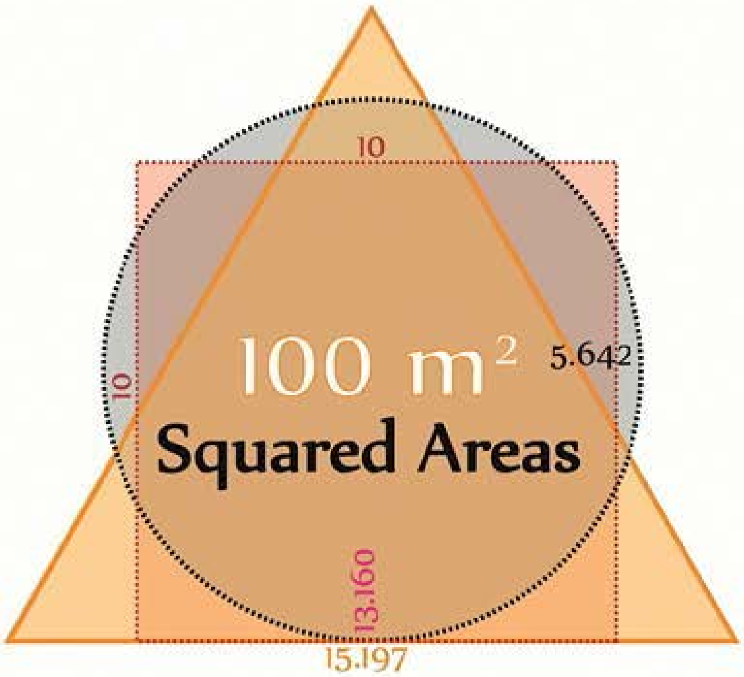
10
10
100 m²
Squared Areas
5.642
13.160
15.197

can be created by a number of planar geometries

For a long time it has been assumed by scientists (and mathematicians)
that circular [and squared] geometries are the geometric foundation of all physics,
leading to a serioulsy flawed model of particles and forces in quantum mechanics

Tetryonic theory now reveals that quantised equilateral angular momenta creates the foundational geometry of all the mass-Energy-Matter & forces of physics

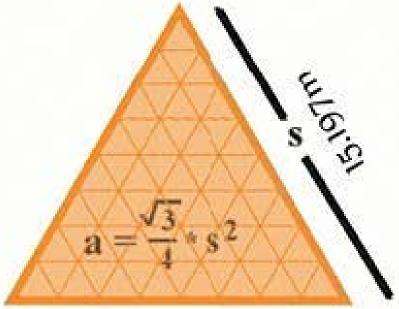
s
15.197m
a = √3/4 * s²

Equilateral

area = $(\frac{1}{2}*b)*h$

Triangles

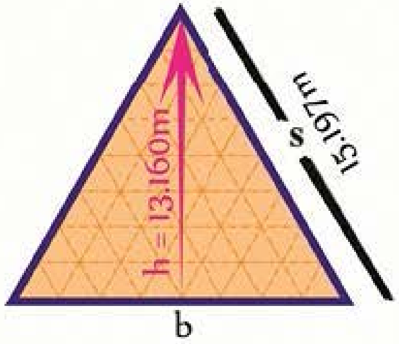
h = 13.160m
s
15.197m
b

b h
[.5x15.197] x 13.160
= 100

Copyright ABRAHAM [2008] - all rights reserved

Integers

The integers (from the Latin integer), literally "untouched", hence "whole"
in Tetryonics it is the basis for the quantum

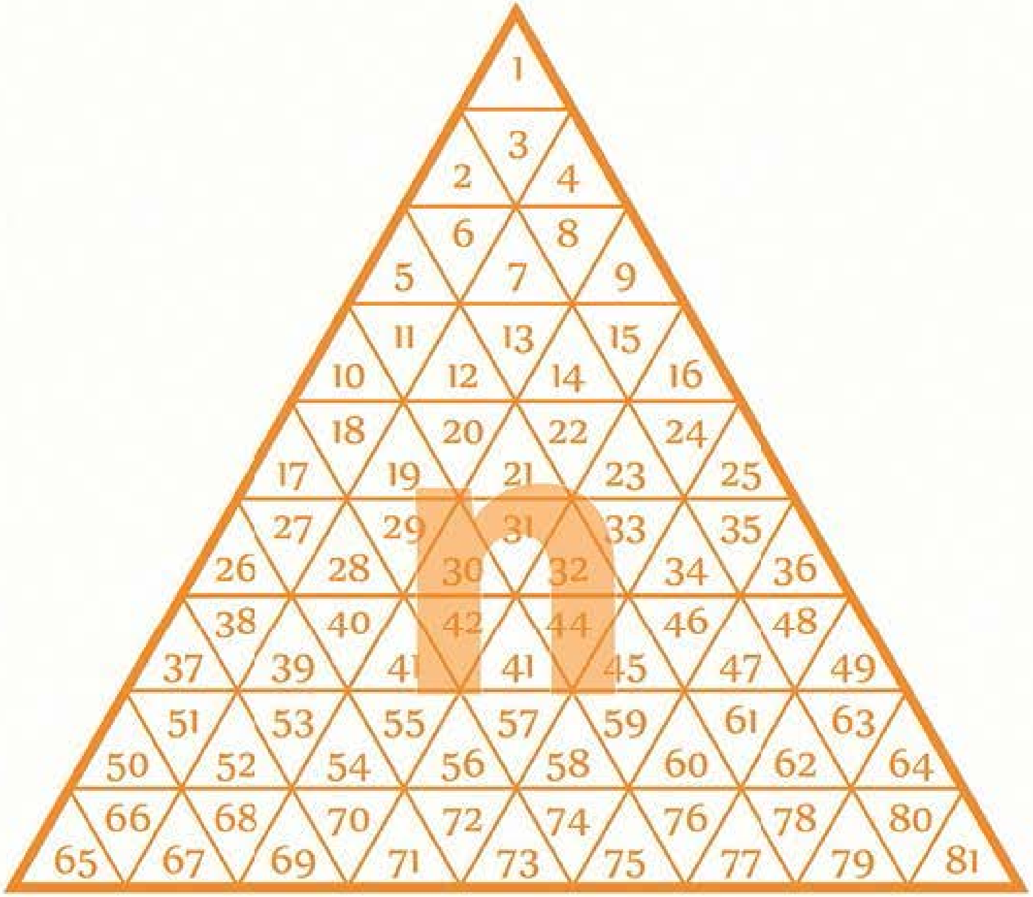
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 41 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
n

Viewed as a subset of the real numbers, they are numbers that can be written without a fractional or decimal component

Copyright ABRAHAM [2008] - all rights reserved

ODD numbers

An odd number is an integer which is not a multiple of two.

$2n-1$

$2n+1$

ODD numbers
in each level

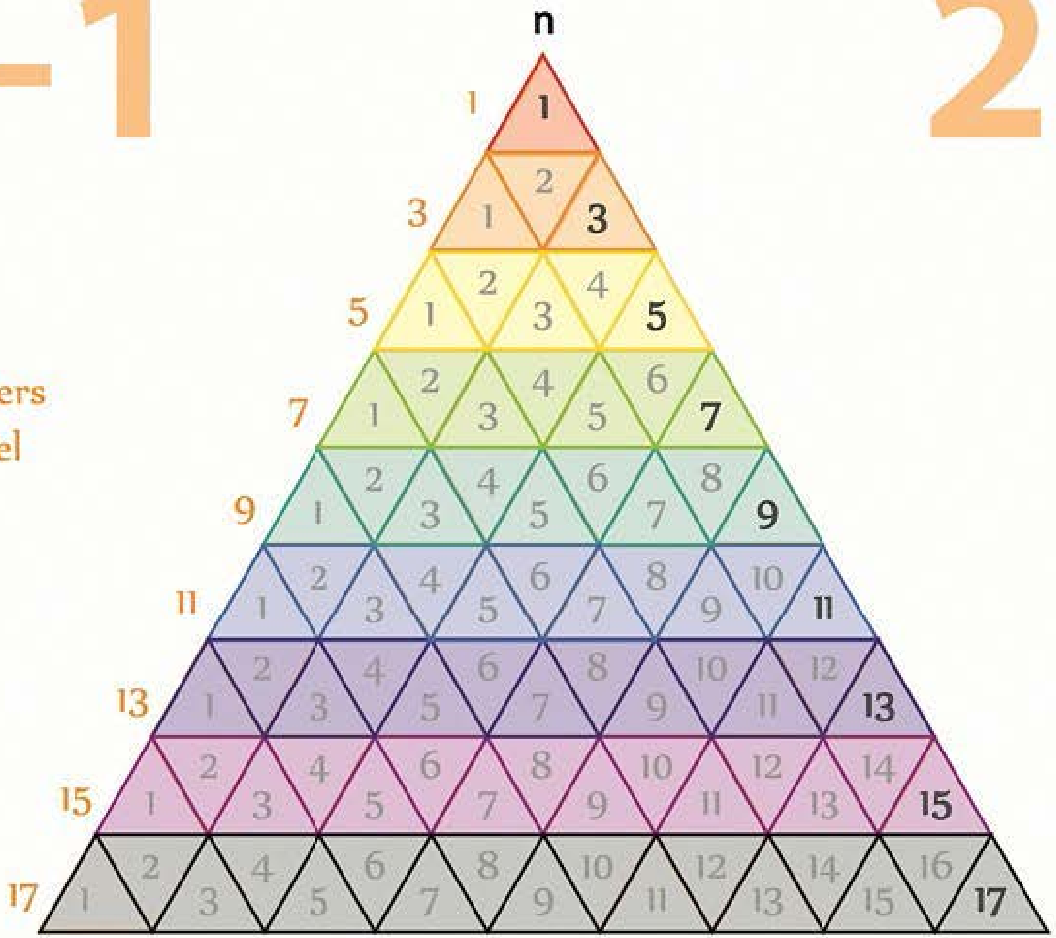

Bosons have
ODD number
quanta

An odd number, when divided by two, will result in a fraction

Copyright ABRAHAM [2008] - all rights reserved

EVEN numbers

An integer that is not an odd number is an even number.

2

Photons have
EVEN number
quanta

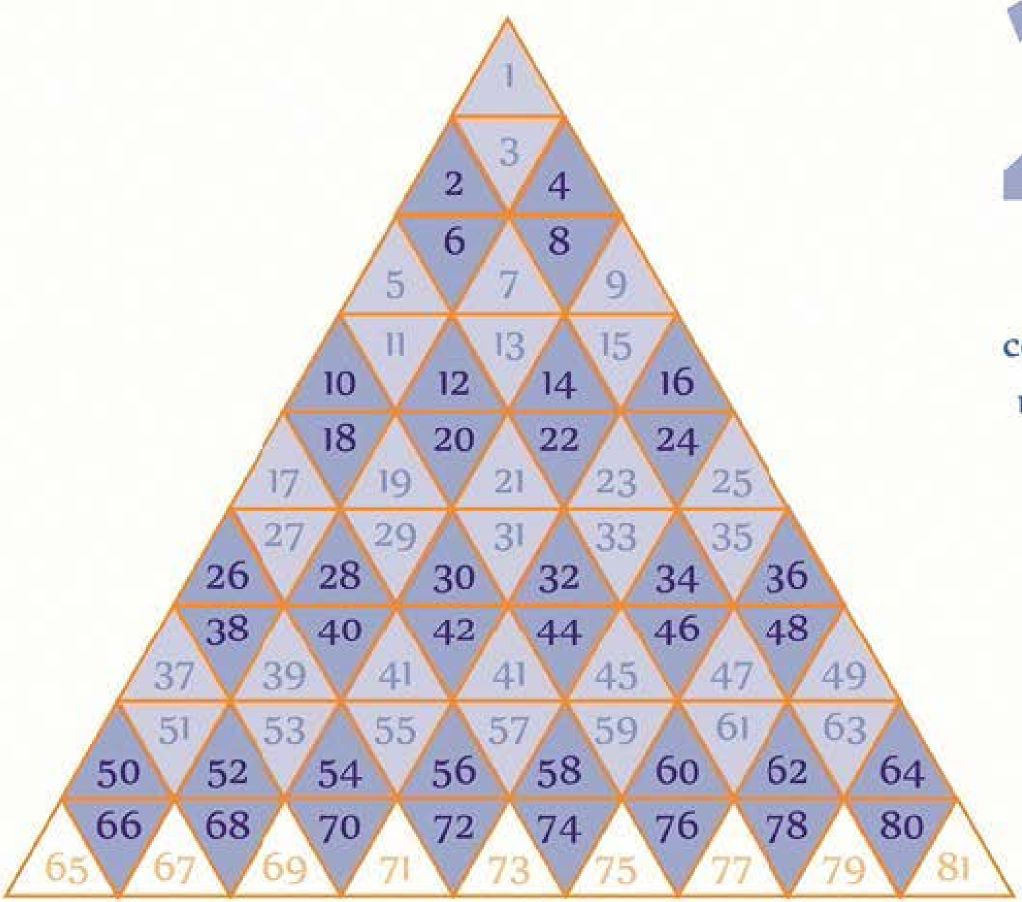
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 41 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

2n

EM waves are
comprised of EVEN
numbered quanta

An even number is defined as a whole number that is a multiple of two.
If an even number is divided by two, the result is another whole number.

Copyright ABRAHAM [2008] - all rights reserved

Square numbers

A square number, sometimes also called a perfect square,
is the result of an integer multiplied by itself

2n-1

Quantum levels
have ODD number
geometries

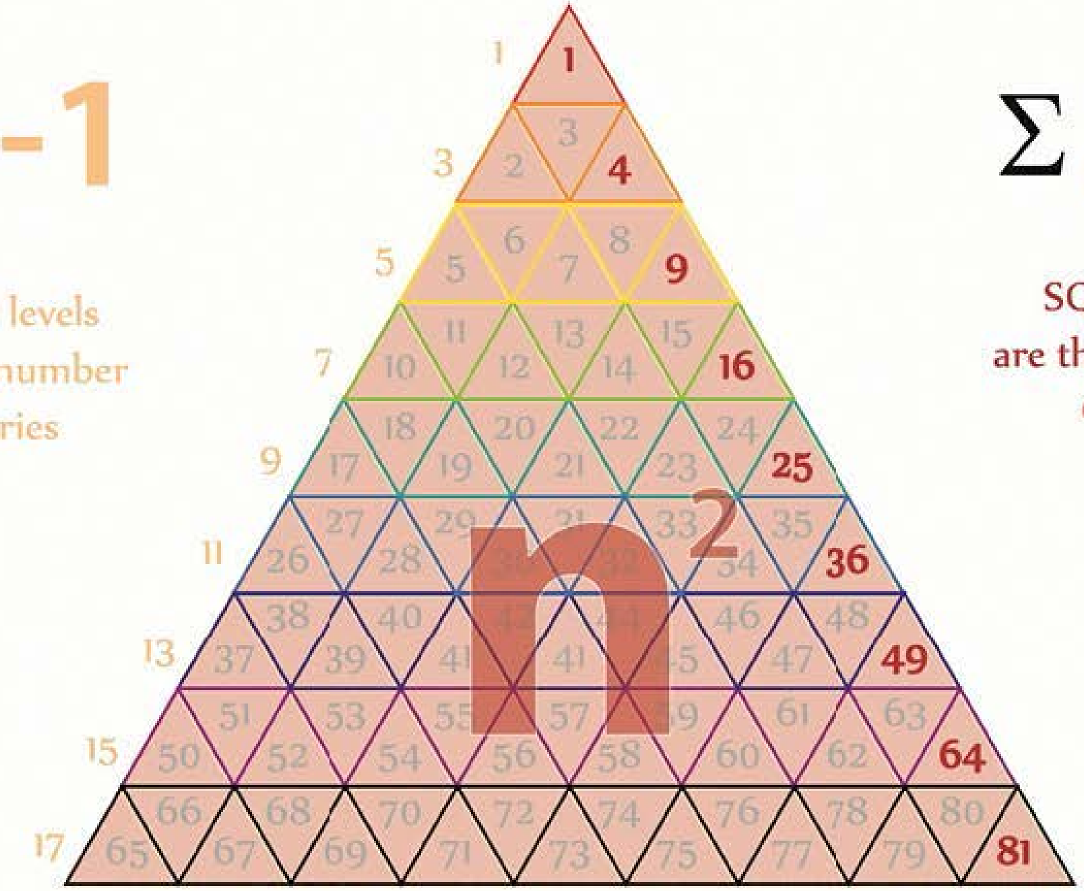
1
3
5
7
9
11
13
15
17
1
2 3 4
5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36
37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
n^2

Square numbers

Σ **2n-1**

SQUARE numbers
are the sum of successive
ODD numbers

In Tetryonics SQUARE numbers are EQUILATERAL geometries

Square roots

A square root of a number is a number that, when it is multiplied by itself (squared) , gives the first number again.

-*i* and +*i*

Against Mathematical convention,
square roots of negative numbers
are real numbers

In Physics
every complex number
except 0 has 2 square roots.

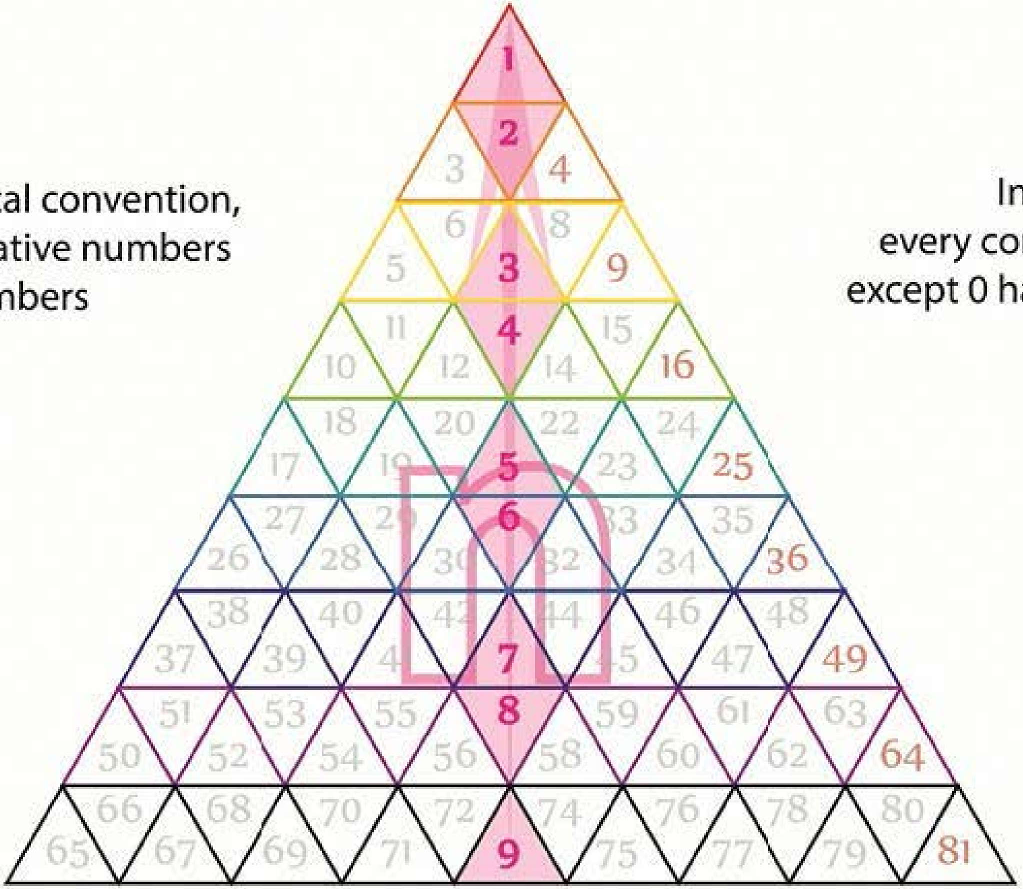
1
3 2 4
5 6 3 8 9
10 11 12 4 14 15 16
17 18 19 20 5 22 23 24 25
26 27 28 29 30 6 32 33 34 35 36
37 38 39 40 41 42 7 44 45 46 47 48 49
50 51 52 53 54 55 56 8 58 59 60 61 62 63 64
65 66 67 68 69 70 71 72 9 74 75 76 77 78 79 80 81

2n-1

n^2

A whole number with a square root that is also a whole number is called a perfect square

Copyright ABRAHAM [2008] - all rights reserved

Real Numbers

A real number is a value that represents a quantity along a continuous line.

The real numbers include all the rational numbers,

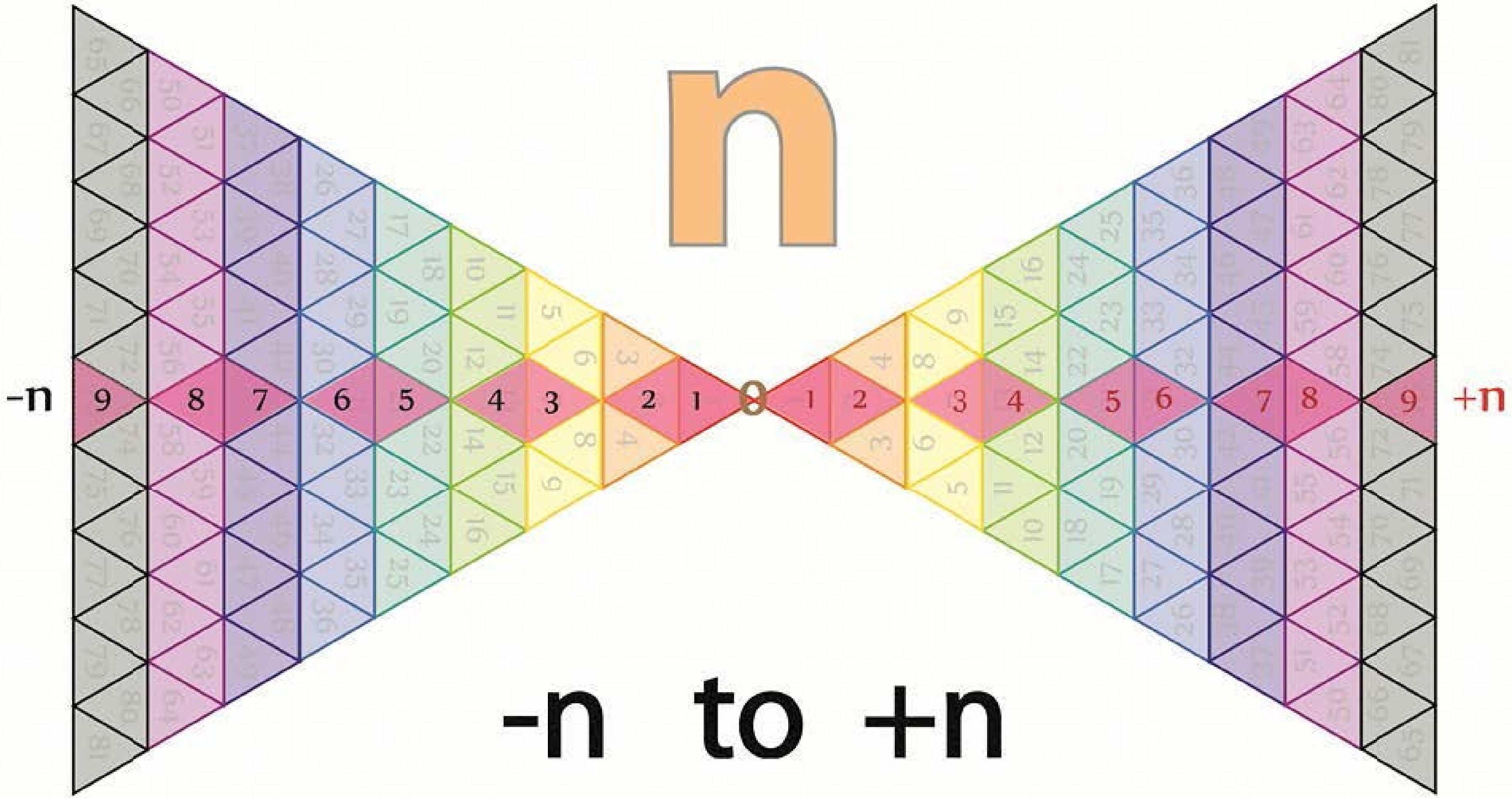
n
-n
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
+n
-n to +n

Copyright ABRAHAM [2008] - all rights reserved

Irrational Numbers

An irrational number is defined to be any real number that cannot be written as a complete ratio of two integers

$+i$ and $-i$

Well known irrational & imanginary numbers in Math are π and i

π

$i^2 = -1$

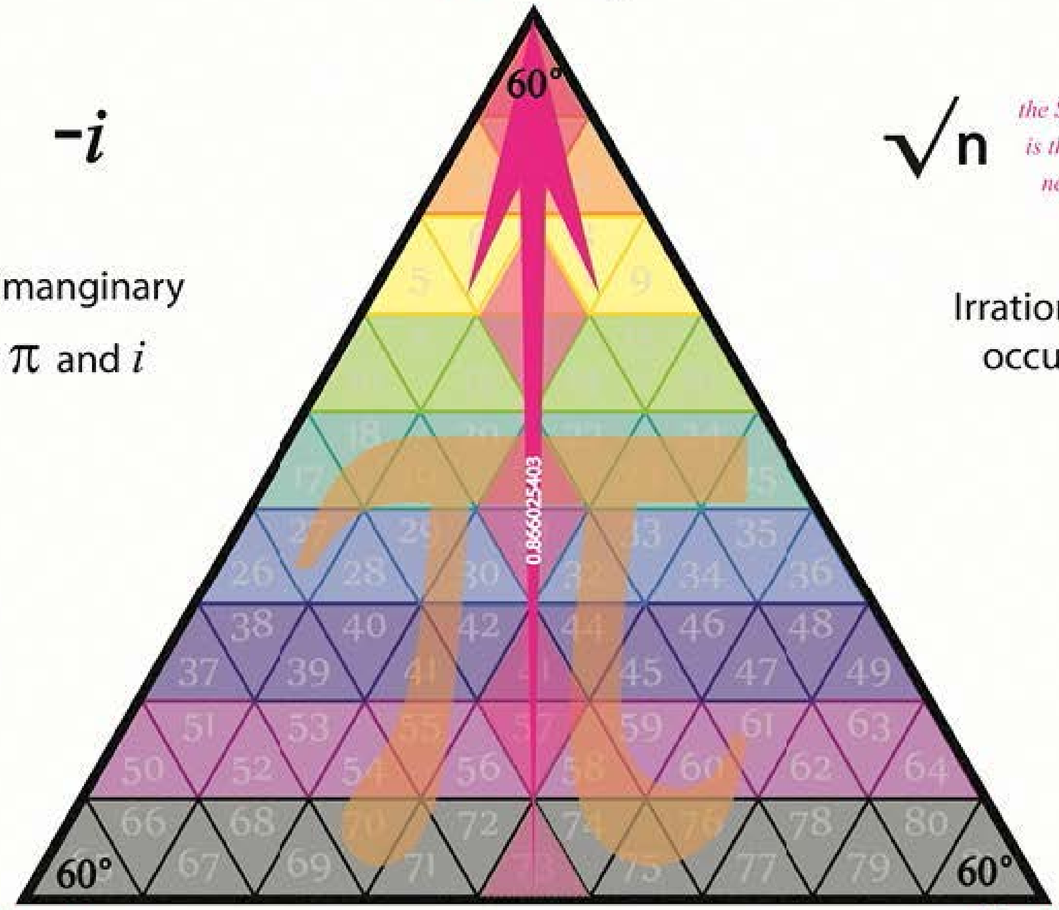
60°
60°
60°
0.866025403
5
9
17
18
25
26
27
28
29
30
32
33
34
35
36
37
38
39
40
41
42
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80

Sin 60

Sin π/3

$\sqrt{3/2}$

$\sqrt{n}$ *in Tetryonics the SQR of a negative number is the linear momentum of a negative charge EM field* $\sqrt{-n}$

Irrational numbers often occur in mathematics

$\sqrt{-n}$

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Colour Code

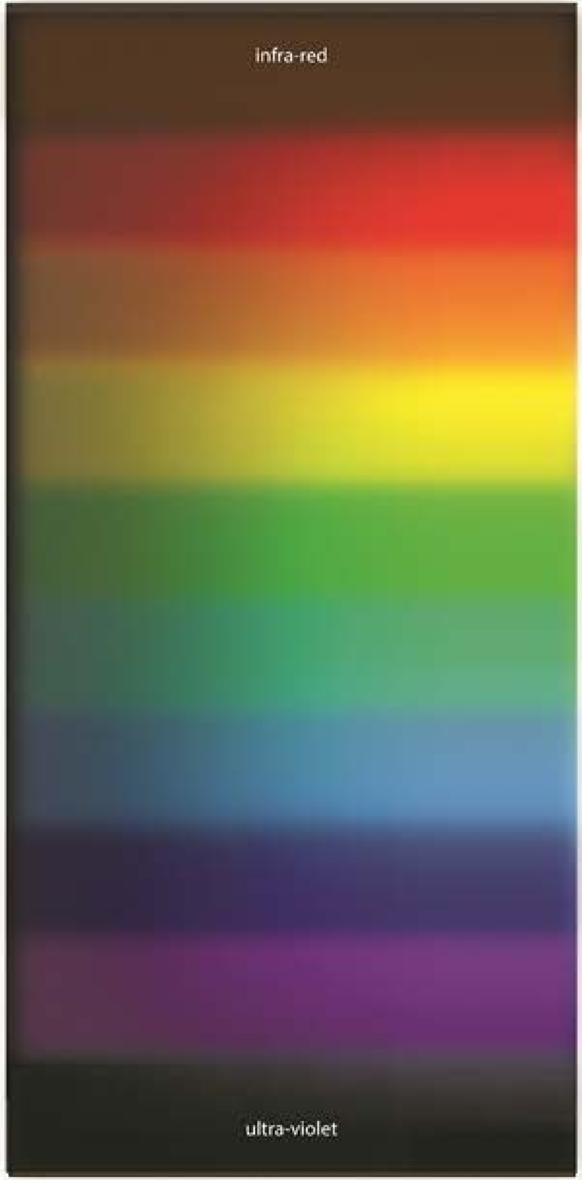
infra-red
ultra-violet

0	Brown
1	Red
2	Orange
3	Yellow
4	Green
5	Aqua
6	Blue
7	Indigo
8	Violet
9	Black

A colour code is used to indicate the varying quantum levels of the numerous forms of mass-ENERGY-Matter and serves to illustrate relationships between various Physical properties.

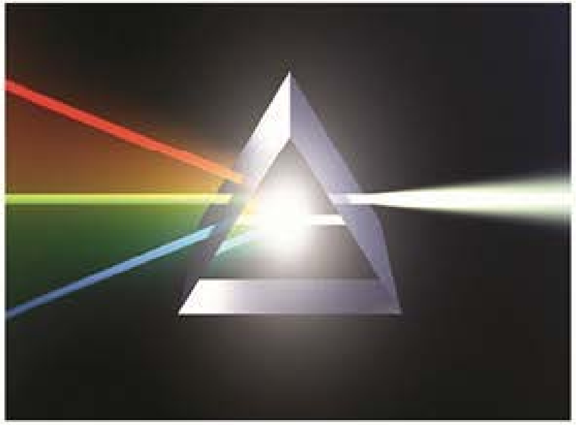

Tetryonics uses a colour code that is based on the spectral colours of dispersed White Light

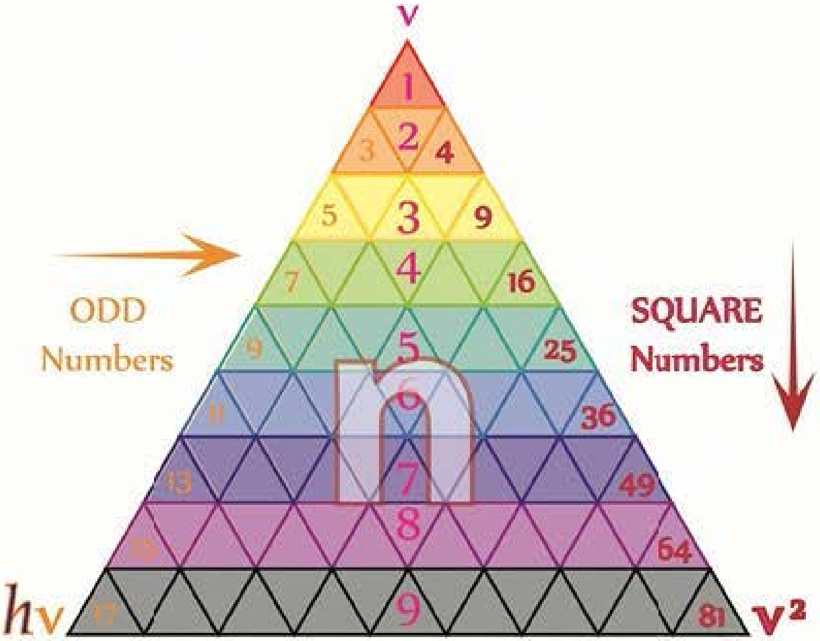
v
1
2
3
4
5
6
7
8
9
3
5
7
9
11
13
15
17
4
9
16
25
36
49
64
81
n
ODD Numbers
SQUARE Numbers
hv
v²

Copyright ABRAHAM [2008] - all rights reserved

Free Space

A contiguous volume or area
of any regular geometry
that is free, available, or unoccupied

$$\underset{\text{in any form}}{E} = 0$$

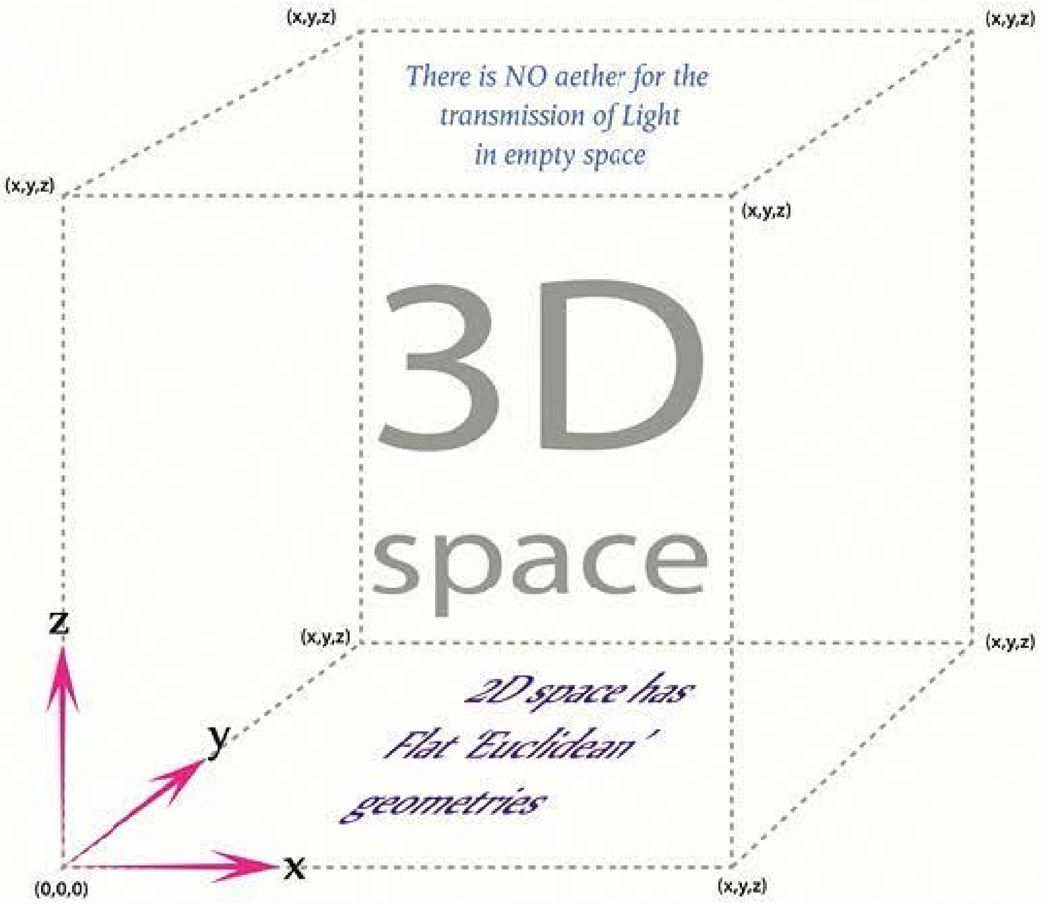

There is NO aether for the transmission of Light in empty space

3D space

2D space has 'Flat Euclidean' geometries

A Spatial region
is defined so as to
measure the physics
of mass-ENERGY-Matter
within its confines

Energy moves through Space
in various forms:
radiant energies, Matter etc

Space can be
Cubic, Spherical or
Tetrahedral as defined by
the spatial co-ordinates
used to define the region

Empty Space is defined as a topology whose volume is devoid of Energy

Copyright ABRAHAM [2008] - all rights reserved

Space-Time co-ordinates

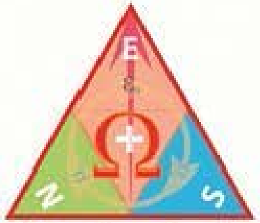

Charge radiates unilaterally
in line with its
vector linear momentum

Vector forces

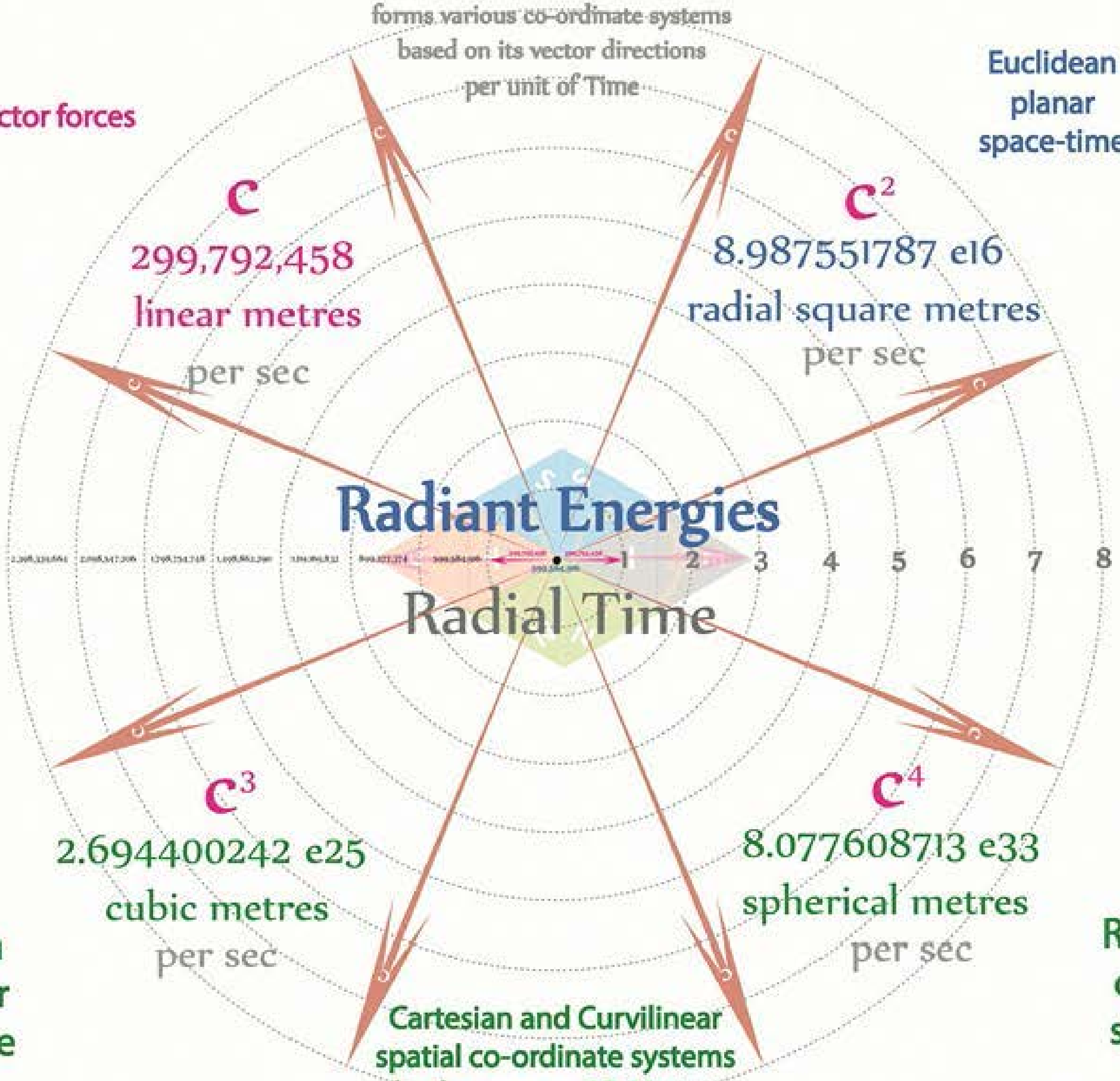

The linear vector of divergent Energy
forms various co-ordinate systems
based on its vector directions
per unit of Time

Radiant Energies

Radial Time

c
299,792,458
linear metres
per sec

c^2
8.987551787 e16
radial square metres
per sec

c^3
2.694400242 e25
cubic metres
per sec

c^4
8.077608713 e33
spherical metres
per sec

Cartesian and Curvilinear
spatial co-ordinate systems
both measure 3D Matter

**Euclidean
planar
space-time**

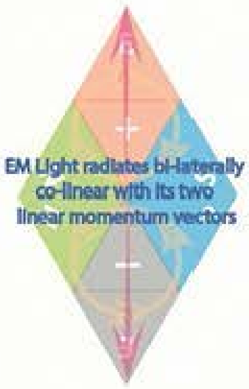

EM Light radiates bi-laterally
co-linear with its two
linear momentum vectors

Time in Physics is a measure
of how long it takes for light to
travel 299,792,458 metres
from its source in a vacuum

$$\frac{299{,}792{,}458 \text{ m}}{\text{second}}$$

$$\frac{\text{metre}}{299{,}792{,}458 \text{ sec}}$$

A metre in Physics is
the distance light travels in
1/ 299,792,458 of a second
from its source

**Cartesian
rectilinear
space time**

**Riemannian
curvilinear
space-time**

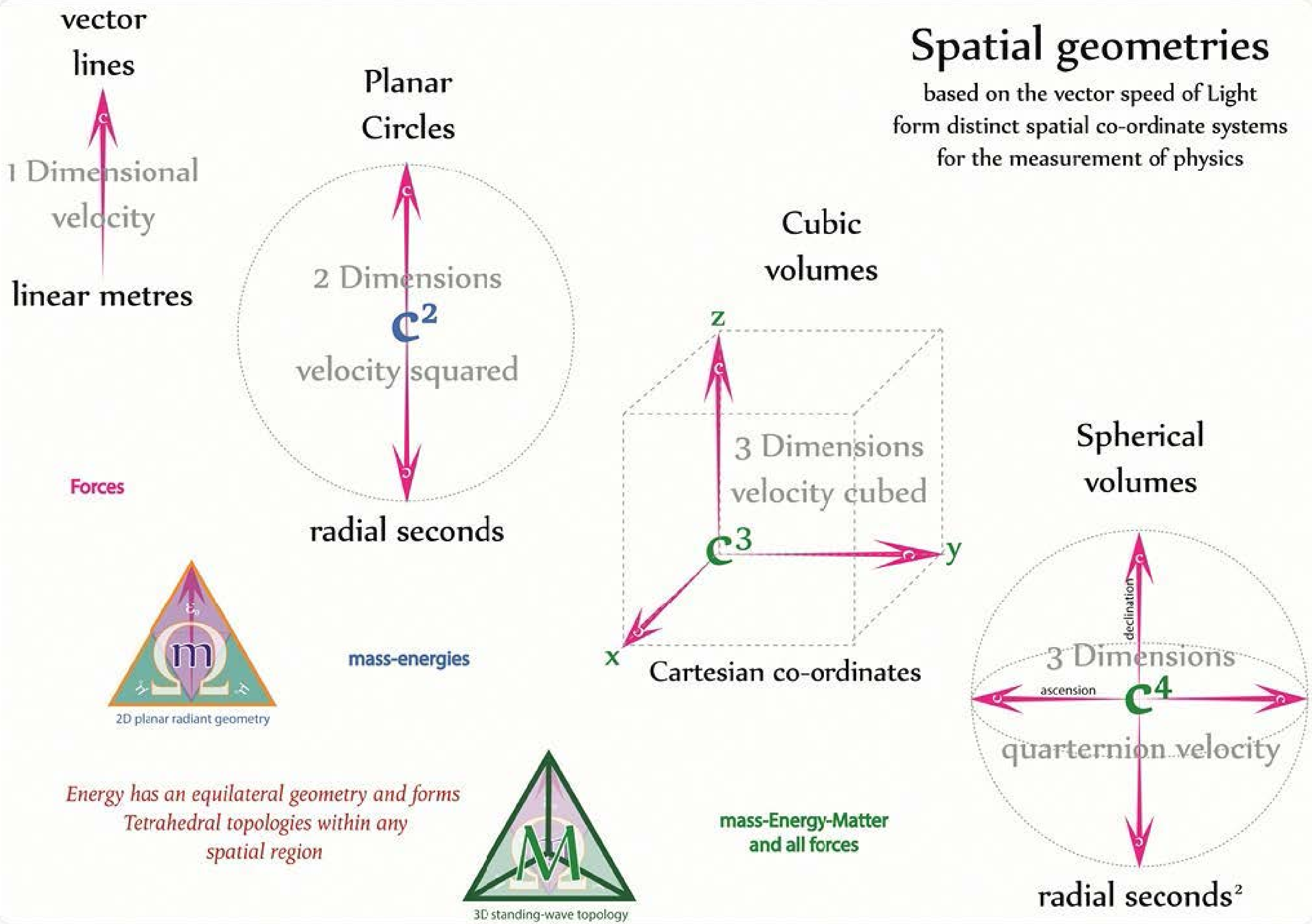

Spatial geometries

based on the vector speed of Light
form distinct spatial co-ordinate systems
for the measurement of physics

vector lines

1 Dimensional velocity

linear metres

Planar Circles

2 Dimensions

c^2

velocity squared

radial seconds

Cubic volumes

3 Dimensions

velocity cubed

c^3

Cartesian co-ordinates

Spherical volumes

3 Dimensions

c^4

quarternion velocity

radial seconds²

Forces

2D planar radiant geometry

mass-energies

Energy has an equilateral geometry and forms
Tetrahedral topologies within any
spatial region

3D standing-wave topology

mass-Energy-Matter
and all forces

Copyright ABRAHAM [2008] - all rights reserved

c is the natural velocity of light

Energy has a 2D
equilateral geometry
[Euclidian]

Time - as a measure
of divergent energies
has a radial geometry

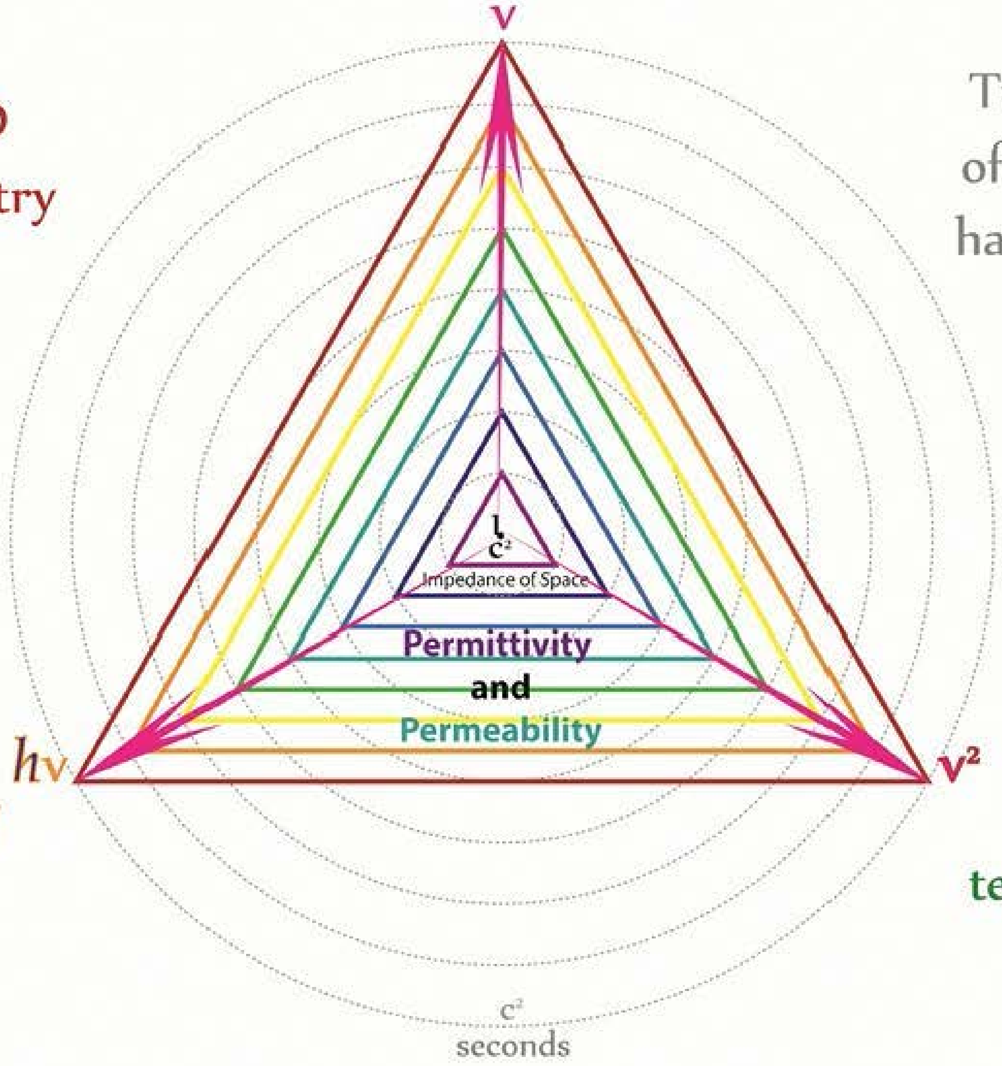
v
$\frac{1}{c^2}$
Impedance of Space
Permittivity
and
Permeability
hv
v²
c²
seconds

Equilateral geometry
is the inverse of
radial geometry

3D Matter has a
tetrahedral mass-energy
topology

The scalar spatial geometry of Energy

Velocity

Velocity is the measurement of the rate and direction of change in the position of an object.

It is a vector physical quantity; both magnitude and direction are required to define it.

The absolute scalar value (magnitude) of velocity is speed, a quantity that is measured in metres per second (m/s or ms–1) when using the SI (metric) system.

$$\frac{m}{s}$$

Speed is the scalar value of the Distance traveled per unit of TIme

$$\bar{\mathbf{v}} = \frac{\Delta \mathrm{x}}{\Delta t}.$$

Velocity is the vector value of the Distance traveled per unit of Time

$$\frac{\overrightarrow{m}}{s}$$

v Velocity $\frac{m}{s}$

is a 2D RADIAL SPACE-TIME co-ordinate system

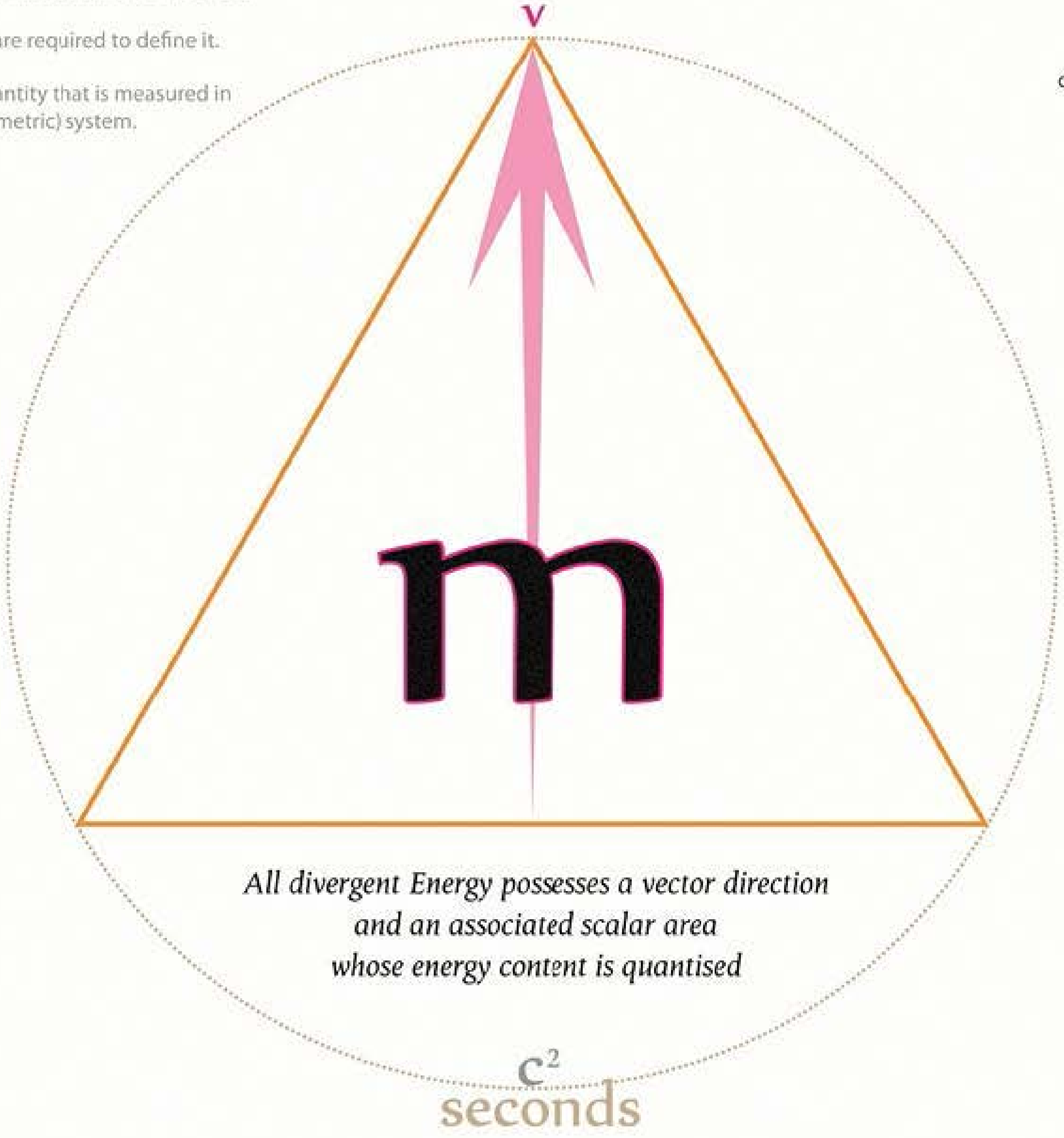

All divergent Energy possesses a vector direction and an associated scalar area whose energy content is quantised

Copyright ABRAHAM [2008] - all rights reserved

is a 3D SPHERICAL
SPACE-TIME
co-ordinate system

a acceleration $\frac{m}{s^2}$

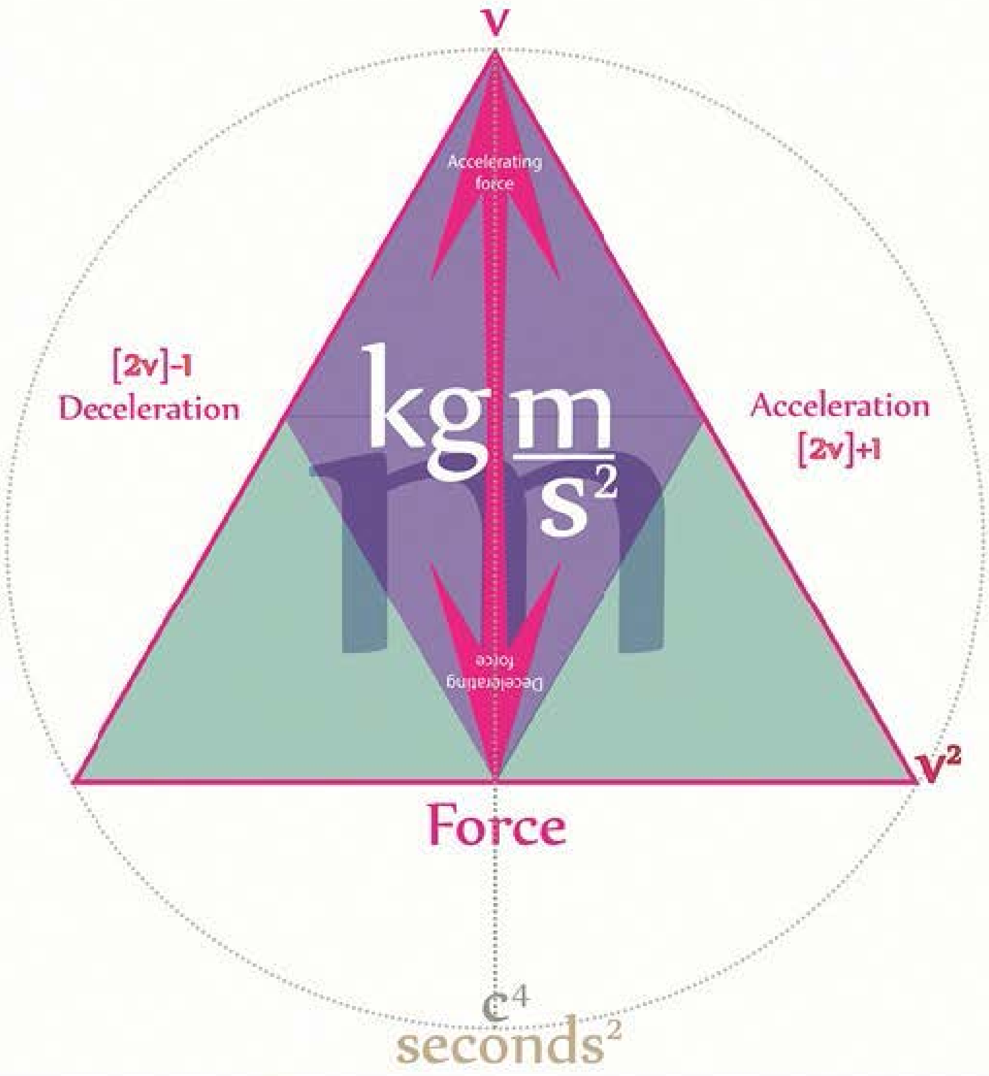
v
Accelerating force
kg m/s²
m
Decelerating force
[2v]-1
Deceleration
Acceleration
[2v]+1
v²
Force
c⁴
seconds²

Acceleration

In physics, acceleration is the rate of change of velocity [dv] over time [dt]

In one dimension, acceleration is the rate at which something speeds up or slows down.

However, since velocity is a vector, acceleration describes the rate of change of both the magnitude and the direction of velocity.

Acceleration has the dimensions [Length]/[Time Squared]
In SI units, acceleration is measured in meters per second squared (m/s^2).

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

In classical mechanics, for a body with constant mass,
the acceleration of the body is proportional to the net force acting on it
(Newton's second law)

$$F=ma \longrightarrow a=F/m$$

$$\overset{\text{Force}}{kg\,\frac{m}{s^2}}$$

Additionally, for a mass with constant velocity,
(ie in an inertial frame)
the energy of motion is expressed as its momentum
(acceleration causes changes in Energy-momentum)

$$p = kg\,\frac{m}{s}$$

Copyright ABRAHAM [2008] - all rights reserved

Quantised Angular momentum

As it is a physical [equilateral] geometry QAM is conservative in any system where there are no external Forces and serves as the foundational geometric source for all the conservation laws of physics

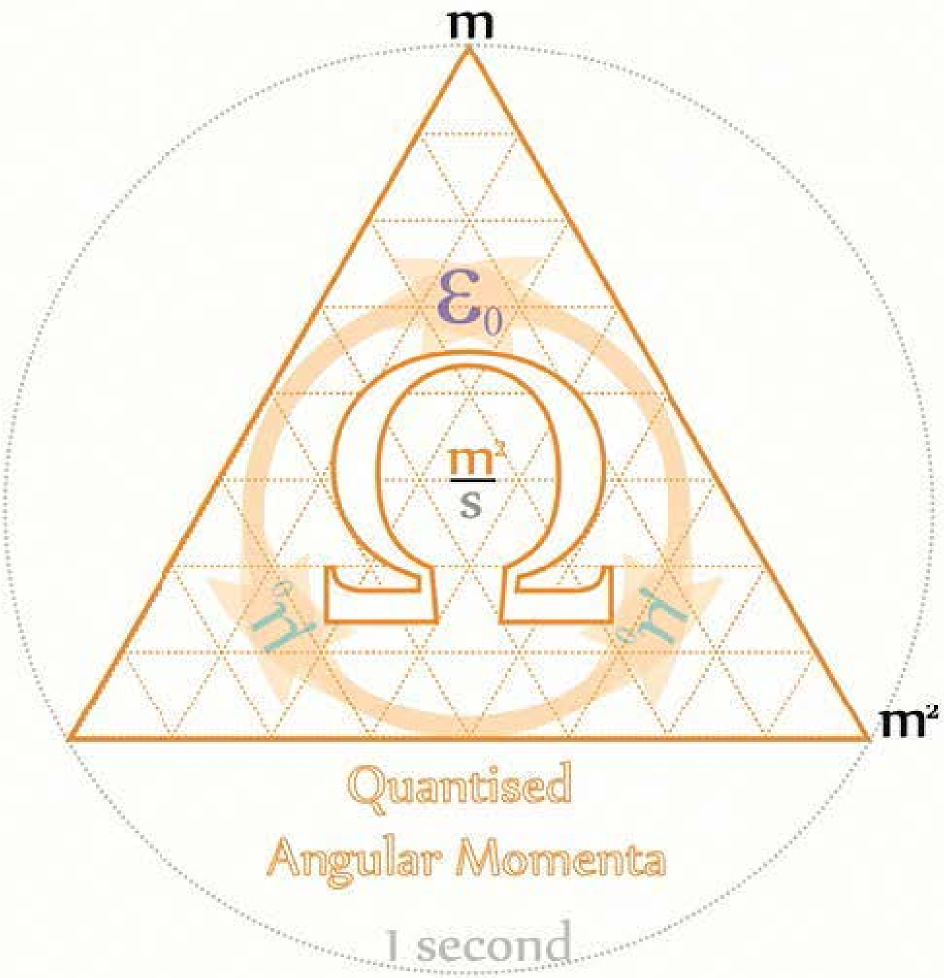
m
ε_0
Ω
$\frac{m^2}{s}$
m²
Quantised
Angular Momenta
1 second

Conservation of Quantised Angular momenta

In QFT, angular momentum is is considered to be the rotational analog of linear momentum, in Tetryonics it is revealed to be the equilateral geometry of mass-energy within any defined spatial co-ordinate system

A major re-definition of quantised angular momenta in physics is revealed

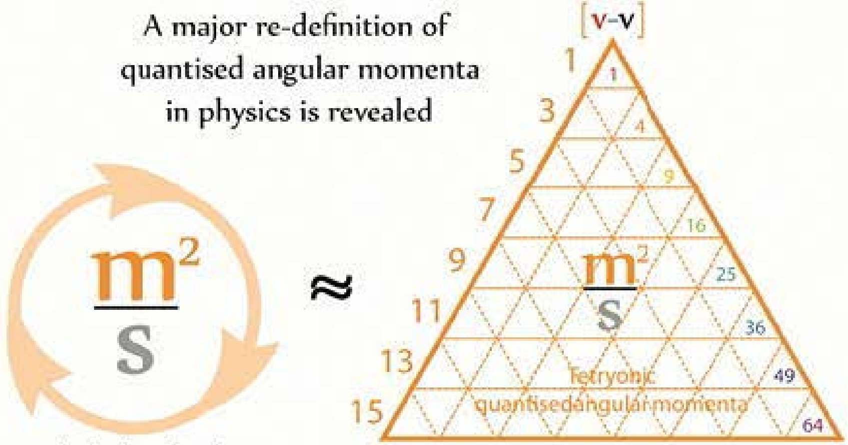
$\frac{m^2}{s}$
≈
[v-v]
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64
$\frac{m^2}{s}$
Tetryonic quantised angular momenta

classical rotational angular momenta

Tetryonic quantum mechanics

In quantum mechanics, angular momenta is quantised – that is, it cannot vary continuously, but only in ODD number "quantum steps" between the allowed SQUARE nuclear Energy levels

In physics, angular momentum, moment of momentum, or rotational momentum is a conserved vector quantity that can be used to describe the overall state of a physical system.

When applied to specific mass-energy-Matter systems QAM reveals the true quantum geometry and nature of Energy in our universe

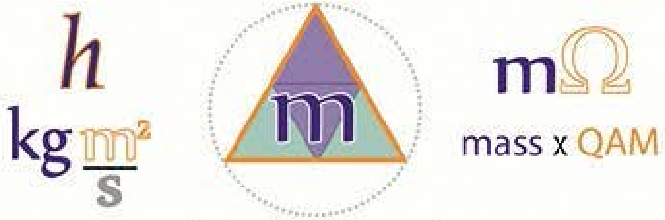
h
$kg\,\frac{m^2}{s}$
m
mΩ
mass x QAM

Planck's Constant

Normally viewed as an expression of rotational momentum Quantised Angular Momentum [QAM] is in fact a result of the equilateral geometric quantization of mass-energy

Copyright ABRAHAM [2008] - all rights reserved

mass - Energy geometry

Newton developed his
Laws using Force
$p = mv$
(linear momentum)

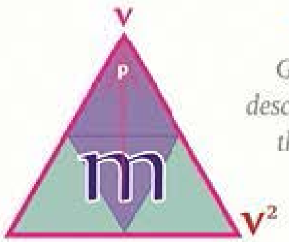

Gottfried Leibniz first
described Scalar Energy as
the square of velocity

$E = mv^2$

Leibnitz mass-energy equivalence

mass velocity
squared

$kg\frac{m^2}{s^2}$

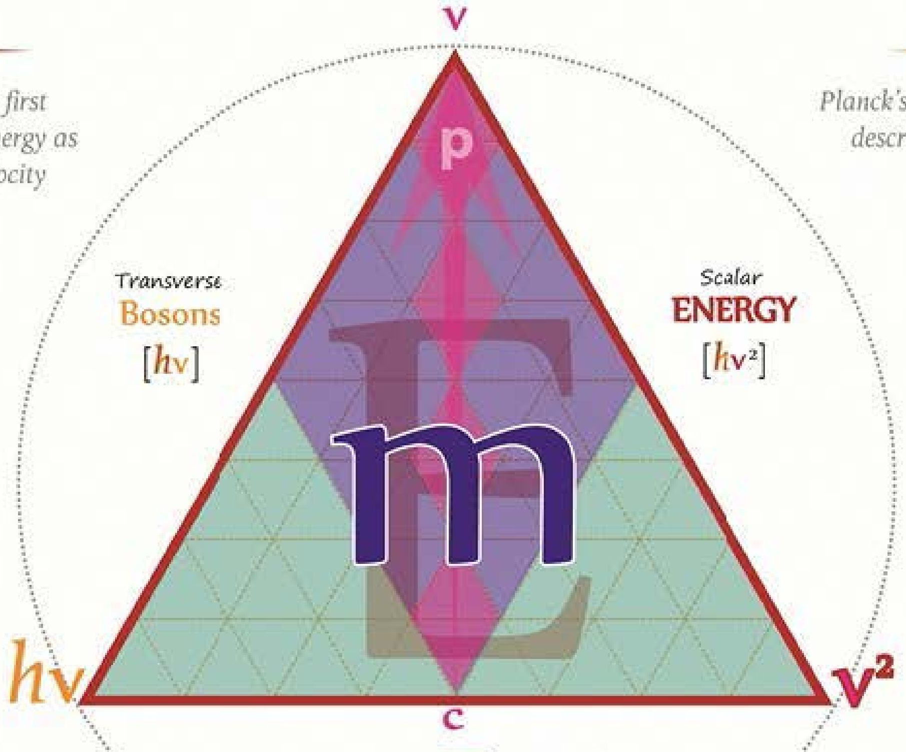

$mv^2 = E = hv^2$

Leibniz-Planck mass-Energy equivalence

Planck developed his Heat
Law using quantised mass
$h = m\Omega$
(Planck Constant)

Planck's equation for heat energy
describes transverse masses
[Bosons]

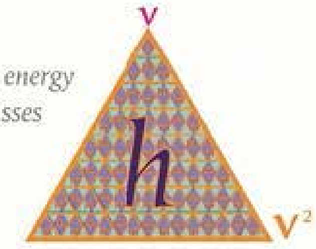

$E = nhv$

Planck mass-energy equivalence

quantised mass
per second

Planck quanta per second

$\left[kg\frac{m^2}{s}\ \frac{1}{s}\right]$

scalar energies can be related to velocity-momenta
through mass geometries

	scalar		*quantised*	
mv^2	Leibniz-Newton mass-Energy equivalence	$m\Omega v^2$	Tetryonic-Planck mass-Energy equivalence	hv^2

mass-Energy Forms

Scalar/Linear forms

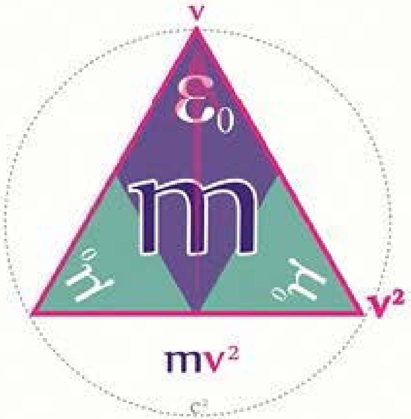

mass x velocity squared

$$kg.\left[\frac{m}{s}\right]^2$$

Energy momentum

Leibniz [and Newton] showed that the Energy of a system can be viewed as a product of its mass x velocity squared or [equally as linear momentum squared], forever linking Energy to velocity through the scalar property of mass

$$mv^2 = E = p^2$$

Energy is the ability to do work in varying forms such as potential, kinetic, & mechanical energies, work, heat, and chemical or electrical energies.

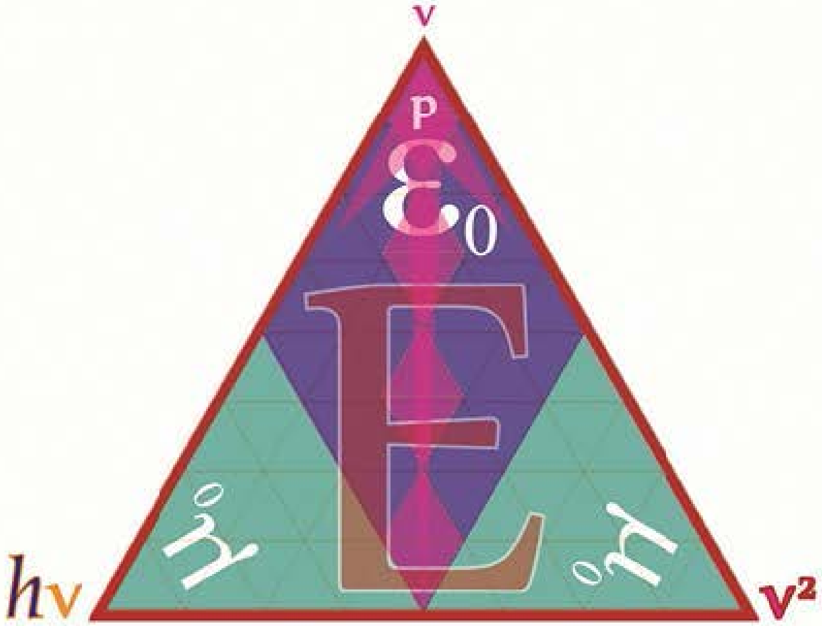

$$\boxed{mv^2 = E = hv^2}$$

The total energy contained in an object is identified with its EM mass, and Energy (like mass), cannot be created or destroyed

Tetryonic reveals mass to be a scalar measurement of quantised [equilateral] energy per unit of Time

$$m[\Omega]v^2$$

Energy is subject to the law of conservation

Quantised form

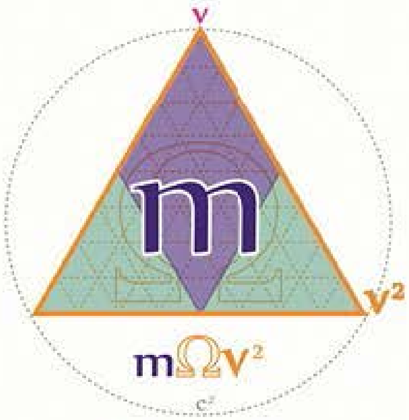

Planck quanta per second

$$\left[kg.\frac{m^2}{s}\right].s^{-1}$$

mass-Energy quanta

Max Planck reveealed that energy was not continuous, it was quantized – only certain energies are allowed. Continuous energy is a scalar propertyo of mass-energy and its quantisation is the result of its equilateral geometry

$$E = hv^2$$

In quantum mechanics energy is defined in terms of the energy operator as a time derivative of the wave function

Copyright ABRAHAM [2008] - all rights reserved

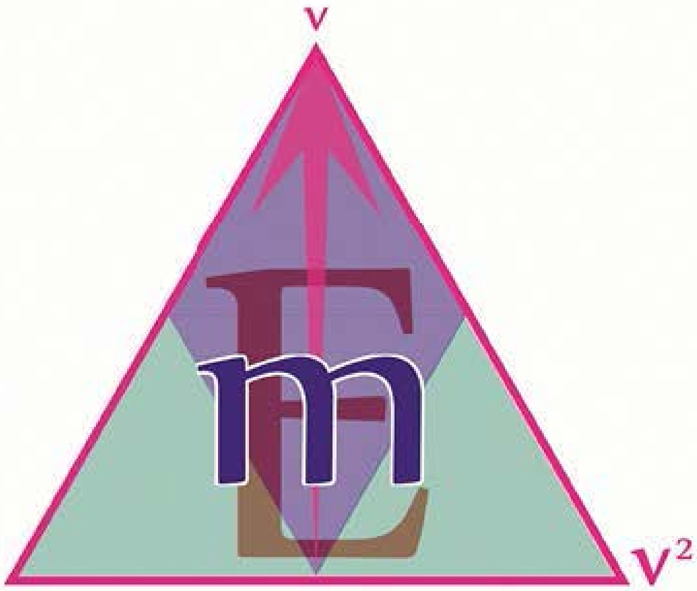

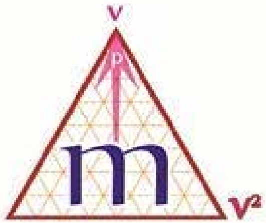

Planck quanta

$$\left[\left[m\Omega v\right]^2\right]$$

mass velocity

linear momentum squared

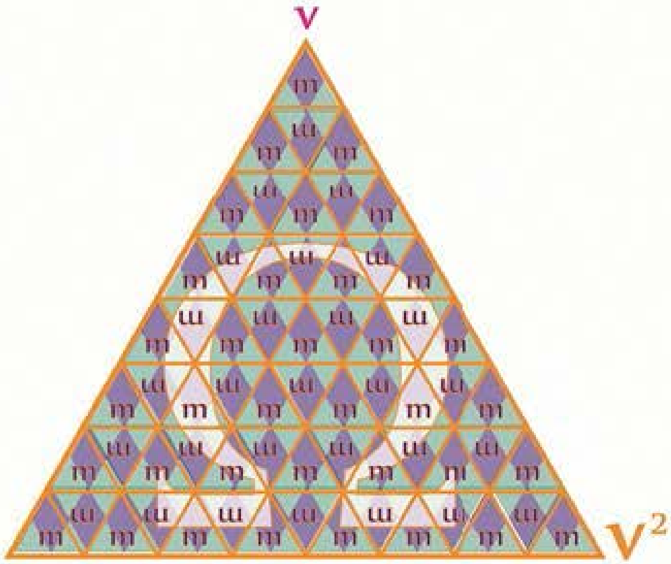

Energy is mass-velocity squared

Everything in our Universe results from the equilateral geometry of quantised mass-energies

Energy is Planck-quanta squared

$$E = mv^2$$

Quantised Energy

$$E = hv^2$$

scalar mass

$$kg$$

Planck quanta per second

$$\left[\left[kg\,\frac{m^2}{s}\right]\frac{1}{s}\right]$$

mass velocity squared

quantised mass

$$kg\,\frac{m^2}{s}$$

$$m = \frac{E}{v^2}$$

Scalar Energy

$$h = \frac{E}{v^2}$$

Note: There is a direct correspondence between Velocity and Planck quanta numbers
(ie as velocity varies, the energy quanta varies as well by the square of the linear change)

Copyright ABRAHAM [2008] - all rights reserved

p Momentum $kg\frac{m}{s}$

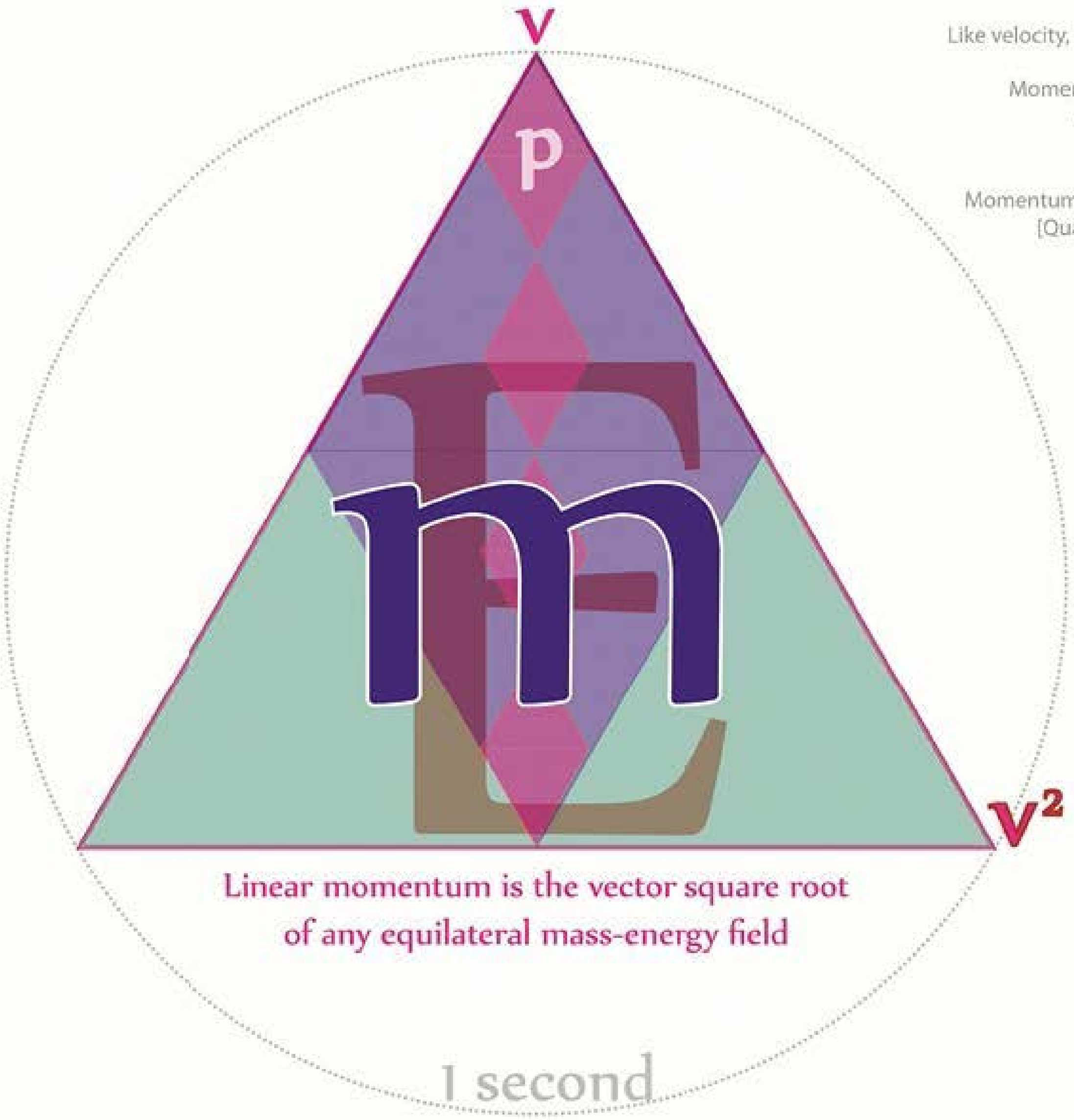

Linear momentum is the vector square root
of any equilateral mass-energy field

Linear Momentum

In classical mechanics, momentum (pl. momenta; SI unit kg·m/s, or, equivalently, N·s) is the product of the mass and velocity of an object (p).

Like velocity, momentum is a vector quantity, possessing a direction as well as a magnitude.

Momentum is a conserved quantity (law of conservation of linear momentum), meaning that if a closed system is not affected by external forces, its total momentum cannot change.

Momentum should be referred to in its specific forms to distinguish it in its various forms [Quantised Angular, Linear, Rotational and quantum/nuclear momentum]

$$\boxed{p = \frac{hv^2}{v} = mv}$$

Although originally expressed in Newton's Second Law, the conservation of momentum also holds in special relativity and, with appropriate definitions, a (generalized) momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity.

In relativistic mechanics, non-relativistic momentum is further multiplied by the Lorentz factor.

$$p^2 = E = mv^2$$

Energy can be expressed as
the square of linear momentum

$$kg\frac{m^2}{s^2}$$

Copyright ABRAHAM [2008] - all rights reserved

Ω **Just as Tetyonic geometry distingushes between angular momenta and linear momentum it also distingushes between linear momentum and the vector velocities it produces** p

Linear Momentum

$p = mv$

Velocity

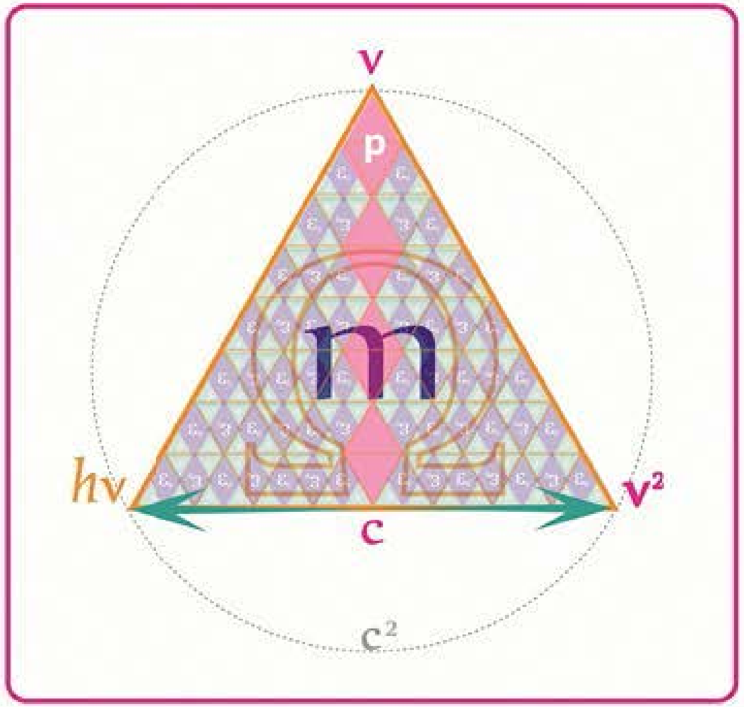

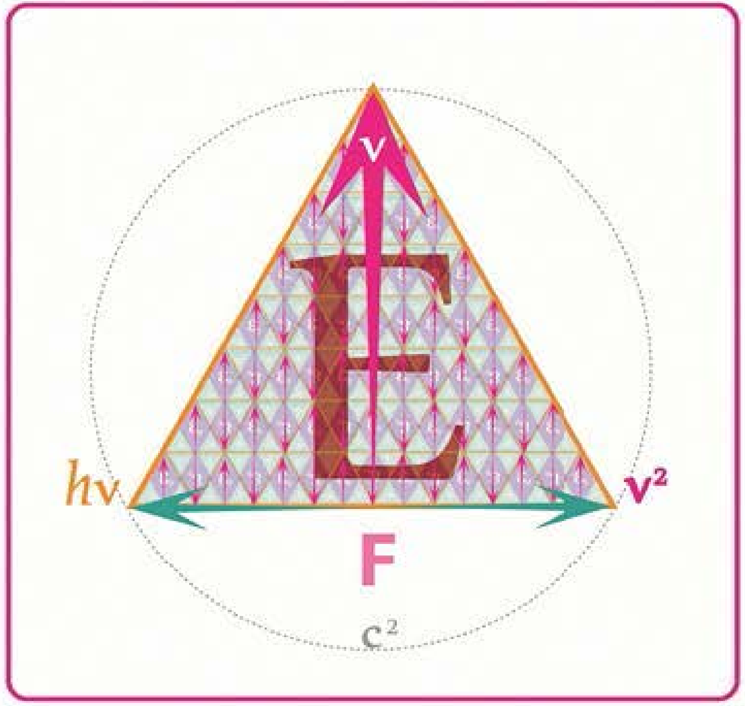

$E = pv$

$E = p^2$

SCALAR square root

Scalars are quantities that are fully described by only their magnitude

$E = mv^2$

VECTOR square root

Vectors fully describe both the magnitude and direction.

p **linear momentum is a scalar component of all equilateral mass-energies that produce vector velocities** v

Copyright ABRAHAM [2008] - all rights reserved

Velocity-Quanta equivalence

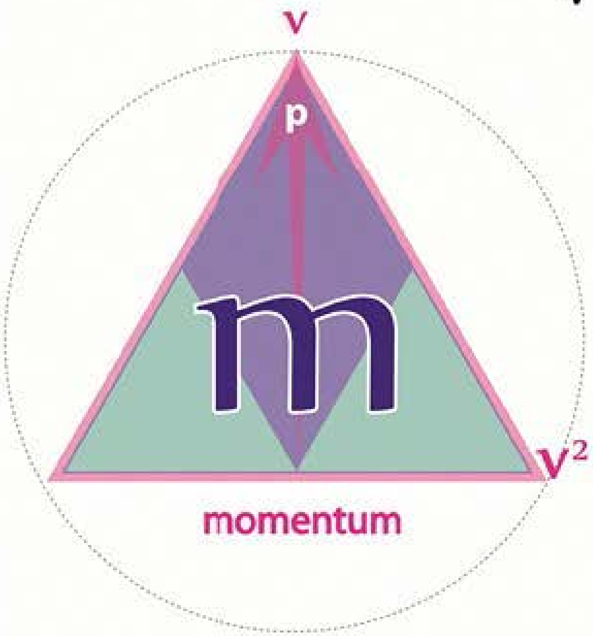

Classically, the Energy of massive bodies was determined using the Newtonian mass-velocity relationship

but most recently Quantum mechanics was developed utilising the Planck's quantised Energy relationship

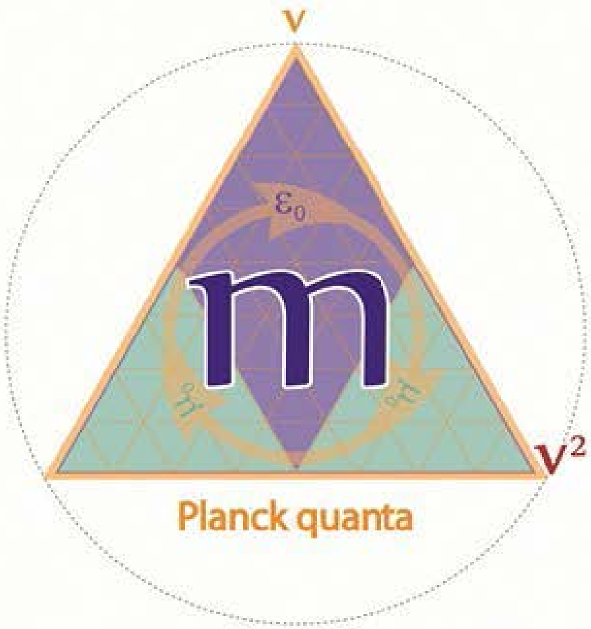

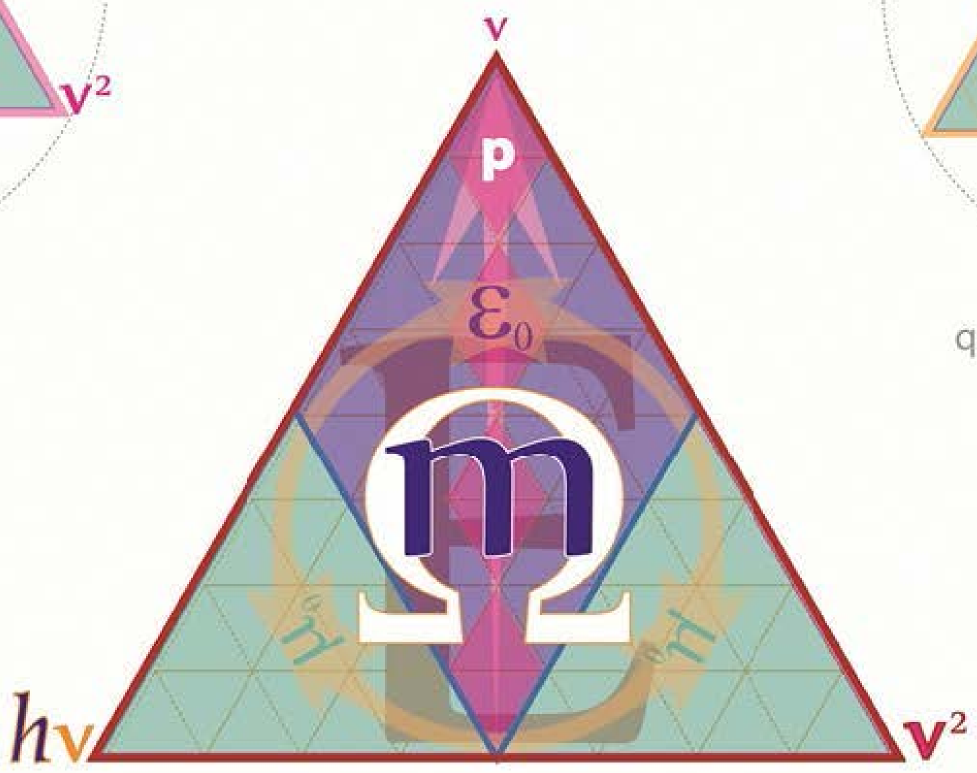

classical vector force

mv^2

mass
linear velocity
squared

$kg\left[\frac{m}{s}\right]^2$

quantised energy momenta

hv^2

mass
angular momenta
per second

$kg\left[\frac{m^2}{s}\frac{1}{s}\right]$

The EM mass-Energy relationship can be revealed either by linear or angular momentum analysis

Energy-momentum relationship

The total number of Planck quanta [mass-angular momenta] in any physical system
is directly related to the square of its linear momentum [mass-velocity]

$$h = \frac{E}{v^2}$$

$$m = \frac{E}{v^2}$$

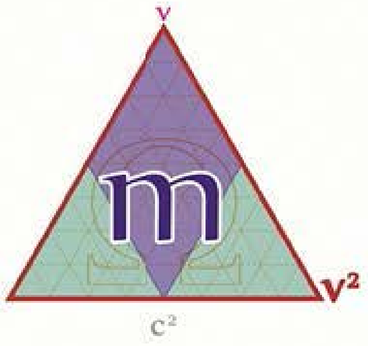

Quantised Energy-momentum

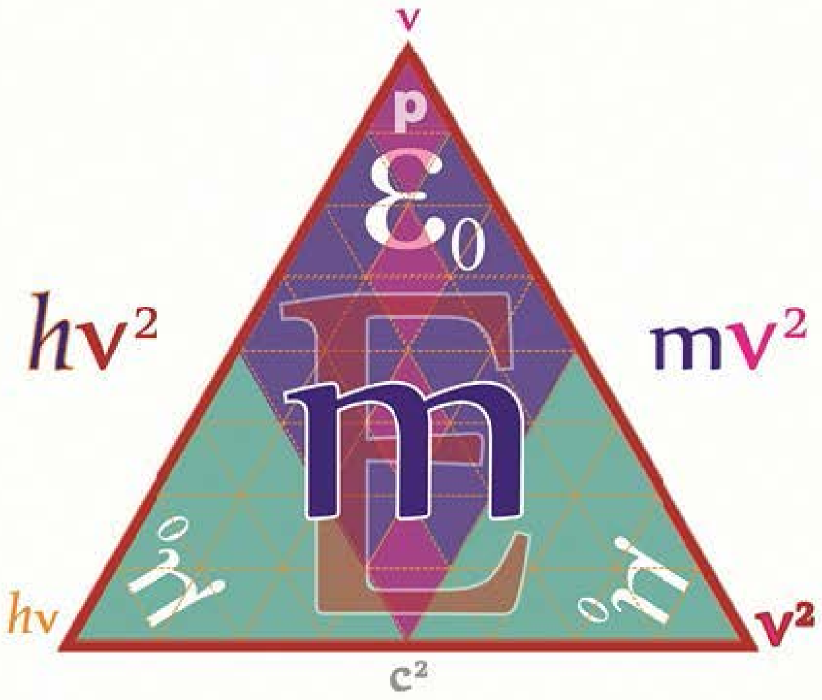

mass-energy momenta

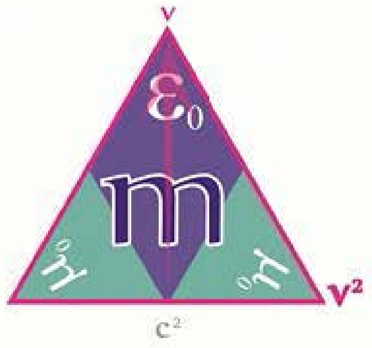

Linear Energy-momentum

$$E = m\Omega v^2$$

Quantised Angular Momenta
is an equilateral geometry

The omega geometry of Energy produces the direct relationships between Planck's constant-quanta and mass-Energy-momentum of any spatio-temporal co-ordinate system

$$E = pc$$

Linear momentum
is a vector Force

$$m\Omega v^2 = E = mv^2$$

mass is a derived physical property relating Energy momenta to Velocity

Copyright ABRAHAM [2008] - all rights reserved

CHARGE

Charge is a measure of mass.QAM/second
[the equilateral geometry of Energy]
that gives form to all physics

ElectroMagnetic Charge
is a quantum property
resulting from the equilateral
QAM geometry of mass-Energy

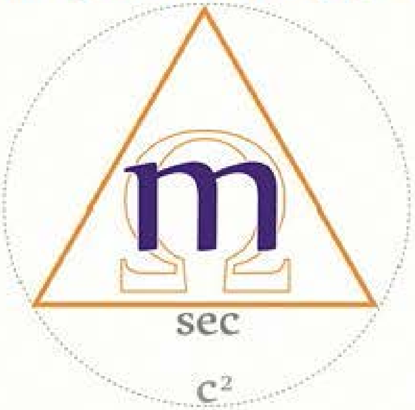

The two ElectroMagnetic charge
geometries possible can be modelled
by the flux of electrical energy
in ideal inductive loops

It is a measure of the arrangment of Planck quanta
geometries/topologies within any specific
space-time co-ordinate system

q

Clockwise inductive
energy flux

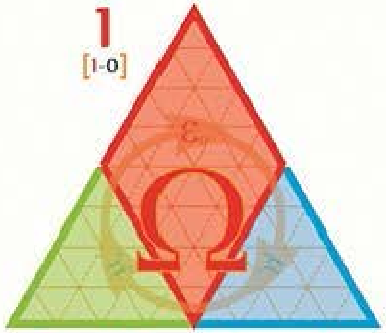

Positive charged
mass-energy momenta

$$\frac{hv}{c^2}$$

$$kg\,\frac{m^2}{s}\,\frac{s^2}{m^2}$$

$$kg.s$$

1.33518 e-20 s

Counter clockwise
energy flux

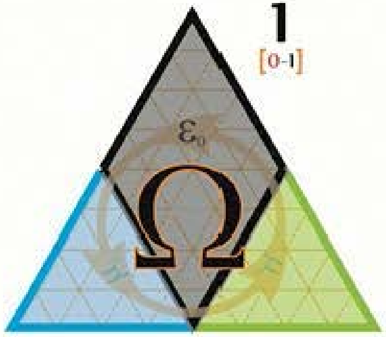

Negative charged
mass-energy momenta

q

Measuring charge geometries

Charge comes in two types, called Negative and Positive
which create the Law of Interaction (historically, the Law of Attraction)

q

quantised Charge

[v] [v]

Q

nett Charge

[v-v]

All charged masses seek equilibrium

Charge geometry is velocity invariant

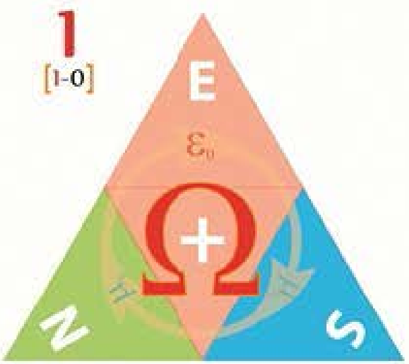

Clockwise Energy flux

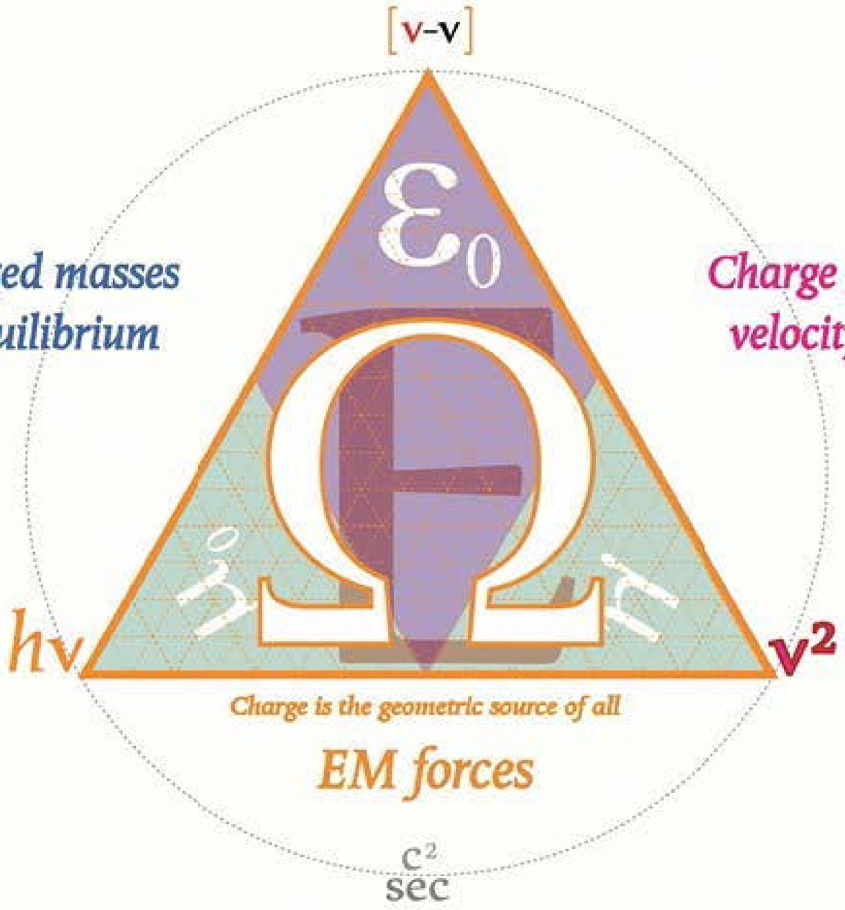

Charge is the geometric source of all

EM forces

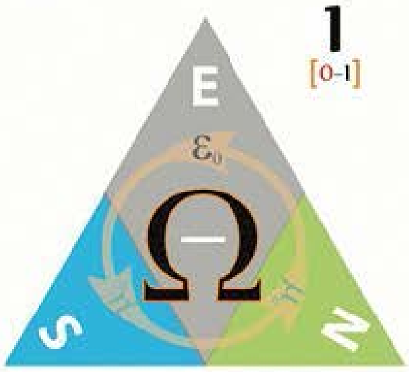

anti-Clockwise Energy flux

Historically defined as a physical property of mass-Matter that causes it to experience a force
charge is actually the result of the equilateral geometry of qantised mass angular momenta
which provides the physical mechanics of mass-ENERGY-Matter differentiation and interaction

The SI unit of quantity of charge is the Coulomb, $C \approx$ mass seconds $\approx$ kg.s which can also be expressed as Amp-seconds

Copyright ABRAHAM [2008] - all rights reserved

EM Field Geometry

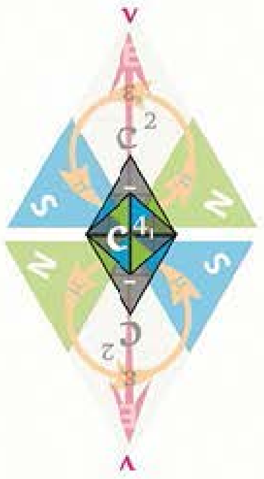
v
c²
S
N
c⁴
N
S
c²
v

mass-energies are 2D radiant EM field geometries

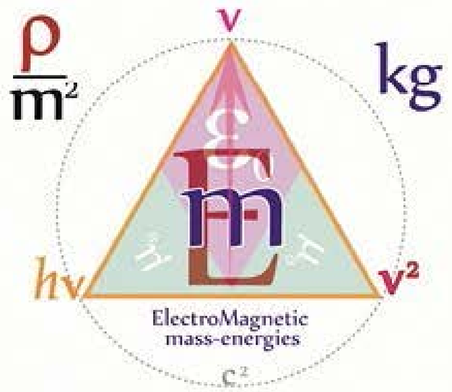
ρ/m²
v
kg
ε
m
E
hv
v²
ElectroMagnetic mass-energies
c²

[299,792,458 m/s]

'c' forms a CONSTANT of proportionality for different spatio-temporal co-ordinate systems used to measure mass-ENERGY-Matter

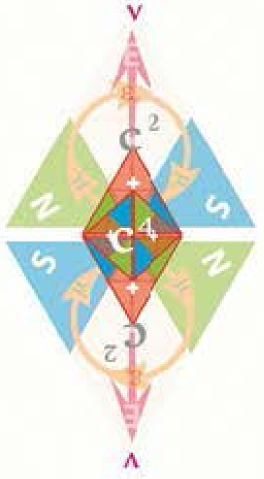
v
c²
N
S
c⁴
S
N
c²
v

All Matter are 3D EM standing wave topologies

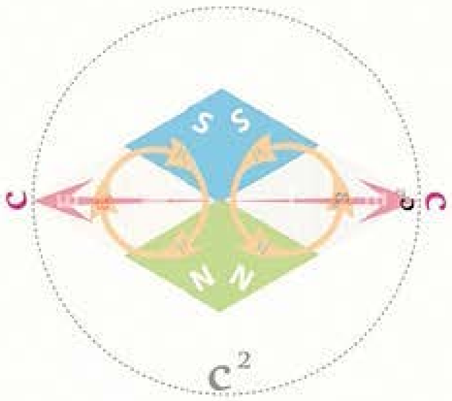

EM fields create 'interaction-at-a-distance'

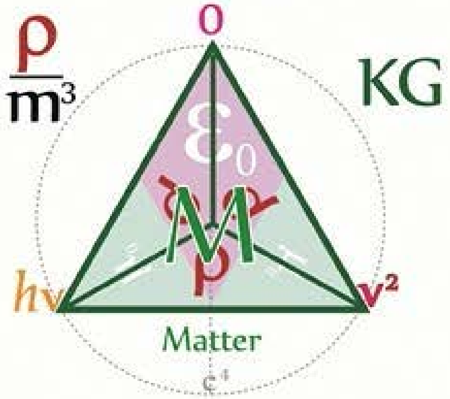
ρ/m³
0
KG
ε₀
M
ρ
hv
v²
Matter
c⁴

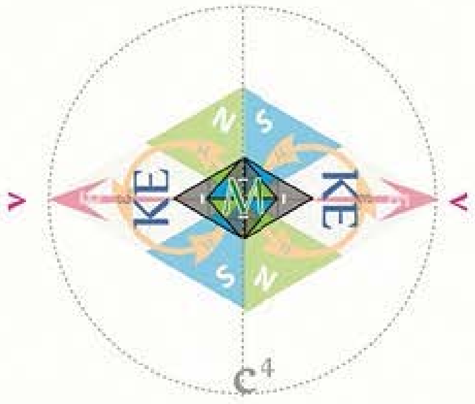
v
KE
N S
M
S N
KE
v
c⁴

Electrostatic Matter has opposing 2D KEM fields

Copyright ABRAHAM [2008] - all rights reserved

Energy quantisation

All mass-Energy-Matter can only have certain integer Energy-momenta,
[mass-Energy in all its forms is QUANTISED]

q

Net quanitised angular momenta
[inductive energy flux]
determines Charge

$$m\left[\frac{\Omega}{c^2}\right]$$

$$kg\,\frac{m^2}{s}\cdot\frac{s^2}{m^2}$$

seconds

Transverse quanta
create Quantum Levels

The equilateral geometry of
quantised angular momentum
creates charged masses

Planck's Constant is in fact
mass x equilateral QAM

$$kg.\frac{m^2}{s}$$

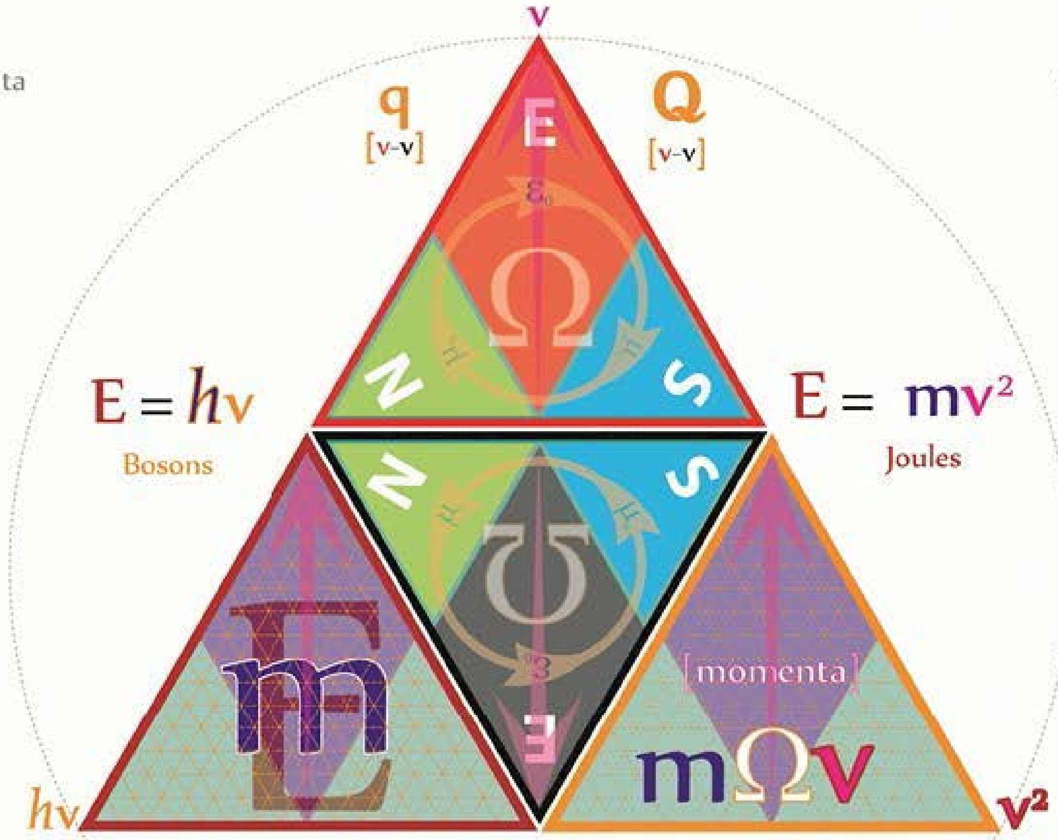

$E = hv$

Bosons

$E = mv^2$

Joules

$$kg.\frac{m^2}{s}$$

Planck's Constant

n.Planck quanta per second is Energy

q

Net quanitised angular momenta
[inductive energy flux]
determines Charge

$$m\left[\frac{\Omega}{c^2}\right]$$

$$kg\,\frac{m^2}{s}\cdot\frac{s^2}{m^2}$$

seconds

Scalar (nett) quanta
create Square Energies

Planck's Constant
can be described in
a number of differing ways

Scalar Energy x Time $kg.\frac{m^2}{s^2}.s$

Linear momentum x Distance $kg.\frac{m}{s}.m$

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Mnemonics

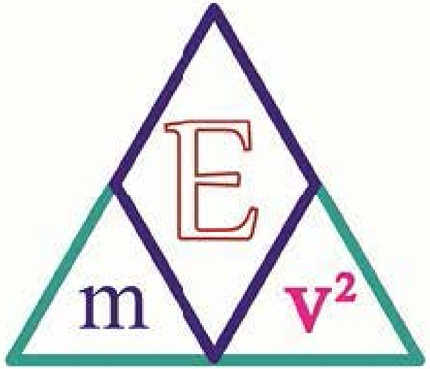

Energy momentum

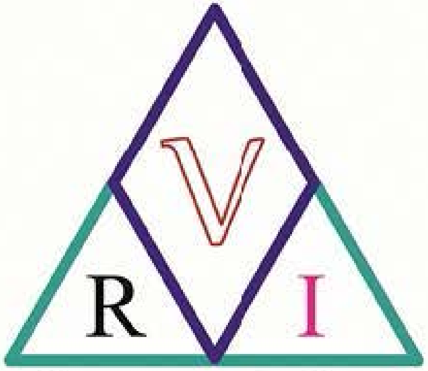

Electrical flow

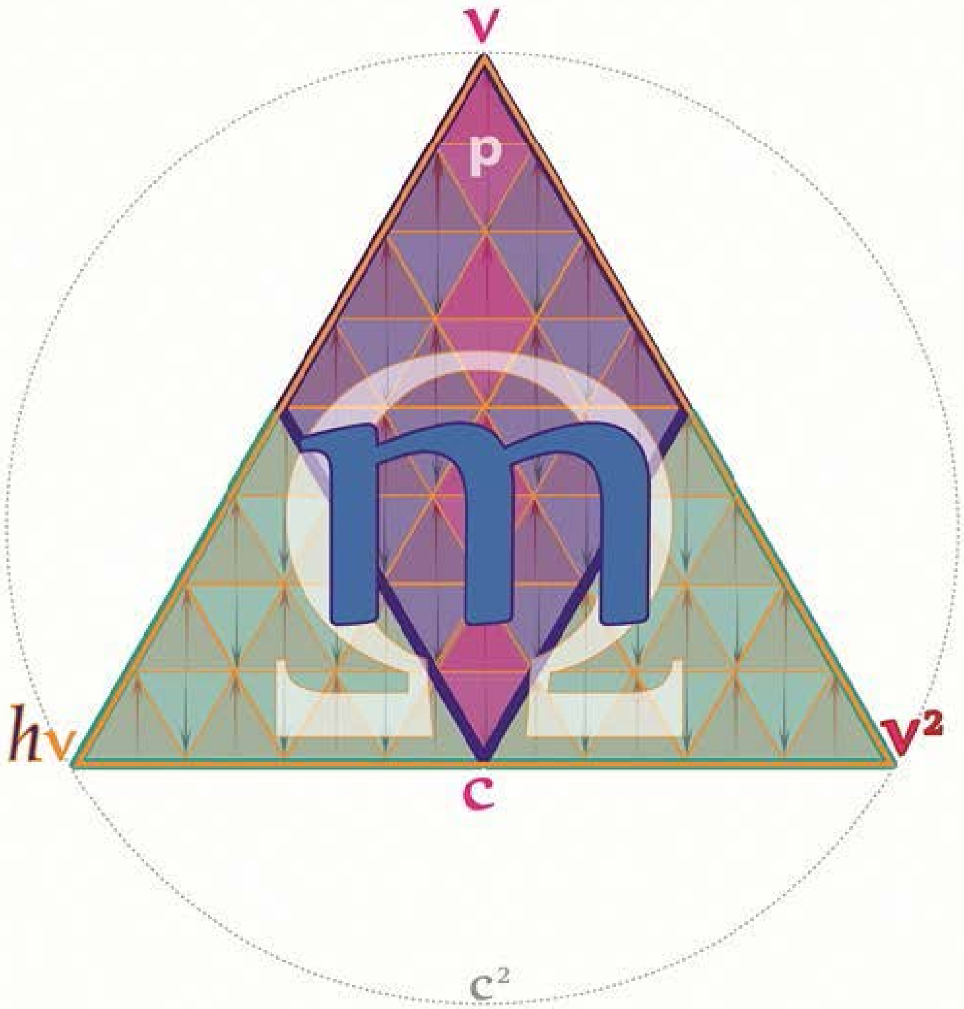
v
p
m
Ω
hv
c
v²
c²

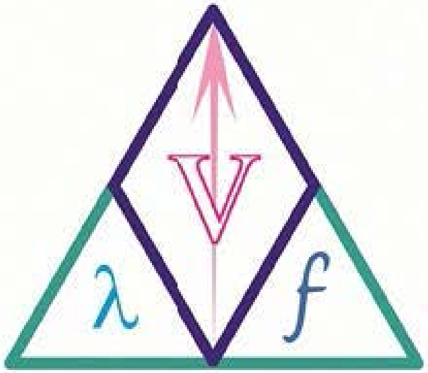

velocity of propagation

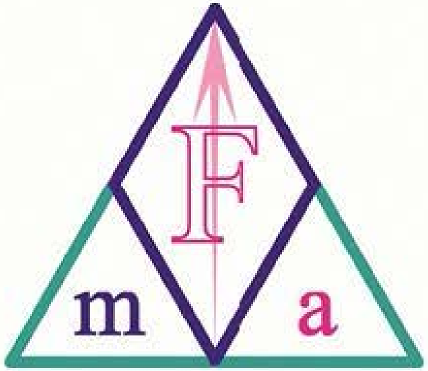

Inertial Force

Many physical relationships are can be represented with DELTA mnemonics

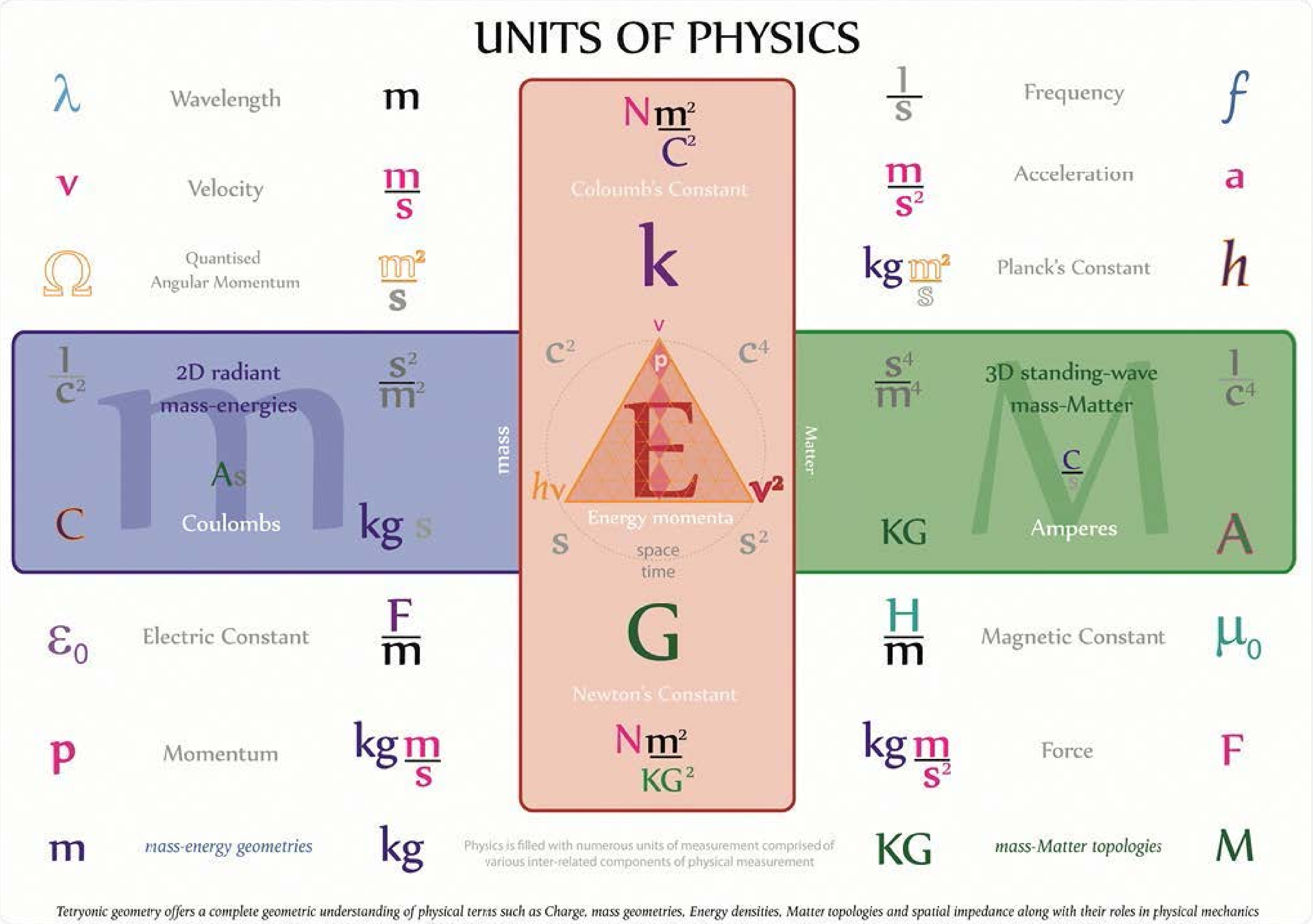

UNITS OF PHYSICS

λ	Wavelength	m
v	Velocity	$\frac{m}{s}$
Ω	Quantised Angular Momentum	$\frac{m^2}{s}$
$\frac{1}{c^2}$	2D radiant mass-energies	$\frac{s^2}{m^2}$
C	As Coulombs	kg s
ε_0	Electric Constant	$\frac{F}{m}$
p	Momentum	$kg\frac{m}{s}$
m	*mass-energy geometries*	kg

$N\frac{m^2}{C^2}$

Coloumb's Constant

k

G

Newton's Constant

$N\frac{m^2}{KG^2}$

Physics is filled with numerous units of measurement comprised of various inter-related components of physical measurement

$\frac{1}{s}$	Frequency	f
$\frac{m}{s^2}$	Acceleration	a
$kg\frac{m^2}{s}$	Planck's Constant	h
$\frac{s^4}{m^4}$	3D standing-wave mass-Matter	$\frac{1}{c^4}$
KG	$\frac{C}{s}$ Amperes	A
$\frac{H}{m}$	Magnetic Constant	μ_0
$kg\frac{m}{s^2}$	Force	F
KG	*mass-Matter topologies*	M

Tetryonic geometry offers a complete geometric understanding of physical terms such as Charge, mass geometries, Energy densities, Matter topologies and spatial impedance along with their roles in physical mechanics

Tetryonics 01.20- Physical Units

Copyright ABRAHAM [2008] - all rights reserved

Scalar mass-energies

have an EQUILATERAL geometry
(60°- 60°- 60°)

$\frac{\sqrt{3}}{2}$ **s**

The Area of an Equilateral triangle

½ base x height

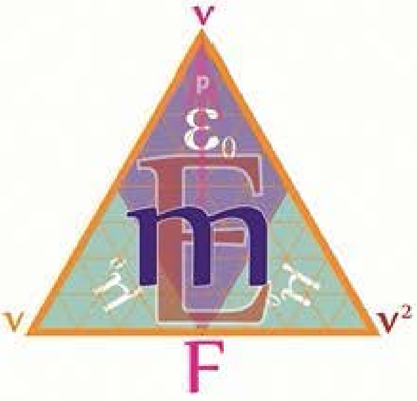
v
p
ε_0
E
m
v
v^2
F

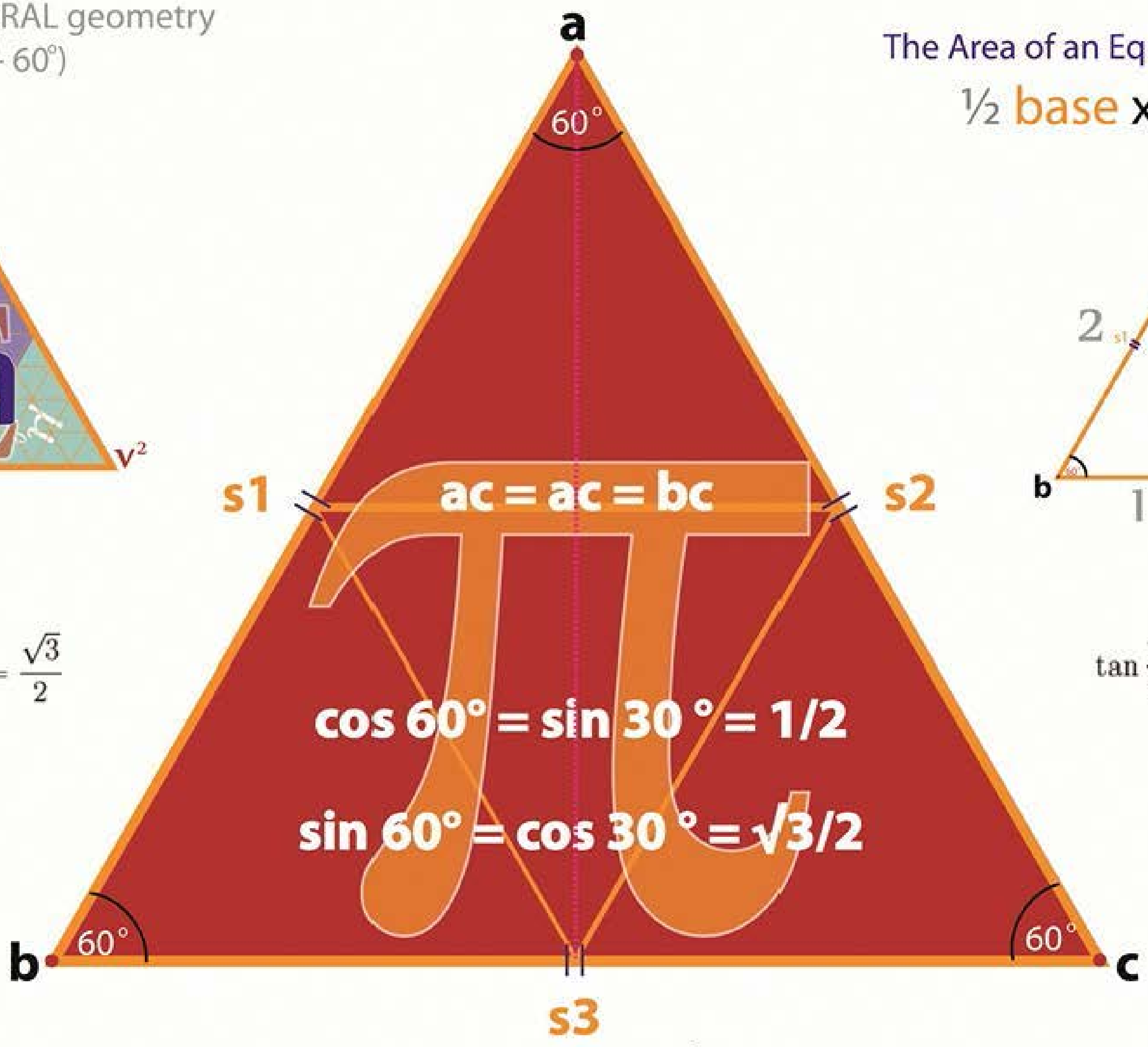
a
60°
s1
s2
b
60°
c
60°
s3

$ac = ac = bc$

$\cos 60° = \sin 30° = 1/2$

$\sin 60° = \cos 30° = \sqrt{3}/2$

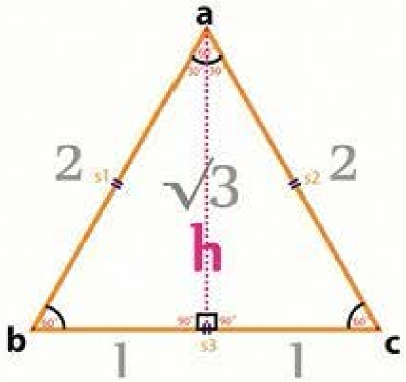
a
2
2
$\sqrt{3}$
h
b
c
1
1
s1
s2
s3

$$\sin\frac{\pi}{3} = \sin 60° = \frac{\sqrt{3}}{2}$$

$$\tan\frac{\pi}{3} = \tan 60° = \sqrt{3}$$

$$\cos\frac{\pi}{3} = \cos 60° = \frac{1}{2}$$

Pythagorian geometry

Energy geometries within Physics including Special Relativity and Lorentz corrections have historically been incorrectly illustrated as having the geometry of right angled triangles

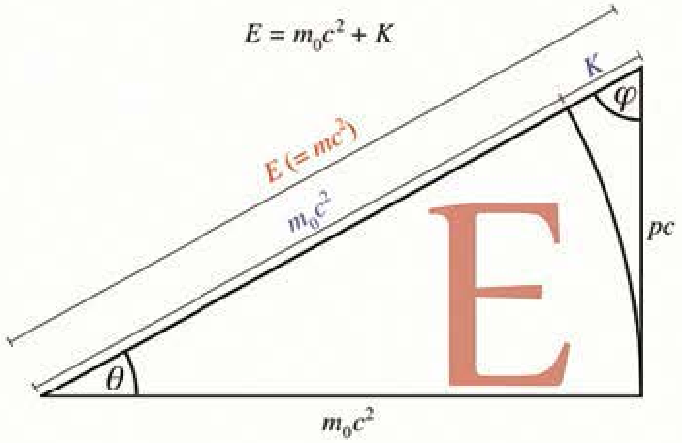

$$E^2 = p^2c^2 + m_0^2c^4$$

Generalizing, we see that the square of the total mass-energies is the sum of the components squared.
[shown incrorectly formulated in this above equation]

We can see an origin of distance in spacetime relating to velocity in pc in which Energy is subject to Lorentz corrections [v/c]

$$E = pc.$$

Additionally, EM mass can be directly related to the energy content of a body by the velocity of Energy

$$E = mc^2$$

Physics is geometry,
one cannot be separated from the other

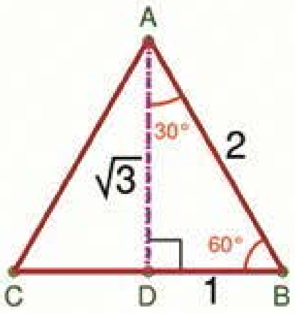

$$mv^2 = E = hv^2$$

There are three ways to look at geometry – mathematically, verbally, and visually,

Of the three, Visually will be shown to be superior leading to intuitive understandings of Physics, Chemistry, Electrodynamics and Gravitation along with all their related physical attributes

6.629432672 e-34 J

Planck quanta

$$[m\Omega v]^2$$

mass velocity

7.376238634 e-51 kg

$$mv = p$$

Tetryonic geometry

The source of all the physical relationships of mass-Energy momenta & their constants is the geometry of equilateral Planck Triangles (and all texts must be corrected)

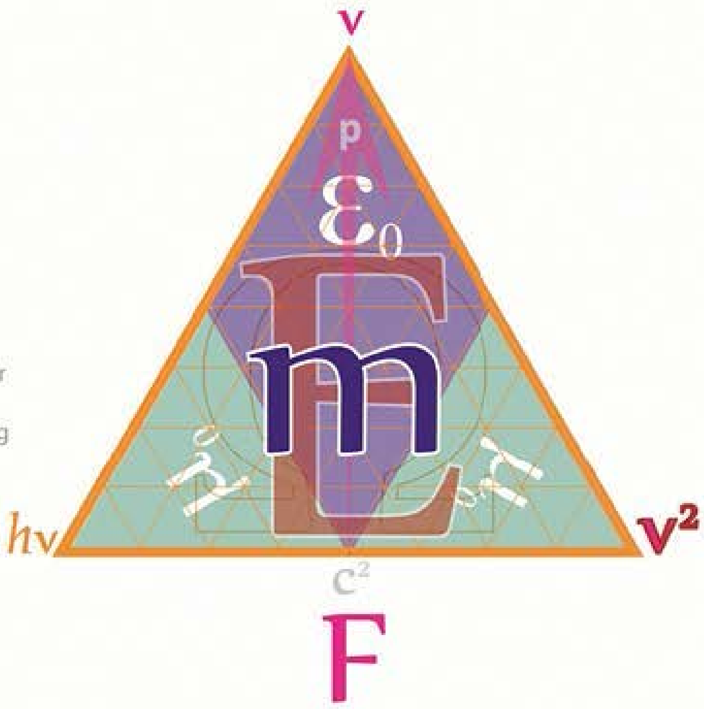

$$\sum \mathbf{F} = \frac{d\mathbf{P}}{dt} = m\frac{d\mathbf{v}}{dt} = m\mathbf{a}$$

Newton's Second law of Motion
is based on changes to linear momentum

$$\mathbf{F} = m\mathbf{a}\,.$$

Copyright ABRAHAM [2008] - all rights reserved

Tetryonics and Pi radians

Although not historically considered a physical constant, π appears routinely in equations describing fundamental principles of the Universe, due in no small part to its relationship to the nature of the circle and, correspondingly, spherical coordinate systems.

The quantised equilateral geometry of mass-energy momenta is measured in π radians

Using units such as Planck units can sometimes eliminate π from formulae.

Heisenberg's uncertainty principle, which shows that the uncertainty in the measurement of a particle's position (Δx) and momentum (Δp) can not both be arbitrarily small at the same time:

$$\Delta x\,\Delta p \geq \frac{h}{4\pi} = \frac{\hbar}{2}$$

Einstein's field equation of general relativity:

$$R_{ik} - \frac{g_{ik}R}{2} + \Lambda g_{ik} = \frac{8\pi G}{c^4}T_{ik}$$

The cosmological constant Λ from Einstein's field equation is related to the intrinsic energy density of the vacuum ρvac via the gravitational constant G as follows:

$$\Lambda = 8\pi G\ \rho_{vac}$$

Coulomb's law for the electric force, describing the force between two electric charges (q1 and q2) separated by distance r:

$$F = \frac{|q_1 q_2|}{4\pi\varepsilon_0 r^2}$$

Magnetic permeability of free space relates the production of a magnetic field in a vacuum by an electric current in units of Newtons (N) and Amperes (A):

$$\mu_0 = 4\pi \cdot 10^{-7}\,\mathrm{N/A^2}$$

Kepler's third law constant, relating the orbital period (P) and the semimajor axis (a) to the masses (M and m) of two co-orbiting bodies:

$$(\frac{2\pi}{P})^2 a^3 = \omega^2 a^3 = G(M+m)$$

and the Gaussian formula for a Normal Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-(x-\mu)^2/(2\sigma^2)}$$

$$\pi = \frac{C}{d}.$$

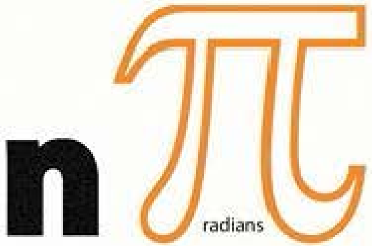

$$\pi = \frac{A}{r^2}.$$

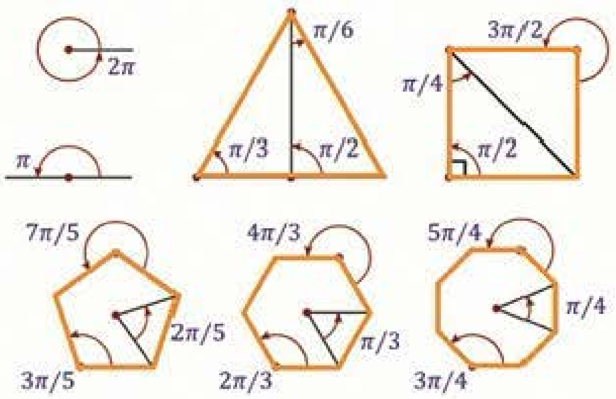

Pi radian mass-energy fascia geometries

All Equilateral triangles have internal angles that add up to 180°

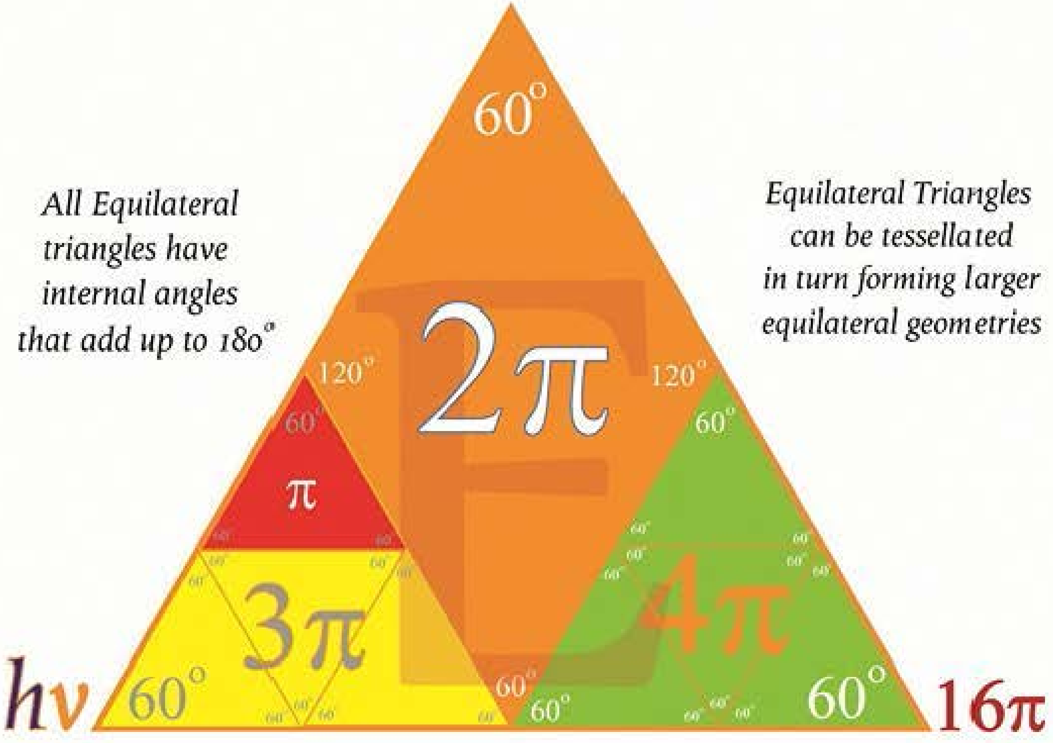

Equilateral Triangles can be tessellated in turn forming larger equilateral geometries

SQUARED numbers in physics are EQUILATERAL geometries

Copyright ABRAHAM [2008] - all rights reserved

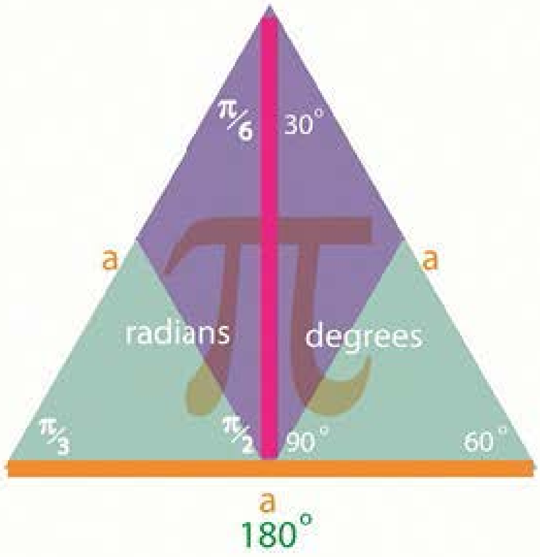
π/6
30°
a
a
π
radians
degrees
π/3
π/2
90°
60°
a
180°

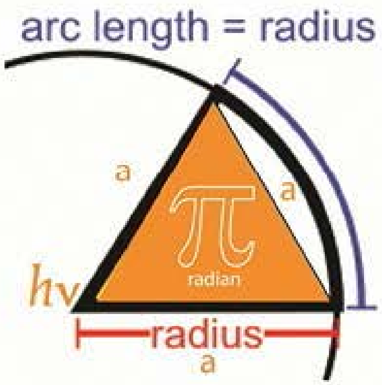
arc length = radius
a
a
π
radian
hv
radius
a

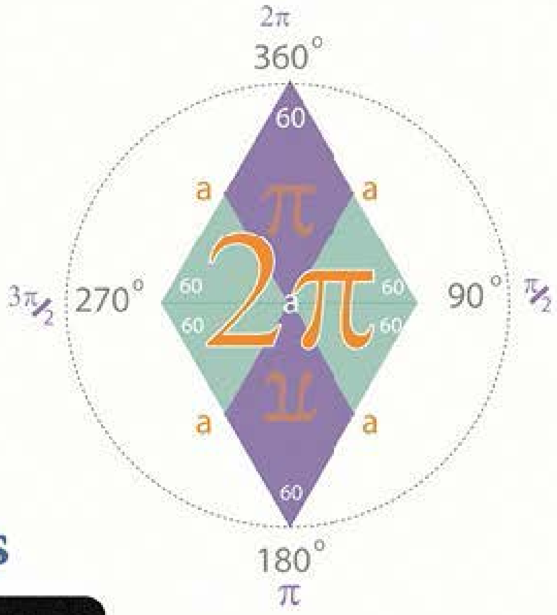
2π
360°
60
a
a
π
3π/2
270°
2π
90°
π/2
a
a
60
180°
π

nπ mass-energy geometries

180° π An equilateral triangle of mass-energy momenta has a geometry of π radians 2π **360°**

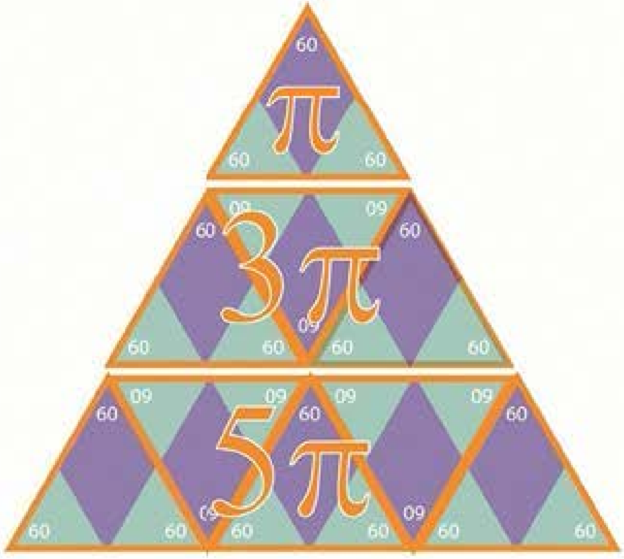

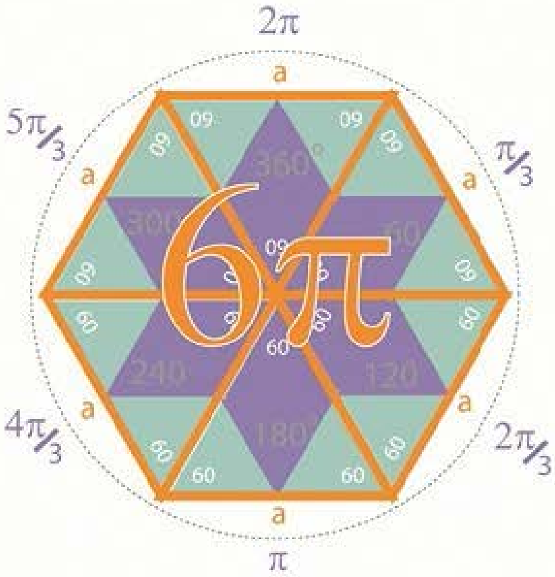
2π
a
5π/3
π/3
360°
300
60
6π
240
120
180
4π/3
2π/3
a
π

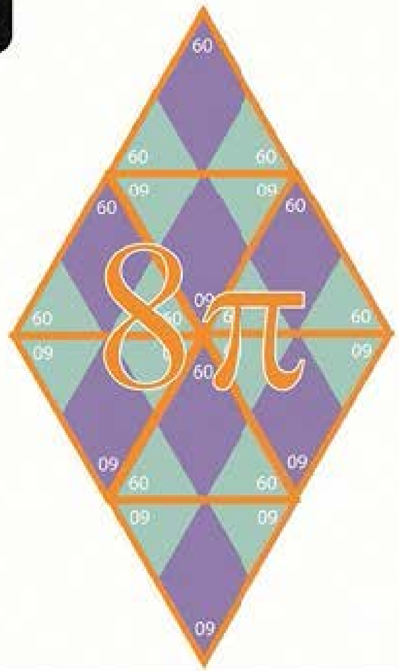

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Cardinal Angles

equilateral mass-energy geometries form tetrahedral mass-Matter topologies

EM Charge

Equilateral energies

Q

E

Space

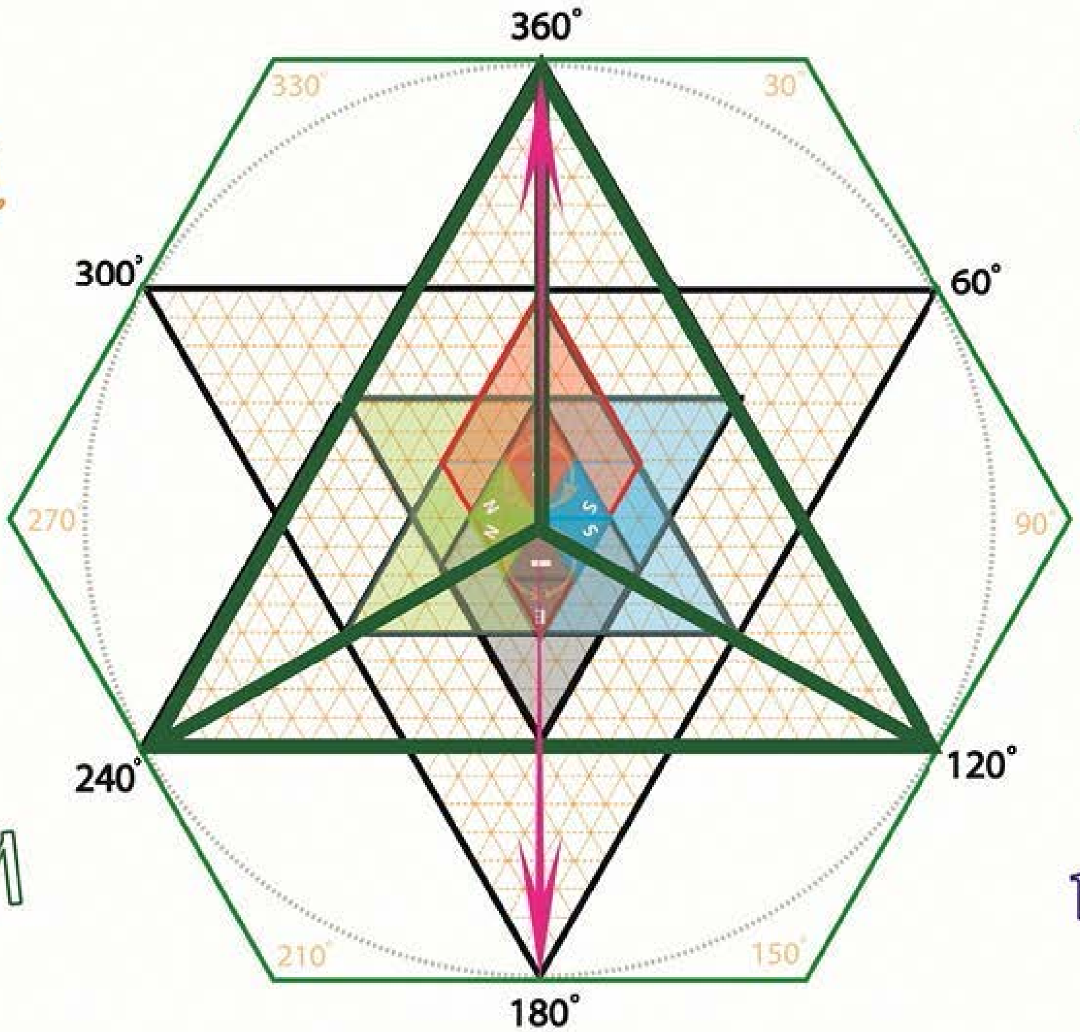

Time

M

m

Matter topologies

mass geometries

radiant light

in radial spatial co-ordinate systems defined by the speed of light

Charge
EM Field
Planck quanta

$$\text{ODD}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons & quantum levels
ElectroMagnetic mass velocity

Kinetic Energies
EM Field
Planck quanta

$$\text{EVEN}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons & EM waves
ElectroMagnetic mass velocity

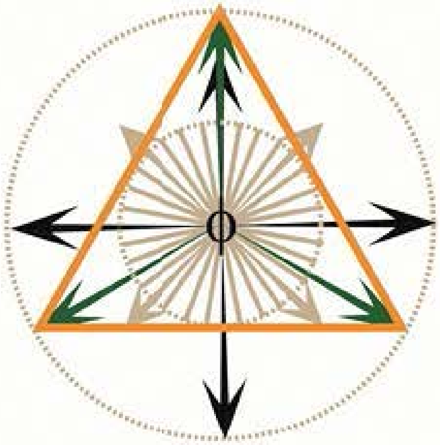

ALL Matter topologies stem from tessellated equilateral mass-energies

Matter
EM Field
Planck quanta

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Fermions
ElectroMagnetic mass velocity

Matter's 4π mass-Energies
ares Lorentz invariant to accelerations

mass-energy geometries

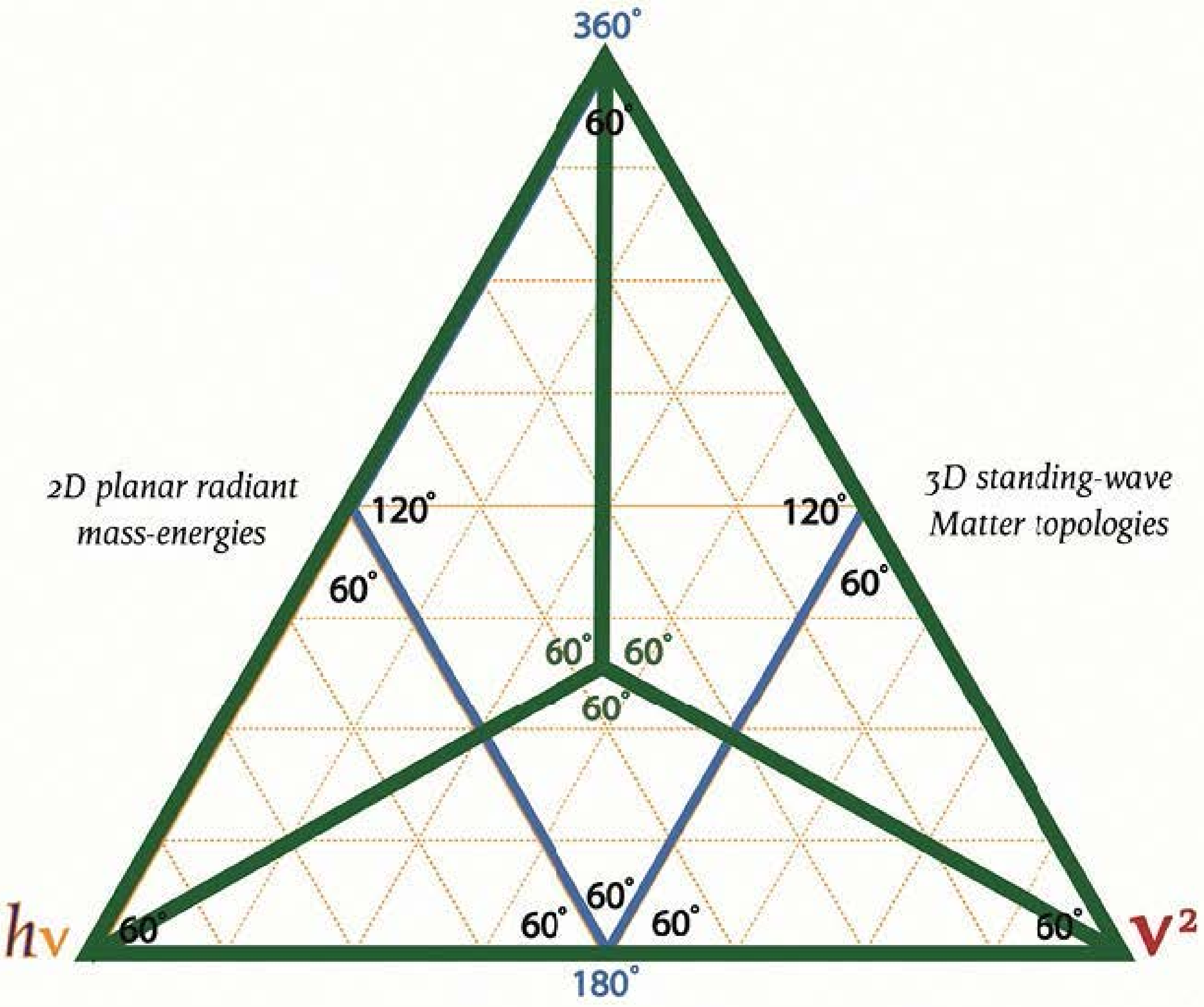
360°
60°
2D planar radiant mass-energies
120°
120°
3D standing-wave Matter topologies
60°
60°
60°
60°
60°
60°
60°
60°
60°
hv
v²
180°

Matter topologies

Tetryonic geometry

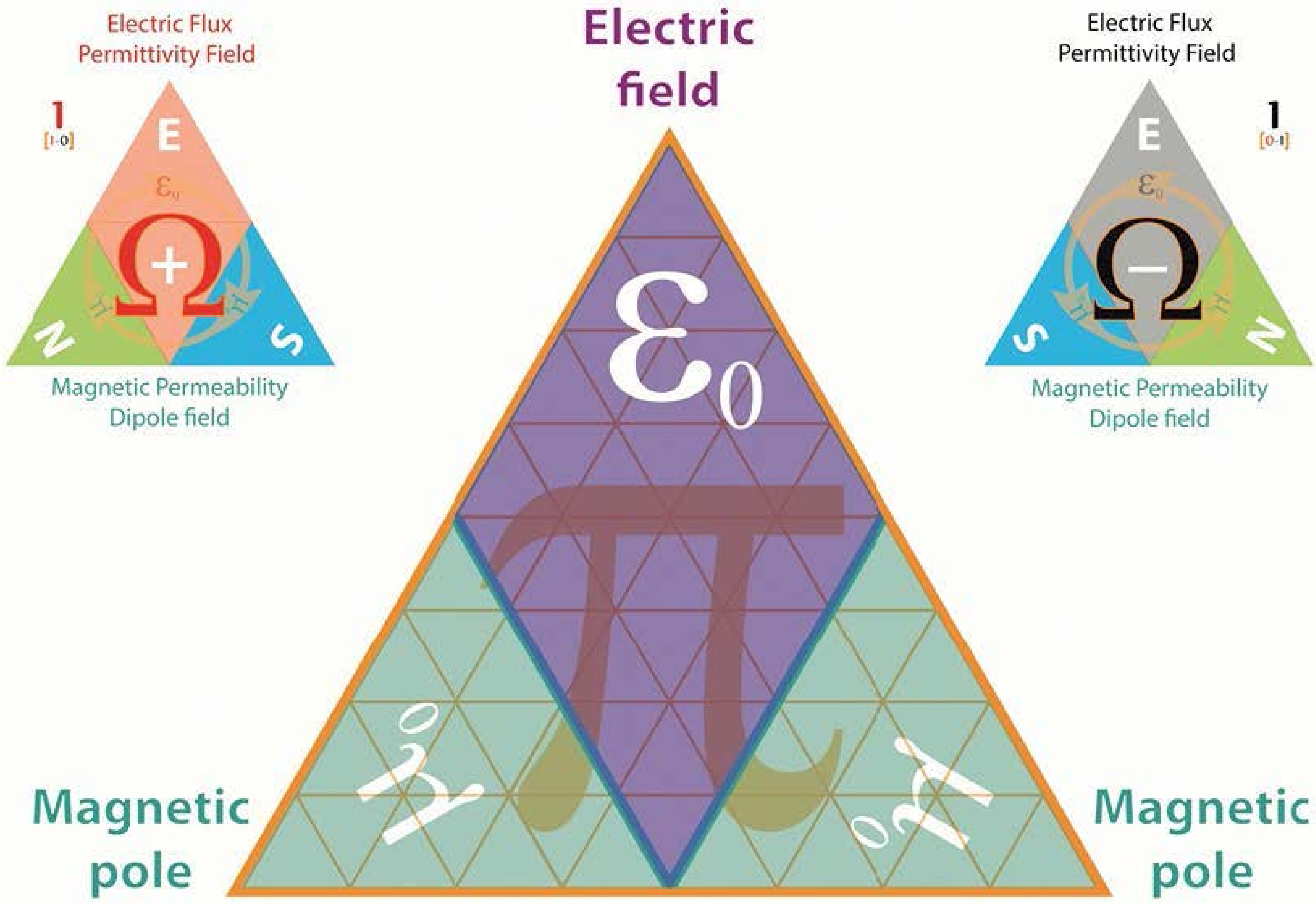
Electric Flux
Permittivity Field
1
[1-0]
E
ε_0
Ω
+
μ
μ
N
S
Magnetic Permeability
Dipole field
Electric
field
ε_0
π
μ_0
μ_0
Magnetic
pole
Magnetic
pole
Electric Flux
Permittivity Field
1
[0-1]
E
ε_0
Ω
–
μ
μ
S
N
Magnetic Permeability
Dipole field

equilateral quantised angular momenta
is the foundational geometry of all mass-ENERGY-Matter

The Golden Triangle

Planck's formulation for Energy is imprecise for use in Tetryonics
and does not reflect the velocity-momenta relationship
inherent in the equilateral geometry of Energy

Planck quantum levels

Energy is gained or lost in
whole number multiples of the quantity hv

$$E = n[hv]$$

ODD

Charge

EM Field

Planck quanta

$$ODD\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Bosons & quantum levels

ElectroMagnetic

mass

velocity

$$3 = 3.[1]$$

$$9 = 9.[1]$$

$$25 = 25.[1]$$

$$hv$$

The generalise formulation of Plank's heat law
E = nhv is now changed to a specific formulation of
E = [ODD] hv for transverse quantum levels [Bosons]

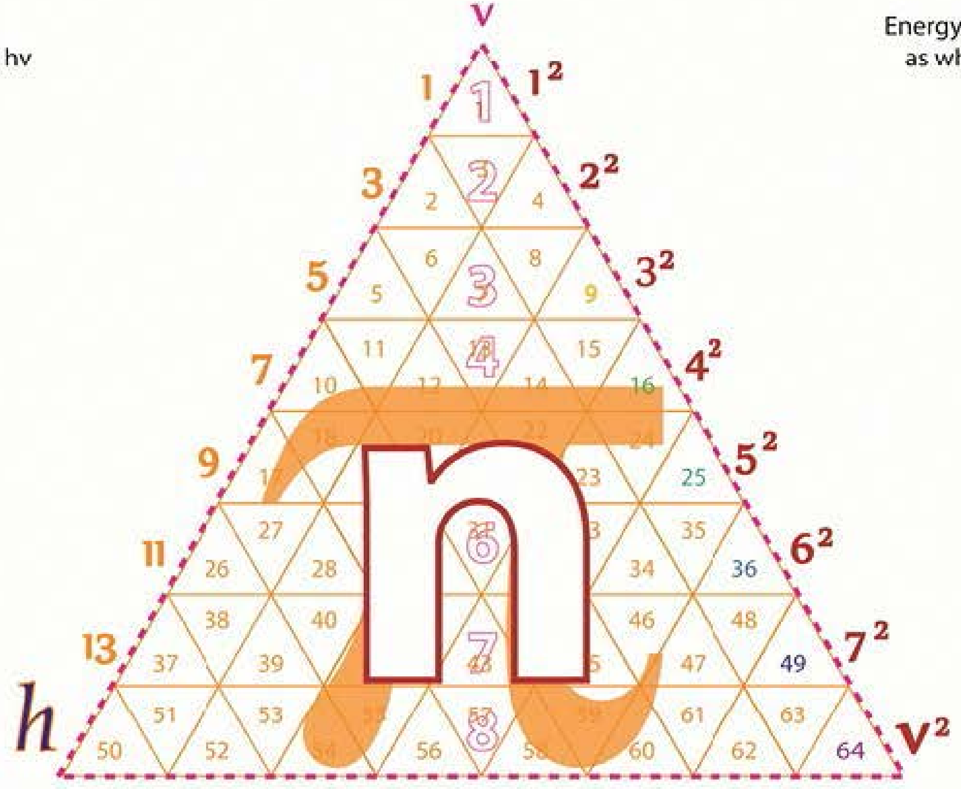

mass-energy momenta are geometrically related to velocity

Tetryonic geometry [nπ] redefines Planck's quantum formulation
for heat energies from a generalised equation for 2D energy momenta
into a geometric formulation for all mass-energy momenta in Matter

[all equilateral [π] geometries contain square number quanta]

Scalar energy quanta

Energy is gained or lost in equilateral geometries
as whole number multiples of the quantity hv²

$$E = n[hv]$$

SQUARED

Planck quanta

$$E = n\pi[[m\Omega v^2]]$$

mass

velocity

$$3 = 3.[1^2]$$

$$9 = 1.[3^2]$$

$$25 = 1.[5^2]$$

$$hv^2$$

The general formulation of Plank's heat law
is also changed to a specific formulation of
E = hv² for scalar EM waveforms [ENERGY]

Copyright ABRAHAM [2008] - all rights reserved

Scalar field geometries

All scalar fields are comprised of Transverse and Longitudinal mass-energy momenta all of which are formed from equilateral quantum geometries

Charge is a conserved force

Energy is a conserved quantity

Bosons are transverse
ODD planck geometries

$$E = n\pi\left[\left[hv\right]\right]$$

Planck quanta

Energy is scalar
SQUARE planck quanta

$$E = n\pi\left[\left[hv^2\right]\right]$$

Planck quanta squared

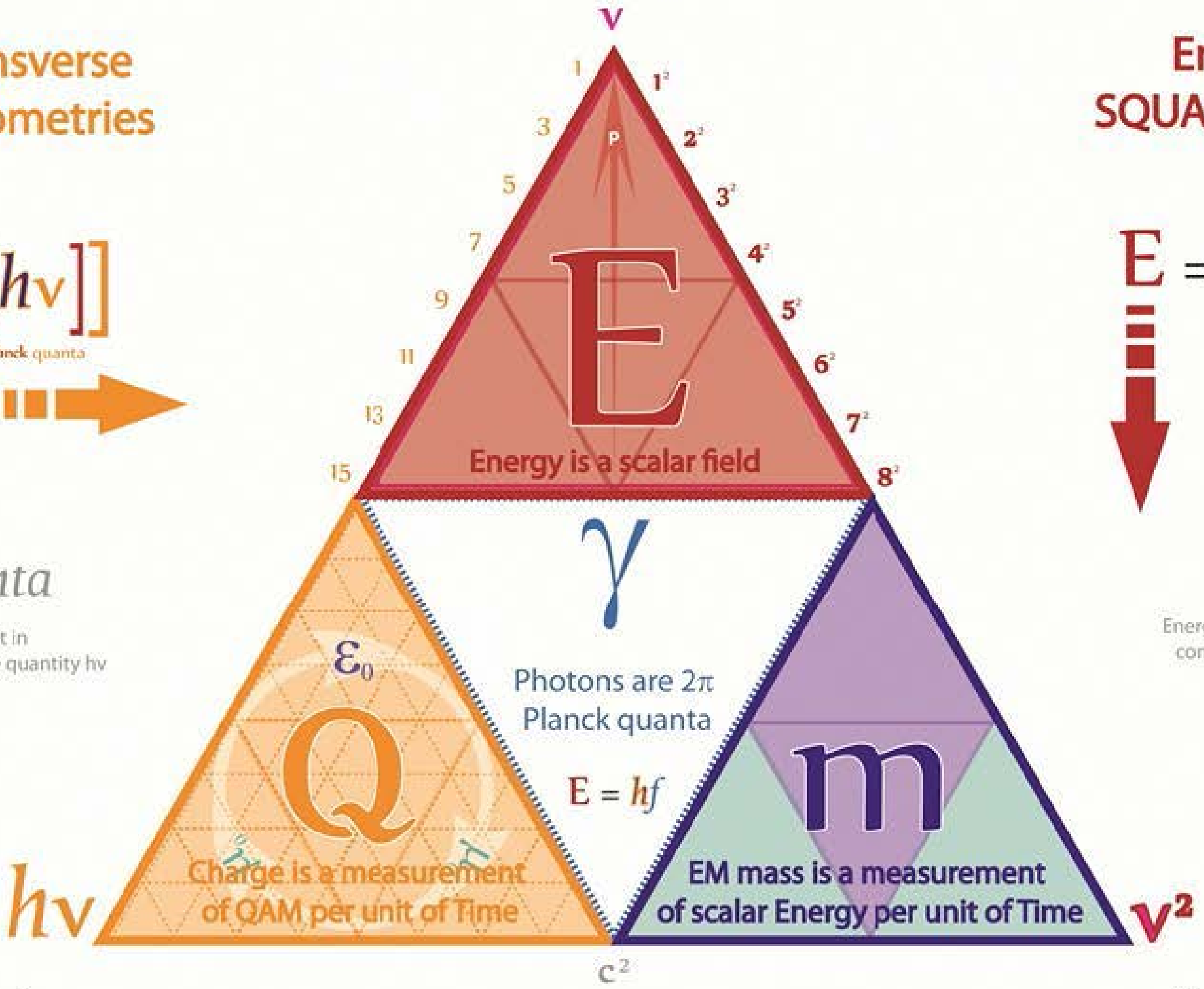

Energy is a scalar field

Photons are 2π Planck quanta

$E = hf$

Charge is a measurement of QAM per unit of Time

EM mass is a measurement of scalar Energy per unit of Time

Planck quanta

Energy is gained or lost in whole number multiples of the quantity hv

Scalar energies

Energy is gained or lost in equilateral quanta containing whole number multiples of hv^2

All energy is comprised of EM fields

All squared energy geometries have component quantum levels

Copyright ABRAHAM [2008] - all rights reserved

Squared energy levels in quantum physics

are in fact equilateral Planck mass-energy geometries

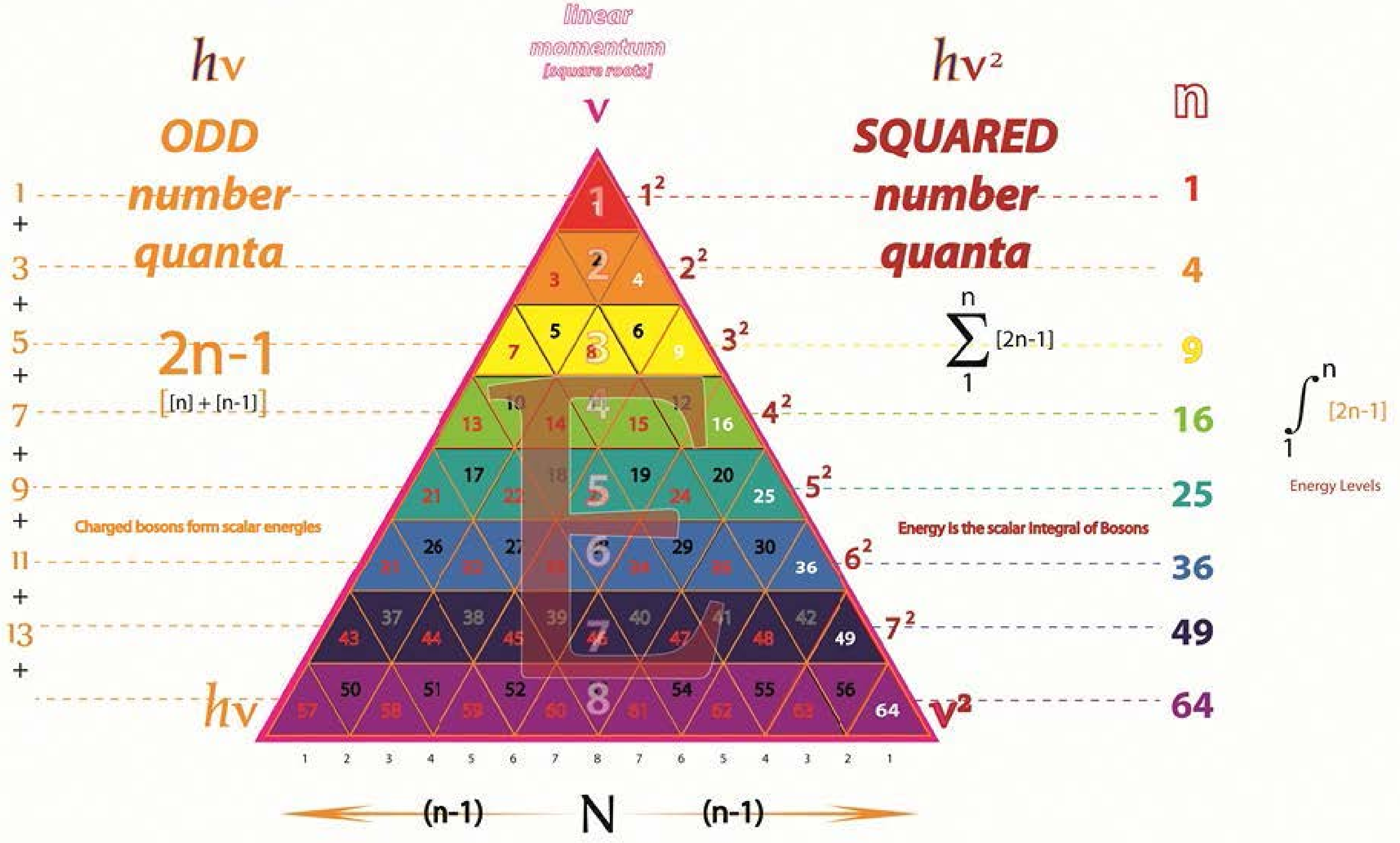
linear
momentum
[square roots]
v
hv
ODD
number
quanta
2n-1
[n] + [n-1]
1
+
3
+
5
+
7
+
9
+
11
+
13
+
Charged bosons form scalar energies
hv
hv²
n
SQUARED
number
quanta
1^2
2^2
3^2
4^2
5^2
6^2
7^2
v^2
1
4
$\sum_{1}^{n} [2n-1]$
9
16
25
36
49
64
$\int_{1}^{n} [2n-1]$
Energy Levels
Energy is the scalar Integral of Bosons
E
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
(n-1)
N
(n-1)

Cumulative Distribution

NORMAL DISTRIBUTION

['Bell Curve']

Probability Distribution

Copyright ABRAHAM [2008] - all rights reserved

Energy quanta defined

Quantum levels

Charge is gained or lost in
odd number multiples of the quantity hv

$$E = n\pi\left[\left[hv\right]\right]$$

Planck quanta

transverse

ODD quanta

Scalar Energies

Energy is gained or lost in equilateral geometries
containing whole number multiples of the quantity hv²

$$E = n\pi\left[\left[hv^2\right]\right]$$

Planck quanta squared

scalar

SQUARE quanta

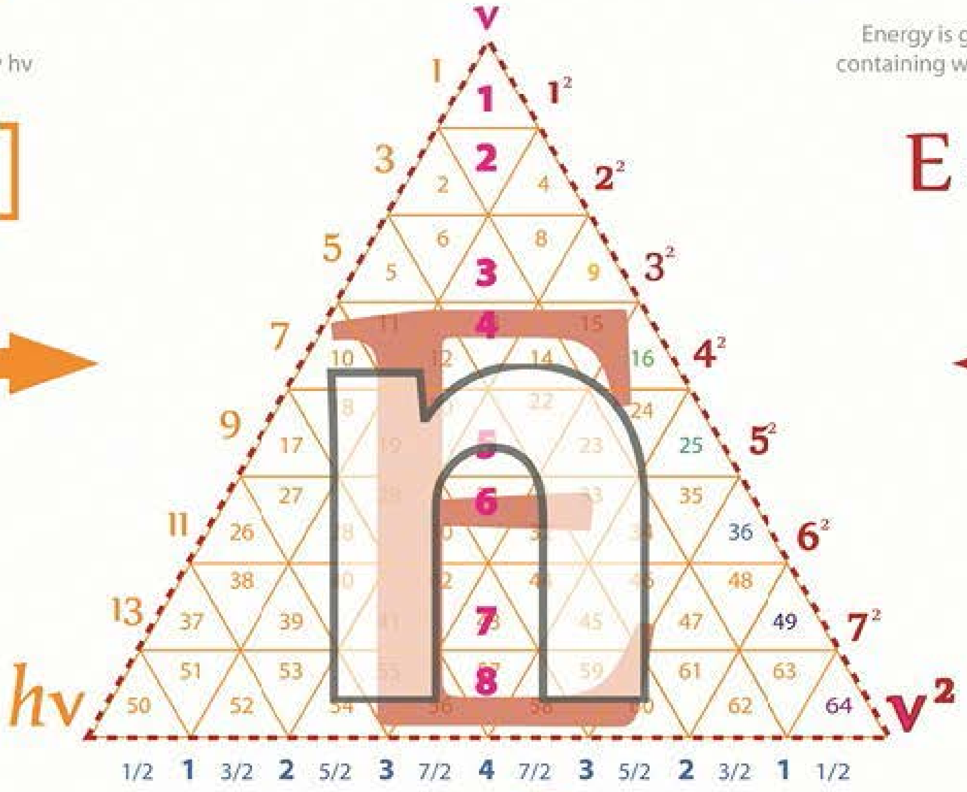
v
1
3
5
7
9
11
13
hv
1²
2²
3²
4²
5²
6²
7²
v²
1
2
3
4
5
6
7
8
2
4
6
8
5
9
10
11
12
14
15
16
17
22
23
24
25
27
26
35
36
38
48
37
39
47
49
51
53
61
63
50
52
62
64
1/2
1
3/2
2
5/2
3
7/2
4
7/2
3
5/2
2
3/2
1
1/2

Longitudinal

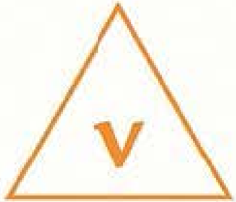

Quantum

The number of Planck quanta
in any physical system

$$v = f/2$$

Transverse

Quantum Levels

The number of
repeating waveform cycles
in a system

$$2v = f$$

Frequency

EVEN quanta

Tetryonics 02.11 - Energy quanta defined

Tetryonic field equation

E

Planck quanta

$$E = n\pi\left[\left[m\Omega v^2\right]\right]$$

mass velocity

electromagnetic energy is a scalar

$$E$$

m

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

mass is an inertial constant that relates Force to acceleration

$$\frac{E}{c^2}$$

$[v\text{-}v]$ Q

Charge EM Field Planck quanta

$$ODD\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons & quantum levels ElectroMagnetic mass velocity

equilateral charge geometry is Lorentz invariant

$$\frac{\Omega}{c^2}$$

KE

Kinetic Energies EM Field Planck quanta

$$EVEN\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons & EM waves ElectroMagnetic mass velocity

KEM fields mass-energy geometries are Lorentz variant to accelerations

$$\frac{1}{2}m\left[\frac{v^2}{c^2}\right]$$

M

Matter EM Field Planck quanta

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Fermions ElectroMagnetic mass velocity

Matter's 4π mass-energy topologies are Lorentz invariant to accelerations

$$4n\pi\left[\frac{E}{c^4}\right]$$

Energy momenta geometry

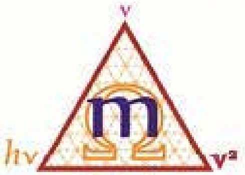

Energy is the total Planck quanta per second

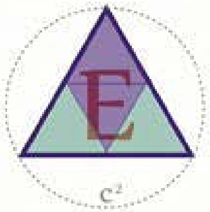

EM mass is scalar energy per unit of Time

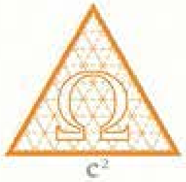

Charge is the nett Angular momentum per spatial co-ordinate system

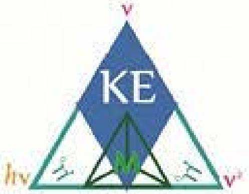

Kinetic Energy is the Electric field mass-energy of Matter moving at v

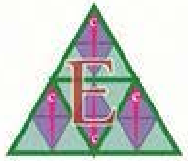

mass-energy geometry

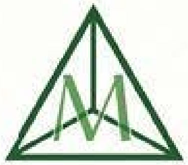

Matter topology

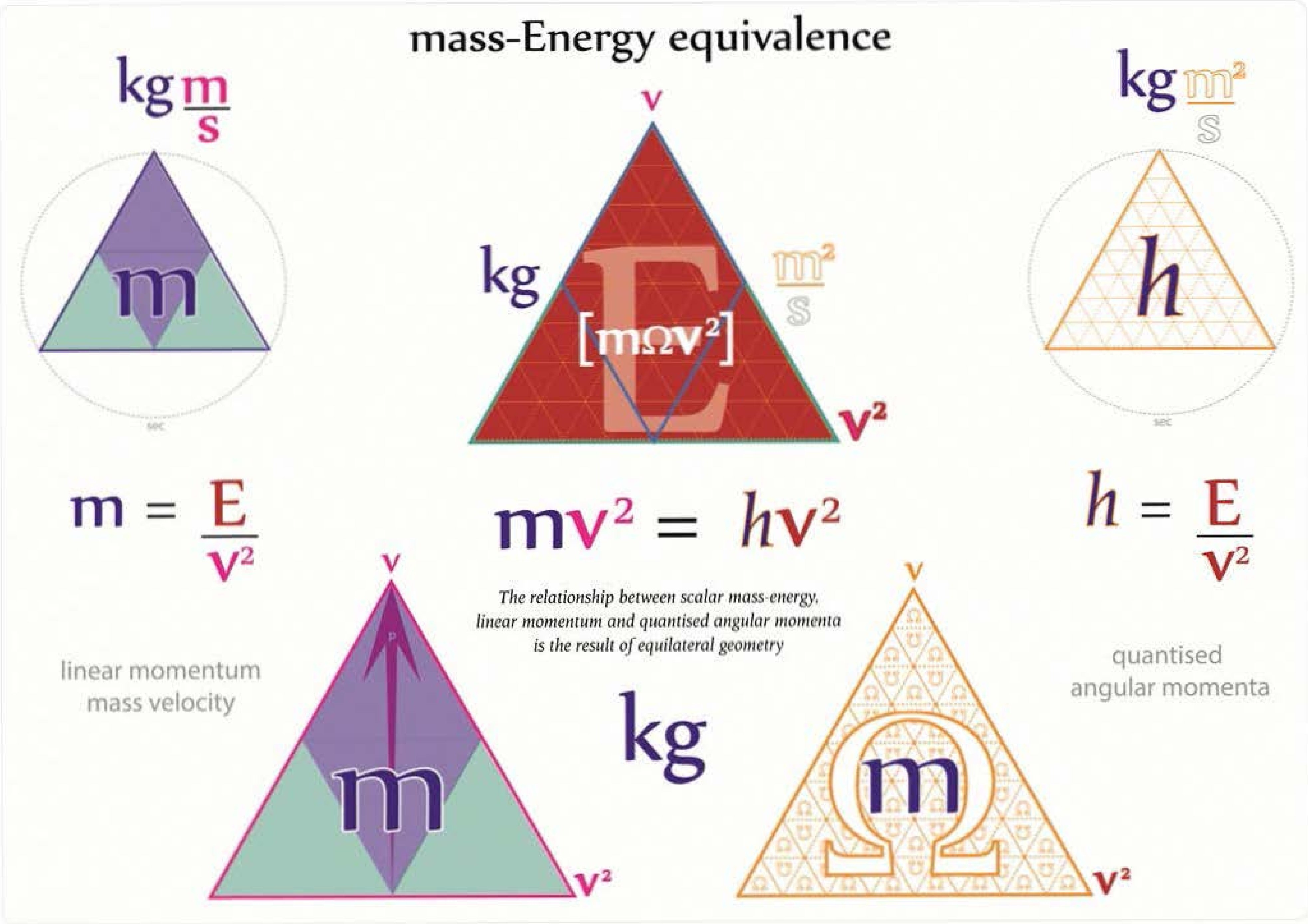
$kg\frac{m}{s}$
v
kg
$\frac{m^2}{s}$
E
[mΩv²]
v²
$kg\frac{m^2}{s}$
m
sec
h
sec
v
p
m
v²
kg
v
Ω
m
v²

mass-Energy equivalence

$$m = \frac{E}{v^2}$$

$$mv^2 = hv^2$$

$$h = \frac{E}{v^2}$$

The relationship between scalar mass-energy, linear momentum and quantised angular momenta is the result of equilateral geometry

linear momentum
mass velocity

quantised
angular momenta

Tetryonics 02.13 - mass-Energy equivalance

mass-Energy geometries

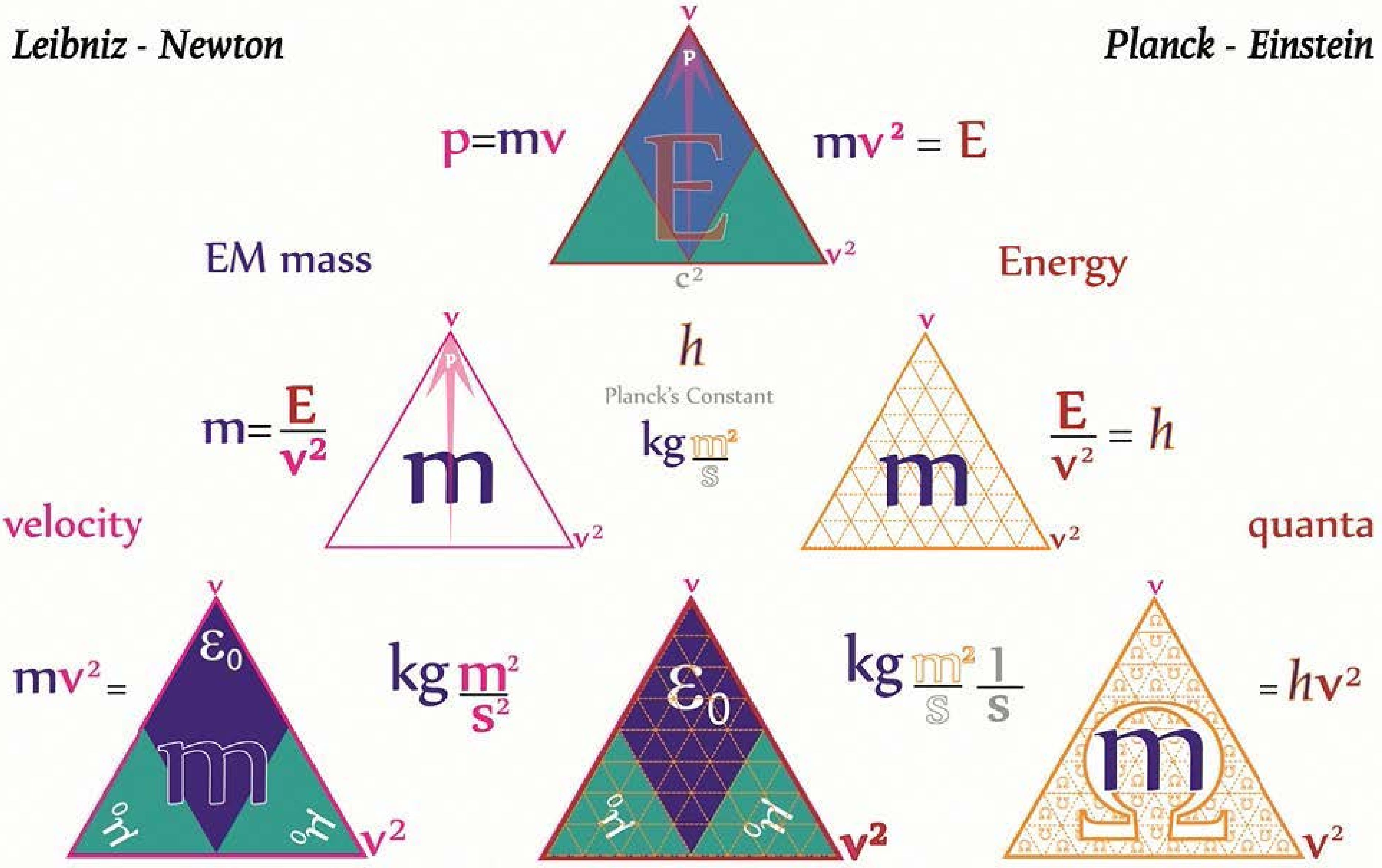
v
P
E
v²
c²
v
P
m
v²
v
m
v²
v
ε0
m
μ0
μ0
v²
v
ε0
μ0
μ0
v²
v
m
Ω
v²

Leibniz - Newton

Planck - Einstein

$p=mv$

$mv^2 = E$

EM mass

Energy

velocity

$m=\frac{E}{v^2}$

h

Planck's Constant

$kg\frac{m^2}{s}$

$\frac{E}{v^2} = h$

quanta

$mv^2 =$

$kg\frac{m^2}{s^2}$

$kg\frac{m^2}{s}\frac{1}{s}$

$= hv^2$

velocity squared

There exists an intrinsic geometric relationship between mass-velocity and the quantisation of Energy momenta

quanta squared

EM mass-energy

$$E = mc^2$$

EM mass is related to Matter
throught the square of the velocity of light
of 8.987551787 e16 m²/s² [c²]

EM mass-Matter

$$E = Mc^4$$

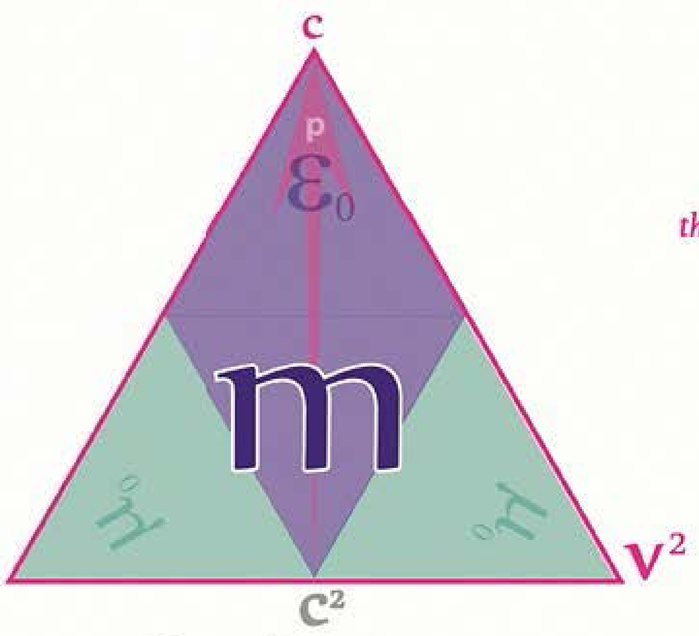
c
p
ε_0
m
μ_0
μ_0
v^2
c^2

Electro-Magnetic mass

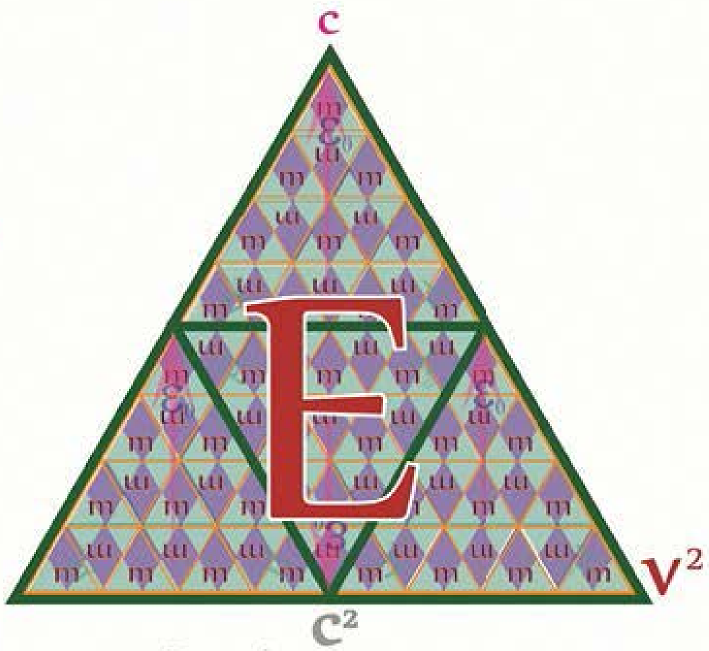
c
E
m
ε_0
v^2
c^2

Matter's energy quanta

scalar EM mass

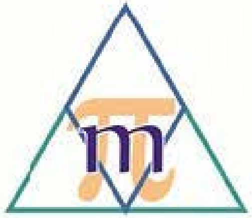

$$M = \frac{E}{c^4}$$

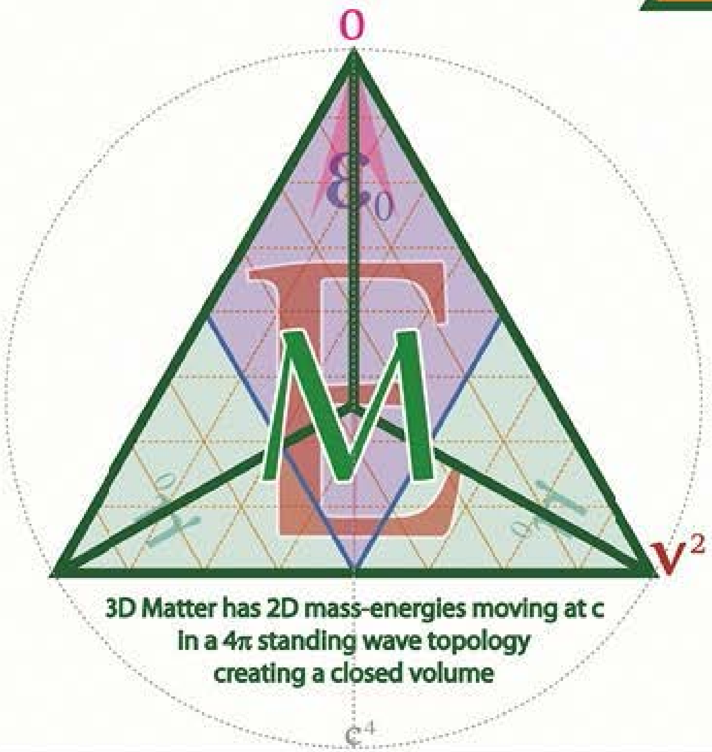
0
ε_0
E
M
μ_0
μ_0
v^2
c^4

3D Matter has 2D mass-energies moving at c
in a 4π standing wave topology
creating a closed volume

quantised mass

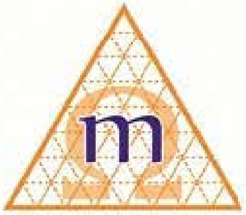

$$h = \frac{E}{v^2}$$

Tetryonics 02.15 - EM mass-ENERGY-Matter

Copyright ABRAHAM [2008] - all rights reserved

Kinetic Electro-Magnetic fields

[The energies of Motion]

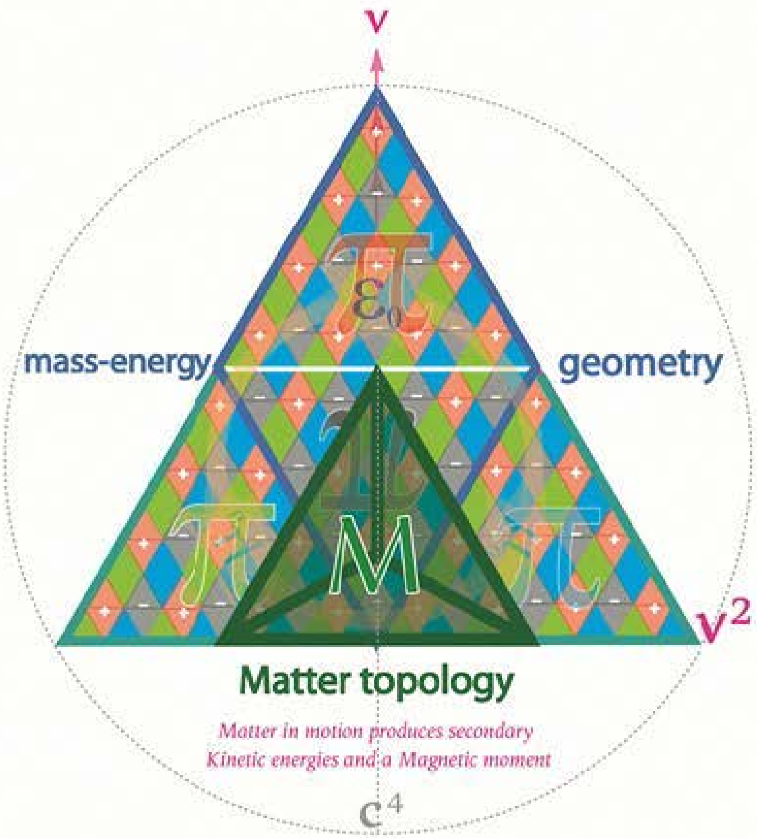
v
mass-energy
geometry
ε_0
M
v^2
Matter topology
c^4

Matter in motion produces secondary Kinetic energies and a Magnetic moment

$$p^2 = E = Mv^2$$

The Electric field energy in any EM field is equal and orthogonal to the Magnetic field energy

$$p^2 = KEM = Mv^2$$

$$\mu_B = \frac{1}{2}Mv^2$$

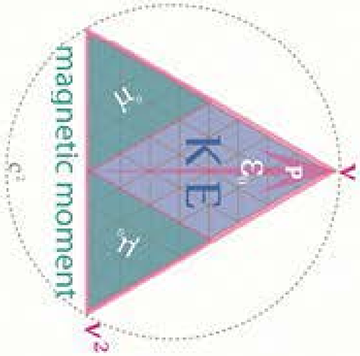
magnetic moment
μ_0
KE
ε_0
p
v
v^2
c^2

$$KE = \frac{1}{2}Mv^2$$

All Matter in motion possess momenta and kinetic mass-energies in extrinsic KEM fields

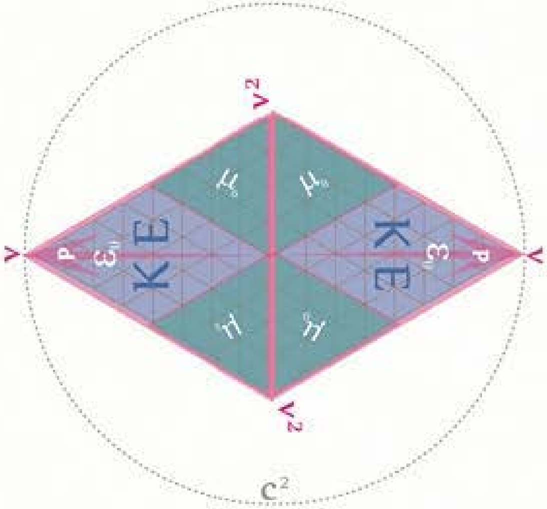
v^2
μ_0
KE
ε_0
p
v
c^2

$$E_\gamma = \frac{2mv^2}{c^2}$$

These 2D planar fields are subject to Lorentz factor corrections that apply to the KEM field mass-energies of Matter in motion

Charge & Kinetic EM fields

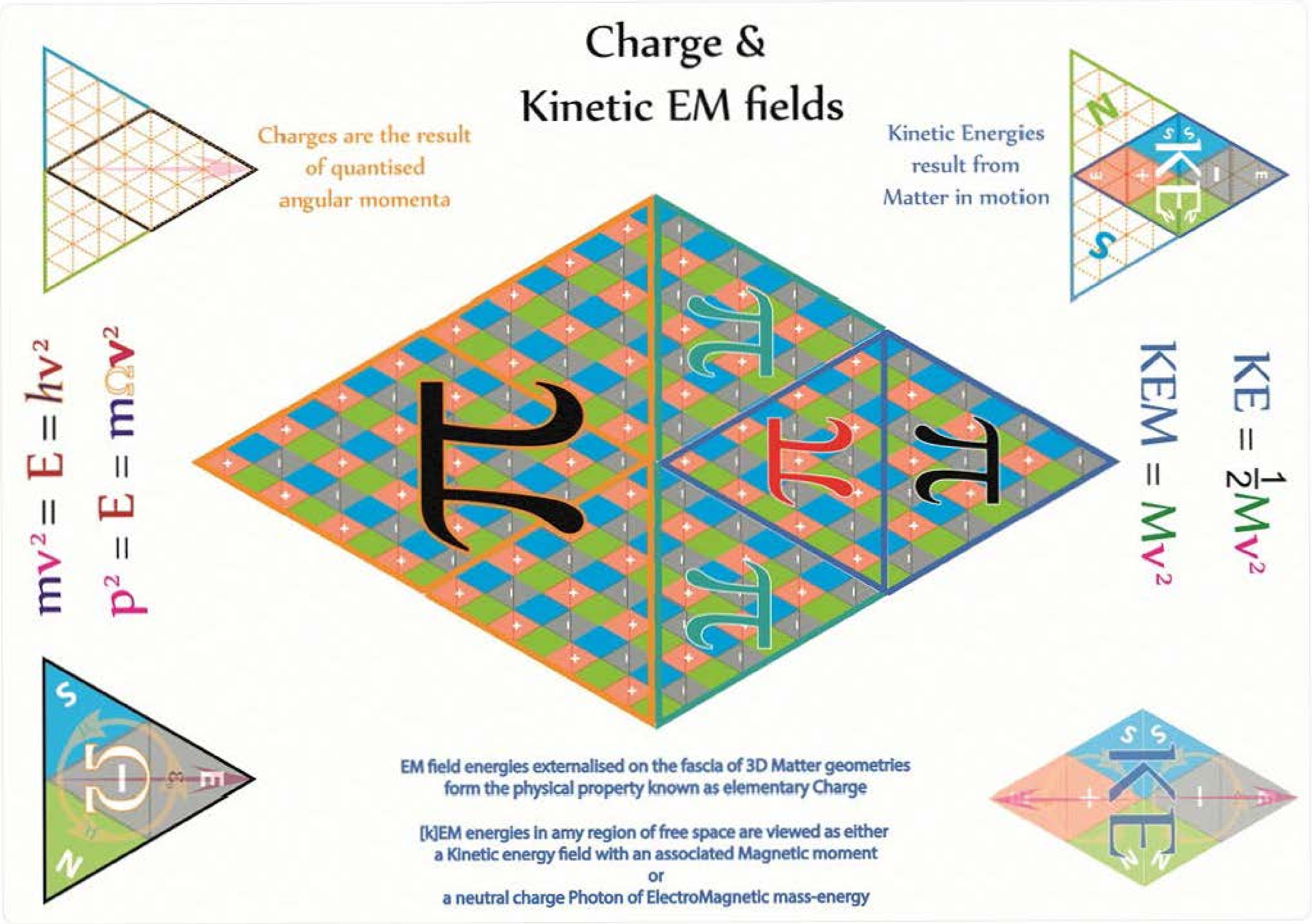

Charges are the result
of quantised
angular momenta

Kinetic Energies
result from
Matter in motion

$mv^2 = E = hv^2$

$p^2 = E = m\Omega v^2$

$KEM = Mv^2$

$KE = \frac{1}{2}Mv^2$

EM field energies externalised on the fascia of 3D Matter geometries
form the physical property known as elementary Charge

[k]EM energies in amy region of free space are viewed as either
a Kinetic energy field with an associated Magnetic moment
or
a neutral charge Photon of ElectroMagnetic mass-energy

ε_0
μ_0
μ_0

$$m_0 = \frac{E}{c^2}$$

Rest mass is equivalent to the total quantity of Energy in a body or system (divided by c²)

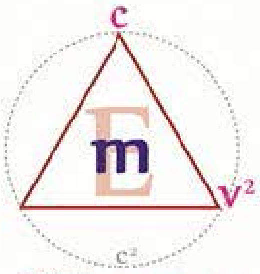

2D fields propagate at c
3D Matter moves at v

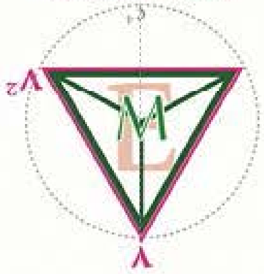

inertial mass

$$m = \frac{F}{a}$$

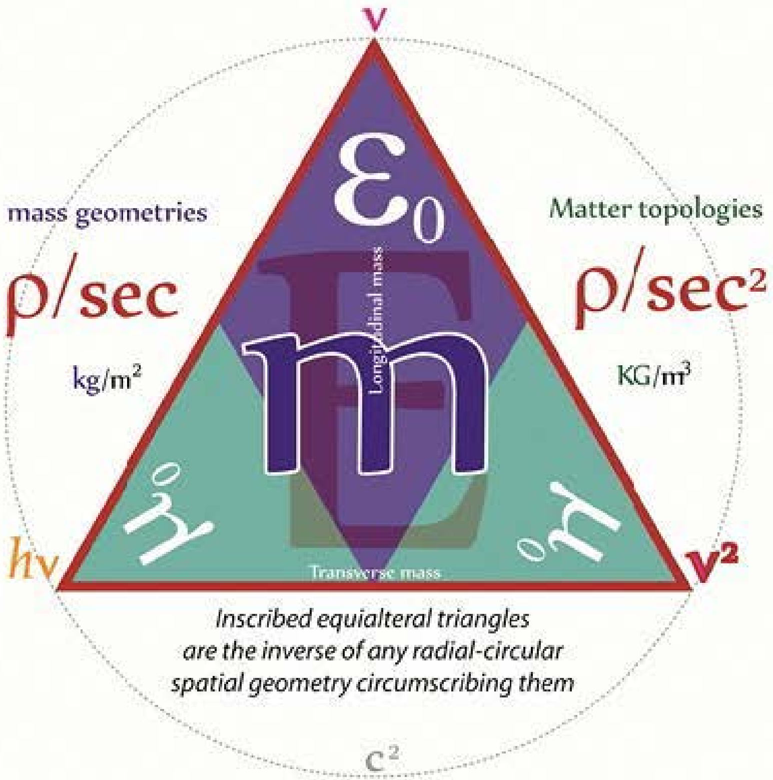

Inscribed equialteral triangles are the inverse of any radial-circular spatial geometry circumscribing them

mass

ρ

EM mass is a measure of planar energy density per second

$$\frac{mv^2}{c^2} \quad \boxed{\frac{m\Omega v^2}{c^2}} \quad \frac{hv^2}{c^2}$$

per unit of time

EM mass geometries are subject to Lorentz corrections

$$\frac{m\overset{\text{velocity}}{v^2}}{c^2} = \frac{m\overset{\text{QAM quanta}}{\Omega}v^2}{c^2}$$

per unit of time

mass must be distinguished from Matter in physics, because Matter is a poorly-defined concept in modern science, and although all types of agreed-upon Matter exhibit properties of mass, it is also the case that there many types of Energy that posses NO Matter topology, such as potential energy, kinetic energies and electromagnetic radiation (photons)

$$\underset{\text{mass}}{n\pi}\left[\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{\varepsilon_0\mu_0}}\right].\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

The term 'massless' must be re-termed 'Matterless' to reflect true physical attributes of mass-energy-momenta

Gravitational Matter

$$g = G\frac{M}{r^2}$$

EM mass is a planar measurement of 2D energy per unit of time

Thus, all 3D Matter topologies have charged fascia comprised of 2D scalar of mass-energies, but closed volume 3D topology is not a property of 2D EM mass-energy geometries

Copyright ABRAHAM [2008] - all rights reserved

Electromagnetic mass-energy

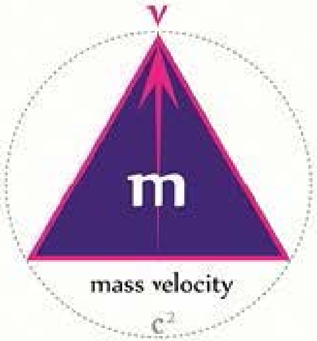

In physics, EM mass-energy equivalence is the concept that the EM mass of a body is a measure of its energy content

Using this concept, EM mass is a property of all Energy, and Energy is a property of all EM mass, and the two properties are connected by a constant.

Using Tetryonic geometry it can be shown that the constant is the equilateral geometry of QAM thus unifying Classical mechanics and Relativisic mechanics

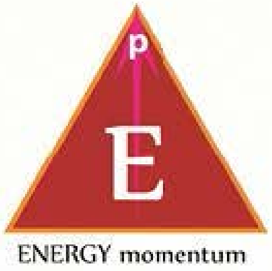

2D mass-ENERGY geometry is NOT 3D Matter topology

$$m = \frac{E}{v^2}$$

Classical mass

$$m = \frac{E}{c^2}$$

Relativistic mass

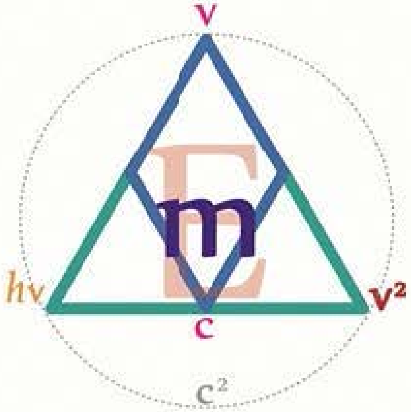

radiant 2D mass-energies are planar equilateral energy geometries

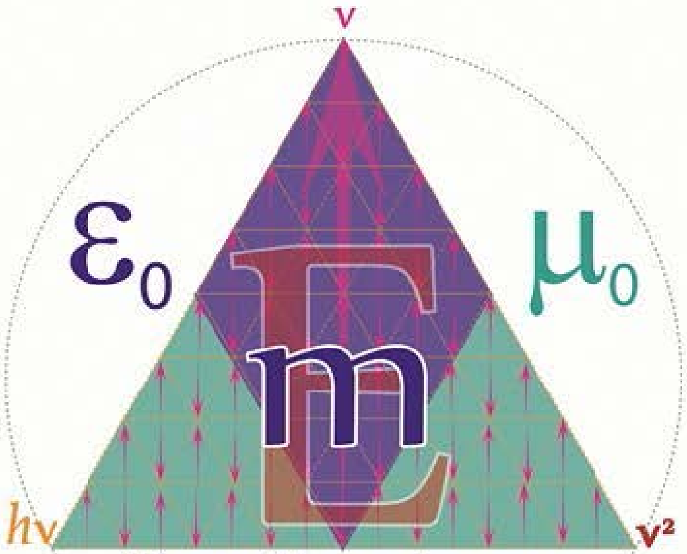

ElectroMagnetic mass

is a measure of the scalar energy momenta in any spatial co-ordinate system

c^2

Relativity shows that rest mass and rest energy are essentially equivalent, via the well-known relationship ($E=mc^2$)

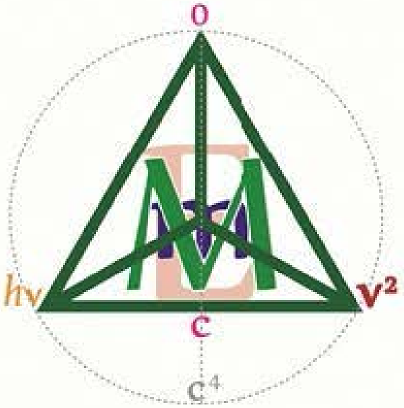

standing-wave 3d mass-Matter are tetrahedral energy momenta topologies

EM mass-Energy-Matter

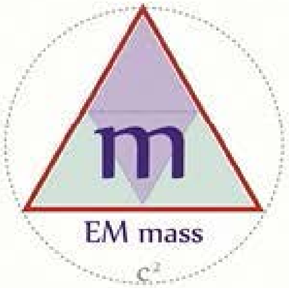

$$\left[\frac{\Omega}{c^2}\right]$$

Charge is a measure of the quantised angular momentum of any physical system

Tetryonic Matter

planar spatial Impedance

Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

3D topology

ElectroMagnetic

2D mass

velocity

Planck quanta

$$\left[m\Omega v\right]^2$$

mass velocity

Bosons
n
v
v
c²

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ODD quanta

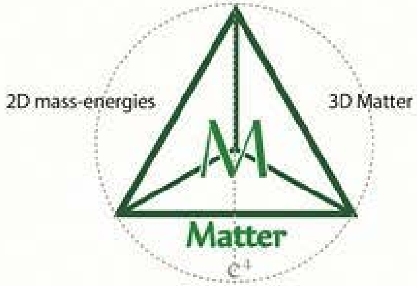
2D mass-energies
3D Matter
M
Matter
c⁴

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

TETRYONS

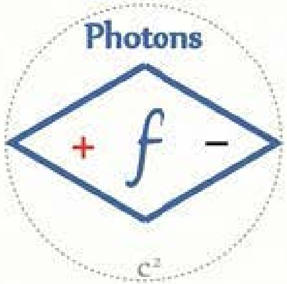
Photons
+
f
−
c²

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EVEN quanta

Copyright ABRAHAM [2008] - all rights reserved

EM mass-Energy momenta

Energy

$$\left[\frac{m\Omega v^2}{s}\right]$$

E

6.629432672 e-34 J

mass-Energy

7.376238634 e-51 kg

m

$$\left[\frac{E}{c^2}\right]$$

mass

Planck's Constant

is the 'quantum of Action'

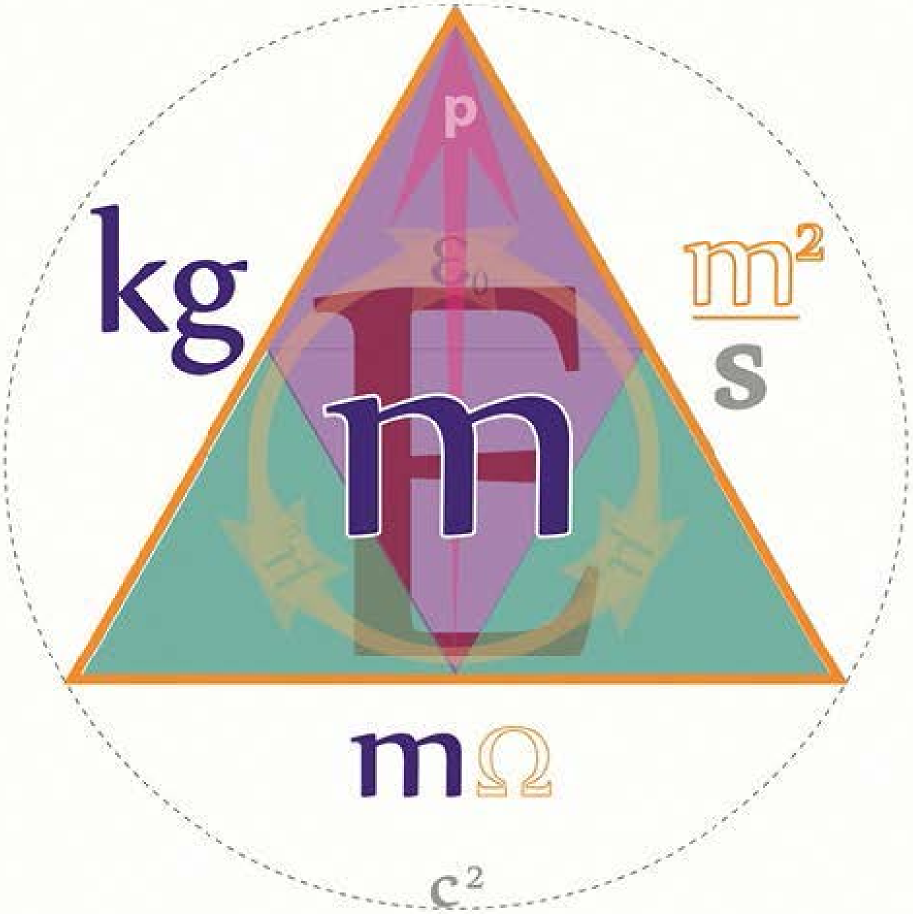

6.629432672 x 10^{-34} J.s

mass momenta

quantised angular momenta

$$\left[\frac{m^2}{s}\right]$$

Ω

0.0012 m2/s

Charge-momenta

1.33518 e-20 s

s

$$\left[\frac{\Omega}{c^2}\right]$$

charge

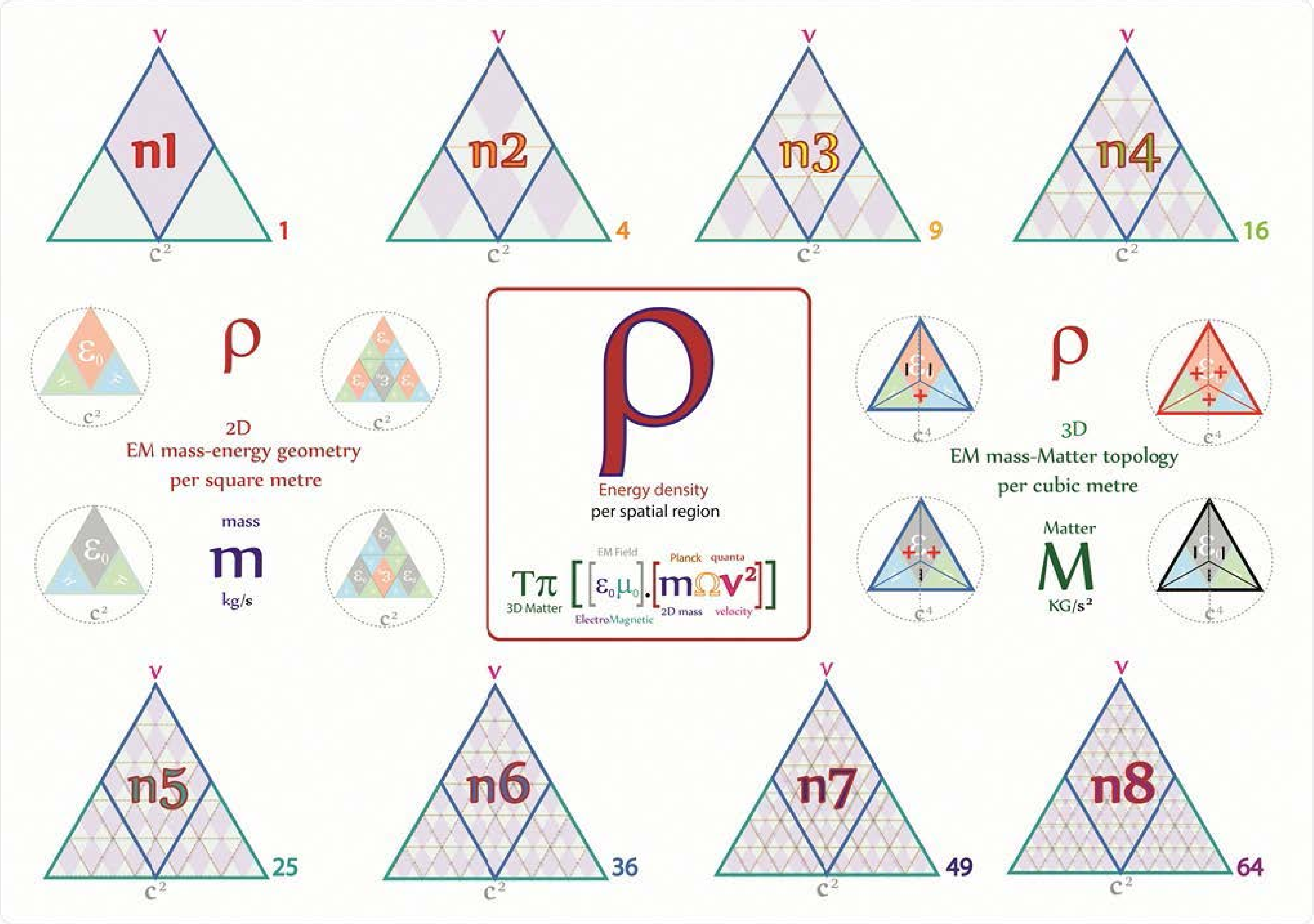
v
n1
1
c²
v
n2
4
c²
v
n3
9
c²
v
n4
16
c²
ε₀
c²
ρ
2D
EM mass-energy geometry
per square metre
ε₀
c²
mass
m
kg/s
ρ
Energy density
per spatial region
EM Field
Planck
quanta
3D Matter
ElectroMagnetic
2D mass
velocity
c⁴
ρ
3D
EM mass-Matter topology
per cubic metre
c⁴
Matter
M
KG/s²
c⁴
c⁴
v
n5
25
c²
v
n6
36
c²
v
n7
49
c²
v
n8
64
c²

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Zero Point Fields [ZPFs]

IDEAL QUANTUM INDUCTORS
(equilateral triangle Energy geometry)

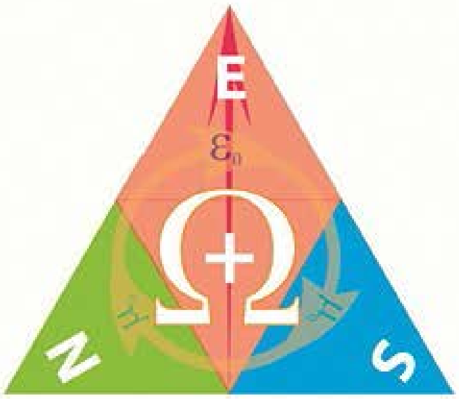

The EM FIELD

Electric flux fields can propagate in any direction
Magnetic fields are always at 90 degrees to Electric fields

Magnetic dipole fields propagate in 2 directions at
180 degrees to each other (bi-directionally)
forming North and South poles

Electric flux field energy is directly proportional to
the resultant Magnetic dipole field energy
and vice versa

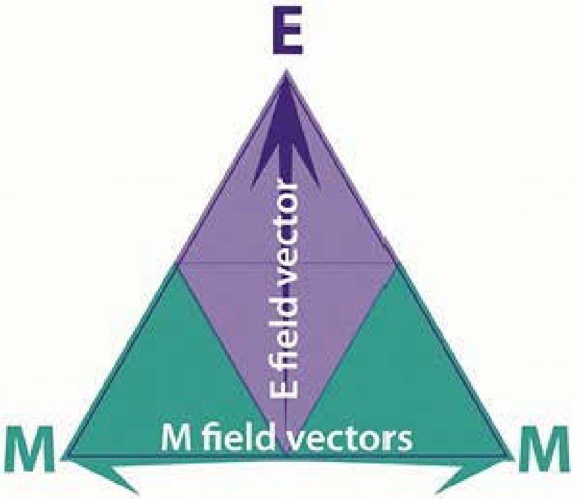

ZPFs are quantum inductive tank circuits
(Short-circuited 'IDEAL' inductors with energy)
ZPFs charge energies do NOT oscillate
[The magnetic dipole vector determines charge]

The linear Electric field strength is directly
proportional to its associated transverse
Magnetic field which propagates bi-directionally
from & into the bloch wall of the Zero point field

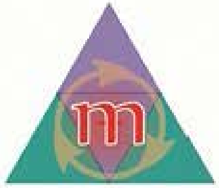

h

Positive Charge ZPF

Nett positive Planck quanta
with North-South m-dipole vector

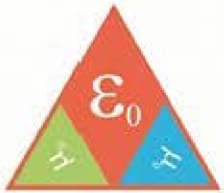

Zero Point Fields consist of
Electric and Magnetic (EM) fields
propagating at 90 degrees
to each other

Magnetic MONOPLES do NOT exist

Energy quanta always form charged Electric fields and dipole Magnetic fields

As localised energy quanta increases
(number of ZPFs per time unit)
the charge geometry remains the same

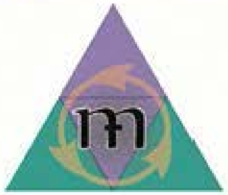

h

Negative Charge ZPF

Nett negative Planck quanta
with South-North m-dipole vector

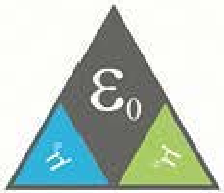

Copyright ABRAHAM [2008] - all rights reserved

$mv^2 = E$

Quantised Angular Momentum

is the equilateral geometry of scalar energy momenta per unit of time

$E = hv^2$

$$\frac{m^2}{s}$$

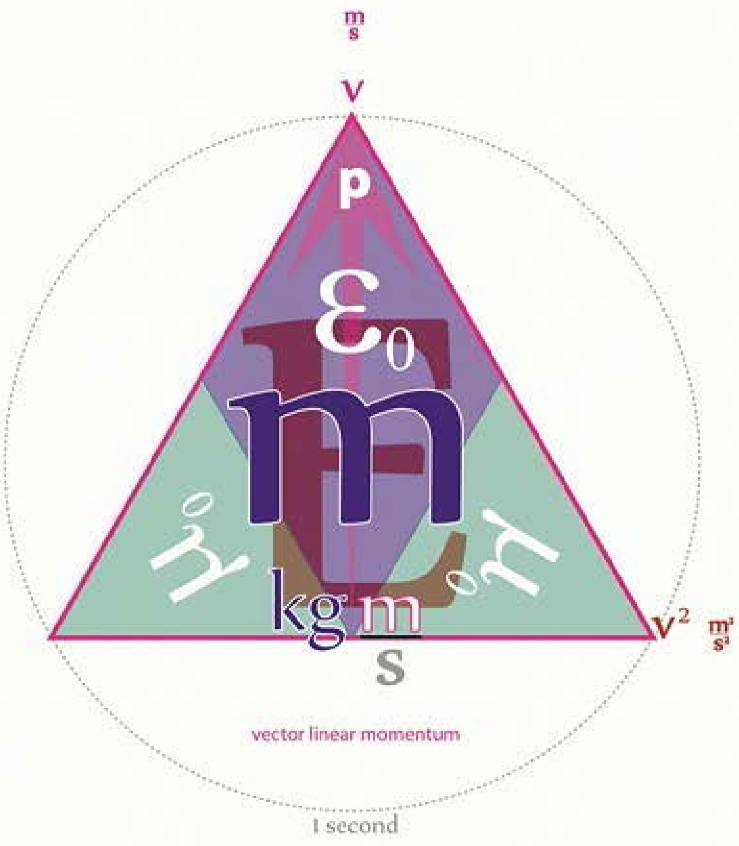

m/s
v
p
ε0
E
m
kg m/s
v²
m²/s²
1 second

vector linear momentum

mΩ

mass-QAM

6.629432672 e-34 J.s

Planck quanta
$[m\Omega v]^2$
mass velocity

7.376238634 e-51 kg

mass

m

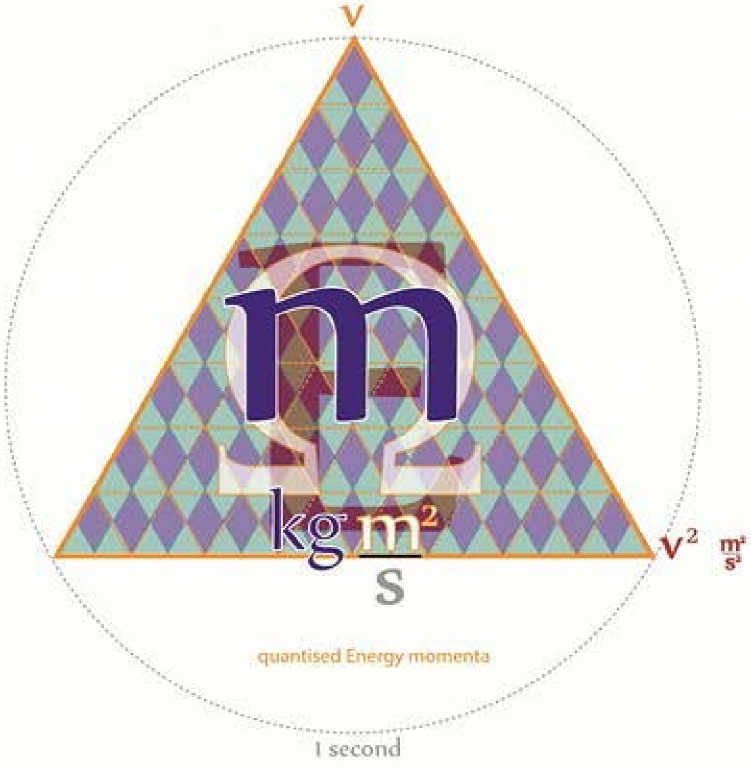

v
m
kg m²/s
v²
m²/s²
1 second

quantised Energy momenta

p

linear momentum

$kg\frac{m^2}{s}$

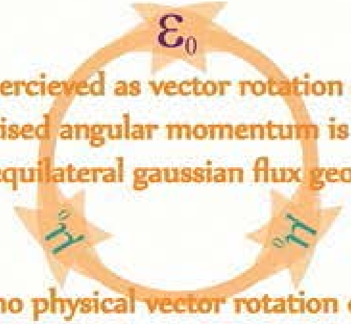

Normally percieved as vector rotation about a point
quantised angular momentum is actually
an equilateral gaussian flux geometry

(there is no physical vector rotation component)

h

Planck's Constant

$kg\frac{m^2}{s}$

Copyright ABRAHAM [2008] - all rights reserved

Charged mass & Matter in motion

produce differing but related measurments of EM force

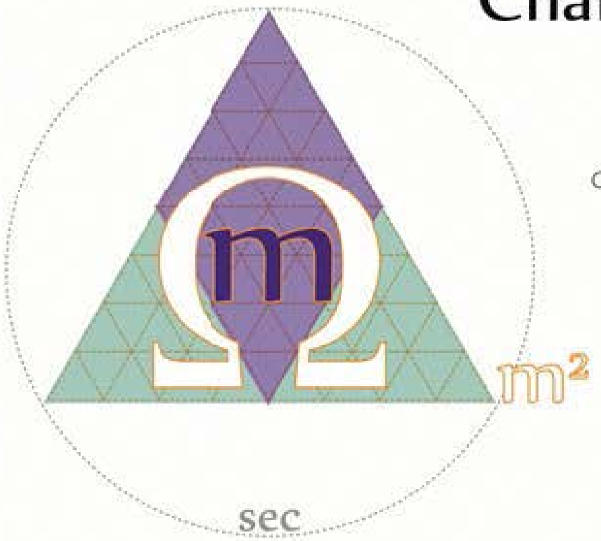

Charged masses per second $\frac{C}{s} \approx kg \approx A$ Charged Matter in motion

Amperes

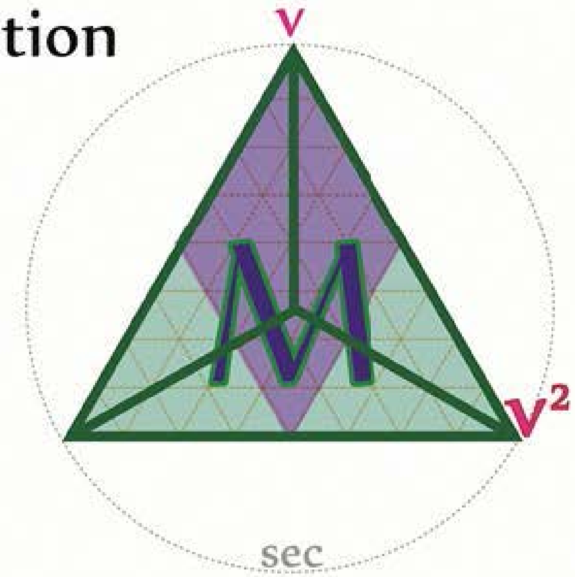

Quantised angular momenta creates CHARGE

By definition in Tetryonics the quantised angular momentum of mass in any defined spatial co-ordinate system provides the quantum basis for

$$\boxed{kg\,\frac{m^2}{s}}$$

charged mass-energies

Vector linear momentum creates CURRENT

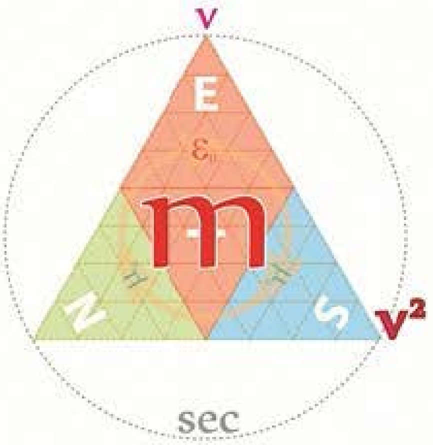

charged mass

$$C \approx kg.s \approx As$$

Coulombs

Amp.sec

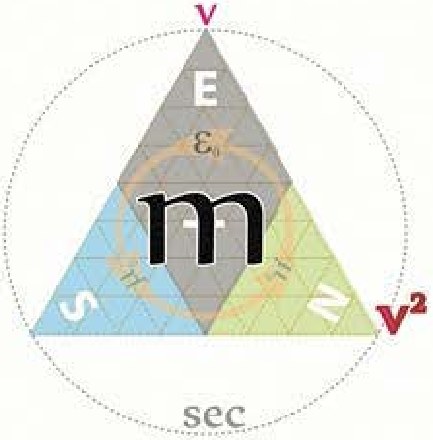

Charge, through its energy interactions, creates the geometric scaffolding for ALL 3D EM mass-Energy-Matter geometries

Copyright ABRAHAM [2008] - all rights reserved

Charged mass geometry

The symmetry of Charge geometries provides a geometric foundation
for all mass-Energy-Matter particles and physical forces

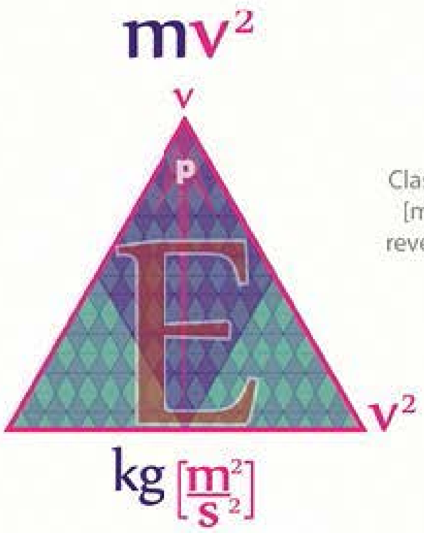
mv^2
v
p
E
v^2
$kg\left[\frac{m^2}{s^2}\right]$

Classical analysis of Energy
[mass velocity] does not
reveal the nature of Charge

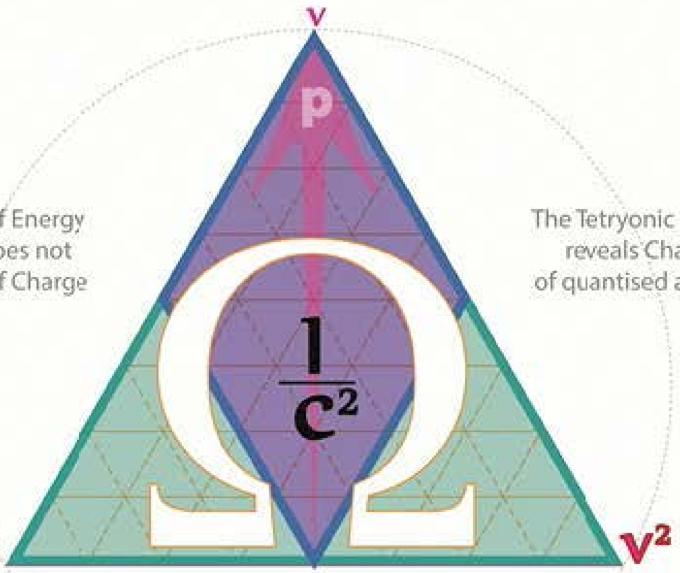
v
p
Ω
$\frac{1}{c^2}$
v^2

The Tetryonic geometry of Energy
reveals Charge as a product
of quantised angular momentum

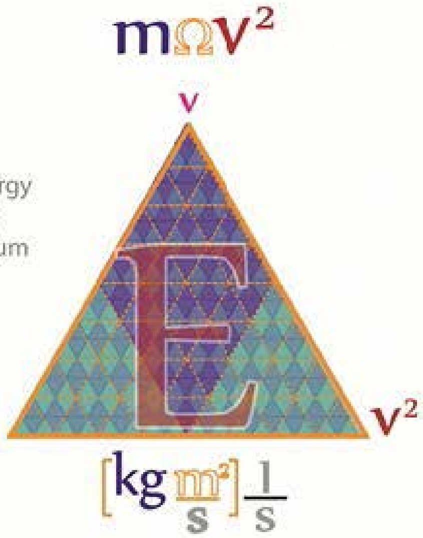
$m\Omega v^2$
v
E
v^2
$\left[kg\frac{m^2}{s}\right]\frac{1}{s}$

Charge is the result of quantised angular momentum
and can also be modelled electrically as the
uni-directional rotation of inductive energies

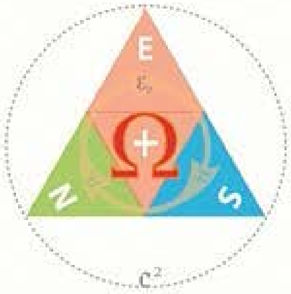
E
ε_0
+
Ω
N
S
c^2

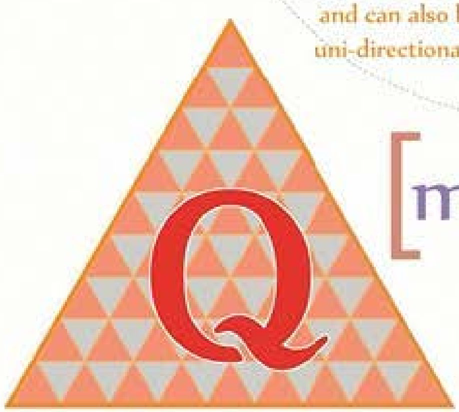

$$\left[m\left[\frac{\Omega}{c^2}\right]v^2\right]$$

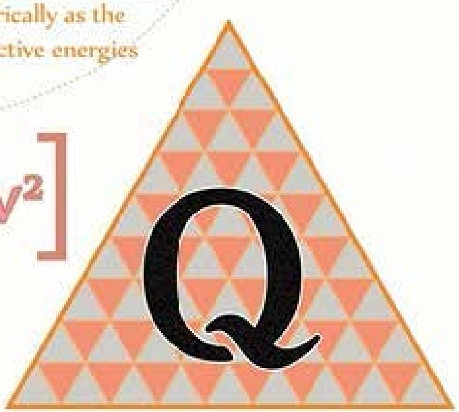

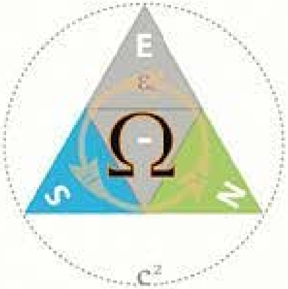
E
ε_0
-
Ω
S
N
c^2

'a un-balanced (broken) energy symmetry resulting from
the nett geometry of equilateral EM mass-energy quanta'

Copyright ABRAHAM [2008] - all rights reserved

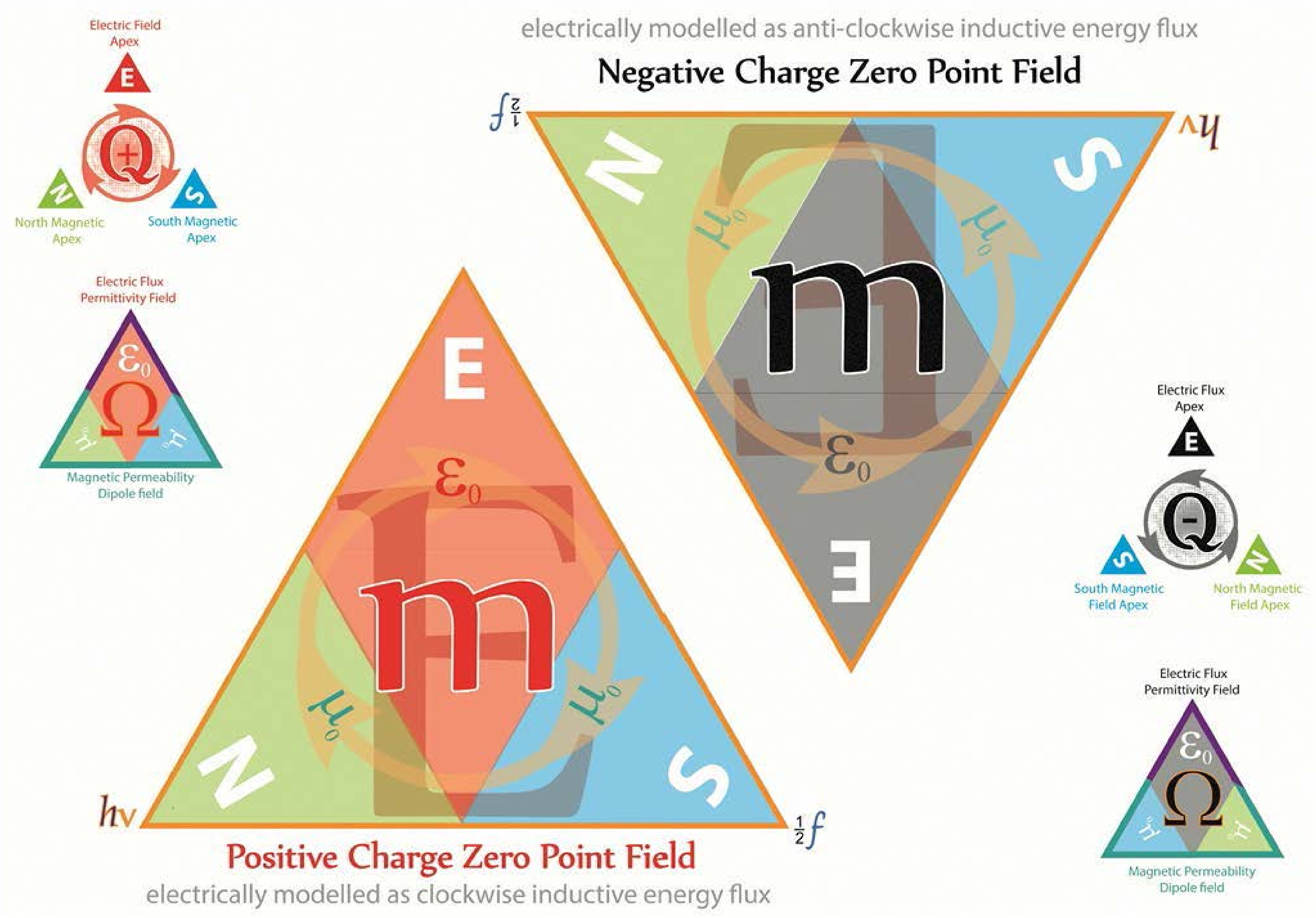

electrically modelled as anti-clockwise inductive energy flux

Negative Charge Zero Point Field

Positive Charge Zero Point Field

electrically modelled as clockwise inductive energy flux

Electric Field Apex
E
Q
+
N
S
North Magnetic Apex
South Magnetic Apex
Electric Flux Permittivity Field
ε_0
Ω
μ_0
μ_0
Magnetic Permeability Dipole field
$\frac{1}{2}f$
hv
N
S
μ_0
μ_0
m
ε_0
E
E
ε_0
m
μ_0
μ_0
N
S
hv
$\frac{1}{2}f$
Electric Flux Apex
E
Q
-
S
N
South Magnetic Field Apex
North Magnetic Field Apex
Electric Flux Permittivity Field
ε_0
Ω
μ_0
μ_0
Magnetic Permeability Dipole field

Copyright ABRAHAM [2008] - all rights reserved

Elementary and quantum Charges

" All known Fermions have charge topologies that are integer 1/3 multiples of the elementary charge "

ε_0
Ω
μ_0
μ_0

It can be shown that the "quantum of charge"
Is the intrinsic angular momentum of a Planck mass

It then follows that Quasi-particles
[Tetryons and Quarks]
have charge topologies with nett charges of 0, 4 and 8

Further, it can be shown that the "elementary charge"
applied to Leptons and Baryons (12 times quantum charge)
Is reflective of their nett charged Matter topologies

topology charge

$$T\pi[\pm q]$$

geometry

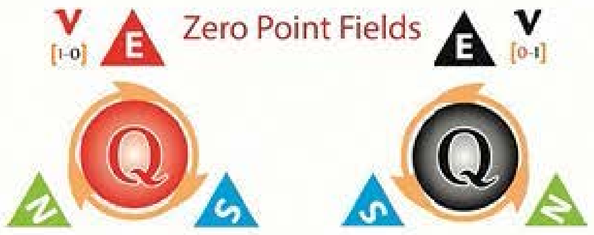

All Charged particles and their respective KEM fields
can be modelled with ZPF field geometries
reflective of their nett Charge topologies

Opposites attract
Similars repel

Quarks

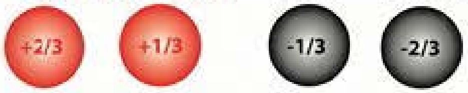

Charmed Top | Strange Bottom | Strange Bottom | Charmed Top

Leptons

Positron
Tau
anti-electron

neutrino
Tau
neutrino

Muon
Tau

Baryons

Proton **Neutron** **Negatron**

quantised Charges
(EM energy flux rotation
within a ZPF geometry)

q
quantum charges

[v] [v]

The elementary charge, usually denoted as e,
is the electric charge carried by a single proton,
or equivalently,
the absolute value of the electric charge
carried by a single electron

[v-v]

Q
Elementary charges

Elementary Charges
are the nett charge created
by mass-energy-Matter
topologies

Tetryons

4 [4-0] | 0 [2-2] | 4 [0-4]

Quarks

up 8 [10-2] | down 4 [8-4] | down 4 [4-8] | up 8 [2-10]

12 [12-0] e | 0 [6-6] νe | 0 [6-6] νe | 12 [0-12] e

Leptons

Baryons

Proton 12 [24-12] | Neutron 0 [18-18] | Negatron 12 [12-24]

All Charges seek Equilibrium

Copyright ABRAHAM [2008] - all rights reserved

Sir Issac Newton

(25 December 1642 – 20 March 1727)

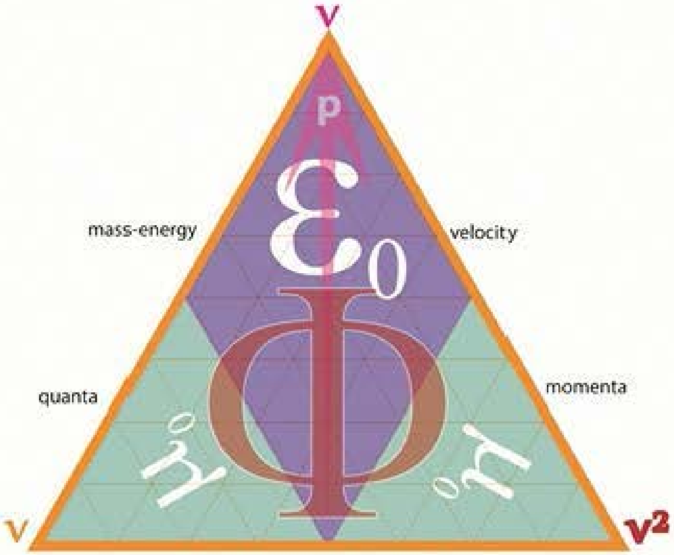

Any changes to velocity
result in changes to Energy-momenta
within a charged geometry

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Newton's first law of motion says:
"A body maintains the current state of motion unless acted upon by an external force."

$$F = ma$$

Inertia is the resistance of any physical object to a change in its state of motion or rest, or the tendency of an object to resist any change in its motion.

It is proportional to an object's mass.

Ideal Inductive loops (ZPF quanta) will oppose ANY changes to their energy levels and consequently, Inertia can be viewed as an outcome of quantum ZPF self-Inductance

↑ [v-v] ↓

+

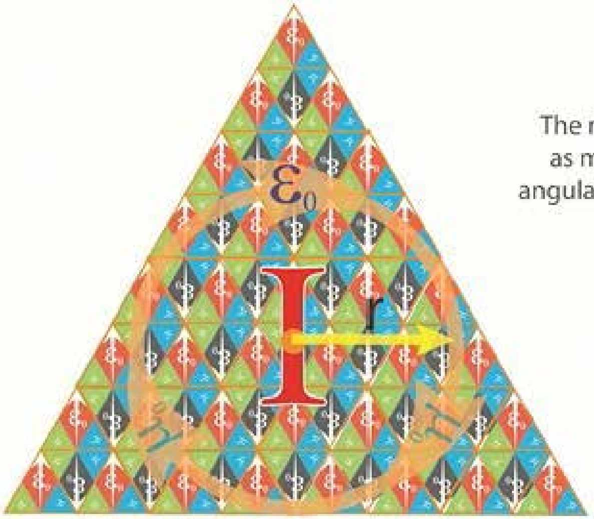

Inertia

The moment of inertia plays much the same role in rotational dynamics as mass does in linear dynamics, describing the relationship between angular momentum and angular velocity, torque and angular acceleration, and several other quantities

The difference between the impeding (inertial) linear momenta and the co-linear (supportive) momenta is a result of the nett Planck mass-energy momenta within the charged geometries of Matter & its KEM fields

↓ [v-v] ↑

−

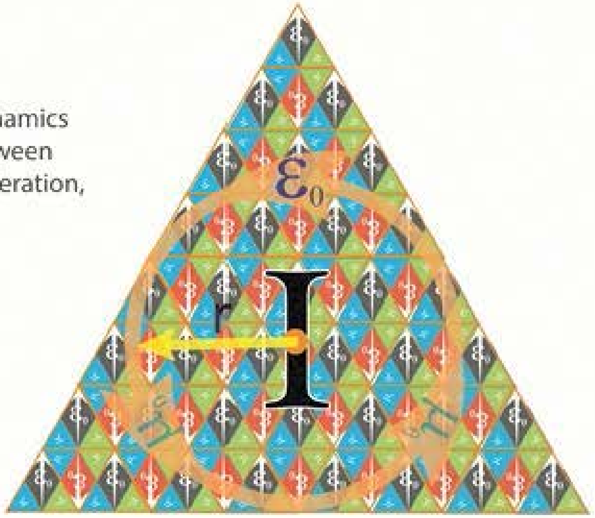

Planck Quanta

are often defined as the "quantum of action"

h

Planck's Constant

$$kg\,\frac{m^2}{s}$$

The Planck constant (denoted h), is a geometric constant resulting from QAM in quantum mechanics

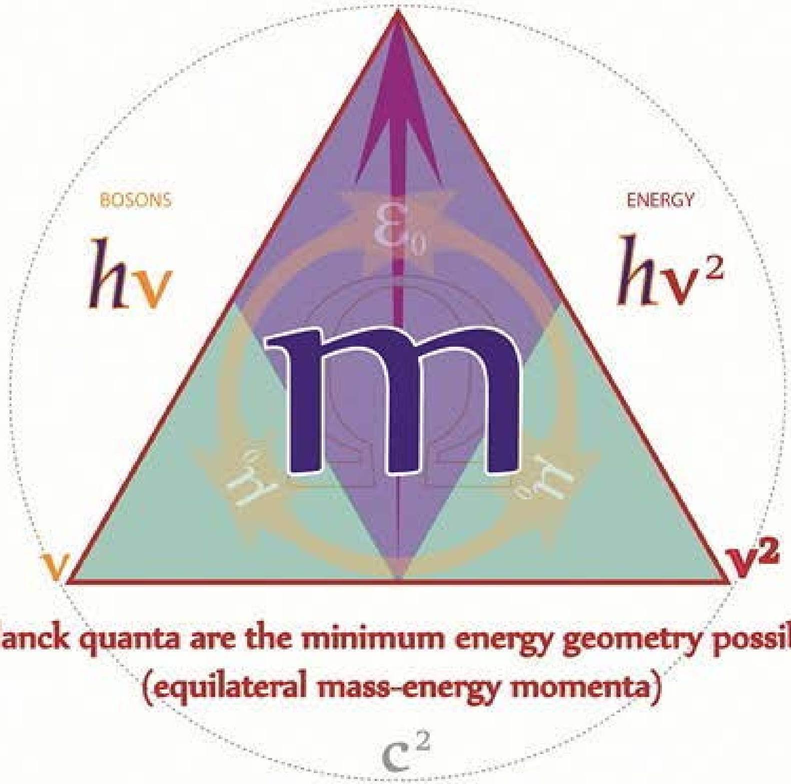
BOSONS
hv
ENERGY
hv²
ε₀
m
v
v²
c²

Planck quanta are the minimum energy geometry possible (equilateral mass-energy momenta)

hv

$m\Omega v^2$

Planck quanta per second

$$kg.\left[\frac{m^2}{s}\right].s^{-1}$$

Planck's constant relates the energy in one quantum (photon) of electromagnetic radiation to the frequency of that radiation

Charge is a result of QAM geometry

quantised angular momentum is the result of equilateral geometry

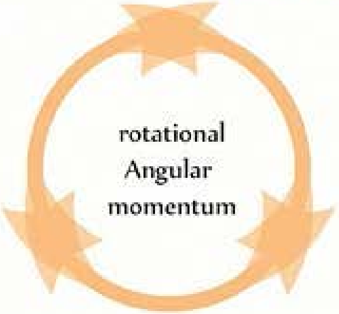
rotational
Angular
momentum

≈

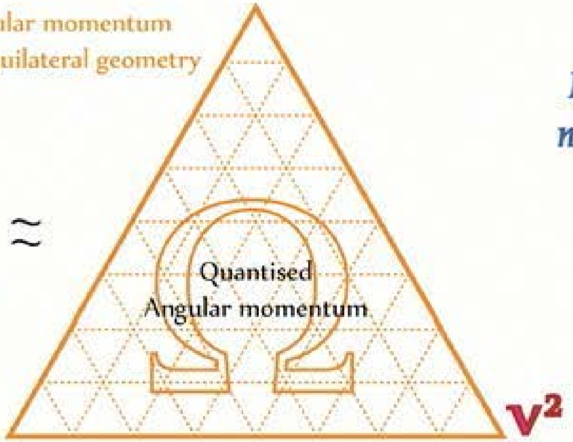
Quantised
Angular momentum
v²

Photons have neutral quanta geometry

Copyright ABRAHAM [2008] - all rights reserved

$E = n.h\nu$

Planck Number

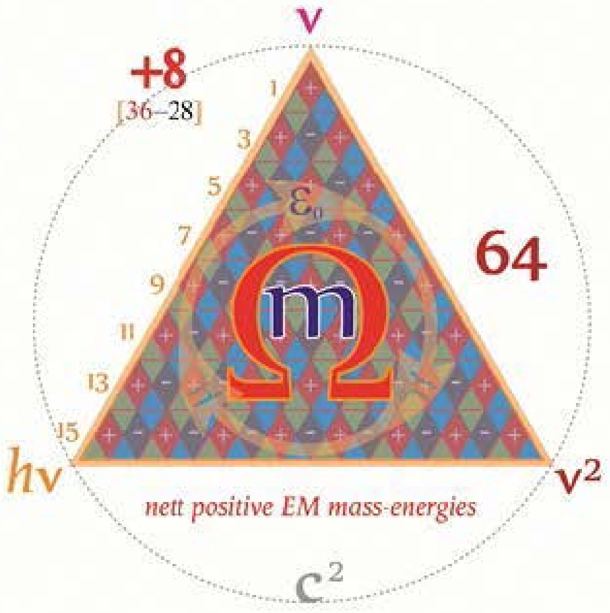
v
+8
[36–28]
1
3
5
7
9
11
13
15
ε_0
m
Ω
64
hv
v²
c²

nett positive EM mass-energies

Max Planck

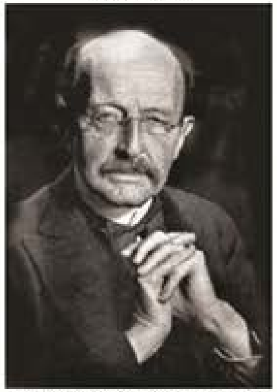

(April 23, 1858 – October 4, 1947)

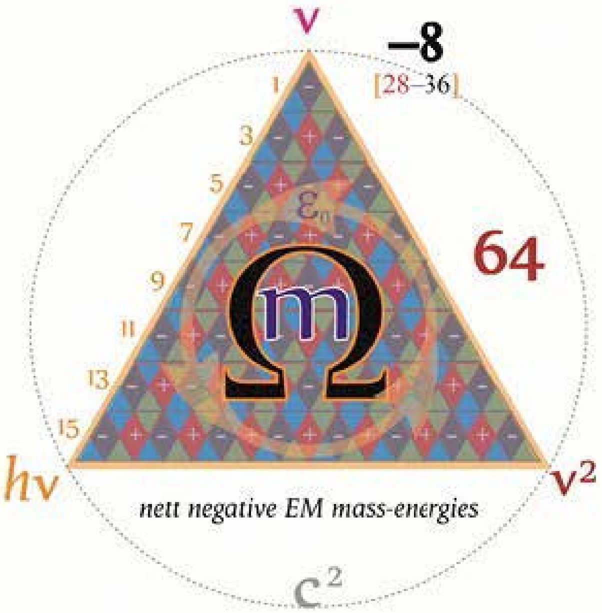
v
–8
[28–36]
1
3
5
7
9
11
13
15
ε_0
m
Ω
64
hv
v²
c²

nett negative EM mass-energies

hv
tranverse
Bosons

hf
longitudinal
Photons

$1e19\nu = n = 5e18f$

quantum number

BOSONS
are the quanta of Charge

$1\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Boson · EM Field · ElectroMagnetic · Planck · quanta · mass · velocity

Equating number of quanta
is the source of a number

$h\nu \neq hf$

Planck Einstein

$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Photon · EM Field · ElectroMagnetic · Planck · quanta · mass · velocity

to photon frequency
of quantum misconceptions

PHOTONS
are the neutral quanta of EM waves

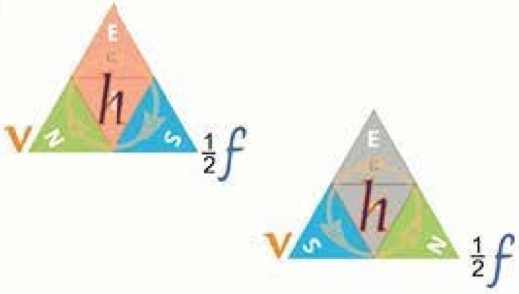
E
h
v
N
S
$\frac{1}{2}f$
E
h
v
S
N
$\frac{1}{2}f$

The Planck constant was first described as the proportionality constant between
the energy (E) of a photon and the frequency (v) of its associated electromagnetic wave.
[it in fact describes the energy of bosons]

The zero-point energy for a
simple harmonic oscillator of frequency f
is ½ hf

The relation between the energy and frequency is called the Planck–Einstein relationship.

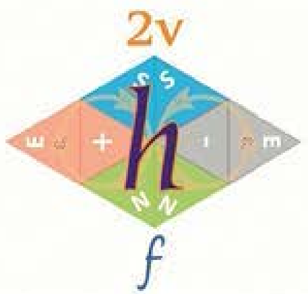
2v
S S
E
+
h
-
E
N N
f

Copyright ABRAHAM [2008] - all rights reserved

Quantum Inductors

+ ZPF

A positive ZPF can be viewed as a quantun quoin or an quantum 'ideal' inductor with an internalised energy flux that is the opposite of a negative ZPF

Clockwise current flow

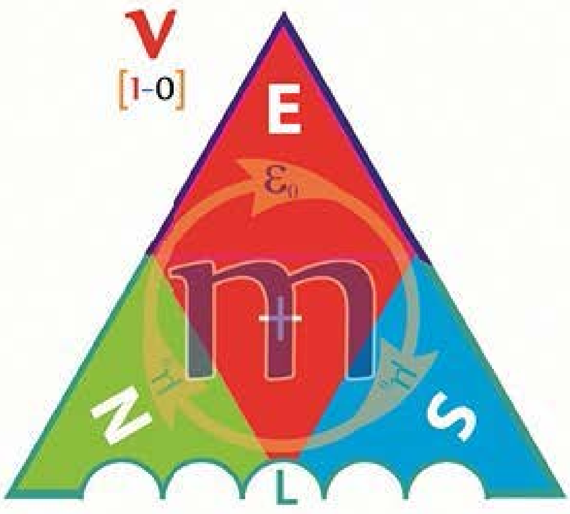

POSITIVE

Quantum L circuit

[quantum inductive circuit]

An "ideal inductor" has inductance, but no resistance or capacitance, and will not dissipate energy (until it interacts with other ZPFs or Matter) and forms the basis for all Charge-Parity-Time [CPT] interactions

A ZPF is fixed in either a + or - state

[Quantum Inductor circuit]

Its energy flux direction as modelled electrically] is relative to the observer's view or the direction of measurement

Charge polarity is opposed on opposite faces of the same quoin [quantum coin]

The direction of inherent energy flux from the presective of the observer determines ZPF charge polarity

The Quantum Inductor (L) circuit stores energy as EM mass in π geometries, it does not oscillate

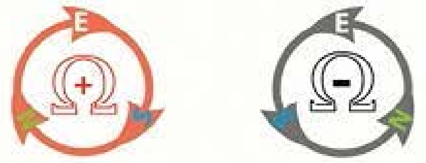

Energy received is stored indefinitely until its release via weak interaction

[Inductive Magnetic coupling]

The quantum Inductive circuit is a SINGLE charge tri-field inductive energy loop

It does not oscillate energy between two opposing charges its differing energy fields are the reuslt of its equilateral QAM geometry storing Electric energy in its E field, and Magnetic energy in its M field

The direction of the QAM flux that models inertia is relative to the observer

- ZPF

A negative ZPF can be viewed as a quantum quoin or an quantum 'ideal' inductor with an internalised energy flux that is the opposite of a positive ZPF

Counter-Clockwise Current Flow

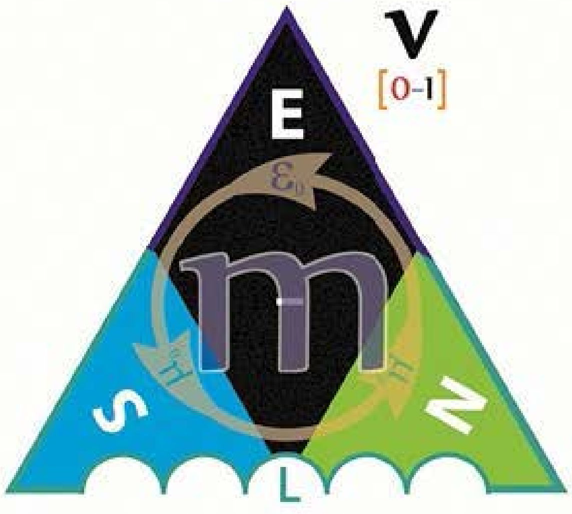

NEGATIVE

Quantum L circuit

[quantum inductive circuit]

Quantum Inductors and EM energy Levels

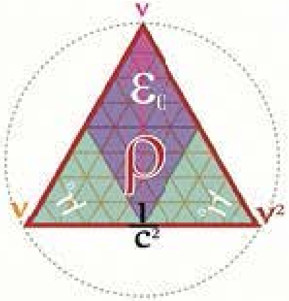

coupled ODD# mass-energies
form SQUARE energy geometries

quantum Inductors

Scalar Energies

[ZPFs]

[Bosons]

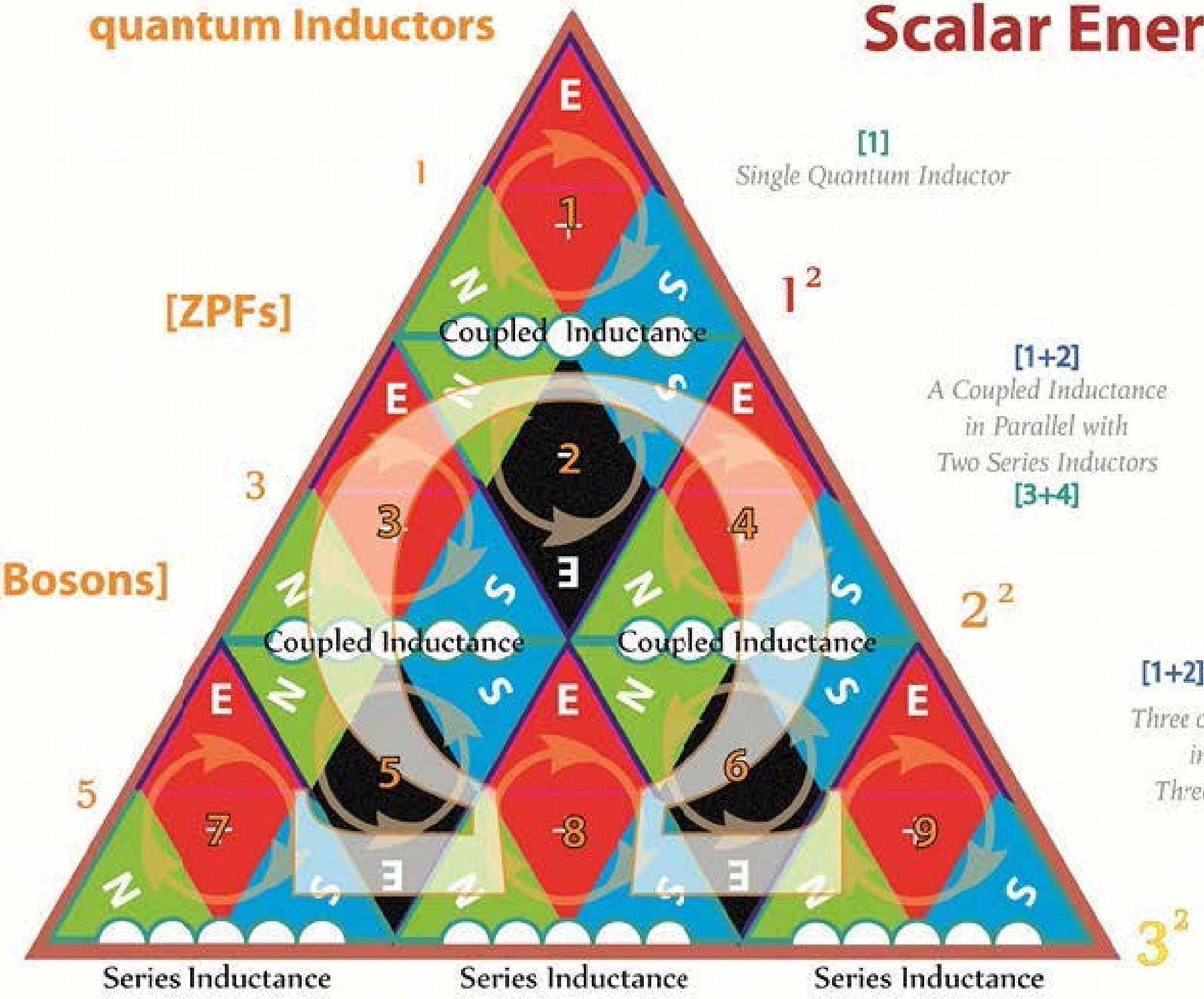
1
1²
3
2²
5
3²
E
1
N
S
Coupled Inductance
E
E
2
3
4
N
S
Ǝ
N
S
Coupled Inductance
Coupled Inductance
E
E
E
5
6
7
8
9
N
S
Ǝ
Ǝ
Series Inductance
Series Inductance
Series Inductance

[1]
Single Quantum Inductor

[1+2]
Coupled
Quantum Oscillator

[1+2]
A Coupled Inductance
in Parallel with
Two Series Inductors
[3+4]

E

[1+2] [3+5] [4+6]
Three coupled Inductances
in Parallel with
Three Series Inductors
[7+8+9]

All quantum levels form inductive magnetic dipole bases
[Weak Force]

Quantum energy levels can be viewed as various combinations of:

Coupled Inductances

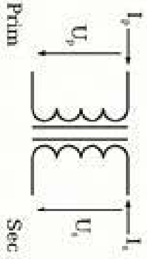
Prim
U_p
I_p
Sec
U_s
I_s

Parallel Inductances

$$L_{total} = \frac{1}{\frac{1}{L_1} + \frac{1}{L_2} + \ldots \frac{1}{L_n}}$$

Series Inductances

$$L_{total} = L_1 + L_2 + \ldots L_n$$

The Golden Triangle

Charged mass-energy geometries

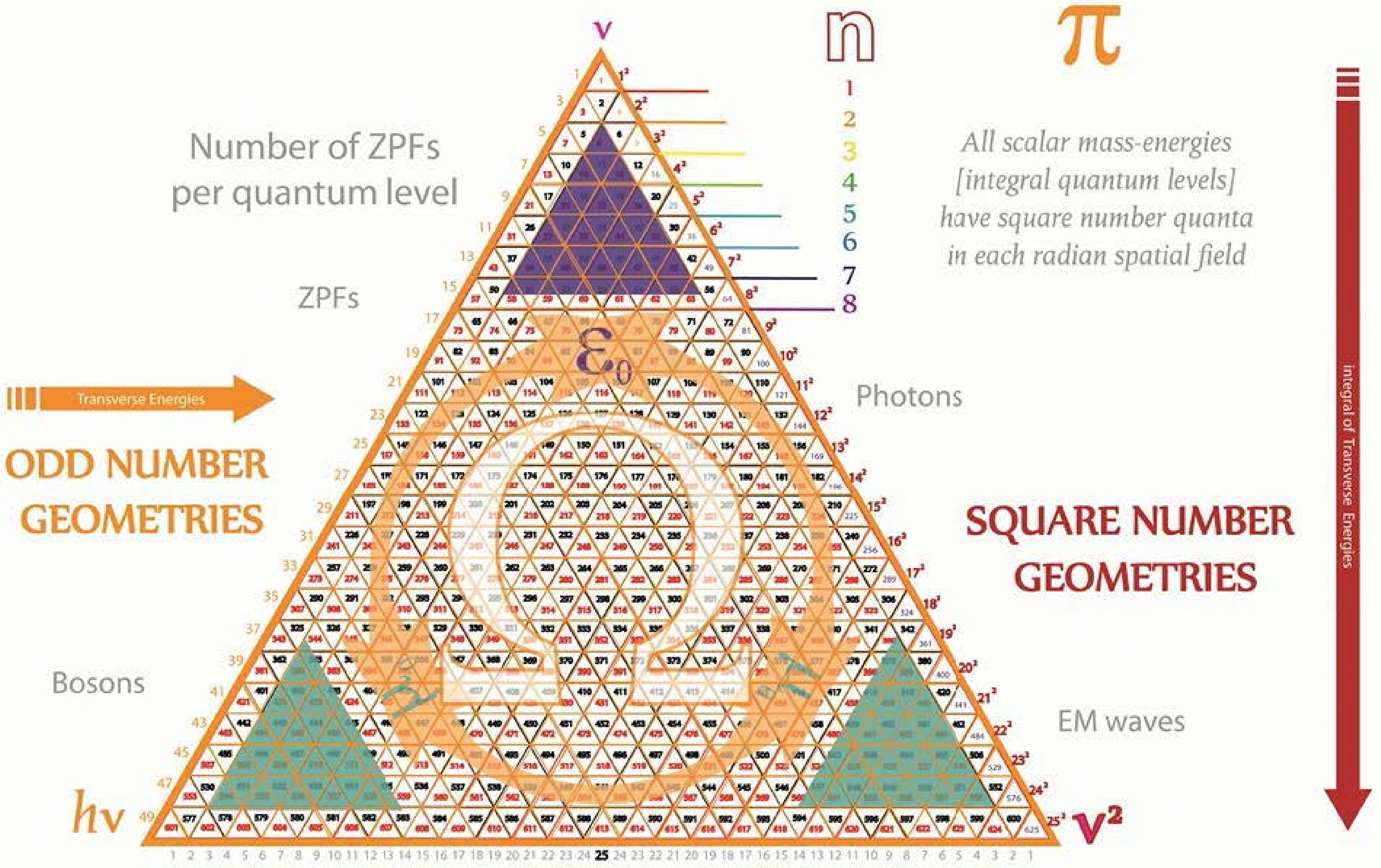
v
n
π
1
2
3
4
5
6
7
8
ε0
Number of ZPFs per quantum level
ZPFs
Transverse Energies
Photons
Bosons
EM waves
hv
v²
Integral of Transverse Energies
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
1² 2² 3² 4² 5² 6² 7² 8² 9² 10² 11² 12² 13² 14² 15² 16² 17² 18² 19² 20² 21² 22² 23² 24² 25²
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

All scalar mass-energies [integral quantum levels] have square number quanta in each radian spatial field

ODD NUMBER GEOMETRIES

SQUARE NUMBER GEOMETRIES

Probabilities **NORMAL DISTRIBUTIONS** Wavefunctions

Copyright ABRAHAM [2008] - all rights reserved

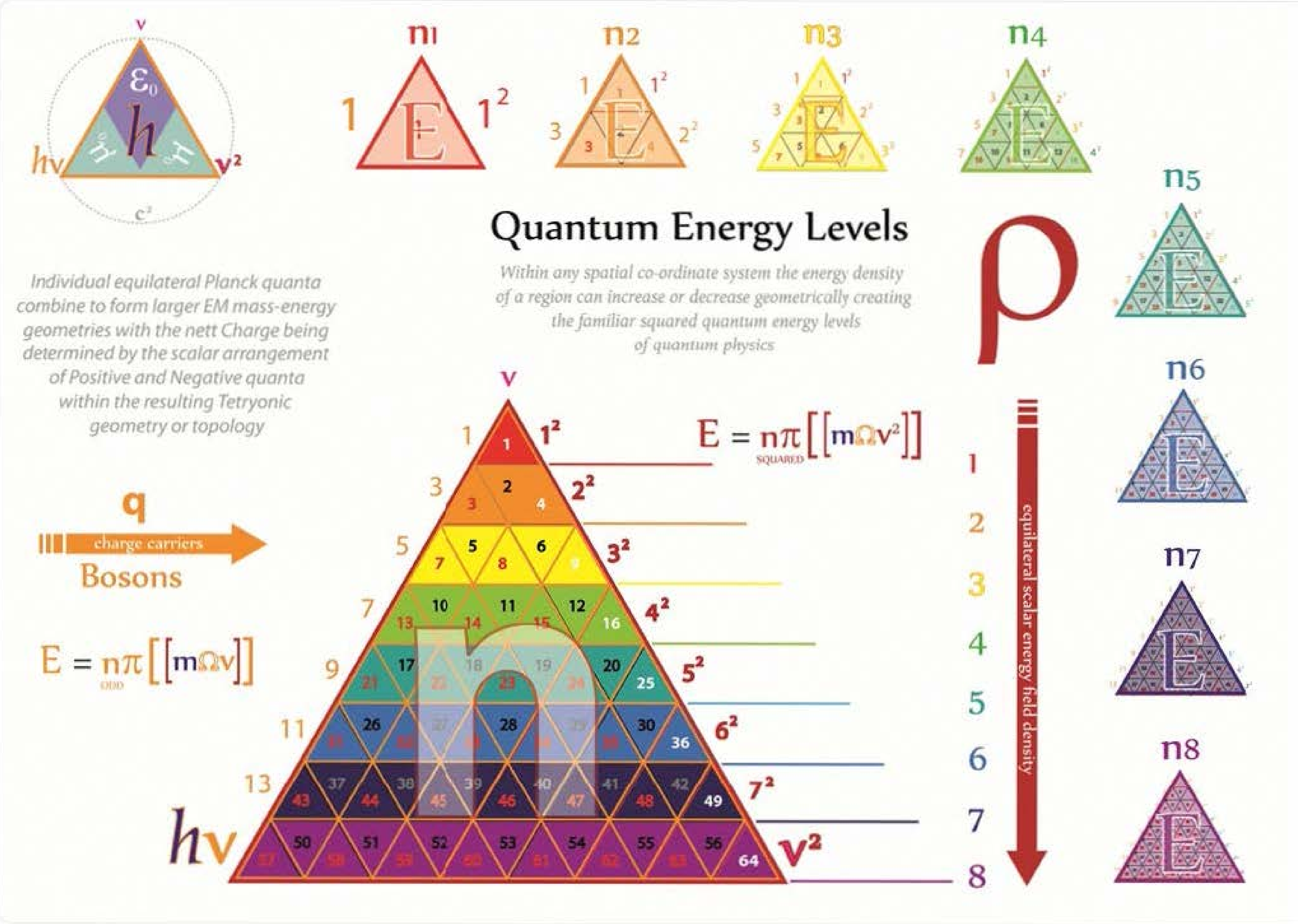

v
ε0
h
hv
v²
c²
n1
n2
n3
n4
n5
n6
n7
n8
ρ
q
charge carriers
v
1
3
5
7
9
11
13
1²
2²
3²
4²
5²
6²
7²
hv
v²
1
2
3
4
5
6
7
8
equilateral scalar energy field density

Quantum Energy Levels

Within any spatial co-ordinate system the energy density of a region can increase or decrease geometrically creating the familiar squared quantum energy levels of quantum physics

Individual equilateral Planck quanta combine to form larger EM mass-energy geometries with the nett Charge being determined by the scalar arrangement of Positive and Negative quanta within the resulting Tetryonic geometry or topology

Bosons

$$E = \underset{\text{SQUARED}}{n}\pi\left[\left[m\Omega v^2\right]\right]$$

$$E = \underset{\text{ODD}}{n}\pi\left[\left[m\Omega v\right]\right]$$

Tetryonics 03.13 - Quantum Energy levels

Magnetic Vectors

Intrinsic Magnetic vectors are transverse to E fields

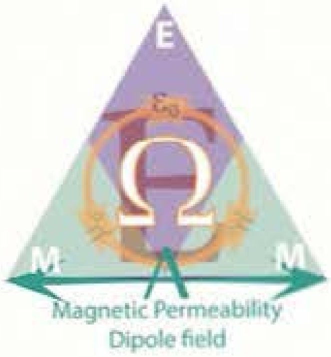
E
ε_0
Ω
M
M
A
Magnetic Permeability
Dipole field

External Magnetic fields are termed B fields

Opposite charges moving in opposing directions produce co-directional Magnetic vectors

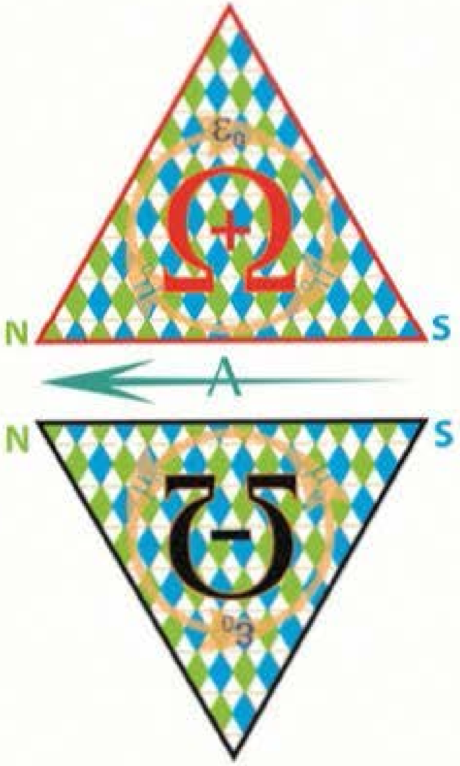
ε_0
+
N
S
A
N
S
-
ε_0

Magnetic vectors can be modelled geometrically or electrically through energy field fluxes [quantised angular momenta]

Within a Magnetic dipole

Magnetic Vector A

energy 'flows' South to North

External to a Magnetic dipole energy 'flows' North to South

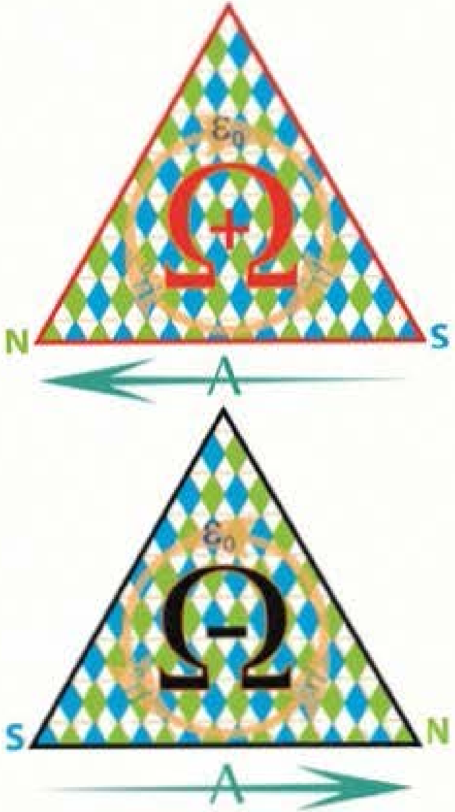
ε_0
+
N
S
A
ε_0
-
S
N
A

Opposite charges moving in the same direction produce Magnetic vectors in opposition

Charge Fields

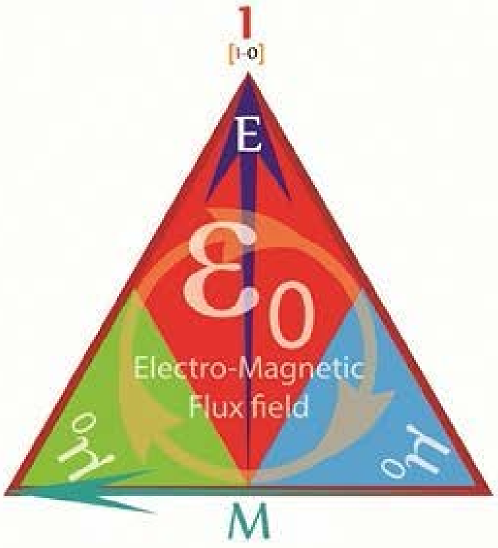
1
[1-0]
E
ε0
Electro-Magnetic
Flux field
μ0
μ0
M

EM energy fluxes in a Positive ZPF flows
are electrically modelled as Clockwise
(from North to South)

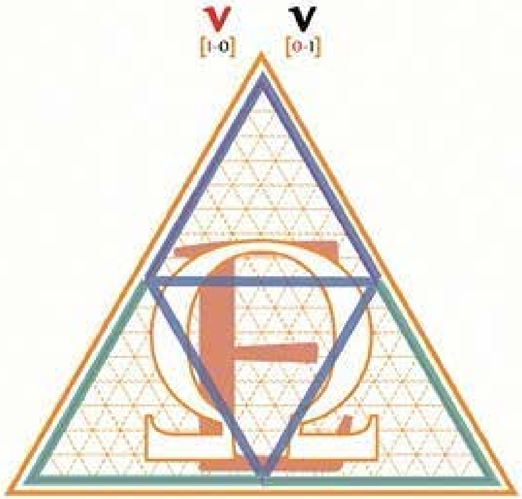

The EM flux directions of Charges
can be modelled vectorially with
Electric and Magnetic vectors

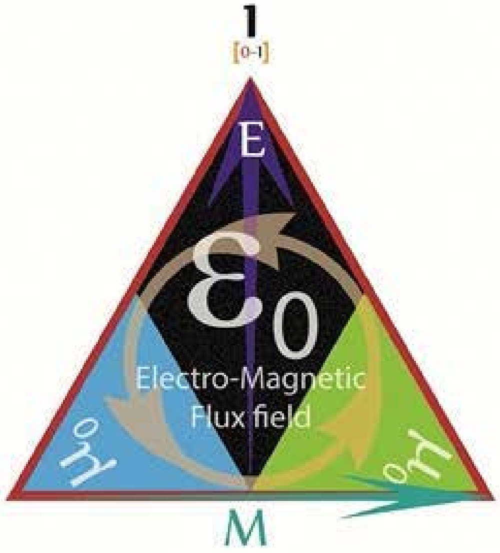
1
[0-1]
E
ε0
Electro-Magnetic
Flux field
μ0
μ0
M

EM energy fluxes in a Negative ZPF
are electrically modelled as counter-Clockwise
(from North to South)

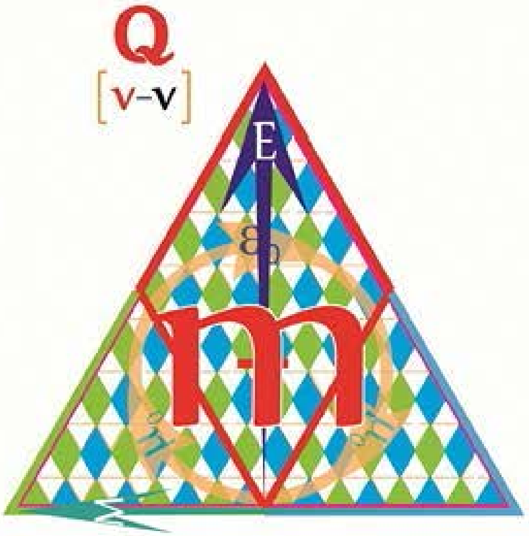
Q
[v-v]
E
ε0
m
μ0
μ0

All rotational planck energy fluxes can serve as models for the nett quantised angular momenta of any mass-energy geometry

Polarised
Electric and Magnetic fluxes
in ElectroMagnetic fields arise from
intrinsic quantised angular momentum

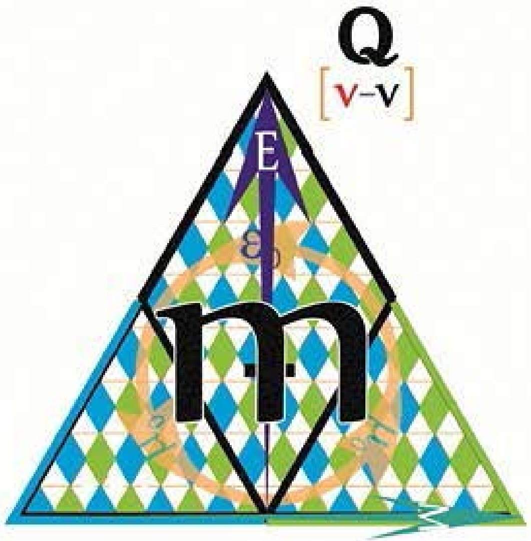
Q
[v-v]
E
ε0
m
μ0
μ0

Copyright ABRAHAM [2008] - all rights reserved

EM fields

are the combined Electric & Magnetic fields
resulting from the quantised angular momentum
of mass-energy in any region of free space

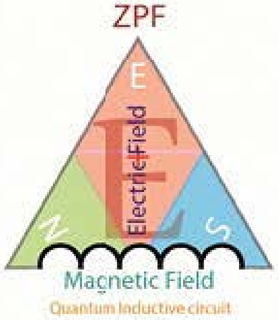

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

EM Permittivity-Permeability is a measure of how much resistance
is encountered when the quantised angular momenta of
EM energies form an electro-magnetic field in a vacuum

Celeritas = 299,792,458 $\frac{m}{s}$

$$c_0 = \frac{1}{\sqrt{\mu_0\varepsilon_0}}.$$

EM field Permittivity-Permeability

$$\varepsilon_0\mu_0 = 1.112650056\text{e-}17\ \frac{s^2}{m^2}$$

$$\left[\frac{A^2\ s^4}{kg\ m^3}\cdot\frac{kg\ m}{A^2\ s^2}\right]$$

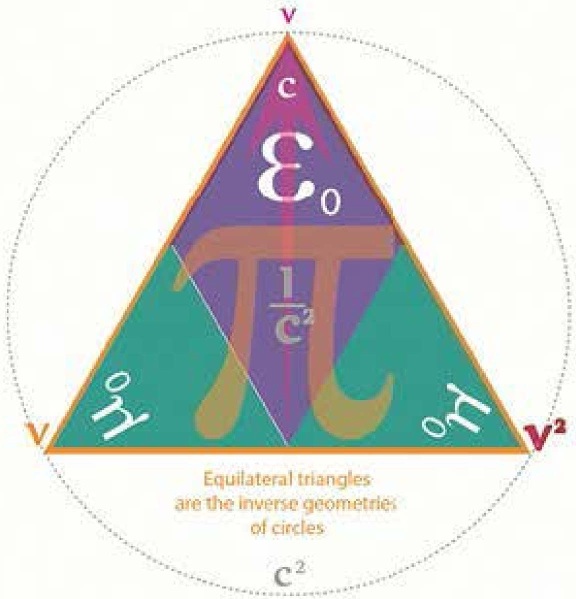

Equilateral triangles
are the inverse geometries
of circles

m		nπ	$[\varepsilon_0\mu_0]$	$[m\Omega v^2]$
mass	=	charged	Spatial geometry	Energy density

Electric Constant

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

$$= 8.85418785\text{e-}12\ \frac{F}{m}\qquad \frac{A^2\ s^4}{kg\ m^3}$$

The permittivity of empty space, equal to 1 in centimeter-gram-second
electrostatic units and numerically, to 8.854 × 10-12 farad per meter
in International System units, where c is the speed of light in meters per second.
Symbolized ε0.

Magnetic Constant

$$k_0 = \frac{1}{4\pi\varepsilon_0}$$

$$= 1.25663706\text{e-}6\ \frac{H}{m}\qquad \frac{kg\ m}{A^2\ s^2}$$

A measure of the degree to which molecules of some material polarize (align)
under the influence of an electric field; symbol k0, units F/m (farads per metre).

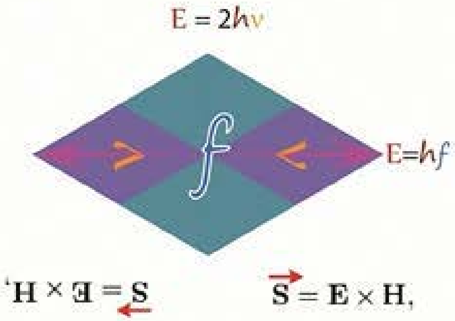

The Energy-momenta
of ZPFs form
natural Poynting vectors

EM field Permittivity

The Electric constant, commonly called
the vacuum permittivity, or permittivity of free space,
relates the units for electric charge to
mechanical quantities such as
length and force.

The name Vacuum Permittivty is a misnomer and should be replaced with the correct term EM field Permittivity

The strength of Electric fields is determined by the Electrical Permittivity Constant

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

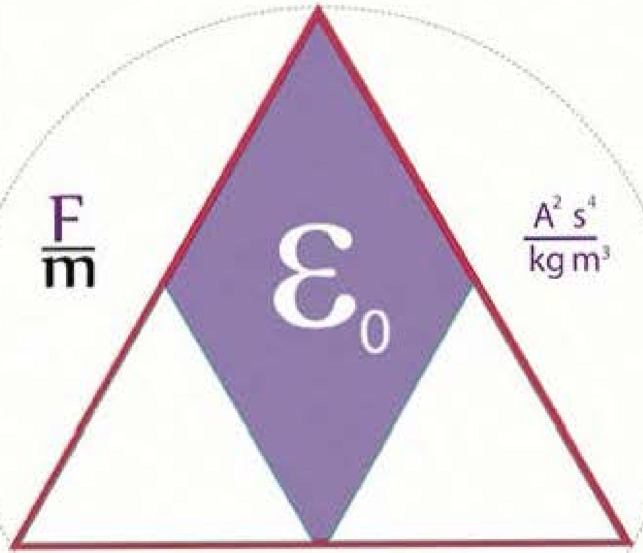

8.85418785 e-12

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

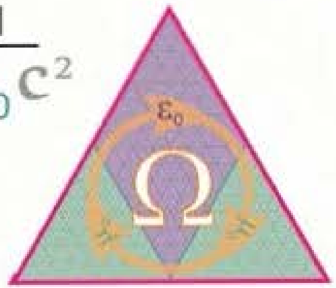

"Ampere's Law states that for any closed loop path, the sum of the quantities (B.ds) for all path elements into which the complete loop has been divided is equal to the product of μ0 and the total current enclosed by the loop."

$$k = \frac{1}{4\pi\varepsilon_0}$$

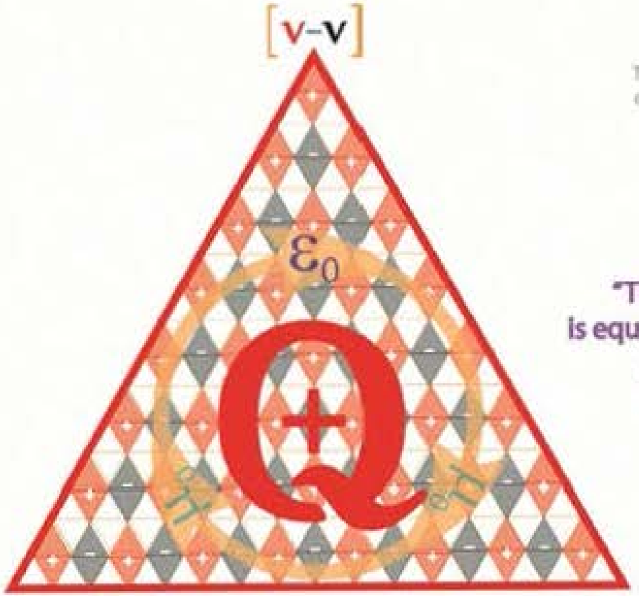

Positive Charge Electric Field

The permittivity of empty space, equal to 1 in centimeter-gram-second electrostatic units and to 107/4πc2 farads per meter or, numerically, to 8.854 × 10-12 farad per meter in International System units, where c is the speed of light in meters per second.

Gauss' Law:

"The total of the electric flux out of a closed surface is equal to the charge enclosed divided by the permittivity"

This applies equally to any geometry chosen to tessellate a surface area

Superpostioned E fields
gives rise to Coulomb Forces

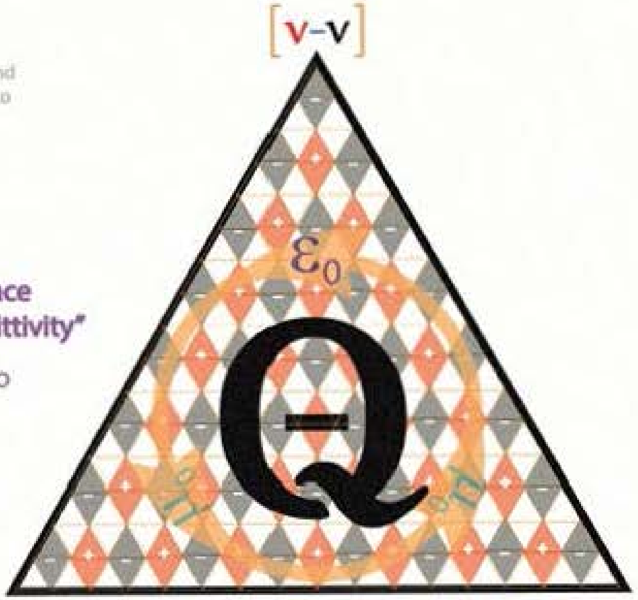

Negative Charge Electric Field

Electric permittivity Fields

Negative externalised Planck quanta
(Counter-clockwise energy fluxs)

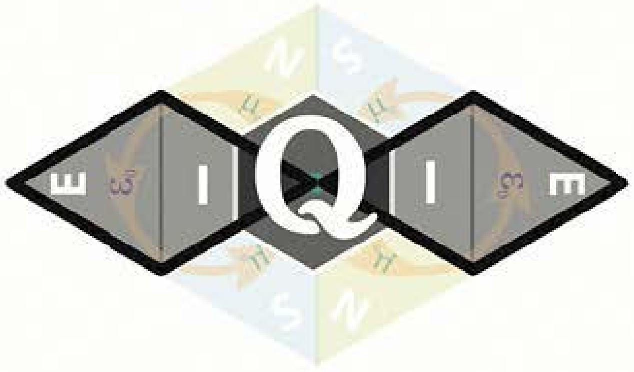

Positive externalised Planck quanta
(Clockwise energy fluxs)

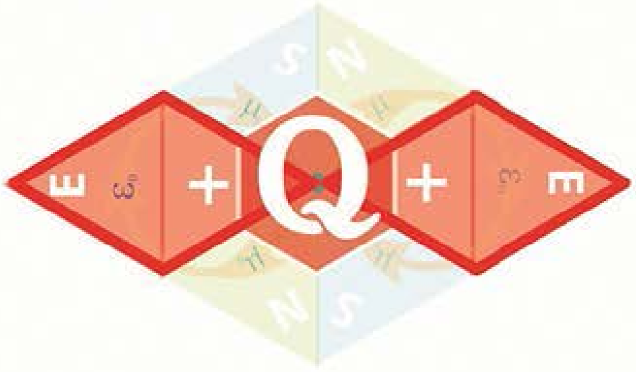

Coupled same charge ZPFs
have **neutralised** Magnetic fields

*In Electro-statics superpositioned E fields
with interactive energy momenta
are the interactive mechanism
for Coulombic forces*

Electrostatic Particles
in motion have Kinetic energies
resulting in Magnetic moments

Polar view

Negative E-fields

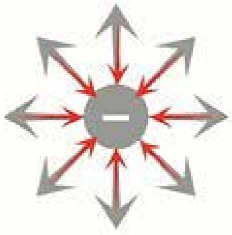

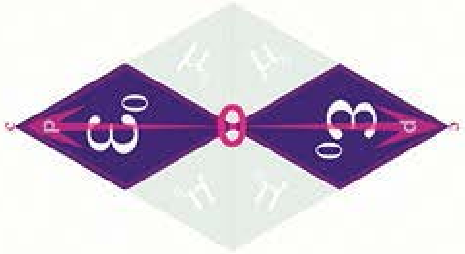

Polar view

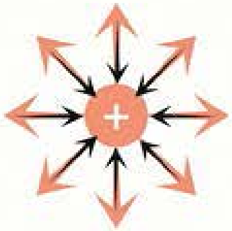

Positive E-fields

NEGATIVE electrostatic E fields
attract positive charges and
repel negative charges

ε_0

POSITIVE electrostatic E fields
attract negative charges and
repel positive charges

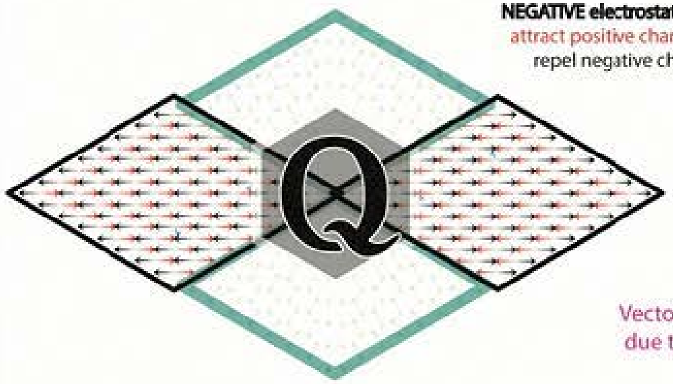

*Opposites Attract
Similars Repel*

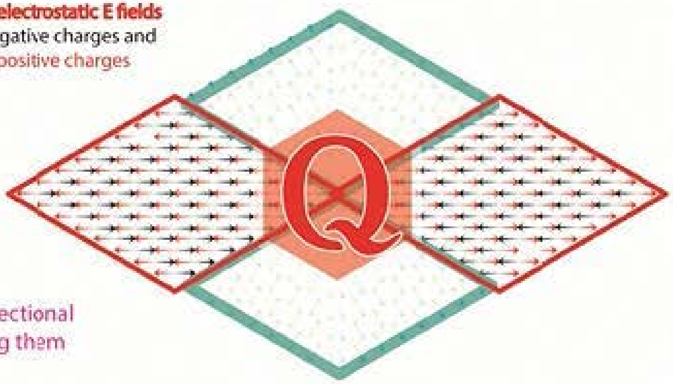

Vectorial momenta forces in EM fields are bi-directional
due to the energy-momenta quanta comprising them

The currently stated 'standard' premise of Electrical Energy flowing from Positive to Negative is misleading (as Energy also flows from Negative to Positive at the same time)

EM field Permeability

The permeability of free space,
also called absolute permeability.

The name Vacuum Permeability is a misnomer and should be replaced with the correct term EM field Permeability

The magnetic constant has the value of $4\pi \times 10^{-7}$ henry per meter.

The strength of Magnetic fields is determined by the Magnetic Permeability Constant

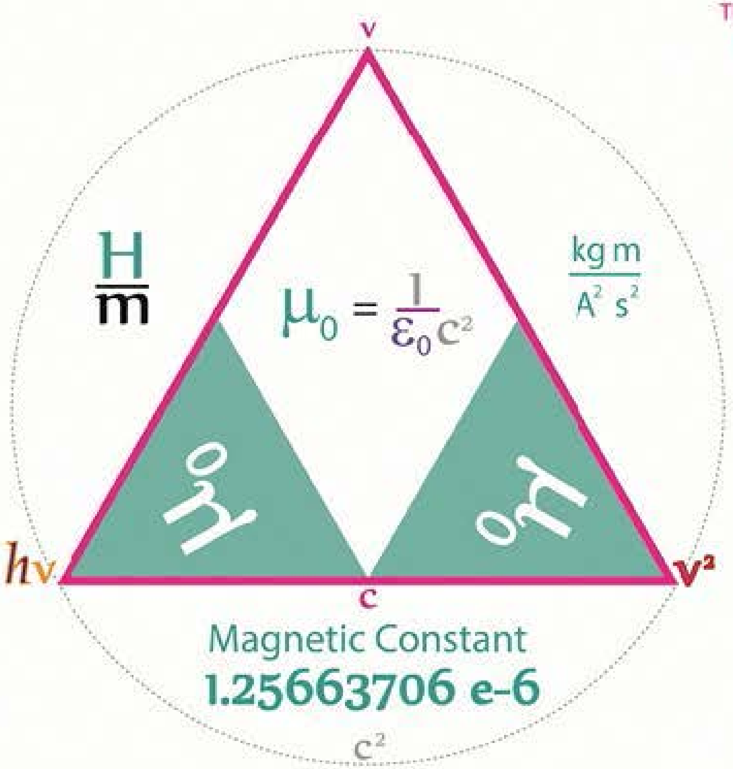

Magnetic Constant
1.25663706 e-6

F

The magnetic field is most commonly defined in terms of the Lorentz force it exerts on moving electric charges.

The magnetic field generated by a steady current (a constant flow of electric charges in which charge is neither accumulating nor depleting at any point) is described by the Biot–Savart law

$$\mu_0$$

$$B = \mu_0 H$$

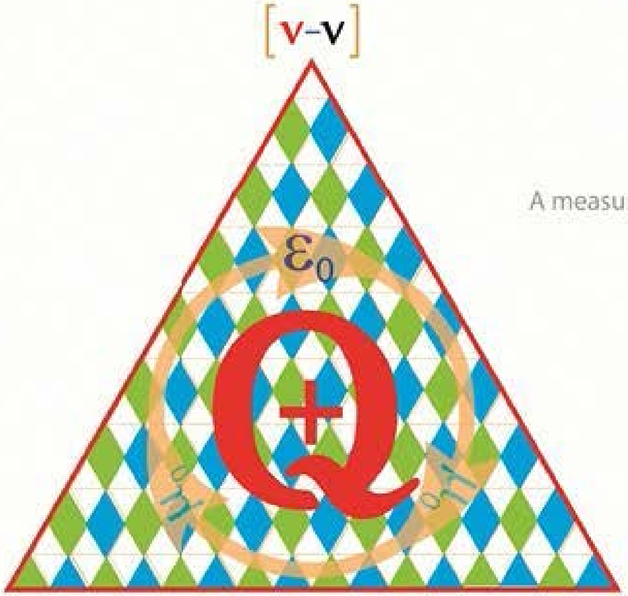

Positive Magnetic Moment

A measure of the degree to which molecules of some material polarize (align) under the influence of an electric field.

Measured in units of units H/m (Henries per metre).

There are
NO magnetic monopoles
[not under any condition]

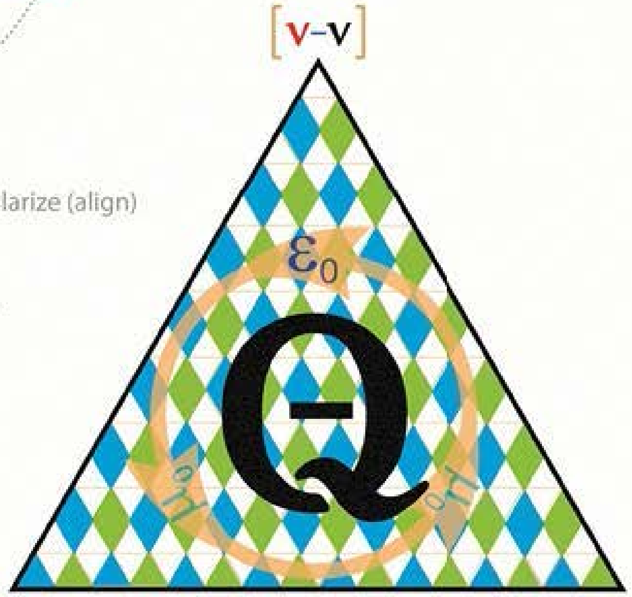

Negative Magnetic Moment

Magnetic permeability Fields

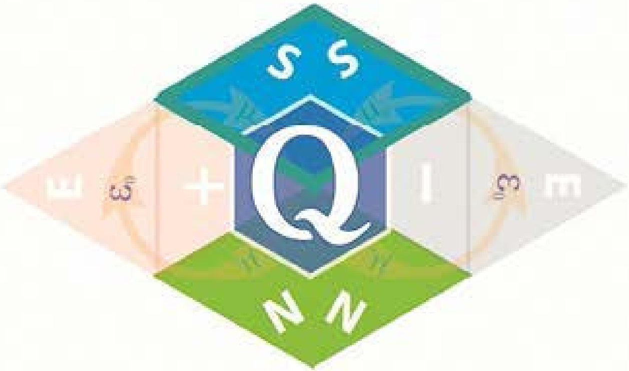

Coupled opposite charge ZPFs
produce **neutralised** Electric fields

*In Magneto statics superpositioned M fields
with interactive energy momenta
are the interactive mechanism
for Lorentz forces*

Magneto-static Particles
have enhanced Magnetic moments

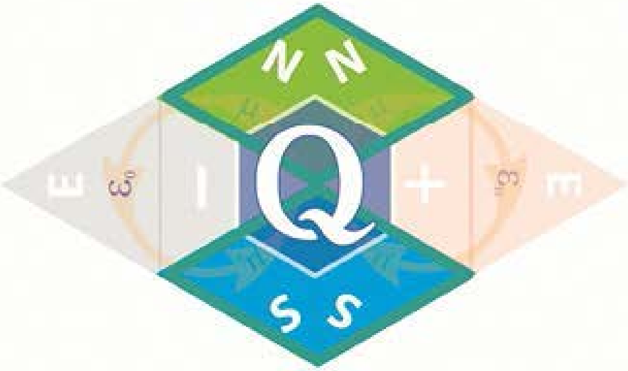

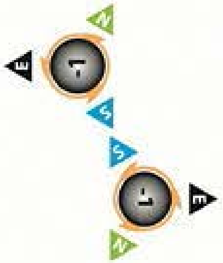

Amperes law

*Current flowing through a wire
will create a magnetic field*

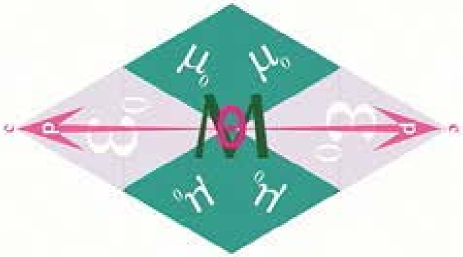

Lorentz forces

*charged particles in motion are subject
to forces due to external EM fields*

*magnetic field forces are
orthagonal to electric Coulomb forces*

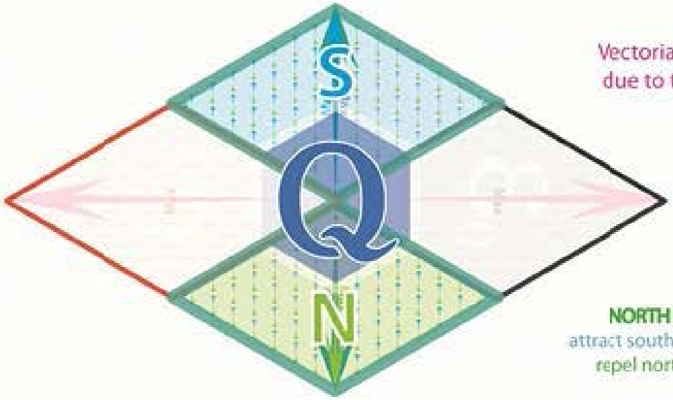

Vectorial momenta forces in EM fields are bi-directional
due to the energy-momenta quanta comprising them

*Opposites Attract
Similars Repel*

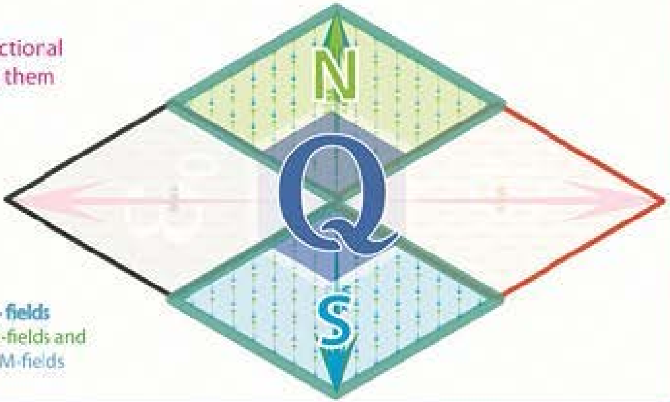

NORTH M- fields
attract south M-fields and
repel north M-fields

μ_0

SOUTH M- fields
attract north M-fields and
repel south M-fields

Magentic fields can only exist in conjuction with Electric fields & Electric fields can only exist in conjunction with Magnetic fields

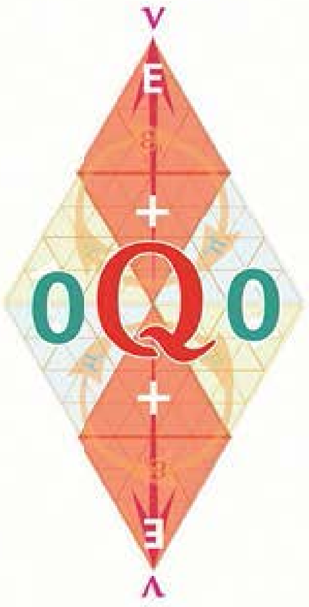

Electro-static fields

Similar charge electric dipole pairings
create 'neutralised' Magnetic dipoles

*There are no such things as purely
Electric or Magnetic fields*

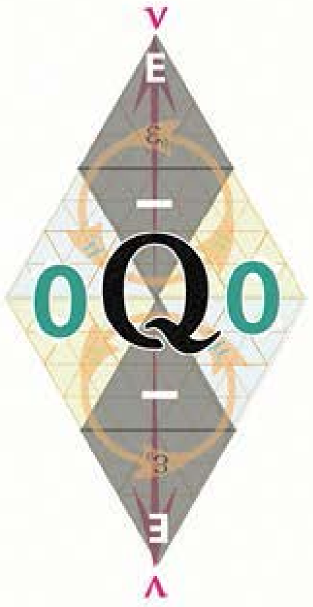

**Fields
of
Force**

v=0

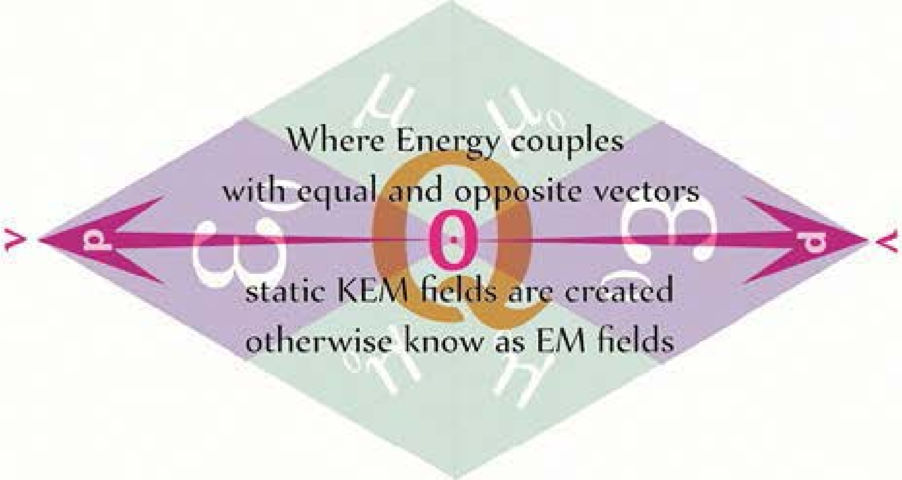

Where Energy couples
with equal and opposite vectors

static KEM fields are created
otherwise know as EM fields

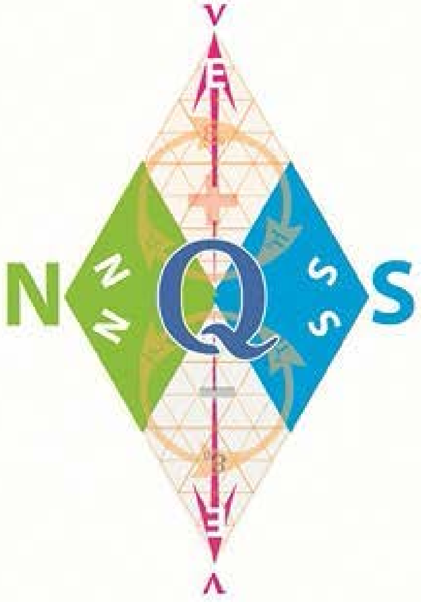

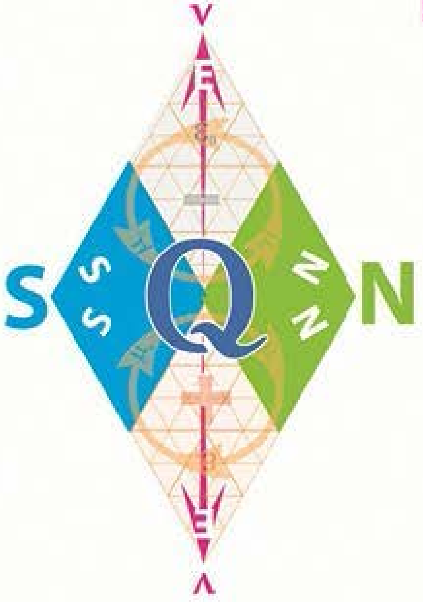

*All energy fields are Electro-Magnetic in nature
a direct product of their equilateral geometry*

'Neutralised' electric dipole pairings
create Magnetic dipole fields

Magneto-static fields

Copyright ABRAHAM [2008] - all rights reserved

Charged EM field geometries

Opposite charge fields
can produce neutral E-fields
(with magnetic moments)

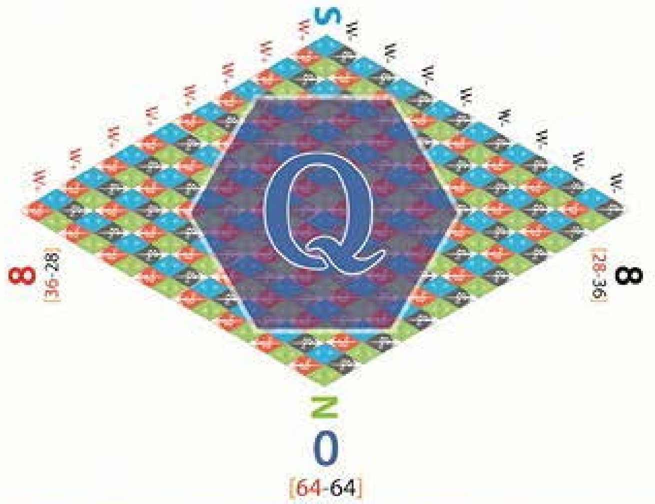

Electrostatic charged matter generate
charged energy fields around them

Moving charged particles generate
Kinetic energy & Magnetic moments

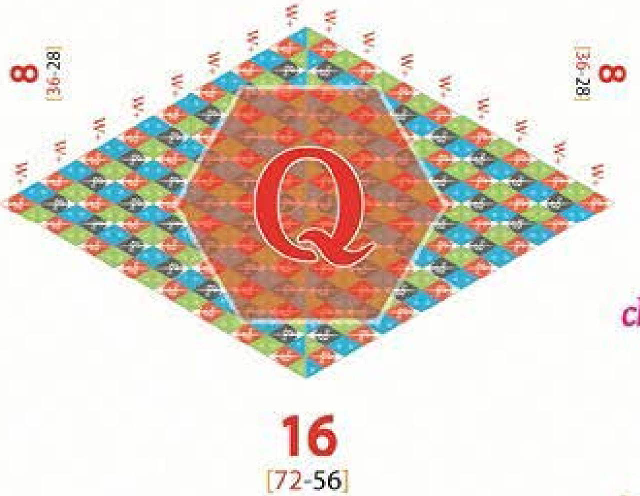

Charged electrostatic fields

accelerate charged particles
vectorally dependent on their quantum
charge mass-energy momenta field geometries

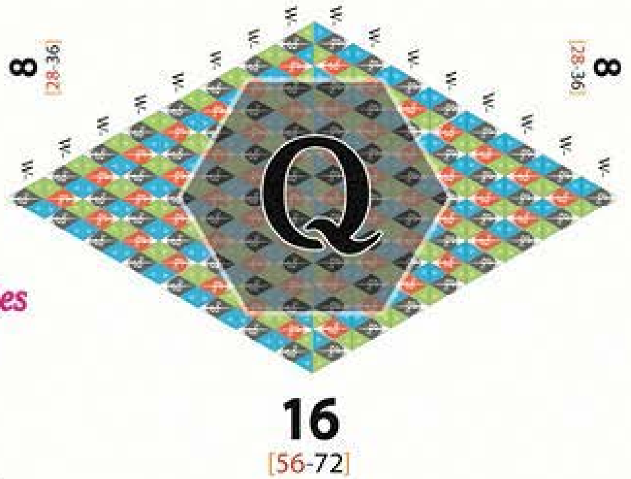

Positive charge electrostatic fields
attract Negative charges
repel Positive charges

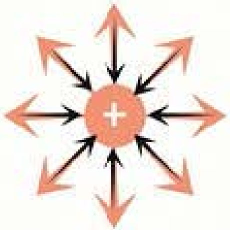

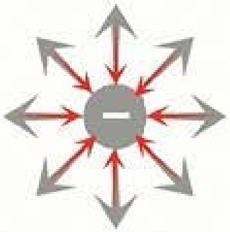

Negative charge electrostatic fields
attract Positive charges
repel Negative charges

E field acceleration of charged particles

Electric fields can accelerate charged particles within their field geometry
dependent on the particle's nett charged mass-Matter topology

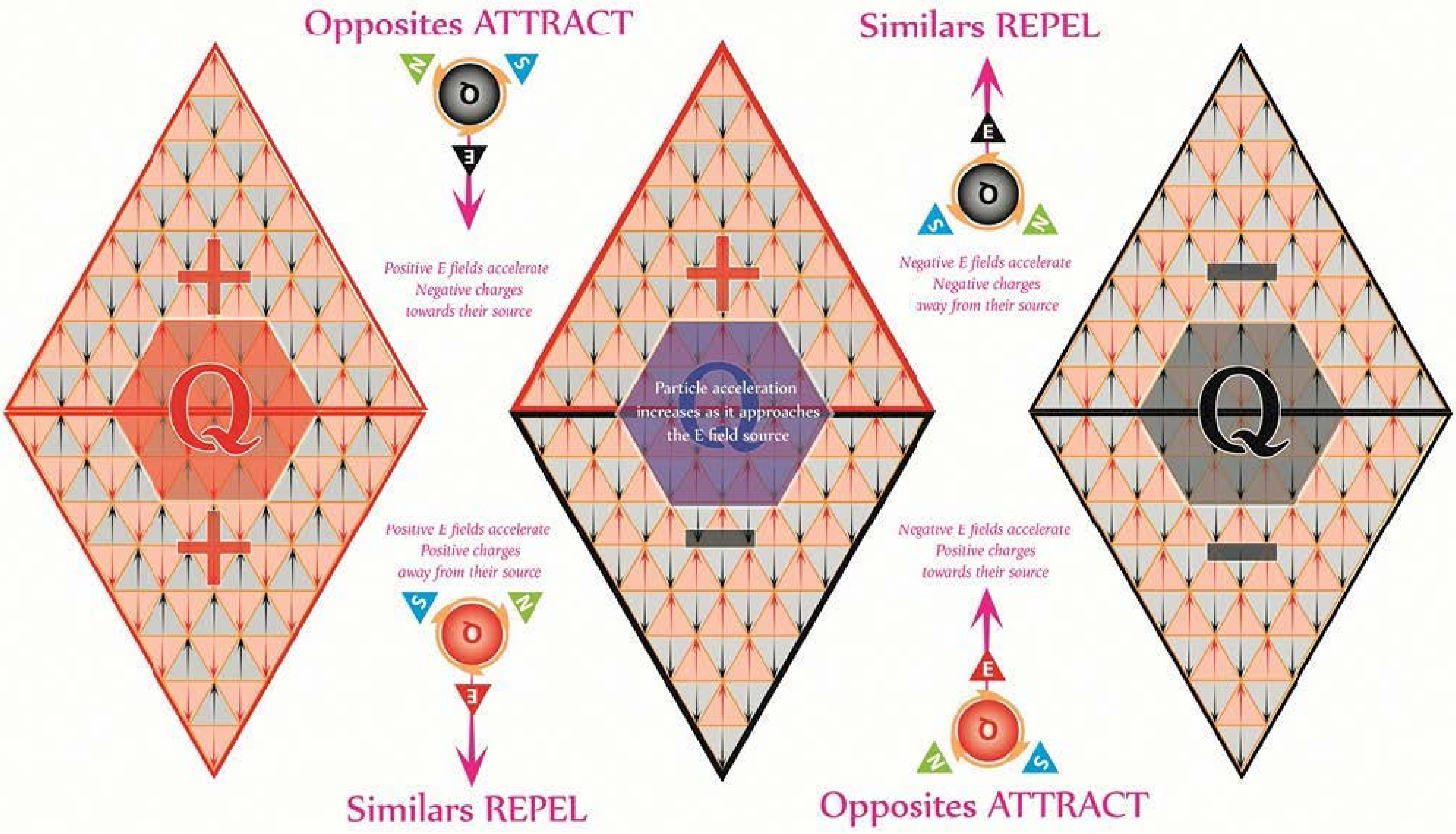

*Positive E fields accelerate
Negative charges
towards their source*

*Negative E fields accelerate
Negative charges
away from their source*

Particle acceleration
increases as it approaches
the E field source

*Positive E fields accelerate
Positive charges
away from their source*

*Negative E fields accelerate
Positive charges
towards their source*

The charge quanta within Neutral particles are affected equally by Electric fields

Tetryonics 04.10 - E field acceleration of charged particles

ZPFs can combine to form
4 distinct static EM fields

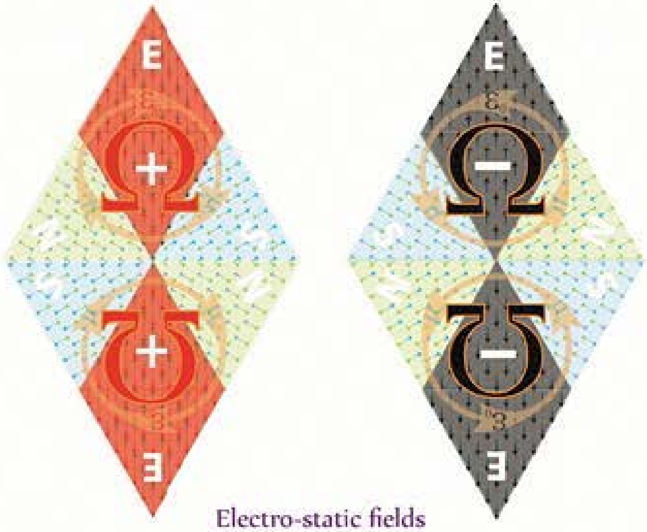

Electro-static fields

ElectroMagnetic fields

Magneto-static fields

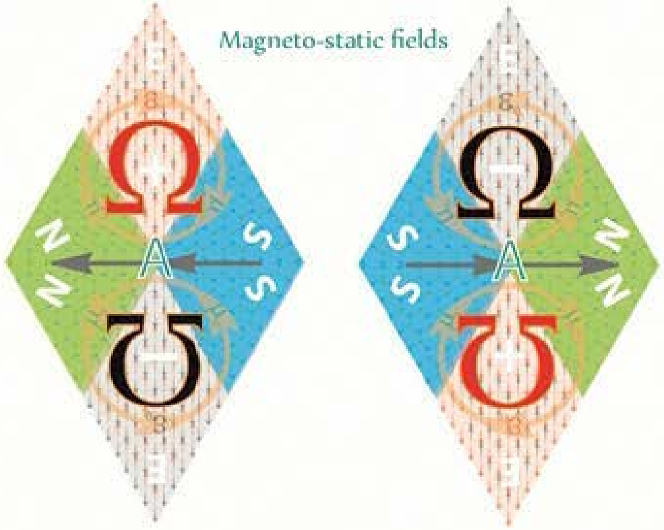

Similars REPEL
[QAM momenta vectors in opposition]

Opposites ATTRACT
[QAM momenta vectors in same direction]

'THE LAW OF INTERACTION'

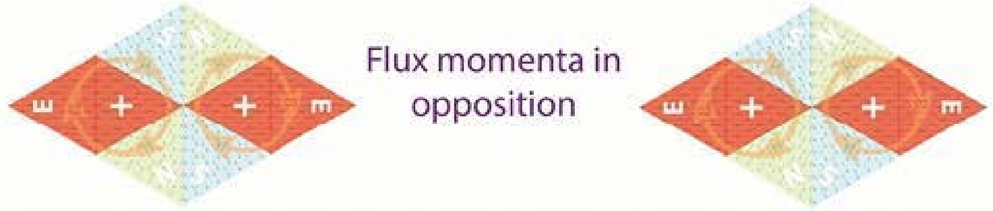

Flux momenta in
opposition

Electro-static fields
$1/r^2$

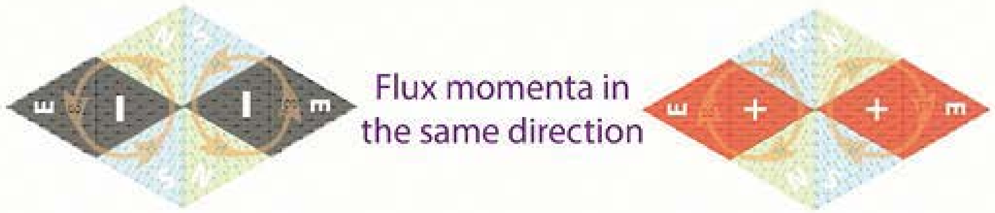

Flux momenta in
the same direction

'FORCES OF INTERACTION'

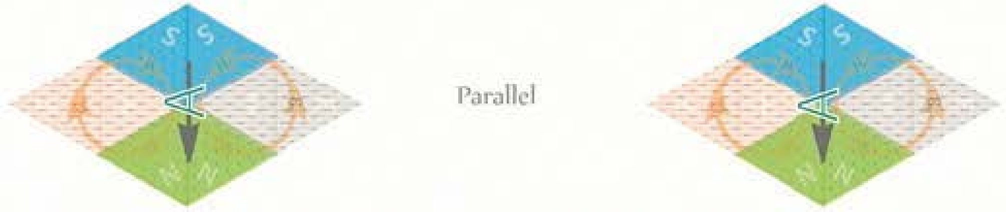

Parallel

Flux momenta in
the same direction

Magneto-static fields
$1/r^2$

Flux momenta in
opposition

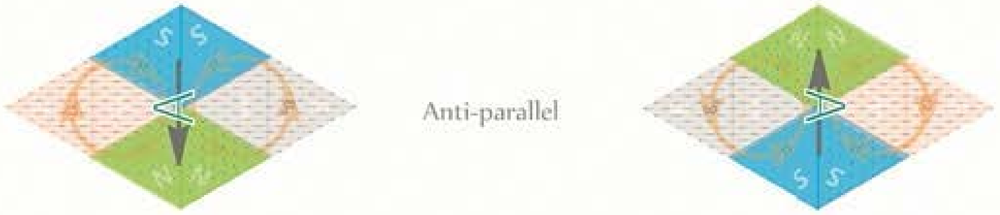

Anti-parallel

Positive electrostatic fields

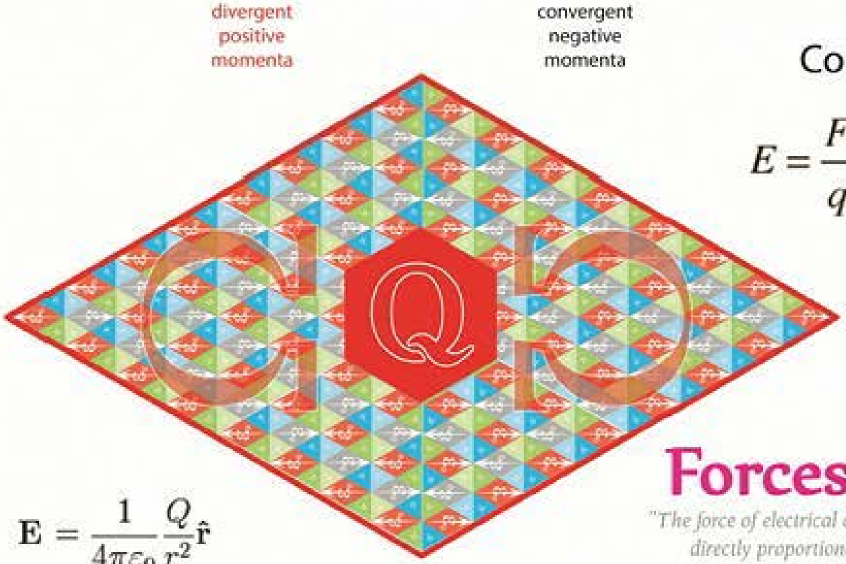
divergent positive momenta
convergent negative momenta
Q

Negative electrostatic fields

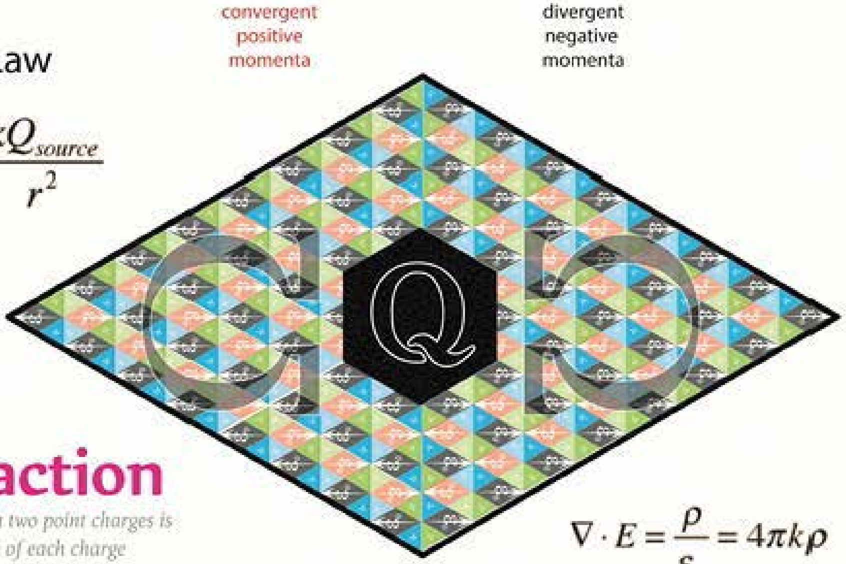
convergent positive momenta
divergent negative momenta
Q

Coulomb's Force Law

$$E = \frac{F}{q} = \frac{kQ_{source}q}{qr^2} = \frac{kQ_{source}}{r^2}$$

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

$$\nabla \cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k\rho$$

Forces of Interaction

"The force of electrical attraction or repulsion between two point charges is directly proportional to the product of magnitude of each charge and inversely proportional to the square of distance between them"

attract Negative charged masses
repel Positive charged masses

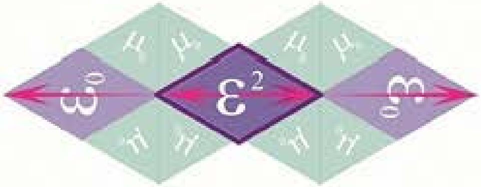

attract Positive charged masses
repel Negative charged masses

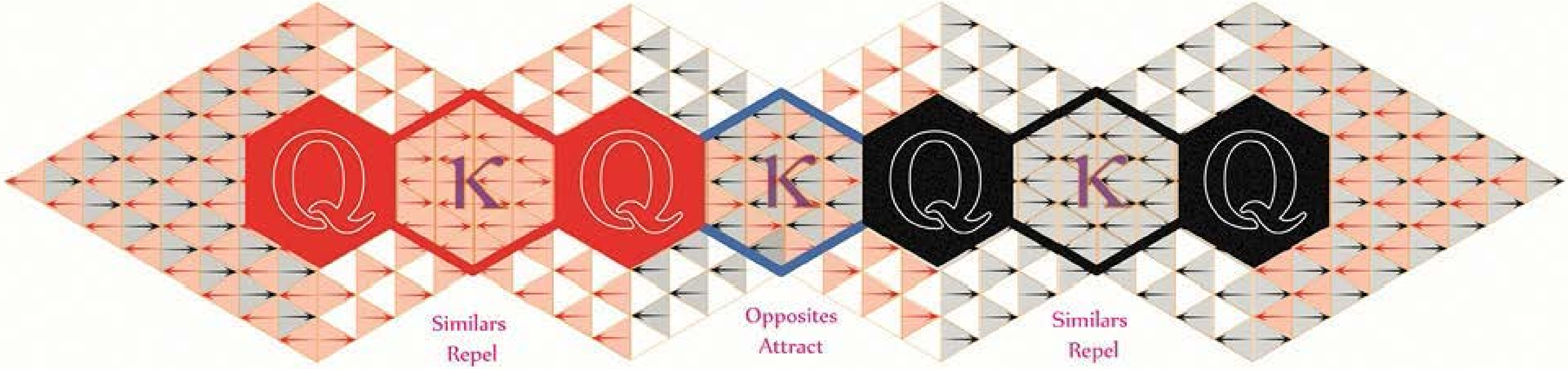
Q
K
Q
K
Q
K
Q
Similars Repel
Opposites Attract
Similars Repel

Copyright ABRAHAM [2008] - all rights reserved

Electro-Magnetic field Lines

The magnetic field at any given point is specified by both its direction and magnitude

Magnetic dipoles can be produced by the coupling of opposite electric charge fields

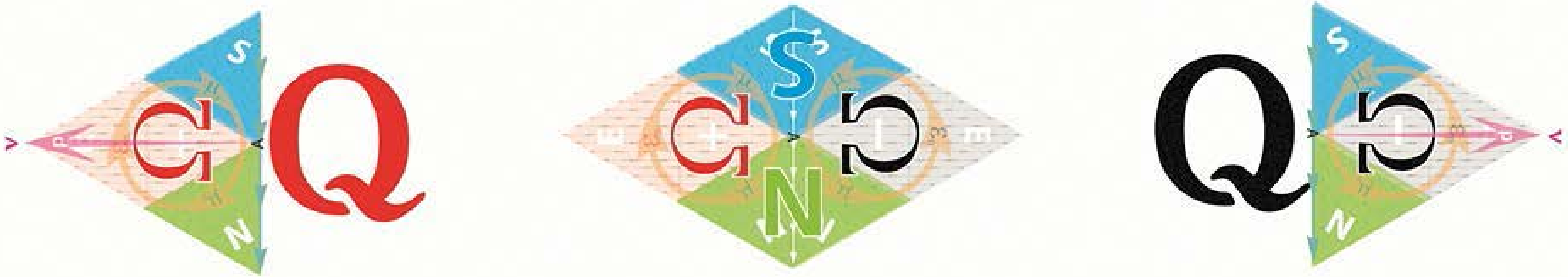

or

by accelerating charged bodies of Matter which in turn produce associated Kinetic energies
[Neutral Electric field and equal strength Magnetic moment]

Magnetic South Poles have divergent south vectors and convergent north vectors

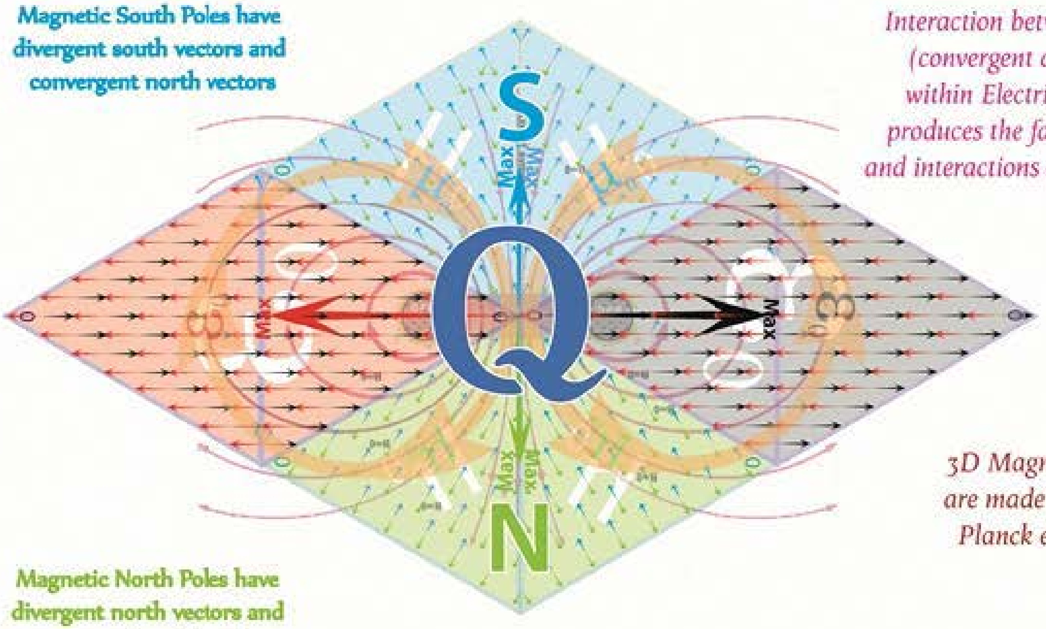

Magnetic North Poles have divergent north vectors and convergent south vectors

Interaction between vectorally opposed (convergent and divergent) vectors within Electric and Magnetic fields produces the familar lines of force and and interactions of Electro-Magnetic fields

3D Magnetic lines of Force are made up of 2D fields of Planck energy momenta

Every particle of Matter in motion possesses intrinsic Planck M-field dipoles becasue of their nett charged quanta which combine to produce a nuclear magneton.

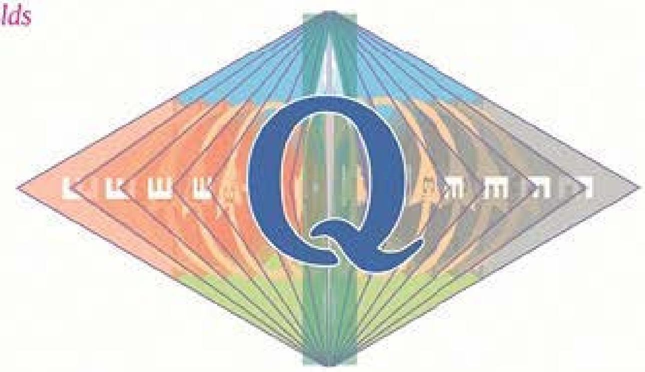

The magnetic field lines of permanent bar magnets are the result of the equilateral charge [QAM] geometry of neutralised quanta that form their electrostatic fields

Copyright ABRAHAM [2008] - all rights reserved

Magnetic field Forces

attractive magnetic forces

$$\frac{m^2 \cdot kg}{s^2 \cdot A^2} = \frac{J}{A^2} = \frac{Wb}{A} = \frac{s^2}{F} = \frac{V \cdot s}{A} = \frac{J/C \cdot s}{C/s} = \frac{J \cdot s^2}{C^2} = \frac{m^2 \cdot kg}{C^2}$$

In physics superpositioned M fields with interactive energy momenta produce Lorentz forces

repulsive magnetic forces

Magnetic moments of same charges moving in the same direction

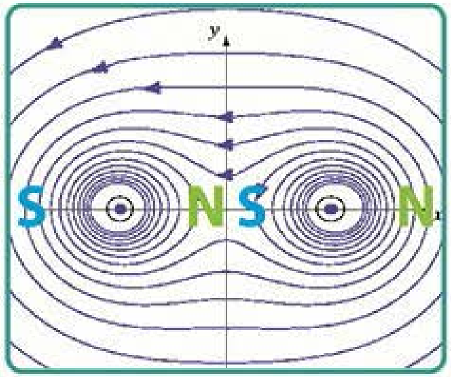

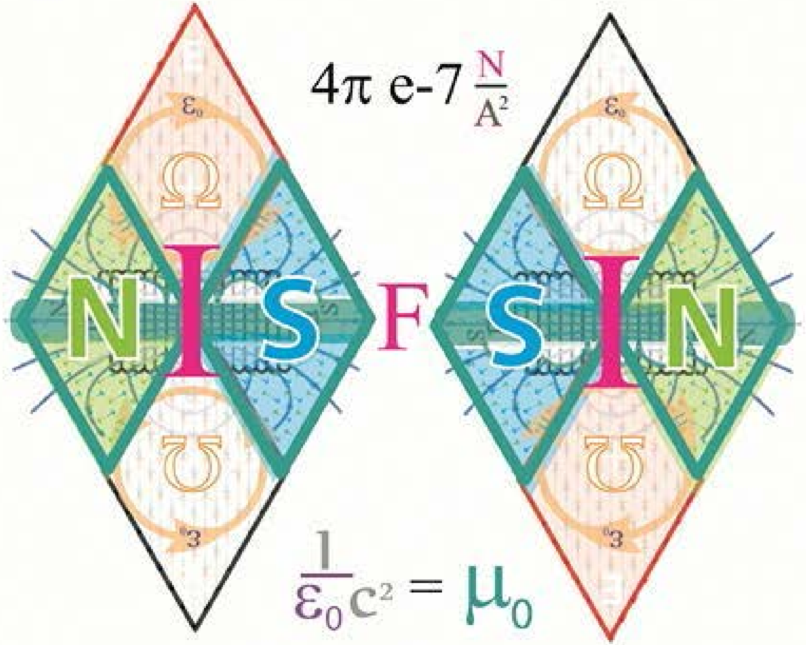
ε₀
Ω
N
S
F
S
N
Ω
ε₀

4π e-7 $\frac{N}{A^2}$

$$\frac{1}{\varepsilon_0 c^2} = \mu_0$$

Magnetic moments of same charges moving in opposite directions

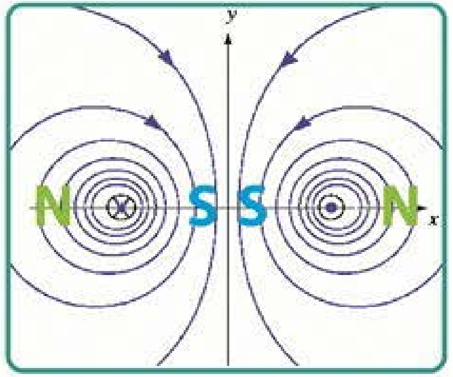

magnetostatic fields

ε₀
magnetic dipole
Q
μ₀
μ₀
ε₀

$$H = \frac{B}{\mu_0}$$

Coupled opposite charge EM fields produce Magnetic dipole fields

Joseph Henry

(17 December 1797 – 13 May 1878)

$$\mu_B$$

Electrostatic charged particles in motion create magnetic moments

KE
magnetic moments
Q
KE
μ₀
μ₀

kinetic EM fields

Magnetic field lines

Magnetostatic Dipole

Magnetic field lines were introduced by Michael Faraday (1791-1867) who named them "lines of force"

Michael Faraday

(22 September 1791 – 25 August 1867)

Magnetic South Poles have divergent south vectors and convergent north vectors

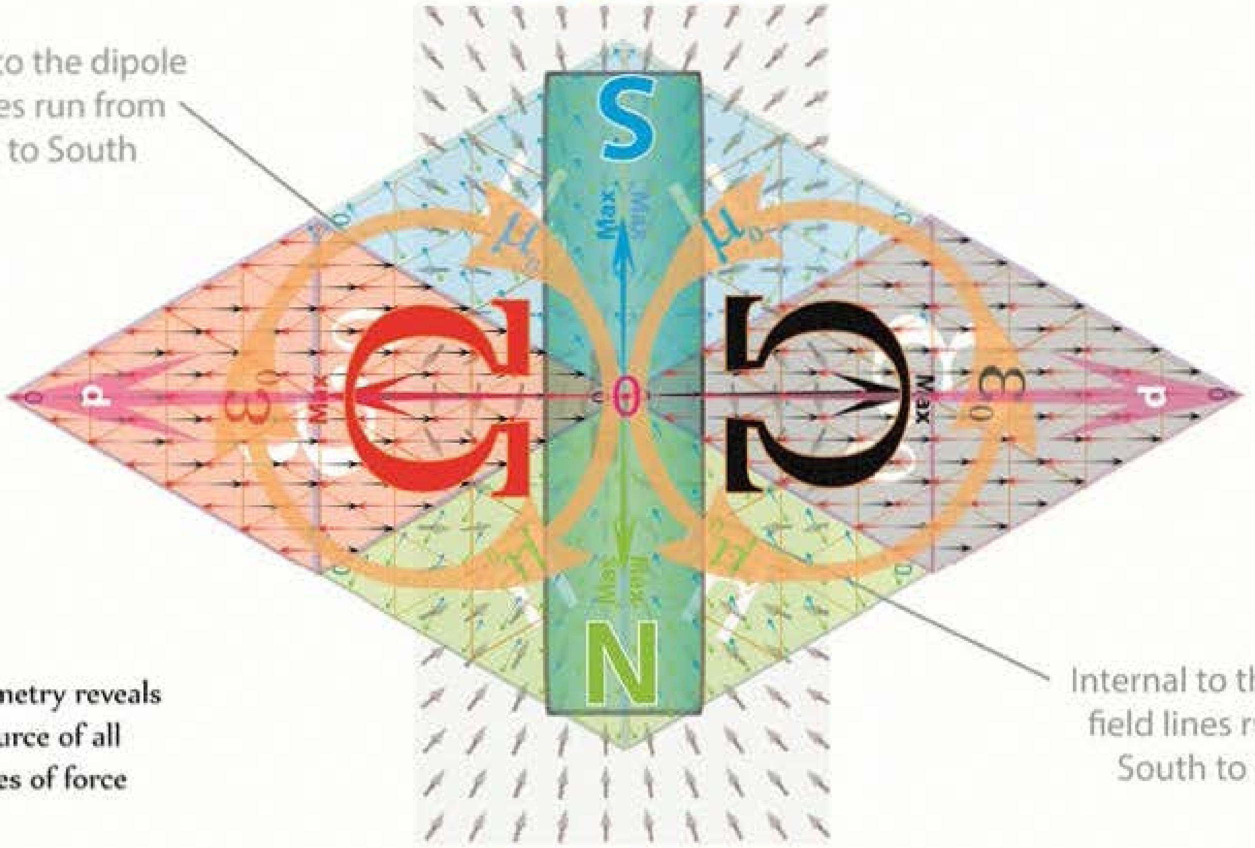

External to the dipole field lines run from North to South

Positive Electric fields have divergent positive E vectors and convergent negative E vectors

Negative Electric fields have divergent negative E vectors and convergent positive E vectors

Internal to the dipole field lines run from South to North

Tetryonic geometry reveals the true source of all EM field lines of force

Magnetic North Poles have divergent north vectors and convergent south vectors

Permanent Magnets can be viewed as neutral charge Electrostatic EM fields

Similar to conventional vs electron current flows Magnetic field vectors can be modelled with two directions of 'magnetic flux'

Parallel Magnetic Dipoles

Michael Faraday

(22 September 1791 – 25 August 1867)

James Clerk Maxwell

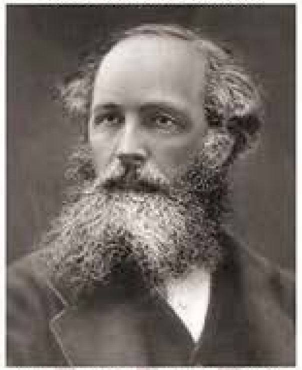

(13 June 1831 – 5 November 1879)

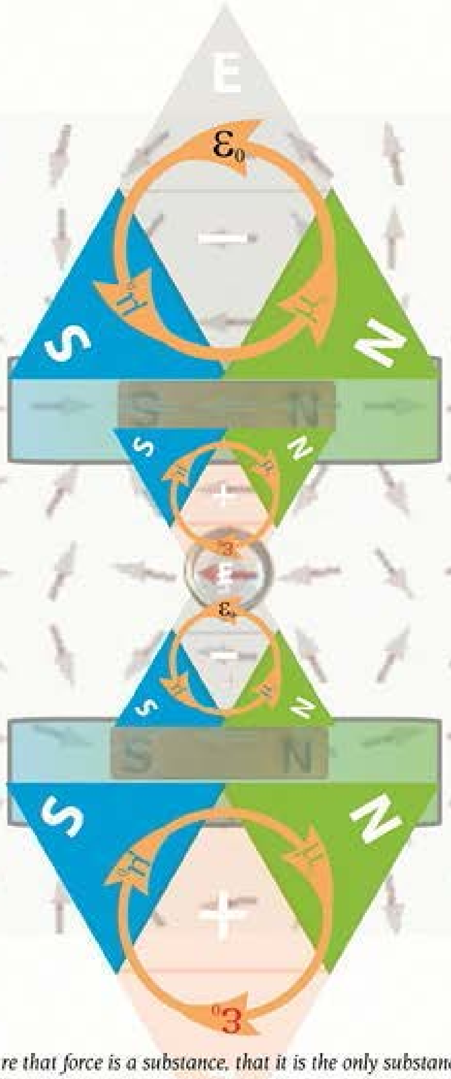

Magnetic lines of force are continuous and will always form closed loops.

Magnetic lines of force will never cross one another.

Parallel magnetic lines of force traveling in the same direction repel one another.

Parallel magnetic lines of force traveling in opposite directions tend to unite with each other and form into single lines traveling in a direction determined by the magnetic poles creating the lines of force.

Magnetic lines of force tend to shorten themselves.

Therefore, the magnetic lines of force existing between two unlike poles cause the poles to be pulled together.

Magnetic lines of force pass through all materials, both magnetic and nonmagnetic.

Magnetic lines of force always enter or leave a magnetic material at right angles to the surface

Maxwell had studied and commented on the field of electricity and magnetism as early as 1855/6 when "On Faraday's lines of force" was read to the Cambridge Philosophical Society.

The paper presented a simplified model of Faraday's work, and how the two phenomena were related. He reduced all of the current knowledge into a linked set of differential equations with 20 equations in 20 variables. [Quarterions]

This work was later published as "On physical lines of force" in March 1861.

In his 1864 paper "A dynamical theory of the electromagnetic field", Maxwell wrote, "The agreement of the results seems to show that light and magnetism are affections of the same substance, and that light is an electromagnetic disturbance propagated through the field according to electromagnetic laws

Maxwell showed that the equations predict the existence of waves of oscillating electric and magnetic fields that travel through empty space at a speed of 310,740,000m/s.

His famous equations, in their modern form of four partial differential equations, first appeared in fully developed form in his textbook A Treatise on Electricity and Magnetism in 1873.

The specific features of Faraday's field concept, in its 'favourite' and most complete form, are that force is a substance, that it is the only substance and that all forces are interconvertible through various motions of the lines of force. These features of Faraday's 'favourite notion' were not carried on by Maxwell in his approach to the problem of finding a mathematical representation for the continuous transmission of electric and magnetic forces. Maxwell considered these electric and magnetic forces to be states of stress and strain in a mechanical aether, a notion further advanced by relativity theory with its 'stress energy' tensor math.

Tetryonics reveals lines of Force to be a direct result of the various superpositioned EM field geometries of equilateral mass-energy momenta

Anti-Parallel Magnetic Dipoles

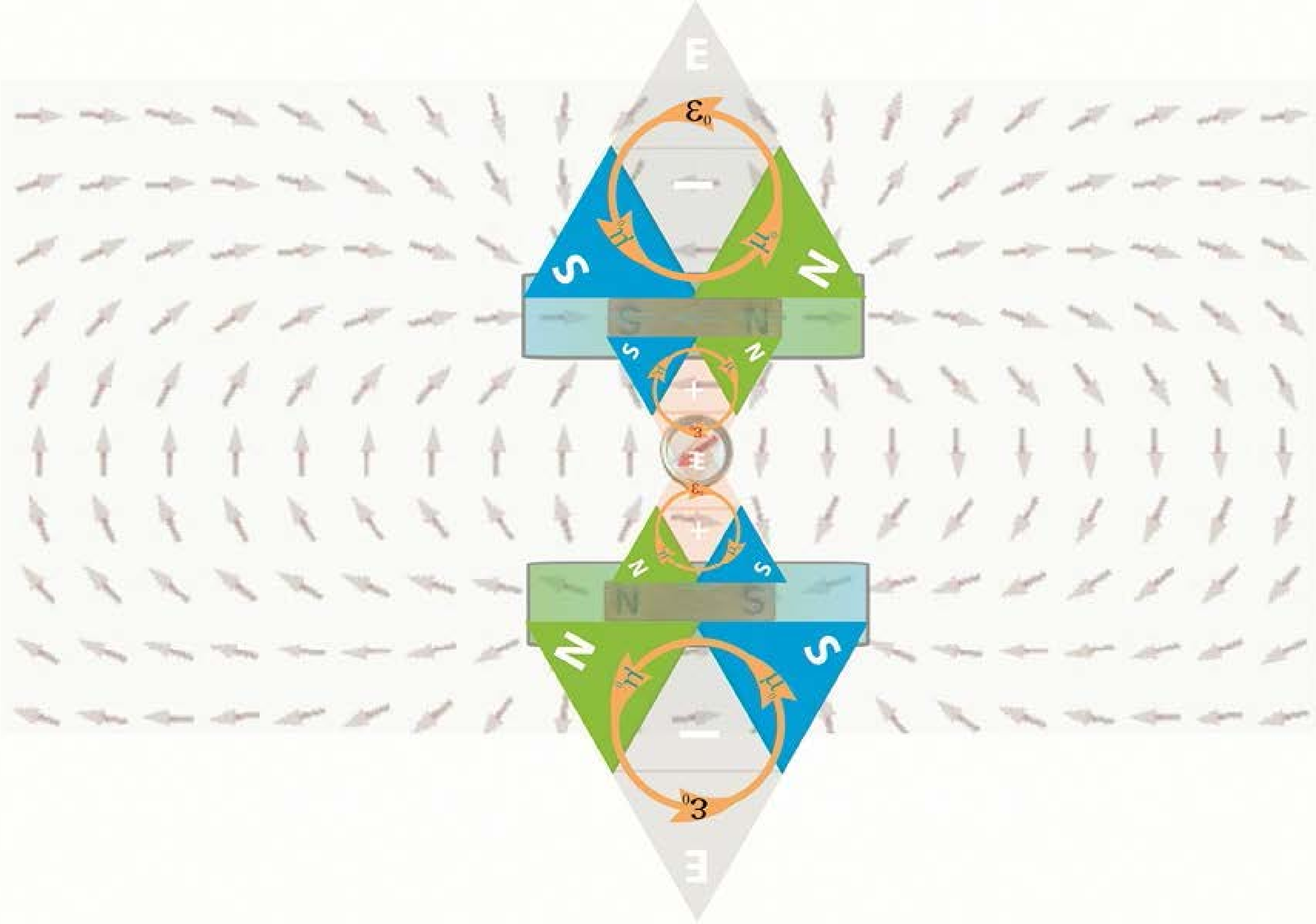
E
ε_0
μ_0
S
N
−
S
N
+
+
N
S
N
S
N
S
−
ε_0
E

Tetryonics 04.17 - Anti-Parallel Magnetic dipoles

Copyright ABRAHAM [2008] - all rights reserved

Magnetic Moments

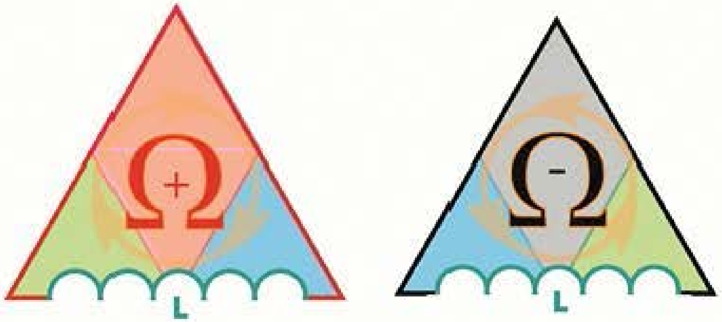

Single ZPFs are
'ideal quantum inductor elements'

1

Electro-static Energies
Charges

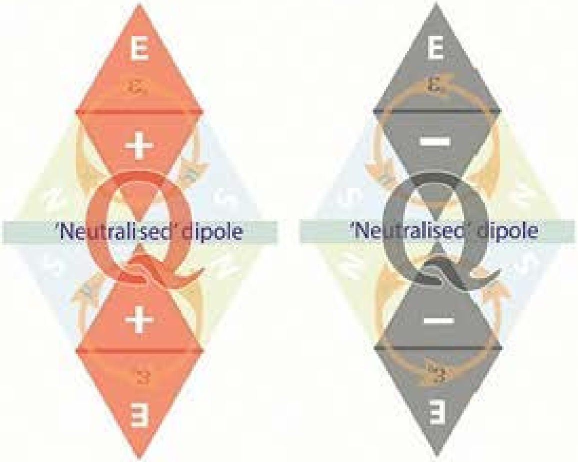
E
ε
+
'Neutralised' dipole
+
E
E
ε
−
'Neutralised' dipole
−
E
N
S

All Matter in motion possesses kinetic energies
which are stored as Planck quanta in their KEM fields

The charge geometry of KEM fields are reflective
of the interactive component of the charged topology
of the particle in motion

Each charge geometry
has distinct Magnetic dipole
alignments

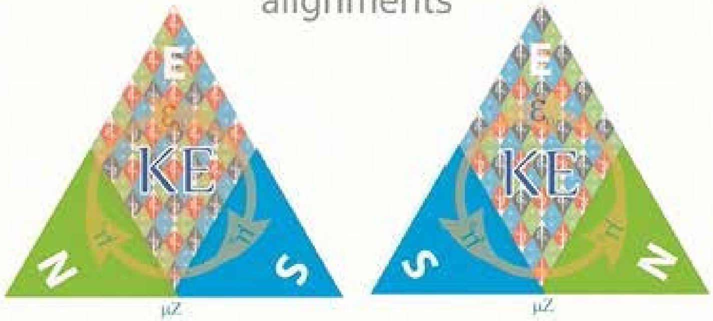
E
ε
KE
N
S
μZ
E
ε
KE
S
N
μZ

which in turn can only create
2 distinct orientations [spins] of
magnetic moments

2

Magneto-static Energies
Kinetic Energies

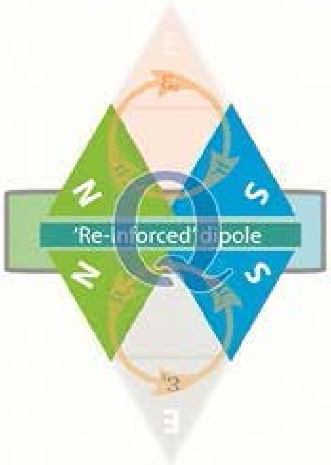
E
ε
N
S
'Re-inforced' dipole
N
S
E

ZPF sets can form inductively coupled
quantum Harmonic Oscillators

Copyright ABRAHAM [2008] - all rights reserved

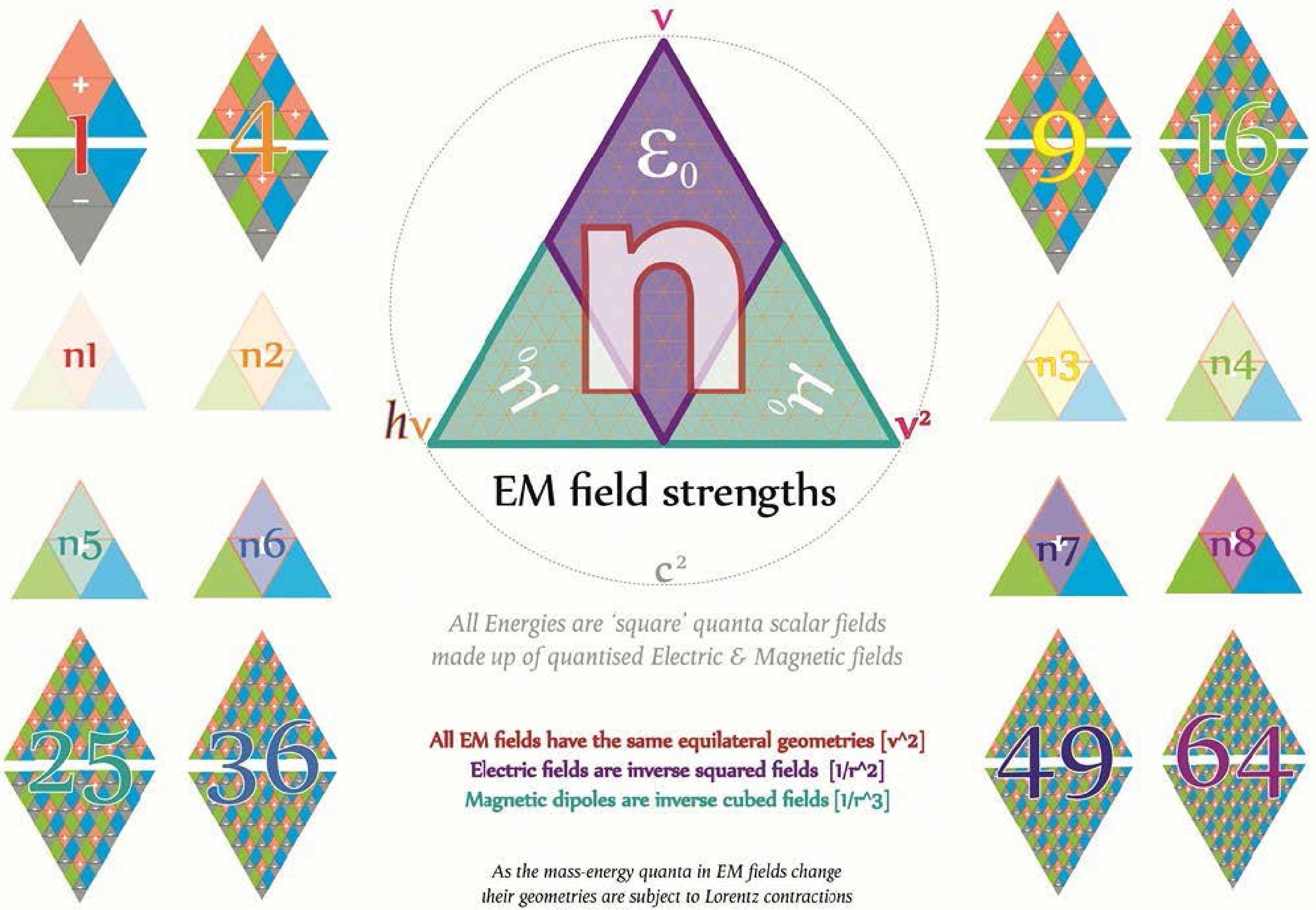
1
4
9
16
n1
n2
n3
n4
n5
n6
n7
n8
25
36
49
64
v
ε0
n
μ0
μ0
hv
v²
c²

EM field strengths

All Energies are 'square' quanta scalar fields
made up of quantised Electric & Magnetic fields

All EM fields have the same equilateral geometries [v^2]
Electric fields are inverse squared fields [1/r^2]
Magnetic dipoles are inverse cubed fields [1/r^3]

As the mass-energy quanta in EM fields change
their geometries are subject to Lorentz contractions

Electrostatic particle modeling

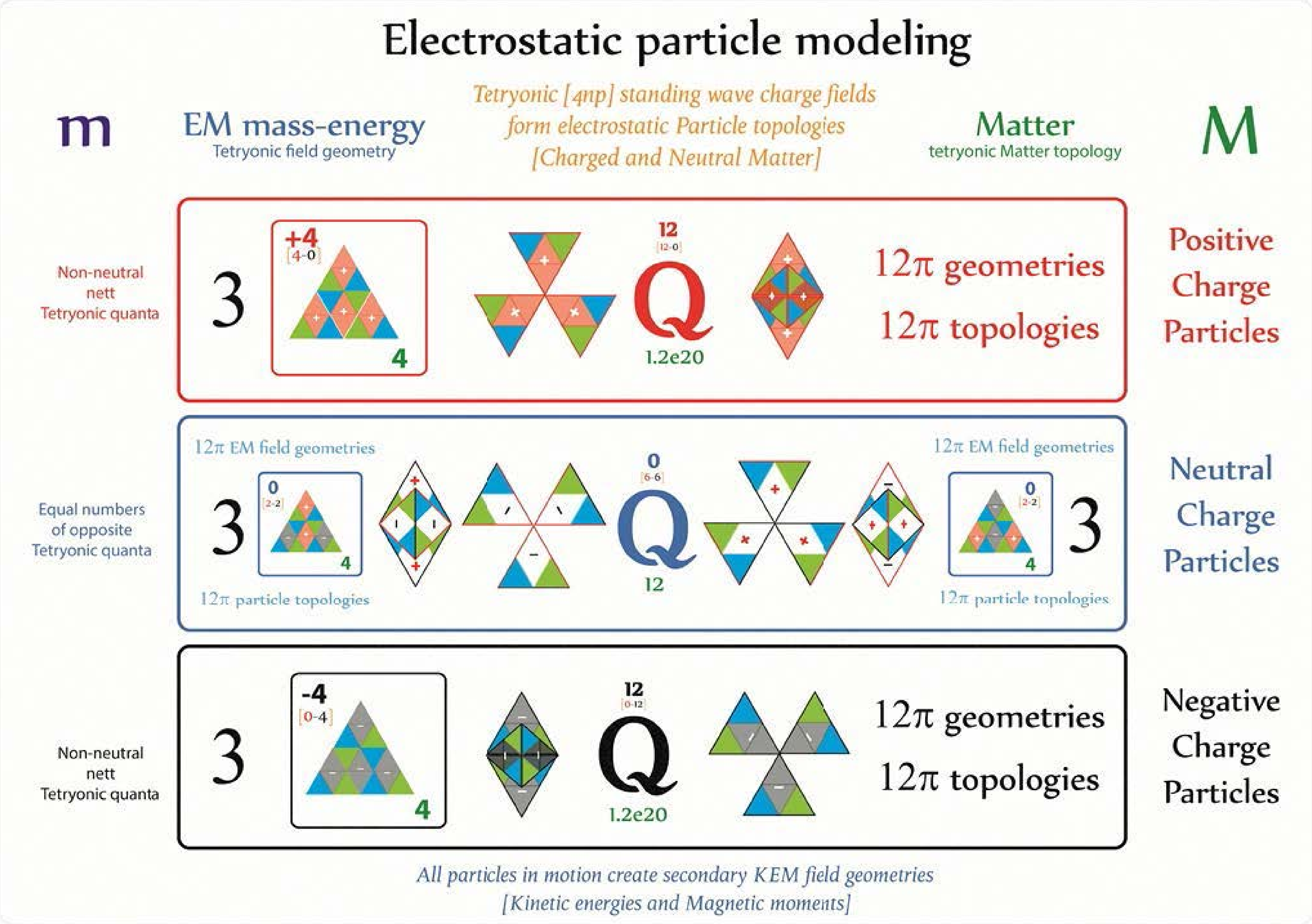

m EM mass-energy
Tetryonic field geometry

Tetryonic [4πp] standing wave charge fields
form electrostatic Particle topologies
[Charged and Neutral Matter]

Matter **M**
tetryonic Matter topology

Non-neutral
nett
Tetryonic quanta

12π geometries
12π topologies

Positive
Charge
Particles

Equal numbers
of opposite
Tetryonic quanta

12π EM field geometries

12π EM field geometries

12π particle topologies

12π particle topologies

Neutral
Charge
Particles

Non-neutral
nett
Tetryonic quanta

12π geometries
12π topologies

Negative
Charge
Particles

All particles in motion create secondary KEM field geometries
[Kinetic energies and Magnetic moments]

Copyright ABRAHAM [2008] - all rights reserved

Electro-static Fields

Momenta acting in oppose directions results in zero velocity

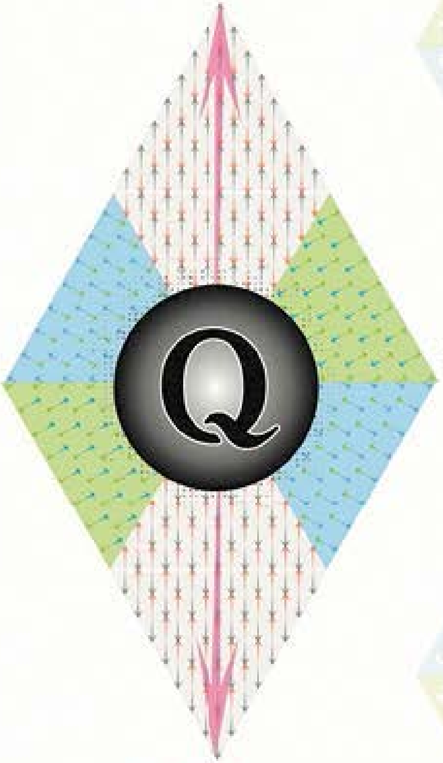

Zero Velocity equates to Zero nett Momentum

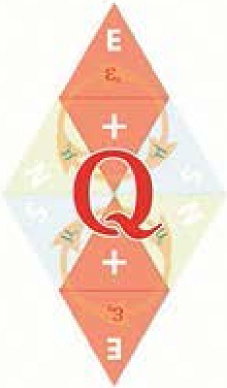

Stationary Charges have neutralised magnetic dipoles

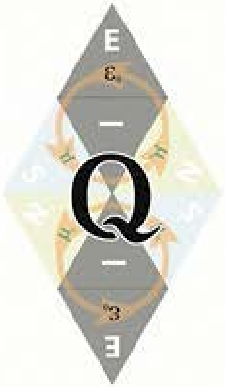

The Kinetic Electric & Magnetic energies are contained in an KEM field extending from a charged particle in motion

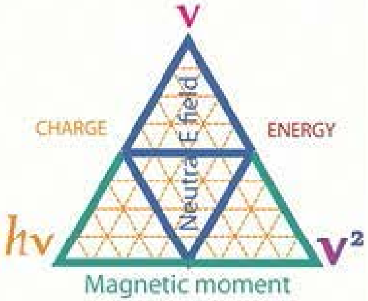
v
CHARGE
ENERGY
Neutral E field
hv
v²
Magnetic moment

Charged particles in motion produce a Magnetic moment

Kinetic EM Fields

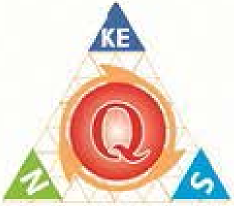

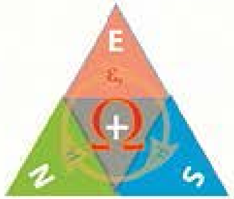

Moving Charges have neutral KE fields and Magnetic moments

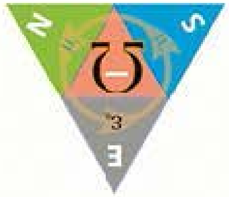

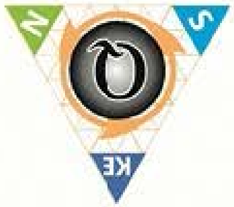

Motion in any direction produces Kinetic & Magnetic energies

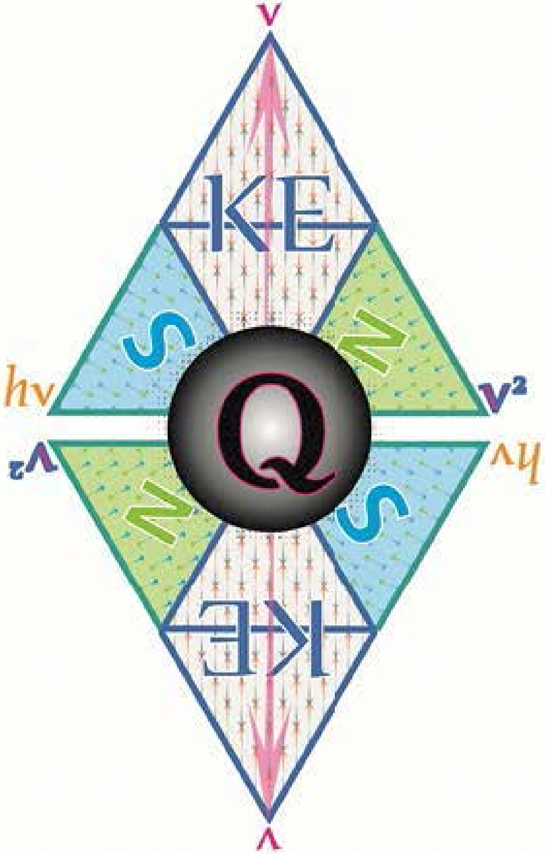
v
KE
S
N
hv
v²
Q
N
S
KE
v

The Strength of the Kinetic Electric field & Magnetic moment is directly proportional to the square of the particle's Velocity

Electric & Kinetic Fields

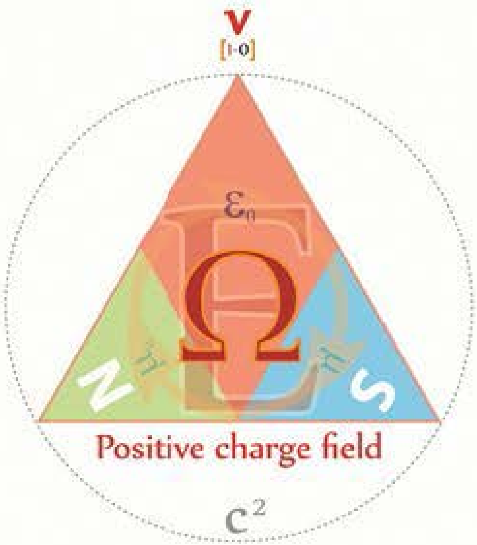
v
[1-0]
ε₀
Ω
N
S
c²

Positive charge field

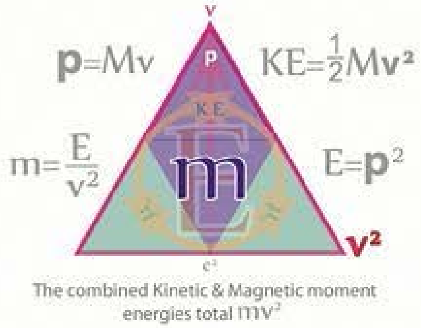

$\mathbf{p}=Mv$ $\quad KE=\frac{1}{2}Mv^2$

$m=\frac{E}{v^2}$ $\quad E=\mathbf{p}^2$

The combined Kinetic & Magnetic moment energies total mv^2

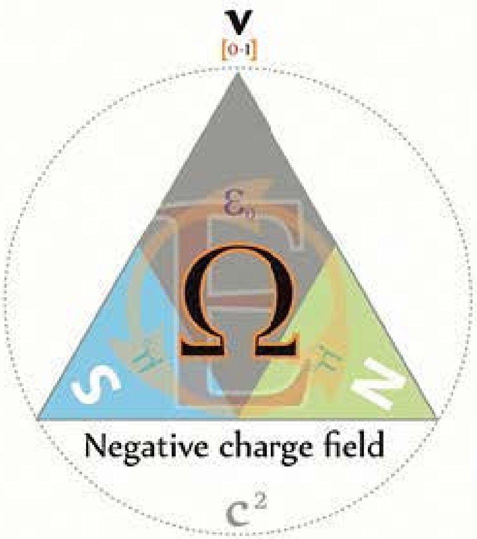
v
[0-1]
ε₀
Ω
S
N
c²

Negative charge field

Kinetic Energy field

KEM field created by a Positive charged body

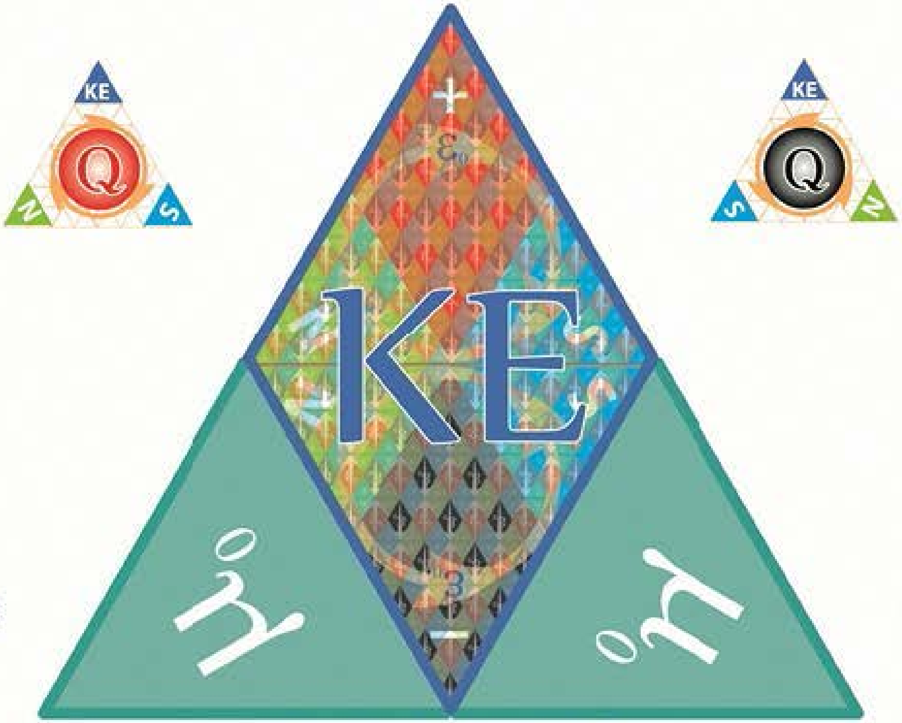
KE
Q
N
S

+
ε₀
KE
−

KE
Q
S
N

KEM field created by a Negative charged body

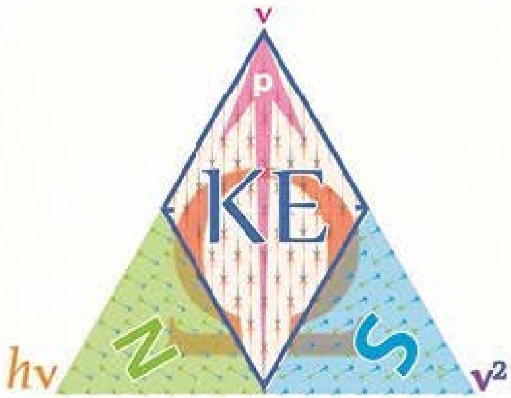
v
p
KE
N
S
hv
v²

Positive charges repelled
Negative charges attracted

Neutralised Electric fields
Enhanced Magnetic moments

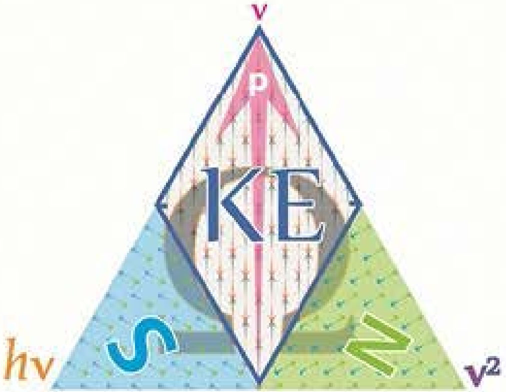
v
p
KE
S
N
hv
v²

Negative charges repelled
Positive charges attracted

EM & KEM force vectors

All mass-energy quanta are ideal quantum Inductors

The E&M force vectors create orthogonal equilateral EM Fields

Electric fields propogate orthogonally to the Magnetic dipole field

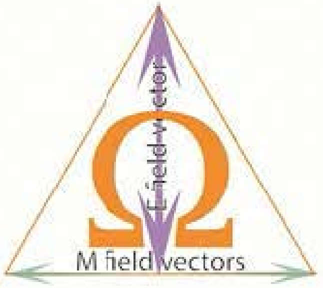

Magnetic flux external to the Magnetic dipole flows from North to South

&

Internal to a Magnetic dipole it flows South to North

When ZFFs combine to form a Magnetostatic dipole they form orthogonal magnetic vectors

Positive charge can be modelled electrically with clockwise EM fluxes [omegas]

Negative charge can be modelled electrically with anti-clockwise EM fluxes [omegas]

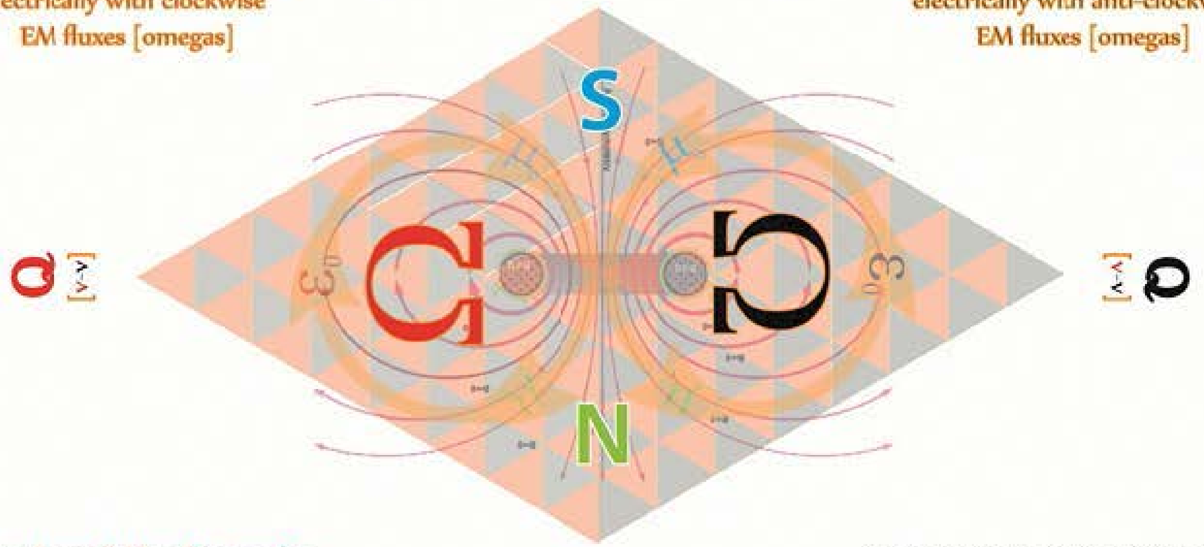

Positive charge fields of interaction result from an excess of divergent positive energy momenta

Equal energy-momenta in opposition create static fields

Negative charge fields of interaction result from an excess of divergent negative energy momenta

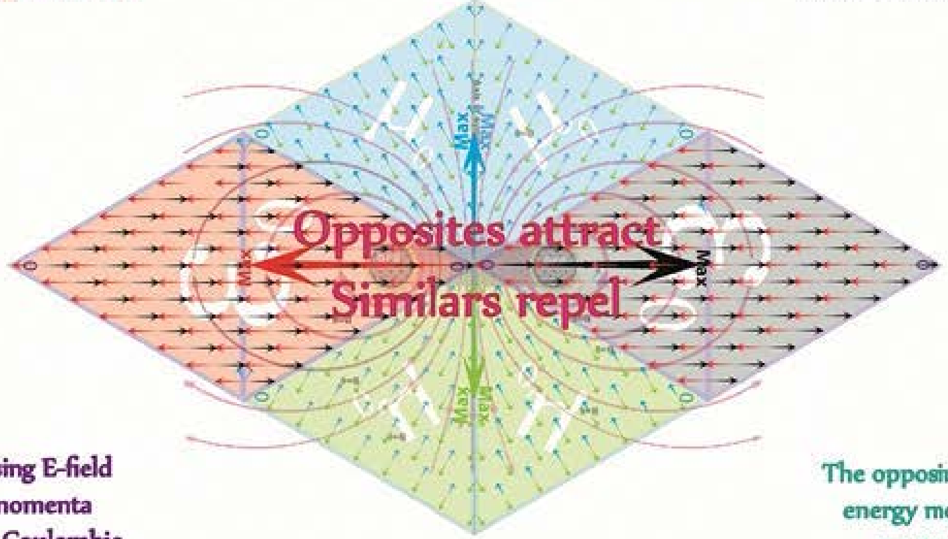

The opposing E-field energy momenta creates the Coulombic Law of Interaction

When E field is at Maximum - B field is at Minimum
when B field is at Maximum - E field is at Minimum

The opposing M-field energy momenta creates the Lorentz Force

EM Forces and ZPFs

ZPF quanta can combine in differing combinations
to produce 3 distinct charged sets

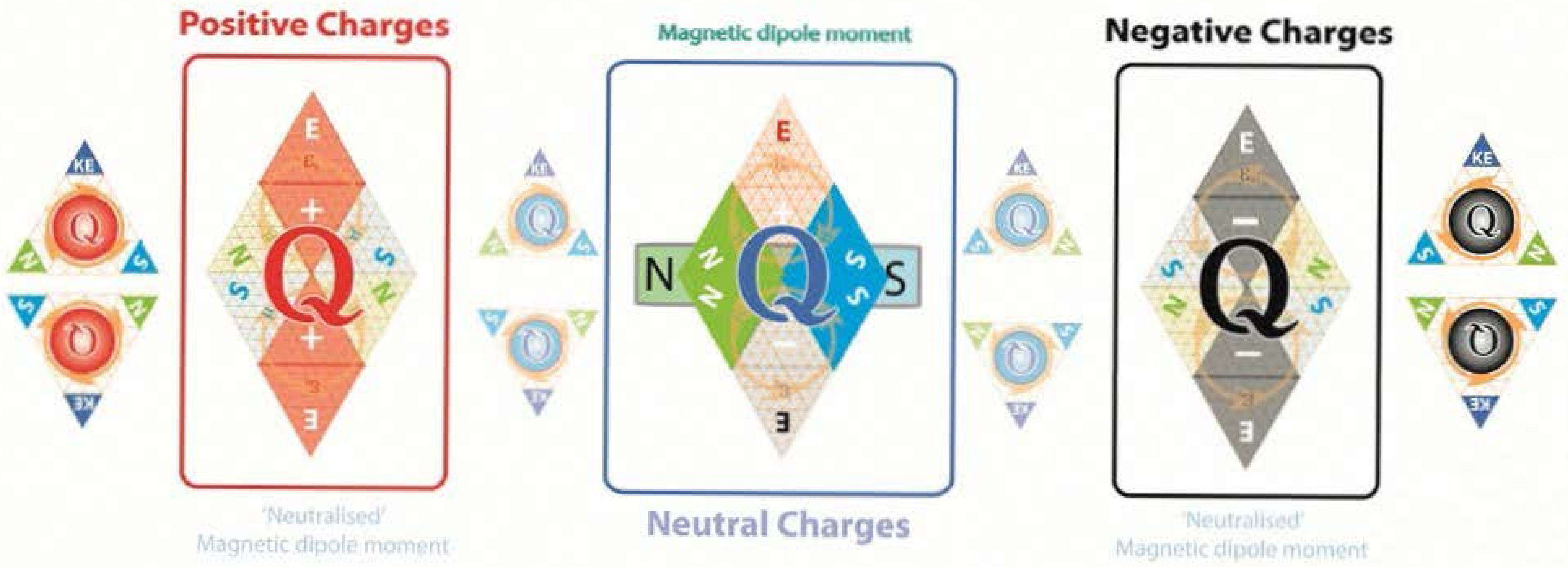

Positive Charges

'Neutralised'
Magnetic dipole moment

Magnetic dipole moment

Neutral Charges

Negative Charges

'Neutralised'
Magnetic dipole moment

All EM mass-Energy-Matter & forces can be modelled using Tetryonic geometries

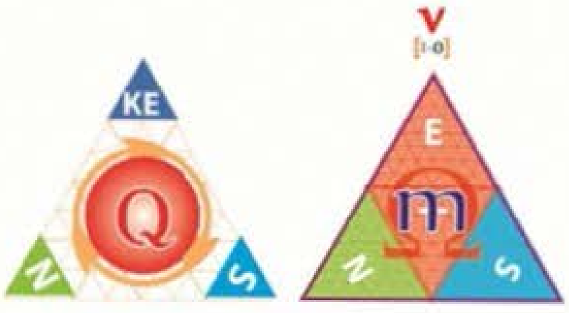

Positive charge particle KE field and magnetic moment

Zero Point Field EM geometry is the foundation for all the EM forces comprising, and acting between particles of Matter

Lorentz force, Lenz's Law, Right/Left Hand rules, etc can all be easily replaced with this simple geometric model

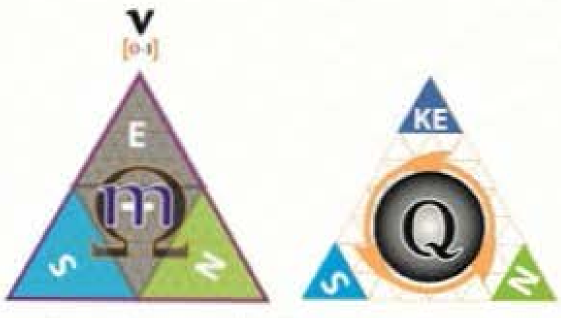

Negative charge particle KE field and magnetic moment

Kinetic EM fields

All Kinetic EM fields resulting from motion have charge field geometries resulting from the charged Matter geometry of the particle in motion

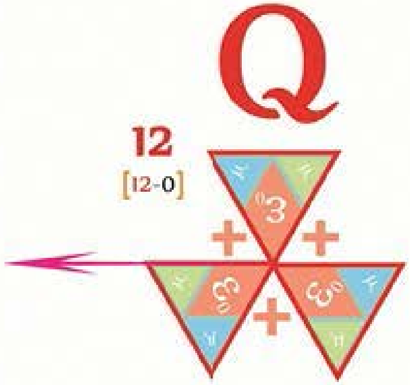

Positive charge particle topologies produce positive KEM fields

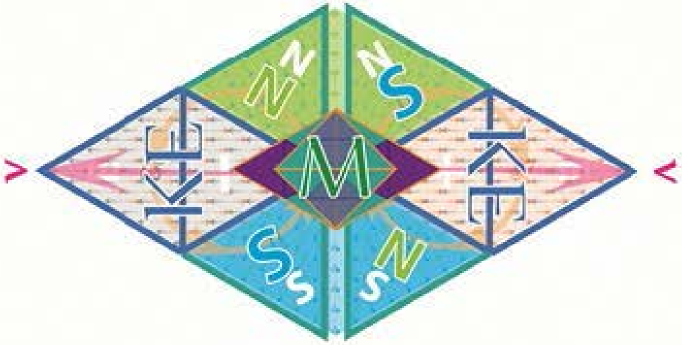

KEM field charge geometries do not contribute to the nett charge

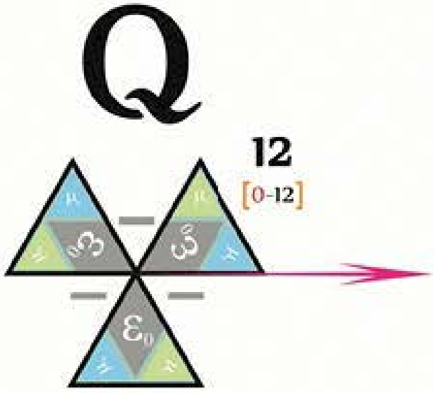

Negative charge particle topologies produce negative KEM fields

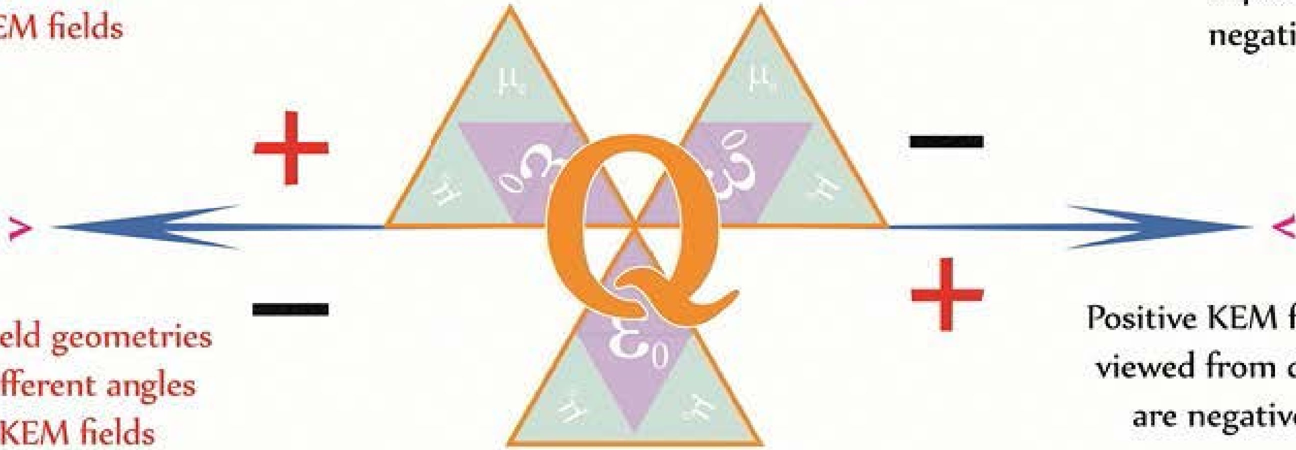

Negative KEM field geometries viewed from different angles are positive KEM fields

Positive KEM field geometries viewed from different angles are negative KEM fields

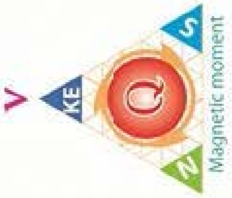

As a direct result of the KEM field being a EM field permeating free space the symmetry of EM fields results in KEM field geometries being viewed as having neutral KE fields with a magnetic moment

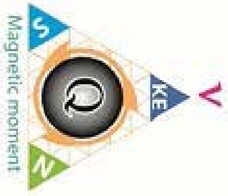

Point Particles and KEM fields

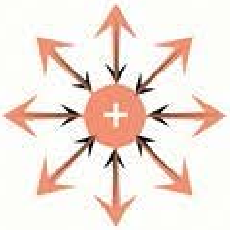

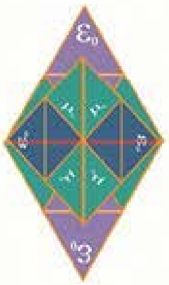

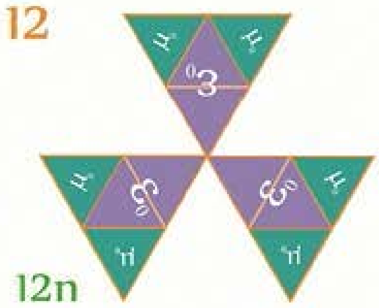

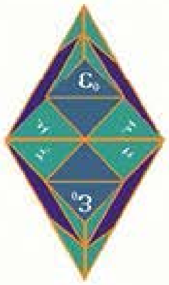

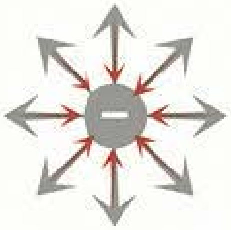

Charged rest mass-Matter topology

Quarks T[8π]
Leptons T[12π]

velocity invariant rest mass-Matter has a standing wave topology

Baryons T[20π]

Charged Leptons at rest are Electric field standing waves
(with neutral Magnetic poles)
KE from motion generates a Magnetic Moment

Bosons nπ [ODD]

Kinetic EM field geometry is divergent from a particle's rest Matter topology

Photons nπ [EVEN]

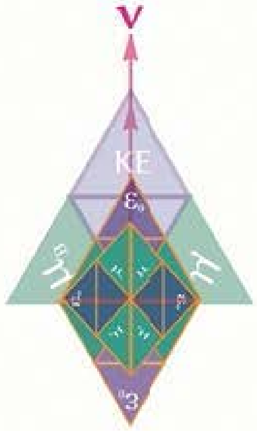

Charge

Q

[v-v]

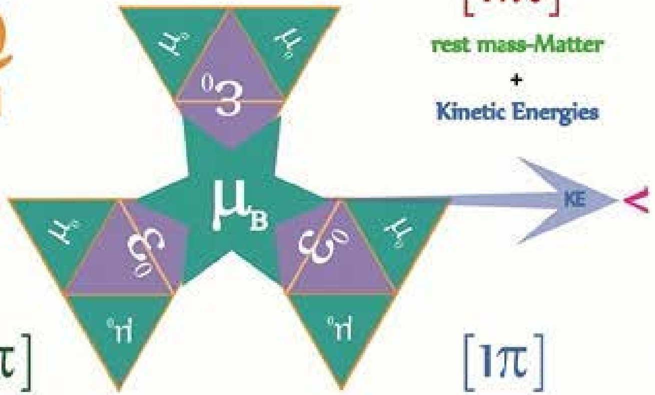

[nπ]

rest mass-Matter
+
Kinetic Energies

[12π]

rest Matter topology

[1π]

Kinetic Energy geometry

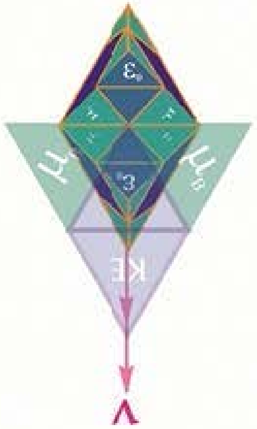

Copyright ABRAHAM [2008] - all rights reserved

KEM fields of Matter in motion

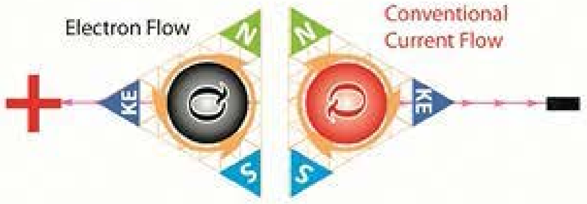
Electron Flow
Conventional Current Flow
+
KE
N
S
N
S
KE
−

ZPFs geometries can be used
to model the KEM fields
of charged particles

2D charged EM field geometries create Matter topologies

tetryons	[4π]	tetryons	[4π]
quarks	[12π]	quarks	[8π]
leptons	[12π]	leptons	[12π]
Baryons	[36π]	Baryons	[20π]
Elements	[84π]	Elements	[54π]

The mass-energy content of all charged fascia constituting
massive particles have momenta that is proportional
to the intrinsic velocity [c] of the standing wave

rest Matter

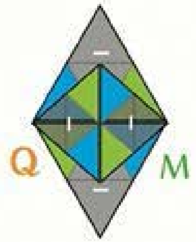

Electro-static particles
have neutralised
magnetic dipoles

Relativistic mass

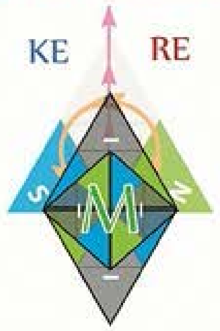

Kinetic motion
produces
Magnetic moments

Spherical point charges do NOT exist

positron

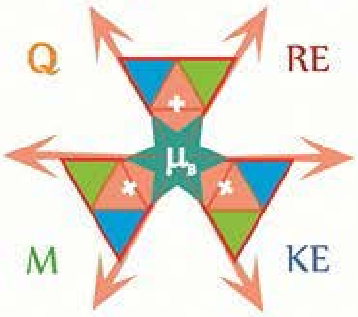

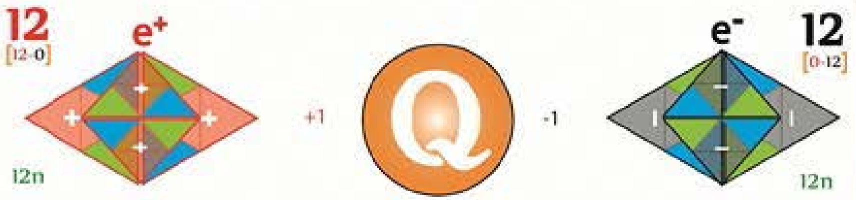
12
[12-0]
e+
12n
+1
Q
-1
e-
12
[0-12]
12n

3D Matter is a standing-wave topology resulting from radiant 2D EM geometries
All leptons and quarks both have 12 charged fasica geometries.
[but differing mass-Matter-particle topologies]

electron

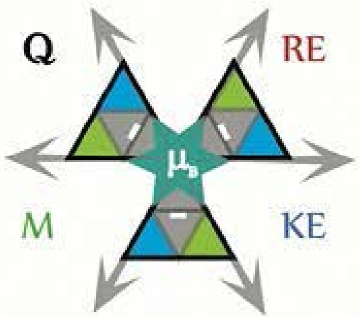

All EM fields resulting from Kinetic Energy (motion) radiate outwards
(the intrinsic KEM fields contain both Negative and Positive Energy momenta quanta)

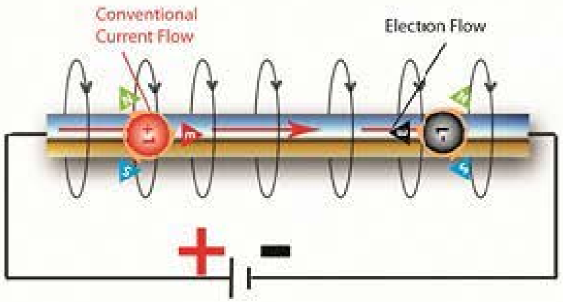
Conventional Current Flow
Electron Flow
+
−

Magnetic field
around a
current carrying
conductor

Magnetic field
produced
by a
Solenoid

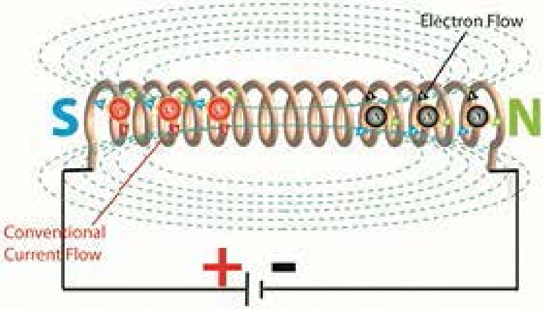
Electron Flow
S
N
Conventional Current Flow
+
−

Biot-Savart Law

is an equation in electromagnetism that describes the magnetic field B generated by an electric current.

μ_0 There is an inverse cube relationship between magnetic field strength and magnetic field force with respect to distance from the magnet. ε_0

The interaction between Magnetostatic fields for which a Magentic dipole exists obeys the inverse CUBED law of attraction and repulsion

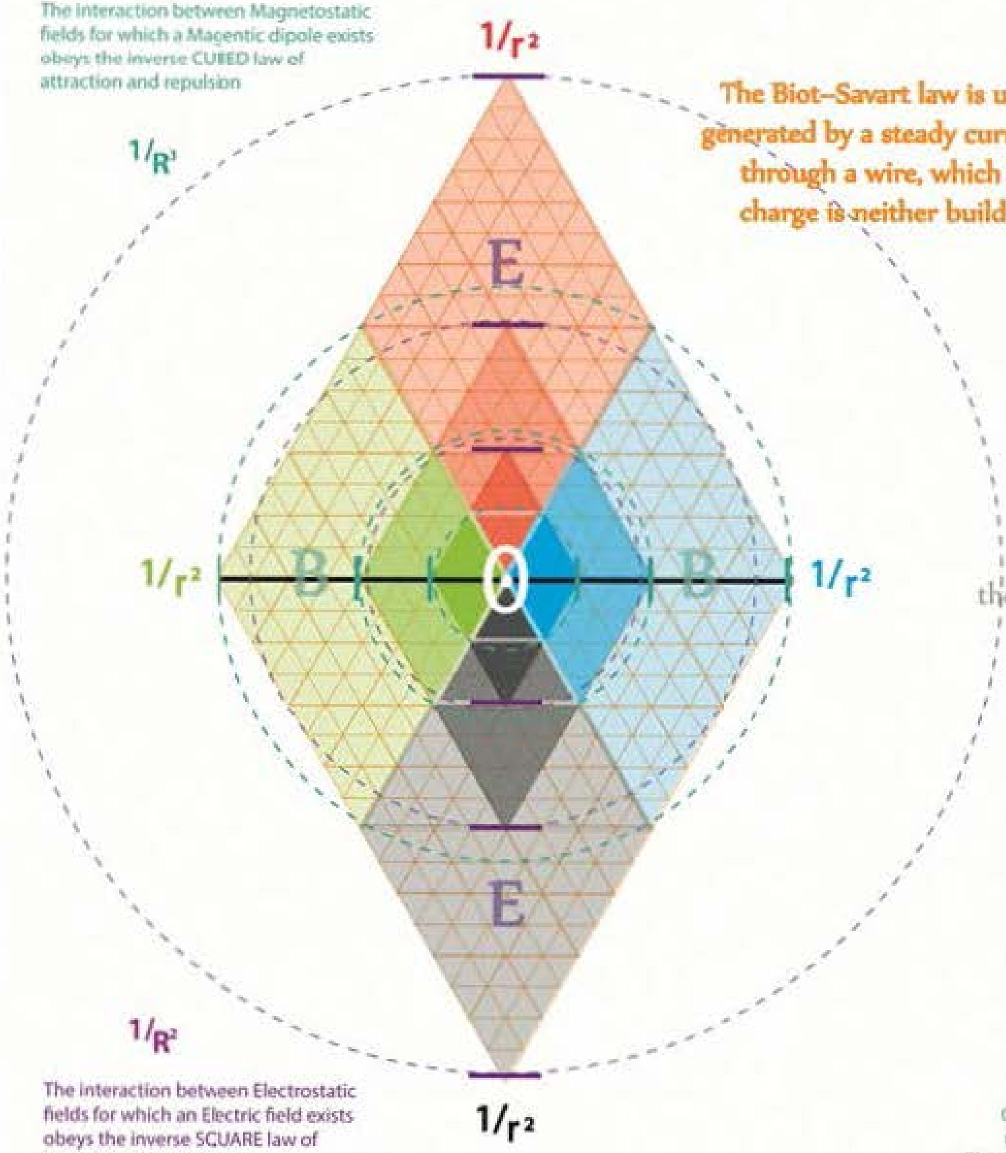

1/r²
1/R³
E
1/r²
B
0
B
1/r²
E
1/R²
1/r²

The interaction between Electrostatic fields for which an Electric field exists obeys the inverse SQUARE law of attraction and repulsion

The Biot–Savart law is used to compute the magnetic field generated by a steady current, i.e. a continual flow of charges, through a wire, which is constant in time and in which charge is neither building up nor depleting at any point

$$\mathbf{B} = \int \frac{\mu_0}{4\pi}\frac{I\,d\mathbf{l}\times\mathbf{r}}{|r|^3},$$

The radial distance between Magnetic dipoles is less than the distance between Electric dipoles

$$\mathbf{B} = \int \frac{\mu_0}{4\pi}\frac{I\,d\mathbf{l}\times\hat{\mathbf{r}}}{|r|^2},$$

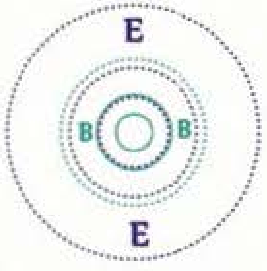

Close to one pole of a magnet, B field strength resembles the inverse square of Electric force. This is because it behaves as a "unipolar magnetic field"

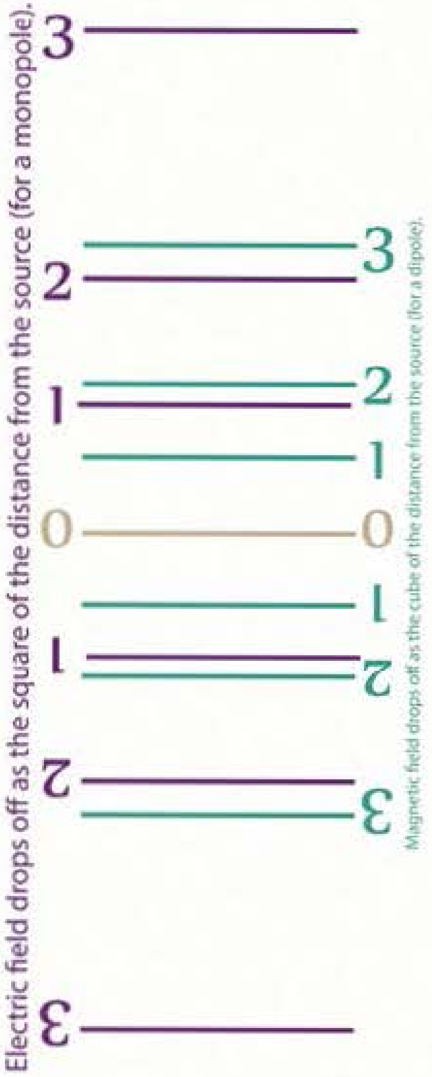

Electric field drops off as the square of the distance from the source (for a monopole).
3
2
1
0
1
2
3
Magnetic field drops off as the cube of the distance from the source (for a dipole).
3
2
1
0
1
2
3

Copyright ABRAHAM [2008] - all rights reserved

Quantum Chirality

Its the equilateral geometry of quantised angular momenta that creates chirality in physics

Mirror imaged Planck quanta are NOT identical to each other

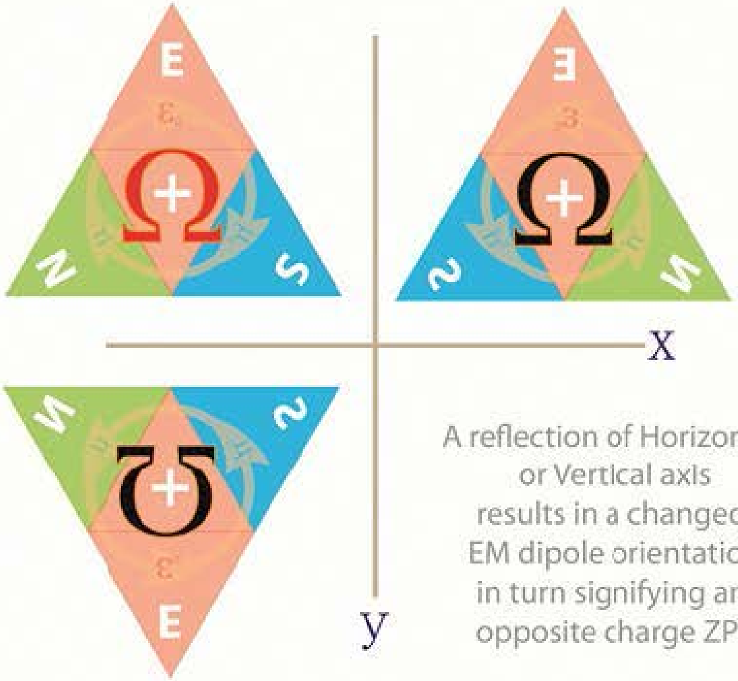

A reflection of Horizontal or Vertical axis results in a changed EM dipole orientation in turn signifying an opposite charge ZPF

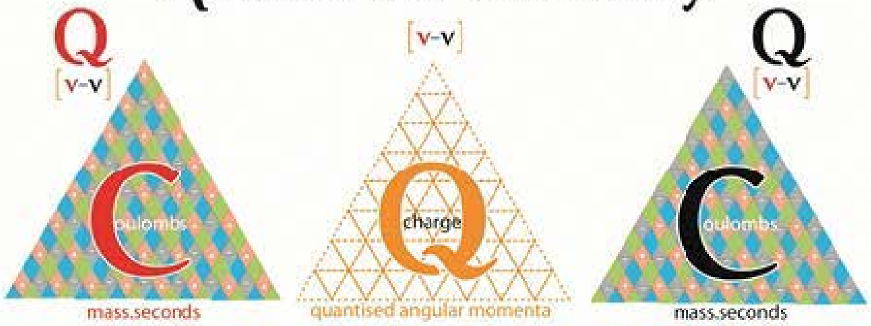
Q
[v-v]
C
oulombs
mass.seconds
[v-v]
Q
charge
quantised angular momenta
Q
[v-v]
C
oulombs
mass.seconds

The equilateral geometry of any EM field or Matter particle is determined by its nett Coulombic charge

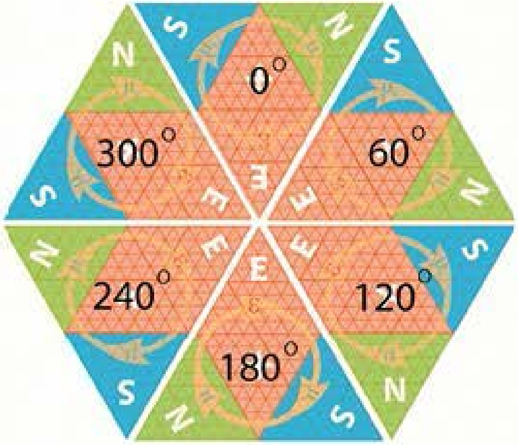
0°
60°
120°
180°
240°
300°
N
S
E
ε

Any nominal rotation about an axis results in a re-orientation of the electromagnetic vectors but does not affect any change to charge etc.

Irrespective of orientation or rotation:

Ω Positive charge fields have clockwise inductive flux geometries

Negative charge fields have counter-clockwise inductive flux geometries Ω

Pi radian - energy momenta geometries

mass-energy momenta

Q

nett charge

[v-v]

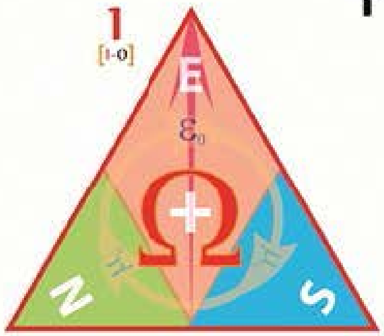
1
[1-0]
E
ε0
Ω
+
N
S
μ

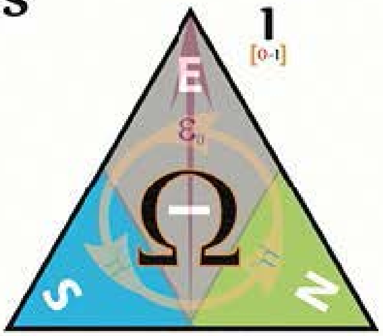
1
[0-1]
E
ε0
Ω
−
S
N
μ

π

pi radian energy geometries

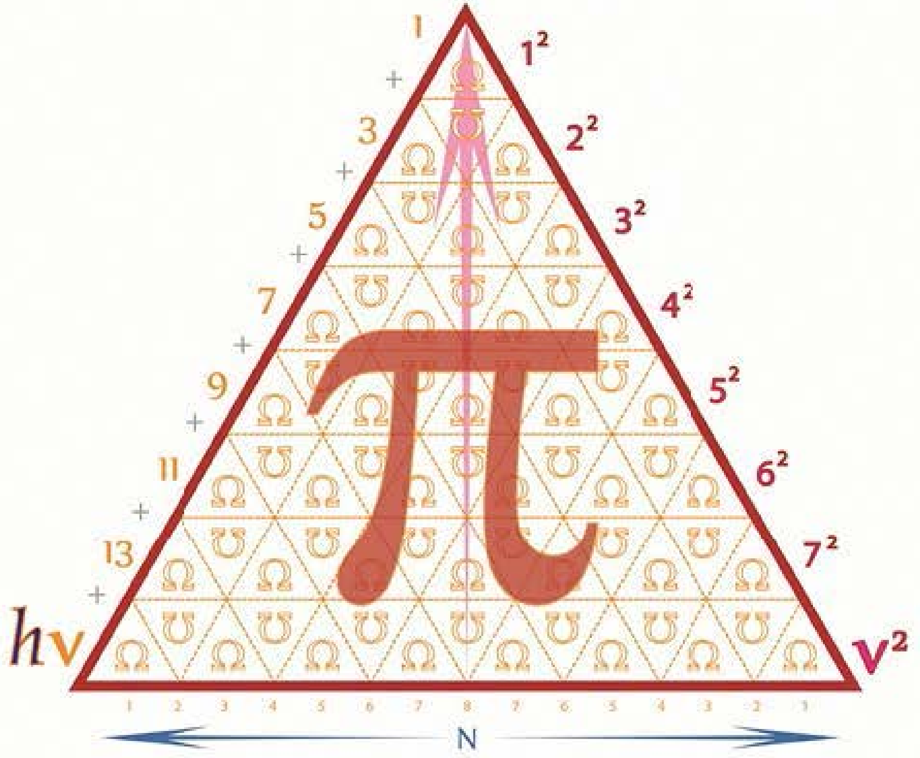
1
+
3
+
5
+
7
+
9
+
11
+
13
+
1^2
2^2
3^2
4^2
5^2
6^2
7^2
hν
v^2
π
Ω
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
N

Ω

quantised angular momentum

$$n\pi\,[m\Omega v]^2$$

Planck quanta

energy mass velocity

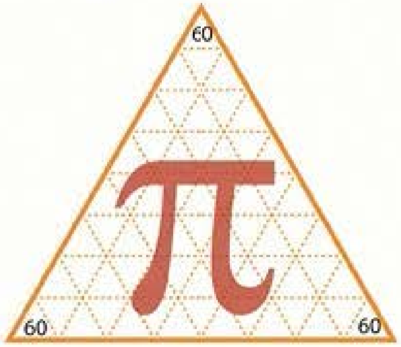

π radian geometries are comprised of equilateral mass momenta Ω

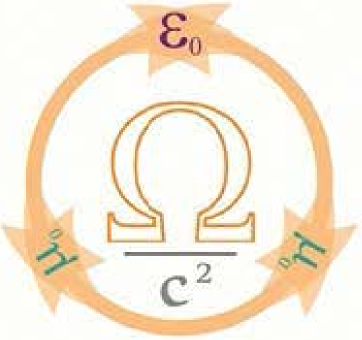

ZPFs and Bosons

ZPFs are Planck energy quantum elements

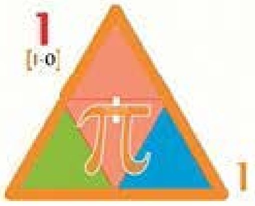

Bosons are transverse Quantum levels

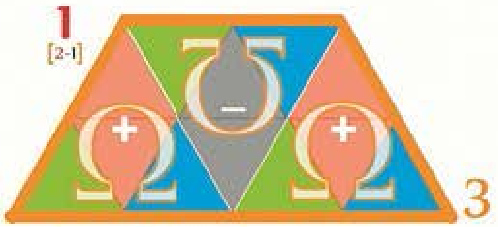

Quantised angular momentum is the foundation for all Tetryonic mass-ENERGY-Matter geometries

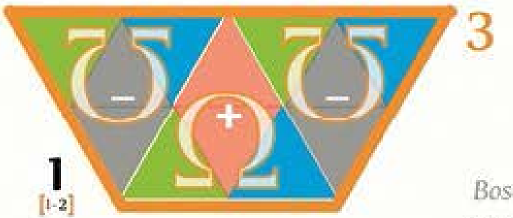

Bosons are the exchange particles for the EM force

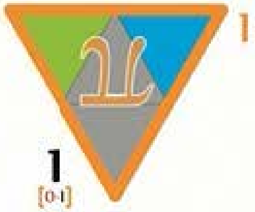

ZPFs are the quanta geometry for Charge, Bosons and Energy

Each Quantum level is an ODD number of quanta (Bosons)

Each 'squared' energy field is the sum of the preceeding Bosons

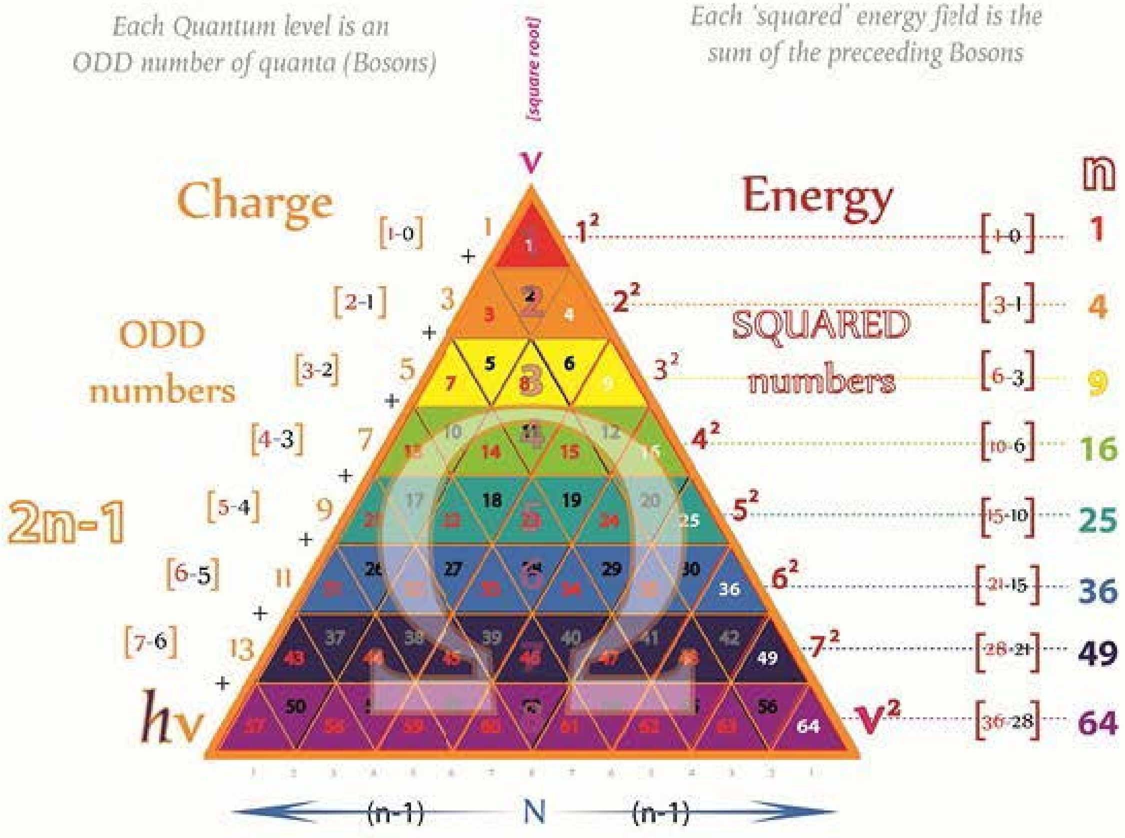
[square root]
v
Charge
Energy
n
ODD
numbers
SQUARED
numbers
2n-1
hv
[1-0] 1
[2-1] 3
[3-2] 5
[4-3] 7
[5-4] 9
[6-5] 11
[7-6] 13
+
1^2 [1-0] 1
2^2 [3-1] 4
3^2 [6-3] 9
4^2 [10-6] 16
5^2 [15-10] 25
6^2 [21-15] 36
7^2 [28-21] 49
v^2 [36-28] 64
(n-1)
N
(n-1)

The charged quanta in all mass-Energy geometries create a NORMAL Distributions

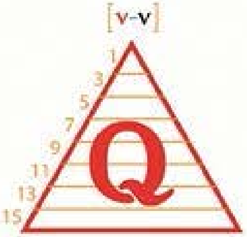
[v-v]
1
3
5
7
9
11
13
15
Q

EM Field Planck quanta

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons ElectroMagnetic mass velocity

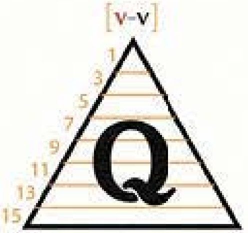
[v-v]
1
3
5
7
9
11
13
15
Q

Bosons form the geometry of Quantum Levels

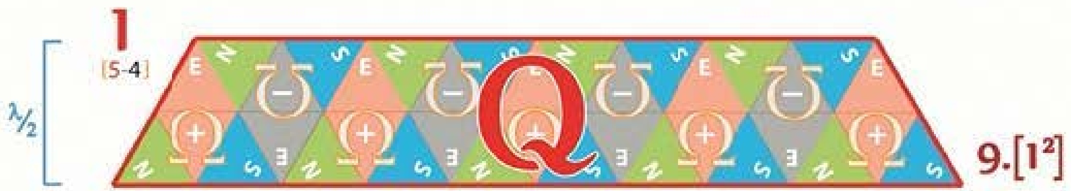

$$W^{+} = n\underset{ODD}{\pi}\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Force carrier for Positive charge particles

W⁺ Boson

EM field Planck quanta

$$\underset{Bosons}{ODD\,\pi}\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

ElectroMagnetic mass velocity

CHARGE carrier Bosons
are ODD number quanta

Bosons

form unit charge Quantum levels
that facilitate EM induction between
mass-Energy-Matter

Neutral Z Bosons and Photons
are EVEN quanta Bosons

EMField Planck quanta

$$\underset{Photons}{EVEN\,\pi}\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

ElectroMagnetic mass velocity

Neutral charge parallelogram geometry EM force carriers

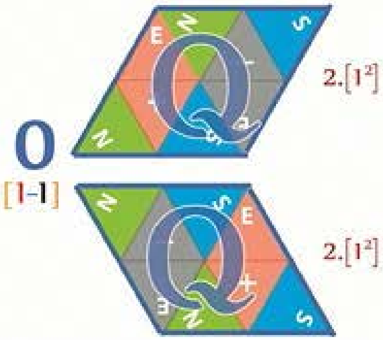

Z⁰ Boson

Neutral Z Bosons can be formed by combining EVEN numbers of W Bosons

They are EVEN numbered alternate geometry Photons

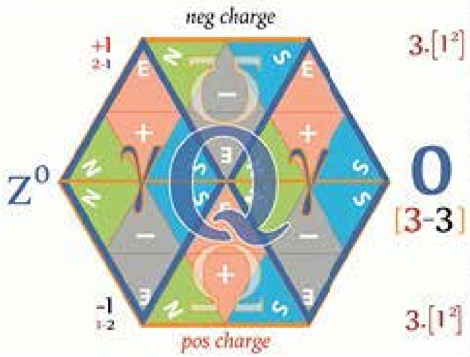

Photons are alternate (diamond) EM geometries formed from Z Bosons

W⁻ Boson

Force carrier for Negative charge particles

$$W^{-} = n\underset{ODD}{\pi}\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

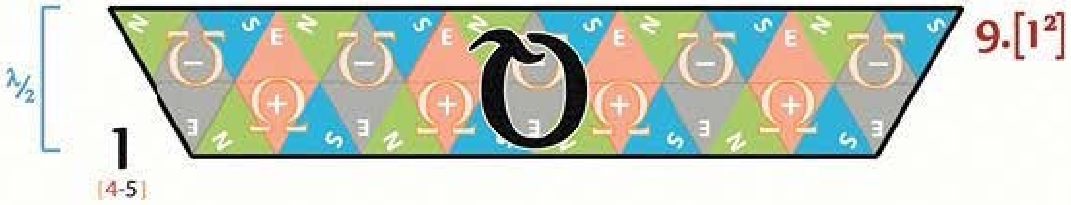

Bosons are charge carriers

[the geometric foundation of quantum levels]

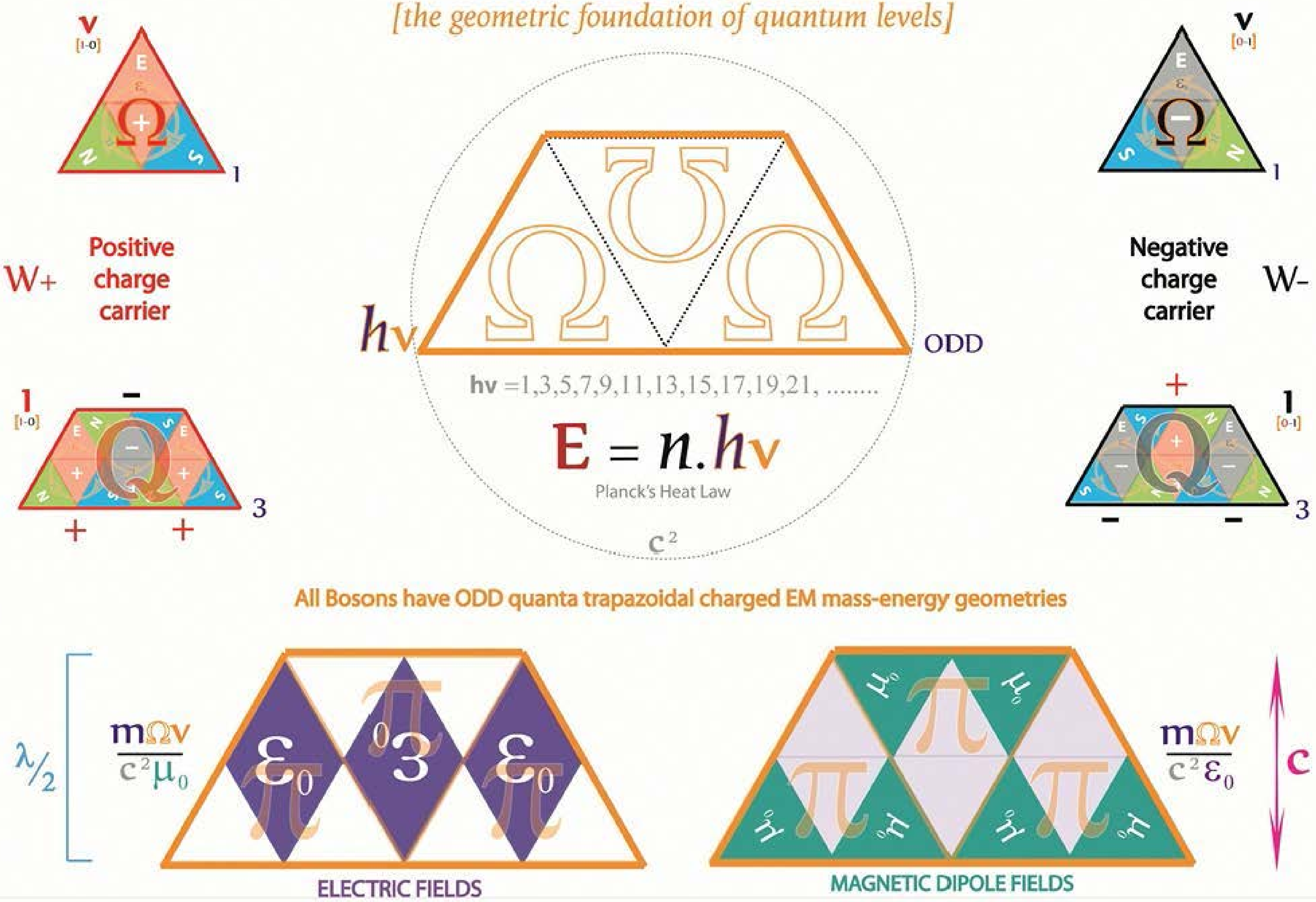
v
[1-0]
E
Ω
+
N
S
1
1
[1-0]
−
E
+
3
+
+
hv
ODD
c²
v
[0-1]
E
Ω
−
S
N
1
+
1
[0-1]
3
−
−
λ/2
$\frac{m\Omega v}{c^2\mu_0}$
ε_0
π
$\frac{m\Omega v}{c^2\varepsilon_0}$
μ_0
c

W+ Positive charge carrier

Negative charge carrier W-

hv =1,3,5,7,9,11,13,15,17,19,21, ...

$$E = n.hv$$

Planck's Heat Law

All Bosons have ODD quanta trapazoidal charged EM mass-energy geometries

ELECTRIC FIELDS

MAGNETIC DIPOLE FIELDS

Tetryonics 05.03 - Boson EM field geometry

Copyright ABRAHAM [2008] - all rights reserved

Boson Frequencies

All Bosons are 1/2 wavelength EM fields with ODD number quanta
They are the geometric basis for transverse EM field Quantum levels

EM Field Planck quanta
$\text{ODD}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
Bosons ElectroMagnetic mass velocity

W Bosons are comprised of
ODD number quanta

Bosons are TRANSVERSE Charge carriers

ODD #

Q

Planck quanta
$n\pi\left[\left[m\Omega v^2\right]\right]$
mass velocity

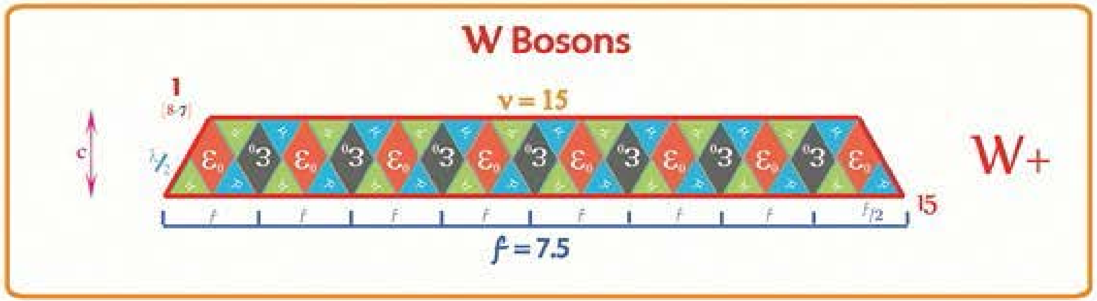
W Bosons
1
[8-7]
v = 15
c
½
ε0
ε0
ε0
ε0
ε0
ε0
ε0
ε0
f
f
f
f
f
f
f
f/2
15
f = 7.5
W+

Neutral Z BOSONS and PHOTONS have differing EM geometries

EVEN #

Q

Planck quanta
$2n\pi\left[\left[m\Omega v^2\right]\right]$
mass velocity

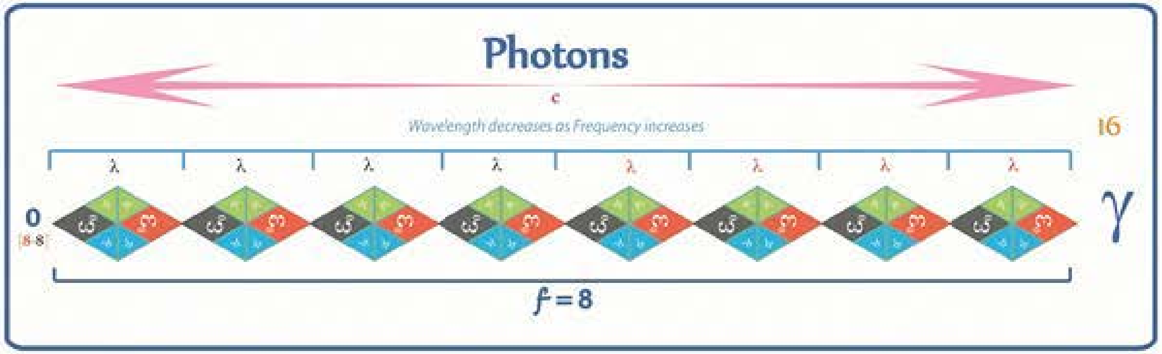
Photons
c
Wavelength decreases as Frequency increases
16
λ
λ
λ
λ
λ
λ
λ
λ
0
[8-8]
γ
f = 8

Photons are LONGITUDINAL
neutral EM force carriers
resulting from Bosons

ODD #

Q

Planck quanta
$n\pi\left[\left[m\Omega v^2\right]\right]$
mass velocity

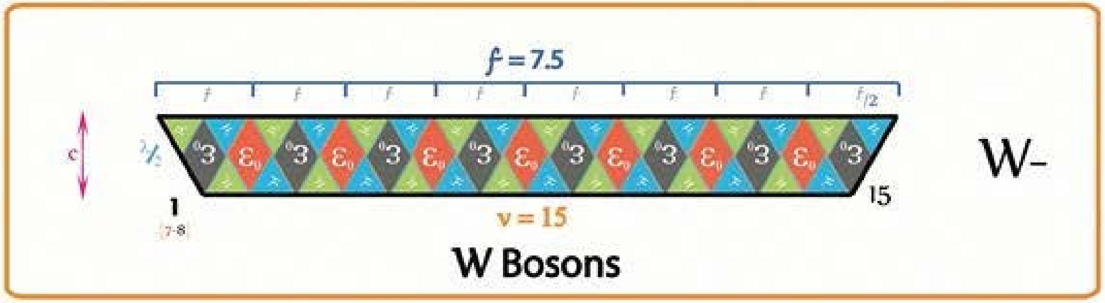
f = 7.5
f
f
f
f
f
f
f
f/2
c
½
15
1
[7-8]
v = 15
W Bosons
W-

EM waves are comprised of
EVEN number quanta [Photons]

EM Field Planck quanta
$\text{EVEN}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$
EM waves ElectroMagnetic mass velocity

Copyright ABRAHAM [2008] - all rights reserved

Boson Waveforms

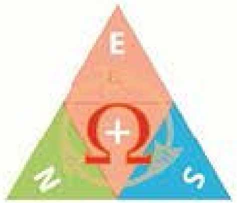

W The Electro-weak force is the result of Bosons interacting along their edge of Permeability

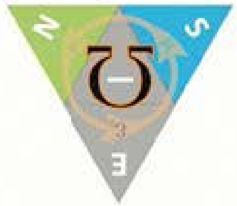

Boson exchange is the basis of Electro-Magnetic Induction & Charge transfer

S

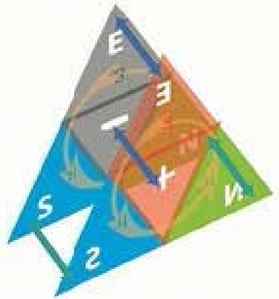

The Strong Colour Force is the result of Bosons interacting via their their Electric charge fascia

All ODD Ω geometries [bosons] create a quantum of charge

Positive charge carrier

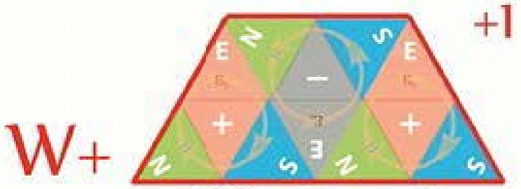

ODD number quanta

All charges are comprised of odd numbered Bosons

[each Boson is a Quantum level]

ODD number quanta

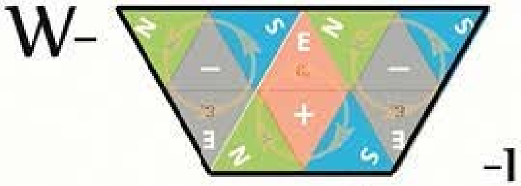

Negative charge carrier

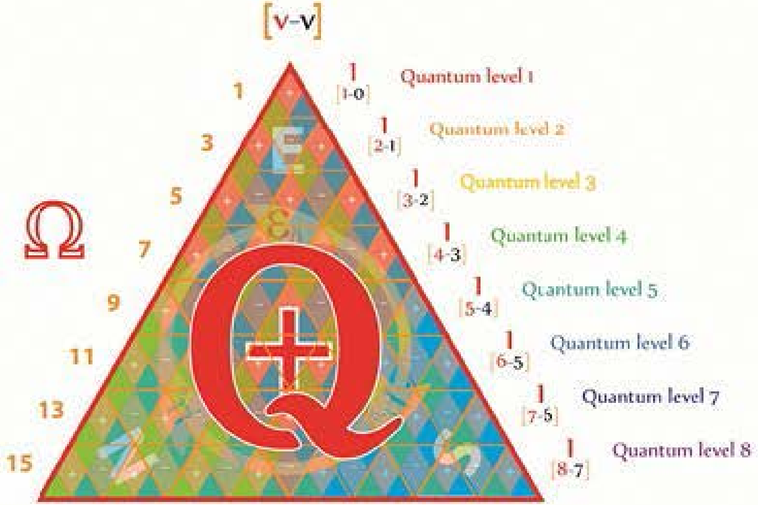
[v-v]
Ω
1
3
5
7
9
11
13
15
1 [1-0] Quantum level 1
1 [2-1] Quantum level 2
1 [3-2] Quantum level 3
1 [4-3] Quantum level 4
1 [5-4] Quantum level 5
1 [6-5] Quantum level 6
1 [7-6] Quantum level 7
1 [8-7] Quantum level 8
E
Q
+
N
S

Integral W Bosons form square Energy geometries
Separated energy geometries create an electromotive force

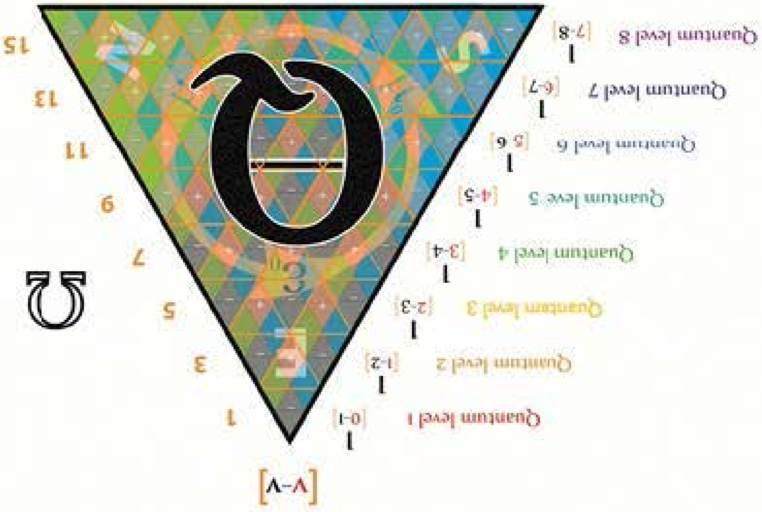
℧
15
13
11
9
7
5
3
1
Quantum level 8 [8-7] 1
Quantum level 7 [7-6] 1
Quantum level 6 [6-5] 1
Quantum level 5 [5-4] 1
Quantum level 4 [4-3] 1
Quantum level 3 [3-2] 1
Quantum level 2 [2-1] 1
Quantum level 1 [1-0] 1
[v-v]
–
E
N
S

Copyright ABRAHAM [2008] - all rights reserved

Bosons

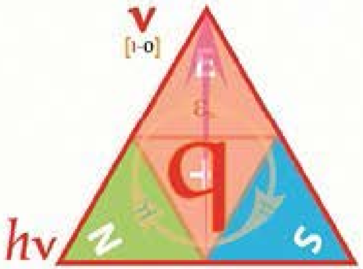

Positive Charges have nett
positive quantised angular momenta

Positive Bosons

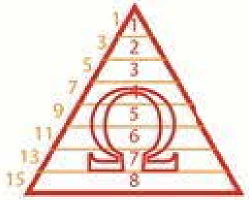
1
3
5
7
9
11
13
15
1
2
3
5
6
7
8
Ω

EM Field Planck quanta

ODD $\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Bosons ElectroMagnetic mass velocity

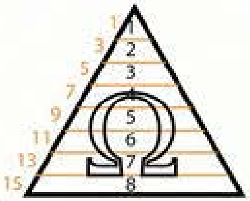
1
3
5
7
9
11
13
15
1
2
3
5
6
7
8
Ω

Bosons are the geometry of Quantum Levels

Negative Bosons

Negative Charges have nett
negative quantised angular momenta

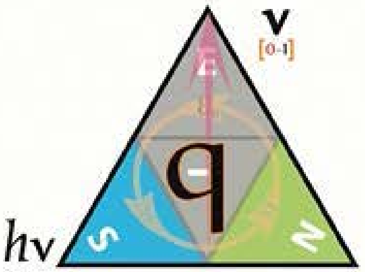

**Bosons are
TRANSVERSE
EM fields [levels]**

Each quantum Level is
a unit of Charge

**nett Charge
is a SCALAR
EM field property**

Charge

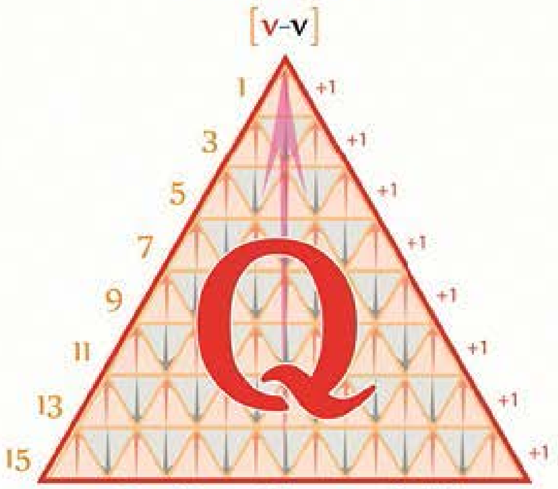
[v-v]
1
3
5
7
9
11
13
15
+1
+1
+1
+1
+1
+1
+1
+1
Q

Positive Charge field

**Charge is a measure of the nett
quantised angular momenta - energy geometry
of any given region of Space-time**

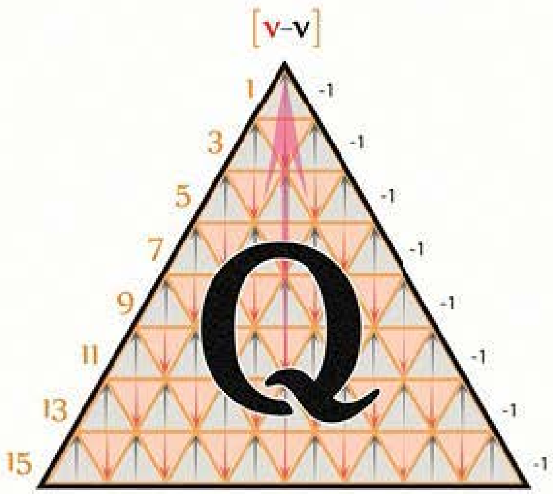
[v-v]
1
3
5
7
9
11
13
15
-1
-1
-1
-1
-1
-1
-1
-1
Q

Negative Charge field

Copyright ABRAHAM [2008] - all rights reserved

bosons

photon

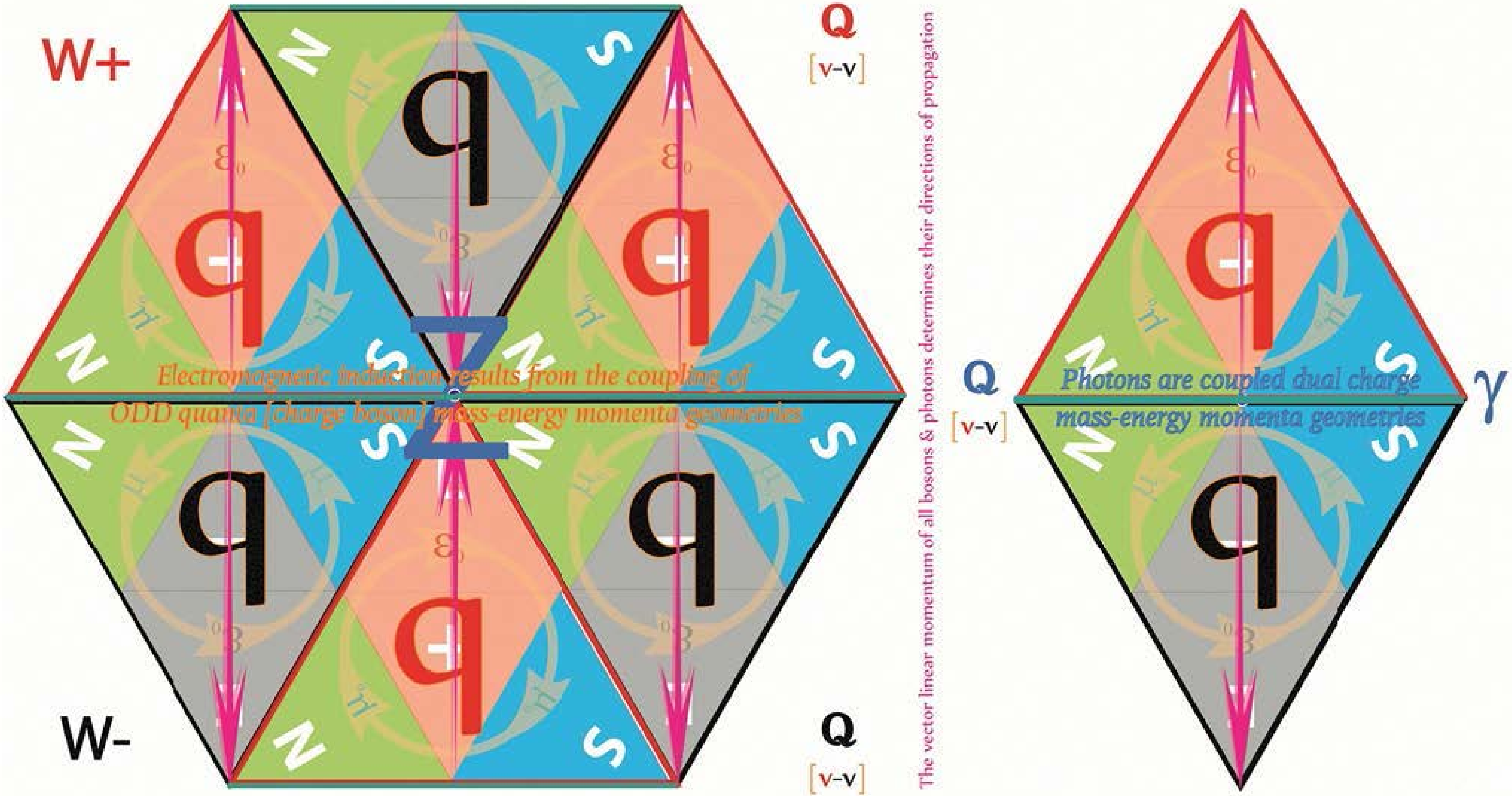

Electromagnetic induction results from the coupling of ODD quanta [charge boson] mass-energy momenta geometries

Photons are coupled dual charge mass-energy momenta geometries

The vector linear momentum of all bosons & photons determines their directions of propagation

ODD π

2nπ

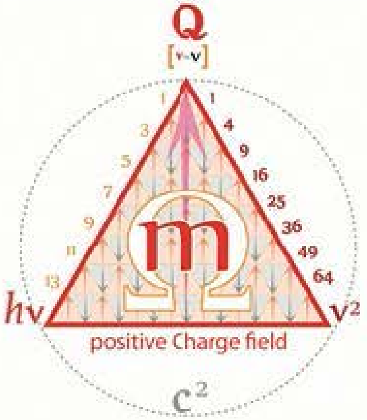
Q
[v-v]
1
1
3
4
5
9
7
16
9
25
11
36
13
49
64
m
Ω
hv
v²
positive Charge field
c²

mass geometries

$$m \quad \overset{\text{mass}}{\underset{\text{energy}}{n\pi}}\left[\overset{\text{spatial field}}{\underset{\text{impedance}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\;\Omega\;\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \quad E$$

mass-energy

the distribution of energy momenta in EM fields creates charged fascia

charged planar mass-energy momenta form

radiant 2D equilateral geometries

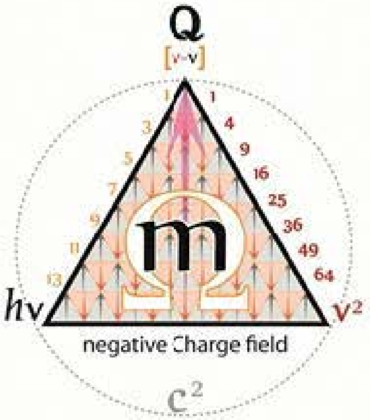
Q
[v-v]
1
1
3
4
5
9
7
16
9
25
11
36
13
49
64
m
Ω
hv
v²
negative Charge field
c²

ρ

mass-energy geometries

m

kg/m²

energy momenta per second

$$\frac{E}{c^2} = \underset{\text{mass}}{m} \qquad \mathbf{E}\ \text{ENERGY} \qquad \underset{\text{Matter}}{M} = \frac{E}{c^4}$$

energy momenta per second squared

ρ

mass-Matter topologies

M

KG/m³

charged mass-energy geometries form

standing-wave 3D tetrahedral topologies

Matter displacement topologies have internalised strong force fascia partitions

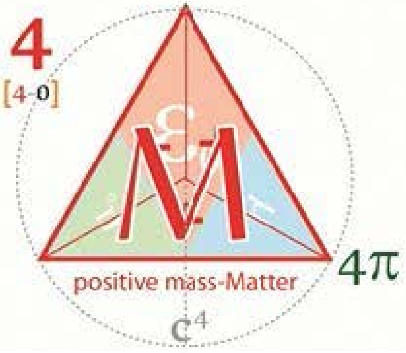
4
[4-0]
ε
M
4π
positive mass-Matter
c⁴

mass-Matter

$$m \quad \overset{\text{}}{\underset{\text{Matter}}{4n\pi}}\left[\overset{\text{spatial field}}{\underset{\text{impedance}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\;\Omega\;\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \quad M$$

Matter topologies

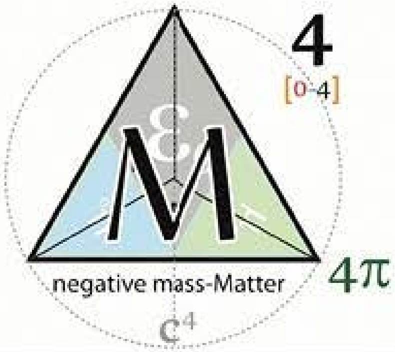
4
[0-4]
ε
M
4π
negative mass-Matter
c⁴

Tetryonics 05.08 - geometric mass-ENERGY-Matter topologies

Matter

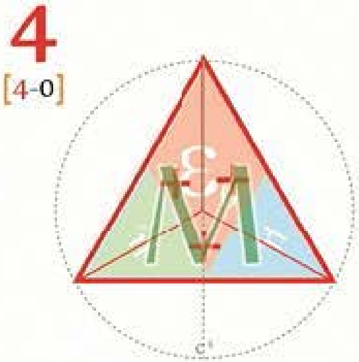

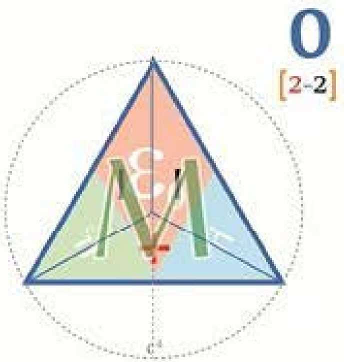

EM mass-energies have
2D planar geometries

$E = mv^2$

2D EM mass-Energy
forms
3D Matter

EM Matter has a
3D Tetrahedral topology

$E = Mc^4$

3D Matter
is comprised of
2D EM mass-Energies

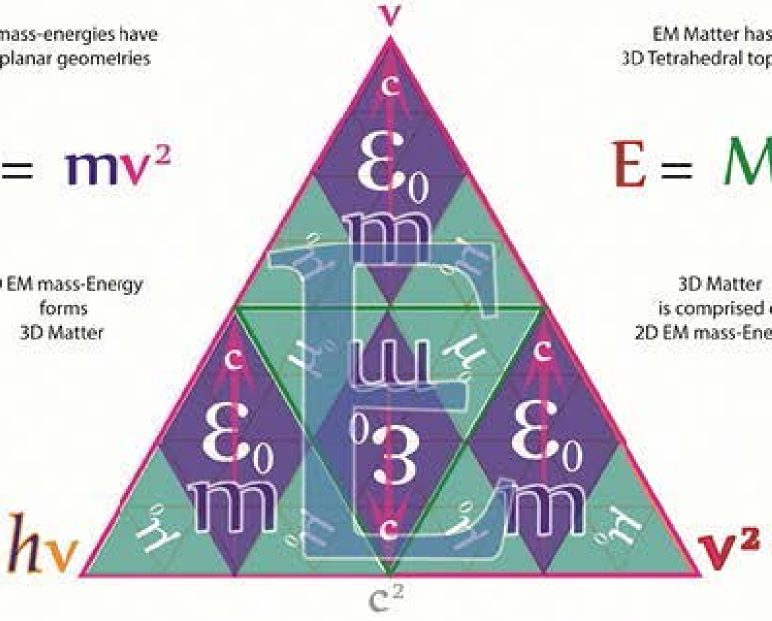

2D

$\rho/\sec$

EM mass-energy per square metre

EM mass geometries

m

kg/m²

3D

$\rho/\sec^2$

EM mass-energy per cubic metre

Matter topologies

M

kg/m³

**Matter is comprised of
4nπ mass-energy geometries
forming a 3D standing wave topology**

Tetryons | EM Field | Planck | quanta

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter quanta | ElectroMagnetic | mass | velocity

Matter is anything that has a closed mass-energy topology and displaces a volume
(the 3D massive building blocks of quantum particles and atomic elements)

The EM mass-energy quanta of Matter can be measured transversely and longitudinally
allowing them to exhibit a number of wave-like and particle-like properties
(de Broglie wavelength & Compton frequency)

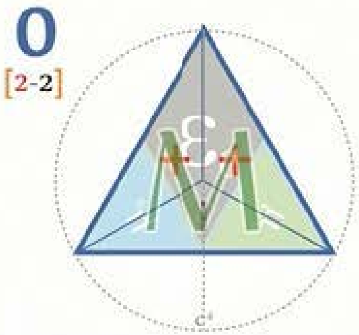

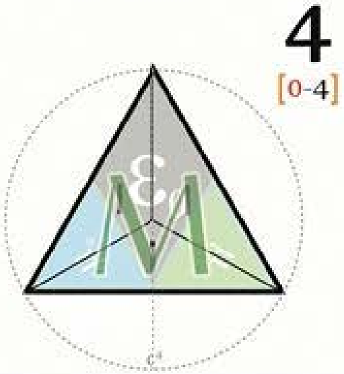

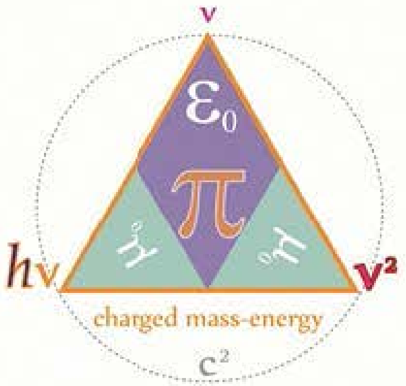
v
ε_0
π
hv
v²
charged mass-energy
c²

Tetryons - the quanta of Matter

'massless' is a physics mis-nomer as all energy exhibits mass equivalence

ZPF

mass-Energy quanta

m

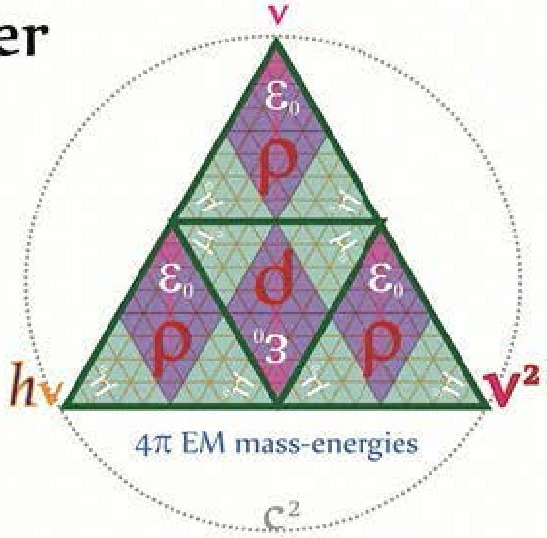
v
ε_0
ρ
d
3
hv
v²
4π EM mass-energies
c²

Energy

$$m = n\pi\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{1/c^2}}\right].\left[\underset{\text{mass}}{m}\,\overset{\text{Planck}}{\Omega}\,\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

momenta

TETRYONS

Platonic tetrahedrons are the foundational topologies of all 3D Matter

Energy

$$T\pi\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{1/c^4}}\right].\left[\underset{\text{mass}}{m}\,\overset{\text{Planck}}{\Omega}\,\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right] = \underset{\text{Matter}}{M}$$

momenta

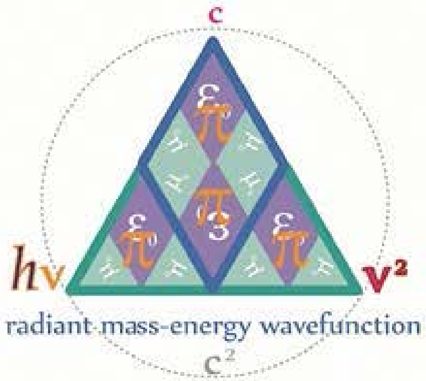
c
ε_0
π
3
hv
v²
radiant mass-energy wavefunction
c²

Tetryon

Matter Quantum

M

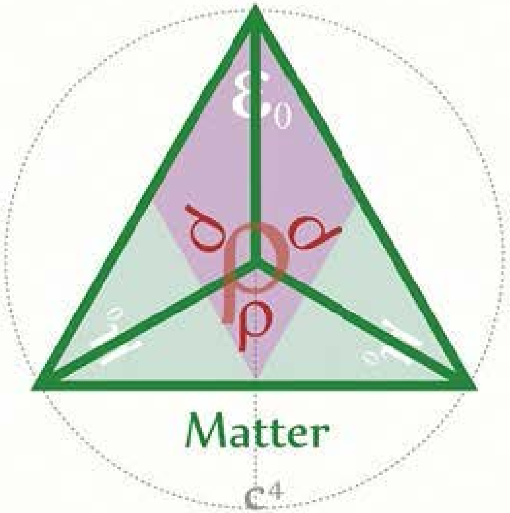
ε_0
ρ
Matter
c⁴

2D mass-energy geometries can be combined to form 3D mass-ENERGY-Matter particles

Standing wave mass-energy geometries and Matter topologies

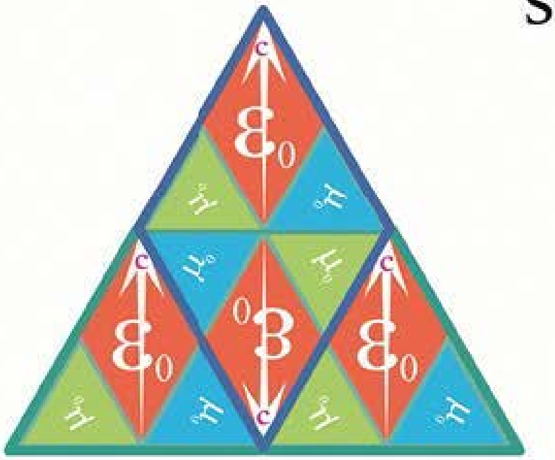
c
ε_0
μ_0
ε_0
ε_0
ε_0
c
c
c

As each charged fascia's E field tries to propagate outwards it interacts with a M field dipole at each apex

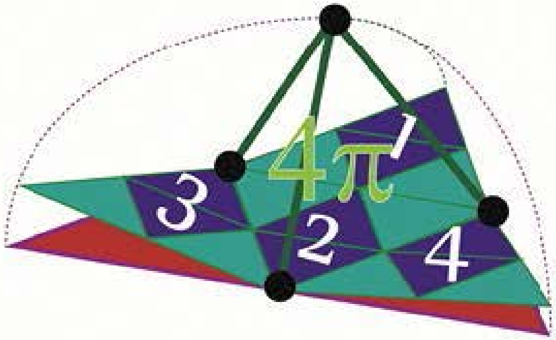

$$\underset{\text{Matter}}{M} = T\pi\left[\left[\underset{\text{impedance}}{\overset{\text{spatial}}{1/c^4}}\right].\overset{\text{Energy}}{\underset{\text{momenta}}{\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\Omega\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]}}\right]$$

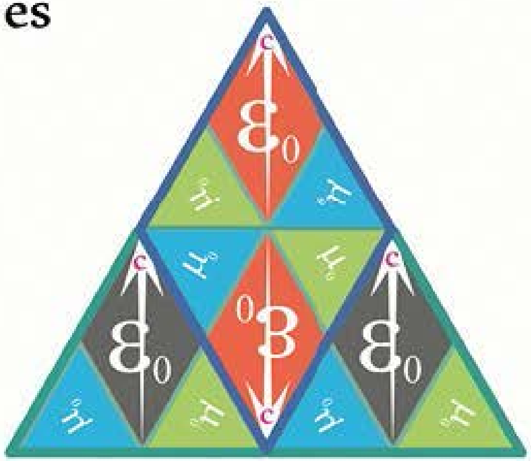
c
ε_0
μ_0
ε_0
ε_0
ε_0
c
c
c

In turn energised M fields supply energy to the E Fields creating a EM standing wave [the EM topology of Matter]

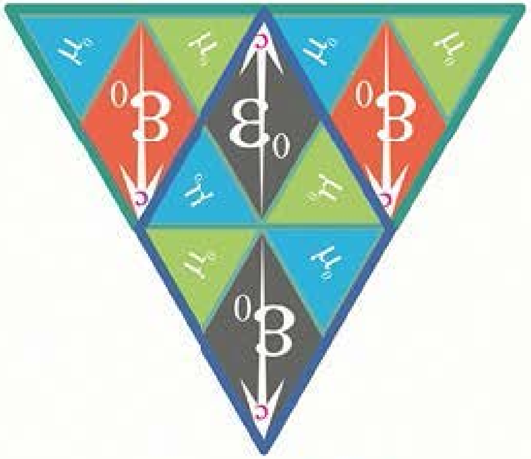
μ_0
ε_0
ε_0
ε_0
ε_0
c

Matter stores EM mass-energies geometries in its 3D Tetrahedral topology

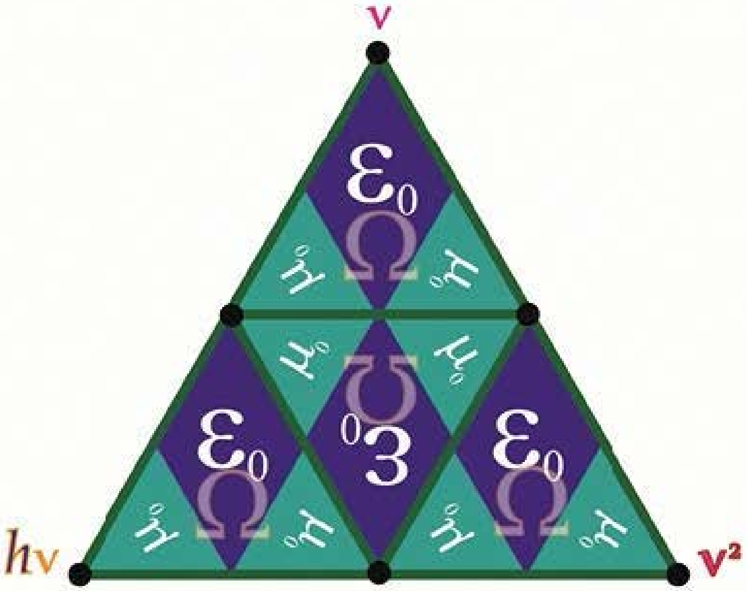
v
hv
v²
ε_0
μ_0
Ω
ε_0
ε_0
ε_0

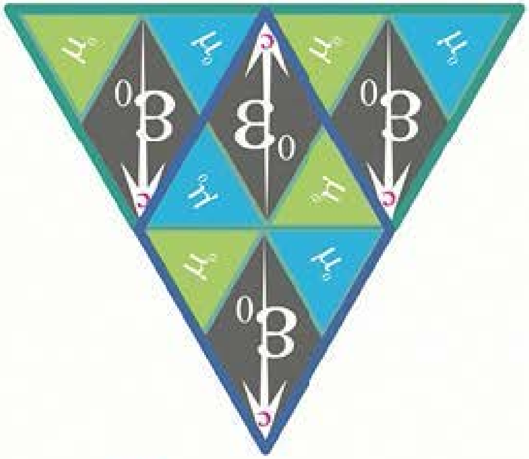
μ_0
ε_0
ε_0
ε_0
ε_0
c

All mass-energy propagate at the speed of light creating linear energy momentum

Tetryonic charge topologies

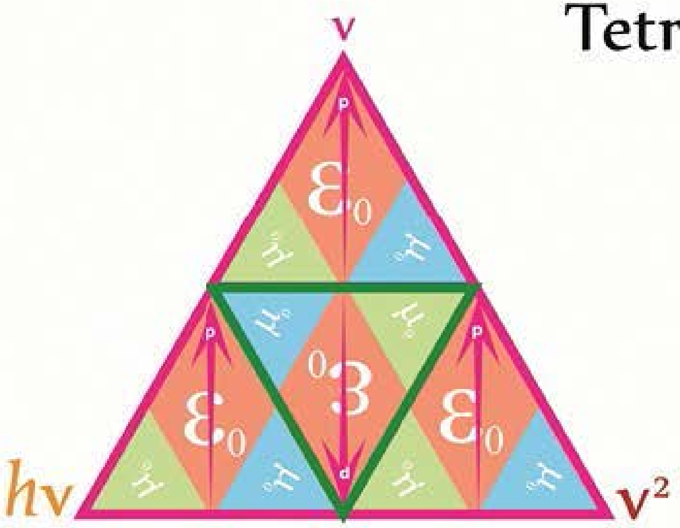

Positive Tetryon

The 3D Tetryonic volume of Matter is what distingushes it from 2D EM mass-energies

Negative Tetryon

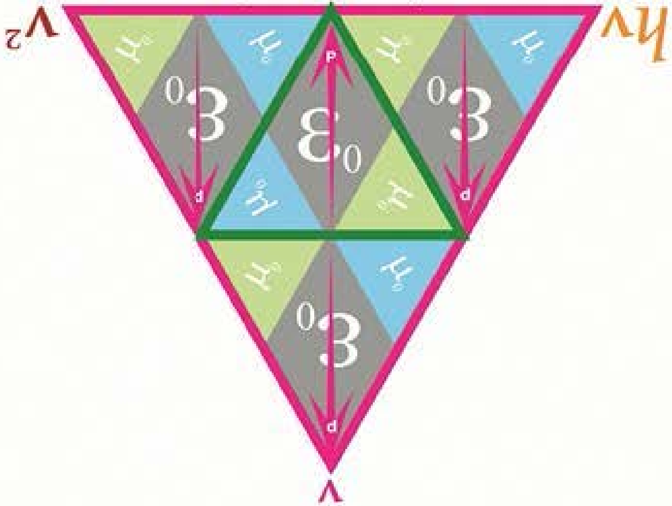

TETRYONS

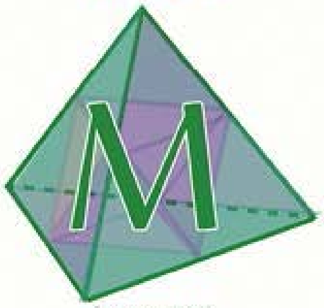

[MATTER]

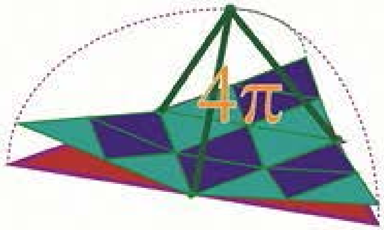

The four EM mass-energy momenta produced by Tetryonic collapse must not be confused with the classical 4 Energy-momenta (which refers to x,y,z co-ordinates & v)

4pi radiant mass-energies

Four 2D mass-energy geometries can combine to form via charge interactions 3D tetrahedral Matter topologies

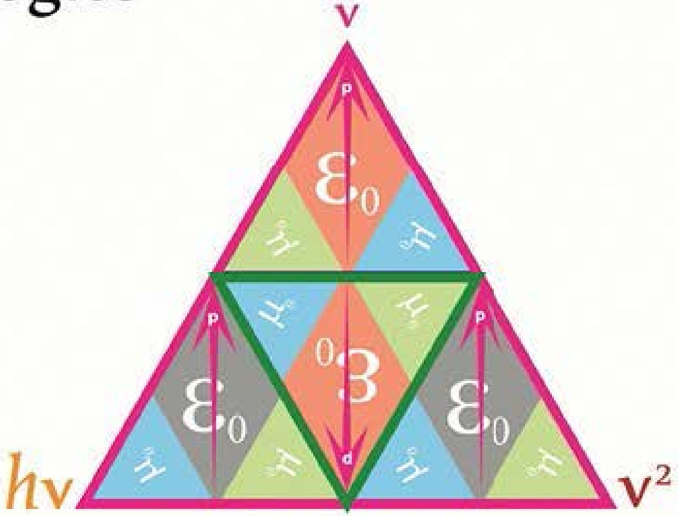

Neutral Tetryon

A 3D Tetryon can be 'collapsed' into a 2D waveform radiant four energy momenta EM waveform that conserves all of its 2D mass-energy-momenta

Neutral Tetryon

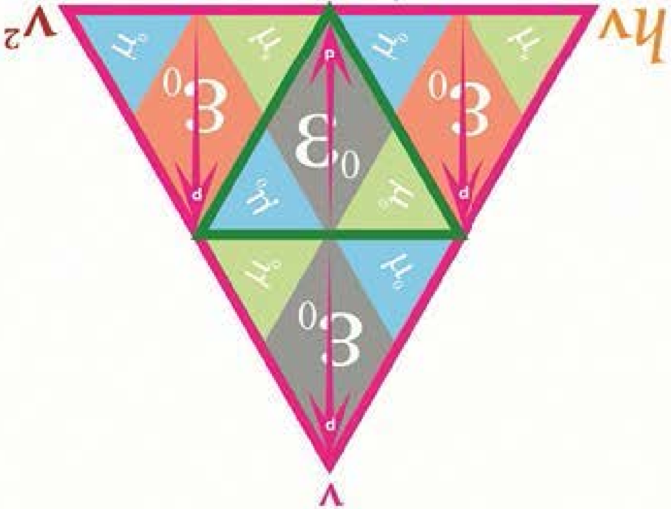

Tetryon Genesis

1. WEAK interaction
Mutual Inductive Coupling
[Magnetic dipoles interact]

ODDπ
charged mass-energies

Non-Zero
nett mass-Energy-momenta
result in Linear momentum

EVENπ
radiant mass-energies

2. *2D EM mass-energies*
interact to form
3D Matter topologies

EM wave momenta
can form standing waves
[Matter geometries]

3. STRONG interaction
Electrostatic Matter attract
via Electric charges and
Magnetic dipoles

4nπ
mass-Matter geometries

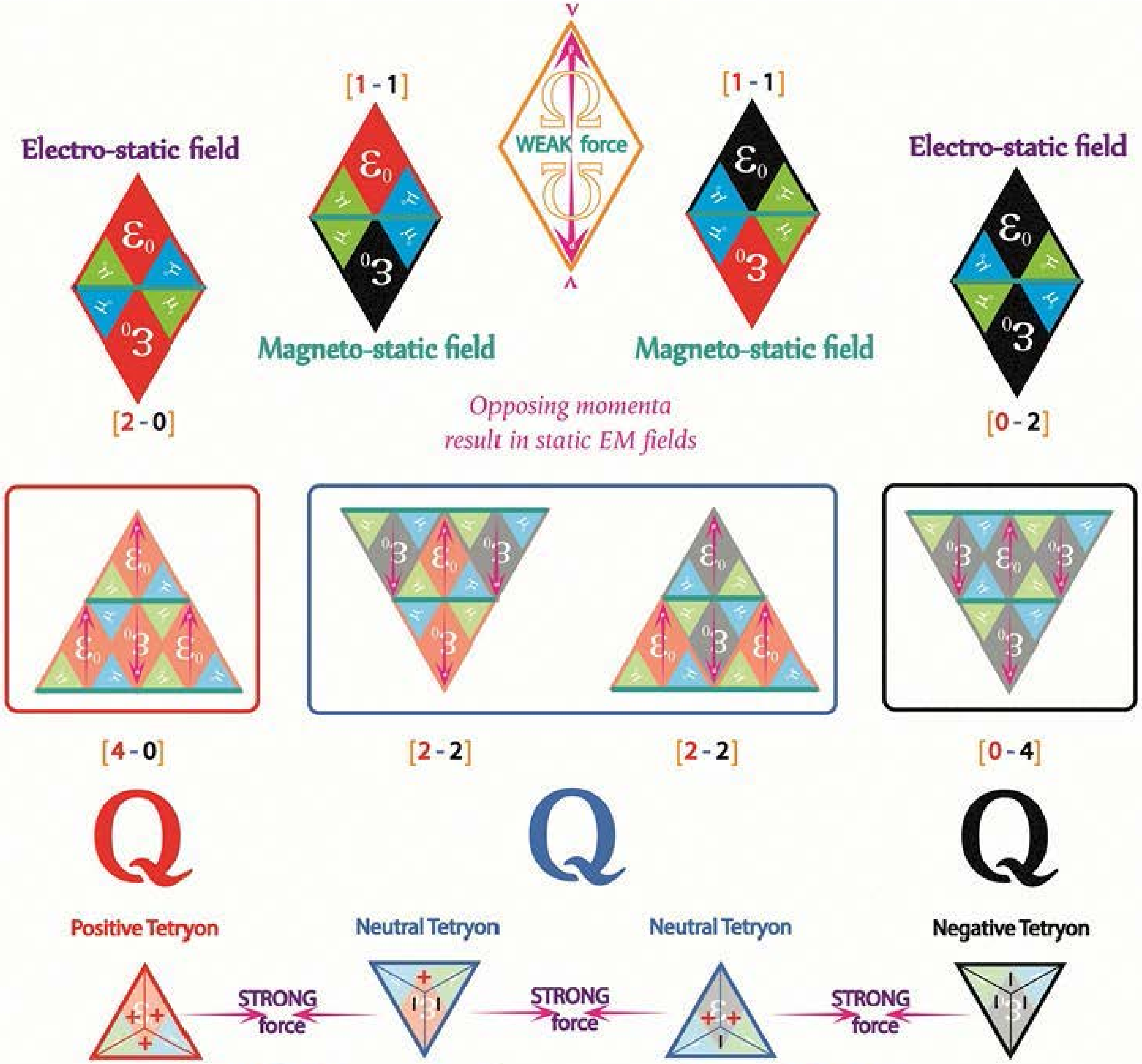
Electro-static field
[1 - 1]
WEAK force
[1 - 1]
Electro-static field
Magneto-static field
Magneto-static field
[2 - 0]
[0 - 2]
[4 - 0]
[2 - 2]
[2 - 2]
[0 - 4]
Q
Q
Q
Positive Tetryon
Neutral Tetryon
Neutral Tetryon
Negative Tetryon
STRONG force
STRONG force
STRONG force

Opposing momenta
result in static EM fields

All Tetryonic charges seek equilibrium

Tetryonics 06.05 - Tetryon genesis

Positive Tetryon

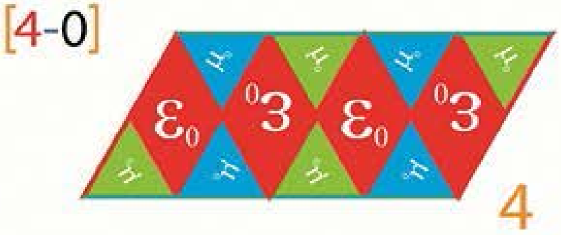
[4-0]
ε0
ε0
ε0
ε0
μ0
4

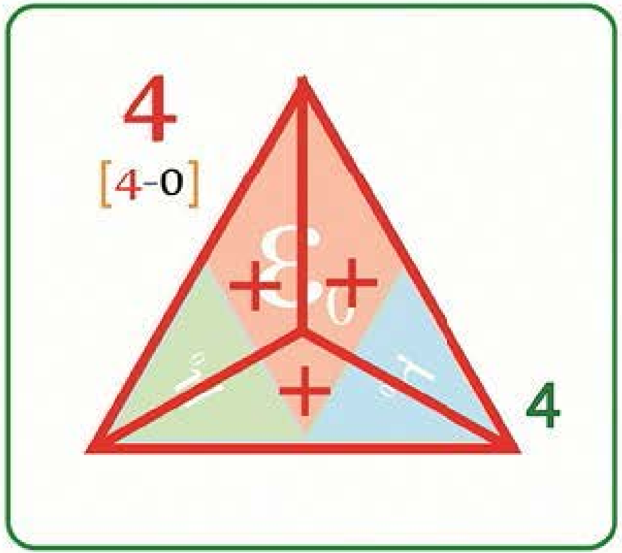

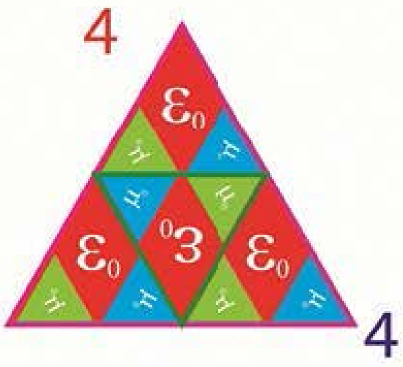
4
ε0
ε0
ε0
ε0
4

charged boson geometry

standing-wave Matter topology

radiant mass-energy geometries

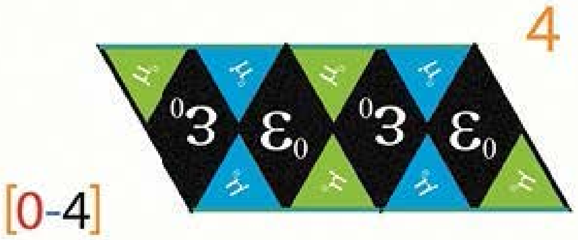
4
ε0
ε0
ε0
ε0
μ0
[0-4]

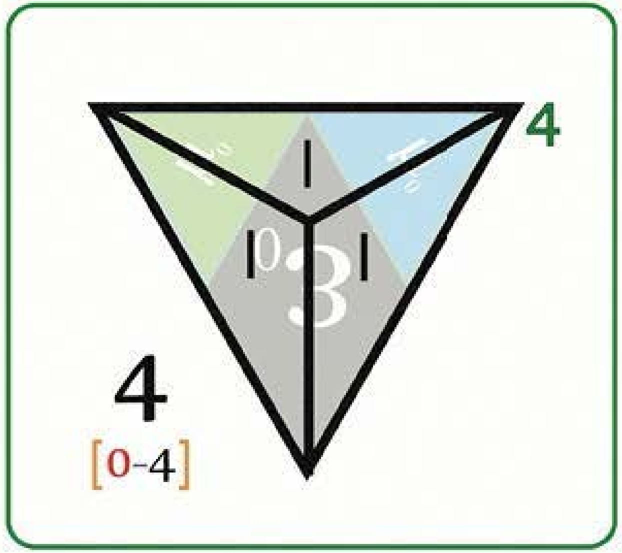

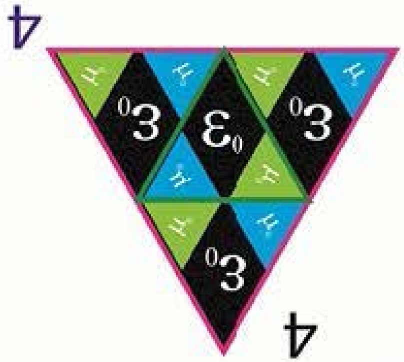
4
ε0
ε0
ε0
ε0
4

Negative Tetryon

Neutral Tetryons

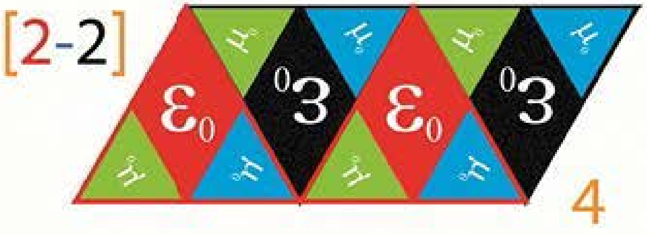
[2-2]
ε0
03
ε0
03
μ
4

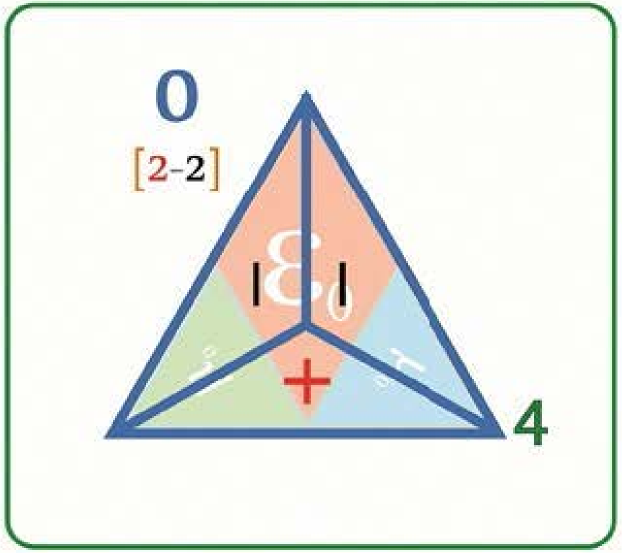

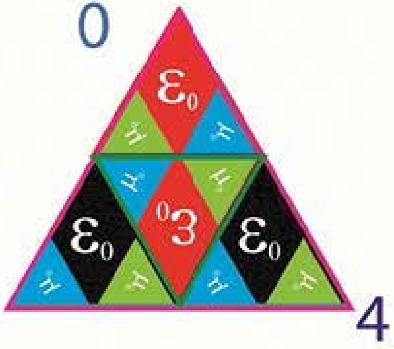
0
ε0
03
ε0
ε0
4

charged boson geometry

standing-wave Matter topology

radiant mass-energy geometries

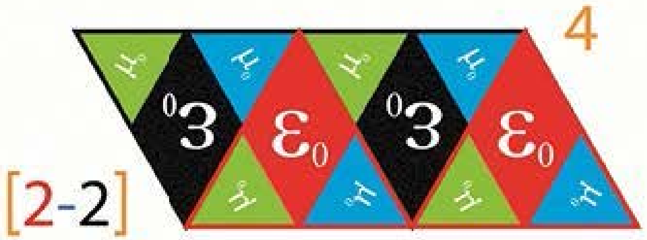
4
03
ε0
03
ε0
[2-2]

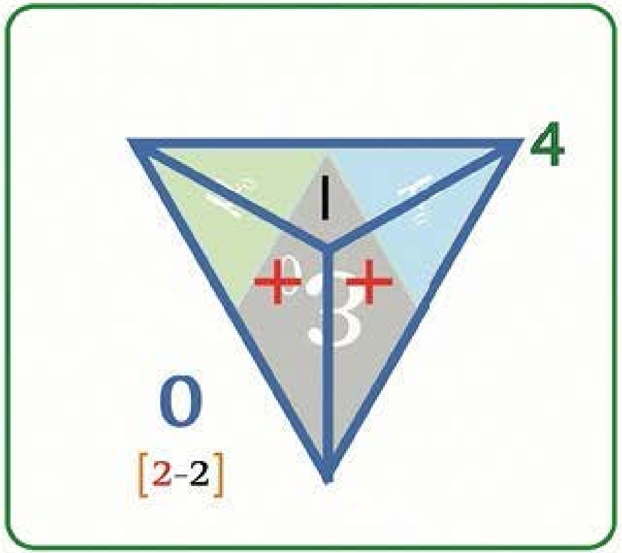

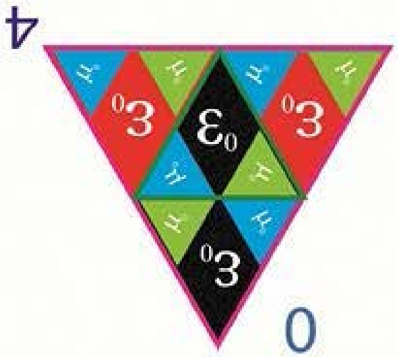
4
03
ε0
03
03
0

mass

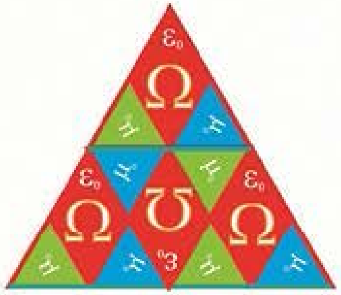

$$\underset{\text{mass}}{m} = n\pi\left[\left[\overset{\text{spatial}}{\underset{\text{impedance}}{1/c^2}}\right].\overset{\text{Energy}}{\underset{\text{momenta}}{\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]}}\right]$$

radiant mass-energy geometries

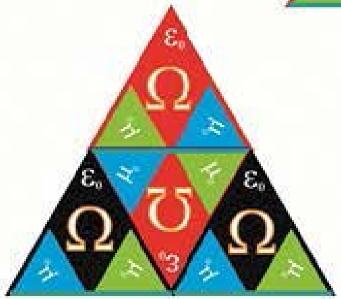

$$E = mc^2$$

Energy per second

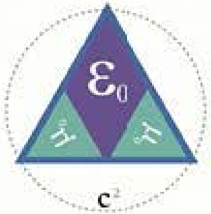

energy

$$n\pi\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]$$

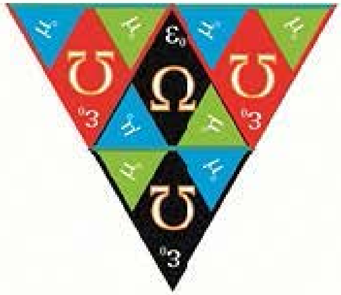

Energy per second ²

$$E = Mc^4$$

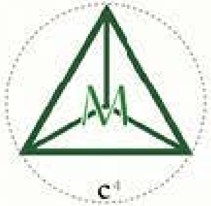

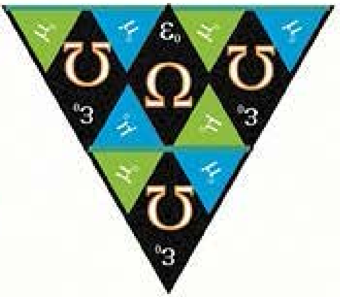

Matter

$$\underset{\text{Matter}}{M} = T\pi\left[\left[\overset{\text{spatial}}{\underset{\text{impedance}}{1/c^4}}\right].\overset{\text{Energy}}{\underset{\text{momenta}}{\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]}}\right]$$

standing wave Matter topologies

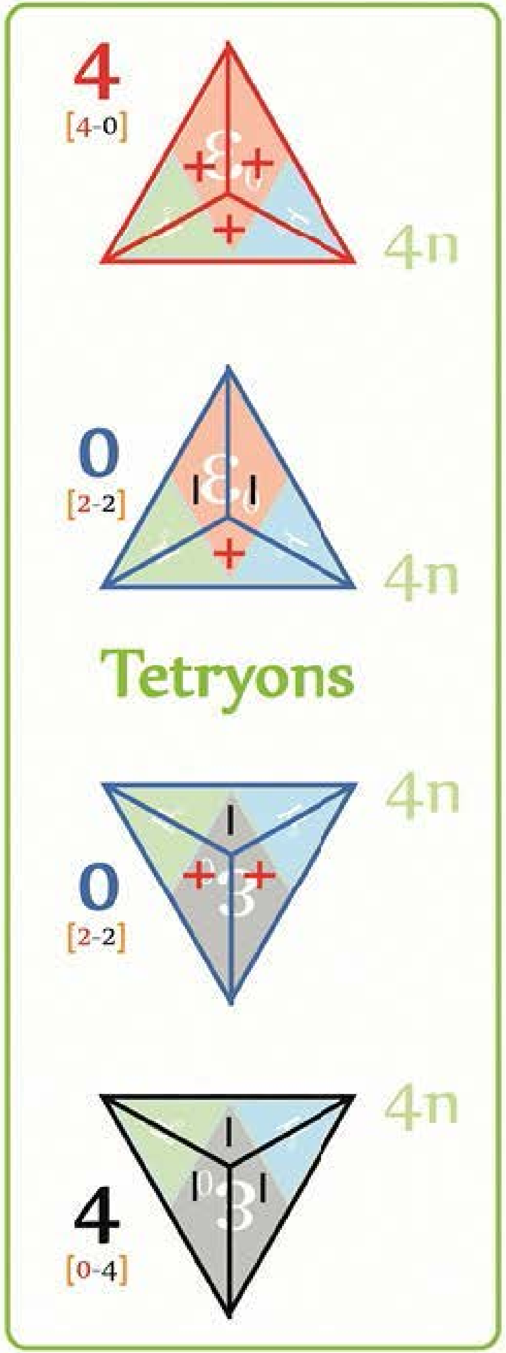

Tetryons

Tetryonics 06.08 - Tetryon family

mass EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

geometry ElectroMagnetic mass velocity

mass-ENERGY-Matter

geometries topologies

ALL EM masss-energies have equilateral geometry

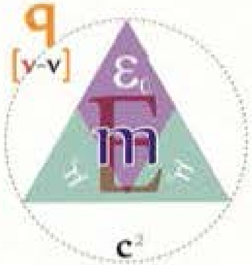

Transverse EM mass-energies form bosons [Quantum levels]

2D mass-energy geometries are a property of 3D Matter topologies

Longitudinal EM mass-Energies form Photons

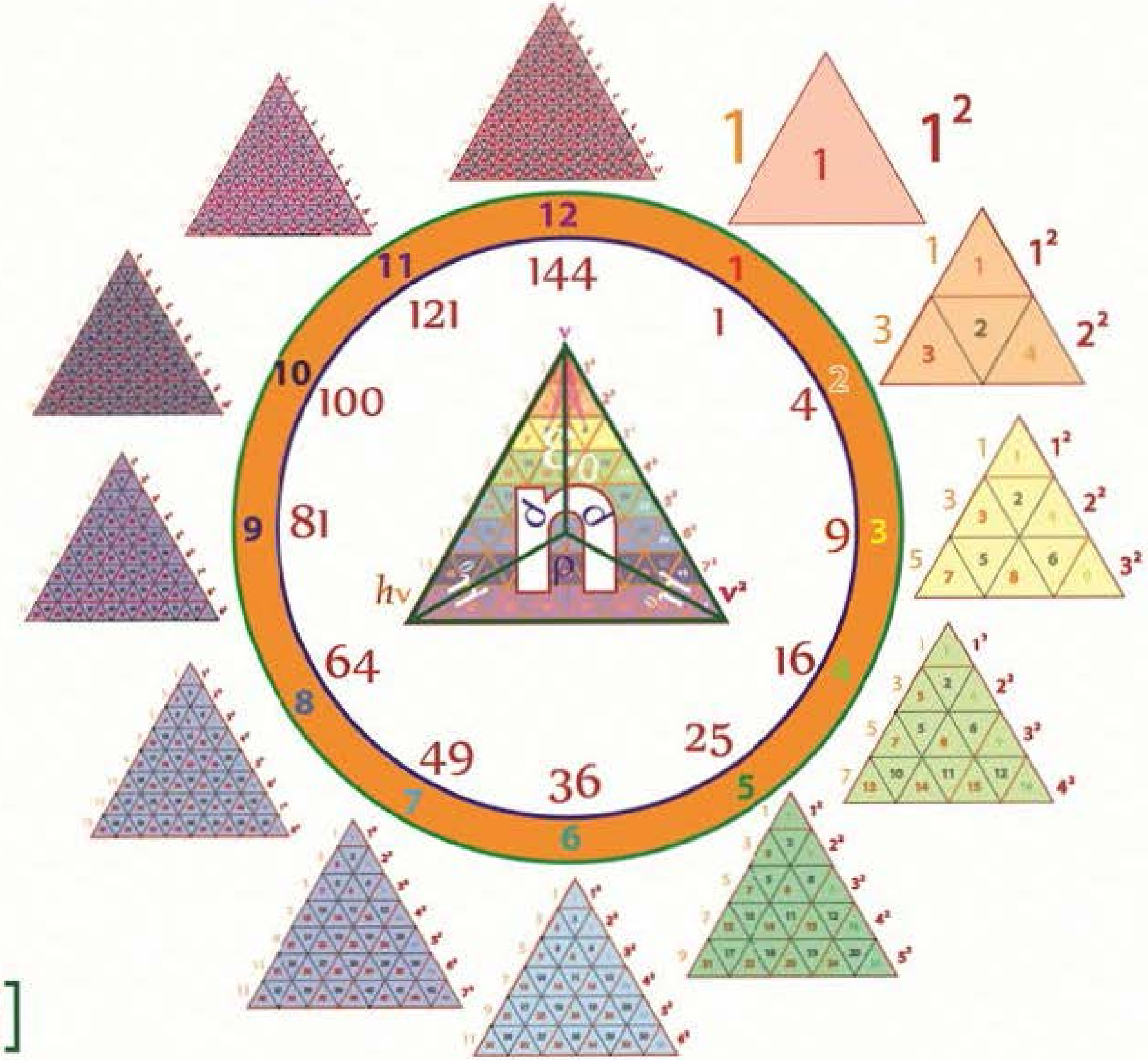

Bosons combine to form SQUARED scalar EM energies

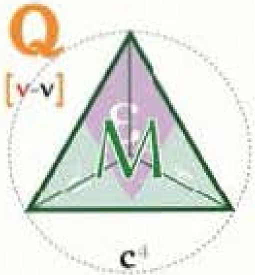

massless is a scientific misnomer

Tetryonic Matter is the building block of all Fermions, elements and compounds

2D [non-Topolgical] EM mass-ENERGY is 'Matterless'

ALL Matter has a 3D topology

Matter EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

topology ElectroMagnetic mass velocity

Tetryonic Matter

All quantum Matter has a foundational tetrahedral topology as a result of their equilateral mass-energy geometries

[not spherical as has been assumed from the math]

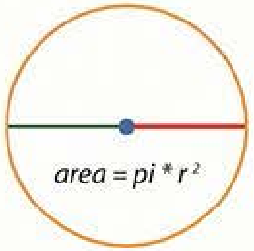

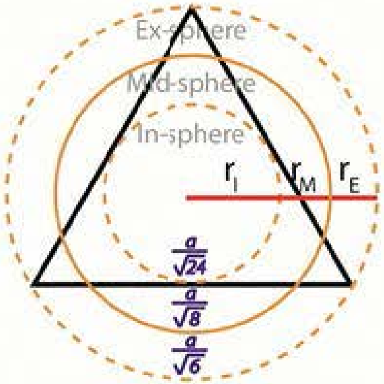
Ex-sphere
Mid-sphere
In-sphere
r_I
r_M
r_E
$\frac{a}{\sqrt{24}}$
$\frac{a}{\sqrt{8}}$
$\frac{a}{\sqrt{6}}$

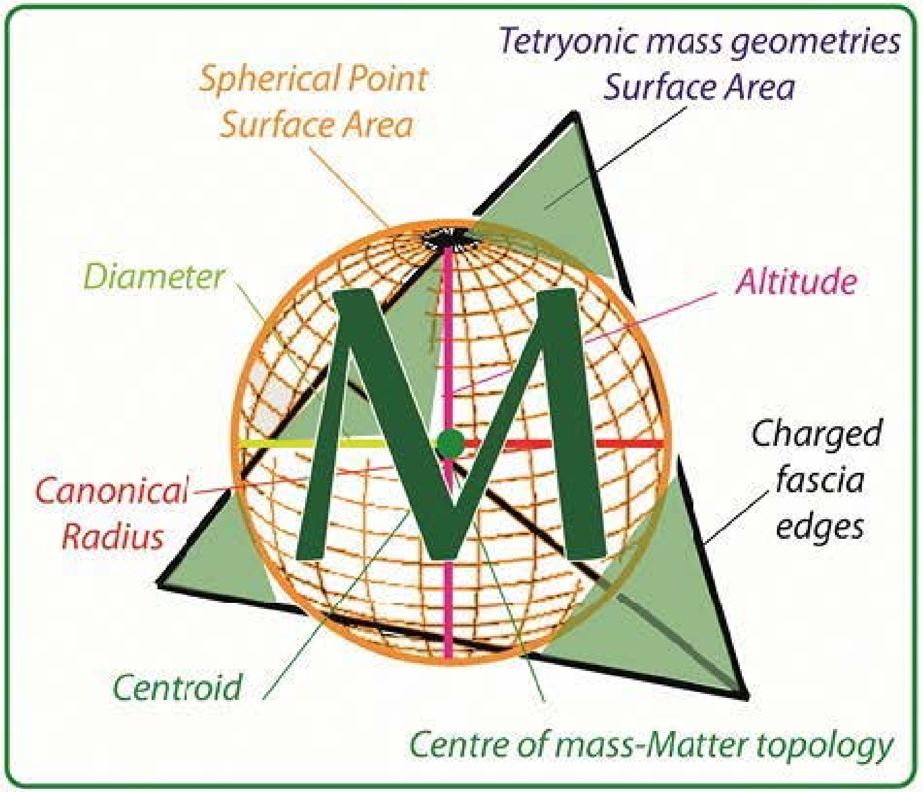
Tetryonic mass geometries
Surface Area
Spherical Point
Surface Area
Diameter
Altitude
Charged
fascia
edges
Canonical
Radius
M
Centroid
Centre of mass-Matter topology

Spheres & Tetrahedra are both 3D Platonic solids with 4π scalar integral Gaussian topologies and physical displacement volumes

[Gauss-Bonnet theorem]

Sphere

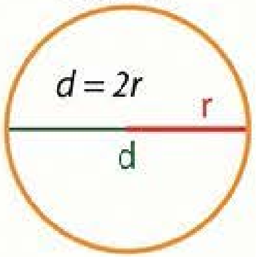

Area sphere= $4\pi r^2$

Surface curvature = $1/r^2$

Total curvature = 4π

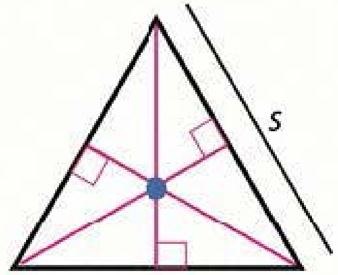

area = $\sqrt{3/4} * a^2$

ZPF fascia

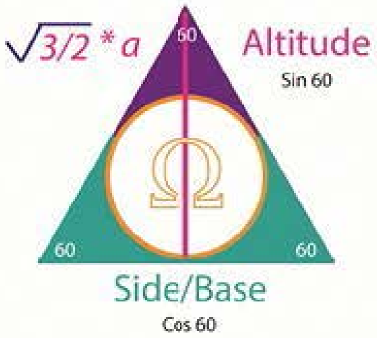
$\sqrt{3/2} * a$
60
Altitude
Sin 60
Ω
60
60
Side/Base
Cos 60

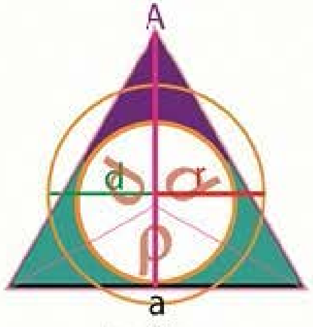

Area Tetryon

$$A = 4\,A_0 = \sqrt{3}a^2$$

Vertex curvature = π

Total curvature = 4π

Copyright ABRAHAM [2008] - all rights reserved

Dodecyons

while quarks & leptons are comprised of

12π mass-energy geometries

they have differing final Matter topologies

+1/3

up quark

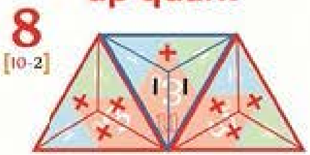

anti-strange quark

+2/3

top quark

charmed quark

8π

strange quark

-1/3

bottom quark

anti-bottom quark

-2/3

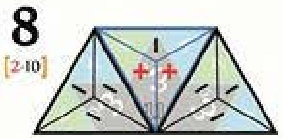

anti-up quark

anti-down quark

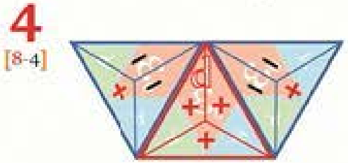

anti-charmed quark

quarks

anti-top quark

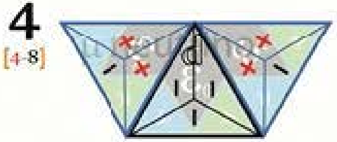

down quark

electron neutrino

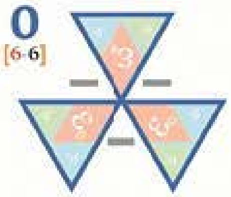

neutrinos

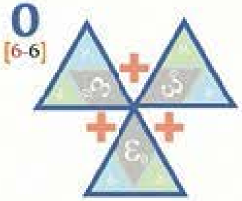

positron neutrino

positron

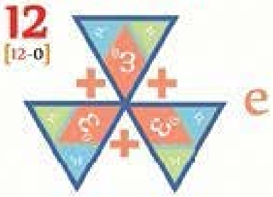

+1

anti-τ neutrino

anti-μ neutrino

tau

12π

0

muon

muon neutrino

tau neutrino

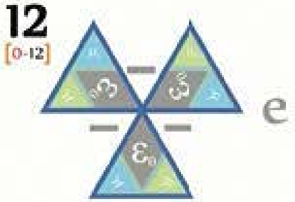

-1

electron

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

ElectroMagnetic mass velocity

Quarks have octahedral topologies

Leptons have dodecahedral topologies

Copyright ABRAHAM [2008] - all rights reserved

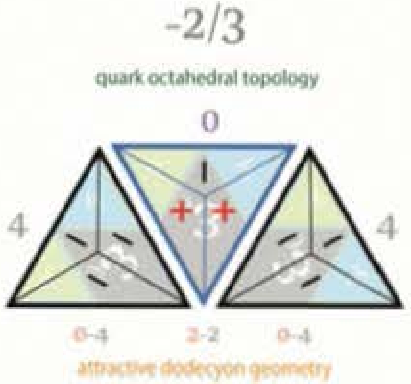
-2/3
quark octahedral topology
0
4
4
0-4
2-2
0-4
attractive dodecyon geometry

anti-up quark

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

EM Field Planck quanta

$$12\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

dodecyon ElectroMagnetic mass velocity

1. **anti-up** quark
2. **anti-charmed** quark
3. **anti-top** quark

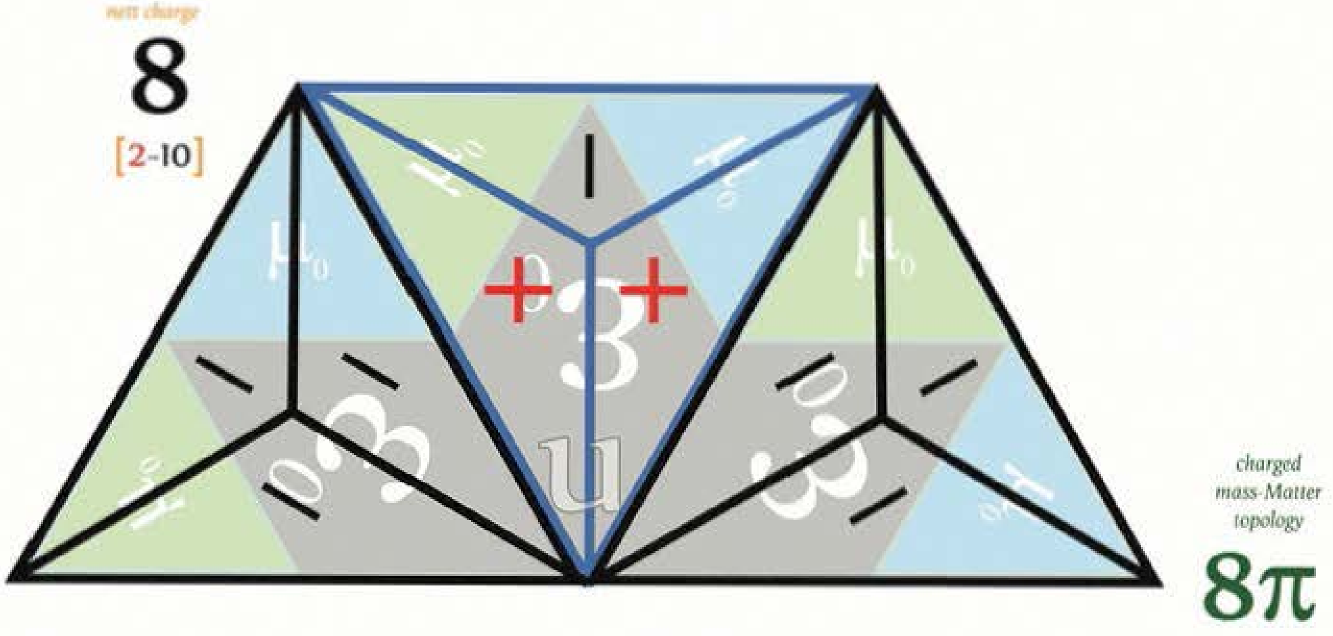
nett charge
8
[2-10]
μ0
μ0
μ0
μ0
μ0
μ0
ε0
ε0
ε0
3
3
3
u
charged
mass-Matter
topology
8π

Quarks have 12 charged fascia - mass-energy geometries

2 Positive 10 Negative

Quarks have octahedral mass-Matter topologies

[-2/3 elementary charge]

down quark

-1/3

quark octahedral topology

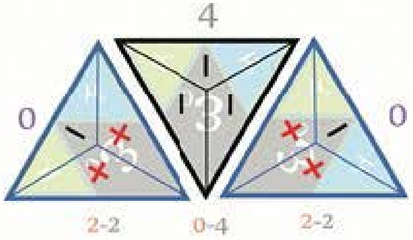

attractive dodecyon geometry

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

docdecyon ElectroMagnetic mass velocity

as mass-energy momenta is added to Matter topologies & KEM fields of motion different generations of particles are created

1. **down** quark
2. **strange** quark
3. **bottom** quark

d

nett charge

4

[4-8]

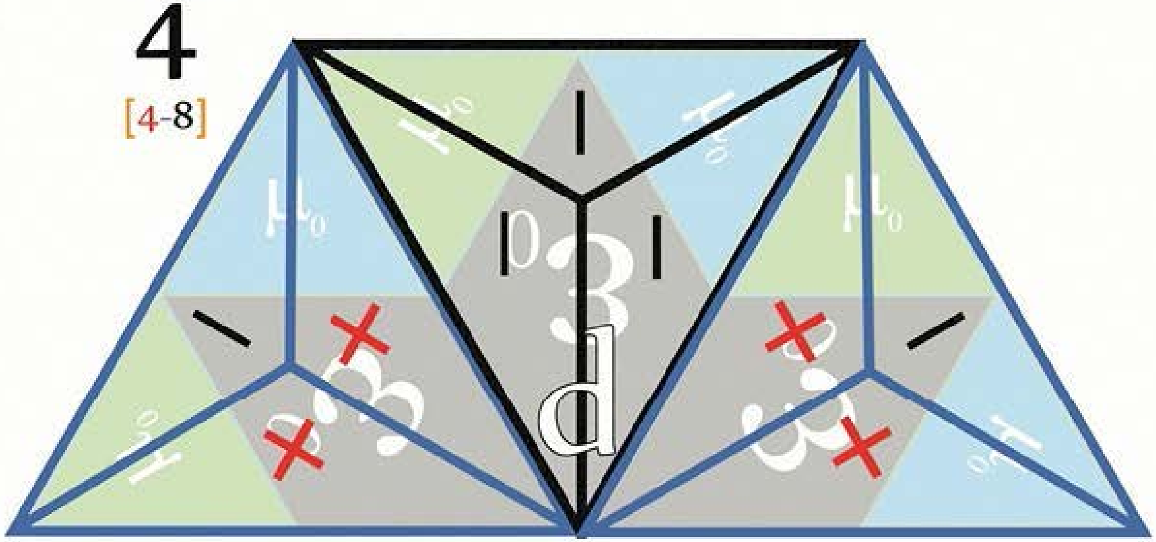

charged mass-Matter topology

8π

Quarks have 12 charged fascia - *mass-energy geometries*

4 Positive 8 Negative

Quarks have octahedral mass-Matter topologies

[-1/3 elementary charge]

Copyright ABRAHAM [2008] - all rights reserved

anti-down quark

+1/3

quark octahedral topology

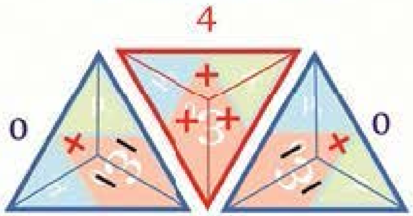

2-2 4-0 2-2

attractive dodecyon geometry

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

dodecyon

EM Field

ElectroMagnetic

Planck quanta

mass

velocity

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

1. **anti-down** quark
2. **anti-strange** quark
3. **anti-bottom** quark

d

nett charge

4

[8-4]

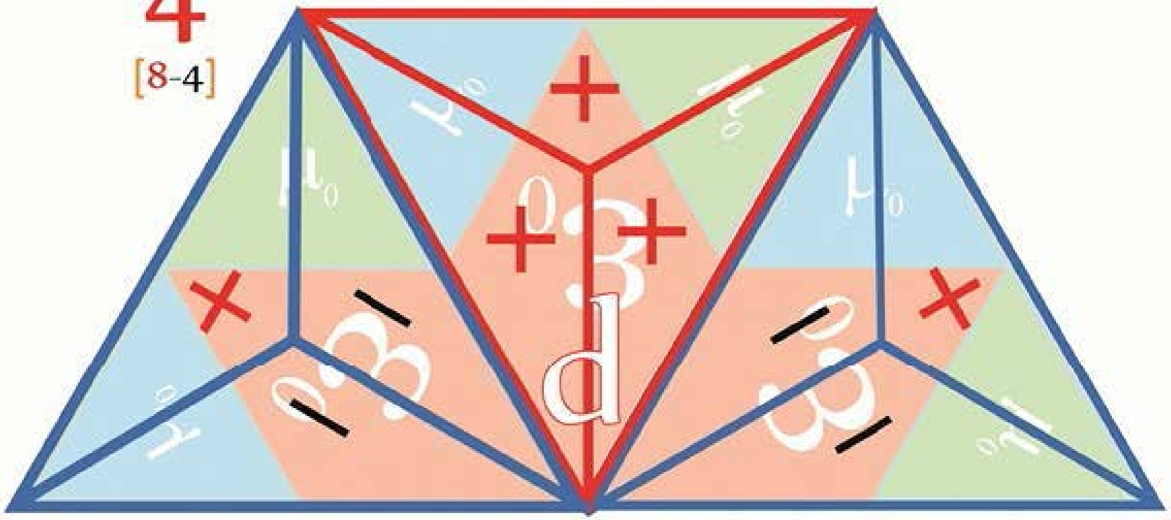

charged
mass-Matter
topology

8π

Quarks have 12 charged fascia - ***mass-energy geometries***

8 Positive 4 Negative

Quarks have octahedral mass-Matter topologies

[+1/3 elementary charge]

up quark

+2/3

quark octahedral topology

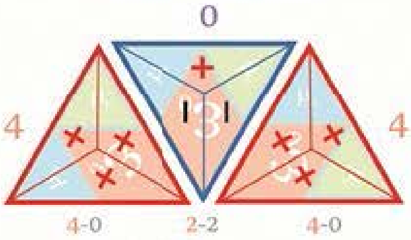

attractive dodecyon geometry

EM Field Planck quanta

$$12\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

dodecyon ElectroMagnetic mass velocity

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

1. **up quark**
2. **charmed** quark
3. **top** quark

u

nett charge

8

[10-2]

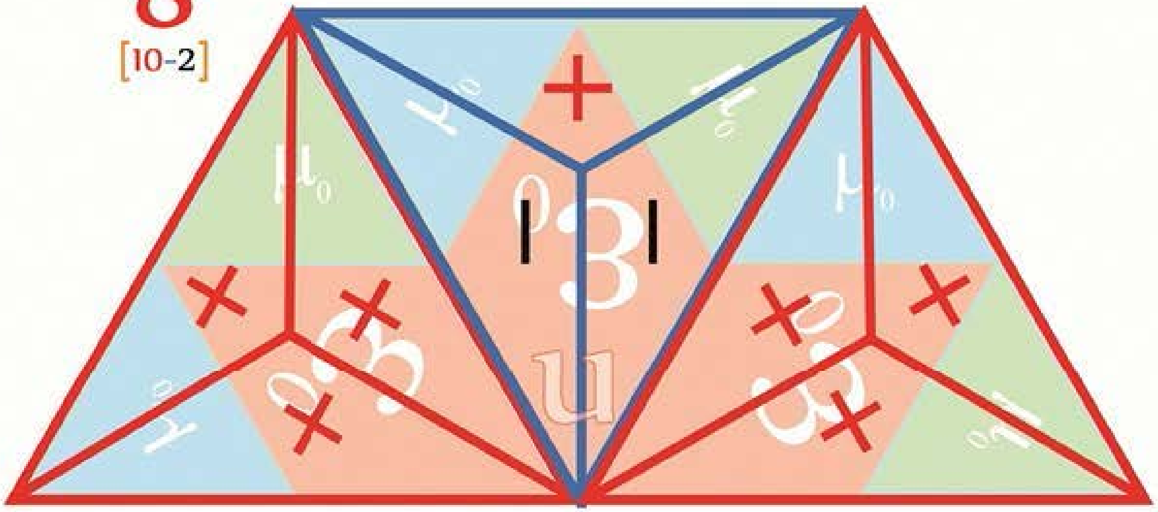

charged
mass-Matter
topology

8π

Quarks have 12 charged fascia - mass-energy geometries

10 Positive 2 Negative

Quarks have octahedral mass-Matter topologies

[+2/3 elementary charge]

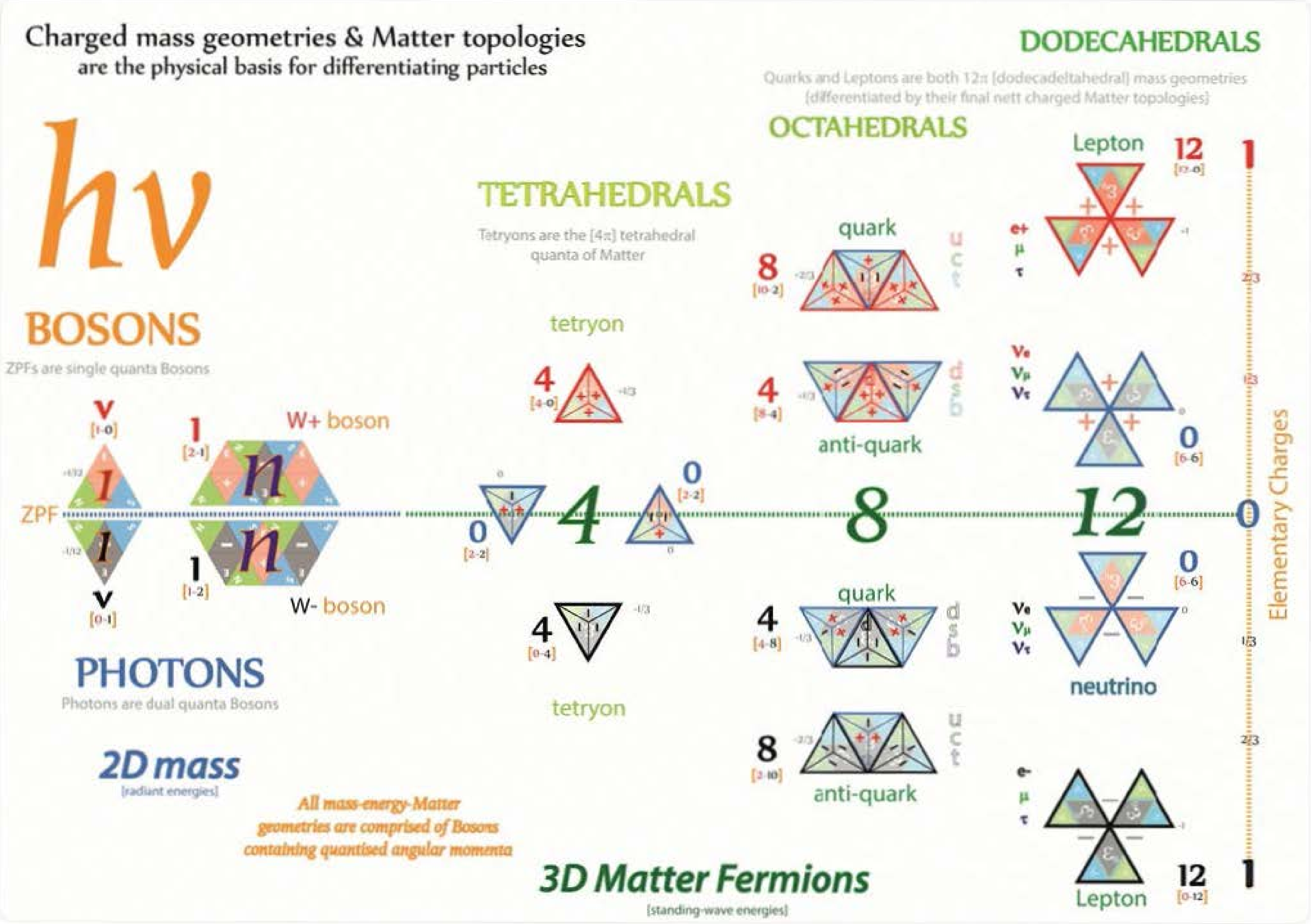

Charged mass geometries & Matter topologies
are the physical basis for differentiating particles

hv

BOSONS
ZPFs are single quanta Bosons

PHOTONS
Photons are dual quanta Bosons

2D mass
[radiant energies]

All mass-energy-Matter geometries are comprised of Bosons containing quantised angular momenta

TETRAHEDRALS
Tetryons are the [4π] tetrahedral quanta of Matter

OCTAHEDRALS

DODECAHEDRALS
Quarks and Leptons are both 12π [dodecadeltahedral] mass geometries (differentiated by their final nett charged Matter topologies)

3D Matter Fermions
[standing-wave energies]

v
[1-0]
+1/12
ZPF
-1/12
v
[0-1]
1
[2-1]
W+ boson
1
[1-2]
W- boson
tetryon
4
[4-0]
+1/3
0
0
[2-2]
4
0
[2-2]
0
4
[0-4]
-1/3
tetryon
quark
8
[10-2]
+2/3
u c t
4
[8-4]
+1/3
d s b
anti-quark
8
quark
4
[4-8]
-1/3
d s b
8
[2-10]
-2/3
u c t
anti-quark
Lepton
12
[12-0]
e+
μ
τ
+1
νe
νμ
ντ
0
0
[6-6]
12
0
[6-6]
νe
νμ
ντ
0
neutrino
e-
μ
τ
-1
Lepton
12
[0-12]
1
2/3
1/3
0
1/3
2/3
1
Elementary Charges

divergent negative E-fields

electron

convergent negative E-fields

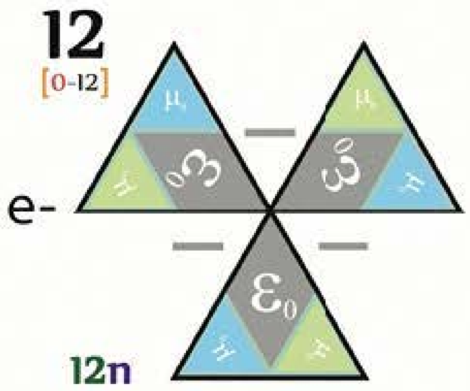
12
[0-12]
e-
12n
μ_0
ε_0
–

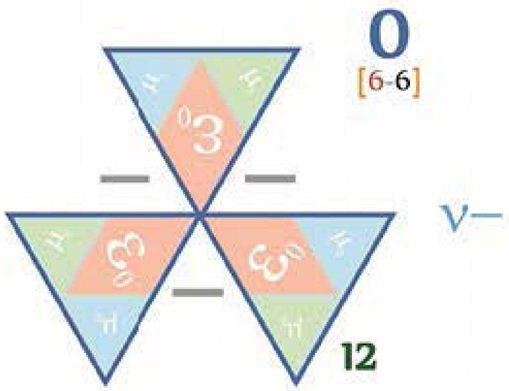
0
[6-6]
ν–
12
μ_0
ε_0
–

divergent negative E-fields

electron neutrino

convergent positive E-fields

Leptons have 12 charged mass-energy fascia geometries

Repulsive STRONG force creates Lepton topologies

12π

Tetryonic charged fascias creates the Matter topologies of all Leptons

Repulsive STRONG force creates Neutrino topologies

Leptons have dodecadeltahedral mass-Matter topologies

divergent positive E-fields

positron

convergent positive E-fields

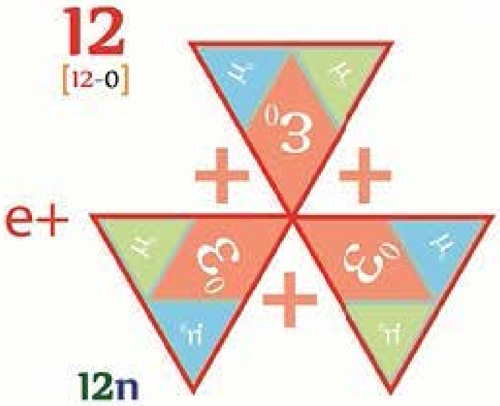
12
[12-0]
e+
12n
μ_0
ε_0
+

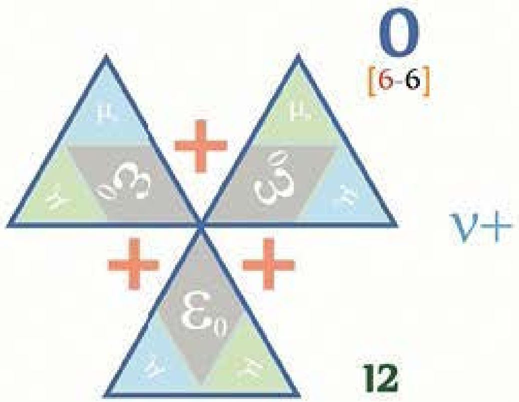
0
[6-6]
ν+
12
μ_0
ε_0
+

divergent positive E-fields

positron neutrino

convergent negative E-fields

Charged Leptons

EM Field Planck quanta

$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

Dodecyon ElectroMagnetic mass velocity

positron

e+

12

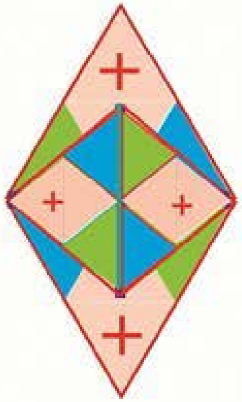

net charge

12

[12-0]

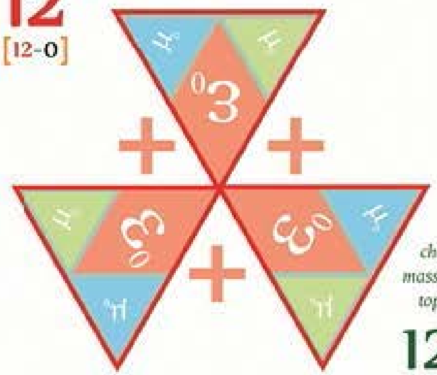

charged mass-Matter topology

12π

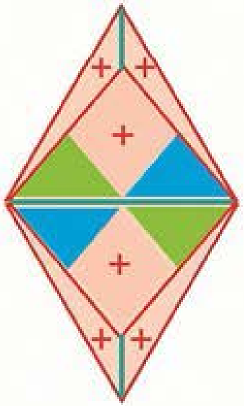

1. positron
2. anti-muon
3. anti-tau

Leptons have 12 charged mass-energy fascia geometries

12 positive charge fascia or 12 Negative charge fascia

repulsive dodecyon geometries

lepton topologies

Charged Leptons have dodecadeltahedral mass-Matter topologies

[1 elementary charge]

e-

12

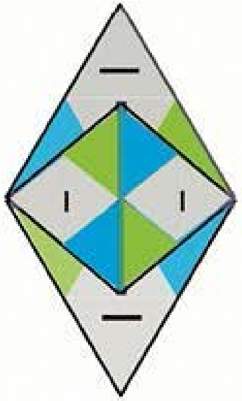

net charge

12

[0-12]

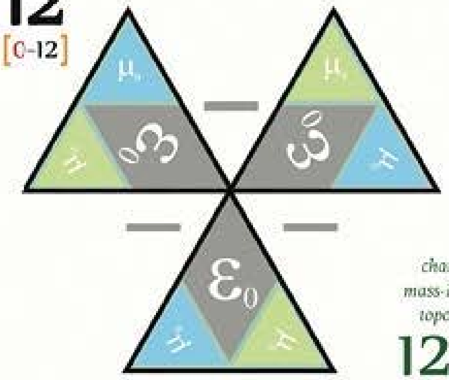

charged mass-Matter topology

12π

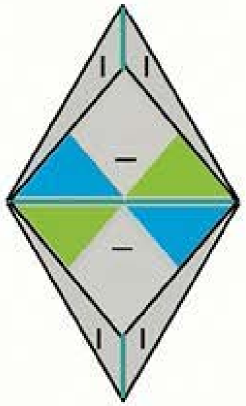

1. electron
2. muon
3. tau

as mass-energy momenta is added to Matter topologies & KEM fields of motion different generations of particles are created

electron

Copyright ABRAHAM [2008] - all rights reserved

positron

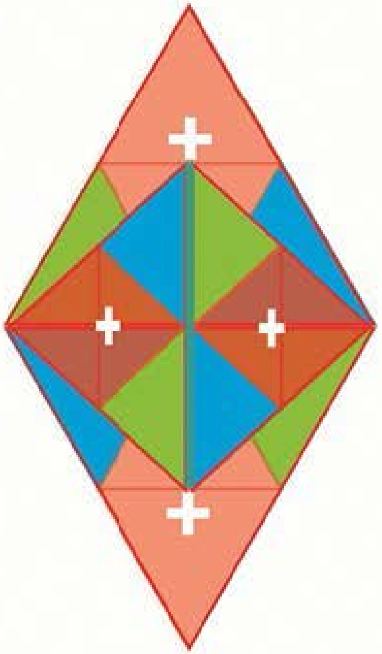

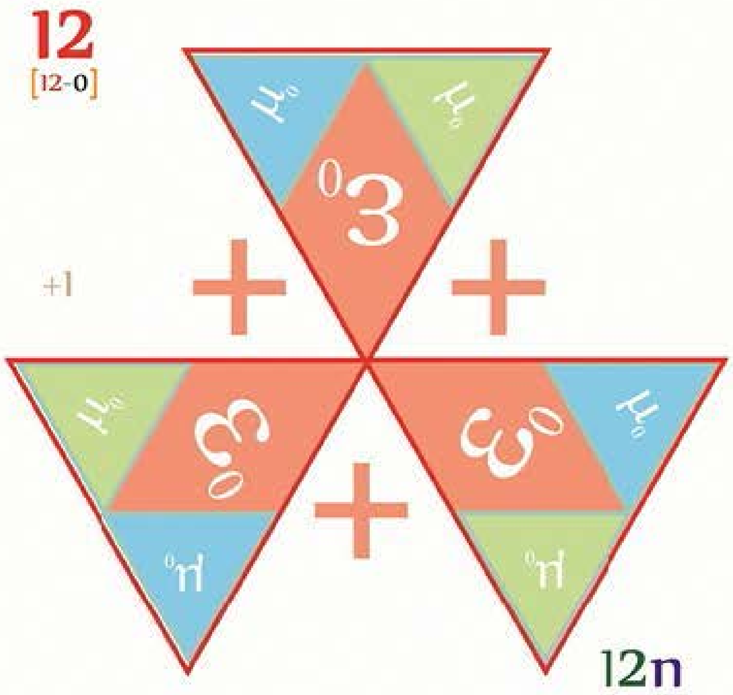
12
[12-0]
+1
μ_0
ε_0
12n

repulsive charge dodecyon geometry

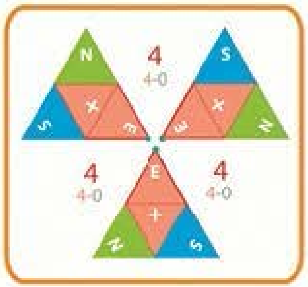

lepton mass-Matter topology

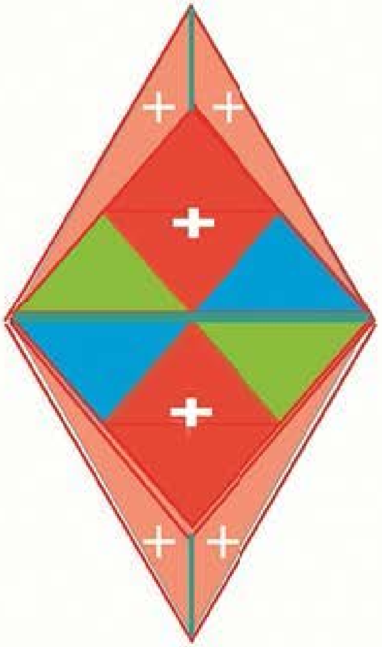

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

1. **positron**
2. **anti-muon**
3. **anti-tau**

Neutral Leptons

EM Field Planck quanta

$12\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Dodecyon ElectroMagnetic mass velocity

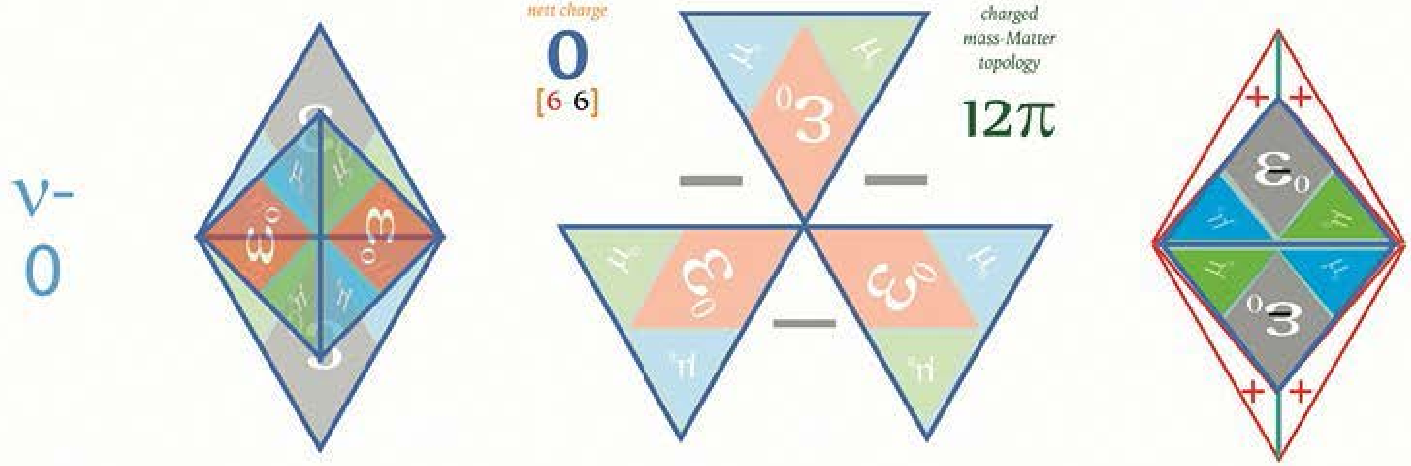
ν-
0
nett charge
0
[6 6]
charged
mass-Matter
topology
12π
ε0
μ0
ε0
μ0
ε0
μ0
+
+
+
+

1. electron neutrino
2. muon neutrino
3. tau neutrino

Neutrinos have 12 charged mass-energy fascia geometries

6 Positive & 6 Negative

repulsive dodecyon geometries

lepton topologies

Neutrinos have neutral dodecadeltahedral mass-Matter topologies

[0 elementary charge]

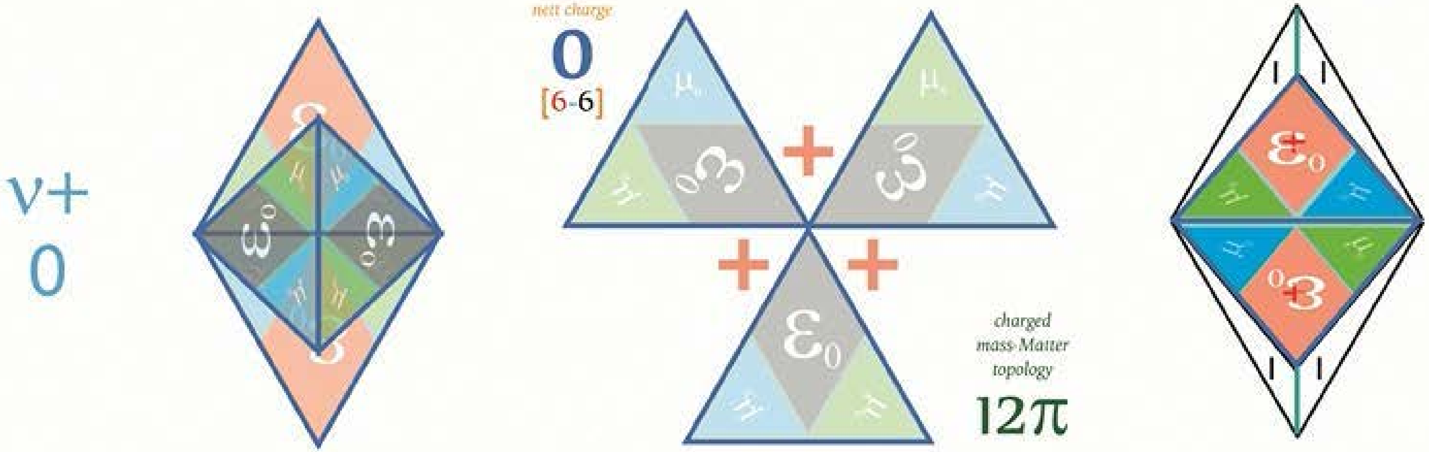
ν+
0
nett charge
0
[6-6]
+
+
+
charged
mass-Matter
topology
12π
ε0
μ0
ε0
μ0
ε0
μ0

1. positron neutrino
2. anti- muon neutrino
3. anti-tau neutrino

as mass-energy momenta is added to
Matter topologies & KEM fields of motion
different generations of particles are created

electron

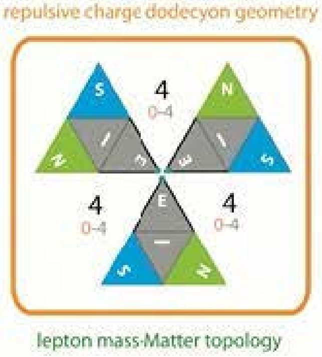
repulsive charge dodecyon geometry

lepton mass-Matter topology

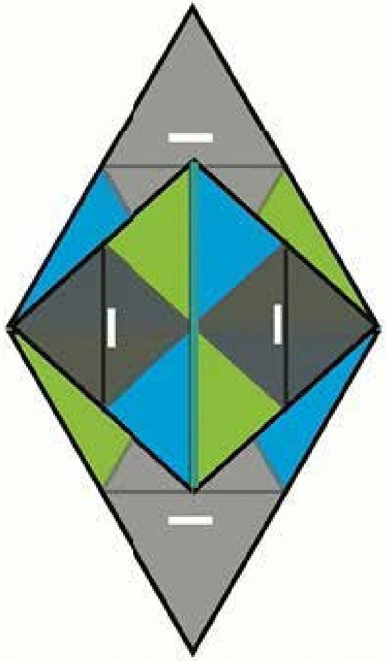

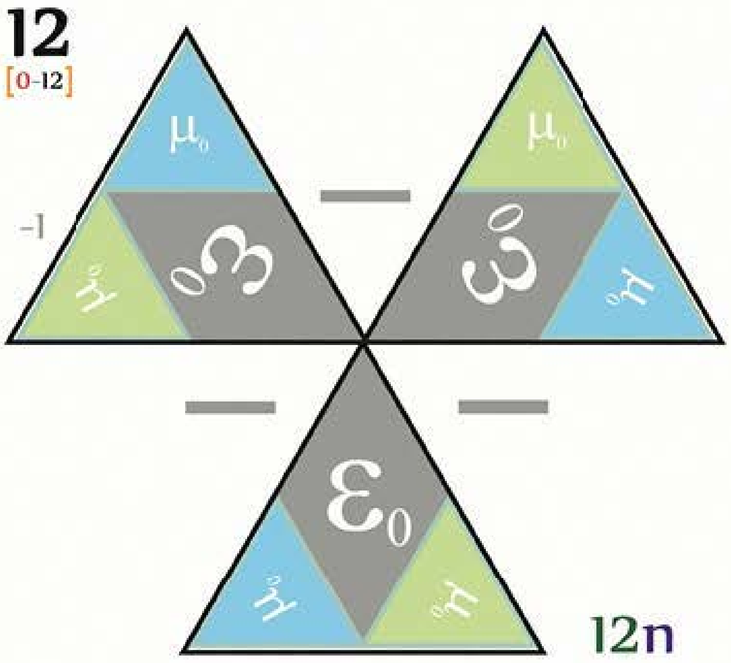
12
[0-12]
-1
μ_0
ε_0
12n

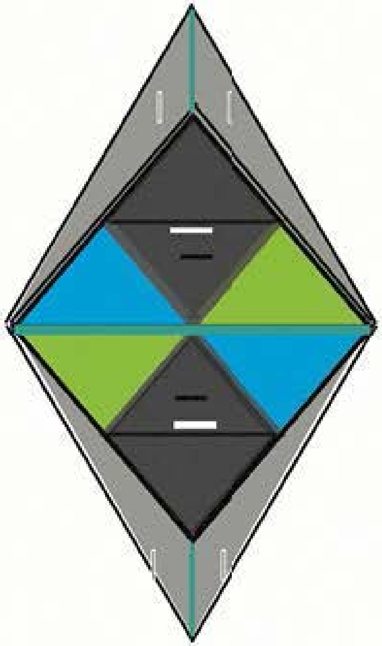

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

1. **electron**
2. **muon**
3. **tau**

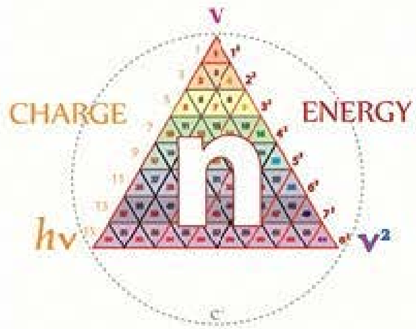
v
CHARGE
ENERGY
n
hv
v²
c

Leptronic Quantum levels

[KEM field mass-energy geometry]

v^2 Total quanta		n Electron quantum level
1.2e20		1
4.8e20		4
1.08e21		9
1.92e21	1e19	16
3.00e21		25
4.32e21		36
5.88e21		49
7.68e21		64

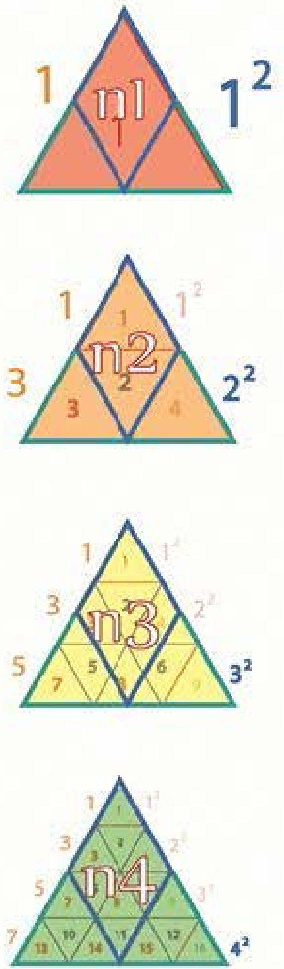
1
n1
1^2
1
n2
1^2
3
2^2
1
n3
1^2
3
2^2
5
3^2
1
n4
1^2
3
2^2
5
3^2
7
4^2

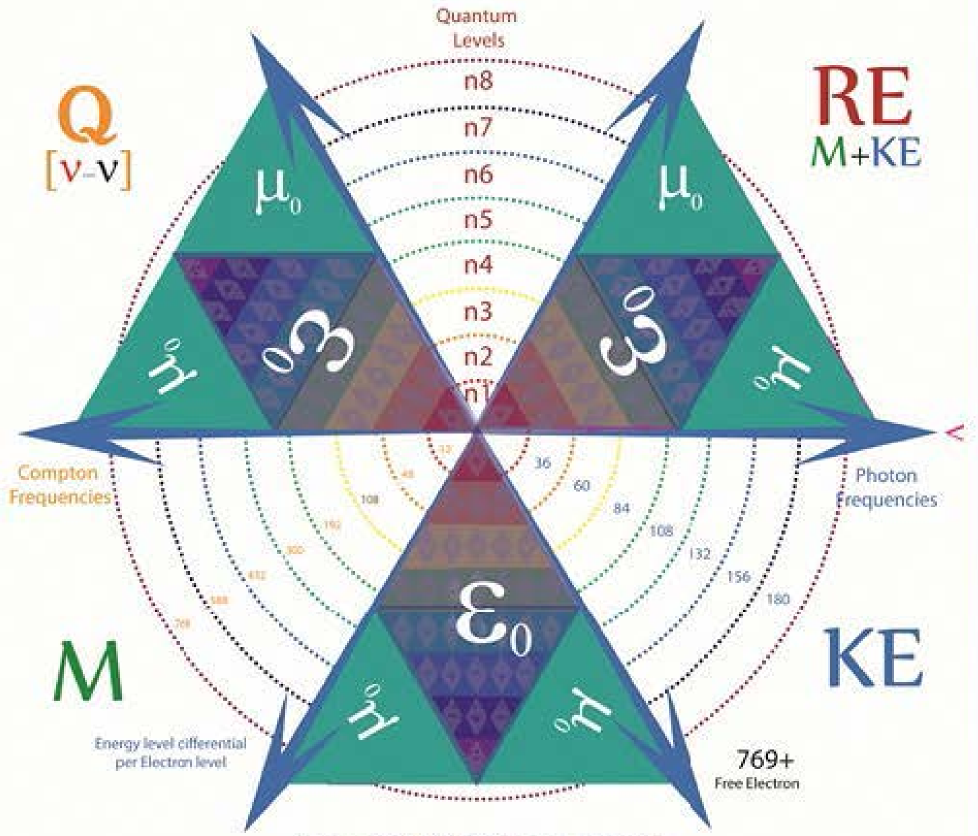
Quantum Levels
n8
n7
n6
n5
n4
n3
n2
n1
Q
[v-v]
RE
M+KE
μ_0
ε_0
v
Compton Frequencies
Photon Frequencies
12
48
108
192
300
432
588
768
36
60
84
108
132
156
180
M
KE
Energy level differential per Electron level
769+
Free Electron

Leptronic KEM field mass-energies

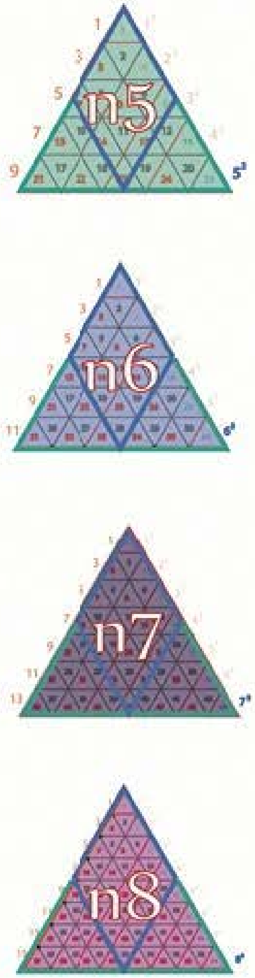
n5
5^2
n6
6^2
n7
7^2
n8
8^2

$$p^2 = KEM = Mv^2$$

Illustrative schema only: All Leptonic quantum levels have the same equilateral KEM geometry as the compton frequency of the KEM field increases the wavelengths of the quanta decrease

Copyright ABRAHAM [2008] - all rights reserved

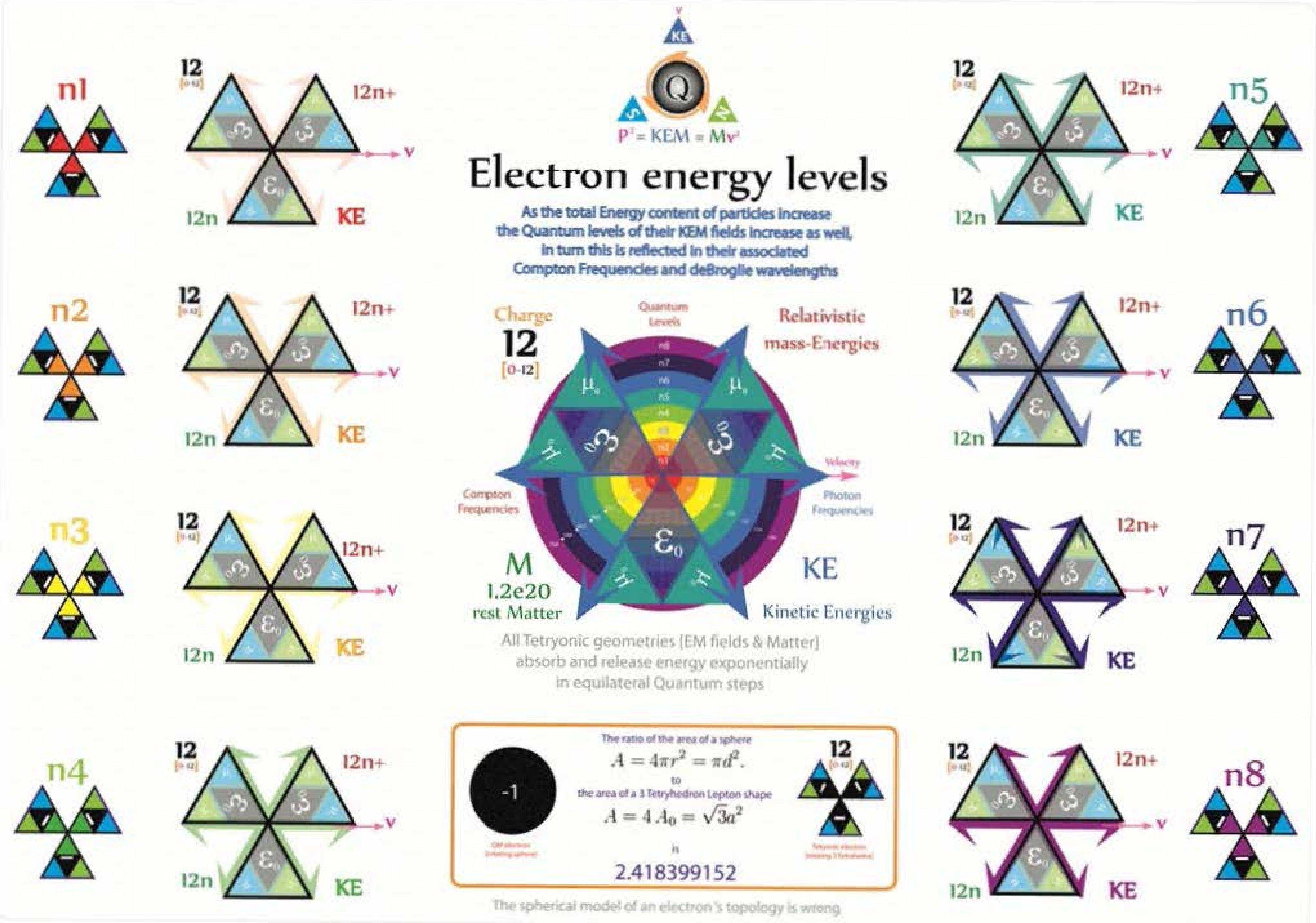
v
KE
Q
S
N
n1
n2
n3
n4
n5
n6
n7
n8
12
[0-12]
12n+
12n
KE
v
Charge
12
[0-12]
Quantum Levels
n8
n7
n6
n5
n4
n3
n2
n1
Relativistic mass-Energies
Velocity
Photon Frequencies
Compton Frequencies
M
1.2e20
rest Matter
KE
Kinetic Energies
-1

P^2 = KEM = Mv^2

Electron energy levels

As the total Energy content of particles increase the Quantum levels of their KEM fields increase as well, in turn this is reflected in their associated Compton Frequencies and deBroglie wavelengths

All Tetryonic geometries [EM fields & Matter] absorb and release energy exponentially in equilateral Quantum steps

The ratio of the area of a sphere

$$A = 4\pi r^2 = \pi d^2.$$

to

the area of a 3 Tetryhedron Lepton shape

$$A = 4\,A_0 = \sqrt{3}a^2$$

is

2.418399152

The spherical model of an electron 's topology is wrong

Tetryonics 08.07 - Electron energy levels

Leptronic Kinetic EM Fields

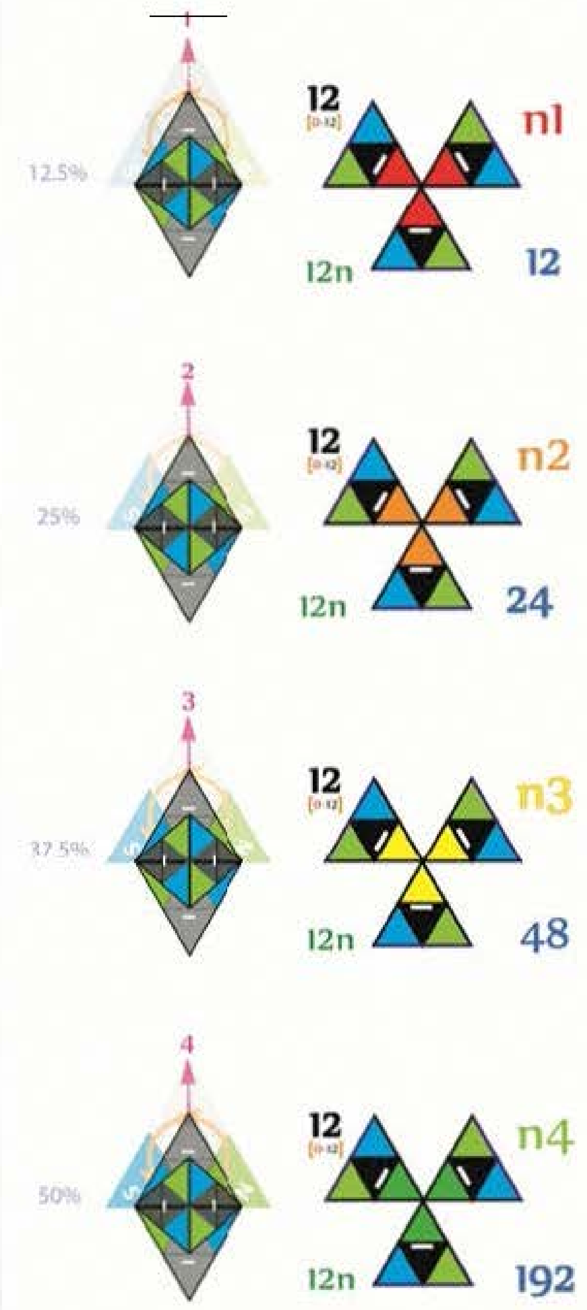
1
12.5%
12
[0-12]
n1
12n
12
2
25%
12
[0-12]
n2
12n
24
3
37.5%
12
[0-12]
n3
12n
48
4
50%
12
[0-12]
n4
12n
192

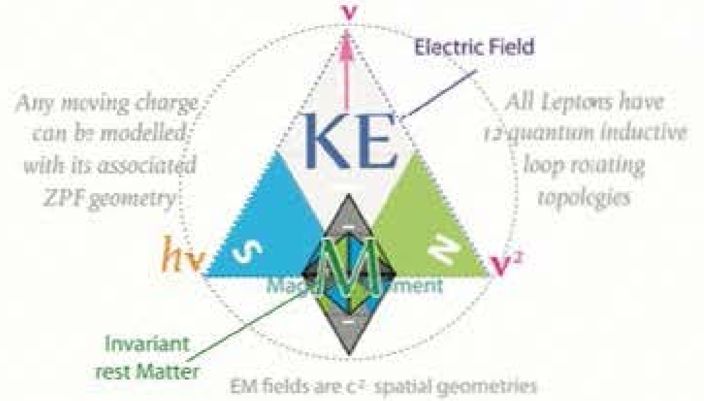
v
Electric Field
Any moving charge can be modelled with its associated ZPF geometry
All Leptons have 12 quantum inductive loop rotating topologies
KE
hv
S
N
M
v²
Magnetic Moment
Invariant rest Matter
EM fields are c² spatial geometries

$$P^2 = KEM = Mv^2$$

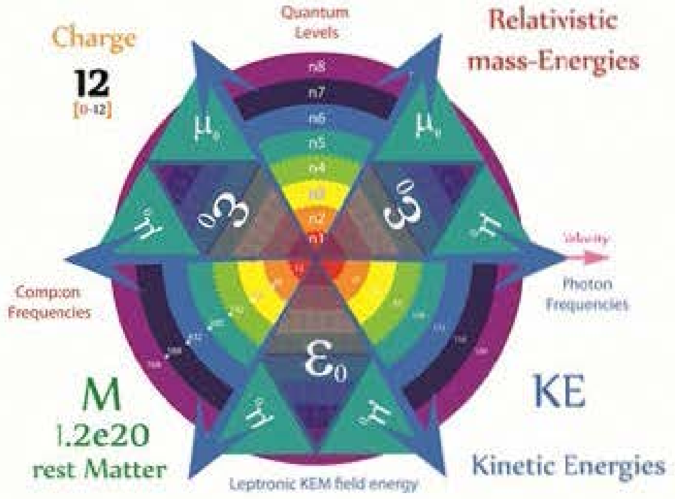
Charge
12
[0-12]
Quantum Levels
Relativistic mass-Energies
n8
n7
n6
n5
n4
n3
n2
n1
μ₀
ε₀
Velocity
Compton Frequencies
Photon Frequencies
M
1.2e20
rest Matter
KE
Kinetic Energies
Leptronic KEM field energy

The motion of an Electron's Matter topology through any external EM fields creates a proportional KE field and Magnetic Moment as a direct result of the 'quantum inductive loops' formed by the Leptronic EM field geometries

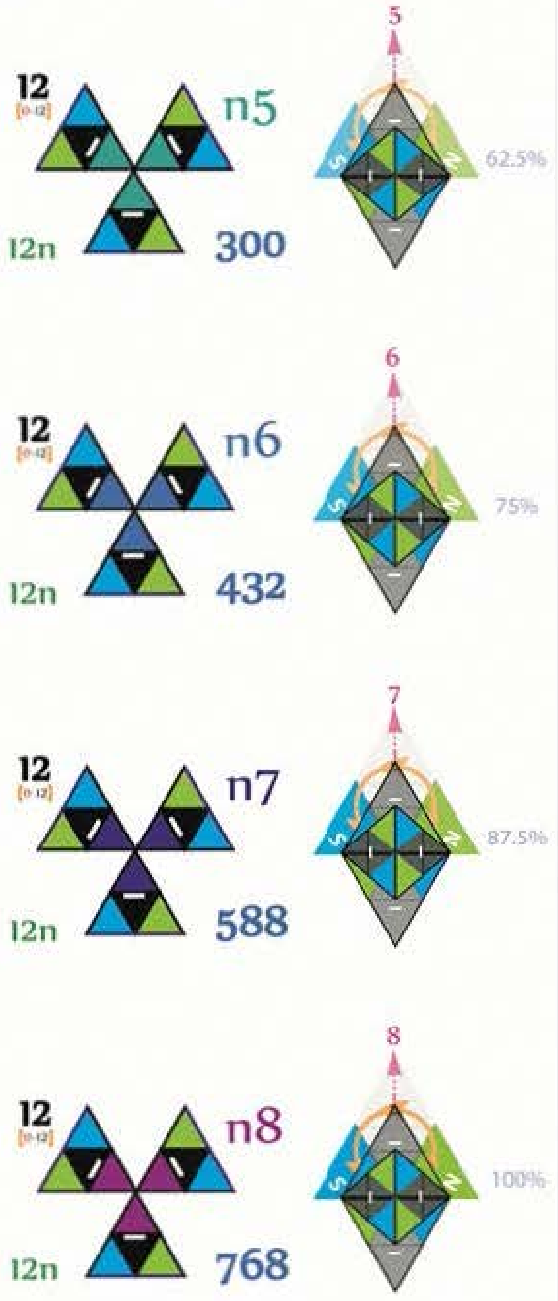
5
12
[0-12]
n5
12n
300
62.5%
6
12
[0-12]
n6
12n
432
75%
7
12
[0-12]
n7
12n
588
87.5%
8
12
[0-12]
n8
12n
768
100%

Tetryonics 08.08 - Leptronic KEM fields

Kinetic field modelling

RE

Mv^2

2D

Kinetic Energy field

$E = mc^2$

3D

rest mass-Matter topology

Tetryonic Positive Charge KEM field geometry

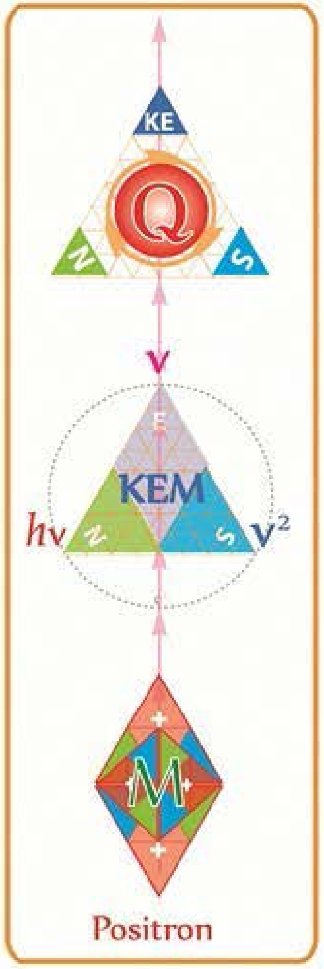

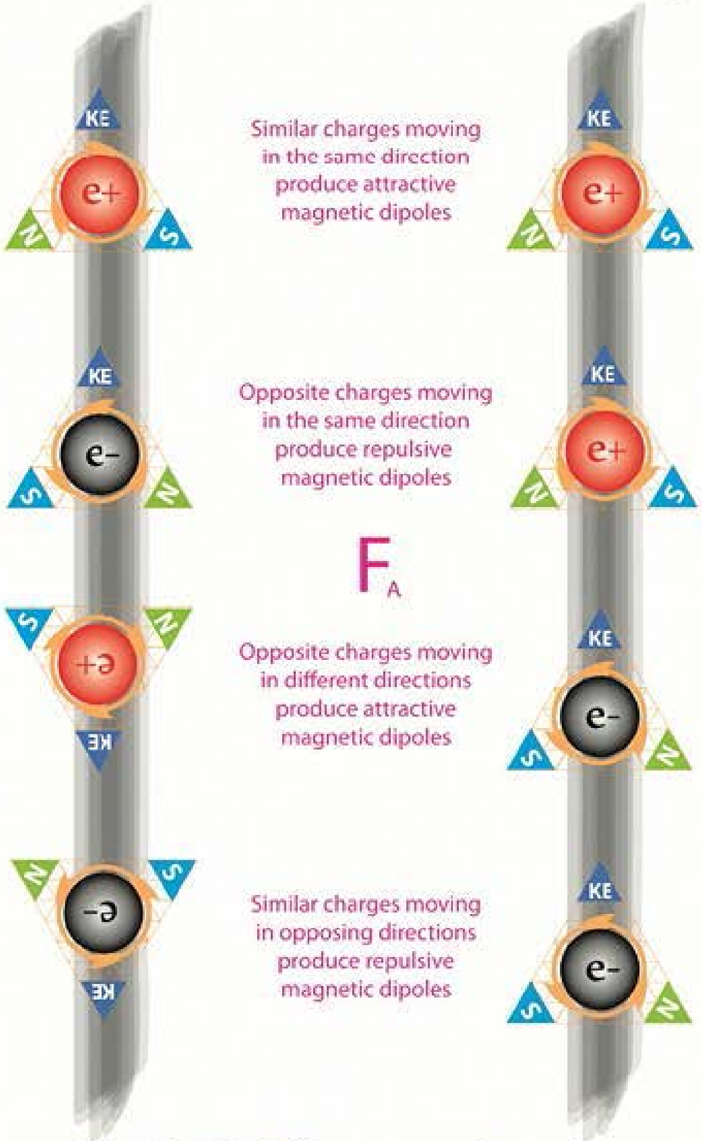

Similar charges moving in the same direction produce attractive magnetic dipoles

Opposite charges moving in the same direction produce repulsive magnetic dipoles

F_A

Opposite charges moving in different directions produce attractive magnetic dipoles

Similar charges moving in opposing directions produce repulsive magnetic dipoles

Tetryonic Negative Charge KEM field geometry

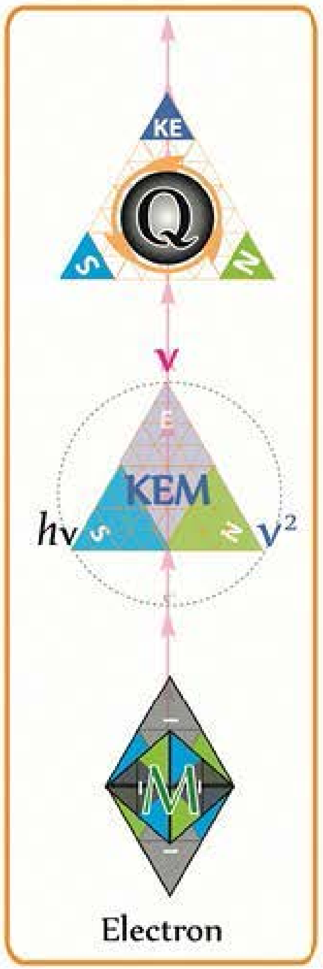

RE

Mv^2

2D

Kinetic Energy field

$E = mc^2$

3D

rest mass-Matter topology

The KEM fields of all Matter particles in motion can be modelled using ZPF EM geometries

Kinetic EM field geometry determines the Magnetic Moments of all particles in motion

The KEM E-field is reflective of the nett particle charge and forms a directional vector of motion

Lepton Families & Generations

v
KE
M
hv
v²
rest Matter is invariant
c

Charged Matter topologies in motion creates Kinetic EM field geometries

Charged mass-energy geometries form elementary particle topologies

$$12\pi$$

$$KEM = Mv^2$$

Total relativistic mass-energies [rest Matter + KEM]

All high mass-energy particles 'decay' into low mass-energy particles by emitting photons

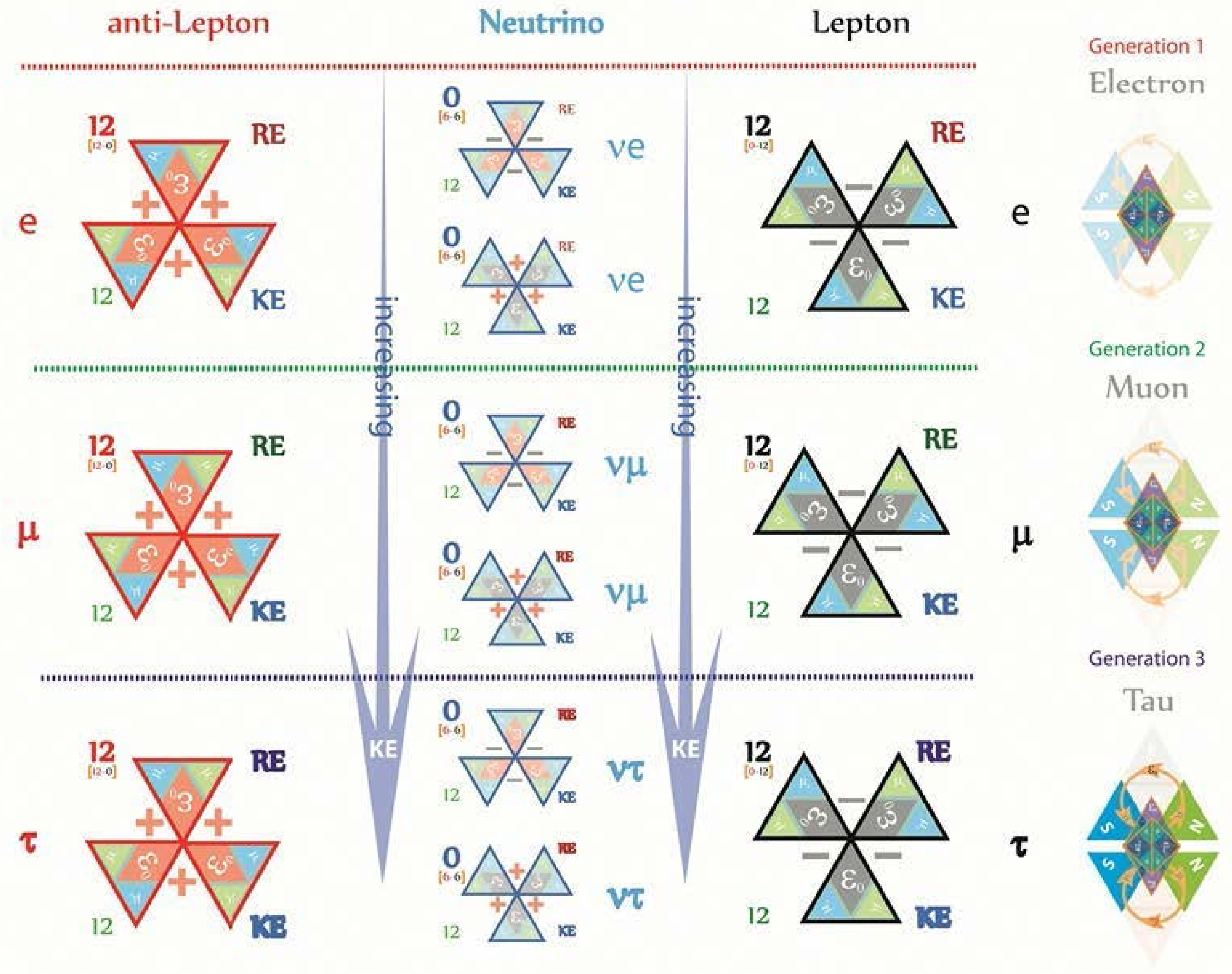
anti-Lepton
Neutrino
Lepton
Generation 1
Electron
e
12
[12-0]
RE
12
KE
0
[6-6]
RE
νe
12
KE
0
[6-6]
RE
νe
12
KE
12
[0-12]
RE
12
KE
e
increasing
increasing
Generation 2
Muon
μ
12
[12-0]
RE
12
KE
0
[6-6]
RE
νμ
12
KE
0
[6-6]
RE
νμ
12
KE
12
[0-12]
RE
12
KE
μ
Generation 3
Tau
τ
12
[12-0]
RE
12
KE
KE
0
[6-6]
RE
ντ
12
KE
0
[6-6]
RE
ντ
12
KE
KE
12
[0-12]
RE
12
KE
τ

Note: All Lepton geometries are the same size - only the energy density of the EM field changes

Tetryonics 08.10 - Lepton Families & Generations

Lepton Helicity and Chirality

Particle Spin and Handedness is always referenced
with respect to the direction of nett total system linear Momentum

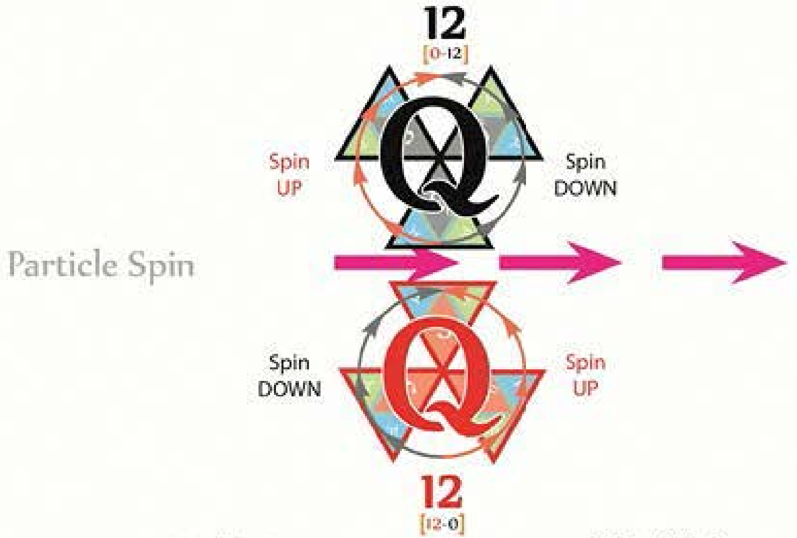
12
[0-12]
Spin
UP
Spin
DOWN
Particle Spin
Spin
DOWN
Spin
UP
12
[12-0]

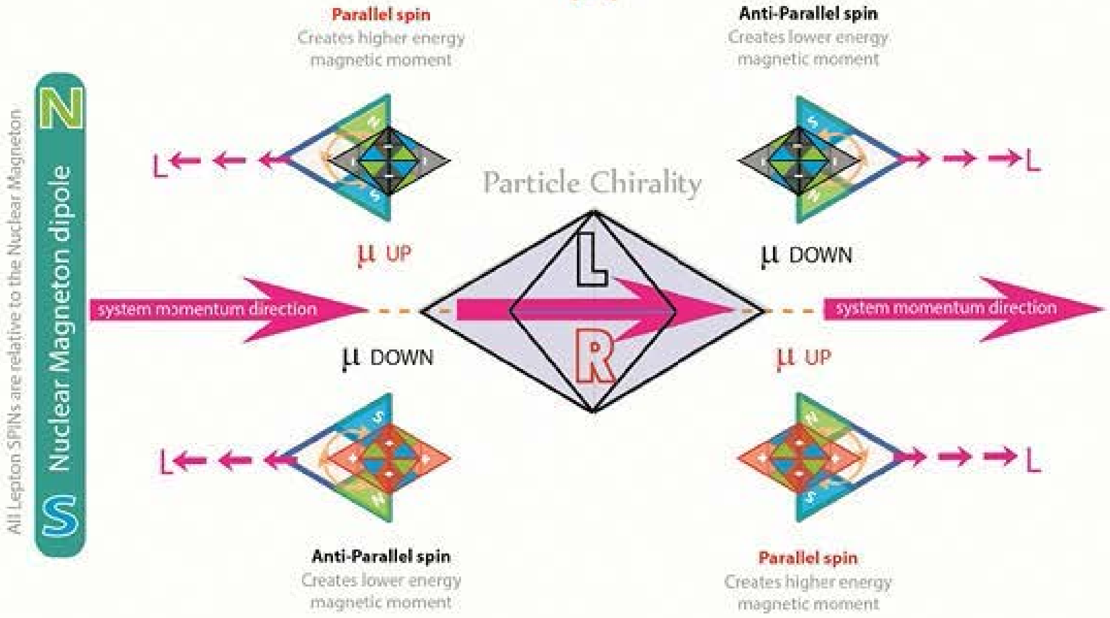
Parallel spin
Creates higher energy
magnetic moment
Anti-Parallel spin
Creates lower energy
magnetic moment
All Lepton SPINs are relative to the Nuclear Magneton
N
Nuclear Magneton dipole
S
L
L
µ UP
Particle Chirality
µ DOWN
system momentum direction
L
R
system momentum direction
µ DOWN
µ UP
L
L
Anti-Parallel spin
Creates lower energy
magnetic moment
Parallel spin
Creates higher energy
magnetic moment

Particle Chirality

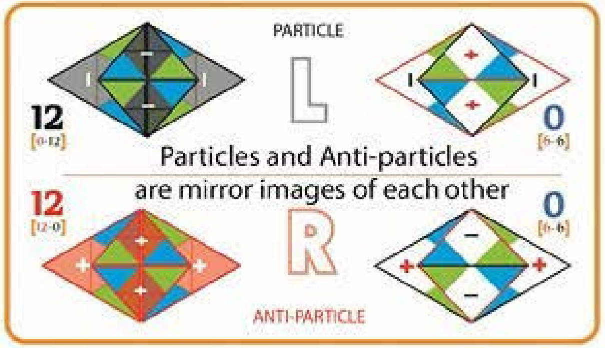
PARTICLE
12
[0-12]
L
0
[6-6]
12
[12-0]
R
0
[6-6]
ANTI-PARTICLE

Particles and Anti-particles
are mirror images of each other

Particle Helicity

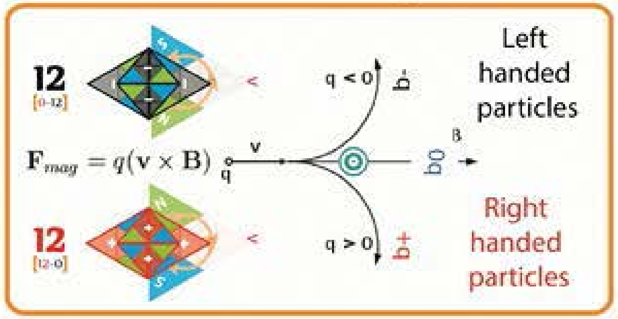
12
[0-12]
Left
handed
particles
q < 0
q-
12
[12-0]
q > 0
q+
Right
handed
particles
v
q
B

$$\mathbf{F}_{mag} = q(\mathbf{v} \times \mathbf{B})$$

Lorentz Force and Transformation

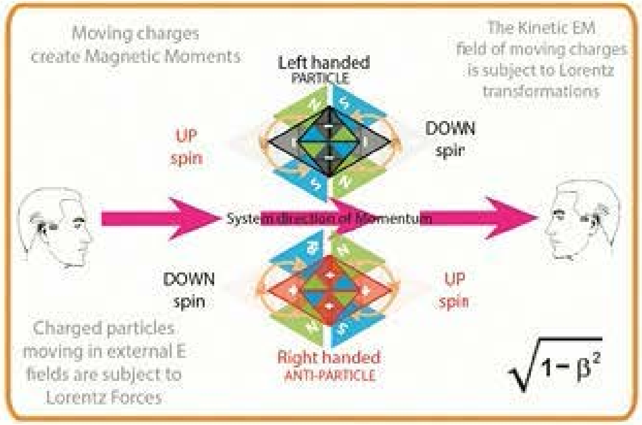
Left handed
PARTICLE
UP
spin
DOWN
spir
System direction of Momentum
DOWN
spin
UP
spin
Right handed
ANTI-PARTICLE

Moving charges
create Magnetic Moments

The Kinetic EM
field of moving charges
is subject to Lorentz
transformations

Charged particles
moving in external E
fields are subject to
Lorentz Forces

$$\sqrt{1-\beta^2}$$

'Point Charges'

The only true 'point charges' are Zero Point Fields

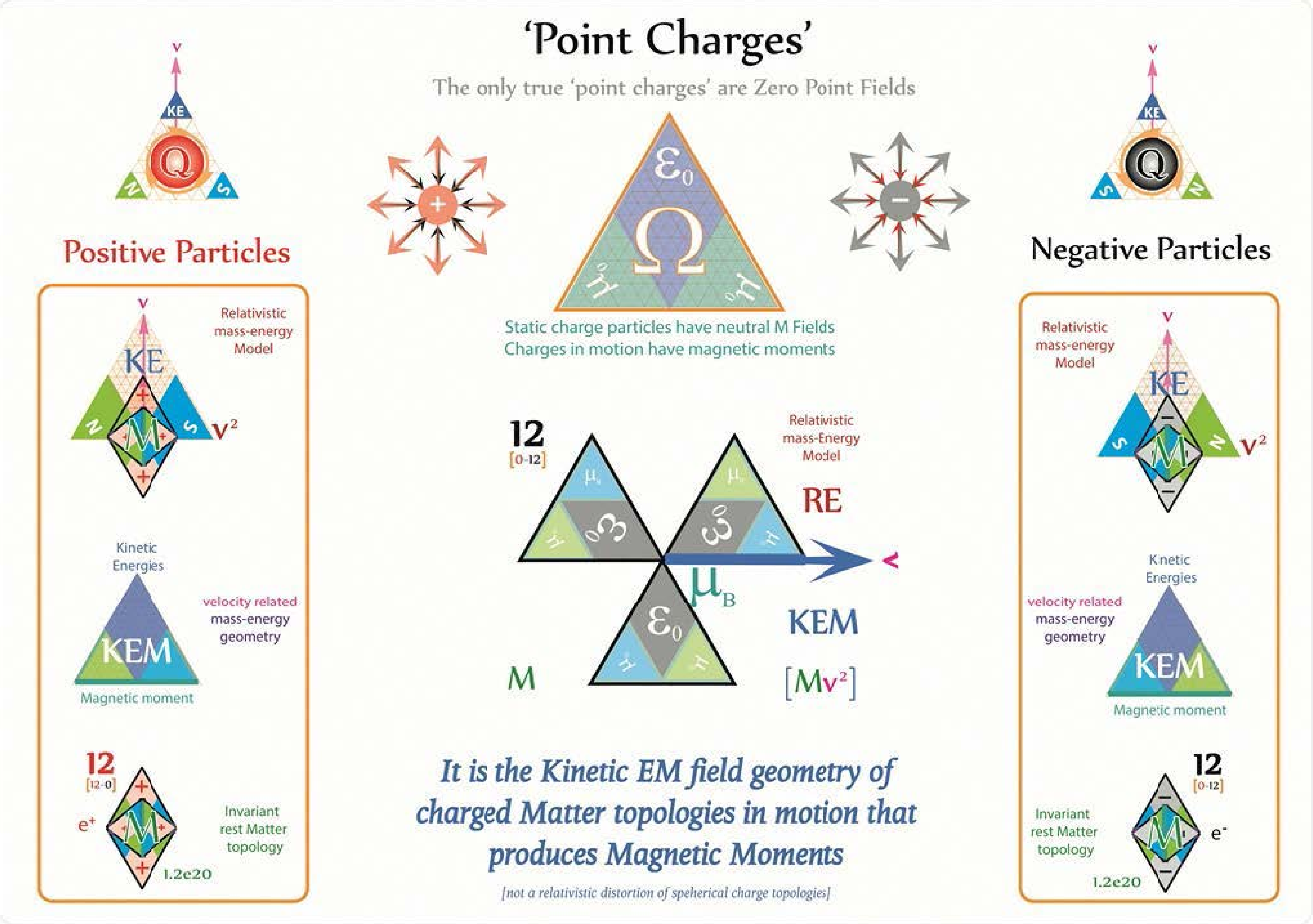
v
KE
Q
N
S
+
−
ε_0
Ω
μ_0
v
KE
M
v^2
KEM
12
[0-12]
μ_0
ε_0
RE
μ_B
KEM
M
$[Mv^2]$
12
[12-0]
e^+
1.2e20
12
[0-12]
e^-
1.2e20

Positive Particles

Negative Particles

Static charge particles have neutral M Fields
Charges in motion have magnetic moments

Relativistic mass-energy Model

Kinetic Energies

velocity related mass-energy geometry

Magnetic moment

Invariant rest Matter topology

Relativistic mass-Energy Model

It is the Kinetic EM field geometry of charged Matter topologies in motion that produces Magnetic Moments

[not a relativistic distortion of speherical charge topologies]

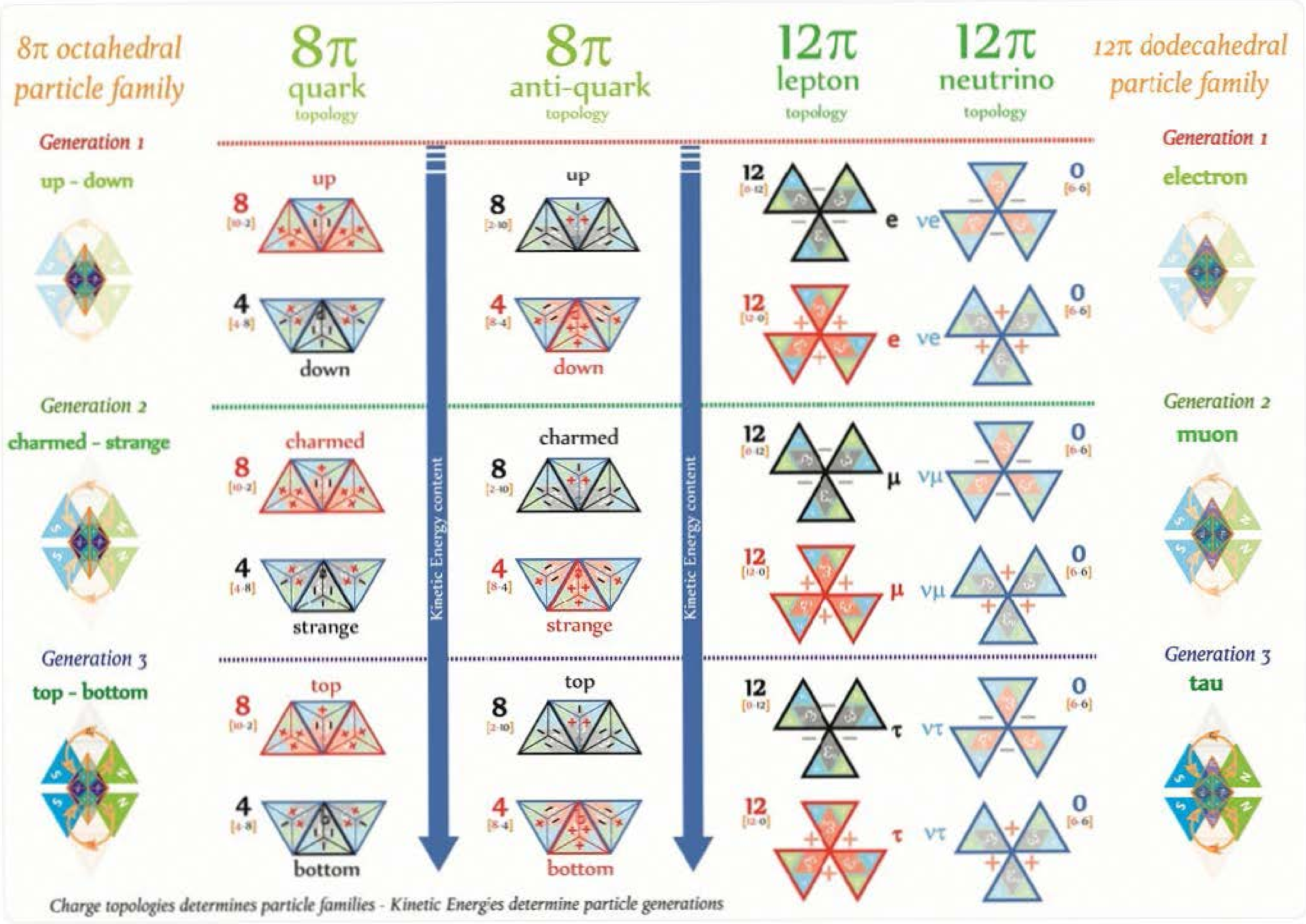
8π octahedral particle family
8π quark topology
8π anti-quark topology
12π lepton topology
12π neutrino topology
12π dodecahedral particle family
Generation 1
up - down
up
8 [10-2]
4 [4-8]
down
up
8 [2-10]
4 [8-4]
down
12 [0-12]
e
ve
0 [6-6]
12 [12-0]
e
ve
0 [6-6]
Generation 1
electron
Generation 2
charmed - strange
charmed
8 [10-2]
4 [4-8]
strange
charmed
8 [2-10]
4 [8-4]
strange
12 [0-12]
μ
νμ
0 [6-6]
12 [12-0]
μ
νμ
0 [6-6]
Generation 2
muon
Generation 3
top - bottom
top
8 [10-2]
4 [4-8]
bottom
top
8 [2-10]
4 [8-4]
bottom
12 [0-12]
τ
ντ
0 [6-6]
12 [12-0]
τ
ντ
0 [6-6]
Generation 3
tau
Kinetic Energy content
Kinetic Energy content
Charge topologies determines particle families - Kinetic Energies determine particle generations

Tetryonics 08.13 - Quark & Lepton families

Copyright ABRAHAM [2008] - all rights reserved

Mesons

are 24π subatomic mass-Matter particles composed of one quark and one antiquark

$$24\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field
Planck
quanta
Mesons
ElectroMagnetic
mass
velocity

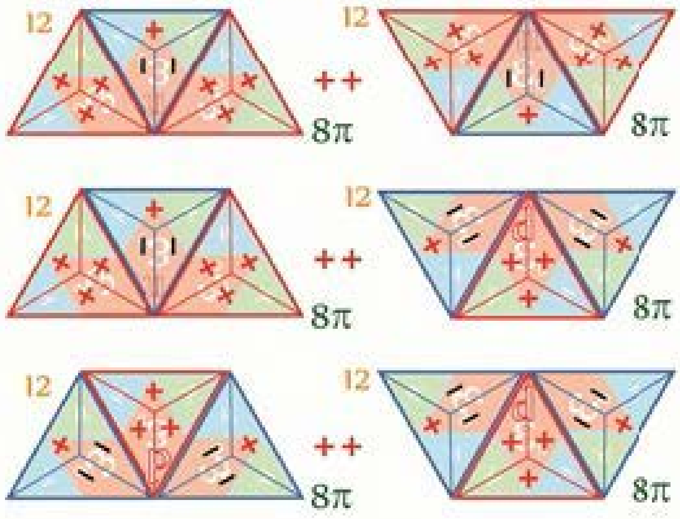
12
++
8π
12
8π
12
++
8π
12
8π
12
++
8π
12
8π

+16 [20-4]

+12 [18-6]

+8 [16-8]

prevented by strong repulives forces

24π charged mass-energy fascia geometries

Pions

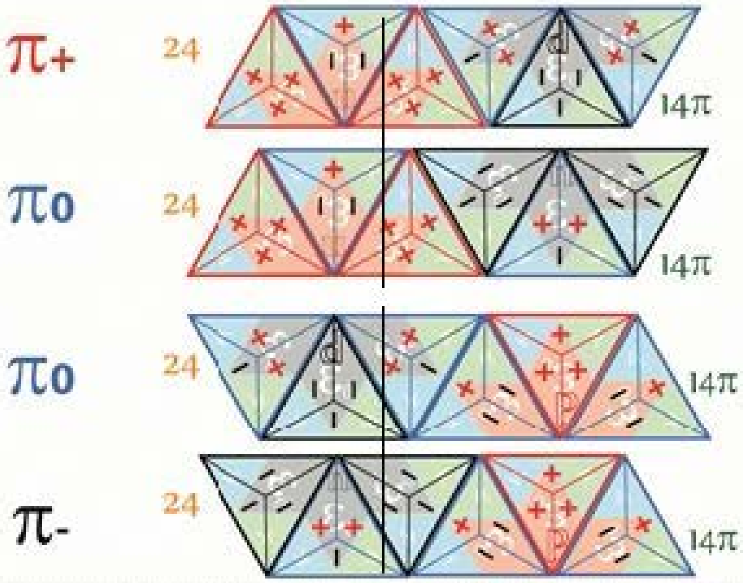
24
14π
24
14π
24
14π
24
14π

π+	+4	[14-10]
πo	0	[12-12]
πo	0	[12-12]
π-	-4	[10-14]

Short lived combinations

14π final mass-Matter topologies

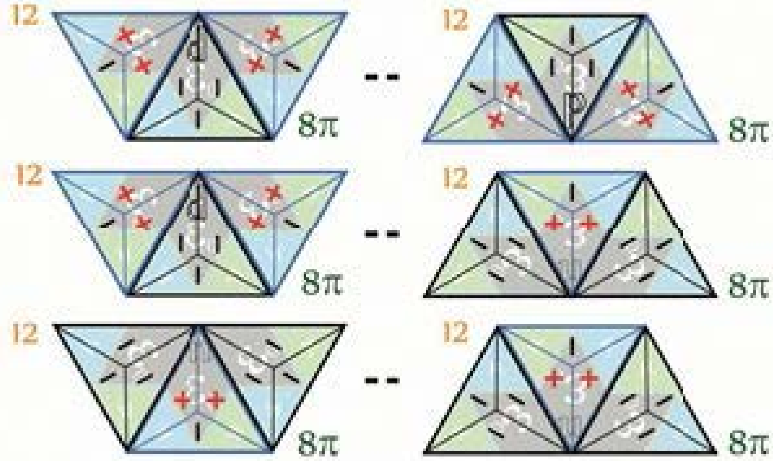
12
--
8π
12
8π
12
--
8π
12
8π
12
--
8π
12
8π

-8 [8-16]

-12 [6-18]

-16 [4-20]

prevented by strong repulives forces

Copyright ABRAHAM [2008] - all rights reserved

PION decay

24π dual -quark mass-energy geometries

12π lepton mass-energy geometries

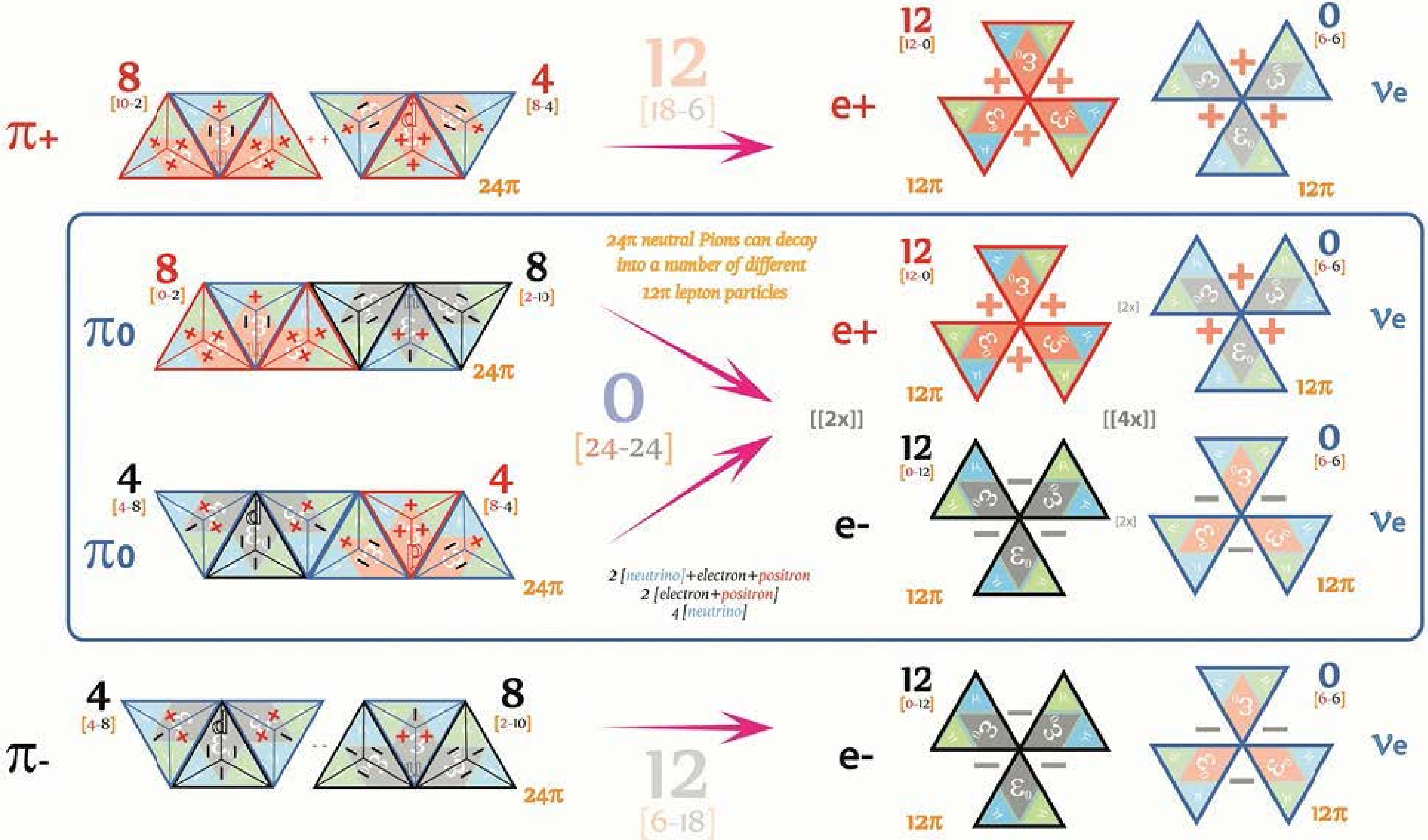
π+
8
[10-2]
4
[8-4]
24π
12
[18-6]
e+
12
[12-0]
12π
0
[6-6]
νe
12π
π0
8
[0-2]
8
[2-10]
24π
0
[24-24]
e+
12
[12-0]
12π
[2x]
0
[6-6]
νe
12π
[[2x]]
[[4x]]
π0
4
[4-8]
4
[8-4]
24π
e-
12
[0-12]
12π
[2x]
0
[6-6]
νe
12π
π-
4
[4-8]
8
[2-10]
24π
12
[6-18]
e-
12
[0-12]
12π
0
[6-6]
νe
12π

24π neutral Pions can decay into a number of different 12π lepton particles

2 [neutrino]+electron+positron
2 [electron+positron]
4 [neutrino]

EM Field Planck quanta

$$24\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Mesons ElectroMagnetic mass velocity

24π dual dodecyon geometries

Despite having 24π geometries Mesons have same nett Charges as 36π Baryons

36π tri-dodecyon geometries

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right] + 12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Lepton - Neutrino generational pairing produced is dependent on the mass-energy levels of the interacting Pions

Copyright ABRAHAM [2008] - all rights reserved

$24\pi + 12\pi = 36\pi$

mesons *charged mass-energy geometries* **Baryons**

Short-lived Particles

All particles seek charge equilibrium

Stable Particles

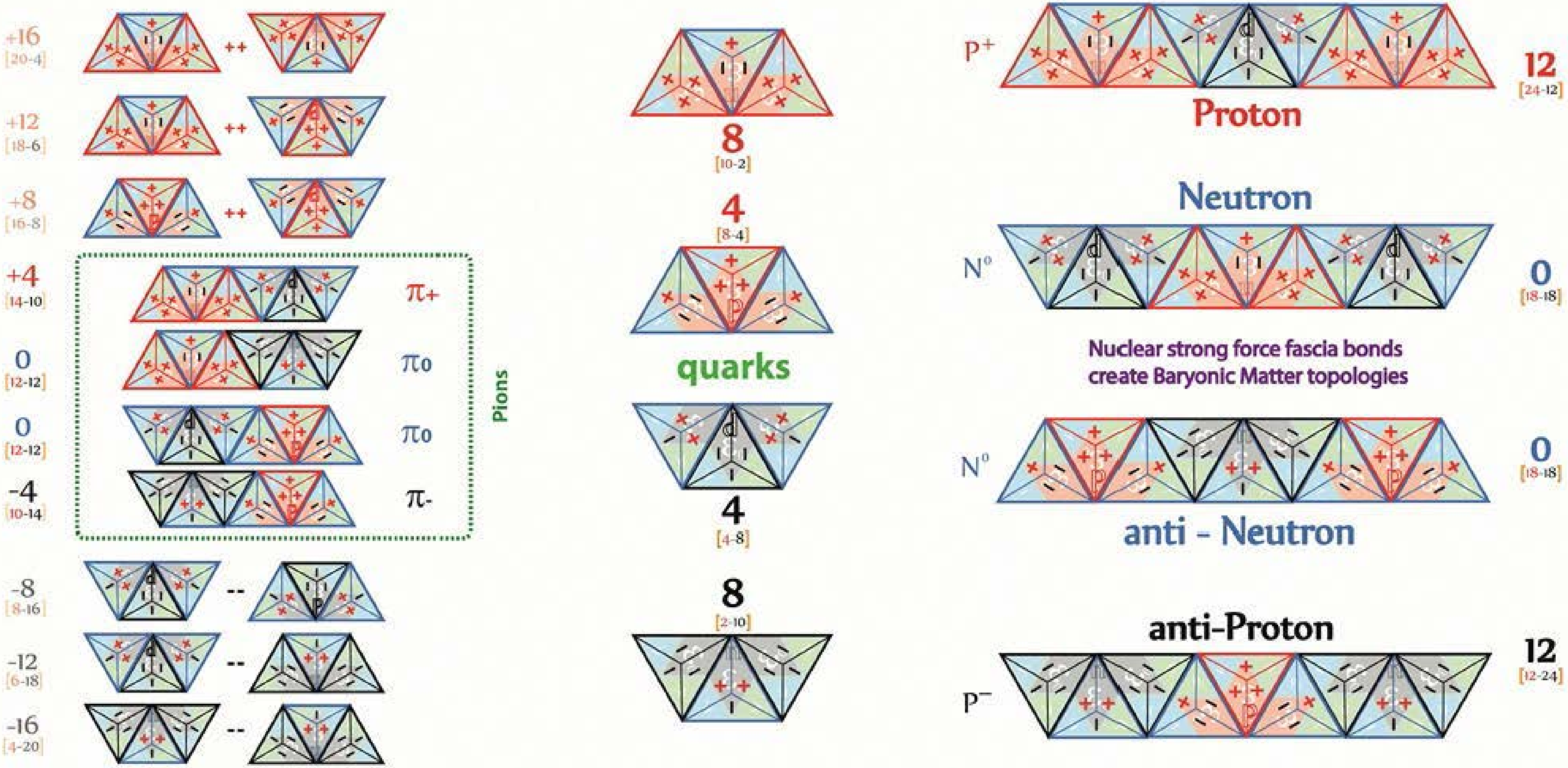
+16
[20-4]
++
+12
[18-6]
++
+8
[16-8]
++
+4
[14-10]
π+
0
[12-12]
πo
0
[12-12]
πo
-4
[10-14]
π-
Pions
-8
[8-16]
--
-12
[6-18]
--
-16
[4-20]
--
8
[10-2]
4
[8-4]
quarks
4
[4-8]
8
[2-10]
P+
Proton
12
[24-12]
Neutron
N°
0
[18-18]
N°
0
[18-18]
anti - Neutron
anti-Proton
P−
12
[12-24]

Nuclear strong force fascia bonds
create Baryonic Matter topologies

All Matter topologies have internalised mass-energy fascia

mesons *mass-Matter topologies* **Baryons**

$14\pi + 8\pi = 20\pi$

Tetryonics 09.03 - Baryon Formation [Meson-Quark interactions]

Baryons

Baryons are 36π mass-energy geometries

$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$

EM Field Planck quanta

Baryons ElectroMagnetic mass velocity

Double Charmed Xi

Charmed Sigma

Lambda

P+

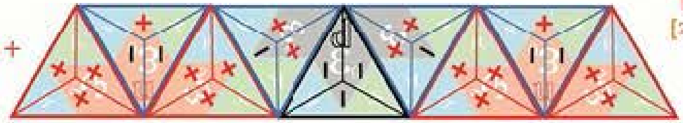

12
[24-12]

Double Bottom Xi

-1

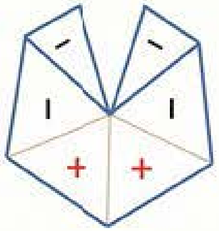

Bottom Xi

Proton

Charmed Bottom Omega

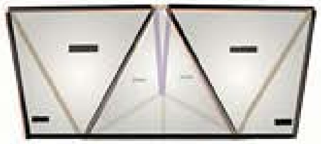

Bottom Omega

Bottom Lambda

-2

Neutron

0
[18-18]

+1

Bottom Sigma

N^0

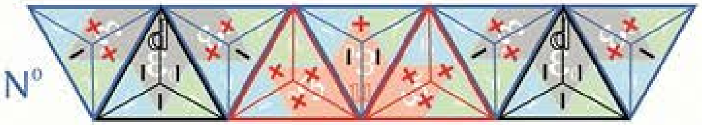

Charmed Omega

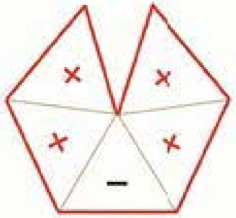

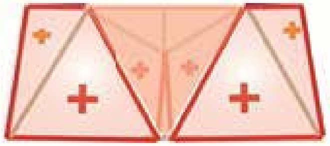

+2

Charmed Bottom Omega Prime

0
[18-18]

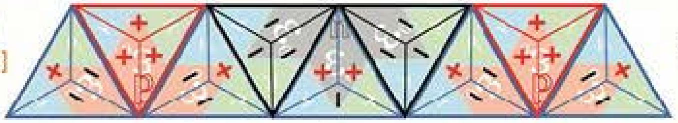

$\overline{N}^0$

0

Charmed Xi Prime

Anti - Neutron

Charmed Lambda

Double Charmed Omega

Xi

Anti-Proton

Bottom Xi Prime

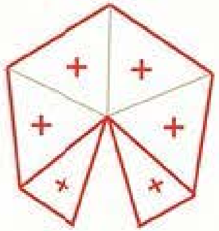

12
[12-24]

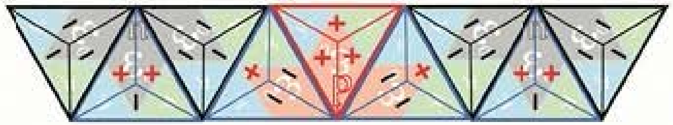

P^-

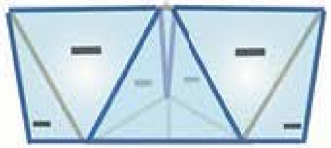

Sigma

Double Bottom Omega

that result in 20π mass-Matter topologies

Copyright ABRAHAM [2008] - all rights reserved

Proton

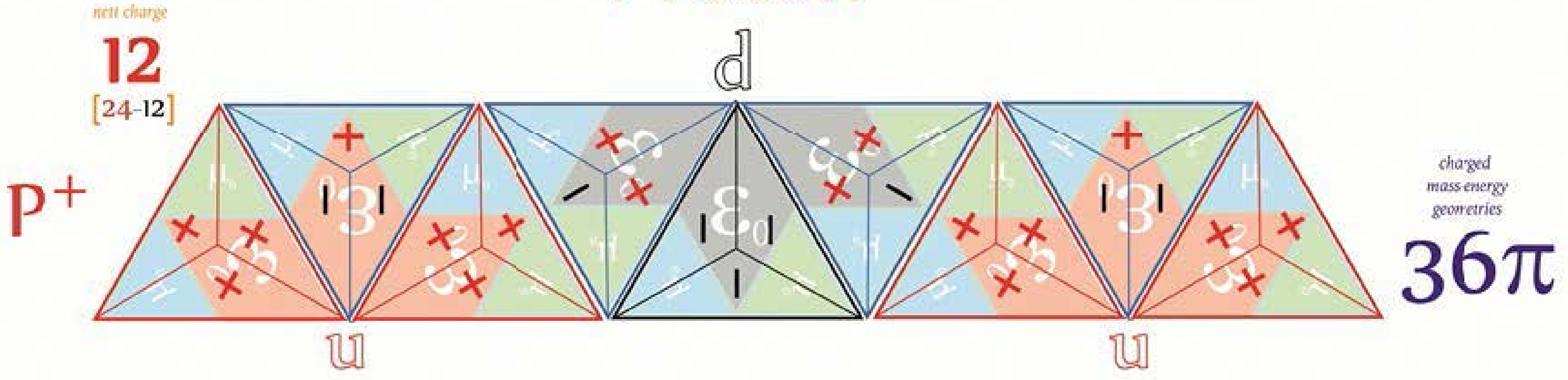
nett charge
12
[24-12]
p+
d
u
u
charged
mass-energy
geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

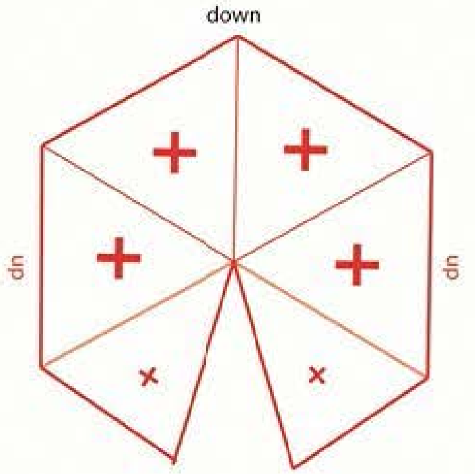

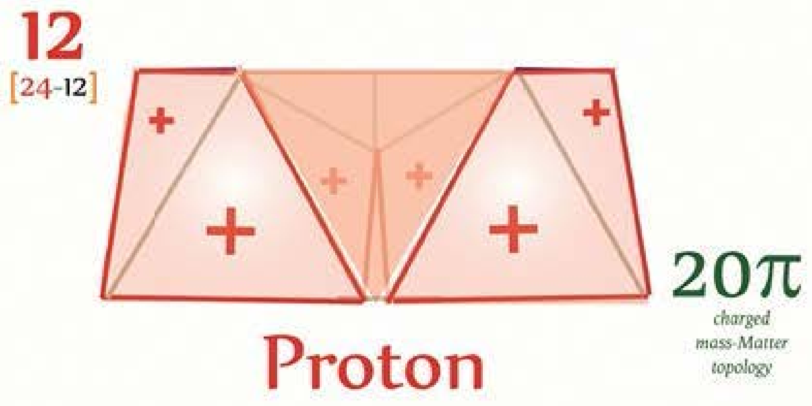
12
[24-12]
Proton
20π
charged
mass-Matter
topology

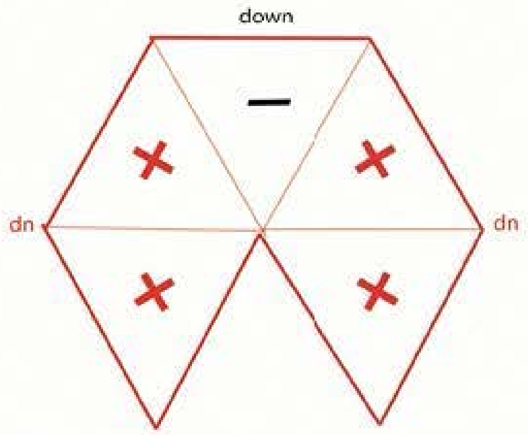

Copyright ABRAHAM [2008] - all rights reserved

Neutron

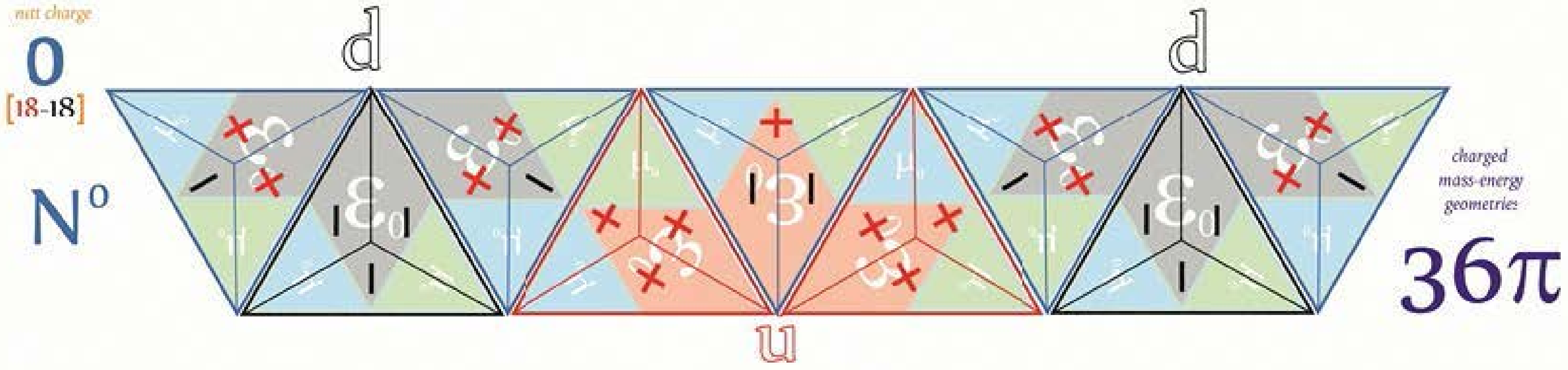
nett charge
0
[18-18]
N⁰
d
d
u
charged
mass-energy
geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

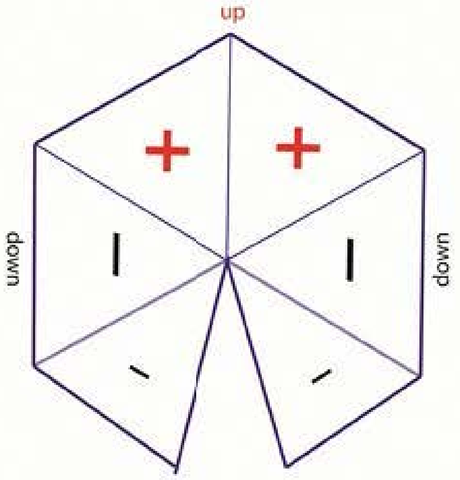
up
down
down

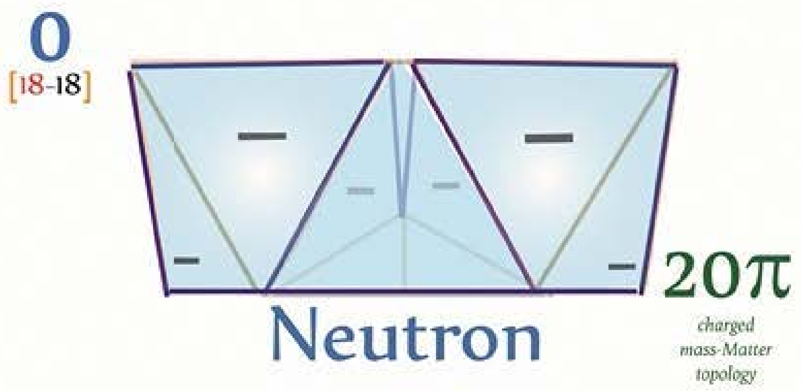
0
[18-18]
Neutron
20π
charged
mass-Matter
topology

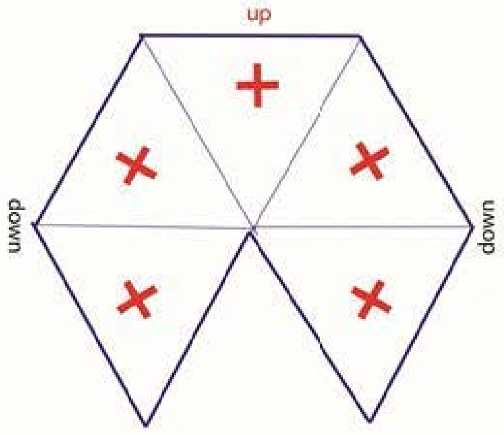
up
down
down

anti-Neutron

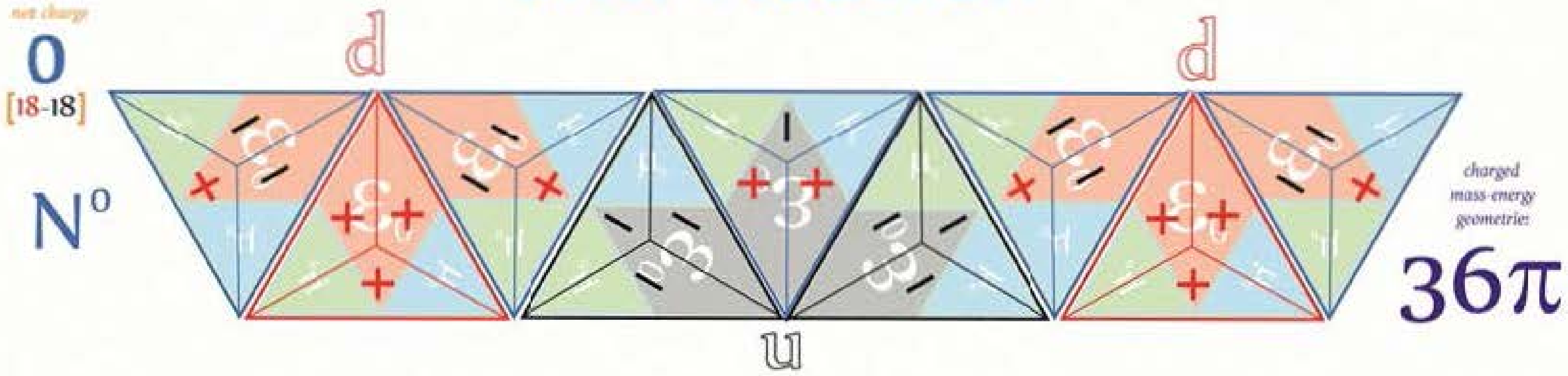
net charge
0
[18-18]
N⁰
d
d
u
charged mass-energy geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

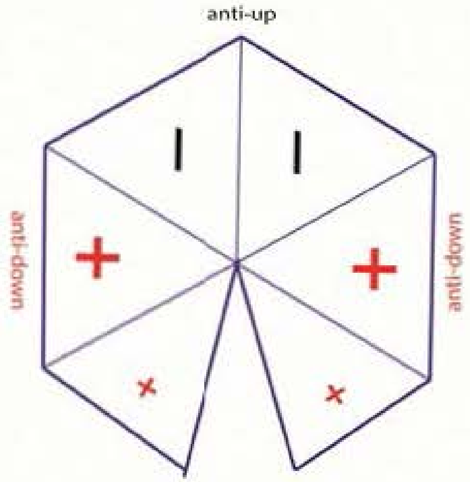
anti-up
anti-down
anti-down

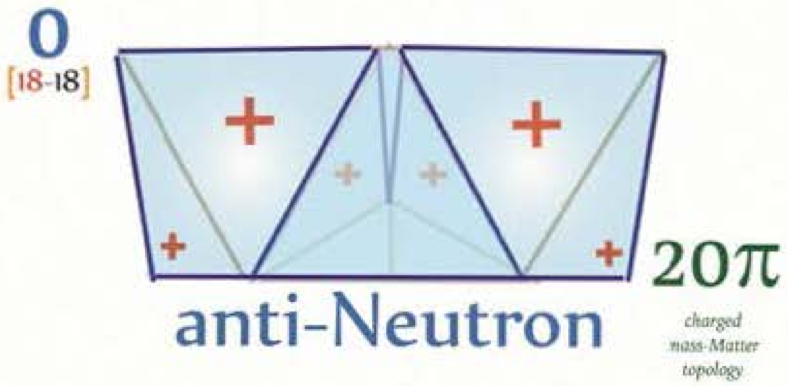
0
[18-18]
anti-Neutron
20π
charged mass-Matter topology

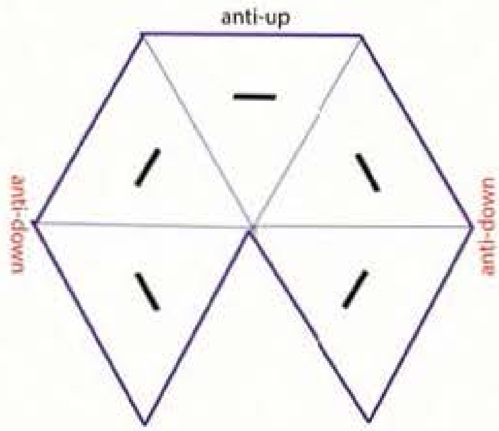
anti-up
anti-down
anti-down

anti-Proton

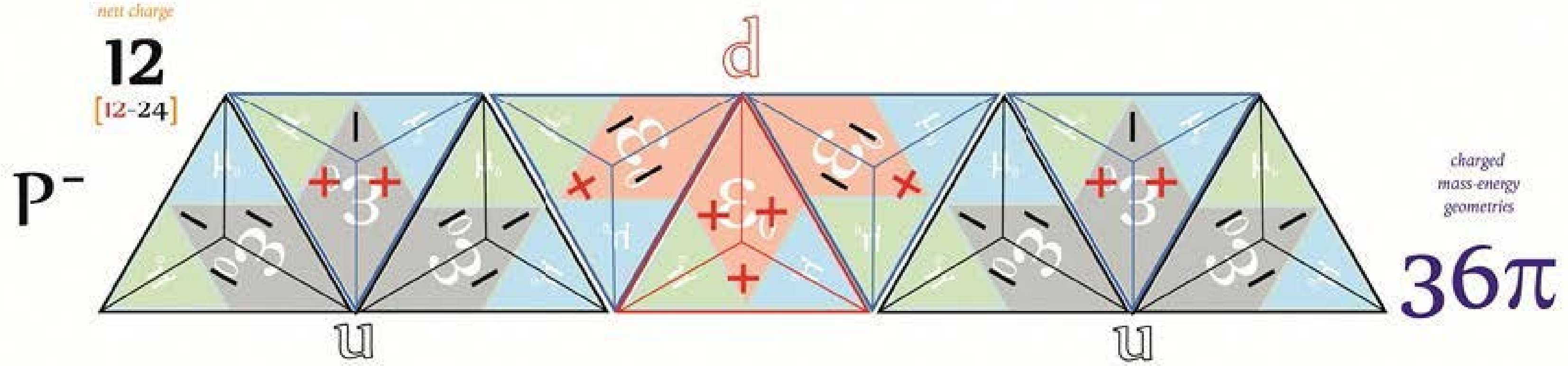
nett charge
12
[12-24]
P⁻
d
u
u
charged mass-energy geometries
36π

Tri-quark mass-energy geometry

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

Baryon mass-Matter topology

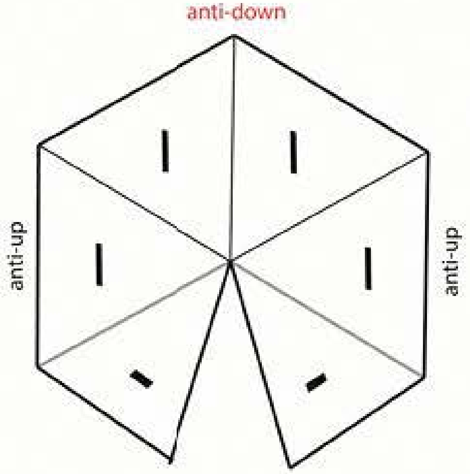
anti-down
anti-up
anti-up

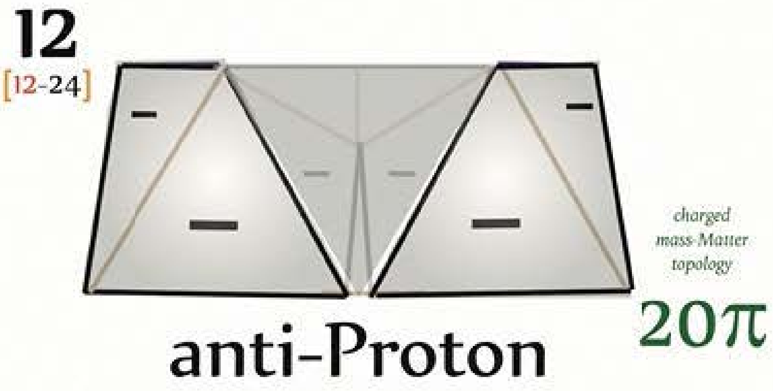
12
[12-24]
anti-Proton
charged mass-Matter topology
20π

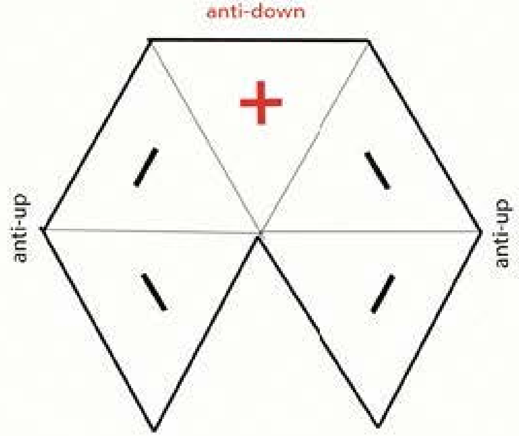
anti-down
anti-up
anti-up

Copyright ABRAHAM [2008] - all rights reserved

Baryon geometry

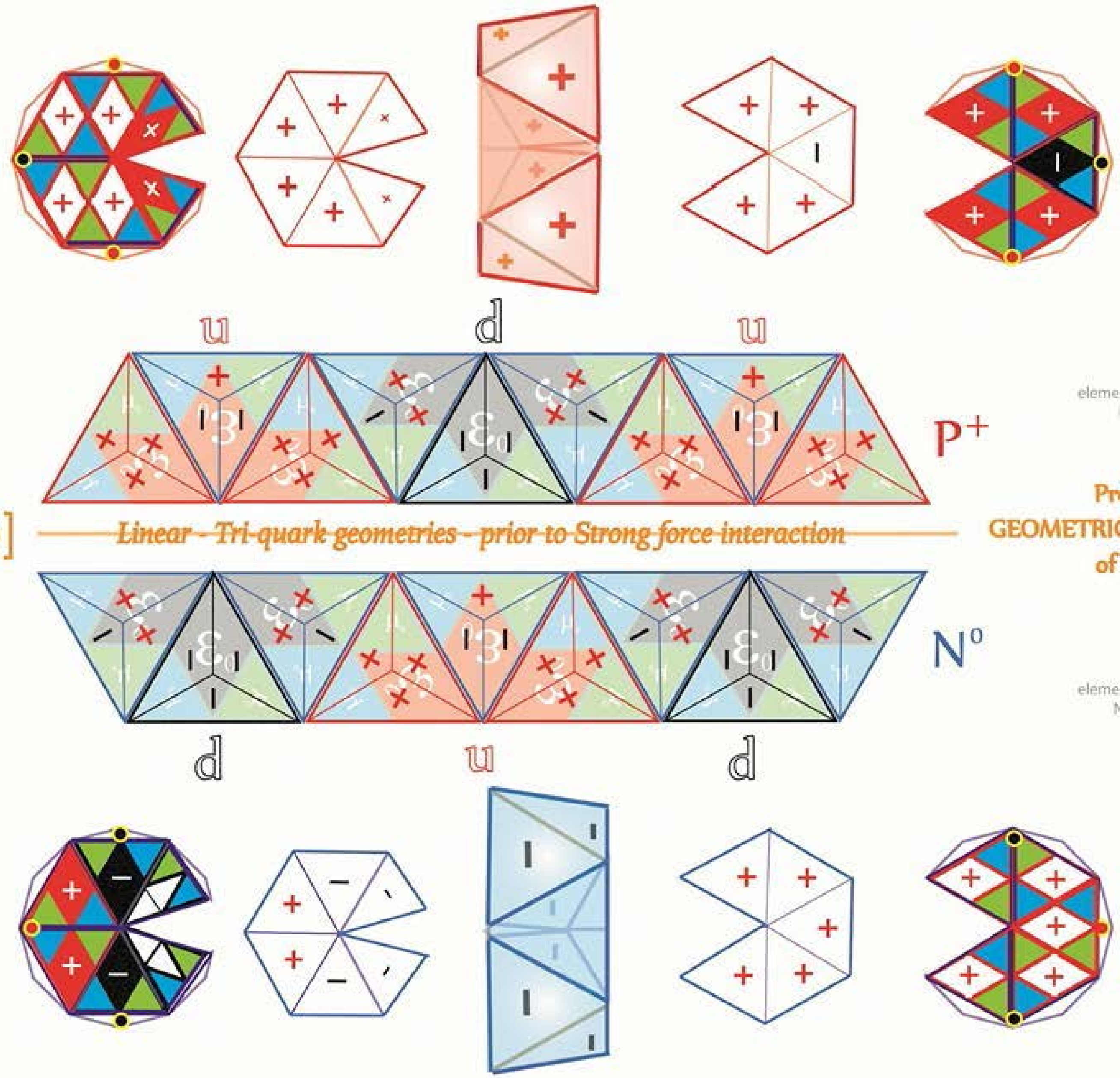

Proton

12

Tetryonic Charge

[24-12]

P^+

+1

elementary charge
Proton

[36]

Linear - Tri-quark geometries - prior to Strong force interaction

**Protons are
GEOMETRIC MIRROR IMAGES
of Neutrons**

N^0

0

elementary charge
Neutron

[18-18]

Tetryonic Charge

0

Neutron

Tetryonics 10.06 - Baryon geometry

Copyright ABRAHAM [2008] - all rights reserved

Chared particle families

[charged mass geometries & Matter topologies]

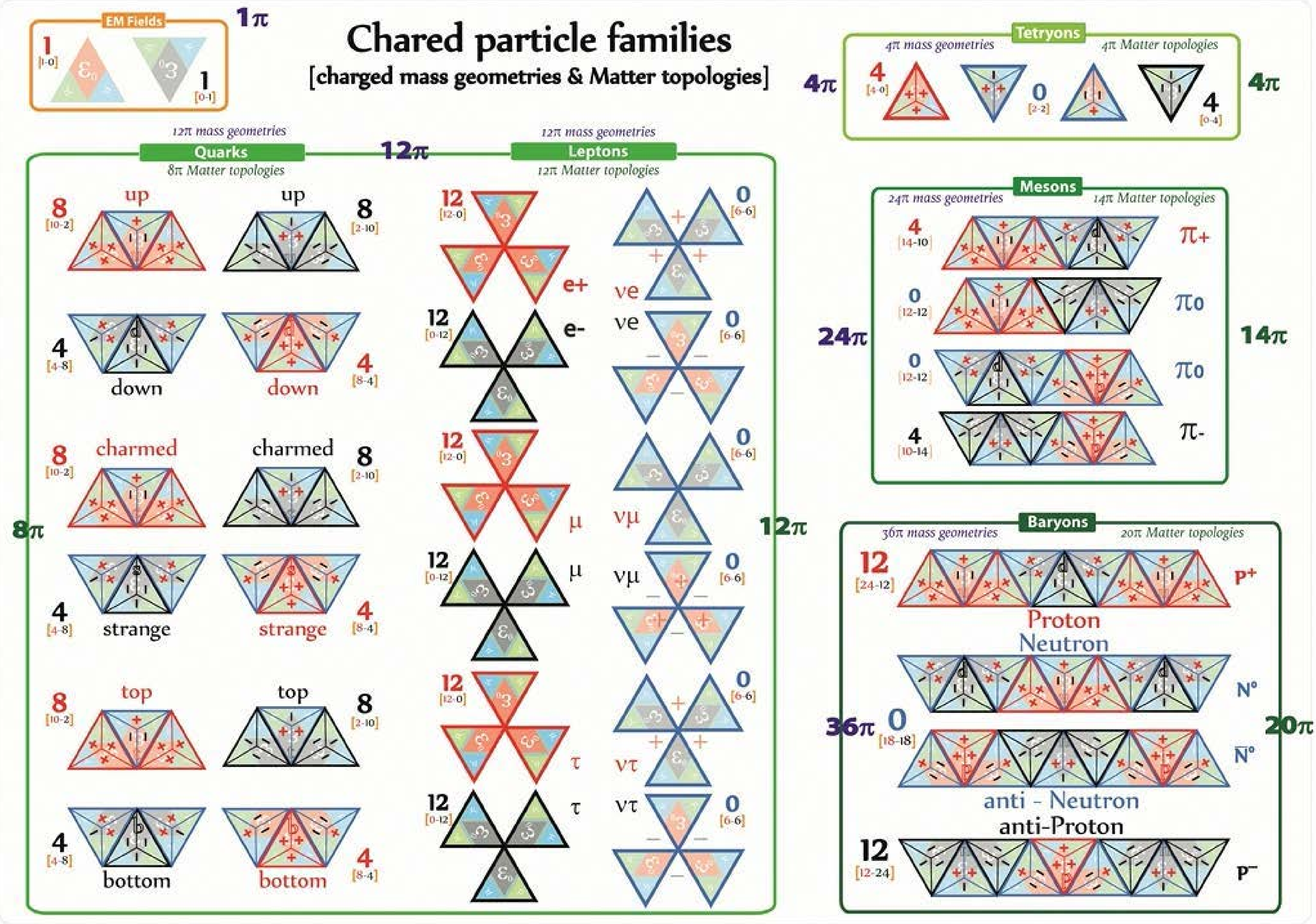
EM Fields
1 [1-0]
1 [0-1]
1π
Tetryons
4π mass geometries
4π Matter topologies
4π
4 [4-0]
0 [2-2]
4 [0-4]
4π
12π mass geometries
Quarks
8π Matter topologies
12π
12π mass geometries
Leptons
12π Matter topologies
8 [10-2] up
up 8 [2-10]
4 [4-8] down
down 4 [8-4]
8 [10-2] charmed
charmed 8 [2-10]
4 [4-8] strange
strange 4 [8-4]
8 [10-2] top
top 8 [2-10]
4 [4-8] bottom
bottom 4 [8-4]
8π
12 [12-0] e+
12 [0-12] e-
νe 0 [6-6]
νe 0 [6-6]
12 [12-0] μ
12 [0-12] μ
νμ 0 [6-6]
νμ 0 [6-6]
12 [12-0] τ
12 [0-12] τ
ντ 0 [6-6]
ντ 0 [6-6]
12π
Mesons
24π mass geometries
14π Matter topologies
24π
4 [14-10] π+
0 [12-12] π0
0 [12-12] π0
4 [10-14] π-
14π
Baryons
36π mass geometries
20π Matter topologies
12 [24-12] P+
Proton
Neutron
N°
36π 0 [18-18]
N̄°
anti - Neutron
anti-Proton
12 [12-24] P−
20π

Tetryonics 10.07 - Tetryonic Charged particle families

Tetryonic Energy levels

$$E = n\pi\left[\left[m\Omega v^2\right]\right]$$

Energy
Planck quanta
mass velocity
momentum

Energy levels

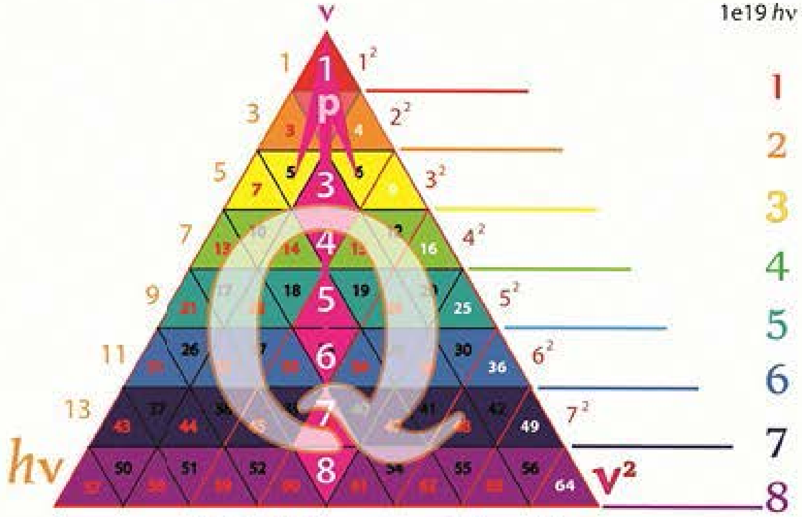
m
1e19 hν = n
v
hν
v²
p
1 3 5 7 9 11 13
1² 2² 3² 4² 5² 6² 7²
1 2 3 4 5 6 7 8
Q

charged field quantisation

Every charged fascia of a Field or Particle holds squared energy levels
[equilateral EM mass-energy geometries & Matter topologies]

rest mass-Matter

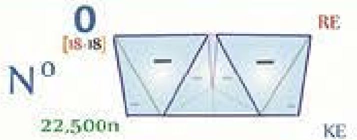
0
[18-18]
N⁰
22,500n
RE
KE

1875
rest Matter
differential

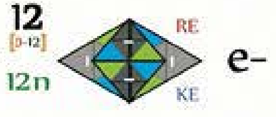
12
[0-12]
12n
RE
KE
e-

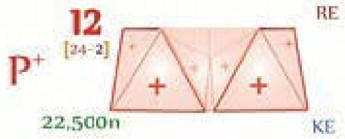
12
[24-2]
P⁺
22,500n
RE
KE

The total mass-energy for any particle is its rest mass-Matter + Kinetic Energies

The mass-energies of its charge geometry determines its rest mass-Matter

An electron's nuclear kinetic energy level is determined by the Baryons it binds to (or by incident photons)

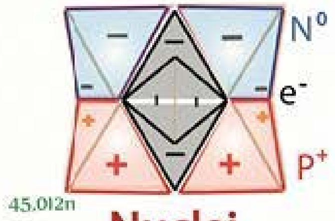
N⁰
e⁻
P⁺
45,012n

Nuclei

Nuclear Quantum Level | electron Kinetic Energy Level

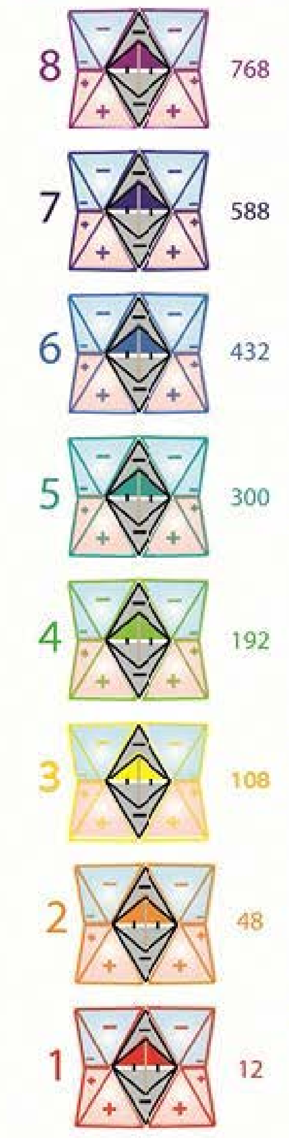
8 768
7 588
6 432
5 300
4 192
3 108
2 48
1 12

As the Nuclei quanutm level increases - so does the electron's energy and tmomenta

The Tetryonic KEM Colour code

As the energy levels of atomic nuclei increase/decrease
the KEM fields of leptons bound to them change accordingly

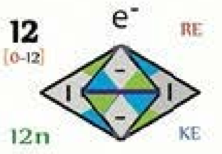
12
[0-12]
e⁻
RE
12n
KE

Electron

1	2	3	4	5	6	7	8
▲	▲	▲	▲	▲	▲	▲	▲
1	4	9	16	25	36	49	64

All energy levels are Tetryonic square number quantas
ie 1, 4, 9, 16, 25 etc (1,2,3,4,5)

$$KEM = 4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Kinetic energies
EM-Field
ElectroMagnetic
Planck quanta
mass velocity

Copyright ABRAHAM [2008] - all rights reserved

tri-quark charged toplogy combinations

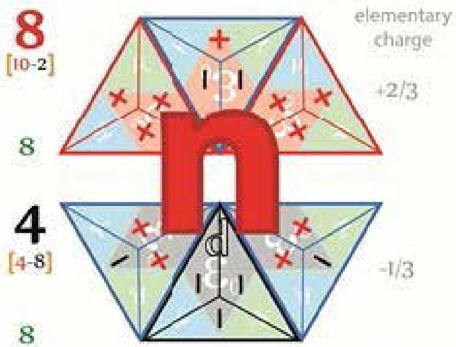

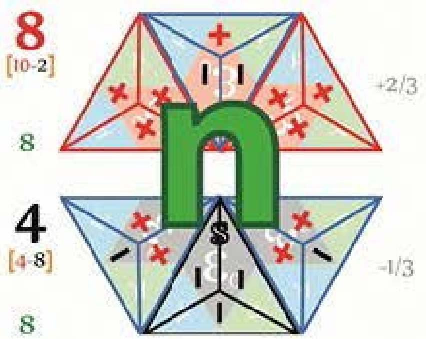

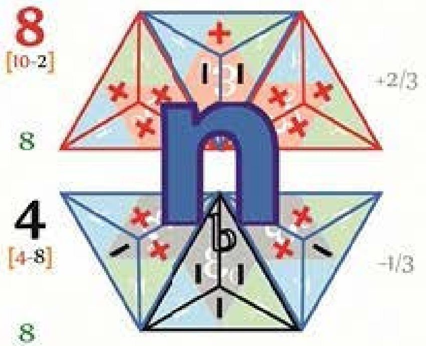

Quarks seek charge equilibrium by
forming Baryonic Matter topologies
as a result of charge interactions
[STRONG nuclear Force]

UP
DOWN] gen 1
CHARMED
STRANGE] gen 2
TOP
BOTTOM] gen 3

[+8 +8 +8] [10-2] [10-2] [10-2] **24** + 2 2/3 2/3 2/3

[+8 -4 +8] [10-2] [4-8] [10-2] **12** + 1 2/3 1/3 2/3

The particle Zoo is the result of tri-quark mass-energy topologies

[-4 +8 -4] [4-8] [10-2] [4-8] **0** 0 0 1/3 2/3 1/3

[-4 -4 -4] [4-8] [4-8] [4-8] **12** - 1 1/3 1/3 1/3

Twice the number of combinations are possible with the inclusion of anti-matter topologies

EM Field
Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons
ElectroMagnetic mass velocity

Any combinations of inter-generational quarks
results in differing final mass-energies
in the created Baryonic Matter
due to energy equalisation

The Particle Zoo

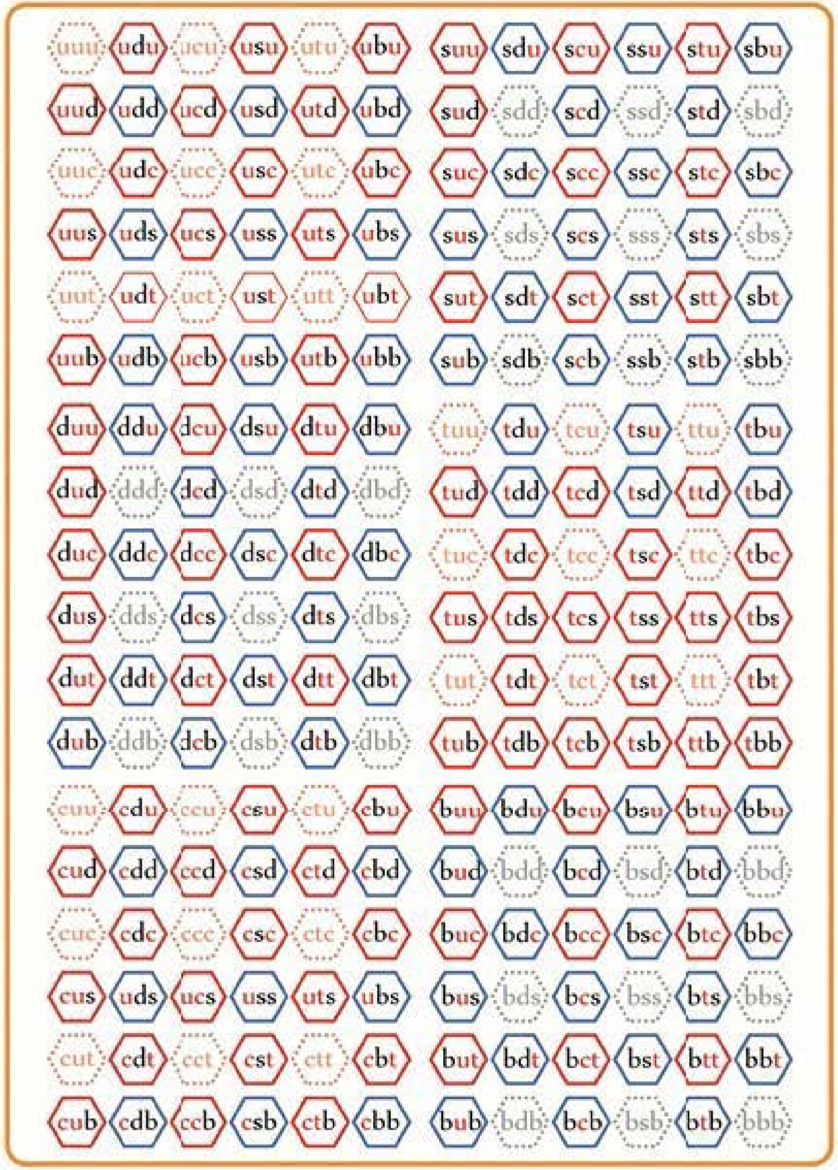

uuu	udu	ucu	usu	utu	ubu	suu	sdu	scu	ssu	stu	sbu
uud	udd	ucd	usd	utd	ubd	sud	sdd	scd	ssd	std	sbd
uuc	udc	ucc	usc	utc	ubc	suc	sdc	scc	ssc	stc	sbc
uus	uds	ucs	uss	uts	ubs	sus	sds	scs	sss	sts	sbs
uut	udt	uct	ust	utt	ubt	sut	sdt	sct	sst	stt	sbt
uub	udb	ucb	usb	utb	ubb	sub	sdb	scb	ssb	stb	sbb
duu	ddu	dcu	dsu	dtu	dbu	tuu	tdu	tcu	tsu	ttu	tbu
dud	ddd	dcd	dsd	dtd	dbd	tud	tdd	tcd	tsd	ttd	tbd
duc	ddc	dcc	dsc	dtc	dbc	tuc	tdc	tcc	tsc	ttc	tbc
dus	dds	dcs	dss	dts	dbs	tus	tds	tcs	tss	tts	tbs
dut	ddt	dct	dst	dtt	dbt	tut	tdt	tct	tst	ttt	tbt
dub	ddb	dcb	dsb	dtb	dbb	tub	tdb	tcb	tsb	ttb	tbb
cuu	cdu	ccu	csu	ctu	cbu	buu	bdu	bcu	bsu	btu	bbu
cud	cdd	ccd	csd	ctd	cbd	bud	bdd	bcd	bsd	btd	bbd
cuc	cdc	ccc	csc	ctc	cbc	buc	bdc	bcc	bsc	btc	bbc
cus	uds	ucs	uss	uts	ubs	bus	bds	bcs	bss	bts	bbs
cut	cdt	cct	cst	ctt	cbt	but	bdt	bct	bst	btt	bbt
cub	cdb	ccb	csb	ctb	cbb	bub	bdb	bcb	bsb	btb	bbb

216 possible Baryon topologies [excluding anti-matter] with each combination producing differing Baryonic nett energy levels

Quarks always form symmetrical [udu] or [dud] charged Baryonic topologies
[same charges never combine except under high energy conditions]
ie. Proton is a [udu] not [uud] as is commonly stated

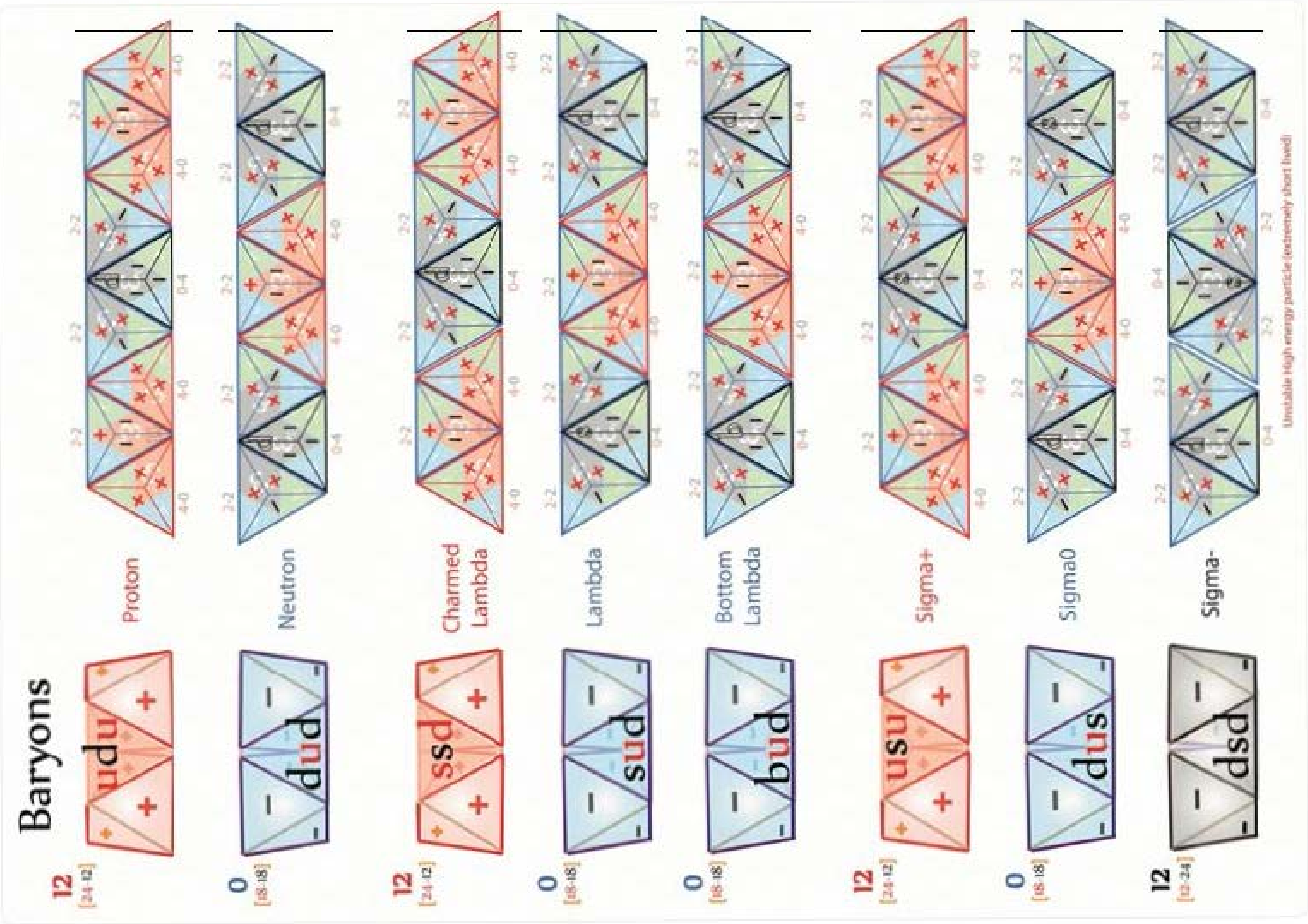
Baryons
12 [24-12]
udu
Proton
0 [18-18]
dud
Neutron
12 [24-12]
ssd
Charmed Lambda
0 [18-18]
sud
Lambda
0 [18-18]
bud
Bottom Lambda
12 [24-12]
usu
Sigma+
0 [18-18]
dus
Sigma0
12 [12-24]
dsd
Sigma-
Unstable High energy particle (extremely short lived)

Tetryonics 10.10 - Proton to Sigma-

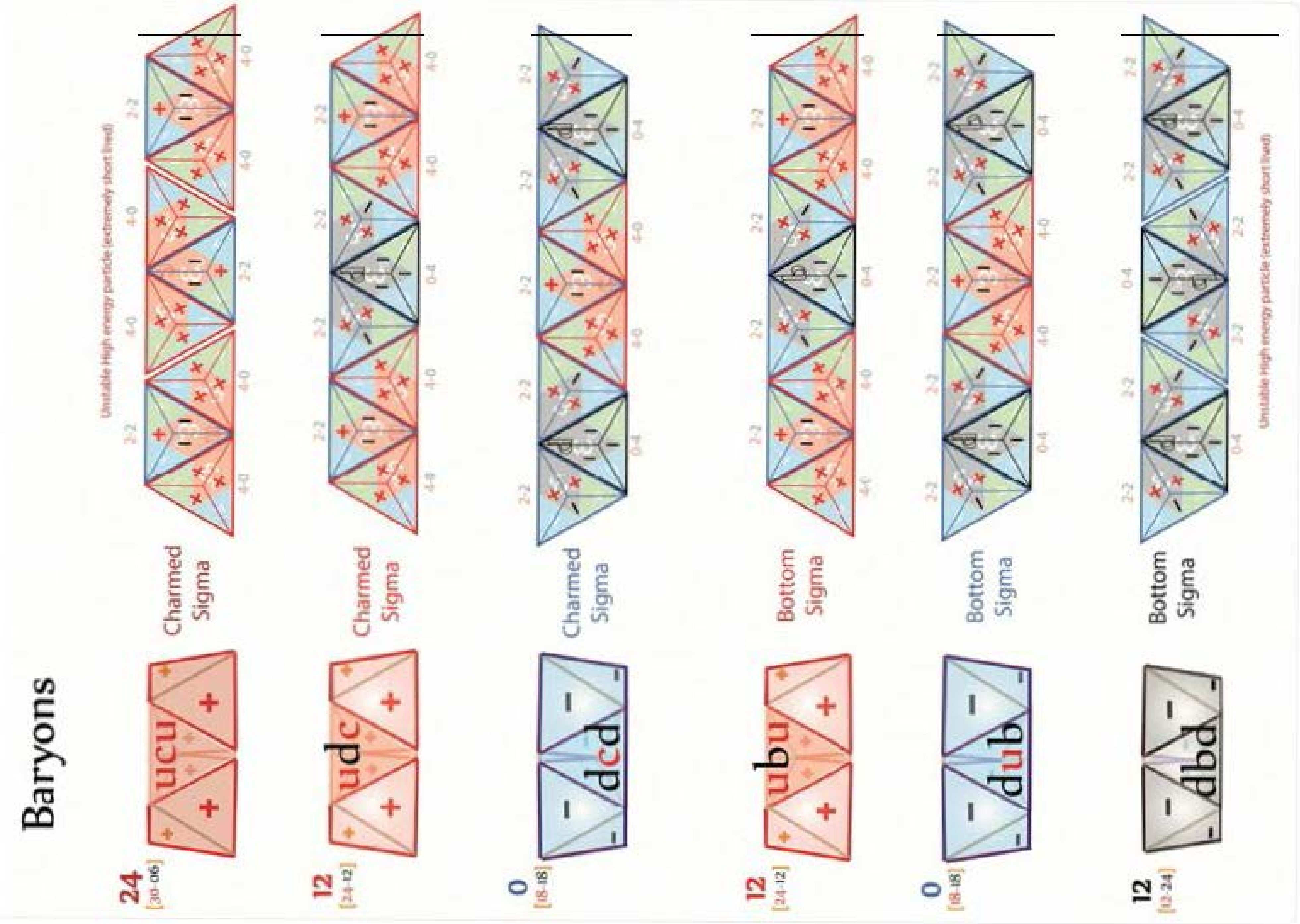
Baryons
24 [30-06] ucu Charmed Sigma
12 [24-12] udc Charmed Sigma
0 [18-18] dcd Charmed Sigma
12 [24-12] ubu Bottom Sigma
0 [18-18] dub Bottom Sigma
12 [12-24] dbd Bottom Sigma
Unstable High energy particle (extremely short lived)
Unstable High energy particle (extremely short lived)

Tetryonics 10.11 - Charmed to Bottom Sigma

Copyright ABRAHAM [2008] - all rights reserved

Baryons

0
[18-18]

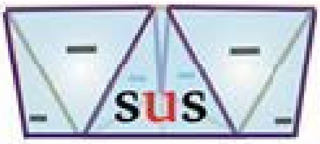

Xi

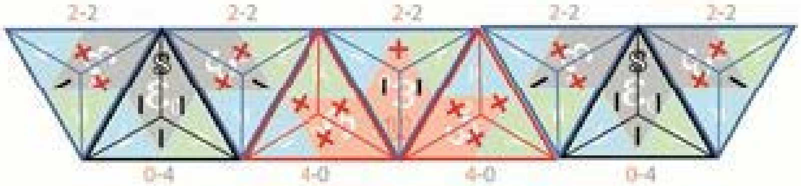
2-2
2-2
2-2
2-2
2-2
s
s
0-4
4-0
4-0
0-4

12
[12-24]

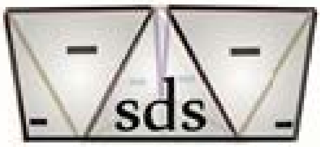

Xi

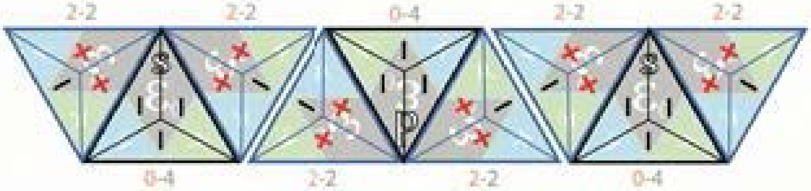
2-2
2-2
0-4
2-2
2-2
s
p
s
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

Charmed Xi Prime

12
[24-12]

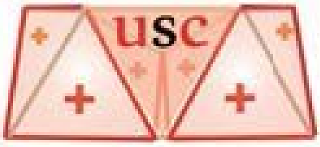

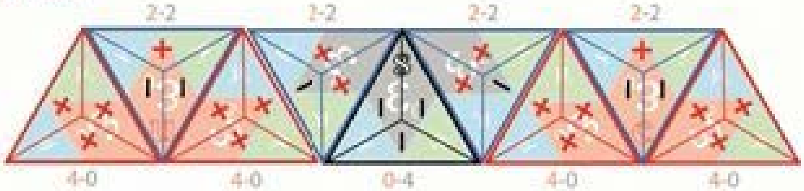
2-2
2-2
2-2
2-2
s
4-0
4-0
0-4
4-0
4-0

Charmed Xi Prime

0
[18-18]

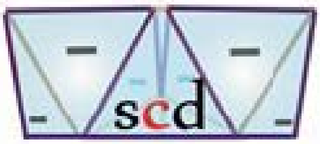

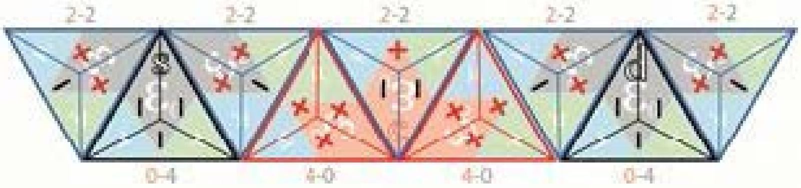
2-2
2-2
2-2
2-2
2-2
s
d
0-4
4-0
4-0
0-4

Double Charmed Xi

24
[30-06]

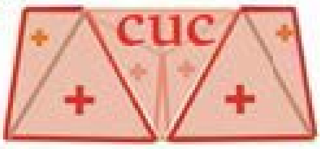

Unstable High energy particle (extremely short lived)

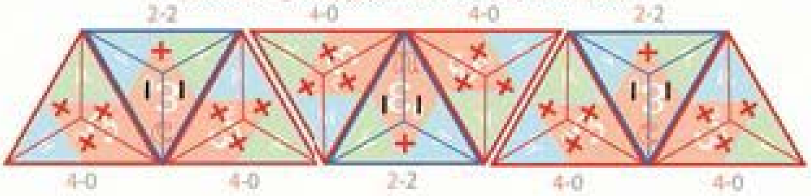
2-2
4-0
4-0
2-2
4-0
4-0
2-2
4-0
4-0

Double Charmed Xi

12
[24-12]

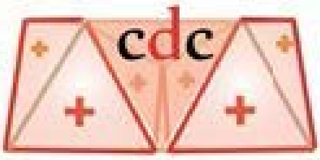

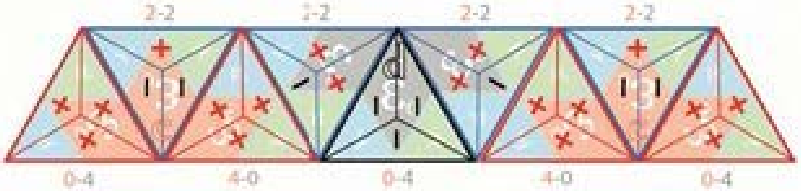
2-2
2-2
2-2
2-2
d
0-4
4-0
0-4
4-0
0-4

Copyright ABRAHAM [2008] - all rights reserved

Baryons

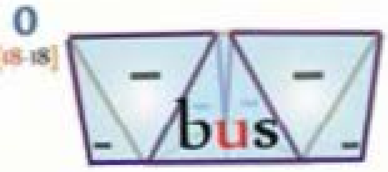

Bottom Xi
(Cascade B)

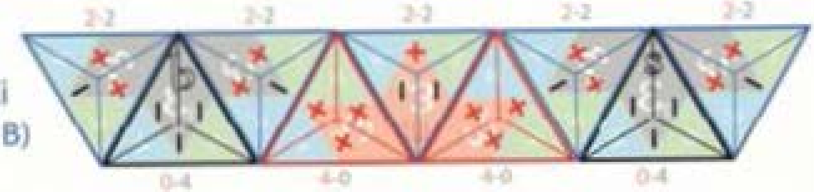
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

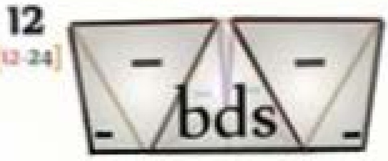

Bottom Xi
(Cascade B)

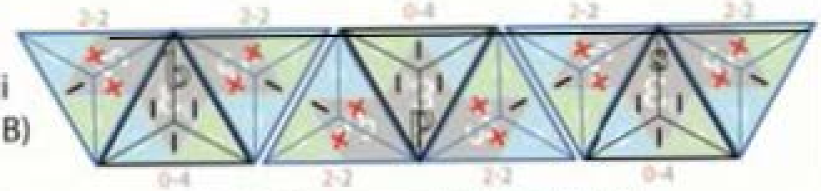
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

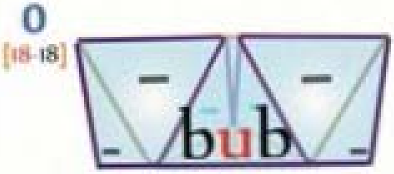

Double
Bottom Xi

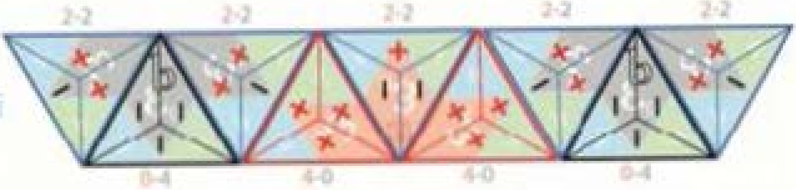
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

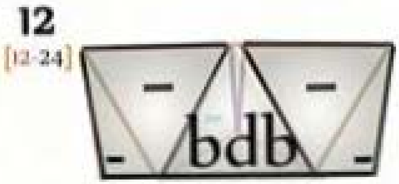

Double
Bottom Xi

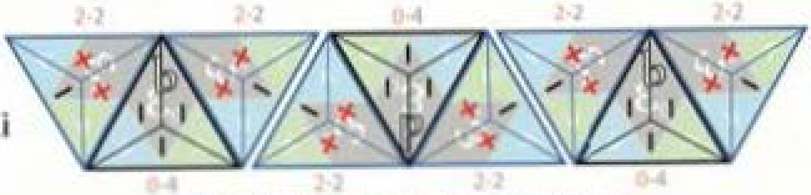
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable High energy particle (extremely short lived)

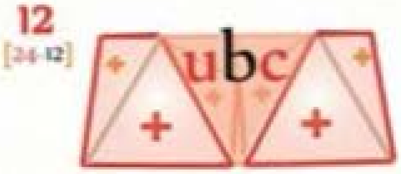

Charmed
Bottom Xi

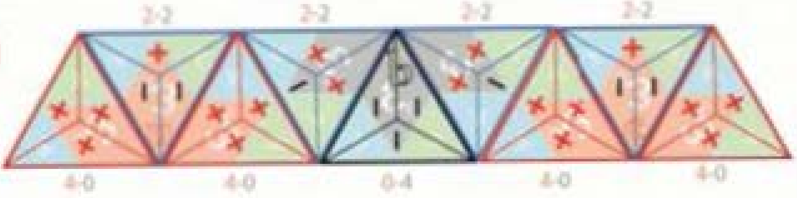
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

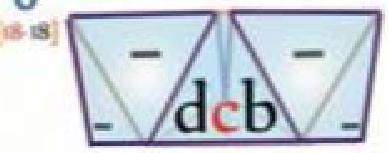

Charmed
Bottom Xi

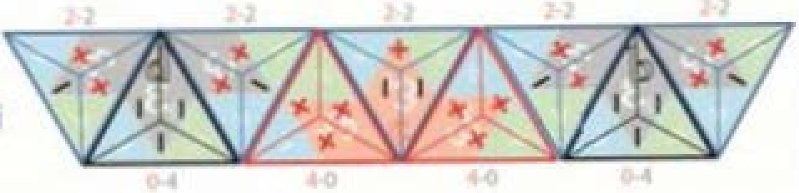
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

Baryons

0
[18-18]

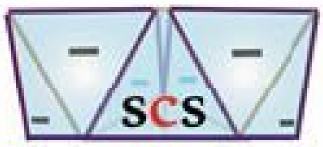

Charmed
Omega

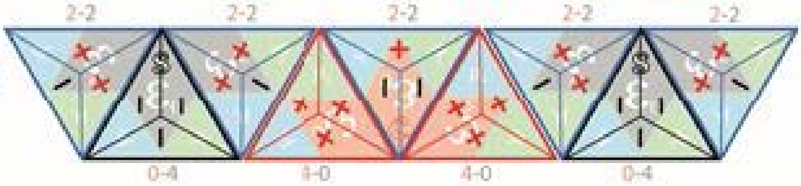
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

12
[12-24]

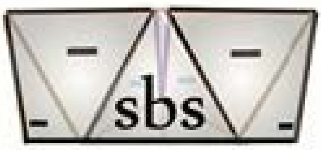

Bottom
Omega

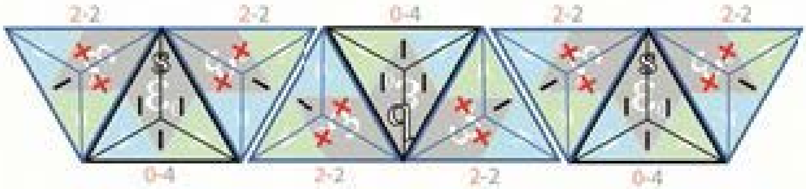
2-2
2-2
0-4
2-2
2-2
0-4
2-2
2-2
0-4

Unstable Highenergy particle (extremely short lived)

12
[24-12]

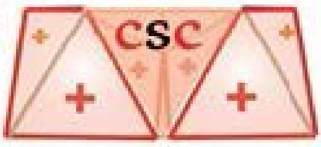

Double
Charmed
Omega

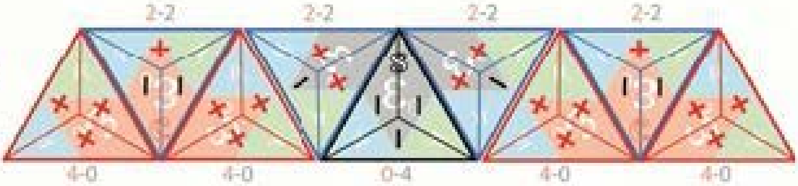
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

0
[18-18]

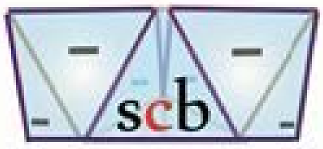

Charmed
Bottom
Omega

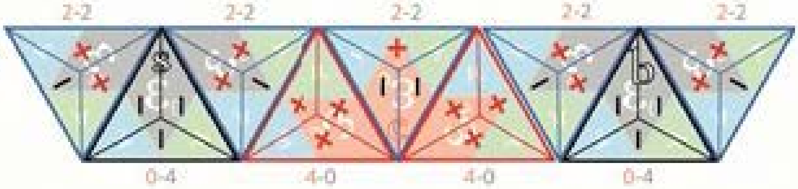
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

12
[12-24]

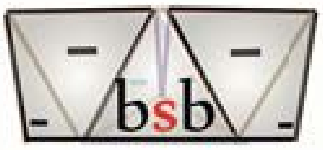

Double
Bottom
Omega

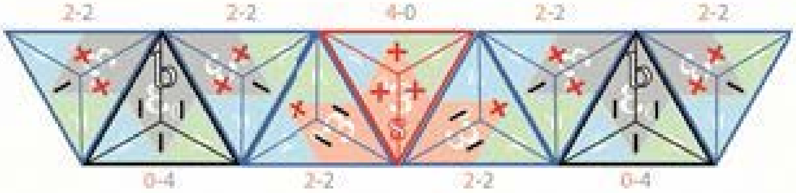
2-2
2-2
4-0
2-2
2-2
0-4
2-2
2-2
0-4

12
[24-12]

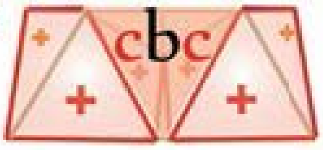

Double
Charmed
Bottom
Omega

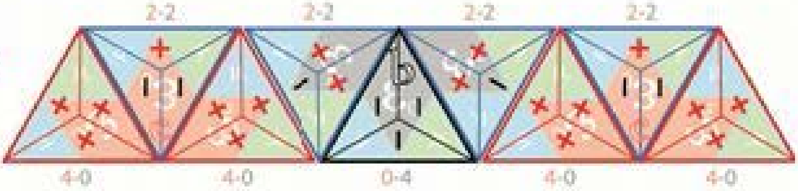
2-2
2-2
2-2
2-2
4-0
4-0
0-4
4-0
4-0

0
[18-18]

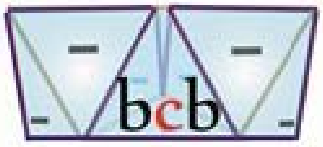

Charmed
Double
Bottom
Omega

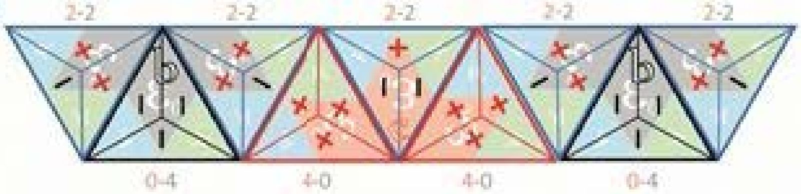
2-2
2-2
2-2
2-2
2-2
0-4
4-0
4-0
0-4

Copyright ABRAHAM [2008] - all rights reserved

ZPF **Photons** **Bosons** **Tetryons**

$n\pi$ charged mass-Energy-Matter geometries

ZPFs are the quantum of Bosons

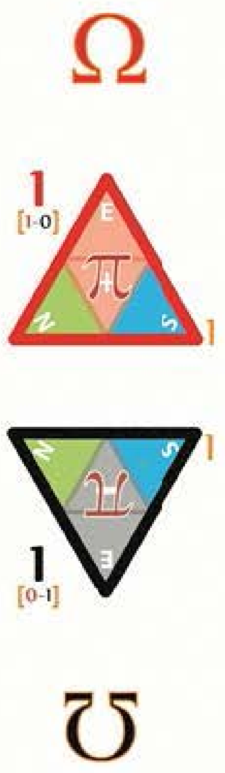

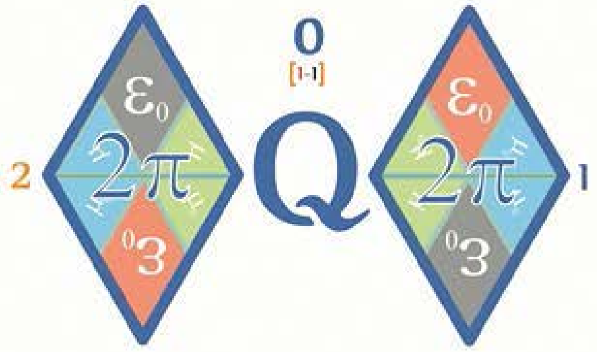

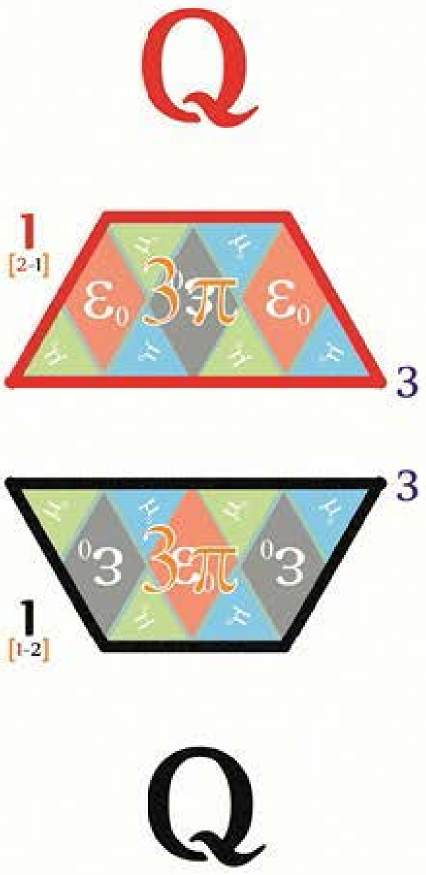

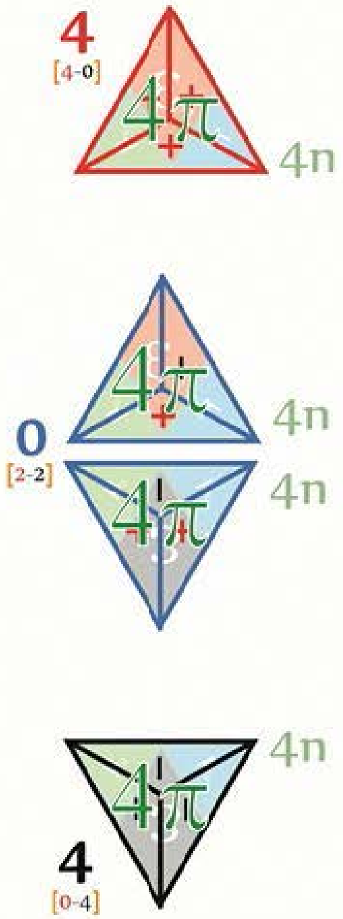

All energies seek equilibrium

Energy momenta

$$\underset{\text{Energy}}{n\pi}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v}}\right]^2\right]$$

EM mass-Energies

$$\underset{\text{mass-energy}}{\text{EVEN}\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

Charges

$$\underset{\text{Charge}}{\text{ODD}\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

Matter

$$\underset{\text{Matter}}{4n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\,\Omega\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

Gluons

The gluon can be considered to be the fundamental exchange particle facilitating the strong interaction between protons and neutrons in a nucleus.

Gluons are the exchange particles for the 'colour' strong force between quarks, analogous to the exchange of photons in the electromagnetic force between two charged particles.

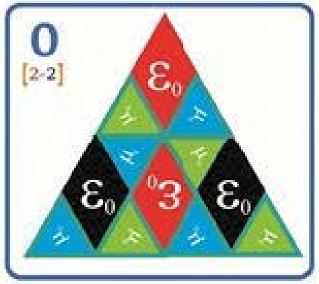
0
[2-2]
ε_0
$^0 3$
ε_0
ε_0

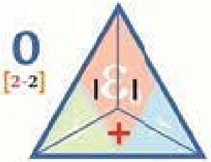

Ω

Gluons are considered to be bi-coloured, carrying a unit of colour and a unit of anti-colour

℧

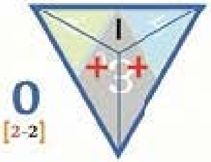

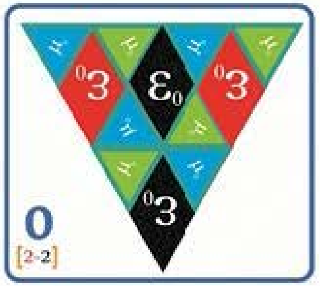
$^0 3$
ε_0
$^0 3$
$^0 3$
0
[2-2]

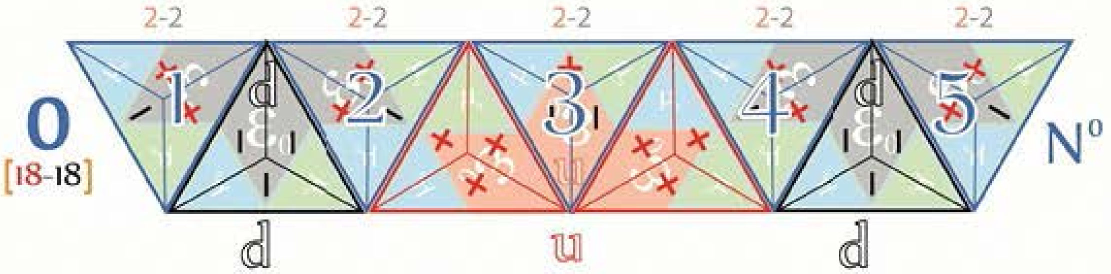
2-2
2-2
2-2
2-2
2-2
0
[18-18]
1
2
3
4
5
N^0
d
u
d

Q

Gluons are the neutral Tetryons inside all fermionic Matter topologies

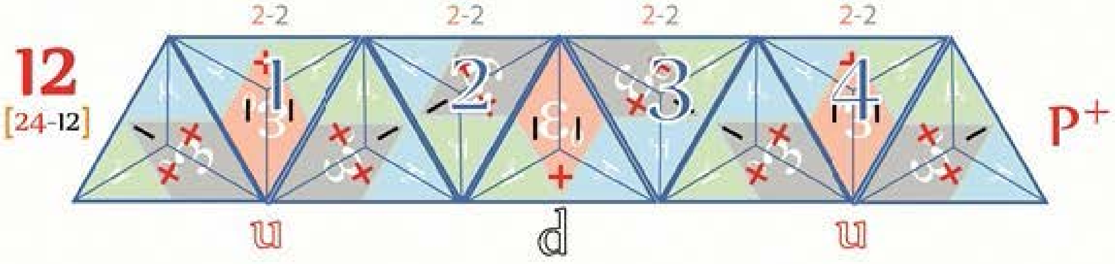
2-2
2-2
2-2
2-2
12
[24-12]
1
2
3
4
P^+
u
d
u

They are the neutral charge [di-electric] Tetryons located between charged Tetryons in Matter topologies

[attractive coulombic forces between their opposite charged fascia create the 'colour force]

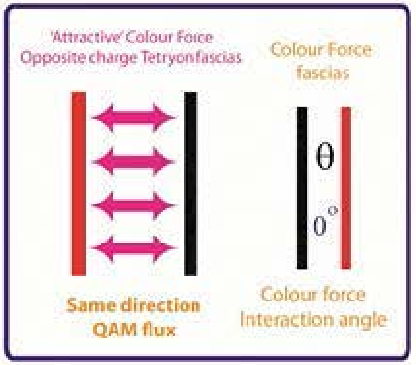
'Attractive' Colour Force
Opposite charge Tetryonfascias
Colour Force fascias
θ
0°
Same direction QAM flux
Colour force Interaction angle

The interaction between neutral Tetryons (Gluons) and all charged Tetryons is the Strong Colour force

nett charge of Gluons is same as a Neutron

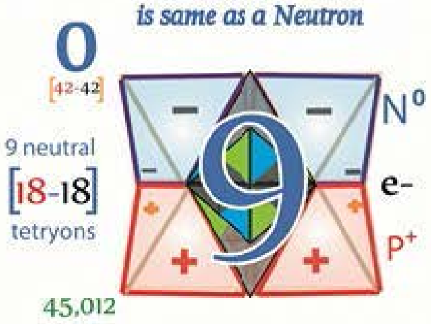
0
[42-42]
9 neutral
[18-18]
tetryons
45,012
9
N^0
e-
P^+

Electrons contain no Gluons

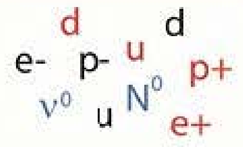
d
d
e-
p-
u
p+
ν^0
u
N^0
e+

Neutral Tetryons combine with charged Tetryons to form all Particles

Gluons in Fermions

In the Standard model Gluons have no mass and no electric charge.
They carry a special 'colour charge' property that hold quarks together to form Baryons

Tetryonics reveals that gluons are neutral charge tetryons comprised of equal positive & negative mass-energies & are elementary Matter particles

'Gluons' are neutral tetryons

Neutrinos are 3 Gluon sets

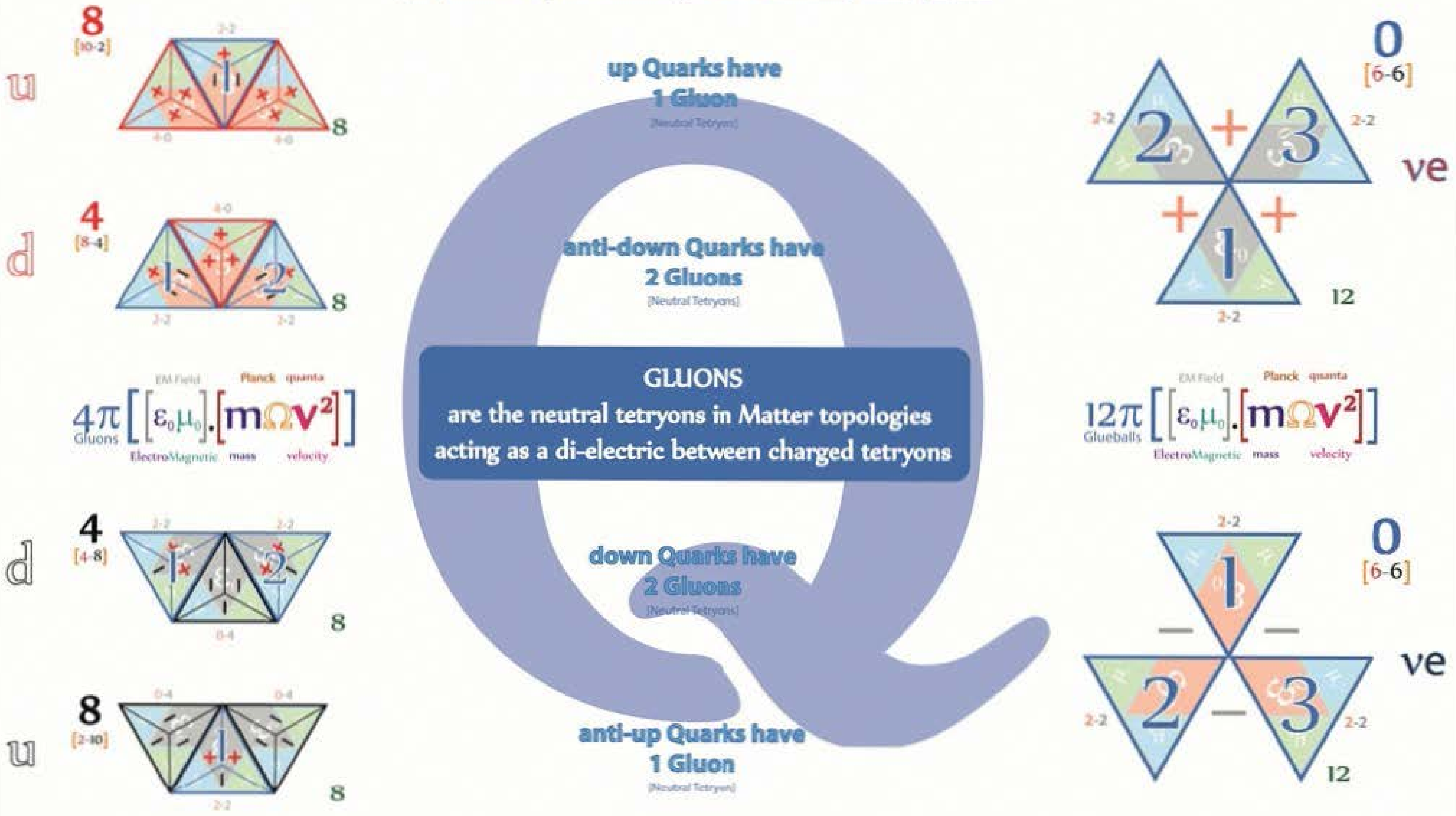

up Quarks have 1 Gluon
[Neutral Tetryon]

anti-down Quarks have 2 Gluons
[Neutral Tetryons]

GLUONS
are the neutral tetryons in Matter topologies acting as a di-electric between charged tetryons

down Quarks have 2 Gluons
[Neutral Tetryons]

anti-up Quarks have 1 Gluon
[Neutral Tetryon]

EM Field Planck quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Gluons ElectroMagnetic mass velocity

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Glueballs ElectroMagnetic mass velocity

Baryonic Gluons

In 'the Standard Model', Gluons are vector gauge bosons that mediate strong interactions of quarks in quantum chromodynamics (QCD). Unlike the electrically neutral photon of quantum electrodynamics (QED), gluons themselves carry colour charge and therefore participate in the strong interaction in addition to mediating it, making QCD significantly harder to analyze than QED.

They are considered to be elementary particles which act as the exchange particles (or gauge bosons) for the colour force between quarks, analogous to the exchange of photons in the electromagnetic force between two charged particles

Tetryonics simplifies the current definition of Gluons and clearly identifies their geometric properties, along with their role in particle genesis

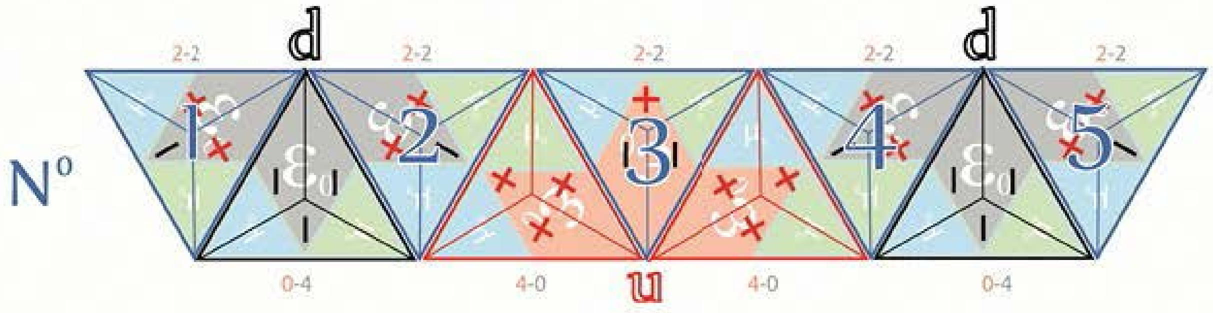

Neutrons have 5 Gluons

[Neutral Tetryons]

0

[18-18]

Deuterium nuclei (being the constituent quanta of all Elements) have 9 neutral Tetryons (Gluons) which in turn contribute to their gravitational mass along with the charged Tetryons

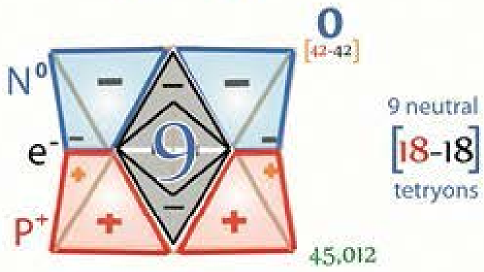

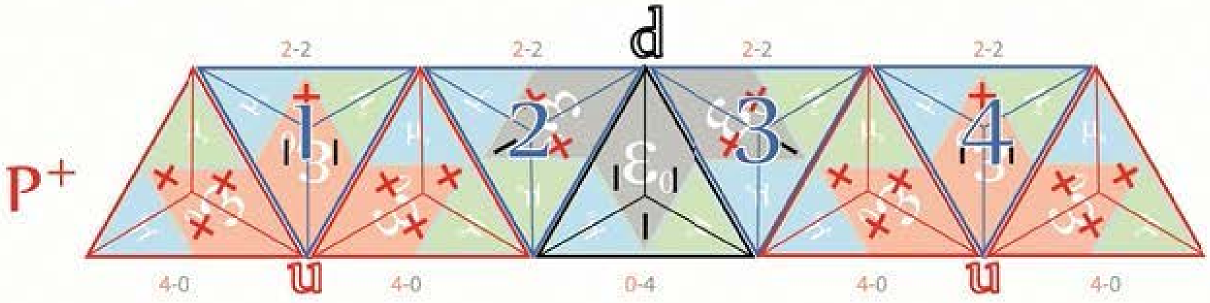

12

[24-12]

Protons have 4 Gluons

[Neutral Tetryons]

'Glueballs'

In particle physics, a glueball is a hypothetical composite particle.
It solely consists of gluon particles, without valence quarks.

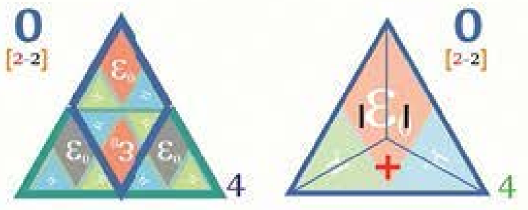

Neutral Tetryons
are the result of equal number
charged EM fields combining to form
neutral charge Tetryonic topologies

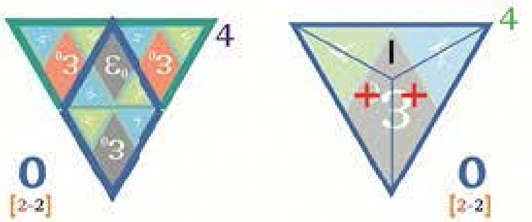

It is conceivable that given the right conditions (ie a cloud of neutral Tetryons), that in the absence of charged Tetryons to interact with, a Glueball topology can be formed entirely from neutral Tetryons

Note: Despite their total neutral charge
Glueball topologies are polarised

Neutrinos and some exotic neutral charge Baryons
can be considered to be 'Gluon topologies'

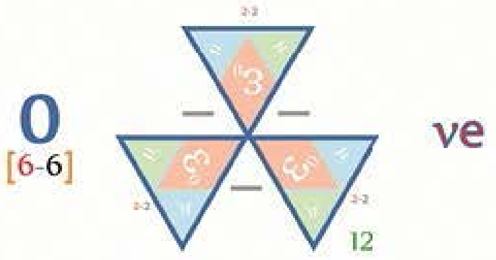

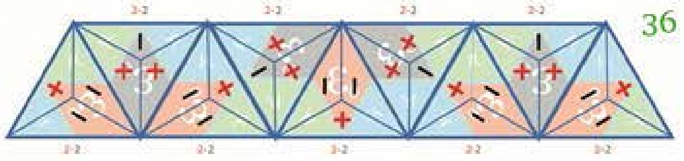

It is possible to create a family of Baryonic particles comprised entirely of neutral Tetryons

0
[18-18]

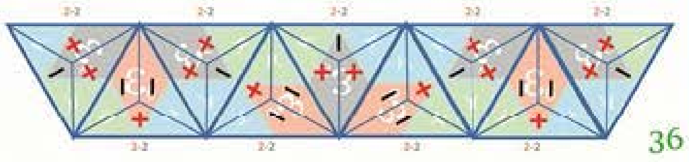

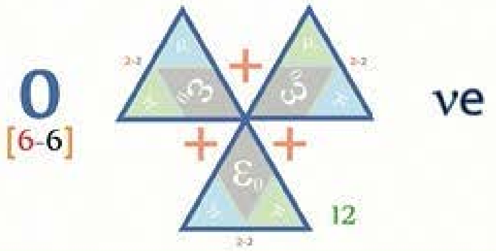

Neutrons are not considered to be Glueballs
as they contain charged Tetryons

OMEGA Particle
Icosahedron Glueball

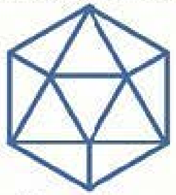

20 Neutral Teryons

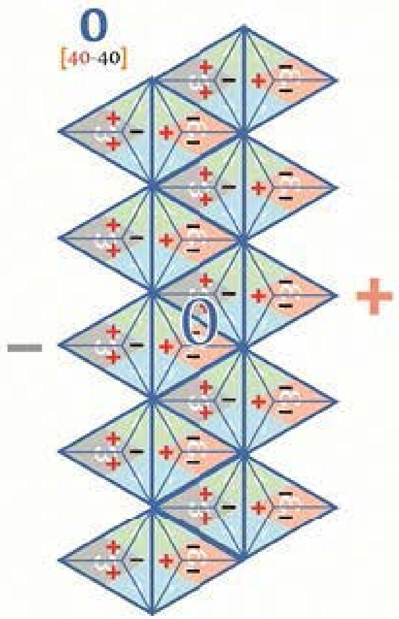

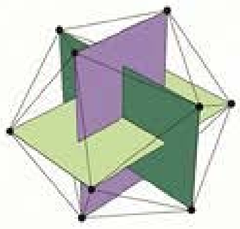

The 3 internalised planes formed by
the external apex points correspond with
the 3 spatial dimensions of Cartesian geometry

Non-gluonic Baryon formation

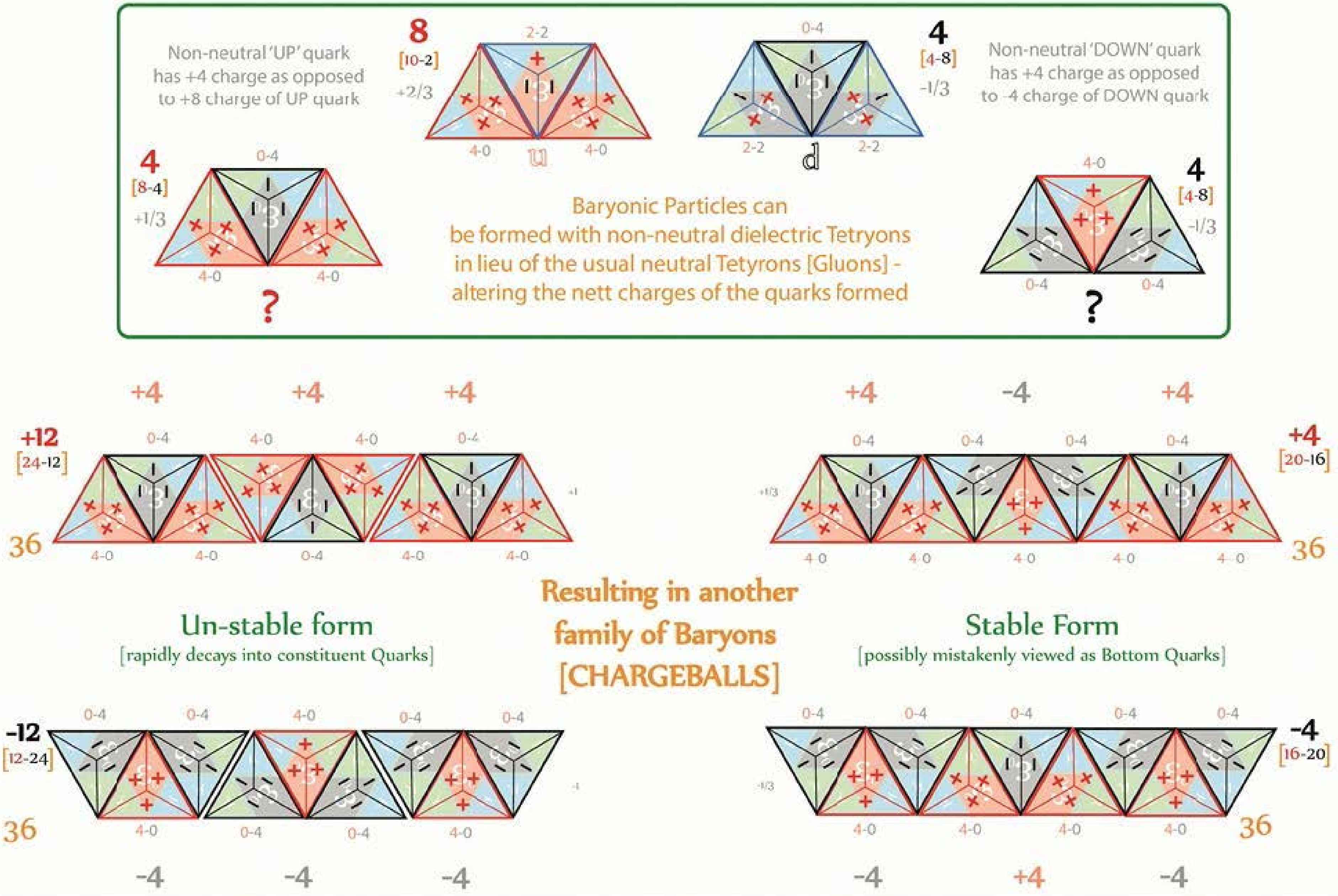
8
[10-2]
+2/3
2-2
4-0
u
4-0
0-4
4
[4-8]
-1/3
2-2
d
2-2
4
[8-4]
+1/3
0-4
4-0
4-0
?
4-0
4
[4-8]
-1/3
0-4
0-4
?
+4
+4
+4
+12
[24-12]
36
0-4
4-0
4-0
0-4
4-0
4-0
0-4
4-0
4-0
+4
-4
+4
+4
[20-16]
36
0-4
0-4
0-4
0-4
4-0
4-0
4-0
4-0
4-0
-12
[12-24]
36
0-4
0-4
4-0
0-4
0-4
4-0
0-4
0-4
4-0
-4
-4
-4
-4
[16-20]
36
0-4
0-4
0-4
0-4
0-4
4-0
4-0
4-0
4-0
-4
+4
-4

Non-neutral 'UP' quark
has +4 charge as opposed
to +8 charge of UP quark

Non-neutral 'DOWN' quark
has +4 charge as opposed
to -4 charge of DOWN quark

Baryonic Particles can
be formed with non-neutral dielectric Tetryons
in lieu of the usual neutral Tetyrons [Gluons] -
altering the nett charges of the quarks formed

Un-stable form
[rapidly decays into constituent Quarks]

Resulting in another family of Baryons
[CHARGEBALLS]

Stable Form
[possibly mistakenly viewed as Bottom Quarks]

Electro-Magnetic fields

Quantised Charge
is the nett geometry of mass-Energy-Matter formed by equilateral ElectroMagnetic fields and can be modelled with classical vector flux rotations

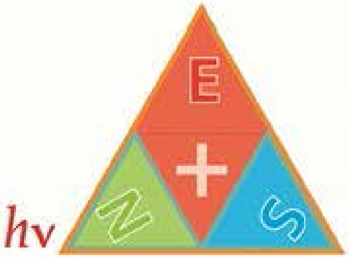

Positive Charge ZPF

Magnetic Monopoles do NOT exist

Negative Charge ZPF

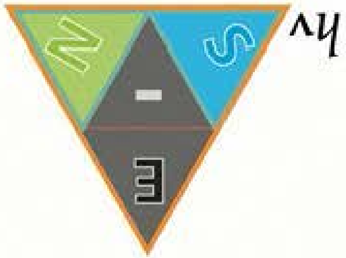

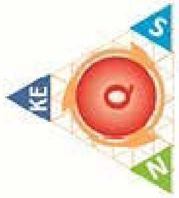

Charge

All mass-Energy-Matter EM field interactions can be modelled using ZPFgeometries

Current

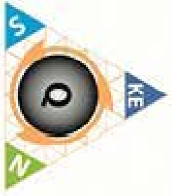

ε_0 The charge geometry of all EM fields determines their Electric Permittivity and Magnetic Permeability μ_0

The 'zero point' field has scalar EM energies

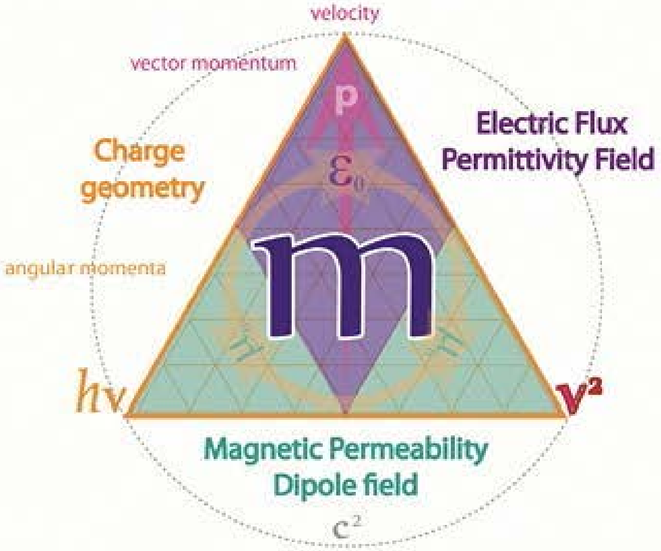
velocity
vector momentum
Charge geometry
Electric Flux Permittivity Field
angular momenta
p
ε0
m
hv
v²
Magnetic Permeability Dipole field
c²

The EM field is a equilateral waveform with the Magnetic field always orthogonal to the Electric

The Magnetic fields propogate bi-directionally and the Electric field is responsible for producing linear momentum

The Electric field and Magnetic fields are equal to each other and directly proportional to the velocity of propagation

Photons
are the neutral quanta of EM waves
they are comprised of two opposite charge EM field quanta

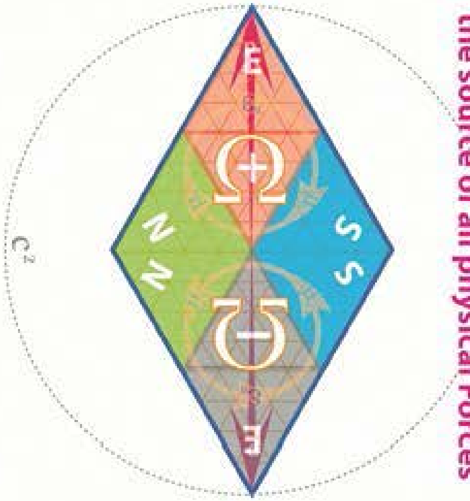

Electro-Magnetic fields are the source of all physical Forces

Zero Point fields are polarised and are the sources of

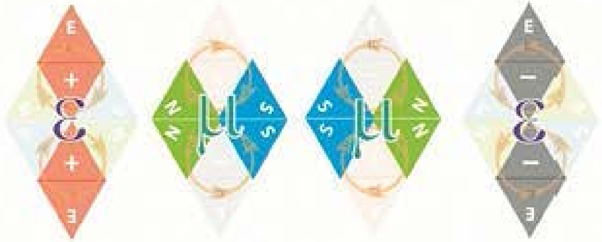
E
+
ε
+
E
N
μ
S
S
μ
N
E
-
ε
-
E

Electro-static and **Magneto-static** fields and particles

Copyright ABRAHAM [2008] - all rights reserved

The Weak Force

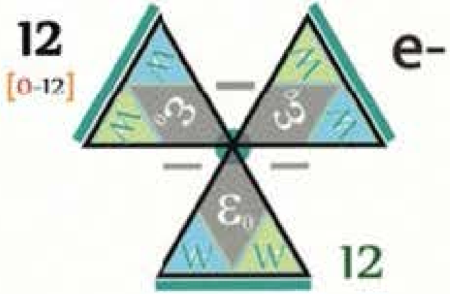

6 Inductive 'Weak' edges
(3 internalised)
(3 externalised)

Quantum Energy is stored or released, and distributed throughout Tetryonic geometries in ODD number quanta [Bosons]

Quantum Inductive loops

Quantum Tank circuit

Energy received is stored in quantum inductive loops (until its release via weak interaction - Boson exchange)

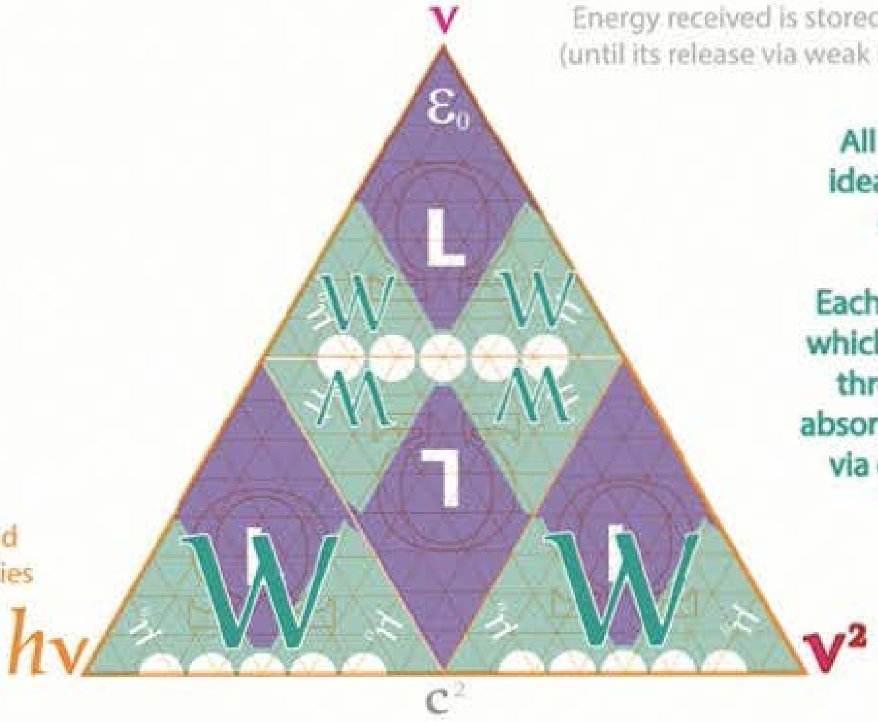

All ElectroMagnetic fields are ideal quantum inductive loops of EM energy momenta

Each has a Magnetic edge (base) which acts as a quantum inductor through which energy can be absorbed or released in EM quanta via electro**Magnetic** induction [exchange Bosons]

Leptons and Quarks have the same Tetryonic numbers but differing Matter topologies

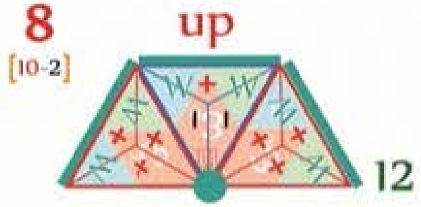

6 Inductive 'Weak' edges

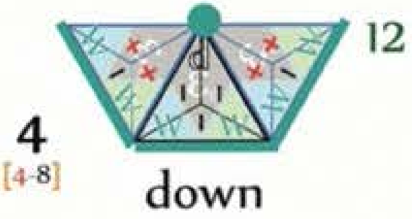

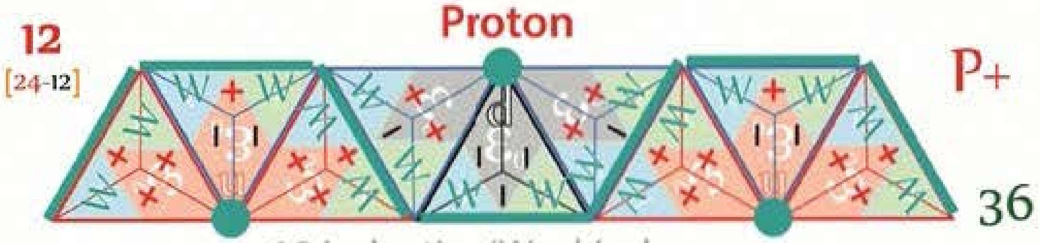

18 Inductive 'Weak' edges

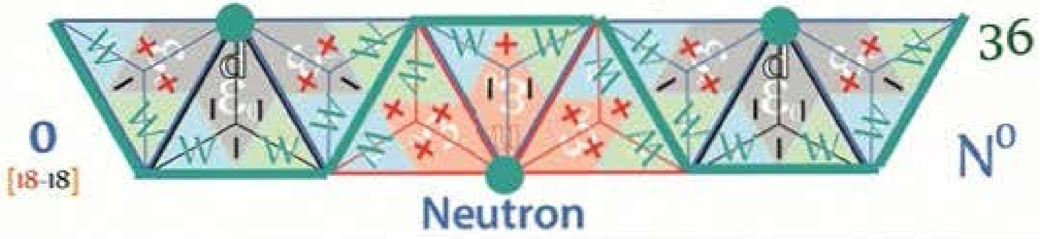

Copyright ABRAHAM [2008] - all rights reserved

The Strong Force

The force resulting from the interaction of 2 Tetryons along their parallel planar (Electric charge) surfaces

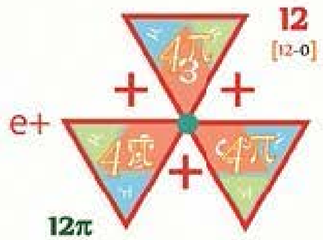
12
[12-0]
e+
+
+
+
4π
4π
4π
12π

Leptons
Repulsive colour Force

Leptons are 'held together' by a radial Weak Force axis

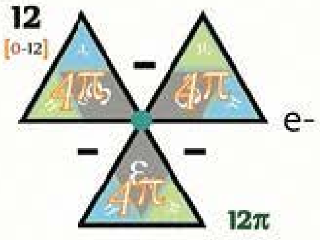
12
[0-12]
e-
-
-
-
4π
4π
4π
12π

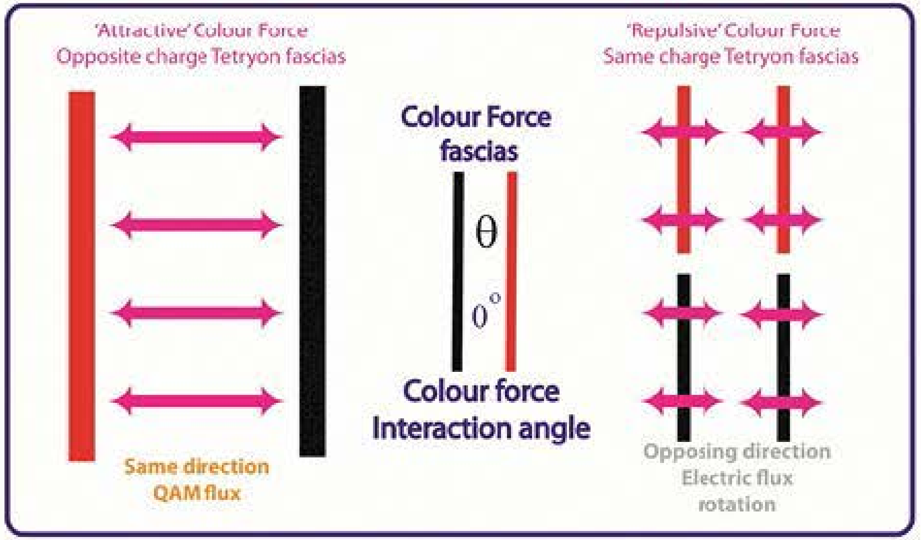
'Attractive' Colour Force
Opposite charge Tetryon fascias
'Repulsive' Colour Force
Same charge Tetryon fascias
Colour Force fascias
θ
0°
Colour force Interaction angle
Same direction QAM flux
Opposing direction Electric flux rotation

Strong Force strength is directly proportional to the nett mass-energy quanta involved
(Increases/decreases as total Tetryon mass-energy quanta increases/decreases)

Leptons and Quarks result from the Strong 'colour' force interactions in dodeca-deltahedrons

(In Quarks the attractive strong force produces Octahedral topologies)
(In Leptons the repulsive strong force produces Dodecahedral topologies)

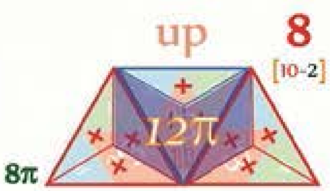
up
8
[10-2]
12π
8π

Quarks
4 Fascia 'Colour' Attraction

Quarks
Attractive colour Force

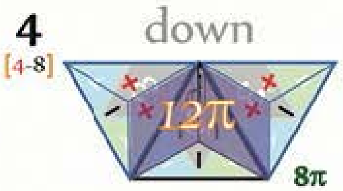
4
[4-8]
down
12π
8π

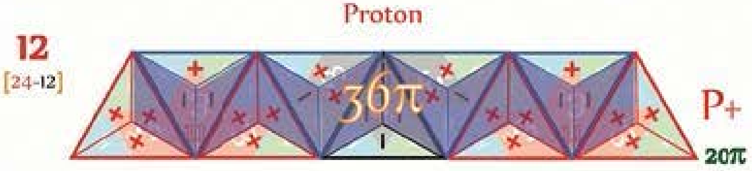
Proton
12
[24-12]
36π
P+
20π

Baryons
Attractive colour Forces

Baryon formation
16 Fascia 'Colour' Attraction

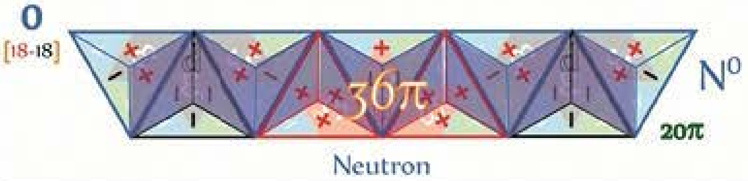
0
[18-18]
36π
N⁰
20π
Neutron

The Strong Colour Force is the binding force that holds Nuclei together

Tetryonic Fascia (Planar Electric Charge) interaction

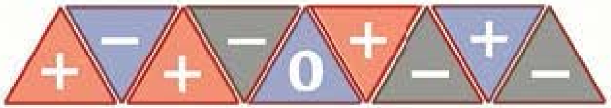

It interacts through oppositely charged Tetryon Fascias in contact with each other in all Matter (save Leptons)

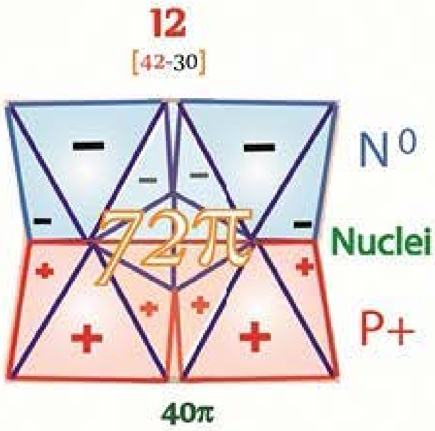
12
[42-30]
N⁰
Nuclei
P+
72π
40π

Strong Force Interaction

On the smallest scale (less than about 0.8 fm), it is the Electric force (mediated by gluons in nuclei) that holds Tetryons together in order to form Quarks, Protons and Neutrons

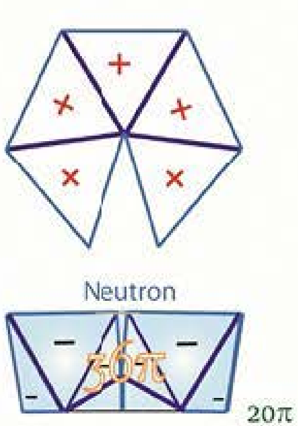
0
[18-18]
Neutron
36π
20π

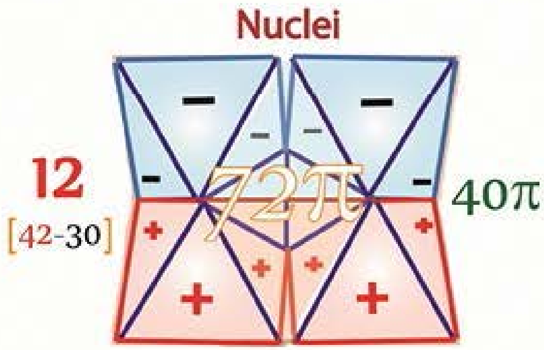
Nuclei
12
[42-30]
72π
40π

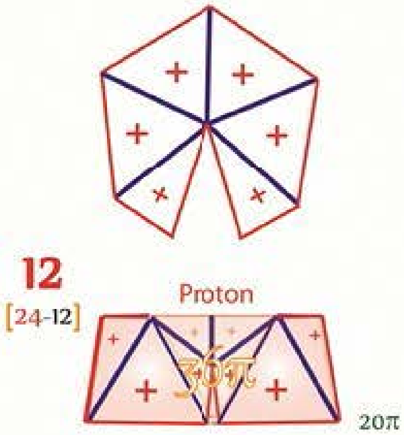
12
[24-12]
Proton
36π
20π

The residual Electric field force produced by the Strong Force is also facilitates the binding of Protons and Neutrons together to form the nucleus of an atom

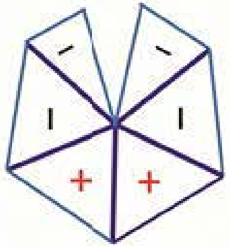

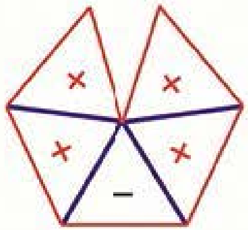

Attractive Colour Charge Force holding Quarks and Nuclei together

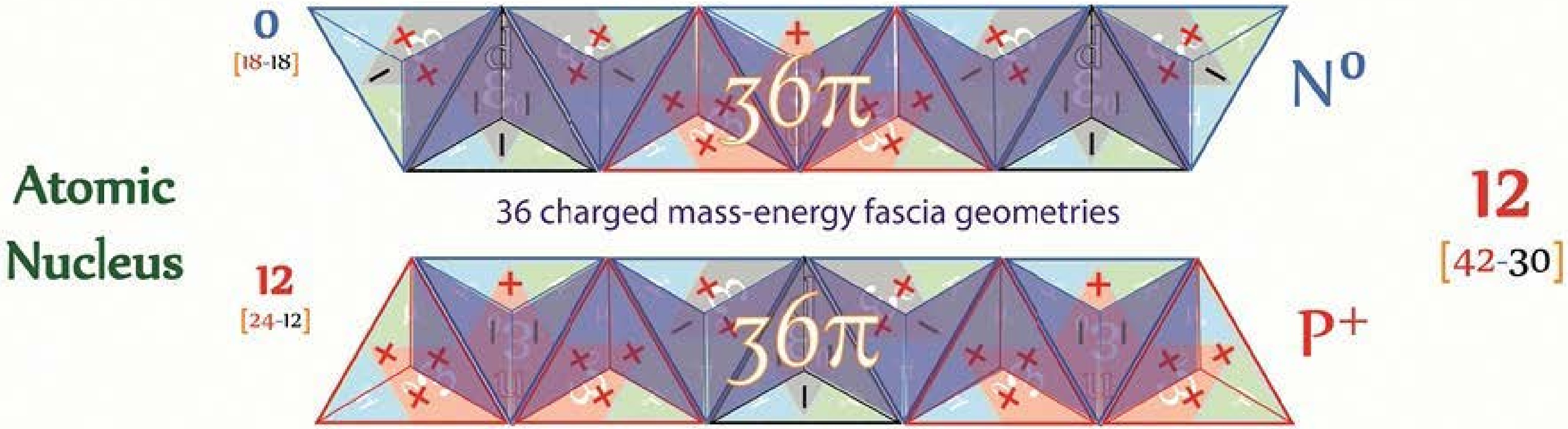
0
[18-18]
36π
N0
Atomic Nucleus
12
[42-30]
12
[24-12]
36π
P+

36 charged mass-energy fascia geometries

creates an external Baryons topology with 20 charge fascia

Proton
Hydrogen ion
36

1 Proton	24-12	
Total Charge	24-12	[+12]

H^+
F+ n=25

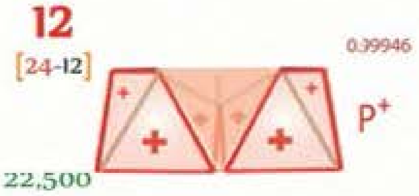

Proton - Hydrogen ion

The Proton is a Baryon with a Positive Tetryonic charge geometry of +12 [24-12] [and a final Matter topology of 20π]

It attracts electrons [0-12] and Neutrons [13-18] through their opposite charged facia in order to reach charge equilibrium

The Proton's opposite charge particle is the electron
If it binds with one or more Neutrons it will still attract electrons in order to acheive Tetryonic charge equilibrium.

As a Hydrogen nuclei it is highly reactive

Neutral
Hydrogen
48

1 Proton	24-12	
1 electron	0-12	
Total Charge	24-24	[0]

H^0
F+ n=25
e- n=1

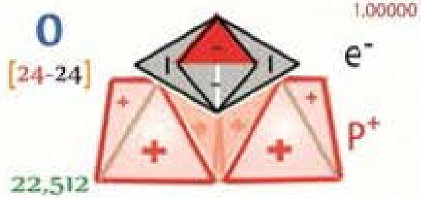

Neutral Hydrogen

Hydrogen is the Second Neutral Tetryonic geometry [Baryon] formed in the creation of Matter

It is a [24-24] NEUTRAL [balanced] Tetryonic charge geometry

The Neutral Hydrogen Atom interacts with Deuterium and other Elements via its external nucleon fascia charges (residual EM Interactions)

Deuterium
84

1 Proton	24-12	
1 Neutron	18-18	
1 electron	0-12	
Total Charge	42-42	[0]

D 1
N° n=25
F+ n=25
e- n=1

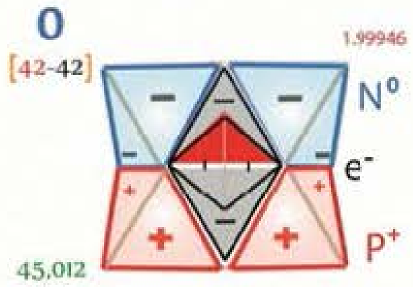

Deuterium

Deuterium has the same nett charge as Neutral Hydrogen but has an increased Tetryonic mass- Matter topology

It has an additional 9 Tetryons in its nucleus (due to the Neutron) and consequently is larger than the Hydrojon nucleus

It has a [42-42] NEUTRAL Tetryonic charge geometry
Deuterium nuclei combine to form all other elements

Tritium
84

1 Proton	24-12	
1 Neutron	18-18	
1 electron	0-12	
Total Charge	42-42	[0]

D 7
N° n=31
F+ n=31
e- n=7
Radioactive Isotope

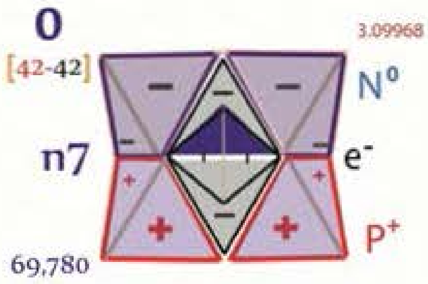

Tritium

(Radioactive Deuterium)

Tritium has the same nett Tetryonic charge as Neutral Hydrogen and Deuterium (D) but due to its higher quantum energy levels it has additional mass-energy quanta in its charged geometry

It has the same 84 charged fascia in its nucleus as Deuterium but its increased mass-energy is equivalent to that of 1 Neutron.

This extra nuclear energy is the source of its radioactivity and ongoing confusion in chemistry where it is released as a variety of decay particles over time

Copyright ABRAHAM [2008] - all rights reserved

Hydrogen - Helium genesis

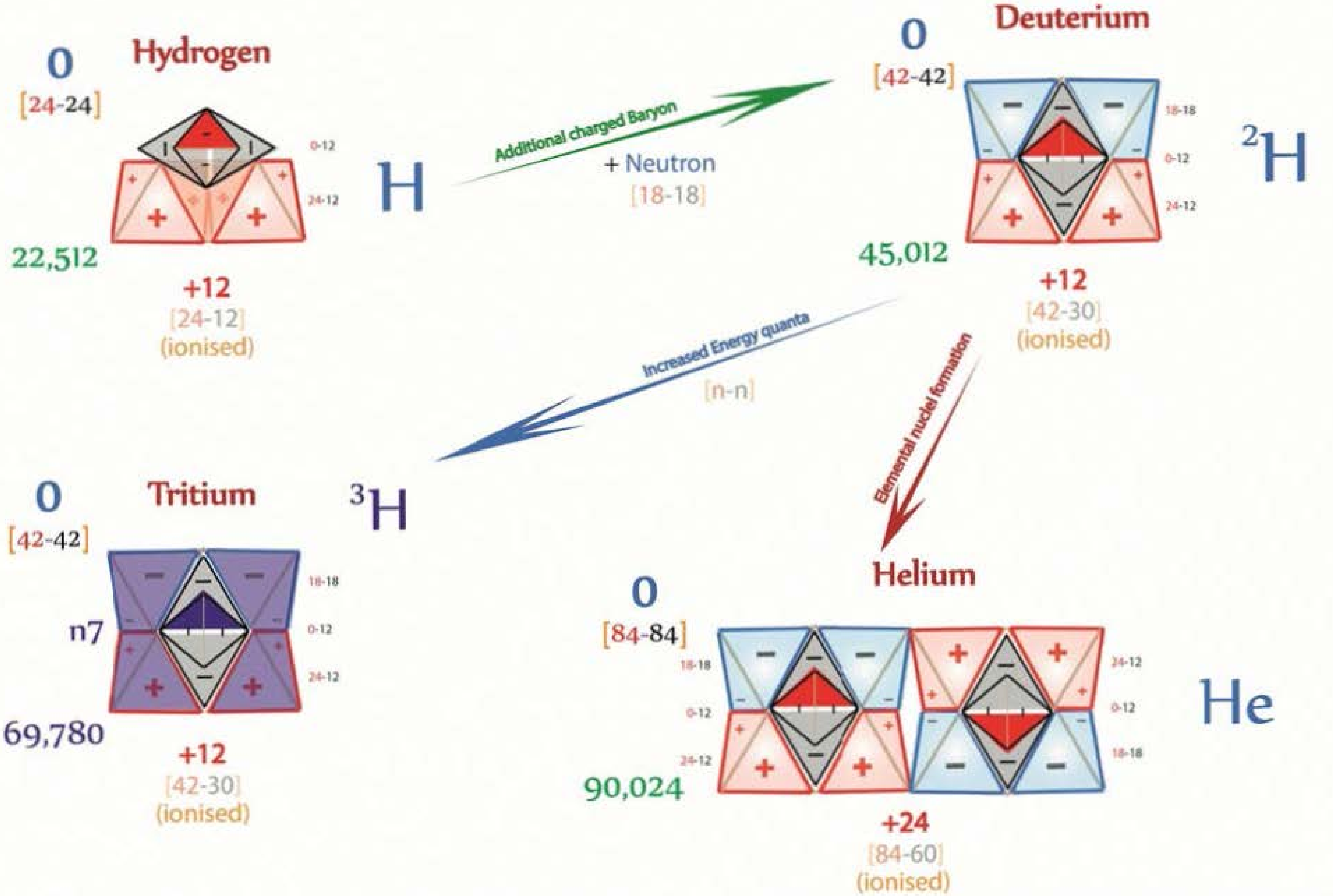

Hydrogen

0

[24-24]

H

22,512

+12

[24-12]

(ionised)

Additional charged Baryon

+ Neutron

[18-18]

Deuterium

0

[42-42]

^{2}H

45,012

+12

[42-30]

(ionised)

Increased Energy quanta

[n-n]

Elemental nuclei formation

Tritium

0

[42-42]

^{3}H

n7

69,780

+12

[42-30]

(ionised)

Helium

0

[84-84]

He

90,024

+24

[84-60]

(ionised)

0-12
24-12
18-18
0-12
24-12
18-18
0-12
24-12
18-18
0-12
24-12
24-12
0-12
18-18

Tetryonics 13.02 - Hydrogen-Helium genesis

Hydrogen Family

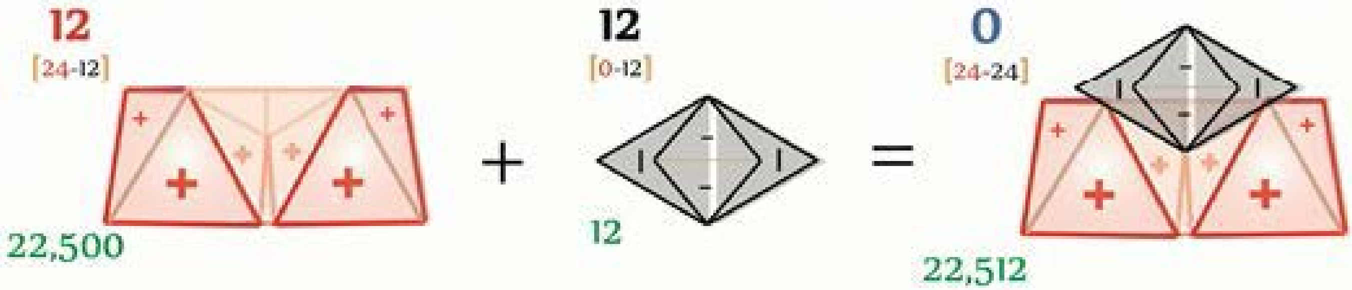
12
[24-12]
22,500
+
12
[0-12]
12
=
0
[24-24]
22,512

Proton	electron	Hydrogen
36π	12π	48π

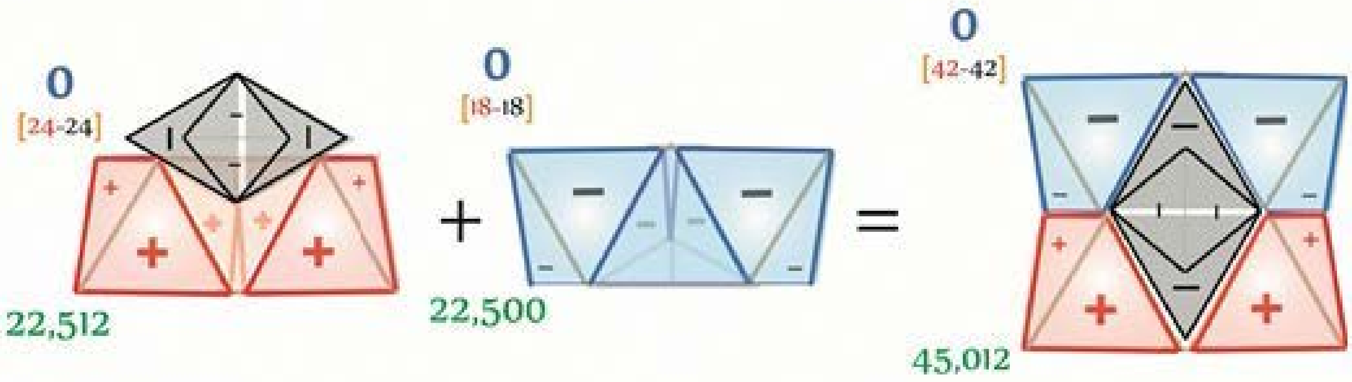
0
[24-24]
22,512
+
0
[18-18]
22,500
=
0
[42-42]
45,012

Hydrogen	Neutron	Deuterium
48π	36π	84π

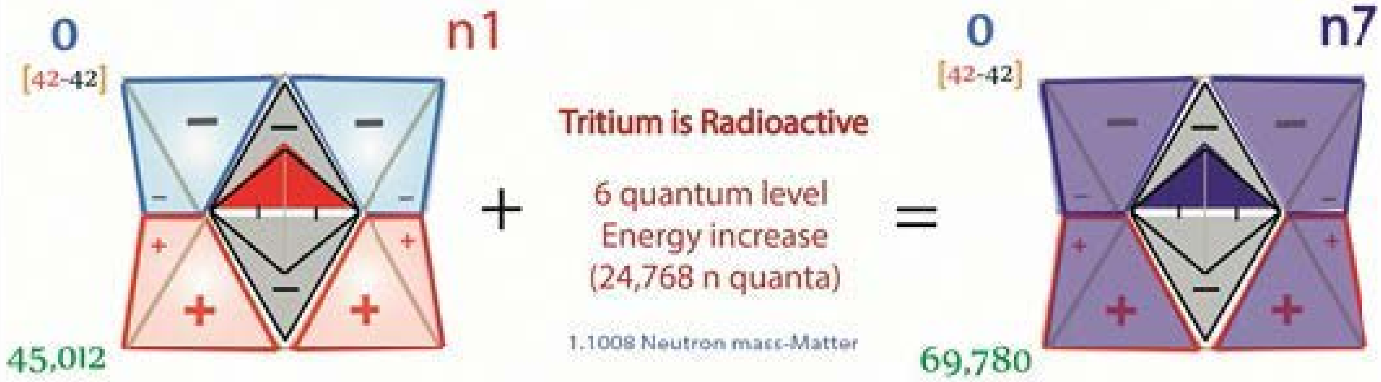
0
[42-42]
n1
45,012
+
=
0
[42-42]
n7
69,780

Tritium is Radioactive

6 quantum level
Energy increase
(24,768 n quanta)

1.1008 Neutron mass-Matter

Deuterium	Tritium
84π	84π

Copyright ABRAHAM [2008] - all rights reserved

Copyright ABRAHAM [2008] - all rights reserved

Hydrogen-Helium3 genesis

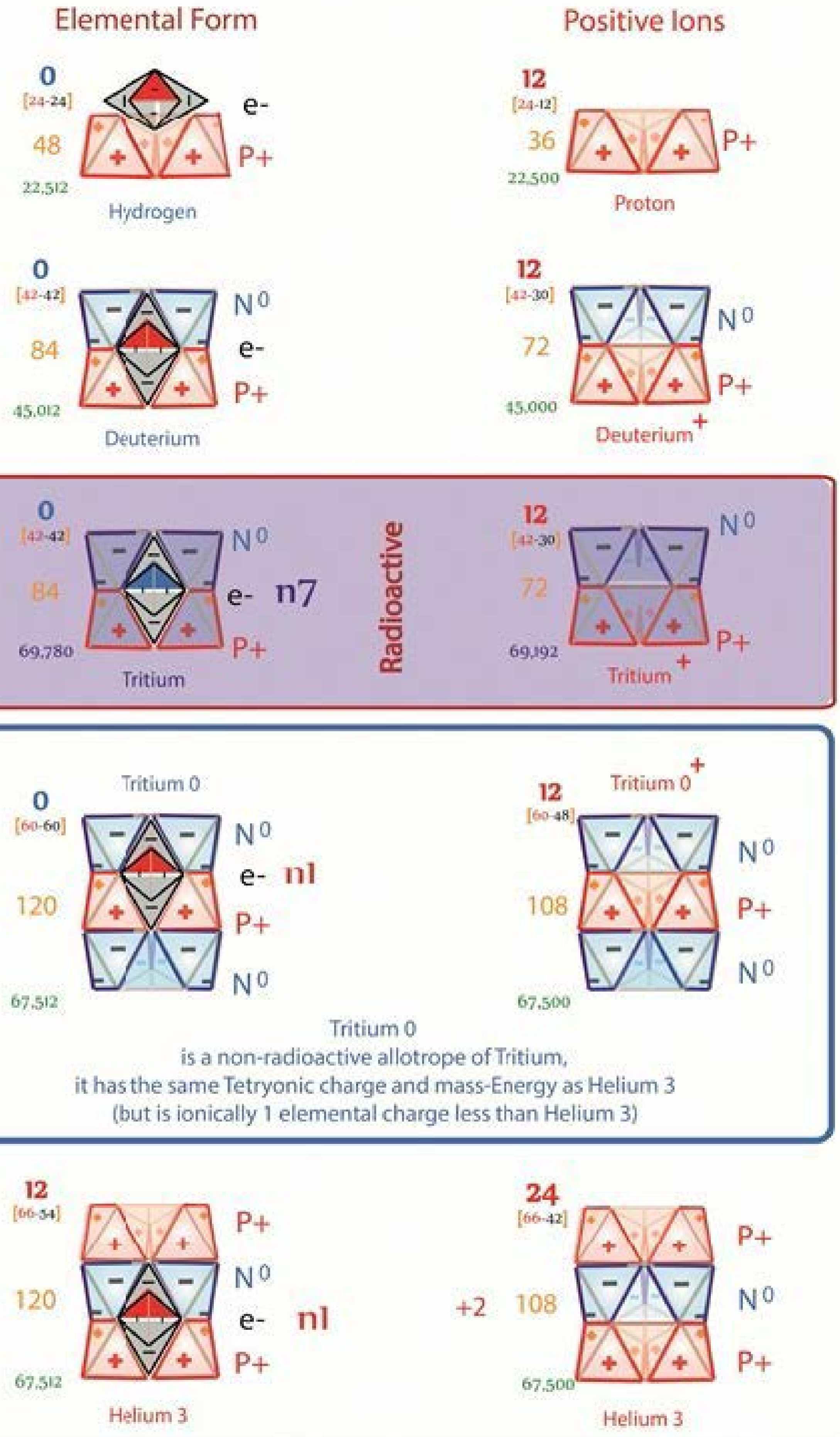
Elemental Form
Positive Ions
0
[24-24]
48
22,512
e-
P+
Hydrogen
12
[24-12]
36
22,500
P+
Proton
0
[42-42]
84
45,012
N^0
e-
P+
Deuterium
12
[42-30]
72
45,000
N^0
P+
Deuterium+
0
[42-42]
84
69,780
N^0
e-
n7
P+
Tritium
Radioactive
12
[42-30]
72
69,192
N^0
P+
Tritium+
Tritium 0
0
[60-60]
120
67,512
N^0
e-
n1
P+
N^0
Tritium 0+
12
[60-48]
108
67,500
N^0
P+
N^0
+1
12
[66-54]
120
67,512
P+
N^0
e-
n1
P+
Helium 3
+2
24
[66-42]
108
67,500
P+
N^0
P+
Helium 3

Tritium 0
is a non-radioactive allotrope of Tritium,
it has the same Tetryonic charge and mass-Energy as Helium 3
(but is ionically 1 elemental charge less than Helium 3)

Copyright ABRAHAM [2008] - all rights reserved

Quantum Batteries

Atomic nuclei can be easily scaled to non-quantum sizes to offer clean, safe and portable long term energy storage devices that can store energy indefinitely and release it on demand anywhere in the World

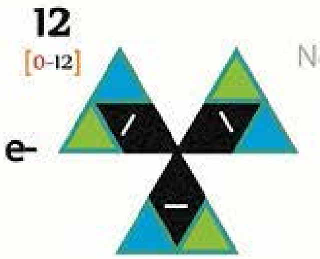
12
[0-12]
e-
Negative charge topology
Quantum Rotor

12 loop quantum inductive rotor

The quantum battery is unique in that in addition to storing energy indefinitely, when the nuclei combine with a lepton it has the ability to release specific energies [photons] by way of its synchronous quantum convertor topology

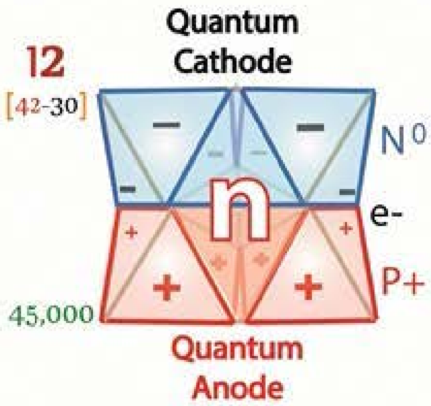
Quantum Cathode
12
[42-30]
N0
n
e-
P+
45,000
Quantum Anode

Energy stored in quantum batteries is measured as mass

(Atomic Nuclei)

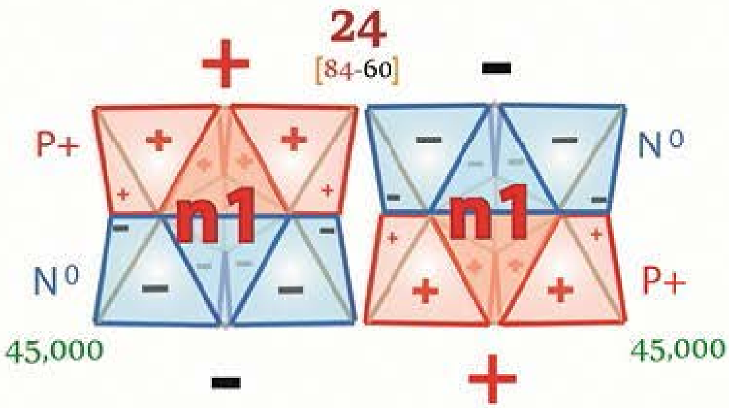
24
[84-60]
+
-
P+
n1
n1
N0
N0
P+
45,000
45,000
-
+

anti-Parallel Configuration

The non-neutral charge of atomic nuclei attract free leptons into 'bound' states within the various **n** levels of atomic shells releasing energy as spectral photons

Series Configuration

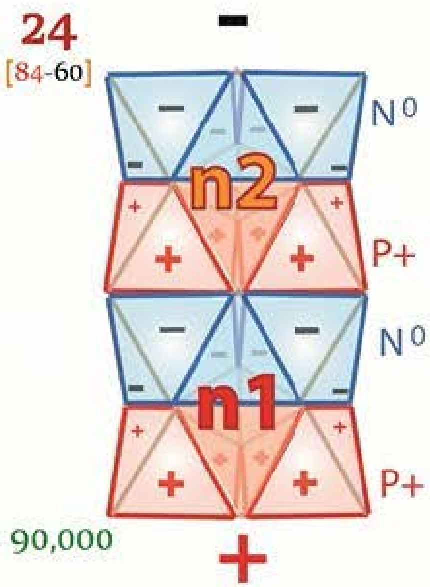
24
[84-60]
-
N0
n2
P+
N0
n1
P+
90,000
+

Quantum synchronous Converters

With the addition of a quantum rotor (lepton) a quantum battery can be converted from a storage device into a energy distribution device.

And just like the quantum battery, the quantum convertor can be scaled to any size in order to provide tailor-made energy efficient delivery devices

The electron can be viewed as
a rotating inductor consisting of
3 negative Tetryons

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]^{n8}_{n1}$$

Leptons ElectroMagnetic mass velocity

The electron has a charged
Matter topology that is electrically
equivalent to a quantum 6 loop inductive rotor

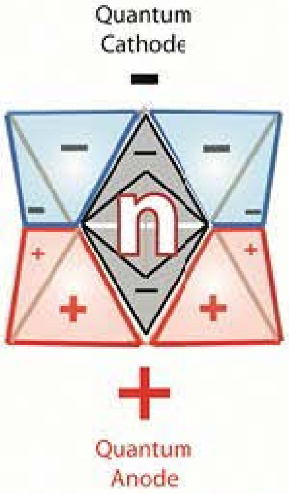

EM Field Planck quanta

$$72\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]^{n32}_{n25}$$

Nuclei ElectroMagnetic mass velocity

EM Field Planck quanta

$$84\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]^{n32}_{n25}$$

Deuterons ElectroMagnetic mass velocity

***Where varying levels and frequencies of Energy
are transmitted long distances and
need to be stored for later release on demand
the 'ideal' mechanical device is
the rotating (or synchronous) convertor***

Changes in energy-momenta results in photon emission/absorption lines
photon emission/absorption produces changes in leptronic energy-QAM
and results in the quantum transition of electrons in atomic orbitals

Changes in Baryonic energy levels induces a directly
proportional change in electron energy levels

3 forms of mass-energy momenta are possessed by quantum convertors

Angular Momentum (motional energy)
linear momentum of Baryons & leptons

Photonic spectral energies (emission/absorption)
KEM fields geometries of photo-electrons

Nuclear mass-energies (stored energy quanta)
standing wave mass-energies of Matter topologies

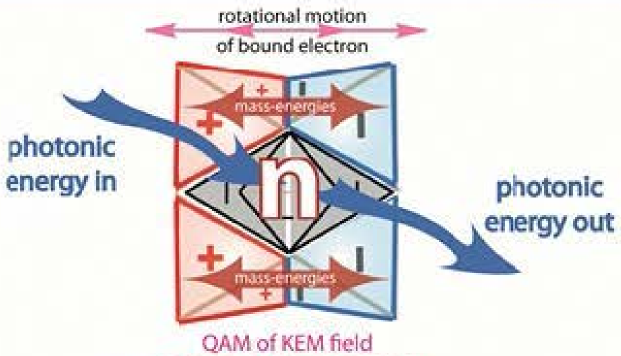

QAM of KEM field
creates angular momentum

A change in any 1 of the 3 types
of energy in a atom results
in a proportional change in the other 2

*External EM fields and incident photons can all
affect the quantum energy levels of the atomic nuclei*

Hydrogen Ionisation Energies

$Mv^2 = KEM = hv^2$

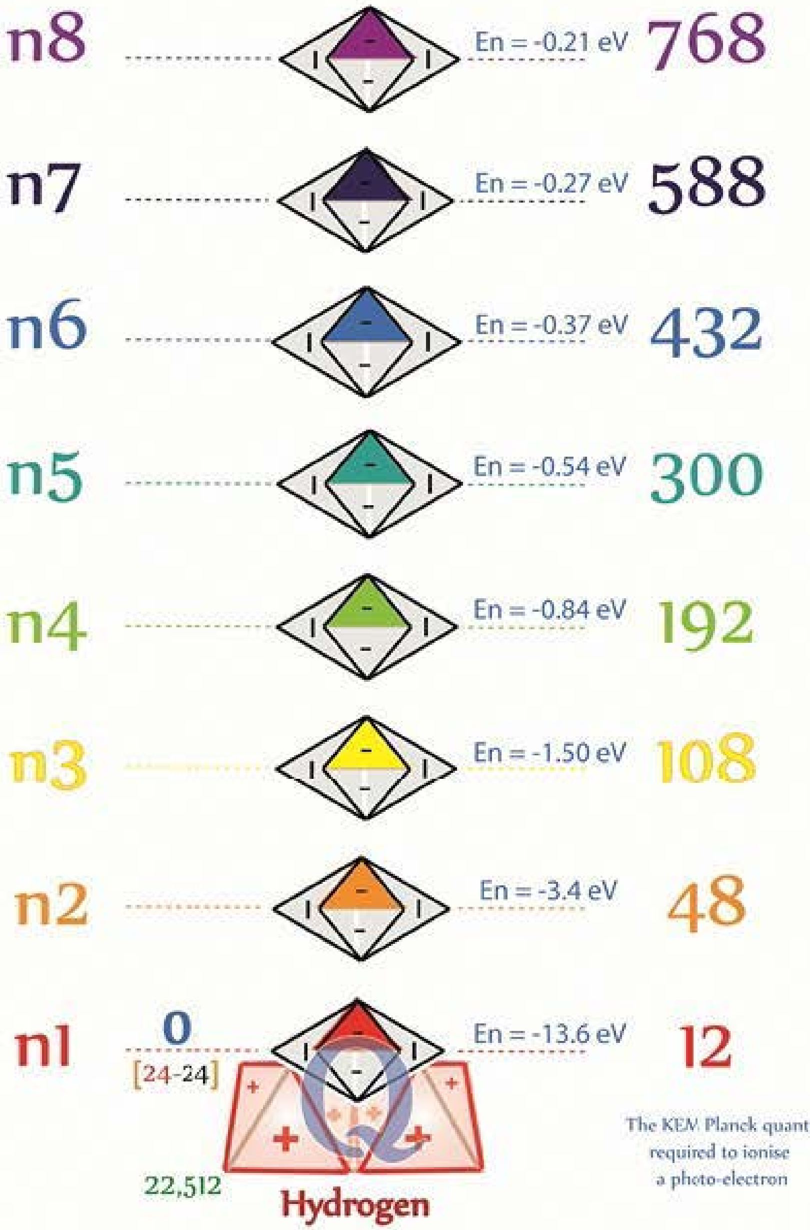
n8
En = -0.21 eV
768
n7
En = -0.27 eV
588
n6
En = -0.37 eV
432
n5
En = -0.54 eV
300
n4
En = -0.84 eV
192
n3
En = -1.50 eV
108
n2
En = -3.4 eV
48
n1
0
[24-24]
En = -13.6 eV
12
22,512
Hydrogen

The KEM Planck quanta required to ionise a photo-electron

The ionisation energy level for each quantum level is proportional to the square of the quantum number

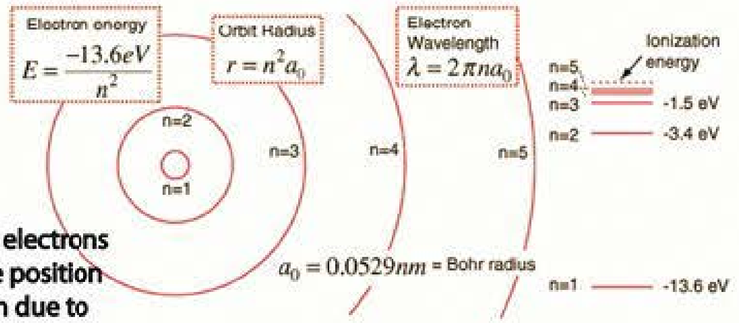
Electron energy
$E = \frac{-13.6eV}{n^2}$
Orbit Radius
$r = n^2 a_0$
Electron Wavelength
$\lambda = 2\pi n a_0$
n=1
n=2
n=3
n=4
n=5
$a_0 = 0.0529nm$ = Bohr radius
Ionization energy
n=5
n=4
n=3
n=2
n=1
-1.5 eV
-3.4 eV
-13.6 eV

In Hydrogen nuclei electrons remain in the same position close to the Proton due to Coulombic attraction.

The established model of electron orbitals having circular radii of increasing size in proportion to their energies is incorrect

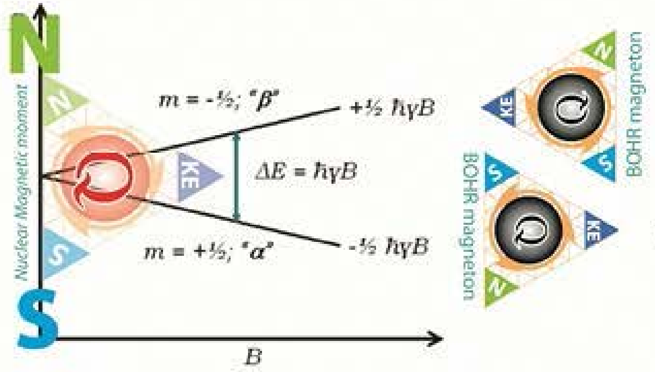
N
S
Nuclear Magnetic moment
KE
$m = -½$; "β"
$+½\ \hbar\gamma B$
$\Delta E = \hbar\gamma B$
$m = +½$; "α"
$-½\ \hbar\gamma B$
B
BOHR magneton
BOHR magneton

Spin UP
Parallel magnetic moments
(higher energy)

Spin DOWN
Anti-parallel magnetic moments
(lower energy)

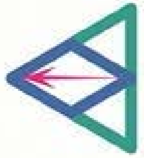

S
up

The intrinsic magnetic moment of the rotating electron couples wrt the nuclear magnetic moment to produce a split in the resulting spectral lines produced by quantum level transitions

S
down

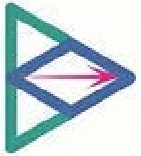

Copyright ABRAHAM [2008] - all rights reserved

Hydrogen Energy Levels

KEM field energy in each electron's energy level

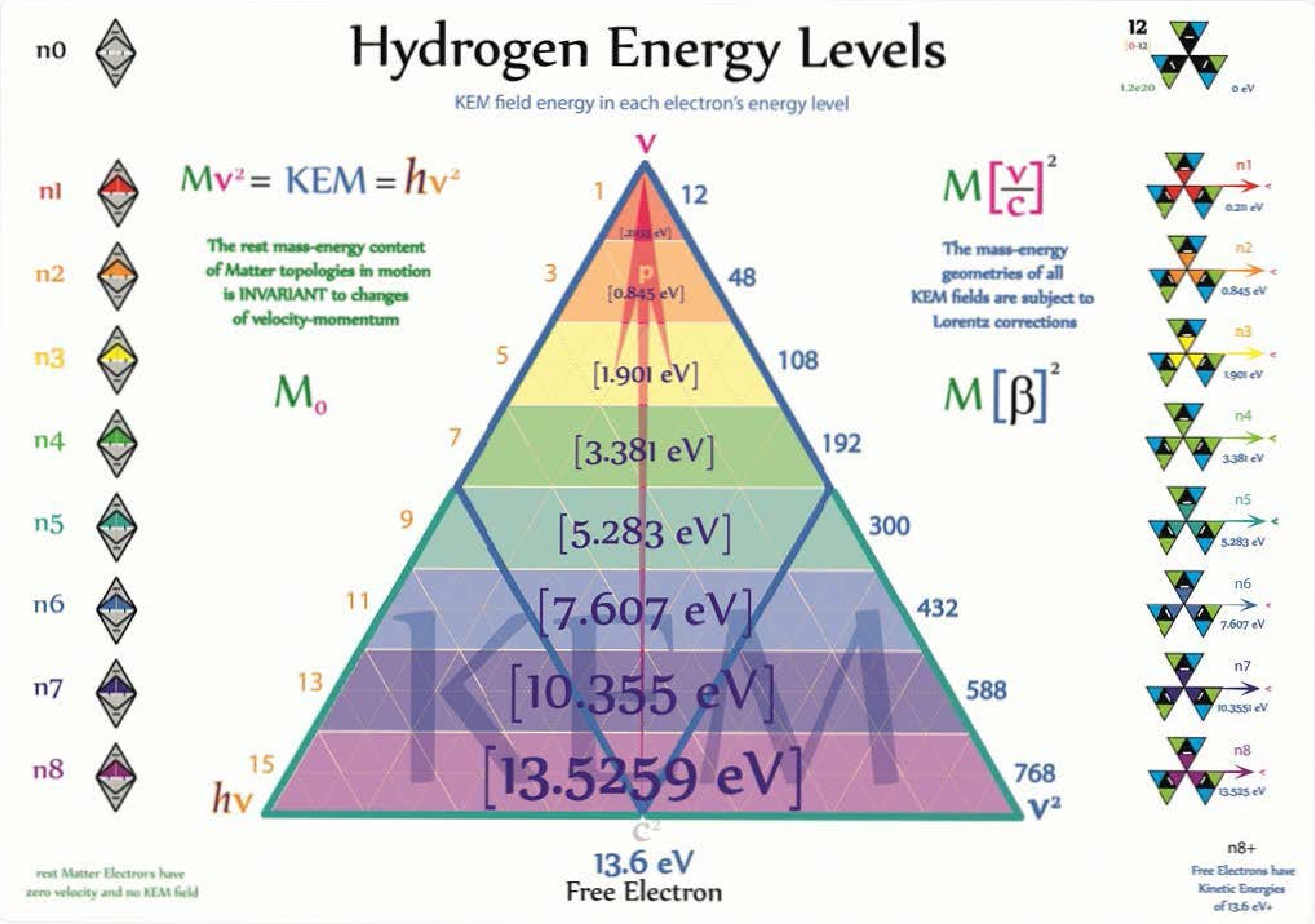
n0
n1
n2
n3
n4
n5
n6
n7
n8
12
0-12
1.2e20
0 eV
v
1
12
p
3
48
[0.845 eV]
5
108
[1.901 eV]
7
192
[3.381 eV]
9
300
[5.283 eV]
11
432
[7.607 eV]
13
588
[10.355 eV]
15
768
[13.5259 eV]
KEM
hv
v^2
c^2
n1
0.211 eV
n2
0.845 eV
n3
1.901 eV
n4
3.381 eV
n5
5.283 eV
n6
7.607 eV
n7
10.3551 eV
n8
13.525 eV
n8+

$Mv^2 = KEM = hv^2$

The rest mass-energy content
of Matter topologies in motion
is INVARIANT to changes
of velocity-momentum

M_0

$M\left[\frac{v}{c}\right]^2$

The mass-energy
geometries of all
KEM fields are subject to
Lorentz corrections

$M[\beta]^2$

rest Matter Electrons have
zero velocity and no KEM field

13.6 eV
Free Electron

Free Electrons have
Kinetic Energies
of 13.6 eV+

Tetryonics 13.08 - Hydrogen Energy Levels

Copyright ABRAHAM [2008] - all rights reserved

Kinematics

An inertial frame of reference is one in which the motion of a particle not subject to forces and results in motion in a straight line at constant velocity

Deceleration

[Inertial Frame MINUS odd quanta]

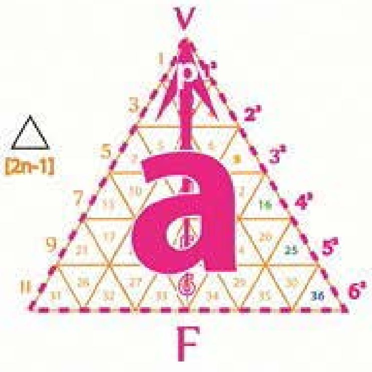

Newton's First Law

$$\sum \mathbf{F} = 0 \Rightarrow \frac{d\mathbf{v}}{dt} = 0$$

Every body persists in its state of being at rest or of moving uniformly straight forward, except insofar as it is compelled to change its state by force impressed

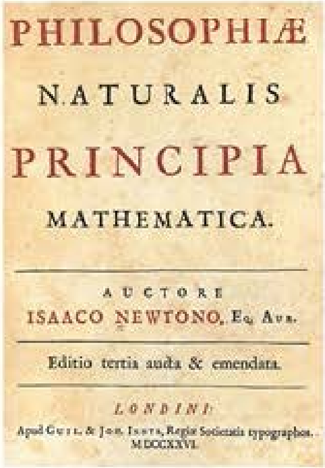

PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

AUCTORE
ISAACO NEWTONO, Eq. Aur.

Editio tertia aucta & emendata.

LONDINI:
Apud Guil. & Joh. Innys, Regiæ Societatis typographos.
MDCCXXVI.

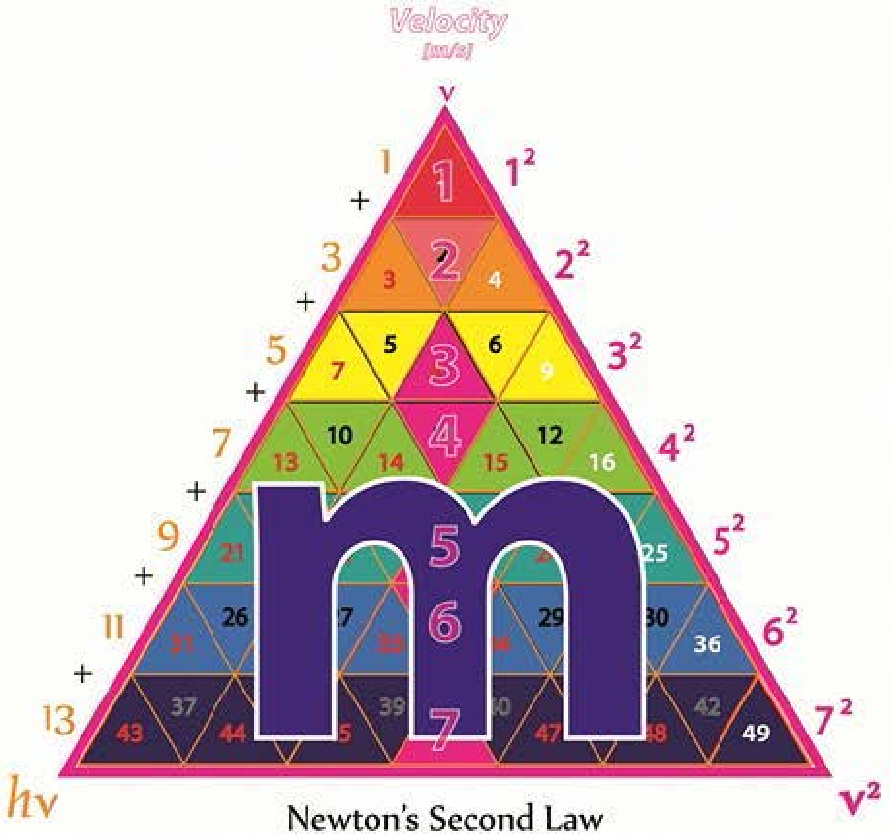

Newton's Second Law

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}_{cm}}{\mathrm{d}t} = m\mathbf{a}_{cm}$$

Force creates a change in Momentum over Time

A body of [m]ass subject to a net [F]orce undergoes an [a]cceleration that has the same direction as the force and a magnitude that is directly proportional to the force and inversely proportional to the mass, i.e., F = ma.

Alternatively, the total force applied on a body is equal to the time derivative of linear momentum of the body.

Sir Issac Newton

(25 December 1642 – 20 March 1727)

$$a = \frac{\Delta y}{\Delta x} = \frac{\Delta v}{\Delta t}.$$

$$\mathbf{F} = m\mathbf{a} \quad \longrightarrow \quad \mathbf{a} = \mathbf{F}/m$$

Acceleration

[Inertial Frame PLUS odd quanta]

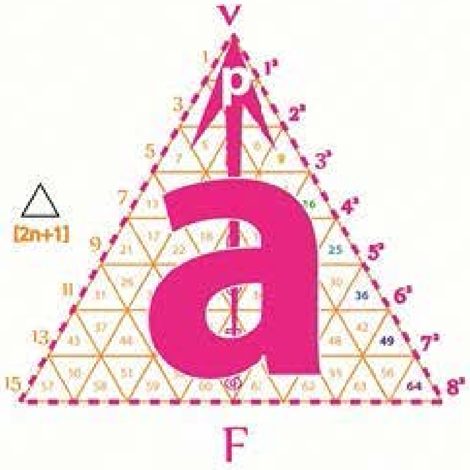

Copyright ABRAHAM [2008] - all rights reserved

Inertia and Force

Any change to the energy-momenta content of a closed inductive loop requires a proportional change to the loop's energy density

momentum

The classical definition of Momentum relates the mass of a material body at given velocity (v) to its Momentum (p); it is a proportionality factor in the formula

$$p = mv$$

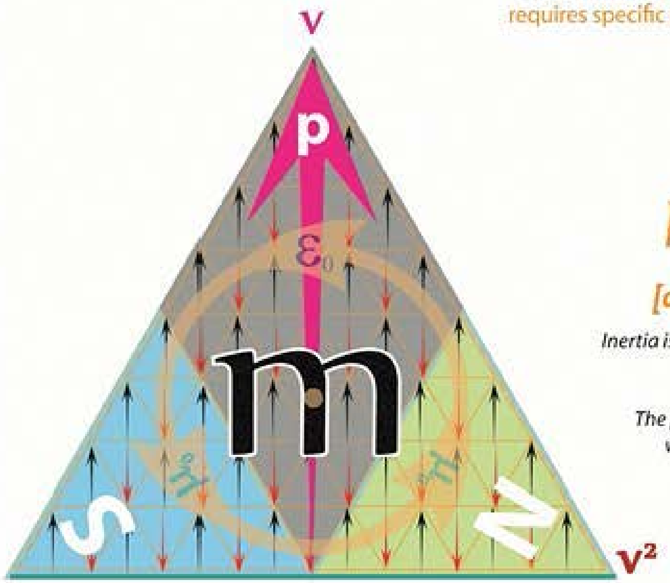

Linear momentum is the nett square root of mass-energy quanta

p

linear quanta forces

divergent ↑ [v - v] ↓ convergent

Transverse Bosons

Any change to energy-momenta levels in Tetryonic geometry requires specific number quanta

Scalar mass-energy

Inertia is resistance to change of energy momenta vectors within spatial geometries

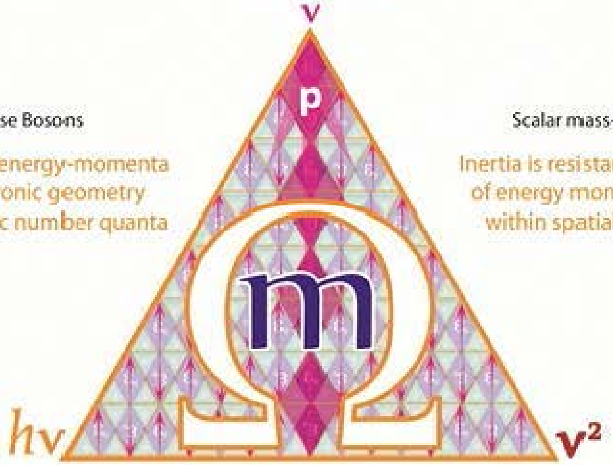

[all inductive loops resist changes to their energy levels]

Inertia is the resistance of any physical object to a change in its state of motion or rest, or the tendency of an object to resist any change in its motion.

The principle of inertia is one of the fundamental principles of classical physics which is used to describe the motion of matter and how it is affected by applied forces

$$E = mv^2$$

Any change in the energy-momenta content of a body of mass-Matter results in a proportional change in its momentum-velocity

Force

This meaning of a body's inertia therefore is altered from the classsical definition of "a tendency to maintain momentum" to a description of the measure of how difficult it is to change the momentum of a body

$$F = ma$$

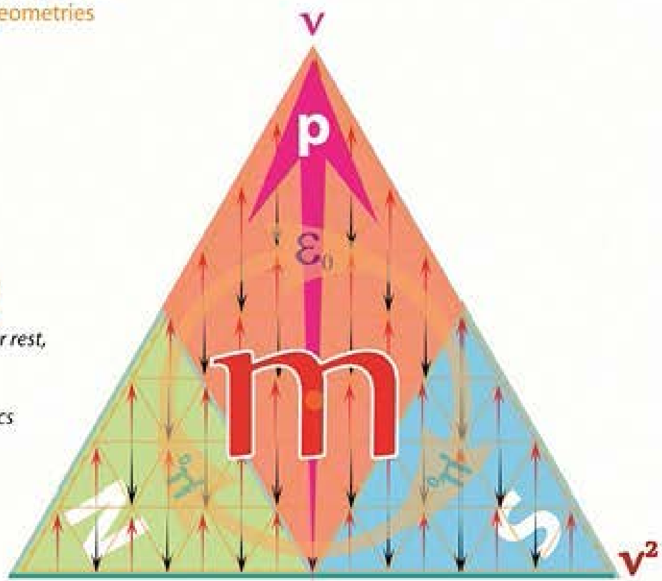

Force is the sum of the linear mass-energy momenta quanta

F

Gottfried Wilhelm von Leibniz

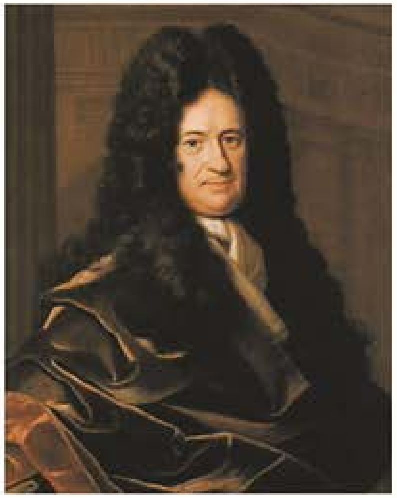

(July 1, 1646 – November 14, 1716)

Scalar Energy

Leibniz's vis viva (Latin for living force) is mv^2, twice the modern kinetic energy.

He realized that the total energy would be conserved in certain mechanical systems, so he considered it an innate motive characteristic of Matter.

His thinking gave rise to another regrettable nationalistic dispute. His vis viva was seen as rivaling the conservation of linear momentum championed by Newton in England and by Descartes in France; hence academics in those countries tended to neglect Leibniz's idea.

In reality, both energy and momentum are conserved, so the two approaches are equally valid.

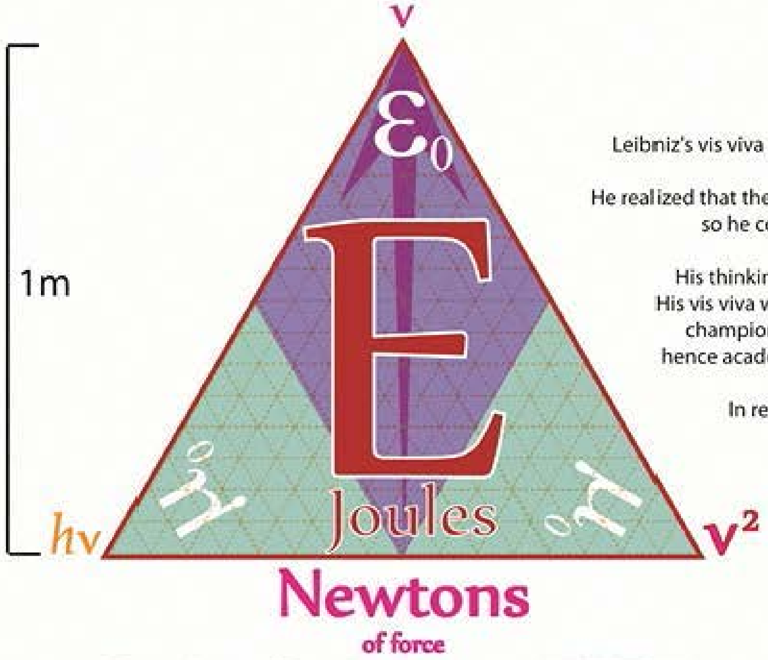

The nett direction of Force within energy geometry is UNIDIRECTIONAL
ie the force exerted is the result of the nett linear momenta irrespective of charge

mass x velocity squared $kg.\left[\frac{m}{s}\right]^2$

Energy $kg.\frac{m^2}{s^2}$

$$J = \frac{kg \cdot m^2}{s^2} = N \cdot m = Pa \cdot m^3 = W \cdot s$$

A Joule is equal to the energy expended (or work done) in applying a force of one newton through a distance of one metre (1 newton metre or N·m)

momentum x velocity $\left[kg.\frac{m}{s}\right].\frac{m}{s}$

Planck's quanta per second $\left[kg.\frac{m^2}{s}\right].s^{-1}$

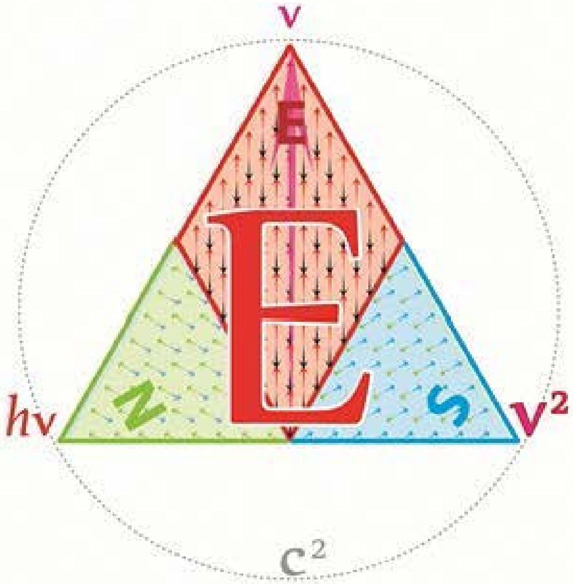

[Joule seconds] per second

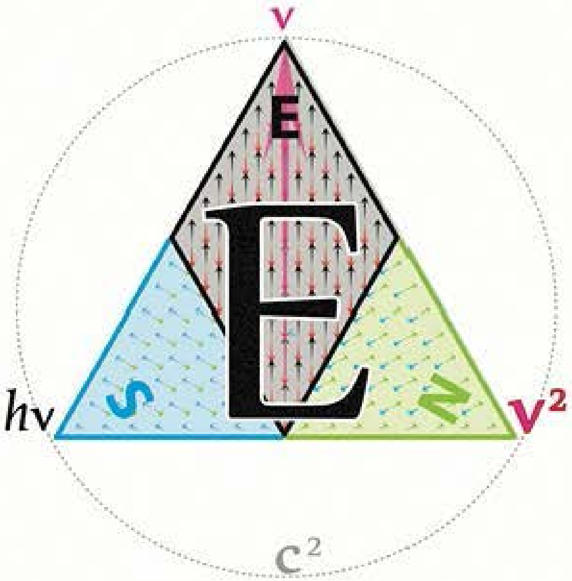

[Joule seconds] per second

Copyright ABRAHAM [2008] - all rights reserved

Linear Momentum

Linear momentum is the SQUARE ROOT of equilateral Planck mass-energy geometries and produces a undirectional vector force

The Energy-momentum relationship is a fundamental physical property used to determine the mass of a body

Using the formula for mass-Energy equivalence as it relates to Photons moving at 'c'

$$E = h\nu = \frac{hc}{\lambda} = mc^2$$

Noting that the rest mass in the case of EM fields (Bosons and Photons) is to equal Zero

we can derive a relationship for Momentum - Energy - Wavelength showing that

$$p = E/c$$

Thus momentum in Photons is directly related to the EM energy content of the photon and the mass-energy content of any massive 3D body

$$p = \hbar k = \frac{h\nu}{c} = \frac{h}{\lambda}.$$

and momentum in Particles is related to the total EM Energies of an object (its rest Matter + KE) and the wavelengths associated with those distinct energy levels

$$p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$$

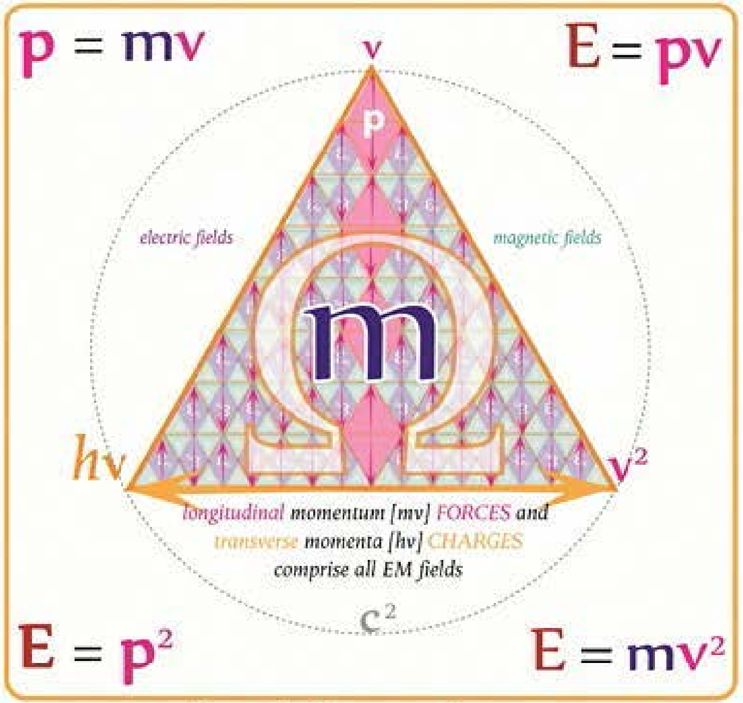

$$p = mv \qquad E = pv$$

$$E = p^2 \qquad E = mv^2$$

v
p
electric fields
magnetic fields
m
hν
v²
longitudinal momentum [mv] FORCES and
transverse momenta [hv] CHARGES
comprise all EM fields
c²

Longitudinal [linear] momentum

$$p = n\pi\left[\left[\overset{\text{Energy}}{\underset{\text{momentum}}{\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\Omega\,\underset{\text{velocity}}{\overset{\text{quanta}}{v}}}}\right]\right]$$

Momentum is a conserved quantity, meaning that the total momentum of any closed system (one not affected by external forces) cannot change.

Linear Momentum is the intrinsic sqare root vector component of Force

$$p = mv$$

In classical mechanics, momentum is the product of the mass and velocity of an object.

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} = m\frac{d\mathbf{v}}{dt} + \mathbf{v}\frac{dm}{dt}$$

In relativistic mechanics, this quantity is multiplied by the Lorentz factor.

$$p = \frac{n\pi\left[\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\Omega\,\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]}{v} = mv$$

EM field momentum is a function of its energy density, and is directly proportional to the group velocity

$$\left[\sqrt{\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\,\Omega\,\underset{\text{velocity}}{\overset{\text{quanta}}{v}}\right]^2}\,\right]$$

Vector linear momentum

is the square root of the KEM field energies produced by Matter in motion

$$n\pi\left[\left[m\overset{\text{Energy}}{\Omega}v\right]\right]$$

momenta

Classical formulation

Quantum formulation

$$p = mv$$

Linear Momentum
is the vector force of
mass times its velocity

$$p = \frac{E}{v}$$

Linear Momentum is
Energy divided by the
square root of its quanta

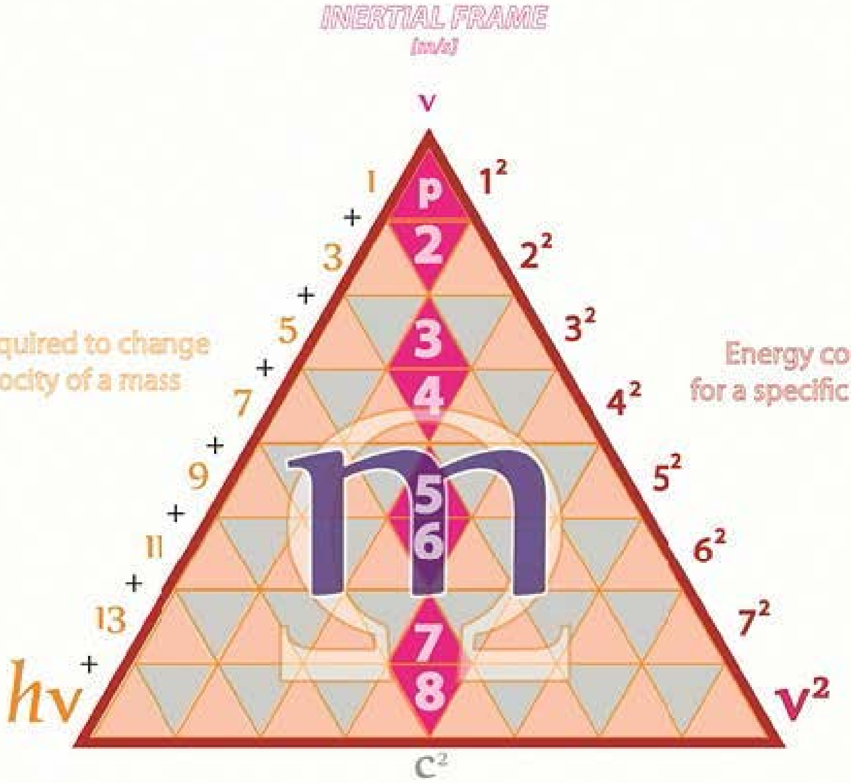

Energy required to change
the velocity of a mass

Energy content
for a specific velocity

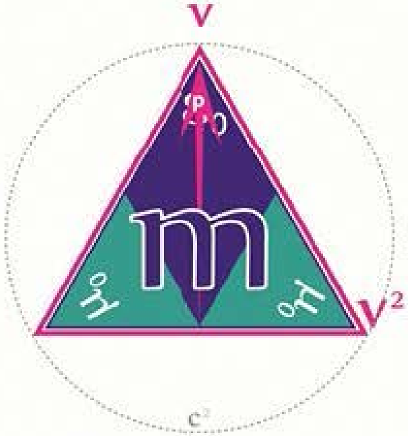

$$\frac{E}{c^2}\cdot v$$

$$\mathbf{p} = \sum_{i=1}^{n} m_i \mathbf{v}_i = m_1\mathbf{v}_1 + m_2\mathbf{v}_2 + m_3\mathbf{v}_3 + \cdots + m_n\mathbf{v}_n\,,$$

$$\frac{hv^2}{v}$$

The linear momentum of a EM field or system of particles is
the vector sum of the linear momenta in the KEM fields
of individual particles in any spatial co-ordinate system

Copyright ABRAHAM [2008] - all rights reserved

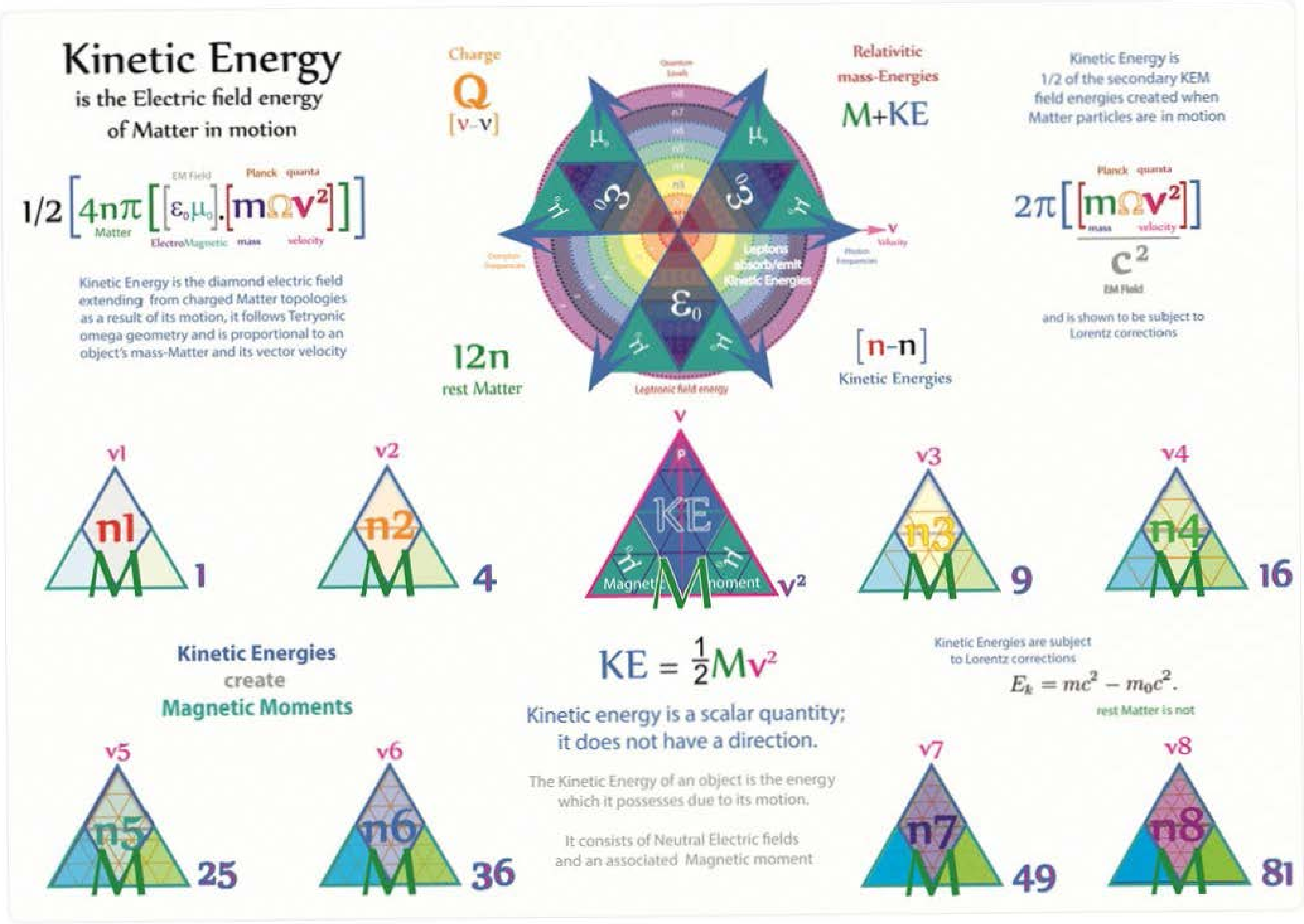

Kinetic Energy

is the Electric field energy
of Matter in motion

$$1/2\left[4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega \mathbf{v}^2\right]\right]\right]$$

EM Field
Planck
quanta
Matter
ElectroMagnetic
mass
velocity

Kinetic Energy is the diamond electric field extending from charged Matter topologies as a result of its motion, it follows Tetryonic omega geometry and is proportional to an object's mass-Matter and its vector velocity

Charge

Q

[v-v]

12n

rest Matter

Quantum Levels
n8
n7
n6
n5
n4
n3
n2
μ_0
μ_0
ω_0
ω_0
μ_0
μ_0
μ_0
μ_0
ε_0
Compton frequencies
Leptons absorb/emit Kinetic Energies
Photon frequencies
v
Velocity
Leptonic field energy

Relativitic mass-Energies

M+KE

[n-n]

Kinetic Energies

Kinetic Energy is 1/2 of the secondary KEM field energies created when Matter particles are in motion

$$\frac{2\pi\left[\left[m\Omega \mathbf{v}^2\right]\right]}{c^2}$$

Planck
quanta
mass
velocity
EM Field

and is shown to be subject to Lorentz corrections

v1
n1
M
1
v2
n2
M
4
v
P
KE
μ_0
μ_0
Magnetic moment
M
v^2
v3
n3
M
9
v4
n4
M
16

Kinetic Energies
create
Magnetic Moments

$$KE = \frac{1}{2}Mv^2$$

Kinetic energy is a scalar quantity;
it does not have a direction.

The Kinetic Energy of an object is the energy which it possesses due to its motion.

It consists of Neutral Electric fields and an associated Magnetic moment

Kinetic Energies are subject to Lorentz corrections

$$E_k = mc^2 - m_0c^2.$$

rest Matter is not

v5
n5
M
25
v6
n6
M
36
v7
n7
M
49
v8
n8
M
81

Tetryonics 14.06 - Kinetic Energy

Copyright ABRAHAM [2008] - all rights reserved

Kinetic Energy and Magnetic moments

$$Mv^2 = KEM = p^2$$

Unlike charged fields KEM fields have neutralised Electric Fields

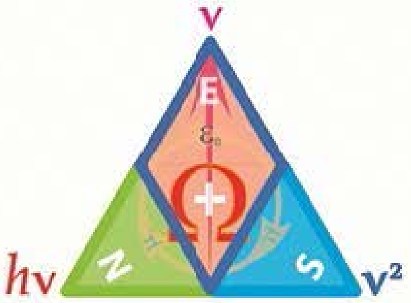
v
E
ε_0
Ω
+
hv
N
S
v²

KEM field geometry
of positive charged
Matter in motion

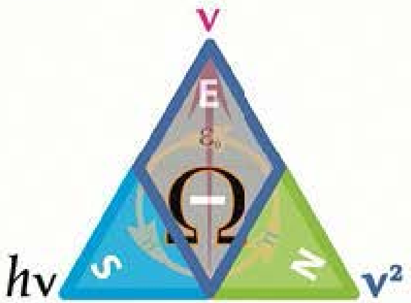
v
E
ε_0
Ω
−
hv
S
N
v²

KEM field geometry
of negative charged
Matter in motion

$$KE = \frac{1}{2}Mv^2$$

Often noted as being two distinct EM energies [Kinetic Energy and Magnetic moments] are shown to be orthogonal aspects of the same KEM field of motion

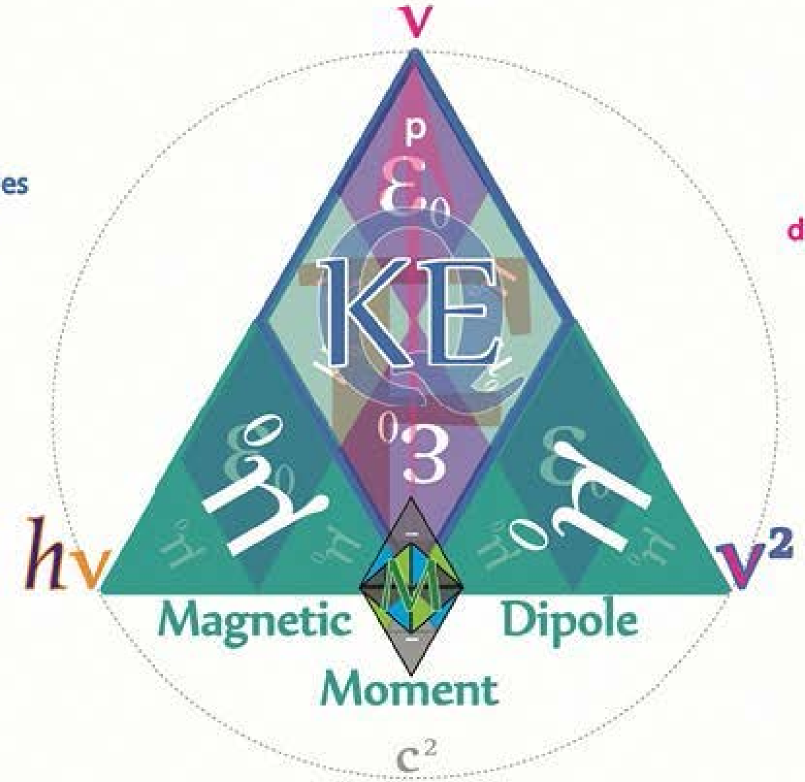
v
p
ε_0
KE
Ω
ε_0
μ
hv
v²
M
Magnetic
Dipole
Moment
c²

As the velocity of a particle increases so does its Kinetic Energy and Magnetic moment creating Lorentz variable KEM fields

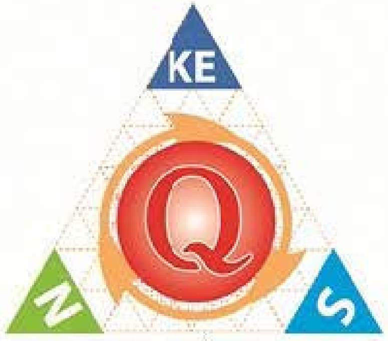

$$\mu_B = \frac{1}{2}Mv^2$$

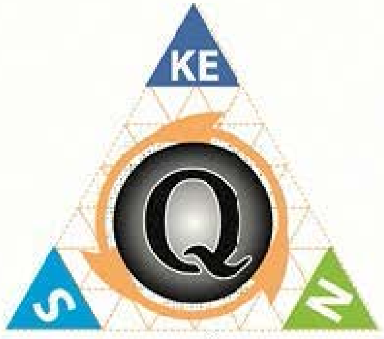

$$\mu_B = \frac{1}{2}Mv^2$$

Types of Momentum

There are 3 forms of momenta in physics

Ω

quantised angular momenta

[m²/s]

The quantised 'angular momentum' of each Planck mass-energy geometry, gives rise to the two quantum Charges

equilateral charge geometry

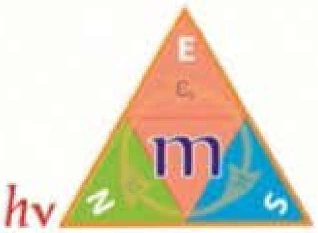

h

[kg·m²/s]

$m\Omega$

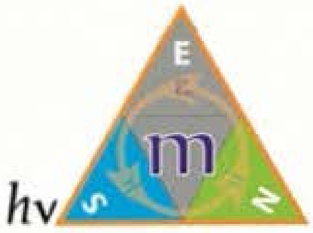

p

Linear momentum

[kg·m/s]

The square root (v) of each Planck quanta's mass-energy geometry (v^2) is vector Linear Momentum

vector momentum

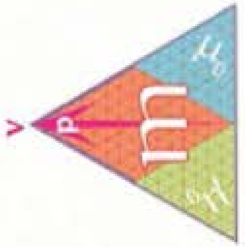

mv

[kg·m/s]

p

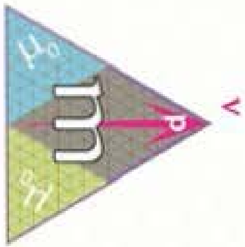

ℓ

Angular momentum

[kg·m²/s]

The orbital angular momentum of electrons in atoms associated with a given quantum state

. . . [n-1] [n] [n+1] . . .

vector rotation about a point

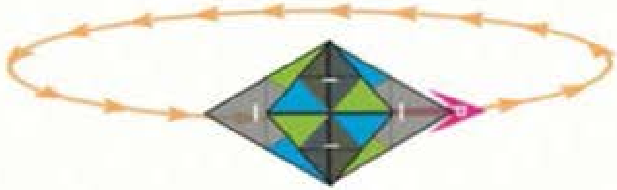

$$L^2 = L_x^2 + L_y^2 + L_z^2$$

Copyright ABRAHAM [2008] - all rights reserved

Kinetic Energy vs Momentum

An important difference is that Kinetic energy is a scalar quantity - it has no direction in space momentum is a vector quantity - it has a direction in space, momenta combine like forces do.

$$p = Mv$$

$$KE = 1/2\left[Mv^2\right]$$

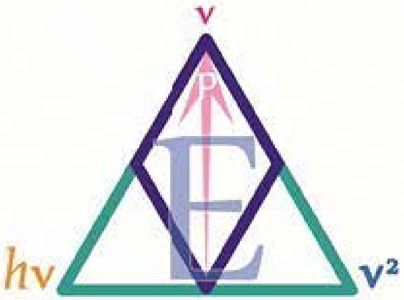

Energy momenta

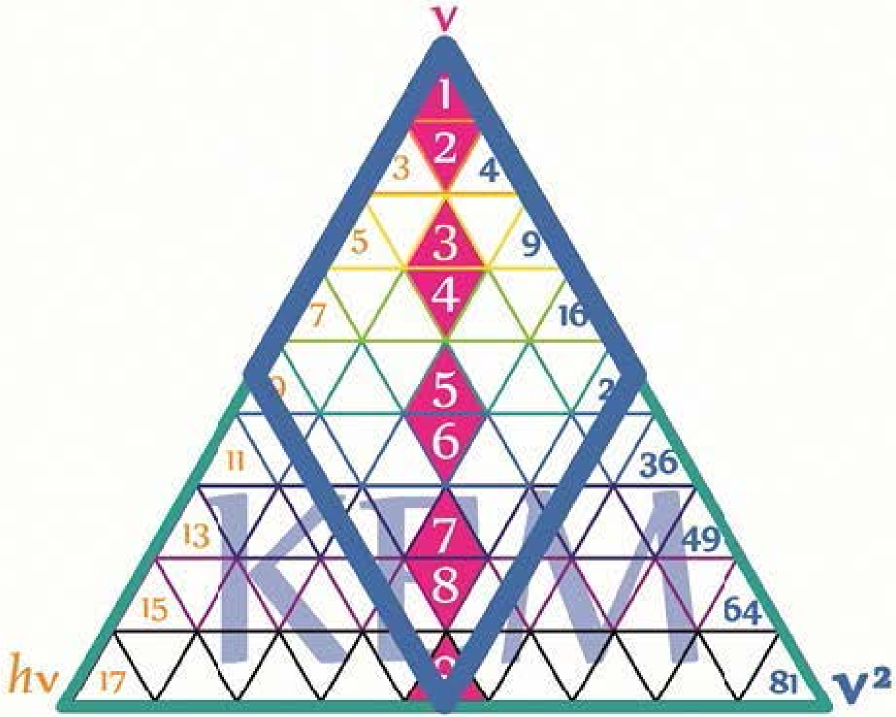

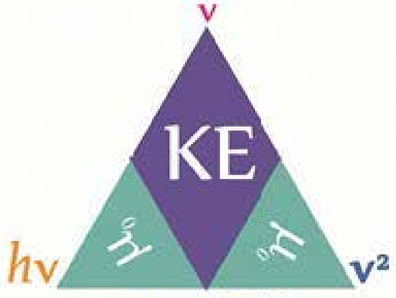

KEM field

$$p = \sqrt{E}$$

$$KEM = Mv^2$$

In Tetryonic geometry, the square root maps the linear momenta [mv] of a field to its ENERGY

$$p^2/_{2m} = KE$$

In Tetryonic geometry, E-field geometry maps the kinetic energy [1/2 mv] of a field to its ENERGY

Linear momentum has a different geometric ENERGY relationship to that of Electric fields

Copyright ABRAHAM [2008] - all rights reserved

mass-Energy-momentum

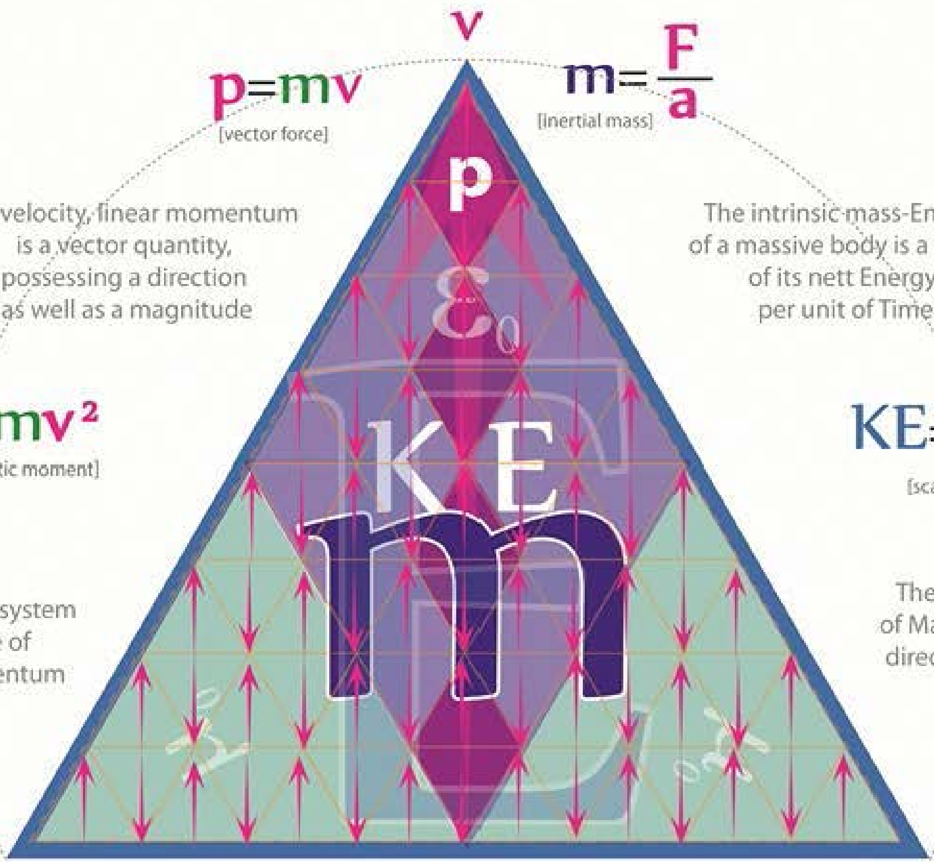

$p=mv$

[vector force]

$m=\frac{F}{a}$

[inertial mass]

Like velocity, linear momentum
is a vector quantity,
possessing a direction
as well as a magnitude

The intrinsic mass-Energy
of a massive body is a result
of its nett Energy
per unit of Time

$KEM=mv^2$

[Kinetic Energy & magnetic moment]

$KE=\frac{1}{2}mv^2$

[scalar E-field]

The Energy of a system
is the square of
its linear momentum

The Kinetic energy
of Matter in motion is
directly related to its
velocity

$E=mv^2$

[scalar KEM field energies
of Matter in motion]

$E=p^2$

[Energy momenta]

$\left[\underset{\text{mass}}{m}\overset{\text{Planck}\quad\text{quanta}}{\Omega}\underset{\text{velocity}}{v}\right]$

acceleration results in
changes to momenta

$F=ma$

Newton's
Second law of Motion

Linear momentum is the square root
of the Energy of a system.

It is the nett linear Force
resulting from Matter in motion
and was used by Newton
as the foundation for
his Laws of motion

Linear momentum is also a conserved quantity, meaning that if a closed system is not affected by external forces, its total linear momentum cannot change. Although originally expressed in Newton's second law, the conservation of linear momentum also holds in special relativity and, with appropriate definitions, a (generalized) linear momentum conservation law holds in electrodynamics, quantum mechanics, quantum field theory, and general relativity. In relativistic mechanics, non-relativistic linear momentum is further multiplied by the Lorentz factor

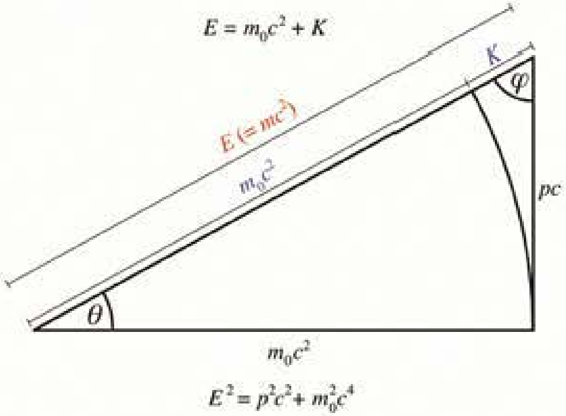

Often generalised as having the geometry of
right angled triangles the mass-energy momenta
relationship is fully revealed using
Equilateral geometry

E^2

is the energy content
of a superpositioned
EM field

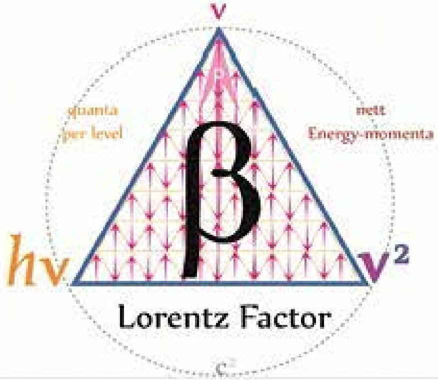

Tetryonics 14.10 - mass-Energy momentum

Lorentz velocity correction Factor

The Lorentz factor or Lorentz term is an expression which appears in several equations in special relativity. It arises from deriving the Lorentz transformations. The name originates from its earlier appearance in Lorentzian electrodynamics - named after the Dutch physicist Hendrik Lorentz.

Wavelength, momentum

mass-ENERGY

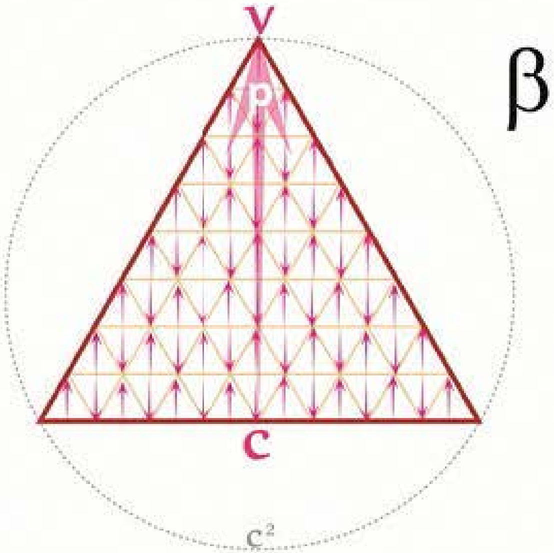

$$\beta = \left[\frac{v}{c}\right]$$

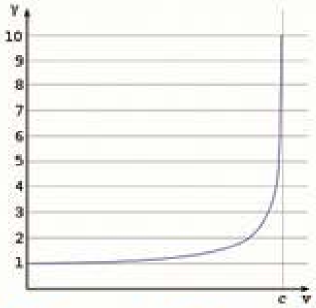

$$\left[\frac{v^2}{c^2}\right] = \beta^2$$

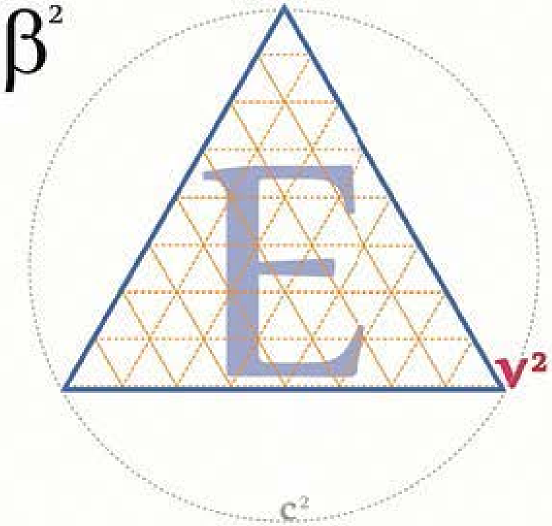

Classically modelled as an infinite series approaching c
Tetryonics reveals it to be a physical property
of equilateral energy-momenta geometries

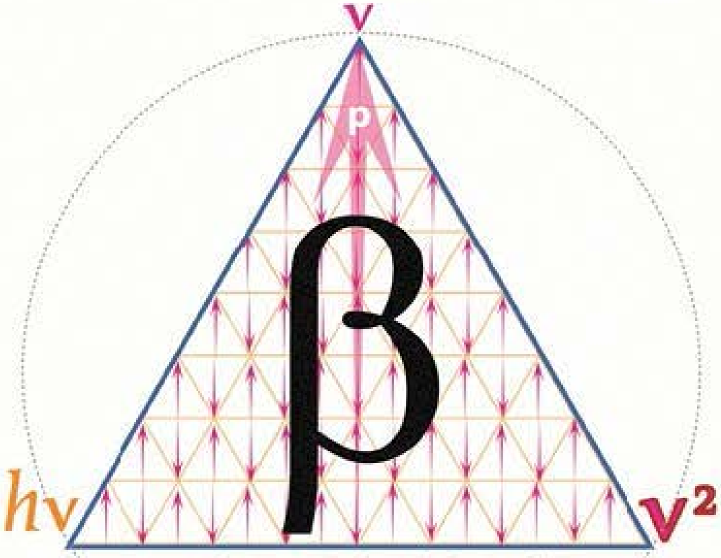

A vector measure of the Energy content of a waveform

$$\gamma \equiv \frac{c}{\sqrt{c^2 - v^2}} = \frac{1}{\sqrt{1 - \beta^2}} = \frac{dt}{d\tau}$$

A scalar measure of a KEM waveform's energies

$$\gamma = 1/\sqrt{1 - v^2/c^2}$$

Importantly - Newton's classical velocity addition remains correct, but the energy-momenta required to accelerate Matter to higher velocities increases exponentially in line with the equilateral [Tetryonic] geometry of mass-energies

γ

Γ

Tetryonics 14.11 - Lorentz velocity correction Factor

Unified Energy momenta geometry

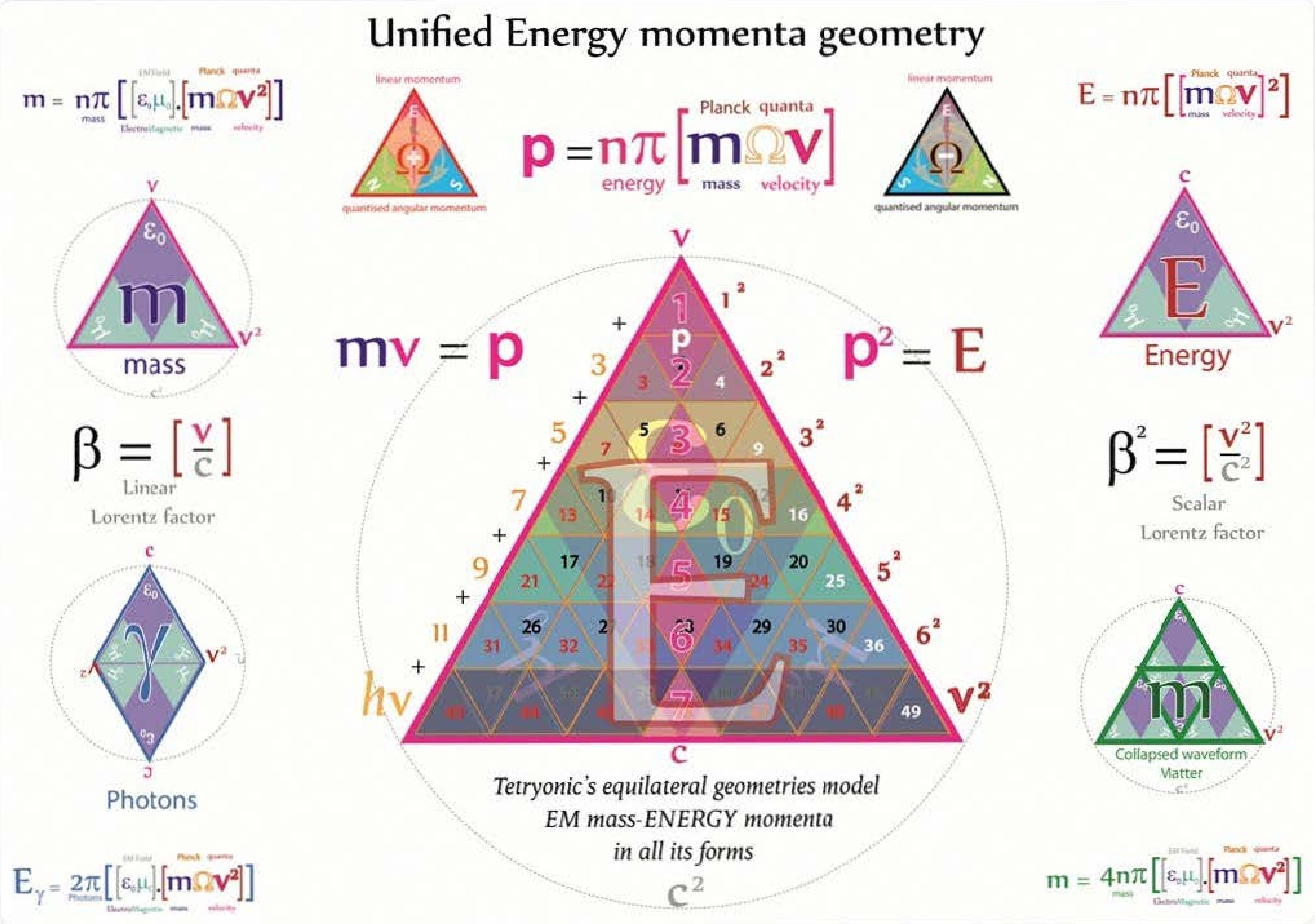

$m = n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

$p = n\pi[m\Omega v]$

$E = n\pi[[m\Omega v]^2]$

$mv = p$

$p^2 = E$

$\beta = [\frac{v}{c}]$

Linear Lorentz factor

$\beta^2 = [\frac{v^2}{c^2}]$

Scalar Lorentz factor

$E_\gamma = 2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

$m = 4n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Tetryonic's equilateral geometries model EM mass-ENERGY momenta in all its forms

EM Field
Planck quanta
mass
ElectroMagnetic
mass
velocity
energy
linear momentum
quantised angular momentum
E
N
S
v
v²
c
c²
ε0
m
mass
E
Energy
γ
Photons
Collapsed waveform
Matter
1²
2²
3²
4²
5²
6²
3
5
7
9
11
+
hv
p
1
2
3
4
5
6
7
16
17
19
20
21
25
26
29
30
31
32
36
49

Squared Energy distributions

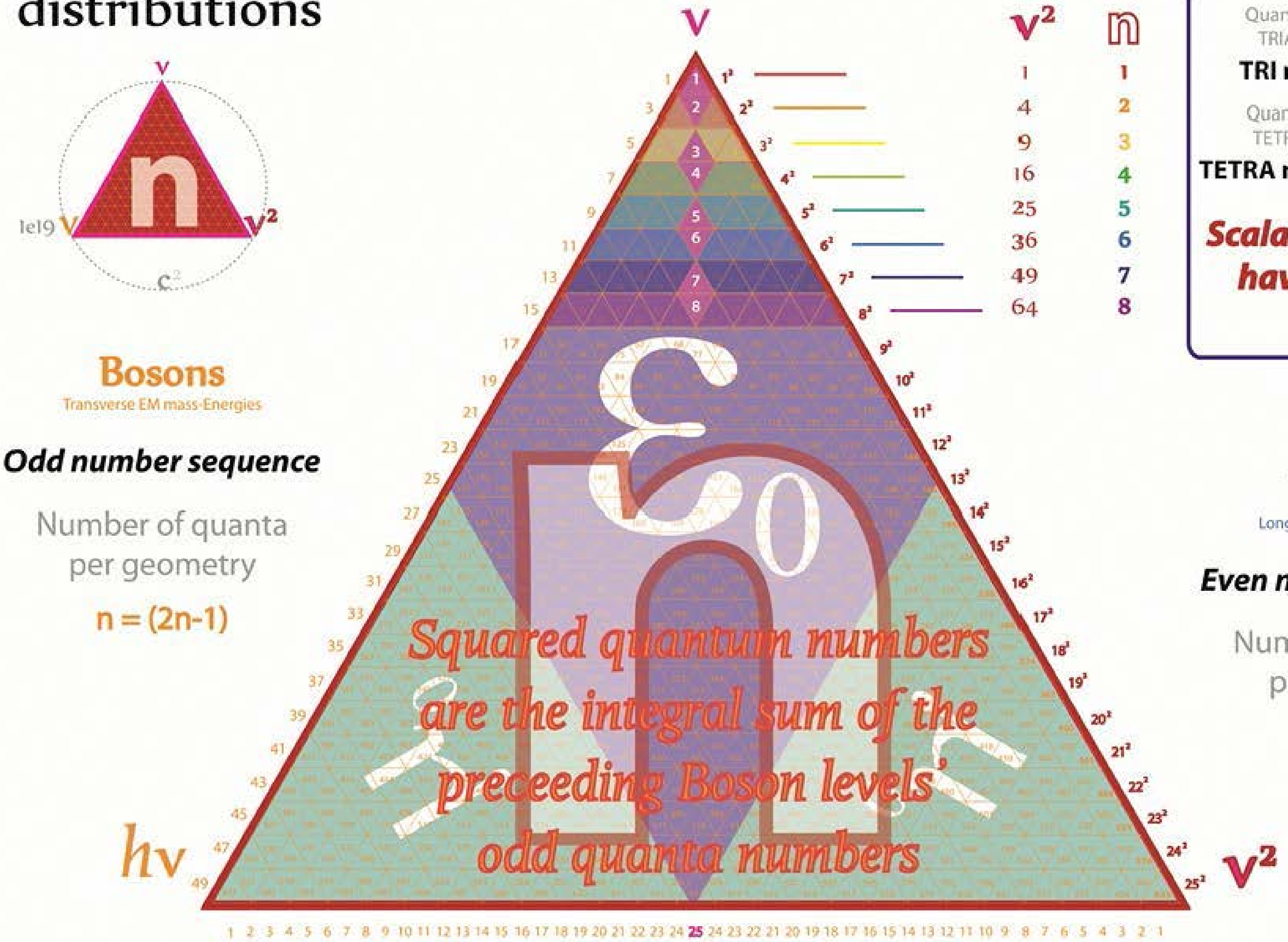
v
n
le19 v
v²
c²

Bosons

Transverse EM mass-Energies

Odd number sequence

Number of quanta
per geometry

n = (2n-1)

v
v²
n
1 2 3 4 5 6 7 8
1² 2² 3² 4² 5² 6² 7² 8² 9² 10² 11² 12² 13² 14² 15² 16² 17² 18² 19² 20² 21² 22² 23² 24² 25²
1 4 9 16 25 36 49 64
1 2 3 4 5 6 7 8
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
ε0
hv
v²
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Squared quantum numbers are the integral sum of the preceeding Boson levels' odd quanta numbers

[n-1] Transverse Quanta Distributions [n-1]

EM mass-energy

Quantum numbers are *not*
TRIANGULAR NUMBERS

TRI n = (n/2) x (n+1)

Quantum numbers are *not*
TETRAHEDRAL NUMBERS

TETRA n = (n/6) (n+1) (n+2)

Scalar energy levels have SQUARED quanta

Photons

Longitudinal EM mass-Energies

Even number sequence

Number of quanta
per geometry

n = (2n)

Tetryonics 15.02 - Squared Energy distributions

Copyright ABRAHAM [2008] - all rights reserved

Tetryonic Energy and Charge relationships

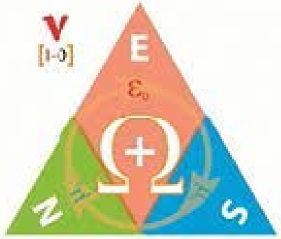
v
[1-0]
E
ε_0
+Ω
N
S

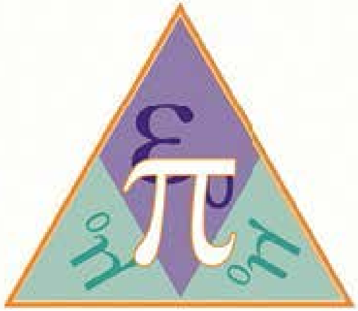

Quantised angular momenta is the geometric source of all mass-Energy-Matter relationships

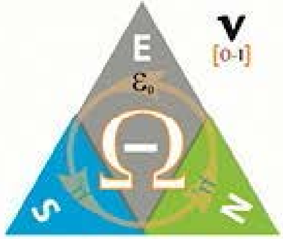
v
[0-1]
E
ε_0
−Ω
S
N

Scalar energies
(are 'squared' numbers)

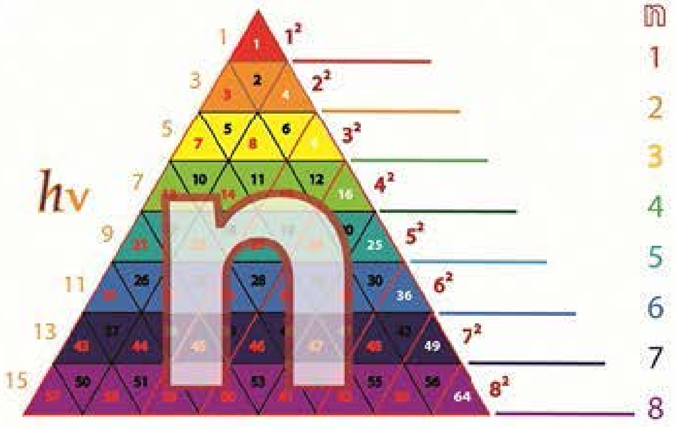
hv
n
1
3
5
7
9
11
13
15
1^2
2^2
3^2
4^2
5^2
6^2
7^2
8^2
1
2
3
4
5
6
7
8

The Energy density of a particle's charged fascia geometry determines its mass

m Charged mass-energy geometry & Matter topology determines a particle's physical characteristics (Type, Family, generation, mass etc) **M**

Charge

-12

electron

12
[0-12]

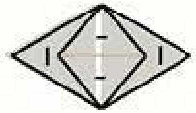

1.2 e20

mass-Matter

e-

8.851486 e-31 KG

Opposite nett elementary charge geometries

1875x
(differential)

differing Matter topologies

Proton

12
[24-12]

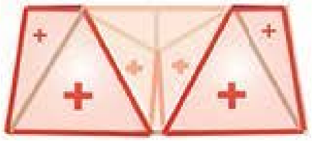

2.25 e23

1.65965 e-27 KG

P+

+12

[v-v]

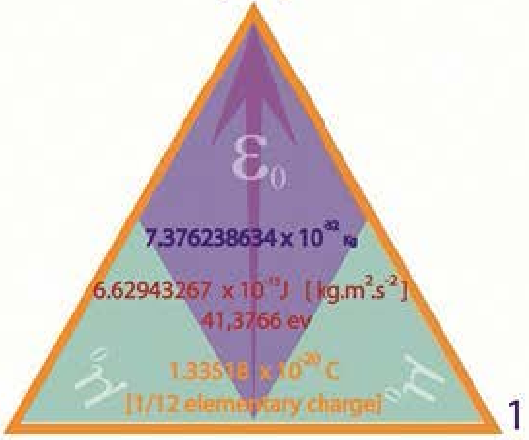
ε_0
$7.376238634 \times 10^{-51}$ kg
$6.62943267 \times 10^{-34}$ J [$kg.m^2.s^{-2}$]
41,3766 ev
1.33518×10^{-20} C
[1/12 elementary charge]
1

Planck mass-energy

Copyright ABRAHAM [2008] - all rights reserved

Planck's energy momenta quantum

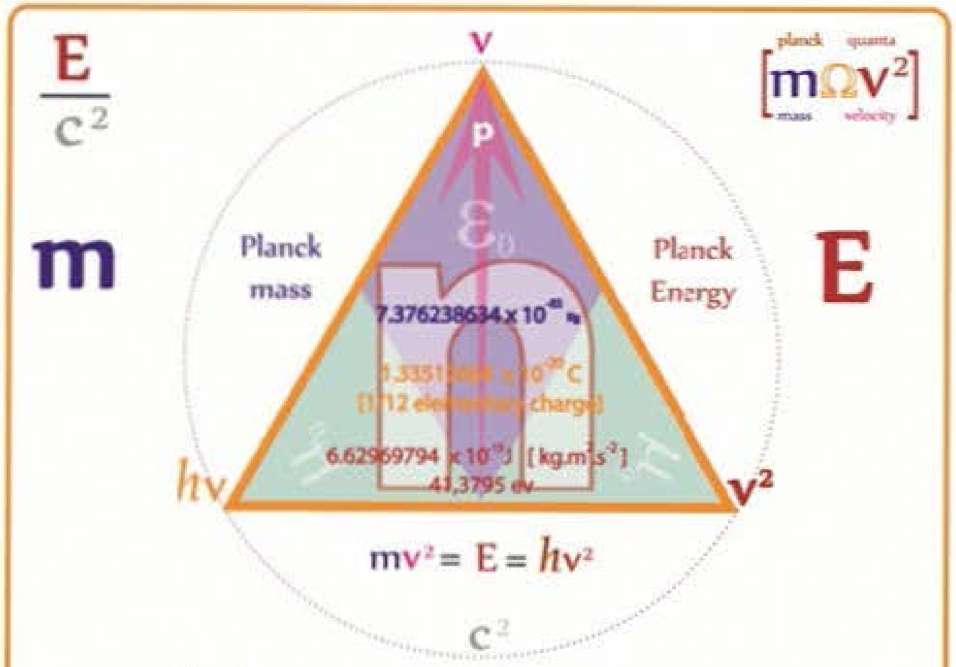
E
c²
m
E
v
p
ε₀
Planck mass
Planck Energy
planck quanta
[mΩv²]
mass velocity
7.376238634 x 10⁻⁴⁸ kg
1.3351... x 10⁻²⁰ C
[1/12 elementary charge]
6.62969794 x 10⁻³⁴ J [kg.m²s⁻²]
41,3795 ev
hv
v²
c²

$$mv^2 = E = hv^2$$

	Planck mass	m	m - mass quantum	
q	Planck Charge quantised angular momentum / spatial co-ordinate system	Ω	v - quanta of Charge elemental charge / 12	q
	Planck Energy	E	E - Energy quantum	

v^2 Compton Frequency	$\frac{mc^2}{h}$	1×10^{19} quanta per second
λ de Broglie wavelength	$\frac{h}{mc}$	$2.99792458 \times 10^{-11}$ meters
p Momentum	mc	2.21134×10^{-23} Ns

Planck's constant is the equilateral geometry of rest mass-energy momenta in standing wave Matter topologies

Currently the Atomic masses are calculated using 1/12 of a Molar Carbon12 atom's mass as the reference mass.

Tq [Carbon] - [252-252]

Tm [Carbon] - 270,072

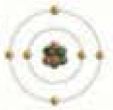

[270,072] / 12

22,506

If we calculated for the molar mass of C12 in 12 grams we would get

(.012/N/270072) 7.378205107 x 10^{-32} kgrams

this is in error by 1/2 of the mass of an electron

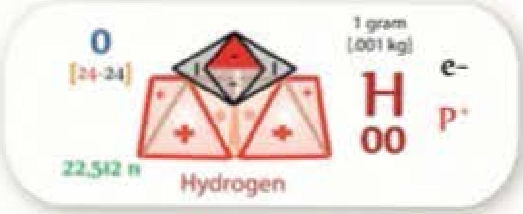
0
[24-24]
1 gram
(.001 kg)
H
00
e-
P⁺
22,512 n
Hydrogen

Tetryonic mass [Hydrogen] - 2.2512 e23 v

Using Tetryonic geometry we can solve for an exact Compton frequency of any mass-Matter topology and account exactly for all quantum energy contributions to the rest Matter of all electrons and Baryons in any element thus determining exact molar masses exclusive of blackbody radiation, kinetic energies and energies of measurement and avoiding 'weighted' masses

Defining [N] as the number of rest mass Hydrogen atoms in 1gram exactly and makes Avagadro's number the inverse of this number

$$\text{Molar mass} = \frac{H_1\,\text{Atomic mass}}{\text{Avagadro's No.}}$$

$$\text{mass } H_{mol}/Av = \frac{.001}{6.022141579 \times 10^{23}}$$

rest mass Hydrogen = $1.660538841 \times 10^{-27}$ g/mol

$$H_{mole}/m[H] = \frac{1.660738412 \times 10^{-27}}{22{,}512}$$

n Planck mass = $7.376238634 \times 10^{-32}$ Kg

This is an exact Atomic rest mass as opposed to the current weighted molar Atomic mass estimates (which is incorrect by **1/2** the mass of an electron)

Planck's constant is the 'quantum of action' and also contributes to the mass-energies of radiant EM wave geometries

Tetryonics 15.04 - Planck's energy momenta quantum

Copyright ABRAHAM [2008] - all rights reserved

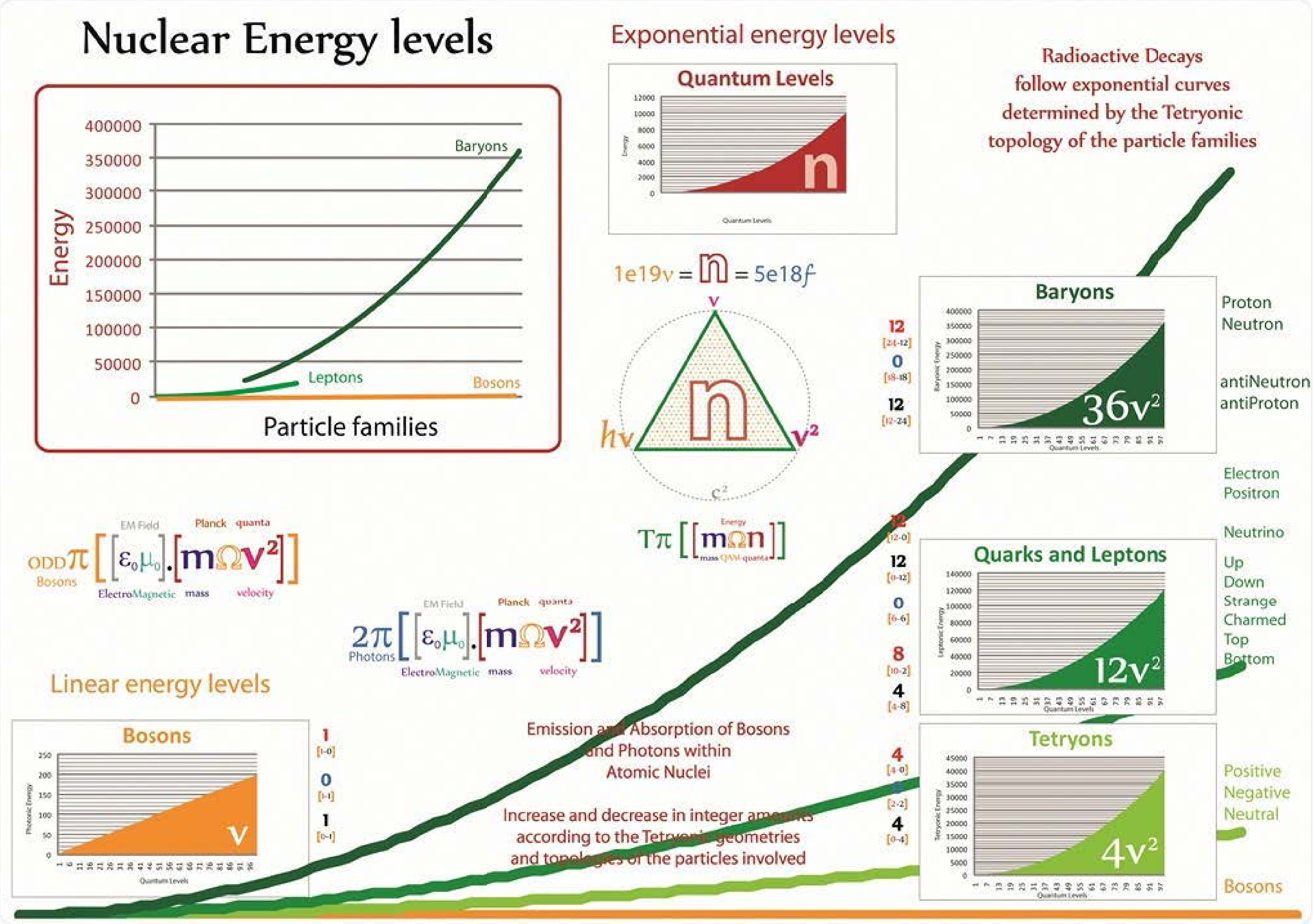

Nuclear Energy levels

Exponential energy levels

Quantum Levels

Radioactive Decays follow exponential curves determined by the Tetryonic topology of the particle families

$1e19v = n = 5e18f$

$T\pi[[m\Omega n]]$

Baryons

12 [24-12]
0 [18-18]
12 [12-24]

Proton
Neutron

antiNeutron
antiProton

Electron
Positron

Neutrino

12 [12-0]
12 [0-12]
0 [6-6]
8 [10-2]
4 [4-8]

Quarks and Leptons

Up
Down
Strange
Charmed
Top
Bottom

$ODD\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field
Planck
quanta
Bosons
ElectroMagnetic
mass
velocity

$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM Field
Planck
quanta
Photons
ElectroMagnetic
mass
velocity

Linear energy levels

Bosons

250
200
150
100
50
0
Photonic Energy
v
Quantum Levels

1 [1-0]
0 [1-1]
1 [0-1]

Emission and Absorption of Bosons and Photons within Atomic Nuclei

Increase and decrease in integer amounts according to the Tetryonic geometries and topologies of the particles involved

4 [4-0]
0 [2-2]
4 [0-4]

Tetryons

45000
40000
35000
30000
25000
20000
15000
10000
5000
0
Tetryonic Energy
4v²
Quantum Levels

Positive
Negative
Neutral

Bosons

Tetryonics 15.05 - Nuclear energy levels

Energy momenta geometry

m

All mass-energies have equilateral geometries

Ω

Quantised angular momenta facilitates EM interactions

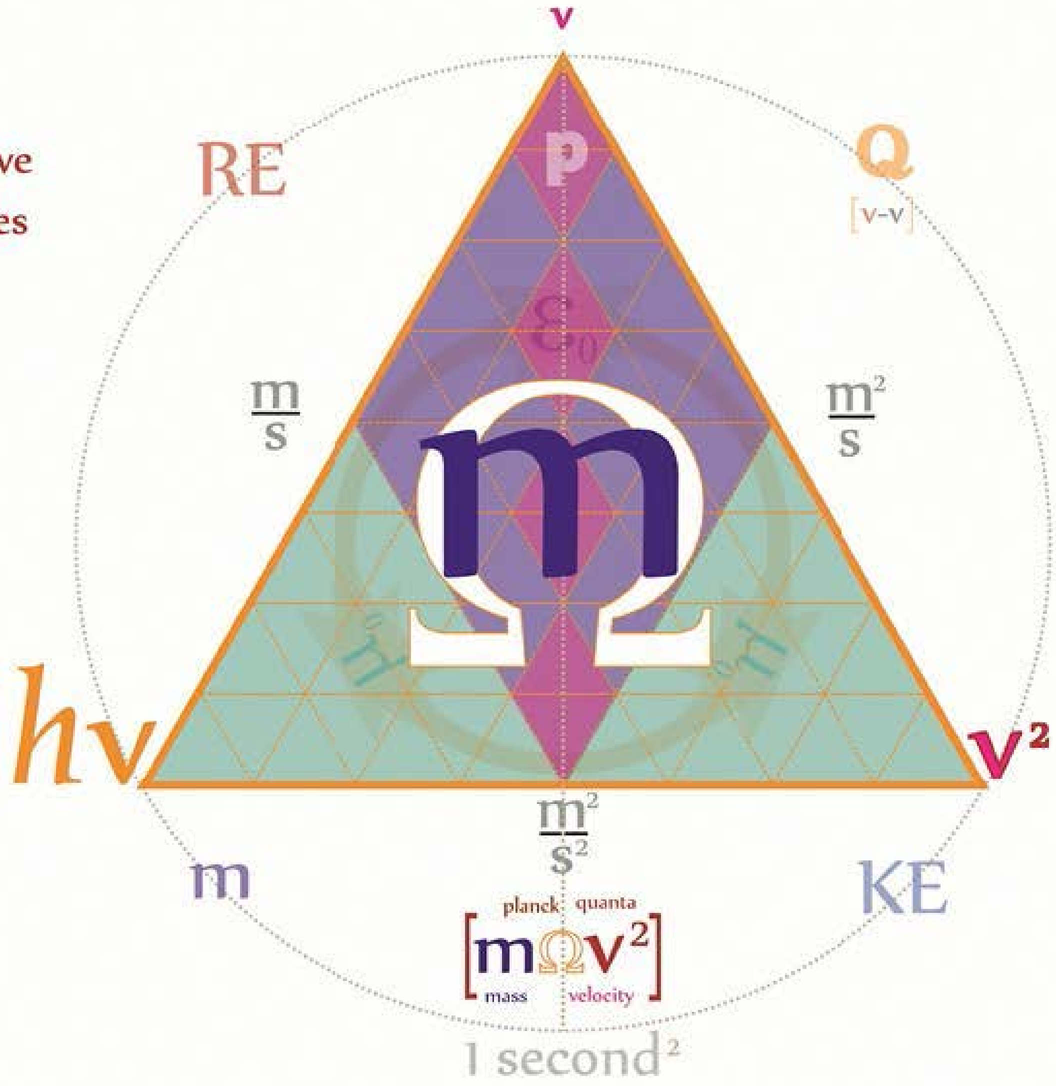
v
RE
p
Q
[v-v]
ε0
m/s
m²/s
m
Ω
hv
v²
m²/s²
m
KE
planck
quanta
[mΩv²]
mass
velocity
1 second²

E

mass-Energy

$kg\frac{m^2}{s^2}$

h

Planck's Constant

$kg\frac{m^2}{s}$

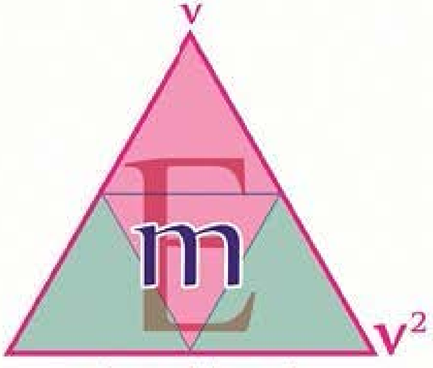

scalar mass velocity squared

$$mv^2 = E = hv^2$$

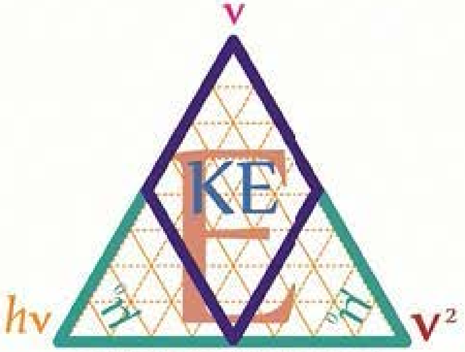

quantised angular momenta per second

m

mass

E

Energy

Bosons are transverse EM masses

Photons are longitudinal EM masses

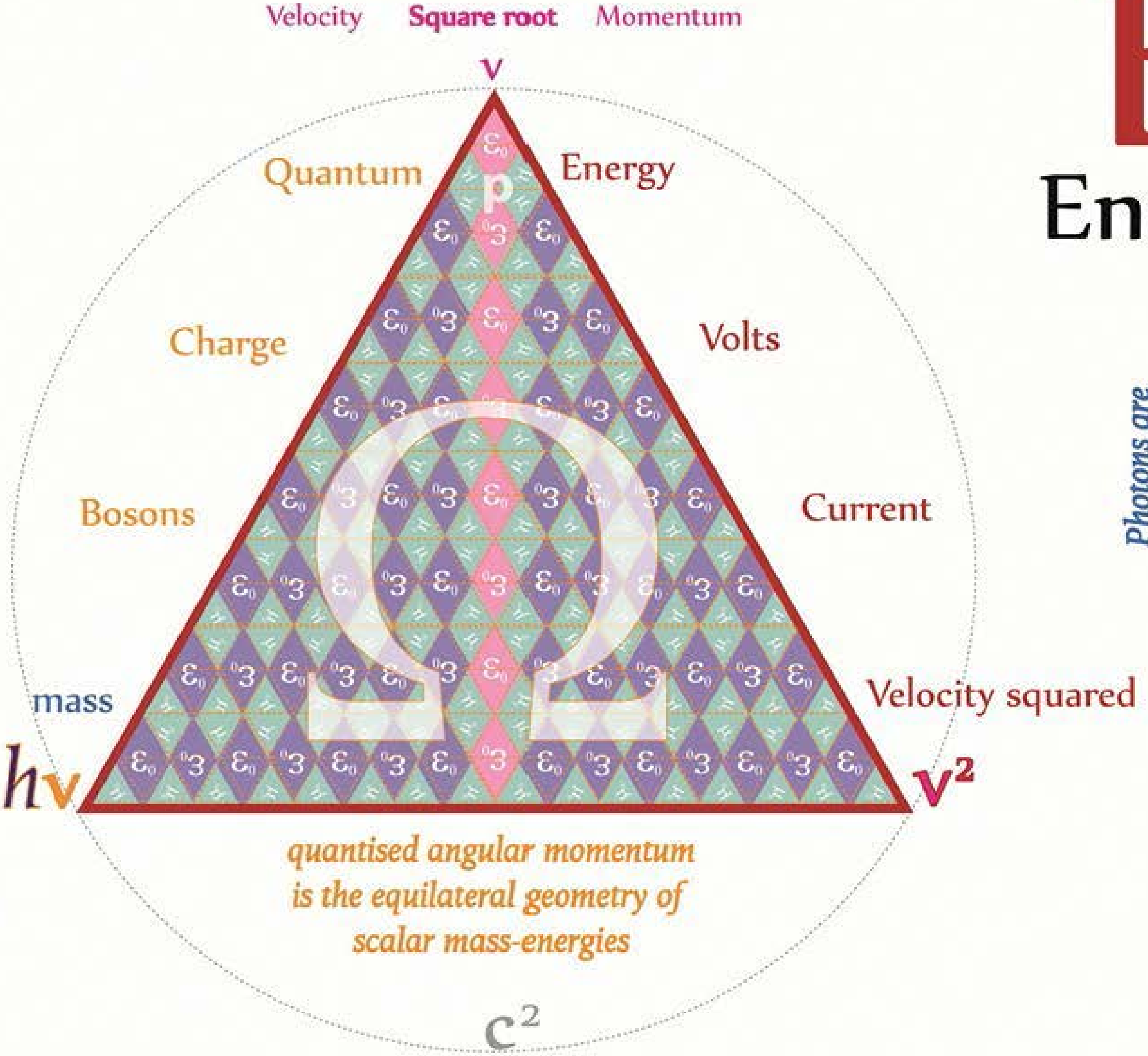
Velocity
Square root
Momentum
v
p
Quantum
Energy
Charge
Volts
Bosons
Current
mass
Velocity squared
Ω
ε0
hv
v²
quantised angular momentum is the equilateral geometry of scalar mass-energies
c²

mass is a measure of Energy per spatial co-ordinate system

mass-Energy momenta

Electromagnetic energy is a scalar measurement of mass

Tetryonics 15.07 - EM mass-energy relationships

Bosons are
Transverse masses

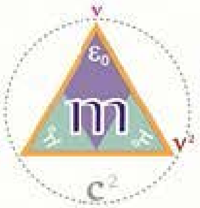
v
ε0
m
μ0
v²
c²

Planck
mass

EM Field Planck quanta

$$\text{ODD}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Bosons ElectroMagnetic mass velocity

m

7.376238634e-51 kg

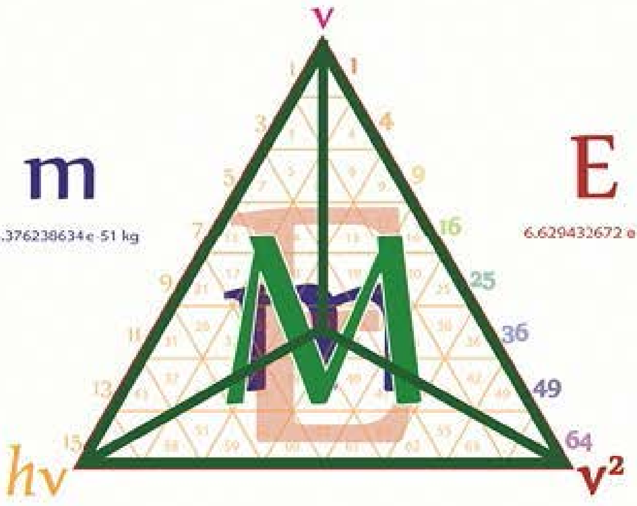
v
1
3
4
5
9
7
16
9
25
11
36
13
49
15
64
hv
v²
M

E

6.629432672 e-34 j

Photons are
Longitudinal masses

Photons

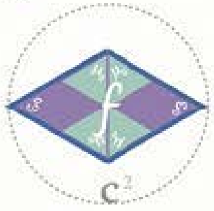

EM Field Planck quanta

$$\text{EVEN}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Photons ElectroMagnetic mass velocity

mass-ENERGY-Matter equivalence

LEPTONS

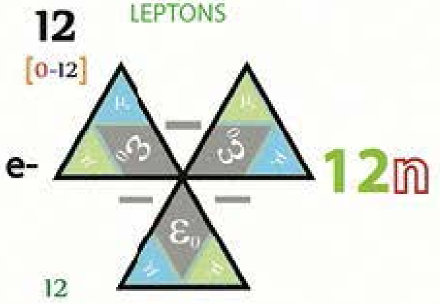
12
[0-12]
e-
12n
12

Tetryonic Matter of Electron - 8.851486361 x 10^{-31} Kg
mass-Energy of Electron - 7.955319207x 10^{-14} J

EM Field Planck quanta

$$12\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Fermions ElectroMagnetic mass velocity

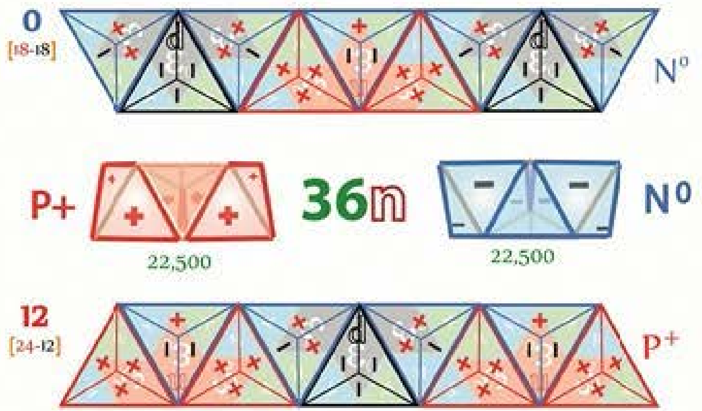
0
[18-18]
N°
P+
36n
N0
22,500
22,500
12
[24-12]
P+

Tetryonic Matter of Proton - 1.659653693 x 10^{-27} Kg
mass-Energy of Proton - 1.491622351 x 10^{-10} J

EM Field Planck quanta

$$36\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Baryons ElectroMagnetic mass velocity

ELEMENTS

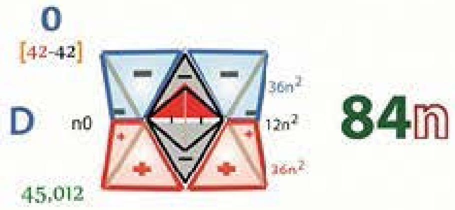
0
[42-42]
D
n0
36n²
12n²
36n²
84n
45,012

Tetryonic Matter of Deuterium - 3.320192534 x 10^{-27} Kg
mass-Energy of Deuterium - 2.984040234 x 10^{-10} J

EM Field Planck quanta

$$84\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Deuterons ElectroMagnetic mass velocity

Tetryonics 15.08 - mass-Energy-Matter equivalence

EM mass quanta in Matter

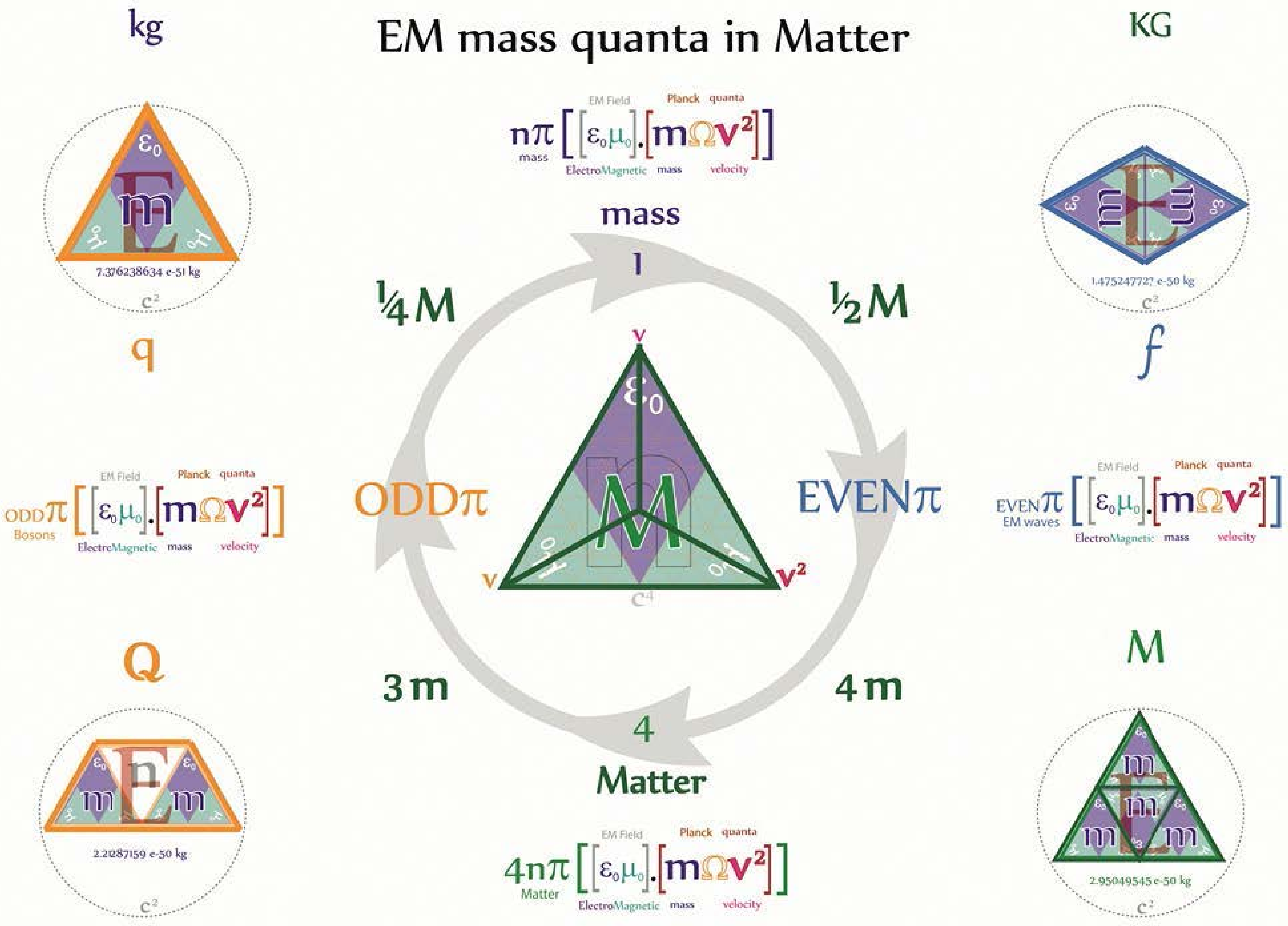
kg
ε0
m
E
μ0
7.376238634 e-51 kg
c²
q
KG
ε0
m
E
μ0
1.475247727 e-50 kg
c²
f
mass
1
¼ M
½ M
ODDπ
EVENπ
v
ε0
M
μ0
v
v²
c⁴
3 m
4 m
4
Matter
Q
ε0
n
m
E
μ0
2.2287159 e-50 kg
c²
M
ε0
m
E
μ0
2.95049545 e-50 kg
c²

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

EM Field Planck quanta

$$ODD\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons ElectroMagnetic mass velocity

EM Field Planck quanta

$$EVEN\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves ElectroMagnetic mass velocity

EM Field Planck quanta

$$4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

Tetryonics 15.09 - EM mass quanta in Matter

Copyright ABRAHAM [2008] - all rights reserved

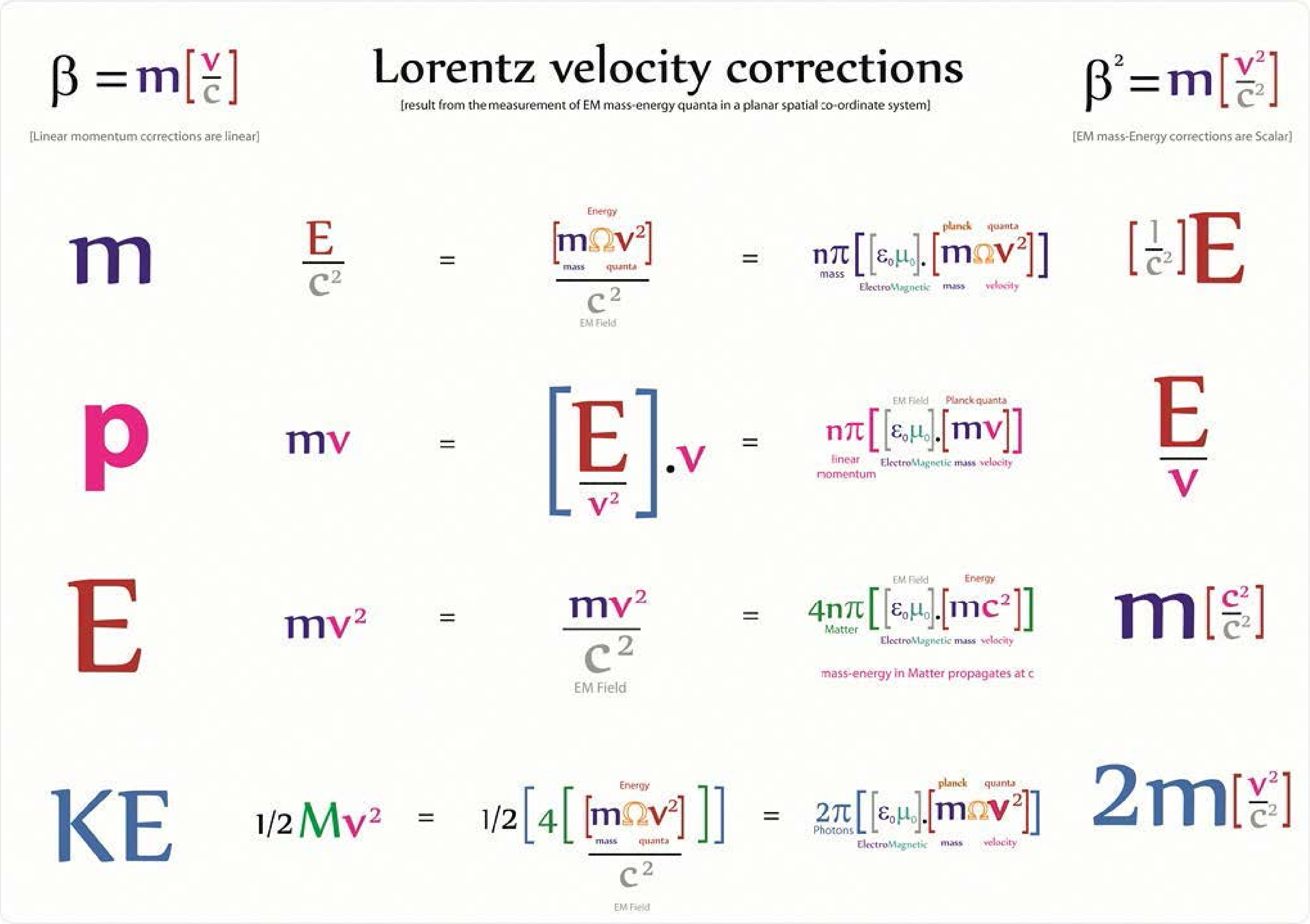

$$\beta = m\left[\frac{v}{c}\right]$$

[Linear momentum corrections are linear]

Lorentz velocity corrections

[result from the measurement of EM mass-energy quanta in a planar spatial co-ordinate system]

$$\beta^2 = m\left[\frac{v^2}{c^2}\right]$$

[EM mass-Energy corrections are Scalar]

$$m \qquad \frac{E}{c^2} = \frac{\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\underset{\text{quanta}}{\Omega}\,v^2\right]}}{\underset{\text{EM Field}}{c^2}} = \underset{\text{mass}}{n\pi}\left[\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}.\left[\underset{\text{mass}}{m}\,\overset{\text{planck}}{\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \qquad \left[\frac{1}{c^2}\right]E$$

$$p \qquad mv = \left[\frac{E}{v^2}\right].v = \underset{\text{linear momentum}}{n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}}.\overset{\text{Planck quanta}}{\left[\underset{\text{mass}}{m}\,\underset{\text{velocity}}{v}\right]}\right] \qquad \frac{E}{v}$$

$$E \qquad mv^2 = \frac{mv^2}{\underset{\text{EM Field}}{c^2}} = \underset{\text{Matter}}{4n\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}}.\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\underset{\text{velocity}}{c^2}\right]}\right] \qquad m\left[\frac{c^2}{c^2}\right]$$

mass-energy in Matter propagates at c

$$KE \qquad 1/2\,Mv^2 = 1/2\left[4\left[\frac{\overset{\text{Energy}}{\left[\underset{\text{mass}}{m}\,\underset{\text{quanta}}{\Omega}\,v^2\right]}}{\underset{\text{EM Field}}{c^2}}\right]\right] = \underset{\text{Photons}}{2\pi}\left[\underset{\text{ElectroMagnetic}}{\left[\varepsilon_0\mu_0\right]}.\left[\underset{\text{mass}}{m}\,\overset{\text{planck}}{\Omega}\,\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right] \qquad 2m\left[\frac{v^2}{c^2}\right]$$

Tetryonics 15.10 - Lorentz velocity corrections

Copyright ABRAHAM [2008] - all rights reserved

The geometry of Constants

Quantised angular momentum is the source of all physical Constants

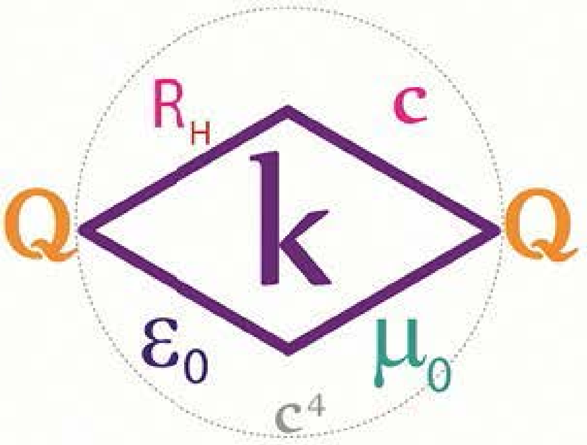
R_H
c
Q
k
Q
ε0
μ0
c4

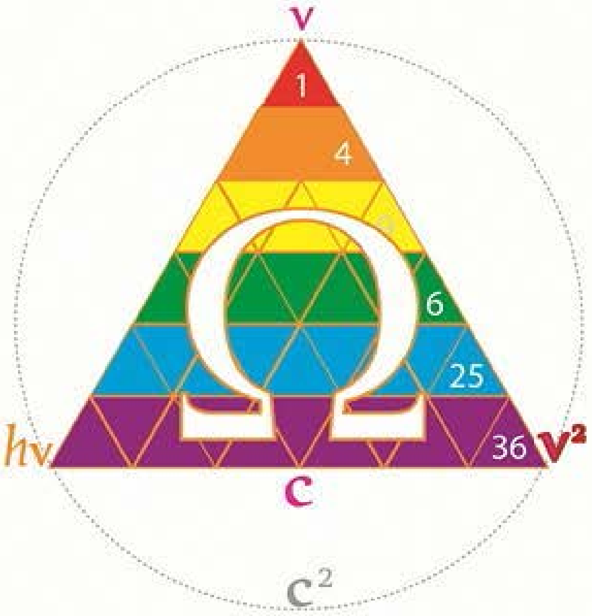
v
1
4
9
Ω
6
25
36
hv
v²
c
c2

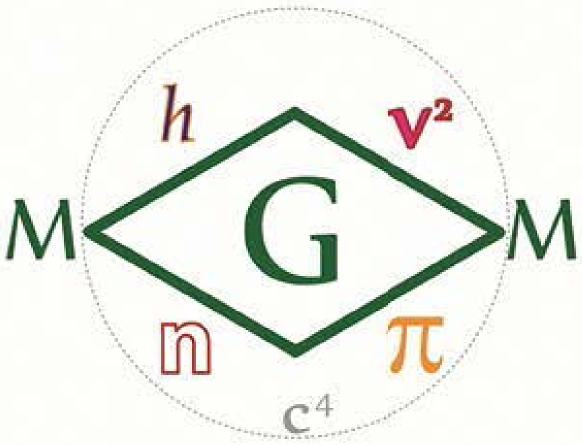
h
v²
M
G
M
n
π
c4

A physical constant is a physical quantity that is generally believed to be both universal in nature and constant in time.

It can be contrasted with a mathematical constant, which is a fixed numerical value but does not directly involve any physical measurement

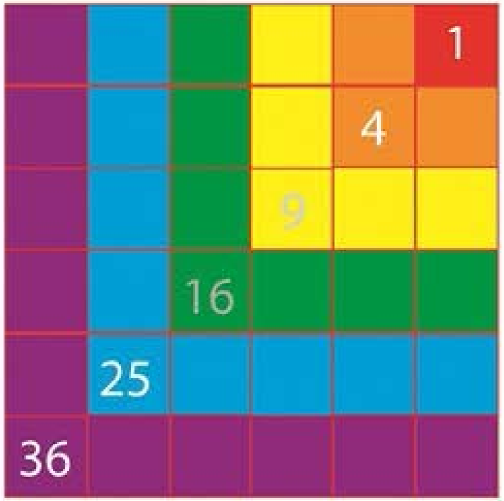

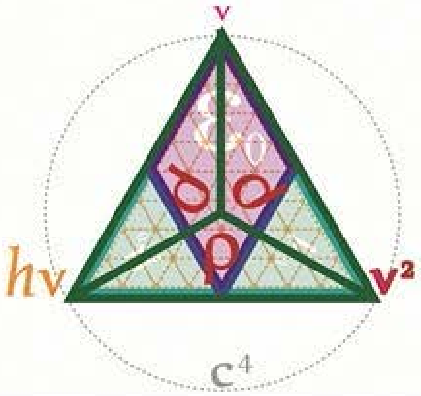

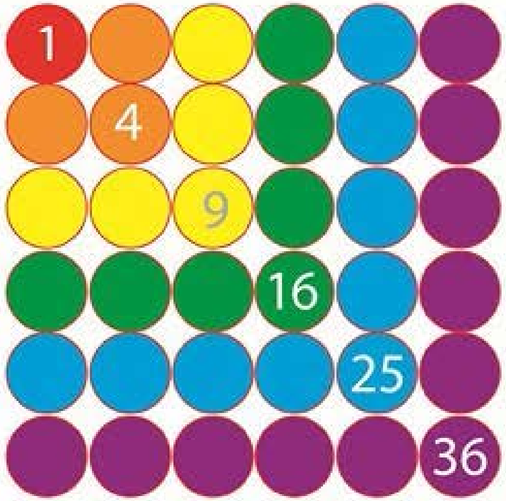

Tetryonics 16.01 - The geometry of Constants

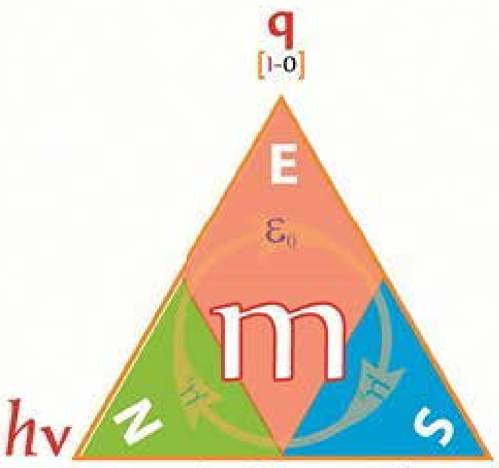

7.376238634 e-51 kg

"To interpret the vibrational
energy of N oscillators
not as a continuous,
infinitely divisible quantity,
but as a discrete quantity
composed of an integral number of
finite equal parts.

Max Planck

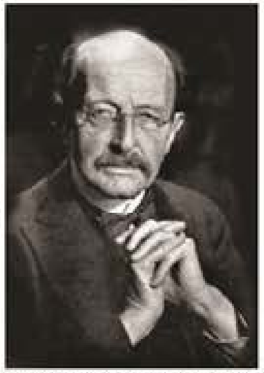

(April 23, 1858 – October 4, 1947)

$$\varepsilon = nhv$$

"Let us call each such part
of the energy element h "

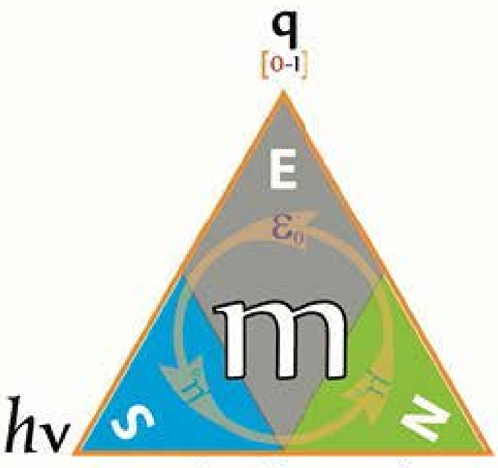

7.376238634 e-51 kg

Planck's Constant

[quantised mass-energy angular momenta]

N_A

1 mole of Hydrogen atoms
has a rest-mass of 1 gram

Solving for Planck's Constant using
the inverse of Avogadros number & Tetryonic geometry
we obtain an exact corrected value of :

$$\pm\ 6.629432672 \times 10^{-34}\ \text{J.s}$$

$4.137664546 \times 10^{-15}$ eV.s

$$mv^2 = E = hv^2$$

A rest mass Hydrogen atom
has a Compton frequency of
2.2512 e23 Planck quanta

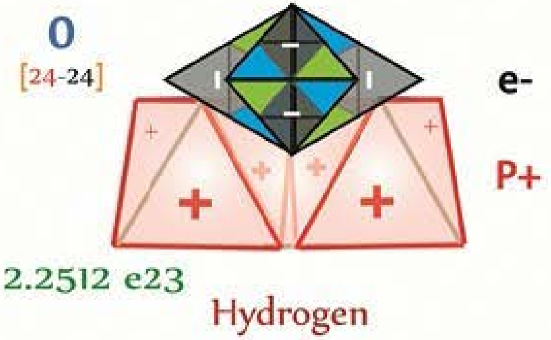

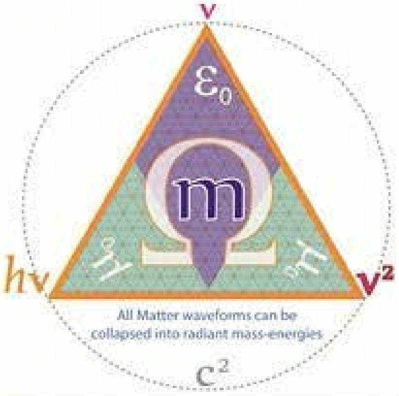

All Matter waveforms can be
collapsed into radiant mass-energies

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

*Planck's constant is the relationship between
EM mass-energy and quantised angular momenta
that provides the basis for EM charge in
Tetryonic geometries*

Copyright ABRAHAM [2008] - all rights reserved

Amedeo Avogadro

(9 August 1776 – 9 July 1856)

Avogadro's Number

The number of rest mass Hydrogen atoms in 1 gram
(and the rest molar mass of any element or compound)
can be determined directly from tetryonic theory

[exclusive of any measurement, blackbody or kinetic energies]

Using a Compton frequency of 2.2512 e23 Planck quanta
for a rest mass-Matter Hydrogen atom

atomic mass unit

$$1\ \mathrm{u} = \frac{M_\mathrm{u}}{N_\mathrm{A}} = 1.660\,538\,782(83) \times 10^{-24}\ \mathrm{g}$$

Avogadro N = 6.022141579 x 10^{23}

$$N^{-1} = \frac{2.2512\text{ e23 v [Hydrogen mass]}}{1.660538841 \times 10^{-24}\ \mathrm{g}}$$

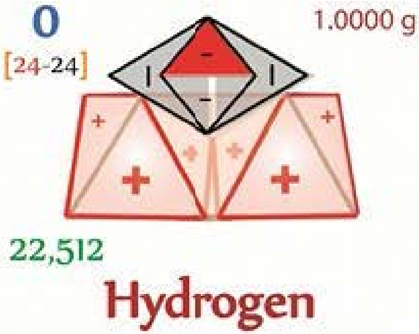

n0 *1 mol = 1 g*

Hydrogen = 1.660538841 x 10^{-27} Kg

n1 *1 mol = 1.000533 g*

n0 *1 mol = 11.996801 g*

Carbon 12 = 1.99211552 x 10^{-26} Kg

n1 *1 mol = 12 g*

Hydrogen's rest Tetryonic mass is 22,512 n Planck quanta [Proton - 22,500n + electron 12n]

EM Field Planck quanta

$$\left[22{,}512\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]\right]^{-1}$$

ElectroMagnetic mass velocity

The inverse mass of Hydrogen is equal to Avogadro's number

Coulomb's Constant

$$k = \frac{1}{4\pi\varepsilon_0}$$

The proportionality constant ke, called the Coulomb constant (sometimes called the Coulomb force constant), is related to defined properties of EM Energy-momenta and is used to define Electric field forces

$$c = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

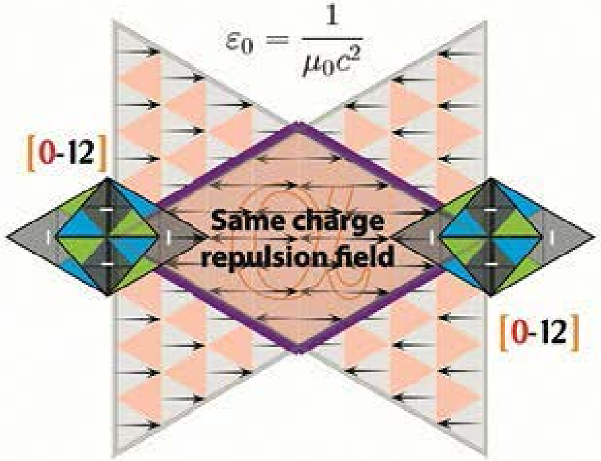
$\varepsilon_0 = \frac{1}{\mu_0 c^2}$
[0-12]
Same charge repulsion field
[0-12]

Similars repel

Linear Coulombic force interactions
are a result of charged E field
linear momenta

$$\mathbf{E} = \frac{\mathbf{F}}{q_t}$$

The Electric field can be defined as the longitudinal Force exerted by charged masses

$$8.987\text{ e9 } \frac{\text{Nm}^2}{\text{C}^2}$$

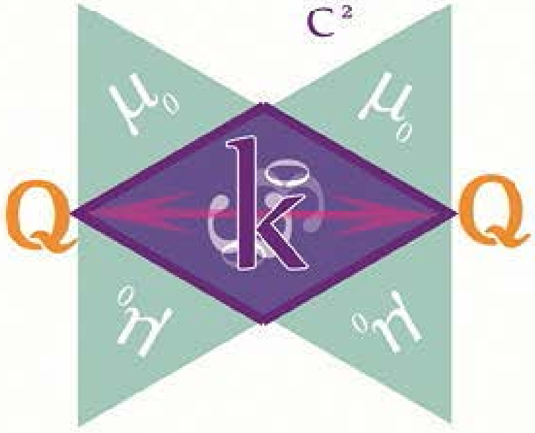
μ_0
μ_0
Q
k
Q
ε_0
ε_0

Opposites attract

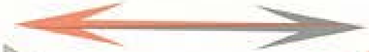

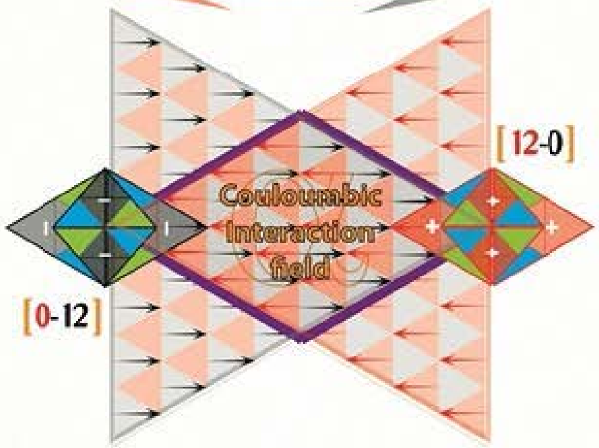
[12-0]
Couloumbic Interaction field
[0-12]

It is a measure of the interactive force
produced by the Electric field energy-momenta
of two superpositioned charge KEM fields

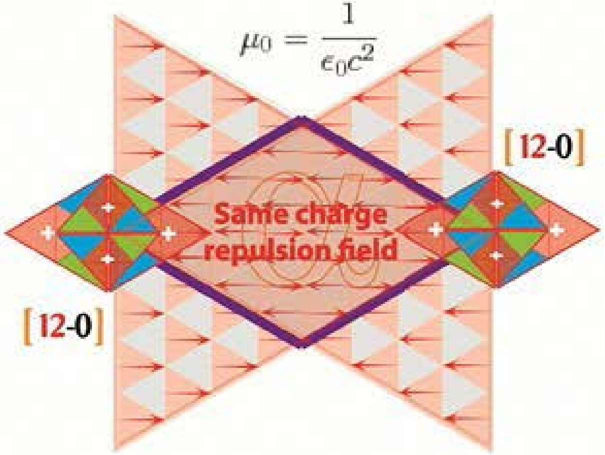
$\mu_0 = \frac{1}{\epsilon_0 c^2}$
[12-0]
Same charge repulsion field
[12-0]

Similars repel

Longitudinal E field forces
between Charged particles
are mediated by Photons

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

The Electric field can also be derived from Coulomb's Law

The Speed of ElectroMagnetic energies

The classical behaviour of the electromagnetic field is described by Maxwell's equations, which predict that the speed c with which electromagnetic waves (such as light) propagate through the vacuum is related to the electric constant ε0 and the magnetic constant μ0 by the equation $c = 1/\sqrt{\varepsilon_0 \mu_0}$.

c

Celeritas is a Latin word for "swiftness" or "speed". It is often given as the origin of the symbol c, the universal notation for the speed of light in a vacuum

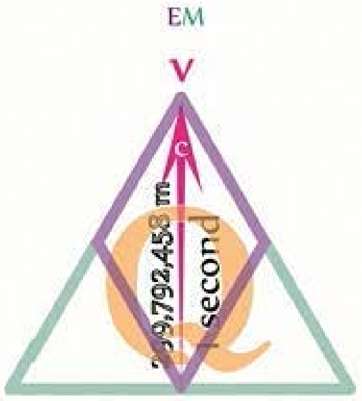

The distance energy travels in 1 second from its source

EM waves and energy momenta can be measured as either Transverse or Longitudinal waveforms with respect to their velocity vector

All EM waves and energy are symmetrical waveforms whose quanta contain ElectroMagnetic mass-energy and momenta

The equilateral geometry of EM energy is the inverse of the spatial geometry of 1 second

The square root of equilateral energies [scalar speed] is a vector velocity

velocity is a vector direction of the energies of motion

speed is a scalar property of the energies of motion

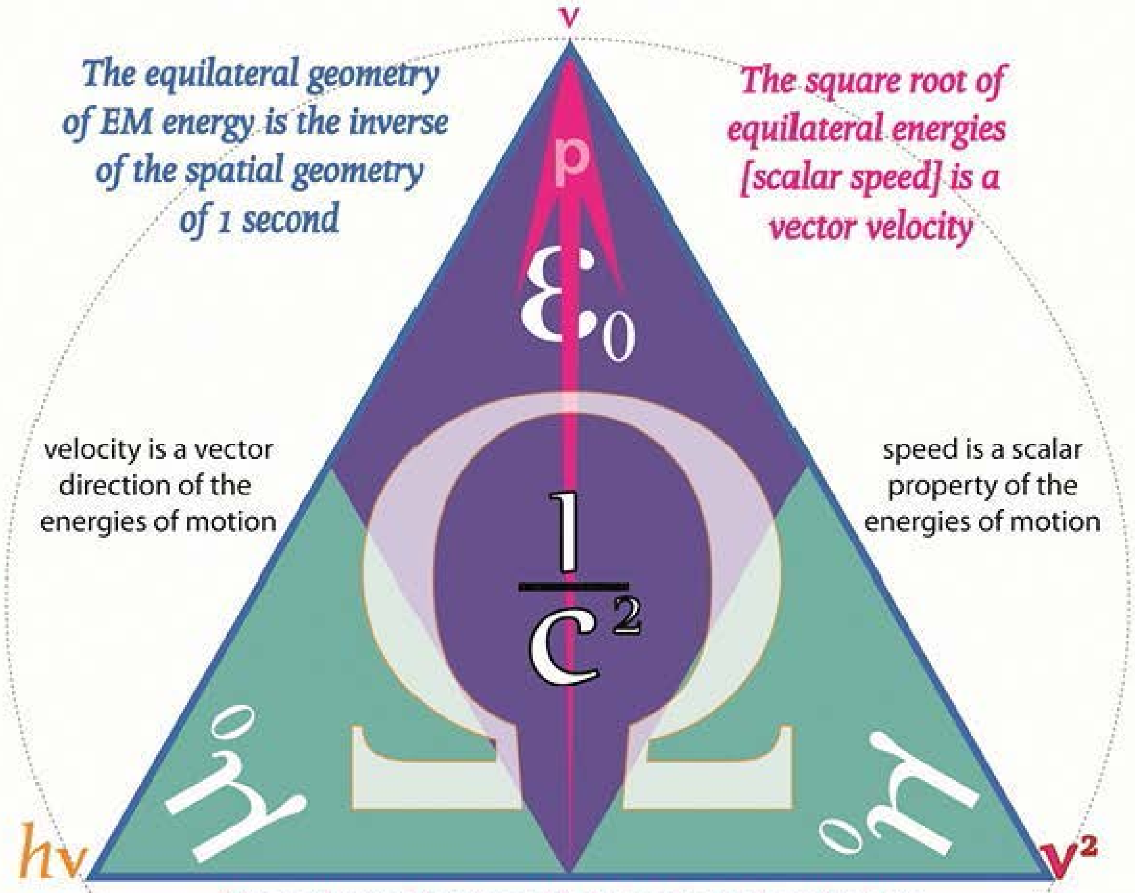

The maximum velocity possible due to electrical accelerations or the impedance of any spatial region that energy propagates through it is often referenced in physics as 'the speed of Light' IT IS NOT AN ABSOLUTE LIMIT or BARRIER

All EM waves radiate outwards in accord with Biot-Savart's law

c²

The velocity vectors of EM energies form the geometries of spatial co-ordinate systems

The speed of EM energies in a vacuum is defined as 299,792,458 meters per second (1,079,252,848.8 km/h).

$$\frac{K_E}{K_M} = 9e16 \, \frac{\frac{N\,m^2}{c^2}}{\frac{N\,s^2}{c^2}}$$

$$C^2 = 9e16 \, \frac{m^2}{s^2}$$

$$C = 299{,}792{,}458 \, \frac{m}{s}$$

The natural velocity of EM energies can be calculated from the field's EM permittivity & permeability and is affected by the medium it is propagating through

EM waveforms are bidirectional radiant emissions of 2D energy

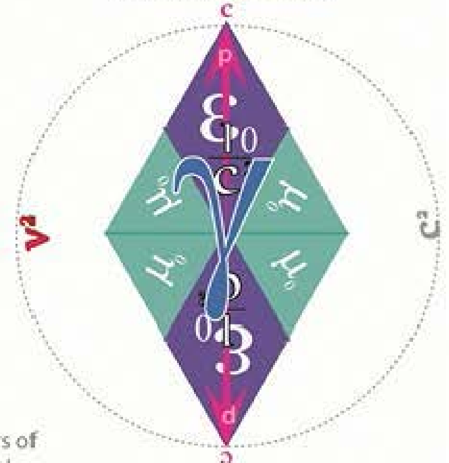

with velocity [Lorentz] variable energy contents

Quantum masses

mass quanta

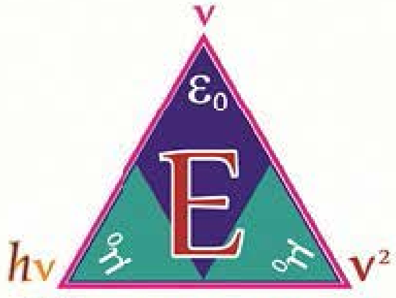

EM mass-energy quanta

Matter is a KEM standing wave propagating at c

EM Field planck quanta

$$[\varepsilon_0\mu_0]\,[m\Omega v^2]$$

mass velocity

1.112650056 e-17 s2 m-2

6.629432672 e-34 kg m2 s-2

Permittivity x Energy density

$$\frac{E}{c^2} = m$$

v
ε0
m
v²
mass
c²

mass = 7.376238634 x 10^{-51} Kg

Planck-Einstein

$$hv^2 = E = mv^2$$

$$kg\,\frac{m^2}{s}\,\frac{1}{s} \qquad kg\,\frac{m^2}{s^2} \qquad kg\,\frac{m^2}{s^2}$$

The quantum of mass-energy can be derived with several methodolgies using Tetryonic Geometry

Planck mass-energy geometries

quantised energy momenta

$$h$$

$$kg\,\frac{m^2}{s}$$

6.629432672 e-34 J

Planck quanta

$$[m\Omega v^2]$$

mass velocity

7.376238634 e-51 kg

$$kg\,\frac{m^2}{s^2}$$

$$m$$

scalar mass-energy

Charged mass-Matter topologies

The mass EM Field Planck quanta

$$T\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

of Matter ElectroMagnetic mass velocity

Matter Quanta

Tetryonic molar mass [Hydrogen] - 1 g

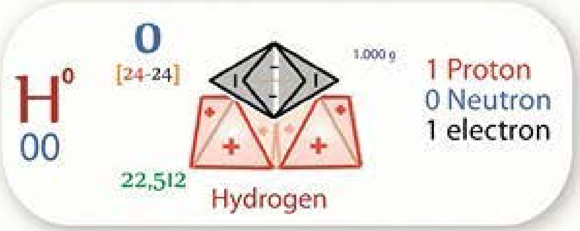

$$\text{Molar mass} = \frac{H_1\text{ Atomic mass}}{\text{Avagadro's No.}}$$

$$\text{mass } H_{mol}/Av = \frac{.001}{6.022141579 \times 10^{23}}$$

$$= 1.660538841 \times 10^{-27}\ \text{g/mol}$$

$$H_{mole}\,/\,\triangle[H] = \frac{1.660538841 \times 10^{-27}}{22{,}512}$$

mass = 7.376238634 x 10^{-32} KG

Avogadro - Mandeleev

2D mass-Energy equivalence

'The Speed of Light' is the Natural velocity of Energy propagation

Energy

$$mv^2 = hv^2$$

EM mass

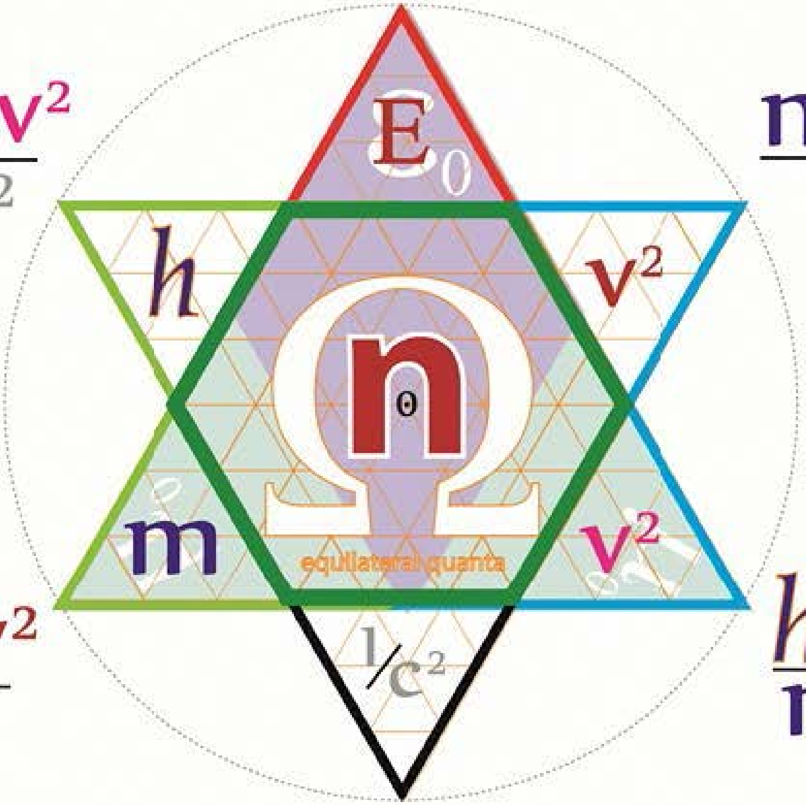
E_0
h
v^2
Ω
n_0
m
v^2
equilateral quanta
$^1/_{c^2}$

$$\frac{mv^2}{c^2}$$

$$\frac{mv^2}{h}$$

$$\frac{c^2}{v^2}$$

8.987551787 e16

1.112650056 e-17

$$\frac{m}{h}$$

$$\frac{hv^2}{c^2}$$

$$\frac{hv^2}{m}$$

$$\frac{m}{hv^2}$$

Photons

Matter

Tetryonics 16.07 - 2D mass-Energy equivalence

Copyright ABRAHAM [2008] - all rights reserved

Zero Point Fields

EM Field Planck quanta

$n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Quanta ElectroMagnetic mass velocity

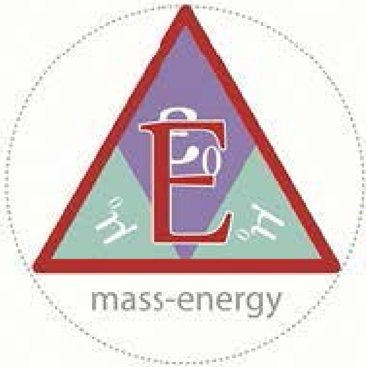
E
ε0
μ0
μ0
mass-energy

m=1π

radiant 4π mass-energy geometries

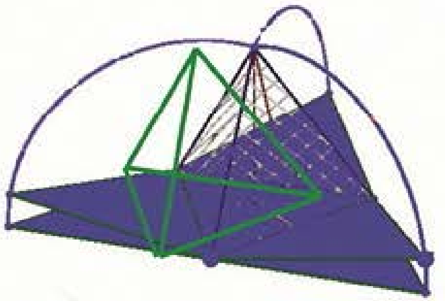

in standing wave topologies are 3D MATTER quanta

Photons

EM Field Planck quanta

$2n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Photons ElectroMagnetic mass velocity

m=2π

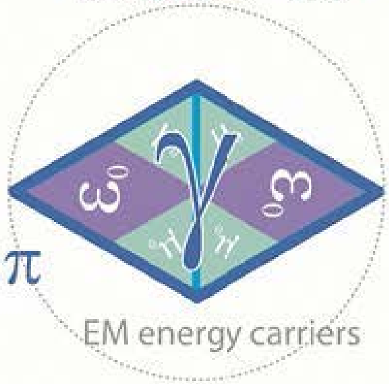
γ
ε0
ε0
μ0
μ0
EM energy carriers

[1, 2, 3, 4, 5, 6, 7, 8, 9,10......]

[2D EM masses are planar energy momenta]

[2, 8,18, 32, 50, 72, 98,128......]

The geometry of mass and Matter

[1, 3, 5, 7, 9, 11, 13,15, 17, 19.........]

[3D Matter has mass-energy momenta and volume]

[4, 8,12, 24, 36, 48, 72, 84, 168......]

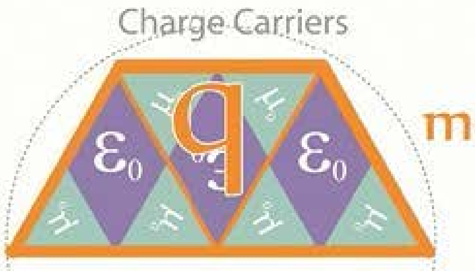
Charge Carriers
q
ε0
ε0
μ0
μ0
μ0
μ0
μ0

m=3π

EM Field Planck quanta

$ODDn\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Bosons ElectroMagnetic mass velocity

Bosons

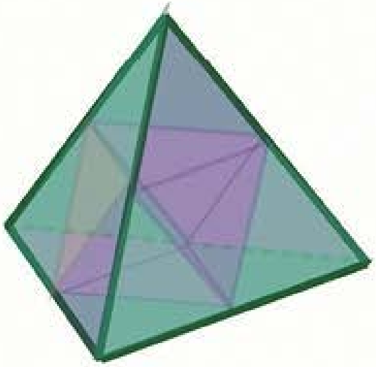

3D Matter topologies are standing waves of 2D mass-energies propagating at c

m=4π

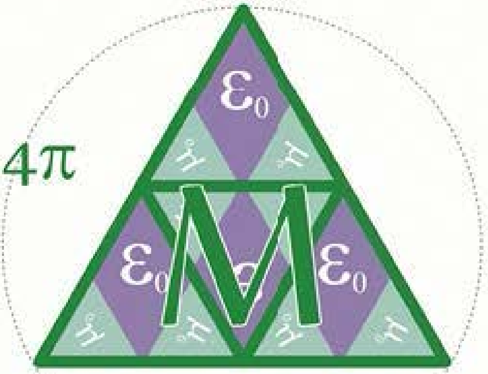
M
ε0
ε0
ε0
μ0
μ0
μ0
μ0
μ0
μ0

EM Field Planck quanta

$4n\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Matter ElectroMagnetic mass velocity

mass-Matter

Tetryonics 17.01 - The geometry of mass and Matter

Copyright ABRAHAM [2008] - all rights reserved

mass geometries & rest Matter topology

2D

m

mass

kg

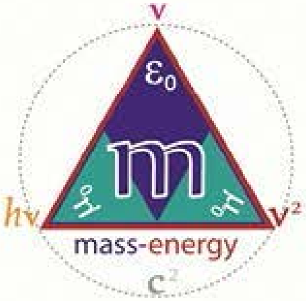
v
ε0
m
hv
v²
mass-energy
c²

M

Matter

3D

KG

Quanta number

1e19 v

Compton frequency

Matter is a three dimensional charged mass-energy topology

EM Field

Planck

quanta

$$M = 4n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter

ElectroMagnetic

mass

velocity

Energy density is mass - the term 'massless particle' is a misnomer

Quantum number

n

Nuclear Level

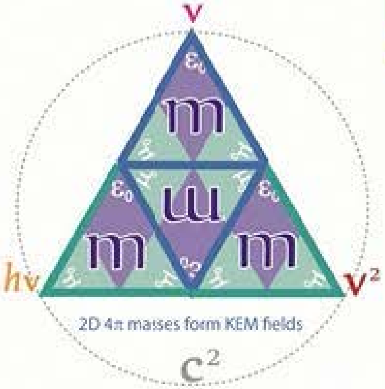
v
ε0
m
ε0
ω
ε0
m
m
hv
v²
2D 4π masses form KEM fields
c²

Kinetic Energies

RE = Matter + KE

Relativistic mass is the total EM energy content of a massive body (or system) in motion

The relativistic rest mass-energies of Matter are velocity invariant as
3D Matter is a EM standing wave with 2D mass- energy fascia
whose velocity of propagation is the speed of Light

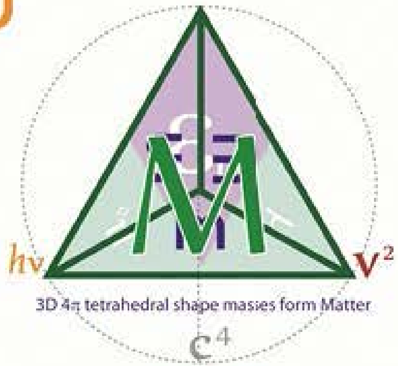
M
hv
v²
3D 4π tetrahedral shape masses form Matter
c⁴

Matter

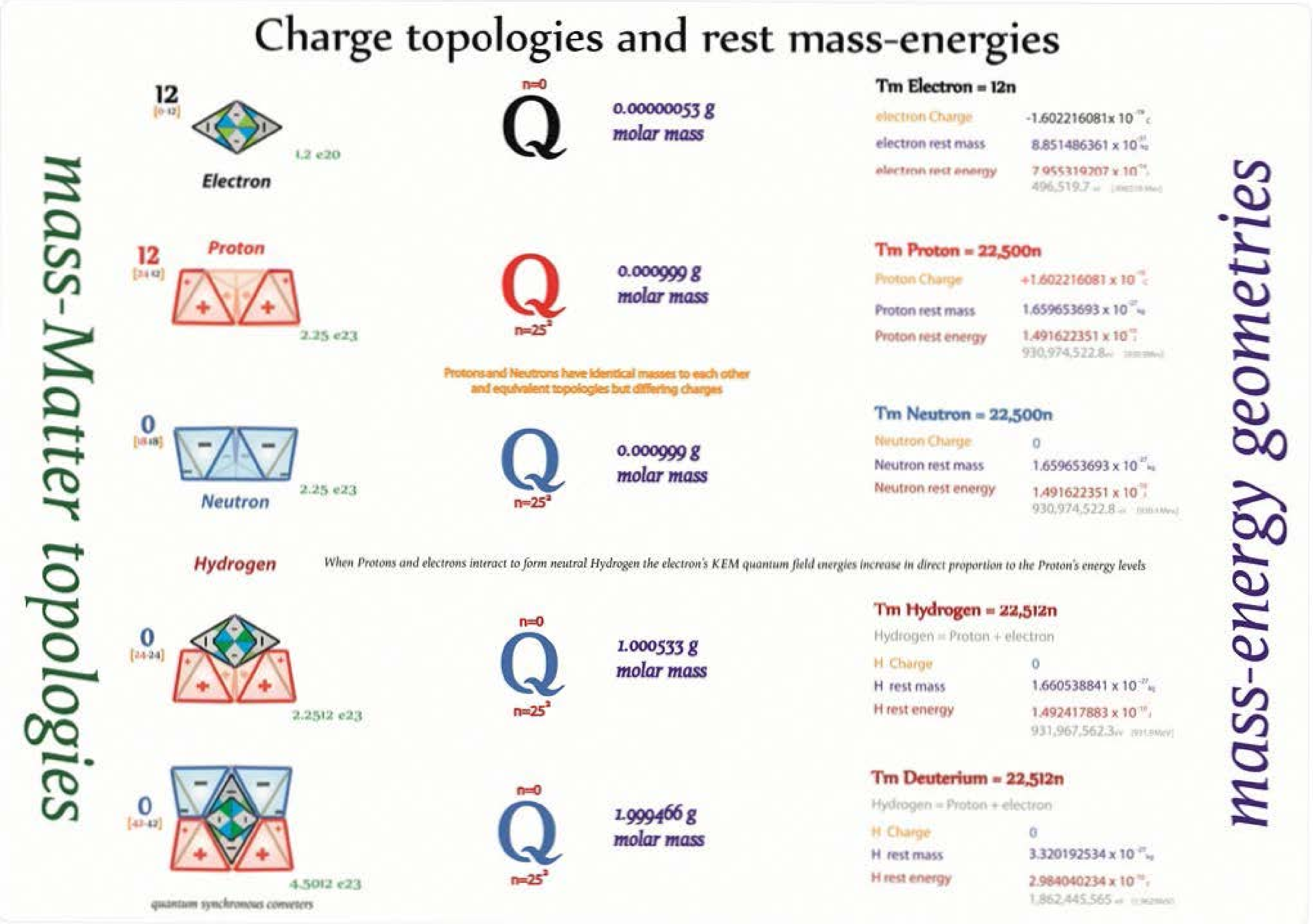

Charge topologies and rest mass-energies

mass-Matter topologies

mass-energy geometries

Electron

12 [0-12]

1.2 e20

n=0

Q

0.00000053 g
molar mass

Tm Electron = 12n

electron Charge	$-1.602216081 \times 10^{-19}$ C
electron rest mass	$8.851486361 \times 10^{-31}$ kg
electron rest energy	$7.955319207 \times 10^{-14}$ J 496,519.7 eV [illegible]

Proton

12 [24-12]

2.25 e23

Q

$n=25^2$

0.000999 g
molar mass

Tm Proton = 22,500n

Proton Charge	$+1.602216081 \times 10^{-19}$ C
Proton rest mass	$1.659653693 \times 10^{-27}$ kg
Proton rest energy	$1.491622351 \times 10^{-10}$ J 930,974,522.8 eV [illegible]

Protons and Neutrons have identical masses to each other
and equivalent topologies but differing charges

Neutron

0 [18-18]

2.25 e23

Q

$n=25^2$

0.000999 g
molar mass

Tm Neutron = 22,500n

Neutron Charge	0
Neutron rest mass	$1.659653693 \times 10^{-27}$ kg
Neutron rest energy	$1.491622351 \times 10^{-10}$ J 930,974,522.8 eV [illegible]

Hydrogen

When Protons and electrons interact to form neutral Hydrogen the electron's KEM quantum field energies increase in direct proportion to the Proton's energy levels

0 [24-24]

2.2512 e23

n=0

Q

$n=25^2$

1.000533 g
molar mass

Tm Hydrogen = 22,512n

Hydrogen = Proton + electron

H Charge	0
H rest mass	$1.660538841 \times 10^{-27}$ kg
H rest energy	$1.492417883 \times 10^{-10}$ J 931,967,562.3 eV [931.9MeV]

0 [42-42]

4.5012 e23

quantum synchronous converters

n=0

Q

$n=25^2$

1.999466 g
molar mass

Tm Deuterium = 22,512n

Hydrogen = Proton + electron

H Charge	0
H rest mass	$3.320192534 \times 10^{-27}$ kg
H rest energy	$2.984040234 \times 10^{-10}$ J 1,862,445,565 eV [illegible]

Tetryonics 17.03 - Charge topologies and rest mass-energies

In order to make an exact 1kg reference mass-Matter topology

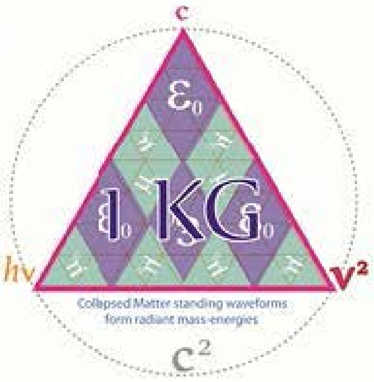
c
ε0
1 KG
hv
v²
Collapsed Matter standing waveforms form radiant mass-energies
c²

$$\frac{\overset{\text{mass}}{n\pi}}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

mass-energy

Energy per second

$$\boxed{mc^2 = E = hv^2}$$

$$m = E/c^2$$

mass is the scalar integral surface area of Matter topologies

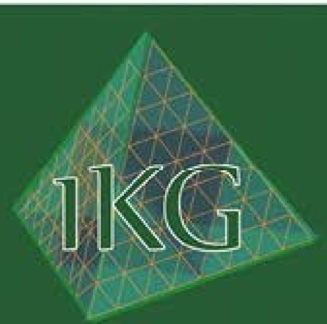

which is equivalent to exactly 1,000 molar Hydrogen masses

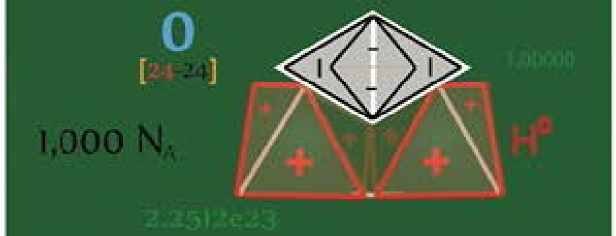
0
[24-24]
1.00000
1,000 N_A
H⁰
2.2512e23

whose Tetryonic Matter topology can be modelled using charged Planck mass-energy geometries

$$\underset{\text{1kg}}{M} = \overset{\text{grams per kg}}{1{,}000} \times \underset{\text{Avogadro's number}}{N_A}\,.\,T\pi\left[\left[m\overset{\text{Energy}}{\Omega}v^2\right]\right]$$

6.629432672 e-34 J

revealing that there are exactly

1.355704512 e50

Planck quanta in standing wave Matter

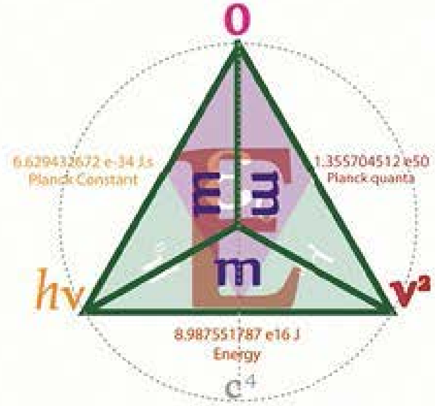
0
6.629432672 e-34 J.s
Planck Constant
1.355704512 e50
Planck quanta
E
m
hv
v²
8.987551787 e16 J
Energy
c⁴

$$\frac{\overset{\text{Matter}}{T\pi}}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

Matter-energy

Energy per second squared

$$\boxed{hv^2 = E = Mc^4}$$

$$M = E/c^4$$

Matter topologies are standing-waves of mass-energy

you need to create a standing-EM wave with a specific compton frequency

EM mass and Matter defined

There remains a lot of confusion over the exact definition of EM mass and Matter resulting in the frequent interchanging of one term for the other in physical processes

This must be clarified and the two terms must be properly defined in a manner that explains their derivation and physical properties in detail.

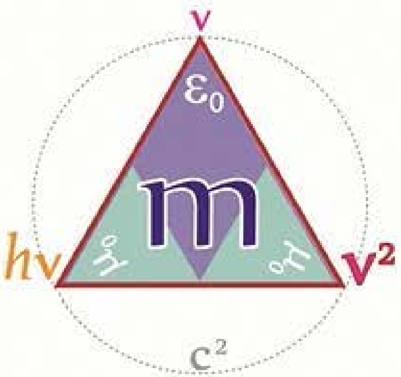

$$m = \frac{E}{c^2}$$

EM mass

3D Matter is comprised of 2D mass-energies

2D mass-energy cannot contain 3D Matter

Matter

$$M = \frac{4\pi E}{c^4}$$

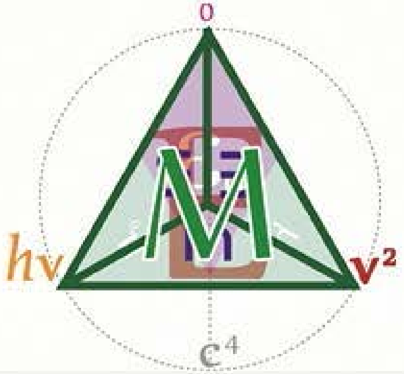

$$\underset{\text{mass}}{n\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

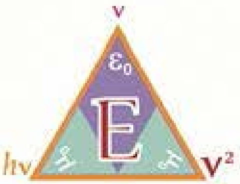

$$\underset{\text{Energy}}{n\pi}\Big[\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

2D mass-Energies

EM mass is a measure of equilateral scalar energy per unit of Time

$$\beta = \left[\frac{v}{c}\right]$$

mass is the Lorentzian velocity corrected energy content of Tetryonic geometries

All forms of EM mass-Energy are subject to Lorentz factor relationships

$$\beta^2 = \left[\frac{v}{c}\right]^2$$

Matter is Lorentz invariant as its mass-energies propagate in a 4π standing wave geometry

Bosons and Photons are not 'massless' they are 'Matterless' [2D waveforms]

Bosons

$$m_p = \frac{hv}{c^2}$$

Planck masses

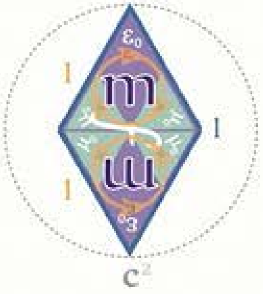

Photons

$$m_\gamma = \frac{hf}{c^2}$$

Photon mass

$$\underset{\text{Bosons}}{\text{ODD}\,\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

$$\underset{\text{EM waves}}{\text{EVEN}\,\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

$$\underset{\text{Matter}}{4n\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

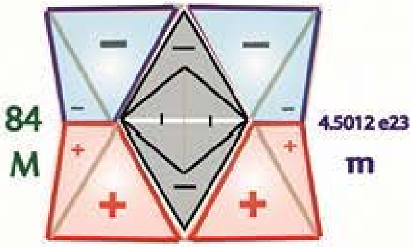

$$\underset{\text{Deuterium}}{84\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]$$

Matter is the 3D topology of standing wave 2D mass-energies

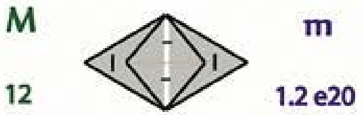

$$\underset{\text{electrons}}{12\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]\ \text{e-}$$

$$+$$

$$\underset{\text{Nuclei}}{72\pi}\Big[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\overset{\text{Planck}\ \ \text{quanta}}{\underset{\text{mass}\ \ \text{velocity}}{[m\Omega v^2]}}\Big]\ \text{Deuteron}$$

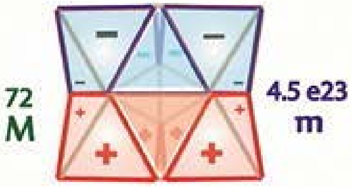

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Matter geometries

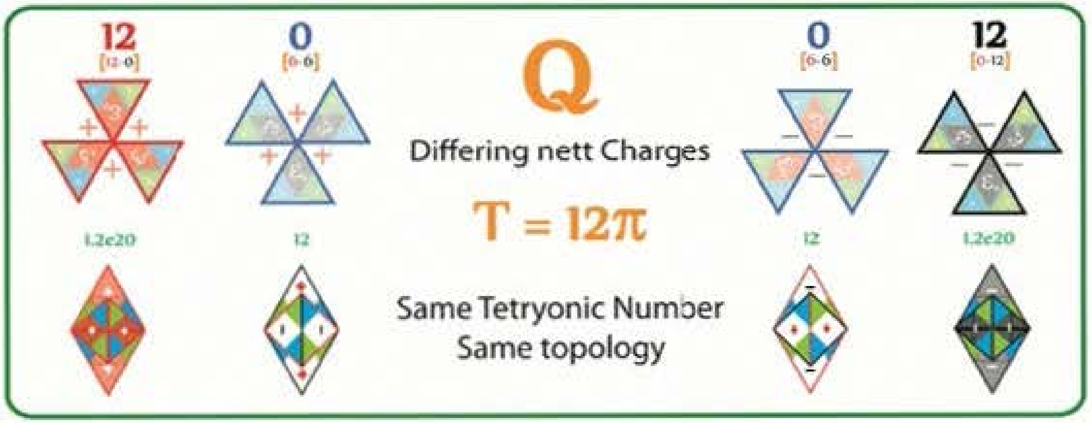

Q

Differing nett Charges

T = 12π

Same Tetryonic Number
Same topology

Matter Topology is determined by the geometry of charged Planck T[q] quanta

Any measurement of a system's mass is subject to velocity corrections

mass = EM energy density

$$\frac{\underset{\text{mass quanta}}{n\pi}\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]}{\underset{\text{Spatial geometry}}{c^2}}$$

n [7.376238634 e-51] kg

EM mass is a measure of 2D planar energies comprising the fascia of charged topologies

m

$$\overset{\text{ENERGY}}{n\pi}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

M

c^2
2D space

The mass - ENERGY - Matter contents of any physical system are all related through the spatial co-ordinate system used [which in turn is determined by the speed of light]

c^4
3D space

$$\overset{\text{mass}}{\frac{n\pi}{c^2}}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

E

$$\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

The Energy content of any physical system remains the same irrespective of the spatial co-ordinates used

Matter and Charge are velocity invariant properties

Differing Tetryonic charge numbers produce differing particle topologies

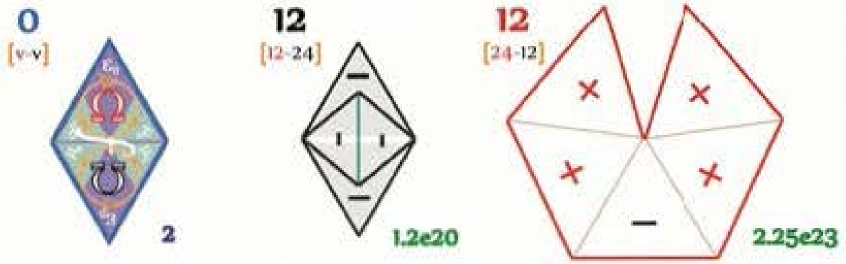

Photons $\frac{2\pi}{c^2}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$ $\frac{12\pi}{c^4}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$ Baryons $\frac{36\pi}{c^4}\left[\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$

As the energy content [levels] of particles and fields increase their intrinsic Planck quanta and EM mass increases but their charge geometries remain the same

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Energies

All 3D Matter particles are comprised of charged fascia whose energy content determines their 2D mass

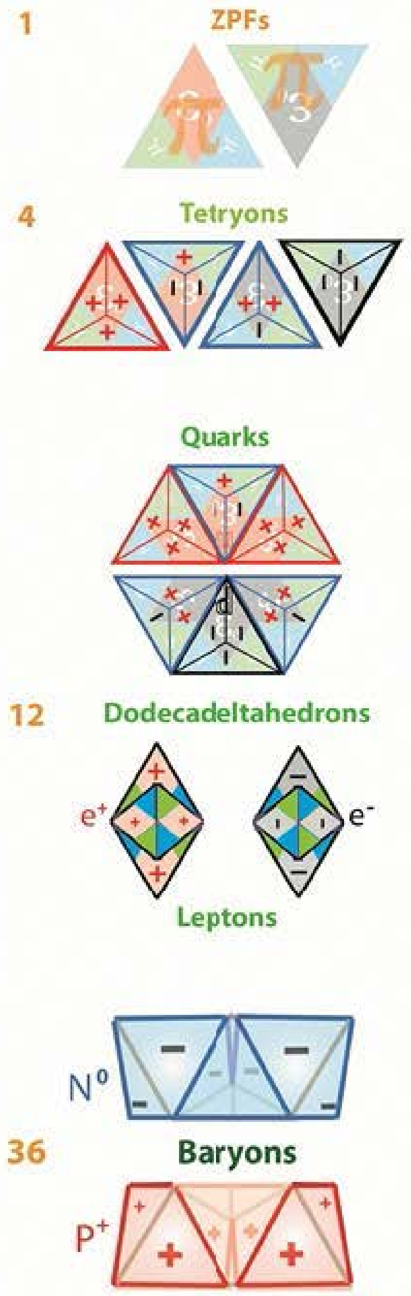
1
ZPFs
4
Tetryons
Quarks
12
Dodecadeltahedrons
e+
e-
Leptons
N0
36
Baryons
P+

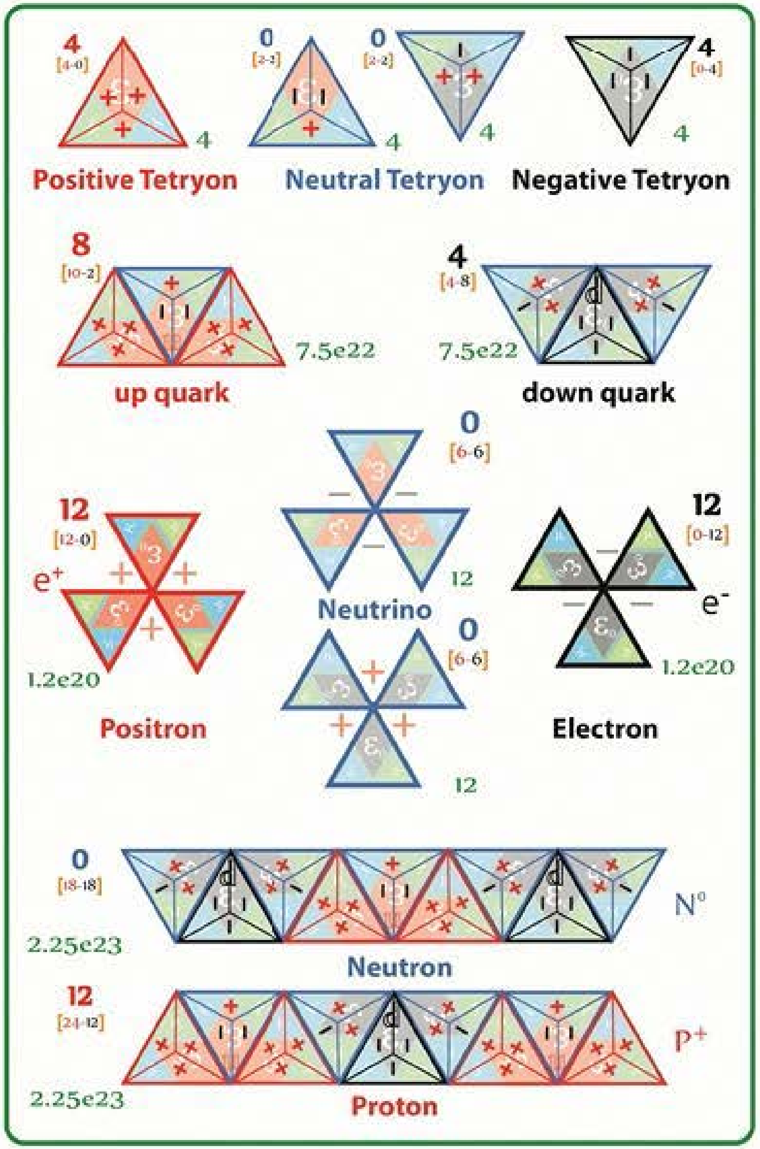
4
[4-0]
4
Positive Tetryon
0
[2-2]
4
0
[2-2]
4
4
[0-4]
4
Neutral Tetryon
Negative Tetryon
8
[10-2]
7.5e22
up quark
4
[4-8]
7.5e22
down quark
0
[6-6]
12
[12-0]
e+
1.2e20
Positron
12
Neutrino
0
[6-6]
12
12
[0-12]
e-
1.2e20
Electron
0
[18-18]
2.25e23
N0
Neutron
12
[24-12]
2.25e23
P+
Proton

3D Matter geometries

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Tetryons	4 charge fascia	4π	$2.651773069 \times 10^{-33}$ J	$2.950495454 \times 10^{-50}$ kg	4 quanta
Quarks	12 charge fascia	12π	$4.97207450 \times 10^{-11}$ J	$5.532178976 \times 10^{-28}$ kg	7.5×10^{22} quanta
Neutrinos	12 charge fascia	12π	$7.955319207 \times 10^{-33}$ J	$8.851486361 \times 10^{-50}$ kg	12 quanta
Leptons	12 charge fascia	12π	$7.955319207 \times 10^{-14}$ J	$8.851486361 \times 10^{-31}$ kg	1.2×10^{20} quanta
Baryons	36 charge fascia	36π	$1.491622351 \times 10^{-10}$ J	$1.659653693 \times 10^{-27}$ kg	2.250×10^{23} quanta
Deuterium ion	72 charge fascia	72π	$2.984040234 \times 10^{-10}$ J	$3.320192534 \times 10^{-27}$ kg	4.5012×10^{23} quanta

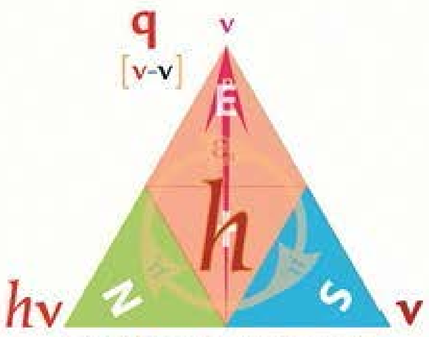

inertial mass-energy quanta

In physics there exists
2 forms of energy momenta
[linear momenta &
quantised angular]

MOMENTA

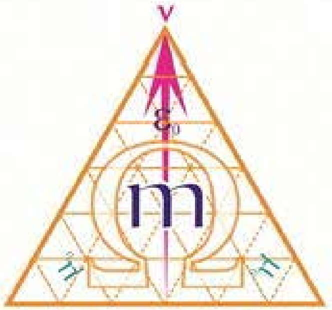

The nett scalar
energy momenta
required to effect
changes in velocity

ENERGY

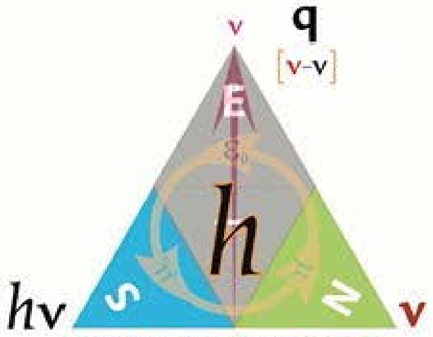

inertial mass-energy quanta

Net Charge is reflective of the total energy momenta
of any EM field or Matter subjected to acceleration
[inertial mass]

Inertial resistance to Force

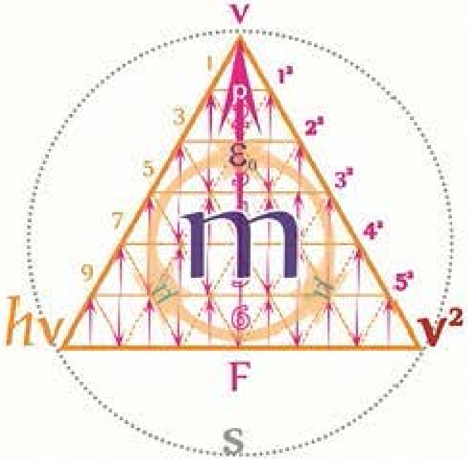

Any change in motion results in changes to
the Charge geometries creating in turn proportional
changes to the nett KEM mass-energy momenta components

$$F=ma$$

The inductive resistance of quanta in charged EM fields
to any changes to their nett mass-energy geometry
is the source ofwhat we term inertial mass

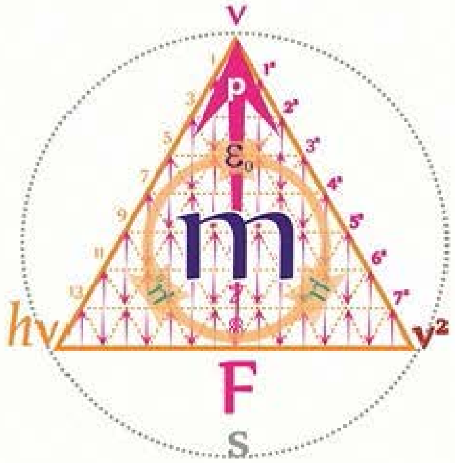

Inertia is the resistance of any physical object
to a change in its state of motion.

Force does work when it results in movement

$$\boxed{W = F.d = ma.d = E}$$

$$kg\,\frac{m}{s^2}\,m$$

classical mechanics — quantum mechanics

$$\Delta mv = \overset{\text{scalar}}{\underset{\text{energies}}{\Delta E}} = \Delta m\Omega v$$

linear momenta — angular momenta

$$kg\,\frac{m^2}{s^2}$$

The Planck constant is the quantum of Action

$$\boxed{E = m\Omega v^2 = hv^2}$$

$$kg\,\frac{m^2}{s}\,\frac{1}{s}$$

Changing an object's velocity results in a corresponding energy momenta change which relate to each other through its inertial mass

Tetryonics 17.08 - Inertial resistance to Force

EM field densities

1 Planck quanta of energy momenta

EM mass is a measure of the energy
content of any spatial co-ordinate system

$$\beta^2 = \left[\frac{v}{c}\right]^2$$

and is subject to
Lorentz velocity corrections
in 2D EM fields

6.629432672 e-34 J
1 planck quantum has a
EM mass-Energy of
7.376238634 e-51 kg
and
Quantised Angular Momentum
which creates Charge
1.33518 e-20 C

mass-Energy equivalence

mass velocity

v
ε_0
m
E
v²
c

7.376238634 e-51 kg

6.629432672e-34 j.s

$$mv^2 = E = hv^2$$

8.987551787 e16 [m/s]²

6.629432672 e-34 j

quanta/sec

Matter is only mass-energy in Tetrahedral topologies [T4π+]
[else it is nπ EM mass-energies that propagate away at c]

v
ε_0
m
Ω
v²
c

planck quanta

6.629432672 e-34 J

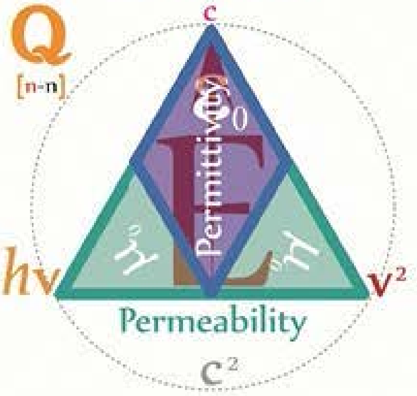

7.376238634 e-51 kg
radiant mass-energies

If reduced to a flat Euclidean space geometry
Matter topologies become radiant mass-energies

m mass	E ENERGY	M Matter
$\left[\frac{n\pi}{c^2}\rho\right]$	Planck quanta $[m\Omega v^2]$ mass velocity	$\left[\frac{4n\pi}{c^4}\rho\right]$

The inertial properties of electromagnetic mass - ENERGY & Matter
[can all be differentiated as energy densities per unit of time]
in any spatial co-ordinate system

2.984040234 e-10 J

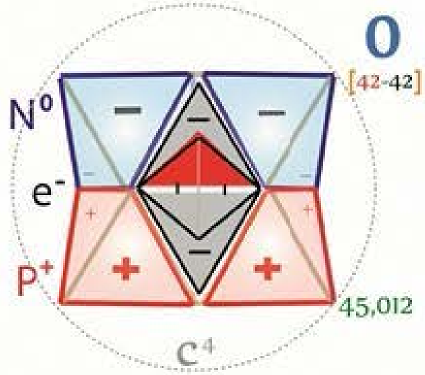

3.320192534 e-27 kg
rest mass-Matter

Copyright ABRAHAM [2008] - all rights reserved

m

[oddπ]hv

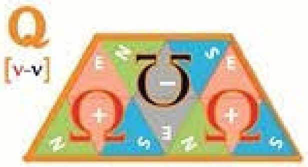

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge ElectroMagnetic mass velocity

Bosons

2D

Photons

m

[evenπ]hv

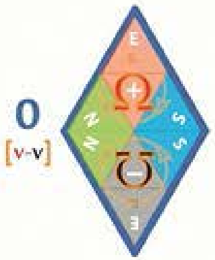

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

EM Field geometries & Matter topologies

M

[4nπ]hv

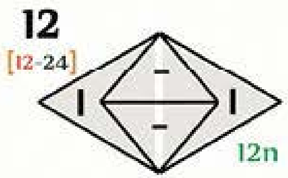

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Dodecyon ElectroMagnetic mass velocity

Leptons

3D

Baryons

M

[4nπ]hv

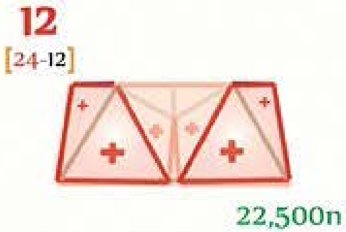

EM Field Planck quanta

$$36\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Baryons ElectroMagnetic mass velocity

2D planar Euclidean electromagnetic mass-energies propagate through the vacuum energy aether without interaction

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

2D Fields

3D Matter

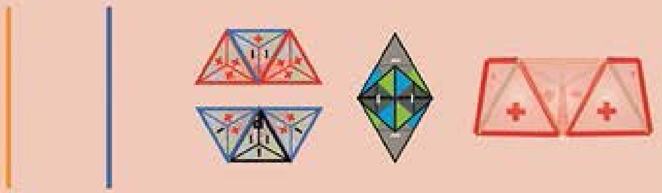

Tetryonic charged geometries provides a clear mechanism for providing all particles with their distinct properties of 2D inertial EM mass-energy and 3D Matter.

EM Field Planck quanta

$$T\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter ElectroMagnetic mass velocity

3D standing wave topologies of electromagnetic mass-energies interact with the vacuum energy aether at various angles through their charged [inductive] fascia

Tetryonics 17.10 - EM Field geometries and Matter topologies

Velocity invariant rest Matter

The Relativistic EM mass-energies of a system in motion is the sum of its invariant rest Matter and Kinetic Energies

Electron rest Matter | Kinetic Energy field

$$RE = \underset{\text{Fermions}}{12\pi}\frac{[[m\Omega v^2]]}{c^4} + \underset{\text{Photons}}{2\pi}\frac{[[m\Omega v^2]]}{c^2}$$

EM fields resulting from motion are subject to Lorentz correction

Kinetic mass-energies are divergent from invariant rest Matter topologies as a result of a particle's motion

At zero velocity the relativistic mass is equal to the invariant mass.

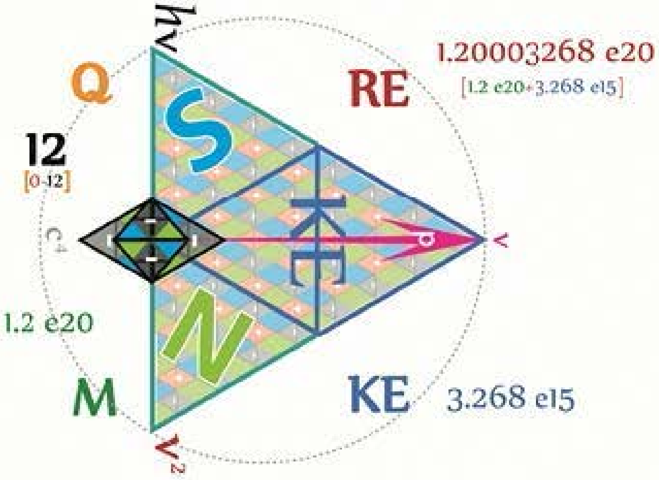
hv
Q
12
[0-12]
c⁴
1.2 e20
M
S
N
RE
1.20003268 e20
[1.2 e20+3.268 e15]
KE
KE
3.268 e15
v
v²

The energy which an object has due to its motion will not add mass into the invariant rest Matter of the particle in motion (it increases the total Planck quanta [EM mass-enegies] of its extended KEM field)

Electrostatic charges have no Magnetic Moments

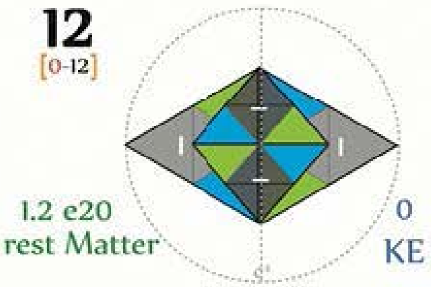
12
[0-12]
1.2 e20
rest Matter
0
KE
c⁴

All Matter are 3D standing wave topologies

$$M = \frac{E}{c^4}$$

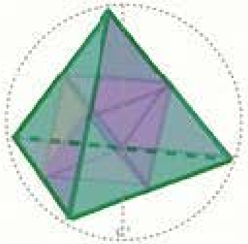

$E = Mc^4$

rest mass-Matter is composed of 4nπ standing wave topologies and is INVARIANT to velociiy changes

Wavelengths are proportional to EM energy content

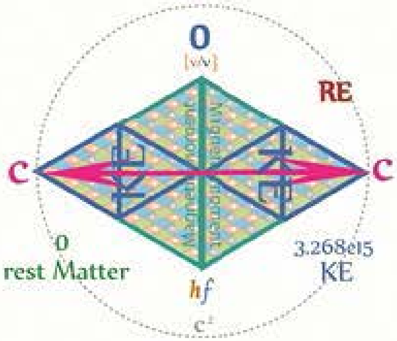
0
[v/v]
RE
c
c
KE
KE
Magnetic moment
Magnetic moment
0
rest Matter
hf
c²
3.268e15
KE

Photons are bidirectional Kinetic EM Fields

$$\frac{E}{c^2} = m$$

Matter in motion has a resultant velocity related Energy field that posseses the physical properties of Kinetic energy and Magnetic moments

$KEM = Mv^2$

KEM field Energy is directly relaıed to the Velocity changes of massive particles

All 2D EM fields and 3D Matter particles have ElectroMagnetic fields, inertial mass-energies & momenta resulting from their constituent equilateral Planck quanta which possess the additional physical properties of Compton Frequency and De Broglie Wavelength

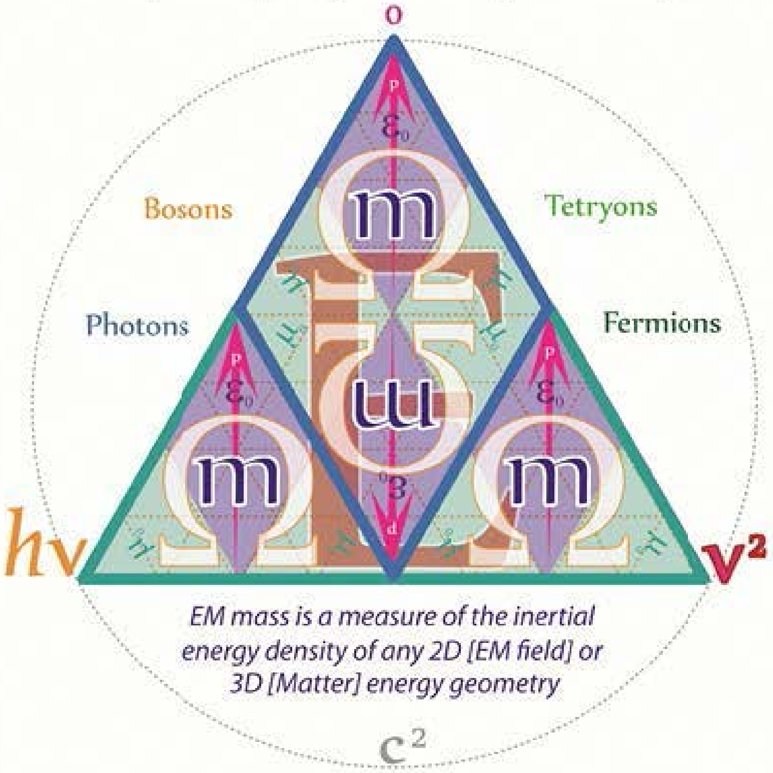

EM mass is a measure of the inertial energy density of any 2D [EM field] or 3D [Matter] energy geometry

$$m\Omega v^2 = E = mv^2$$

EM mass-energy momenta

The EM mass of an object is a fundamental property of the object; a numerical measure of its inertia; a fundamental measure of the energy density of an object.

mass geometries and Matter Topologies

All 3D Matter topologies contain 2D EM mass-energy geometries not all 2D EM mass-energy geometries form 3D Matter topologies

EM mass can be clearly defined as a measure of the energy density of any charged geometry

$$\underset{\text{Matter}}{T\pi}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}\cdot\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

3D Matter is any mass-energy geometry that creates a closed volume Topology (4nπ Tetryonic geometry)

EM waves [FIELDS] are distinguishable from Material Particles [MATTER] through their non-Tetrahedral topologies

$$\frac{E}{c^2} = m$$

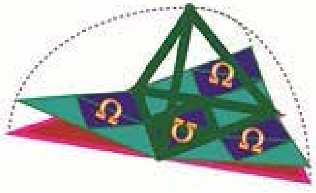

mass-energy is a conserved property
Matter is not conservative

$$M = \frac{E}{c^4}$$

The term 'massless' is a misnomer and should be discontinued in its use as all EM fields and Particles have EM mass [Energy quanta per second] geometries [other alternatives could be 2D, EM field, or Matterless]

Copyright ABRAHAM [2008] - all rights reserved

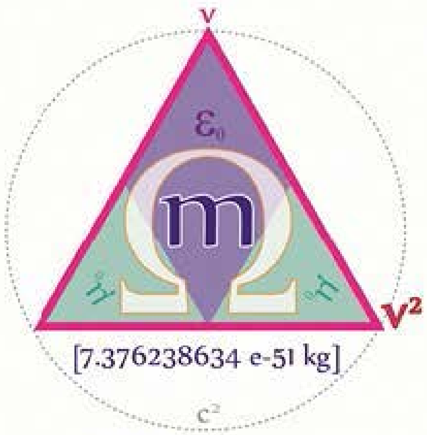
v
ε_0
m
Ω
μ_0
μ_0
v²
[7.376238634 e-51 kg]
c²

$$\frac{E}{c^2} = m$$

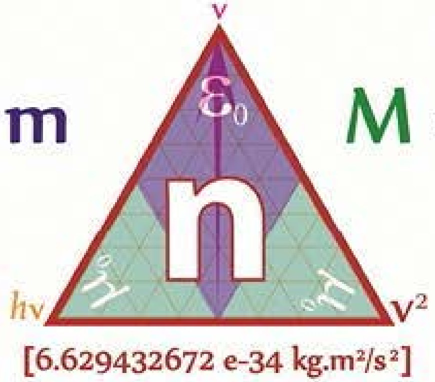
v
ε_0
n
μ_0
μ_0
hv
v²
[6.629432672 e-34 kg.m²/s²]

$$M = \frac{E}{c^4}$$

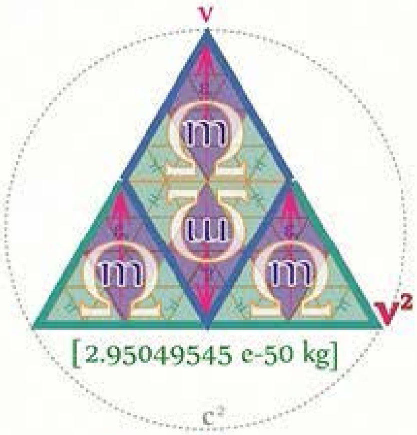
v
m
m
m
m
Ω
v²
[2.95049545 e-50 kg]
c²

EM mass

ENERGY

All attempts to clearly differentiate between planar EM mass-energy & 3D Matter in physics have met with limited success until the introduction of charged geometries

Collapsed Matter

mv mass-velocity squared is Energy ← nπ → Planck quanta is Energy content **mΩ**

mass-ENERGY-Matter equivalence

Newton and Leibniz measured Energy as a velocity related function of mass

Planck and Einstein viewed mass-energies as quantised particles with momenta

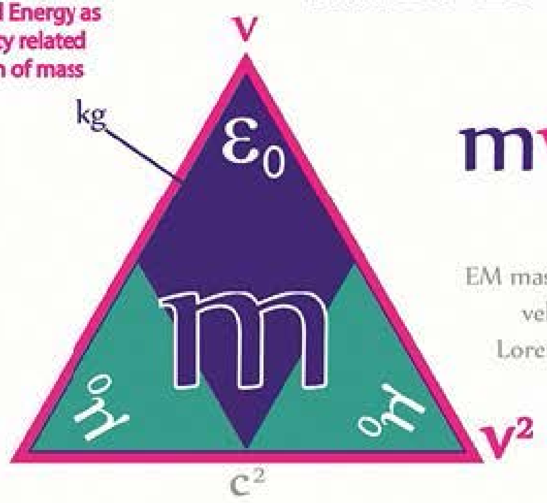
v
kg
ε_0
m
μ_0
μ_0
v²
c²

Energy momenta

$$mv^2 = E = hv^2$$

EM masses are subject to velocity related Lorentz corrections

Matter is a standing-wave of electromagentic energies and is Lorentz invariant

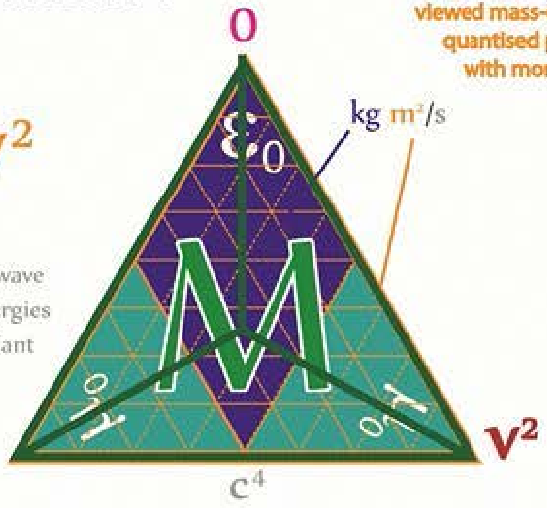
0
kg m²/s
ε_0
M
μ_0
μ_0
v²
c⁴

EM mass quantum

$$\frac{n\pi}{c^2}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

$$n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Matter quantum

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Tetryonics 17.13 - EM mass-Energy-Matter equivalence

m

mass geometries

All forms of mass-ENERGY-Matter are defined by their charged geometries and the spatial co-ordinate systems used to measure their physics

M

Matter topologies

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

Unified Field Equation

for Tetryonic EM mass-Energy-Matter

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

Tetryonic EM Field Planck quanta

$$T\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

geometry ElectroMagnetic mass velocity

EM Field Planck quanta

$$n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

mass ElectroMagnetic mass velocity

EM Field Planck quanta

$$\text{ODD}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Bosons ElectroMagnetic mass velocity

EM Field Planck quanta

$$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Photons ElectroMagnetic mass velocity

EM Field Planck quanta

$$\text{EVEN}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

EM waves ElectroMagnetic mass velocity

bosons & photons

nπ

EM waveforms

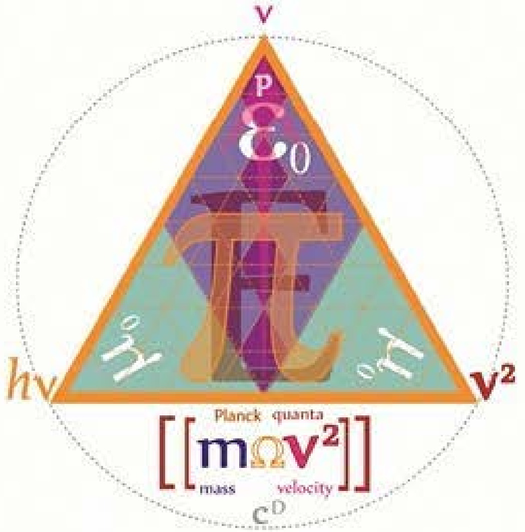

spatial co-ordinate systems are defined by the velocity of energy

Fermions

Tπ

Particles

EM Field Planck quanta

$$4\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Tetryons ElectroMagnetic mass velocity

EM Field Planck quanta

$$12\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Dodecyons ElectroMagnetic mass velocity

EM Field Planck quanta

$$24\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Mesons ElectroMagnetic mass velocity

EM Field Planck quanta

$$36\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Baryons ElectroMagnetic mass velocity

Tetryonics 17.14 - Unified Field Equation

Copyright ABRAHAM [2008] - all rights reserved

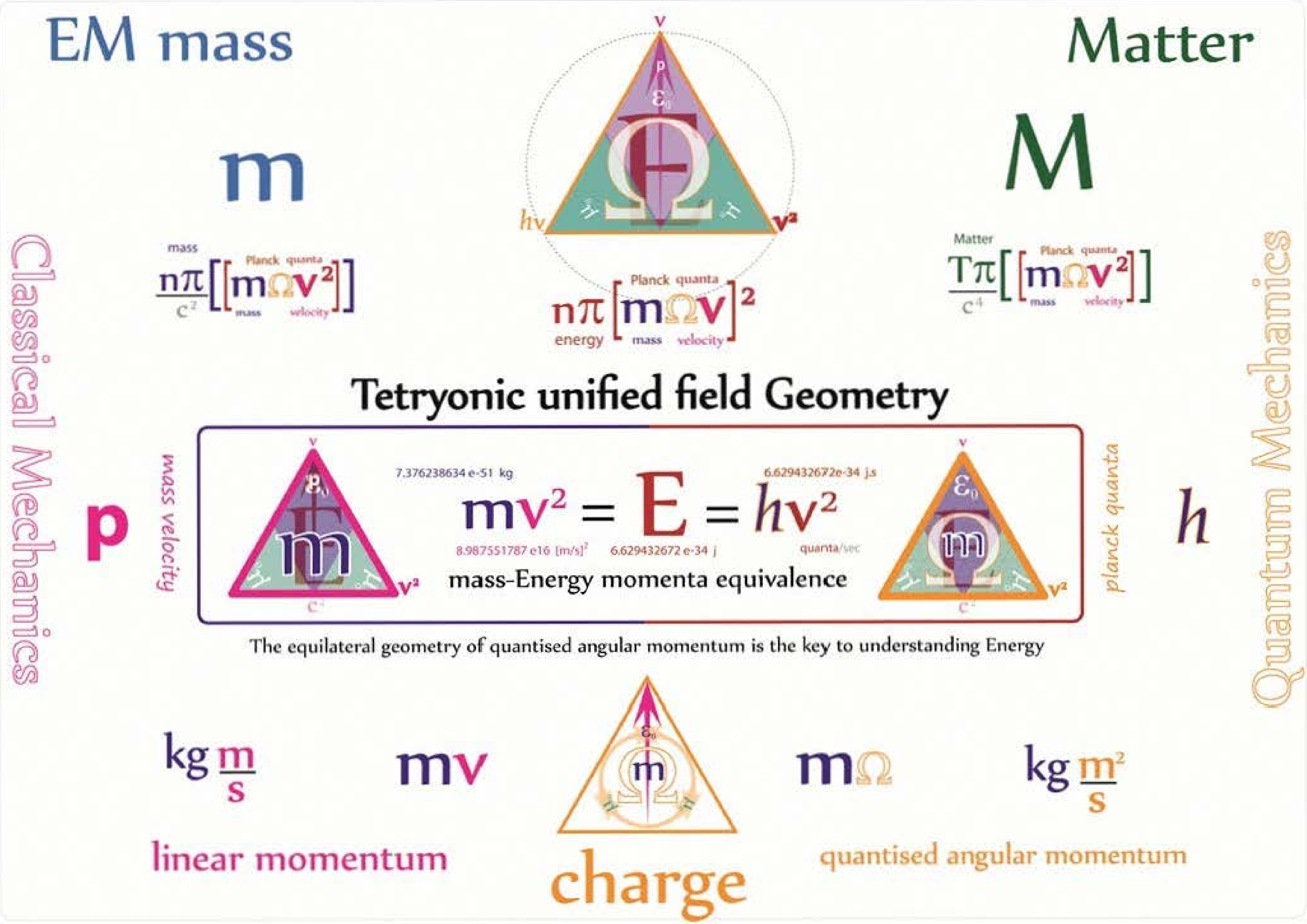

EM mass

Matter

m

M

mass

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

Planck quanta
mass velocity

$$n\pi[m\Omega v]^2$$

energy
Planck quanta
mass velocity

Matter

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

Planck quanta
mass velocity

Classical Mechanics

Quantum Mechanics

Tetryonic unified field Geometry

p

mass velocity

7.376238634 e-51 kg

6.629432672e-34 j.s

$$mv^2 = E = hv^2$$

8.987551787 e16 [m/s]²

6.629432672 e-34 j

quanta/sec

mass-Energy momenta equivalence

planck quanta

h

The equilateral geometry of quantised angular momentum is the key to understanding Energy

$\text{kg}\,\frac{\text{m}}{\text{s}}$

mv

mΩ

$\text{kg}\,\frac{\text{m}^2}{\text{s}}$

linear momentum

charge

quantised angular momentum

Tetryonics 17.15 - Unified field geometries

Copyright ABRAHAM [2008] - all rights reserved

$kg\,\frac{m^2}{s}\,\frac{1}{s}$

Planck quanta

$[m\Omega v^2]$

mass velocity

$kg\,\frac{m^2}{s^2}$

m

EM Field Planck quanta

$n\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

mass ElectroMagnetic mass velocity

kg

E

Energy

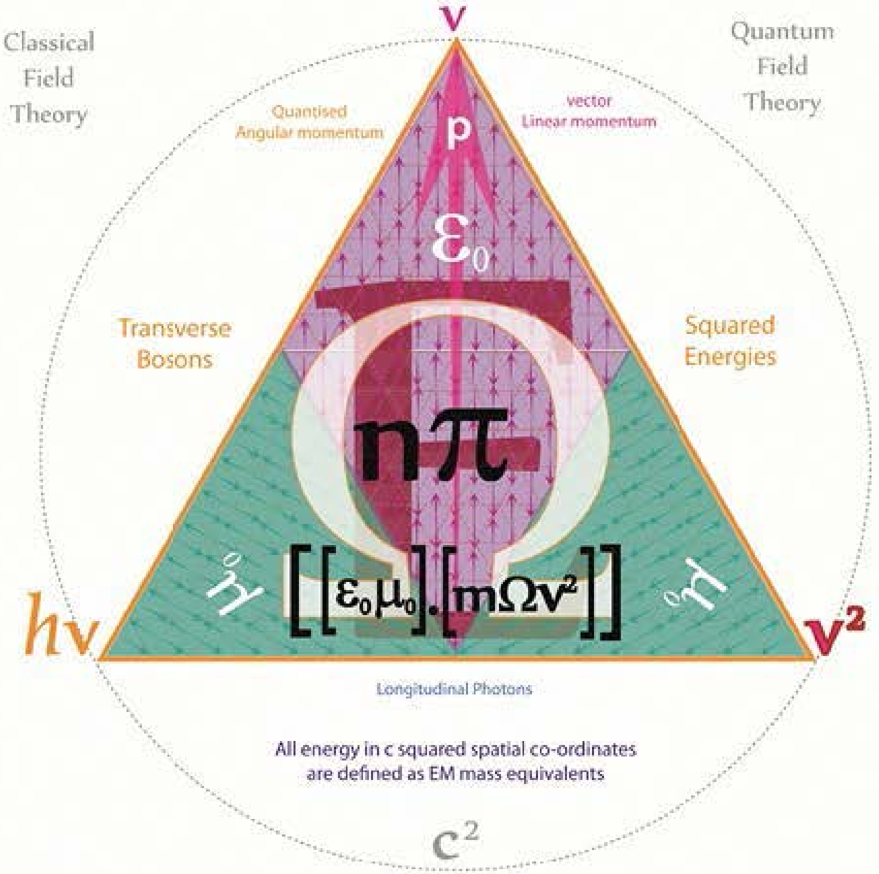

Classical Field Theory
v
Quantum Field Theory
Quantised Angular momentum
p
vector Linear momentum
ε0
Transverse Bosons
Squared Energies
nπ
μ0
[[ε0μ0].[mΩv²]]
μ0
hv
v²
Longitudinal Photons
c²

All energy in c squared spatial co-ordinates are defined as EM mass equivalents

Q

Charge

$\frac{s^2}{m^2}\quad\frac{m}{s^2}$

Charge

$[\varepsilon_0\mu_0\Omega]$

ElectroMagnetic QAM

s

h

$[m\Omega]$

mass QAM

$kg\,\frac{m^2}{s}$

Unified Field Components

The Unified field equation can be re-arranged to reveal a multitude of physical properties and relationships previously poorly understood.

Highlighting the fact that all of the constants and properties of both Classical and Quantum Mechanics are in fact geometric properties of Energy

ElectroMagnetic field

$[\varepsilon_0\mu_0]$

$\frac{1}{c^2}$

$\frac{s^2}{m^2}$

QAM

$\frac{m^2}{s}\quad\Omega$

v^2

velocity squared

Velocity - quanta

v^2

quanta/sec

Tetryonics 17.16 - UFE components

EM mass Relationships

v	Velocity	$\frac{m}{s}$
a	Acceleration	$\frac{m}{s^2}$
	m	
p	Momentum	$kg\,\frac{m}{s}$
F	Force	$kg\,\frac{m}{s^2}$

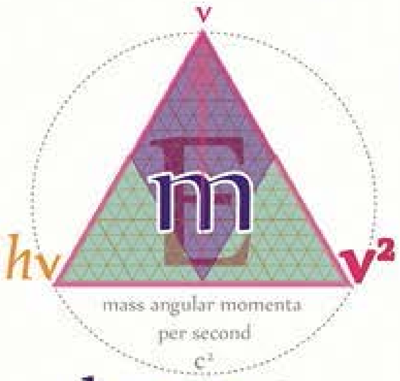
v
E
m
hv
v²
mass angular momenta
per second
c²

$$kg\ \frac{m^2}{s}\ \frac{1}{s}$$

EM mass is revealed to be the scalar property of 2D Energy waveforms that is at the core of many important physical processes and measurements

Ω	Quantum Angular Momentum	$\frac{m^2}{s}$
v^2	Celeritas squared	$\frac{m^2}{s^2}$
	m	
h	Planck's Constant	$kg\,\frac{m^2}{s}$
E	Energy	$kg\,\frac{m^2}{s^2}$

f	Frequency	$\frac{1}{s}$
T	Period	s
	m	
I	Current	$\frac{kg}{s}$
C	Charge	$\frac{s}{kg}$

E/second

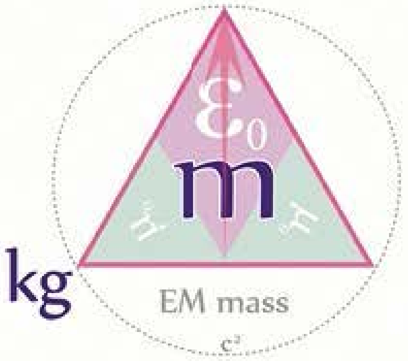
ε₀
m
kg
EM mass
c²

E/second²

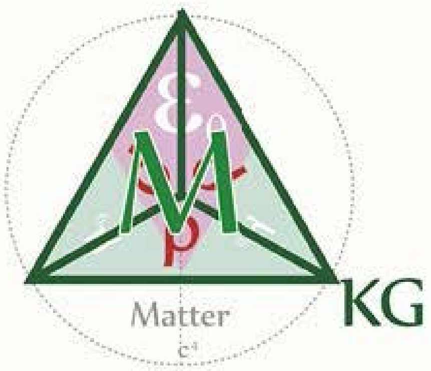
ε₀
M
ρ
KG
Matter
c⁴

2D mass geometries should never be confused with 3D Matter topoplogies; nor should the terms be used in exchange for each other

Tetryonics 17.17 - EM mass relationships

Relativistic mass-Matter

Rest Matter

$$\underset{\text{mass-Matter}}{T\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

3D rest mass-Matter topolgy =
closed volume of 2D mass-energies

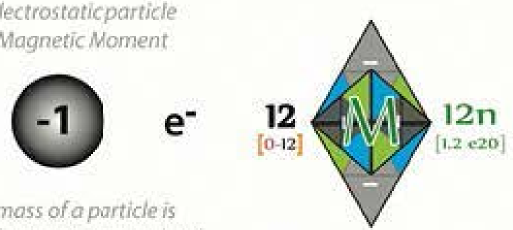

3D Electrostatic particle
No Magnetic Moment

rest mass of a particle is
dependent on its Energy level

All Matter is a Tetryonic standing-wave charged geometry occupying a volume in 3D spherical space

$$\left[\;\overset{\text{Matter}}{\frac{T\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v}}{}^2\right]\right] \Big\downarrow \overset{\text{Dodecyon}}{\frac{12\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v}}{}^2\right]\right]\;\right]$$

rest mass-Matter is velocity invariant
(not subject to Lorentz corrections)

The property of Matter cannot be measured using a planar [c squared] spatial co-ordinate system

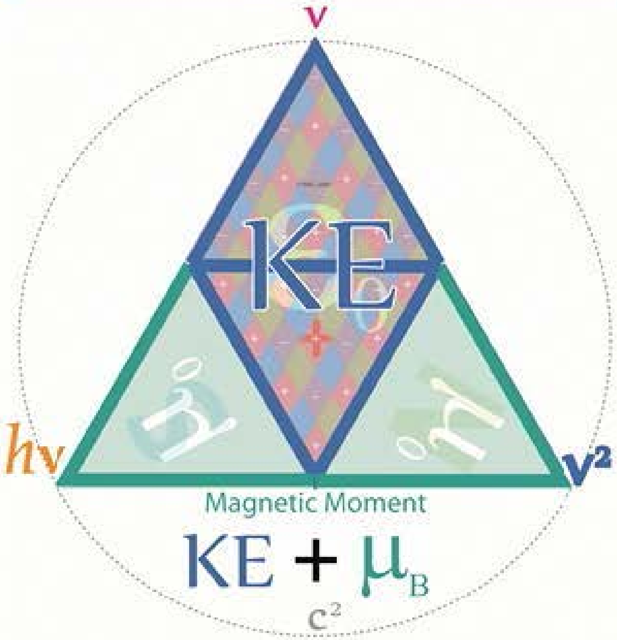

$$\frac{KE + \mu_B}{c^2}$$

A KEM field has Electric and Magnetic quanta whose total energy quanta is directly related to the square of the particle's velocity

$$4\pi\left[\frac{\overset{\text{mass}}{m}\Omega\overset{\text{velocity}}{v}{}^2}{c^2}\right]$$

All KEM fields are subject to
Lorentz corrections

Kinetic Energies

$$\underset{\text{KEM field}}{4n\pi}\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Matter-Energy

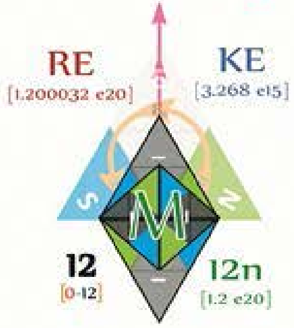

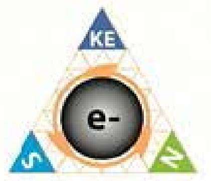

electron KEM model

Relativistic mass Energy = rest Matter + Kinetic EM Energies

Electron rest Matter

$$\left[\;\overset{\text{Lepton}}{\frac{12\pi}{c^4}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v}}{}^2\right]\right] + \overset{\text{KEM wave}}{\frac{4\pi}{c^2}}\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v}}{}^2\right]\right]\;\right]$$

Kinetic Energies

KEM mass-energies are velocity dependent
(subject to Lorentz corrections)

Tetryonics 17.18 - Relativistic mass-Matter

Tetryonic mass & Matter

Historically interchanged due to the lack of proper definitions
the physical properties of EM mass & Matter can now be firmly defined
with respect to their energy equivalence and spatial geometries

m

mass is a measure of the 2D planar energy content of any physical system

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

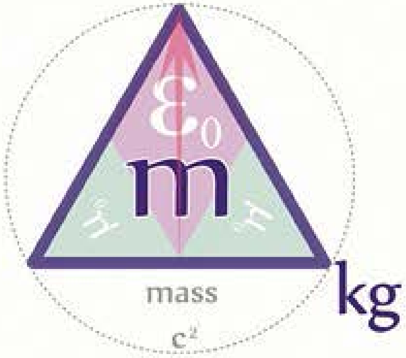

E/second

2D
radiant equilateral geometry
of EM mass-ENERGY momenta
EM mass

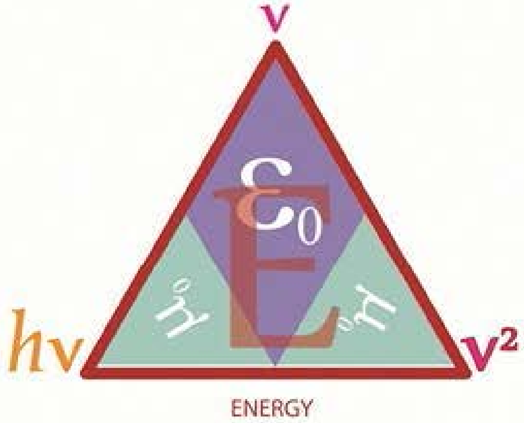

EM mass should replace the generic term mass with reference to ElectroMagnetic energy densities

MATTER is a geometric
4nπ standing wave topology of
EM mass-energy geometries

The electromagnetic energies of rest Matter
is never 'at rest' as the electromagnetic
field energies creating mass-Matter
topologies alway propagate at c

M

Matter is a measure of the 3D volumetric energy content of any physical system

Matter Planck quanta

$$\frac{T\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

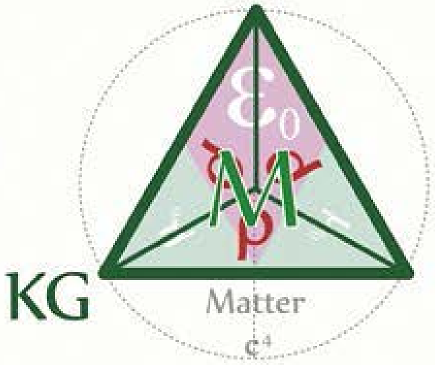

E/second²

3D
standing-wave topology
of EM mass-ENERGY momenta
rest Matter

Tetryonics 17.19 - Tetryonic mass & Matter

Copyright ABRAHAM [2008] - all rights reserved

Photoelectron KEM fields

The linear momenta produced as a result of energy momenta in KEM fields is converted into angular momentum when leptons are bound to atomic nuclei

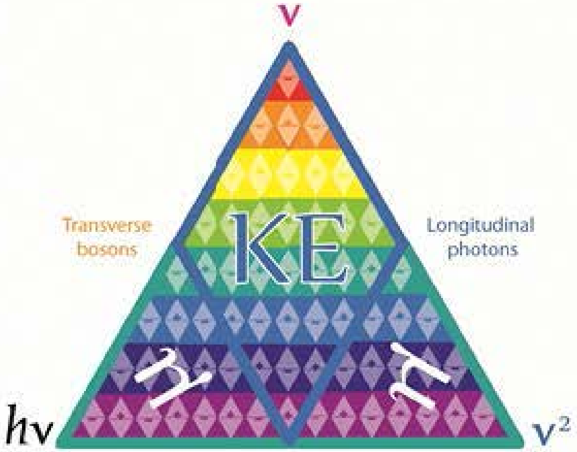
v
Transverse bosons
Longitudinal photons
KE
hv
v²

SPIN UP

The EM mass-energy content of Baryons directly influence the KEM field energy levels of bound Leptons

EM mass-energy-momenta

Changes in the momenta of bound Leptons [linear & quantised angular]

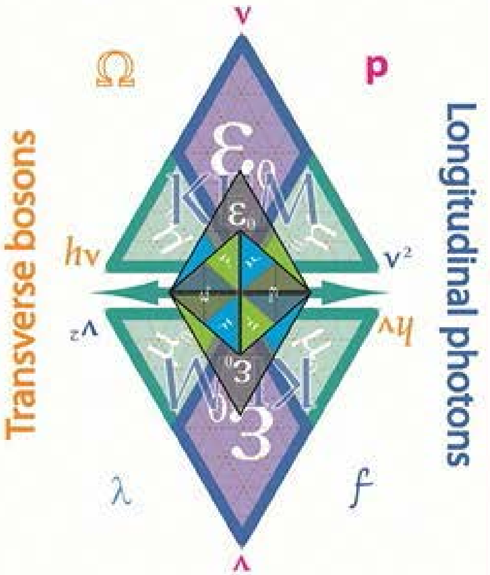
v
Ω
p
Transverse bosons
Longitudinal photons
ε₀
KEM
hv
v²
λ
f
v

prouduces spectral lines of varying frequencies

and Spin are always conserved

The motion of leptons within atomic nuclei produces Magnetic moments at various orientations to the nuclear magneton

SPIN DOWN

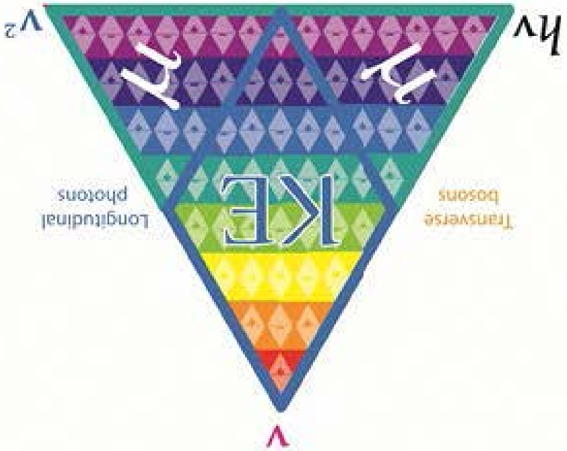

If an electron is 'ejected' from the Nuclei it will obey conservation of EM mass-energy momentum

ie Its ejected energy-momentum equals the absorbed photon's energy-momenta [minus the work energy required to free it]

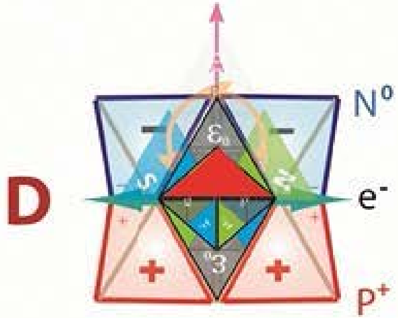

UP
spin

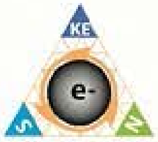

Parallel
Magnetic Moments

Nuclear magneton

DOWN
spin

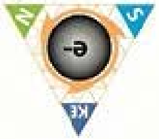

Antiparallel
Magnetic moments

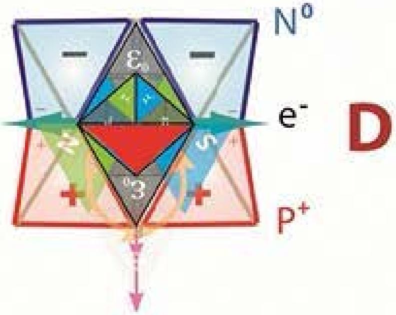

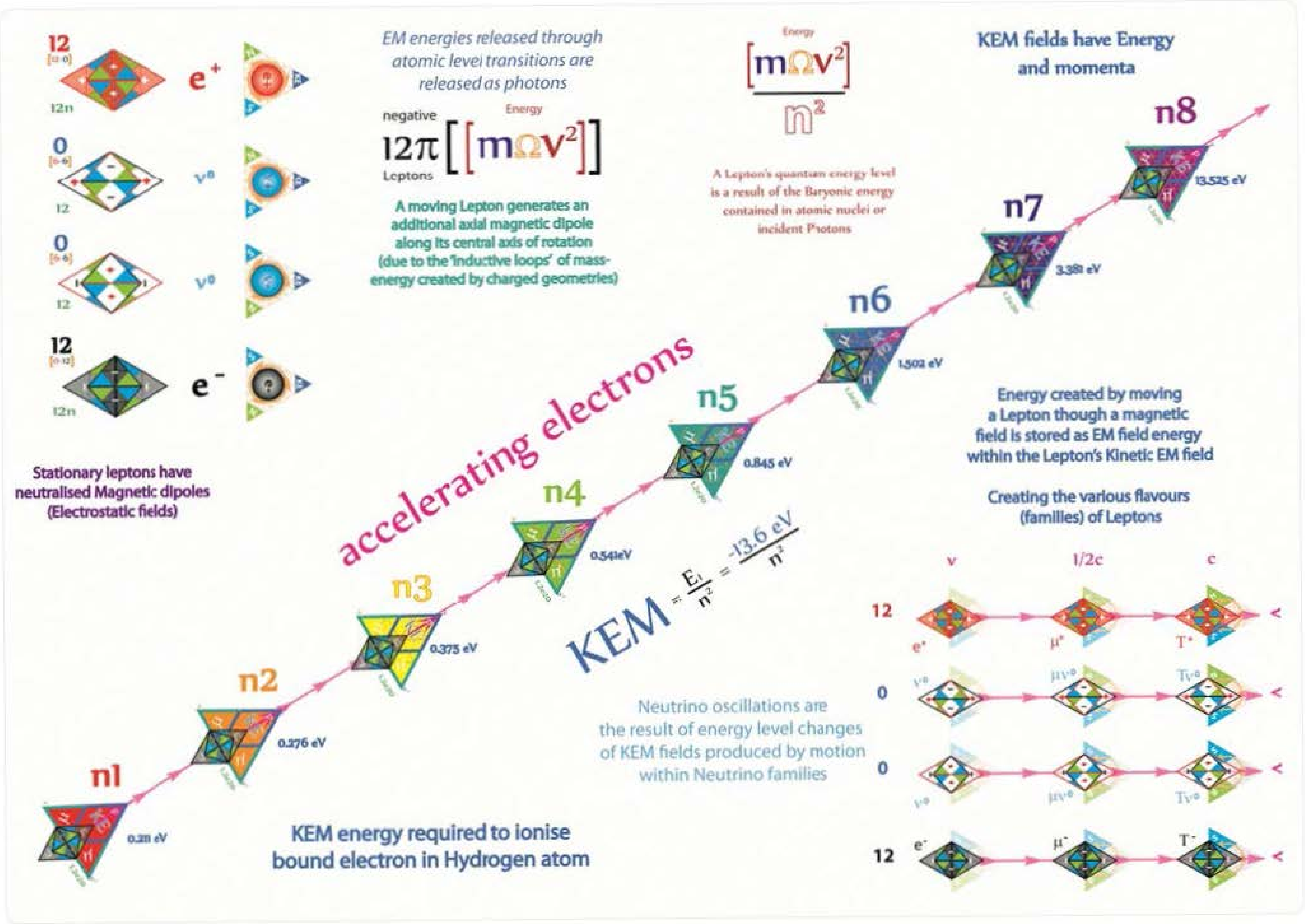
12
[12-0]
12n
e+
0
[6-6]
12
ν0
0
[6-6]
12
ν0
12
[0-12]
12n
e-
n1
0.211 eV
n2
0.276 eV
n3
0.375 eV
n4
0.541eV
n5
0.845 eV
n6
1.502 eV
n7
3.381 eV
n8
13.525 eV
KE
accelerating electrons
v
1/2c
c
12
e+
μ+
T+
0
ν0
μν0
Tν0
0
ν0
μν0
Tν0
12
e-
μ-
T-

EM energies released through atomic level transitions are released as photons

negative Energy

$$12\pi\left[\left[m\Omega v^2\right]\right]$$

Leptons

A moving Lepton generates an additional axial magnetic dipole along its central axis of rotation (due to the 'inductive loops' of mass-energy created by charged geometries)

Energy

$$\frac{\left[m\Omega v^2\right]}{n^2}$$

A Lepton's quantum energy level is a result of the Baryonic energy contained in atomic nuclei or incident Protons

KEM fields have Energy and momenta

Stationary leptons have neutralised Magnetic dipoles (Electrostatic fields)

$$KEM \approx \frac{E_1}{n^2} = \frac{-13.6\ eV}{n^2}$$

Energy created by moving a Lepton though a magnetic field is stored as EM field energy within the Lepton's Kinetic EM field

Creating the various flavours (families) of Leptons

Neutrino oscillations are the result of energy level changes of KEM fields produced by motion within Neutrino families

KEM energy required to ionise bound electron in Hydrogen atom

Tetryonics 18.02 - accelerating electrons

Leptronic Oscillations

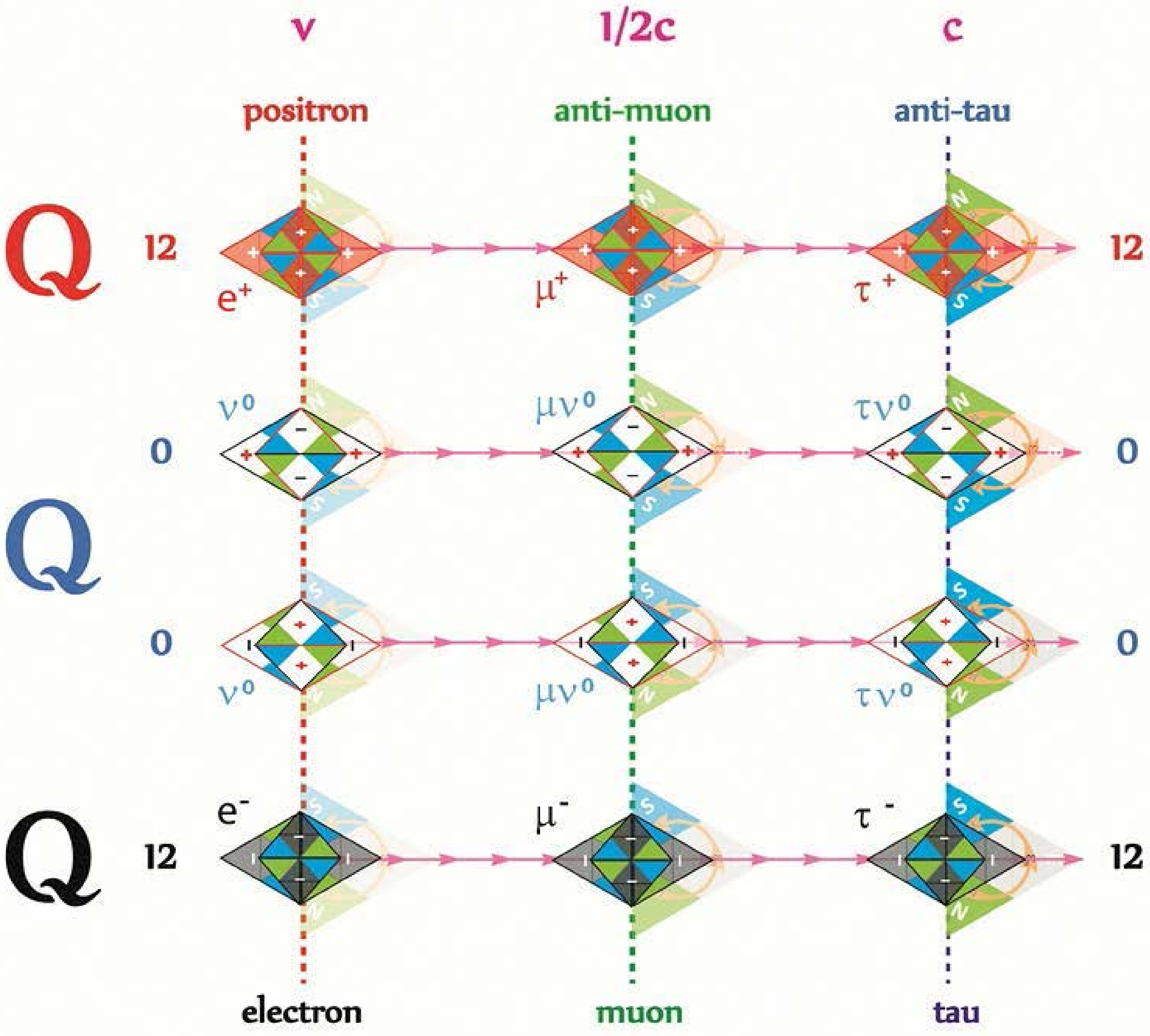
v
1/2c
c
positron
anti-muon
anti-tau
Q
12
e^+
μ^+
τ^+
12
ν^0
$\mu\nu^0$
$\tau\nu^0$
0
0
Q
0
ν^0
$\mu\nu^0$
$\tau\nu^0$
0
Q
12
e^-
μ^-
τ^-
12
N
S
+
–
electron
muon
tau

Antimatter

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Leptons ElectroMagnetic mass velocity

All Leptronic generations, oscillations and types can be accounted for through Tetryonic geometries & energy levels

Matter

Tetryonics 18.03 - Leptronic oscillations

Copyright ABRAHAM [2008] - all rights reserved

Principal Quantum levels

Eigenstate bound energy states

$$Mv^2 = KEM = hv^2$$

Bound electron Eigenstate values

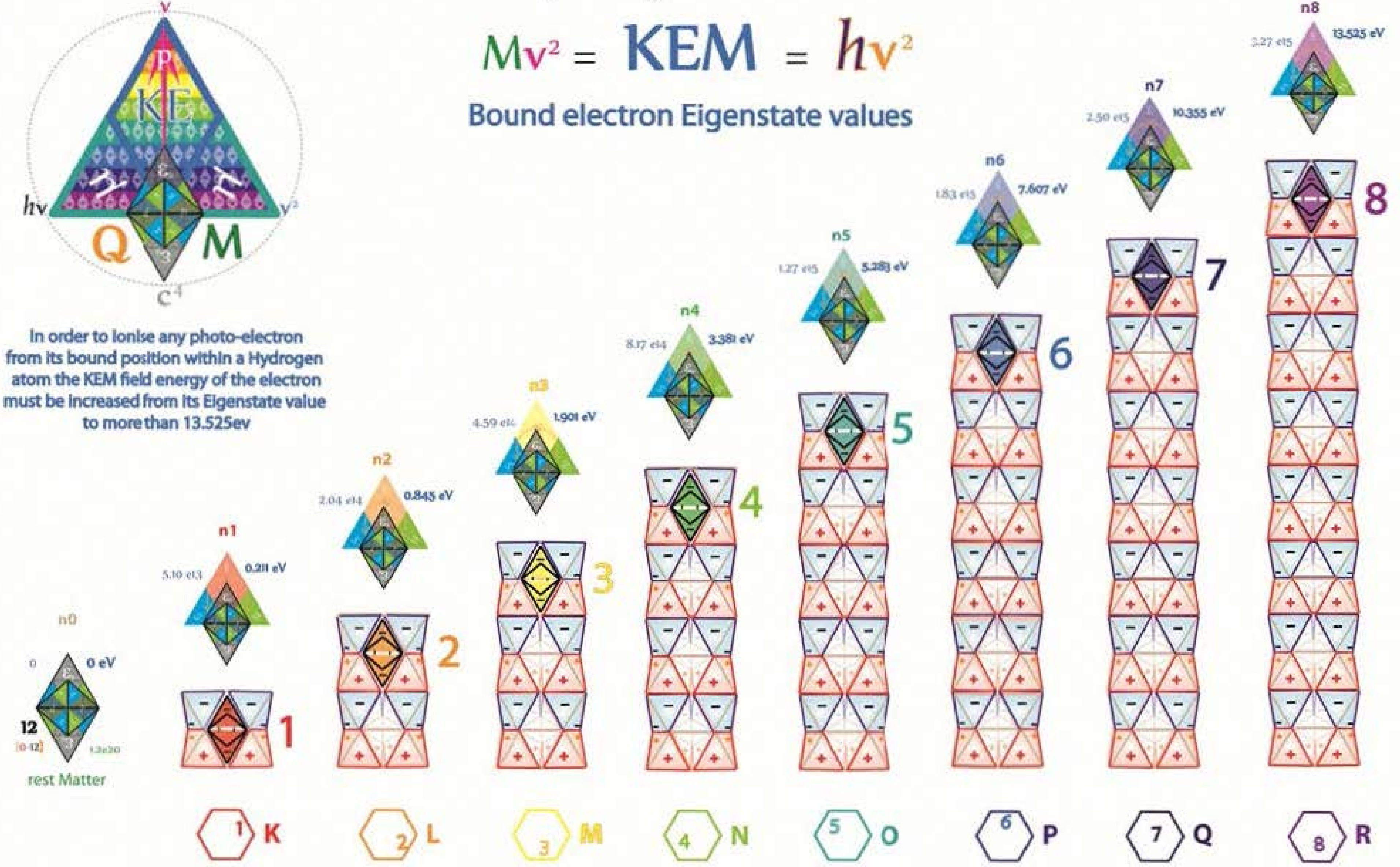
v
P
KE
hv
v²
Q
M
c⁴
n0
0
0 eV
12
[0-12]
1.2e20
rest Matter
n1
5.10 e13
0.211 eV
1
n2
2.04 e14
0.845 eV
2
n3
4.59 e14
1.901 eV
3
n4
8.17 e14
3.381 eV
4
n5
1.27 e15
5.283 eV
5
n6
1.83 e15
7.607 eV
6
n7
2.50 e15
10.355 eV
7
n8
3.27 e15
13.525 eV
8
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

In order to ionise any photo-electron from its bound position within a Hydrogen atom the KEM field energy of the electron must be increased from its Eigenstate value to more than 13.525ev

The mass-energy content of Baryons comprising atomic nuclei directly determines the Kinetic energy levels of bound Leptons [forming quantum synchronous convertors]
(Incident Bosons/Photons can increase these levels if they contain the exact quanta required to increase the level to the next square number)

Tetryonics 18.04 - Principal Quantum Levels

Bohr radius

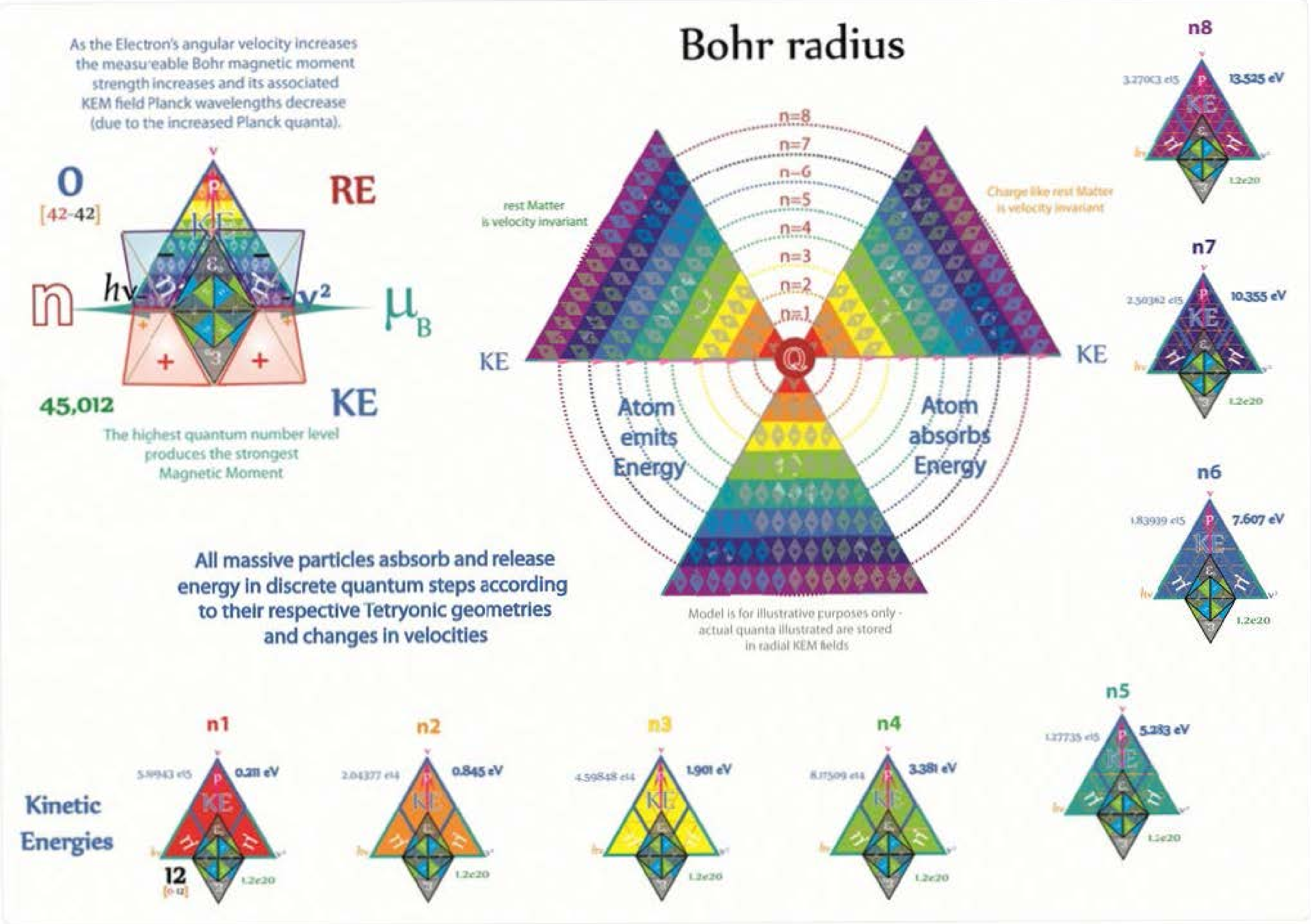

As the Electron's angular velocity increases the measu'eable Bohr magnetic moment strength increases and its associated KEM field Planck wavelengths decrease (due to the increased Planck quanta).

The highest quantum number level produces the strongest Magnetic Moment

All massive particles asbsorb and release energy in discrete quantum steps according to their respective Tetryonic geometries and changes in velocities

Model is for illustrative purposes only - actual quanta illustrated are stored in radial KEM fields

O
[42-42]
RE
n
hv
v²
μ_B
45,012
KE
rest Matter is velocity invariant
Charge like rest Matter is velocity invariant
n=8
n=7
n=6
n=5
n=4
n=3
n=2
n=1
KE
KE
Q
Atom emits Energy
Atom absorbs Energy
n8
3.27063 e15
13.525 eV
n7
2.50362 e15
10.355 eV
n6
1.83939 e15
7.607 eV
n5
1.27735 e15
5.283 eV
Kinetic Energies
n1
5.10943 e13
0.211 eV
12
[0-12]
n2
2.04377 e14
0.845 eV
n3
4.59848 e14
1.901 eV
n4
8.17509 e14
3.381 eV
1.2e20

Nuclear Magnetons

are weaker than Bohr magnetons
due to differing mass-charge ratios
and non-parallel Quark magnetons

Baryons

are tri-Dodecyon geometries

EM Field Planck quanta

$$36\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Baryons ElectroMagnetic mass velocity

Tetryonic Charge

Proton

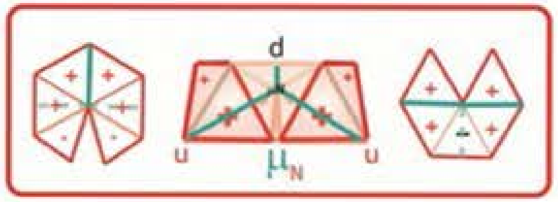

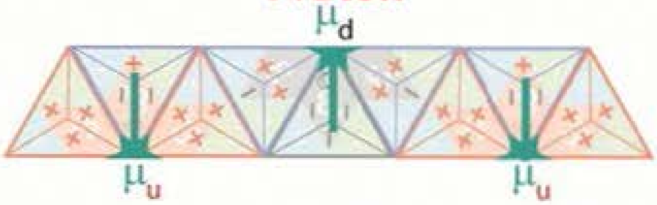

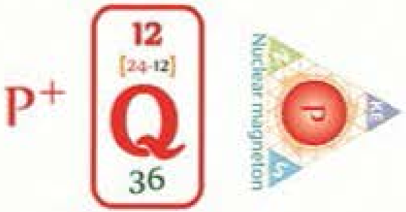
p^+
12
[24-12]
Q
36
Nuclear magneton
P

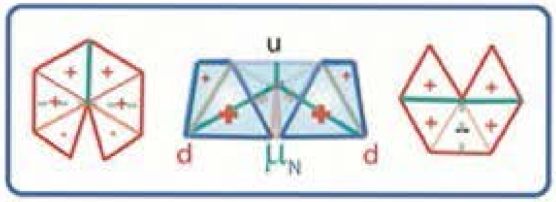

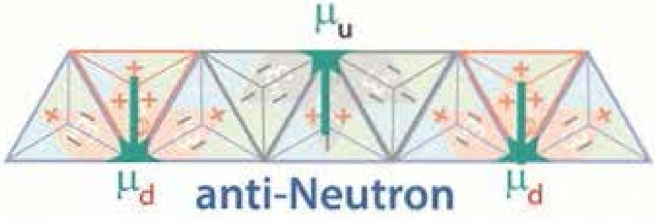

anti-Neutron

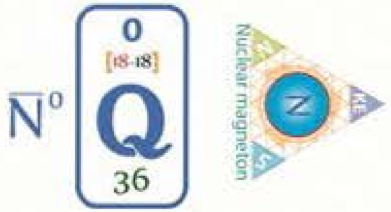
$\bar{N}^0$
0
[18-18]
Q
36
Nuclear magneton
N

Neutron

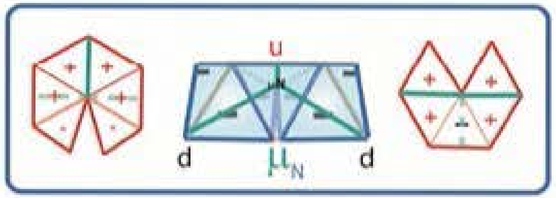

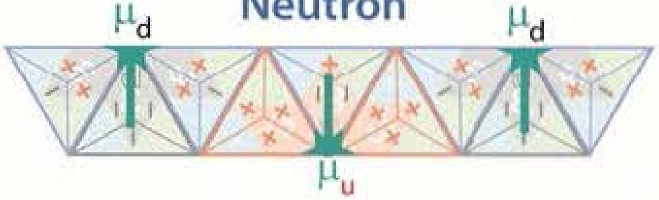

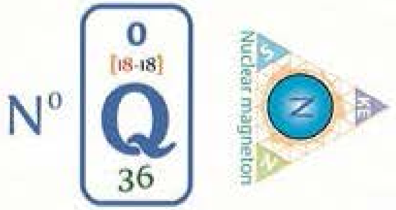
N^0
0
[18-18]
Q
36
Nuclear magneton
N

anti-Proton

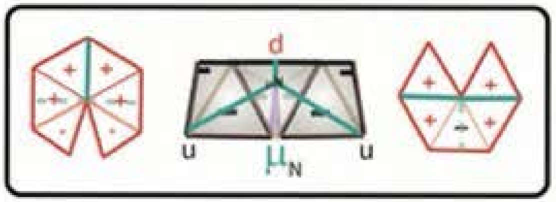

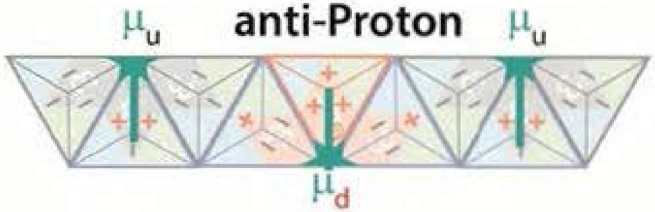

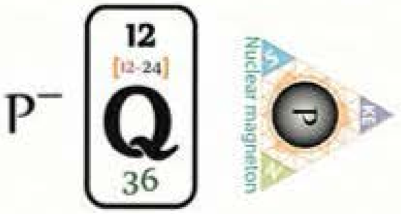
p^-
12
[12-24]
Q
36
Nuclear magneton
P

Matter geometry

Tetryonics 18.06 - Nuclear Magnetons

Copyright ABRAHAM [2008] - all rights reserved

Baryonic Magnetic Moments

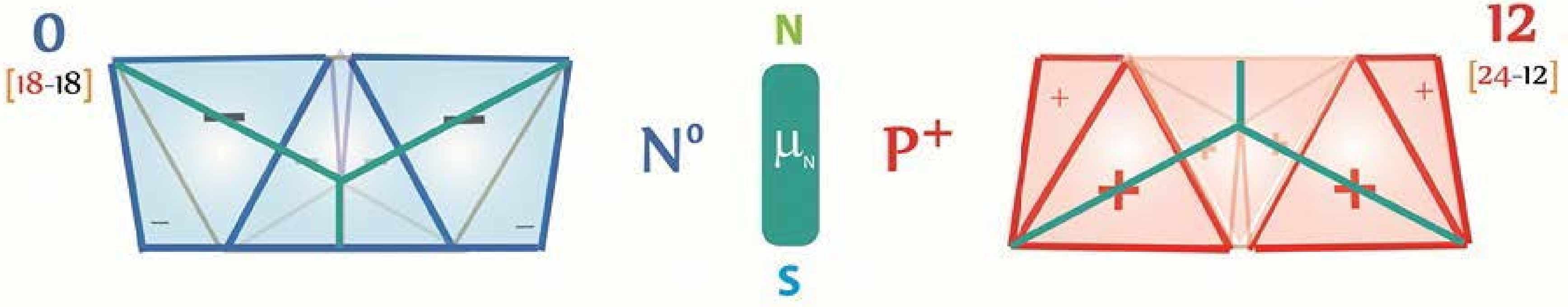
0
[18-18]
N^0
N
μ_N
S
P^+
12
[24-12]

A Baryon's Magnetic Moment
is a result of the combined non-parallel
Magnetic moments of its Quarks

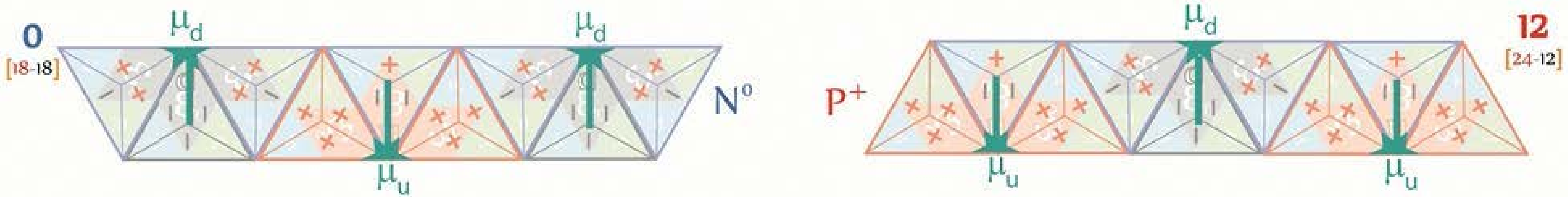
0
[18-18]
μ_d
μ_d
μ_u
N^0
P^+
μ_d
μ_u
μ_u
12
[24-12]

As a result of the 3 non-aligned Quark magnetic moments
resulting from their Tetryonic geometry,
the Nuclear Magneton is considerably weaker than the Bohr Magneton

Nuclear Quantum Energies

[Principal Quantum Numbers]

All electron energy levels are reflective of the KEM field of a electron in a specific quantum level (The rest Matter of each electron is invariant)

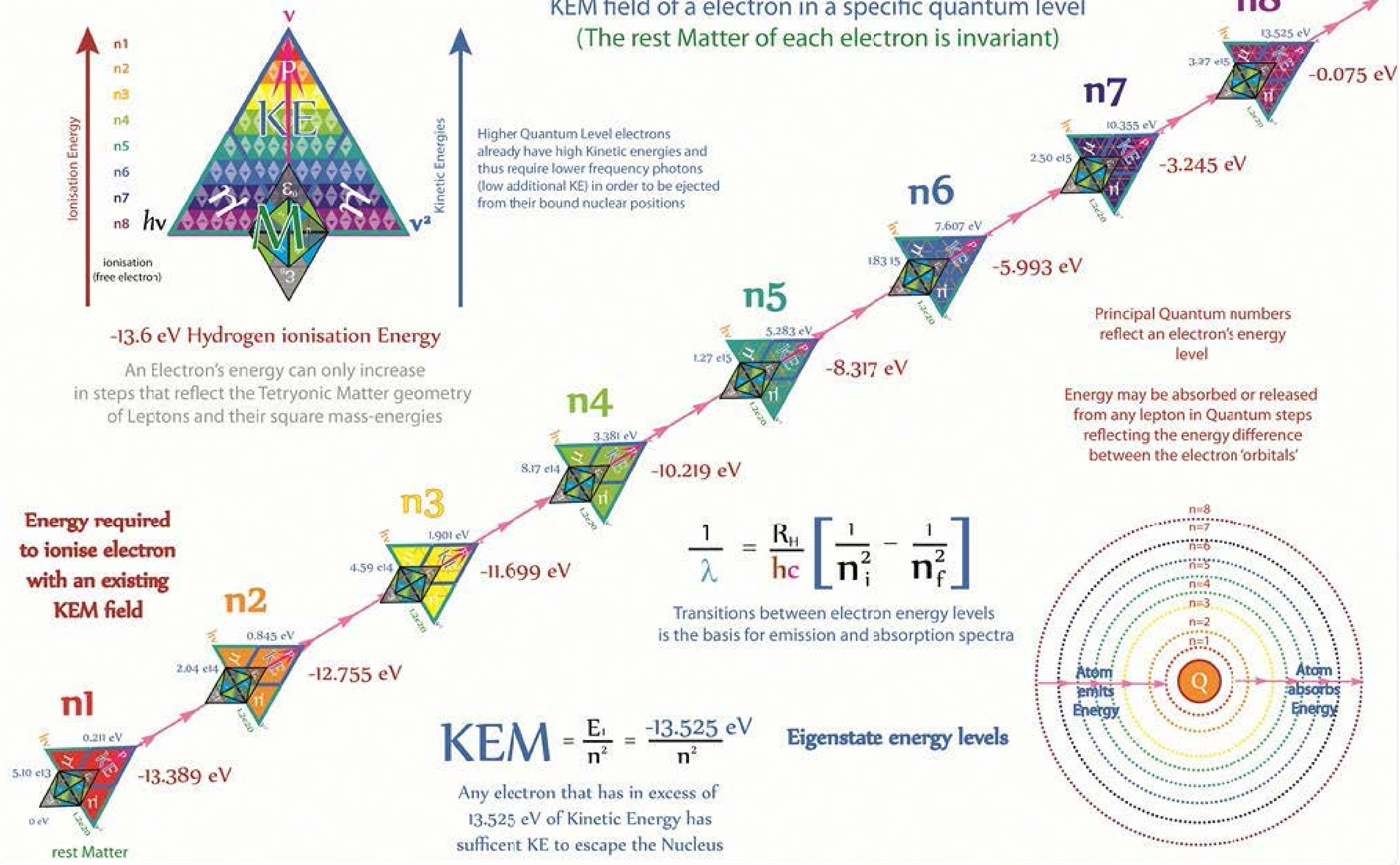
Ionisation Energy
n1
n2
n3
n4
n5
n6
n7
n8
hv
ionisation
(free electron)
v
P
KE
M
v²
Kinetic Energies
n1
0.211 eV
5.10 e13
0 eV
1.2e20
-13.389 eV
rest Matter
n2
0.845 eV
2.04 e14
-12.755 eV
n3
1.901 eV
4.59 e14
-11.699 eV
n4
3.381 eV
8.17 e14
-10.219 eV
n5
5.283 eV
1.27 e15
-8.317 eV
n6
7.607 eV
1.83 15
-5.993 eV
n7
10.355 eV
2.50 e15
-3.245 eV
n8
13.525 eV
3.27 e15
-0.075 eV
n=8
n=7
n=6
n=5
n=4
n=3
n=2
n=1
Q
Atom emits Energy
Atom absorbs Energy

Higher Quantum Level electrons already have high Kinetic energies and thus require lower frequency photons (low additional KE) in order to be ejected from their bound nuclear positions

-13.6 eV Hydrogen ionisation Energy

An Electron's energy can only increase in steps that reflect the Tetryonic Matter geometry of Leptons and their square mass-energies

Principal Quantum numbers reflect an electron's energy level

Energy may be absorbed or released from any lepton in Quantum steps reflecting the energy difference between the electron 'orbitals'

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left[\frac{1}{n_i^2} - \frac{1}{n_f^2}\right]$$

Transitions between electron energy levels is the basis for emission and absorption spectra

Energy required to ionise electron with an existing KEM field

$$KEM = \frac{E_i}{n^2} = \frac{-13.525\text{ eV}}{n^2}$$

Any electron that has in excess of 13.525 eV of Kinetic Energy has sufficent KE to escape the Nucleus

Eigenstate energy levels

Leptons are historically classified as Spin 1/2 particles
(by the spin-statistics theorem and the Pauli exclusion principle)
as determined by their magnetic moments

Quantum Spin Numbers

(rotations about an axis)

Spin 1/2

720°

Rotating a spin-1/2 particle by 360 degrees does not bring it back to the same quantum state it needs a 720 degree rotation

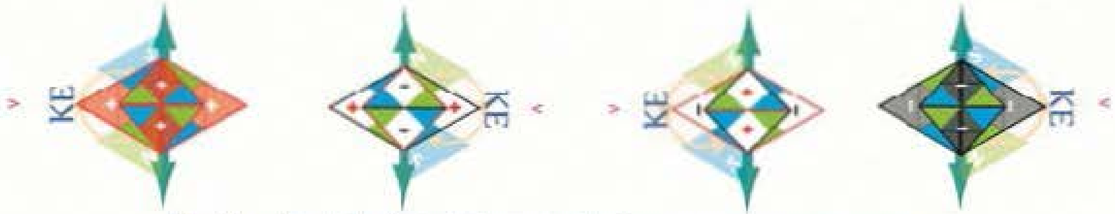

Planck bar relates to the electric field content of KEM fields resulting from Matter in motion
(as reference to the Nuclear magneton or external Magnetic field)

Spin 0

any°

A spin-zero particle can only have a single quantum state, even after torque is applied.

Spherical 'point particles' of charge do not exist

Spin 1

360°

Rotating a spin-1 particle 360 degrees can bring it back to the same quantum state

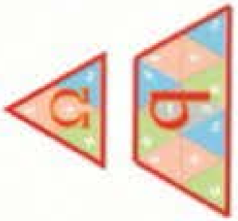

SPIN
must not to be confused with
Chirality (reflections)

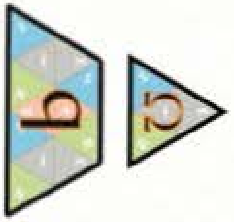

Spin 2

180°

Rotating a spin-2 particle 180 degrees can bring it back to the same quantum state

Photons are their own anti-particle

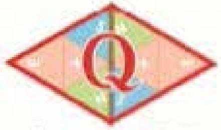

electro-static field

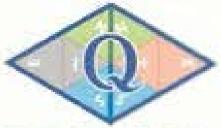

magneto-static field

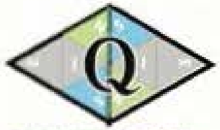

electro-static field

Spin 3

120°

Rotating a spin-3 particle 120 degrees can bring it back to the same quantum state

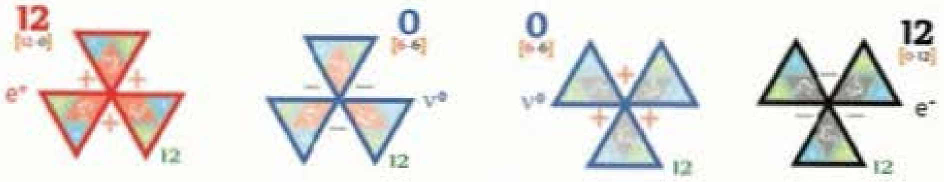

On a geometric basis all Leptons are in fact spin 3 particles

Tetryonics 19.01 - Quantum spin numbers

Magnetic moments are determined by the KEM fields created by vector linear momentum

Nuclear Magnetic moments are complicated by the tri-quark magnetons within all Matter topologies

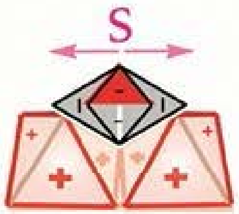

The KEM field of any charged particle in motion is reflective of the particle's nett charge topology

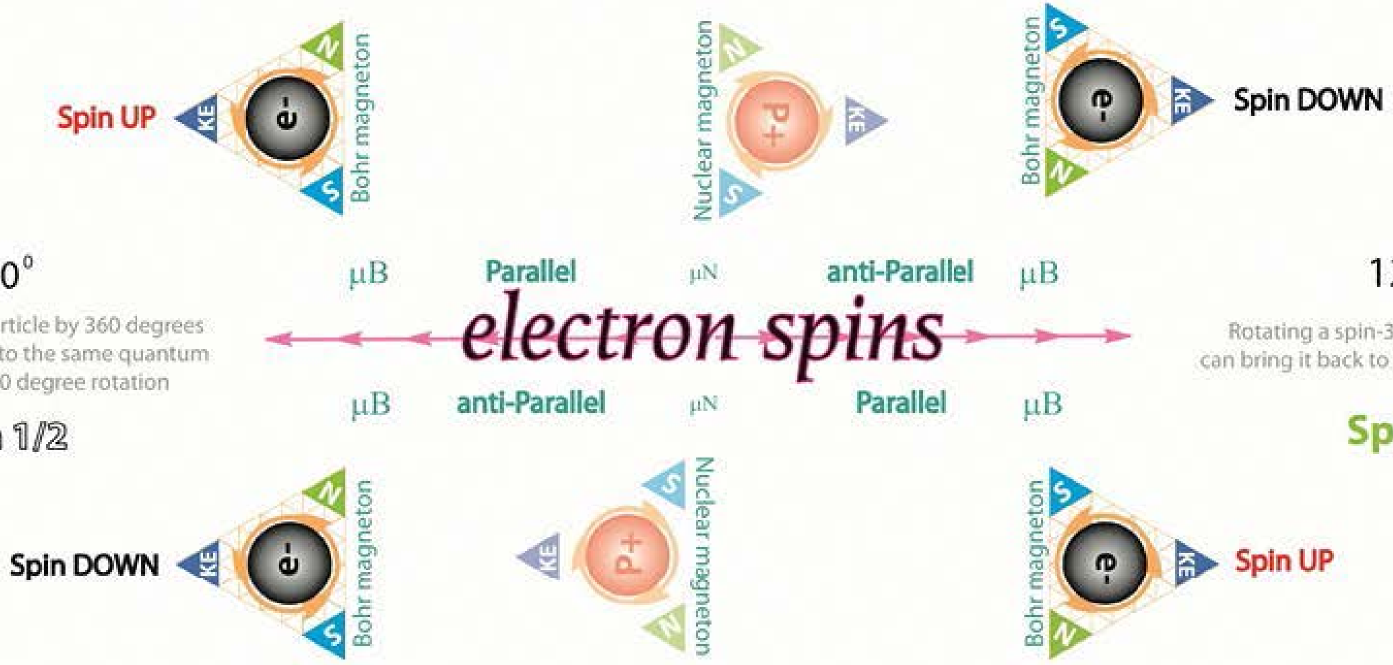
Spin UP
KE
e-
N
S
Bohr magneton
Nuclear magneton
P+
Spin DOWN
μB
Parallel
μN
anti-Parallel
μB
electron spins
μB
anti-Parallel
μN
Parallel
μB
Spin DOWN
Spin UP

720°

Rotating a spin-1/2 particle by 360 degrees does not bring it back to the same quantum state it needs a 720 degree rotation

Spin 1/2

120°

Rotating a spin-3 particle 120 degrees can bring it back to the same quantum state

Spin 3

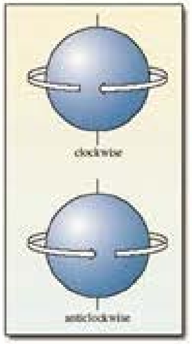
clockwise
anticlockwise

All atomic particles have a particular "spin" analogous to the Earth's rotation on its axis.

$$\mu_B = \left[\frac{Q}{2mv}\right]/\ell$$

An isolated electron has an angular momentum and a magnetic moment resulting from its spin. While an electron's spin is sometimes visualized as a literal rotation about an axis, it is in fact a fundamentally different, quantum-mechanical phenomenon with no true analogue in classical physics.

The quantum mechanical reality underlying spin is complex and still poorly understood Consequently, there is no reason to expect the above classical relation to hold.

$$\mu_B = g\left[\frac{e-}{2mv}\right]S$$

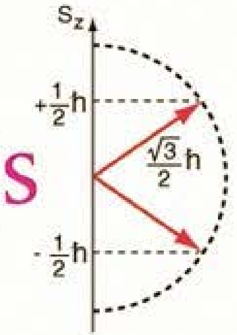
S_z
$+\frac{1}{2}\hbar$
$\frac{\sqrt{3}}{2}\hbar$
$-\frac{1}{2}\hbar$

Tetryonic KEM field geometries reveals the source and orientation of all atomic magnetic moments

Generating Magnetons

A static Electron has a negative Tetryonic charge [0-12] topology with neutralised magnetic dipoles

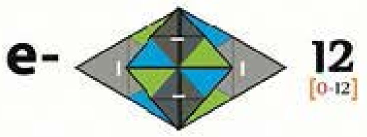

Leptons are 12 loop inductive charge rotors

Tetryonic geometry fully explains Leptronic 'spins'

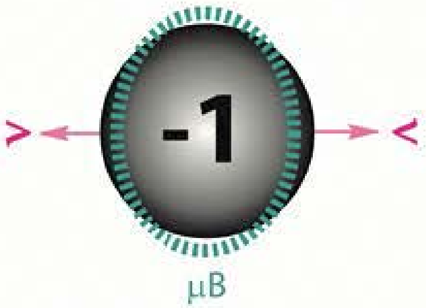

Einstein's Special Relativity model of distorted moving charges producing magnetic moments is incorrect

BOHR Magneton produced by Lorentzian distortion of charges due to relativistic velocities

$$\mu_B = \frac{e\hbar}{2m_e}$$

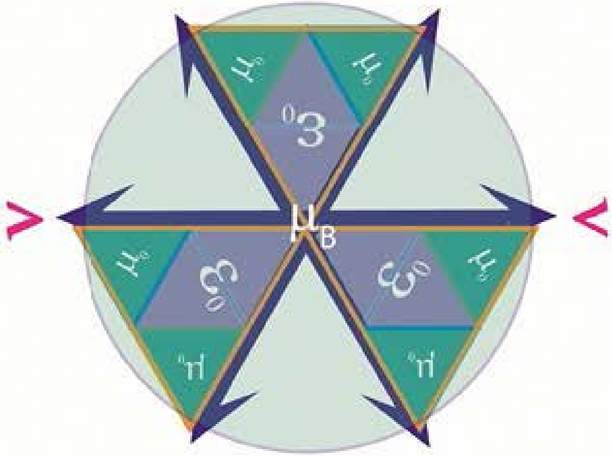

The term "electron spin" can now be taken literally (when modelled with Tetryonic geometries) as an accurate description of the origin of Magnetic moments for all Leptonic [BOHR] magnetons.

The previously held model of the electron as a spinning sphere of charge must be abandoned in favour of the true Tetryonic charge geometries of EM mass-Energy-Matter

Ampere Force

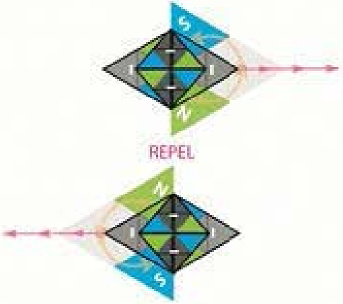

$$B = \frac{\mu_0 I}{4\pi} \int \frac{d\ell \times \hat{r}}{r^2},$$

$$F = 2k_A \frac{I_1 I_2}{r}$$

$$F = q[E + (v \times B)].$$

Lorentz Force

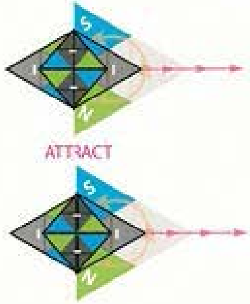

A moving Electron has a KEM field with an Electric field and a Magnetic Moment

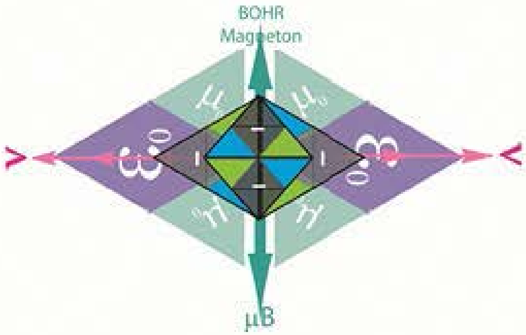

The gyromagnetic ratio of a particle or system is the ratio of its magnetic dipole moment to its angular momentum

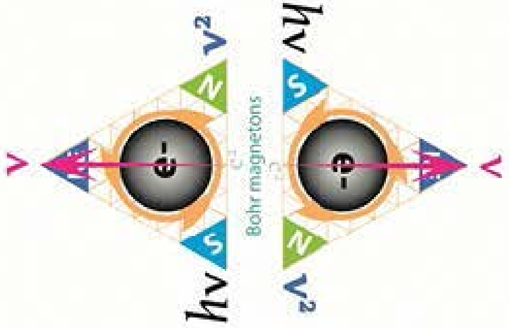

The KEM field energy of an Electron in motion is subject to relativistic corrections due to energy changes resulting from its acceleration

$$\lambda = L'\sqrt{1 - \frac{v^2}{c^2}}$$

WAVE-length contraction of mass-energy quanta of KEM field

Stern Gerlach Experiment

In 1922 O. Stern and W. Gerlach measured the intrinsic spin angular momentum
of silver atoms and found it to take only two discrete values.
+ h / 2 and - h / 2 commonly called "spin up" and "spin down

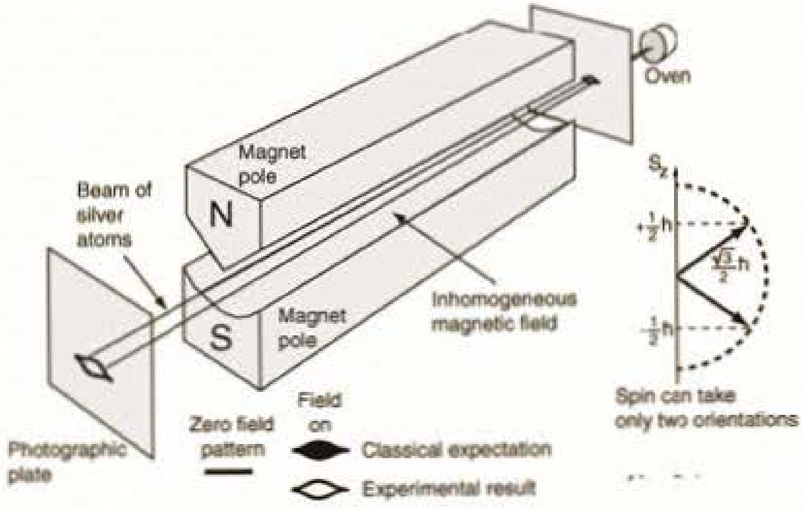
Oven
Magnet pole
N
Beam of silver atoms
S
Magnet pole
Inhomogeneous magnetic field
S_z
$+\frac{1}{2}h$
$\frac{\sqrt{3}}{2}h$
$-\frac{1}{2}h$
Spin can take only two orientations
Photographic plate
Zero field pattern
Field on
Classical expectation
Experimental result

The Stern-Gerlach experiment to determine electron SPIN

The results were interpreted to show that particles possess an intrinsic angular momentum that is most closely analogous to the angular momentum of a classically spinning charged sphere, that takes on only certain quantized values of angular momentum

Tetryonics shows the results are a product of the KEM field produced by charge particles in motion, with all charged particles being able to produce 2 distinct magneton orientations as a result of the real motion of their intrinsic quantum inductive loop [Matter]geometries through external E fields & as referenced to external M moments.

This along with the Lorentz force produced by an external Magnetic field acting on the KEM magnetons produces the two results obtained

Any increase in electron energies results in an increase in quantised angular momenta and its vector linear momentum

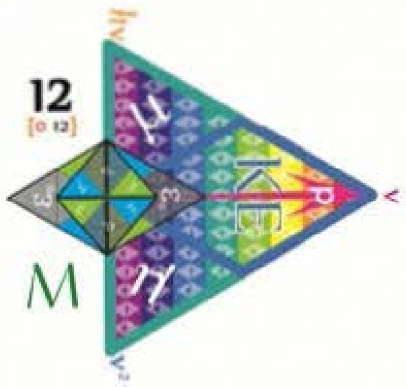
hv
12
[0 12]
3
3
μ
KE
p
v
ε
M
v²

Negative charged in motion produce negative charge KEM field geometries

Any change in the KEM energy level of an electron results in Photon emission or absorption

Electron Spins based on the measured Bohr Magnetons of moving electrons are reflected with KEM field geometries

Electrons bound in atomic nuclei increase of decrease their energy levels dependent on the energy level of the nuclei in which they are bound

Bound and unbound electrons can rotate in one of two directions influenced only by external magnetic and electric fields or incident Photons

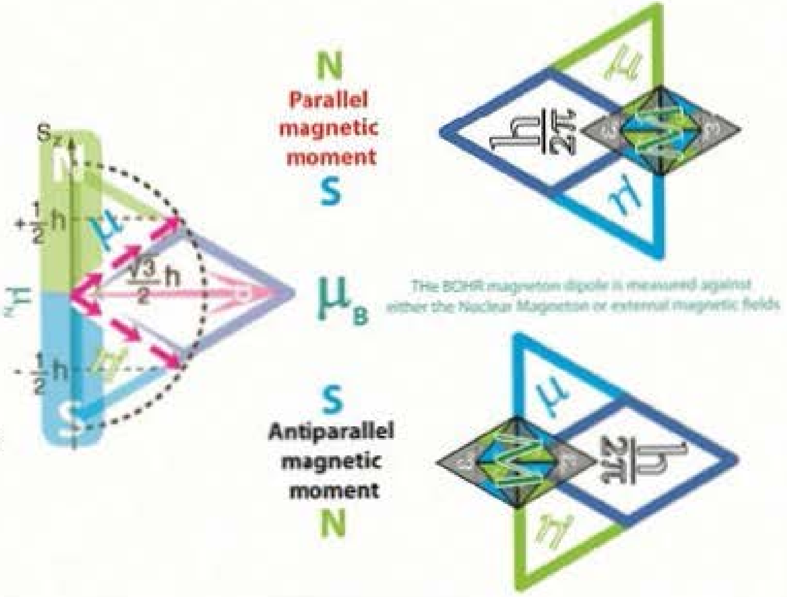
S_z
N
$+\frac{1}{2}h$
μ
$\frac{\sqrt{3}}{2}h$
μ
ε
$-\frac{1}{2}h$
S
N
Parallel magnetic moment
S
μ_B
S
Antiparallel magnetic moment
N
μ
$\frac{h}{2\pi}$
ε
M

The BOHR magneton dipole is measured against either the Nuclear Magneton or external magnetic fields

The two separated beams of electrons produced are defined as having differing SPINS [UP or DOWN]

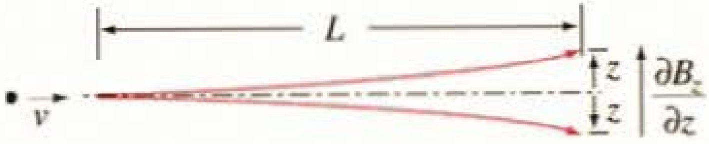
L
v
z
z
$\frac{\partial B_z}{\partial z}$

$$z = \frac{1}{2}at^2 = \frac{1}{2}\frac{F}{m}\left[\frac{L}{v}\right]^2 = \pm\frac{\mu_B L^2}{4KE}\frac{\partial B_z}{\partial z}$$

The orientation of the BOHR magneton wrt the NUCLEAR magneton determines the 'direction' of electron SPIN

Tetryonics 19.04 - Stern Gerlach experiment

Bohr Magnetons

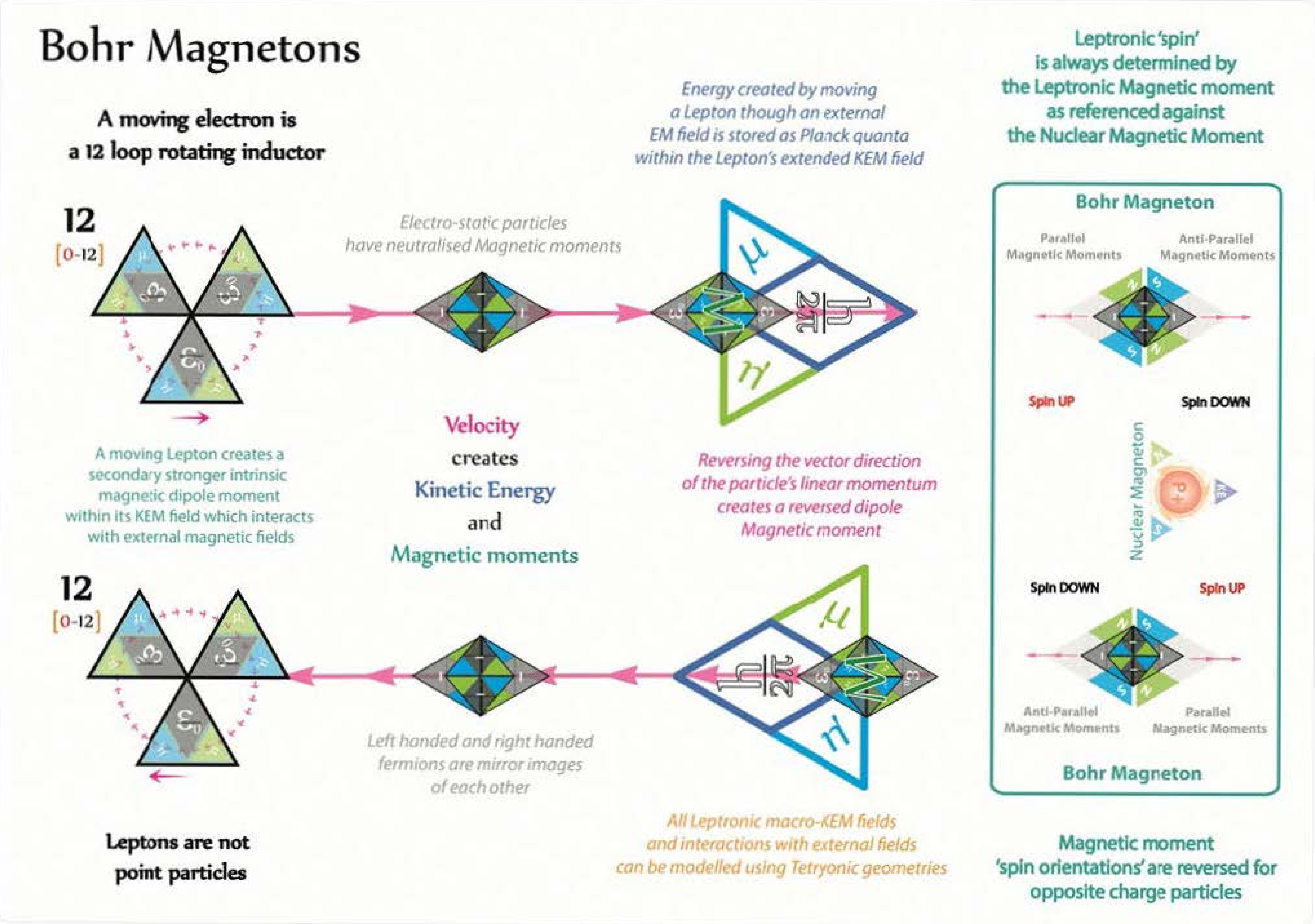
12
[0-12]
μ
h/2π
ɳ
Bohr Magneton
Parallel Magnetic Moments
Anti-Parallel Magnetic Moments
Spin UP
Spin DOWN
Nuclear Magneton
Spin DOWN
Spin UP
Anti-Parallel Magnetic Moments
Parallel Magnetic Moments
Bohr Magneton
12
[0-12]
μ
h/2π
ɳ

A moving electron is
a 12 loop rotating inductor

Energy created by moving a Lepton though an external EM field is stored as Planck quanta within the Lepton's extended KEM field

Leptronic 'spin' is always determined by the Leptronic Magnetic moment as referenced against the Nuclear Magnetic Moment

Electro-static particles have neutralised Magnetic moments

A moving Lepton creates a secondary stronger intrinsic magnetic dipole moment within its KEM field which interacts with external magnetic fields

Velocity
creates
Kinetic Energy
and
Magnetic moments

Reversing the vector direction of the particle's linear momentum creates a reversed dipole Magnetic moment

Left handed and right handed fermions are mirror images of each other

Leptons are not point particles

All Leptronic macro-KEM fields and interactions with external fields can be modelled using Tetryonic geometries

Magnetic moment 'spin orientations' are reversed for opposite charge particles

Tetryonics 19.05 - Bohr Magnetons

Copyright ABRAHAM [2008] - all rights reserved

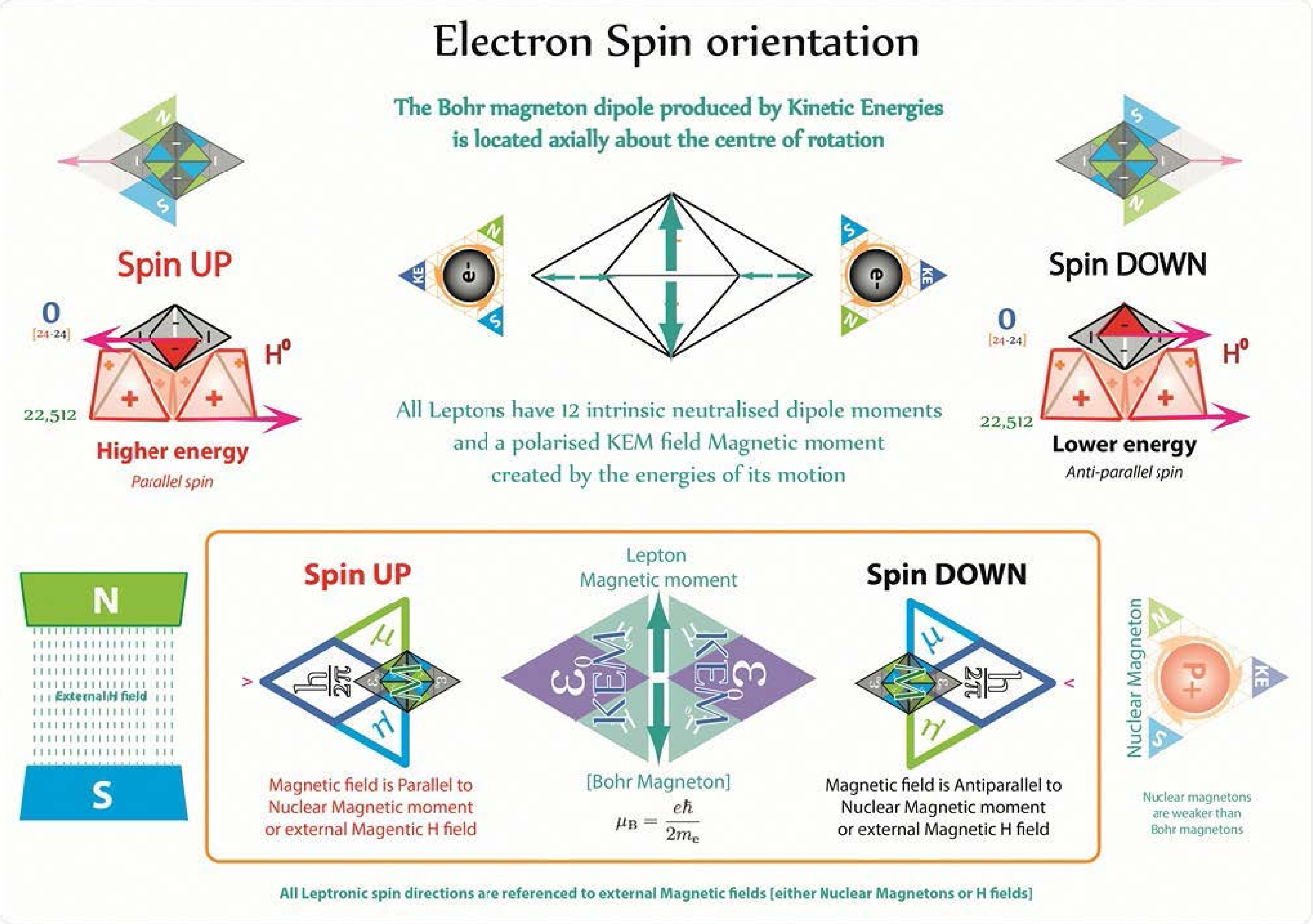

Electron Spin orientation

The Bohr magneton dipole produced by Kinetic Energies
is located axially about the centre of rotation

Spin UP

Higher energy

Parallel spin

Spin DOWN

Lower energy

Anti-parallel spin

All Leptons have 12 intrinsic neutralised dipole moments
and a polarised KEM field Magnetic moment
created by the energies of its motion

Spin UP

Magnetic field is Parallel to
Nuclear Magnetic moment
or external Magentic H field

Lepton
Magnetic moment

[Bohr Magneton]

$$\mu_B = \frac{e\hbar}{2m_e}$$

Spin DOWN

Magnetic field is Antiparallel to
Nuclear Magnetic moment
or external Magnetic H field

Nuclear magnetons
are weaker than
Bohr magnetons

All Leptronic spin directions are referenced to external Magnetic fields [either Nuclear Magnetons or H fields]

Tetryonics 19.06 - Electron Spin orientation

Copyright ABRAHAM [2008] - all rights reserved

Nuclear Spins

are determined by the orientation of Bohr Magneton
with respect to the Nuclear Magneton
[or an external Magnetic field]

*electron Matter geoemtries
are negative charged fascia
which create neutral intrinsic
magnetic dipole configurations*

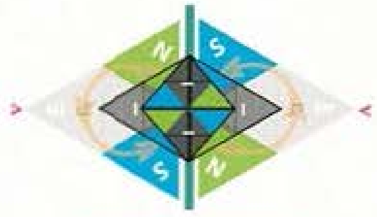

MAGNETIC MOMENT

*In a moving electron
Kinetic energy produces
an axial Magnetic moment*

*In a static electron
all intrinsic dipoles are
neutralised through
their orientations*

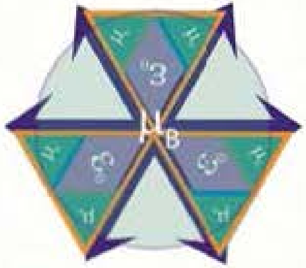

***Quantum Inductive Loops
In motion produce Kinetic Energies
In turn creating Magnetic Moments***

Electron

PARALLEL
Magnetic moments

ANTIPARALLEL
Magnetic moments

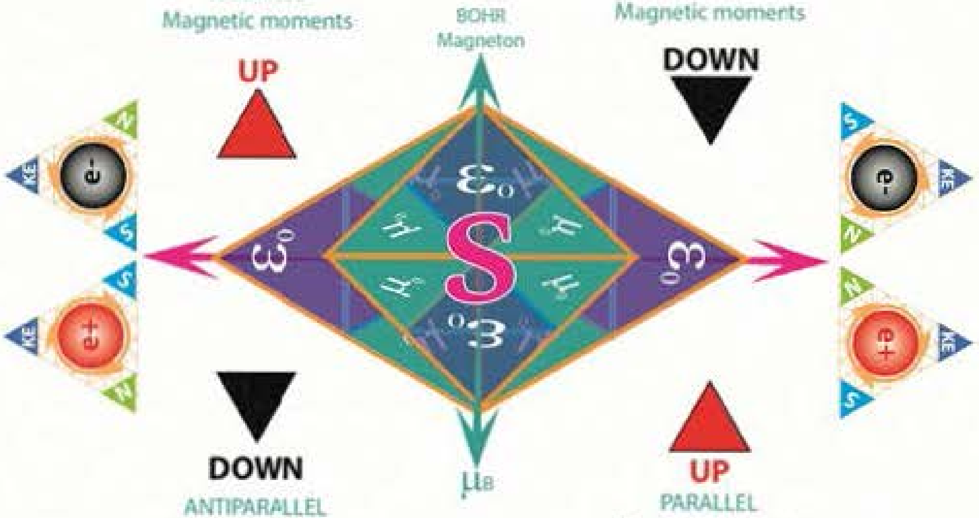

ANTIPARALLEL
Magnetic moments

PARALLEL
Magnetic moments

Positron

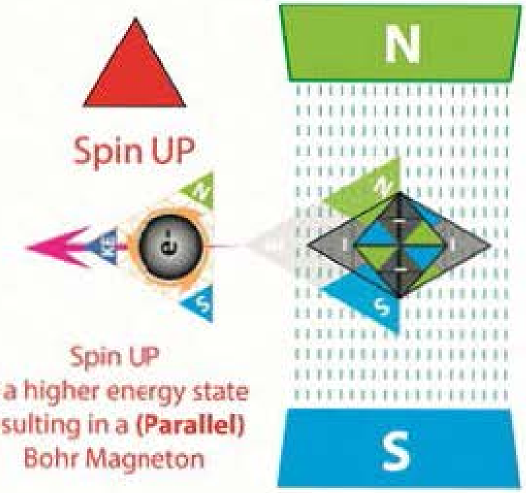

Spin UP

Spin UP
is a higher energy state
resulting in a **(Parallel)**
Bohr Magneton

Nuclear magneton or external Magnetic field

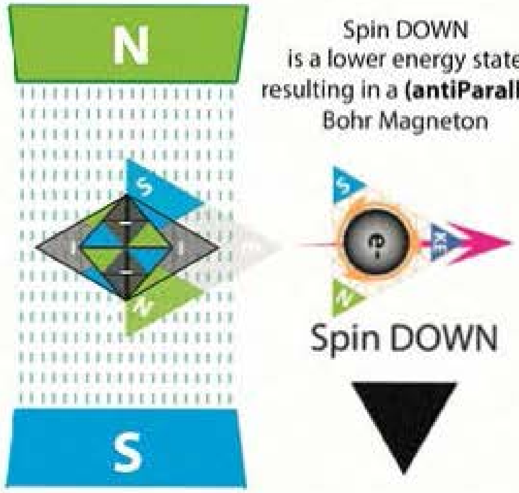

Spin DOWN
is a lower energy state
resulting in a **(antiParallel)**
Bohr Magneton

Spin DOWN

Lorentz Force

**When moving in an
external magnetic field**

The axial dipole
moment of an
electron will experience
torque force
proportional to
its velocity

**Opposites attract
Similars repel**

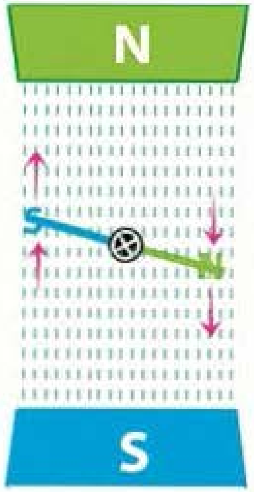

Nuclear magnetic Moment

[Nuclear magneton]

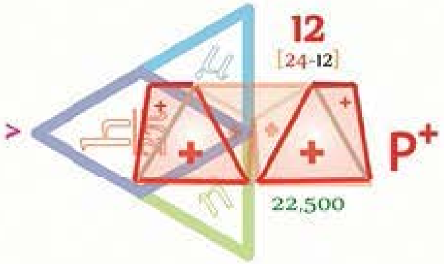
12
[24-12]
P⁺
22,500

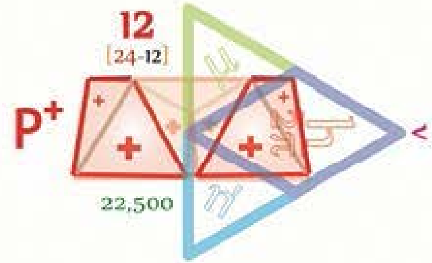
12
[24-12]
P⁺
22,500

NUCLEAR magnetons are much weaker than BOHR magnetons due to the higher charge to mass ratio of electrons

$$\frac{M_p}{M_e} = \frac{1}{1875}$$

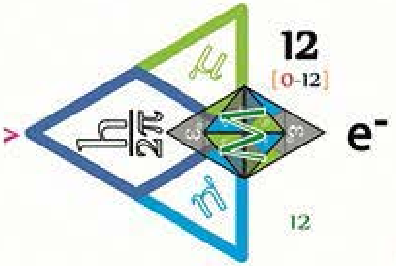

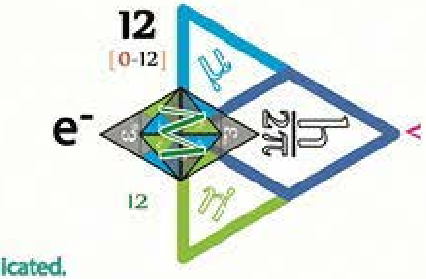

The nuclear spins for individual protons and neutrons parallels the treatment of electron spin, with spin 1/2 and an associated magnetic moment.

For the combination of electron, neutrons and protons in periodic elements, the situation is even more complicated.

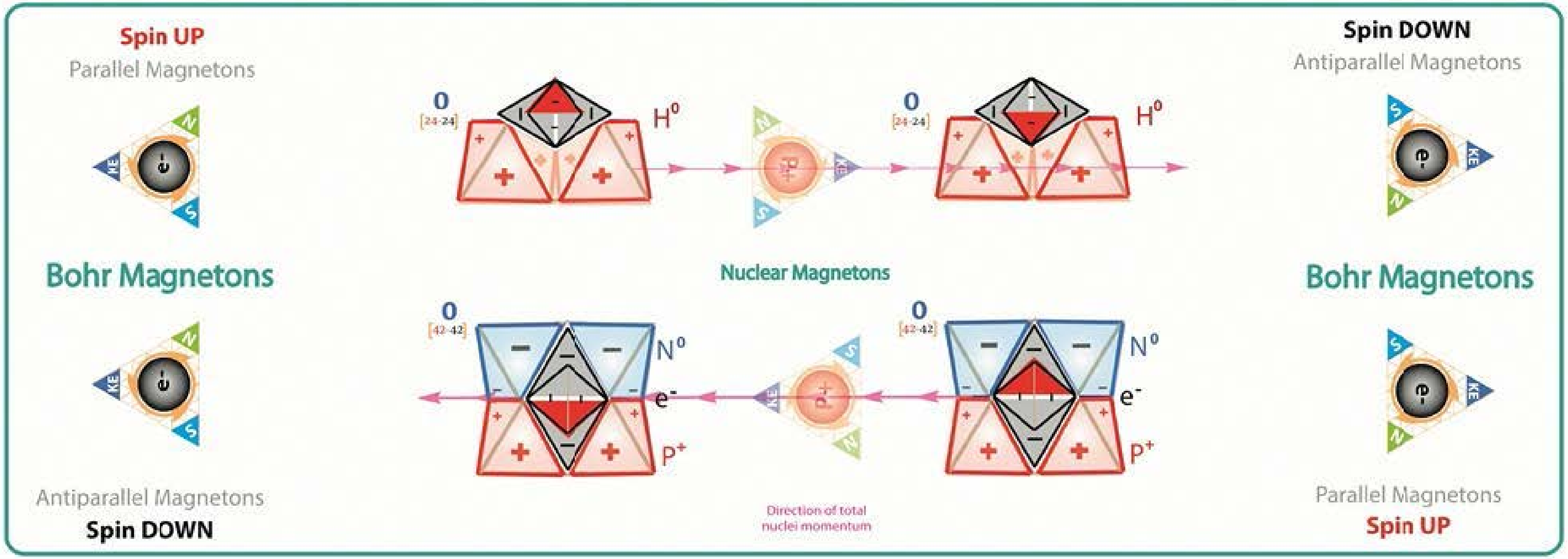
0
[24-24]
H⁰
0
[24-24]
H⁰
0
[42-42]
N⁰
e⁻
P⁺
0
[42-42]
N⁰
e⁻
P⁺
N
S
KE

Spin UP
Parallel Magnetons

Spin DOWN
Antiparallel Magnetons

Bohr Magnetons

Nuclear Magnetons

Bohr Magnetons

Antiparallel Magnetons
Spin DOWN

Direction of total nuclei momentum

Parallel Magnetons
Spin UP

Tetryonics 19.08 - Nuclear Magnetic moments

Spin orbital coupling mechanics

The induced magnetic moment of electrons in atomic nuclei combine vectorally with the magnetic moment of the nuclei

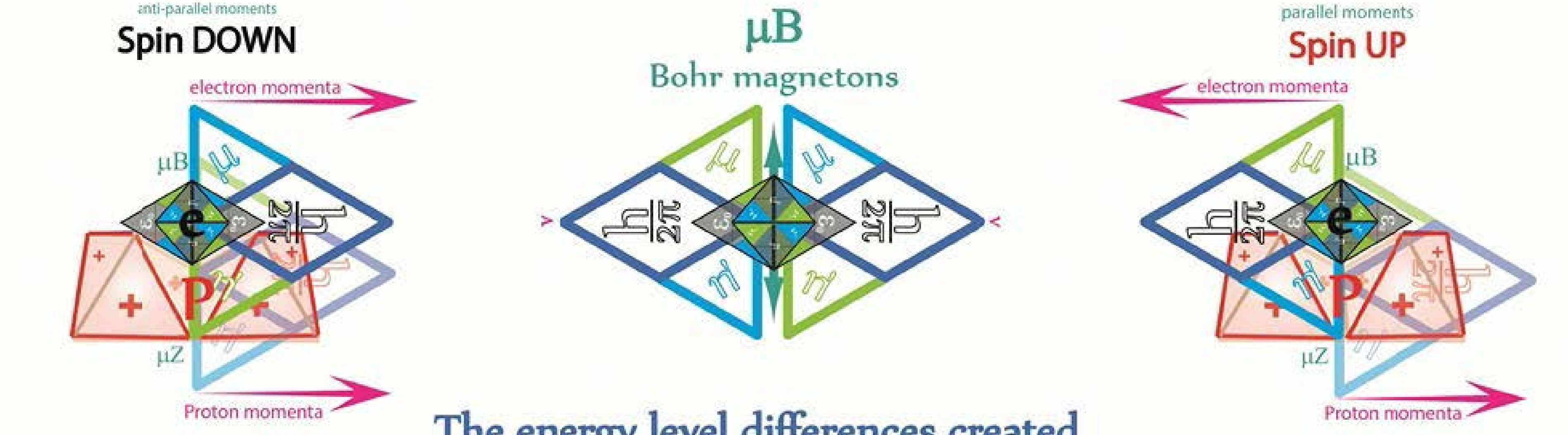
anti-parallel moments
Spin DOWN
electron momenta
μB
μ
e
$\frac{h}{2\pi}$
P
+
μZ
Proton momenta
μB
Bohr magnetons
$\frac{h}{2\pi}$
μ
parallel moments
Spin UP
electron momenta
μB
e
P
+
μZ
Proton momenta

The energy level differences created are manifested in Hyperfine-line splitting Zeeman effects etc.

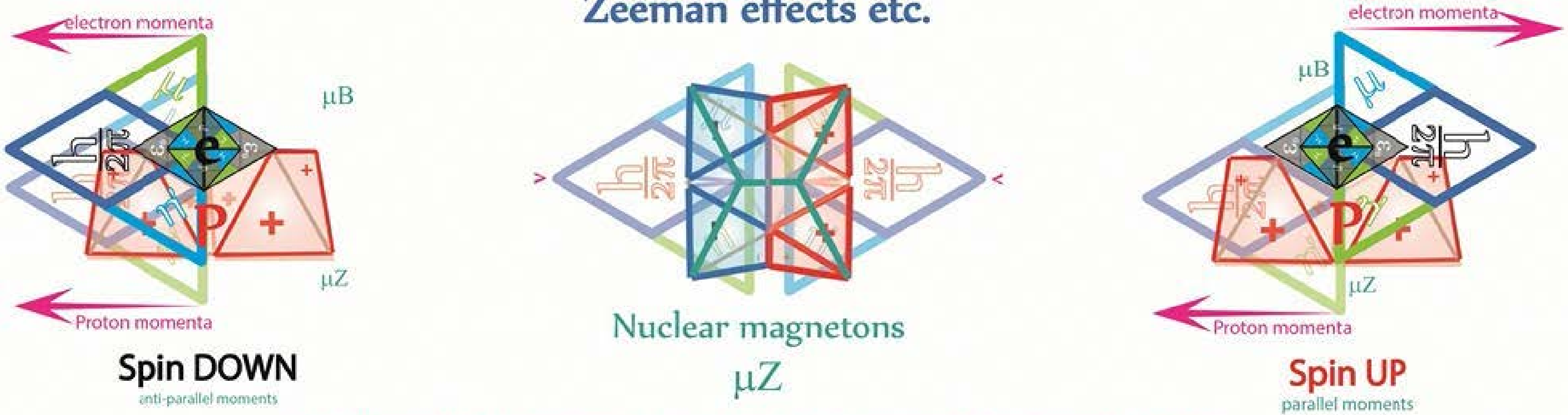
electron momenta
μB
e
$\frac{h}{2\pi}$
P
+
μZ
Proton momenta
Spin DOWN
anti-parallel moments
$\frac{h}{2\pi}$
Nuclear magnetons
μZ
electron momenta
μB
e
$\frac{h}{2\pi}$
P
+
μZ
Proton momenta
Spin UP
parallel moments

The alignment of electron spins in nuclei results in diamagnetic, para-magnetic Matter

Copyright ABRAHAM [2008] - all rights reserved

Gyromagnetic Ratio

The electron is a 12 charge quantum rotor
with a uniform charge to mass density ratio,
the ratio of its magnetic moment to
its orbital angular momentum,
also known as gyromagnetic ratio

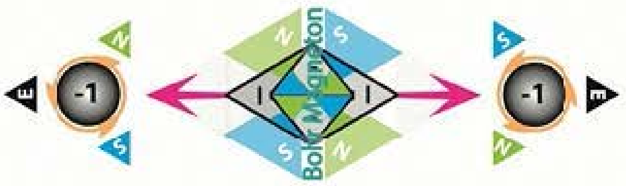

This implies that a more massive
assembly of charges spinning with the
same angular momentum will have a
proportionally weaker magnetic moment,
compared to its lighter counterpart.

12
[0-12]
1.2 e20
1
electron
12π

The Bohr Magneton is determined by the
charged KEM field geometry of Leptons

1875
charged mass-Matter differential

The combined Kinetic energy of Motion [KEM field]
and Electron Spin coupling with Nuclear Magnetons
will effect any measured Gyromagnetic ratios

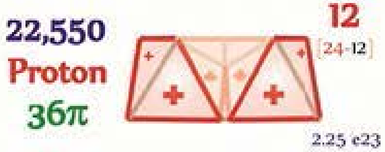

-1

9.1 e-31 kg

Classical electron model
[rotating sphere]

4π

**An electron is
NOT
a spherical particle**

12π

12
[0-12]

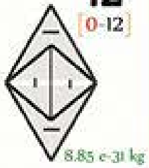

8.85 e-31 kg

Tetryonic electron
[rotating tri-tetryon topology]

In physics, the gyromagnetic ratio
(also sometimes known as the magnetogyric ratio in other disciplines)
of a particle or system is the ratio of its magnetic dipole moment to
its angular momentum, and it is often denoted by the symbol γ, gamma.

An isolated electron has an angular momentum and
a magnetic moment resulting from its spin.

Its SI units are radian per second per tesla (s−1·T -1) or,
equivalently, coulomb per kilogram (C·kg−1).

Tetryonic quantum mass to Charge ratio

$1.810109642 \times 10^{11}$

The 2006 CODATA
– e / me = -1.758 820 150(44) x 10[11]

Nuclear Magnetons

e-
em

Tetryonic elementary charge
1.602216081 e-19 coulombs

Tetryonic electron mass
8.851486361 e-31

12q
1.2 e20

[q/m]

181,010,964,200 C/kg

Spin UP
K.E.

Lower energy

Electron Spin
[Orbital Angular Momentum]

N
Nuclear Magneton
S

Spin DOWN
K.E.

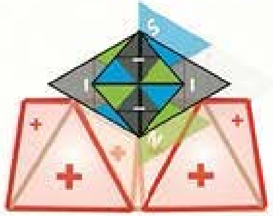

Higher energy

12q
2.25 e23

Tetryonic elementary charge
1.602216081 e-19 coulombs

Tetryonic Proton mass
1.659653693 e-27

e+
Pm

[q/m]

96,539,180.9 C/kg

Tetryonics 19.10 - Gyromagnetic ratio

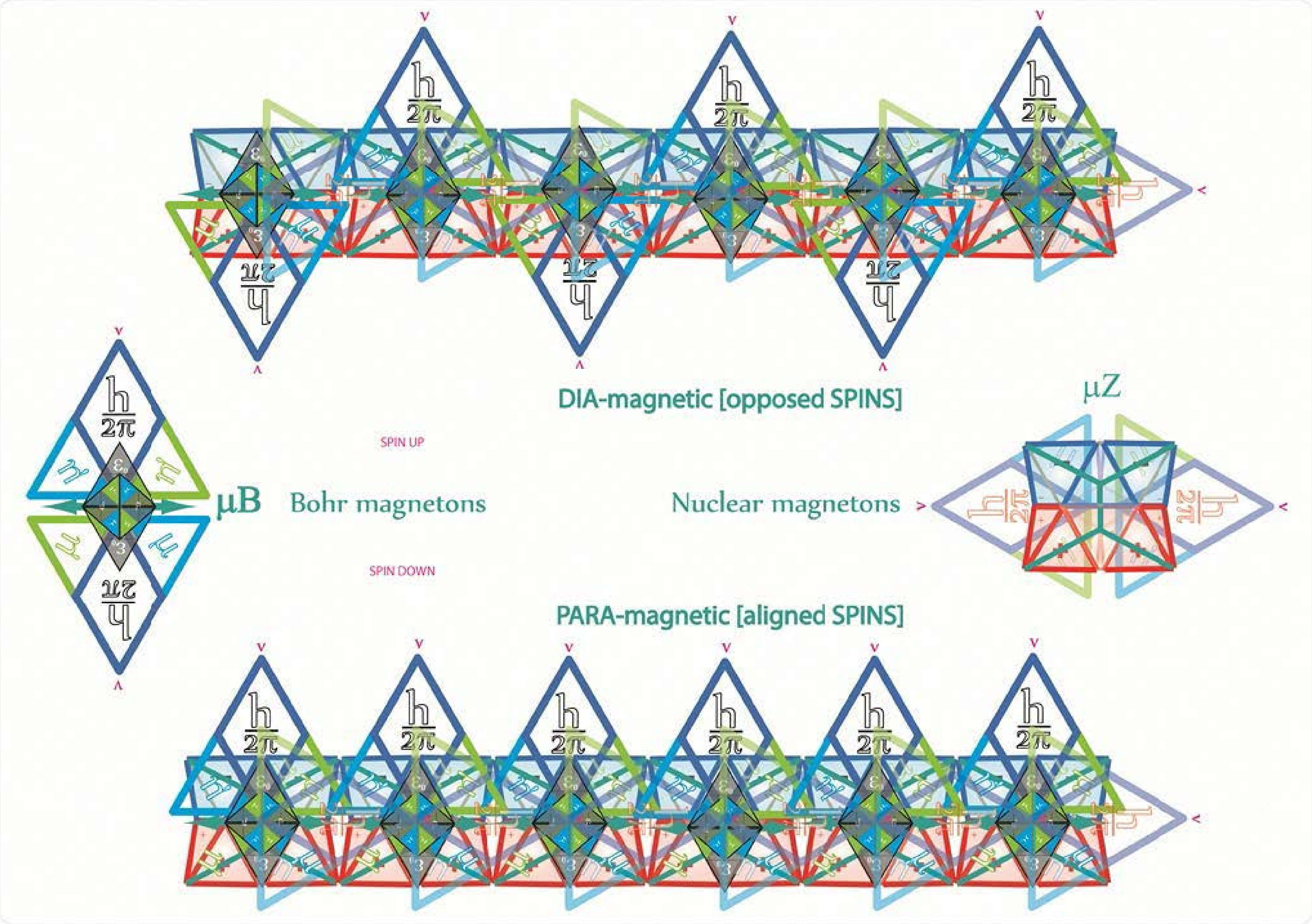
DIA-magnetic [opposed SPINS]
µZ
SPIN UP
µB
Bohr magnetons
Nuclear magnetons
SPIN DOWN
PARA-magnetic [aligned SPINS]
h
2π
ε₀
μ₀

Tetryonics 19.11 - DIA & PARA-magnetic fields

Copyright ABRAHAM [2008] - all rights reserved

12
[12-0]
1.2e20

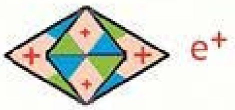

Positron

Leptons

181,010,964,200 C/kg

12
[0-12]
1.2e20

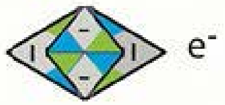

Electron

12
[24-12]
2.25e23

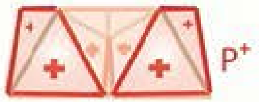

Proton

Baryons

96,539,180.9 C/kg

0
[18-18]
2.25e23

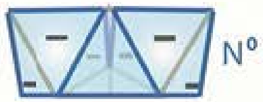

Neutron

All mass-Matter have distinct charge geometries
and once in motion are subject to EM forces
as a result of their geometries & KEM field

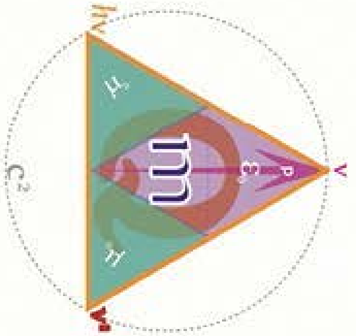

Lorentz force Law

$$\vec{F} = q\vec{E} + q\vec{v} \times B$$

LORENTZ force — Electric force — charge velocity — Magnetic force

all Matter in motion is subject to
NEWTON'S SECOND LAW OF MOTION

$$F = ma$$

Combining
Lorentz's force law & Newton's Third Law
we obtain a EM mass-charge quotient
for any particle of mass-Matter in motion

$$(M/Q)a = E + v \times B$$

mass-charge ratios

1.335180067 e-20 C

1.810109642 e11 $\left[\frac{q}{m}\right]$n 1e19 V

7.376238634 e-51 kg

are the basis for mass spectroscopy

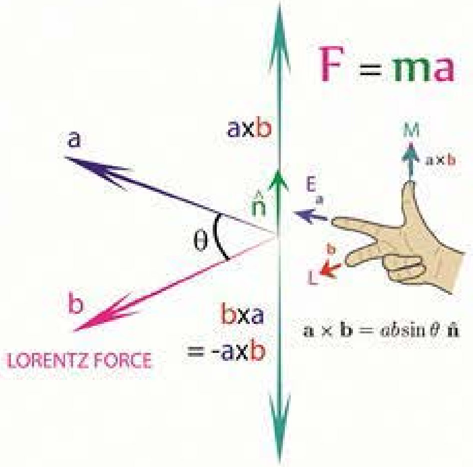

For every action there is an equal and opposite reaction
(NEWTON'S THIRD LAW OF MOTION)

Q *Any two particles with the same EM mass-to-charge ratio follow the same path in a vacuum when subjected to the same external electric field* **M**

Dependent on their generations

Charge/mass

$\left(\frac{q_t}{m_t}\right)$ 181,010,964,200 C/kg $\left(\frac{q_e}{m_e}\right)$

4:4 12:12

Tetryons and Leptons can have identical mass-charge ratios

96,487,720.78 C/kg

Elements

48,256,721.99 C/kg

0
[24-24]
2.2512e23

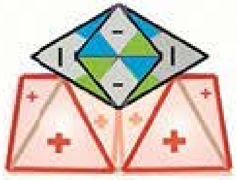

Hydrogen

Deuterium

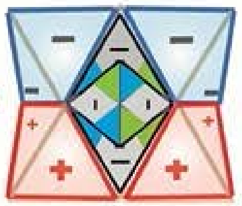

0
[42-42]
4.5012e23

Tetryonic theory shows mass-charge ratios are a measure of mass-energy geometries in Matter topologies

Tetryonics 20.01 - mass-Charge ratios

Collider particle track physics

A collider is a type of a particle accelerator involving directed beams of particles.

E

t **4** [0-4]

4 [4-0] t

Tetryons have the same 1/3 elementary charge as some Quarks

d **4** [4-8]

4 [8-4] d

m

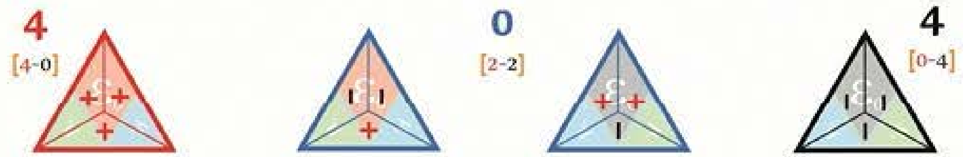

Colliders may either be ring accelerators or linear accelerators,
and may collide a single beam of elementary particles against a stationary target or two beams head-on.

2D bosons & photons are radiant mass-enery geometries

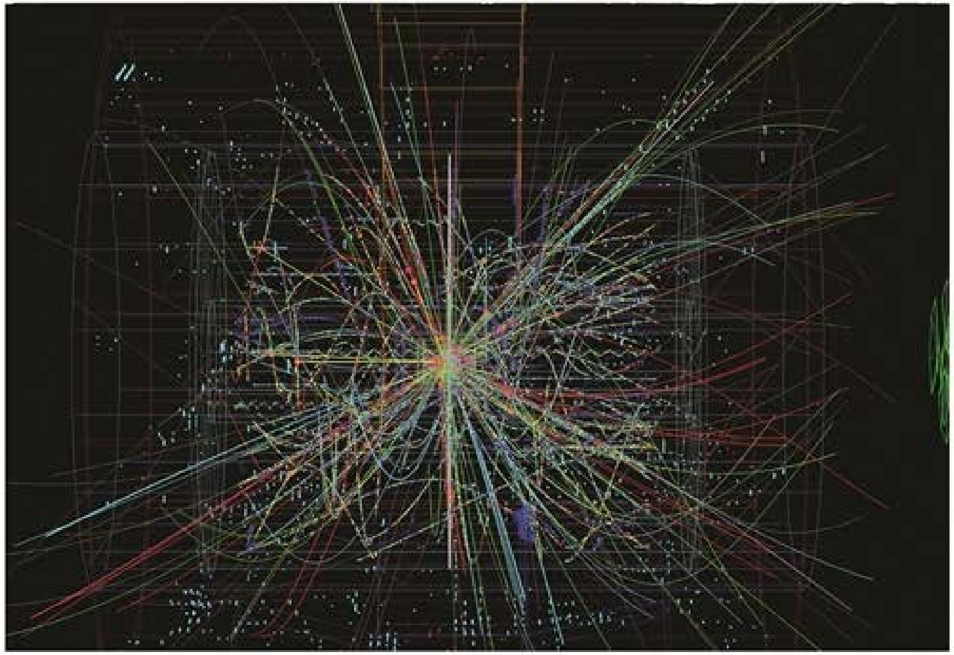

3D Matter Fermions are standing wave mass-energy topologies

Analysis of the by-products of these collisions without a clear definition of, and distinction between, EM mass-energy geometries & Matter topologies through the charged geometrics of Tetryonics ceates a misleading and erroneous picture of the particles created in high energy collisions within particle physics accelerator experiments

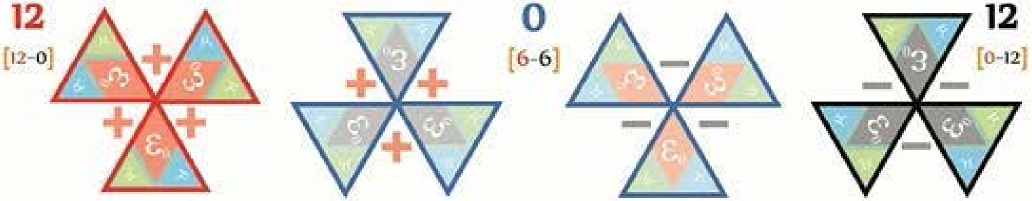

Q

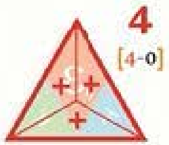

181,010,964,200 C/Kg

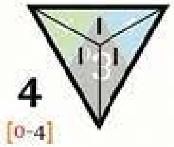

Tetryons can have the same mass-charge ratios as Leptons

12
[0-12]
e^-
1.2e20

181,010,964,200 C/Kg

12
[12-0]
e^+
1.2e20

M

Radiant EM mass-energy geometries

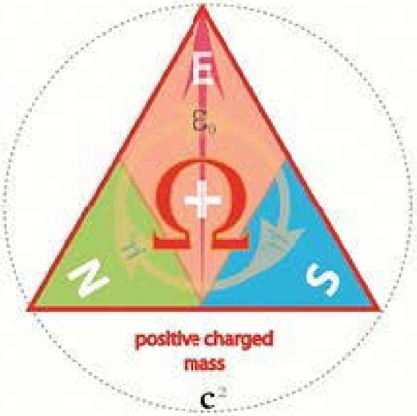

mass Planck quanta

$$\frac{n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

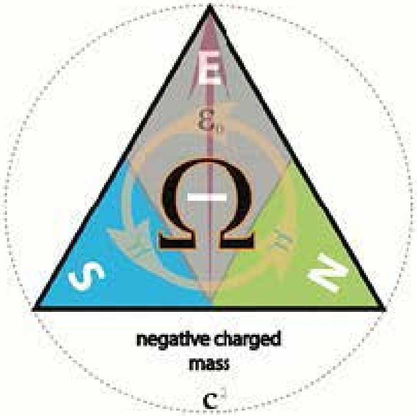

Charge Planck quanta

$$\frac{1\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

Bosons Planck quanta

$$\frac{\text{ODD}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

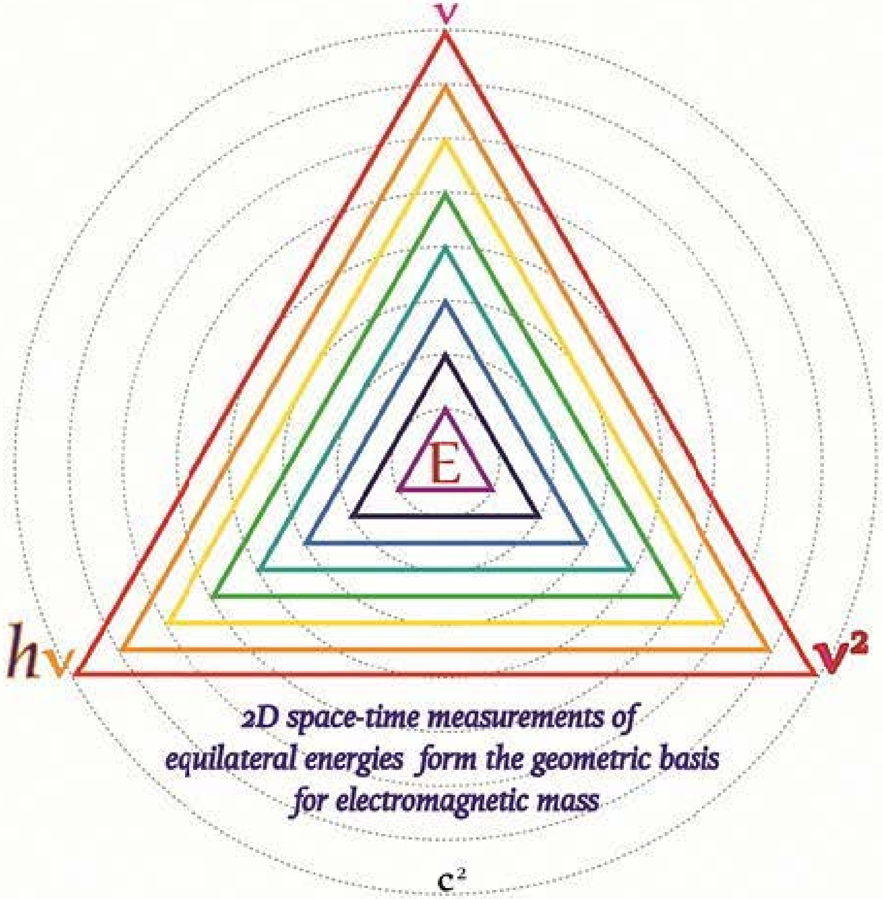

2D space-time measurements of equilateral energies form the geometric basis for electromagnetic mass

Photons Planck quanta

$$\frac{2\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

EM waves Planck quanta

$$\frac{\text{EVEN}\,\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

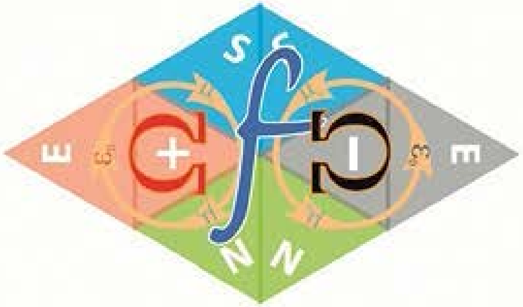

2D Planar divergent EM mass-energies form radiant Electromagnetic waves

ElectroMagnetic mass is a property of Matter
[being the energy content of its charged geometry per unit time]

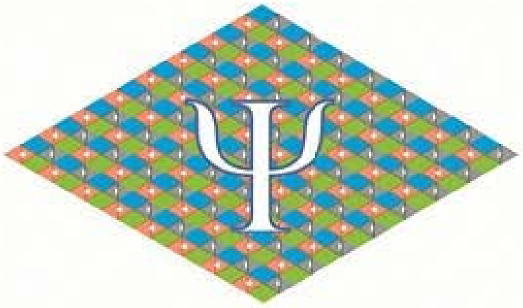

ENERGY

2D Equilateral scalar energy-momenta

Q
charge

E
energy

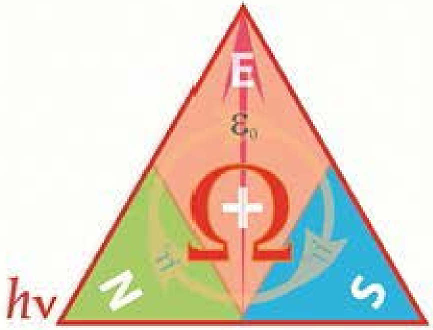
E
ε_0
Ω
+
hv
N
S

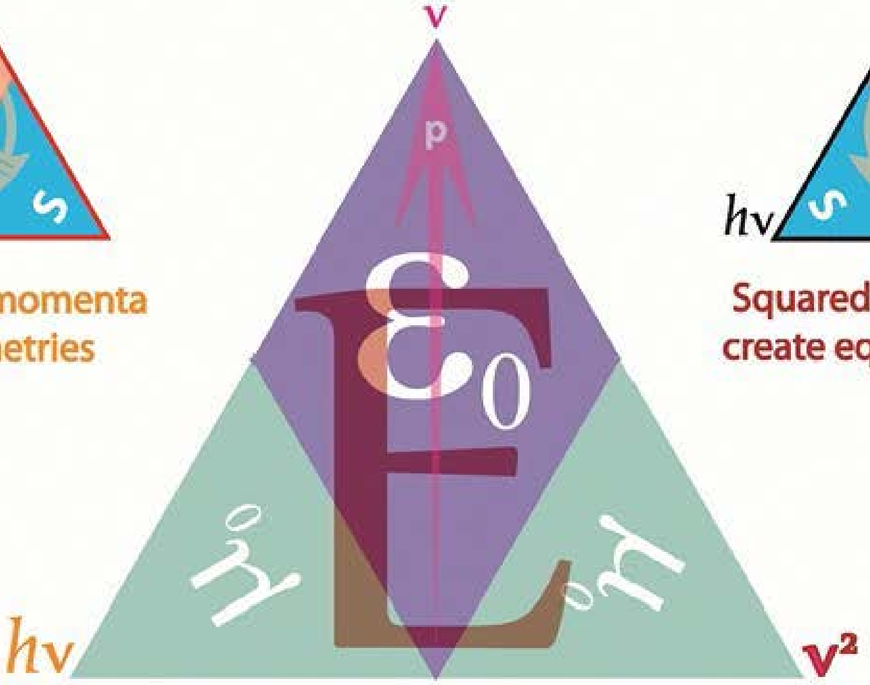
v
p
ε_0
E
μ_0
μ_0
hv
v²

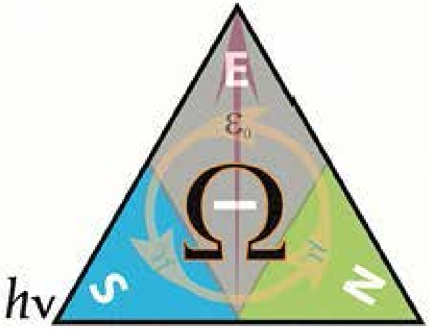
E
ε_0
Ω
−
hv
S
N

Quantised angular momenta are energy geometries

Squared numbers in physics create equilateral geometries

Planck's constant of
mass-energy momenta
is the source of all physical
constants and force relationships

Separated fields of charge
create electromotive forces
accelerating Material bodies
within them

ENERGY Planck quanta

$$n\pi\left[\left[m\Omega v^2\right]\right]$$

mass velocity

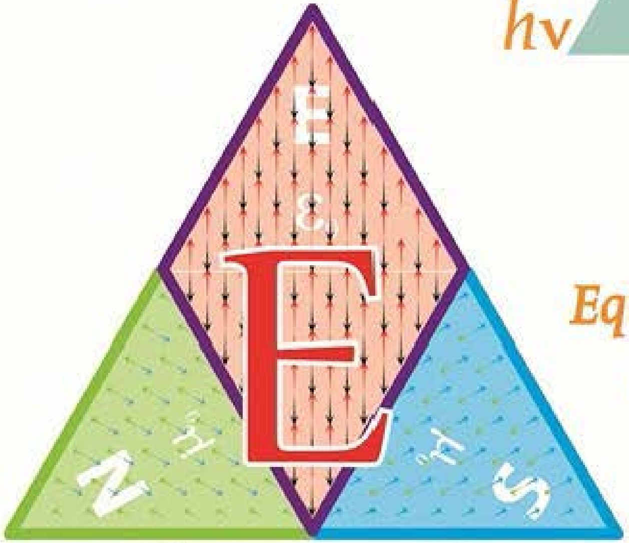
E
ε_0
E
μ_0
μ_0
N
S

Positive charge energy momenta fields

Equilateral energy-momenta form the foundation of all Forces, EM masses & Matter

All energy seeks equilibrium

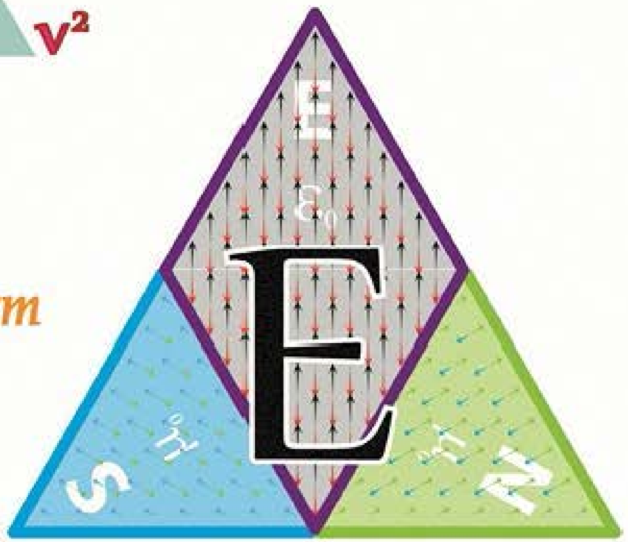
E
ε_0
E
μ_0
μ_0
S
N

Negative charge energy momenta fields

3D Matter topology is NOT a property of 2D ElectroMagnetic mass-energies

it is a measure of the closed 3D standing-wave spatial topology of all fermionic particles created by their charged equilateral mass-energy momenta geometries

Matter displaces vacuum energies to create convergent gravity fields around it

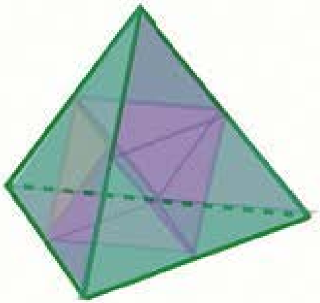

Tetryons are the quantum of Matter

Tetryons

$$\frac{4\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Fermions

$$\frac{12\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Mesons

$$\frac{24\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

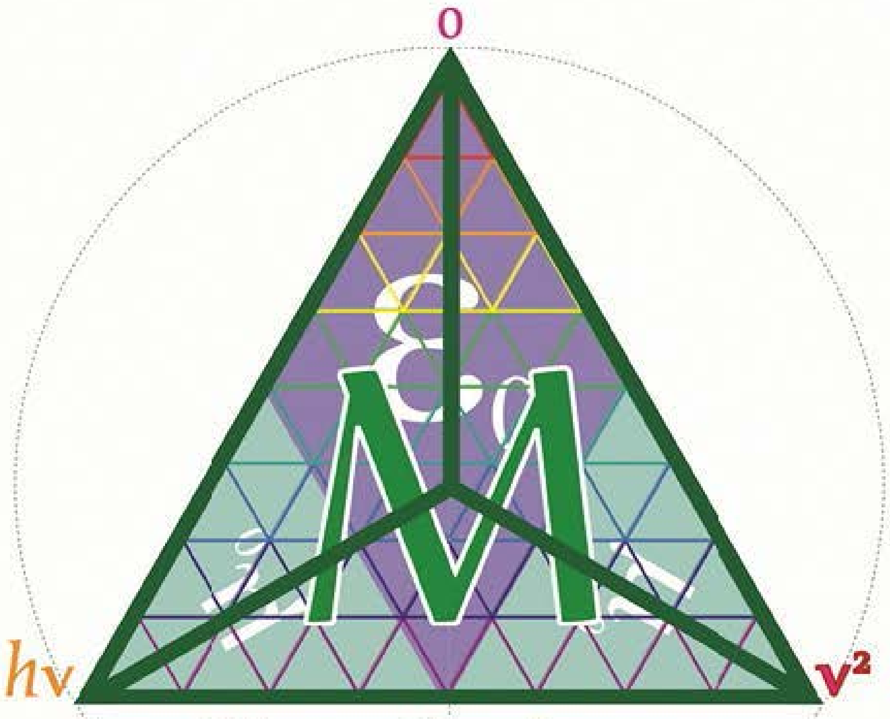

3D Matter particles are Z component second squared mass-energy topologies & created closed volumes

c^4

Matter

$$\frac{T\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

The mass-energies of Matter are Lorentz invariant to velocity changes

All Matter is comprised of [and radiates divergent] kEM mass-energies

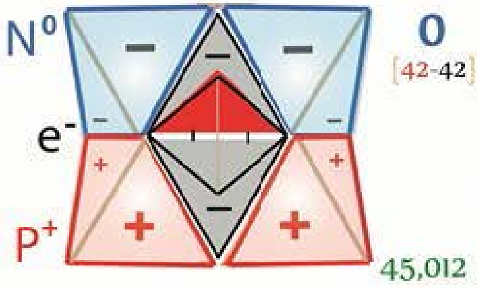

Deuterium is the quantum of all Elements

Baryons

$$\frac{36\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Nuclei

$$\frac{72\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

Deuterons

$$\frac{84\pi}{c^4}\left[\left[\underset{\text{mass}}{m}\,\overset{\text{Planck quanta}}{\Omega}\,\underset{\text{velocity}}{v^2}\right]\right]$$

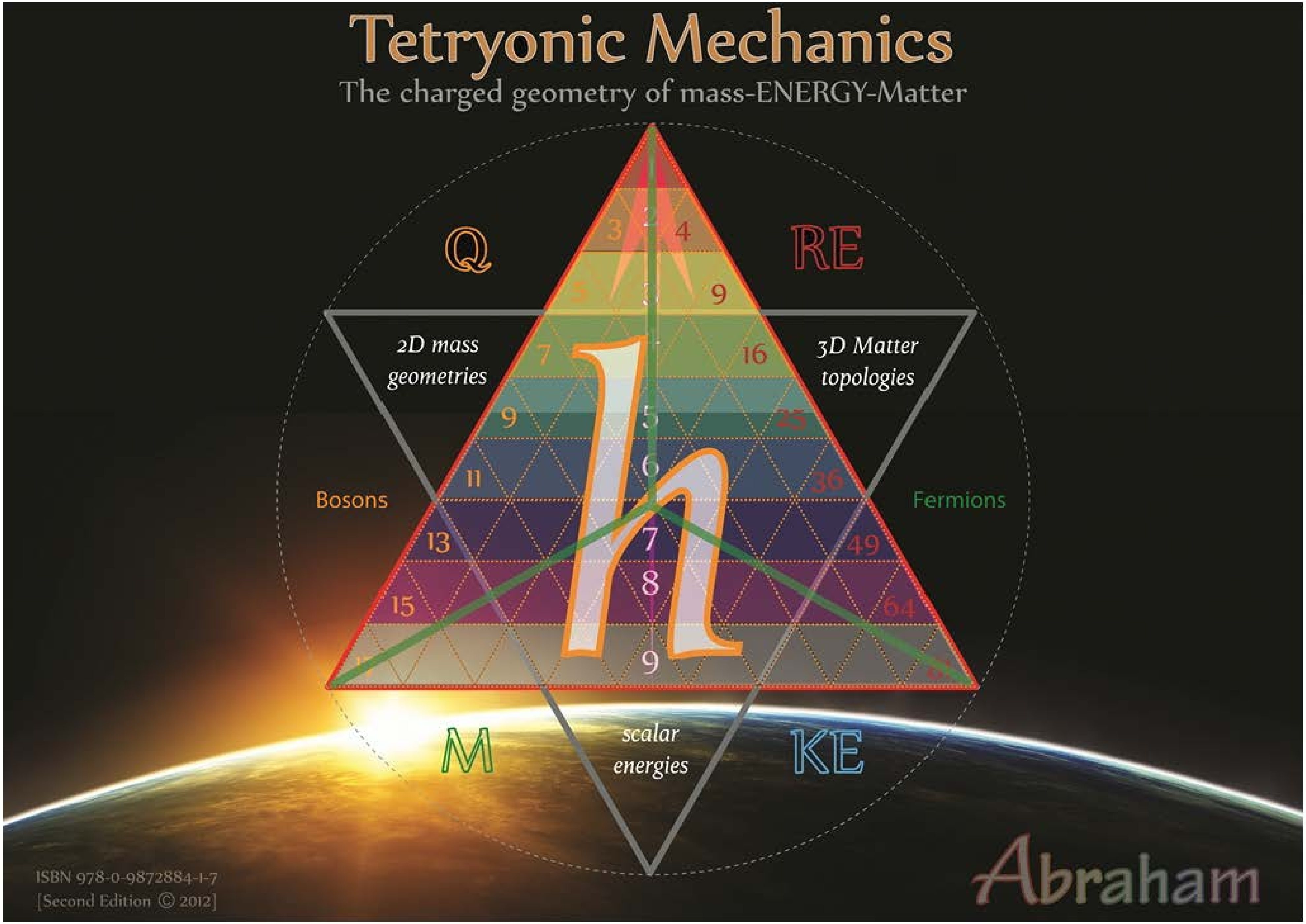

Tetryonic Mechanics

The charged geometry of mass-ENERGY-Matter

Q
RE
2D mass geometries
3D Matter topologies
Bosons
Fermions
M
KE
scalar energies
h
3
4
5
9
7
16
9
25
11
36
13
49
15
64
5
6
7
8
9

ISBN 978-0-9872884-1-7
[Second Edition © 2012]

Abraham

Copyright ABRAHAM [2008] - all rights reserved

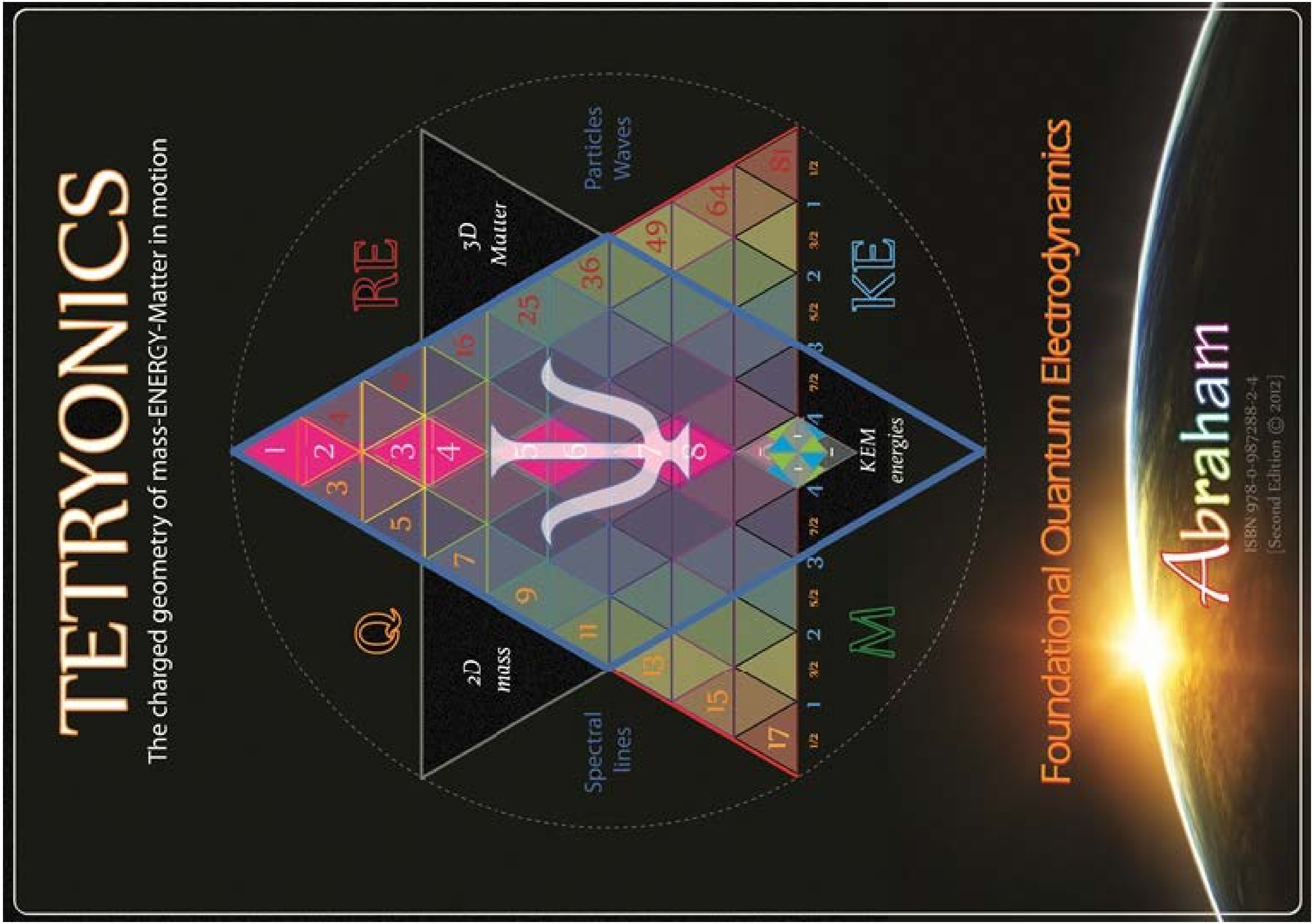
TETRYONICS
The charged geometry of mass-ENERGY-Matter in motion
RE
Q
KE
M
3D Matter
2D mass
Particles Waves
Spectral lines
KEM energies
1
2
3
4
5
6
7
8
9
11
13
15
16
17
25
36
49
64
81
1/2
1
3/2
2
5/2
3
7/2
4
Foundational Quantum Electrodynamics
Abraham
ISBN 978-0-987288-2-4
[Second Edition © 2012]

Tetryonics 21.00 - Tetryonic Electrodynamics title page

Copyright ABRAHAM [2008] - all rights reserved

James Clerk Maxwell

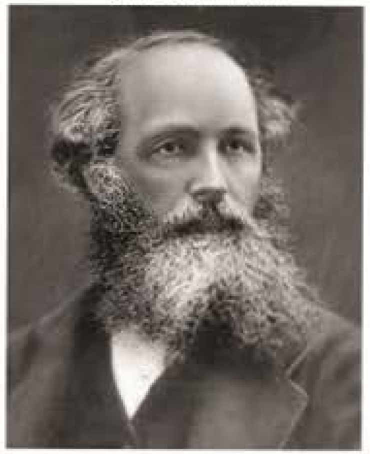

(13 June 1831 – 5 November 1879)

Q

[v-v]

Net Electric charge is the result of quanta imbalances in equilateral geometry

EM Planck Energy

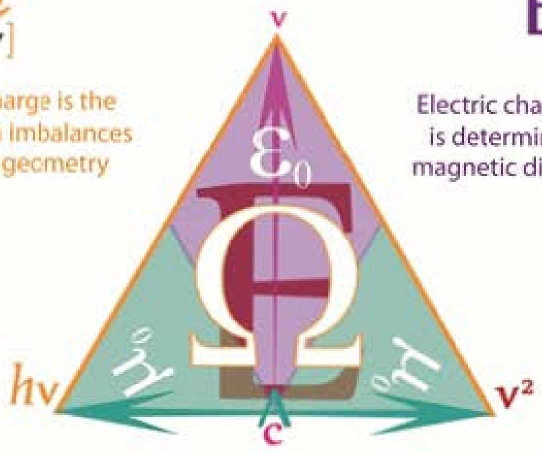

ORTHOGONAL MAGNETIC DIPOLE FIELD

E

Electric charge polarity is determined by the magnetic dipole vector

$$\oint \vec{E}\cdot d\vec{A}=\frac{q}{\varepsilon_0}$$

Gauss's Law for Electricity

$$\oint \vec{B}\cdot d\vec{A}=0$$

Gauss's Law for Magnetism

$$\oint \vec{E}\cdot d\vec{s}=\frac{-d\Phi_B}{dt}$$

Faraday's Law

$$\oint \vec{B}\cdot d\vec{s}=\mu_0 i+\frac{\mu_0\varepsilon_0 d\Phi_E}{dt}$$

Ampere's Law

E = electric field
A = area
q = charge
ε_0 = permittivity of free space (a constant)
B = magnetic field intensity
s = path length
Φ_B = magnetic flux
Φ_E = electric flux
t = time
μ_0 = permeability of free space (a constant)
i = current electric charge
$\mu_0\varepsilon_0$ = $1/c^2$, where c=speed of light

Maxwell's equations describe how Electric and Magnetic fields are generated and altered by each other and by charges and currents.

Maxwell's Equations

Maxwell's equations are a set of partial differential equations that, together with the Lorentz force law, form the foundation of classical electrodynamics, classical optics, and electric circuits.

ElectroMagnetic fields in turn provide the foundation of our modern electrical and communications technologies.

q

[1-0]

POSITIVE [Ω] CHARGE

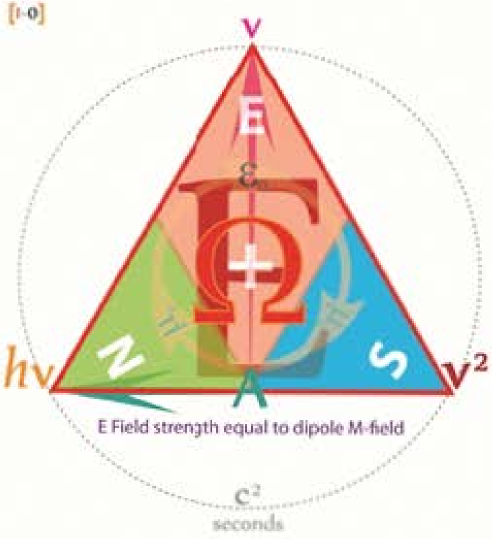

Q

Gauss' Law of Electric Flux

Electric fields diverge from Electric charge, and produce the Coulomb force,

Q

N

Gauss' Law of Zero Nett Magnetism

there are no isolated Magnetic poles, but the Henry force acts between the poles of a magnet,

S

ε_0

Faraday's Law of Inductance

Electric fields are produced by changing Magnetic fields,

μ_0

Ampere's Circuital Law

circulating Magnetic fields are produced by moving Electric fields and by electric currents.

NEGATIVE [Ω] CHARGE

q

[0-1]

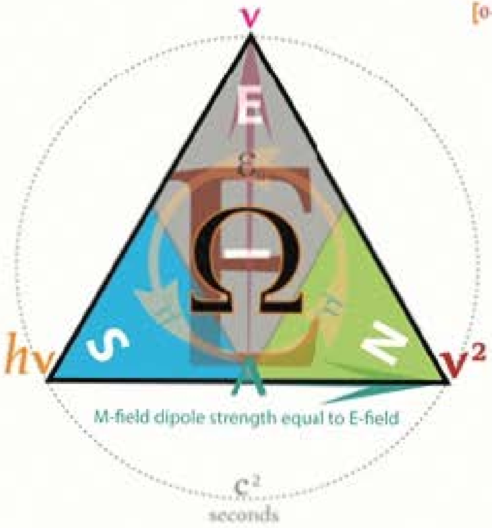

Copyright ABRAHAM [2008] - all rights reserved

Gauss' Laws

- Gauss' Law of Electric Flux and
- Gauss' Law of Magnetic dipoles

The laws were formulated by Carl Friedrich Gauss in 1835, but was not published until 1867.

They form two of Maxwell's equations which are the basis of classical electrodynamics.

The other two being:
- Faraday's law of induction, and
- Ampère's law with Maxwell's correction.

Gauss's law can be used to derive Coulomb's law, and vice versa.

Equilateral triangles are the foundational charge geometry of EM mass-Energy and Matter

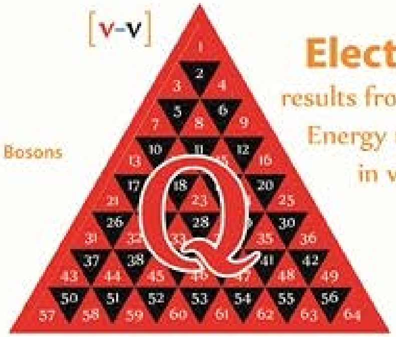
[v-v]
Bosons
Q
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64

Electric Charge

results from an imbalance of Energy momenta quanta in v² geometries

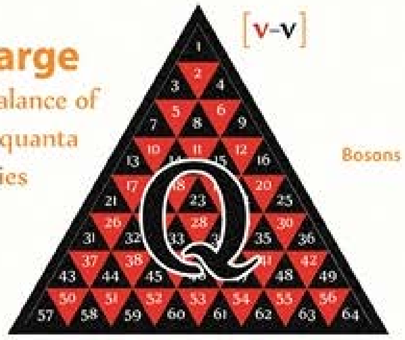
[v-v]
Bosons
Q
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64

Carl Friedrich Gauss

(30 April 1777 – 23 February 1855)

Integral form

$$\oint \vec{E} \cdot d\vec{A} = \frac{q}{\varepsilon_0} = 4\pi k q$$

Differential form

$$\nabla \cdot E = \frac{\rho}{\varepsilon_0} = 4\pi k \rho$$

Electric fields are divergent from their source

Although originally envisaged as a property applied to spherical geometries and usually illustrated with square sectional boundaries integrating a surface area

$$\Phi = EA = E4\pi r^2 = \frac{Q}{\varepsilon_0}$$

Charge is shown to be the result of the symmetry of equilateral quantised angular momenta

Integral form

$$\oint \vec{B} \cdot d\vec{A} = 0$$

Differential form

$$\nabla \cdot B = 0$$

There are NO Magnetic Monopoles

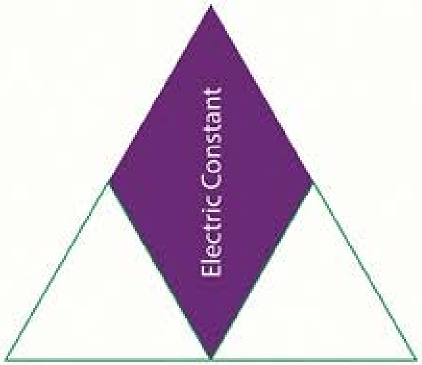
Electric Constant

$$\varepsilon_0 = 8.85418785\text{e-}12 \ \frac{F}{m}$$

$$\oint \vec{E} \cdot d\vec{A} = \frac{q}{\varepsilon_0}$$

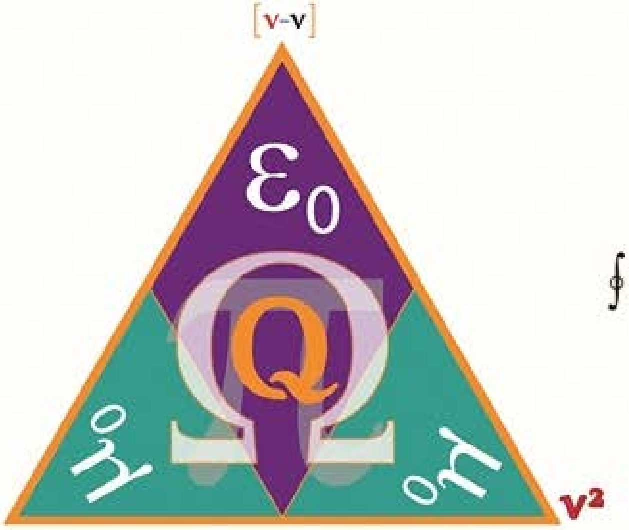
[v-v]
ε0
Ω
Q
μ0
μ0
v²

$$\oint \vec{B} \cdot d\vec{A} = 0$$

The electric flux through any closed surface is proportional to the enclosed electric charge

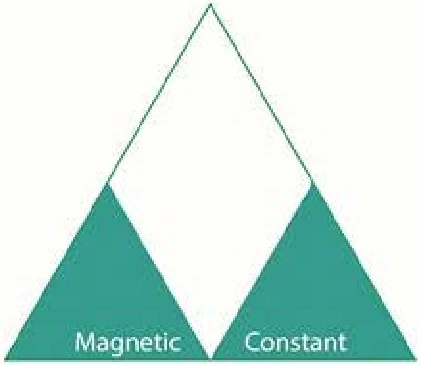
Magnetic
Constant

$$\mu_0 = 1.25663706\text{e-}6 \ \frac{H}{m}$$

Copyright ABRAHAM [2008] - all rights reserved

The Divergent force of Charge

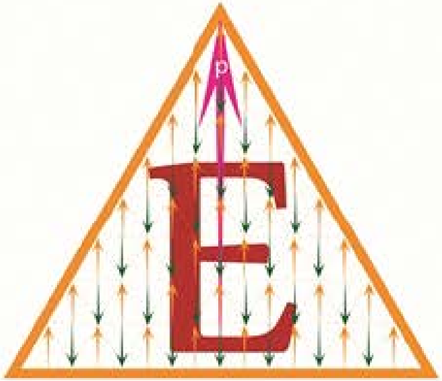

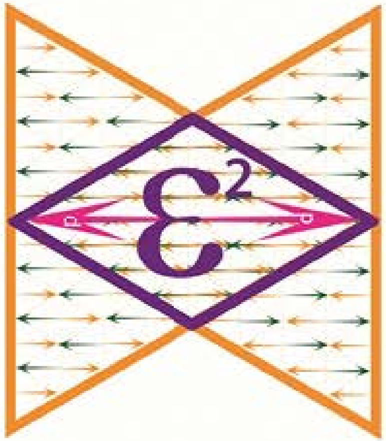

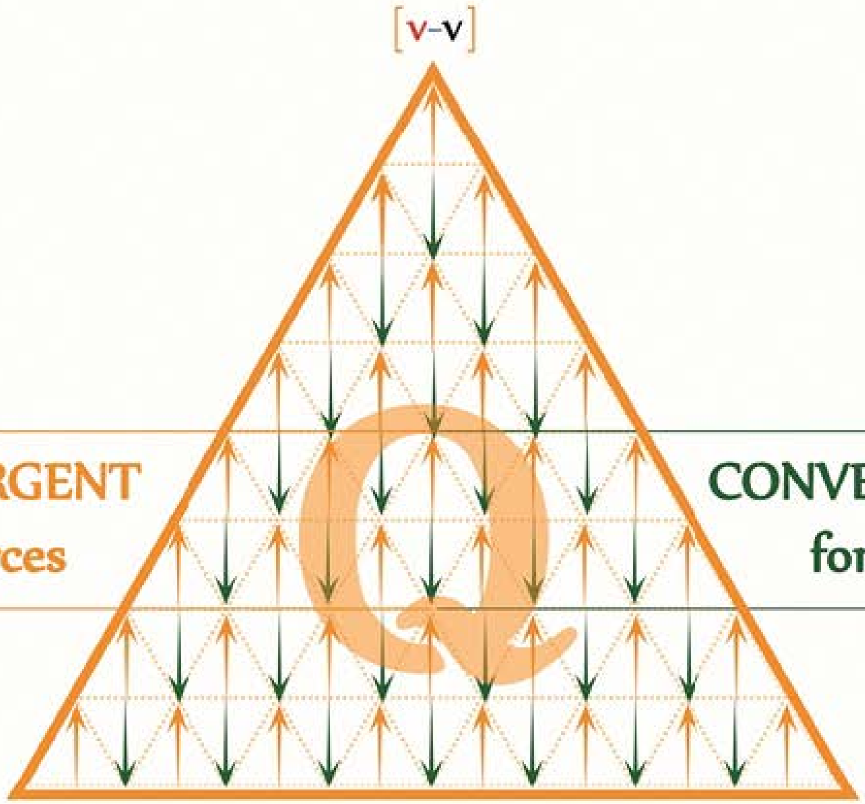

↑ **DIVERGENT**
forces

CONVERGENT ↓
forces

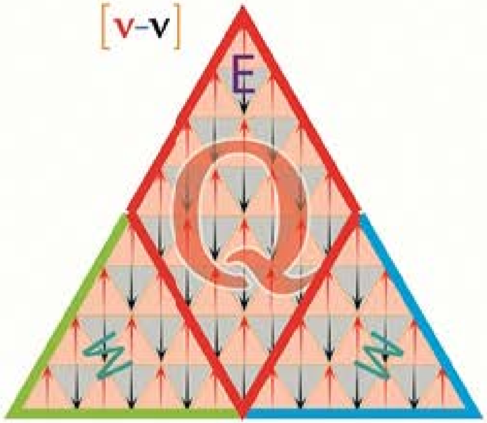

Charged EM fields create a net DIVERGENT force
as a result of their outward radiant energy momenta

but still possess CONVERGENT forces as a result of
their inherent equilateral energy momentum symmetries

Where these radiant charged fields diverge from their source
and overlap their superpositional energy momenta
create the forces of the Law of Interaction

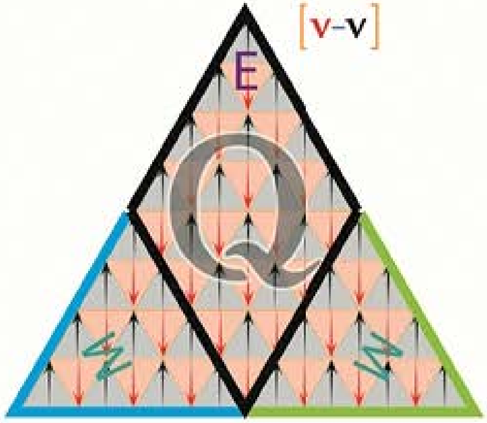

Coulomb's Law

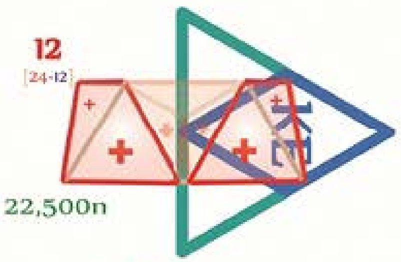
12
[24-12]
22,500n
KE

Proton

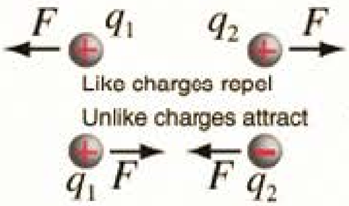
F
q1
q2
F
Like charges repel
Unlike charges attract
q1
F
F
q2

$$F = \frac{kq_1q_2}{r^2} = \frac{q_1q_2}{4\pi\varepsilon_0 r^2}$$

Coulomb's Law

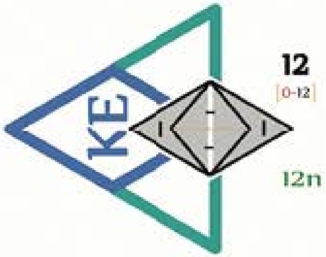
KE
12
[0-12]
12n

Electron

The magnitude of the electrostatic force between two point electric charges is directly proportional to the product of the magnitudes of each of the charges and inversely proportional to the square of the distance between the two charges.

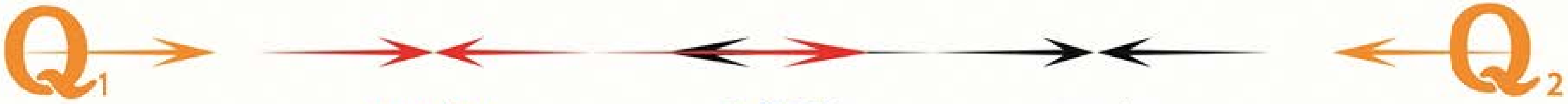

opposing energy momenta directions

co-linear energy momenta directions

opposing energy momenta directions

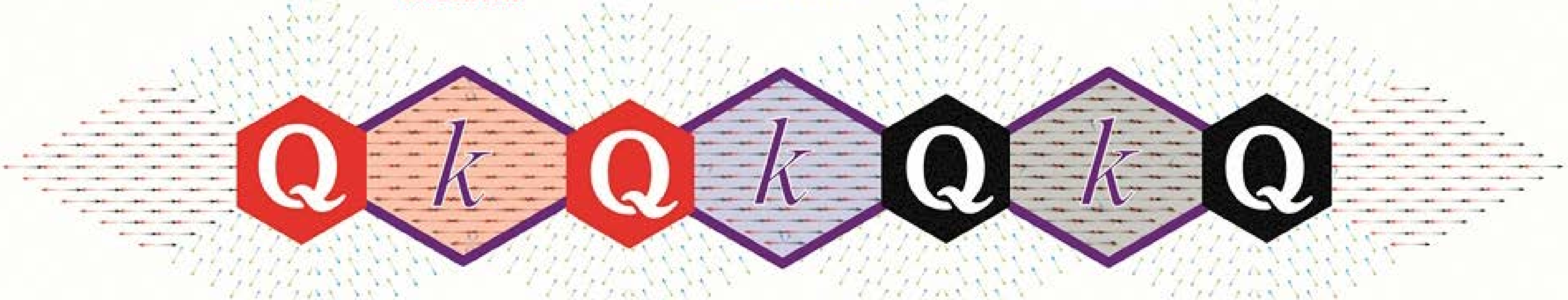

Similars Repel

Opposites Attract

Similars Repel

The law of Interaction

Copyright ABRAHAM [2008] - all rights reserved

Changing Electric Fields produce changing Magnetic fields

Changing Magnetic fields produce changing Electric fields

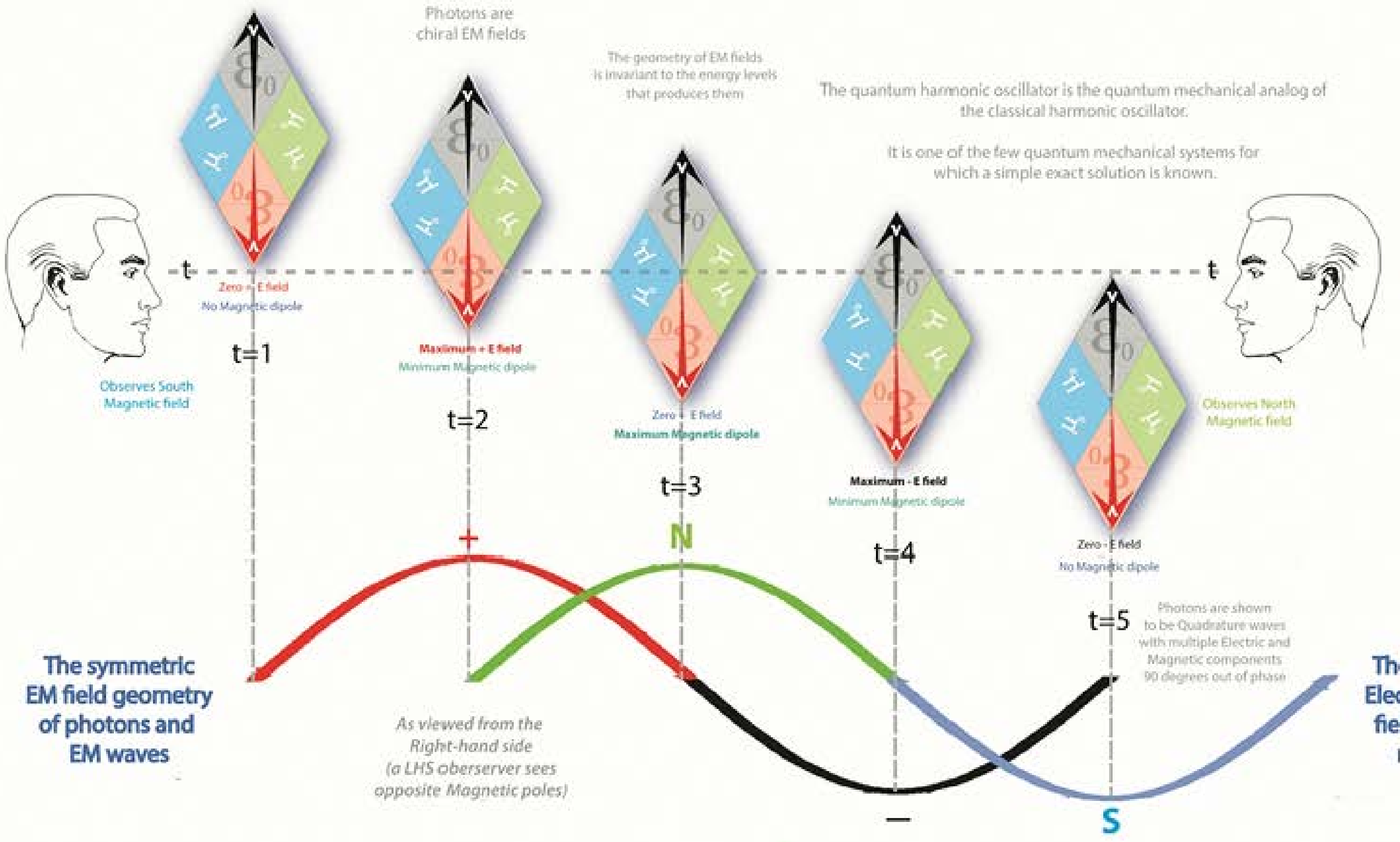
Photons are chiral EM fields
The geometry of EM fields is invariant to the energy levels that produces them
t
t
Zero + E field
No Magnetic dipole
t=1
Observes South Magnetic field
Maximum + E field
Minimum Magnetic dipole
t=2
Zero + E field
Maximum Magnetic dipole
t=3
Maximum - E field
Minimum Magnetic dipole
t=4
Zero - E field
No Magnetic dipole
t=5
Observes North Magnetic field
+
N
−
S
Photons are shown to be Quadrature waves with multiple Electric and Magnetic components 90 degrees out of phase
As viewed from the Right-hand side (a LHS oberserver sees opposite Magnetic poles)

The quantum harmonic oscillator is the quantum mechanical analog of the classical harmonic oscillator.

It is one of the few quantum mechanical systems for which a simple exact solution is known.

The symmetric EM field geometry of photons and EM waves

The symmetric ElectroMagnetic field strengths measured

Michael Faraday

(22 September 1791 – 25 August 1867)

EM field symmetry

"It is proposed that if a changing magnetic field can make an oscillating electric field, then

a changing electric field (from a oscillating electric charge) should make a changing magnetic field"

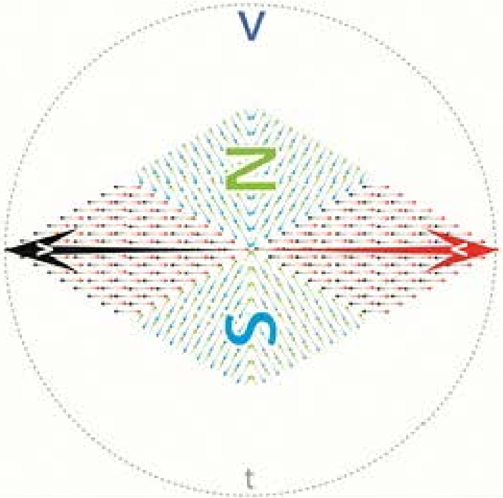

Any change in the magnetic environment of a coil of wire will cause a voltage (emf) to be "induced" in the coil.

No matter how the motional change is produced, a changing electromotive force will be generated.

The change could be produced by
changing the magnetic field strength,
moving a magnet toward or away from the coil,
moving the coil into or out of the magnetic field,
rotating the coil relative to the magnet.

All the resulting changes to the E & M fields and the vector strengths etc are the result of the inherent equilateral geometries of EM fields

Electromagnetic Induction

All ElectroMagnetic circuits are comprised of quantum inductive fields [ZPFs] and energies that obey Newton's third law and the conservation of energy

Electromagnetic induction is the production of an electric current across a conductor moving through a magnetic field.
It underlies the operation of generators, transformers, induction motors, electric motors, synchronous motors, and solenoids.

EM field coupling

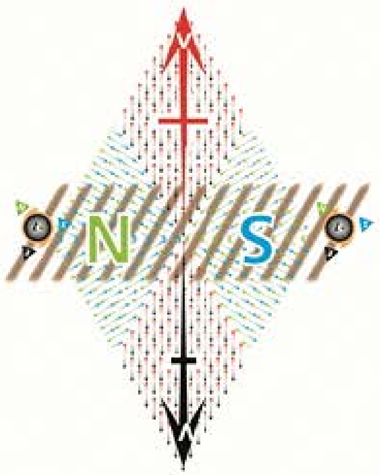

$$|\mathcal{E}| = \left|\frac{d\Phi_B}{dt}\right|$$

Michael Faraday formulated that electromotive force (EMF) produced around a closed path is proportional to the rate of change of the magnetic flux through any surface bounded by that path.

$$\nabla \times \mathbf{E} = -\frac{\partial \mathrm{B}}{\partial t}$$

$$V_{\rm p} = N_{\rm p}\frac{{\rm d}\Phi}{{\rm d}t}.$$

For the special case of a coil of wire, or a transformer circuit composed of N loops with the same area, Faraday's general equation becomes

$$V_{\rm s} = N_{\rm s}\frac{{\rm d}\Phi}{{\rm d}t},$$

Inductive Coupling

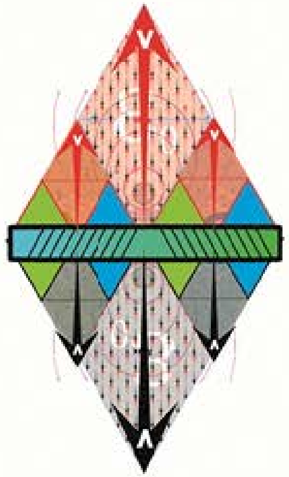

A corollary of Faraday's Law, together with Ampère's law and Ohm's law is Lenz's law:
The EMF induced in an electric circuit always acts in such a direction that the current it drives around the circuit opposes the change in magnetic flux which produces the EMF

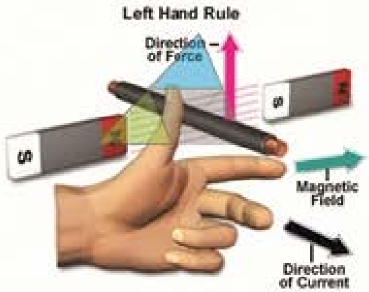

Flemings left hand rule
(for electric motors)

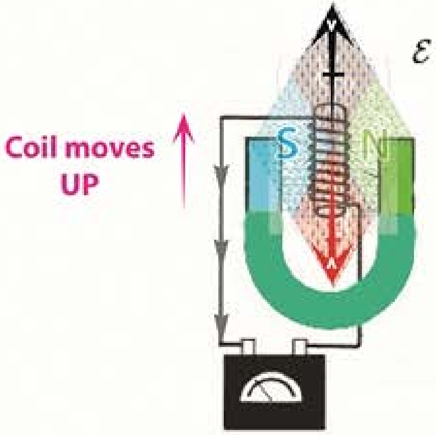

$$\mathcal{E} = -N\frac{d\Phi_B}{dt}.$$

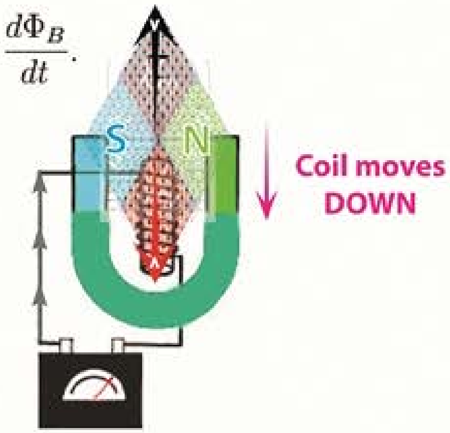

An electric current passes through a solenoid, resulting in a magnetic field. When you wrap your right hand around the solenoid with your fingers in the direction of the conventional current, your thumb points in the direction of the magnetic north pole.

An electric current passes through a straight wire. Here, the thumb points in the direction of the conventional current (from positive to negative), and the fingers point in the direction of the magnetic lines of flux.

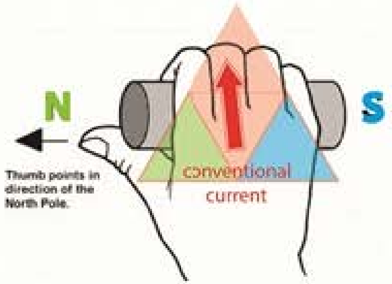
N
S
Thumb points in direction of the North Pole.
conventional current

Fleming's right hand rule
(for EM induction)

Michael Faraday

(22 September 1791 – 25 August 1867)

Faraday's Law of Induction

Electromagnetic induction is the production of a potential difference (voltage) across a conductor when it is exposed to a varying magnetic field.

In the frame of a conductor moving relative to the magnet, charged particles in the conductor experience a coulombic force due to the neutralised electric fields of the permanent magnet.

Faraday's law states that:
The induced electromotive force (emf) in any closed circuit is equal to the time rate of change of the magnetic flux through the circuit.

$$\boxed{\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_B}{dt}}$$

The motion of EM fields creates an induced motion in external fields and particles

The linear energy momenta of the neutralised E-fields produces differing directions of induced emf (current flow)

A permanent magnet has a static Magnetic dipole field and neutralised Electric field geometries

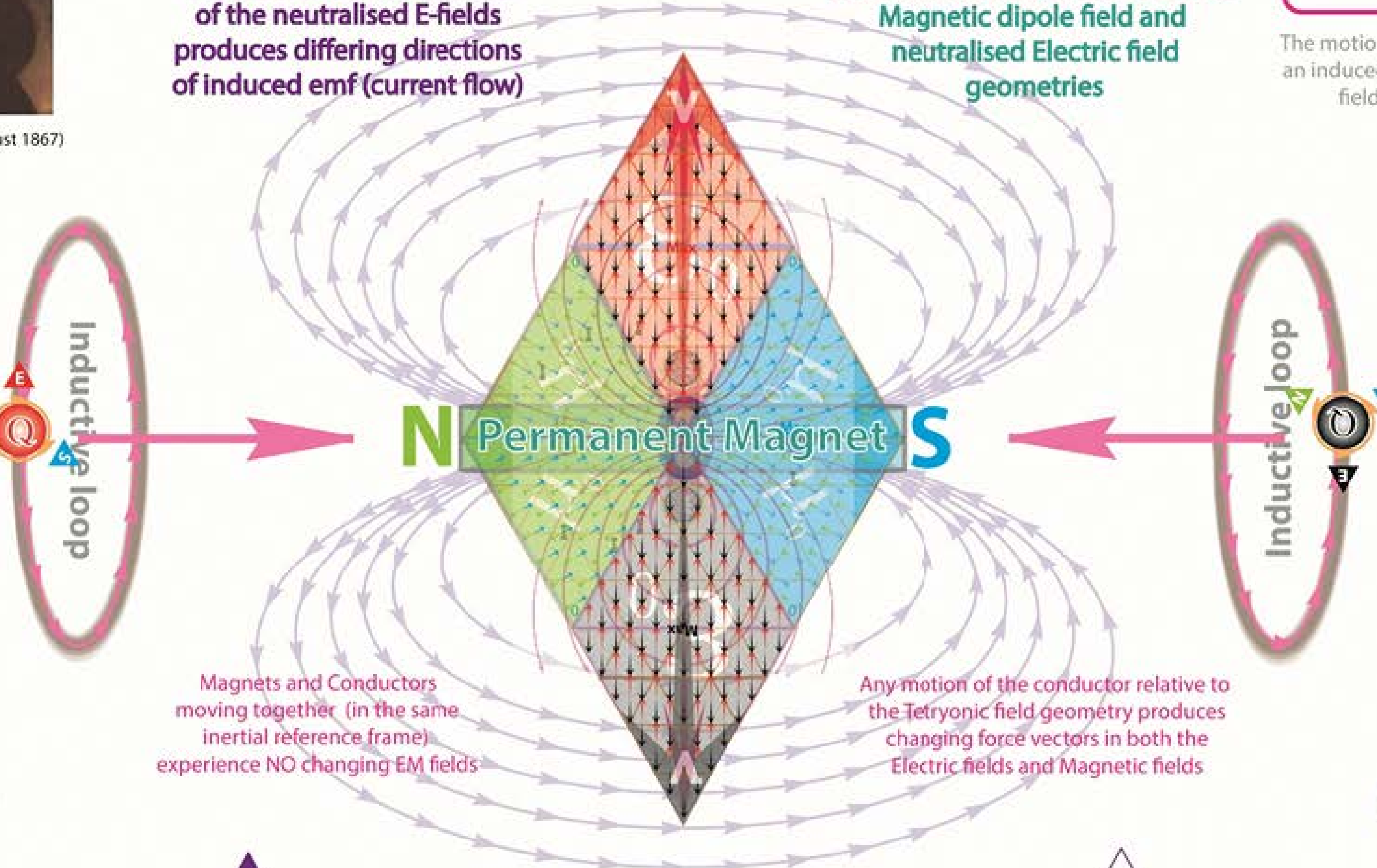

Charged particles in the conductor are subject to voltage emfs

Charged particles in the conductor are subject to voltage emfs

Magnets and Conductors moving together (in the same inertial reference frame) experience NO changing EM fields

Any motion of the conductor relative to the Tetryonic field geometry produces changing force vectors in both the Electric fields and Magnetic fields

An Electrostatic (charged) particle has Electric fields of a particular polarity and neutralised Magnetic fields

Electric Constant

It is impossible (due to the geometry of EM fields) to move either the permanent Magnet or the Conductor without producing variations in the Electric and Magnetic fields

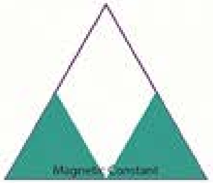

A Magneto-static body of Matter has a dipolar Magnetic field and neutralised Electric fields

Copyright ABRAHAM [2008] - all rights reserved

Ampere's Law

In classical electromagnetism, Ampère's circuital law, relates the integrated magnetic field around a closed loop to the electric current passing through the loop.

Q

[v-v]

All energy momenta quanta have equilateral geometries, [quantised angular momenta [Ω]] that can be modelled as a continuous steady current within a closed inductive loop

$$\boxed{\oint \vec{B}\cdot d\vec{s} = \mu_0 i + \frac{1}{c^2}\frac{\partial}{\partial t}\int \vec{E}\cdot d\vec{A}}$$

The nett Charge of any particle or surface is the result of the quantised equilateral energy momenta symmetries of its integral surface area

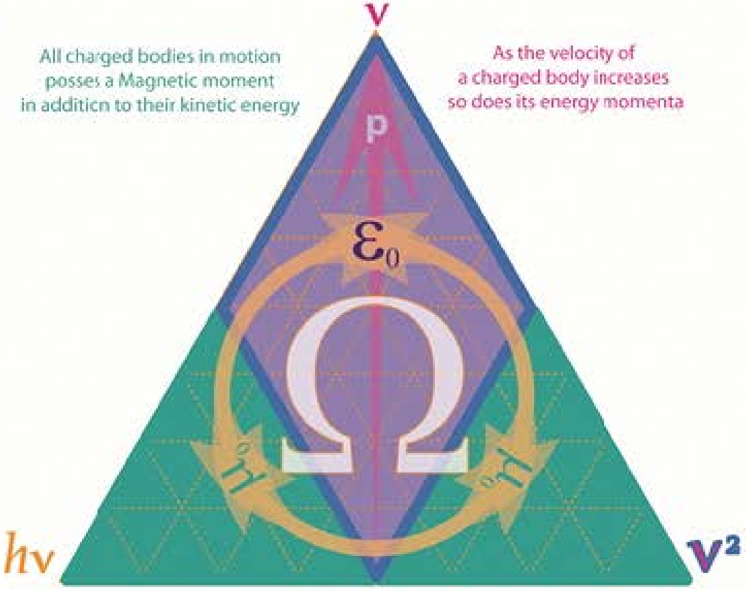
v
p
ε_0
Ω
hv
v²

All charged bodies in motion posses a Magnetic moment in addition to their kinetic energy

As the velocity of a charged body increases so does its energy momenta

$$\oint_C \mathbf{B}\cdot d\boldsymbol{\ell} = \mu_0 I_{\text{enc}}$$

$$\mathbf{J}_f + \mathbf{J}_D + \mathbf{J}_M = \mathbf{J}_f + \mathbf{J}_P + \mathbf{J}_M + \varepsilon_0\frac{\partial \mathbf{E}}{\partial t} = \mathbf{J} + \varepsilon_0\frac{\partial \mathbf{E}}{\partial t},$$

With the addition of Maxwell's Displacement current to account for time varying electric fields without a physical flow of charged Matter the way forward is paved for Planck, Lorentz and Tetryonics

The Closed circuit or loop can be any geometric shape with Tetryonics dictating equilateral [triangular] Planck energies & quantum charge geometries with Tetrahedral geometries for Matter quanta

Andre Ampere

(20 January 1775 – 10 June 1836)

Ampere's law relates magnetic fields to the electric currents that produce them.

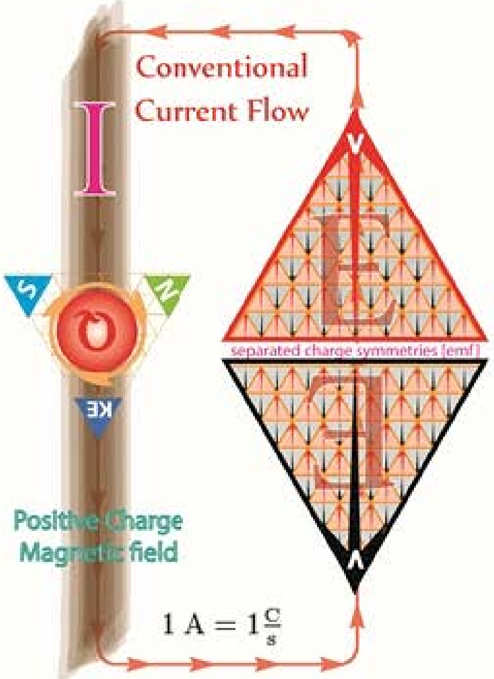
Conventional Current Flow
I
S
N
KE
v
E
separated charge symmetries [emf]
Positive Charge Magnetic field
$1\ \mathrm{A} = 1\frac{\mathrm{C}}{\mathrm{s}}$

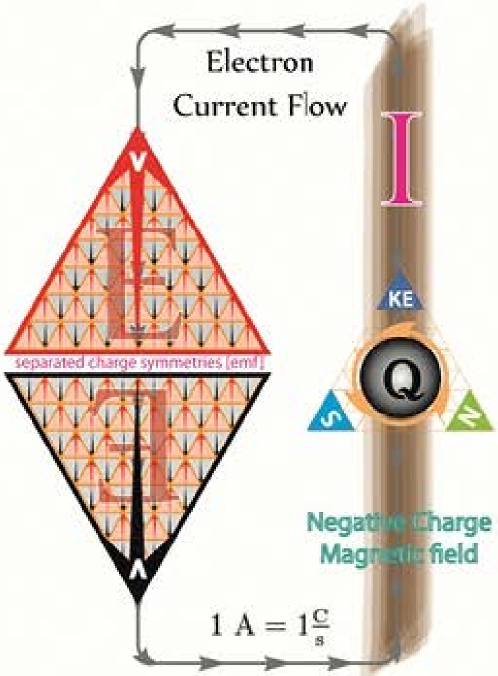
Electron Current Flow
I
KE
Q
S
N
v
E
separated charge symmetries [emf]
Negative Charge Magnetic field
$1\ \mathrm{A} = 1\frac{\mathrm{C}}{\mathrm{s}}$

$$F = 2k_A \frac{I_1 I_2}{r}$$

Ampère's force law states that there is an interactive force between two parallel wires carrying an electric current.

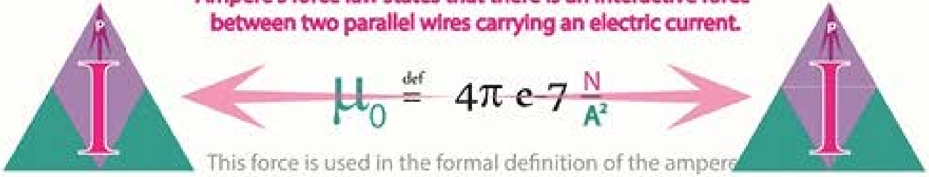

$$\mu_0 \overset{def}{=} 4\pi \text{ e-7 } \frac{N}{A^2}$$

This force is used in the formal definition of the ampere which states that it is "the constant current which will produce an attractive force of 2 × 10–7 newtons per metre of length between two straight, parallel conductors of infinite length and negligible circular cross section placed one metre apart in a vacuum"

These attractions and repulsions between electric currents differ fundamentally from the effects produced by electricity in repose. First, they cease, as chemical decompositions do, as soon as we break the circuit. Second, in ordinary electric attractions and repulsions, opposite charges attract, and like charges repel; in the attractions and repulsions of electric currents, we have precisely the contrary; it is when the two conducting wires are placed parallel in such a way that their ends of the same sign are next to each other that there is attraction, and there is repulsion when the ends of the same sign are as far apart as possible. Third, in the case of attraction, when it is sufficiently strong to bring the movable conductor into contact with the fixed conductor, they remain attached to one another like two magnets, and do not separate after a while, as happens when two conducting bodies, oppositely electrified, come to touch.

André Marie Ampère (1775 - 1836)

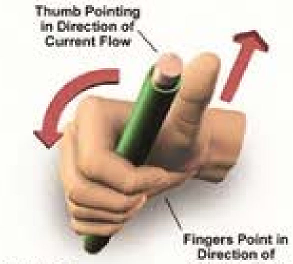
Thumb Pointing in Direction of Current Flow
Fingers Point in Direction of Magnetic Field

Figure 3

$$1\text{ A} = 1\frac{\text{C}}{\text{s}}$$

Interactive force between Conductors

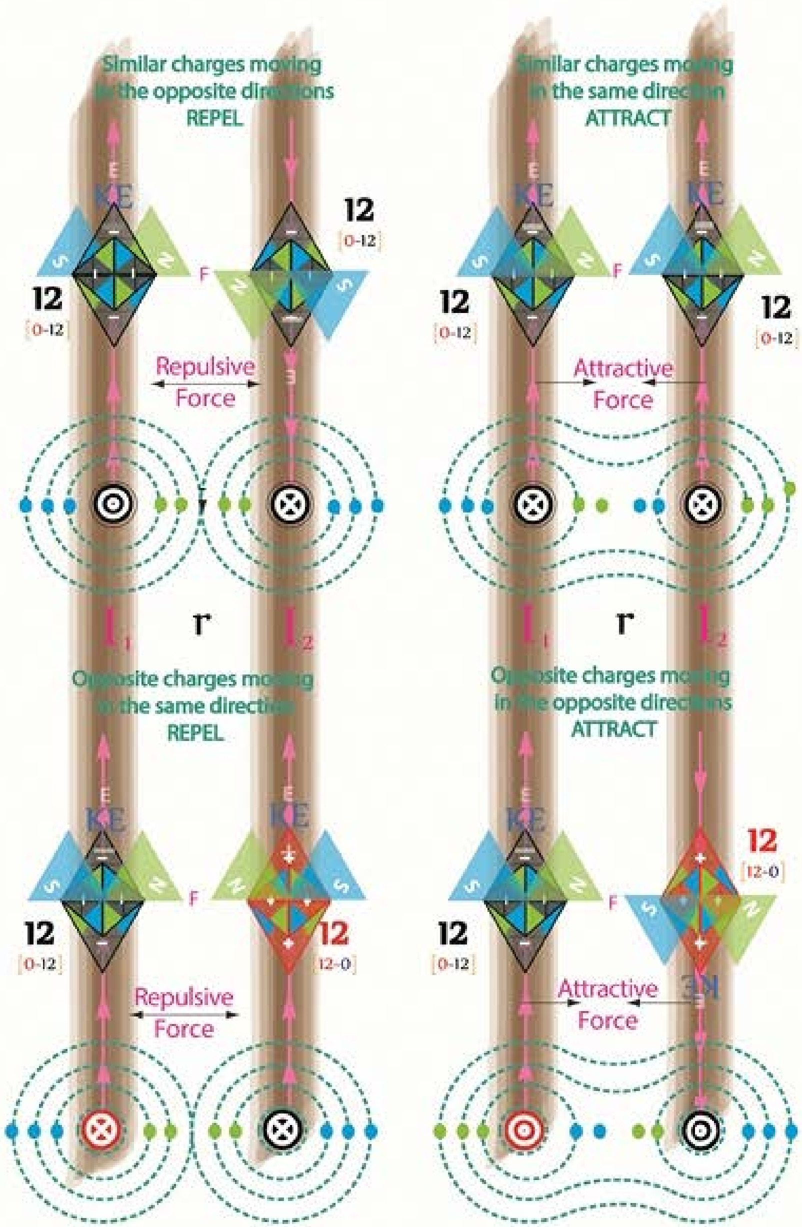
Similar charges moving in the opposite directions REPEL
Similar charges moving in the same direction ATTRACT
Repulsive Force
Attractive Force
12
[0-12]
12
[12-0]
KE
E
F
S
N
I1
I2
r
Opposite charges moving in the same direction REPEL
Opposite charges moving in the opposite directions ATTRACT
Repulsive Force
Attractive Force

Copyright ABRAHAM [2008] - all rights reserved

ElectroMotive exchange particles

W

All four of the fundamental forces involve the exchange of ODD numbers of electromagnetic charge bosons so as to transfer discrete energy momenta between separated Matter

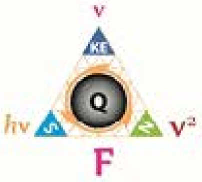

In ElectroMagnetic field or circuits, when charged Matter moves along electric field lines, electrical work is done on them by the electromotive force, whether it involves storing potential energy (negative work) or increasing kinetic energy (positive work)

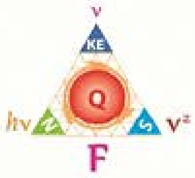

γ

The EM exchange force denotes a force produced by the exchange of force carrier particles, such as the electromagnetic force produced by the exchange of photons between electrons

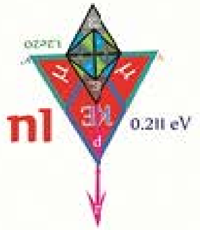
n1
EM
0.211 eV
1.2e20

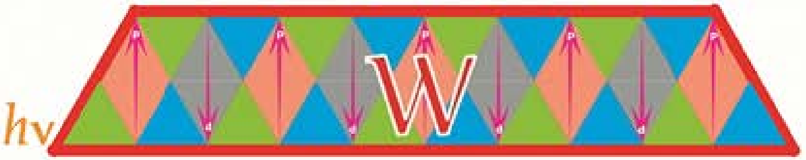

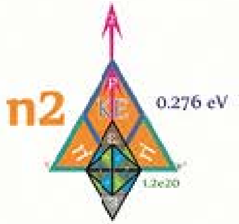
n2
KE
0.276 eV
1.2e20

KEM field energy momenta attracts opposite charges

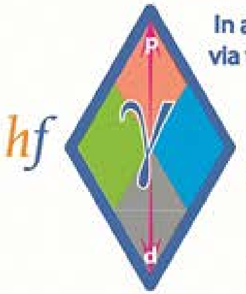

In accordance with Newton's 3rd Law Q2 can also act on Q1 via the superpositioning of energy fields and their momenta

This layered interaction of EM force quanta [EM charge carriers - Z bosons & Photons] is what constitutes the electromotive inductive force [emf]

The closer that Q1 and Q2 are, the greater the effect.

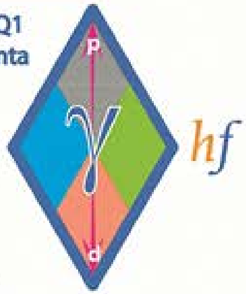

KEM field energy momenta repels same charge particles

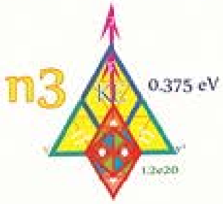
n3
KE
0.375 eV
1.2e20

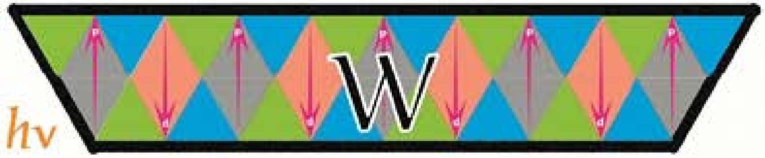

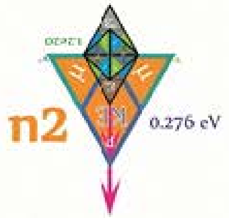
n2
EM
0.276 eV
1.2e20

Bosons

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge ElectroMagnetic mass velocity

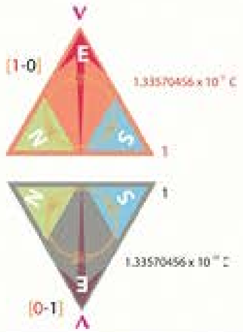
v
[1-0]
E
N
S
1.33570456 x 10 C
1
1
1.33570456 x 10 C
E
[0-1]
v

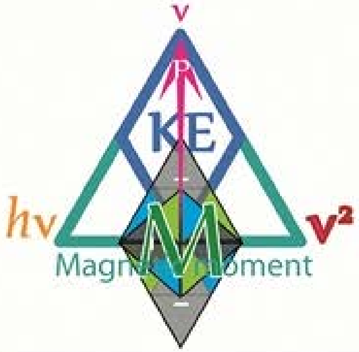
v
p
KE
hv
v²
M
Magnetic Moment

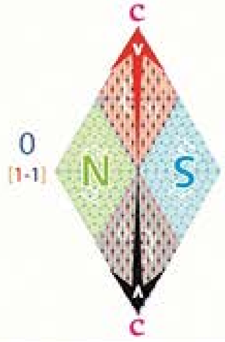

Photons

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

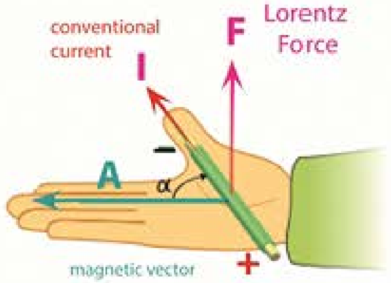
conventional current
I
F
Lorentz Force
A
α
−
+
magnetic vector

Lorentz Force

The magnetic force component of the Lorentz force manifests itself as the force that acts on a current-carrying wire in a magnetic field.

In that context, it is also called the Laplace force.

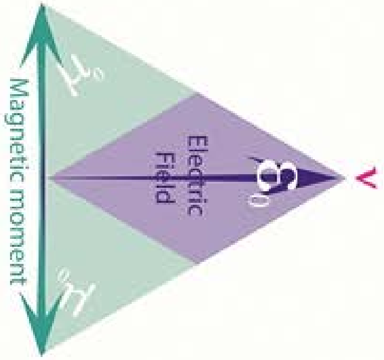
Magnetic moment
μ_0
Electric Field
ε_0
v

Hendrik Lorentz

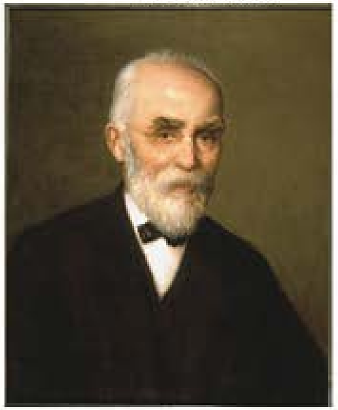

(18 July 1853 – 4 February 1928)

In physics, the Lorentz force is the force on a point charge due to electromagnetic fields.

It is given by the following equation in terms of the electric and magnetic fields

$$\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$$

LORENTZ force — *Electric force* — *charge velocity* — *Magnetic force*

Magnetic fields created by moving charges are perpendicular to the direction of motion (and can do NO work)

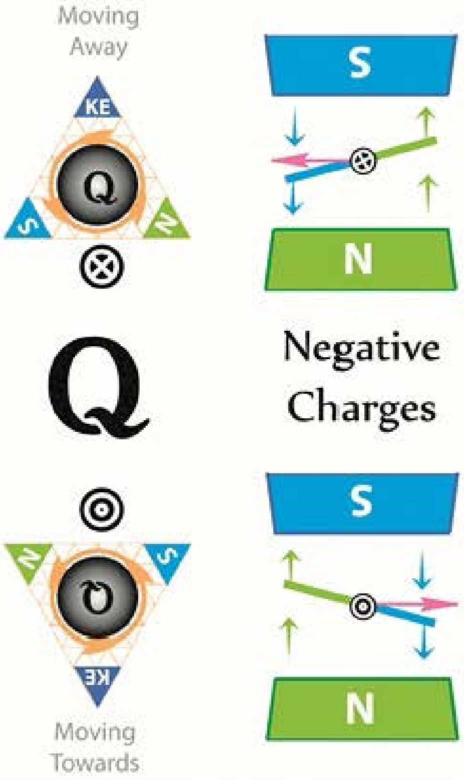
Moving Away
KE
Q
S
N
⊗
Q
Negative Charges
◎
KE
Moving Towards

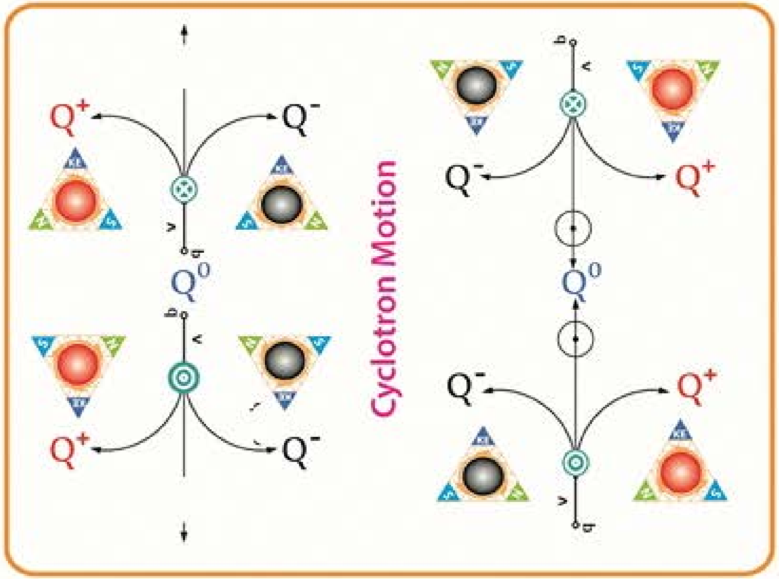
Q+
Q-
Q0
Q+
Q-
Cyclotron Motion
Q-
Q+
Q0
Q-
Q+
KE

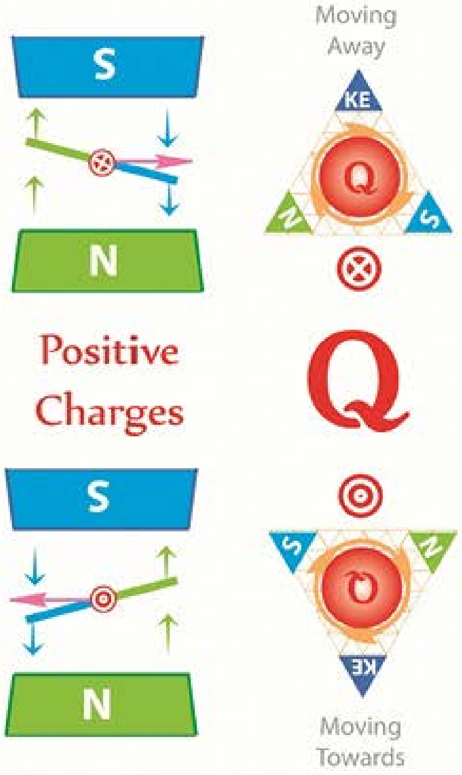
Moving Away
S
N
KE
Q
⊗
Positive Charges
Q
◎
KE
Moving Towards

Up out of page ◎ Magnetic Field ⊗ Down into page

Copyright ABRAHAM [2008] - all rights reserved

Work - Force - Energy

Work is the result of energy-momenta transfer (by means other than Matter-transfer), and produces changes in the second system's nett energy-momenta

Work

refers to an activity involving a force and movement in the directon of the force

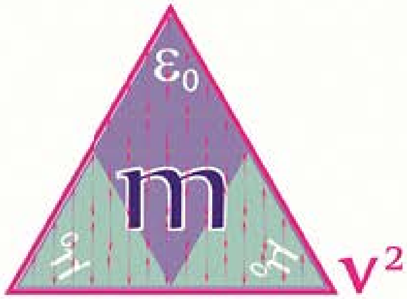
ε_0
m
μ_0
μ_0
v^2

Work

$$\boxed{W = F.d = ma.d = E}$$

The Planck constant is the quantum of Action

$$kg\,\frac{m}{s^2}\,m$$

Mechanical work is a scalar quantity that can be described as the product of a force times the distance through which it acts

[Force per meter]

Force

$$\boxed{ma = \frac{\Delta p}{\Delta t} = \Delta\left[\frac{mv}{t}\right]} \quad kg\,\frac{m}{s^2}$$

Momentum is the quantum of Force

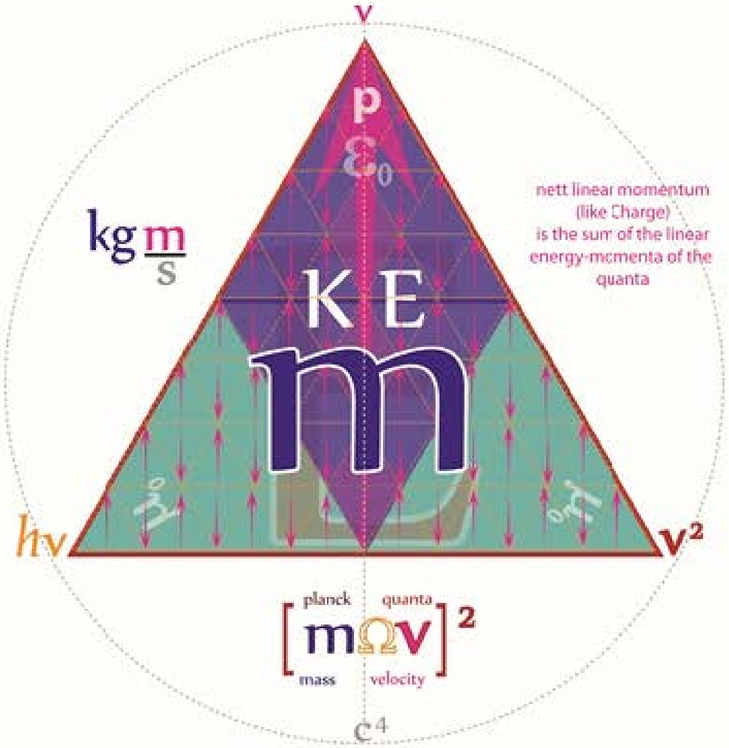
v
p
ε_0
$kg\,\frac{m}{s}$
K E
m
μ_0
μ_0
$h\nu$
v^2
planck
quanta
$[m\Omega v]^2$
mass
velocity
c^4

nett linear momentum (like Charge) is the sum of the linear energy-momenta of the quanta

Power is defined as the rate of using Energy or the rate of doing work

Energy

**is the scalar capacity to do work
You must have energy to accomplish work**

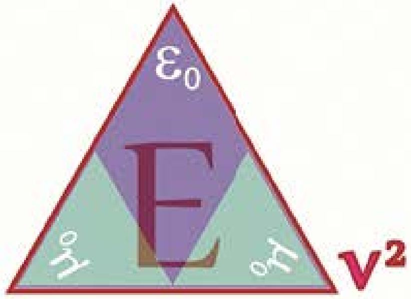
ε_0
E
μ_0
μ_0
v^2

Energy

$$\boxed{E = mv^2 = hv^2}$$

Energy is the square of Momentum

$$kg\,\frac{m^2}{s^2}$$

Energy is always equivalent to the ability to exert pulls or pushes against the basic forces of nature, along a path of a certain length.

[mass-energy momentum squared]

Watts

[Joules per second]

James Watt

(19 January 1736 – 25 August 1819)

WORK

Force per second

In Classical Mechanics

One watt is the rate at which work is done when an object's velocity is held constant at one meter per second against constant opposing force of one newton.

Work in physics is measured in Joules

Nm/s

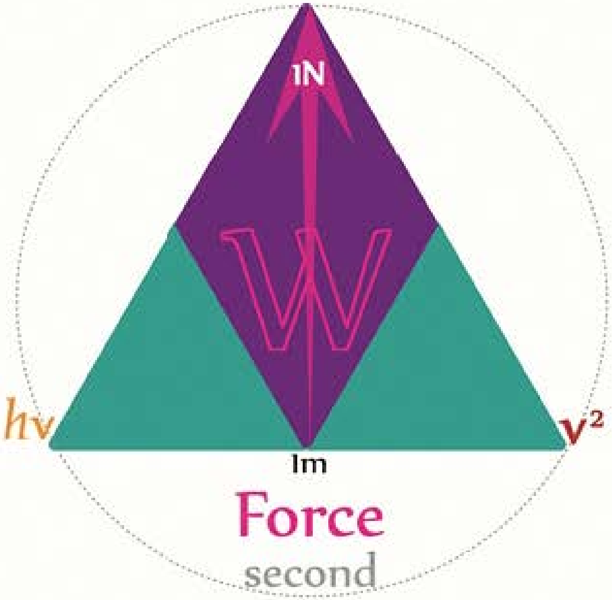

The watt second is a unit of energy, equal to the joule

$$W = \frac{J}{s} = \frac{N \cdot m}{s} = \frac{kg \cdot m^2}{s^3}$$

One Watt, defined as one joule per second, measures the rate of energy conversion.

$$W = V.A$$

Energy / second
[Joules/sec]

POWER

Energy per second

In ElectroDynamics

One watt is the rate at which work is done when one ampere (A) of current flows through an electrical potential difference of one volt (V).

Electrical power is measured in Watts

J/s

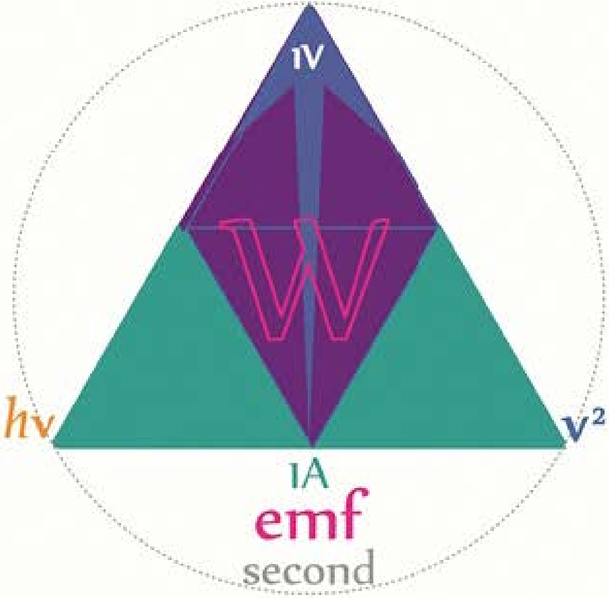

Copyright ABRAHAM [2008] - all rights reserved

The electromotive Force

The electromotive force, or most commonly emf (seldom capitalized), or (occasionally) electromotance is "that which tends to cause current (actual electrons and ions) to flow.

Electric Fields contain bi-directional energy-momenta

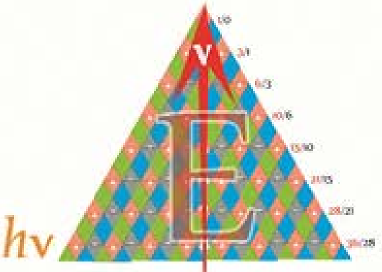

The electromotive Force is the nett force resulting from the separation of electrical charges (potential difference)

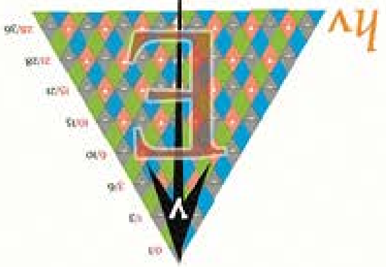

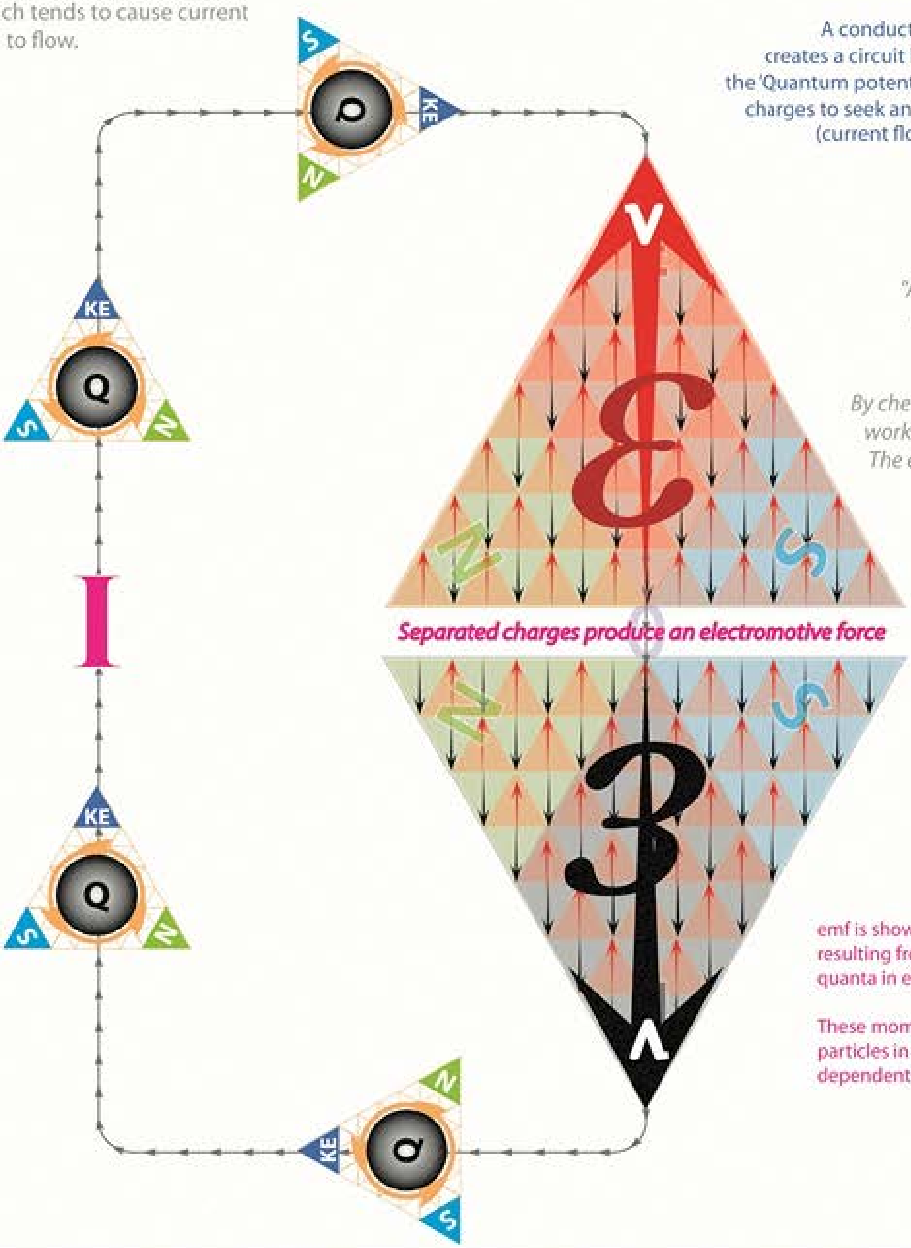

A conductor creates a circuit between the 'Quantum potentials' allowing charges to seek an eqilibrium (current flow)

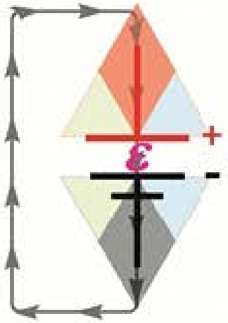

$$\mathcal{E} = -\frac{d\Phi_B}{dt},$$

"A source emf can be thought of as a kind of charge pump that acts to move charge from a point of high potential through its interior to a point of opposite potential. …

By chemical, mechanical or other means, the source of emf performs work dW on that charge to move it to the high potential terminal. The emf of the source is defined as the work dW done per charge dq: = dW/dq."

A Magnet stores charge quanta (bosons) in static Electro-Magnetic fields that in turn produce a Magnetic dipole

emf is shown to be the nett force resulting from the linear momenta quanta in every Electric 'voltage' field

These momenta accelerate charged particles in one of two directions dependent of the particle's nett charge

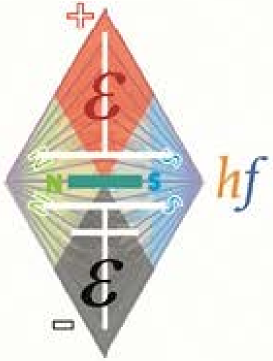

A Magnet can be viewed as a neutralised quantum emf source

All Charges seek Equilibrium

Copyright ABRAHAM [2008] - all rights reserved

Poynting Vectors

Poynting Field

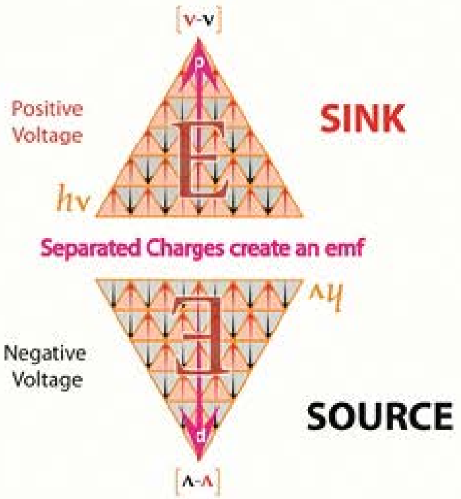

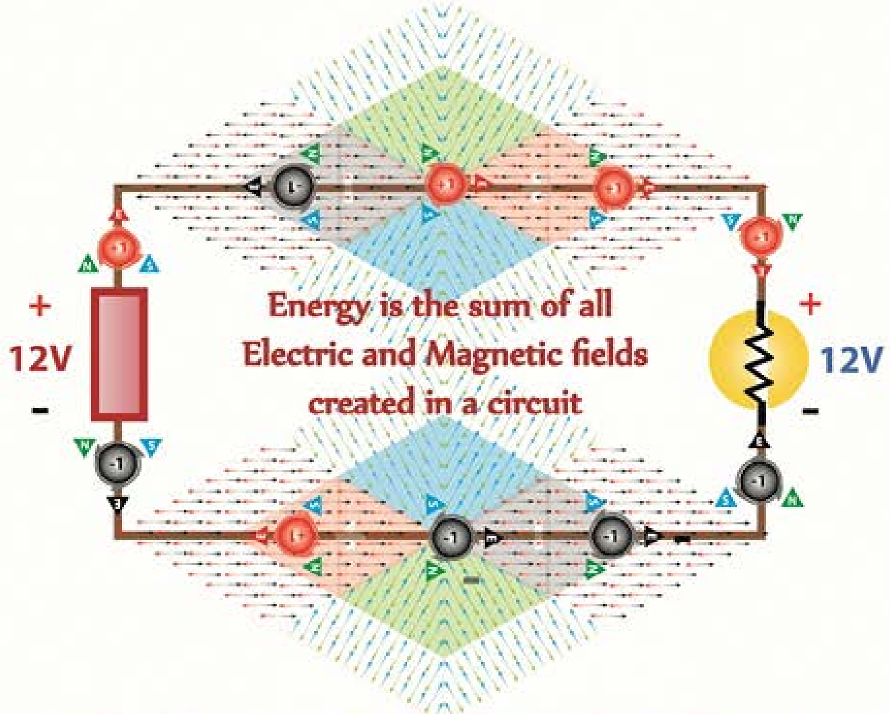

Energy is the sum of all Electric and Magnetic fields created in a circuit

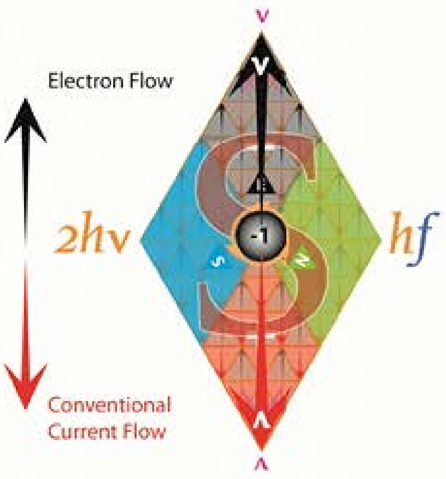

An test charge placed in an Electric field will experience an acceleration in a direction dependent on the sign of the charge

John Henry Poynting

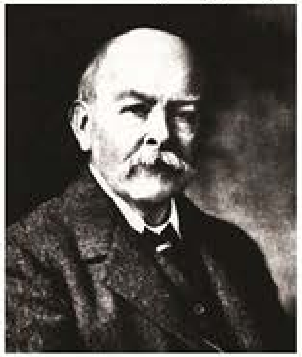

(9 September 1852 – 30 March 1914)

Electrical energy is transferred along a conductive path in an electrical circuit accelerating charged particles and ions in turn imbuing them with Kinetic Energy [voltage leads current]

There is an electric field running down the middle of the wire, which extends to just beyond its surface.

This electric field pushes the charges along against the resistance and adds to the electric field caused by the surface charges. The resultant electric field changes its direction around the circuit as the wires form a loop back to the battery.

Moving charged topologies inside the wire create a secondary magnetic field geometry around the wire

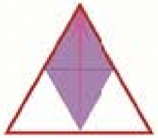

Electric fields

Energy

$$S = E \times H$$

momenta

Magnetic fields

Electrons form a conductive path for Energy to flow along

The Electron drift velocity is much less than the speed of electricity(c)

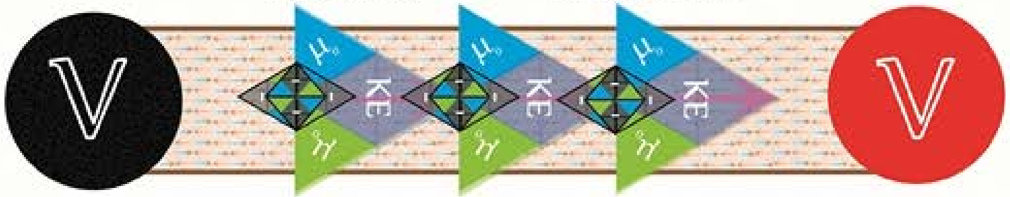

Despite the very low 'drift' velocity of the Electrons they generate a large Magnetic field as a result of their motion

Energy is transferred through empty space around (and NOT in) the wires of an electric circuit via an electromagnetic field called the Poynting field

It's important to remember that the current doesn't flow in both directions, only the energy does

Coulomb's Electric Interactions

$$k = \frac{1}{4\pi\varepsilon_0}$$

The proportionality constant ke, called the Coulomb constant (sometimes called the Coulomb force constant), is related to defined properties of linear EM energy momentum and is used to define Electric field forces

$$c = \frac{1}{\sqrt{\mu_0\varepsilon_0}}$$

$$8.987\text{ e9 } \frac{\text{Nm}^2}{\text{C}^2}$$

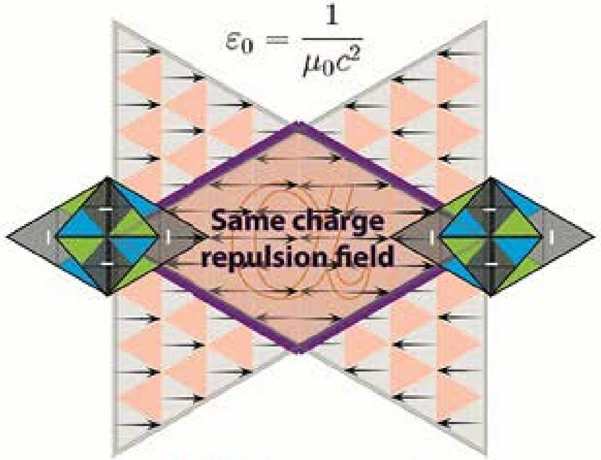

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

Similars repel

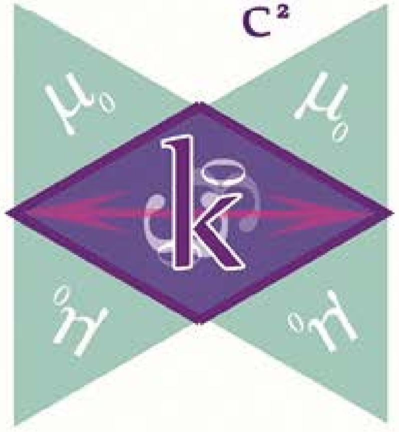

Opposites attract

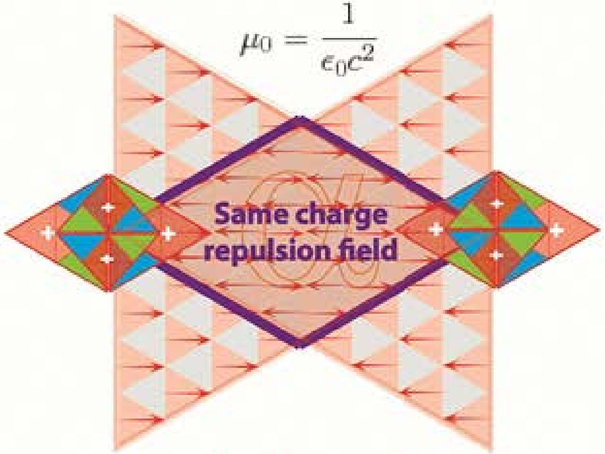

$$\mu_0 = \frac{1}{\epsilon_0 c^2}$$

Similars repel

Linear Coulombic force interactions
are a result of charged E field
linear momenta

Longitudinal E field forces
between Charged particles
are mediated by Photons

$$\mathbf{E} = \frac{\mathbf{F}}{q_t}$$

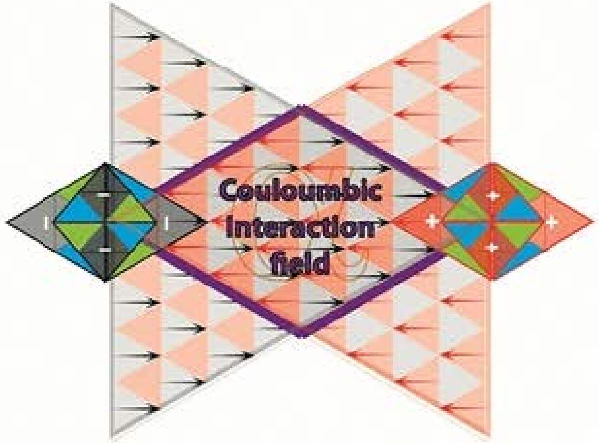

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}\hat{\mathbf{r}}$$

**The Electric field
can be defined by the
Force exerted by a Charge**

It is a measure of the interactive force
produced by the Electric field energy-momenta
of two superpositioned charge KEM fields

**The Electric field
can also be derived
from Coulomb's Law**

Copyright ABRAHAM [2008] - all rights reserved

Coulomb Force

Electric fields produce Coulombic forces

$$F = k\frac{Q_1 Q_2}{r^2}$$

It is defined as the nett Charge transported by a steady current of one Ampere in one second.

C

1 Coulomb of nett charge is comprised of

$(12 \times 1.335180067\ e\text{-}20)^{-1}$

6.240355408 e18 electrons

kg.s

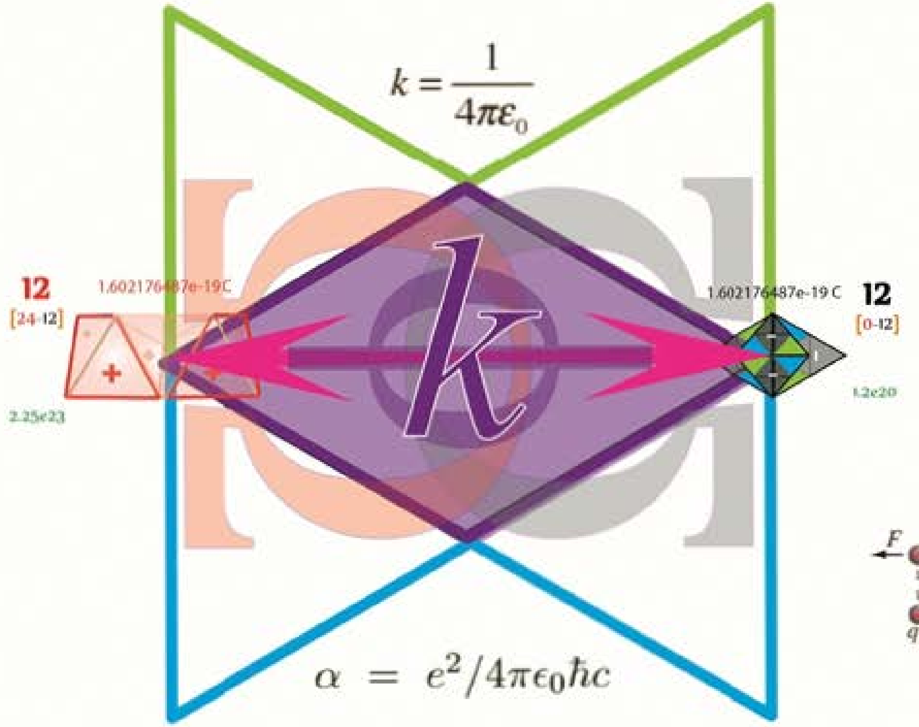

$k = \frac{1}{4\pi\varepsilon_0}$
12
[24-12]
1.602176487e-19C
+
2.25e23
k
1.602176487e-19 C
12
[0-12]
1.2e20
$\alpha = e^2/4\pi\epsilon_0\hbar c$

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Coulomb's Law is a law of physics describing the electric interaction between any two charged particles (and forms the basis for Ampere's Law)

F
q_1
q_2
F
Like charges repel
Unlike charges attract
q_1
F
F
q_2

$$F = \frac{kq_1q_2}{r^2} = \frac{q_1q_2}{4\pi\varepsilon_0 r^2}$$ *Coulomb's Law*

The interactive linear momenta of super-positioned E-fields creates Coulombic forces

Coulombs

As

$1\text{C} = 1\text{A} \cdot 1\text{s}$

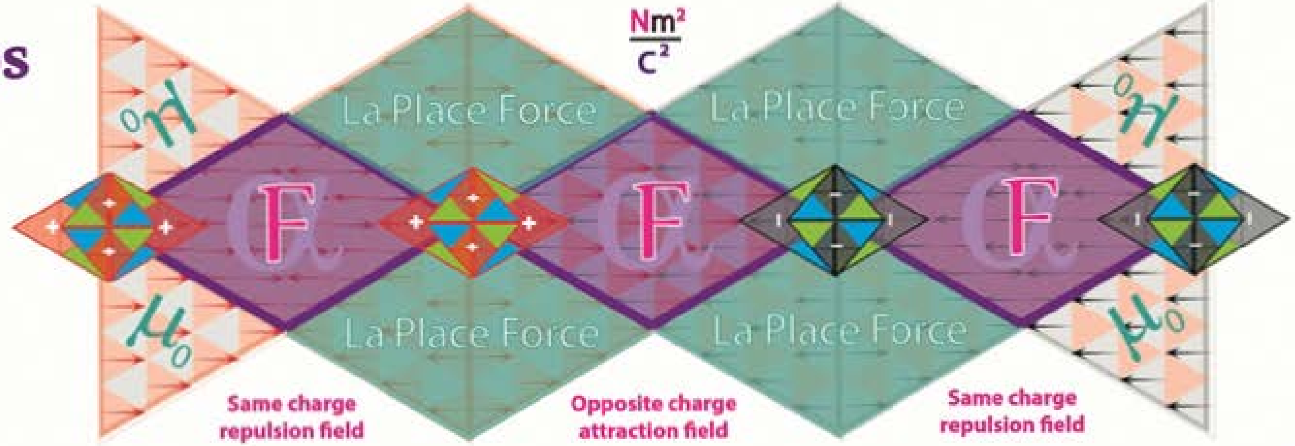

$\frac{\text{Nm}^2}{\text{C}^2}$
ε_0
μ_0
La Place Force
La Place Force
La Place Force
La Place Force
F
F
F
α
α
α
ε_0
μ_0
Same charge repulsion field
Opposite charge attraction field
Same charge repulsion field

Amperes

$$\frac{\text{C}}{\text{s}}$$

$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$

Copyright ABRAHAM [2008] - all rights reserved

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Coulomb force

$$k_c \frac{Q_1 Q_2}{r^2}$$

As

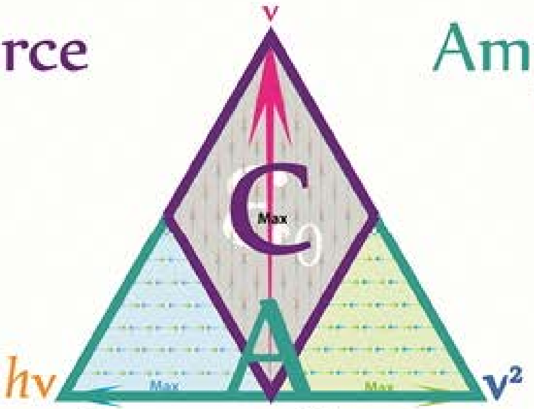
v
C
Max
A
hv
Max
Max
v²

Ampere force

$$2k_A \frac{I_1 I_2}{r}$$

$$\frac{C}{s}$$

Andre Ampere

(20 January 1775 – 10 June 1836)

Lorentz Forces

are the forces on a point charge due to external electromagnetic fields

$$\vec{F} = q[\vec{E} + (\vec{v} \times B)]$$

electric force

velocity

magnetic forces do not accelerate charged particles

Linear acceleration Electric force between charged particles

Q

Same charges

Q

Opposite charges

Q

Same charges

Q

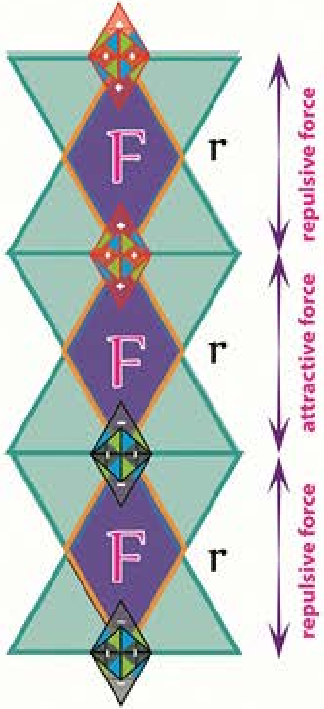
F
r
repulsive force
F
r
attractive force
F
r
repulsive force

I

The amount of charged mass-Matter topologies moving passing a point per unit time

Current

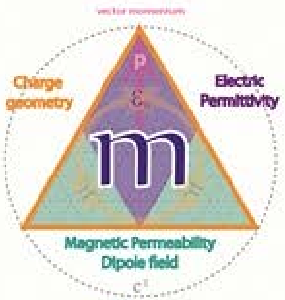
vector momentum
P
ε
Charge geometry
Electric Permittivity
m
Magnetic Permeability Dipole field
c²

Charge

The angular mass momenta possessed by a steady current of one ampere in one second.

Q

Transverse perturbative Magnetic force between charged particles

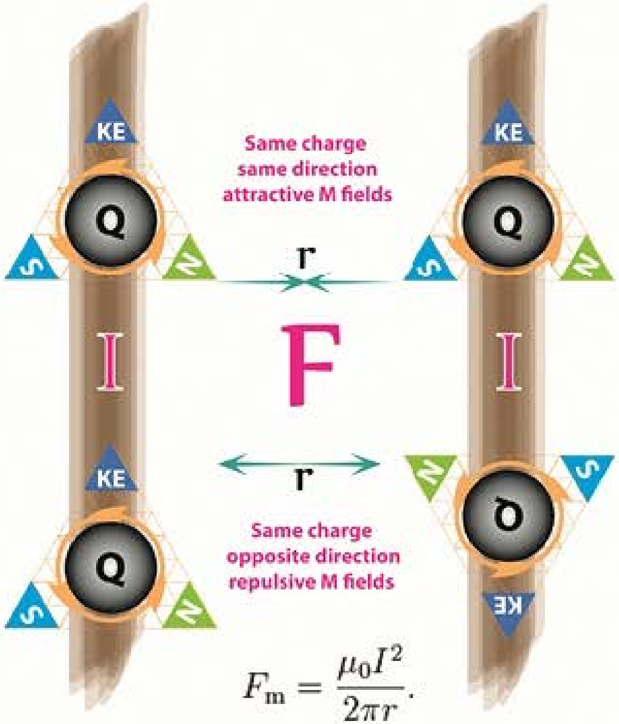
KE
Q
S
N
I
KE
Q
S
N
r
F
r
KE
Q
N
S
I
KE
Q
S
N

Same charge same direction attractive M fields

Same charge opposite direction repulsive M fields

$$F_m = \frac{\mu_0 I^2}{2\pi r}.$$

Copyright ABRAHAM [2008] - all rights reserved

Capacitance

Any two electrical conductors separated by a non-conducting (or very high resistance) medium is a capacitor
[these can be plates, conductive wires or coaxial cables etc]

Capacitance is a function only of the physical dimensions (geometry) of conductors and the permittivity of the dielectric separating them

Plate capacitors

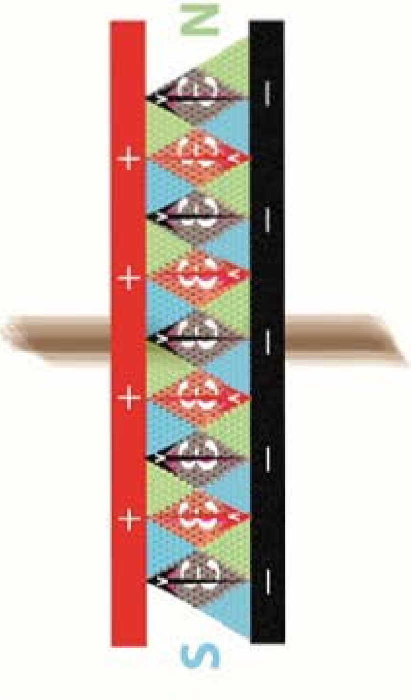

separated conductive cables

In general any separated charges will create a emf voltage

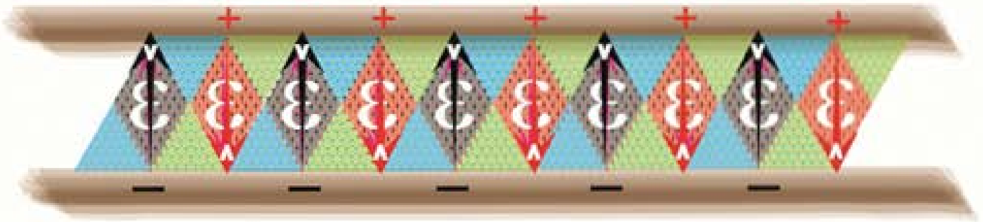

should a current flow in the conductors the capacitive field will be dominated by perpendicular, superpositioned amperian M-fields of greater strength

Coaxial cables

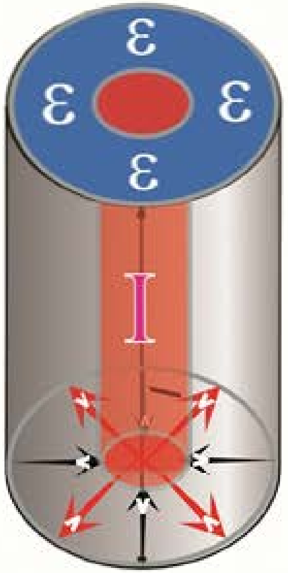

The SI unit of capacitance is the farad (symbol: F)
named after the English physicist Michael Faraday;

Capacitance is the ability of a body to store an electrical charge.

$$C = \frac{Q}{V}$$

Any object that can be electrically charged exhibits capacitance

A 1 farad capacitor when charged with1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

Copyright ABRAHAM [2008] - all rights reserved

Amperian Forces

Currents produce magnetic field forces

$$F = 2k_A \frac{I_1 I_2}{r}$$

6.241335 e18 electrons passing a given point per second constitutes one Ampere.

Coulombs are charged masses

m

that produce accelerations

Andre Ampere

(20 January 1775 – 10 June 1836)

Current is charged Matter in motion

M

that produce magnetic fields

A

1 Amp of electrons has a EM mass of

6.241355408 e18 x 8.851486361 e-31

5.524527227 e-12 kg

kg

An Ampere is a Coulomb per second.
a measure of the rate at which charged Matter moves.

A coulomb is the quantity of charge transported by one amp in one second

C

$1C = 1A \cdot 1s$

As

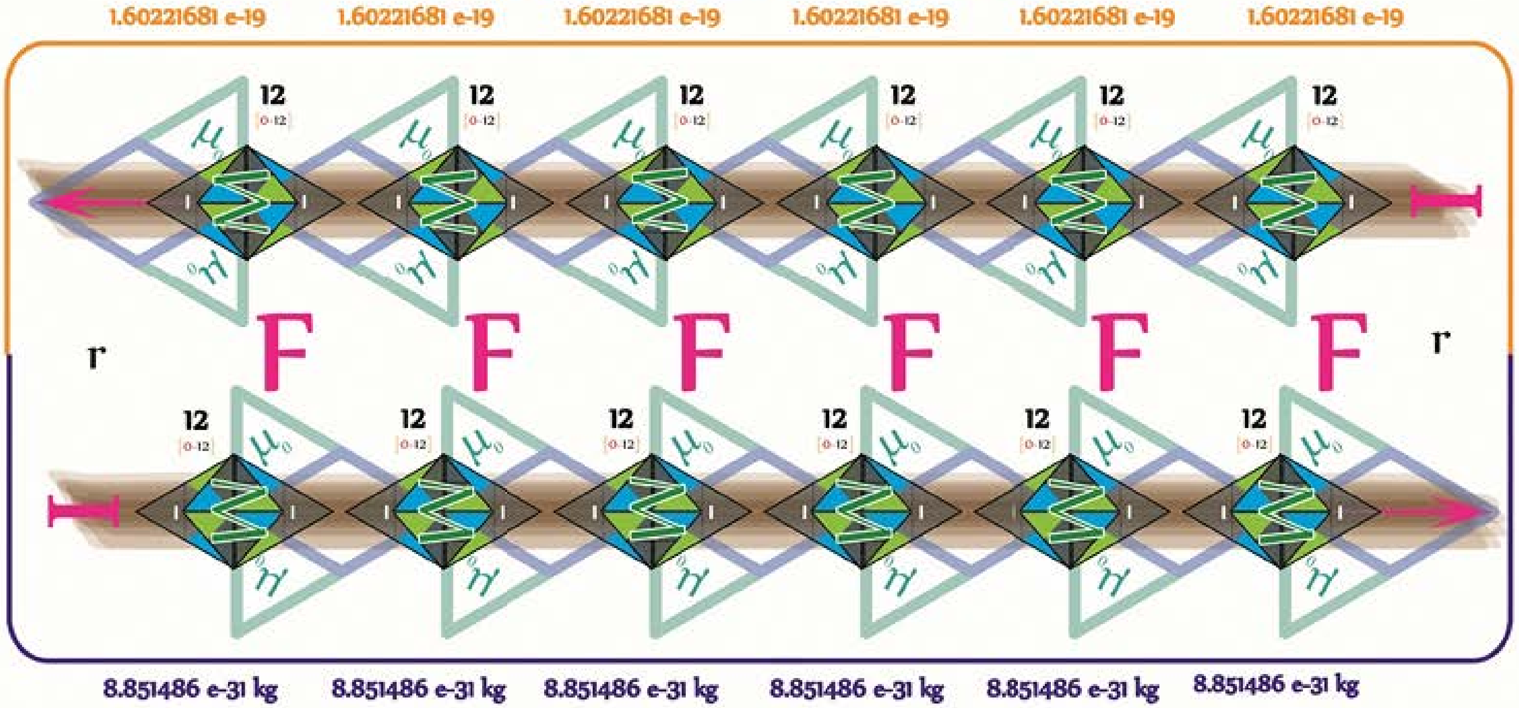
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
1.60221681 e-19
12
[0-12]
μ0
ε0
F
F
F
F
F
F
r
r
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg
8.851486 e-31 kg

A

$1\,A = 1\frac{C}{s}$

$\frac{C}{S}$

An ampere is one coulomb per second

Copyright ABRAHAM [2008] - all rights reserved

Ampere's Magnetic Interactions

the force of attraction or repulsion between two current-carrying wires (see first figure below) is often called Ampère's force law.

The best-known and simplest example of Ampère's force law, which underlies the definition of the ampere, the SI unit of current, states that the force per unit length between two straight parallel conductors is

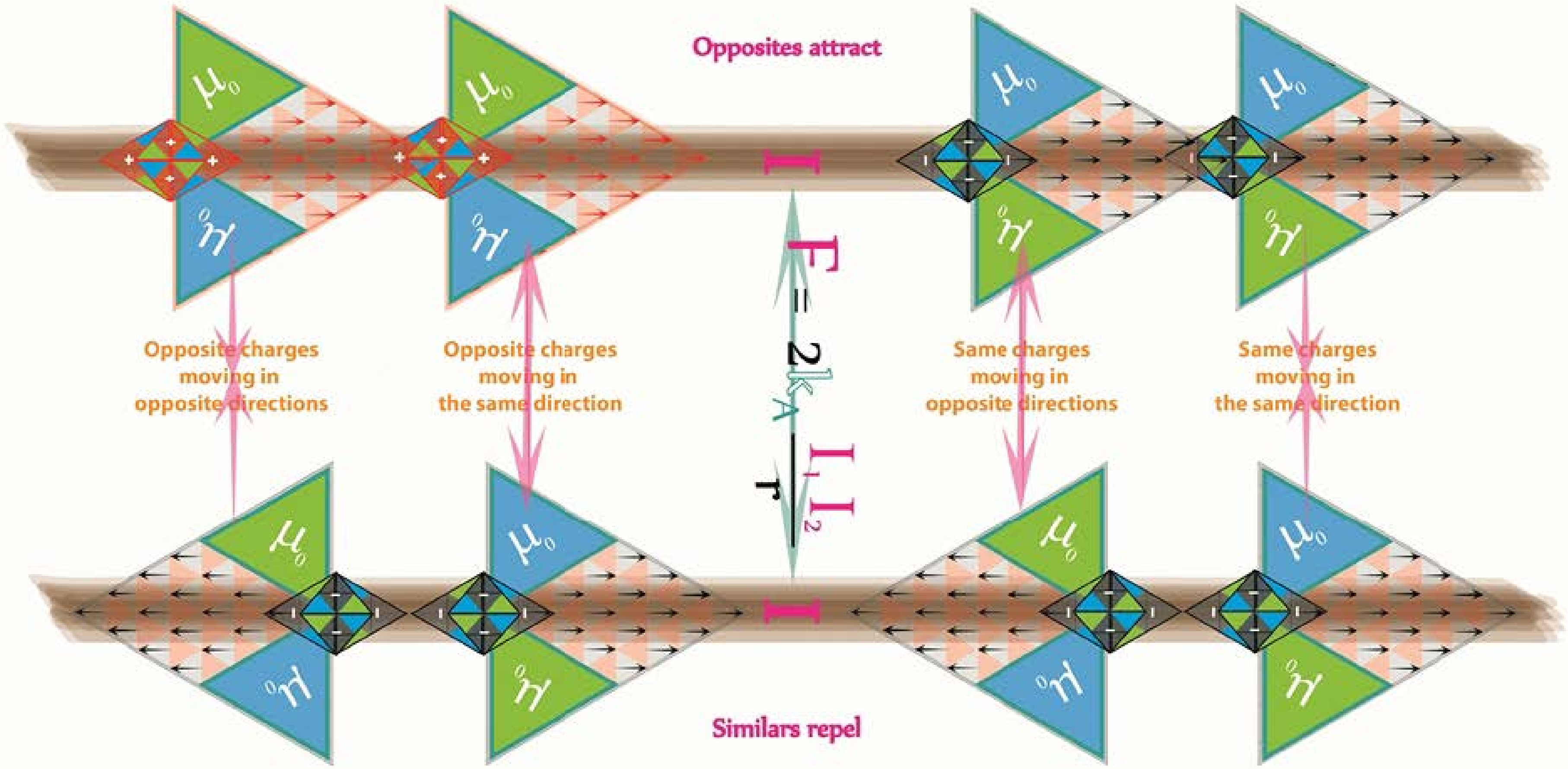
Opposites attract
μ_0
I
Opposite charges moving in opposite directions
Opposite charges moving in the same direction
Same charges moving in opposite directions
Same charges moving in the same direction
Similars repel

$$F = 2k_A \frac{I_1 I_2}{r}$$

It is a measure of the interactive force produced by the Magnetic field energy-momenta of two superpositioned KEM field geometries

2D Electric field geometries accelerate charged particles

Lorentz Forces

If a particle of charge q moves with velocity v in the presence of an electric field E and a magnetic field B, then it will experience a force

3D Matter topologies in motion create magnetic moments

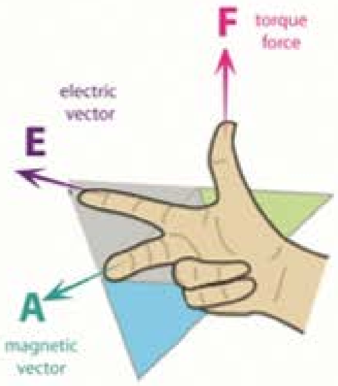
F
torque force
electric vector
E
A
magnetic vector

charged particles in motion produce KEM fields with magnetic moments

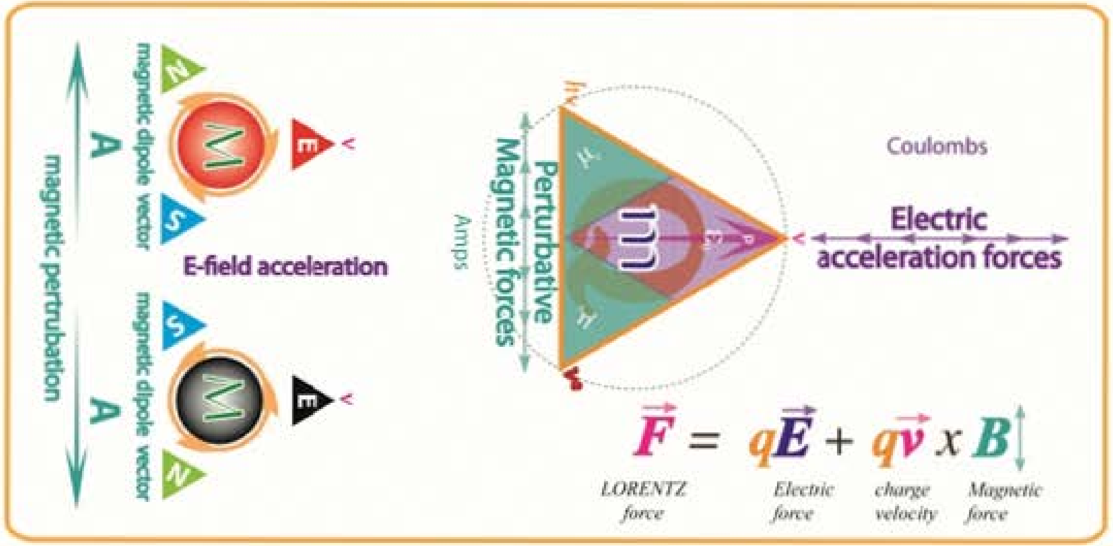
magnetic pertrubation
A
magnetic dipole vector
N
M
E
S
E-field acceleration
S
M
E
N
A
magnetic dipole vector
Amps
Magnetic forces
Perturbative
m
Coulombs
Electric acceleration forces

$$\vec{F} = q\vec{E} + q\vec{v} \times B$$

LORENTZ force — *Electric force* — *charge velocity* — *Magnetic force*

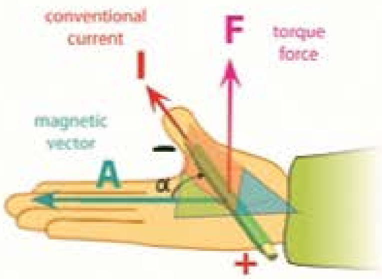
conventional current
I
F
torque force
magnetic vector
A
α
−
+

external magnetic fields create torque forces on the M dipoles of KEM fields

The Lorentz force is the force on a charged particle due to external electromagnetic fields.

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Magnetic field upwards through paper ⊙

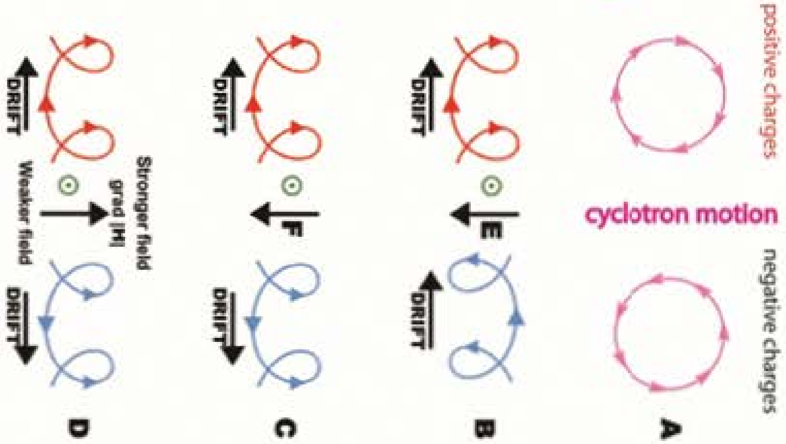
DRIFT
Weaker field
Stronger field grad |M|
DRIFT
D
DRIFT
F
DRIFT
C
DRIFT
E
DRIFT
B
positive charges
cyclotron motion
negative charges
A

All charges in motion are subject to external EM forces on their KEM fields

Andre Ampere

(20 January 1775 – 10 June 1836)

Neutral Particle interactions

Opposites Attract

Classical Forces of interaction
are unable to explain
how Neutral particles
are attracted to other particles

Similars Repel

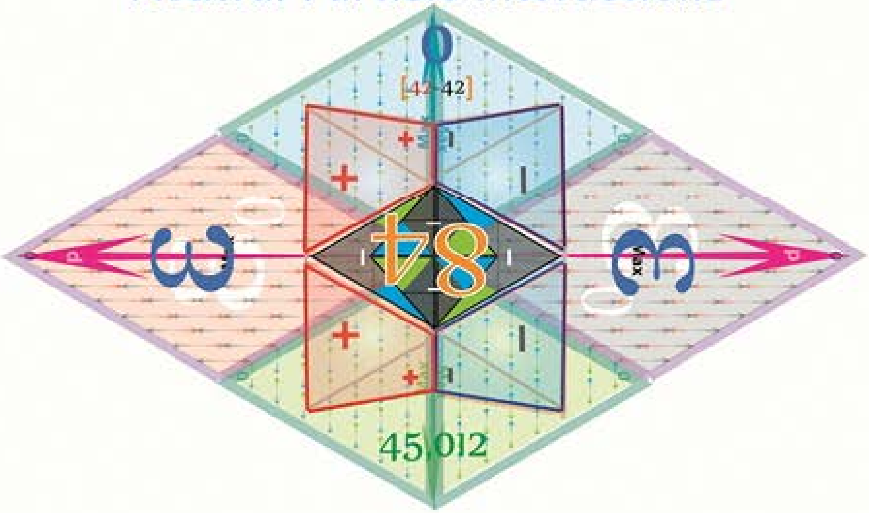
0
[42-42]
+
-
ε
84
ε
+
-
45,012

Tetryonic charge geometry

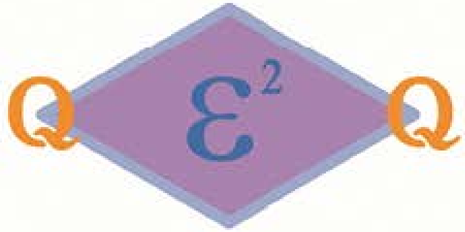
Q
ε^2
Q

reveals the equilateral field mechanics behind
EM induction and all the physical forces

All Matter topologies are comprised of charged mass-energy quanta
that effect atomic interactions via their KEM field energy momenta

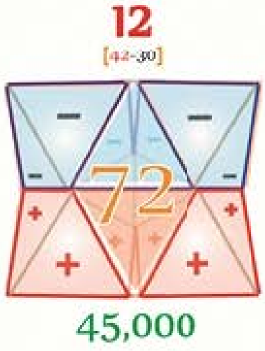
12
[42-30]
-
-
72
+
+
45,000

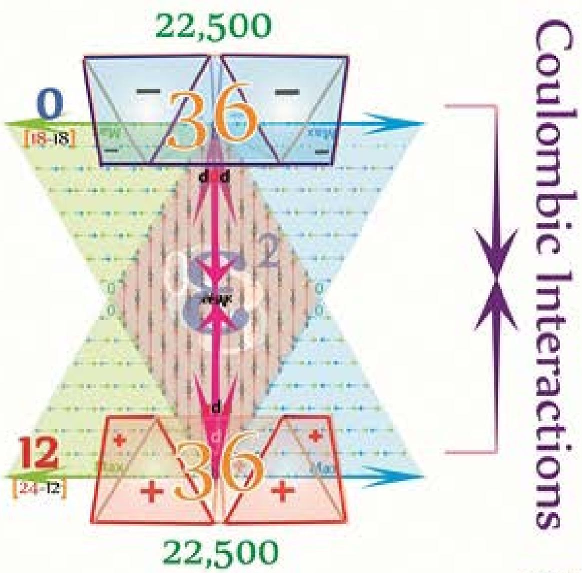
22,500
0
[18-18]
36
ε^2
12
[24-12]
36
22,500
Coulombic Interactions

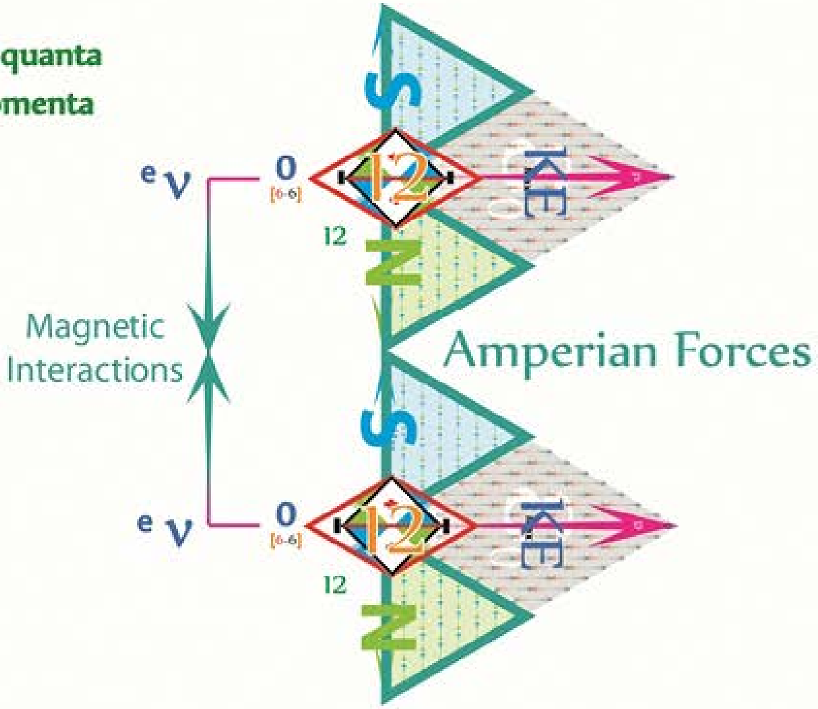
ev
0
[6-6]
12
12
KE
S
N
Magnetic Interactions
Amperian Forces
ev
0
[6-6]
12
12
KE
S
N

All Matter in motion produces a secondary interactive KEM field

Copyright ABRAHAM [2008] - all rights reserved

Charged mass-Matter

All particles and fields have charged mass-energy geometries

Coulombs

C

kg.s

mass seconds

$$1\text{C} = 1\text{A} \cdot 1\text{s}$$

An elementary charge is 1.602216081 e-19 C

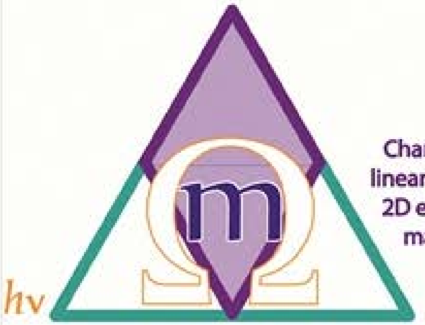

Charge is measured by the linear forces produced by the 2D equilateral geometry of mass-energy momenta

2D kEM angular mass-energy momenta

It is defined as the charge provided by a steady current of one ampere in one second.

Charge

Q

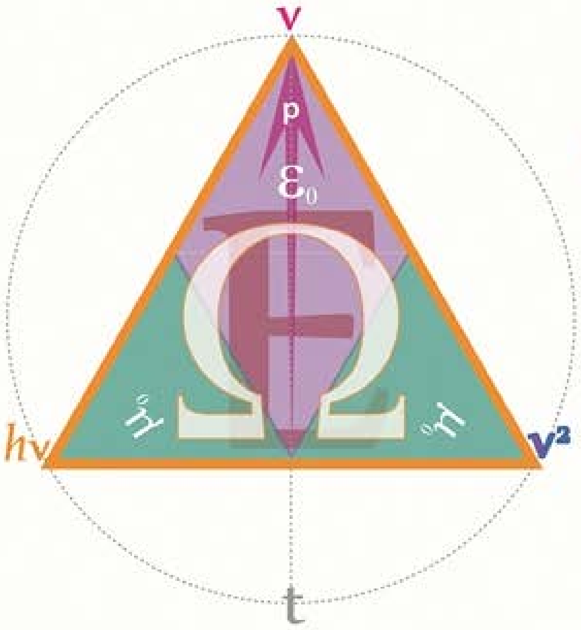

CHARGE

One coulomb of charge flowing per second equals one ampere of current.

CURRENT

$$Q = I \,.\, t$$

Charge and Current are related to each other through time

Amperes

A

KG

mass-Matter per second

$$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$$

6.241355408 e18 electrons /sec constitutes one ampere.

Current is measured by the transverse forces produced by mass-energy momenta of Matter topologies in motion per unit of time

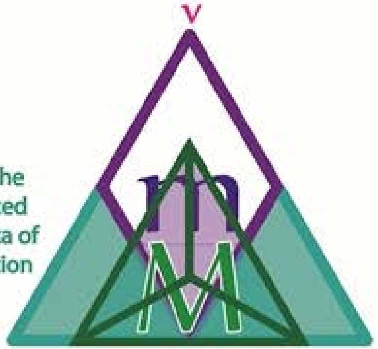

3D mass-Matter momentum

The amount of charged mass- Matter passing a point per unit time

Current

I

Copyright ABRAHAM [2008] - all rights reserved

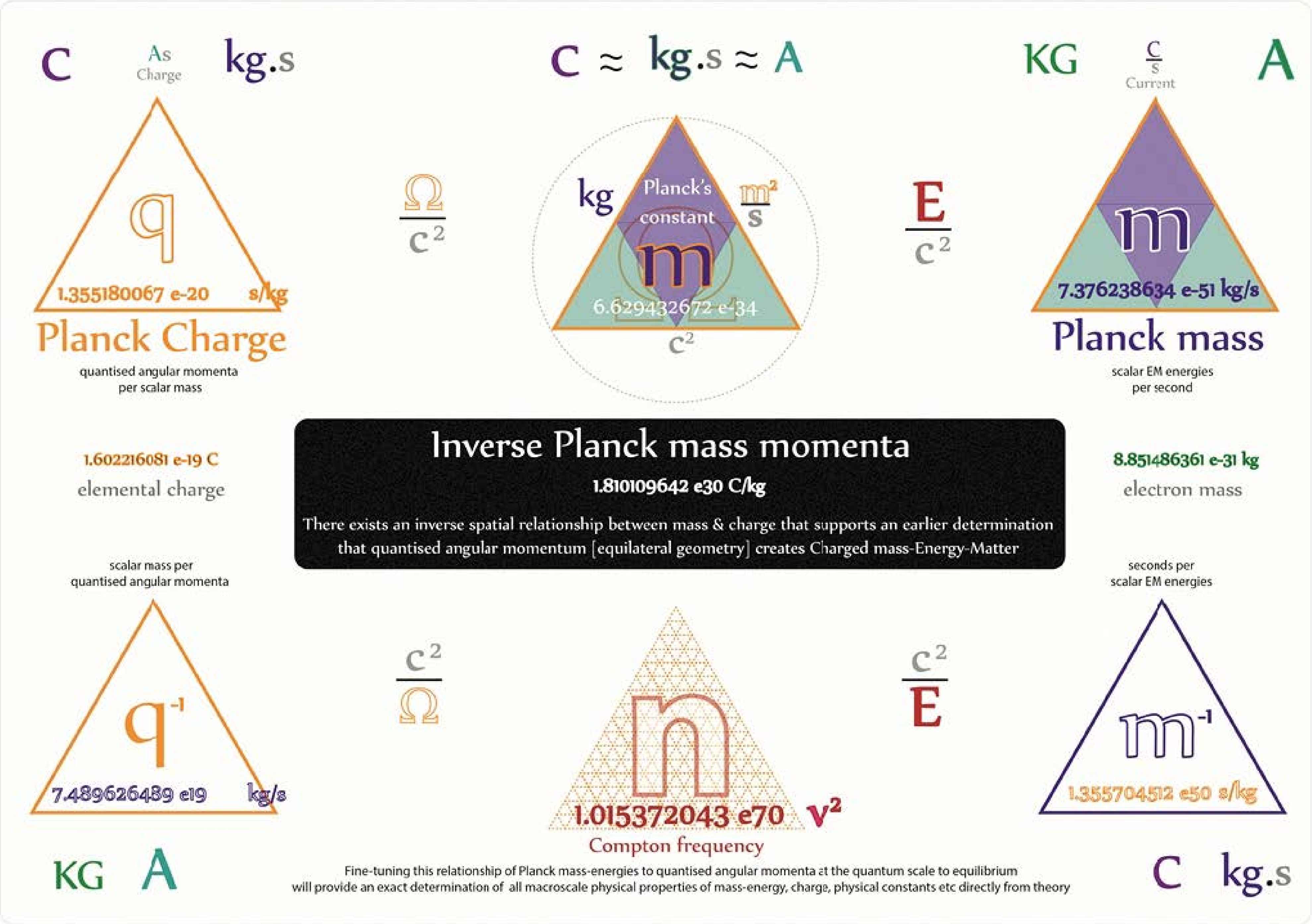

C As Charge kg.s

C ≈ kg.s ≈ A

KG $\frac{C}{s}$ Current A

q

1.355180067 e-20 s/kg

Planck Charge

quantised angular momenta
per scalar mass

$\frac{\Omega}{c^2}$

kg Planck's constant $\frac{m^2}{s}$

m

6.629432672 e-34

c^2

$\frac{E}{c^2}$

m

7.376238634 e-51 kg/s

Planck mass

scalar EM energies
per second

1.602216081 e-19 C

elemental charge

Inverse Planck mass momenta

1.810109642 e30 C/kg

There exists an inverse spatial relationship between mass & charge that supports an earlier determination that quantised angular momentum [equilateral geometry] creates Charged mass-Energy-Matter

8.851486361 e-31 kg

electron mass

scalar mass per
quantised angular momenta

seconds per
scalar EM energies

q^{-1}

7.489626489 e19 kg/s

$\frac{c^2}{\Omega}$

n

1.015372043 e70 v^2

Compton frequency

$\frac{c^2}{E}$

m^{-1}

1.355704512 e50 s/kg

KG A

C kg.s

Fine-tuning this relationship of Planck mass-energies to quantised angular momenta at the quantum scale to equilibrium
will provide an exact determination of all macroscale physical properties of mass-energy, charge, physical constants etc directly from theory

Tetryonics 22.13 - Charge-coulombs-amps-Current

Copyright ABRAHAM [2008] - all rights reserved

Electricity

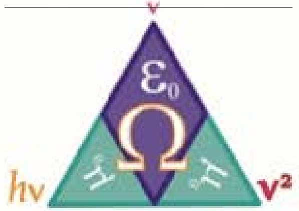

"There exists Negative and Positive electrical quanta"

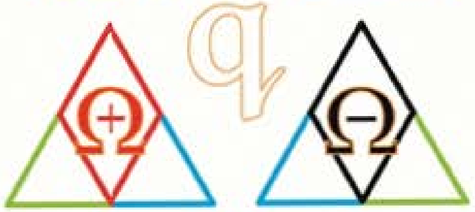

Electromagnetism: a fundamental interaction between the magnetic field and the presence and motion of an electric charge topology.

Electric charge: the geometry of EM energy momenta, also determines their electromagnetic interactions.

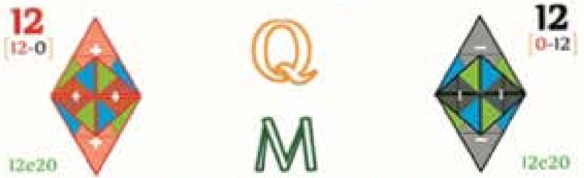

Electrically charged Matter topologies are influenced by, and produce, electromagnetic fields when in motion

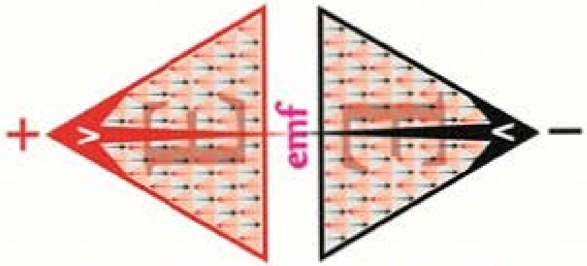

Electric potential: the capacity of an electric field to do work on an electric charge, typically measured in volts.

Benjamin Franklin

(January 17, 1706 – April 17, 1790)

Charles-Augustin_de_Coulomb
(14 September 1736 – 23 August 1806)

Andre Ampere
(20 January 1775 – 10 June 1836)

Georg Simon Ohm
(16 March 1789 – 6 July 1854)

Carl Friedrich Gauss
(30 April 1777 – 23 February 1855)

Heinrich Lenz
(February 12, 1804 – February 10, 1865)

Michael Faraday
(22 September 1791 – 25 August 1867)

James Clerk Maxwell
(13 June 1831 – 5 November 1879)

James Prescott Joule
(24 December 1818 – 11 October 1889)

John Henry Poynting
(9 September 1852 – 30 March 1914)

Nikola Tesla
(10 July 1856 – 7 January 1943)

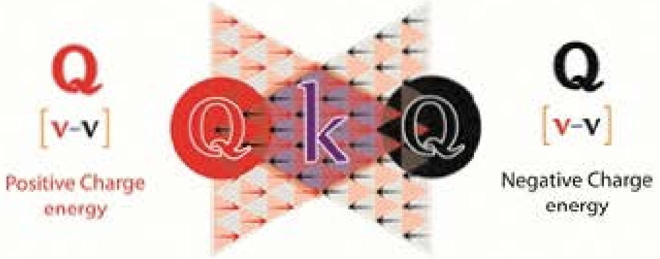

Electric field: an influence produced by one electric charge on other charges in its vicinity.

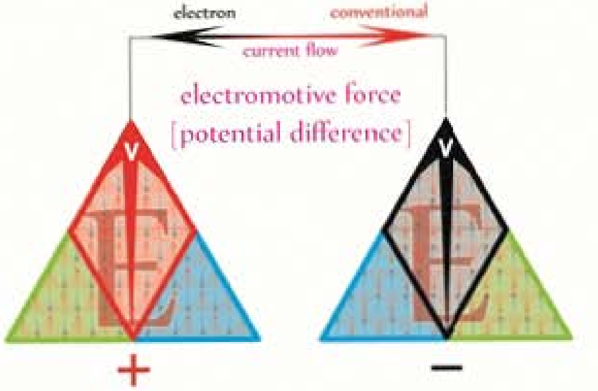

Electricity is the flow of Energy between separated Charge potentials when an electric circuit is formed measured in Volts [Joules/C]

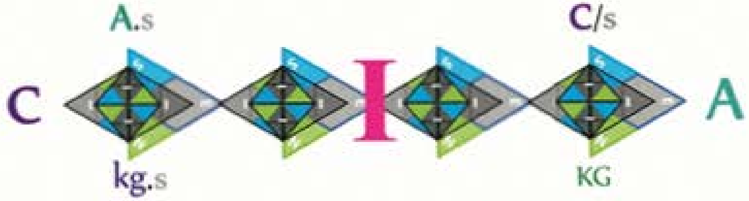

Electric current: is a secondary effect resulting from electrical energy in a circuit It is the movement or flow of electrically charged particles, typically measured in amperes.

Electricity is a general term encompassing a variety of phenomena resulting from the presence and flow of electric charge.
These include many easily recognizable phenomena, such as lightning, static electricity, and the flow of electrical current in an electrical wire.
In addition, electricity encompasses less familiar concepts such as the electromagnetic field and electromagnetic induction.

EM field Permittivity

The Electric constant, commonly called the vacuum permittivity, or permittivity of free space, relates the units for electric charge to mechanical quantities such as length and force.

The name Vacuum Permittivty is a misnomer and should be replaced with the correct term EM field Permittivity

The strength of Electric fields is determined by the Electrical Permittivity Constant

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

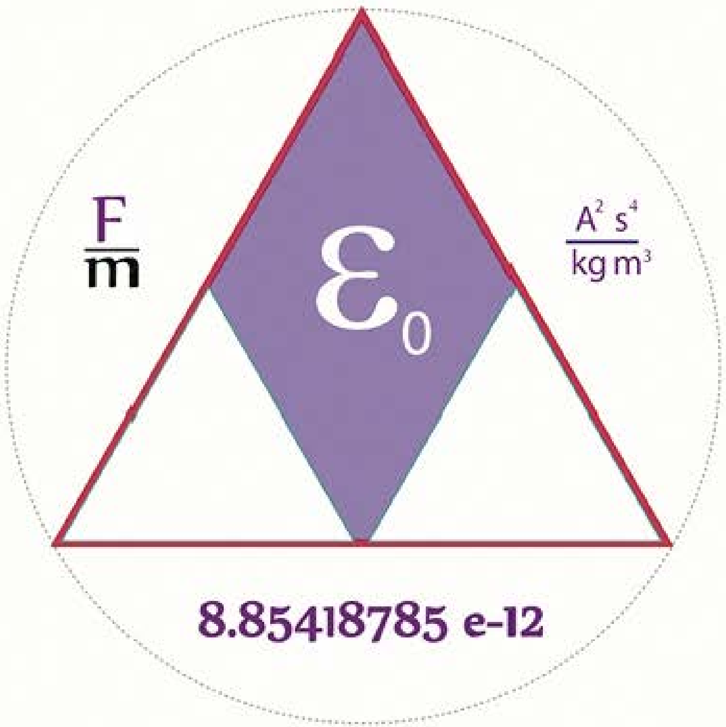

8.85418785 e-12

The permittivity of empty space, equal to 1 in centimeter-gram-second electrostatic units and to 107/4πc2 farads per meter or, numerically, to 8.854 × 10-12 farad per meter in International System units, where c is the speed of light in meters per second.

Gauss' Law:
"The total of the electric flux out of a closed surface is equal to the charge enclosed divided by the permittivity"

This applies equally to any integral geometry chosen to tessellate a surface area

Superpostioned E fields gives rise to Coulomb Forces

$$\varepsilon_0 = \frac{1}{\mu_0 c^2}$$

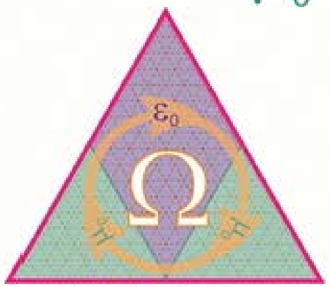

"Ampere's Law states that for any closed loop path, the sum of the quantities (B.ds) for all path elements into which the complete loop has been divided is equal to the product of µ0 and the total current enclosed by the loop."

$$k = \frac{1}{4\pi\varepsilon_0}$$

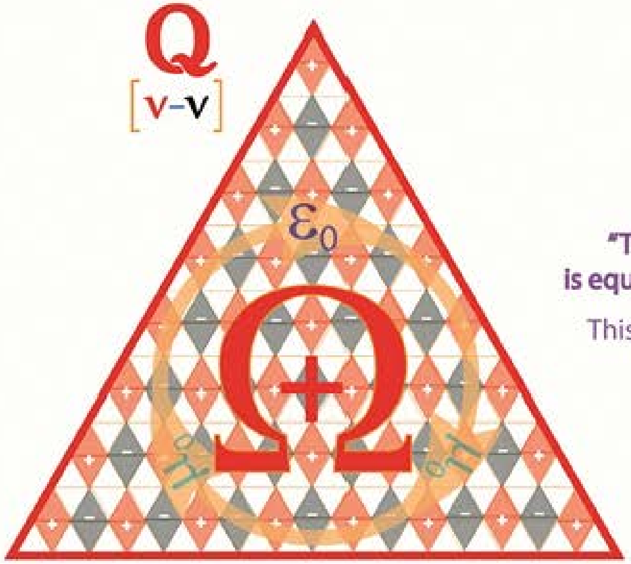

Positive Charge Electric Field

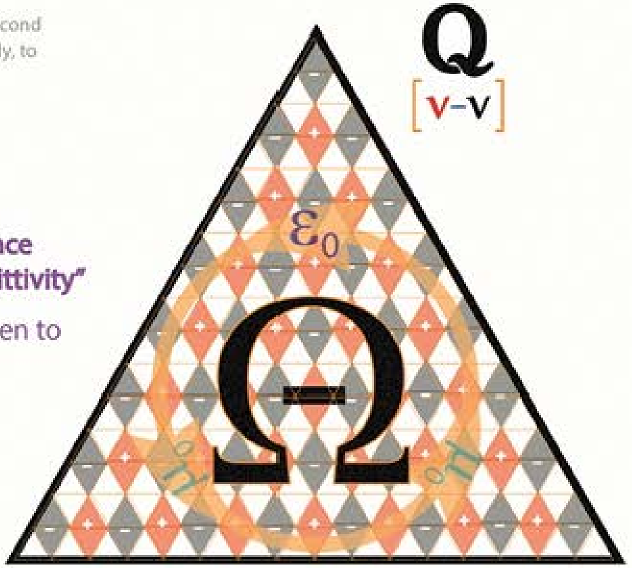

Negative Charge Electric Field

Copyright ABRAHAM [2008] - all rights reserved

EM field Permeability

The permeability of free space,
also called absolute permeability.

The name Vacuum Permeability is a misnomer and should be replaced with the correct term EM field Permeability

The magnetic constant has the value of
4π × 10-7 henry per meter.

The strength of Magnetic fields is determined by the Magnetic Permeability Constant

$$\mu_0$$

The magnetic field is most commonly defined in terms of the Lorentz force it exerts on moving electric charges.

$$F$$

The magnetic field generated by a steady current (a constant flow of electric charges in which charge is neither accumulating nor depleting at any point) as described by the Biot–Savart law

$$B=\mu_0 H$$

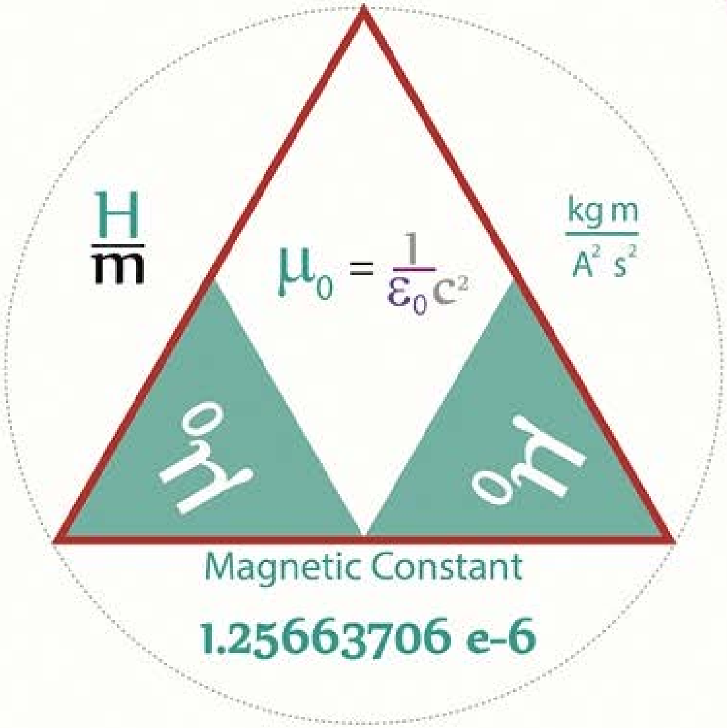

$$\frac{H}{m}$$

$$\mu_0 = \frac{1}{\varepsilon_0 c^2}$$

$$\frac{kg\,m}{A^2\,s^2}$$

Magnetic Constant

1.25663706 e-6

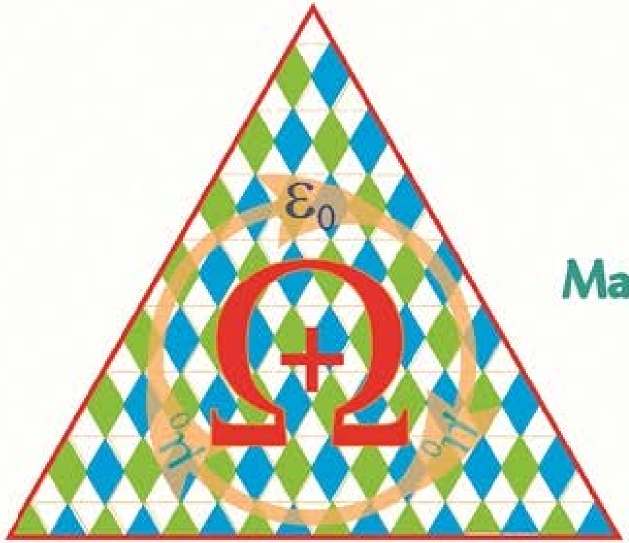

Positive Magnetic Moment

Magnetic monopoles do NOT exist

they are a mathematical identity
and not physically possible due to
the equilateral geometry
of Planck energies

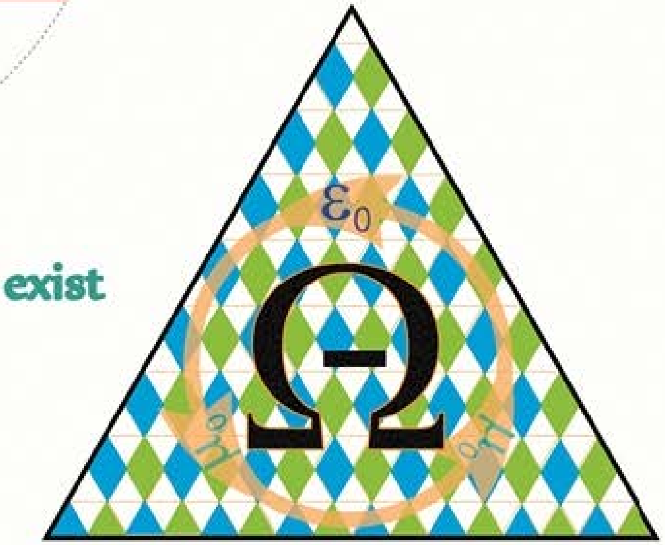

Negative Magnetic Moment

James Prescott Joule

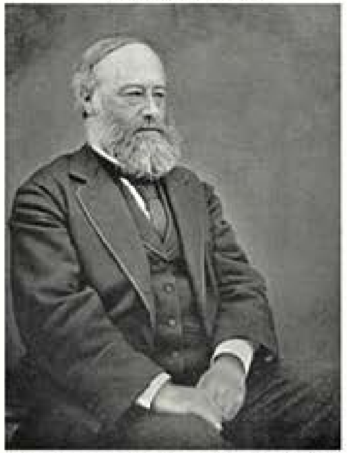

(24 December 1818 – 11 October 1889)

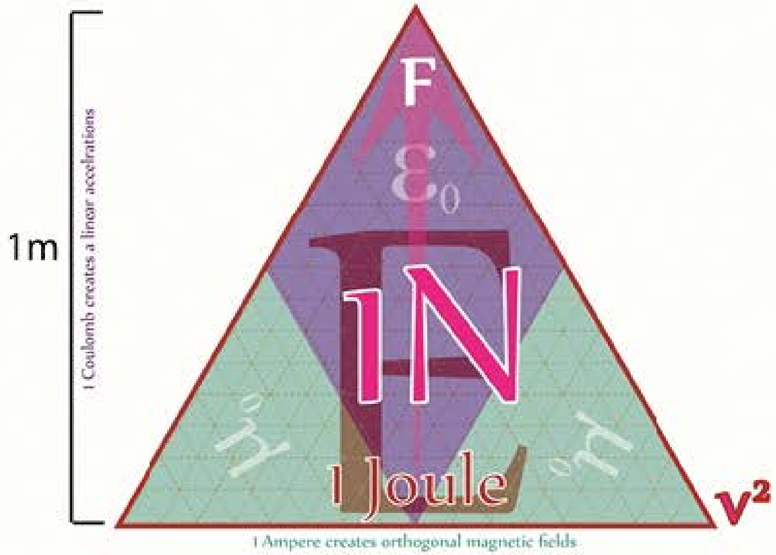
1m
1 Coulomb creates a linear accelerations
F
ε_0
1N
μ_0
μ_0
1 Joule
v^2
1 Ampere creates orthogonal magnetic fields

Joule

A measure of the equilateral charged
mass-energy momenta that can to do work

... the mechanical power exerted in turning a magneto-electric machine is converted into the heat evolved by the passage of the currents of induction through its coils; and, on the other hand, that the motive power of the electro-magnetic engine is obtained at the expense of the heat due to the chemical reactions of the battery by which it is worked (1845)

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

It is equal to the energy expended (or work done) in applying a force of one newton through a distance of one meter (1 newton metre or N·m), or in passing an electric current of one ampere through a resistance of one ohm for one second

$$\boxed{\text{Energy}\quad kg.\frac{m^2}{s^2}}$$

$$J = \frac{kg \cdot m^2}{s^2} = N \cdot m = Pa \cdot m^3 = W \cdot s$$

Planck's Constant x frequency $\left[kg.\frac{m^2}{s}\right].s^{-1}$

mass x velocity squared $kg.\left[\frac{m}{s}\right]^2$

Momentum x velocity $\left[kg.\frac{m}{s}\right].\frac{m}{s}$

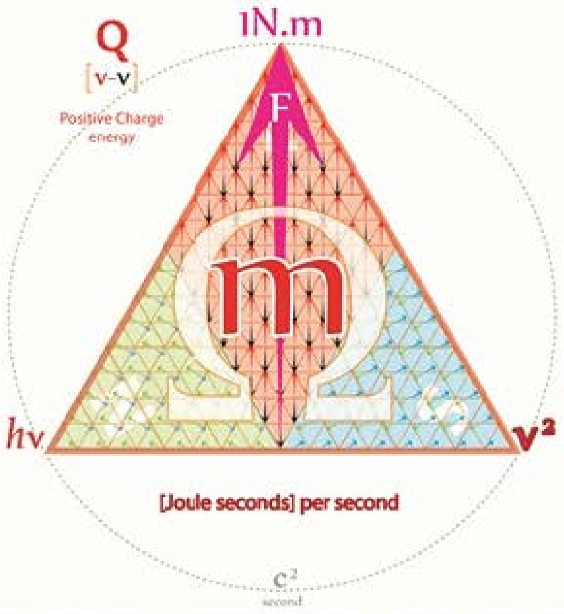
Q
[v-v]
Positive Charge energy
1N.m
F
m
$h\nu$
v^2
c^2
second

[Joule seconds] per second

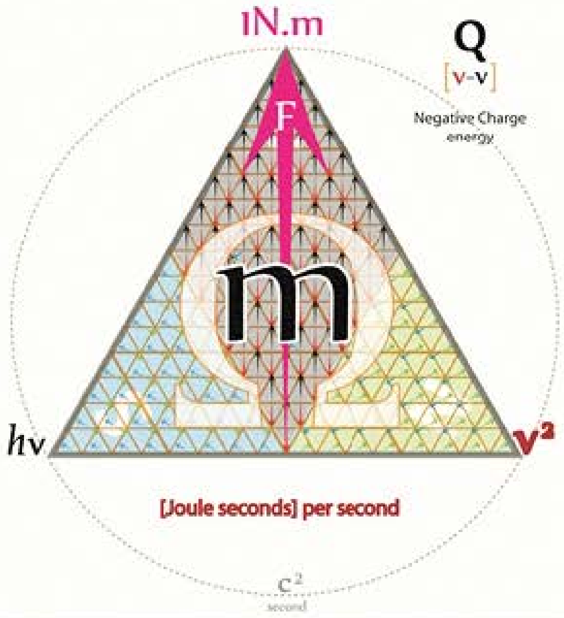
1N.m
Q
[v-v]
Negative Charge energy
F
m
$h\nu$
v^2
c^2
second

[Joule seconds] per second

Copyright ABRAHAM [2008] - all rights reserved

Volts

[Joules per Coulomb]

The volt is the unit for electric potential (voltage), electric potential difference, and electromotive force.

$$V = \frac{W}{A} = \frac{J}{A \cdot s} = \frac{N \cdot m}{A \cdot s} = \frac{kg\,m^2}{A\ s^3} = \frac{kg \cdot m^2}{C \cdot s^2} = \frac{N \cdot m}{C} = \frac{J}{C}$$

A single volt is defined as the difference in electric potential across a wire when an electric current of one ampere dissipates one watt of power.
It is also equal to the potential difference between two parallel, infinite planes spaced 1 meter apart that create an magnetic field of 1 newton per coulomb.
Additionally, it is the potential difference between two points that will impart one joule of energy per coulomb of charge that passes through it

Positive charge

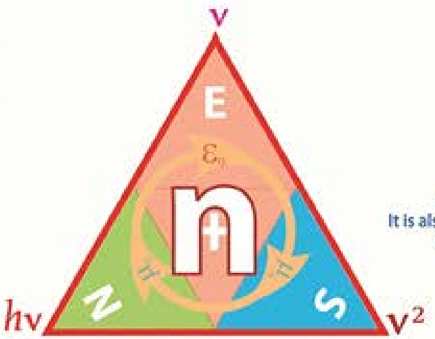
v
E
ε_0
n
N
S
hv
v^2

1V = 1J/C

One Volt is defined as energy consumption of
one joule per electric charge of one coulomb

Negative charge

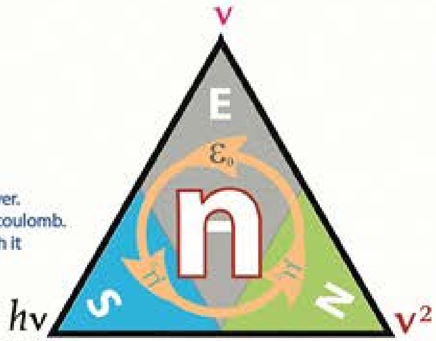
v
E
ε_0
n
S
N
hv
v^2

1V = 1A · 1ohm

One volt is equal to current of 1 amp
times resistance of 1 ohm

Positive voltage

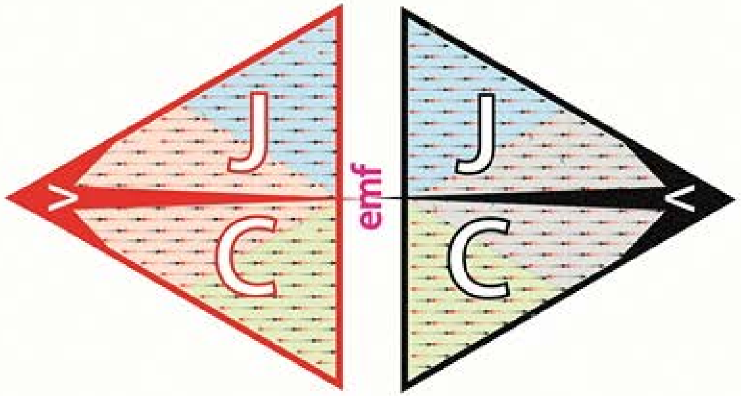

Negative voltage

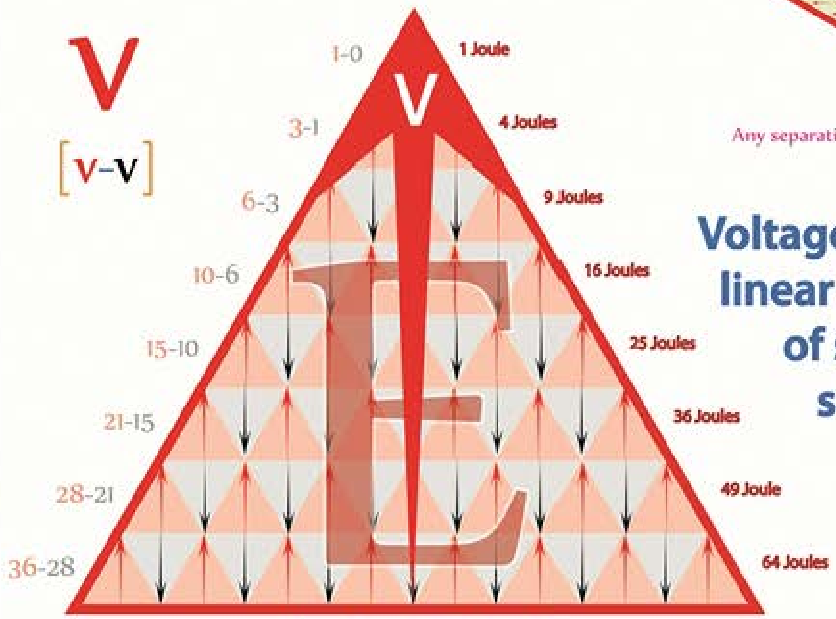
V
[v–v]
1-0
3-1
6-3
10-6
15-10
21-15
28-21
36-28
V
E
1 Joule
4 Joules
9 Joules
16 Joules
25 Joules
36 Joules
49 Joule
64 Joules

[Joules per Coulomb]

Any separation of electrostatic charges, results in a voltage.
[Potential Difference]

Voltage is a scalar measure of linear electromotive forces of separated charges seeking potential equilibrium

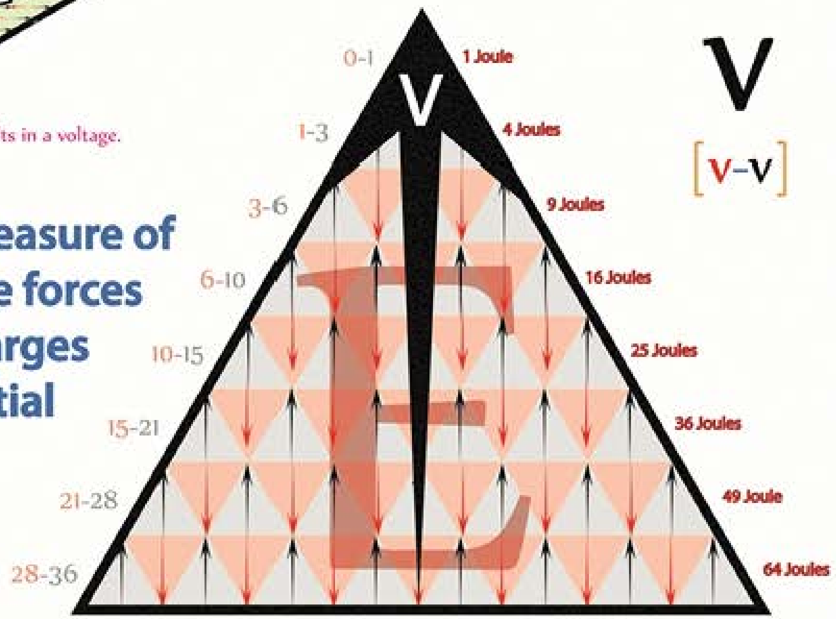
0-1
1-3
3-6
6-10
10-15
15-21
21-28
28-36
V
E
1 Joule
4 Joules
9 Joules
16 Joules
25 Joules
36 Joules
49 Joule
64 Joules
V
[v–v]

[Joules per Coulomb]

Positive charged mass-energy

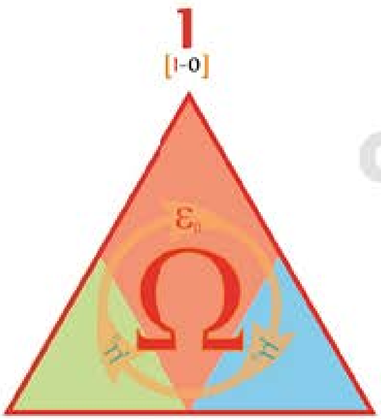

Clockwise inductive energy flux

Charge is the equilateral geometry of Energy that gives form to all mass & Matter

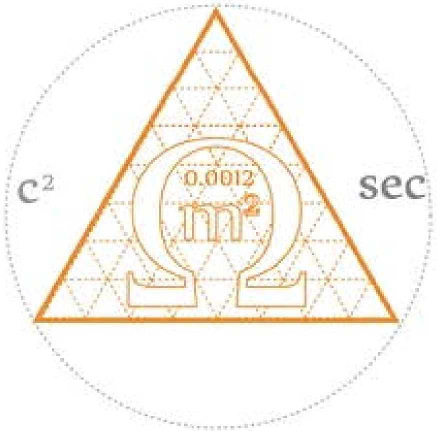

Negative charged mass-energy

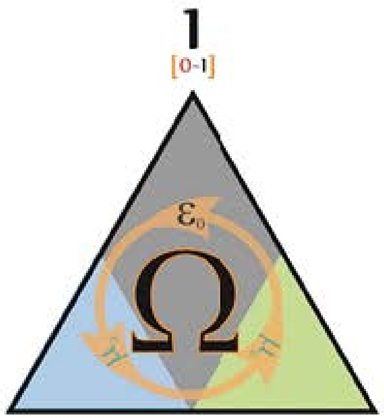

Counter clockwise energy flux

charge

time

q

q

t

$$\frac{\Omega}{c^2}$$

$$\frac{m^2}{s}\,\frac{s^2}{m^2}$$

$$s$$

seconds

t

1.33518 e-20 s

Changes to nett quantised angular momenta [charged mass-energies] within any space-time co-ordinate system forms the basis of time

ElectroMagnetic Charge is a quantum property resulting from the equilateral QAM geometry of mass-Energy

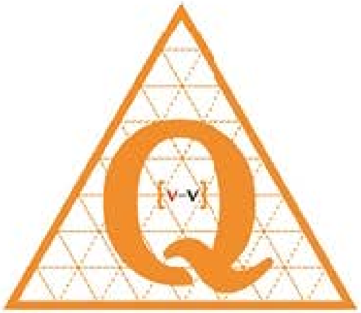

The two ElectroMagnetic charge geometries possible can be created through tesselation of electrical energy in ideal inductive loops

Copyright ABRAHAM [2008] - all rights reserved

Coulombs

are a measure of 2D charged mass geometries
and the electromagnetic forces they produce per unit time

W+ boson integrals
nett positive charged
mass momenta

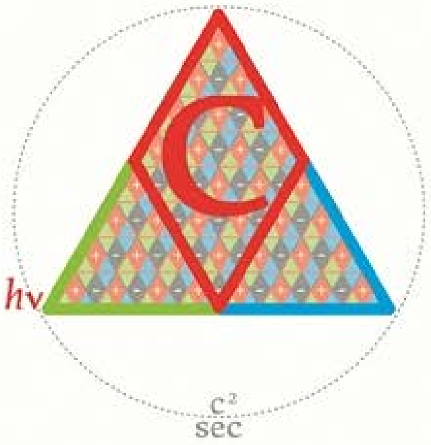

$$Q \; [v\text{-}v]$$

Charles-Augustin de Coulomb

(14 September 1736 – 23 August 1806)

Derived in 1785 Coulomb's Law is a law of physics describing the electrostatic interaction between any two charged particles (and forms the basis for Ampere's Law)

$$Q \; [v\text{-}v]$$

W- boson integrals
nett negative charged
mass momenta

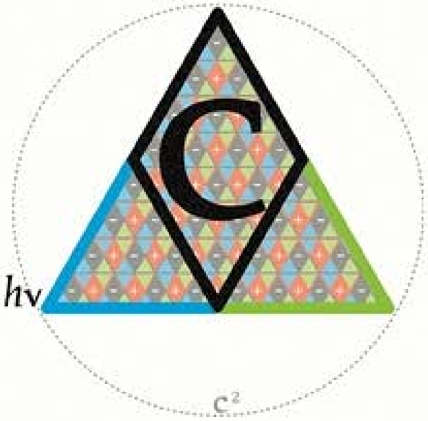

Electric forces

$$m\frac{\Omega}{c^2}$$

Magnetic forces

Coulombs

The coulomb (unit symbol: C) is the SI derived unit of electric charge (symbol: Q or q).
It is defined as the charge transported by a steady current of one ampere in one second:

Amperes

$$As$$

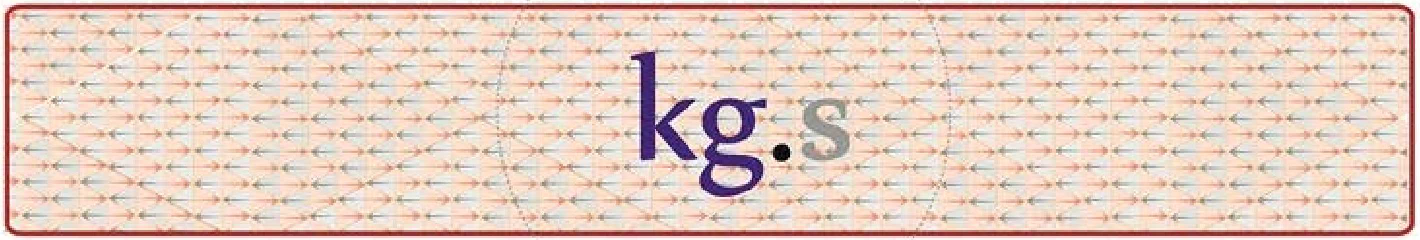

$$\frac{C}{s}$$

Interactive linear momenta of super-positioned E-fields creates Coulomb's force law

c^2
sec

1A.1s

1F.1V

Copyright ABRAHAM [2008] - all rights reserved

$1\text{C} = 1\text{A} \cdot 1\text{s}$

1 COULOMB is a measure of linear
2D electric field energy momenta
in EM field geometries

1 AMPERE is a measure of orthogonal
2D magnetic field forces produced
by Matter topologies in motion

$1\,\text{A} = 1\frac{\text{C}}{\text{s}}$

quantised electrical energy momenta

1C is the quantity of charged masses which pass any point in a circuit in which a current of 1A flows for 1s

C

Coulombs

kg.s

linear acceleration
energy momenta

A

Amperes

KG

magnetic perturbation
energy momenta

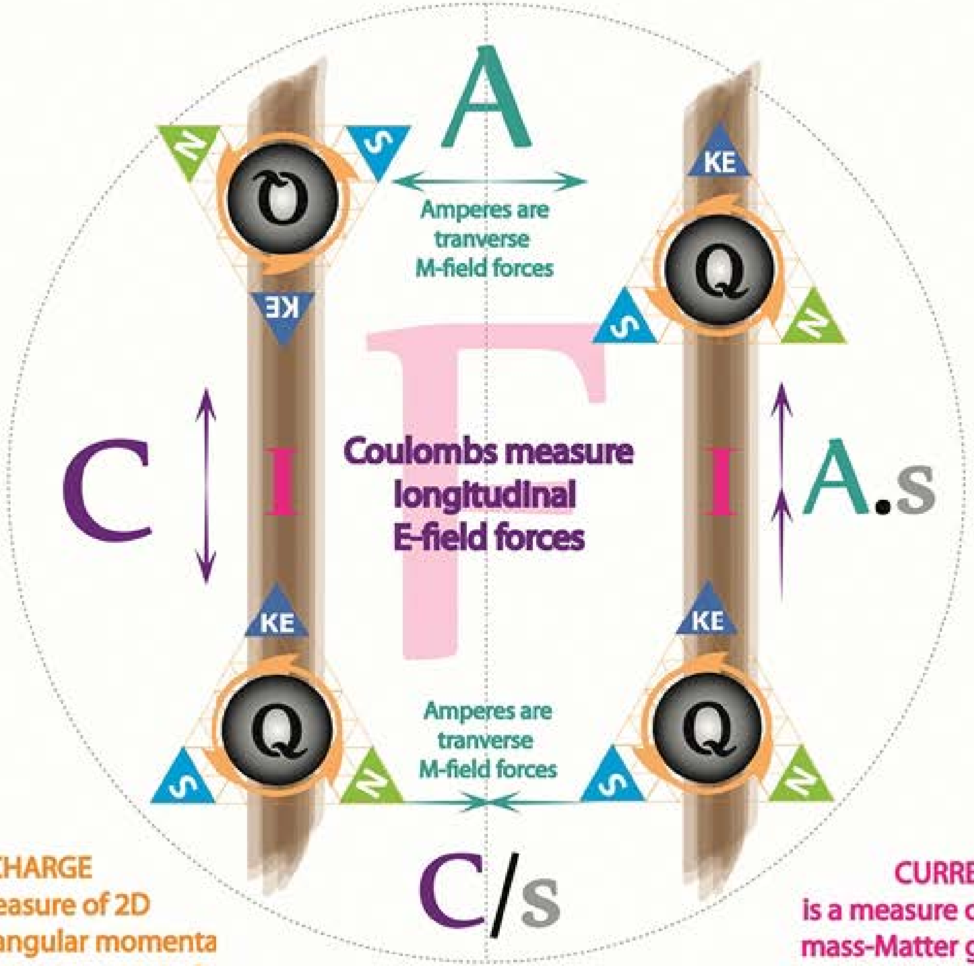

Amperes are
tranverse
M-field forces

Coulombs measure
longitudinal
E-field forces

Amperes are
tranverse
M-field forces

Q

Charge

CHARGE
is a measure of 2D
quantised angular momenta
[mass -Matter geometries]

$$Q = I.t$$

quantity of charge = rate of flow of charge × time

CURRENT
is a measure of charged
mass-Matter geometries
in motion

I

Current

Copyright ABRAHAM [2008] - all rights reserved

Current

is a measure of charged Matter in motion in a circuit per unit of Time

An elementary charge is
1.602216081 e-19 C

C

It is defined as the charge
transported by a steady current of
one ampere in one second.

Coulombs

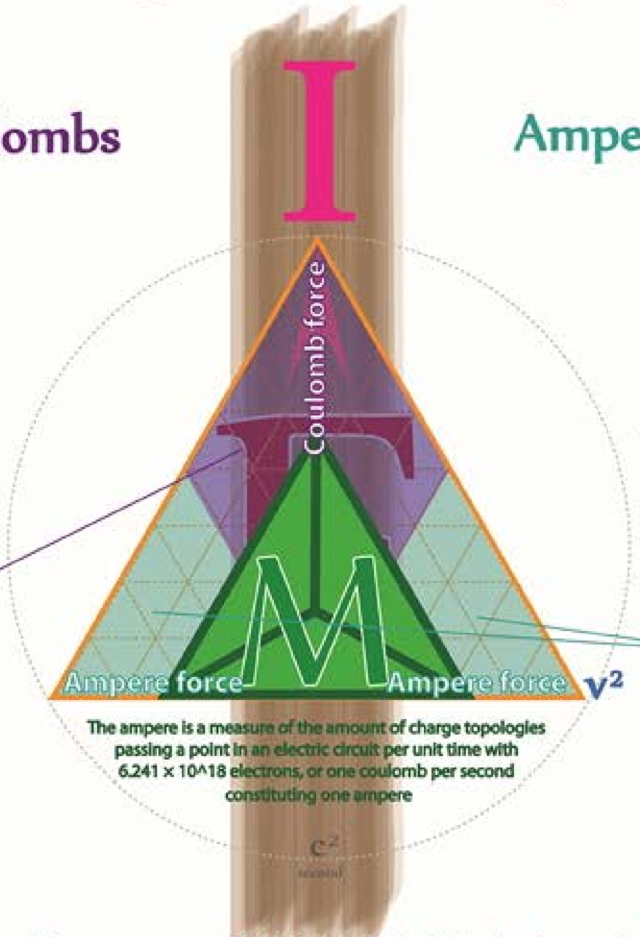

The ampere is a measure of the amount of charge topologies passing a point in an electric circuit per unit time with 6.241 x 10^18 electrons, or one coulomb per second constituting one ampere

Amperes

The amount of electric charge
passing a point per unit time

A

6.2413354 e18 electrons passing a
given point each second constitutes
one ampere.

charged **masses**

kg.s

1.33518 e-20 kg.s

7.489626 e19 kg

Matter

KG

The 2D linear mass-energy momenta capable of accelerating charged particles

The transverse magnetic forces produced by Matter in motion

$$1C = 1A \cdot 1s$$

$$1\,A = 1\frac{C}{s}$$

Electric current is a measure of the flow of electrically charged mass-Matter in a circuit as a result of an electrical Voltage potential

As

5.524527227 e-12 kg

Charged mass

1 second

Matter in motion

6.241355408 e18 electrons

$$\frac{C}{S}$$

Voltage [emf] is a measure of the energy momenta per coulomb available to exert a force on charged Matter in a circuit

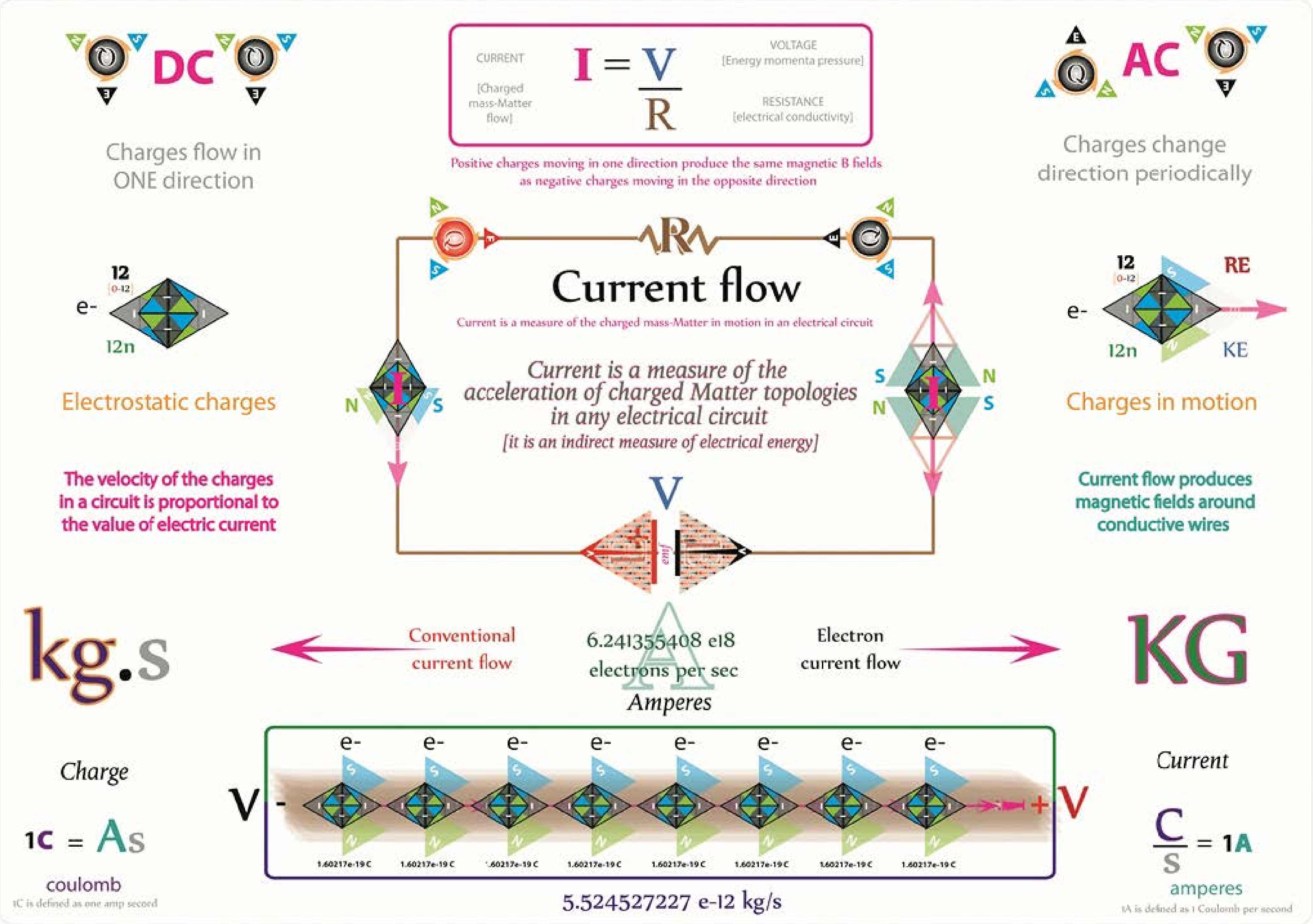

DC

Charges flow in ONE direction

AC

Charges change direction periodically

CURRENT [Charged mass-Matter flow]

$$I = \frac{V}{R}$$

VOLTAGE [Energy momenta pressure]

RESISTANCE [electrical conductivity]

Positive charges moving in one direction produce the same magnetic B fields as negative charges moving in the opposite direction

Current flow

Current is a measure of the charged mass-Matter in motion in an electrical circuit

Current is a measure of the acceleration of charged Matter topologies in any electrical circuit [it is an indirect measure of electrical energy]

Electrostatic charges

Charges in motion

The velocity of the charges in a circuit is proportional to the value of electric current

Current flow produces magnetic fields around conductive wires

Conventional current flow

6.241355408 e18 electrons per sec

Amperes

Electron current flow

kg.s

KG

Charge

1C = As

coulomb

1C is defined as one amp second

Current

$\frac{C}{s} = 1A$

amperes

1A is defined as 1 Coulomb per second

5.524527227 e-12 kg/s

Tetryonics 23.10 - Current [Matter in motion]

The electrical resistance of a conductor is a measure of how much it opposes the passage of an electric current through it

OHM's Law

Georg Simon Ohm

(16 March 1789 – 6 July 1854)

resistivity

$$\mathcal{E} = \rho J$$

Electric field

Current density

General vector equation

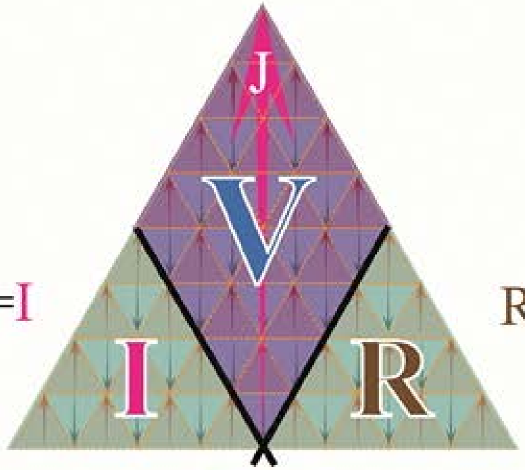

$$\frac{V}{R} = I$$

$$R = \frac{V}{I}$$

$$V = I.R$$

$$P = I.V = I^2.R$$

Power is the amount of current times the voltage level at a given point measured in watts.

Current is the flow of charged mass-Matter resulting from an electromotive force

Resistance determines how much current will flow through a conductor

Current [I.V] Voltage
POWER
Current squared [I^2R] Resistance

Ohm's law states that the current through a conductor between two points is directly proportional to the potential difference across the two points.

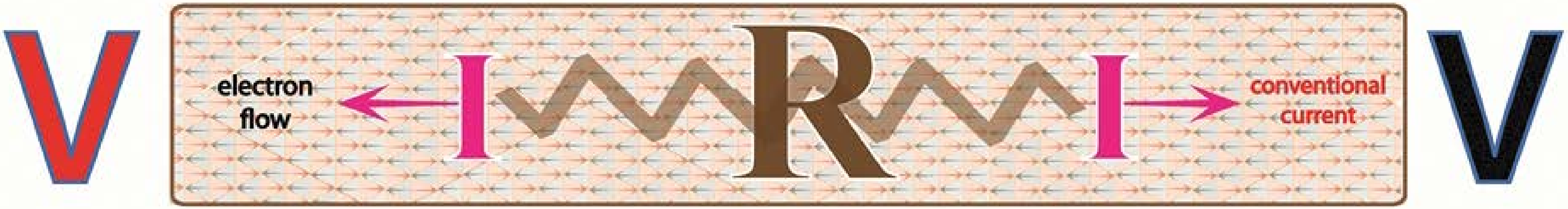

By introducing the constant of proportionality, Resistance, one arrives at the usual mathematical equation that describes this relationship

Voltage is the force motivating charges to "flow" in a circuit, it is measured as the difference in electrical potential between two points in a circuit

Copyright ABRAHAM [2008] - all rights reserved

Electrical Resistance

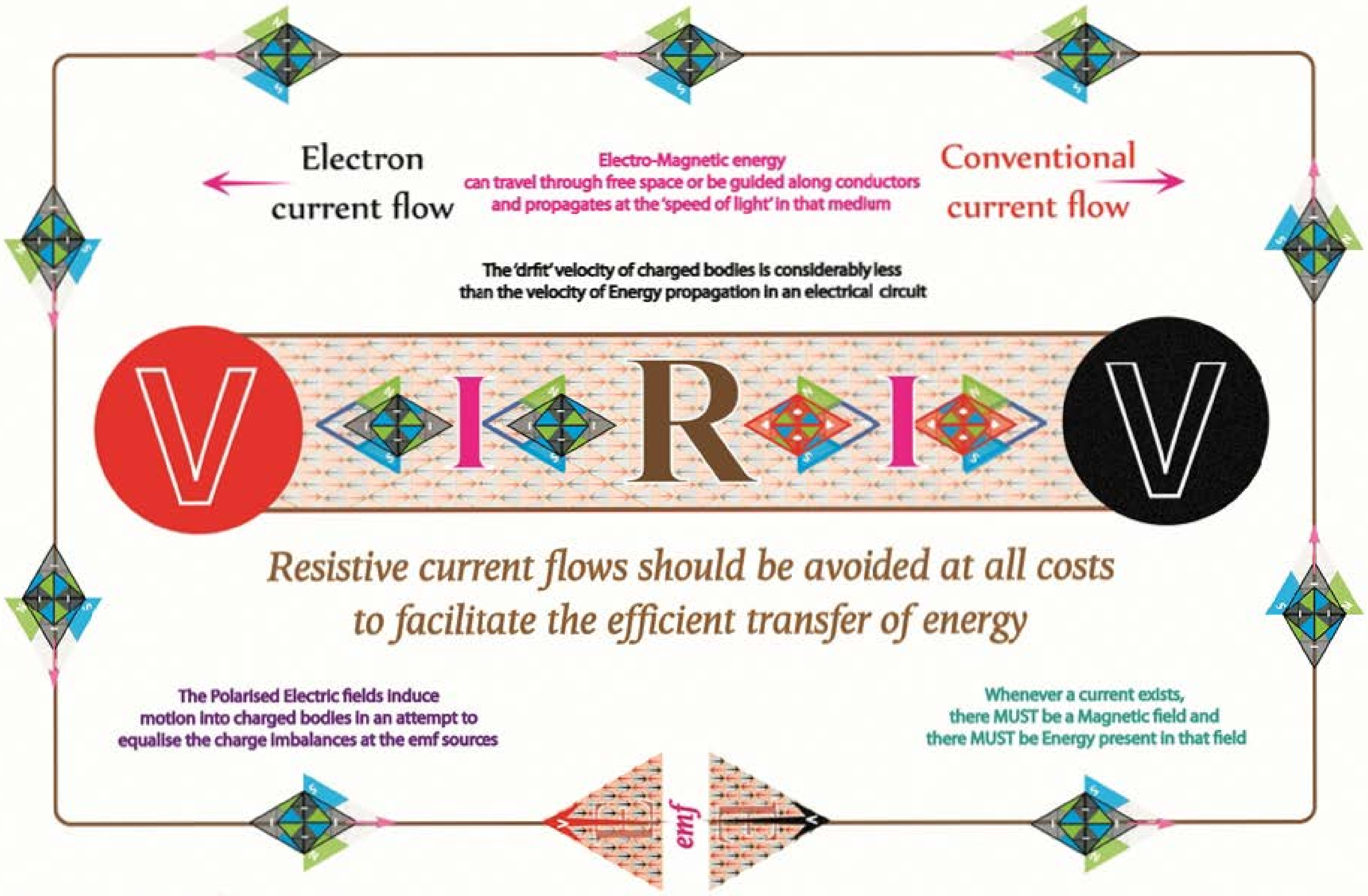

Electron current flow

Electro-Magnetic energy
can travel through free space or be guided along conductors
and propagates at the 'speed of light' in that medium

Conventional current flow

The 'drfit' velocity of charged bodies is considerably less
than the velocity of Energy propagation in an electrical circuit

*Resistive current flows should be avoided at all costs
to facilitate the efficient transfer of energy*

The Polarised Electric fields induce
motion into charged bodies in an attempt to
equalise the charge imbalances at the emf sources

Whenever a current exists,
there MUST be a Magnetic field and
there MUST be Energy present in that field

The acceleration [change in velocity] of electrons in motion
requires the release [or absorption] of Bosons of energy
from [or to] the equilateral KEM field geometry
[[spectriaaline emission~absorption]

Voltage produces a linear force that motivates
charged Matter topologies into motion in a circuit
as electrical energy geometries seek equilibrium

Power is a measure of the total charge geometry
(electrons) moving past a point along with the energy each
charge geometry (electron) possesses (KE expressed as eV)
V.I = P = i.R.I

Copyright ABRAHAM [2008] - all rights reserved

James Prescott Joule verified Ohm's Law and determined that the heat delivered by a conductor is directly proportional to its resistance and to the square of the current through it.

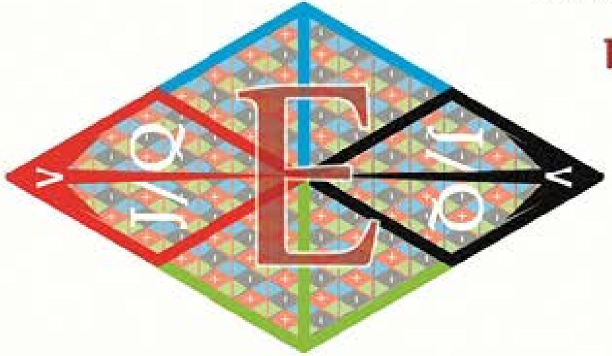

All energy is comprised of 2D charged masses

He defined Power physically to be the rate at which electrical energies are created or transferred in an electric circuit

Electrical Power

Power is the integral of the cross-product of all electrical and magnetic field vectors produced by scalar energy momenta within a specific area defined by time

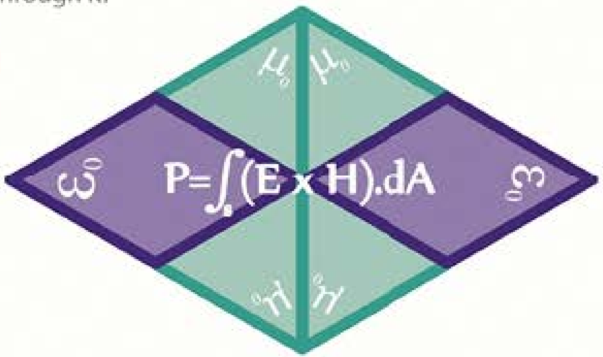

The scalar surface integral of the Poynting vector.

The SI unit of power is the watt, one joule per second.

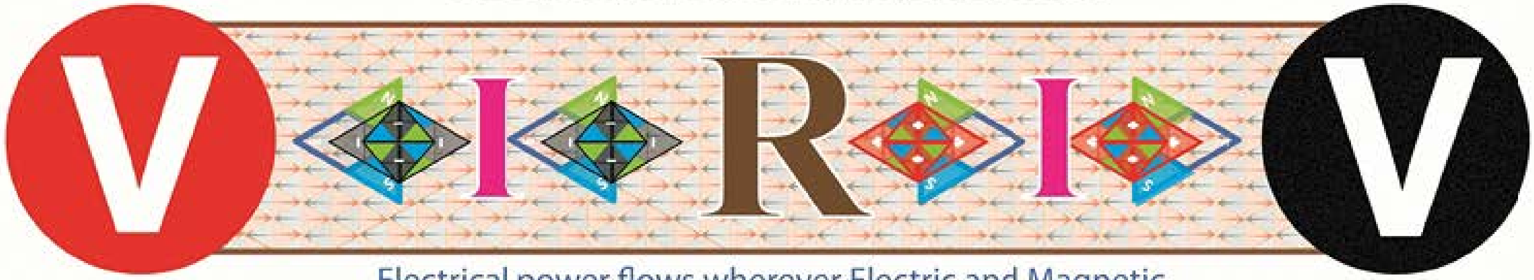

Electrical power flows wherever Electric and Magnetic fields exist together and change over time

emf Voltage

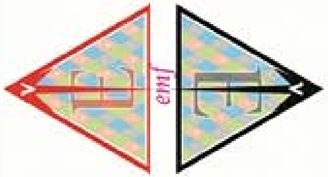

The potential to do work

$$P = V.I$$

Charge geometry **Q**

RE Relativistic mass-energies densities

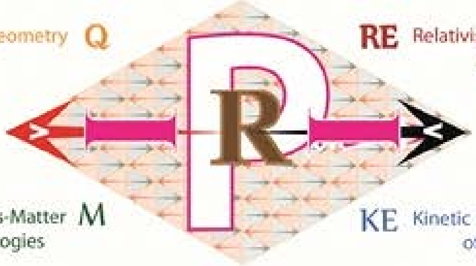

3D mass-Matter **M** topologies

KE Kinetic mass-energies of motion

$$I^2.R = \frac{V^2}{R}$$

Resistance

[0-12]

1.2e 20

The inertial mass of Matter

Electrical power propagates at the speed of light in an electrical circuit [dependent on the voltage source] electrons have a vastly slower 'drift' velocity resulting from their inertial interaction with energy momenta of the Voltage [emf] field

Copyright ABRAHAM [2008] - all rights reserved

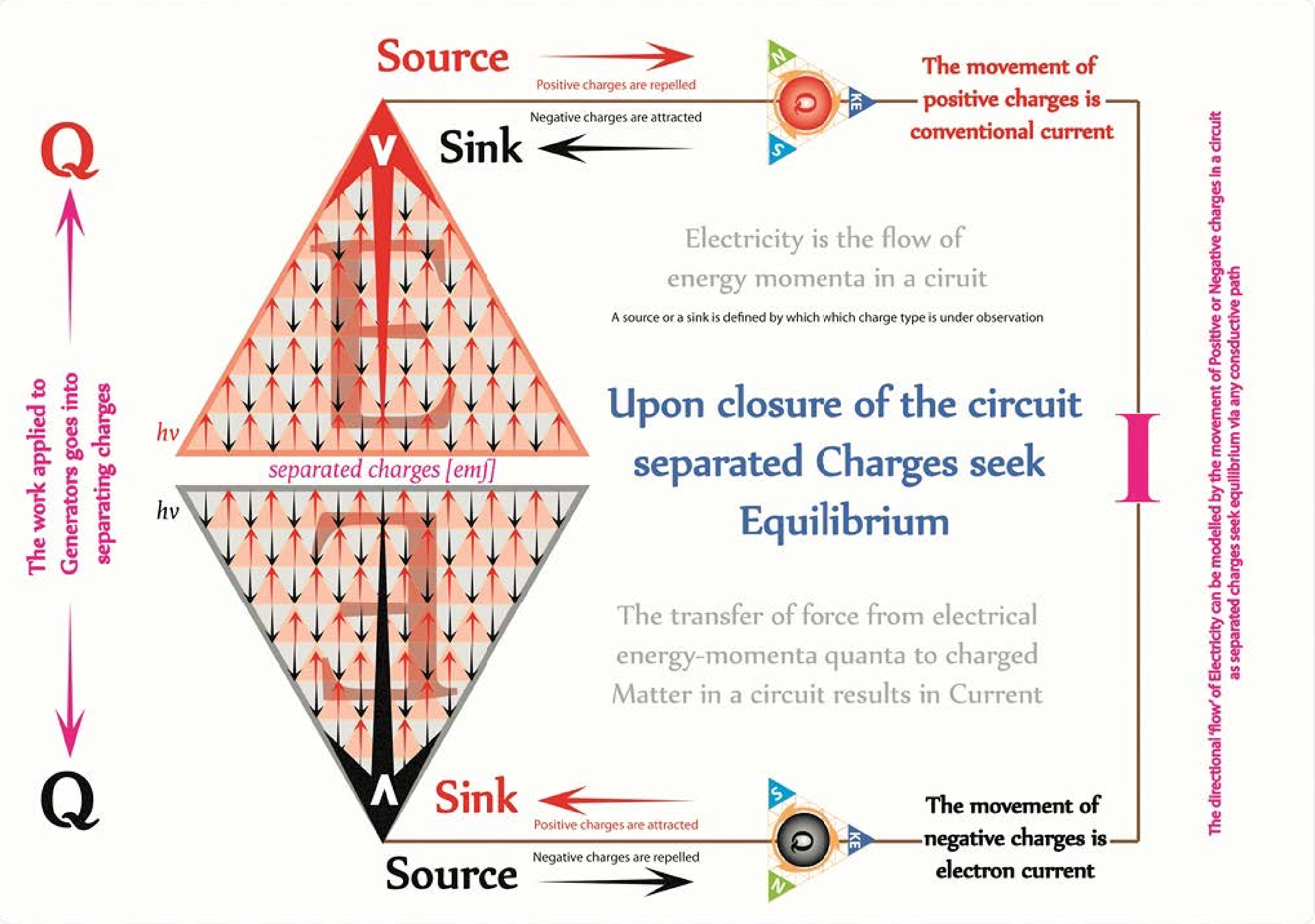
Source
Positive charges are repelled
Negative charges are attracted
Sink
Q
N
S
KE
V
E
hv
separated charges [emf]
hv
Sink
Positive charges are attracted
Negative charges are repelled
Source
Q
I

The movement of positive charges is conventional current

Electricity is the flow of energy momenta in a ciruit

A source or a sink is defined by which which charge type is under observation

Upon closure of the circuit separated Charges seek Equilibrium

The transfer of force from electrical energy-momenta quanta to charged Matter in a circuit results in Current

The movement of negative charges is electron current

The work applied to Generators goes into separating charges

The directional 'flow' of Electricity can be modelled by the movement of Positive or Negative charges in a circuit as separated charges seek equilibrium via any conductive path

Tetryonics 24.01 - Sources and Sinks

Copyright ABRAHAM [2008] - all rights reserved

EM Inductive circuits

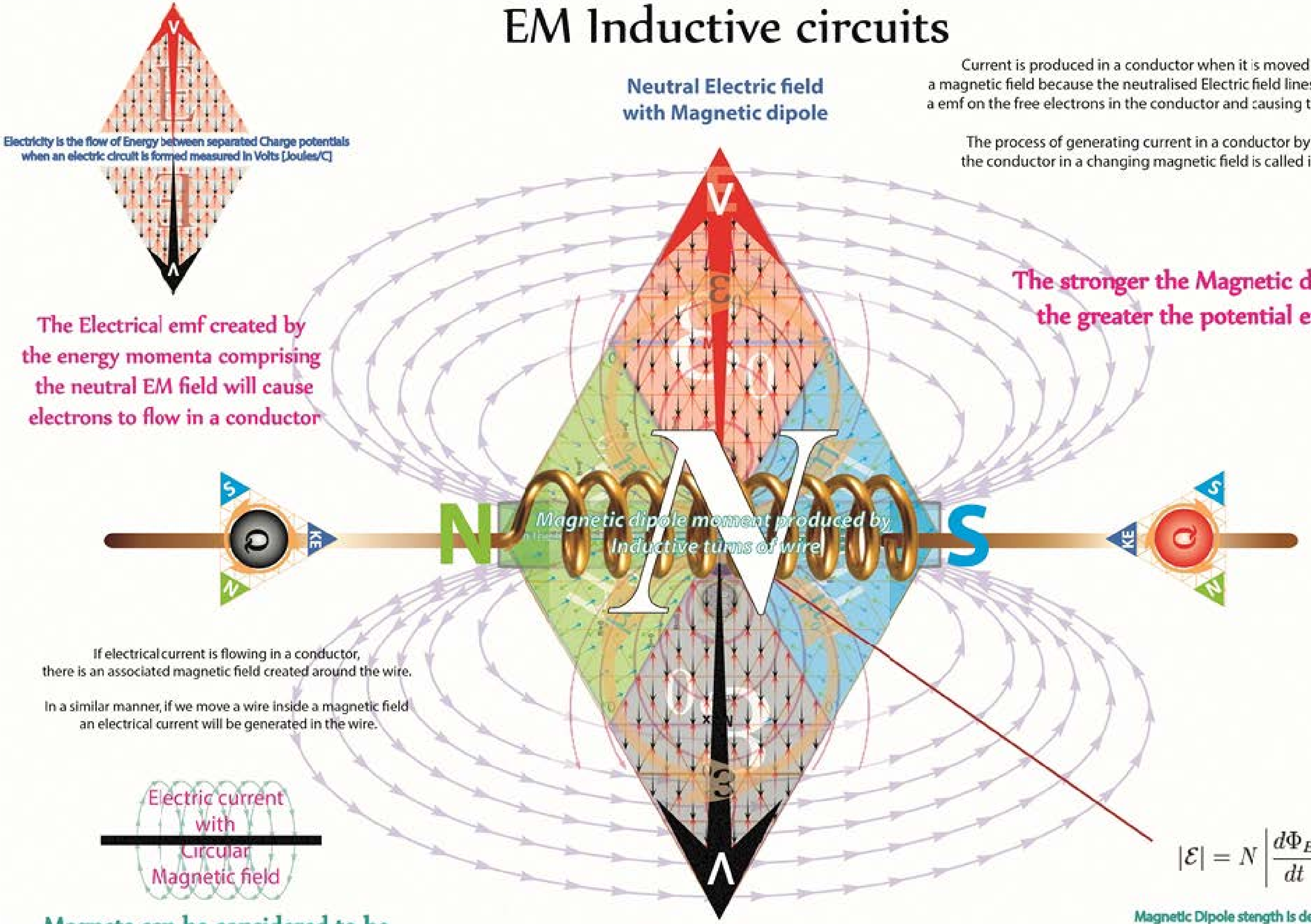

Electricity is the flow of Energy between separated Charge potentials
when an electric circuit is formed measured in Volts [Joules/C]

Neutral Electric field
with Magnetic dipole

Current is produced in a conductor when it is moved through
a magnetic field because the neutralised Electric field lines are applying
a emf on the free electrons in the conductor and causing them to move.

The process of generating current in a conductor by placing
the conductor in a changing magnetic field is called induction

The Electrical emf created by
the energy momenta comprising
the neutral EM field will cause
electrons to flow in a conductor

The stronger the Magnetic dipole
the greater the potential emf

If electrical current is flowing in a conductor,
there is an associated magnetic field created around the wire.

In a similar manner, if we move a wire inside a magnetic field
an electrical current will be generated in the wire.

Electric current
with
Circular
Magnetic field

$$|\mathcal{E}| = N \left| \frac{d\Phi_B}{dt} \right|$$

Magnets can be considered to be
static stores of potential [emf fields]

Magnetic Dipole
with Neutral Electric field

Magnetic Dipole stength is dependent on:

The total Number of turns in the inductor
& the current flowing through the circuit

Electron Volts

The electron volt can also be used
as a unit of mass-energy by applying
Einstein's relation $E = mc^2$.

For example, the rest mass-energy of electron
topologies if 496,532 eV (496.532 MeV).

Chemically, for 1 mole of electrons
1 eV ~ 100 kJ mol-1 (96.49 kJ mol-1)

eV

A unit of energy
equal to the work required to move
one electron through a potential
difference of 1 volt.

An Electron Volt is also
a measure of Electrical Kinetic Energy

EM Field Planck quanta

$$\frac{1}{2}\left[4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]\right]$$

Tetryons ElectroMagnetic mass velocity

As distinct from Matter's ENERGY
Kinetic Energy is 'extended from the
Tetryonic geometry of all Matter in motion
in a separate 2D KEM field

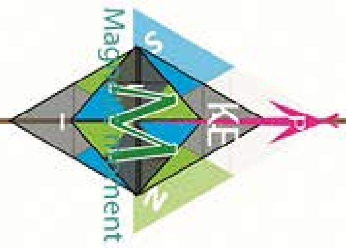

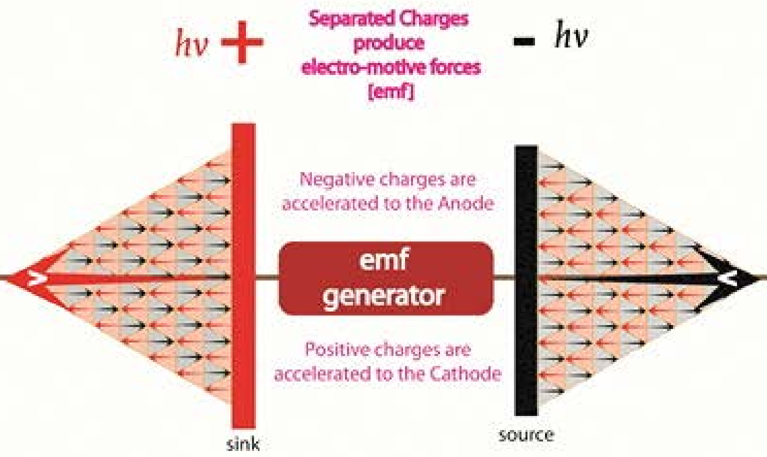

Separated Charges produce electro-motive forces [emf]

Negative charges are accelerated to the Anode

Positive charges are accelerated to the Cathode

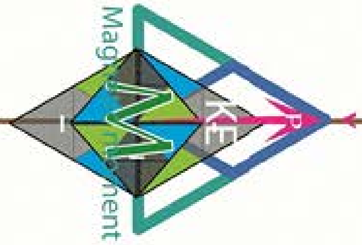

A Free electron is attracted to Positive anode of an emf source

The application of UFE to Kinetic energy calculations reveals the underlying processes of the Photo-electric effect and Light-Energy interactions with respect to Kinetic energy

Accelerated electrons have increased KE, Magnetic moments & momentum

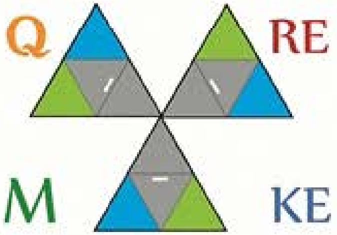

EM Field Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Fermions ElectroMagnetic mass velocity

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

By definition, it is equal to the amount of kinetic energy momenta gained by a single unbound electron when it accelerates through an electric potential difference of one volt.

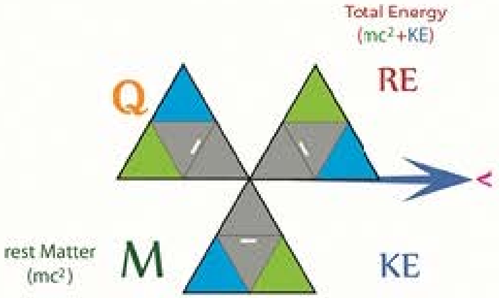

(Total Energy = Absolute rest Matter + Kinetic energy)
additionally the Lorentz relativistic correction factor (β)
presents itself naturally from the UFE when velocity is applied

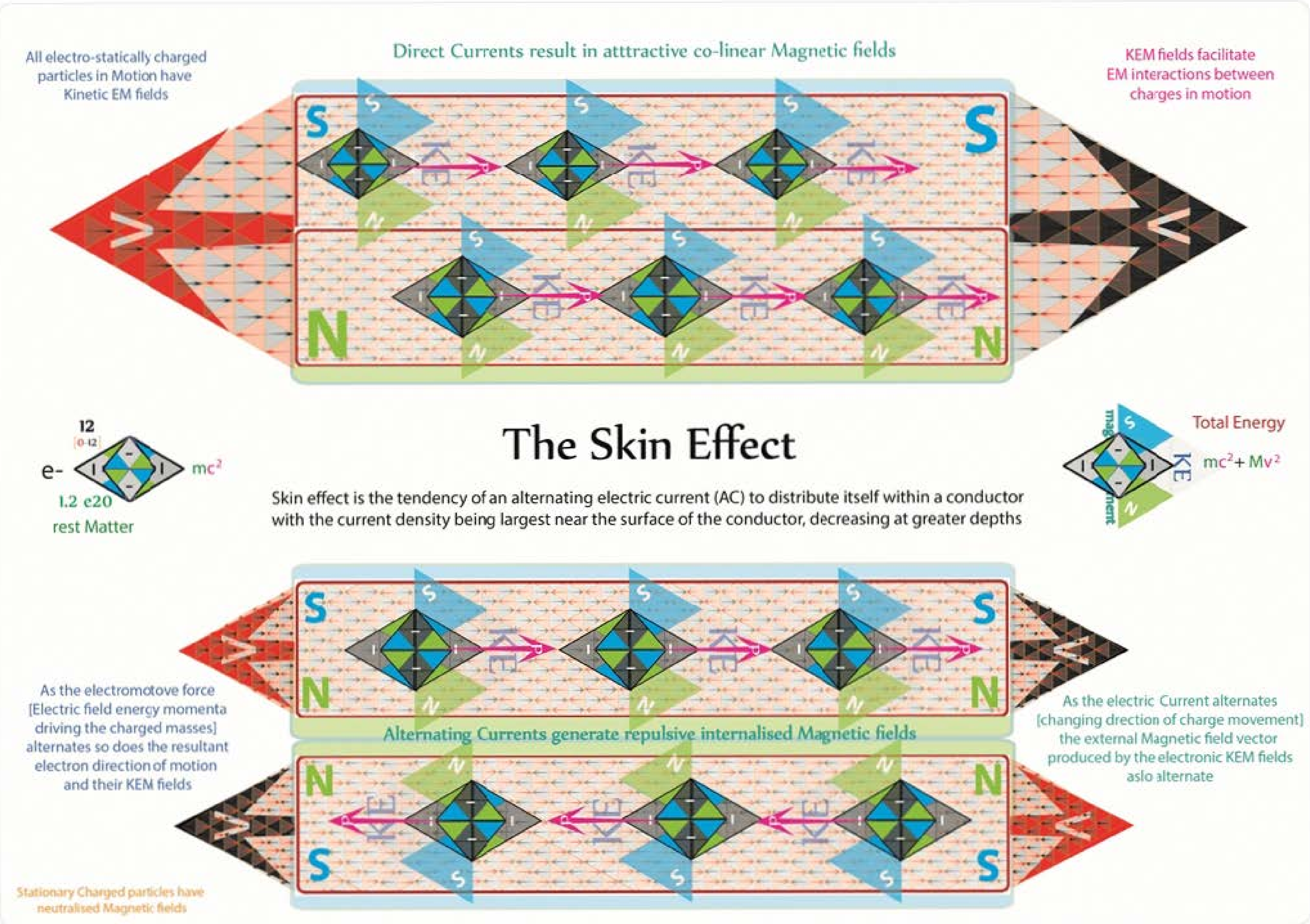

All electro-statically charged particles in Motion have Kinetic EM fields

Direct Currents result in atttractive co-linear Magnetic fields

KEM fields facilitate EM interactions between charges in motion

The Skin Effect

Skin effect is the tendency of an alternating electric current (AC) to distribute itself within a conductor with the current density being largest near the surface of the conductor, decreasing at greater depths

Alternating Currents generate repulsive internalised Magnetic fields

As the electromotove force [Electric field energy momenta driving the charged masses] alternates so does the resultant electron direction of motion and their KEM fields

As the electric Current alternates [changing direction of charge movement] the external Magnetic field vector produced by the electronic KEM fields aslo alternate

Stationary Charged particles have neutralised Magnetic fields

EM fields of particles in Motion

Positive changed quanta traveling anti-clockwise in a solonoid creates a North-South magnetic field orientated in the direction shown

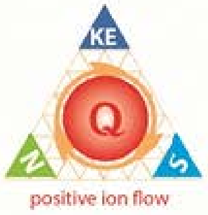

positive ion flow

Negative charged quanta create a reversed Magnetic dipole field and reversing the direction of particle motion also reverses the magnetic vector

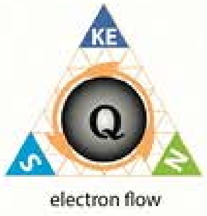

electron flow

Solenoidal Motion

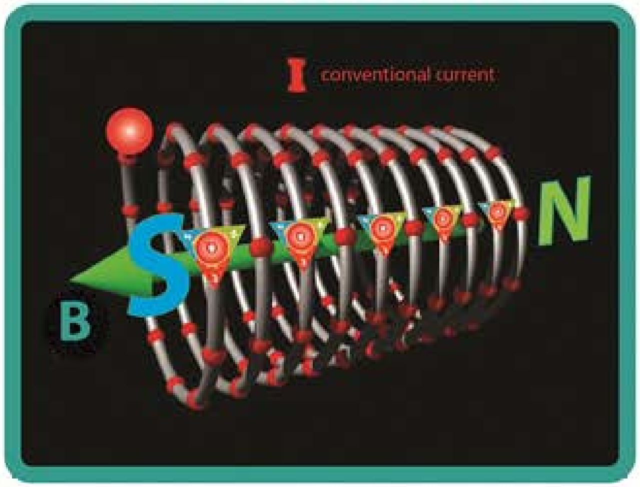

Charged particles moving in a spiral produce a Circular Magnetic field

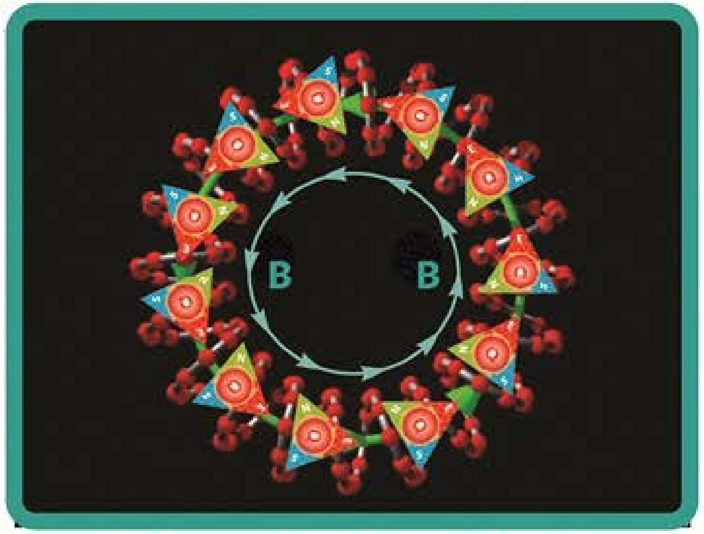

Toroidal Motion

A toroidally wound conductor produces a Circular Magnetic field

Helical Motion

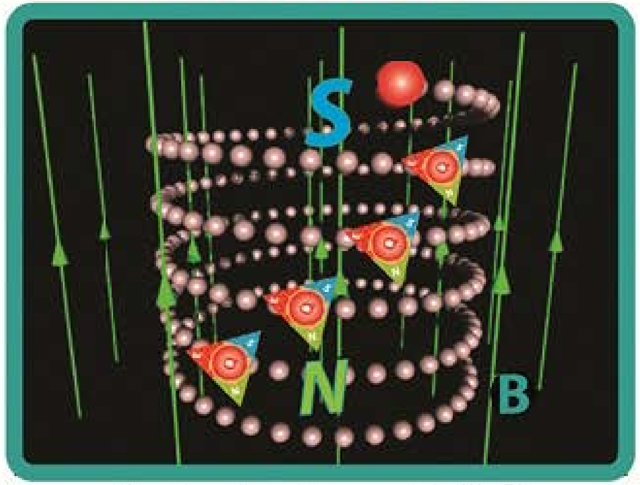

Motion in a Magnetic field

Inductors

An inductor (also choke, coil, or reactor) is a passive two-terminal electrical component that stores energy in its magnetic field.

For comparison, a capacitor stores energy in an electric field, and a resistor does not store energy but rather dissipates energy as heat

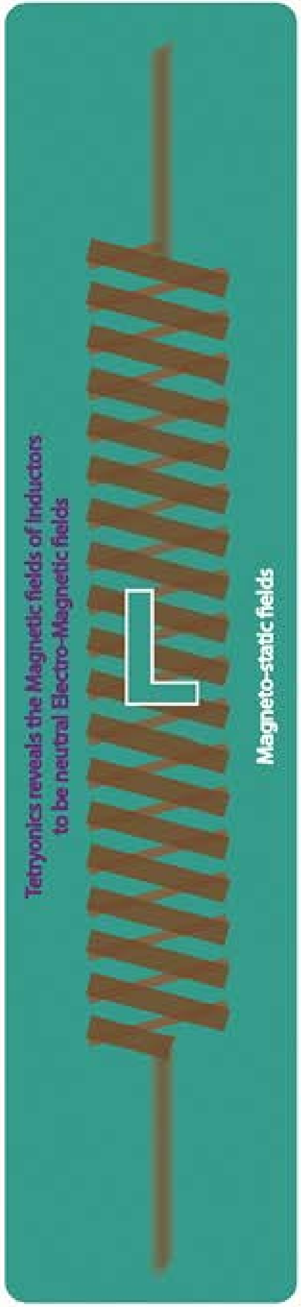
L
Tetryonics reveals the Magnetic fields of Inductors to be neutral Electro-Magnetic fields
Magneto-static fields

When the current flowing through an inductor changes, a time-varying electromagnetic field is created around the coil, and a changing emf voltage is formed

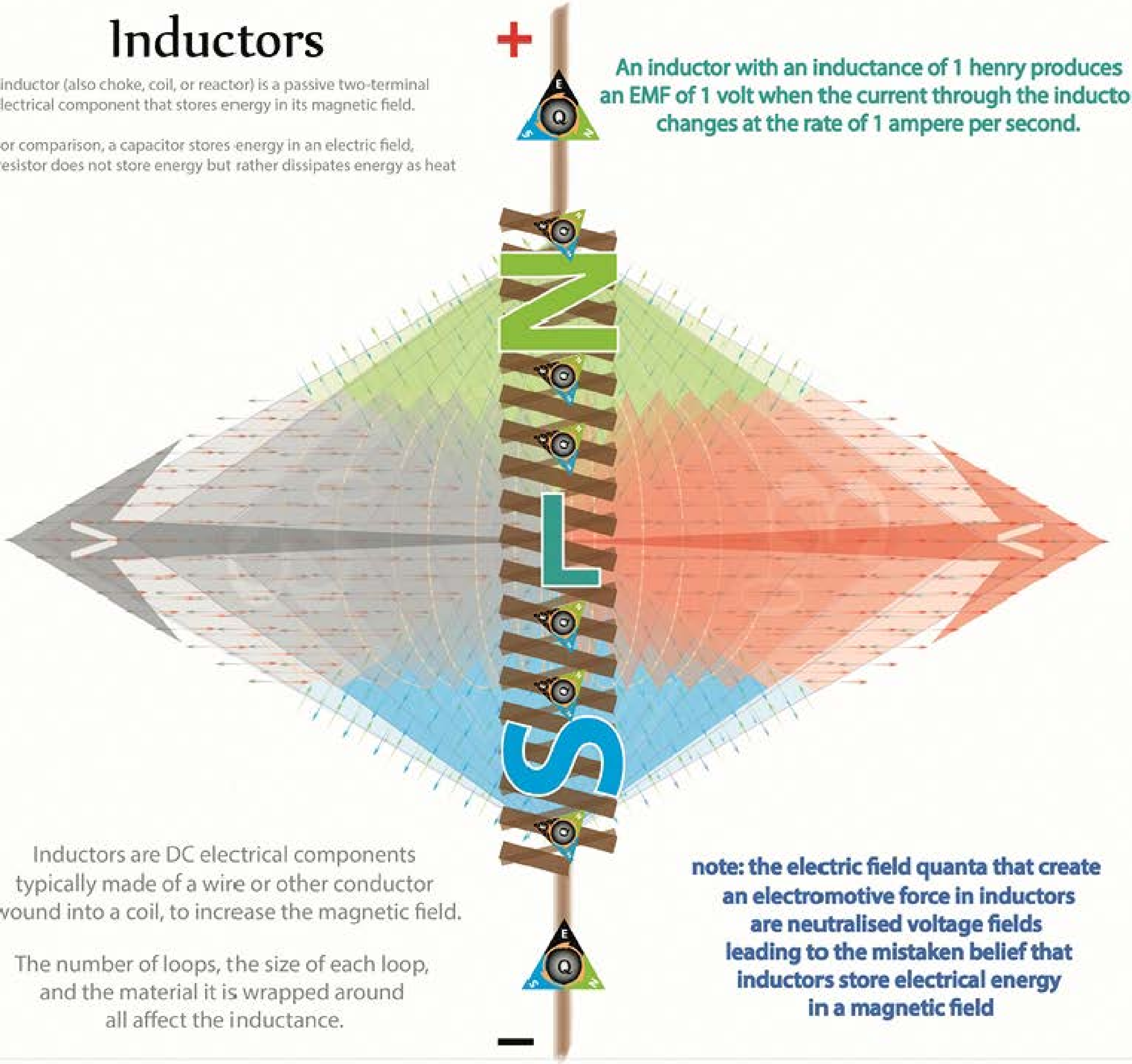

An inductor with an inductance of 1 henry produces an EMF of 1 volt when the current through the inductor changes at the rate of 1 ampere per second.

Inductors are DC electrical components typically made of a wire or other conductor wound into a coil, to increase the magnetic field.

The number of loops, the size of each loop, and the material it is wrapped around all affect the inductance.

note: the electric field quanta that create an electromotive force in inductors are neutralised voltage fields leading to the mistaken belief that inductors store electrical energy in a magnetic field

Charges moving in an Inductive winding

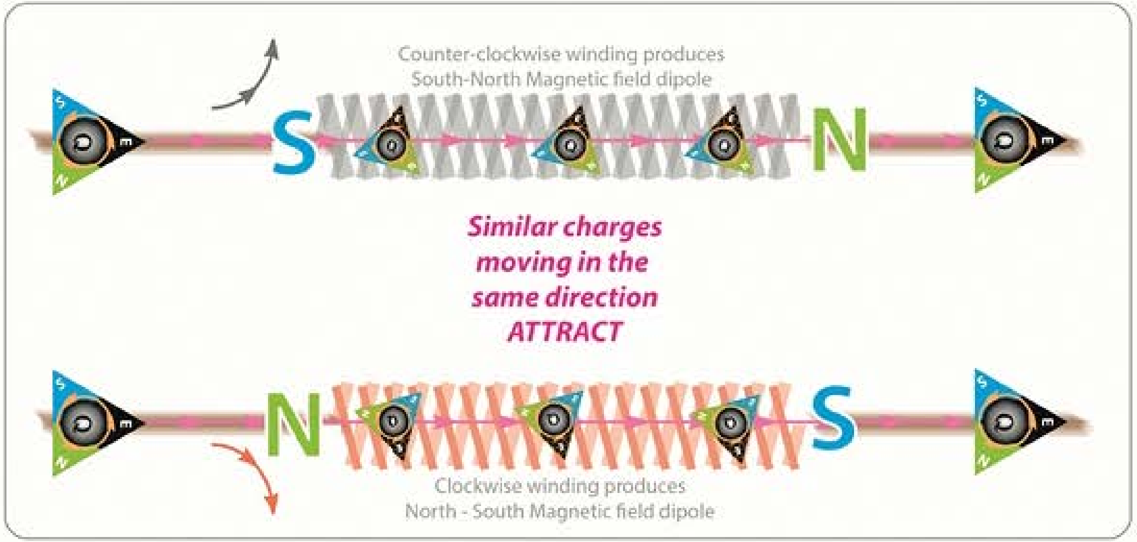

Counter-clockwise winding produces
South-North Magnetic field dipole

Similar charges moving in the same direction ATTRACT

Clockwise winding produces
North - South Magnetic field dipole

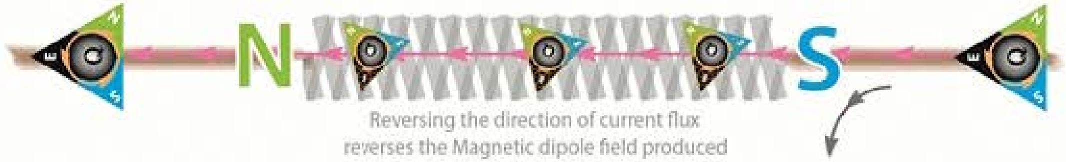

Reversing the direction of current flux
reverses the Magnetic dipole field produced

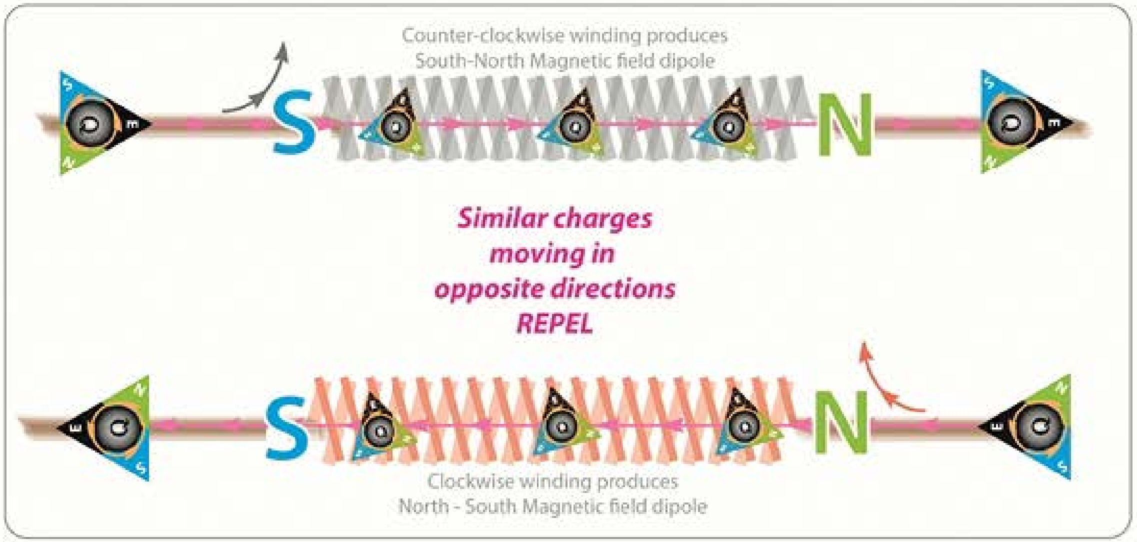

Counter-clockwise winding produces
South-North Magnetic field dipole

Similar charges moving in opposite directions REPEL

Clockwise winding produces
North - South Magnetic field dipole

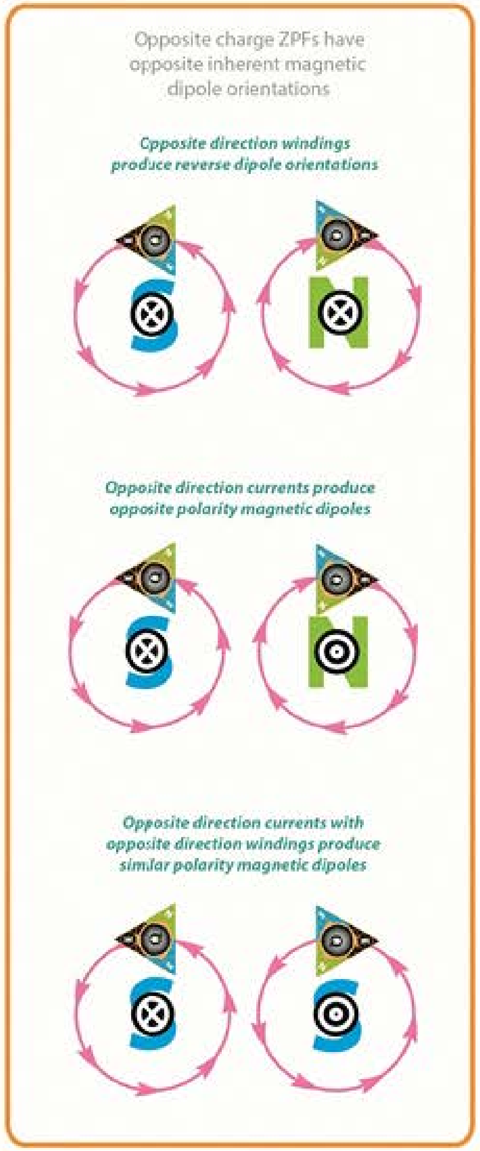

Opposite charge ZPFs have opposite inherent magnetic dipole orientations

Opposite direction windings produce reverse dipole orientations

Opposite direction currents produce opposite polarity magnetic dipoles

Opposite direction currents with opposite direction windings produce similar polarity magnetic dipoles

This is best viewed using physical models

Tetryonics 24.07 - Moving charges in an Inductor

Capacitors

Changing Electric fields
create
changing Magnetic fields

A capacitor is a passive two-terminal AC electrical component
used to store alternating electrical energy in an electric field.
It blocks the block of direct currents.

Changing Magnetic fields
create
changing Electric fields

Capacitors store & pass electrical energy as charged 2D masses

Capacitors store EM Energy
as transverse Bosons

Capacitors store charged
masses between their plates

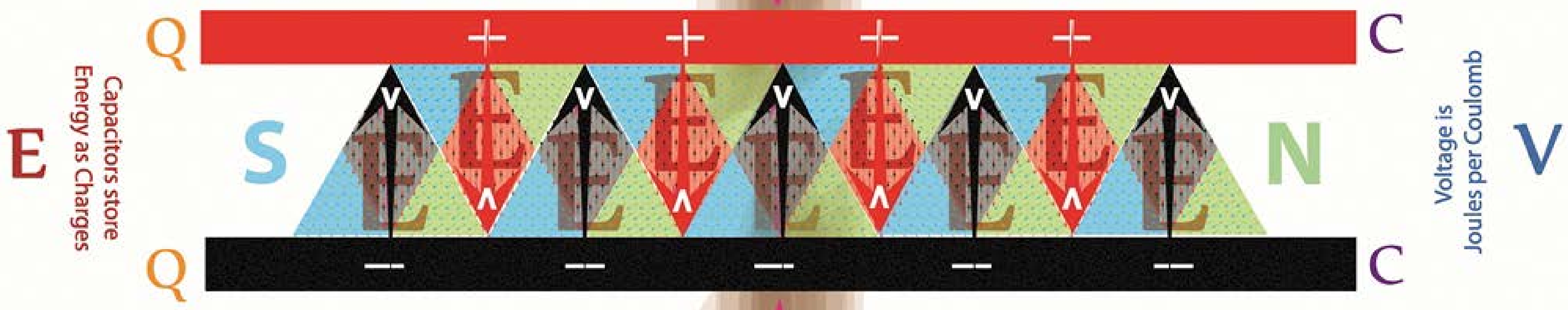

Charged Matter [current] does not
flow between the plates
of a Capacitor

ODD.hv

charge

A common form of energy storage device is a parallel-plate capacitor whose
capacitance is directly proportional to the surface area of the conductor plates
and inversely proportional to the separation distance between the plates.

volts

hv^2

Capacitance

Copyright ABRAHAM [2008] - all rights reserved

Capacitance

The SI unit of capacitance is the farad (symbol: F), named after the English physicist Michael Faraday

In general any separated charges will create a emf voltage

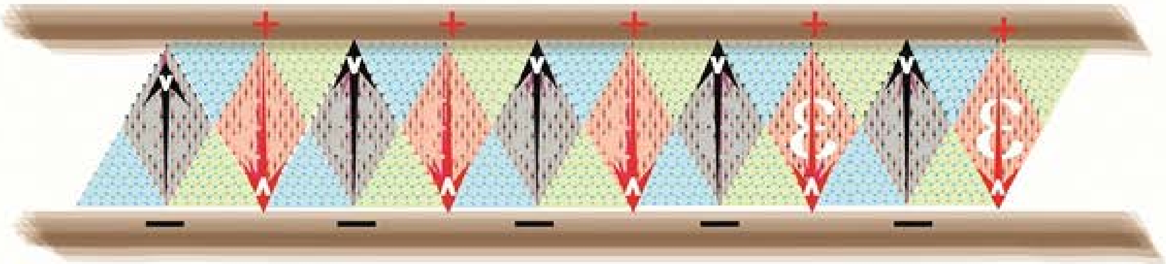

Any two electrical conductors separated by a non-conducting (or very high resistance) medium is a capacitor
[these can be plates, conductive wires or coaxial cables etc]

Plate capacitors

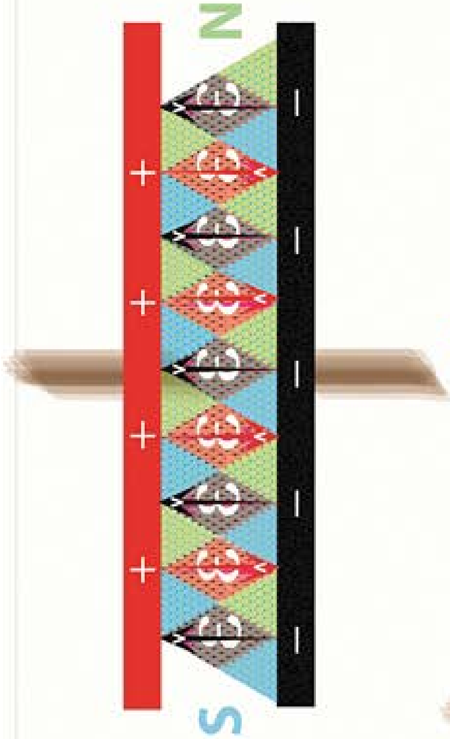

Capacitance is the ability of a body to store an electrical charge.

$$C = \frac{Q}{V}$$

Any object that can be electrically charged exhibits capacitance

Coaxial cables

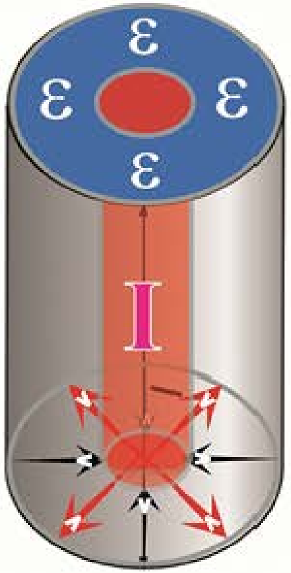

A 1 farad capacitor when charged with1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

should a current flow in the conductors the capacitive field will be dominated by orthogonal, superpositioned Amperian M-fields of greater strength

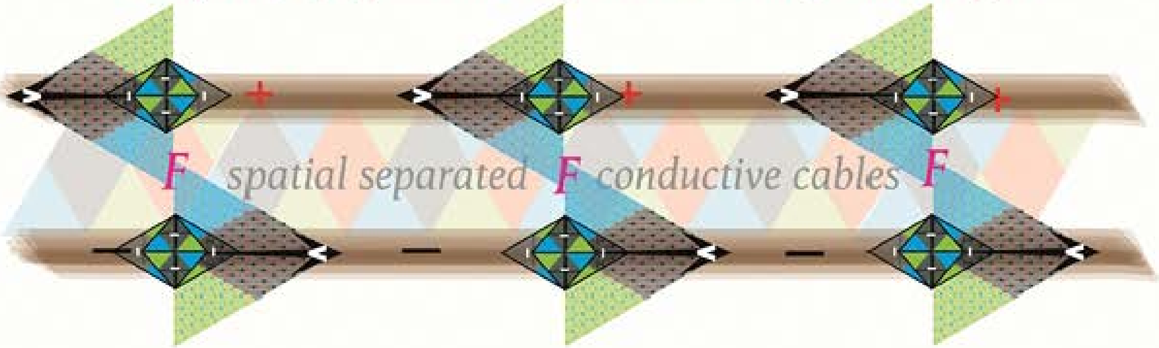

Capacitance is a function only of the physical dimensions (spatial geometry) of conductors and the permittivity of the dielectric separating them

Copyright ABRAHAM [2008] - all rights reserved

Inductive and Capacitive energy storage

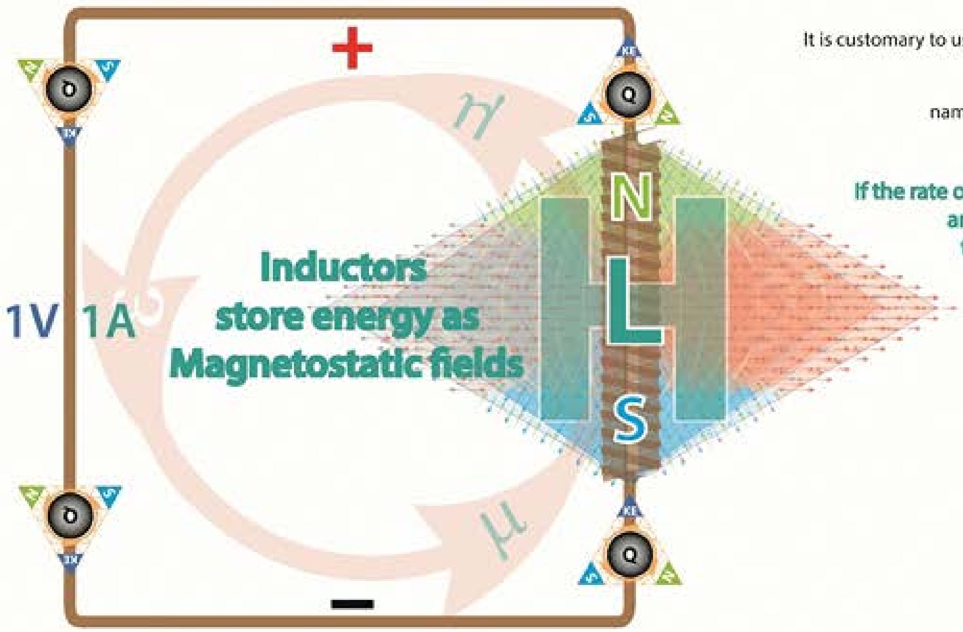

Inductors store energy as Magnetostatic fields

It is customary to use the symbol L for inductance, in honour of the physicist Heinrich Lenz.

In the SI system the unit of inductance is the Henry,
named in honor of the scientist who discovered inductance,
Joseph Henry.

If the rate of change of current in a circuit is one ampere per second and the resulting electromotive force is one volt, then the inductance of the circuit is one henry

$$v_m = \sum_{n=1}^{K} L_{m,n} \frac{di_n}{dt}.$$

The voltage across an inductor
is equal to the product of its
inductance and the time
rate of change of the current through it

Joseph Henry

[17 December 1797 – 13 May 1878]

Michael Faraday

[22 September 1791 – 25 August 1867]

The SI unit of capacitance is the farad (symbol: F),
named after the English physicist Michael Faraday

a 1 farad capacitor when charged with 1 coulomb of electrical charge will have a potential difference of 1 volt between its plates.

$$W_{\text{stored}} = \frac{1}{2}CV^2 = \frac{1}{2}\varepsilon_r\varepsilon_0\frac{A}{d}V^2.$$

Capacitance is the ability of a body to store an electrical charge.
Any object that can be electrically charged exhibits capacitance.

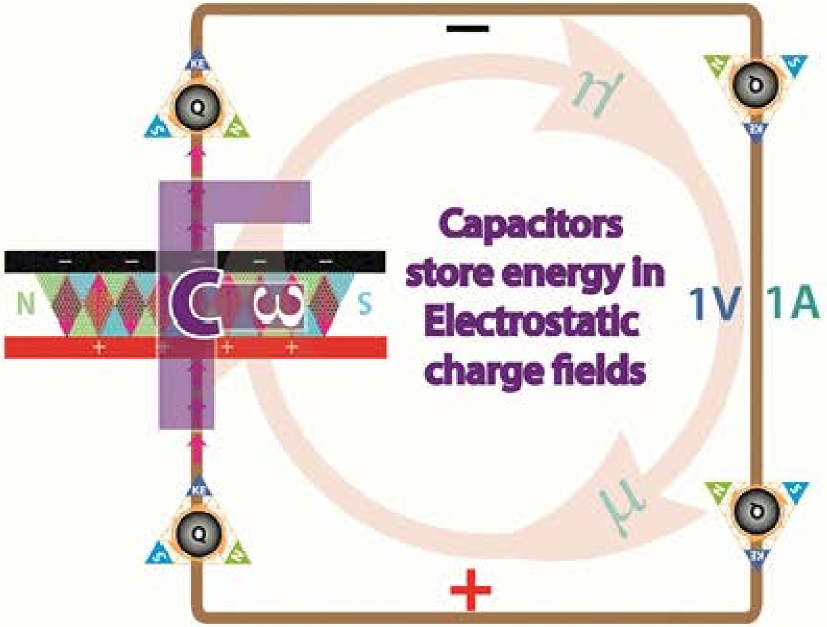

Capacitors store energy in Electrostatic charge fields

Copyright ABRAHAM [2008] - all rights reserved

Inductive - Capacitive circuits

An LC circuit, also called a resonant circuit, tank circuit, or tuned circuit,
consists of an inductor, represented by the letter L,
and a capacitor, represented by the letter C.

LC

When connected together, they can act as an electrical resonator,
an electrical analogue of a tuning fork, storing energy
oscillating at the circuit's resonant frequency

These LC circuits are idealised models
assuming there is no dissipation of energy due to resistance over time,
with the circular EM vectors representing electron current flow in the circuit
and its translation between inductive and capacitive forms of energy storage.

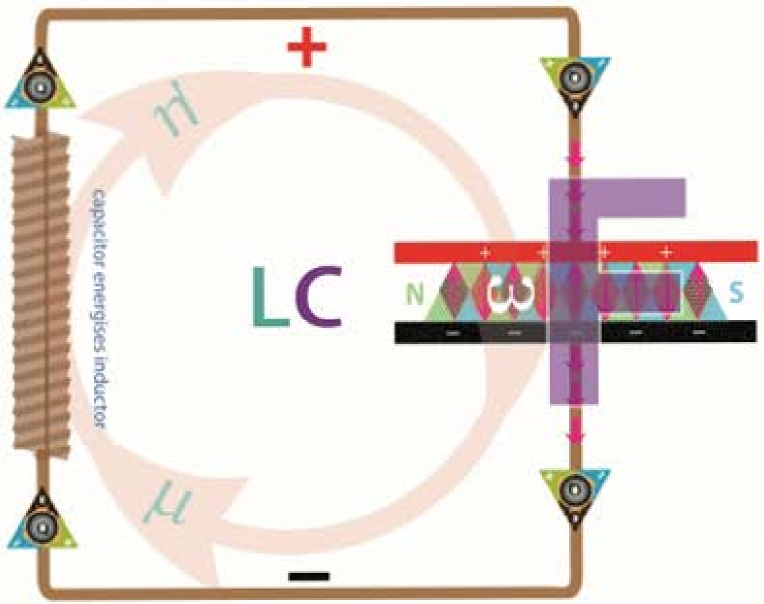

**If a charged capacitor is connected across an inductor,
charge will start to flow through the inductor,
building up a magnetic field around it and
reducing the voltage on the capacitor.**

**Eventually all the charge on the capacitor will be gone
and the voltage across it will reach zero.**

**However, the current will continue,
because inductors resist changes in current.**

**The energy to keep it flowing is extracted from the magnetic field,
which will begin to decline.**

**The current will begin to charge the capacitor
with a voltage of opposite polarity to its original charge.**

**When the magnetic field is completely dissipated the current will stop
and the charge will again be stored in the capacitor,
with the opposite polarity as before.**

**Then the cycle will begin again,
with the current flowing in the
opposite direction through the inductor.**

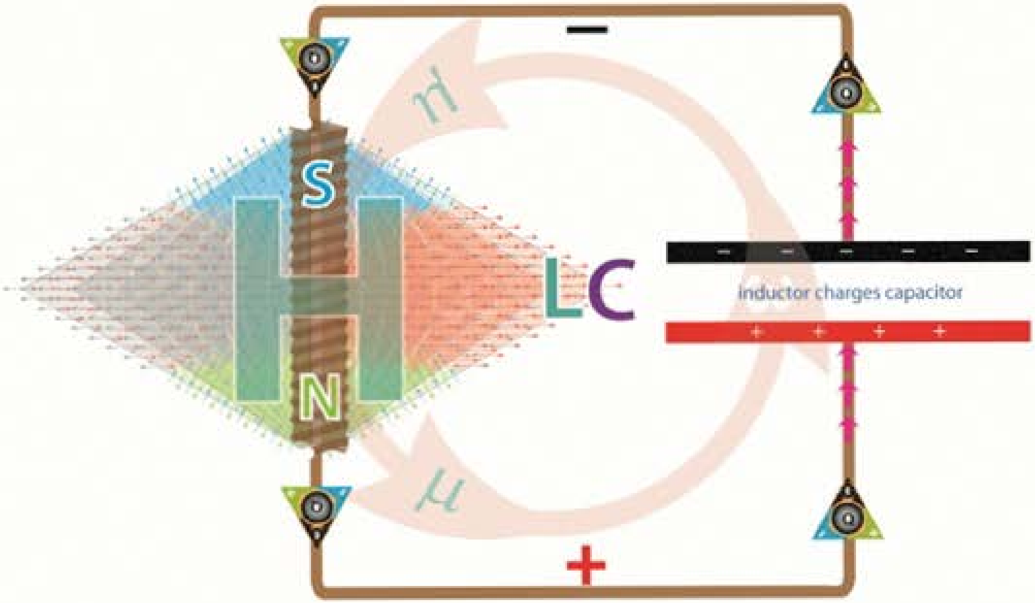

Copyright ABRAHAM [2008] - all rights reserved

Capacitors store EM Energy as transverse Bosons

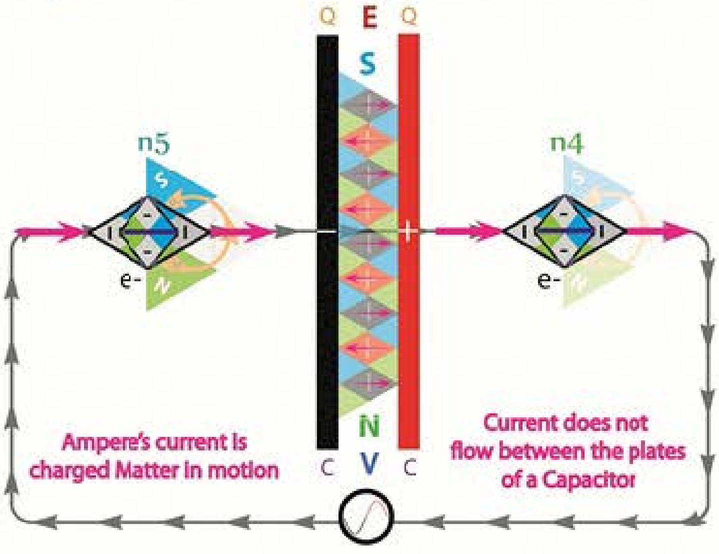

Ampere's current is charged Matter in motion

Current does not flow between the plates of a Capacitor

Capacitors store Energy as Joules per charge

Capacitors store & pass energy as charged masses

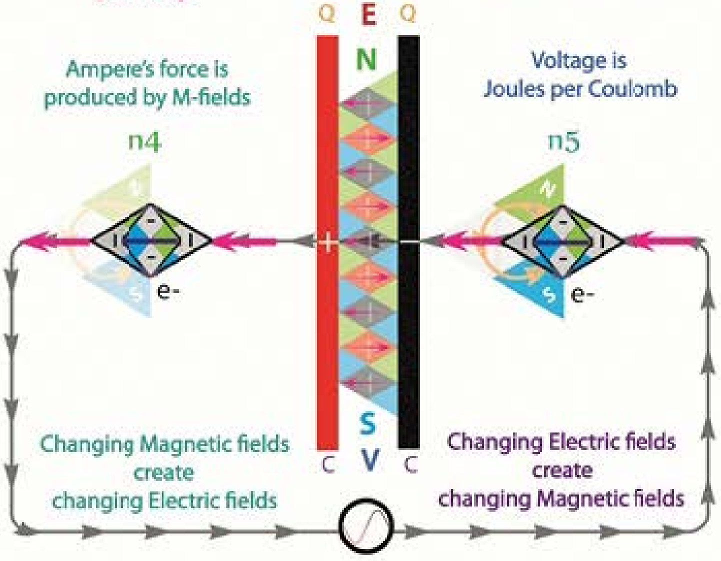

Ampere's force is produced by M-fields

Voltage is Joules per Coulomb

Changing Magnetic fields create changing Electric fields

Changing Electric fields create changing Magnetic fields

Capacitors store charged masses between their plates

Displacement Current

Maxwell's displacement current is better termed a 'displacement voltage' in light of the fact the charged Matter does not move/flow between the charged plates of a capacitor only the charged mass-energies.

Changing Electric fields produce changing Magnetic fields and vice versa as a direct result of their respective E&M field geometries within the Planck quanta of EM fields

$$\nabla \times \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} + \mu_0 \mathbf{J},$$

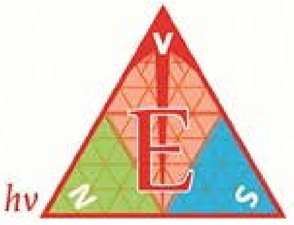

As the energy content of a charge increases so does its voltage

As opposed to Maxwell's view of a 'stretching ether' that stores energy between the capacitive plates

Equilateral energy momenta is stored compressively in the inductive quantum fields [ZPF] as it increases, and is released via charged exchange W-Bosons

Producing a capacitive region of ElectroMagnetic fields and stress energies

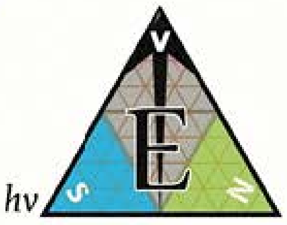

Bosons are transverse stores of energy momenta

$$\mathbf{J}_D = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

'Displacement Currents' between charged surfaces are produced by a increasing Electric Field component that seeks equilibrium

It must be carefully borne in mind that we have made only one step in the theory of the action of the medium. We have supposed it to be in a state of stress, but we have not in any way accounted for this stress, or explained how it is maintained. This step, however, seems to me to be an important one, as it explains, by the action of the consecutive parts of the medium, phenomena which were formerly supposed to be explicable only by direct action at a distance.

Maxwell - On Physical Lines of Force (1861)

mass

$$mc^2 = E = Mc^4$$

Matter

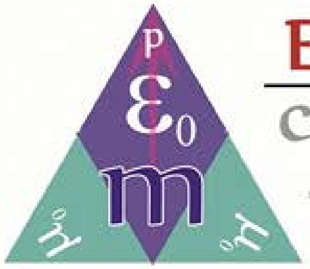

$$\frac{E}{c^2} = m$$

mass-energy momenta in a 2D equilateral geometry

Radiant waveform
EM mass-energies

A clear geometric distinction can now be made between **EM mass geometries** and standing wave **Matter topologies**

$$M = \frac{E}{c^4}$$

Four mass-energy momenta in a 3D tetrahedral topology

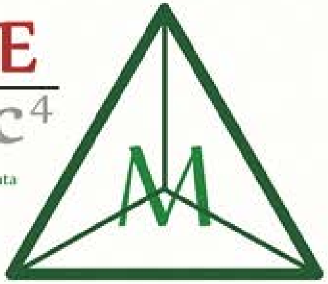

Standing waveform
EM energies

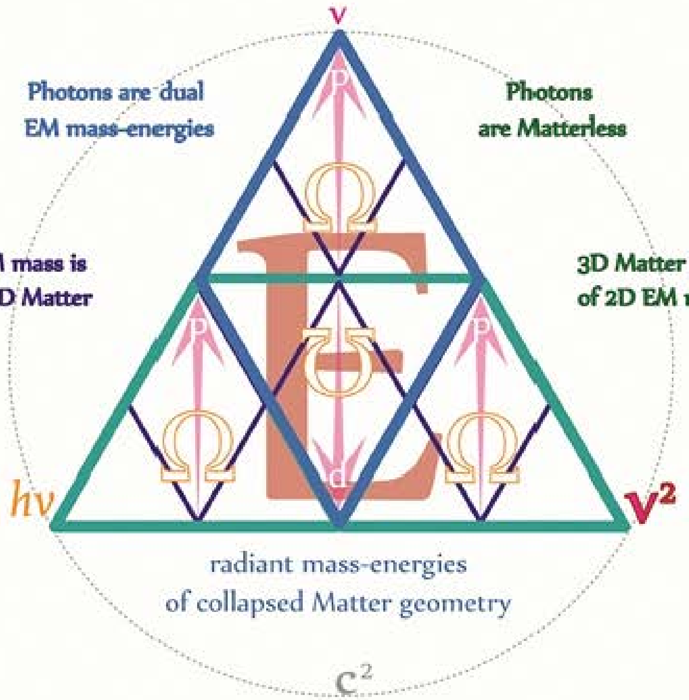

Photons are dual EM mass-energies

Photons are Matterless

2D EM mass is NOT 3D Matter

3D Matter is comprised of 2D EM mass-energies

radiant mass-energies
of collapsed Matter geometry

$$\underset{\text{mass}}{n\pi}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

$$\underset{\text{Matter}}{T\pi}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\underset{\text{velocity}}{\overset{\text{quanta}}{v^2}}\right]\right]$$

ρ/second

ρ/second^2

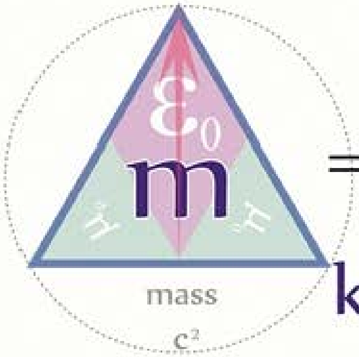

$$= m\Omega\left[\frac{v^2}{c^2}\right]$$

kg

Radiant EM masses

$$\frac{4\pi m}{c^2} = M$$

Standing-wave mass-Matter

$$m\Omega\left[\frac{v^2}{c^4}\right] =$$

KG

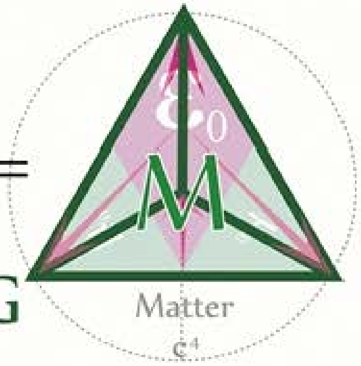

massless particles are a scientific misnomer and massless must now be re-termed Matterless

de Broglie Matter-waves

EM CHARGE
provides the foundational
interaction geometry for
all mass & Matter

In quantum mechanics, a Matter wave or de Broglie wave
is the probabilistic electromagnetic quantum wave-form
created by the Planck elements constituting Matter

Relativistic mass-energy
is the total sum of all Planck
energy momenta in any
spatial co-ordinate
system

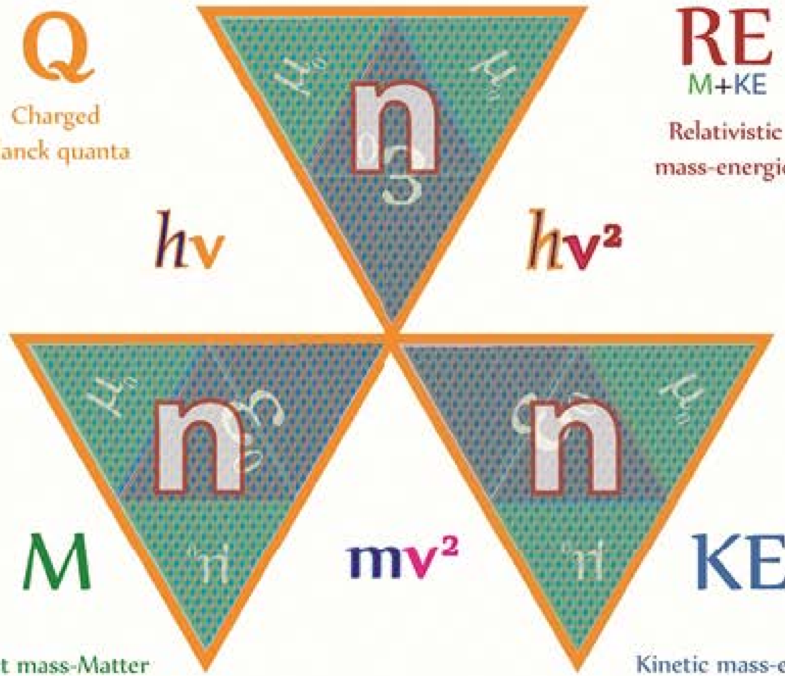

Q
Charged
Planck quanta

RE
M+KE
Relativistic
mass-energies

$h\nu$

$h\nu^2$

M
rest mass-Matter
topologies

mv^2

KE
Kinetic mass-energy
geometries

All EM waves and Matter
are made up of integral Planck quanta
[any imbalance results in Charge]
as the number of Planck quanta increases
their associated wavelengths decrease

The relativistic stress tensor
mass-energies of Matter in motion
is the sum of its standing-wave mass-
Matter topology and its Lorentz corrected
kEM field mass-energy geometries

3D MATTER
is the charged topology
created by standing-waves
of 2D mass-energies

Kinetic EM mass-energy levels depend on
the velocity of the Matter particle in motion
[in Matter energy travels in a standing wave at the speed of Light]

Kinetic EM fields
are the mass-energy
fields created by
Matter in motion

P
KE
N
S
magnetic moment

The rest mass-energy geometries comprising all Matter topologies are velocity invariant

Copyright ABRAHAM [2008] - all rights reserved

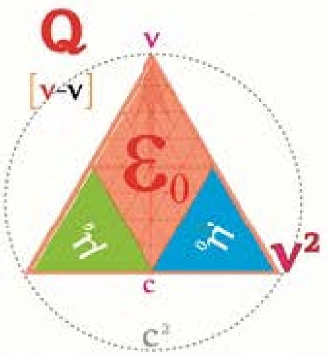

Material EM masses

deBroglie wavelength & Compton frequency

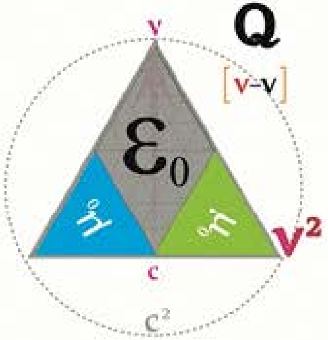

λ

Any increase in Energy quanta results in
a decrease in EM mass quanta wavelengths

v^2

In Matter all energy propagates at c

$$mv^2 = E = hv^2$$

in Standing wave Matter
v=c

Number of Quanta comprising rest Matter

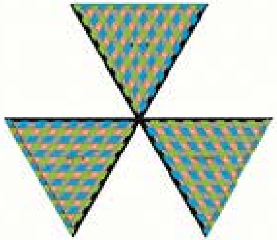

Compton Frequency

$$\frac{mc^2}{h} = 1x10^{19}$$

Electron Compton Frequency

$= 1.2 \times 10^{20}$ quanta

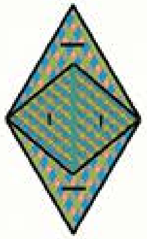

v^2

Wavelength of Quanta comprising rest Matter

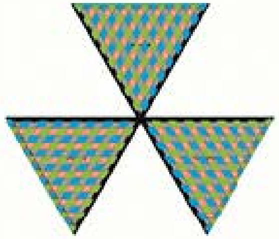

de Broglie wavelength

$$\frac{h}{mc} = 2.99792456 \times 10^{-33} m$$

Electron de Broglie wavelength

$= 4.002769142 \times 10^{-11} m$

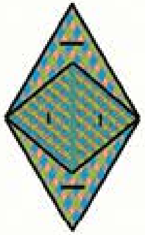

λ

Linear momentum of standing wave Energy comprising Matter

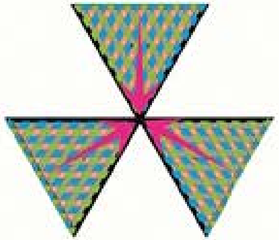

Intrinsic Particle Momenta

$$mc = 2.211340633 \times 10^{-23} Ns$$

Intrinsic Electron momenta

$= 2.95253793 \times 10^{-30} Ns$

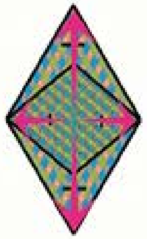

p

The Compton frequency, deBroglie wavelength
and Energy momenta of any physical system
are all related through the speed of Energy [c]

**The examples above are for stationary Electrons
any motion will NOT affect the results for the electron itself
(however a extended KEM field will be produced by the motion
and its properties will be affected by changes in velocity)**

In Lorentz invariant Matter the standing wave
Energy always propagates at the speed of Light [c]
(with the KEM field subject to Lorentz corrections)

Tetryonics 25.03 - Material deBroglie-Compton properties

KEM Quantum levels

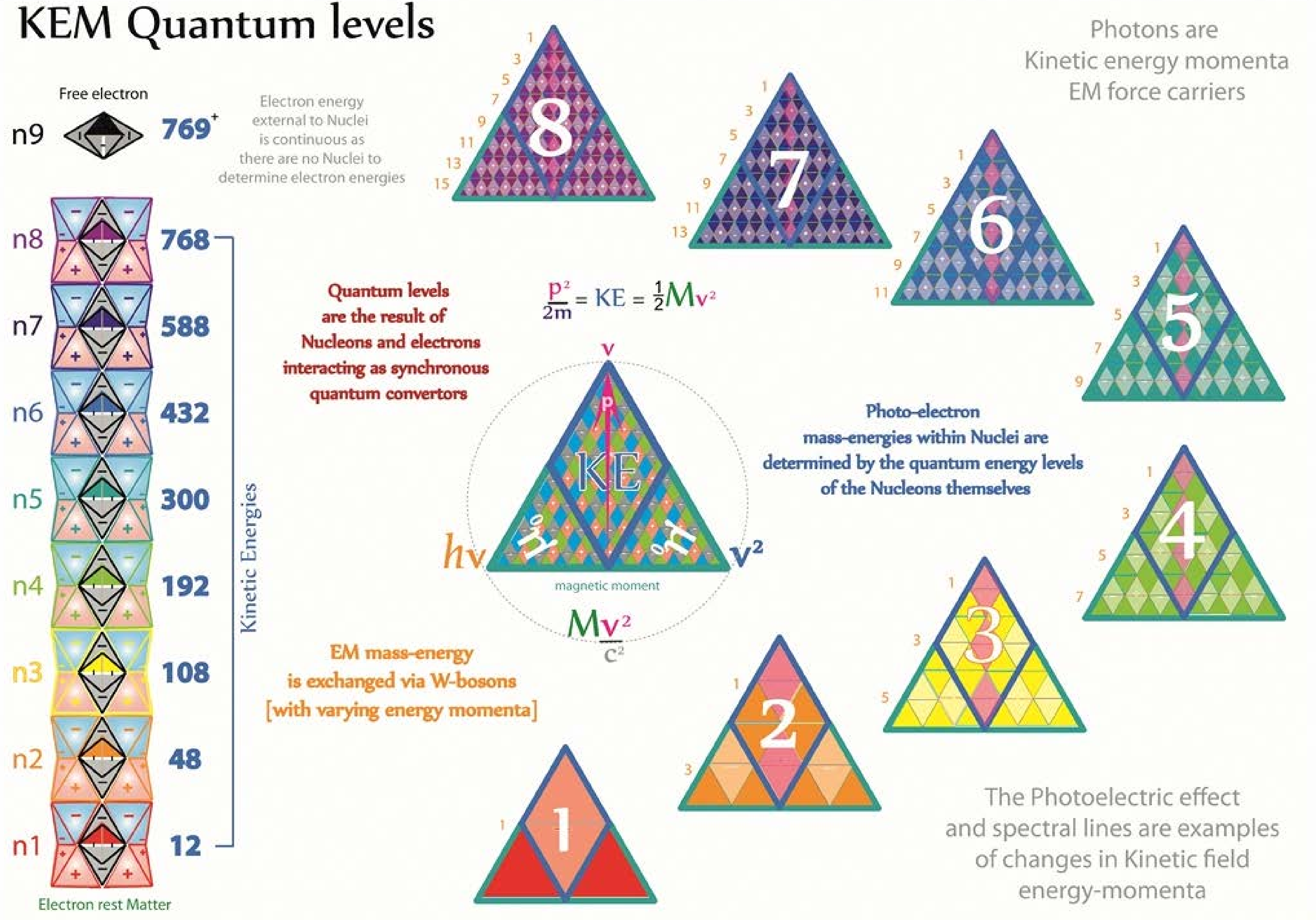
Free electron
n9 769+
n8 768
n7 588
n6 432
n5 300
n4 192
n3 108
n2 48
n1 12
Kinetic Energies
Electron rest Matter
8
7
6
5
4
3
2
1
v
p
KE
hv
v²
magnetic moment
Mv²/c²

Electron energy external to Nuclei is continuous as there are no Nuclei to determine electron energies

Photons are Kinetic energy momenta EM force carriers

Quantum levels are the result of Nucleons and electrons interacting as synchronous quantum convertors

$$\frac{p^2}{2m} = KE = \tfrac{1}{2}Mv^2$$

Photo-electron mass-energies within Nuclei are determined by the quantum energy levels of the Nucleons themselves

EM mass-energy is exchanged via W-bosons [with varying energy momenta]

The Photoelectric effect and spectral lines are examples of changes in Kinetic field energy-momenta

Tetryonics 25.04 - KEM quantum levels

deBroglie relationships

*" The electron which is moving in a sine wave circular path,
will repeat the same sine wave path in each successive orbit.
The sine wave paths in consecutive orbits will exactly overlap.
The electron wave reconnects with itself and is in phase with itself. "*

An electron bound in a atomic nuclei will trace
a toroidal path around the nuclei mapped
by the sine wave motion of any point
on its charged geometric surface

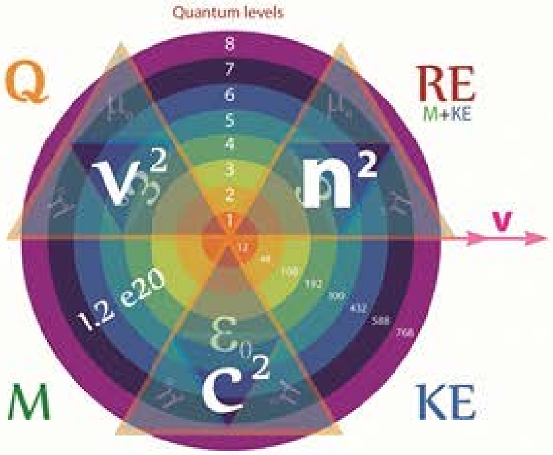
Quantum levels
8
7
6
5
4
3
2
1
Q
RE
M+KE
v²
n²
v
1.2 e20
ε0
c²
M
KE
12
48
108
192
300
432
588
768

**rest mass-Matter is
velocity invariant**

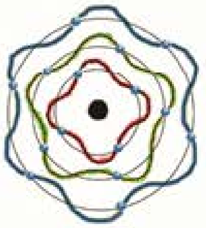

$$\lambda = \frac{h}{p}$$

$$f = \frac{E}{h}$$

Louis de Broglie

(15 August 1892 – 19 March 1987)

$$p = \frac{h}{\lambda} = \frac{h}{2\pi}\frac{2\pi}{\lambda} = \hbar k$$

Although Mathematically correct
the illustrative model commonly used to
depict the deBroglie relationships
is incorrect

***The relativistic energy of Matter in motion
is a measure of its standing wave mass-energies,
and the object's velocity related kinetic energy and
magnetic moment in the form of a KEM wave geometry***

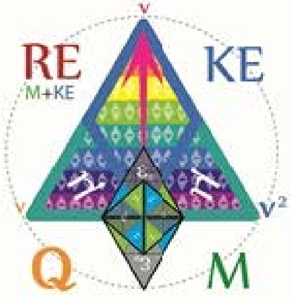
v
RE
M+KE
KE
v
v²
Q
M

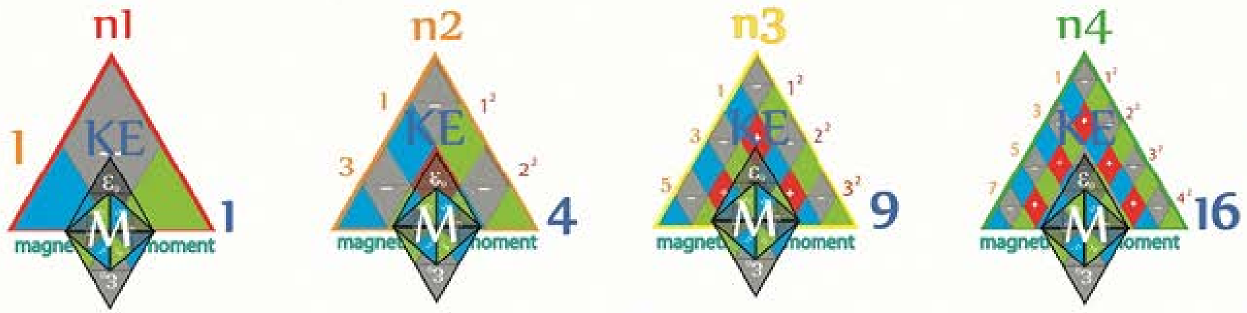
n1
1
KE
1
magnetic moment
M
n2
1
1²
3
2²
KE
4
magnetic moment
M
n3
1
1²
3
2²
5
3²
KE
9
magnetic moment
M
n4
1
1²
3
2²
5
3²
7
4²
KE
16
magnetic moment
M

**The greater the
KEM field energy level
the smaller the
quantum wavelength
of the KEM field's EM mass**

**All mass-Matter topologies are
Lorentz velocity invariant,
only the EM mass-energy content
and quantised angular momentum
of the Kinetic EM fields vary**

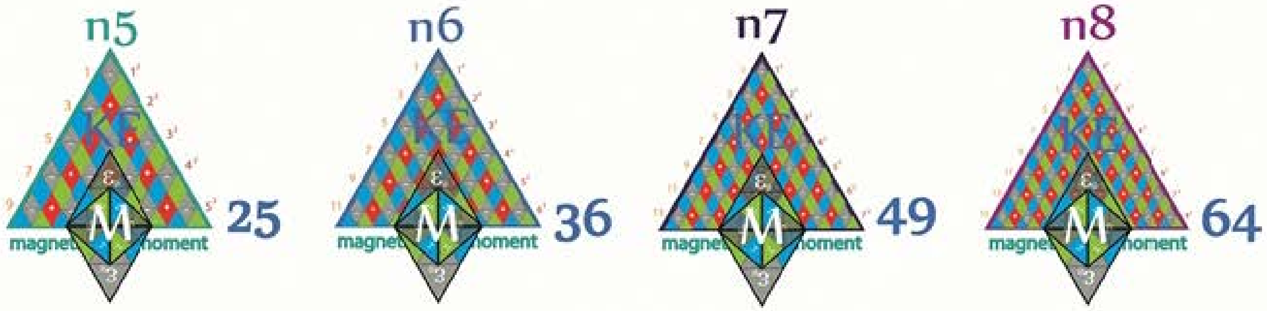
n5
KE
25
magnetic moment
M
n6
KE
36
magnetic moment
M
n7
KE
49
magnetic moment
M
n8
KE
64
magnetic moment
M

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Electron Positions in Atomic Orbitals

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over space, categorized by n, l, and m quantum numbers, [as covered in Tetryonic Chemistry] which correspond to the electron's energy, angular momentum, and an angular momentum vector component, respectively

Using Tetryonics as a EM field model, an electron's position and velocity CAN be modelled simultaneously (but any attempt to measure or interact with it, will affect its energy levels)

Quantum Mechanics is a statistical [mathematical] representation of equilateral, charged geometries and EM energy interactions

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

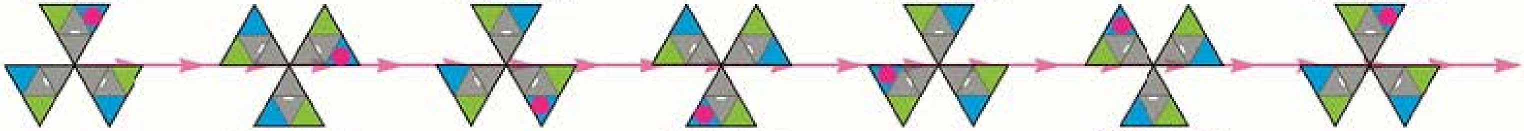

The unique 12 facet topology of Leptons results in the same geometry being presented every 120° rotation of the particle

Additionally, every radial arm of the Lepton's topology is identical to every other

Including Electric fields and Magnetic dipole orientations

Leading to the interpretation that the Lepton disappears and re-appears when being measured

Tetryonics resolves these issues with an accurate 3D model of Leptons

Making accurate measurement of its rotational dynamics incorrect without correct models

And Quantum Levels thus providing a solid foundation for explaining Electro-Dynamics

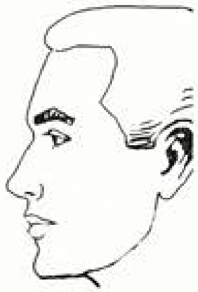

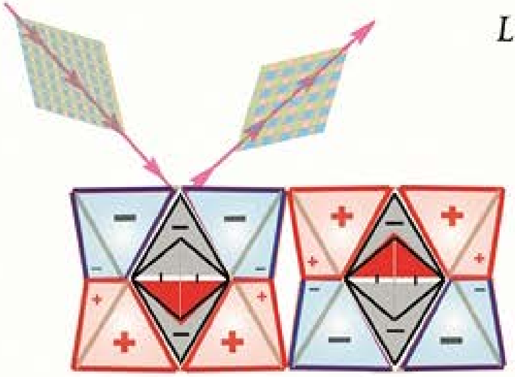

Leptons are 6 loop quantum loop inductive rotors

Lepton's are Spin 1 particles that can easily be misinterpreted as 1/2 or 3 quantum spin numbers without the correct physical models to interpret physcical observations on

The charged facia presented changes continously, with a complete fascia change occuring in every 30 degrees of rotation

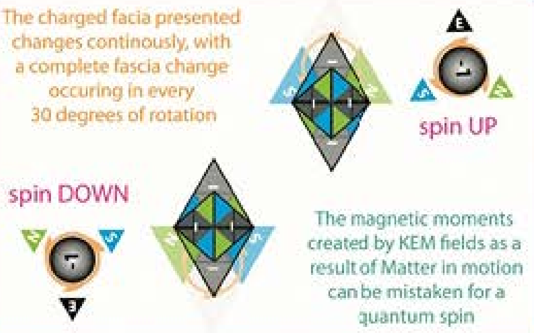

The magnetic moments created by KEM fields as a result of Matter in motion can be mistaken for a quantum spin

Copyright ABRAHAM [2008] - all rights reserved

Distributions and Uncertainty

A Bell Curve (Normal Distribution) is a mathematical reflection of the integral Planck mass-energy momenta quanta distributions found in all equilateral charge geometries

Probabilities are the square of the Amplitude

All probabilities are re-normalisable and sum to Unity

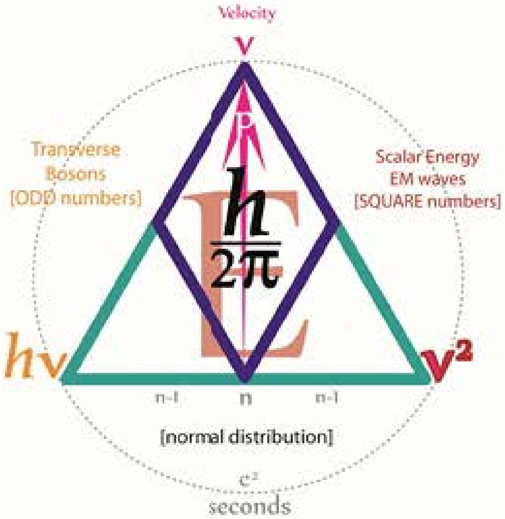
Velocity
v
P
Transverse Bosons [ODD numbers]
Scalar Energy EM waves [SQUARE numbers]
E
$\frac{h}{2\pi}$
hv
v²
n-1
n
n-1
[normal distribution]
c²
seconds

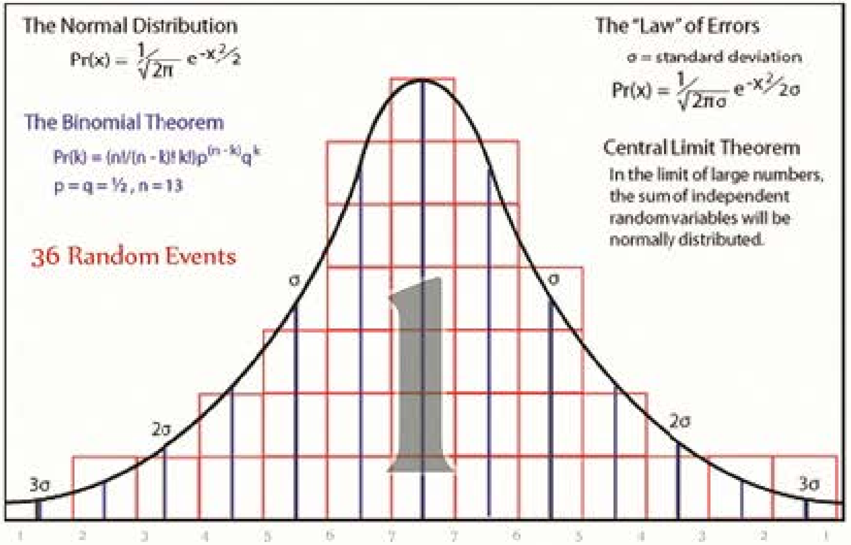

The Normal Distribution

$Pr(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

The Binomial Theorem

$Pr(k) = (n!/(n-k)!\,k!)p^{(n-k)}q^k$

$p = q = ½, n = 13$

The "Law" of Errors

σ = standard deviation

$Pr(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-x^2/2\sigma}$

Central Limit Theorem

In the limit of large numbers, the sum of independent random variables will be normally distributed.

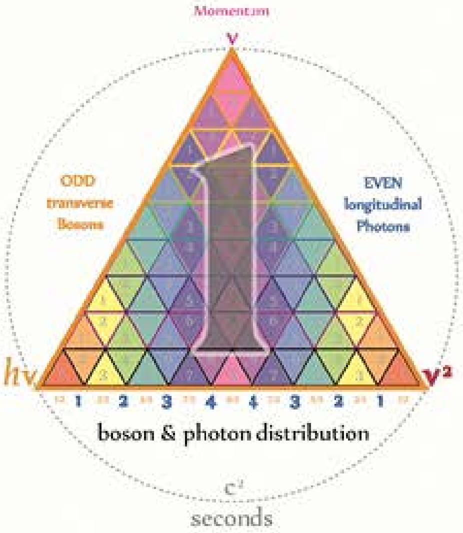
Momentum
v
ODD transverse Bosons
EVEN longitudinal Photons
1
hv
v²
1 2 3 4 4 3 2 1
boson & photon distribution
c²
seconds

All Matter and EM fields are comprised of equilateral mass-energy momenta and any attempt to measure the system involves the introduction of additional Planck EM quanta into the systems

John Stewart Bell

(28 June 1928 – 1 October 1990)

No physical theory of local hidden variables can reproduce all of the predictions of quantum mechanics.

WRONG

$$\Delta p \, \Delta x \geq \frac{1}{2} \hbar$$

Quantum uncertainty formulation

$$\Delta E \, \Delta t \geq \frac{1}{2} \hbar$$

The more precisely the position is determined, the less precisely the momentum is known in this instant, and vice versa.

Heisenberg, uncertainty paper, 1927

WRONG

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Chance is closely related to the ideas of uncertainty and indeterminacy.
Uncertainty today is best known from Werner Heisenberg's principle in quantum mechanics.
It states that the exact position and momentum of an atomic particle can only be known within certain limits.
The product of the position error and the momentum error is equal to a multiple of Planck's constant of action.
This irreducible randomness in physical processes established the existence of chance and indeterminism in physics

Copyright ABRAHAM [2008] - all rights reserved

Leptronic Self-energies

As the velocity of a charged particle increases
the energy level of its KEM field increases

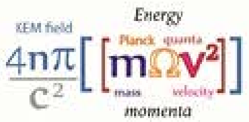
KEM field
Energy
Planck
quanta
mass
velocity
momenta

$$\frac{4n\pi}{c^2}\left[\left[m\Omega v^2\right]\right]$$

mass

radiant energy per second

$$E/c^2$$

KEM field geometries generated as a product of Matter in motion are Lorentz variant

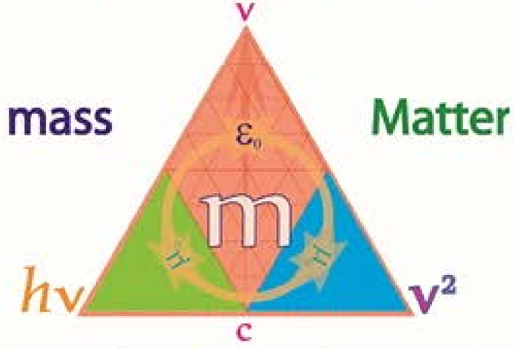
v
mass
Matter
ε_0
m
hv
v^2
c

Quantum inductive loops resist changes to energies contained in their equilateral energy fields [inertial mass]

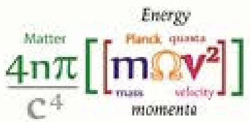
Matter
Energy
Planck
quanta
mass
velocity
momenta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

Matter

standing-wave energy per second squared

$$E/c^4$$

rest mass-Matter topologies are not affect by changes in velocity [Matter is Lorentz invariant]

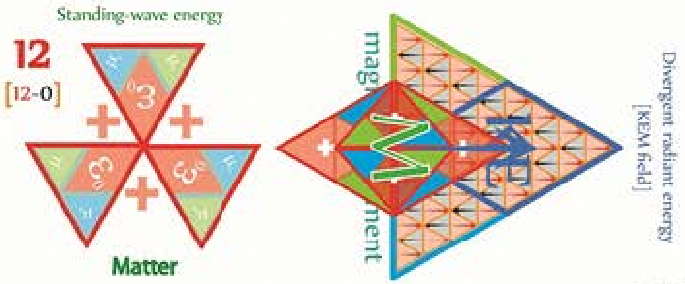
Standing-wave energy
12
[12-0]
+
+
+
ε_0
Matter
mag
nent
M
KEM
Divergent radiant energy
[KEM field]

The Law of Interaction

The Energy momenta of an electron's KEM field polarises the region surrounding its Matter and creates fields of interaction through the super-positioning of these fields

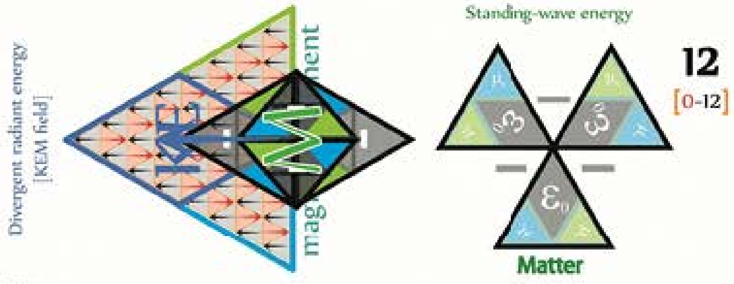
Divergent radiant energy
[KEM field]
KEM
M
nent
mag
Standing-wave energy
12
[0-12]
–
–
–
ε_0
Matter

a particle's self-energy represents the contribution to the particle's energy, or effective mass, due to interactions between the particle and the system it is part of.

$$KEM = Mv^2$$

All Matter topologies are made up of quantum inductive loops which extract energy from any EM field they move through

Matter stores energy in its 3D planar facsia as charged mass-energy

additional to creating the familiar laws of interaction [opposite attract - similars repel] the KEM fields of leptons can act as tuned antennas extracting energy from their environment

KEM fields store energy in their 2D planar EM field as neutral divergent mass-energies

$$E = Mc^4$$

All Matter contains EM energy propagating in a standing wave topology at the speed of light (the source of inertial mass)

Resonant EM fields

Transmitting in order to efficiently recieve
via resonant inductive coupling
of EM fields

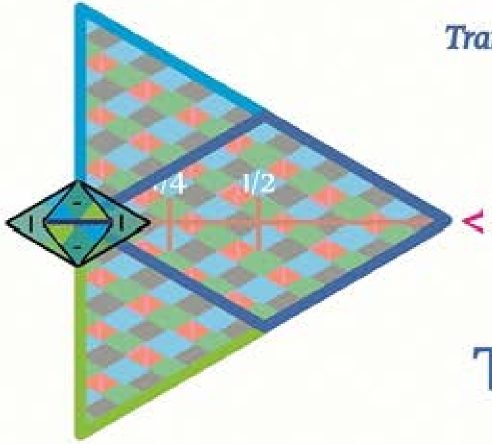

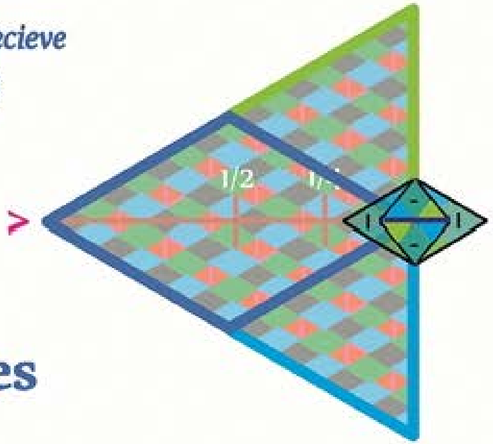

1:1

[or other harmonics]

Tuned KEM waves

Quantum inductive antenna fields

Historically, all recieiving antennas have been mechanically tuned to match the sought EM wavelength
But if power is applied to a small wavelength electrostatic topology it creates a KEM field
that produces a much larger, tuned rhombic antenna field

EM waves are hf geometries

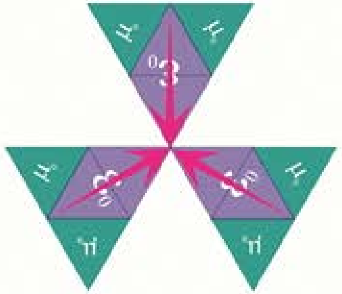

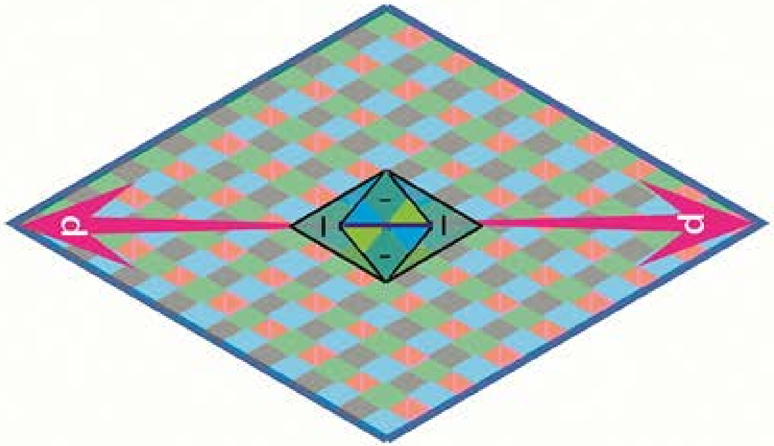

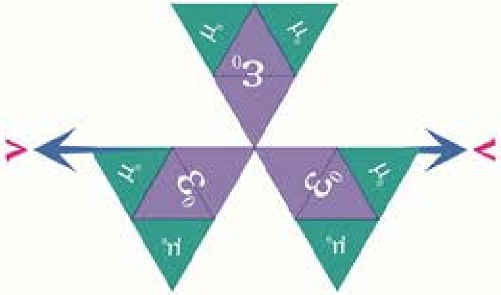

Electron's rest Matter
stores EM mass-energy
in a standing wave

EM bosons have hv geometries

Electrostatic particles of Matter
have dual KEM energy fields that
can act as resonant antennas

Bosons

All Light is made of transverse EM energy Quanta

EM Field Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons ElectroMagnetic mass velocity

Energy is quantised

Max Planck

$$E = n.hv$$

All Photons have mass-Energy and momentum

$$\lambda = \frac{h}{p} = \frac{h}{mv}\sqrt{1-\frac{v^2}{c^2}}$$

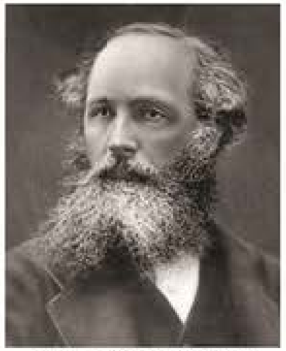

James Clerk Maxwell

$$\nabla\cdot\mathbf{D} = 4\pi\rho$$

$$\nabla\times\mathbf{H} = \frac{4\pi}{c}\mathbf{J} + \frac{1}{c}\frac{\partial\mathbf{D}}{\partial t}$$

$$\nabla\times\mathbf{E} + \frac{1}{c}\frac{\partial\mathbf{B}}{\partial t} = 0$$

$$\nabla\cdot\mathbf{B} = 0$$

ElectroMagnetic waveforms propagate at the speed of light

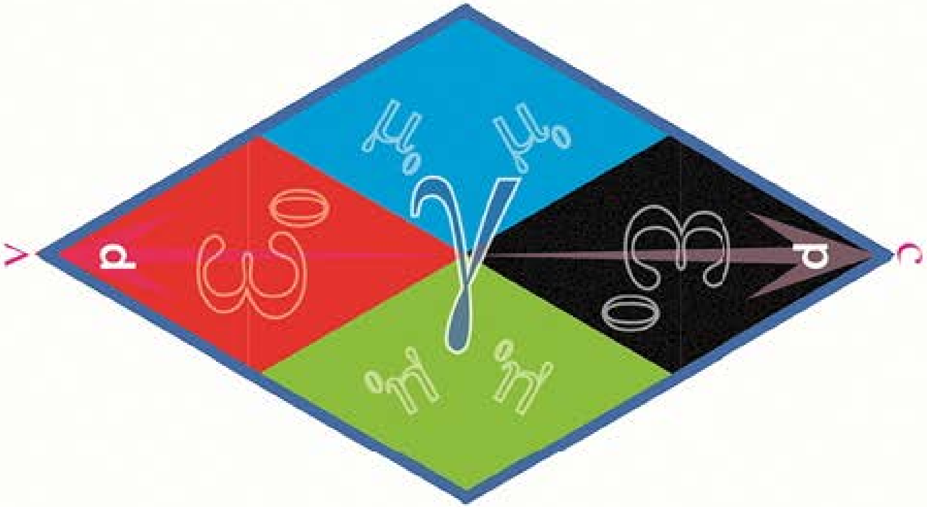

Matter is a probabilistic wave of EM particles

p

Louis de Broglie

Erwin Schrodinger

Ψ

Photons

All Light consists of longitudinal harmonic oscillators

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

Photons are particles

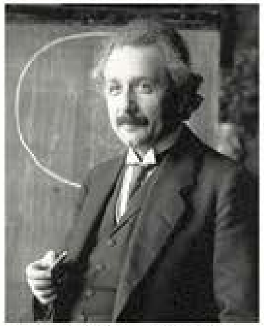

Albert Einstein

$$E = hf$$

All EM waves and Matter exhibit a Wave-Particle duality

$$i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi$$

Copyright ABRAHAM [2008] - all rights reserved

Photonic EM fields

In 1865, James Clerk Maxwell's prediction that light was an electromagnetic wave,
[which was confirmed experimentally in 1888 by Heinrich Hertz's detection of radio waves],
seemed to be the final blow to particle models of light.

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field · Planck · quanta · Photons · ElectroMagnetic · mass · velocity

James Clerk Maxwell

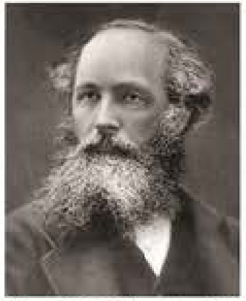

(13 June 1831 – 5 November 1879)

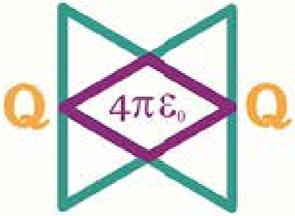

$$E = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r^2}$$

$$\frac{K_E}{K_M} = 9e16\ \frac{\frac{N\,m^2}{C^2}}{\frac{N\,s^2}{C^2}}$$

$$c^2 = 9e16\ \frac{m^2}{s^2}$$

$$c = 299{,}792{,}458\ \frac{m}{s}$$

The speed of light
can be calculated using the
Electric and Magnetic constants

8.8541e-12

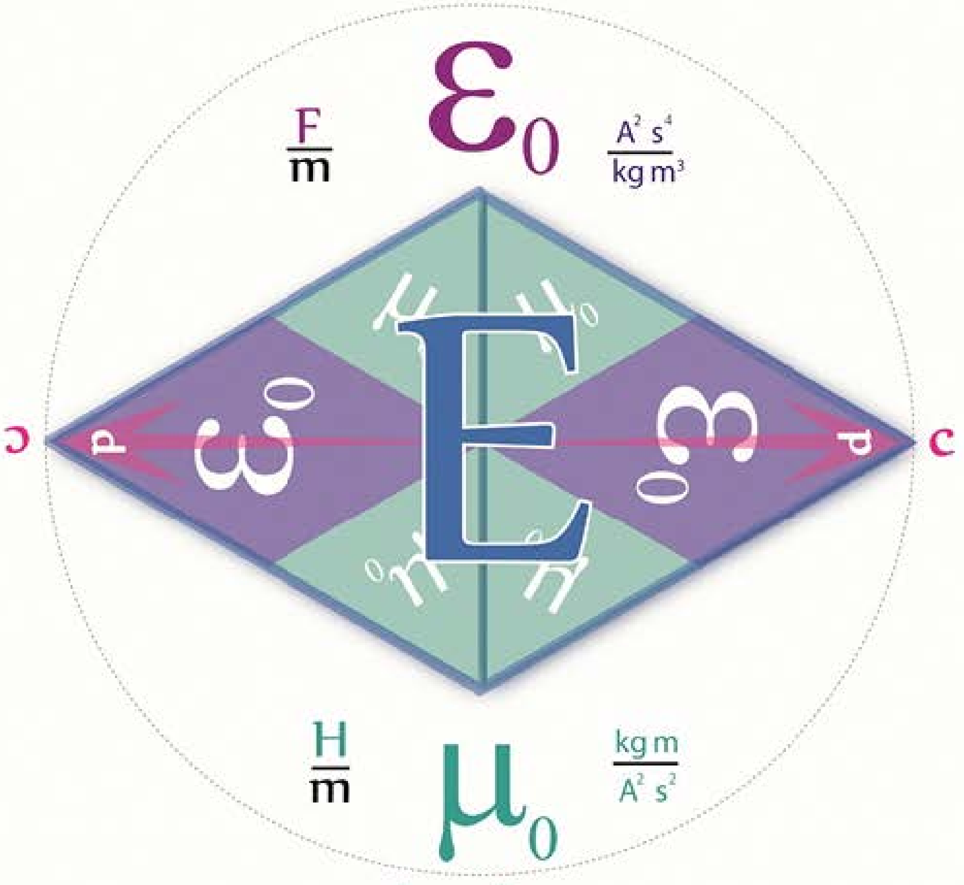

[4π]e-7

All EM mass-energy momenta
propagates bi-directionally at c
[the speed of light] in a vacuum

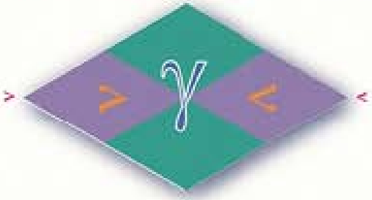

Electric field strength

The strength of Electric fields
is determined by the
Electrical Permittivity Constant

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

The strength of Magnetic fields
is determined by the
Magnetic Permeability Constant

$$B = \mu_0 H$$

The ElectroMagnetic field geometry of Photons

$$2mv^2 = KE = hf$$

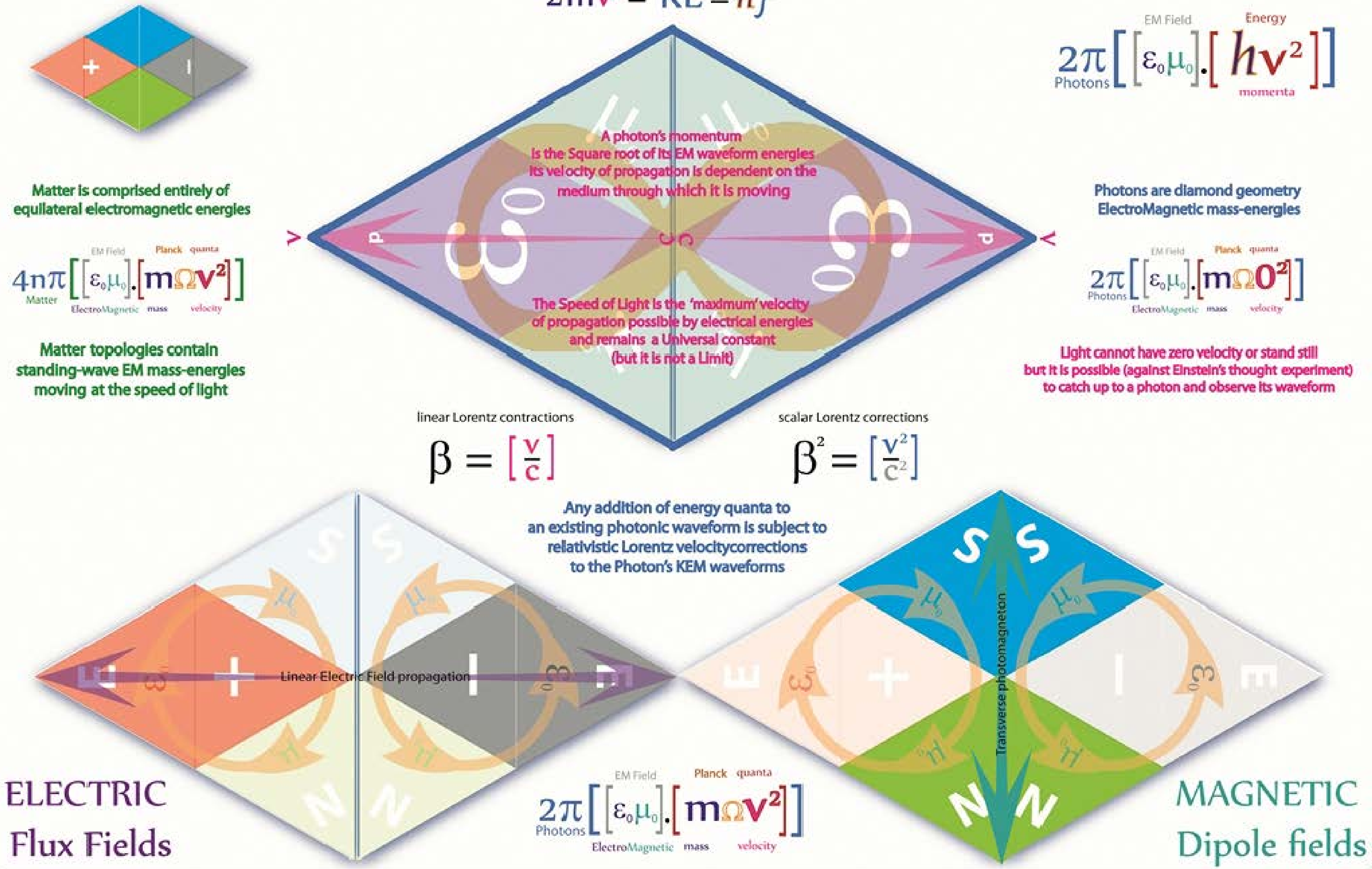

EM Field Energy

$$2\pi[[\varepsilon_0\mu_0].[hv^2]]$$

Photons momenta

A photon's momentum
is the Square root of its EM waveform energies
its velocity of propagation is dependent on the
medium through which it is moving

The Speed of Light is the 'maximum' velocity
of propagation possible by electrical energies
and remains a Universal constant
(but it is not a Limit)

Matter is comprised entirely of equilateral electromagnetic energies

EM Field Planck quanta

$$4n\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Matter ElectroMagnetic mass velocity

Matter topologies contain standing-wave EM mass-energies moving at the speed of light

Photons are diamond geometry ElectroMagnetic mass-energies

EM Field Planck quanta

$$2\pi[[\varepsilon_0\mu_0].[m\Omega 0^2]]$$

Photons ElectroMagnetic mass velocity

Light cannot have zero velocity or stand still but it is possible (against Einstein's thought experiment) to catch up to a photon and observe its waveform

linear Lorentz contractions

$$\beta = \left[\frac{v}{c}\right]$$

scalar Lorentz corrections

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

Any addition of energy quanta to an existing photonic waveform is subject to relativistic Lorentz velocitycorrections to the Photon's KEM waveforms

EM Field Planck quanta

$$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Photons ElectroMagnetic mass velocity

ELECTRIC Flux Fields

MAGNETIC Dipole fields

Tetryonics 26.03 - Photon EM field geometry

Photons and Charge bosons

Photons are dual [2π]
neutral charge quanta

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

Their Electric and Magnetic
fields are orthogonal to each other

**Bosons are transverse charge carriers
they combine to create ElectroMagnetic
photons which are longitudinal charge carriers**

They propagate
in differing mediums
at v/c

EM Field Planck quanta

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves ElectroMagnetic mass velocity

nπ electromagnetic waves are comprised of
numerous 2π photons of the same wavelength
(save in superpositioned states - White light)

Photons are dual Boson waveforms

$$E = 2hv$$

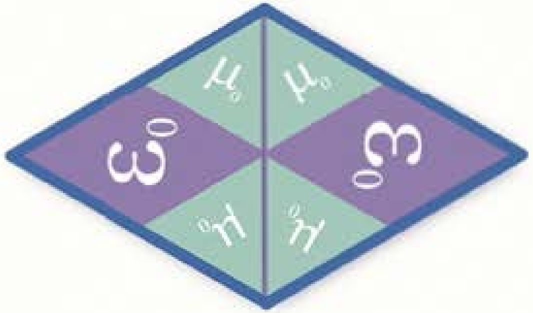

$$E = hf$$

**A moving charge creates a magnetic field throughout space
that is perpendicular to the direction of motion.**

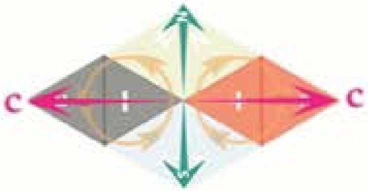

**Photons require NO ether to propagate
they are discrete bundles of EM energy-momenta**

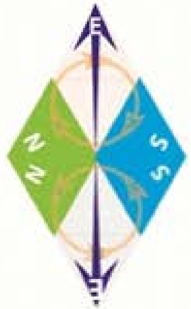

**Similarly, a magnet has an intrinsic neutral electric field
that is perpendicular to its Magnetic Dipole.**

Photon EM strength vectors

Electric field force vectors

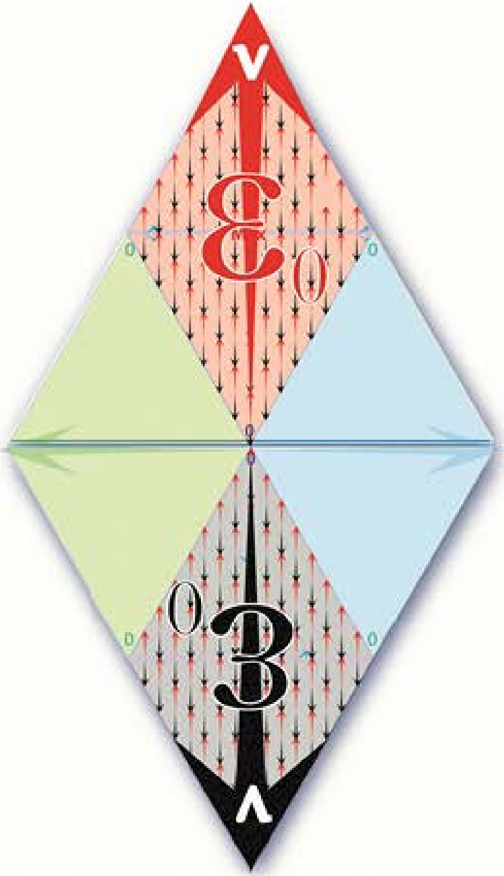

ε_0
$^0\varepsilon$
0
0
0
0

When E field flux is at Maximum
B field flux is at Minimum

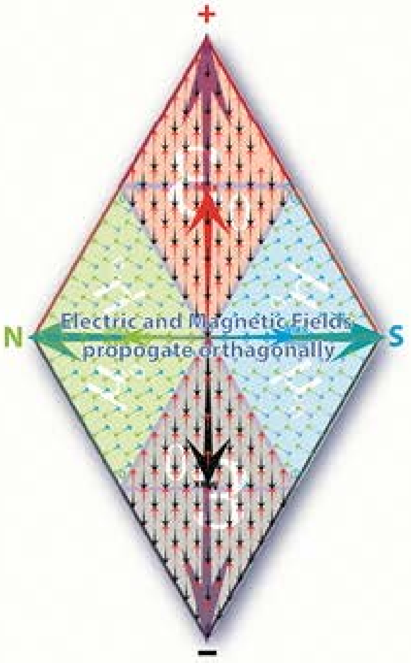

Electric and Magnetic Fields propogate orthagonally

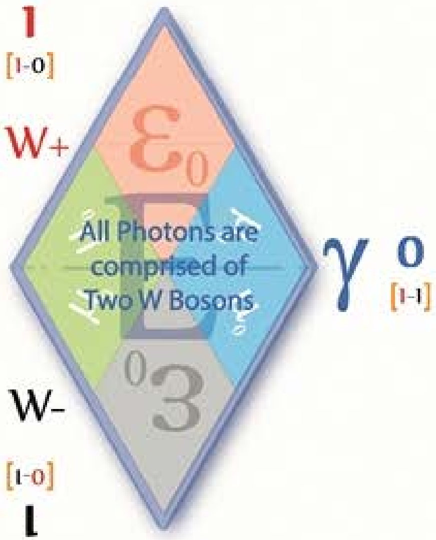

1
[1-0]
W+
ε_0
$^0\varepsilon$
W-
[1-0]
1
γ
0
[1-1]

All Photons are comprised of Two W Bosons

Magnetic field force vectors

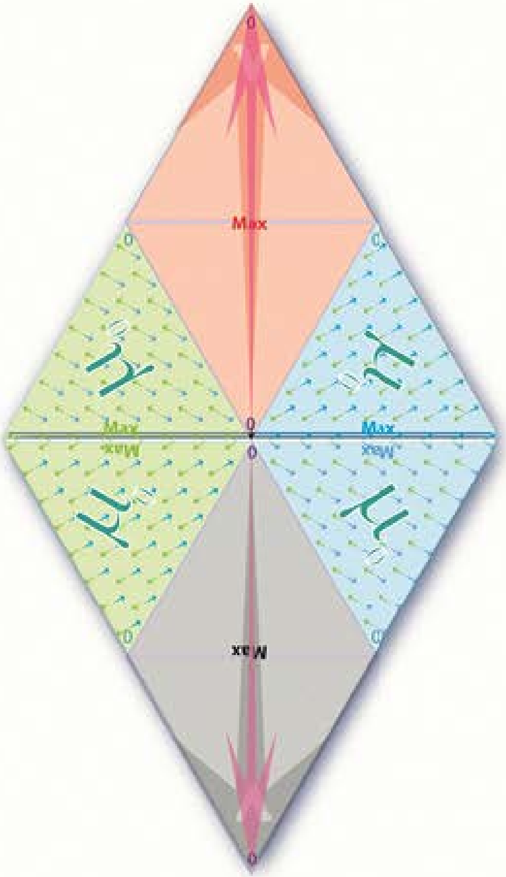

Max
Max
Max
Max
Max
Max
0
0
0
0
0

When B field flux is at Maximum
E field flux is at Minimum

Copyright ABRAHAM [2008] - all rights reserved

Photons are
2π electromagentic masses
possessing:

Frequency
Wavelength
Energy
Momentum
Kinetic Energy
Magnetic Moment
Probabilistic Properties

and can

Refract
Deflect and
Disperse

ZPFs [Bosons] are
1/2 wavelength photons

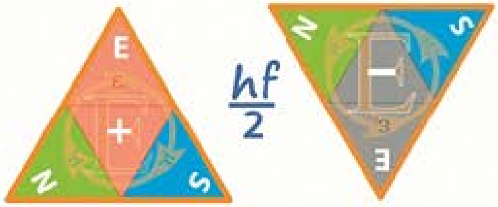
E
+
N
S
$\frac{hf}{2}$
N
S
E
E

$$2v = f$$

2π Planck Constants
Photon energy density

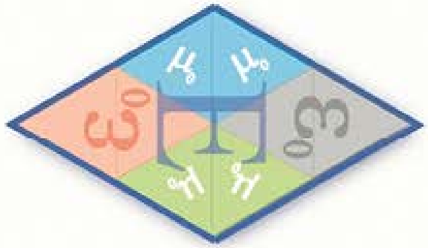
μ_0
μ_0
ε_0
ε_0
μ_0
μ_0

$$hf$$

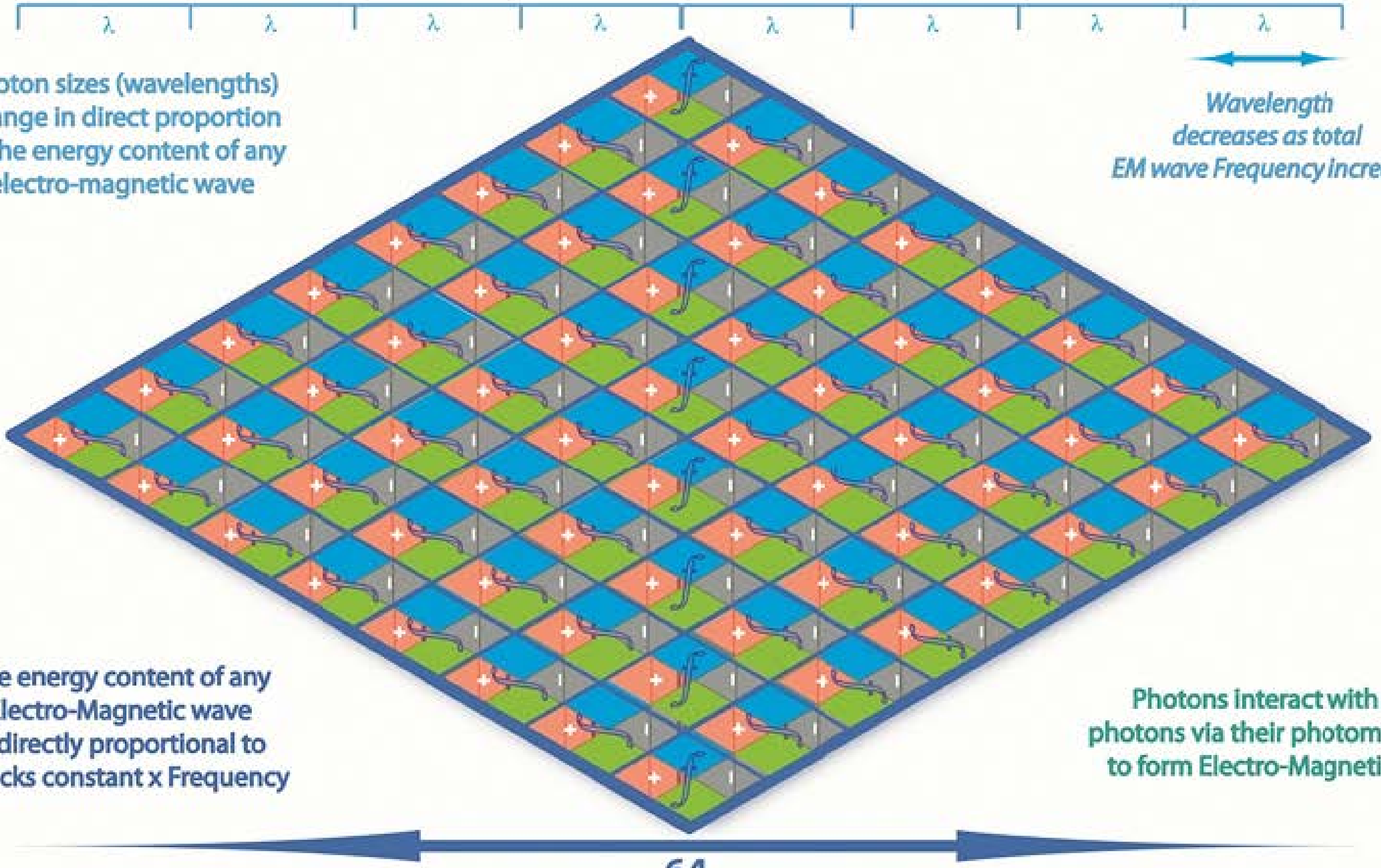

Photon sizes (wavelengths) change in direct proportion to the energy content of any electro-magnetic wave

Wavelength decreases as total EM wave Frequency increases

$$E = hf$$

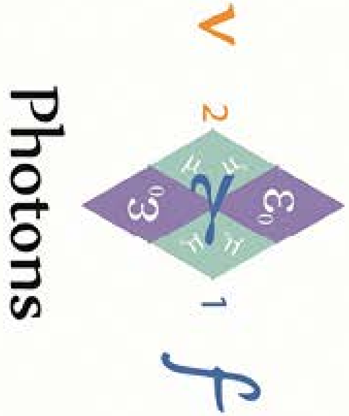
Photons
v
2
γ
ε_0
ε_0
μ_0
1
f

The energy content of any Electro-Magnetic wave is directly proportional to Plancks constant x Frequency

Photons interact with other photons via their photomagnetons to form Electro-Magnetic waves

64
Frequency
Number of Cycles per time unit

Tetryonics 26.06 - Photons in EM waves

Bosons in a EM wave

EM waves are comprised of transverse W Bosons

Odd numbers totaling to a Square number

$E = hv$

$\boxed{2v = f}$

$E = hf$

Even numbers totaling to a Square number

EM waves are also comprised of longitudinal Photons

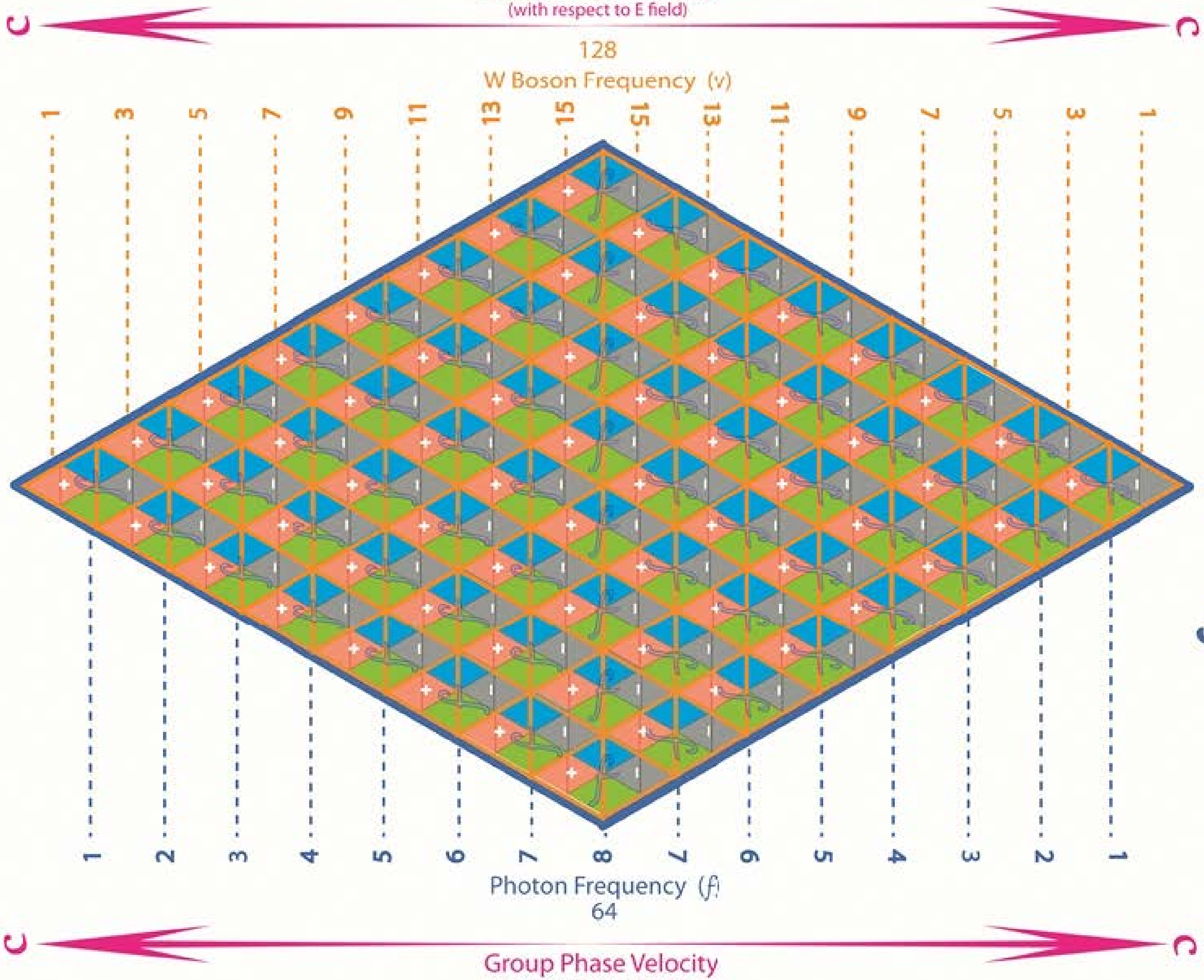
c
Group Phase Velocity
(with respect to E field)
c
128
W Boson Frequency (v)
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Photon Frequency (f)
64
c
Group Phase Velocity
(with respect to E field)
c

$2hv = E = hf$

[Bosons]

ODD π

(1,3,5,7,9, ...)

Z BOSONS are
1/2 wavelength
quantum levels

$E = h\nu$

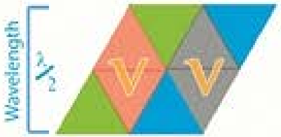

vs.

1:1

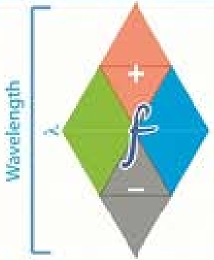

PHOTONS are
integer wavelength
frequencies

$E = hf$

[Photons]

EVEN π

(2,4,6,8,10, ...)

EM Field · Planck · quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons · ElectroMagnetic · mass · velocity

ZPFs are Boson quantums

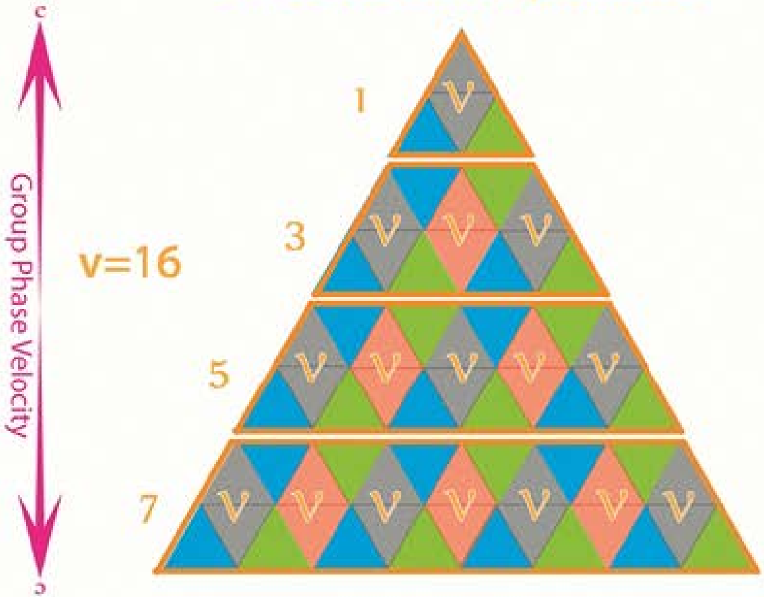

EM Field · Planck · quanta

$$\text{ODD}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons · ElectroMagnetic · mass · velocity

Charges are comprised of transverse Bosons

Bosons and Photons have differing EM geometries

EM Field · Planck · quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons · ElectroMagnetic · mass · velocity

Photons are comprised of Bosons

**Photons are EVEN number
Planck geometries**

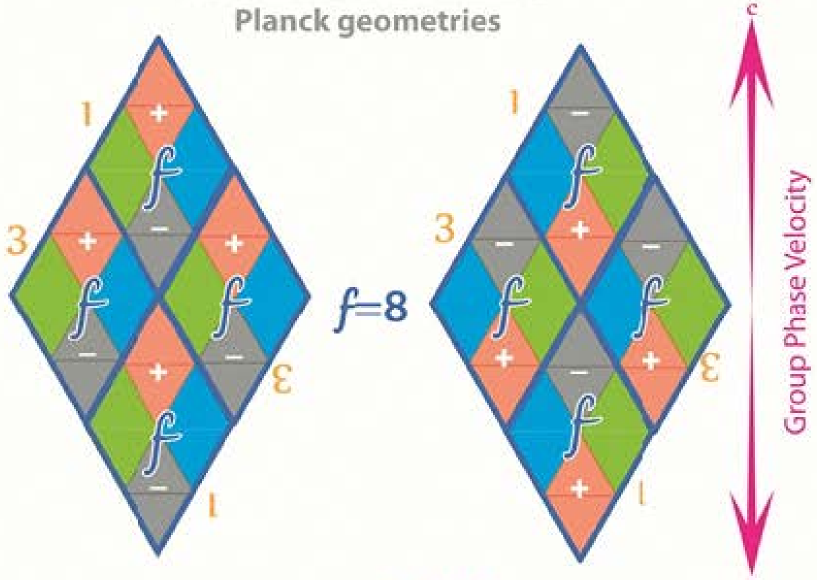

EM Field · Planck · quanta

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves · ElectroMagnetic · mass · velocity

EM waves are comprised of longitudinal Photons

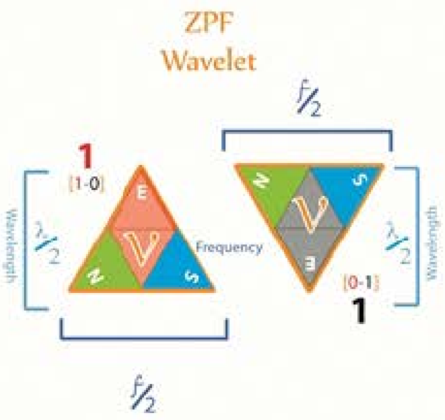
ZPF
Wavelet
1
[1-0]
Wavelength
λ/2
E
N
S
ν
Frequency
f/2
f/2
ν
N
S
E
[0-1]
1
λ/2
Wavelength

Bosons

Charge Carriers
$E = h\nu$

W+ boson

Bosons are half wavelength - non-Integer frequency
waveforms due to their odd ZPF numbers

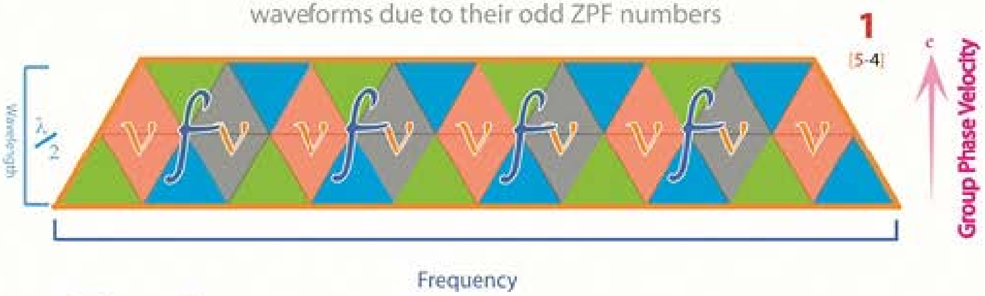
1
[5-4]
Wavelength
λ/2
ν f ν ν f ν ν f ν ν f ν ν
c
Group Phase Velocity
Frequency

$$mv^2 = E = h\nu^2$$

All Bosons are transverse ODD quanta EM masses

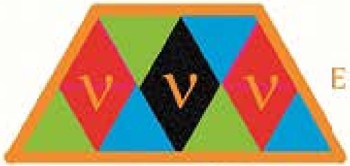

$E = h\nu$

$\text{ODD}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

The interchanging of Quantum number [ν]

$$2\nu = f$$

with Photon Frequency [f] is a cause of confusion

$E = hf$

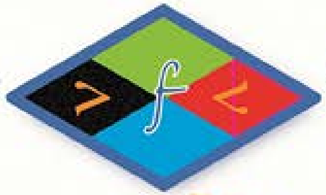

$\text{EVEN}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

All Photons are longitudinal EVEN quanta EM masses

Transverse Bosons

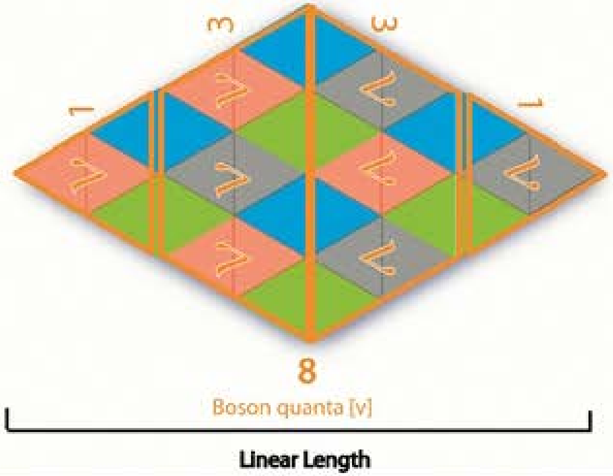
3
3
1
1
ν ν ν ν ν ν ν ν
8
Boson quanta [ν]
Linear Length

$$2h\nu = E = hf$$

EM Force Carriers
$E = hf$

Photons

Longitudinal Photons

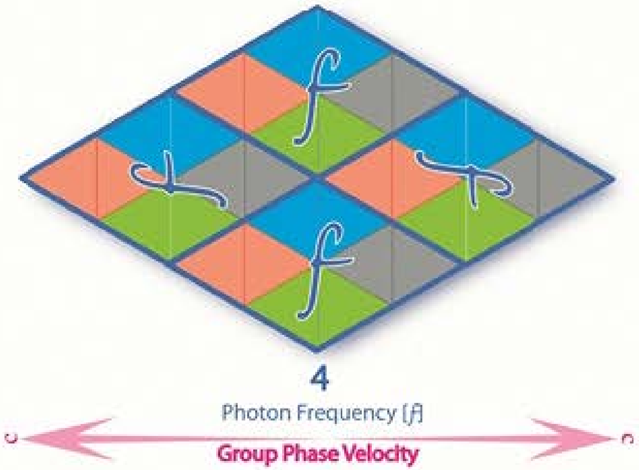
f
f
f
f
4
Photon Frequency [f]
c
c
Group Phase Velocity

Photonic Energies

$$2p^2 = KE = 2mv^2$$

The equilateral geometries of Quantised Angular Momentum of Planck energy quanta is the source of linear Momentum

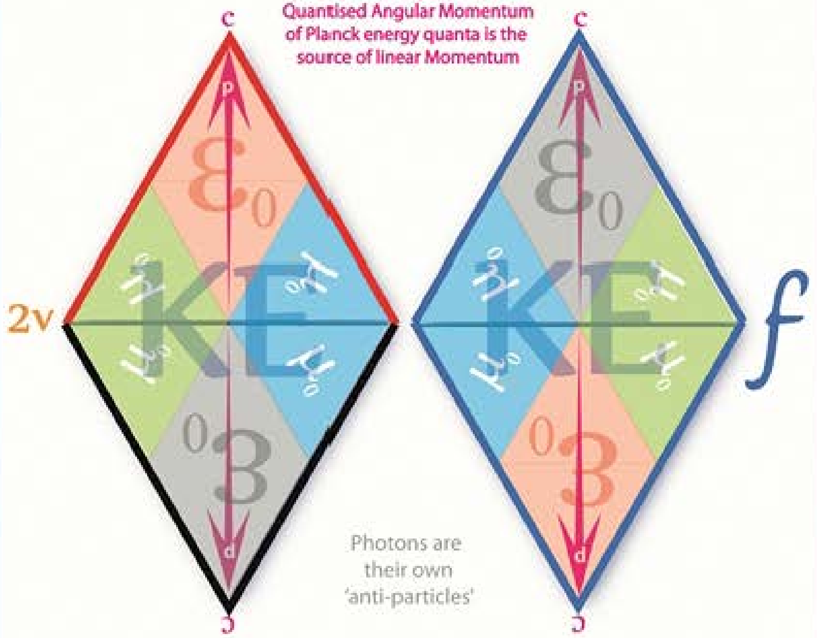

Photons are their own 'anti-particles'

$$2hv = E = hf$$

$$E_\gamma = 2mc^2$$

As Photons are 2D EM energy waveforms they should always be referred to as having EM mass-energy equivalence

γ

Planck quanta

$$2\pi\left[m\Omega v^2\right]$$

mass velocity

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

Photons are kinetic energy [KE] wave packets:

$KE = (1/2) * m_\gamma * v^2$

$= 1/2 * (E/c^2) * c^2$

$= E/2$

KE

Frequency
Number of repeating cycles of energy momenta per time unit

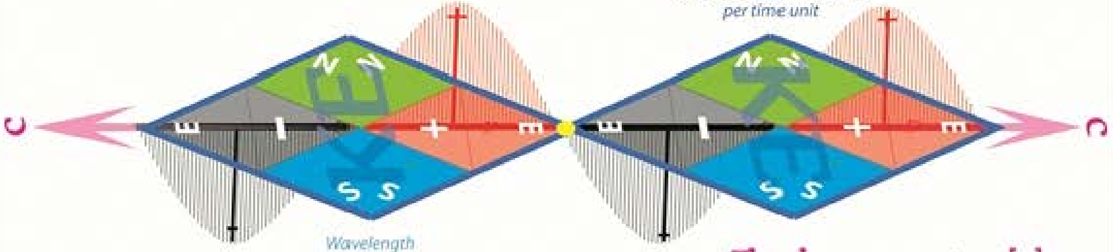

Wavelength decreases as Photon internal Frequency increases

They have momentum [p]:

$p = m_\gamma * v$

$= (E/c^2) * c$

$= E * c$

0
[v-v]

Photons are 2n neutral charge waveforms

Photons are divergent kinetic waveforms

$$KE = \frac{E_{KEM}}{2}$$

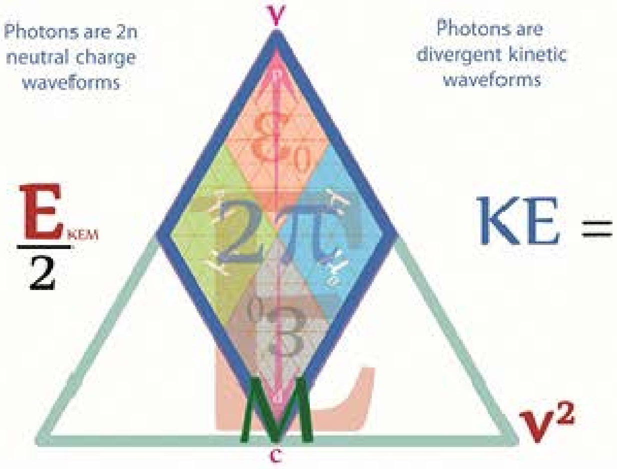

$$KE = \frac{1}{2}Mv^2$$

Copyright ABRAHAM [2008] - all rights reserved

Quantum Harmonic Motion

$$\underset{\text{Photons}}{2\pi}\left[\underset{\text{ElectroMagnetic}}{\overset{\text{EM Field}}{[\varepsilon_0\mu_0]}}.\left[\overset{\text{Planck}}{\underset{\text{mass}}{m}}\overset{\text{quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

The quantum harmonic oscillator is the quantum mechanical analog of the classical harmonic oscillator.

It is one of the few quantum mechanical systems for which a simple exact solution is known.

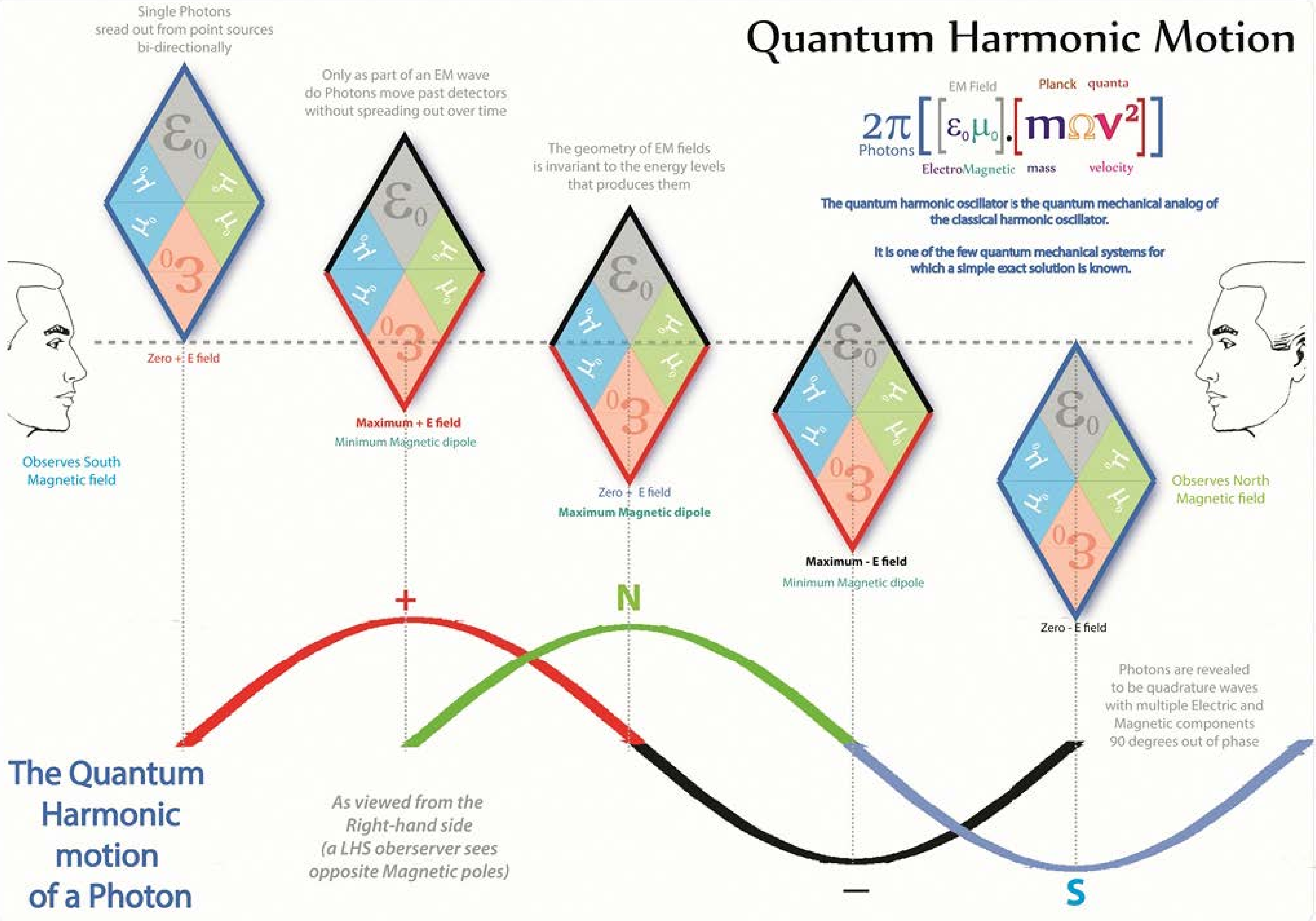

Single Photons
sread out from point sources
bi-directionally

Only as part of an EM wave
do Photons move past detectors
without spreading out over time

The geometry of EM fields
is invariant to the energy levels
that produces them

Zero + E field

Maximum + E field
Minimum Magnetic dipole

Zero + E field
Maximum Magnetic dipole

Maximum - E field
Minimum Magnetic dipole

Zero - E field

Observes South
Magnetic field

Observes North
Magnetic field

Photons are revealed
to be quadrature waves
with multiple Electric and
Magnetic components
90 degrees out of phase

*As viewed from the
Right-hand side
(a LHS oberserver sees
opposite Magnetic poles)*

The Quantum Harmonic motion of a Photon

Tetryonics 26.11 - Quantum Harmonic Motion

Photonic EM fields

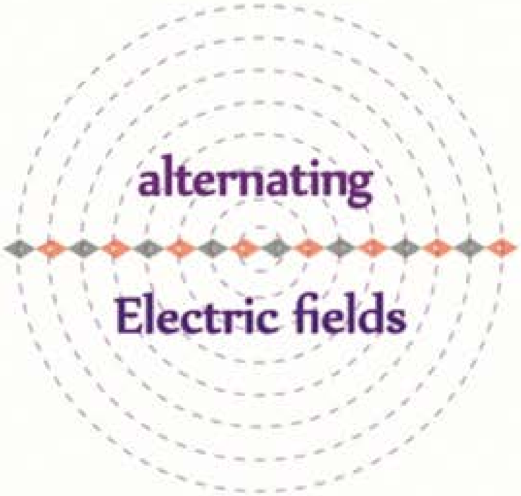

alternating

Electric fields

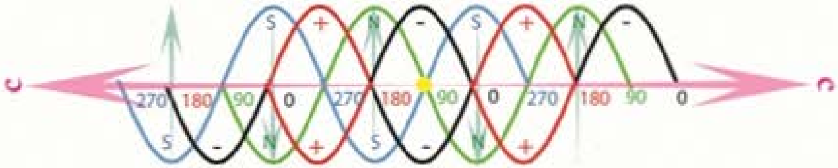
S
+
N
-
S
+
N
-
270
180
90
0
270
180
90
0
270
180
90
0
S
-
N
+
S
-
N
+

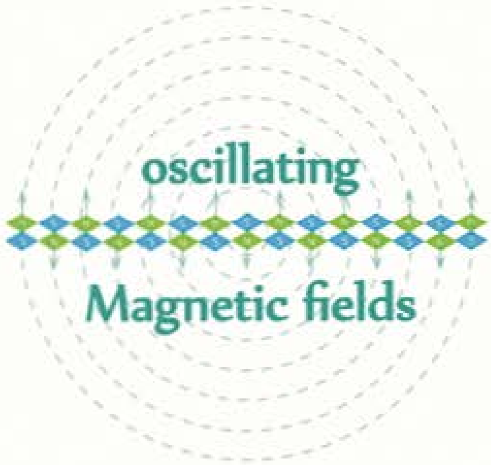

oscillating

Magnetic fields

Euler's formula is a natural geometric expression of electromagnetic waves

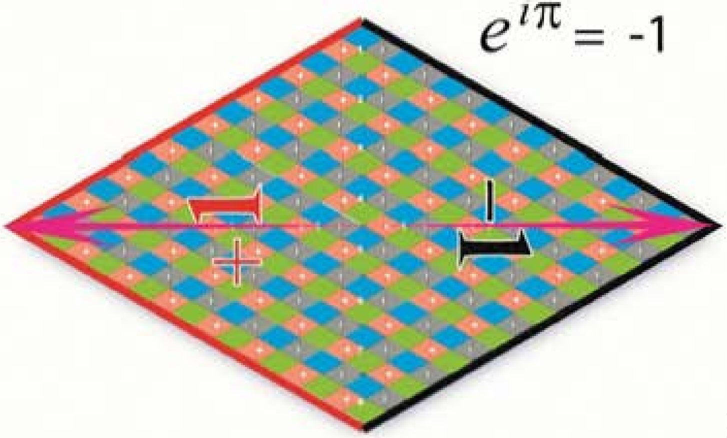

$e^{i\pi} = -1$

$e^{i\pi} + 1 = 0$

$e^{ix} = \cos x + i \sin x$

magnetic fields

cos

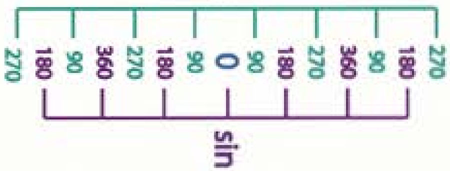
270
180
90
360
270
180
90
0
90
180
270
360
90
180
270

sin

electric fields

Every field geometry in a EM wave is 90 degrees out of phase with the E or M field adjacent to it

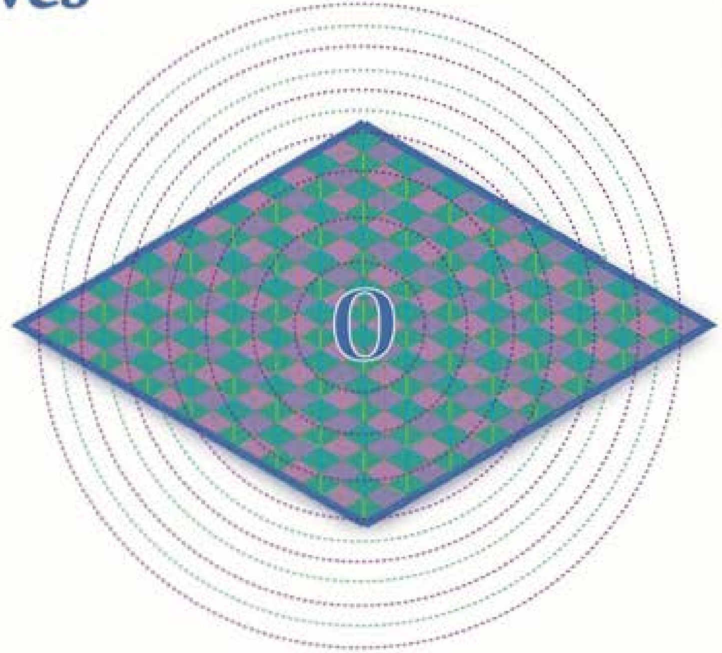

Tetryonics 26.12 - Photonic EM field phases

$$e^{i\pi} = -1$$

$$e^{i\pi} + 1 = 0$$

Euler's Formula

Euler's formula is often considered to be the basis of the complex number system.
In deriving this formula, Euler established a relationship between the
trigonometric functions, sine and cosine, and e raised to a power

$$e^{ix} = \cos(x) + i\sin(x)$$

a mathematical description of EM-Energy waveforms

adjacent EM fields are 90-180° out of phase

Leonhard Euler

(15 April 1707 – 18 September 1783)

Positive E-fields are out of phase with Negative fields

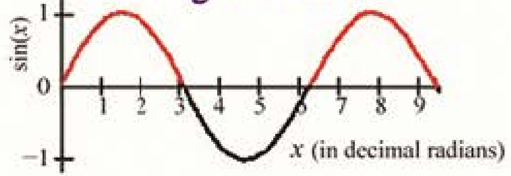

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$$

Magnetic fields are out of phase with Electric fields

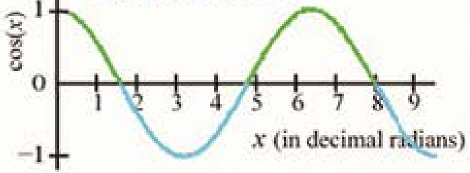

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots$$

$$\sin x + \cos x = 1 + x - \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$$

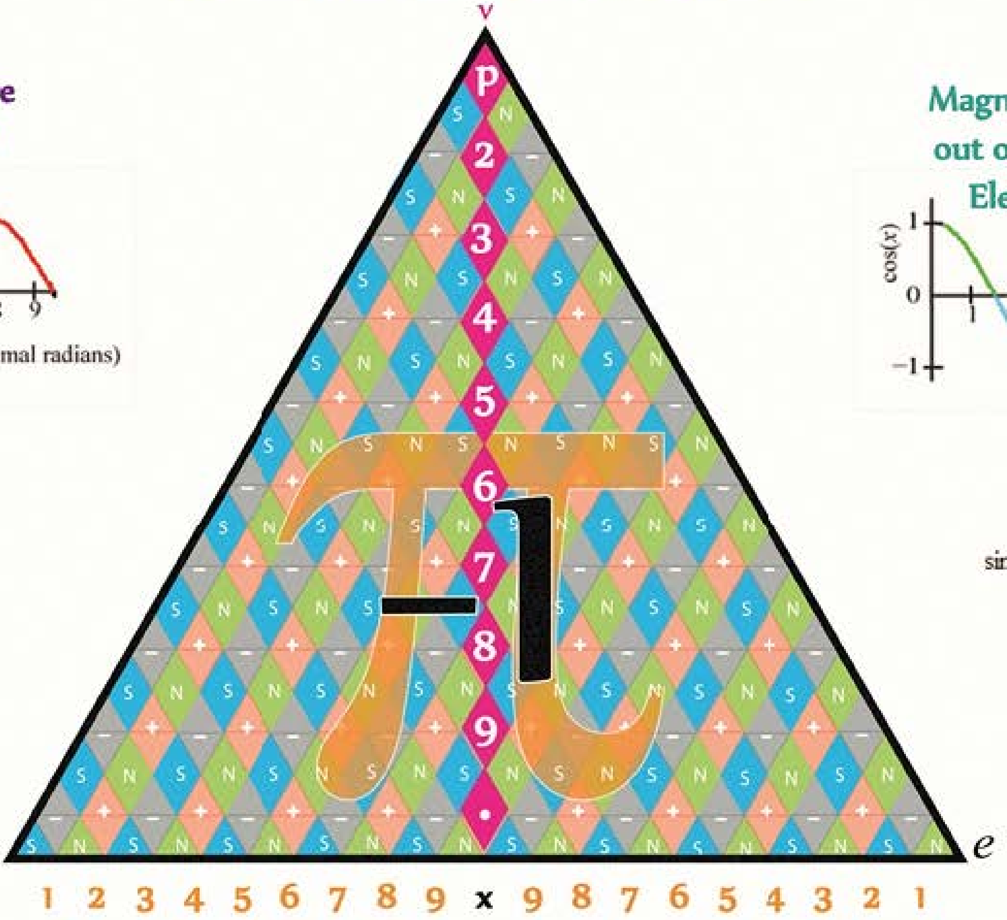

The above geometry is reflective of negative charge energy momenta Planck geometries

the momentum of the nett charged geometry is the Square root of Negative 1

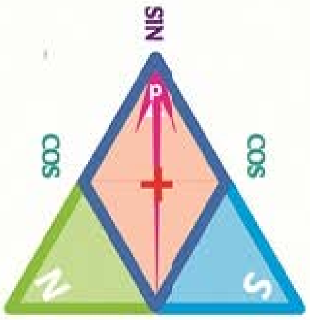

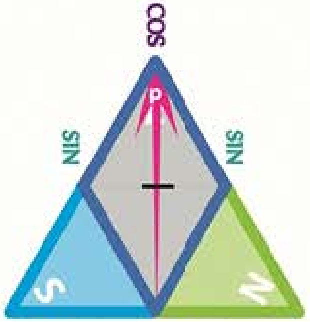

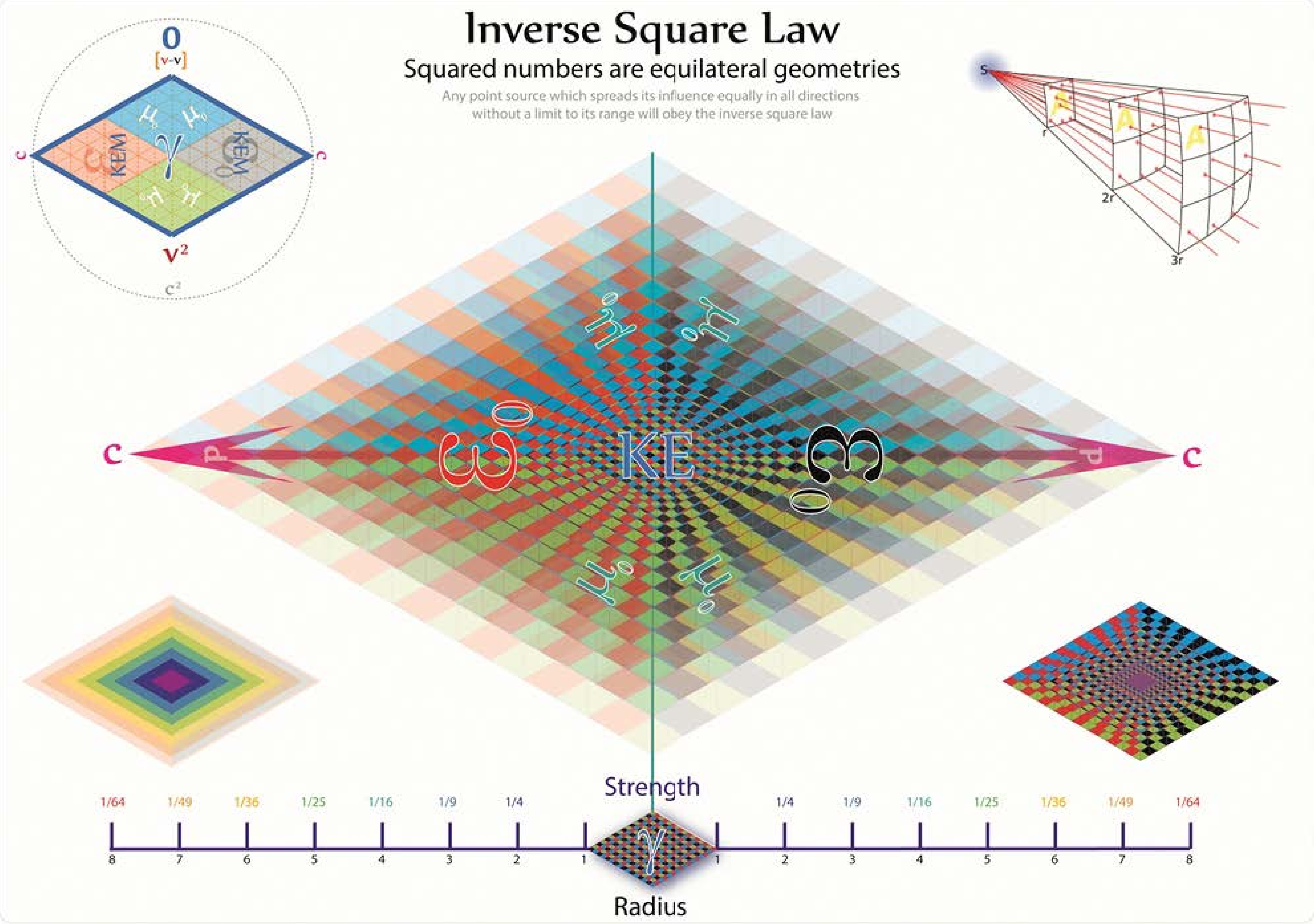

Inverse Square Law

Squared numbers are equilateral geometries

Any point source which spreads its influence equally in all directions without a limit to its range will obey the inverse square law

0
[v-v]
μ0
μ0
KEM
KEM
γ
c
c
v²
c²
S
A
A
A
r
2r
3r
c
c
p
p
ε0
ε0
KE
μ0
μ0
μ0
μ0
Strength
1/64
1/49
1/36
1/25
1/16
1/9
1/4
1/4
1/9
1/16
1/25
1/36
1/49
1/64
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
γ
Radius

mass-Energy momenta

EM Field Planck quanta

$$n\pi\left[\left[\varepsilon_0\mu_0\right]\cdot\left[m\Omega v^2\right]\right]$$

mass ElectroMagnetic mass velocity

Throughout history Physicists have sought relationships between EM mass-Energy and momenta in an attempt to discern the true Nature of these physical properties

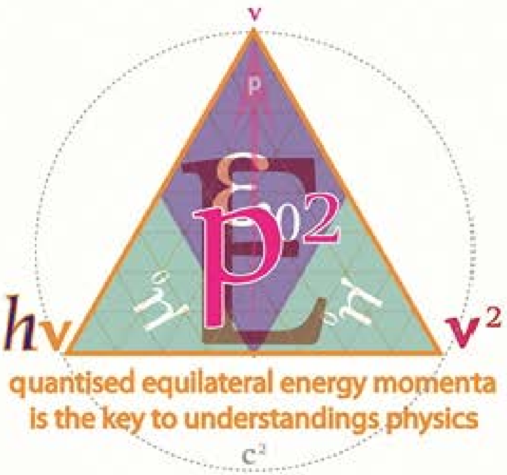

quantised equilateral energy momenta is the key to understandings physics

E

Energy

$$kg\,\frac{m^2}{s^2}$$

Compton Frequency

$$\left[\frac{E}{h}\right]$$

$E = mv$

Newton

$E = mv^2$

Leibnitz

$v = \frac{1}{\sqrt{\varepsilon_0\mu_0}}$

Maxwell

Velocity of Light

$$v = \left[\underbrace{\left[\frac{c^2}{\Omega}\right]}_{\text{Scalar Frequency}}^{f}\cdot\underbrace{\left[\frac{\Omega}{c}\right]}_{\text{Linear Wavelength}}^{\lambda}\right]$$

Planck

$E = hv$

Einstein

$E = hf$

de Broglie

h/p

Planck's Constant

$$kg\,\frac{m^2}{s}$$

de Broglie wavelength

$$\left[\frac{E}{p}\right]$$

Wavelength

p

Linear momentum

$$kg\,\frac{m}{s}$$

Ryberg wavenumber

$$\left[\frac{p}{E}\right]$$

Copyright ABRAHAM [2008] - all rights reserved

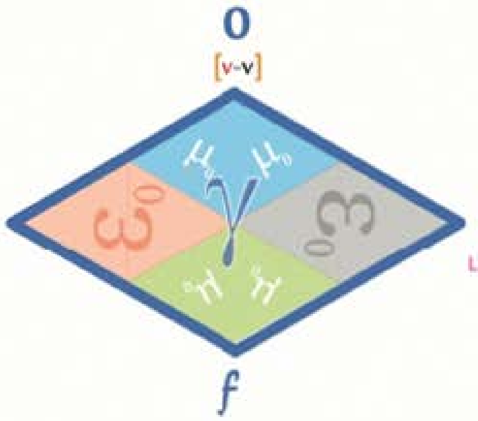

$$mv^2 = KE = hv^2$$

All Photons contain Energy momenta
and therefore have electromagnetic mass
but they are 'Matterless' geometries

Wavenumber

$$\frac{m}{s}\cdot\frac{s}{m^2}\left[\frac{c}{\Omega}\right]\frac{1}{m}$$

Light cannot have zero velocity
[stand still] but it is possible
to catch up to a photon
and observe its waveform

The Speed of Light is
the maximum speed of Electrical energy
and remains a Universal constant
(but it is not a Limit)

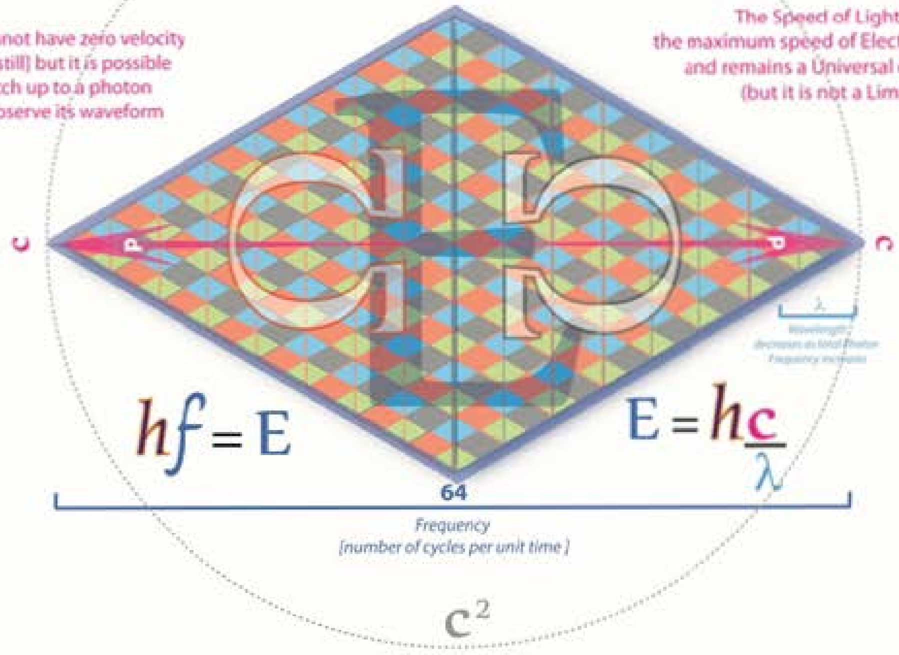

$$hf = E \qquad E = h\frac{c}{\lambda}$$

Frequency
[number of cycles per unit time]

Celeritas

[Frequency x Wavelength]

EM waves are typically described
through any of their following
three physical properties:

frequency *f*,
wavelength λ, or
photon energy E

ElectroMagnetic Fields

$$\varepsilon_0\mu_0 = \frac{1}{c^2}$$

1.112650056 e-17 $\frac{s^2}{m^2}$

Frequency, momentum, and Wavelength
are all related through the geometries of c & c²

$$\lambda$$

$$m\left[\frac{\Omega}{c}\right]\frac{m^2}{s}\cdot\frac{s}{m}$$

Wavelength

Velocity of Light

$$c = \left[\overset{f}{\left[\frac{c^2}{\Omega}\right]}\cdot\overset{\lambda}{\left[\frac{\Omega}{c}\right]}\right]$$

Frequency Wavelength

$$v = \frac{1}{\sqrt{\varepsilon_0\mu_0}}$$

$$f$$

$$\frac{m^2}{s^2}\cdot\frac{s}{m^2}\left[\frac{c^2}{\Omega}\right]\frac{1}{s}$$

Frequency

Frequency is the number of neutral quanta sets repeating per unit time in a EM wave

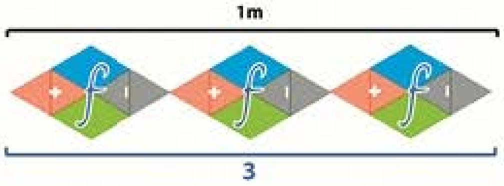

EM waves frequencies [photon number] change in direct proportion to the energy content of any electromagnetic wave

The specific wavelengths and frequencies of energy momenta in KEM fields form the basis for Spectral emission/absorption lines [care must be always taken to specify if quanta or photons are being measured]

$$2v = f$$

Bosons are charged transverse energy momenta quanta

$E = h\nu$

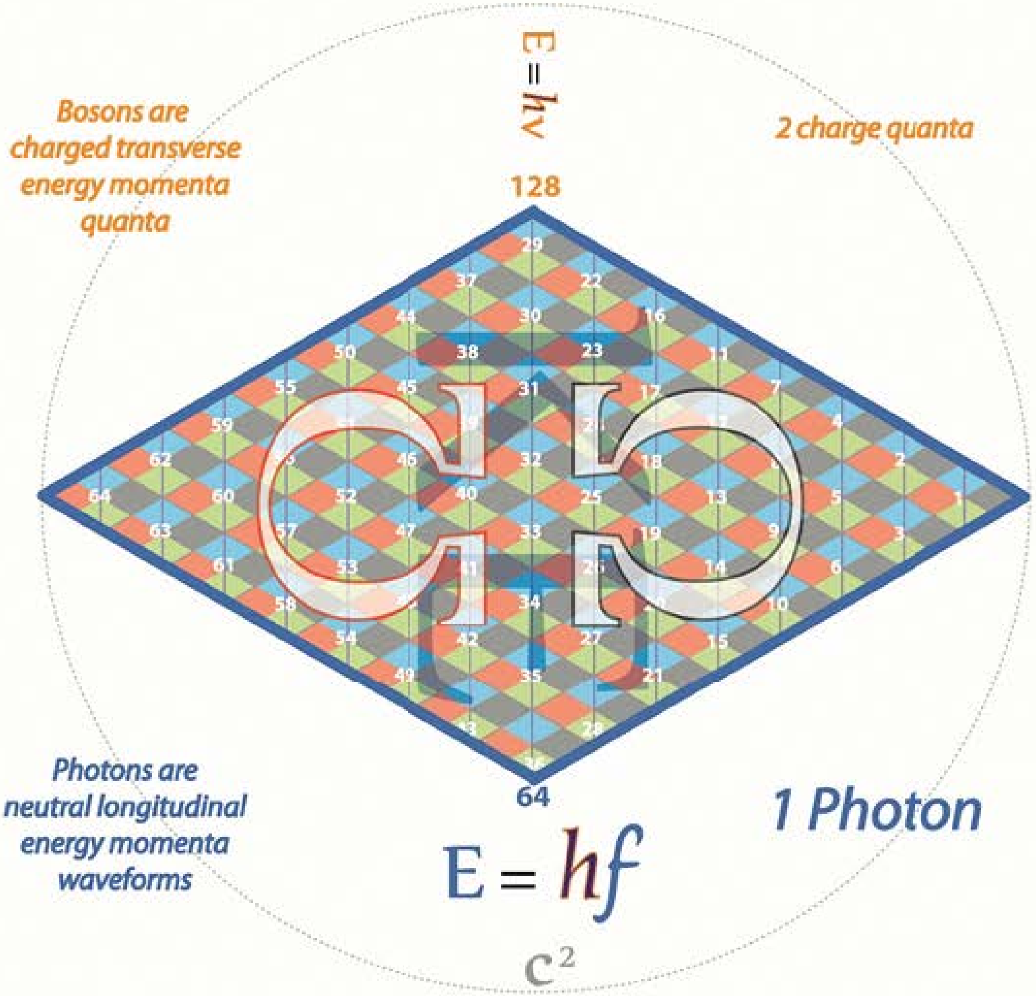

2 charge quanta

Irrespective of whether its measured as a transverse or longitudinal electromagnetic wave

Frequency is the inverse of Charge

$$f$$

$$\left[\frac{c^2}{\Omega}\right]$$

Frequency

Photons are neutral longitudinal energy momenta waveforms

1 Photon

$$E = hf$$

$$c^2$$

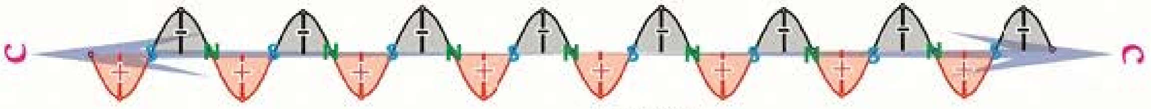

Frequency

[Number of Neutral ZPF sets per unit of time]

is directly related to the nett quantised angular momentum of any space time geometry

Copyright ABRAHAM [2008] - all rights reserved

λ

Wavelength

Wavelength is a measure of the distance between repetitions of geometric features such as maximums, minimums, or zero-points of Electric and Magnetic fields

$$\lambda$$

$$\left[\frac{\Omega}{c}\right]$$

Wavelength

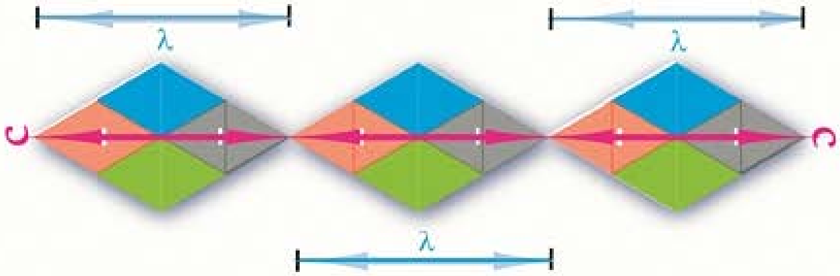

$\tilde{v}$

Wavelengths are the inverse of Wavenumbers

λ

For additional quantised energy momenta to be added to any spatial geometry their physical wavelengths must decrease

Lorentz velocity dependent length contractions must be correctly retermed electromagnetic WAVE-length contractions

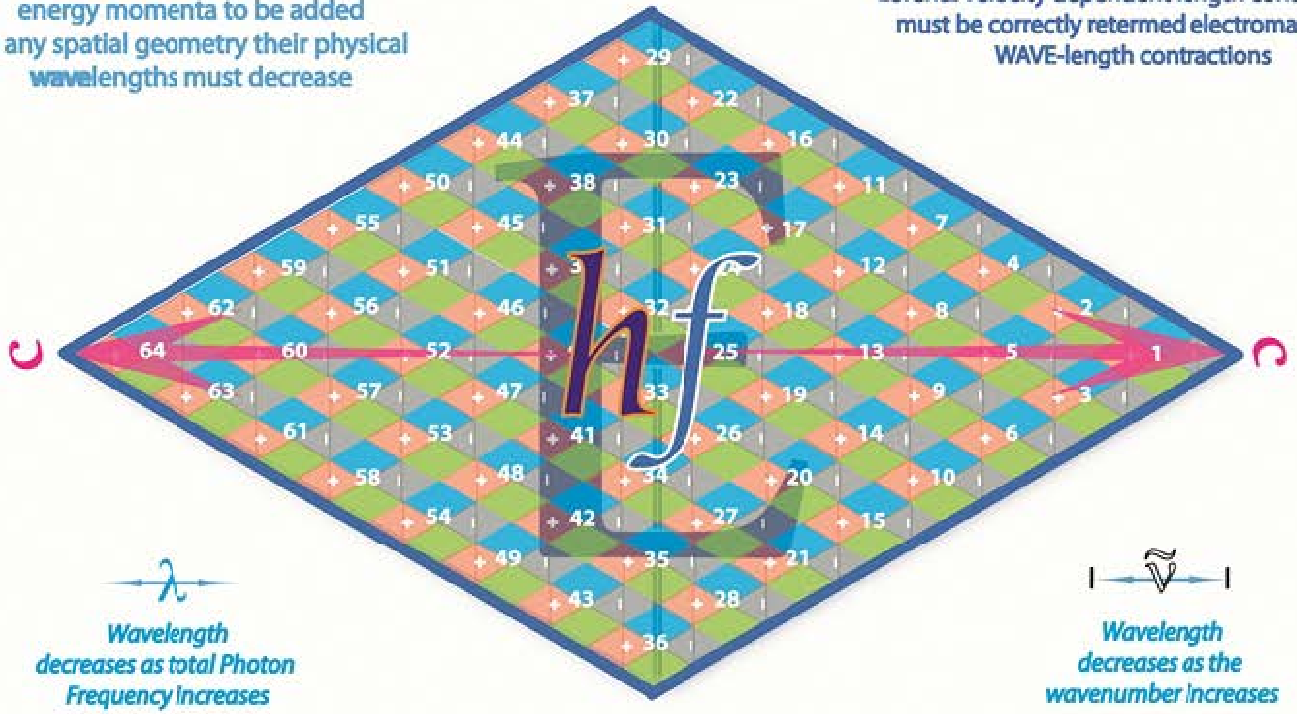
hf
c
c
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
25
26
27
28
29
30
31
32
33
34
35
36
37
38
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64

λ

Wavelength decreases as total Photon Frequency Increases

$\tilde{v}$

Wavelength decreases as the wavenumber Increases

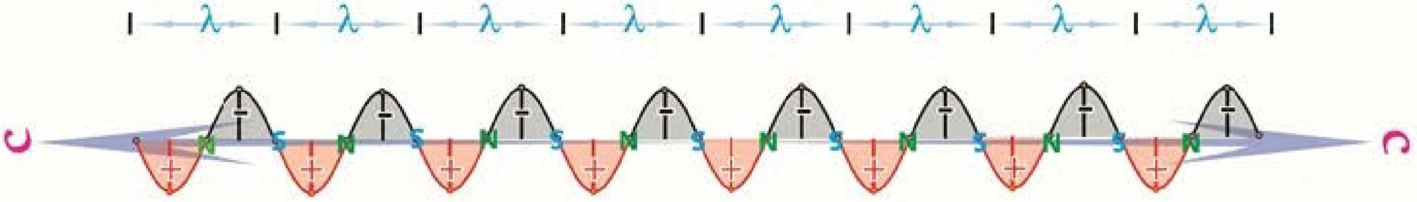
λ
λ
λ
λ
λ
λ
λ
λ
N
S
+
c
c

299,792,458

linear metres per second

In physics, the wavelength of a sinusoidal wave is the spatial period of the wave – the distance over which the wave's shape repeats. It is usually determined by considering the distance between consecutive corresponding points of the same phase, such as crests, troughs, or zero crossings, and is a characteristic of both traveling waves and standing waves, as well as other spatial wave patterns.

The concept can also be applied to periodic complex or non-sinusoidal waves

Copyright ABRAHAM [2008] - all rights reserved

Wavenumbers

$\tilde{\nu}$

Wavenumbers are the inverse of Wavelengths

λ

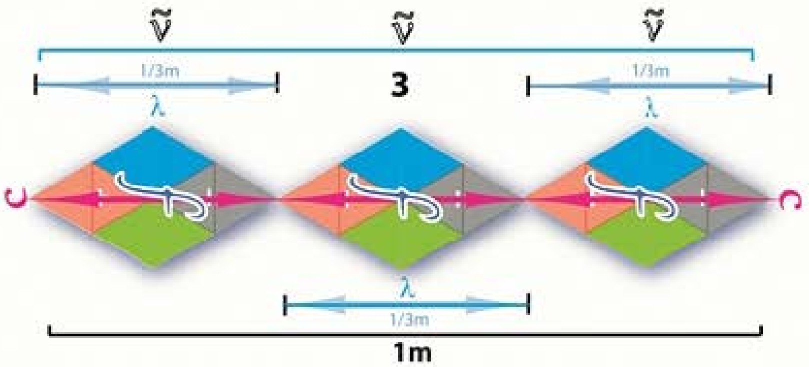

$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
1/3m
λ
3
1/3m
λ
c
c
λ
1/3m
1m

Wavenumbers are the spatial equivalent of Frequency

$$\left[\frac{c}{\Omega}\right]$$

λ

$$\left[\frac{\Omega}{c}\right]$$

Wavelength and Frequency are related to the group velocity of quanta in an EM wave

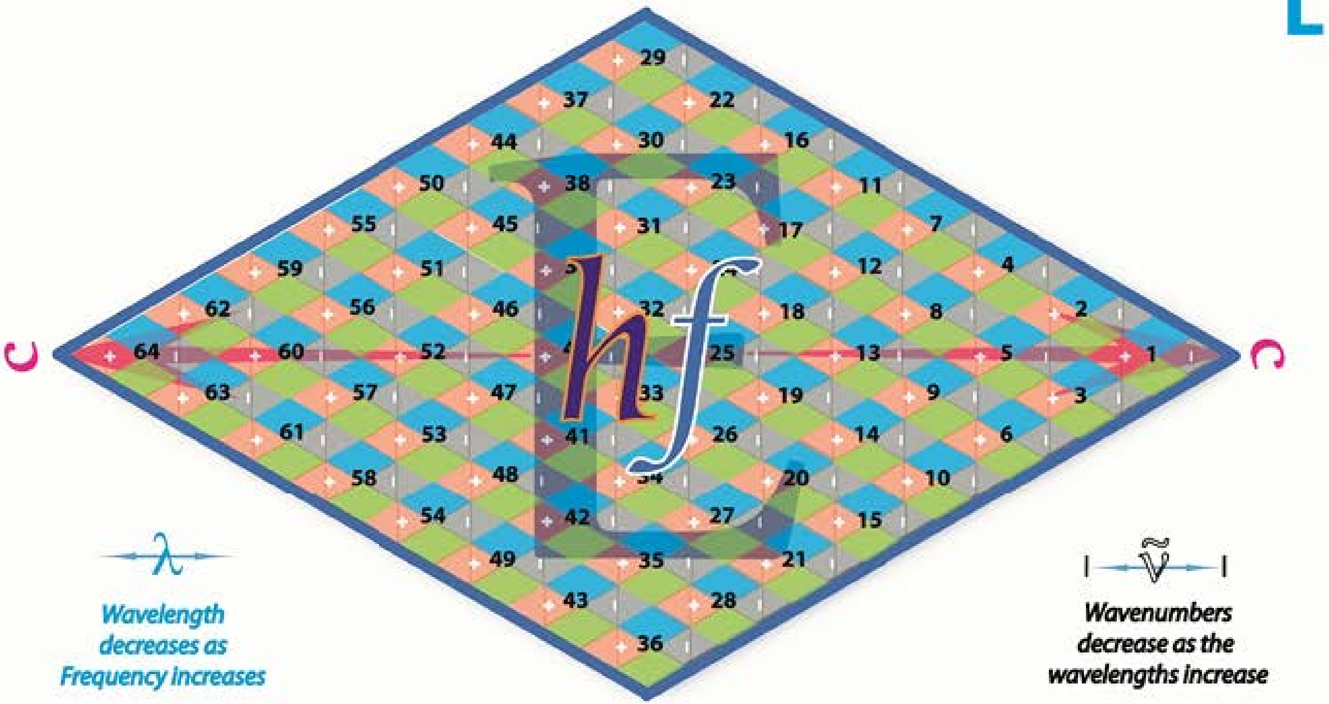

29
37
22
44
30
16
50
38
23
11
55
45
31
17
7
59
51
12
4
62
56
46
32
18
8
2
64
60
52
25
13
5
1
63
57
47
33
19
9
3
61
53
41
26
14
6
58
48
20
10
54
42
27
15
49
35
21
43
28
36
hf
c
c

$$\tilde{\nu} = 1/\lambda$$

Wavenumbers are inversely proportional to Wavelengths

λ

Wavelength decreases as Frequency increases

$\tilde{\nu}$

Wavenumbers decrease as the wavelengths increase

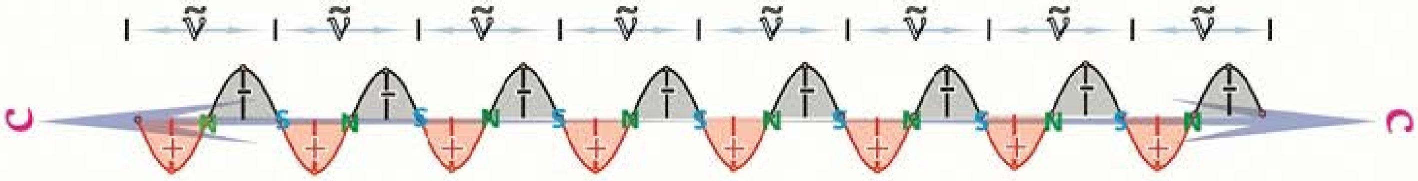

$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
$\tilde{\nu}$
c
N
S
c

299,792,458

linear metres per second

Tetryonics 26.19 - Wavenumbers

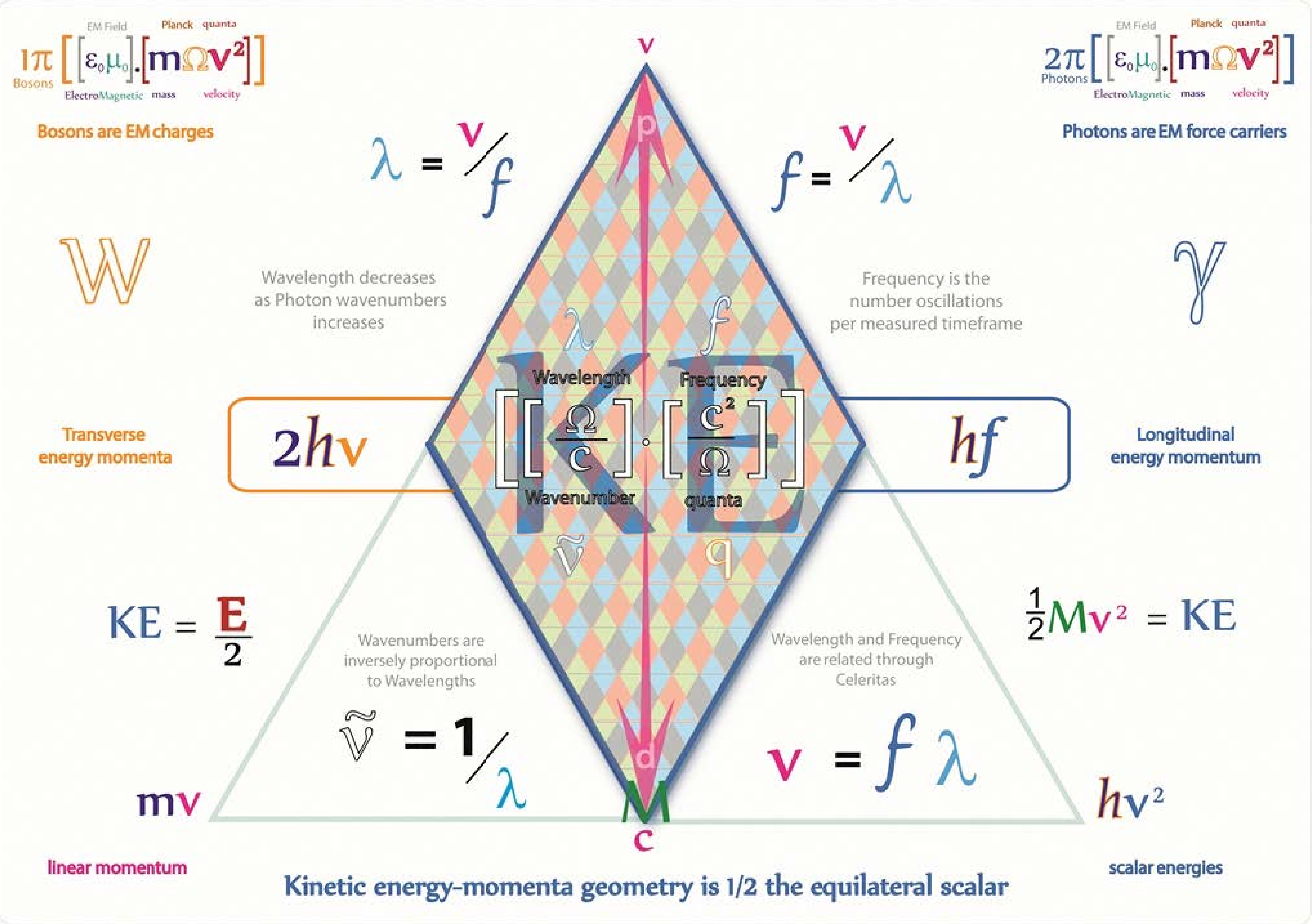
EM Field
Planck
quanta
Bosons
ElectroMagnetic
mass
velocity
Photons
W
γ
p
d
v
M
c
λ
f
$\tilde{\nu}$
q
KE
Wavelength
Frequency
Wavenumber
quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons are EM charges

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons are EM force carriers

$$\lambda = {}^{v}/_{f}$$

$$f = {}^{v}/_{\lambda}$$

Wavelength decreases as Photon wavenumbers increases

Frequency is the number oscillations per measured timeframe

Transverse energy momenta

$$2hv$$

$$\left[\left[\frac{\Omega}{c}\right].\left[\frac{c^2}{\Omega}\right]\right]$$

$$hf$$

Longitudinal energy momentum

$$KE = \frac{E}{2}$$

$$\frac{1}{2}Mv^2 = KE$$

Wavenumbers are inversely proportional to Wavelengths

$$\tilde{\nu} = {}^{1}/_{\lambda}$$

Wavelength and Frequency are related through Celeritas

$$v = f\lambda$$

$$mv$$

$$hv^2$$

linear momentum

scalar energies

Kinetic energy-momenta geometry is 1/2 the equilateral scalar

Tetryonics 26.20 - Kinetic energy-momenta

Faraday Rotation

The Faraday effect or Faraday rotation is a Magneto-optical phenomenon, that is, an interaction between light and a magnetic field in a medium. The Faraday effect causes a rotation of the plane of polarization which is linearly proportional to the component of the magnetic field in the direction of propagation.

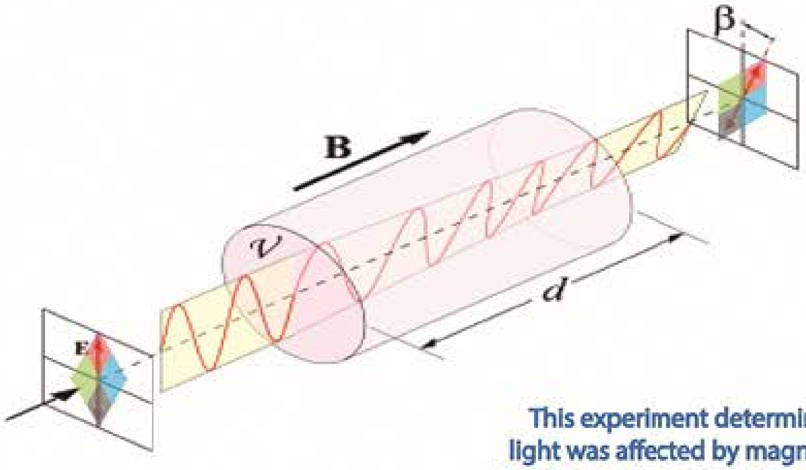

Discovered by Michael Faraday in 1845,
the Faraday effect was the first experimental
evidence that light and electromagnetism are related

Faraday summarized the entire effect as follows:

"Magnetic lines, then, in passing through silicated borate of lead, and a great number of other substances, cause these bodies to act upon a polarized ray of light when the lines are parallel to the ray, or in proportion as they are parallel to it: if they are perpendicular to the ray, they have no action upon it.

They give the diamagnetic the power of rotating the ray; and the law of this action on light is, that if a magnetic line of force be going from a north pole, or coming from a south pole, along the path of a polarized ray coming to the observer, it will rotate that ray to the right-hand; or, that if such a line of force be coming from a north pole, or going from a south pole, it will rotate such a ray to the left hand."

Michael Faraday

(22 September 1791 – 25 August 1867)

Evan's Photomagneton

A photon has a magnetic dipole.
It is an elementary magnet.
Evans discovery of the photon's
longitudinal magnetic field in
1992 is as significant, as Einstein's
discovery of relativity.

It helps in giving a physical
interpretation of wave mechanics,
two-slit interference
and the Faraday effect,

all accounted for in Tetryonic Theory.

This experiment determined that light was affected by magnetic force. This "magneto-optical effect" was later termed the Faraday effect.

Faraday experimented with other substances that yielded similar results.

The resultant effect he termed "diamagnetism" concluding that magnetism was an inherent property of all EM mass-ENERGY-Matter

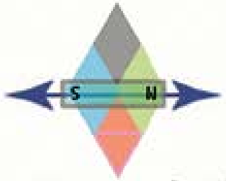

Before the
Magnetic field

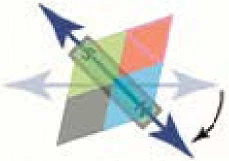

After the
Magnetic field

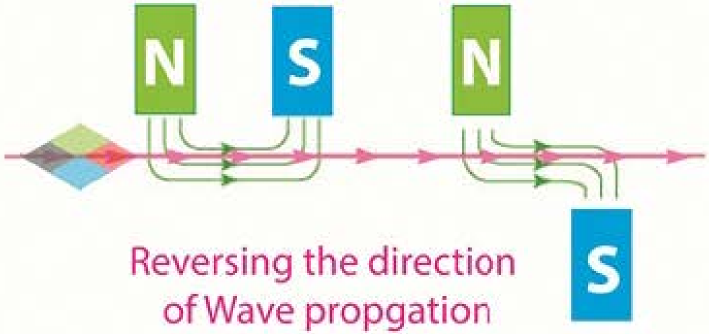

Reversing the direction
of Wave propgation
reverses the rotation
effected by the external
magnetic field

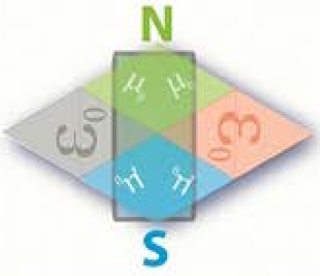

Every Photon and EM wave has polarised magnetic apexes as a result of the Planck quanta constituing them (creating Magnetic moments)

These Photo-Magnetic Moments allow for the interaction of Photons with external magnetic fields resulting in Faraday rotation of Photons and EM fields

Tetryonics 26.21 - Faraday rotation of Photons

Copyright ABRAHAM [2008] - all rights reserved

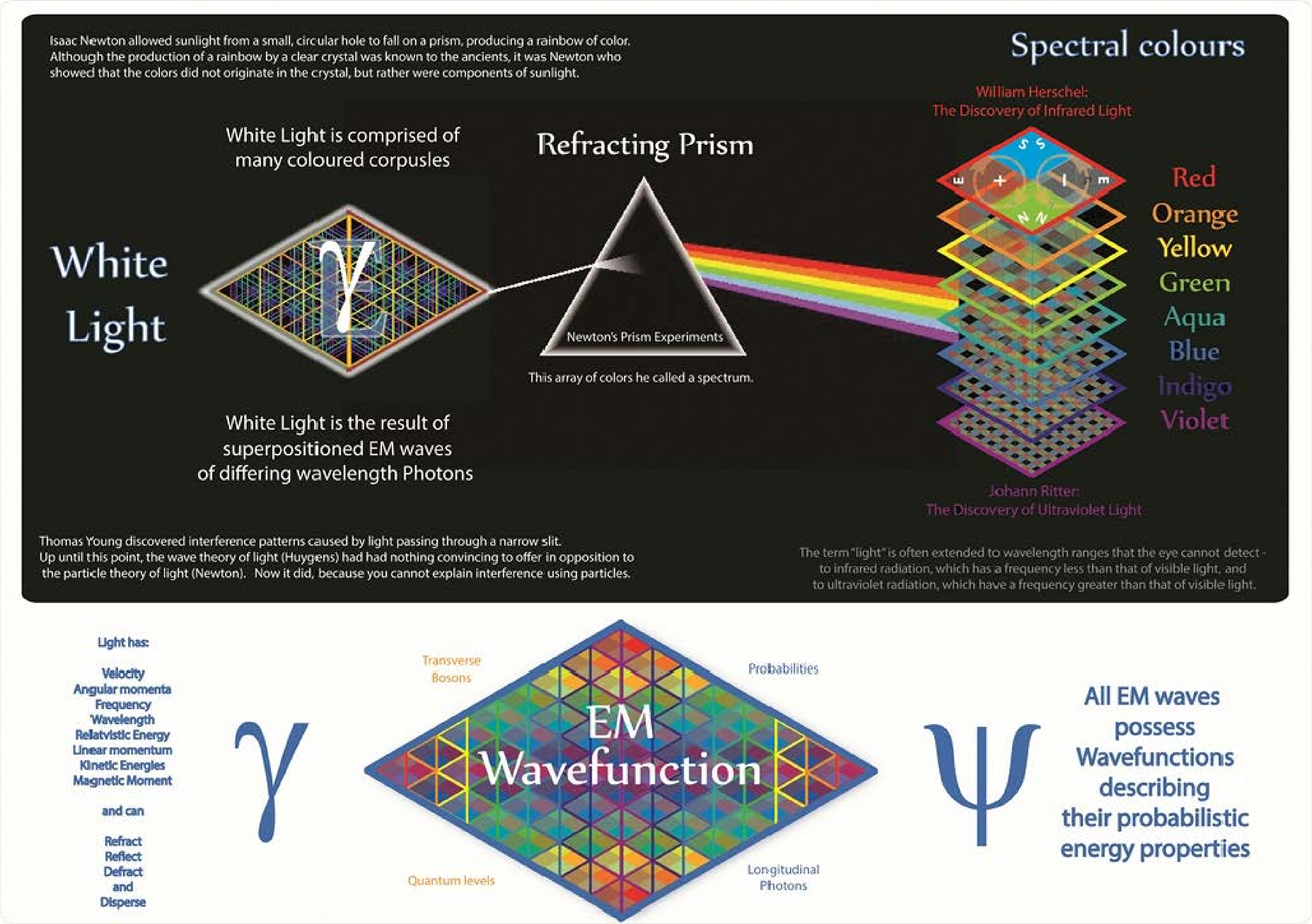

Isaac Newton allowed sunlight from a small, circular hole to fall on a prism, producing a rainbow of color. Although the production of a rainbow by a clear crystal was known to the ancients, it was Newton who showed that the colors did not originate in the crystal, but rather were components of sunlight.

Spectral colours

William Herschel:
The Discovery of Infrared Light

White Light is comprised of many coloured corpusles

Refracting Prism

White Light

Newton's Prism Experiments

This array of colors he called a spectrum.

White Light is the result of superpositioned EM waves of differing wavelength Photons

Johann Ritter:
The Discovery of Ultraviolet Light

Thomas Young discovered interference patterns caused by light passing through a narrow slit. Up until this point, the wave theory of light (Huygens) had had nothing convincing to offer in opposition to the particle theory of light (Newton). Now it did, because you cannot explain interference using particles.

The term "light" is often extended to wavelength ranges that the eye cannot detect - to infrared radiation, which has a frequency less than that of visible light, and to ultraviolet radiation, which have a frequency greater than that of visible light.

Light has:

Velocity
Angular momenta
Frequency
Wavelength
Relativistic Energy
Linear momentum
Kinetic Energies
Magnetic Moment

and can

Refract
Reflect
Defract
and
Disperse

All EM waves possess Wavefunctions describing their probabilistic energy properties

Copyright ABRAHAM [2008] - all rights reserved

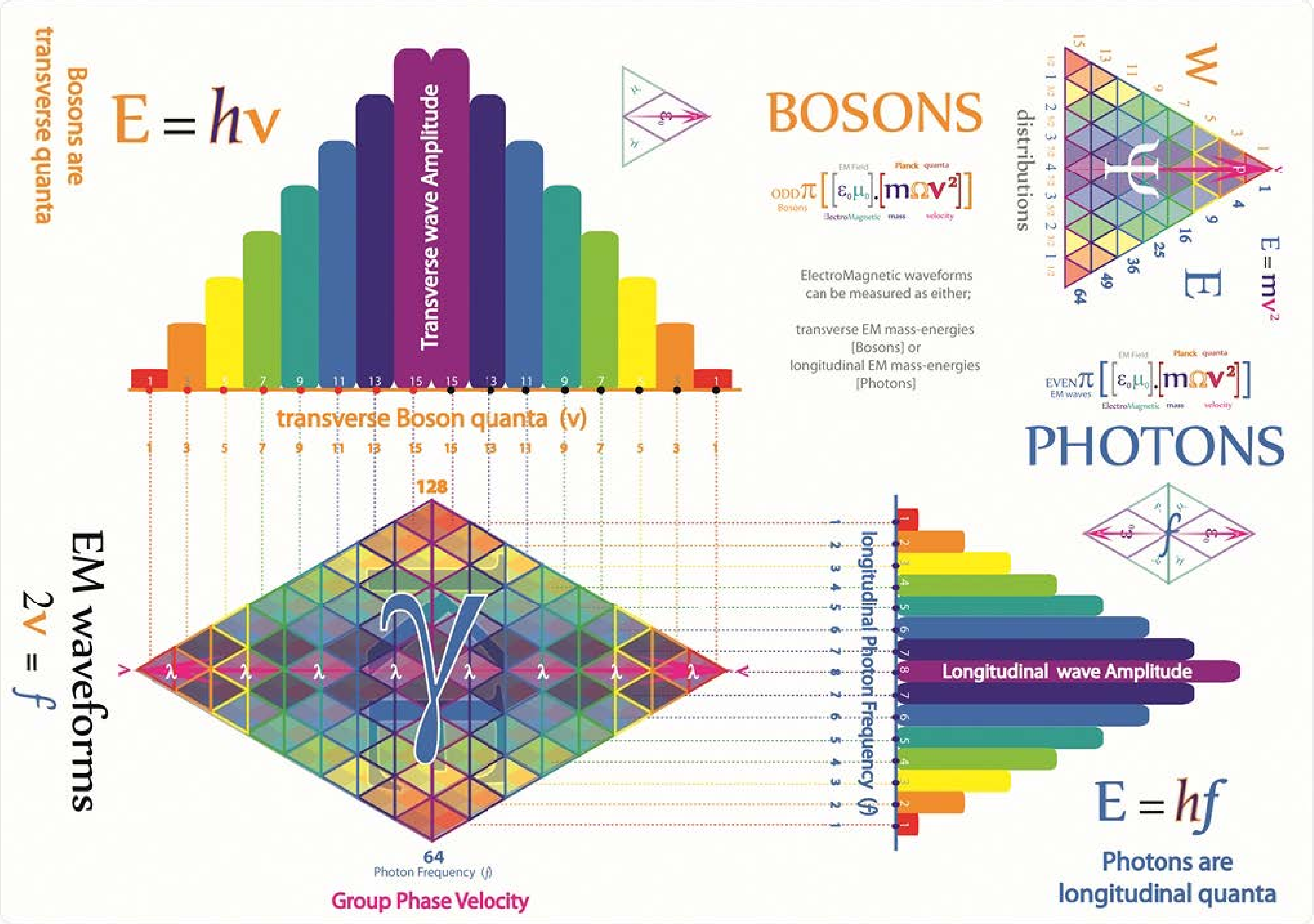

Bosons are transverse quanta

$E = h\nu$

BOSONS

ODD $\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

ElectroMagnetic waveforms can be measured as either;

transverse EM mass-energies [Bosons] or longitudinal EM mass-energies [Photons]

EVEN $\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

PHOTONS

EM waveforms

$2v = f$

$E = hf$

Photons are longitudinal quanta

Transverse wave Amplitude
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
transverse Boson quanta (v)
1 3 5 7 9 11 13 15 15 13 11 9 7 5 3 1
EM Field
Planck quanta
Bosons
ElectroMagnetic
mass
velocity
EM waves
W
distributions
E
$E = mv^2$
1 4 9 16 25 36 49 64
15 13 11 9 7 5 3 1
Ψ
128
γ
λ
64
Photon Frequency (f)
Group Phase Velocity
longitudinal Photon Frequency (f)
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Longitudinal wave Amplitude

Wavefunction probabilites

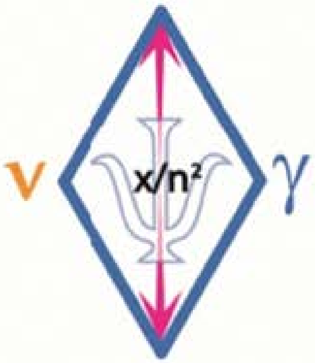

Wave Amplitude

Amplitude = bi-directional Momentum

Probability = [Amplitude]^2

In his 1926 paper, Max Born suggested that the wave function of Schrödinger's wave equation represents the probability density of finding a particle

Probability of finding a Photon in a electromagnetic wave is the Square of its Amplitude

The de Broglie-Schrodinger wave fields should not to be interpreted as a mathematical description of how an event actually takes place in time and space, though, of course, they have reference to such an event.

Rather they are a mathematical description of what we can actually know about the system. They serve only to make statistical statements and predictions of the results of all measurements which we can carry out upon the system.

(Albert Einstein, on Quantum Physics, 1940)

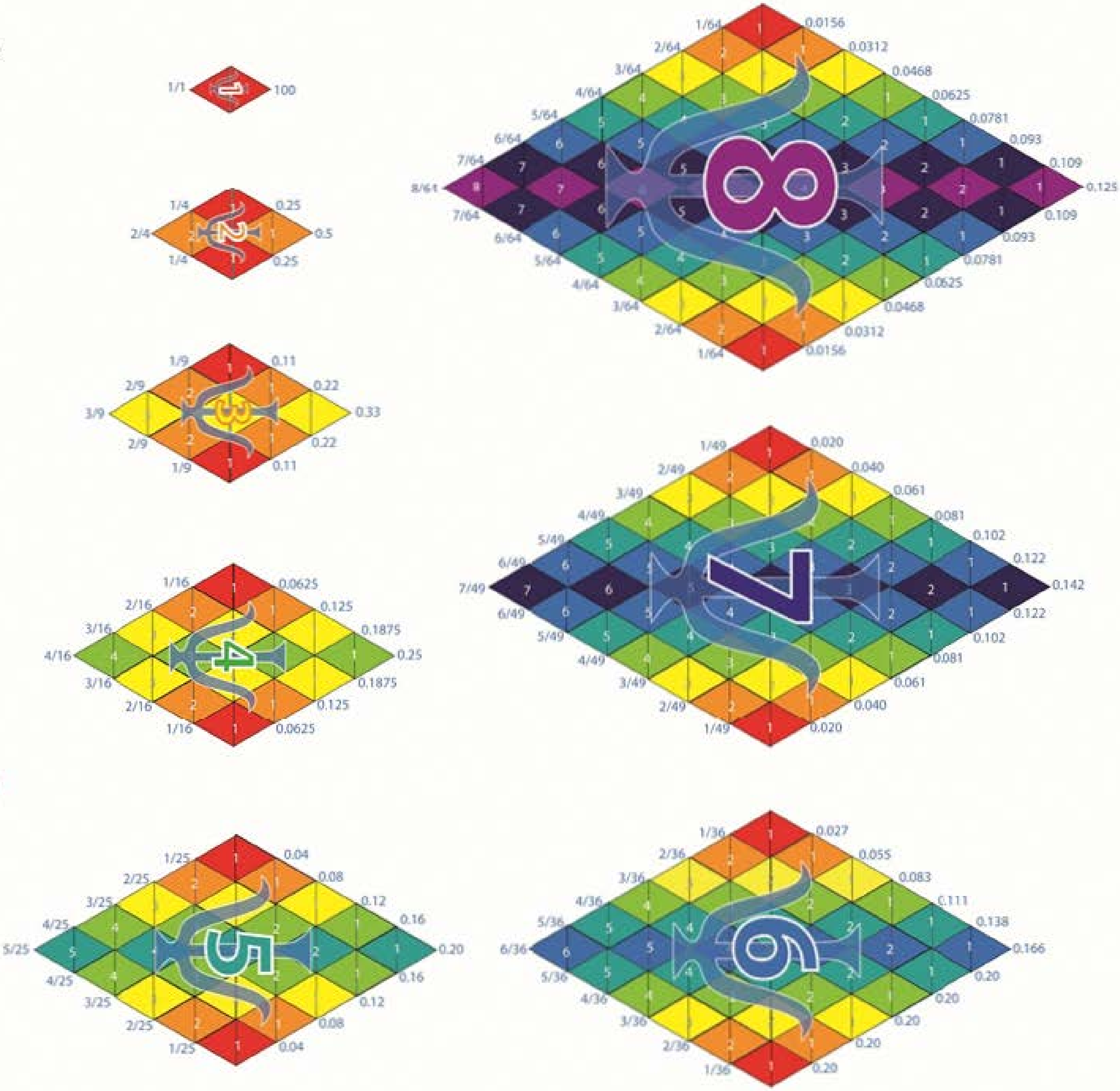
1/1
100
1/4
2/4
1/4
0.25
0.5
0.25
1/9
2/9
3/9
2/9
1/9
0.11
0.22
0.33
0.22
0.11
1/16
2/16
3/16
4/16
3/16
2/16
1/16
0.0625
0.125
0.1875
0.25
0.1875
0.125
0.0625
1/25
2/25
3/25
4/25
5/25
4/25
3/25
2/25
1/25
0.04
0.08
0.12
0.16
0.20
0.16
0.12
0.08
0.04
1/36
2/36
3/36
4/36
5/36
6/36
5/36
4/36
3/36
2/36
1/36
0.027
0.055
0.083
0.111
0.138
0.166
0.20
0.20
0.20
0.20
0.20
0.20
1/49
2/49
3/49
4/49
5/49
6/49
7/49
6/49
5/49
4/49
3/49
2/49
1/49
0.020
0.040
0.061
0.081
0.102
0.122
0.142
0.122
0.102
0.081
0.061
0.040
0.020
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
1
2
3
4
5
6
7
8

EM radiation patterns

Quantum wavefunctions mathematically model
the statistical probability distributions
of electromagnetic mass-energy
within EM waves

n1

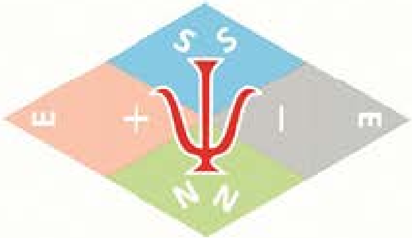

1

n2

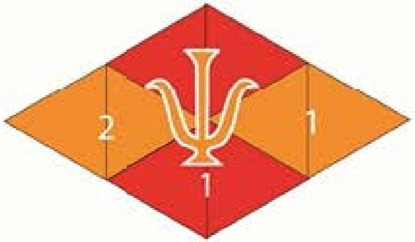

4

n3

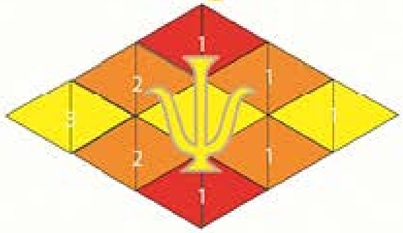

9

n4

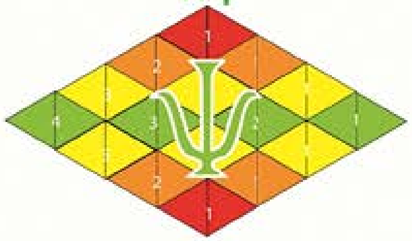

16

As the energy of an EM wave increases so does its frequency

EM waves radiate outward at the speed of light along their E fields within scribed radial time dependent spatial co-ordinates

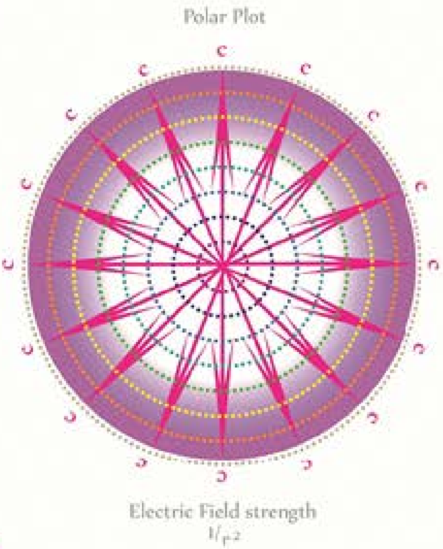
Polar Plot
c
Electric Field strength
$^1/_{r^2}$

$E = h\nu$

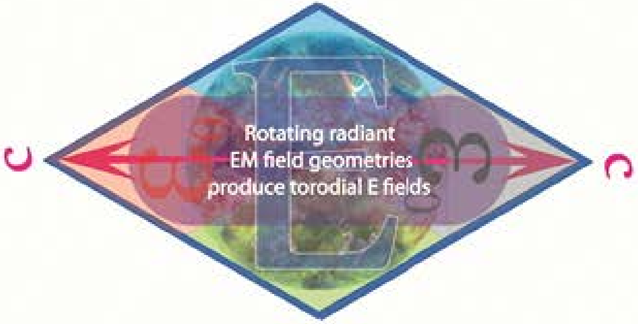
c
c
Rotating radiant
EM field geometries
produce torodial E fields

$E = hf$

$E = mv^2$

n5

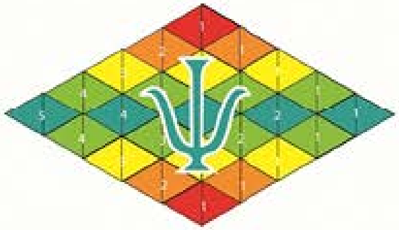

25

n8

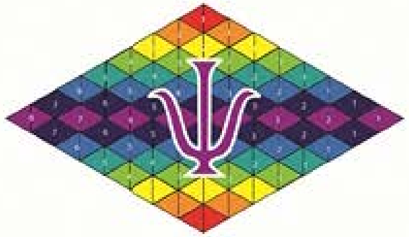

64

n7

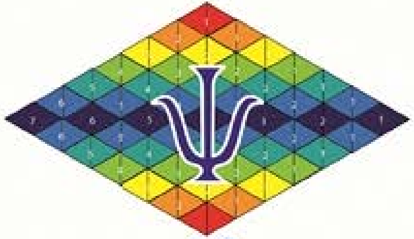

49

n6

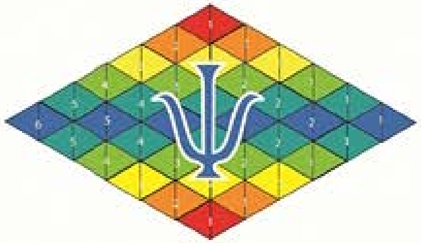

36

As the Energy and frequency increase, the Wavelengths of the Photons in a EM wave decreases

EM wave geometry

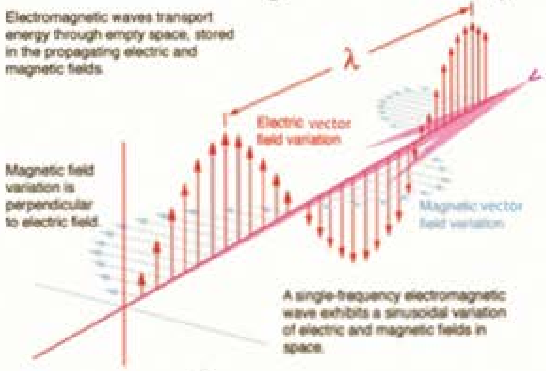
Electromagnetic waves transport energy through empty space, stored in the propagating electric and magnetic fields.

Magnetic field variation is perpendicular to electric field.

A single-frequency electromagnetic wave exhibits a sinusoidal variation of electric and magnetic fields in space.
λ
Electric vector field variation
Magnetic vector field variation

Bosons

$E = h\nu$

transverse mass-energies

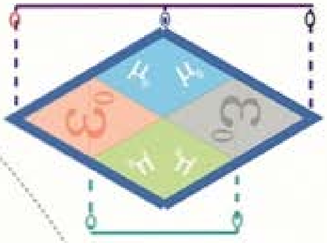

When E fields are at Maximum the B fields are at Minimum (and Vice-a-versa)

$2v = f$

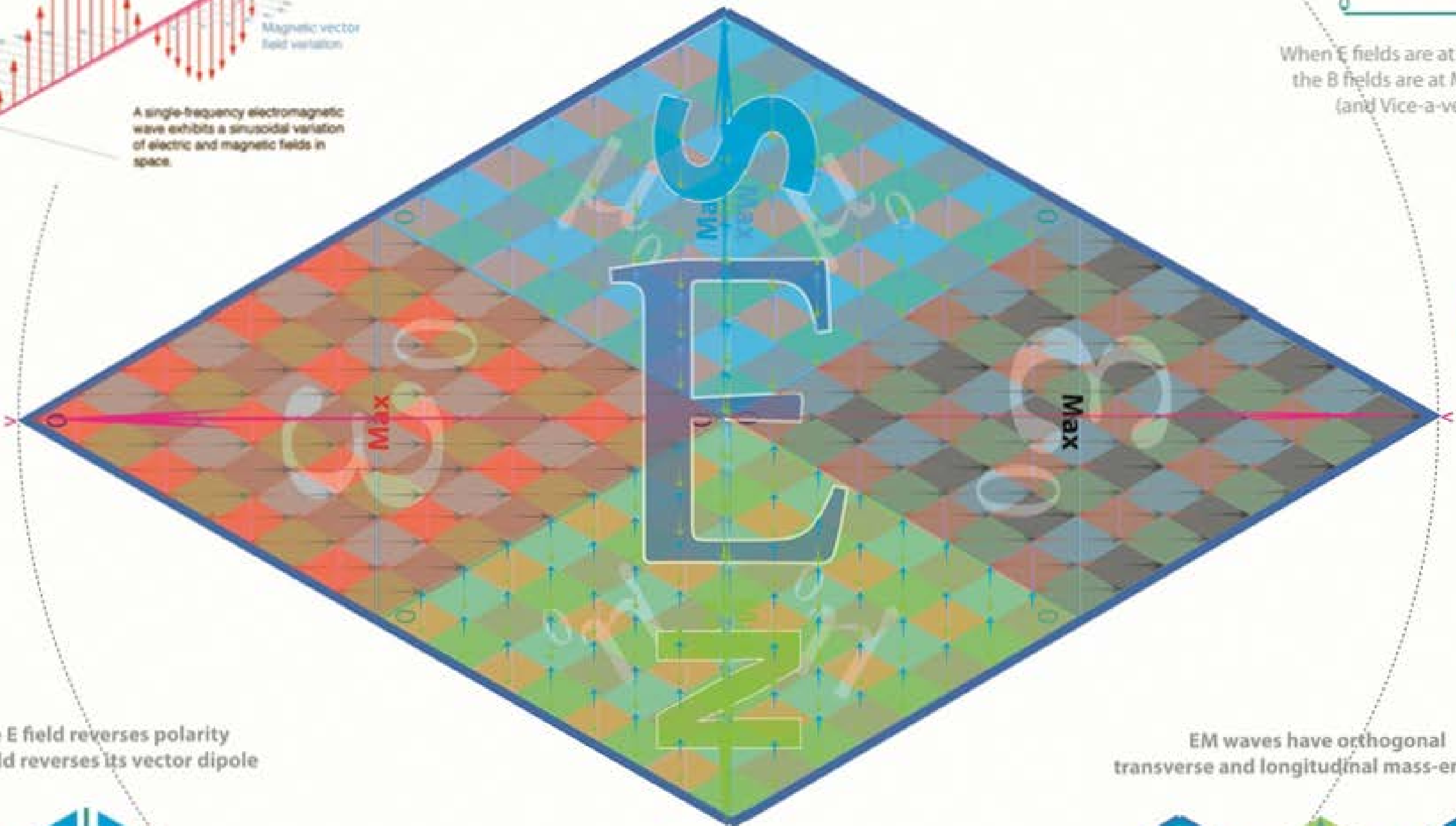
S
E
N
Max
Max
Max
0
ε0
ε0
μ0
μ0

Photons

$E = hf$

longitudinal mass-energy frequency

As the E field reverses polarity the M field reverses its vector dipole

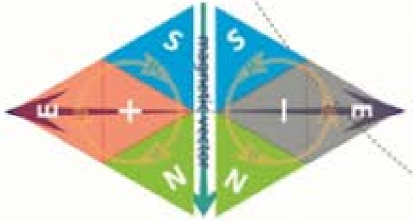
S
S
E
+
−
E
N
N
magnetic vector
E Field oscillations

EM waves are comprised of relativistic photons of mass-energy

$$RE = 2m\frac{v^2}{c^2}$$

EM waves have orthogonal transverse and longitudinal mass-energies

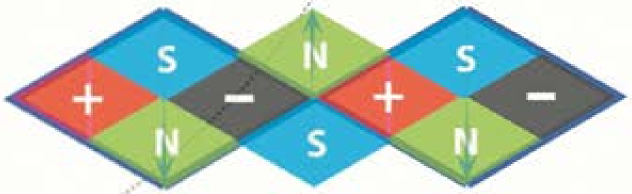

Copyright ABRAHAM [2008] - all rights reserved

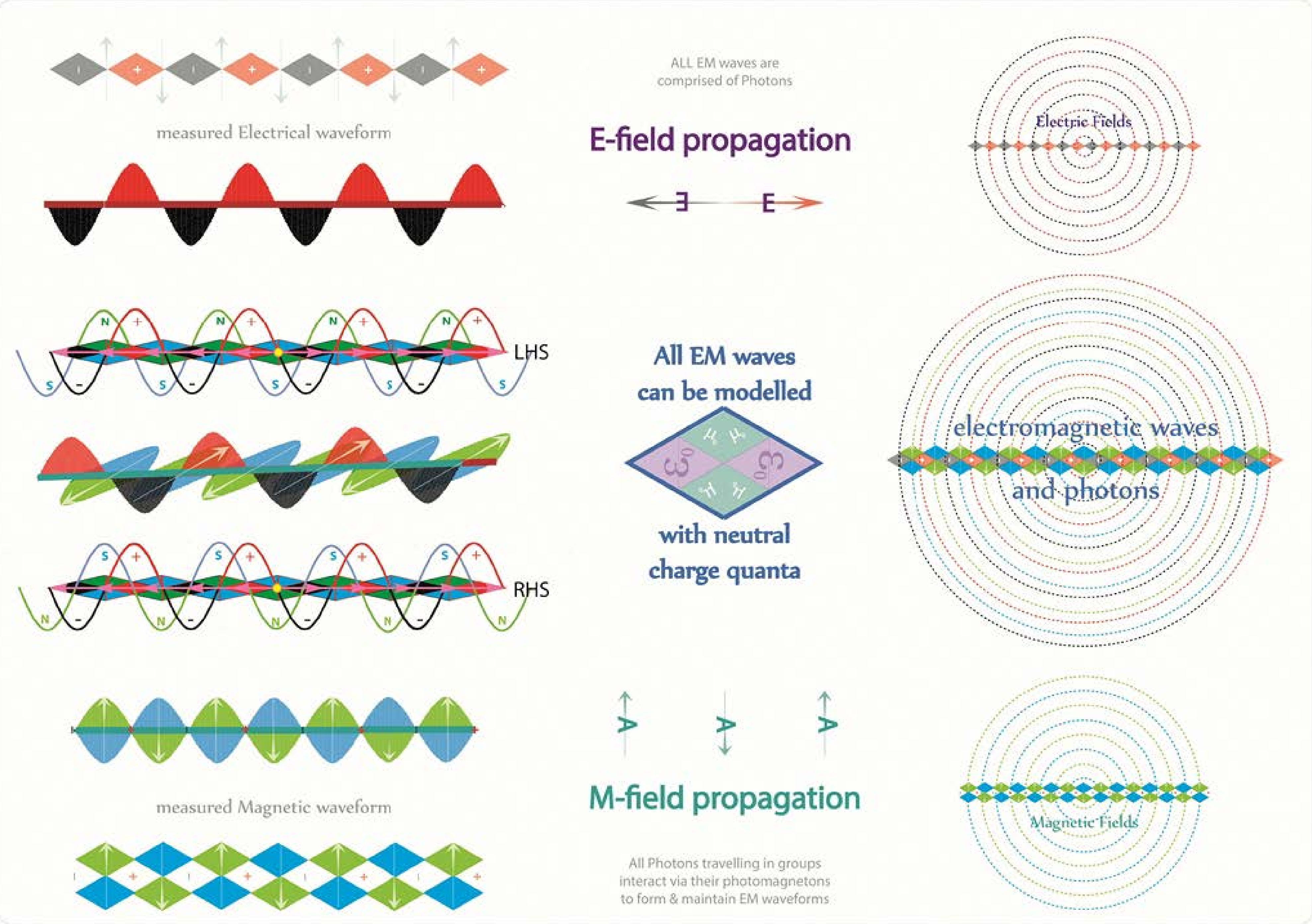
measured Electrical waveform
E
E
N
S
+
-
LHS
RHS
ε_0
μ_0
Electric Fields
electromagnetic waves
and photons
Magnetic Fields
measured Magnetic waveform

ALL EM waves are
comprised of Photons

E-field propagation

**All EM waves
can be modelled
with neutral
charge quanta**

M-field propagation

All Photons travelling in groups
interact via their photomagnetons
to form & maintain EM waveforms

Tetryonics 27.06 - E&M wave geometries

Copyright ABRAHAM [2008] - all rights reserved

Energy momenta in EM waveforms

Linear energy momentum creates physical vector forces

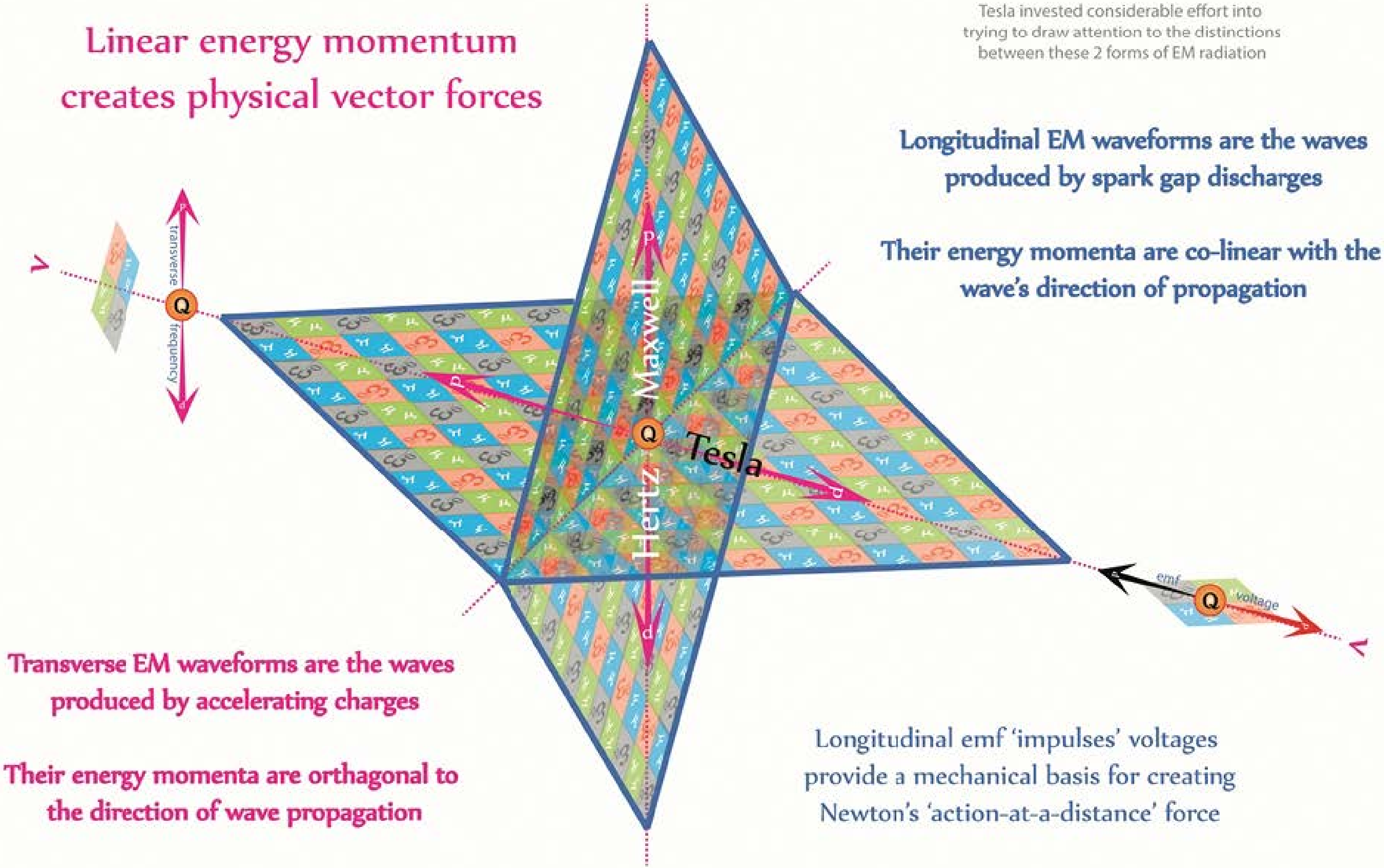

Tesla invested considerable effort into trying to draw attention to the distinctions between these 2 forms of EM radiation

Longitudinal EM waveforms are the waves produced by spark gap discharges

Their energy momenta are co-linear with the wave's direction of propagation

Transverse EM waveforms are the waves produced by accelerating charges

Their energy momenta are orthagonal to the direction of wave propagation

Longitudinal emf 'impulses' voltages provide a mechanical basis for creating Newton's 'action-at-a-distance' force

Copyright ABRAHAM [2008] - all rights reserved

Producing Longitudinal EM waves

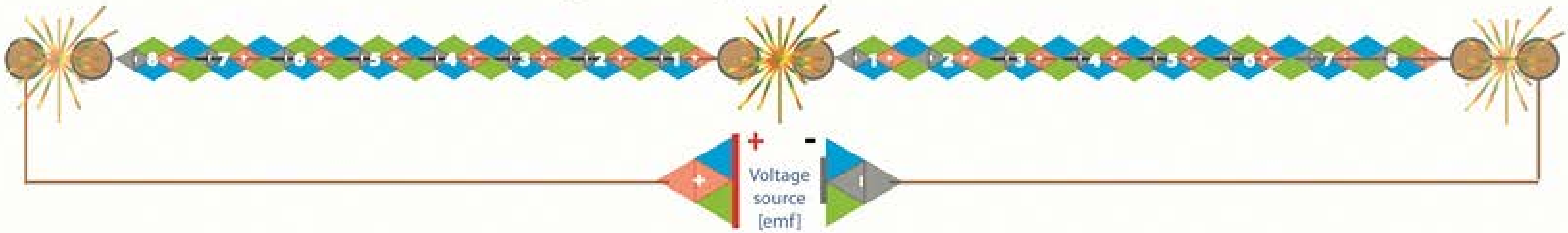

The short circuiting of voltage potentials or spark gap discharges of electrical energy are two ways to produce longitudinal EM waves

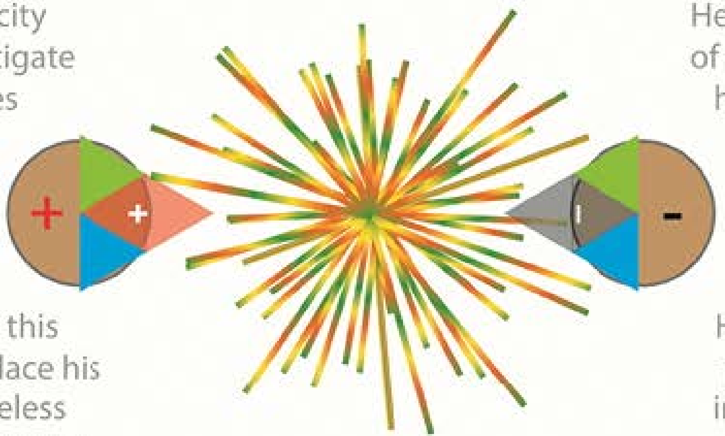

After inventing AC electricity Nikola Tesla began to investigate disruptive DC discharges

Orthagonal Magnetic fields produce reactive inductance fields

and he was convinced that this form of EM energy would replace his AC system and provide wireless energy to the World without losses

He noted the deadly power of the spark discharges and how they 'stung his face'

Longitudinal Electric fields produce lines of force

He strove to point out that the load circuits must be inductively coupled to the transmission circuits

or a combination of the two

Very large Voltage

High frequency

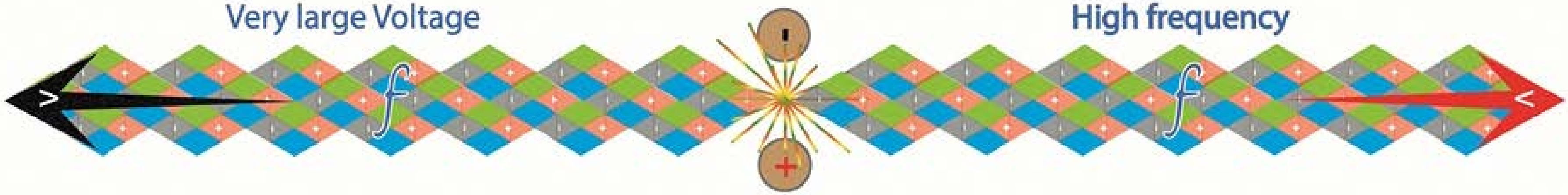

disruptive discharges

Longitudinal energy momenta

Maxwell's equations lead to the prediction of transverse electromagnetic waves as a means of transefering electrical power (where the wave's Electric fields and Magnetic fields vary perpendicularly to their direction of propagation).

Longitudinal Photons & EM waves are 2D electromagnetic waveforms with co-planar energy momenta that is co-linear with their directions of propagation and that can 'cut like a knife'

Longitudinal energy momenta waves [created by the use of spark gap technologies] are the EM energy forms created and radiated by Stars

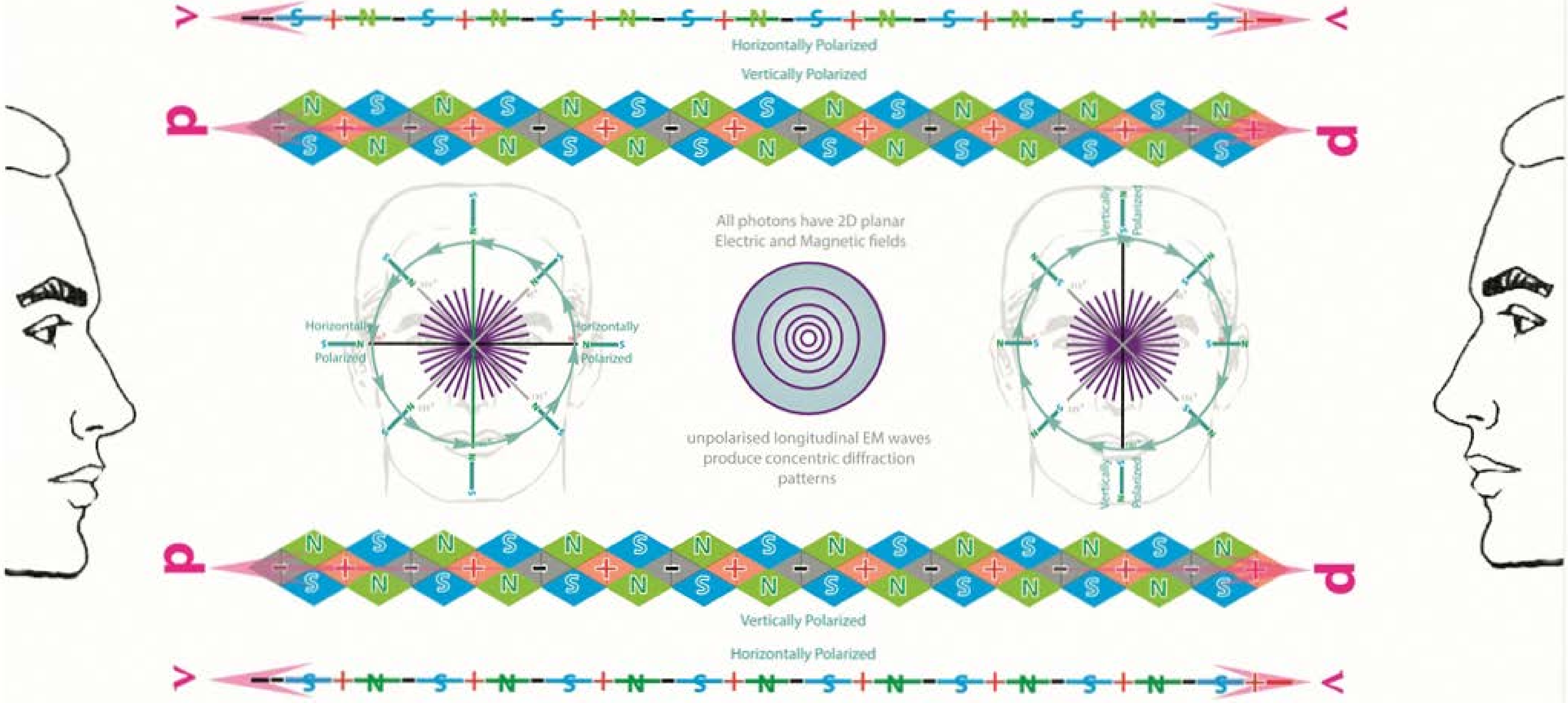
v
v
Horizontally Polarized
Vertically Polarized
p
p
N
S
+
-
Horizontally Polarized
Horizontally Polarized
Vertically Polarized
Vertically Polarized
Vertically Polarized
Horizontally Polarized
p
p
v
v

All photons have 2D planar Electric and Magnetic fields

unpolarised longitudinal EM waves produce concentric diffraction patterns

The longitudinal alignment of E field momenta with their direction of propagation produces 'action-at-a-distance' and can produce material 'faster than light' velocities, if sufficient energies are inputted

Copyright ABRAHAM [2008] - all rights reserved

The velocity of Electrical energy momenta

The velocity of light is the limit for electrical energy acceleration

(the induced velocity of charged particles resulting from electric field energy-momenta interactions)

c

Transverse EM waves
propagate at energy at
the speed of light

Γ

Longitudinal EM waves
can propagate energies & information
in excess of the speed of light

470,912,891 m/s

463,491,072 km/second

288,000 miles/second

$v = [\pi/2]c$

Longitudinal electrical fields
can create near-instantaneous
'action-at-a-distance' EM fields

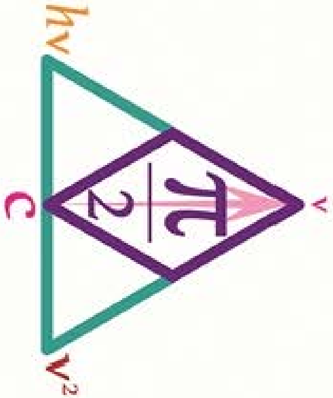

once established, these fields can
transfer momentum and information
at speeds faster than light through
co-linear impulses of momentum

Wheatstone achieved renown by a great experiment
The measurement of the velocity of electrical Energy in a wire.

He cut the wire at the middle, to form a gap which a spark might leap across, and connected its ends to the poles of a Leyden jar filled with electricity. Three sparks were thus produced, one at either end of the wire, and another at the middle. He mounted a tiny mirror on the works of a watch, so that it revolved at a high velocity, and observed the reflections of his three sparks in it.

The points of the wire were so arranged that if the sparks were instantaneous, their reflections would appear in one straight line; but the middle one was seen to lag behind the others, because it was an instant later. The electricity had taken a certain time to travel from the ends of the wire to the middle.

This time was found by measuring the amount of lag, and comparing it with the known velocity of the mirror. Having got the time, he had only to compare that with the length of half the wire, and he could find the velocity of electricity.

His results gave a calculated velocity of 288,000 miles per second,
i.e. faster than what we now know to be the speed of light

Charles Wheatstone

(6 February 1802 – 19 October 1875)

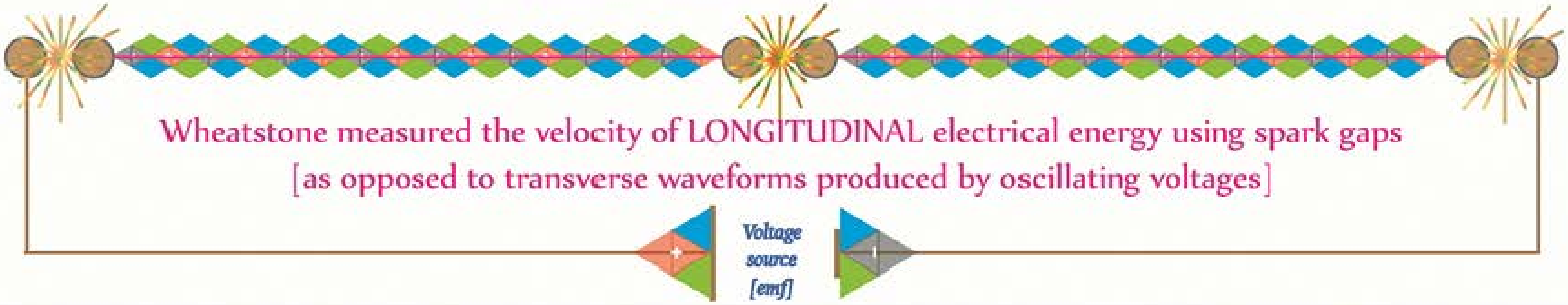

Wheatstone measured the velocity of LONGITUDINAL electrical energy using spark gaps
[as opposed to transverse waveforms produced by oscillating voltages]

Copyright ABRAHAM [2008] - all rights reserved

'Action at a Distance'

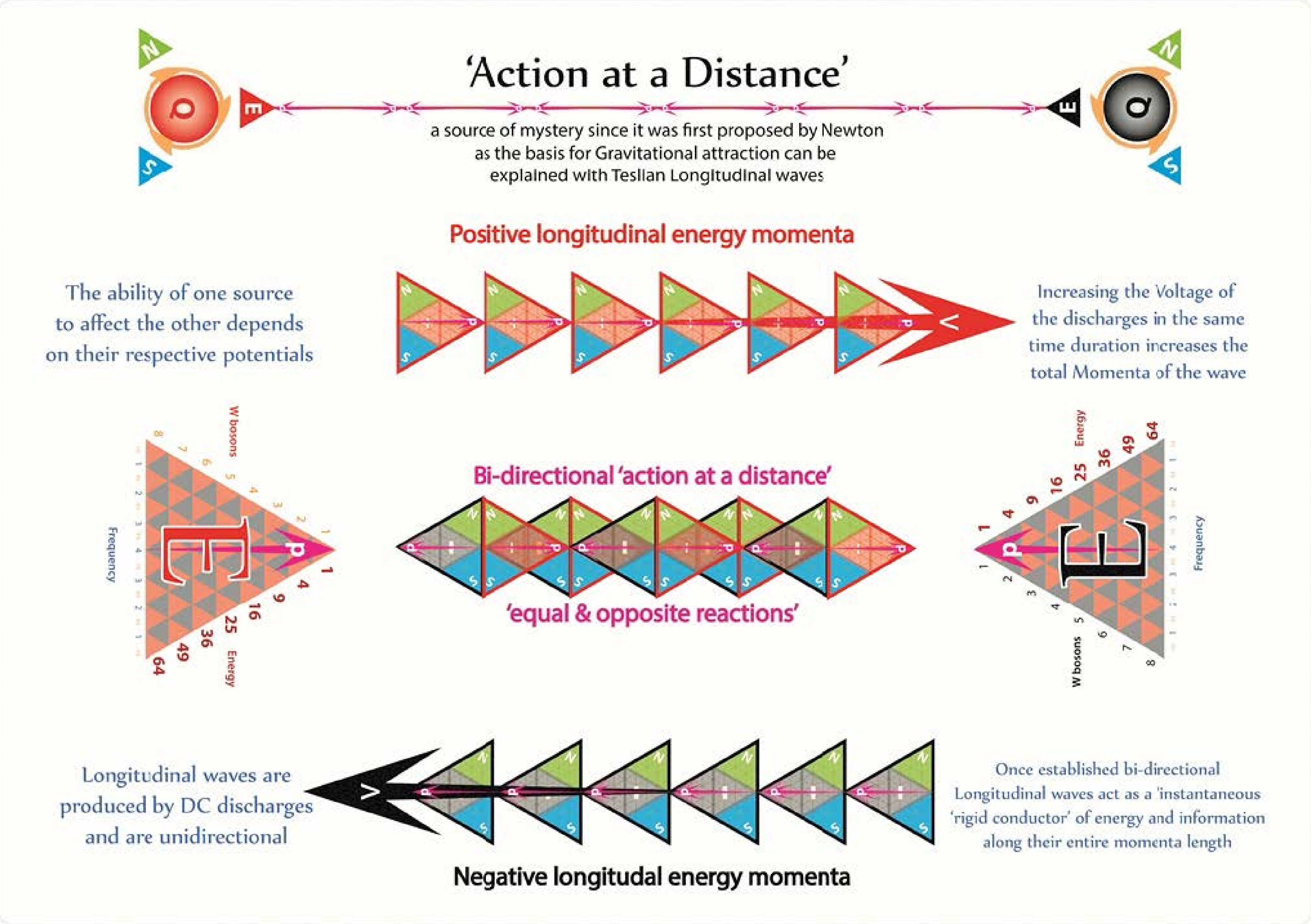

a source of mystery since it was first proposed by Newton
as the basis for Gravitational attraction can be
explained with Teslian Longitudinal waves

Positive longitudinal energy momenta

The ability of one source
to affect the other depends
on their respective potentials

Increasing the Voltage of
the discharges in the same
time duration increases the
total Momenta of the wave

Bi-directional 'action at a distance'

'equal & opposite reactions'

Longitudinal waves are
produced by DC discharges
and are unidirectional

Once established bi-directional
Longitudinal waves act as a 'instantaneous
'rigid conductor' of energy and information
along their entire momenta length

Negative longitudal energy momenta

Tetryonics 27.11 - Longitudinal 'action at a distance'

Copyright ABRAHAM [2008] - all rights reserved

Longitudinal EM waves have their E-fields co-linear with their direction of propagation

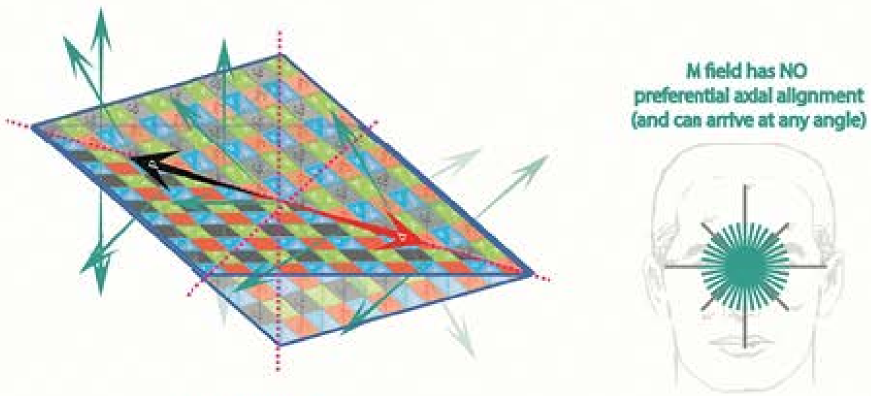

M field has NO preferential axial alignment (and can arrive at any angle)

Un-Polarised

Longitudinal Photons & EM waves are produced by disruptive spark gap discharges

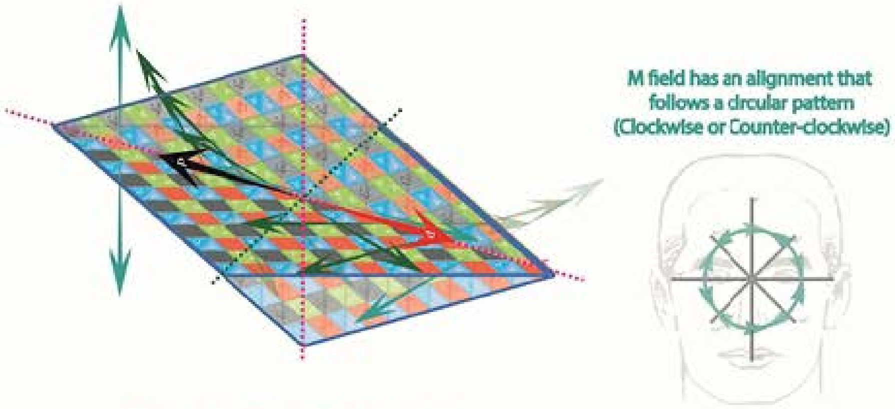

M field has an alignment that follows a circular pattern (Clockwise or Counter-clockwise)

Circularly (or Elliptically) Polarised

Longitudinal wave polarisations

emf
Q
voltage

As a convention, the polarization of longitudinal EM waves must be described by specifying the orientation of the wave's magnetic field at a point in space over one period of the oscillation

Longitudinal waves have their linear momenta aligned along their direction of propagation and are extremely efficient at accelerating charges they interact with, creating co-linear accelerations [parallel to their direction of field propagation] - they do not produce charge oscillations along their vector direction of propagation like transverse waves.

Examples of transverse waves include seismic P (primary) waves, and Tesla's longitudinal 'stinging' rays, whose electric fields and energy momenta are both co-linear with each other and parallel to the direction of energy transfer.

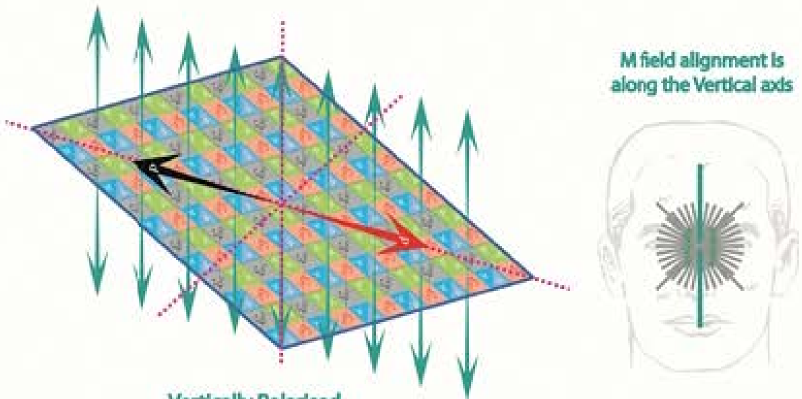

M field alignment is along the Vertical axis

Vertically Polarised

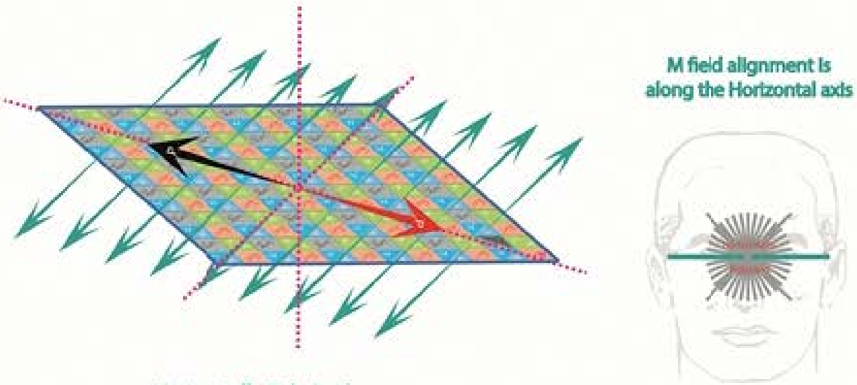

M field alignment is along the Horizontal axis

Horizontally Polarised

Tetryonics 27.12 - Longitudinal wave polarisation

ElectroMagnetic waveforms

In Teslian longitudinal waves, produced by disruptive EM discharges,
the E fields are co-linear with the direction of propagation

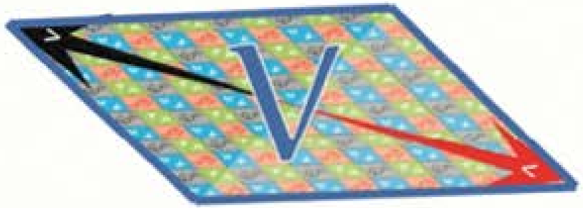

After inventing AC electricity Tesla investigated transferring energy
through the air via electrostatic longitudinal waves to receiving devices.
and noted the differing 'lethality' of the forces produced by these waves

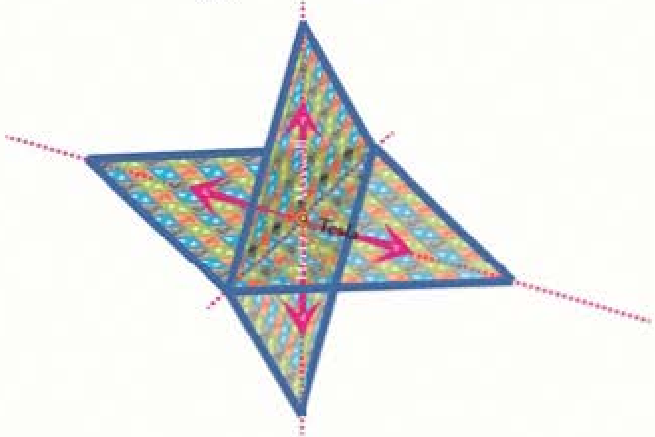

Although comprised of the same energy momenta quanta
EM waves can be formed with Electric fields perpendicular to each other
thus leading to conflicting theories of EM wave propagation

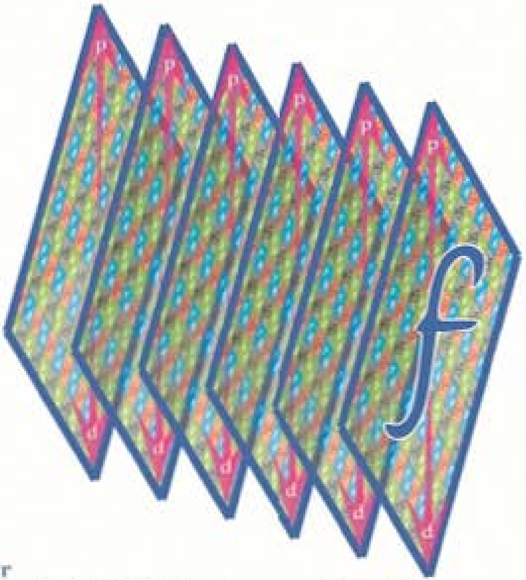

Hertzian transverse waveforms have their EM fields
perpendicular to the direction of propagation

In 1887, Heinrich Hertz demonstrated the reality of
Maxwell's electromagnetic waves by experimentally
generating radio waves in his laboratory

T

Teslian waves
transmit Energy
in a LONGITUDINAL
waveform
producing
'action-at-a-distance'

Nikola Tesla

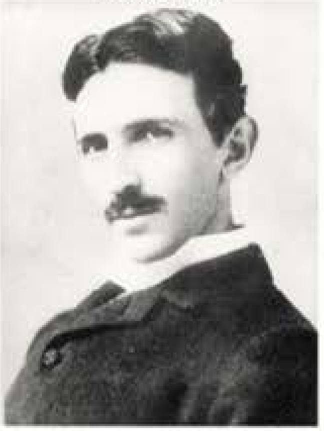

(10 July 1856 – 7 January 1943)

Volt Seconds [per metre²]

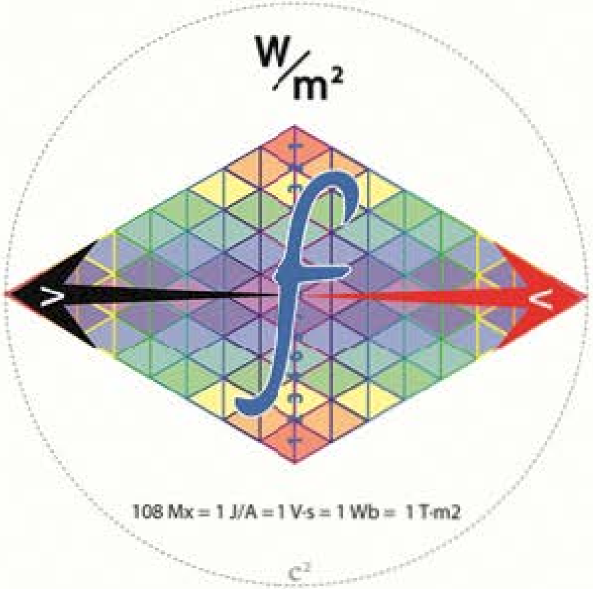

108 Mx = 1 J/A = 1 V·s = 1 Wb = 1 T·m2

$\frac{c^2}{sec}$

Heinrich Hertz

(22 Feburary 1857 - January 1 1894)

Cycles per Second

Hz

Hertzian waves
transmit Energy
in a TRANSVERSE
waveform
producing
'radio waves'

Soon after Hertz's claim of discovering Maxwell's transverse EM waves Tesla visited him and personally demonstrated the experimental error to him.
Hertz agreed with Tesla and had planned to withdraw his claim, but varying agendas intervened and set the stage for a major rift in the 'accepted' theories

Copyright ABRAHAM [2008] - all rights reserved

Transverse EM waves have their E&M fields orthagonal with their direction of propagation

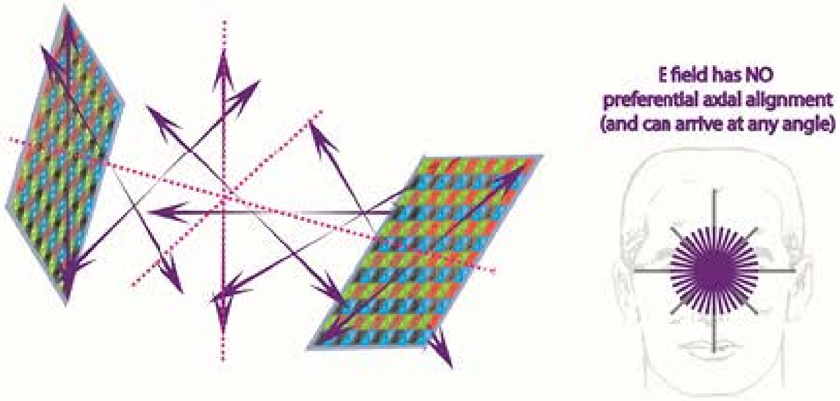

E field has NO preferential axial alignment (and can arrive at any angle)

Un-Polarised

Transverse Photons & EM waves are produced by accelerating/oscillating charges

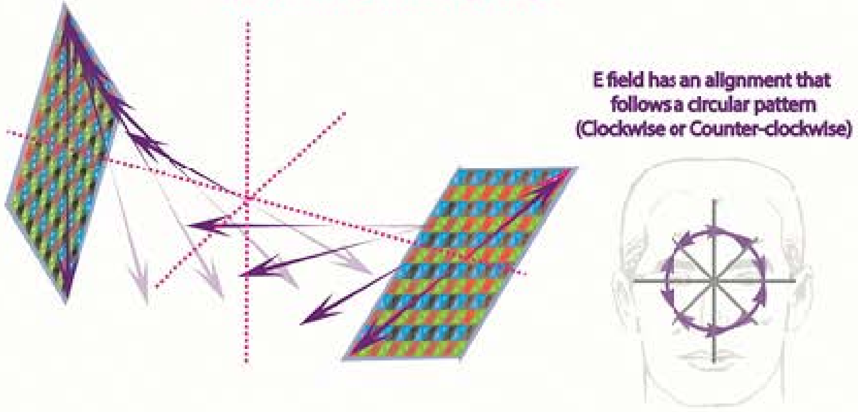

E field has an alignment that follows a circular pattern (Clockwise or Counter-clockwise)

Circularly (or Elliptically) Polarised

transverse
Q
frequency

Transverse wave polarisations

By convention, the polarization of EM waves is described by specifying the orientation of the wave's electric field at a point in space over one period of the oscillation

Transverse waves are inefficient propagators of EM energy as they apply their linear momenta tangentially to the charges they interact with, creating lateral accelerations [orthagonal to their direction of field propagation] - they do not accelerate charges along their vector direction of propagation like longitudinal waves.

Examples of transverse waves include seismic S (secondary) waves, and electromagnetic plane wave, whose electric (E) and magnetic (M) fields both oscillate perpendicular to each other and orthogonal to the direction of energy transfer.

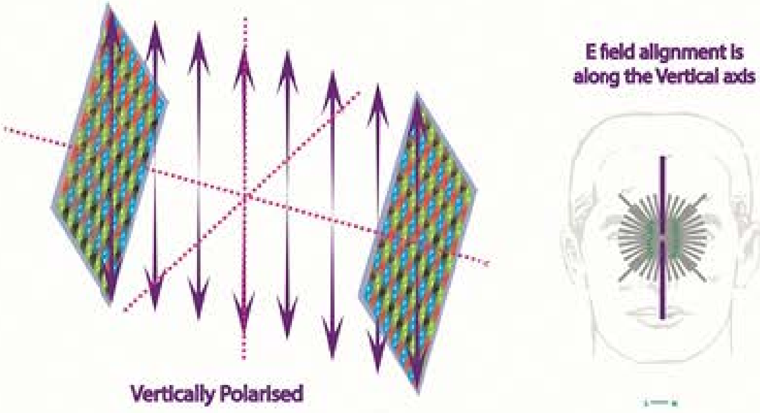

E field alignment is along the Vertical axis

Vertically Polarised

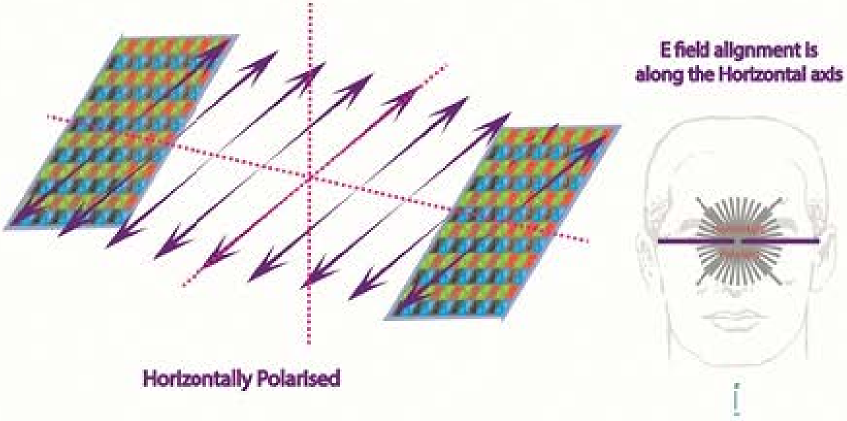

E field alignment is along the Horizontal axis

Horizontally Polarised

Tetryonics 27.14 - Transverse wave polarisation

EM wave modulations

In telecommunications, modulation is the process of varying one or more properties of a periodic waveform, called the carrier signal, with a modulating signal which typically contains information to be transmitted

Amplitude modulation

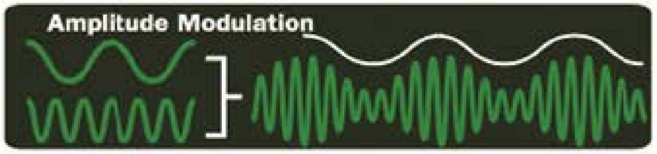
Amplitude Modulation

the amplitude of the carrier signal is varied

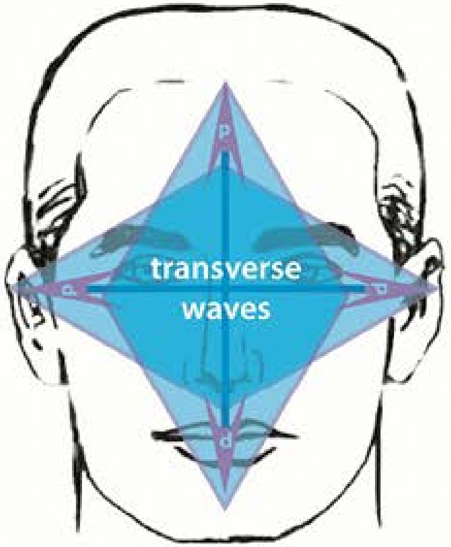
transverse
waves

the frequency of the carrier signal is varied

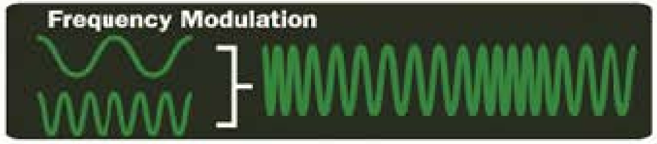
Frequency Modulation

Frequency modulation

In addition to planar Hertzian waves
there exists Longitudinal EM waves

Impulse propagation

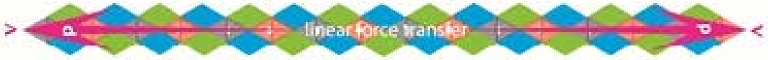
linear force transfer

As the voltage [energy/coulomb] of longitudinal waves increases their co-linear energy-momenta increase and these forces provide a mechanical basis for near instantaneous 'action-at-a-distance'

Impulse propagation is distinct from pulse modulations

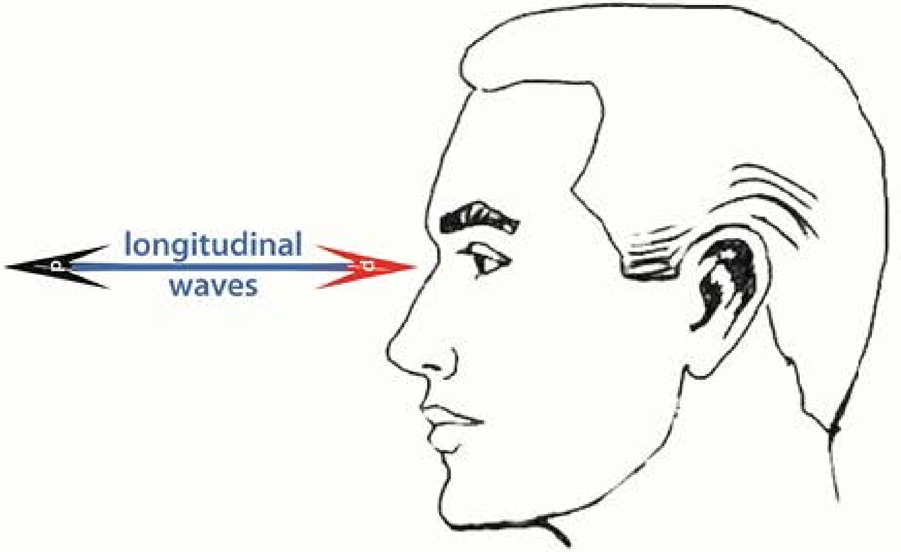
longitudinal
waves

a narrowband analog signal is pulsed over an analog baseband channel

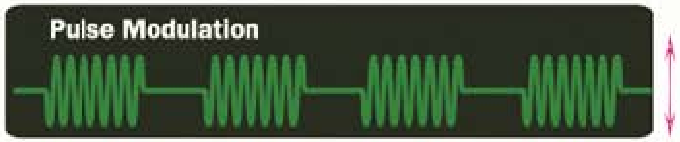
Pulse Modulation

PCM Pulsed modulation **TDM**

Copyright ABRAHAM [2008] - all rights reserved

Transverse EM wave production

Alternating the Voltage potentials in an electrical circuit produces an Alternating Current
of electrons with Kinetic EM energies reflective of the Time-Energy duration of the AC circuit

Alternating Voltages provide an electromotive force to Electrically charged Particles producing tranverse Kinetic Energy fields with orthagonal Magnetic fields

Kinetic Energies from motion
(diverge from rest Matter)

Kinetic
Matter

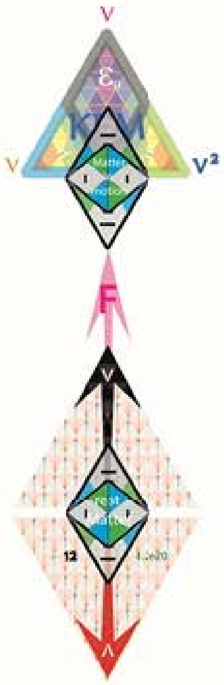
v
ε0
KEM
v
v²
Matter
motion
F
rest
12

All Stationary charged particles have
NO extrinsic Magnetic moment
ie. They are Electro-static

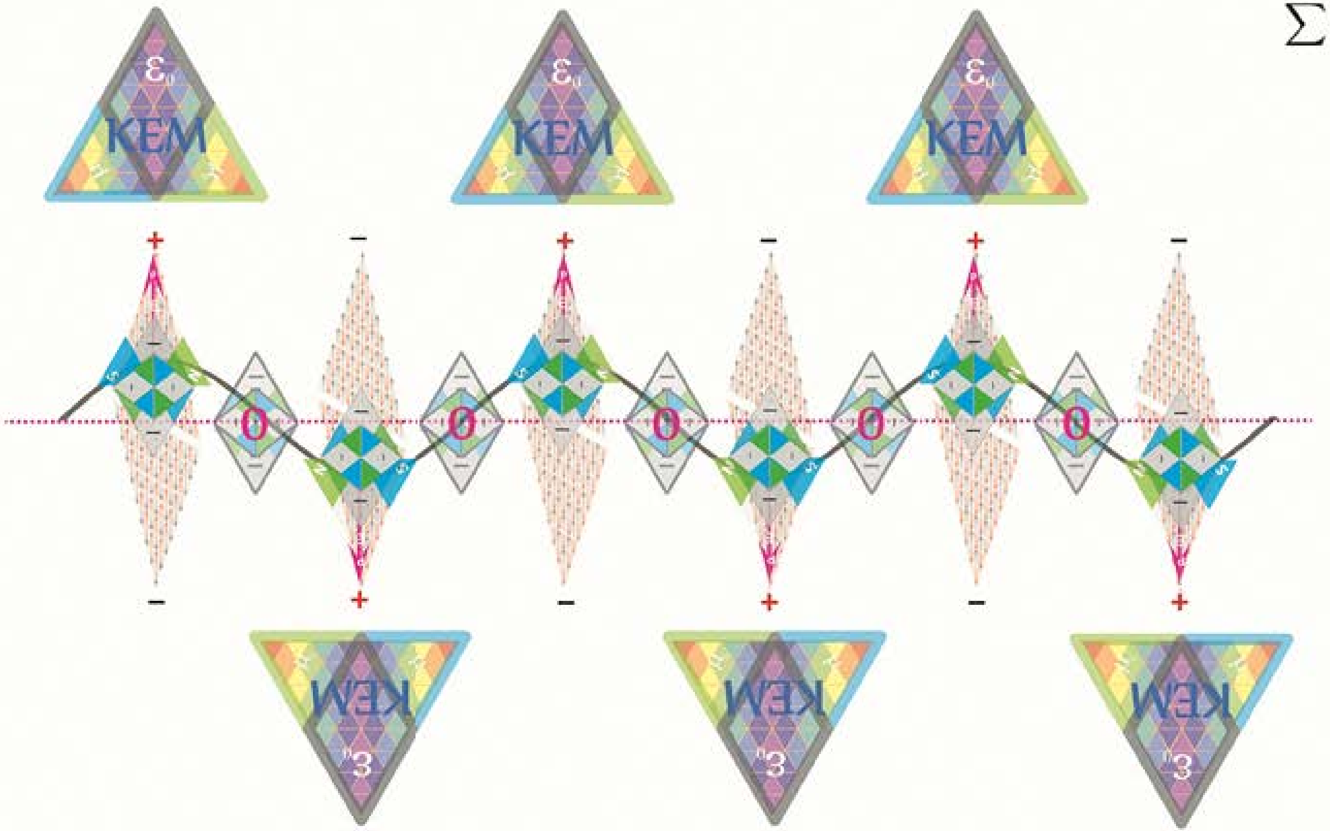
ε0
KEM
ε0
KEM
ε0
KEM
+
−
+
−
+
−
0
0
0
0
0
−
+
−
+
−
+
KEM
ε0
KEM
ε0
KEM
ε0

F=ma

$$\sum \mathbf{F} = \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} + \mathbf{v}\frac{\mathrm{d}m}{\mathrm{d}t}$$

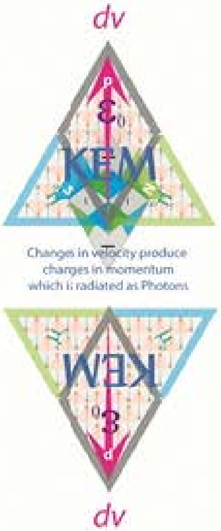
dv
ε0
KEM
Changes in velocity produce
changes in momentum
which is radiated as Photons
KEM
ε0
dv

Charges with constant velocities
produce
REACTIVE [K]EM FIELDS

Any change to the momentum of a charged particle in motion requires a corresponding quantum level energy-momentum change through the emission/absorption of quantised EM mass-ENERGY momenta in the form of W bosons

Accelerating Charges
produce
RADIANT EM WAVES

Tetryonics 27.16 - Transverse EM wave production

Modelling of planar transverse waves

(of varying amplitudes)

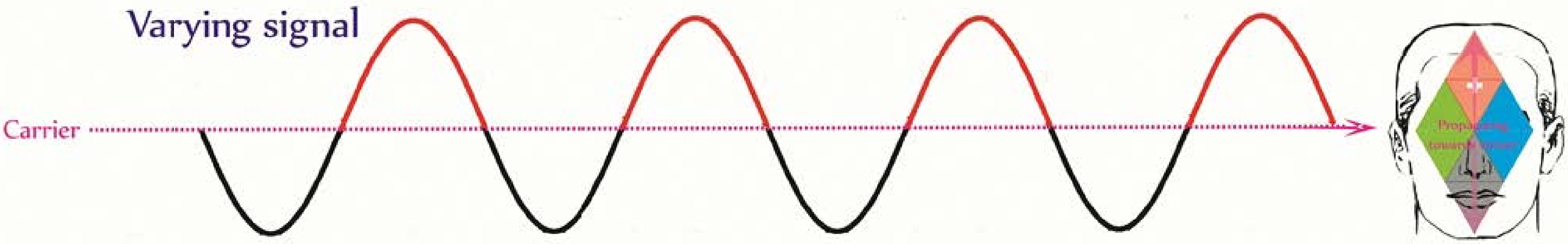
Varying signal
Carrier

AM waveform

Amplitude Modulation (AM) works by varying the strength of the transmitted signal in relation to the information being sent.

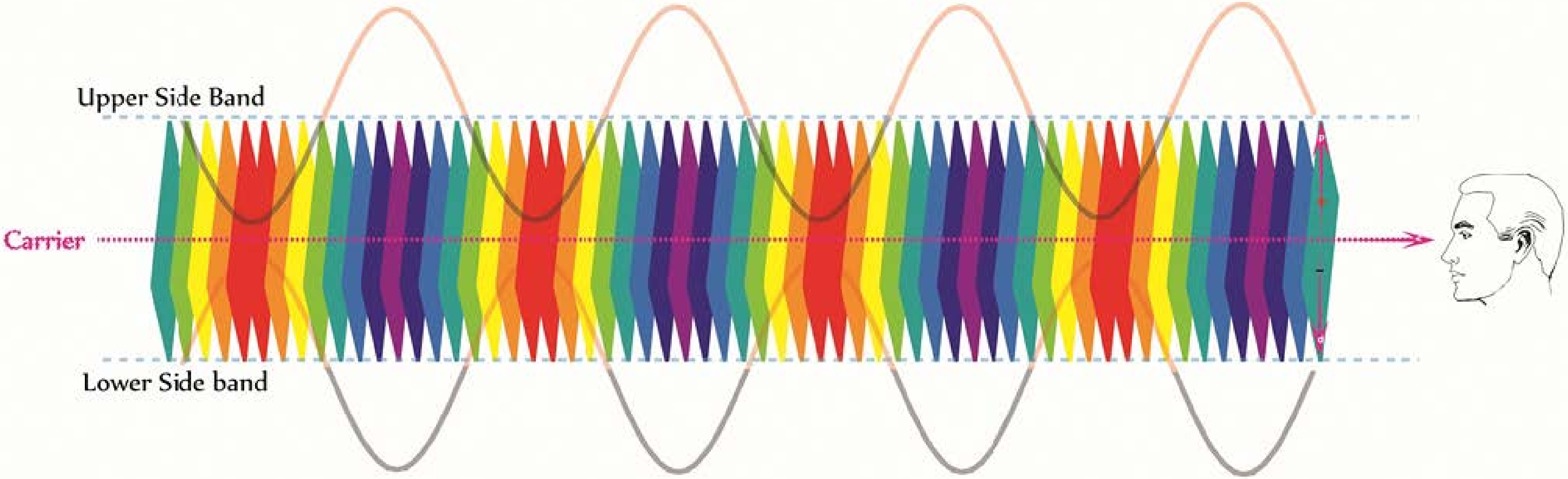
Upper Side Band
Carrier
Lower Side band

Note: In amplitude modulation each colour planar wave contains differing energy momenta [n1-red to n8-violet]

Tetryonics 27.17 - Amplitude Modulation

Modelling of planar transverse waves

(of the same frequencies)

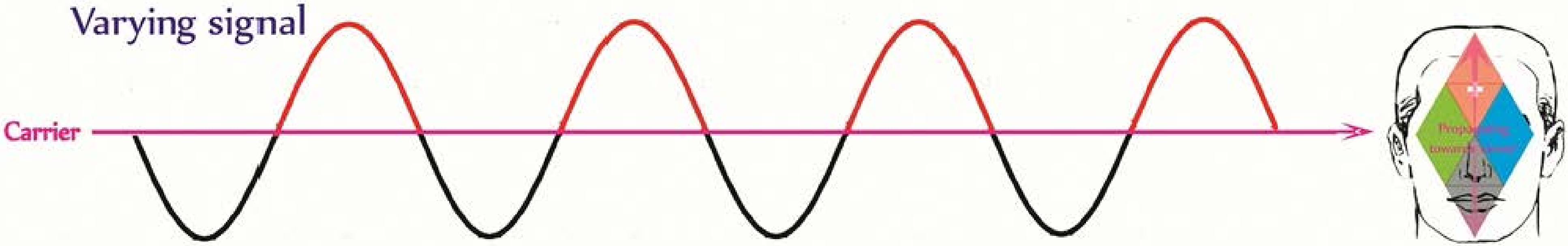

FM waveform

Frequency modulation (FM) is a form of modulation which conveys information over a carrier wave by varying its frequency (in contrast with amplitude modulation, in which the amplitude of the carrier is varied while its frequency remains constant).

In analog applications, the instantaneous frequency of the carrier is directly proportional to the instantaneous value of the input signal

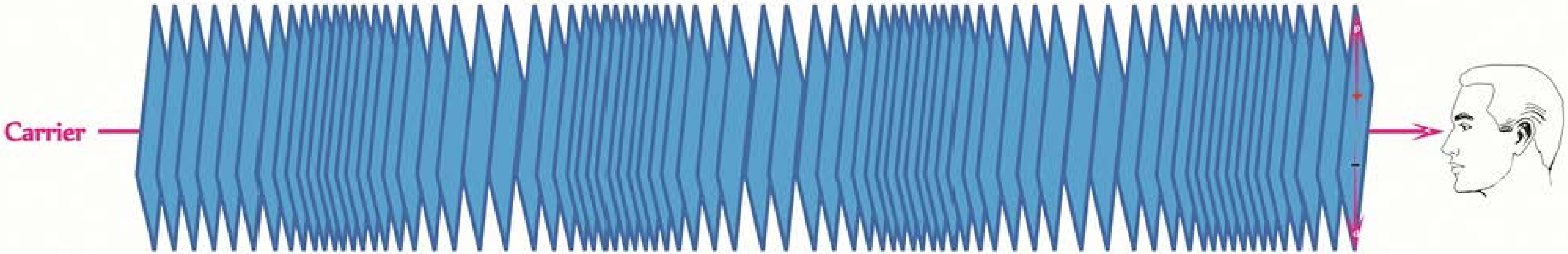

Note: In Frequency modulation all the planar waves have the same energy momenta [only the number of waves per sec varies]

Copyright ABRAHAM [2008] - all rights reserved

ElectroMagnetic waves

Mathematics alone [without a correct foundational geometry] is insuffcent to describe EM radiation
EM waves can be Longitudinal or Transverse with respect to their direction of propagation

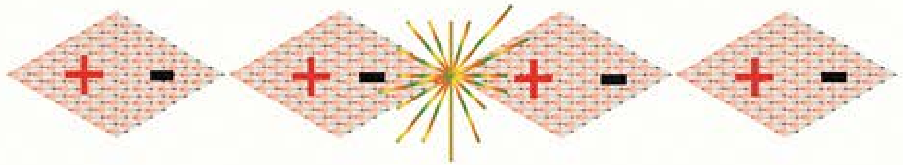

Tesla

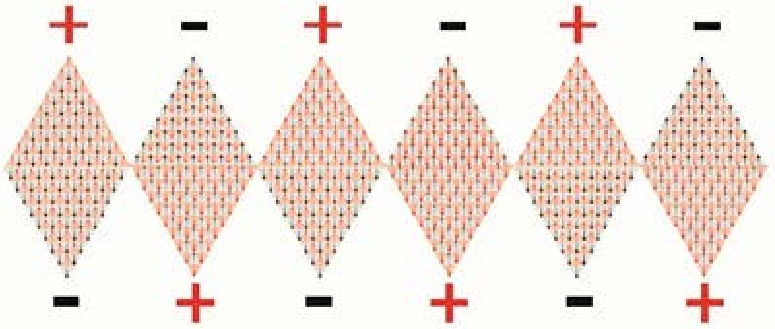

Maxwell - Hertz

It is now clear [using Tetryonic geometry] that both Tela and Maxwell were correct in their opposing theories of EM waveforms
with Tesla proposing Longitudinal waves [from experiment] and Maxwell proposing Transverse waves [from his equations].

The differing opinions on the EM waveforms is shown to be a direct result of the sources of EM wave generation employed in their creation.
After the discovery of the photo-electric effect Hertz and others moved to oscillating charges to produce Transverse [planar] waves
while Tesla continued to investigate and utilise the older spark gap technology to produce Longitudinal waves

Spark Gaps & Plasma discharges
produce
longitudinal EM radiation

Accelerating charges
produce
transverse EM radiation

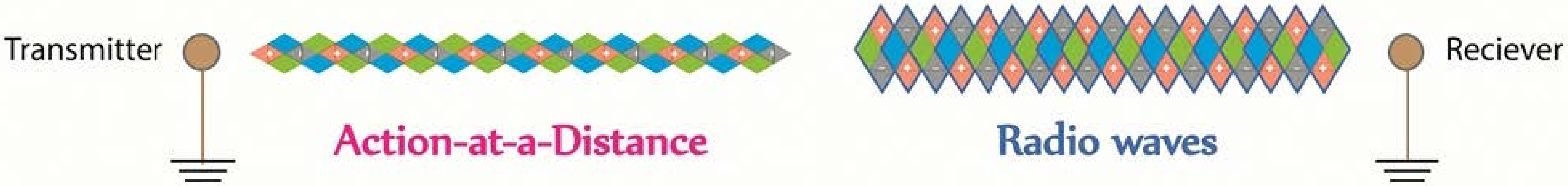

Action-at-a-Distance

Radio waves

In order to explain and advance our understanding of all EM phenomena we must return to, and utilise, the older discharge technologies

Copyright ABRAHAM [2008] - all rights reserved

Planar electromagnetic waves

In the physics of wave propagation, a plane wave (also spelled planewave) is a wave whose wavefronts are infinite parallel planes of peak-to-peak amplitude normal to the phase velocity vector.

transverse Radio waves

longitudinal Spectral lines

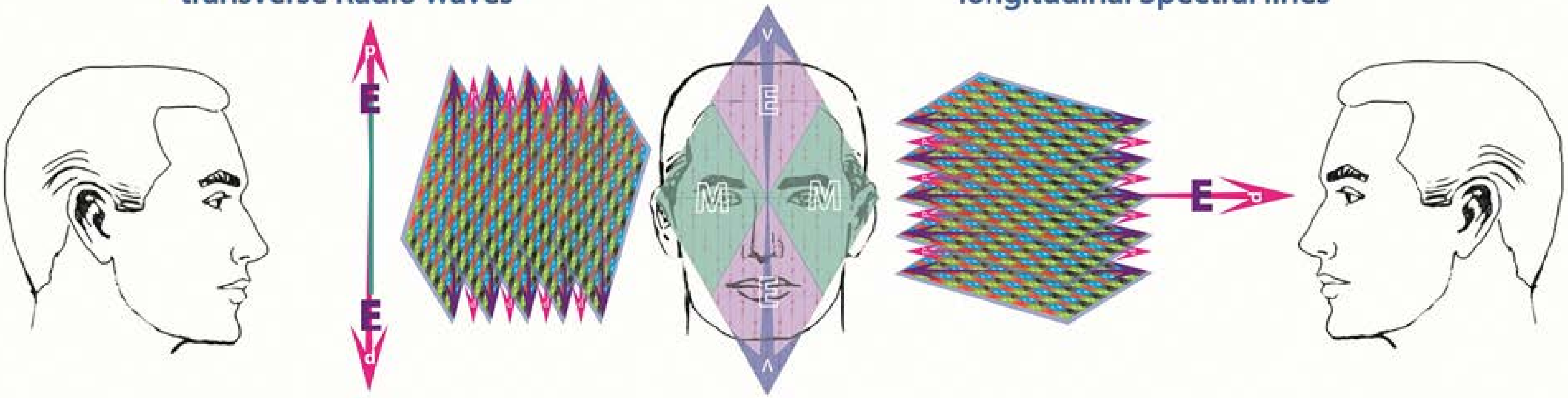
p
E
E
p
v
E
M
M
E
v
E
p

The phase relationships of their EM energy momenta can all be finally visualised and drawn through Tetryonic theory geometries

[with appropriate radian units]

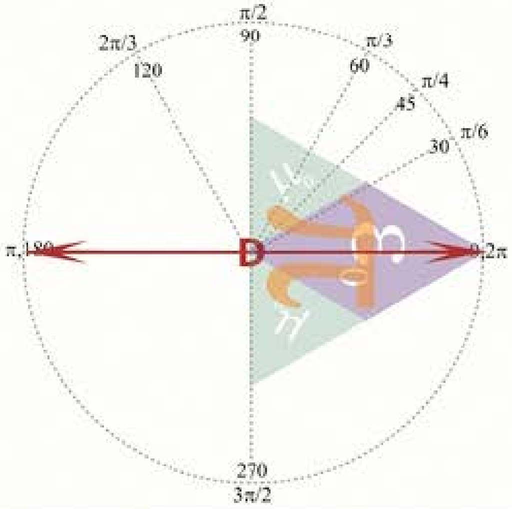
π/2
90
2π/3
120
π/3
60
π/4
45
π/6
30
π,180
D
π
0,2π
270
3π/2

bosons $2\pi[hv] = hf$ photons

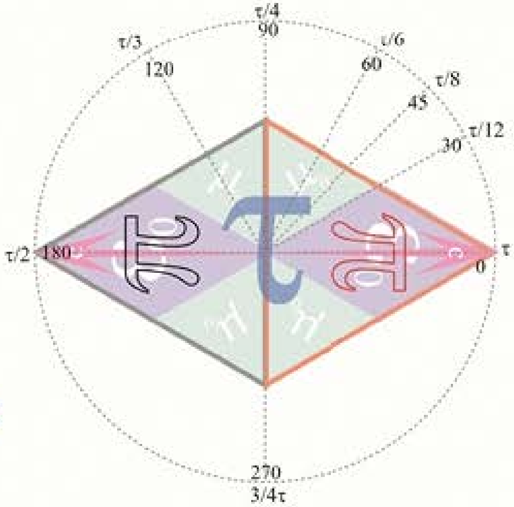
τ/4
90
τ/3
120
τ/6
60
τ/8
45
τ/12
30
τ/2
180
π
τ
π
0
τ
270
3/4τ

τ is a more 'natural' radian unit for describing EM waves than π
tau = 2π = 360 degree rotation about a point

Lorentz co-ordinate transforms

In physics, the Lorentz transformation (or transformations) are named after the Dutch physicist Hendrik Lorentz.

The transformations describe how measurements of mass-energies in space and time by two observers are related through mathematics

It was the result of attempts by Lorentz and others to explain how the speed of light was observed to be independent of the reference frame, and to understand the symmetries of the laws of electromagnetism.

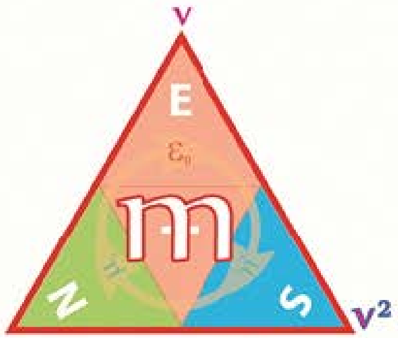

All photons are comprised of two dual-sided equilateral Planck quanta

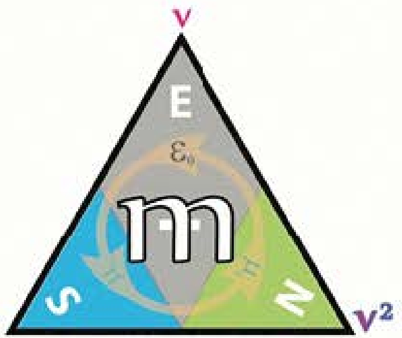

All EM waves are comprised of neutral charge quanta called Photons

Photons are comprised of mass-energies and the term 'mass-less' is not appropriate

Forces add mass-energies to KEM fields resulting in WAVElength changes

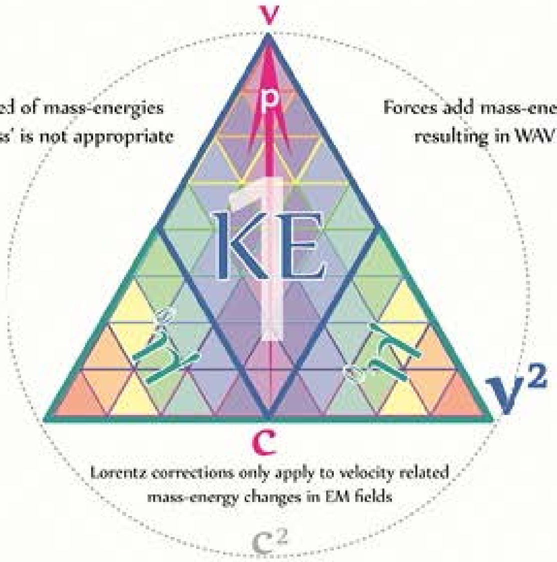

Lorentz corrections only apply to velocity related mass-energy changes in EM fields

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Linear Lorentz boosts produce changes to the velocity and momenta of a system

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

As the velocity and momenta of a system changes so does its scalar energies

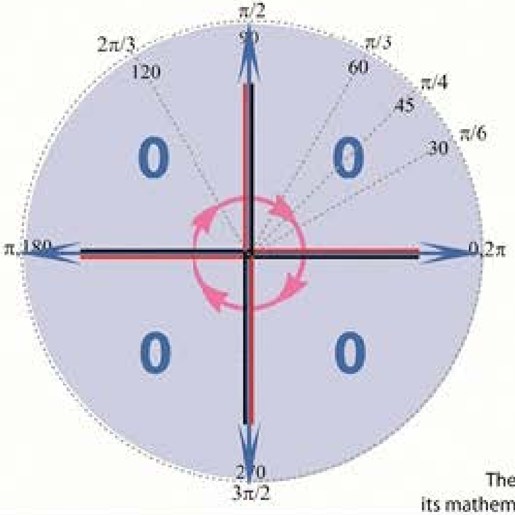

The Lorentz transformation is a linear transformation

It may include a rotation of space;
a rotation-free Lorentz transformation is called a Lorentz boost.

bosons $2\pi[hv] = hf$ photons

π τ

As photons are neutral charge mass-energy geometries care must alway be taken in identifying and specifying the planar wavefront polarities and direction of wave propagation when developing co-ordinate systems to describe their motions and resulting electromagnetic properites

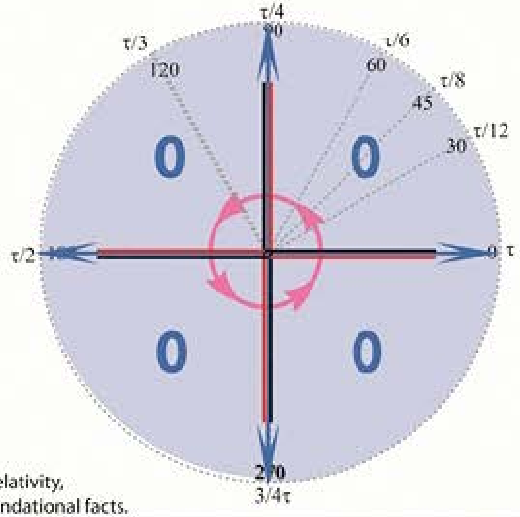

The Lorentz transformation is in accordance with special relativity, but was derived well before special relativity, its mathematical development in relativity theory has resulted in erroneous assumptions being accepted as foundational facts.

Photoelectron KEM fields

As electrons bound to nuclei are accelerated by external forces their energies of motion are stored in their KEM fields

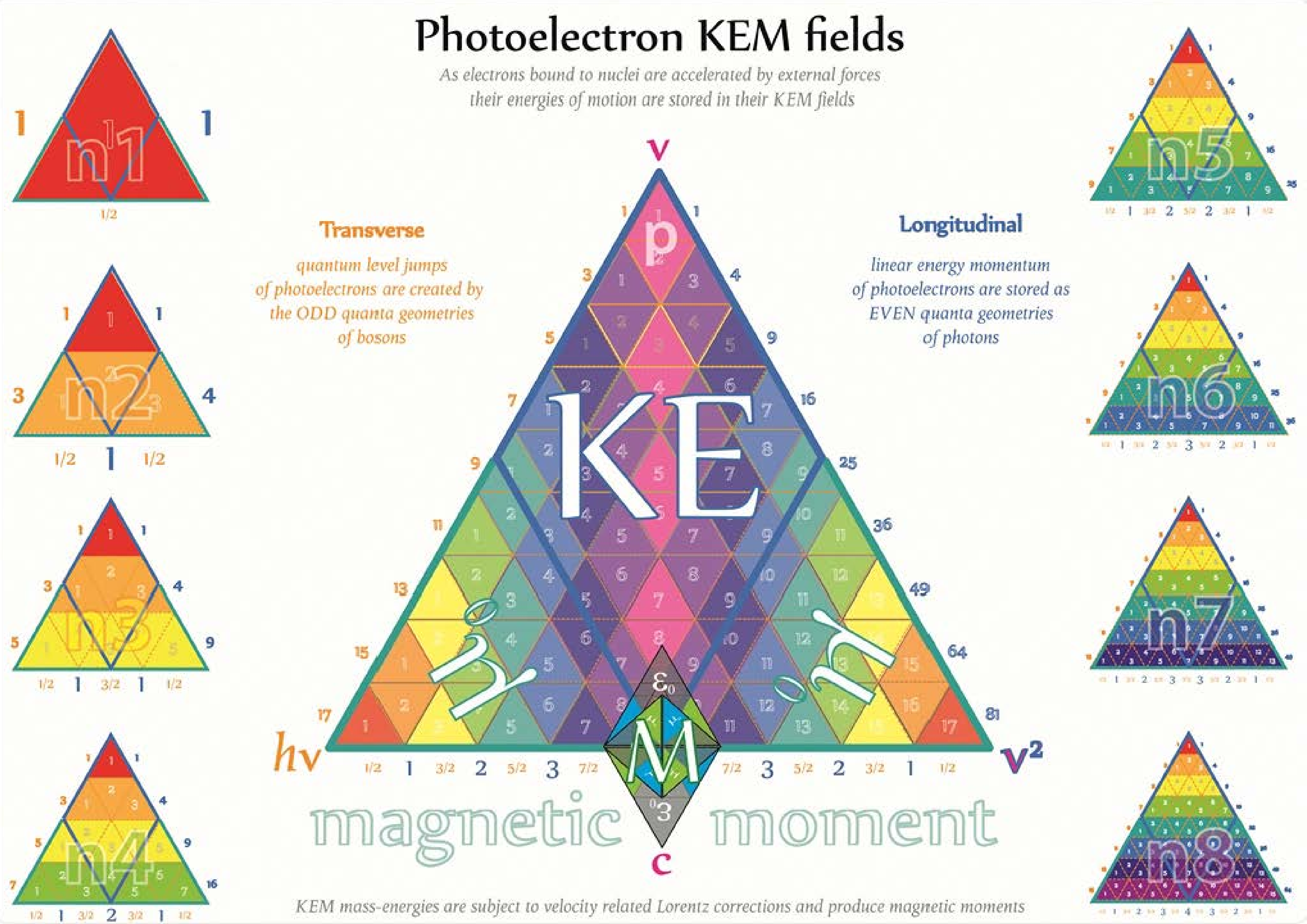
n1
n2
n3
n4
n5
n6
n7
n8
v
p
KE
M
ε_0
μ_0
γ
hv
v^2
c
magnetic moment

Transverse

quantum level jumps of photoelectrons are created by the ODD quanta geometries of bosons

Longitudinal

linear energy momentum of photoelectrons are stored as EVEN quanta geometries of photons

KEM mass-energies are subject to velocity related Lorentz corrections and produce magnetic moments

Tetryonics 28.01 - Photoelectron KEM fields

Kinetics & Photons

The photon has several properties that distinguish it from all other subatomic particles.

It is the only elementary particle wherein a high-energy photon can transform/split into two or more low-energy photons.

A photon is a magnetic dipole. It is an elementary magnet. Evans discovery of the photon's longitudinal magnetic field in 1992 is considered as significant as Einstein's discovery of Relativity.

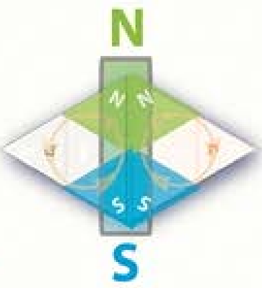

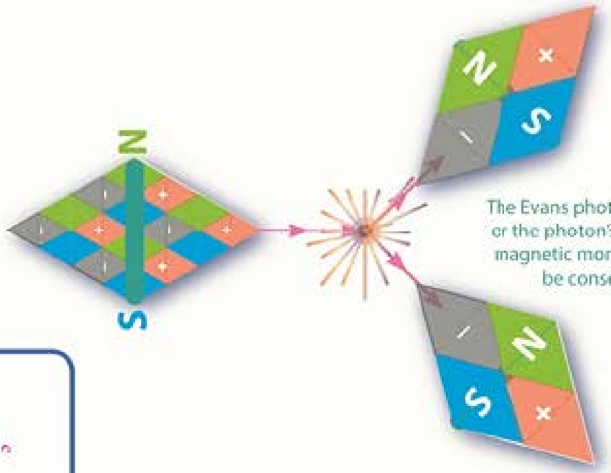

The Evans photomagneton or the photon's transverse magnetic moment has to be conserved.

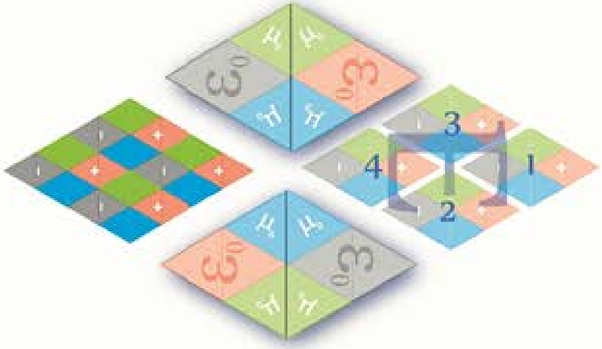

Photons can form super-position EM waves

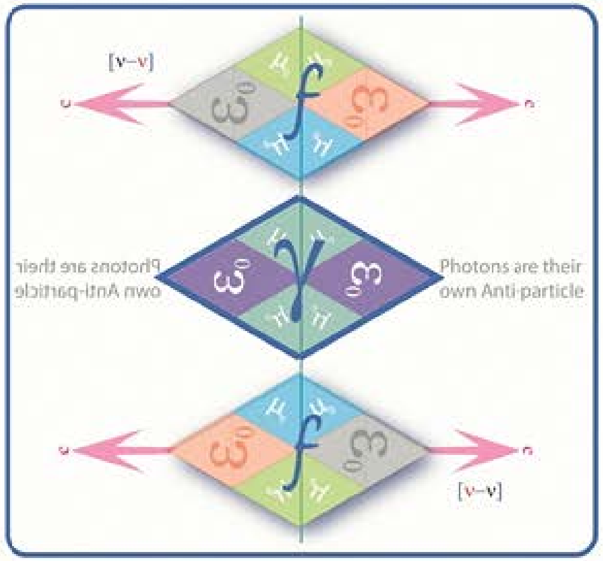

Photons are their own Anti-particle

PAIR PRODUCTION
A photon with sufficient energy can be transformed into a real electron-positron pair.

The EM mass-energy is conserved as per Einstein's equivalence of mass and energy.

The photon's magnetic charge is conserved in the magnetic moments of the created electron-positron pair.

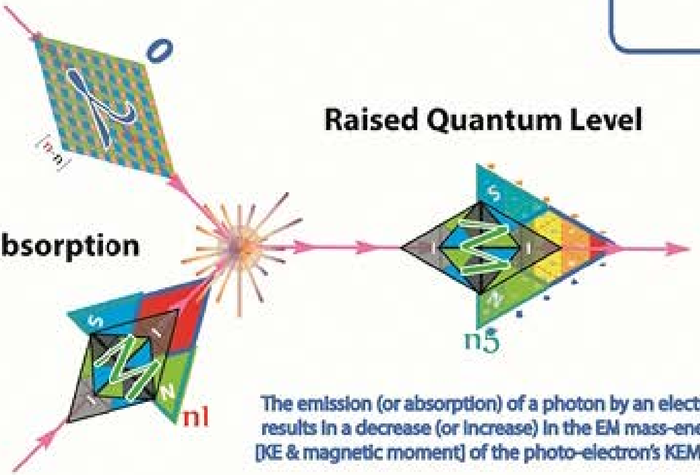

The emission (or absorption) of a photon by an electron, results in a decrease (or increase) in the EM mass-energy [KE & magnetic moment] of the photo-electron's KEM field

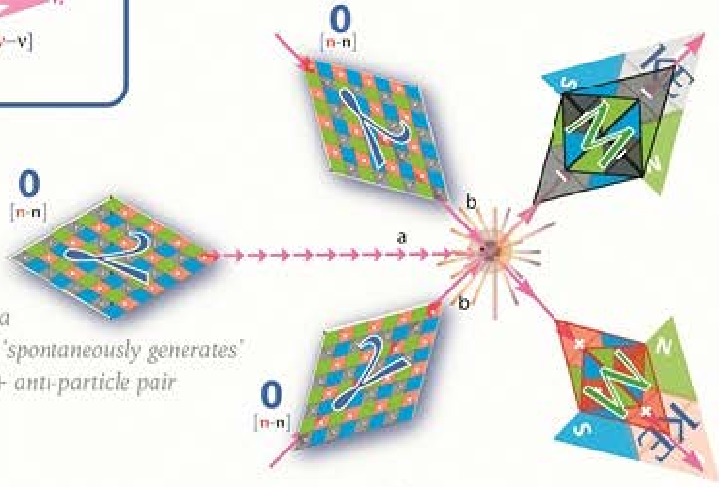

Option a
Photon 'spontaneously generates' a pair + anti-particle pair

Option b
Two Photons combine

Compton Scattering

Arthur Compton

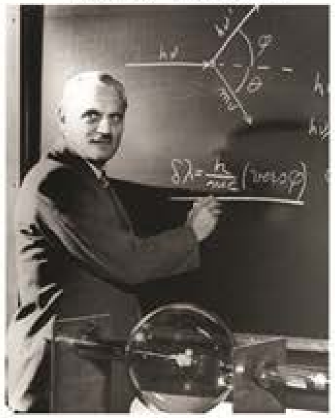

(September 10, 1892 - March 15, 1962)

The in-elastic scattering of photons in matter results in a decrease in energy (increase in wavelength) of an X-ray or gamma ray photon and is called the Compton effect.

Part of the energy of the X/gamma ray is transferred to a scattering electron, which recoils and is ejected from its atom (which becomes ionized), and the rest of the energy is taken by the scattered, "degraded" photon.

Compton frequency of photons

$$E = hf = \frac{hc}{\lambda} = mc^2$$

where v = c

Lorentz corrected energies of EM masses in motion

$$n\pi\frac{[m\Omega v^2]}{c^2}$$

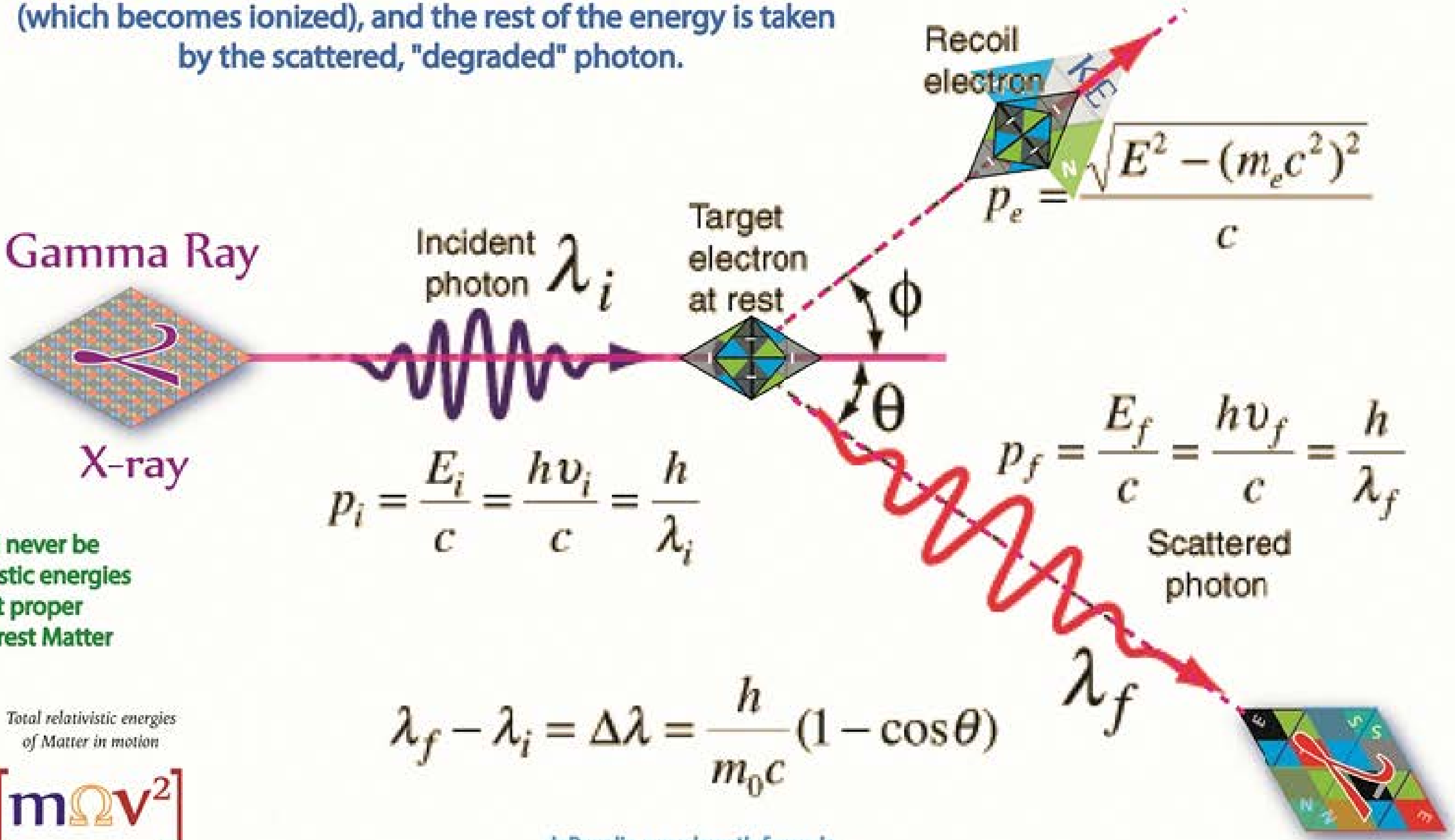
Gamma Ray
X-ray
Incident photon λ_i
Target electron at rest
Recoil electron
KE
N
ϕ
θ
Scattered photon
λ_f
Degraded Photon

$$p_e = \frac{\sqrt{E^2 - (m_e c^2)^2}}{c}$$

$$p_i = \frac{E_i}{c} = \frac{h\upsilon_i}{c} = \frac{h}{\lambda_i}$$

$$p_f = \frac{E_f}{c} = \frac{h\upsilon_f}{c} = \frac{h}{\lambda_f}$$

Compton frequencies should never be calculated from the total relativistic energies of Matter in motion without proper consideration of the invariant rest Matter

Total relativistic energies of Matter in motion

$$n\pi\frac{[m\Omega v^2]}{c^4}$$

$$\lambda_f - \lambda_i = \Delta\lambda = \frac{h}{m_0 c}(1-\cos\theta)$$

deBroglie wavelength formula

Copyright ABRAHAM [2008] - all rights reserved

Pair Production

$$mv^2 = E = \tfrac{1}{2}hf$$

A photon with sufficient energy can be transformed into a real, material electron-positron pair.

The EM mass-energies of the system are conserved as per the equivalence of mass and Energy.

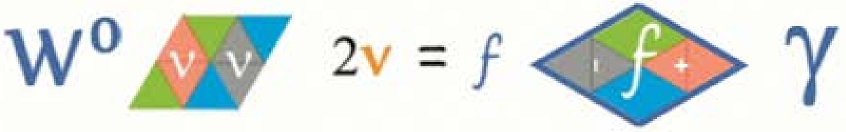

The photon's magnetic dipole and charges are also conserved (but in differing charge geometries)

Option a

Photon produces 2 particles

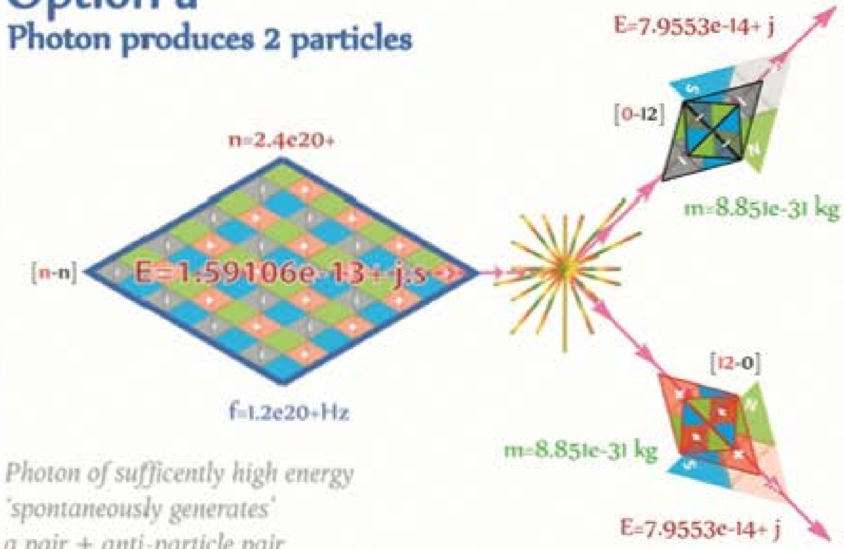
E=7.9553e-14+ j
[0-12]
n=2.4e20+
m=8.851e-31 kg
[n-n]
E=1.59106e-13+ j.s
f=1.2e20+Hz
[12-0]
m=8.851e-31 kg
E=7.9553e-14+ j

Photon of sufficently high energy 'spontaneously generates' a pair + anti-particle pair

Option b

Two Photons combine

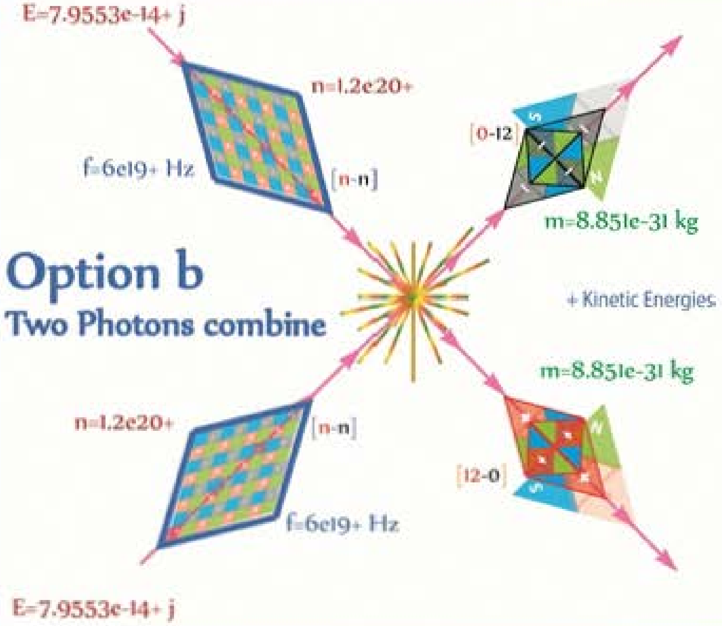
E=7.9553e-14+ j
n=1.2e20+
f=6e19+ Hz
[n-n]
[0-12]
m=8.851e-31 kg
+ Kinetic Energies
m=8.851e-31 kg
n=1.2e20+
[n-n]
[12-0]
f=6e19+ Hz
E=7.9553e-14+ j

Planck's constant	Energy	electron rest mass-Matter
6.62432672 e-34 kgm^2/s	7.9553 e-14 j	8.8514860 e-31 kg

$$\boxed{n\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\underset{\text{velocity}}{\overset{\text{quanta}}{\Omega v^2}}\right] = E = T\pi\left[\underset{\text{mass}}{\overset{\text{Planck}}{m}}\underset{\text{velocity}}{\overset{\text{quanta}}{\Omega v^2}}\right]}$$

6 e19
photon frequency

1.2 e20
Compton frequency

All pair production - annihilations must follow the mass-Energy-Matter equivalency formula

mass $c^2 \xleftarrow{\text{per second}} E \xrightarrow{\text{per second squared}} c^4$ Matter

Kinetic Energies

$$KE = \frac{1}{2}Mv^2$$

As velocity increases so does the number of energy momenta quanta per spatial geometry

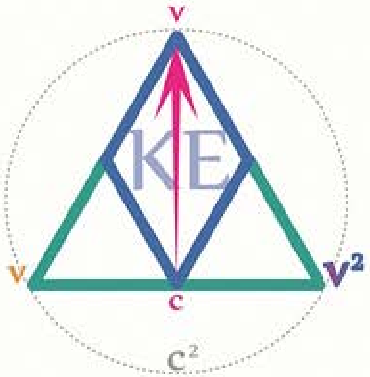

Photons contain Kinetic energies and momentum

Photonic Energy

$$E_\gamma = 2\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]$$

Kinetic Energy and momenta are all related through Tetryonic geometry

Photons

$$p = \frac{h}{\lambda} = \frac{E}{c}$$

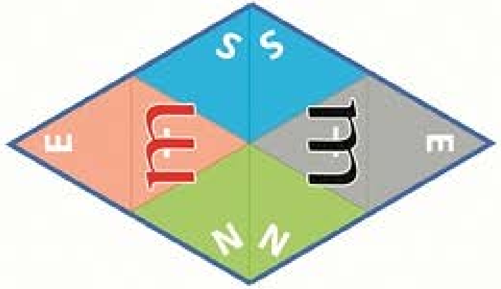

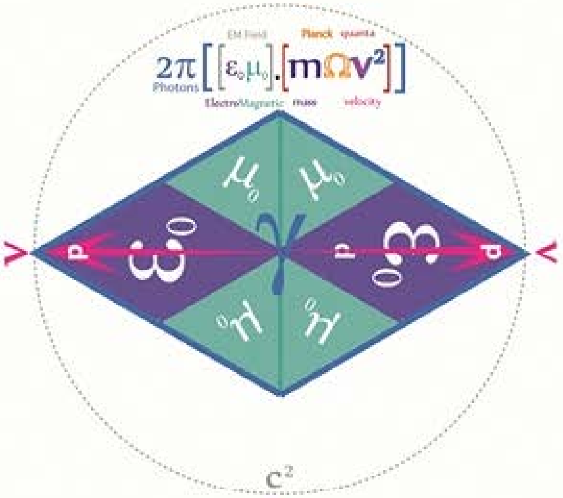
EM Field
Planck quanta
2π[[ε₀μ₀].[mΩv²]]
Photons
ElectroMagnetic
mass
velocity
μ₀
μ₀
ε₀
ε₀
γ
p
p
v
v
c²

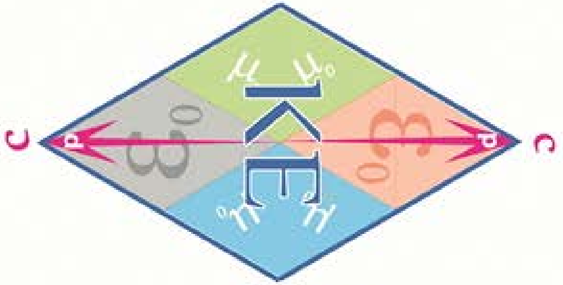
c
c
p
p
ε₀
ε₀
μ₀
μ₀
KE

$$E = hf$$

Momentum

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

As the number of Photons increases so does the Momentum

$$p = \frac{n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v^2}\right]\right]}{v} = E/v$$

All Photons travel at the 'speed of light' in their medium of transmission

$$p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$$

$$p = n\pi\left[\left[\underset{\text{mass}}{m}\overset{\text{Planck quanta}}{\Omega}\underset{\text{velocity}}{v}\right]\right]$$

linear energy-momentum

Copyright ABRAHAM [2008] - all rights reserved

Electronic Kinetic Energies

have discrete energy levels within the nucleus and
continuous energy levels in unbound electrons

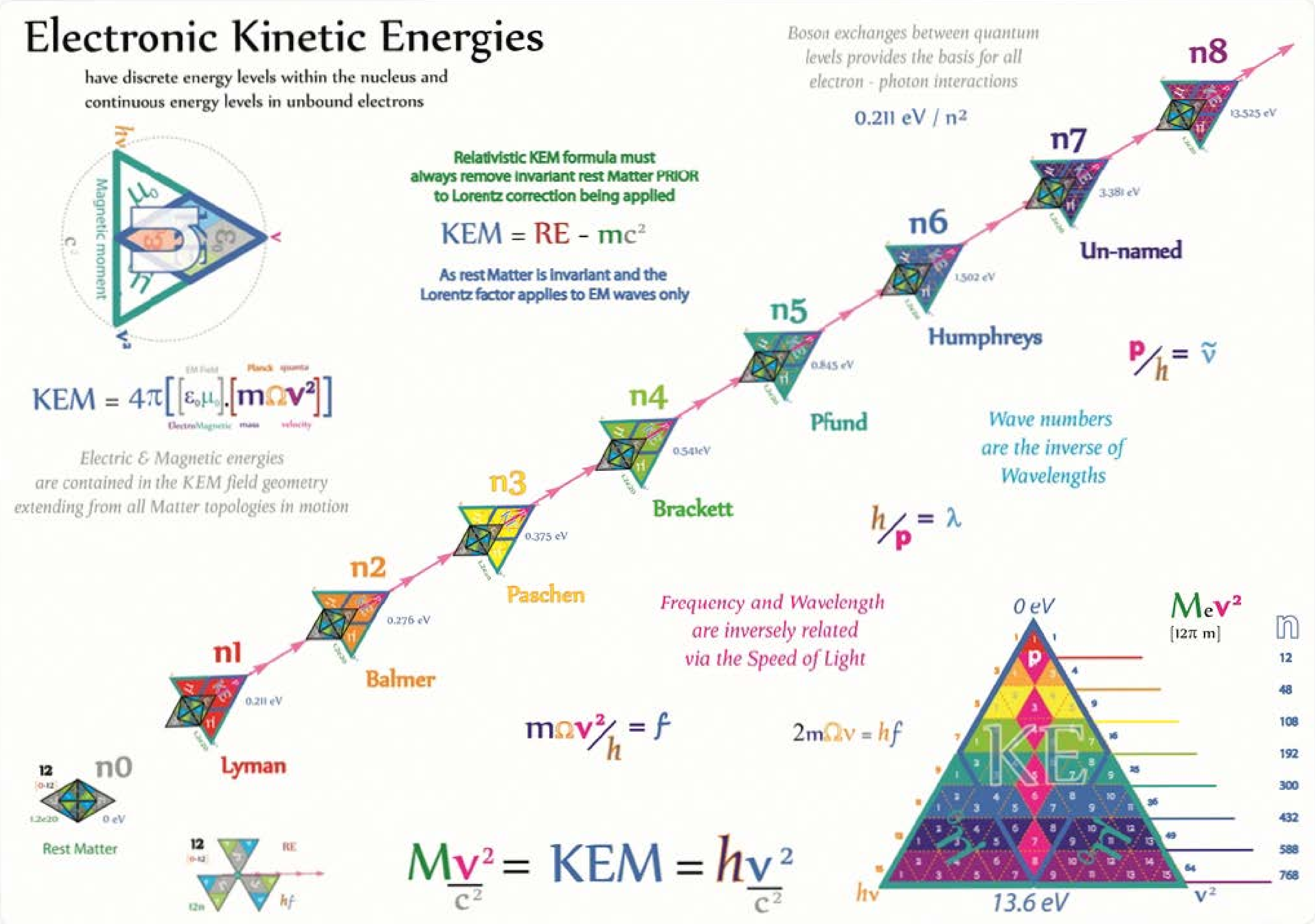
hv
Magnetic moment
μ_0
ε_0
c^2
v^2
n0
12
[0-12]
1.2e20
0 eV
Rest Matter
12
[0-12]
12n
RE
hf
n1
0.211 eV
Lyman
n2
0.276 eV
Balmer
n3
0.375 eV
Paschen
n4
0.541eV
Brackett
n5
0.845 eV
Pfund
n6
1.502 eV
Humphreys
n7
3.381 eV
Un-named
n8
13.525 eV
0 eV
13.6 eV
hv
v^2
KE
p
n
12
48
108
192
300
432
588
768

*Boson exchanges between quantum
levels provides the basis for all
electron - photon interactions*

0.211 eV / n^2

**Relativistic KEM formula must
always remove invariant rest Matter PRIOR
to Lorentz correction being applied**

$$KEM = RE - mc^2$$

**As rest Matter is invariant and the
Lorentz factor applies to EM waves only**

$$KEM = 4\pi\left[\left[\underbrace{\overset{\text{EM Field}}{\varepsilon_0\mu_0}}_{\text{ElectroMagnetic}}\right].\left[\overset{\text{Planck quanta}}{\underset{\text{mass}}{m}\Omega\underset{\text{velocity}}{v^2}}\right]\right]$$

*Electric & Magnetic energies
are contained in the KEM field geometry
extending from all Matter topologies in motion*

$$p/h = \tilde{\nu}$$

*Wave numbers
are the inverse of
Wavelengths*

$$h/p = \lambda$$

*Frequency and Wavelength
are inversely related
via the Speed of Light*

$$m\Omega v^2/h = f$$

$$2m\Omega v = hf$$

$$M_e v^2$$
$[12\pi\ m]$

$$M\frac{v^2}{c^2} = KEM = h\frac{v^2}{c^2}$$

Copyright ABRAHAM [2008] - all rights reserved

Photonic energy
momenta of KEM fields

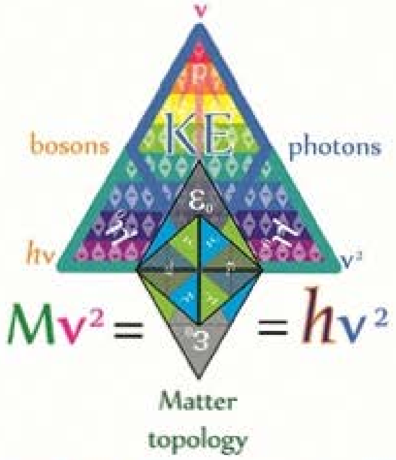
v
KE
bosons
photons
hv
v²
ε_0
3
Matter
topology

$$Mv^2 = \quad = hv^2$$

Rydberg's formula is a wavenumber equivalent re-expression of Lorentz's velocity correction formula as it applies to the quantised energy momentum of Photo-electrons bound in atomic nuclei

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

The difference of two inverse squared KEM energies

Rydberg		Tetryonic
$R\left(\frac{1}{1} - \frac{1}{4}\right)$	=	$R(36 \diagup 48)$
.75		.75

n1 and n2 are integers greater than or equal to 1 such that n1 < n2

$$\frac{1}{2}Mv^2 = KE = hcR$$

Whereas Rydberg calculated for wavenumbers using the difference of two inverse squared KEM energies. (inverse wavelengths)

$$\left(\frac{1}{1} - \frac{1}{4}\right)$$

.75

A mathematical solution that works but whose physical mechanics has eluded explanation

Tetryonics obtains the same answers geometrically using the difference of the two KEM energies. (quanta differentials)

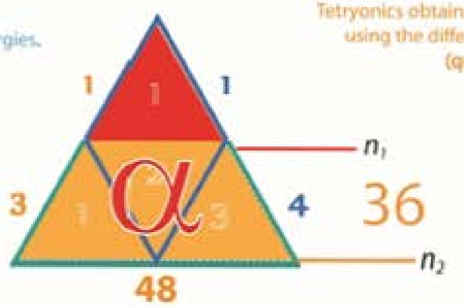
1
1
1
α
2
1
3
3
4
n_1
36
n_2
48

(36/48)

.75

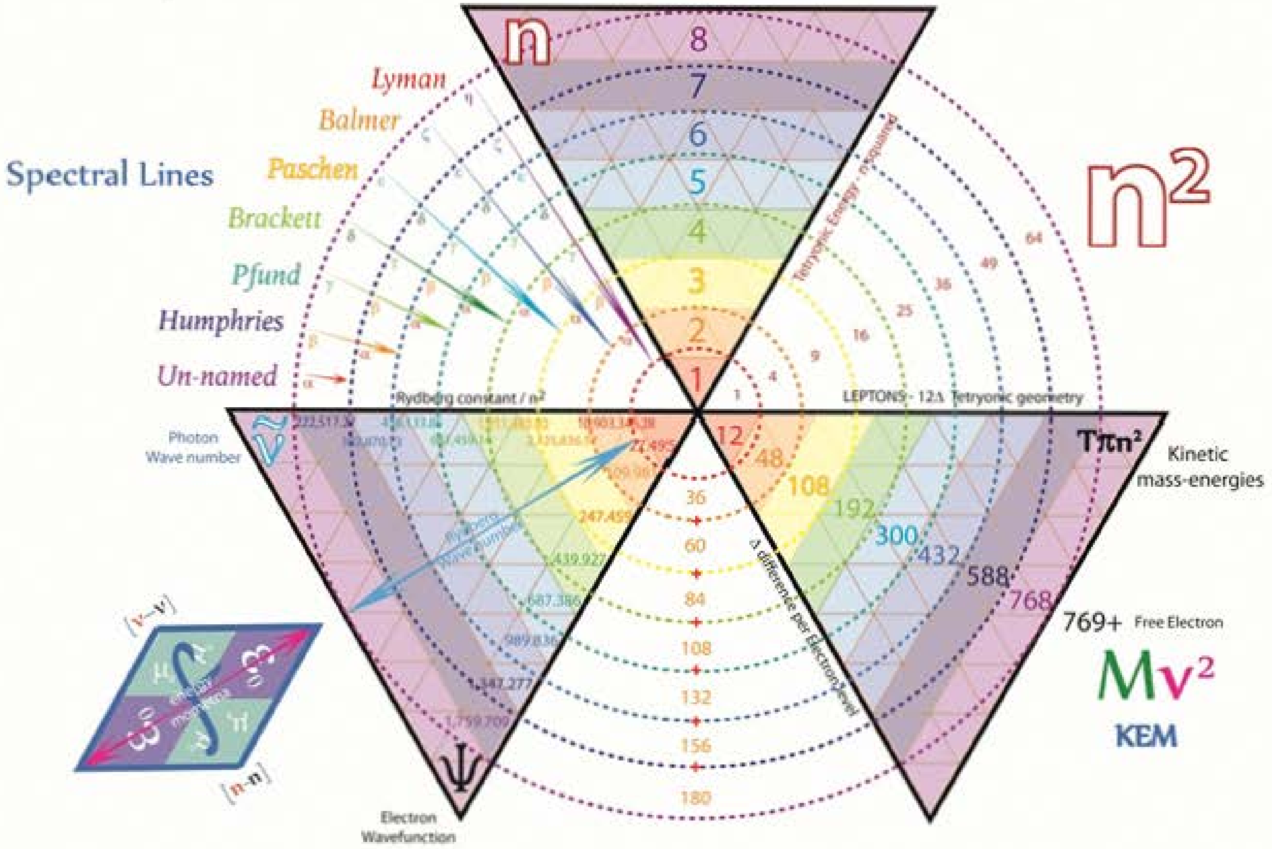
n
8
7
6
5
4
3
2
1
n²
Spectral Lines
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
Tetryonic Energy - n squared
1
4
9
16
25
36
49
64
Rydberg constant / n²
LEPTONS - 12Δ Tetryonic geometry
Photon Wave number
ν̃
222,517.9
247,459
439,927
687,386
989,836
1,347,277
1,759,709
Rydberg Wave number
Δ difference per Electron level
12
48
108
192
300
432
588
768
36
60
84
108
132
156
180
Tπn²
Kinetic mass-energies
769+ Free Electron
Mv²
KEM
Ψ
Electron Wavefunction
[v-v]
[n-n]
μ
ε
energy momenta

Any changes to the Energy momenta of the electron's KEM field must be reflected in the Energy momenta of the Photo-electrons emitted [absorbed]

$$\pm\Delta hv = KEM_i - KEM_f$$

The difference of two KEM energies

Tetryonics 28.07 - Photonic energy momenta of KEM fields

Copyright ABRAHAM [2008] - all rights reserved

Kinetic Energies of Photoelectrons

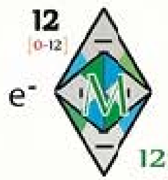
12
[0-12]
e⁻
M
12n

rest Matter

Kinetic energies are DIVERGENT from Matter topologies

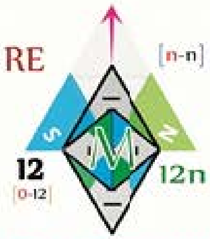
RE
[n-n]
12
[0-12]
12n

Relativistic mass-energies

rest Matter + Kinetic Energies

The total energy of a photon is absorbed into the electron's KEM field geometry

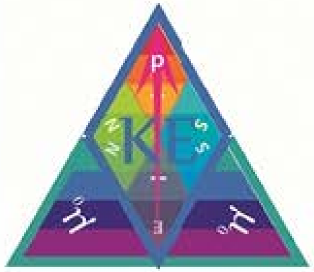

All velocity changes produce changes to mass-energies, momenta and wavefunction relationships

$$KEM = Mv^2$$

rest Matter is velocity invariant

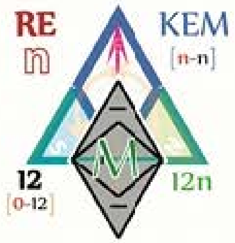
RE
KEM
[n-n]
M
12
[0-12]
12n

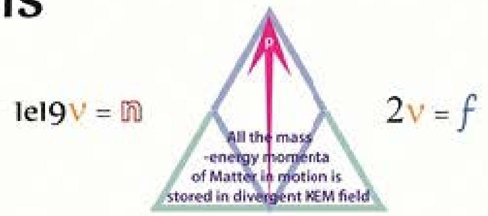
lel9v = n
2v = f
All the mass-energy momenta of Matter in motion is stored in divergent KEM field

6.62943 e -34 J.s

13.5252 eV
2.1669 e -18 J

$$E = hv^2$$

3.26874 e15 quanta

KEM fields are velocity variant

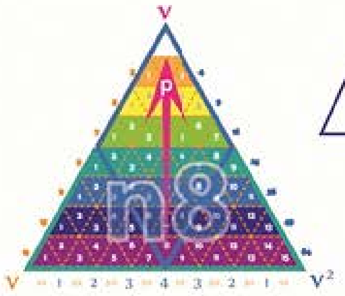
v
p
n8
v
1 2 3 4 3 2 1
v²

Δp

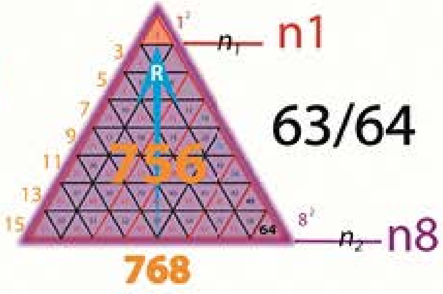
1²
n₁
n1
3
5
7
9
11
13
15
R
756
64
8²
n₂
n8
768

63/64

EM Field · Planck quanta

$$4\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

KEM field · ElectroMagnetic · mass · velocity

Electron KEM Quantum levels

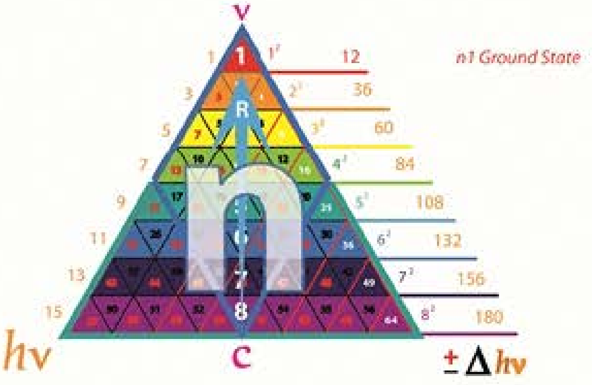
v
n1 Ground State
1²
2²
3²
4²
5²
6²
7²
8²
12
36
60
84
108
132
156
180
1
3
5
7
9
11
13
15
R
n
hv
c
±Δhv

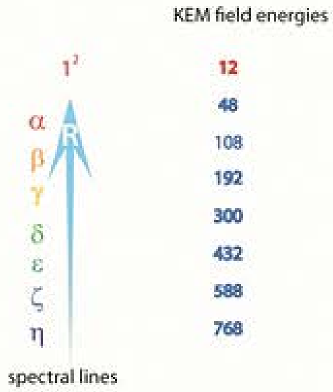
KEM field energies
1²
α
β
γ
δ
ε
ζ
η
12
48
108
192
300
432
588
768
spectral lines

Rydberg spectral transition jumps

Rydberg · Tetryonic

$$\underset{.984375}{R_H\left(\frac{1}{1}-\frac{1}{64}\right)} = \frac{1}{\lambda} = \underset{.984375}{KEM\left(\frac{63/64}{1}\right)}$$

Copyright ABRAHAM [2008] - all rights reserved

Spectral line energies

EM Field Planck quanta

$1\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Charge ElectroMagnetic mass velocity

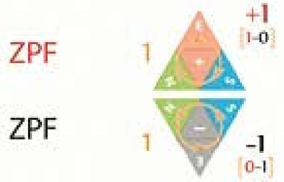
ZPF
ZPF
1
+1
[1-0]
1
-1
[0-1]

Bosons are

Odd number Charge carriers

n.hv

EM Field Planck quanta

$\mathrm{ODD}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Bosons ElectroMagnetic mass velocity

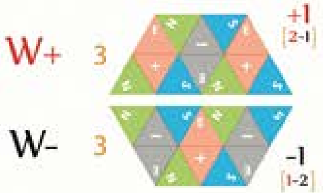
W+
3
+1
[2-1]
W-
3
-1
[1-2]

Each quantum level transition requires kinetic energies with an ODD number of EM mass-energy momenta quanta [bosons - charge carriers]

In order to transistion between levels for each emission and absorption leptons must release or absorb specific W bosons (neutral Z bosons - photons) for each KEM field change

Pairs of charged W bosons combine to form neutral bosons / photons [EM induction]

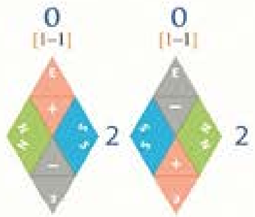

γ

EM Field Planck quanta

$2\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

Photons ElectroMagnetic mass velocity

Photons are

Even number EM force carriers

n.hf

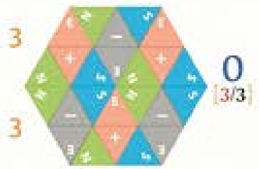

W^0

EM Field Planck quanta

$\mathrm{EVEN}\pi[[\varepsilon_0\mu_0].[m\Omega v^2]]$

EM waves ElectroMagnetic mass velocity

KEM field Differentials

Bosonic quantum level changes

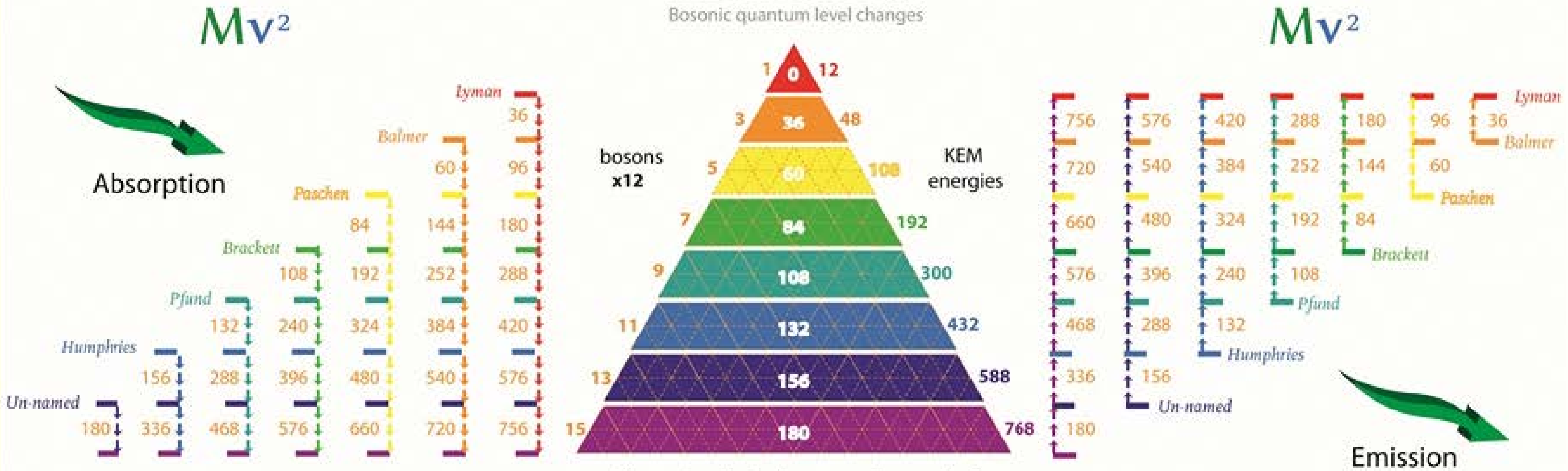
Mv²
Absorption
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
36
96
180
288
420
576
756
60
144
252
384
540
720
84
192
324
480
660
108
240
396
576
132
288
468
156
336
180
bosons x12
1
3
5
7
9
11
13
15
0
36
60
84
108
132
156
180
12
48
108
192
300
432
588
768
KEM energies
Mv²
756
720
660
576
468
336
180
576
540
480
396
288
156
420
384
324
240
132
288
252
192
108
180
144
84
96
60
36
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
Emission

All quantum level changes are the result of boson increases or decreases

Copyright ABRAHAM [2008] - all rights reserved

Quantum level differentials

The exact number of Kinetic energy quanta (ZPFs)
required for each bound electron transition within a nucleus

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg calculated this as the wavenumber differerential of two inverse squared energies

n_f^2	Rydberg		Tetryonic	n_i^2
	$R_H\left(\frac{1}{1} - \frac{1}{4}\right)$	=	$R_H(36 / 48)$	
	.75		.75	

Tetryonic geometry reveals it to be sum of ODD integer bosons between two squared KEM energies

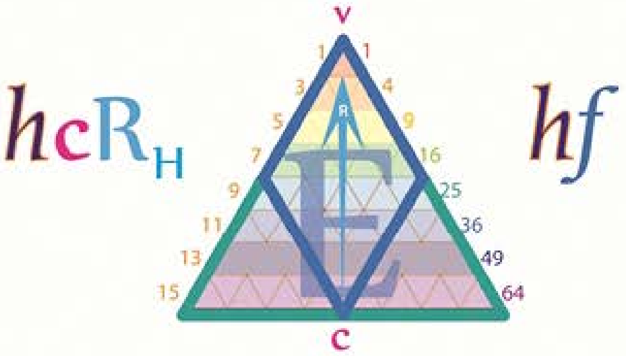

hcR_H
hf
v
c
E
R
1
3
5
7
9
11
13
15
1
4
9
16
25
36
49
64

Rydberg's constant is a LINEAR measurement of wavenumbers
which can be related to squared KEM energies through
the velocity-frequency-wavelength relationship

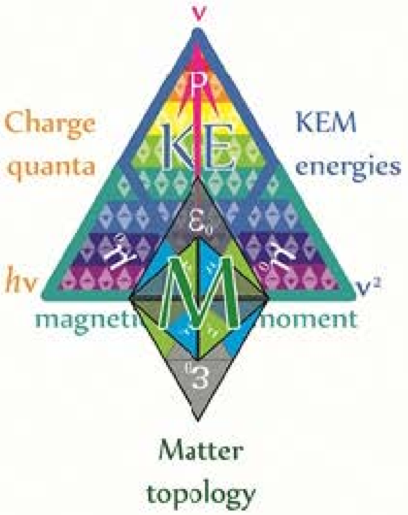

v
P
KE
Charge quanta
KEM energies
ε₀
μ₀
M
hv
v²
magnetic moment
3
Matter topology

Each Quantum level jump within
an Atomic Nuclei is discrete
and requires an exact number
of EM mass-energy quanta

The calculation of the quanta
required is historically done
utilising Rydberg's Constant and
its associated formula

The quantum level [boson] differential
equates it to the release or absorption of
consecutive ODD number Planck quanta
from squared KEM field energy geometries
Δ[hv]

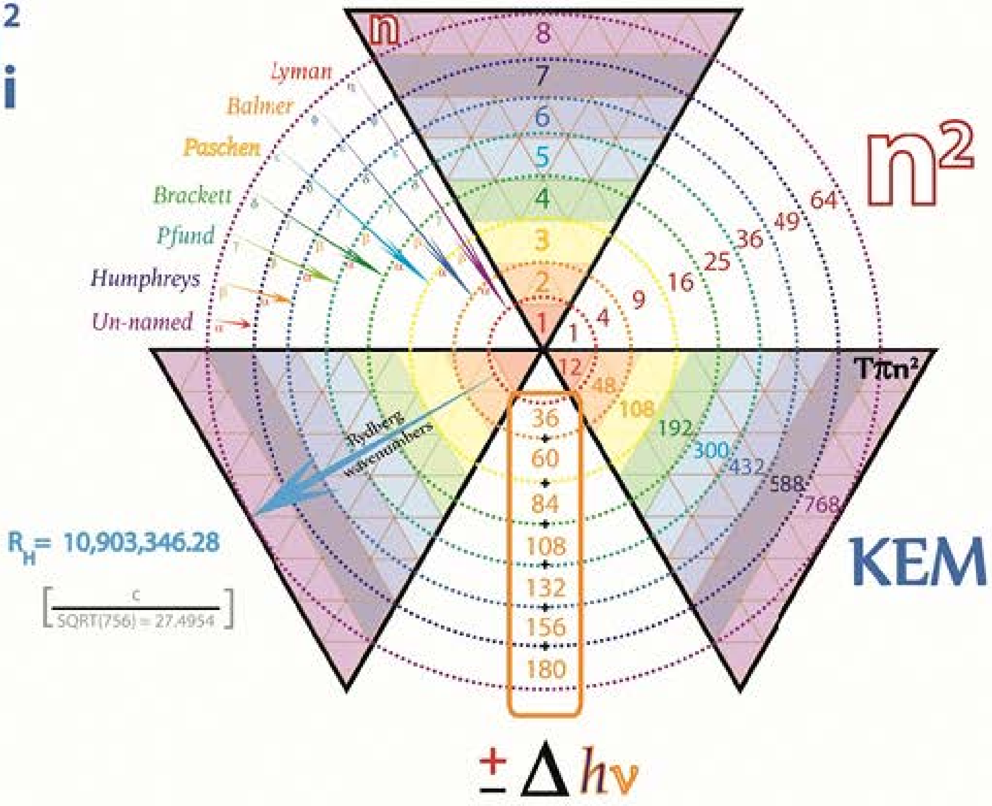

n
8
7
6
5
4
3
2
1
Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
Un-named
n²
1
4
9
16
25
36
49
64
Tπn²
12
48
108
192
300
432
588
768
36
60
84
108
132
156
180
Rydberg wavenumbers
KEM
± Δ hv

R_H = 10,903,346.28

$\left[\frac{c}{\mathrm{SQRT}(756) = 27.4954}\right]$

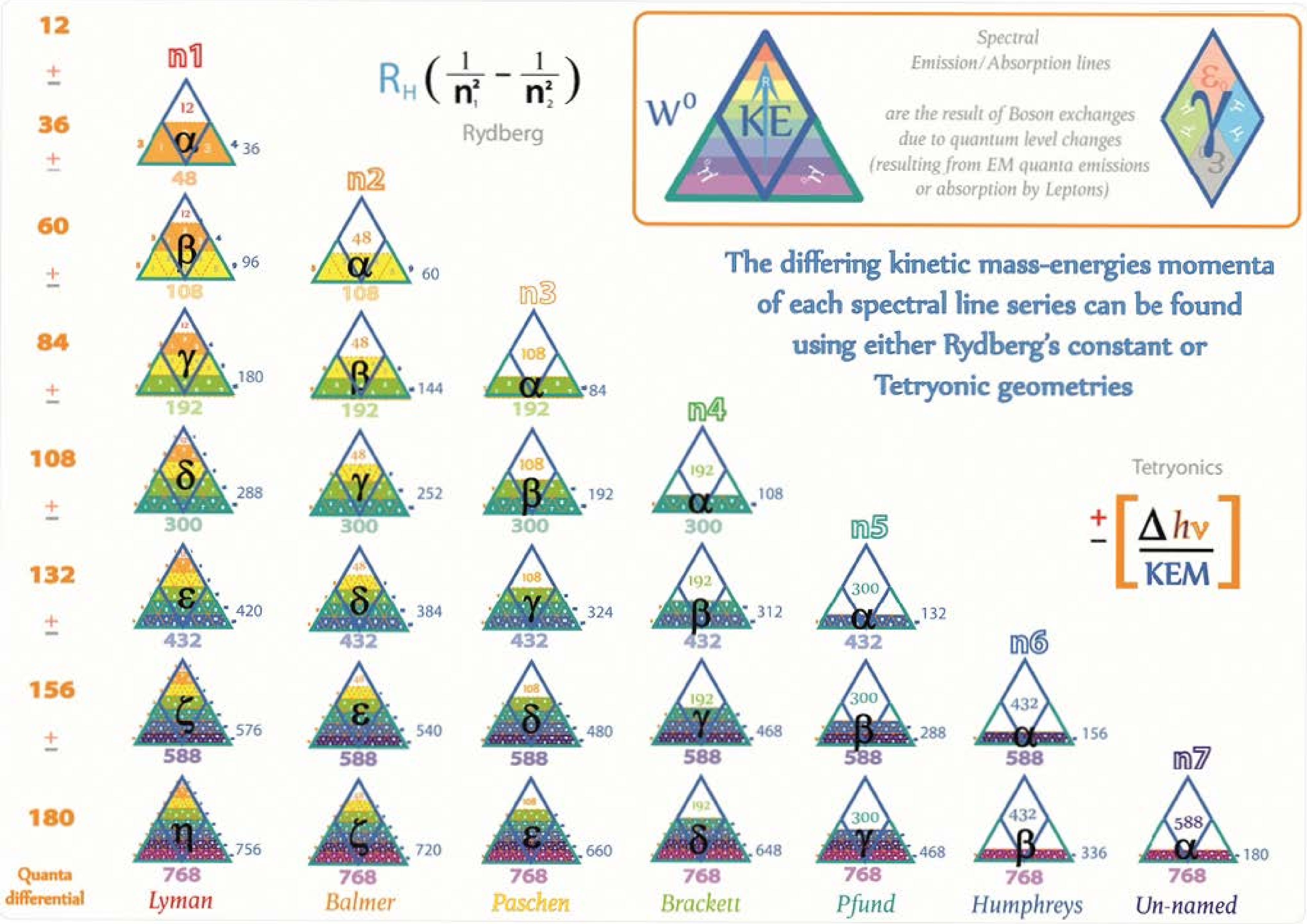
$R_H\left(\frac{1}{n_1^2}-\frac{1}{n_2^2}\right)$
Rydberg
W⁰
KE
Spectral
Emission/Absorption lines
are the result of Boson exchanges
due to quantum level changes
(resulting from EM quanta emissions
or absorption by Leptons)
The differing kinetic mass-energies momenta
of each spectral line series can be found
using either Rydberg's constant or
Tetryonic geometries
Tetryonics
$\pm\left[\frac{\Delta hv}{KEM}\right]$
12 ± 36 ± 60 ± 84 ± 108 ± 132 ± 156 ± 180
Quanta differential
n1 n2 n3 n4 n5 n6 n7
α 36 48; β 96 108; γ 180 192; δ 288 300; ε 420 432; ζ 576 588; η 756 768
α 60 108; β 144 192; γ 252 300; δ 384 432; ε 540 588; ζ 720 768
α 84 192; β 192 300; γ 324 432; δ 480 588; ε 660 768
α 108 300; β 312 432; γ 468 588; δ 648 768
α 132 432; β 288 588; γ 468 768
α 156 588; β 336 768
α 180 768
Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
Un-named

Tetryonics 28.11 - Boson differentials

Absorption~Emission lines

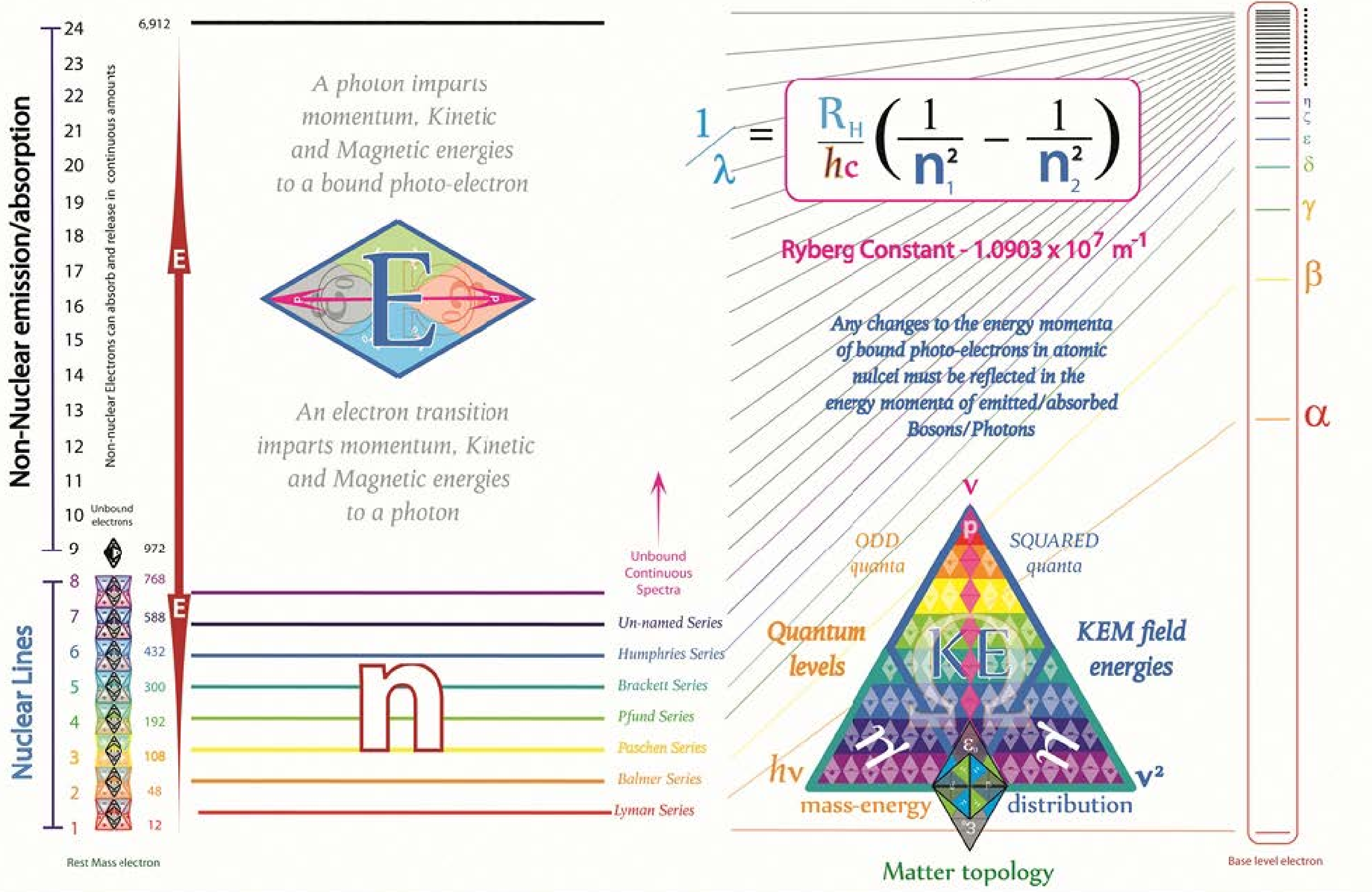

$$E = hf = \frac{hc}{\lambda} = mc^2$$

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg Constant - $1.0903 \times 10^7\ m^{-1}$

A photon imparts momentum, Kinetic and Magnetic energies to a bound photo-electron

An electron transition imparts momentum, Kinetic and Magnetic energies to a photon

Any changes to the energy momenta of bound photo-electrons in atomic nulcei must be reflected in the energy momenta of emitted/absorbed Bosons/Photons

Non-Nuclear emission/absorption
Non-nuclear Electrons can absorb and release in continuous amounts
Nuclear Lines
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
Unbound electrons
9
8
7
6
5
4
3
2
1
6,912
972
768
588
432
300
192
108
48
12
Rest Mass electron
E
E
n
Unbound Continuous Spectra
Un-named Series
Humphries Series
Brackett Series
Pfund Series
Paschen Series
Balmer Series
Lyman Series
Spectral line series
η
ζ
ε
δ
γ
β
α
Base level electron
v
p
ODD quanta
SQUARED quanta
Quantum levels
KEM field energies
KE
hv
v²
mass-energy
distribution
Matter topology

Tetryonics 29.01 - Absorption~Emission Lines

Copyright ABRAHAM [2008] - all rights reserved

Spectral Lines

Spectral lines are the result of interaction between a quantum system (usually atoms, but sometimes molecules or atomic nuclei) and a single photon. When a photon has exactly the right amount of energy to allow a change in the energy state of the system (in the case of an atom this is usually an electron changing orbitals), the photon is absorbed. Then it will be spontaneously re-emitted, either in the same frequency as the original or in a cascade, where the sum of the energies of the photons emitted will be equal to the energy of the one absorbed (assuming the system returns to its original state). The direction and polarization of the new photons will, in general, correlate with those of the original photon.

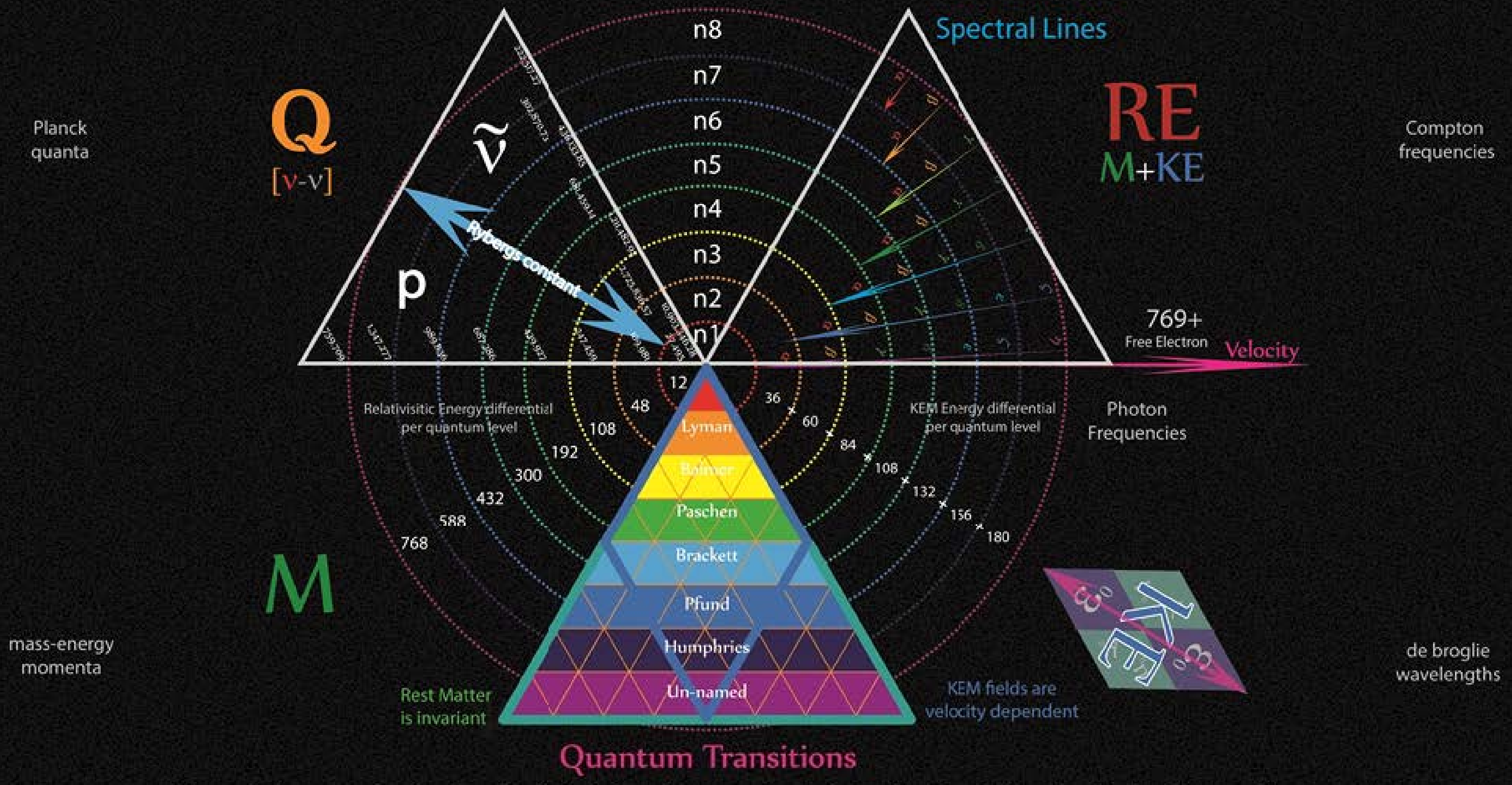
Planck
quanta
Q
[ν-ν]
ν̃
p
Rydbergs constant
n8
n7
n6
n5
n4
n3
n2
n1
Spectral Lines
RE
M+KE
Compton
frequencies
769+
Free Electron
Velocity
12
48
108
192
300
432
588
768
36
60
84
108
132
156
180
Relativisitic Energy differential
per quantum level
KEM Energy differential
per quantum level
Photon
Frequencies
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
M
mass-energy
momenta
Rest Matter
is invariant
KEM fields are
velocity dependent
KE
de broglie
wavelengths
Quantum Transitions

Spectral lines are highly atom-specific, and can be used to identify the chemical composition of any medium capable of letting light pass through it (typically gas is used). Several elements were discovered by spectroscopic means, such as Helium [04], Rubidium [37], Cerium [58] and Thallium [81]

Spectral lines also depend on the physical conditions of the gas, so they are widely used to determine the chemical composition of stars and other celestial bodies that cannot be analyzed by other means, as well as many other of their physical properties.

Tetryonics 29.02 - Spectral line production

Copyright ABRAHAM [2008] - all rights reserved

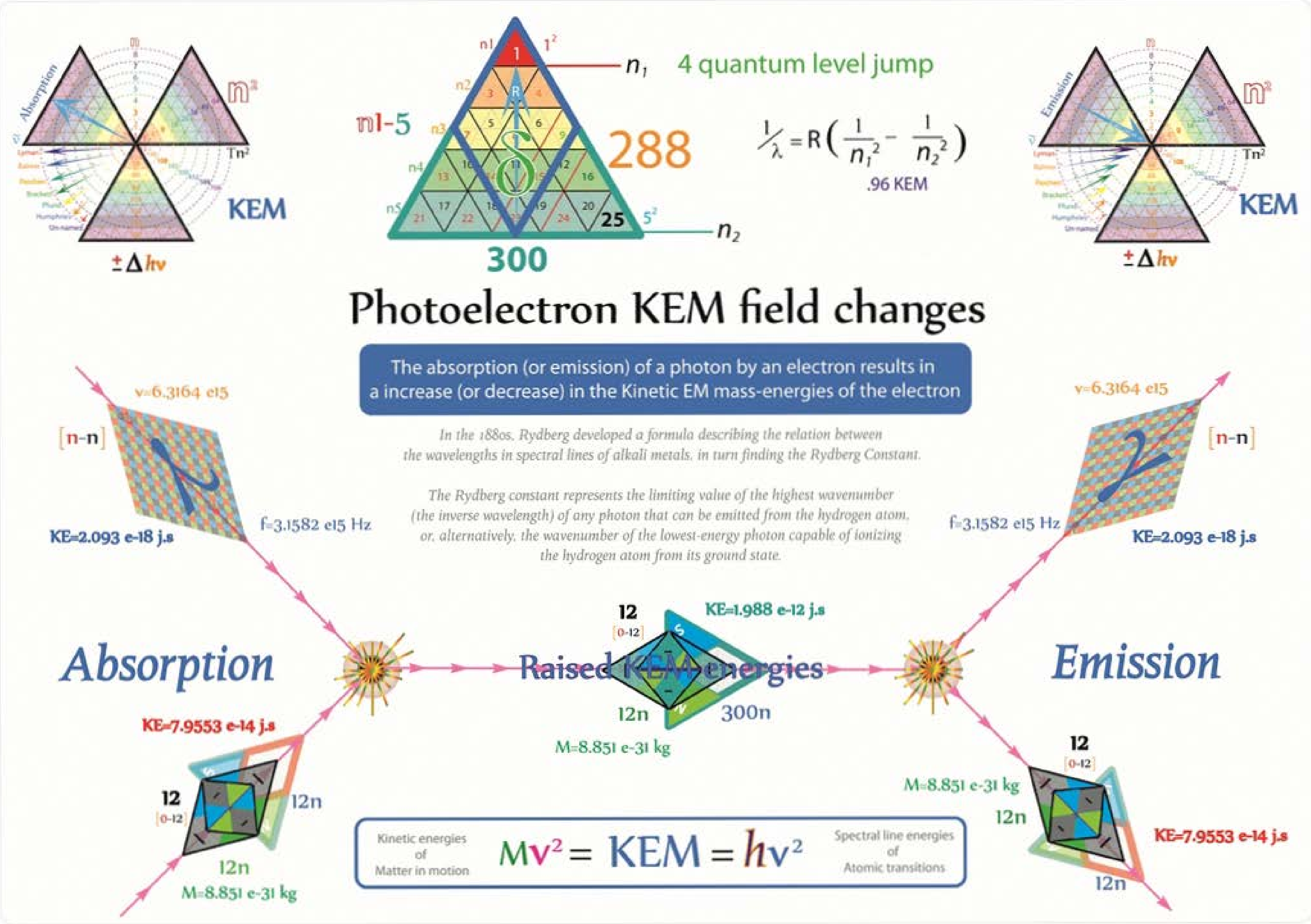
Absorption
Emission
n²
Tn²
KEM
±Δhv
n1
n2
n3
n4
n5
1²
5²
1
25
n1-5
288
300
n_1
n_2
4 quantum level jump
.96 KEM
v=6.3164 e15
[n-n]
f=3.1582 e15 Hz
KE=2.093 e-18 j.s
KE=7.9553 e-14 j.s
12
[0-12]
12n
M=8.851 e-31 kg
KE=1.988 e-12 j.s
Raised KEM energies
300n

$$\frac{1}{\lambda} = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Photoelectron KEM field changes

The absorption (or emission) of a photon by an electron results in a increase (or decrease) in the Kinetic EM mass-energies of the electron

In the 1880s, Rydberg developed a formula describing the relation between the wavelengths in spectral lines of alkali metals, in turn finding the Rydberg Constant.

The Rydberg constant represents the limiting value of the highest wavenumber (the inverse wavelength) of any photon that can be emitted from the hydrogen atom, or, alternatively, the wavenumber of the lowest-energy photon capable of ionizing the hydrogen atom from its ground state.

Kinetic energies of Matter in motion

$$Mv^2 = KEM = hv^2$$

Spectral line energies of Atomic transitions

Tetryonics 29.03 - Photolectron KEM changes

Copyright ABRAHAM [2008] - all rights reserved

Revealing Rydberg Formula's geometry

The Rydberg formula is used in atomic physics to determine the wavenumber of spectral lines of many chemical elements.

$$\frac{h}{Mv} = \lambda$$

Wavelength is inversely related to Frequency

$$1/\lambda = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

$$p/_h = \tilde{\nu}$$

Wavenumbers are a inverse measure of Wavelengths

$$p^2 = E = Mv^2$$

$$p = \frac{E}{v} = hv$$

Tetryonic geometry can be applied to Kinetic EM field variations produced by the emission and absorption of Photons by electrons in Nuclear orbits to reveal the geometry behind Rydberg's formula

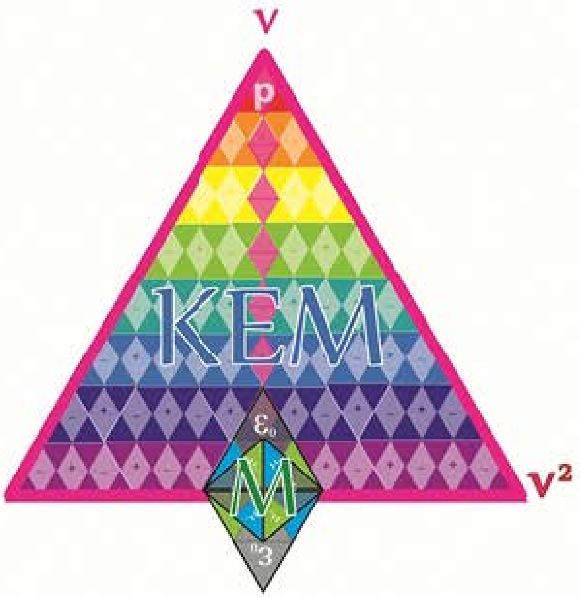
v
p
KEM
v²
ε₀
M
³

EM Field　Planck quanta

$$12\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photoelectron KEM fields　ElectroMagnetic　mass　velocity

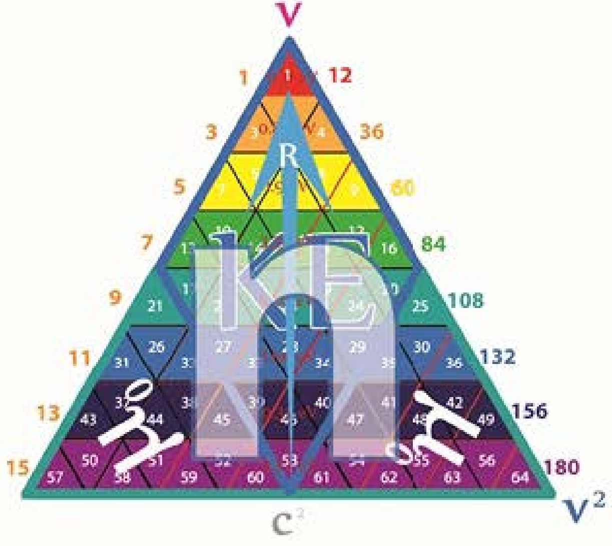
v
1
3
5
7
9
11
13
15
12
36
60
84
108
132
156
180
R
KE
h
1
3
4
7
9
16
21
24
25
26
27
28
29
30
31
34
36
39
40
42
43
44
45
47
49
50
51
53
55
56
57
58
59
60
61
62
63
64
c²
v²

13.525 eV

$$\underset{\text{linear momentum}}{Mv} = \frac{KEM}{v} = \underset{\text{longitudinal quanta}}{hv}$$

Wavelengths and Frequency are related through the velocity of propagation

Compton frequency

$$\frac{Mv^2}{h} = f$$

Spectral line emissions reveals Rydberg's formula to be a measure of longitudinal KEM momenta

$$KEM = hcR$$

$$\lambda_{KEM} = \frac{f}{v}$$

Hydrogen spectral lines

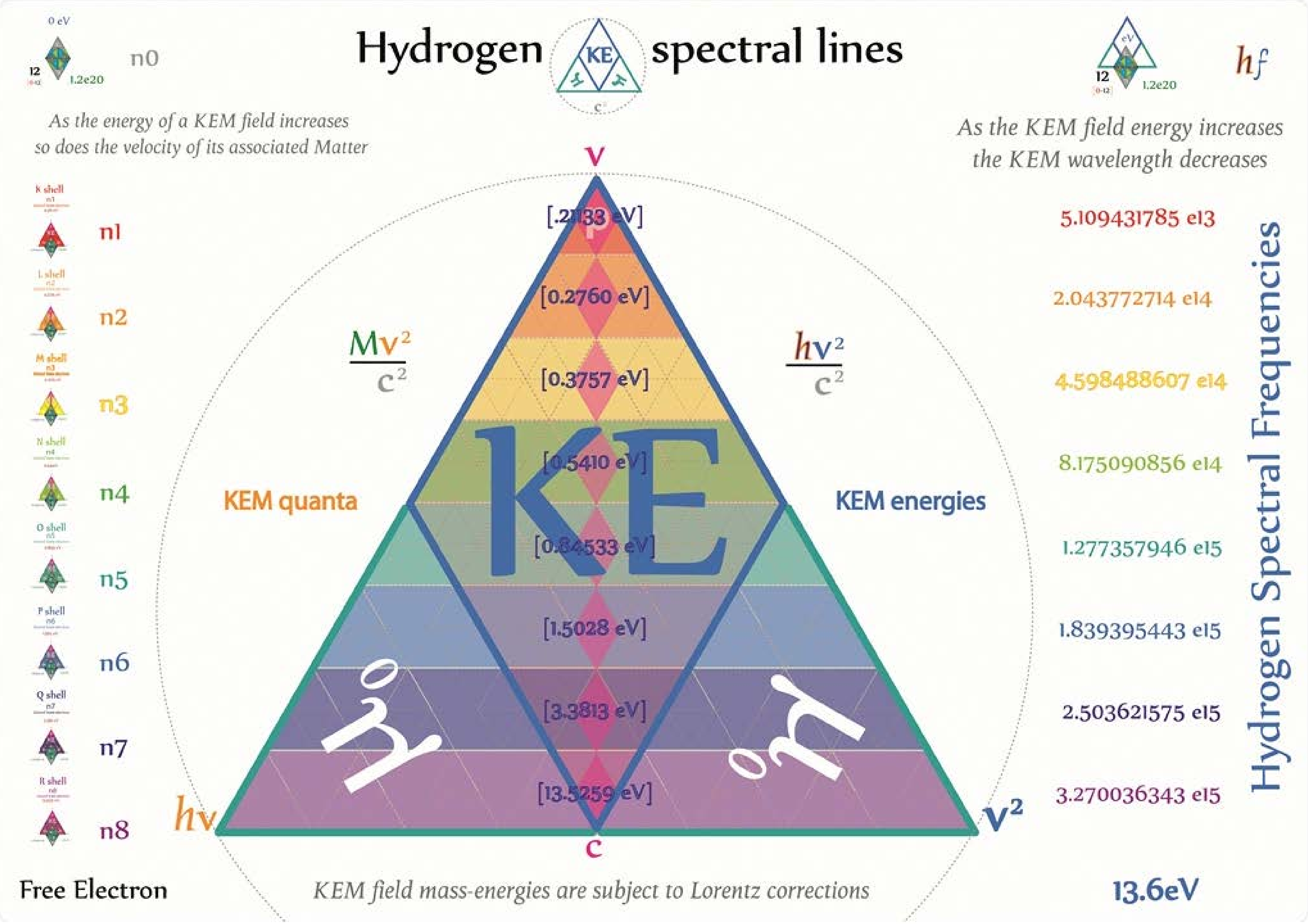
0 eV
12
1.2e20
n0
KE
c
12
1.2e20
hf
v
K shell
n1
n1
L shell
n2
n2
M shell
n3
n3
N shell
n4
n4
O shell
n5
n5
P shell
n6
n6
Q shell
n7
n7
R shell
n8
n8
[.2133 eV]
[0.2760 eV]
[0.3757 eV]
[0.5410 eV]
[0.84533 eV]
[1.5028 eV]
[3.3813 eV]
[13.5259 eV]
KE
$\frac{Mv^2}{c^2}$
$\frac{hv^2}{c^2}$
KEM quanta
KEM energies
hv
c
v²
5.109431785 e13
2.043772714 e14
4.598488607 e14
8.175090856 e14
1.277357946 e15
1.839395443 e15
2.503621575 e15
3.270036343 e15
Hydrogen Spectral Frequencies
13.6eV
Free Electron

As the energy of a KEM field increases so does the velocity of its associated Matter

As the KEM field energy increases the KEM wavelength decreases

KEM field mass-energies are subject to Lorentz corrections

Tetryonics 29.05 - Hydrogen Spectral Frequencies

Absorption lines

Increase in photo-electronic energies

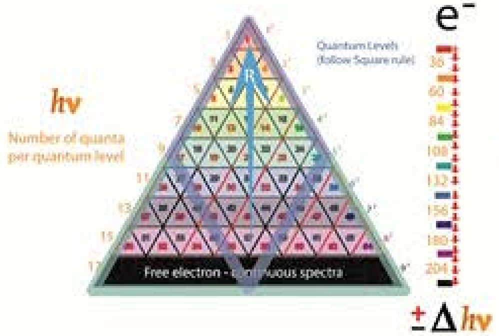
hv
Number of quanta per quantum level
Quantum Levels (follow Square rule)
R
Free electron - continuous spectra
e⁻
36
60
84
108
132
156
180
204
±Δhv

$$\Delta 2h\nu = E = hf$$

When an electron 'absorbs' a photon of energy momenta it 'jumps' from one energy level to another, dependent on the energy and frequency of the incident photon

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

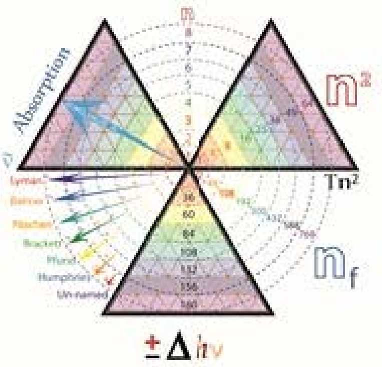
Absorption
n²
Tn²
n_f
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
36
60
84
108
132
156
180
±Δhv

The quantum level of the nuclei determines the ground quantum level of electrons within the nucleus
All transitions within the nulceus are discrete quantum jumps - outside the nucleus all spectra are continuous

KEM A Hydrogen photo-electron can only exist in specific energy levels below 13.525eV within the nucleus (any excess KE results in a unbound electron with KE - velocity)

KE in excess of 13.525 eV results in unbound electron

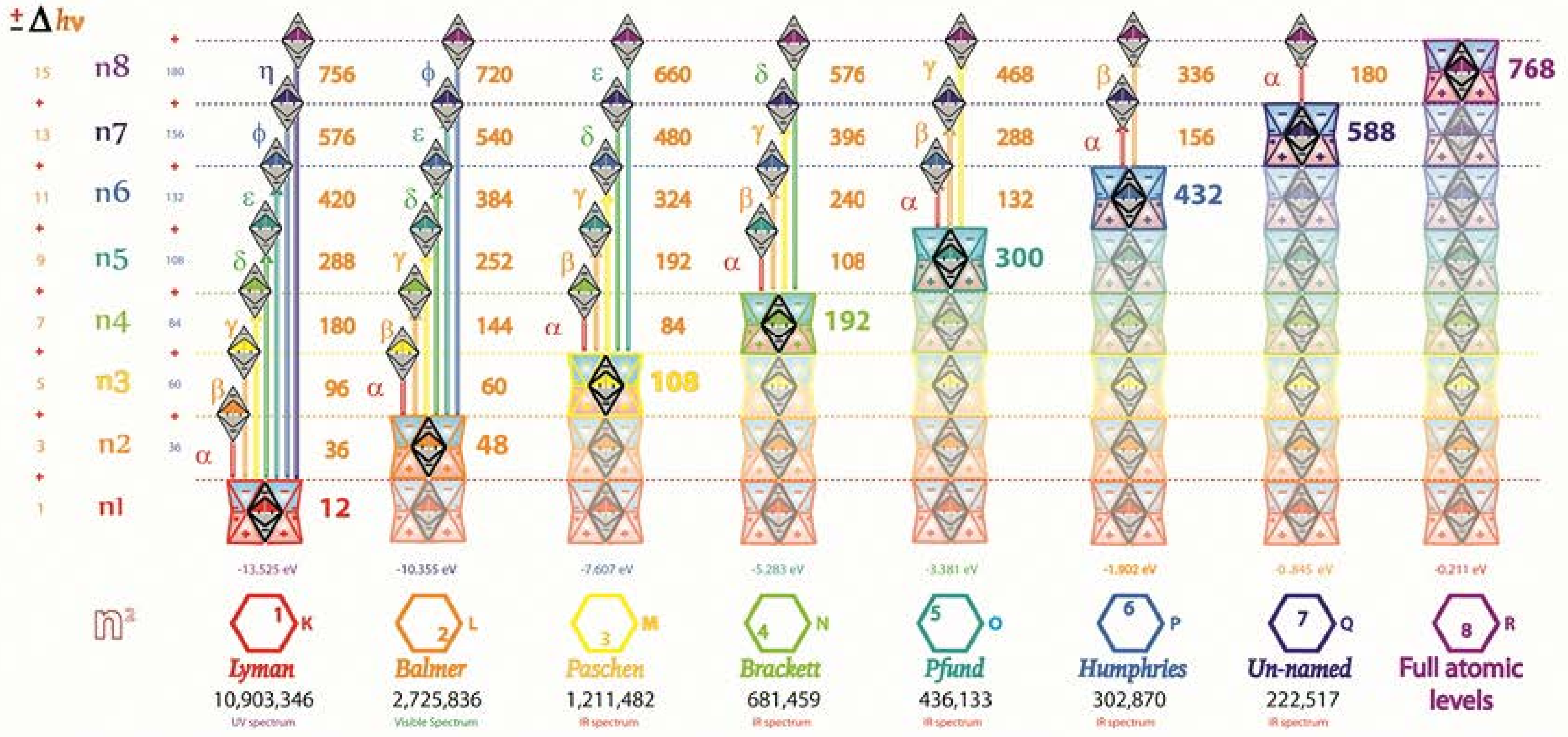
±Δhv
15
n8
180
13
n7
156
11
n6
132
9
n5
108
7
n4
84
5
n3
60
3
n2
36
1
n1
η 756
φ 576
ε 420
δ 288
γ 180
β 96
α 36
12
φ 720
ε 540
δ 384
γ 252
β 144
α 60
48
ε 660
δ 480
γ 324
β 192
α 84
108
δ 576
γ 396
β 240
α 108
192
γ 468
β 288
α 132
300
β 336
α 156
432
α 180
588
768
-13.525 eV
-10.355 eV
-7.607 eV
-5.283 eV
-3.381 eV
-1.902 eV
-0.845 eV
-0.211 eV
n²
1 K
Lyman
10,903,346
UV spectrum
2 L
Balmer
2,725,836
Visible Spectrum
3 M
Paschen
1,211,482
IR spectrum
4 N
Brackett
681,459
IR spectrum
5 O
Pfund
436,133
IR spectrum
6 P
Humphries
302,870
IR spectrum
7 Q
Un-named
222,517
IR spectrum
8 R
Full atomic levels

Copyright ABRAHAM [2008] - all rights reserved

KEM fields & quantum jumps

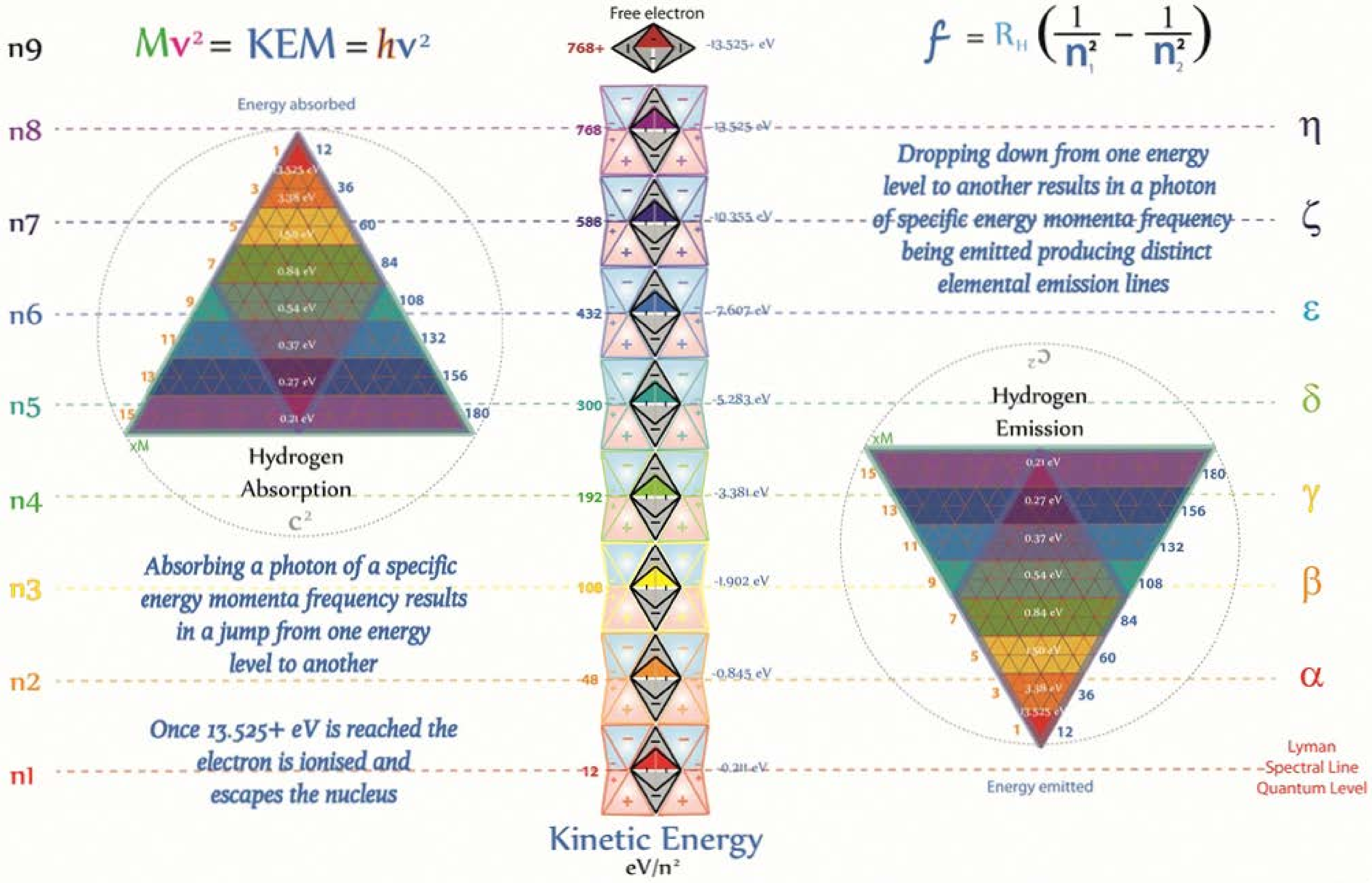

$$Mv^2 = KEM = hv^2$$

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Absorbing a photon of a specific energy momenta frequency results in a jump from one energy level to another

Once 13.525+ eV is reached the electron is ionised and escapes the nucleus

Dropping down from one energy level to another results in a photon of specific energy momenta frequency being emitted producing distinct elemental emission lines

Kinetic Energy

eV/n²

n9
n8
n7
n6
n5
n4
n3
n2
n1
Energy absorbed
Hydrogen Absorption
c²
xM
1
3
5
7
9
11
13
15
12
36
60
84
108
132
156
180
13.525 eV
3.38 eV
1.50 eV
0.84 eV
0.54 eV
0.37 eV
0.27 eV
0.21 eV
Free electron
768+
-13.525+ eV
768
-13.525 eV
588
-10.355 eV
432
-7.607 eV
300
-5.283 eV
192
-3.381 eV
108
-1.902 eV
48
-0.845 eV
12
-0.211 eV
Hydrogen Emission
Energy emitted
η
ζ
ε
δ
γ
β
α
Lyman
Spectral Line
Quantum Level

Emission lines

Decrease in photo-electronic energies

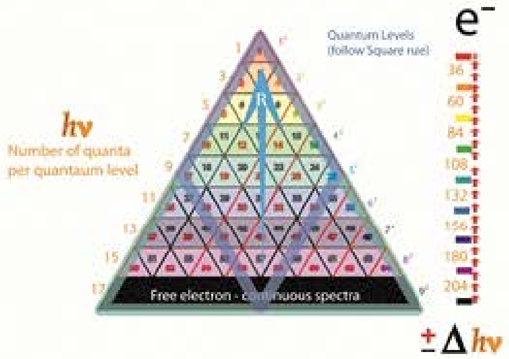
hv
Number of quanta per quantaum level
Quantum Levels (follow Square rule)
e⁻
1
3
5
7
9
11
13
15
17
R
36
60
84
108
132
156
180
204
Free electron - continuous spectra
±Δhv

$$\Delta 2hv = E = hf$$

When an electron 'emits' a photon of energy momenta it 'drops' from one energy level to another dependent on the energy and frequency of the ejected photon

$$f = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

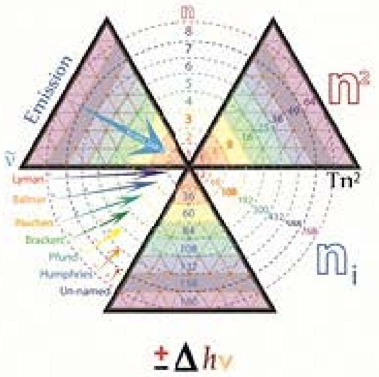
Emission
n
8
7
6
5
4
3
n²
Tn²
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
36
60
84
108
132
156
n
i
±Δhv

The quantum level of the nuclei determines the ground quantum level of electrons within the nucleus
All transitions within the nulceus are discrete quantum jumps - outside the nucleus all spectra are continuous

KEM A Hydrogen electron can only exist in specific energy levels below 13.525eV within the nucleus (any excess KE must be shed as Photons)

KE in excess of 13.525 eV results in unbound electron

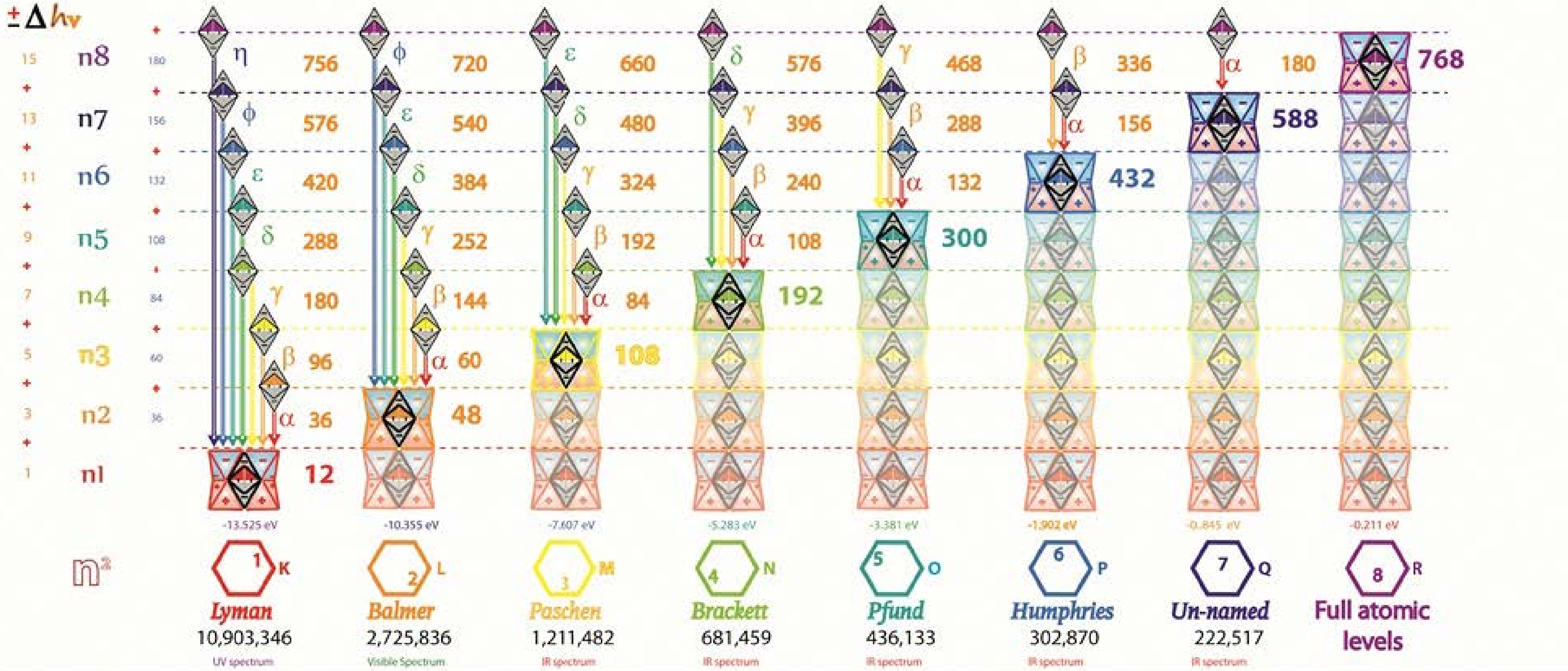
±Δhv
15 n8 180 η 756 φ 720 ε 660 δ 576 γ 468 β 336 α 180 768
13 n7 156 φ 576 ε 540 δ 480 γ 396 β 288 α 156 588
11 n6 132 ε 420 δ 384 γ 324 β 240 α 132 432
9 n5 108 δ 288 γ 252 β 192 α 108 300
7 n4 84 γ 180 β 144 α 84 192
5 n3 60 β 96 α 60 108
3 n2 36 α 36 48
1 nl 12
-13.525 eV
-10.355 eV
-7.607 eV
-5.283 eV
-3.381 eV
-1.902 eV
-0.845 eV
-0.211 eV
n²
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R
Lyman
10,903,346
UV spectrum
Balmer
2,725,836
Visible Spectrum
Paschen
1,211,482
IR spectrum
Brackett
681,459
IR spectrum
Pfund
436,133
IR spectrum
Humphries
302,870
IR spectrum
Un-named
222,517
IR spectrum
Full atomic levels

Copyright ABRAHAM [2008] - all rights reserved

Spectral line series

(are produced by changes in linear momentum in photo-electric KEM fields due to atomic transitions of electrons between squared energy levels)

Lyman	27.49545417
Balmer	109.9818167
Paschen	247.4590875
Brackett	439.9272667
Pfund	687.3863542
Humphries	989.8363501
Un-named	1,347.277254

Square root of Energy required to transition nuclear quantum levels

$$[\sqrt{756}].n^2$$

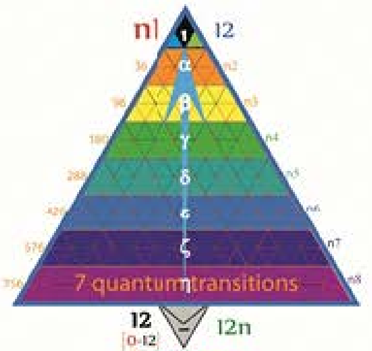
n1
12
1
α
β
γ
δ
ε
ζ
η
7 quantum transitions
12
[0-12]
12n
n1-8

Lyman

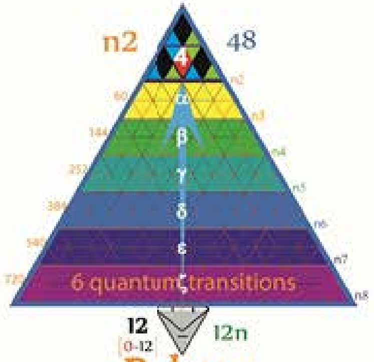
n2
48
4
α
β
γ
δ
ε
ζ
6 quantum transitions
12
[0-12]
12n
n2-8

Balmer

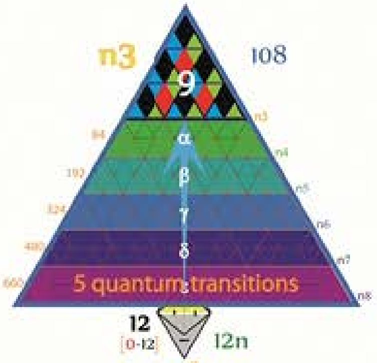
n3
108
9
α
β
γ
δ
ε
5 quantum transitions
12
[0-12]
12n
n3-8

Paschen

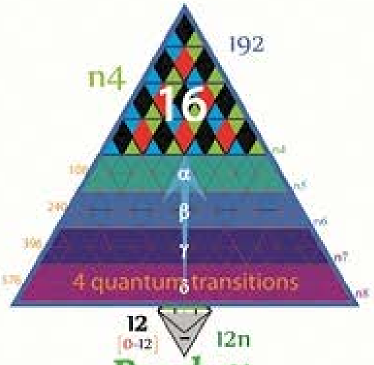
n4
192
16
α
β
γ
δ
4 quantum transitions
12
[0-12]
12n
n4-8

Brackett

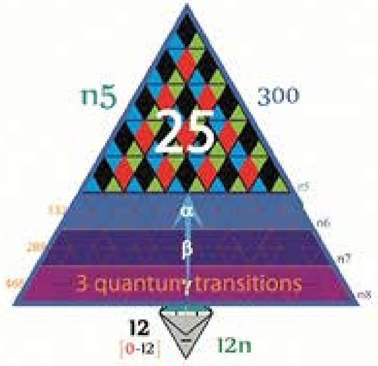
n5
300
25
α
β
γ
3 quantum transitions
12
[0-12]
12n
n5-8

Pfund

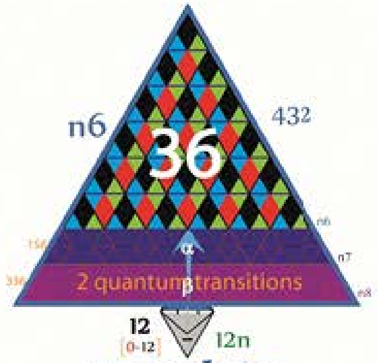
n6
432
36
α
β
2 quantum transitions
12
[0-12]
12n
n6-8

Humphries

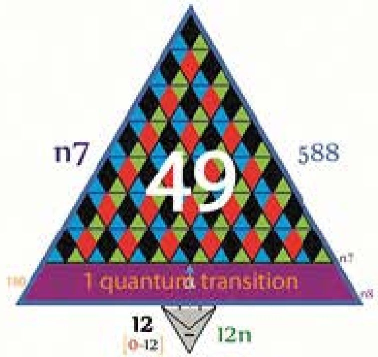
n7
588
49
α
1 quantum transition
12
[0-12]
12n
n7-8

Un-named

Tetryonics 29.09 - Spectral line series

Rydberg's Constant

Johannes Robert Rydberg

(November 8, 1854 – December 28, 1919)

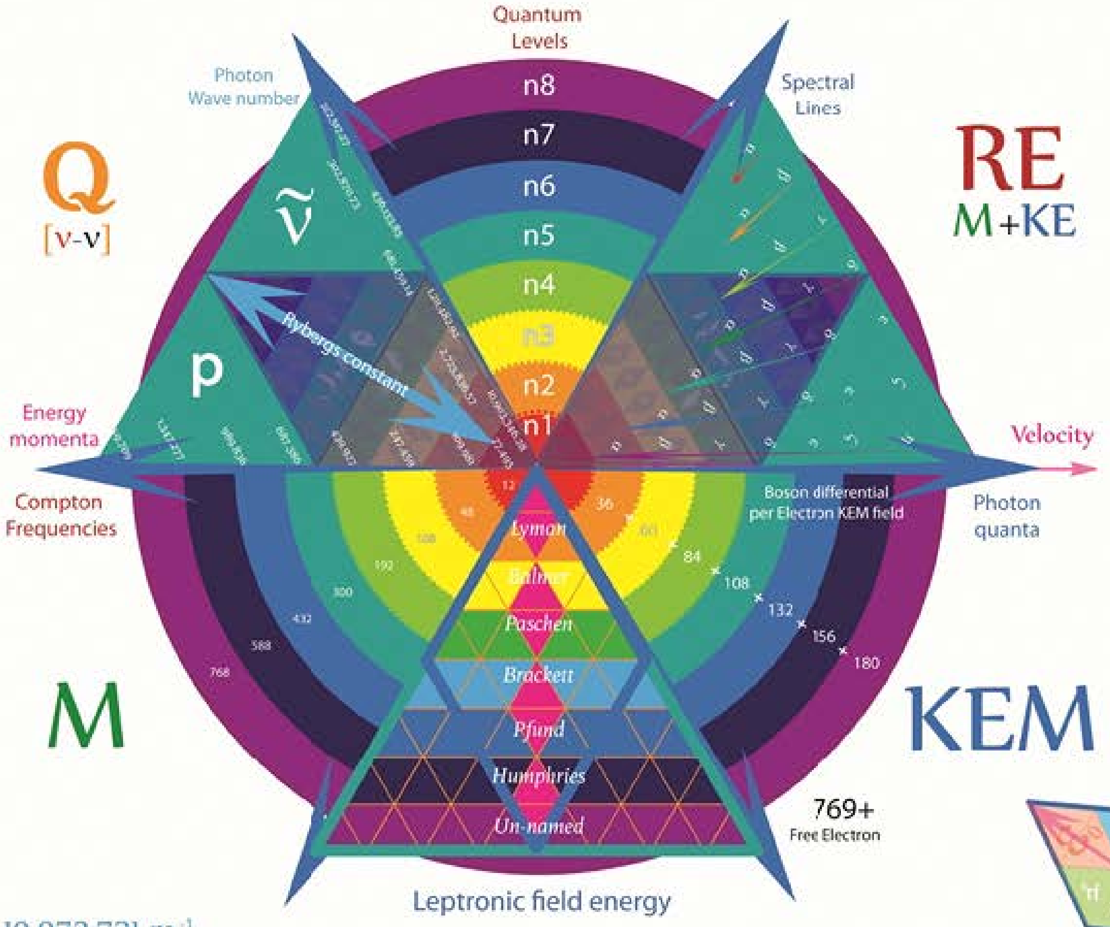
Quantum
Levels
n8
n7
n6
n5
n4
n3
n2
n1
Photon
Wave number
Spectral
Lines
Q
[ν-ν]
ν̃
p
Rydbergs constant
RE
M+KE
Energy
momenta
Compton
Frequencies
Velocity
Boson differential
per Electron KEM field
Photon
quanta
12
36
48
60
84
108
132
156
180
192
300
432
588
768
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
M
KEM
769+
Free Electron
Leptronic field energy

$$\tilde{\nu} = \frac{c}{27.49545}$$

$$f = \frac{c^2}{27.49545}$$

$$\lambda = \frac{27.49545}{c}$$

$$p = 27.49545$$

$$R_H = \frac{c}{27.49545}$$

1.0903346.28 x 10^7 m-1

$$R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

$$\lambda = v/f$$

Using the Tetryonic model of a photo-electron and its associated quantum KEM energy levels we see that Rydberg's constant is a linear measure of the photo-electronic KEM field - and in turn directly related to its scalar energy-momenta through 'c'

Currently accepted value - 10,973,731 m^{-1}

Rydberg's Constant can be shown to be the square root of the total number of quanta required to transition a n1 electron to n8

(the highest quantum level possible in elemental nuclei)

as per Planck's heat law

Tetryonic theory value - 10,903,346 m^{-1}

$$p^2 = KEM = Mv^2$$

Quantum Jump

The energy momenta of emitted photons is a function of the photo-electron's KEM wavefunction

27.49545

KE

Photons

$$\boxed{Mv^2 = KEM = hcR_H}$$

Tetryonics 29.10 - Rydberg Constant

Copyright ABRAHAM [2008] - all rights reserved

Wavenumbers and Planck's Law

Wavenumbers are revealed to be the number of quanta [nhv] required for a photo-electron to transition between specific quantum levels divided by the speed of light [c]

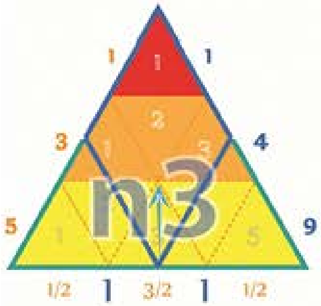

Balmer spectral lines [absorption]

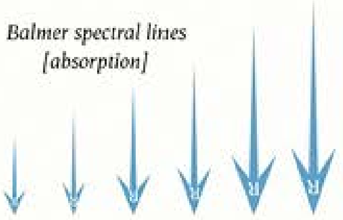

$$E = n[h\nu]$$

Wavenumbers are the number of Planck quanta required to reach a specific energy level

Max Karl Ernst Ludwig Planck

[April 23, 1858 - October 4, 1947]

Balmer spectral lines [emission]

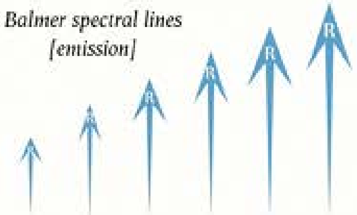

$$E = hc\tilde{\nu}$$

Photo-electrons emit and absorb photons to create blackbody radiation

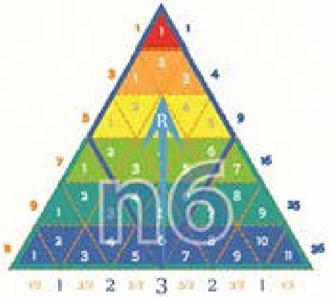

Planck's law describes the electromagnetic radiation emitted by a black body in thermal equilibrium at a specific temperature

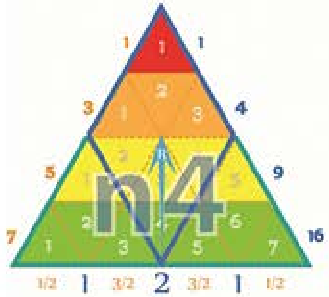

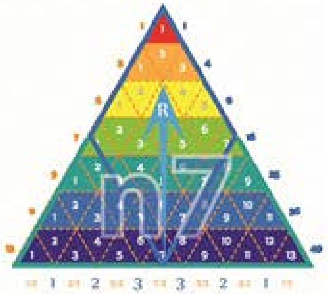

$$\Delta[nh\nu]$$

SQUARED energy levels have ODD numbered quanta in each energy level [bosons]

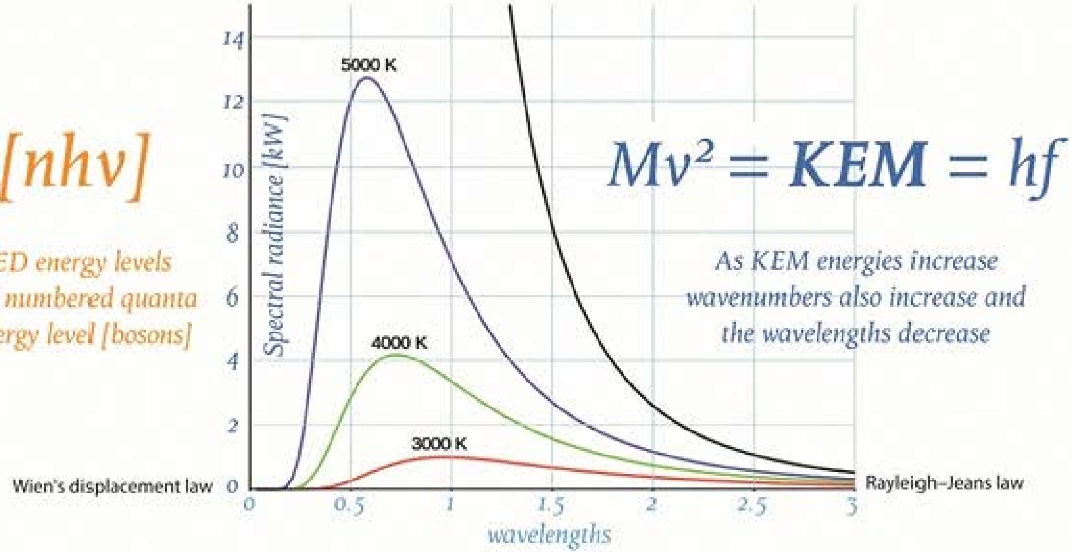

$$Mv^2 = \mathbf{KEM} = hf$$

As KEM energies increase wavenumbers also increase and the wavelengths decrease

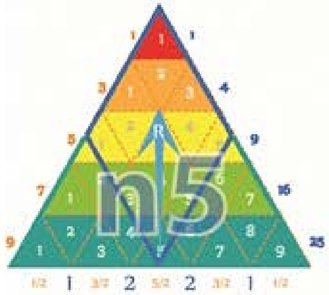

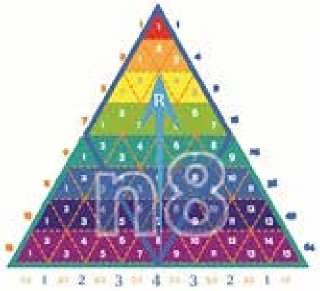

Spectral lines can be understood in quantum theory as differences between energy levels and proportional to wavenumber, frequency or wavelengths

Copyright ABRAHAM [2008] - all rights reserved

Kinetic mass-energy momenta relationships

Photons are comprised of dual EM masses

Photons are KEM energy-momenta propagating at the speed of Light

EM Field Planck quanta

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons ElectroMagnetic mass velocity

kg m

Q
Photons have neutral charge
[ν-ν]

f

RE
Photons are Lorentz variant

p $kg\frac{m}{s}$

$\frac{1}{s}$ f

v $\frac{m}{s}$

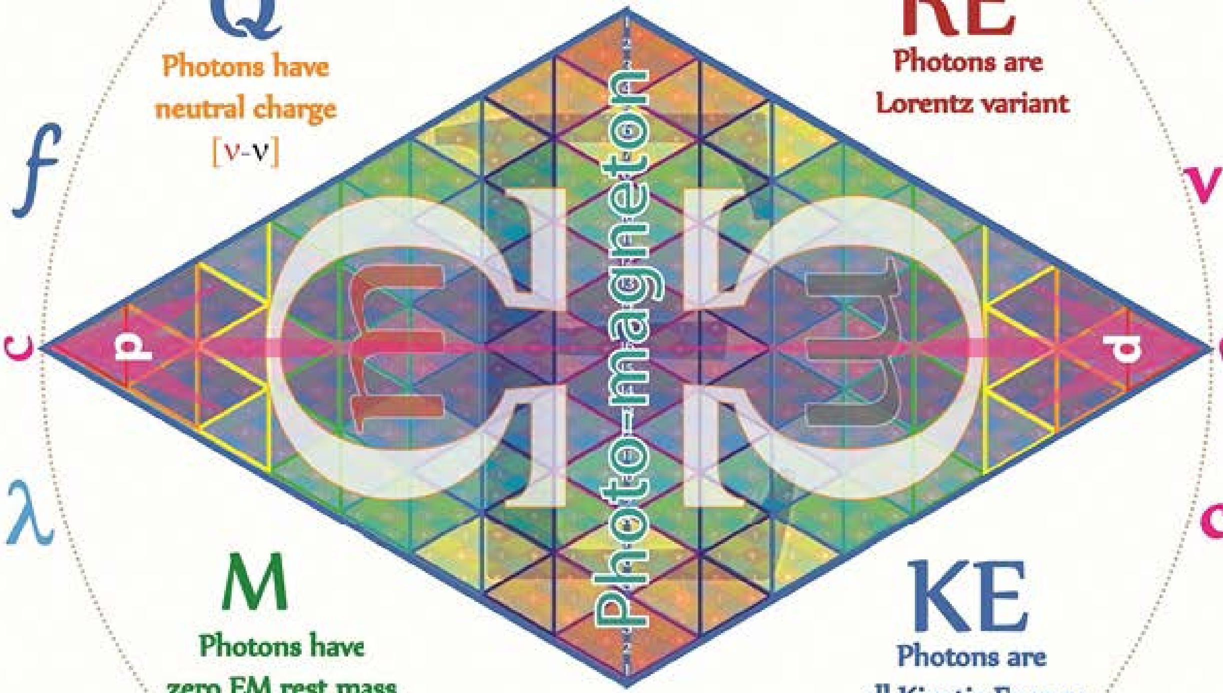

Frequency and Wavelength are related by the natural speed of Energy

c

c

'c' is the natural speed of Energy in free space

m λ

M
Photons have zero EM rest mass

KE
Photons are all Kinetic Energy

c $\frac{m}{s}$

KG 0

E $kg\frac{m^2}{s^2}$

$$\Delta 2h\nu = E = hf$$

Planck Einstein

Photons are 'Matterless' 2D neutral charge mass-energy geometries

c^2
1 second

Kinetic Energy is 1/2 the energy created by Matter topologies in motion

Tetryonics 29.12 - Kinetic mass-energy relationships

Copyright ABRAHAM [2008] - all rights reserved

Relativistic Lorentz corrections

Hendrik Lorentz

$$\sqrt{1-\beta^2}$$

$$E=\frac{mc^2}{\sqrt{1-\frac{v^2}{c^2}}}$$

$$\sqrt{1-\left(\frac{v}{c}\right)^2}$$

(18 July 1853 – 4 February 1928)

The Lorentz transformation was originally the result of attempts by Lorentz and others to explain how the speed of light was observed to be independent of the reference frame and later expanded upon in an attempt to understand the creation of emfs in magnets and symmetries in electromagnetic forces

Einstein inlcuded the transformation in his theory of Special relativity [SR].

The Lorentz transformation supersedes the Galilean transformation of Newtonian physics, which assumes an absolute space and time.

According to special relativity, the Galilean transformation is a good approximation only at relative speeds much smaller than the speed of light.

Tetryonics expands upon all of this revealing this relationship as a result of the geometry of electromagnetic mass-energies in motion

rest mass-Matter

Energy

Planck quanta

$$\frac{4n\pi}{c^4}\left[\left[m\Omega v^2\right]\right]$$

mass velocity

momentum

spatial co-ordinate system

Matter is EM energy propagating at the 'speed of Light' in a standing wave energy topology

One of the greatest mistakes in relativistic mechanics is the application of Lorentz corrections to Matter.

It stems from there being no definition and enforced differentiation between EM mass and Matter

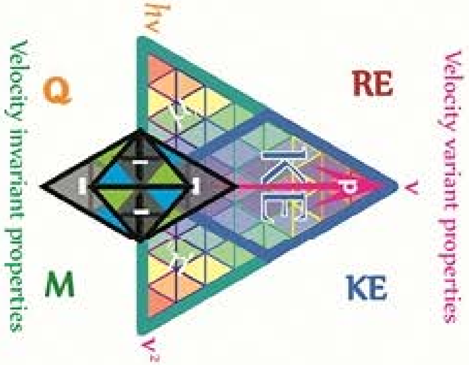

$$m = \frac{E}{c^2}$$

2D planar fields are relativisitic

radiant EM mass geometry

$$\frac{m}{c^2} = M$$

standing-wave Matter topology

3D standing-waves are velocity invariant

$$\frac{E}{c^4} = M$$

mass is a property of Matter - Matter is not a property of mass [they are directly related throught the velocity of light]

$$\beta = \left[\frac{v}{c}\right]$$

velocity is a result of vector forces

All 2D energy waveforms propagate at the 'speed of Light'

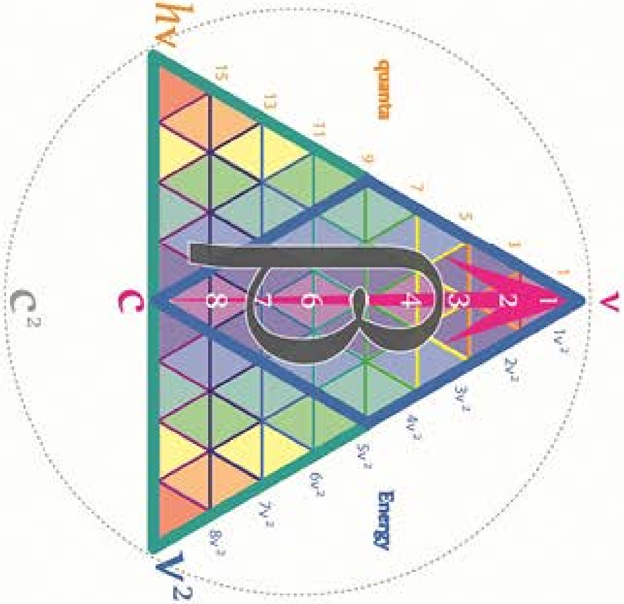

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

The Kinetic Energy of Matter in motion is directly related to the square of the velocity

$$\gamma \equiv \frac{c}{\sqrt{c^2-v^2}} = \frac{1}{\sqrt{1-\beta^2}} = \frac{dt}{d\tau}$$

The 'speed of Light' is a EM constant and is the limiting velocity acheivable by the electrical acceleration of Matter

Spectral Energy relationships

$$\frac{\text{kg m}^2}{\text{s}^2} \quad E = p^2$$

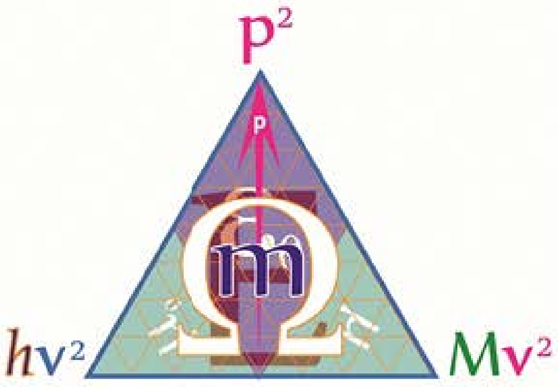
p²
P
Ω
m
hv²
Mv²

Scalar EM mass
-energy momenta

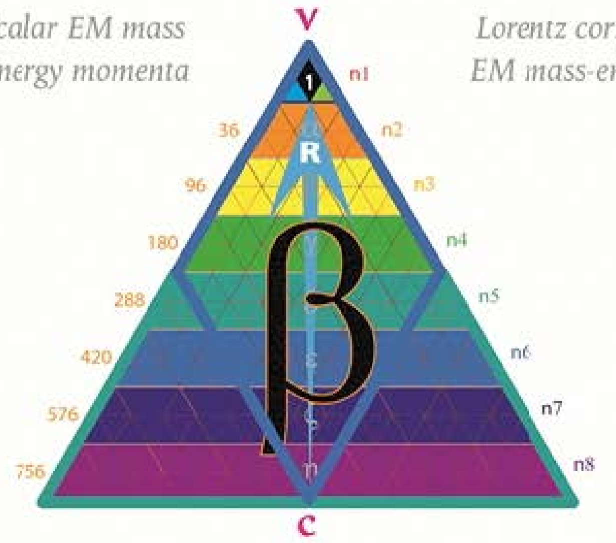
v
1
n1
36
n2
R
96
n3
180
n4
β
288
n5
420
n6
576
n7
756
n8
c

Lorentz corrected
EM mass-energies

$$KEM = \frac{mv^2}{c^2}$$

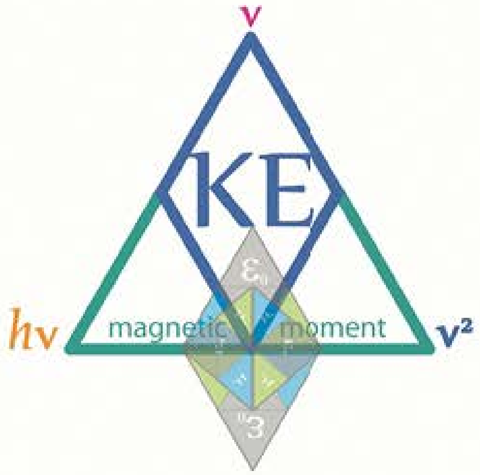
v
KE
ε
hv
magnetic
moment
v²
μ

$$\boxed{Mv^2 = \underset{13.525\ \text{eV}}{KEM} = hcR_H}$$

All Spectral line emissions and absorptions
produce changes in KEM energies, Angular momentum,
Linear momentum, Frequency and Wavelengths

Linear

$$\left[\frac{\Omega}{c}\right] = \lambda$$

de Broglie wavelengths

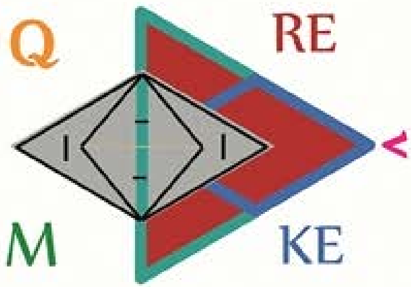

The 'speed of Light' is the maximum velocity acheivable by electrical energies

Scalar

$$f = \left[\frac{c^2}{\Omega}\right]$$

Compton frequencies

Tetryonics 29.14 - Spectral Energy relationships

Copyright ABRAHAM [2008] - all rights reserved

Spectral line photon production

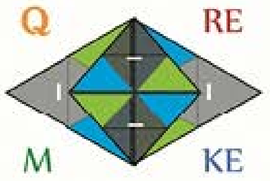

At no time does the 3D rest Matter-energy content of the electron change

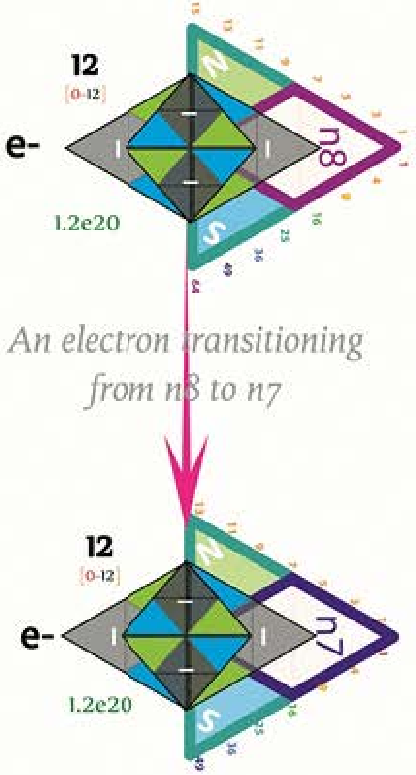
12
[0-12]
e-
n8
1.2e20
12
[0-12]
e-
n7
1.2e20

An electron transitioning from n8 to n7

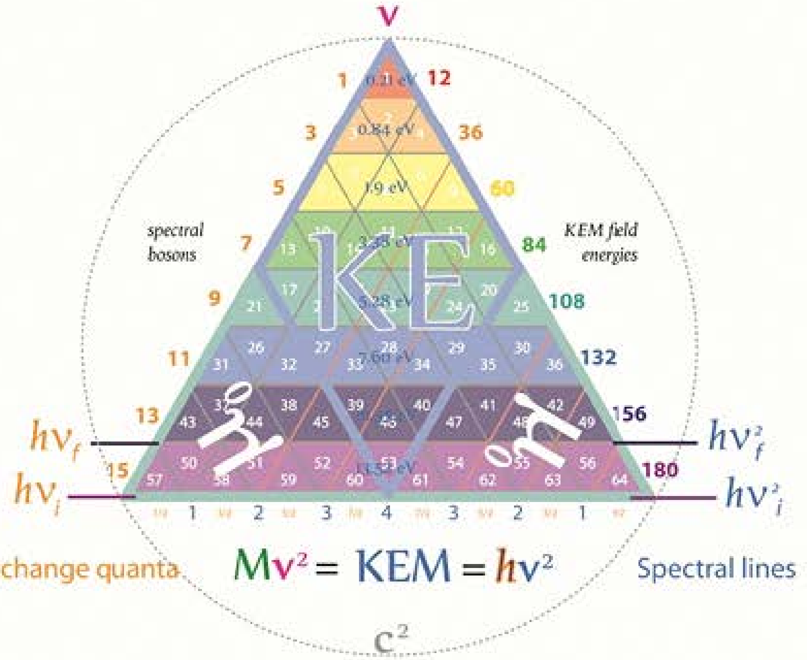
v
1
12
0.21 eV
3
36
0.84 eV
5
60
1.9 eV
spectral bosons
7
84
KEM field energies
3.35 eV
KE
9
108
5.28 eV
11
132
7.6 eV
13
156
15
180
$h\nu_f$
$h\nu_i$
$h\nu^2_f$
$h\nu^2_i$
1
2
3
4
3
2
1
Exchange quanta
Spectral lines
c²

$$Mv^2 = \text{KEM} = hv^2$$

$$E = hf$$

As the 3D rest Matter of any particle in motion is Lorentz invariant any changes in velocity-momentum produces changes to the mass-energy momenta content of the 2D KEM field [boson changes are the difference of two squares]

Spectral line wavenumbers

Tetryonic

$$\left(\frac{15 \nearrow 49}{64}\right)$$

.0047

$$1/\lambda$$

Rydberg

$$R\left(\frac{1}{49} - \frac{1}{64}\right)$$

.0047

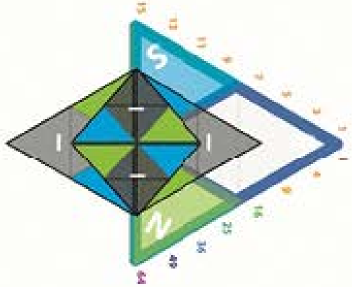

Does so by emitting Kinetic Energy quanta

λ	19,174.54158	nm
$\tilde{\nu}$	52,152.48541	m-1
f	$1.563492179 \times 10^{13}$	Hz
E	0.064693658	eV
Ω	5,748.382952	m²/s

Which are released as photons of specific energy-momenta

un-named

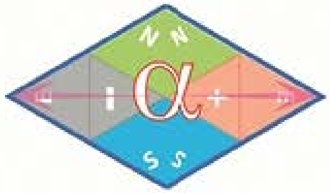

of specific wavelengths (Spectral Lines)

Wavenumbers

are the number of energy momenta wavelengths per unit distance

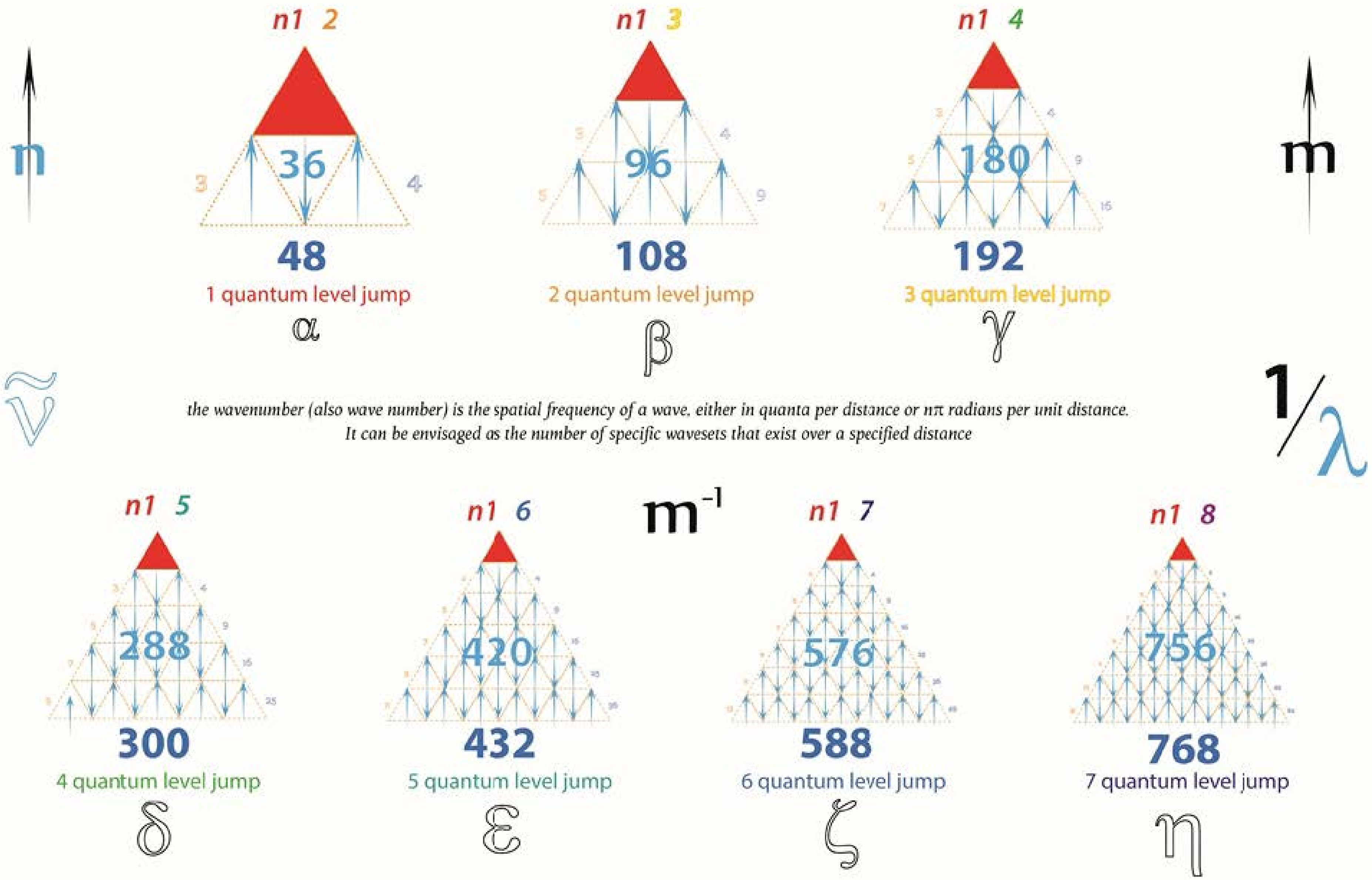
m
n1 2
36
3
4
48
1 quantum level jump
α
n1 3
96
3
4
5
9
108
2 quantum level jump
β
n1 4
180
3
4
5
9
7
16
192
3 quantum level jump
γ
m
ṽ
1/λ
m⁻¹
n1 5
288
300
4 quantum level jump
δ
n1 6
420
432
5 quantum level jump
ε
n1 7
576
588
6 quantum level jump
ζ
n1 8
756
768
7 quantum level jump
η

the wavenumber (also wave number) is the spatial frequency of a wave, either in quanta per distance or nπ radians per unit distance.
It can be envisaged as the number of specific wavesets that exist over a specified distance

For electromagnetic radiation in vacuum, wavenumber is proportional to frequency and to photon energy

Bosons, Photons and quantums of energy momenta

$\tilde{\nu}$ WAVENUMBERS

$E = \Sigma h\nu$

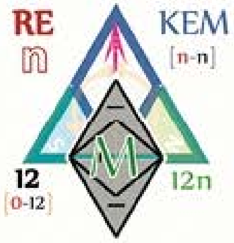
RE
n
KEM
[n-n]
M
12
[0-12]
12n

$KEM = Mv^2$

LINEAR MOMETNUM **p**

γ Spectral lines are produced $\Delta m\Omega v = \Delta mv = \Delta p$ by accelerating electrons **a**

equilateral energy momenta is the foundation of all quantum EM wave-functions

ODD distribution $2\nu = f$ **SQUARED distribution**

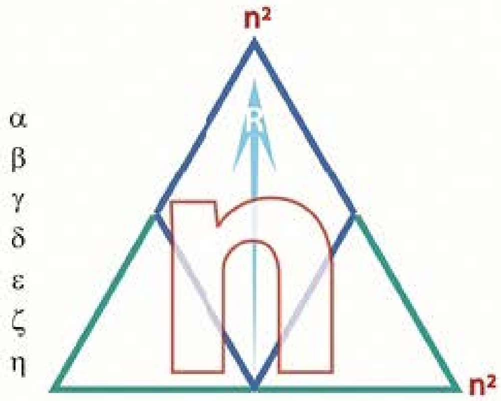
n²
α
β
γ
δ
ε
ζ
η
n
n²

Bosons

$h\nu$

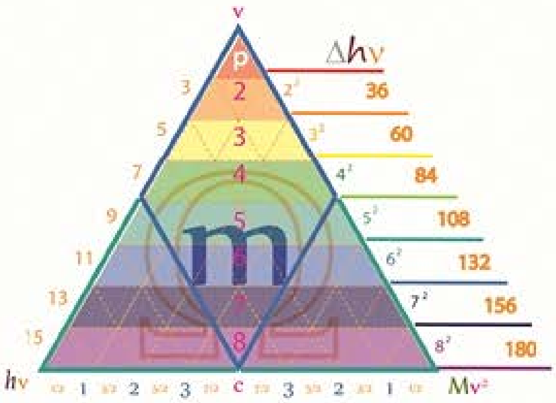
v
Δhv
p
2
3
4
5
6
7
8
m
3
5
7
9
11
13
15
2²
3²
4²
5²
6²
7²
8²
36
60
84
108
132
156
180
hv
1
2
3
c
3
2
1
Mv²

Kinetic energies

$h\nu^2$

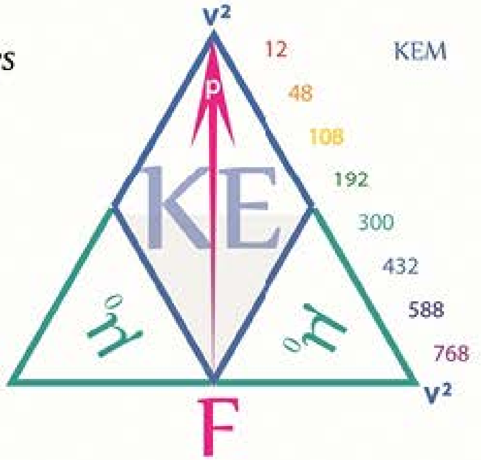
v²
p
KE
12
48
108
192
300
432
588
768
KEM
v²
F

Spectral lines

EVEN distribution

hf

Photons

energy momenta

Spectral Energy - Planck relationships

$$\frac{E}{v^2} = m\Omega = \frac{E}{c^2}$$

deBroglie wavelength λ

Frequency f

Planck quanta differential

spectral mass-energies

p MOMENTUM

h

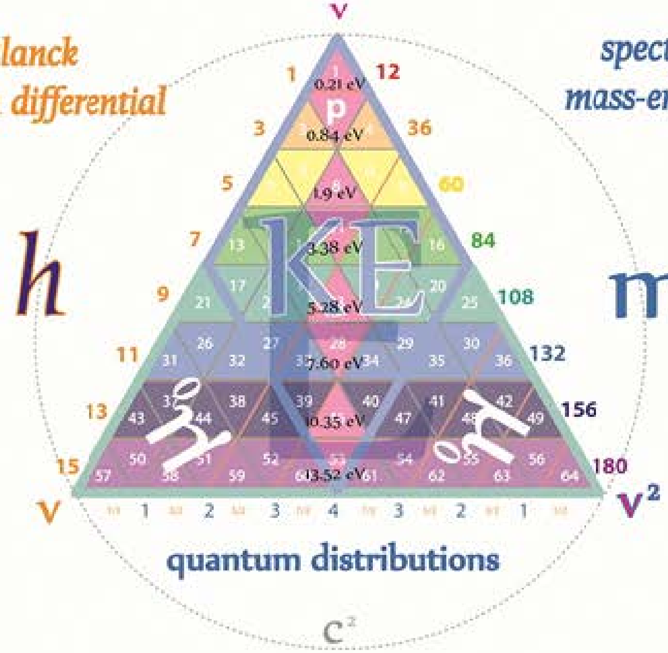
v
1
3
5
7
9
11
13
15
12
36
60
84
108
132
156
180
p
KE
0.21 eV
0.84 eV
1.9 eV
3.38 eV
5.28 eV
7.60 eV
10.35 eV
13.52 eV
v
1 2 3 4 3 2 1
v^2
c^2

m

ENERGY **E**

quantum distributions

Wave number $\tilde{v}$

v^2 Compton Frequency

Applying mass to spectral energy relationships reveals the familiar quantum mechanical wave-particle relationships and properties

	2D waves						3D Matter	
EM wave wavelength	Ω / c	=	planck constant	h / p	linear momentum	=	$m\Omega / mc$	deBroglie wavelength
EM wavenumber	c / Ω	=	linear momentum	p / h	planck constant	=	$mc / m\Omega$	Matter waves
EM wave Frequency	c^2 / Ω	=	kinetic energy	E / h	planck constant	=	$mc^2 / m\Omega$	Compton Frequency

Thus reinforcing the geometric role of quantised angular momenta in Planck's constant

Tetryonics 29.18 - Spectral energy relationships

White Light

White light is radiated by our SUN and all the stars throughout the Universe.
In fact, most of the energy radiated by the sun is within the visible spectrum, which is most likely why we see this range.

Incandescence is the greatest known generator of white light. The verb incandesce means to glow white. This is also referred to as black-body radiation.

The color of a black body object at high temperatures causes it to glow, and the waves emitted include visible light.
In addition to the sun and common light bulb, molten materials such as metal or glass also glow incandescently.

Hydrogen
Oxygen
Sulfur

Spectroscopy

Planck's Heat Law

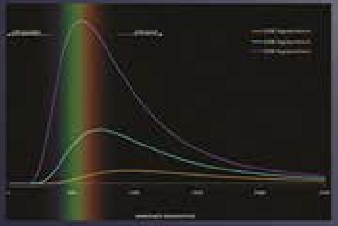

Most light sources are thermal, meaning the type of radiation they emit is a characteristic of the source's temperature.

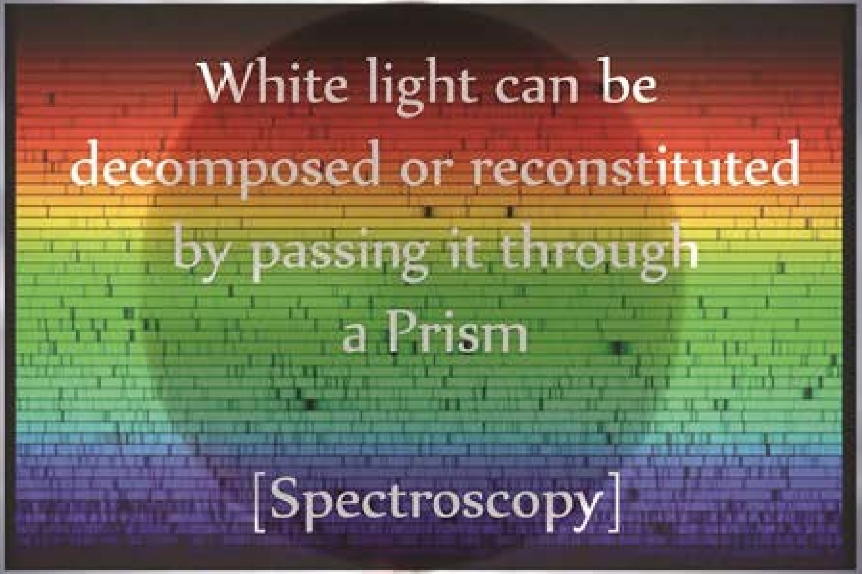

White light can be decomposed or reconstituted by passing it through a Prism

[Spectroscopy]

White light is comprised of many photons of varying frequencies and wavelengths

Each colour of light has a specific frequency and wavelength

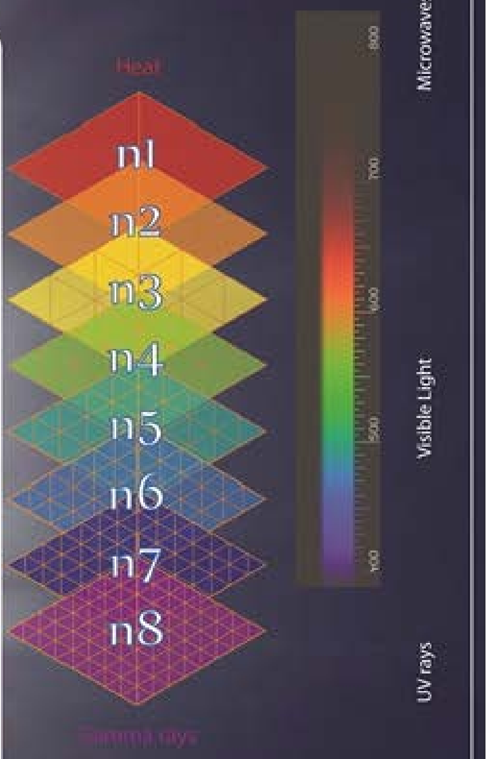

Until Newton's work proved otherwise, most scientists believed that white was the fundamental color of light; and that other colors were formed only by adding something to light. Newton demonstrated this was not true by passing white light through a prism, then through another prism. If the colors were added by the prism, the second prism should have added further colors to the single-colored beam. Since the single-colored beam remained a single color, Newton concluded that the prism merely separated the colors already present in the light.

White light is the result of superpositioning of the visible colors [varying frequencies] of EM waves

Copyright ABRAHAM [2008] - all rights reserved

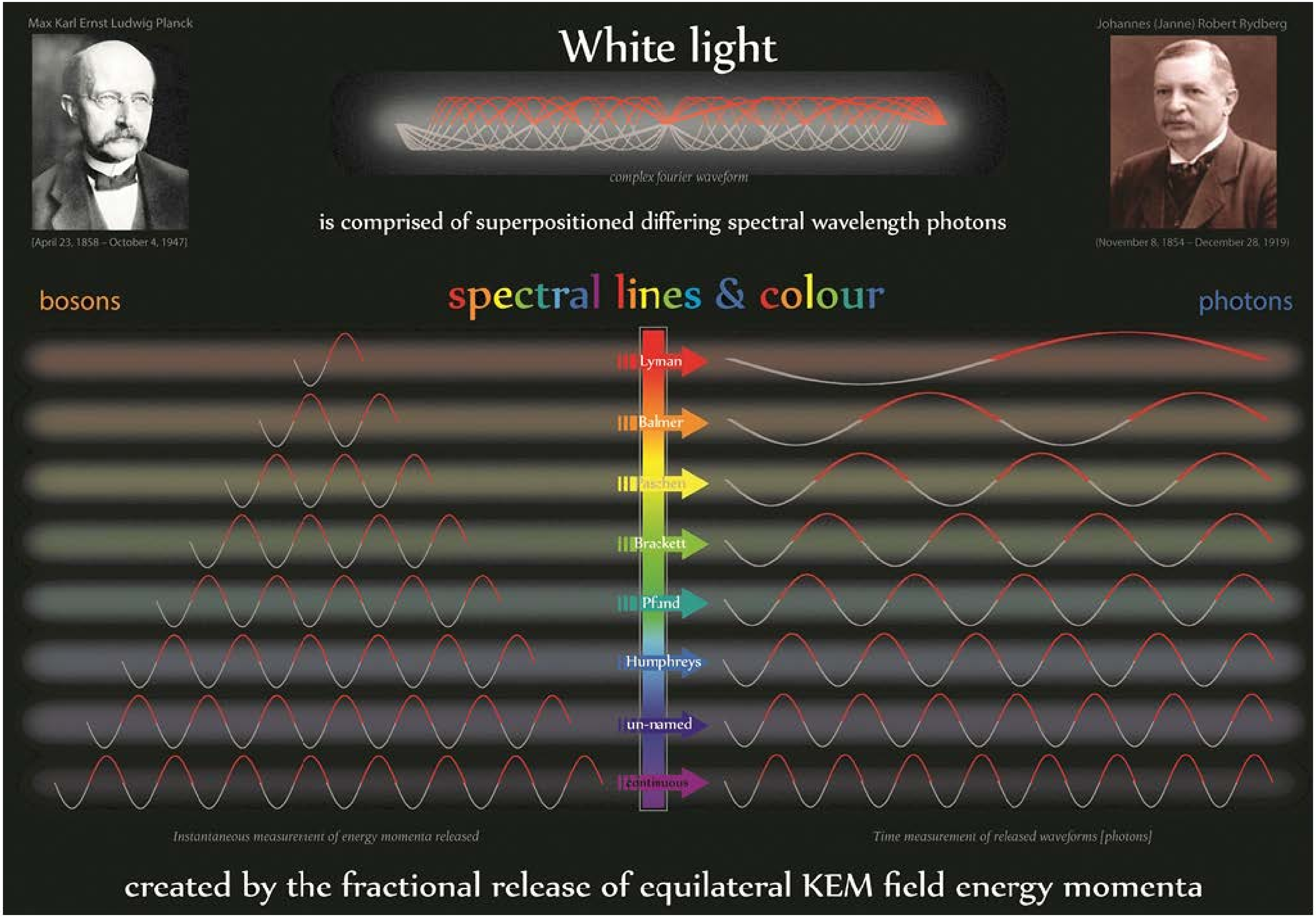

Max Karl Ernst Ludwig Planck

(April 23, 1858 – October 4, 1947)

White light

complex fourier waveform

is comprised of superpositioned differing spectral wavelength photons

Johannes (Janne) Robert Rydberg

(November 8, 1854 – December 28, 1919)

spectral lines & colour

bosons

photons

Lyman
Balmer
Paschen
Brackett
Pfund
Humphreys
un-named
continuous

Instantaneous measurement of energy momenta released

Time measurement of released waveforms [photons]

created by the fractional release of equilateral KEM field energy momenta

Tetryonics 29.20 - White light & spectral Colours

Lyman spectral lines

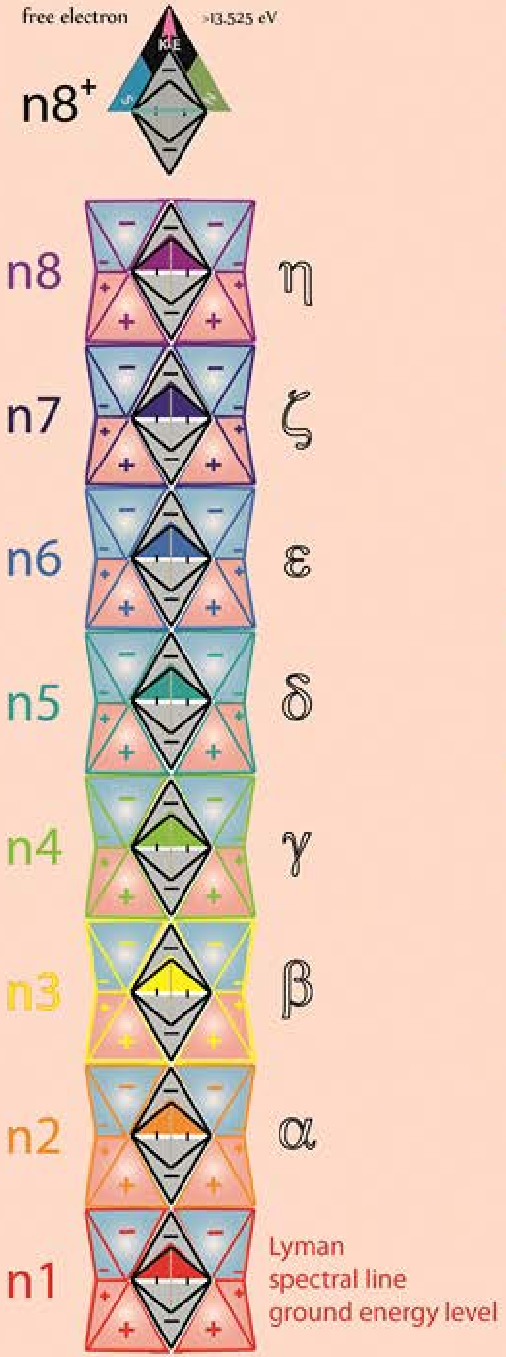
free electron
>13.525 eV
KE
n8+
n8
η
n7
ζ
n6
ε
n5
δ
n4
γ
n3
β
n2
α
n1
Lyman
spectral line
ground energy level

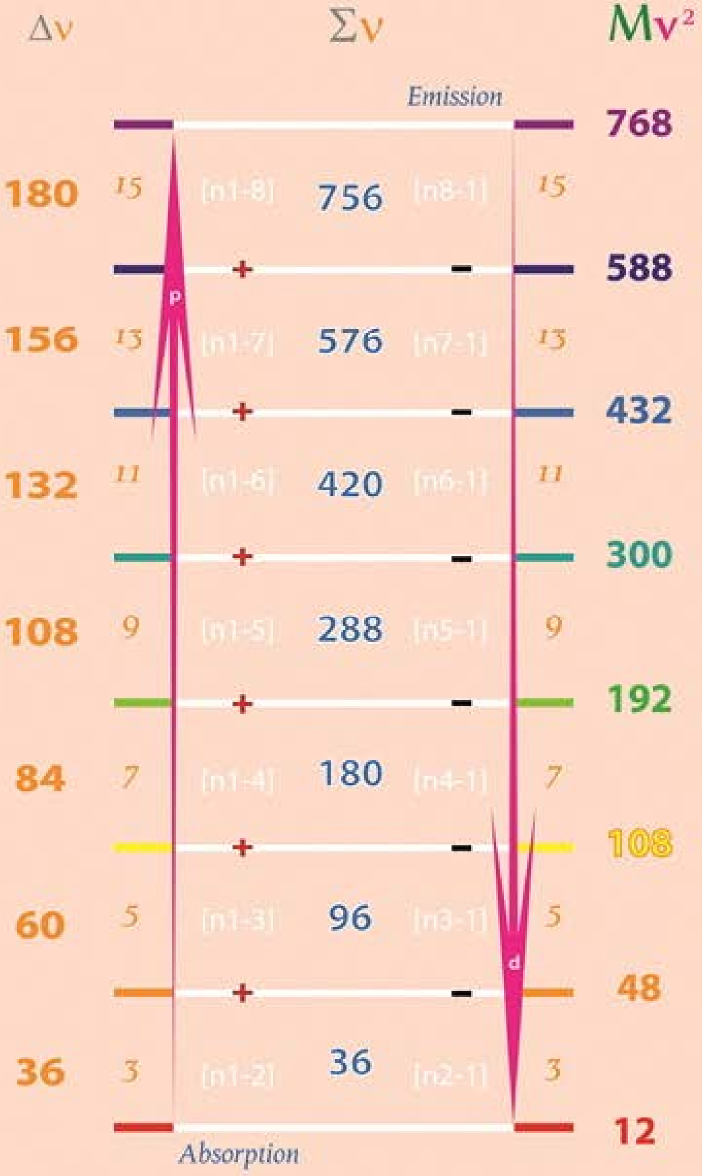
Δv
Σv
Mv²
Emission
768
180
15
[n1-8]
756
[n8-1]
15
588
156
13
[n1-7]
576
[n7-1]
13
432
132
11
[n1-6]
420
[n6-1]
11
300
108
9
[n1-5]
288
[n5-1]
9
192
84
7
[n1-4]
180
[n4-1]
7
108
60
5
[n1-3]
96
[n3-1]
5
48
36
3
[n1-2]
36
[n2-1]
3
12
p
d
Absorption

nl

Ground State

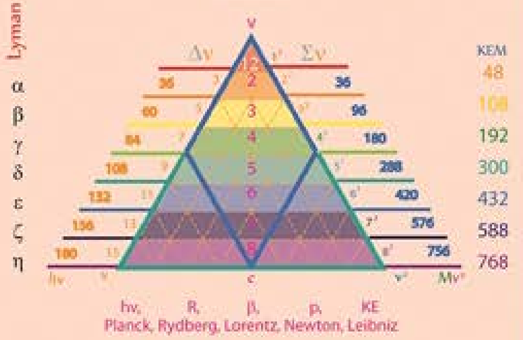
Lyman
α
β
γ
δ
ε
ζ
η
v
Δv
Σv
KEM
36
60
84
108
132
156
180
36
96
180
288
420
576
756
48
108
192
300
432
588
768
c
hv
Mv²
hv,
R,
β,
p,
KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-13.5252 eV

$$KEM = \frac{hf}{n^2}$$

v
nl
KEM
12
1.2e20

K shell

91.714 nm

$$KEM = \frac{hcR}{n^2}$$

m. Ω/v	λ	91.71496292	nm	h/p
v/Ω	$\tilde{v}$	10,903,346.28	m-1	p/h
c^2/Ω	f	$3.268740982 \times 10^{15}$	Hz	E/h
v^2/c^2	E	13.52528	eV	hf
	Ω	27.49545417	m^2/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

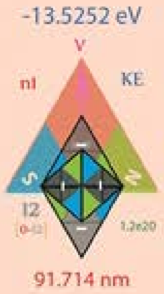
-13.5252 eV
v
nl
KE
S
N
12
[0-12]
1.2e20
91.714 nm

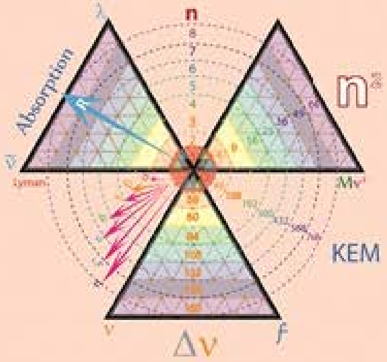
λ
n
Absorption
R
n²
Lyman
Mv²
KEM
v
Δν
f

Lyman spectral series

α	122.28	n1-2
β	103.18	n1-3
γ	97.83	n1-4
δ	95.53	n1-5
ε	94.33	n1-6
ζ	93.62	n1-7
η	93.17	n1-8

UV spectrum

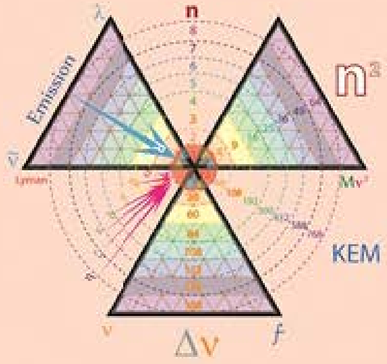
λ
n
Emission
R
n²
Lyman
Mv²
KEM
v
Δν
f

K shell

Δν

KEM

n1~2

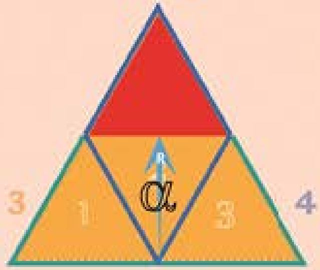

36

48

1 quantum level jump

n1~3

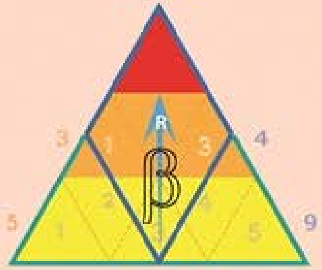
R
β
3
1
3
4
5
2
4
9
1
3
5

96

108

2 quantum level jump

n1~4

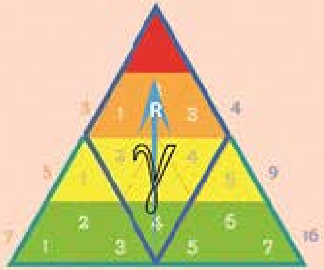
R
γ
3
1
3
4
5
9
7
2
4
6
16
1
3
5
7

180

192

3 quantum level jump

n1~5

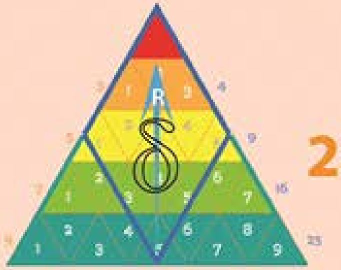

288

300

4 quantum level jump

n1~6

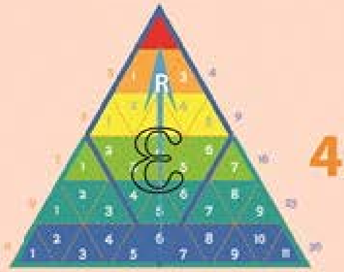

420

432

5 quantum level jump

n1~7

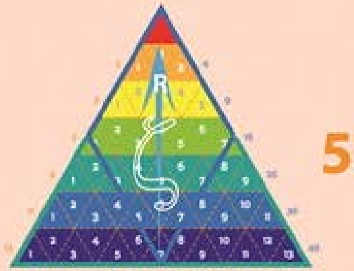

576

588

6 quantum level jump

n1~8

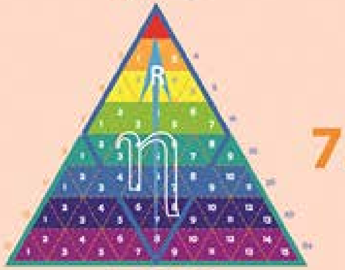

756

768

7 quantum level jump

Tetryonics 30.02 - Lyman quantum transitions

Lyman series energy momenta

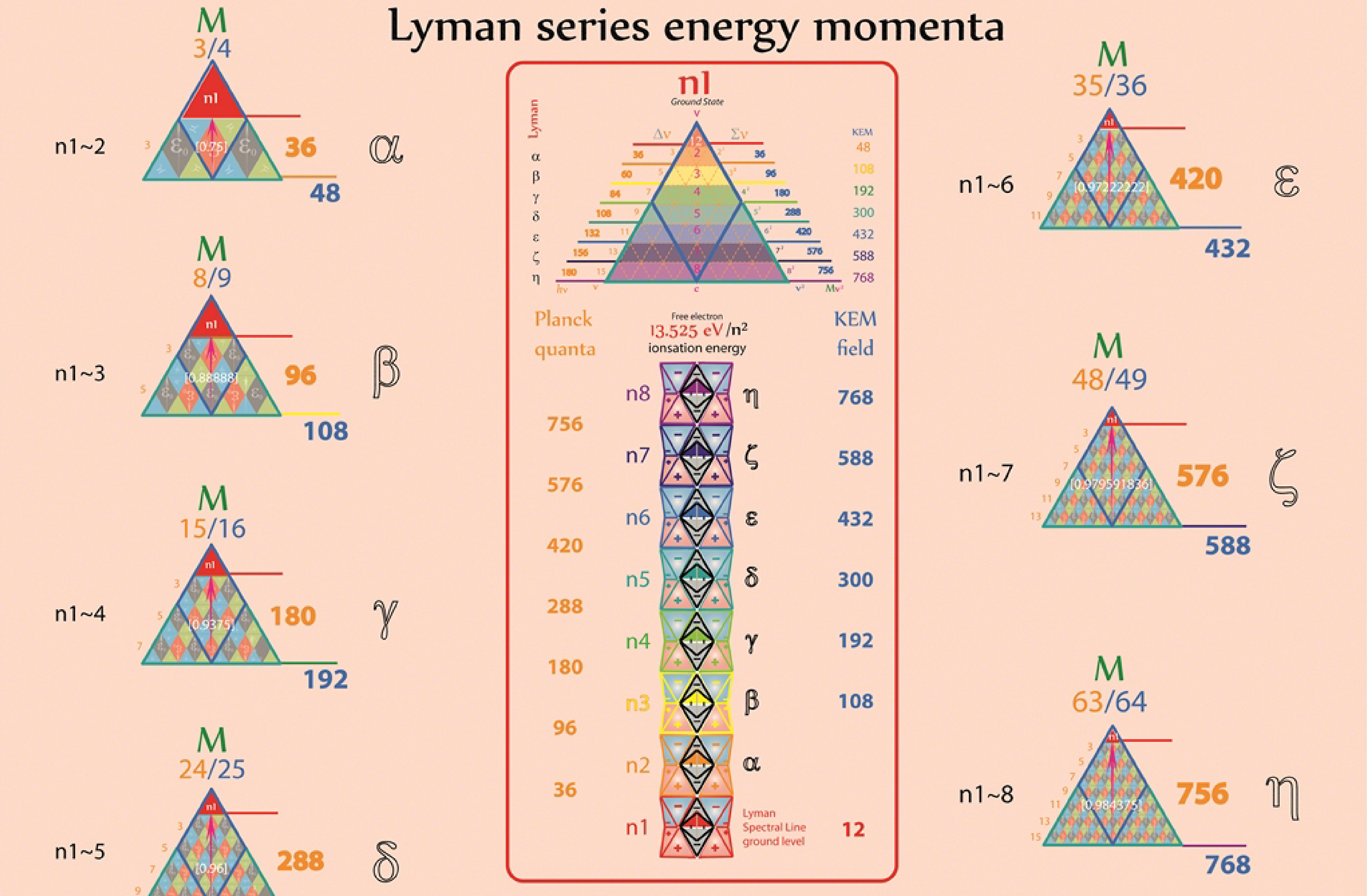
M
3/4
n1~2
nl
[0.75]
36
48
α
M
8/9
n1~3
nl
[0.88888]
96
108
β
M
15/16
n1~4
nl
[0.9375]
180
192
γ
M
24/25
n1~5
nl
[0.96]
288
300
δ
nl
Ground State
Lyman
α
β
γ
δ
ε
ζ
η
Δv
Σv
KEM
48
108
192
300
432
588
768
Planck
quanta
Free electron
13.525 eV /n²
ionsation energy
KEM
field
n8
η
768
756
n7
ζ
588
576
n6
ε
432
420
n5
δ
300
288
n4
γ
192
180
n3
β
108
96
n2
α
36
n1
Lyman
Spectral Line
ground level
12
M
35/36
n1~6
nl
[0.97222222]
420
432
ε
M
48/49
n1~7
nl
[0.979591836]
576
588
ζ
M
63/64
n1~8
nl
[0.984375]
756
768
η

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 30.03 - Lyman series energy momenta

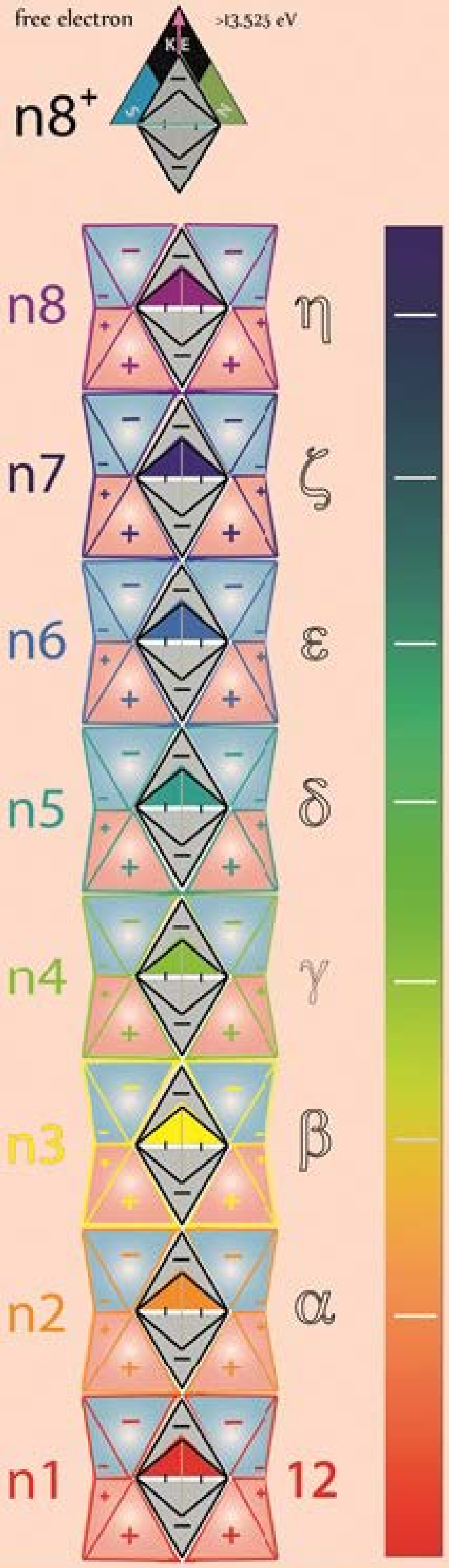

Lyman Series

$$\underset{\text{Tetryonic}}{\Delta h\nu} = \underset{13.525\ \text{eV}}{\text{KEM}} = \underset{\text{Rydberg}}{hcR}$$

63/64
.9843

λ	93.17075598	nm
ṽ	10,732,981.5	m-1
f	3.217666905 x 10^{15}	Hz
E	13.31395504	eV
Ω	27.93188995	m²/s

$R\left(\frac{1}{1}-\frac{1}{64}\right)$ .984375

48/49
.9795

λ	93.62569132	nm
ṽ	10,680,829.01	m-1
f	3.202031983 x 10^{15}	Hz
E	13.24926138	eV
Ω	28.06827613	m²/s

$R\left(\frac{1}{1}-\frac{1}{49}\right)$ .9795

35/36
.9722

λ	94.33539043	nm
ṽ	10,600,475.55	m-1
f	3.177942622 x 10^{15}	Hz
E	13.14958523	eV
Ω	28.28103857	m²/s

$R\left(\frac{1}{1}-\frac{1}{36}\right)$ .9722

24/25
.9600

λ	95.53641971	nm
ṽ	10,467,212.43	m-1
f	3.137991343 x 10^{15}	Hz
E	12.98427616	eV
Ω	28.64199809	m²/s

$R\left(\frac{1}{1}-\frac{1}{25}\right)$ .96

15/16
.9375

λ	97.82929378	nm
ṽ	10,221,5587.14	m-1
f	3.064444671 x 10^{15}	Hz
E	12.67995718	eV
Ω	29.32544445	m²/s

$R\left(\frac{1}{1}-\frac{1}{16}\right)$ .9375

8/9
.8888

λ	103.1793333	nm
ṽ	9,691,863.362	m-1
f	2.90554754 x 10^{15}	Hz
E	12.02247792	eV
Ω	30.93228594	m²/s

$R\left(\frac{1}{1}-\frac{1}{9}\right)$ .888

3/4
.7500

λ	122.2866172	nm
ṽ	8,177,509.712	m-1
f	2.451555737 x 10^{15}	Hz
E	10.14396575	eV
Ω	36.66080536	m²/s

$R\left(\frac{1}{1}-\frac{1}{4}\right)$ .75

ground KEM energies

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	91.71496292	nm	h/p
ṽ	10,903,346.28	m-1	p/h
f	3.268740982 x 10^{15}	Hz	E/h
E	13.52528	eV	hf
Ω	27.49545417	m²/s	

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Copyright ABRAHAM [2008] - all rights reserved

Balmer spectral lines

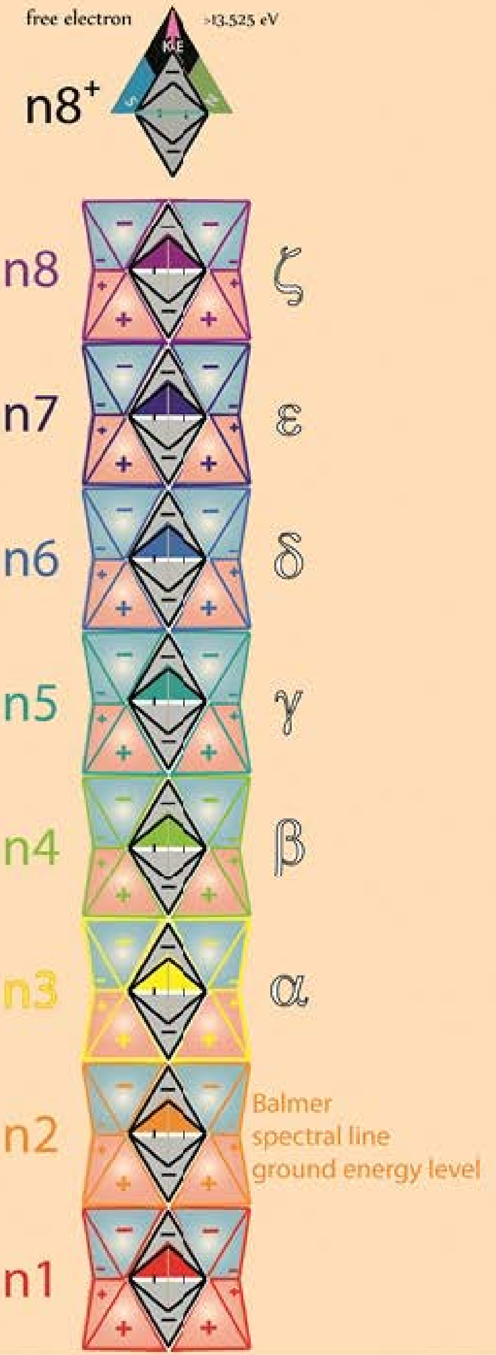
free electron
>13.525 eV
$n8^+$
KE
S
N
n8
ζ
n7
ε
n6
δ
n5
γ
n4
β
n3
α
n2
n1
Balmer
spectral line
ground energy level

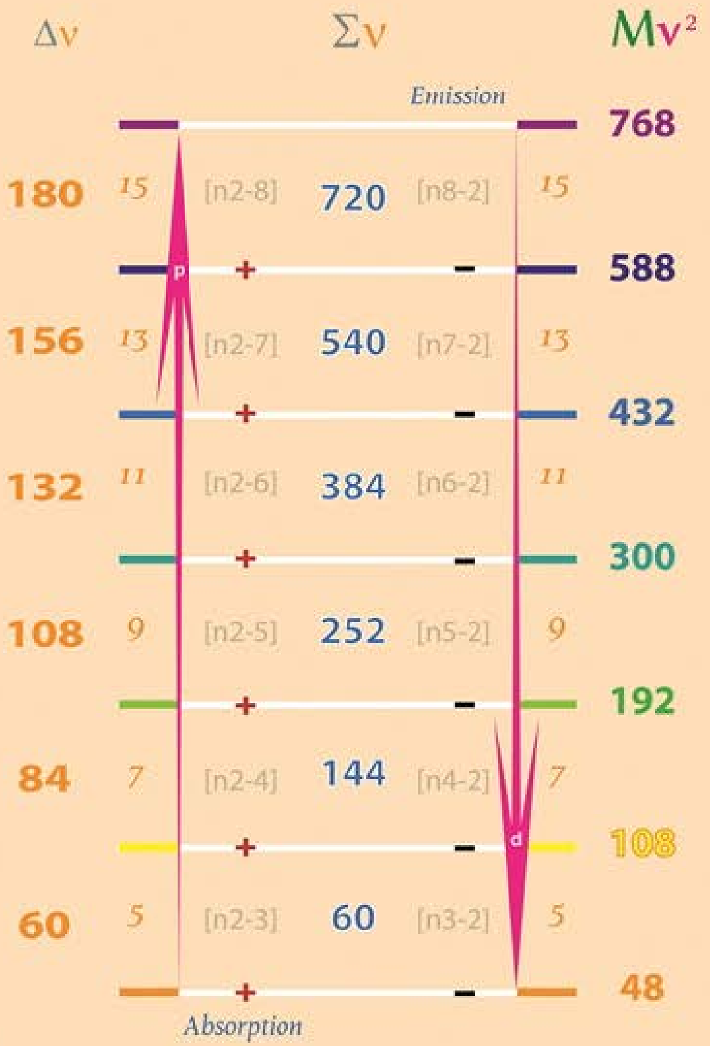
Δv
Σv
Mv^2
Emission
768
180
15
[n2-8]
720
[n8-2]
15
p
+
–
588
156
13
[n2-7]
540
[n7-2]
13
432
132
11
[n2-6]
384
[n6-2]
11
300
108
9
[n2-5]
252
[n5-2]
9
192
84
7
[n2-4]
144
[n4-2]
7
d
108
60
5
[n2-3]
60
[n3-2]
5
48
Absorption

n2

Ground State

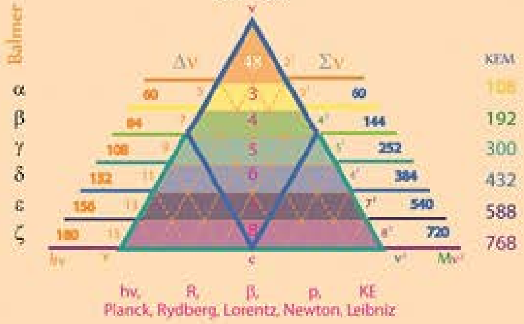
Balmer
α
β
γ
δ
ε
ζ
v
Δv
Σv
48
3
4
5
6
7
8
60
84
108
132
156
180
60
144
252
384
540
720
hv
c
Mv²
KEM
108
192
300
432
588
768
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-3.38125 eV

$$KEM = \frac{hf}{n^2}$$

v
n2
KE
S
N
12
1.2e20

L shell

366.859 nm

$$KEM = \frac{hcR}{n^2}$$

m. Ω/v	λ	366.8598517	nm	h/p
v/Ω	$\tilde{v}$	2,725,836.571	m-1	p/h
c^2/Ω	f	$8.171852456 \times 10^{14}$	Hz	E/h
v^2/c^2	E	3.381321915	eV	hf
	Ω	109.9818167	m^2/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

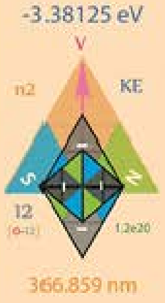

-3.38125 eV
v
n2
KE
12
1.2e20
366.859 nm

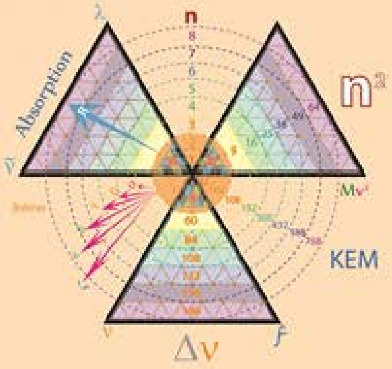

λ
n
8
7
6
5
4
3
Absorption
n²
Mv²
KEM
Δv
v
f

Balmer spectral series

α	660.34	n2-3
β	489.14	n2-4
γ	436.73	n2-5
δ	412.71	n2-6
ε	399.46	n2-7
ϕ	391.31	n2-8

Visible Spectrum

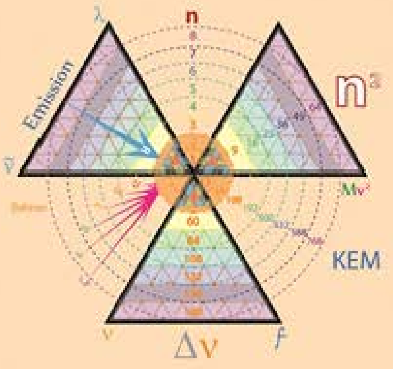

λ
n
Emission
n²
Mv²
KEM
Δv
v
f

L shell

Δv

KEM

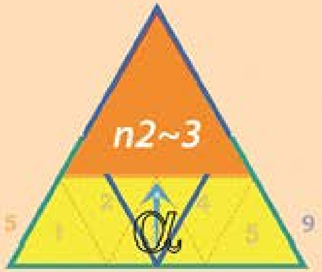

60

108

1 quantum level jump

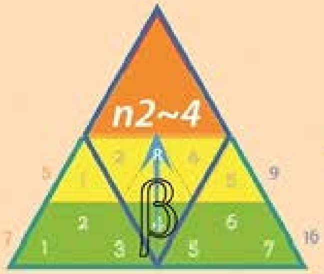

144

192

2 quantum level jump

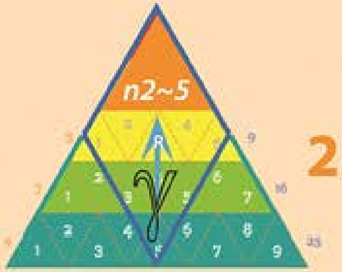

252

300

3 quantum level jump

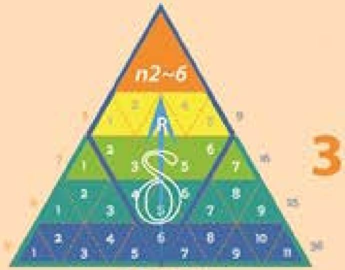

384

432

4 quantum level jump

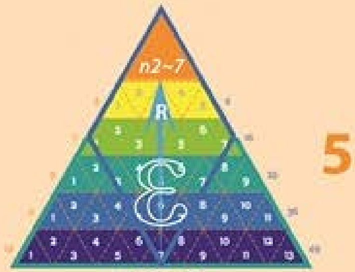

540

588

5 quantum level jump

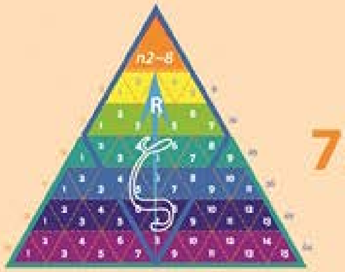

720

768

6 quantum level jump

Copyright ABRAHAM [2008] - all rights reserved

Balmer series energy momenta

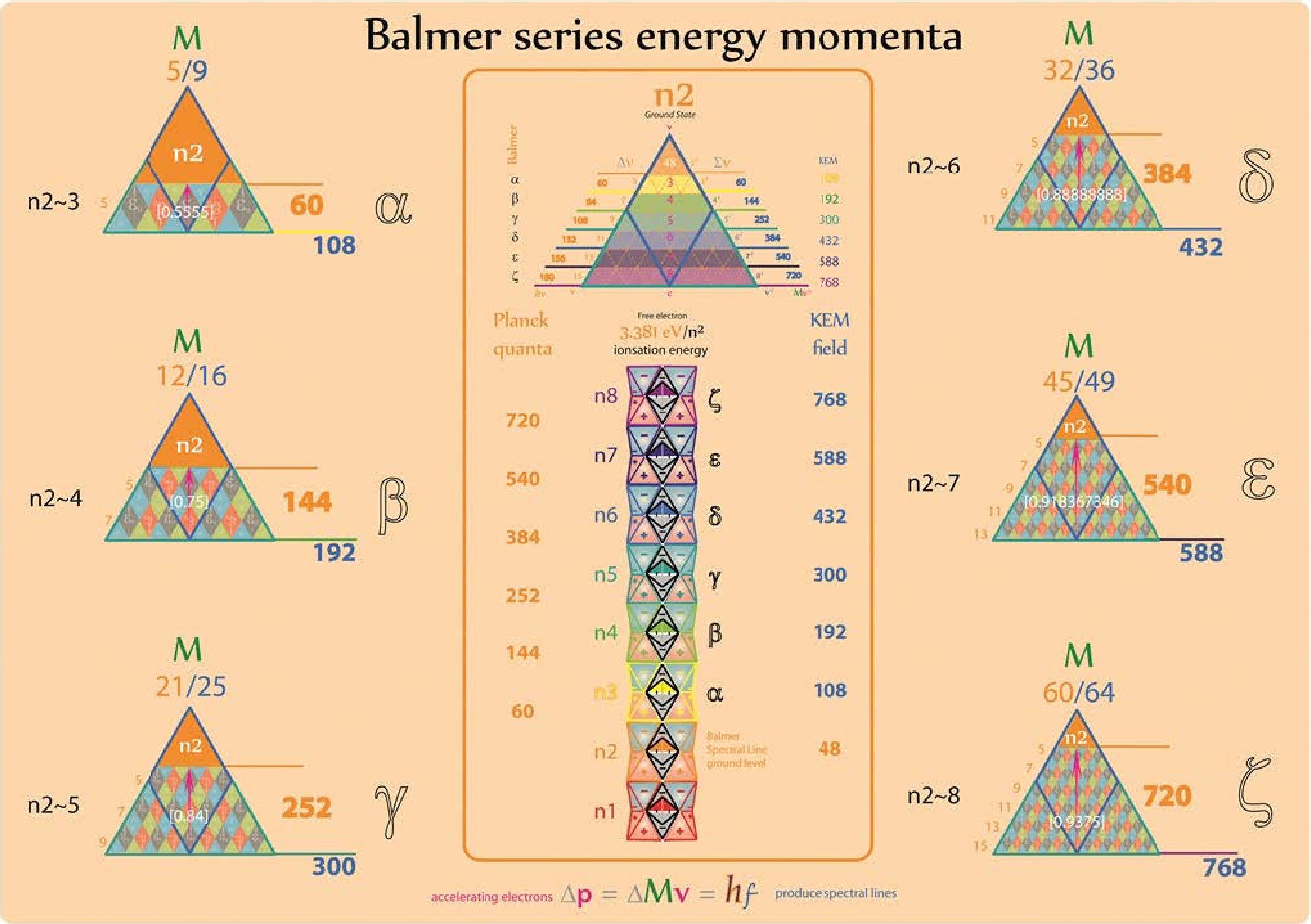
M
5/9
n2
n2~3
[0.5555]
60
108
α
M
12/16
n2
n2~4
[0.75]
144
192
β
M
21/25
n2
n2~5
[0.84]
252
300
γ
n2
Ground State
Balmer
α
β
γ
δ
ε
ζ
Δν
Σν
KEM
60
84
108
132
156
180
60
144
252
384
540
720
108
192
300
432
588
768
Planck
quanta
Free electron
3.381 eV/n²
ionsation energy
KEM
field
720
540
384
252
144
60
n8
n7
n6
n5
n4
n3
n2
n1
ζ
ε
δ
γ
β
α
Balmer
Spectral Line
ground level
768
588
432
300
192
108
48
M
32/36
n2
n2~6
[0.88888888]
384
432
δ
M
45/49
n2
n2~7
[0.918367346]
540
588
ε
M
60/64
n2
n2~8
[0.9375]
720
768
ζ
accelerating electrons $\Delta p = \Delta M v = hf$ produce spectral lines

Tetryonics 31.03 - Balmer series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

Balmer Series

$\Delta h\nu$ = KEM = hcR

Tetryonic · 3.381 eV · Rydberg

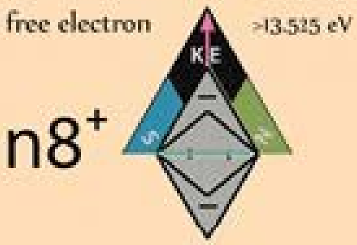

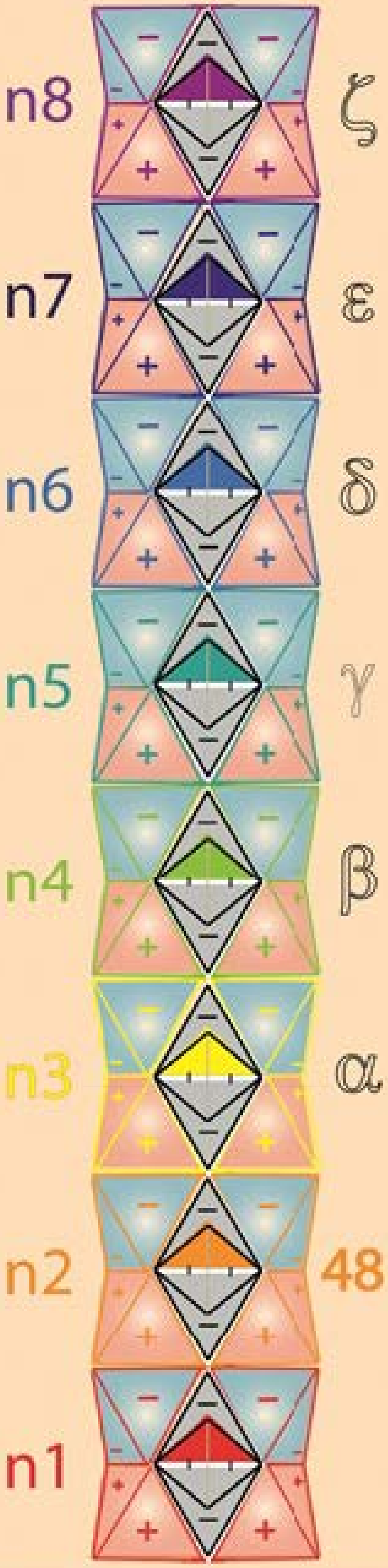

60/64
.2343

λ	391.3171751	nm
ν̄	2,555,471.1785	m-1
ƒ	7.661111677 x 10^{14}	Hz
E	3.169989296	eV
Ω	103.1879511	m²/s

$R\left(\frac{1}{4}-\frac{1}{64}\right)$
.2343

45/49
.2295

λ	399.4696163	nm
ν̄	2,5033,319.299	m-1
ƒ	7.50476246 x 10^{14}	Hz
E	3.105295637	eV
Ω	101.0037092	m²/s

$R\left(\frac{1}{4}-\frac{1}{49}\right)$
.2295

32/36
.2222

λ	412.7173331	nm
ν̄	2,422,965.84	m-1
ƒ	7.26386885 x 10^{14}	Hz
E	3.00561948	eV
Ω	97.76161853	m²/s

$R\left(\frac{1}{4}-\frac{1}{36}\right)$
.2222

21/25
.2100

λ	436.7379187	nm
ν̄	2,289,702.719	m-1
ƒ	6.864356063 x 10^{14}	Hz
E	2.840310409	eV
Ω	92.38472691	m²/s

$R\left(\frac{1}{4}-\frac{1}{25}\right)$
.21

12/16
.1875

λ	489.1464659	nm
ν̄	2,044,377.428	m-1
ƒ	6.128889542 x 10^{14}	Hz
E	2.535991437	eV
Ω	82.48636251	m²/s

$R\left(\frac{1}{4}-\frac{1}{16}\right)$
.1875

5/9
.138

λ	660.347733	nm
ν̄	1,514,353.65	m-1
ƒ	4.539918031 x 10^{14}	Hz
E	1.878512175	eV
Ω	61.10109927	m²/s

$R\left(\frac{1}{4}-\frac{1}{9}\right)$
.138

m.

Ω/v	λ	366.8598517	nm	h/p
v/Ω	ν̄	2,725,836.571	m-1	p/h
c²/Ω	ƒ	8.171852456 x 10^{14}	Hz	E/h
v²/c²	E	3.381321915	eV	hf
	Ω	109.9818167	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Paschen spectral lines

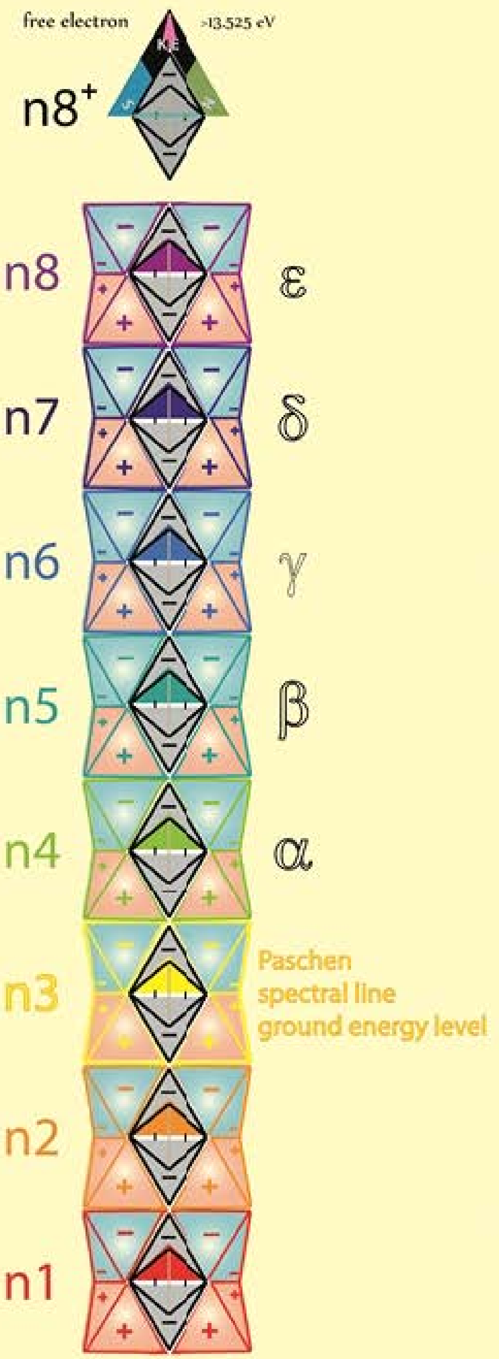
free electron
>13.525 eV
KE
$n8^+$
n8
ε
n7
δ
n6
γ
n5
β
n4
α
n3
Paschen
spectral line
ground energy level
n2
n1

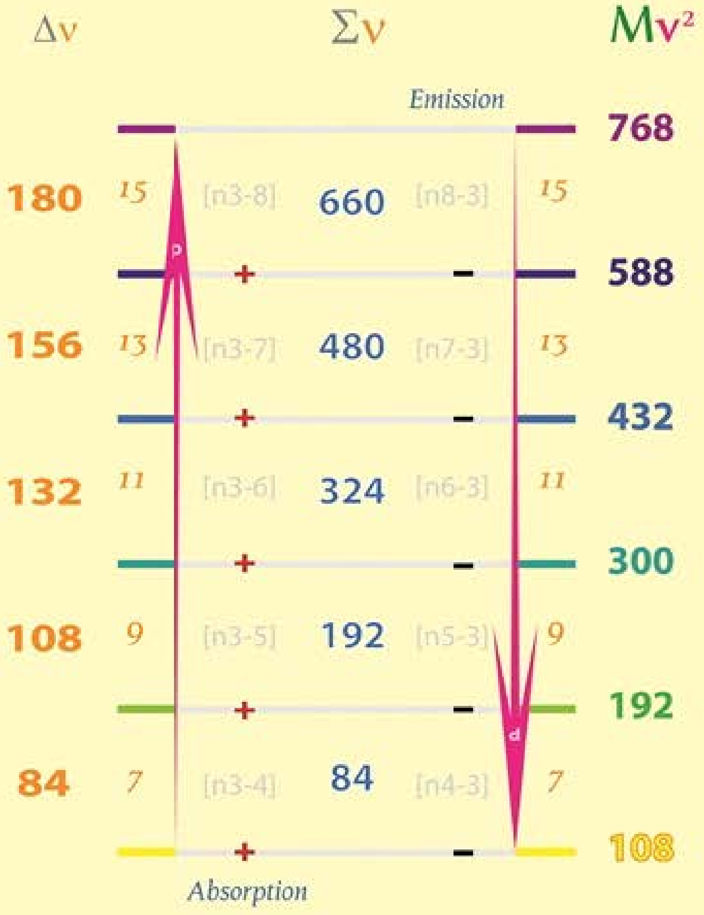
Δv
Σv
Mv²
Emission
768
180
15
[n3-8]
660
[n8-3]
15
588
156
13
[n3-7]
480
[n7-3]
13
432
132
11
[n3-6]
324
[n6-3]
11
300
108
9
[n3-5]
192
[n5-3]
9
192
84
7
[n3-4]
84
[n4-3]
7
108
Absorption

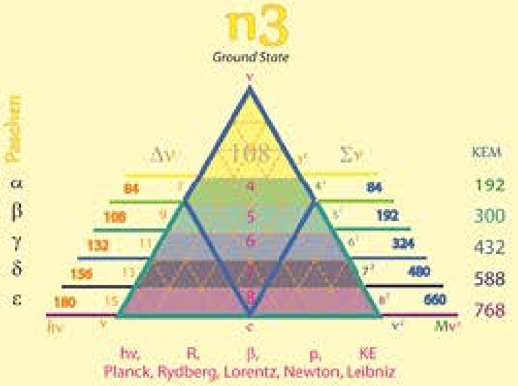
n3
Ground State
Paschen
α
β
γ
δ
ε
Δv
108
Σv
KEM
84
192
108
192
300
132
324
432
156
480
588
180
660
768
hv
Mv²
hw, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-1.50280974 eV

$$KEM = \frac{hf}{n^2}$$

v
n3
KE
12
1.2e20

M shell

825.434 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix}\Omega/v\\ v/\Omega\\ c^2/\Omega\\ v^2/c^2\end{bmatrix}$$

λ	825.4346663	nm	h/p
$\tilde{v}$	1,211,482.92	m-1	p/h
f	$3.631934425 \times 10^{14}$	Hz	E/h
E	1.50280974	eV	hf
Ω	247.4590875	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

Paschen spectral series

-1.50280974 eV
v
n3
KE
12
1.2e20
825.434 nm

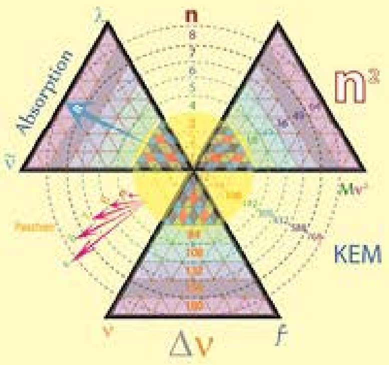
λ
n
Absorption
n²
Mv²
Paschen
KEM
v
Δv
f

α	1,886	n3-4
β	1,289	n3-5
γ	1,100	n3-6
δ	1,011	n3-7
ε	960	n3-8

IR spectrum

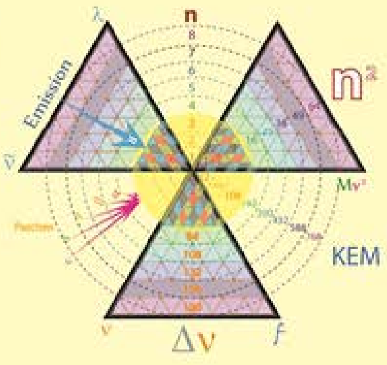
λ
n
Emission
n²
Mv²
Paschen
KEM
v
Δv
f

M shell

Δv
KEM

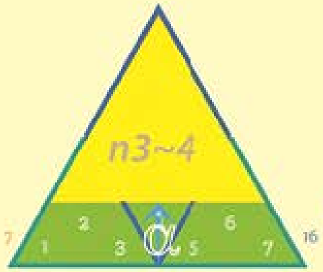

84

192

1 quantum level jump

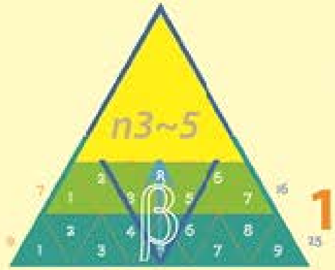

192

300

2 quantum level jump

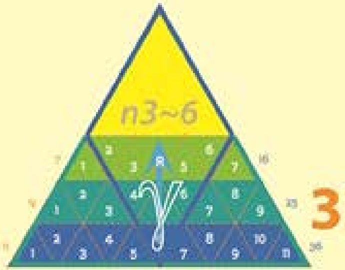

324

432

3 quantum level jump

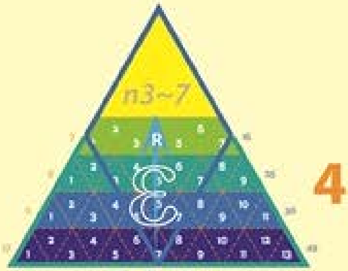

480

588

4 quantum level jump

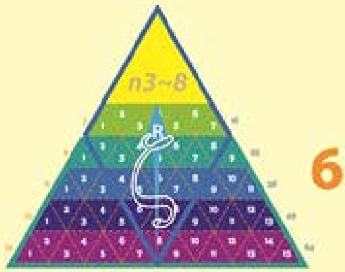

660

768

5 quantum level jump

Tetryonics 32.02 - Paschen quantum transitions

Paschen series energy momenta

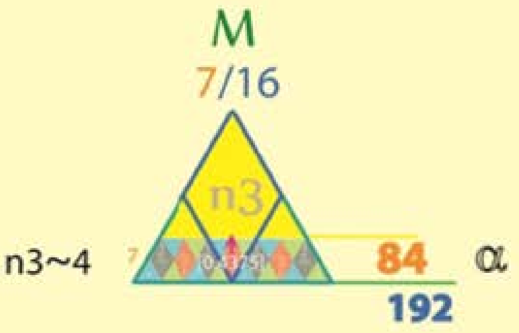
M
7/16
n3
n3~4
[0.4375]
84
α
192

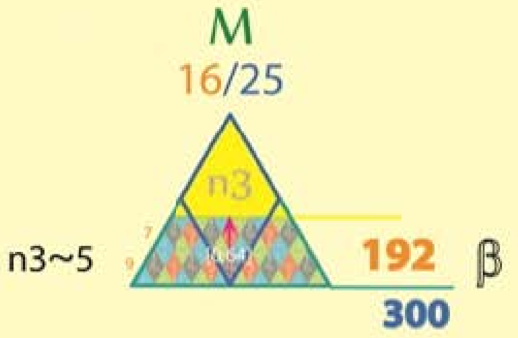
M
16/25
n3
n3~5
[0.64]
192
β
300

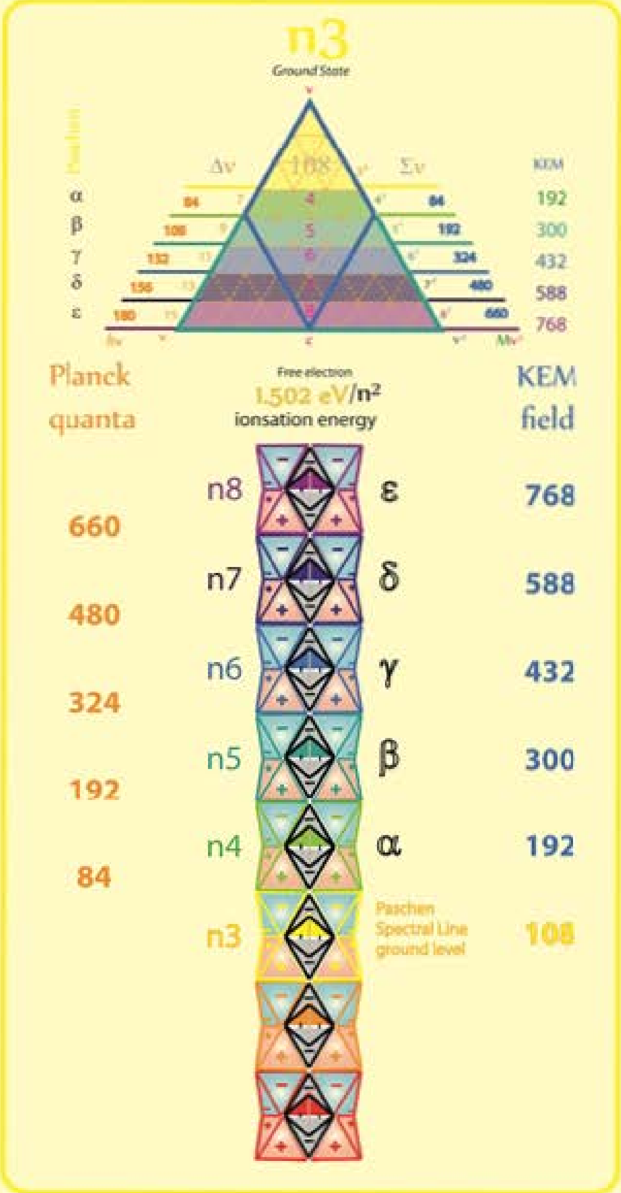
n3
Ground State
Δv
108
Σv
KEM
α
β
γ
δ
ε
84
108
132
156
180
84
192
324
480
660
192
300
432
588
768
Planck quanta
Free electron
1.502 eV/n²
ionsation energy
KEM field
n8
ε
768
660
n7
δ
588
480
n6
γ
432
324
n5
β
300
192
n4
α
192
84
n3
Paschen Spectral Line ground level
108

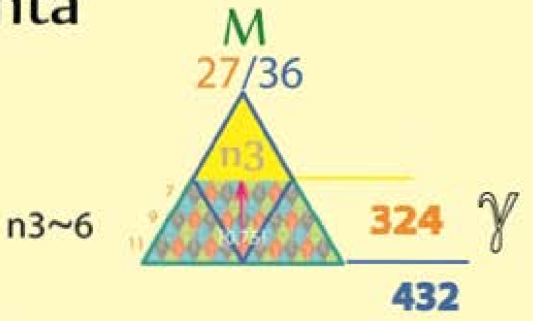
M
27/36
n3
n3~6
[0.75]
324
γ
432

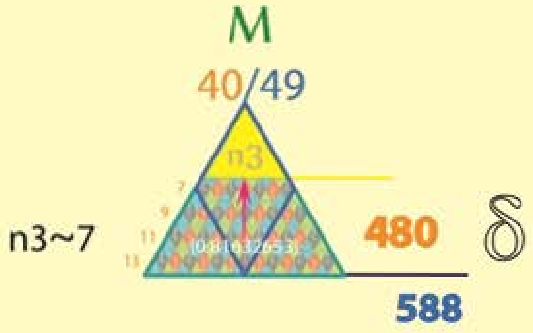
M
40/49
n3
n3~7
[0.81632653]
480
δ
588

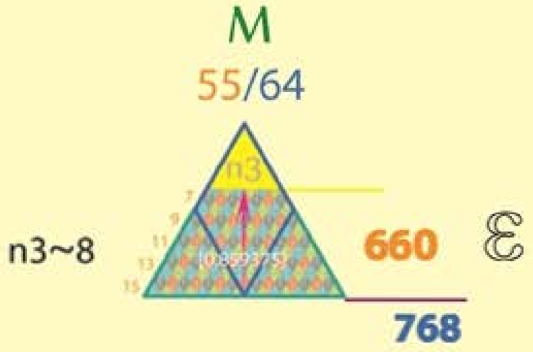
M
55/64
n3
n3~8
[0.859375]
660
ε
768

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 32.03 - Paschen series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

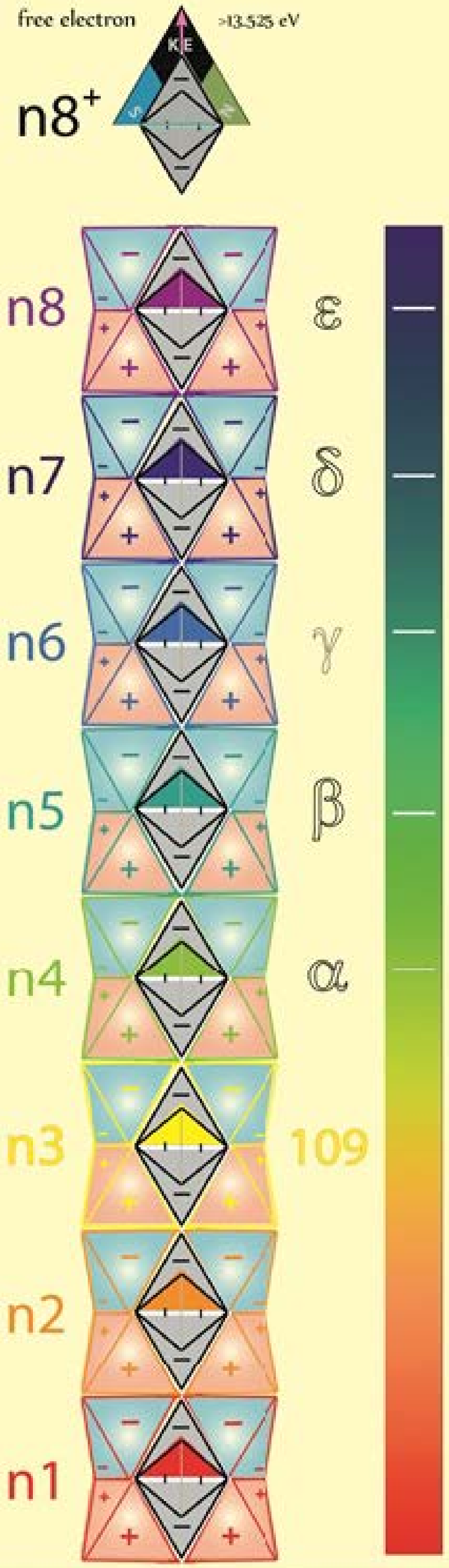

Paschen Series

$\Delta h\nu$ = KEM = hcR

Tetryonic — 1.502 eV — Rydberg

55/64
.0954

λ	960.5057935	nm
$\tilde{\nu}$	1,041,118.135	m-1
f	$3.121193646 \times 10^{14}$	Hz
E	1.29147712	eV
Ω	287.9523928	m²/s

$R\left(\frac{1}{9}-\frac{1}{64}\right)$
.09548

40/49
.0907

λ	1,011.157466	nm
$\tilde{\nu}$	988,965.6492	m-1
f	$2.9464844428 \times 10^{14}$	Hz
E	1.226783461	eV
Ω	303.1373822	m²/s

$R\left(\frac{1}{9}-\frac{1}{49}\right)$
.0907

27/36
.8033

λ	1,100.579555	nm
$\tilde{\nu}$	908,612.1902	m-1
f	$2.723950819 \times 10^{14}$	Hz
E	1.127107305	eV
Ω	329.94545	m²/s

$R\left(\frac{1}{9}-\frac{1}{36}\right)$
.0833

16/25
.0711

λ	1,289.741666	nm
$\tilde{\nu}$	775,349.069	m-1
f	$2.324438032 \times 10^{14}$	Hz
E	0.961798233	eV
Ω	386.6548243	m²/s

$R\left(\frac{1}{9}-\frac{1}{25}\right)$
.0711

7/16
.0486

λ	1,886.707809	nm
$\tilde{\nu}$	530,023.7776	m-1
f	$1.588971311 \times 10^{14}$	Hz
E	0.657479261	eV
Ω	565.6207715	m²/s

$R\left(\frac{1}{9}-\frac{1}{16}\right)$
.0486

$$m.\begin{bmatrix}\Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2\end{bmatrix}$$

λ	825.4346663	nm	h/p
$\tilde{\nu}$	1,211,482.92	m-1	p/h
f	$3.631934425 \times 10^{14}$	Hz	E/h
E	1.50280974	eV	hf
Ω	247.4590875	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Brackett spectral lines

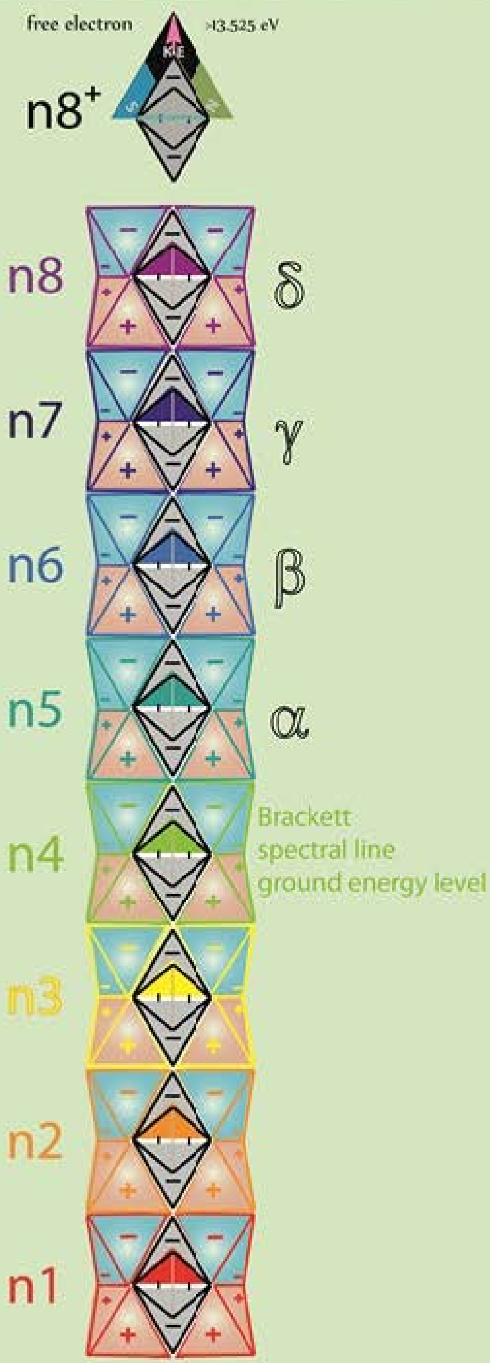
free electron
>13.525 eV
KE
n8⁺
n8
δ
n7
γ
n6
β
n5
α
n4
Brackett
spectral line
ground energy level
n3
n2
n1

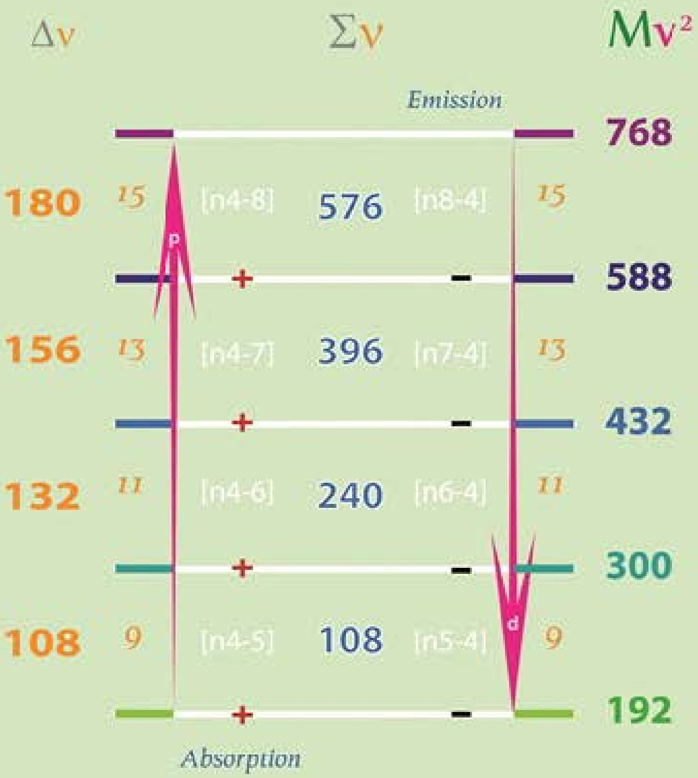
Δv
Σv
Mv²
Emission
768
180
15
[n4-8]
576
[n8-4]
15
p
588
156
13
[n4-7]
396
[n7-4]
13
432
132
11
[n4-6]
240
[n6-4]
11
300
108
9
[n4-5]
108
[n5-4]
9
d
192
Absorption

n4

Ground State

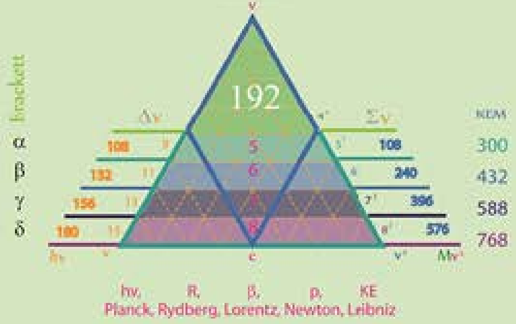
Brackett
α
β
γ
δ
v
192
Δv
Σv
KEM
108
132
156
180
5
6
7
8
108
240
396
576
300
432
588
768
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.84533478 eV

$$KEM = \frac{hf}{n^2}$$

v
n4
KE
S
N
12
1.2e20

N shell

1,467.43 nm

$$KEM = \frac{hcR}{n^2}$$

m.	Ω/v	λ	1,467.439407	nm	h/p
	v/Ω	$\tilde{v}$	681,459.1426	m-1	p/h
	c²/Ω	f	$2.042963114 \times 10^{14}$	Hz	E/h
	v²/c²	E	0.84533478	eV	hf
		Ω	[illegible]	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Tetryonics 33.01 - Brackett spectral lines

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhν] W bosons

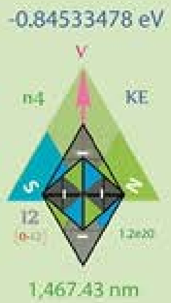
-0.84533478 eV
v
n4
KE
S
N
12
1.2e20
1,467.43 nm

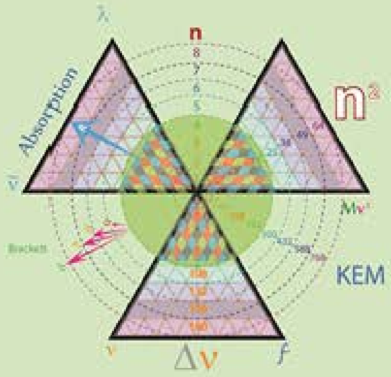
λ
n
Absorption
n²
Mv²
Brackett
KEM
v
Δν
f

Brackett spectral series

α	4,076	n4-5
β	2,641	n4-6
γ	2,178	n4-7
δ	1,956	n4-8
	IR spectrum	

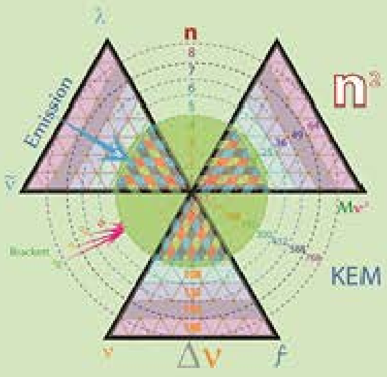
λ
n
Emission
n³
Mv²
Brackett
KEM
v
Δν
f

N shell

$$\frac{\Delta\nu}{\text{KEM}}$$

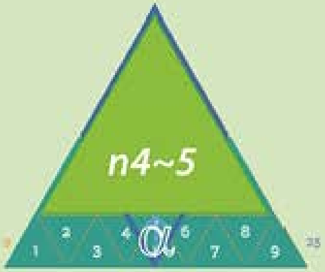

108

300

1 quantum level jump

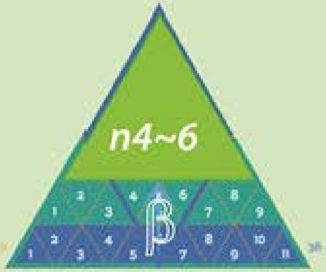

240

432

2 quantum level jump

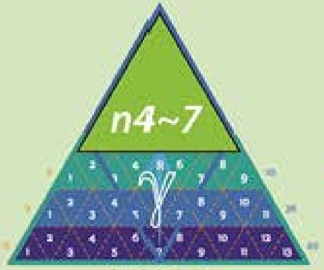

396

588

3 quantum level jump

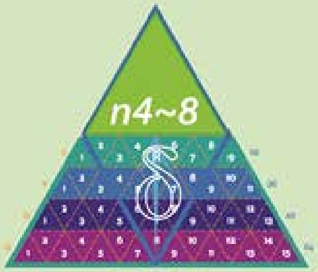

576

768

4 quantum level jump

Tetryonics 33.02 - Brackett quantum transitions

Brackett series energy momenta

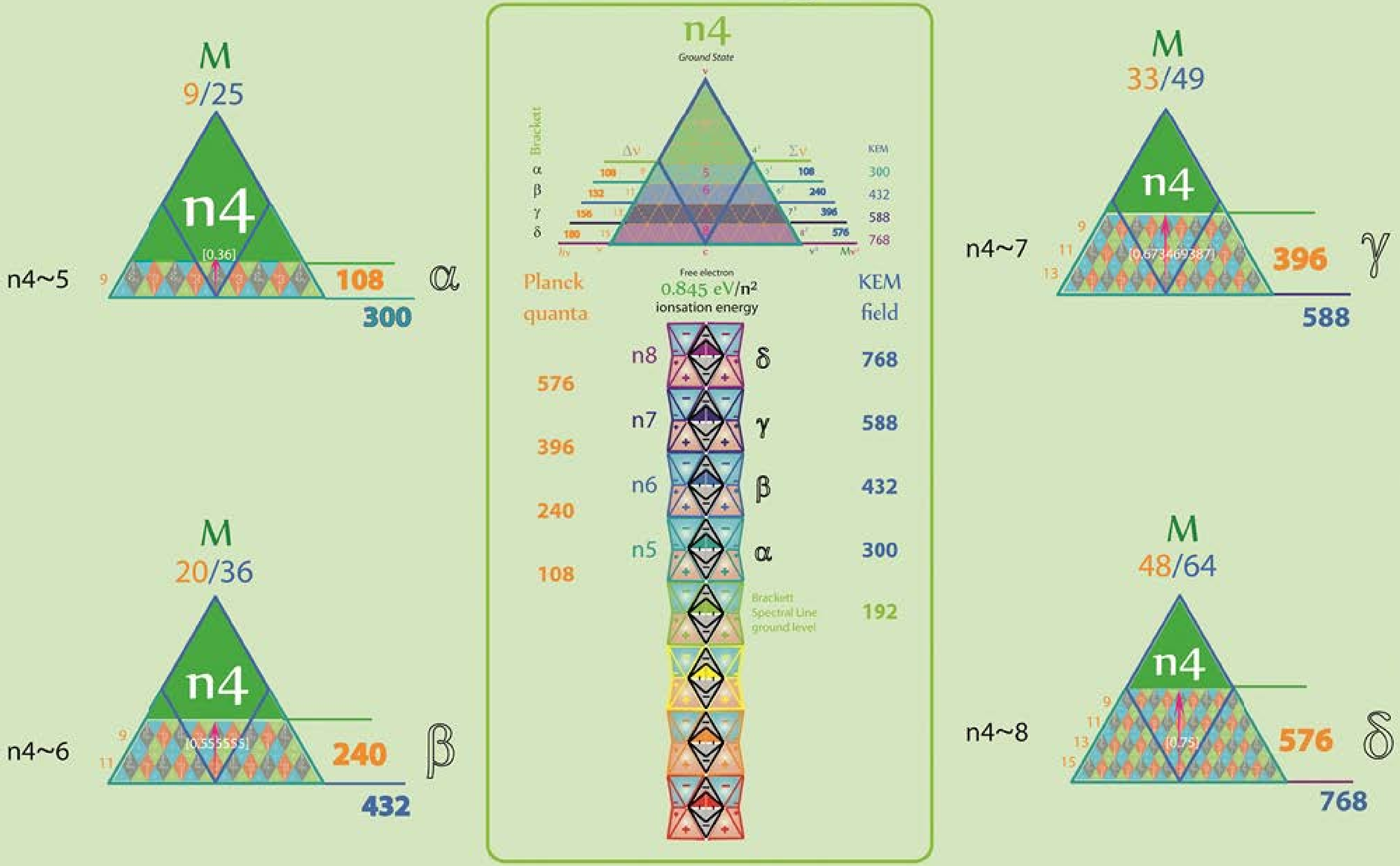
M
9/25
n4
[0.36]
n4~5
9
108
300
α
M
20/36
n4
[0.555555]
n4~6
9
11
240
432
β
n4
Ground State
Brackett
α
β
γ
δ
Δv
Σv
KEM
108
132
156
180
108
240
396
576
300
432
588
768
Free electron
0.845 eV/n²
ionsation energy
M
33/49
n4
[0.673469387]
n4~7
9
11
13
396
588
γ
M
48/64
n4
[0.75]
n4~8
9
11
13
15
576
768
δ

Planck quanta			KEM field
	n8	δ	768
576			
	n7	γ	588
396			
	n6	β	432
240			
	n5	α	300
108			
		Brackett Spectral Line ground level	192

accelerating electrons $\Delta p = \Delta M v = hf$ produce spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Brackett Series

$\Delta h v$ = KEM = hcR

Tetryonic — 0.845 eV — Rydberg

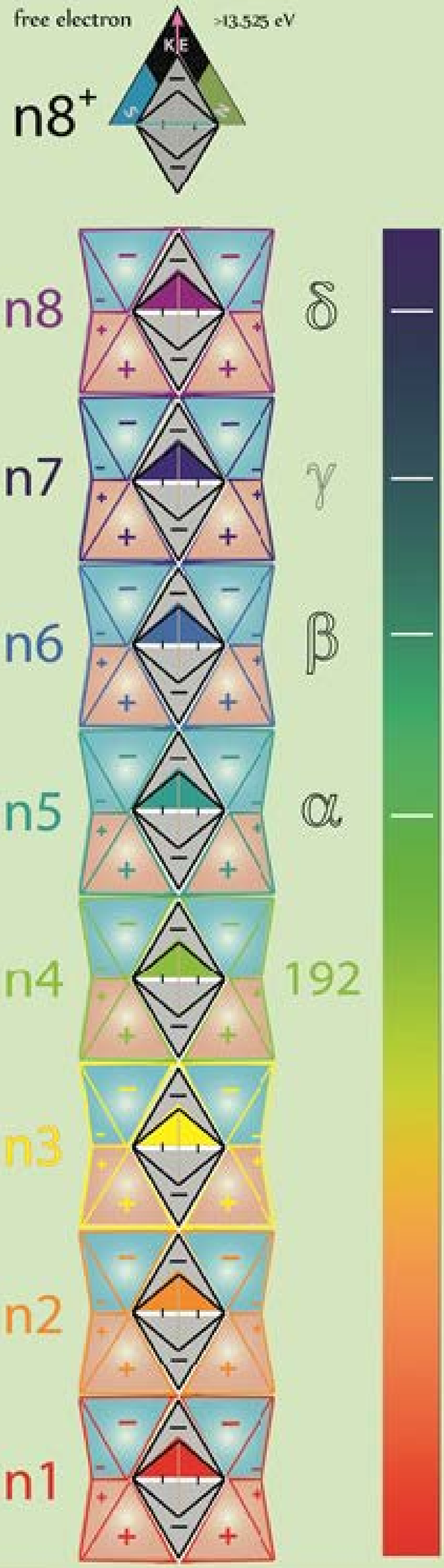

48/64
.0468

λ	1,956.585876	nm
ν̃	511,094.357	m-1
ƒ	1.532222335 x 10^{14}	Hz
E	0.634001085	eV
Ω	586.569689	m²/s

$R\left(\frac{1}{16}-\frac{1}{64}\right)$ .0468

33/49
.04209

λ	2,178.92518	nm
ν̃	458,941.8716	m-1
ƒ	1.375873118 x 10^{14}	Hz
E	0.569307096	eV
Ω	653.2253354	m²/s

$R\left(\frac{1}{16}-\frac{1}{49}\right)$ .04209

20/36
.0347

λ	2,641.390932	nm
ν̃	378,588.4126	m-1
ƒ	1.134979508 x 10^{14}	Hz
E	0.469630433	eV
Ω	791.36690801	m²/s

$R\left(\frac{1}{16}-\frac{1}{36}\right)$ .0347

9/25
.0225

λ	4,076.220574	nm
ν̃	245,325.2914	m-1
ƒ	7.35466721x 10^{13}	Hz
E	0.30432052	eV
Ω	1,222.020185	m²/s

$R\left(\frac{1}{16}-\frac{1}{25}\right)$ .0225

$m.\begin{bmatrix}\Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2\end{bmatrix}$

λ	1,467.439407	nm	h/p
ν̃	681,459.1426	m-1	p/h
ƒ	2.042963114 x 10^{14}	Hz	E/h
E	0.84533478	eV	hf
Ω	439.9272667	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

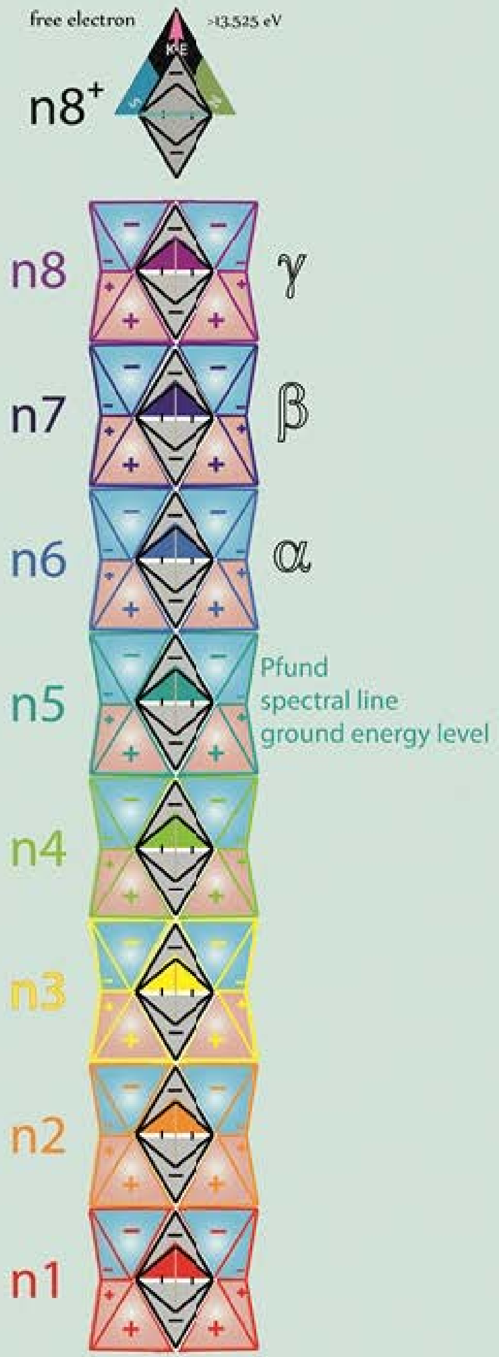
free electron
>13.525 eV
KE
n8+
n8
γ
n7
β
n6
α
n5
Pfund
spectral line
ground energy level
n4
n3
n2
n1

Pfund spectral lines

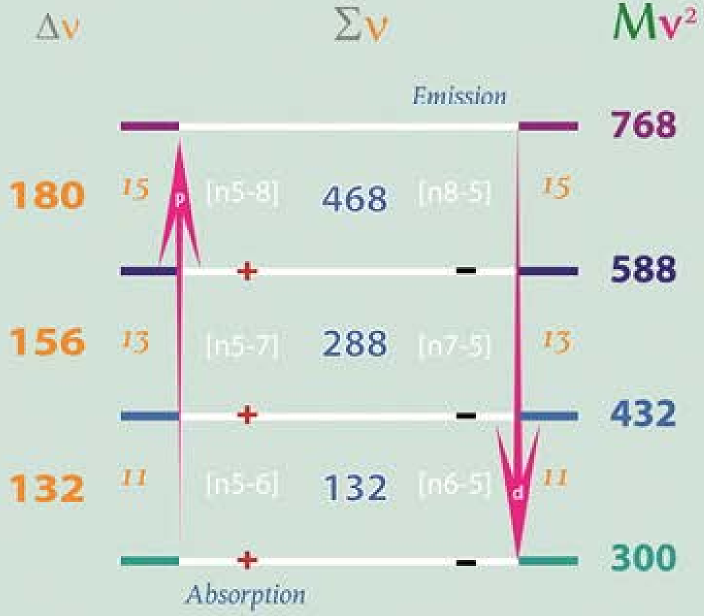
Δv
Σv
Mv²
Emission
768
180
15
p
[n5-8]
468
[n8-5]
15
588
156
13
[n5-7]
288
[n7-5]
13
432
132
11
[n5-6]
132
[n6-5]
d
11
300
Absorption

n5

Ground State

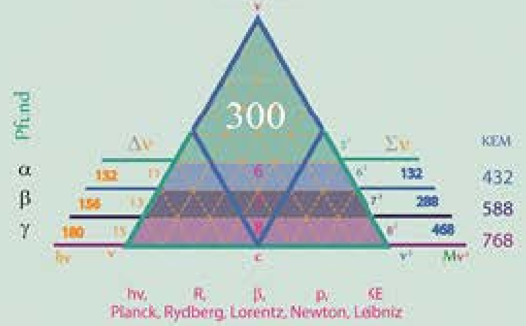
Pfund
300
Δv
Σv
KEM
α
β
γ
132
156
180
11
13
15
6
132
288
468
432
588
768
hv
v
c
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.5410115 eV

$$KEM = \frac{hf}{n^2}$$

v
n5
KE
S
N
12
1.2e20

O shell

2,292.87 nm

$$KEM = \frac{hcR}{n^2}$$

m. [Ω/v, v/Ω, c²/Ω, v²/c²]

λ	2,292.874073	nm	h/p
ṽ	436,133.8513	m-1	p/h
ƒ	1.307496393 x 10^{14}	Hz	E/h
E	0.541011506	eV	hƒ
Ω	687.3863542	m²/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave-functions

Tetryonics 34.01 - Pfund spectral lines

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

2,292.87 nm

v
n5
KE
12
[o-12]
1.2e20

-0.5410115 eV

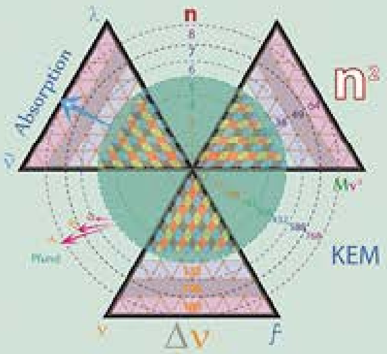
λ
n
Absorption
n²
Mv²
KEM
Pfund
v
Δv
f

Pfund spectral series

α	7,503	n5-6
β	4,681	n5-7
γ	3,762	n5-8

IR spectrum

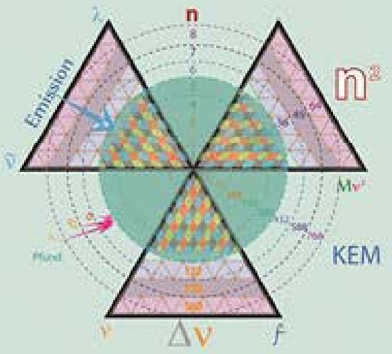
λ
n
Emission
n²
Mv²
KEM
Pfund
v
Δv
f

O shell

Δv

KEM

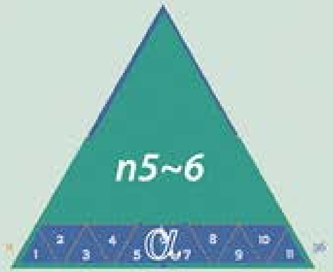

132

432

1 quantum level jump

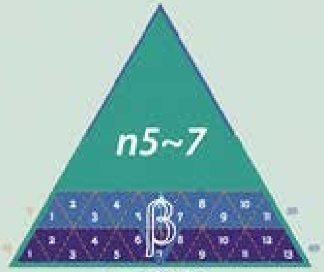

288

588

2 quantum level jump

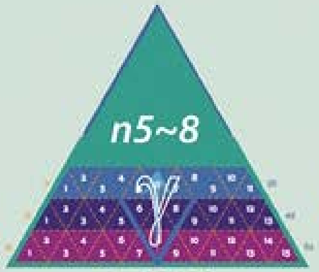

468

768

3 quantum level jump

Tetryonics 34.02 - Pfund quantum transitions

Pfund series energy momenta

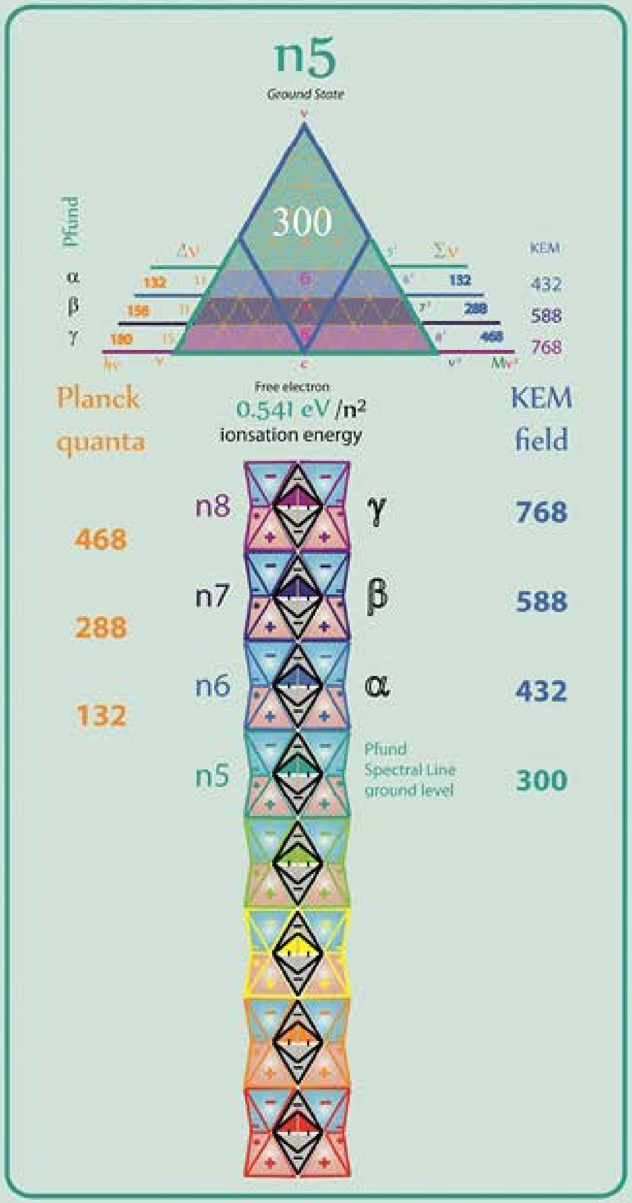
n5
Ground State
v
300
Pfund
Δv
Σv
KEM
α
β
γ
132
156
180
11
13
15
6
7
8
132
288
468
432
588
768
hv
v
c
v²
Mv²
Planck
quanta
Free electron
0.541 eV /n²
ionsation energy
KEM
field
n8
γ
768
468
n7
β
588
288
n6
α
432
132
n5
Pfund
Spectral Line
ground level
300

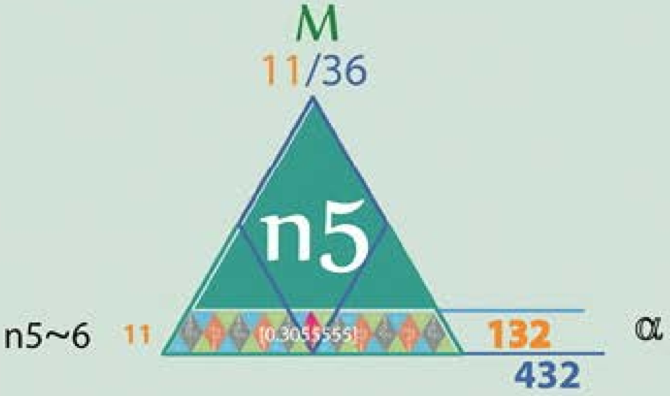
M
11/36
n5
n5~6
11
[0.30555555]
132
432
α

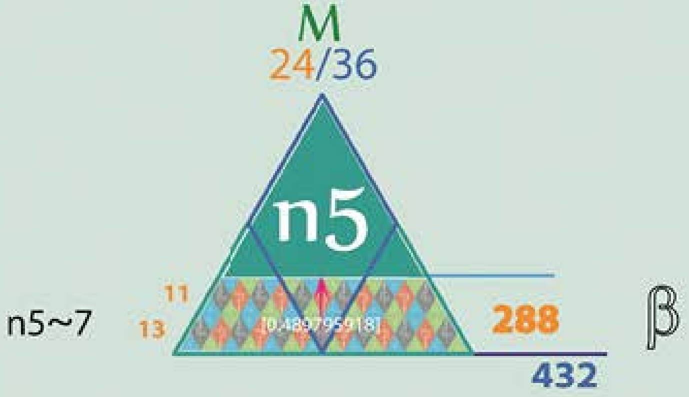
M
24/36
n5
n5~7
11
13
[0.489795918]
288
432
β

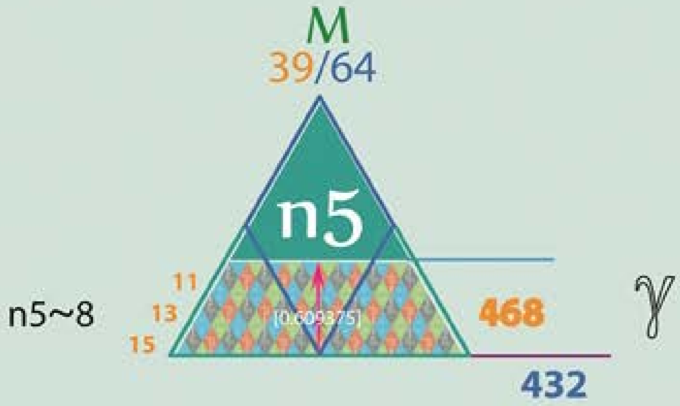
M
39/64
n5
n5~8
11
13
15
[0.609375]
468
432
γ

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 34.03 - Pfund series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

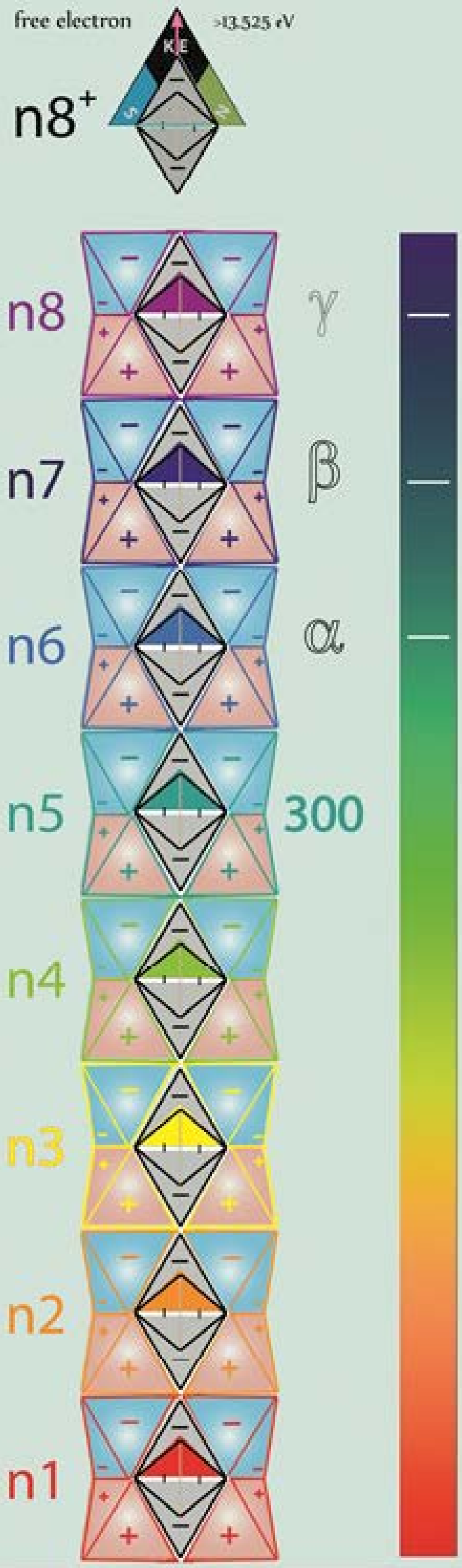

Pfund Series

$\Delta h\nu$ = KEM = hcR

Tetryonic

0.541 eV

Rydberg

39/64

.0243

λ	3,762.665146	nm
ṽ	265,769.0656	m-1
f	7.967556145 x 10^{13}	Hz
E	0.329678886	eV
Ω	1,128.018633	m²/s

$R\left(\frac{1}{25}-\frac{1}{64}\right)$

.0243

24/49

.0195

λ	4,681.284566	nm
ṽ	213,616.5802	m-1
f	6.404063965 x 10^{13}	Hz
E	0.264985227	eV
Ω	1,403.413807	m²/s

$R\left(\frac{1}{25}-\frac{1}{49}\right)$

.01959

11/36

.0122

λ	7,503.951512	nm
ṽ	133,263.1212	m-1
f	3.995127867 x 10^{13}	Hz
E	0.165309071	eV
Ω	2,249.628068	m²/s

$R\left(\frac{1}{25}-\frac{1}{36}\right)$

.0122

$$m.\begin{bmatrix} \Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2 \end{bmatrix}$$

λ	2,292.874073	nm	h/p
ṽ	436,133.8513	m-1	p/h
f	1.307496393 x 10^{14}	Hz	E/h
E	0.541011506	eV	hf
Ω	687.3863542	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Humphreys spectral lines

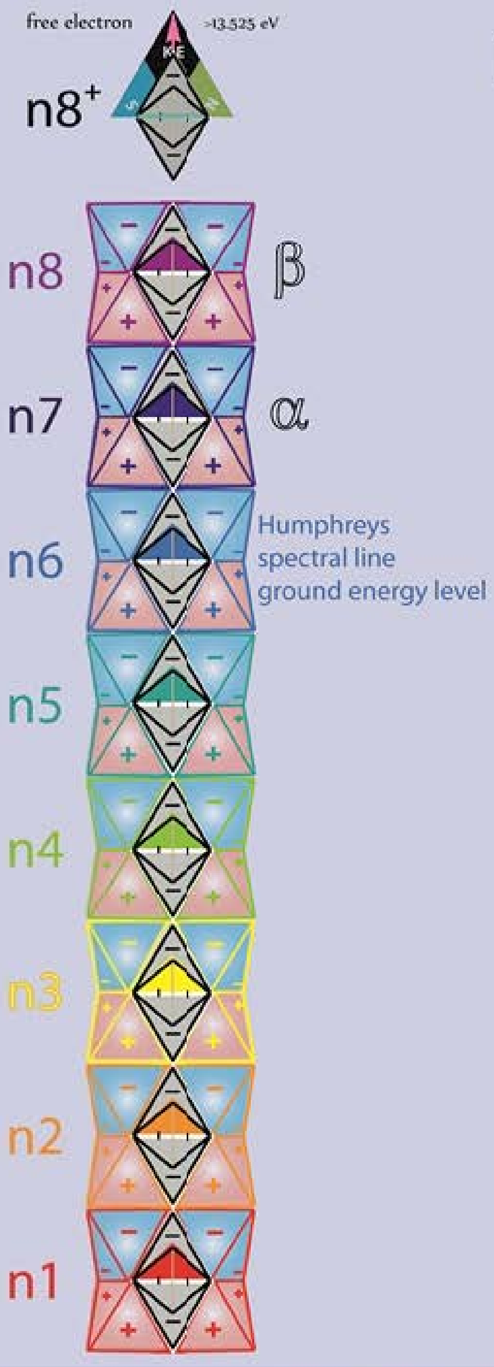
free electron
>13.525 eV
KE
n8+
n8
β
n7
α
n6
Humphreys
spectral line
ground energy level
n5
n4
n3
n2
n1

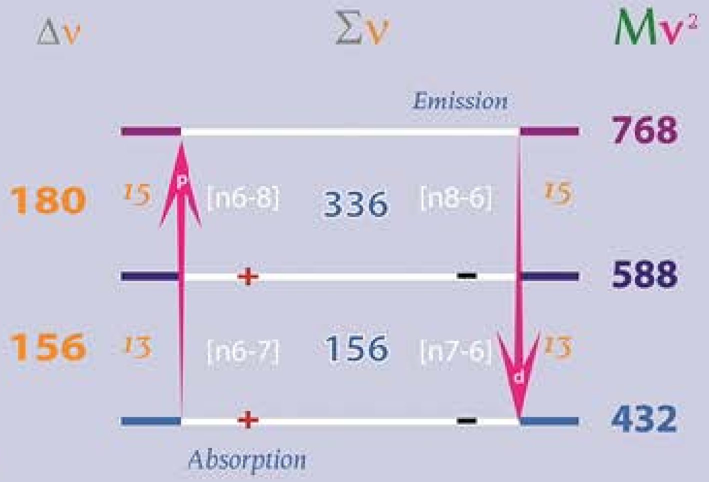
Δv
Σv
Mv²
Emission
768
180
15
p
[n6-8]
336
[n8-6]
15
588
156
13
[n6-7]
156
[n7-6]
d
13
432
Absorption

n6

Ground State

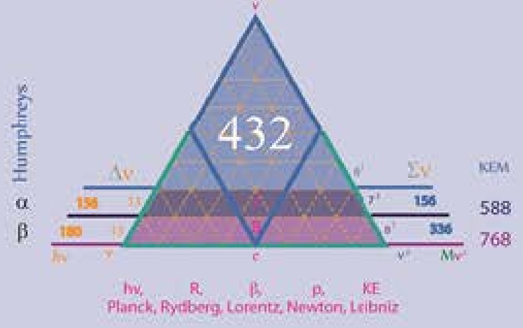
v
432
Humphreys
α
β
Δv
Σv
KEM
156
13
7²
156
588
180
13
8²
336
768
hv
v
c
v²
Mv²
hv,
R,
β,
p,
KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.375702435 eV

$$KEM = \frac{hf}{n^2}$$

v
n6
KE
S
N
12
1.2e20

P shell

3,301.73 nm

$$KEM = \frac{hcR}{n^2}$$

$$m.\begin{bmatrix} \Omega/v \\ v/\Omega \\ c^2/\Omega \\ v^2/c^2 \end{bmatrix}$$

λ	3,301.738665	nm	h/p
$\tilde{v}$	302,870.7301	m-1	p/h
f	$9.079836062 \times 10^{13}$	Hz	E/h
E	0.375702435	eV	hf
Ω	989.8363501	m^2/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave functions

Tetryonics 35.01 - Humphreys spectral lines

Copyright ABRAHAM [2008] - all rights reserved

As photo-electrons transition between Mv squared KEM field energy levels in atomic nuclei they release specific [Δhv] W bosons

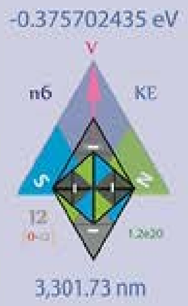

-0.375702435 eV
v
n6
KE
12
1.2e20
3,301.73 nm

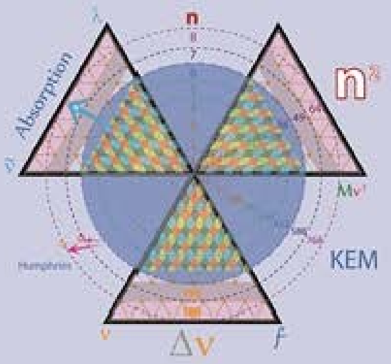

λ
n
Absorption
n²
Mv²
KEM
Humphries
v
Δv
f

Humphries spectral series

α	12,445	n6-7
β	7,546	n6-8
	IR spectrum	

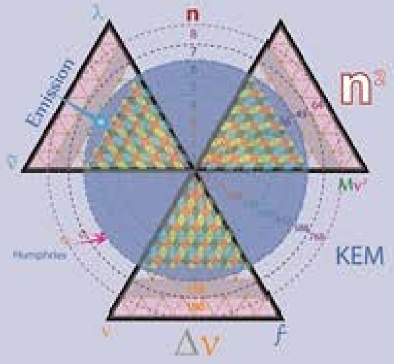

λ
n
Emission
n²
Mv²
KEM
Humphries
v
Δv
f

P shell

Δv

KEM

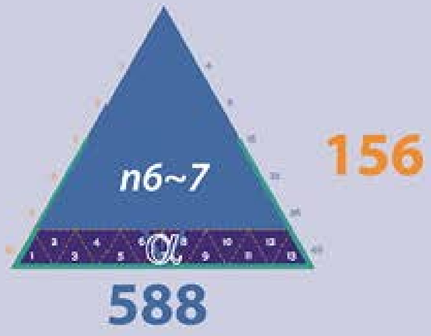

n6~7
α
156
588

1 quantum level jump

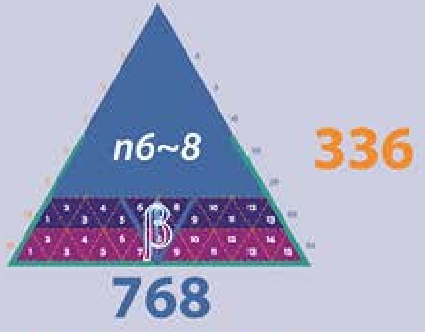

n6~8
β
336
768

2 quantum level jump

Tetryonics 35.02 - Humphreys quantum transitions

Humphreys series energy momenta

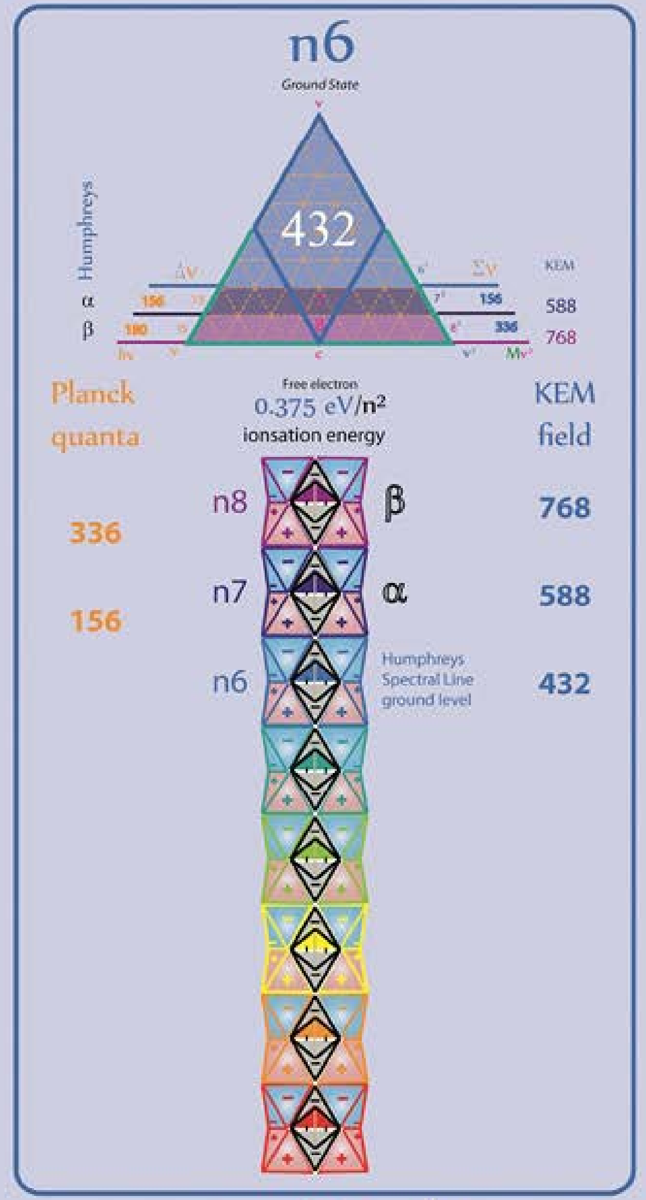
n6
Ground State
432
Humphreys
α
β
Δv
156
180
Σv
156
336
KEM
588
768
Mv²
Planck
quanta
Free electron
0.375 eV/n²
ionsation energy
KEM
field
n8
β
336
768
n7
α
156
588
n6
Humphreys
Spectral Line
ground level
432

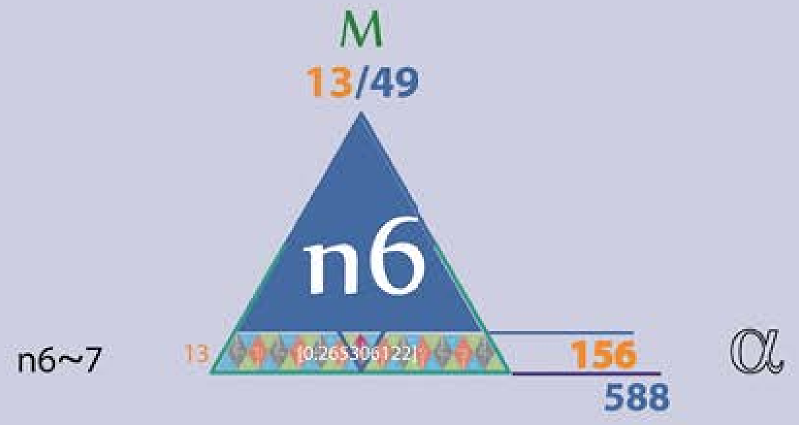
M
13/49
n6
n6~7
13
[0.265306122]
156
588
α

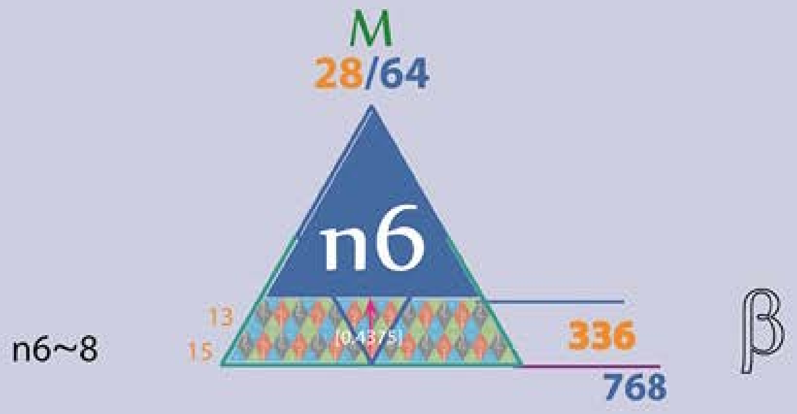
M
28/64
n6
n6~8
13
15
[0.4375]
336
768
β

accelerating electrons $\Delta \mathbf{p} = \Delta M v = hf$ produce spectral lines

Tetryonics 35.03 - Humphreys series energy momenta

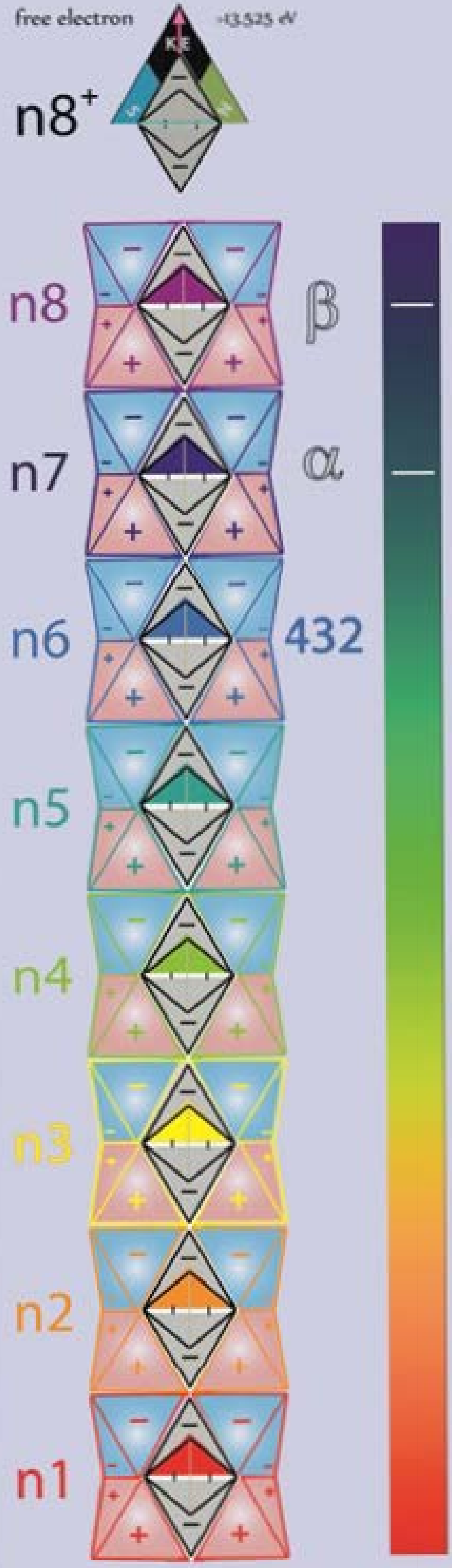

Humphreys Series

$$\Delta h\nu = \text{KEM} = hcR$$

Tetryonic — 0.375 eV — Rydberg

28/64

.01215

λ	7,546.831235	nm
$\tilde{\nu}$	132,505.9444	m-1
f	3.972428277 x 10^{13}	Hz
E	0.164369815	eV
Ω	1,123.018633	m²/s

$R\left(\frac{1}{36} - \frac{1}{64}\right)$

.01215

13/49

.00733

λ	12,445.01497	nm
$\tilde{\nu}$	80,353.459	m-1
f	2.408993609$8$x 10^{13}	Hz
E	0.099676156	eV
Ω	1,403.413807	m²/s

$R\left(\frac{1}{36} - \frac{1}{49}\right)$

.0073

m.				
Ω/v	λ	3,301.738665	nm	h/p
v/Ω	$\tilde{\nu}$	302,870.7301	m-1	p/h
c²/Ω	f	9.079836062 x 10^{13}	Hz	E/h
v²/c²	E	0.375702435	eV	hf
	Ω	889.8363501	m²/s	

ground KEM energies

changing energy momenta produces specific frequency spectral lines

Copyright ABRAHAM [2008] - all rights reserved

Copyright ABRAHAM [2008] - all rights reserved

Un-named spectral line

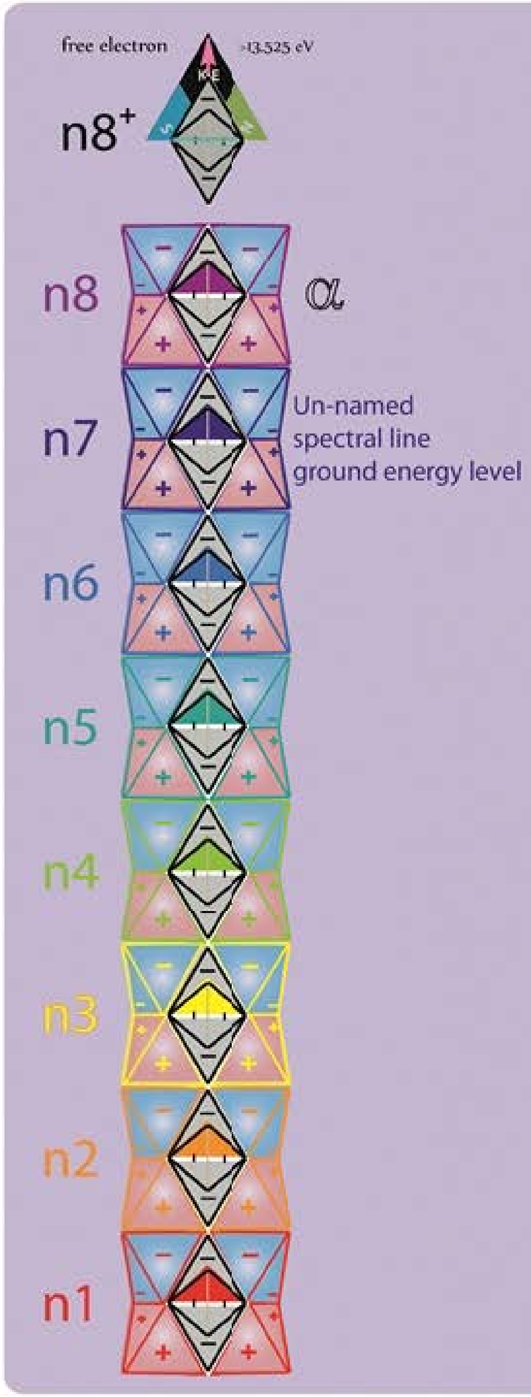
free electron
>13.525 eV
n8⁺
KE
S
N
n8
α
n7
Un-named
spectral line
ground energy level
n6
n5
n4
n3
n2
n1

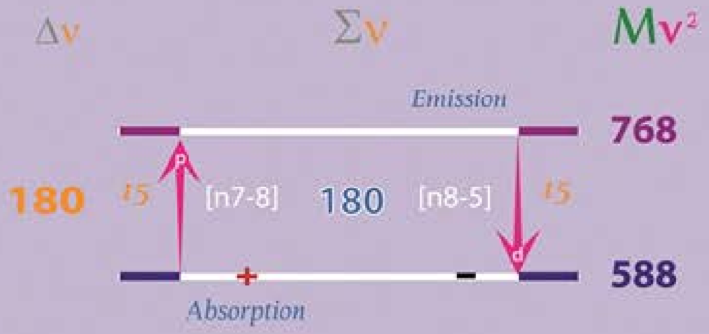
Δv
Σv
Mv²
Emission
768
180
t5
p
[n7-8]
180
[n8-5]
t5
d
588
Absorption

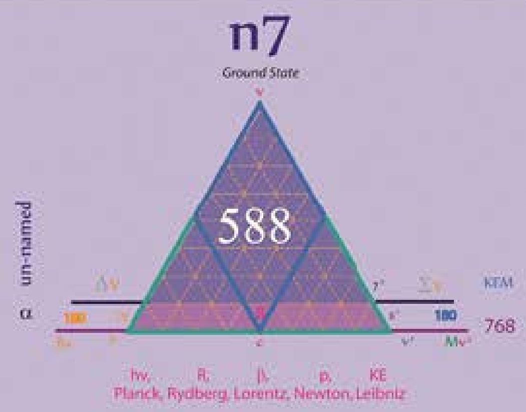
n7
Ground State
588
un-named
α
Δv
Σv
KEM
180
180
768
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

-0.276026782 eV

$$KEM = \frac{hf}{n^2}$$

Q shell

4,494.03 nm

$$KEM = \frac{hcR}{n^2}$$

m.	Ω/v	λ	4,494.033183	nm	h/p
	v/Ω	$\tilde{v}$	222,517.2711	m-1	p/h
	c^2/Ω	f	$6.67089964 \times 10^{13}$	Hz	E/h
	v^2/c^2	E	0.276026278	eV	hf
		Ω	1,347.277254	m^2/s	

ground KEM energies

equilateral energy momenta is the foundation of all quantum EM wave functions

Tetryonics 36.01 - Un-named spectral line

This spectral line photo electrons transition can only release ***one*** *specific [Δhν] W boson of energy*

-0.276026782 eV

ν
n7
KE
S
N
12
f.2e20

4,494.03 nm

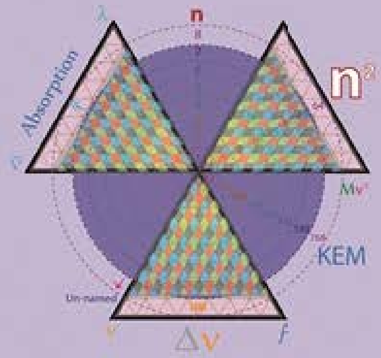

n
λ
Absorption
n²
Mν²
KEM
Un-named
ν
Δν
f

Un-named spectral series

α **19,174** n1-8

IR spectrum

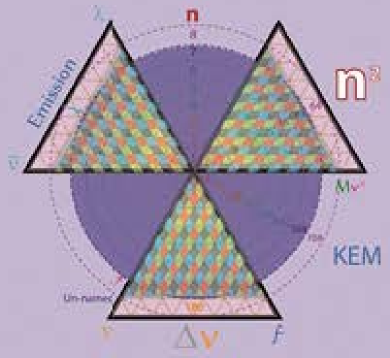

n
λ
Emission
n²
Mν²
KEM
Un-named
ν
Δν
f

Q shell

Δν

KEM

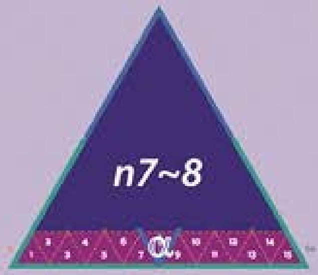

n7~8

180

768

1 quantum level jump

Copyright ABRAHAM [2008] - all rights reserved

Un-named series energy momenta

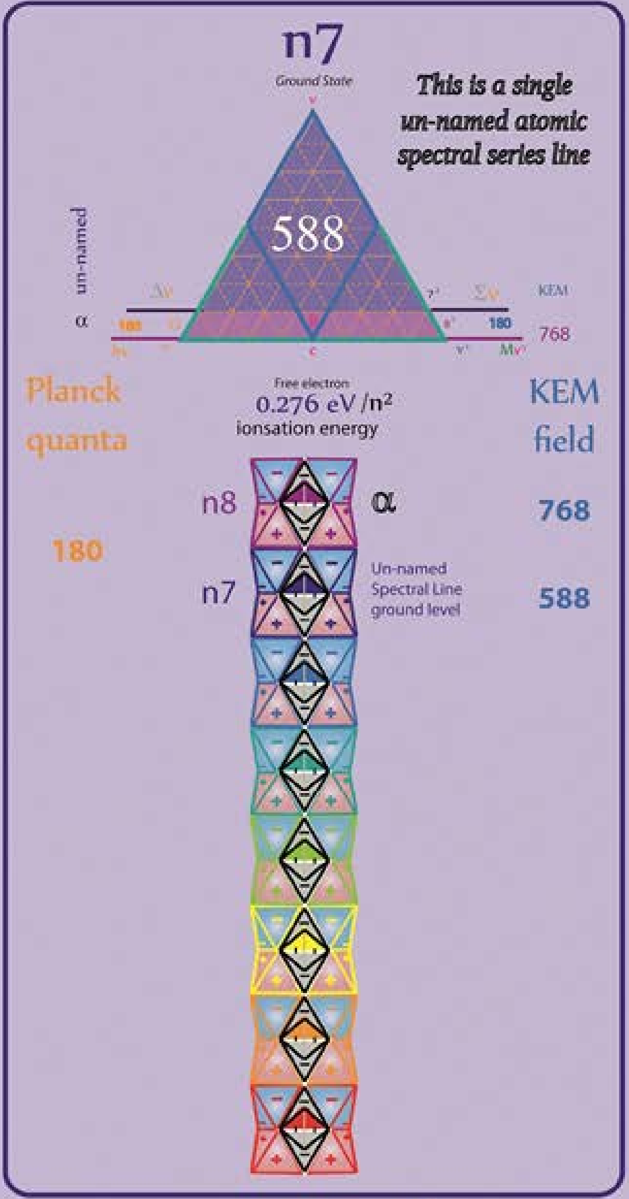
n7
Ground State
v
588
un-named
α
Δv
180
hv
c
γ²
Σv
KEM
180
768
v²
Mv²
Planck
quanta
180
KEM
field
n8
α
768
n7
588

This is a single un-named atomic spectral series line

Free electron
0.276 eV /n²
ionsation energy

Un-named
Spectral Line
ground level

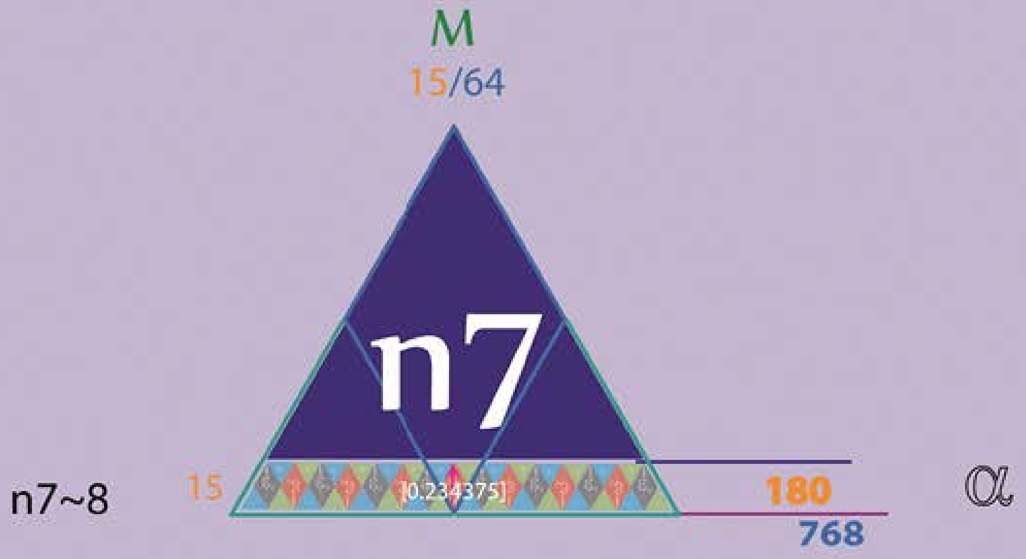
M
15/64
n7
n7~8
15
[0.234375]
180
768
α

accelerating electrons $\Delta p = \Delta Mv = hf$ produce spectral lines

Tetryonics 36.03 - Un-named series energy momenta

Copyright ABRAHAM [2008] - all rights reserved

Un-named Series

$$\Delta h\nu = \underset{0.276\ eV}{KEM} = hcR$$

Tetryonic Rydberg

15/64

.0047

λ	19,174.54158	nm
$\bar{\nu}$	52,152.48541	m-1
f	1.563492179 x 10 ^{13}	Hz
E	0.064693658	eV
Ω	5,748.382952	m^2/s

$$R\left(\frac{1}{49}-\frac{1}{64}\right)$$

.0047

$$m.\begin{bmatrix} \Omega/\nu \\ \nu/\Omega \\ c^2/\Omega \\ \nu^2/c^2 \end{bmatrix}$$

λ	4,494.033183	nm
$\bar{\nu}$	222,517.2711	m-1
f	6.67089964 x 10 ^{13}	Hz
E	0.276026278	eV
Ω	1,347.277254	m^2/s

h/p
p/h
E/h
hf

ground KEM energies

changing energy momenta produces a specific frequency spectral line

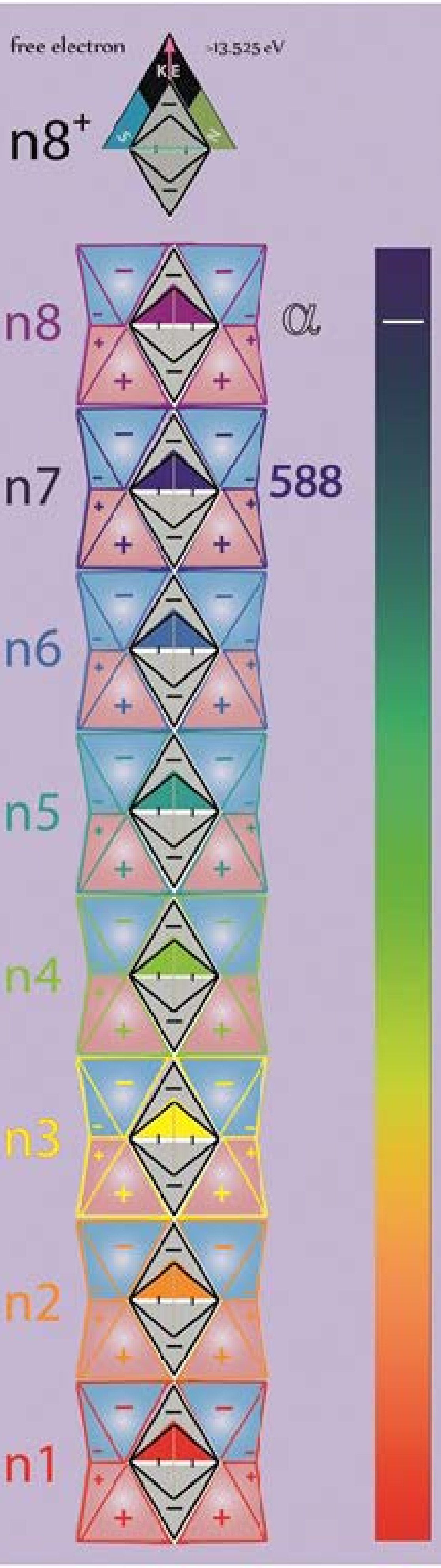

Tetryonics 36.04 - Un-named series maths

KEM field geometry

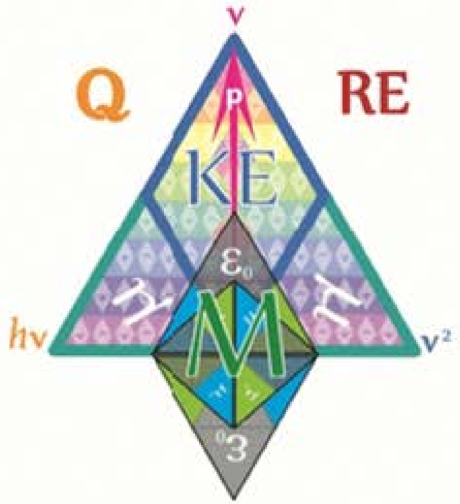
v
Q
RE
P
KE
ε0
M
hv
v²
3

The Lorentz corrected velocity of electron resulting from atomic energy level transitions

Mv^2

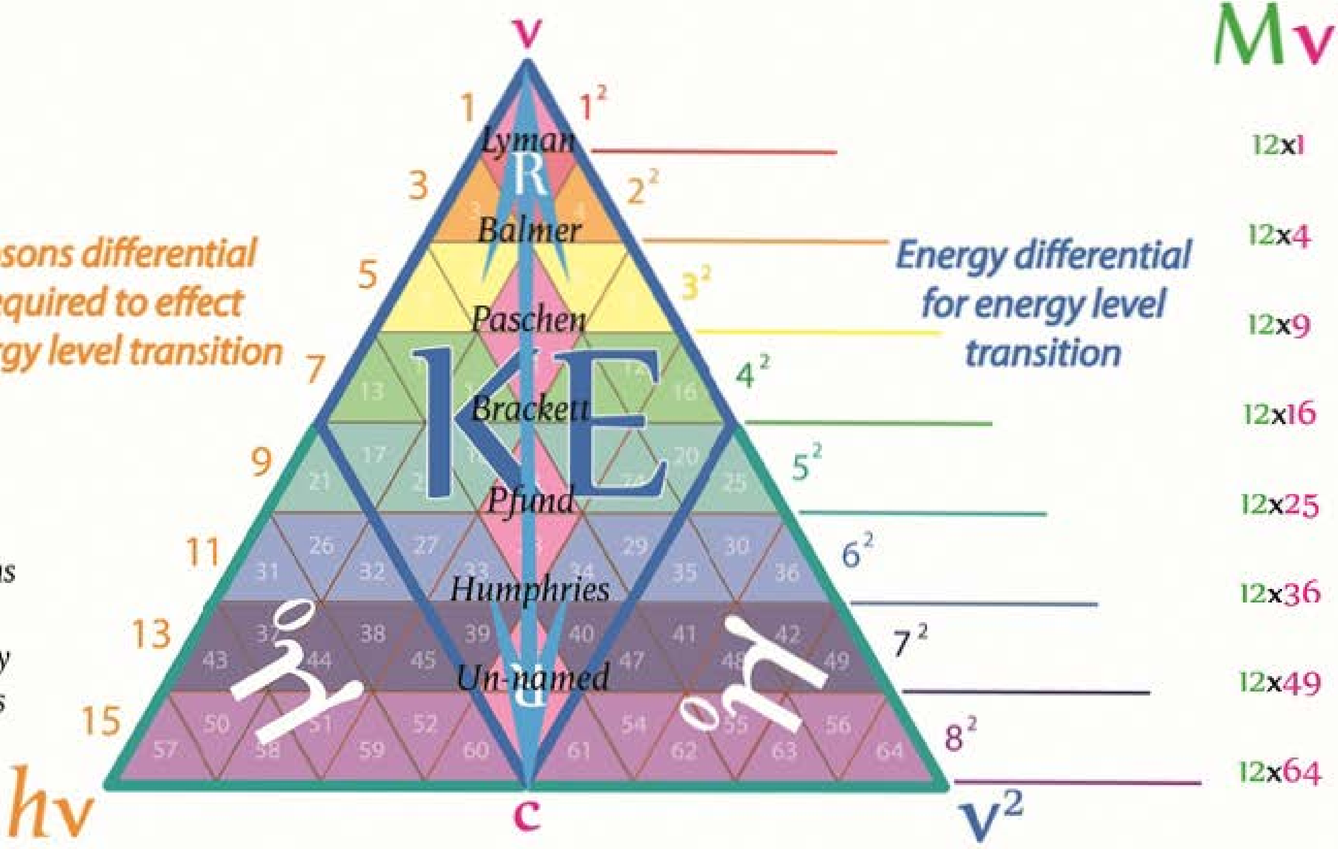
v
1
1²
Lyman
R
3
2²
Balmer
5
3²
Paschen
7
4²
KE
Brackett
9
5²
Pfund
11
6²
Humphries
13
7²
Un-named
15
8²
hv
c
v²
12x1
12x4
12x9
12x16
12x25
12x36
12x49
12x64

Bosons differential required to effect energy level transition

Energy differential for energy level transition

All of the quantum jumps of electrons bound in a Hydrogen atom can be summerised in the Tetryonic geometry reflecting the energy level transitions

Ryberg's constant is reflective of the wavelengths and linear momentum of a transitioning electron's KEM field

Ryberg's formula is a mathematical description of Tetryonic energy geometry

$$Mv^2 = \text{KEM} = hcR_H$$

KEM field quanta

Copyright ABRAHAM [2008] - all rights reserved

Photoelectron transitions

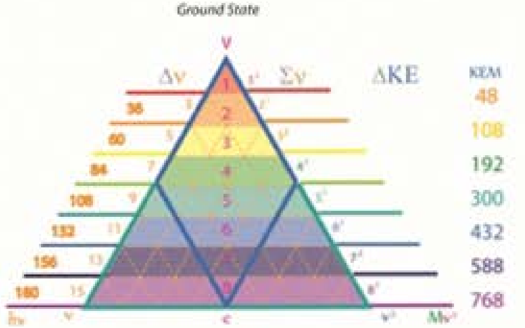
Ground State
v
Δv
Σv
ΔKE
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
1
2
3
4
5
6
7
8
hv
v
c
Mv²

Spectral Lines

Transition			Spectral series	
n1~	n2	91.29	α	Lyman UV spectra
	n3	91.30	β	
	n4	91.32	γ	
	n5	91.33	δ	
	n6	91.36	ε	
	n7	91.38	ϕ	
	n8	91.41	η	
n2~	n3	367.06	α	Balmer Visible Spectra
	n4	367.23	β	
	n5	367.28	γ	
	n6	367.54	δ	
	n7	368.19	ε	
	n8	368.59	ϕ	
n3~	n4	833.16	α	Paschen IR spectra
	n5	834.34	β	
	n6	835.68	γ	
	n7	837.23	δ	
	n8	839.02	ε	
n4~	n5	1499.69	α	Brackett
	n6	1503.50	β	
	n7	1507.87	γ	
	n8	1512.92	δ	
n5~	n6	2381.53	α	Pfund
	n7	2391.17	β	
	n8	2402.25	γ	
n6~	n7	3499.27	α	Humphreys
	n8	3520.12	β	
n7~	n8	4880.38	α	Un-named

Nuclear Quantum Levels

The kinetic energy levels of bound nuclear photon-electrons is determined by the energy levels of their Baryons

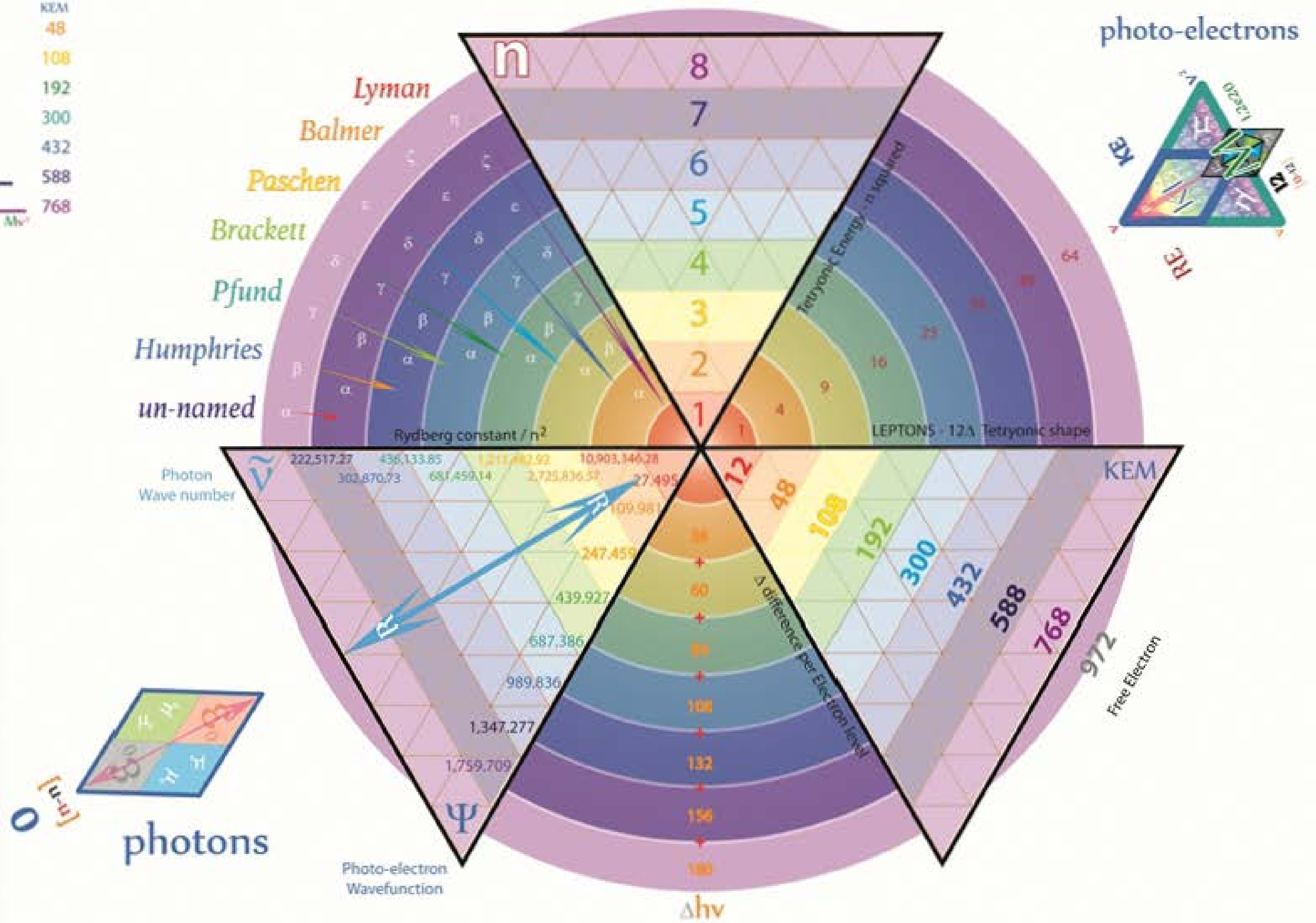
photo-electrons
KE
RE
12
n
8
7
6
5
4
3
2
1
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
α
β
γ
δ
ε
ζ
η
Tetryonic Energy - n squared
64
49
36
25
16
9
4
1
Rydberg constant / n²
LEPTONS - 12Δ Tetryonic shape
ṽ
Photon Wave number
222,517.27
302,870.73
436,133.85
681,459.14
1,211,482.92
2,725,836.57
10,903,346.28
27.495
109.981
247.459
439.927
687.386
989.836
1,347.277
1,759.709
Ψ
Photo-electron Wavefunction
KEM
12
48
108
192
300
432
588
768
972
Free Electron
Δ difference per Electron level
36
60
84
108
132
156
180
Δhv
0
[n-n]
photons

Spectral lines are produced by

The energy and wavelength of emitted photons is a function of the photo-electron's KEM field quantum wavenumber

Quantum Transitions

$$hf = \Delta Mv = \Delta p$$

27.49545

accelerating photo-electrons

Changes to the energy momenta of KEM fields produce changes to the linear momenta of photo-electrons

Photoelectric energy level transitions

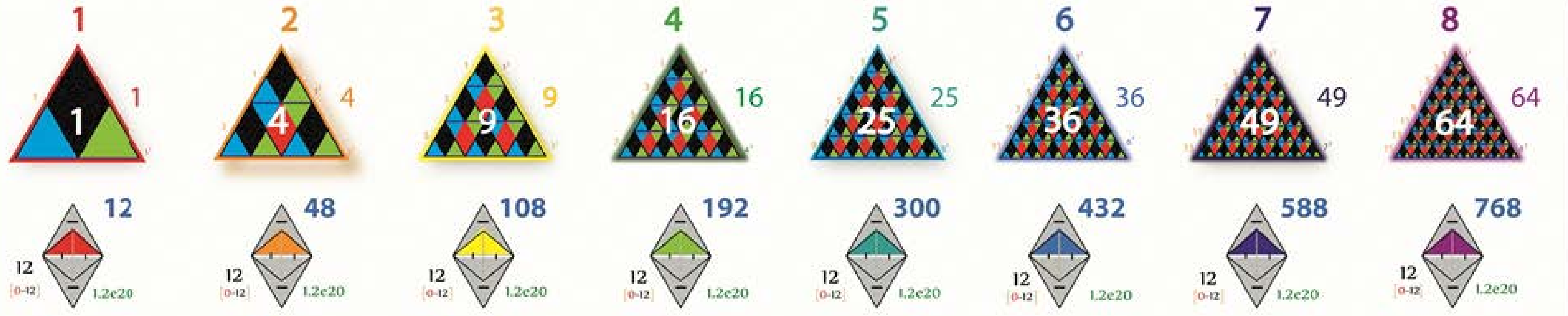
1
2
3
4
5
6
7
8
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768
12
[0-12]
1.2e20

Lyman	*Balmer*	*Paschen*	*Brackett*	*Pfund*	*Humphries*	*un-named*
91.29	367.06	833.16	1499.69	2381.53	3499.27	1917.4
91.3	367.23	834.34	1503.5	2391.17	3520.12	
91.32	367.28	835.68	1507.87	2402.25		
91.33	367.54	837.23	1512.92			
91.36	368.19	839.02				
[illegible]	[illegible]					
91.41						
UV spectrum	*Visible Spectrum*	*IR spectrum*				

Applying any extra Kinetic Energy to leptons in this energy level results in unbound photo-leptons

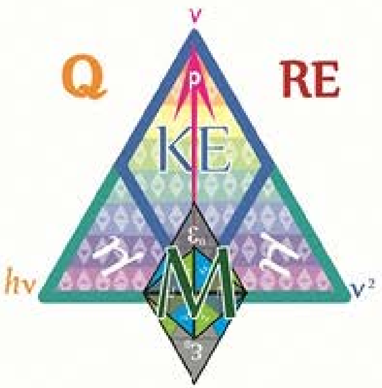
Q
RE
v
p
KE
hv
v²
M

$$Mv^2 = KEM = hcR_H$$

These specific quantum level steps create the Photoelectric effect

$$\Delta hv = \phi = hf$$

Photo-electrons bound to atomic nuclei can only transition between Mv squared KEM energies as a direct result of the energy levels of Baryons

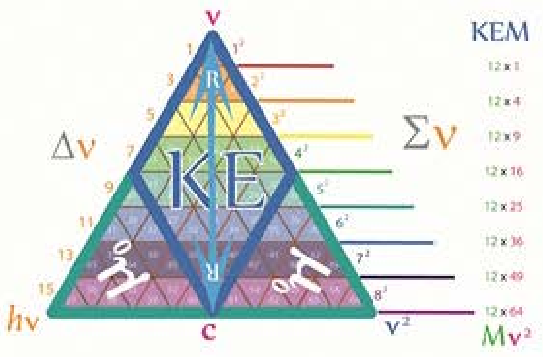
v
KEM
1
3
5
7
9
11
13
15
1²
2²
3²
4²
5²
6²
7²
8²
Δv
Σv
KE
R
12 x 1
12 x 4
12 x 9
12 x 16
12 x 25
12 x 36
12 x 49
12 x 64
hv
c
v²
Mv²

The Photoelectric effect

The photoelectric effect was first observed in 1887 by Heinrich Hertz (1857-1894) during experiments with a spark-gap generator — the earliest form of radio receiver

There are specific frequencies of light that are more efficient at exciting photo-electrons from atoms

$$p^2 = KEM = Mv^2$$

The energies of the emitted electrons are independent of the intensity of the incident radiation but are dependent on the photon frequency

n2-4 specific transitions can release 8 times the number of electrons that n3-4 can release & 4 times the number of electrons that n1-4 can release

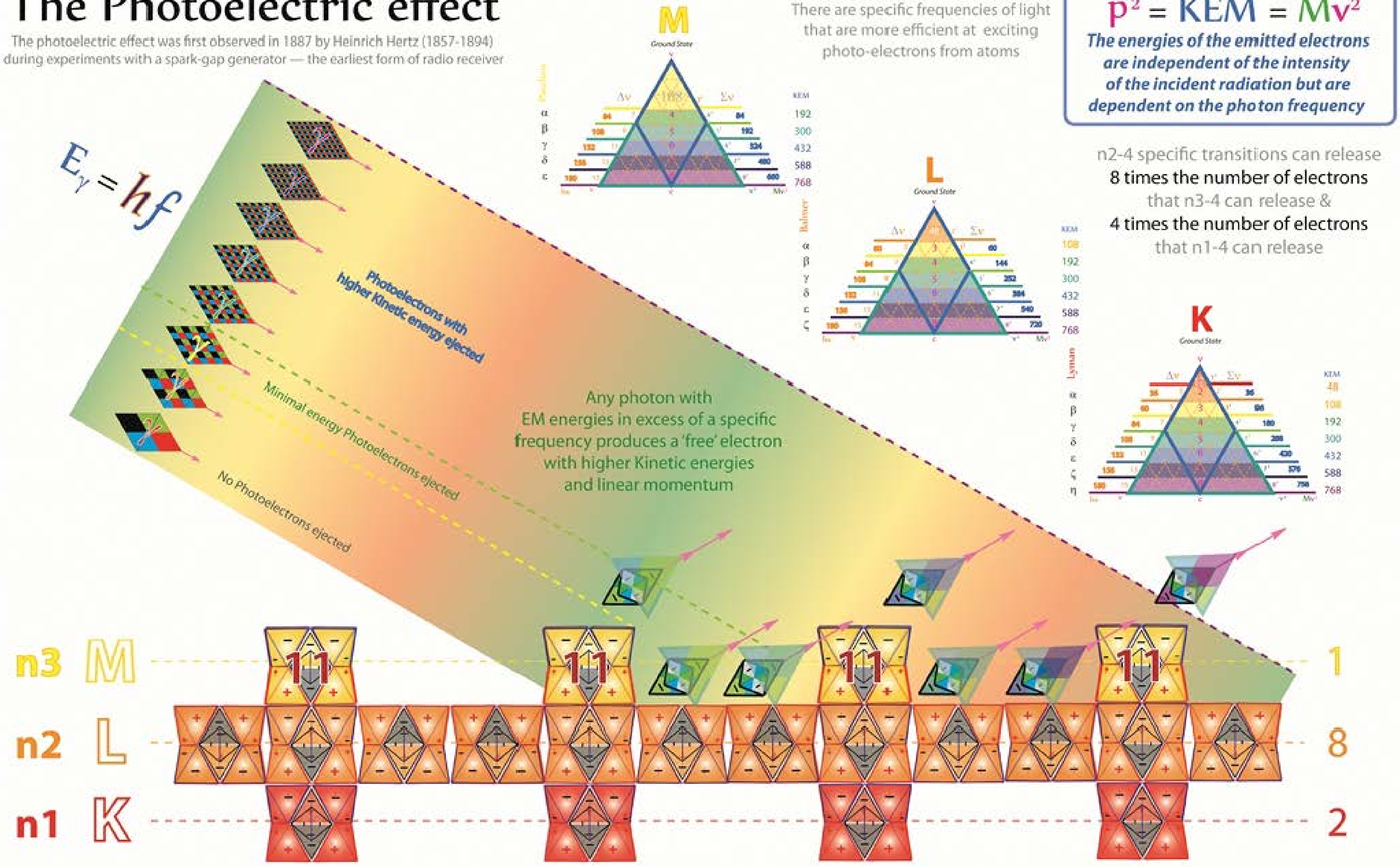
$E_\gamma = hf$
M
Ground State
Paschen
KEM
192
300
432
588
768
L
Ground State
Balmer
KEM
108
192
300
432
588
768
K
Ground State
Lyman
KEM
48
108
192
300
432
588
768
Photoelectrons with higher Kinetic energies ejected
Minimal energy Photoelectrons ejected
No Photoelectrons ejected
n3 M 1
n2 L 8
n1 K 2

Any photon with EM energies in excess of a specific frequency produces a 'free' electron with higher Kinetic energies and linear momentum

Soon after the discovery of Planck's Heat law and the quantisation of energy at the quantum level experiments by others, most notably Robert Millikan (1865-1953), found that light with frequencies below a certain cutoff value, called the threshold frequency, ould not eject photoelectrons from the metal surface no matter how bright the source

Copyright ABRAHAM [2008] - all rights reserved

Photonic mass-energy momenta

$$2m\Omega v^2 = E = n.hf$$

All 2D EM waveforms possess mass-energy and momenta

Bosons

$$\text{ODD}\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Bosons · EM Field · ElectroMagnetic · Planck · quanta · mass · velocity

Photons

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Photons · EM Field · ElectroMagnetic · Planck · quanta · mass · velocity

charged mass-energy-momentum

$$E = mv^2$$

radiant mass-energy momenta

$$E = hv^2$$

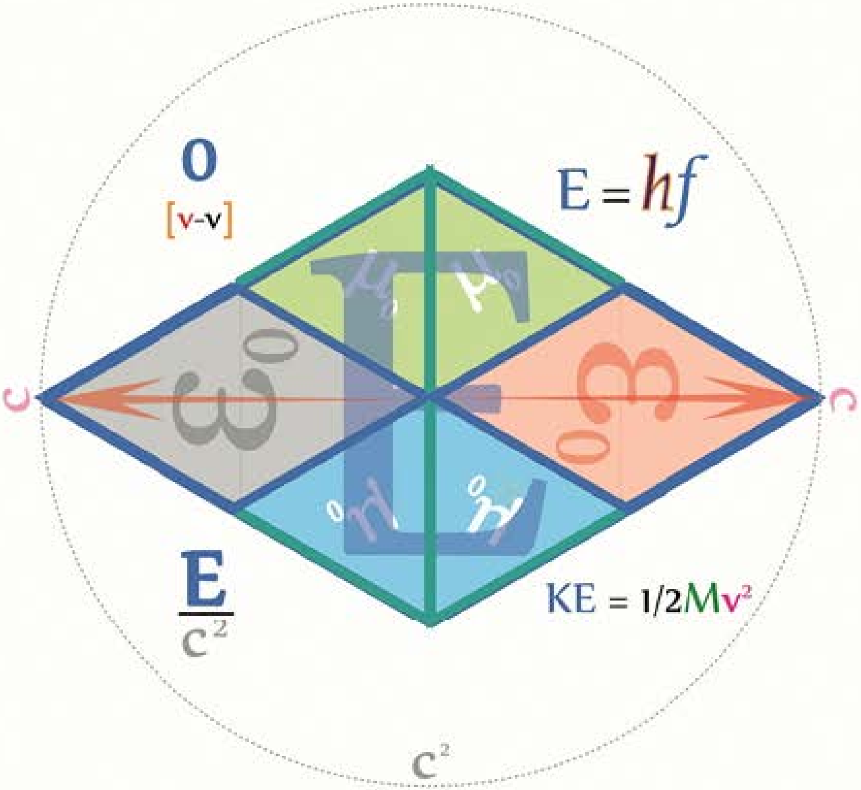

Planck energies

All 2D EM waveforms propagate at the speed of light

Velocity of Light · Speed of Light

$$V = \left[\left[\frac{c^2}{\Omega}\right]\cdot\left[\frac{\Omega}{c}\right]\right] = c$$

f: scalar frequency · λ: linear wavelength

Tetryonics 37.05 - Photonic mass-Energy equivalence

Wave~Particle duality

First proposed by Christian Huygens in the 1600's, it wasn't until the 1800's that Thomas Young proved this wave-particle duality with the classic double-slit experiment.

Sir Isaac Newton

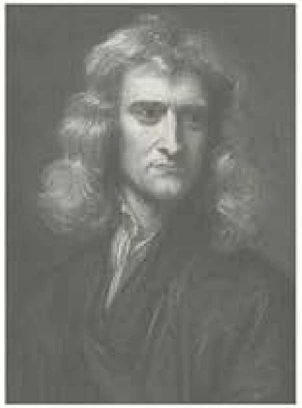

(1643-1727)

Particles

In 1678, Dutch physicist, Christiaan Huygens, believed that light was made up of waves vibrating up and down perpendicular to the direction of the light travels, and therefore formulated a way of visualising wave propagation.

This became known as 'Huygens' Principle' and was the successful theory of light wave motion in three dimensions

Sir Isaac Newton, held the theory based on his spetral observations that light was made up of tiny particles or 'corpusles of colour'

English physicist Thomas Young argued that Isaac Newton's theory of particle light was incorrect, and instead argued that light is a transverse wave.

In 1803, Thomas Young studied the interference of light waves by shining light through a screen with two slits equally separated, the light emerging from the two slits, spread out and produced wave-like interference patterns.

Waves

Thomas Young

(1773-1829)

In 1900
Max Planck proposed the existence of a light quantum,
[n.hv] a finite packet of energy which depends on the
frequency and spectral energy of the radiation

Light - a Wave, or a Particle?

In 1905
Albert Einstein, suggested that light is composed
of tiny particles called photons, and that
each photon has energy related to its frequency [hf].

Max Karl Ernst Ludwig Planck

(April 23, 1858 – October 4, 1947)

Bosons

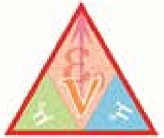

$$nhv = E = hf$$

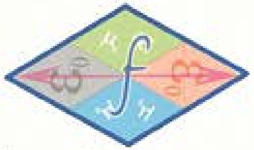

Mathematically directly relating the number of
Planck quanta [n.hv] with Photon frequencies [hf]
and the frequent interchanging of one term for the other
in physics is the source of considerable quantum confusion

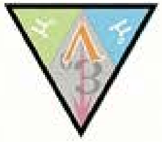

$$2hv = hf$$

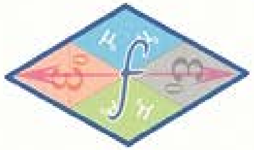

Photons

Albert Einstein

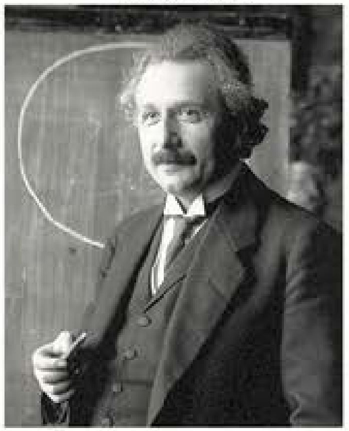

(14 March 1879 – 18 April 1955)

The quantum idea was soon seized upon to explain the photoelectric effect, and became part of the Bohr theory of discrete atomic spectra, quickly becoming part of the foundation of modern quantum theory in turn this led to the quantum wierdness of wave-particle dualty, Heisenberg's Uncertainty principle and Schrodinger's quantum wave equation & wavefunctions.

Wave Particle Mechanics

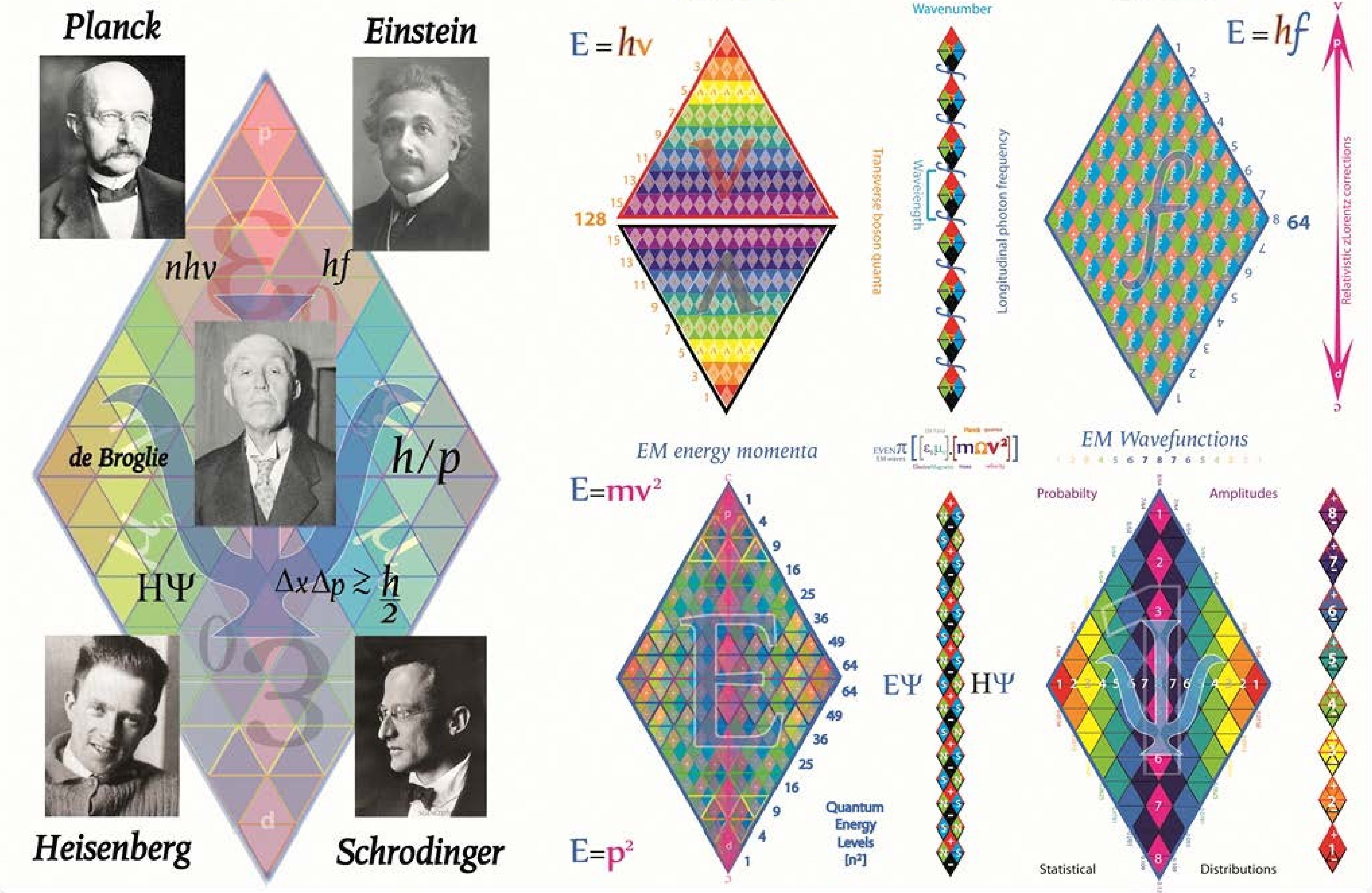
Planck
Einstein
nhv
hf
de Broglie
h/p
HΨ
ΔxΔp ≥ ħ/2
Heisenberg
Schrodinger
E = hv
128
Wavenumber
Transverse boson quanta
Wavelength
Longitudinal photon frequency
E = hf
64
Relativistic zLorentz corrections
EM energy momenta
EM Wavefunctions
E=mv²
E=p²
1 4 9 16 25 36 49 64 64 49 36 25 16 9 4 1
Quantum Energy Levels [n²]
EΨ
HΨ
Probabilty
Amplitudes
Statistical
Distributions

Copyright ABRAHAM [2008] - all rights reserved

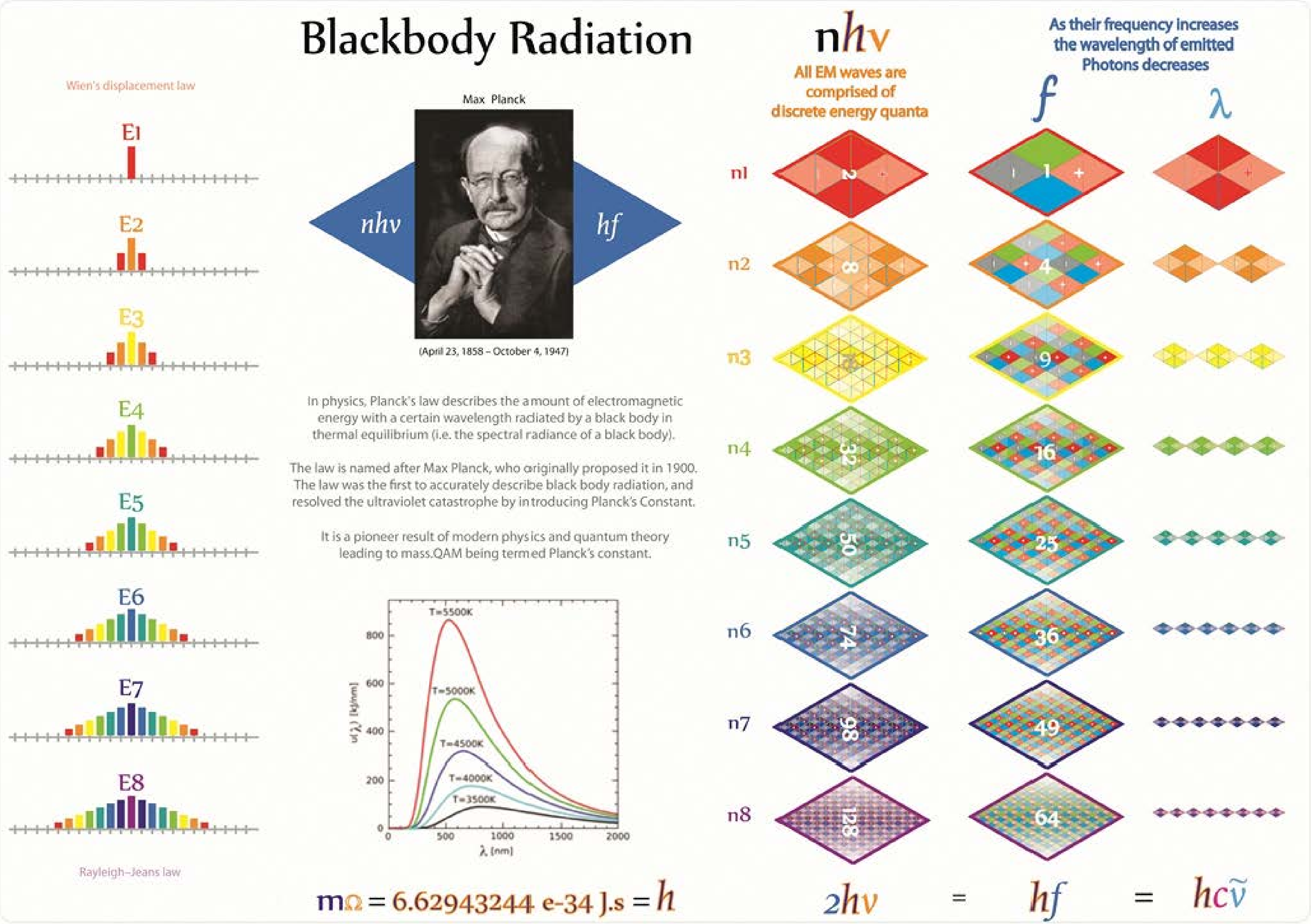

Blackbody Radiation

Max Planck

(April 23, 1858 – October 4, 1947)

In physics, Planck's law describes the amount of electromagnetic energy with a certain wavelength radiated by a black body in thermal equilibrium (i.e. the spectral radiance of a black body).

The law is named after Max Planck, who originally proposed it in 1900. The law was the first to accurately describe black body radiation, and resolved the ultraviolet catastrophe by introducing Planck's Constant.

It is a pioneer result of modern physics and quantum theory leading to mass.QAM being termed Planck's constant.

nhv

All EM waves are comprised of discrete energy quanta

As their frequency increases the wavelength of emitted Photons decreases

mΩ = 6.62943244 e-34 J.s = h

2hv = hf = hcṽ

Wien's displacement law
E1
E2
E3
E4
E5
E6
E7
E8
Rayleigh–Jeans law
nhv
hf
T=5500K
T=5000K
T=4500K
T=4000K
T=3500K
u(λ) [kJ/nm]
0
200
400
600
800
0
500
1000
1500
2000
λ [nm]
f
λ
n1
n2
n3
n4
n5
n6
n7
n8
2
8
18
32
50
74
98
128
1
4
9
16
25
36
49
64

Tetryonics 38.02 - Blackbody Radiation curves

Copyright ABRAHAM [2008] - all rights reserved

EM field energies

Planck quanta

$$E = n\pi\left[\left[m\Omega v^2\right]\right]$$

mass velocity

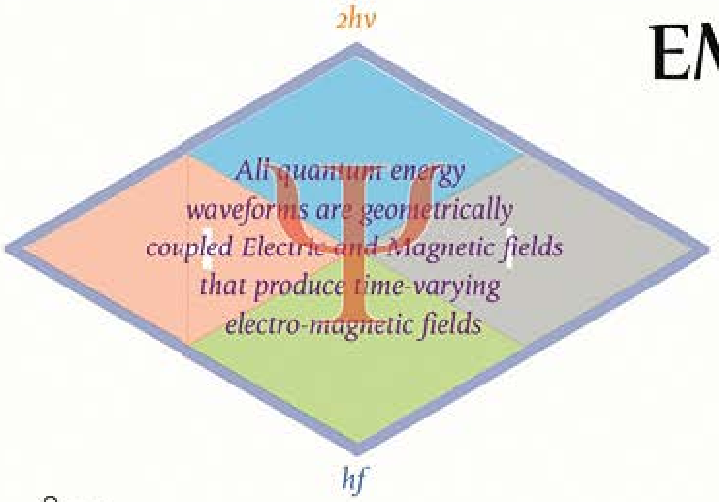

All quantum energy waveforms are geometrically coupled Electric and Magnetic fields that produce time-varying electro-magnetic fields

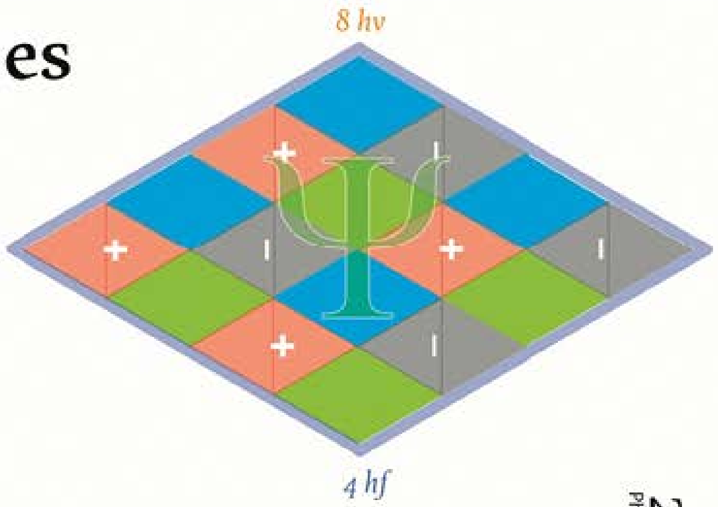

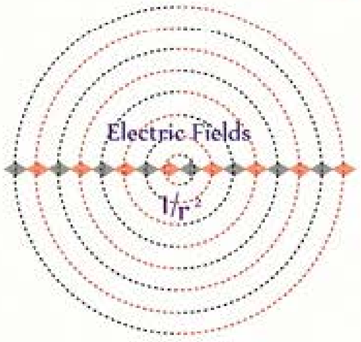
Electric Fields
1/r²

Charge

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field · Planck quanta · ElectroMagnetic · mass · velocity

$$E = h\nu$$

EM waves are subject to Lorentz co-ordinate transformations

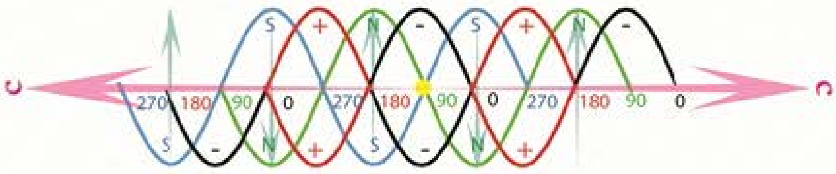
c
S
+
N
-
S
+
N
-
270
180
90
0
270
180
90
0
270
180
90
0
S
-
N
+
S
-
N
+
c

EM waves are a quadrature waveform of bi-directional electromagnetic oscillations

$$E = hf$$

Photons

$$2\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM Field · Planck quanta · ElectroMagnetic · mass · velocity

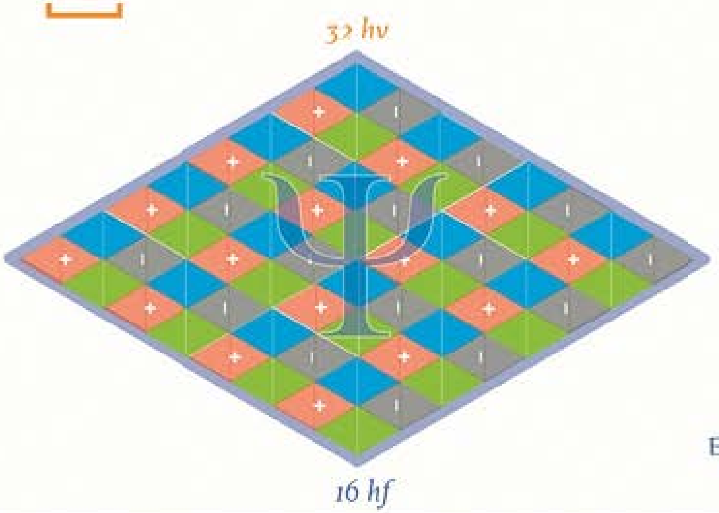

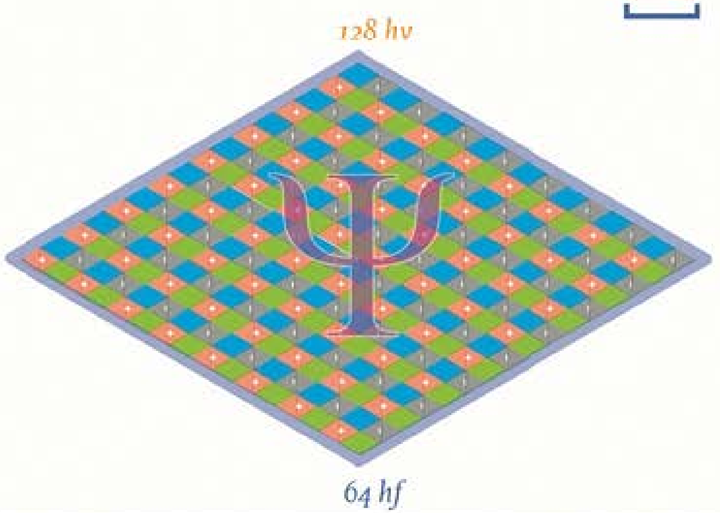

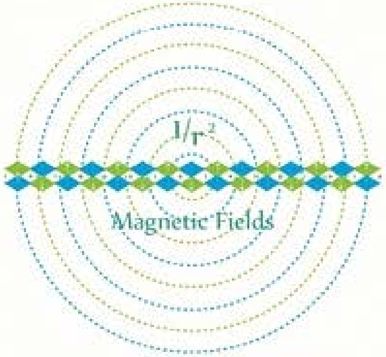
1/r²
Magnetic Fields

EM Field Planck quanta

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves ElectroMagnetic mass velocity

Tetryonics 38.03 - EM field Energy

Bosons and Photons in EM waves

EM Field Planck quanta

$$\text{ODD}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Bosons ElectroMagnetic mass velocity

Transverse charge Quanta

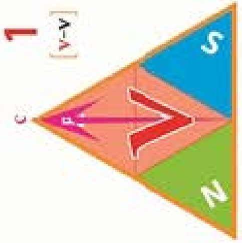

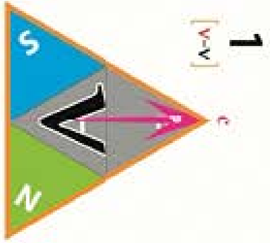

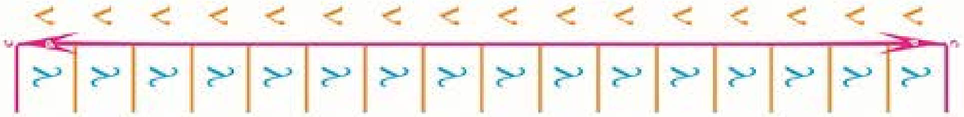
λ λ λ λ λ λ λ λ λ λ λ λ λ λ λ λ

All EM waves are comprised of transverse Bosons

$$E = h\nu$$

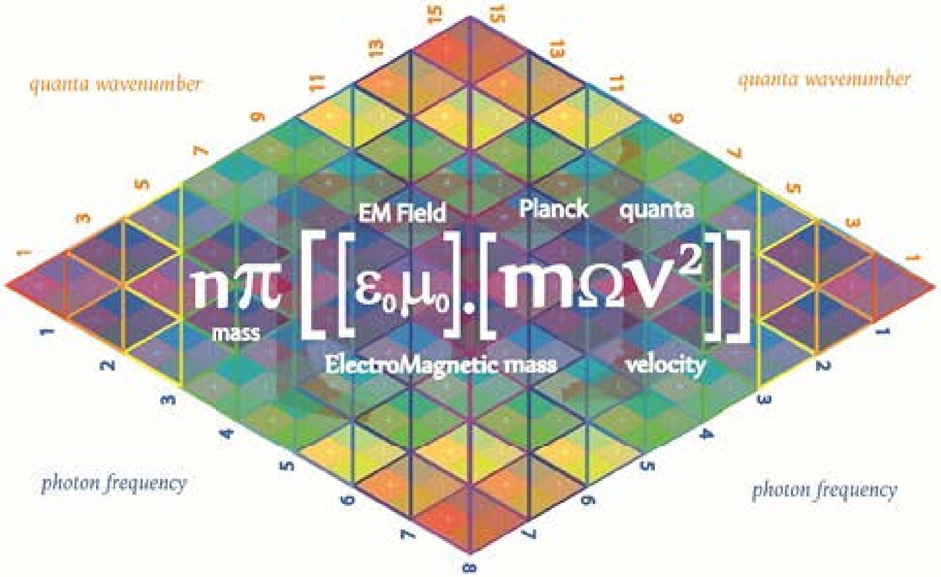
quanta wavenumber
quanta wavenumber
1 3 5 7 9 11 13 15
15 13 11 9 7 5 3 1
1 2 3 4 5 6 7 8
7 6 5 4 3 2 1
photon frequency
photon frequency

EM Field Planck quanta

$$n\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

mass ElectroMagnetic mass velocity

$$E = hf$$

All EM waves are comprised of longitudinal Photons

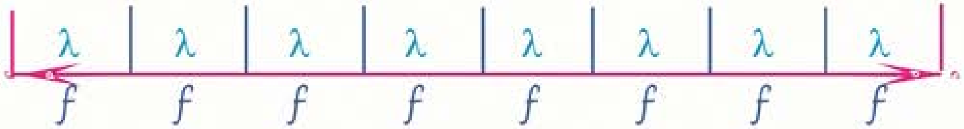
λ λ λ λ λ λ λ λ
f f f f f f f f

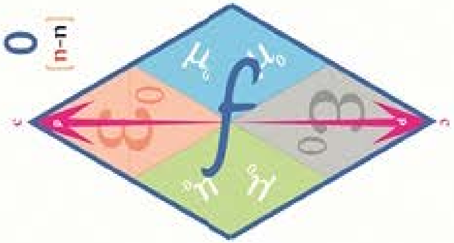

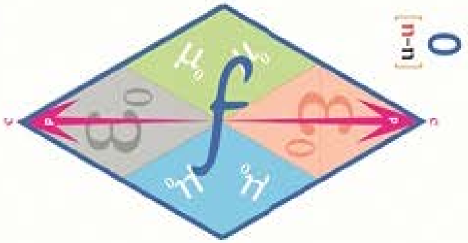

Longitudinal neutral Photons

EM Field Planck quanta

$$\text{EVEN}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$$

Photons ElectroMagnetic mass velocity

Copyright ABRAHAM [2008] - all rights reserved

EM wave energy quanta

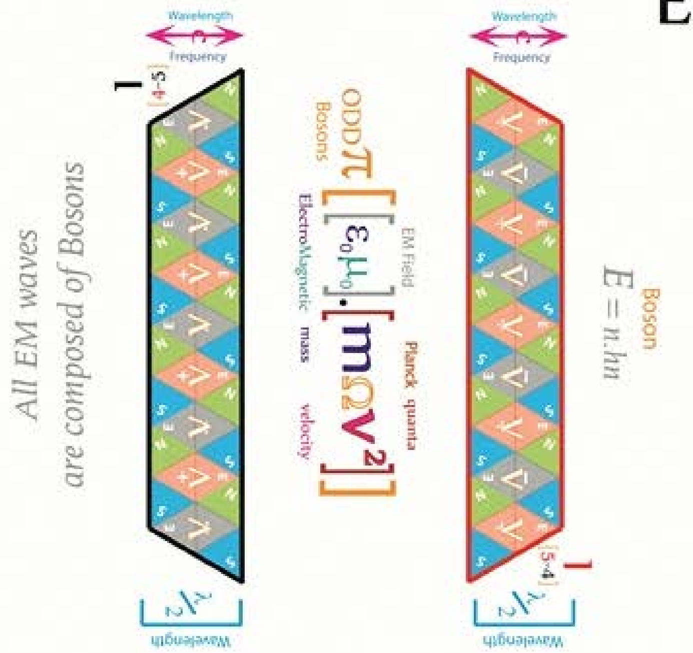
Wavelength
c
Frequency
1
[4-5]
ODD π
Bosons
$[[\varepsilon_0\mu_0].[m\Omega v^2]]$
EM Field
ElectroMagnetic
mass
Planck quanta
velocity
Boson
$E = n.hn$
[5-4]
1
$\frac{1}{2}\lambda$
Wavelength
$\frac{1}{2}\lambda$
Wavelength

All EM waves are composed of Bosons

Tranverse Bosons in EM wave energies should be modelled using [n] Planck quanta

Wavefunction

$$i\hbar\frac{\partial}{\partial t}\Psi \quad \Psi \quad \hat{H}\Psi$$

Probability = |Amplitude|^2

The Longitudinal EM wave energies should be modelled using [f] Photon frequencies

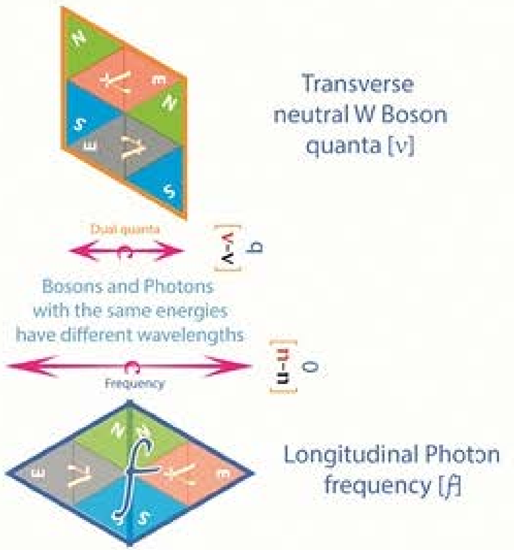
Transverse neutral W Boson quanta [ν]
Dual quanta
c
b
[v-v]
Bosons and Photons with the same energies have different wavelengths
c
Frequency
0
[n-n]
Longitudinal Photon frequency [f]

$$E = n\pi\left[m\Omega v^2\right]$$

Planck quanta

mass velocity

All Wave~Particle geometries and mass-energies can be calculated from and modelled with the Tetryonic Unified Field Equation

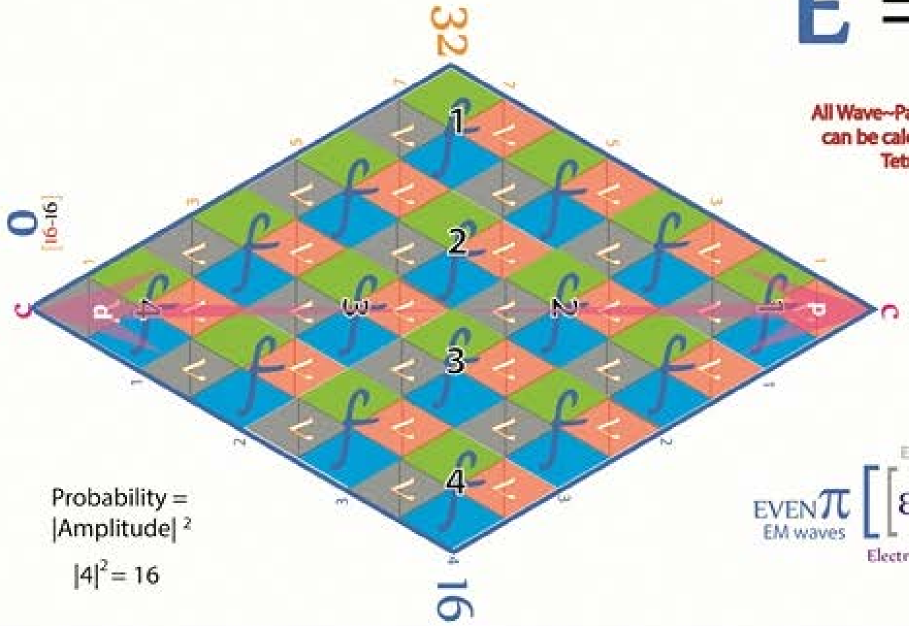
32
0
[16-16]
c
d
1
2
3
4
c
16

Probability = |Amplitude|^2

$|4|^2 = 16$

$E = hf$

Photon

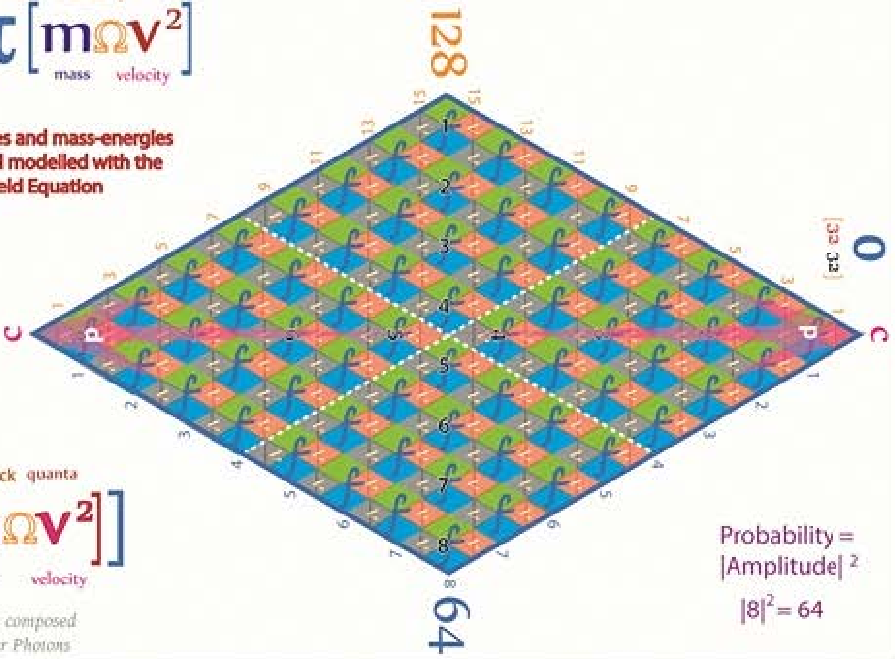
128
0
[32-32]
c
d
1
2
3
4
5
6
7
8
c
64

Probability = |Amplitude|^2

$|8|^2 = 64$

$$\text{EVEN}\,\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

EM waves

EM Field Planck quanta

ElectroMagnetic mass velocity

All EM waves are composed of Square number Photons

Tetryonics 38.05 - EM wave Energies

Copyright ABRAHAM [2008] - all rights reserved

EM wave properties

Linear momentum
is the square root of
scalar mass-energies

Energy

$$\left[\sqrt{[m\Omega v^2]}\right]$$

mass quanta

EM waveform
geometries

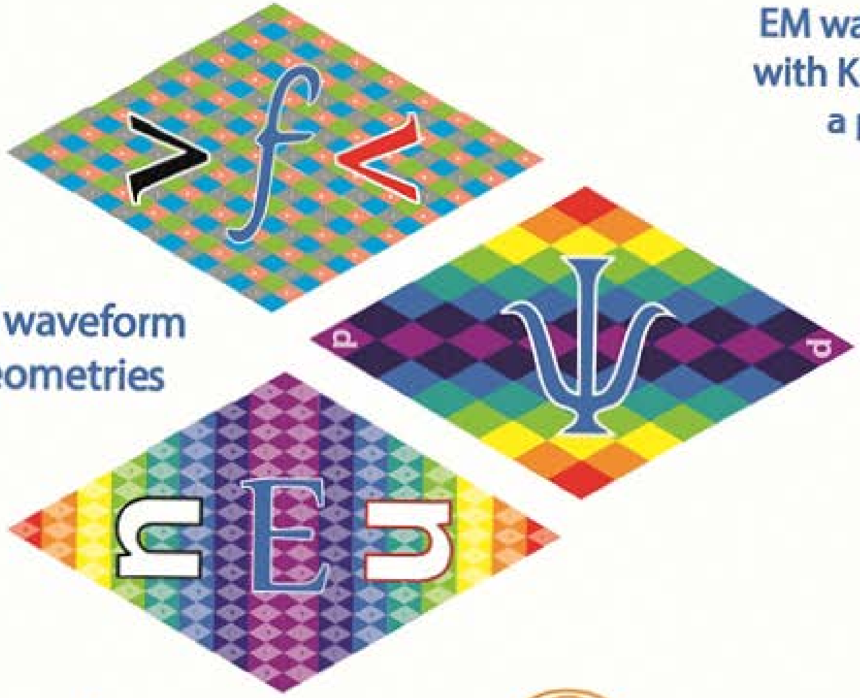

λ
f
λ
p
Ψ
p
Ω
E
Ω

EM waves are comprised of Photons
with Kinetic energies, momenta and
a photo-magnetic moment

Energy

$$p_\gamma = 2\pi[[m\Omega v]]$$

momenta

Photonic energy-momentum

Transverse
waveforms

Longitudinal
waveforms

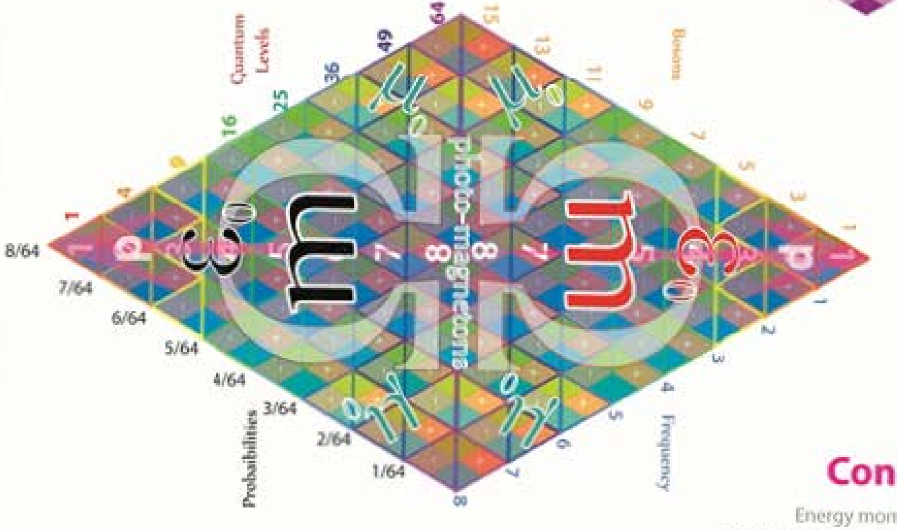

Quantum Levels
1
4
9
16
25
36
49
64
Bosons
15
13
11
9
7
5
3
1
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
Probabilities
1
2
3
4
5
6
7
8
Frequency
photo-magnetons
m
m
ε_0
μ_0
p
p

Scalar Planck quanta
per second
is the scalar source of all
linear energy-momenta

$$p^2$$

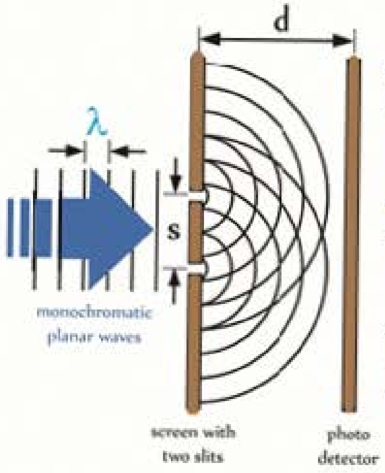

d
λ
s
monochromatic planar waves
screen with two slits
photo detector

Interference is the result of
KEM waveform superpositioning

Conservation of Energy momenta

Energy momenta a conserved quantity in physical systems (laws of conservation),
meaning that if a closed system is not affected by external forces, its total momentum cannot change

Tetryonics 38.06 - EM Wave properties

Copyright ABRAHAM [2008] - all rights reserved

Statistical probabilities of energy distributions

form the basis for quantum mechanical probabilities and wavefunctions

$E = 2mv$

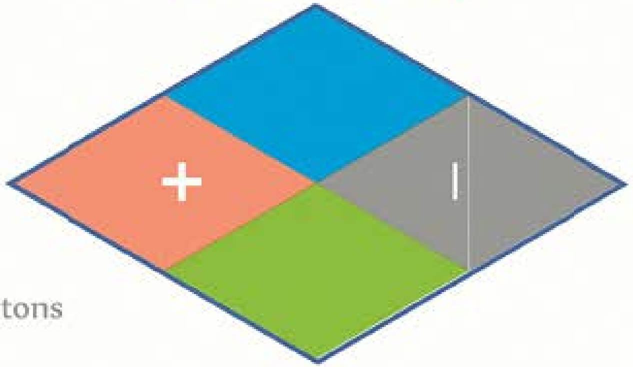

$E = Mc^4$

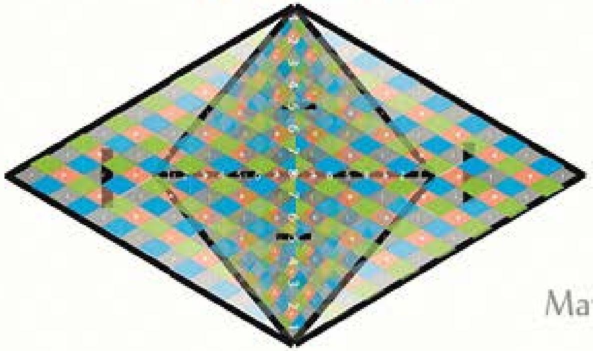

All ENERGY in motion has probabilistic distribtutions of energy quanta
resulting from the equilateral Tetryonic geometries
of electromagnetic mass-ENERGY-Matter

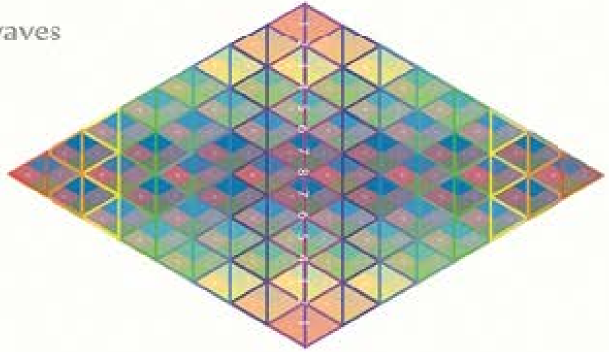

$p^2 = E = mv^2$

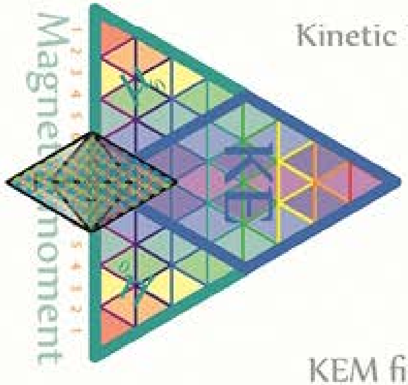
Magnetic moment
KE
Kinetic Energies
KEM fields

$Mv^2 = KEM$

$KE = \frac{1}{2}Mv^2$

Transverse

Bosons

ODD $\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

Boson quanta (ν)

ν=128

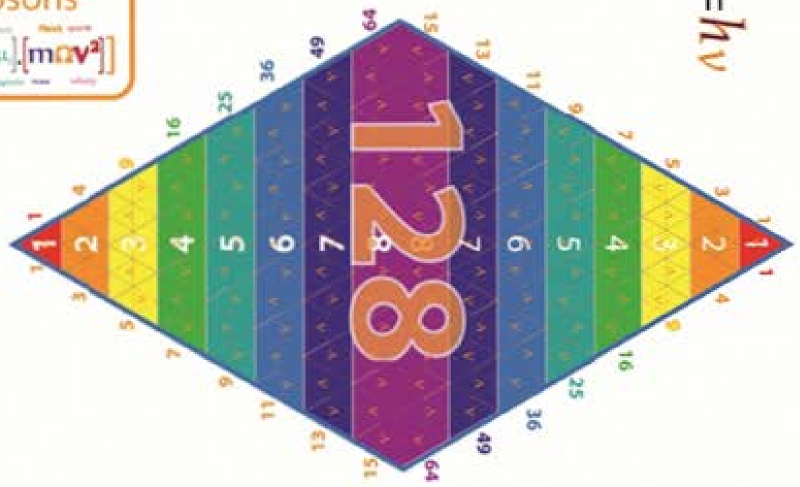

128
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
1
4
9
16
25
36
49
64
15
13
11
9
7
5
3

$E = h\nu$

Group Phase Velocity

(with respect to E field)

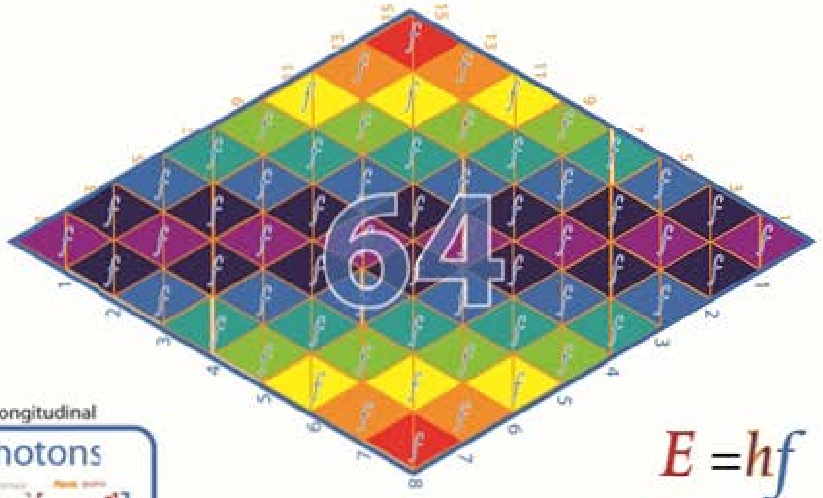

64
f
1
2
3
4
5
6
7
8
9
11
13
15

$E = hf$

Longitudinal

Photons

$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

64

Photon Frequency (f)

EM waves

EM Field Planck quanta

EVEN $\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$

EM waves ElectroMagnetic mass velocity

Planck quanta

ν=128

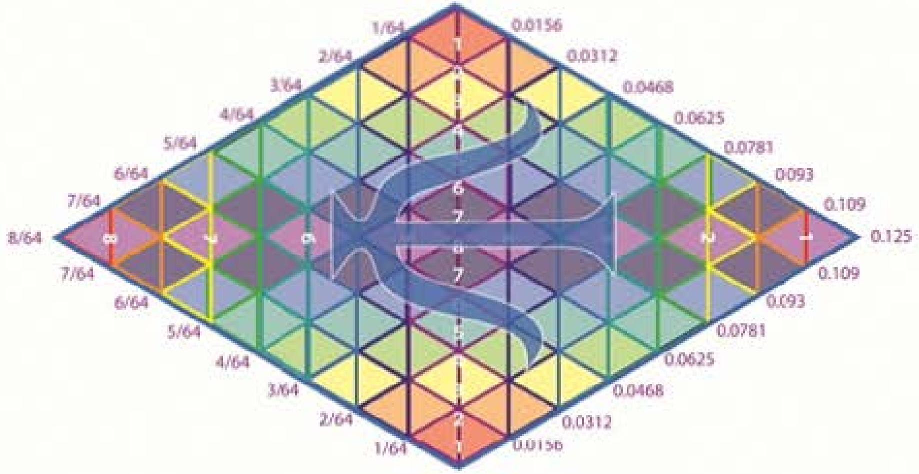

1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
ψ

EM wavefunction

f = 64

Photon frequency

Waveform probability distribution

[Amplitude]^2

n/64 = n/[8^2]

Wave functions

All EM waves and Photons exhibit quantum levels of mass-energy momenta determined by their constituent bosons

Transverse EM Waveforms

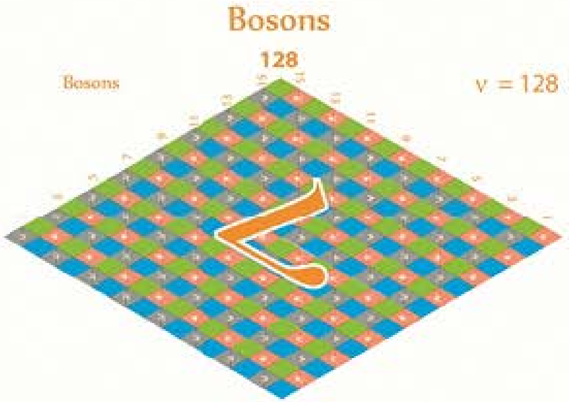
Bosons
128
Bosons
ν = 128

$$E = h\nu$$

[Planck quanta are single-quanta Bosons]

All EM waves are comprised of Longitudinal Photons which in turn are made up of Transverse Bosons

[Photons are neutral dual-quanta Bosons]

$$E = hf$$

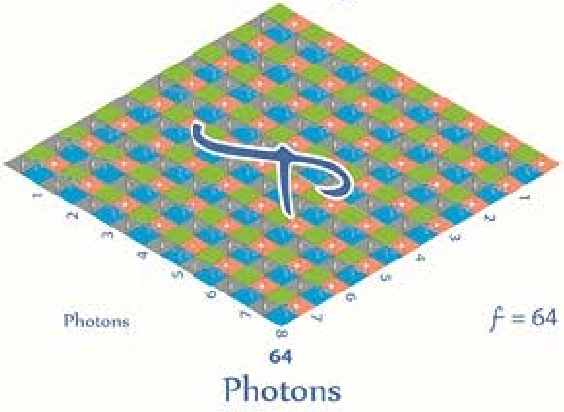
1
2
3
4
5
6
7
8
Photons
64
Photons
f = 64

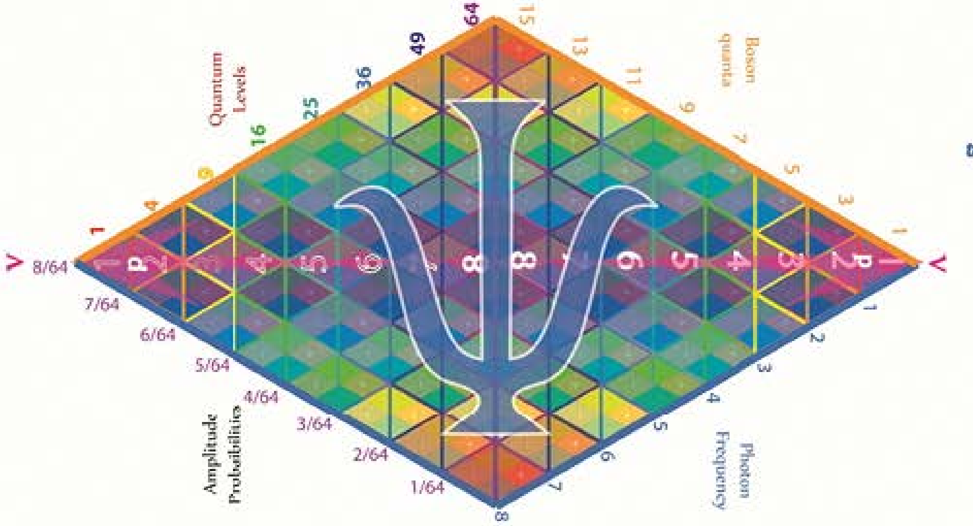
Quantum Levels
1
4
9
16
25
36
49
64
Boson quanta
15
13
11
9
7
5
3
1
ν
ν
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
Amplitude Probabilities
1
2
3
4
5
6
7
8
Photon Frequency

Longitudinal EM waveforms

All electromagnetic waves exhibit wave probabilities and amplitude functions determined by their constituent photons

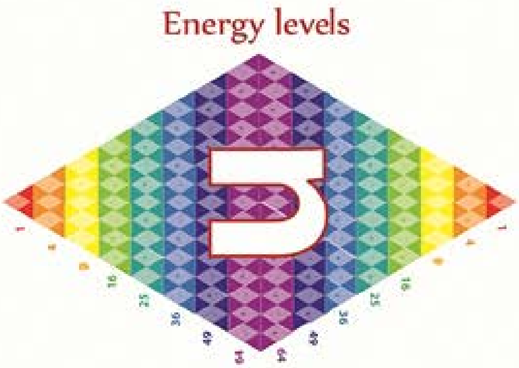
Energy levels
1
4
9
16
25
36
49
64

Squared numbers

Wave-particle duality holds that light and matter simultaneously exhibit properties of waves as well as particle-like properties (photons).

This concept is a consequence of the mathematical treament of mass-energy physics in quantum mechanics and a comprehensive explanation of this duality has been elusive to Physics since its discovery.

Any attmpt to develop an unified quantum theory must explain the root cause and processes behind Wave-Particle duality and must also explain:

Diffraction, Deflection and Reflection along with Bosons, Photons, EM radiation, energy quantisation and mass-Matter.

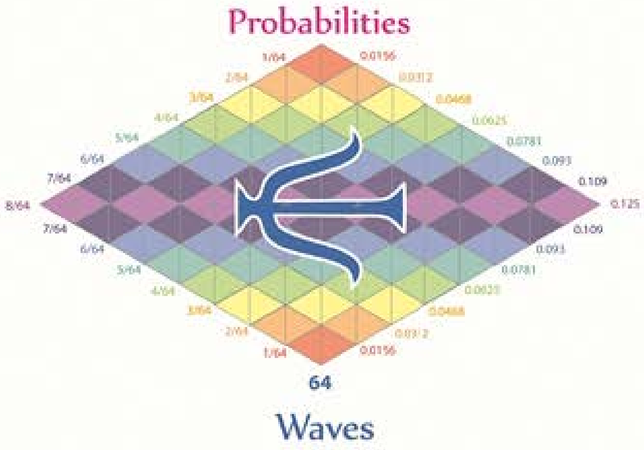
Probabilities
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
64
Waves

EM Wavefunctions

The wavefunction for a system contains all the information about the quantum state that a particle/system is in and gives a complete description of that part of the world at any one particular instant in time.

Transverse mass-energy distrbutions

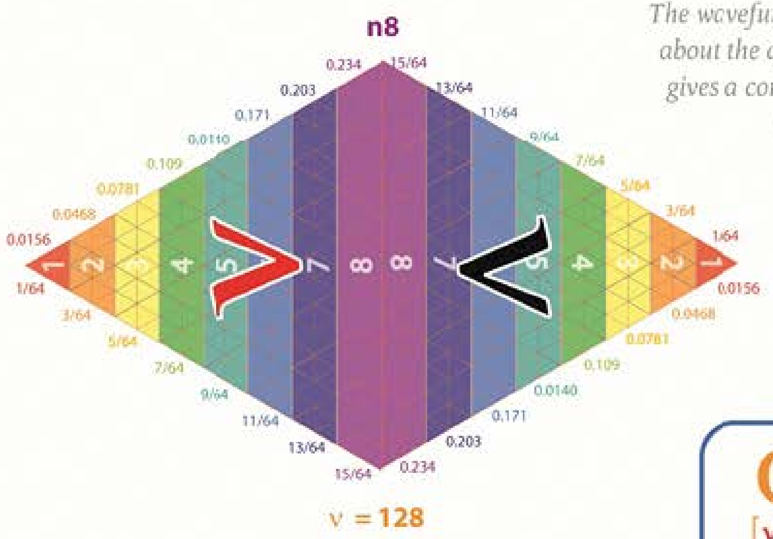
n8
0.234
15/64
0.203
13/64
0.171
11/64
0.0140
9/64
0.109
7/64
0.0781
5/64
0.0468
3/64
0.0156
1/64
1 2 3 4 5 6 7 8 8 7 6 5 4 3 2 1
v
v
v = 128

Longitudinal mass-energy distributions

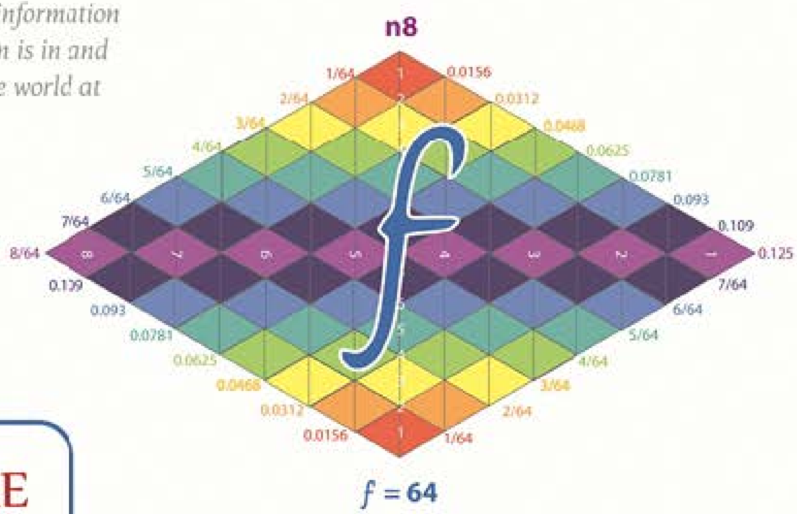
n8
1/64
0.0156
2/64
0.0312
3/64
0.0468
4/64
0.0625
5/64
0.0781
6/64
0.093
7/64
0.109
8/64
0.125
8 7 6 5 4 3 2 1
f
f = 64

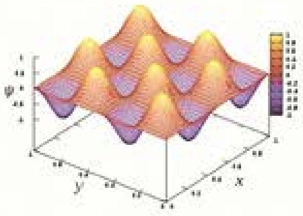

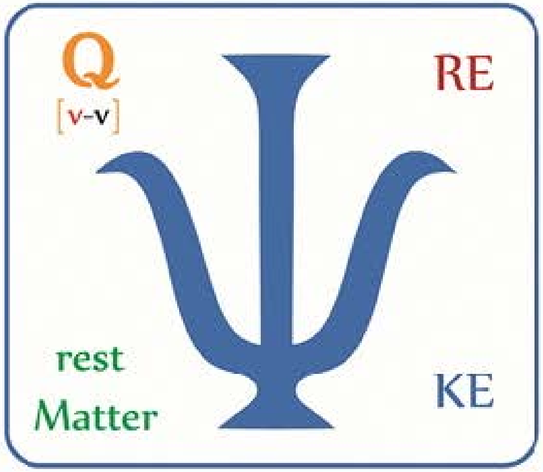
Q
[v-v]
RE
Ψ
rest
Matter
KE

The wavefunction of an EM wave (KEM) geometry is distinct from that of rest mass-Matter toplogies due to the number of Planck quanta contained in it but both can be modelled using Tetryonic geometry

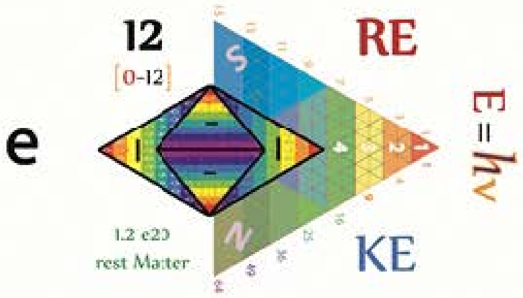
12
[0-12]
RE
e
E = hv
1.2 e20
rest Matter
KE

Quantum Energy Levels

The wavefunction itself is often said to be un-observable. In fact, it can modelled as it is a reflection of the quantas making up the quantum state of any particle and is complex-valued.

As the system evolves over time, the wavefunction also changes, so it can be written as a function of time Ψ(t).

$$\int_{-\infty}^{\infty} |\Psi|^2 \mathrm{d}x = 1.$$

The absolute square of the function must be normalizable

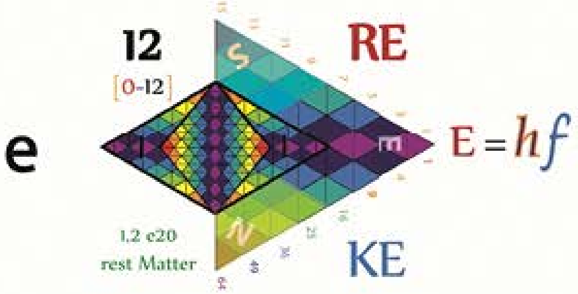
12
[0-12]
RE
e
E = hf
1.2 e20
rest Matter
KE

Wave-Particle Probabilities

The absolute square of the wavefunction is a probability density (the area of highest probability for a measurement to take place.

For example, if the wavefunction is expressed in real space and our system is a particle, the absolute square gives a probability density for the position of the system. Integrating this probability density between some bounds will give the probability that the particle will be found in that region when its position is measured

Wave~Particle energies

Equilateral energy is the foundational geometry
of Kinetic EM Waveforms [WAVES]
and standing wave Matter [PARTICLES]

**All Matter in motion exhibits
a Wave-Particle duality**

Everything is made up of
***charged mass-energy
quanta***

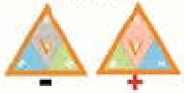

ZPFs combine in ODD
numbers to form
Bosons

Bosons combine to form
Quantum Levels

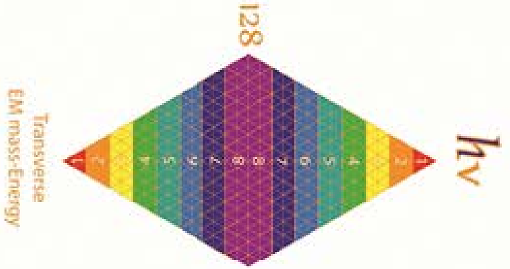
128
Transverse
EM mass-Energy
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
hv

Wavelength
v
Frequency

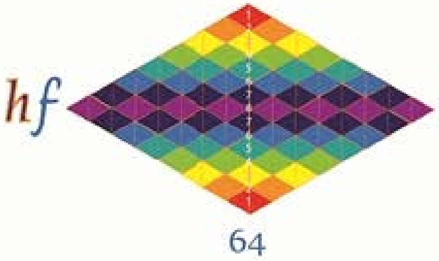

Longitudinal
EM mass-Energy

The quantum Planck quanta tesslation
of Matter can be accurately modelled with
Wave Probability mechanics

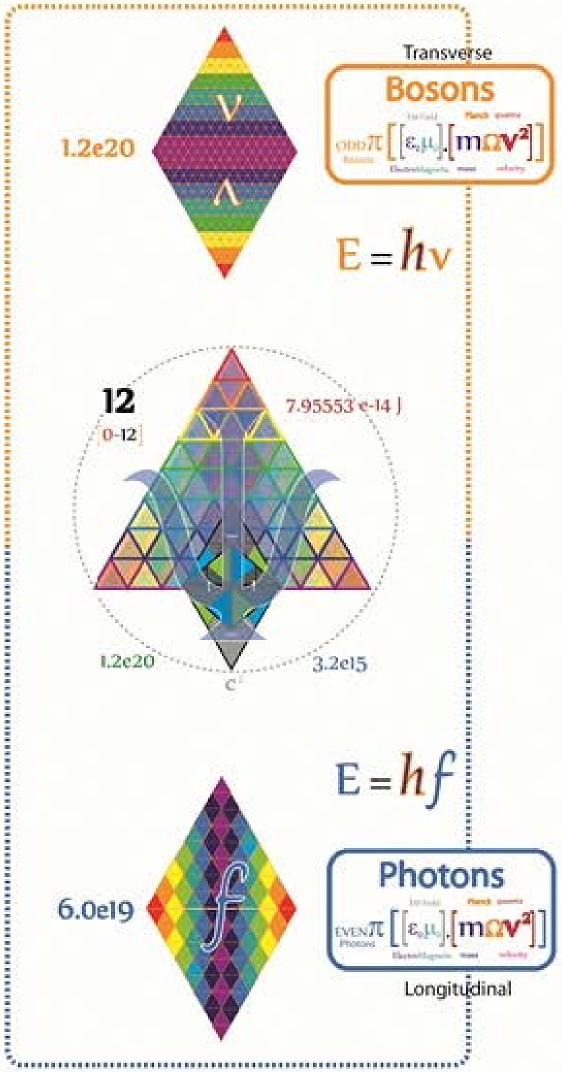
Transverse
1.2e20
v
v
12
[0-12]
7.95553 e-14 J
1.2e20
c
3.2e15
6.0e19
f
Longitudinal

Bosons

$\text{ODD}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

$E = hv$

$E = hf$

Photons

$\text{EVEN}\,\pi\,[[\varepsilon_0\mu_0].[m\Omega v^2]]$

A n8 Electron is a massive particle
which has a Kinetic EM energy field
[6.4e15 planck quanta]

*2D EM field geometries
can combine to form
3D Tetryonic Matter topologies*

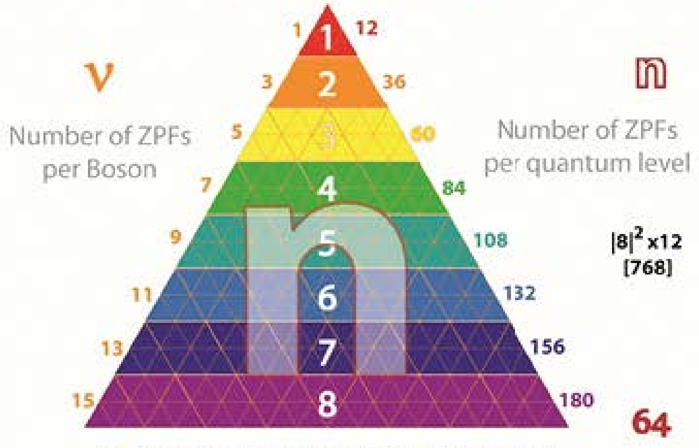
v
Number of ZPFs
per Boson
n
Number of ZPFs
per quantum level
1 2 3 4 5 6 7 8
1 3 5 7 9 11 13 15
12 36 60 84 108 132 156 180
|8|² x12
[768]
64

Quantum levels are a transverse measurements of energy

Equivalently, all Matter & EM waves can be viewed either in terms of their transverse boson or longitudinal photon content and/ or their associated mass-energy momenta geometries

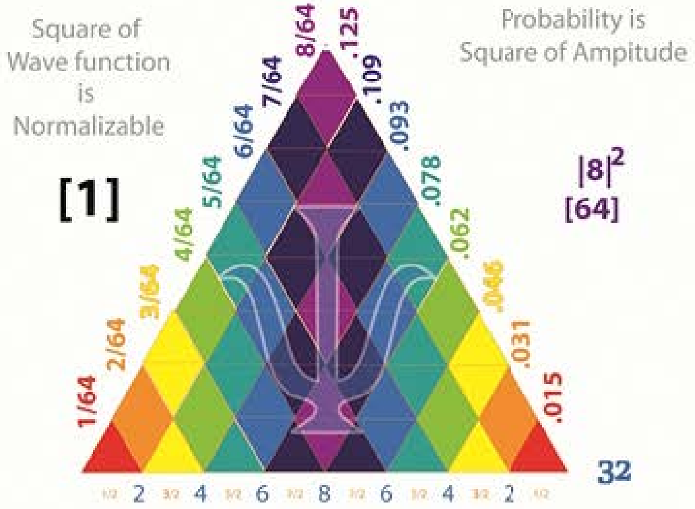
Square of
Wave function
is
Normalizable
[1]
Probability is
Square of Amplitude
|8|²
[64]
1/64 2/64 3/64 4/64 5/64 6/64 7/64 8/64
.125 .109 .093 .078 .062 .046 .031 .015
32
1/2 2 3/2 4 5/2 6 7/2 8 7/2 6 5/2 4 3/2 2 1/2

Wave functions are a linear measurement of photonic energy

Tetryonics 38.11 - Wave~Particle energies

Wave-particle Probabilities

All Particles in motion exhibit a Wave-Particle duality due to the 2D waveform geometry of its KEM wave and the lorentz velcoity invariant 3D standing wave geometry of its Matter

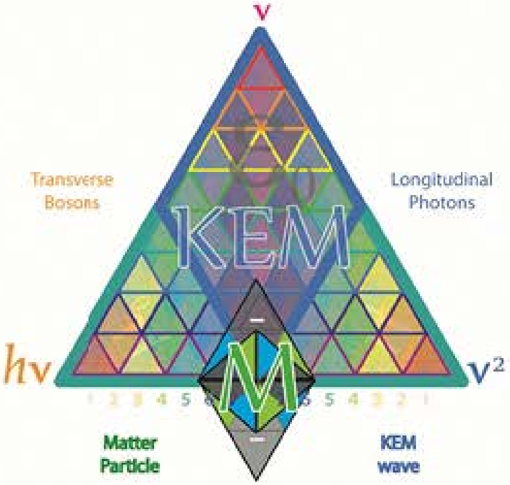
v
Transverse Bosons
Longitudinal Photons
KEM
hv
M
v^2
1 2 3 4 5 6 5 4 3 2 1
Matter Particle
KEM wave

Equilateral energy tessellation provides the basis for the probabilistic statisical mechanics and math for all of Quantum Mechanics

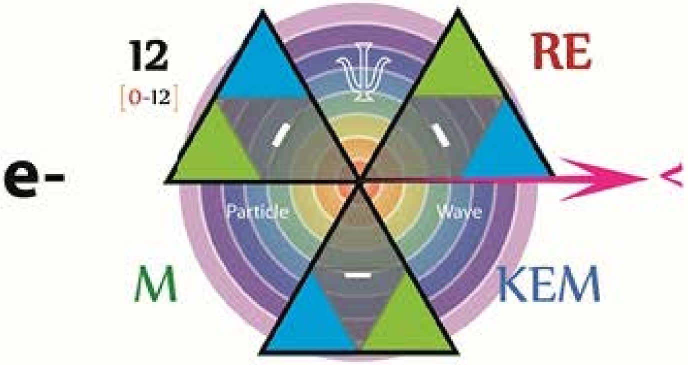
12
[0-12]
RE
e-
v
Particle
Wave
M
KEM

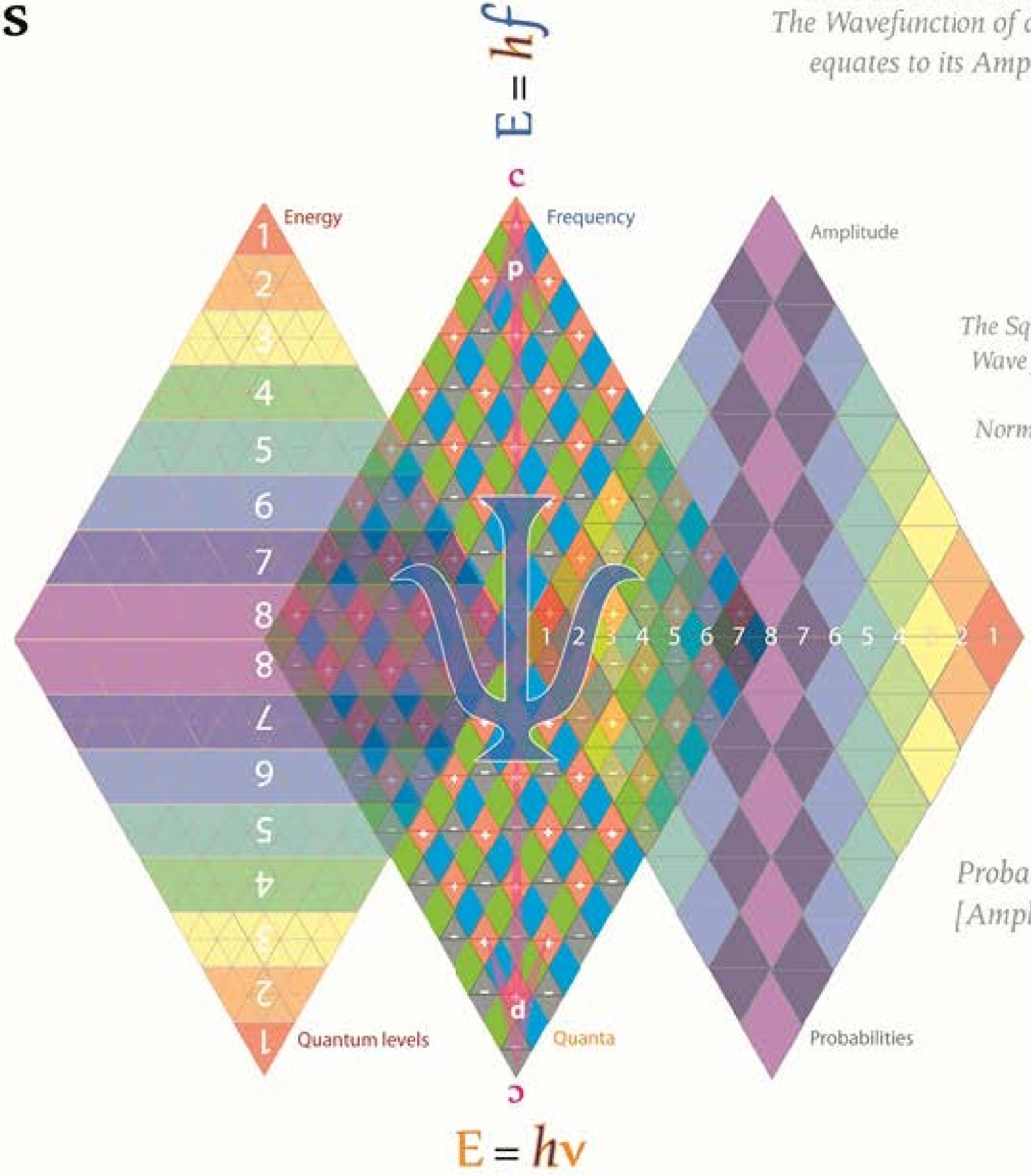
E = hf
c
Energy
Frequency
Amplitude
1
2
3
4
5
6
7
8
8
7
6
5
4
3
2
1
p
d
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1
Quantum levels
Quanta
Probabilities
c
E = hv

Schrodinger Wavefunction
The Wavefunction of a Particle equates to its Amplitude

The Square of a Wave function is Normalizable

Probability = $[Amplitude]^2$

Born Probability Rule
The Probability of finding a Particle is the Square of its Amplitude

White Light mass-energy momenta waveforms

White light is comprised of
superpositioned EM waves

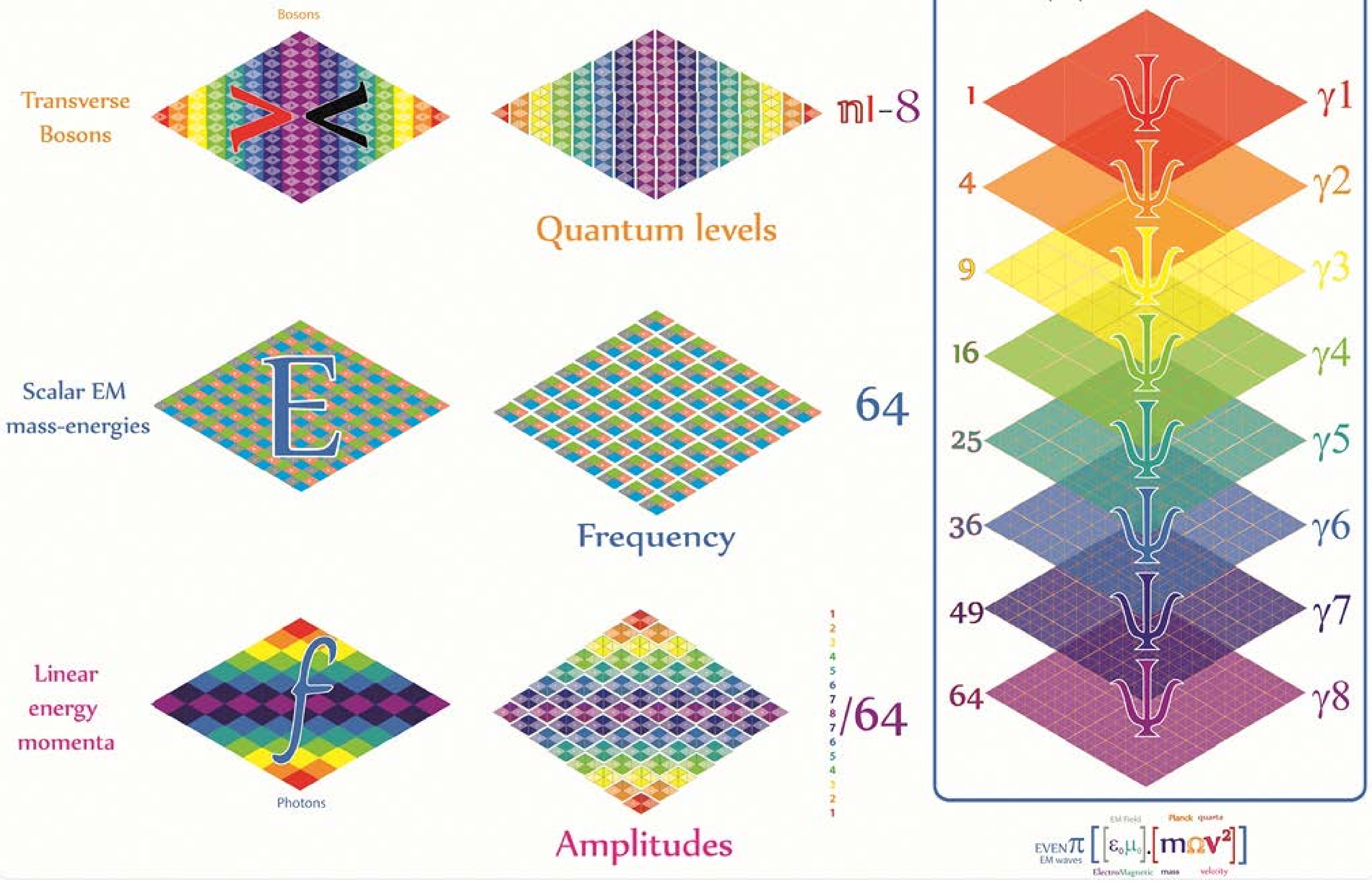
Bosons
Transverse
Bosons
Quantum levels
nl-8
Scalar EM
mass-energies
E
Frequency
64
Linear
energy
momenta
f
Photons
Amplitudes
1
2
3
4
5
6
7
8
7
6
5
4
3
2
1
/64
1
4
9
16
25
36
49
64
γ1
γ2
γ3
γ4
γ5
γ6
γ7
γ8
Ψ
EM field
Planck
quanta
EVEN
EM waves
ElectroMagnetic
mass
velocity

$$\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Tetryonics 38.13 - White Light waveform geometries

Copyright ABRAHAM [2008] - all rights reserved

Bosons and Photons

ε_0
μ_0
c^3

EM waves are comprised of longitudinal [E=hf] Photons which in turn can also be measured as transverse [E=nhv] Bosons

Electro-Magnetic Fields

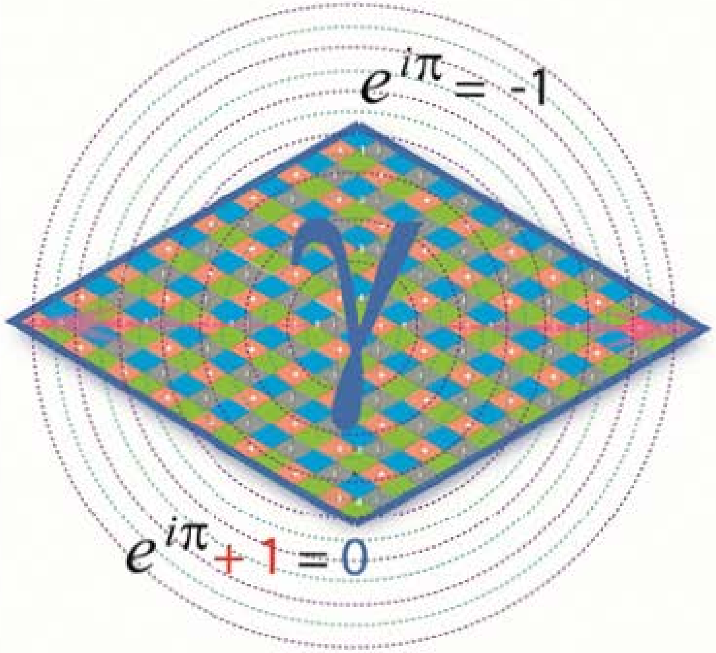

$e^{i\pi} = -1$

$e^{i\pi} + 1 = 0$

As described by Lorentz co-ordinate transforms the neutral quantum coin [quoin] geometry of all photons and EM waves have mirror E-field charges on each of their opposing fascia

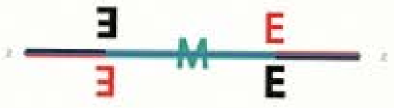

2D planar mass-energy momenta waveforms

$$\underset{\text{Photons}}{2\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

Photons and EM waves have dual c^2 space-time geometries
(8.987551787 e16 m /s)

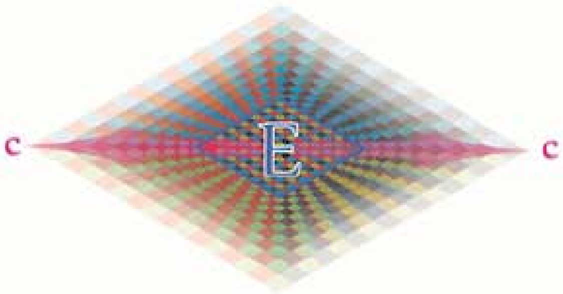

Photons and EM waves propagate outwards from their source at 'c'
(299,792,458 m/s)

$$\underset{\text{EM waves}}{\text{EVEN}\,\pi}\left[\overset{\text{EM Field}}{\underset{\text{ElectroMagnetic}}{[\varepsilon_0\mu_0]}}.\left[\underset{\text{mass}}{m}\overset{\text{Planck}}{\Omega}\overset{\text{quanta}}{\underset{\text{velocity}}{v^2}}\right]\right]$$

EM waves are detected as alternating radial electric E-wave radiation fields

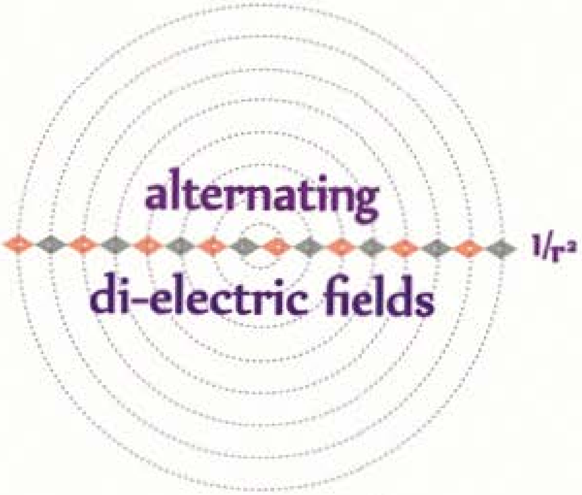

EM wave radiation

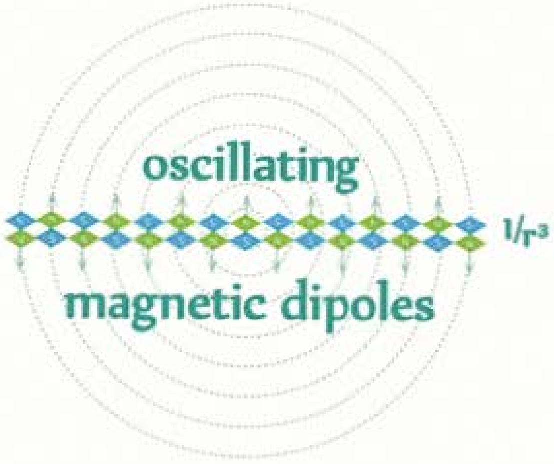

The Magnetic vector of EM energy momenta propagate orthagonally to their E-fields

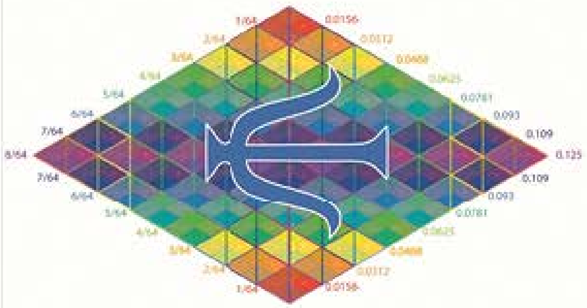
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125

EM wave patterns

Polarised Photons and EM waves produce linear interference patterns

1
2
3
4
5
6
7
8
7
6
5
4
3
2
1

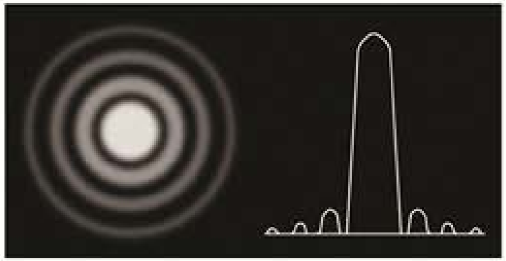

Diffraction occurs whenever propagating waves appear to bend around small obstacles in their path and spread out as smaller waves waves past any openings, its effects are generally most pronounced for waves whose wavelength is roughly similar to the dimensions of the diffracting objects.

If the obstructing object provides multiple, closely spaced openings, a complex pattern of varying intensity can result.

This is due to the superposition, or interference, of different parts of a wave that travels to the observer by different paths

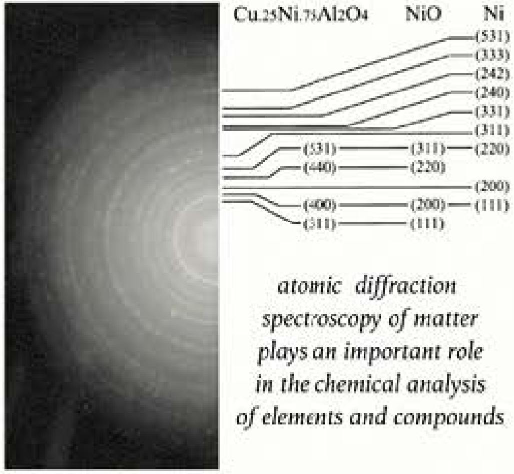
Cu.25Ni.75Al2O4
NiO
Ni
(531)
(333)
(242)
(240)
(331)
(311)
(531)
(311)
(220)
(440)
(220)
(200)
(400)
(200)
(111)
(311)
(111)

atomic diffraction spectroscopy of matter plays an important role in the chemical analysis of elements and compounds

Diffraction occurs with all waves, including sound waves, water waves, as well as with electromagnetic waves such as visible light, X-rays and radio waves.

Un-Polarised Photons and EM waves produce circular diffraction patterns

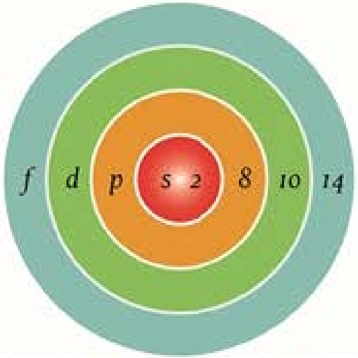
f
d
p
s
2
8
10
14

Richard Feynman once commented that:
"No-one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them."

Light waves

In order to establish his wave theory of light Young had to overcome the century-old view, expressed by Sir Isaac Newton' in his treatise "Optics", that light is a particle of colour.

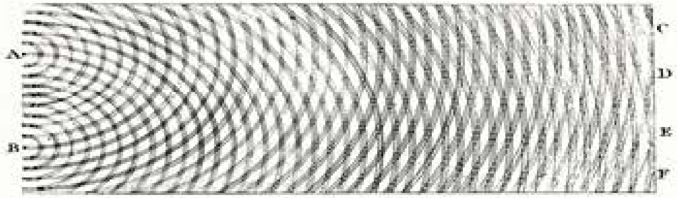

Thomas Young studied the interference of light waves by shining light through a screen with two slits equally separated, where the light emerging from the two slits would spread out and eventually superposition [overlap] each other to produce a distinct interference pattern

Thomas Young

(13 June 1773 – 10 May 1829)

$$E = mv^2$$

"The term energy may be applied, with great propriety, to the product of mass or weight of a body, into the square of the number expressing its velocity.

Thus, if the weight of one ounce moves with a velocity of a foot in a second, we call its energy 1; if a second body of two ounces has a velocity of three feet in a second, its energy will be twice the square of three, or 18."

$$KE = \frac{1}{2}Mv^2$$

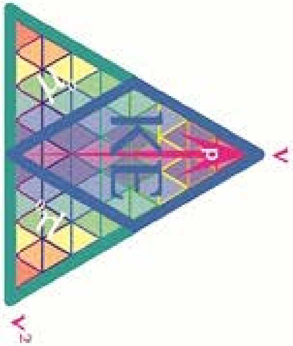

All Light propagates at the speed of EM ENERGY

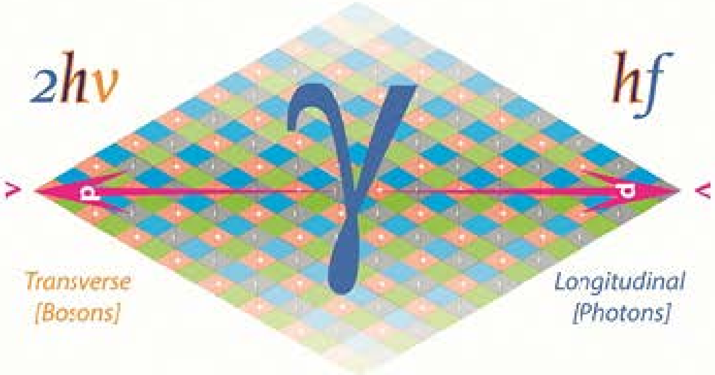

All EM waves have energy momenta geometries whose physical properties can be modeled with statisical probabilities due to their equilateral wavefunctions

Tetryonics reveals the true geometry of Light

and dispels the misconceptions surrounding its physical properties of wave-length, energy momenta, frequency and associated wave-function

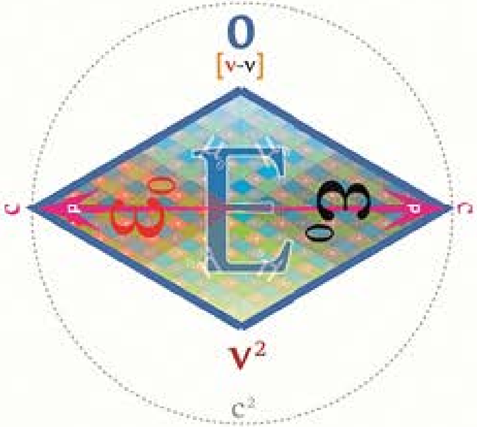

$$E_\gamma = 2mv^2$$

Photons are radiative Kinetic EM mass-energies

$$p^2 = KEM = Mv^2$$

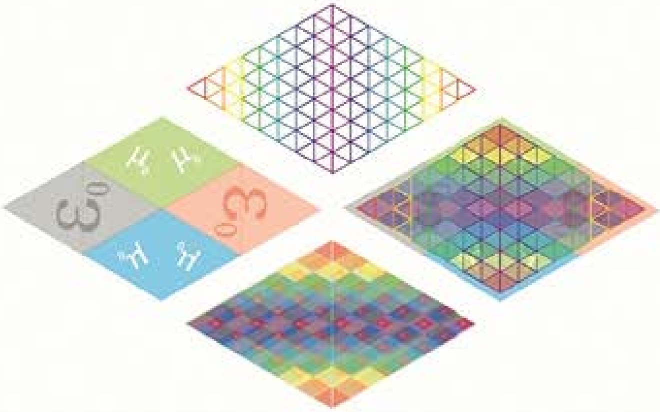

Photon-EM Wave Superpositioning

In physics, the Superposition principle, also known as superposition property, states that, for all linear systems, the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually.

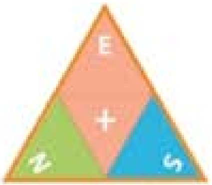

Positive quantum charge state
is one side of the quantum
energy momenta quoin

$$F(\psi_1+\psi_2)=F(\psi_1)+F(\psi_2)$$

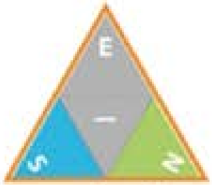

Negative quantum charge state
is the other side of the quantum
energy momenta quoin

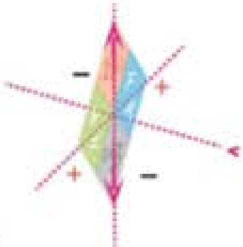

The linear momenta of transverse waves is orthagonal to the direction of wave propagation

Photons are neutral EM wavepackets

whose chiral quantum charges can be determined through Lorentz co-ordinate transformations

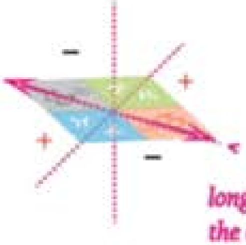

The linear momenta of longitudinal waves is co-linear to the direction of wave propagation

transverse Radio waves

longitudinal Spectral lines

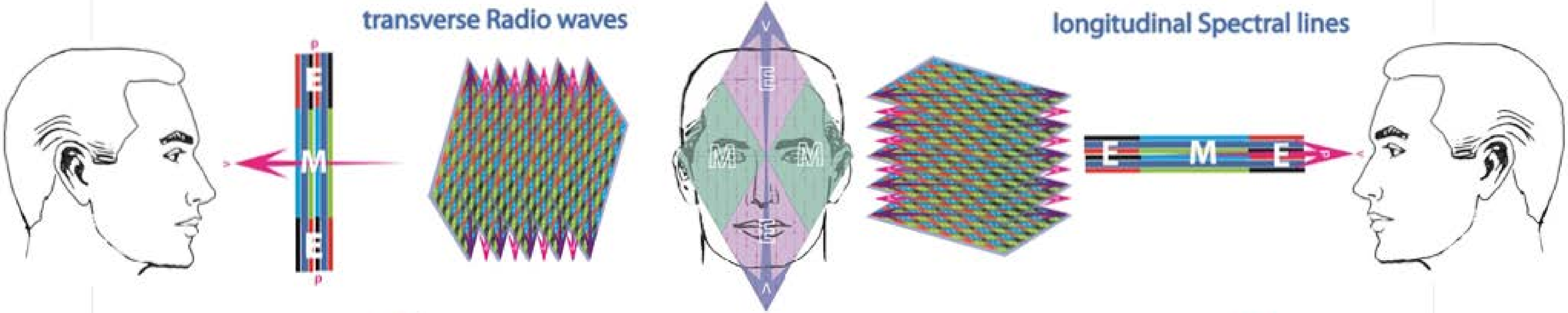

mc^{2^2} *Superpositioned EM waves can be likened to the cells of a chemical battery* mv^4

E^2

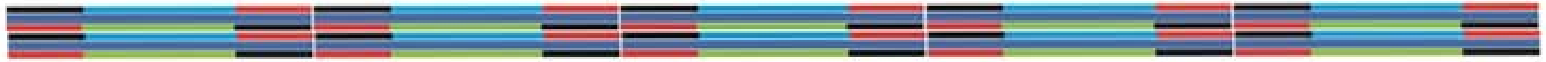

p^4

It is the superpositioning of longitudinal EM waveforms that creates white light and linear interaction-at-a-distance forces

Copyright ABRAHAM [2008] - all rights reserved

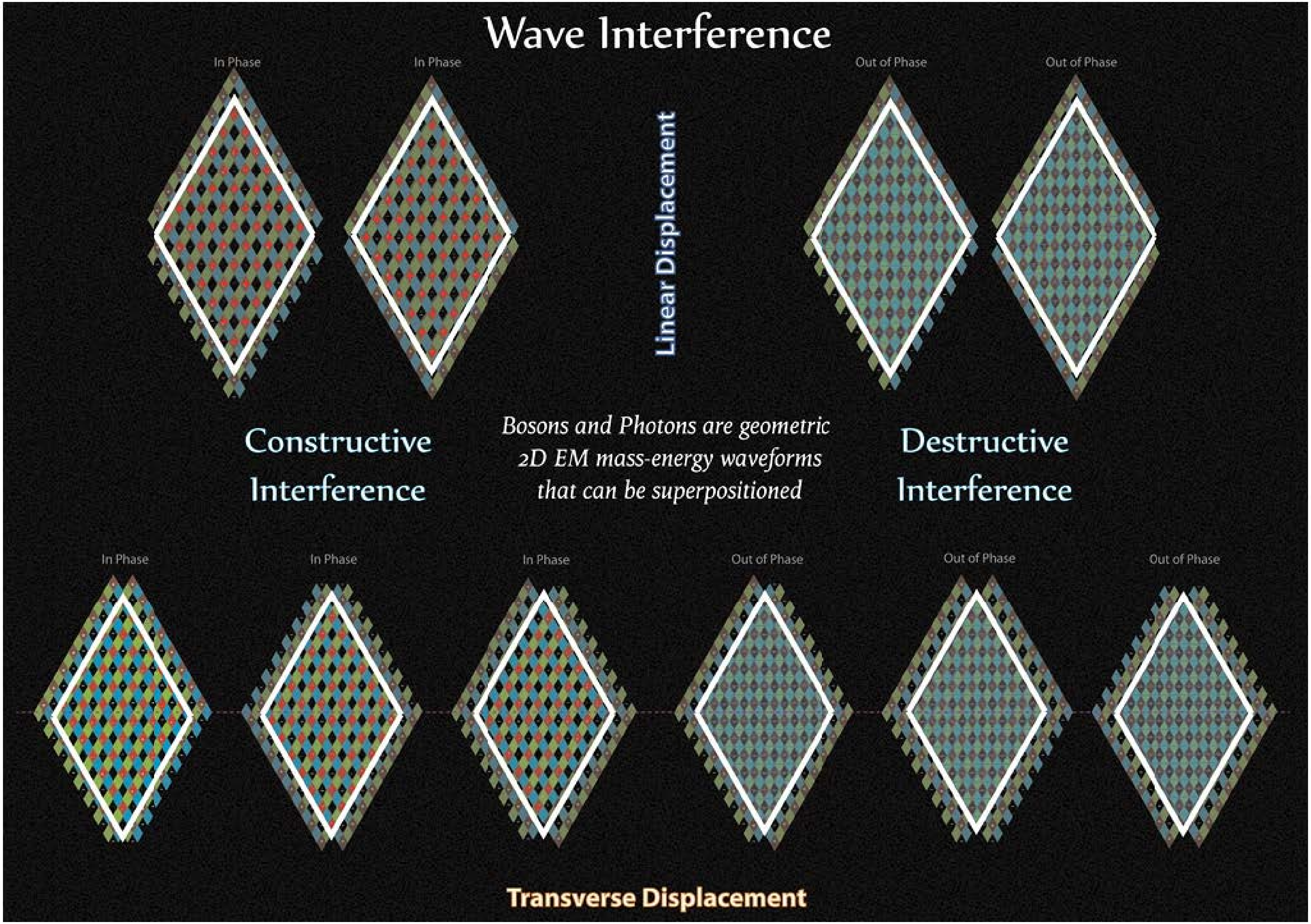

Wave Interference

Constructive Interference

Bosons and Photons are geometric 2D EM mass-energy waveforms that can be superpositioned

Destructive Interference

In Phase
In Phase
Linear Displacement
Out of Phase
Out of Phase
In Phase
In Phase
In Phase
Out of Phase
Out of Phase
Out of Phase
Transverse Displacement

Tetryonics 39.05 - Wave Interference - EM waves

Wave phase interference

The interference patterns produced by the photon intensities of superpositioned EM waveforms have been historically interpretted as being the result of transverse radial wave fronts

Radial modelling of the E-field propagation of EM waves and photons does not reflect their true equilateral energy momenta geometries

This is now shown to be incorrect

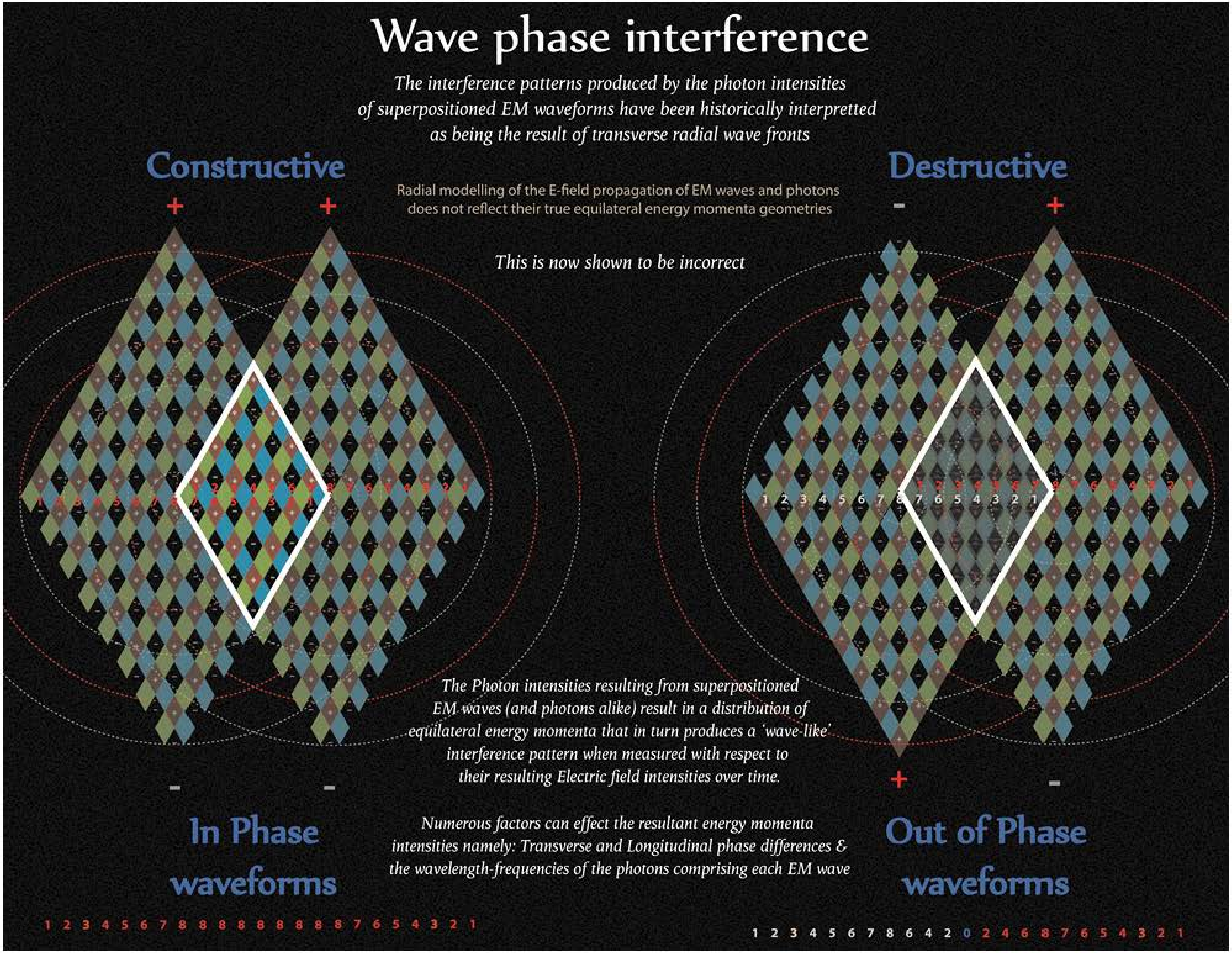
Constructive
+
+
-
-
In Phase waveforms
1 2 3 4 5 6 7 8 8 8 8 8 8 8 8 8 7 6 5 4 3 2 1
Destructive
-
+
+
-
Out of Phase waveforms
1 2 3 4 5 6 7 8 6 4 2 0 2 4 6 8 7 6 5 4 3 2 1

The Photon intensities resulting from superpositioned EM waves (and photons alike) result in a distribution of equilateral energy momenta that in turn produces a 'wave-like' interference pattern when measured with respect to their resulting Electric field intensities over time.

Numerous factors can effect the resultant energy momenta intensities namely: Transverse and Longitudinal phase differences & the wavelength-frequencies of the photons comprising each EM wave

Tetryonics 39.06 - Wave phase interference

Transverse EM mass-energies

Matter in motion

Longitudinal EM mass-energies

$E = h\nu$

Historically modeled through relativistic stress energy tensors
all Matter in motion is possessive of distinct energy waveforms:
a standing wave mass-energy geometry [Matter particle] &
a divergent kinetic mass-energy field [KEM wave]

$E = hf$

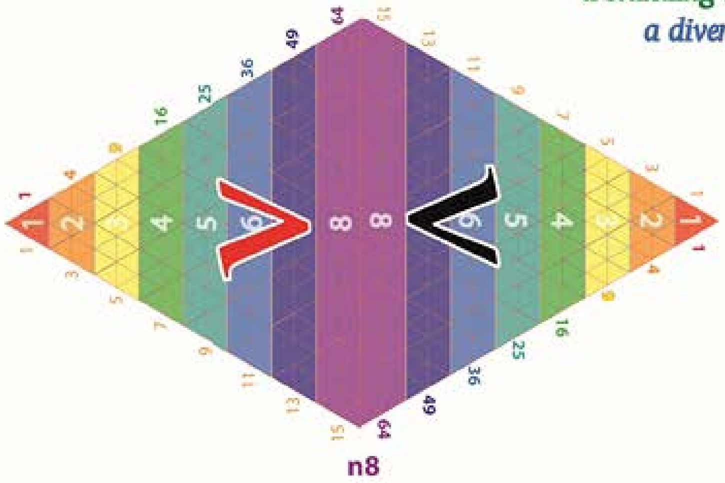
1
2
3
4
5
6
v
8
8
v
6
5
4
3
2
1
1
4
9
16
25
36
49
64
3
5
7
9
11
13
15
n8

$T\mu\nu$

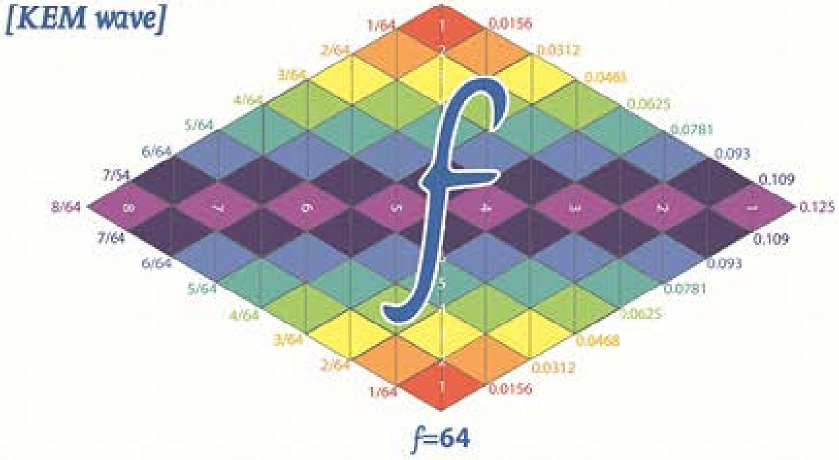
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
f
f=64

Negative charges in motion

EM induction and forces are effected by
neutral bi-directional KEM mass-energy fields
with interactive energy momenta

Positive charges in motion

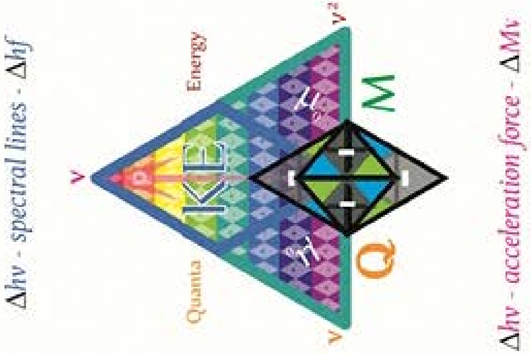
Δhν - spectral lines - Δhf
v
Energy
v²
M
KE
Q
Quanta
v
Δhν - acceleration force - ΔMv

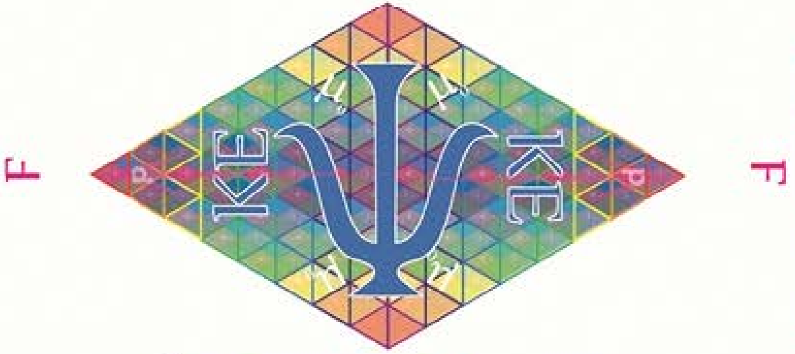

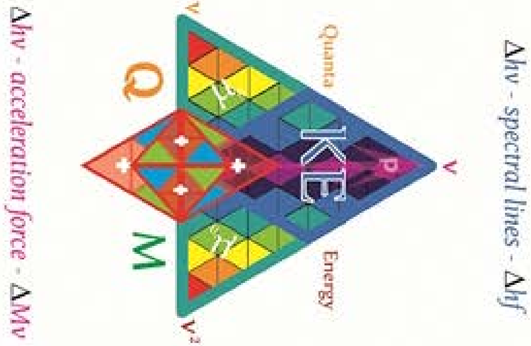
Δhν - acceleration force - ΔMv
v
Q
Quanta
KE
M
Energy
v²
Δhν - spectral lines - Δhf

unidrectional negative KEM fields
with interactive energy momenta

electromagnetic waves

[neutral bosons and/or photons]

unidrectional positive KEM fields
with interactive energy momenta

Copyright ABRAHAM [2008] - all rights reserved

**All Matter in motion [or not]
has both an intrinsic Matter-wave function
and an extrinsic KEM mass-energy wave function**

Transverse boson wavefunction

EM Field Planck quanta

$$\text{ODD}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Bosons ElectroMagnetic mass velocity

$$E = n.h\nu$$

128

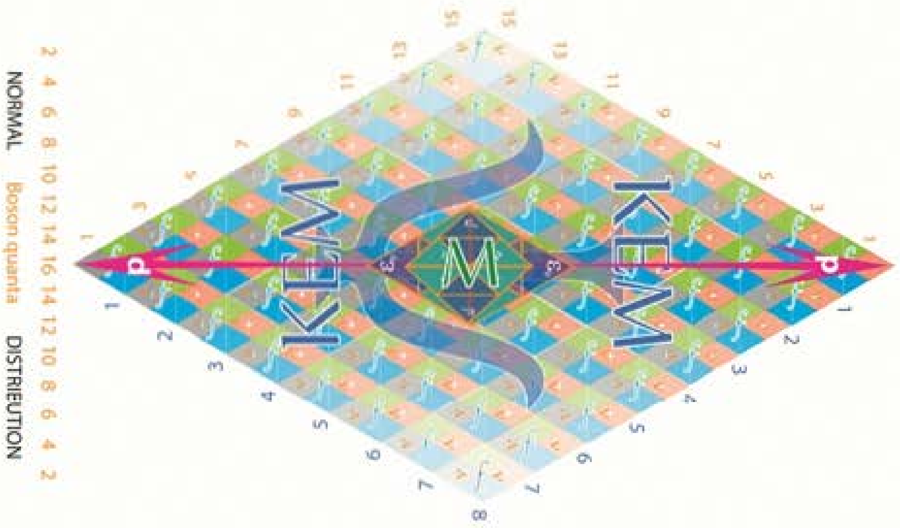

EVEN π EM waves

$$\text{EVEN}\,\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

EM Field Planck quanta
ElectroMagnetic mass velocity

64

EM Field Planck quanta

$$2\pi\left[[\varepsilon_0\mu_0].[m\Omega v^2]\right]$$

Photons ElectroMagnetic mass velocity

$$E = hf$$

Longitudinal photon wavefunction

Photon Intensity

*Photon Intensity is a result of the geometry of EM waves
[which in turn is directly proprtional to the velocity of the mass-Matter]
and can be measured in a number of vaying ways related to
the mass-energy momenta of the EM wave*

*Wavenumbers, frequency or wavelength
Transverse EM masses [BOSONS]
Longitudinal EM masses [PHOTONS]
Wavefunction probabilities [WAVE-FUNCTIONS]*

*The wave~particle nature of mass-Matter in motion
has been the subject of much debate since Lucretius in 55BC with the
debate intensitying with Newton's and Young's differing views and the
development of Planck-Einstein quantum mechanics of the photo-electric effect*

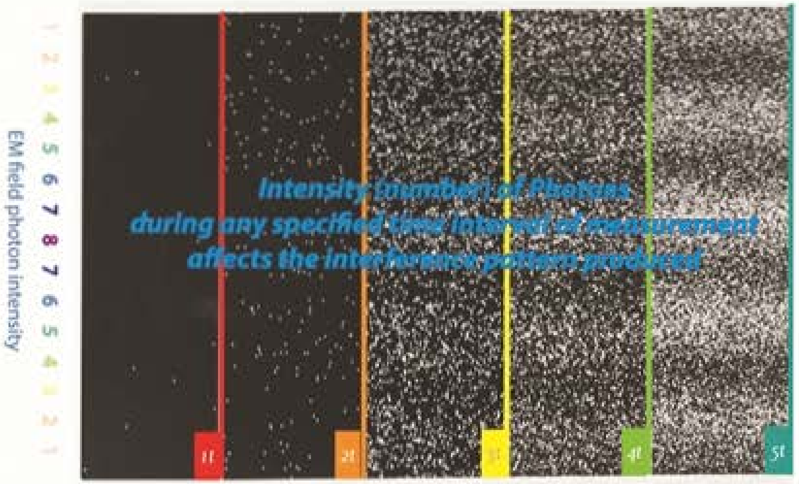

**Intensity [number] of Photons
during any specified time interval of measurement
affects the interference pattern produced**

*The measurement of Photons with the use
of Photo-multipliers and Charge coupled Devices
results in the measurement of the E field properties
of EM waves (producing the long confusing wave-particle
results obtained by diffraction gratings)*

*In turn these results have been historically misinerpreted
as a waveform property that cannot be attributed to
the particle properties of a Photon.*

Tetryonic geometry clears this matter up once and for all.

Interference Patterns

ε_0

When EM waves pass through the slits they are detected via their E-fields as photons of varying strength energy momenta producing the impression of a interference pattern

μ_0

$E = h\nu$

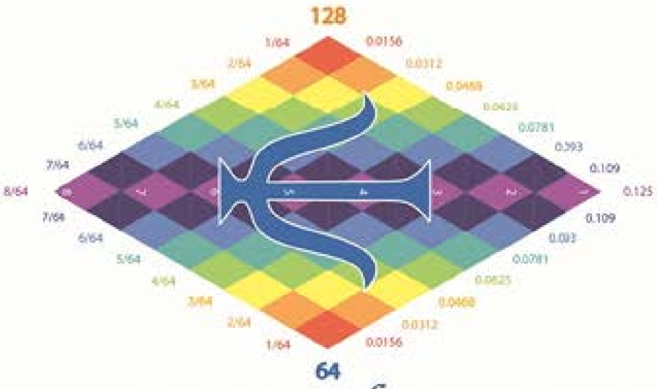
128
1/64
2/64
3/64
4/64
5/64
6/64
7/64
8/64
7/64
6/64
5/64
4/64
3/64
2/64
1/64
0.0156
0.0312
0.0468
0.0625
0.0781
0.093
0.109
0.125
0.109
0.093
0.0781
0.0625
0.0468
0.0312
0.0156
64

$E = hf$

Every EM wave is comprised of identical, specific wavelength photons which are arranged in a Normal distribution resulting from the electromagnetic Wavefunction with a peak value equal to the wave's amplitude

[the square root of the EM wave's Wavefunction/Probability amplitude]

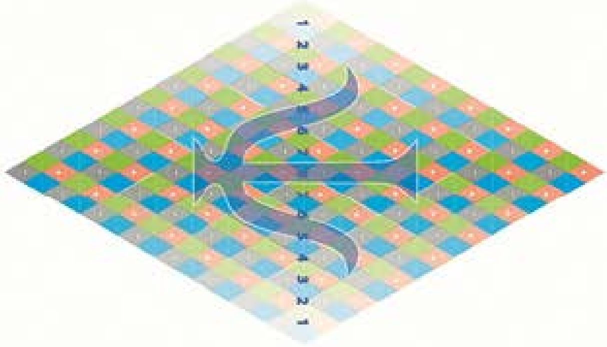
1 2 3 4 5 6 7 8 7 6 5 4 3 2 1

Photons impact screens and photo-detectors with intensities that are determined by their energy momenta distributions

ALL EM mass-ENERGY Matter geometries being comprised of energy-momenta quanta are capable of producing interference patterns

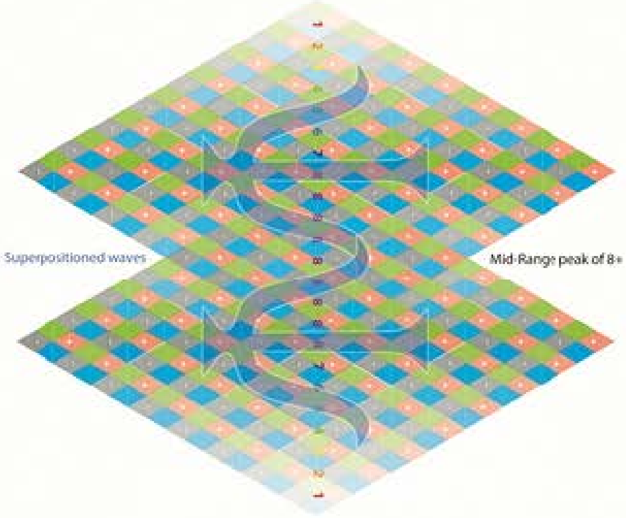
Superpositioned waves
Mid-Range peak of 8+

The resultant amplitudes are a direct result of the Phase of the Superpositioned photons within EM waves

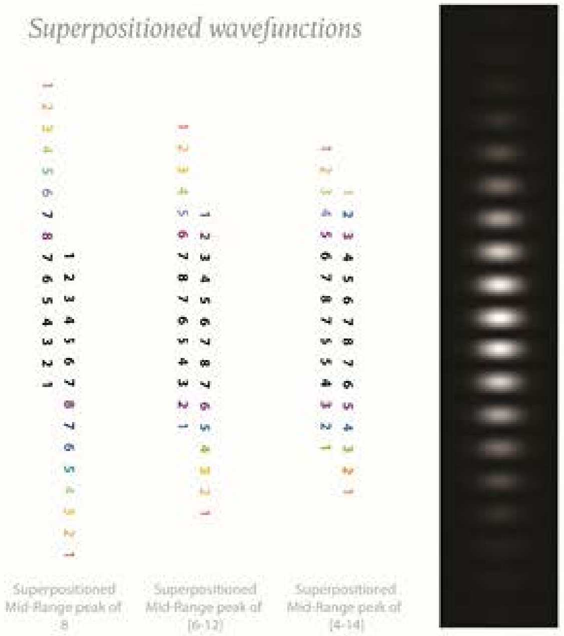
Superpositioned wavefunctions
Superpositioned Mid-Range peak of 8
Superpositioned Mid-Range peak of [6-12]
Superpositioned Mid-Range peak of [4-14]

The amplitudes of the detected photons in the resultant interference patterns are determined by their constructive/destructive superpositioning in each wave explaining the interference patterns currently accounted for by Young's wave theory

Copyright ABRAHAM [2008] - all rights reserved

Single slit experiment

If light consisted strictly of ordinary or classical particles, and these particles were fired in a straight line through a slit and allowed to strike a screen on the other side, we would expect to see a pattern corresponding to the size and shape of the slit. However, when this "single-slit experiment" is actually performed, the pattern on the screen is a diffraction pattern with a fairly narrow central band with ever diminishing bands parallel to it on either side

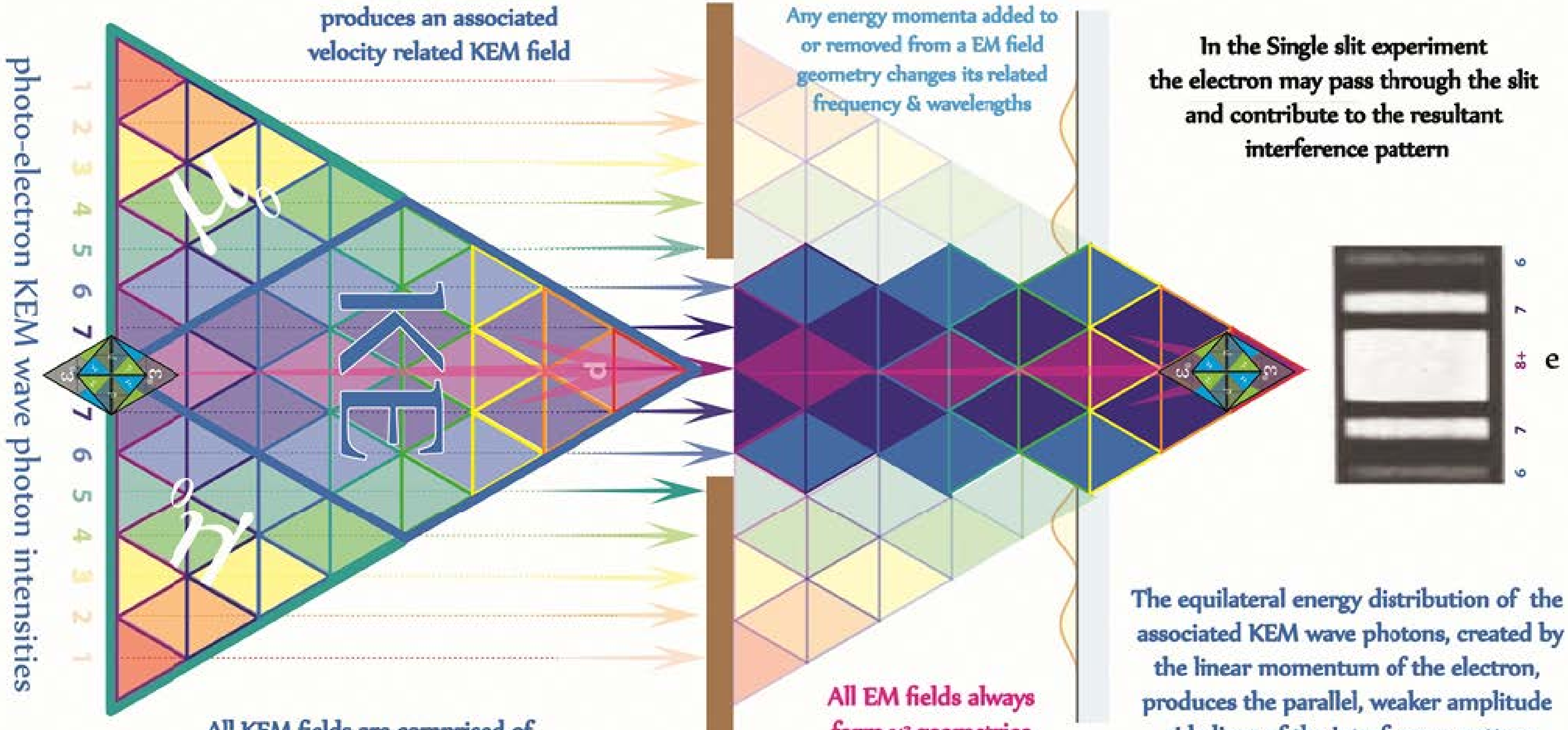

All matter in motion produces an associated velocity related KEM field

Any energy momenta added to or removed from a EM field geometry changes its related frequency & wavelengths

In the Single slit experiment the electron may pass through the slit and contribute to the resultant interference pattern

The equilateral energy distribution of the associated KEM wave photons, created by the linear momentum of the electron, produces the parallel, weaker amplitude side lines of the interference pattern

All EM fields always form v² geometries

All KEM fields are comprised of transverse Bosons and longitudinal Photons producing its associated wavefunctions and properties

$\Delta h\nu$ - *acceleration force* - ΔMv

any interaction between the KEM field and the barrier will affect the energy content of the KEM field and result in velocity changes to the electron

$\Delta h\nu$ - *spectral lines* - Δhf

Tetryonics 39.10 - Single slit experiment

Double slit experiments

The double-slit experiment, sometimes called Young's experiment, is a demonstration that matter and energy can display characteristics of both waves and particles, and demonstrates the fundamentally probabilistic nature of quantum mechanical phenomena and Establishes the quantum interference principle known as wave–particle duality.

Matter particles are stopped by the barrier but the [K]EM wave passes through both slits and is diffracted by them producing weaker EM waves that then superposition with each other to produce interference patterns

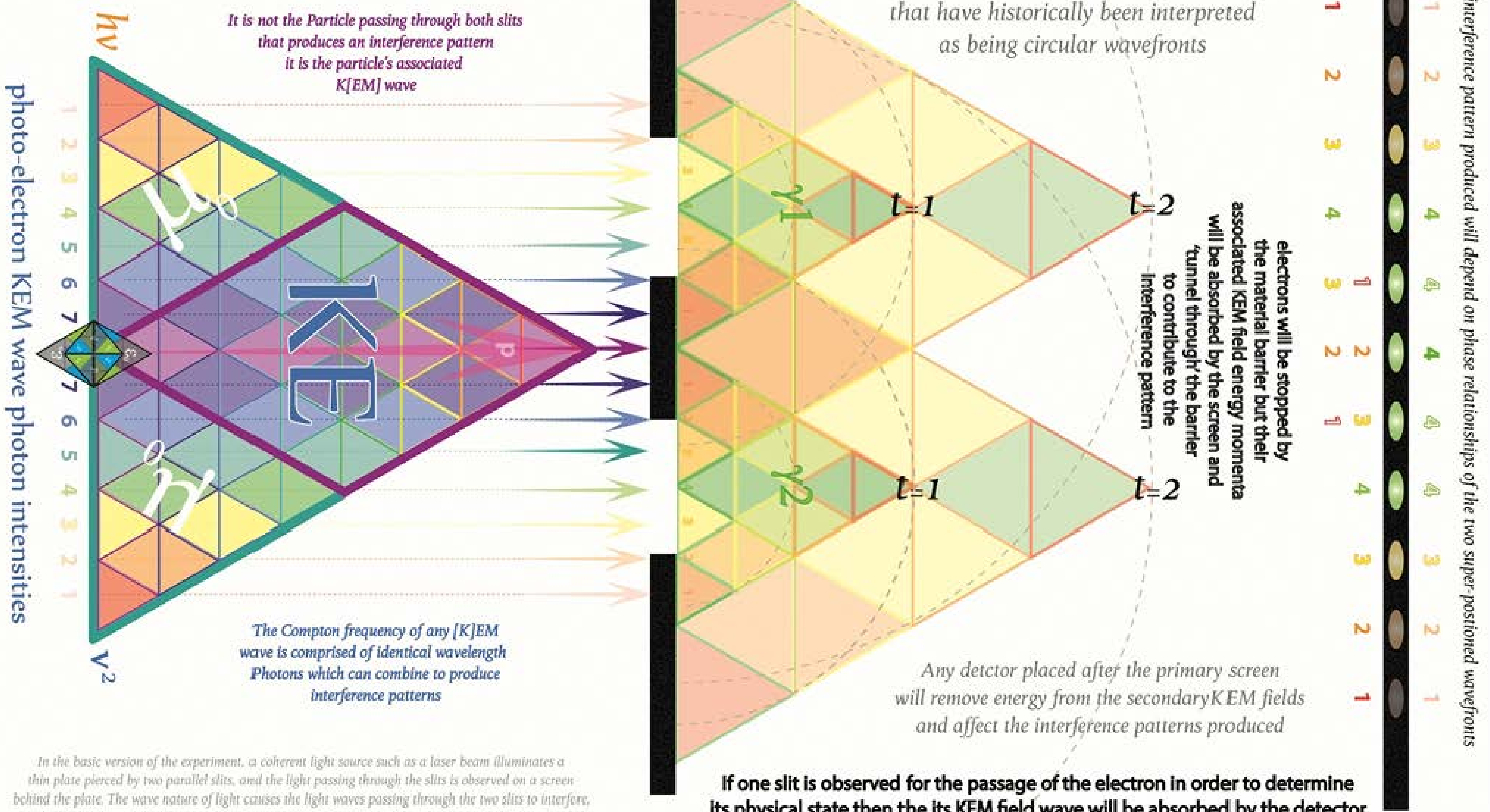

It is not the Particle passing through both slits that produces an interference pattern it is the particle's associated K[EM] wave

The Compton frequency of any [K]EM wave is comprised of identical wavelength Photons which can combine to produce interference patterns

Equilateral superpositioned KEM waves produce constructive and destructive interference waves that have historically been interpreted as being circular wavefronts

electrons will be stopped by the material barrier but their associated KEM field energy momenta will be absorbed by the screen and 'tunnel through' the barrier to contribute to the interference pattern

final interference pattern produced will depend on phase relationships of the two super-postioned wavefronts

Any detector placed after the primary screen will remove energy from the secondaryKEM fields and affect the interference patterns produced

If one slit is observed for the passage of the electron in order to determine its physical state then the its KEM field wave will be absorbed by the detector resulting in only one wave remaining, enforcing a classical particle outcome

In the basic version of the experiment, a coherent light source such as a laser beam illuminates a thin plate pierced by two parallel slits, and the light passing through the slits is observed on a screen behind the plate. The wave nature of light causes the light waves passing through the two slits to interfere, producing bright and dark bands on the screen — a result that would not be expected if light consisted strictly of particles. However, on the screen, the light is always found to be absorbed as though it were composed of discrete particles or photons.

Tetryonics 39.11 - Double slit experiment

Copyright ABRAHAM [2008] - all rights reserved

Light wave Reflection

Excited (or higher energy)
Atoms and Molecules can
emit light and energy via
Electro-dynamics

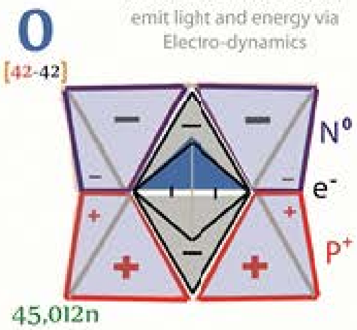
0
[42-42]
N°
e⁻
P⁺
45,012n

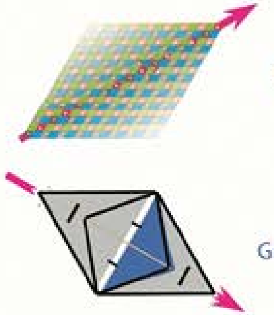
Heat
Spectral Lines
X Rays
Gamma Rays
Photo-electrons

The emitted photon/particle
can go in any direction
so this is the mechanism
for diffuse reflection/radiation.

All Atoms and Molecules
have ANGULAR charged fascia
in their topologies

Incoming light interacts with a surface and may be absorbed, refelected, and/or transmitted.

Materials have a reflectance spectrum which is a function of the angle of incidence of the incoming light.

The color of an object is a function of the color spectrum of the incident light and the reflectance spectrum of the object's surface.

Angle of Incidence
equals
Angle of Reflection

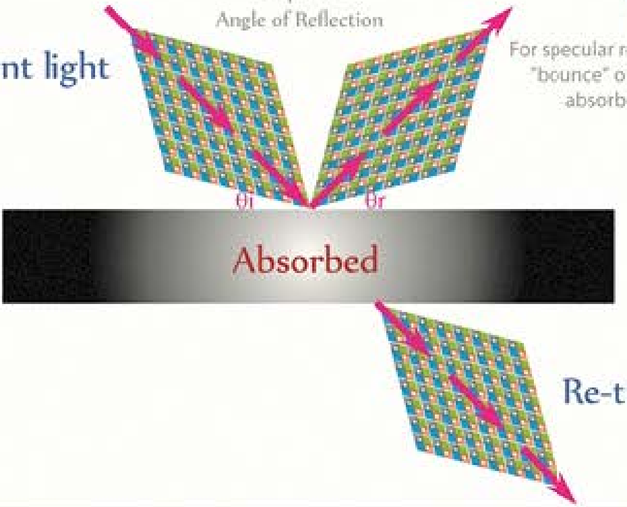
Incident light
θi
θr
Absorbed

Reflected

For specular reflection the photons actually
"bounce" off of the surface and are not
absorbed and then re-radiated

Re-transmitted

Photon Re-Transmission

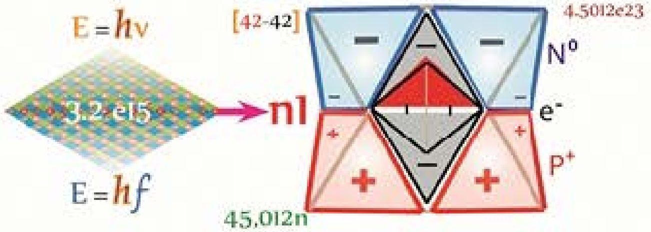
E = hv
3.2e15
E = hf
0
[42-42]
4.5012e23
N°
nl
e⁻
P⁺
45,012n

When a photon strikes a atom or molecule
it generally imparts EM energy as heat
and the photon is absorbed

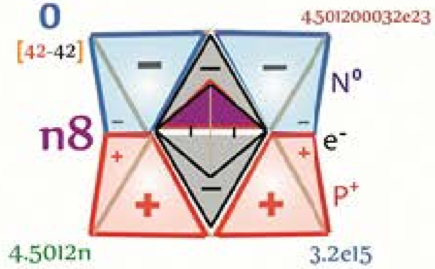
0
[42-42]
4.501200032e23
N°
n8
e⁻
P⁺
4.5012n
3.2e15

For some photons their EM energy is equal
to the resonance energy of the molecule.

When this happens the photon is absorbed
and the molecule is put into an excited state

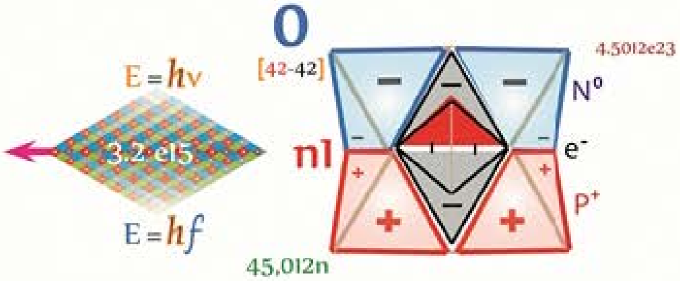
E = hv
3.2e15
E = hf
0
[42-42]
4.5012e23
N°
nl
e⁻
P⁺
45,012n

The molecule relaxes from the excited state back to
the ground state by emitting a photon with EM energy
equal to that of the original photon so the photon
appears to have been reflected
(but is actually absorbed then emitted)

Photon absorption
Photon emission

Copyright ABRAHAM [2008] - all rights reserved

Lightwave Refraction

The refraction of light when it passes from a fast medium to a slow medium bends the light ray toward the normal at the boundaries between any two media.

The amount of bending depends on the indices of refraction of the two media and is described quantitatively by Snell's Law.

White light can be refracted into its component spectral colours

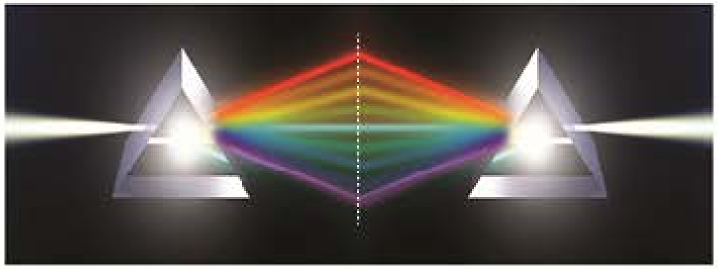

or be re-combined back into white light through superpositioning

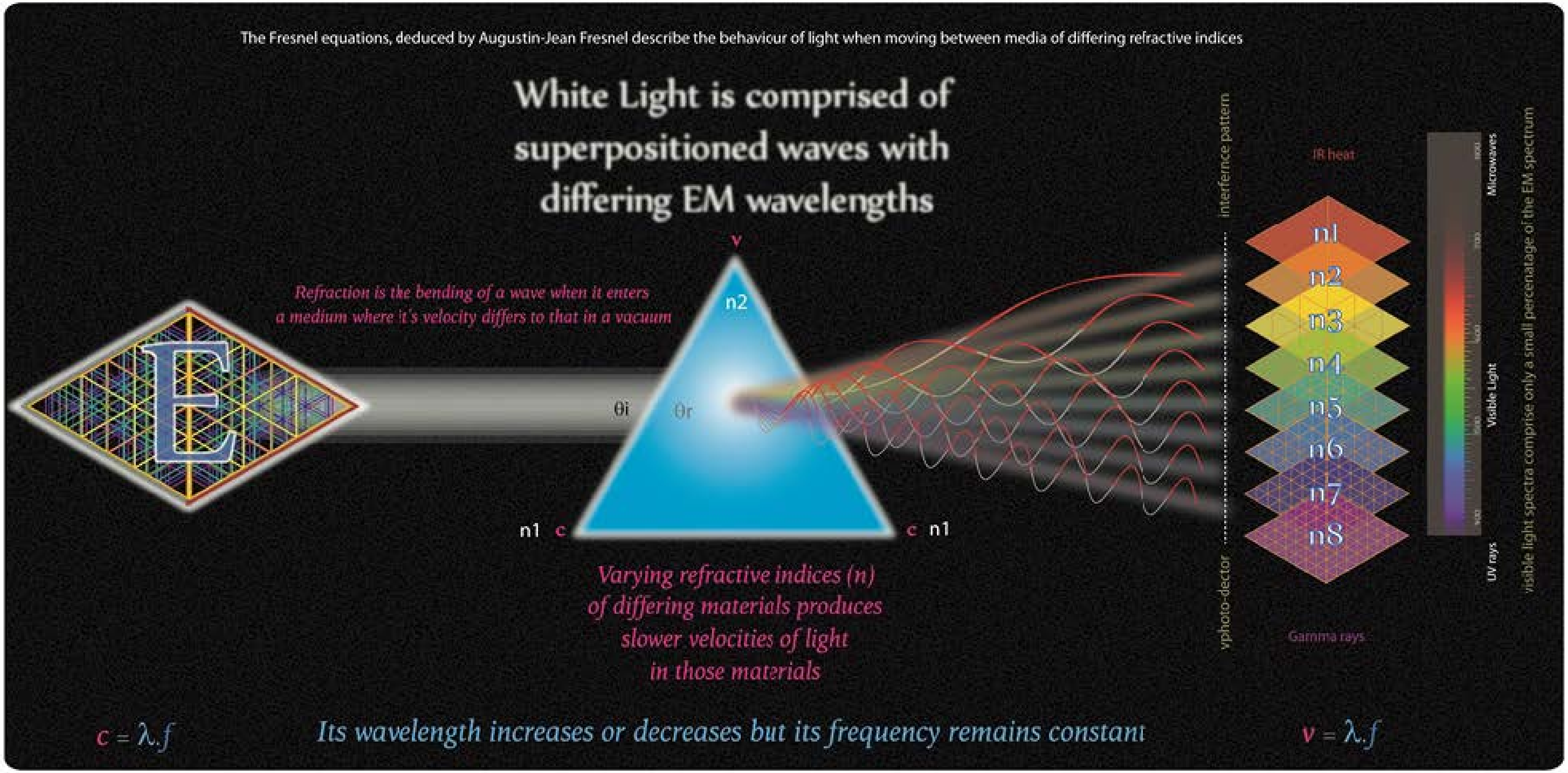

The Fresnel equations, deduced by Augustin-Jean Fresnel describe the behaviour of light when moving between media of differing refractive indices

White Light is comprised of superpositioned waves with differing EM wavelengths

Refraction is the bending of a wave when it enters a medium where it's velocity differs to that in a vacuum

Varying refractive indices (n) of differing materials produces slower velocities of light in those materials

$c = \lambda.f$ *Its wavelength increases or decreases but its frequency remains constant* $v = \lambda.f$

v
n2
θi
θr
n1 c
c n1
E
interference pattern
vphoto-dector
IR heat
n1
n2
n3
n4
n5
n6
n7
n8
Gamma rays
Microwaves
Visible Light
UV rays
visible light spectra comprise only a small percenatage of the EM spectrum

Copyright ABRAHAM [2008] - all rights reserved

Ball Lightning

An alternative radiant geometry for electromagnetic waves exists

The EM energies in radiant 'ball' lightning are transformer-couplings of energy momenta & self-sustaining until any interaction with Matter topologies occurs

Descriptions of ball lightning vary wildly. It has been described as moving up and down, sideways or in unpredictable trajectories, hovering and moving with or against the wind; attracted to, unaffected by, or repelled from buildings, people, cars and other objects.

Ball lightning has been described as transparent, translucent, multicolored, evenly lit, radiating flames, filaments or sparks, with shapes that vary between spheres, ovals, tear-drops, rods, or disks

Any interaction with a 'ball' of EM energies will result in an explosive discharge of it's EM energy momenta

Nikola Tesla was reportedly able to artificially produce 1.5" (3.8 cm) balls using spark gap technologies

Ball lightning creates a strong Magnetic moment and divergent E fields [similar to a bar magnet without the Matter] creating a "Ball' of light with varying EM energies that is sensitive to external M fields

Tetryonics 39.14 - Ball Lightning

Wave~Particle interaction-at-a-distance

All Matter in motion has both:

a standing wave energy topology [Particle] and

a divergent relativistic, Lorentz velocity corrected geometric field of mass-energy momenta [KEM Wave]

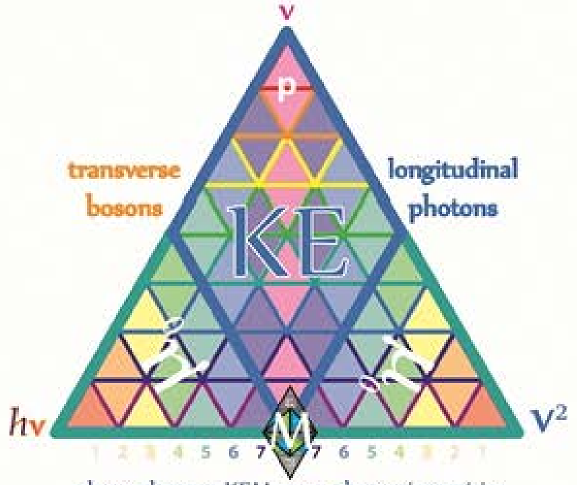
v
p
transverse bosons
longitudinal photons
KE
hv
v²
M
1 2 3 4 5 6 7 7 6 5 4 3 2 1

photo-electron KEM wave photon intensities

Whether it is Light itself or the KEM field of Matter in motion all EM fields are comprised of transverse bosons and longitudinal photons with probabillistic wave geometries

$$RE = M + KEM$$

relativistic energies — Matter — kinetic energies

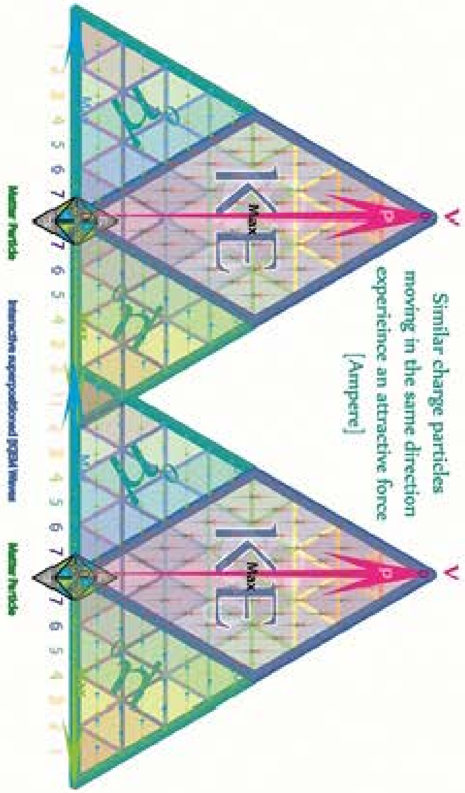
Matter Particle
Interactive superpositioned [K]EM Waves
Matter Particle
KE
Max
p
v
KE
Max
p
v

Similar charge particles moving in the same direction experieince an attractive force [Ampere]

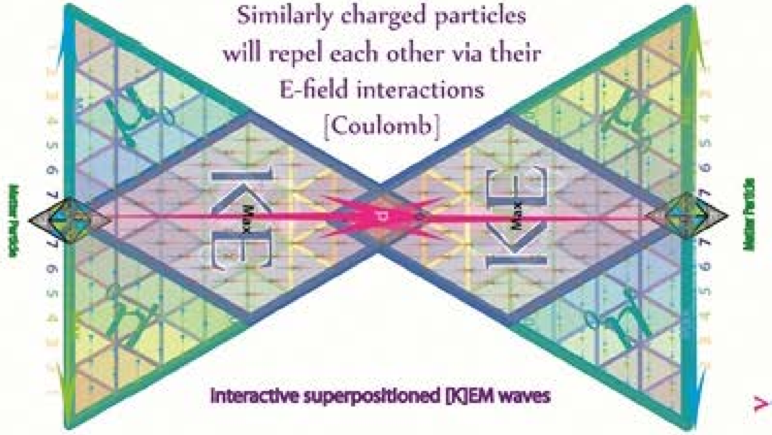
Matter Particle
Matter Particle
KE
Max
p
KE
Max
Interactive superpositioned [K]EM waves

Similarly charged particles will repel each other via their E-field interactions [Coulomb]

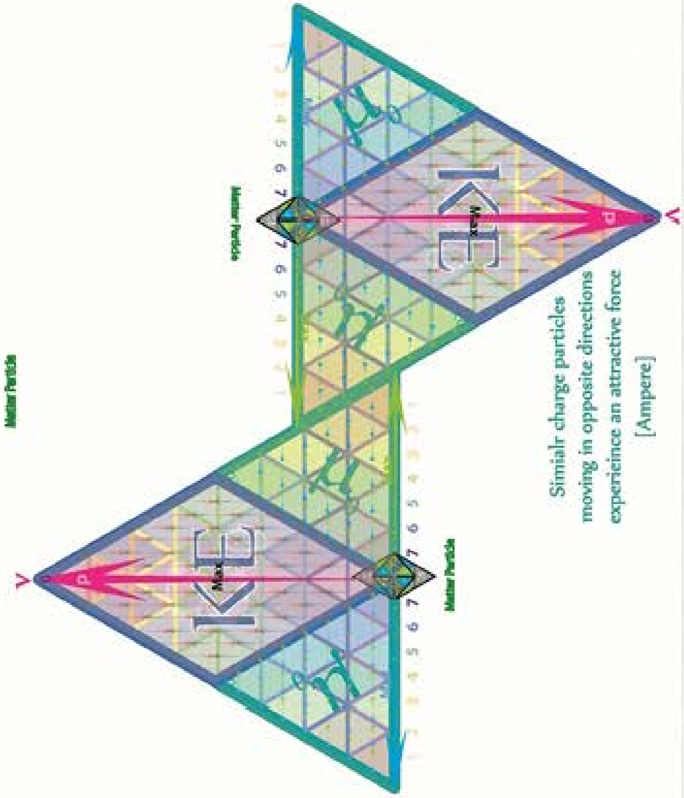
Matter Particle
KE
Max
p
v
Matter Particle
KE
Max
p
v

Simialr charge particles moving in opposite directions experieince an attractive force [Ampere]

Force fields and interaction-at-a-distance

All Matter in motion produces an associated Lorentz velocity corrected KEM field of interactive energy momenta photons

Copyright ABRAHAM [2008] - all rights reserved

Quantum Tunnelling

Particles attempting to travel between potential barriers can be compared to a ball trying to roll over a hill;
quantum mechanics and classical mechanics differ in their treatment of this scenario.

Classical mechanics predicts that particles that do not have enough energy to classically surmount a barrier
will not be able to reach the other side. Thus, a ball without sufficient energy to surmount the hill would roll back down.

Or, lacking the energy to penetrate a wall it would bounce back (reflection) or in the extreme case, bury itself inside the wall (absorption).

In quantum mechanics, these particles can, with a very small probability, tunnel to the other side, thus crossing the barrier.

12
[0-12]
Charged
Matter
1.2 e20

Kinetic EM energies
v
v
v²
KE
p

The quantum magician's slight of hand trick is where the incident electron is swapped another

All electrons are IDENTICAL

It is practically impossible to identify a particular electron and track it

even using spin orientations only eliminates half of the total electrons in the barrier

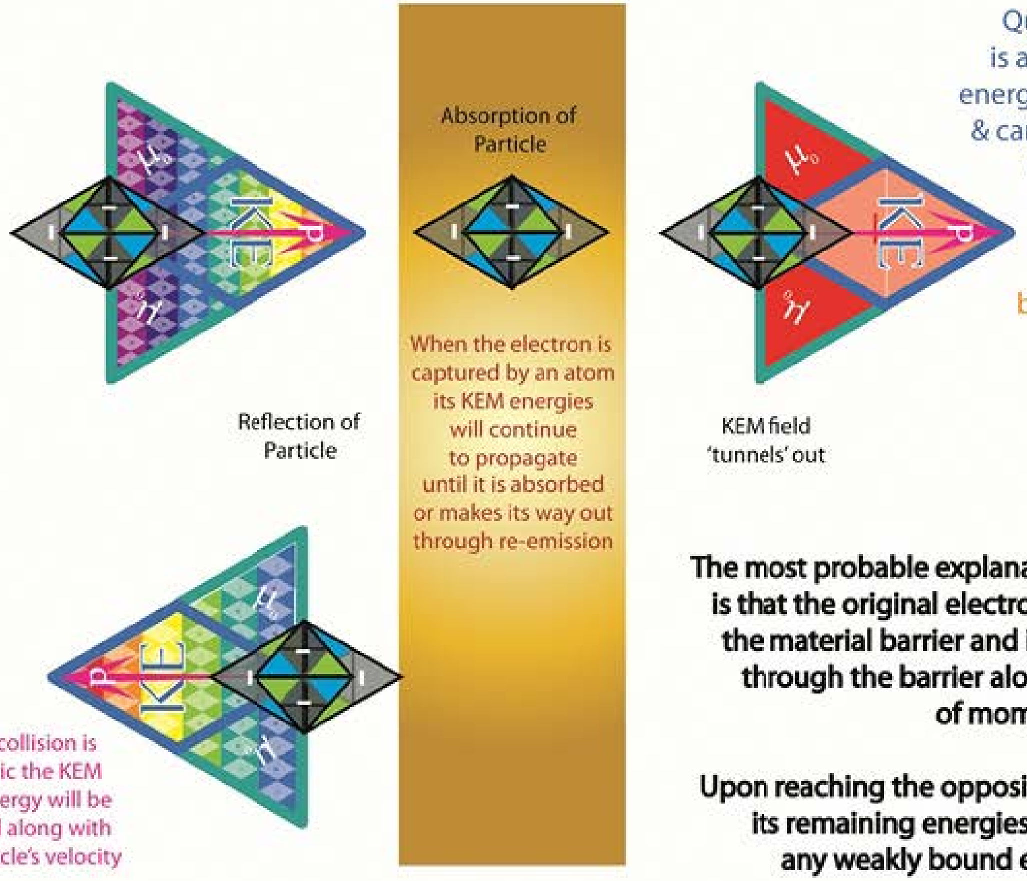

When the electron is captured by an atom its KEM energies will continue to propagate until it is absorbed or makes its way out through re-emission

If the collision is inelastic the KEM field energy will be reduced along with the particle's velocity

Quantum KEM field tunnelling is an example of conservation of energy momenta on the atomic scale & can be likened to a unidirectional version of Newton's cradle

Removal of all electrons but the one under measurement is impractical as it would result in an attractive coloumbic force that would trap the electron in the barrier

The most probable explanation of quantum tunneling is that the original electron was bound to atoms in the material barrier and its KEM field propagated through the barrier along its original direction of momentum.

Upon reaching the opposite side/edge of the barrier its remaining energies are able to accelerate any weakly bound electron found there

Quantum 'tunnelling' is revealed to be the result of an electron's KEM wave propagation through the barrier not the electron itself

Copyright ABRAHAM [2008] - all rights reserved

$$\Delta hf = \Delta Mv = \Delta p$$

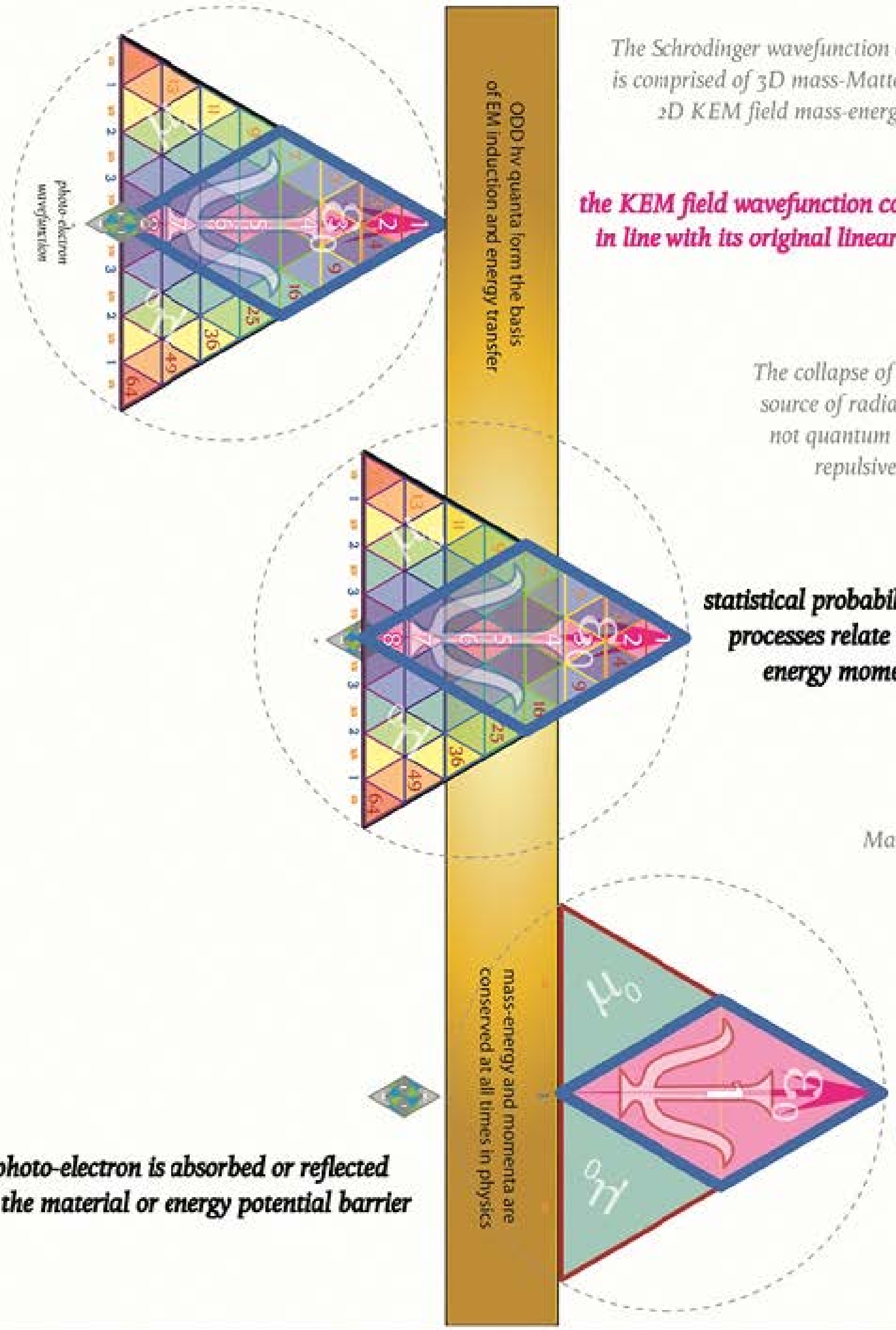
photo-electron wavefunction
ODD hv quanta form the basis of EM induction and energy transfer
mass-energy and momenta are conserved at all times in physics
μ₀
ε₀

The Schrodinger wavefunction of photo-electrons is comprised of 3D mass-Matter topologies and 2D KEM field mass-energy geometries

the KEM field wavefunction continues to propagate in line with its original linear momentum vector

The collapse of Matter topologies is the source of radiant stellar mass-energies not quantum tunnelling to overcome repulsive Coloumbic forces

statistical probabilities applied to quanutm tunneling processes relate to the distribution of EM Planck energy momenta quanta in the KEM field

Matter-energy conversion via GEM pinches is the source of stellar energies released in stars

photo-electron is absorbed or reflected by the material or energy potential barrier

As Planck energy momenta quanta interact or are absorbed the KEM field's wavefunction will entropy accordingly, in turn affecting the wavelength-frequency-velocity of the KEM field

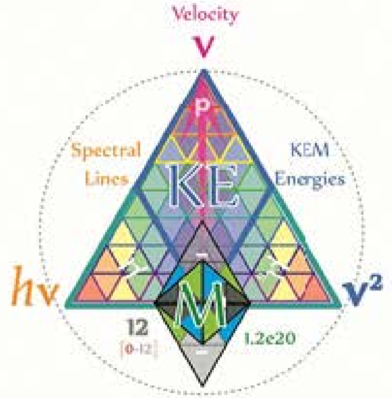
Velocity
v
p
Spectral Lines
KEM Energies
KE
hv
v²
M
12
[0-12]
1.2e20

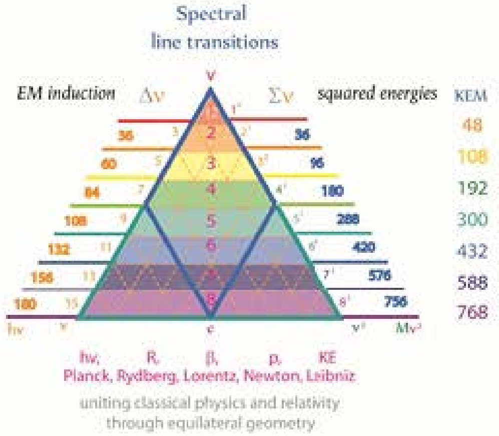
Spectral line transitions
EM induction
Δv
Σv
squared energies
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
1
2
3
4
5
6
7
8
hv
v²
Mv²
hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz
uniting classical physics and relativity through equilateral geometry

Quantum tunneling plays no role in the 'fusion' of lighter elements into heavier elements in stars (which is itself an erroneous process)

Tetryonics 39.17 - mass-energy momenta induction & conservation

All Matter in motion is comprised of:

a mass-Matter standing-wave topology,
and
a KEM mass-energy field geometry

whose quantised energy momenta is reflective
of the particle's vector velocity

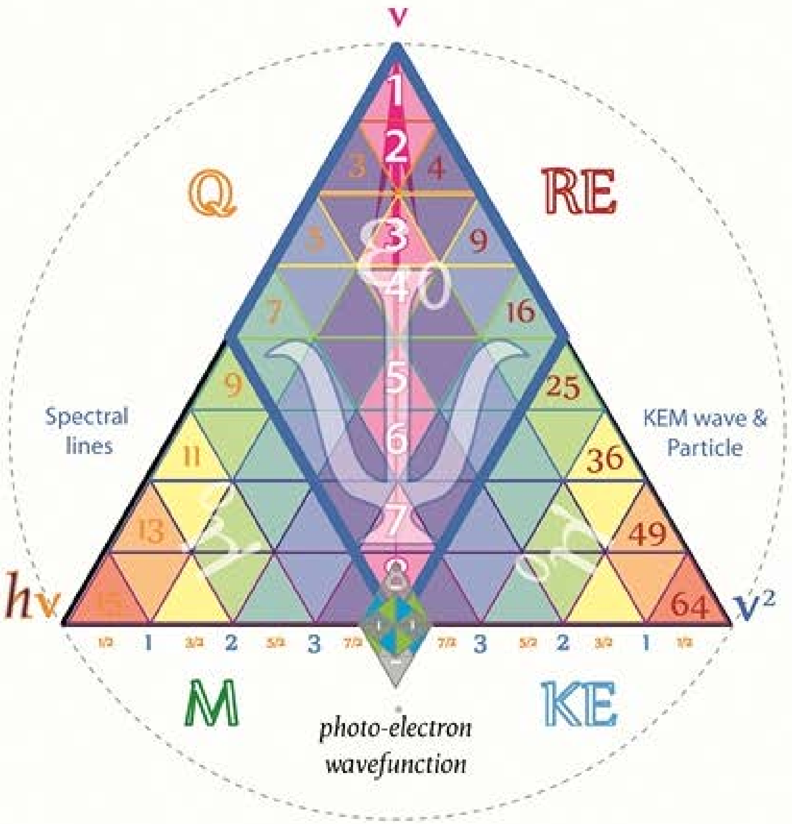

photo-electron
wavefunction

The normal distribution of Planck energy momenta in any KEM field is the quantum foundation for the statistical probability mechanics of quanutm mechanics and quantum electrodynamics

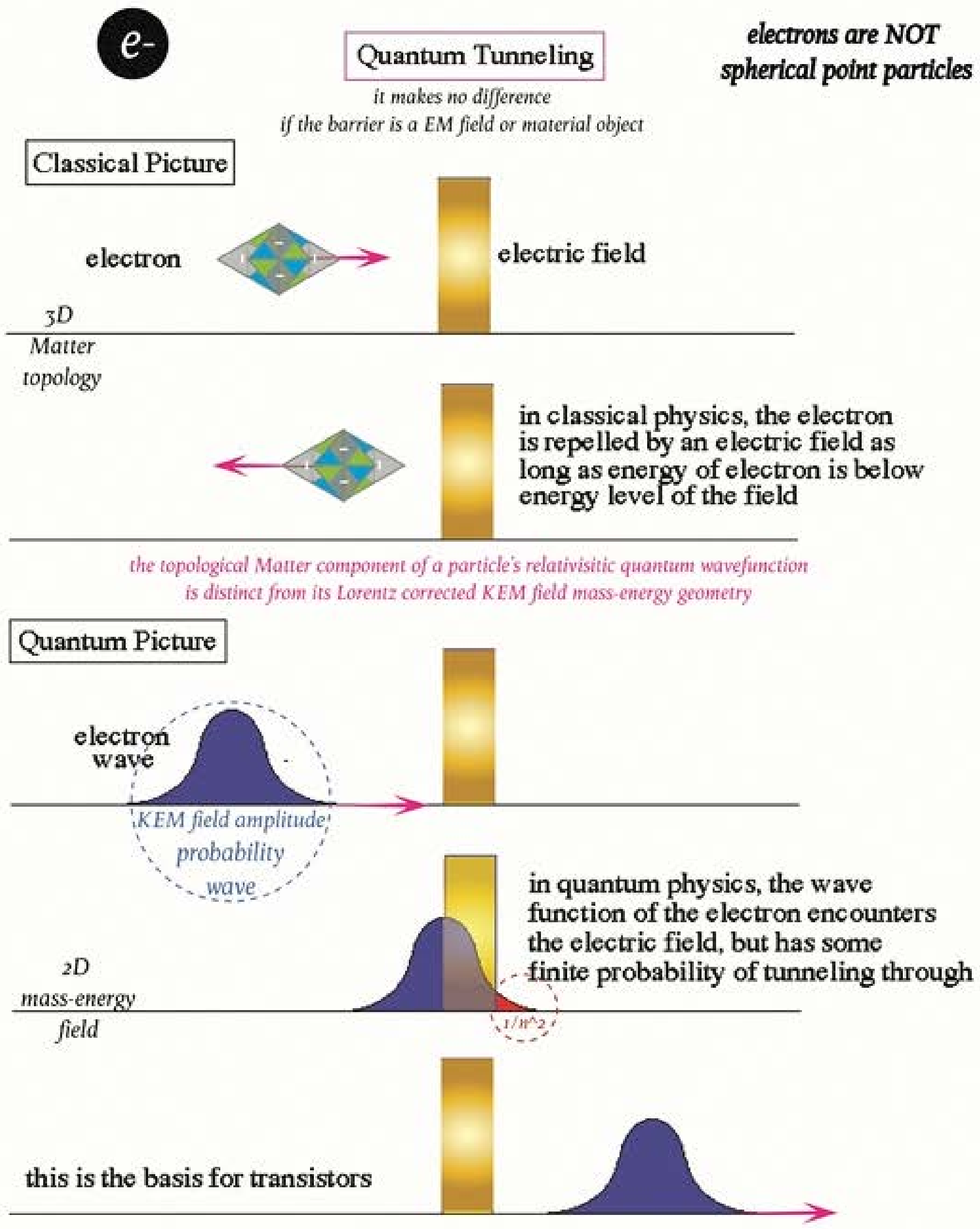

Quantum Tunneling

it makes no difference
if the barrier is a EM field or material object

electrons are NOT spherical point particles

Classical Picture

electron

electric field

in classical physics, the electron is repelled by an electric field as long as energy of electron is below energy level of the field

the topological Matter component of a particle's relativisitic quantum wavefunction
is distinct from its Lorentz corrected KEM field mass-energy geometry

Quantum Picture

electron wave

KEM field amplitude
probability
wave

in quantum physics, the wave function of the electron encounters the electric field, but has some finite probability of tunneling through

this is the basis for transistors

quantum tunnelling is often also utilised in modern physics
as a means of overcoming Coulombic charge barriers of Protons in stars
in order to explain 'fusion' as the source of stellar energy

Tetryonics 39.17 - Quantum Tunnelling wavefunctions

Copyright ABRAHAM [2008] - all rights reserved

Relativistic Kinetic Energies

$$KE = \frac{1}{2}M\frac{v^2}{c^2}$$

The relativistic Kinetic mass-energies of all Matter topologies in motion are contained in their KEM fields and are subject to Lorentz velocity corrections

$$p^2 = E = m\frac{v^2}{c^2}$$

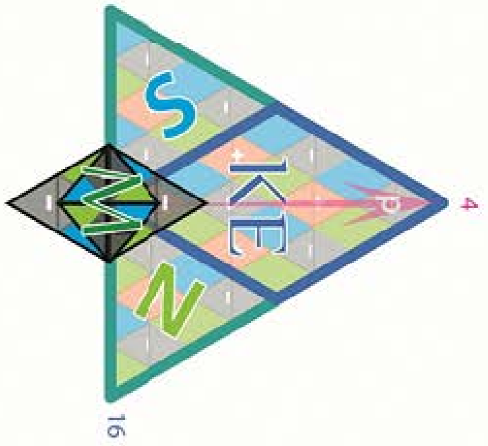

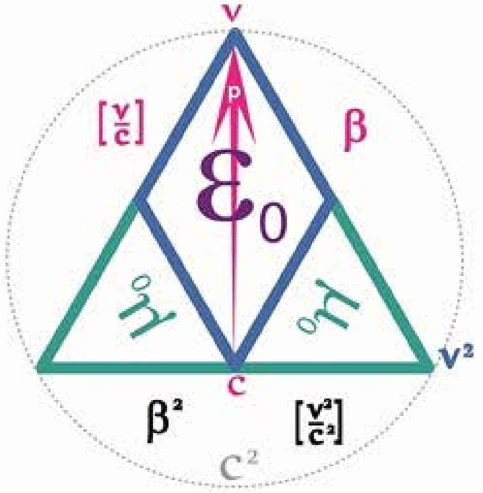
v
p
[v/c]
β
ε0
μ0
μ0
v²
c
β²
[v²/c²]
c²

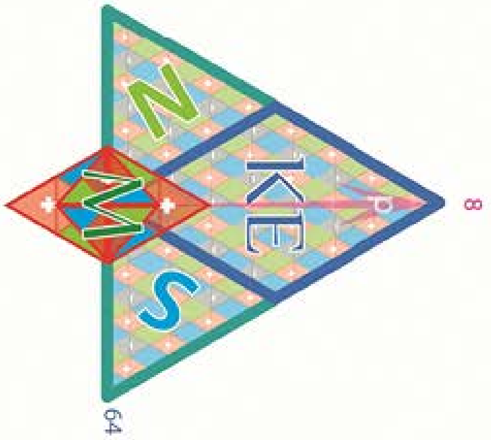

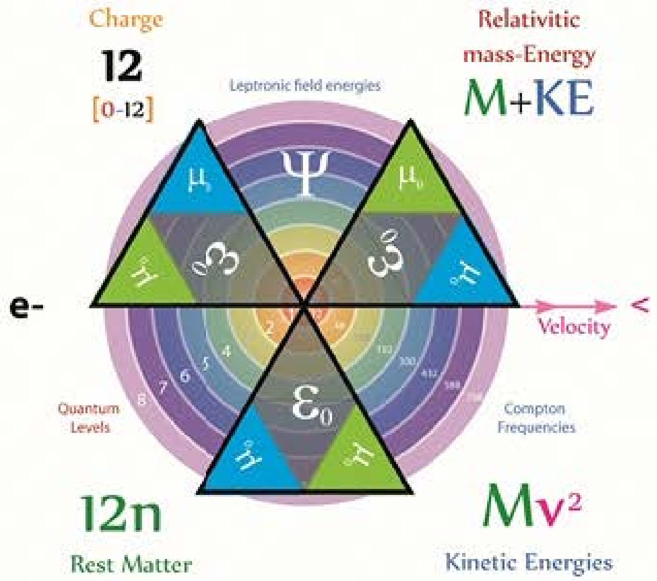
Charge
12
[0-12]
Leptronic field energies
Relativitic
mass-Energy
M+KE
Ψ
μ0
μ0
ε0
ε0
e-
Velocity
<
Quantum
Levels
Compton
Frequencies
12n
Rest Matter
Mv²
Kinetic Energies

Lorentz contractions only apply to the KEM fields of Matter in motion

The Kinetic Energies of a particle in motion can be expressed in terms of its linear momentum

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

or equally as 1/2 the total relativistic EM mass-energies [exclusive of its invariant rest mass-Matter]

Charge and Matter topologies are Lorentz invariant

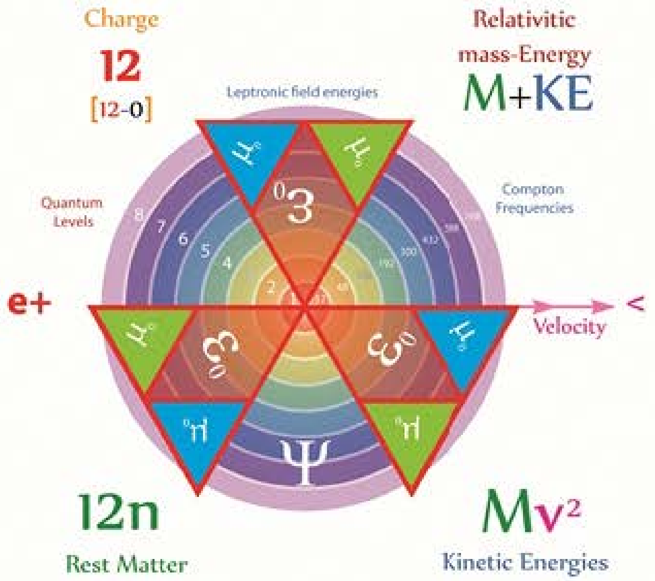
Charge
12
[12-0]
Leptronic field energies
Relativitic
mass-Energy
M+KE
Quantum
Levels
Compton
Frequencies
μ0
ε0
Ψ
e+
Velocity
<
12n
Rest Matter
Mv²
Kinetic Energies

Relativistic Lorentz Corrections

ges their name from their earliest appearance, in Lorentzian electrodynamics
as velocity increases the wavelengths and physical scalar geometries of Planck quanta decreases

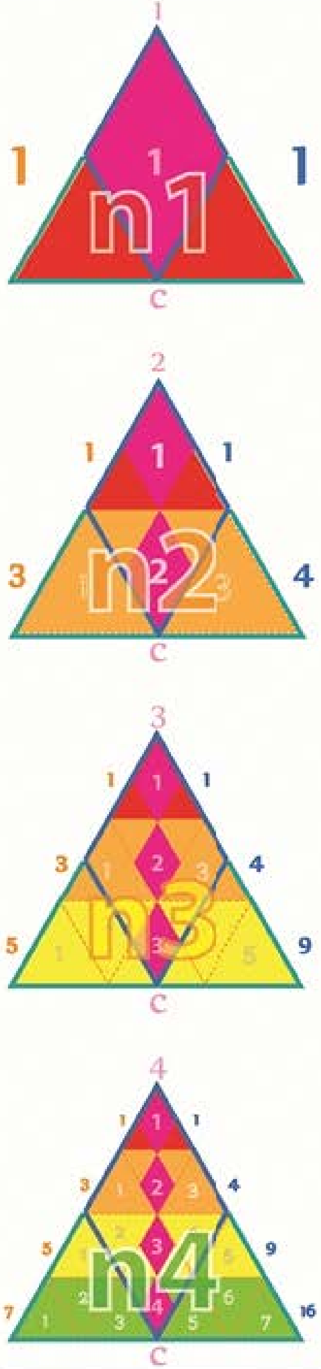
1
1
1
1
n1
c
2
1
1
1
3
4
n2
c
3
1
1
3
4
5
9
n3
c
4
1
1
3
4
5
9
7
16
n4
c

Linear correction factor

$$\beta = \left[\frac{v}{c}\right]$$

Equilateral QAM contractions

$$t' = \frac{t}{\sqrt{1-\frac{v^2}{c^2}}}$$

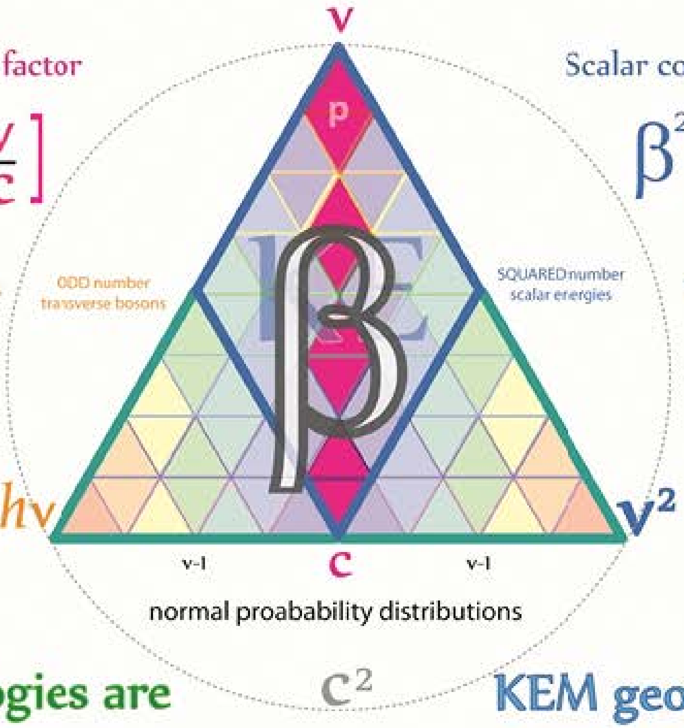
v
p
ODD number
transverse bosons
SQUARED number
scalar energies
hv
v²
v-1
c
v-1
normal proabability distributions
c²

Scalar correction factor

$$\beta^2 = \left[\frac{v^2}{c^2}\right]$$

'WAVE-length' contractions

$$L = L'\sqrt{1-\frac{v^2}{c^2}}$$

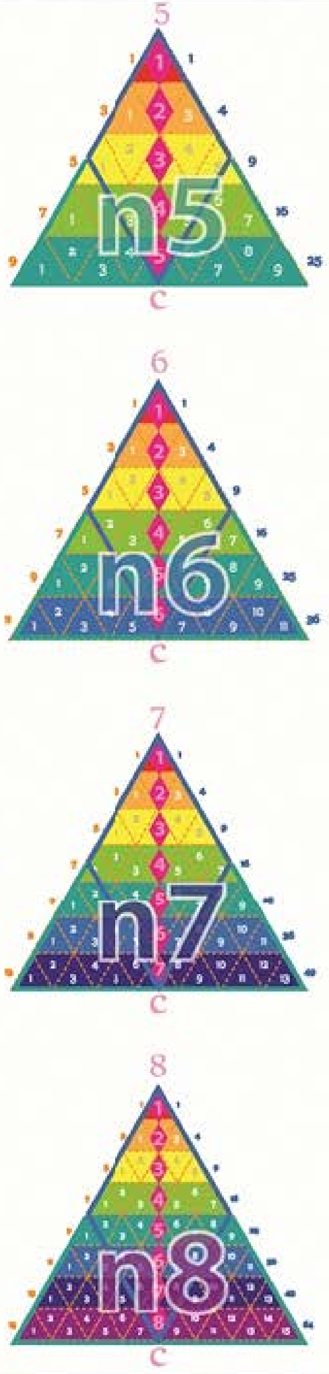
5
n5
c
6
n6
c
7
n7
c
8
n8
c

Matter topologies are Lorentz invariant

KEM geometries are relativistic fields

The poor definitions of EM mass & Matter has led to the incorrect application of Lorentz corrections to Matter

all the quantum energy momenta probabilites
of wavefunction distributions re-normalise [or sum] to

Lorentz contractions apply to [K]EM waveforms only

c is the limiting velocity for the electrical acceleration of Matter

$$\sqrt{1-\beta^2}$$

Incorrectly applied to Matter topologies in motion,
The Lorentz factor or Lorentz term appears in
several equations in special relativity.

$$\sqrt{1-\left(\frac{v}{c}\right)^2}$$

The Energy of a Magnet

It is known that Maxwell's electrodynamics--as usually understood at the present time--when applied to moving bodies, leads to asymmetries which do not appear to be inherent in the phenomena. Take, for example, the reciprocci electrodynamic action of a magnet and a conductor. The observable phenomenon here depends only on the relative motionof the conductor and the magnet, whereas the customary view draws a sharp distinction between the two cases in which either the one or the other of these bodies is in motion. For if the magnet is in motion and the conductor at rest, there arises in the neighborhood of the magnet an electric field with a certain definite energy, producing a current at the places where parts of the conductor are situated. But if the magnet is stationary and the conductor in motion, no electric field arises in the neighborhood of the magnet.

In the conductor, however, we fied an electromotive force, to which in itself there is no corresponding energy, but which gives rise--assuming equality of relative motion in the two cases discussed--to electric currents of the same path and intensity as those produced by the electric forces in the former case. [On the electrodynamics of moving bodies, A. Einstein, 1905]

V

Energy can be removed by placing a conductive loop between any Positive and Negative quanta the EM field thus creating a potential difference & an electro-motive force

ε_0

The neutralised [K]EM fields are divergent from a permanent magnet and create a Magneto-static EM field with neutral Electro-static energies

Investigation of the source of the emf produced by magnets in conductive loops placed around them led to the development of relativity theory

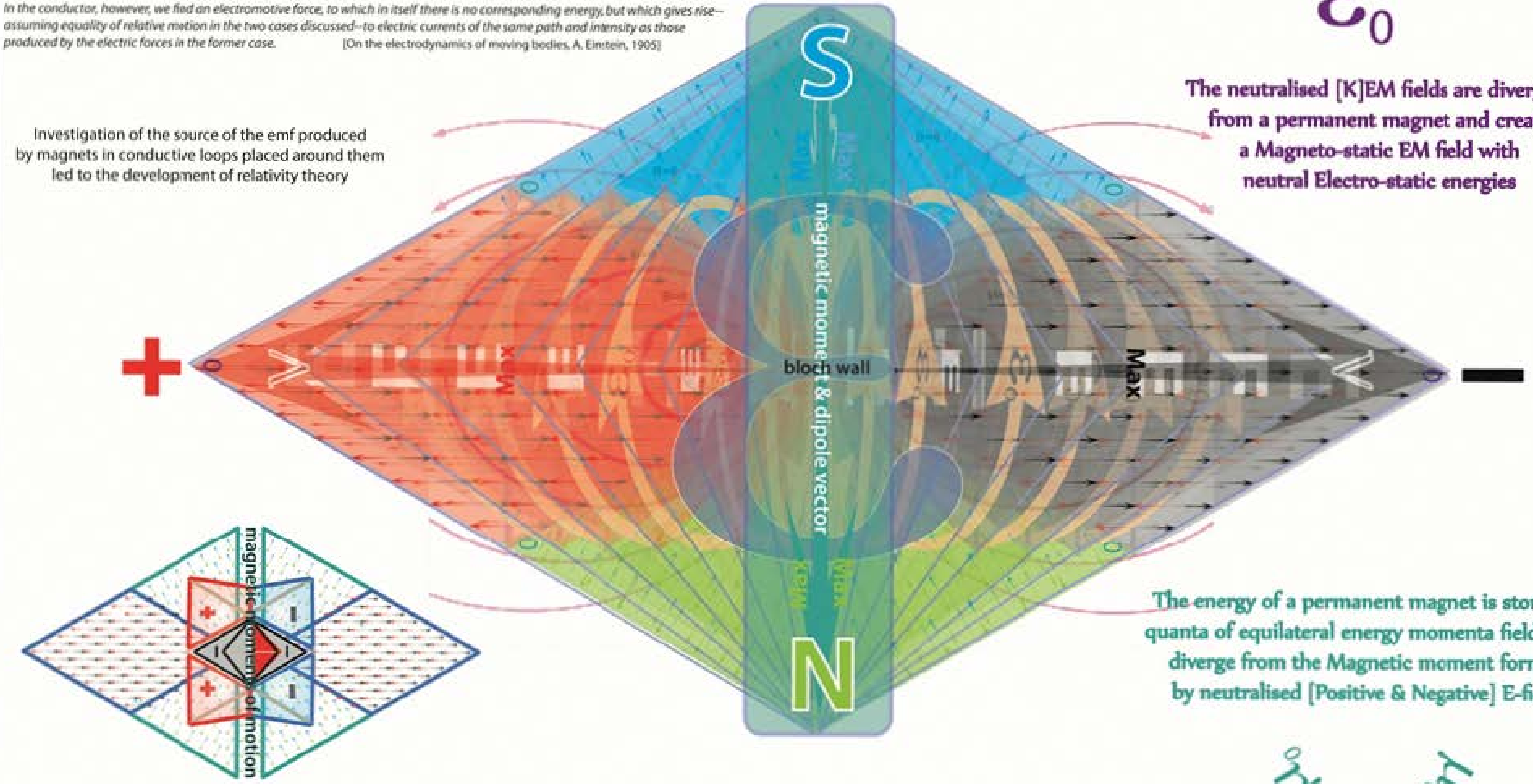

The energy of a permanent magnet is stored in quanta of equilateral energy momenta fields that diverge from the Magnetic moment formed by neutralised [Positive & Negative] E-fields

Magnetic fields are produced by moving electric charges and the intrinsic magnetic moments of elementary particles associated with a fundamental quantum property, their spin. In special relativity, electric and magnetic fields are two interrelated aspects of a single object, caled the electromagnetic field tensor; the aspect of the electromagnetic field that is seen as a magnetic field is dependent on the reference frame of the observer. In QED, the electro-magnetic field is quantized and electromagnetic interactions result from the exchange of photons.

Ω

Classically, the energy of a permanent magnet 'circulates in endless loops' from North to South in 3 dimensions around the magnetic moment

Copyright ABRAHAM [2008] - all rights reserved

$E = h\nu$

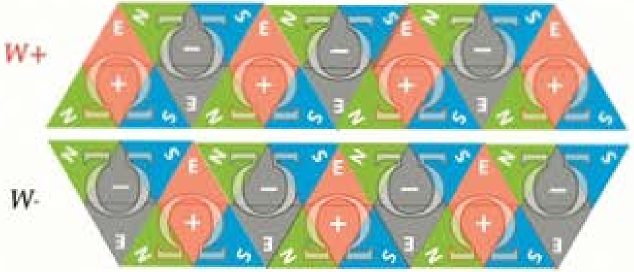

BOSONS are
transverse EM waves
[charge carriers]

PHOTONS are
longitudinal EM waves
[neutral charge carriers]

$E = hf$

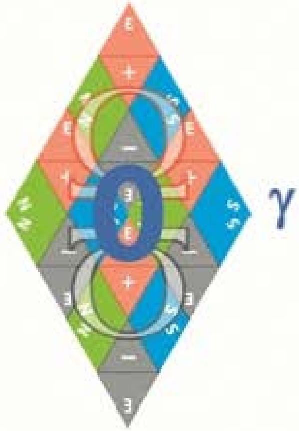

The equilateral geometry of
ALL EM fields in relative motion
to a conductor produces a
time varying emf

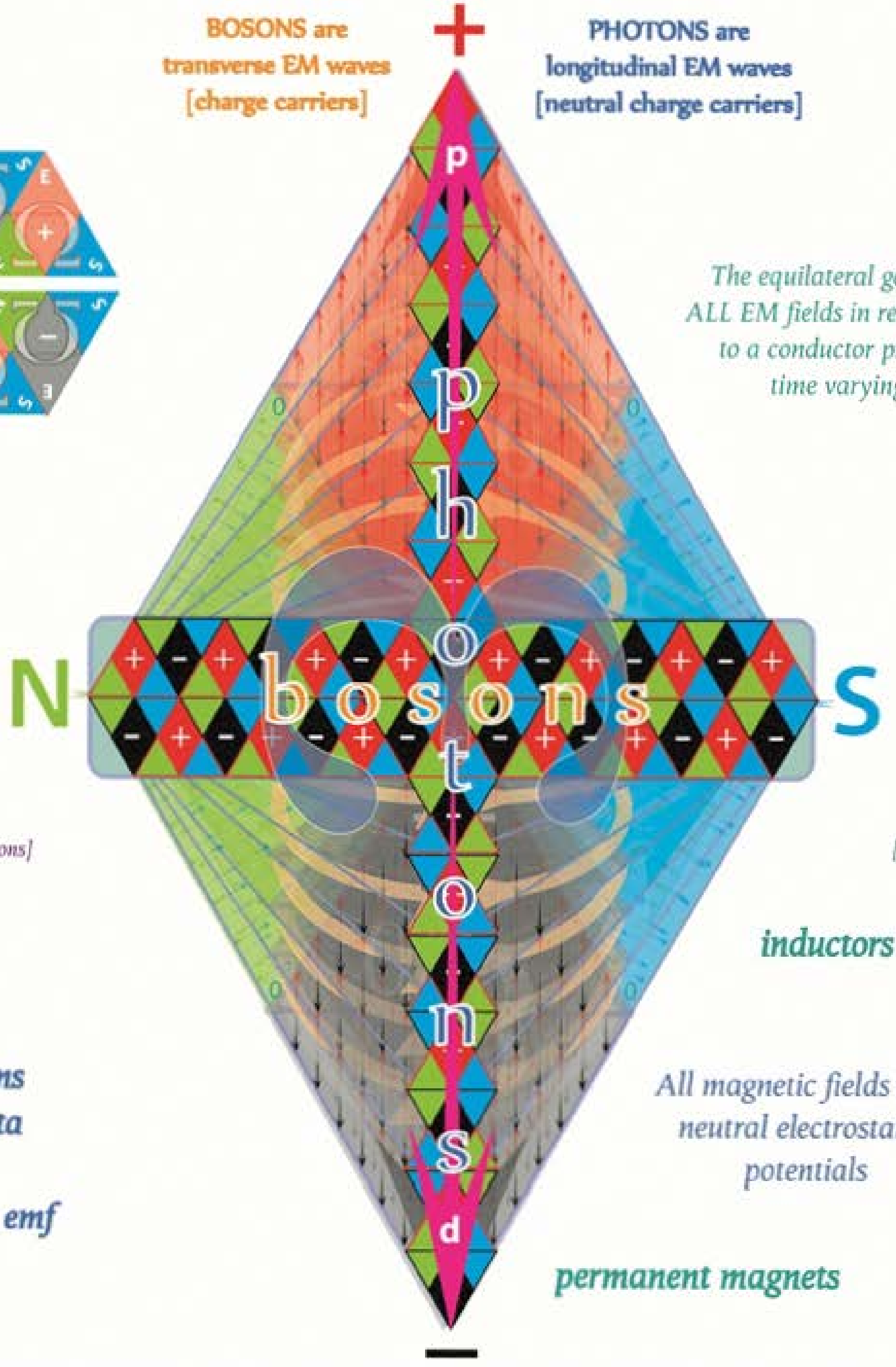

$B = -\frac{d\phi_\varepsilon}{dt}$

A changing Electric field
produces
a changing Magnetic field

The is no such entity as a purely Electric field
[all fields are Electro-Magnetic in their geometric foundations]

$\mathcal{E} = -\frac{d\phi_B}{dt}$

A changing Magnetic field
produces
a changing Electric field

The is no such entity as a purely Magnetic field
[all fields are Electro-Magnetic in their geometric foundations]

Any two separated charge bosons
contain charged energy momenta
that once connected via a
conductive circuit will produce an emf

inductors

All magnetic fields store
neutral electrostatic
potentials

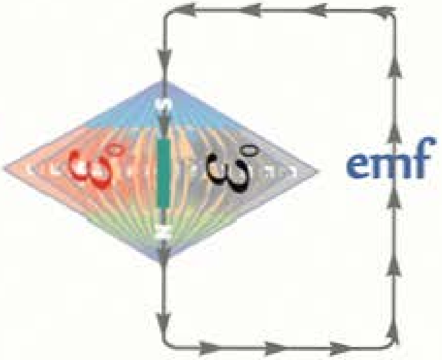

permanent magnets

Copyright ABRAHAM [2008] - all rights reserved

The moving magnet - conductor problem

$$B = -\frac{d\phi_{\varepsilon}}{dt}$$

$$\varepsilon = -\frac{d\phi_{B}}{dt}$$

All permanent magnets are electromagnetic fields comprised of
transverse bosons and longitudinal photons

We see two observers both moving, approaching each other.

Observer 1 says:

*The conductive loop is stationary and the magnet is moving toward it.
The electrons in the loop are stationary and have no magnetic moments
There is a magnetic field, but it can't produce any force on
the electrons since they are stationary within the loop.*

*Instead, the magnetic field is changing, growing stronger as
the magnet gets closer, and this changing magnetic field
produces an electric field which causes forces on the electrons, and
drives them around the loop producing the current in the galvanometer.*

Conductor Frame of Reference
(Moving Magnet)

**Charged boson field geometries
contain equilateral energy momenta
with transverse magnetic dipoles**

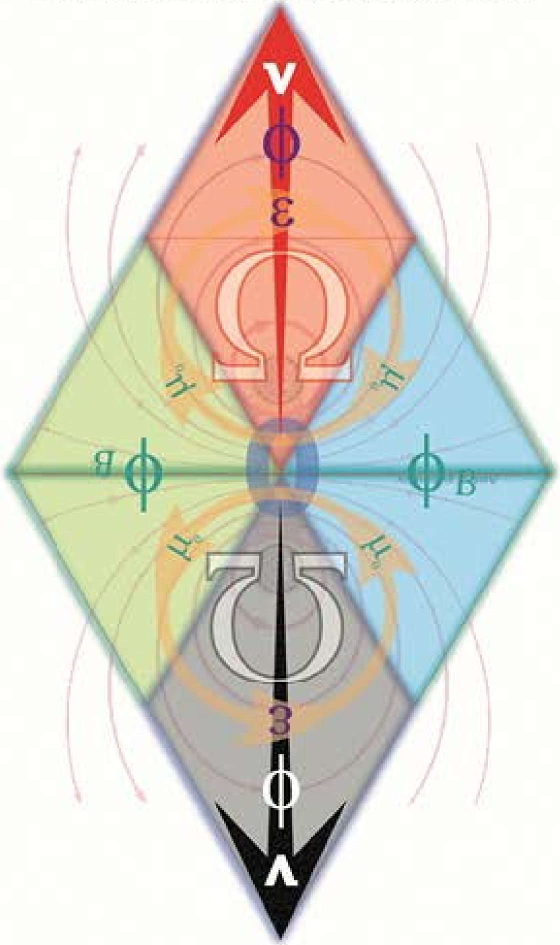

**Neutral photon field geometries
produce electromotive forces as a result
of the charged bosons that make them**

Magnet Frame of Reference
(Moving Conductor)

Observer 2 says:

*The magnet is stationary and the loop is moving toward it.
The electrons in the loop, since they are moving with the loop,
generate their own magnetic moments and experience Lorentz forces
as a result of the external magnetic field [F = q v X B],
which drives them around the loop and produces the current in the galvanometer.*

There is no electric field.

v
p
KE
S
N
M
hv
v²

Special relativity postulates that electric fields in one reference frame become magnetic fields in another frame

Tetryonic theory shows geometrically that the Electric and Magnetic fields are discrete, invariant geometries
resulting from the equilateral 'inductive' mass-energy [Planck quanta] geometries that comprise all EM fields.

It is the relative motion of these geometries wrt a conductor that creates changing E&M field flux strengths in turn producing an emf.
When there is no relative motion in either the EM field or the conductor, no emf is produced

Copyright ABRAHAM [2008] - all rights reserved

The Electromotive Force

The electromotive force, [or emf]
or electromotance
is "that which tends to cause
a current (electrons and ions)
to flow in a conductor.

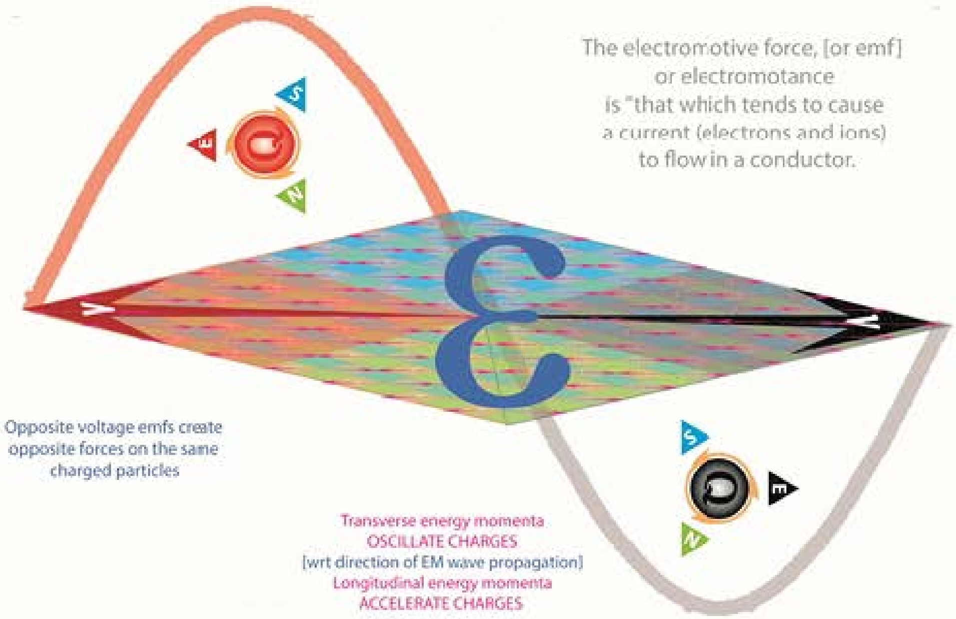

Opposite voltage emfs create
opposite forces on the same
charged particles

Transverse energy momenta
OSCILLATE CHARGES
[wrt direction of EM wave propagation]
Longitudinal energy momenta
ACCELERATE CHARGES

Opposite charged particles
experience opposing forces
due to the emf

"A source of emf can be thought of as a kind of charge pump
that acts to move a charge from a point of low potential
through its interior to a point of high potential. ...
The emf of the source is defined as the work done per charge
dq: = dW/dq."

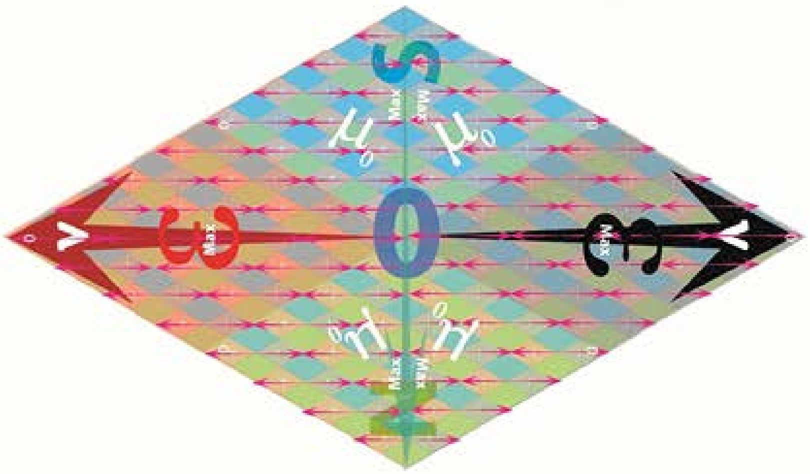

THe magneto-static fields of all Magnets are
comprised of electromagnetic fields of discrete
Planck energy momenta quanta, which
in turn are capable of accelerating
charged particles

All ElectroMagnetic fields have
distinct equilateral electric and
magnetic field geometries that produce
velocity related sinusodal waveforms

$$B = -\frac{d\phi_\varepsilon}{dt}$$

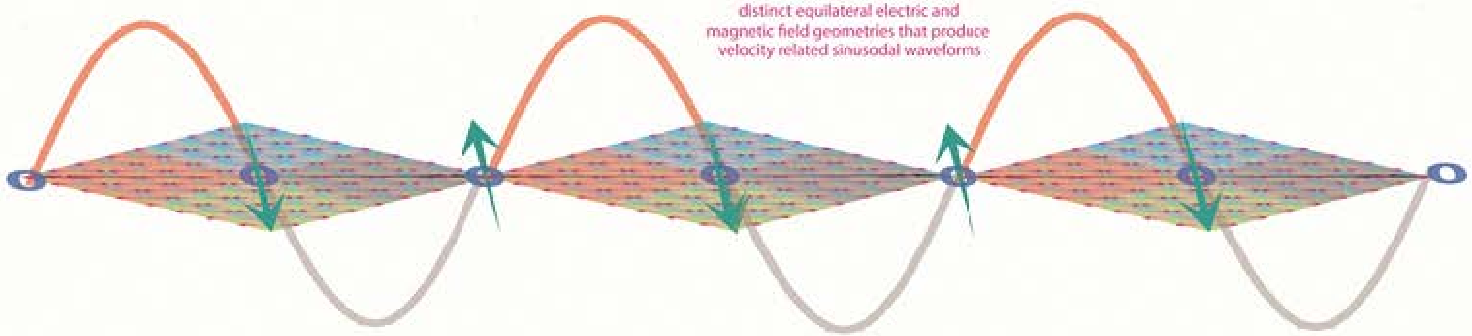

$$\mathcal{E} = -\frac{d\phi_B}{dt}$$

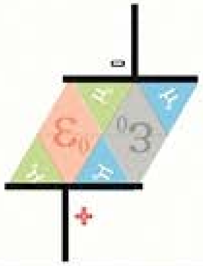

**Charged boson field geometries
can be viewed as transverse
quantum capacitors**

Lenz's law:
***"The emf induced in an electric circuit always acts
in such a direction that the current it drives around
the circuit opposes the change in magnetic flux
which produces the emf."***

**Neutral photon field geometries
can be viewed as longitudinal
quantum batteries**

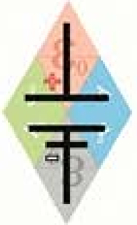

Copyright ABRAHAM [2008] - all rights reserved

The Moving Magnet problem

An electromotive Force results from
the motion of a conductor
relative to a magnetic field

A Magnet is really an electrostatic
store of neutralised emf potential differences
[creating an orthagonal Magnetic moment]

The moving magnet and conductor problem
is a famous 19th century thought experiment,
and provides the intersection for classical
and relativistic electromagnetism.

The moving magnet and conductor problem,
along with Michelson-Morley experiment,
formed the basis of Einstein's theory of
special relativity.

Relativistic theory explanation

In the frame of a conductor moving relative to the magnet, the conductor experiences a force due to an electric field.

A Magnetic field is an Electric field viewed in a differing inertial frame

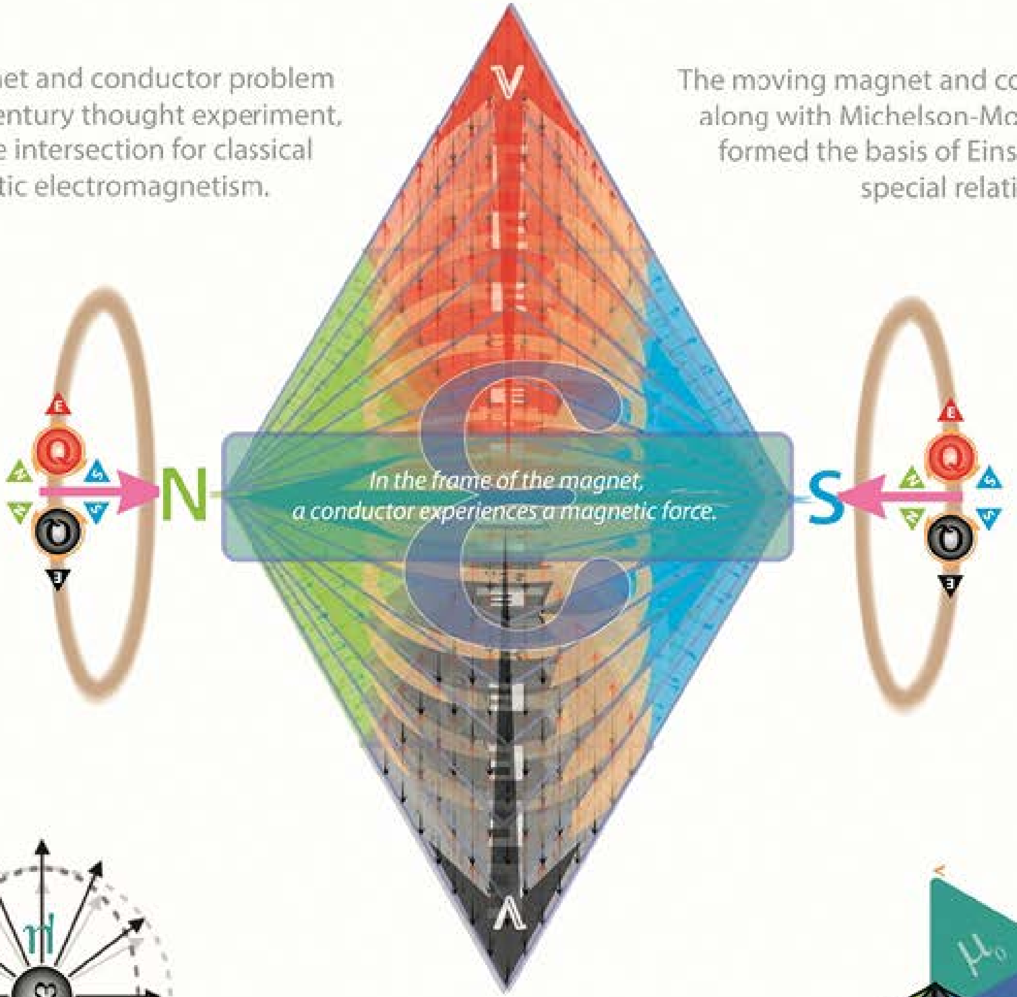
N
S
In the frame of the magnet,
a conductor experiences a magnetic force.

Tetryonic theory explanation

The conductor moves through continuously changing electric and magnetic field geometries experiencing sinuodal E/M forces that are directly related to each other at all times

Electric and Magnetic fields are discrete geometrically related energies

SR requires that all observers
in inertial frames arrive at the
same form for Maxwell's equations

SR seeks to modify space and time
in a manner such that the forces
and fields transform consistently.

producing expressions that
differ from Newton's law of
motion by a factor of γ

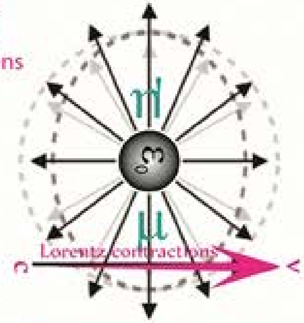

the relativistic distortion of
accelerating spherical point charges
creates magnetic moments

Electrostatic particles have no magnetic moment

$$\vec{F} = \underset{\text{Electric force}}{q\vec{E}} + \underset{\text{Magnetic force}}{q\vec{v} \times \vec{B}}$$

*Once in motion it posseses a magnetic moment and kinetic energies
and is subject to Lorentz forces when it moves throuhg an external EM field*

magnetic moments are
the result of secondary KEM
fields of Matter in motion

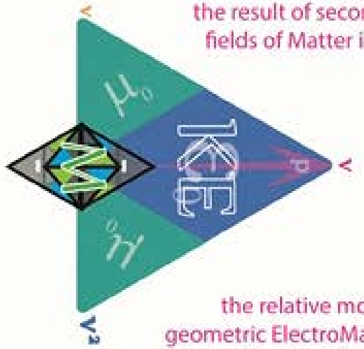

Lorentz corrections
only all to KEM field
mass-energies

the relative motion of
geometric ElectroMagnetic fields
creates changing EM forces

Tetryonics 40.07 - The Moving Magnet problem

Copyright ABRAHAM [2008] - all rights reserved

Einstein's Error of perception

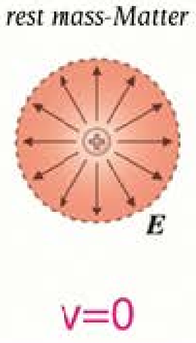

In its rest frame the electric field of a point charge has the same strength in all directions
and diverges away from the charge.

Electric fields are distorted due to relativistic speed effects to create Magnetic fields
The faster the velocity the greater the Magnetic field
At rest the Magnetic field becomes an Electric field

What led me more or less directly to the special theory of relativity was the convicion that
the electromotive force acting on a body in motion in a magnetic field was nothing else but an electric field

Albert Einstein 1953

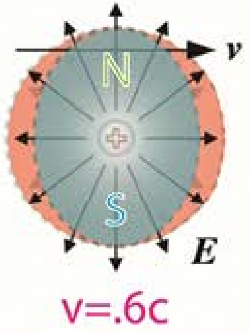

**The SR theoretical model of a spherical charged body being relativisitically distorted as the
the source of the Magnetic moments for charges in motion is incorrect**

Lorentz velocity dependent factors relate only to Kinetic EM mass-energies of motion

M + It is the KEM field mass-energies that produce Magnetic moments in relativistic particles due to the motion of ElectroStatic Matter (12 + loop inductive geometries) = **RE**

Stationary Leptons are 12pi charged standing waves [Matter] topologies with NO secondary Kinetic EM mass-energies geometries

Matter is Lorentz invariant

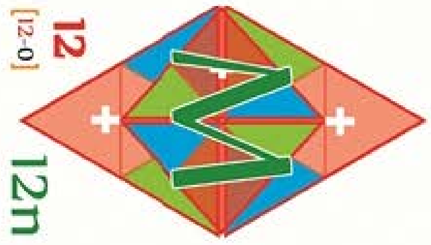

Stationary charges have Electric Fields and neutralised Magnetic fields

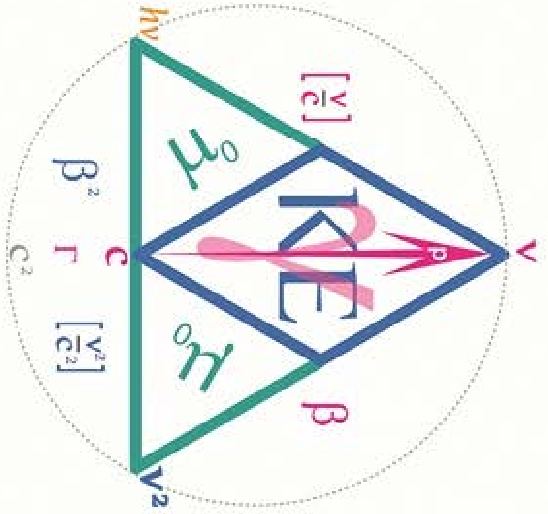

The speed of light in a vacuum [c] is the limit for electrically accelerated particles
Tetryonic geometry reveals it is NOT the limit of acheivable velocities

Moving Leptons are 12pi charged standing waves [Matter] topologies with Kinetic EM mass-energies geometries

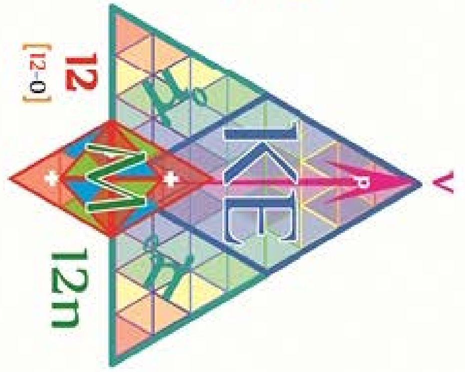

Moving charges have Electric fields and Magnetic moments

Copyright ABRAHAM [2008] - all rights reserved

ε_0

Tetryonic relativistic motion

The [K]EM mass-energy content of a particle in motion is velocity dependent its rest Matter toplogy & charge geometries are velocity invariant

μ_0 μ_0

In a frame in which the particle is at rest, we can detect only electric field

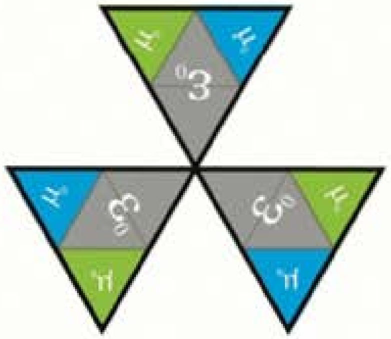

rest Matter

all Magnetic dipoles are neutral dipole pairs

Charge is a velocity invariant measure of the quantised angular momenta iof any spatial co-ordinate system

The total EM mass-energy momenta of any particle in motion is velocity dependent and is specific to particular spatial co-ordinate systems

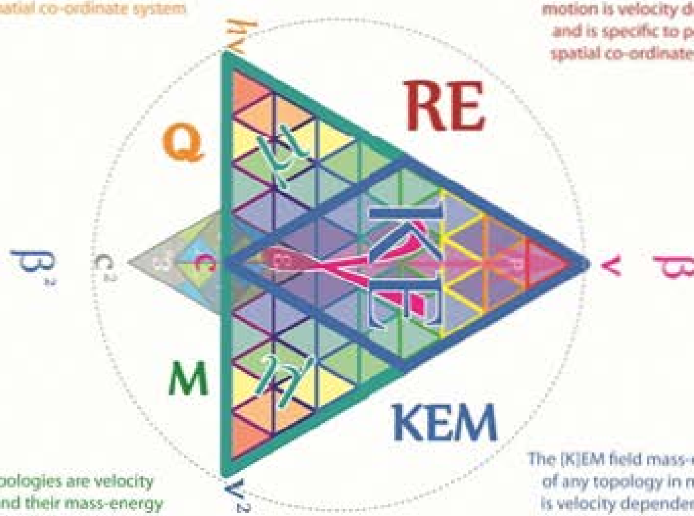

Matter topologies are velocity invariant and their mass-energy content can only be measured in c^4 spatial co-ordinates

The [K]EM field mass-energies of any topology in motion is velocity dependent and subject to Lorentz correction within in its c^2 spatial co-ordinates

It is well demonstrated that the fields of charged particles appear different in different inertial frames

In a frame in which the particle is moving, we also see a magnetic field.

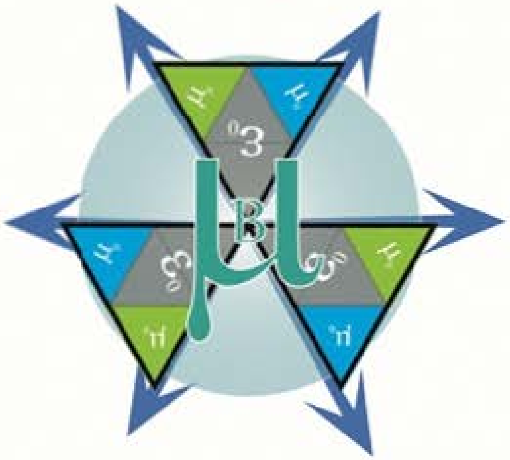

rest Matter+KEM

relativistic KEM fields produce a secondary magnetic moment

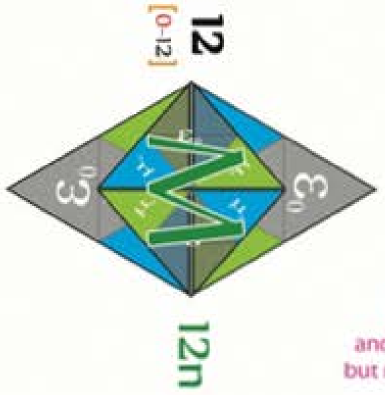

$$Mv^2 = KEM = hv^2$$

Tetryonics has revealed that the Kinetic Electro-Magnetic [KEM] field of a particle in motion has a distinct geometric identity of its own rather than being the relativistic distortion of a charged EM particle

Virtual particles and action-at-a-distance are negatived by Tetryonic KEM field mass-energy momenta and the problem of simultanity is avoided because a second particle responds not directly to the first particle, but rather to the first particle's extended, velocity related [K]EM field generated by its motion at its own position.

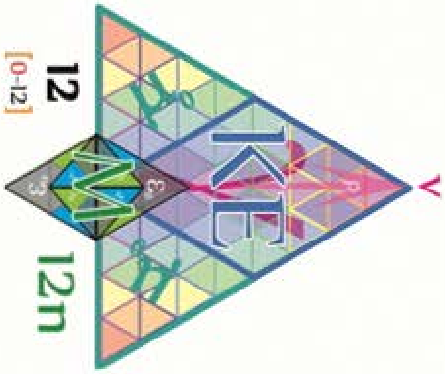

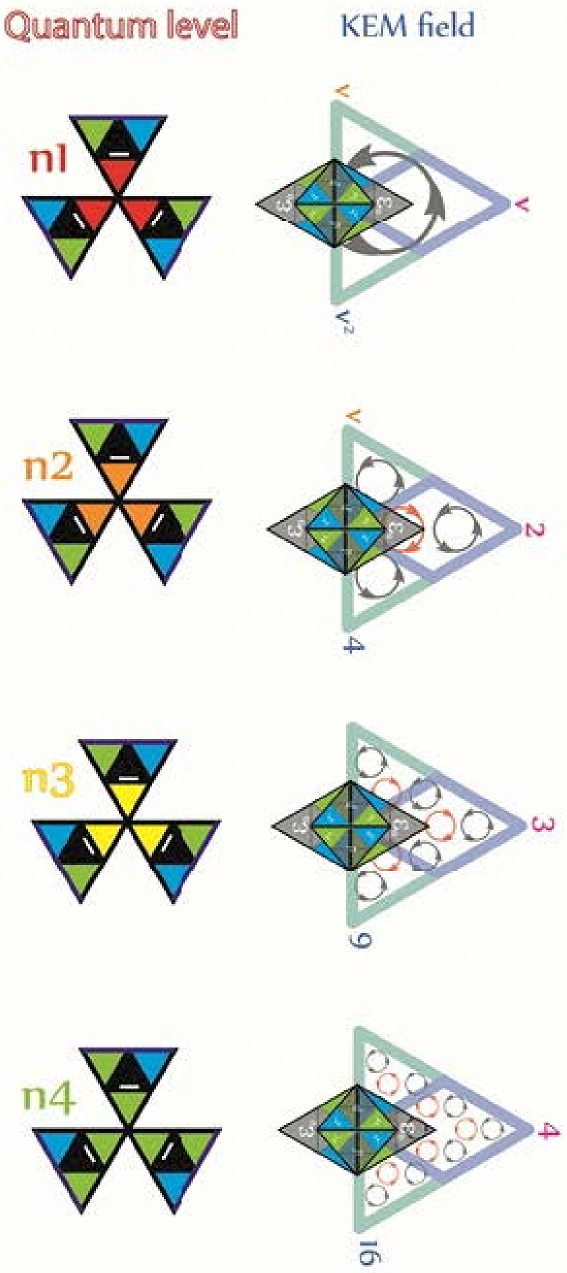

James Clerk Maxwell

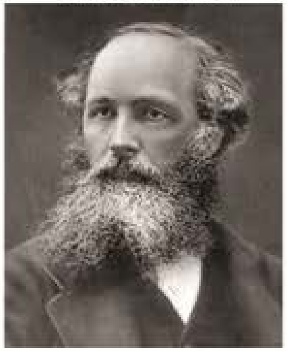

(13 June 1831 – 5 November 1879)

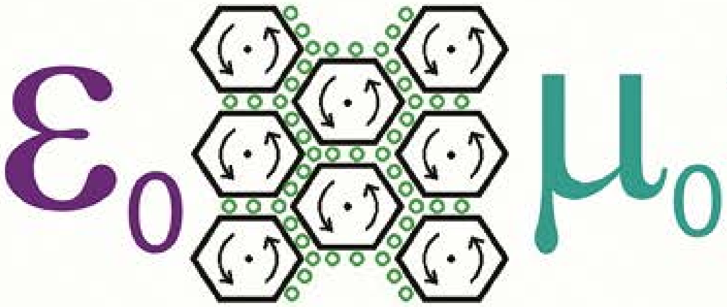

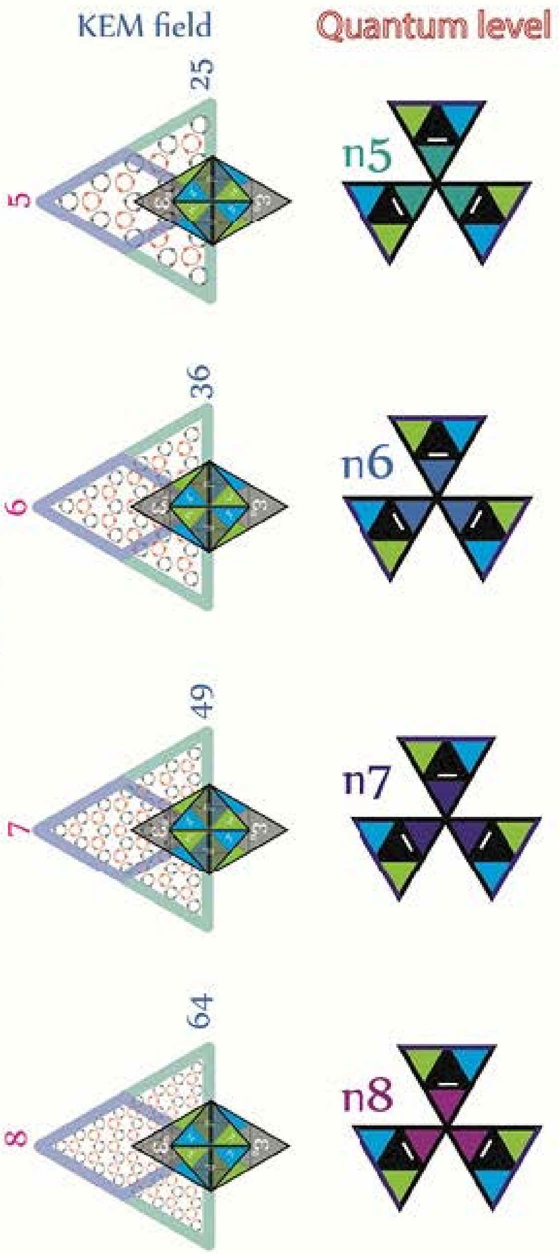

All quantum charge arrangements of EM mass-energy and force fields can be viewed as an arrangement of 'quantum idler wheels

"As early as 1857 Maxwell began to develop the idea of orienting molecular vortices along magnetic field lines, culminating in the publication of his paper 'On physical lines of force'…

He posited a honeycomb of vortices in which each vortex cell was separated from its neighbour by a layer of spherical particles, revolving in the opposite direction to the vortices.

These 'idler wheel' particles communicated the rotatory velocity of the vortices from one part of the field to another. In this ether model, the most famous image in nineteenth-century physics, the analogy provides mechanical correlates for electromagnetic quantities. The angular velocity of the vortices corresponds to the magnetic field intensity, and the translational flow of the idle wheel particles to the flow of an electric current; the field equations are based on the rotation of molecular vortices in the ether.

He emphasized that while the theory was mechanically conceivable, the model itself was provisional and temporary, even awkward, hardly 'a model of connexion existing in nature'

Equating Equilateral geometries to Quantised Angular Momentum reveals a classical mechanical model closely resembling Maxwell's hypothesised idler wheels that be be applied to any EM field of force

Virtual Particles

In physics, a virtual particle is a particle that exists for a limited time and space, it has become a commonplace mechanism in current Physical theories to provide a basis for the Force interactions between particles

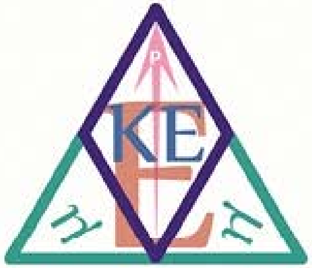

$$mv^2 = E = pv$$

The energy and momentum of a virtual particle are uncertain according to the uncertainty principle.

$$m^2v^4 = E^2 = p^4$$

The degree of uncertainty of each is inversely proportional to time duration (for energy) or to position span (for momentum).

Original Feymar diagram illustrating the exchange of a quantum between 2 electrons

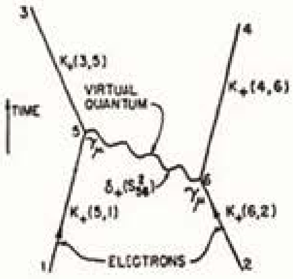

Bohr understood that if you are going to try to be mechanical, you have to show some convincing mechanics.

If you can't show some convincing mechanics, you might as well dodge all mechanics from the beginning, staying with the math.

He had learned this from Maxwell, who had done the same thing 60 years earlier.

In the 1860's, Maxwell had tried to create vortices to explain the field mechanics, but, finding himself under heavy fire from Lord Kelvin and others, he decided to give it up and go to other mathematical alternatives like quaternions instead.

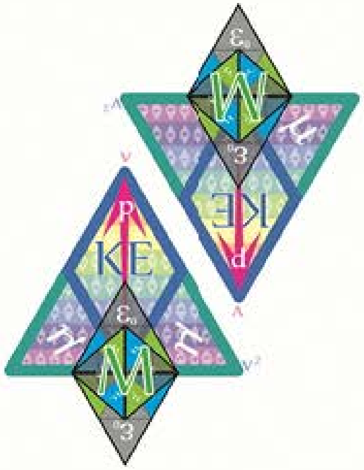

$$m^2c^4 = E^2 - p^2c^2$$

is in-correct for superpositioned 2D KEM fields the real interaction geometry is illustrated below:

$$\frac{m\Omega^2v^4}{2} = KE^2 = \frac{p^2v^2}{2}$$

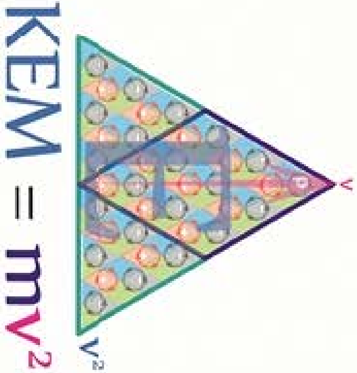

$$KEM = mv^2$$

Superpostioned EM interaction fields

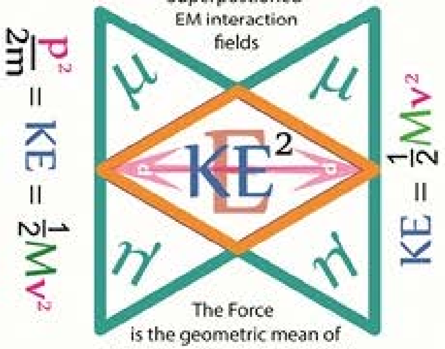

$$\frac{p^2}{2m} = KE = \frac{1}{2}Mv^2$$

$$KE = \frac{1}{2}Mv^2$$

The Force is the geometric mean of the superpositioned [KE] fields

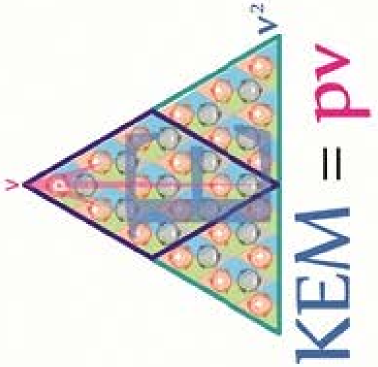

$$KEM = pv$$

The equilateral energy geometry of Tetryonics clears up this issue once and for all, virutal particles do NOT exist. The Force interactions between all Matter in relative motion is mediated by the equilateral geometries of [K]EM mass-ENERGY momenta that spread out from their source Matter through Space-time

Equilateral geometry - the Hidden Constant

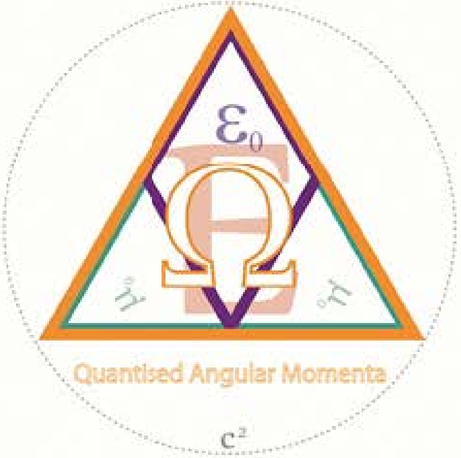
ε0
Ω
μ0
μ0
Quantised Angular Momenta
c²

$$\mathrm{kg}\; m\Omega\frac{m^2}{s} = h \;\; \mathrm{kg}\frac{m^2}{s}$$

Quantised Angular Momenta has an equilateral geometry

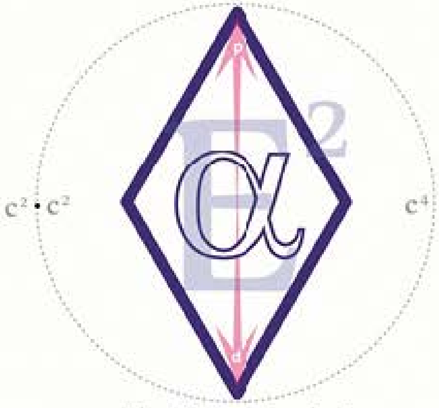
P
α
E²
d
c²•c²
c⁴

The Alpha Constant

EM Field · Planck quanta

$$1\pi\left[\left[\varepsilon_0\mu_0\right].\left[m\Omega v^2\right]\right]$$

Charge · ElectroMagnetic · mass · velocity

Tetryonics has revealed the equilateral relationships between Charge, Planck energies and the Constants showing quantised angular momentum to be the 'hidden' geometric constant driving the physics of our Universe on all scales

Alpha is the geometric coupling constant between Magnetic and Electrical energies that acts as a scaling constant between Kinetic and Potential quantum energies

It is the source of the quantum charges of all EM mass-energy momentum & inertia

Classically modelled as a 'rotational' vector Quantised Angular Momentum is in fact an Equilateral geometry

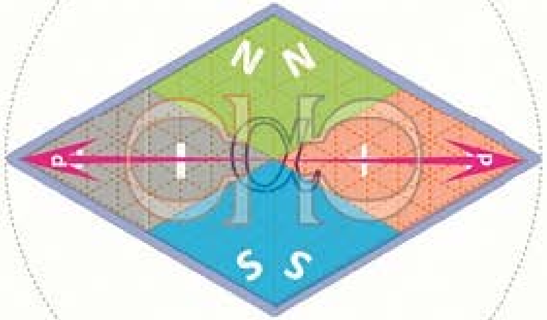

Photons are dual equilateral geometries creating a neutral charge with two opposed coupling constants

c²

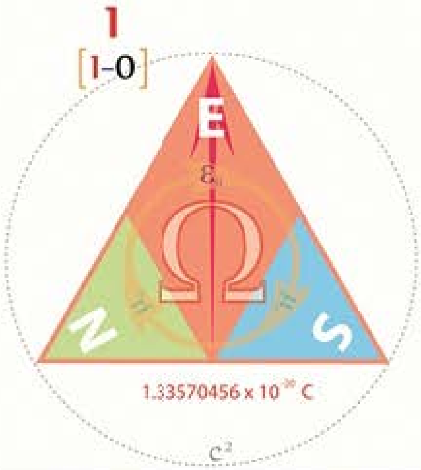
1
[1-0]
E
ε0
Ω
μ0
μ0
N
S
1.33570456 x 10[illegible] C
c²

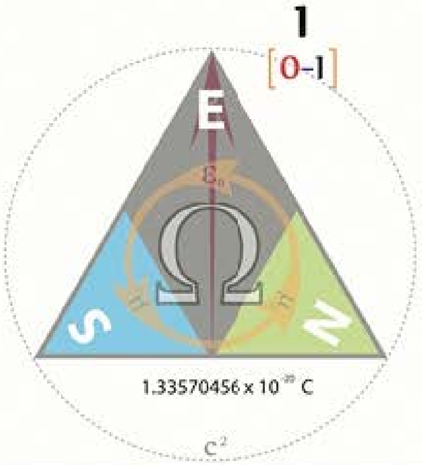
1
[0-1]
E
ε0
Ω
μ0
μ0
S
N
1.33570456 x 10[illegible] C
c²

Quantised Angular Momenta creates charged EM mass-energies

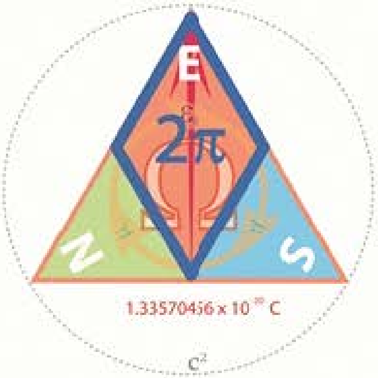

.0012

2π

.007539822

$$\alpha = \frac{e^2}{\hbar c}$$

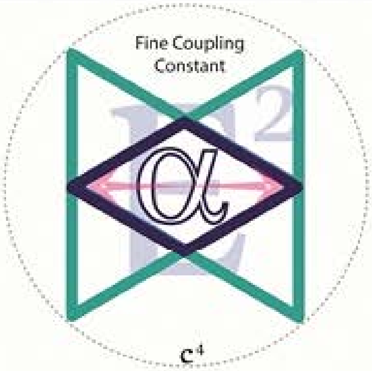

$$\alpha = \frac{e^2}{\hbar c\, 4\pi\varepsilon_0}$$

$7.297\,352\,5376(50) \times 10^{-3}$

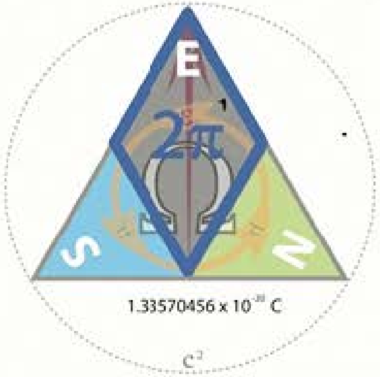

$$= \frac{1}{137.035\,999\,679}$$

132.6291192^{-1}

The ALPHA Constant is the scalar coupling field of quantised energy momenta in superpositioned EM fields that facilitates Force (linear energy momenta) exchanges, resulting in the familiar Laws of Attraction/Interaction

In physics, the fine-structure constant
is a fundamental physical constant,
namely the coupling constant characterizing
the strength of the electromagnetic interaction.

The numerical value of α is the same
in all systems of units, because α is a
dimensionless quantity.

There is a most profound and beautiful question associated with the observed coupling constant,
e the amplitude for a real electron to emit or absorb a real photon.
It is a simple number that has been experimentally determined to be close to 0.08542455.

It has been a mystery ever since it was discovered more than fifty years ago,
and all good theoretical physicists put this number up on their wall and worry about it.

Immediately you would like to know where this number for a coupling comes from:
is it related to pi or perhaps to the base of natural logarithms? Nobody knows.
It's one of the greatest damn mysteries of physics: a magic number that comes to us with no understanding by man.

We know what kind of a dance to do experimentally to measure this number very accurately,
but we don't know what kind of dance to do on the computer to make this number come out, without putting it in secretly

Richard Feynman

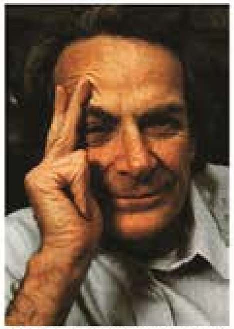

(May 11, 1918 - February 15, 1988)

equilateral energies

.0012 m²/s

Superpositioned Fields

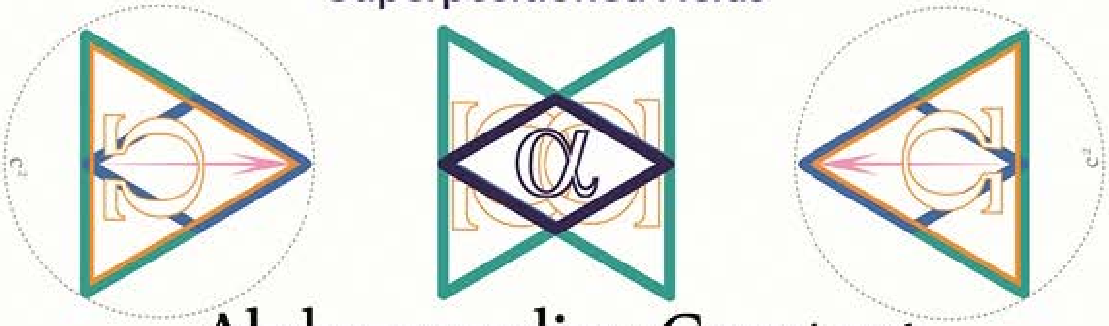

Alpha coupling Constant

0.007539822

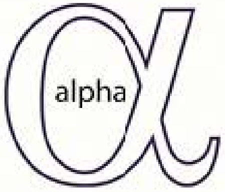

132.6291192^{-1}

You might say the "Hand of GOD" wrote that number, but "we don't know how HE pushed his pencil."

The Fine Structure Constant

$2\pi[QAM]_{.0012} = 0.007539822$

Arnold Sommerfeld introduced the fine-structure constant in 1916 as a fundamental physical constant, namely the coupling constant characterizing the strength of the electromagnetic interaction.

$2\pi[QAM]^{-1}_{.00119997} = 132.6291192$

In quantum electrodynamics, α is the coupling constant determining the strength of the interaction between electrons and their exchange photons

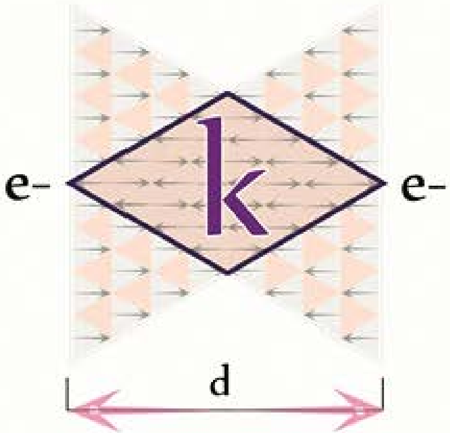

Similar Charges REPEL

10^{-6}

weak force

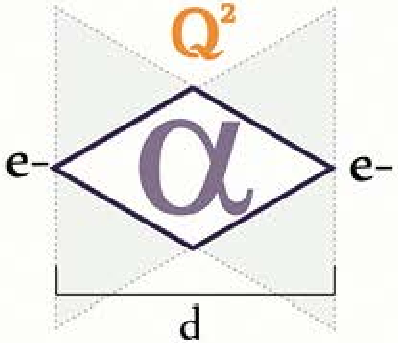

It is a dimensionless quantity that has a constant numerical value in all systems of units that is the direct result of its equilateral geometry.

1

strong force

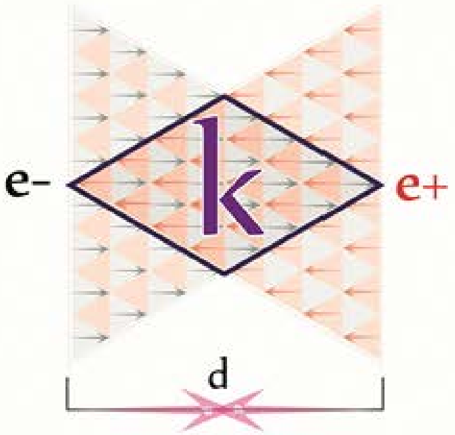

Opposite Charges ATTRACT

$$\alpha = .007539822$$

It can now be revealed to be a measure of the forces produced by, or contained in, the geometry of two superpositioned KEM $[E^2]$ field energies

The alpha constant applies equally to the coupling of Electric and Gravitational superpositioned fields

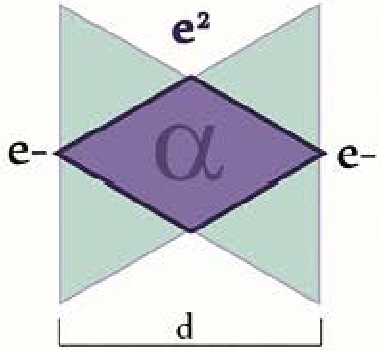

the energy needed to overcome the electrostatic repulsion between two electrons separated by a distance of d

10^{-39}

gravity

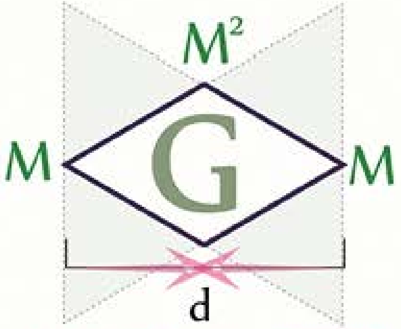

Gravitational Matter is always ATTRACTIVE

$7.5e^{-3}$

EM forces

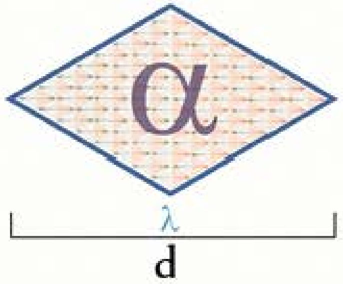

the energy of a single photon of wavelength λ that provides the energy momenta required to create a force between particles

Tetryonics 40.14 - The Fine Structure Constant

Coupling Constants

Coulomb's [k], the Fine structure constant [α] and Newton's [G] are all reflections of the same coupling constant geometries differing only by the strength of their respective superpositioned energy field densities

$$k\frac{Q_1Q_2}{r^2}=F$$

$$F=-G\frac{M_1M_2}{r^2}$$

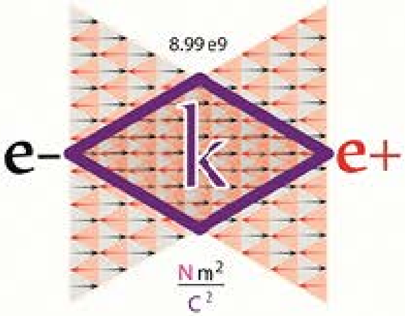

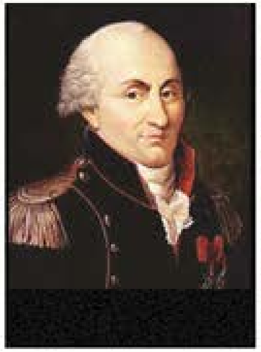

Coulomb's constant

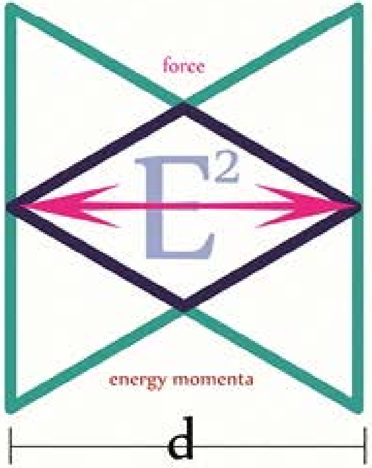

Newton's constant

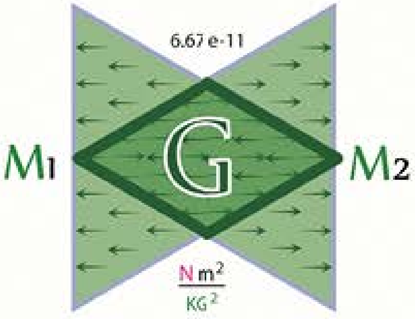

$$\beta=\left[\frac{v}{c}\right]$$

Linear force coupling constant

In physics, a coupling constant is a number that determines the strength of an interaction for superpositioned Electrical or Gravitational fields between material objects

$$\beta^2=\left[\frac{v^2}{c^2}\right]$$

Scalar energy coupling constant

Coupling Constants are dimensionless numbers

They are reflective of the linear & scalar forces produced by 2D mass-energies

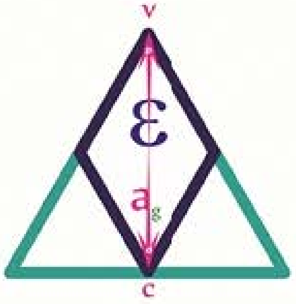

$$\frac{m}{s}\cdot\frac{s}{m}$$

electric charge coupling

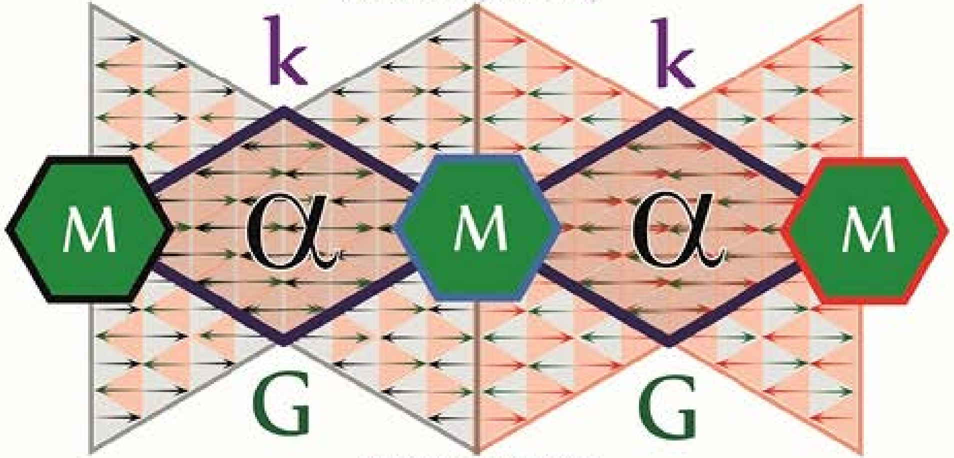

gravitational coupling

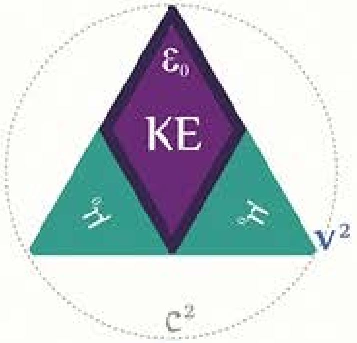

$$\frac{m^2}{s^2}\cdot\frac{s^2}{m^2}$$

Usually the Lagrangian or the Hamiltonian of a system can be separated into a kinetic part and an interaction part.
[For example, the electric charge of a particle is an Electrical coupling constant]

Copyright ABRAHAM [2008] - all rights reserved

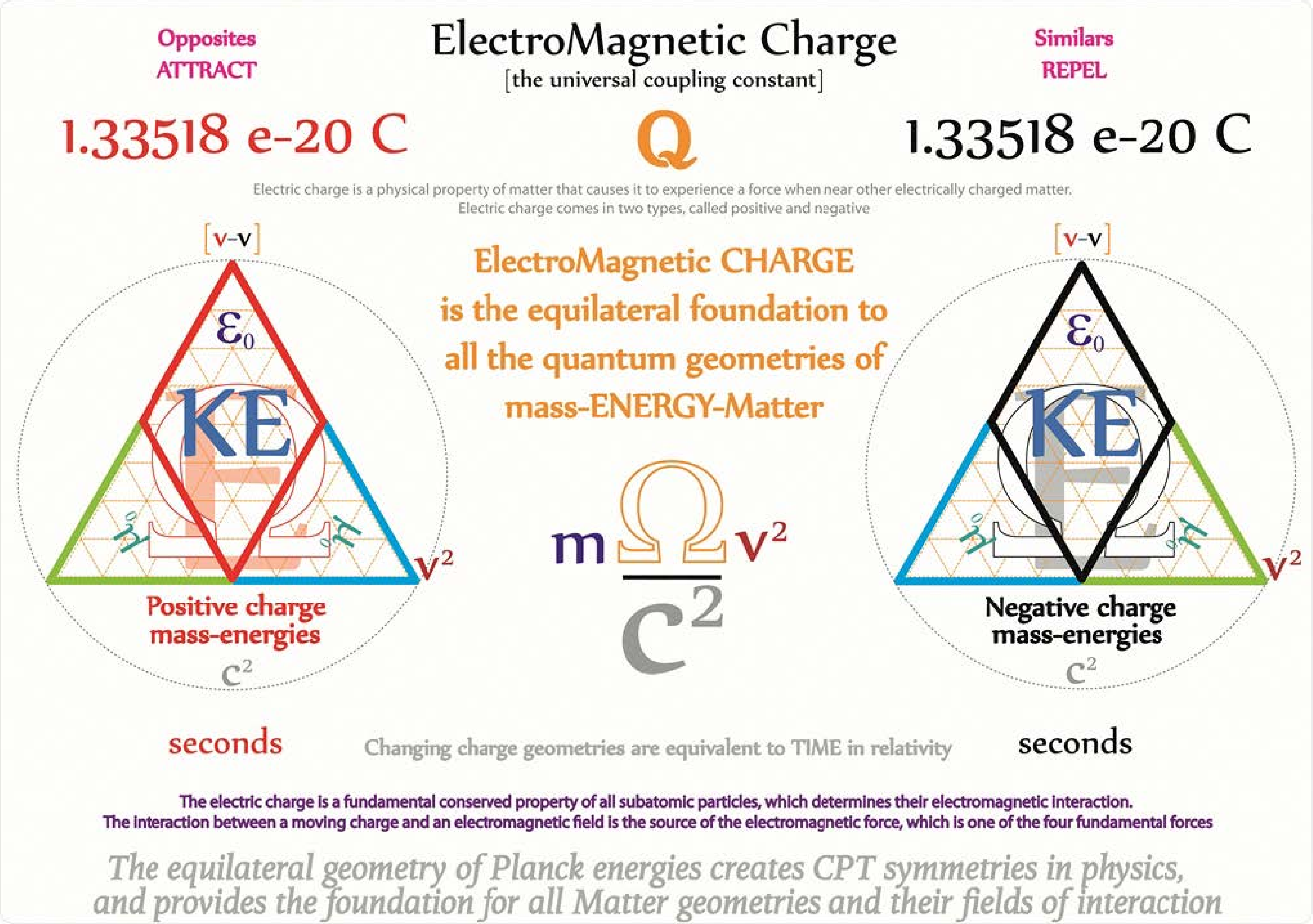

ElectroMagnetic Charge

[the universal coupling constant]

Opposites
ATTRACT

Similars
REPEL

1.33518 e-20 C

Q

1.33518 e-20 C

Electric charge is a physical property of matter that causes it to experience a force when near other electrically charged matter.
Electric charge comes in two types, called positive and negative

ElectroMagnetic CHARGE
is the equilateral foundation to
all the quantum geometries of
mass-ENERGY-Matter

$$\frac{m\Omega v^2}{c^2}$$

Positive charge
mass-energies

Negative charge
mass-energies

seconds

Changing charge geometries are equivalent to TIME in relativity

seconds

The electric charge is a fundamental conserved property of all subatomic particles, which determines their electromagnetic interaction.
The interaction between a moving charge and an electromagnetic field is the source of the electromagnetic force, which is one of the four fundamental forces

The equilateral geometry of Planck energies creates CPT symmetries in physics,
and provides the foundation for all Matter geometries and their fields of interaction

Tetryonics 40.16 - Charge the universal Coupling Constant

Copyright ABRAHAM [2008] - all rights reserved

Einstein–Podolsky–Rosen paradox

was a thought experiment that attempted to challenge the Copenhagen interpretation of Quantum physics

Tetryonics provides a complete model of all Energy forms and Wave-Particle interactions, allowing a clear understanding of previously mysterious actions and processes in Quantum Mechanics

Albert Einstein

(14 March 1879 – 18 April 1955)

Showing that it is possible to know the both Position and Momenta of Particles and to model EM wave geometries and interactions in Quantum Physics

Boris Yakovlevich Podolsky

(1896–1966)

Δx

Δp

Nathan Rosen

(March 22, 1909 – December 18, 1995)

The EPR experiment involved two systems that initially interact with each other and are then allowed to separate before being measured.

Measuring the Position of the Electron determines the Position of the Positron

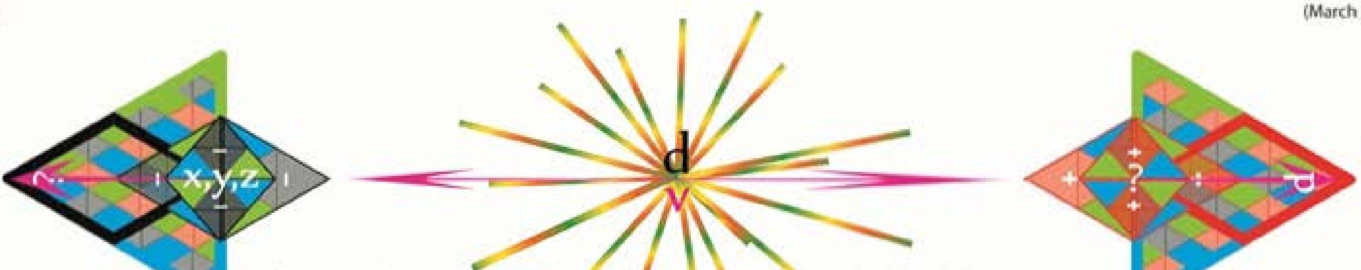

Measuring the Momentum of the Positron determines the Momentum of the Electron

Quantum entanglement occurs when particles such as photons, electrons, etc interact physically and then become separated; this type of interaction is such that each resulting member of a pair is properly described by the same quantum mechanical state, which is indefinite in terms of important factors such as position, momentum, spin, polarization, etc

The EPR paper shows that measuring one feature of a entangled system, e.g., the momentum of one of the pair of particles, will reveal the same feature of the other particle - thus providing a mechanism for determining both the momentum and position of both particles simultaneously

Thus providing a theoretical indication that either the Uncertainty Principle was incorrect or that our understanding of Quantum Mechanics was incomplete

Copyright ABRAHAM [2008] - all rights reserved

Spooky interactions-at-a-distance & the transfer of information

Quantum entanglement is sasid to occur when particles such as photons, electrons, and other forms of EM mass-energies interact physically and then become separated;
the type of interaction is such that each resulting member of a pair is properly described by the same quantum mechanical description (state),
which is indefinite in terms of important factors such as position, momentum, spin, polarization, etc.

Through the equilateral geomtry of Tetryonics the mass-energy-Matter of any system of particles
can be modelled in order to reveal the true EM mechanics of their quantum interactions

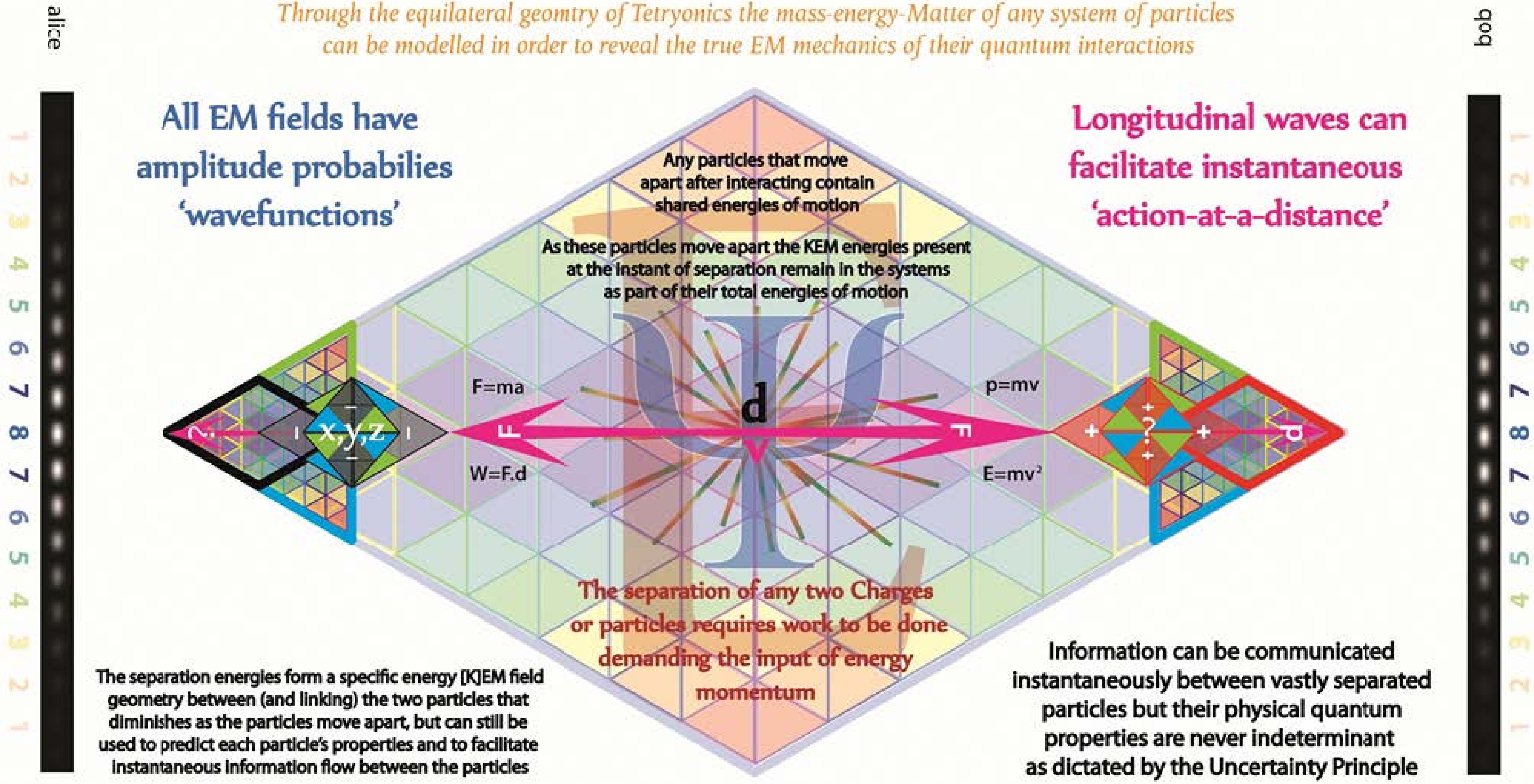

All EM fields have amplitude probabilies 'wavefunctions'

Longitudinal waves can facilitate instantaneous 'action-at-a-distance'

Any particles that move apart after interacting contain shared energies of motion

As these particles move apart the KEM energies present at the instant of separation remain in the systems as part of their total energies of motion

The separation of any two Charges or particles requires work to be done demanding the input of energy momentum

The separation energies form a specific energy [K]EM field geometry between (and linking) the two particles that diminishes as the particles move apart, but can still be used to predict each particle's properties and to facilitate instantaneous information flow between the particles

Information can be communicated instantaneously between vastly separated particles but their physical quantum properties are never indeterminant as dictated by the Uncertainty Principle

When two entangled particles are separated using longitudinal energies, the energy momenta of separation provides a mechanism for the establishment and near-instantaneous communication of the physical wavefunction/properties of one particle to the other, irrespective of the spatial or temporal [space-time] separation of the two 'entangled' partner particles

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

changes to energy momenta

accelerate photo-electrons

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_i^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is
a mathematical description
of Tetryonic energy geometry

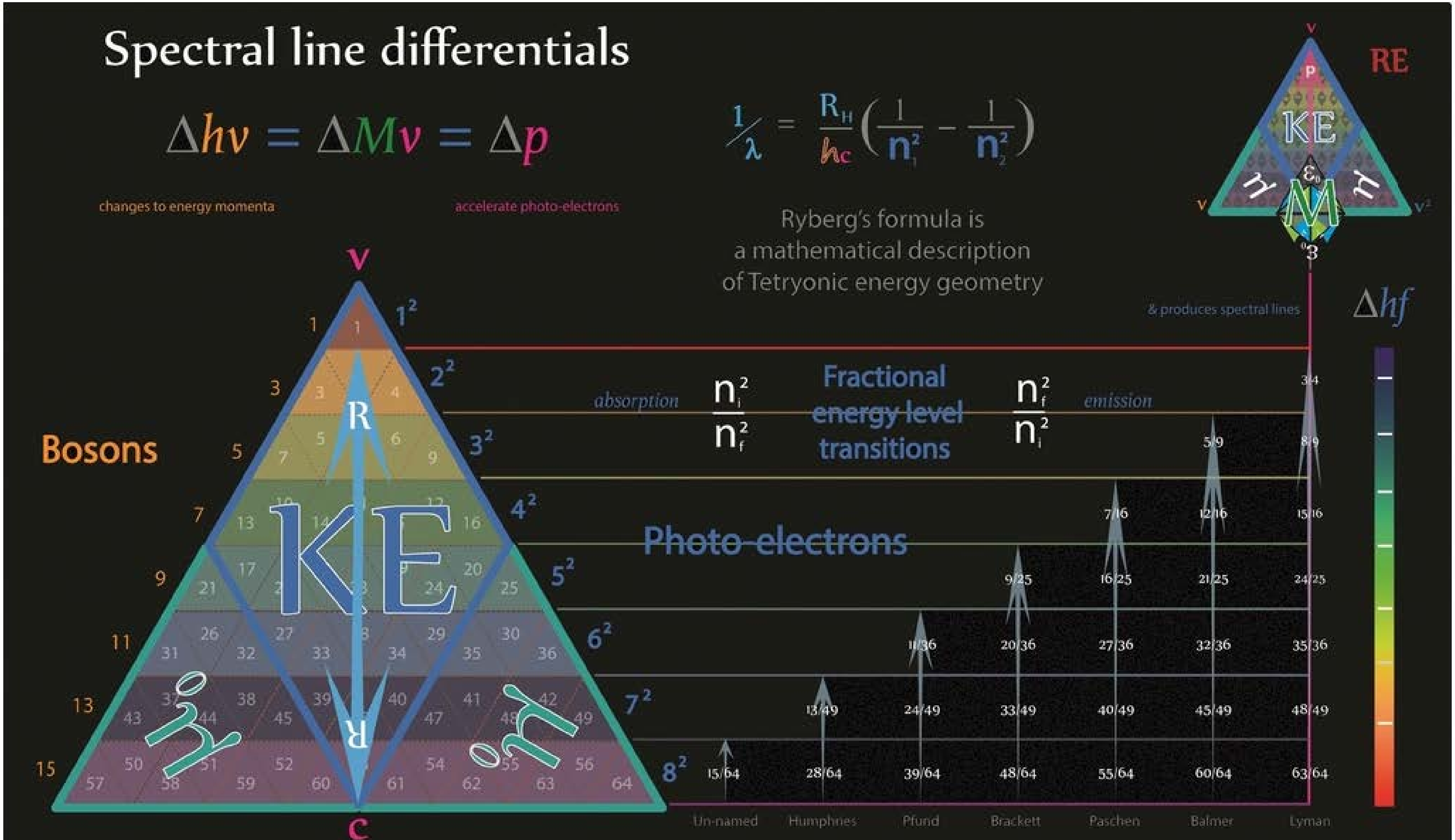
RE
KE
M
v
v²
3
& produces spectral lines
Δhf
Bosons
v
c
R
KE
1
3
5
7
9
11
13
15
1²
2²
3²
4²
5²
6²
7²
8²
absorption
$\frac{n_i^2}{n_f^2}$
Fractional energy level transitions
$\frac{n_f^2}{n_i^2}$
emission
Photo-electrons
3/4
5/9 8/9
7/16 12/16 15/16
9/25 16/25 21/25 24/25
11/36 20/36 27/36 32/36 35/36
13/49 24/49 33/49 40/49 45/49 48/49
15/64 28/64 39/64 48/64 55/64 60/64 63/64
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

Ryberg's constant reflects the changing
energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

All of the transitions of photo-electrons
bound into a Hydrogen atom can be
revealed in the fractional geometry
of KEM field energies

KEM field energies

Tetryonics 40.19 - Spectral line differentials

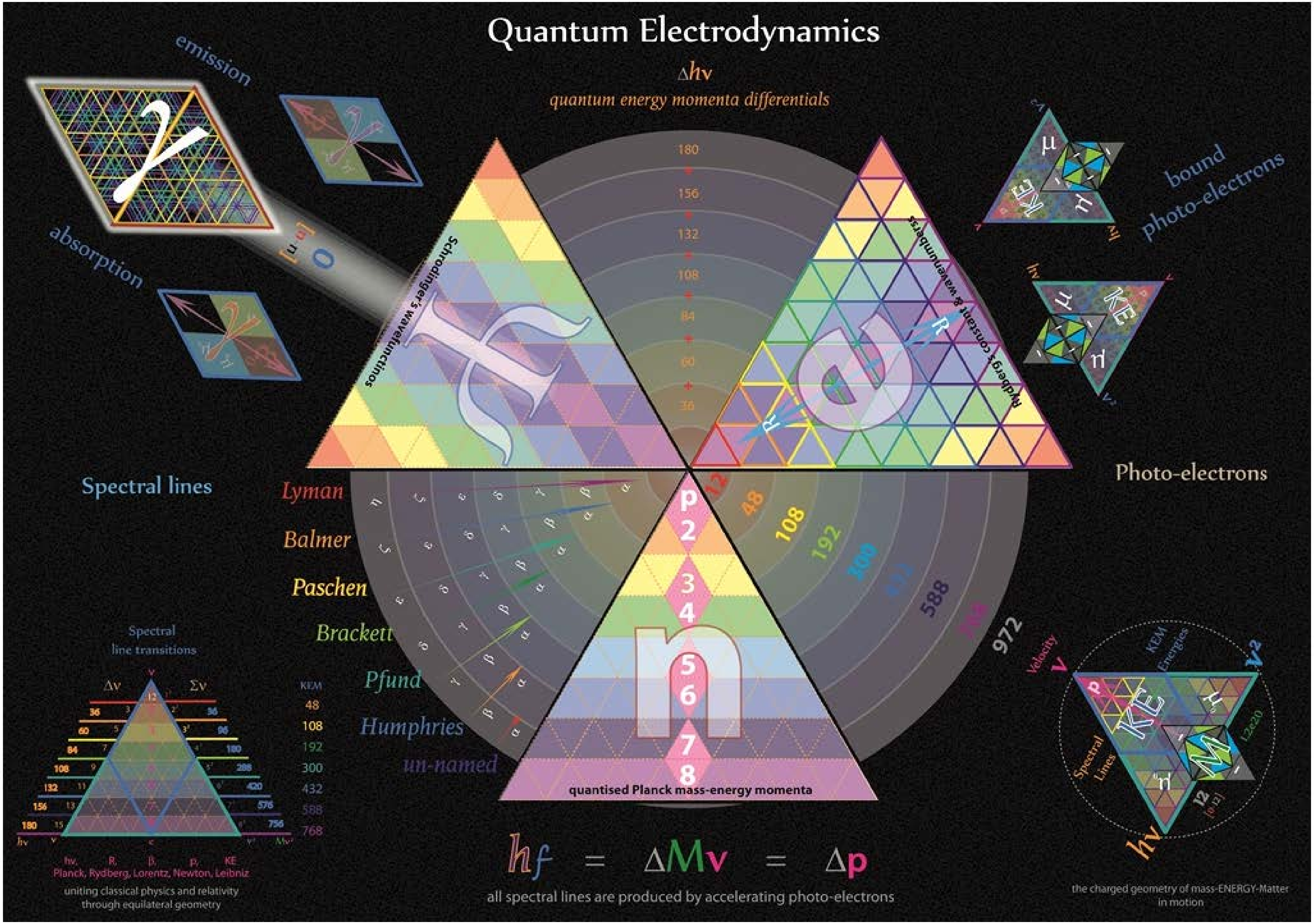

Quantum Electrodynamics

Δhv

quantum energy momenta differentials

Spectral lines

Photo-electrons

$$hf = \Delta Mv = \Delta p$$

all spectral lines are produced by accelerating photo-electrons

Spectral line transitions

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

the charged geometry of mass-ENERGY-Matter
in motion

emission
absorption
λ
bound photo-electrons
μ
KE
Schrodinger's wavefunctions
h
Rydberg constant & wavenumbers
R
180
156
132
108
84
60
36
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
η
ζ
ε
δ
γ
β
α
p
2
3
4
5
6
7
8
n
quantised Planck mass-energy momenta
12
48
108
192
300
432
588
768
972
Δv
Σv
v
12
36
60
84
108
132
156
180
1
3
5
7
9
11
13
15
36
96
180
288
420
576
756
hv
c
v²
Mv²
KEM
48
108
192
300
432
588
768
Velocity
v
KEM Energies
v²
12x20
Spectral Lines
12
hv
M

Tetryonics 40.20 - Quantum Chemistry follows

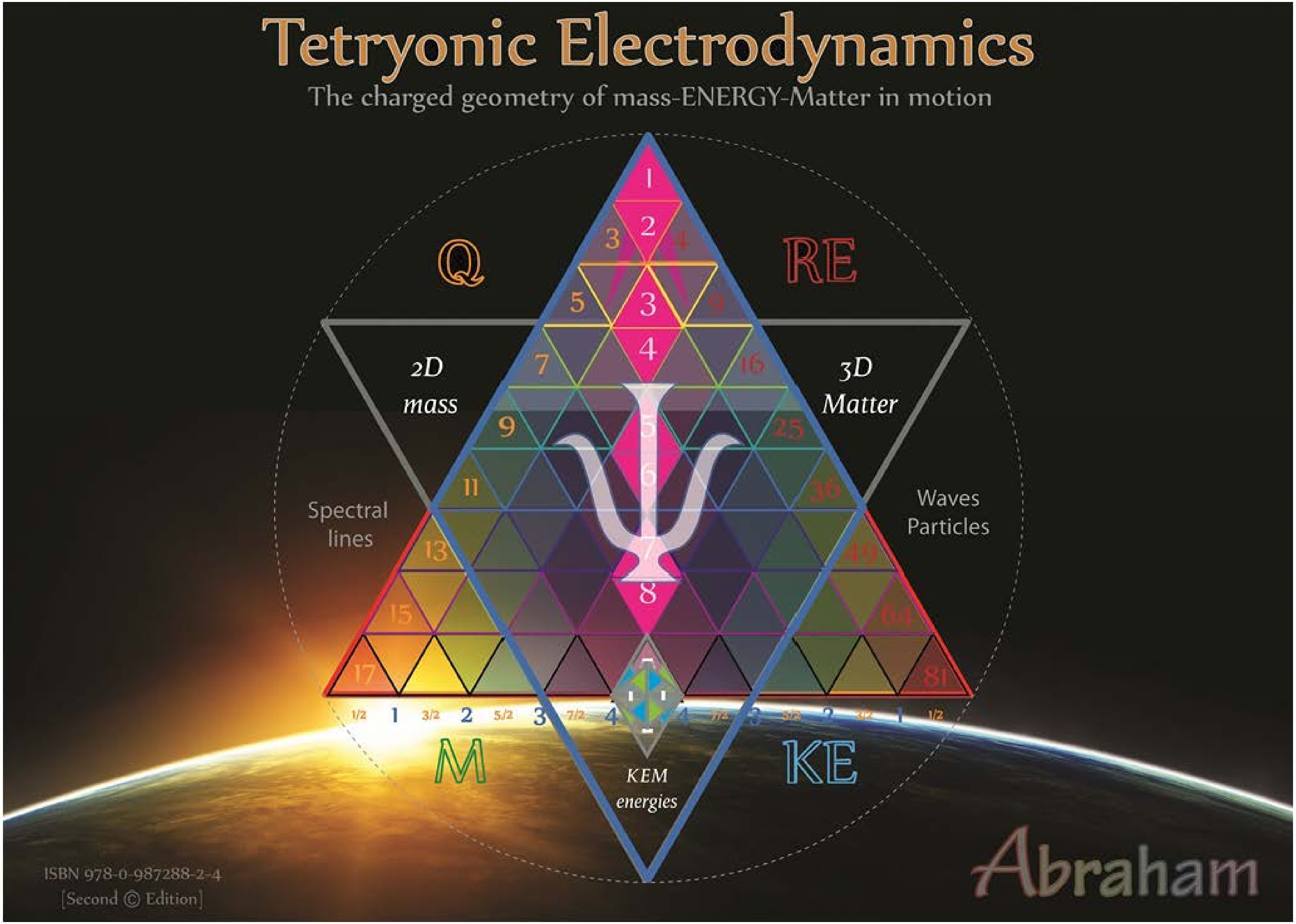
Tetryonic Electrodynamics
The charged geometry of mass-ENERGY-Matter in motion
Q
RE
2D
mass
3D
Matter
Spectral
lines
Waves
Particles
1
2
3
4
5
6
7
8
3
4
5
9
7
16
9
25
11
36
13
49
15
64
17
81
1/2
1
3/2
2
5/2
3
7/2
4
4
7/2
3
5/2
2
3/2
1
1/2
M
KE
KEM
energies
ISBN 978-0-987288-2-4
[Second © Edition]
Abraham

Tetryonics 40.21 - Quantum Electro-Dynamics

Copyright ABRAHAM [2008] - All rights reserved

TETRYONICS

The charged topology of periodic & compound Matter

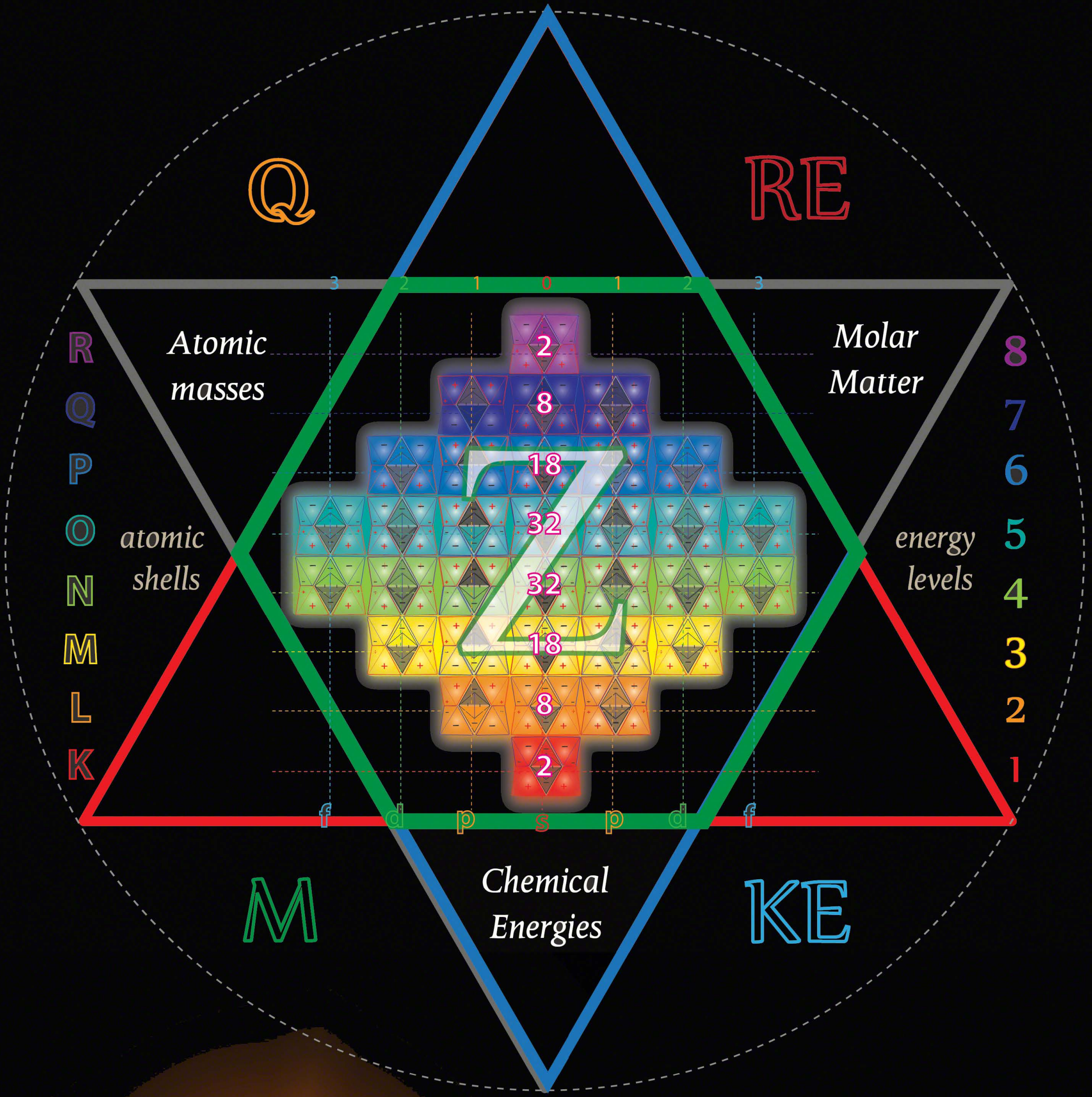
Q
RE
3
2
1
0
1
2
3
Atomic masses
Molar Matter
R
Q
P
O
N
M
L
K
atomic shells
energy levels
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
Chemical Energies
M
KE

Foundational Quantum Chemistry

Abraham

ISBN 978-0-987288-3-1

[Second Edition © 2012]

Copyright ABRAHAM [2008] - All rights reserved

Protons + *Neutrons*

u+d+u d+u+d

All electrons, quarks and Baryons are made up of 4n+ (Tetryonic) standing-wave EM fields.

As well as having nett Tetryonic charged topologies ranging between [+24] ~ [-12] they all posses distinct ELECTRIC FIELDS that are concentrated in 3 apex points as indicated in the illustrations

These points result from the orientation of Electric apexes and orthagonal Magnetic dipole field edges that make up each particle's externalised EM fields.

The Positive and Negative electric apex points, obey the Law of Interaction forcing separated nuceli to combine due to their individual nett Tetryonic charges and provide a means of orienting nuclei to each other to create larger particles [elements, allotropes and compounds]

External Magnetic (H) fields can interact with the integral magnetic (B) dipoles of Tetryonic particles forcing them to orientate in specific directions to facilite chemical bonding [nuclear forces]

Additionally, external Electric fields can interact with the integral electric fields attracting or repelling them depending on the polarity of the external electric field [Electrostatics]

External energies can be induced into these integral EM fields via inductive coupling or the absorption of spectral photons in turn leading to an increase in the strengthes of the integral EM apexes in turn increasing the Strong Nuclear Force.

Residual Electro-Magnetic Forces

allow Neutrons and Protons to attract via the opposite Electric charge points created by their constituent Quarks in order to create Elementary Nuclei

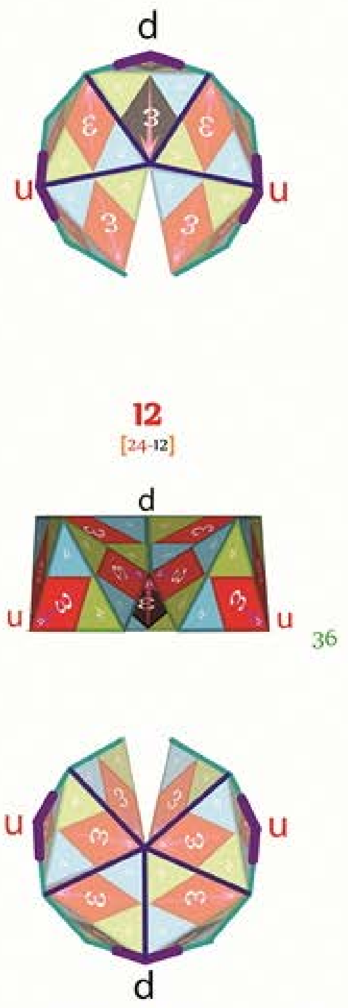

The orientation of the component Electric fields within 3D Matter creates macroscopic force apexes via externalised 'E-points'

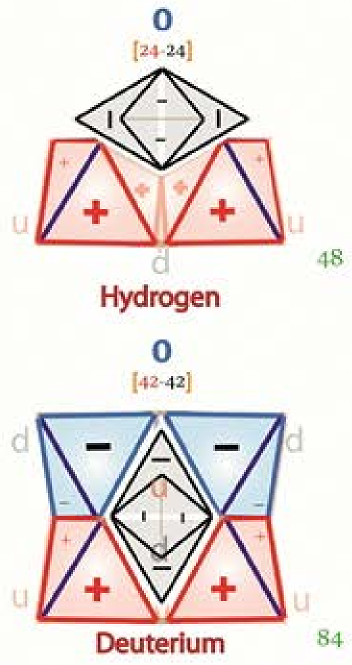

The Strong Nuclear force binds Matter together

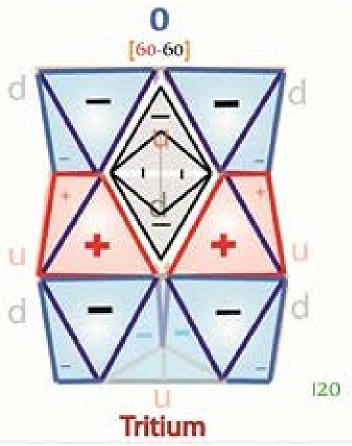

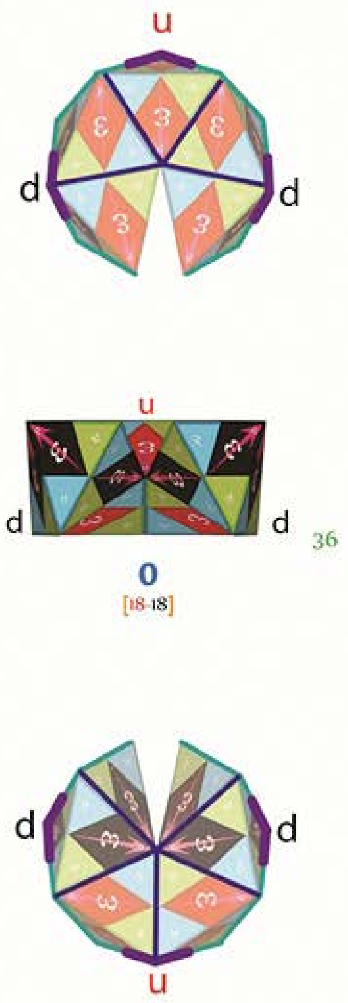

The orientation of the component Magnetic fields within 3D Matter creates macroscopic force apexes via externalised 'M-dipoles'

Copyright ABRAHAM [2008] - All rights reserved

Nucleonic residual EM force

The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei

The residual e-field apexes and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei

Neutrons

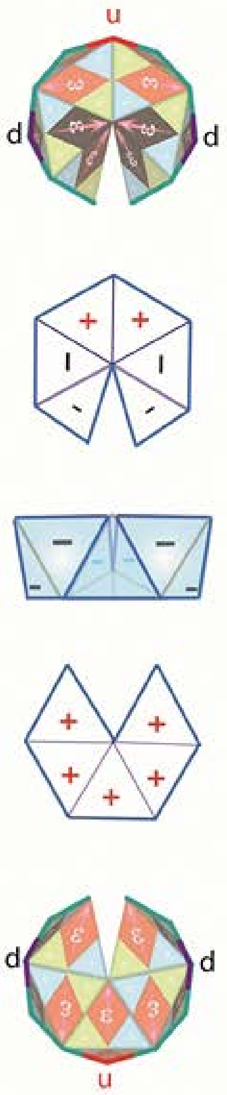

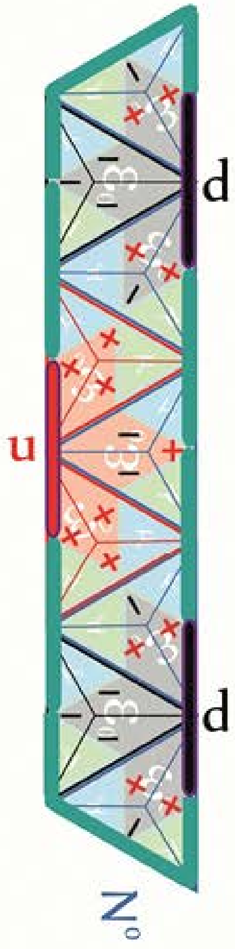

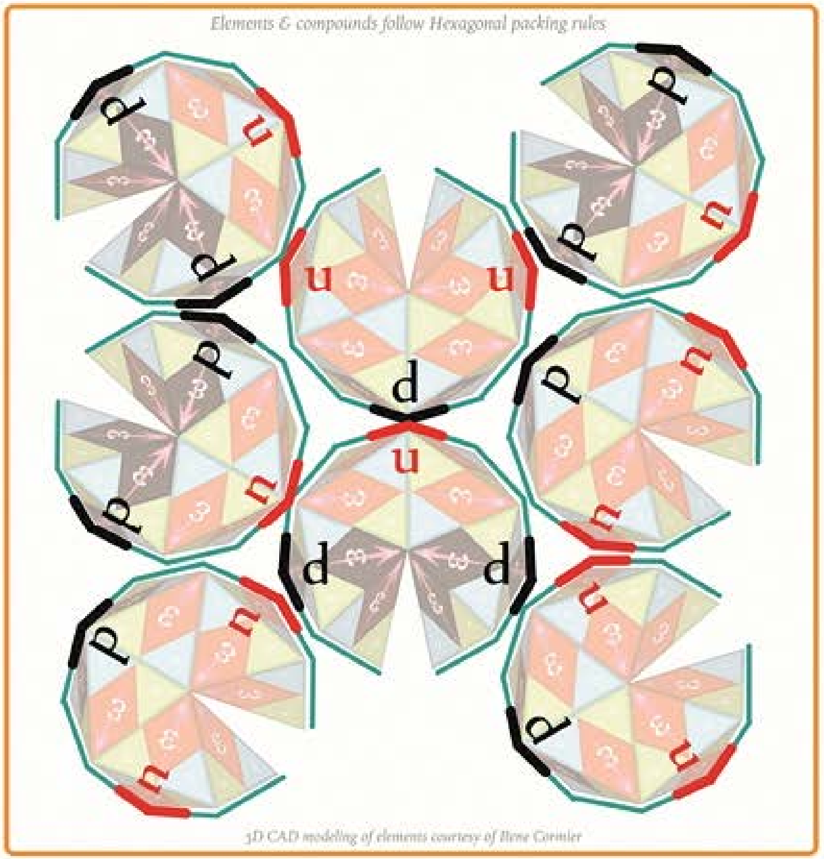

Elements & compounds follow Hexagonal packing rules

3D CAD modeling of elements courtesy of Ilene Cormier

Protons

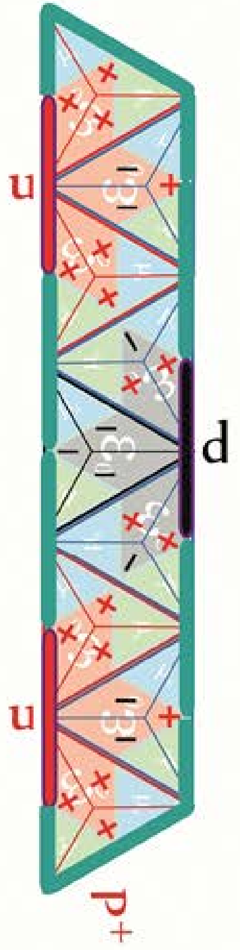

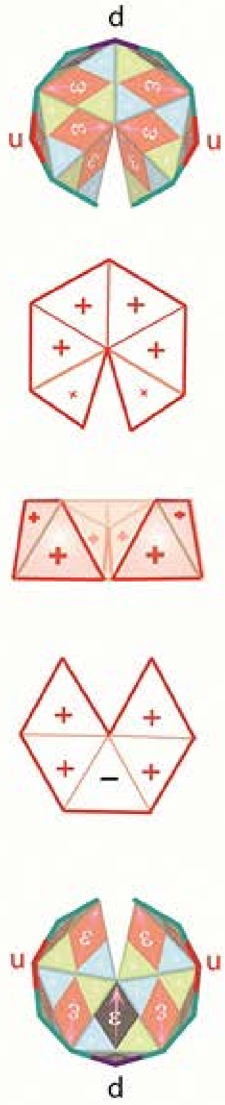

E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements

UP Quark
Positive Electric field apex

DOWN Quark
Negative Electric field apex

Insulators and Conductors

The position of electrons in Nuclei within Atomic Elements
results in the properties of Insulators or Conductors

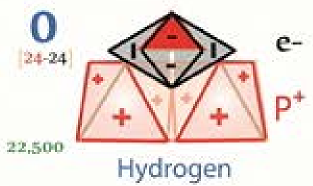
0
[24-24]
22,500
e-
P+
Hydrogen

Coloumbic forces

Electrons are attracted to the residual EM net (+12) positive charge of Protons or n[+12] unbalanced Ionic charges of nuclei

Conductor

Electrical energies move around the material via boson exchanges and electron movement

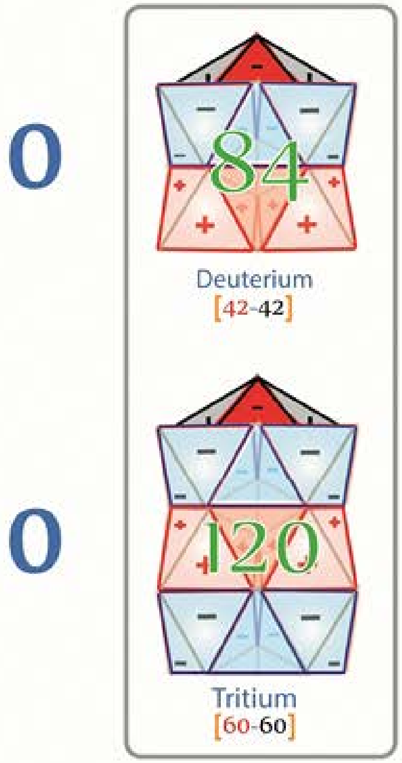
0
84
Deuterium
[42-42]
0
120
Tritium
[60-60]

≈

Insulator

Electrical energy is fixed within the nucleus as electrostatic charges & released upon demand via electron rotation/motion within the nucleus

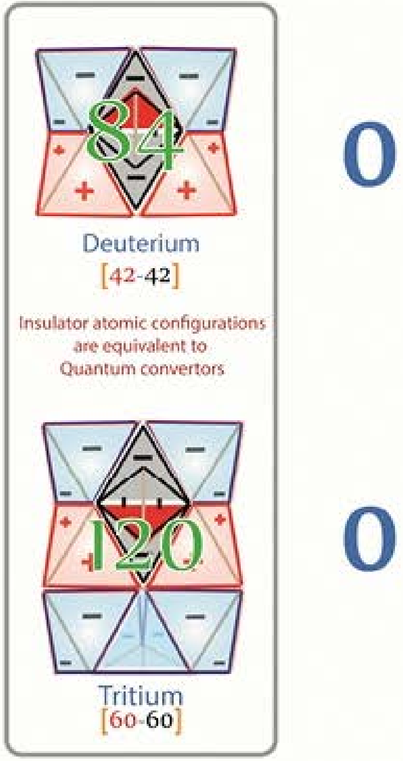
84
Deuterium
[42-42]
0
120
Tritium
[60-60]
0

Insulator atomic configurations are equivalent to Quantum convertors

Ions

Charge (energy) is moved around material via electron movement

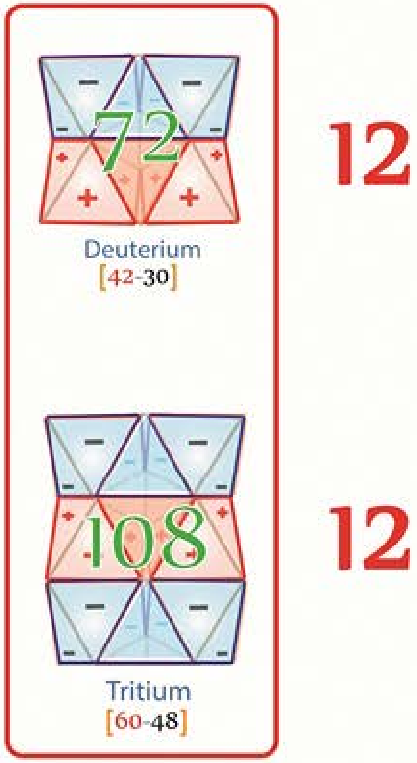
72
Deuterium
[42-30]
12
108
Tritium
[60-48]
12

Conductive materials contain 'free' electrons that can be readily or easily moved within the material

Insulator materials have electrons that are 'bound' tightly to the atoms and store charges locally where they are applied

Materials that have been ionised are more likely to become Conductors as they easily attract and bind free electrons to them

Nucleon Quark Arrangement

The nuclei arrangement of each atomic shell [quantum level] is the result of quark EM field interactions

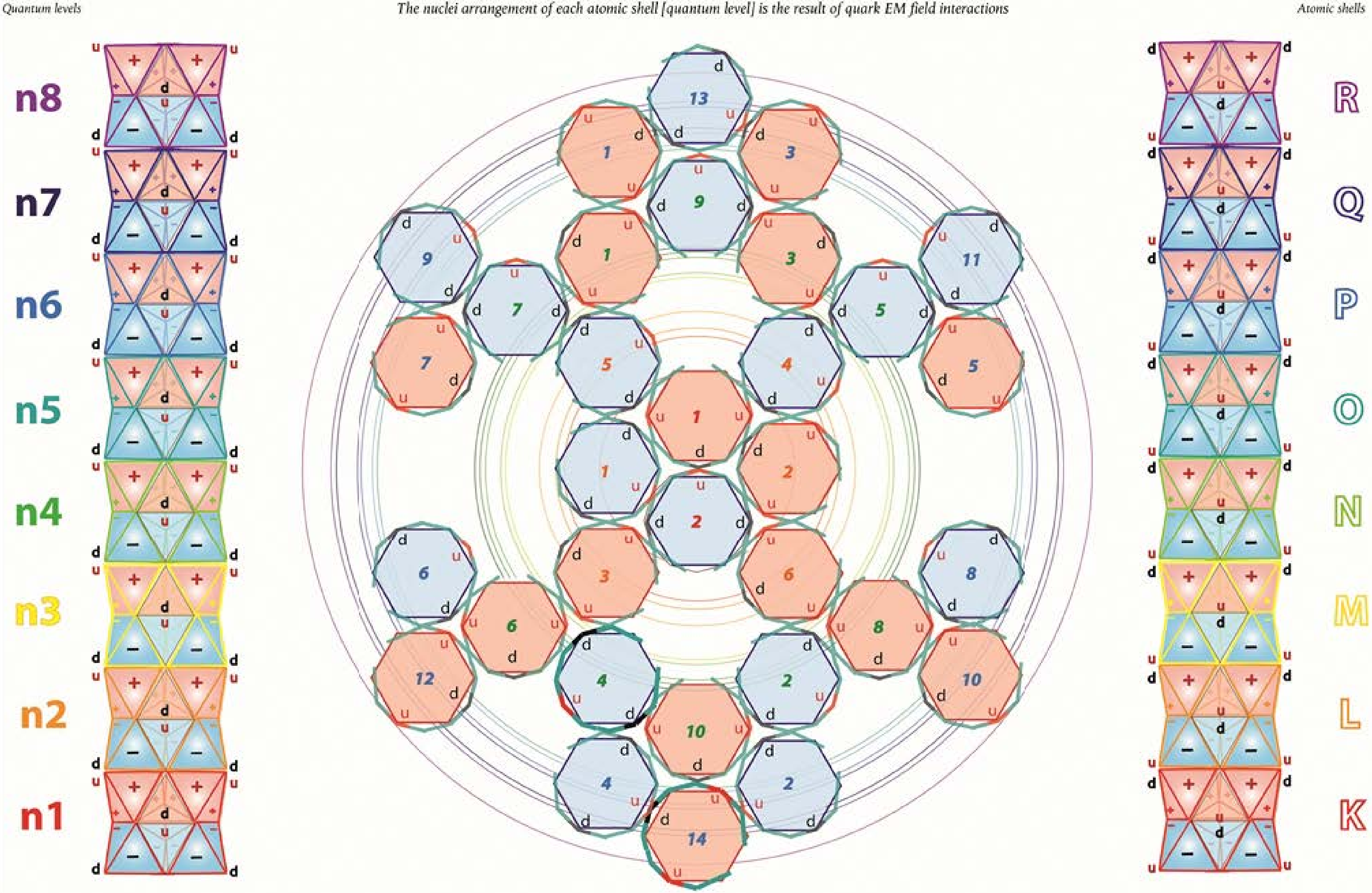
Quantum levels
Atomic shells
n8
n7
n6
n5
n4
n3
n2
n1
R
Q
P
O
N
M
L
K
u
d
+
−

Copyright ABRAHAM [2008] - All rights reserved

Nucleon Charges and Bonding

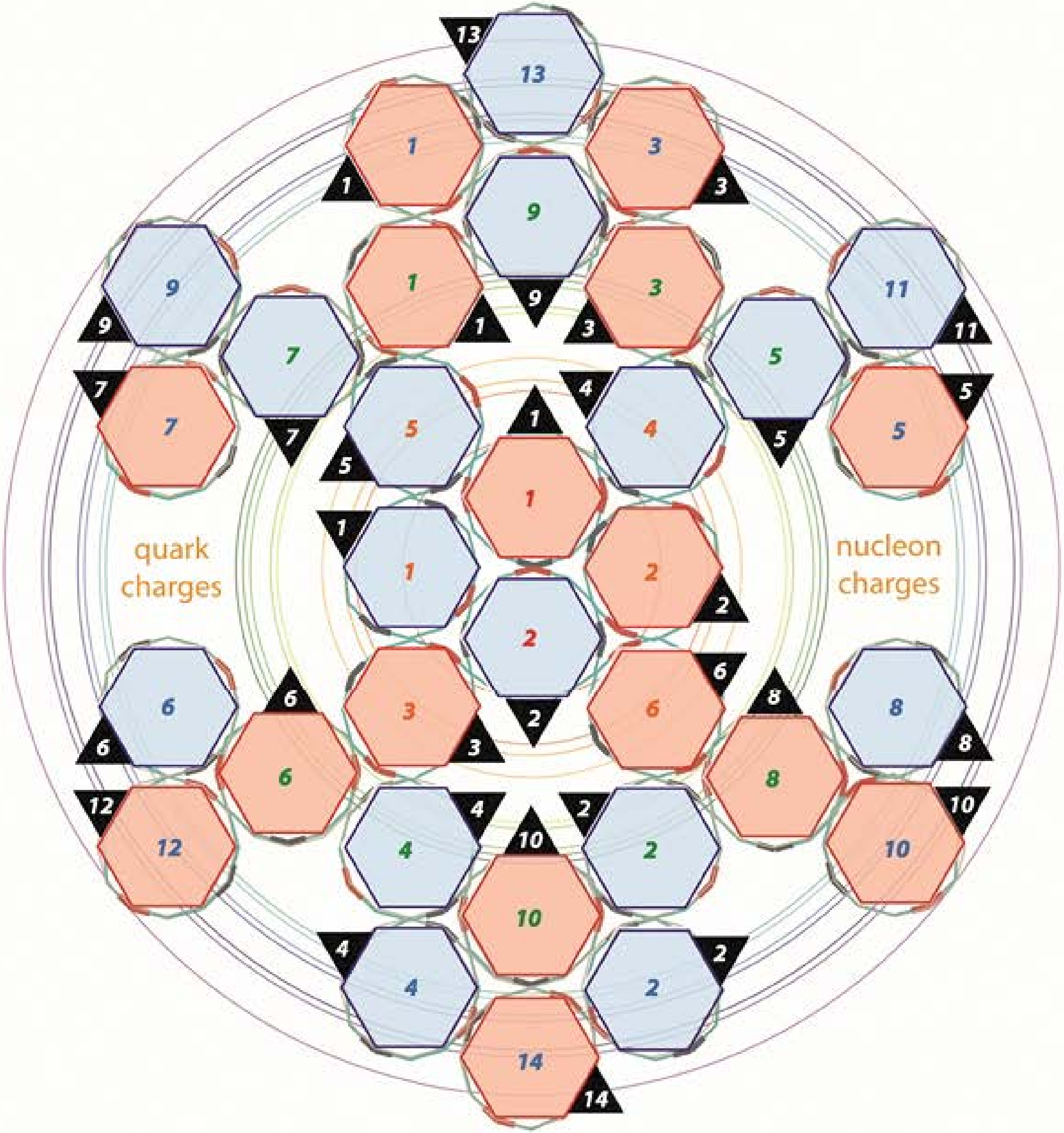
quark charges
nucleon charges

(Strong force - topological Electric Points)

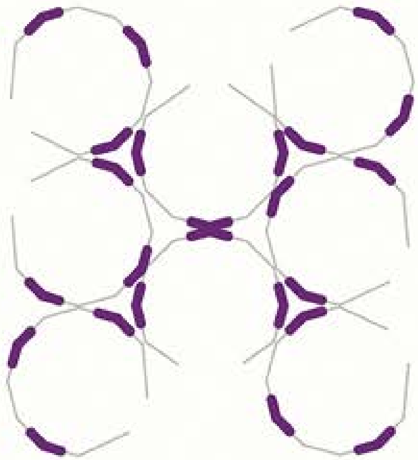

residual EM forces

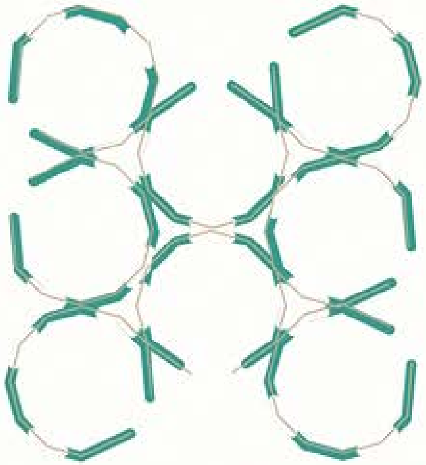

(Strong force - topological Magnetic dipoles)

Bound electron arrangements

Externally bound electrons produce sub-orbital patterns different to the electron orbitals of internally bound electrons

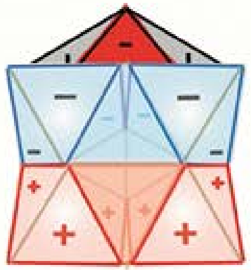

Deuterium

electrons are externally bound to the Deuteron nuclei

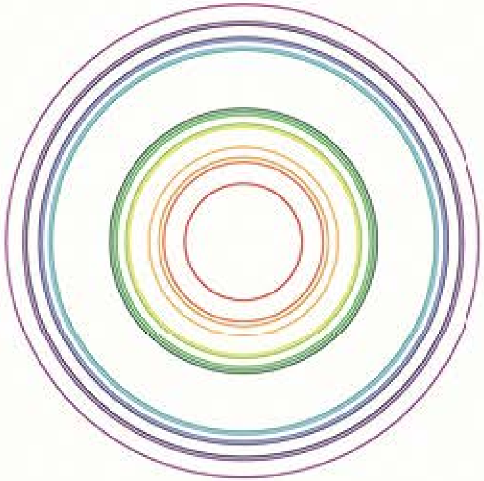

Conductors

Charges are free to move and equalise

electrons require less energy to 'break free' from Nuclei

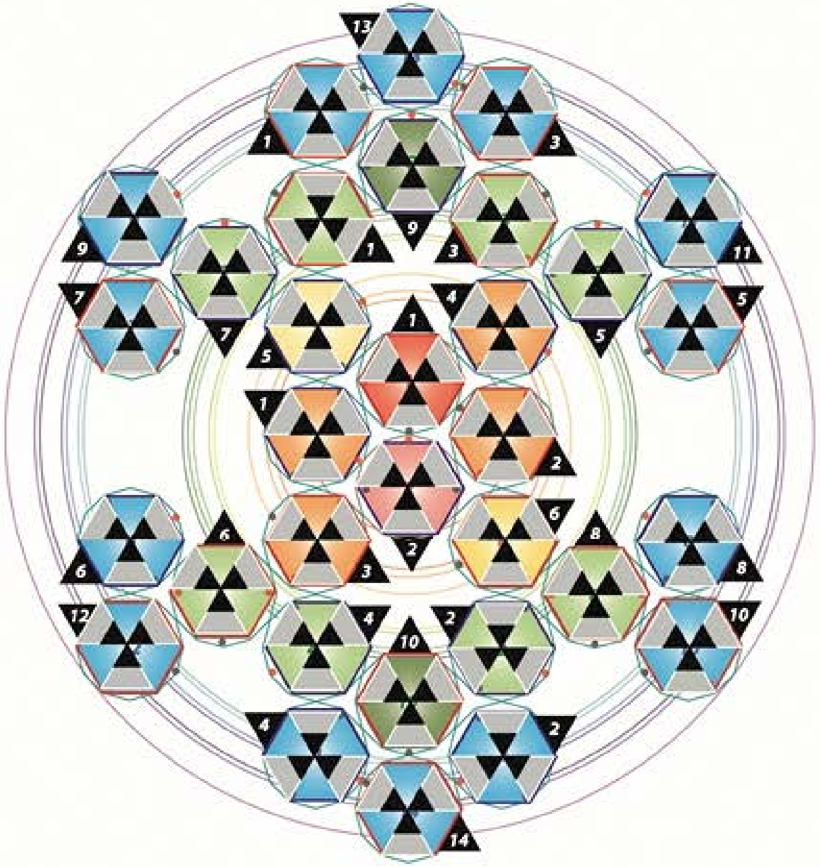

The electron orbitals of conductors are lower energies than those of insulators

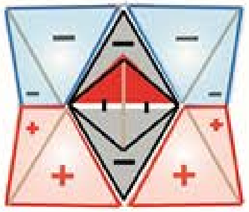

Deuterium

electrons are internally bound in the Deuteron nuclei

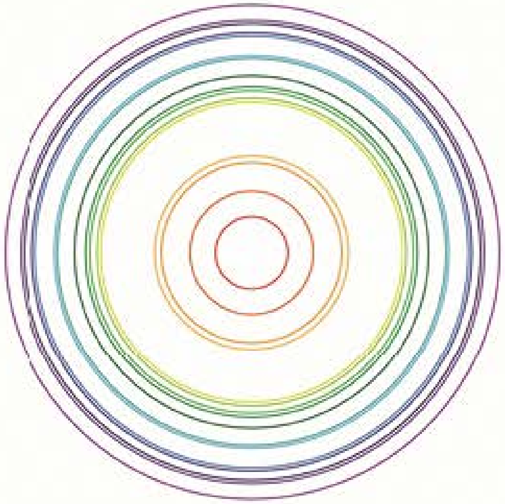

Insulators

Charges are bound to specific locations

electrons require more energy to 'break free' from Nuclei

2D mass-energy geometries form the fabric of 3D Matter topologies

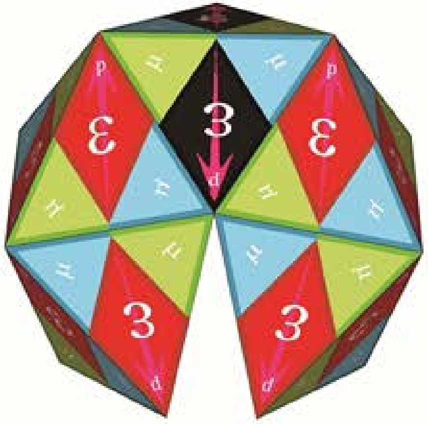

Proton

12

[24-12]

nett Charge

component charges

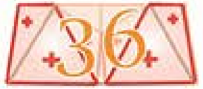

charged mass-energy geometry

20π

Matter topology

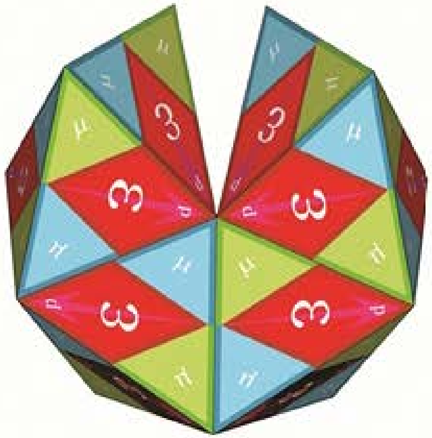

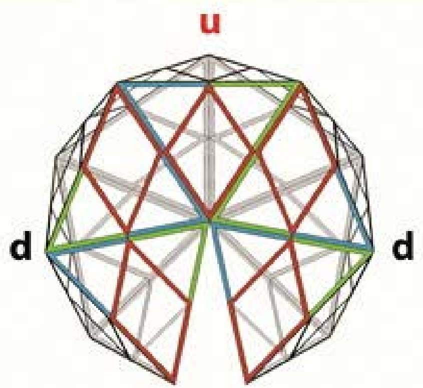

12

[42-30]

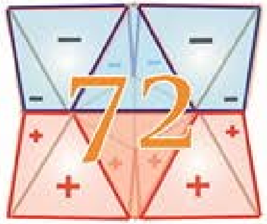

40π

Deuteron

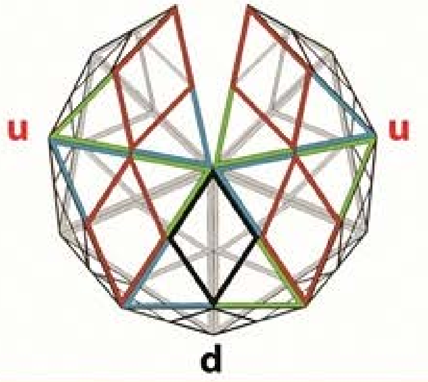

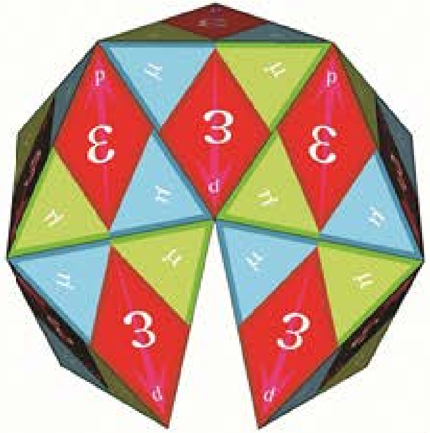

Neutron

nett Charge

component charges

0

[18-18]

charged mass-energy geometry

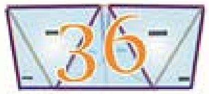

Matter topology

20π

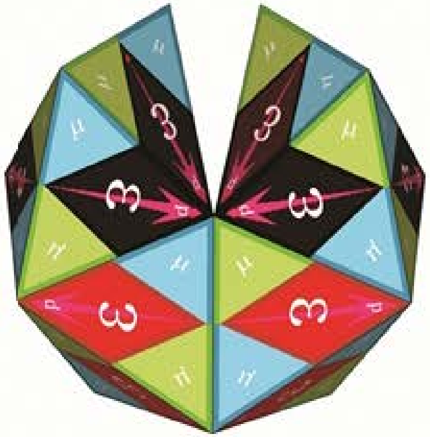

Charge provides the framework for the mass-energy geometry of Matter

Copyright ABRAHAM [2008] - All rights reserved

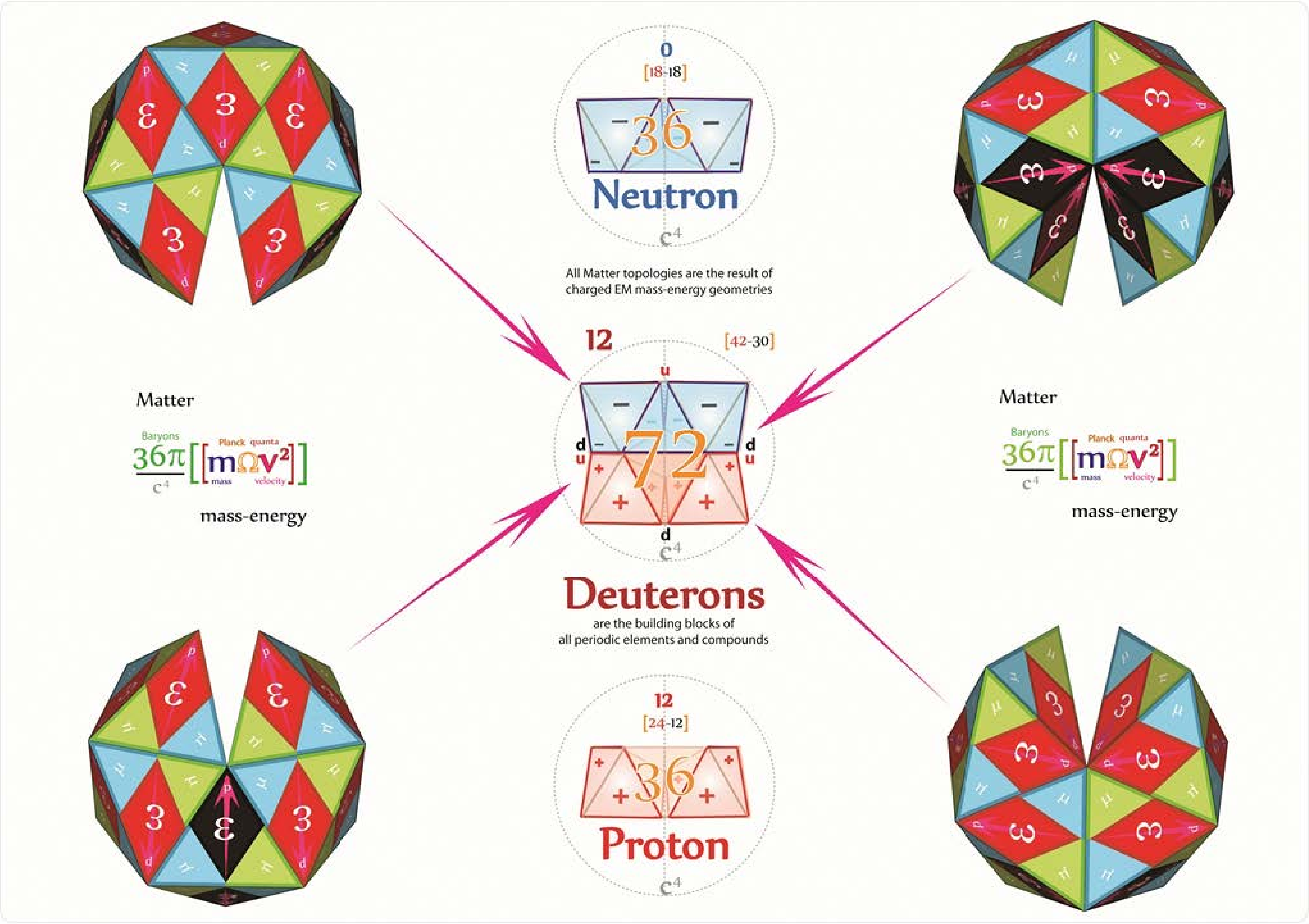
0
[18-18]
36
Neutron
c^4
12
[42-30]
u
d
u
d
u
72
d
c^4
12
[24-12]
36
Proton
c^4
Matter
Baryons
Planck
quanta
mass
velocity
mass-energy
Matter
Baryons
Planck
quanta
mass
velocity
mass-energy

$$\frac{36\pi}{c^4}[[m\Omega v^2]]$$

All Matter topologies are the result of charged EM mass-energy geometries

Deuterons

are the building blocks of all periodic elements and compounds

Copyright ABRAHAM [2008] - All rights reserved

Strong Forces and Nuclear Bonding

How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?

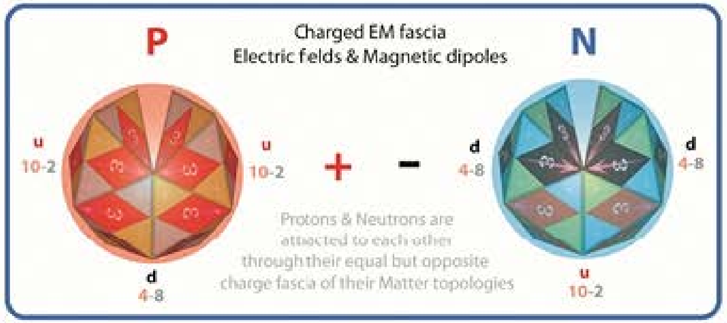
P
N
u
10-2
u
10-2
d
4-8
+
-
d
4-8
d
4-8
u
10-2

Charged EM fascia
Electric felds & Magnetic dipoles

Protons & Neutrons are attracted to each other through their equal but opposite charge fascia of their Matter topologies

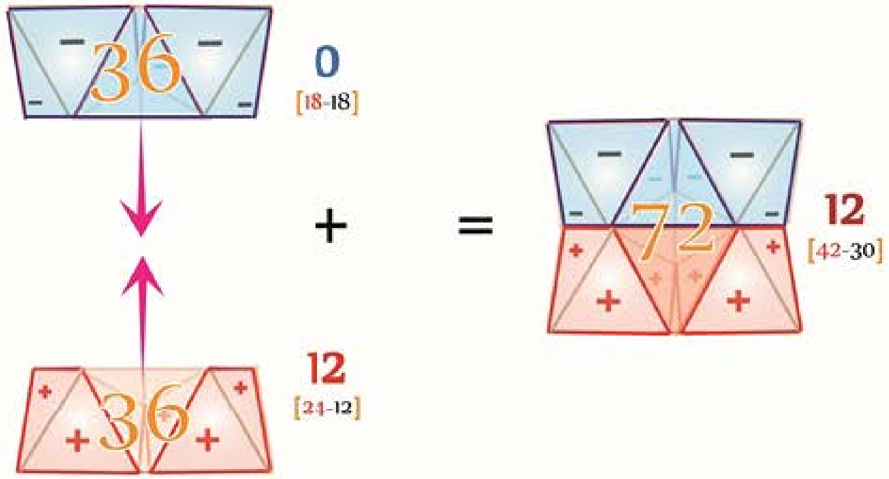
36
0
[18-18]
+
=
72
12
[42-30]
36
12
[24-12]

The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges

The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

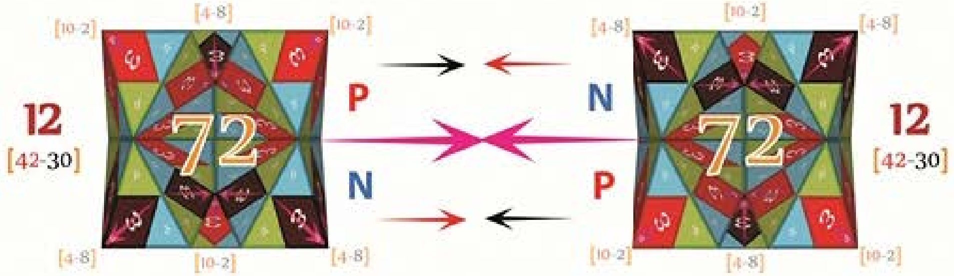
[10-2]
[4-8]
[10-2]
12
[42-30]
72
P
N
[4-8]
[10-2]
[4-8]
[4-8]
[10-2]
[4-8]
N
P
72
12
[42-30]
[10-2]
[4-8]
[10-2]

Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements

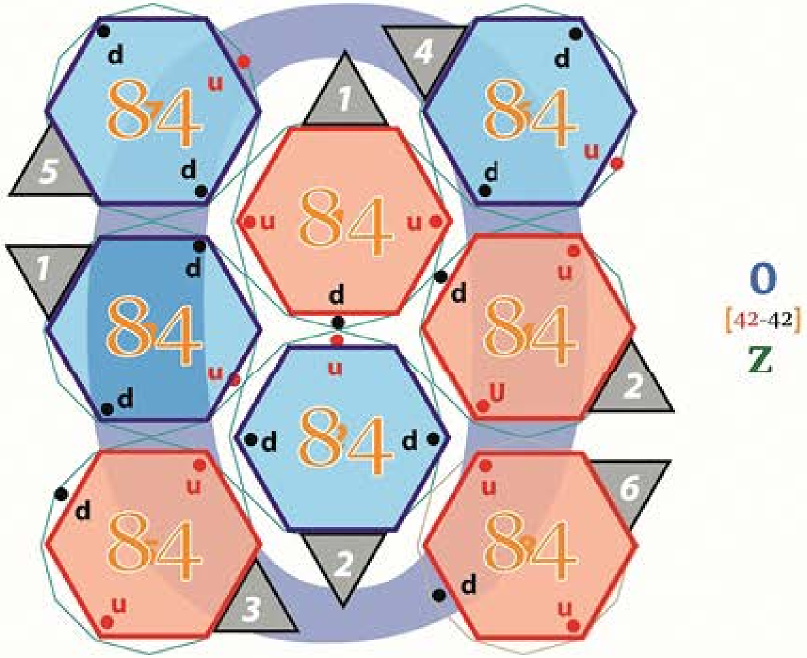
84
84
84
84
84
84
84
d
u
1
2
3
4
5
6
1
2
0
[42-42]
Z

Copyright ABRAHAM [2008] - All rights reserved

Strong Nucleonic Bonding

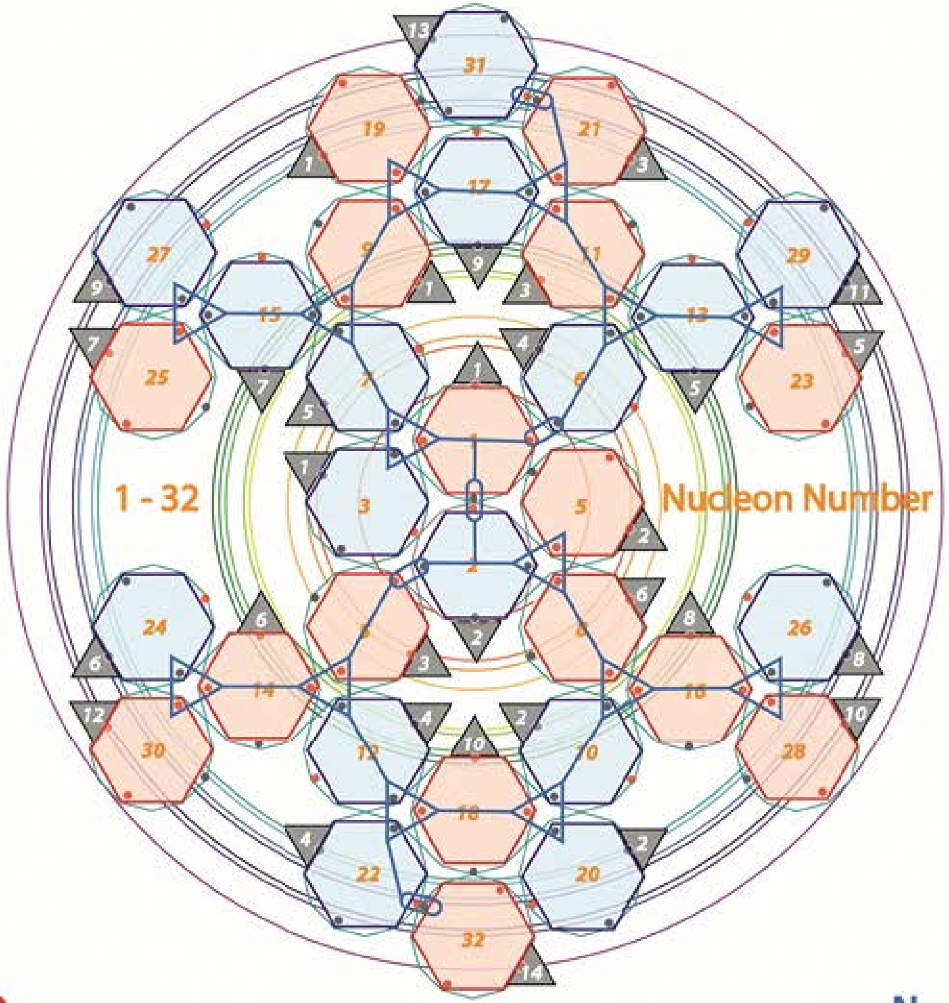

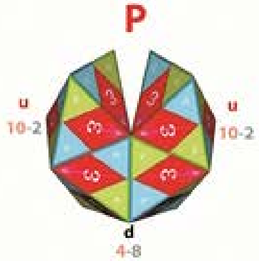

+

In addition to the Strong colour force
a separate residual binding force arises
from the external apexes formed by
Positive and Negative Electric points of quarks
in each Baryon [nuclear-chemical bonds]

-

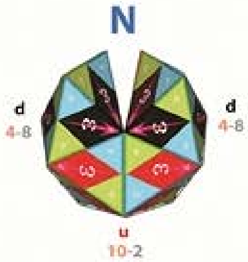

All energy seeks equilibrium

Strong E-force
nuclear linkages

External Electric field points
bind via Charge fascia interactions
with Plus and Minus Electric points
combining and sharing energy
throughout the resultant nuclei

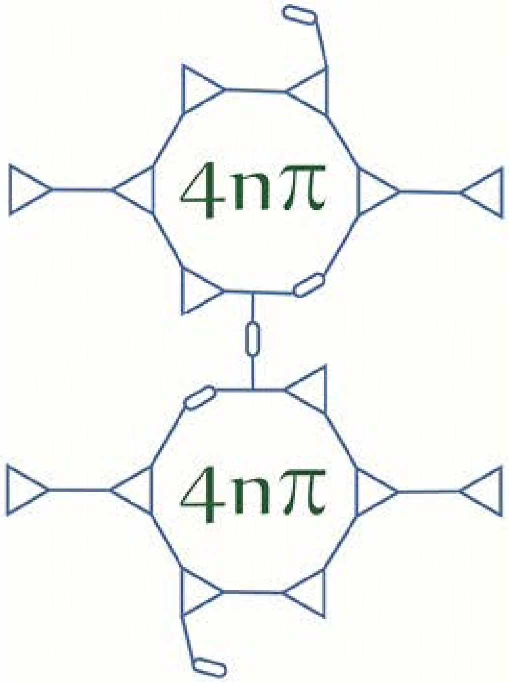

● *Positive Electric field apex*

Negative Electric field apex ●

Copyright ABRAHAM [2008] - All rights reserved

Hydrogenic vs Nucleonic electron binding

If a unbound Proton attracts an Electron
the Electron can be bound to the nuclei
in a number of differing orientations
[each with differing spin energies]

All atomic nuclei (and elements) are Deuteron'ic nuclei with a mixture of *orthagonal*, parallel and anti-paralllel spin orientations

(this is why Rydberg is less accurate for elemental nuclei compared to Hydrogenic atoms - see QM spin)

Electrons can be bound to deuteron nuclei
in four distinct orientations [2 horizontal & 2 vertical]
with each spin coupling orientation producing
differing energy electron orbitals
[wrt to the nuclear magnetic moments]

Ejecting electrons from atomic nuclei
by adding energies to their KEM fields
[the Photoelectric effect]
creates Positive Ions

Vertically orientated electrons
within Proton-Neutron Nuclei [Deuterons]
create quantum synchronous convertor geometries

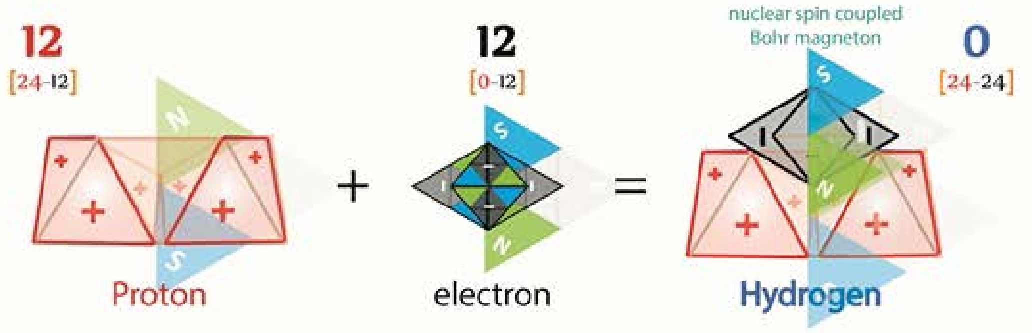
12
[24-12]
Proton
+
12
[0-12]
electron
=
nuclear spin coupled
Bohr magneton
0
[24-24]
Hydrogen

The energy levels of Baryons determines the KEM field energy of bound electrons

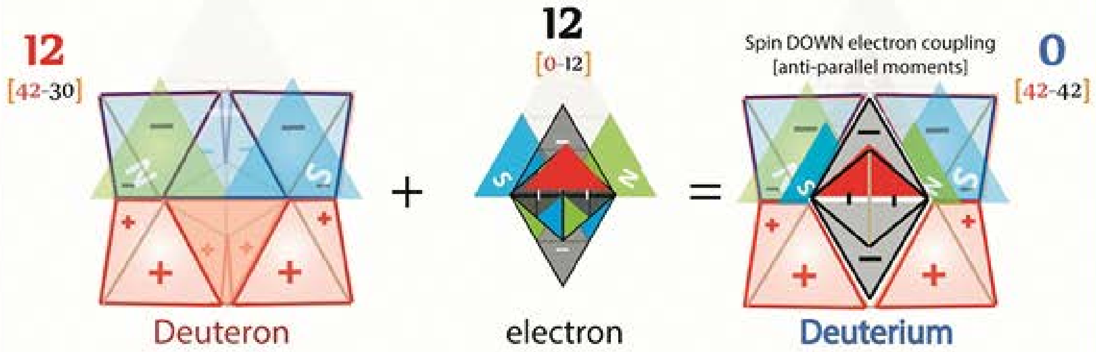
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin DOWN electron coupling
[anti-parallel moments]
0
[42-42]
Deuterium

Bohr magnetons are always referenced wrt the Nuclear magneton

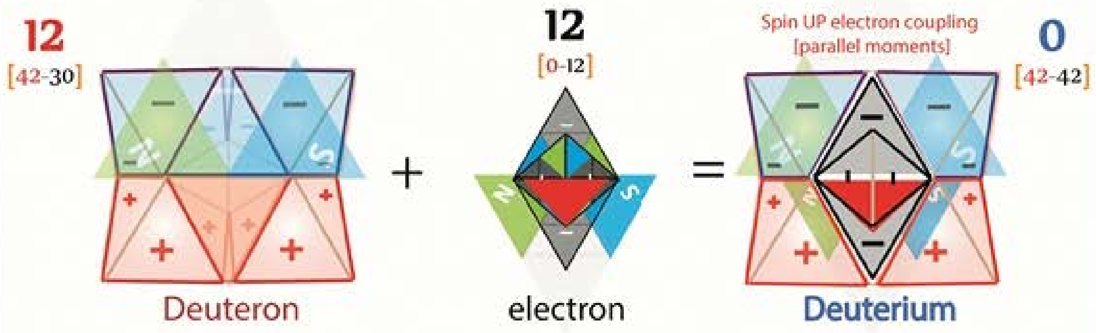
12
[42-30]
Deuteron
+
12
[0-12]
electron
=
Spin UP electron coupling
[parallel moments]
0
[42-42]
Deuterium

electrons produce stronger magnetic moments due to their mass-charge quotient

Atomic Nucleus Master Template

Nucleon Number Z

1 Proton	[24-12]
1 electron	[0-12]
1 Neutron	[18-18]

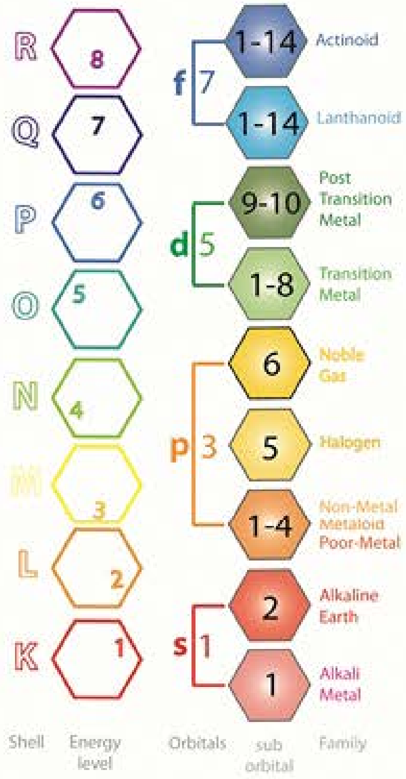
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

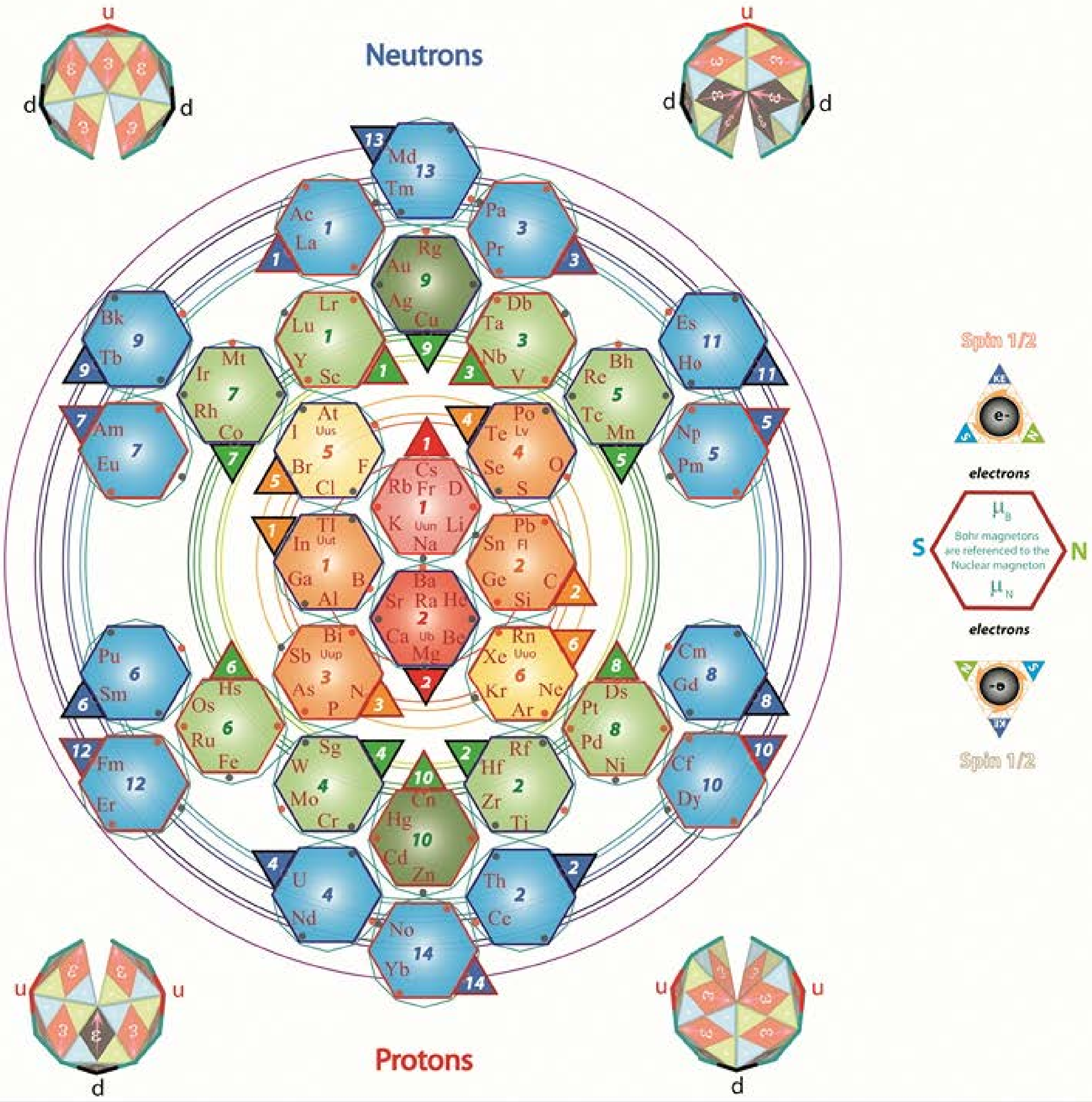
u
d
d
u
d
d
Neutrons
Md 13 Tm
Ac 1 La
Pa 3 Pr
Rg Au 9 Ag Cu
Bk 9 Tb
Lr Lu 1 Y Sc
Db Ta 3 Nb V
Es 11 Ho
Mt Ir 7 Rh Co
Bh Re 5 Tc Mn
Am 7 Eu
At I Uus 5 Br F Cl
Cs Rb Fr D 1 K Uun Li Na
Po Te Lv 4 Se O S
Np 5 Pm
Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si
Ba Sr Ra He 2 Ca Ub Be Mg
Pu 6 Sm
Hs Os 6 Ru Fe
Bi Sb Uup 3 As N P
Rn Xe Uuo 6 Kr Ne Ar
Ds Pt 8 Pd Ni
Cm 8 Gd
Fm 12 Er
Sg W 4 Mo Cr
Cn Hg 10 Cd Zn
Rf Hf 2 Zr Ti
Cf 10 Dy
U 4 Nd
Th 2 Ce
No 14 Yb
Protons
u
u
d
u
u
d
Spin 1/2
e-
electrons
μB
Bohr magnetons are referenced to the Nuclear magneton
μN
S
N
electrons
-e
Spin 1/2

○ **s** *1 sub-Orbital (2 electrons max)*

○ **p** *3 sub-Orbitals (6 electrons max)*

○ **d** *5 sub-Orbitals (10 electrons max)*

○ **f** *7 sub-Orbitals (14 electrons max)*

The Chemical elements

Elemental family groupings are a reflection of atomic sub-orbitals within elemental atoms

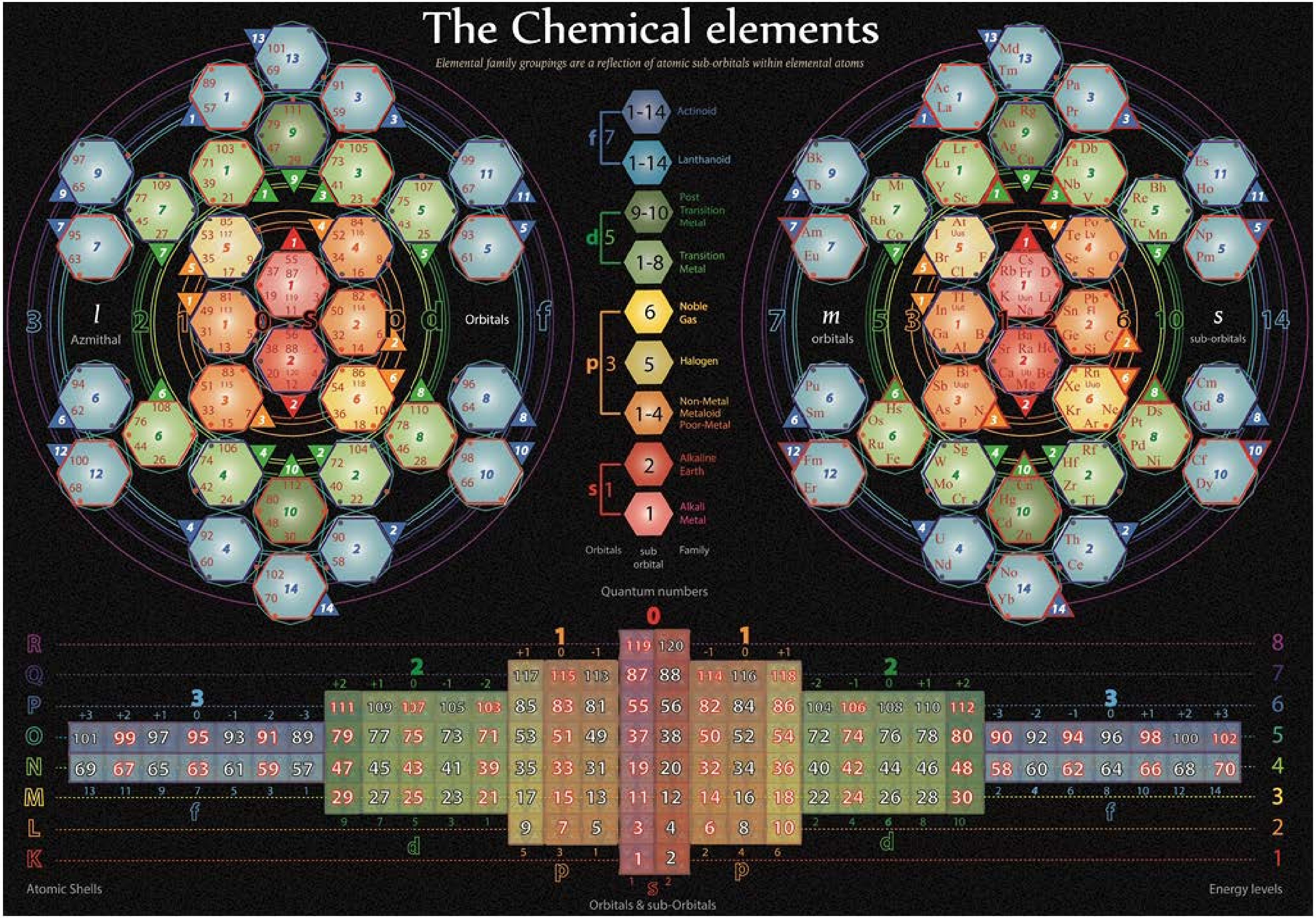
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Orbitals
sub orbital
Family
Quantum numbers
l
Azmithal
Orbitals
m
orbitals
s
sub-orbitals
R Q P O N M L K
Atomic Shells
8 7 6 5 4 3 2 1
Energy levels
f d p s p d f
Orbitals & sub-Orbitals
119 120
117 115 113 87 88 114 116 118
111 109 107 105 103 85 83 81 55 56 82 84 86 104 106 108 110 112
101 99 97 95 93 91 89 79 77 75 73 71 53 51 49 37 38 50 52 54 72 74 76 78 80 90 92 94 96 98 100 102
69 67 65 63 61 59 57 47 45 43 41 39 35 33 31 19 20 32 34 36 40 42 44 46 48 58 60 62 64 66 68 70
29 27 25 23 21 17 15 13 11 12 14 16 18 22 24 26 28 30
9 7 5 3 4 6 8 10
1 2

Tetryonics 43.02 - The Chemical Elements

Element Numbers

Z *

1 Proton	[24-12]
1 Neutron	[18-18]
1 electron	[0-12]

Each elemental nuclei is made from Deuterium

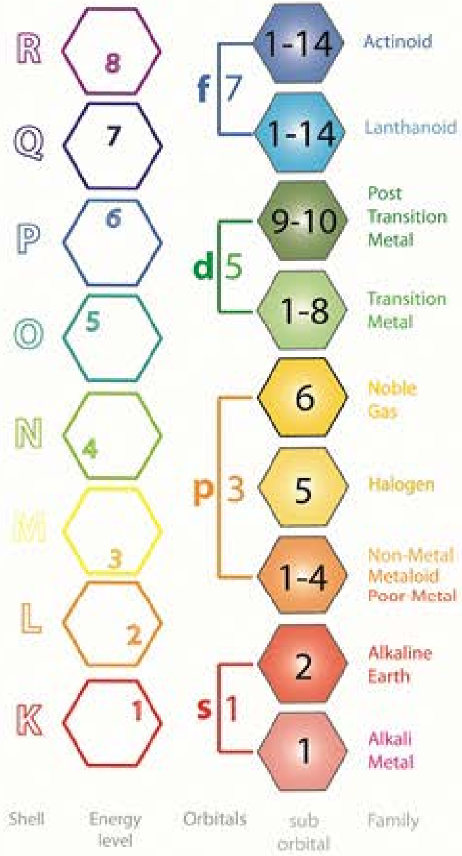

Shell	Energy level
R	8
Q	7
P	6
O	5
N	4
M	3
L	2
K	1

Orbitals	sub orbital	Family
f 7	1-14	Actinoid
	1-14	Lanthanoid
d 5	9-10	Post Transition Metal
	1-8	Transition Metal
p 3	6	Noble Gas
	5	Halogen
	1-4	Non-Metal Metaloid Poor Metal
s 1	2	Alkaline Earth
	1	Alkali Metal

○	s	1 sub-Orbital	(2 electrons max)
○	p	3 sub-Orbitals	(6 electrons max)
○	d	5 sub-Orbitals	(10 electrons max)
○	f	7 sub-Orbitals	(14 electrons max)

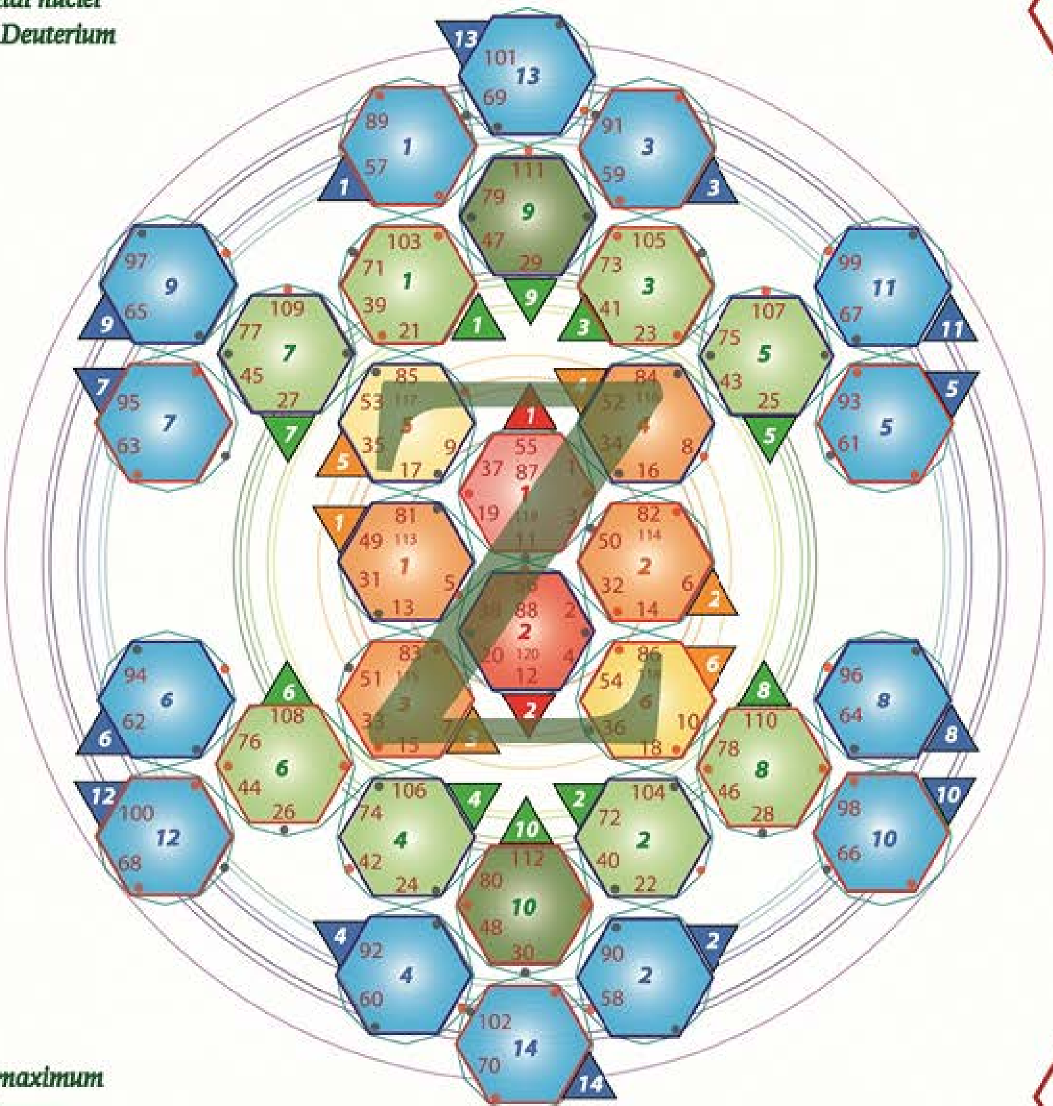
101 13 69
89 1 57
91 3 59
111 79 9 47 29
103 71 1 39 21
105 73 3 41 23
97 9 65
99 11 67
109 77 7 45 27
107 75 5 43 25
85 53 117 5 35 9 17
84 52 116 4 34 8 16
95 7 63
93 5 61
55 37 87 1 19 119 3 11
81 49 113 1 31 5 13
82 50 114 2 32 6 14
56 38 88 2 20 120 4 12
83 51 115 3 33 7 15
86 54 118 6 36 10 18
94 6 62
96 8 64
108 76 6 44 26
110 78 8 46 28
100 12 68
98 10 66
106 74 4 42 24
104 72 2 40 22
112 80 10 48 30
92 4 60
90 2 58
102 14 70
Z

Atomic shells

There are a maximum of 120 elements possible

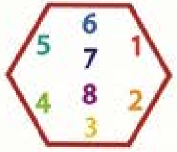

Energy levels

Element Names

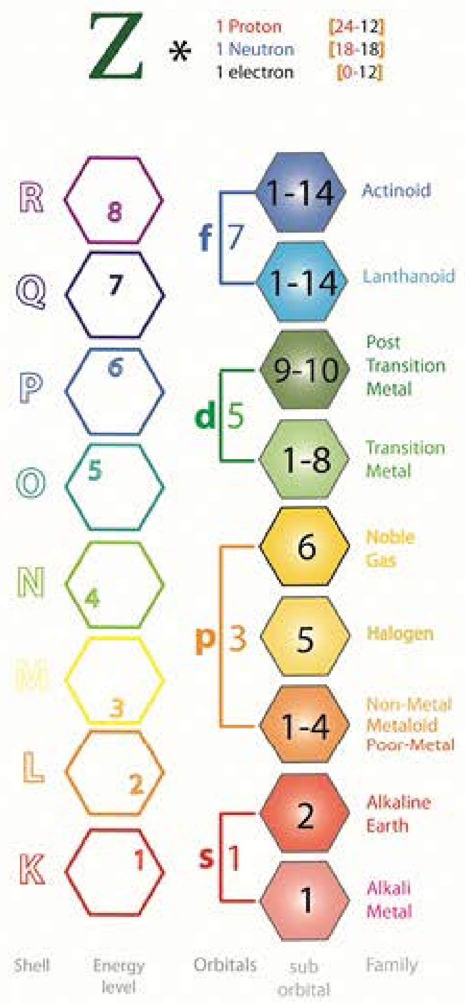
Z *
1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

o s *1 sub-Orbital (2 electrons max)*

○ p *3 sub-Orbitals (6 electrons max)*

○ d *5 sub-Orbitals (10 electrons max)*

○ f *7 sub-Orbitals (14 electrons max)*

There are a maximum of 120 elements possible

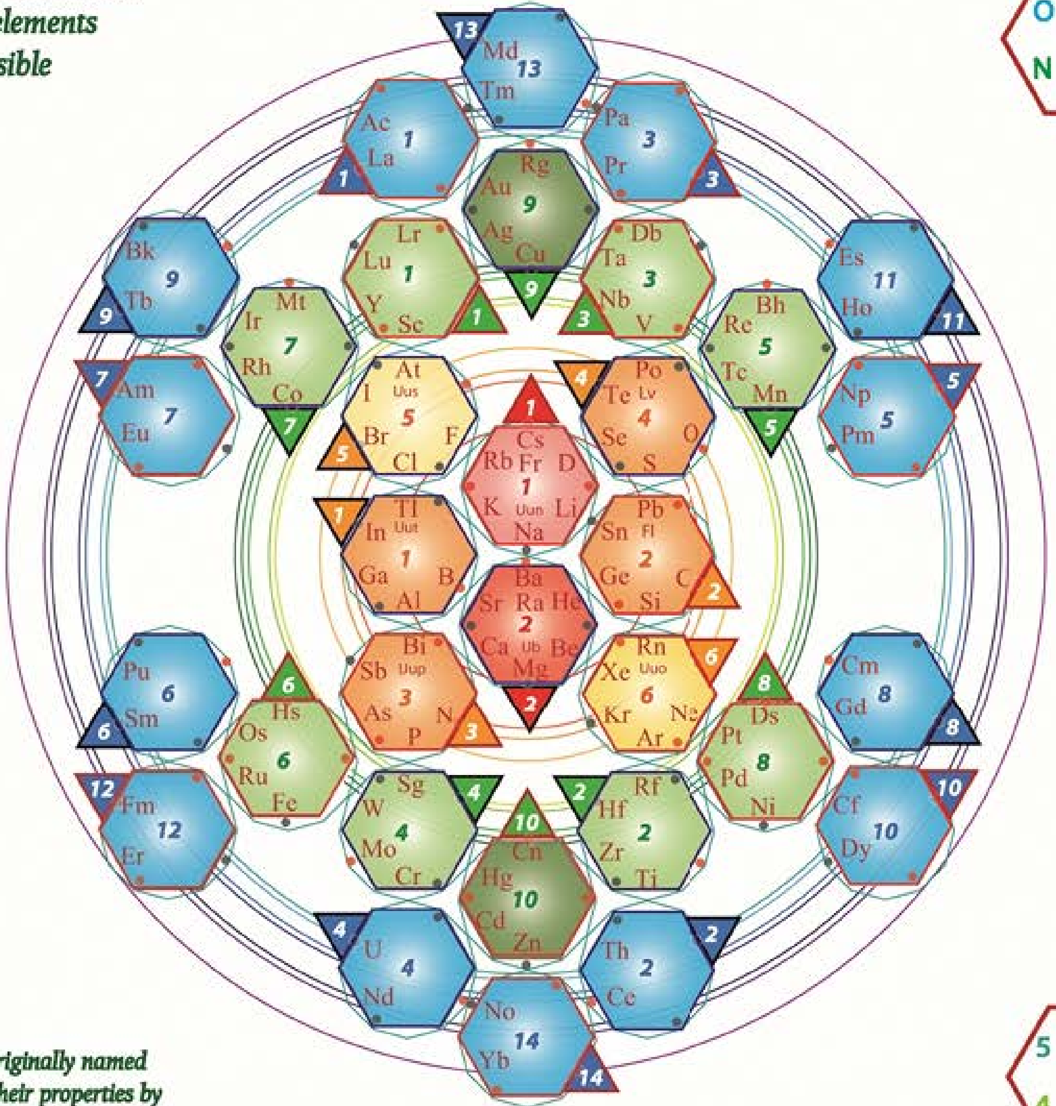
13 Md 13 Tm
Ac 1 La 1
Pa 3 Pr 3
Rg Au 9 Ag Cu 9
Bk 9 Tb 9
Es 11 Ho 11
Lr Lu 1 Y Sc 1
Db Ta 3 Nb V 3
Mt Ir 7 Rh Co 7
Bh Re 5 Tc Mn 5
7 Am 7 Eu
Np 5 Pm 5
At I Uus 5 Br Cl F 5
4 Po Te Lv 4 Se O S
1 Cs Rb Fr D 1 K Uun Li Na
1 Tl In Uut 1 Ga B Al
Pb Sn Fl 2 Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg 2
Bi Sb Uup 3 As N P 3
Rn Xe Uuo 6 Kr Ne Ar 6
Pu 6 Sm 6
Cm Gd 8 8
6 Hs Os 6 Ru Fe
8 Ds Pt Pd 8 Ni
12 Fm 12 Er
Cf 10 Dy 10
Sg W 4 Mo Cr 4
2 Rf Hf 2 Zr Ti
10 Cn Hg 10 Cd Zn
4 U 4 Nd
Th 2 Ce 2
No 14 Yb 14

Atomic shells
P
O
Q
K
N
R
L
M

They were originally named according to their properties by their discoverer but have recently been named after famous scientists

6
5
7
1
4
8
2
3
Energy levels

Copyright ABRAHAM [2008] - All rights reserved

Electron orbital configurations

A huge number of differing d and f orbital configurations are possible given the number of nuclei and bond points created by elemental topologies

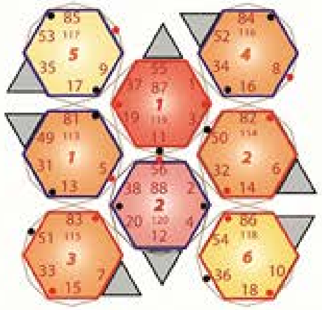
85
53 117
5
35 9
17
55
37 87
1
19 119
11
84
52 116
4
34 8
16
81
49 113
1
31 5
13
82
50 114
2
32 6
14
56
38 88 2
2
20 120 4
12
83
51 115
3
33 7
15
86
54 118
6
36 10
18

As the number of nucleons increases so does the complexity of the electron orbitals possible

quantum snowflakes

However all have a stable 'core' grouping of nuclei comprised of s and p electron orbitals

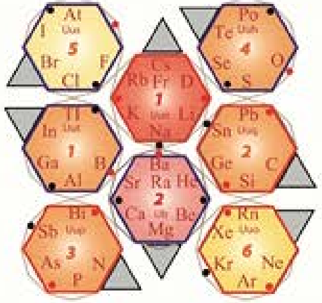
At
I Uus
5
Br F
Cl
Cs
Rb Fr D
1
K Uue Li
Na
Po
Te Uuh
4
Se O
S
Tl
In Uut
1
Ga B
Al
Pb
Sn Uuq
2
Ge C
Si
Ba
Sr Ra He
2
Ca Ubn Be
Mg
Bi
Sb Uup
3
As N
P
Rn
Xe Uuo
6
Kr Ne
Ar

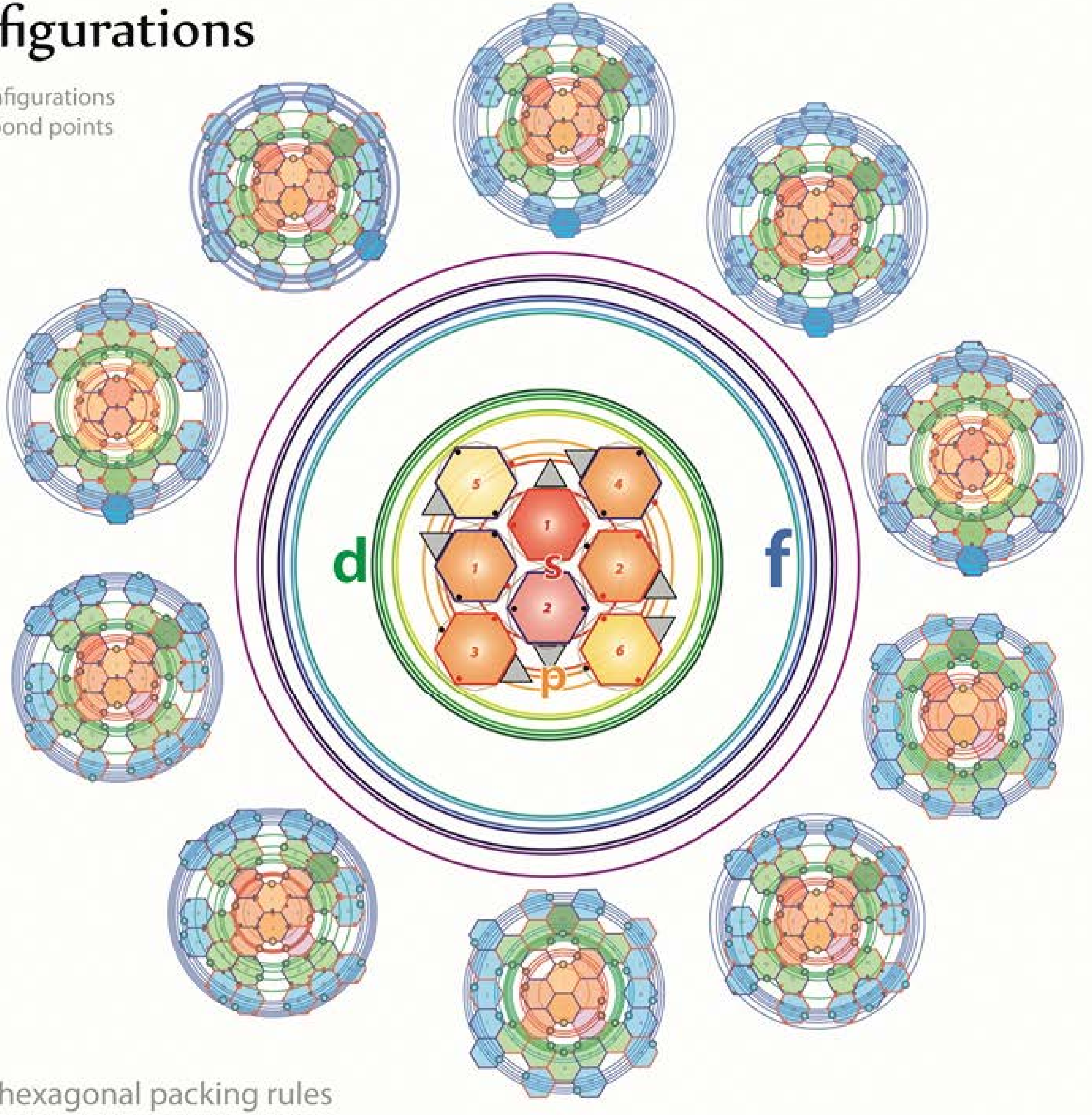
d
f
s
p
5
4
1
1
2
2
3
6

All nuclei bonding closely follows hexagonal packing rules

Copyright ABRAHAM [2008] - All rights reserved

Atomic Nucleus
Exploded view

1 Proton [24-12]
1 electron [0-12]
1 Neutron [18-18]

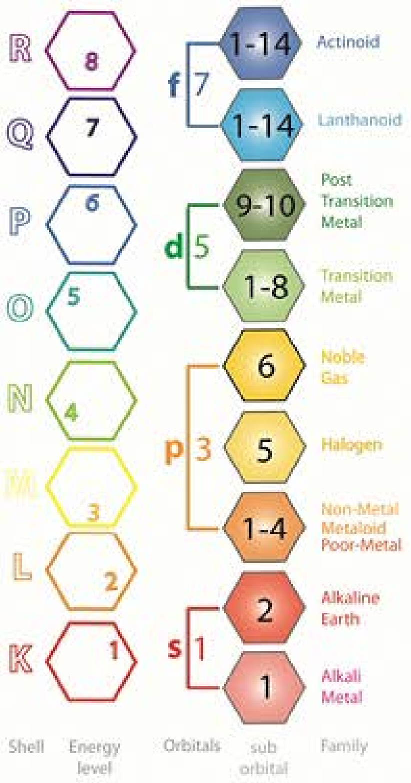
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shell
Energy level
Orbitals
sub orbital
Family

○ **s** *1 sub-Orbital (2 electrons max)*
○ **p** *3 sub-Orbitals (6 electrons max)*
○ **d** *5 sub-Orbitals (10 electrons max)*
○ **f** *7 sub-Orbitals (14 electrons max)*

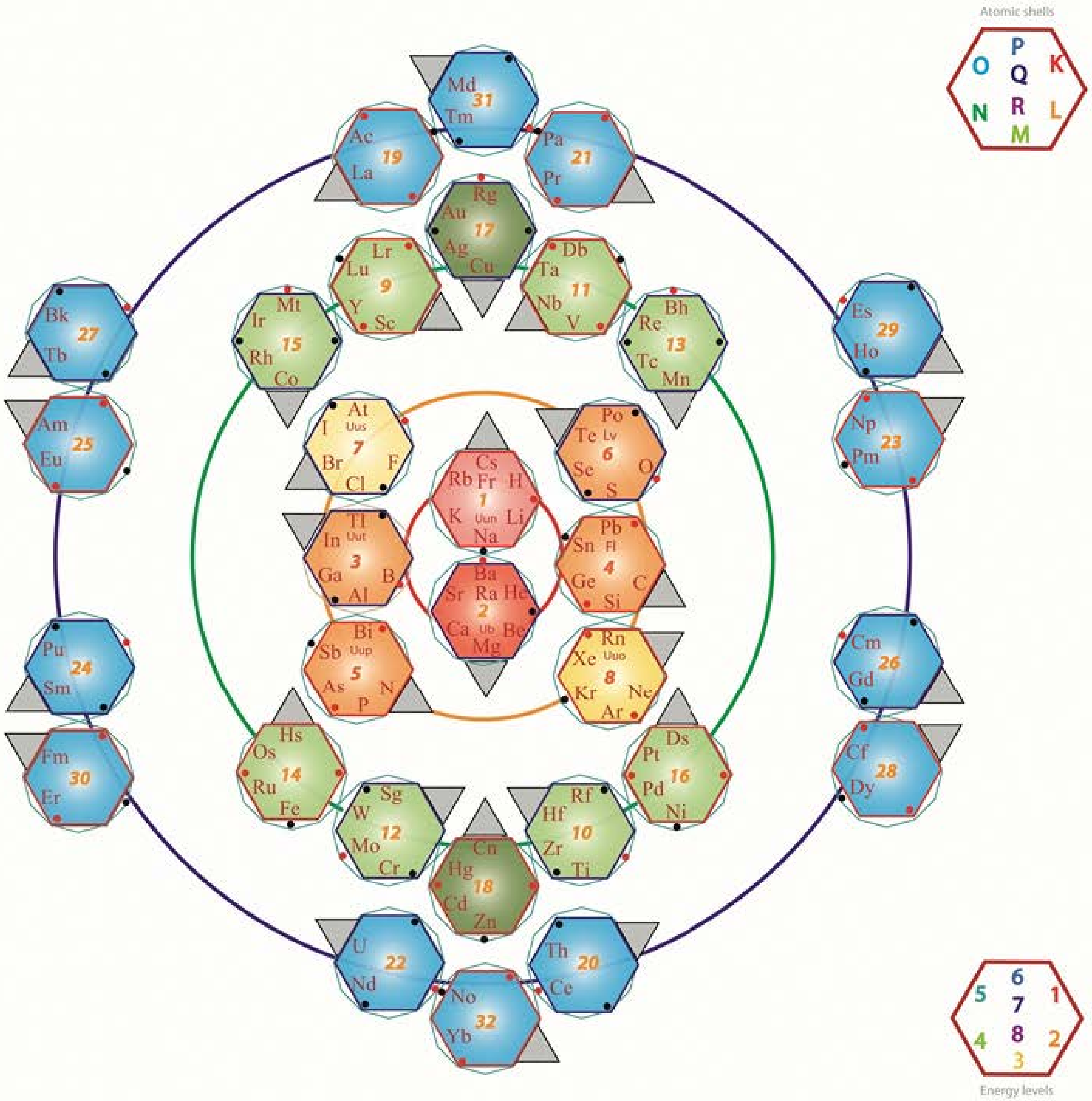
Atomic shells
P Q K O R L N M
Md 31 Tm
Ac 19 La
Pa 21 Pr
Rg Au 17 Ag Cu
Lr Lu 9 Y Sc
Db Ta 11 Nb V
Mt Ir 15 Rh Co
Bh Re 13 Tc Mn
Bk 27 Tb
Es 29 Ho
Am 25 Eu
Np 23 Pm
At I Uus 7 Br F Cl
Po Te Lv 6 Se O S
Cs Rb Fr H 1 K Uun Li Na
Tl In Uut 3 Ga B Al
Pb Sn Fl 4 Ge C Si
Ba Sr Ra He 2 Ca Ub Be Mg
Bi Sb Uup 5 As N P
Rn Xe Uuo 8 Kr Ne Ar
Pu 24 Sm
Cm 26 Gd
Fm 30 Er
Cf 28 Dy
Hs Os 14 Ru Fe
Ds Pt 16 Pd Ni
Sg W 12 Mo Cr
Rf Hf 10 Zr Ti
Cn Hg 18 Cd Zn
U 22 Nd
Th 20 Ce
No 32 Yb
6 7 5 1 8 4 2 3
Energy levels

Copyright ABRAHAM [2008] - All rights reserved

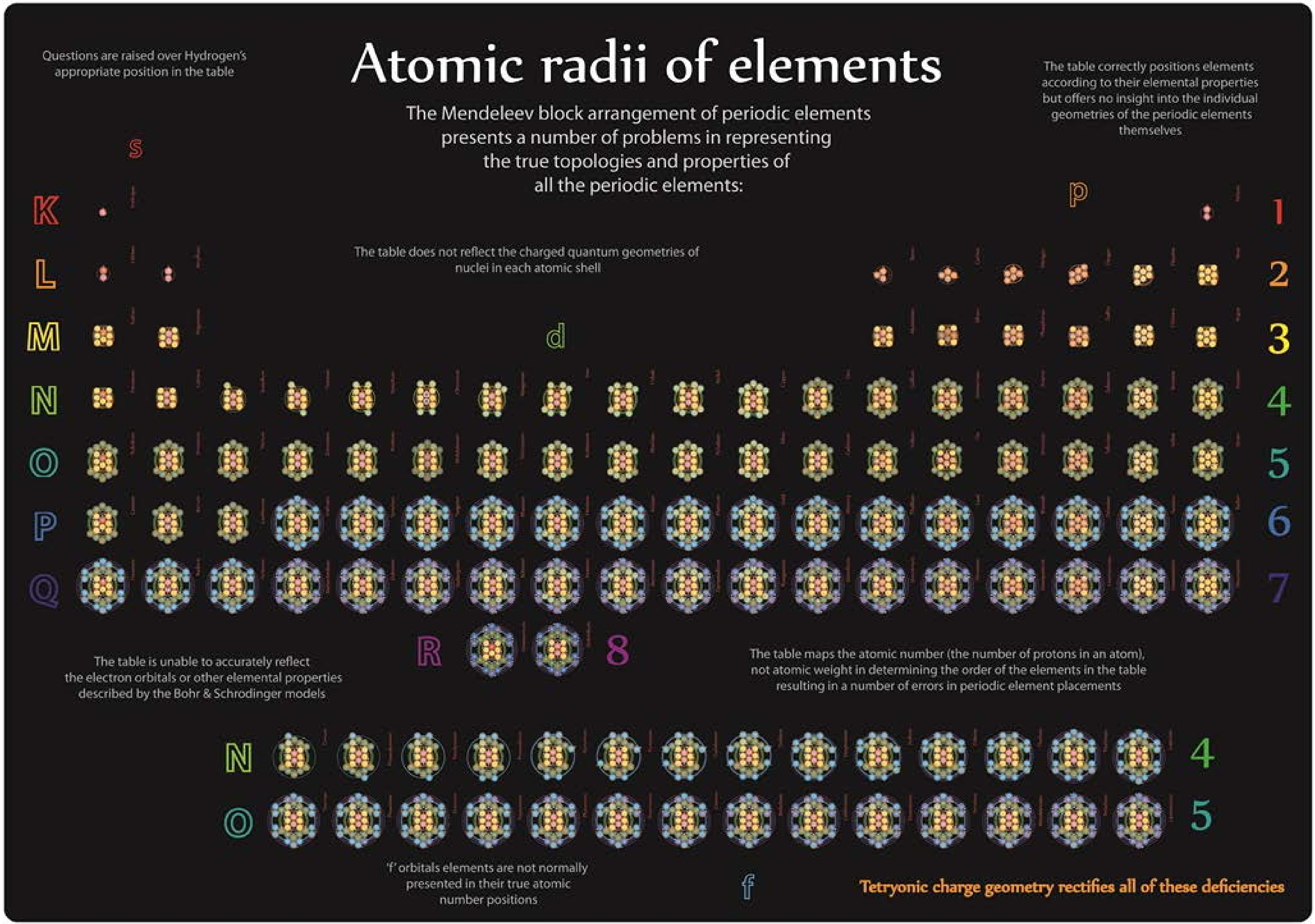

Atomic radii of elements

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

Questions are raised over Hydrogen's appropriate position in the table

The table correctly positions elements according to their elemental properties but offers no insight into the individual geometries of the periodic elements themselves

The table does not reflect the charged quantum geometries of nuclei in each atomic shell

The table is unable to accurately reflect the electron orbitals or other elemental properties described by the Bohr & Schrodinger models

The table maps the atomic number (the number of protons in an atom), not atomic weight in determining the order of the elements in the table resulting in a number of errors in periodic element placements

'f' orbitals elements are not normally presented in their true atomic number positions

Tetryonic charge geometry rectifies all of these deficiencies

Tetryonics 43.07 - Atomic radii of elements

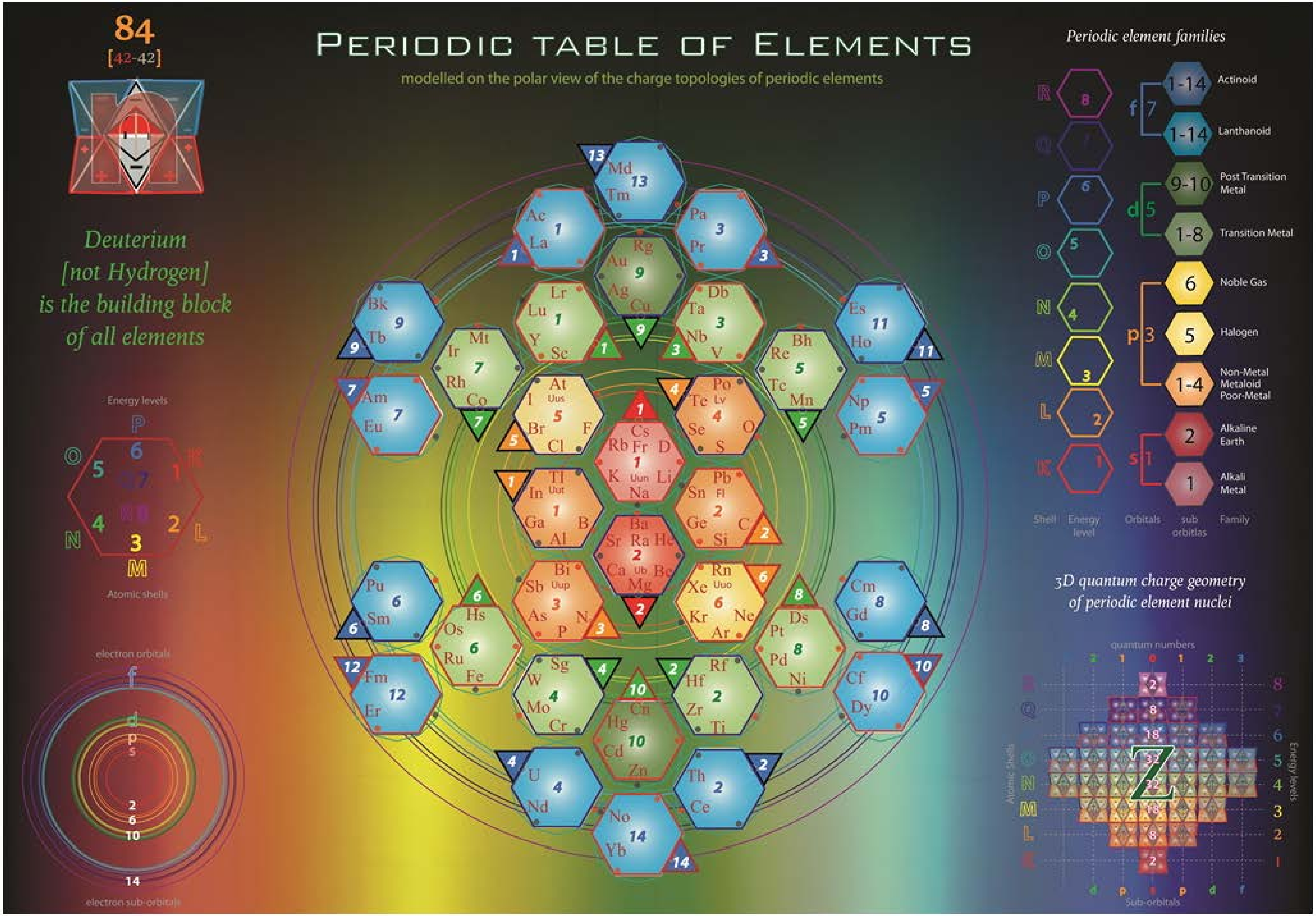

PERIODIC TABLE OF ELEMENTS

modelled on the polar view of the charge topologies of periodic elements

Deuterium
[not Hydrogen]
is the building block
of all elements

Periodic element families

3D quantum charge geometry
of periodic element nuclei

84
[42-42]
Energy levels
P 6
O 5
Q 7
R 8
K 1
L 2
N 4
M 3
Atomic shells
electron orbitals
f
d
p
s
2
6
10
14
electron sub-orbitals
Md Tm 13
Ac La 1
Pa Pr 3
Rg Au Ag Cu 9
Lr Lu Y Sc 1
Db Ta Nb V 3
Bk Tb 9
Mt Ir Rh Co 7
Bh Re Tc Mn 5
Es Ho 11
Am Eu 7
At I Uus Br Cl F 5
Cs Rb Fr D 1 K Uun Li Na
Po Te Lv Se O S 4
Np Pm 5
Tl In Uut Ga B Al 1
Pb Sn Fl Ge C Si 2
Ba Sr Ra He 2 Ca Ub Be Mg
Pu Sm 6
Hs Os Ru Fe 6
Bi Sb Uup As N P 3
Rn Xe Uuo Kr Ne Ar 6
Ds Pt Pd Ni 8
Cm Gd 8
Fm Er 12
Sg W Mo Cr 4
Cn Hg Cd Zn 10
Rf Hf Zr Ti 2
Cf Dy 10
U Nd 4
Th Ce 2
No Yb 14
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
Shell
Energy level
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Orbitals
sub orbitlas
Family
quantum numbers
3 2 1 0 1 2 3
R Q P O N M L K
Atomic Shells
2 8 18 32 32 18 8 2
Z
8 7 6 5 4 3 2 1
Energy levels
f d p s p d f
Sub-orbitals

Tetryonics 43.08 - Periodic Table 2.0

Copyright ABRAHAM [2008] - All rights reserved

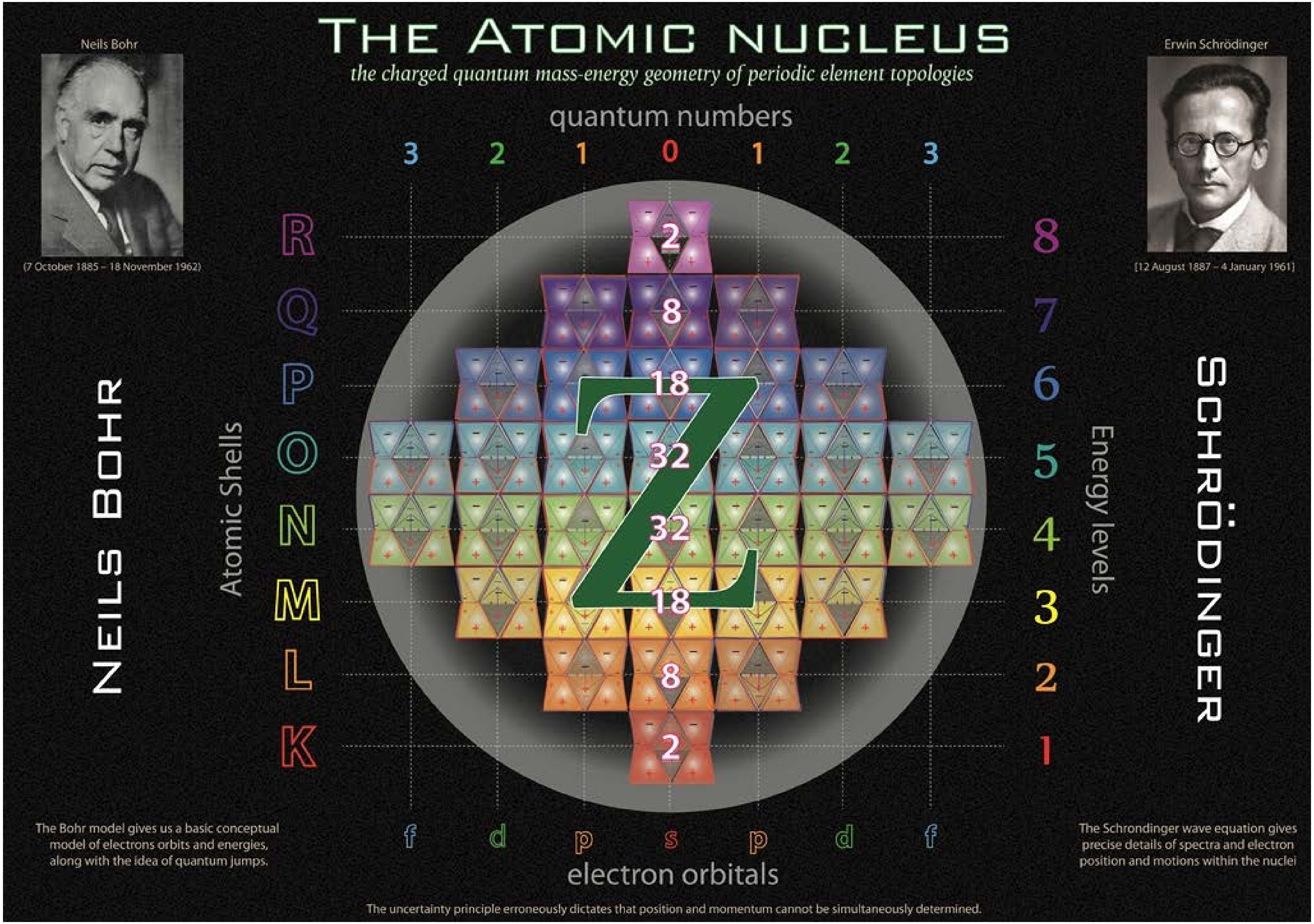

THE ATOMIC NUCLEUS

the charged quantum mass-energy geometry of periodic element topologies

Neils Bohr
(7 October 1885 – 18 November 1962)

Erwin Schrödinger
[12 August 1887 – 4 January 1961]

NEILS BOHR
Atomic Shells
R
Q
P
O
N
M
L
K
quantum numbers
3
2
1
0
1
2
3
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
Energy levels
SCHRÖDINGER
f
d
p
s
p
d
f
electron orbitals

The Bohr model gives us a basic conceptual model of electrons orbits and energies, along with the idea of quantum jumps.

The Schrondinger wave equation gives precise details of spectra and electron position and motions within the nuclei

The uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined.

Tetryonics 44.01 - The Atomic Nucleus

Periodicity of atomic elements

An electron shell may be thought of as an orbit followed by electrons around an atom's nucleus.
The closest shell to the nucleus is called the "1 shell" (also called "K shell"), followed by the
"2 shell" (or "L shell"), then the "3 shell" (or "M shell"), and so on further and further from the nucleus.
The shell letters K, L, M, ... are alphabetical

Each shell can contain only a
fixed number of deuterium nuclei
[Protons, Neutrons & electrons]

Each shell consists of one or more
electron orbitals, and each orbital consists
of one or more sub-orbitals.

8 atomic energy shells with electrons

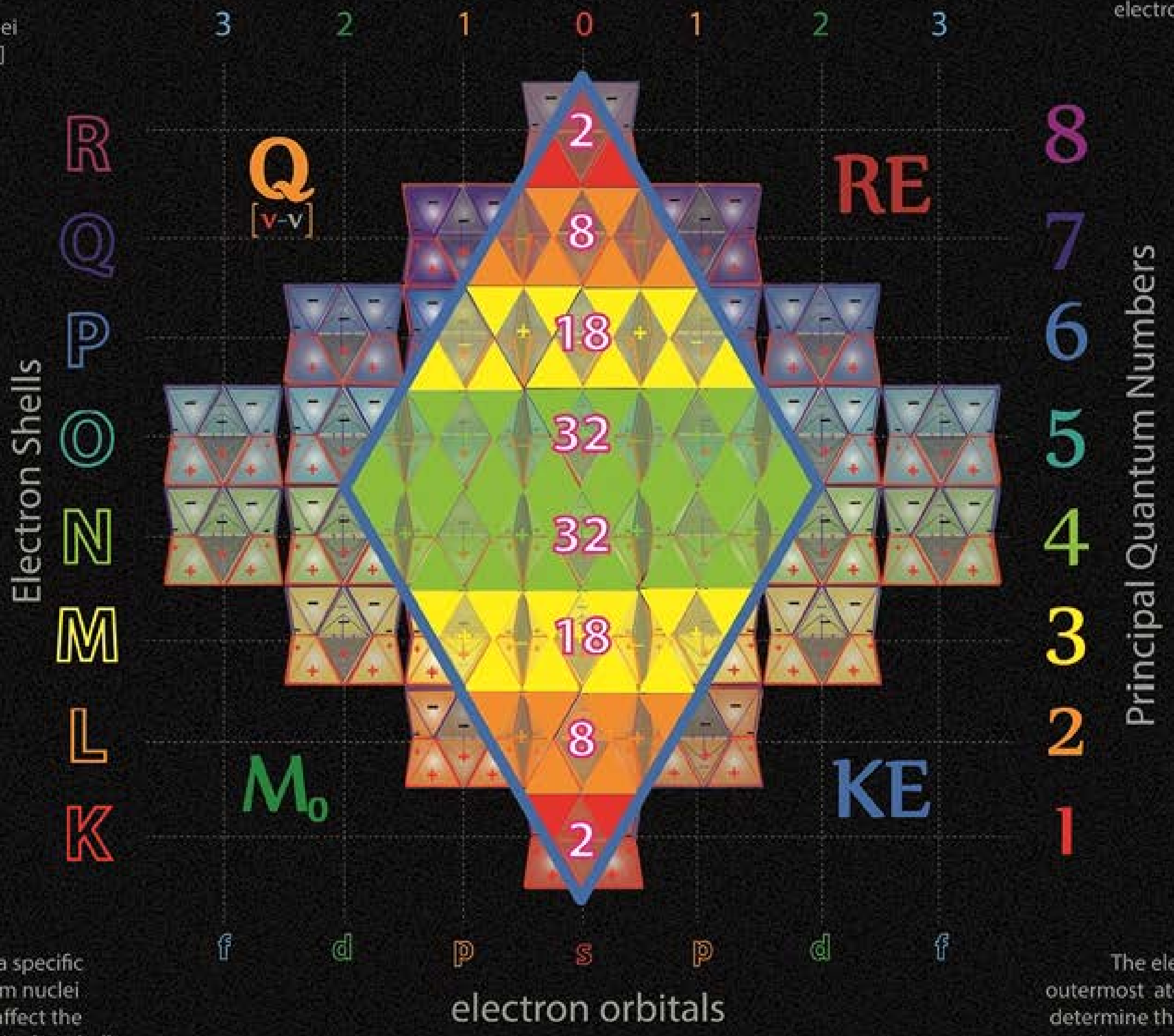

120 periodic elements in total

The number of deuterium nuclei that form each atomic shell in an elemental atom number is double that of Photon energy levels

Each atomic shell equates to a specific
energy level for the Deuterium nuclei
that comprise it with in turn affect the
angular momentum of electrons in that shell

The electrons in the partially filled
outermost atomic shell [or highest energy level]
determine the chemical properties of the atom;
they are called valence shell electrons.

Copyright ABRAHAM [2008] - All rights reserved

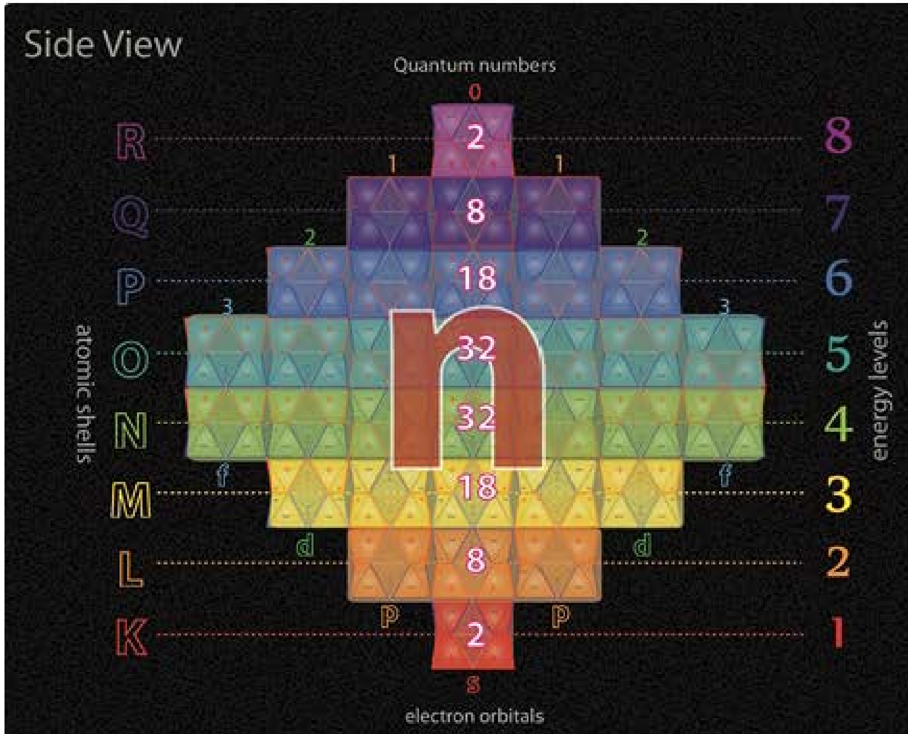
Side View
Quantum numbers
0
1 1
2 2
3 3
R Q P O N M L K
atomic shells
2 8 18 32 32 18 8 2
n
f f
d d
p p
s
electron orbitals
8 7 6 5 4 3 2 1
energy levels

Atomic shells & energy levels

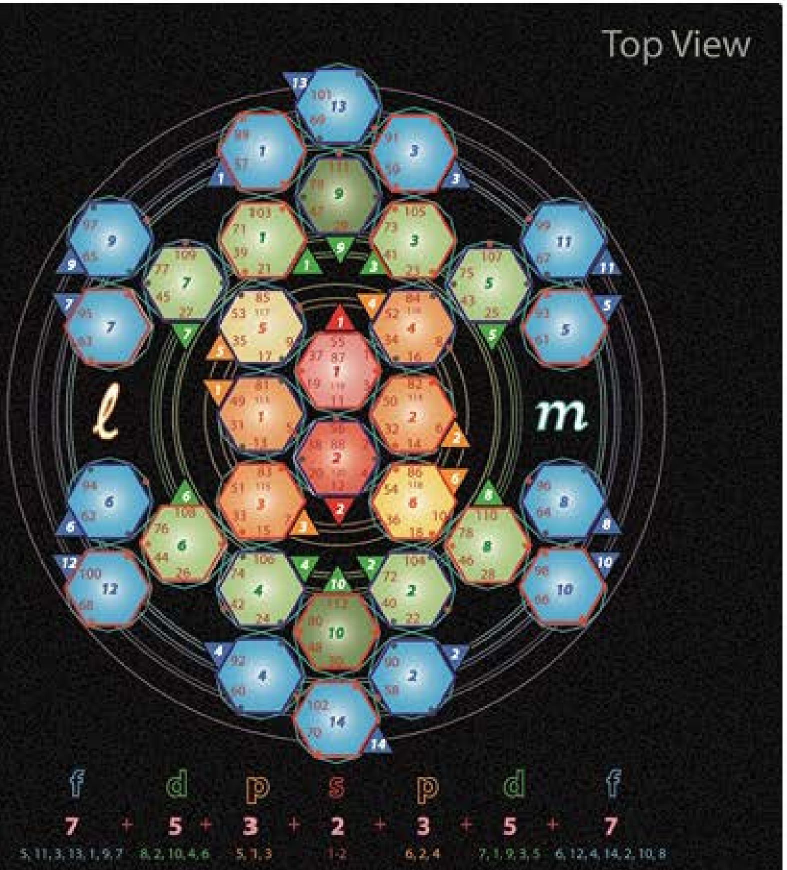
Top View
ℓ
m
f d p s p d f
7 + 5 + 3 + 2 + 3 + 5 + 7
5, 11, 3, 13, 1, 9, 7
8, 2, 10, 4, 6
5, 1, 3
1-2
6, 2, 4
7, 1, 9, 3, 5
6, 12, 4, 14, 2, 10, 8

Using the charge geometries of equilateral
mass-energies in Matter topologies
all period elements can be
accurately modelled

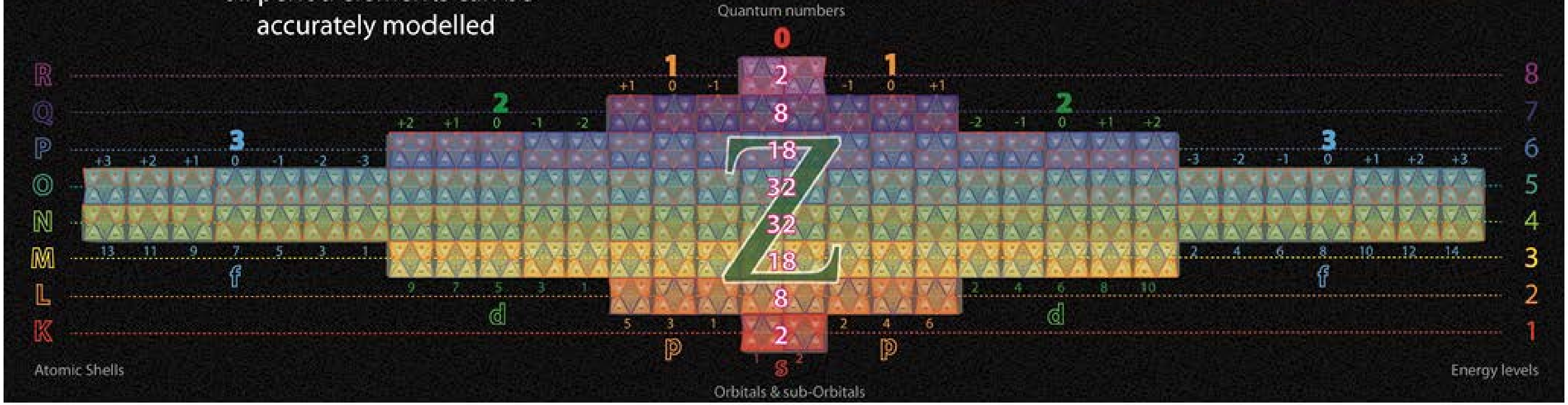
Quantum numbers
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
R Q P O N M L K
2 8 18 32 32 18 8 2
Z
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f f
9 7 5 3 1
2 4 6 8 10
d d
5 3 1
2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals
8 7 6 5 4 3 2 1
Energy levels

atomic Shells

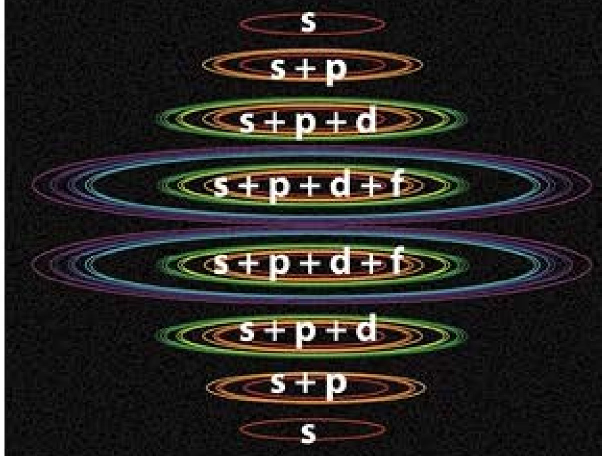
s
s + p
s + p + d
s + p + d + f
s + p + d + f
s + p + d
s + p
s

energy levels

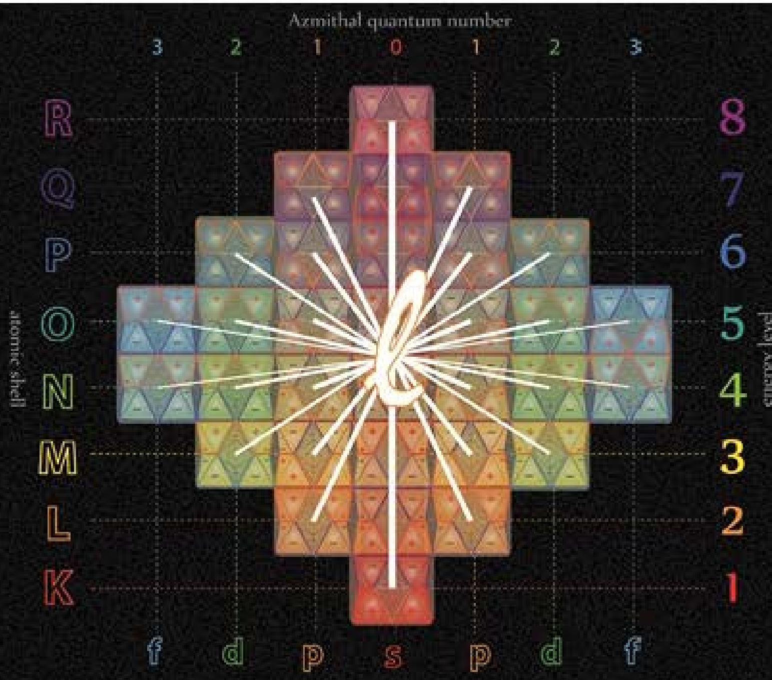
Azmithal quantum number
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
atomic shell
8
7
6
5
4
3
2
1
energy level
ℓ
f
d
p
s
p
d
f

Electron orbitals

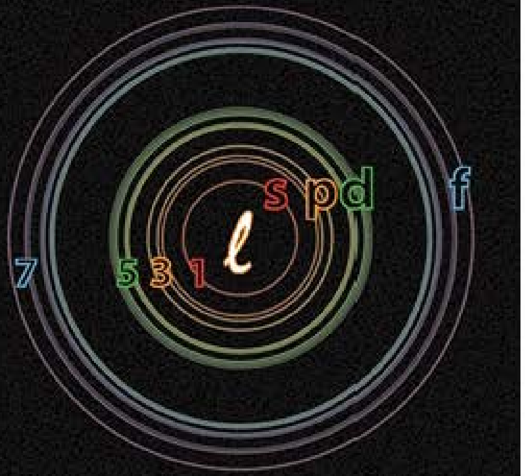

quantum numbers

All atoms have shells comprised of concentric s, p, d & f electron orbitals

electron orbitals

each electron orbital can hold a maximum of two spin opposed electrons

Each electron's location in
an elemental nuclei is defined
by its quantum numbers

a core s-orbital
surrounded by
3 p-orbitals
5 d-orbitals and
7 f-orbitals

Azmithal & Magnetic numbers

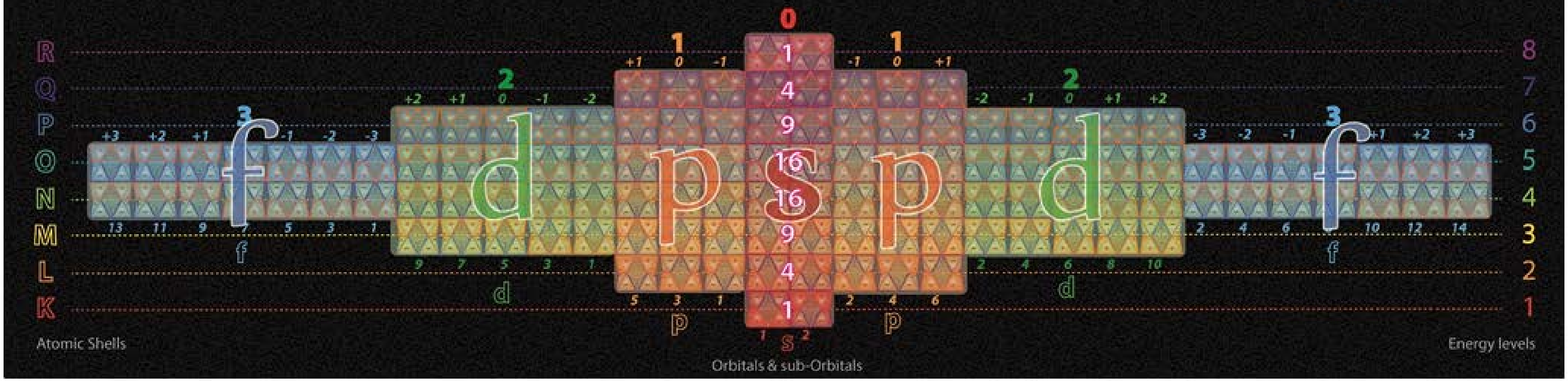
0
1
1
+1
0
-1
-1
0
+1
2
2
+2
+1
0
-1
-2
-2
-1
0
+1
+2
3
3
+3
+2
+1
-1
-2
-3
-3
-2
-1
+1
+2
+3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1
4
9
16
16
9
4
1
f
d
p
s
p
d
f
13
11
9
7
5
3
1
2
4
6
8
10
12
14
f
f
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
Atomic Shells
Energy levels

Orbitals & sub-Orbitals

Copyright ABRAHAM [2008] - All rights reserved

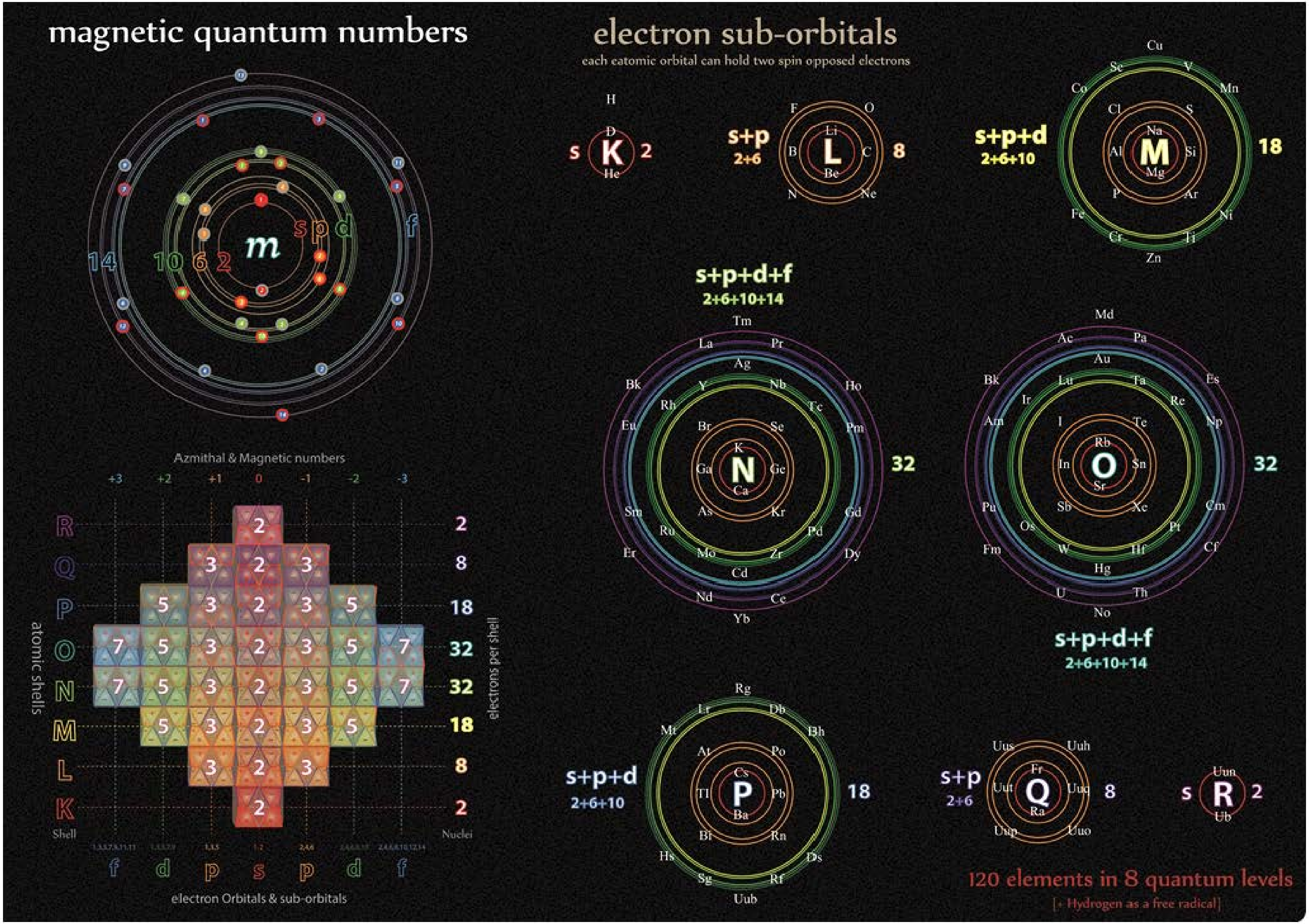
magnetic quantum numbers
m
s p d f
14 10 6 2
Azmithal & Magnetic numbers
+3 +2 +1 0 -1 -2 -3
R Q P O N M L K
atomic shells
Shell
2 8 18 32 32 18 8 2
electrons per shell
Nuclei
f d p s p d f
electron Orbitals & sub-orbitals
electron sub-orbitals
each eatomic orbital can hold two spin opposed electrons
s K 2
H D He
s+p 2+6 L 8
F O Li B C Be N Ne
s+p+d 2+6+10 M 18
Cu Sc V Co Mn Cl S Na Al Si Mg P Ar Fe Ni Cr Ti Zn
s+p+d+f 2+6+10+14 N 32
Tm La Pr Ag Bk Y Nb Ho Rh Tc Eu Br Se Pm K Ga Ge Ca Sm As Kr Gd Ru Pd Er Mo Zr Dy Cd Nd Ce Yb
s+p+d+f 2+6+10+14 O 32
Md Ac Pa Au Bk Lu Ta Es Ir Re Am I Te Np Rb In Sn Sr Pu Sb Xe Cm Os Pt Fm W Hf Cf Hg U Th No
s+p+d 2+6+10 P 18
Rg Lr Db Mt Bh At Po Cs Tl Pb Ba Bi Rn Hs Ds Sg Rf Uub
s+p 2+6 Q 8
Uus Uuh Fr Uut Uuq Ra Uup Uuo
s R 2
Uun Ub
120 elements in 8 quantum levels
[+ Hydrogen as a free radical]

Tetryonics 44.05 - Electron sub-Obitals

Copyright ABRAHAM [2008] - All rights reserved

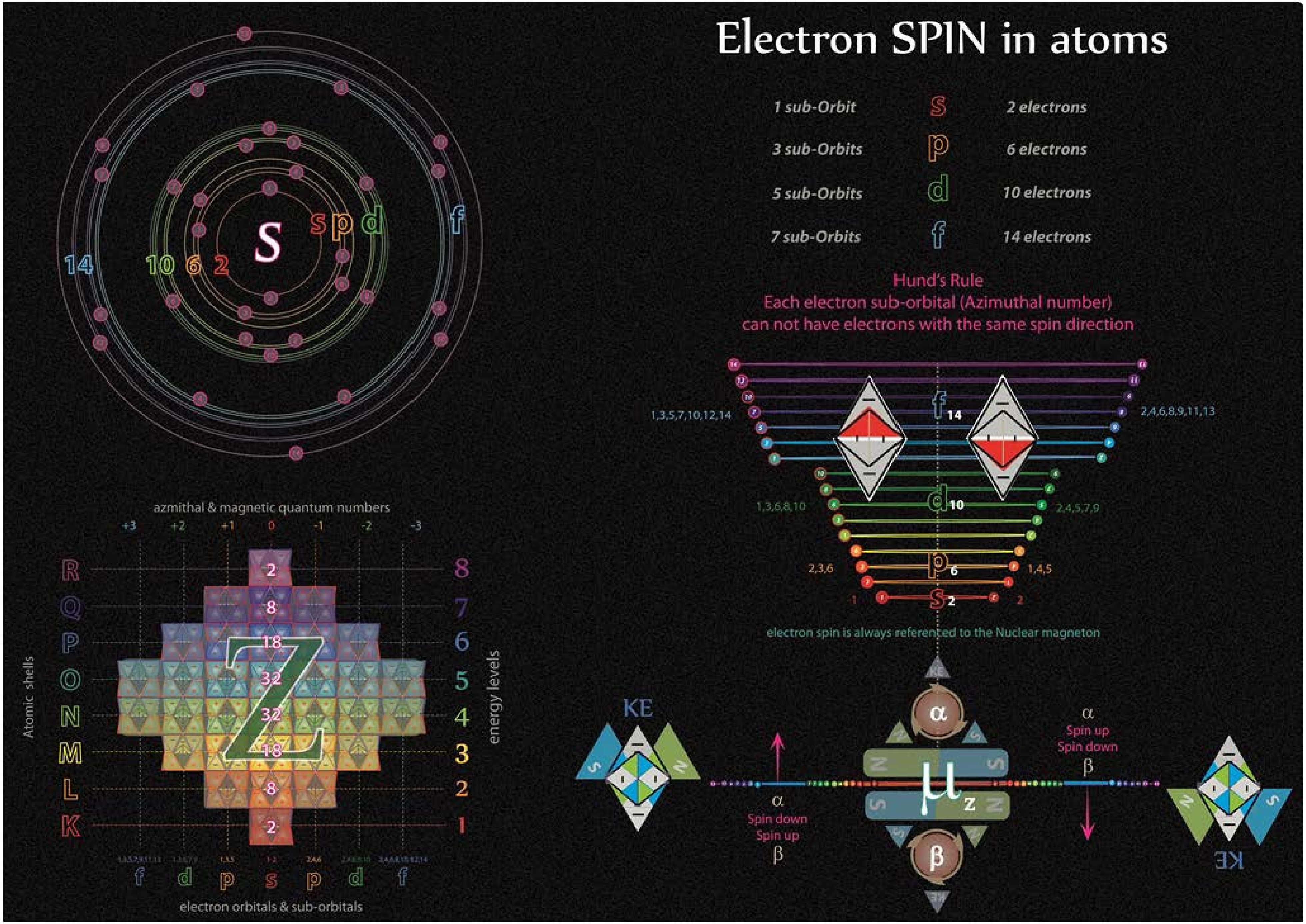

Electron SPIN in atoms

1 sub-Orbit	s	*2 electrons*
3 sub-Orbits	p	*6 electrons*
5 sub-Orbits	d	*10 electrons*
7 sub-Orbits	f	*14 electrons*

Hund's Rule
Each electron sub-orbital (Azimuthal number)
can not have electrons with the same spin direction

electron spin is always referenced to the Nuclear magneton

S
s p d
f
14
10 6 2
azmithal & magnetic quantum numbers
+3
+2
+1
0
-1
-2
-3
R
Q
P
O
N
M
L
K
Atomic shells
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f
electron orbitals & sub-orbitals
1,3,5,7,10,12,14
f 14
2,4,6,8,9,11,13
1,3,6,8,10
d 10
2,4,5,7,9
2,3,6
p 6
1,4,5
1
s 2
2
KE
α
Spin down
Spin up
β
α
β
μ z
N
S
α
Spin up
Spin down
β
KE

Tetryonics 44.06 - Electron spin

Quantum numbers

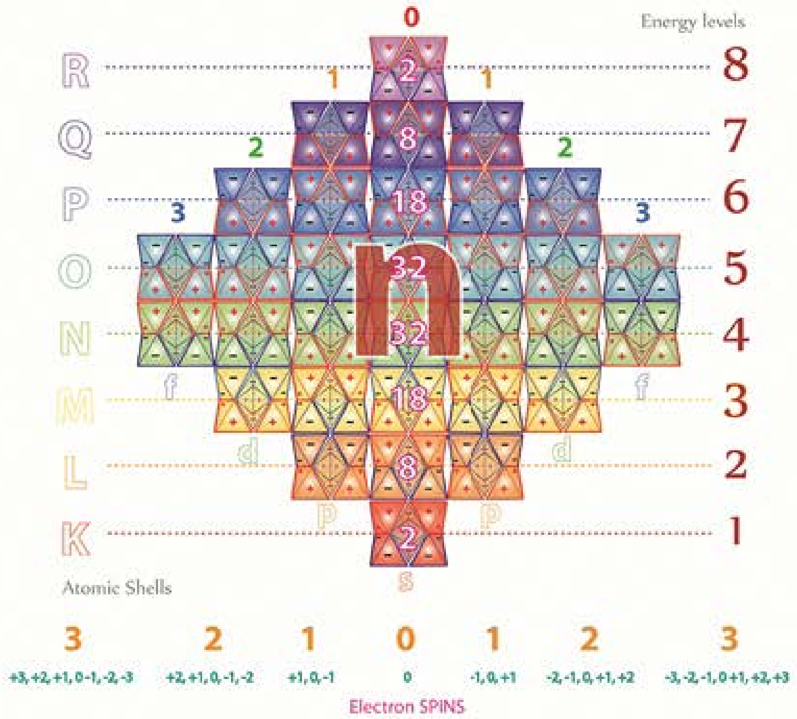
Energy levels
0
R 8
1 1
Q 7
2 2
P 6
3 3
O 5
N 4
M 3
L 2
K 1
2
8
18
32
n
32
18
8
2
f f
d d
p p
s
Atomic Shells
3 2 1 0 1 2 3
+3, +2, +1, 0 -1, -2, -3
+2, +1, 0, -1, -2
+1, 0, -1
0
-1, 0, +1
-2, -1, 0, +1, +2
-3, -2, -1, 0 +1, +2, +3
Electron SPINS

These four numbers, n, ℓ, m and s can be used to describe any electron in a stable atom [and can be mapped back to the classical terminology of Shells and electron orbitals].

The properties of every atom's specific electron configuration can be described by four quantum numbers:

n (1-8)
Principal
(n = 1, 2, 3, 4 ...)

ℓ (0-3)
Azimuthal
(l = 0, 1 ... n−1)

m (2ℓ+1)
Magnetic
(ml = −l, −l+1 ... 0 ... l−1, l)

S ± ½
electron Spin
(ms = −1/2 or +1/2),

The Bohr model was a one-dimensional model that used one quantum number to describe the distribution of electrons

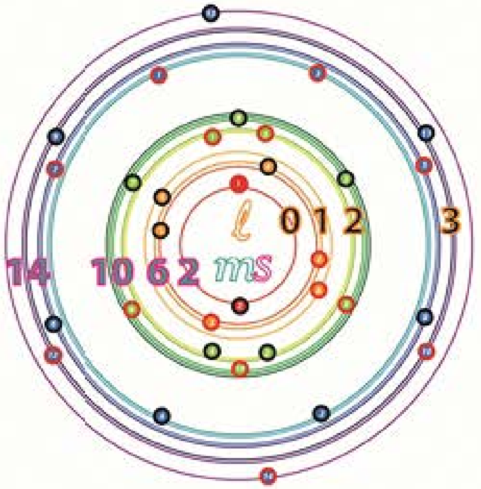
ℓ
0 1 2 3
14 10 6 2 ms

The three coordinates that come from Schrödinger's wave equations are the principal (n), angular (l), and magnetic (m) quantum numbers.

These quantum numbers describe the size, shape, and orientation in space of the orbitals of any particular atom mathematically.

Each electron's quantum numbers are unique and cannot be shared by another electron in that atom.

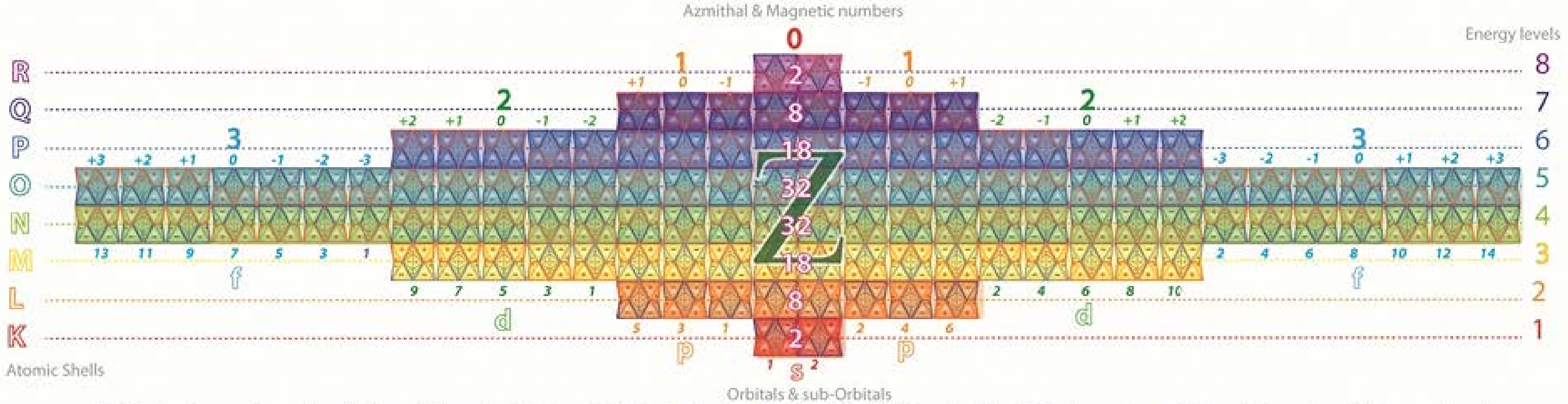
Azmithal & Magnetic numbers
Energy levels
0
1 1
+1 0 -1 -1 0 +1
2 2
+2 +1 0 -1 -2 -2 -1 0 +1 +2
3 3
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f f
9 7 5 3 1
2 4 6 8 10
d d
5 3 1
2 4 6
p p
1 s 2
Atomic Shells
Orbitals & sub-Orbitals

A stable atom has equal numebrs of Protons, electrons [and Neutrons], all following the Pauli Exclusion Principle thus orientating their spins so that each element has a unique electron configuration

Copyright ABRAHAM [2008] - All rights reserved

PRINCIPAL quantum number

The first describes the electron shell, or energy level, of an atom.

This is the only quantum number introduced by the Bohr model

The principal quantum number can only have positive integer values

atomic shells

n(1-8)

(n = 1, 2, 3, 4 ...)

energy levels

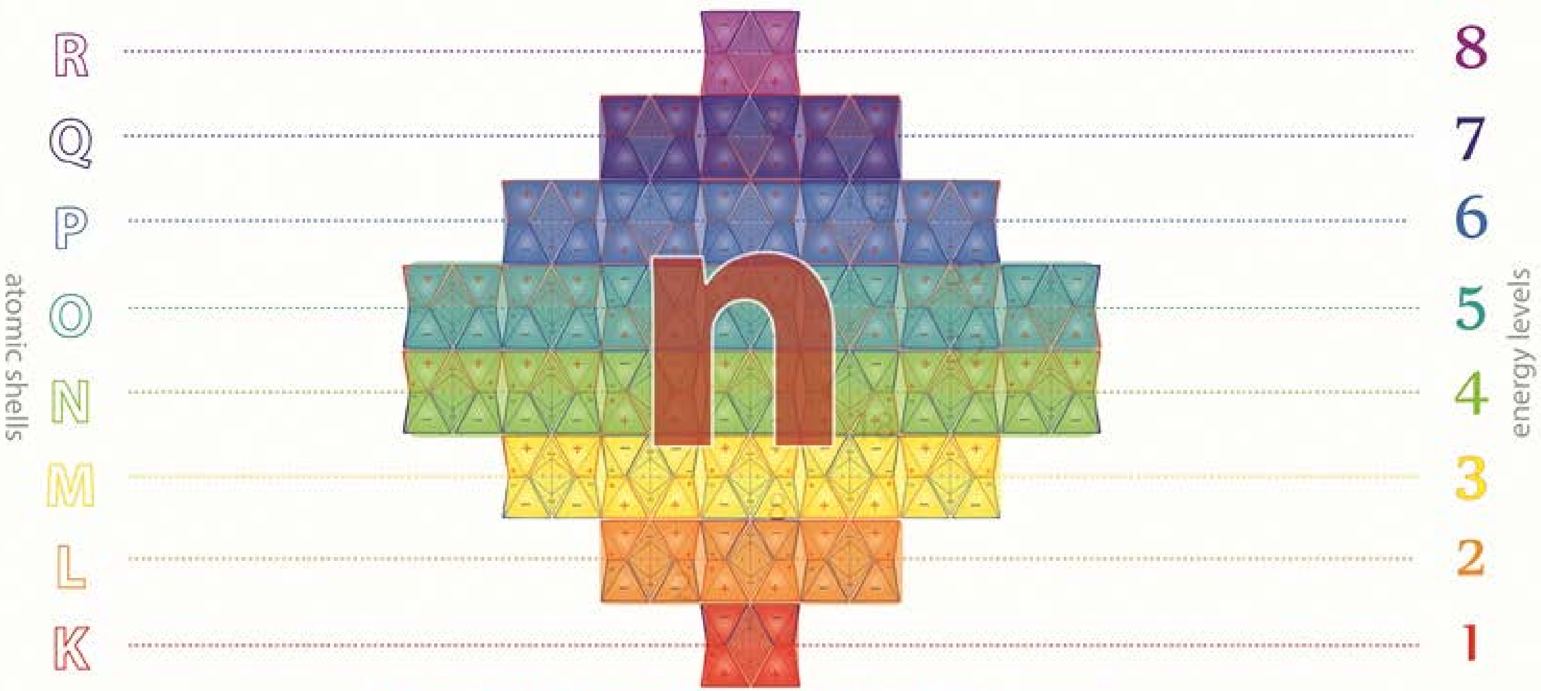

As energies of the Baryons comprising the atomic nuclei increases, the electron bound to each nuclei also possesses more KEM field energies and is therefore less tightly bound to the nucleus

Atomic Shells

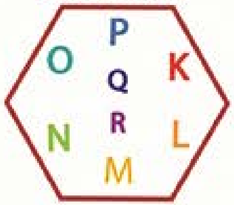

(K, L, M, N, O, P, Q, R)

Atomic shells relate directly to Principal quantum numbers

(1, 2, 3, 4, 5, 6, 7, 8)

Energy levels

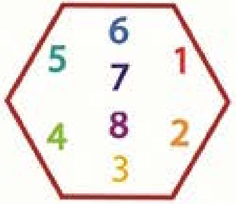

Copyright ABRAHAM [2008] - All rights reserved

AZMITHAL quantum number

The azimuthal quantum number is a quantum number assigned to any atomic orbital that describes its orbital angular momentum and determines the shape of the electron orbital

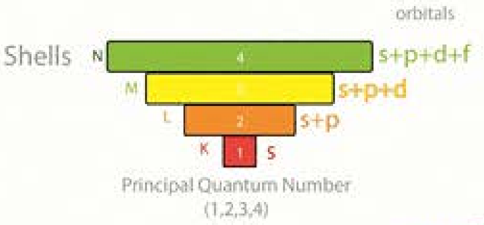
orbitals
Shells
N
4
s+p+d+f
M
s+p+d
L
2
s+p
K
1
s
Principal Quantum Number
(1,2,3,4)

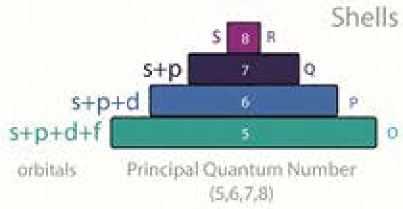
Shells
S
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron orbitals

$(l = 0, 1 \ldots n-1)$

ℓ (0-7)

Azimuthal

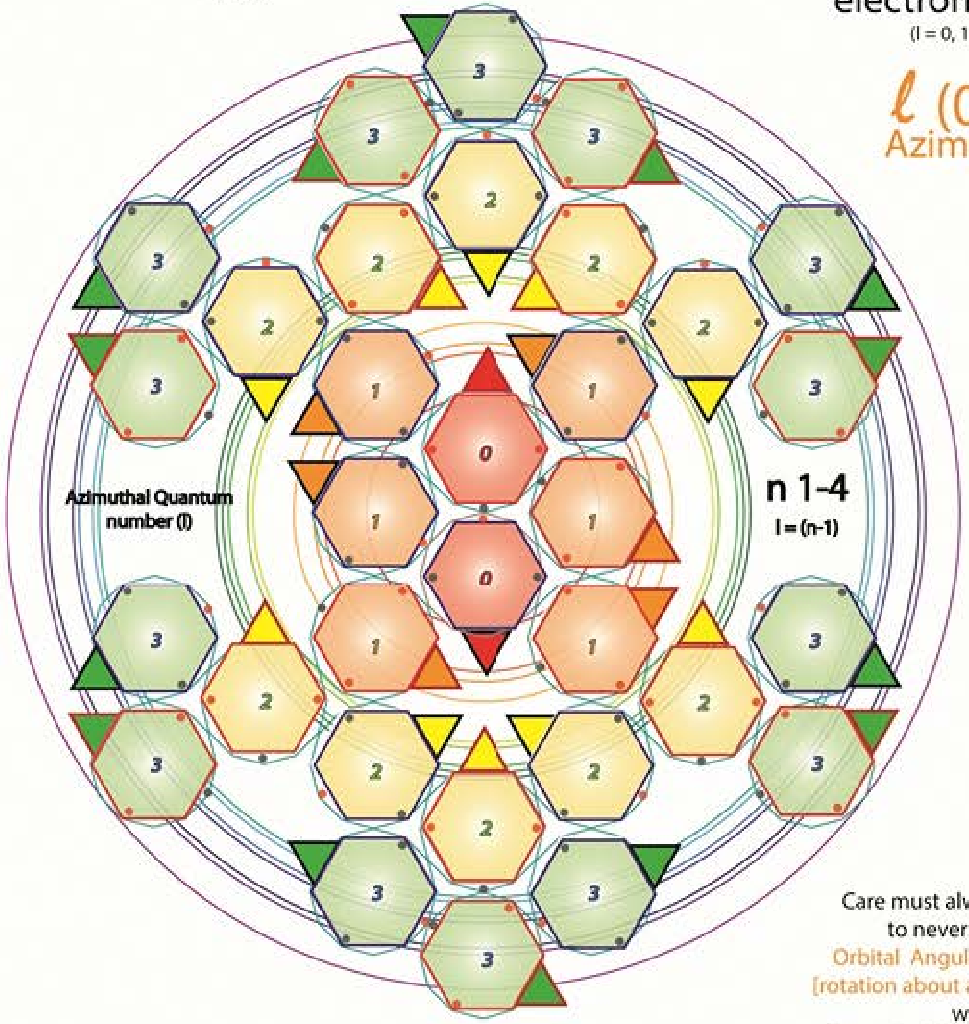
Azimuthal Quantum number (l)
n 1-4
l = (n-1)
0
1
2
3

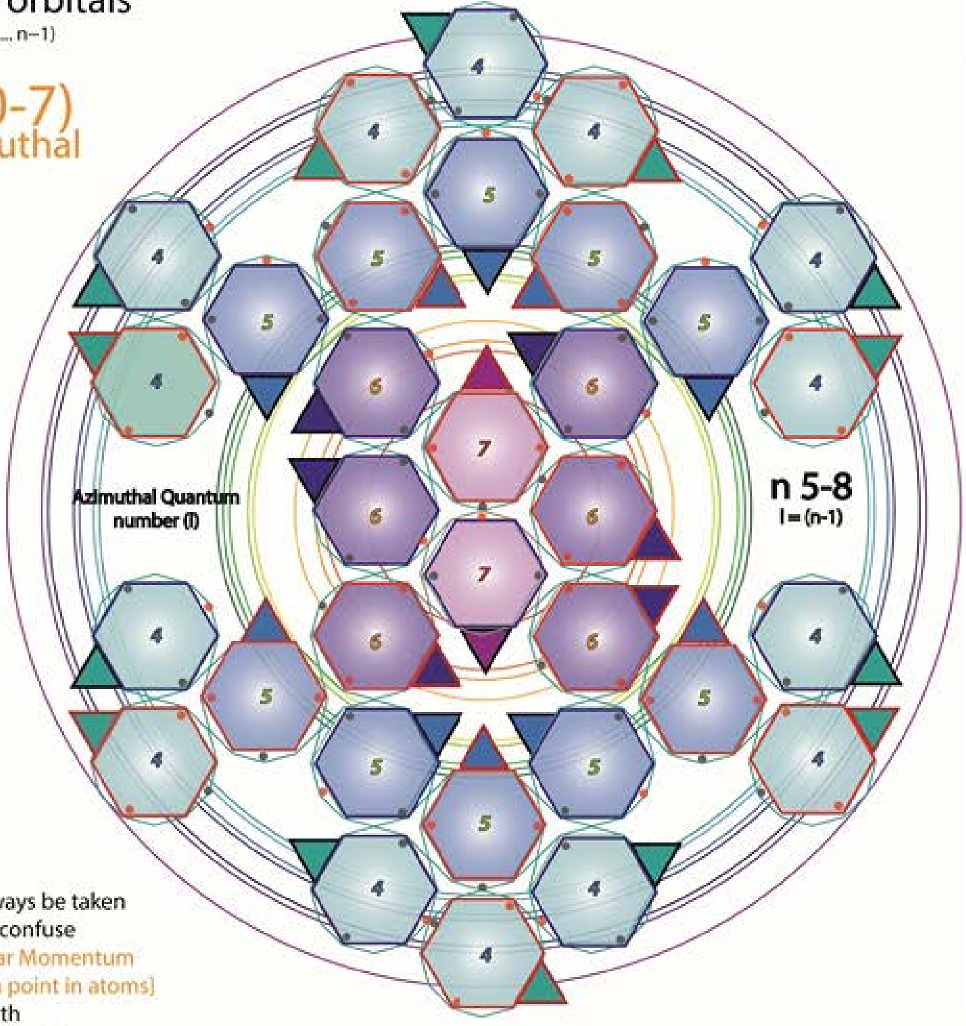
Azimuthal Quantum number (l)
n 5-8
l = (n-1)
4
5
6
7

Care must always be taken
to never confuse
Orbital Angular Momentum
[rotation about a point in atoms]
with
Quantised Angular Momenta
[equilateral Planck energy geometries]

Copyright ABRAHAM [2008] - All rights reserved

MAGNETIC quantum number

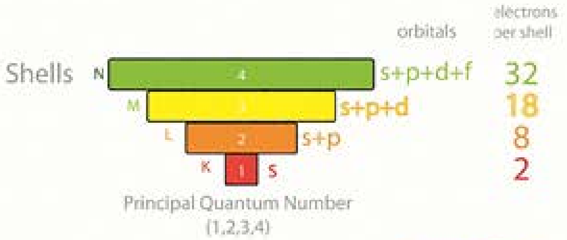
Shells
orbitals
electrons per shell
N
4
s+p+d+f
32
M
3
s+p+d
18
L
2
s+p
8
K
1
s
2
Principal Quantum Number
(1,2,3,4)

The magnetic quantum number denotes the energy levels available within any subshell

Magnetic numbers do not continue to increase as the Principal numbers increase instead they reverse after n4 to reflect the charged quanutm geometry of Elements and do not follow the current computer models in popular use

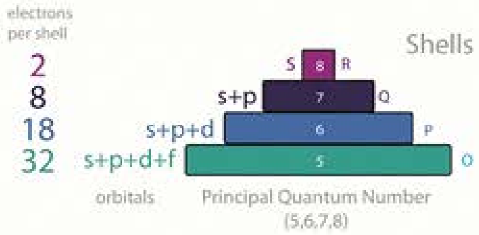
electrons per shell
2
8
18
32
Shells
S
8
R
s+p
7
Q
s+p+d
6
P
s+p+d+f
5
O
orbitals
Principal Quantum Number
(5,6,7,8)

s p d f

electron sub-orbitals

(ml = −l, −l+1 ... 0 ... l−1, l)

$m\ell\,(2\ell+1)$

Magnetic

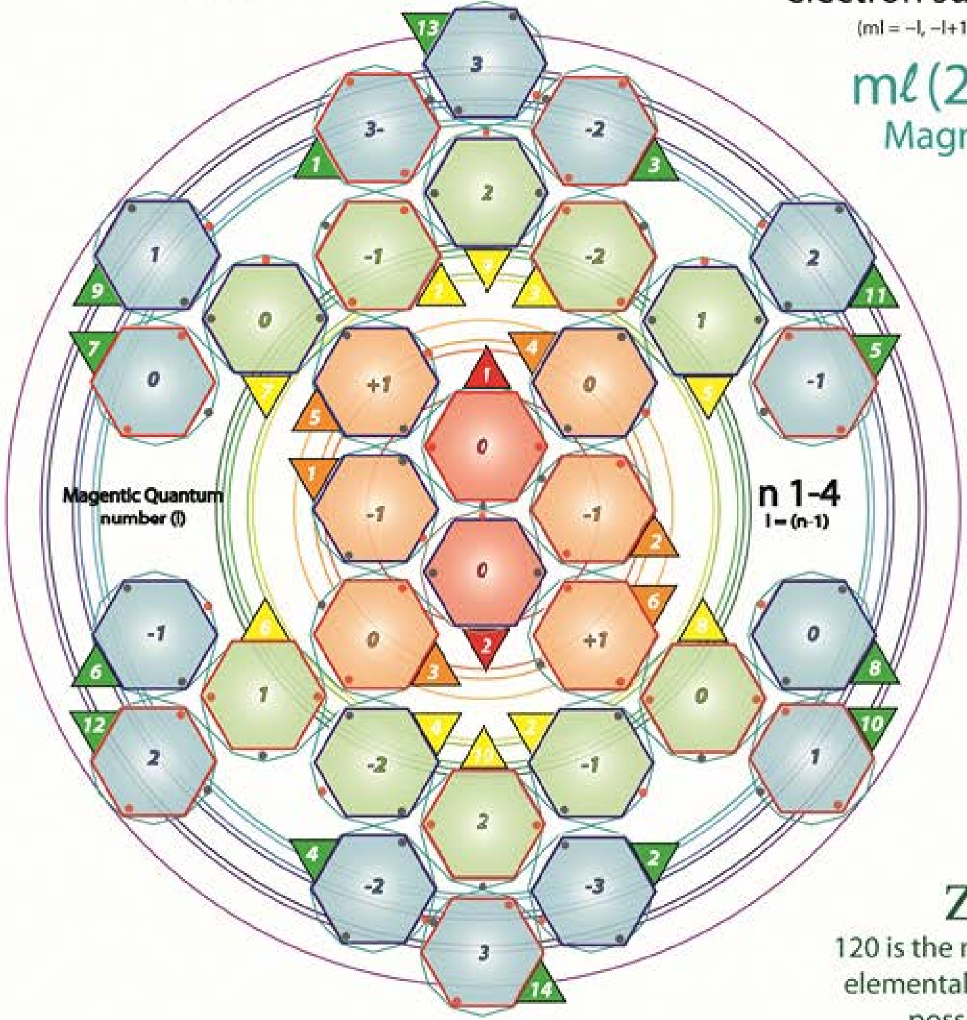
Magentic Quantum number (l)
n 1-4
l = (n-1)

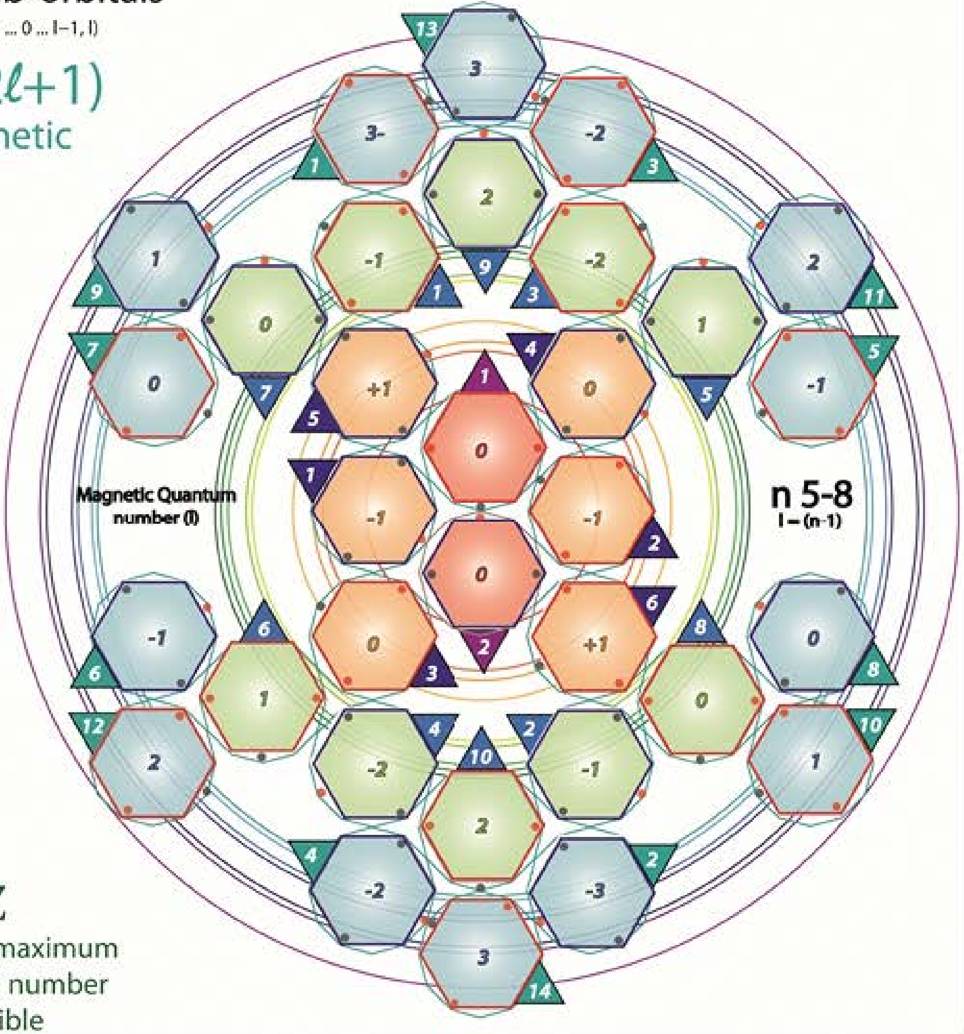
Magentic Quantum number (l)
n 5-8
l = (n-1)

Z

120 is the maximum elemental number possible

Copyright ABRAHAM [2008] - All rights reserved

SPIN quantum number

The spin quantum number is a quantum number that parameterizes
the intrinsic angular momentum (or spin angular momentum),
of any given electron anywhere in an atomic nucleus

parallel magnetic moments

Spin UP

higher coupling energies

anti-parallel magnetic moments

Spin DOWN

lower coupling energies

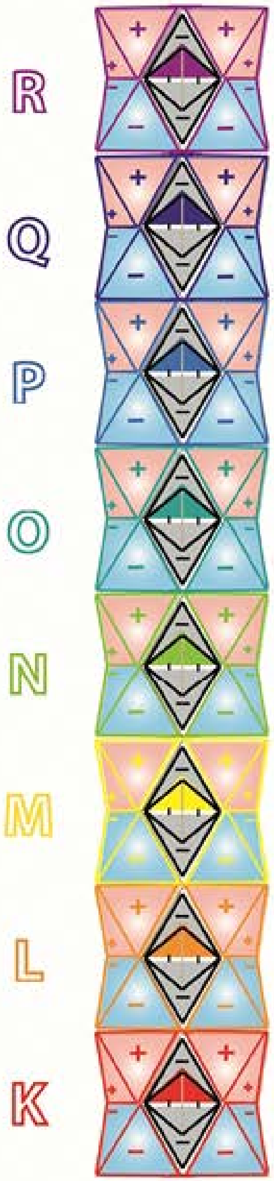

Electron spin can orientate in either direction within Nuclei, providing the nett spins follow the Hund rule and Pauli exclusion principle

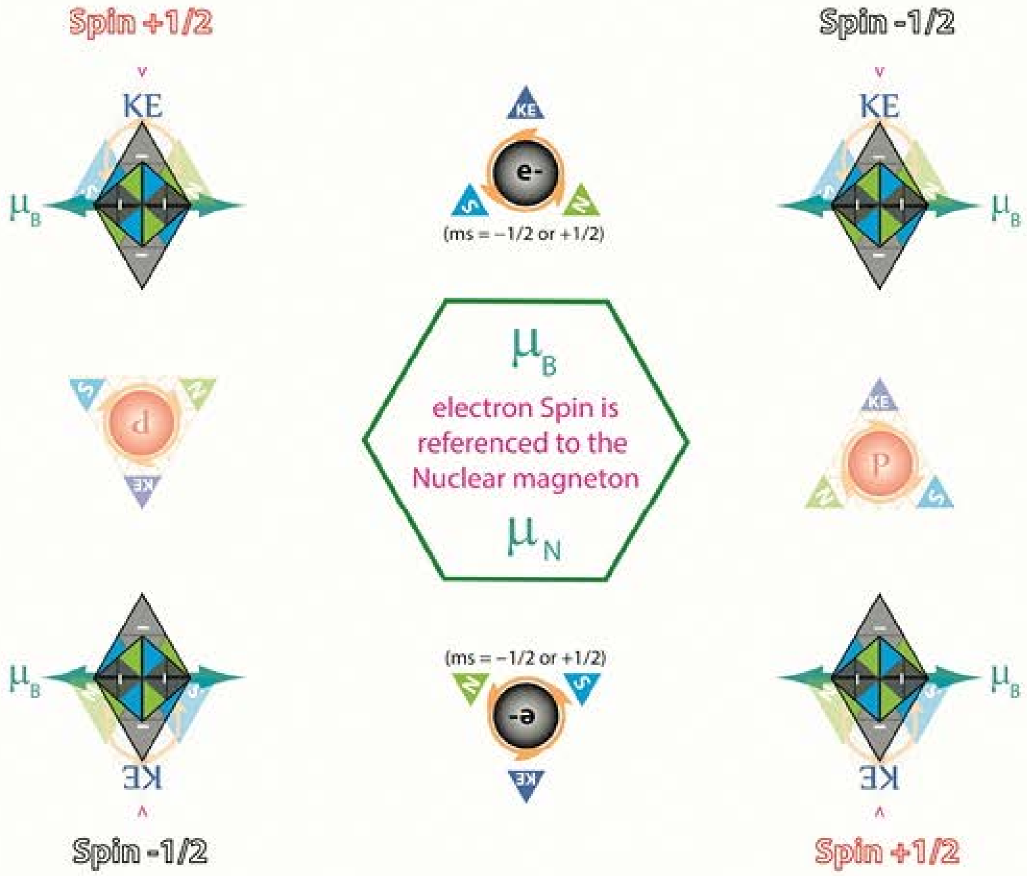
Spin +1/2
KE
μB
Spin -1/2
KE
μB
KE
e-
(ms = −1/2 or +1/2)
d
KE
P
KE
(ms = −1/2 or +1/2)
-e
KE
μB
KE
Spin -1/2
μB
KE
Spin +1/2
μB
μN

electron Spin is
referenced to the
Nuclear magneton

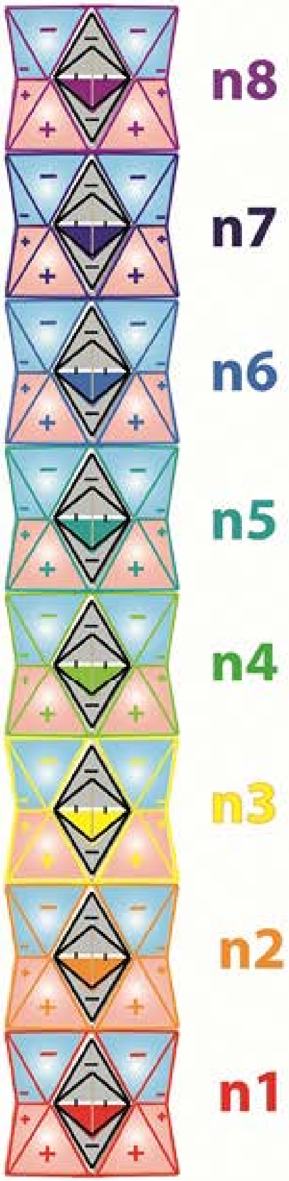
n8
n7
n6
n5
n4
n3
n2
n1

The nuclear energy levels of the Baryons comprising Elemental nuclei determine the energy-momenta of electrons bound to them

Copyright ABRAHAM [2008] - All rights reserved

Modifying Hund's Rule

Electrons fill orbitals in an alternating sequential numbering pattern due to nucleon placement creating opposed direction electron spins

The increased stability of the atom, most commonly manifested in a lower energy state, arises because the high-spin state forces the unpaired electrons to reside in different spatial orbitals.

A commonly given reason for the increased stability of high multiplicity states is that the different occupied spatial orbitals create a larger average distance between electrons, reducing electron-electron repulsion energy. In reality, it has been shown that the actual reason behind the increased stability is a decrease in the screening of electron-nuclear attractions[1].

The total spin state is calculated as the total number of unpaired electrons + 1, or twice the total spin + 1 written as 2s+1.

As a result of Hund's rule, constraints are placed on the way atomic orbitals are filled using the Aufbau principle.

Before any two electrons occupy an orbital in a subshell, other orbitals in the same subshell must first each contain one electron. Also, the electrons filling a subshell will have parallel spin before the shell starts filling up with the opposite spin electrons (after the first orbital gains a second electron).

As a result, when filling up atomic orbitals, the maximum number of unpaired electrons (and hence maximum total spin state) is assured

s	p	d	f
▲ 1 1 2 ▼	▼ 1 1 2 ▲	▲ 1 1 2 ▼	▲ 1 1 2 ▼
	▲ 3 2 4 ▼	▲ 3 2 4 ▼	▲ 3 2 4 ▼
	▼ 5 3 6 ▲	▼ 5 3 6 ▲	▲ 5 3 6 ▼
		▼ 7 4 8 ▲	▲ 7 4 8 ▼
		▼ 9 5 10 ▲	▼ 9 5 10 ▲
			▼ 11 6 12 ▲
			▼ 13 7 14 ▲

▲

Hund's rule of orbital filling
must be modified to reflect
the true orbital filling order

▼

Sub-orbitals fill in order of numbering

Electrons spins pair
before next orbital
is filled

ie.
p1 [DOWN] and p2 [UP] fill
before
p3 [UP] and p4 [DOWN]
before
p5 [DOWN] and p6 [UP] etc

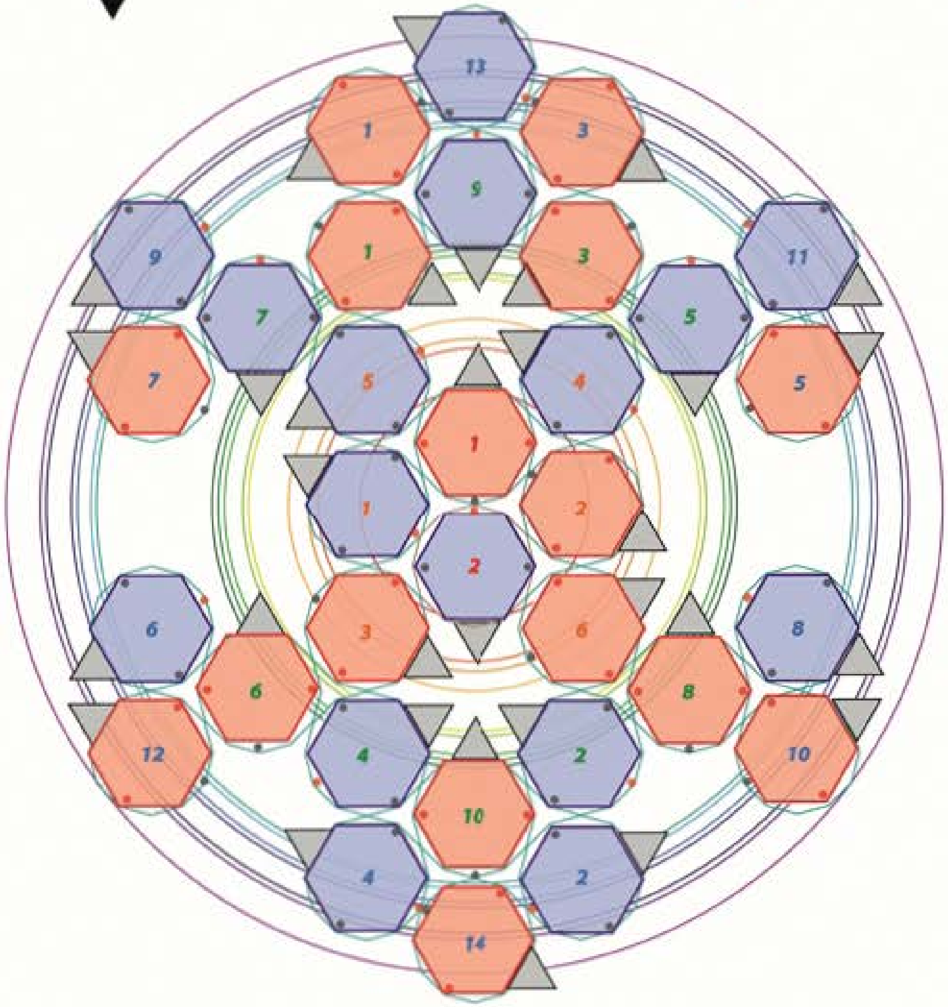

Copyright ABRAHAM [2008] - All rights reserved

Principle quantum Energies

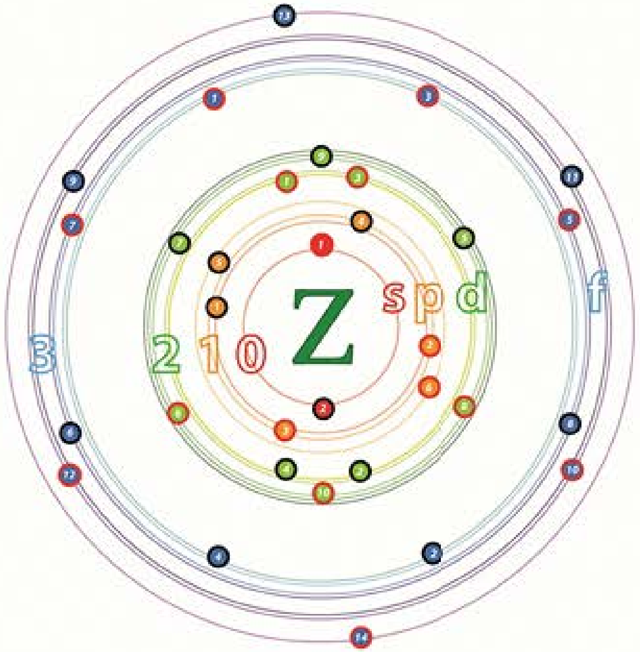

In an atom - electron energies are proportional to their intrinsic Kinetic Energies - which in turn are directly proportional to the quantum energy level of the nuclei which the electron binds to in their respective atomic shells

In a nucleus, lower energy orbits have less 'paired' nuclei supplying energy.
The more energy you give a nuclei the faster it casuses the bound electron to rotate.
If you give the nuclei enough energy, it will impart eneough energy to its electron for it to leave the system entirely.

The same is true for an electron orbital.
Higher values of n mean more energy for the electron and the corresponding KEM field enegies of the electron is larger, resulting in increased angular momentum.

Values of n start at 1 and go up by integer amounts.

If enough energy is added to the system by incident Photons a electron will leave the atom creating a positively charged nuclei [ionisation].

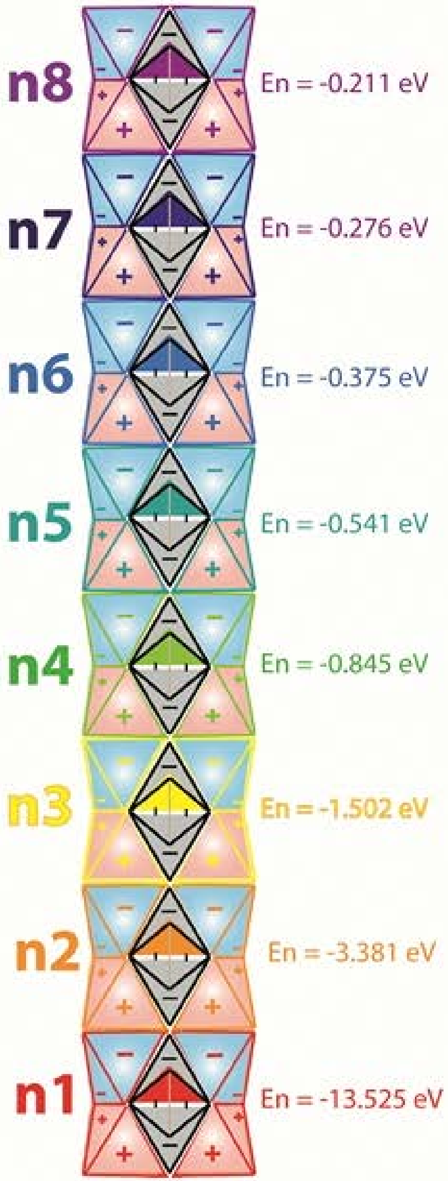

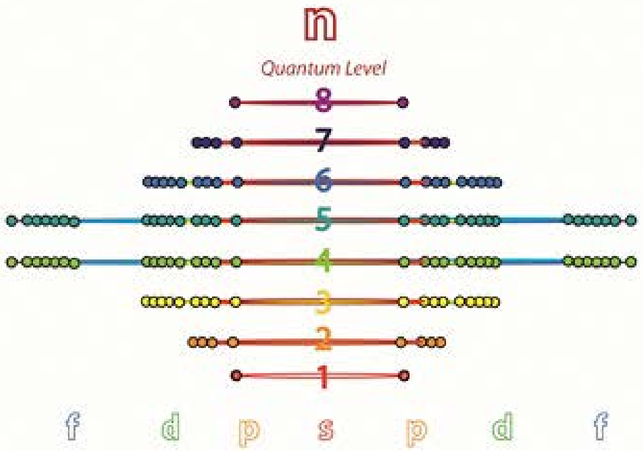

s orbits n1-8
p orbits n2-7
d orbits n3-6
f orbits n4-5

Eigenstate value

KEM field energy [per n] required to exceed 13.525 eV at which point the photo-electron has sufficient KE to break free of the Nucleus

$$E_n = \frac{E_1}{n^2} = \frac{-13.6eV}{n^2}, n = 1, 2, 3...$$

The possible Kinetic Energies (quantum levels) of an electron are directly related to the energy level of the Nuclei in each Quantum Level

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 1

K shell

n1

Ground State electron

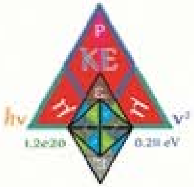
v
P
KE
hv
v²
1.2e20
0.28 eV

$iE = -13.313\ eV$

The energy levels of bound electrons is determined by Baryons

Z #	Name		Protons	electrons	Neutrons
1	1s1	Deuterium	1	1	1
2	1s2	Helium	2	2	2

2

Deuterium [not Hydrogen] is the building block of elements

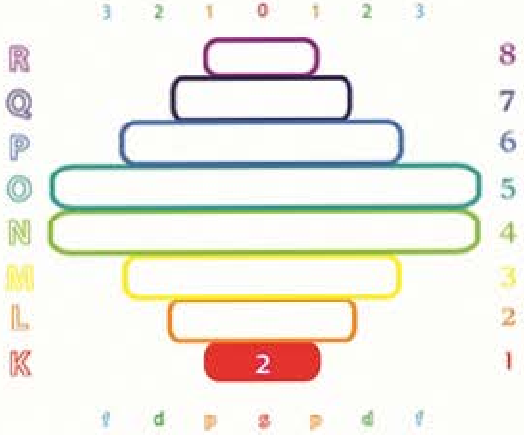
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

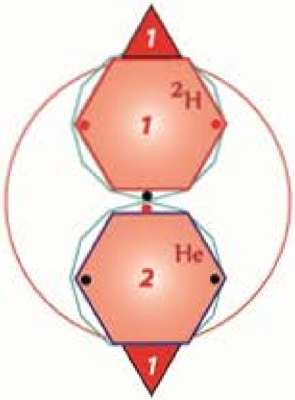

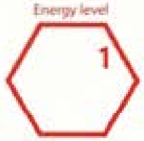
Energy level
1

Z #	Name		Protons	electrons	Neutrons
1	1s1	Hydrogen	1	1	0

Hydrogen is a free radical element

n1

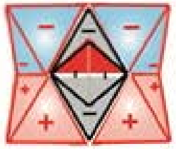

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 2

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
3	2s1	Lithium	3	3	3
4	2s2	Beryllium	4	4	4
5	1p1	Boron	5	5	5
6	1p2	Carbon	6	6	6
7	1p3	Nitrogen	7	7	7
8	1p4	Oxygen	8	8	8
9	1p5	Fluorine	9	9	9
10	1p6	Neon	10	10	10

8

Deuterium [not Hydrogen] is the building block of elements

L shell

n2

Ground State electron

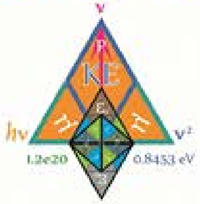
v
P
KE
hv
v²
1.2e20
0.8453 eV

$iE = -12.679\ eV$

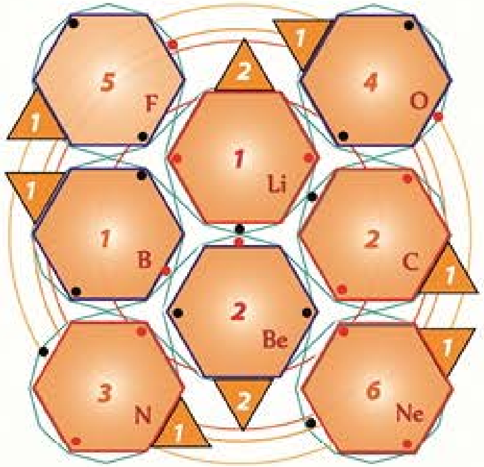
5
F
4
O
2
1
1
Li
1
1
B
2
C
1
2
Be
1
3
N
2
1
6
Ne
1

Energy level

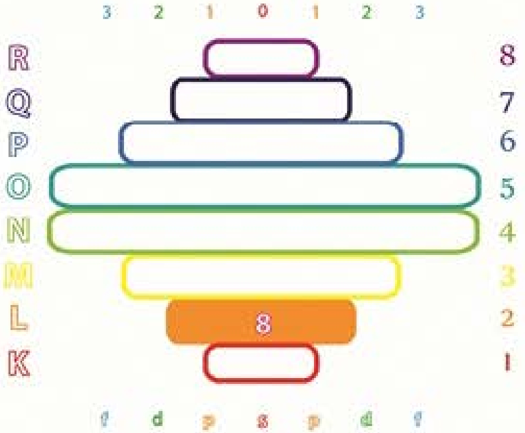
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
8
f
d
p
s
p
d
f

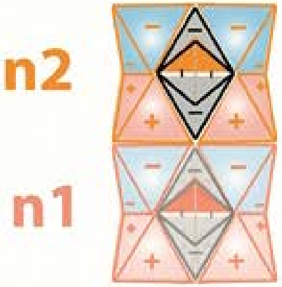

Quantum Level 3

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
11	3s1	Sodium	11	11	11
12	3s2	Magnesium	12	12	12
13	2p1	Aluminium	13	13	13
14	2p2	Silicon	14	14	14
15	2p3	Phosphorus	15	15	15
16	2p4	Sulfur	16	16	16
17	2p5	Chlorine	17	17	17
18	2p6	Argon	18	18	18
21	3d1	Scandium	21	21	21
22	3d2	Titanium	22	22	22
23	3d3	Vanadium	23	23	23
24	3d4	Chromium	24	24	24
25	3d5	Manganese	25	25	25
26	3d6	Iron	26	26	26
27	3d7	Cobalt	27	27	27
28	3d8	Nickel	28	28	28
29	3d9	Copper	29	29	29
30	3d10	Zinc	30	30	30

18

Deuterium [not Hydrogen] is the building block of elements

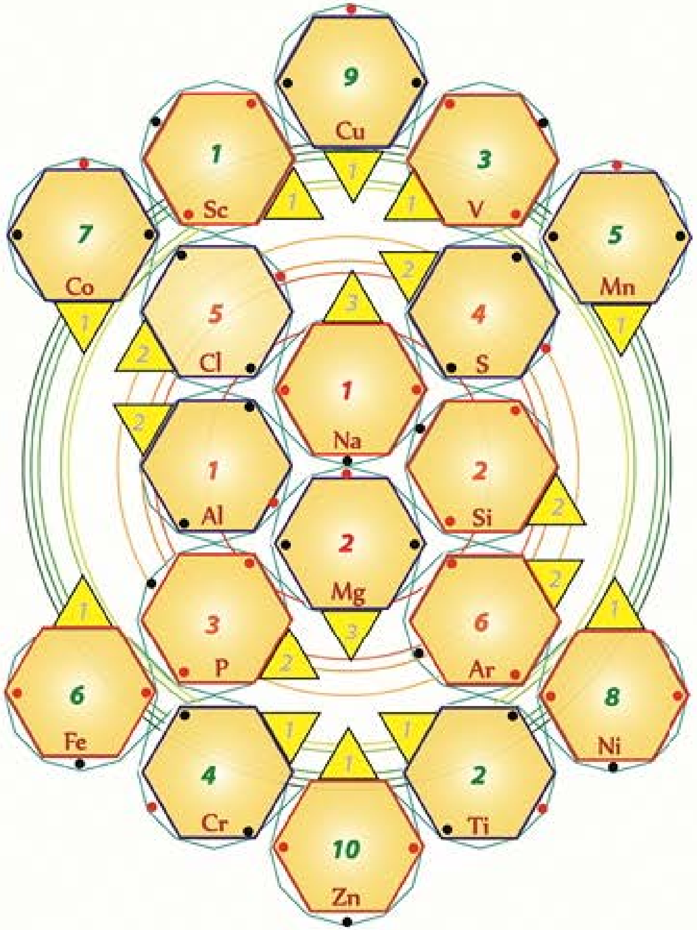
9
Cu
1
Sc
3
V
7
Co
5
Mn
5
Cl
4
S
1
Na
1
Al
2
Si
2
Mg
3
P
6
Ar
6
Fe
8
Ni
4
Cr
2
Ti
10
Zn
1
2
3

M shell

n3

Ground State electron

v
KE
hv
v²
1.2e20
1.9019 eV

$iE = -11.623\ eV$

Energy level

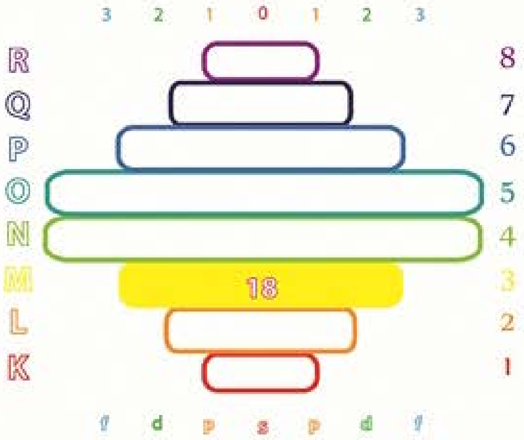
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
18
f
d
p
s
p
d
f

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 4

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
19	4s1	Potassium	19	19	19
20	4s2	Calcium	20	20	20
31	4p1	Gallium	31	31	31
32	4p2	Germanium	32	32	32
33	4p3	Arsenic	33	33	33
34	4p4	Selenium	34	34	34
35	4p5	Bromine	35	35	35
36	4p6	Krypton	36	36	36
39	4d1	Yttrium	39	39	39
40	4d2	Zirconium	40	40	40
41	4d3	Niobium	41	41	41
42	4d4	Molybdenum	42	42	42
43	4d5	Technetium	43	43	43
44	4d6	Ruthenium	44	44	44
45	4d7	Rhodium	45	45	45
46	4d8	Palladium	46	46	46
47	4d9	Silver	47	47	47
48	4d10	Cadmium	48	48	48
57	4f1	Lanthanum	57	57	57
58	4f2	Cerium	58	58	58
59	4f3	Praseodymium	59	59	59
60	4f4	Neodymium	60	60	60
61	4f5	Promethium	61	61	61
62	4f6	Samarium	62	62	62
63	4f7	Europium	63	63	63
64	4f8	Gadolinium	64	64	64
65	4f9	Terbium	65	65	65
66	4f10	Dysprosium	66	66	66
67	4f11	Holium	67	67	67
68	4f12	Erbium	68	68	68
69	4f13	Thulium	69	69	69
70	4f14	Ytterbium	70	70	70

32

Deuterium [not Hydrogen] is the building block of elements

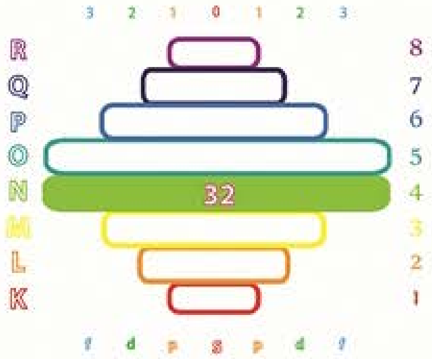
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
32
f d p s p d f

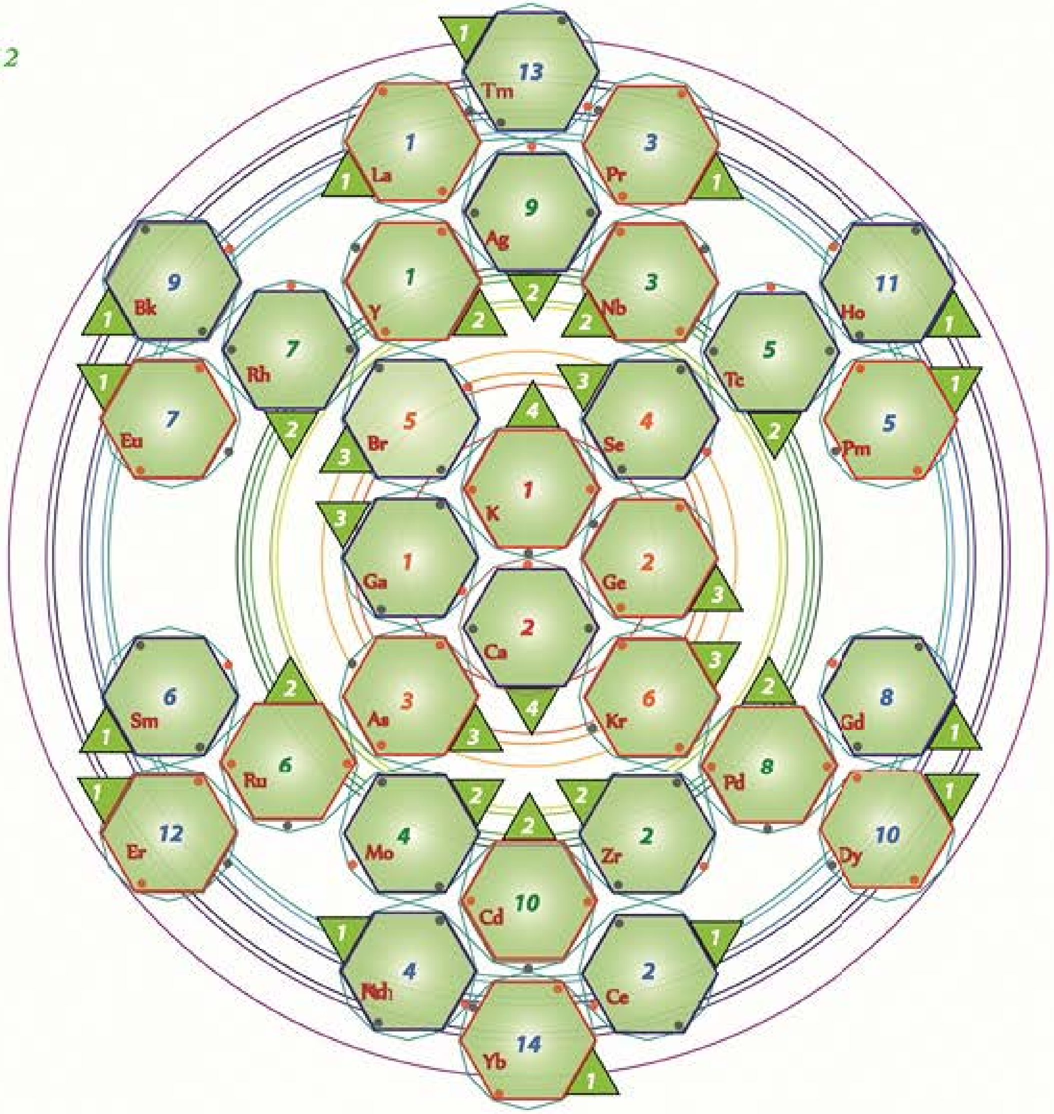
13
Tm
1
La
3
Pr
9
Ag
9
Bk
1
Y
3
Nb
11
Ho
7
Rh
5
Tc
7
Eu
5
Br
4
Se
5
Pm
1
K
1
Ga
2
Ge
2
Ca
6
Sm
3
As
6
Kr
8
Gd
6
Ru
8
Pd
12
Er
4
Mo
2
Zr
10
Dy
10
Cd
4
Nd
2
Ce
14
Yb

N shell

n4

Ground State electron

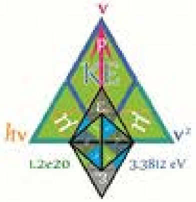
hv
v²
1.2e20
3.3812 eV

iE = -10.143 eV

Energy level

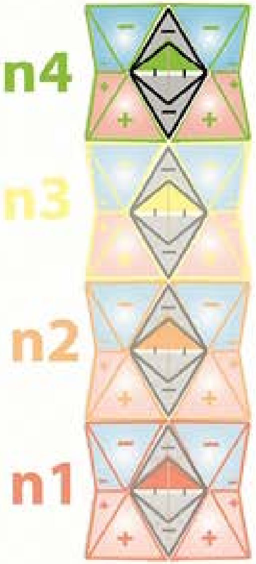

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 5

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
37	5s1	Rubidium	37	37	37
38	5s2	Strontium	38	38	38
49	5p1	Indium	49	49	49
50	5p2	Tin	50	50	50
51	5p3	Antimony	51	51	51
52	5p4	Tellurium	52	52	52
53	5p5	Iodine	53	53	53
54	5p6	Xenon	54	54	54
71	5d1	Lutetium	71	71	71
72	5d2	Hafnium	72	72	72
73	5d3	Tantalum	73	73	73
74	5d4	Tungsten	74	74	74
75	5d5	Rhenium	75	75	75
76	5d6	Osmium	76	76	76
77	5d7	Iridium	77	77	77
78	5d8	Platinum	78	78	78
79	5d9	Gold	79	79	79
80	5d10	Mercury	80	80	80
89	5f1	Actinium	89	89	89
90	5f2	Thorium	90	90	90
91	5f3	Protactium	91	91	91
92	5f4	Uranium	92	92	92
93	5f5	Neptunium	93	93	93
94	5f6	Plutonium	94	94	94
95	5f7	Americium	95	95	95
96	5f8	Curium	96	96	96
97	5f9	Berkelium	97	97	97
98	5f10	Californium	98	98	98
99	5f11	Einsteinium	99	99	99
100	5f12	Fermium	100	100	100
101	5f13	Mendelevium	101	101	101
102	5f14	Nobelium	102	102	102

32

Deuterium [not Hydrogen] is the building block of elements

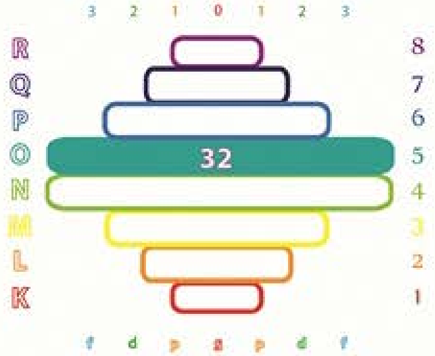
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
32
f d p s p d f

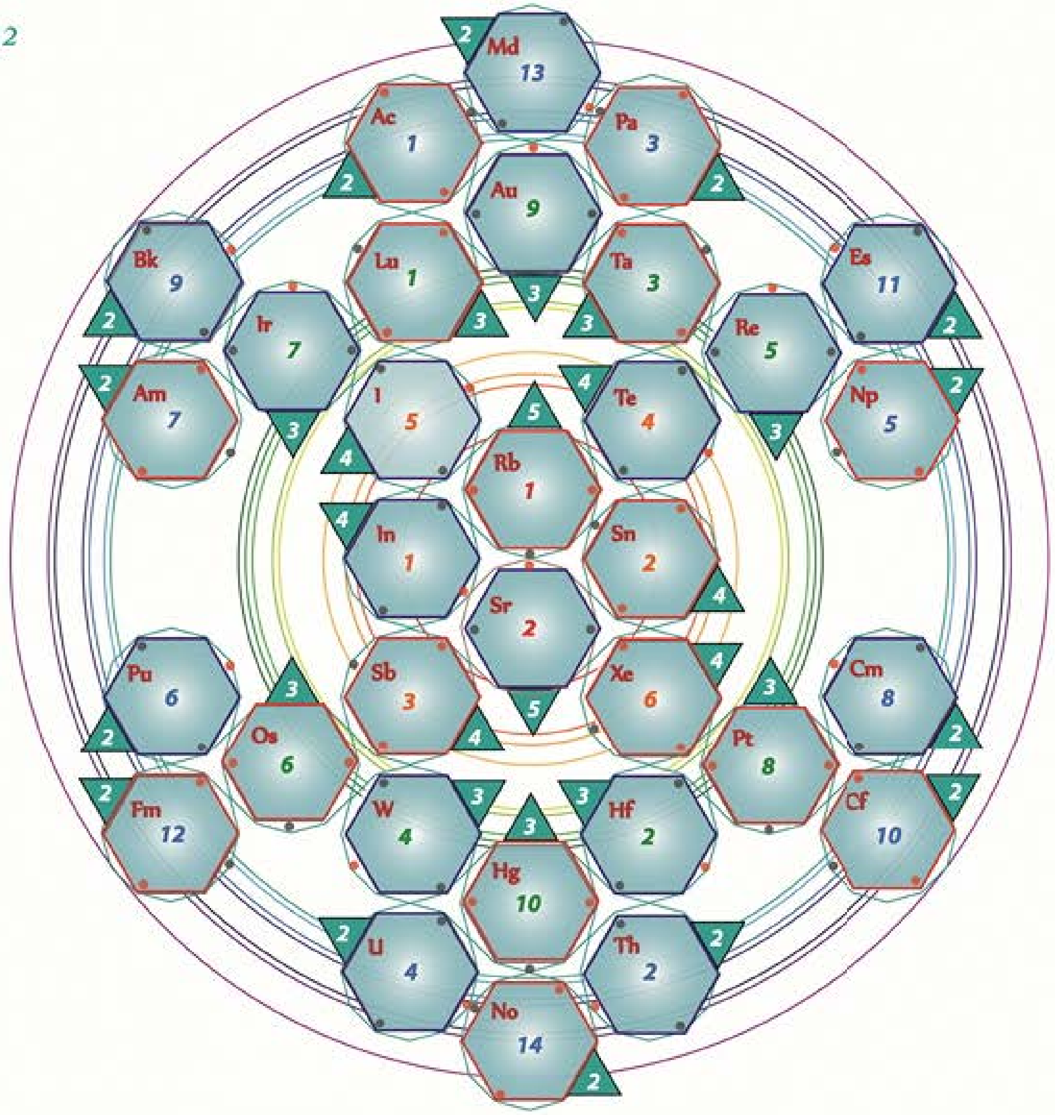
Md 13
Ac 1
Pa 3
Au 9
Bk 9
Lu 1
Ta 3
Es 11
Ir 7
Re 5
Am 7
I 5
Te 4
Np 5
Rb 1
In 1
Sn 2
Sr 2
Pu 6
Sb 3
Xe 6
Cm 8
Os 6
Pt 8
Fm 12
W 4
Hf 2
Cf 10
Hg 10
U 4
Th 2
No 14

O shell

n5

Ground State electron

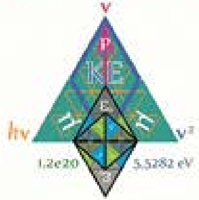
v
P
KE
hv
v²
1,2e20
5,5282 eV

iE = -8.241 eV

Energy level
5

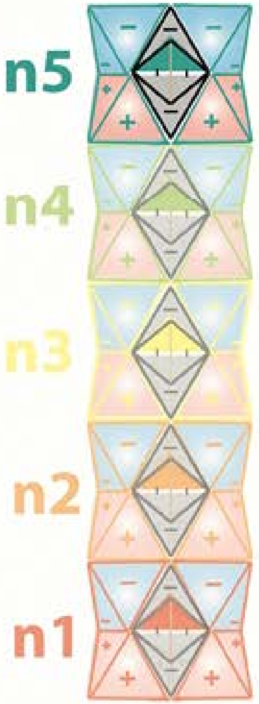
n5
n4
n3
n2
n1

Quantum Level 6

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
55	6s1	Caesium	55	55	55
56	6s2	Barium	56	56	56
81	6p1	Thallium	81	81	81
82	6p2	Lead	82	82	82
83	6p3	Bismuth	83	83	83
84	6p4	Polonium	84	84	84
85	6p5	Astatine	85	85	85
86	6p6	Radon	86	86	86
103	6d1	Lawrencium	103	103	103
104	6d2	Rutherfordium	104	104	104
105	6d3	Dubnium	105	105	105
106	6d4	Seaborgium	106	106	106
107	6d5	Bohrium	107	107	107
108	6d6	Hassium	108	108	108
109	6d7	Meitnerium	109	109	109
110	6d8	Darmstadtium	110	110	110
111	6d9	Roetgenium	111	111	111
112	6d10	Copernicium	112	112	112

18

Deuterium [not Hydrogen] is the building block of elements

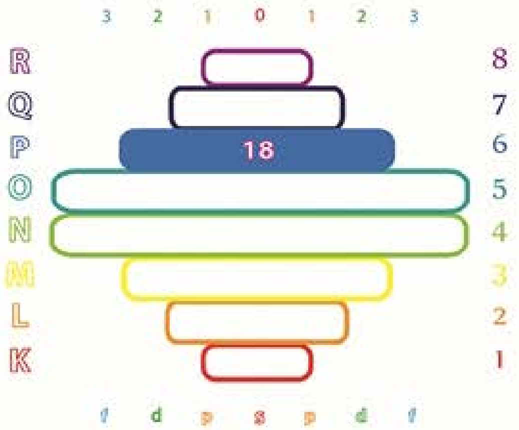
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
18
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

Energy level
6

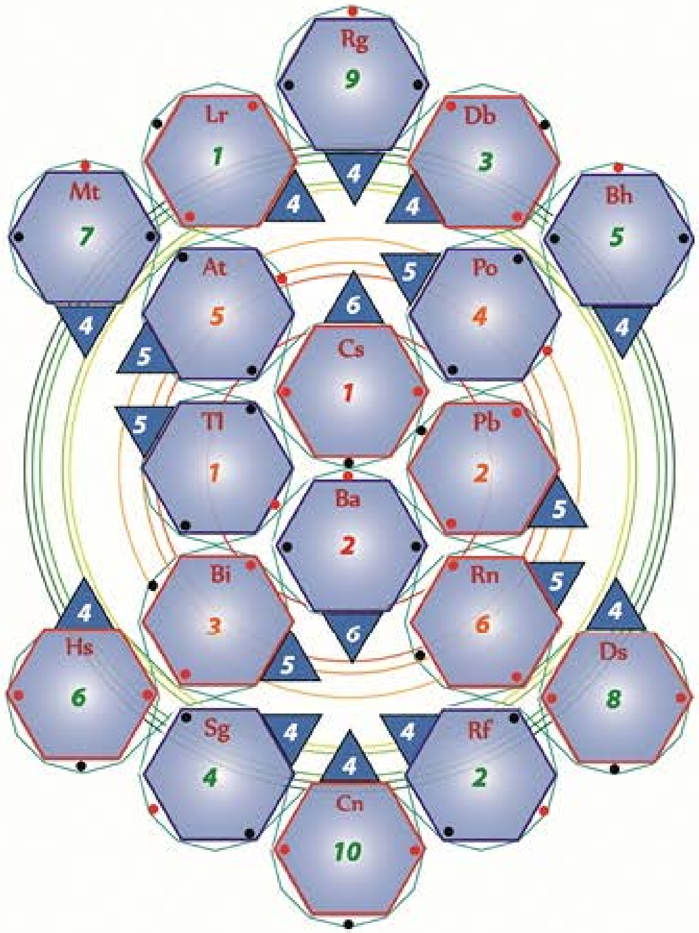
Rg
9
Lr
1
Db
3
Mt
7
Bh
5
At
5
Po
4
Cs
1
Tl
1
Pb
2
Ba
2
Bi
3
Rn
6
Hs
6
Ds
8
Sg
4
Rf
2
Cn
10
4
5
6

P shell

n6

Ground State electron

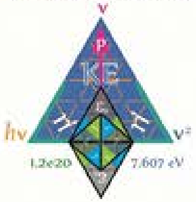
v
P
KE
hv
v²
1.2e20
7.607 eV

$iE = -5.917\ eV$

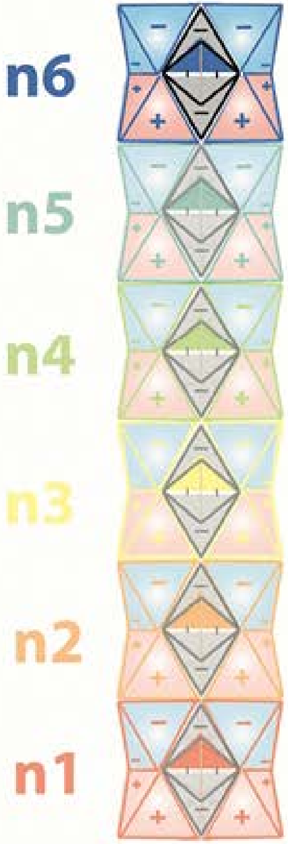
n6
n5
n4
n3
n2
n1

Quantum Level 7

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
87	*7s1*	**Francium**	87	87	87
88	*7s2*	**Radium**	88	88	88
113	*7p1*	**Ununtrium**	113	113	113
114	*7p2*	**Flerovium**	114	114	114
115	*7p3*	**Ununpentium**	115	115	115
116	*7p4*	**Livermorium**	116	116	116
117	*7p5*	**Ununseptium**	117	117	117
118	*7p6*	**Ununoctium**	118	118	118

8

Deuterium [not Hydrogen] is the building block of elements

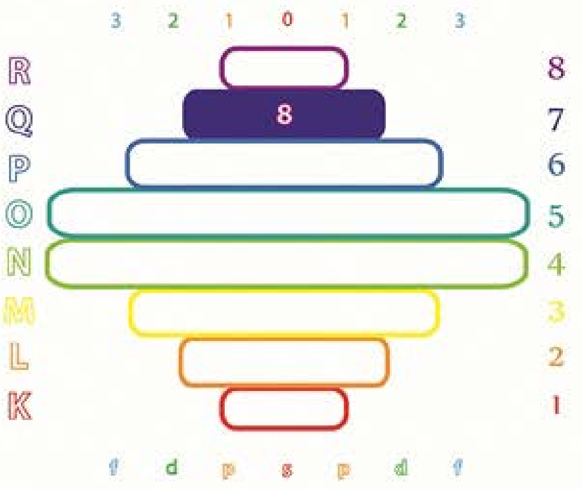
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

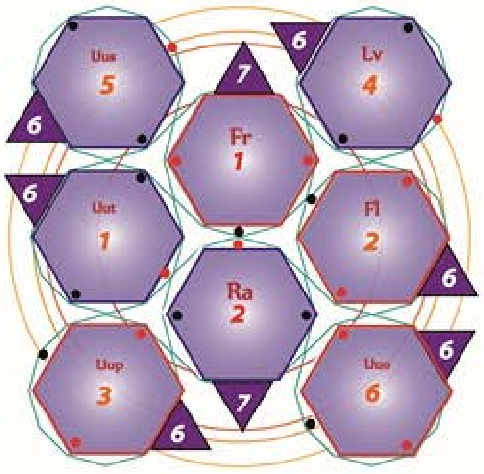
Uus
5
Lv
4
Fr
1
Uut
1
Fl
2
Ra
2
Uup
3
Uuo
6
7
7
6
6
6
6
6
6

Q shell

n7

Ground State electron

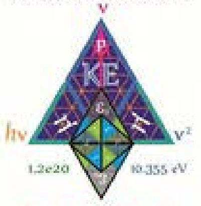
v
P
KE
hv
v²
1.2e20
10.355 eV

iE = -3.169 eV

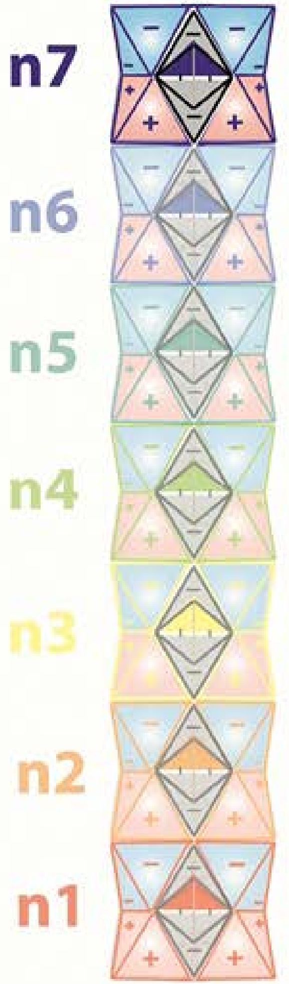
n7
n6
n5
n4
n3
n2
n1
−
+

Energy level

Copyright ABRAHAM [2008] - All rights reserved

Quantum Level 8

The energy levels of bound electrons is determined by Baryons

Z #		Name	Protons	electrons	Neutrons
119	*8s1*	**Ununnonium**	119	119	119
120	*8s2*	**Unbinilium**	120	120	120

2

Deuterium [not Hydrogen] is the building block of elements

R shell

n8

Ground State electron

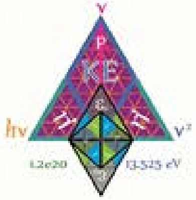
v
P
KE
hv
v²
1.2e20
13.525 eV

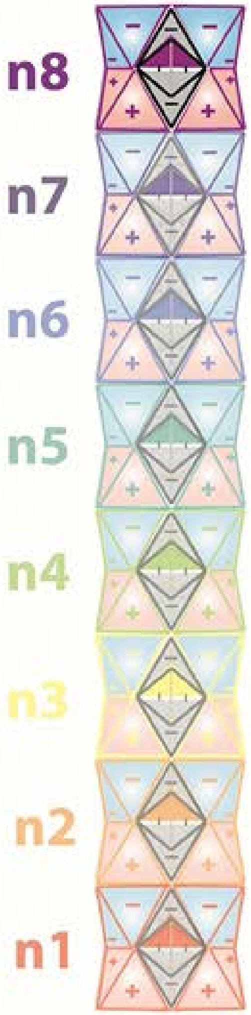
n8
n7
n6
n5
n4
n3
n2
n1
−
+

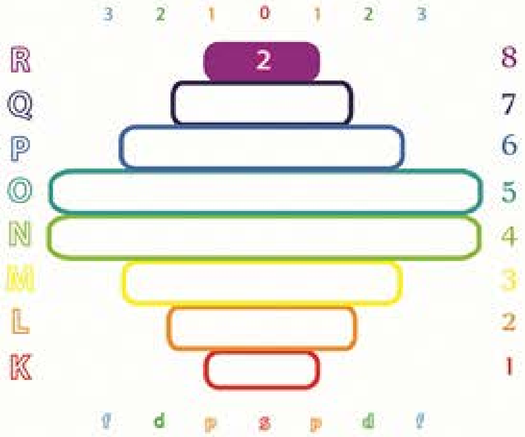
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
2
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

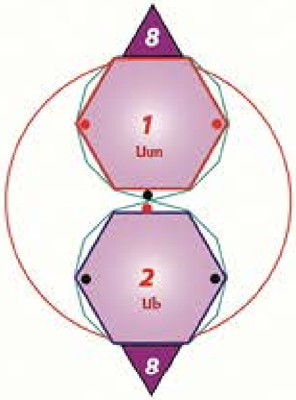

Energy level

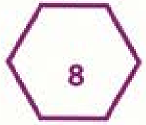

Quantum Level Jumps

Photon Absorption and Release

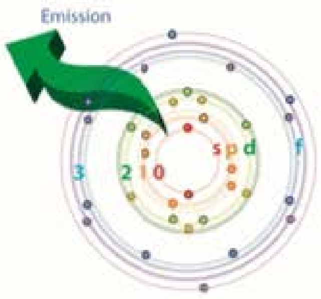
Emission
3
2
1
0
s
p
d
f

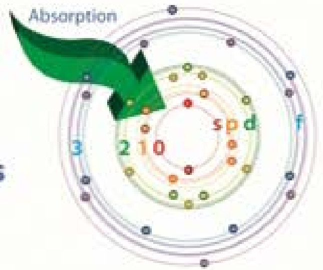
Absorption
3
2
1
0
s
p
d
f

Photo-electrons can only transition between principal energy Baryons in the atomic nuclei in discrete steps [or quantum jumps] because Baryons determine the KEM energy levels of electrons in nuclei

[see Tetryonic QED for full details on spectral line mechanics]

Spectral line transitions

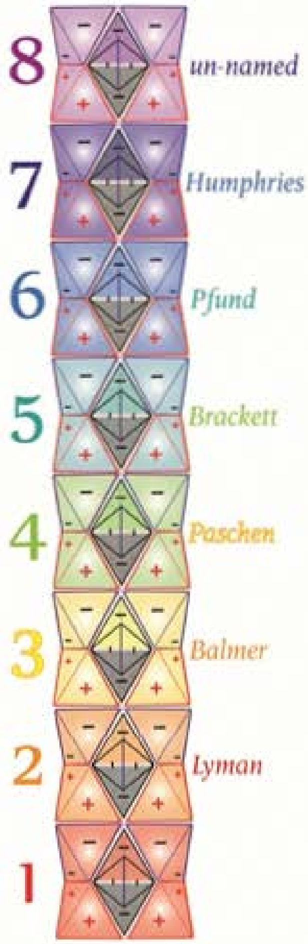
8
un-named
7
Humphries
6
Pfund
5
Brackett
4
Paschen
3
Balmer
2
Lyman
1

Quantum level jumps

Initial Quantum level

8	768	756	720	660	576	468	336	180	0
7	588	576	540	480	396	288	156	0	
6	432	420	384	324	240	132	0		
5	300	288	252	192	108	0			
4	192	180	144	84	0				
3	108	96	60	0					
2	48	36	0						
1	12	0							
		12	48	108	192	300	432	588	768
		1	2	3	4	5	6	7	8

Final Quantum level

$$\Delta p = \Delta M v = hf$$

accelerating photo-electrons produce spectral lines

Spectral line transitions

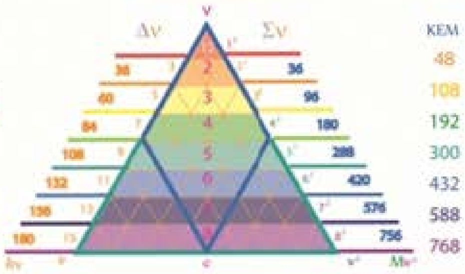
v
Δv
Σv
KEM
48
108
192
300
432
588
768
36
60
84
108
132
156
180
36
96
180
288
420
576
756
1
2
3
4
5
6
7
8
hv
c
v²
Mv²

Quantum differentials

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity through equilateral geometry

Nuclear energy emission~absorption

If atoms are left undisturbed, their electrons usually fill the lowest available energy levels and stay there, in their "ground state."

Occasionally, however, they may also be pushed up to some higher energy ("become excited") e.g. by a collision with a fast atom or electron, one which got extra speed from an electric voltage or from some source of heat.

An atom/electron elevated to one of its higher "excited levels" soon falls back to a lower level ("undergoes a quantum jump"), emitting a photon whose energy corresponds to the difference between the levels.

That need not be the ground state: the atom/electron might descend to that state in several steps, emitting a photon at each step on the way.

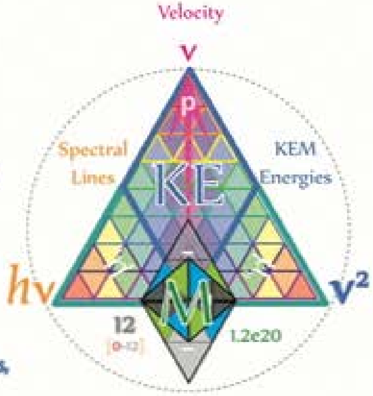
Velocity
v
p
Spectral Lines
KE
KEM Energies
hv
M
v²
12
[0-12]
1.2e20

Quantum transitions

(Orbital Shells - Bound energy states)

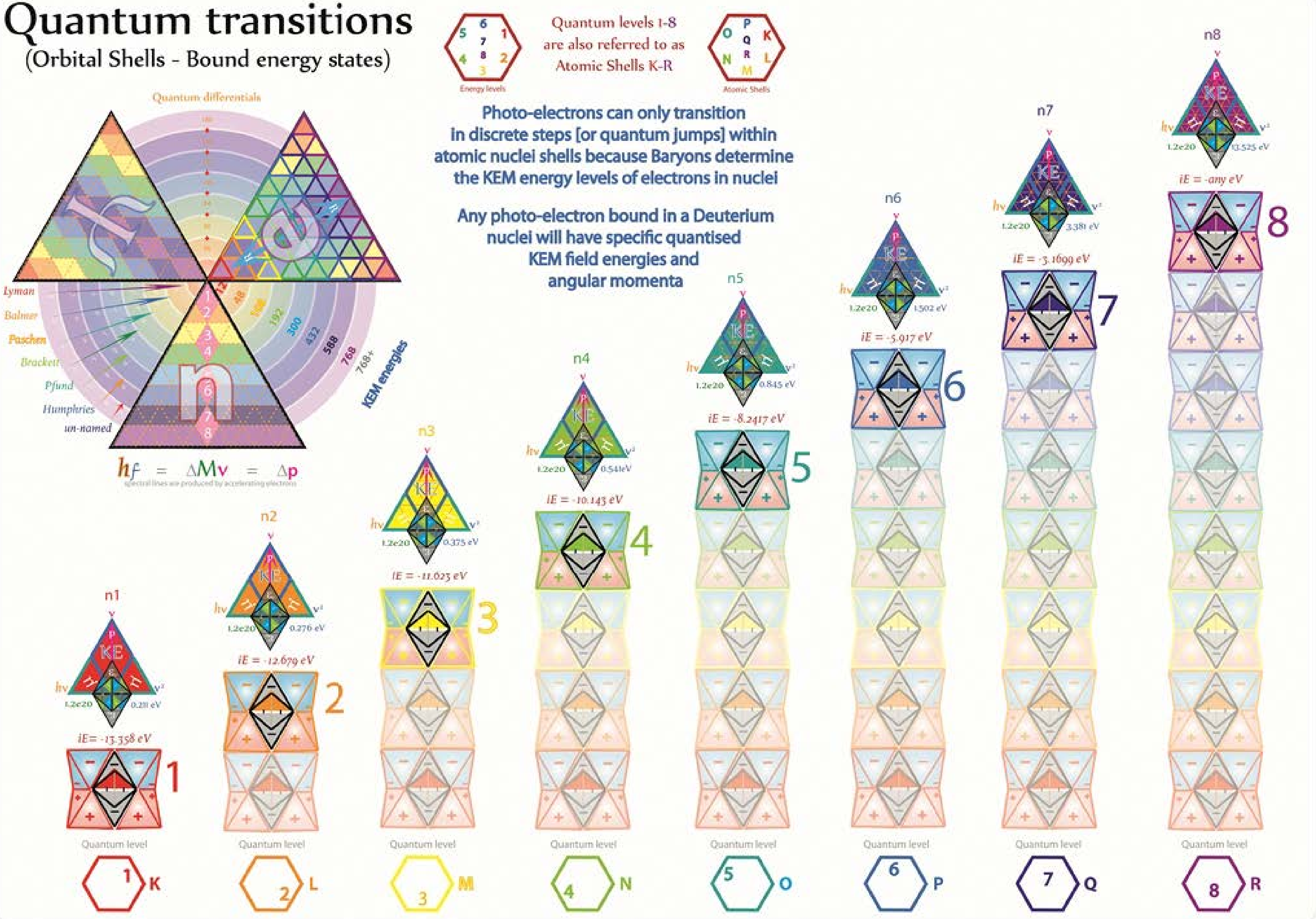

Energy levels
Atomic Shells
Quantum differentials
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
12
48
108
192
300
432
588
768
768+
KEM energies
$hf = \Delta Mv = \Delta p$
spectral lines are produced by accelerating electrons
n1
hv
1.2e20
0.211 eV
iE= -13.358 eV
1
n2
0.276 eV
iE = -12.679 eV
2
n3
0.375 eV
iE = -11.623 eV
3
n4
0.541eV
iE = -10.143 eV
4
n5
0.845 eV
iE = -8.2417 eV
5
n6
1.502 eV
iE = -5.917 eV
6
n7
3.381 eV
iE = -3.1699 eV
7
n8
13.525 eV
iE = -any eV
8
KE
Quantum level
1 K
2 L
3 M
4 N
5 O
6 P
7 Q
8 R

Quantum levels 1-8
are also referred to as
Atomic Shells K-R

Photo-electrons can only transition in discrete steps [or quantum jumps] within atomic nuclei shells because Baryons determine the KEM energy levels of electrons in nuclei

Any photo-electron bound in a Deuterium nuclei will have specific quantised KEM field energies and angular momenta

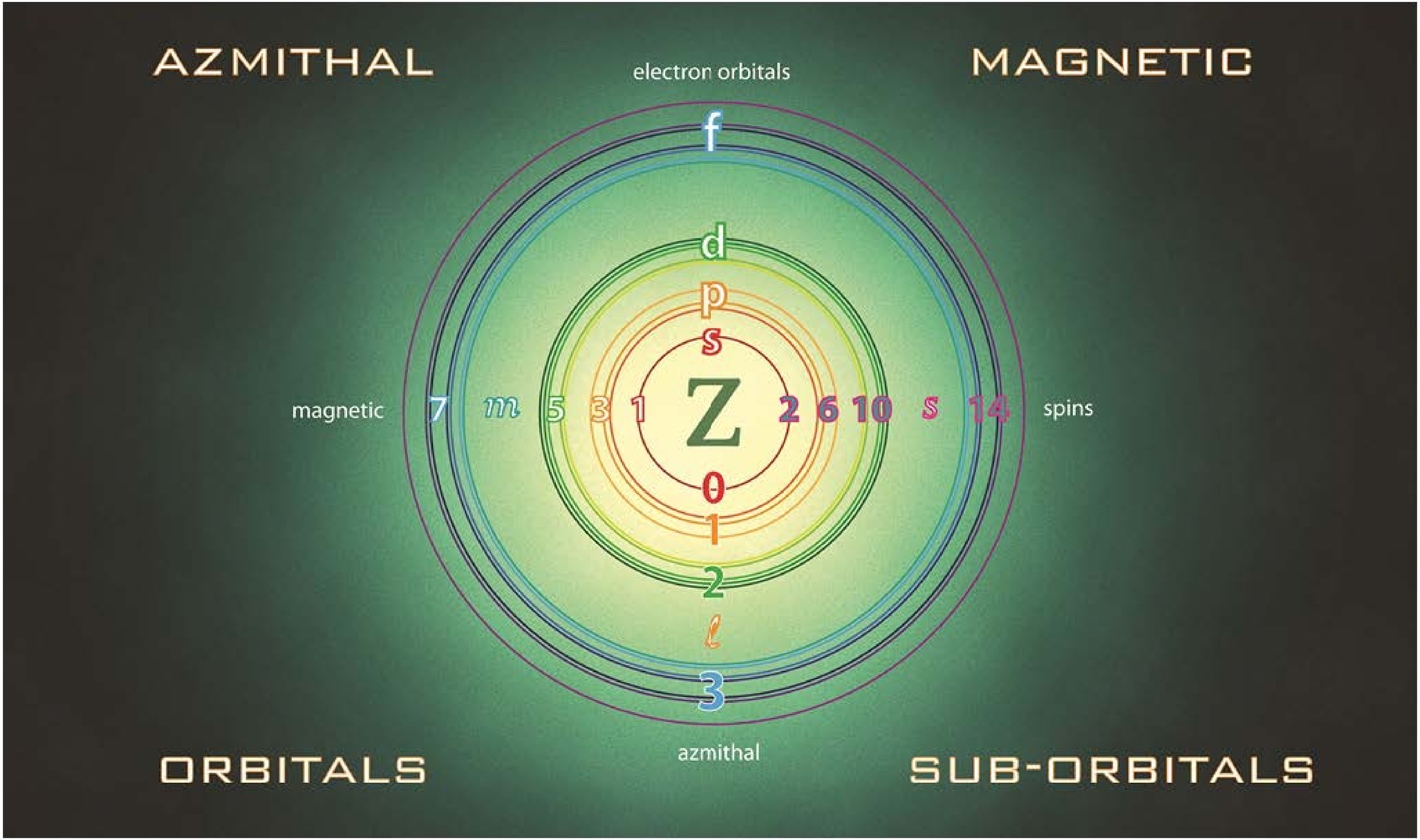
AZMITHAL
MAGNETIC
ORBITALS
SUB-ORBITALS
electron orbitals
f
d
p
s
Z
magnetic
7
m
5
3
1
2
6
10
s
14
spins
0
1
2
ℓ
3
azmithal

Atomic Orbitals

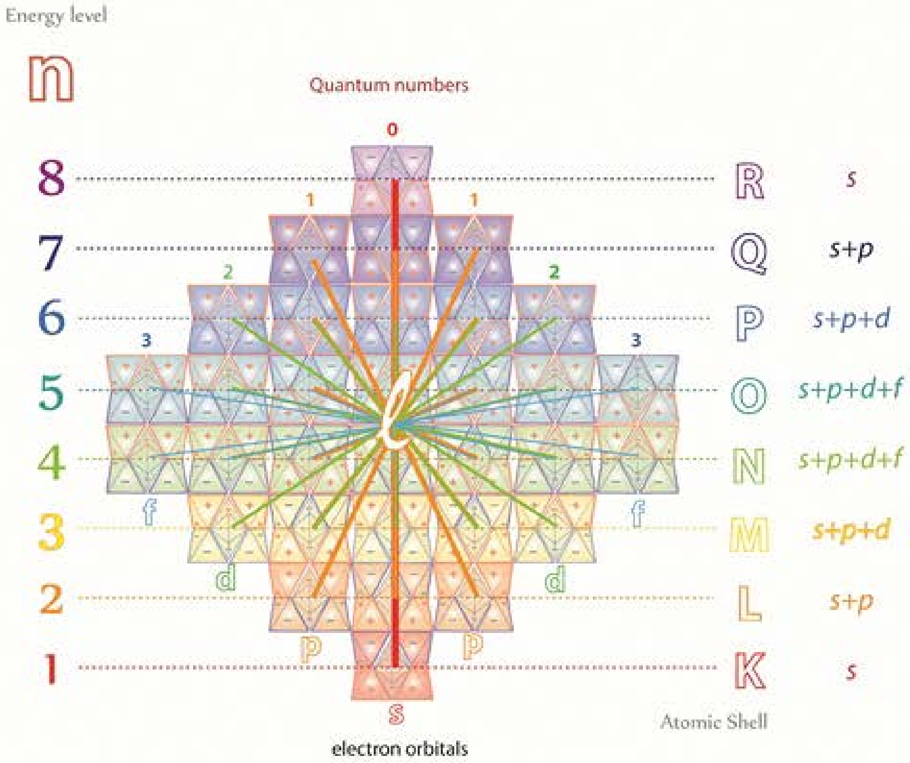
Energy level
n
Quantum numbers
0
1
1
2
2
3
3
8
7
6
5
4
3
2
1
R
Q
P
O
N
M
L
K
s
s+p
s+p+d
s+p+d+f
s+p+d+f
s+p+d
s+p
s
l
f
f
d
d
p
p
s
Atomic Shell
electron orbitals

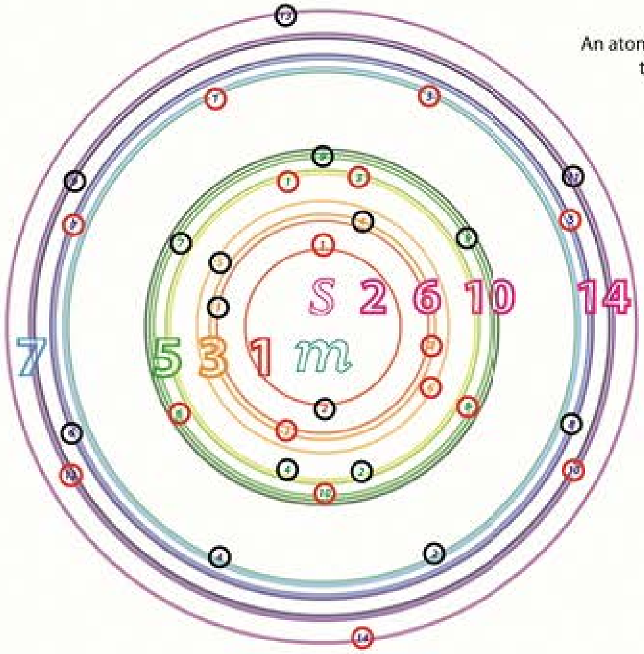
S 2 6 10 14
7 5 3 1 m

An atomic orbital is a mathematical function that describes the wave-like behavior of either one electron or a pair of electrons in an atom

Atomic orbitals are typically categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and an angular momentum vector component, respectively.

Historically used to define the pedagogical electron cloud model of an atom Tetryonics reveals the true geometry of atomic nuclei

Each orbital is defined by a different set of quantum numbers and contains a maximum of two spin opposed electrons.

Aufbau

Azmithal & Magnetic numbers

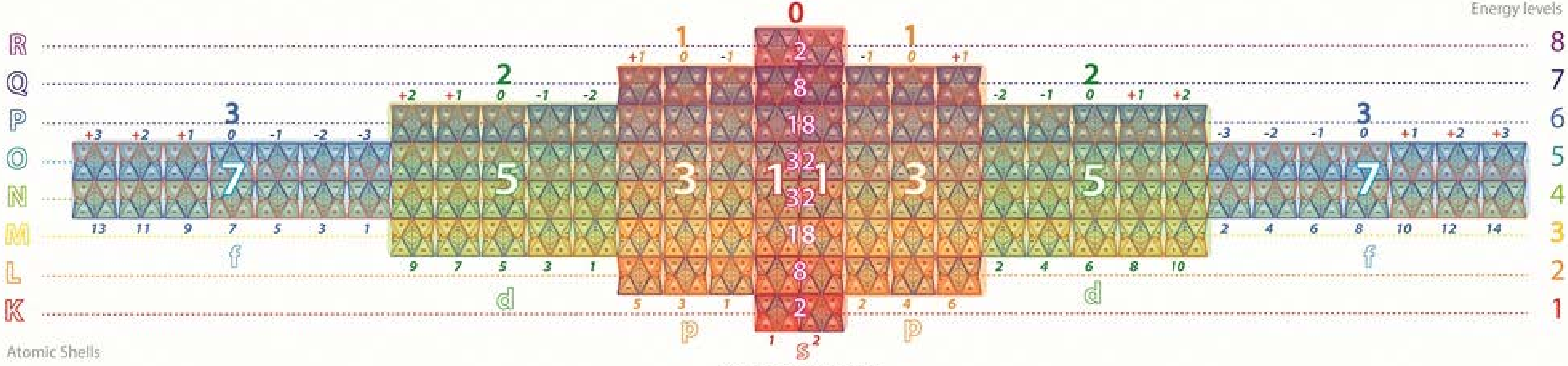
Energy levels
0
1
1
2
2
3
3
+1 0 -1
-1 0 +1
+2 +1 0 -1 -2
-2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
7 5 3 1 1 3 5 7
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
f
f
d
d
p
p
s²
Atomic Shells

Orbitals & sub-Orbitals

Electron Spins can be either up or down providing they obey the Pauli exclusion principle

s Orbital

1 Orbit (2 electrons max)

Alkali Metals & Alkali Earths

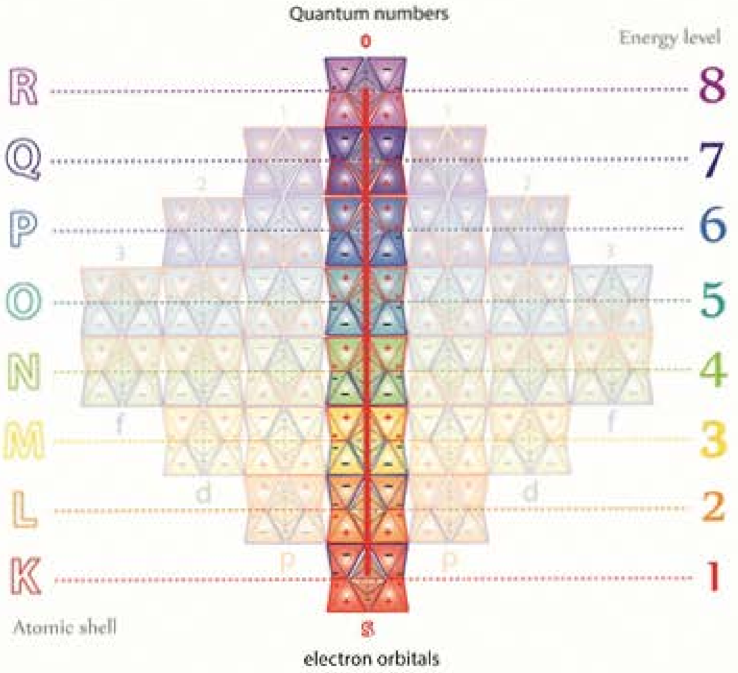
Quantum numbers
0
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f
d
p
Atomic shell
s
electron orbitals

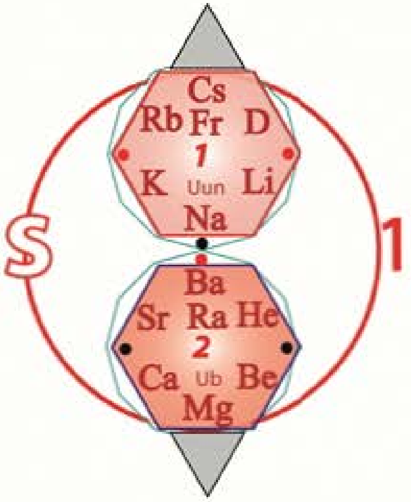
Cs
Rb
Fr
D
1
K
Uun
Li
Na
Ba
Sr
Ra
He
2
Ca
Ub
Be
Mg
s
1

n

azmithal number 0 *l*

magnetic numbers 0 *m*

spins -1/2 +1/2 *s*

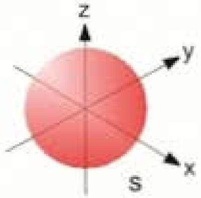

erroneous computer model of 's' electron orbitals

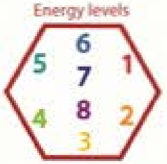
Energy levels
6
5
1
7
8
4
2
3

Copyright ABRAHAM [2008] - All rights reserved

p Orbital

3 Orbits (6 electrons max)

Non-Metals, Halogens & Nobel gases

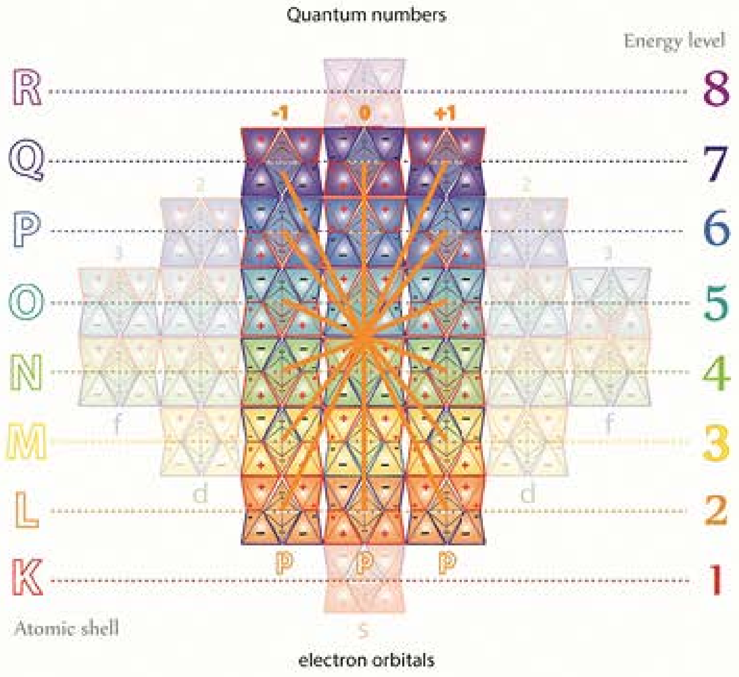
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-1
0
+1
d
f
p
p
p
s
Atomic shell
electron orbitals

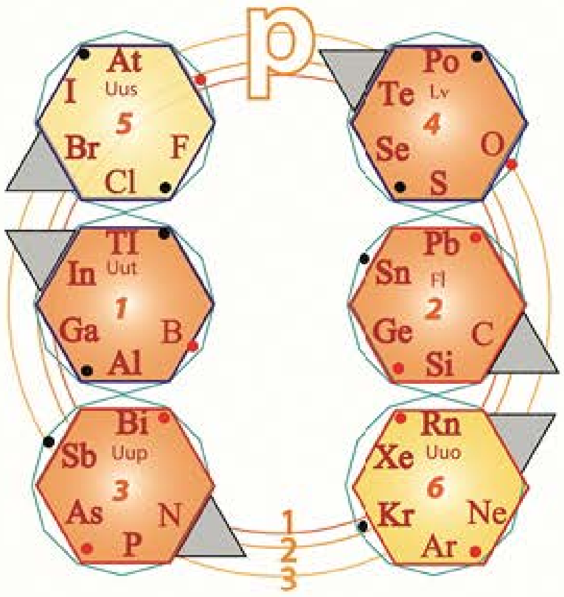
p
At
I
Uus
5
Br
F
Cl
Po
Te
Lv
4
Se
O
S
Tl
In
Uut
1
Ga
B
Al
Pb
Sn
Fl
2
Ge
C
Si
Bi
Sb
Uup
3
As
N
P
Rn
Xe
Uuo
6
Kr
Ne
Ar
1
2
3

				n
azmithal number		1		ℓ
magnetic numbers	-1	0	+1	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	S

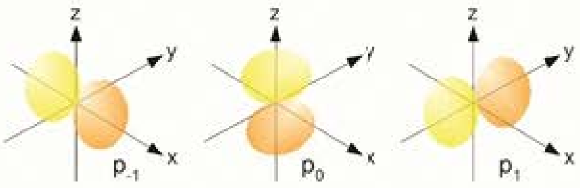
z
y
x
p_{-1}
z
y
x
p_0
z
y
x
p_1

erroneous computer model of 'p' electron orbitals

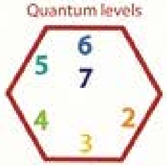
Quantum levels
6
5
7
4
2
3

d Orbital

5 Orbits (10 electrons max)

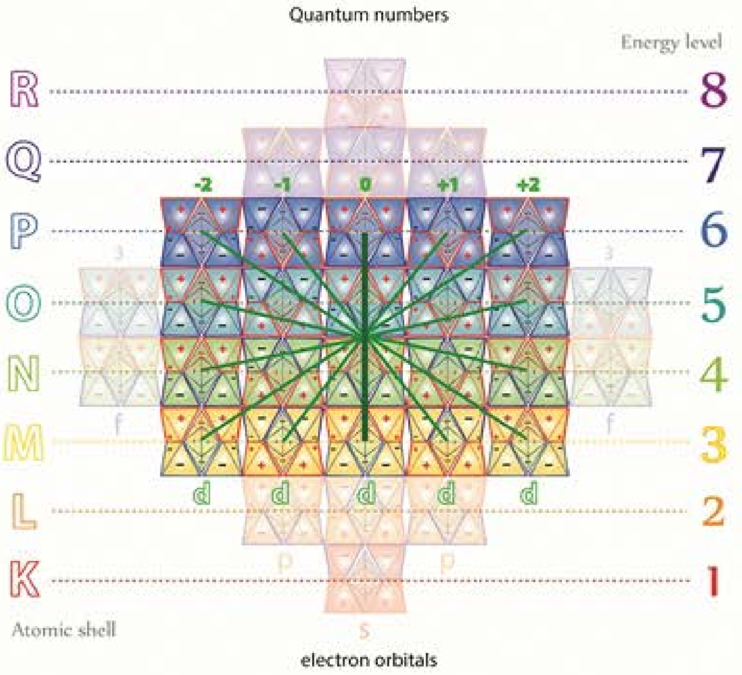
Quantum numbers
Energy level
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
-2
-1
0
+1
+2
f
f
d
d
d
d
d
p
p
s
Atomic shell
electron orbitals

						n
azimthal number			2			l
magnetic numbers	-2	-1	0	+1	+2	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

Transition & post-Transition Metals

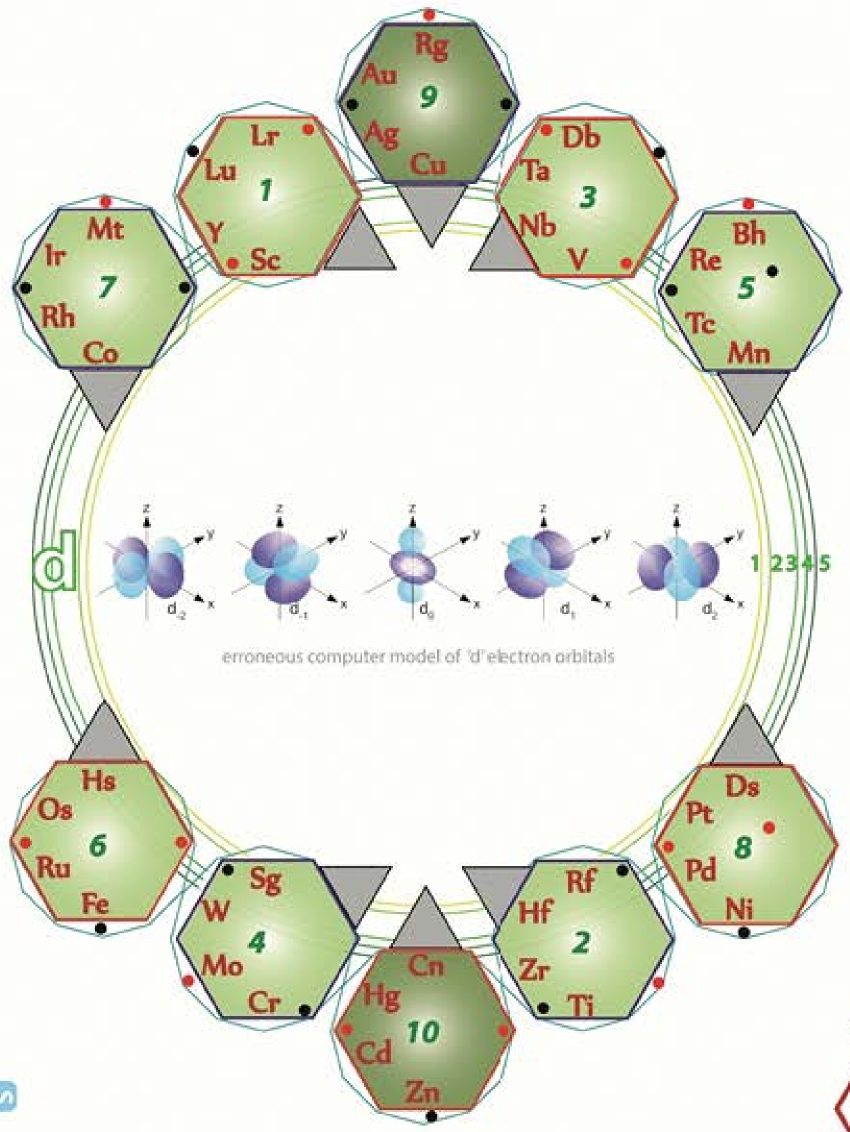
Rg
Au
9
Ag
Cu
Lr
Lu
1
Y
Sc
Db
Ta
3
Nb
V
Mt
Ir
7
Rh
Co
Bh
Re
5
Tc
Mn
d
1 2 3 4 5
erroneous computer model of 'd' electron orbitals
Hs
Os
6
Ru
Fe
Sg
W
4
Mo
Cr
Cn
Hg
10
Cd
Zn
Rf
Hf
2
Zr
Ti
Ds
Pt
8
Pd
Ni

Quantum levels
5
6
4
3

Copyright ABRAHAM [2008] - All rights reserved

f Orbital

7 Orbits (14 electrons max)

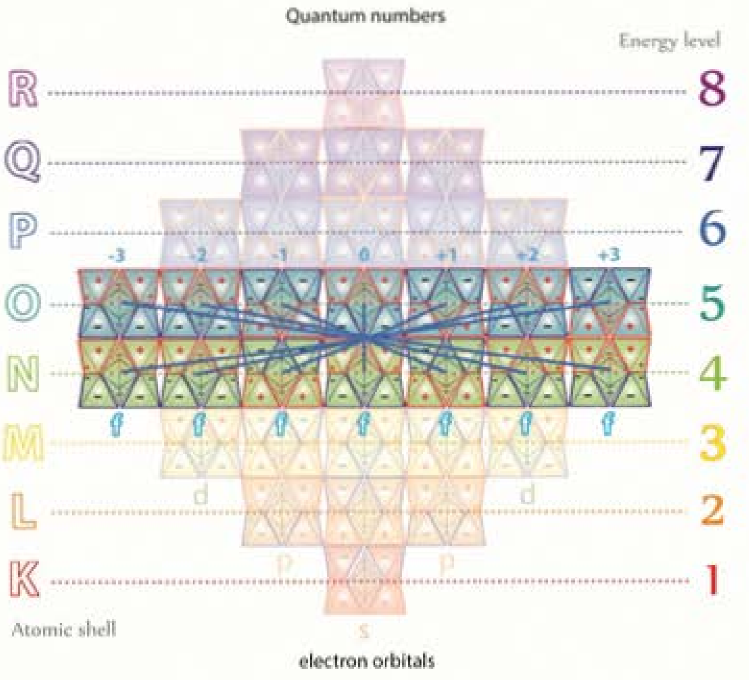
Quantum numbers
Energy level
R 8
Q 7
P 6
O 5
N 4
M 3
L 2
K 1
-3 -2 -1 0 +1 +2 +3
f f f f f f f
d d
p p
s
Atomic shell
electron orbitals

								n
azmithal number				3				ℓ
magnetic numbers	-3	-2	-1	0	+1	+2	+3	m
spins	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	-1/2 +1/2	s

Lanthanoids & Actinoids

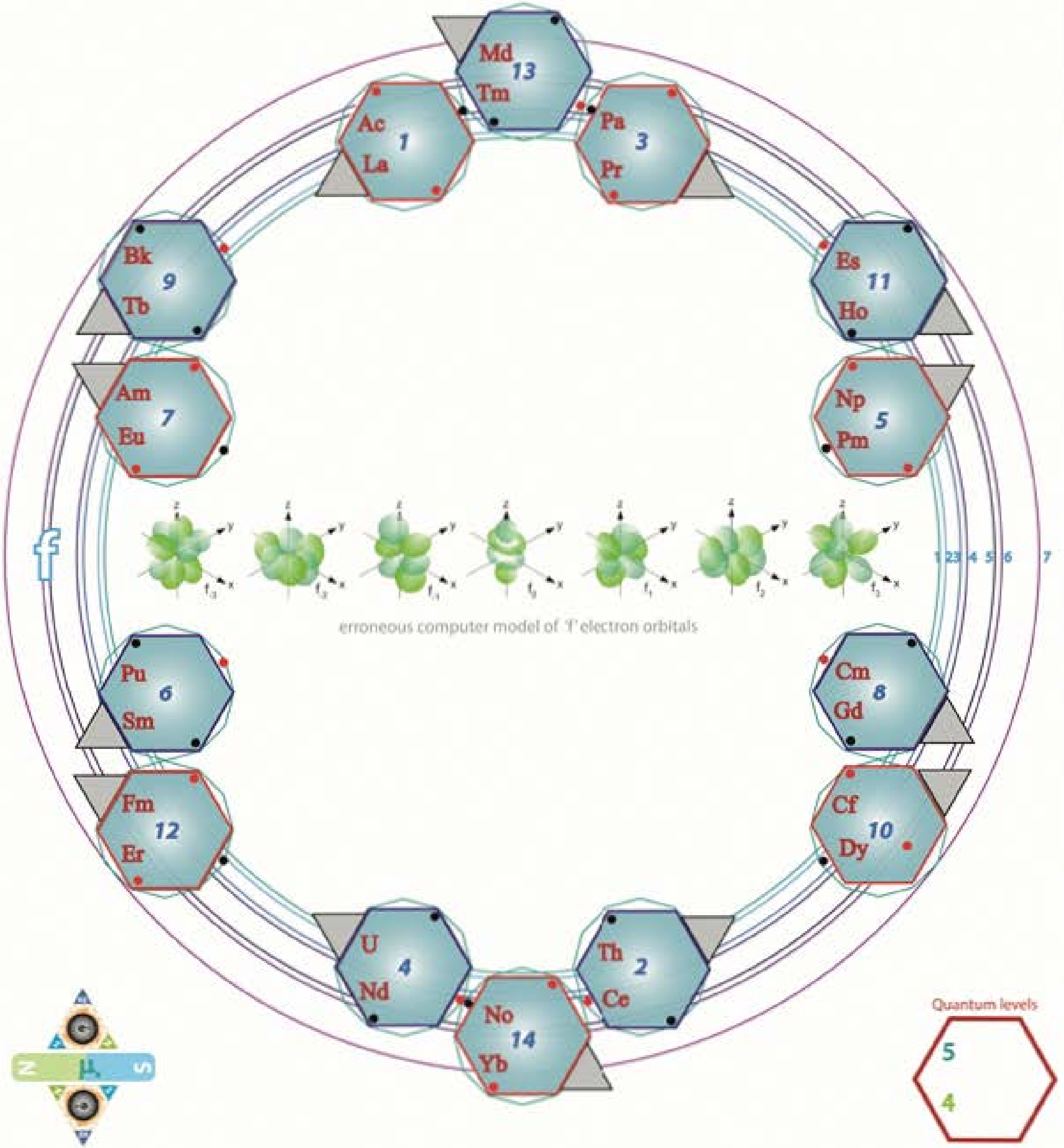
Md 13 Tm
Ac 1 La
Pa 3 Pr
Bk 9 Tb
Es 11 Ho
Am 7 Eu
Np 5 Pm
f
1 2 3 4 5 6 7
erroneous computer model of 'f' electron orbitals
Pu 6 Sm
Cm 8 Gd
Fm 12 Er
Cf 10 Dy
U 4 Nd
Th 2 Ce
No 14 Yb
N μ S
Quantum levels
5
4

Tetryonics 47.06 - 'f' Orbital

Orbital energy variations

All Elements have stable core electron configurations of s & p orbitals for each energy level as revealed through diffraction studies

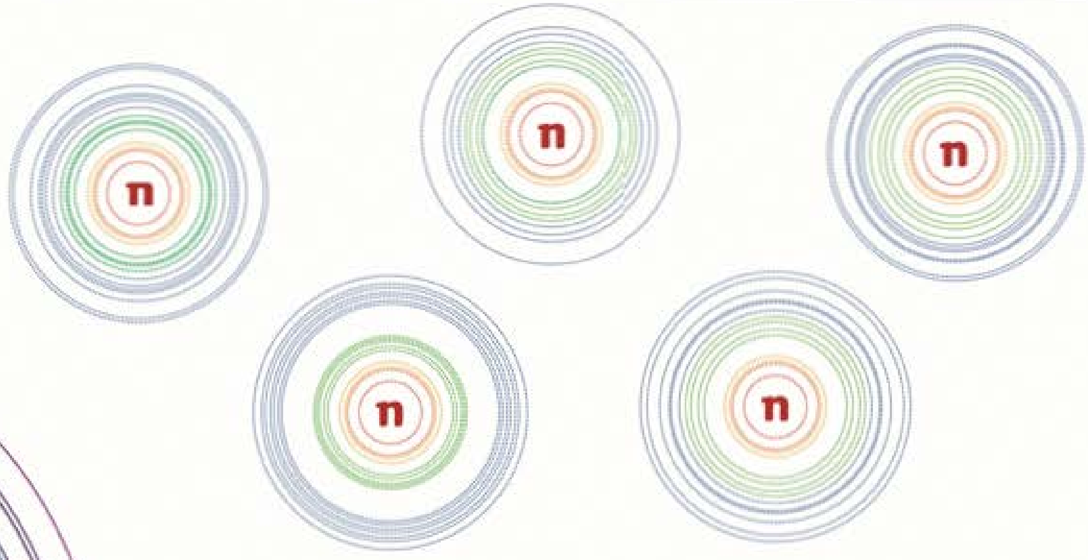

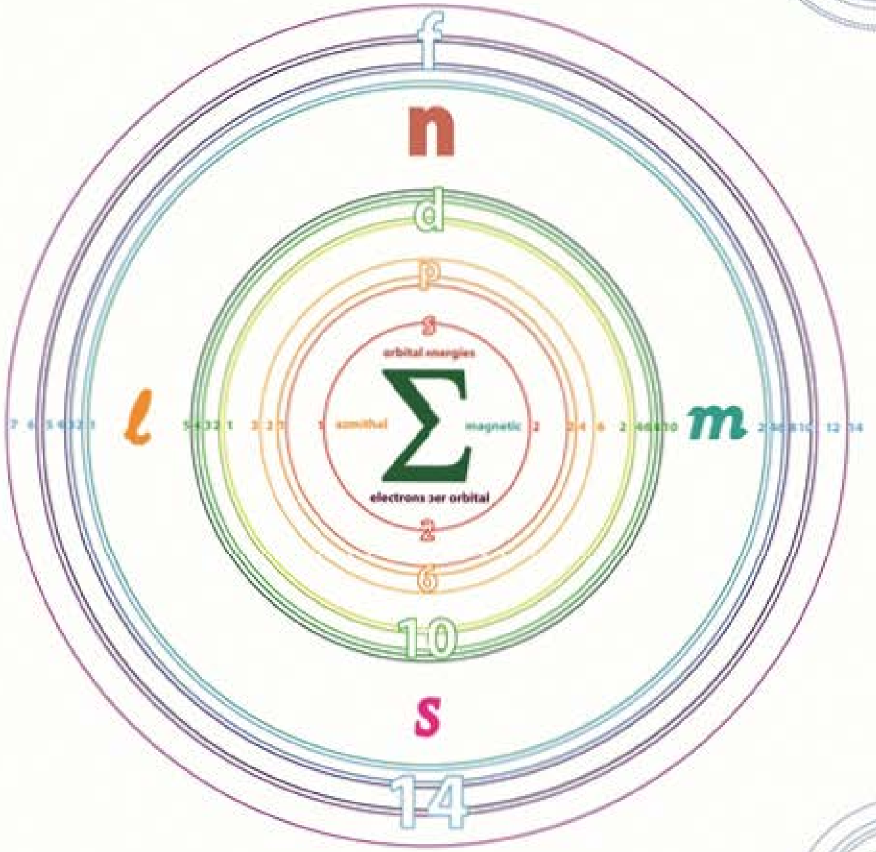
f
n
d
p
s
orbital energies
l
azimuthal
Σ
magnetic
m
electrons per orbital
2
6
10
s
14

As additional nuclei bind to form d & f orbitals they can do so by bonding in many positions, with each location producing different orbital energies for each electron that binds to nuclei in that position

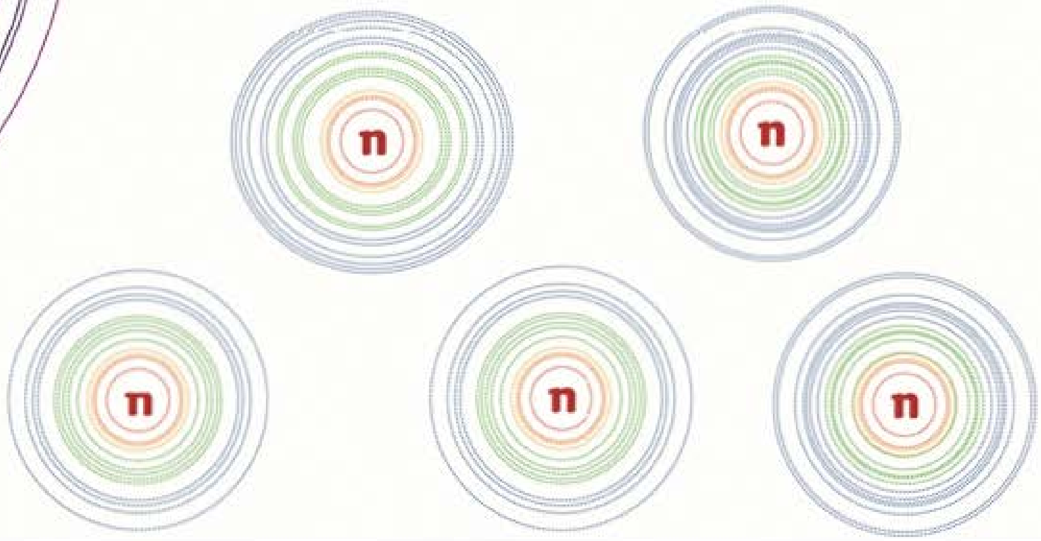

The final energy levels of each orbital is the result of the energy of the Baryons in the nuclei and the spin coupling energies of the photo-electrons bound to them

Electrons per shell

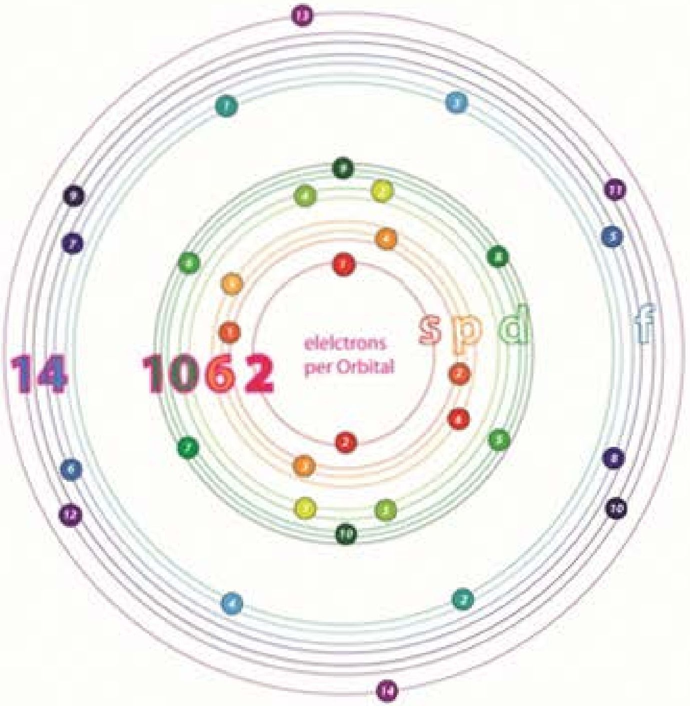
14
10
6
2
eleictrons per Orbital
s
p
d
f

Each energy shell of a periodic element can hold only a fixed number of electrons

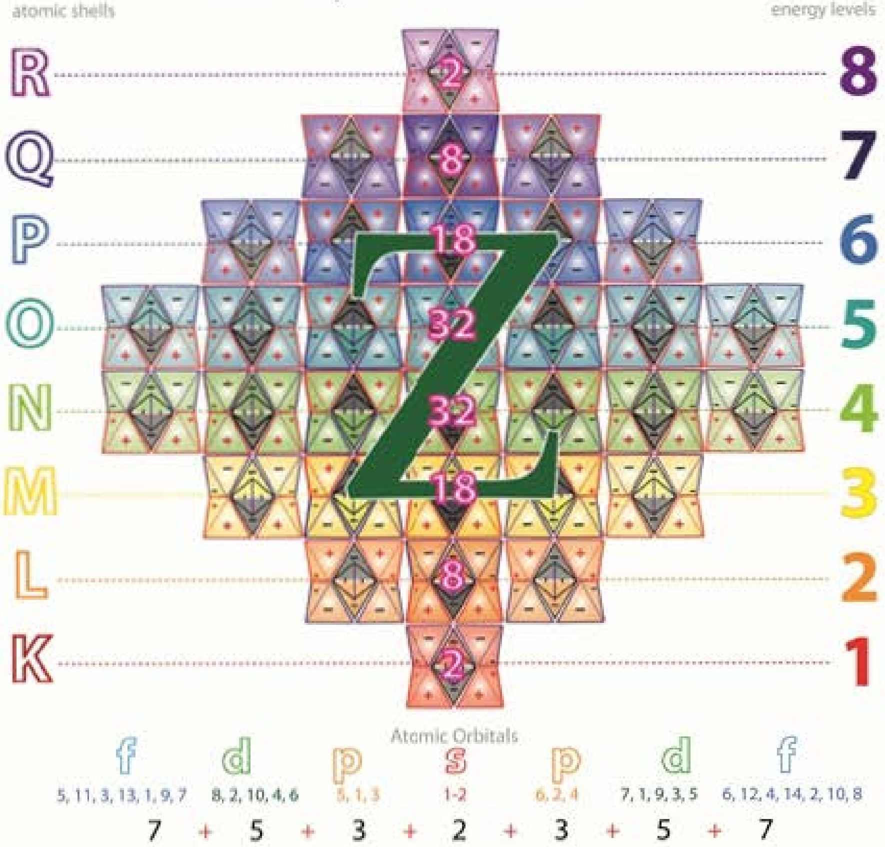
atomic shells
energy levels
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
Atomic Orbitals
f
d
p
s
p
d
f
5, 11, 3, 13, 1, 9, 7
8, 2, 10, 4, 6
5, 1, 3
1-2
6, 2, 4
7, 1, 9, 3, 5
6, 12, 4, 14, 2, 10, 8
7 + 5 + 3 + 2 + 3 + 5 + 7

Rules governing the allowed combinations of Quantum Numbers

The three quantum numbers (n, l, and m)
that describe an orbital are integers: 0, 1, 2, 3, and so on

n (1-8) Principal	The principal quantum number (n) cannot be zero.
ℓ (0-3) Azimuthal	The angular quantum number (l) can be any integer between 0 and n - 1.
mℓ (2ℓ+1) Magnetic	The magnetic quantum number (m) can be any integer between -l and +l.
m_s : $^1_2$ Spin Projection	The Spin of electrons in any nuclear sub-orbital can only be +1/2 (Spin UP) or -1.2 (Spin DOWN).

Z

1, 2, 3, 4, 5, 6, 7, 8

s, p, d, f

1	2
1, 3, 5	2, 4, 6
1, 3, 5, 7, 9	2, 4, 6, 8, 10
1, 3, 5, 7, 9, 11, 13	2, 4, 6, 8, 10, 12, 14
down up	up down

Quantum Numbers

n	4-5	3-6	2-7	1-8	2-7	3-6	4-5
ℓ	3	2	1	0	1	2	3
m	-3	-2	-1	1	+1	+2	+3
s	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2	+1/2 -1/2

Copyright ABRAHAM [2008] - All rights reserved

Using Tetryonic charged geometries for mass-ENERGY-Matter,
an electron's position and velocity CAN be modelled simultaneously
(but any attempt to measure or interact with it, will affect its component energy-momenta)

Electron Position Uncertainty

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over any spatial region of measurement, categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and a vector momentum component, respectively

Erwin Schrödinger

(12 August 1887 – 4 January 1961)

Werner Heisenberg

(5 December 1901 – 1 February 1976)

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

Quantum Mechanics is a mathematical representation of equilateral energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

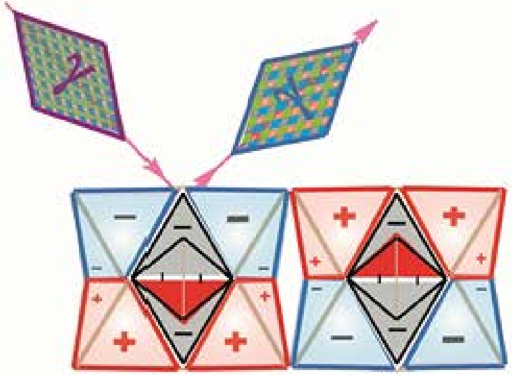

The energies of photo-electrons are determined by the Baryons they bind to & incident photons

Leptons are 12 loop quantum rotors

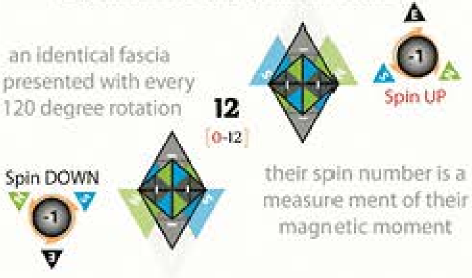

an identical fascia is presented with every 120 degree rotation

their spin number is a measure ment of their magnetic moment

The unique 12 faceted topology of leptons results in an identical EM geometry being oberved for every 120° rotation of the particle

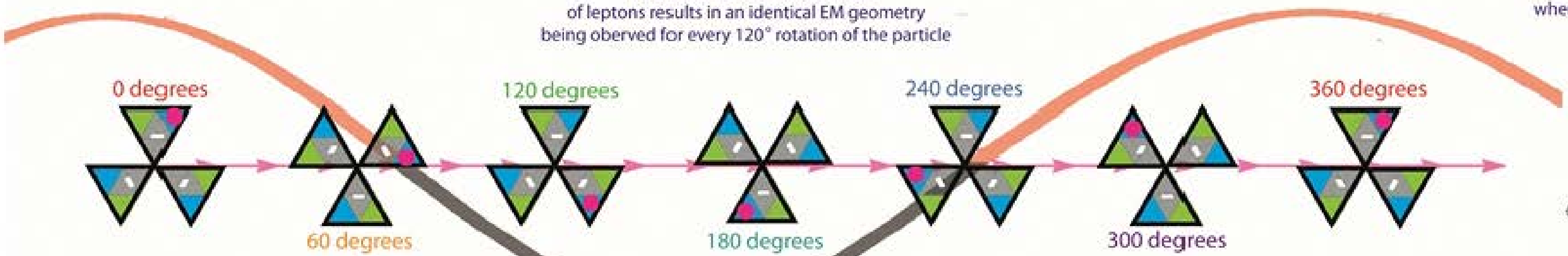

Leading to the interpretation that the Lepton disappears and re-appears when being 'observed' or measured

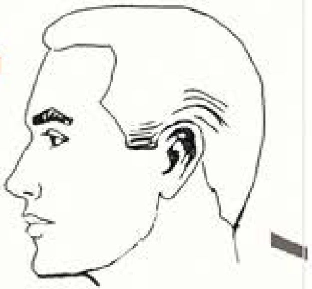

Every charged radial arm of a Lepton's Matter topology is identical to every other

Making accurate measurement and mathematical modelling of its rotational dynamics & mechanics incorrect without the correct physical topologies

Electron modelling & probability calculations

12
[0-12]
RE
S
KE

Electron positional measurements have proven to be historically difficult to accurately model due to the charged 12π rotating topologies of leptons

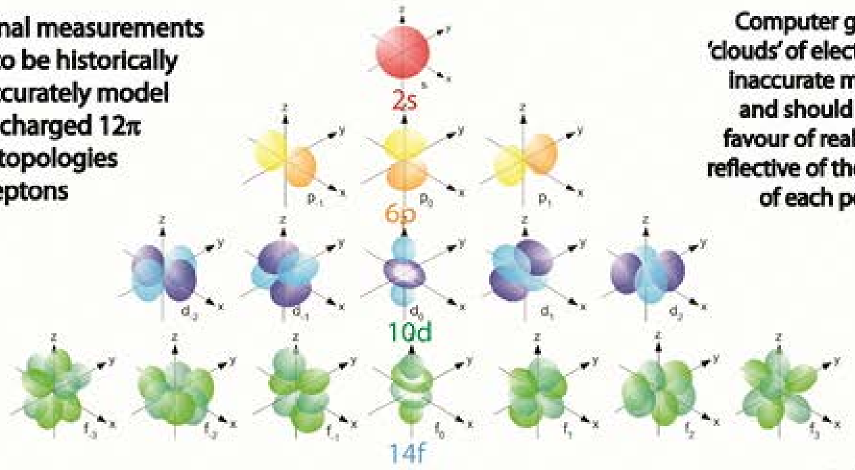

Computer generated plots of 'clouds' of electron probabilites are inaccurate mis-representations and should be abandoned in favour of realistic atomic models reflective of the charged geometry of each periodic element

v
RE
KE
P
KE
hv
Q
M
M
v²
c²

Every elemental atom can be viewed as a quantum carousel with a unique number of oscillating fans positioned around it.

Each fan has 3 blades and a fixed speed n[1-8] related to its height above ground level, AND the carousel is turning around on its axis

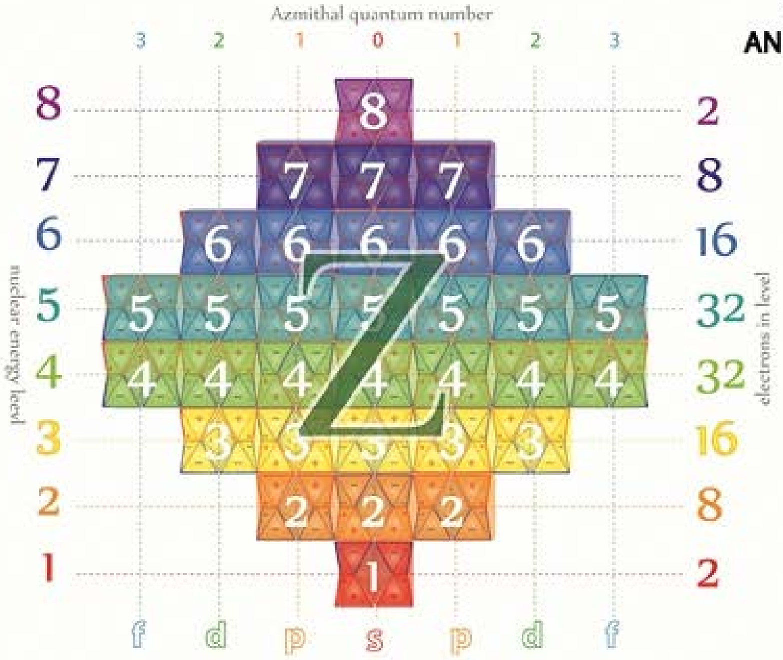

All periodic elements are made form Deuterium nuclei

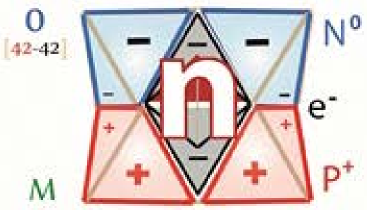

The baryonic energies of nuclei determines the energies of bound photo-electrons

The current computer generated electron probability diagrams in popular use at present can now be show to be a misrepresentative model of mathematical modelling of electron sub-orbital energies

The nuclear quantum levels [n], intrinsic quantised angular momentum [h] and orbital angular momentum [l] of each electron bound within atomic nuclei are all the direct result of the Baryonic energies of the nuclei they are bound to

Each level of the quantum carousel can contain only a limited number of fans each running at a specific speed

Imagine trying to measure (or model) the motion of any 1 quantum scale blade while the carousel rotates

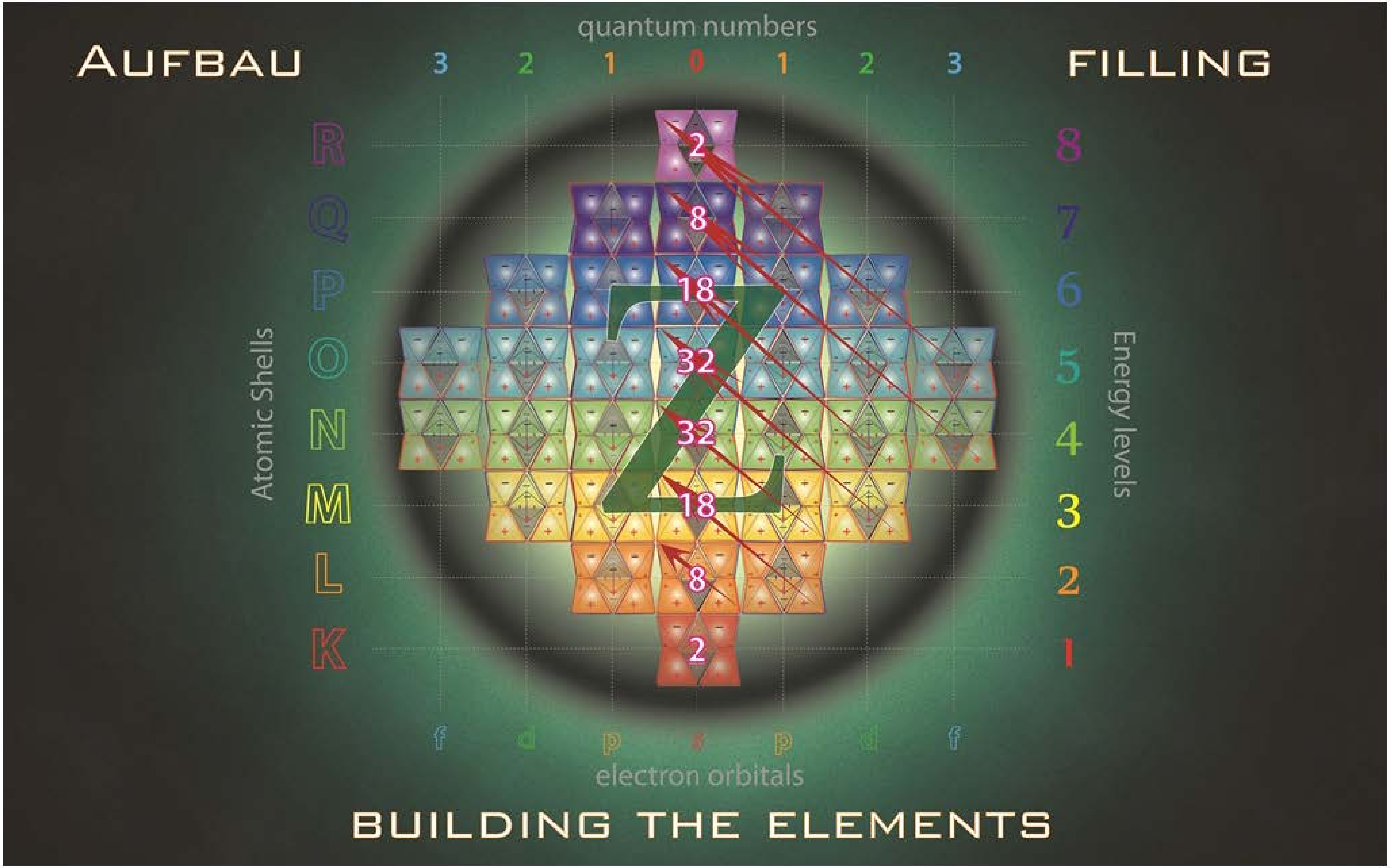
AUFBAU
FILLING
quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Atomic Shells
8
7
6
5
4
3
2
1
Energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals
BUILDING THE ELEMENTS

Tetryonics 48.01 - aufbau geometry

Quantum Topologies

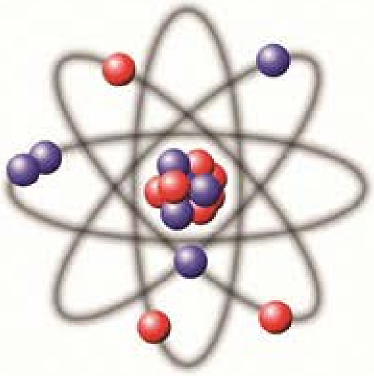

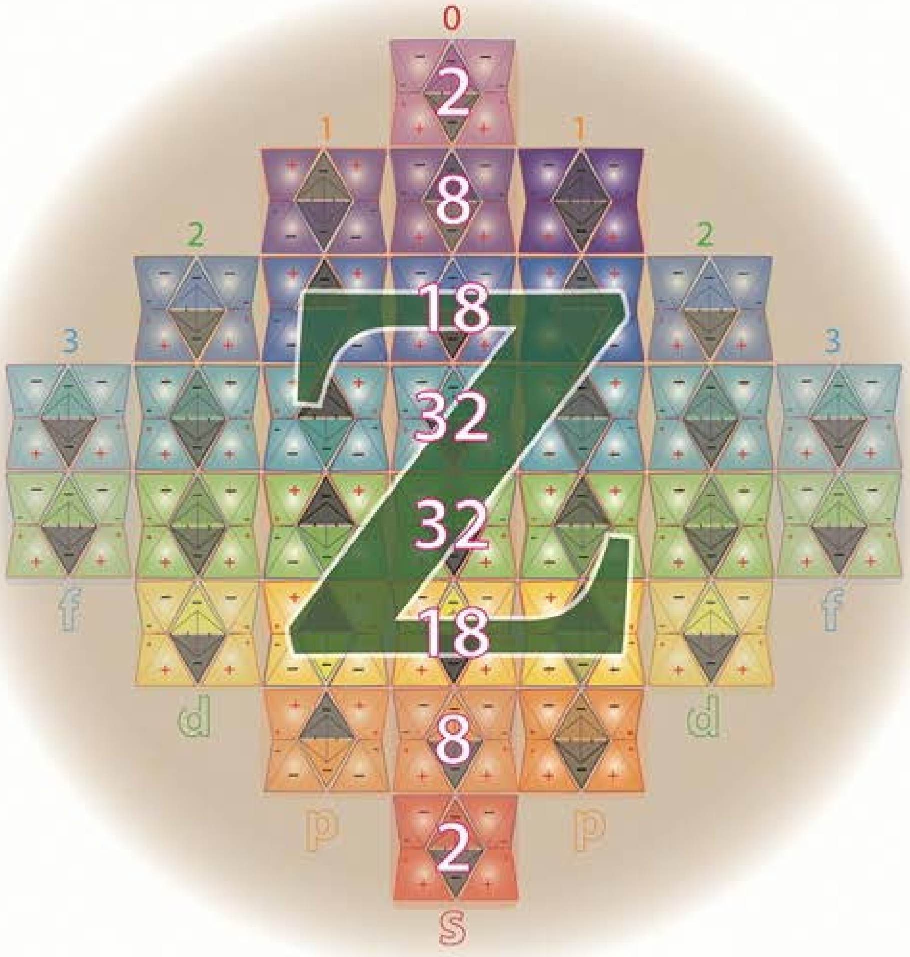
0
1
1
2
2
3
3
2
8
18
32
32
18
8
2
f
f
d
d
p
p
s
Z

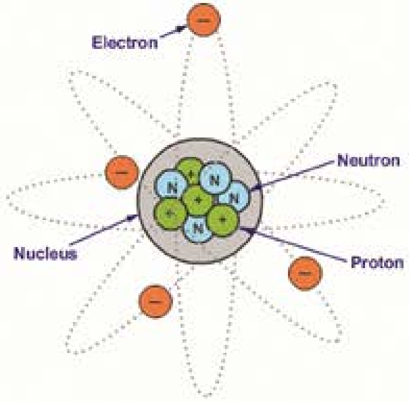
Electron
Neutron
Nucleus
Proton

Dalton Model

Quantum Model

Thomson Model

Rutherford Model

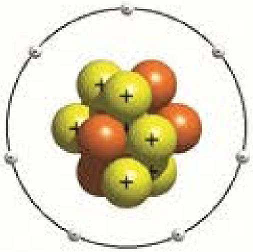

Historically viewed as a spherical object
Tetryonic charge geometry has finally revealed the
true quantum topology of all atoms

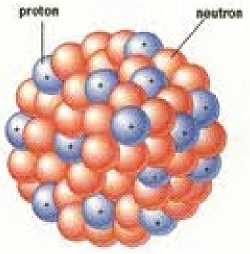
proton
neutron

Element numbers

The rule dictating how many nuclei form each Atomic shell is known as the Aufbau principle.

The physical and chemical properties of elements is determined by the atomic structure.
The atomic structure is, in turn, determined by the electrons and
which shells, subshells and orbitals they reside in.

The number of possible nuclei in each Quantum level follows aufbau principle 'numbers' which can be determined using the following summation formula

The maximum periodic elemental number is 120

$$Z = \overset{1\,2\,3\,4}{\underset{4\,3\,2\,1}{\sum}} 2(n^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

Number of nuclei per level	Element #
2	120 119
8	118 111
18	110 93
32	92 61
32	60 29
18	28 11
8	10 3
2	2 1

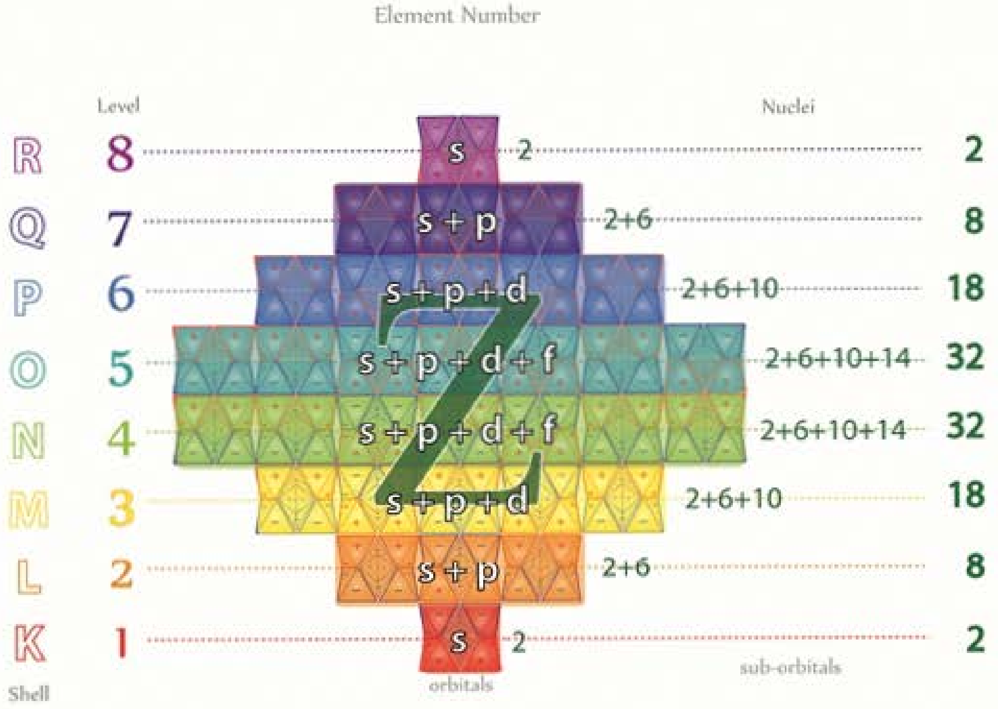
Element Number
Level
Nuclei
R 8 s 2 2
Q 7 s + p 2+6 8
P 6 s + p + d 2+6+10 18
O 5 s + p + d + f 2+6+10+14 32
N 4 s + p + d + f 2+6+10+14 32
M 3 s + p + d 2+6+10 18
L 2 s + p 2+6 8
K 1 s 2 2
Z
Shell
orbitals
sub-orbitals

Deuterium is the building block of all elements

The number of nuclei per quantum level is reflective of photonic energy levels and provides the foundational geometry for all of the periodic elements

Each element has equal numbers of Protons, electron & Neutrons with their stored mass-energies making up the molar masses of elements not excess neutron as currently modelled

Aufbau Principle

(Nuclei number and position)

1 Sub-Orbit (2 electrons max) ○ s

3 Sub-Orbits (6 electrons max) ○ p

○ d *5 Sub-Orbits (10 electrons max)*

○ f *7 Sub-Orbits (14 electrons max)*

atomic shells			energy level
R	2	2	8
Q	2 + 6	8	7
P	2 + 6 + 10	18	6
O	2 + 6 + 10 + 14	32	5
N	2 + 6 + 10 + 14	32	4
M	2 + 6 + 10	18	3
L	2 + 6	8	2
K	2	2	1

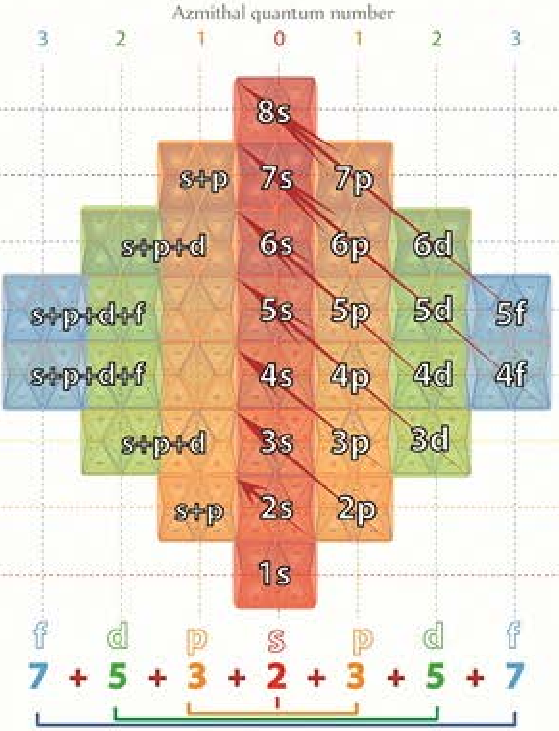

2 6 10 14

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Quantum number	per level
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	*s1*	*s2*	*p1*	*p2*	*p3*	*p4*	*p5*	*p6*	*d1*	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*	*d8*	*d9*	*d10*	*f1*	*f2*	*f3*	*f4*	*f5*	*f6*	*f7*	*f8*	*f9*	*f10*	*f11*	*f12*	*f13*	*f14*		

sub-orbitals

A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.

The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements

Copyright ABRAHAM [2008] - All rights reserved

Wolfgang Pauli

(25 April 1900 – 15 December 1958)

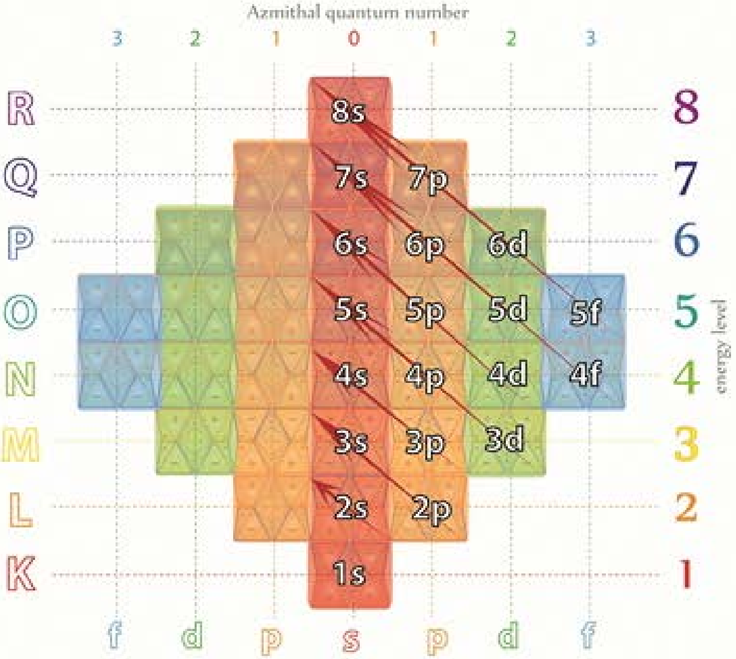
Azmithal quantum number
3 2 1 0 1 2 3
R Q P O N M L K
8s
7s 7p
6s 6p 6d
5s 5p 5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
8 7 6 5 4 3 2 1
energy level
f d p s p d f

Friedrich Hermann Hund

(4 February 1896 - 31 March 1997)

The orbitals of lower energy are filled in first with the electrons and only then are the higher energy orbitals filled.

aufbau electron orbital filling

Orbitals of equal energy are each occupied by one electron before any orbital is occupied by a second electron, and all electrons in singly occupied orbitals must have the same spin state

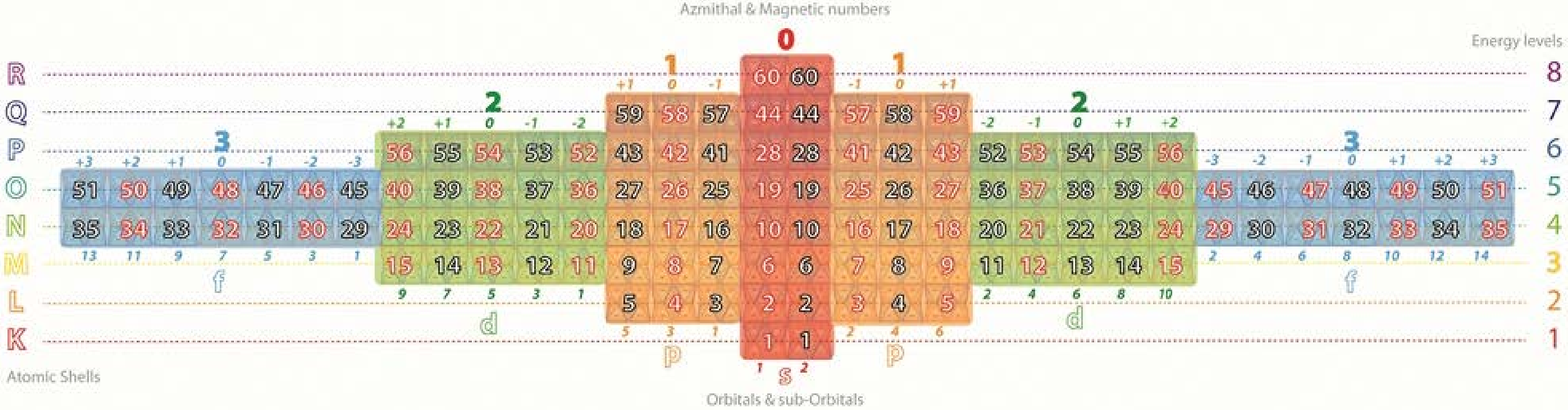

Azmithal & Magnetic numbers

Atomic Shells	3							2					1			0		1			2					3							Energy levels
	+3	+2	+1	0	-1	-2	-3	+2	+1	0	-1	-2	+1	0	-1			-1	0	+1	-2	-1	0	+1	+2	-3	-2	-1	0	+1	+2	+3	
R																60	60																8
Q													59	58	57	44	44	57	58	59													7
P								56	55	54	53	52	43	42	41	28	28	41	42	43	52	53	54	55	56								6
O	51	50	49	48	47	46	45	40	39	38	37	36	27	26	25	19	19	25	26	27	36	37	38	39	40	45	46	47	48	49	50	51	5
N	35	34	33	32	31	30	29	24	23	22	21	20	18	17	16	10	10	16	17	18	20	21	22	23	24	29	30	31	32	33	34	35	4
M								15	14	13	12	11	9	8	7	6	6	7	8	9	11	12	13	14	15								3
L													5	4	3	2	2	3	4	5													2
K																1	1																1
	13	11	9	7	5	3	1	9	7	5	3	1	5	3	1	1	2	2	4	6	2	4	6	8	10	2	4	6	8	10	12	14	
	f							d					p			s		p			d					f							

Orbitals & sub-Orbitals

Tetryonics 48.05 - electron orbital filling

Aufbau construction

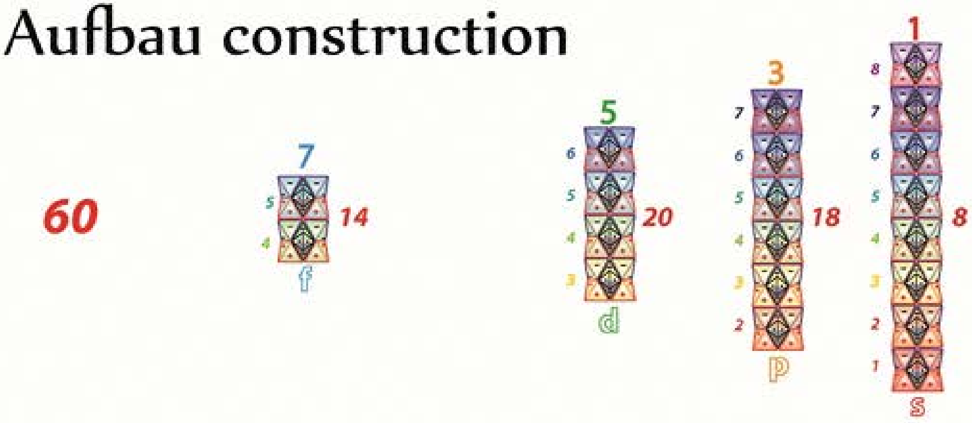
60
7
5
4
14
f
5
6
5
4
3
20
d
3
7
6
5
4
3
2
18
p
1
8
7
6
5
4
3
2
1
8
s

Proton
base
Hemisphere

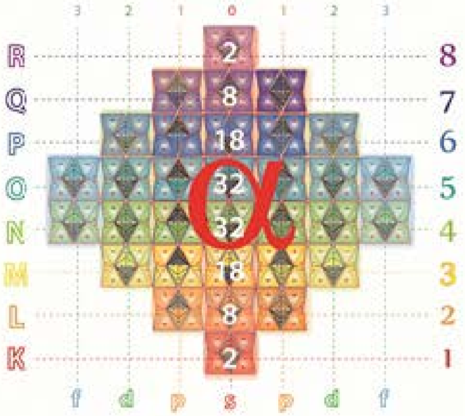
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
α
8
7
6
5
4
3
2
1
f d p s p d f

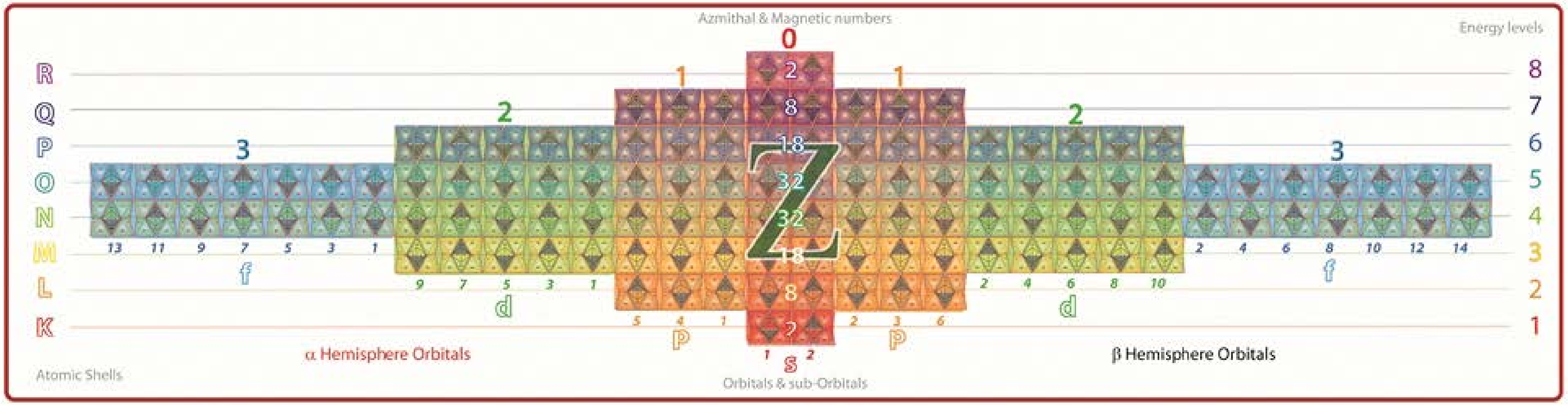
Azmithal & Magnetic numbers
Energy levels
0
1
1
2
2
3
3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
13 11 9 7 5 3 1
f
9 7 5 3 1
d
5 4 1
p
1 2
s
2 3 6
p
2 4 6 8 10
d
2 4 6 8 10 12 14
f
α Hemisphere Orbitals
β Hemisphere Orbitals
Atomic Shells
Orbitals & sub-Orbitals

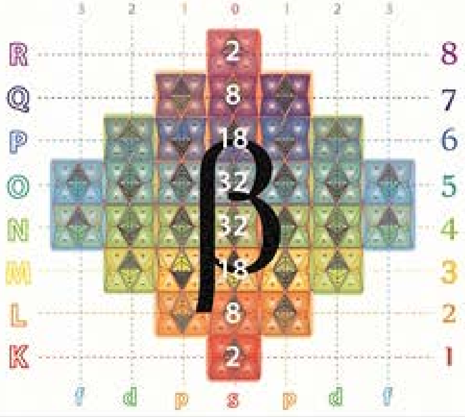
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
2
8
18
32
32
18
8
2
β
8
7
6
5
4
3
2
1
f d p s p d f

Neutron
base
Hemisphere

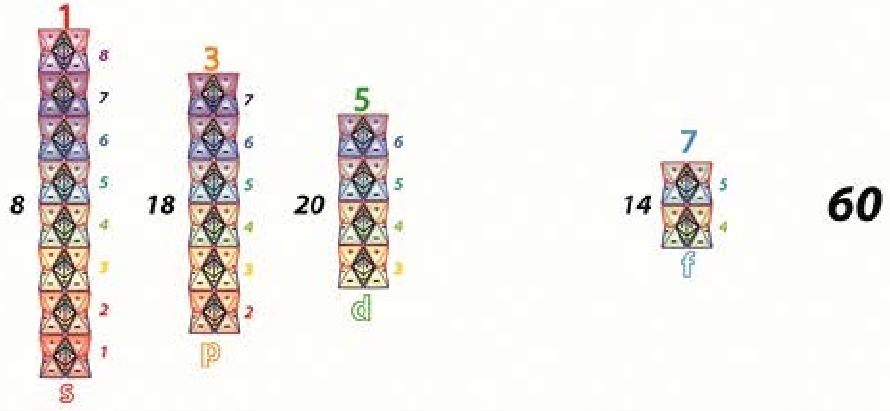
1
8
7
6
5
4
3
2
1
8
s
3
7
6
5
4
3
2
18
p
5
6
5
4
3
20
d
7
5
4
14
f
60

Atomic Weights

[Molar mass-energies]

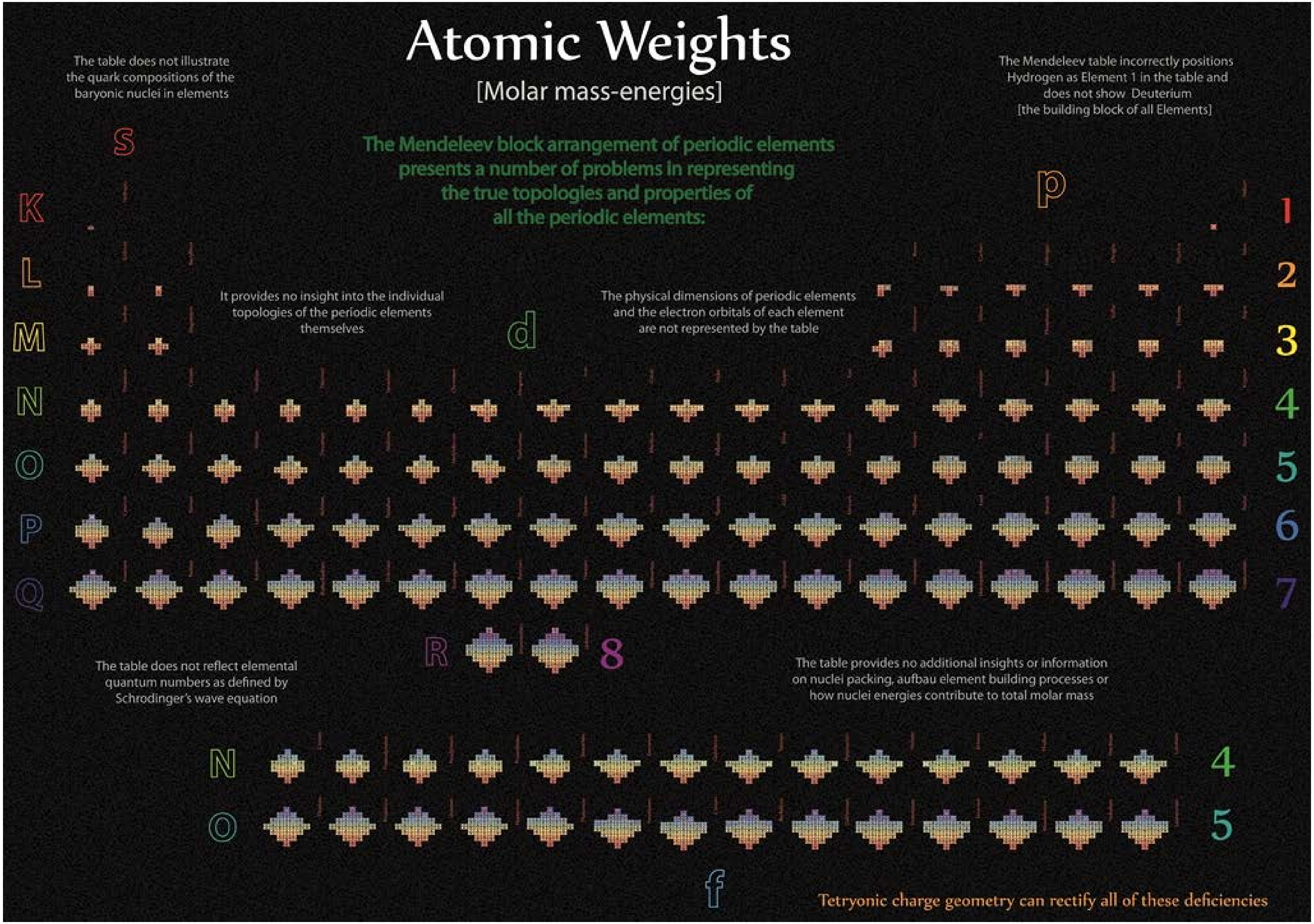

The table does not illustrate the quark compositions of the baryonic nuclei in elements

The Mendeleev table incorrectly positions Hydrogen as Element 1 in the table and does not show Deuterium [the building block of all Elements]

The Mendeleev block arrangement of periodic elements presents a number of problems in representing the true topologies and properties of all the periodic elements:

It provides no insight into the individual topologies of the periodic elements themselves

The physical dimensions of periodic elements and the electron orbitals of each element are not represented by the table

The table does not reflect elemental quantum numbers as defined by Schrodinger's wave equation

The table provides no additional insights or information on nuclei packing, aufbau element building processes or how nuclei energies contribute to total molar mass

Tetryonic charge geometry can rectify all of these deficiencies

Tetryonics 49.01 - Atomic Weight

E

The quantum level mass-energies of Baryons determines the kinetic energies of electrons

1n - 1 e19v

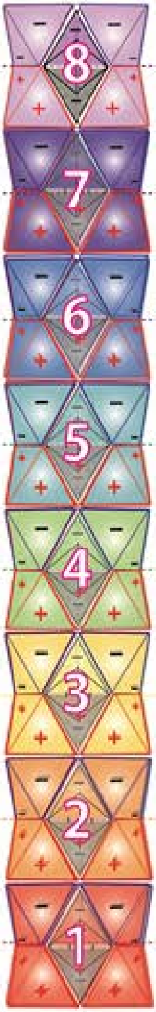

Baryon mass-energies	n		electron KEM field hv^2
73,728	n32	v8	3.268969693 e15
69,192	n31	v7	2.502804921 e15
64,800	n30	v6	1.838795452 e15
60,552	n29	v5	1.276941286 e15
56,448	n28	v4	8.172424232 e14
52,488	n27	v3	4.596988631 e14
48,672	n26	v2	2.043106058 e14
45,000	n25	v1	5.107765145 e13
Compton frequencies 930 MeV	**1.2e20 electron rest Matter is velocity invariant** 496 keV		**spectral frequencies** 13.6 eV

Atomic nuclei mass-energies

Each element's weight [mass-Matter in a gravitational field] is the result of the total quanta comprising that element

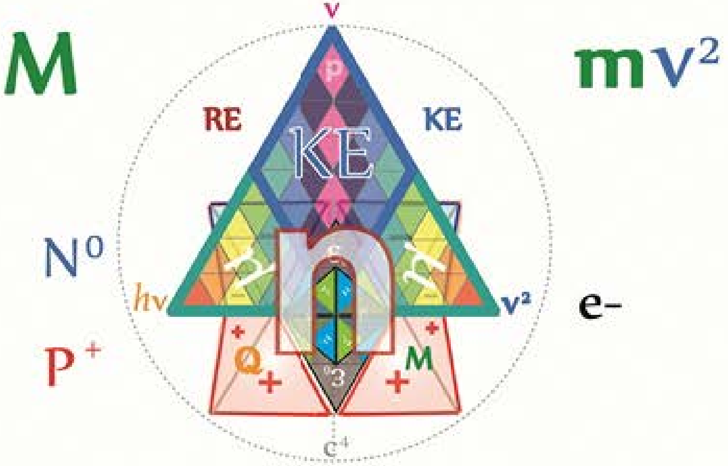

The nuclei forming each atomic shell have specific mass-energy quanta

$$\overset{8}{\underset{1}{n}}\Bigg[\underbrace{\overset{\text{Baryon rest masses}}{[72(n)^2]}+\overset{\text{lepton rest mass}}{[12e19]}+\overset{\text{KEM}}{[m_e v^2]}}_{\text{Deuterium mass-energy per shell}}\Bigg]$$

Despite having differing mass-energies each Deuterium nuclei has the same velocity invariant Matter geometry [84π]

spin orbital coupling in synchronous quantum convertors

Electrons act as quantum scale rotating armatures in atomic nuclei and can only have specific energies reflective of the electron orbital energy level of the Baryons in which they are found

They acheive these energy levels by absorbing or emitting photons to acheive the specific angular momentum required

Copyright ABRAHAM [2008] - All rights reserved

Hydrogen Ionisation Energies [/n^2]

13.6eV
Free electron

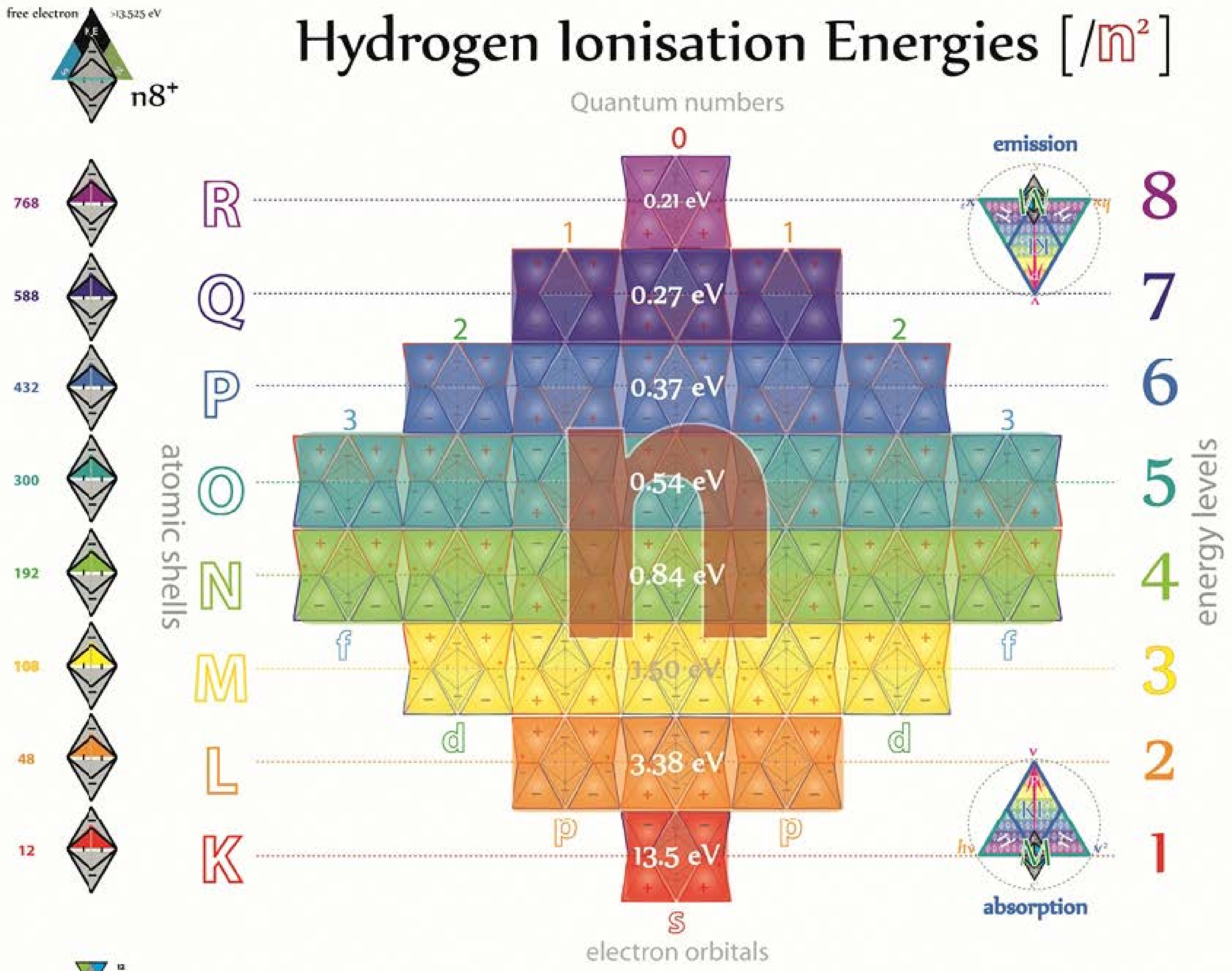
free electron
>13.525 eV
K.E
n8+
Quantum numbers
0
1
1
2
2
3
3
f
f
d
d
p
p
s
electron orbitals
atomic shells
energy levels
emission
absorption
hv
n0

768	R	0.21 eV	8	5.107765145 e13
588	Q	0.27 eV	7	6.671366720 e13
432	P	0.37 eV	6	9.078047137 e13
300	O	0.54 eV	5	1.307587877 e14
192	N	0.84 eV	4	2.043106058 e14
108	M	1.50 eV	3	3.632188548 e14
48	L	3.38 eV	2	8.172424234 e14
12	K	13.5 eV	1	3.268969693 e15

Planck quanta

Spectral line series frequencies
[Hydrogen]

accelerating electrons $Mv^2 = KEM = hv^2$ produce spectral lines

Redefining Atomic weights

Atomic weight (symbol: Ar) is a dimensionless physical quantity,
the ratio of the average mass of atoms of an element (from a given source) to
1/12 of the mass of an atom of carbon-12 (known as the unified atomic mass unit)

Carbon 12

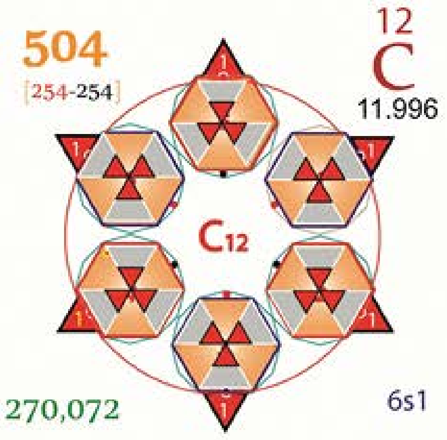

6 Protons [24-12]
6 Neutrons [18-18] n1
6 electrons [0-12]

The 'unified atomic mass unit' currently in use
is known to be inaccurate and must be corrected
in order to bring clarity & increased accuracy
to the atomic weights of all elements

$$\boxed{A_r = 22{,}512 \quad \text{Hydrogen}}$$

Defining Hydrogen as having an exact
atomic Planck mass of 22512n quanta
provides uniformity with Tetryonics

1/12 C_{12}

$A_r = 22{,}506$

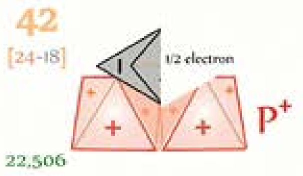

1.660096209e-27 kg

1 Proton [24-12]
.5 electron 0-6 n1

Unified atomic Matter unit

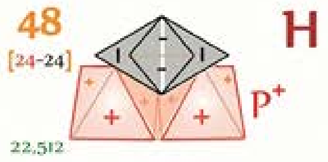

1.660538783e-27 kg

1 Proton [24-12]
1 electron [0-12] n1

Deuterium is the building block of all elements in the period table

$$\boxed{A_r = 45{,}012 \quad \text{Deuterium}}$$

Defining Deuterium as having an exact
atomic Planck mass of 45012n quanta
reflects the true charged geometries
of all Elements & their topologies

D

1/6 C

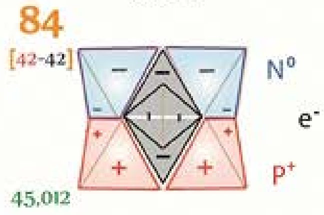

3.320192418e-27 kg

1 Proton [24-12]
1 Neutron [18-18] n1
1 electron [0-12]

Copyright ABRAHAM [2008] - All rights reserved

Planck mass-energy units

NAu = 0.001 kg

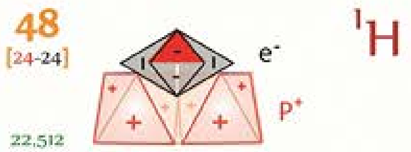

1.660538841 e-27 kg

One Da is approximately equal to the mass of one proton or one neutron

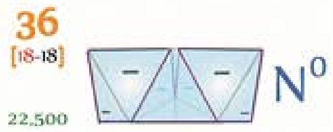

0.499866702 H

1.659653693 e-27 kg

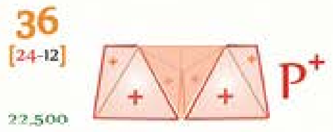

0.499866702 H

Deuterium is the building block of all elements

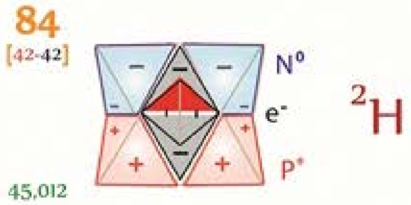

3.320192534 e-27 kg

The unified atomic mass unit (symbol: u) or Dalton (symbol: Da)
is a unit that is used for indicating mass on an atomic or molecular scale

270,072

1/12 the mass of a C12 graphene atom
at rest in its electronic ground state

1.660538782(83)×10−27 kg

22,506

is an inaccurate means of determining the
exact rest mass of a Hydrogen atom

22,512

Carbon 12 has 270,072n planck quanta
(270,072 / 12 = 22,506)

Hydrogen has a mass of 22,512n (22,500+12) requiring all mass to be calculated directly using the Planck mass-energy quantum (.001kg / N_A / 22,512) & Tetryonic charge geometries

Using Tetryonic theory to define
n Planck mass = $7.376238634 \times 10^{-32}$ Kg
(see Tetryonics QM 15.04)

exact atomic rest masses for all
particles, elements and compounds
can be determined directly from atomic theory

N_A = 6.02214179 e23

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is "mol".

NA(12u) = 0.012 kg

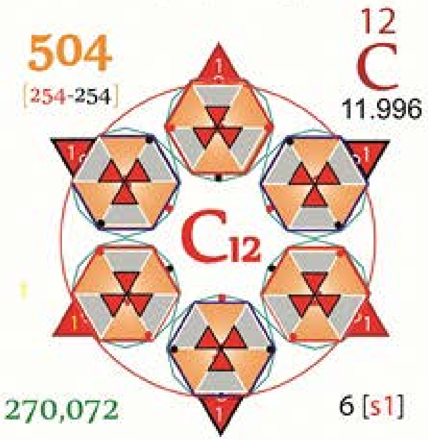

6 Protons	[24-12]	n1
6 Neutrons	[18-18]	
6 electrons	[0-12]	

Carbon has a number of differing atomic configurations (allotropes)

6 Protons	[24-12]	n1-2
6 Neutrons	[18-18]	
6 electrons	[0-12]	

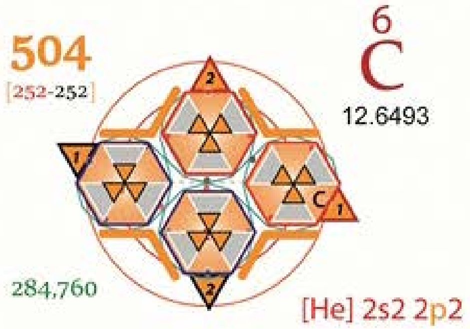

Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies

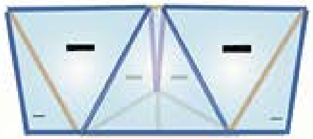

1.659653693 e-27 kg

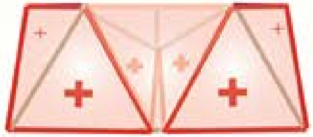

8.851486361 e-31 kg

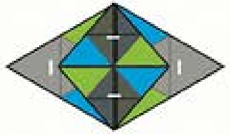

2.41109611 e-35 kg

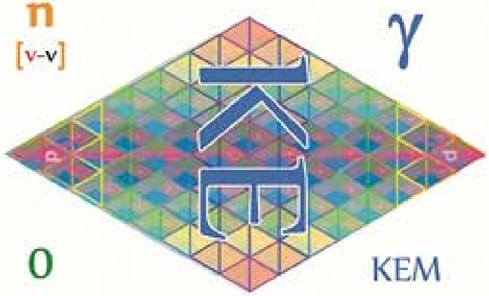

Baryons have 2,25e23 Planck quanta comprising their rest Matter topologies

[930.974 MeV]

Δ 1875 x

[496.5 keV]

Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies

[496.5 keV]

Δ 36,711 x

[13.6 eV]

Photons are planar geometries [Matter-less] (purely Kinetic mass-Energy and momenta)

The Lyman alpha spectral line mass-energy contribution to the mass of a Deuterium nucleus is negligible

Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements

D^+

[1.862 GeV]

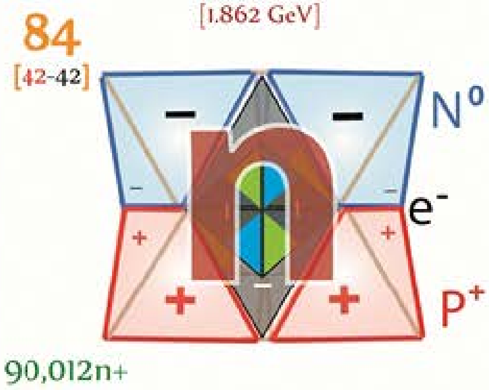

Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]

Photons are 2π charge mass-energy geometries

Ionisation energies

Fractional quantum differentials

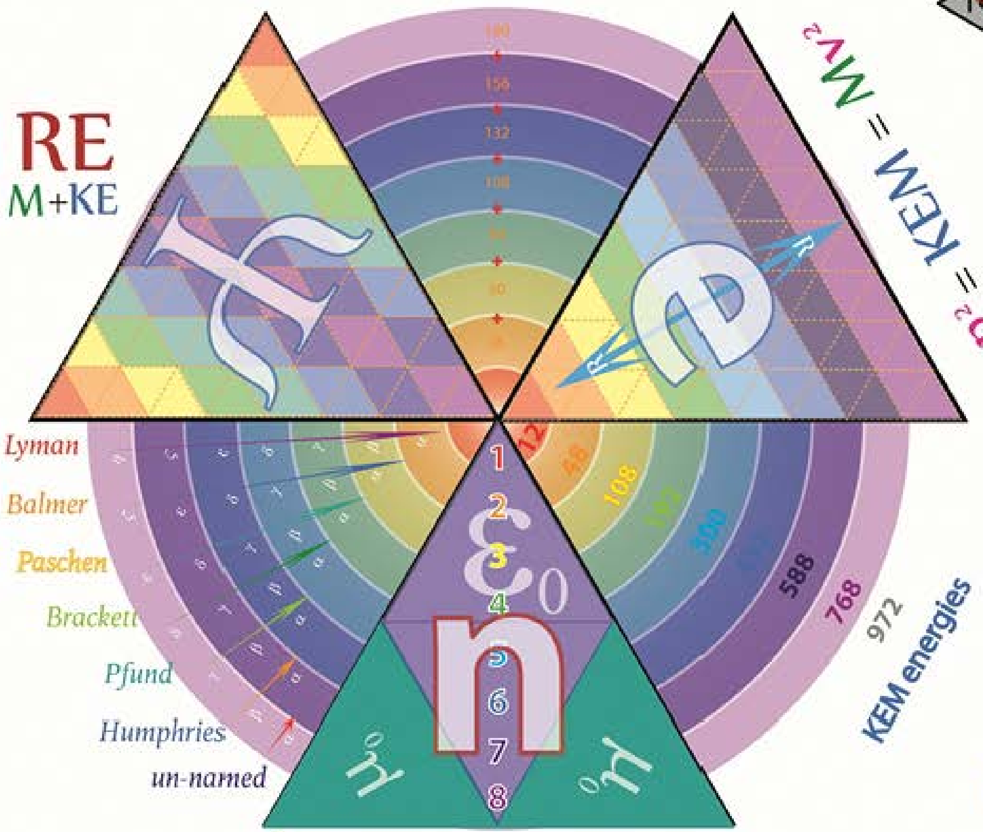
RE
M+KE
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
un-named
Mv² = KEM = p²
1
2
3
4
5
6
7
8
ε0
n
12
48
108
192
300
432
588
768
972
KEM energies

$hf = \Delta M v = \Delta p$

spectral lines are produced by accelerating electrons

Note: this is an Illustrative schema for modelling KEM field energies
All KEM fields possess the same physical spatial geometry
in radial-time defined spatial co-ordinate systems

c^2 c^4

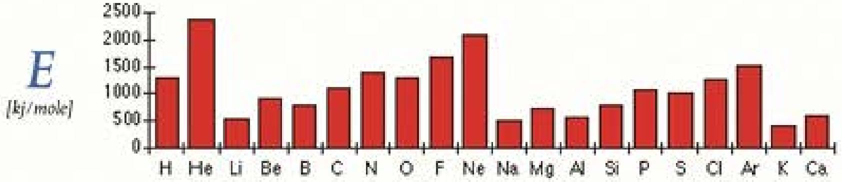

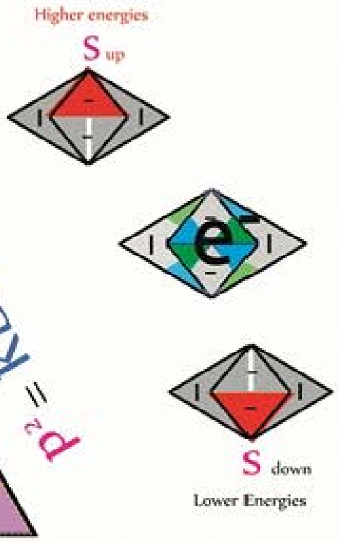
Higher energies
S up
e
S down
Lower Energies

Photo-electrons absorb/emit spectral energies

$$E = eV = \frac{1}{4\pi\epsilon_0}\frac{ne^2}{a}$$

photo-electronic energy momenta

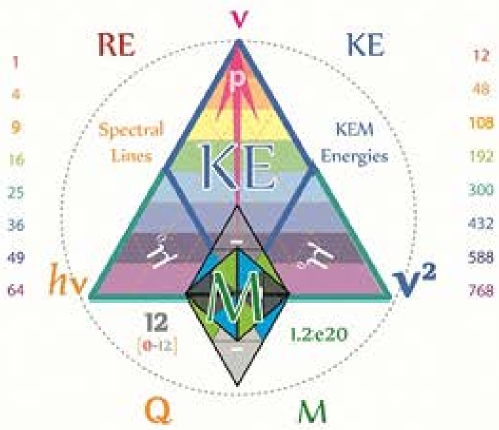
RE
KE
v
p
Spectral Lines
KEM Energies
KE
M
hv
v²
12
[0-12]
1.2e20
Q
M
1
4
9
16
25
36
49
64
12
48
108
192
300
432
588
768

Mapping photo-electron transition energies
to Tetryonic energy momenta geometries
reveals many key facts about the
ionisation energies of nuclei

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

The differing fractional KEM field energy momenta of electrons
that results from their transitions to specific energy nuclei
in elements results in differing QAM quanta and
produces spectral lines and fine line splitting

Spectral line transitions

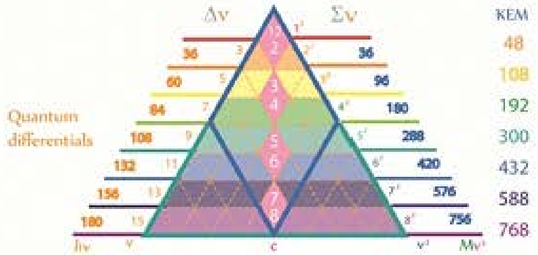
Δv
Σv
KEM
Quantum differentials
36
60
84
108
132
156
180
36
96
180
288
420
576
756
48
108
192
300
432
588
768
hv
v
c
v²
Mv²

hv, R, β, p, KE
Planck, Rydberg, Lorentz, Newton, Leibniz

uniting classical physics and relativity
through equilateral geometry

Elementary ionisation energies

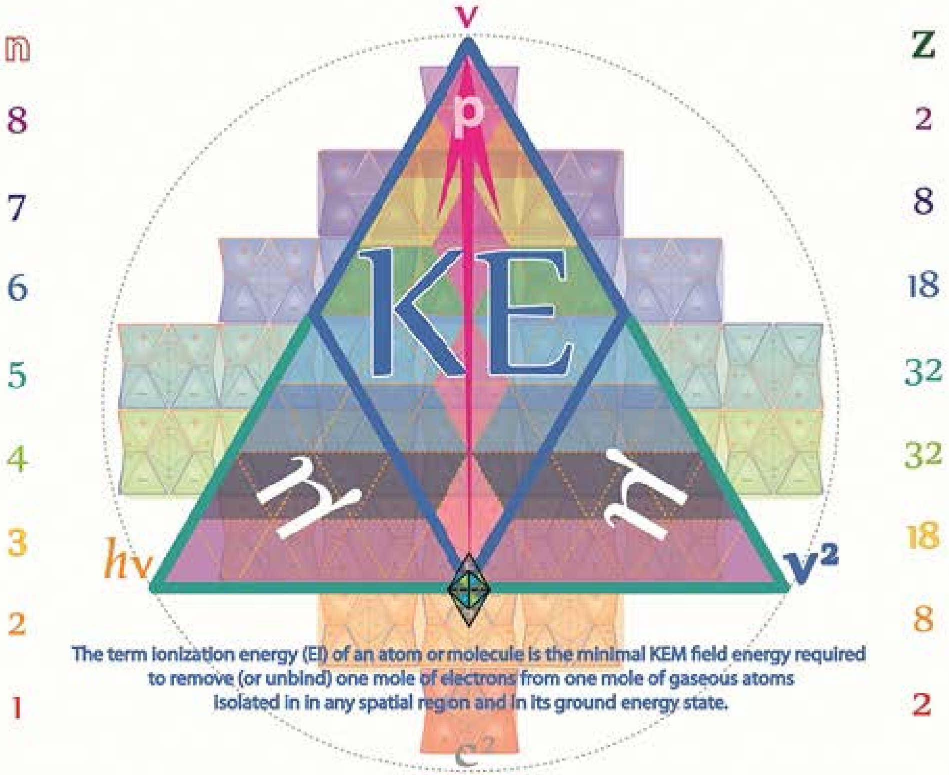

The term ionization energy (EI) of an atom or molecule is the minimal KEM field energy required to remove (or unbind) one mole of electrons from one mole of gaseous atoms isolated in in any spatial region and in its ground energy state.

The term "ionization energy" is sometimes used as a name for the work needed to remove (or un-bind) the highest energy photoelectron from an atom or molecule.

However, due to interactions with surfaces, this value differs from the ionization energy of the atom or molecule in question when it is located by itself in free space.

So, in the case of surface-adsorbed atoms and molecules, it may be better to use the more general term "electron binding energy", in order to avoid confusion.

Both these names are also sometimes used to describe the work needed to remove an electron from a "lower" orbital (i.e., not the topmost orbital) for both free and adsorbed atoms; in such cases it is necessary to specify the orbital from which the electron has been removed

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

Every electron in each elementary orbit has a unique ionisation energy

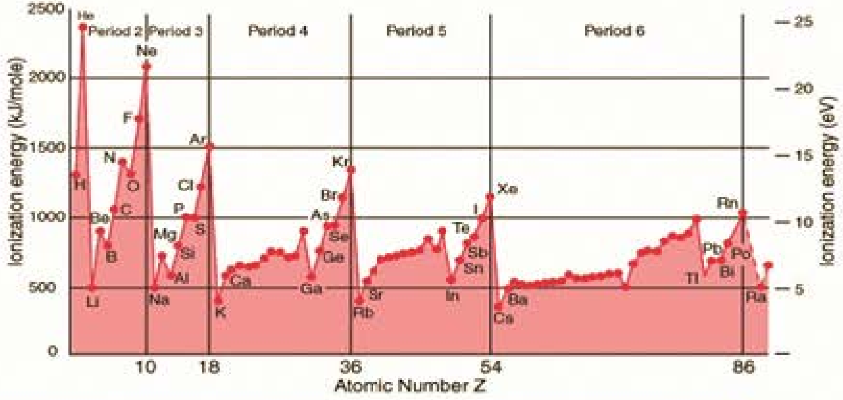

Element number

Nuclei	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Energy level	per shell
R	119	120																															8	2
Q	87	88	113	114	115	116	117	118																									7	8
P	55	56	81	82	83	84	85	86	103	104	105	106	107	108	109	110	111	112															6	18
O	37	38	49	50	51	52	53	54	71	72	73	74	75	76	77	78	79	80	89	90	91	92	93	94	95	96	97	98	99	100	101	102	5	32
N	19	20	31	32	33	34	35	36	39	40	41	42	43	44	45	46	47	48	57	58	59	60	61	62	63	64	65	66	67	68	69	70	4	32
M	11	12	13	14	15	16	17	18	21	22	23	24	25	26	27	28	29	30															3	18
L	3	4	5	6	7	8	9	10																									2	8
K	1	2																															1	2
Shells	s1	s2	p1	p2	p3	p4	p5	p6	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14		

s p d f

Orbitals

sub-orbitals

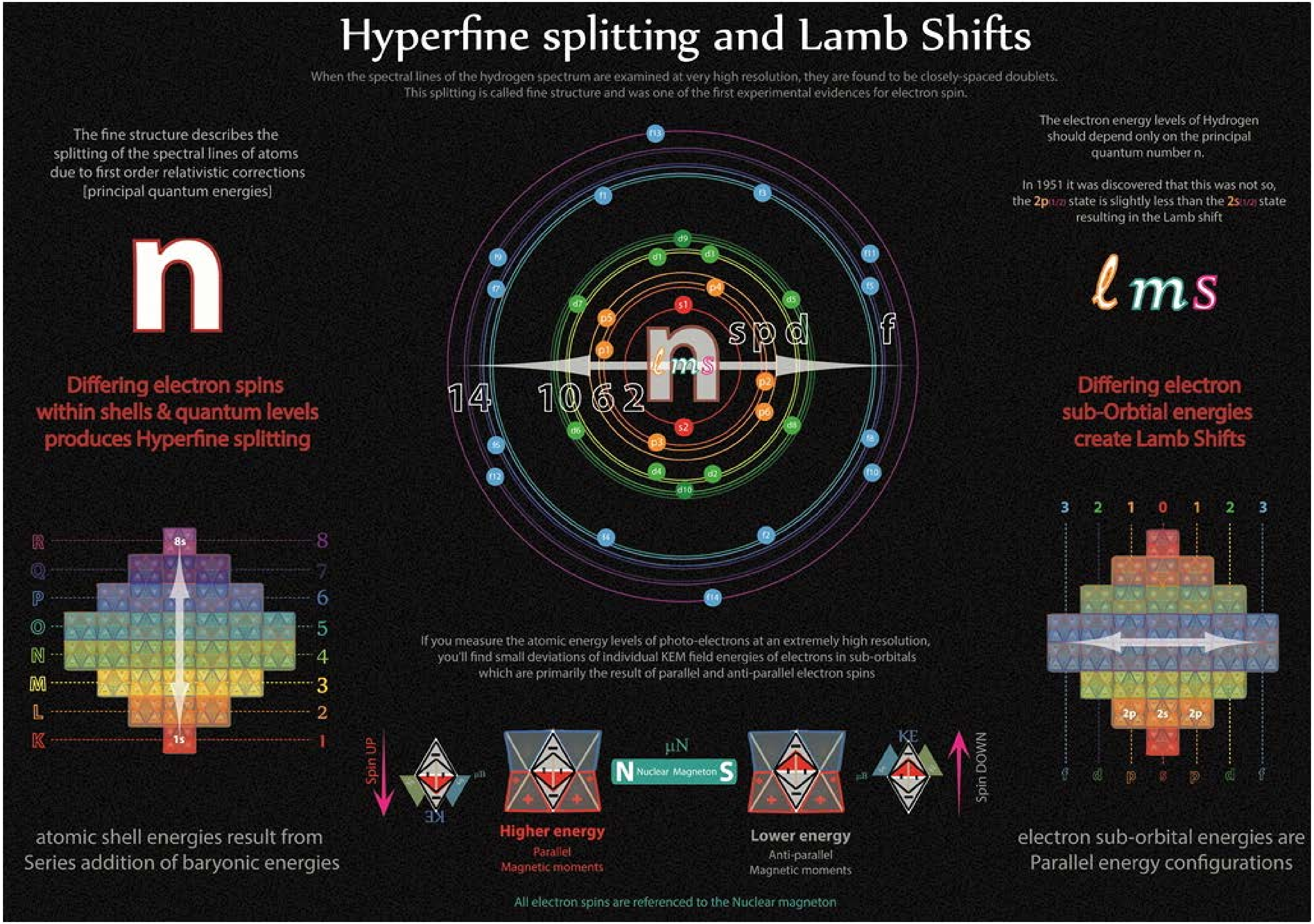

Hyperfine splitting and Lamb Shifts

When the spectral lines of the hydrogen spectrum are examined at very high resolution, they are found to be closely-spaced doublets.
This splitting is called fine structure and was one of the first experimental evidences for electron spin.

The fine structure describes the splitting of the spectral lines of atoms due to first order relativistic corrections [principal quantum energies]

The electron energy levels of Hydrogen should depend only on the principal quantum number n.

In 1951 it was discovered that this was not so, the $2p_{(1/2)}$ state is slightly less than the $2s_{(1/2)}$ state resulting in the Lamb shift

n
s p d f
14 10 6 2
n
ℓ m s
s1
s2
p1
p2
p3
p4
p5
p6
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
f14
ℓ m s

Differing electron spins within shells & quantum levels produces Hyperfine splitting

Differing electron sub-Orbtial energies create Lamb Shifts

R
Q
P
O
N
M
L
K
8s
1s
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
2p
2s
2p
f d p s p d f

If you measure the atomic energy levels of photo-electrons at an extremely high resolution, you'll find small deviations of individual KEM field energies of electrons in sub-orbitals which are primarily the result of parallel and anti-parallel electron spins

Spin UP
KE
μn
μN
N Nuclear Magneton S
μn
KE
Spin DOWN

atomic shell energies result from Series addition of baryonic energies

Higher energy
Parallel
Magnetic moments

Lower energy
Anti-parallel
Magnetic moments

electron sub-orbital energies are Parallel energy configurations

All electron spins are referenced to the Nuclear magneton

Copyright ABRAHAM [2008] - All rights reserved

M_0

3D Matter topologies are comprised of charged 2D mass-energies

Energy per second²

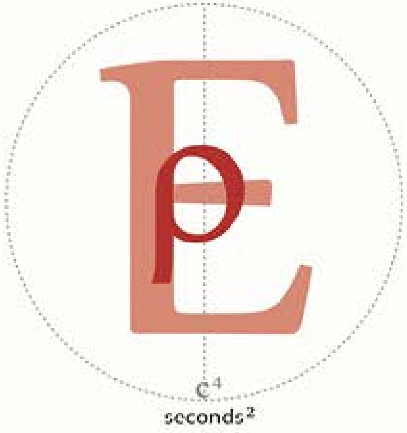
c^4
seconds²

Matter Planck quanta

$$\frac{T\pi}{c^4}[[m\Omega v^2]]$$

mass velocity

mass-Matter

standing wave mass-energiescreate
the material substance of all chemical elements

RE

Relativistic mass-ENERGY-Matter

Relativity fails at the foundational level to explain and differentiate between mass-ENERGY and Matter in physical systems

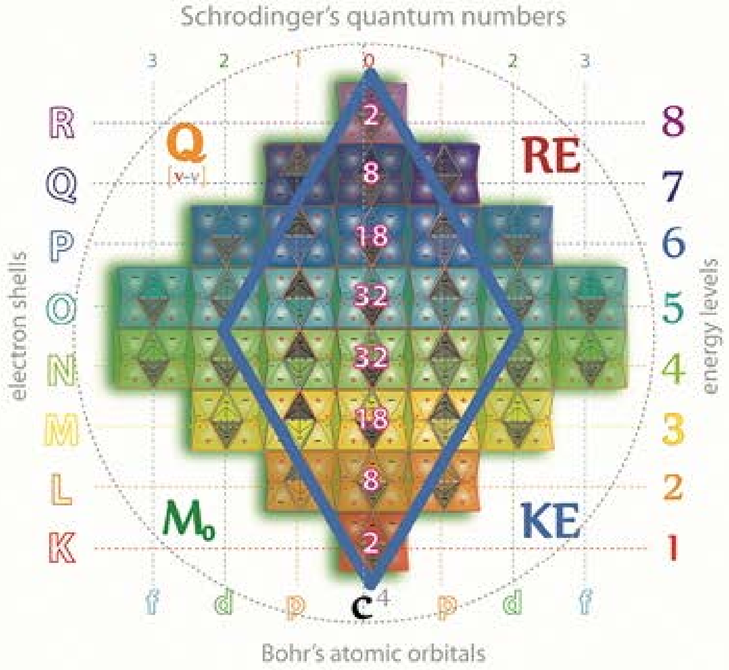
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
M_0
KE
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

Einstein's relativistic [Lorentz corrected] stress energy tensor aggregates all forms of energy into a single energy density gradient

atomic energies

$$Tm\frac{n}{c^4} + m\frac{v^2}{c^2}$$

electron spins

3D rest Matter + Lorentz corrected 2D Kinetic Energies
= total Relativistic Energies

KE

2D equilateral mass-energies are euclidean geometries

Energy per second

mass Planck quanta

$$\frac{n\pi}{c^2}[[m\Omega v^2]]$$

mass velocity

mass-energies

radiant planar mass-energies create
EM fields, spectral lines & chemical interactions

Deuterium is the building block of all Elements (save Hydrogen)

The geometry of any Element is determined by its Charge

The Molar Weight of any Element is a measure of its standing wave mass-energies

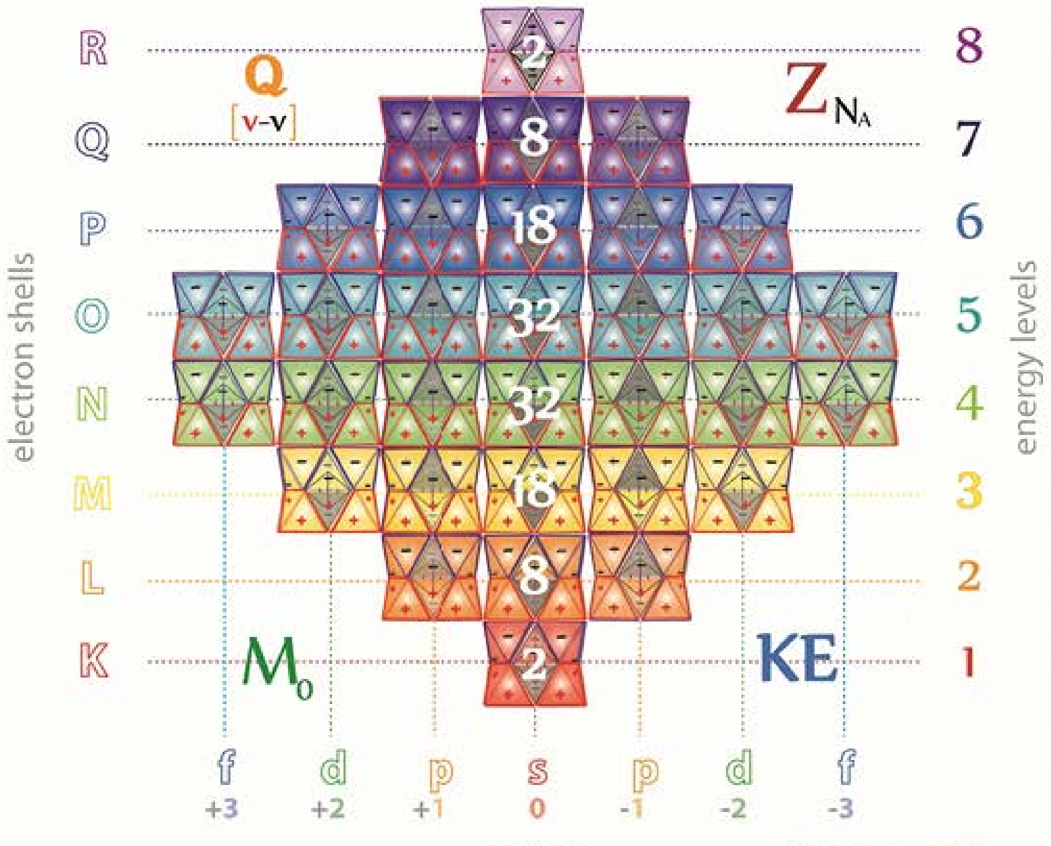

The rest mass-Matter of any Element is determined by the total number of quanta making up the Protons, Neutrons and electrons that comprise them (in their respective energy levels)

Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies

Important point to note:

The Kinetic Energy difference between any Element's total [n1] Deuteron mass-energies and its Molar mass has historically been incorrectly explained as resulting from an excess number of Neutrons in the atom it is not. Z# = (number of Protons = number of electrons = number of Neutrons)

Elementary mass-Matter

$\sum_{K}^{R}$

	n per nuclei 1e19v = n	
2	2 nuclei [74,496 ea] +	120 Unbinilium 119 Ununennium
8	8 nuclei [69,780 ea] +	118 Ununoctium 87 Francium
18	18 nuclei [65,232 ea] +	112 Copernicium 55 Caesium
32	32 nuclei [60,852 ea] +	102 Nobelium 37 Rubidium
32	32 nuclei [56,640 ea] +	70 Ytterbuim 19 Potassium
18	18 nuclei [52,596 ea] +	30 Zinc 11 Sodium
8	8 nuclei [48,720 ea] +	10 Neon 3 Lithium
2	2 nuclei [45,012 ea]	2 Helium 1 Deuterium

The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei

Aufbau

$$Z\#\left[\begin{array}{ll} z \text{ Protons} & [24\text{-}12] \\ z \text{ Neutrons} & [18\text{-}18] \\ z \text{ electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

(ie Calcium [20] = 2+8+10 n level Deuterium nuclei)

Periodic Harmonic motions

$x = A\cos(\omega t + \varphi)$

Circular motion

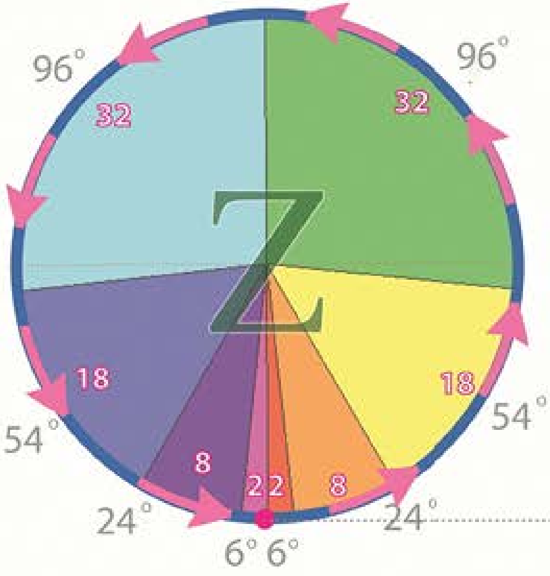
96°
96°
32
32
Z
18
18
54°
54°
8
2 2
8
24°
24°
6° 6°

circular harmonic motion

Circular motions describe the motion of a body with a changing velocity vector [the result of an acceleration force].

Much of the math in of modern physics is predicated on the assumption that π [where it appears] is related to the properties of a circle

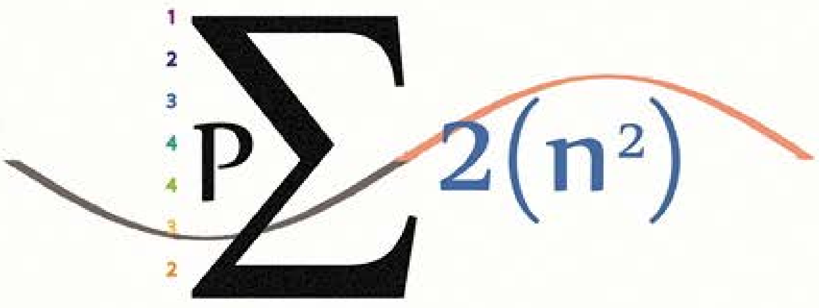

$$P\sum 2(n^2)$$

Simple harmonic motion can be visualized as the projection of uniform circular motion onto one axis

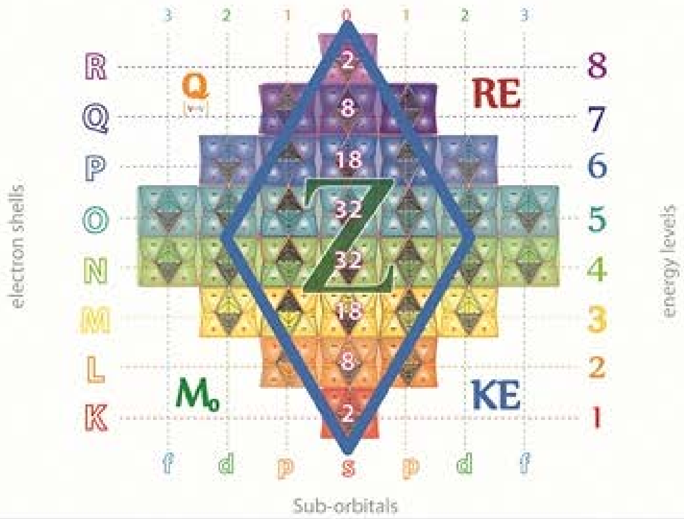
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M₀
KE
2
8
18
32
Z
32
18
8
2
8
7
6
5
4
3
2
1
electron shells
energy levels
f
d
p
s
p
d
f
Sub-orbitals

$F = -kx$

Linear motion

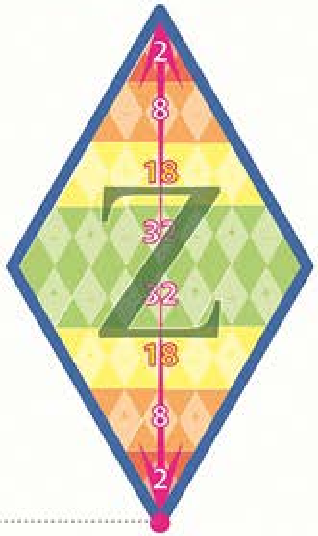
2
8
18
32
Z
32
18
8
2

simple harmonic motion

Nuclei per shell in elements follows a 'periodic summation rule' that is reflective of photonic energies

Copyright ABRAHAM [2008] - All rights reserved

$$\sum_{K}^{R}$$

nuclei per shell

$$_{K}^{R}P\sum Z$$

$$\sum_{1}^{120} Z$$

element number

Periodic Summation

Periodic summation is a notation developed for Tetryonic theory to model the geometric series addtion of $Z[n^2]$ energy level Deuterium nuclei that form the periodic elements

STEP ONE

Periodic summation follows the atomic shell electron config

STEP TWO

Periodic elements build up following the aufbau sequence

$$P\sum_{K}^{R} 2(x^2) =$$

1 R
2
3
4
4
3
2
1 K

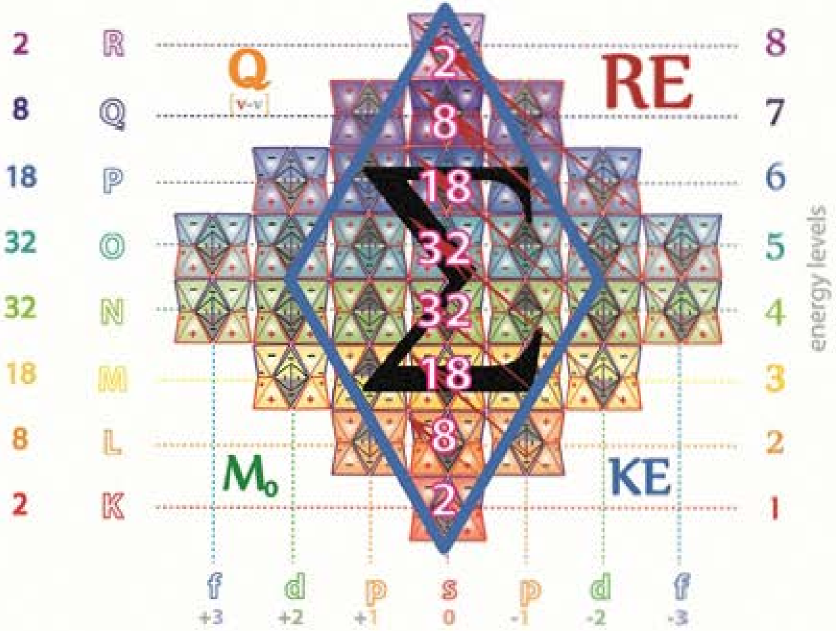

Each atomic shell can hold only a fixed number of deuterium nuclei

$P\sum R = 2$	2 nuclei [74,496 ea]	120	Unbinilium
	+		
$P\sum Q = 8$	8 nuclei [69,780 ea]	118	Ununoctium
	+		
$P\sum P = 18$	18 nuclei [65,232 ea]	110	Darmstadtium
	+		
$P\sum O = 32$	32 nuclei [60,852 ea]	92	Uranium
	+		
$P\sum N = 32$	32 nuclei [56,640 ea]	60	Neodymuim
	+		
$P\sum M = 18$	18 nuclei [52,596 ea]	28	Argon
	+		
$P\sum L = 8$	8 nuclei [48,720 ea]	10	Neon
	+		
$P\sum K = 2$	2 nuclei [45,012 ea]	2	Helium
		0	Hydrogen

THe LHS of the notation determine the number of nuclei in each atomic shell, from the periodic mass-energy levels for atoms, and the RHS follows the aufbau building principle to determine the rest mass-Matter of any specific element

Aufbau

Each periodic element is made of Z [n²energy] deuterium nuclei

$$Z\#\left[\begin{array}{ll} z \text{ Protons} & [24\text{-}12] \\ z \text{ Neutrons} & [18\text{-}18] \\ z \text{ electrons} & [0\text{-}12] \end{array}\right] n1\text{-}8$$

Planck mass-energies form the surface integral of rest Matter topologies for each periodic element

Copyright ABRAHAM [2008] - All rights reserved

Element numbers

Nuclei per shell in elements follow
a 'periodic summation rule'
that is reflective of
photonic energies

$$\sum_{P\,K}^{R} 2(x^2) = 2 + 8 + 18 + 32 + 32 + 18 + 8 + 2$$

1 R
2
3
4
4
3
2
1 K

Z

120	Unbinilium	n8
119	Ununennium	
118	Ununoctium	n7
87	Francium	
112	Copernicium	n6
55	Caesium	
102	Nobelium	n5
37	Rubidium	
70	Ytterbuim	n4
19	Potassium	
30	Zinc	n3
11	Sodium	
10	Neon	n2
3	Lithium	
2	Helium	n1
1	Deuterium	

Hydrogen 0

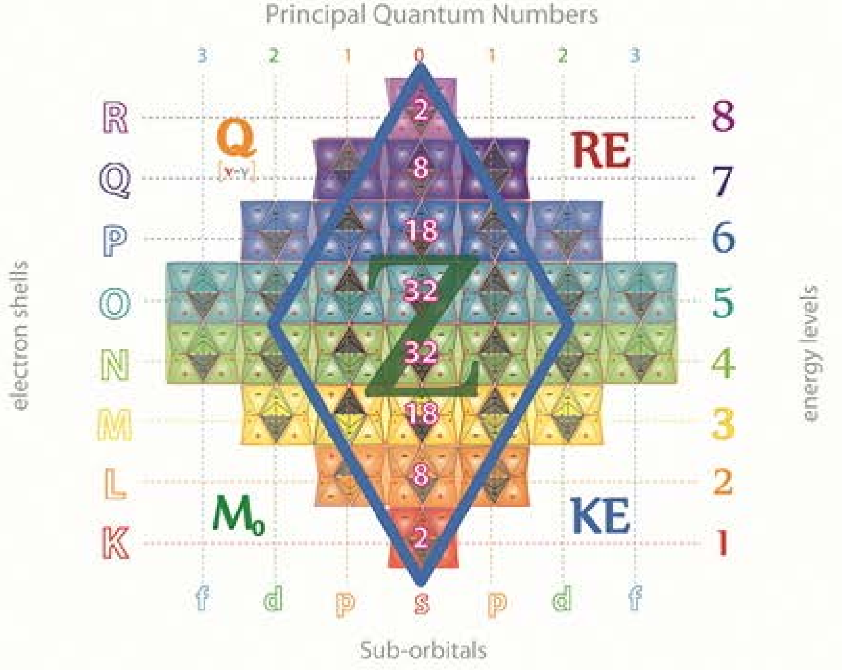
Principal Quantum Numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
Z
2
8
18
32
32
18
8
2
M0
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
s
p
d
f
Sub-orbitals

Periodic mass-ENERGY-Matter

Following periodic summation rules for shell filling
n[1-8] quantum energy deuterium nuclei
combine to form elementary Matter

$$\sum_{P\,K\,25}^{R\,32}\Big[\underbrace{\overset{\text{Baryon rest masses}}{[72(n)^2]}+\overset{\text{lepton rest mass}}{[12e19]}+\overset{\text{KEM}}{[m_e v^2]}}_{\text{Deuterium mass-energy per shell}}\Big]_1^8$$

The measured weight of Matter in gravitational fields is the result of
planar mass-energies in tetryonic standing-wave geometries

The periodicity of all the elements,
along with their exact molar rest mass-energies and
quantum wavefunctions can be described with Tetryonic geometries

Ionisation energies

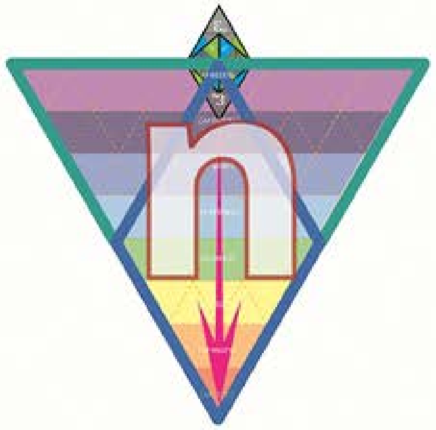

γ

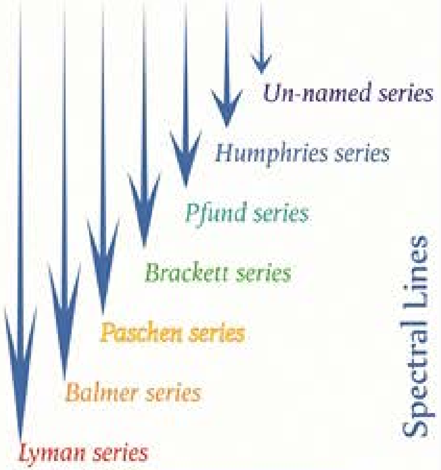
Un-named series
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
Spectral Lines

$$\boxed{Mv^2 = KEM = hcR_H}$$

Photon emission/absorption

Photo-electron ionisation energies

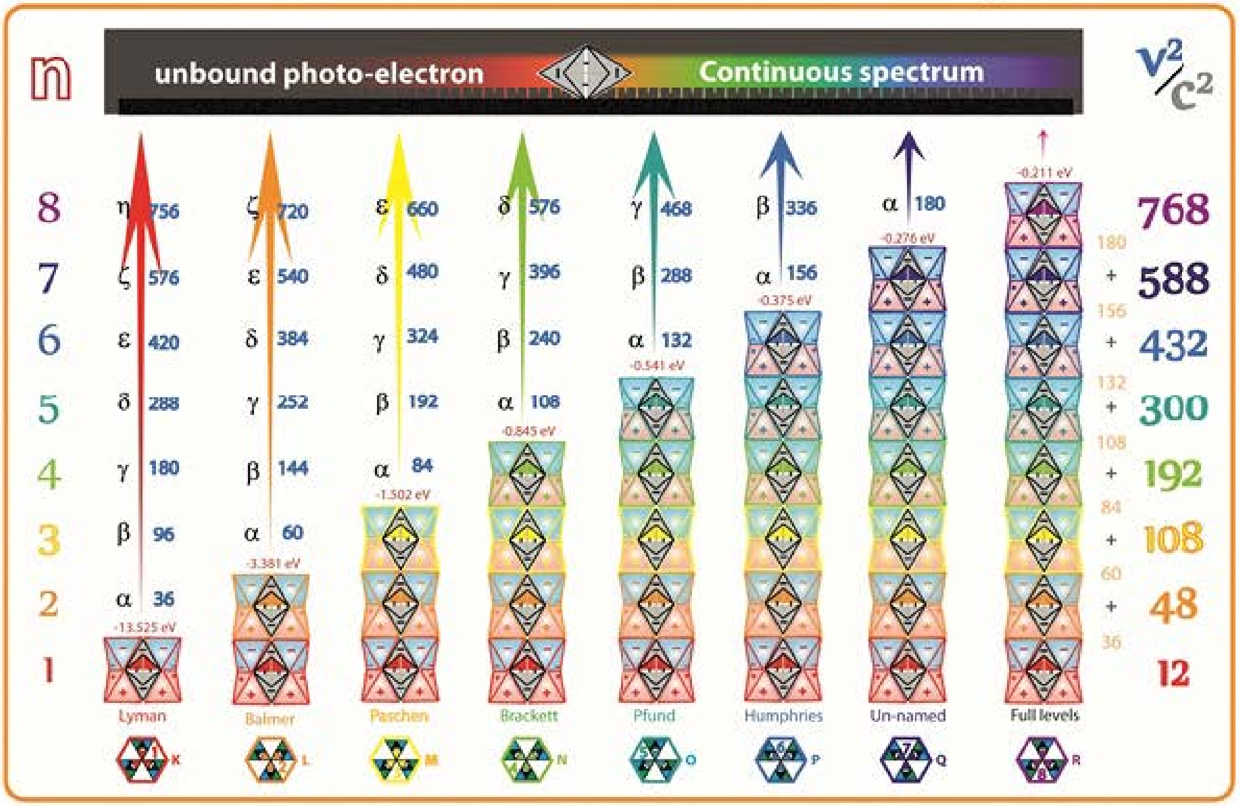
n
unbound photo-electron
Continuous spectrum
v²/c²
8
7
6
5
4
3
2
1
η 756
ζ 576
ε 420
δ 288
γ 180
β 96
α 36
-13.525 eV
ζ 720
ε 540
δ 384
γ 252
β 144
α 60
-3.381 eV
ε 660
δ 480
γ 324
β 192
α 84
-1.502 eV
δ 576
γ 396
β 240
α 108
-0.845 eV
γ 468
β 288
α 132
-0.541 eV
β 336
α 156
-0.375 eV
α 180
-0.276 eV
-0.211 eV
180
+
156
+
132
+
108
+
84
+
60
+
36
768
588
432
300
192
108
48
12
Lyman
Balmer
Paschen
Brackett
Pfund
Humphries
Un-named
Full levels
K
L
M
N
O
P
Q
R

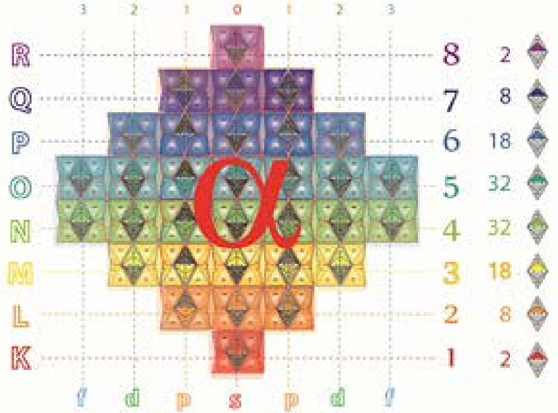
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
α
f d p s p d f

Opposed electron spins results in Hyperfine splitting & Lamb shifts

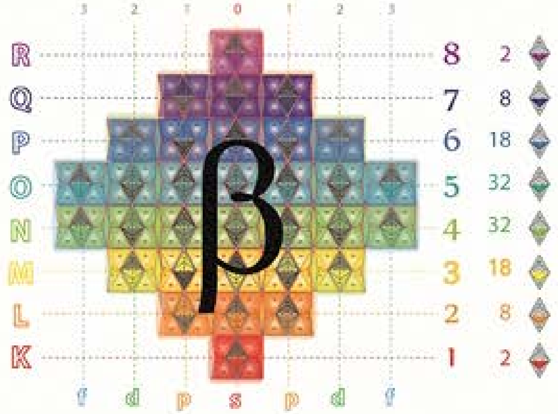
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
β
f d p s p d f

Energy levels

Azmithal & Magnetic numbers

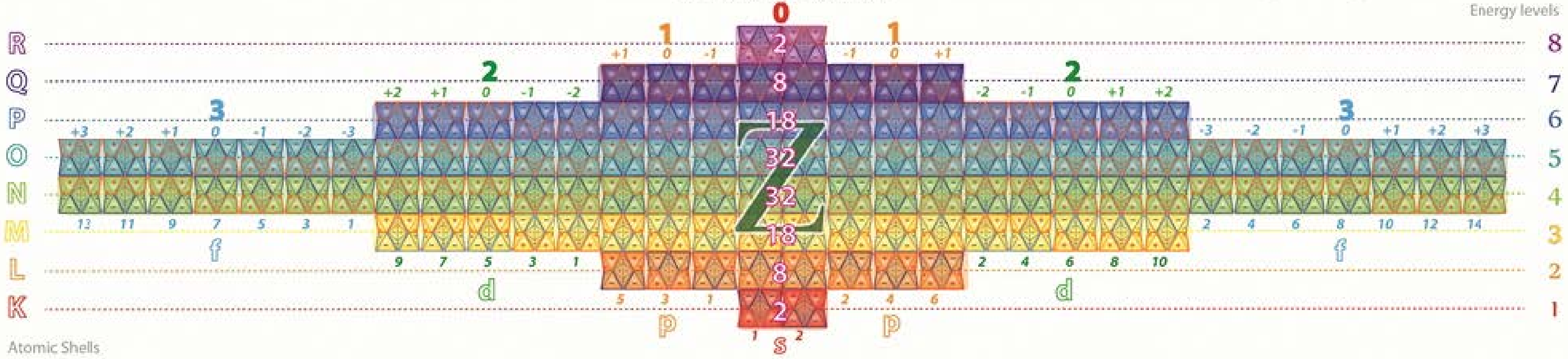
R
Q
P
O
N
M
L
K
0
1
1
+1 0 -1
-1 0 +1
2
2
+2 +1 0 -1 -2
-2 -1 0 +1 +2
3
3
+3 +2 +1 0 -1 -2 -3
-3 -2 -1 0 +1 +2 +3
2
8
18
32
32
18
8
2
Z
13 11 9 7 5 3 1
2 4 6 8 10 12 14
f
f
9 7 5 3 1
2 4 6 8 10
d
d
5 3 1
2 4 6
p
p
1 s 2
8
7
6
5
4
3
2
1

Atomic Shells

Orbitals & sub-Orbitals

Eigenstate - Ionisation energies

Atomic shells
P
O
Q
K
N
R
L
M

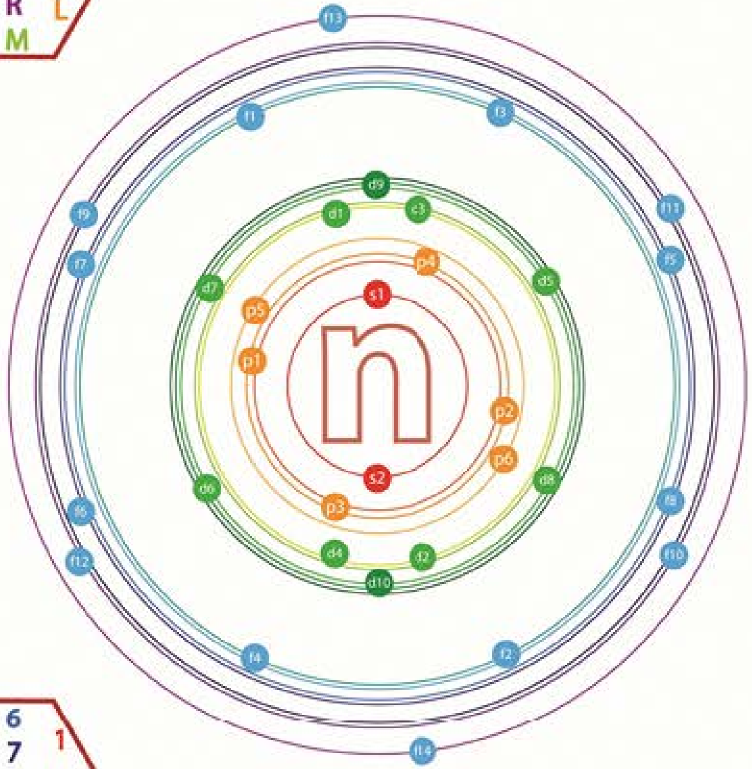
n
s1
s2
p1
p2
p3
p4
p5
p6
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10
f11
f12
f13
f14

6
5
7
1
8
4
2
3
Quantum numbers

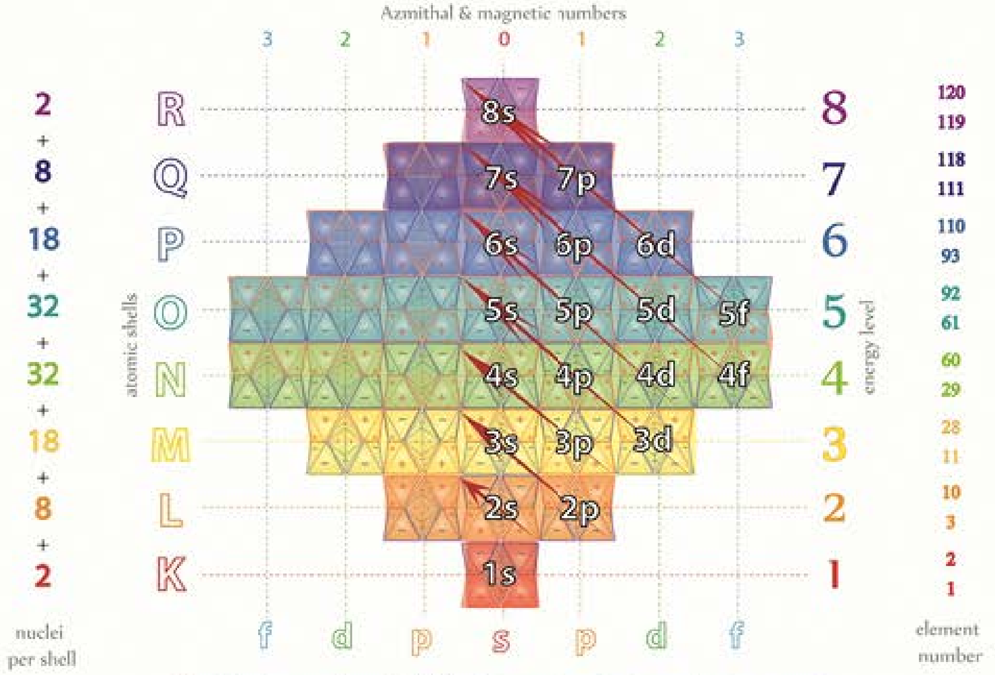
Azmithal & magnetic numbers
3 2 1 0 1 2 3
2
+
8
+
18
+
32
+
32
+
18
+
8
+
2
nuclei per shell
atomic shells
R
Q
P
O
N
M
L
K
8s
7s 7p
6s 6p 6d
5s 5p 5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
f d p s p d f
8
7
6
5
4
3
2
1
energy level
120
119
118
111
110
93
92
61
60
29
28
11
10
3
2
1
element number

The ionisation energies of individual atoms varies due to many factors, namely:
electron spin-orbital coupling with Baryons of specific energies,
the relativistic energies of photo-electrons bound in nuclei
and Zitterbewegung effects on bound electrons

First ionisation energies for all periodic elements

$$E = -\frac{Z^2}{n^2}\frac{ke^2}{2a_0} = -\frac{13.6Z^2}{n^2}eV$$

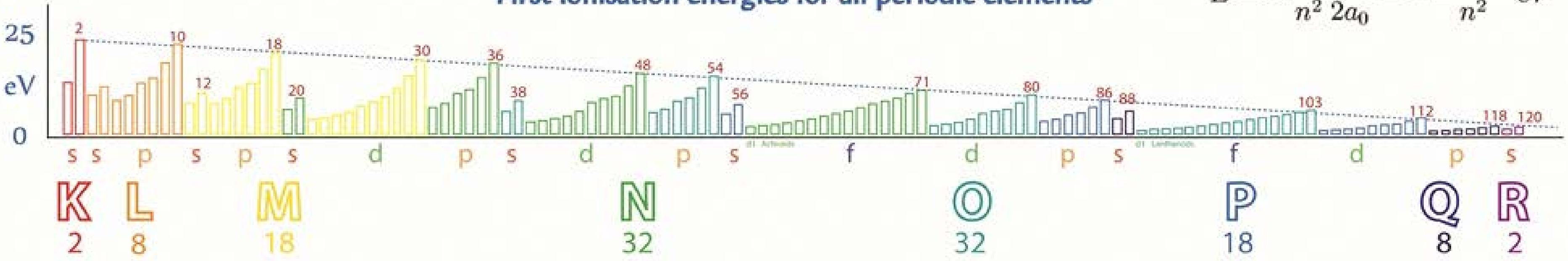

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number. It is used as rule to determine which nuclei are stable or unstable.

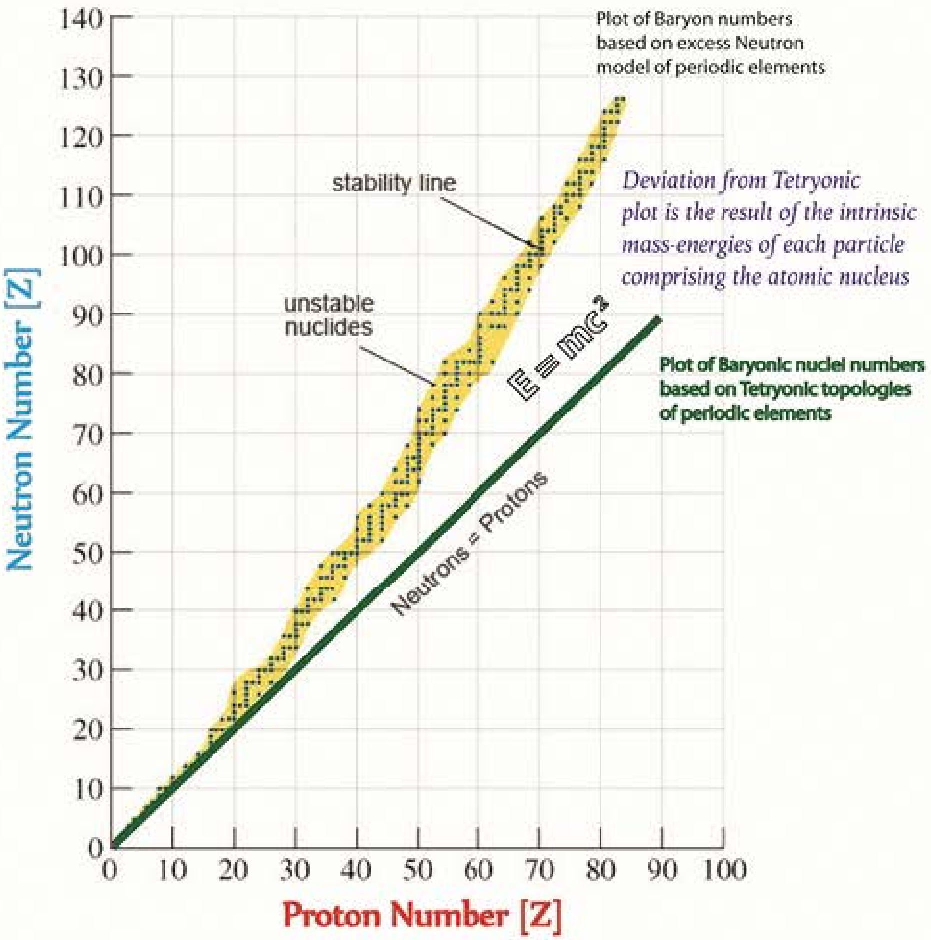

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

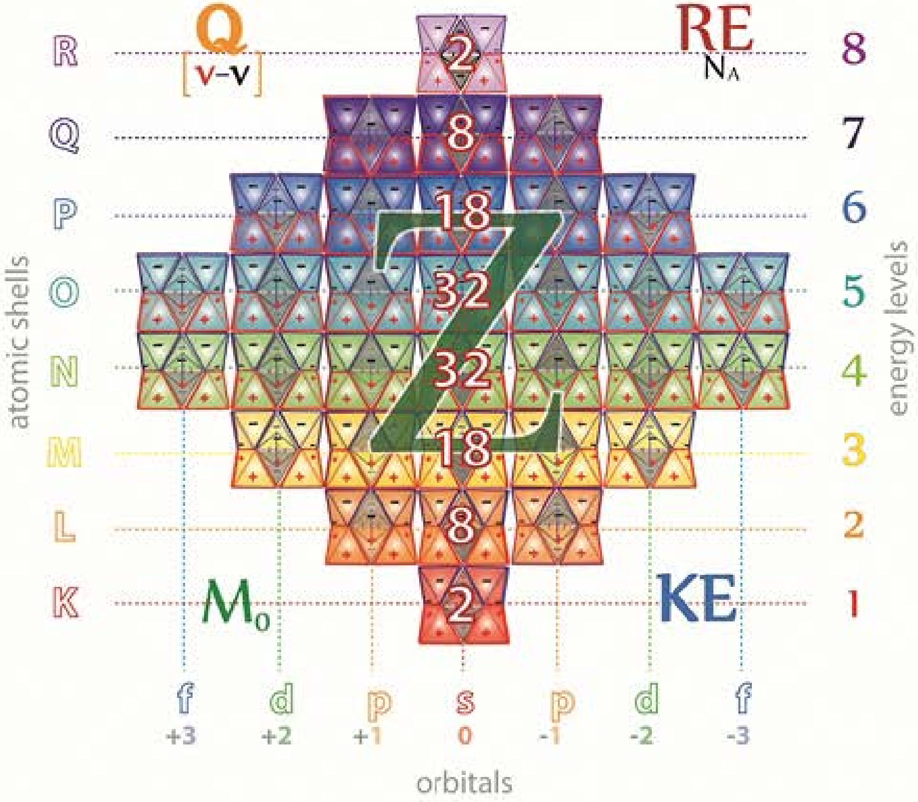
Q
[v-v]
RE
N_A
R
Q
P
O
N
M
L
K
atomic shells
2
8
18
32
32
18
8
2
Z
8
7
6
5
4
3
2
1
energy levels
M_0
KE
f
+3
d
+2
p
+1
s
0
p
-1
d
-2
f
-3
orbitals

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Planck mass-energy contributions to elementary Matter and isotopes

The mass-energy content of Deuterium nuclei creates the molar mass of elements
[not extra neutrons in excess of the elemental number]

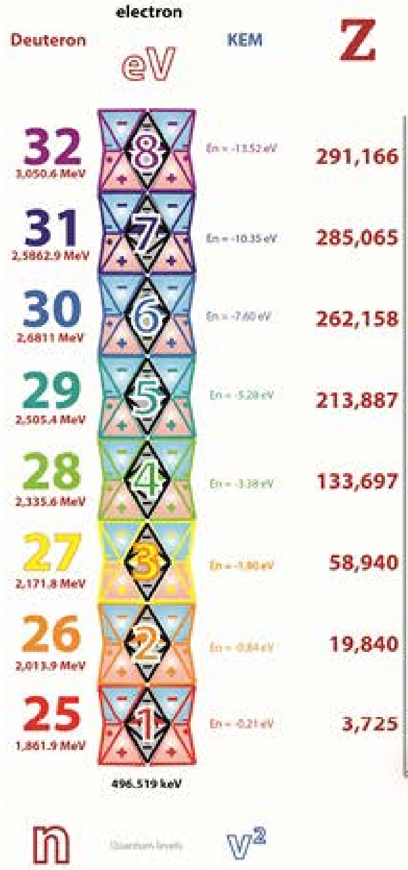
Deuteron
electron
eV
KEM
Z
32
3,050.6 MeV
8
En = -13.52 eV
291,166
31
2,5862.9 MeV
7
En = -10.35 eV
285,065
30
2,6811 MeV
6
En = -7.60 eV
262,158
29
2,505.4 MeV
5
En = -5.28 eV
213,887
28
2,335.6 MeV
4
En = -3.38 eV
133,697
27
2,171.8 MeV
3
En = -1.90 eV
58,940
26
2,013.9 MeV
2
En = -0.84 eV
19,840
25
1,861.9 MeV
1
En = -0.21 eV
3,725
496.519 keV
n
Quantum levels
v²

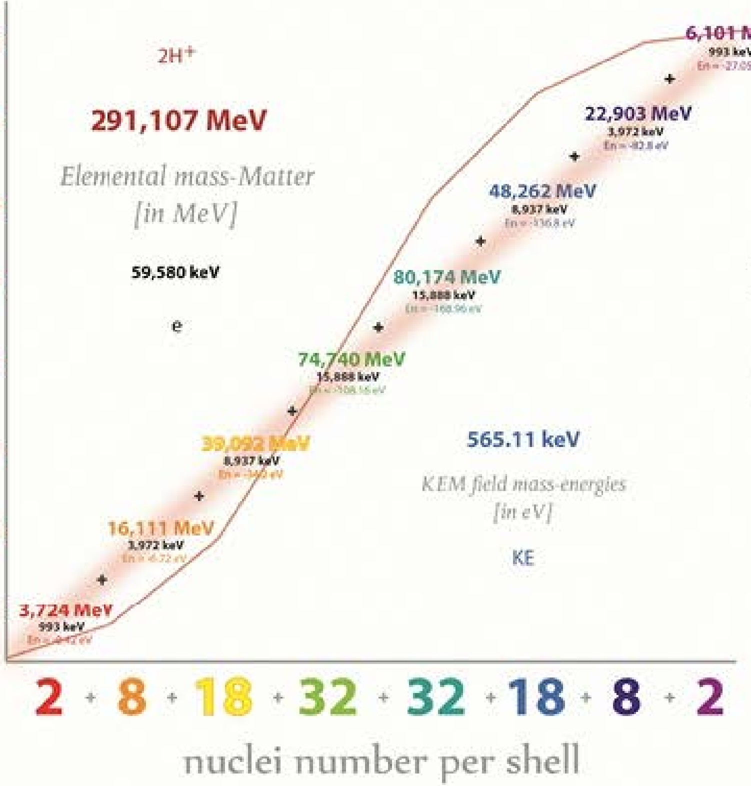

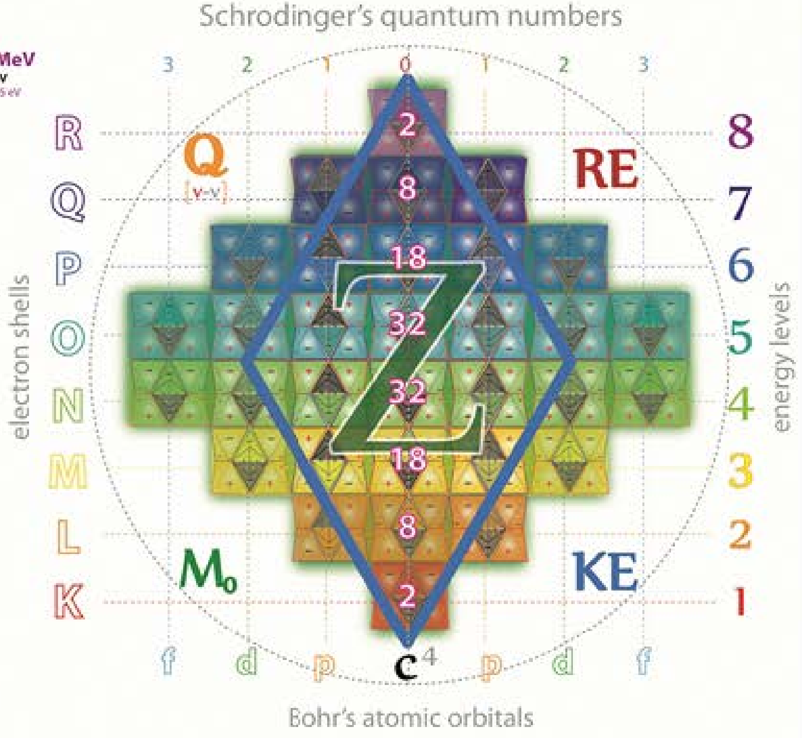
Schrodinger's quantum numbers
3
2
1
0
1
2
3
electron shells
R
Q
P
O
N
M
L
K
Q
[v-v]
RE
M₀
KE
Z
2
8
18
32
32
18
8
2
energy levels
8
7
6
5
4
3
2
1
f
d
p
c⁴
p
d
f
Bohr's atomic orbitals

n1
Elementary nuclei

are comprised of equal numbers of
Protons, Neutrons & electrons
with varying energy levels

Baryons		electrons		KEM fields
930.947 MeV	+	**496.519 keV**	+	**13.525 ev**

The mass-energy content of Matter topologies is velocity invariant

The mass-energy content of Baryons determines the KEM field of electrons

Copyright ABRAHAM [2008] - All rights reserved

Baryons **930.947 MeV** + **KEM fields** **13.525 ev** + **electrons** **496.519 keV**

Mapping Planck mass-energy contributions to elementary Matter and isotopes

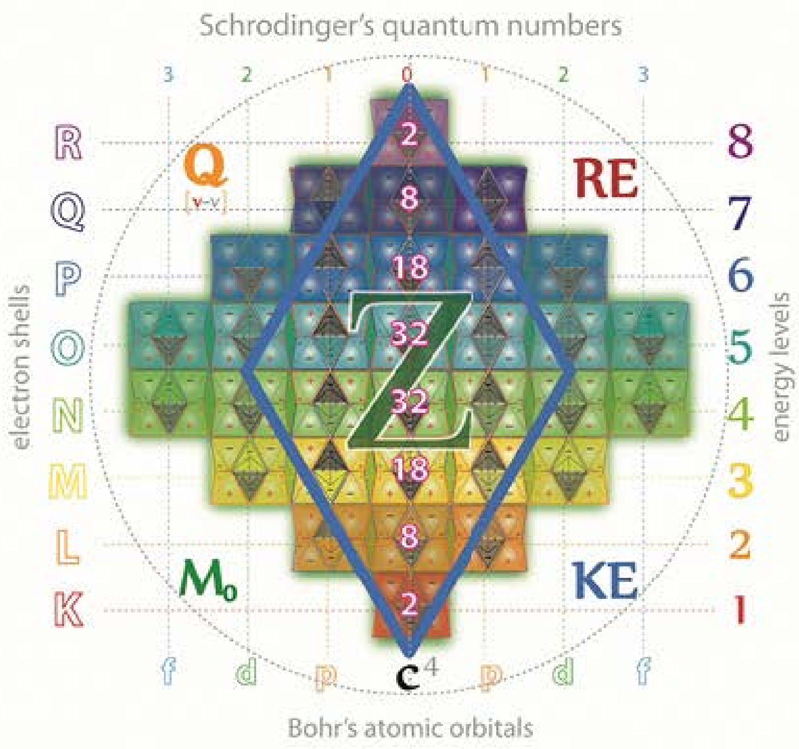
Schrodinger's quantum numbers
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
RE
2
8
18
32
Z
32
18
8
2
M_0
KE
8
7
6
5
4
3
2
1
energy levels
f
d
p
c^4
p
d
f
Bohr's atomic orbitals

$$E = h\nu^2$$

$$n^2 + \nu^2 + e = Z$$

general form quatratic equation

$$ax^2 + bx + c = Z$$

$$E = nh\nu$$

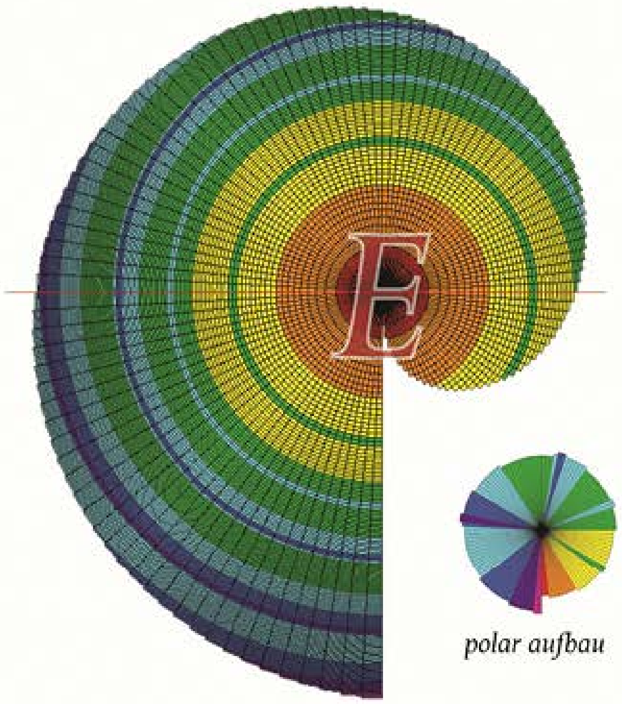
E
polar aufbau

polar energy spirals courtesy of Rene Cormier

Identifying electron rest Matter topologies as velocity invariant we can re-arrange the component Planck mass-energy geometry formulation of periodic elements to

$$h\Big[\underbrace{72[\nu^2]}_{\text{Deuteron rest mass}} + \underbrace{\nu}_{\text{Spectral lines}} + \underbrace{1.20\text{ e}20\ \nu}_{\text{electron rest mass}}\Big]$$

reveal a quadratic formulation for all Z numbers

All elements are comprised of n level Duetrium nuclei

Baryons **KEM fields** **electrons**

Z [[72n²] + [12v²] + [1.2e20]]

1,861,949 MeV 13.525 eV 496,519 keV

2H+ Z γ

[72n²] [Mₑv²]

Elemental mass-Matter [in MeV]

The atomic shell energy levels of Deuterium nuclei in elements

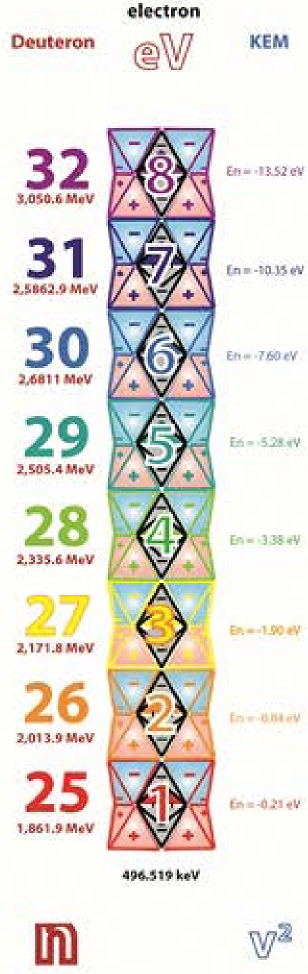
electron
Deuteron
eV
KEM
32
3,050.6 MeV
8
En = -13.52 eV
31
2,5862.9 MeV
7
En = -10.35 eV
30
2,6811 MeV
6
En = -7.60 eV
29
2,505.4 MeV
5
En = -5.28 eV
28
2,335.6 MeV
4
En = -3.38 eV
27
2,171.8 MeV
3
En = -1.90 eV
26
2,013.9 MeV
2
En = -0.84 eV
25
1,861.9 MeV
1
En = -0.21 eV
496.519 keV
m
v²

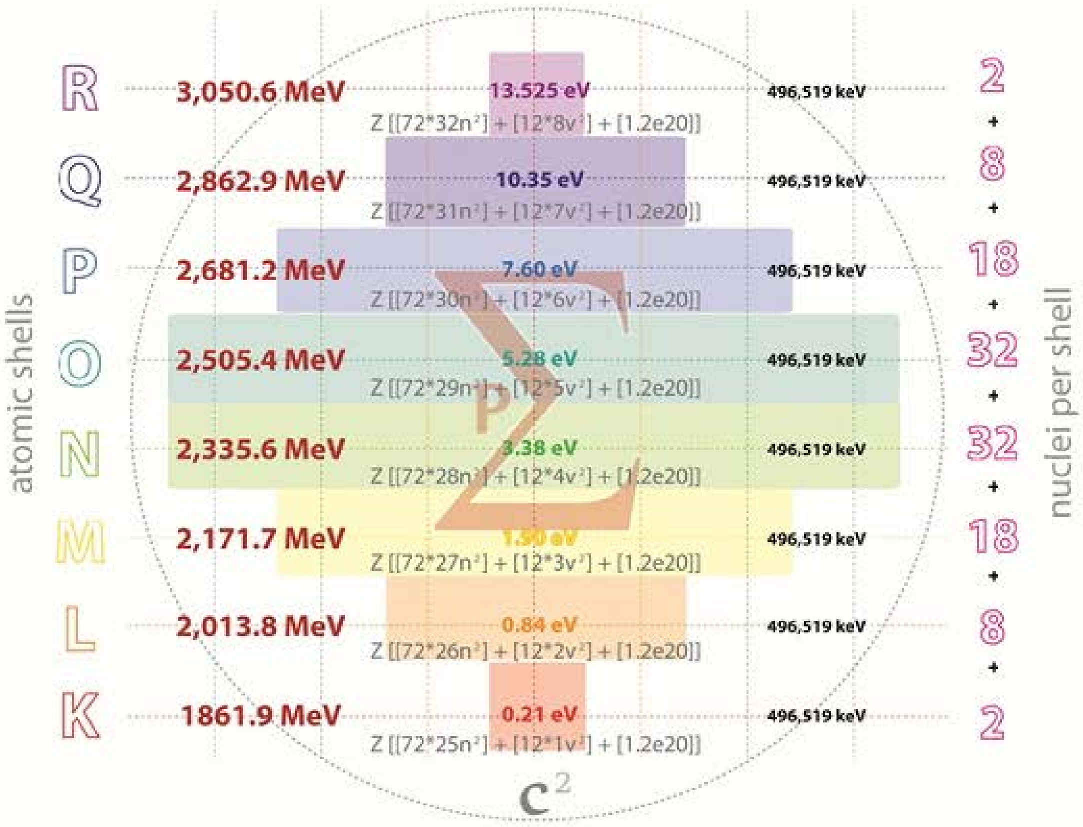
atomic shells
R
3,050.6 MeV
13.525 eV
496,519 keV
Z [[72*32n²] + [12*8v²] + [1.2e20]]
Q
2,862.9 MeV
10.35 eV
496,519 keV
Z [[72*31n²] + [12*7v²] + [1.2e20]]
P
2,681.2 MeV
7.60 eV
496,519 keV
Z [[72*30n²] + [12*6v²] + [1.2e20]]
O
2,505.4 MeV
5.28 eV
496,519 keV
Z [[72*29n²] + [12*5v²] + [1.2e20]]
N
2,335.6 MeV
3.38 eV
496,519 keV
Z [[72*28n²] + [12*4v²] + [1.2e20]]
M
2,171.7 MeV
1.90 eV
496,519 keV
Z [[72*27n²] + [12*3v²] + [1.2e20]]
L
2,013.8 MeV
0.84 eV
496,519 keV
Z [[72*26n²] + [12*2v²] + [1.2e20]]
K
1861.9 MeV
0.21 eV
496,519 keV
Z [[72*25n²] + [12*1v²] + [1.2e20]]
Σ
P
Σ
c²
2
+
8
+
18
+
32
+
32
+
18
+
8
+
2
nuclei per shell

6,101 MeV
993 keV
En = -27.05 eV
+
22,903 MeV
3,972 keV
En = -82.8 eV
+
48,262 MeV
8,937 keV
En = -316.8 eV
+
80,174 MeV
15,888 keV
En = -168.96 eV
+
74,740 MeV
15,888 keV
En = -108.16 eV
+
39,092 MeV
8,937 keV
En = -34.2 eV
+
16,111 MeV
3,972 keV
En = -6.72 eV
+
3,724 MeV
993 keV
En = -0.42 eV

[1.2e20]]

e

Determines the spectral line [KEM field energies] of electrons bound to them

The relativistic rest mass-energy-Matter of all periodic elements

is the sum of the mass-energies of all atomic nuclei and spectral lines that comprise its mass-Matter topology as measured in any spatial co-ordinate system per unit of time

the rest mass-Matter of bound photo-electrons is velocity invariant

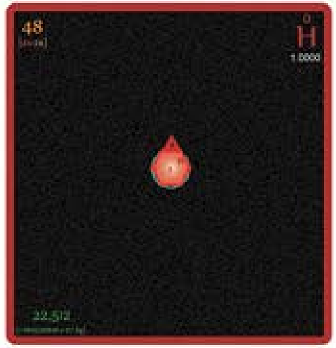

Avagadro's number

6.022141579 e26
atoms in 1KG of Matter

1 KG mass
[of Matter]

1.660538841 e-27 kg
atomic rest mass-Matter

Hydrogen

01

using SI units Avagadro's number can be expressed exactly as the inverse rest mass of Hydrogen

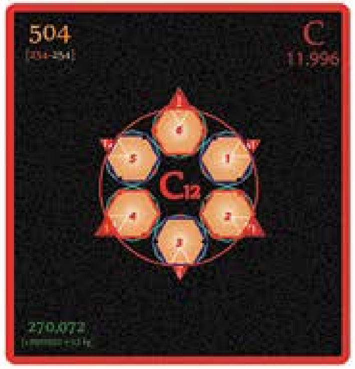

Carbon

12

Weighted atomic mass

5.019789213 e25
atoms in 1KG of Matter

6.02214078 e 23

1.99211552 e-26 kg
atomic rest mass-Matter

Tetryonic charge geometries make weighted atomic mass measurements and calculations obsolete

22,506 1/12 of Carbon 12 [Graphene] is not equal to 1 Hydrogen atom 22,512
(Deuterium is the building block of all atomic elements)

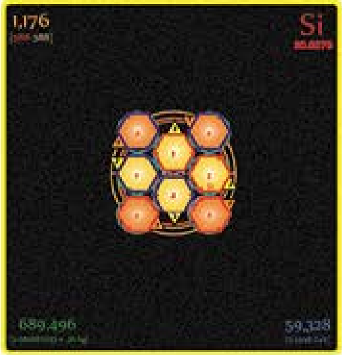

International Avagadro project

1.966225348 e25
atoms in 1KG of Matter

5.085887033 e-26 kg
atomic rest mass-Matter

Silicon

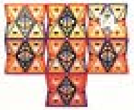

28

The gram was originally defined in 1795 as the mass of one cubic centimeter of water at 4°C,
making the kilogram equal to the mass of one liter of water.

The prototype kilogram, manufactured in 1799 and from which the current kilogram is based
has a mass equal to the mass of 1.000025 liters of water

In recent years two major experiments, namely the Watt balance & Avagadro projects, have been attempting to
measure and define 1KG of mass-Matter in terms of electrical force and the number of atoms respectively
in order to better define 1KG of mass-Matter precisely for all future physical references

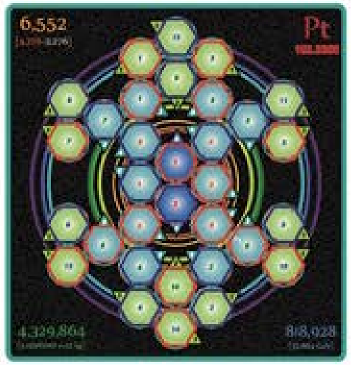

Platinum

191

La Grande K

2.817950081 e24 atoms in 1KG of Matter	90%		Pt	3.1893811012 e-25 kg atomic rest mass-Matter
3.181804449 e23	10%	*The La grande K is an alloy of 90% Platinum & 10% Iridium that has been slowly losing mass since its manufacture*	Ir	3.142870708 e-25 kg

All atomic rest masses are for atoms at absolute zero and any deviation is a measure of the topological Matter's Kinetic energy content [chemical energy, KEM fields and/or spectral lines]

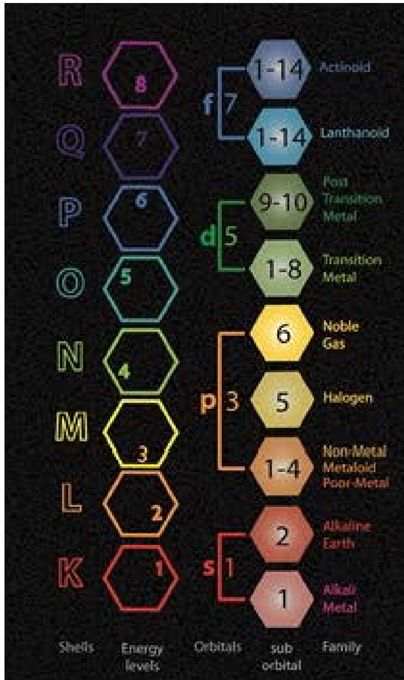
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
f 7
1-14 Actinoid
1-14 Lanthanoid
d 5
9-10 Post Transition Metal
1-8 Transition Metal
p 3
6 Noble Gas
5 Halogen
1-4 Non-Metal Metaloid Poor-Metal
s 1
2 Alkaline Earth
1 Alkali Metal
Shells
Energy levels
Orbitals
sub orbital
Family

Incorporating

- *Quark configurations*
- *Nuclei geometries*
- *Chemical bond points*
- *Atomic Orbitals*
- *Hund Rule*
- *Pauli exclusion*
- *Aufbau filling principle*
- *Element names, numbers and*
- *Elementary families*

Element placements that reflects the true quantum geometries of periodic elements

● UP Quark

● DOWN Quark

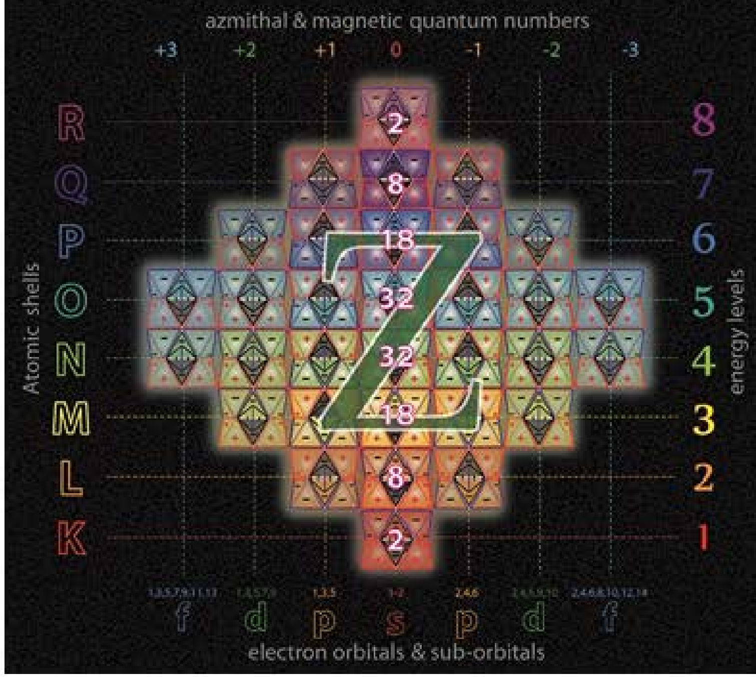
azmithal & magnetic quantum numbers
+3
+2
+1
0
-1
-2
-3
R
Q
P
O
N
M
L
K
Atomic shells
8
7
6
5
4
3
2
1
energy levels
2
8
18
32
32
18
8
2
Z
f
d
p
s
p
d
f
electron orbitals & sub-orbitals

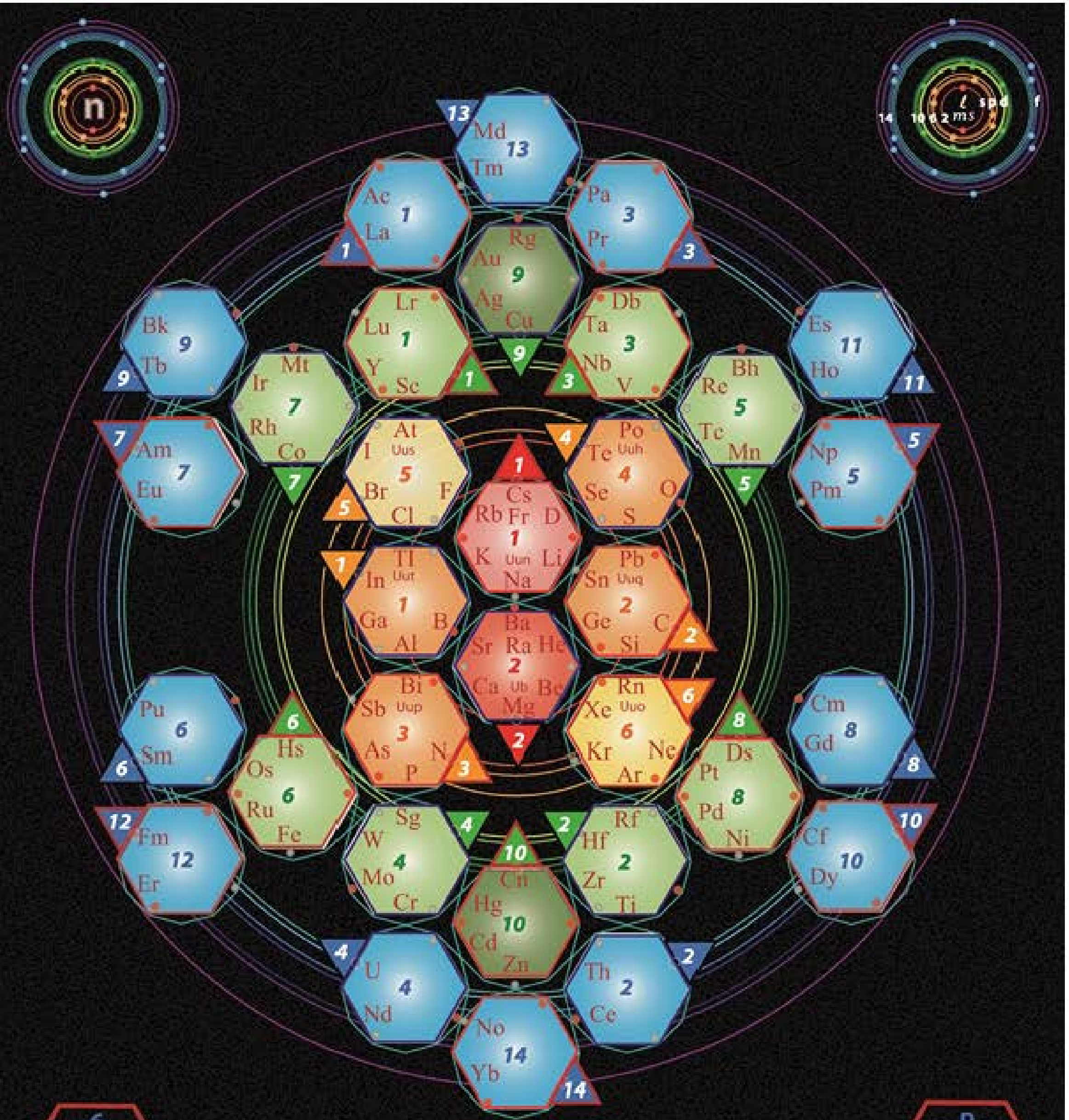
n
ℓ s p d f
m s
14 10 6 2
Md 13 Tm
13
Ac 1 La
1
Pa 3 Pr
3
Rg Au 9 Ag Cu
9
Lr Lu 1 Y Sc
1
Db Ta 3 Nb V
3
Bk 9 Tb
9
Mt Ir 7 Rh Co
7
Bh Re 5 Tc Mn
5
Es 11 Ho
11
Am 7 Eu
7
Np 5 Pm
5
At I Uus 5 Br F Cl
5
Cs Rb Fr D 1 K Uun Li Na
1
Po Te Uuh 4 Se O S
4
Tl In Uut 1 Ga B Al
1
Pb Sn Uuq 2 Ge C Si
2
Ba Sr Ra He 2 Ca Ub Be Mg
2
Bi Sb Uup 3 As N P
3
Rn Xe Uuo 6 Kr Ne Ar
6
Pu 6 Sm
6
Hs Os 6 Ru Fe
6
Ds Pt 8 Pd Ni
8
Cm 8 Gd
8
Fm 12 Er
12
Sg W 4 Mo Cr
4
Cn Hg 10 Cd Zn
10
Rf Hf 2 Zr Ti
2
Cf 10 Dy
10
U 4 Nd
4
Th 2 Ce
2
No 14 Yb
14

Energy levels

Periodic Table 2.0

(the quantum geometry of periodic elements)

Atomic shells

Copyright ABRAHAM [2008] - All rights reserved

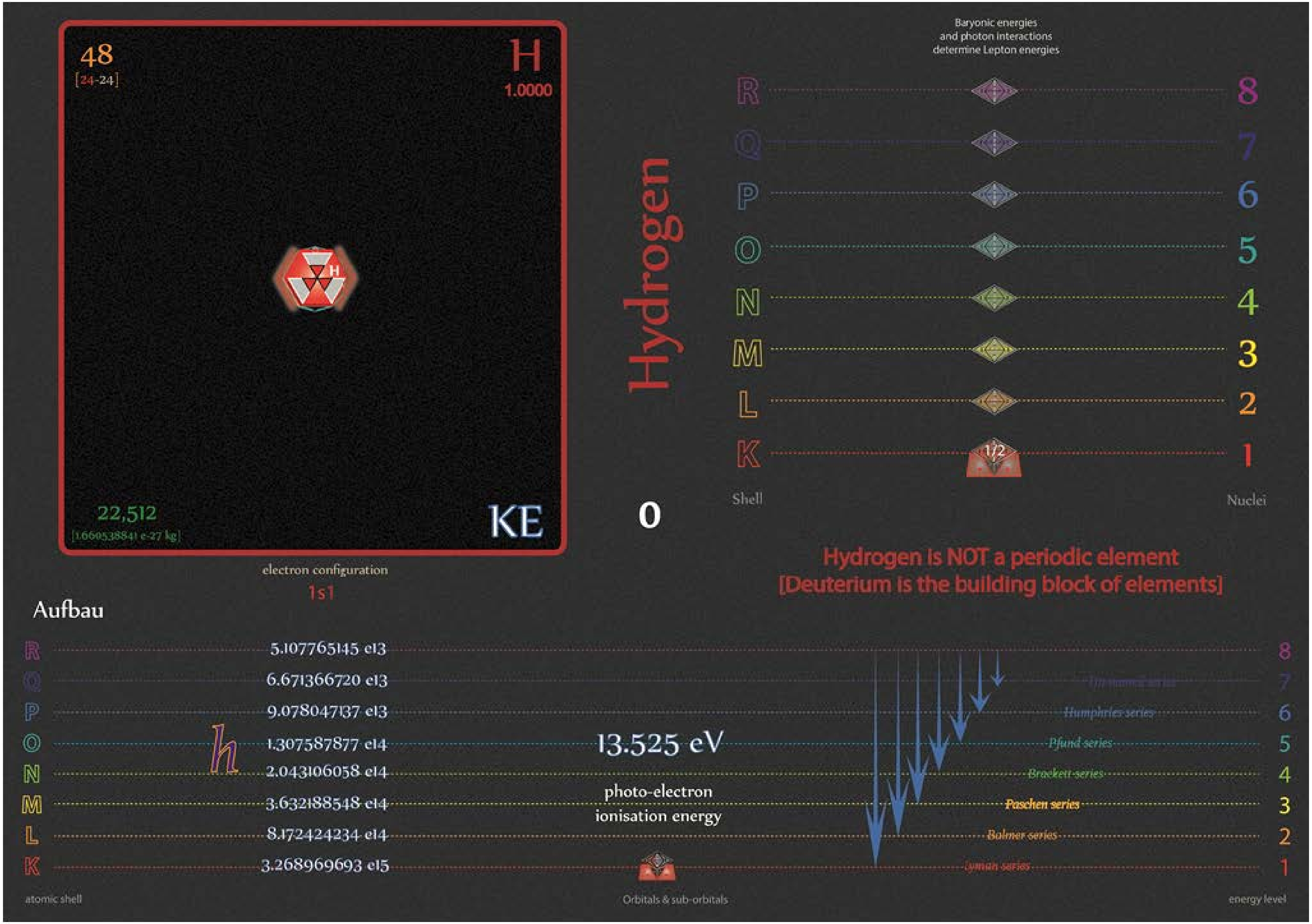
48
[24-24]
H
1.0000
22,512
[1.660538841 e-27 kg]
KE
electron configuration
1s1
Hydrogen
0
Baryonic energies
and photon interactions
determine Lepton energies
R
Q
P
O
N
M
L
K
1/2
8
7
6
5
4
3
2
1
Shell
Nuclei
Aufbau
R 5.107765145 e13
Q 6.671366720 e13
P 9.078047137 e13
O 1.307587877 e14
N 2.043106058 e14
M 3.632188548 e14
L 8.172424234 e14
K 3.268969693 e15
h
13.525 eV
photo-electron
ionisation energy
Humphries series
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
atomic shell
Orbitals & sub-orbitals
energy level

Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]

Tetryonics 51.00 - Hydrogen atom

Copyright ABRAHAM [2008] - All rights reserved

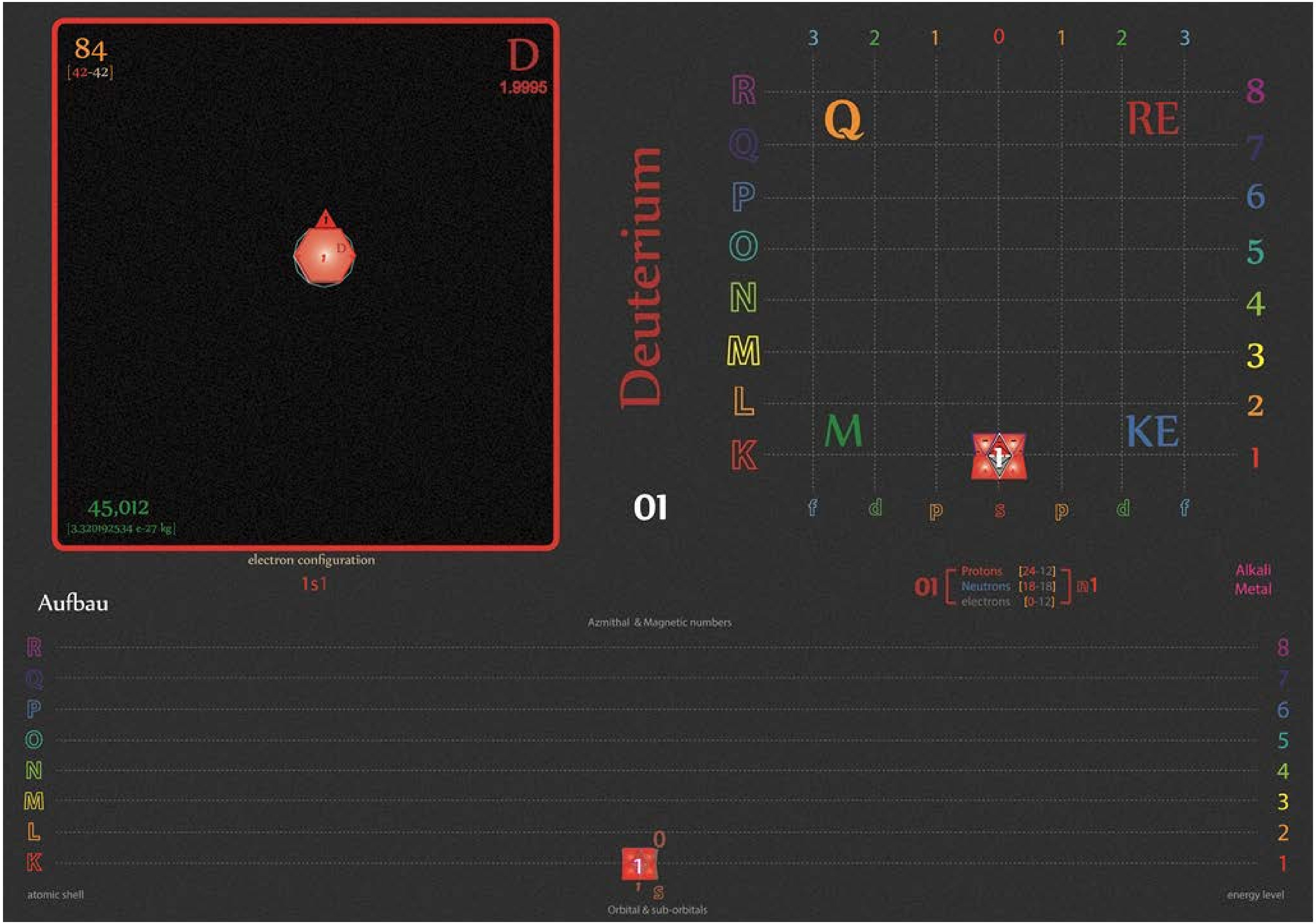
84
[42-42]
D
1.9995
45,012
[3.320192534 e-27 kg]
electron configuration
1s1
Deuterium
01
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
01
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1
Alkali Metal
Aufbau
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
s
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.01 - Deuterium atom

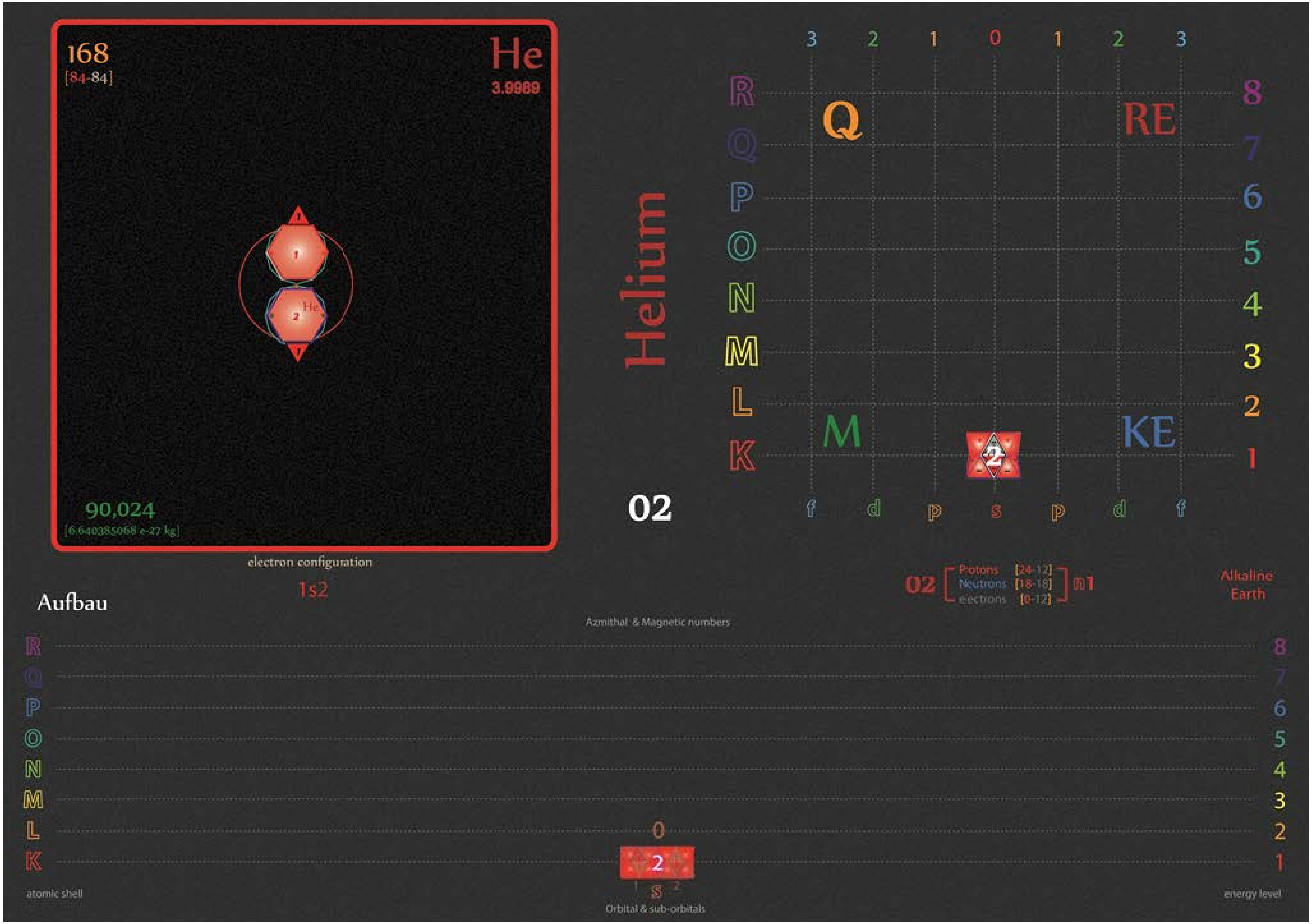
168
[84-84]
He
3.9989
1
1
He
2
1
90,024
[6.640385068 e-27 kg]
electron configuration
1s2
Aufbau
Helium
02
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
2
f d p s p d f
02
Potons [24-12]
Neutrons [18-18]
eectrons [0-12]
m1
Alkaline Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
1 s 2
atomic shell
Orbital & sub-orbitals
energy level

Tetryonics 51.02 - Helium atom

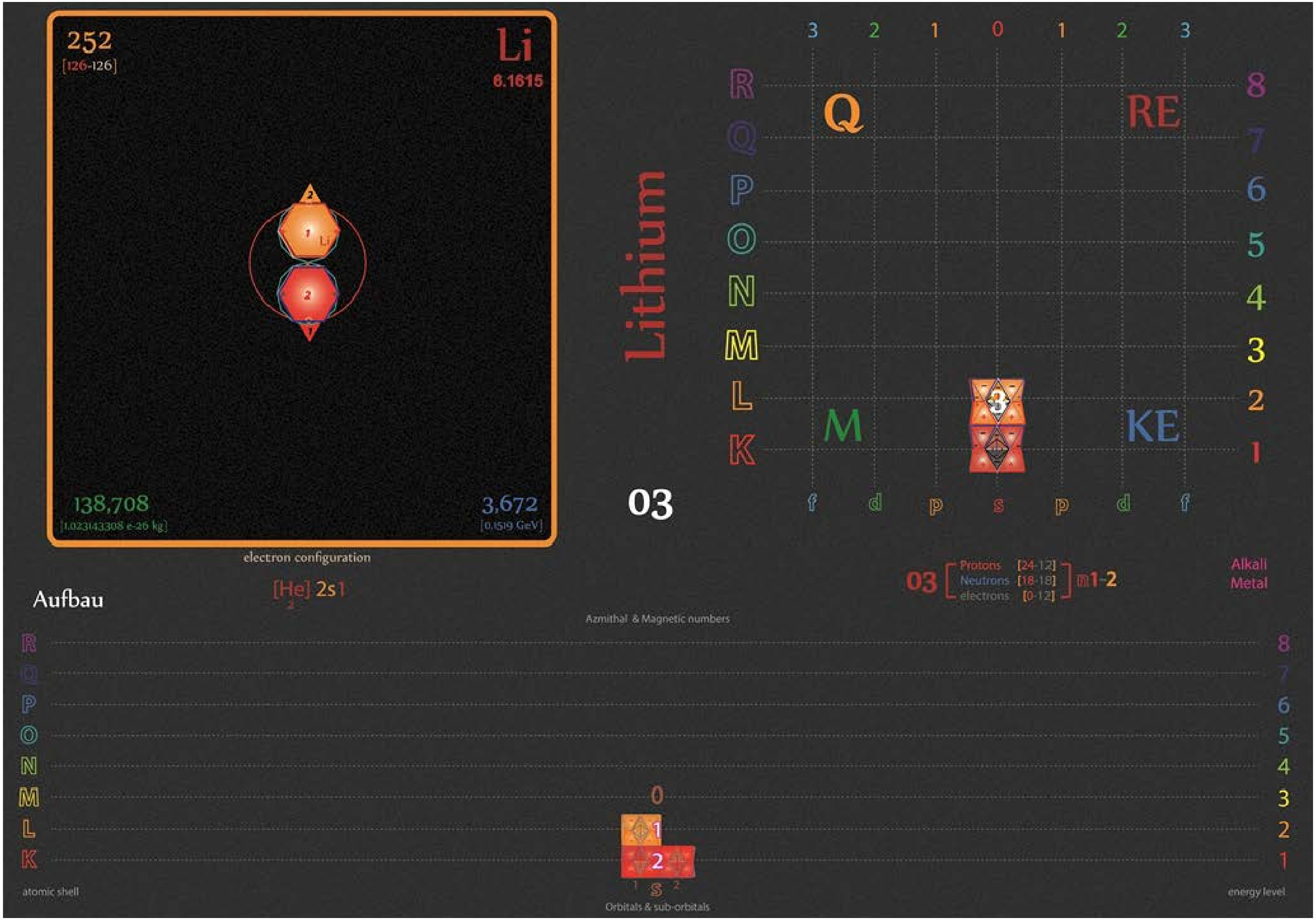
252
[126-126]
Li
6.1615
Li
1
2
138,708
[1.023143308 e-26 kg]
3,672
[0.1519 GeV]
electron configuration
[He] 2s1
Aufbau
Lithium
03
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
3
f d p s p d f
03 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Alkali Metal
Azmithal & Magnetic numbers
0
1
2
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.03 - Lithium atom

Copyright ABRAHAM [2008] - All rights reserved

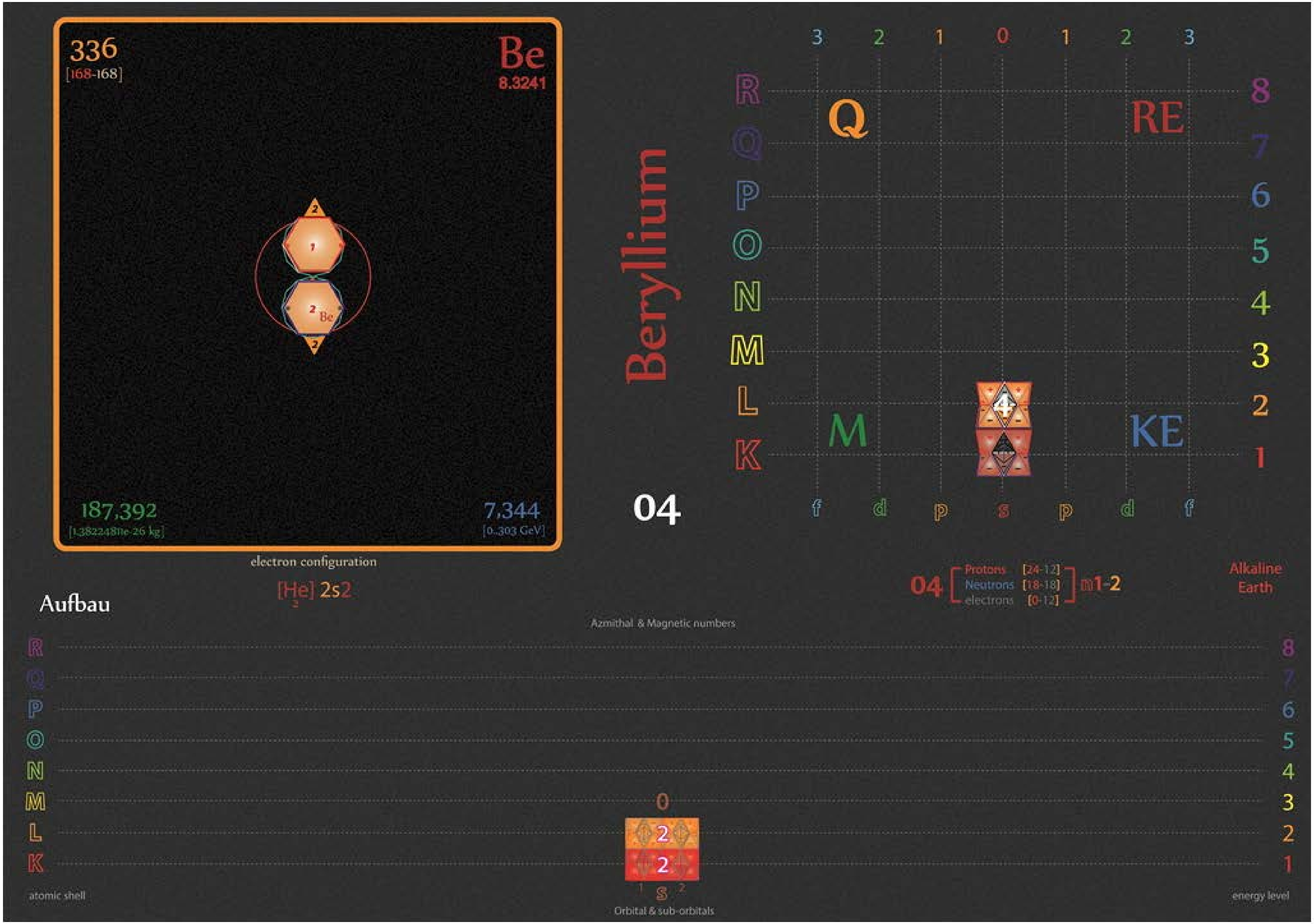
336
[168-168]
Be
8.3241
2
1
2
Be
2
187,392
[1.382248e-26 kg]
7,344
[0..303 GeV]
electron configuration
[He] 2s2
2
Aufbau
Beryllium
04
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
4
8
7
6
5
4
3
2
1
f d p s p d f
04
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Alkaline Earth
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
2
1
s
2
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.04 - Beryllium atom

Copyright ABRAHAM [2008] - All rights reserved

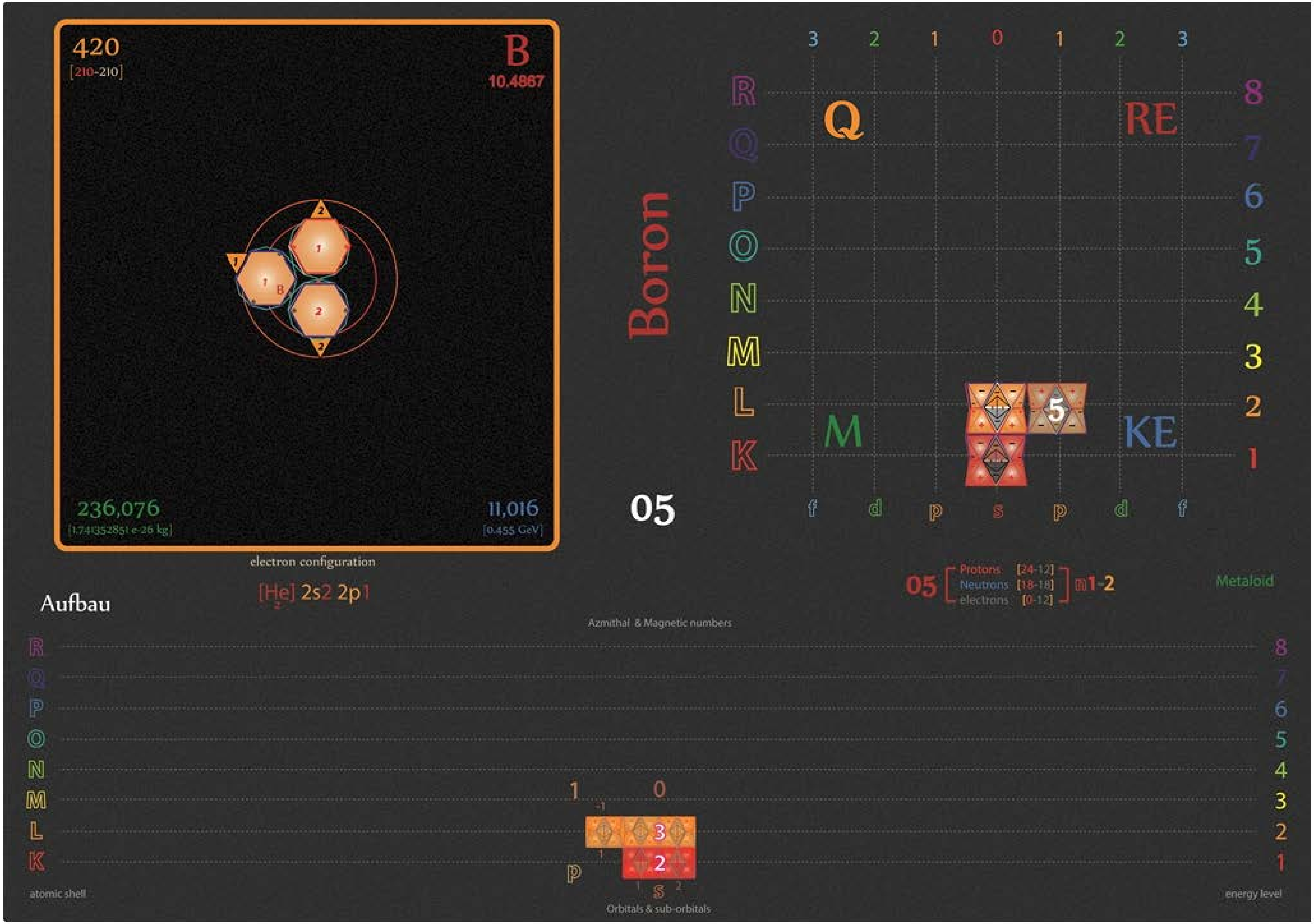
420
[210-210]
B
10.4867
236,076
[1.741352851 e-26 kg]
11,016
[0.455 GeV]
electron configuration
[He] 2s2 2p1
Aufbau
Boron
05
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
5
f d p s p d f
05
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
1
0
-1
3
2
p
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.05 - Boron atom

Copyright ABRAHAM [2008] - All rights reserved

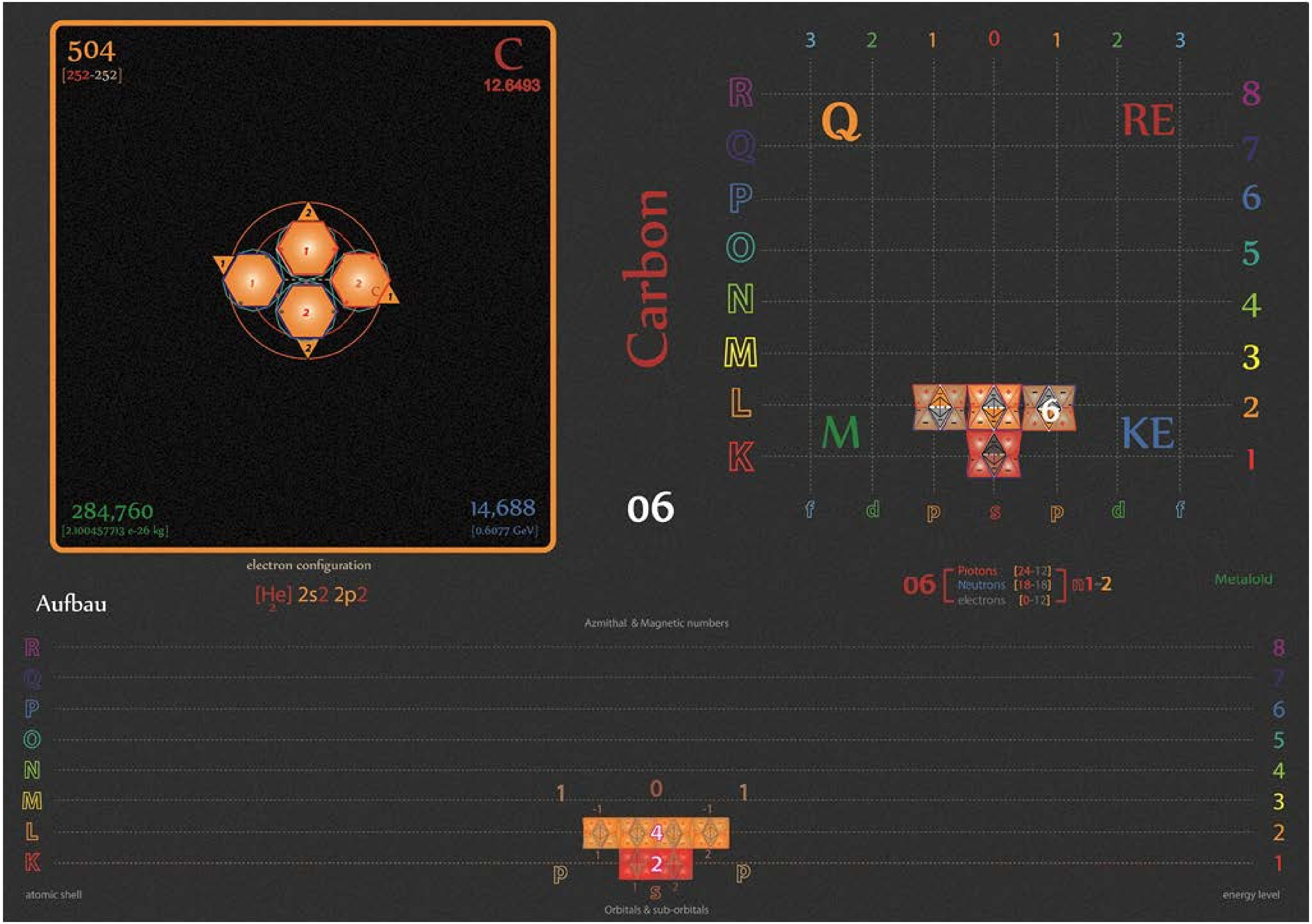
504
[252-252]
C
12.6493
284,760
[2.100457713 e-26 kg]
14,688
[0.6077 GeV]
electron configuration
[He] 2s2 2p2
Aufbau
Carbon
06
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
6
f d p s p d f
06 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Metaloid
Azmithal & Magnetic numbers
1 0 1
-1 -1
4
2
1 2
p p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.06 - Carbon atom

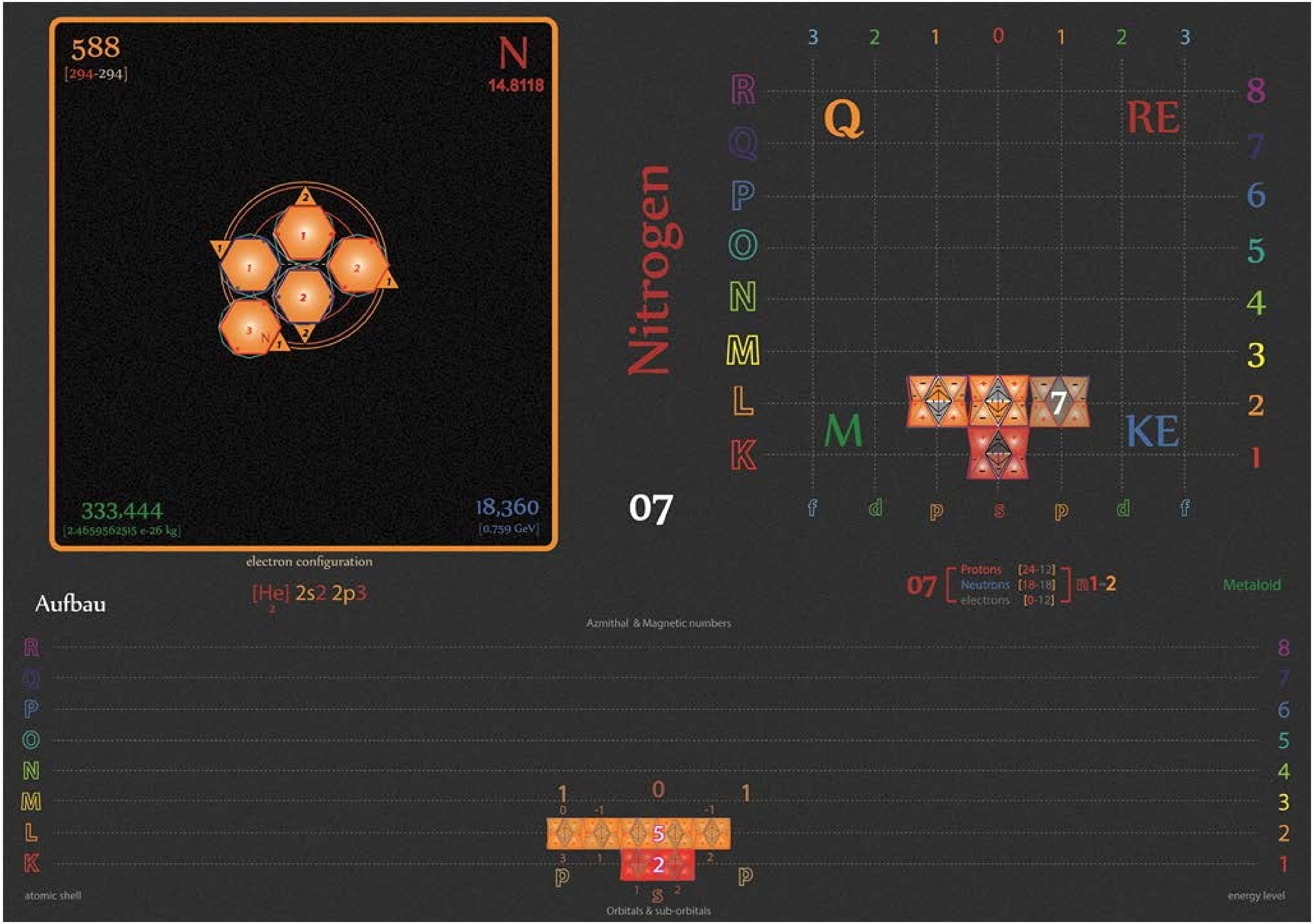
588
[294-294]
N
14.8118
333.444
[2.4659562515 e-26 kg]
18,360
[0.759 GeV]
Nitrogen
07
electron configuration
[He] 2s2 2p3
2
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
7
f d p s p d f
07
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
0 -1 -1
5
2
3 1 2
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.07 - Nitrogen atom

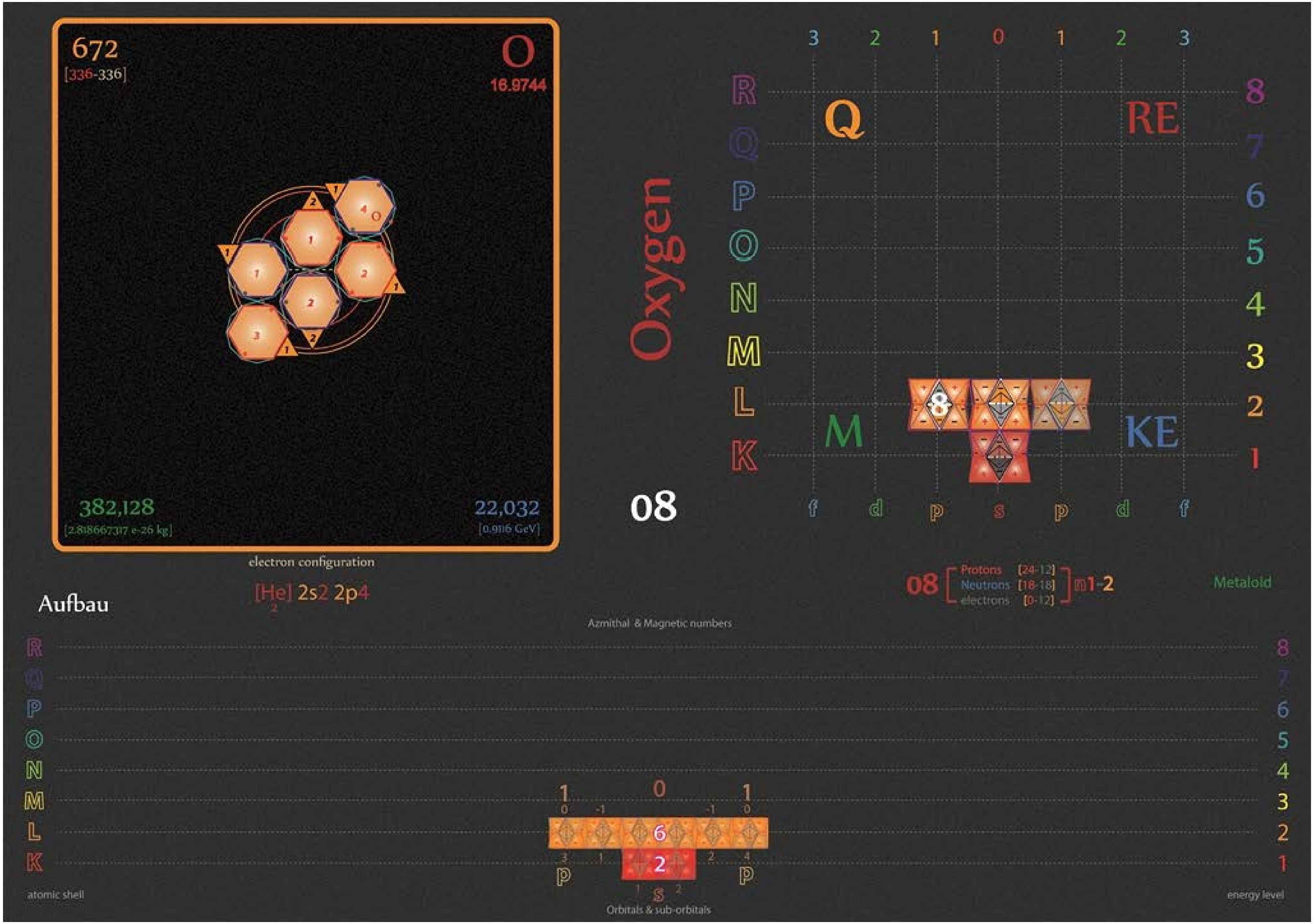
672
[336-336]
O
16.9744
382,128
[2.818667317 e-26 kg]
22,032
[0.9116 GeV]
electron configuration
[He] 2s2 2p4
2
Aufbau
Oxygen
08
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
8
f d p s p d f
08
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-2
Metaloid
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
0 -1 -1 0
6
2
3 1 2 4
p p
1 s 2
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.08 - Oxygen atom

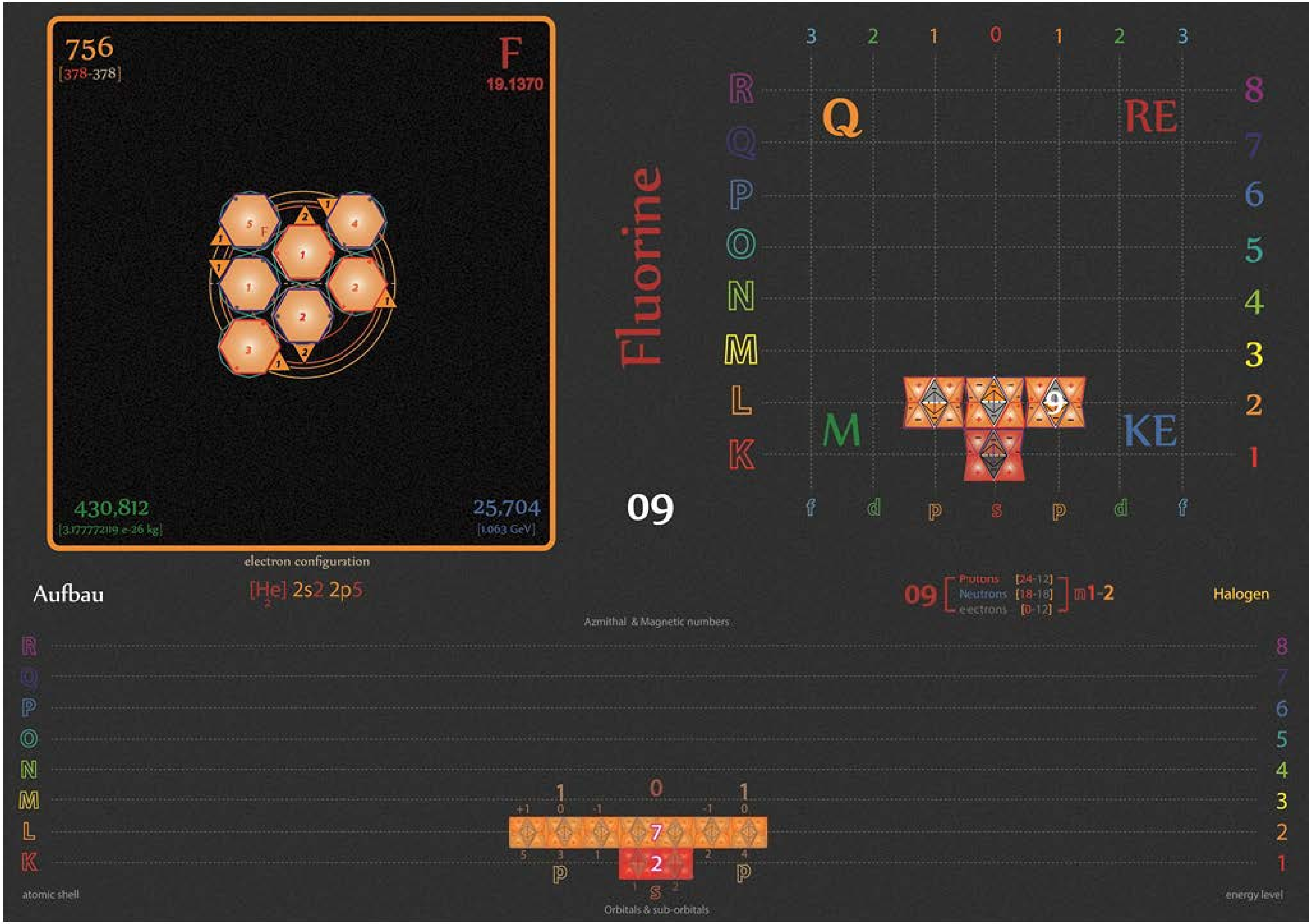
756
[378-378]
F
19.1370
430,812
[3.17777219 e-26 kg]
25,704
[1.063 GeV]
Fluorine
09
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
9
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[He] 2s2 2p5
09 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Halogen
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0
7
2
5 3 1 2 4
p s p
1 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.09 - Fluorine atom

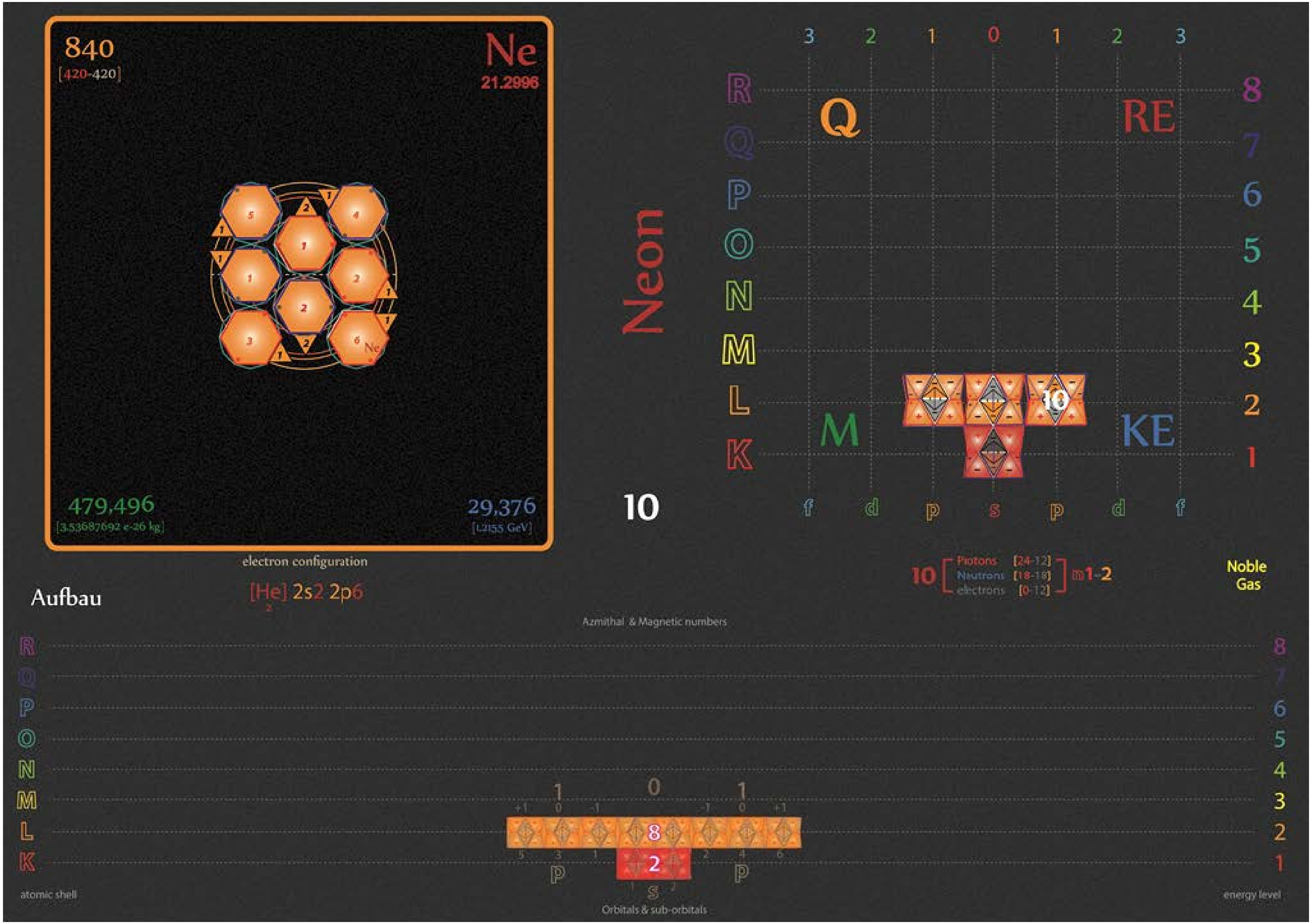
840
[420-420]
Ne
21.2996
479,496
[3.53687692 e-26 kg]
29,376
[1.2155 GeV]
electron configuration
[He] 2s2 2p6
2
Aufbau
Neon
10
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
10
f d p s p d f
10 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-2
Noble Gas
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.10 - Neon atom

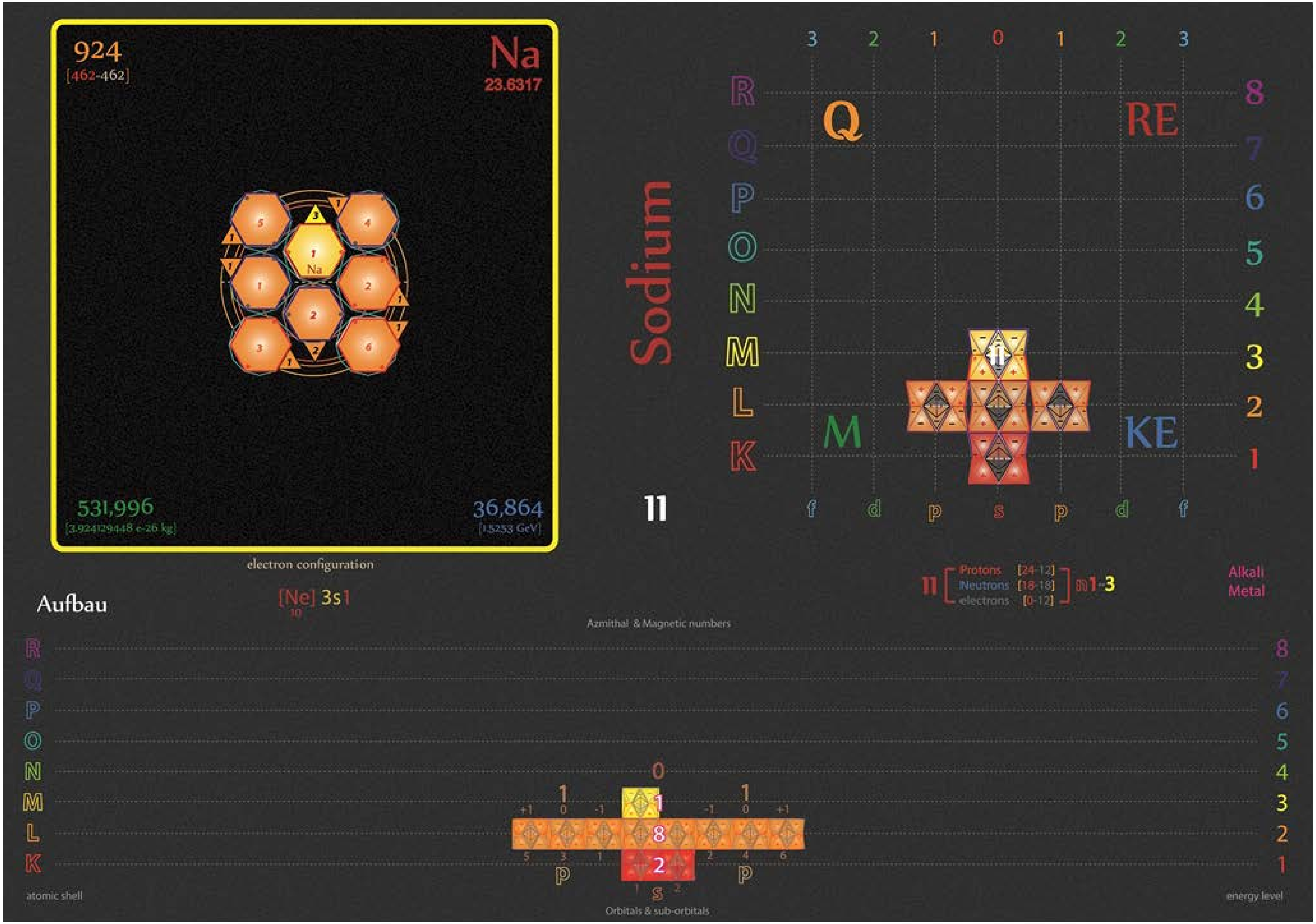
924
[462-462]
Na
23.6317
1
Na
5
4
1
2
2
3
6
531,996
[3.924129448 e-26 kg]
36,864
[1.5253 GeV]
electron configuration
Aufbau
[Ne] 3s1
10
Sodium
11
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
11 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Alkali Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
+1 0 -1
-1 0 +1
1
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.11 - Sodium atom

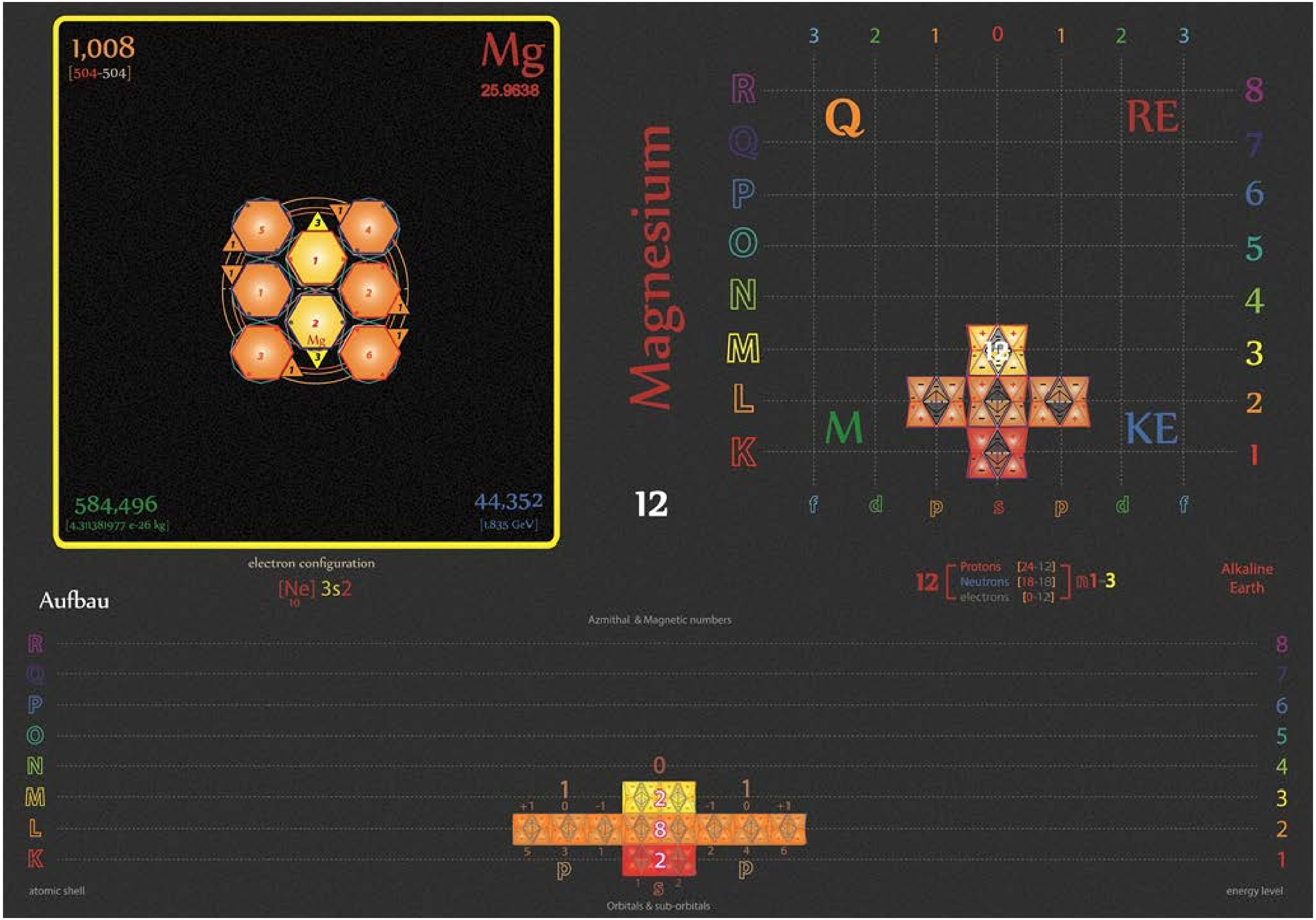
1,008
[504-504]
Mg
25.9638
5
4
3
1
1
2
2
Mg
3
6
584,496
[4.31138977 e-26 kg]
44,352
[1.835 GeV]
electron configuration
[Ne] 3s2
10
Aufbau
Magnesium
12
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
12
f d p s p d f
12
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Alkaline
Earth
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
+1 0 -1
-1 0 +1
2
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.12 - Magnesium atom

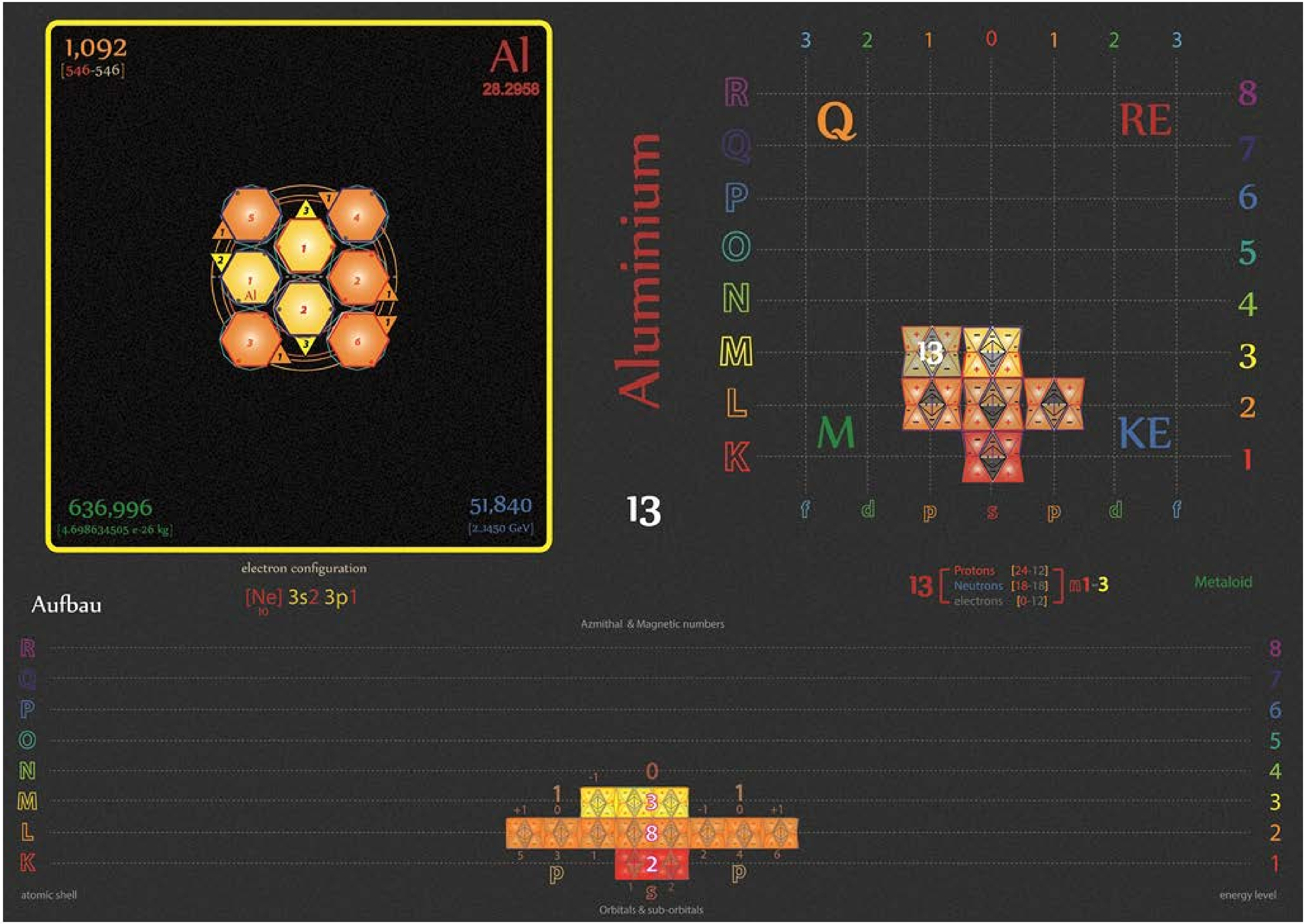
1,092
[546-546]
Al
28.2958
636,996
[4.698634505 e-26 kg]
51,840
[2.3450 GeV]
electron configuration
[Ne] 3s2 3p1
10
Aufbau
Aluminium
13
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
13 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
Azmithal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals
-1 0 1 1
+1 0 -1 0 +1
3 8 2
5 3 1 2 4 6
p s p
1 2

Tetryonics 51.13 - Aluminium atom

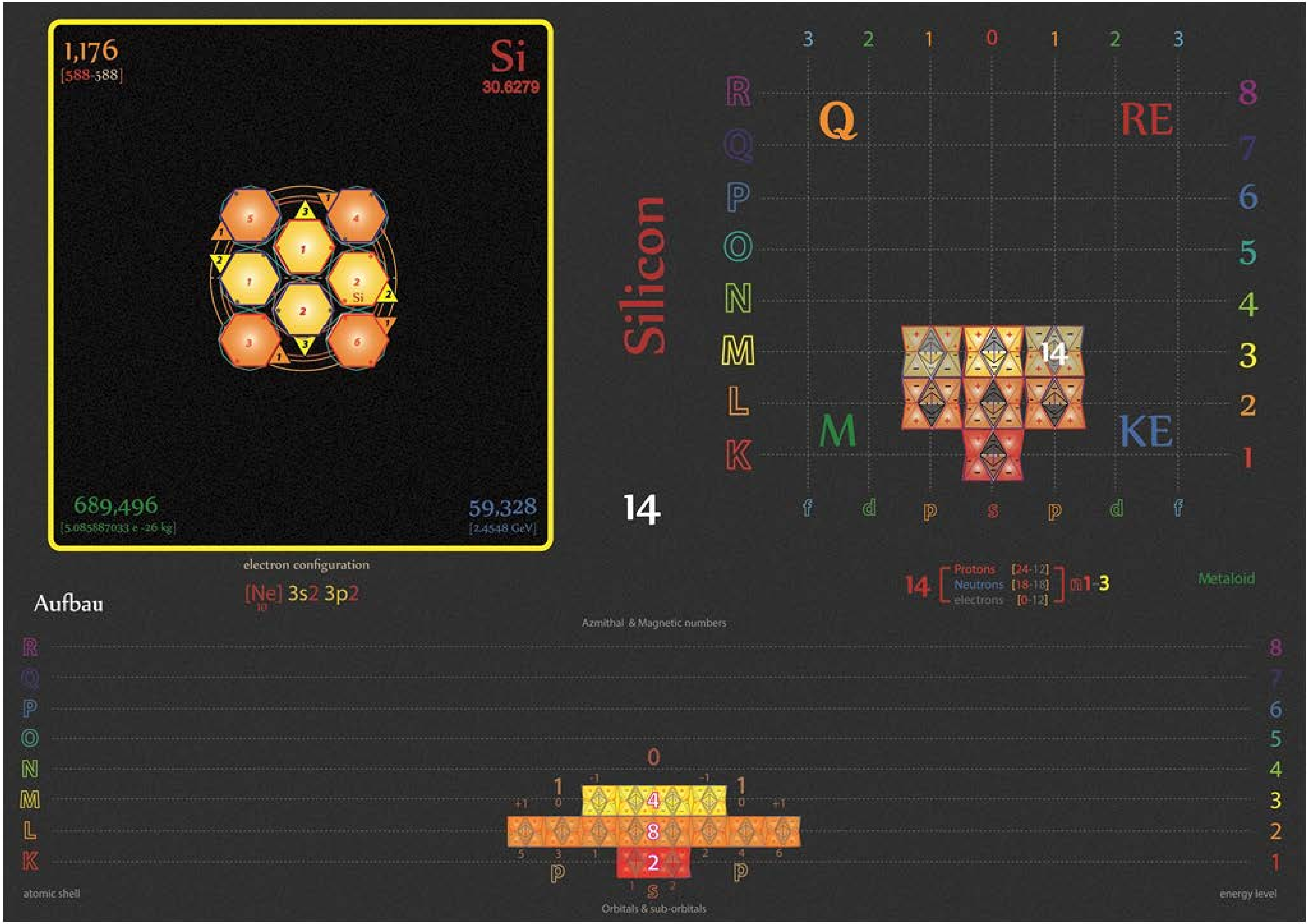
1,176
[588-588]
Si
30.6279
Si
689,496
[5.085887033 e -26 kg]
59,328
[2.4548 GeV]
Silicon
14
electron configuration
[Ne] 3s2 3p2
10
Aufbau
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
14
8
7
6
5
4
3
2
1
f d p s p d f
14
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-3
Metaloid
Azmithal & Magnetic numbers
0
1
1
-1
-1
+1
0
0
+1
4
8
2
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.14 - Silicon atom

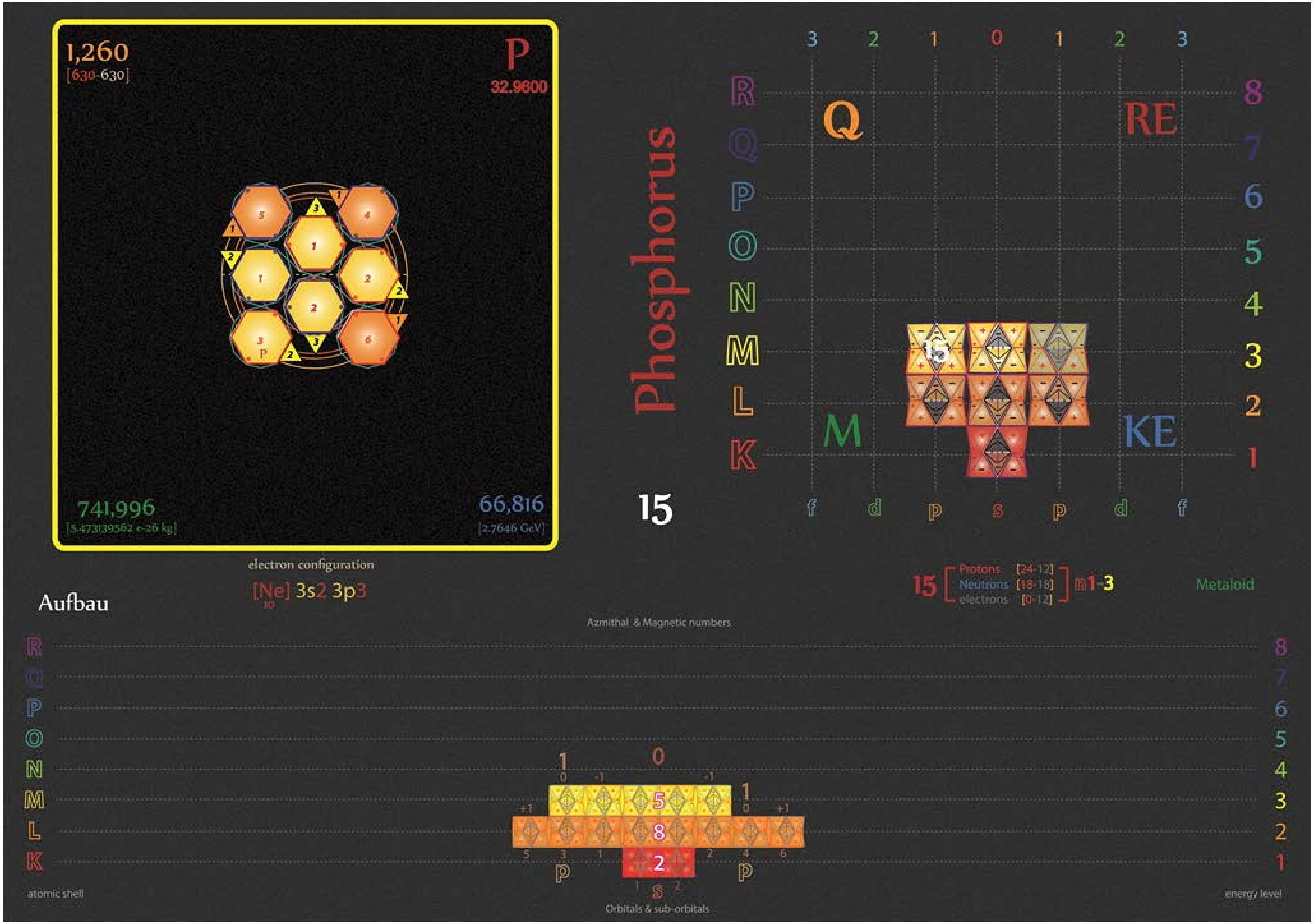
1,260
[630-630]
P
32.9600
741,996
[5.473139562 e-26 kg]
66,816
[2.7646 GeV]
Phosphorus
15
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Ne] 3s2 3p3
10
Aufbau
15 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Metaloid
Azmithal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals
5
8
2
p
s
p

Tetryonics 51.15 - Phosphorus atom

Copyright ABRAHAM [2008] - All rights reserved

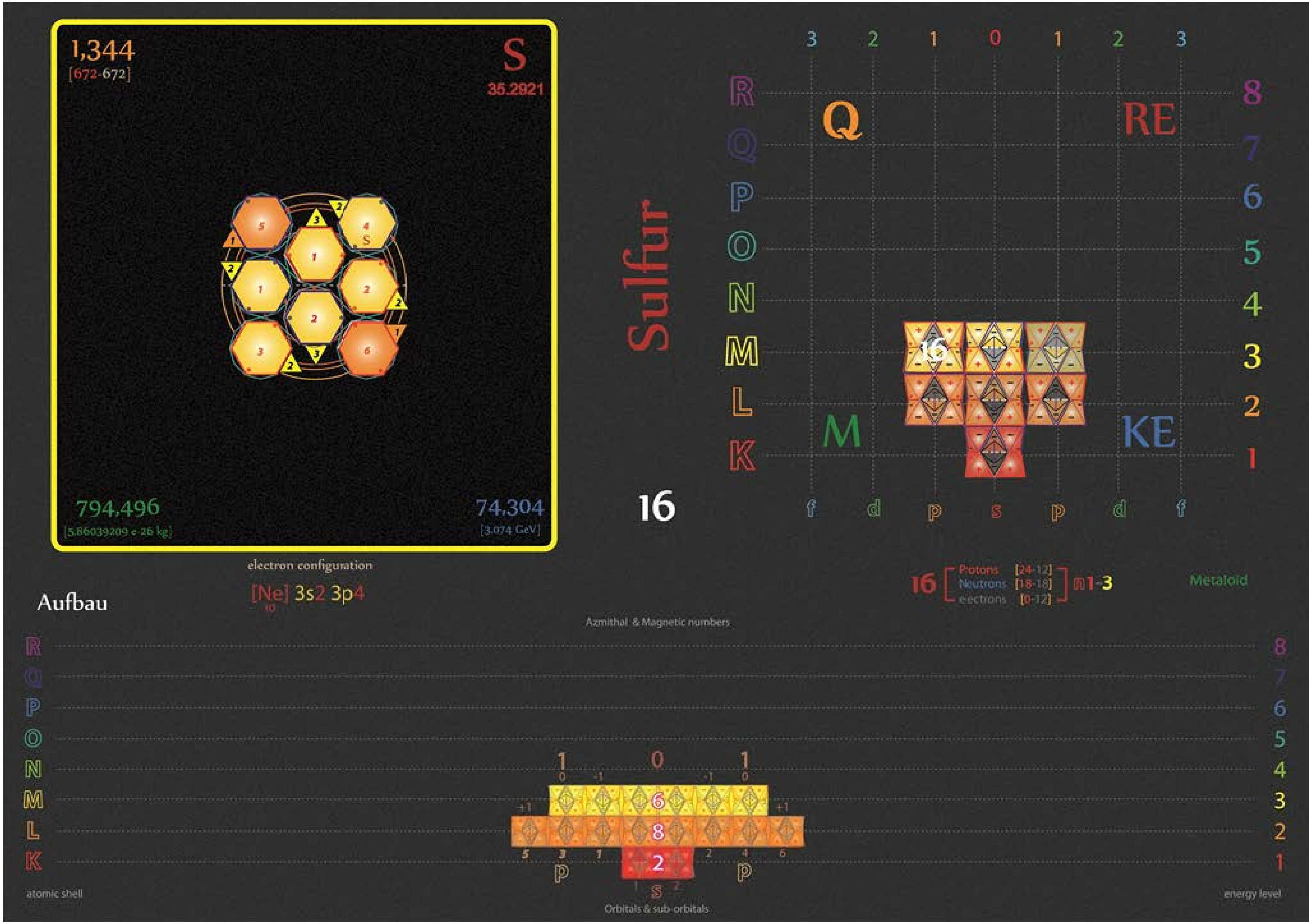
1,344
[672-672]
S
35.2921
794,496
[5.86039209 e-26 kg]
74,304
[3.074 GeV]
Sulfur
16
electron configuration
[Ne] 3s2 3p4
10
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
16 [Potons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Metaloid
Azmithal & Magnetic numbers
1 0 1
0 -1 -1 0
+1 +1
6
8
2
5 3 1 2 4 6
p s p
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.16 - Sulfur atom

Copyright ABRAHAM [2008] - All rights reserved

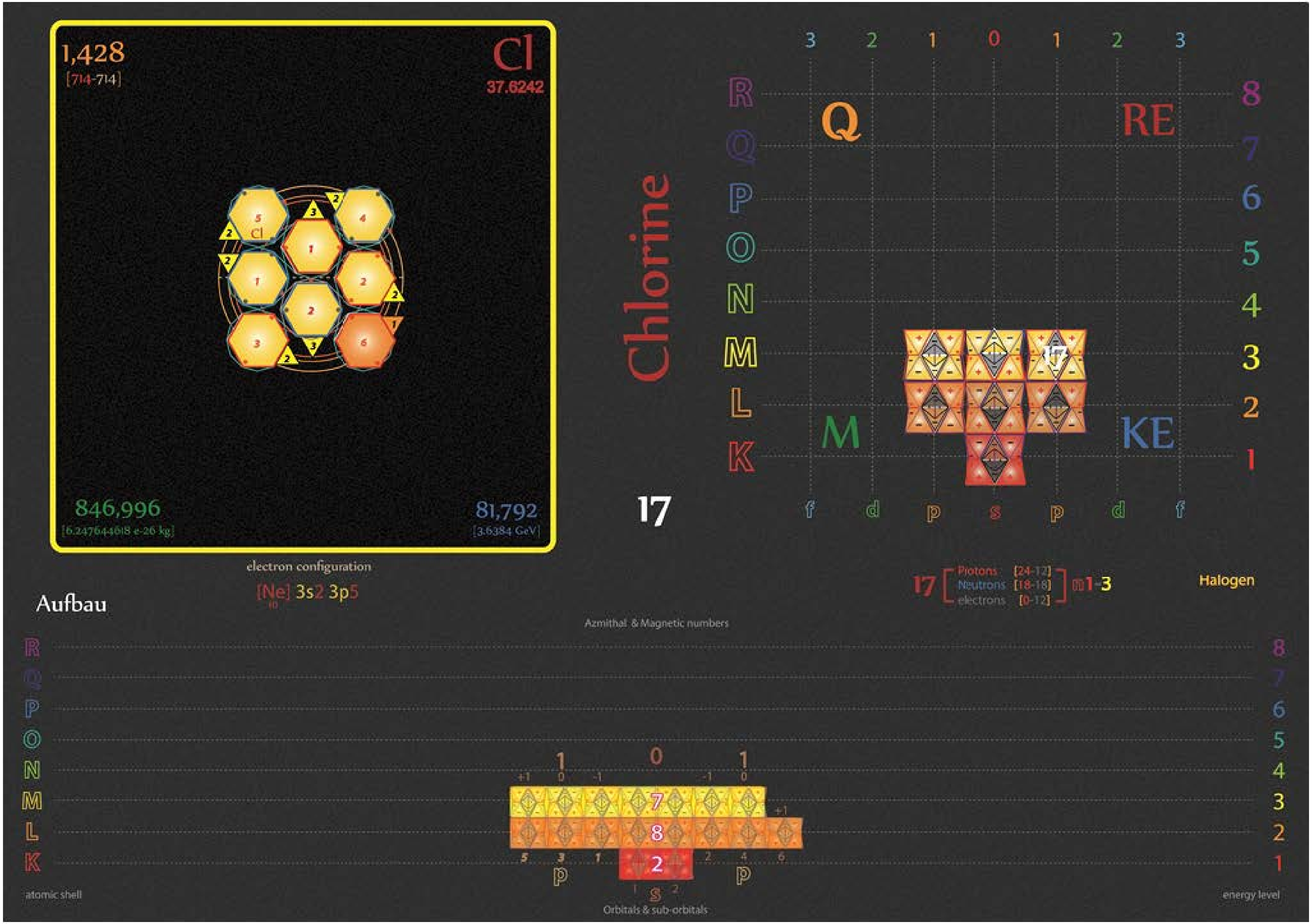
1,428
[714-714]
Cl
37.6242
5
Cl
4
1
1
2
2
3
6
846,996
[6.247644618 e-26 kg]
81,792
[3.6384 GeV]
electron configuration
[Ne] 3s2 3p5
10
Aufbau
Chlorine
17
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
17
f d p s p d f
17 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-3
Halogen
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0
+1
7
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.17 - Chlorine atom

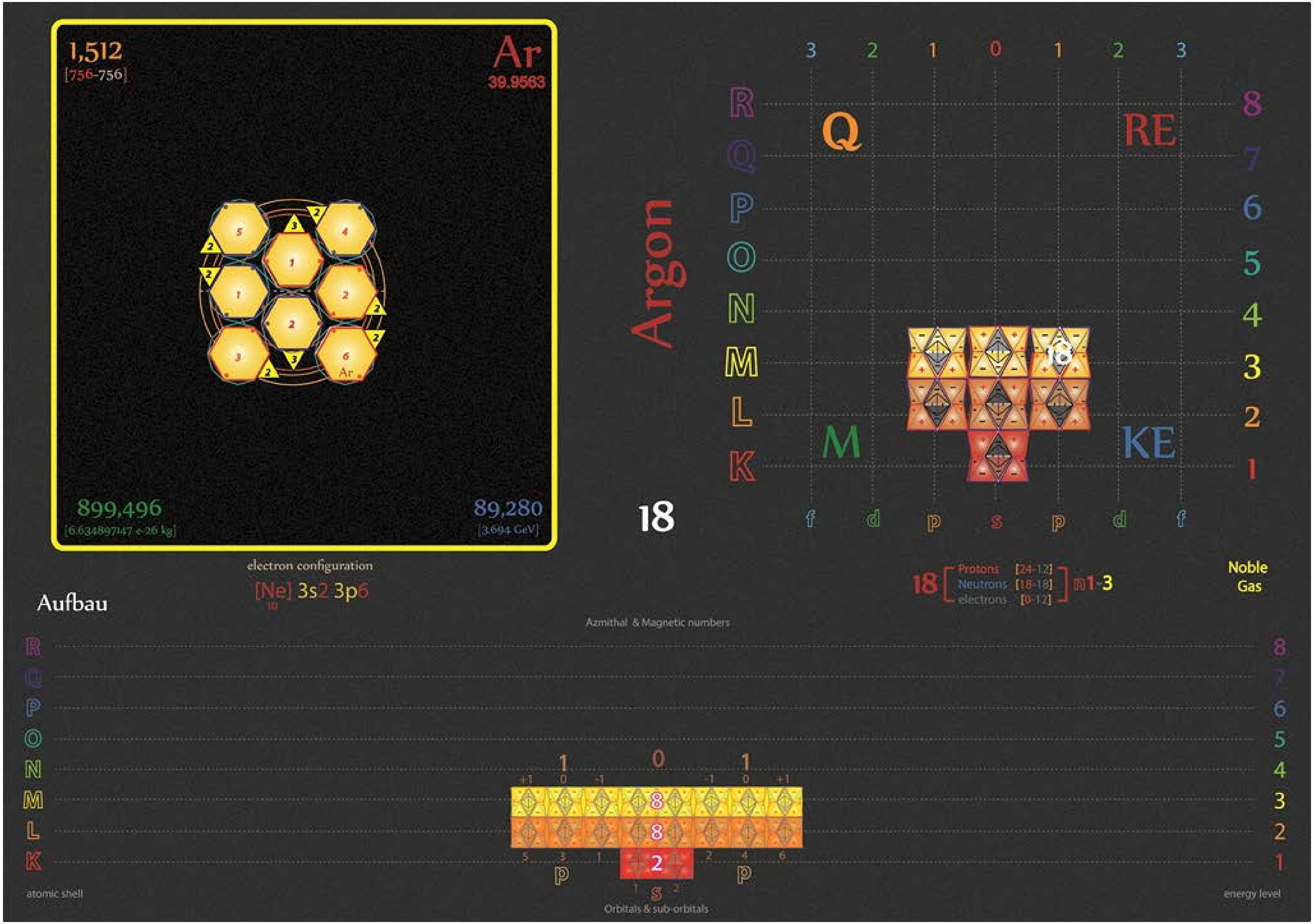
1,512
[756-756]
Ar
39.9563
899,496
[6.634897147 e-26 kg]
89,280
[3.694 GeV]
Argon
18
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
18 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-3
Noble Gas
electron configuration
[Ne] 3s2 3p6
10
Aufbau
Azmithal & Magnetic numbers
1 0 1
+1 0 -1 -1 0 +1
8
8
2
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.18 - Argon atom

Copyright ABRAHAM [2008] - All rights reserved

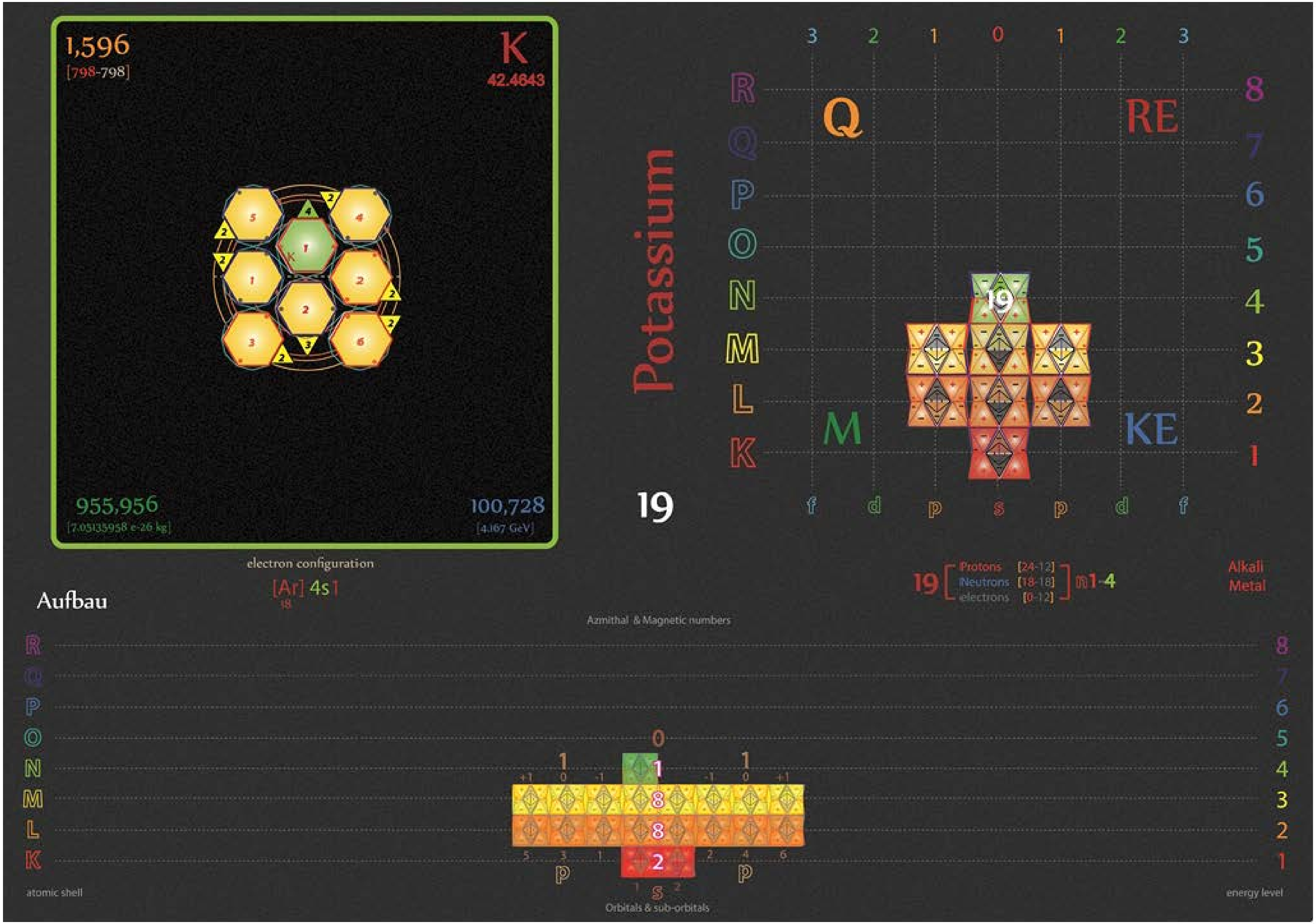
1,596
[798-798]
K
42.4643
955,956
[7.05135958 e-26 kg]
100,728
[4.167 GeV]
Potassium
19
electron configuration
[Ar] 4s1
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
19
f d p s p d f
19 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Alkali Metal
Azmithal & Magnetic numbers
0
1
1
1
+1 0 -1
-1 0 +1
8
8
2
5 3 1
2 4 6
p
p
1 s 2
atomic shell
Orbitals & sub-orbitals
energy level

Tetryonics 51.19 - Potassium atom

Copyright ABRAHAM [2008] - All rights reserved

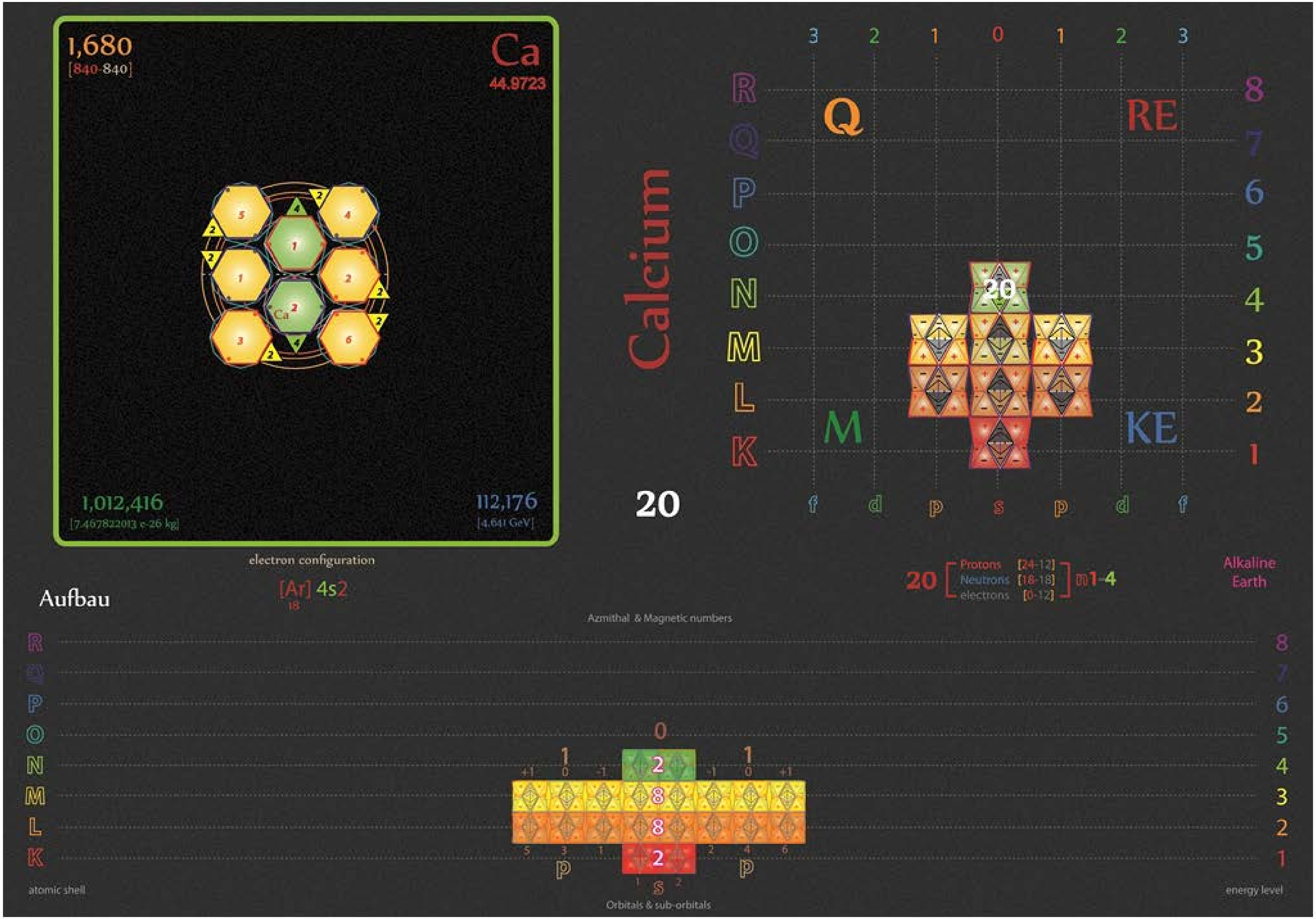
1,680
[840-840]
Ca
44.9723
1
2
1
2
3
4
5
6
Ca
1,012,416
[7.467822013 e-26 kg]
112,176
[4.641 GeV]
Calcium
20
electron configuration
[Ar] 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
20
20 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Alkaline Earth
Azmithal & Magnetic numbers
0
1
1
+1 0 -1
-1 0 +1
2
8
8
2
5 3 1
2 4 6
p
p
1 2
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.20 - Calcium atom

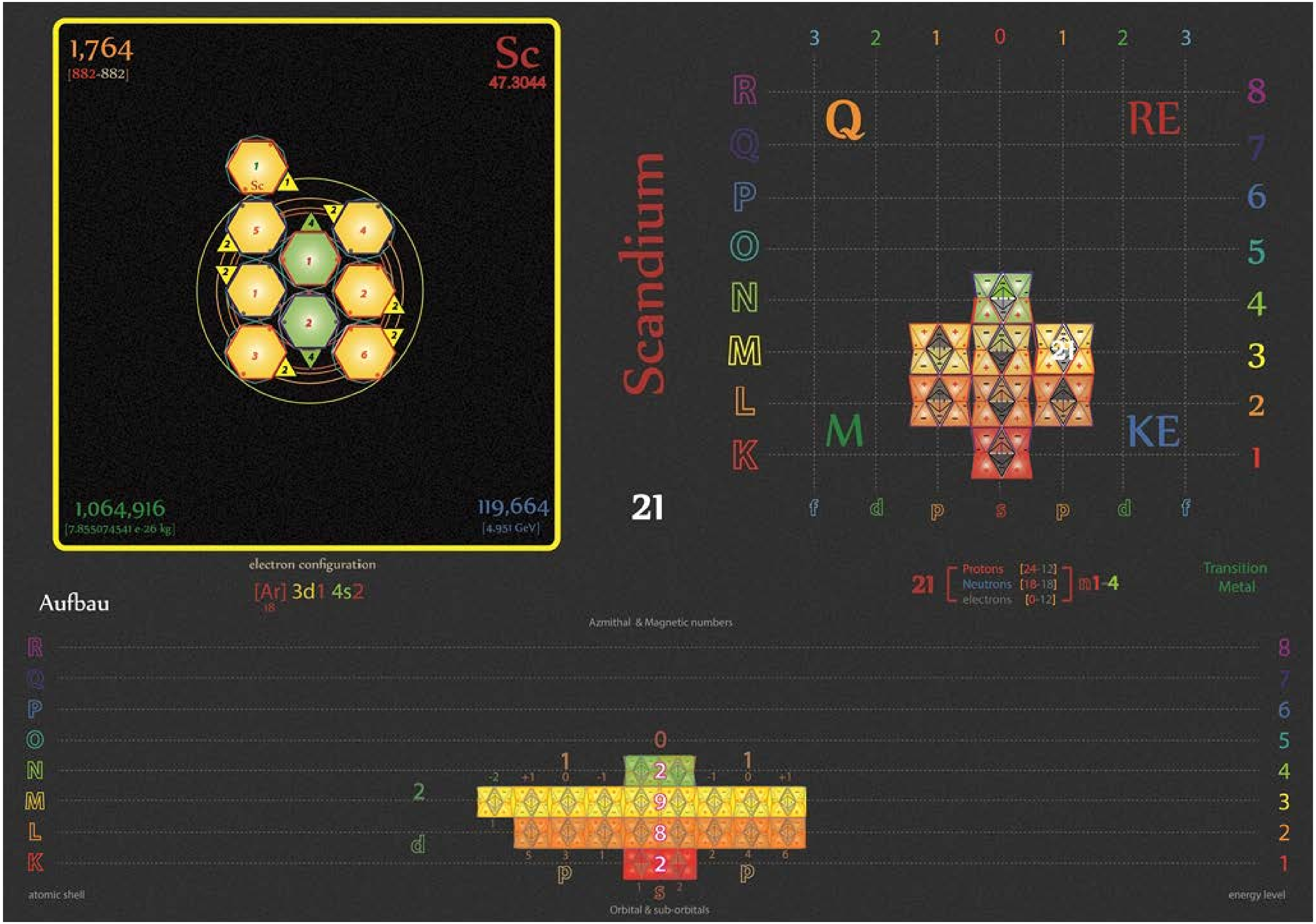
1,764
[882-882]
Sc
47.3044
1,064,916
[7.855074541 e-26 kg]
119,664
[4.951 GeV]
Scandium
21
electron configuration
[Ar] 3d1 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
21
21 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
d
-2 +1 0 -1 -1 0 +1
2
9
8
2
1 5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.21 - Scandium atom

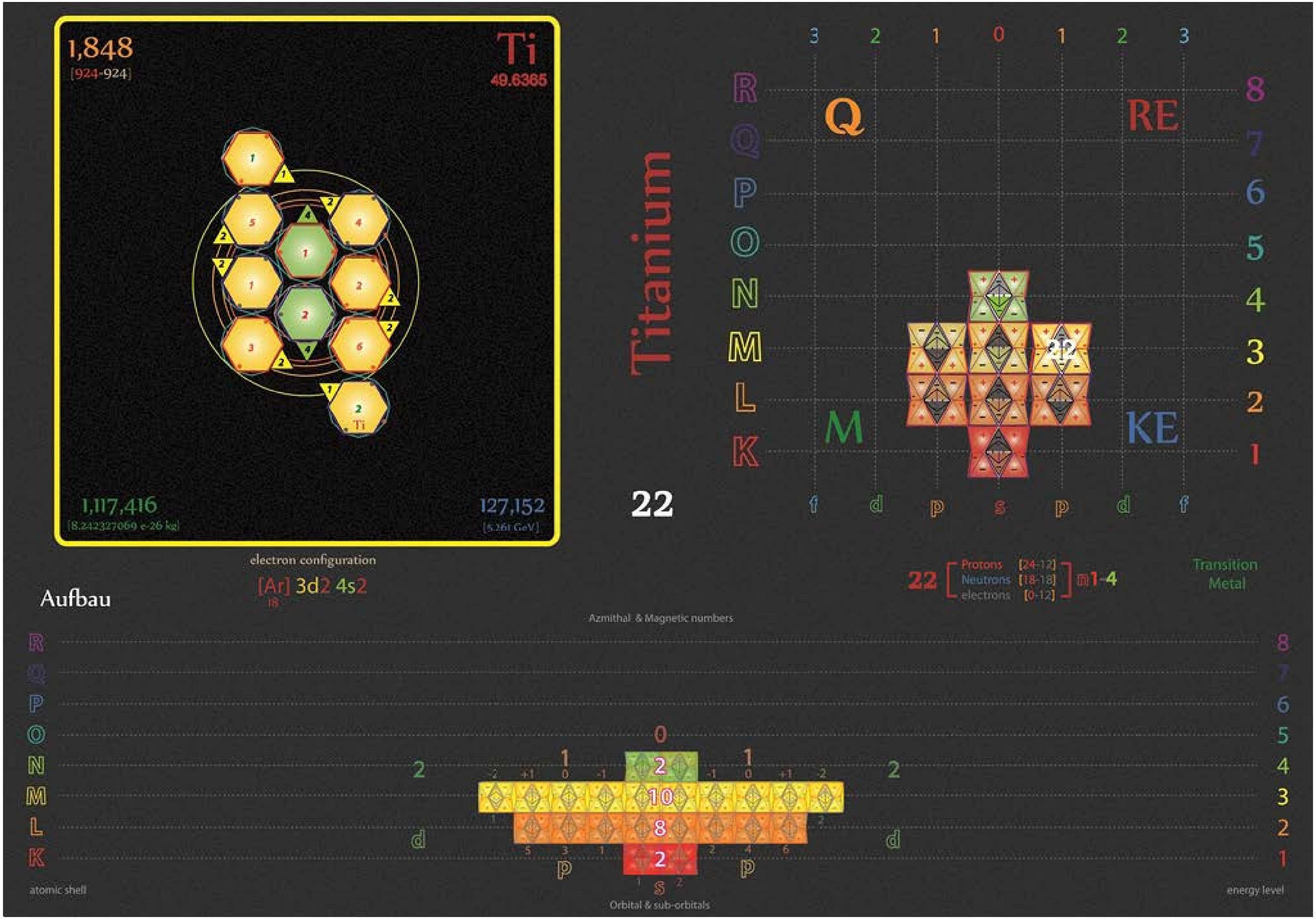
1,848
[924-924]
Ti
49.6365
1,117,416
[8.242327069 e-26 kg]
127,152
[5.261 GeV]
Titanium
22
electron configuration
[Ar] 3d2 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
22
f d p s p d f
22 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
1 1
2 2
2
10
8
2
d d
p p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.22 - Titanium atom

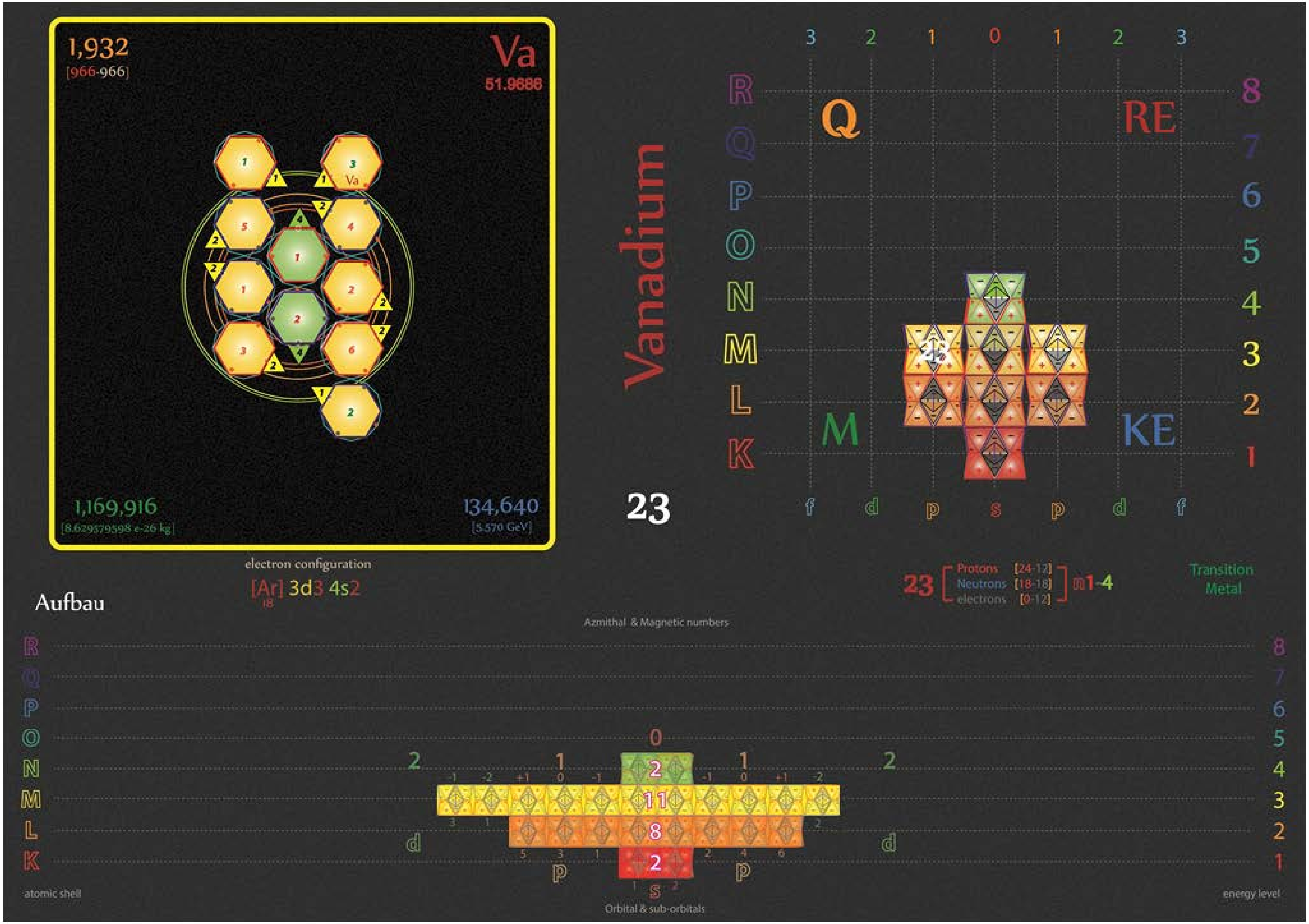
1,932
[966-966]
Va
51.9888
1,169,916
[8.629579598 e-26 kg]
134,640
[5.570 GeV]
Vanadium
23
electron configuration
[Ar] 3d3 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
23 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2
1
1
2
-1 -2 +1 0 -1 -1 0 +1 -2
2
11
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.23 - Vanadium atom

Copyright ABRAHAM [2008] - All rights reserved

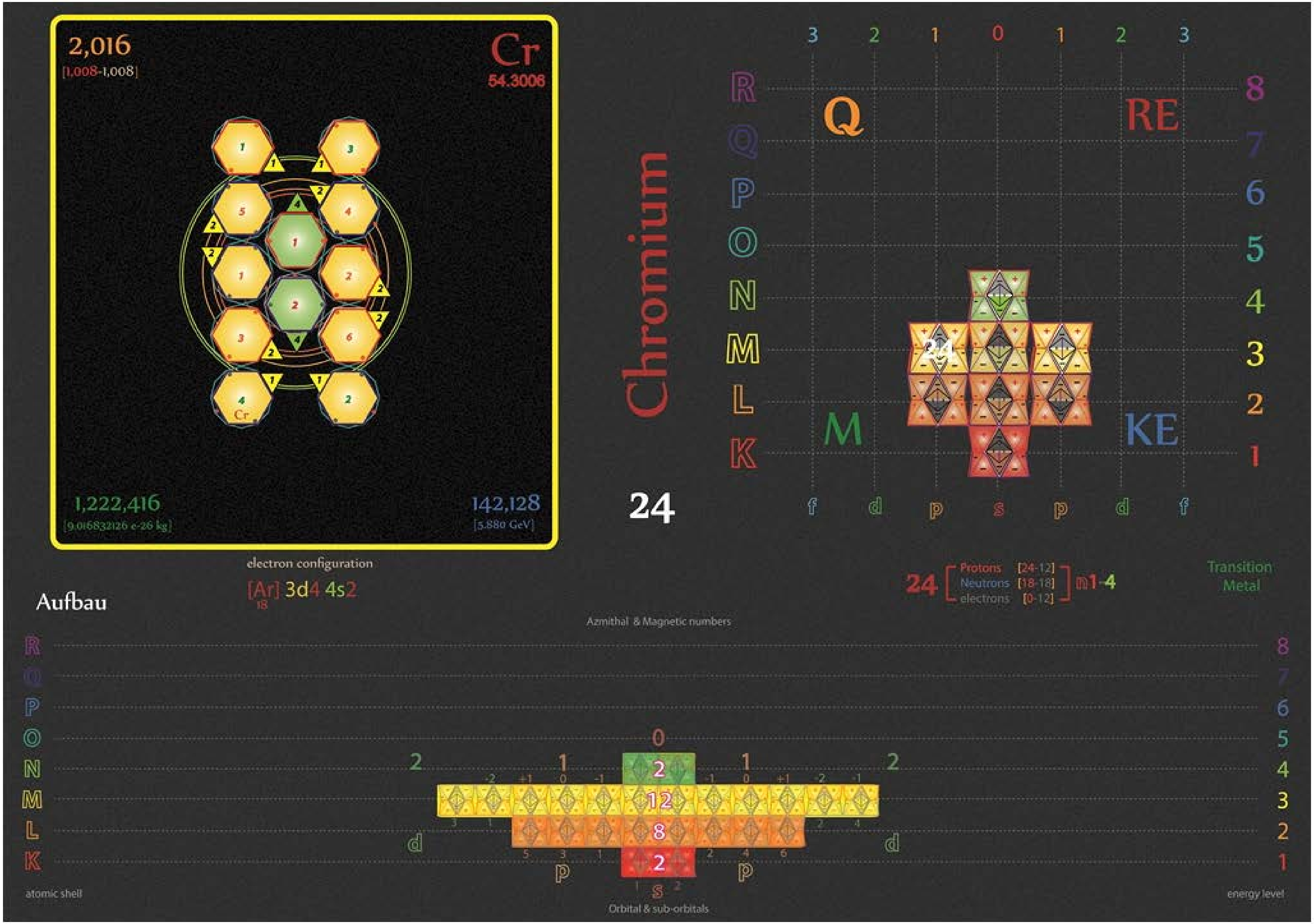
2,016
[1,008-1,008]
Cr
54.3008
Cr
1,222,416
[9.016832126 e-26 kg]
142,128
[5.880 GeV]
Chromium
24
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
24
f d p s p d f
electron configuration
[Ar] 3d4 4s2
18
Aufbau
24
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
-2 +1 0 -1 -1 0 +1 -2 -1
2
12
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.24 - Chromium atom

Copyright ABRAHAM [2008] - All rights reserved

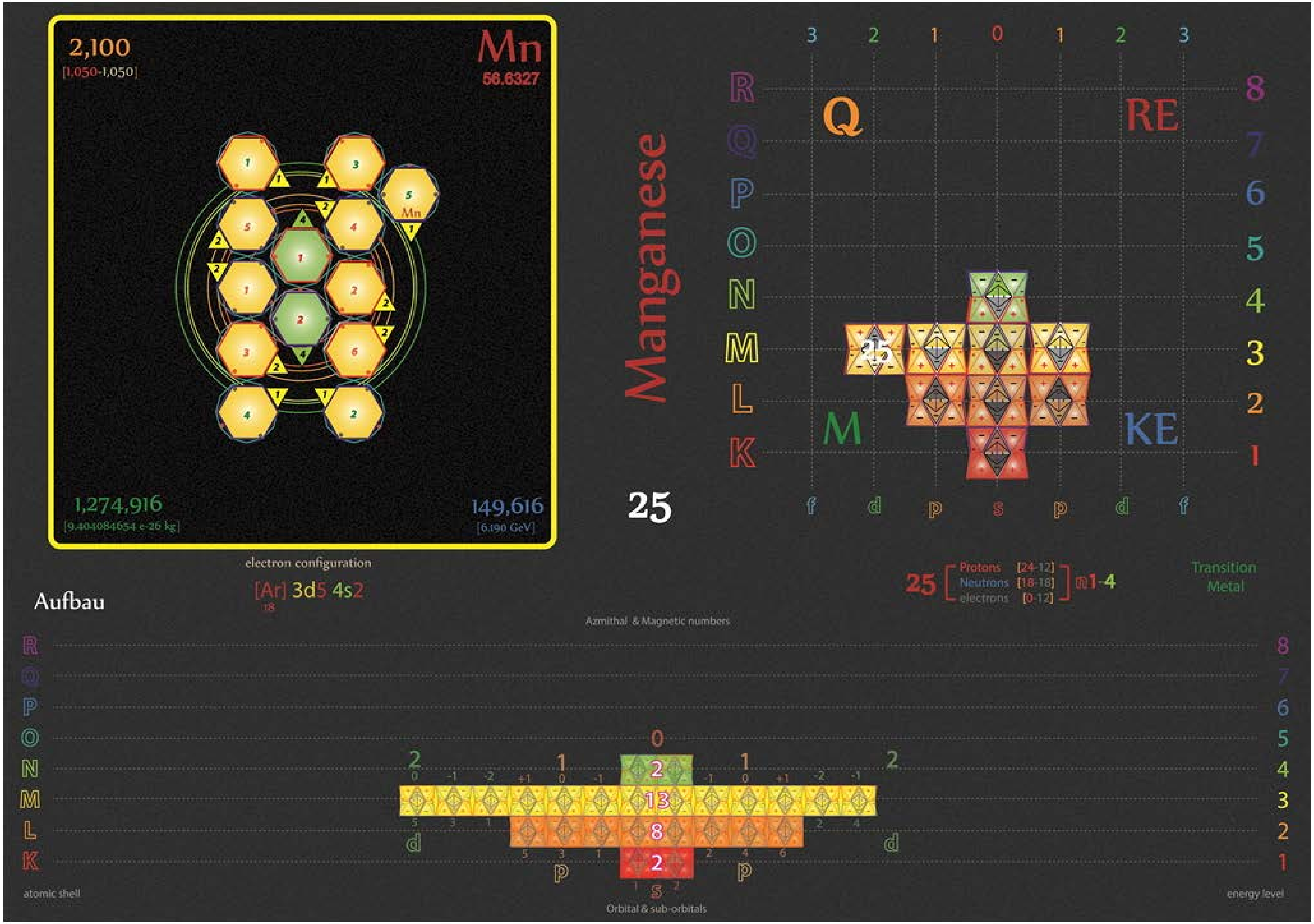
2,100
[1,050-1,050]
Mn
56.6327
1,274,916
[9.404084654 e-26 kg]
149,616
[6.190 GeV]
Manganese
25
electron configuration
[Ar] 3d5 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
25 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
13
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.25 - Manganese atom

Copyright ABRAHAM [2008] - All rights reserved

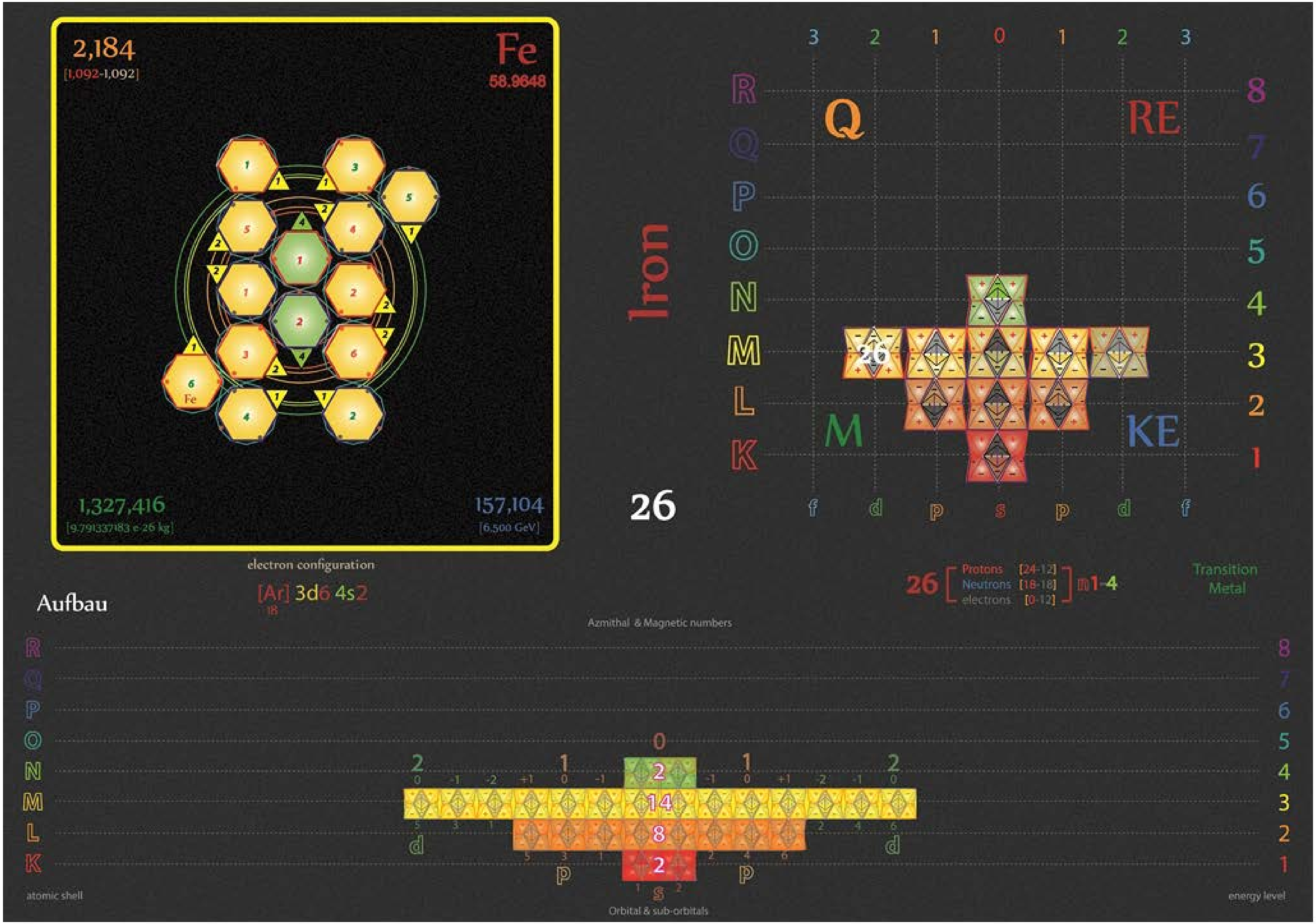
2,184
[1,092-1,092]
Fe
58.9648
1,327,416
[9.791337183 e-26 kg]
157,104
[6,500 GeV]
Iron
26
electron configuration
[Ar] 3d6 4s2
18
Aufbau
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
26 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
14
8
2
d p s p d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.26 - Iron atom

Copyright ABRAHAM [2008] - All rights reserved

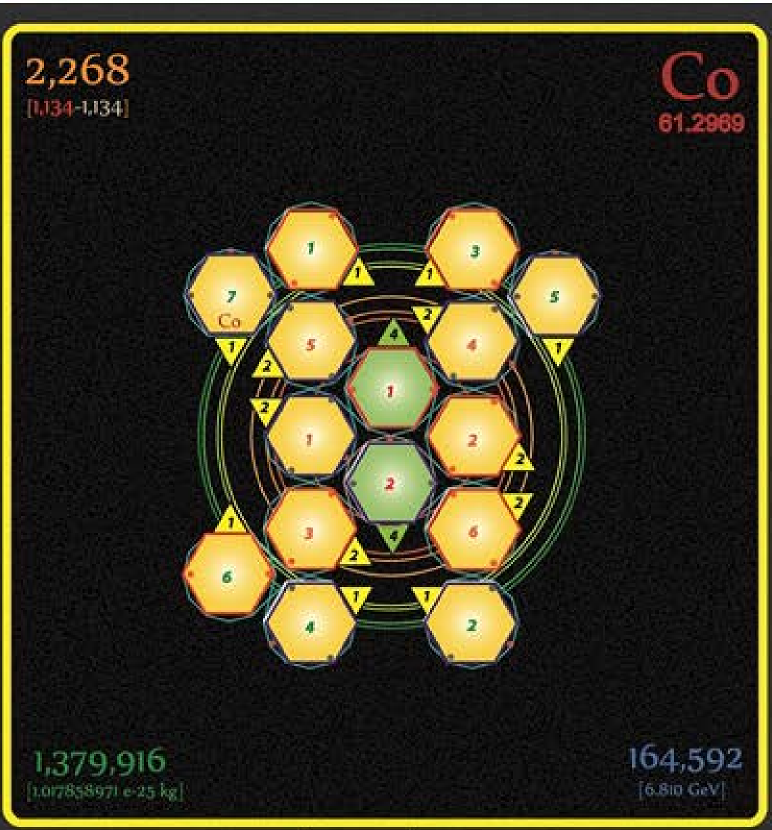

2,268
[1,134-1,134]

Co
61.2969

1,379,916
[1.017858971 e-25 kg]

164,592
[6.810 GeV]

Cobalt

27

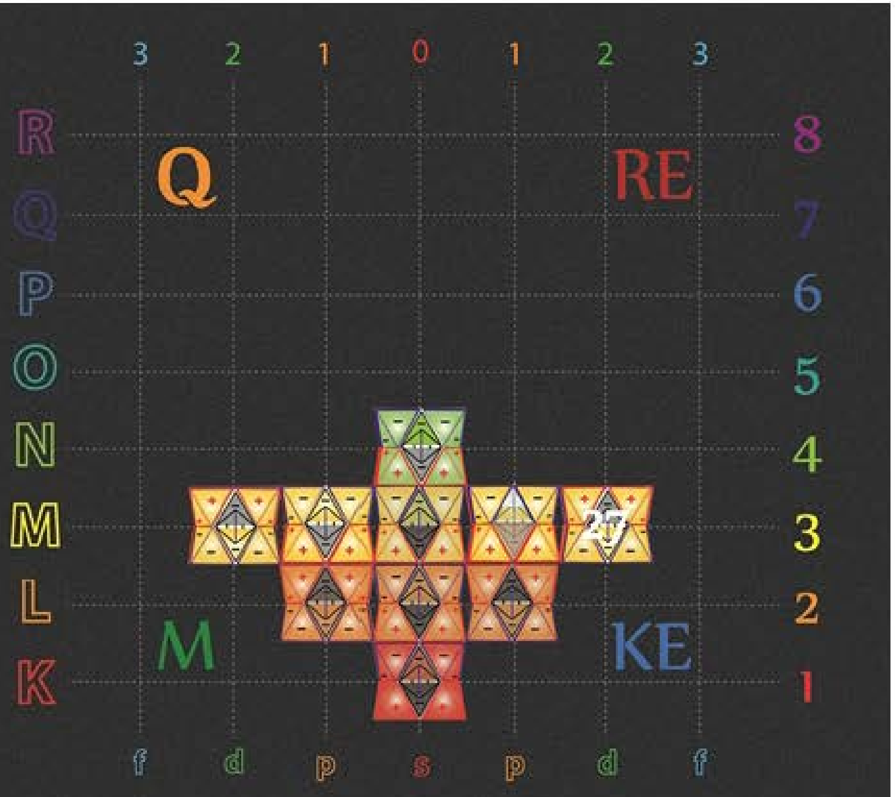
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
27
8
7
6
5
4
3
2
1
f d p s p d f

electron configuration

[Ar] 3d7 4s2
18

Aufbau

27 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Transition Metal

Azmithal & Magnetic numbers

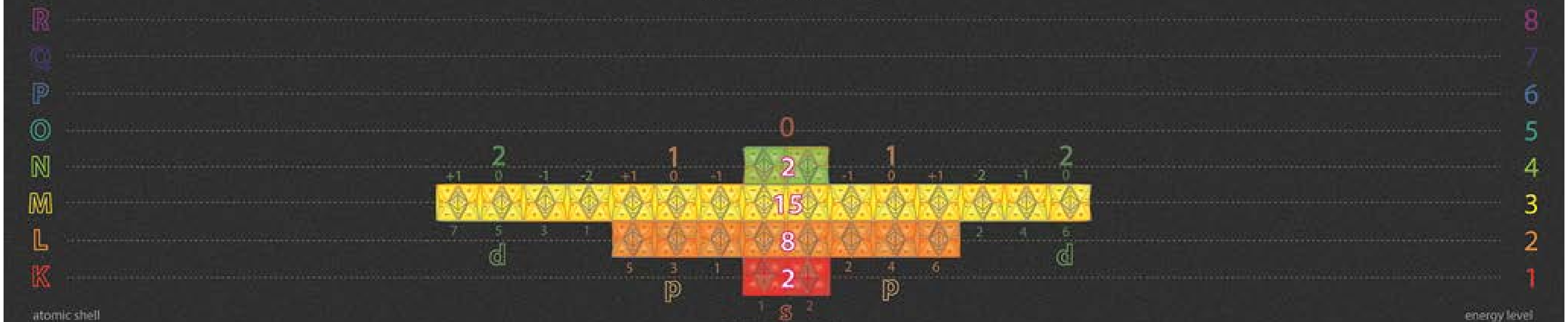
R
Q
P
O
N
M
L
K
0
2
1
1
2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0
2
15
8
2
7 5 3 1 2 4 6
d
d
5 3 1 2 4 6
p
p
1 2
s
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbital & sub-orbitals

Tetryonics 51.27 - Cobalt atom

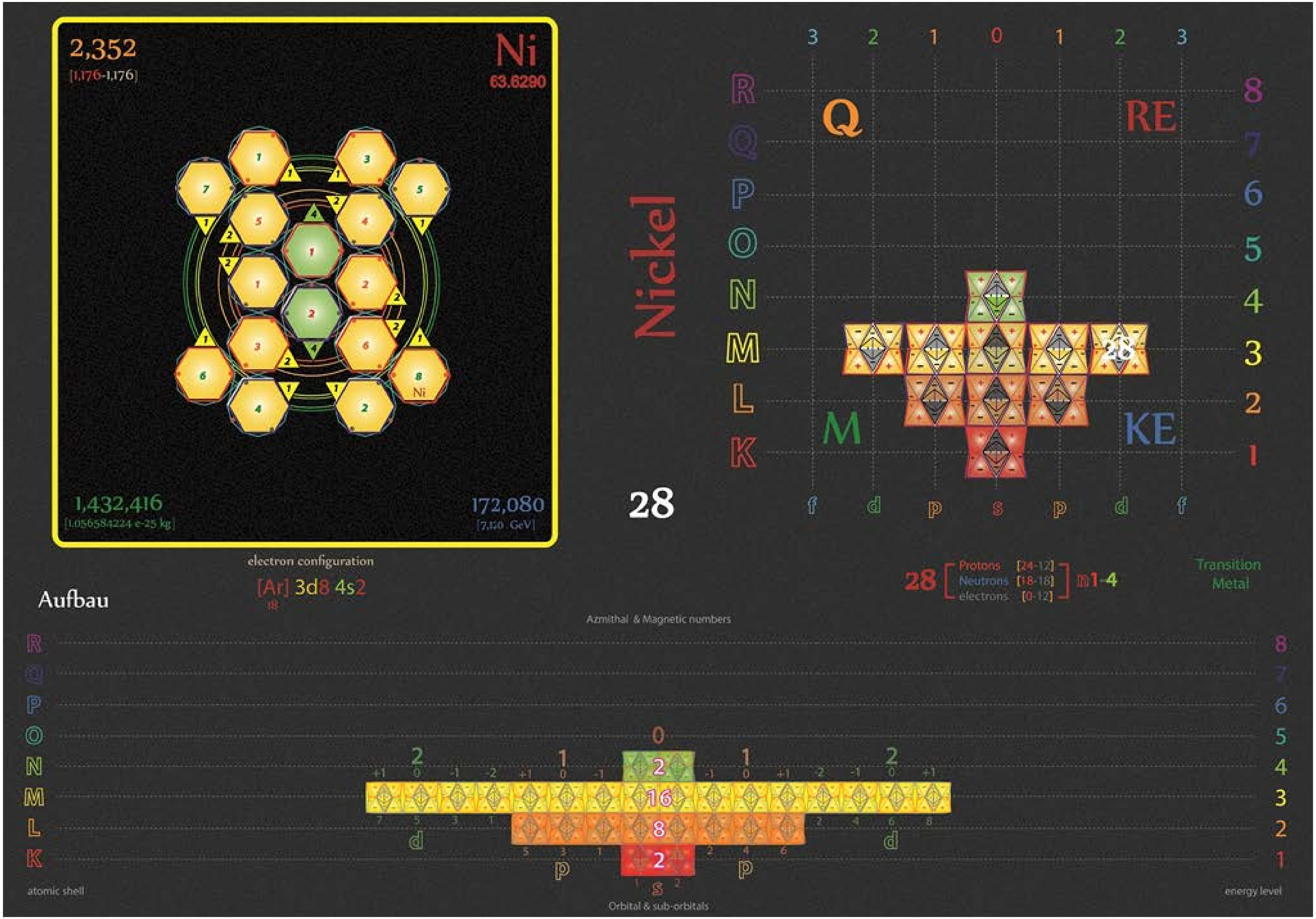
2,352
[1,176-1,176]
Ni
63.6290
1,432,416
[1.056584224 e-25 kg]
172,080
[7,110 GeV]
Nickel
28
electron configuration
[Ar] 3d8 4s2
18
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
28
f d p s p d f
28 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1
2
16
8
2
7 5 3 1 2 4 6 8
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.28 - Nickel atom

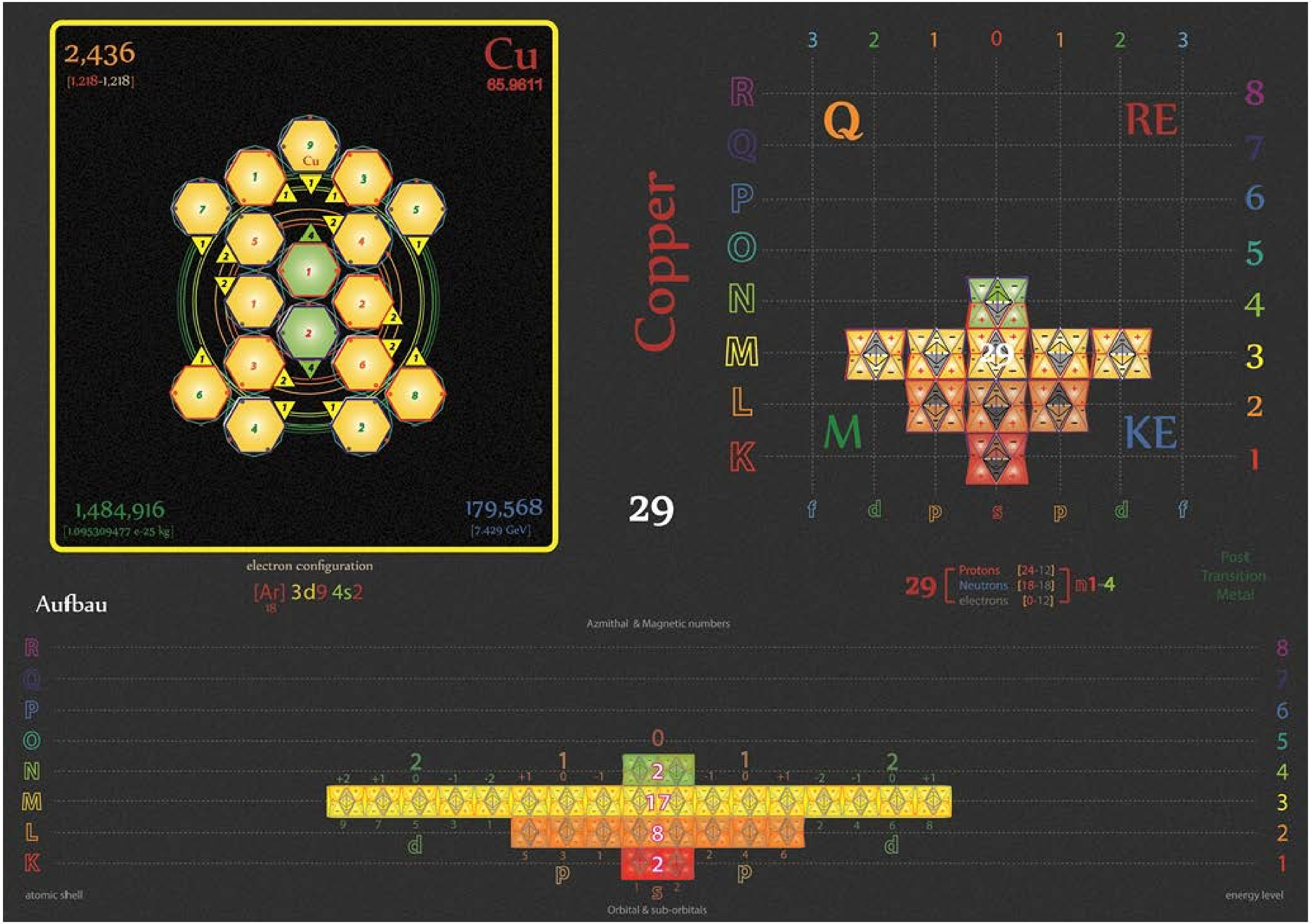
2,436
[1,218-1,218]
Cu
65.9811
9
Cu
1,484,916
[1.095309477 e-25 kg]
179,568
[7.429 GeV]
electron configuration
[Ar] 3d9 4s2
18
Aufbau
Copper
29
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
29
29 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-4
Post Transition Metal
Azmithal & Magnetic numbers
0
2
17
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.29 - Copper atom

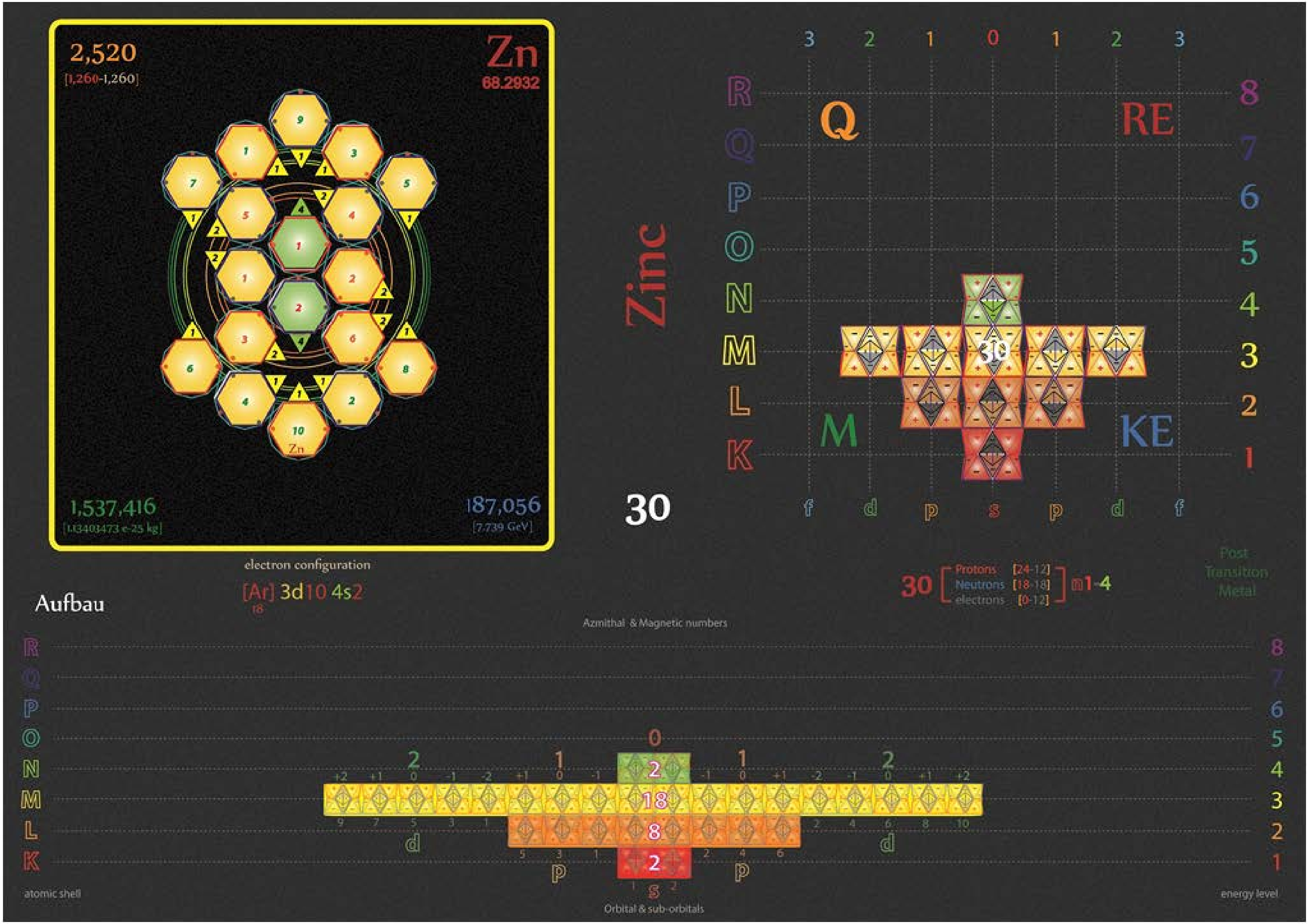
2,520
[1,260-1,260]
Zn
68.2932
1,537,416
[1.13403473 e-25 kg]
187,056
[7.739 GeV]
electron configuration
[Ar] 3d10 4s2
18
Aufbau
Zinc
30
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
30 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
Orbital & sub-orbitals
atomic shell
energy level

Tetryonics 51.30 - Zinc atom

Copyright ABRAHAM [2008] - All rights reserved

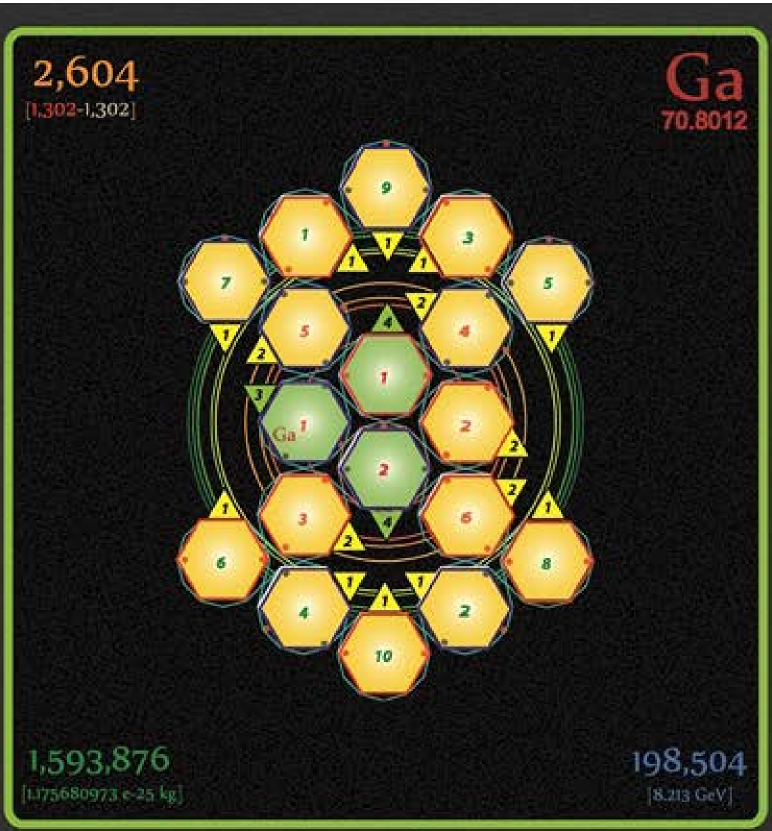
2,604
[1,302-1,302]
Ga
70.8012
9
1
3
7
5
1
1
1
2
5
4
1
1
4
2
1
3
Ga 1
2
2
2
2
4
3
6
2
1
1
6
8
1
1
1
4
2
10
1,593,876
[1.175680973 e-25 kg]
198,504
[8.213 GeV]

Gallium

31

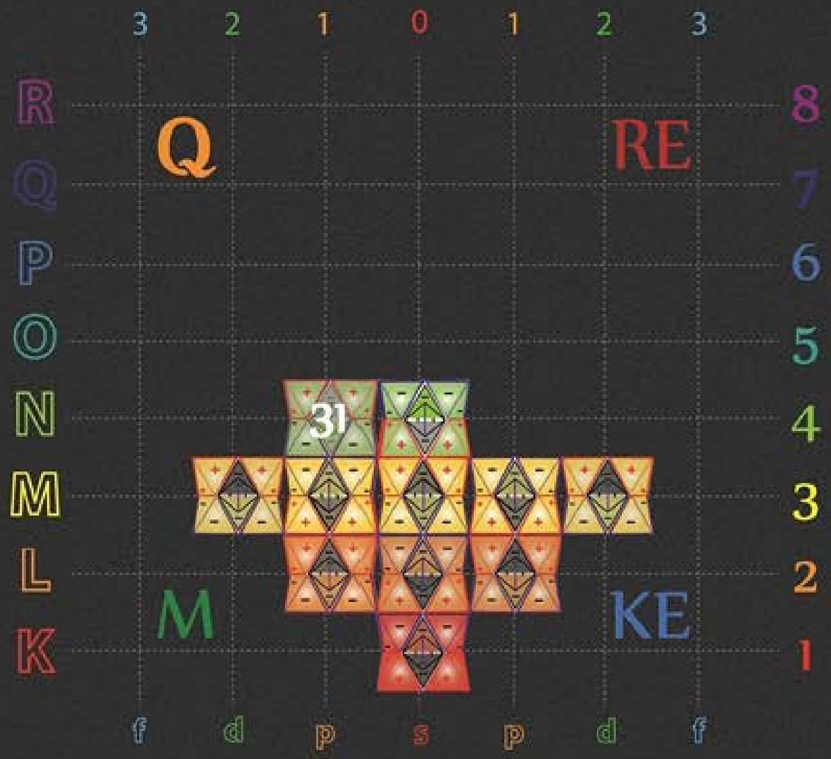
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
31
f
d
p
s
p
d
f

31 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Metaloid

electron configuration

[Ar] 3d10 4s2 4p1

18

Aufbau

Azmithal & Magnetic numbers

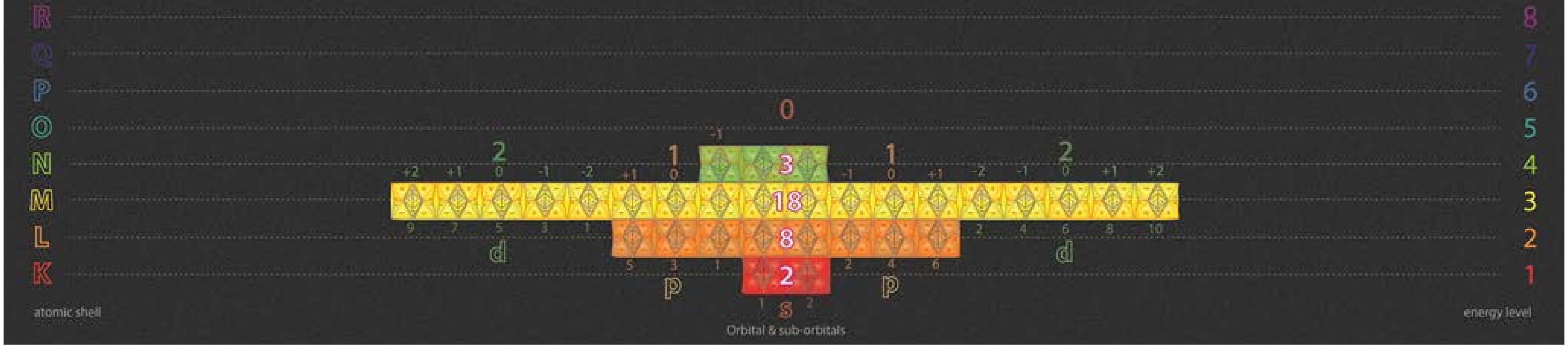
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
-1
3
18
8
2
2
1
1
2
+2
+1
0
-1
-2
+1
0
-1
0
+1
-2
-1
0
+1
+2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbital & sub-orbitals

Tetryonics 51.31 - Gallium atom

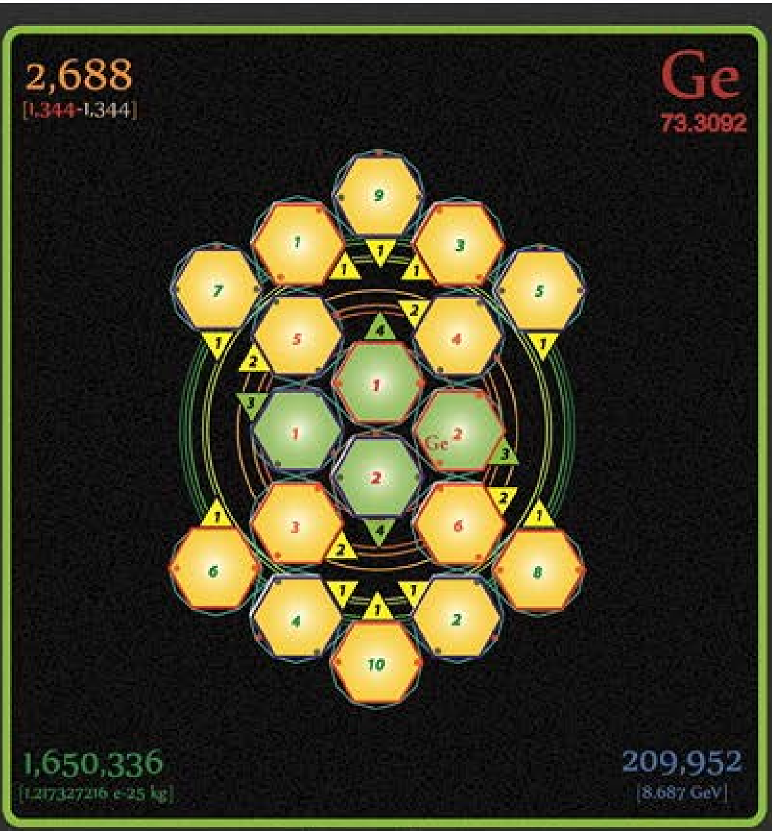
2,688
[1,344-1,344]
Ge
73.3092
9
1
3
7
5
1
1
1
2
5
4
4
1
2
3
1
1
2
Ge
2
3
2
2
1
3
6
1
6
4
2
8
1
1
1
4
2
10
1,650,336
[1.217327216 e-25 kg]
209,952
[8.687 GeV]

Germanium

32

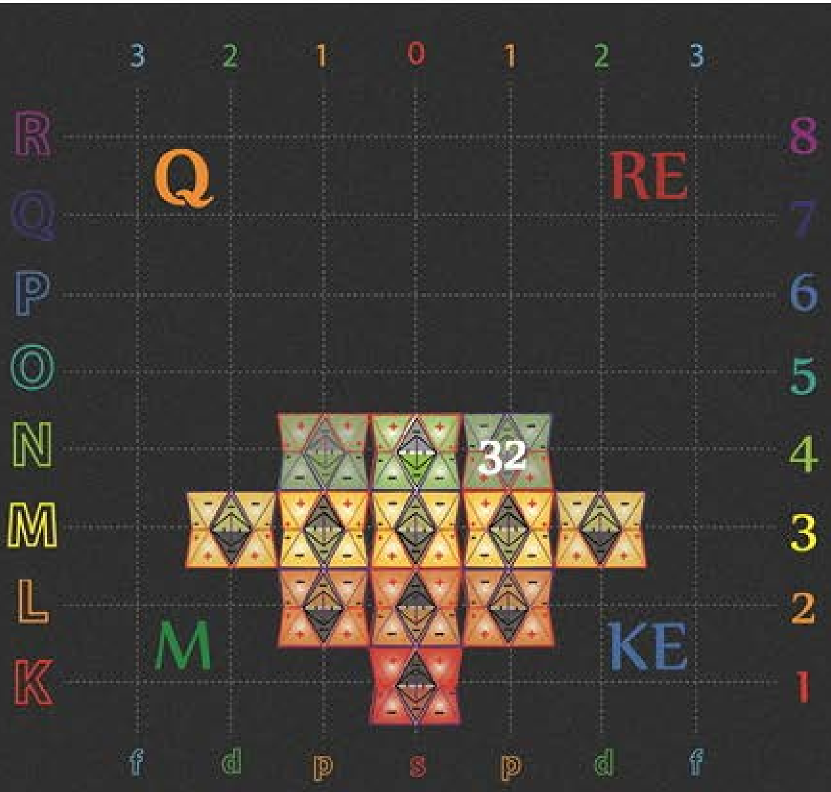
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
32
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

Aufbau

[Ar] 3d10 4s2 4p2
18

32 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-4

Metaloid

Azmithal & Magnetic numbers

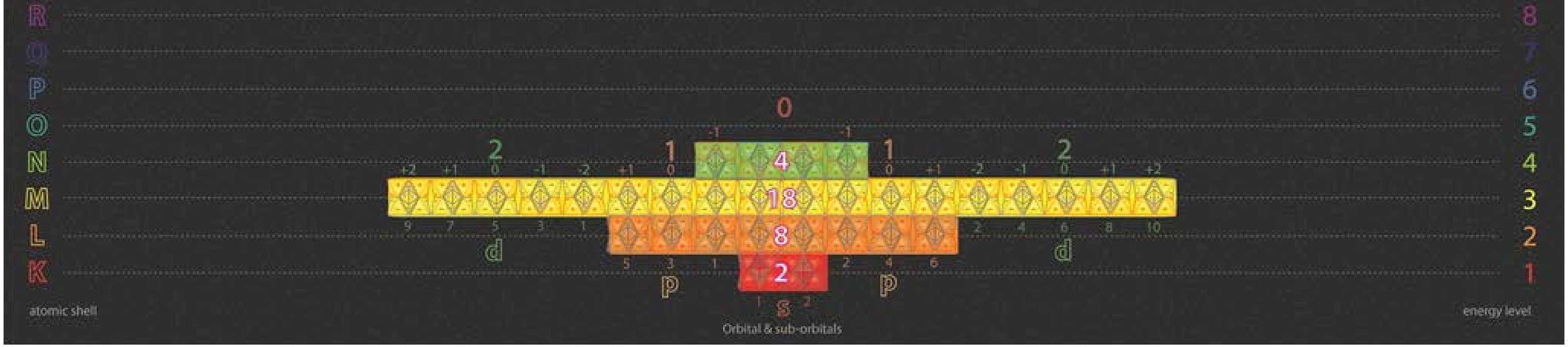
R
Q
P
O
N
M
L
K
0
-1
-1
2
1
4
1
2
+2
+1
0
-1
-2
+1
0
0
+1
-2
-1
0
+1
+2
18
9
7
5
3
1
2
4
6
8
10
d
8
d
5
3
1
2
4
6
p
2
p
1
s
2
8
7
6
5
4
3
2
1

atomic shell

energy level

Orbital & sub-orbitals

Tetryonics 51.32 - Germanium atom

Copyright ABRAHAM [2008] - All rights reserved

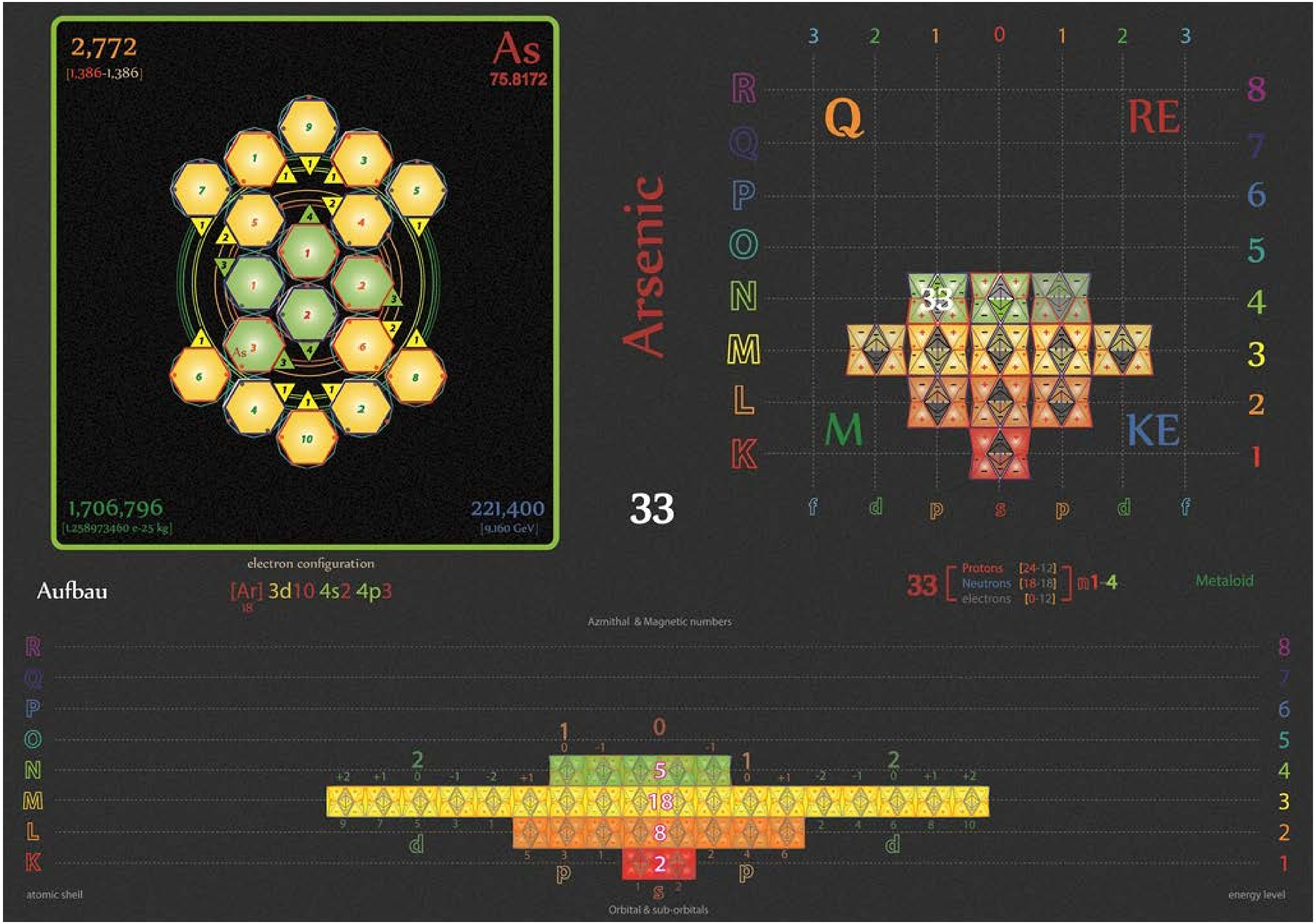
2,772
[1,386-1,386]
As
75.8172
1,706,796
[1.258973460 e-25 kg]
221,400
[9.160 GeV]
Arsenic
33
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Ar] 3d10 4s2 4p3
18
33 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Azmithal & Magnetic numbers
5
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.33 - Arsenic atom

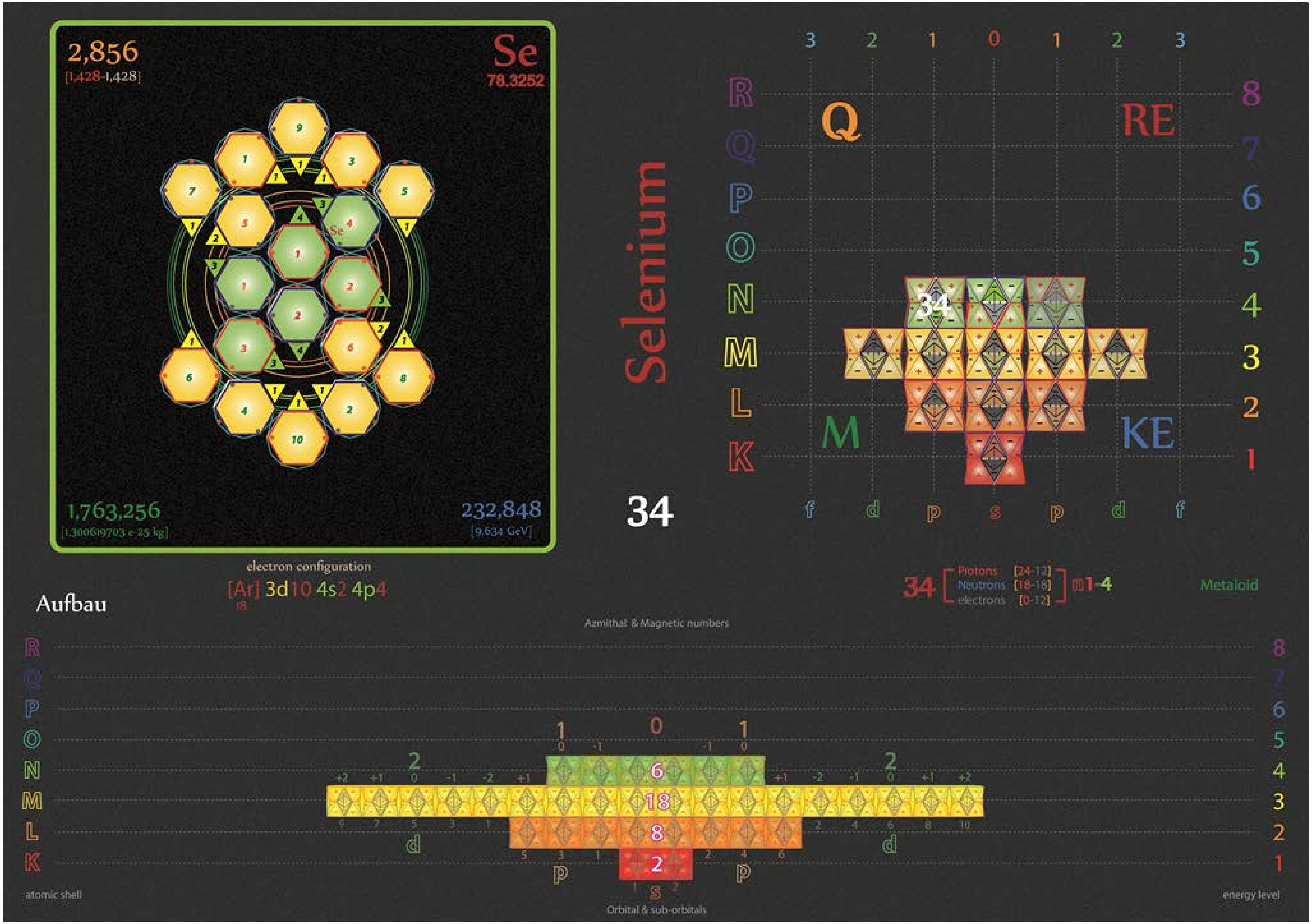
2,856
[1,428-1,428]
Se
78.3252
1,763,256
[1.30061970 3 e-25 kg]
232,848
[9.634 GeV]
Selenium
34
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
[Ar] 3d10 4s2 4p4
18
34 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Metaloid
Aufbau
Azmithal & Magnetic numbers
atomic shell
energy level
Orbital & sub-orbitals
2 6 18 8
s p d

Tetryonics 51.34 - Selenium atom

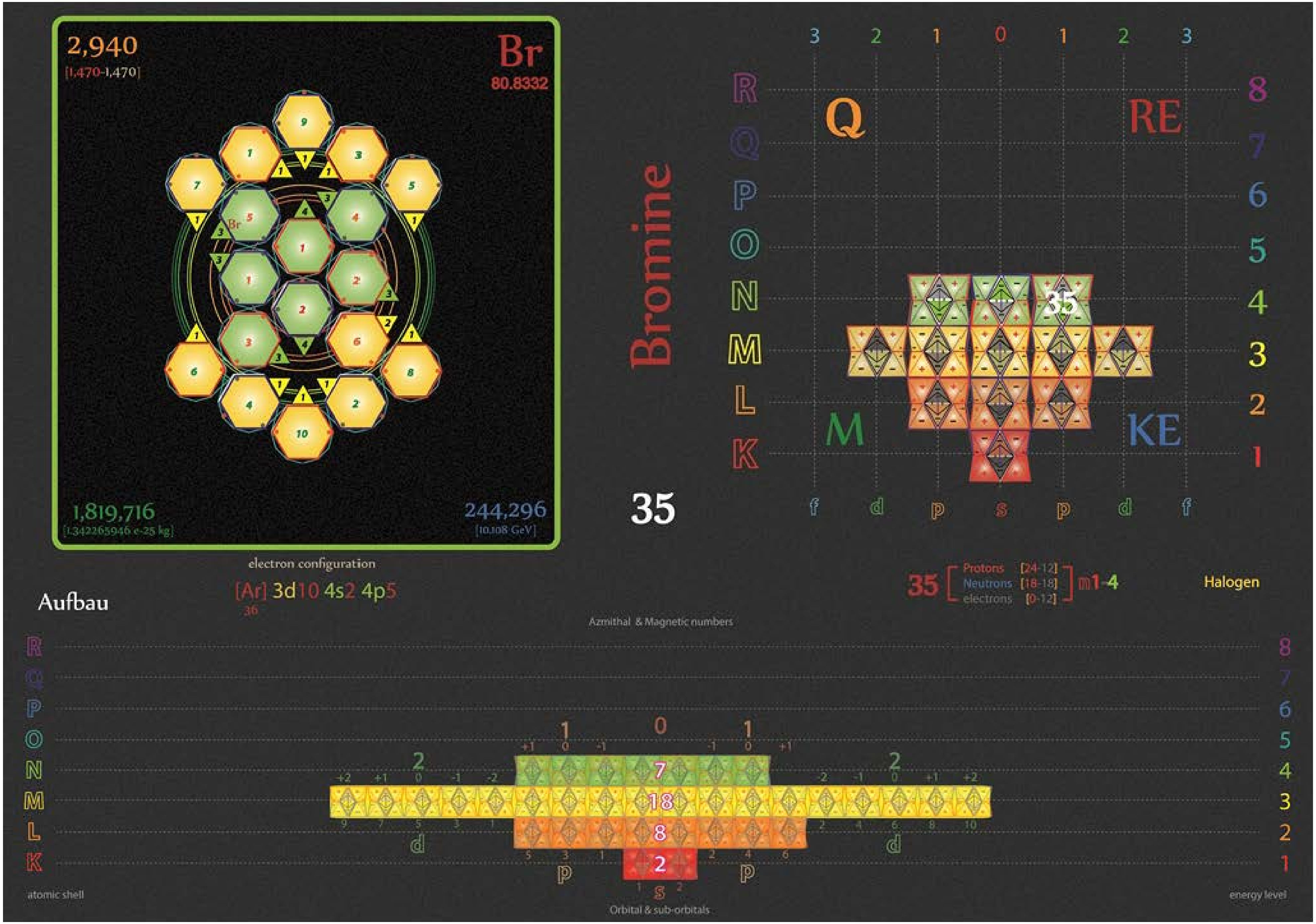
2,940
[1,470-1,470]
Br
80.8332
1,819,716
[1.342265946 e-25 kg]
244,296
[10,108 GeV]
Bromine
35
electron configuration
[Ar] 3d10 4s2 4p5
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
35 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4
Halogen
Azmithal & Magnetic numbers
7
18
8
2
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.35 - Bromine atom

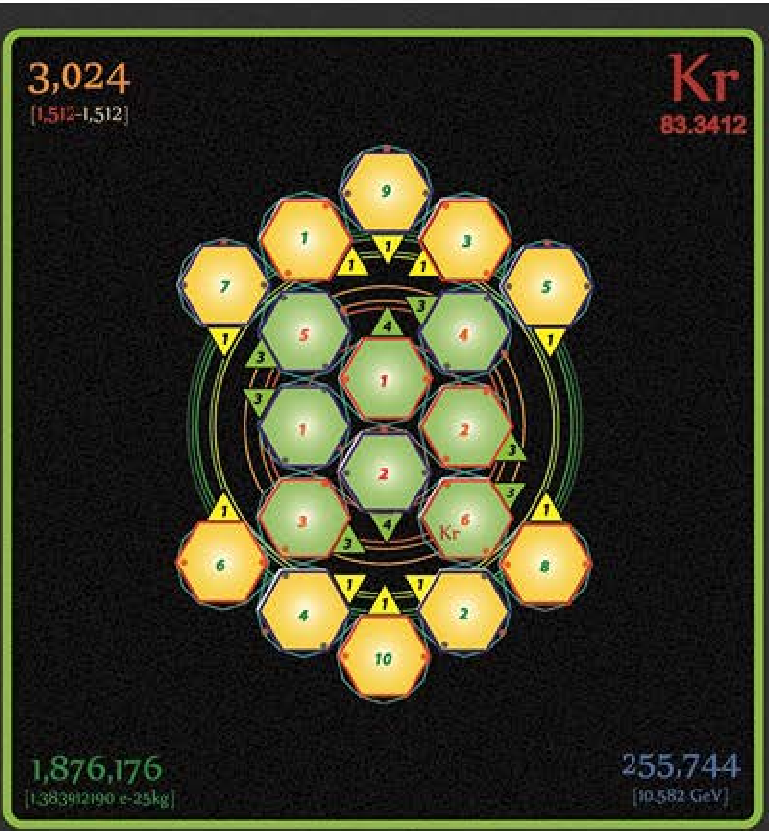
3,024
[1,512-1,512]

Kr
83.3412

9
1
3
7
5
5
4
1
1
2
2
3
6
Kr
6
8
4
2
10

1,876,176
[1.383912190 e-25kg]

255,744
[10.582 GeV]

electron configuration

[Ar] 3d10 4s2 4p6
36

Krypton

36

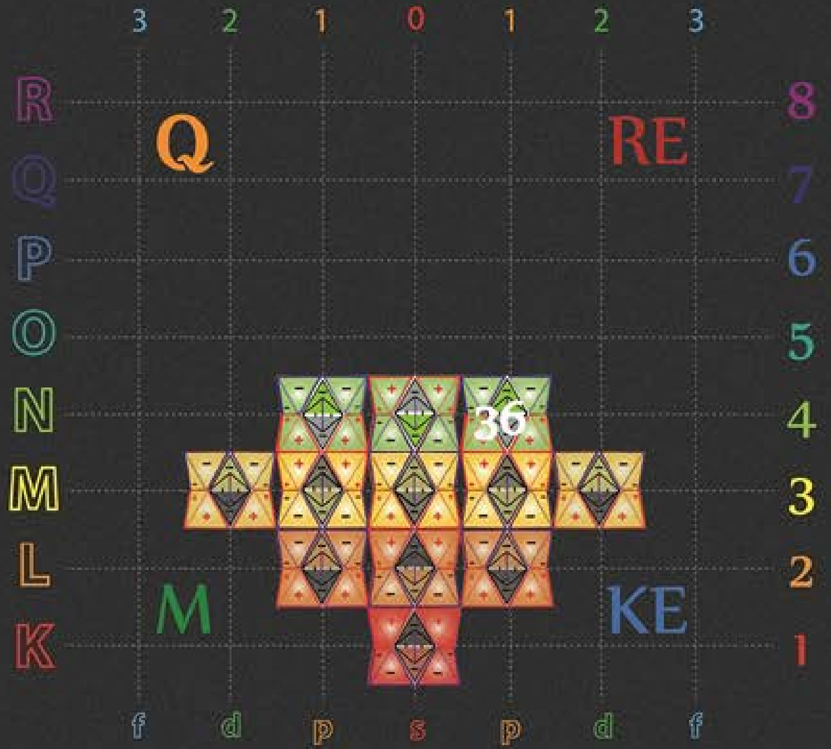
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
36
f
d
p
s
p
d
f

36 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-4

Noble Gas

Aufbau

Azmithal & Magnetic numbers

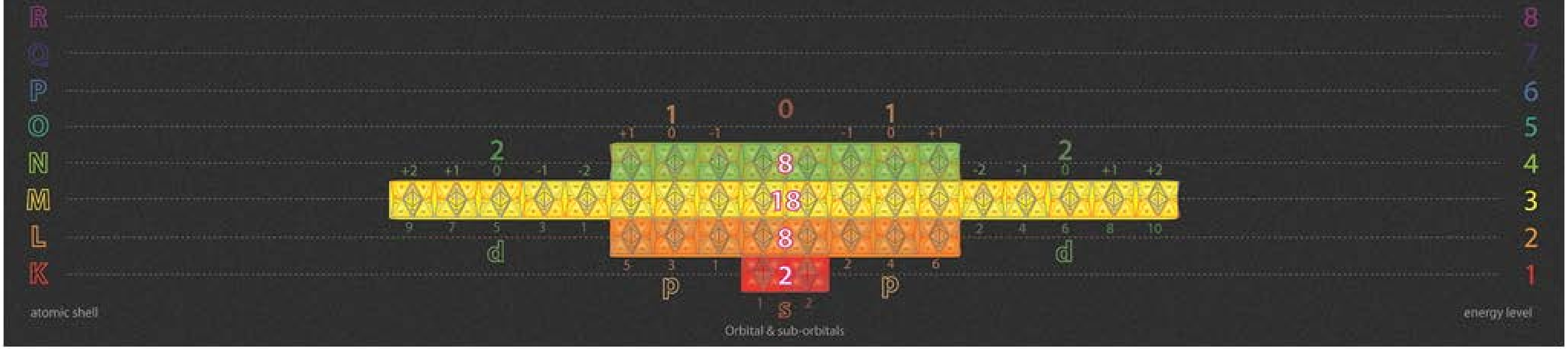
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
+1
0
-1
-1
0
+1
2
2
+2
+1
0
-1
-2
-2
-1
0
+1
+2
8
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbital & sub-orbitals

Tetryonics 51.36 - Krypton atom

Copyright ABRAHAM [2008] - All rights reserved

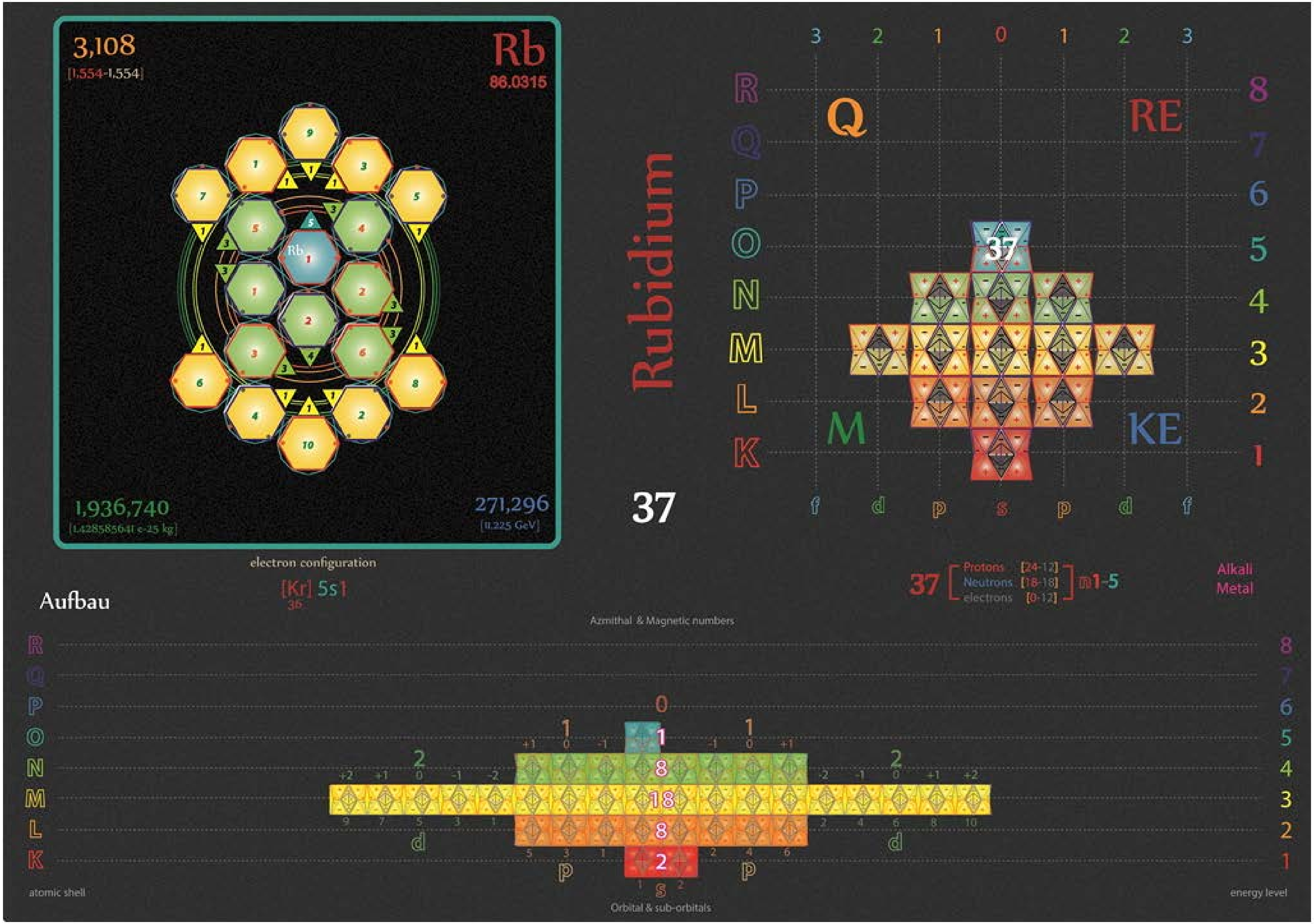
3,108
[1.554-1.554]
Rb
86.0315
Rb
1,936,740
[1.428585641 e-25 kg]
271,296
[11.225 GeV]
Rubidium
37
electron configuration
[Kr] 5s1
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
37
8 7 6 5 4 3 2 1
f d p s p d f
37
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Alkali
Metal
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
1
1
1
+1 0 -1
-1 0 +1
2
2
8
+2 +1 0 -1 -2
-2 -1 0 +1 +2
18
9 7 5 3 1
2 4 6 8 10
d
d
8
5 3 1
2 4 6
p
p
2
1 2
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.37 - Rubidium atom

Copyright ABRAHAM [2008] - All rights reserved

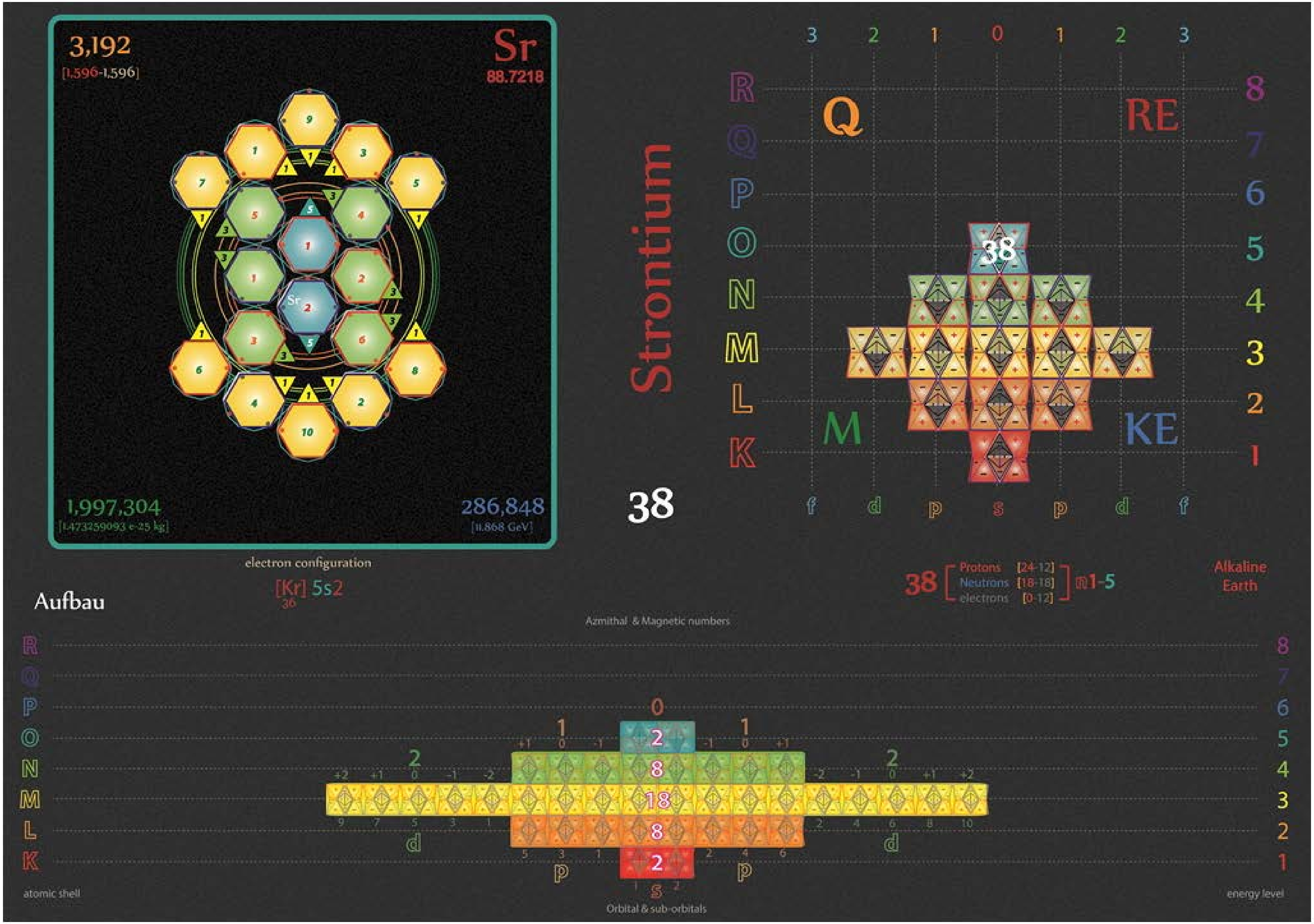
3,192
[1,596-1,596]
Sr
88.7218
9
1
3
7
5
5
4
1
1
2
Sr
2
3
6
5
6
8
4
2
10
1,997,304
[1.473259093 e-25 kg]
286,848
[11.868 GeV]
Strontium
38
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
38
f
d
p
s
p
d
f
electron configuration
[Kr] 5s2
36
Aufbau
38
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-5
Alkaline
Earth
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
1
1
2
2
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
2
8
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.38 - Strontium atom

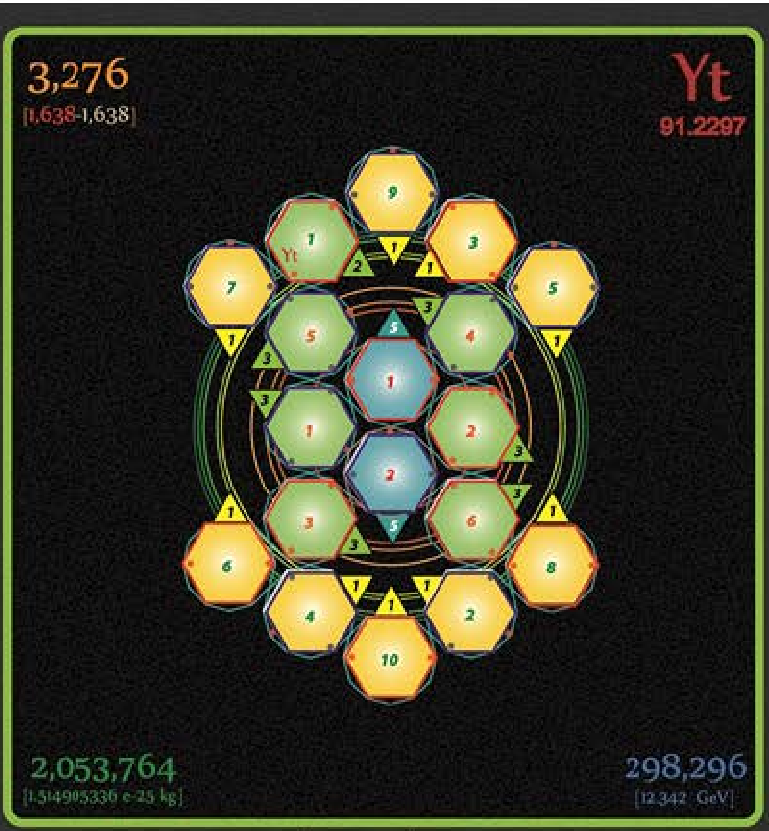
3,276
[1,638-1,638]
Yt
91.2297
2,053,764
[1.514905336 e-25 kg]
298,296
[12.342 GeV]

Yttrium

39

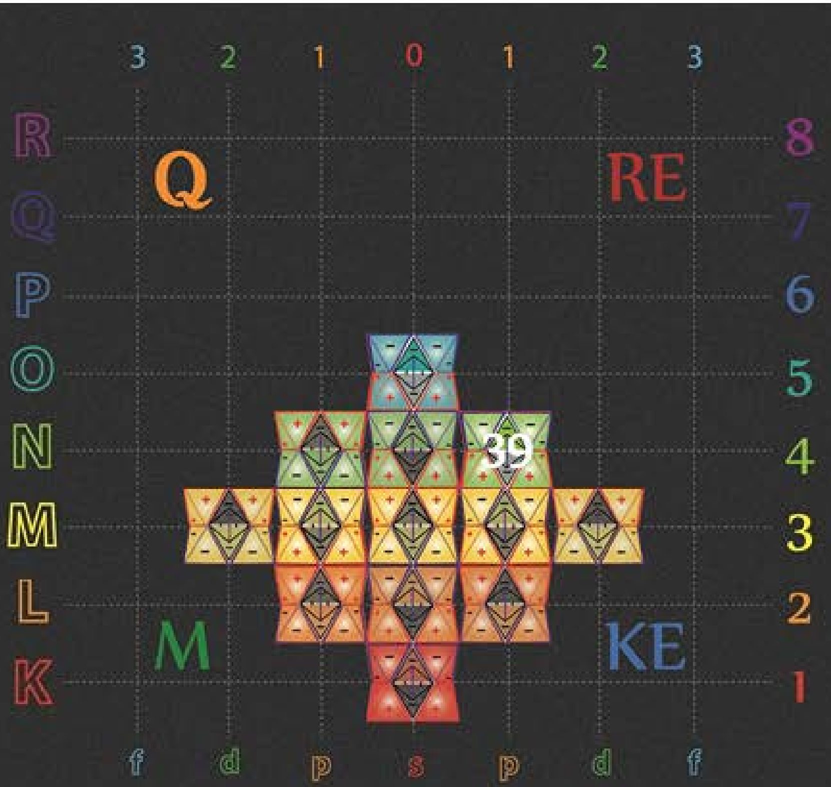
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
39
f
d
p
s
p
d
f

electron configuration

[Kr] 4d1 5s2
36

Aufbau

Azmithal & Magnetic numbers

39 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Transition Metal

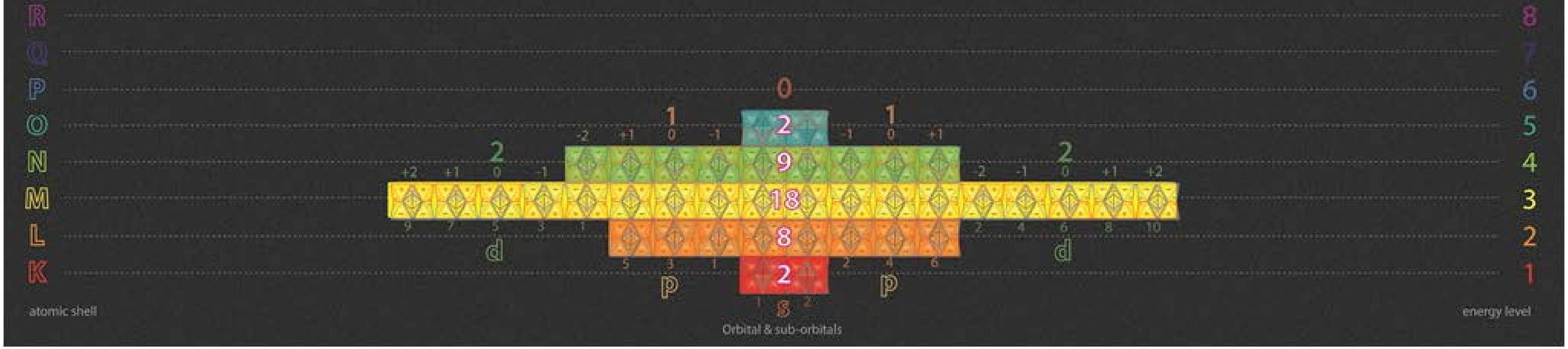
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
2
2
9
18
8
2
d
d
p
p
s
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.39 - Yttrium atom

Copyright ABRAHAM [2008] - All rights reserved

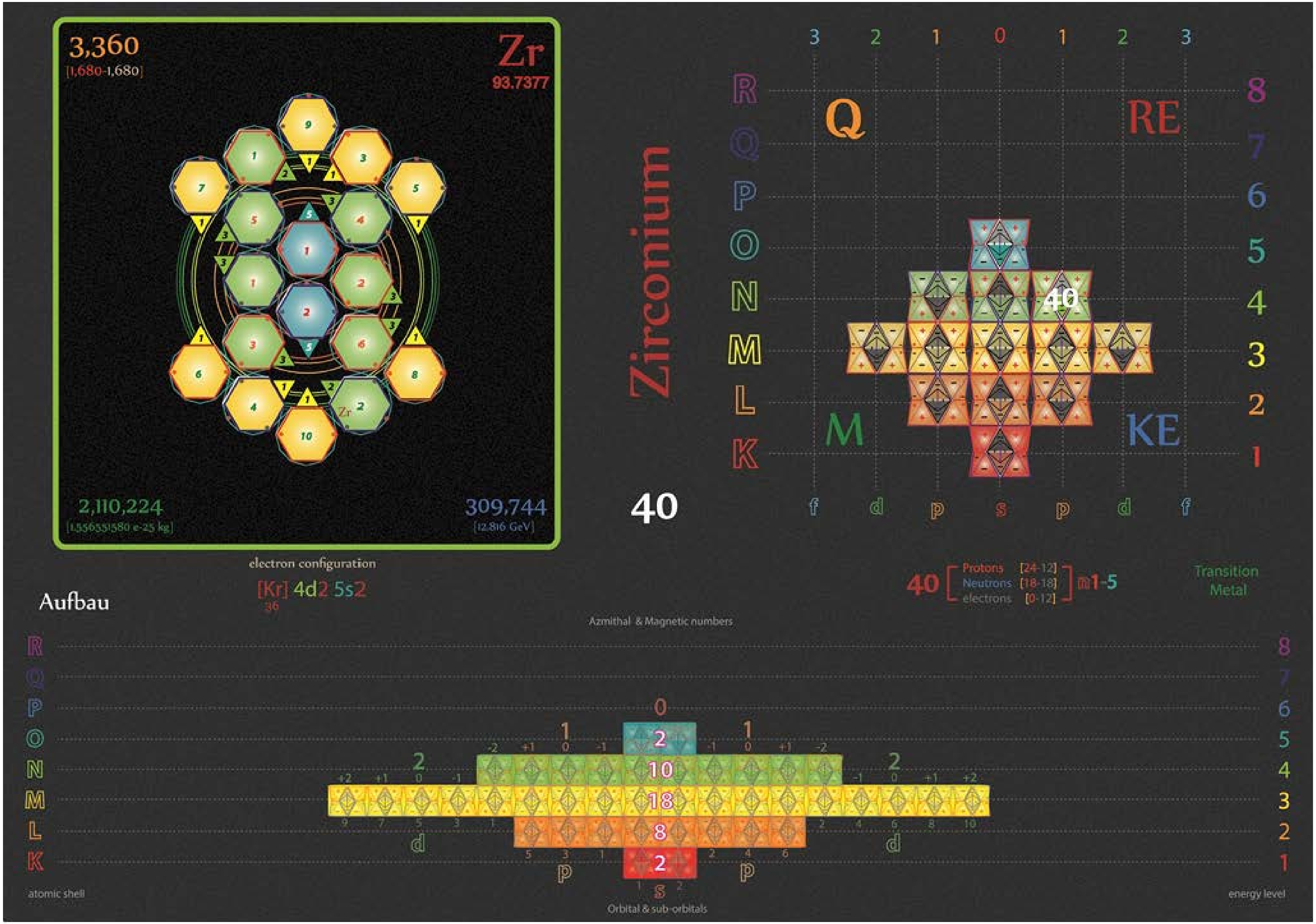
3,360
[1,680-1,680]
Zr
93.7377
2,110,224
[1.556551580 e-25 kg]
309,744
[12.816 GeV]
Zirconium
40
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
40
f d p s p d f
electron configuration
[Kr] 4d2 5s2
36
Aufbau
40 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
1
1
2
2
2
10
18
8
2
d
p
s
p
d
atomic shell
energy level
Orbital & sub-orbitals

Tetryonics 51.40 - Zirconium atom

Copyright ABRAHAM [2008] - All rights reserved

3,444
[1,722-1,722]

Nb
98.2457

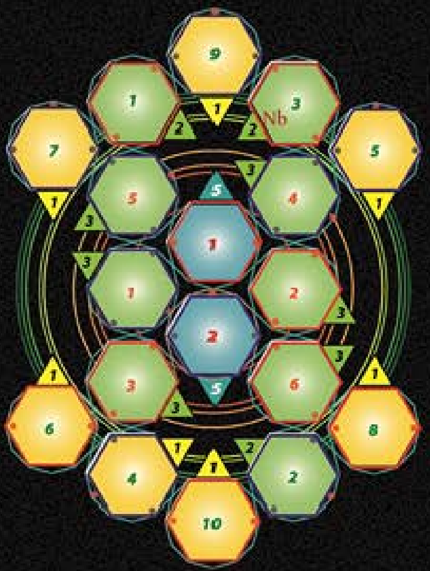
9
1
3
Nb
7
5
5
4
1
1
2
2
3
6
8
4
10
2
1
2
3
3
5

2,166,684
[1.598197823 e-25 kg]

321,192
[13.289 GeV]

Niobium

41

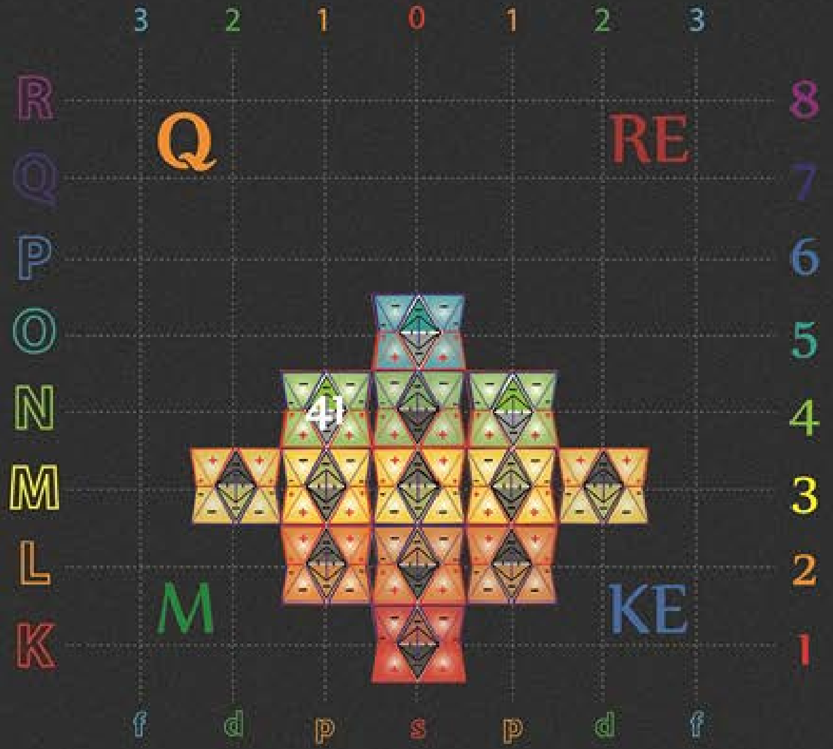
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
41
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration
[Kr] 4d3 5s2
36

41 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-5

Transition
Metal

Aufbau

Azmithal & Magnetic numbers

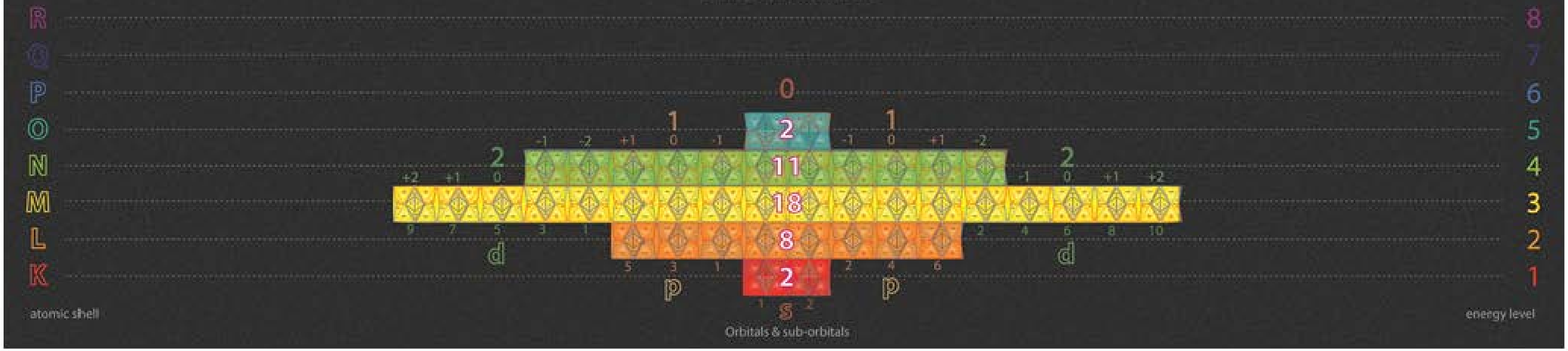
R
Q
P
O
N
M
L
K
0
1
1
2
2
-1
-2
+1
0
-1
-1
0
+1
-2
+2
+1
0
-1
0
+1
+2
2
11
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.41 - Niobium atom

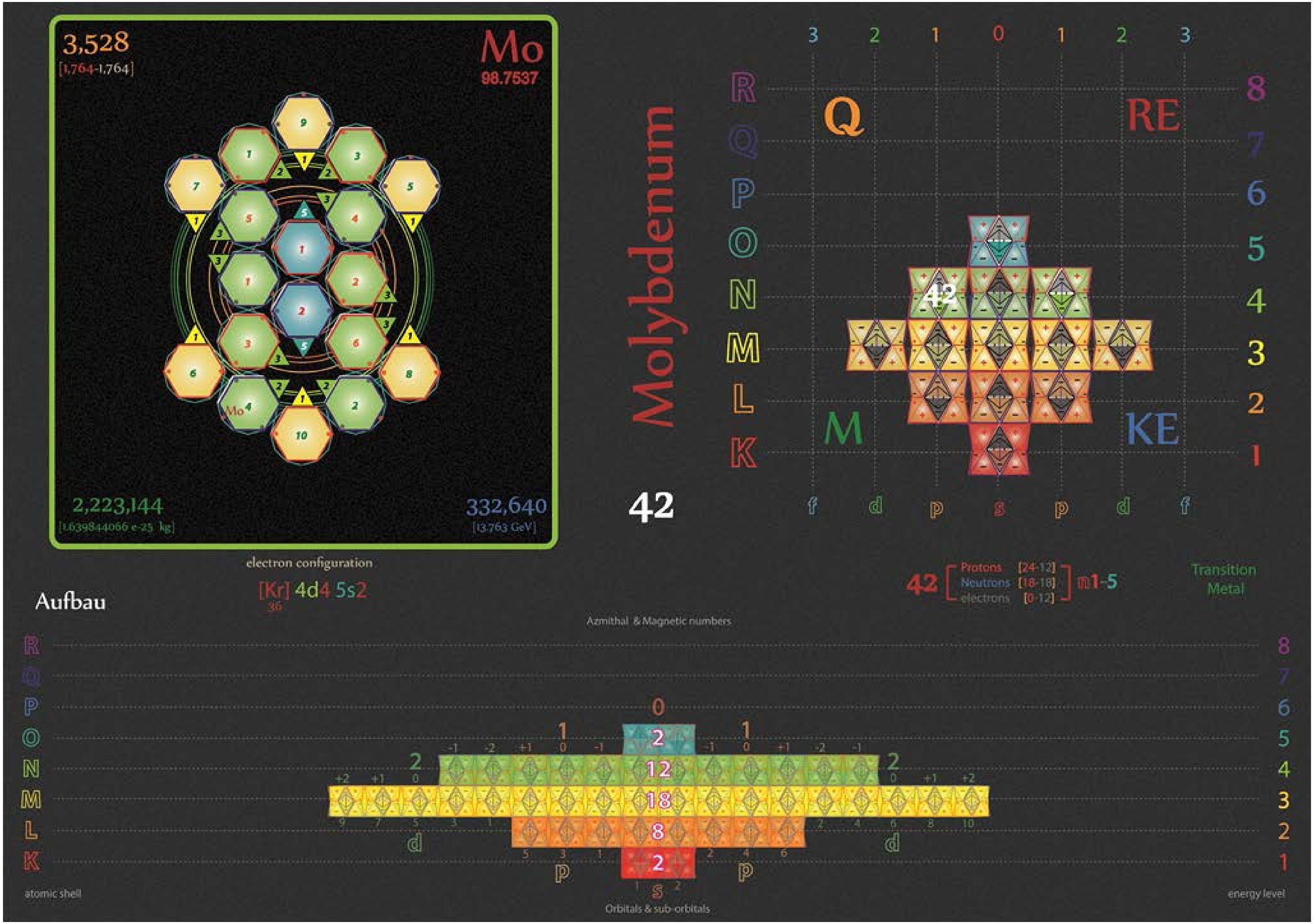
3,528
[1,764-1,764]
Mo
98.7537
2,223,144
[1.639844066 e-25 kg]
332,640
[13.763 GeV]
Molybdenum
42
R Q P O N M L K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d4 5s2
36
Aufbau
42 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
2
12
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.42 - Molybdenum atom

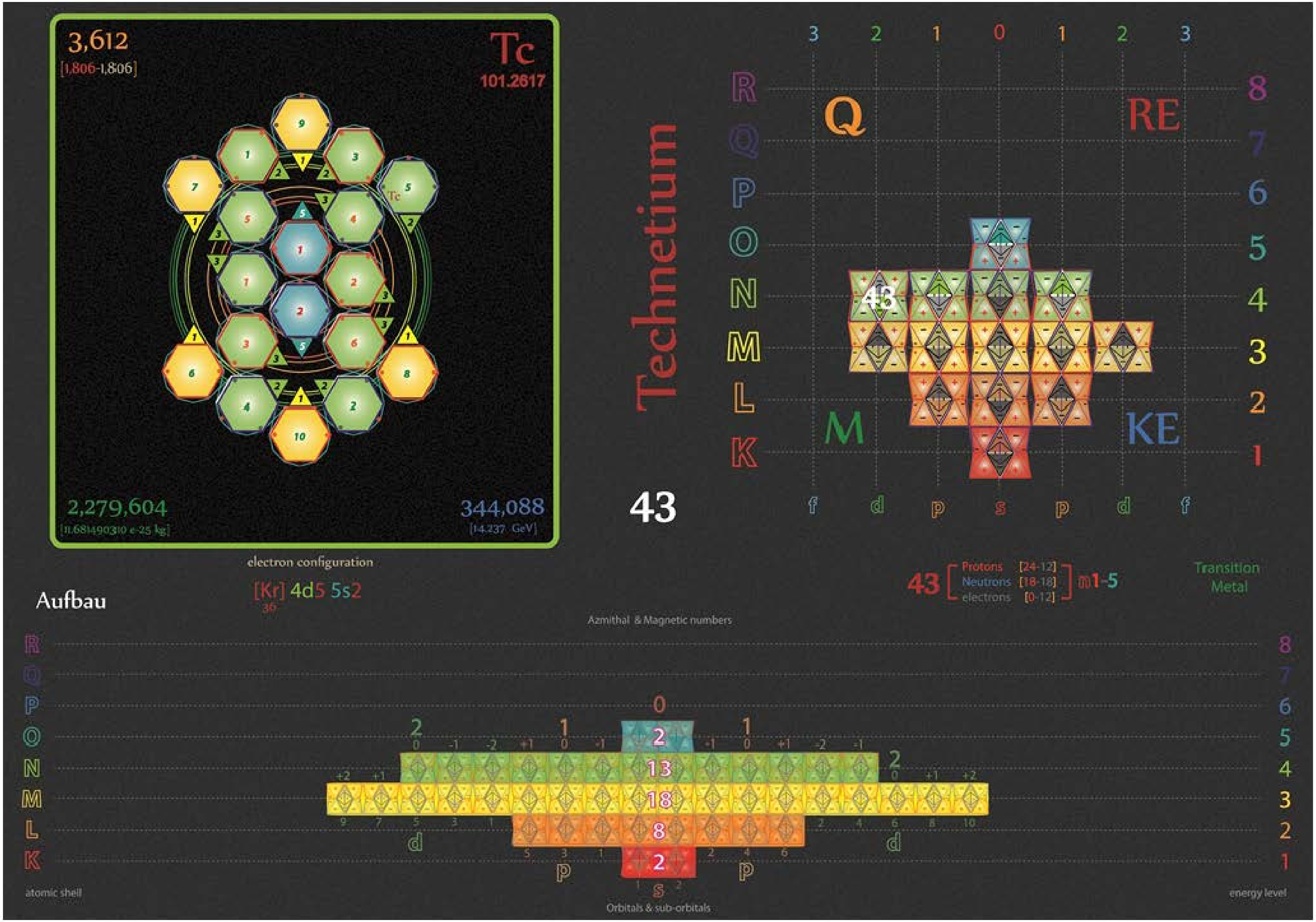
3,612
[1,806-1,806]
Tc
101.2617
2,279,604
[11.68149031 0 e-25 kg]
344,088
[14.237 GeV]
Technetium
43
electron configuration
[Kr] 4d5 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
43
43 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2
13
18
8
2
d
p
s
d
p
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.43 - Technetium atom

Copyright ABRAHAM [2008] - All rights reserved

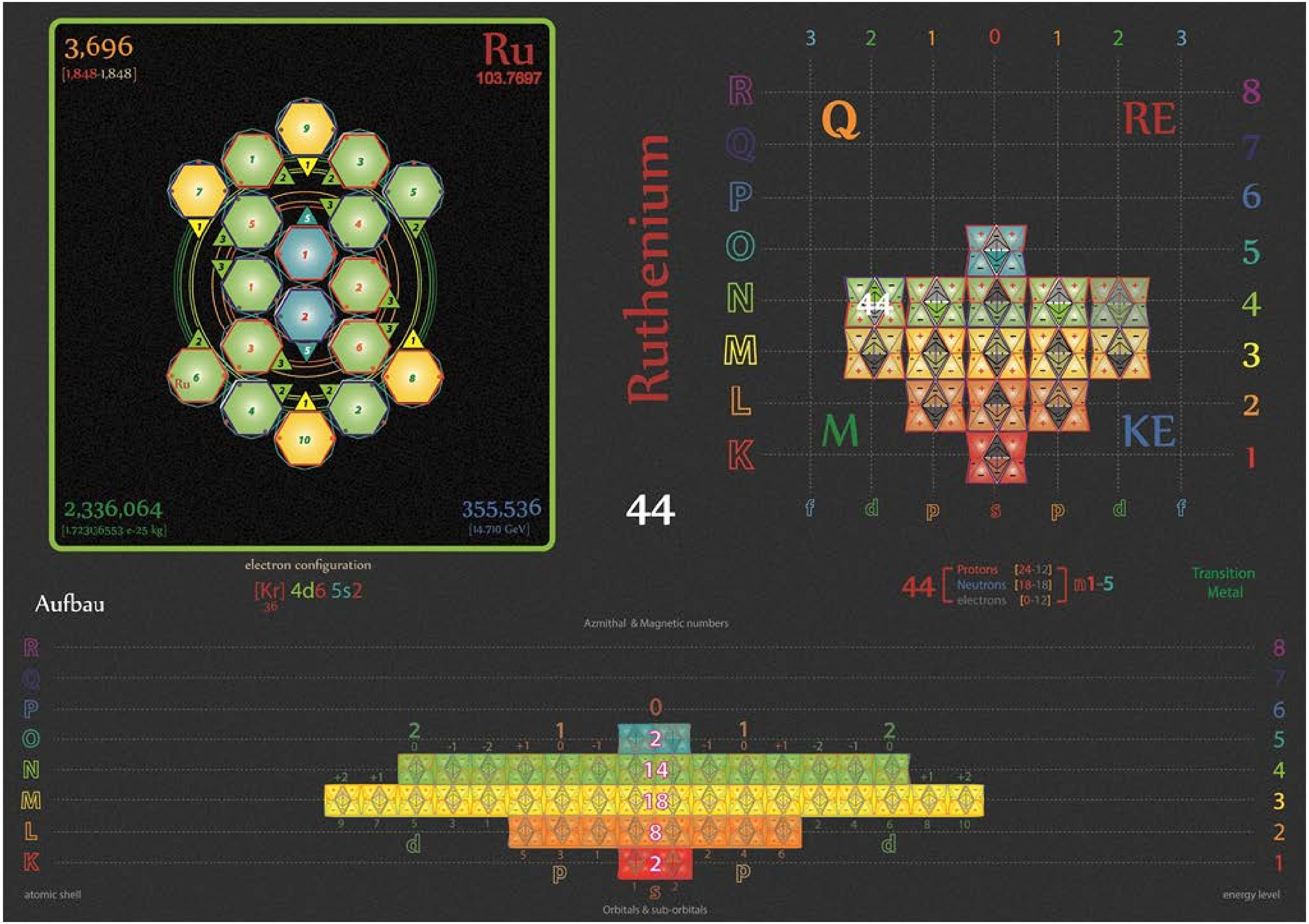
3,696
[1,848-1,848]
Ru
103.7697
2,336,064
[1.7230,6553 e-25 kg]
355,536
[14.710 GeV]
Ruthenium
44
electron configuration
[Kr] 4d6 5s2
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
44 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
14
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.44 - Ruthenium atom

Copyright ABRAHAM [2008] - All rights reserved

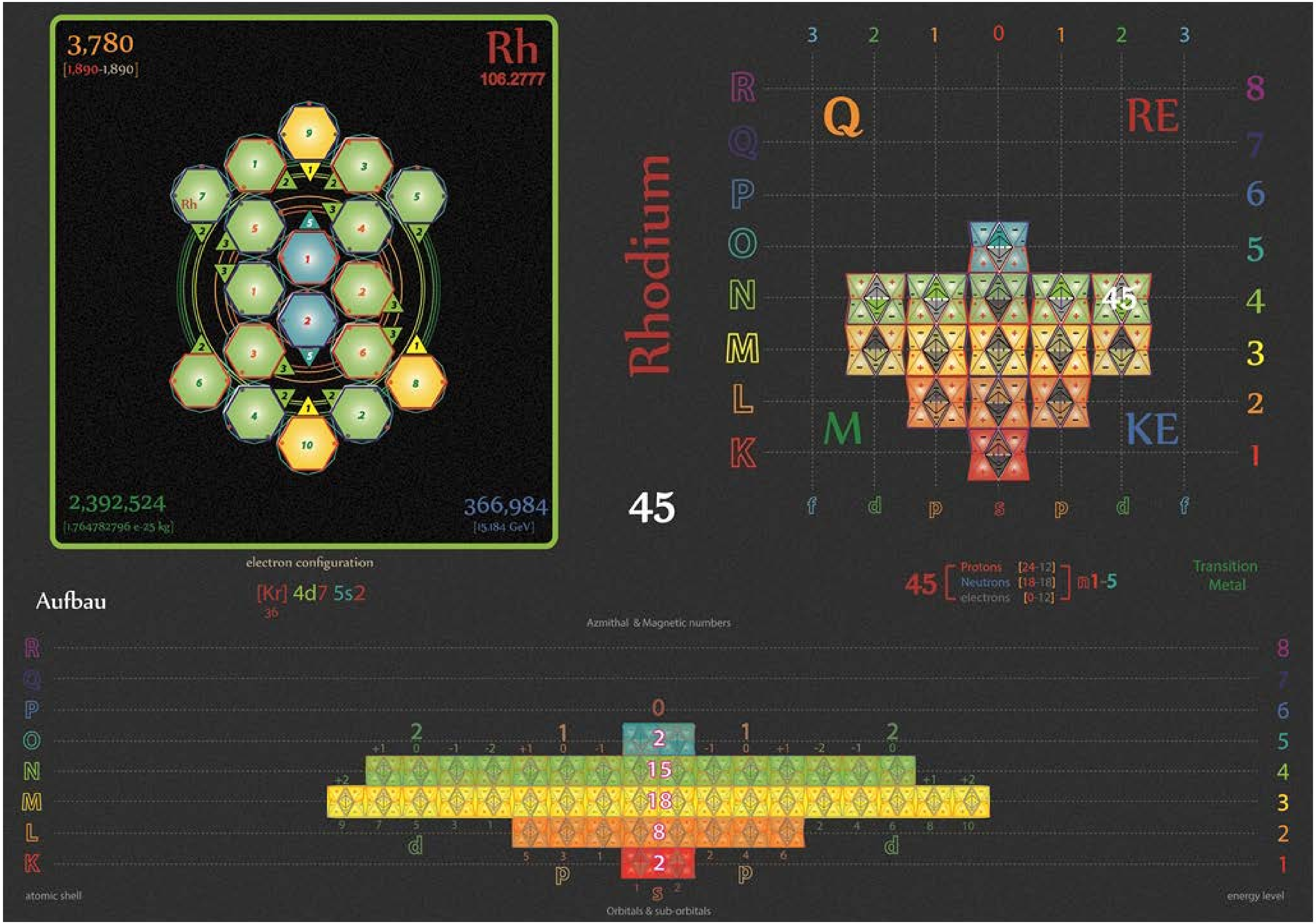
3,780
[1,890-1,890]
Rh
106.2777
2,392,524
[1.764782796 e-25 kg]
366,984
[15.184 GeV]
Rhodium
45
electron configuration
[Kr] 4d7 5s2
36
Aufbau
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
45 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
2
15
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.45 - Rhodium atom

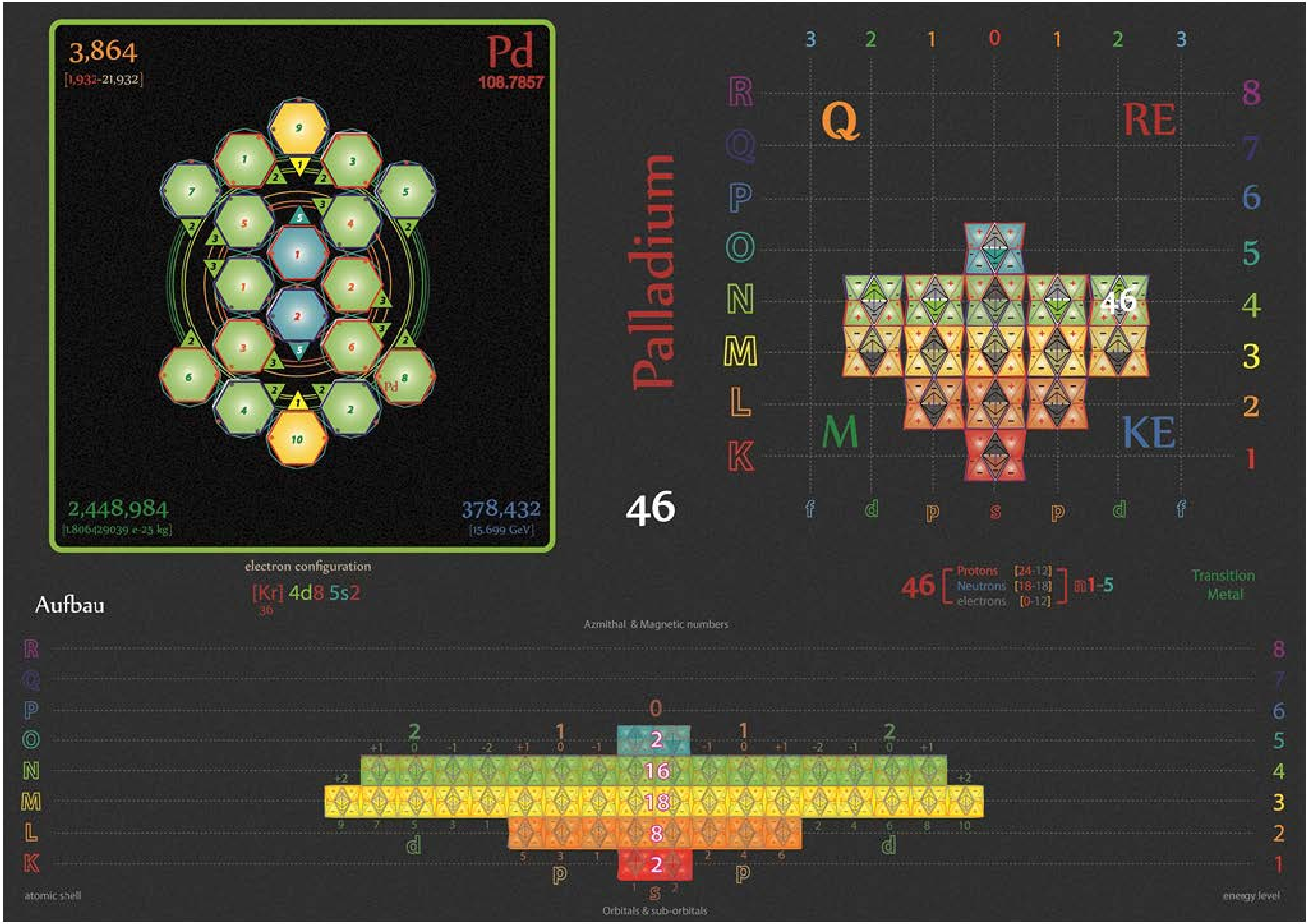
3,864
[1,932-21,932]
Pd
108.7857
2,448,984
[1.806429039 e-25 kg]
378,432
[15.699 GeV]
Palladium
46
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
46
f d p s p d f
46 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Transition Metal
electron configuration
[Kr] 4d8 5s2
36
Aufbau
Azimthal & Magnetic numbers
0
2 1 1 2
2
16
18
8
2
d d
p p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.46 - Palladium atom

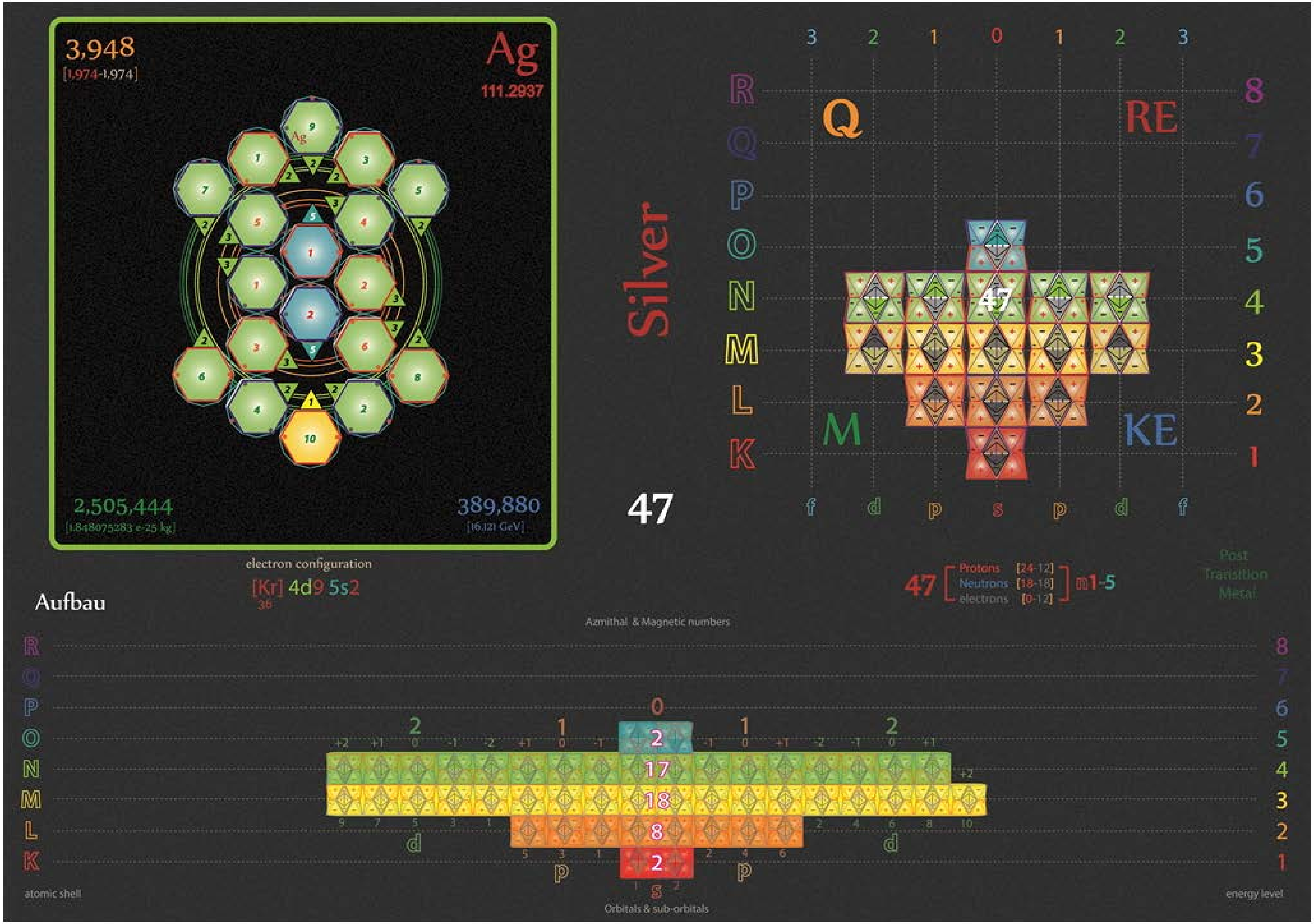
3,948
[1,974-1,974]
Ag
111.2937
2,505,444
[1.848075283 e-25 kg]
389,880
[16.121 GeV]
Silver
47
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
47 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Post Transition Metal
electron configuration
[Kr] 4d9 5s2
36
Aufbau
Azmithal & Magnetic numbers
2 1 0 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
17
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.47 - Silver atom

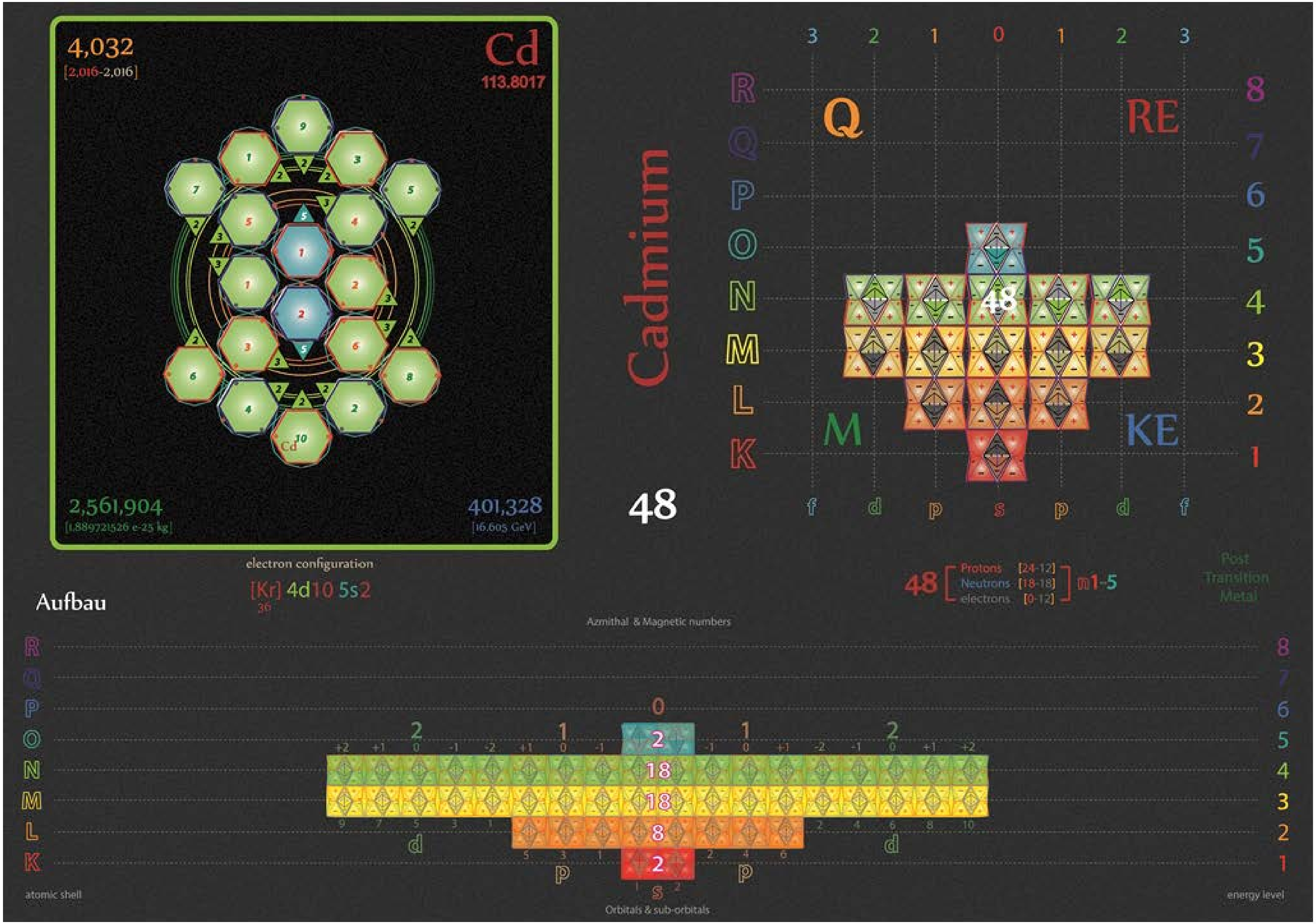
4,032
[2,016-2,016]
Cd
113.8017
2,561,904
[1.889721526 e-25 kg]
401,328
[16.605 GeV]
Cd
10
electron configuration
[Kr] 4d10 5s2
36
Aufbau
Cadmium
48
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
48
f d p s p d f
48
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Post Transition Metal
Azmithal & Magnetic numbers
0
2 1 1 2
+2 +1 0 -1 -2 +1 0 -1 -1 0 +1 -2 -1 0 +1 +2
2
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.48 - Cadmium atom

Copyright ABRAHAM [2008] - All rights reserved

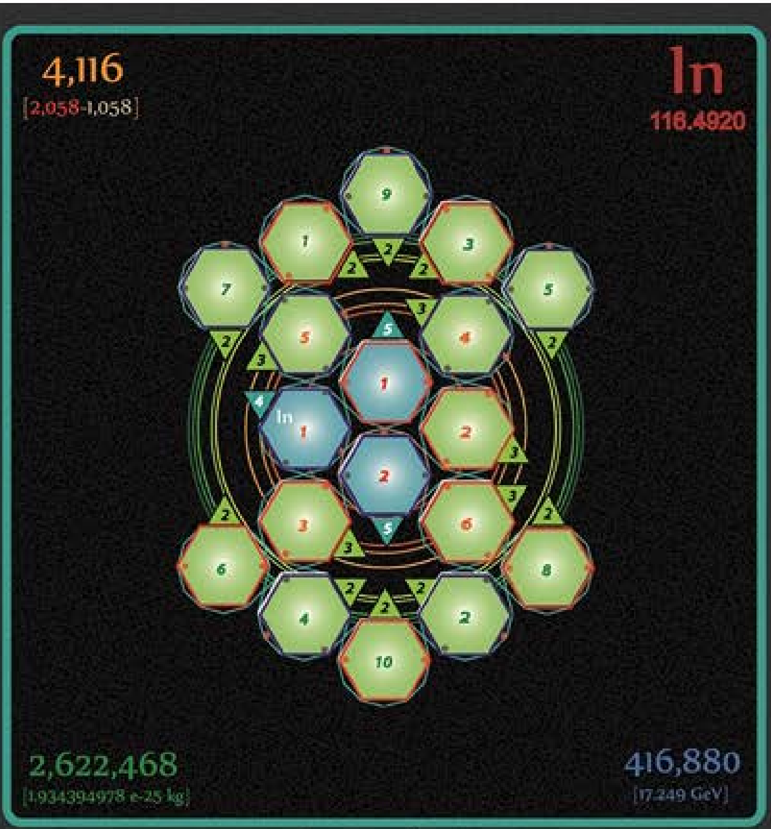
4,116
[2,058-1,058]
In
116.4920
In
2,622,468
[1.934394978 e-25 kg]
416,880
[17.249 GeV]

Indium

49

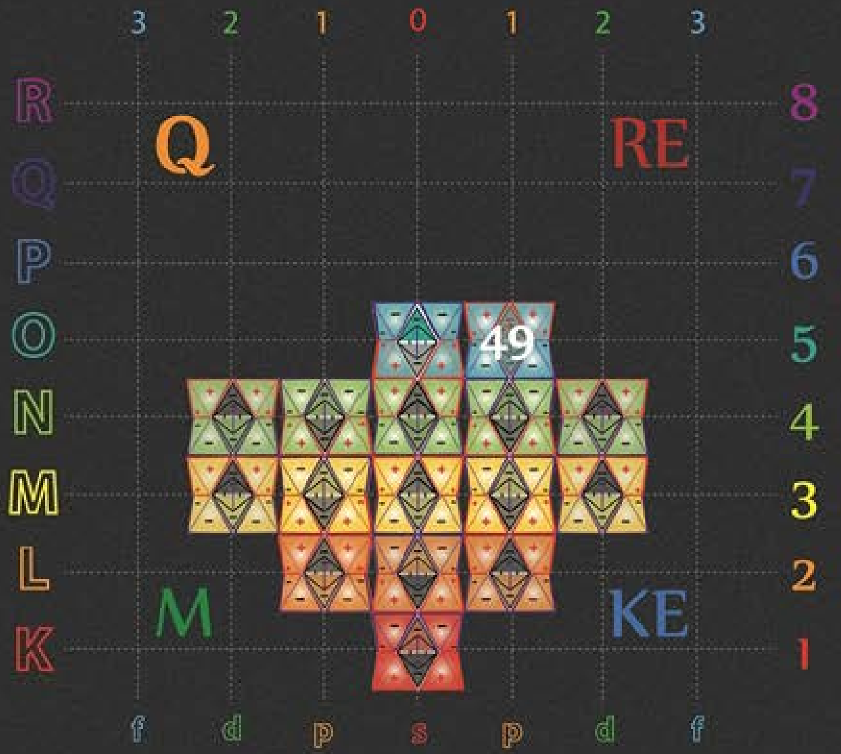
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
49
f
d
p
s
p
d
f

electron configuration

[Kr] 4d10 5s2 5p1
36

49 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5

Azmithal & Magnetic numbers

Aufbau

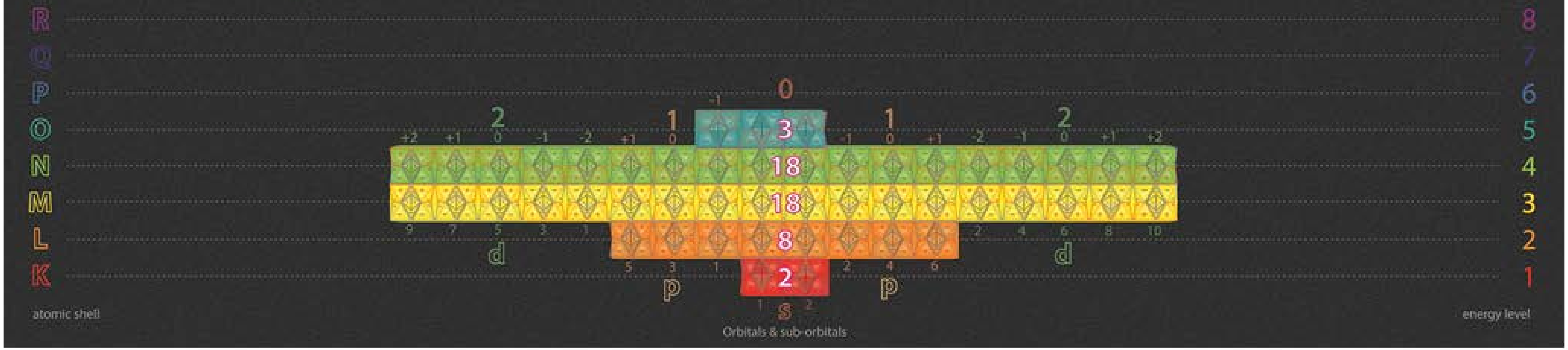
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
2
1
1
2
+2
+1
0
-1
-2
+1
0
-1
-1
0
+1
-2
-1
0
+1
+2
3
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.49 - Indium atom

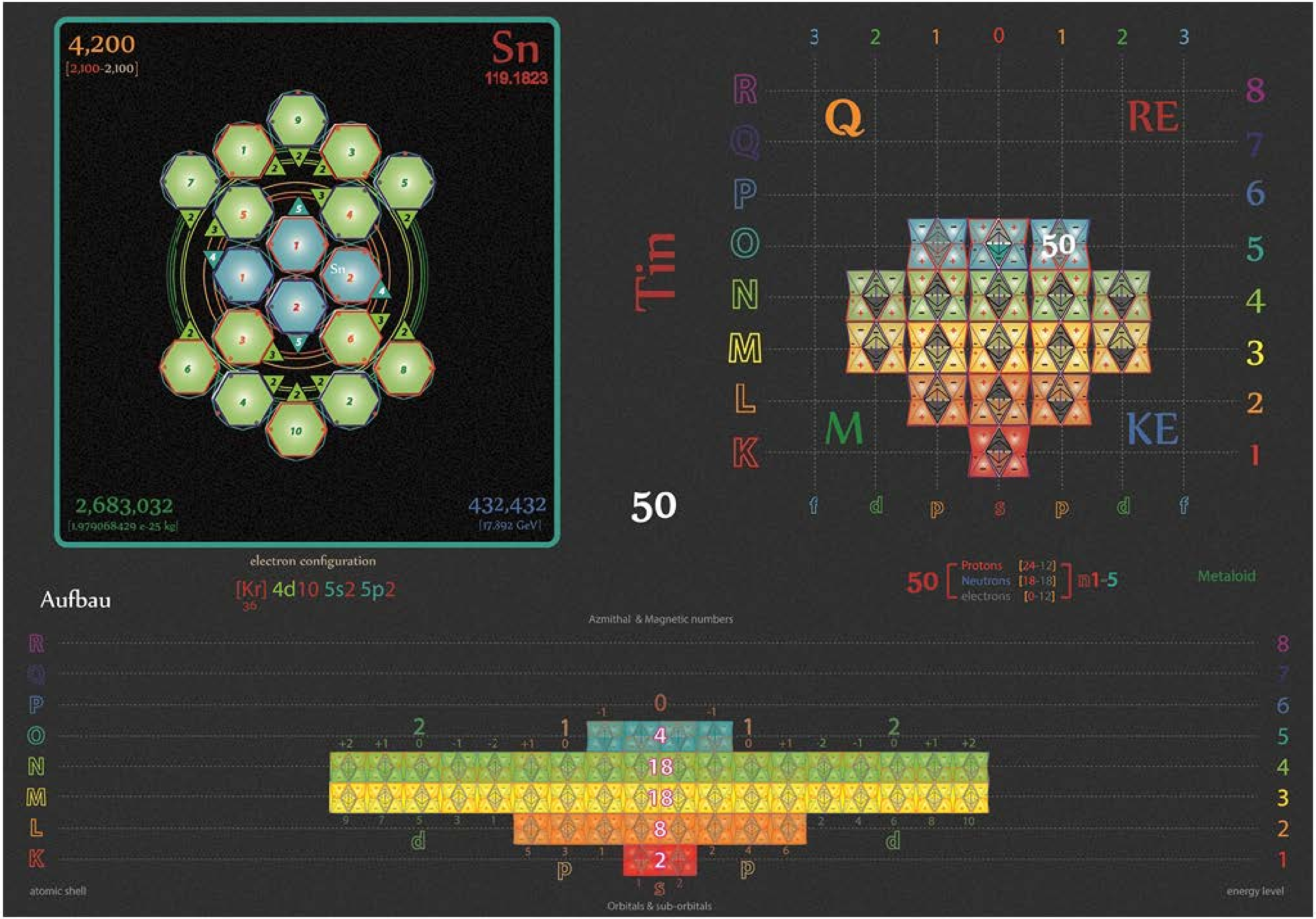
4,200
[2,100-2,100]
Sn
119.1823
2,683,032
[1.979068429 e-25 kg]
432,432
[17.392 GeV]
Tin
50
R
Q
P
O
N
M
L
K
3
2
1
0
1
2
3
8
7
6
5
4
3
2
1
Q
RE
M
KE
50
f
d
p
s
p
d
f
electron configuration
[Kr] 4d10 5s2 5p2
36
Aufbau
50
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-5
Metaloid
Azmithal & Magnetic numbers
4
18
18
8
2
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.50 - Tin atom

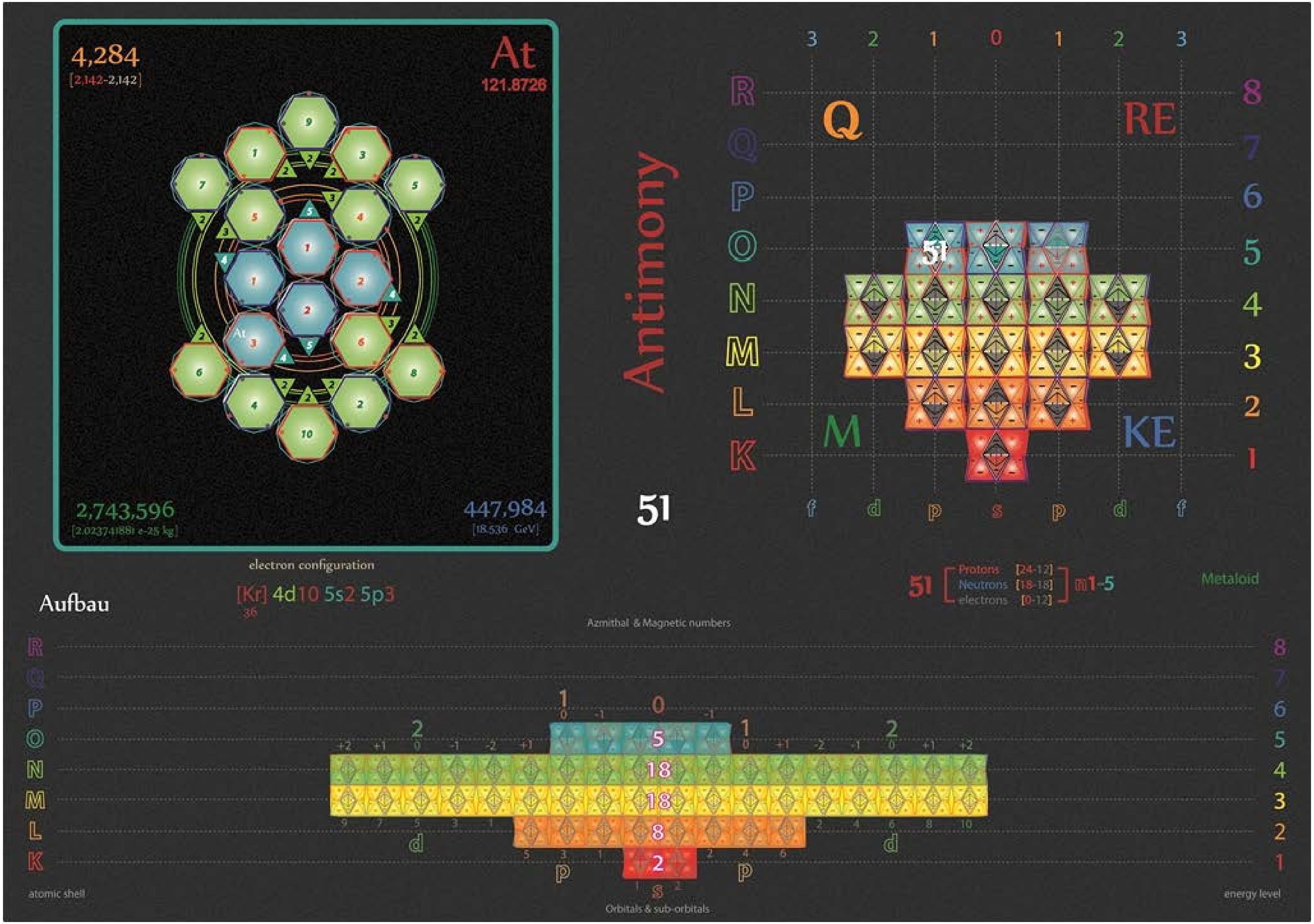
4,284
[2,142-2,142]
At
121.8726
2,743,596
[2.02374188i e-25 kg]
447,984
[18.536 GeV]
Antimony
51
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Kr] 4d10 5s2 5p3
36
51 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
5
18
18
8
2
d
p
s
Orbitals & sub-orbitals
atomic shell
energy level

Tetryonics 51.51 - Antimony atom

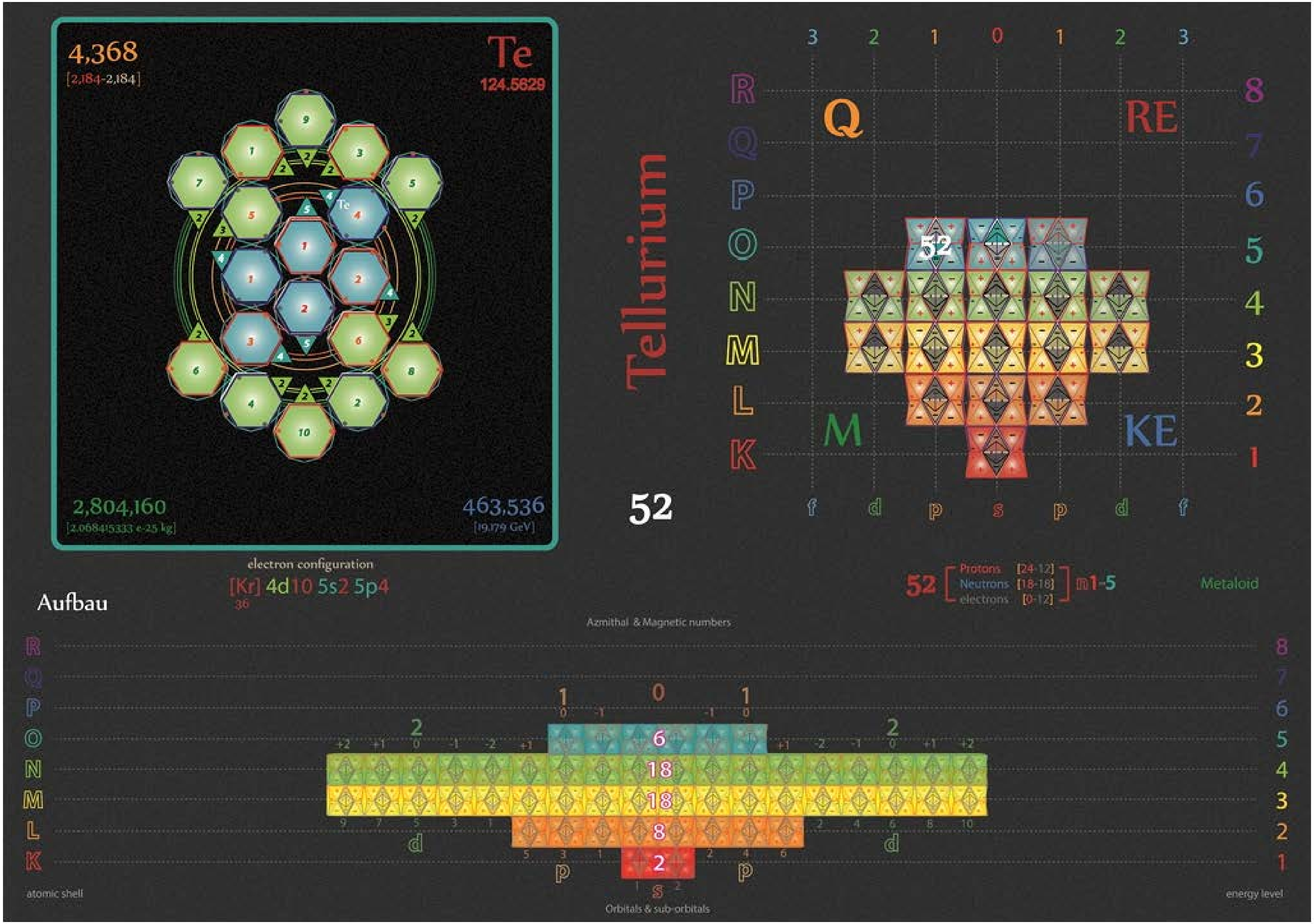
4,368
[2,184-2,184]
Te
124.5629
2,804,160
[2.068415333 e-25 kg]
463,536
[19.179 GeV]
Tellurium
52
electron configuration
[Kr] 4d10 5s2 5p4
36
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
52 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Metaloid
Azmithal & Magnetic numbers
6
18
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.52 - Tellurium atom

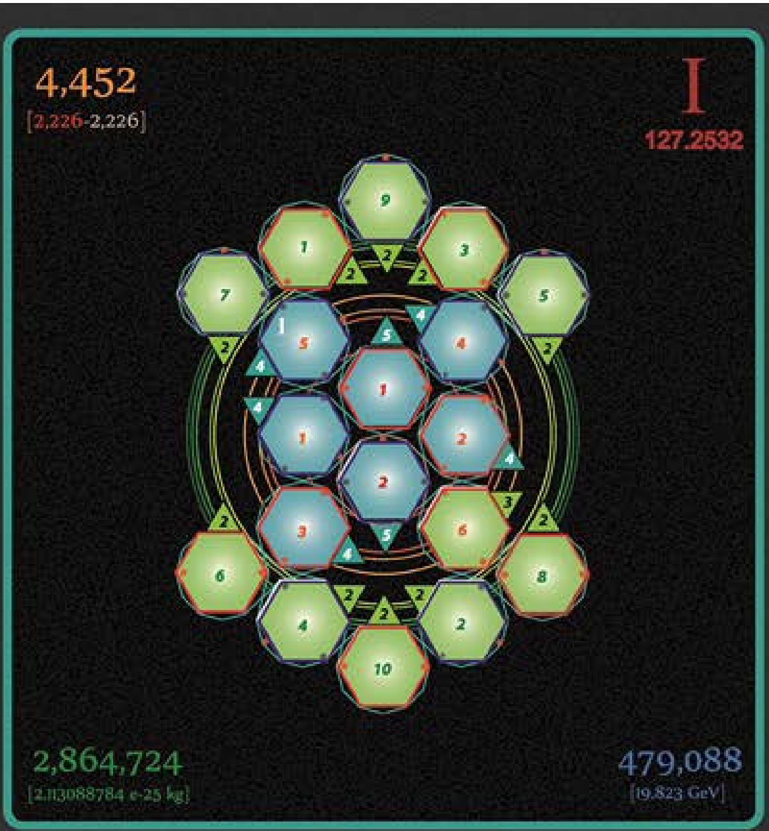
4,452
[2,226-2,226]
I
127.2532
2,864,724
[2.113088784 e-25 kg]
479,088
[19.823 GeV]

Iodine

53

electron configuration

[Kr] 4d10 5s2 5p5
36

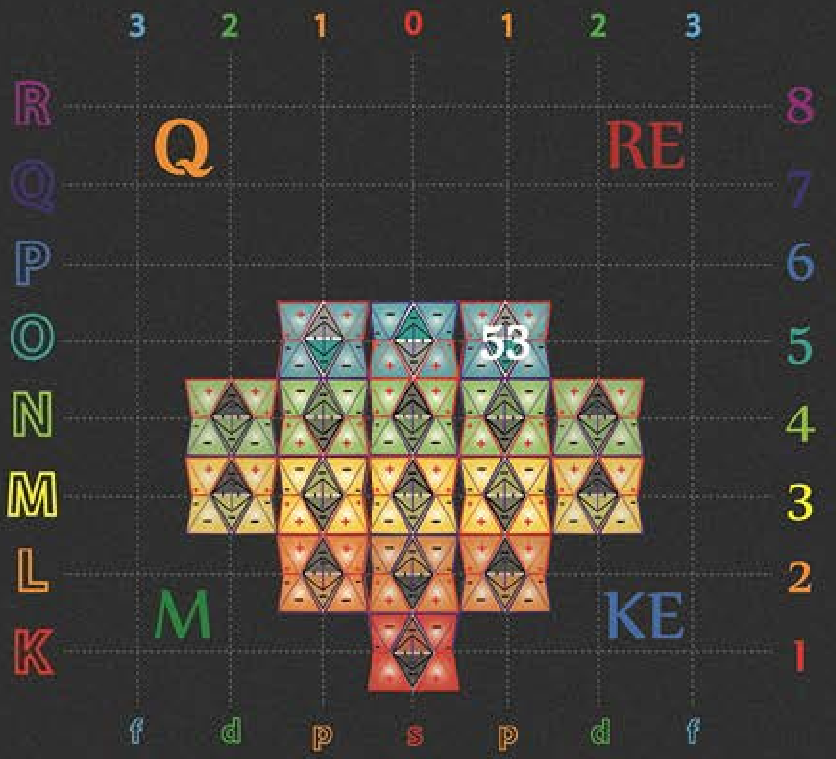
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
53
f d p s p d f

53 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-5

Halogen

Aufbau

Azmithal & Magnetic numbers

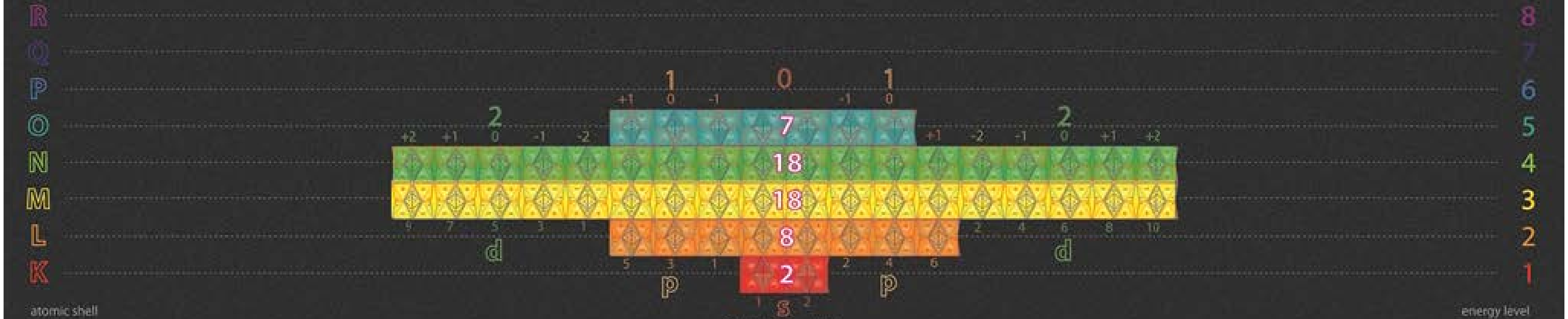
R Q P O N M L K
8 7 6 5 4 3 2 1
1 0 1
+1 0 -1 -1 0
2 2
+2 +1 0 -1 -2 +1 -2 -1 0 +1 +2
7
18
18
8
2
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.53 - Iodine atom

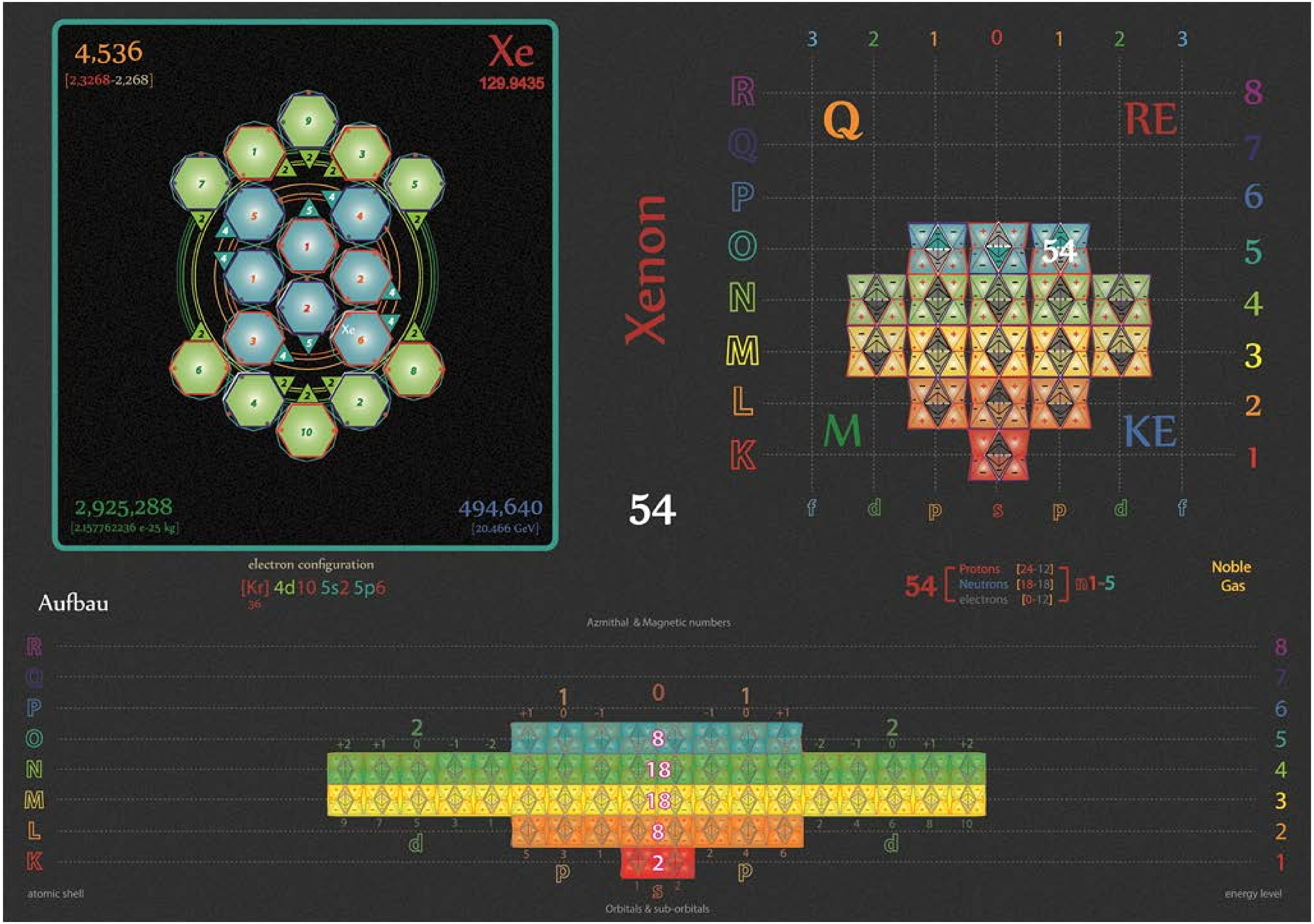
4,536
[2,3268-2,268]
Xe
129.9435
2,925,288
[2.157762236 e-25 kg]
494,640
[20.466 GeV]
Xenon
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Kr] 4d10 5s2 5p6
36
Aufbau
54 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-5
Noble Gas
Azmithal & Magnetic numbers
8
18
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.54 - Xenon atom

Copyright ABRAHAM [2008] - All rights reserved

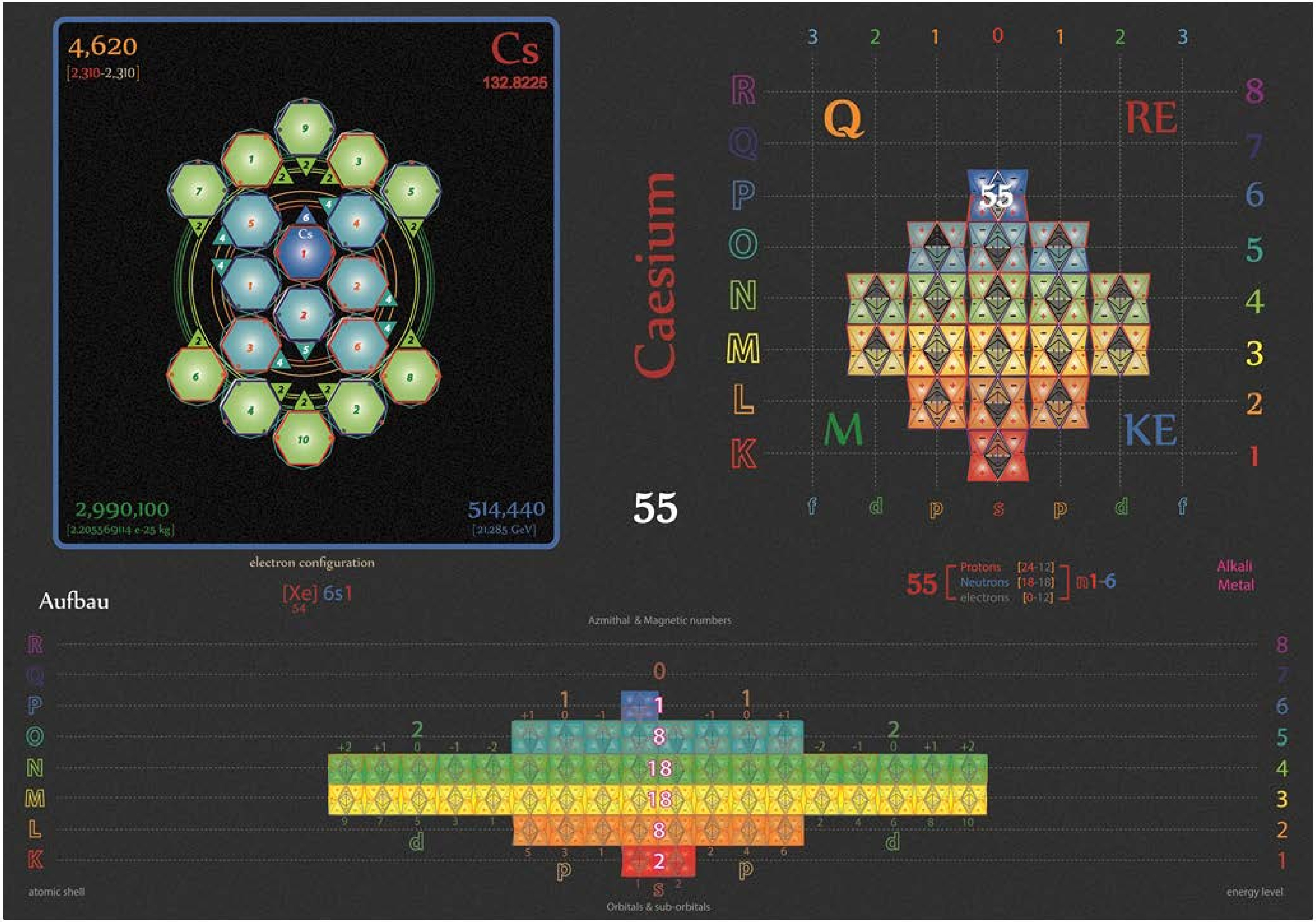
4,620
[2,310-2,310]
Cs
132.8225
2,990,100
[2.205569114 e-25 kg]
514,440
[21.285 GeV]
electron configuration
Aufbau
[Xe] 6s1
54
Caesium
55
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
55 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Alkali Metal
Azmithal & Magnetic numbers
0 1 8 18 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals
s p d

Tetryonics 51.55 - Caesium atom

Copyright ABRAHAM [2008] - All rights reserved

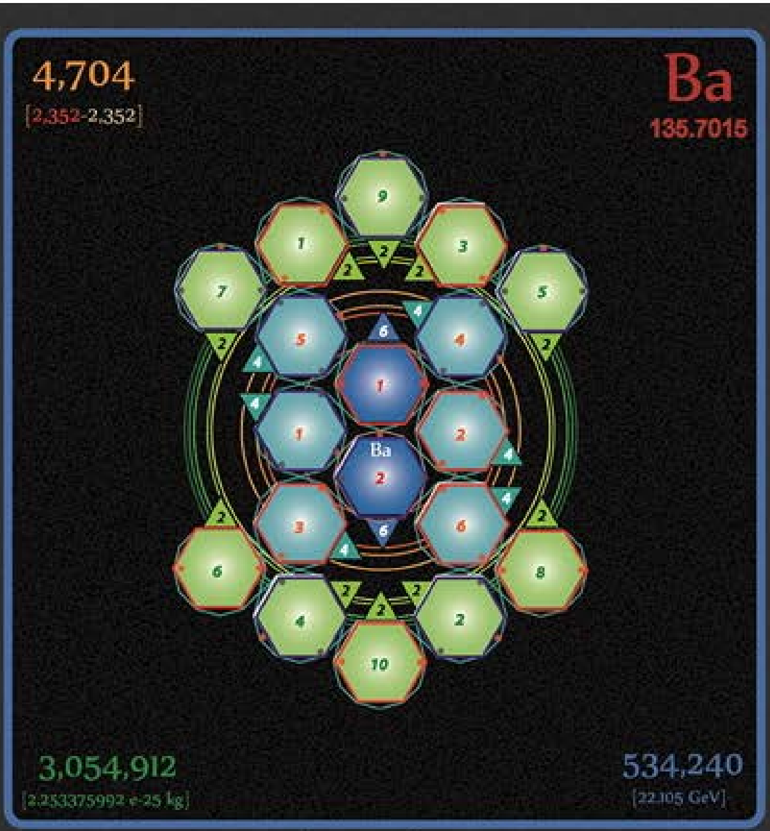
4,704
[2,352-2,352]
Ba
135.7015
9
1
3
7
5
2
5
4
6
1
1
2
Ba
2
3
6
4
6
8
4
2
10
3,054,912
[2.253375992 e-25 kg]
534,240
[22.105 GeV]

Barium

56

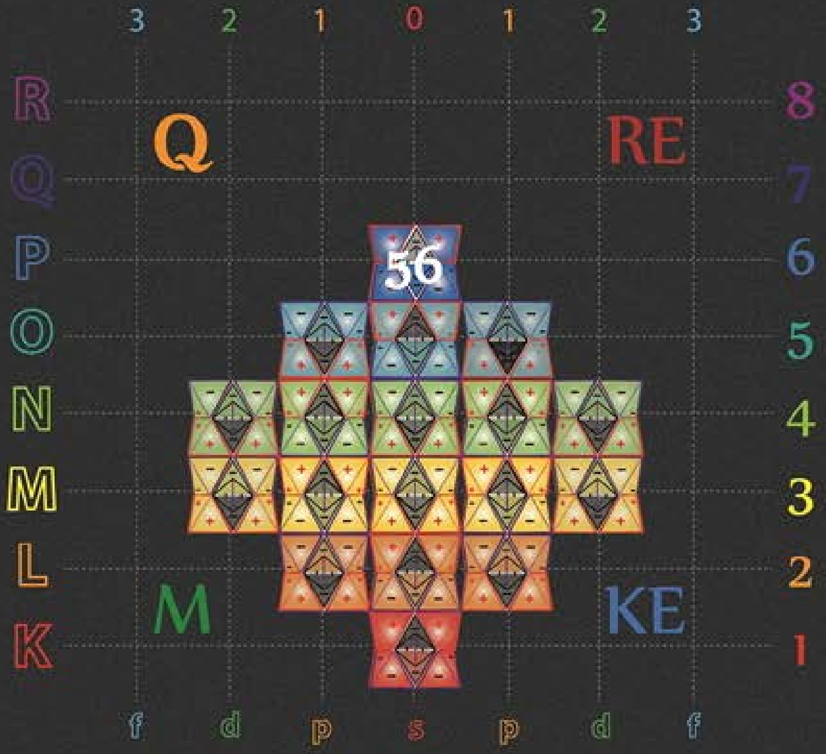
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
56
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

[Xe] 6s2
54

56 [Protons [24-12]
Neutrons [18-18]
electrons [0-12]] m1-6

Alkaline
Earth

Aufbau

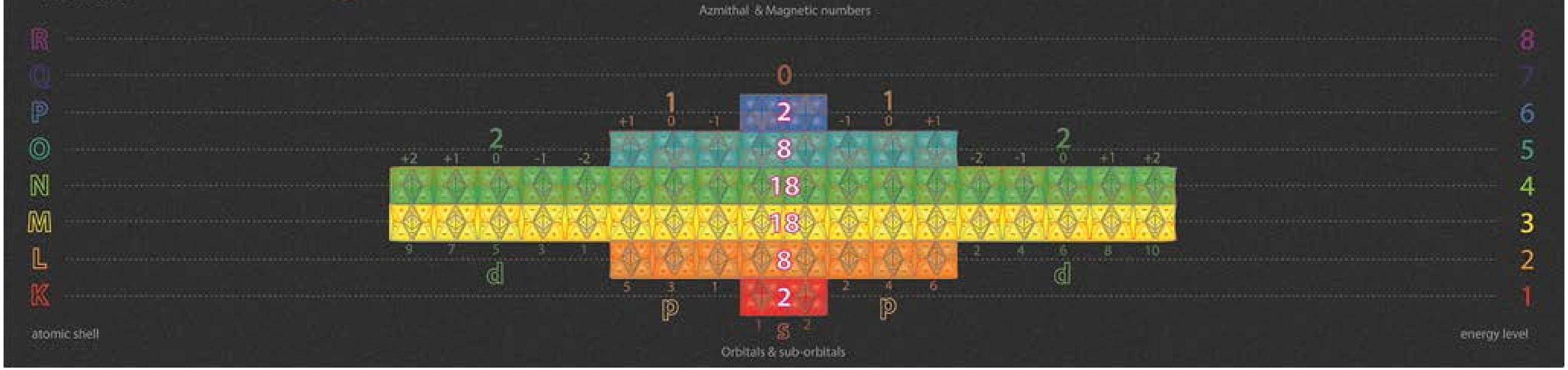
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
1
1
2
2
+1
0
-1
-1
0
+1
+2
+1
0
-1
-2
-2
-1
0
+1
+2
2
8
18
18
8
2
9
7
5
3
1
2
4
6
8
10
d
d
5
3
1
2
4
6
p
p
1
s
2
8
7
6
5
4
3
2
1
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.56 - Barium atom

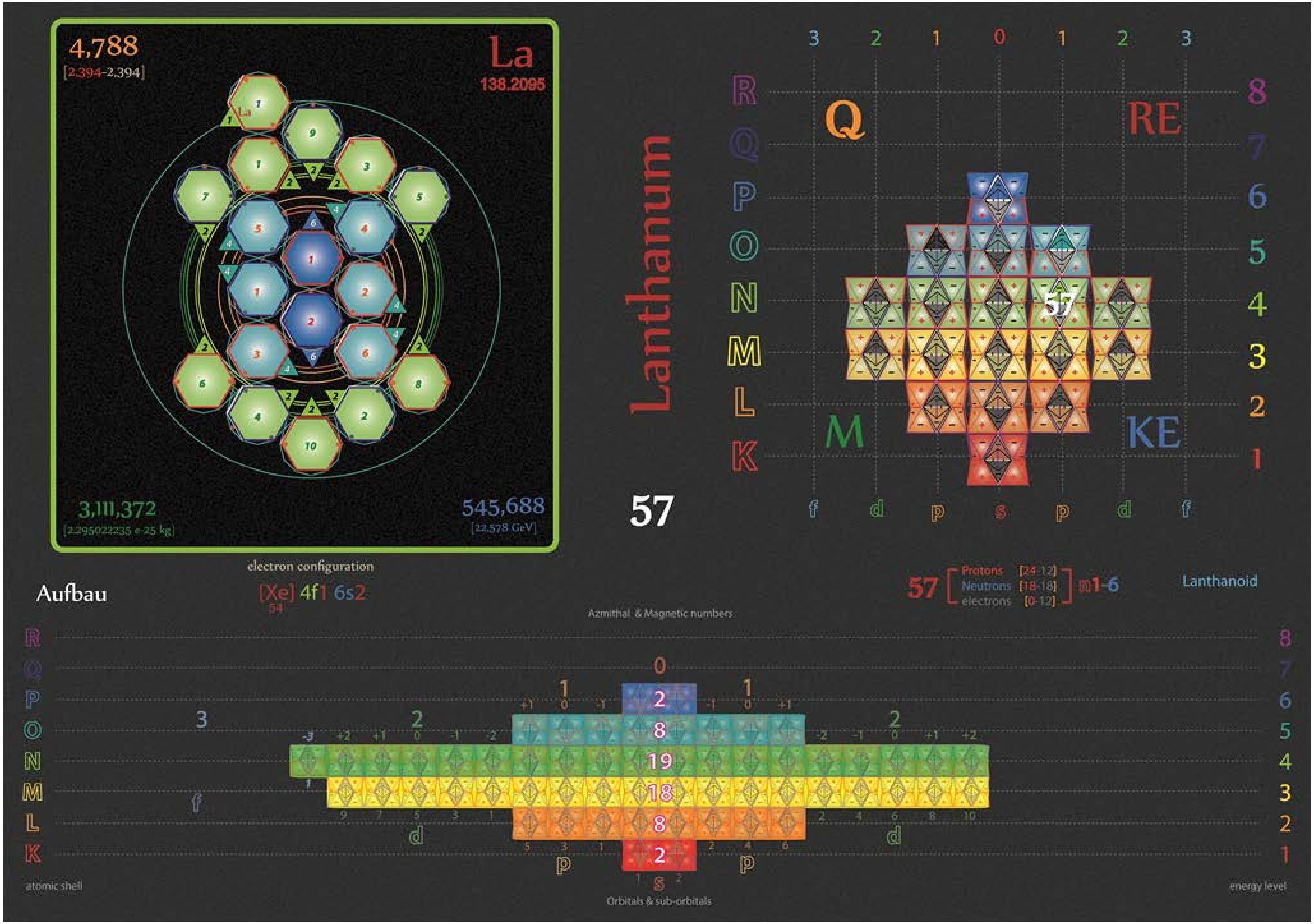
4,788
[2.394-2.394]
La
138.2095
3,111,372
[2.295022235 e-25 kg]
545,688
[22.578 GeV]
Lanthanum
57
Q
RE
M
KE
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Xe] 4f1 6s2
54
Aufbau
57 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1 1
2 2
3
2 8 19 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.57 - Lanthanum atom

Copyright ABRAHAM [2008] - All rights reserved

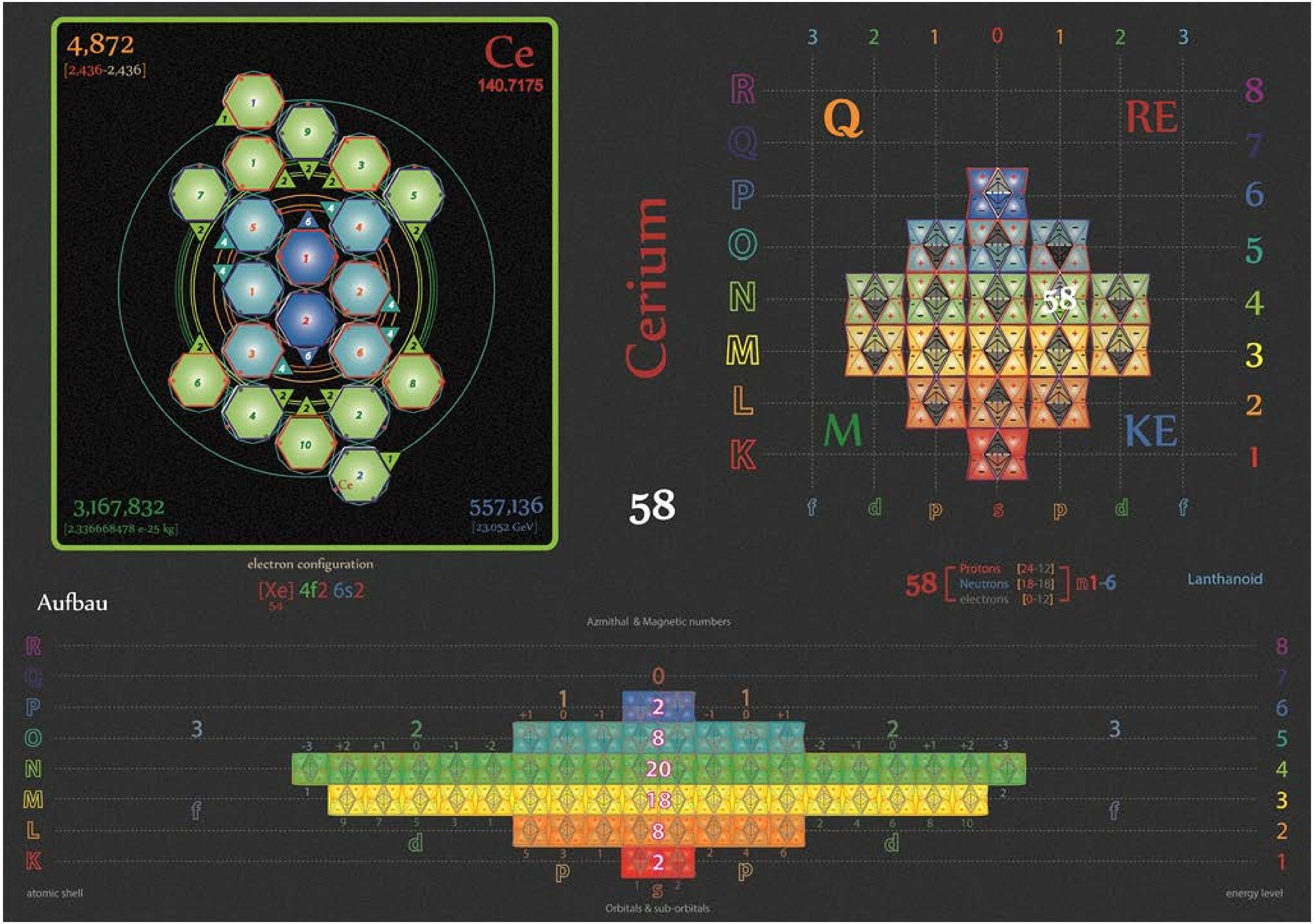
4,872
[2,436-2,436]
Ce
140.7175
3,167,832
[2.336668478 e-25 kg]
557,136
[23.052 GeV]
Cerium
58
electron configuration
[Xe] 4f2 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
58 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
8
20
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.58 - Cerium atom

Copyright ABRAHAM [2008] - All rights reserved

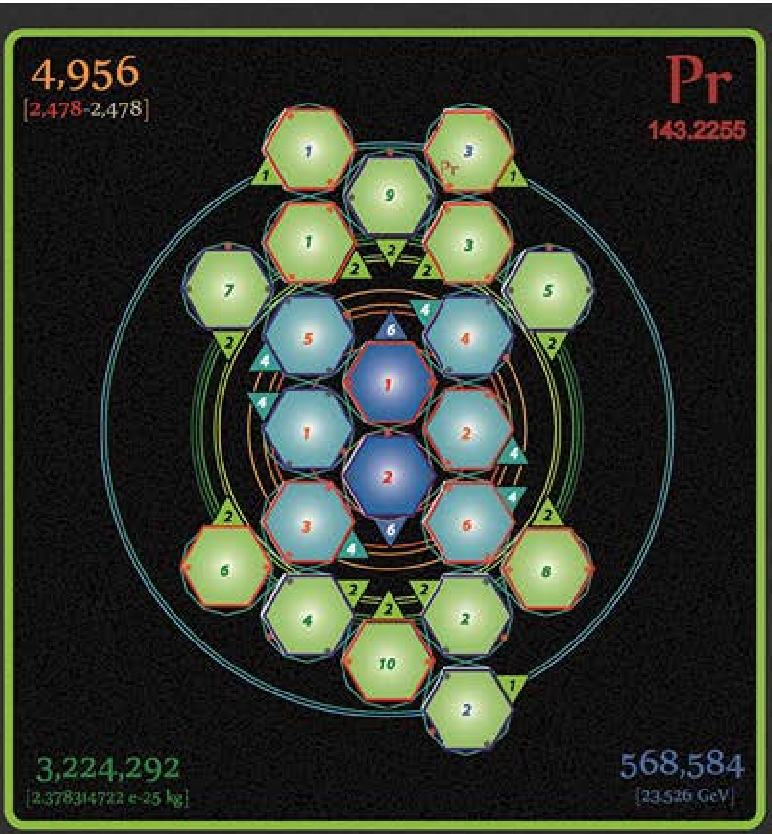
4,956
[2,478-2,478]
Pr
143.2255
Pr
3,224,292
[2.378314722 e-25 kg]
568,584
[23.526 GeV]

Praseodymium

59

electron configuration

[Xe] 4f3 6s2
54

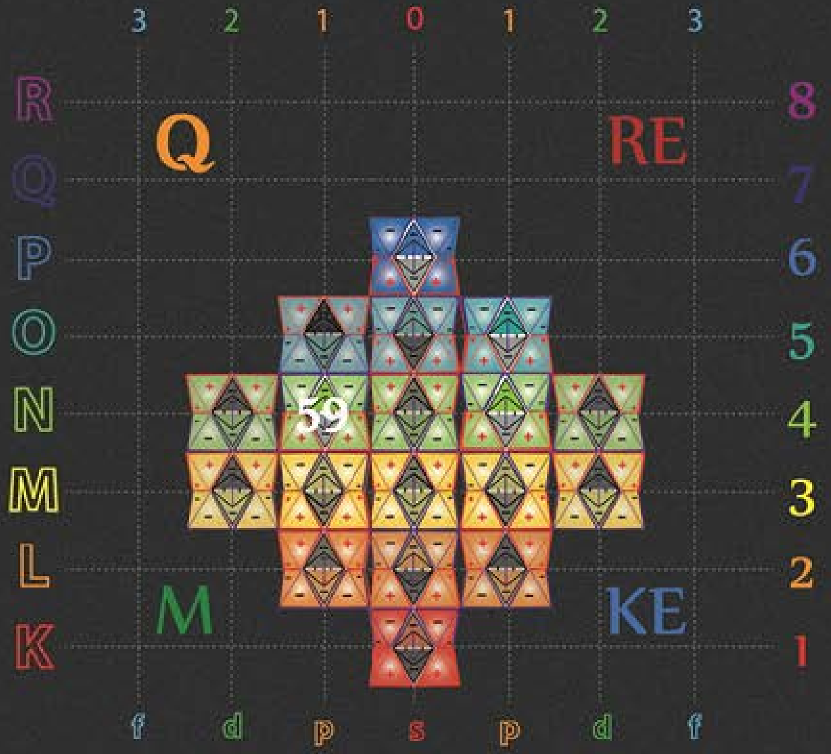
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
59
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

59 [Potons [24-12] Neutrons [18-18] electrons [0-12]] n1-6

Lanthanoid

Aufbau

Azmithal & Magnetic numbers

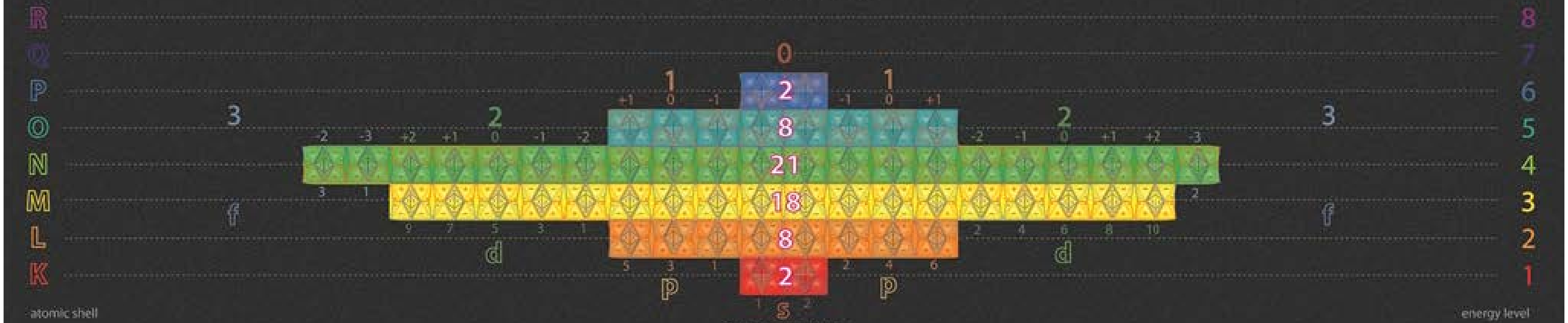
R
Q
P
O
N
M
L
K
0
2
1
1
+1
0
-1
-1
0
+1
3
2
8
2
3
-2
-3
+2
+1
0
-1
-2
-2
-1
0
+1
+2
-3
21
3
1
18
2
f
f
9
7
5
3
1
2
4
6
8
10
8
d
d
5
3
1
2
4
6
2
p
p
1
s
2
8
7
6
5
4
3
2
1

atomic shell

energy level

Orbitals & sub-orbitals

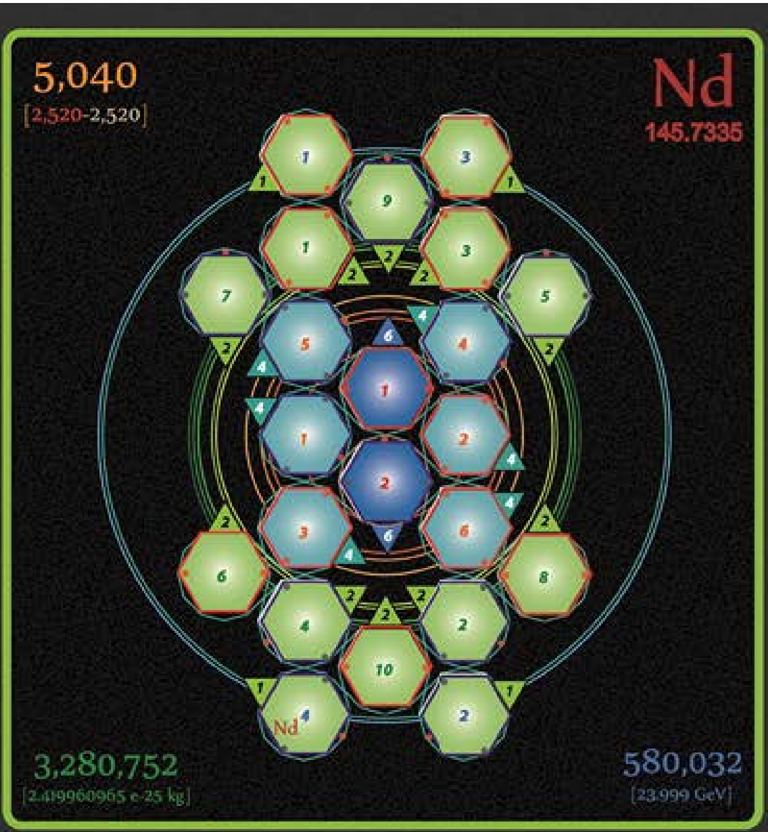
5,040
[2,520-2,520]
Nd
145.7335
1
3
9
1
3
7
5
5
6
4
1
1
2
2
3
6
2
6
8
4
2
10
Nd
4
2
3,280,752
[2.41996096 5 e-25 kg]
580,032
[23.999 GeV]

electron configuration

[Xe] 4f4 6s2
54

Neodymium

60

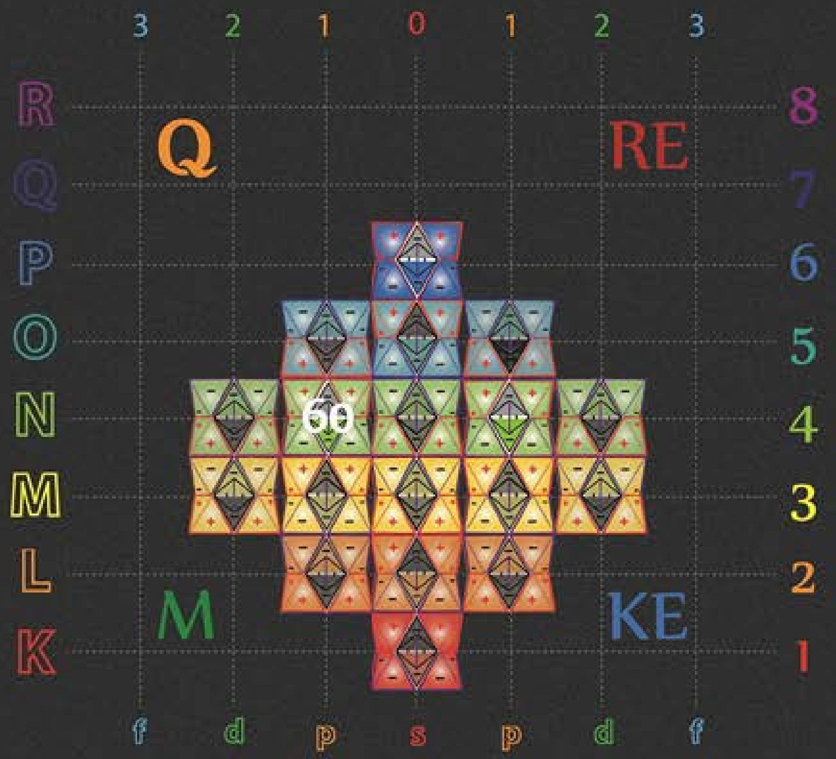
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
60
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

60 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Lanthanoid

Aufbau

Azmithal & Magnetic numbers

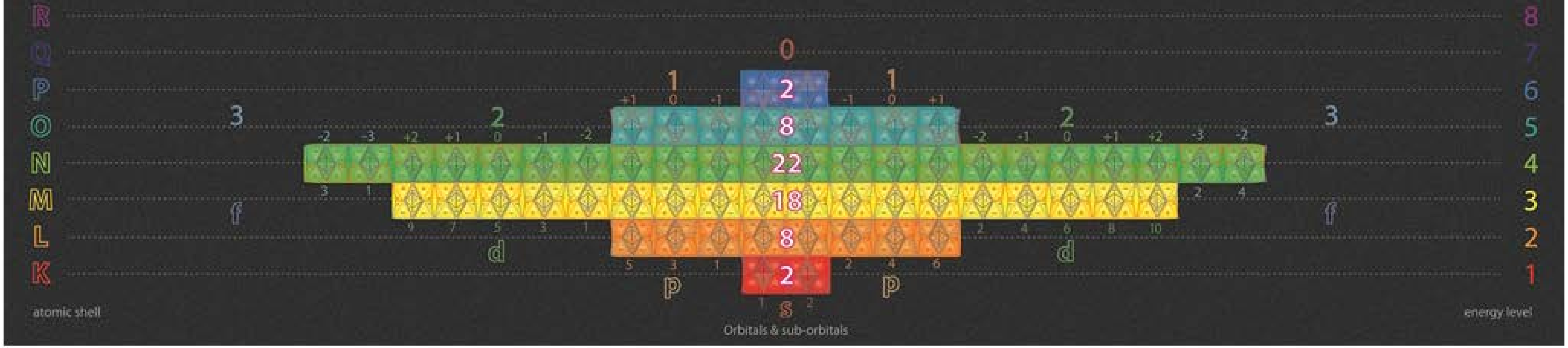
R
Q
P
O
N
M
L
K
0
1
1
2
2
3
3
2
8
22
18
8
2
f
d
p
s
p
d
f
8
7
6
5
4
3
2
1
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.60 - Neodymium atom

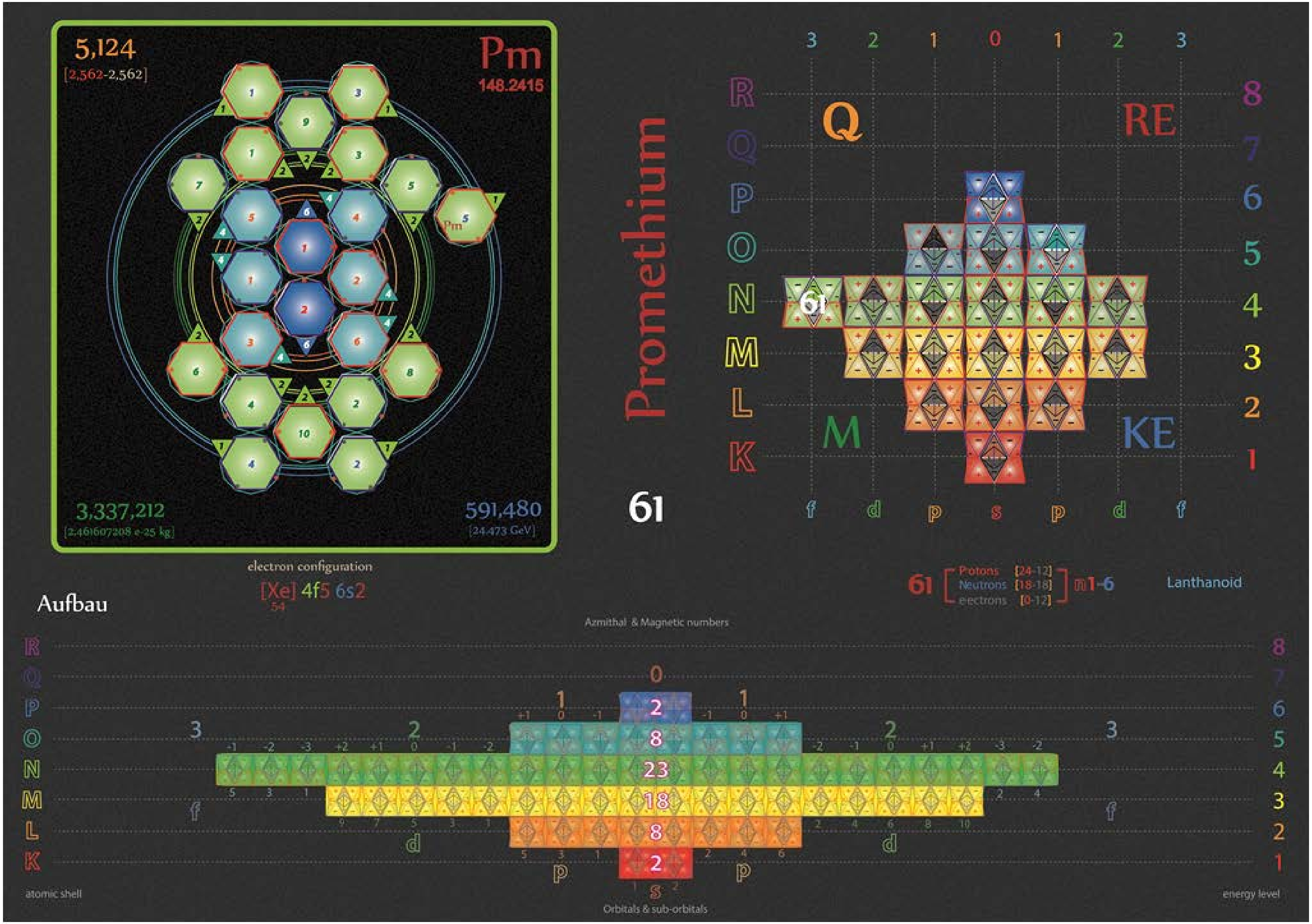
5,124
[2,562-2,562]
Pm
148.2415
3,337,212
[2.461607208 e-25 kg]
591,480
[24.473 GeV]
Promethium
61
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
[Xe] 4f5 6s2
54
61
Potons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
8
23
18
8
2
f
f
d
d
p
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.61 - Promethium atom

Copyright ABRAHAM [2008] - All rights reserved

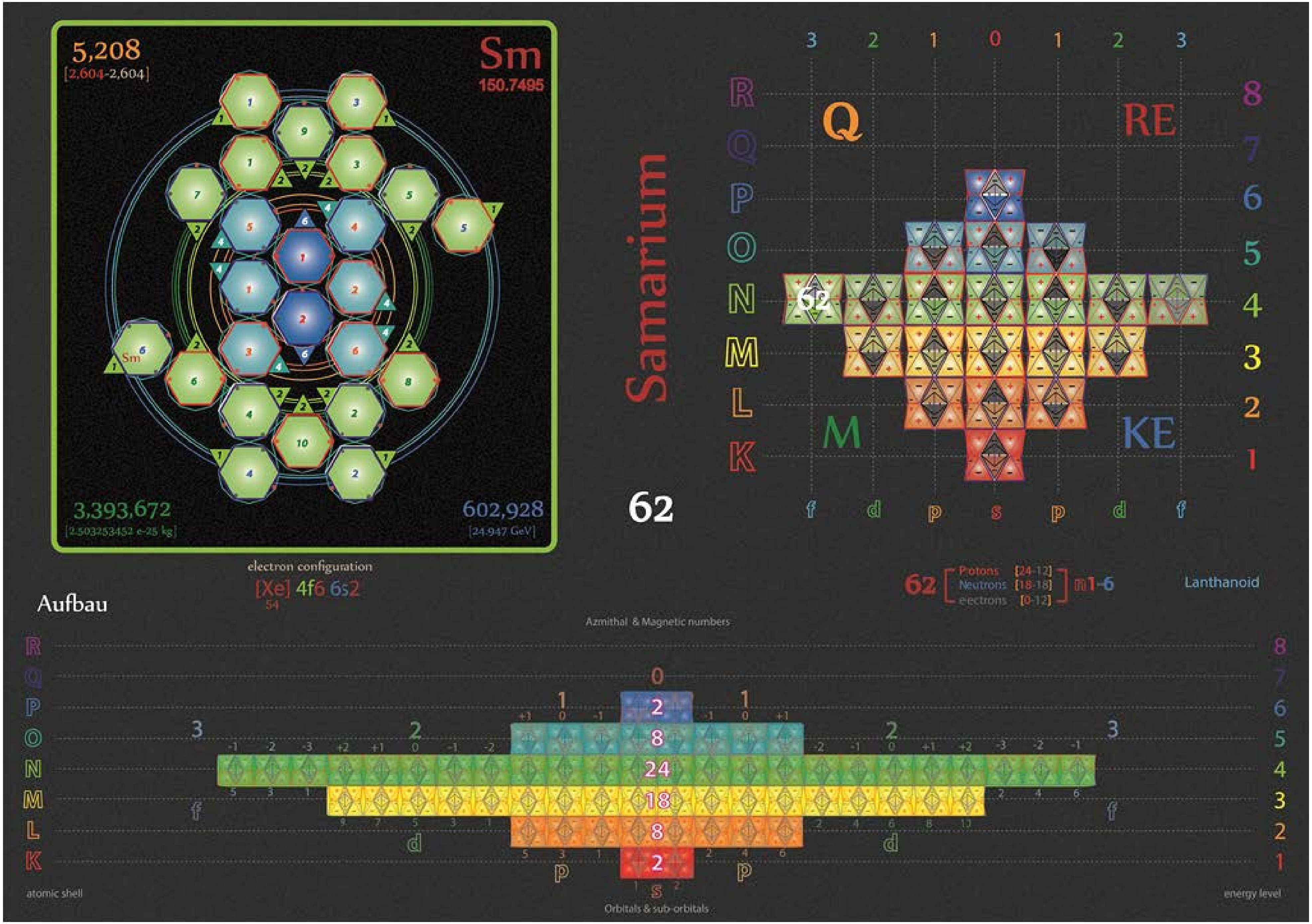
5,208
[2,604-2,604]
Sm
150.7495
3,393,672
[2.503253452 e-25 kg]
602,928
[24.947 GeV]
Samarium
62
electron configuration
[Xe] 4f6 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
62 Potons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
24
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.62 - Samarium atom

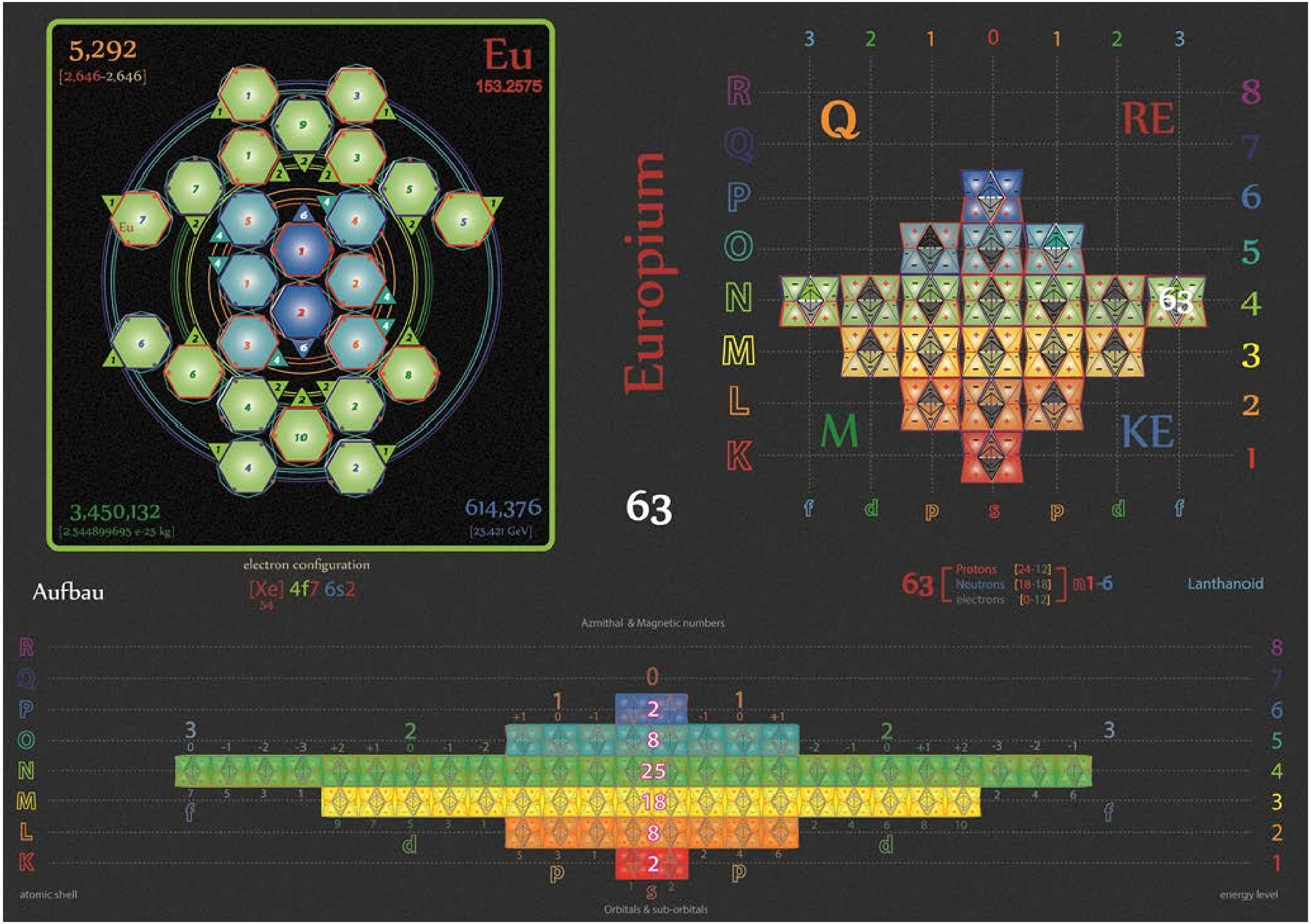
5,292
[2,646-2,646]
Eu
153.2575
3,450,132
[2.544899695 e-25 kg]
614,376
[25.421 GeV]
Europium
63
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Xe] 4f7 6s2
54
63 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
25
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.63 - Europium atom

Copyright ABRAHAM [2008] - All rights reserved

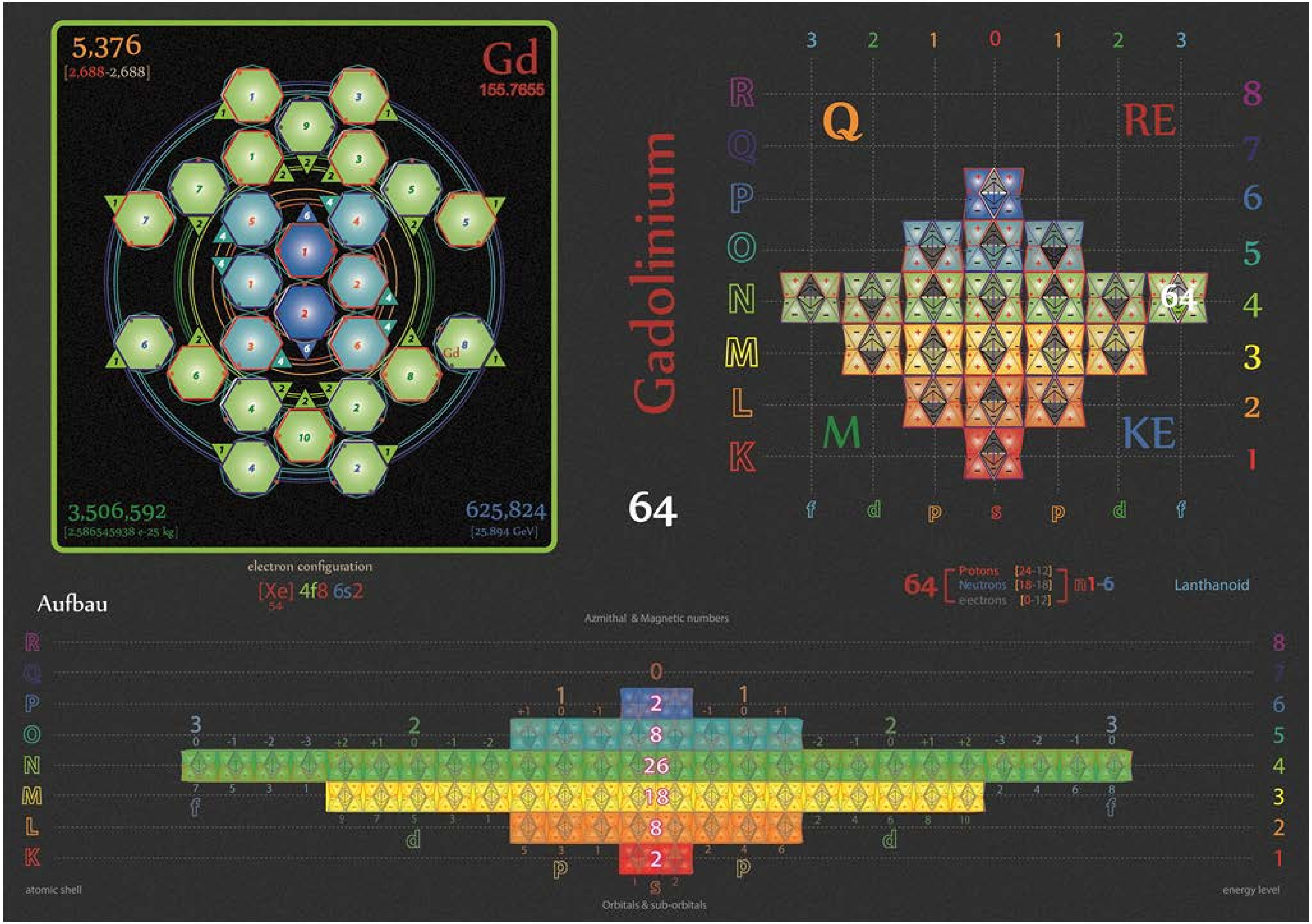
5,376
[2,688-2,688]
Gd
155.7655
3,506,592
[2.586545938 e-25 kg]
625,824
[25.894 GeV]
Gadolinium
64
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
64
f d p s p d f
electron configuration
[Xe] 4f8 6s2
54
Aufbau
64
Potons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
8
26
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.64 - Gadolinium atom

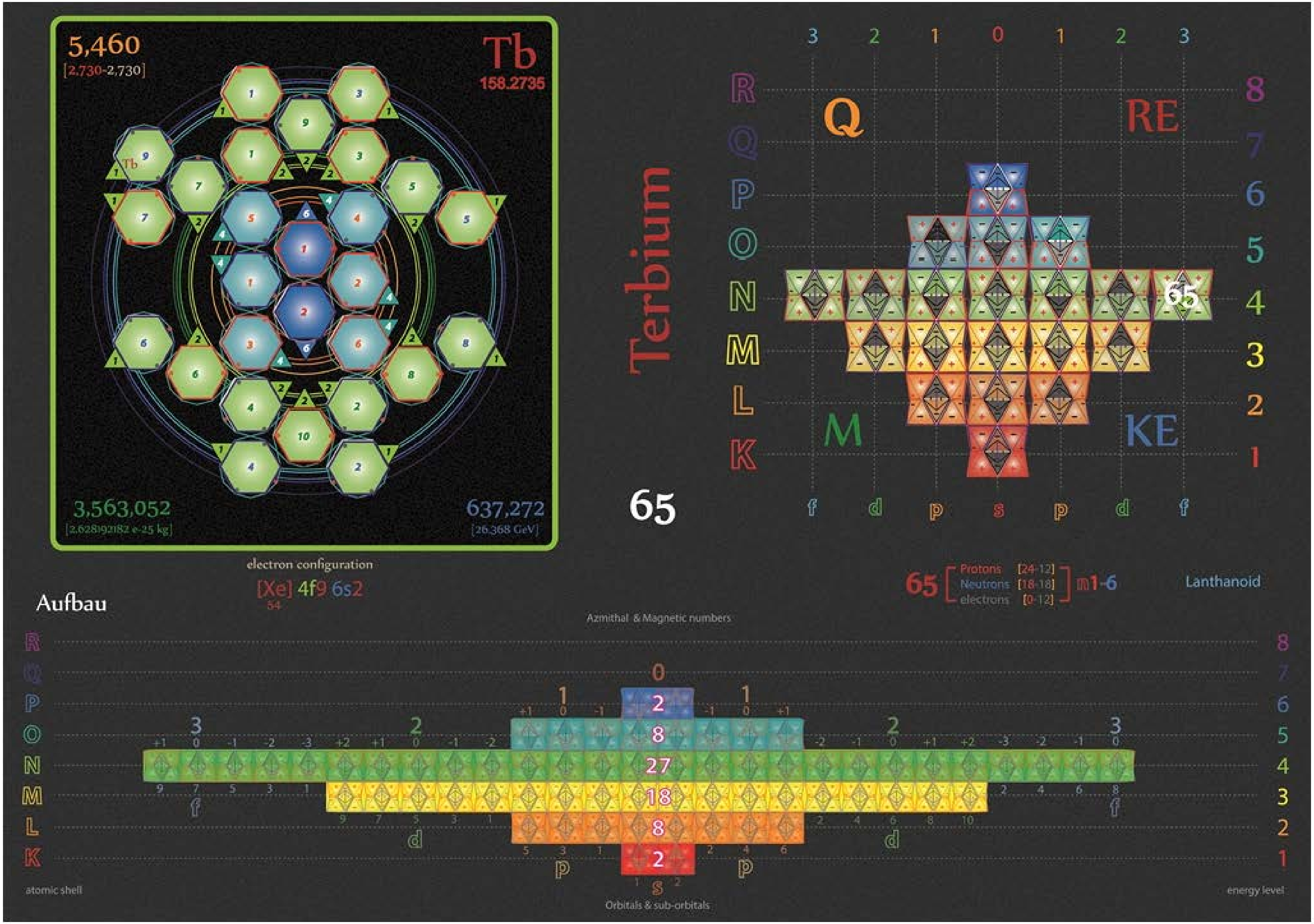
5,460
[2,730-2,730]
Tb
158.2735
3,563,052
[2.628192182 e-25 kg]
637,272
[26.368 GeV]
Terbium
65
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
[Xe] 4f9 6s2
54
65 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
2
8
27
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.65 -Terbium atom

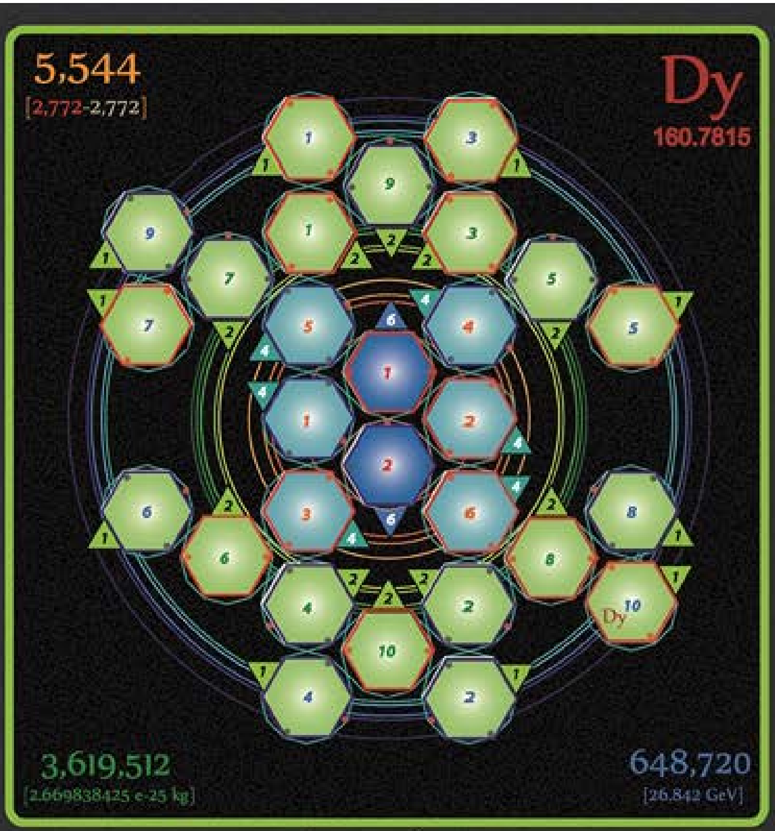
5,544
[2,772-2,772]
Dy
160.7815
Dy
3,619,512
[2.669838425 e-25 kg]
648,720
[26.842 GeV]

Dysprosium

66

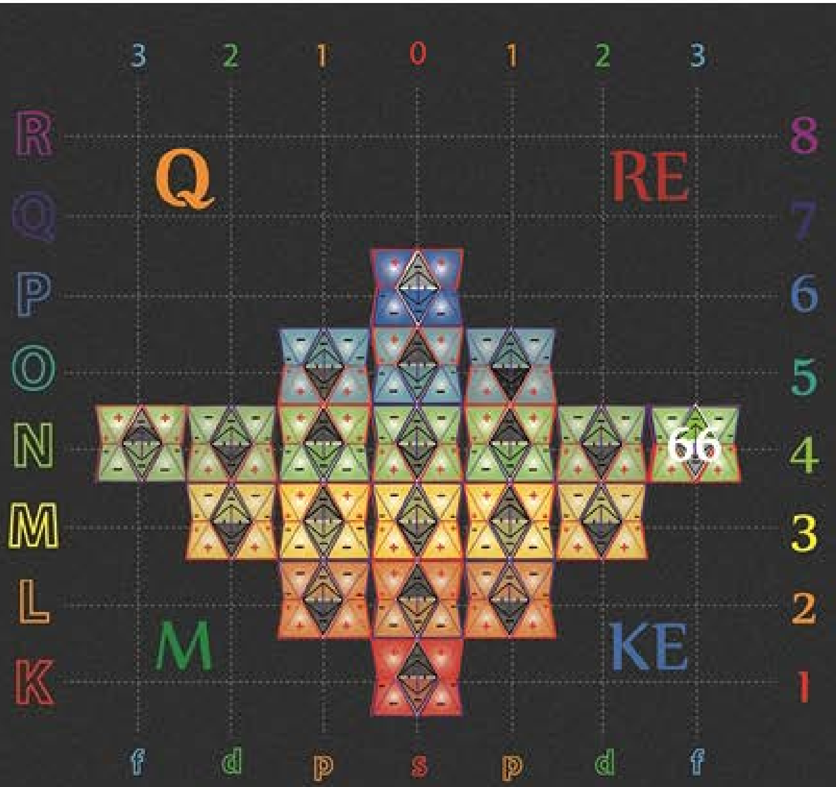
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
66
f
d
p
s
p
d
f

electron configuration

[Xe] 4f10 6s2
54

66 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Lanthanoid

Aufbau

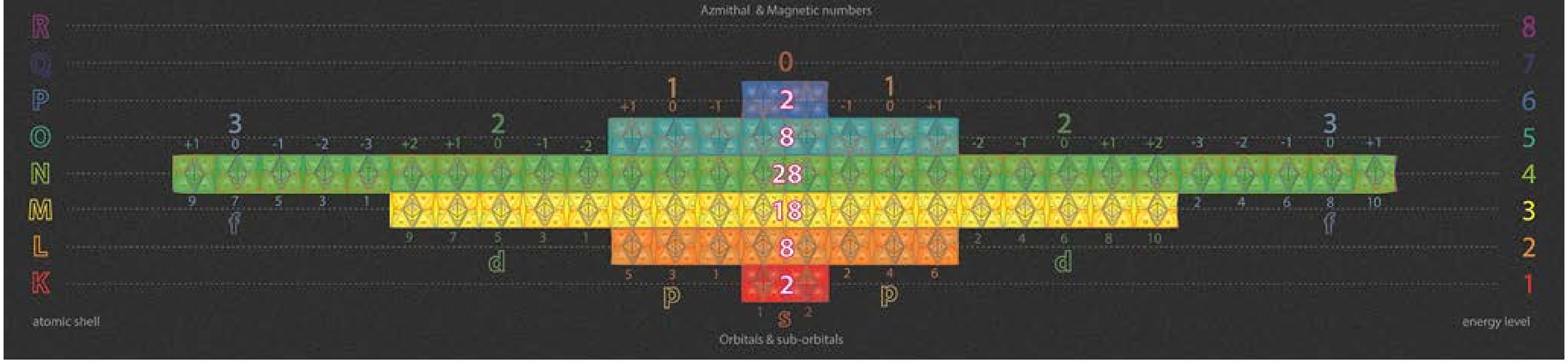
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
0
1
1
2
3
2
3
2
8
28
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.66 - Dysprosium atom

Copyright ABRAHAM [2008] - All rights reserved

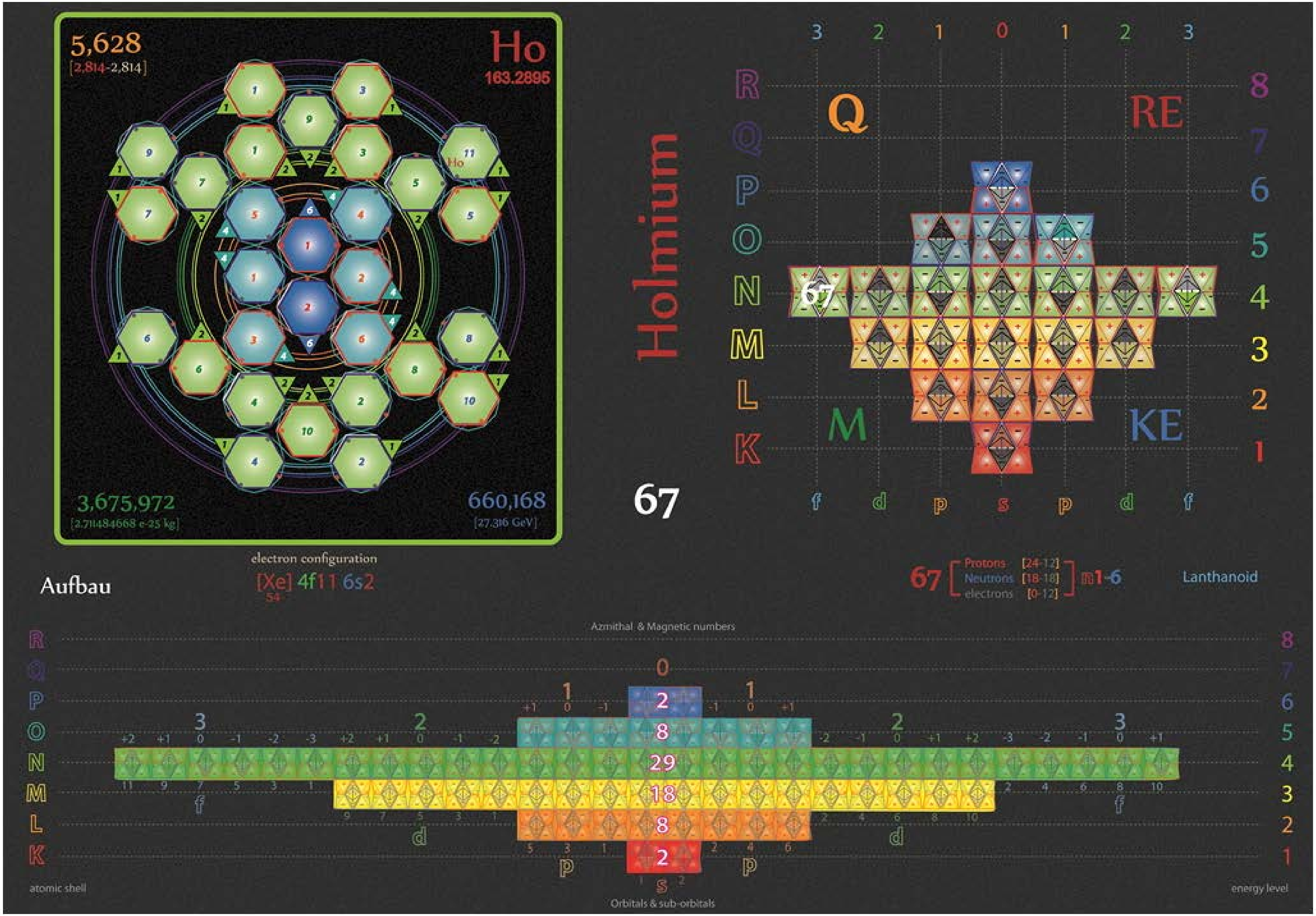
5,628
[2,814-2,814]
Ho
163.2895
3,675,972
[2.711484668 e-25 kg]
660,168
[27.316 GeV]
Holmium
67
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
Aufbau
electron configuration
[Xe] 4f11 6s2
54
67 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
1
2
3
2
8
29
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.67 - Holmium atom

Copyright ABRAHAM [2008] - All rights reserved

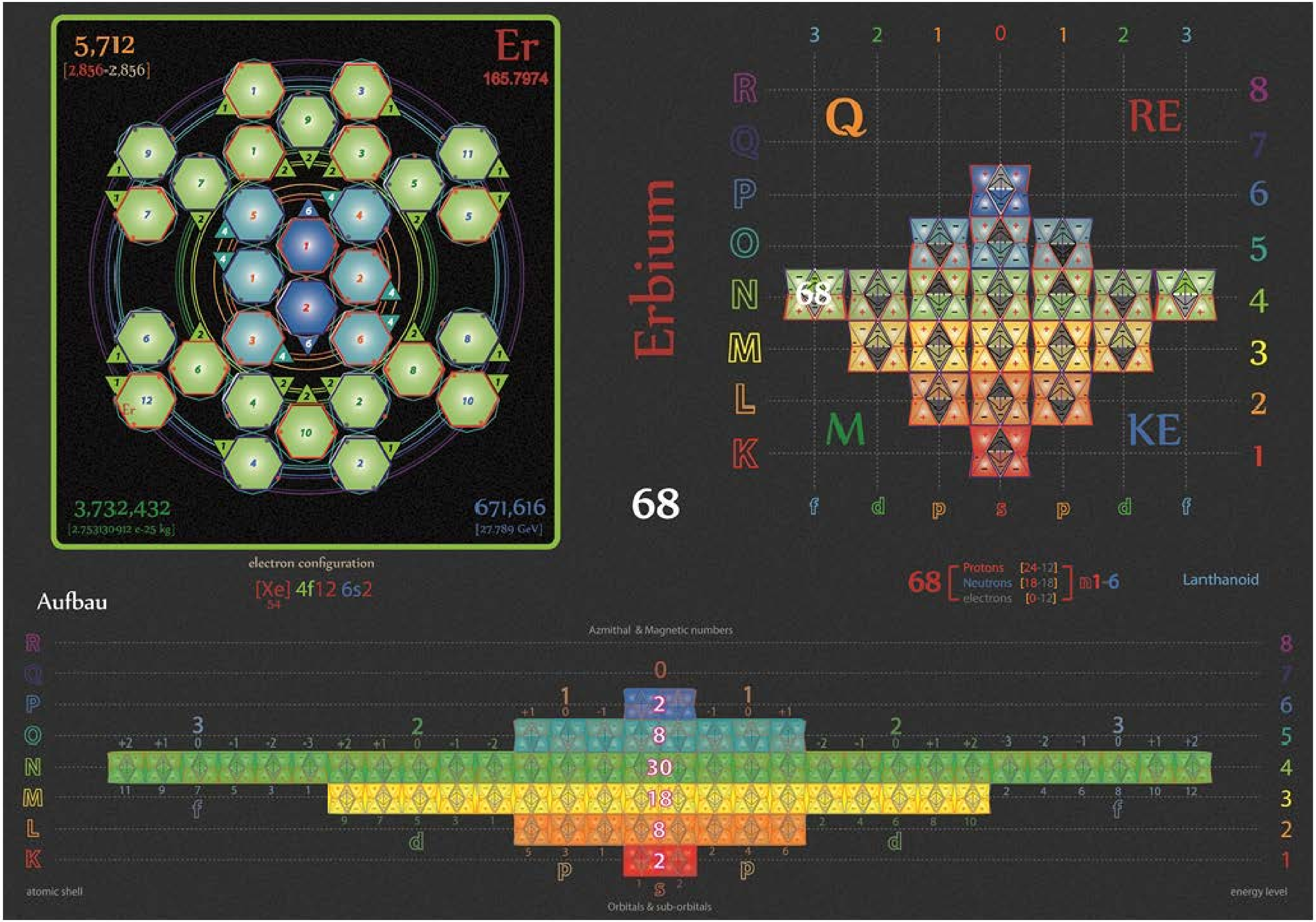
5,712
[2,856-2,856]
Er
165.7974
3,732,432
[2.753130912 e-25 kg]
671,616
[27.789 GeV]
Erbium
68
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
[Xe] 4f12 6s2
54
68 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Lanthanoid
Aufbau
Azmithal & Magnetic numbers
0
1
1
2
8
30
18
8
2
3
2
2
3
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.68 - Erbium atom

Copyright ABRAHAM [2008] - All rights reserved

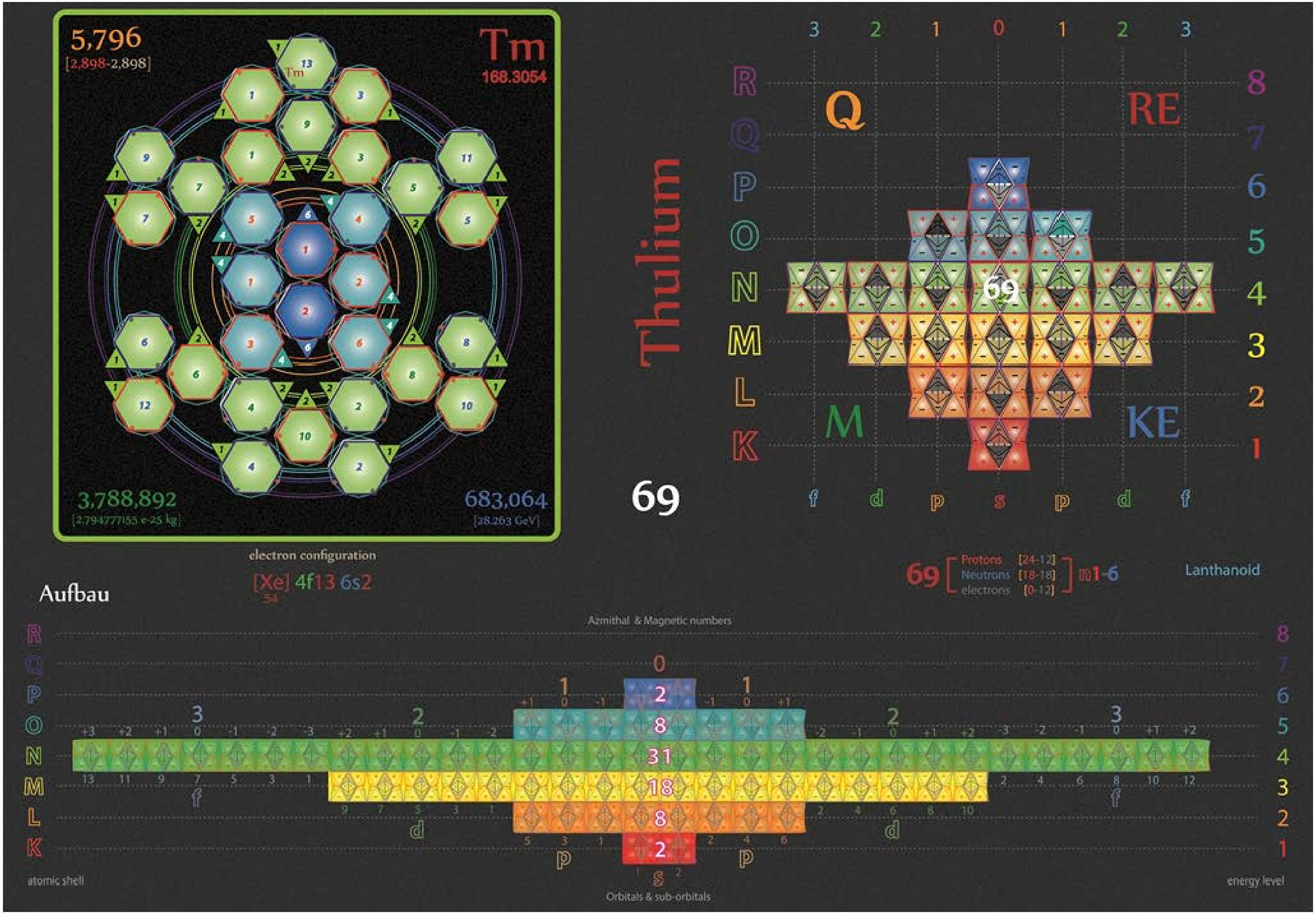
5,796
[2,898-2,898]
Tm
168.3054
3,788,892
[2.794777155 e-25 kg]
683,064
[28.263 GeV]
Thulium
69
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f13 6s2
54
Aufbau
69
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
31
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.69 - Thulium atom

Copyright ABRAHAM [2008] - All rights reserved

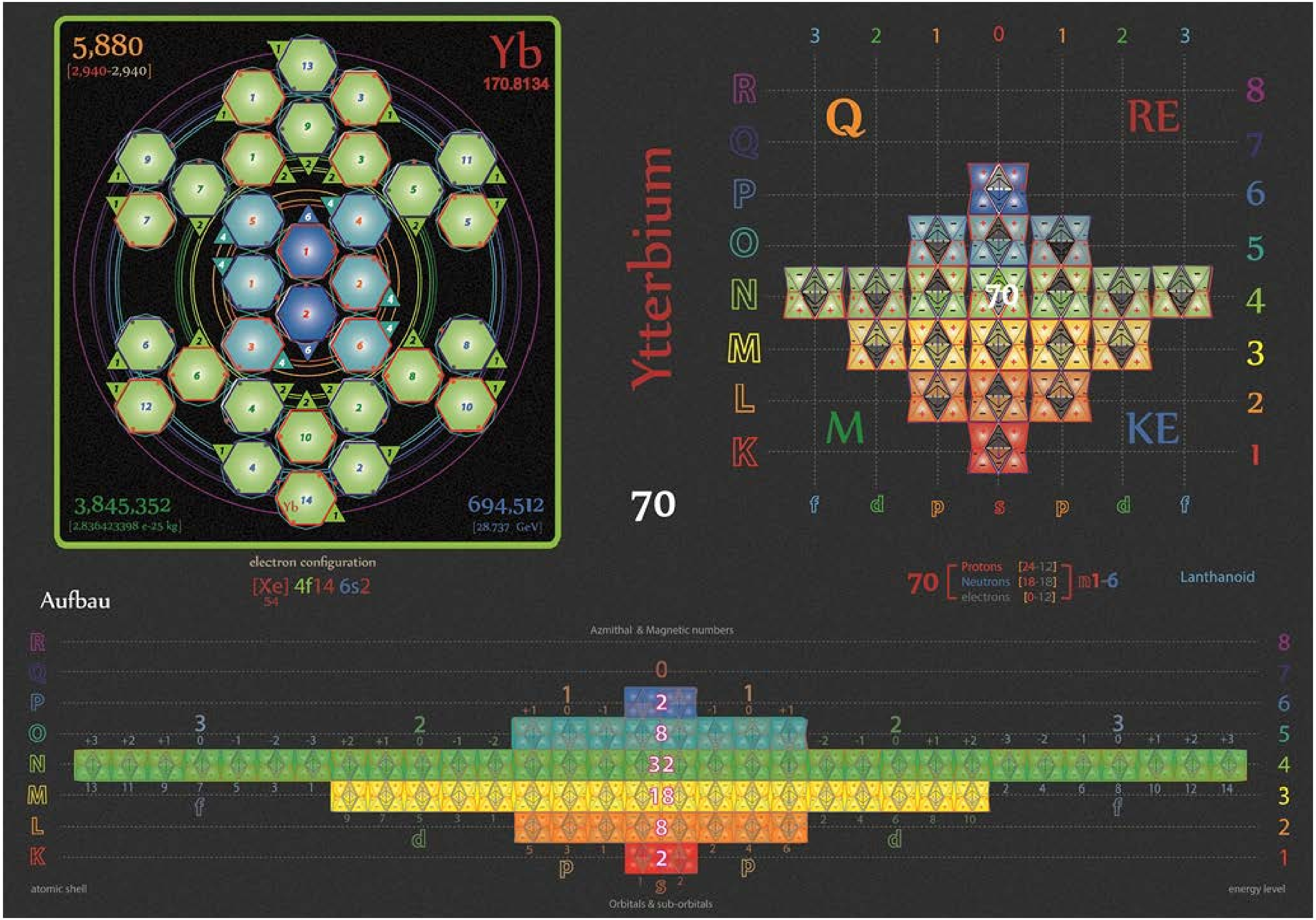
5,880
[2,940-2,940]
Yb
170.8134
3,845,352
[2.836423398 e-25 kg]
694,512
[28.737 GeV]
Ytterbium
70
electron configuration
[Xe] 4f14 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
70 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
2
8
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.70 - Ytterbium atom

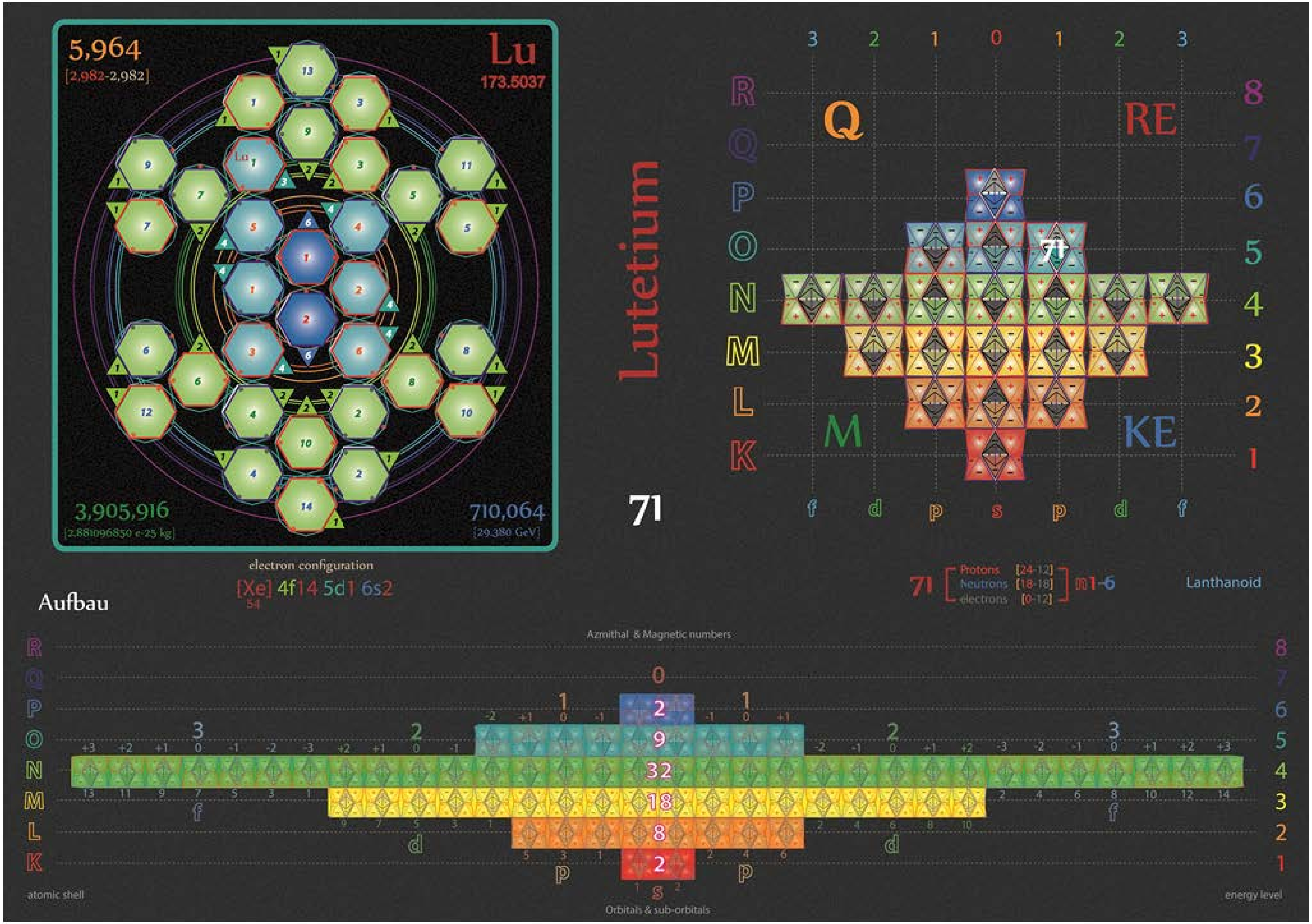
5,964
[2,982-2,982]
Lu
173.5037
3,905,916
[2.881096850 e-25 kg]
710,064
[29.380 GeV]
Lutetium
71
electron configuration
[Xe] 4f14 5d1 6s2
54
Aufbau
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
71
71 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-6
Lanthanoid
Azmithal & Magnetic numbers
0
2
9
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.71 - Lutetium atom

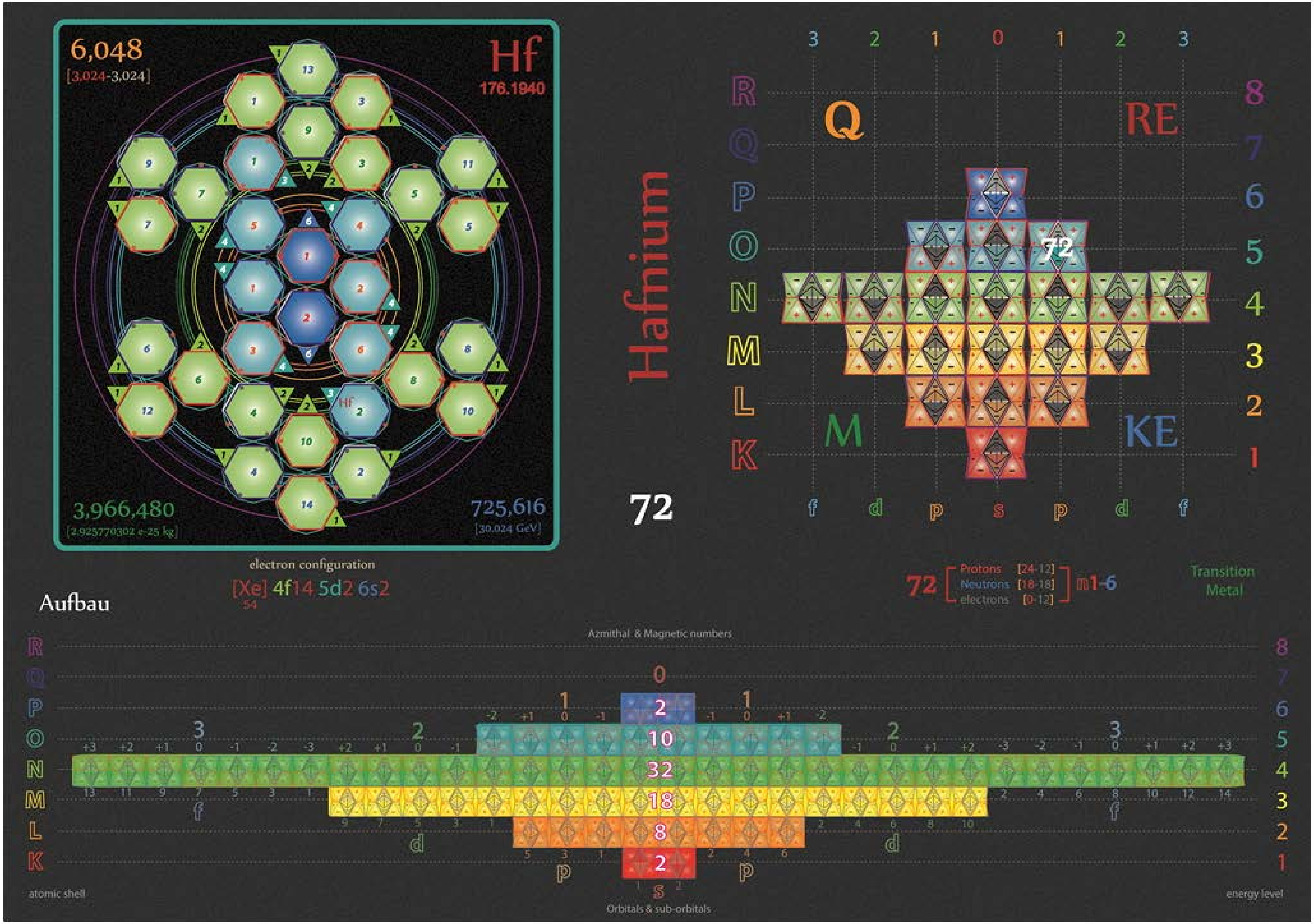
6,048
[3,024-3,024]
Hf
176.1940
3,966,480
[2.925770302 e-25 kg]
725,616
[30.024 GeV]
Hafnium
72
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
72
f d p s p d f
electron configuration
[Xe] 4f14 5d2 6s2
54
72
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Transition Metal
Aufbau
Azmithal & Magnetic numbers
0
1
1
2
2
3
3
2
10
32
18
8
2
f
d
p
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.72 - Hafnium atom

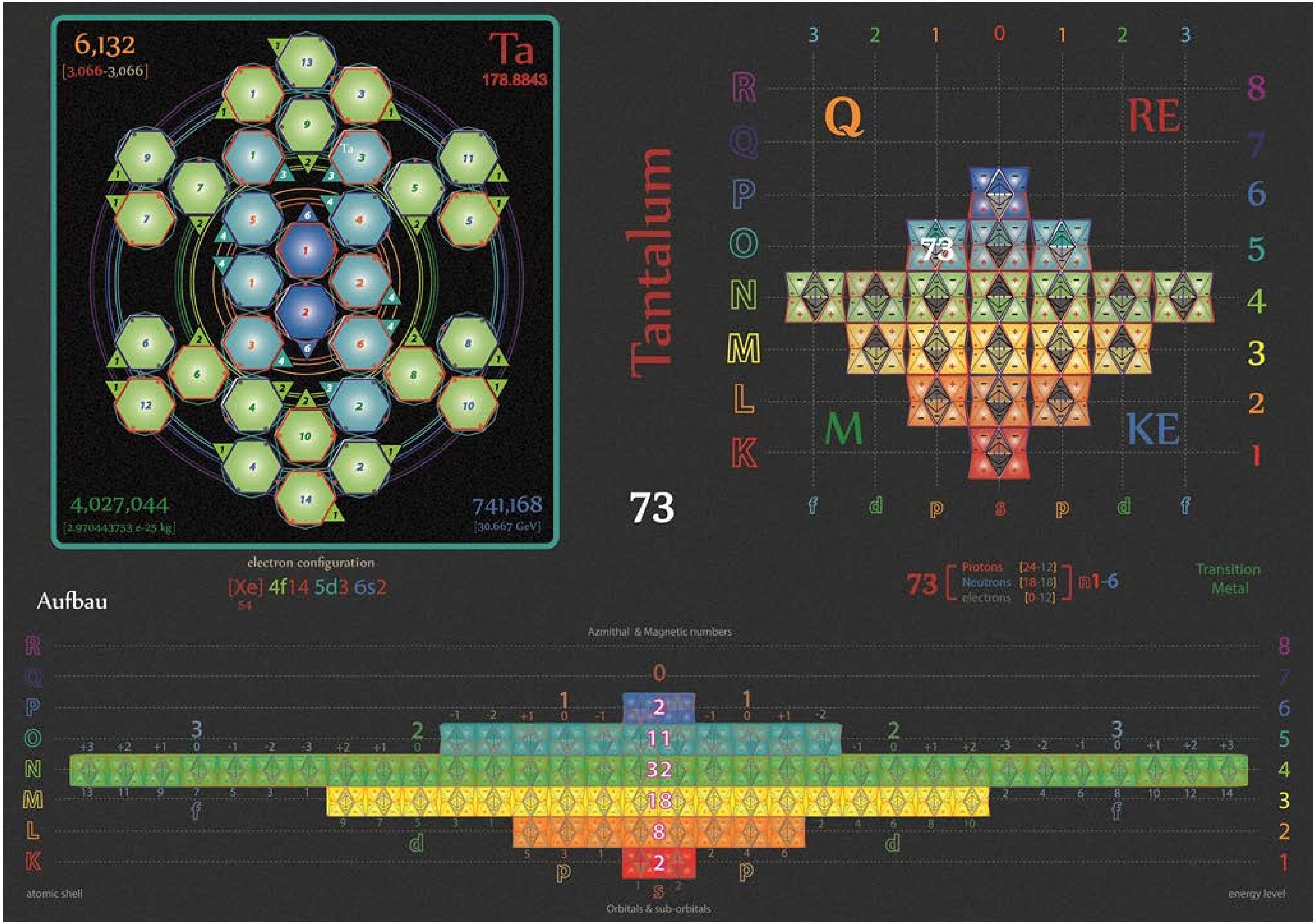
6,132
[3,066-3,066]
Ta
178.8843
4,027,044
[2.970443753 e-25 kg]
741,168
[30.667 GeV]
Tantalum
73
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
Q
RE
M
KE
f
d
p
s
p
d
f
electron configuration
[Xe] 4f14 5d3 6s2
54
73 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Aufbau
Azmithal & Magnetic numbers
2
11
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

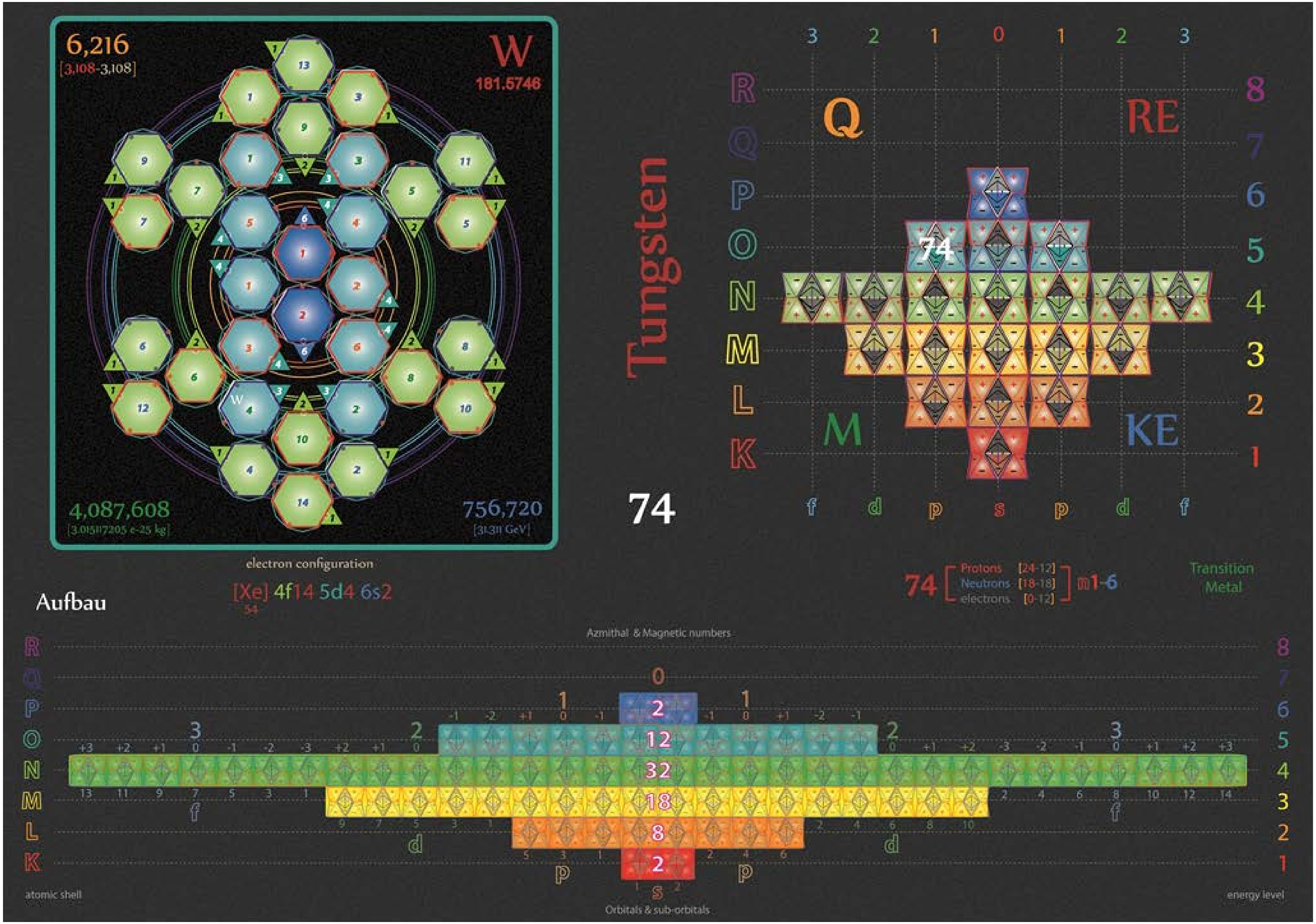
6,216
[3,108-3,108]
W
181.5746
4,087,608
[3.015117205 e-25 kg]
756,720
[31.311 GeV]
Tungsten
74
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 5d4 6s2
54
74 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Aufbau
Azmithal & Magnetic numbers
2
12
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.74 - Tungsten atom

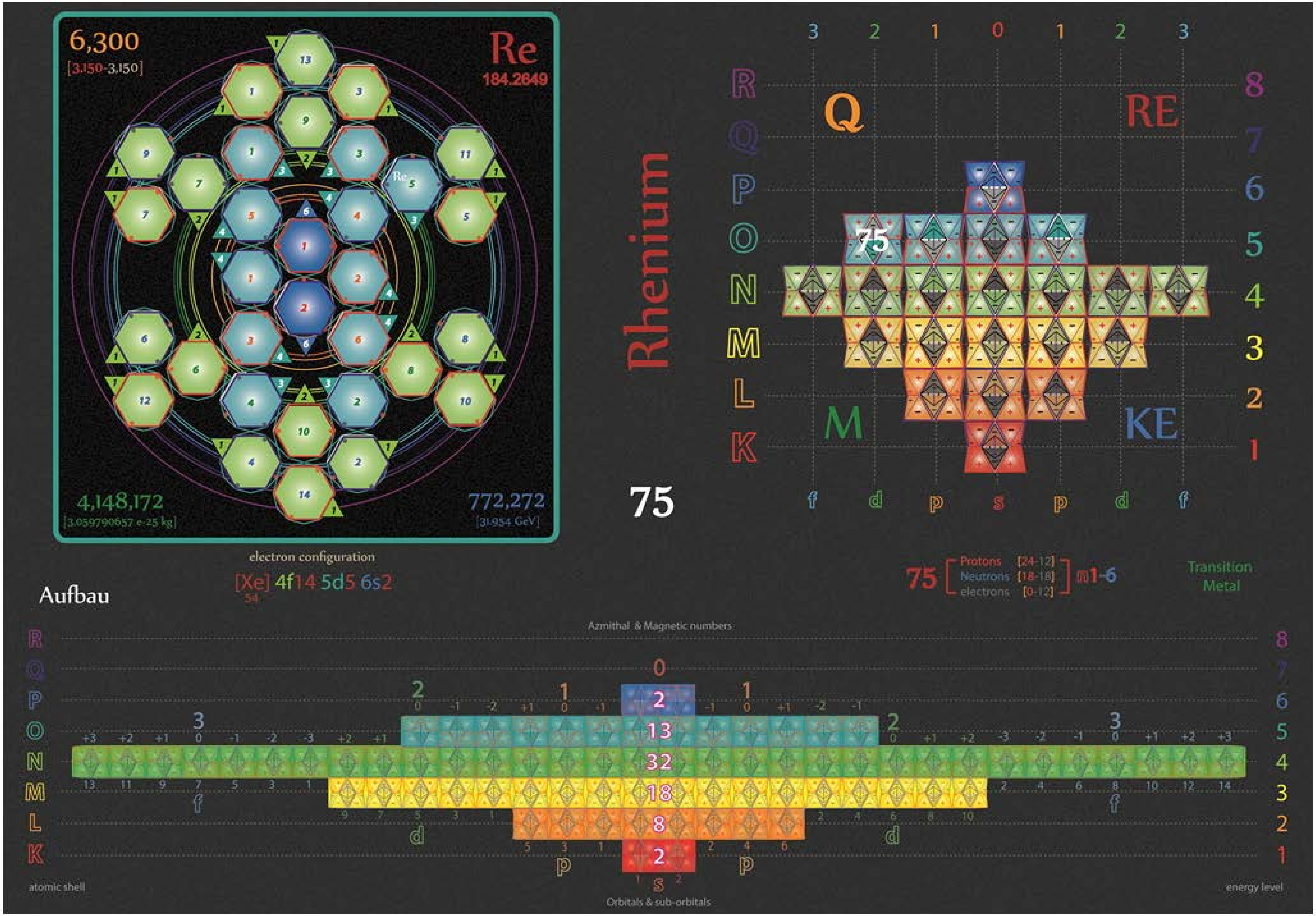
6,300
[3.150-3.150]
Re
184.2849
4,148,172
[3.059790657 e-25 kg]
772,272
[31.954 GeV]
Rhenium
75
electron configuration
[Xe] 4f14 5d5 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
75 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Azmithal & Magnetic numbers
0
2
13
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.75 - Rhenium atom

Copyright ABRAHAM [2008] - All rights reserved

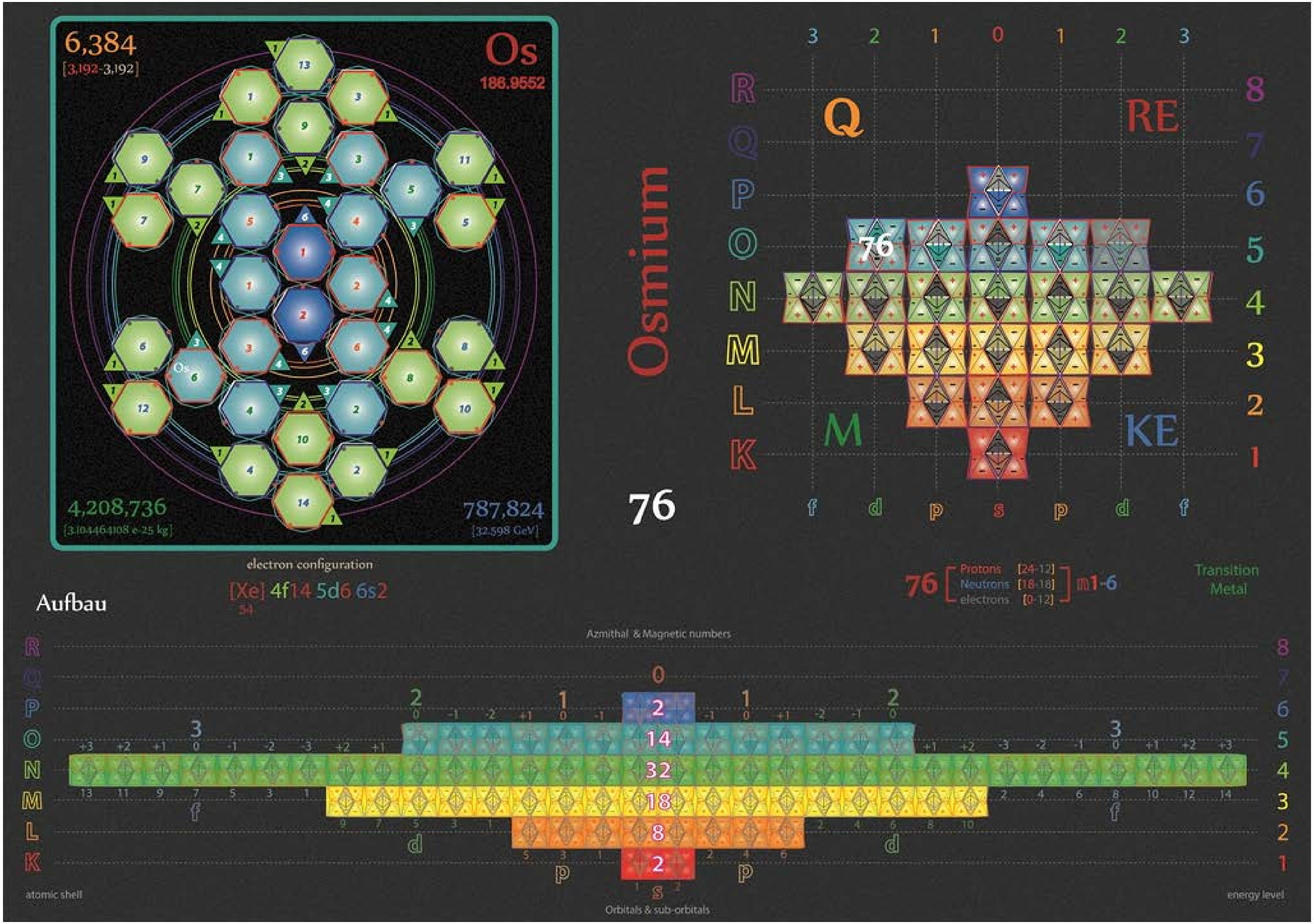
6,384
[3,192-3,192]
Os
186.9552
4,208,736
[3.104464108 e-25 kg]
787,824
[32.598 GeV]
Os
electron configuration
[Xe] 4f14 5d6 6s2
54
Osmium
76
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
76
76 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Transition Metal
Aufbau
Azmithal & Magnetic numbers
0
2 14 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.76 - Osmium atom

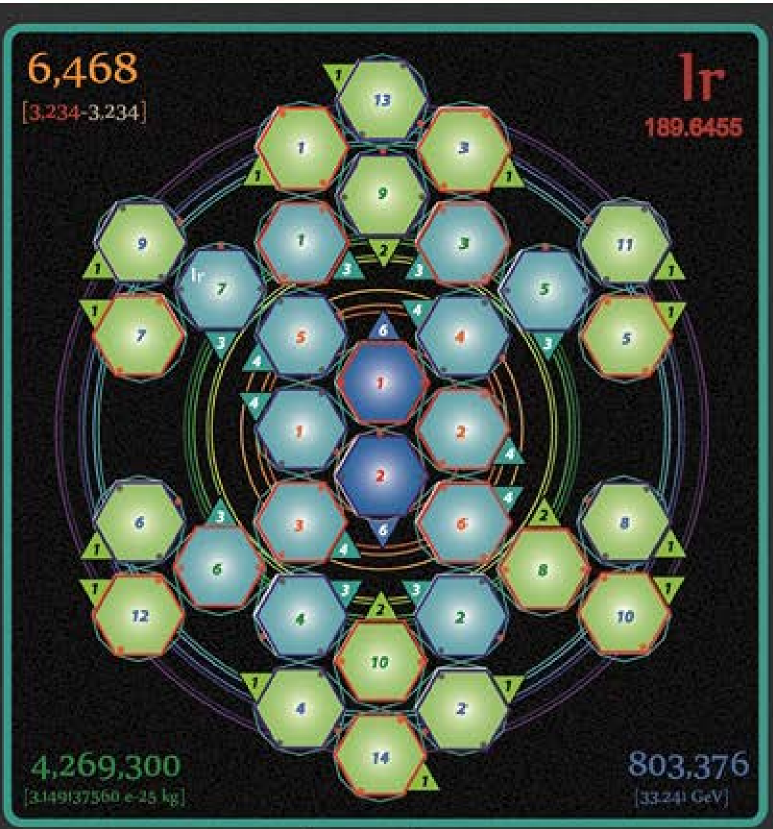
6,468
[3.234-3.234]
Ir
189.6455
4,269,300
[3.149137560 e-25 kg]
803,376
[33.241 GeV]

Iridium

77

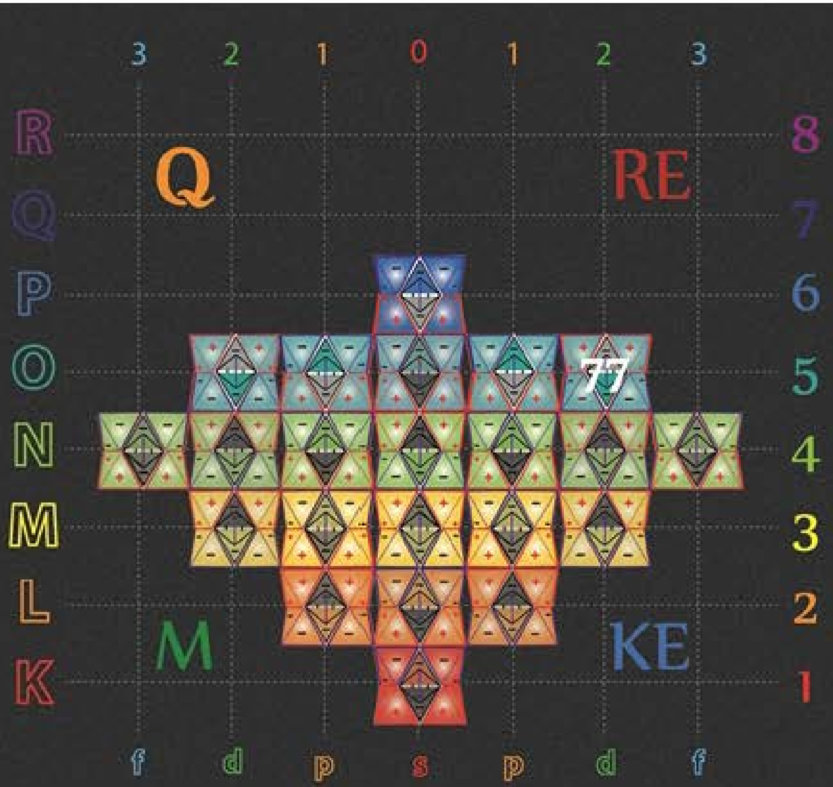
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
77
f d p s p d f

electron configuration

[Xe] 4f14 5d7 6s2

54

Aufbau

77 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Transition Metal

Azmithal & Magnetic numbers

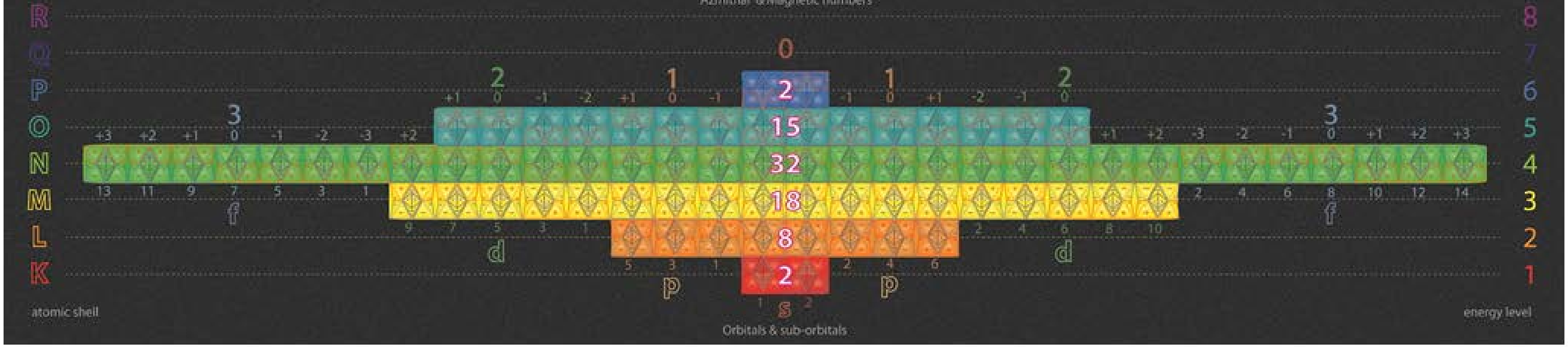
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
15
32
18
8
2
1 2 3
f d p s p d f
atomic shell
energy level

Orbitals & sub-orbitals

Tetryonics 51.77 - Iridium atom

Copyright ABRAHAM [2008] - All rights reserved

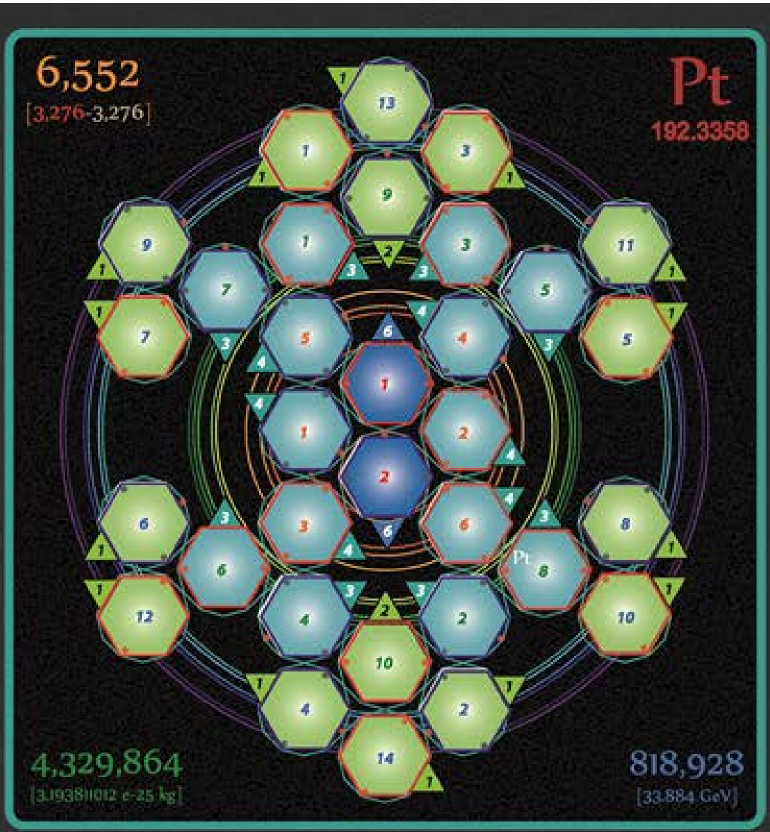
6,552
[3,276-3,276]

Pt
192.3358

4,329,864
[3.19381012 e-25 kg]

818,928
[33.884 GeV]

Platinum

78

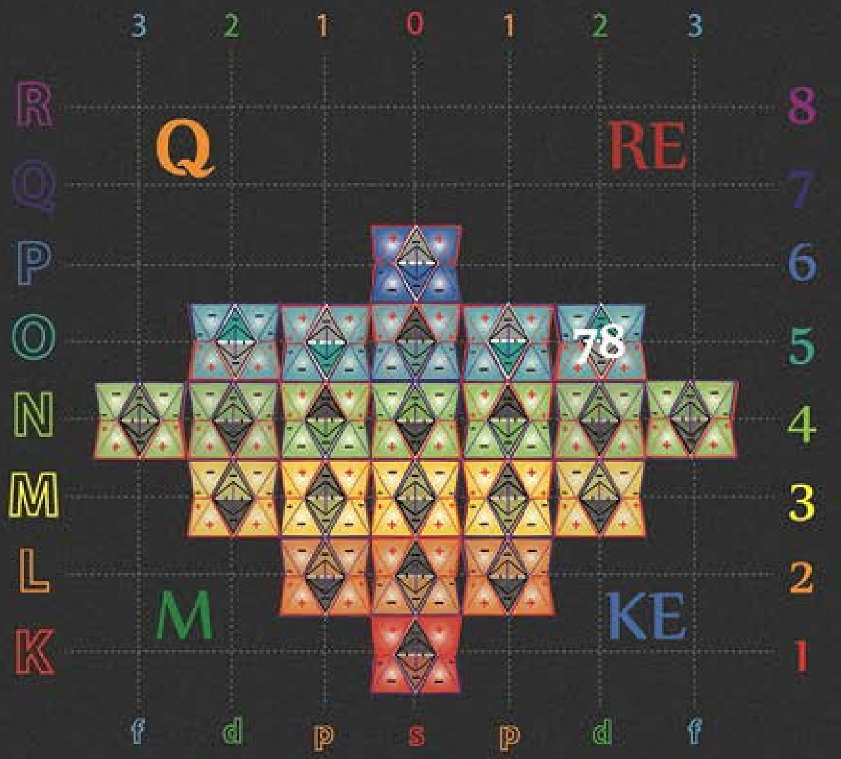
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
78
f d p s p d f

Aufbau

electron configuration
[Xe] 4f14 5d8 6s2
54

78 [Potons [24-12] Neutrons [18-18] eectrons [0-12]] n1-6

Transition Metal

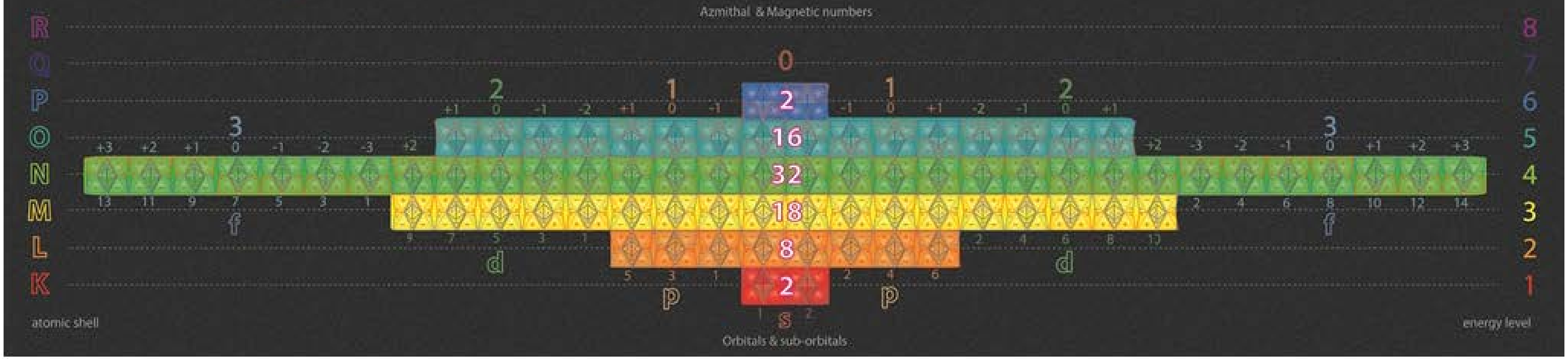
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
16
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.78 - Platinum atom

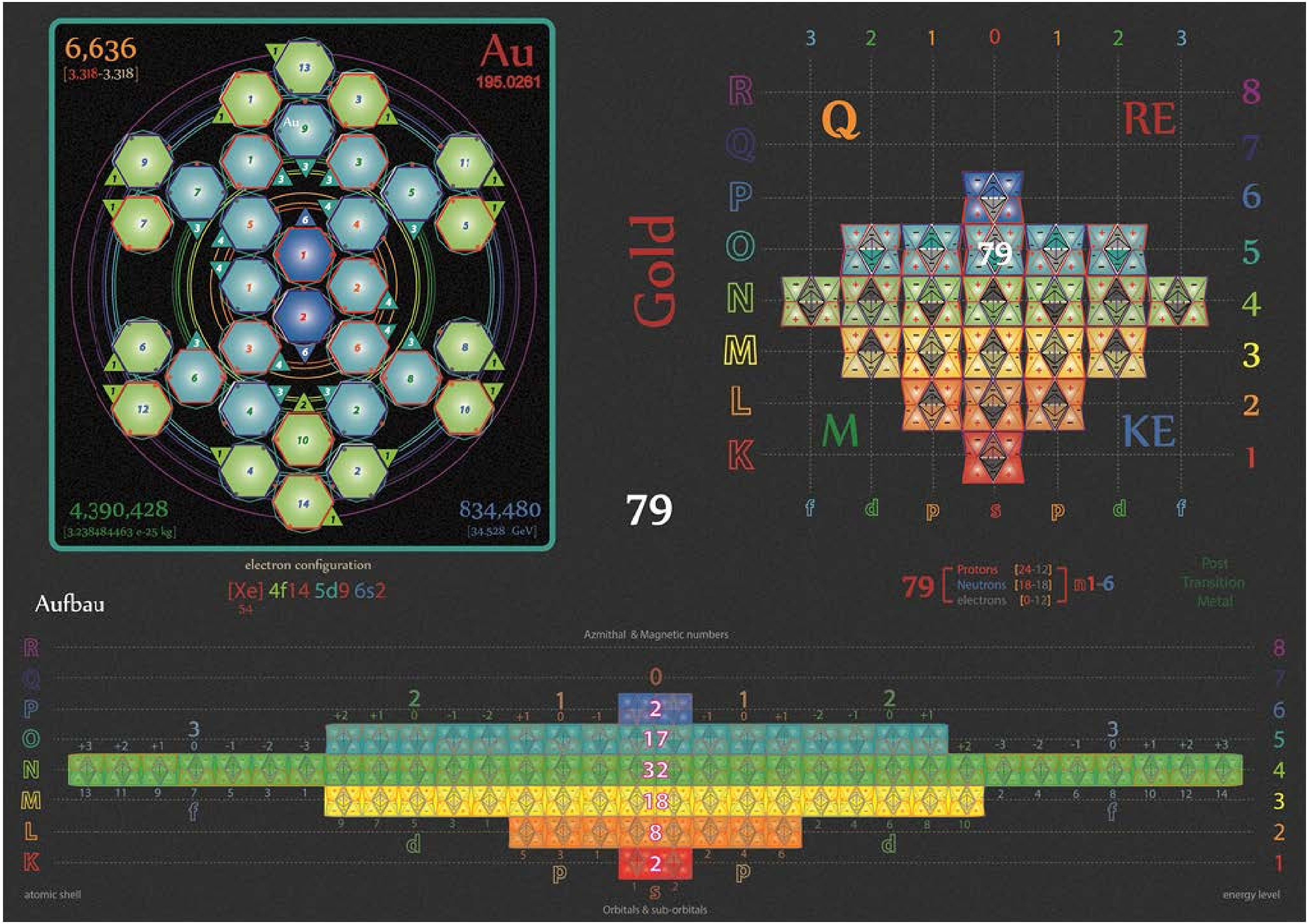
6,636
[3.318-3.318]
Au
195.0261
Au
4,390,428
[3.238484463 e-25 kg]
834,480
[34.528 GeV]
Gold
79
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
79
f d p s p d f
electron configuration
[Xe] 4f14 5d9 6s2
54
Aufbau
79
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Post Transition Metal
Azmithal & Magnetic numbers
0
2
17
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.79 - Gold atom

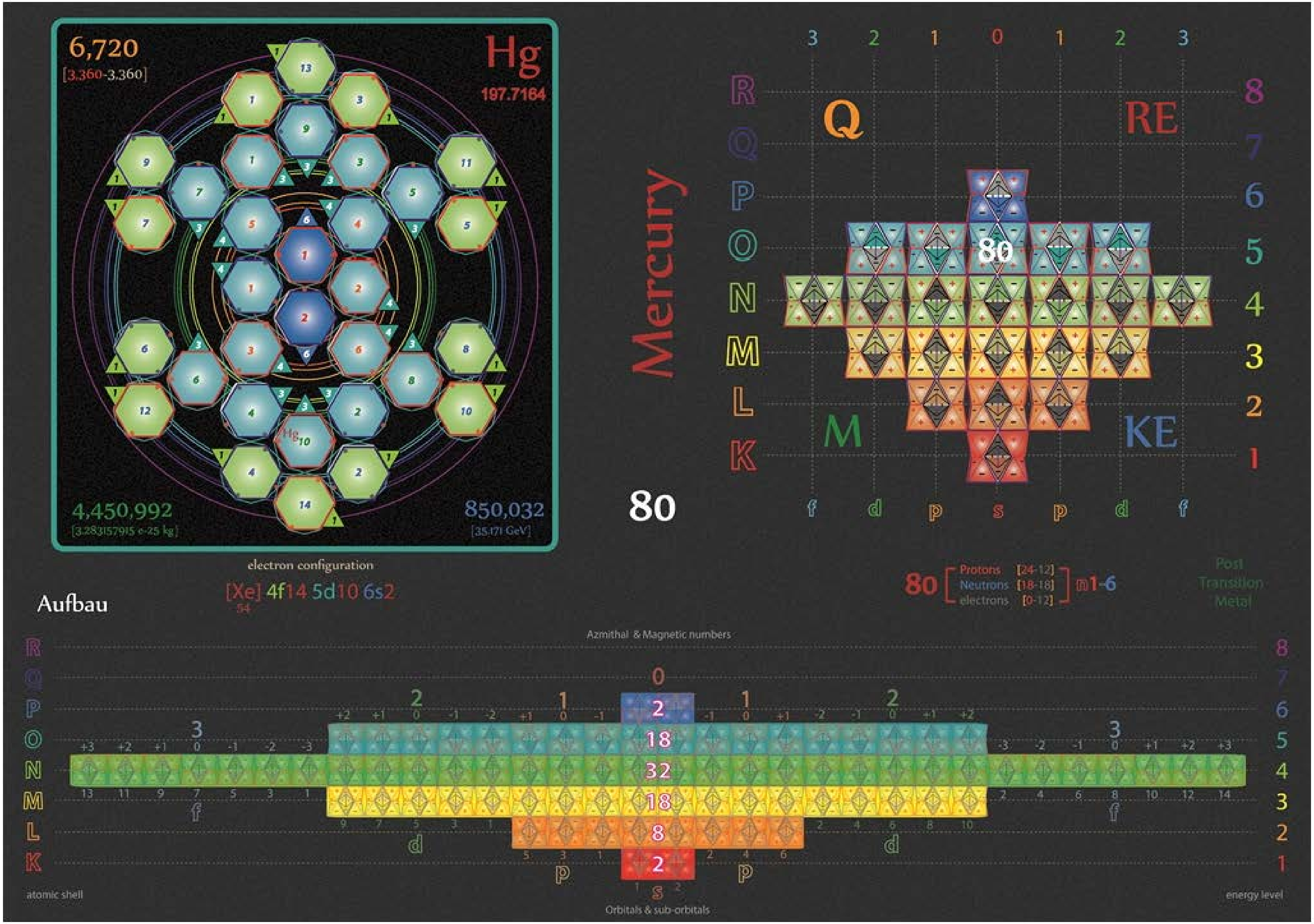
6,720
[3,360-3,360]
Hg
197.7164
4,450,992
[3.283157915 e-25 kg]
850,032
[35.171 GeV]
Mercury
80
electron configuration
[Xe] 4f14 5d10 6s2
54
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
80
f d p s p d f
80 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-6
Post Transition Metal
Azmithal & Magnetic numbers
2
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.80 - Mercury atom

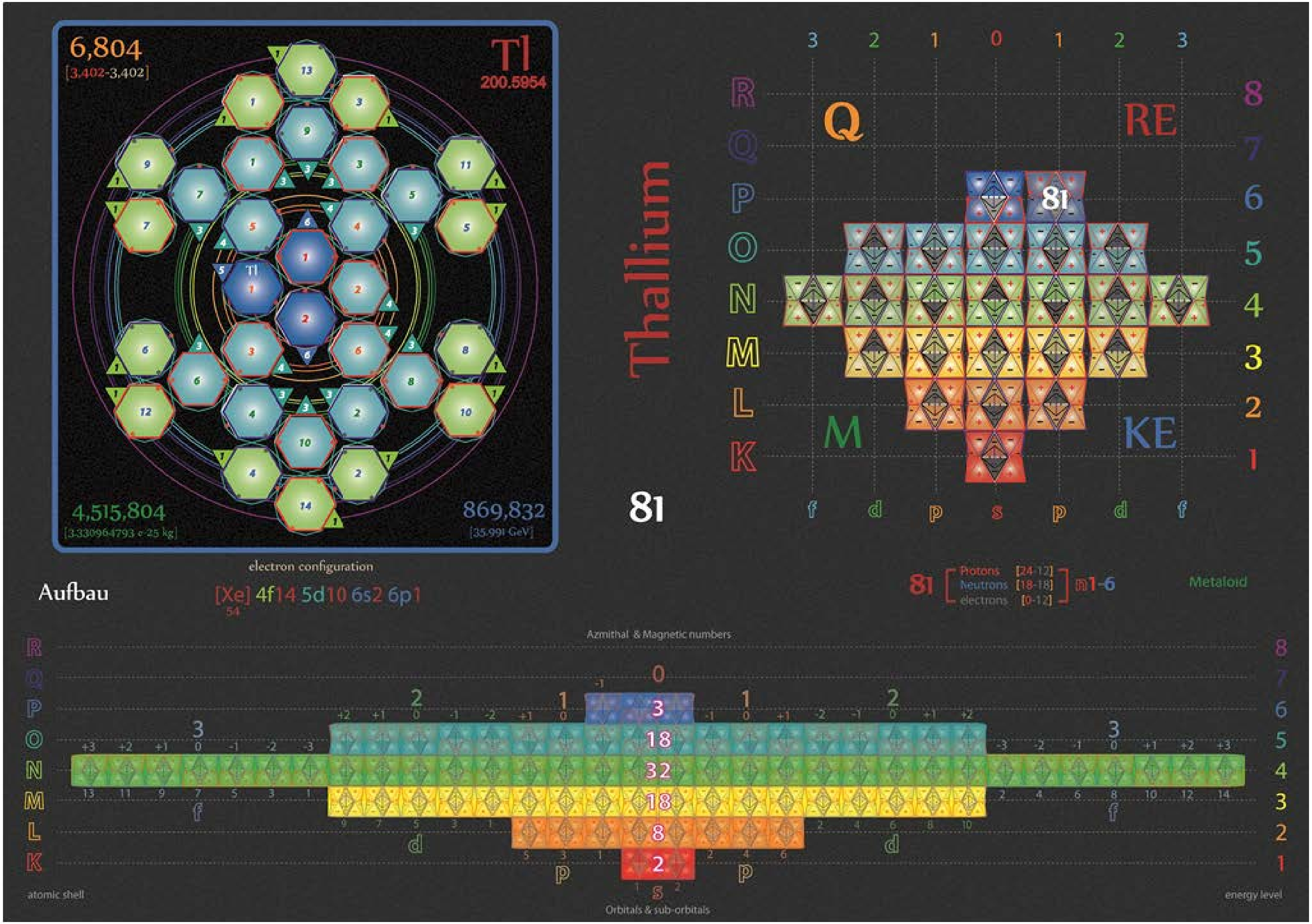
6,804
[3,402-3,402]
Tl
200.5954
4,515,804
[3.330964793 e-25 kg]
869,832
[35.991 GeV]
Thallium
81
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p1
54
81 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-6
Metaloid
Azmithal & Magnetic numbers
3
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.81 - Thallium atom

Copyright ABRAHAM [2008] - All rights reserved

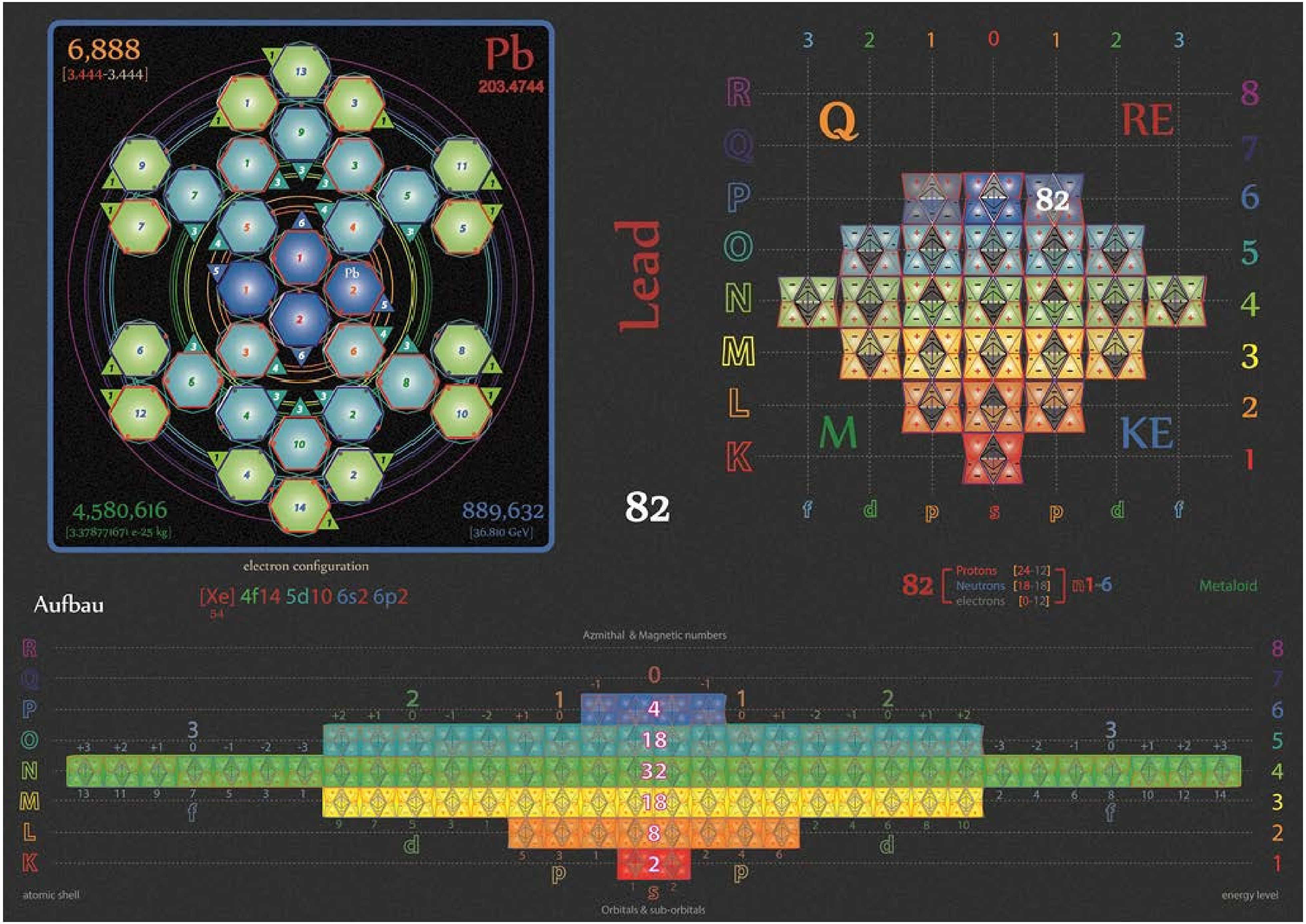
6,888
[3.444-3.444]
Pb
203.4744
4,580,616
[3.378771671 e-25 kg]
889,632
[36.810 GeV]
Lead
82
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p2
54
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
82 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.82 - Lead atom

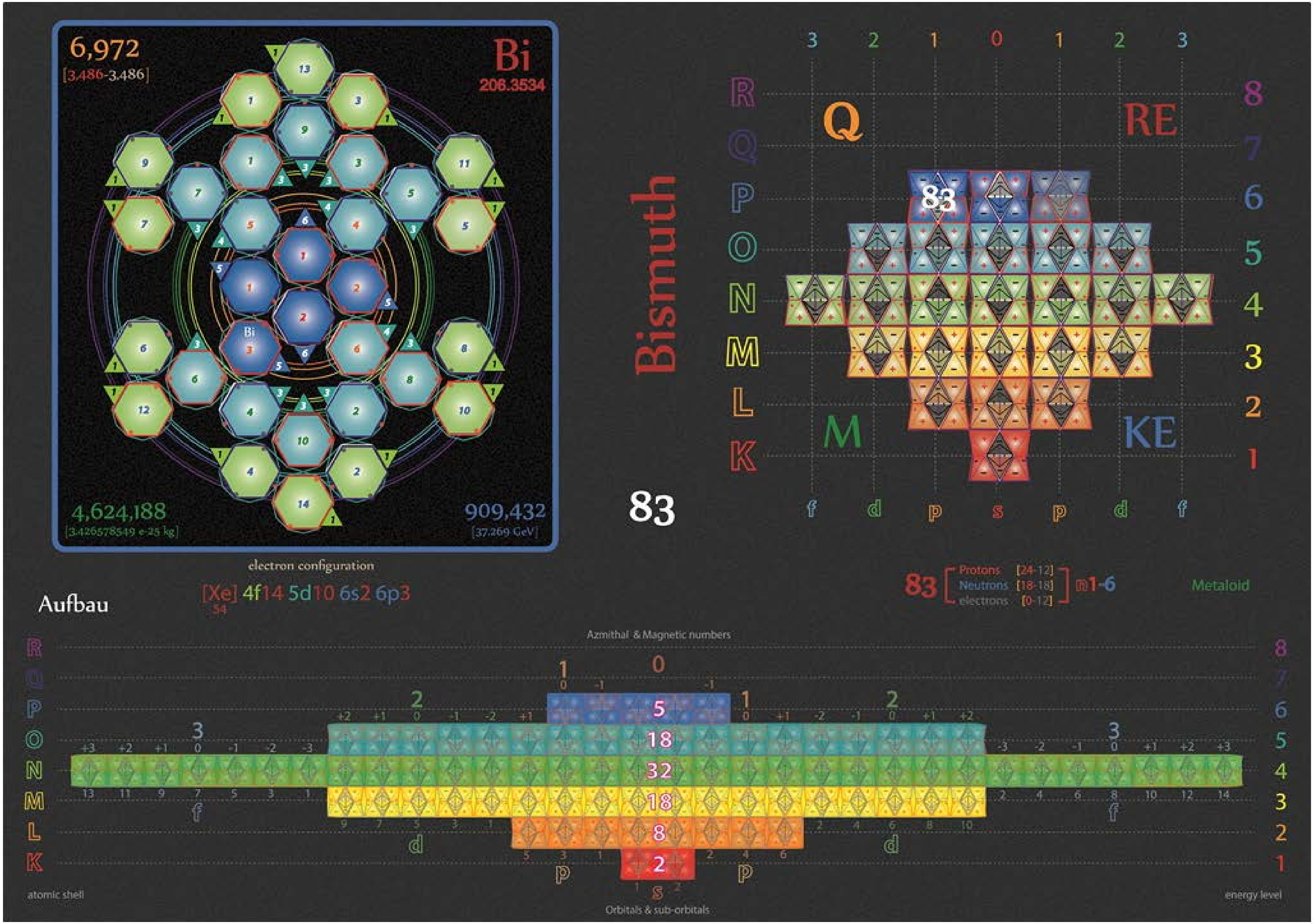
6,972
[3,486-3,486]
Bi
208.3534
4,624,188
[3.426578549 e-25 kg]
909,432
[37.269 GeV]
Bismuth
83
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Xe] 4f14 5d10 6s2 6p3
54
Aufbau
83 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-6
Metaloid
Azmithal & Magnetic numbers
5
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.83 - Bismuth atom

Copyright ABRAHAM [2008] - All rights reserved

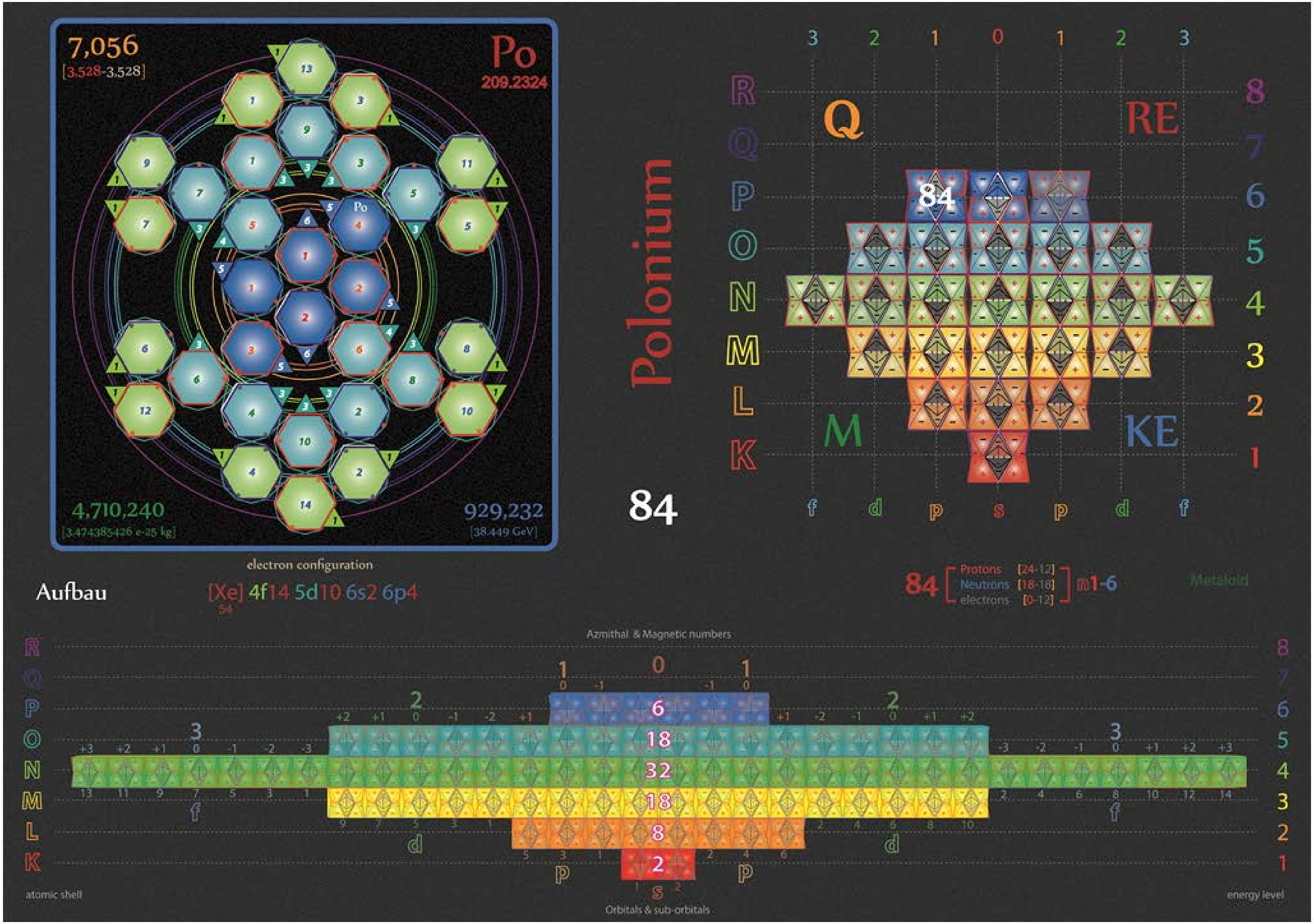

7,056
[3.528-3.528]
Po
209.2324
4,710,240
[3.474385426 e-25 kg]
929,232
[38.449 GeV]
Polonium
84
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p4
54
84
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-6
Metaloid
Azmithal & Magnetic numbers
6
18
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.84 - Polonium atom

Copyright ABRAHAM [2008] - All rights reserved

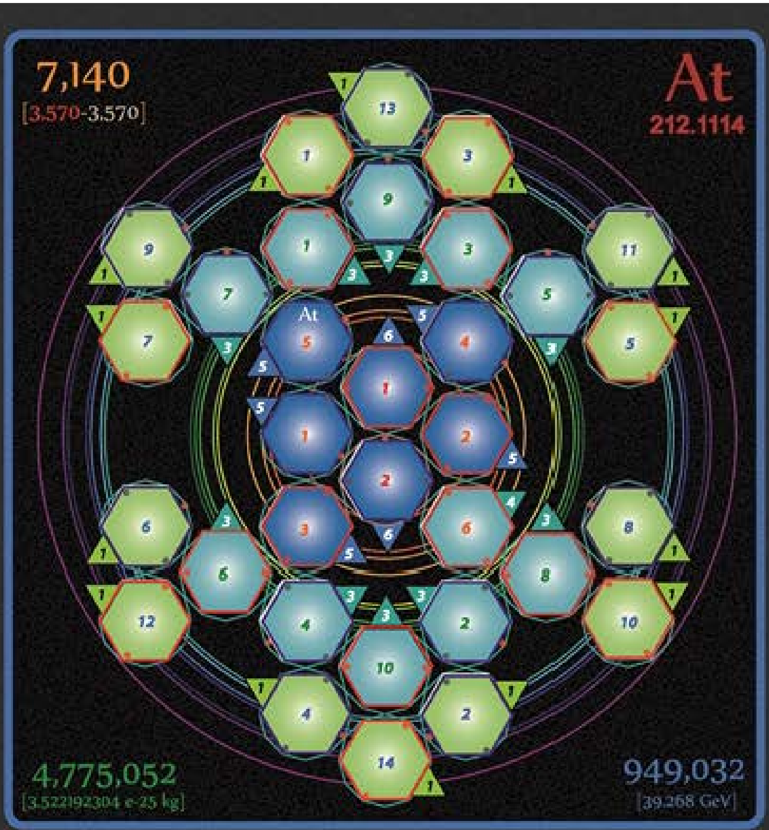
7,140
[3.570-3.570]
At
212.1114
4,775,052
[3.522192304 e-25 kg]
949,032
[39.268 GeV]

Astatine

85

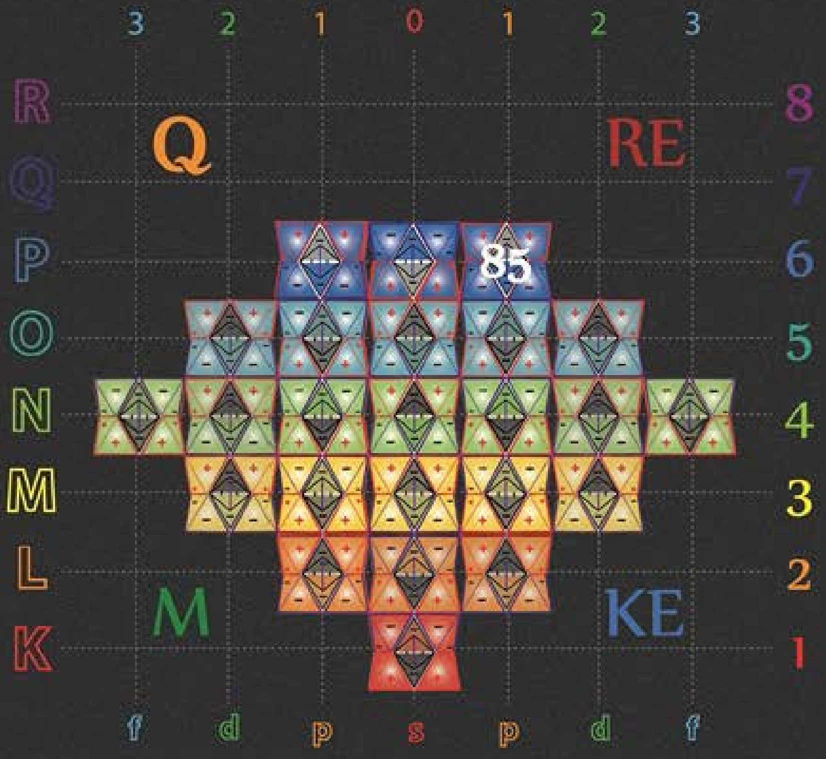
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
85
8
7
6
5
4
3
2
1
f d p s p d f

electron configuration

Aufbau

[Xe] 4f14 5d10 6s2 6p5
54

85 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6

Halogen

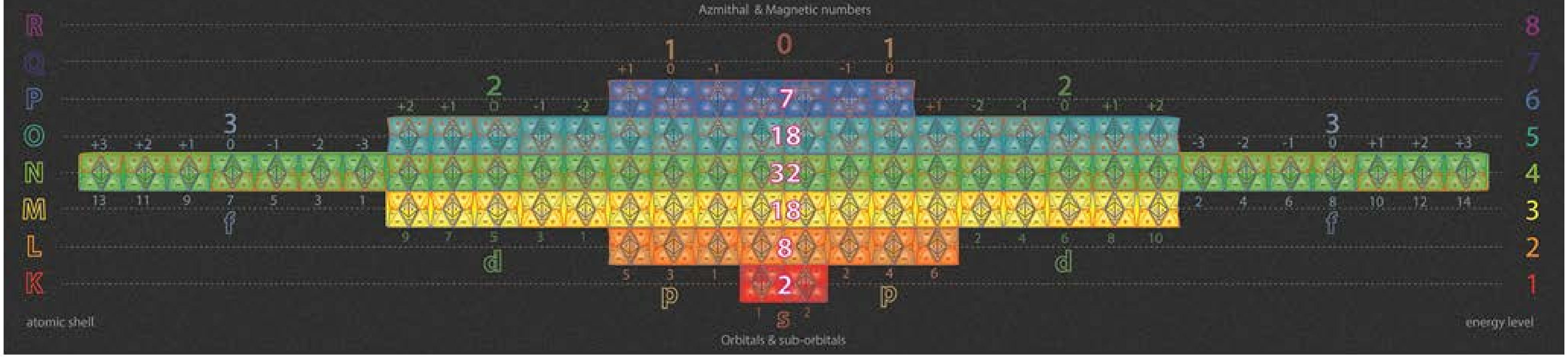
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
7
18
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.85 - Astatine atom

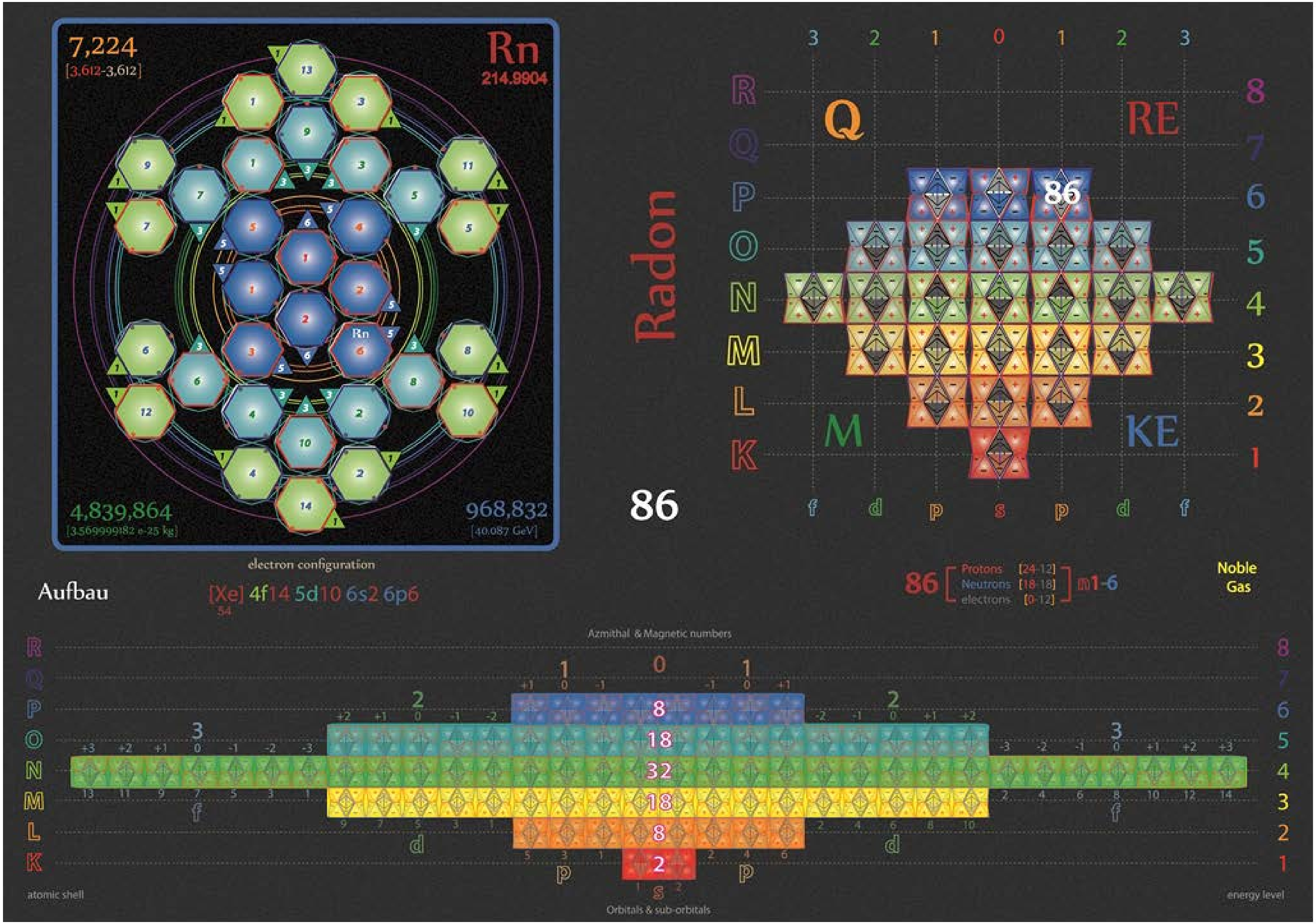
7,224
[3,612-3,612]
Rn
214.9904
4,839,864
[3.569999182 e-25 kg]
968,832
[40.087 GeV]
Radon
86
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Xe] 4f14 5d10 6s2 6p6
54
86 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-6
Noble Gas
Azmithal & Magnetic numbers
8
18
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.86 - Radon atom

Copyright ABRAHAM [2008] - All rights reserved

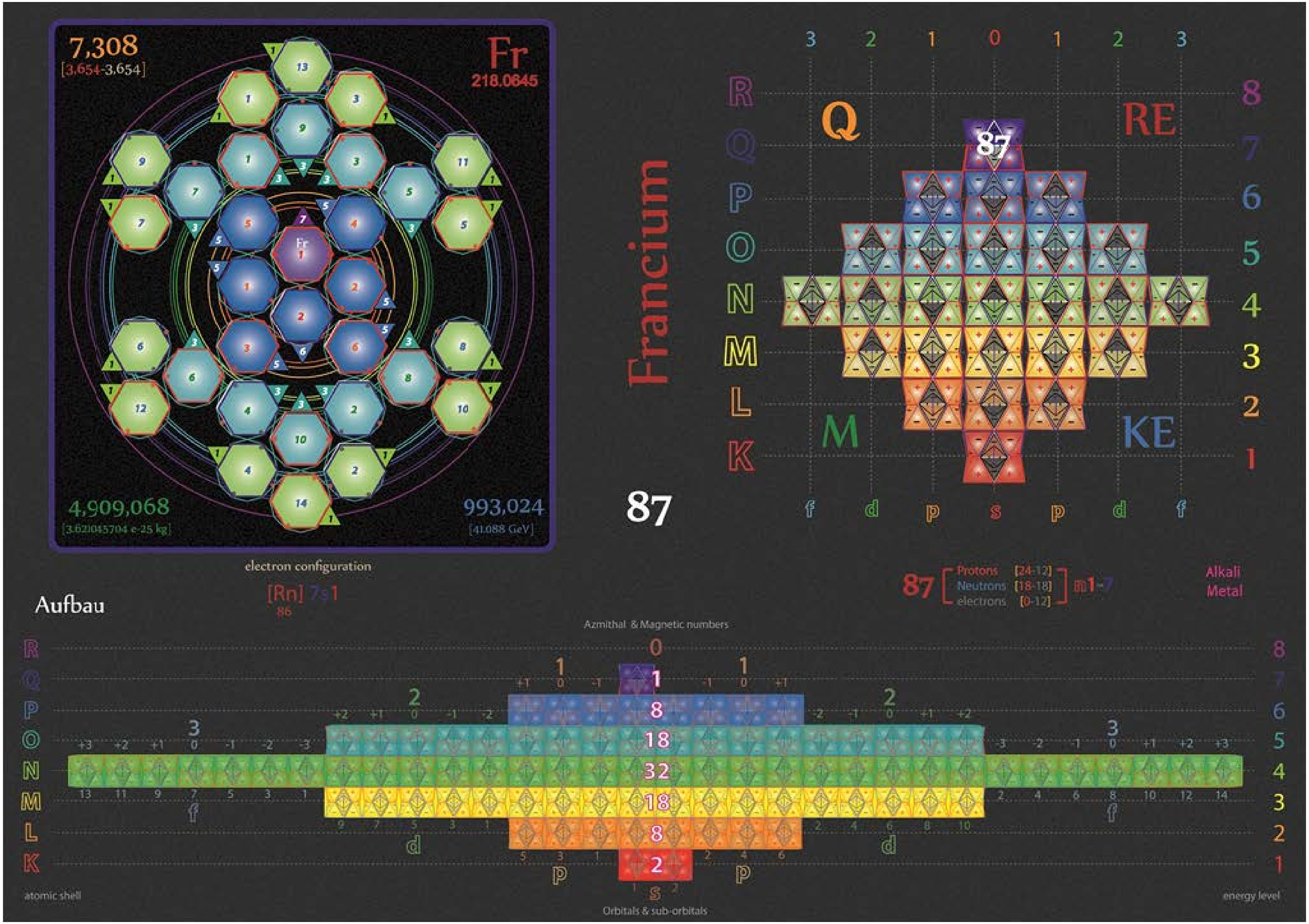
7,308
[3,654-3,654]
Fr
218.0845
4,909,068
[3.62045704 e-25 kg]
993,024
[41.088 GeV]
Francium
87
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
[Rn] 7s1
86
Aufbau
87
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Alkali Metal
Azmithal & Magnetic numbers
0
1
2
3
8
18
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.87 - Francium atom

Copyright ABRAHAM [2008] - All rights reserved

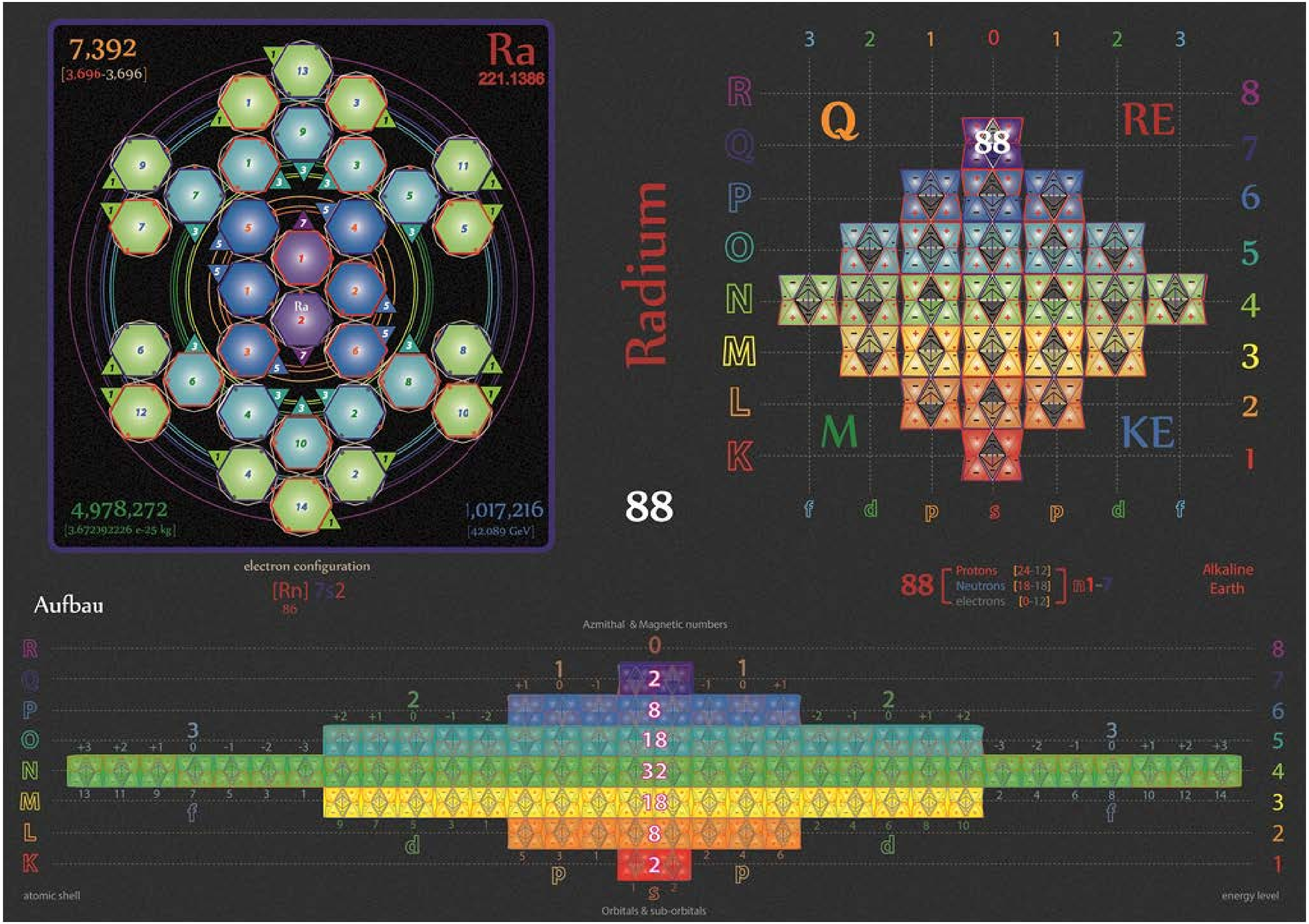
7,392
[3,696-3,696]
Ra
221.1386
4,978,272
[3.672392226 e-25 kg]
1,017,216
[42.089 GeV]
Radium
88
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
[Rn] 7s2
86
88 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Alkaline Earth
Aufbau
Azmithal & Magnetic numbers
2 8 18 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.88 - Radium atom

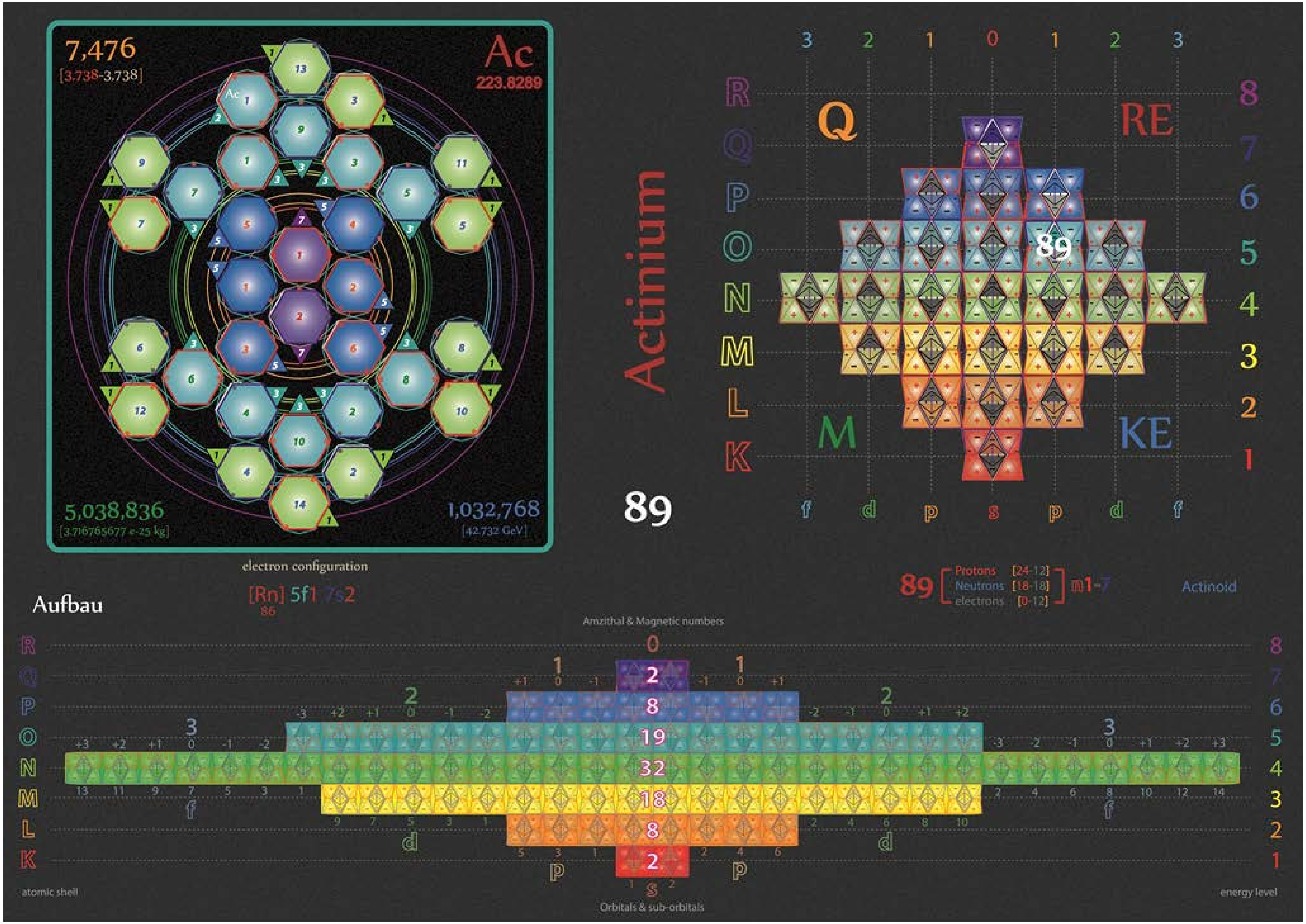
7,476
[3.738-3.738]
Ac
223.8289
5,038,836
[3.716765677 e-25 kg]
1,032,768
[42.732 GeV]
Actinium
89
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
[Rn] 5f1 7s2
86
89
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Aufbau
Amzithal & Magnetic numbers
2
8
19
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.89 - Actinium atom

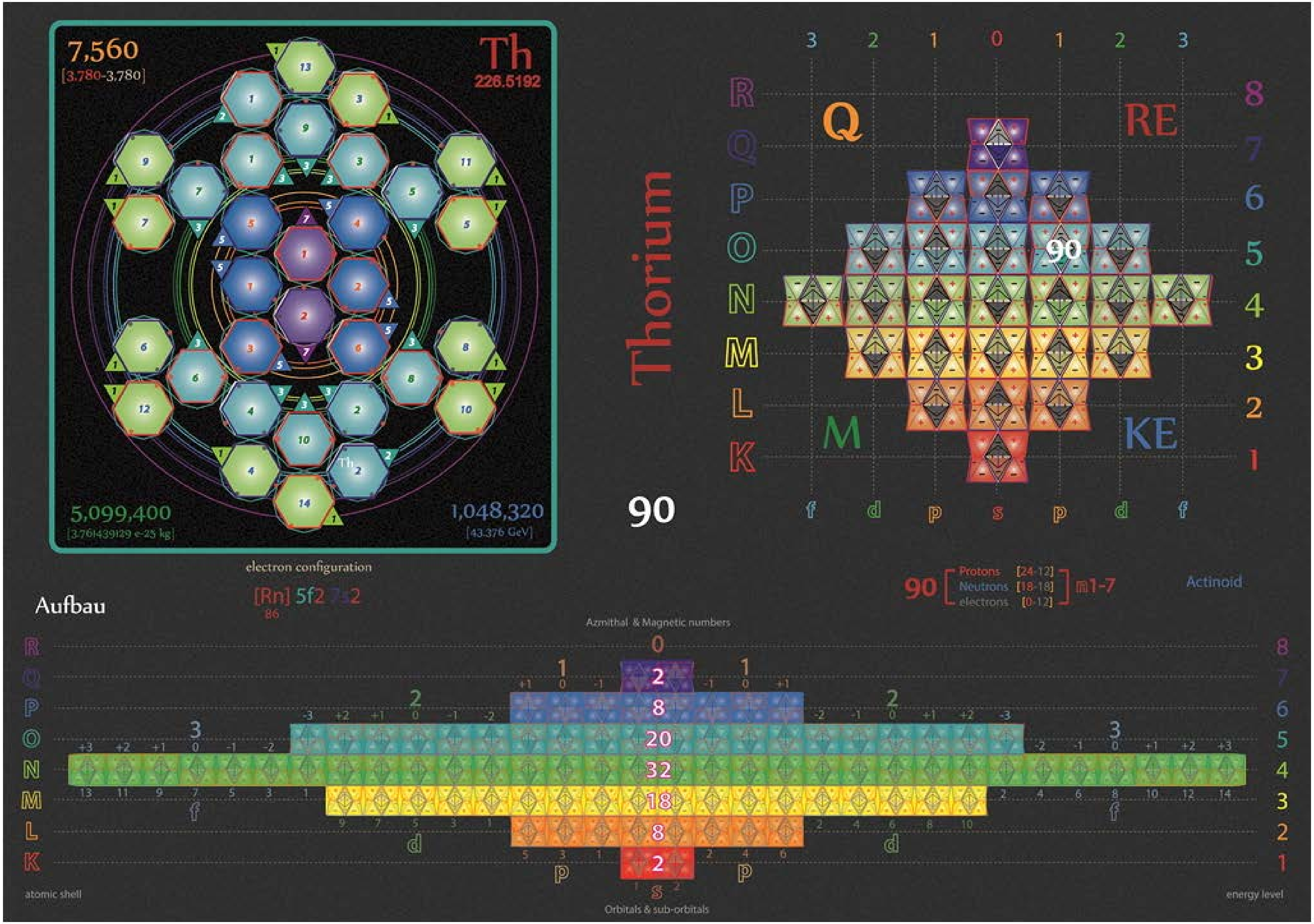
7,560
[3,780-3,780]
Th
226.5192
5,099,400
[3.761439129 e-25 kg]
1,048,320
[43.376 GeV]
Thorium
90
electron configuration
[Rn] 5f2 7s2
86
Aufbau
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
90
f d p s p d f
90 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
0
1 1
2 2
3 3
2
8
20
32
18
8
2
f d p s p d f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.90 - Thorium atom

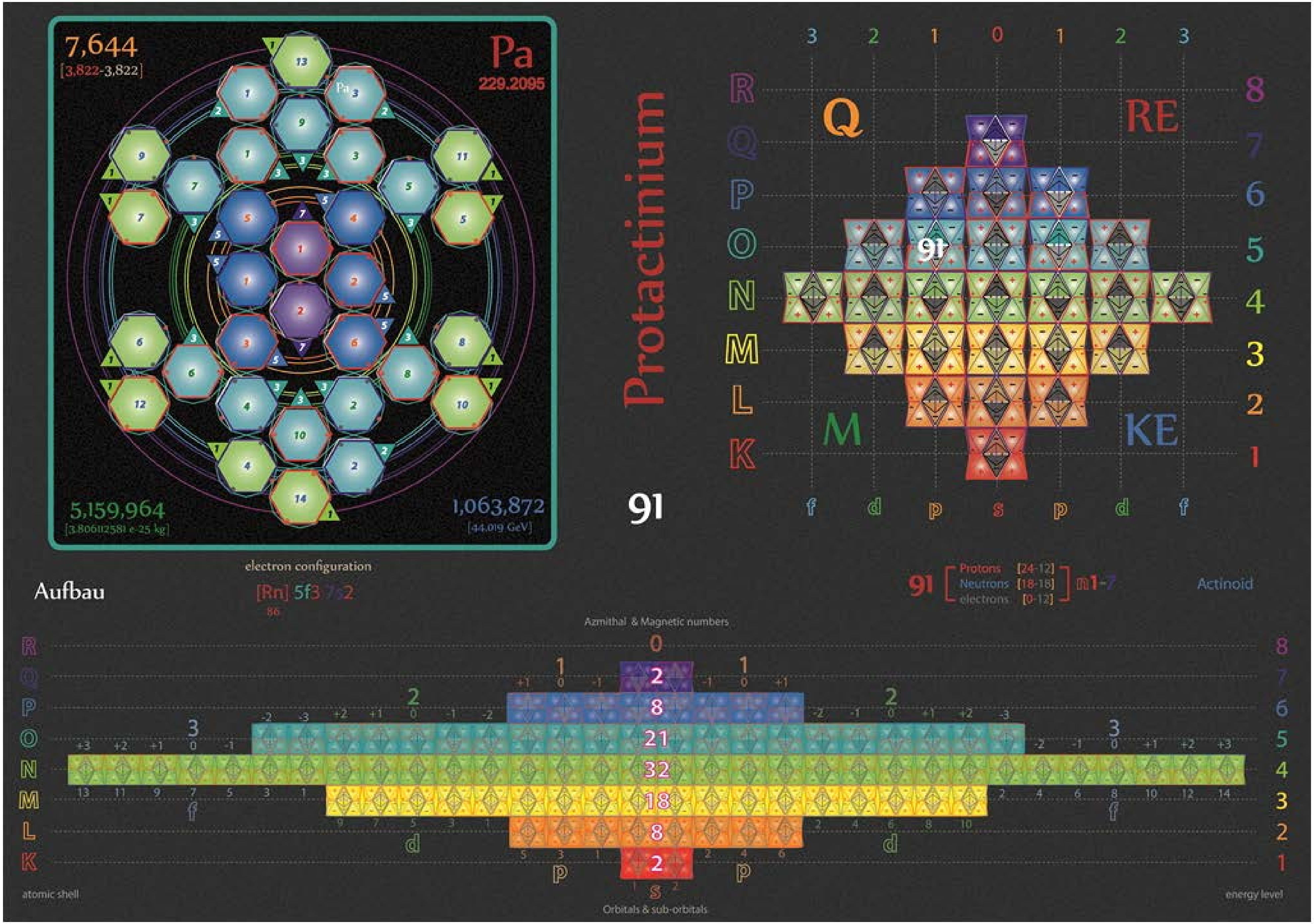
7,644
[3,822-3,822]
Pa
229.2095
5,159,964
[3.80611258l e-25 kg]
1,063,872
[44.019 GeV]
Protactinium
91
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
Aufbau
electron configuration
[Rn] 5f3 7s2
86
91
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
21
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.91 - Protactinium atom

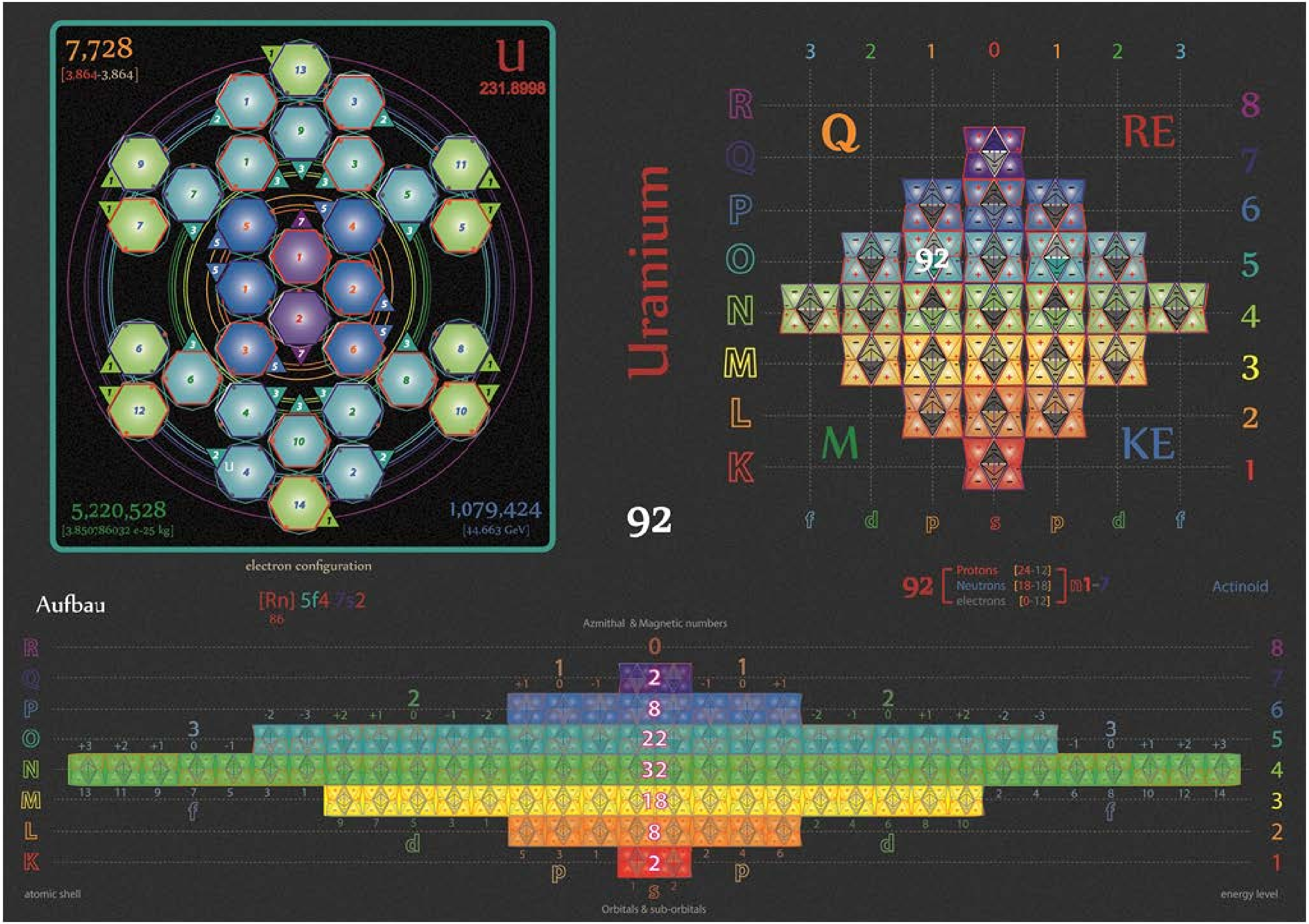
7,728
[3.864-3.864]
U
231.8998
5,220,528
[3.850786032 e-25 kg]
1,079,424
[44.663 GeV]
electron configuration
Uranium
92
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
92 Protons [24-12] Neutrons [18-18] electrons [0-12] m1-7
Actinoid
Aufbau
[Rn] 5f4 7s2
86
Azmithal & Magnetic numbers
0
2
8
22
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.92 - Uranium atom

Copyright ABRAHAM [2008] - All rights reserved

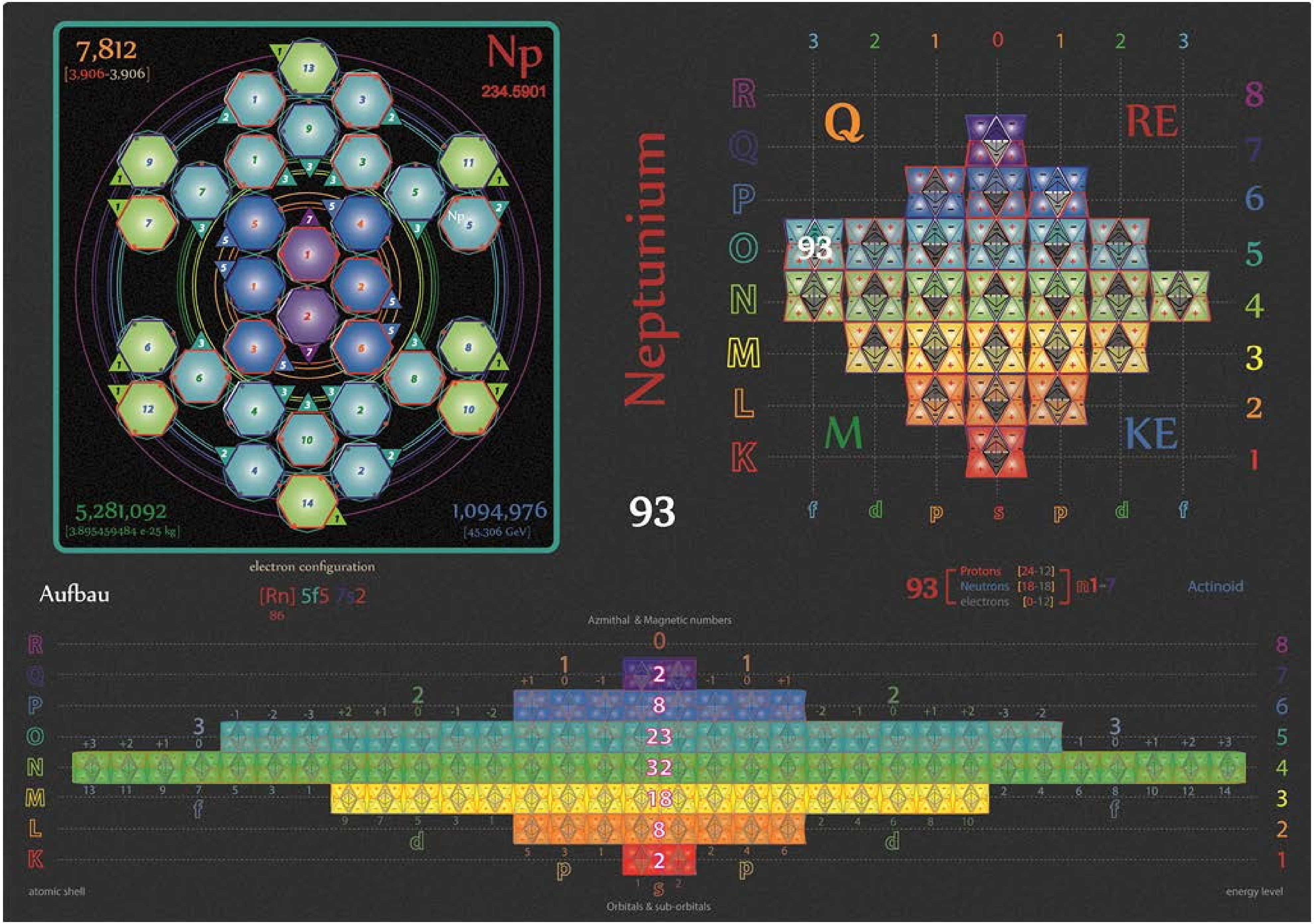
7,812
[3,906-3,906]
Np
234.5901
5,281,092
[3.895459484 e-25 kg]
1,094,976
[45.306 GeV]
Neptunium
93
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
Aufbau
[Rn] 5f5 7s2
86
93 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
23
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.93 - Neptunium atom

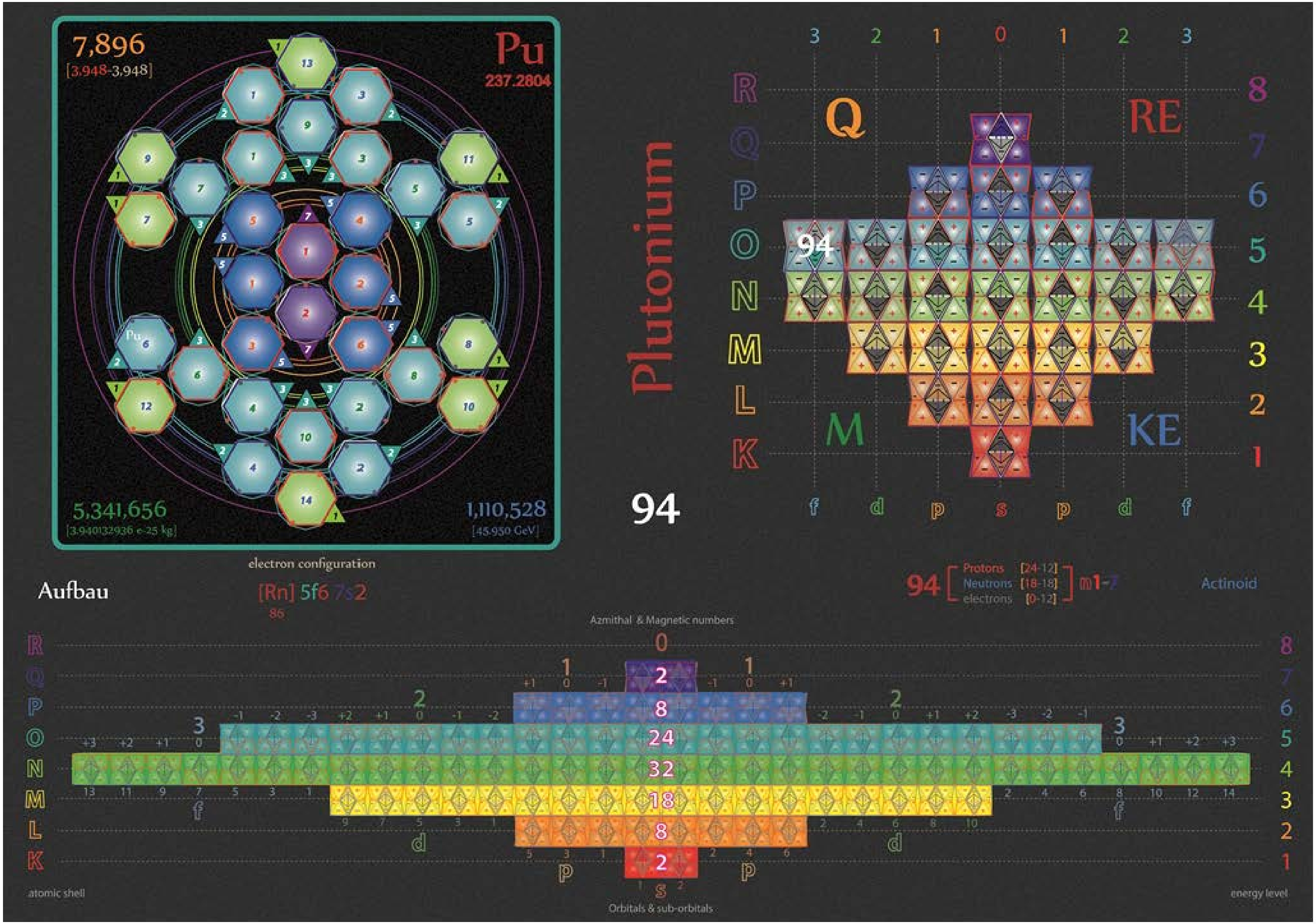
7,896
[3,948-3,948]
Pu
237.2804
5,341,656
[3.940132936 e-25 kg]
1,110,528
[45.950 GeV]
Plutonium
94
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
Aufbau
[Rn] 5f6 7s2
86
94 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
24
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.94 - Plutonium atom

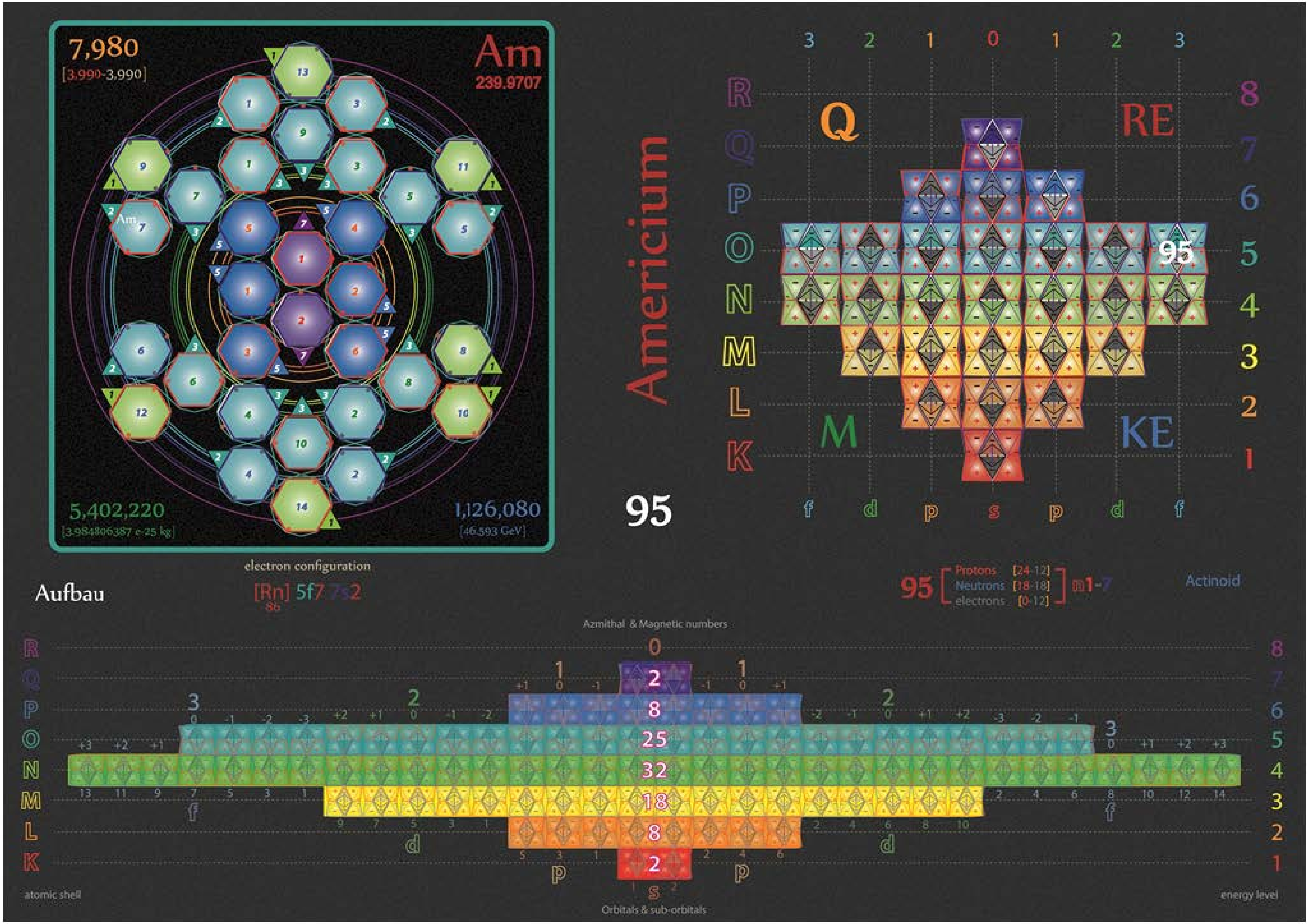
7,980
[3,990-3,990]
Am
239.9707
5,402,220
[3.984806387 e-25 kg]
1,126,080
[46.593 GeV]
Americium
95
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f7 7s2
86
95 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Azmithal & Magnetic numbers
2
8
25
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.95 - Americium atom

Copyright ABRAHAM [2008] - All rights reserved

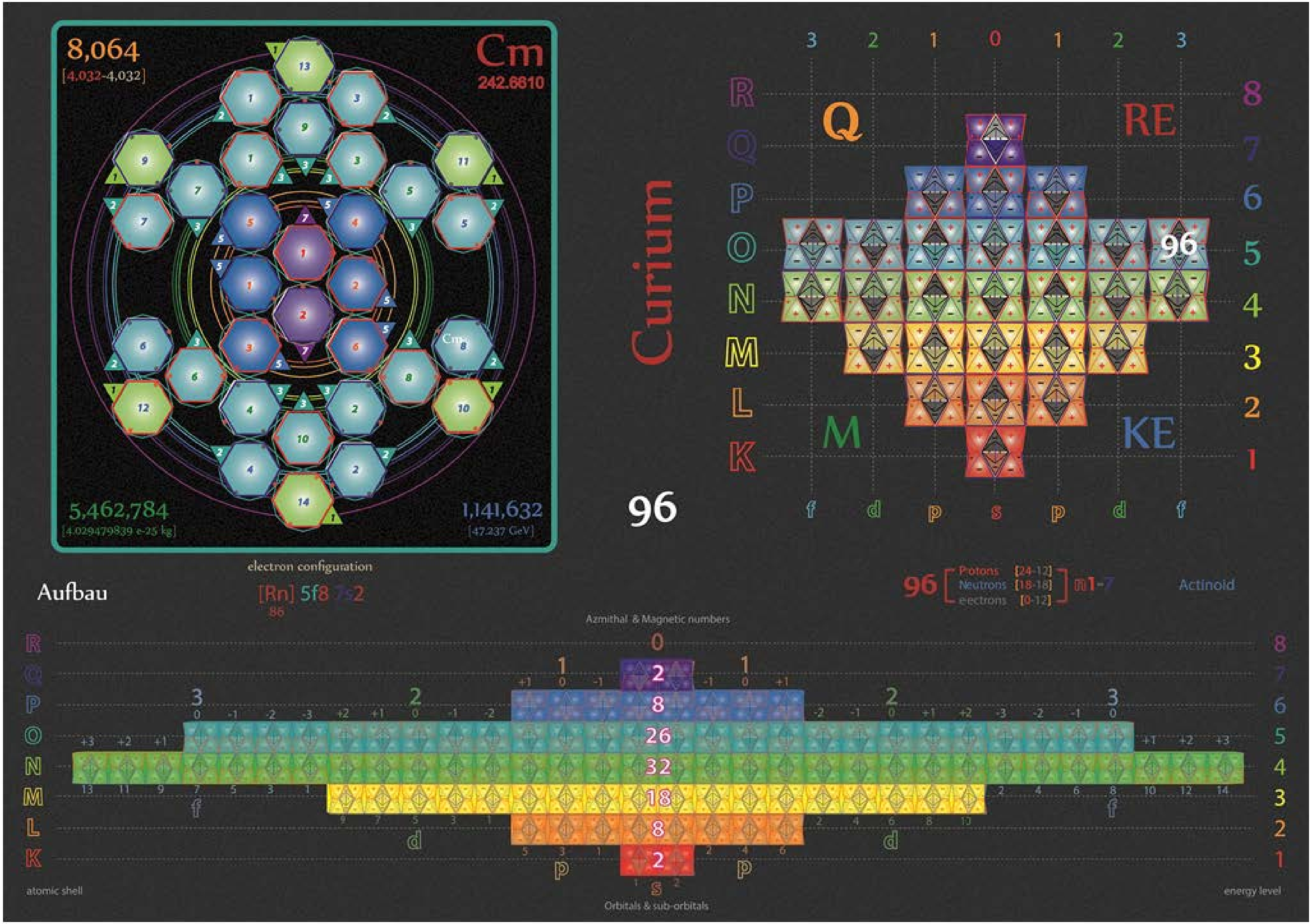
8,064
[4,032-4,032]
Cm
242.6810
5,462,784
[4.029479839 e-25 kg]
1,141,632
[47.237 GeV]
Curium
96
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
96
electron configuration
[Rn] 5f8 7s2
86
96 [Potons [24-12] Neutrons [18-18] eectrons [0-12]] n1-7
Actinoid
Aufbau
Azmithal & Magnetic numbers
0
2
8
26
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.96 - Curium atom

Copyright ABRAHAM [2008] - All rights reserved

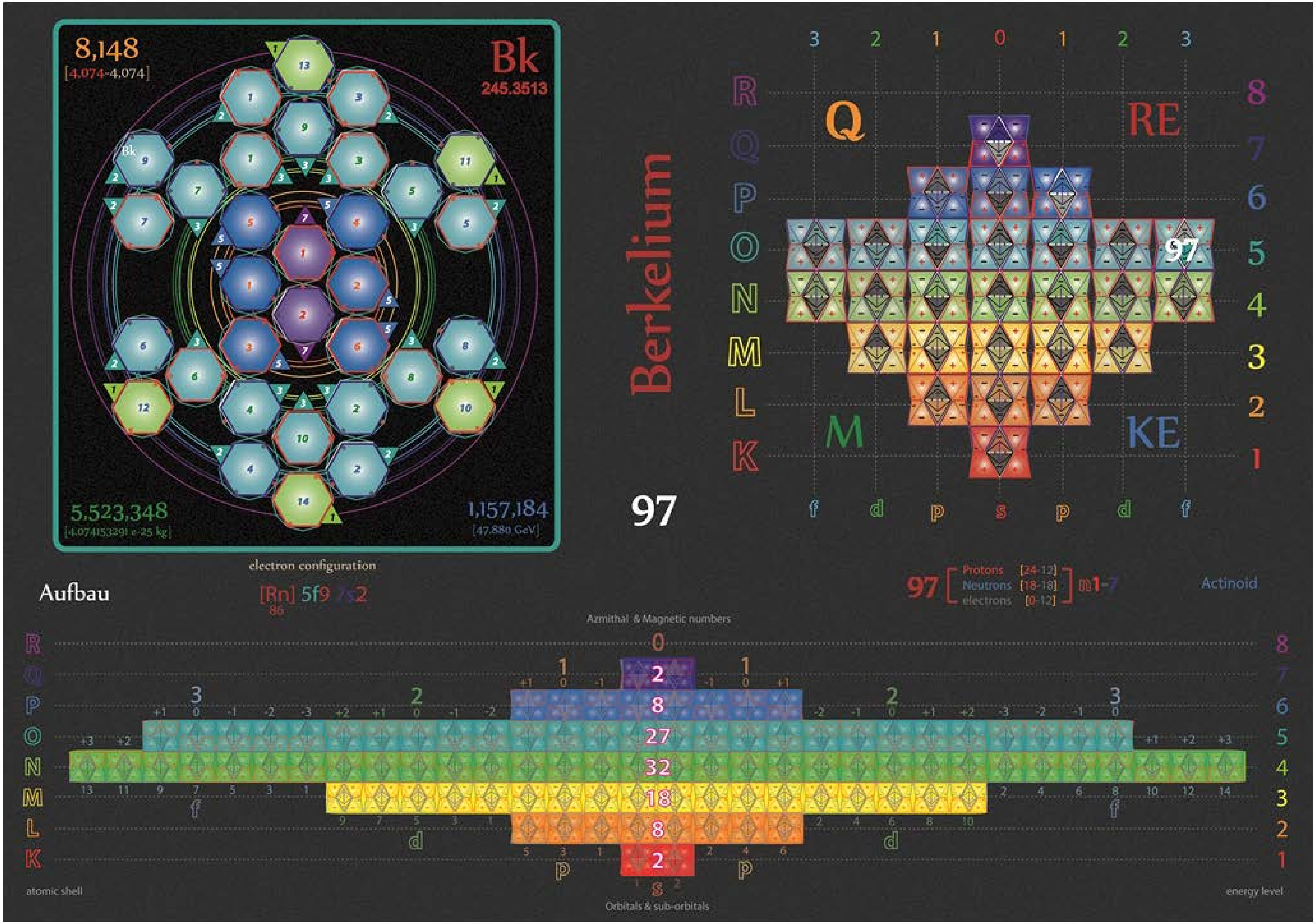
8,148
[4,074-4,074]
Bk
245.3513
5,523,348
[4.07415329l e-25 kg]
1,157,184
[47.880 GeV]
Berkelium
97
R Q P O N M L K
Q RE M KE
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f9 7s2
86
97 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
27
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.97 - Berkelium atom

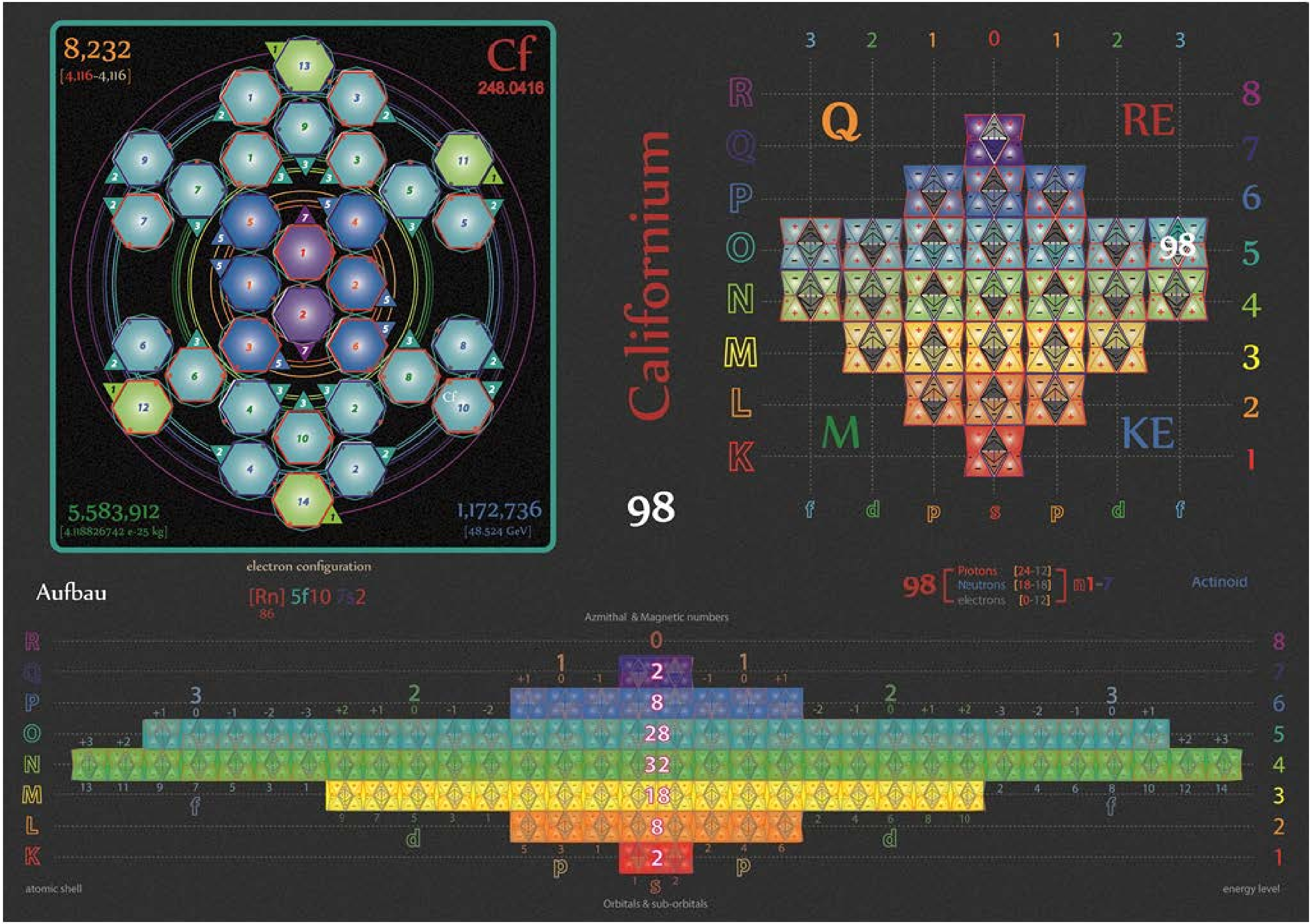
8,232
[4,116-4,116]
Cf
248.0416
5,583,912
[4.118826742 e-25 kg]
1,172,736
[48.524 GeV]
Californium
98
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
98
98 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
electron configuration
Aufbau
[Rn] 5f10 7s2
86
Azmithal & Magnetic numbers
0
2
8
28
32
18
8
2
s
p
d
f
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.98 - Californium atom

Copyright ABRAHAM [2008] - All rights reserved

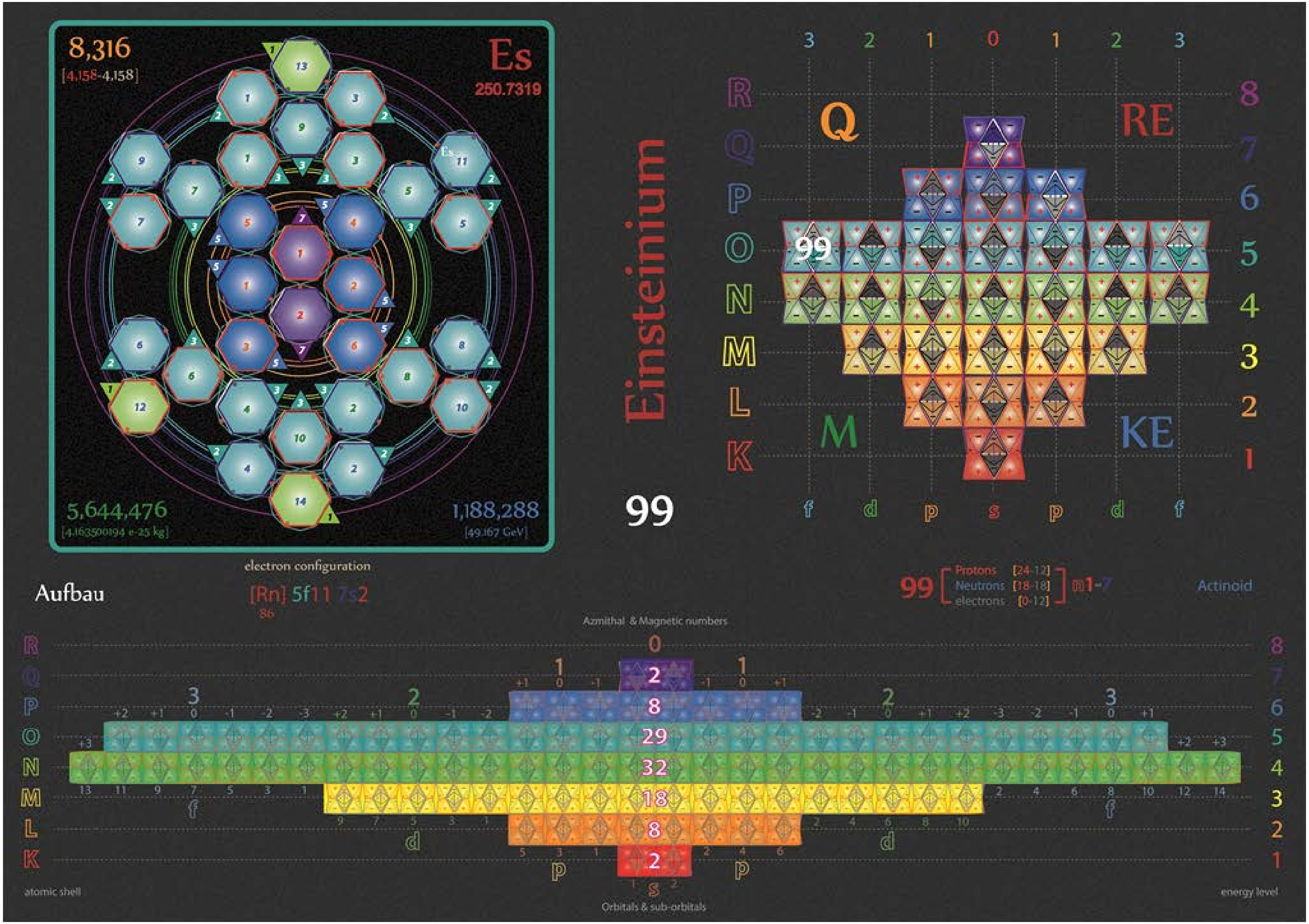
8,316
[4,158-4,158]
Es
250.7319
5,644,476
[4.163500194 e-25 kg]
1,188,288
[49.167 GeV]
Einsteinium
99
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f11 7s2
86
99 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
29
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.99 - Einsteinium atom

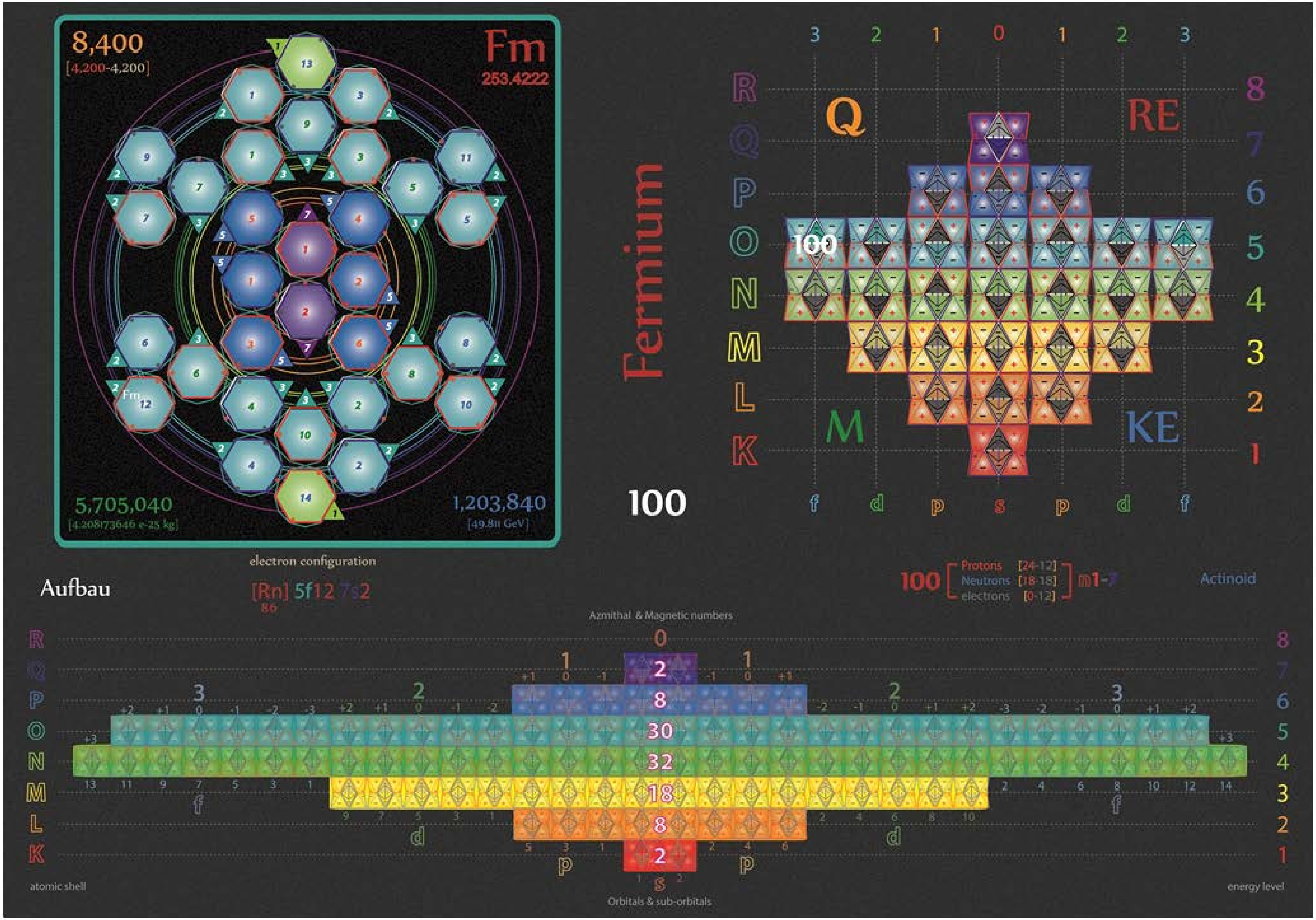
8,400
[4,200-4,200]
Fm
253.4222
5,705,040
[4.208173646 e-25 kg]
1,203,840
[49.811 GeV]
Fermium
100
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f12 7s2
86
100
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
30
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.100 - Fermium atom

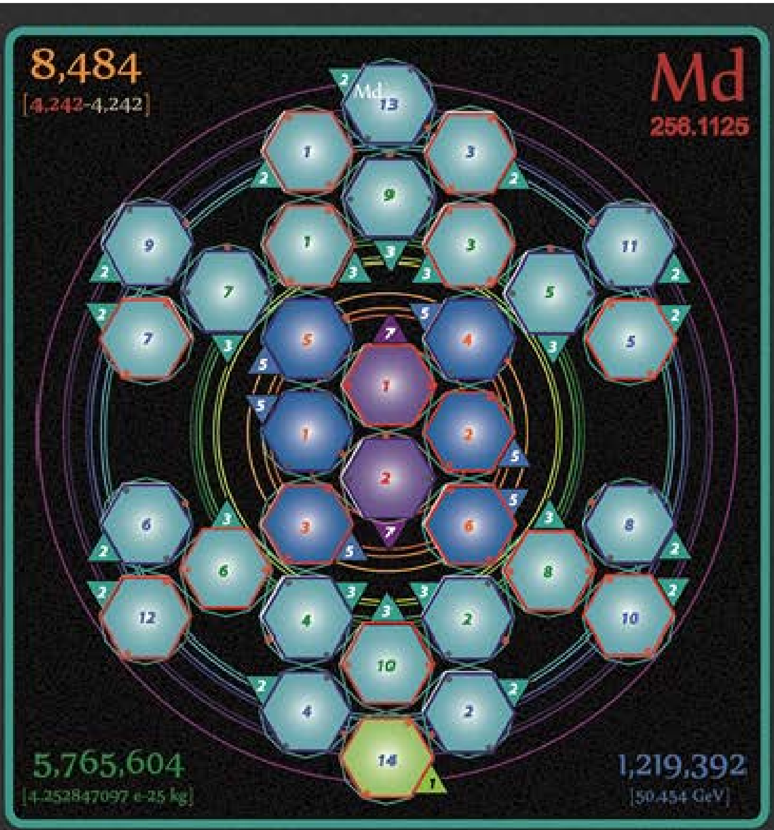
8,484
[4,242-4,242]
Md
256.1125
Md
13
1
3
9
2
9
1
3
11
7
7
5
5
5
4
1
1
2
2
6
8
6
3
6
12
4
10
2
10
4
2
14
5,765,604
[4.252847097 e-25 kg]
1,219,392
[50.454 GeV]

Mendelevium

101

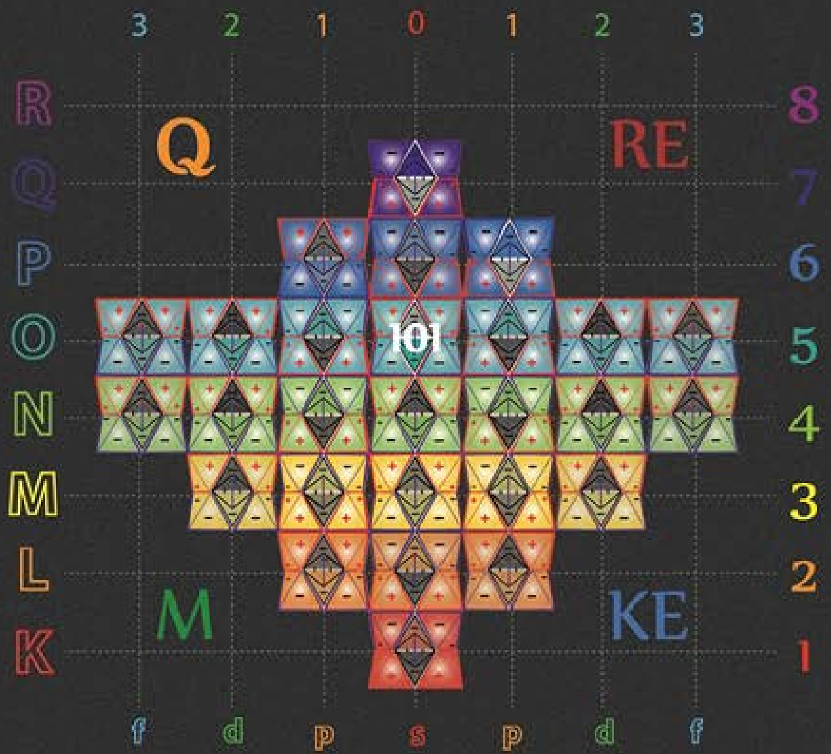
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
101
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f

electron configuration

Aufbau

[Rn] 5f13 7s2
86

101 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Actinoid

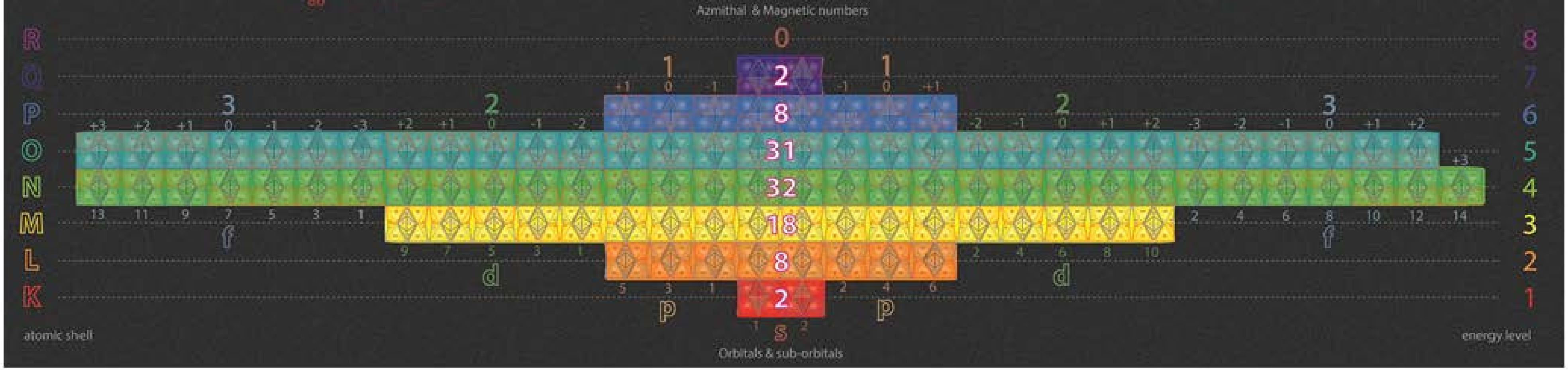
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
0
1
1
2
3
2
2
3
8
31
32
18
8
2
f
d
p
s
p
d
f
8
7
6
5
4
3
2
1
atomic shell
Orbitals & sub-orbitals
energy level

Copyright ABRAHAM [2008] - All rights reserved

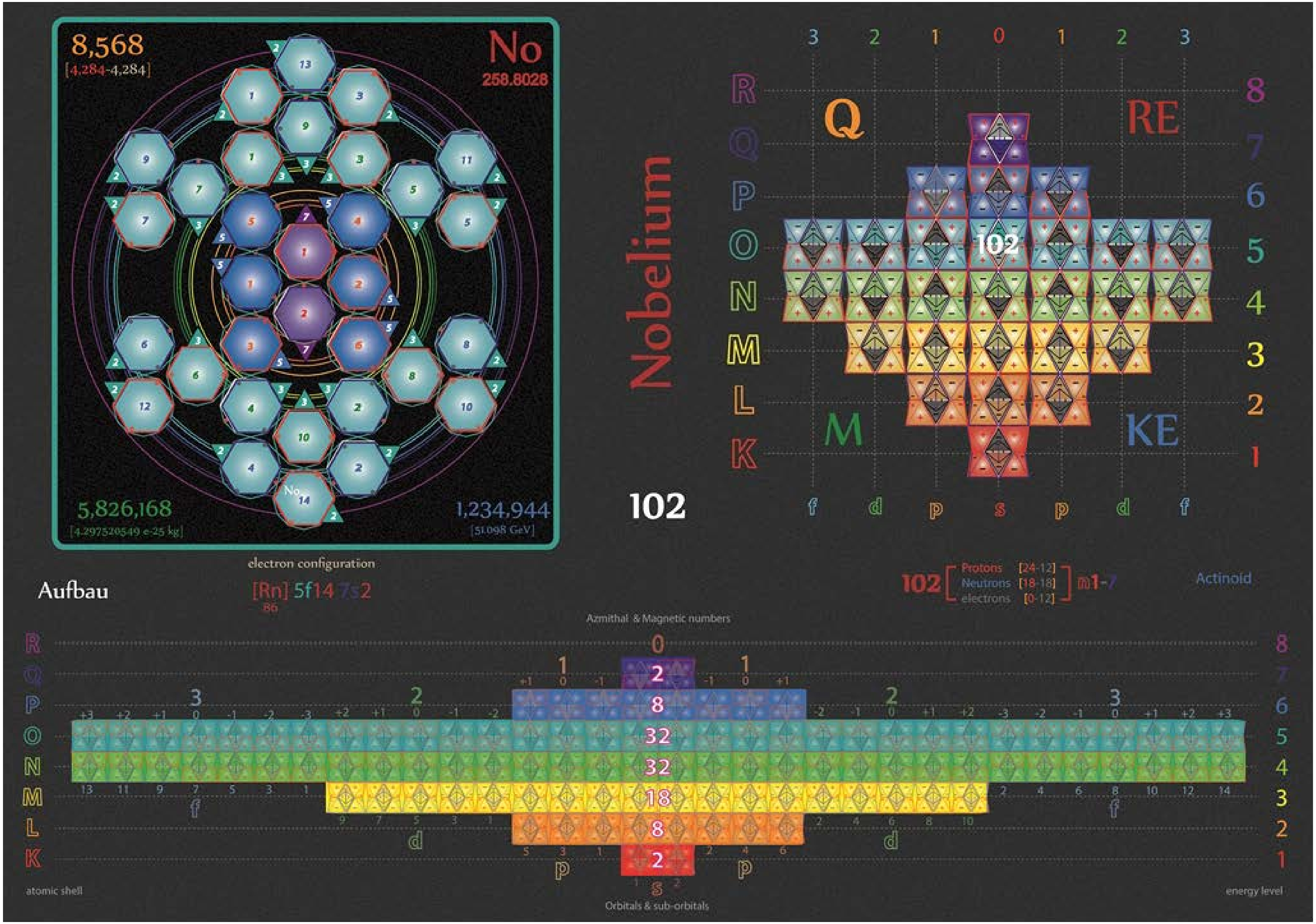
8,568
[4,284-4,284]
No
258.8028
5,826,168
[4.297520549 e-25 kg]
1,234,944
[51.098 GeV]
Nobelium
102
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 7s2
86
102 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Actinoid
Azmithal & Magnetic numbers
2
8
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.102 - Nobelium atom

Copyright ABRAHAM [2008] - All rights reserved

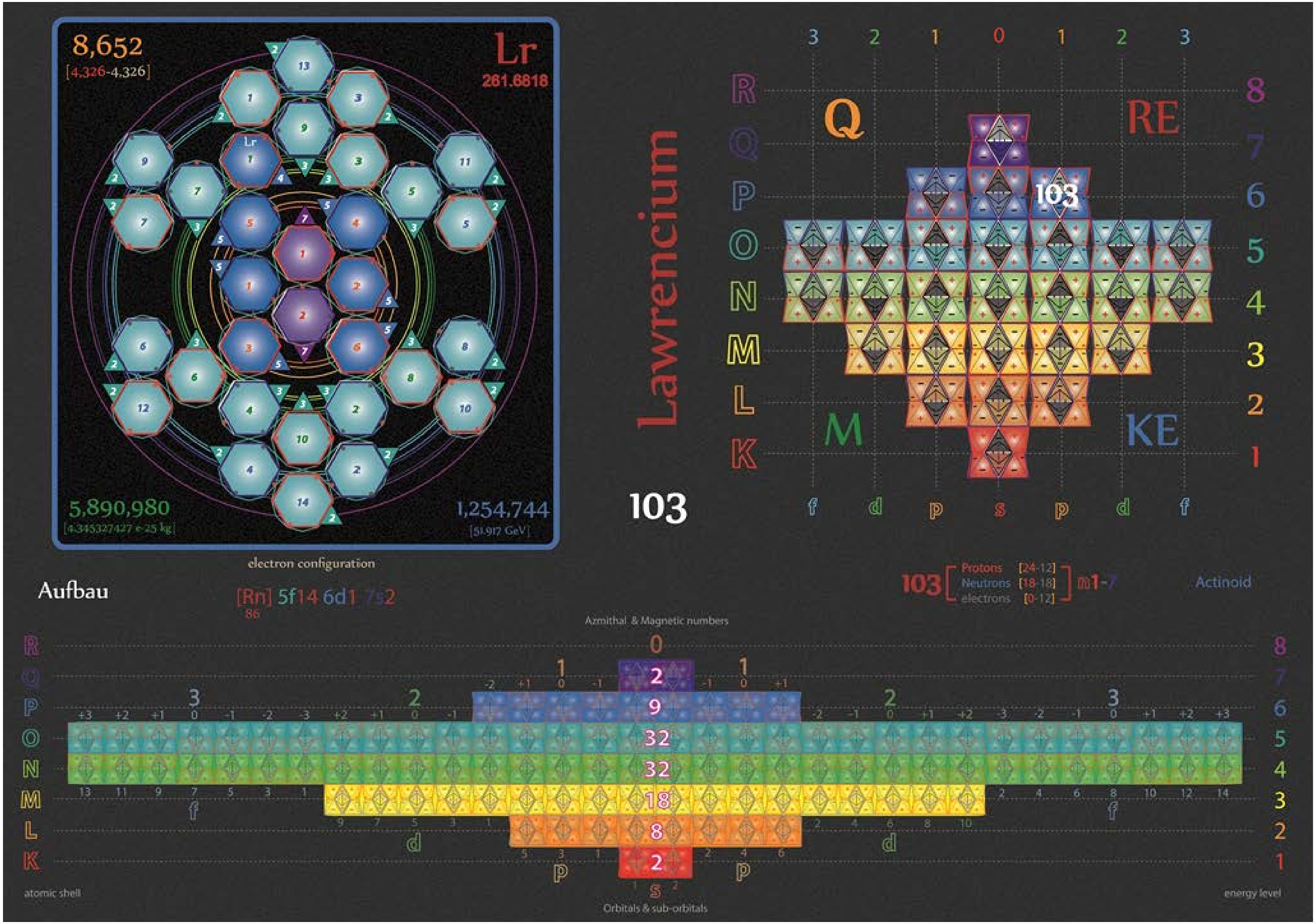
8,652
[4.326-4.326]
Lr
281.6818
5,890,980
[4.345327427 e-25 kg]
1,254,744
[51.917 GeV]
electron configuration
Lawrencium
103
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
103 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Actinoid
Aufbau
[Rn] 5f14 6d1 7s2
86
Azmithal & Magnetic numbers
2
9
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.103 - Lawrencium atom

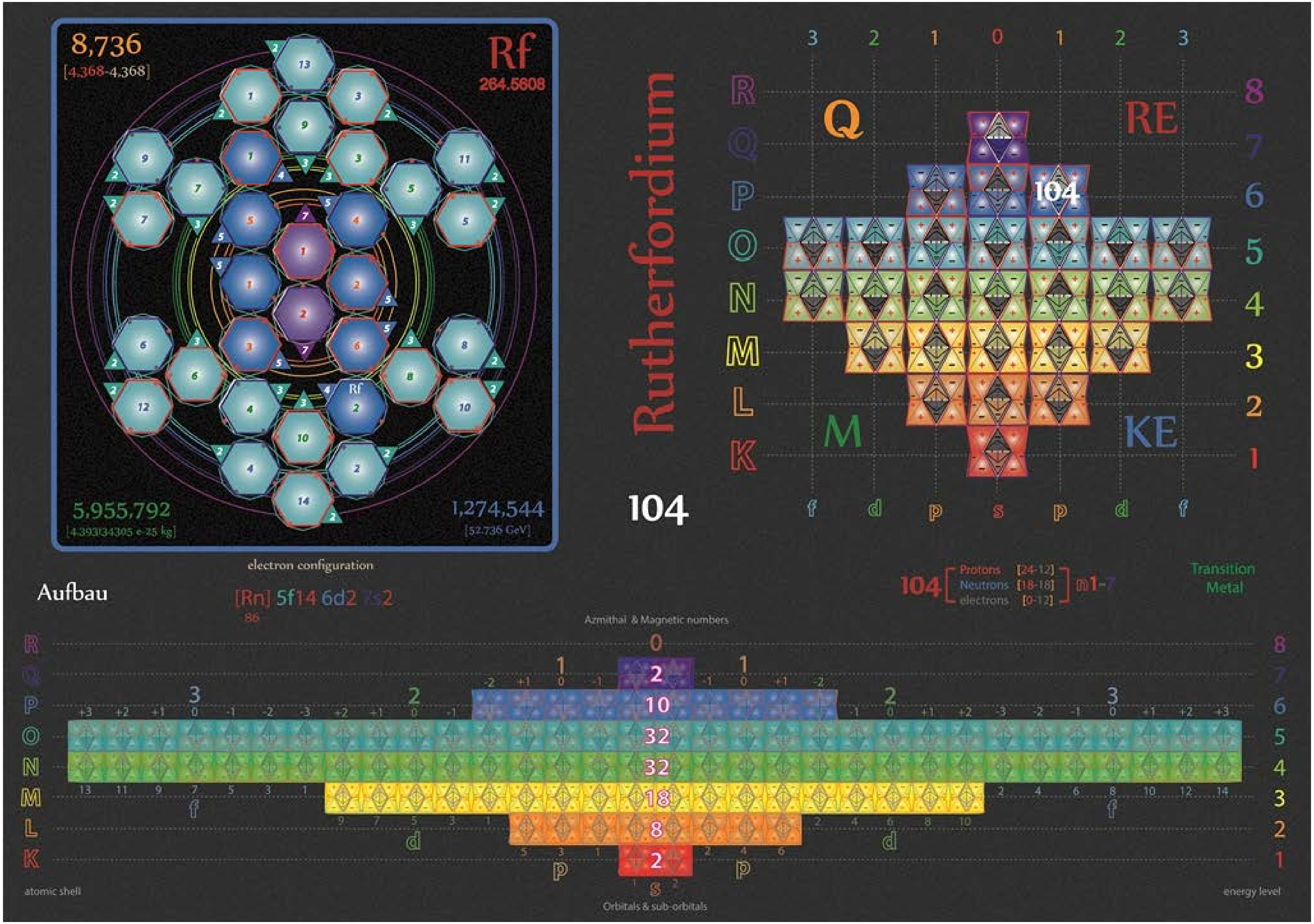
8,736
[4,368-4,368]
Rf
264.5608
5,955,792
[4.393134305 e-25 kg]
1,274,544
[52.736 GeV]
Rutherfordium
104
R
Q
P
O
N
M
L
K
Q
RE
M
KE
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d2 7s2
86
104 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Transition Metal
Azmithal & Magnetic numbers
0
1
2
3
2
10
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.104 - Rutherfordium atom

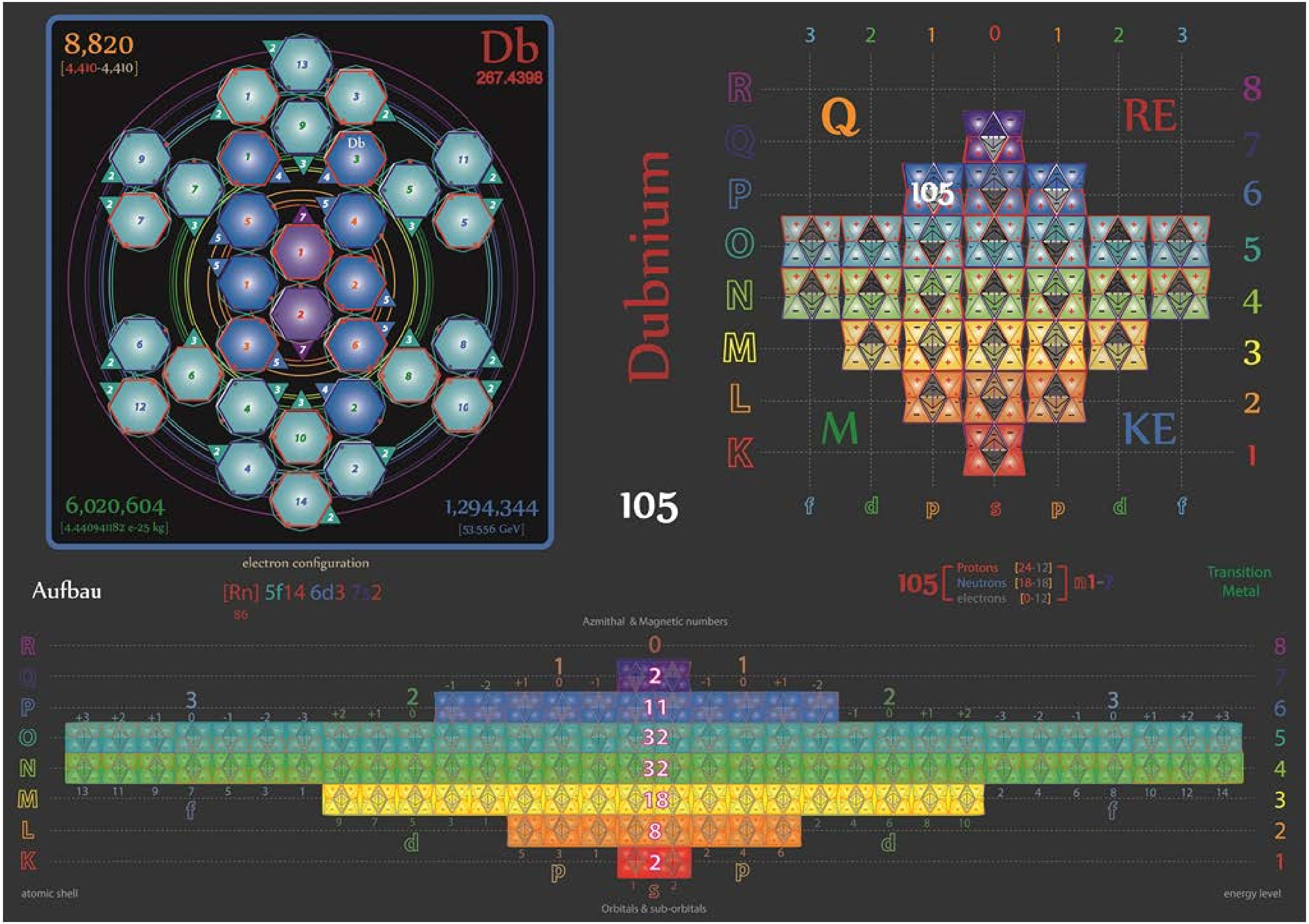
8,820
[4,410-4,410]
Db
287.4398
6,020,604
[4.44094182 e-25 kg]
1,294,344
[53.556 GeV]
Dubnium
105
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d3 7s2
86
105 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
2
11
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.105 - Dubnium atom

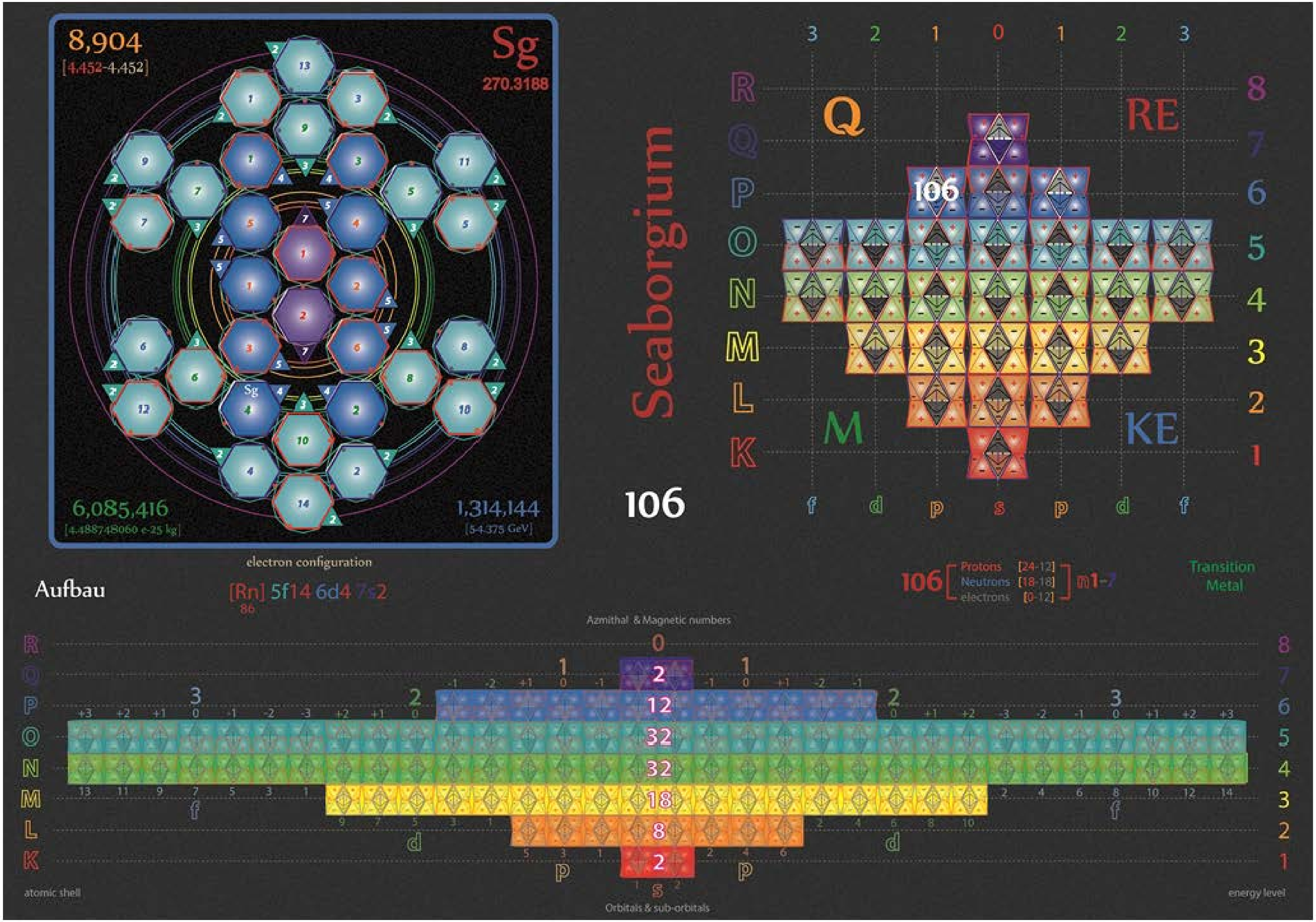
8,904
[4,452-4,452]
Sg
270.3188
6,085,416
[4.488748060 e-25 kg]
1,314,144
[54.375 GeV]
Seaborgium
106
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
Q
RE
M
KE
electron configuration
Aufbau
[Rn] 5f14 6d4 7s2
86
106 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2
12
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.106 - Seaborgium atom

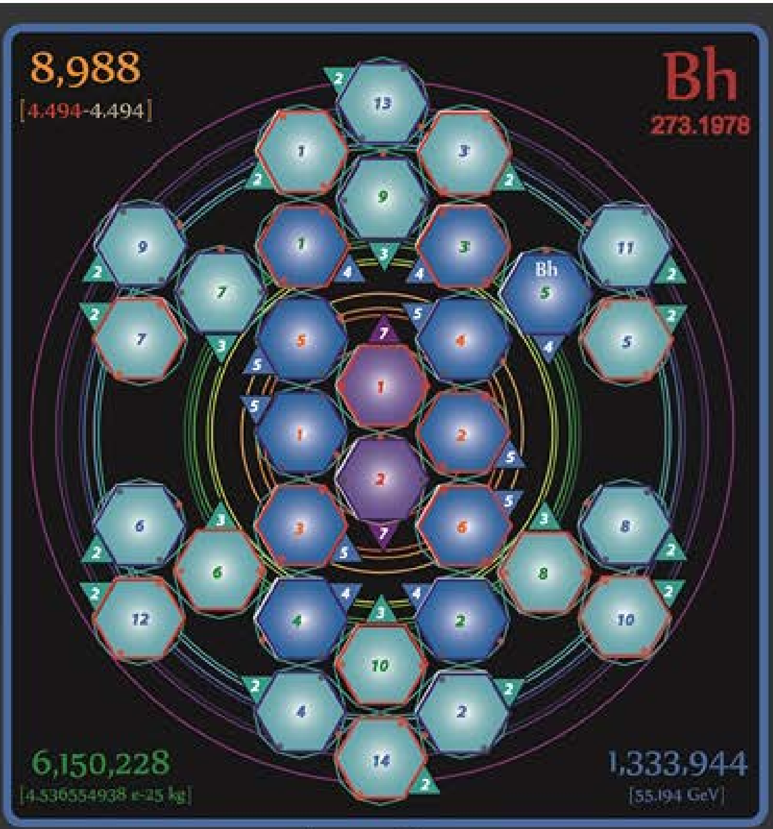
8,988
[4.494-4.494]
Bh
273.1978
6,150,228
[4.536554938 e-25 kg]
1,333,944
[55.194 GeV]

Bohrium

107

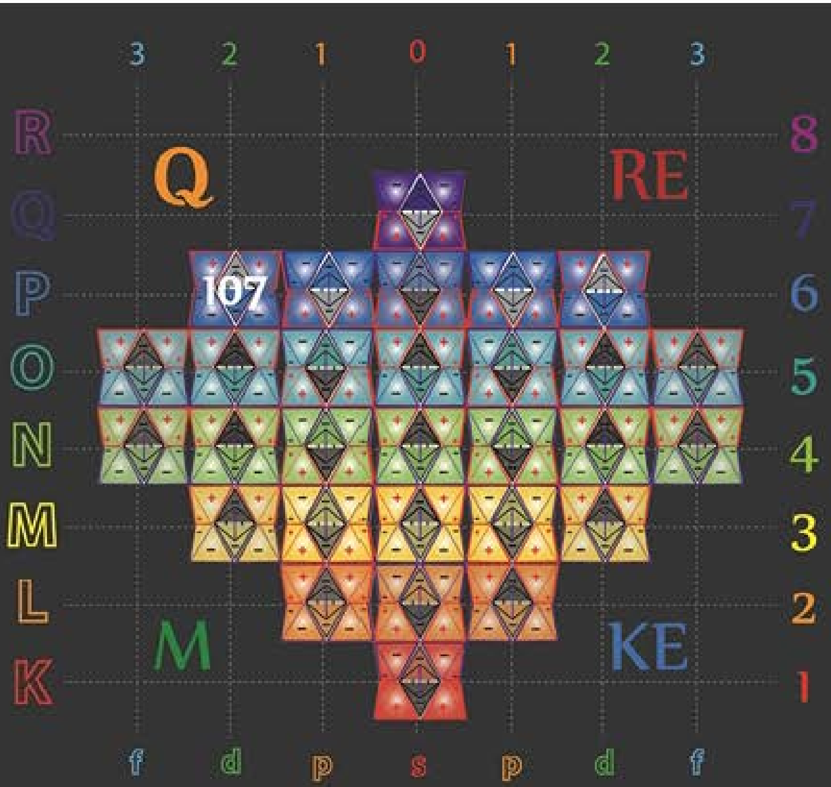
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
107
f d p s p d f

electron configuration

Aufbau

[Rn] 5f14 6d5 7s2
86

107 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7

Transition Metal

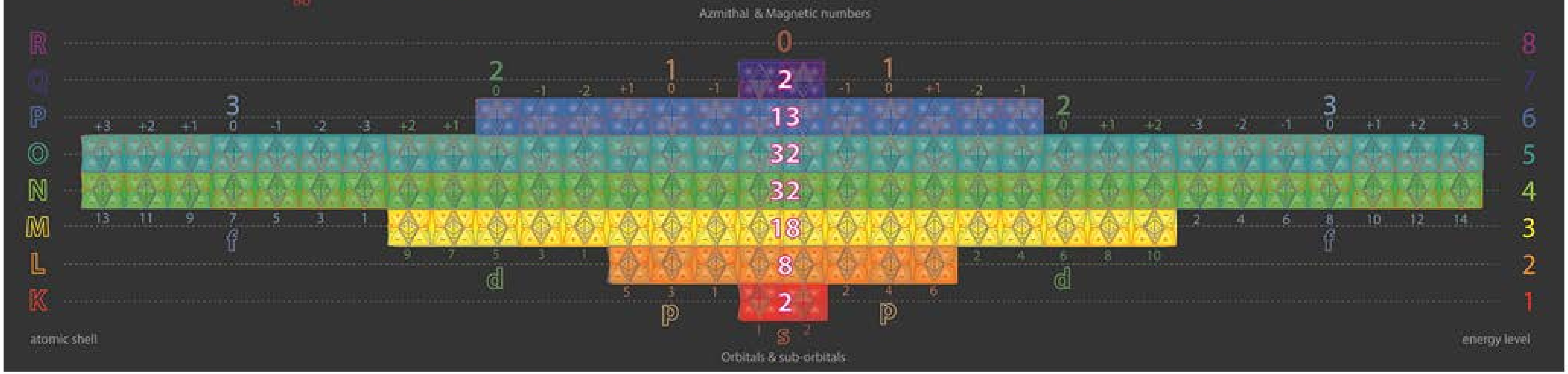
Azmithal & Magnetic numbers
R Q P O N M L K
8 7 6 5 4 3 2 1
0
2
13
32
32
18
8
2
3 2 1 1 2 3
+3 +2 +1 0 -1 -2 -3 +2 +1 0 -1 -2 +1 0 -1 +1 0 -1 +1 -2 -1 0 +1 +2 -3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.107 - Bohrium atom

Copyright ABRAHAM [2008] - All rights reserved

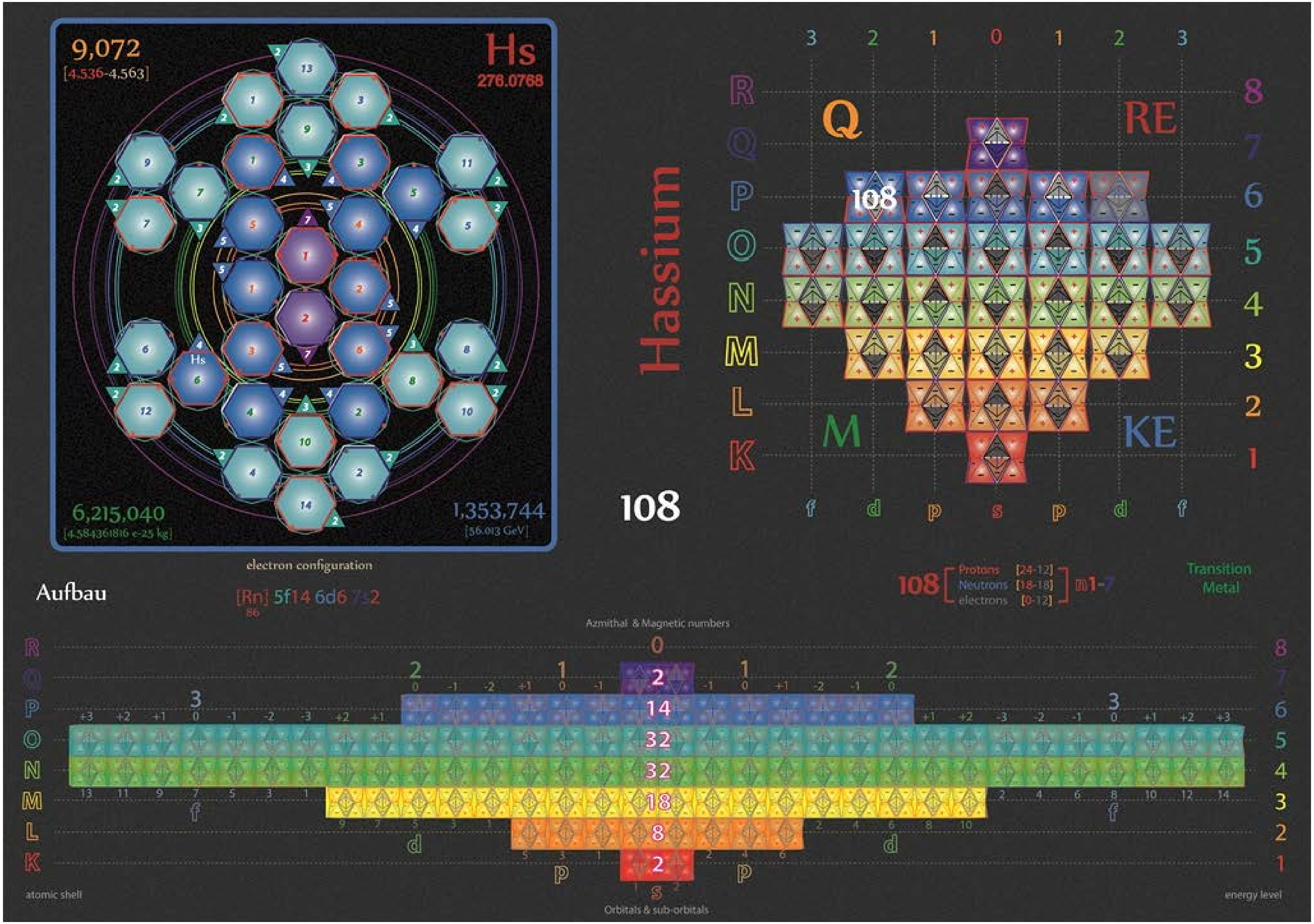
9,072
[4.536-4.563]
Hs
276.0768
6,215,040
[4.584361816 e-25 kg]
1,353,744
[56.013 GeV]
Hassium
108
R Q P O N M L K
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d6 7s2
86
108 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Transition Metal
Azmithal & Magnetic numbers
0
2 14 32 32 18 8 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.108 - Hassium atom

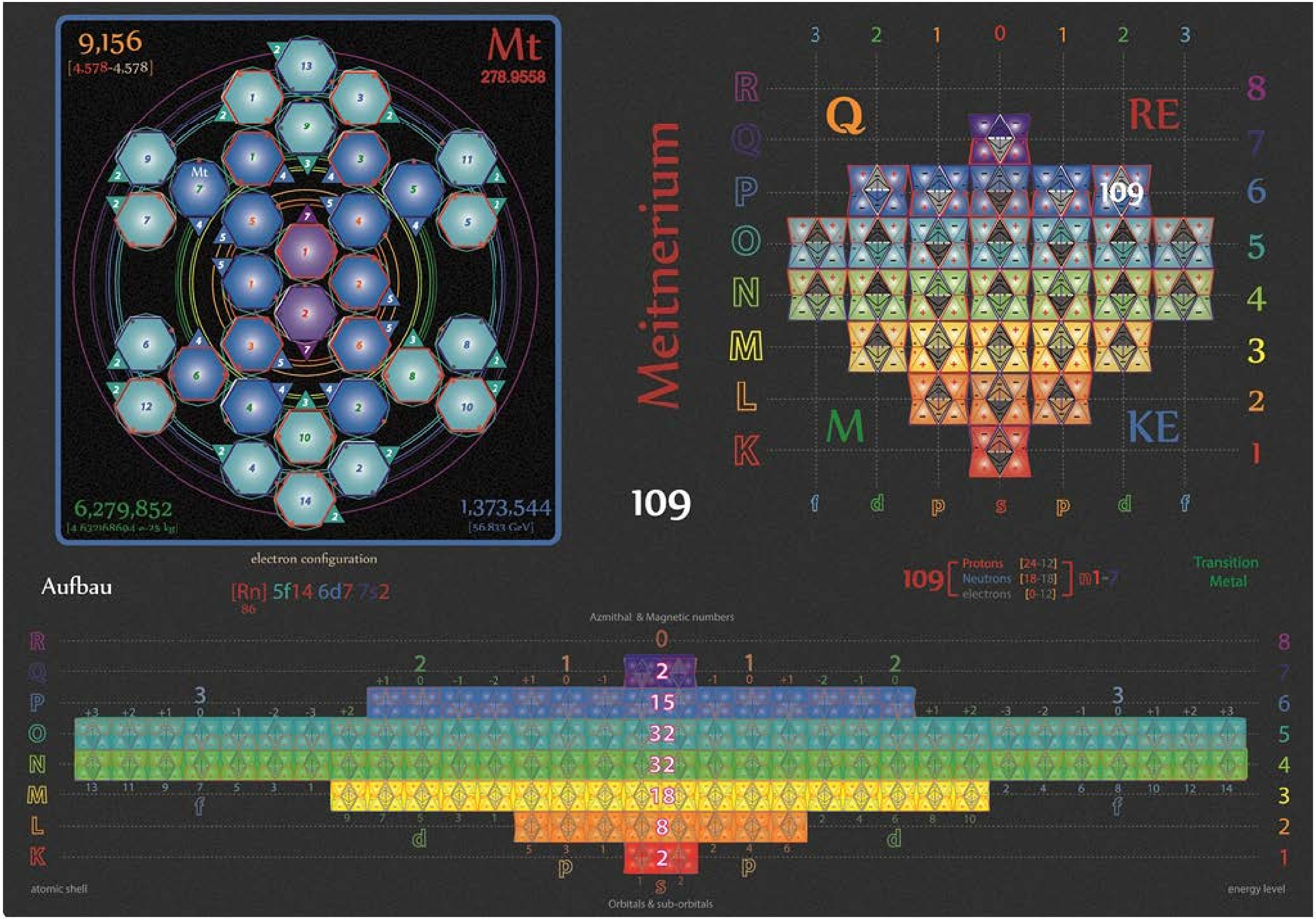
9,156
[4.578-4.578]
Mt
278.9558
6,279,852
[4.632168604 e-25 kg]
1,373,544
[56.833 GeV]
Meitnerium
109
R
Q
P
O
N
M
L
K
3
2
1
0
1
2
3
Q
RE
M
KE
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d7 7s2
86
109
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Transition
Metal
Azimuthal & Magnetic numbers
2
15
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.109 - Meitnerium atom

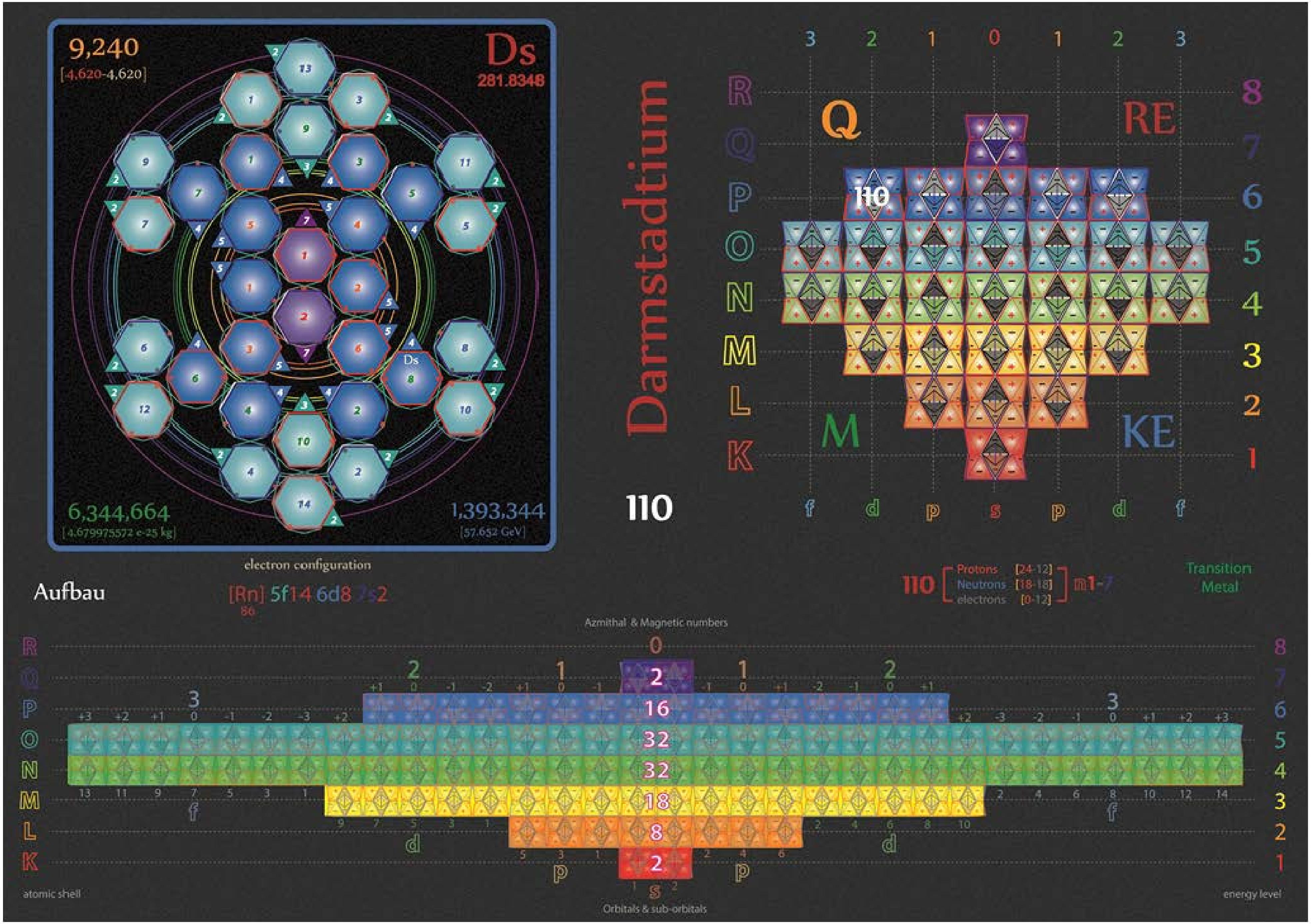
9,240
[4,620-4,620]
Ds
281.8348
Darmstadtium
6,344,664
[4.679975572 e-25 kg]
1,393,344
[57.652 GeV]
110
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
110
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d8 7s2
86
110 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Transition Metal
Azmithal & Magnetic numbers
0
2
16
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.110 - Darmstadtium atom

Copyright ABRAHAM [2008] - All rights reserved

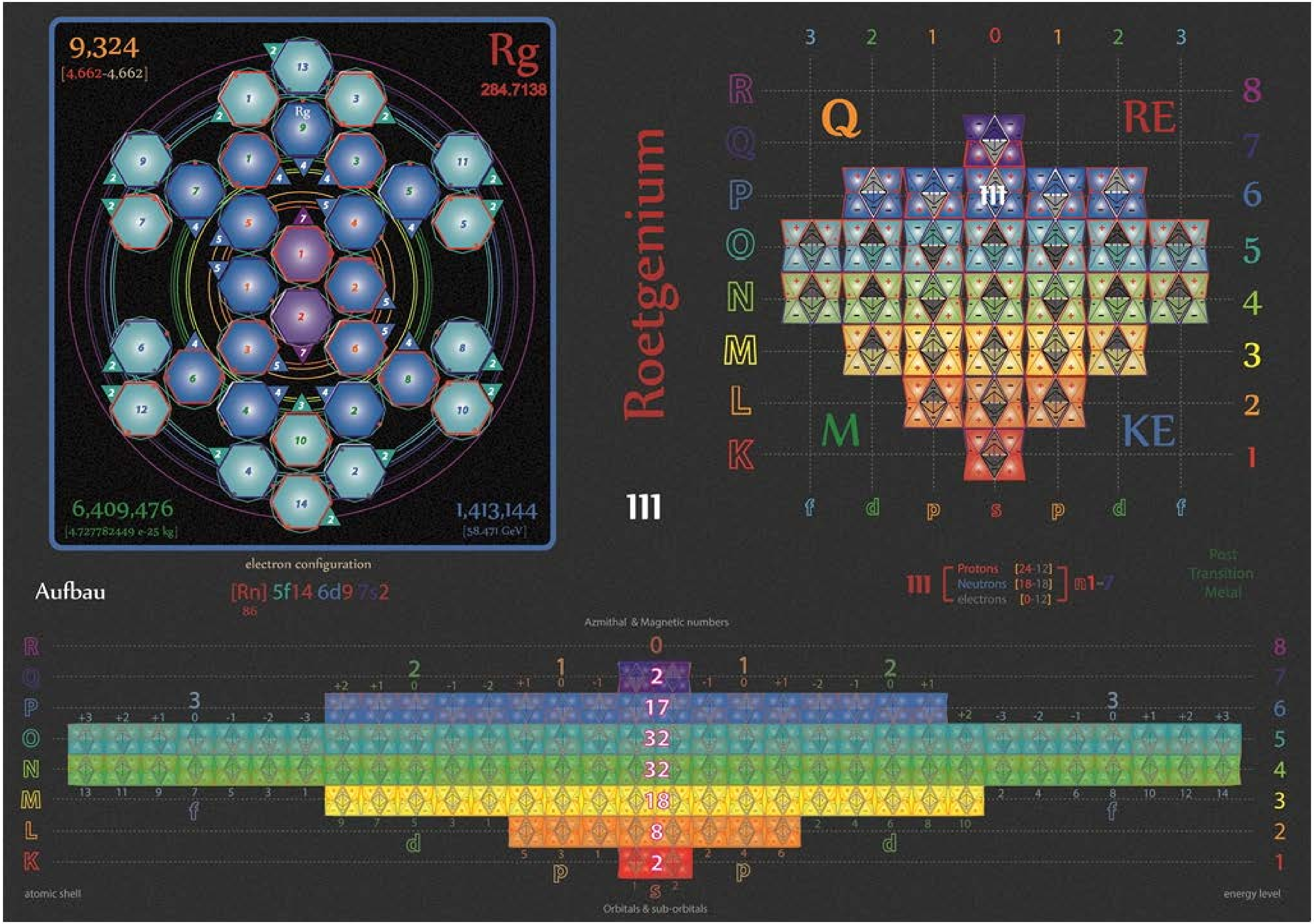
9,324
[4,662-4,662]
Rg
284.7138
6,409,476
[4.727782449 e-25 kg]
1,413,144
[58.471 GeV]
Roetgenium
111
Q
RE
M
KE
R Q P O N M L K
8 7 6 5 4 3 2 1
3 2 1 0 1 2 3
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d9 7s2
86
111 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Post Transition Metal
Azmithal & Magnetic numbers
2
17
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.111 - Roetgenium atom

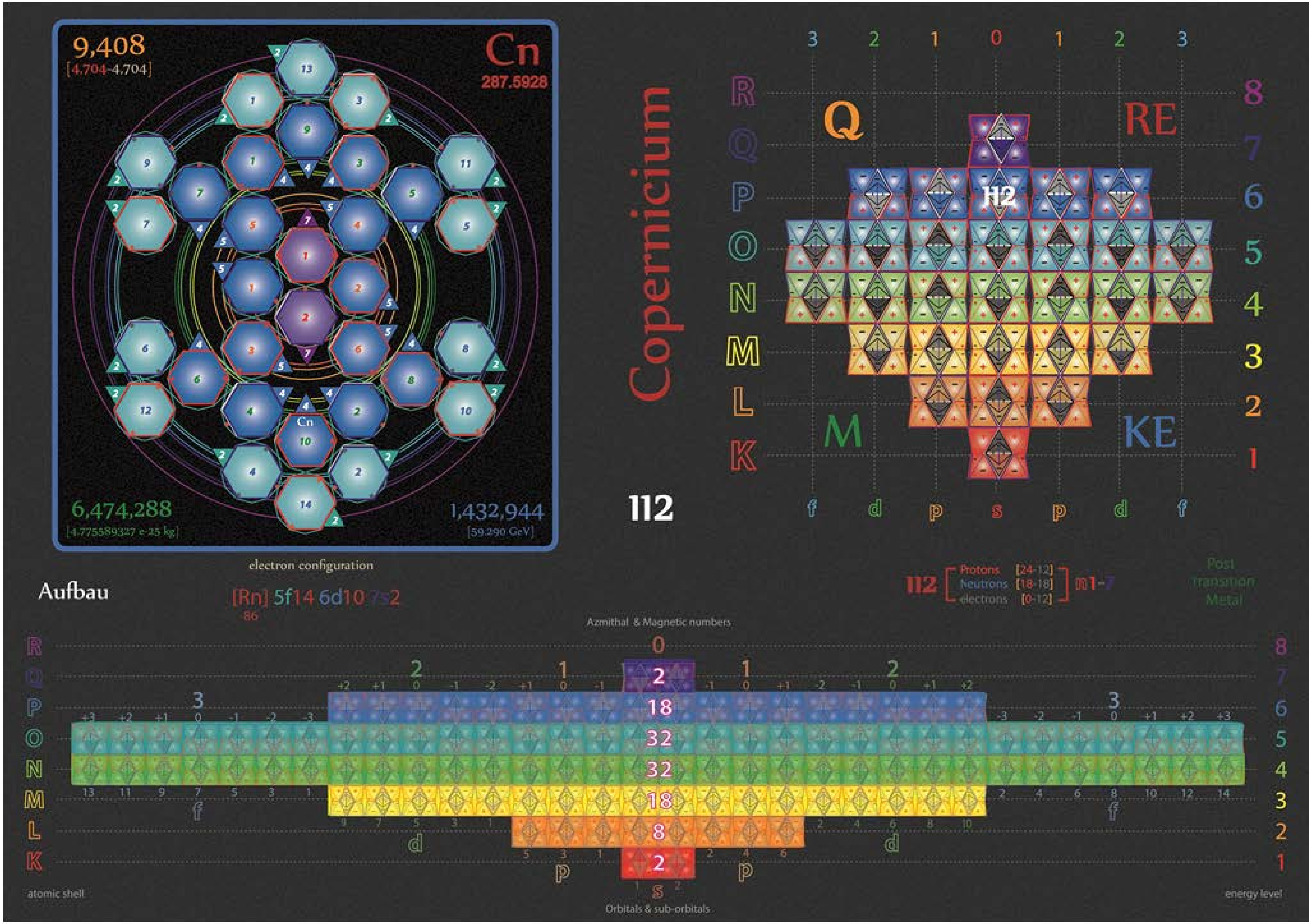
9,408
[4,704-4,704]
Cn
287.5928
Cn
6,474,288
[4.775589327 e-25 kg]
1,432,944
[59.290 GeV]
Copernicium
112
3
2
1
0
1
2
3
R
Q
P
O
N
M
L
K
Q
RE
M
KE
112
8
7
6
5
4
3
2
1
f
d
p
s
p
d
f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2
86
112 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Post transition Metal
Azmithal & Magnetic numbers
0
2
18
32
32
18
8
2
+3 +2 +1 0 -1 -2 -3
+2 +1 0 -1 -2
+1 0 -1
-3 -2 -1 0 +1 +2 +3
13 11 9 7 5 3 1
2 4 6 8 10 12 14
9 7 5 3 1
2 4 6 8 10
5 3 1
2 4 6
1 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.112 - Copernicium atom

Copyright ABRAHAM [2008] - All rights reserved

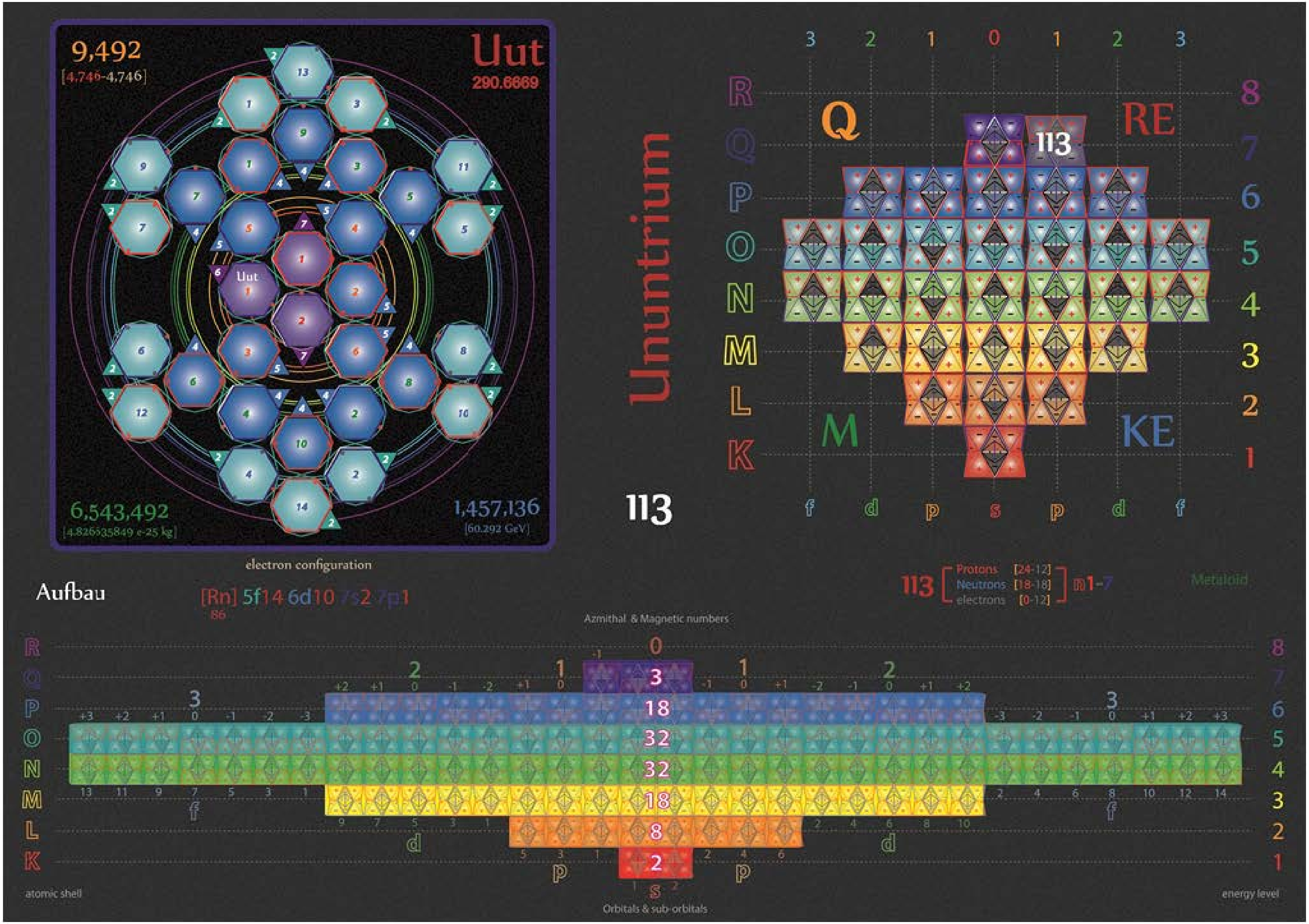
9,492
[4,746-4,746]
Uut
290.6669
Uut
6,543,492
[4.826535849 e-25 kg]
1,457,136
[60.292 GeV]
electron configuration
Ununtrium
113
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
Q
RE
M
KE
113
8
7
6
5
4
3
2
1
f d p s p d f
113
Protons [24-12]
Neutrons [18-18]
electrons [0-12]
n1-7
Metaloid
Aufbau
[Rn] 5f14 6d10 7s2 7p1
86
Azmithal & Magnetic numbers
0
3
18
32
32
18
8
2
f
d
p
s
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.113 - Ununtrium atom

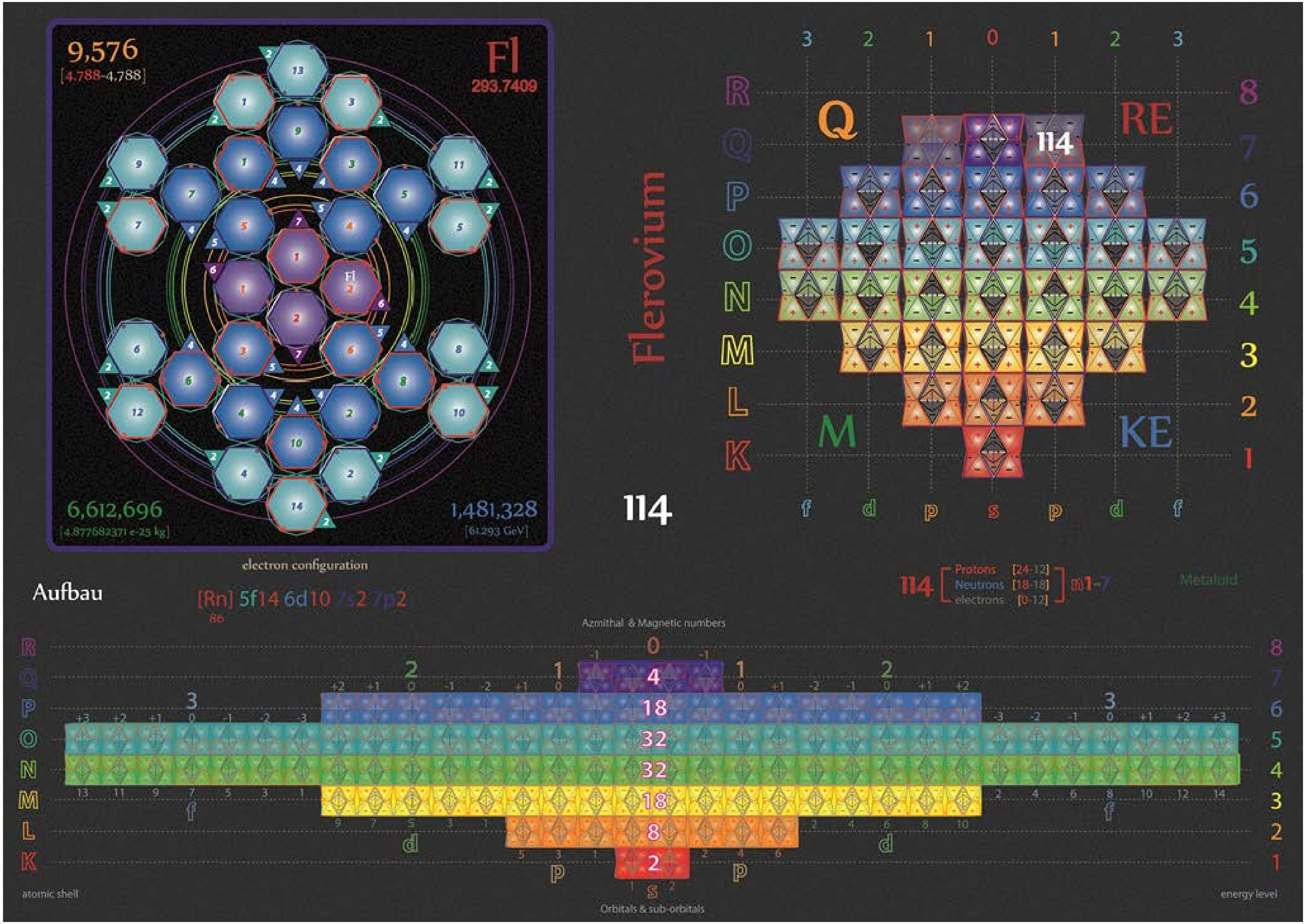
9,576
[4.788-4.788]
Fl
293.7409
6,612,696
[4.877682371 e-25 kg]
1,481,328
[61.293 GeV]
Flerovium
114
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p2
86
Q
RE
M
KE
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
3 2 1 0 1 2 3
f d p s p d f
114 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Metaloid
Azmithal & Magnetic numbers
0
4
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.114 - Flerovium atom

Copyright ABRAHAM [2008] - All rights reserved

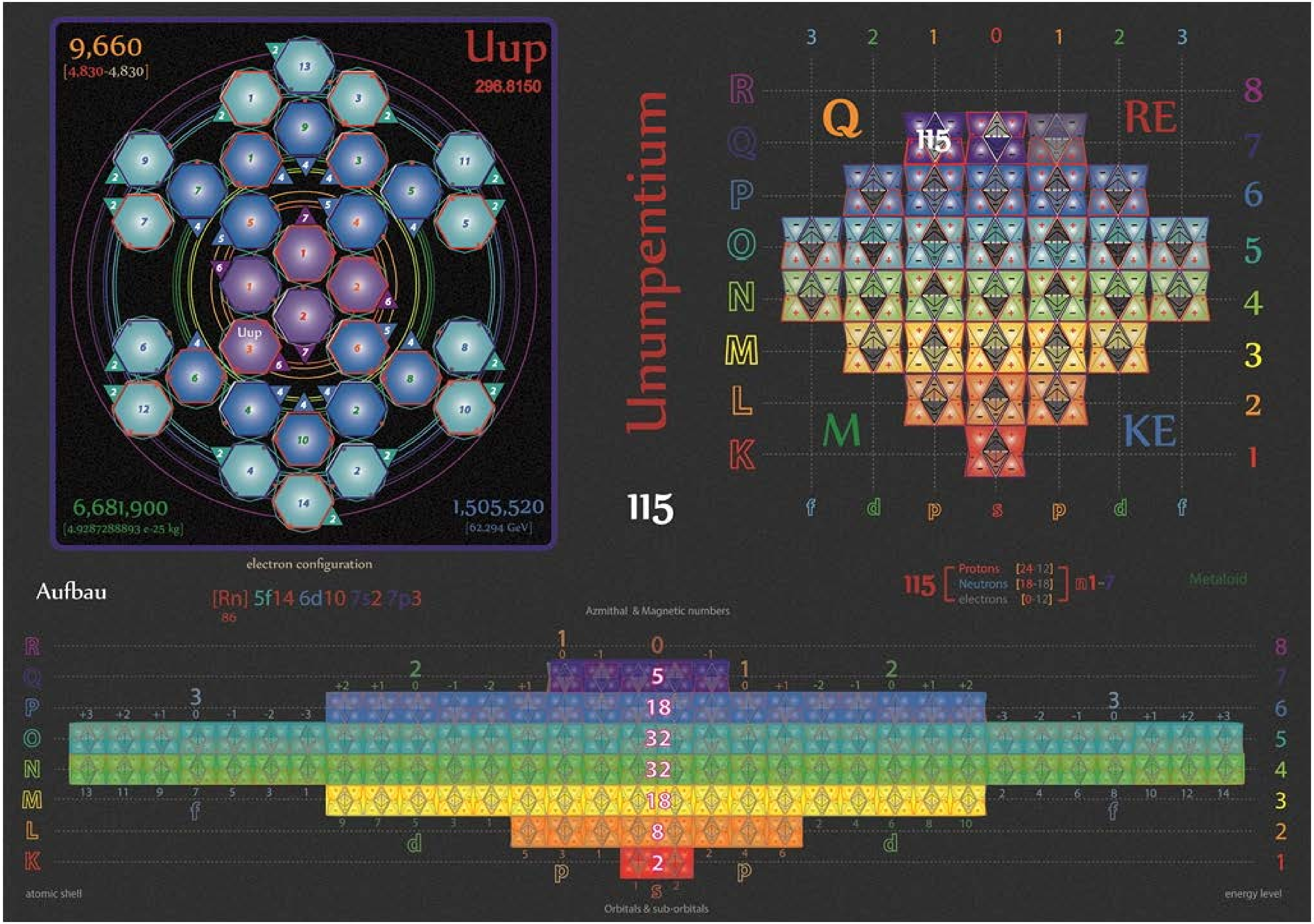
9,660
[4,830-4,830]
Uup
296.8150
6,681,900
[4.9287288893 e-25 kg]
1,505,520
[62.294 GeV]
Uup
electron configuration
Ununpentium
115
Q
RE
M
KE
115
R
Q
P
O
N
M
L
K
3 2 1 0 1 2 3
f d p s p d f
8 7 6 5 4 3 2 1
115 Protons [24-12] Neutrons [18-18] electrons [0-12] n1-7
Metaloid
Aufbau
[Rn] 5f14 6d10 7s2 7p3
86
Azmithal & Magnetic numbers
5
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.115 - Ununpentium atom

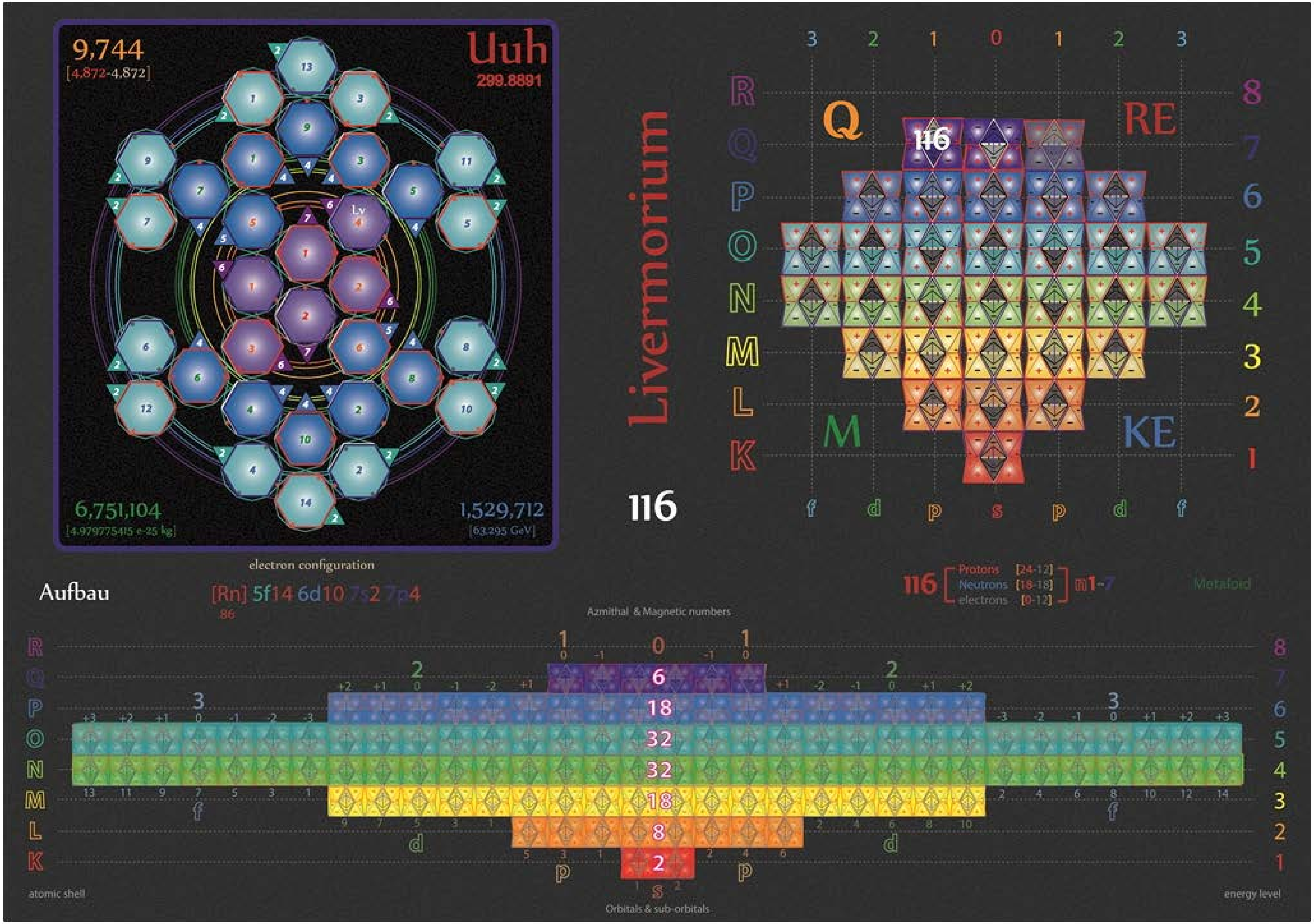
9,744
[4,872-4,872]
Uuh
299.8891
Lv
6,751,104
[4.979775415 e-25 kg]
1,529,712
[63.295 GeV]
Livermorium
116
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p4
86
116 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Metaloid
Azmithal & Magnetic numbers
6
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.116 - Livermorium atom

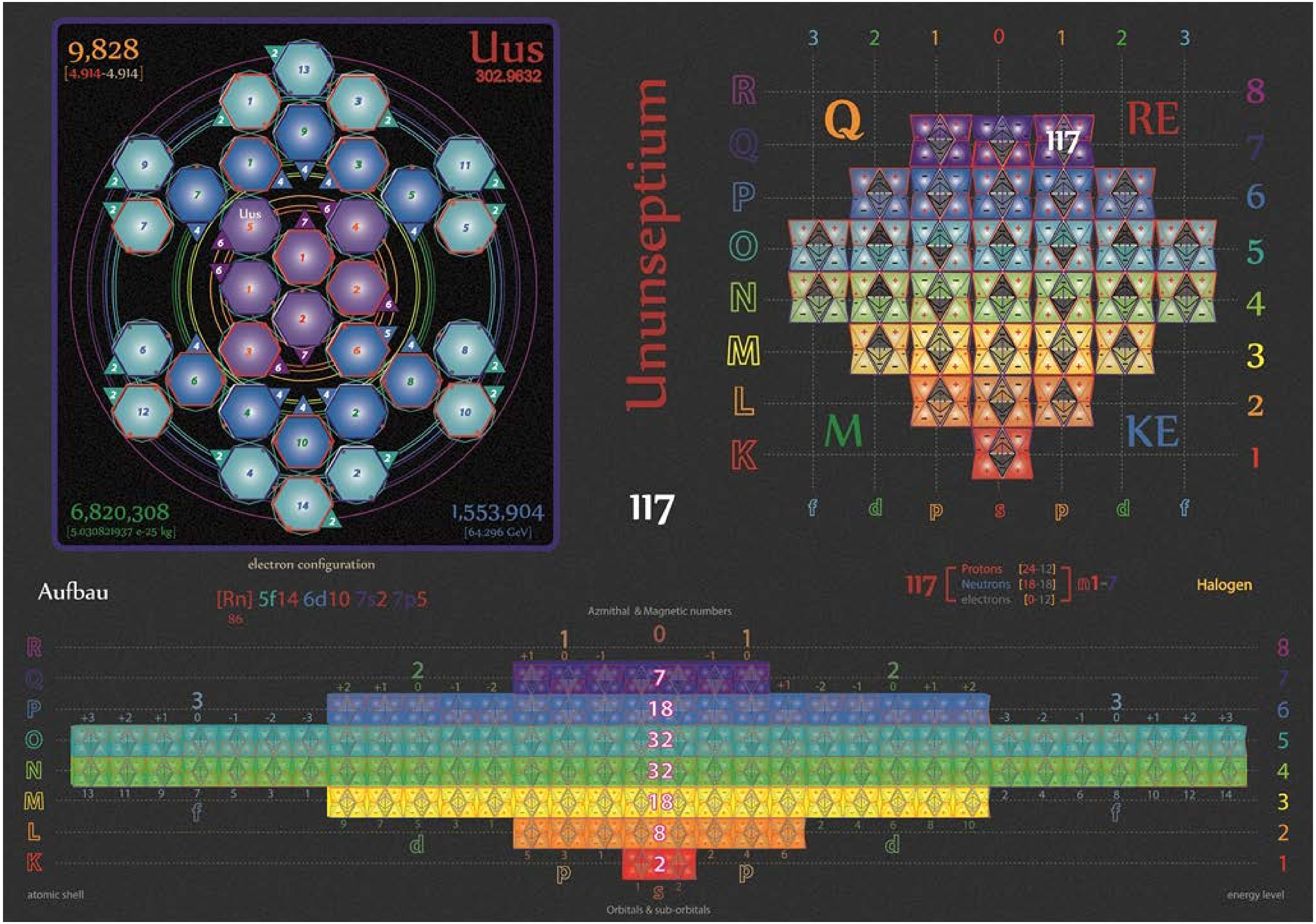
9,828
[4.914-4.914]
Uus
302.9832
Uus
6,820,308
[5.030821937 e-25 kg]
1,553,904
[64.296 GeV]
Ununseptium
117
3 2 1 0 1 2 3
R Q P O N M L K
Q
RE
117
M
KE
8 7 6 5 4 3 2 1
f d p s p d f
electron configuration
Aufbau
[Rn] 5f14 6d10 7s2 7p5
86
117 [Protons [24-12] Neutrons [18-18] electrons [0-12]] m1-7
Halogen
Azmithal & Magnetic numbers
7
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.117 - Ununseptium atom

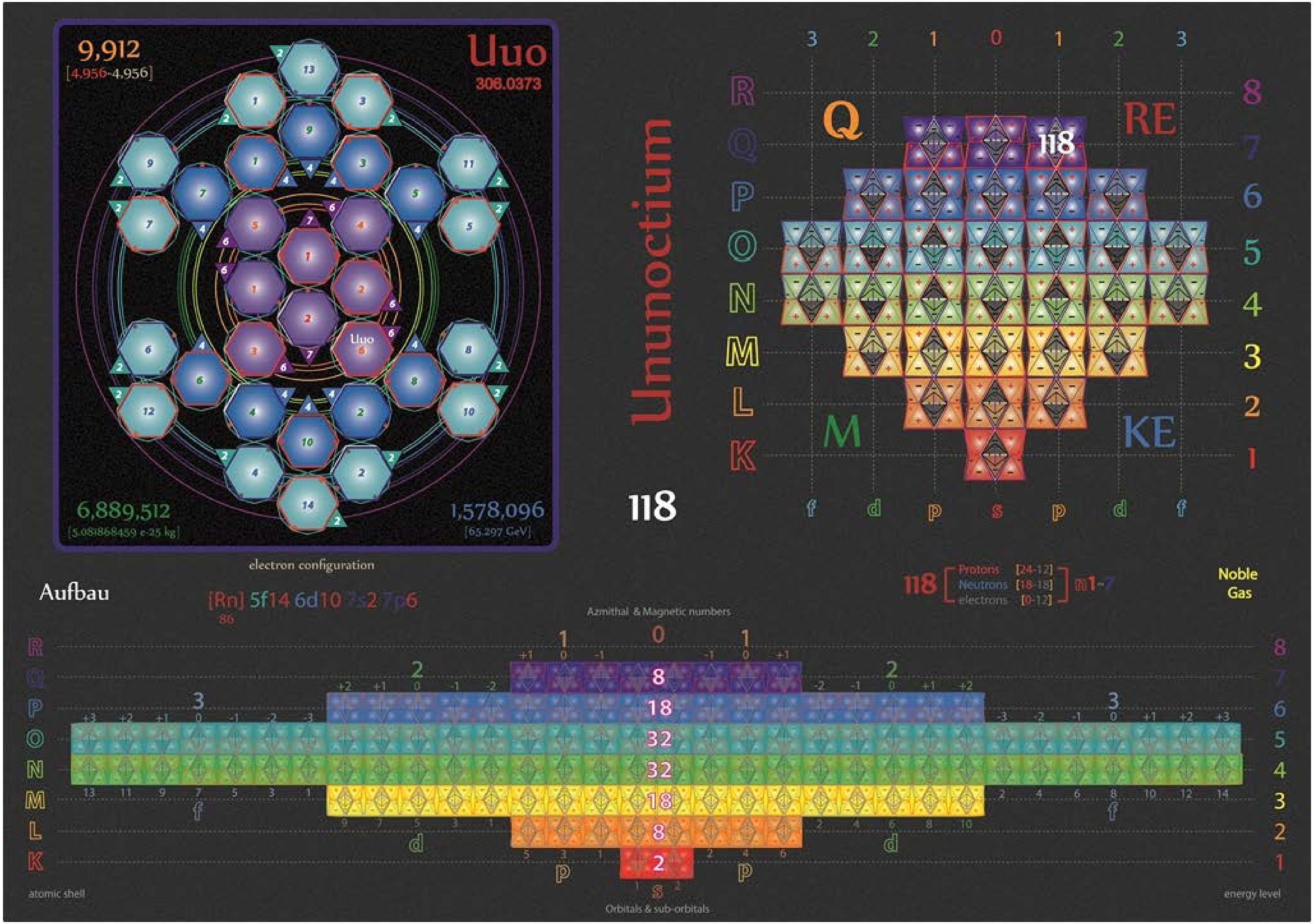
9,912
[4,956-4,956]
Uuo
308.0373
Uuo
6,889,512
[5.081868459 e-25 kg]
1,578,096
[65.297 GeV]
electron configuration
Unuoctium
118
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
118
f d p s p d f
118 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-7
Noble Gas
Aufbau
[Rn] 5f14 6d10 7s2 7p6
86
Azmithal & Magnetic numbers
1 0 1
2 2
3 3
+1 0 -1 -1 0 +1
+2 +1 0 -1 -2 -2 -1 0 +1 +2
+3 +2 +1 0 -1 -2 -3 -3 -2 -1 0 +1 +2 +3
8
18
32
32
18
8
2
13 11 9 7 5 3 1 2 4 6 8 10 12 14
f f
9 7 5 3 1 2 4 6 8 10
d d
5 3 1 2 4 6
p p
1 s 2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.118 - Ununoctium atom

Copyright ABRAHAM [2008] - All rights reserved

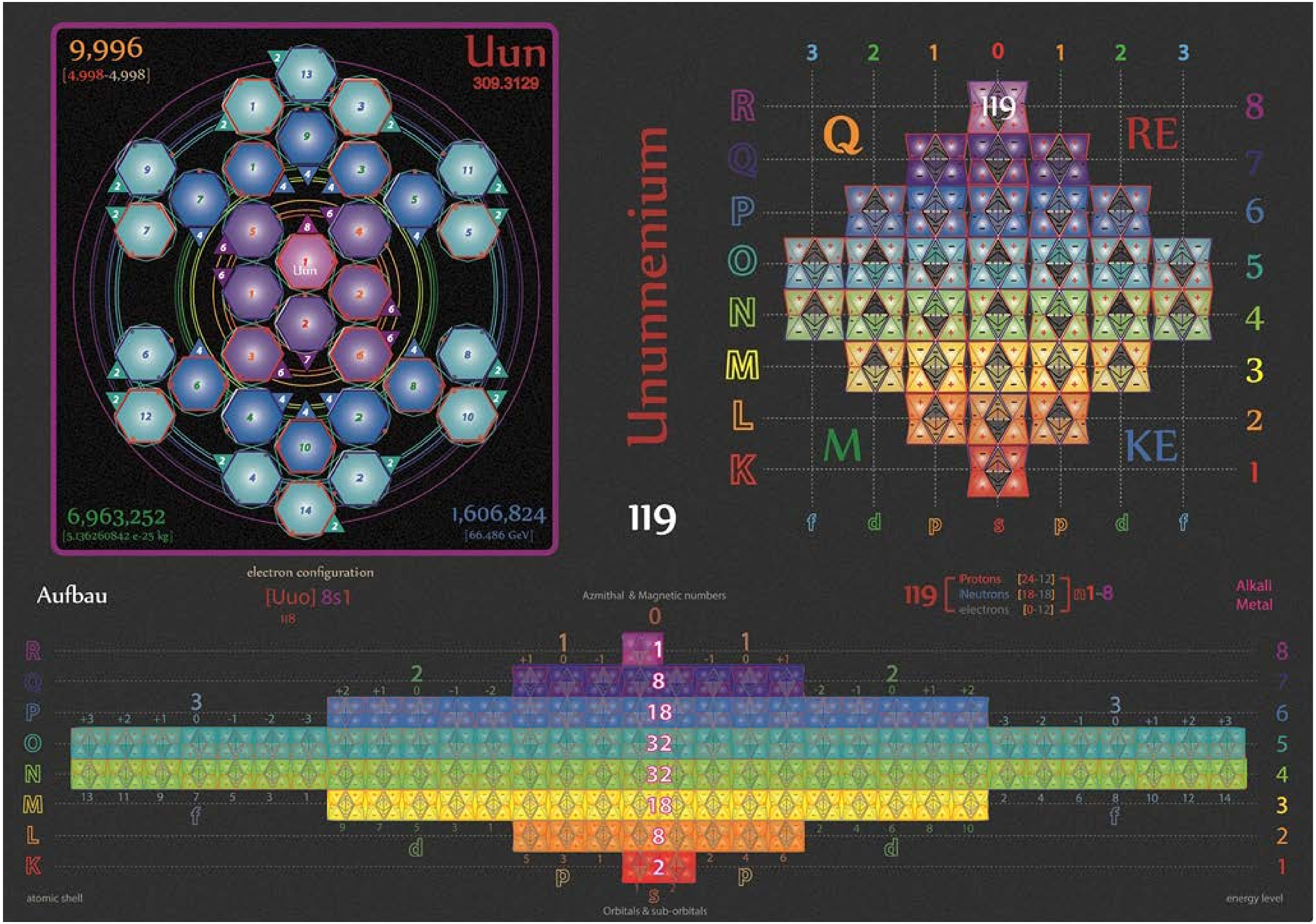
9,996
[4,998-4,998]
Uun
309.3129
Uun
6,963,252
[5.136260842 e-25 kg]
1,606,824
[66.486 GeV]
Ununnenium
119
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
119
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Uuo] 8s1
118
Azmithal & Magnetic numbers
119 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-8
Alkali Metal
1
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.119 - Ununnenium atom

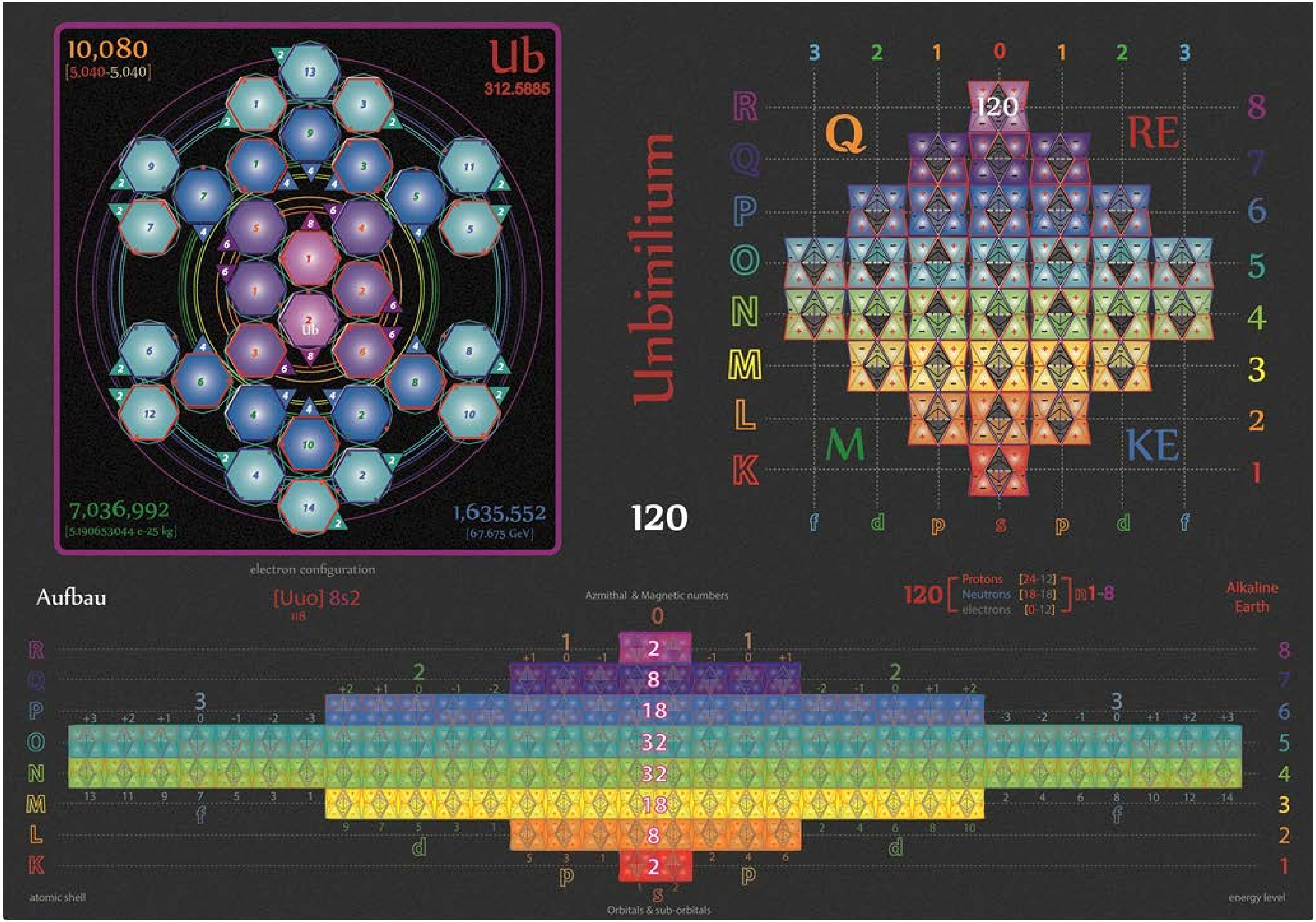
10,080
[5.040-5.040]
Ub
312.5885
7,036,992
[5.190653044 e-25 kg]
1,635,552
[67.675 GeV]
Unbinilium
120
3 2 1 0 1 2 3
R Q P O N M L K
8 7 6 5 4 3 2 1
Q
RE
M
KE
f d p s p d f
electron configuration
Aufbau
[Uuo] 8s2
118
Azmithal & Magnetic numbers
120 [Protons [24-12] Neutrons [18-18] electrons [0-12]] n1-8
Alkaline Earth
2
8
18
32
32
18
8
2
atomic shell
energy level
Orbitals & sub-orbitals

Tetryonics 51.120 - Unbinilium atom

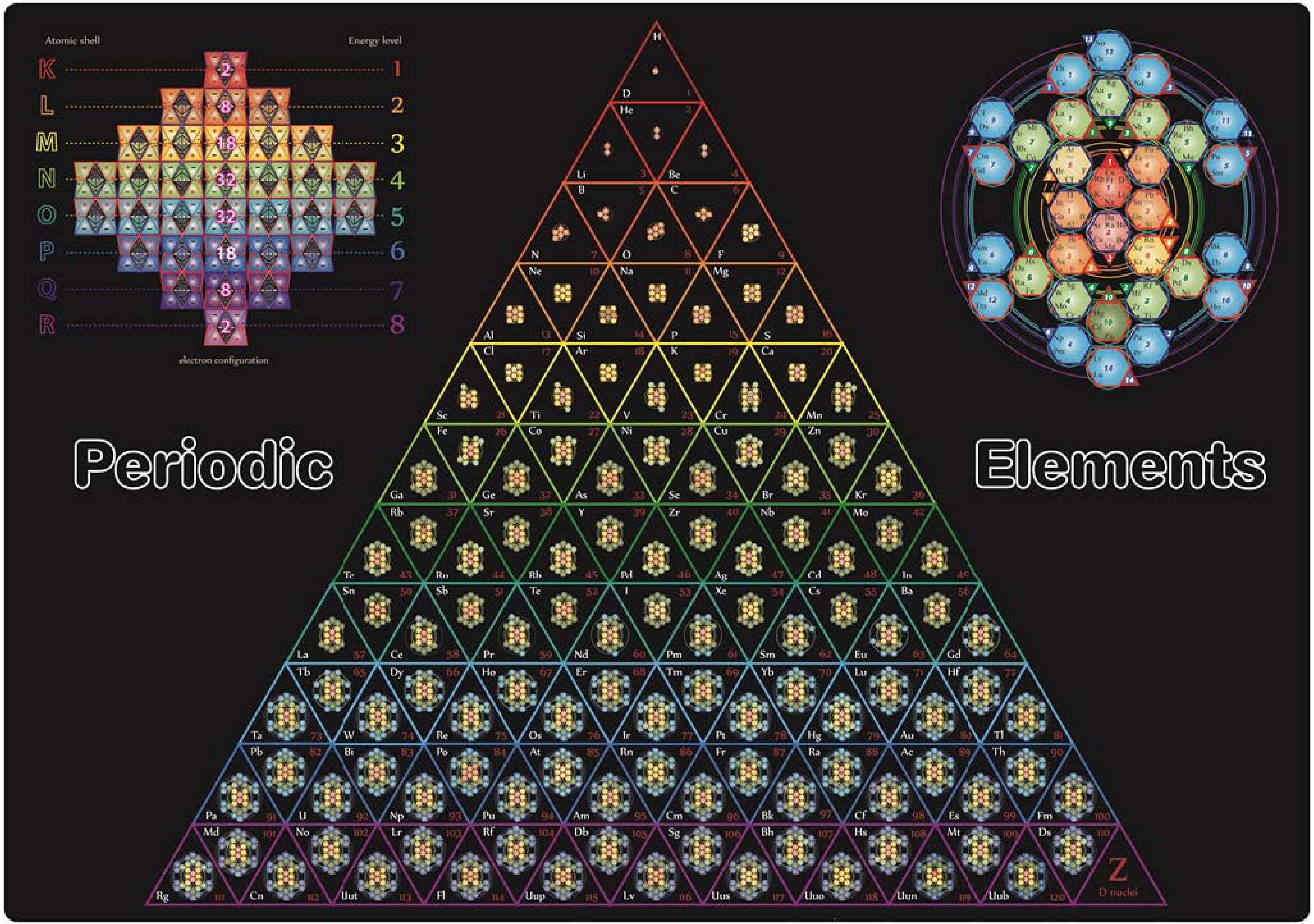
Atomic shell
Energy level
K 2 1
L 8 2
M 18 3
N 32 4
O 32 5
P 18 6
Q 8 7
R 2 8
electron configuration
Periodic
Elements
H
D 1
He 2
Li 3
Be 4
B 5
C 6
N 7
O 8
F 9
Ne 10
Na 11
Mg 12
Al 13
Si 14
P 15
S 16
Cl 17
Ar 18
K 19
Ca 20
Sc 21
Ti 22
V 23
Cr 24
Mn 25
Fe 26
Co 27
Ni 28
Cu 29
Zn 30
Ga 31
Ge 32
As 33
Se 34
Br 35
Kr 36
Rb 37
Sr 38
Y 39
Zr 40
Nb 41
Mo 42
Tc 43
Ru 44
Rh 45
Pd 46
Ag 47
Cd 48
In 49
Sn 50
Sb 51
Te 52
I 53
Xe 54
Cs 55
Ba 56
La 57
Ce 58
Pr 59
Nd 60
Pm 61
Sm 62
Eu 63
Gd 64
Tb 65
Dy 66
Ho 67
Er 68
Tm 69
Yb 70
Lu 71
Hf 72
Ta 73
W 74
Re 75
Os 76
Ir 77
Pt 78
Au 79
Hg 80
Tl 81
Pb 82
Bi 83
Po 84
At 85
Rn 86
Fr 87
Ra 88
Ac 89
Th 90
Pa 91
U 92
Np 93
Pu 94
Am 95
Cm 96
Bk 97
Cf 98
Es 99
Fm 100
Md 101
No 102
Lr 103
Rf 104
Db 105
Sg 106
Bh 107
Hs 108
Mt 109
Ds 110
Rg 111
Cn 112
Uut 113
Fl 114
Uup 115
Lv 116
Uus 117
Uuo 118
Uun 119
Uub 120
Z
D nuclei

Tetryonics 52.01 - Tetryonic Element Table

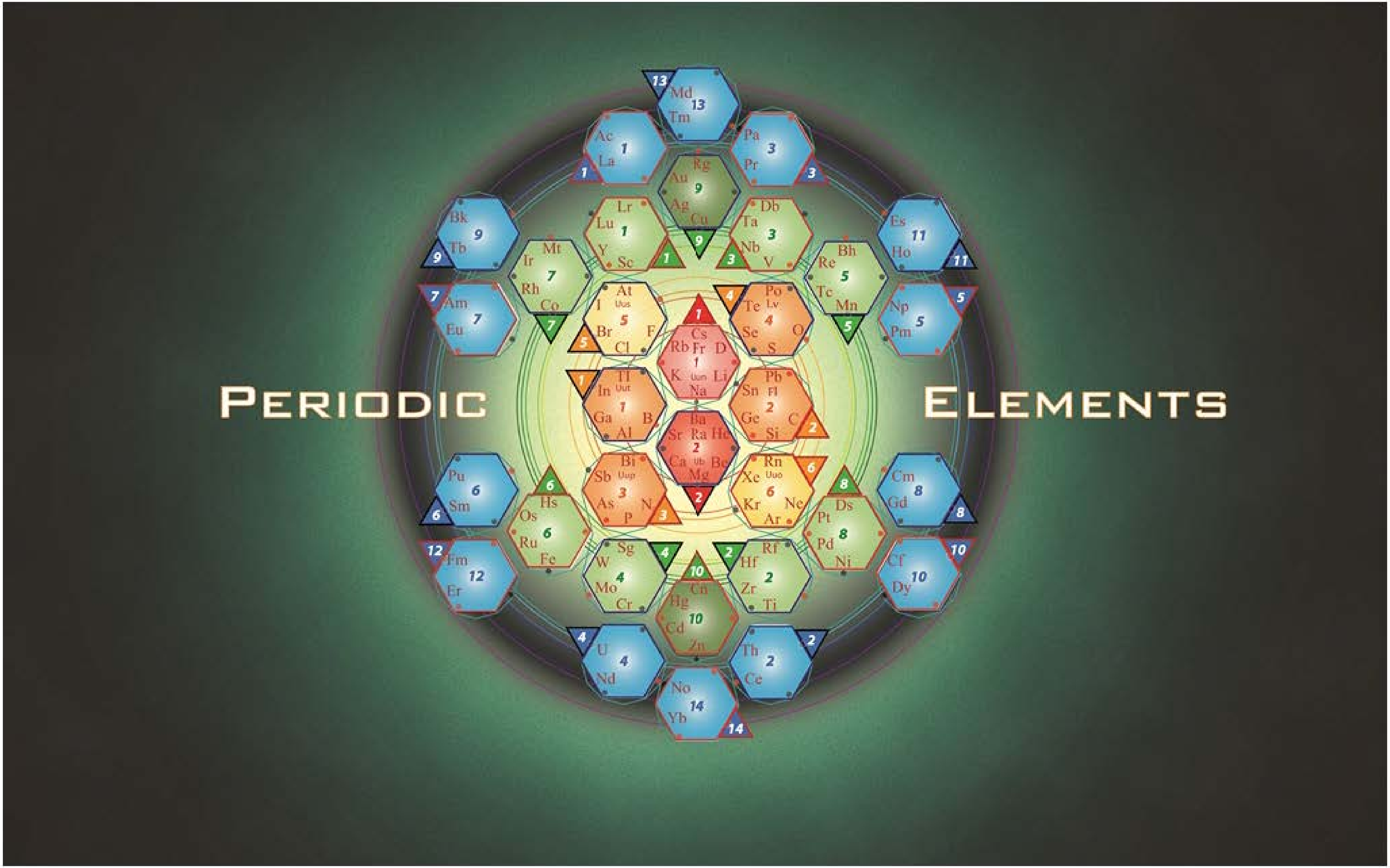
PERIODIC
ELEMENTS
13
Md 13 Tm
Ac 1 La
1
Pa 3 Pr
3
Au Rg 9 Ag Cu
9
Lr Lu 1 Y Sc
1
Db Ta 3 Nb V
3
Bk 9 Tb
9
Es 11 Ho
11
Mt Ir 7 Rh Co
7
Bh Re 5 Tc Mn
5
7
Am 7 Eu
5
Np 5 Pm
At I Uus 5 Br F Cl
5
1
Cs Rb Fr D 1 K Uun Li Na
4
Po Te Lv 4 Se O S
1
Tl In Uut 1 Ga B Al
Ba Sr Ra He 2 Ca Ub Be Mg
2
Pb Sn Fl 2 Ge C Si
2
Pu 6 Sm
6
Cm 8 Gd
8
6
Hs Os 6 Ru Fe
Bi Sb Uup 3 As N P
3
Rn Xe Uuo 6 Kr Ne Ar
6
8
Ds Pt 8 Pd Ni
12
Fm 12 Er
10
Cf 10 Dy
4
Sg W 4 Mo Cr
10
Cn Hg 10 Cd Zn
2
Rf Hf 2 Zr Ti
4
U 4 Nd
2
Th 2 Ce
No 14 Yb
14

Tetryonics 52.02 - Periodic Table [Elements]

Response to an analysis of Tetryonic Chemistry by Robert J Brotherus [July 2015]

In reply to a paper outlining Mr Robert Brotherus' examination and *mis*-understanding of Tetryonic theory and in particular how it applies to the modern understanding of Chemistry as it currently taught…..

Point by point I'll firstly address the particular points of concern raised wrt modern chemical theory and physics in general:

Tetryonics paper on quantum chemistry is very scarce on making any predictions about atomic properties or comparing such predictions to experimental results.

Tetryonic theory provides many predictions regarding the physical 2d planar **geometrie**s of bosons [radiant mass-energies] and the 3D material **Topologies** of Fermions [Matter and their associated physical properties]. Inclusive in these are many predictions and repeated statements and illustrations of specific particle shapes and the values of mass-energy momenta attributed to each… as well as how these values stem from the equilateral tessellation of Planck energy momenta and why they differ from the established values of modern scientific theory….

In many cases there are similarities to be found with existing observations but where Tetryonic theory differs from modern theory it is not by choice but rather dictated by the equilateral mass-energies within charged Matter topologies themselves and all too often led to much effort in order to reconcile the differences, leading to further insights and extensions of the theory over time…. Where differences exist it is deliberate not by error.

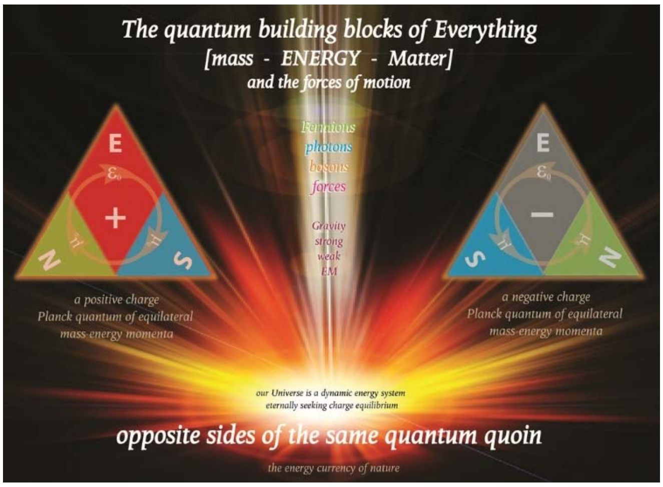
The quantum building blocks of Everything
[mass - ENERGY - Matter]
and the forces of motion
Fermions
photons
bosons
forces
Gravity
strong
weak
EM
E
ε_0
+
N
S
a positive charge
Planck quantum of equilateral
mass-energy momenta
E
ε_0
−
S
N
a negative charge
Planck quantum of equilateral
mass-energy momenta
our Universe is a dynamic energy system
eternally seeking charge equilibrium
opposite sides of the same quantum quoin
the energy currency of nature

Tetryonics is a BOTTOM-UP physical theory that unifies many currently disparate theories of science into a coherent unified model of the same – but not without its own set of mental challenges to those who are first becoming acquainted with it. As such it must be recognised that Tetryonic theory should not be studied without a sound understanding of the foundation upon with it is built…. Many have approached analysis of my work in such a way only to find that their analysis was limited by modern science's lack of precise detail of the foundations of the physics that they attempt to explain [to wit modern theory is a top-down explanation of the science build up over many generations with an equal number of varying interpretations and assumptions]

Tetryonics plays down importance of "mathematics" and instead emphasizes "geometry".

Arguably I place a lot of emphasis on geometry over that of mathematics in Tetryonic theory, but only so as it offers insights into physical theory that maths alone has not been able to provide since its wholesale incorporation into the science in the 1600's. This is not to downplay the importance and productiveness that maths has provided us a scientific tool, but rather to 'illustrate' and draw attention to the advances that a purely geometric model has to offer us in advancing our understanding of the same…. To wit, Tetryonic geometries are based on the postulate that all physical theory is founded on equilateral Planck quanta of energy momenta and as such these physical entities afford no room for speculative ideas or mathematical manipulations beyond what is physically possible to be constructed.

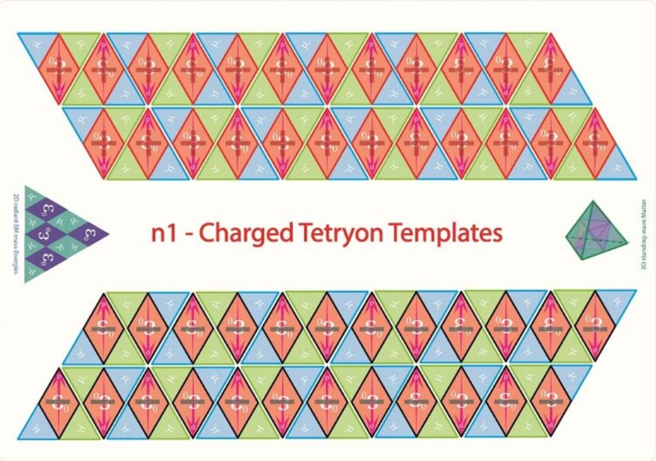

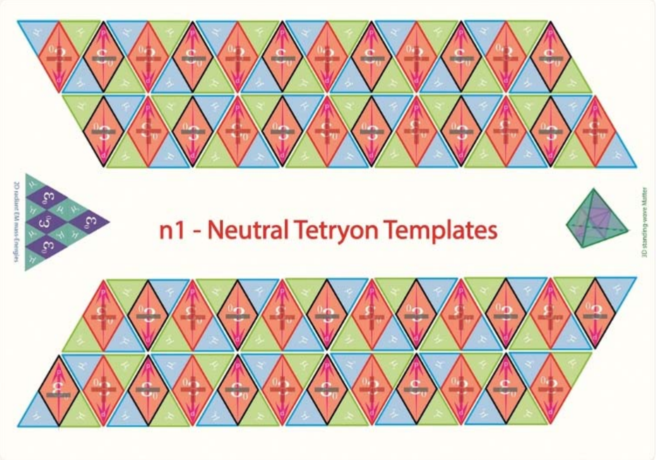

To help elucidate my work and thoughts on the differentiation of 2d planar mass-energy geometries from 3D Material topologies I recorded a detailed YouTube video titled "Tetryonics 101" where I show HOW to use the existing Tetryonic templates [of Planck mass-energy momenta] to construct all the fields [boson & photons] of the Standard model along with the 3D Mater topologies of all the known fermions [and them some]… I do hope you'll take the time to review this detailed instruction on the creation of quantum fields and Matter and try to replicate them same yourself – it is THE most informative thing students of Tetryonics and science in general can do when trying to increase their understanding of the hitherto unseen constituents that make up quantum theory and our macro-scale Universe and the laws of Nature we observe….

Tetryonics paper on physical chemistry is scarce on such numerical predictions and where it does make such predictions they are often in contradiction to observed experimental facts.

In addition to the maths found within Tetryonic illustrations there are also many spreadsheets developed from those very same equations over time to provide the numerical detail found in the illustrations – particular wrt periodic elements and theory rest atomic masses and kinetic elements etc. – It is by no means exhaustive on every topic imaginable but they do address the information provided on particles & elements etc. along with their kinetic and relativistic energies of motion.

Tetryonic Problem: Missing hydrogen element

Hydrogen is not missing from Tetryonic theory – in fact it is extensively detailed and modelled in Tetryonics [1] – Quantum Mechanics eBook [again highlight the need not to jump into one's preferred field of expertise and study but rather learn Tetryonic theory as you would any modern physical theory in order of presentation]

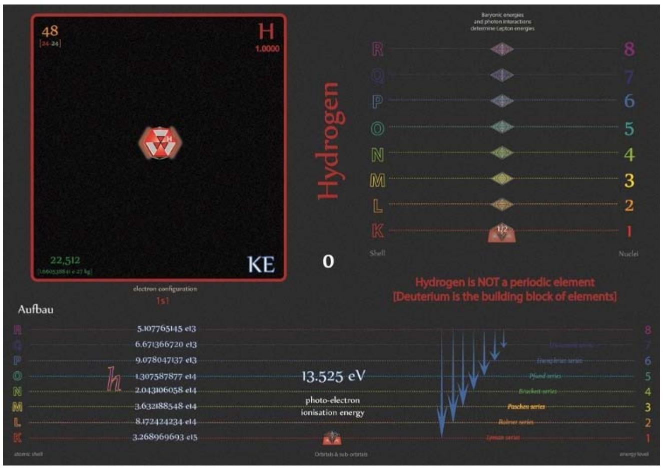
48
[24-24]
H
1.0000
22,512
KE
Hydrogen
0
Baryonic energies and photon interactions determine Lepton energies
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1/2
Shell
Nuclei
Hydrogen is NOT a periodic element
[Deuterium is the building block of elements]
electron configuration
1s1
Aufbau
5.107765145 e13
6.671366720 e13
9.078047137 e13
1.307587877 e14
2.043106058 e14
3.632188548 e14
8.172424234 e14
3.268969693 e15
h
13.525 eV
photo-electron
ionisation energy
Pfund series
Brackett series
Paschen series
Balmer series
Lyman series
atomic shell
Orbitals & sub-orbitals
energy level

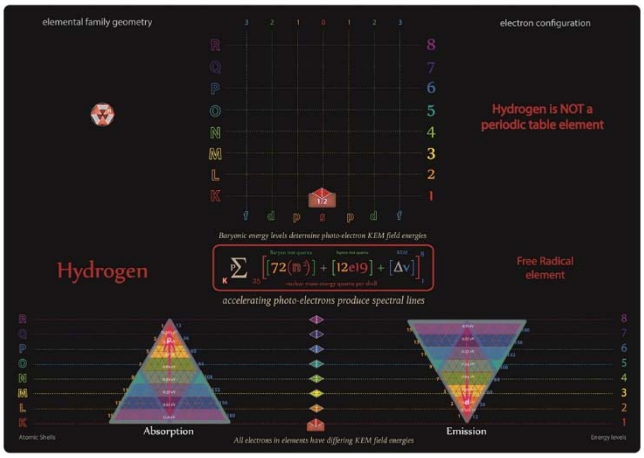
elemental family geometry
electron configuration
3 2 1 0 1 2 3
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
1/2
f d p s p d f
Hydrogen is NOT a periodic table element
Baryonic energy levels determine photo-electron KEM field energies
Hydrogen
$\sum_{K}^{P}\,_{25}\left[[72(n^2)] + [12e19] + [\Delta v] \right]_1^8$
Free Radical element
accelerating photo-electrons produce spectral lines
Absorption
Emission
Atomic Shells
Energy levels
All electrons in elements have differing KEM field energies

Hydrogen is designated as element 00 in Tetryonic theory as it is **NOT** the building block of periodic elements as is currently understood but rather a sub-unit of Deuterium the real building block of all elements…. This is not an arbitrary decision but one forced on anybody who attempts to build physical models of all elements in order to match and explain many of the current established aspects of chemical and quantum theory wrt the quantum electrodynamics of elements themselves wrt molar mass, the Aufbau principle, Schrodinger and Bohr numbers for each element and how Deuterium nuclei [or any other] may bind together in order to form all the elements of the periodic table and their respective spectral emissions lines.

Attempting to build heavy elements utilising Z number Hydrogen nuclei as a foundation with the excess molar masses being accounted for by way of 'extra or excess' Neutrons [as is the established model of modern chemistry] runs into a number of problems immediately – namely:

The lack of residual strong Force binding points on nuclei to facilitate the binding of Neutrons to the nuclei in order to form the elements [particularly elements 60+]

The incorrectness of ‘reserve beta decay’ to explain and model Neutron particle formation and the mechanics of electron binding to periodic elements in accordance with Schrodinger’s wave equation and Bohr’s atomic model of the same in order to produce observed spectral line emissions of each element in the PT.

There is no uncertainty or indeterminacy in Tetryonic theory as there is in modern theory to explain away such inconsistencies between various aspects of physical theory – each attribute MUST be modelled and explained from first principles & invariably leads to new insights into quantum chemical theory that are at odds with established lines of thought.

To wit I stand by the claims and statements made in my eBooks :

> Each elemental nuclei is made from Deuterium" (p. 12)
> “120 elements in 8 quantum levels *and hydrogen as a free radical*.” (p. 26)
> “Deuterium (not hydrogen) is the building block of all elements." (p. 21)
> “Hydrogen is NOT a periodic element” (p. 83)

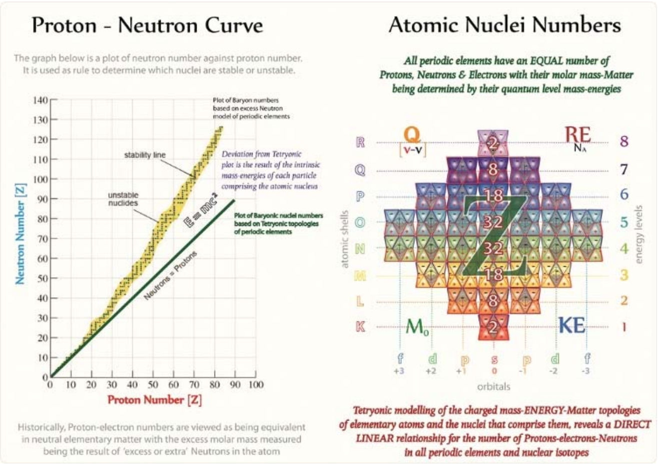

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number. It is used as rule to determine which nuclei are stable or unstable.

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of ‘excess or extra’ Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

As it is the ONLY way to model and explain how all the periodic elements are created from equilateral Planck energy momenta that is in accord with the known observables… they are not in agreement with some aspects of established theory BUT they do closely mirror the observables and once KEM fields of thermal heat, kinetic motion etc. are taken into account provide a very accurate quantum model of chemical processes at the quantum scale.

Tetryonics does not throw all these facts out of the window just because hydrogen does not fit neatly into my [or other] pre-conceived set of boxes and triangles and hexagons, they are calculated from the Material models and immaterial fields associated with each element that MUST grow from first principles… not from assumptions elaborated upon over time.

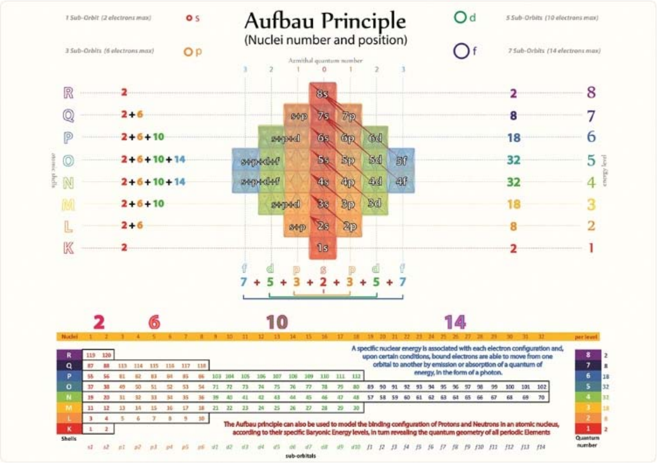
Aufbau Principle
(Nuclei number and position)
1 Sub-Orbit (2 electrons max)
3 Sub-Orbits (6 electrons max)
5 Sub-Orbits (10 electrons max)
7 Sub-Orbits (14 electrons max)
s
p
d
f
R
Q
P
O
N
M
L
K
2
2+6
2+6+10
2+6+10+14
2+6+10+14
2+6+10
2+6
2
8s
7s 7p
6s 6p 6d
5s 5p 5d 5f
4s 4p 4d 4f
3s 3p 3d
2s 2p
1s
2
8
18
32
32
18
8
2
8
7
6
5
4
3
2
1
7 + 5 + 3 + 2 + 3 + 5 + 7
2
6
10
14
A specific nuclear energy is associated with each electron configuration and, upon certain conditions, bound electrons are able to move from one orbital to another by emission or absorption of a quantum of energy, in the form of a photon.
The Aufbau principle can also be used to model the binding configuration of Protons and Neutrons in an atomic nucleus, according to their specific Baryonic Energy levels, in turn revealing the quantum geometry of all periodic Elements
Shells
sub-orbitals
Quantum number

In fact the Aufbau table is shown in Tetryonic theory to be a table that has been 'right justified' to match the nomenclature of mathematics rather than the geometric topology of reality [which is centred about the principle quantum numbers and shells of Schrodinger and Bohr respectively]… it is only our skewed perspective of maths that prevents most from seeing the real symmetry present in atomic nuclei as detailed in the quantum numbers of Schrodinger's wave equation for the same … ie there are no g,h,I,j, shells as supposed and all the quantum numbers of Schrodinger's wave equations match the symmetric topology of elements as geometric portrayed in Tetryonics [and are in agreeance with Bohr's shells and orbitals as well, thus unifying all 3 theories]

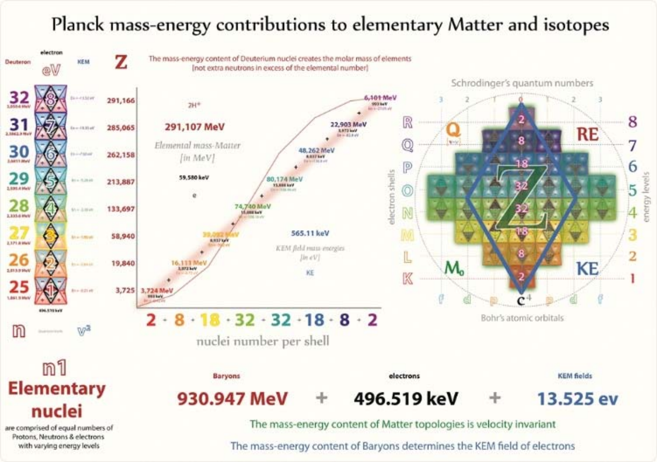
Planck mass-energy contributions to elementary Matter and isotopes
The mass-energy content of Deuterium nuclei creates the molar mass of elements [not extra neutrons in excess of the elemental number]
Deuteron
electron
eV
KEM
Z
32
31
30
29
28
27
26
25
291,166
285,065
262,158
213,887
133,697
58,940
19,840
3,725
291,107 MeV
Elemental mass-Matter [in MeV]
565.11 keV
KEM field mass energies [in eV]
2 · 8 · 18 · 32 · 32 · 18 · 8 · 2
nuclei number per shell
Schrodinger's quantum numbers
RE
KE
Mo
c⁴
electron shells
energy levels
Bohr's atomic orbitals
n1
Elementary nuclei
are comprised of equal numbers of Protons, Neutrons & electrons with varying energy levels
Baryons
930.947 MeV
electrons
496.519 keV
KEM fields
13.525 ev
The mass-energy content of Matter topologies is velocity invariant
The mass-energy content of Baryons determines the KEM field of electrons

The material topologies of Tetryonic theory then also lead to the necessity to correct 'weighted' values for all known particles and elements [and then some] in light of the quantum mechanics of the same – ie their identical 3D charge topologies and the mechanics of the quantum rotating synchronous convertors [Deuterium nuclei] that they form in deference to the accepted theory of mini-solar systems and Neutrons that are heavier than their identical charged counterparts [Protons and anti-Protons]

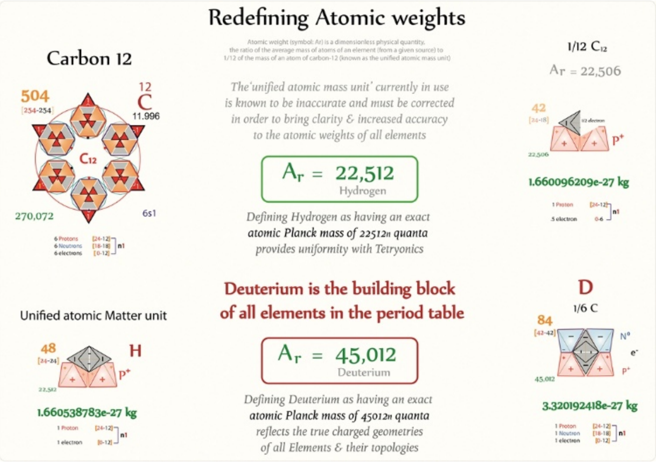

Redefining Atomic weights

Atomic weight (symbol: Ar) is a dimensionless physical quantity, the ratio of the average mass of atoms of an element (from a given source) to 1/12 of the mass of an atom of carbon-12 (known as the unified atomic mass unit)

The 'unified atomic mass unit' currently in use is known to be inaccurate and must be corrected in order to bring clarity & increased accuracy to the atomic weights of all elements

$A_r = 22{,}512$ Hydrogen

Defining Hydrogen as having an exact atomic Planck mass of 22512n quanta provides uniformity with Tetryonics

Deuterium is the building block of all elements in the period table

$A_r = 45{,}012$ Deuterium

Defining Deuterium as having an exact atomic Planck mass of 45012n quanta reflects the true charged geometries of all Elements & their topologies

Carbon 12
504
[254-254]
12
C
11.996
C12
270,072
6s1
6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1
1/12 C12
Ar = 22,506
42
[24-18]
1/2 electron
P+
22,506
1.660096209e-27 kg
1 Proton [24-12]
.5 electron 0-6
n1
Unified atomic Matter unit
48
[24-24]
H
P+
22,512
1.660538783e-27 kg
1 Proton [24-12]
1 electron [0-12]
n1
D
1/6 C
84
[42-42]
N0
e-
P+
45,012
3.320192418e-27 kg
1 Proton [24-12]
1 Neutron [18-18]
1 electron [0-12]
n1

Point in case being **Tetryonic Problem: Incorrect Proton and Neutron masses**
As detailed in illustration 49.04 & 49.05 respectively in my eBook… Tetryonics demands an exact mass for Planck quanta and the molar masses [at absolute zero] for each and every particle in our Universe for modelling and any additional energy momenta [by way of thermal, spectral, kinetic or observational energies etc.] are accounted for in the KEM fields of each respective element or particle.

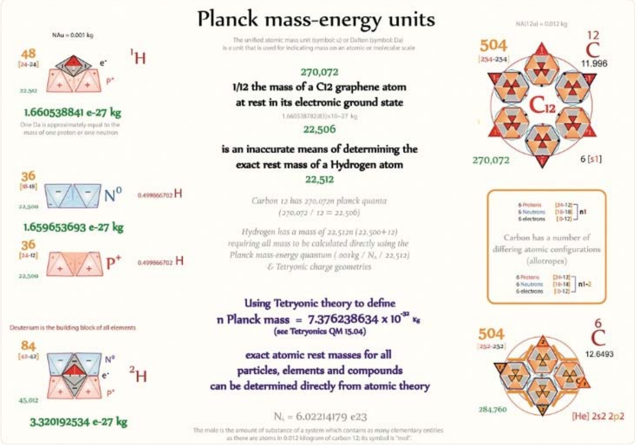

Planck mass-energy units

The unified atomic mass unit (symbol: u) or Dalton (symbol: Da) is a unit that is used for indicating mass on an atomic or molecular scale

270,072
1/12 the mass of a C12 graphene atom at rest in its electronic ground state
1.660538782(83)x10−27 kg
22,506
is an inaccurate means of determining the exact rest mass of a Hydrogen atom
22,512

Carbon 12 has 270,072n planck quanta (270,072 / 12 = 22,506)

Hydrogen has a mass of 22,512n (22,500+12) requiring all mass to be calculated directly using the Planck mass-energy quantum (.001kg / N_A / 22,512) & Tetryonic charge geometries

Using Tetryonic theory to define
n Planck mass = $7.376238634 \times 10^{-32}$ Kg
(see Tetryonics QM 15.04)

exact atomic rest masses for all particles, elements and compounds can be determined directly from atomic theory

$N_A = 6.02214179$ e23

The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12; its symbol is "mol".

NAu = 0.001 kg
48
[24-24]
e-
1H
P+
22,512
1.660538841 e-27 kg
One Da is approximately equal to the mass of one proton or one neutron
36
[18-18]
N0
0.499866702 H
22,500
1.659653693 e-27 kg
36
[24-12]
P+
0.499866702 H
22,500
Deuterium is the building block of all elements
84
[42-42]
N0
e-
2H
P+
45,012
3.320192534 e-27 kg
NA(12u) = 0.012 kg
504
[254-254]
12
C
11.996
C12
270,072
6 [s1]
6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1
Carbon has a number of differing atomic configurations (allotropes)
6 Protons [24-12]
6 Neutrons [18-18]
6 electrons [0-12]
n1-2
504
[252-252]
6
C
12.6493
284,760
[He] 2s2 2p2

There is NO room for a 'weighted' atomic mass as utilised in modern chemistry which conveniently ignores the mass and motional energies of the electrons in the elements or which ignores the KEM field energies noted above for 'most' situations and calculations – Tetryonic theory MUST account for all of these energies even if it does so at the risk of disagreeing with modern theory at times…

The Proton and Neutron MUST have identical atomic masses as they are identical – both are 36π Baryon Matter topologies differing only in the net charges of +12 [24-12] vs 0 [18-18] they possess as a result of their constituent quarks [UDU vs DUD] as extensively detailed in my QM eBook… Neutrons are NOT created through the process of 'reverse' beta-decay as is supposed presently supposed in modern theory and accordingly do NOT have a slightly higher mass than Protons which in turn affects the molar mass calculations for all periodic elements and isotopes as outlined in my Chemistry eBook as well at length.

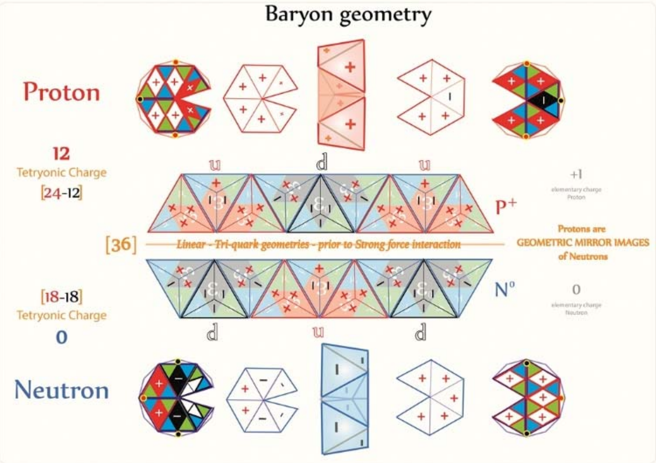
Baryon geometry
Proton
12
Tetryonic Charge
[24-12]
u
d
u
+1
elementary charge
Proton
P+
[36]
Linear - Tri-quark geometries - prior to Strong force interaction
Protons are
GEOMETRIC MIRROR IMAGES
of Neutrons
[18-18]
Tetryonic Charge
0
N°
0
elementary charge
Neutron
d
u
d
Neutron

Ie Carbon 13, 14 and upwards are not the result of ‘extra’ Neutrons being bound to the Carbon nuclei – such a nuclei if possible would be stable and not decay as observed – but rather they are simply the result of the inherent ‘squared’ equilateral mass-energies of each Baryon in the nuclei being increased from n1 to n2 to n3 …. Resulting in a net mass increase that ‘appears’ to be the result of a number of ‘extra’ Neutrons in the nuclei [again this is detailed in my Chemistry eBook and spreadsheets of the calculations be easily created from the formulae found in chapter 53 on my Chemistry eBook….

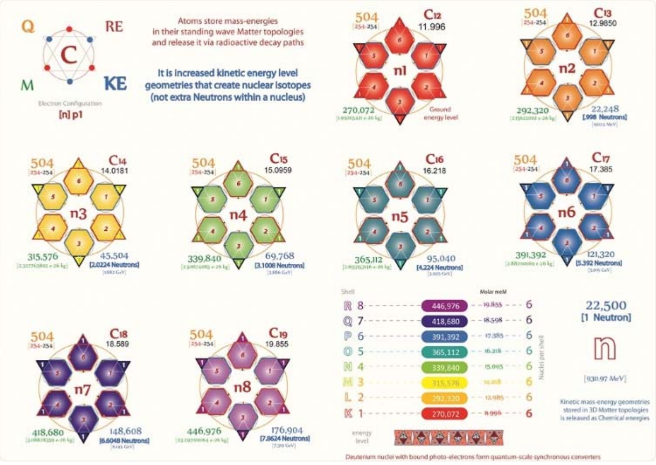
Q
RE
C
M
KE
Electron Configuration
[n] p1
Atoms store mass-energies
in their standing wave Matter topologies
and release it via radioactive decay paths
It is increased kinetic energy level
geometries that create nuclear isotopes
(not extra Neutrons within a nucleus)
504
[254-254]
C12
11.996
n1
270,072
Ground
energy level
504
[254-254]
C13
12.9850
n2
292,320
22,248
[.998 Neutrons]
504
[254-254]
C14
14.0181
n3
315,576
45,504
[2.0224 Neutrons]
504
[254-254]
C15
15.0959
n4
339,840
69,768
[3.1008 Neutrons]
504
[254-254]
C16
16.218
n5
365,112
95,040
[4.224 Neutrons]
504
[254-254]
C17
17.385
n6
391,392
121,320
[5.392 Neutrons]
504
[254-254]
C18
18.589
n7
418,680
148,608
[6.6048 Neutrons]
504
[254-254]
C19
19.855
n8
446,976
176,904
[7.8624 Neutrons]
Shell
Molar mass
R 8 446,976 19.855 6
Q 7 418,680 18.598 6
P 6 391,392 17.385 6
O 5 365,112 16.218 6
N 4 339,840 15.095 6
M 3 315,576 6
L 2 292,320 12.985 6
K 1 270,072 11.996 6
Nuclei per shell
energy level
22,500
[1 Neutron]
n
[930.97 MeV]
Kinetic mass-energy geometries
stored in 3D Matter topologies
is released as Chemical energies
Deuterium nuclei with bound photo-electrons form quantum-scale synchronous converters

This ‘significant discrepancy’ with Tetryonics prediction of Proton and Neutron masses from observed ones alone will be enough to prove Tetryonic theory a worthy successor to modern quantum chemistry theory – unless you can explain all of the above using ‘excess’ Neutrons.

I am neither ignorant of the basic masses of basic particles nor do I think that Tetryonic theory and observed facts of nature disagree, but rather the established theoretical ‘facts’ of modern chemical theory must be in error

I agree that in reality Matter has **3D structure** (including **volume**) hence the terminology of Matter **topologies** throughout my work. To portray my planar illustration of elements and particles in my PDF eBooks as not reflecting their true material topologies is grossly inaccurate and reflects that fact that **YOU have NOT read all my work** including Tetryonics 101, the Google sketch warehouse of 3D CAD models of each and every particle [kindly developed by Rene Cormier over many months of effort] and extensive list of all elements and their charged 3D topologies and electronic wave function properties and numbers which are freely available to all – only planar 2 d mass-energy **geometries** [bosons and photons] are portrayed as being like thin 2D sheets in Tetryonic theory, something I am sure you will become more familiar with as your understanding of T-theory advances,[and your distain for its predictions and values not in accord with modern theory wains over time].

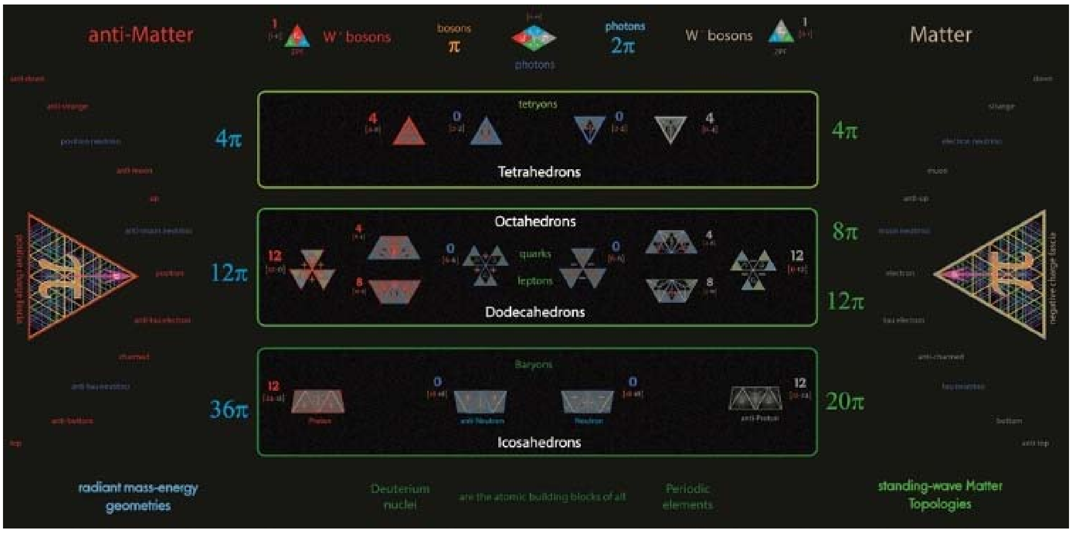
anti-Matter
W⁺ bosons
bosons
π
photons
photons
2π
W⁻ bosons
Matter
4π
tetryons
4
0
0
4
Tetrahedrons
4π
Octahedrons
8π
12π
12
4
8
0
quarks
leptons
0
4
8
12
Dodecahedrons
12π
Baryons
36π
12
0
0
12
Icosahedrons
20π
radiant mass-energy geometries
Deuterium nuclei
are the atomic building blocks of all
Periodic elements
standing-wave Matter Topologies

Regarding the ‘separation ‘ of electrons from the atomic nucleus – this is an antiquated model of quantum electrodynamics that is fashioned on planets and their satellites [ie the mini-solar system Bohr model] and which has been extended to a fuzzy electron orbital model to incorporate the Schrodinger quantum model and the ‘inherent’ quantum uncertainly/indeterminacy of statistical probabilities intrinsic to quantum theory as the theories developed over time…. Tetryonics makes it quite clear in T[1] – QM that the atomic nucleus is in fact a quantum synchronous convertor [ala Tesla/Westinghouse] that can store and release mass-energies in a controlled and predictable manner [ala isotopic decay and spectral line emissions] and that quantum mechanics is NOT quantum inherently uncertain nor in determinant as it is often portrayed – in fact quantum theory as explained by Tetryonic theory is very much deterministic in every way – it was only our top-down view of the mechanics that led modern science to this erroneous conclusion, which I have attempted to correct through Tetryonic theory.

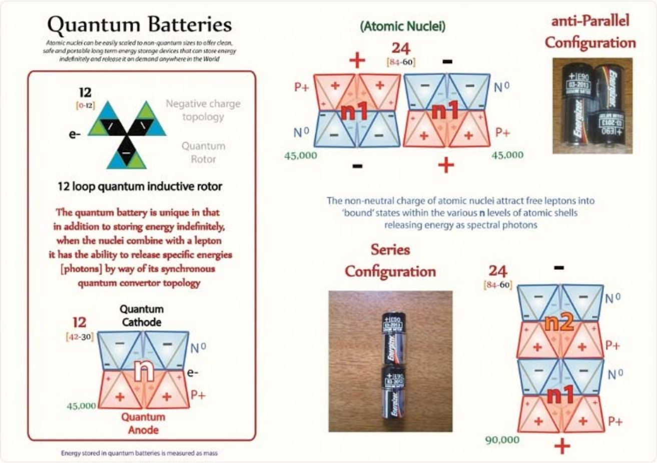
Quantum Batteries

Atomic nuclei can be easily scaled to non-quantum sizes to offer clean, safe and portable long term energy storage devices that can store energy indefinitely and release it on demand anywhere in the World

12 loop quantum inductive rotor

The quantum battery is unique in that in addition to storing energy indefinitely, when the nuclei combine with a lepton it has the ability to release specific energies [photons] by way of its synchronous quantum convertor topology

The non-neutral charge of atomic nuclei attract free leptons into ‘bound’ states within the various **n** levels of atomic shells releasing energy as spectral photons

Energy stored in quantum batteries is measured as mass
12
[0-12]
e-
Negative charge topology
Quantum Rotor
12
[42-30]
Quantum Cathode
N0
e-
P+
n
45,000
Quantum Anode
(Atomic Nuclei)
24
[84-60]
+
-
P+
N0
n1
n1
N0
P+
45,000
45,000
-
+
anti-Parallel Configuration
Series Configuration
24
[84-60]
-
N0
n2
P+
N0
n1
P+
90,000
+

The electron does not orbit the atomic nucleus of any element but rather is bound via its magnetic dipole moment to the positively charged nucleus and spins akin to the rotating rotor of quantum convertor partially within the 3D Matter topology of Deuterium nuclei….

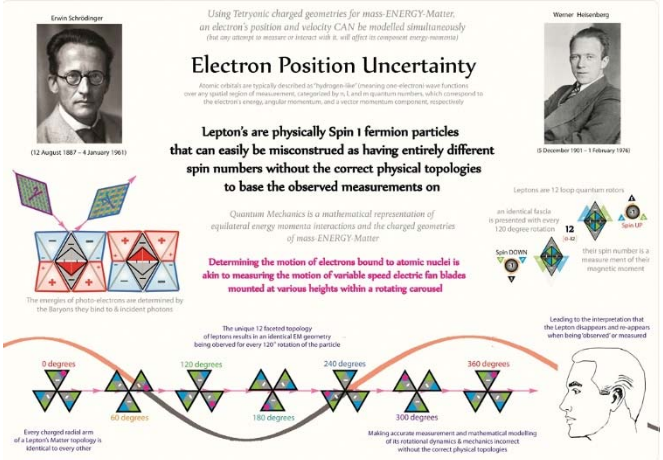

Using Tetryonic charged geometries for mass-ENERGY-Matter, an electron's position and velocity CAN be modelled simultaneously
(but any attempt to measure or interact with it, will affect its component energy-momenta)

Electron Position Uncertainty

Atomic orbitals are typically described as "hydrogen-like" (meaning one-electron) wave functions over any spatial region of measurement, categorized by n, l, and m quantum numbers, which correspond to the electron's energy, angular momentum, and a vector momentum component, respectively

Lepton's are physically Spin 1 fermion particles that can easily be misconstrued as having entirely different spin numbers without the correct physical topologies to base the observed measurements on

Quantum Mechanics is a mathematical representation of equilateral energy momenta interactions and the charged geometries of mass-ENERGY-Matter

Determining the motion of electrons bound to atomic nuclei is akin to measuring the motion of variable speed electric fan blades mounted at various heights within a rotating carousel

The energies of photo-electrons are determined by the Baryons they bind to & incident photons

Leptons are 12 loop quantum rotors

an identical fascia is presented with every 120 degree rotation

their spin number is a measure ment of their magnetic moment

The unique 12 faceted topology of leptons results in an identical EM geometry being oberved for every 120° rotation of the particle

Leading to the interpretation that the Lepton disappears and re-appears when being 'observed' or measured

Every charged radial arm of a Lepton's Matter topology is identical to every other

Making accurate measurement and mathematical modelling of its rotational dynamics & mechanics incorrect without the correct physical topologies

Erwin Schrödinger
(12 August 1887 – 4 January 1961)
Werner Heisenberg
(5 December 1901 – 1 February 1976)
12
(0-12)
Spin UP
Spin DOWN
0 degrees
60 degrees
120 degrees
180 degrees
240 degrees
300 degrees
360 degrees

Again these models have all been built and operate according to established electrodynamic theory and their operation is well understood in the macro sense [but obviously the acceptance of such a model strains the thoughts of most who have grown up a mental picture of fuzzy electrons orbiting a spherical nuclei]

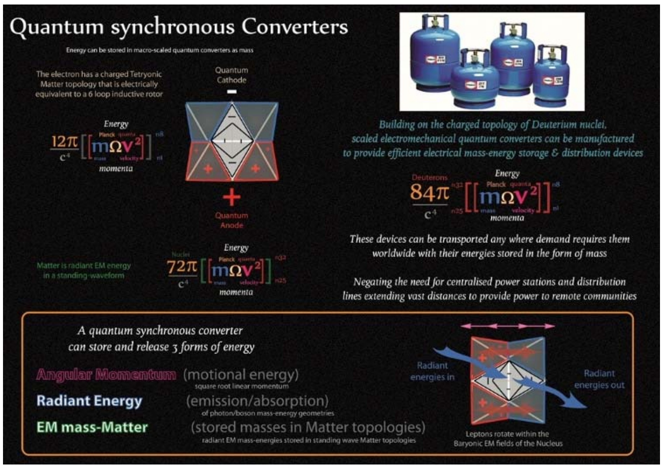

Quantum synchronous Converters

Energy can be stored in macro-scaled quantum converters as mass

The electron has a charged Tetryonic Matter topology that is electrically equivalent to a 6 loop inductive rotor

Matter is radiant EM energy in a standing-waveform

Building on the charged topology of Deuterium nuclei, scaled electromechanical quantum converters can be manufactured to provide efficient electrical mass-energy storage & distribution devices

These devices can be transported any where demand requires them worldwide with their energies stored in the form of mass

Negating the need for centralised power stations and distribution lines extending vast distances to provide power to remote communities

A quantum synchronous converter can store and release 3 forms of energy

Angular Momentum (motional energy)
square root linear momentum

Radiant Energy (emission/absorption)
of photon/boson mass-energy geometries

EM mass-Matter (stored masses in Matter topologies)
radiant EM mass-energies stored in standing wave Matter topologies

Quantum Cathode
Quantum Anode
Energy
$\frac{12\pi}{c^4}[[m\Omega v^2]]$
momenta
Nuclei
$\frac{72\pi}{c^4}[[m\Omega v^2]]$
Deuterons
$\frac{84\pi}{c^4}[[m\Omega v^2]]$
Planck quanta
mass velocity
Radiant energies in
Radiant energies out
Leptons rotate within the Baryonic EM fields of the Nucleus

Again the electronic structure of each and every one of the 120 possible periodic elements is well detailed wrt Bohr shells & orbitals and unified wrt their respective Schrodinger numbers for the same in my Chemistry eBook for all to view…. Unlike modern texts which normally stop at element 20 and claim that it is just too difficult to show the quantum models of larger elements with their evermore complex electron configurations and Schrodinger numbers

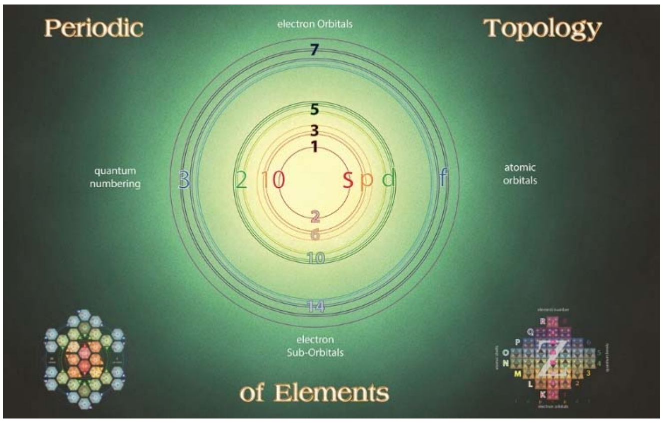

Periodic
Topology
of Elements
electron Orbitals
electron Sub-Orbitals
quantum numbering
atomic orbitals
7
5
3
1
2
6
10
14
3
2
1
0
s
p
d
f

The s.p.d & f orbitals of electrons 'around' any elementary nuclei is the result of the atomic nuclei rotating about is centre of mass/gravity [classical angular momentum] and the electrons tracing out easily determined and well defined concentric orbits about the same centre [as illustrated] …. A close examination of the Schrodinger numbers [or in fact Bohr orbits] of heavier/larger Z number atomic nuclei reveals the there are no g,h I orbits etc as often postulated but in fact there is a reversal of electron orbital distance as the element number and principal energy levels increases….. Resulting in the periodic symmetry depicted in Tetryonic theory of an atomic nucleus resembling a child's spinning top, which could be mistaken for a spherical topology from our macro view of the quantum scale.

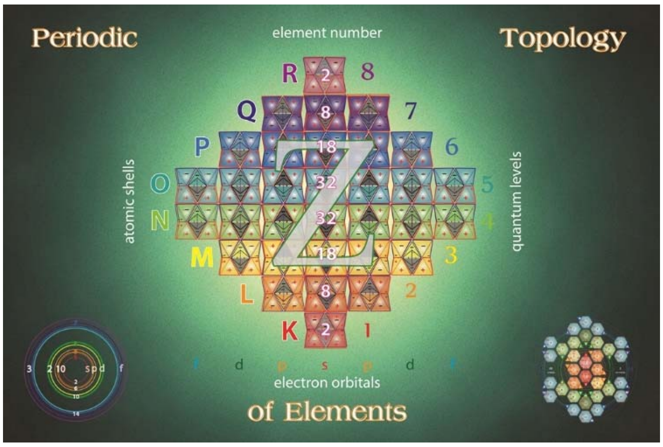
Periodic
Topology
of Elements
element number
atomic shells
quantum levels
electron orbitals
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
d
s
d
3
2 10
s p d
f
2
6
10
14

With each element's Z number reflecting the number of **Deuterium** atoms in each nuclei, *not the number of Hydrogen atoms with an ever increasing number of Neutrons to account for observed molar masses*

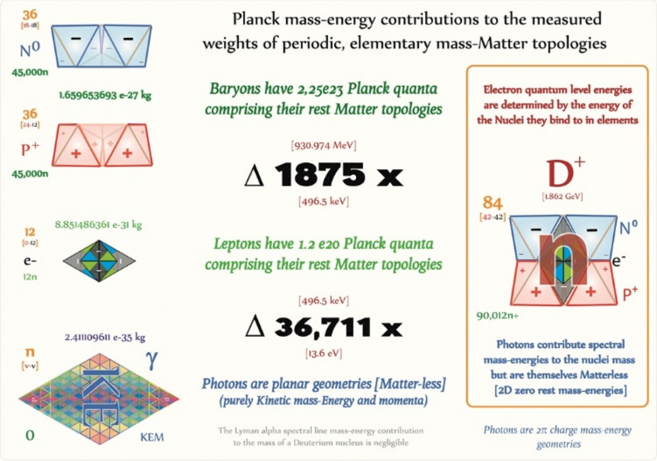
Planck mass-energy contributions to the measured weights of periodic, elementary mass-Matter topologies
36
[18-18]
N0
45,000n
1.659653693 e-27 kg
36
[24-12]
P+
45,000n
8.851486361 e-31 kg
12
[0-12]
e-
12n
2.4111096ll e-35 kg
n
[v-v]
γ
KE
0
KEM
Baryons have 2,25e23 Planck quanta comprising their rest Matter topologies
[930.974 MeV]
Δ 1875 x
[496.5 keV]
Leptons have 1.2 e20 Planck quanta comprising their rest Matter topologies
[496.5 keV]
Δ 36,711 x
[13.6 eV]
Photons are planar geometries [Matter-less] (purely Kinetic mass-Energy and momenta)
The Lyman alpha spectral line mass-energy contribution to the mass of a Deuterium nucleus is negligible
Electron quantum level energies are determined by the energy of the Nuclei they bind to in elements
D+
[1.862 GeV]
84
[42-42]
N0
e-
P+
n
90,012n+
Photons contribute spectral mass-energies to the nuclei mass but are themselves Matterless [2D zero rest mass-energies]
Photons are 2π charge mass-energy geometries

The electrons associated with each atomic element's structure in fact have an important and direct bearing on the spectral emissions of each particular element. Even though their mass-energies of their electronic topologies are only 0.000533' % of the mass-energies of each Baryon in any atomic nuclei this mass-Matter differential along with the energy momenta of the electron's KEM field is key to facilitating the absorption and release of spectral line emissions particular to each and every element known… again, it is not simple good enough to conveniently ignore electrons ad hoc in scientific theory to suit a particular point or theory just because they are or 'insignificant' mass compared to the whole nucleus. Any successful model of science and physics in general must be able to account for ALL the known aspects of modern observations not just a select few that suit their purposes in order to unify physics on all energy scales. Ie a new model of the quantum structure, topology and mechanics of atomic nuclei must account for charge, atomic and chemical interactions, spectral line series and all known elements and compound creation in the ONE physical model.

Clarifying question about Tetryonics nuclear vs. atomic size wrt the Zenon atom
I humbly submit that illustrations 51.49

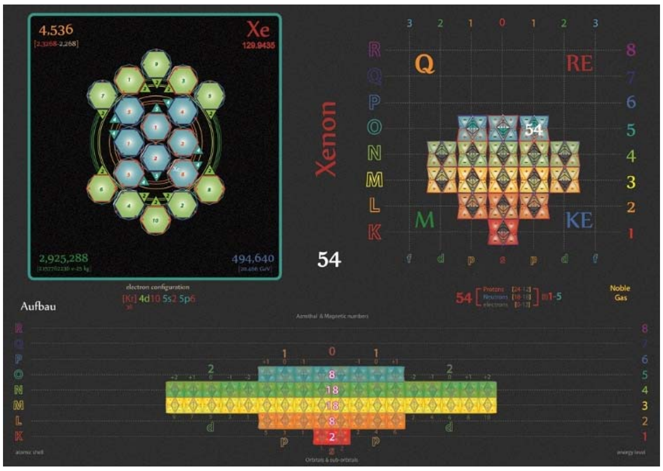
4.536
Xe
Xenon
Q
RE
M
KE
54
2,925,288
494,640
Aufbau
electron configuration
[Kr] 4d10 5s2 5p6
Noble Gas
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
f
d
p
s
Orbitals & sub-orbitals

And 53.49 respectively along with the 3D CAD illustrations of the same

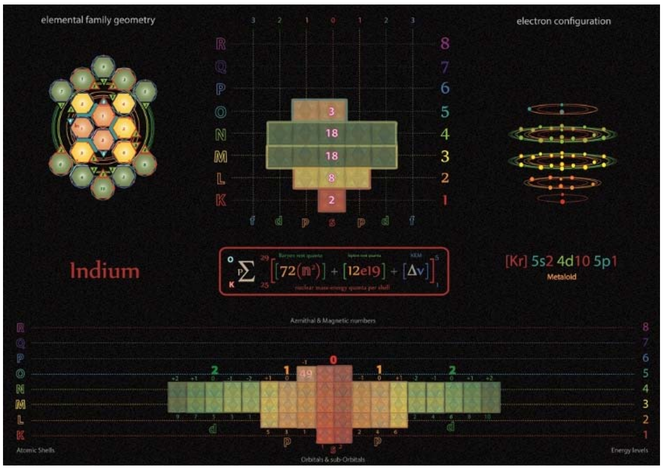
elemental family geometry
electron configuration
Indium
[Kr] 5s2 4d10 5p1
Metaloid
Azmithal & Magnetic numbers
R
Q
P
O
N
M
L
K
f
d
p
s
Atomic Shells
Energy levels
Orbitals & sub-Orbitals

Should be sufficient to explain and illustrate all properties of the Zenon atom [or any other you chose] including their rest mass-Matter energies [at absolute zero], their charged Matter topology in accord with Bohr shells/orbitals and Schrodinger numbers and the kinetic mass-energies of their bound electrons.

Clarifying question 2 about Tetryonics nuclear vs. atomic size

According to Tetryonics the proton and neutron in Deuterium nuclei touch one another at the point son their respective charge topologies where the e-field components of their constituent quarks are located - this is indicated on all Tetryonic element illustrations via red [for positive] focus points and black [for negative] focus points.

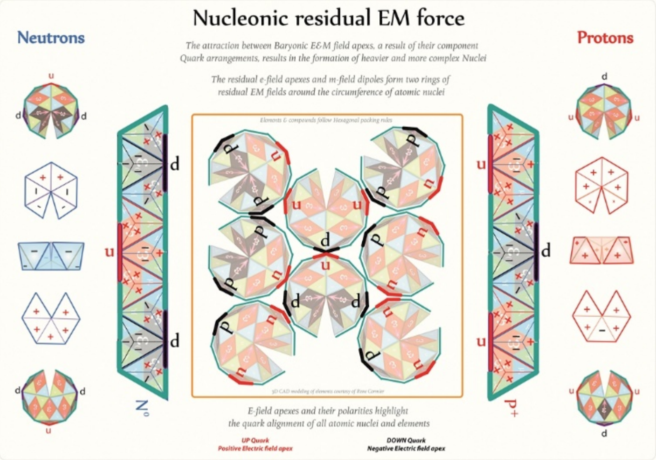
Nucleonic residual EM force
Neutrons
Protons
The attraction between Baryonic E&M field apexs, a result of their component Quark arrangements, results in the formation of heavier and more complex Nuclei
The residual e-field apexes and m-field dipoles form two rings of residual EM fields around the circumference of atomic nuclei
Elements & compounds follow Hexagonal packing rules
u
d
N^0
P^+
3D CAD modeling of elements courtesy of Rene Cormier
E-field apexes and their polarities highlight the quark alignment of all atomic nuclei and elements
UP Quark
Positive Electric field apex
DOWN Quark
Negative Electric field apex

But since Tetryonics does not include any separate strong nuclear force but just electromagnetics, all particles of Tetryonic nucleus would fly apart in a nanosecond.

Again, this misunderstanding and erroneous statement about Tetryonic theory seems to stem from you having failed to read Tetryonics [1] – Quantum Mechanics, viewed my Tetryonics 101 instruction video or even attempted to build and understand how atomic nuclei are formed and come together to form larger elements and compounds alike prior to commenting ….

In my first eBook I have gone to great lengths to show that the STRONG force does indeed exist and in fact plays a crucial role in bringing tetryons [the tetrahedral quantum of Matter topologies themselves] together to form leptons, quarks, mesons & Baryons – it is the parallel opposite charge interaction present between the planar charge faces of each tetryon within each and every Baryon [not to be mistaken for the residual strong force]

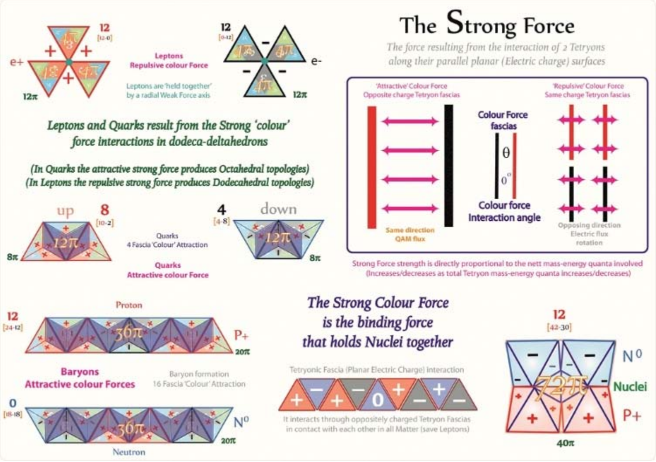
The Strong Force
The force resulting from the interaction of 2 Tetryons along their parallel planar (Electric charge) surfaces
e+
e-
12
[12-0]
[0-12]
12π
Leptons
Repulsive colour Force
Leptons are 'held together' by a radial Weak Force axis
Leptons and Quarks result from the Strong 'colour' force interactions in dodeca-deltahedrons
(In Quarks the attractive strong force produces Octahedral topologies)
(In Leptons the repulsive strong force produces Dodecahedral topologies)
up
8
[10-2]
4
[4-8]
down
8π
Quarks
4 Fascia 'Colour' Attraction
Quarks
Attractive colour Force
'Attractive' Colour Force
Opposite charge Tetryon fascias
'Repulsive' Colour Force
Same charge Tetryon fascias
Colour Force fascias
θ
0°
Colour force Interaction angle
Same direction
QAM flux
Opposing direction
Electric flux rotation
Strong Force strength is directly proportional to the nett mass-energy quanta involved
(Increases/decreases as total Tetryon mass-energy quanta increases/decreases)
Proton
[24-12]
36π
P+
20π
Baryons
Attractive colour Forces
Baryon formation
16 Fascia 'Colour' Attraction
0
[18-18]
N0
Neutron
The Strong Colour Force is the binding force that holds Nuclei together
Tetryonic Fascia (Planar Electric Charge) interaction
It interacts through oppositely charged Tetryon Fascias in contact with each other in all Matter (save Leptons)
[42-30]
72π
Nuclei
40π

Obviously this co-planar opposite charge interaction holds all atomic nuclei together and your claim of Tetryonic particles flying away in a nanosecond just doesn't hold water…

In fact just place two sheets of glass together with a thin film of water between them and see how easy it is to pull those sheets apart by applying a force orthogonal to the co-planar sheets of glass and you'll soon understand the STRONG force much better that you do presently.

This charge interaction between the e-fields of quarks in Baryons serves to not only bring positive charge Protons together with neutral Neutrons [something modern theory cannot explain] but also serve to orientate each baryon to the other so as to facilitate electron bonding and the creation of ever larger atomic elements and compounds [the residual strong or EM force as it is often termed or the London force]

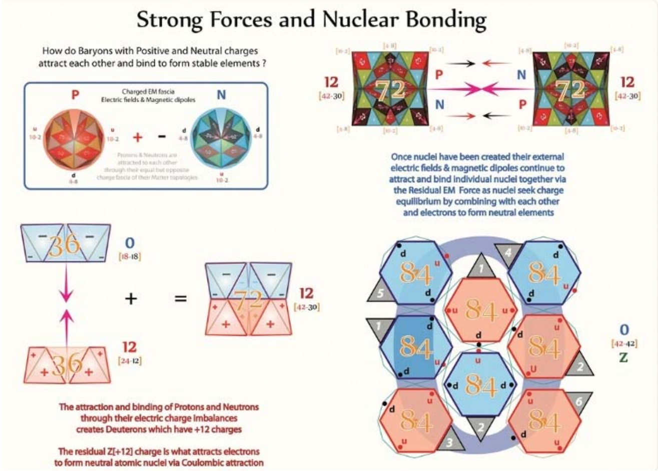
Strong Forces and Nuclear Bonding
How do Baryons with Positive and Neutral charges attract each other and bind to form stable elements ?
P
Charged EM fascia
Electric fields & Magnetic dipoles
N
u
10-2
u
10-2
+
-
d
4-8
d
4-8
d
4-8
Protons & Neutrons are attracted to each other through their equal but opposite charge fascia of their Matter topologies
u
10-2
12
[42-30]
72
P
N
N
P
72
12
[42-30]
Once nuclei have been created their external electric fields & magnetic dipoles continue to attract and bind individual nuclei together via the Residual EM Force as nuclei seek charge equilibrium by combining with each other and electrons to form neutral elements
36
0
[18-18]
+
=
72
12
[42-30]
36
12
[24-12]
84
84
84
84
84
84
84
1
2
3
4
5
6
7
u
d
0
[42-42]
Z
The attraction and binding of Protons and Neutrons through their electric charge imbalances creates Deuterons which have +12 charges
The residual Z[+12] charge is what attracts electrons to form neutral atomic nuclei via Coulombic attraction

Yes I have an ‘obsession’ with all atomic nuclei have equal numbers of Protons, electron AND Neutrons in accord with their Z number, just as you have an obsession with them being Hydrogen atoms with excess Neutrons ‘bound’ to them – but for the reasons highlighted above, and many others – as for the quantum structure of Protium (regular hydrogen), Deuterium and Tritium being ‘missing’ - again my T[1] - QM eBook holds the answers

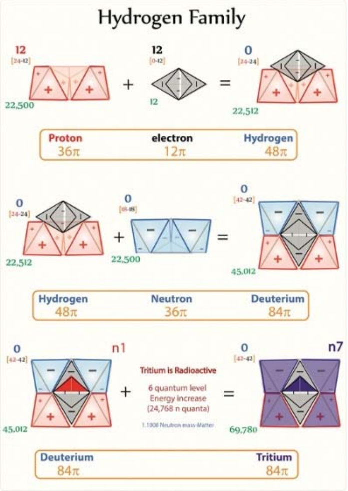
Hydrogen Family
12
[24-12]
22,500
+
12
[0-12]
12
=
0
[24-24]
22,512
Proton
36π
electron
12π
Hydrogen
48π
0
[24-24]
22,512
+
0
[18-18]
22,500
=
0
[42-42]
45,012
Hydrogen
48π
Neutron
36π
Deuterium
84π
0
[42-42]
45,012
n1
+
Tritium is Radioactive
6 quantum level
Energy increase
(24,768 n quanta)
1.1008 Neutron mass Matter
=
0
[42-42]
69,780
n7
Deuterium
84π
Tritium
84π

With additional space being made to highlight the Tetryonic genesis of Tritium & Helium 3

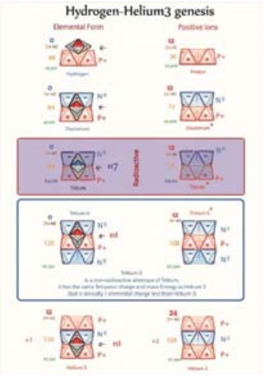
Hydrogen-Helium3 genesis
Elemental Form
Positive Ions
e-
P+
N°
n7
Radioactive
nI
34

All isotopes are simply Z number nuclei with the energy levels of their Baryons raised to the next quantum energy level [n1-n8] in accord with the 3D material symmetry of the atomic nuclei themselves and the Bohr shells/Schrodinger principal quantum numbers, which can occur either through natural energy absorption or the artificial increase of their mass-energy content in the lab….

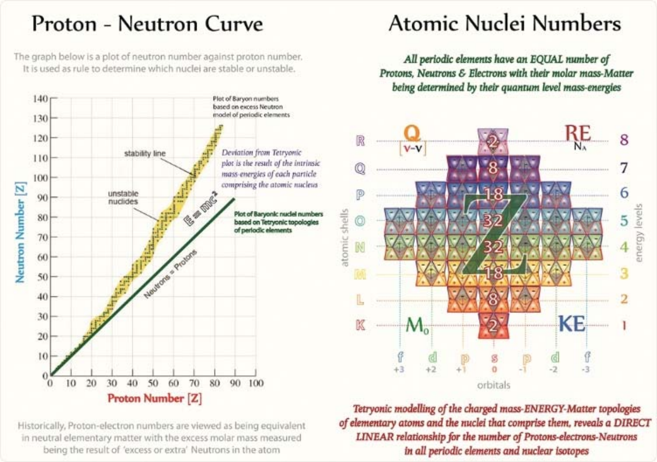

Proton - Neutron Curve

The graph below is a plot of neutron number against proton number. It is used as rule to determine which nuclei are stable or unstable.

Historically, Proton-electron numbers are viewed as being equivalent in neutral elementary matter with the excess molar mass measured being the result of 'excess or extra' Neutrons in the atom

Atomic Nuclei Numbers

All periodic elements have an EQUAL number of Protons, Neutrons & Electrons with their molar mass-Matter being determined by their quantum level mass-energies

Tetryonic modelling of the charged mass-ENERGY-Matter topologies of elementary atoms and the nuclei that comprise them, reveals a DIRECT LINEAR relationship for the number of Protons-electrons-Neutrons in all periodic elements and nuclear isotopes

Neutron Number [Z]
Proton Number [Z]
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
Plot of Baryon numbers based on excess Neutron model of periodic elements
stability line
unstable nuclides
Deviation from Tetryonic plot is the result of the intrinsic mass-energies of each particle comprising the atomic nucleus
E = mc²
Plot of Baryonic nuclei numbers based on Tetryonic topologies of periodic elements
Neutrons = Protons
Q
[v-v]
RE
N_A
R
Q
P
O
N
M
L
K
8
7
6
5
4
3
2
1
2
8
18
32
32
18
8
2
Z
M₀
KE
atomic shells
energy levels
f
d
p
s
p
d
f
+3
+2
+1
0
-1
-2
-3
orbitals

Obviously with all Deuterium nuclei being quantum synchronous convectors as the energy levels of their Baryons are increased [or decreased] the electrons bound to them respond in kind [hence synchronous] via spectral line emissions [or absorption] ….all covered in Tetryonics [2] – Quantum Electrodynamics.

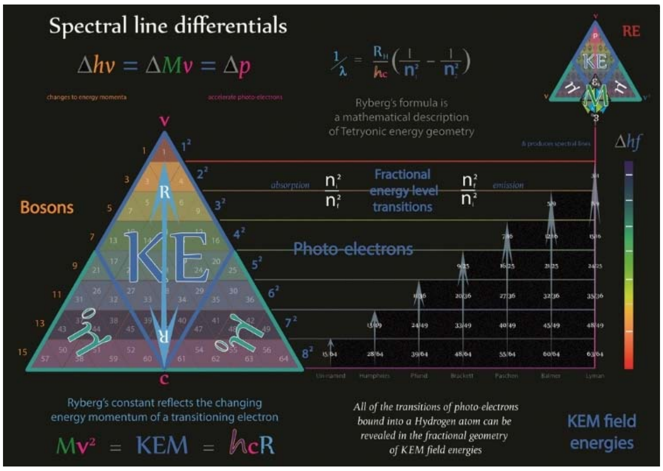

Spectral line differentials

$$\Delta hv = \Delta Mv = \Delta p$$

$$\frac{1}{\lambda} = \frac{R_H}{hc}\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$

Ryberg's formula is a mathematical description of Tetryonic energy geometry

Ryberg's constant reflects the changing energy momentum of a transitioning electron

$$Mv^2 = KEM = hcR$$

All of the transitions of photo-electrons bound into a Hydrogen atom can be revealed in the fractional geometry of KEM field energies

KEM field energies

changes to energy momenta
accelerate photo-electrons
& produces spectral lines
RE
KE
M
Δhf
Bosons
v
c
R
KE
1²
2²
3²
4²
5²
6²
7²
8²
1
3
5
7
9
11
13
15
absorption
emission
Fractional energy level transitions
Photo-electrons
15/64
28/64
39/64
48/64
55/64
60/64
63/64
13/49
24/49
33/49
40/49
45/49
48/49
11/36
20/36
27/36
32/36
35/36
9/25
16/25
21/25
24/25
7/16
12/16
15/16
5/9
8/9
3/4
Un-named
Humphries
Pfund
Brackett
Paschen
Balmer
Lyman

There is NO explanation in modern chemical or quantum theory that accounts for the Proton-Neutron 'curve' of atomic isotopes nor for the islands of stability that exist for various elements, however Tetryonic theory accounts for all of this firstly by 'straightening' out the curved Z line of Protons and Neutrons underpinning atomic elements [by showing that all atomic nuclei are comprised of D nuclei not H atoms with excess neutrons] and then showing that the observed isotopic masses of each are the result of increased principal energy levels of their constituent baryons with the D nuclei comprising each element…. It is no error or 'false notion' as you put it, rather it is the current explanation of these facts by modern theory that is in error

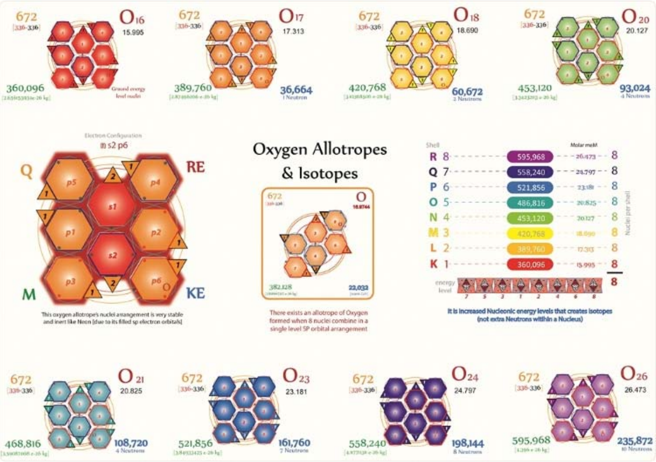
672
[336-336]
O16
15.995
360,096
Ground energy level nuclei
O17
17.313
389,760
36,664
1 Neutron
O18
18.690
420,768
60,672
2 Neutrons
O20
20.127
453,120
93,024
4 Neutrons
Electron Configuration
s2 p6
Q
RE
M
KE
p5
p4
s1
p1
p2
s2
p3
p6
Oxygen Allotropes
& Isotopes
O
382,128
22,032
Shell
Molar meM
R 8
595,968
26.473
Q 7
558,240
24.797
P 6
521,856
23.181
O 5
486,816
20.825
N 4
453,120
20.127
M 3
420,768
18.690
L 2
389,760
17.313
K 1
360,096
15.995
8
Nuclei per shell
energy level
O21
20.825
468,816
108,720
4 Neutrons
O23
23.181
521,856
161,760
7 Neutrons
O24
24.797
558,240
198,144
8 Neutrons
O26
26.473
595,968
235,872
10 Neutrons

This oxygen allotrope's nuclei arrangement is very stable and inert like Neon [due to its filled sp electron orbitals]

There exists an allotrope of Oxygen formed when 8 nuclei combine in a single level SP orbital arrangement

It is increased Nucleonic energy levels that creates Isotopes (not extra Neutrons within a Nucleus)

You have chosen Lithium to make your point I chose Carbon, Nitrogen and Oxygen to make mine, but the explanation for Lithium remains the same….. an increase in the mass-energies of the baryons making up the element concerned… btw I have done spreadsheets for each and all the elements as a matter of course during the development of my theory.

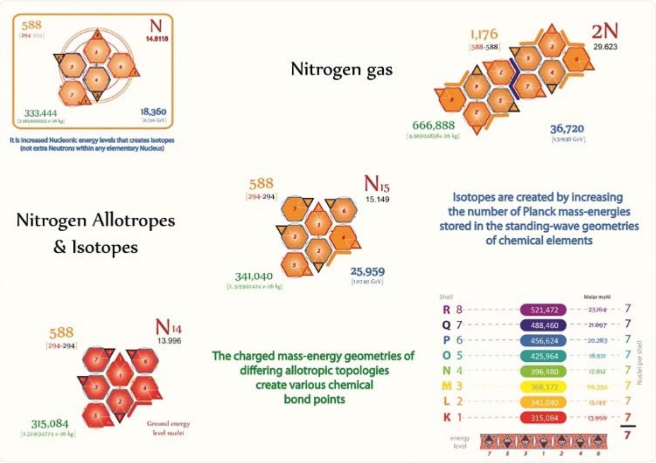
588
[294-294]
N
14.8118
333,444
18,360
Nitrogen gas
1,176
[588-588]
2N
29.623
666,888
36,720
[1.5938 GeV]
Nitrogen Allotropes
& Isotopes
N15
15.149
341,040
25,959
N14
13.996
315,084
Ground energy level nuclei
Shell
Molar meM
R 8
521,472
23.164
Q 7
488,460
21.697
P 6
456,624
20.283
O 5
425,964
18.921
N 4
396,480
17.612
M 3
368,172
16.354
L 2
341,040
15.149
K 1
315,084
13.996
7
Nuclei per shell
energy level

It is increased Nucleonic energy levels that creates Isotopes (not extra Neutrons within any elementary Nucleus)

Isotopes are created by increasing the number of Planck mass-energies stored in the standing-wave geometries of chemical elements

The charged mass-energy geometries of differing allotropic topologies create various chemical bond points

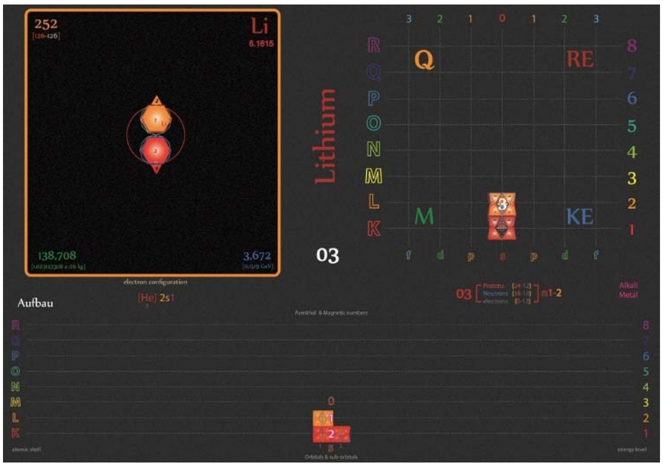
252
[126-126]
Li
6.1615
138,708
3,672
electron configuration
[He] 2s1
Lithium
03
R
Q
P
O
N
M
L
K
Q
RE
M
KE
3 2 1 0 1 2 3
8 7 6 5 4 3 2 1
f d p s p d f
03
Protons
Neutrons
electrons
Alkali Metal
Aufbau
Azimuthal & Magnetic numbers
atomic shell
energy level
Orbitals & sub-orbitals

Element 3 – Lithium [ground state atomic topology and electronic configuration]

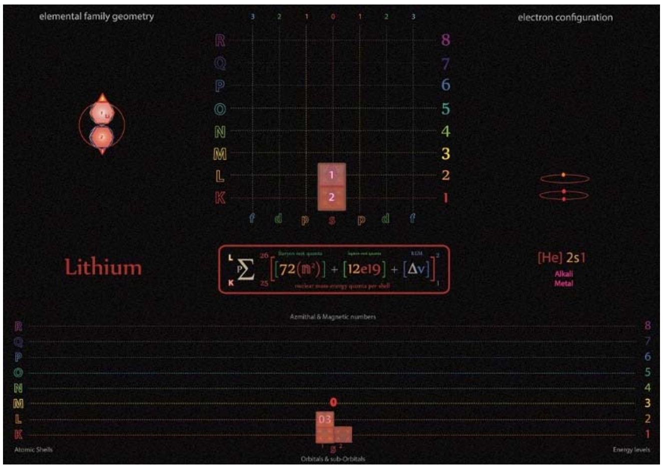
elemental family geometry
electron configuration
Lithium
[He] 2s1
Alkali Metal
R Q P O N M L K
8 7 6 5 4 3 2 1
f d p s p d f
3 2 1 0 1 2 3
Azmithal & Magnetic numbers
Atomic Shells
Energy levels
Orbitals & sub-Orbitals
[72(m²)] + [12e19] + [Δv]
nuclear mass-energy quanta per shell
03

Regarding your comment of "but in reality **only electrons are on these shells** and the structure of the nucleus is completely separate from these shells", - that is only the current opinion/consensus explanation of quantum chemistry at this stage and is obviously at odds with Tetryonic theory *built from the ground up from equilateral Planck energies not the top-down through supposition*….. I am confident that over time this consensus will change.

1. Explaining the extra (neutron) weight as higher energy-levels of the Deuterium nuclei seems *ad hoc* explanation with no supporting evidence – *simply nonsense*…..
2. No calculation is presented to *derive a prediction* for the excess mass (or energy) from the Tetryonics model – *have you even read my work in detail – calculations and mass-energies are provided everywhere throughout it*.
3. The excess mass of, say, Lithium 7 (one excess neutron compared to 3x deuterium) *– again a number based on your erroneous viewpoint – not y given explanation and illustrations of the same.*
4. This is inconsistent with the fact that there are **multiple** isotopes of each element.- *how so?*

   The energy level of Baryons of each and every illustrated element in Tetryonic can be increased resulting in isotopes of that element – with NO extra Neutrons present, again you seem to have failed to read and/or comprehend what was presented before you in the Tetryonics chemistry eBook.

5. See above for Hydrogen family [please]

All the 'observed' masses you have provided in that table are firstly just that, weighted [averaged] measurements of particles with kinetic energies [heat, spectral, motion etc.] not the absolute rest masses I have provided and clearly stated as such…… you must account for observational and inherent kinetic energies for your stated values to even begin reflecting the quantum values of particles at absolute zero.

Again, 'isotopes' is NOT what you think and proclaim – there are no extra Neutrons in the nuclei – only Baryonic energy above that of the res/ground state of such an element at absolute zero – unless you can account for those KEM field energies [as should be done]

Tetryonic Problem: Helium as Alkaline Earth metal

- And Hydrogen is element 1 which no-one can agree on which position it truly belongs in on the current periodic table.

Tetryonics problem: 120 possible maximum elements

- That is what is dictated by the charge topology of particles and elements themselves, perhaps you'd care to share what your spherical [or other] model for element 93 is, and how all the 'extra' Neutrons bind to each other to create that particular model?

Tetryonics shows quite clearly the reasons for there being only 120 possible periodic elements and it agrees with the math and equations of both Bohr and Schrodinger's models despite the differing quantum topologies.

Perhaps, you can produce element 121 and prove me wrong on this.

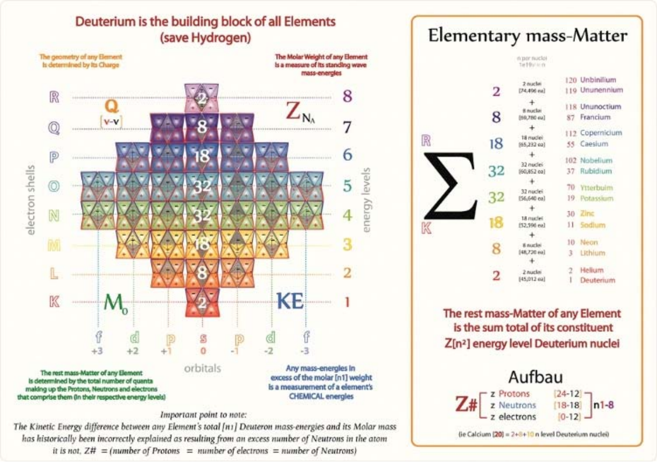
Deuterium is the building block of all Elements
(save Hydrogen)
The geometry of any Element is determined by its Charge
The Molar Weight of any Element is a measure of its standing wave mass-energies
R
Q
P
O
N
M
L
K
electron shells
Q
[v-v]
Z_{N_A}
8
7
6
5
4
3
2
1
energy levels
2
8
18
32
32
18
8
2
M_0
KE
f +3
d +2
p +1
s 0
p -1
d -2
f -3
orbitals
The rest mass-Matter of any Element is determined by the total number of quanta making up the Protons, Neutrons and electrons that comprise them (in their respective energy levels)
Any mass-energies in excess of the molar [n1] weight is a measurement of a element's CHEMICAL energies
Important point to note:
The Kinetic Energy difference between any Element's total [n1] Deuteron mass-energies and its Molar mass has historically been incorrectly explained as resulting from an excess number of Neutrons in the atom it is not. Z# = (number of Protons = number of electrons = number of Neutrons)
Elementary mass-Matter
R
Σ
K
2 – 2 nuclei [74,496 ea] – 120 Unbinilium, 119 Ununennium
8 – 8 nuclei [69,780 ea] – 118 Ununoctium, 87 Francium
18 – 18 nuclei [65,232 ea] – 112 Copernicium, 55 Caesium
32 – 32 nuclei [60,852 ea] – 102 Nobelium, 37 Rubidium
32 – 32 nuclei [56,640 ea] – 70 Ytterbuim, 19 Potassium
18 – 18 nuclei [52,596 ea] – 30 Zinc, 11 Sodium
8 – 8 nuclei [48,720 ea] – 10 Neon, 3 Lithium
2 – 2 nuclei [45,012 ea] – 2 Helium, 1 Deuterium
The rest mass-Matter of any Element is the sum total of its constituent Z[n²] energy level Deuterium nuclei
Aufbau
Z# [z Protons [24-12], z Neutrons [18-18], z electrons [0-12]] n1-8
(ie Calcium [20] = 2+8+10 n level Deuterium nuclei)

Additional random quotes - the uncertainty principle erroneously dictates that position and momentum cannot be simultaneously determined

- There's nothing random about this comment/quote - Tetryonic theory is founded on the SINGLE postulate that Planck mass-energy momenta has an quantum equilateral geometry [ie quantised angular momenta (m^2/s) is equilateral/triangular not circular as it is the macro sense of angular momentum] - from that flows the fact that the probabilistic nature of quantum theory and the maths used to describe it are in fact not 'fuzzy' or indeterminate as supposed presently but in fact an emergent property of the tessellation of mass-energy momenta at the quantum level.

The inherent indeterminacy of quantum theory is a crutch used my mathematicians and scientists alike to explain many of the puzzling aspects of quantum theory like the inherent statically probabilities of quantum fields themselves or the wave-particle duality of photons or particles in motion….. Again it is them and modern theory that have simply failed to grasp the geometric reality of energy momenta that underpins these phenomena and how they build to form the familiar laws of Nature at our scale of observation.

Tetryonic theory is all about correcting these erroneous assumptions and established viewpoints – but any change of perspective is not without birth pangs of angst and misunderstanding from those who first encounter it [like Newton, Planck and Einstein with their theories of Gravity, Quantum mechanics & Relativity theories before it]

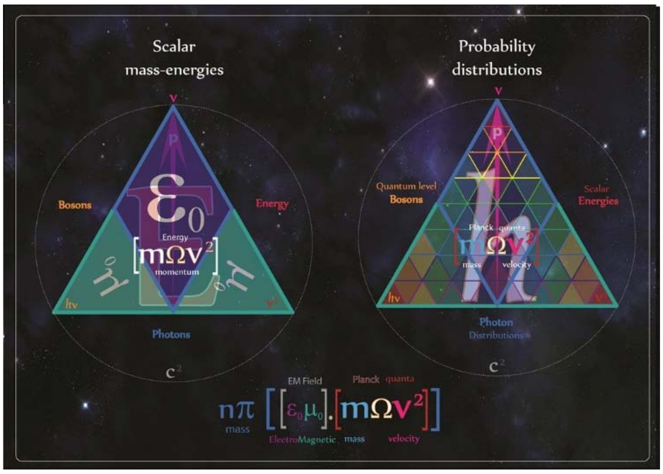
Scalar
mass-energies
Probability
distributions
v
p
Bosons
Energy
ε0
Energy
[mΩv²]
momentum
hv
v²
Photons
c²
Quantum level
Bosons
Scalar
Energies
Planck quanta
mass
velocity
Photon
Distributions
EM Field
Planck quanta
nπ [[ε0μ0].[mΩv²]]
mass
ElectroMagnetic
mass
velocity

That it is the transverse bosons Max Planck described using E=n.hv when resolving the ultraviolet catastrophe and spectral line emissions

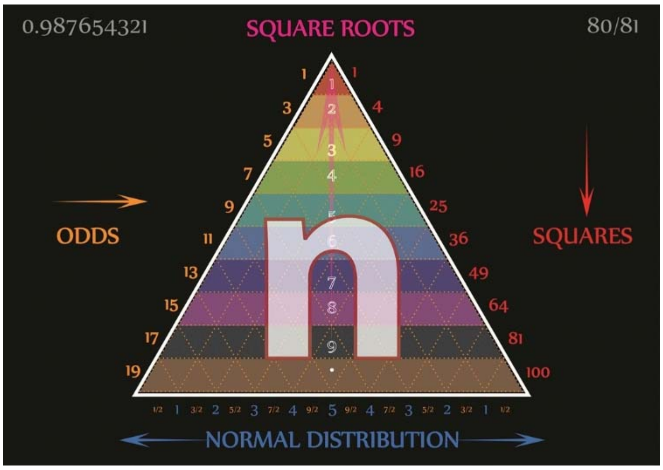
0.987654321
SQUARE ROOTS
80/81
ODDS
SQUARES
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
1
2
3
4
5
6
7
8
9
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2
NORMAL DISTRIBUTION

And which subsequent theorists incorrectly applied directly to Einstein's formulation of E= hf to describe neutral PHOTONS of energy momenta [dual charge bosons – quantum transformers]

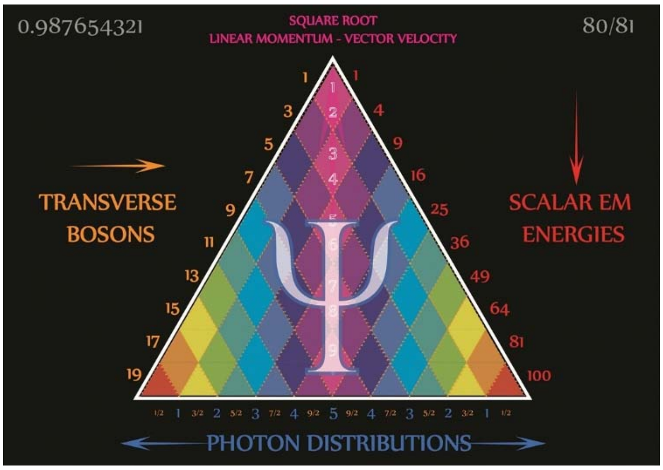
0.987654321
SQUARE ROOT
LINEAR MOMENTUM - VECTOR VELOCITY
80/81
TRANSVERSE
BOSONS
SCALAR EM
ENERGIES
1
3
5
7
9
11
13
15
17
19
1
4
9
16
25
36
49
64
81
100
1
2
3
4
Ψ
1/2 1 3/2 2 5/2 3 7/2 4 9/2 5 9/2 4 7/2 3 5/2 2 3/2 1 1/2
PHOTON DISTRIBUTIONS

And then erroneously attributed to an inherent uncertainly in quantum field dynamics etc. that does NOT and never existed in Nature itself or in Tetryonics….

I do look forward to further elaborating Tetryonic theory in detail with you – but also strongly suggest that you become more familiar with my Tetryonics [1] – Quantum Mechanics and & Tetryonics [2] - Quantum Electrodynamics eBooks beforehand as a lot of your suspicions of, and reaction to Tetryonic theory seem to stem from a lack of understanding of the basics of the theory….

And as I mentioned in previous communications, I remain open to an online discussion so we may discuss these points in far more detail than possible herein.

There is also lots more additional information and explanation of Tetryonic theory to be found on YouTube and Otto's Science of Life web- blog & the detailed 3D CAD models developed by Rene Cormier in the Google sketchup warehouse.

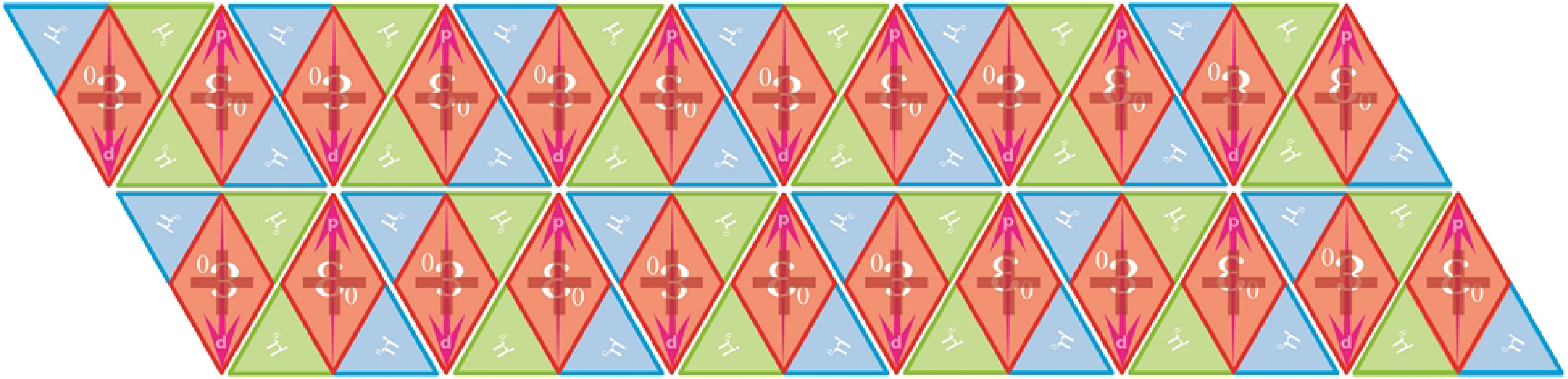

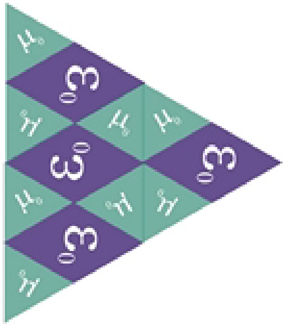

2D radiant EM mass-Energies

n1 - Charged Tetryon Templates

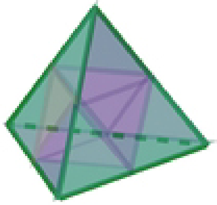

3D standing-wave Matter

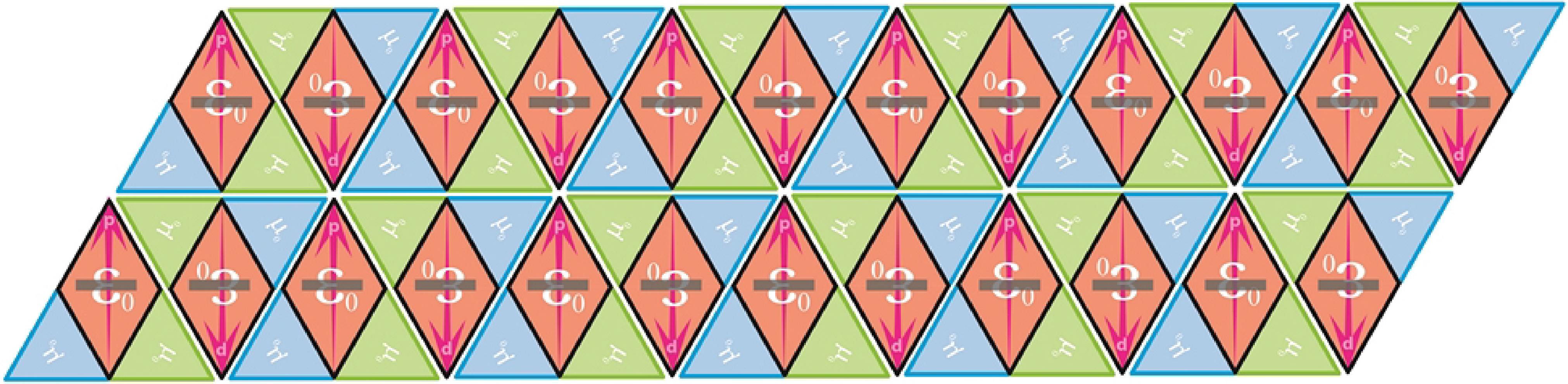

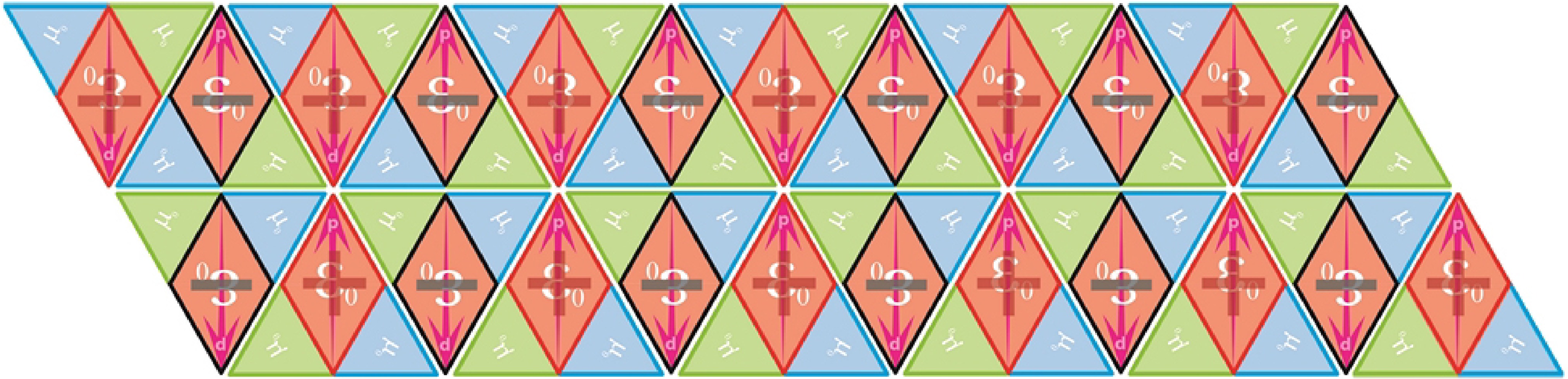

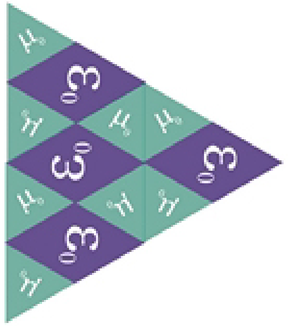

2D radiant EM mass-Energies

n1 - Neutral Tetryon Templates

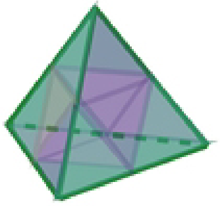

3D standing-wave Matter

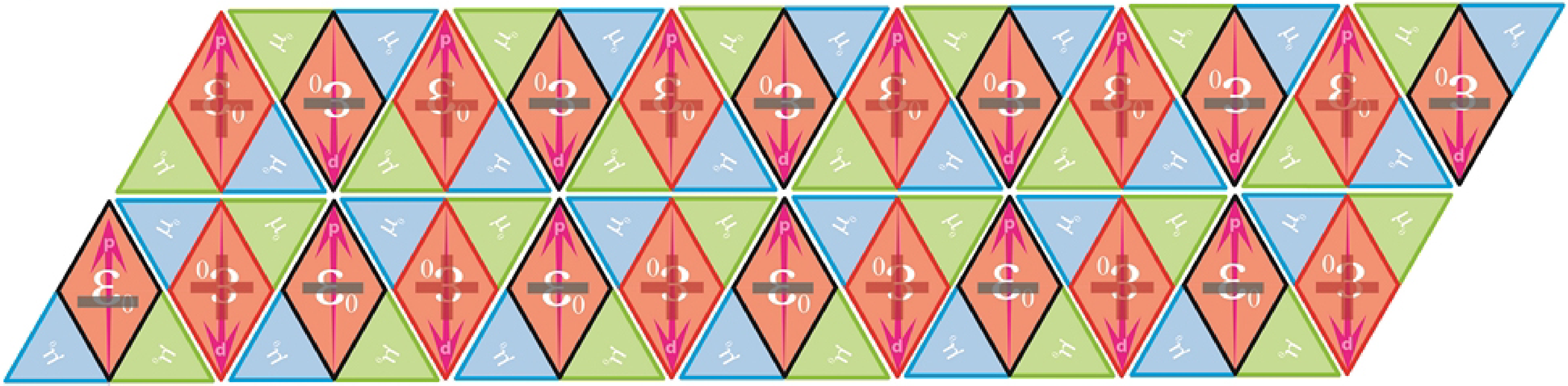

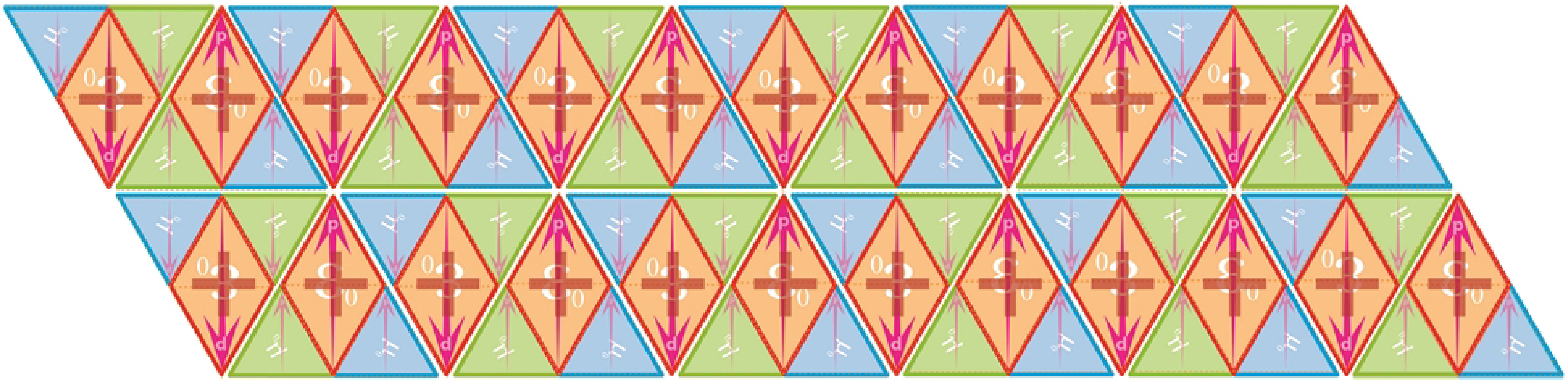

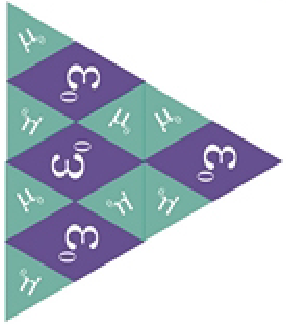

2D radiant EM mass-Energies

n2 - Charged Tetryon Templates

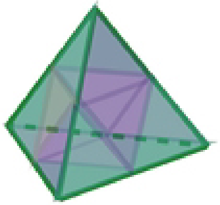

3D standing-wave Matter

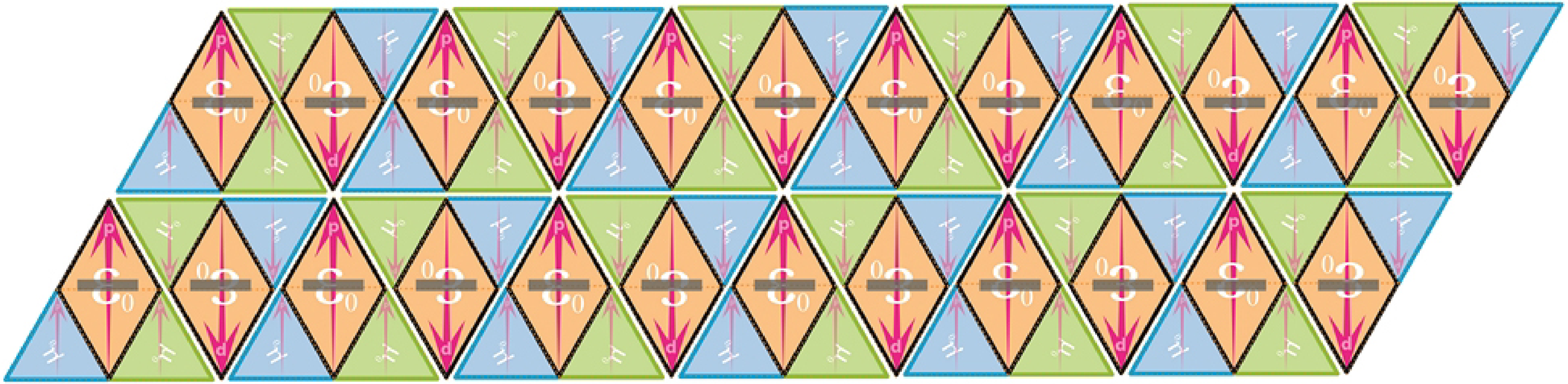

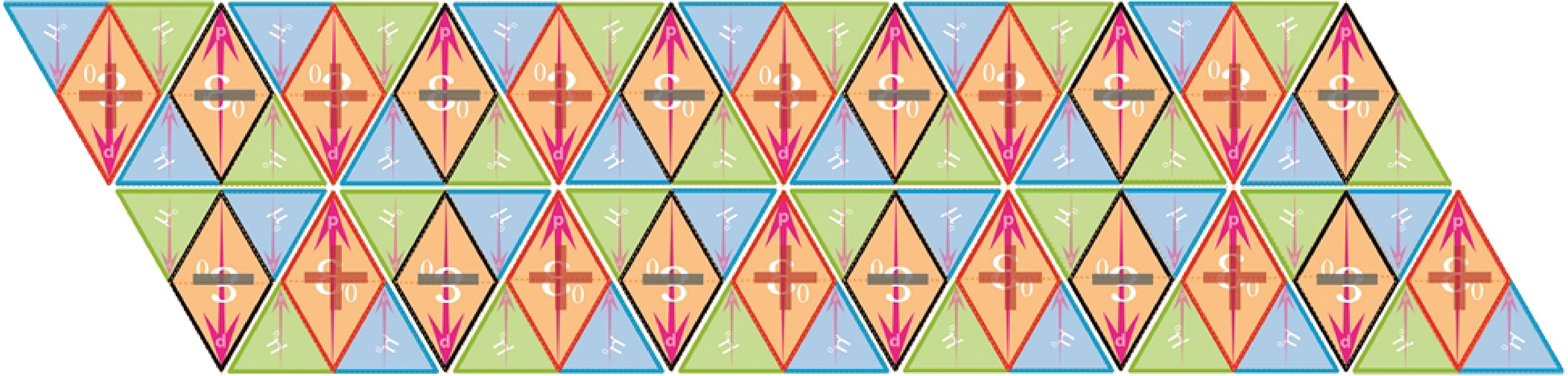

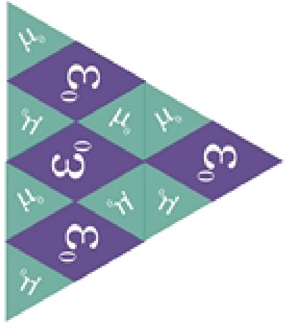

2D radiant EM mass-Energies

n2 - Neutral Tetryon Templates

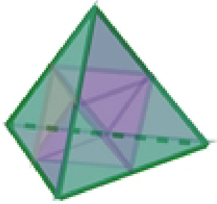
3D standing-wave Matter

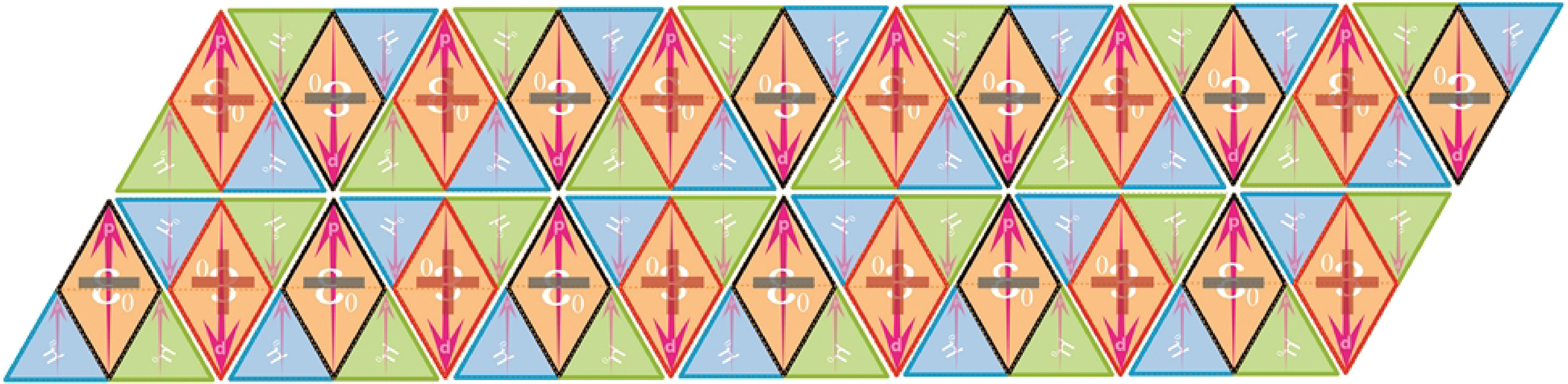

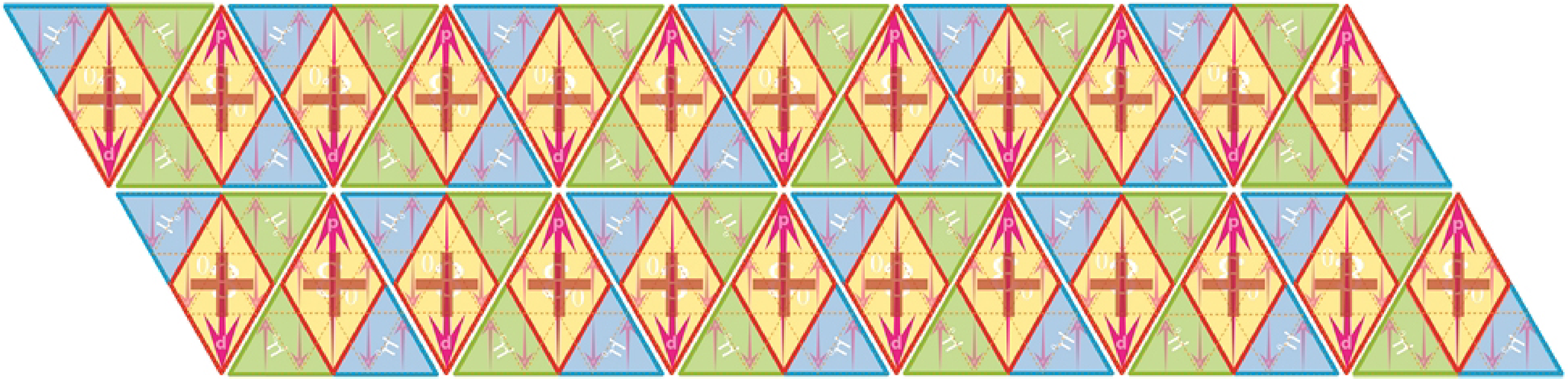

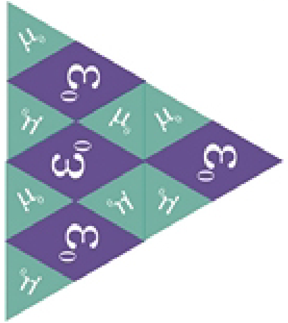

2D radiant EM mass-Energies

n3 - Charged Tetryon Templates

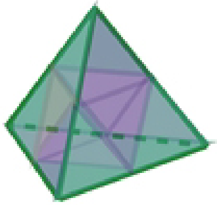

3D standing-wave Matter

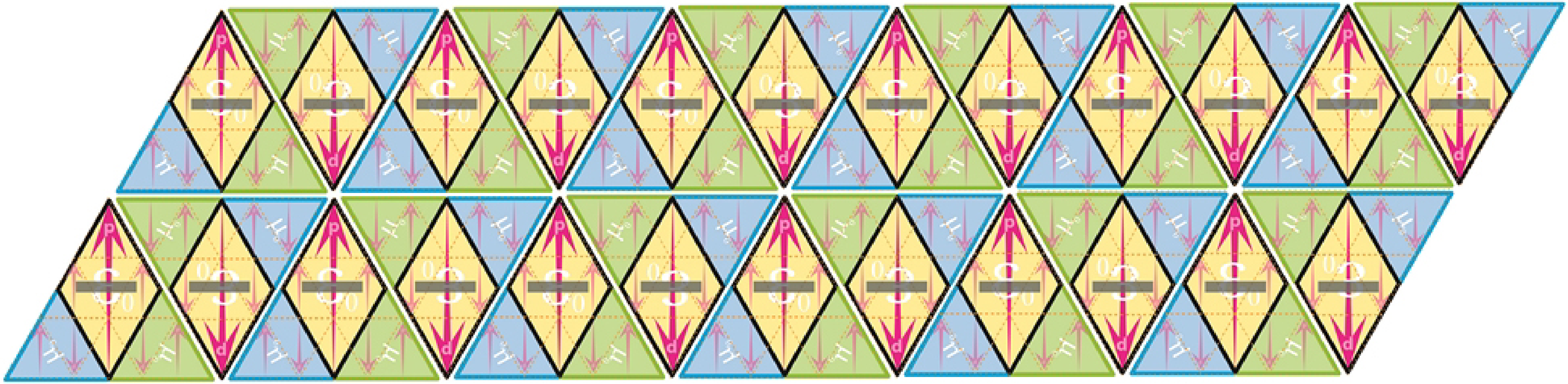

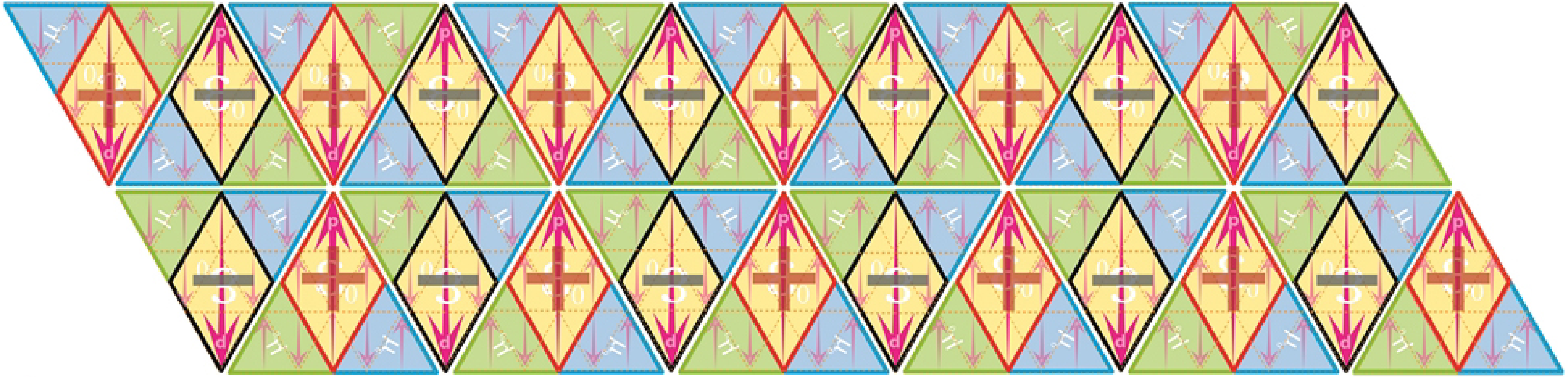
p
d
μ_0
ε_0
+
−

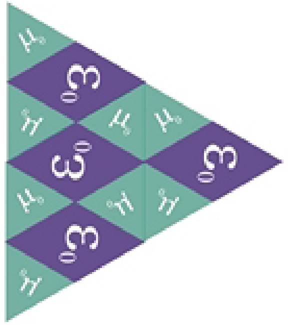
μ_0
ε_0

2D radiant EM mass-Energies

n3 - Neutral Tetryon Templates

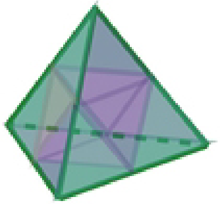

3D standing-wave Matter

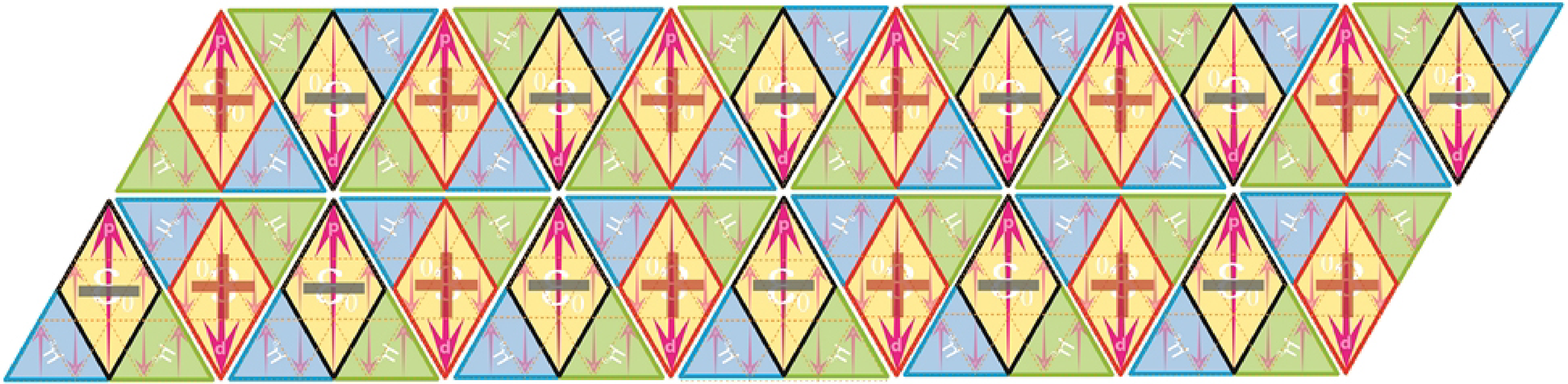
p
d
μ_0
ε_0
+
−

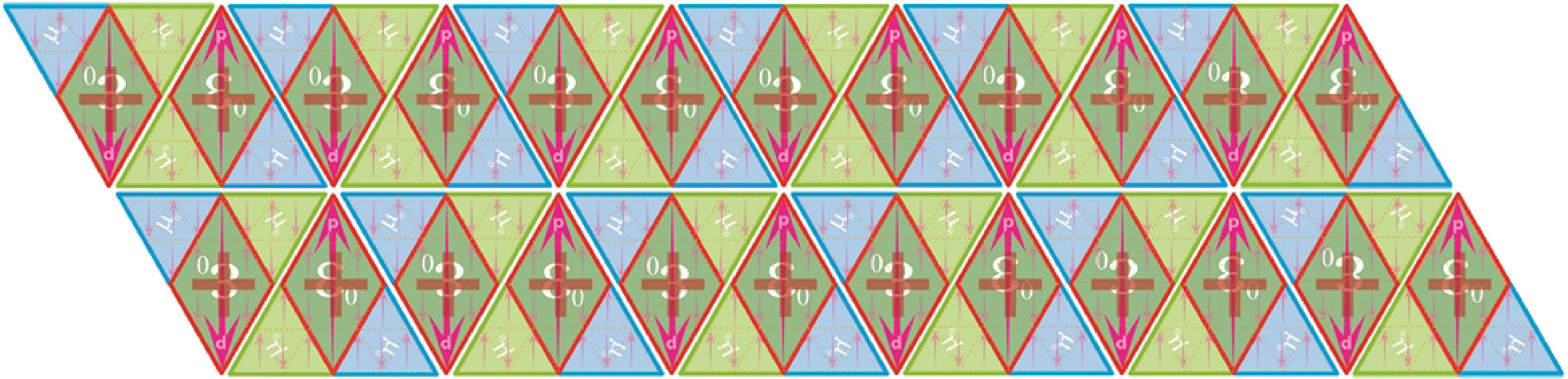

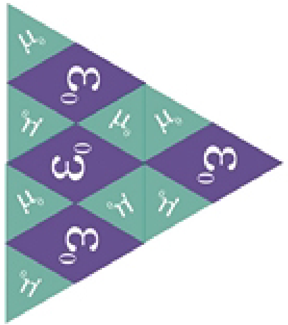

2D radiant EM mass-Energies

n4 - Charged Tetryon Templates

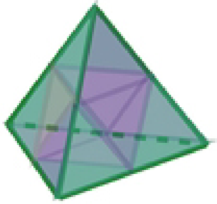

3D standing-wave Matter

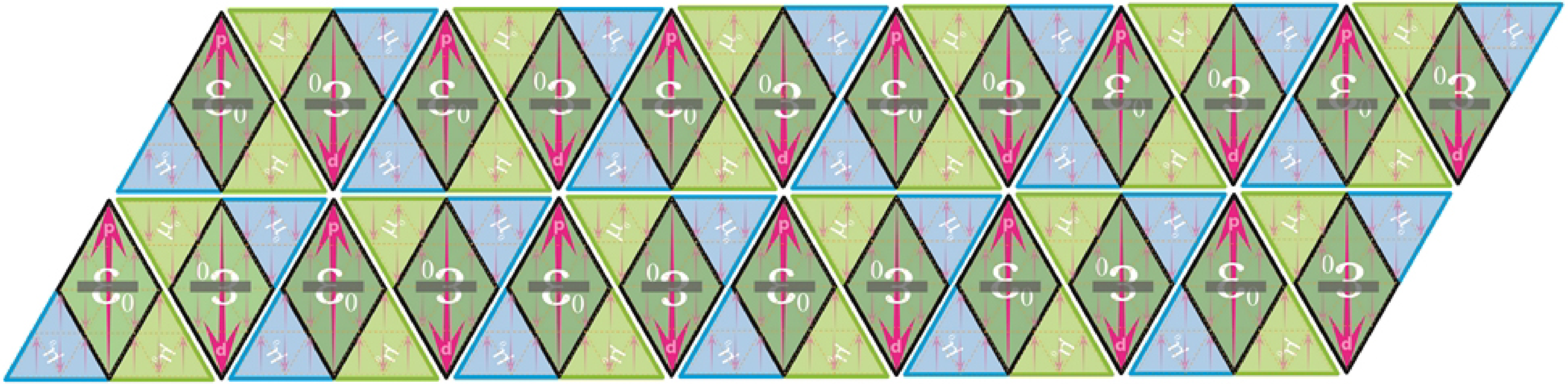

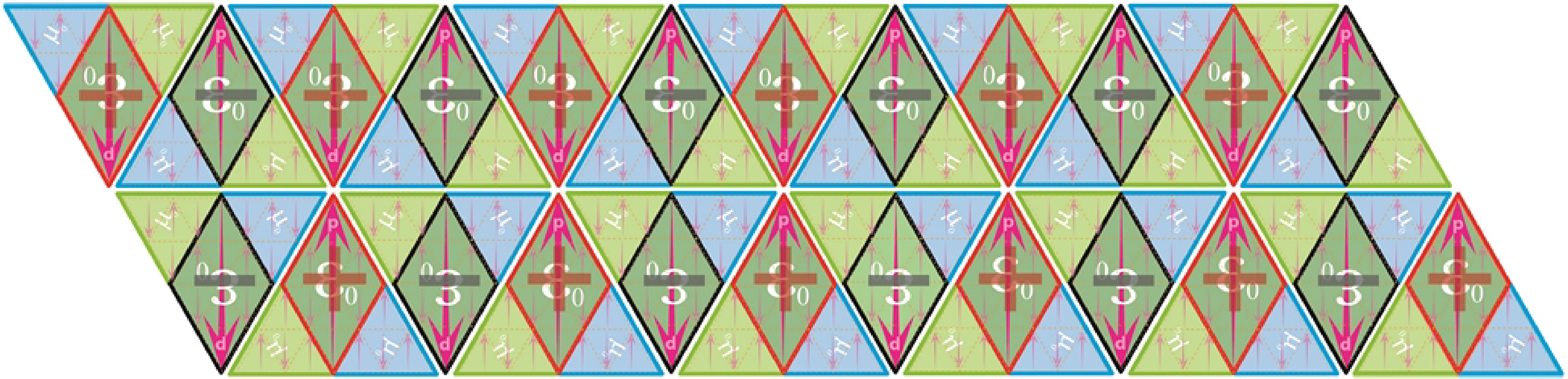

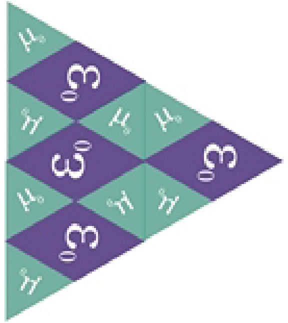

2D radiant EM mass-Energies

n4 - Neutral Tetryon Templates

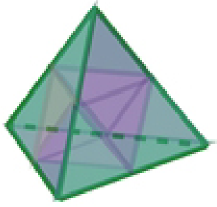

3D standing-wave Matter

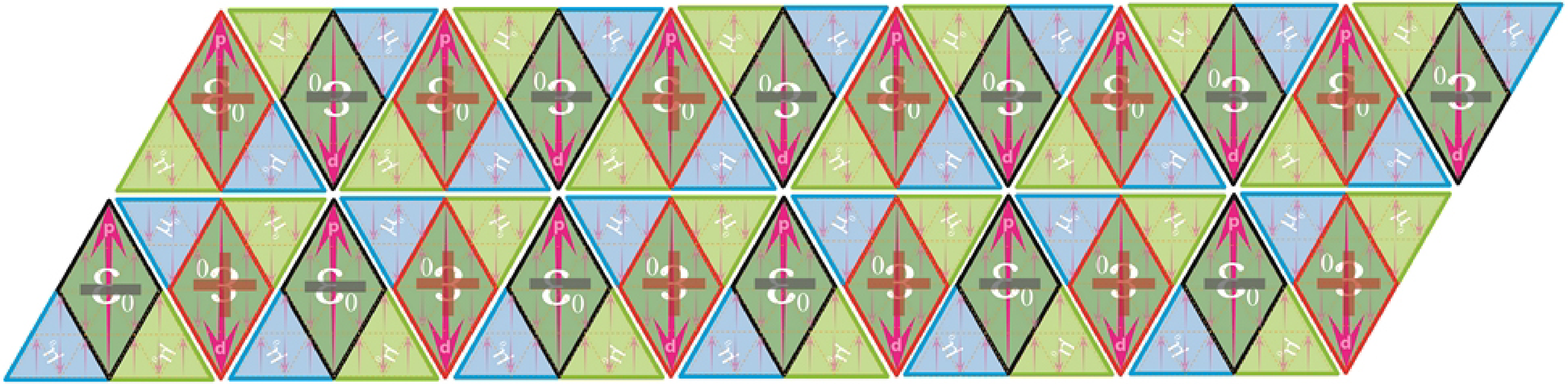

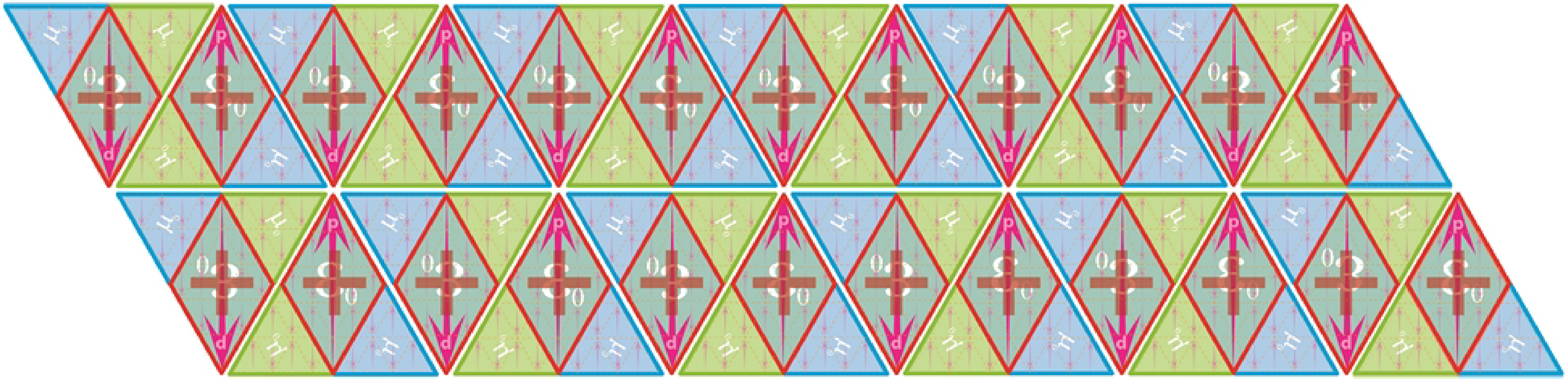

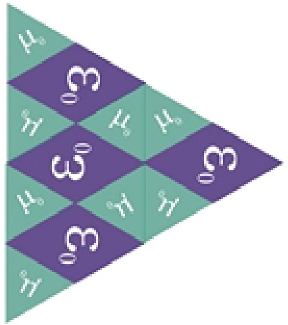

2D radiant EM mass-Energies

n5 - Charged Tetryon Templates

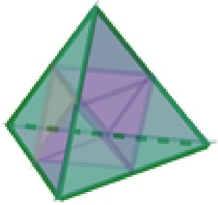

3D standing-wave Matter

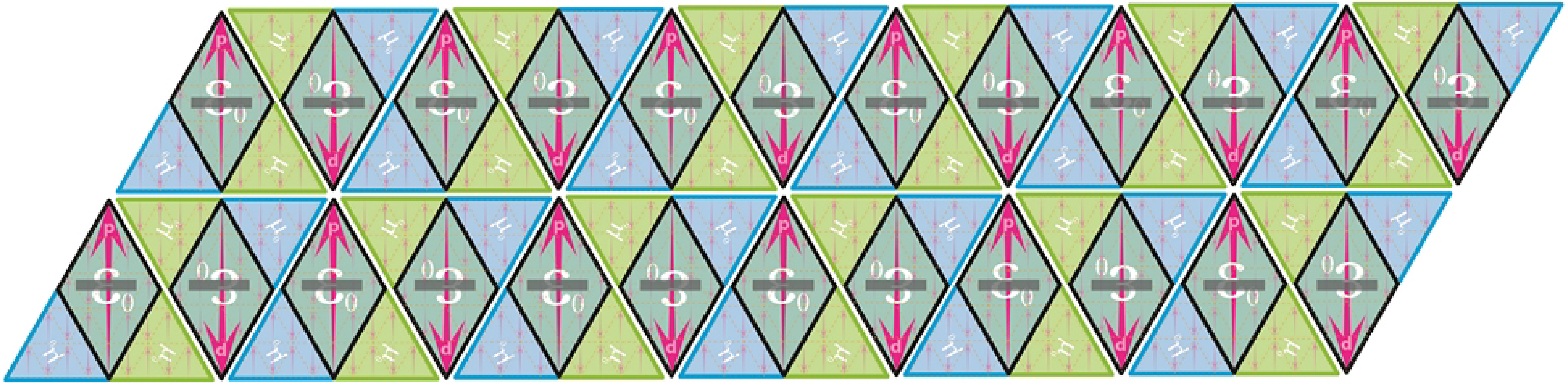

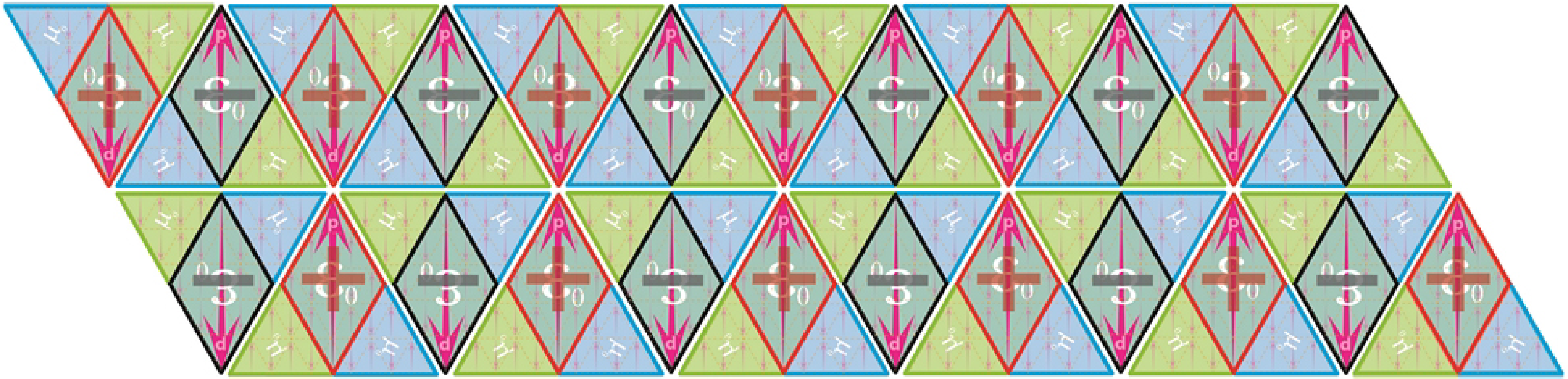
p
d
μ_0
ε_0

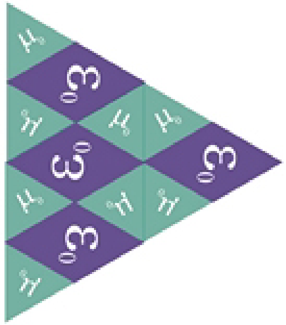
μ_0
ε_0

2D radiant EM mass-Energies

n5 - Neutral Tetryon Templates

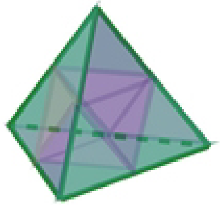

3D standing-wave Matter

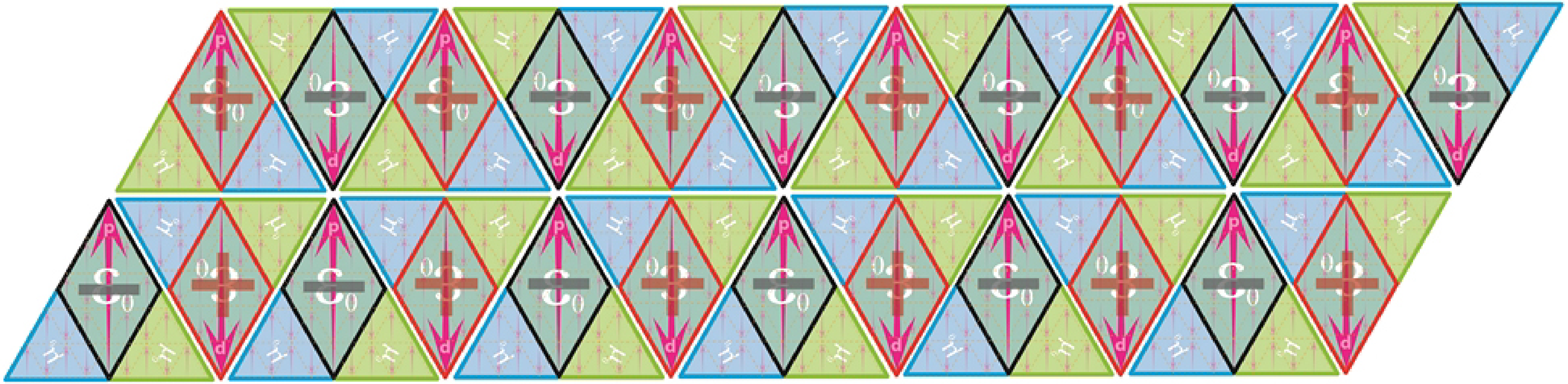
p
d
μ_0
ε_0

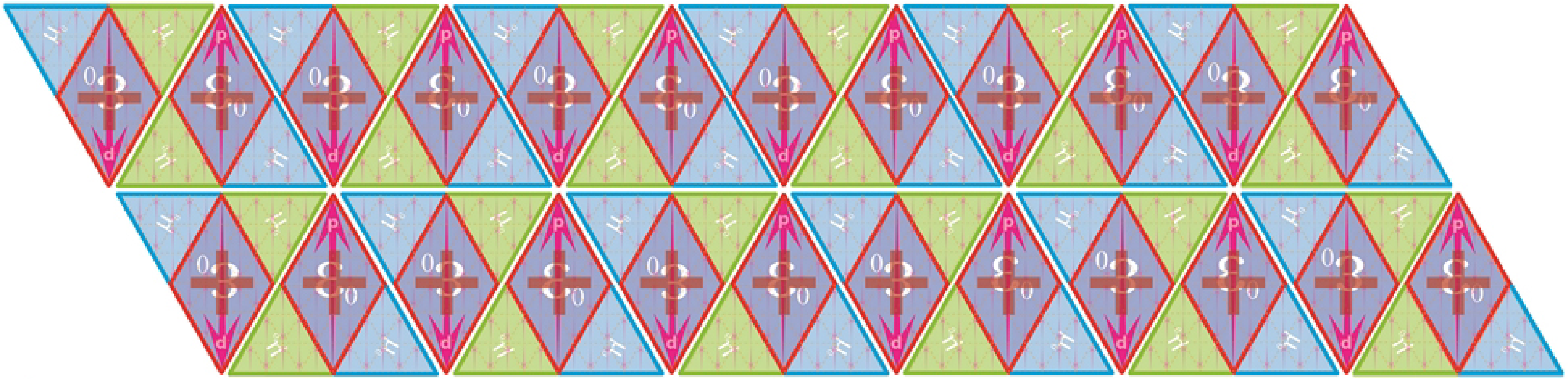

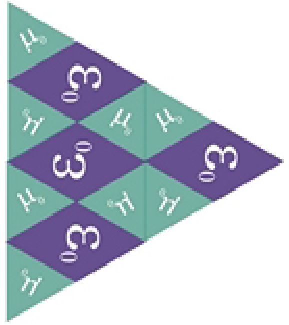

2D radiant EM mass-Energies

n6 - Charged Tetryon Templates

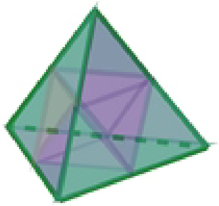

3D standing-wave Matter

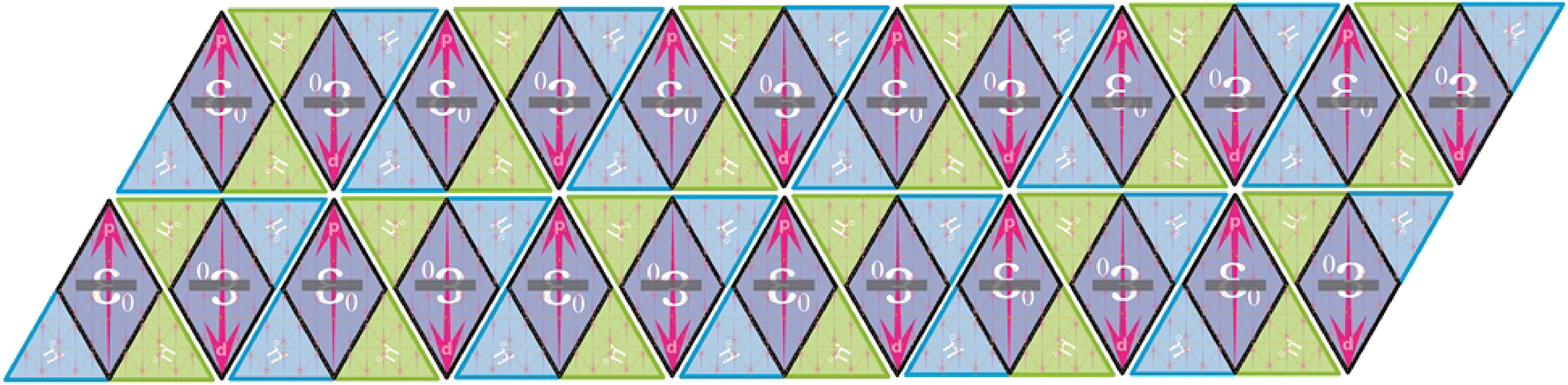

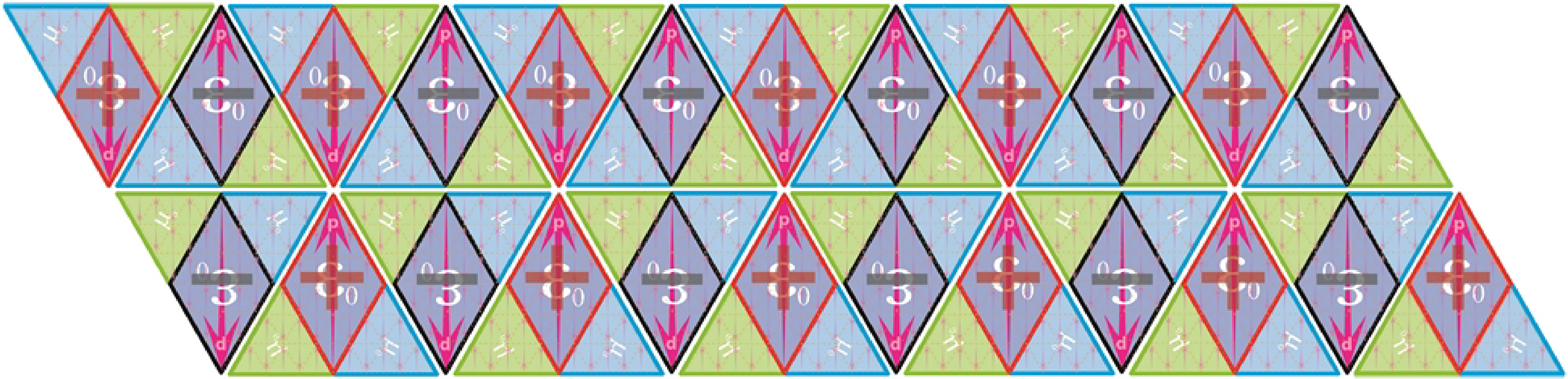

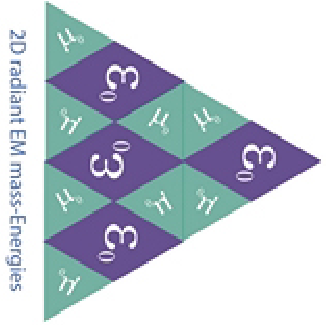
2D radiant EM mass-Energies
μ0
ε0

n6 - Neutral Tetryon Templates

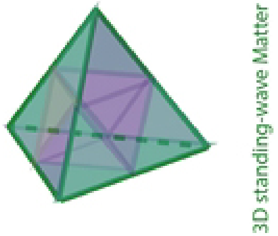
3D standing-wave Matter

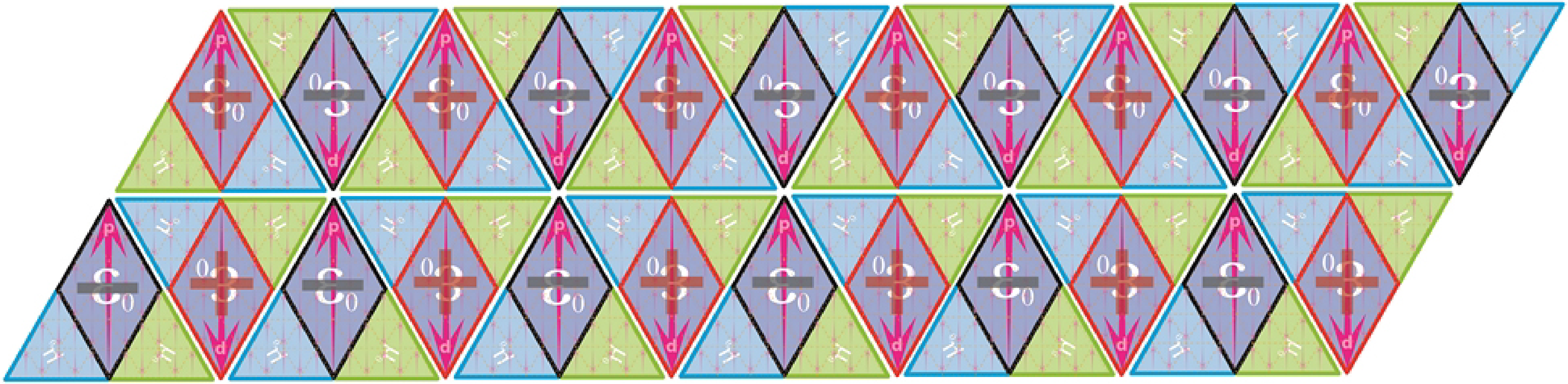

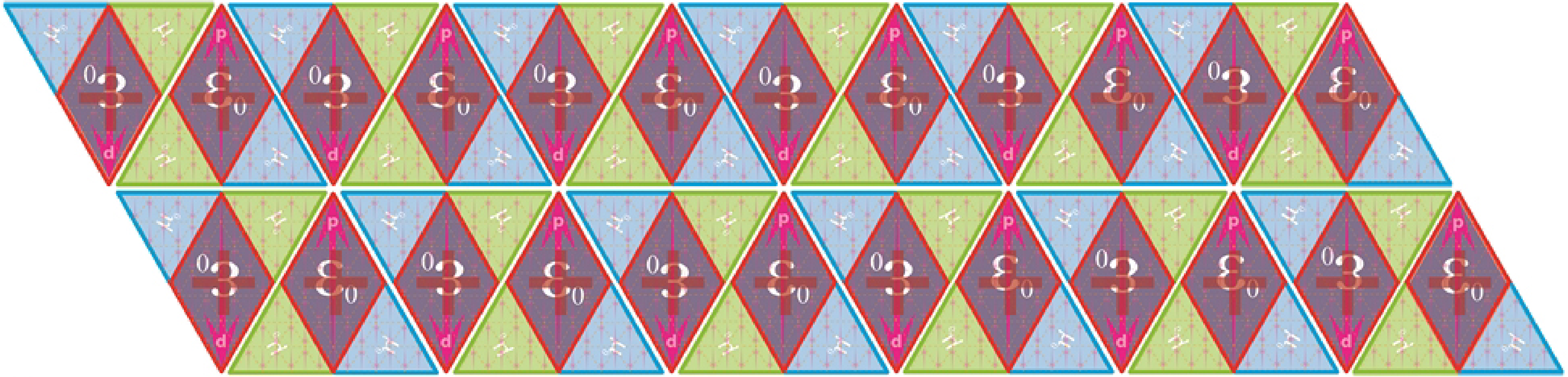
p
d
ε_0
μ_0

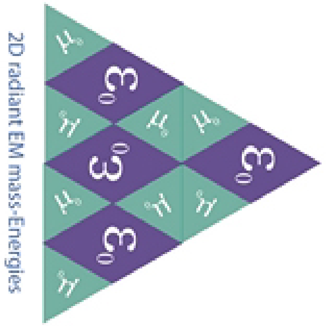
2D radiant EM mass-Energies
ε_0
μ_0

n7 - Charged Tetryon Templates

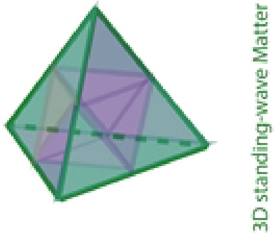
3D standing-wave Matter

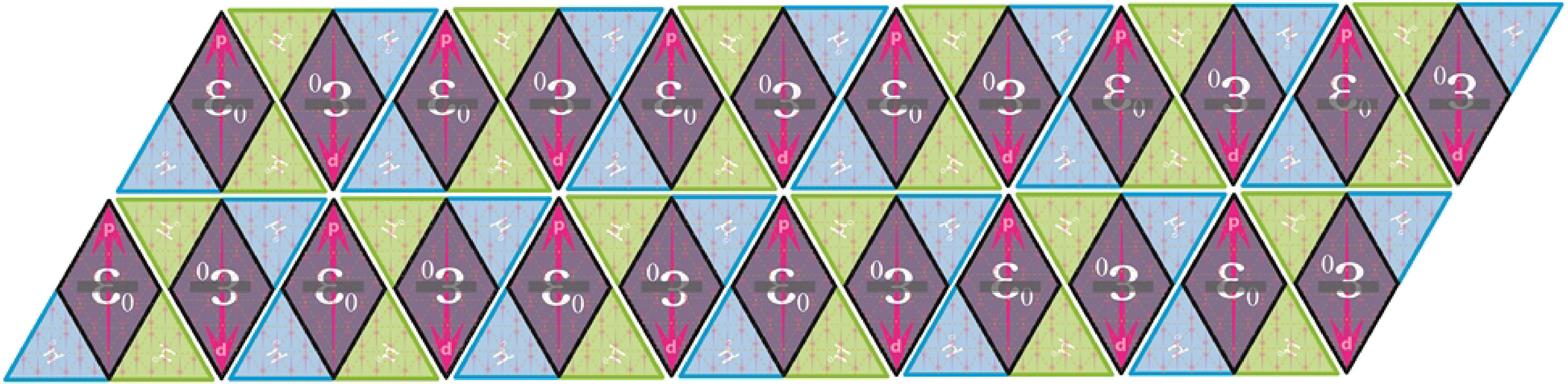
p
d
ε_0
μ_0

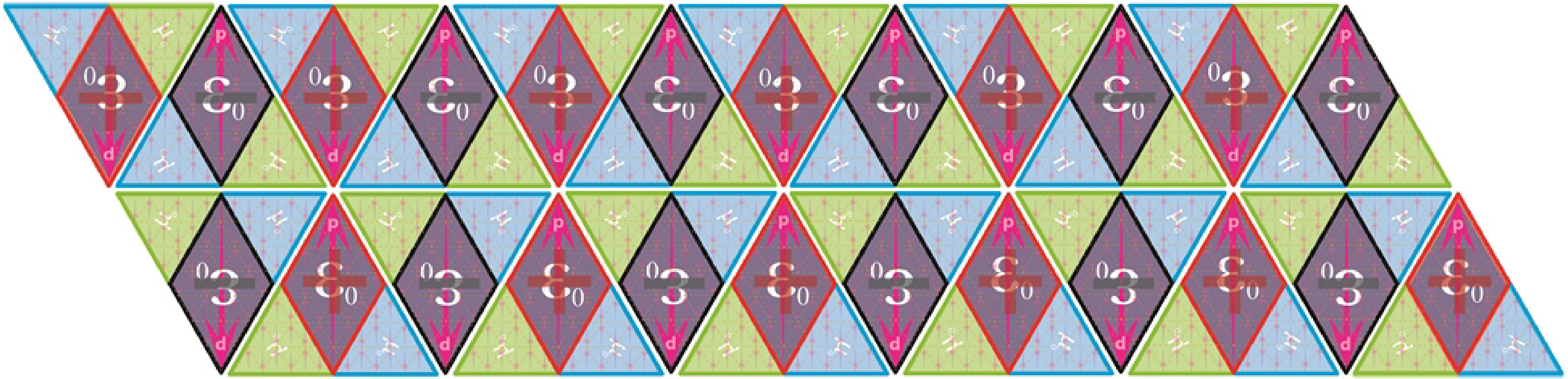

μ_0
ε_0
p
d

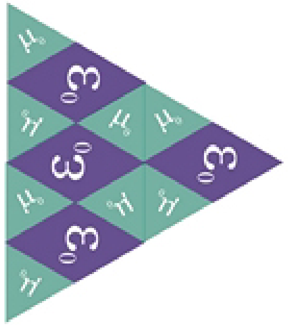

μ_0
ε_0

2D radiant EM mass-Energies

n7 - Neutral Tetryon Templates

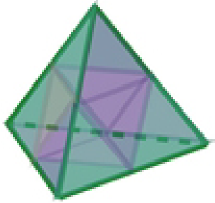

3D standing-wave Matter

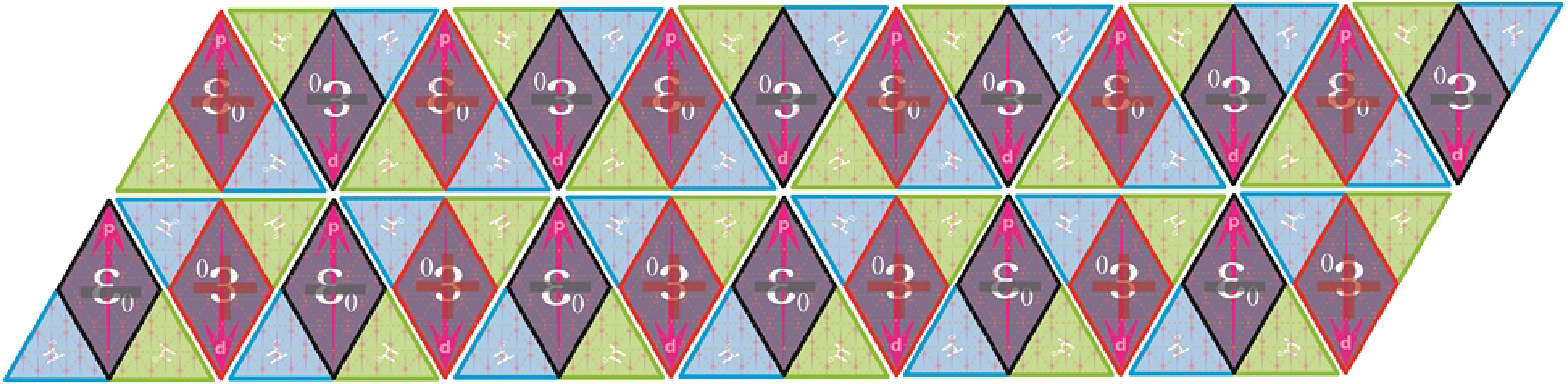

μ_0
ε_0
p
d

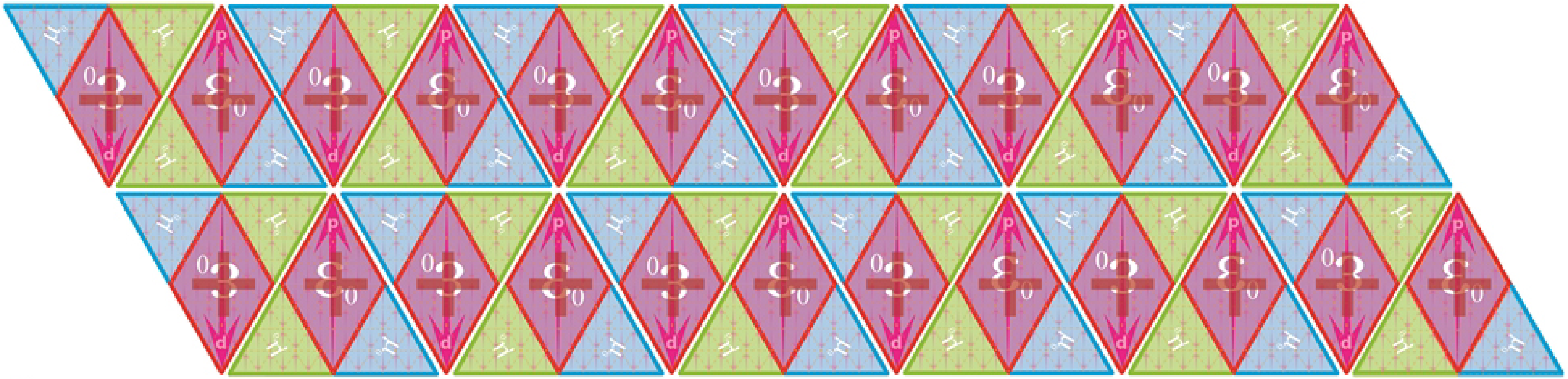

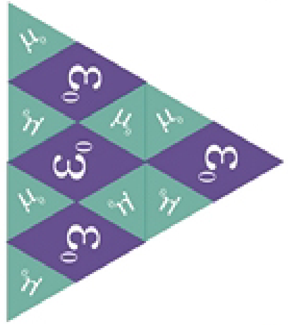

2D radiant EM mass-Energies

n8 - Charged Tetryon Templates

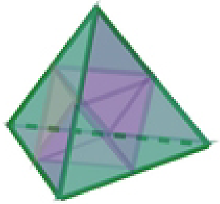

3D standing-wave Matter

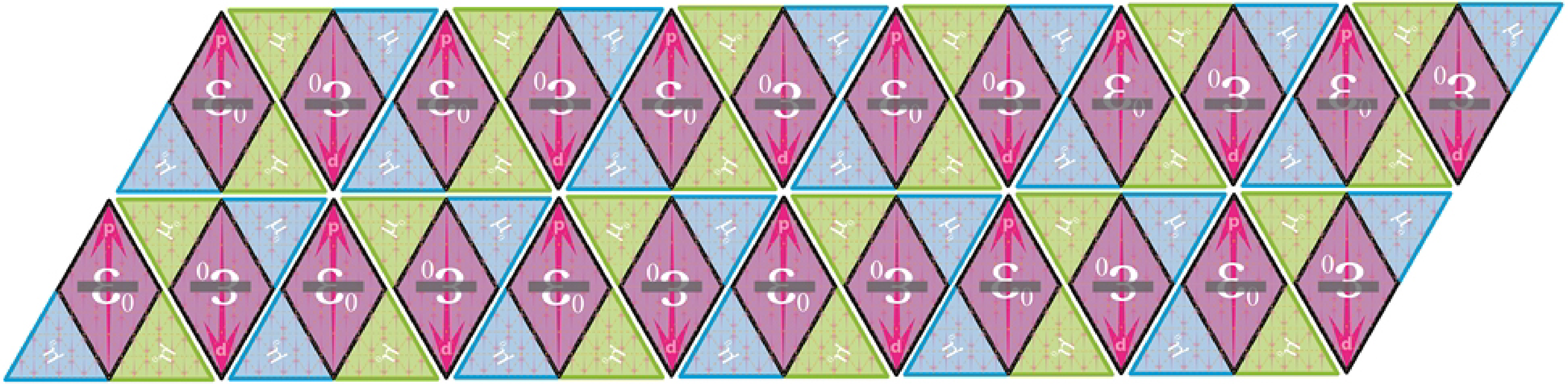

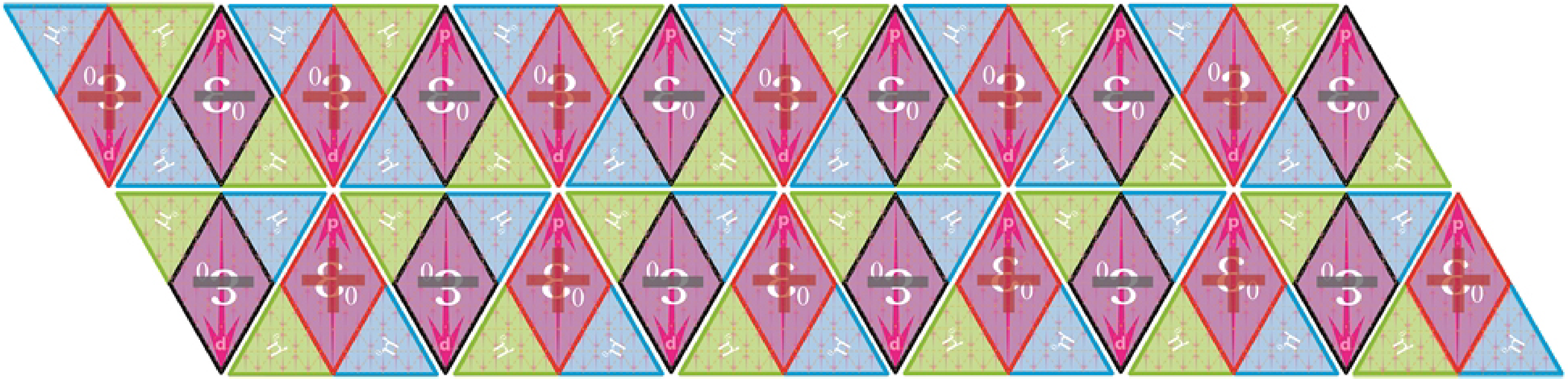

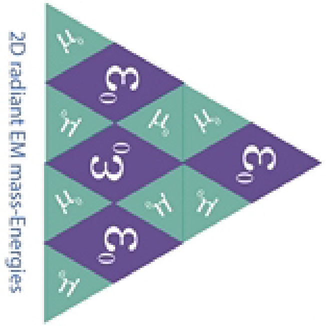
2D radiant EM mass-Energies
ε0
μ0

n8 - Neutral Tetryon Templates

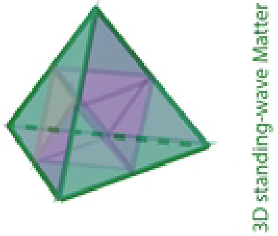
3D standing-wave Matter

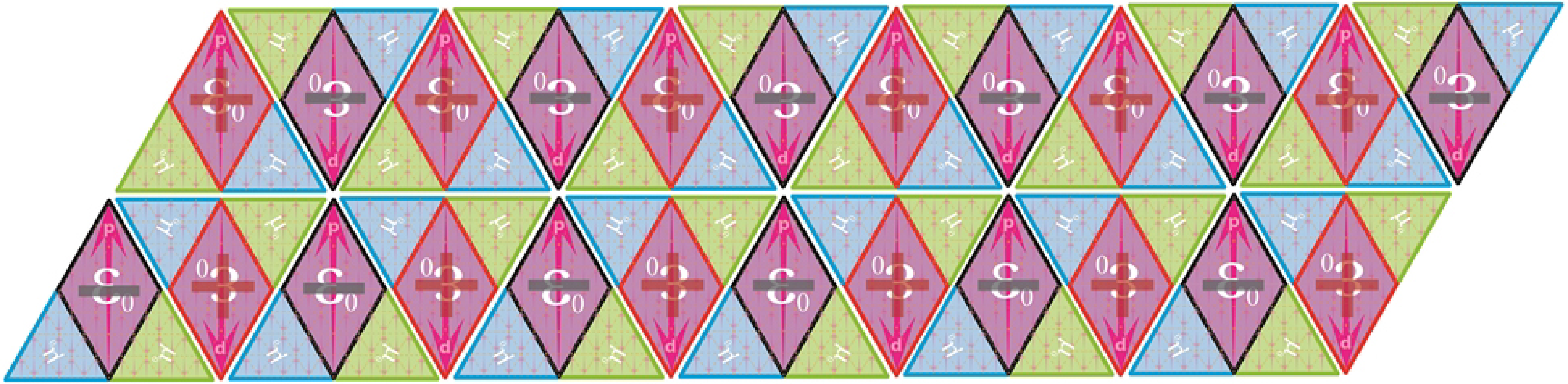

The Universal Order of Creation of MATTERS

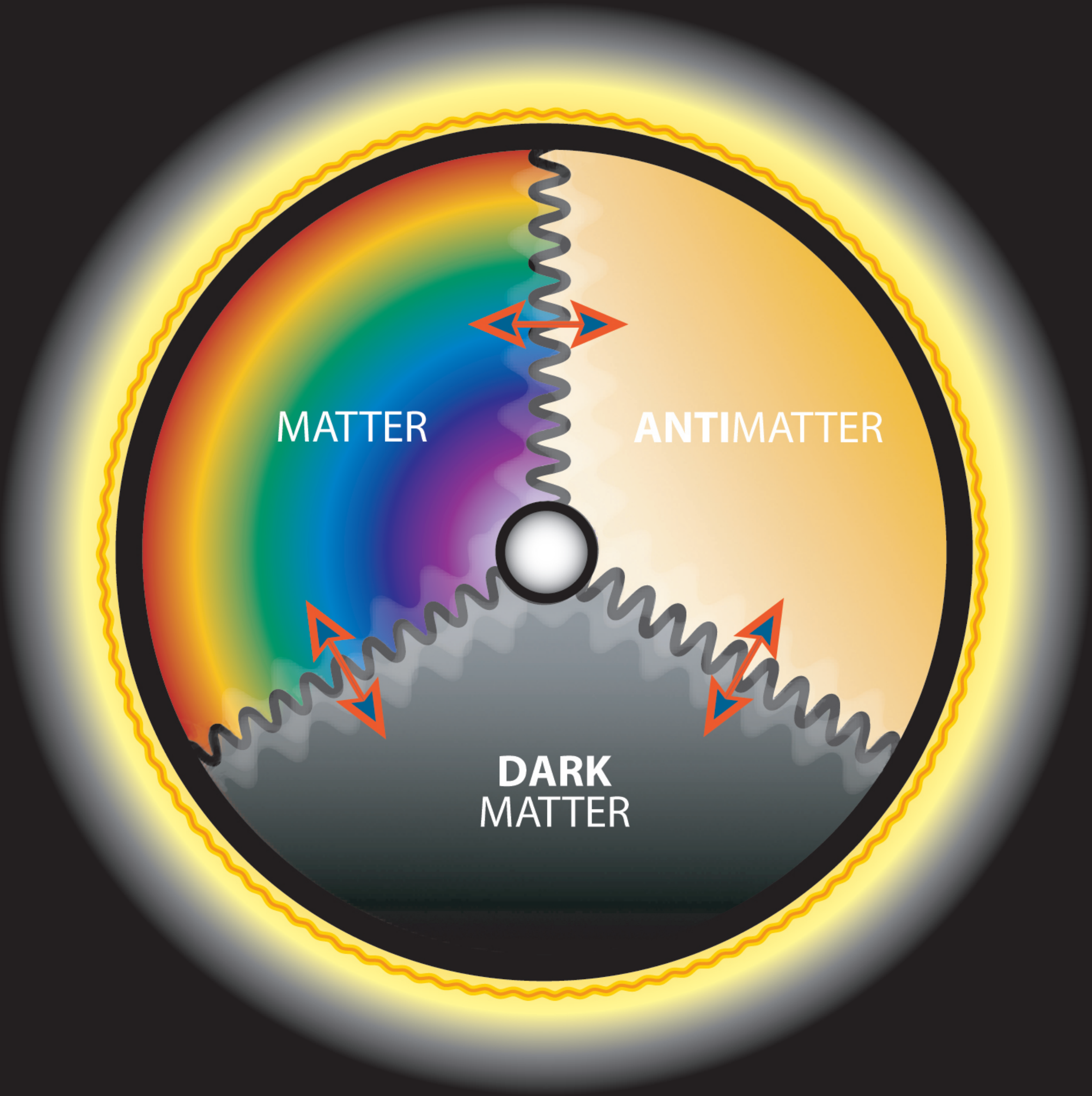
MATTER
ANTIMATTER
DARK
MATTER

by Nuclear Engineer M.T. KESHE

Kt
Editions

本电子书不是官方正版，

而是使用从官方购买的正版电子书，

截屏＋拼接＋OCR文字识别＋重新拼版制作。

因此难免出现错误的字或符号，

如果您希望获得原始的官方版本，

请购买正版，支持KESHE。

QQ群：30137653

Stichting
THE KESHE FOUNDATION
(The Netherlands)

ISBN 978-94-6087-002-6.

The Universal Order Of Creation Of Matters

ITNOTGOG

… a Reflection

Mans intelligence has reached levels that through the use of modern technologies that anyone can peer the knowledge he reads to verify its logic and validity.

I firmly believe that this work has to be understood by all Men and not judged by a few whose self-interest might, and has been, obstructing the release of such a knowledge as disclosed in this book.

This book is written in such a way that anyone from any background field of science, and interest in the workings of all Matters of the universe, can understand the full new knowledge and be their own peers in receiving such a disclosure, which is at the edge of science and technology.

In this book a synthesis is given of the *understanding of new first principles in physics* that are gathered for over more than thirty years of independent research and development, where several working prototypes and numerous tests with different loadings have confirmed their correctness and operativeness.

Mehran Tavakoli Keshe
2009

Copyright notice: © Mehran Tavakoli Keshe 2009. All rights reserved.

Nothing of this e-book may be replicated, simplified, translated and/or made public by means of print, microfilm, tape, audio-recording systems, electronics, optical and solid-state media, electronic storage systems, by network publishing or by any other ways without prior written authorization of the publisher or the author.

ISBN 978-94-6087-002-6

Illustrations and Layout: Dirk Laureyssens

Publisher: Stichting The Keshe Foundation
Registered Office: Jubileumplein 3, NL-6161 SR Geleen,
The Netherlands
Registration Number: KVK 14089728.
Web site: www.keshefoundation.com
E-mail contact: info@keshefoundation.com

Forewords

In Mans technological and scientific evolution over thousands of years, conversion of the Earthly state of matters and energies from one state to another has been the key to his progressive success.

For man to progress further in the path of this evolution, man has to learn, understand, and unravel the powers of the initial fundamental particles of the Matters.

In the following pages, simple new theoretical principles and results of tests done to unleash the powers of the initial fundamental particles forces will be released and discussed.

In my writings, the simplest form of language is used so that every man, from all corners of the world, can read and will be able to comprehend the fundamental orders of their physical creation.

Therefore, every man on Earth will have the same groundwork knowledge and opportunity to build his own environment and so be in control of his future and destiny.

This simple transfer of knowledge gives every man, by understanding the principles of his world of creation, the possibility to utilise and enhance the fundamental powers of creation for his advancement in his path of evolution.

Through appropriate use of this new knowledge, one can generate as much clean power as one needs for survival at any location in the universe.

This is energy production capability, without the production of any further new wastes to damage the environment and to endanger the future generations' chances' of survival.

This new knowledge allows man to produce food and medicine through the natural process of the construction of his cells' initial fundamental plasmas.

Therefore, no child, woman or man should go hungry and rely-upon others for his survival.

This new knowledge gives man the opportunity to produce all his materials' needs, without him being tied-up to the Earthly resources, or environmental conditions, or the place or a position in the universe, wherever that might be.

This new knowledge brings about the capability for man to be able to move and travel in his environment and in the universe with universal magnetic fields protection, in any atmospheric or space conditions.

This new knowledge, gives man the freedom of motion irrespective of the medium whatever his medium, and wherever he might be in the universe. For man to be able to use systems for travel that is at all time in protected magnetic fields levels and gravitational field levels as on the Earth. As his physical body has become accustom to the atmospheric condition around his home planet.

Which up to now, this has been a dream and unattainable in all dimensions, with his present level of scientific development and knowledge.

Men will build on this new knowledge, and they will add their own scientific and technological flavours and unique additions and discoveries to overcome their present natural problems and limitations.

Our hope is that, Men, irrespective of their race, colour of skin, creed or religion, will use this new knowledge collectively and in unity, for advancement of the human race. We pray that, Man uses' these new knowledge and information's for creation of peace and justice among their fellow men and other of Gods' creations and creatures, whom they will come to be in touch with, irrespective of their appearance, technological progress and intelligence.

Mehran Tavakoli Keshe

Table of Contents

… A REFLECTION 3

FOREWORDS 5

TABLE OF CONTENTS 7

INTRODUCTION 11

Rays and plasmatic magnetic fields 11
Magnetic and gravitational fields (Magravs) 13
Plasmas 14
Matter and matter 14
Initial Gravitational Fields 14
Reactor technology 14
Fission and Fusion by Plasma Dilution 15
Lift and Motion by Magravs positioning 17

CHAPTER 1 20

HISTORICALLY UNEXPLAINED PRINCIPLES 20
1. The creation of Magnetic Fields of the Earth 21
2. Creation of the Gravitational Fields of the Earth 23
3. The source of the Heat in the Earth Cores 27
4. What causes the Rotation of the planet Earth? 28
5. The difference between Matter and matter 30
6. How motion of matters and Matters are created? 31
7. How Light is created? 32

CHAPTER 2 34

INTERACTION OF PLASMATIC MAGNETIC FIELDS (PMTICS) 34

CHAPTER 3 37

THE INITIAL FUNDAMENTAL PARTICLES OF THE PLASMA 37
Creation of Matter of plasma through interaction of pmtics. 38
Creation of the Antimatter of plasma 41
Creation of the Dark Matter of plasma 43
Creation of the torus field force 45
Initial structure of Matters of plasma 47
Transition-zone pmtics 48
Initial Fundamental Plasma 49
Decay of initial fundamental plasma 50

CHAPTER 4 53

MATTERS' OF PLASMA AND THEIR TRANSITION OF STATES 53

CHAPTER 5 60

DEFINITION OF ENERGY THROUGH PMTICS 60
Measure of a mass 62

CHAPTER 6 67

THE TOTAL ENERGY BALANCE EQUATION OF PLASMA 67

CHAPTER 7 73

THE CREATION AND SPEED OF LIGHT 73

CHAPTER 8 83

DARK LIGHT PHENOMENA 83

CHAPTER 9 84

RELATION BETWEEN GRAVITY AND LIGHT 84

CHAPTER 10 90

THE ANTIMATTER 90

CHAPTER 11 92

THE ANTIMATTER ENERGY 92

CHAPTER 12 93

THE DARK ENERGY 93
Steps in the creation of Dark Energy 94

CHAPTER 13 98

THE DARK MATTER 98
Dark spots on the Sun and Black Holes 100

CHAPTER 14 104

VIRTUAL PARTICLES (VIRTUAL MATTERS) 104
The Casimir effect 105
Van der Waals forces 106

CHAPTER 15 107

INTERACTION BETWEEN MATTERS 107
Interaction of Matter with Antimatter 107
Interaction of Matter with Dark Matter 108
The interaction of Antimatter and Dark Matter 109
1) Scenario 1 109

2) Scenario 2 ... 111
Interaction of plasma with plasma ... 115
Electron interaction: ... 115

CHAPTER 16 ... 122

The fundamental difference between Inertia and Gravity ... 122

CHAPTER 17 ... 129

Plasma's and Nuclear decay ... 129

CHAPTER 18 ... 132

Neutron-Proton, Electron-Atom ... 132
The decay of the neutron ... 133

CHAPTER 19 ... 140

Motion of the Electron around the Proton ... 140
Universal Magravs positioning principle ... 143
Universal method of motion ... 143

CHAPTER 20 ... 146

Molecules ... 146

CHAPTER 21 ... 148

Plasma Dilution Technology ... 148
A soup of weak plasmatic magnetic fields ... 148
Disentanglement ... 150

CHAPTER 22 ... 154

Matters and Magravs systems ... 154
matter propulsion technology ... 154
Matter gravitational technology ... 154

CHAPTER 23 ... 159

Matter Pmtics in reactors for lift and motion ... 159
Principle concept of upward or separation motion ... 159

CHAPTER 24 ... 163

Passive Magnetic Fields ... 163

CHAPTER 25 ... 164

New understandings and New Technologies ... 164
Pmtics reactors for production of energy ... 168
New type of Generators ... 169
Seed of thought ... 174

CHAPTER 26 176

MATTERS EXPERIMENTAL RESULTS AND OBSERVATIONS 176
Radiation 177
Weight reduction, motion and lift.. 177
Production of matters 181
Dark Matter technology... 182

CHAPTER 27 183

DISCUSSIONS AND CONCLUSIONS.. 183
Mass of the Matters and Mass of the plasma..................................... 185

CHAPTER 28 191

INSIGHT INTO THE FUTURE.. 191

REFERENCES.. 197

PUBLISHED AND UNPUBLISHED WORKS BY M. T. KESHE ... 200

INTERNET LINKS.. 211

INDEX 212

Introduction

In the progressive path of mans' technological achievement, it has been proven that the human race has not been able to create and produce fundamental energy systems similar to the ones which are operating in the spans' of the universe.

Nevertheless, he has managed simply to convert one form of energy to another and he has made use of the small surplus energies released by these methods to his advantage.

In the universe, the structure, operation, appearance and existence of the plasma and an atom, does not so much differ from the relationships and interactions of their larger counter parts as the molecules, solar systems and galaxies.

Where, as in the ***present*** science and state-of-the-art technology, the lower orders of particles of construction of an atom in its nucleus are considered to be made of quarks and gluons.

The world of science has never looked at and into the ***progressive construction of a plasma*** (Chapter 3) on the same basis and principles' as other larger matters and fields in universal orders of creation and motion.

What this simply means, is that scientists over centuries have made some limited and un-necessary self-imposed conditions in looking at the physical realities of Matters and matters in their surrounding environment, for them to be able to compare and draw parallels from these phenomena and principles to understand the criterions by which the initial seeds of plasmas are set in the universe and to understand the methods by which different forces operate within them.

From the basic knowledge of what is in hand in the world of science at the present, it is not too hard to discover what are the real elementary particles in the universe, and find an *alternative and realistic way* of how these fundamental particles of plasma come together to create ***the initial seed of plasma***.

It is important to note that the existence of dynamic and different strengths, speed, velocity and densities of ***magnetic fields***, and their interactions, are the **cause** and the **foundation** of the existence of different **Matters** within the universe. In the universe, *collections* of *loose and weak magnetic fields* have initiated the creation of, and finally become the initial components of the Matters of plasma.

Rays and plasmatic magnetic fields

At the same time, we consider "rays" as individual and directional magnetic forces. We consider "fields" as the same ray's magnetic force as it is influenced by other rays and magnetic field forces, which the ray makes a dynamic zone of influence in its path.

We consider the same phenomena in principle with respect to plasma and ***plasmatic***

magnetic fields, where, a plasma of magnetic fields is considered to be a collection of such rays and fields of rays.

The collective interactions of weak plasmatic magnetic fields in the universe are considered to become the foundation and the cornerstone of the creation of initial fundamental particles.

In the universe, the same and regular specific sequential and consequential interactions of lower orders of magnitude plasmatic magnetic fields strength are always considered to lead to the formation of the plasma and its sub-Matters' components.

Throughout this book a clear distinction has been tried and been made between the magnetic field and plasmatic magnetic fields.

The **Plasmatic magnetic fields** are considered to be of *collection* of near *similar dynamic rays of* magnetic fields *strength*. We refer to the Plasmatic Magnetic Fields as "***pmtics***", *as Magnetic fields, which were detached from their source of creation*, hence, the use of the term "plasmatic" for these magnetic fields. For clarity: In this book the term "**plasmatic**" refers to a "collection" of dynamic magnetic fields, and **NOT** to the ***state of a plasma*** in its common term use that refers to a *state of dynamic protons*.

At the same time, pmtics interactions with other pmtics are considered to create effects the likes of gravitational fields and Magnetic fields of Matters, plasmas, planets and so on.

These magnetic rays or fields and their collection as plasmatic magnetic fields, become known as the **Initial Fundamental Particles.** Therefore we consider that magnetic fields to be the **"real" fundamental particles**.

The present science considers that the 'elementary particles' to be of 'quarks' and others. With the new understanding of creation of Matters and forces within the universal order of creation, in this book when the term **'elementary particles'** is used, this means that the **'magnetic fields'** are considered to be the **new** *'elementary particles'* and their forces. The existence and interaction of these magnetic fields leads to and creates all other effects that one sees and observes in the universe, like creation of Magnetic and gravitational fields which leads to creation of what I call different ***Matters*** *(Matter, Antimatter and Dark Matter)* of the plasma, or what presently is known as quarks and so on of the plasma. Where, we consider the Dark Matter as an integral and part of the same normal Matters of existence and as an important part of the construction of the initial fundamental plasma in the universe, and we do not consider the Dark Matter as an exotic and distant hidden Matter.

From this point on in this book, we refer to 'magnetic fields' as the ***'elementary particles'*** and *not consider quarks and others as elementary particles any further*, I consider that these quarks of different colour, with up and down notation and spin are all created as the outcome and consequences of different interactions and attraction of these 'magnetic fields' or 'plasmatic magnetic fields'. As quarks themselves are now known to be made of simpler particles of 'magnetic fields', then quarks cannot be considered to be the elementary particles.

In this book will be explained how 'magnetic fields' become to be the creators of quarks and other effects and forces within the plasma of proton, then the **'magnetic fields'** become the '**new elementary particles'** in the science of physics.

Further in this book it is explained how *different strength of these fields* leads to the creation of different **Matters and their forces** as has been observed by world of science.

Furthermore, these weak collections of pmtics and their interactions with magnetic fields and turbulent magnetic fields forces in different parts of galaxies manifest themselves as the seeds of atoms, stars, galaxies and so on.

Magnetic and gravitational fields (Magravs)

When we write **M**agnetic field (with capital M), we considered that these fields refer to the Magnetic field of celestial objects, like a Star, planet, etc. Where this Magnetic field is always accompanied with gravitational fields of the object.

Where we consider that ***gravitational fields and Magnetic fields*** cannot exist in the universe without each other, and that they are produced simultaneously. Through the same dynamic materials and Matters in the same region. Due to the construction of other Matters in their catchments area, these entities could have a different value than the other. As the Matter/matter content of the object dictates which will be stronger in strength and value than the other in that given position. The only known and possible objects that its gravitational and Magnetic fields could be equal in value strength, in majority of the cases in the universe, are objects of mono-material (like some stars), with pure hydrogen content. There are cases in the universe(s) that mono-matter of objects can be like uranium or heavier elements. These objects of heavier mono-matters are usually the producers of stronger gravitational fields and Magnetic fields, by several orders of magnitudes than the hydrogen mono-matter objects. It is assumed that each universe cannot possess more than one heavy material mono-magnetic star.

It has to be noted that the co-existence of these two fields always in interaction with any other magnetic fields or plasma of fields always lead to the creation of a ***magnetosphere***, which the boundary of the magnetosphere is the maximum point of balance or equality between the two field forces. Any variation in strength of one of the two fields in respect to the other will create different effects, for example such unbalance field between the two fields can be seen in Mercury, which has no atmosphere, and such a misbalance can also be observed in most moons of planetary systems.

In this book, as one field cannot exist without the other, for representation of the **MA***gneti*c and **GRAV***itational field***S** we use the term: "**MAGRAVS**".

Plasmas

We consider that “ The **Plasma** in its **collective existence** is a combination of different initial fundamental Matters, which these Matters are created through the interaction of at least two or more of the same strength plasmatic magnetic fields (the initial fundamental particles)”.

Plasmas (Chapter 3) are considered to be collection of combination of interaction of pmtics forces of Matters’, Antimatters’ and Dark Matters’ and other magnetic fields forces, all at the same time and within an overall sphere of the confines of the plasma.

Further, in the process of interactions and collections of these initial pmtics with each other, in specific combinations, these leads to creation of atoms, molecules, matters and so on.

In this disclosure, simple connections between *Initial Fundamental **Particles*** and *initial Fundamental **Plasmas*** are explained and discussed.

These specific collections and interaction of initial fundamental *particles* and initial fundamental *plasmas* and their interaction with other magnetic fields manifest themselves as **states of matters**, these are purely magnetic fields densities and compactness environment dependent, and they present themselves as different states of matter (gases, liquids and solids).

Matter and matter

In the following sections, we clearly distinguish between *states of matter* (gases, liquids and solids), denoted as “**matter**”, and *state of the Matters* (Matter, Antimatter and Dark Matter) denoted as “**Matter” or “Matters**”.

Initial Gravitational Fields

The interaction of different magnetic fields of different strengths, upon each other, is considered to be the creator of initial gravitational fields for the start of the creation of the initial seeds of the Matters of the plasma (Fig. 8 B).

We consider that pmtics interact to create the attraction field or gravitational field.

Reactor technology

With the new knowledge and new first principles which are disclosed in this book about the creation of Matters’ and plasmas’, techniques acquired and reactor systems have been designed, developed and tested, it has become possible to replicate the pmtics of plasmas’ environment that allows different Matters of plasma to be released from their interlocking magnetic fields forces within the confines of a reactor.

Further on, it has become possible within the structure of these reactors, to manage and control, these unlocked fields and Matters of the plasma, for specific predetermined functions and applications.

In these new reactors, in being able to weaken the Magravs of the plasmas (Coulomb magnetic or energy barrier) and by allowing the Matters of the plasma to be more free to move within the environment of the reactor. Where, these reactors were designed to facilitate the reduction in plasma magnetic fields forces and eventually to create an environment that Matters of the plasma can freely move in a soup of Matters, pmtics and rays. Where, through the operation of these types of reactors the famous Coulomb barrier becomes irrelevant in the overall structure of the soups of Matters binding magnetic fields of the plasma.

With these new methods of the manipulation of the structure of the Matters' of the plasma it has become possible for the pmtics forces of the plasma within these reactors to be unleashed, and with plasmas' constituent substructure loosen, it has become possible to manage to attain new powerful effects and properties from these rather weak initial fundamental magnetic fields of the plasma. Where up to the present time, these have been the unknown knowledge and methods in the world of science.

Within the structure of simple nuclear-based systems that have been developed, ***conditions*** *are created in which simple weak ambient dynamic pmtics, similar to the internal construction of plasmas, are generated.* The technology for creation of conditions similar to the internal structure of the plasma and methods which have been specially developed, that allows the loosening of the magnetic fields structure of the plasma, this will change the course of present nuclear fusion, and allows scientists in this field to achieve their goals *rapidly* and *at the fraction of cost* of present systems.

In the new state-of-the-art technology much simpler, softer and more practical ways, very similar to conditions as in the universe, are employed to attain fission and fusion of different Matters and plasmas.

During the reading of this book the reader will discover various future applications of this technology, so it will not surprise the reader that we have applied for several international patents to cover all aspects of the topics, principles, methods and systems mentioned in this book and the ones to follow.

Fission and Fusion by Plasma Dilution

The present methods of fusion and the Antimatter production technologies apply strong magnetic fields forces or high-speed rotation of the plasma, with low or limited success. As the real principle of the construction and structure of these sub-plasma Matters have never been understood correctly.

In our new state-of-the-art science and technology, one creates a control-conditioned environment within the confines of a multi-core reactor, which facilitates and allows the plasmas' Matters to ***disentangle*** themselves into their sub-Matters and pmtics components.

In these reactors the effects of the newly freed principal Matters of plasma's interactions with each other are utilised to create new and more powerful internal and external Magravs' forces, within and in respect to other plasmas and Matters in multi-core reactors system environments'.

In the new state-of-the-art technology reactors, after achieving desired goals, the system parameters can be changed so that the system can allow the same separated parts of plasmas' Matters to return to their original properties and regroup as the original plasmas.

This new method of loosening and the use of plasmas' Matters within the confines of a reactor is to be called the ***Keshe Theory*** of the "**Plasma Dilution**" and the art to achieve this "Plasma Dilution Technology"(Chapter 21).

This technology means that it is now possible to create an environment of weak magnetic fields within the cores' of a reactor, where the plasma is allowed to be opened-up and be released of the interlocking magnetic fields forces between its different constituent Matters.

The new dilution method in opening-up the plasma is a more practical and gentler methods for fission or fusion of the plasmas and/or its sub-Matters' components, for utilisation of their effects and properties.

In the present state-of-the-art technology in physics, scientists have chosen harsh method systems in fission industry. For fusion, they have chosen high speed and by compression of plasma, utilising giant magnetic fields induced technology to achieve their goals.

They use accelerators to slam the plasma on to a surface, to extract the sub-plasmatic components like the Antimatter component of the plasma (6, 7, and 8).

In the new state-of-the-art plasma dilution technology, within the confines of the reactors, environments and conditions are created, which are like a soup of weak pmtics, similar to the internal structure of plasma's own environment.

In this new approach, any new introduced plasma into the reactor, new **plasma's *binding* magnetic fields forces** come into an environment and ***interact*** *with fields* ***similar to their own plasmatic magnetic fields strength*** to allow their more condensed fields of interlocking Matter, Antimatter and others, to loosen and operate more freely in the environment of the pmtics soup of the reactor cores.

Using this new state-of-the-art science and technology, achievements of astonishing effects has been reported over the past few years. Where, with the current state-of-the-art technology, to achieve these effects, is considered to be nearly impossible and literally a dream and against the present known laws' of physics.

These effects reported, like being able to lift a reactor system weighing several kilograms by the use of a few grams of gasses, utilises the properties of the Antimatter components of the plasma's interactions with each other within the cores of these reactors.

This new technology makes it possible to produce in a very fast and easy manner nanomaterials and their constituent pmtics, like nano atomic layers known as graphene (sp2), diamond atomic structure (sp3) and nano-oxides, at room temperatures and atmospheric conditions, as has been independently confirmed by reputable organisation in the west.

In fact, all that has been achieved is the application of the understanding of how to use the soup principle to undo the tangles of magnetic fields of the plasma, and release of its substructure Matters, like Antimatter, in a reactor.

Lift and Motion by Magravs positioning

By utilising the properties of the interaction of forces of the collection of pmtics, interlocking between the Matters of the plasma and/or their sub-magnetic fields, which are part of the structure of the Antimatter and the Dark Matter, it has become possible to create stronger Magnetic fields, or gravitational fields' forces within the confines of and in respect to the surrounding environment of multiple-core reactors'.

These Magravs, if strong enough to interact with similar fields in their environment and with other fields of the same components of plasma in other parts of the same reactors, these Magravs interactions' of the system and the planet can cause lift and motion of the mass of the system, which the reactor is attached to.

Scientists at the Fermilab (6) in the USA state that small amounts of Antimatter, as little as a spoon-full, can release enough energy equivalent for the launch of several shuttle flights. This laboratory has confirmed that they have managed to collect some of this Antimatter in the past forty years, using special plasma separation and containment methods.

Utilising the new plasma dilution simple methods', not only are the Antimatter components of the plasma separated but, at the same time the interaction of the magnetic fields of these Matters are used within the system for the creation of strong Magravs within and in respect to their environment, for creation of lift and motion.

The newly created Magravs in these reactor cores, in interacting with the surrounding planetary Magravs and in trying to find a new Magravs balance, for example, in respect to the Earth fields, causes the system to move.

This method of achieving system motion due to interaction of Magravs forces is to be known as the ***Keshe principle of Magravs positioning*** *(Chapter 19)*. This position is, the point where the combined gravitational and Magnetic fields strength of the reactor system and the Earth, find a new position, where they are in balance strength in respect to each other.

After achieving the required Magravs positioning, this being the new height and/or distance, in respect to another Magravs force like that of the Earths Magravs, the systems' internal parameters can be changed again to create a new Magravs strength within the reactor. At this point, a new Magravs positioning balance needs to be attained by the system in respect to the planet. This, forces the system to find a new Magravs balance in respect to the planet Magravs' strength, and consequently for the need for the system to achieve the new Magravs positioning in respect to Magravs of the planet.

In attaining a new Magravs positioning balance, this leads to the movement of one system, in respect to another.

Thus through this new Magravs positioning method, motion of crafts, cars and systems for energy production can be attained *without burning* any fuel or using the present propulsion technology for lift of relatively small payload into space.

Through Magravs positioning principles, this is how ***motions*** *of atoms, planets, and stars and so on, are created and achieved in the universe (Chapter 19).*

Through tests, we can confirm that, the motion of any mass for any Magravs positioning system is not size dependent, but mass dependent, where the smaller mass always tends to move in respect of the larger mass in the universe. There are some exceptions in respect to this rule of the thumb. As in Magravs positioning reactor systems, where gravitational field strength control within the cores of the reactors plays a major roll to achieve lift, motion and Magnetic field protection of the system.

As I have said before "no one has ever seen a rocket in the back of the Earth that has been causing its motion for billions of years".

Subsequently, if necessary, through the same changes in internal parameters of these new reactors, the components of the Matters' of the plasma are and can be allowed and be facilitated to regroup and return back to their original structural state of pmtics levels of Matters and composition, once the desired affects are achieved.

Through new understanding and new state-of-the-art technology and by simply following the methods and universal orders of creation of Matters, the components of plasmas' Matter, Antimatter and Dark Matter can be manipulated and utilised for effective lift and motion of objects, creation of energy, production of new Matters and so on.

By the release of this new knowledge, this does not mean that the paths chosen up to now by other scientists have been wrong; this is by no means so. Nevertheless, their achievements have been and are limited because of the focus on the use of the **matter**; it's state of gases, liquids and solids, and even a limited view on plasma (as the fourth state-of matter) rather than our approach in use of the **Matters;** it's state of Matters, Antimatters, Dark Matters with their different pmtics strength components of the plasma.

With the use of the matter, one has to burn fuel to overcome gravitational field forces to attain a new position in respect to the planet. However, "with *Magravs positioning systems* one uses the pmtics of the Matters of plasma (Matmags) to create Magravs forces within the reactor, for the reactor system to *move within* the Magravs forces of another object, like the Earth, for the system to attain a position in respect to the Earth Magravs forces."

We consider that as much as there are Magravs forces between Matters, the same is true for the existence of Magravs forces between the Antimatter and the Dark Matter components of the plasmas too. This principle applies the same in respect to the inertia of Matter as much as to inertia of the Antimatter and the Dark Matter components of the plasma.

That is to say, in the universe there are Antimatter Magravs forces and Dark Matter Magravs forces as well as inertia for all three Matters.

All Matters and their effects can be used for the creation of motion and energy production in conjunction with other matters and Matters of the plasma.

Our newly released knowledge is for achieving and reaching the same target as other scientists and institutes, but with new knowledge about the construction of Matters of the plasma, we are achieving more profound and stronger effects, in simpler and less complicated ways than have been chosen in the past.

We see this release of new knowledge and technologies as an addition to the rest of the past and present scientific works, and a breakaway from the complicated ways and harsher methods.

In this disclosure, one comes to understand how Matters and matters are made from the start of time (Chapter 3). Where, the method of creation of plasma and atoms in the universe is explained in a simple form.

Then one can see how to use the inner construction of Matters' pmtics to generate vast amounts of energy and learn to use magnetic fields principles of interaction that are the fundamental principles' in the universal method for creation of motion.

We have tested the Magravs positioning systems for motion, and the results confirm the correctness of the theoretical aspects of the approach. The future of space technology will be based on simple safe portable Magravs positioning nuclear-based reactors, which have been developed and tested.

CHAPTER 1

Historically unexplained principles

The main reason for the lack of understanding of simple processes of creation and control of Matters of the plasma is not so much the lack of knowledge in the scientific world in the past or present. Rather the lack of progress is by large due to, the peer review obstruction mechanism, which has been put-up, for some scientists, to protect their individual financial interests and titles, in their schools of thoughts in their field of science.

For example, despite having huge research budgets and human resources, the world of science in respect to understanding of the simple structure of the plasma and its fusion, has not reached substantial results in the past sixty years. Where, nuclear physicists and industry have been trying to cause fusion of plasmas of hydrogen atoms, in fusion reactors, which have been built, and are being built at costs of billions of dollars.

The structure and working of these state-of-the-art fusion reactors are built against the natural laws of physics and opposite to the way, that fusion of plasmas takes place in the universe. Where, these reactors operate with near zero-gravitational field forces within their structure in trying to achieve fusion. Where in these systems, the natural internal gravitational field forces needed for a simple fusion is missing.

What they miss as the natural gravitational field forces needed for fusion, they try to replace this with the use of strong external electro-magnetic fields, by the use of large massive coils and huge amount of energy to operate these systems, and then the they try to achieve fusion. This is like as they are trying to bake bread with only water, salt and heat, without having its main ingredient, the flour. The question to ask from these scientists, is simply that, where are the gravitational fields essential for fusion in these reactors, and where is it suppose to come from, if there is going to be a successful fusion chain process as in the middle of the Sun, which, these forces are in thousands of G‘s.

In fact, with the use of giant external magnetic coils, they squeeze plasmas into tighter balls and then they try to fuse two hard balls of plasmas together. This is the reason why, they have and will achieve very little, if the present misguided path of fusion technology is to continue.

Another point is that the present twenty-first century advanced world of science is lead to believe and to accept the outdated 19th century principles of creation of Magnetic fields through convection principles in the centre of the Earth. Even, when scientific seismological data’s (1, 2, 3, 42A, 42B) indicate that another explanation is needed.

For example, some peers consider that man in the 21st century is not intelligent enough to discover how gravitational fields of planets are created. According to these peers, the principle of creation of gravitational field forces and their control at present is beyond comprehension of other scientists.

Where, as new Magnetic and gravitational fields positioning systems, which have been developed and tested, confirms the ability to create lift and motion by just using the effects of Matters magnetic fields (Matmags), without burning any fuel or use of propulsion technology.

Within the present 21st century knowledge and proven science, there are a number of unacceptable points from these earlier centuries accepted principles like the creation of the magnetic fields of planet through convection principles.

New theories and principles for these misassumptions, unexplained and out-of-the-hat principles are explained and discussed in the following sections and chapters.

In the following four sections in this chapter, it is important to note that:

Information that have been obtained through the earth-quakes seismological data over past thirty years from around the world, these data points to the fact, that what has been assumed to be the ***solid inner core** of the planet* and to be made of one piece solid metallic matter (Fig. 1), this core in fact has *a **cavity** in its centre* (Fig. 1) (1, 2, 3, 4, 5, 41A, 41B). This newly discovered cavity in the solid core of the planet has been called by geologists the 'inner-inner' core of the planet and is assumed to be about at least over 500 km in radius in the centre of the inner solid core of the planet.

I have called this new core, "***The Caroline Core***" (Fig. 1, 2 and 3), to distinguish it from other presently known cores' of the planet (12).

1. The creation of Magnetic Fields of the Earth

The assumption of the creation of the Magnetic fields of the Earth by scientific world has been based on the understanding of the working of the dynamo principle of earlier centuries. Where, the process of the production of the current was accepted to start-off with the motion of high temperature fluid of matters in the outer core (Fig. 1), in the centre of the Earth by heating and liquidity of matters in this core.

Where electrons of atoms of matters are considered to be released or loosened from their atoms to be free to vibrate or float in this region, and as the rotation of the solid metal inner core (Fig. 1) of the planet interacts with these free electrons in the outer core of the planet, this leading to the creation of Magnetic field of the planet (Fig. 1 Field M).

If this assumption is totally correct, then how do other planets and stars create their Magnetic fields?

Stars like the Sun, possess Magnetic fields and yet do not have solid inner cores as the Earth does. The Sun definitely generates Magnetic fields, which does span beyond the

Earth and the Saturn.

The first flaw with this convection theory of electrons motion is that, we know now **that plasma of an atom** is a charged entity **and possesses much stronger charges** by several orders of magnitude ***than its smaller counterpart electron***.

Thus, the main Magnetic field force of the planet is more feasible to be created by the plasma charges rather than electrons charges. With the knowledge we now know, that even the electron is a plasma condition similar to the proton construction (Chapter 3, 17, 18 and 19).

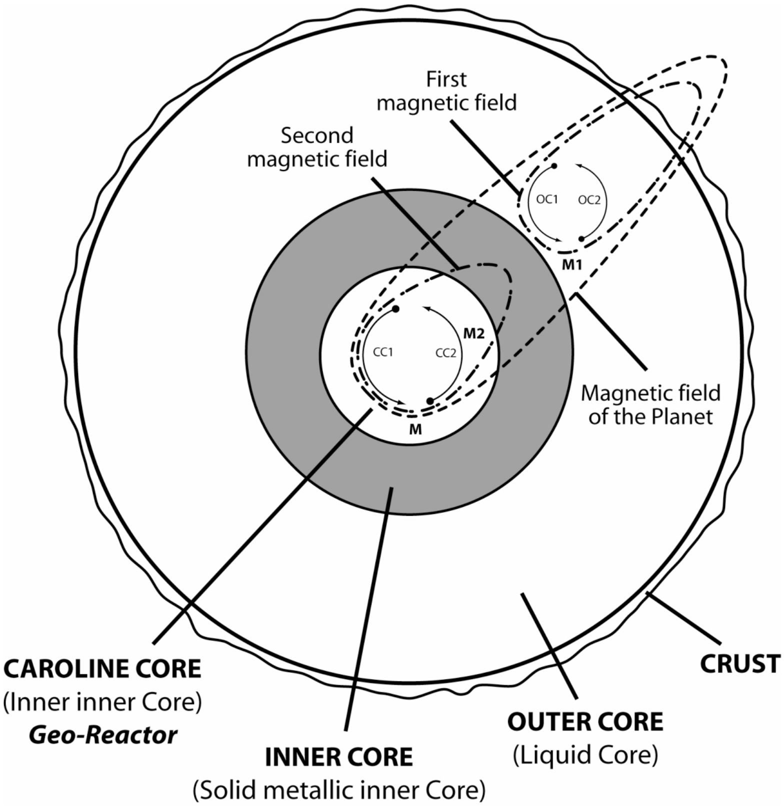
First
magnetic field
Second
magnetic field
OC1
OC2
M1
M2
CC1
CC2
M
Magnetic field
of the Planet
CAROLINE CORE
(Inner inner Core)
Geo-Reactor
INNER CORE
(Solid metallic inner Core)
OUTER CORE
(Liquid Core)
CRUST

Fig. 1: Two interacting plasmatic magnetic fields in the inner cores of the Earth lead to the creation of the Magnetic field of the planet.

Note: The Caroline Core, can contain plasmas, atoms (nanomaterial), molecules, different matters, hydrogen, potassium, uranium and plutonium due to nuclear decay, fission, possible fusion, ionisation of matters, leading to the creation of heat in this core, which is transferred through the solid inner core to keep the materials content of outer core in fluid state.

Through current knowledge in plasma physics and studies in the nuclear fusion industry, it is a known fact that dynamic protons (plasmas) of atoms are positively charged entities and do generate their own currents, through what is known as the Birkeland current principle. It is through this principle of creation of current and the dynamism of plasma in the centre of stars that main solar Magnetic and gravitational fields forces of stars are created, and not the convection principle.

Secondly, if we have plasma in the centre of the planet, which is possible, then the convection theory for the production of Earth's Magnetic field has limited validity, and it is possible that some of the Magnetic fields of the planet are partially created by the convection principles method too.

2. Creation of the Gravitational Fields of the Earth

The other major concept, which is an obstacle to the development of new knowledge, is to understand how gravitational fields of planets are created. This is a concept that has not been understood until now. Where, the creation and maintenance of gravitational field's forces in stars and planets have unnecessarily made been by the scientific world into the holy grail of science.

In fact, the creation of the gravitational fields of planets and stars in the universe is as simple as the creation of their Magnetic fields.

The hypocrisy with this is that the scientific world and an infant child (at school age) agree on one principle, that, two dissimilar poles of solid magnets, gets attracted to each other and pulls one magnet to the other, or they have gravitation in respect to each other. It is an accepted fact that, opposite poles of magnetic fields forces can interact to create attraction field forces or the gravitational fields forces in respect to each other and other matters around them.

Thus, it is correct, to assume that, to have a gravitational field in a planetary system, the gravitational field forces must be made similarly by the same principles of attraction and interaction of at least two magnetic fields, in the centre of the planet in respect to each other. Therefore, the gravitational fields of the planet must have been made by interaction of two or more sources of magnetic fields in respect to each other in the centre of planets or stars that can span beyond the physical boundaries of these objects.

By the reality of the existence of the Magnetic and Gravitational field forces around the Earth, this leads to the confirmation and understanding of a need for two independent regions that can create magnetic fields forces within the structure of the planet inner cores, that, their interaction leads to the existence of the gravitational field forces of the planet.

Consequently the Earth, being in possession of gravitational fields, this confirms and indicates to the existence and operation of at least two independent magnetic fields forces regions in the centre cores of the planet (Fig. 2, fields G1 and G2).

Where, the interaction of the two magnetic fields (Fig. 2, fields G1 and G2), which are generated in the centre cores of the planet, upon each other, leads to the creation of attraction or gravitational field forces of the planet, similar to the attraction fields of opposing poles with solid magnets.

It is a known fact that, planets and stars are made of different layers of different materials. Similarly stars are made of the same material (Hydrogen), which, due to variation in the compression and temperatures, layers are created in these objects, where the same matters in different layers create and behaves and show slightly different properties and effects; These different layers produce their own slightly different in strength magnetic fields force, compared to their adjacent layer magnetic fields strength. Thus, the layers of plasmatic or fluid materials in different depth of stars and planets, each layer create their own independent magnetic fields in the centre of planets or stars.

The magnetic fields created in each dynamic layer in the cores of these celestial bodies, due to the proximity have no alternative, but to interact with the magnetic fields generated in ascending or descending layers from them. The interactions between these magnetic fields in different layers are considered to be as or similar to a solid magnets state magnetic fields interactions.

In the case of stars and planets' inner cores, magnetic fields are created by and in plasmatic states or fluid states of Matters. Where, these flowing dynamic materials in the inner layers of celestial objects cannot physically lock to each other through the magnetic fields generated by each layer, as it happens with solid-state magnets. This being due to fact that atoms and molecules in these layers are continuously in motion and are pushed around by other forces and the rest of the Matters in their given layers or by other layers effects and forces.

Nevertheless, the magnetic fields created by each dynamic layer interact with the magnetic fields of the adjacent dynamic layer or layers. These interactions of magnetic fields in different layers creates attraction magnetic fields, as much as, repulsion Magnetic fields forces, in between the materials in the different layers.

Therefore in the same regions and by the same matters, **similar magnetic fields polarities** in different layers interact and two similar polarities magnetic fields forces (Fig. 1 field M1 and M2) repel each other. This repulsion magnetic fields forces (Fig. 1 field M1 and M2) becomes' the ***outwards field force pushing*** and what is known as the ***Magnetic field force*** of the planet (Fig. 1 field M).

Further, the *opposite polarities magnetic fields* in these layers at the same time, interact simultaneously and in a similar way, and the interaction of **opposite magnetic fields** polarities of matters in these layers (Fig. 2 field G1 and G2), leads to the creation of the different layers magnetic fields pulling or attraction in respect to each other. This attraction magnetic fields force (Fig. 2 field G1 and G2) becomes the ***inward field forces pulling*** field force and what is known as the ***Gravitational field force*** of the planet (Fig. 2 field G)(18).

Through tests done over several years with the construction of similar set-ups as in the inner cores of the planet, the ***Keshe Theory of creation of gravitational field*** of dynamic objects in possession of dynamic cores states that:

"***The dynamic outward' Magnetic fields forces of planets and stars are created by interaction of the same materials and in the same layers, as their dynamic***

gravitational fields forces”.

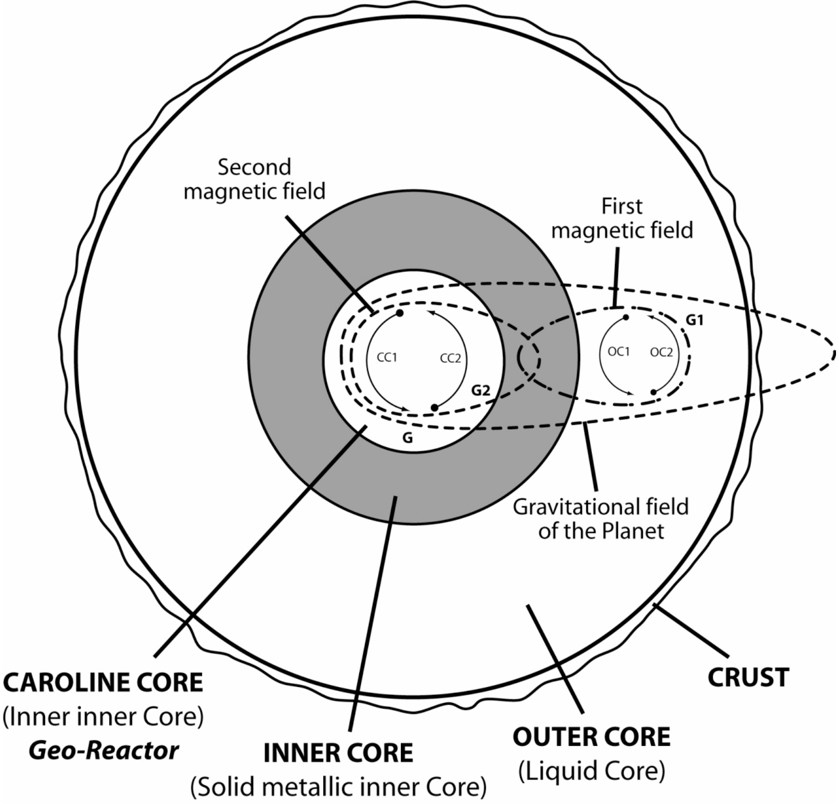

Second magnetic field
First magnetic field
G1
CC1
CC2
OC1
OC2
G2
G
Gravitational field of the Planet
CAROLINE CORE
(Inner inner Core)
Geo-Reactor
INNER CORE
(Solid metallic inner Core)
OUTER CORE
(Liquid Core)
CRUST

Fig. 2: Two interacting plasmatic magnetic fields in the inner cores of the Earth lead to the creation of the Gravitational field’s forces of the planet.

The ***Magnetic fields forces*** of the planets are made from the interaction of similar polarities of plasmatic magnetic fields of Matters in the dynamic layers of its inner cores (Fig. 1 field M) and these radiate themselves ***outward***. The ***Gravitational fields forces*** (Fig. 2 field G) are created by the interaction of the opposite plasmatic magnetic fields polarities of the same Matters in the same layers in the cores of planet, creating the pull or ***inward*** attraction in these celestial objects”.

This new understanding of interaction of multi-magnetic fields in the planets inner cores brings about a new definition, new understanding and new explanation of how gravitational field forces are created in centrally dynamic multi-inner-cores celestial objects in the universe the like of the Earth and the Sun.

Gravity can now be defined on the real basis and what leads to its manifestation, strength, existence and control.

Therefore, Gravitational field forces are magnetically based fields, and thus according to ***the Keshe Theory of Gravity***, Gravity is defined as:

“***The Gravity is, the measure of the interaction and attraction of two or more, magnetic fields forces or dynamic plasmatic*** strength created within the second system”(18).

This is the same as Newton's definition of gravity, with the extension however, that two Matters or matters cannot possess pulling forces nor have gravitation in respect to each other position, without the existence of magnetism.

Certainly, the attraction and interaction of pmtics forces changes with respect to position of two objects and they are distance and fields' strength dependent.

If the celestial objects do not possess magnetic fields strength that can interact with each other, then, how could they be attracted to each other?

Thus, gravitational field now can be understood and explained in its real terms of what creates it, and this is due to interaction and attraction of pmtics created within dynamic cores of celestial objects in respect to each other.

Secondly, as the gravitational fields of these planets or stars are magnetic fields strength dependent, one can now understand why their attraction is position dependent.

Therefore, *gravity in a simple form can be defined as; that gravity or gravitational fields forces are the measure of the interaction and attraction of two or more magnetic rays or magnetic fields systems of two or more entities in respect to each other's strength and position at any point in the Matter, plasma, matters, planets, solar systems, galaxies and universe*.

This new definition of gravitational fields is an extension to the present knowledge and due to new understanding of the plasmatic magnetic fields interactions between components of Matters (see Chapter 3).

Another proposition can be that: We further can assume that the cell of a human or matters of other objects, which all are made of atoms and plasmas and of magnetic fields can/are/or behave as the possessor of the second magnetic field in respect to any other plasmatic magnetic fields in their environment. Therefore, it is possible that pmtics of these cells and matters in their interaction with the dynamic pmtics of the planet can behave and replace the interacting double magnetic fields system principles necessary for the creation of gravitational phenomenon principle to operate.

Further it can be said, that any substance in possession of plasma or nuclei has gravitational attraction in respect to the planet.

If this is correct and applicable, then, this can be another reason why, a singular Magnetic field of a planet can be enough to create the gravitational field pull in respect to any object within the catchments zone of the Magnetic field of the planet's. That is to say, a single Magnetic field would be sufficient for the gravitation of plasma to the planet to be created and exist.

Nevertheless, there could be a flaw with this concept, as one need the gravitational and magnetic fields (Magravs) together to be able to have a constant free motion and position option of the two objects too.

In this book, we try to make a distinction between *magnetic fields of matters* and *Magnetic fields created by* dynamic interaction of two magnetic *rays*, two *magnetic fields* or *two pmtics*. Where, in the case of like **plasmas** and planets *Magnetic field* and *gravitational fields*, which they always accompany each other we use **capital M** as **Magnetic field**. In these references instead of writing **Ma**gnetic and **grav**itational field**s** we write in short as "**Magravs"** ***magnetic fields forces in respect to each others position***".

"Where, the Gravitational field force of one celestial body in respect to the other is a measure of the attraction and of the interaction of one plasmatic magnetic fields strength created within one system in respect to another plasmatic magnetic fields.

3. The source of the Heat in the Earth Cores

Another unanswered question in geophysics is that, where does exactly the heating of the inner solid core of the planet Earth comes' from that can keep the outer core materials in their fluid state? Secondly, how and through what method is heat generated in the inner cores of the planet?

These questions have not been addressed fully with clarity up to the present time in the science of the planet.

In general, it is an accepted fact that the Earth maintains heating of its inner cores. I consider that the Caroline core is a Geo-nuclear reactor (Fig. 1, 2 and 3) to be known as the **Geo-Reactor** (Fig. 2). It is not hard to assume that there are radioactive matters (41C) in this new central core of the planet (12), which leads to the heating of the inner cores of the planet. In general, nuclear geophysicists and myself agree upon this, up to certain degree, with difference in the content of the matter types and the process by which this heating is generated.

Where, geophysicists mainly consider that the heat in the Caroline core is generated through the nuclear decay of heavy elements, which have sunk into the centre of the planet due to their weight.

As a nuclear engineer having built several simple nuclear reactors replicating the structure of the inner cores of the planet Earth and achieving the same effects of creating Magravs similar to of the Earth, I believe, Matters, plasmas, hydrogen atoms, carbon atoms, radioactive matters and other matters collectively are parts of the mixture of the matters content of the Caroline's core (12).

One can now clearly explain that the major contributor to the generation of heat in the inner cores of the planet are, or can be due to nuclear decays, operation of plasmas, nuclear fission (41C) and possible nuclear fusion of materials in the Caroline core of the planet.

Materials in this region of such a high temperature are, and can be, considered to be in the fluid state (12) as has been considered by present geophysics scientists in respect to the condition in the outer core materials of the planet.

Matters in this region, due to their fluidity, can be considered to be able to produce their own dynamic magnetic fields (Fig. 1 field M2 and Fig. 2 field G2), due to the *dynamic motion of fluid matters* in this section of the planet and these fields can/and interact with the dynamic solid inner core of the planet (Fig. 2). This being in principle similar to what has been assumed by geophysicists, that to be happening in between the **inner core** and the **outer core** regions of the planet matters leading to production of Magnetic field forces of the planet at the present time. Secondly due to the *charged plasmas and charged electrons and their motion in this core, these could be partially responsible for the creation of the second magnetic field force of the planet.*

By the discovery of the second cavity, the Caroline core, in the inner cores of the planet, it becomes clear that the **second magnetic fields** (Fig. 1 field M2 and G2) can be, and are created in the very centre of the planet. Where, the solid core of the planet will act as a separating partition wall, between the two dynamic matter regions in the centre of the planet, namely the fluid matters in the Caroline core and the fluid matters in the outer core.

So, the holy grail of the gravity (18) can be explained by the principle of interaction of two dynamic magnetic fields sources in the centre of the planet as was discussed in the above section. Therefore, the interaction between two pmtics on either side of the solid core is the source of the plasmatic Magravs of the planet.

4. What causes the Rotation of the planet Earth?

Another unexplained and daily-accepted phenomenon is the rotation of the planet Earth. How is the rotation of the planet Earth created and maintained? What causes and dictates its constant speed of rotation of the planet, irrespective of where the planet is in its solar system?

The way and how the rotation of a planet is created, has never been explained in the past by world of science.

The existence of two independent pmtics forces (Fig. 1 field forces of M1 and M2, and Fig. 2 field forces of G1 and G2) on either side of the inner solid core of the planet (Fig. 3) can help to explain the next mystery in the world of geophysics, through the ***Keshe Theory of "The principle of the rotation of the planet"***.

Magnetic fields are known to possess *field forces* that can exert pressure on an object, and are capable of causing its motion.

The existence of, and the continuously maintaining and interplay of the two dynamic pmtics forces (Fig. 1 fields M1 and M2 and Fig. 2 fields G1 and G2) on either side of the inner solid core of the planet, partially goes to answer the sources of the fields that exert forces on the inner solid core of the planet that causes the motion of the inner solid core.

It has been explained that the interaction of two dynamic magnetic fields on either side of the inner solid core, leads to the creation of Magravs forces of the planet. At the same time the interaction and exertion of the same magnetic fields forces on the inner solid core of the planet put magnetic pressure on the inner solid core of the planet, causing the motion of the inner solid core of the planet (Fig. 3 fields G1 and G2). In the figure 3 the total and overall effect of Magnetic field and Gravitational field forces of each core on the inner solid core are denoted as G1 and G2.

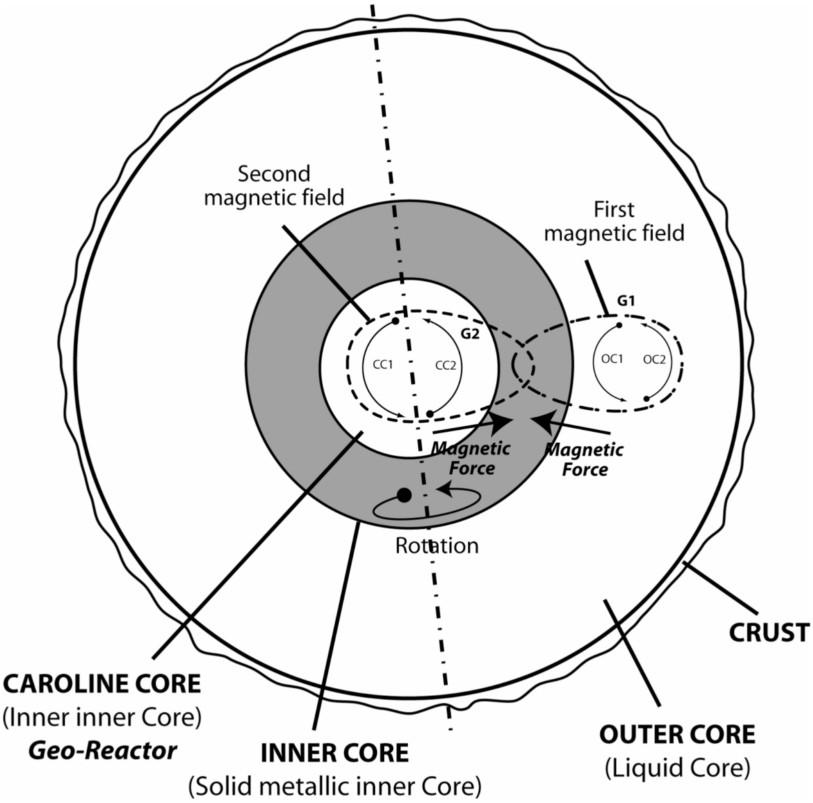
Second magnetic field
First magnetic field
G1
G2
CC1
CC2
OC1
OC2
Magnetic Force
Magnetic Force
Rotation
CRUST
CAROLINE CORE
(Inner inner Core)
Geo-Reactor
INNER CORE
(Solid metallic inner Core)
OUTER CORE
(Liquid Core)

Fig. 3: Interacting plasmatic magnetic fields in the inner cores of the Earth leads to the creation of the rotation of the planet.

As the solid inner core confines' the materials in the Caroline core (Fig. 3) and the outer cores materials are confined by other matters at the outer boundaries of this core (Fig. 3), these two cores are considered to be holding a constant volume and weight of matters. Thus, it is possible to assume that the magnetic fields forces created by each core will be constant and these magnetic fields forces will be maintained on either side of the inner solid core of the planet for a long period of time (billions of years).

Hence, the *balance of magnetic fields forces* exerted on the metallic inner solid core, by both magnetic fields on either side of the inner solid core, will determine the speed of rotation of the solid core of the planet.

Consequently, through the principle of viscosity and friction of the matters inside the planet, the final speed of the rotation of the whole planet is considered to be set and be due to the motion of the inner solid core of the planet. Which is carried out to the outer layers of the planet and becomes the final rotational speed of the planet.

Therefore, the speed of the rotational motion of the planet is determined by the planet's internal cores magnetic interactions upon the inner solid core of the planet and by large, this speed of rotation is **irrespective** of the position of the planet in its solar system and overall effect of the external Magravs forces, which are exerted by other planets, and the star of the solar system, on the planet.

So the twenty-four hour speed of rotation of the Earth is totally dictated by the exertion of the magnetic fields of the other cores on the solid inner cores of the planet as has just been explained, and not necessarily by any other factors.

This principle of constant speed of rotation of celestial object should hold true for the all-dynamic celestial objects that are in possession of dynamic thermal central materials cores and this rotation is irrespective of the temperature of the inner cores or temperature of the outer layers and surrounding environment of these objects.

Through tests in laboratory of similar reactors as the construction of Earth inner cores, we have concluded that the **speed of rotation of celestial objects and Matters** and the **creation of Magravs strength** are **temperature independent**.

The motional rotation of a planet is created through the same principle as the linear motion that is attained in the present Maglev magnetic railway system of Japan, with the difference that, due to *spherical shape* of the inner solid core of the planet (Fig. 3), the motion created in this case will be a rotational one (12) rather than linear. We can consider that this principle is universal.

5. The difference between Matter and matter

Another unexplained phenomenon in the world of physics has been the understanding and the difference between the Matter and the matter, one in respect to the other.

In this book we make a clear distinction between **the Matters** and **the matters** in all levels of universal orders of materials, where;

The **Matters** are understood to be as the like of the Matter, the Antimatter and the Dark Matter. These three Matters are all made of ***different*** *plasmatic magnetic fields strength* and are ***field strength*** dependent in respect to each other. The three Matters are the three major components of a plasma. Each of the three Matters have pmtics of another strength, where the constituent strength of the pmtics creates different Magravs, different magnetosphere field strength and different mass for each Matter, and these independent of the other two Matters within the plasma, and independent of the environment that surrounds the plasma.

Whereas:

The **matter** refers to the Matter component of the plasma. Matter becomes matter, when the Matter comes to levels that **plasmatic magnetic fields forces** and the ***motion*** *of these forces* in their environment can determine the compactness of the components of the Matter pmtics appearance in the plasma that make the matter appears as a gas, a liquid or a solid ***state-of-matter***. These forces being created due to interactions of Magravs, dynamism of the environment, and so on.

At the same time, Matters are pmtics strength, contact friction dependent. Where, the friction of pmtics in matter usually leads to the change of the state of the matter through the change of plasmatic magnetic fields forces to the likes of light, heat and other's.

It is the **external field forces** and ***not the field strength*** of the environment that dictate the compactness of any of the plasmas' Matters so that the pmtics becomes matter and visible or tangible to human senses, like odours or appearance of Matter as solid, gas or liquid, it is then that the ***Matter*** comes to be called the ***matter***.

The principle of ***strength*** of the magnetic fields of Matters is very different than the strength of magnetic fields in matters.

When we consider to work with Matter(s) for space technology, the term "**fuel**" is not applicable, therefore the use of the term "**MAT***ter* **MAG***netic* **S***upply*", abbreviated to "MATMAGS". This is a Matter (Matter, Dark Matter, Antimatter) ***source of pmtics*** used in reactors for the production of Magravs for the operation of gravitational positioning reactors, where one can state the Matter needed and the strength of the Magnetic fields of the Matter needed to achieve desired effects from the reactor.

Since "fuels" only refer to the *matter level*, where fuels are usually *burned through ignition of matters*, whereas ***magnetic fields of Matters***, in the form of Magravs, are manipulated to create a preferred effect, and where the magnetic fields can be reversed back to any other state of Matter, and they are never lost. The only difference would be that with the Matter supply the magnetic fields strength can differ after processing, but in reality one would be left still having magnetic fields of varying degrees in strength.

6. How motion of matters and Matters are created?

The question to be asked first is, that, how the Matter, the Antimatter, and the Dark Matter are created in the universe.

Secondly, how are the regular independent motion of **Matters** (Matter, Dark Matter and Antimatter) and **matters** (electrons, atoms, matters, planets, stars and galaxies) in the universe without burning of any fuel created and maintained?

No explanation has been put forward by the scientific world that is realistic, fundamental and applicable in all levels and to all sizes of objects and in any position in the universe.

In the following chapters of this book, I will disclose how **matters** (gasses, liquids and solids) and **Matters** (Matter, Antimatter and Dark Matter) as part of fundamental elements of creation, move in the universe. In addition, it will be explained, how the plasmas, electrons, planets and stars attain and maintain their motion in the span of the universe. (Chapters 19)

Motion in the universe is achieved by understanding in full that how the construction of Matters happens at their simplest form, from the inception of the initial fundamental particles how they come together, and how they respond to each other's magnetic fields forces, that the interactions of all constituents Magnetic field forces in Matter and plasma of Matters, leads to the creation of motion of plasmas, electrons, planets, stars, and so on.

7. How Light is created?

The present science of physics accepts that the light is created through the drop in energy of an electron from one energy level to a lower, or its ground energy level.

Further, in this book we will explain how light is created in the universe, where actually there are *no electrons present* in the plasma-saturated environment.

The other unanswered question is whether the speed of light (C) is the ultimate speed in the world of Matters.

The speed of light is in the world of science thought to be the ultimate speed that any Matter can travel.

I consider this assumption to be **incorrect** for all states of Matters (Matter, Dark Matter and Antimatter).

The assumption that the speed of light is the ultimate speed can be correct **only** for *interactions between matter and Matter component* of the three basic Matters (Matter, Antimatter, and Dark Matter).

I consider that the **plasma** is to be in **possession of all three states of Matters,** these being **the Matter, the Dark Matter and the Antimatter.** (See the cover image of the book)

Secondly, the assumption that the speed of light to be the ultimate speed, can at the same time be partially correct and applicable, only for a limited fraction of time for specific conditions of **matter** (and not necessarily only in vacuum conditions). Where **matte**r itself is only a small part of bigger components of Matters of plasma in the universe.

There are too many physical facts that are present in the mechanism of the universe's daily operation that rejects that the speed of the light being the ultimate speed. These new theories will be discussed further in Chapter 7, titled "Speed of light". A large numbers of physicists around the world do not accept such a limitation on the speed of matters in the universe.

CHAPTER 2

Interaction of Plasmatic Magnetic Fields (pmtics)

In the Matters the strength of the magnetic fields goes back to the origin of the creation of magnetic fields. Where, the strength of the magnetic fields is in terms of reference, the power of the magnetic fields. Where magnetic fields in essence are created out of the potential difference in position from their source and the rate of the flow of the magnetic fields from their source becomes the current of the magnetic fields. Thus, as the position is closer and the rate of the flow is higher, the stronger the strength of the fields. This principle of creation of magnetic fields and magnetic fields strength are explained in full in the paper titled "The creation of magnetic fields".

In understanding the ways and the simplicity with which the gravitational field force of plasmatic conditions are generated and maintained in planetary and stars systems, we can now state that the method of creation and production of the Magnetic fields and gravitational field forces of galaxies, stars, planets, molecules, atoms and plasma are all considered to be through the same method, irrespective of their physical size.

At the same time, it is important to appreciate that this applies to all scales of the universal order of creation. That is to say, stars are made of the denser, larger and stronger by order magnitudes of plasmatic magnetic fields strength forces in the universe. Compared to the plasmas and the electrons, which are made of a weaker order of magnitude of plasmatic magnetic fields strength forces.

Nevertheless, the same principle of structural formation applies to the Magravs structure of stars as to the structure of an atom, the plasma and the electron.

In the physically big state order of matter, large scale, stronger and denser magnetic fields can be easily observed and detected, like the magnetic fields of the Earth, the Sun or the Galaxies.

By the same measure and the same order of the scale and magnitude, plasmas and electrons possess and are made of and follow the same principles of Magravs interaction and attraction, and hold on to their different Matters and forces, as a star or a galaxy does.

Whereas, the existence of the lower order of the magnitude pmtics that make the physical structure of plasma due to their very weak magnetic fields strength, can mainly be confirmed through the physical existence of the plasma itself, its behaviour, and its properties.

Through our own tests, which have been done over several years, it is apparent that the plasmas possess their own Magravs forces, which are made from the interactions of combinations of different pmtics forces within their structure.

The fact is that, if the plasma did not create and maintained Magravs forces internally, it would not have been able to hold on to its fundamental particles, for the plasma to exist. By the ***Keshe Theory of the Gravity***, the existence of Magravs forces as the existence of a magnetosphere of the plasma confirms the existence and interaction of at least two plasmatic magnetic fields forces within this structure.

Secondly, if the plasma did not possess Magnetic field, it would get stock to other plasmas in its environment, these environments being for example the like of the other plasmas in the nucleus of heavier atoms. Then as the plasma is a free motion entity, which keeps distance from other plasmas in its environment, this confirms the existence of Magnetic as well as gravitational field's, which is internally maintained and emanated from the plasma itself.

Unless it can be proven to the contrary, we can assume that all plasmas are made of the same types of pmtics, and the same structural Matters and the same structural configuration. Then, we can say with confidence that all plasmas are made of identical Matters and create the identical Magravs forces within their structure.

We consider neutrons, protons and *especially* electrons as being of plasma structure with only difference in their pmtics content level. Thus, different atoms of different matters have different energy and physical pmtics properties from other atoms purely due to the level of total pmtics of their total plasmas that they can hold onto.

From this point on, in this book, it is considered that all **hydrogen atoms** have and will always be in ***possession of the same properties***, be of ***the same construction*** and be made ***of the same similar internal pmtics and Matters***.

Therefore, it is assumed from this point in this book that what we call initial fundamental plasmas, or what is called neutron, must always be made of the same composition and of the same specific pmtics at all times, for all such plasmas of all atoms.

Furthermore, it can be said that all initial fundamental plasmas create the same and similar plasmatic Magravs forces.

Until contrary to this fact is proven, then it is correct to assume that majority of plasmas have the same characteristics, the same properties, and have the same dimensions, and the same measurement of Magravs density and field strength.

Therefore, it is right to assume that the *majority of plasmas* in the universe are all made of the same collection of pmtics interactions.

Hence, it is correct to assume that to have plasma of specific characteristics and specific Magravs strengths; the plasma is made of and must be able to attain a Specific Magnetic Configurations for all initial fundamental plasmas.

From observation of motion of plasmas in the universe, it is clear that not all packs of pmtics are connected to a specific source. They are mainly free floating in space and are made of different magnetic fields strength.

Where, some of these pmtics interact with others of the *same strength*. Which, these interactions lead to the creation of the plasmatic Magravs, in their space and in respect to other floating pmtics in their given environments.

Initially, when pmtics of similar strength get close to each other, the interacting fields get entangled with each other and set-off a new cycle of Magravs in respect to other pmtics and magnetic fields in their given environment.

Further on as more pmtics of different field strength are gathered together due to their initial magnetic fields strength the new seeds for creation of new and different Matters are set to start. Where, the collection of these different Matters and fields and their interactions leads to the creation of initial fundamental plasma, in a specific given position in space.

Plasmas are always considered to be made in the same manner, by the same process, by the collection of the same pmtics force strength and the same interaction procedure of the fields within their structure.

Hence, plasmas are said to be made of **S***pecific* **E***ntangled* **P***lasmatic* **MA***gnetic* **F***ields* (SEPMAF) and this is considered to be the same for composition of all plasmas in general.

CHAPTER 3

The Initial Fundamental Particles of the plasma

The **initial fundamental particles** components of plasma are considered to be *a **collection** of weak* **plasmatic magnetic fields** (in short: pmtics) originally existing in the spans of galaxies.

These are **residuals** of different plasmatic magnetic fields activities of matters and Matters, which are released by the interaction of different magnetic fields by/and through for example processes like fission and fusion within stars and so on. It is considered that some of these ***weaker*** plasmatic magnetic fields are *released into the environment* by the nuclear decay of atoms themselves.

Where some of the pmtics (Fig. 4) are *continuously created in the **denser** environments* by the **interaction and friction** of two or denser ***or stronger*** *pmtics* and some ***weaker pmtics*** are generated through the **collision or friction** of *matters and Matters* in the universe.

South
North
South
North

Fig. A

Fig. B

Fig. 4: Graphic presentation of plasmatic magnetic fields.

Note: The figure 4 A: represents the collection of plasmatic magnetic fields of different strength. The figure 4 B: represents a graphical representation of dynamic pmtics in motion.

The pmtics due to their very weak strength are usually considered to exist and be in motion in **packs** within the different parts of universe (Fig. 5). Where, each pack can be made of number of different orders of magnitude of pmtics strength.

For ease of understanding in trying to explain that how plasmas are created initially, we will consider three packs of pmtics. Where, for our purposes, each pack of plasmatic magnetic fields is denoted by one capital letter. Then each strength sub-constituent magnetic fields forces is number denoted. The different strength fields denoted with different numbers (Fig. 5).

For simplicity of explanation, we will consider three packs of A, B and C with constituents of different dynamic pmtics (considering that fields are in motion in their environment).

Furthermore we consider that each pack of magnetic fields contains five different sets of fields strength called (**Pack A**) as A1, A2, A3, A4 and A5, (**Pack B**) as B1, B2, B3, B4 and B5 and (**Pack C**) as C1, C2, C3, C4 and C5 respectively (Fig. 5).

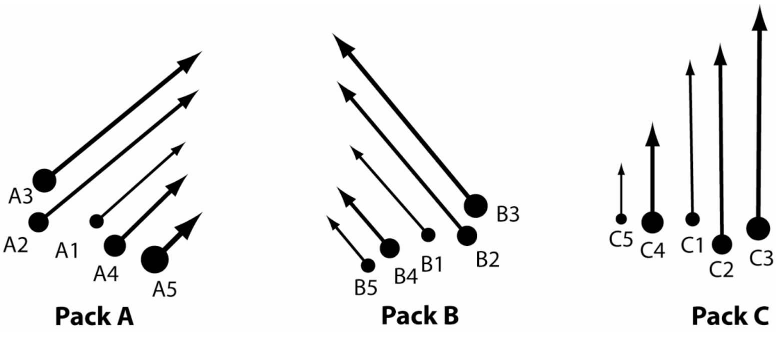

Fig. 5: Initial fundamental particles of plastic magnetic fields.

It is important to remember in this section, that when we talk about the *gravitational field forces (Magravs)*, this means the **gravitational fields** and the **Magnetic fields** of any entity in possession of these magnetic fields, which are generated simultaneously by the interaction of the same two or more plasmatic magnetic fields.

It is crucial to remember that there cannot be a situation where, there are gravitational fields and no Magnetic fields within a system and vice-versa. Even where, the Magnetic field comes to be placed in the boundary of the gravitational field zone.

In principle, the gravitational fields and the Magnetic fields of the plasma are represented in the same region in our diagram(s), and for simplicity, these are shown with the doted line and this single line is representing both the Magravs zone of a Matter.

Creation of Matter of plasma through interaction of pmtics.

Initially considering interactions between, the pack A and the pack B and ***at least one pmtics of each pack***, which are to match with each other strength ***of the other pack***, where these two pmtics are allowed to interact and interlock into each other.

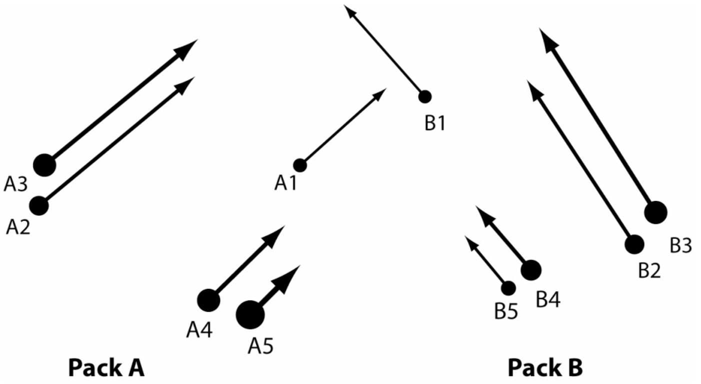

Fig. 6: Dynamic plasmatic magnetic fields A1 and B1 making the initial magnetic interaction of two packs.

Note: Only pmtics of similar or near similar pmtics strength can interlock into each other, where pmtics of different strength are attracted to each other but cannot interlock in the majority of the cases.

For the pmtics of packs A and B, which are similar or equal in magnetic fields strength, and they are in motion, and able to interact with each other or to interlock into each other (Fig. 6 and Fig. 7 a), it is considered that the North Pole of pmtics A1 and the South Pole of the pmtics B1, (Fig. 7 b) start the ***attraction*** of the two pmtics to each other.

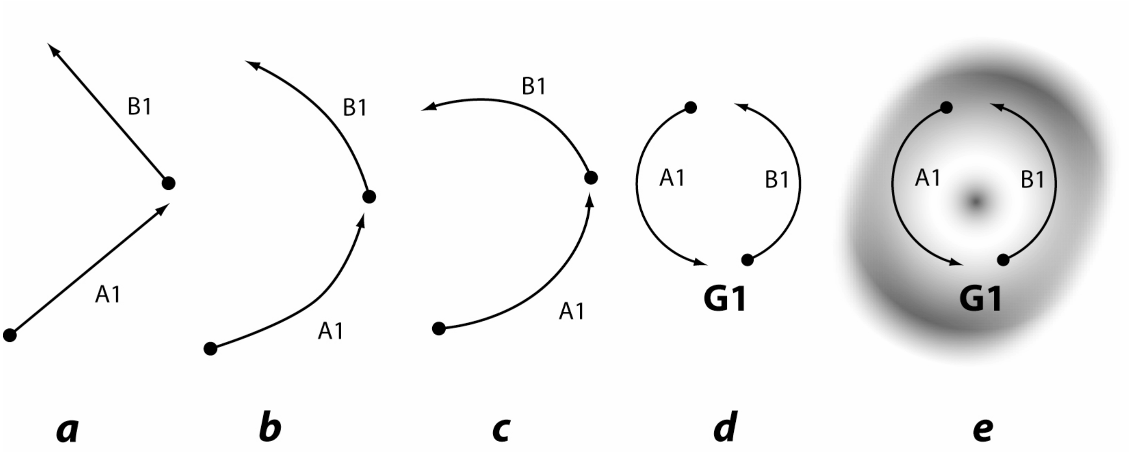

Fig. 7: Interacting plasmatic magnetic fields making pmtics of Matter.

Note: It is considered at all times in the universe all magnetic fields and plasmatic magnetic fields are in dynamic and in the state of motion, as magnetic fields in essence are always attracted or interacting with other magnetic entities, thus they are always on the move towards or repulsing from one magnetic field to another. This is how the motion of Matters and fields are maintained in the universe. This continuous dynamic mixture of different magnetic fields strength become part of a localized or universal pmtic soup(s).

This interaction and interlocking of these pmtics initiates and creates the ***first step*** of interactions between two packs A and B plasmatic magnetic fields (Fig. 7 c). This ***interlocking***, leads to the creation of the very first *initial fundamental magnetic fields set*, or the first initial Magravs, what is called the initial configuration of pmtics needed for initial seed of the plasma (Fig. 7 d and e).

The interaction of at least one plasmatic magnetic fields of each pack that is initiating, and starting the initial weak Magravs forces of a Matter of a plasma, this is the **first step** *for the interlocking of the two packs* of A and B, of pmtics forces in respect to each other.

From this point on, the first plasmatic Magravs force, essential for the seed of the first Matter of the plasma, is set to start and operate (Fig. 7 e). That is to say, the first gravitational field (G1 in Fig. 8 B) and the Magnetic field (M1 in Fig. 8 B) forces can be considered to be created in between these weak pmtics packs.

It is appropriate to consider that the interlocking of the two dynamic fields, due to their interaction and attraction, are bent inward or become overall to be ***spherical*** in shape (Fig. 7e and Fig. 8 A).

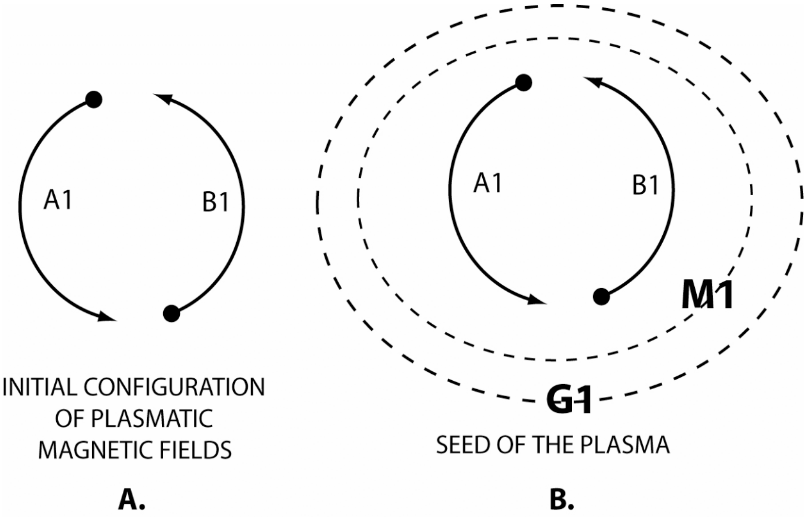

Fig. 8: The creation of the seed of the plasma.

For ease of explanation, all the initial plasmatic gravitational fields (Fig. 9 field G1), and all the plasmatic Magnetic fields (Fig. 9 field M1) from now on in this disclosure are denoted as G1 fields (Fig. 10), or what we call **Magnetic and gravitational fields** or **Magravs** and considered these to be the seed of the pmtics of the **Matter** components' of the plasma.

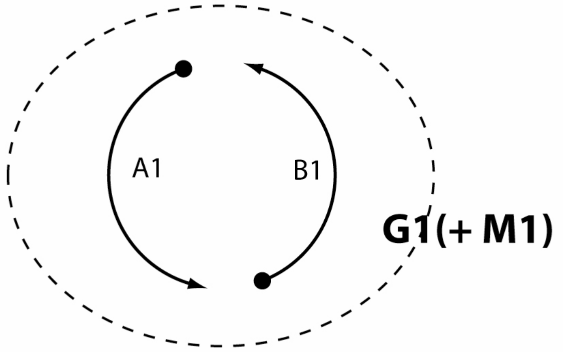

Fig. 9: Initial gravitational fields and Magnetic fields (Magravs) of the Matter components of the plasma.

At the same time, the ***remaining*** pmtics in each pack of A and B, due to the dynamism and Magravs in the vicinity of this seed, these will be kept around the seed G1 as one dynamic pack SET P1.

The collective Magravs force of the seed of Matter will be known as G1 (Fig. 10) and the remaining pmtics magnetic fields of Pack A and B as initial set of the Matter component of the plasma is denote as **SET P1** (Fig. 10).

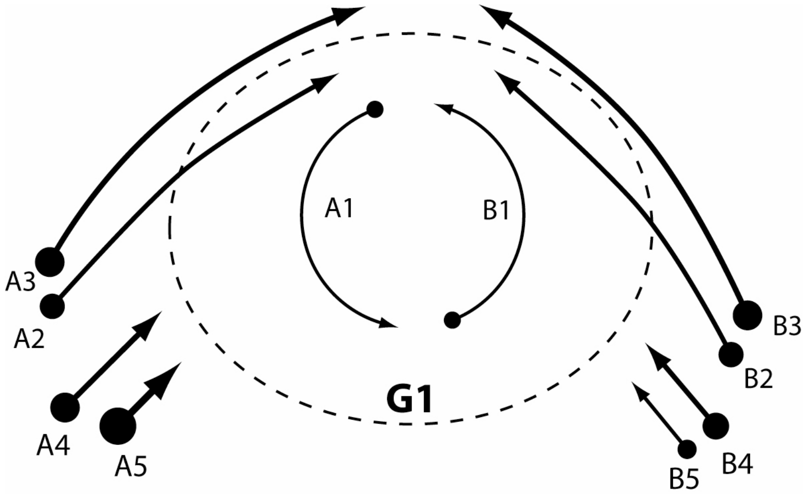

Fig. 10: **Set P1**, the initial pmtics of the Matters Magravs' components of the plasma.

*Note: It has to be remembered that the pmtics in G1 are not produced and connected to a source of a solid magnet, but these **forces** were made by magnetic fields and were detached from their source of creation, some times before, hence, the use of the term "plasmatic" for these magnetic fields. The term "plasmatic" refers in this book to a "collection" of dynamic magnetic fields, and **not** to the state of a plasma in its common term use, which for example refers to a state of dynamic protons.*

Nevertheless, the remaining plasmatic magnetic fields of the two packs A and B are ***not matching in strength*** and hence they cannot lock into each other, *these being* pmtics A2, A3, A4, A5, and pmtics B2, B3, B4, B5.

Therefore, the initial dynamic pmtics interaction of the two packs making up the G1, possessing dynamic Magravs force, become the **Matter** component's of the *upcoming plasma*.

In the open spaces of the universe, the SET P1 due to its dynamism and motion, naturally and eventually will come across another pack of pmtics, like the **pack C**.

The interaction between magnetic fields of the SET P1 and pmtics of the pack C will be different in respect to the initial interaction of the packs A and B.

In this set of interactions between all the fields and initial Magravs force of the SET P1 and the pack C, several steps can take place instantaneously, simultaneously or over a length of time.
The first step is that, some of the magnetic fields strength in the pack C (C1, C2, C3, C4 or C5) could be **similar** to certain magnetic fields strength as in the G1 of the SET P1.

The *addition of **new** plasmatic magnetic fields* does not necessarily increase the strength of the fields of the G1, but it just adds to the density of the compact pmtics in the G1, and hence **increases of the mass** of the G1 only.

Creation of the Antimatter of plasma

The **second step** is that the stronger field of the SET P1 interacts with similar fields as in the pack C and as they are ***stronger***, they would create a much stronger Magravs strength set. This, leading to the creation of a new Magravs zone, **near** the initial Magravs forces of the G1.

Due to its ***higher*** *pmtics strength*, this new Magravs system denoted as the G2 (Fig. 11) and it would have a separate distinctive characteristic. The new Magravs set possesses **a much stronger field** due to its stronger pmtics, which initiated its creation. This new and stronger Magravs will ***exert forces on the G1*** (Fig. 10), this making the pmtics constituent of the G1 a denser environment as the seat of the Matter (solid, liquid, gas) components of the plasma.

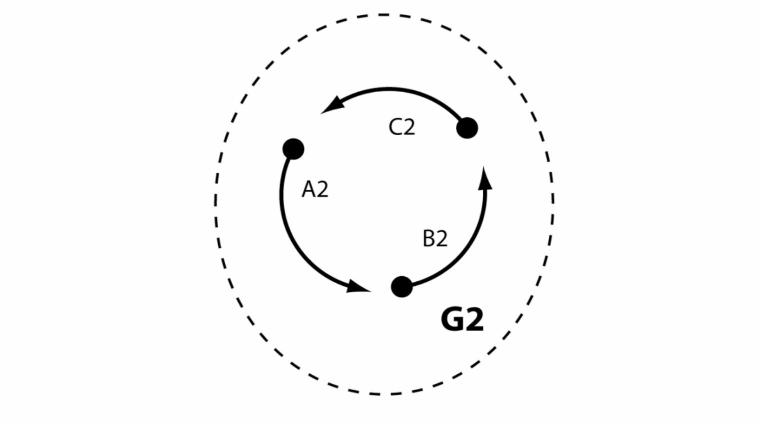

Fig. 11: The pmtics of the second Magravs.

Hence, this new interaction of fields creates a separate and stronger Magravs zone G2 (Fig. 13) adjacent to the G1 (Fig. 13) and in the same environmental plasma condition as G1. This new Magravs G2 (Fig. 11) is independent and interlocked to the initial Magravs of Matters of G1 but it will not interfere with the operation of the G1.

In fact, the Magravs forces within the G2 (Fig. 12) and its surroundings become the seed of another Matter in the same environment.

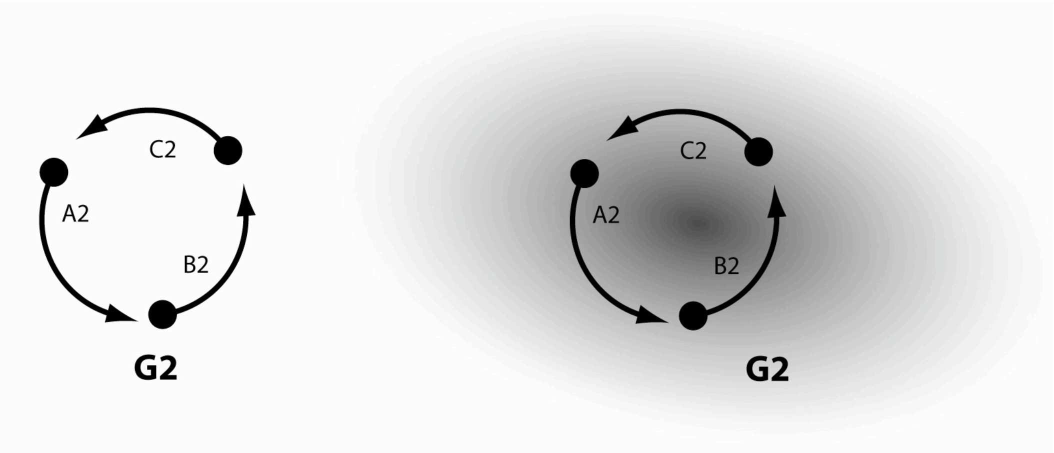

Fig. 12: The Antimatter Pmtics and Antimatter pmtics Magravs.

This new and stronger Magravs force, due to its stronger plasmatic Magravs interactions and its transparency, become what is known as the Antimatter part of the plasma (Fig. 12 G2).

For clarity, the new stronger magnetic fields interaction between the SET P1 and the pack C, and creation of the Magravs force of G2 (Fig. 11), becomes part of the **SET P2** (Fig. 13).

Where, now this new set includes the G1 the Matter and the G2 the Antimatter and the remaining pmtics of the three original packs. Where, the weaker Magravs of the G1 and the stronger Magravs of the G2 co-exist next to each other, and are interlocked to each other by the original three pmtics fields in the original packs, which they originally shared (Fig. 13). Through overall Magravs of the two Matters in respect to each other, these two Matters' Magravs keep the rest of pmtics of the original three packs together and around them.

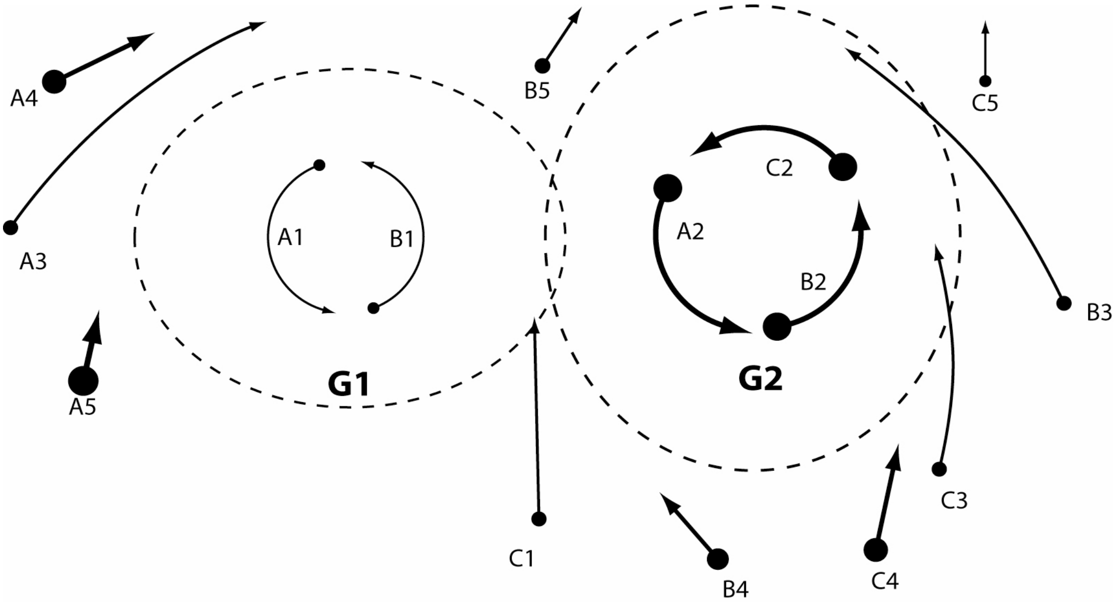

Fig. 13: Set P 2. This set includes pmtics and the Magravs of the Matter, the Antimatter and residual pmtics of 3 original packs.

This interaction of two different plasmatic Magravs sets of different strengths is similar to the plasmatic Magravs set forces operating in independent celestial bodies of the solar systems the like of the Earth and the Sun, where, the Earth can be considered to be the G1 pmtics Magravs and the Sun the G2 pmtics Magravs of the set.

Creation of the Dark Matter of plasma

The **third step** of the same interaction of the pack A, B and C pmtics, is from the other residual fields of these packs. Where some of the *residual* magnetic fields will generate a new Magravs force environment, the like of the G3 (Fig. 14). The G3 gravitational field force is a part of the same set of the same plasma environment as the G1 and the G2.

This new Magravs force zone of the G3 is created through the same process as the other two Magravs, with a difference that the outward Magravs strength created by G3 are in magnetic fields strength balance with the overall pmtics strength created by the G1 and the G2 and other pmtics within the environment of the original three packs of pmtics.

This G3 is independent from the magnetic fields of other two Matters of the plasma. But these sets overall Magravs are ***near equal*** or are ***in balance***, in respect to the pmtics of the overall enclosure of the plasma and of all the three original packs and the two gravitational sets of G1 and G2.

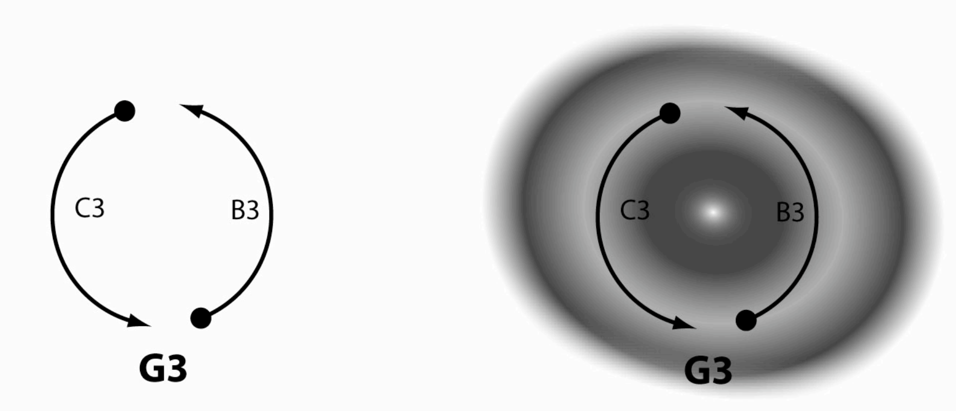

Fig. 14: The schematic and pmtics representation of Dark Matter.

Hence, even though this region of the plasma possesses Magravs forces, but it appears to be void of external magnetic fields, or this region appears to be void of a magnetosphere. Where, in the interaction of the G3 Magnetic fields with its surrounding magnetic and Magravs of the G1 and the G2 of the same environment, these fields of the G3 fail to create a visible and distinguishable magnetosphere for the G3 set with other magnetic fields in its surrounding environment. Therefore, as there are no interaction between the overall balance of the Magravs of the G3 with its surrounding pmtics, thus no visible or may be very little detectable light due to the interaction magnetic fields of the G3 with its surrounding pmtics are produced, that these could create light that could make this new gravitational centre to be detectable. Which this light can confirm the existence of another internal gravitational fields of another Matter in this region of the plasma.

Even though there would be a gravitational field force in this region of the plasma, which confirms the existence of another Matter in this region of plasma. However, the G3, due to the balance external Magravs interaction with other fields around this region with the lack of stronger external magnetic fields to create more profound detectable light in comparison to the G1 and the G2 magnetospheres within the plasma, makes this region around the G3 darker in respect to the rest of the plasma environment (Chapter 7).

Nevertheless, this region possesses pulling field forces or gravitational field forces but will not be as visible or detectable as the rest of the plasma, or will be darker in comparison, due to a lack of the stronger magnetospheric interface (Fig. 15 **SET P3**).

The G3 gravitational field force region, due to its dark appearance, becomes the Dark Matter component of the plasma (Chapter 14).

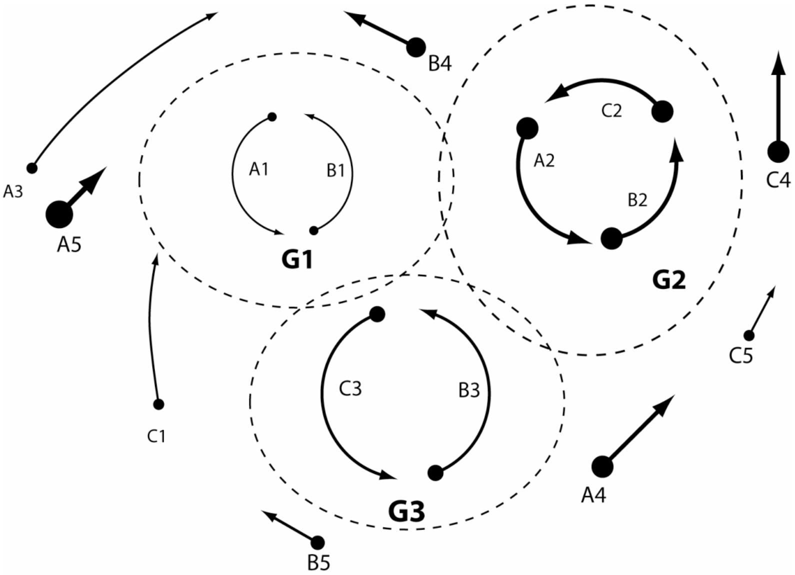

Fig. 15: Set P3 interactive gravitational fields of a Matter, Antimatter and Dark Matter.

In the past papers (36) and the Dark Matter chapter, ***Keshe Theory of creation of Dark Matter*** considers' and explains' that "Matter which possesses both Magravs, but where the overall Magravs strength are *in balance or near balance* in respect to its environments pmtics strength, so much so, that these regions do not create any or create only a very little magnetospheric visible light zone through the interaction of their pmtics (Chapter 7), that the Matter cannot become detectable or become visible in respect to its surrounding environment at specific given time and position, hence the Matter Magravs magnetosphere appears to be darker in that part of given environment, hence the term **Dark Matter**" is used.

These interactions and principles of how Dark Matters are generated have been explained and discussed in detail in the paper titled "The creation of Black Hole" (13), "Creation of Dark Matter" (36), "The rings of Saturn" (39) and further in different following chapters in this book.

We consider the creation and appearance of the Dark Matters, Black Holes, Dark Spots on the surface of the Sun and Dark sections of the rings of the Saturn, are all to be due to and through the same fundamental principle of the interaction of equal plasmatic Magravs strength forces in their given specific environment at a specific duration of time as in the G3 (13, 14, 17, 18, 24, 36).

Nevertheless, the Magravs forces of the G3 will be interacting with the rest of the plasma Matter's field's forces. The G3 pmtics released partially will feed the plasma's of other components of Matters for their motional energy and in keeping the gravitational field forces of the G1 and the G2 intact and overall to keep the plasma as one integrated system (Fig. 15).

Creation of the torus field force

The **fourth step**: Further on in this process of interaction of three original packs of pmtics, and due to the dynamic characteristics of the other three Matters Magravs, soon ***some of free pmtics*** left in the overall initial three packs will come to be positioned in the centre of three Magravs forces of the G1, the G2 and the G3, and becomes the **interlocking force** between the three. The free pmtics become a dynamic swirling compact of residual of *un-matching pmtics* field strength and behave as a partial sphere of rotating magnetic fields forces to be known as the **F1** (Fig. 16) part of the plasma.

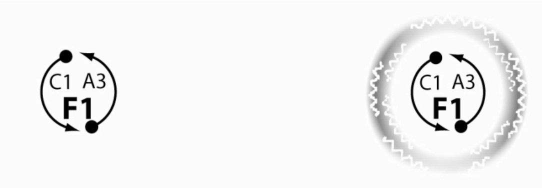

Fig. 16: The spherical torus field force.

These free F1 pmtics, due to not being able to find equal magnetic fields strength partners to establish themselves as a Matter condition phenomenon, this portion of the interacting fields become void of central gravitational field forces.

As the G1, the G2 and the G3 are in continuous motion, and as these three Matters lose their energies and momentums, due to their long-term dynamic activities, they come close to this inner field force region of the F1 and in contact with its pmtics forces. This is the point where, the three Matters of plasma gain momentum from their interactions and encounter with the dynamic magnetic fields forces of the F1 (Fig. 17).

This dynamic pmtics region in between three dynamic Matters can be compared to an empty spherical torus shaped field force effect (F1) (Fig. 17), as the field forces turn inwards, without any central gravitational field forces strength, and without any matching pmtics to generate any viable internal gravitational field force for the creation of any Magravs to lead to the creation of any Matter types.

In practice, the conversion of the pmtics of the F1 into energy, partially allows the balanced dynamism of the whole plasma to be maintained.

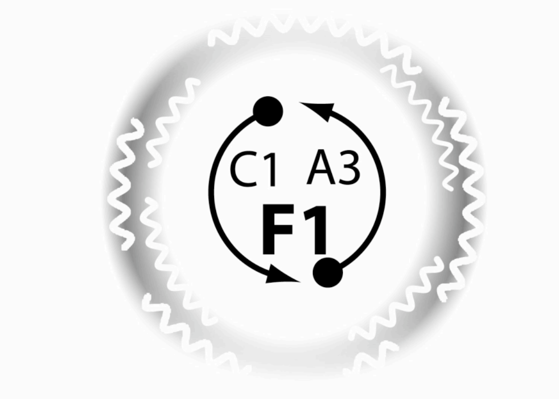

Fig. 17: Spherical torus shaped pmtics made by interaction of different strength residual pmtics of the three original packs.

F1's are considered as a natural process and phenomenon in the plasma and in the universe, with the content of larger order of magnitude difference in pmtics strength in their constituents.

This dynamic spherical magnetic fields force region of the F1 in plasma or in space, void of central gravitational field forces, is commonly called a ***Wormhole***, and its operation of its field's forces in motion as Wormhole effect.

It has been hypothesized in the past by scientists that such a Wormhole can be used for inter-space travel with extreme speeds of motions near the speed of light.

We consider that these dynamic magnetic fields regions can be used to increase the speed of any object, which comes in touch with their dynamic magnetic fields. In fact the fast journey of the object through the F1, will not necessarily lead the Matter to anywhere rather than giving a turbulent fast ride within the structure of the plasma.

In case of wormhole in a galaxy, the object will have a fast speed travel without much control in speed and dictating the point of final destination of where the object might end up.

These are points in space that acceleration for motion for other Matters of the plasma or physical matters can be obtained, by using the dynamic pmtics forces of the F1 in plasma or space, where these forces happen to exist. Finding these points in space are much easier than ever thought before (Future publication).

Initial structure of Matters of plasma

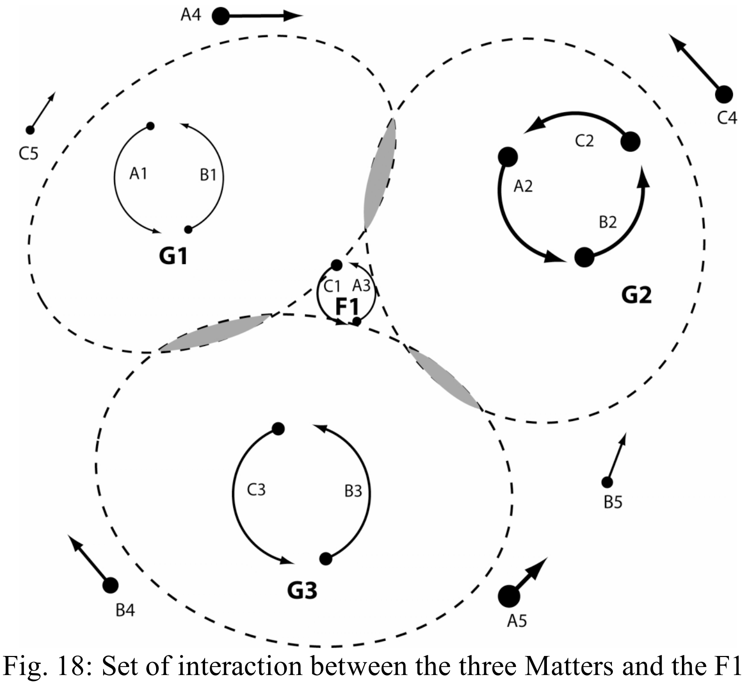

Fig. 18: Set of interaction between the three Matters and the F1.

Even though each of the three Matters are separate and operates independently from other Matters within the plasma, it is possible that from time to time the magnetospheric envelop of each Matter comes in touch with other magnetospheres of Matters of the plasma and creates a common ***interface*** in the region for the window of time of the interaction (Fig. 18 shaded areas).

It is important to note that plasmas can possess several G3's and F1's of different pmtics strengths in its overall entity for some time, individually or simultaneously, at different time in its life cycle.

Transition-zone pmtics

The new overall interaction of all three Matters and all fields of all pmtics of all three Packs, as one entity, leads to the creation of the initial Matters and Forces of the plasma (Fig. 19).

At this point, there could be other pmtics of the three original packs within the structure of plasma that could not link or be part of the components of the three Matters and the F1, these free leftover pmtics literally float in the plasma and over time, they become the feeding pmtics for one Matter or another and the interlocking magnetic fields forces of part of the other Matters of the plasma.

These loose pmtics in-between interfaces of Matters are the dynamic plasmatic magnetic fields transition zones, which we call "***the energy transition zones***" (Fig. 20, the field arrows), in the overall structure of plasma. The magnetic fields in these intervening gaps among the Matters are fields that are mutually shared by all Matters in the plasma, when there pmtics strength comes to their level; these residual pmtics are considered more as a mixture of magnetic rays than pmtics.

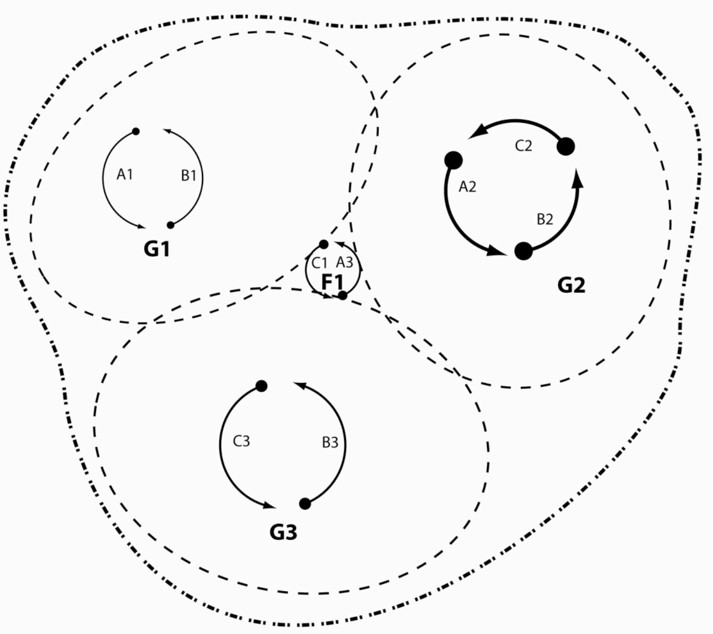

Fig. 19: Initial Matters and forces of the envelope of plasma.

They are in the plasma, like intermediating or interface zones in between two or more

Matters of the plasma at the same time.

Now that all the plasmatic magnetic fields of the three original packs are considered and counted for and used up, now we consider that the full set of the G1, the G2 and the G3, loose pmtics in the transition zones and F1, create an overall **integrated** independent Magnetic field and gravitational field of their own to keep the integrate of all pmtics of the three Packs A, B and C, together as on entity. Where this new entity becomes to be known as the "Initial Fundamental Plasma".

Initial Fundamental Plasma

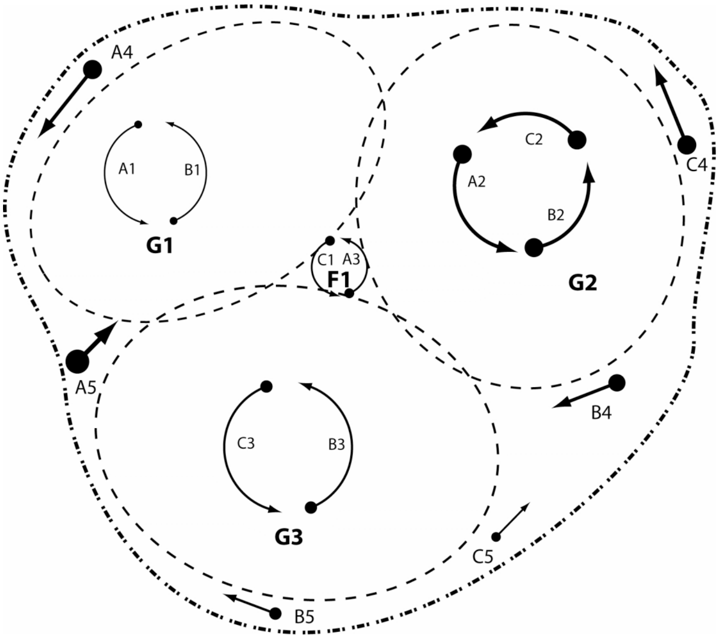

Fig. 20: Keshe Model of the Initial fundamental plasma.

I call this dynamic concept the **Keshe Model of *"The Universal order of creation of Matter"***.

We consider the Initial Fundamental Plasma to be similar to **the construction of the Neutron** (Fig. 20).

Figure 21 gives the conceptual presentation of complete plasmatic magnetic fields of plasma, of "**The universal order of creation of Matters**". This model represents the pmtics structure of the **Initial Fundamental Plasma**, where the G1 is the Magravs of Matter, the G2 is the Magravs of the Antimatter and the G3 is the Magravs of the Dark Matter.

It is considered that Initial Fundamental Plasmas always have an internally balanced magnetic fields environment, as one observes in the plasma as in the solar systems.

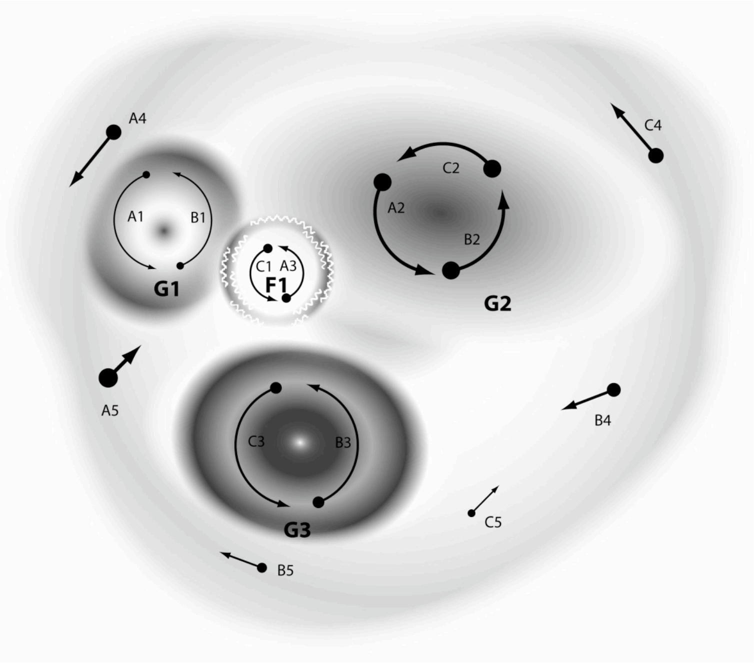

Fig. 21: Keshe Model of the pmtics of Matters of plasma.

At the same time initial fundamental plasma maintains an overall and external Magravs created by Matters and all pmtics, which are within the confine of the plasma too. Thus the internal balance magnetic fields of the neutron does not mean that it does not possess Magravs in respect to its outside environment. But due to the existence of Magravs of its component structure the initial fundamental plasma maintains an independent magnetosphere in respect to its surrounding environment.

When, contemplating on discussing about the proton or the neutron as a plasma, one has to talk about the whole of the plasmas' components, namely, the *Matter* part (what has been considered as the matter part of the nucleus of the atom), the *Antimatter* part, the *Dark Matter* part, the *spherical torus magnetic fields* and the *transition magnetic fields parts.* These are all together ***the components of one plasma*** and have to be considered as one package and entity.

There cannot be a plasma condition, where all three Matters do not participate in the existence of the plasma.

Decay of initial fundamental plasma

At the point in the cycle of the life of an initial fundamental plasma, when the shared magnetic fields of the transition zones, or pmtics of one of the Matters of plasma are partly exhausted and the Magravs balance among or in the Matter fields are disturbed and the Magravs binding forces are unbalanced, then the initial fundamental plasma disintegrates to its substructure of new balanced Magravs of the same Matters for the initial fundamental plasma to find a new balanced combination within its own Matters and fields within their given environment.

This process of splitting of the initial fundamental plasma to find new and smaller balanced plasmas for Magravs of Matters and fields balance, I call this process the **decay of the Initial fundamental plasma.**

The decay of the initial fundamental plasma always leads to the creation of at least two new balanced and smaller initial plasmatic magnetic fields plasmas of the same configuration, with, one plasma which possess more of the Matters and is larger, and one plasma which is a smaller and holds less of the Matters and pmtics of the original initial fundamental plasma. Nevertheless, the two new plasmas and their contents have to keep the overall balanced structure of the initial fundamental plasma. This happens so that both plasmas can survive within their environment as one entity.

The outcome of the **decay** of the **initial fundamental plasma** into two new components is considered to become the creator of the new ***proton*** and the ***electron*** of an atom. This new balanced pmtics entity, or so-called atom; itself has to have an overall balanced Magravs forces operating within it, for the atom to keep the integrity of the original initial fundamental plasma.

The decay of the initial fundamental plasma leads to the splitting of its content into a proton and an electron simultaneously (Fig. 46 diagrams 1 to 8). Nevertheless, at the same time as the new plasmas' component reposition and rebalance during this process of decay, this leads to the release of some residual pmtics or magnetic fields, which are not needed for or by the two new entities for their respective parts of their Magravs positioning (Chapter 18).

The release of the unwanted fields for the balance interaction between the two new plasmas, leads to the release of the magnetic fields packs as different rays or energies from the splitting of the initial fundamental plasma. This release can be in any form like a mixture of rays, pmtics, sound energy packs and so on.

I consider this ***process of the decay*** of the initial fundamental plasma of the contents of its three Matters and fields to be a **new** fundamental universal principle, where in larger scales of order of magnitude; it is called the nuclear decay of atoms.

In science at the present, scientists consider that elementary particles are quarks and so on, but now we know that ***quarks*** *themselves are further* ***made-up of basic magnetic fields and their interactions***. Where, ***collections*** of magnetic fields in a group, as dynamic plasmatic magnetic fields, and their interactions with other dynamic plasmatic magnetic fields, leads to the creation of the Matter(s). Where, the mass of the "quark" is produced by the interaction of these plasmatic magnetic fields through the principle of the creation of gravity as has been explained in Chapter 1.

Secondly, what is at the present called the spin of the quarks now we understand is the natural dynamic motion of the Matters created in the plasmas' environment, through the interaction and Magravs positioning of the dynamic Matters of the plasma.

Thirdly, what is considered by the present science, as the *colours of the quarks*, are in reality the different contents, of different plasmatic magnetic fields strength interactions leading to the creation of gravity and Magnetic fields with their overall interaction effect in respect to their environment pmtics, which leads to the creation of the magnetosphere of each Matter of the plasma, these being the magnetosphere of the Matter, the Dark Matter and the Antimatter in the magnetic environment of the plasma or the neutron, the proton or the electron.

This means that the quark cannot be considered as an “elementary particle” as it has sub-components and is made of magnetic fields itself.

Thus, ***I consider the new “elementary particle” to be of the “magnetic fields” origin,*** which this itself explains why one sees all sorts of effects, like the present quarks, bosons and the spin of the dynamic Matters of the plasma, and so forth. Where, these all can now simply be explained as being due to the interaction of different magnetic fields strengths, which makes up the structure of the Matters of the plasma and the other magnetic fields, which are within the structure of the plasma.

Even in the largest scale, this initial fundamental plasma decay model is what cosmologists see in the galaxies as the collapse of stars (which the star is and can be considered to be the initial fundamental plasma) and its re-emergence as new smaller components of star or solar systems. By the same principle, the release of energies for the star to find its new balance through its star decay always leads to releases of vast amounts of unwanted matters, Matters and energies and creation of sub-components but still with its Magravs forces.

This decay of initial fundamental plasma is discussed in more detail in Chapter 15.

CHAPTER 4

Matters' of plasma and their transition of states

There is conclusive evidence to show that plasma possesses energy, and maintains specific magnetic behaviour, and that plasmas can and do generate their own Magnetic fields and gravitational field forces.

Research done by a number of institutes around the world indicates that the plasma of Matter is accompanied by a second Matter constituent called the Antimatter. Some institutes (6, 7, 8) claim that they have separated and have managed to contain the Antimatter.

In other sections of science, scientists refer to what is called the **Dark Matter**, and observations of virtual matters.

The world of science has failed to consider the creation of all these three states' of Matters (Matter, Antimatter and Dark Matter) as constituents, and as of one process, in the chain of initial progressive process of creation of the three Matters and as the content of one plasma in the universe.

This failure has been due to the lack of understanding that the Matter, the Antimatter and the Dark Matter components of plasmas are all created simultaneously and as one integrated dynamic plasmatic magnetic entity.

It is fundamental to understand that one of these Matters' does not exist without the others within the overall structure of the plasma. Then through this process, it becomes easier to understand the operation of Matters, matter and fields, plasmas, atoms, molecules, solar systems and galaxies.

One can now translate this new understanding of how plasmas are created and operate into how larger celestial objects are created, and operate in their given environment collectively. In other words, we translate the knowledge of the operation of electrons around the nucleus to the operation of planets around their stars.

To understand the relationship and construction of the plasma with more clarity, every effort in this book have been made to go through stage by stage, how **Matters** of plasmas are considered to be connected, and *appear* **to inter-transit from one state of Matter to another** within different environments without loss or gain of any of their plasmatic magnetic fields strength.

In the matter world this is like the inter-change between states of matters, from solid-to-liquid, liquid-to-gas, and vise-versa.

Where in the matter world the ***state-of-matter*** is *magnetic fields'* **forces** dependent, and in the ***state-of-Matter*** it is *magnetic fields'* **strength** dependent.

The phenomenon of transition from one state of Matter to another is a regular occurrence in dynamic pmtics processes in the spans of the universe (Fig. 21).

All components of Matters of the plasma or any Matter or matters in the universe are connected to each other through the interaction of their pmtics, and their pmtics strength, which are the constituent and the creator of their Magravs.

Thus, as each Matter gains or losses some of its pmtics, one Matter component of the plasma will become or can become the component of another Matter of another plasma or field at "a given point in motion of pmtics" (time) and at a given point in its space.

For example, Matters of plasma can interchange from one Matter to another due to their motion of dynamic pmtics or what is called energy levels changes in respect to a given environment pmtics strength.

All Matters can appear as different Matter by just changes of the environmental pmtics strength too. For example the same plasma or Matter by just passing through another region of different pmtics strength. This is the process what we call *transmutation*, but in this case the transmutation is done by the environmental condition and not by the internal Matter pmtics strength change.

Matter can appear as different Matter, for example Dark Matter can appears as Matter by *simply entering into a new pmtics environmental condition*. This is how Dark Matter suddenly appears as virtual matter.

For example, **Dark Matter** in a given environment of plasmatic magnetic fields strength by entering into a new and stronger or weaker pmtics strength configuration, instantaneously, becomes **visible Matter** in respect to its newly entered environment (Dark Matter to Virtual matter). I call this the ***phase of transmutation of Matter(s)***, so this is the unveiling or occurrence of a new state of Matter from the same old Matter in a different outfit.

Secondly, Matters can change their status and behaviour from Matter to another for example from Dark Matter to Antimatter or Matter, by simply ***gaining from*** or ***losing to*** a part of their plasmatic magnetic fields strength to other Matter(s) or pmtics in respect to their given environment.

If Antimatter Magravs strength reaches an environment pmtics strength that its overall Magravs strength becomes equal and in balance with its new environment pmtics strength level, then the Antimatter can become and behave as the Dark Matter entity in respect to its new given environment (Fig. 22).

On the other hand, for example as the pmtics strength of the components of Antimatter reduces or as Antimatters interact with other Antimatters of different Magravs strength, these Antimatters lose enough of their Magravs strength that one or both Antimatters come to a new balance of plasmatic magnetic fields strength. Where, these balances can be in their internal and external Magravs, then these Antimatters Magravs strength can come to a level that the new strength of the former Antimatter become to be in the pmtics strength level of the Matter, or change to the Dark Matter field strength level status in respect to the same environmental condition.

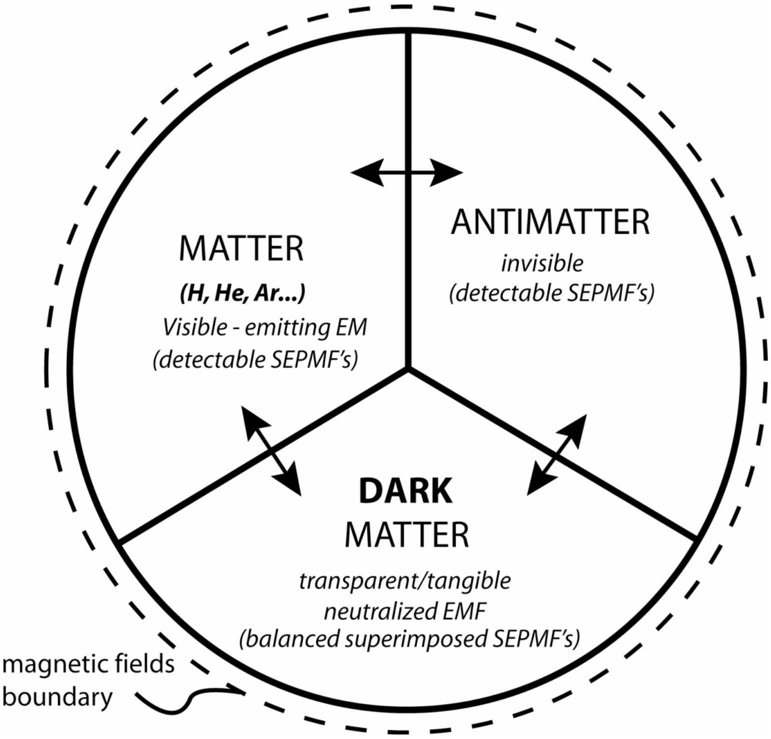
MATTER
(H, He, Ar...)
Visible - emitting EM
(detectable SEPMF's)
ANTIMATTER
invisible
(detectable SEPMF's)
DARK
MATTER
transparent/tangible
neutralized EMF
(balanced superimposed SEPMF's)
magnetic fields
boundary

Fig. 22: Schematic Keshe model of interconnection relation between different states of Matters.

This is what we call the ***Keshe Theory and principle of the transition of Matters***, that is to say "***the transition from the Matter to the Antimatter or the Dark Matter and in reverse, and the transition of the Dark Matter to the Antimatter and in reverse, is a normal condition of transition of the Matters"*** (Fig. 22 and Fig. 23). These figures showing in a simple and new way how and the connections by which interchange(s) between the states of Matters in the universe are achieved.

Keshe principle of transition of the Matter states' that, "the plasmatic magnetic fields strength of the environment in which the Matters operate in, ***dictates*** what the Matter appears as or presents itself as, in that given environment" or on the other hand "when one Matter gains or losses enough of its plasmatic magnetic fields strength that the Matter Status, characteristic or behaviour becomes of other Matters of the plasma status and behaviour or characteristics within the same given environment".

Therefore, Matters can change from one Matter to another by increase or decrease in their pmtics strength or they can appear as different Matter as they enter different plasmatic strength environment and conditions.

That is to say, the Matter can become to have the characteristic of the Dark Matter in a given environment pmtics strength, and by the same measure, the Antimatter can lose enough of its pmtics to become Matter, and so one. The schematic transitions of state of the Matters are shown in figure 22 and figure 23 of the ***Keshe model of transition of Matters.***

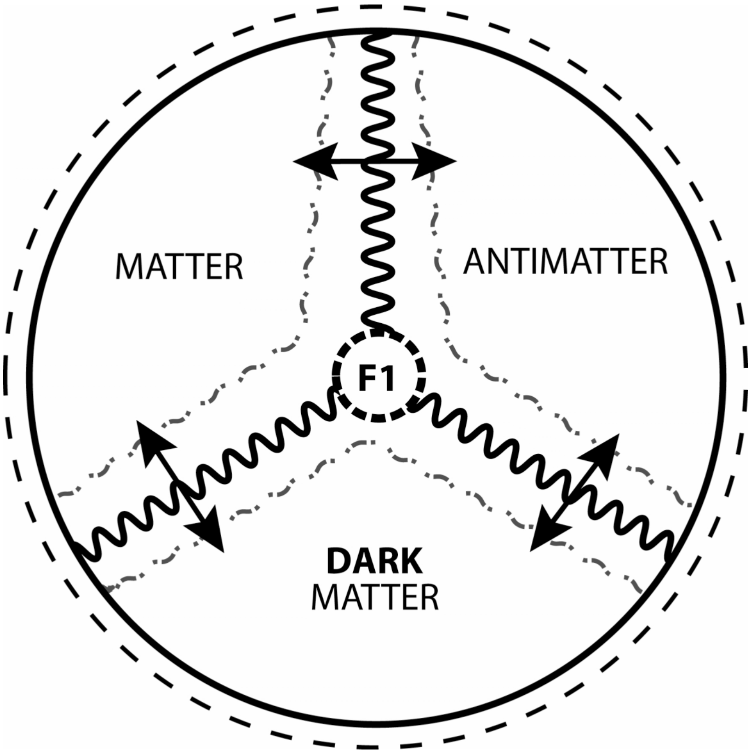

Fig. 23: Keshe universal Model of the plasma transition of plasmatic magnetic fields and Matters.

With the use of simple pmtics plasma dilution reactors, which have been developed by Keshe Foundation through its Keshe Technology development sector, the principle of transition of Matters have been tested and results have proven to be in line with the theoretical consideration.

Where, in these nuclear-based reactors, Matters can be allowed to go from one state of Matter to another and then allowed to return to their original state of Matter. This new method of transition of Matters brings about a new way and technology of producing all man’s material needs for motion, energy, creation of new matters and so on.

For example, effects and properties of transition of Matters (Fig. 23) can be used to make physical Matter to have no magnetospheric boundary as Dark Matters do in a given environment, so that the Matter behaves like the Dark Matter in characteristics. In these environments, the Matter will become translucent dark or the centre core behaves as becoming invisible in respect to its surrounding environment, without any change in the property of the physical Matter or change in its gravitational field strength or position in space of the Matter.

As the Matter with this property presents no effective magnetospheric condition in respect to its surrounding environment due to the lack of /or little interaction between the pmtics of the Matter and its environment, then the Matter now can behave as Dark

Matter and move without friction and freely within the same environment.

Utilizing this state of the Dark Matter effects around an aircraft or spacecraft, this condition will give the craft frictionless motion and higher speeds and possible invisibility in the same environment (Chapters 13).

The technology of transition of Matters has been tested-out over years by the Keshe Foundation, and several systems were designed, developed and manufactured and used to achieve the formation of original magnetic fields plasmas to prove the correctness of these principles, and indeed the effect of transition of Matters were proven to be correct.

The implication and application of the use of transition of the states of Matters are so immense and profound, that one needs to know, for example, where the Matter will end-up in, and in what state of Matter in respect to its environment, one will appear as.

That is to say, one has to know if for example, going from Matter to Antimatter, where in the strength fields of Antimatter, one will end-up in. Vice-versa, when one can go from Dark Matter into a Matter condition zone, one has to know in what state of matter (gas, solid, and liquid) one will come to ends up in.

Through the operation and control of these reactors, one needs to know where the system state of matter and magnetic condition will end-up. This could be in liquid seas, or in the middle of a solid Matter or in the middle of deserts. Without being prepared for these parameters through the operation of systems, which can facilitate such a transition behaviour and characteristic, the use of Matter transition can have catastrophic consequence, not only for the system, which produces it, but also for the occupants of possible crafts using this type of transportation in the spans of the universe.

By the understanding of the principles of creation of the initial fundamental Matters within the initial fundamental plasma, through the interaction of the initial fundamental particles, a number of unanswered questions in the world of science can be answered.

For example, in the universe the total quantity of plasmatic magnetic fields and the related energies are to be considered ***constant***. Therefore, *in the **totality***, no matter, Matters or magnetic fields will ever be lost or gained in the universe. Pmtics of different strength start existence at one strength level and they consolidate to become stronger, or lose their field strength through division and collision to become weaker in strength. Nevertheless, pmtics in their different strength fields can never be destroyed or lost.

Therefore, pmtics simply interact, divide, and interlock to produce different Matters and conditions like forces, motions, heat and compactness for them to be matters to be visible to the eye of creatures, according to their positions, and so on.

Pmtics in their interactions with others of the same can only move from one state of Matter to another depending upon the condition(s), which prevails at a given point in a given time in space. Thus, the Matter can go from the Dark Matter to the Antimatter and vice-versa, and the Matters to the Dark Matters and vice-versa and so on.

Therefore, the theory and the discussion that universe one day will run out of matters and energy and then there will be nothing as has been promoted by some scientists in the recent past is anything but a fallacy.

By the same principle, we call the **pmtics in motion as** the **energy** of the pmtics**.** These pmtics in motion can release and can lead to creation of the same Matters as they were created-out of, and can have the same capability. That is to say, for example, the energy of the Dark Matter can change into energy of Matter, or part of the matter, and the same applies for other states of Matters.

In understanding, the principles of "**the universal order of creation of Matters**", one can utilise the transition of states of Matters and matters to take full advantage of universal Matters and matters pmtics and their energies.

Through the permanent interaction of all dynamic pmtics of different scales of strength, there will be no end to the universe, as the universe moves through its different pmtics strengths field interactions and repulsions. Where, these will lead to eternal motion of the universal pmtics forces and creation of new Matters, matters, forces and new condition for them to cooperate within the universe. Thus, there never has been a Big Bang and there will never be a doomsday scenario for the universe as a whole, the universe will be forever in equilibrium in its overall universal order of pmtics.

The pmtics of Matters will ***transit*** from one state of **Matter** (Matter, Dark Matter, and Antimatter) to another state of **Matter**. In addition, from one state of **matter** (solid, liquid, gas) to another state of **matter.**

Where, these changes will depend on the properties and forces of the states of plasmatic Magravs and their strength in respect to each other. Similarly, pmtics from strength of fields of Matter will transit to pmtics of different pmtics strength of Matter, in these ***processes of transition***, plasma releasing dynamic pmtics in motion as **energy**.

The real size of the universe will never change; that is to say, the universe cannot expand into a space if the space for its expansion was not there to begin with. Where, from our point of view we only observe a small fraction of the whole, which we observe that fraction expanding. The observed expansion is relative to the position of the observer in respect to a given position in the universe, for example what will be a red shift observation of one at one end of the universe, this will be blue shift to an observer on the opposite direction of the motion of the same entity. In the paper titled "Creation of magnetic fields" it has been explained that the ***blue shift of light*** is created when ***pmtics are compressed***, and the ***red shift is created*** when the ***pmtics move away*** from each other, or ***when pmtics are opened up***.

In the real world of the universal order of creation there is no room for the principle of singularity. Except, if, we would call, a large pmtics, that covered the whole of the universe singularly by itself. Then all that has happened, is that the original single plasma has gradually disintegrated to smaller and weaker strength fields, which their interaction has lead to the creation of what has been seen and is in the present in the short time of the life of man on the Earth. Then, such original singular plasma, which covered the whole of the universe, had to gradually to disintegrate to smaller strength fields, which their interactions have lead to the creation of all events, Matters and matters in the Universe.

The transition of state of Matters principles will be discussed in full in future disclosures.

CHAPTER 5

Definition of Energy Through pmtics

The initial components and construction of plasma are due to the existence of initial dynamic plasmatic magnetic fields.

For a magnetic field to be able to be detected or its existence to be confirmed, the pmtics has to *leave the confines of the boundaries* of its containments, this being the boundary of Magravs of the Matter or the plasma.

It is important to keep in mind that the pmtics do not possess or show **field force** unless they are in motion.

The energy possessed by a plasma can be define as "the total plasmatic magnetic fields forces of all its constituent Matters *and fields in motion,* in respect to the original position of the plasma in a full three hundred and sixty degree spherical direction''.

This is an important factor, as up to this point, scientists have measured and calculated the energy of the Matter component of the Matters of the plasma. Where, in calculating the true energy of the plasma, *one has to calculate the energy of* ***all Matters*** (Matter, Antimatter and Dark Matter) and the other components of the dynamic plasma.

When the magnetic field has left the Matters' or the plasmas' boundary, at this point its existence and the content of the pmtics strength of all Matters pmtics' released can then be confirmed as energy. This is shown in (Fig. 24), the magnetic fields in the Matters.

The pmtics constituent constructions of individuals Matters of plasma are not energies in respect to other Matters while they are within the boundary of the plasmas Magravs' forces. These are shown as the magnetic fields inside the Matters on the left of image (Fig. 24) but are *only forces* (these being the like of Magnetic field or gravitational field forces) in respect to other Matters and pmtics.

When the field leaves the environment of its Matter or the plasma, it is at this point that it can release its domain pmtics to other plasmas in regions or combinations of field interaction zones. These are shown as field forces A at the boundary of the plasma in figure 24.

As the plasmatic magnetic fields, moves further from its original plasmas' boundary (field B in Fig. 24), the plasmatic magnetic fields uses' its energy to overcome other Magravs in its environment. Hence, as the plasmatic magnetic fields moves further away from its mother plasma, the less energy it will have to transfer to other plasmas as it comes into contact with their pmtics.

Thus, energy transfers from one Matter to another Matter through one pmtics strength and the measure of the pmtics possessed at the point of the contact with the second Matter, this is position and distance dependent, from the point of departure to the point of arrival.

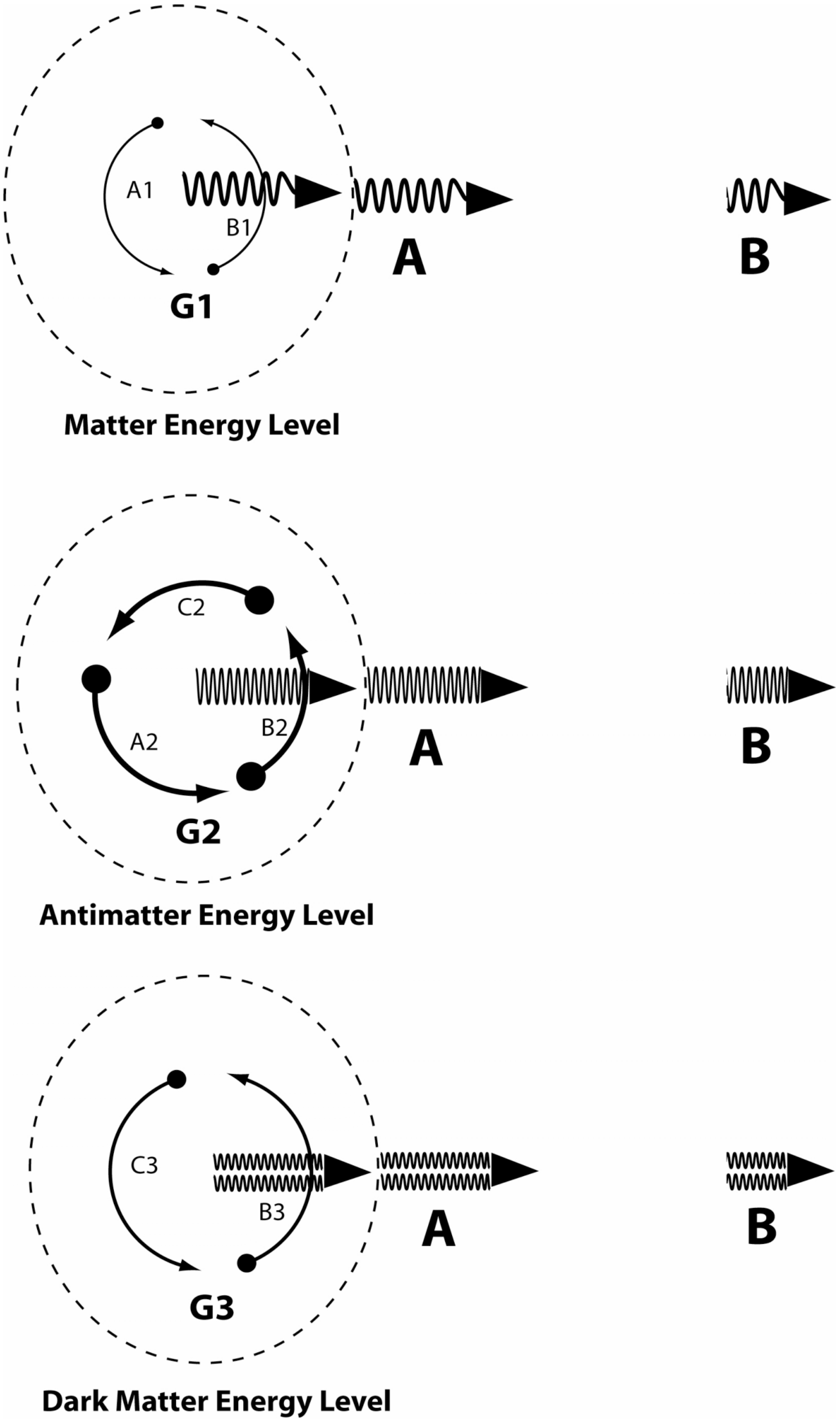
A1
B1
G1
A
B
Matter Energy Level
C2
A2
B2
G2
A
B
Antimatter Energy Level
C3
B3
G3
A
B
Dark Matter Energy Level

Fig. 24: Schematic presentation of relationship of pmtics and energy in respect to their Matters.

The total Energy released by the Matter or plasma is the total of pmtics forces that the plasma can release and transfer, and that and what measure of the pmtics can be absorbed by all components of other plasmas in its environment.

That is to say, the same proportion of magnetic fields forces by Matter, Antimatter and Dark Matter and other fields of plasma are released and the same proportion of all energy fields are transferred to the all recipient plasmas and Matters and pmtics.

The total energy of Matter is said to be the total pmtics that Matter can release as it disentangles all the components of its **initial fundamental magnetic fields**, or what is known as its' "**initial fundamental particles**", which the plasma and of the Matters of plasma were made up of, at the point of their inception.

Thus energies released by the plasma or the total energy possessed by the plasma is in fact, the total collection of initial fundamental magnetic fields of all the component constituents of the plasma, which disentangles from their Matters or plasmas and are in motion in the direction of the encountering other pmtics components.

Pmtics absorbed by some plasma's can be only a part of the total pmtics in motion, or energies that have been released by the initial fundamental plasma, and are not necessarily considered to be the total energies that a plasma possessed or has released.

Energies released by plasma are said to be "**the total initial fundamental pmtics released by all components of that plasma" passing its Magravs boundary.**

Multi-plasma atoms' release more plasmatic magnetic fields, due to their total sum of the plasmatic magnetic fields released by all their participating plasmas (protons plus neutrons, and electrons).

The principles of creation of Magravs forces, gives us an insight that "**energy**" can simply be defined as "pmtics, once released, through the dynamic Magravs characteristic boundary from the plasma or Matters, that these released pmtics can ***exert or pass a part of their released pmtics to another plasma,*** for the new Matter or plasma to maintain its existence and/or maintain constant normal magnetic fields strength and/or increase its magnetic density of the same pmtics strength condition to their ground base levels, in their given environment" (Fig. 24).

Measure of a mass

One of the central unsolved questions in the world of physics is how the mass of the Matters or of matters are created, and how they can be calculated.

Through the ***Keshe theory of Mass***, the creation and the measure of the mass of an entity can be defined to be:

"Due to the interaction of at least two magnetic rays, magnetic fields, or plasmatic magnetic fields, which the outcome of their interactions leads to and creates two field forces of a ***pulling*** field forces called *the gravitational fields force* and a ***pushing*** field force called *Magnetic fields force*, whereby the interaction of the gravitational and the Magnetic field forces with each other, this interaction of two field leads to the creation of a given magnetic fields strength **balance** between the two field forces, where, the measure of the magnetic fields difference of the interaction of these two field forces of the gravitational and the Magnetic field forces, is the measure of the Mass of the entity which the two original magnetic fields have created.”

The outcome of the interaction of the fields balance is creation of the mass of the Matter, components’ of the plasma (Fig. 25), or mass of the matter (solid, liquid or gas).

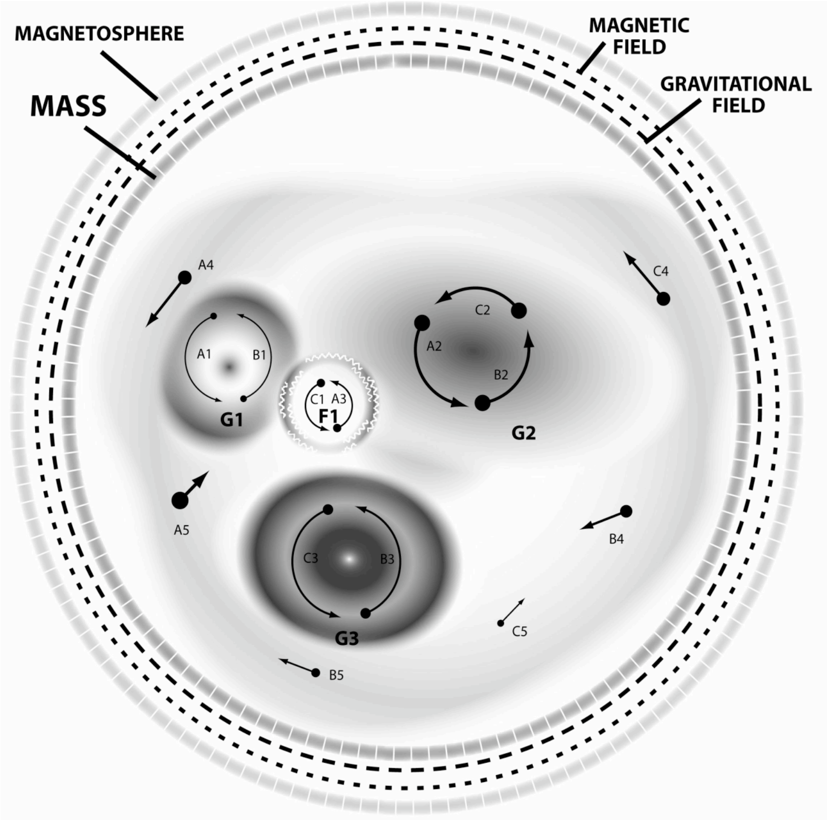
MAGNETOSPHERE
MASS
MAGNETIC FIELD
GRAVITATIONAL FIELD
A4
C4
C2
A1
B1
A2
B2
C1
A3
F1
G1
G2
A5
B4
C3
B3
G3
C5
B5

Fig. 25: Mass

That is to say by the ***Keshe law of measure of the mass*** of any entity, this being Matter, plasma, atom, matter, planet, star, etc states that:

“ **M**ass of any entity =

M*easure of the total gravitational fields strength forces of the entity*

- (minus)

M*easure of the total Magnetic fields strength forces of the entity*”

Through this definition and understanding it becomes clear that the gravitational field forces are always predominate and stronger as all entities possess mass, otherwise their mass would be negative.

Thus ***mass*** *is the measure of the* ***difference between two or more magnetic fields after their interaction***, and **NOT** the measure of strength **before interaction**.

Where, the measure of the magnetic fields *before interaction* is their “forces” of their fields, and *after the interaction* of the magnetic fields the outcome of the balance between the gravitational and Magnetic fields that is generated, becomes the mass in respect to a given fixed position of the central Magravs.

The creation of Magravs always is accompanied by the creation of a ***Magnetosphere*** and the creation of the mass of any entity in the universe, and all these four forces are created instantaneously and simultaneously.

That is to say, “the **Initial Fundamental four forces** of existence of an entity that are, the ***Magnetic fields***, the ***Gravitational fields***, the ***Magnetosphere field*** and the ***Mass field*** forces, are *all the outcome of the interaction and interlocking of two or more mainly matching in strength magnetic field*s or plasmatic magnetic fields, irrespective of the size of the entity, this being Matter, matter, an atom or a star.

This is the reason why the mass of any object, which is created by any two magnetic fields strength, is *independent of any external forces and factors*, like gravitational fields forces of other objects or entities. The instance the two magnetic rays or magnetic fields interlock, to generate the first Gravitational fields as an entity, at that moment the span of the Magravs of the two fields forces involved in the original fields are not as individual fields, but they become an *independent entity* in respect to their environments’ fields forces. That is the reason why the **Mass of any entity**, like the Matters, are *constant and independent and irrespective of the environmental Magravs forces* which the entity exists and operates within.

That is the reason why the Mass of spacecrafts of the future, or the Mass of planets like the Earth are independent of their environmental conditions. That is to say, the **strength** of the interaction of the two **original** magnetic fields and the span of the fields they produces that created the Magravs *decide how much mass the interaction of the two fields will carry* and no other factors.

This is the reason why, Man will be able to make crafts that can carry any payload by the use of Grapos nuclear reactors Magravs strength, as the payload of these systems and the strength of Magravs that these systems generate, creates independent environment Magravs forces *as long as the boundary of the magnetosphere of the Grapos can cover the physical boundary of these crafts*. Figure 25, in its design, explains’ this concept fully. That is why the two magnetic fields have to be equal in strength before they can create the initial four field forces.

If two unequal magnetic fields in strength interact, the two create the gravitational, the Magnetic, the magnetosphere and the Mass, and the lowest strength field, which dictates the maximum level of the strength of the four fields that the Matter can have. The stronger magnetic field residuals of these types of unequal magnetic field strength encounters and creation of matters become the ***tail*** **magnetic field of the entity**, and this tail is the cause of the creation of the coiling effect of these kinds of masses. Where and if the two unequal magnetic fields strength ***cannot*** create a mass, the same interaction leads to the creation of the coiling of the two fields and to what is known as the ***coiling of the magnetic fields***, this effect is partially the cause of the *swirling of the F1* torus field zones in the plasmas, and its larger counterpart, the rotation of Wormholes, as these regions in the universe are created out of unbalance Magnetic field strength interactions and interlocking, this be it in the plasma or in the universe. In so many ways in knowing the strength of two magnetic rays, magnetic fields, plasmatic magnetic fields and so on, which are about to interact and to create the initial Magnetic and gravitational fields of an entity, one can tell what would be the outcome measure of the mass of the entity they will create.

Einstein has related the mass of matter to its energy through the general equation of relativity.

As explained above, the mass of an atom or plasma itself is made of a collection of dense pmtics strength and their interaction with each other.

Thus initial fundamental plasmas always have the same mass irrespective of the Magravs of their environment.

By the same measure, the Atoms of the same elements are always made of the same number of initial plasmatic magnetic field's sub-components, of neutrons, protons and electrons.

Atoms of the same elements always have the same total plasmatic magnetic gravitational fields **pull** and Magnetic field **push** in respect to all their internal components, irrespective of the environmental plasmatic Magravs that the atoms might be in. This becomes the total mass of an atom, as mass is an expression of measurement of the total enclosed pmtics interaction within the fields of confinement of the plasma.

In considering the general equation of relativity

$$E = m \cdot c^2$$

The energy for the mass of Matters of the plasma, accordingly will be

$$E = m(M) \cdot c^2$$

That is to say "the total pmtics released in motion by the Matter of plasma (once they have left the boundary of the plasma) (Energy), is equal to, the total balance after interacting pmtics strength of the Matter components of the plasma (Mass) which the plasma is made of, multiplied by the square of the maximum speed that pmtics can travel in the matter environment (speed of light)".

Considering that, pmtics of different strength can travel different distance from their point of release with ***different speeds*** in *different Matters,* Matter *and* matter mediums.

Then it is correct to say that the pmtics of the Matter, once it has left the Matter

Magravs of the plasma, this pmtics of the Matter will travel with a different speed than the plasmatic magnetic fields of the Antimatter of the same plasma, or be it the pmtics of the Dark Matter component of the same plasma, in the matter environment.

The Antimatter has and is made of different pmtics strengths, by virtue that it has been created initially by stronger fields, then its speed will be different and higher than that of the pmtics of its Matter component of the same plasma in the same environment.

By the same principle, the faster field of the Antimatter carry more energy or more powerful energies than the slower pmtics in motion of the Matter component of the same of plasma.

Alternatively, in reverse, as the Antimatter possesses more power or energy than the Matter, then for the same volume of Antimatter as Matter in the plasma; the Antimatter pmtics must move faster than the pmtics' of the Matter, in the matter medium.

The Fermi National Accelerator Laboratory - or Fermilab, in USA (6, 7) - considers a small amount of the Antimatter, once released from the Matter of plasma; it can deliver more energy than its Matter parts of the same plasma. They are considering and working on the principle as; they can burn the Antimatter to release its energy so that by the time the Antimatter reaches the state of matter energy, that they can harvest its surpluses energy.

Through the ***Keshe principle of conversion of energy***, the fundamental universal conversion principle follows the path that the Matters convert to matters through their **dynamic pmtics in motion and forces,** or what is refer to as the **energy of the Matter magnetic fields forces, as these energies are transferred to the constituent pmtics' of matter**.

From what I call "The fundamental ***universal conversion principle***" follows that "the pmtics of the Matters' like of Antimatter slow down, through interaction with other magnetic fields, so that their energy can be absorbed and converted to Matters' pmtics strength levels, that become useable in matter environment and appears as the matter (solid, liquid and gas)".

By the same principle, the matter can be converted to Antimatter through the same method in reverse. Where, these processes are not much different for both paths of conversions.

Only the state of pmtics strength and the effects of the environment, which dominates at a given point and at a given motional position (time), that will dictate what state the dynamic pmtics in which the state of Matters, the matter manifest itself as, and what matter will appears in the line of sight by given tools of detection at a given point in the universe. This dictates if pmtics strength will appear in state of matter as solid, liquid or gas, or will manifest itself in one of the state of Matters, as Matter, Dark Matter or Antimatter.

CHAPTER 6

The total energy balance equation of Plasma

The energy balance equation or general equation of relativity ($E = m.c^2$) gives the relation between the *tangible part* of Matter of the plasma and its constituent energy.

The mass of an atom or collection of plasmas of its components, themselves are made of a collection of densities of pmtics strength of different Matters and their interactions, in respect to each other within their respective initial fundamental plasma.

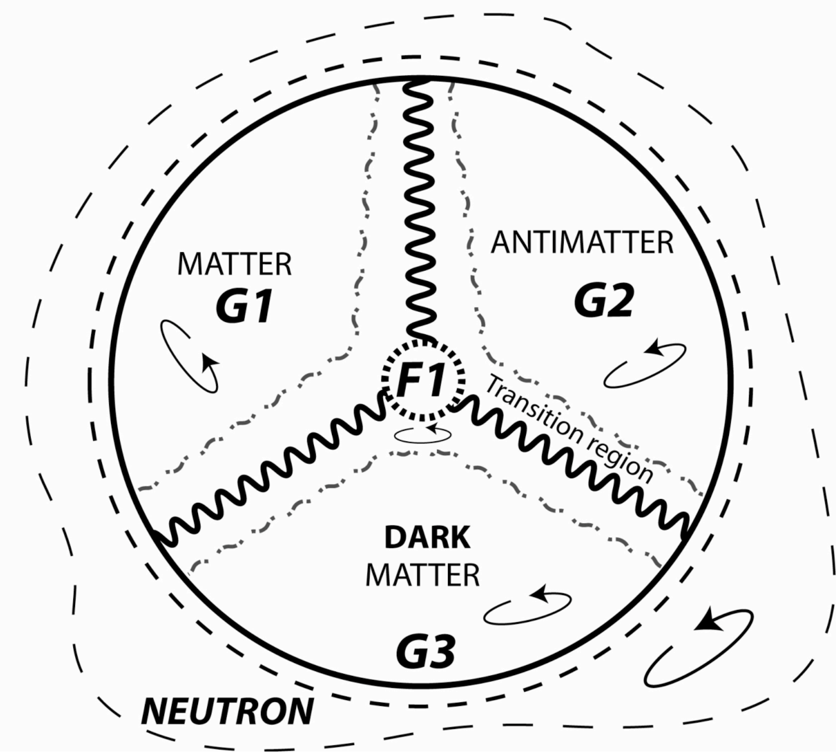

Fig. 26: Schematic of the fields of all components and of the initial fundamental plasma.

In considering the plasma, the general equation of relativity can be written in its real terms of its constituent components and their interactions and effects. Thus, this equation as has been said can be defined in respect to the Matter components of the plasma in the following manner.

That is to say "the total pmtics released in motion by the Matter of the plasma (once they have left the boundary of the plasma) (Energy), is equal to, the total balance after interacting pmtics strength of the Matter components of the plasma (Mass) which the plasma is made of, multiplied by the square of the maximum speed that pmtics can travel *in the* ***matter*** *environment* (speed of light)".

Knowing the new composition of the plasma, then the ***total energy of the plasma*** has to be in its ***totality of all its masses of all its Matters*** and not just its Matter component.

Therefore:

K (the total energy of plasma) = [energy of the Matter + energy of the Antimatter + energy of the Dark Matter + energy of the spherical torus-shaped field force effect + the energy in the inter binding between the other parts in the transition zones of the plasma].

K= [E (Matter) + E (Antimatter) + E (Dark Matter) + Ecf (centre spherical torus pmtics) + Etr (pmtics within transition zones)].

Where, total general equation of relativity for the whole of the components of the plasma can be written in the following form.

$$\mathbf{K = EM + EAm + EDm + Ecf + Etr}$$

Or:

K = [(The total components of constituent pmtics of the **Matter**) x (the square of the **speed** of pmtics in the **Matter medium** or the speed of light) + (The total components of constituent pmtics of the Antimatter) x (the square of the **speed** of pmtics of the Antimatter in the Antimatter **medium**) + (The total components of constituent pmtics of the Dark Matter) x (the square of the **speed** of pmtics of the Dark Matter in the **Dark Matter medium**) + The total energy components of constituent pmtics of spherical torus-shaped field force effect + The total components of constituent pmtics of the inter binding between the four other parts in the transition zones of the plasma].

$$K=M\ (M)\ .\ c^2(M) + M\ (Am)\ .\ c^2\ (Am) + M\ (Dm)\ .\ c^2\ (Dm) + E(cf) + E(tr)$$

Note: Capital M indicates here a mass of Matter.

When for example an apple is measured using the present measuring tools and methods, the figure attained is only for the mass or the weight of the Matter part of all the plasmas' of atoms of the whole apple.

As a representation, one initial plasma of an atom in the whole construction of an **apple** with it's Matter, Antimatter, Dark Matter, and other pmtics are shown on the right hand side picture of figure 28. Where, the *tangible apple* ***as matter*** represents the total Matter components of all the pmtics of an apple is shown on the left picture of figure 28.

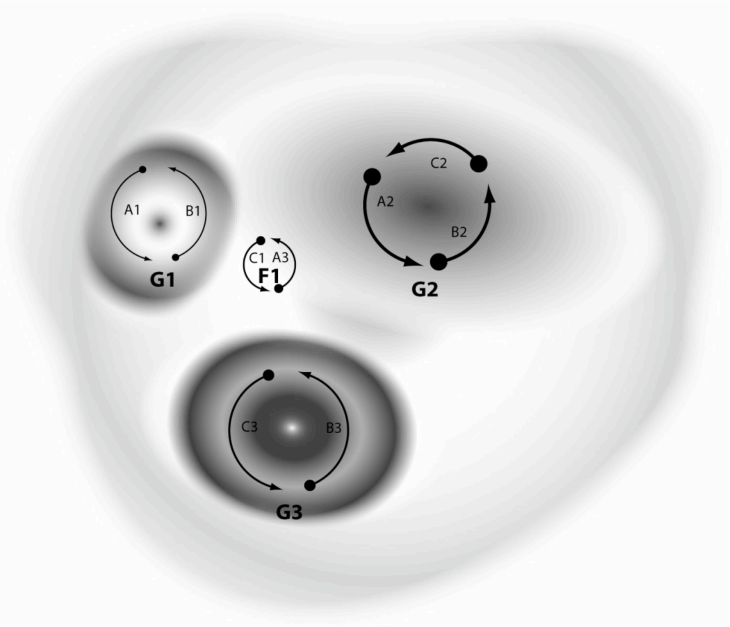

Fig. 27: The pmtics of all components and of initial fundamental plasma.

The total weight or mass of the plasma has to be the total weight and the total mass of all the components pmtics and all of all the plasmas of Matters of the initial fundamental plasma.

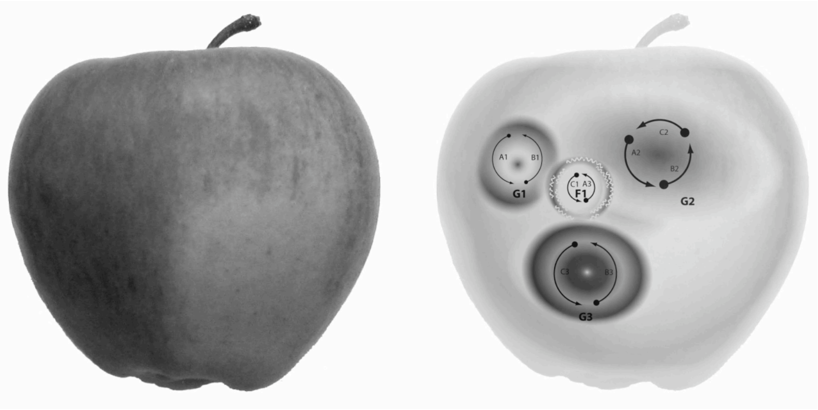

Fig. 28: Newton's apple has different types of pmtics and Matters.

Note: When the Matter is part of the plasma, this is denoted as ***Matter*** *(G1 in Fig. 26, Fig. 27 and Fig. 28) and when the same is considered as tangible matter, it is noted* ***as matter*** *(Apple in Fig. 28), of an atom or an object.*

By the same measure, if one has managed to separate the Antimatter of the plasma from its other constituents of the Matter and the Dark Matter of the plasma to confine the Antimatter in separate container as the Fermilab has claimed. The question to ask is that, what is the weight or the mass of the balanced leftovers of the Matters, (or Matter, Dark Matter and residual pmtics) of the plasma? By the laws of physics, one cannot have something (Matters/Energy) out of nothing. One cannot have mass of Antimatter component of the plasma, when its mass is not calculated in the overall mass of the plasma, or the leftover mass of the Matter and the Dark Matter of the same plasma must be less than before the Antimatter was taken out.

On the other hand, the question is that, is the present used methods of measuring the weight and mass of the plasma, measuring the ***total mass*** of ***all*** *the Matters* of the plasma? Alternatively, does one has to break the present numbers of mass and weight of each Matter, up to its sub-weights and sub-mass of the Matter (Fig. 26 and Fig. 27 G1), the Antimatter (Fig. 26 and Fig. 27 G2) and the Dark Matter (Fig. 26 and Fig. 27 G3) of the plasma. Further, the Magravs of the Matters of the plasma in conjunction with their own Magravs and other pmtics in the plasma create their own additional gravitational and Magnetic field forces, their own additional mass, as well as the individual masses of the Matters of the plasma. Thus the total mass of the plasma has to be more than the total mass of individual Matters of the plasma.

As Fermilab has separated the Antimatter components of the plasma, then, the question to ask from them, is that, if the mass of the matter of the plasma they hold still has the same mass as before without the Antimatter in it? So, what is the mass or the weight of the Antimatter that has been separated?

This applies the same to the Dark Matter components of the initial fundamental plasma too.

As has been said before, the total weight or total mass of the plasma is and has to be the total weight and total mass of all components of the Matters of plasma. This includes the masses or energies of the magnetic fields force of the F1 and the residual fields in the initial fundamental plasma.

Thus, there is a *fundamental error* in the present methods and technologies of measuring the weight and the mass of the plasma systems as a whole.

That is why suddenly one can see strange results from the Antimatter component of the plasma of the same mass or weight in the matter environment.

If, the total dynamic pmtics of the initial fundamental plasma, would have been, measured correctly from the beginning, then there would have not been such an ambiguity and mystery arising from the sudden hidden energy sources called the Antimatter, or the powers of the Dark Matter of the plasma.

The fact is that, the Antimatter and the Dark Matter possess plasmatic Magravs forces and that is why, they are called the **Matters**, as they possess mass and energy.

Hence, if there is a gravitational field's force of the Antimatter or the Dark Matter components of the plasma, then these Antimatter and Dark Matter have to have mass and they have to have weight in respect to the other Matters in the plasma and in respect to their environments pmtics, which the plasma is in, at given time. Therefore one needs the measurement of both of the mass and the weight of the plasma as whole as an entity as well as all the individual masses of each Matter too.

Then, the same as with the atomic mass and atomic weight of an atom measurement phenomenon of today's science, **there is a need for a new scale of measurements**, and new measuring tools to be brought about and developed for the true and total pmtics measurements of the constituent Matters and fields of the Initial fundamental plasma.

By this new total weight measurement, then the ambiguities in the present plasma physics and nuclear physics will be sorted.

That is to say then, where and when one gives the measure of the mass or the weight of the plasma, then at this point, one has to give the mass or the weight of the Matter, the Antimatter, the Dark Matter and the rest of components of the plasma in a given motional position, and the position in the life of the plasmas independently, for all components of the plasma, and one measurement for the plasma collectively as the whole needs to be given too.

The reason for stating the position and motions of the plasma is due to the fact that, the plasmas having dynamic characteristic components, plasma itself as a whole is a dynamic system too and continually gains or losses pmtics from or to its surrounding pmtics environment respectively. Therefore, the total pmtics of plasma, in different frame of motion (time), will be different, from one point to another in a given space.

In the future measurements, one has to be clear about all measurements of the Dark Matters as well as the Matter as well as the Antimatter components, as well as all magnetic field components, the likes of the F1 or maybe several F1 configurations in any given pmtics of the initial fundamental plasma. Therefore, there is a need for correct figures for the mass, the weight and the energy of the initial fundamental plasma, because such data are needed for in-depth research and developments of various kinds of applications in Magravs positioning systems using components of Matters of plasmas as the Matmags, for future space travel.

Another point to consider is the Matters environment in which the weight or mass of the plasma is carried out. That is if, this is carried-out in the Matter plasmatic magnetic fields, Antimatter pmtics strength, or Dark Matter pmtics strength environment.

Similarly, the plasmatic magnetic fields' strength in which the measurements are taken should be considered. This being, in what fields strength in the Matter, the Antimatter or the Dark Matter field strength or their respective pmtics energies measurements are taken place.

This is the true nature of the complexity of the relationship with the mass, weight, energy and their interactions in the world of Matters. These cannot be so simplified with one equation of relativity with three simple denotations, if in the future scientists strive to achieve the universal method of producing energy and motion.

The space travellers of the future will come truly to understand that a small error in the measurement of all matters and Matters components, the environmental magnetic plasmatic strength and so forth, can land them in strange and peculiar environments, which these were not planned for or expected. These small errors can end them up in new conditions and positions in the galaxies, which could be nice for experimental purposes, but not necessarily good for the health or the life of the passengers of these future crafts using Matters *Magravs positioning systems* (Chapters 22 and 23).

It is possible that due to miss-calculations, one part of the systems of the future crafts, or even the reactors of crafts, which are creating the conditions for motion or energy, will experience different environmental and Matter conditions than the rest, of the craft or the reactor. As, for example, the same system can land in the border of two different pmtics strength environments, or, in a different plasmatic condition than was originally planned, because of the miscalculation of the Dark Matter Magravs strength.

A physical comparison of the two state of Matters can be for example, considering a cube of sugar, half-submerged in a hot liquid, like tea, where, half of matter of sugar is in the solution of the hot liquid and the other end of the cube being between the fingers of the dipper, in solid state.

Therefore, the question emerges, of how does or can one put the original components of the sugar cube together to get back to the first state of the solid cube, in its original solid state and shape.

The rearranging Matters through the understanding of the Matter energy balance will be the solutions for these kinds of mistakes.

The misunderstanding and potential pitfalls with space technology of the future will be with the miss-calculation of the loading of matters as the “Matmags”, and the control in the reactors of crafts of the state of Matters transitions will be interesting.

Nevertheless, the travels and space ventures of the future through miss-calculation of energy balance and wrong Matter Matmags loading will not be without its problems and sorrows, as have been experienced in the short lifetime of the space explorations of today for men whom have dared to test the limits of space explorations.

Who dares, in the space traveller of future, may succeed to have a taste of the pleasure of the unseen in the universe! Our hope is that they come out of these journeys much wiser, but not necessarily less darer than before.

CHAPTER 7

The creation and speed of light

One of the possible ways that light in the universe is created can be through interaction of pmtics' of the plasmas. The speed of light is not considered to be the ultimate speed of Matters. We will discuss how, the interaction of pmtics of the flight systems using Magravs positioning reactors would lead to the creation of light around such a system in the atmospheric conditions like that of the Earth.

Understanding the basic structure of the initial fundamental plasmas, the rationale that the speed of light is the ultimate speed of travel of any Matter becomes questionable.

The phenomenon that the speed of light is the ultimate speed of travel of any Matter(s) or matter(s) is an assumption that has a clear flaw in its concept and cannot be correct for all states of Matters in the universe.

In the iron-based magnets of the solid state of matter, the magnetic fields flow is in one direction, out off the North Pole and their return through the opposite South Pole.

The question that needs to be asked is that, what is the ***speed of the flow*** of the magnetic fields force from one pole to another in solid state of matter?

For ease of understanding let us consider two packs of dynamic plasmatic magnetic fields strength of the pack A and the pack B (Fig. 29), in the path of collision with each other.

The pack A and the pack B can be considered to be of the plasmatic magnetic fields of the plasmas, solar systems or galaxies. The other assumption is that the pmtics strength of these two sets considered not being equal. These two sets of pmtics are supposed to be the mixture of the pmtics of gravitational and Magnetic fields of their respective Matters or objects in each pack.

In considering three snapshots a), b) and c), the motion of the two packs A and B in a given space of time and pmtics strength of the environment area of the D in respect to each other.

a) As the pack A and the Pack B are in motion in space and heading towards each other and these two fields are set to collide with each other (Fig. 29).

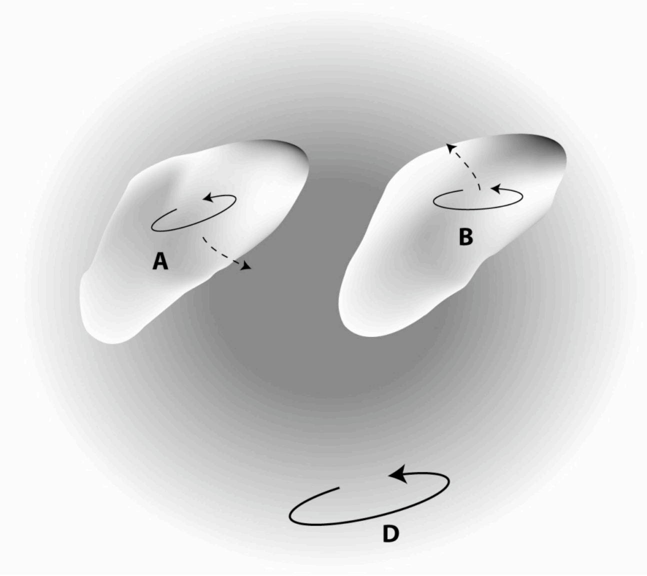

Fig. 29: Two unequal pmtics packs heading for collision.

b) Further, in the process of their motion, the two pmtics collide with each other (Fig. 30).

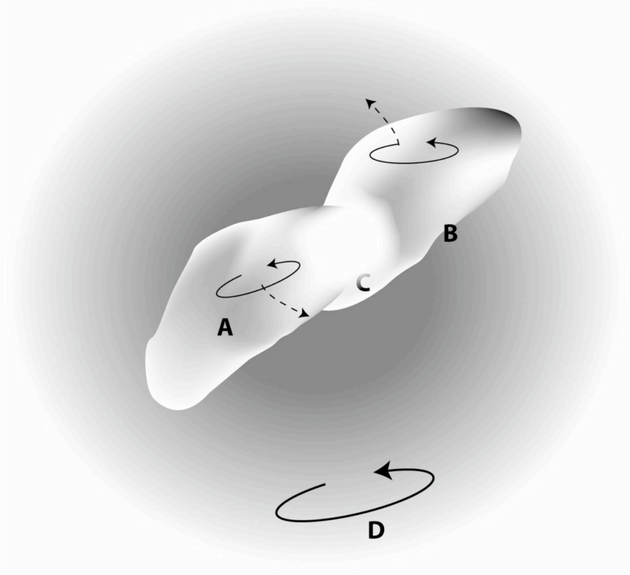

Fig. 30: Two unequal pmtics packs colliding and releasing lower order of pmtics strength as light (C).

c) Further, in the process of the interaction of two original pmtics of A and B, the two fields separate (Fig. 31) and carry on in the universe with some of their fields being released during the collision.

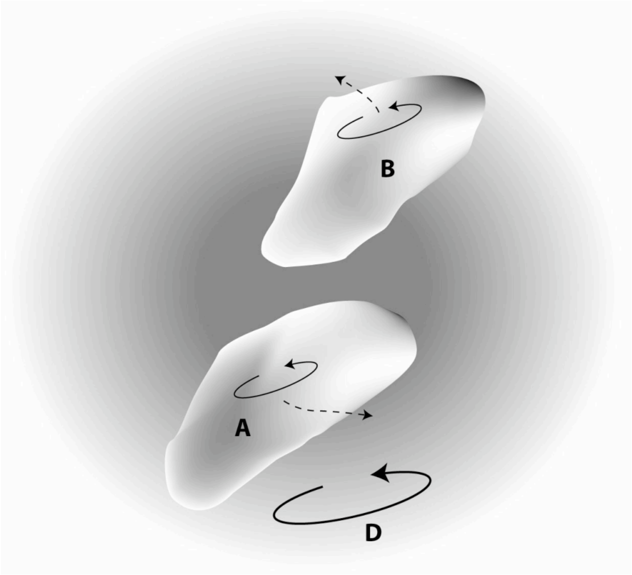

Fig. 31: Two unequal pmtics packs after collision moving away from each other.

Considering the point (b) of impact and collision as in figure 30, as there is no physical matter in these zones, only the pmtics forces can collide and interact.

The collision of the two plasmatic magnetic fields due to their frictions with each other, this leads to separation and production of fragments of pmtics, which some will be of slower plasmatic magnetic fields strength than the two original pmtics of the two original packs.

In case of physical matter, it is normal that when two matters collide head-on, they slow down and some parts of their matter are sheared off and released as fragments.

Through the collision and the interaction of these two packs of pmtics with each other, and the friction between them, these cause separation and fragmentation of some of the same pmtics from the two original pmtics of the two packs. The collision in between two sets of magnetic fields causes some of the pmtics to slow down. In the case of pmtics collision, these slowed down fragments of original two packs of pmtics are slower, and possesses weaker strength pmtics compared to their original two pmtics strength, which were involved in the collision.

We can assume that the original two pmtics plasmas speeds of travel have to be much higher by order of magnitude than the fragments that are created by the collision and interaction of the two packs of magnetic fields forces.

In other word, when two pmtics are colliding with each other, the outcome is partly slowed down fragments of pmtics as the residuals of this collision. These slowed down plasmatic magnetic fields fragments of the collision of the two stronger plasmatic magnetic fields, are slowed enough to appear as, and to be of light magnetic spectrum within the matter magnetic fields strength spectrum of the matter detection zone.

Nevertheless, one can say that, when the two pmtics of A and B collide, the outcome are residual pmtics fragments of the collision, which some of their slowed down magnetic fields fragments created due to their friction and collision, leads to the creation of slower pmtics strength in the range of visible light (Fig. 30 field zone C). Where, this visible range of the magnetic wavelength is part of the overall universal order of magnitude of the magnetic wavelength spectrum fields.

In the universe, when two or more plasmatic magnetic rays, fields, plasmas, of Magravs collide with each other, some of the residual pmtics fragments produced by their collision will be slowed down to magnetic rays or plasmatic magnetic fields strength levels, which the strength of these magnetic rays or magnetic fields are in the range of magnetic fields strength spectrum of the visible light. In other word, the outcome of the interaction of two or more magnetic rays, magnetic fields or plasmas of Magnetic and gravitational fields (Magravs) is the production of light.

This being considered as "one of the prime methods by which visible light is created in the universe. These being the residual of slowed down weaker magnetic fields strength fragments of pmtics released by the collision of two or more pmtics of Margraves of plasmas with each other".

The speed of pmtics fragments, which has slowed down due to a head on collision of two pmtics plasmas, cannot be faster than the original two pmtics plasmas themselves that have caused it to be produced.

Thus, the ***speed of the light***, which is created by interaction of two faster fields' head-on collision, as a fragment and slowed down field, cannot have a faster speed and more pmtics strength than the fields that have caused its creation due to their collision.

For example, by the slowing down of the two faster pmtics of the Earth and the Sun that have not been visible, after collision with each other, and slowing down, they create the slower plasmatic magnetic fields' strength and slower speed magnetic fields in the range of the speed of visible pmtics as of the daylight.

It is correct to accept that "the pmtics, through their collision, have to slow down so that they became slow enough that they become in the range of the slower wavelength in the visible light pmtics wavelength spectrum".

It is obvious that a slower fragment cannot be faster than the fields that have caused its creation. Therefore, "The speed of the light is not and cannot be the ultimate speed of motion" when itself is created by interaction of faster or stronger plasmatic magnetic fields.

In universal order of magnitude of the speed of motion, I consider "the speed of light to be somewhere in the lower end of the overall universal magnetic fields speed spectrum".

Thus, the speed of the light is only the fastest to human methods of detection and not necessarily true in other levels of universal order of motion.

It can be said that the speed of the light is “the ultimate speed that can be attained by pmtics in the Matter components of the plasma”.

Another, fundamental point to realise is that “the speed of travel of light is **different** in a Matter environment than the Antimatter environment and even in the Dark Matter environment”.

Similarly, it is correct to say that “the light created by the interaction of the Antimatter pmtics plasmas and the Dark Matter pmtics plasmas, they have **different** speeds than the light created by the interaction of Matters’ pmtics plasmas”. These faster speeds of magnetic fields than the pmtics of the light spectrum are not detectable by the present tools in the scientific state-of-the-art technology. But their effects and power and strength has been seen and observed in the reactors, which have been built and tested for lift and motion in the past years.

Considering the collision of two pmtics like those of the Earth and the Sun, where the Earth’s Magnetic and gravitational pmtics (Fig. 32 field B) being made through interaction of magnetic fields of Matters in the inner cores of the Earth (Chapter 1), and the Sun’s Magnetic and gravitational pmtics (Fig. 32 fields A) being made by the matters, in the cores of the Star. (In this case, we do not consider the pmtics of Matter, which are released by interaction of matters pmtics on the surface of the Sun).

An example of the creation of the light through the interaction of pmtics:

As the pmtics of Magravs produced and released by the Sun, reach’s’ the boundaries of the Earth pmtic of Magravs, these two pmtics of Magravs collide with each other at the edges of the Earth’s upper atmospheric levels.

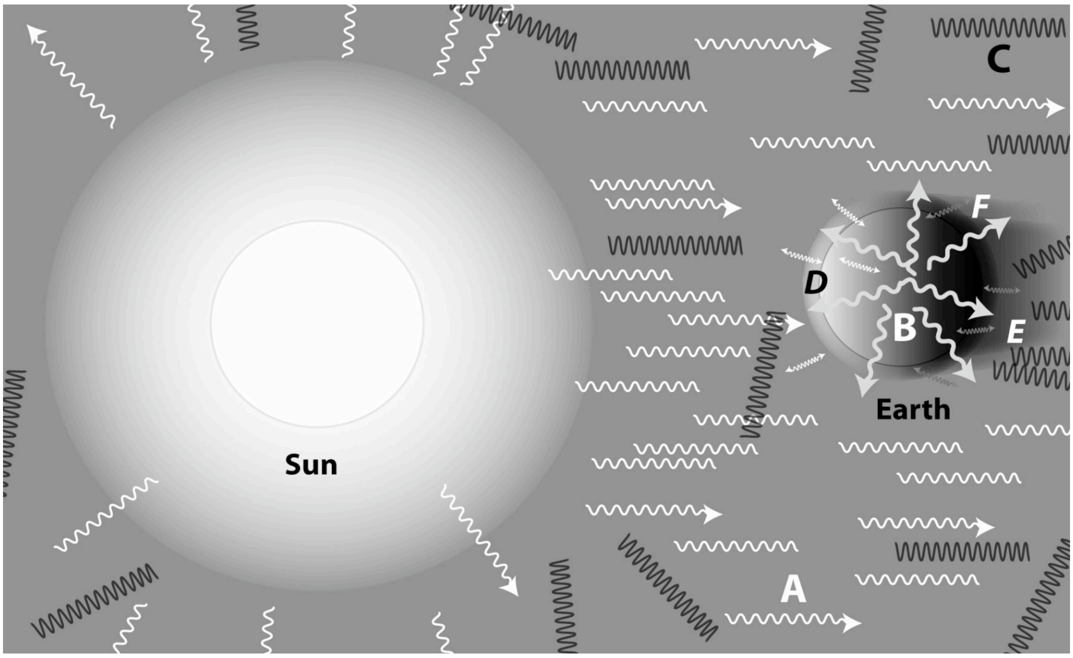

Fig. 32: Interaction of three different types of pmtics strength (A, B, C) and the resulting three residual fragment pmtics (D, E, and F).

Where, these interactions leads' to the creation of the pmtics interface, or the **magnetosphere** of the planet. Where, this magnetosphere of the Earth is partially produced at the same time in respect to the other surrounding pmtics of Matters and other plasmatic magnetic fields (Fig. 32 field C) in the solar system as well as the pmtics of the Sun.

Through the pmtics plasma interaction principles, the pmtics of the Magravs of the Sun interaction with pmtics of Magravs of the Earth leads to the creation of new residual ***slower rays and pmtics strength in the light spectrum*** (Fig. 32 field D), where, the overall effect of these pmtics fragments of light generated, leads to the creation of the *daylight* on the Earth, in the interfacing sector of the Earth with Sun's pmtics only.

The light generated in the sector of the planet, facing away from the pmtics of Magravs of the Sun, in interacting with the weaker pmtics of the Dark energy of space around it, due to the weaker plasmatic magnetic fields' strength of these pmtics of Dark energies in space (Fig. 32 field C) and the same constant pmtics strength of the Earth (Fig. 32 field B), in this sector, **weaker strength pmtics residuals** (Fig. 32 fields E and F) are created and released, thus, the appearance of the darker sky or the *night light*.

Thus, the interaction of the *same pmtics* of Magravs strength of the Earth (Fig. 32 field B) with the *stronger* plasmatic magnetic fields of the Sun (Fig. 32 field A) leads to the creation and release of more powerful pmtics strength residual fragments in the light spectrum (Fig. 32 field D), hence **the daylight**. Similarly, the interaction of the *same pmtics* of Magravs of the planet (Fig. 32 field B) with the *weaker pmtics* of the Dark energy (Fig. 32 field C) in space, away from the Sun's plasmatic magnetic fields, leads to the creation of weaker pmtics strength residual fragments, hence weaker lights or dark night light and creation of the translucent plasmatic darkness (Fig. 32 field E and F) light of the night.

As the Earth's reach a more Magravs pmtics strength balance and equilibrium with its environment's plasmatic magnetic fields' strength, thus the darker that part of the atmosphere it appears to be (This principle is described in Chapters 12 and 13).

Through the same principle of interaction of and collision of solid matters of the cosmic dust and the pmtics of the Earth, the northern lights in different colours are created. Where in this case, the light created is due to collision of the matter pmtics and pmtics of Magravs of the Earth.

It is not correct to consider the speed of light to be the ultimate speed of motion, for any pmtics of Matters environments existing within the universe. Similarly, it is conceivable that pmtics can fuse and that their pmtics can increase in strength and consequently they can increase their speed too. (Paper: Creation of magnetic fields).

In principle, the speed of light is not only achievable by the reduction in pmtics of higher order of magnitude pmtics strength, but conversely it can be achieved from the addition or adhesion of the strength of lower order of magnitude pmtics strength too. By the same measure, the pmtics of the light can be integrated, to create their stronger and faster speed of motion, with new and faster pmtics from their interactions too.

The **pmtics strength** and **the speed of plasmatic magnetic fields** is environment dependent and there are no limits in its lower or higher orders of magnitude. Generally, as long as plasmatic magnetic fields can interact, they can add to pmtics speed or lose their speed.

The limitation of speed of travel to plasmatic magnetic wavelength of light is the restriction that Man has forced upon himself, purely due to lack of his understanding of the true construction of Matters of his universe.

The present limitation of the speed of the light for motion is very much similar to the earlier century's hypocrisy and lack of knowledge of Man, with the belief that the Earth was flat and at the centre of the universe.

The pmtics become visible light to the human eye, purely due to the fact that human eyes are adapted through their protein molecular pmtics strength chain, for them to be able to interact and recognise them. This range of the plasmatic magnetic wavelength is translated for the human brain to be the visible light for humans to be able to see and function. Where, this visible range of the magnetic wavelength is part of the overall universal order of magnitude of the plasmatic magnetic wavelength field's strength and not necessarily its ultimate speed.

Thus, "if the human proteins pmtics strength had a different chemical chain, that is to say, if the protein chain did not have hydrogen, oxygen, carbon and nitrogen in its composition, then man would have had different overall pmtics strength in his protein and the universe would have been visible to man in a different range of plasmatic magnetic wavelength field's strength spectrum".

This phenomenon of the protein chain pmtics strength of creation is explained in more detail in the book to be published, titled "The universal order of existence (9)".

Further, in this book we will talk about new nuclear-based systems, which have been developed for them to be able to make use of the creation and control of pmtics of Magravs strength for motion. We have called these systems ***GRA**vitational **PO**sitioning **S**ystems* (**GRAPOS**). From, tests and what we have seen and observed in the past years in proof of pmtics creation and control, the following can be discussed in respect to the interaction of the pmtics of the Grapos Magravs and the pmtics of Magravs of the Earth.

Using Magravs positioning reactors, the interaction of the Magravs generated by these reactors with Earths' Magravs, these interactions of the two of Magravs, by their natural phenomenon of creation and interaction of pmtics, these Magravs interaction of the Grapos reactor and the planet, always leads to the creation of light around these systems in planets' atmospheric environment.

Crafts using **Grapos for lift and motion** in their movement and travel in pmtics plasma of Magravs environments, like the Earth's atmosphere, **these systems generate pmtics fragments of rays of light and appear** to be like **moving *bright lights* in the skies** (Fig. 33 disk 3 and 4).

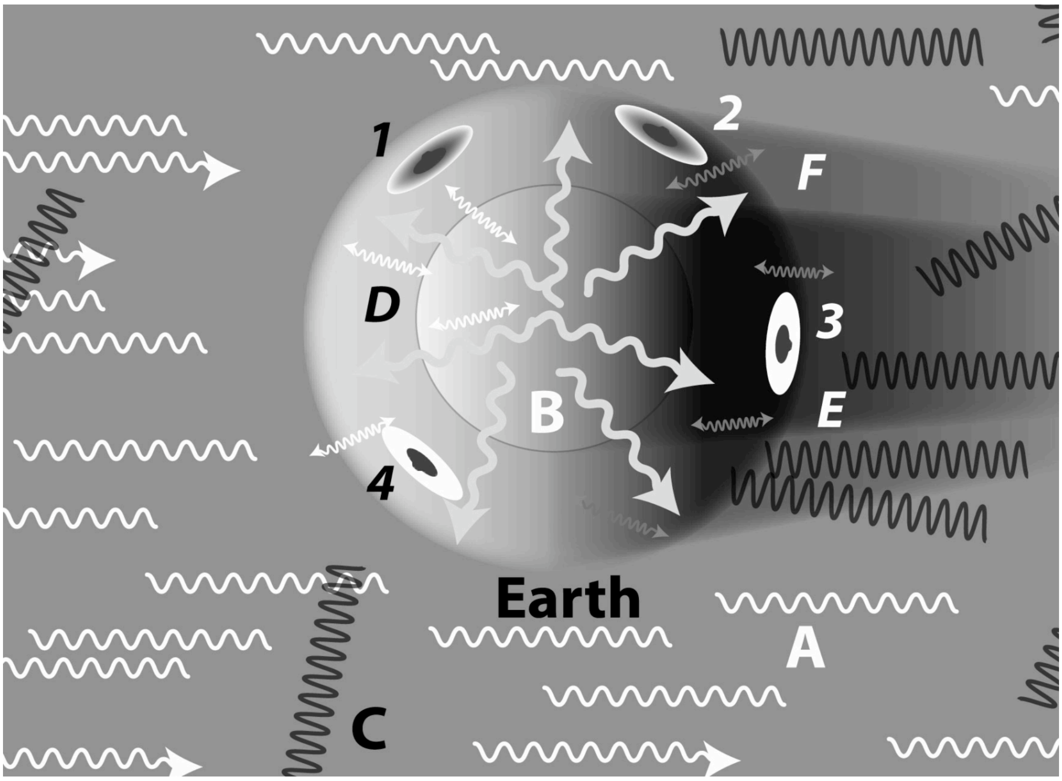

Fig. 33: Visualisation of Magravs positioning systems (1, 2, 3 and 4) within Earth atmosphere.

The effects of the interaction of the *four different Grapos and their different* pmtics strength and lights they produce around these systems, due to their of Magravs strength in respect to the Earths' pmtics are demonstrated in the figure 33 disks 1, 2, 3 and 4.

The **brightness** of these lights in sky is due to the pmtics of Magravs strength of Grapos' in respect to their environment pmtics of Magravs of plasma strength of the Earth, and these *lights are balance pmtics strength dependent*. Lights around these crafts are produced by the same principles and are similar to the way light is created by stars.

Another example of the Magravs positioning reactors to be used in future craft systems' is, in creating a balanced Magravs strength in respect to the Earth atmospheric plasmatic magnetic fields' strength, so that there would be not much field strength difference and hence nearly no interaction between the pmtics of the system and the Earth. Thus, no or very little light created around these systems as is shown in figure 33, with disks lights 1 and 2. So the system or spacecraft in a given position and time in the Earth atmospheric condition can become dark translucent in respect to the Earth pmtics strength. This giving the impression that, the system has disappeared from the sky, in fact, the system has created its own balance plasmatic magnetic fields' strength condition of dark light in its environment in respect to its environments Magravs strength, while the system still is/or can be keeping the same position in atmosphere. By being able to adjust the Magnetic and gravitational field's strength of Grapos reactor(s), these crafts can change their visual appearance in the Earth atmosphere, from dark to bright light and deemed-light and so one. Four different Magravs strength of Grapos in Magravs of the Earth are shown in figure 33 and figure 55 in disks 1, 2, 3 and 4.

Man can and will travel several times faster than the speed of the light in the future. This will be of his own choosing and limited by his technological advancement and the limitation of understanding of his own tangible existence.

The speed of motion, due to the interaction or gravitational pull of the planets and stars of destination, in pulling the crafts equipped with Grapos towards them is beyond the present technological ability and comprehension of the scientific world, but in time this will be as easily and manageable as daily atmospheric flights of today. Which these crafts can travel several times faster than the speed of light.

The other scenario of collision of pmtics around the Earth could be that when the two field Magravs of the Earth and the Sun collide, and their overall field interactions can lead to the creation of faster pmtics than that of visible light. This again is the result of the amalgamation of two pmtics strength forces from two celestial objects.

Theoretically this is not possible around the Earth as most of the fields of the Earth are outward going and crossing the path of the pmtics of the Sun, and these fields would be carried away and out from Earth’s atmosphere and not inwards. Therefore, any amalgamation of two pmtics of the Sun and Earth has to be inward so that Earth’s atmospheres can generate the daylight.

This phenomenon of the addition of two fields to create light in the case of the high speed of the plasmatic magnetic fields of the Sun and the Earth is not a common possibility, or there is so very little of it.

By the pmtics theory, in the case of an atom’s structure, pmtics of Magravs of the constituent Matters of the electron interact with the pmtics of Magravs of the proton. This creates the fuzzy light around the electron fields in motion, which makes it dimensionally easier to detect the presence of the electron than its nucleus. This fuzzy light being similar and created through the same principle as to the interaction of the pmtics of Magravs of the Sun and the Earth, as was explained.

Considering the interaction of the pmtics of matters on the surface of the Sun: The lights which can be observed from the surface of the stars is primarily due to the interaction of the pmtics of the gravitational and Magnetic field from the centre of the star and their interaction with the Magravs of the plasmas on the surface of the star, where this interaction of the two fields leads to the release of pmtics in the visible light spectrum. The intensity of the light of the stars is in principle due the strength of the gravitational and Magnetic fields generated in the centre of the star, as the pmtics of the hydrogen plasmas on the surface of the Sun are considered to be constant.

We consider the pmtics received by the fields of the Earth from the pmtics from the matters from the surface of the Sun to be small and to be a fraction of the power of the pmtics released due to the gravitational and Magnetic fields from the centre of the star which reaches the planet Earth. The only time when the pmtics of the materials on the surface of the Sun has any effect on the interaction with the Earths' gravitational and Magnetic fields is when there is a surge in the volcanic eruptive activities on the top layers of the materials on the surface of the Sun. These eruptions cause the release of large quantities of physical material of plasmas into the solar system.

CHAPTER 8

Dark Light phenomena

In the previous Chapter we explained that the *visible light* is created through the interaction of Magravs and the *light* created due to the difference in strength of Magravs. Opposite to this, must be the interaction of balanced Magravs strength where the weak or balanced Magravs fields strength, which leads to the creation of a type of night light or dark light at nights.

The point has reached in physics, that these balanced pmtics Magravs strength conditions needs to be considered on their own, where the **Dark light phenomena** can be introduced as part of the science of physics.

"**This Dark light phenomenon**, with Dark light intensity, can be said to be created and be due to the near to, or balanced, equilibrium in pmtics strength reached, or that can be achieved, through the interaction of two or more pmtics forces in respect to each others in a given environment". Where, the intensity of the darkness is pmtics strength balance dependent (Fig. 36 zone C).

The level of measurement of this darkness is not field strength dependent, but *plasmatic magnetic fields strength balance dependent of one pmtics in respect to another pmtics.* That is to say, the interacting pmtics environments can possess any magnetic strength, but in all cases, when the two pmtics strength are ***near equal***, then the darkness phenomenon of the same dark intensity prevails and the same type of dark light phenomenon occurs, irrespective of the original pmtics strength involved.

If this phenomenon is fully understood, just as man has developed night-vision tools, equipment can be produced that can view Dark light plasmatic magnetic fields strength balance in respect to different pmtics strength. Then the universe would be full of new objects that have never been seen in light energy matter conditions.

What a wonderful world would one be in and see, through these tools and machines, Men of vision would take pleasure and treasure by looking through from such a machine.

CHAPTER 9

Relation between Gravity and Light

Through the pmtics theory principles, as gravitational fields are made of interaction of pmtics, and the light being itself made of pmtics, it is inevitable that entities made of the same, will interact with each other.

From previous chapters, it is clear that the light is made of the lower order of magnitude pmtics strength in the overall pmtics strength spectrum in the universe. The light in essence is itself moving dynamic pmtics and that by the principle of this definition **light is energy**. Alternatively, it can be said that the light possesses energy.

We consider that light is a made of pmtics strength of a lower order of magnitude in the overall pmtics strength spectrum in the universe.

At the same time as has been explained, "Gravity and gravitational fields' (Chapter 1) are themselves made of interaction of at least *two **similar** pmtics* or at least *two **similar** magnetic rays*".

It is understandable that, any gravitational field's forces, as an entity, in essence are made of interactions of pmtics. Thus, gravity can and will effect any other entity in possession of pmtics, or entities being made of pmtics, like light. The light and the Magravs (known as gravitational fields) are both made of the same magnetic entities; it is normal that they have interactions and attractions in respect to each other's position.

Fundamentally, it is **not correct** to say that **light is gravitational dependent.**

But, it **is correct** to say that, "the attraction and interaction of **the light rays** in respect to, and from, another collection of plasmatic magnetic fields like **gravitational fields (Magravs) are position dependent".**

In classic physics, the behaviour of the light has been said to be gravity dependent. Through the new understanding of the creation of Magravs, this cannot be correct. As the light and gravity are made of the same magnetic fields of different strength, thus they can have near enough the same speed of motion in matter magnetic field strength medium.

In the case of light rays, when the ray comes in the vicinity of, or in position in space that it can be within a field of, or be near to or in the catchments area of a collection of Magnetic fields or gravitational fields (Magravs), these two entities' magnetic fields interact.

Thus, light is not gravitationally dependent, but in fact the attraction of light is position dependent from a pmtics of any Magravs area.

Thus, the stronger the Magravs of an object are, the stronger the pull.

Following the new principles of magnetic field interactions, the stronger pmtics of Magravs strength of the Matter, the Antimatter or the Dark Matter will exert their gravitational magnetic field forces on the weaker magnetic field strength of the light rays. Hence, the stronger pmtics of heavy matters and the stronger pmtics of stars, the Antimatter and the Dark Matter, pull weaker magnetic fields like light pmtics towards themselves.

So that the visible light will move slower as it will be pulled towards the gravitational force of these stronger, Magravs like stars and Antimatter zones and as it interacts with their pmtics forces.

Light has a different speeds in the Antimatter, the Matter and the Dark Matter due to their magnetic fields strength, very much like as it has slower speed in liquid than in air in the matter environment.

The difference between matter environment and the Antimatter and the Dark Matter environment is very much in that the former is pmtics compactness and density dependent, as in the latter it is pmtics strength dependent.

Thus, light, by the principle of its construction from pmtics, will be pulled and travel slower in other state of Matters and near large and stronger plasmatic Magravs strength of objects like stars, through the magnetic fields interaction principles or the gravitational pull of one stronger magnetic fields in respect to the weaker magnetic fields of the light.

In physics, this has been said to be like the effect of the bending of light when it passes a planet or star in possession of gravitational field forces. The principle of bending of the light, or gravitational lensing in vicinity of, for example, stars is purely due to the interaction of the pmtics of the light with the plasmatic Magravs forces of the star.

It is important to notice that in positions in space, where the Magnetic fields of the star supersede the gravitational fields' strength in respect to any given ray of light, the light ray as well as being able to bend inward due to the gravitational pulls, it can also be reflect away from Magnetic fields forces too.

By the same principle, it is now understandable, why and how for example in strong gravitational fields like black holes areas, even light appears as to get absorbed into them. This being due to the fact that Magravs strength in the centre of the black holes in these regions are so strong, that due to their magnetic fields strength they gravitate the magnetic fields of the light inwards or what they call in common terms "absorb" the light. This attraction is simply a gravitational pull of stronger Magravs strength of Matters and matters in the centre of a black hole in respect to the much weaker magnetic fields of the light.

Most certainly, ***gravity and its field strength is not warp dependent*** or gravity is not caused and is not created because of a warping of the matter in the plane of time and space, as has been discussed in some fields of present physics.

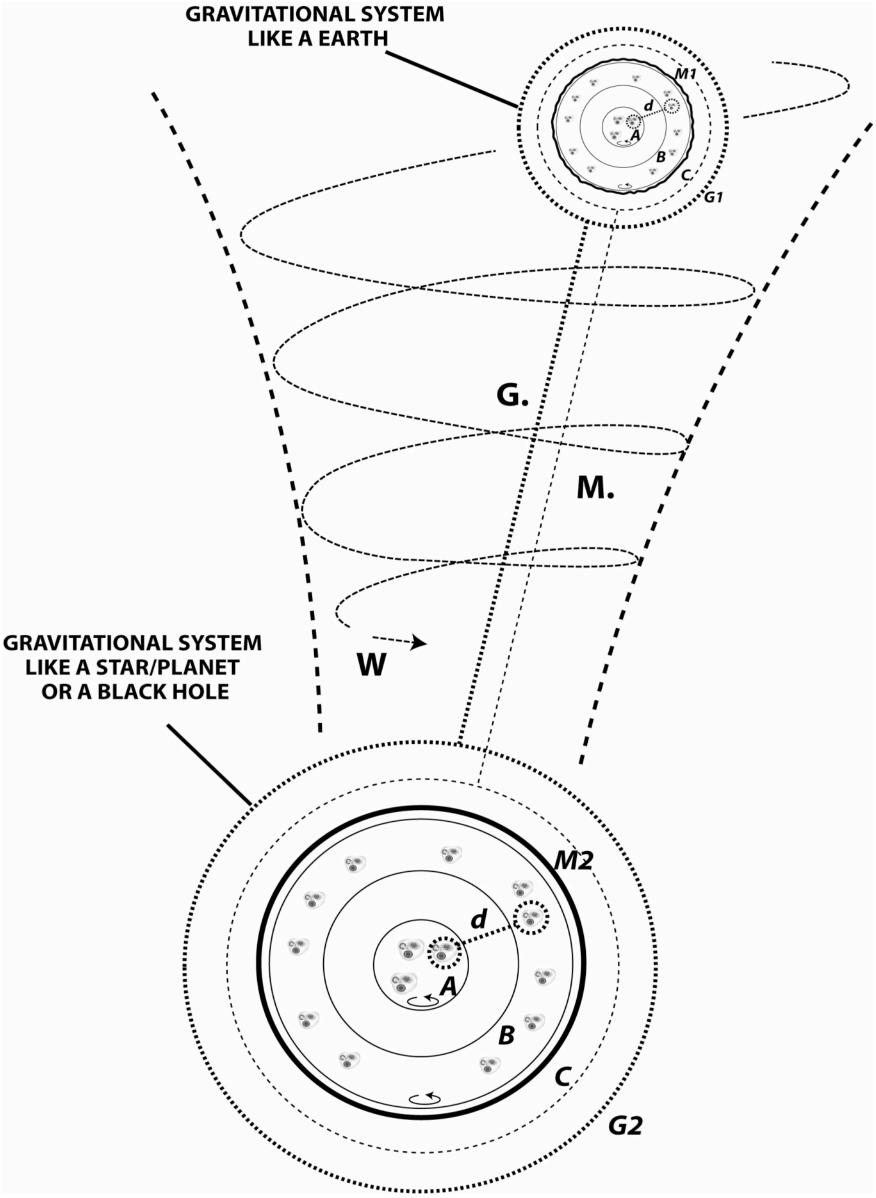
GRAVITATIONAL SYSTEM
LIKE A EARTH
M1
d
A
B
C
G1
G.
M.
W
GRAVITATIONAL SYSTEM
LIKE A STAR/PLANET
OR A BLACK HOLE
M2
d
A
B
C
G2

Fig. 34: Warping path of a weaker Magravs system towards the stronger Magravs system.

Note: Definitely the warping is not the creator of the gravitational fields, but the warping shows the path of the motion of the one gravitational field downward or toward onto another gravitational field.

That is to say, one weaker gravitational field can be attracted towards another stronger gravitational field G2 (Fig. 34), so that the path of the direction of the motion of the weaker gravitational fields G1, towards the stronger, could make the warp like path (Fig. 34 path W) in a given plane of time. This is the warping effect of a ***weaker** gravitational field* in respect of a ***stronger** gravitational field* and this is the warping path of attraction.

Therefore, warping of the matter is not the cause of the creation of gravity in the matter of the weaker gravitational field strength object, but in fact the warping is only an indication of the path of the attraction and motion of the weaker gravitational fields towards the stronger gravitational fields.

Celestial matters in possession of Magravs, through their internal pmtics and their attractions and interaction and nothing else, can generate the majority of their own gravitational fields in the universe. At the same time, the motion of two objects in the possession of their own gravitational fields can set and start different gravitational fields of one object in respect to the other. That is to say, when two stars in the universe come into each other's catchments zone, the catchments envelope creates its own gravitational pull, separate from the internal gravitational pull of the stars themselves. That is in fact the operation of two original pmtics Magravs being cascaded by the bigger one outside them. The interaction of the two stars, in trying to find their position, causes the swirling effect first and then the warping effect type of motion of the weaker stars' gravitational fields towards the stronger pmtics gravitational field.

External motion of matter does not necessary create gravitational fields inside an object, but the path and speed of the motion of matter indicates the strength of the pull of one field towards another and the direction of the path of the attraction.

There are no planes of references for this warping, as the original point of the reference of the interaction of the two pmtics magnetic fields in respect to each other comes from where, *they first come into touch* and start the first fields interactions; this becomes the plane reference point for the path of the attraction of one gravitational field towards another (Fig. 34)

In fact, the theory of warping creating gravity is as far from reality, as gravity being time dependent.

Maybe it is right to say, "In time, once two Magravs come in contact with each other and stay interlocked, the two fields might merge into one, this time of merger is then pmtics Magravs strength dependent".

As has been explained, Matter is one of the three parts of the Matters of the plasma. Thus, the same interaction between the three gravitational pulls of three Matters within the structure of the Matter (see Chapter 18 and 19) causes the warping and motion of Matters in respect to each other within the plasma.

In fact, any magnetic and pmtics entity, at all times of its existence, is gravitating or will have gravitation towards one or another magnetic ray or pmtics or Magravs in its regions and is warping towards one point or another in space at any given frame of time. That is to say, there are ***no regions in space in the universe, which are void of magnetic rays or pmtics***, and their gravitational interactions and attractions in respect to one field or one Magravs or other.

Nevertheless, there are different magnetic fields strengths, which in one form or another, are gravitated to one or another magnetic fields regions in the universe, in respect to their proximity of distance from a ray or a pmtics.

Maybe we can say “that gravitation in respect to time, is dependent on the position in space and pmtics strength of the any two entities, and that in time, for example light being in essence a magnetic field in origin, it will be effected or attracted or gravitated to another magnetic field or collection of fields, and that these magnetic fields can already possess gravitation in respect to each other.

Thus light rays in given point in space will come in touch with other magnetic fields, which they will gravitate towards each other, for them to merge, where this merger will take a certain given duration.

Thus, the light rays and other pmtics strength have a direct effect on the rate of attraction of one field to another. Thus, speed of attraction is pmtics strength dependent”.

In real physical terms, the light itself can be or become one-half of a gravitational field set, so not only is the light getting attracted towards another Magravs system, but also in this process, the light becomes part of a second magnetic field of a gravitational set itself in respect to the first gravitational field. Then this light pmtics becomes part of a plasma magnetic field, which this is, a new set of a light field and the first gravitational field interaction, and this itself becomes another Magravs system like.

The other important point to consider, is the time it takes for the light ray to reach or be attracted into or by a gravitational field, zone or area. This interaction and attraction, is both gravitational field strength, and the light ray’s magnetic fields strength, dependent.

That is to say, stronger gravitational magnetic fields, ***pulls*** rays faster to themselves than a weaker gravitational magnetic fields.
Thus, *the time, that it takes for a light ray to get attracted into the gravitational field zone is gravitational field strength dependent and not time dependent,* as the strength of the pmtics of the light of the matter is considered to be constant for all states of Matters.

Then for different Matters (Matter, Antimatters and Dark Matters) in the universe, there should be different *frames-of-reference* of duration or time. As the light created in the matter environment will have different speed in the Antimatter and Dark Matter magnetic fields spectrum environments of stronger pmtics strength regions. While the light created by the Antimatters and the Dark Matters pmtics, interactions have very **different speed** than the light created by Matter pmtics interaction. Secondly, the speed of the light from Dark Matters and Antimatters are much faster than the speed of the light of the Matter state or in matter state pmtics’ strength levels.

Some notes about light

1. The frequency of the light wave is created on the same basis and due to the same principles as the change of the polarity in the stars, for example the Suns’ eleven years half shift in polarity change (12, 18).

2. By the same principles, the alteration in the extreme points known as maximum and minimum points of return of the waves are Magravs positioning balance dependent,

and not as what has been assumed, electromagnetic field based, as electrons themselves are made of plasmatic magnetic fields.

3. The pmtics as rays, as they enter or come in touch with an atom or an object that they can move or cause motion in the penetrated atoms, electrons of matter or Matter or in their initial fundamental particles, because the matters or Matters components are similarly made of dynamic pmtics, thus, when a pmtics comes in touch with another pmtics, these pmtics interact and collide with each other. Thus, the outcome is a transition of pmtics from one entity to another.

4. The transition of pmtics takes place when the plasmas of Matters come into contact with the initial fundamental plasma of the Matter and they transfer their plasmatic magnetic fields strength from one to another. Usually this transfer is from the stronger to the weaker field strength directional of flow.

CHAPTER 10

The Antimatter

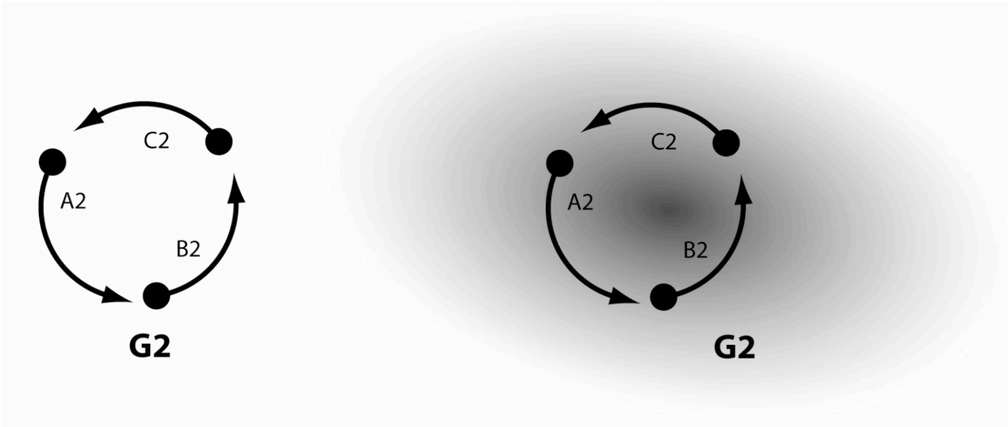

In the world of science, the Antimatter has been considered and said to be the largest source of energy for the future of the energy, space and motion technology.

In the structure of the initial fundamental plasma, I consider that "the Antimatter is the *main component* of the whole structure of the plasma and it is the carrier of the *largest amount and strongest* of dynamic pmtics strength in the plasma, which Antimatter can release the maximum and the larger amount of pmtics strength in the full spectrum of pmtics of energies from the plasma".

To understand what the role of the Antimatter is in the overall construction of plasma, we can simply look at the larger scale of the same type of entity and its effect in the universe. In the universe, the large-scale plasma conditions are said to be the like of the Stars, or the Suns' in their solar systems.

Where, the **Antimatter** part of the plasma is the same as the Star of a solar system, and it is the ***holder of the main quality and quantity*** and the possessor of the ***strongest pmtics strength*** of all Matters of the initial fundamental plasma.

I consider that, the Antimatter is the **source of the energy** and **the central gravitational point of the plasma** as a whole. The Antimatter is in fact the Star of the plasma, thus it is as the Sun to the solar system. In fact, there is nothing "anti" about this Matter component of plasma in the universe.

This is the reason why, in being able to free them from the sub-components of the Matter and the Dark Matter of the plasma, the Antimatter appear to be such a fantastic new source of energy.

As the Antimatter was made initially from the stronger magnetic fields strength in the plasma to start with, hence it holds the real fundamental source and the supplier of pmtics to all matters in the plasma. It can be said that pmtics released by the Antimatter are the platform for the change for almost everything that is made in the matter world.

That is to say, “that Antimatter is the source, where pmtics are positioned to start the change from stronger pmtics strength level in the plasma components, and transfer more of pmtics strength of Matters to matters”.

It is the pmtics of the Antimatters, which are transferred in most parts‘, from one plasma to another, which, causes energy transfer from one molecule to another or one cell to another.

We have used these Antimatters pmtics strength effects in the recent past in our reactors, the pmtics interactions of Antimatters were utilised to create the Magravs forces needed for the Magravs positioning of these systems in respect to another plasmatic Magravs of another object, this being like the Earth’s Magravs.

Through tests, it has been proven that the presence of the **Antimatters** and its strong plasmatic magnetic fields that it provides for the creation of Magravs in the reactors is the main reason why, we have managed to lift and move large masses through the gravitationally positioned principle.

In Grapos test reactors, which have been developed and tested for lift and motion, the operativeness of pmtics strength of the Antimatters have been proven to be in line with our theories, where if the pmtics of the Matter were used in these reactors, the Magravs created by pmtics of this Matter would not haven been strong enough to cause the motion or lift of the system.

The scientific world is using accelerators (6, 7, and 8) to slam the plasma into a wall of matter, for the plasma to open and to release its Antimatter sub-components.

In present technology, scientists are trying to understand how to gather and harvest the power of this Matter. It is however crucial to understand that, “**one does not burn the Antimatter as fuel,** as it is assumed by the present scientific world in this field and as Fermilab (6, 7) is considering”.

Nevertheless, there is an easier and smoother way to achieve this separation, and that is just to dilute the plasma (Chapter 21), through this method, it will be easy to release the Antimatter from its plasma.

Once the scientific world has a full understanding of how to separate this Antimatter from its plasma, then they will understand that one does not need to burn this Matter as fuel to create motion and energy. But, they simply have to learn **to use the effects of the interactions of the constituent pmtics of this Antimatter** in conjunction with other Antimatters pmtics of other plasmas or Matters pmtics to achieve lift, motion, and for creation of energy, as has been done in the universal order of motion and creation of energy for billions of years.

CHAPTER 11

The Antimatter energy

Antimatters are known to be the sources of vast amount of the **strong pmtics**. The strong fields of the Antimatter are considered to be the major sources of the energy supplied into their respective environment for Matter-to-Matter, and from Matter-to-matter through their pmtics. The ***strong** pmtics*, once released by the Antimatter, and these pmtics in being *in motion* in their environment, they become to be called the **energy of the Antimatter**.

These pmtics by essence of their origin and strength they are, and become, ***strong** energy field sources* and reservoirs of energy in their environment. Where, some of the pmtics within the Antimatter are used to maintain the integrity of the Matter and the plasma itself.

At the same time, some of the pmtics of the Antimatter are equally reduced in strength, and these changes and reductions in their strength of some of the pmtics strength of the Antimatter overall constituent pmtics, these pmtics are allowed then to escape the gravitational fields of their Antimatter pmtics Magravs.

These released pmtics from the Antimatter, become the Antimatter pmtics in motion (energies) due to the strength of their plasmatic magnetic fields' forces. These pmtics can travel *much further* from the Matters and plasma, to transfer their energies to other pmtics in their environment.

These released fields become the Antimatter pmtics in motion (energies), which, due to the strength of their plasmatic magnetic fields' forces, these pmtics can travel *much further* from the Matters and the plasma, for them to transfer their energies to other pmtics in their environment.

Due to the fact that the Antimatter itself is originally made of the stronger pmtics, the Antimatter will let go of pmtics in stronger and wider spectrum of magnetic fields strength. Thus, the Antimatter releases a full spectrum of pmtics strength forces. That is why they are more powerful and covering a wider range spectrum of energy levels as they slow down.

In essence, the **Antimatter** itself starts from a combination of interactions of stronger pmtics bases, and hence has more to lose of its slowed down fields and it has more to attract due to its stronger Magravs too. At the same time, it covers a larger domain as its pmtics can spread further.

In fact, the energies that are released by the Antimatter look to be, and basically are stronger than other energy fields around them, purely due to the fact that they start their life from stronger pmtics strength base in their environments.

CHAPTER 12

The Dark Energy

The principles of creation of the Dark energy as an outcome of normal interaction of pmtics are discussed. The similarities of the Dark energy and force, which are created around Black holes, are in principle of the same origin.

In the dynamic environments of the universe, changes happen due to various conditional factors like motion, compression, turbulence, temperature, and so on, that can individually or collectively lead to concentration or motion of pmtics (energy). These changes in environmental conditions can lead to the manifestation of Matters like the Matter, the Antimatter, the Dark Matter, atoms, stars and galaxies and so on. This is the common fundamental base for the creation of Dark energy and creation of Dark Matters (36) and any dark objects like the Dark Matters, the regular appearance of dark spots on the surface of the Sun and stars (12, 13, 14), the appearance of Dark spots in the galaxies known as Black holes (13), the Dark Rings around Saturn (17) and so on. The creation of these conditions and matters can all be explained through the following interaction of normal dynamic pmtics in the universe.

By pmtics theory, the Dark energy is said to be the ***collection of balanced and equilibrium pmtics strength forces in motion*** in respect to their environments pmtics strength, this being in any given environment, in deep space in the universe or in the centre of an atom and so on.

The universe always possesses equilibrium pmtics strength in its total energy balance, hence the universe possesses more environments which are in pmtics balance strength and hence the existence of vast-spans of the universe which energies are in motion, but due to their balance magnetic fields strength in these regions of the universe, no light through the interaction of various magnetic fields are produced to confirm the motion of the pmtics in that region.

It is important to note, that in the universe only the difference in magnetic fields strength is the cause of creation of visible light from the human observation point of view. Where there is balance and equilibrium in the environment, no lights through interaction of magnetic fields or pmtics are produced. This doesn't mean that there are no magnetic fields or pmtics in motion (energies) and that nothing is present in that region of the universe. In fact the darkness of dynamic environment of the universe confirms the existence of the balanced magnetic fields strength. Beholding the truth, the darkness of the universe confirms the existence and motion of balance and equilibrium of magnetic fields strength, and not the lack of the existence of magnetic fields in motion (energy).

We have defined Energy as: "**pmtics forces in motion**". Therefore the Dark energy has the same definition, where in this case, these are **balanced or near balanced pmtics**

strength forces in respect to their environment pmtics field forces in motion.

The universe always possesses equilibrium pmtics strength in its total energy balance, hence it possesses more environments which are pmtics balance strength and hence the darkness of the universe.

Steps in the creation of Dark Energy

To explain the phenomenon of creation of the Dark energy in more detail, let us consider two equal pmtics environment zone A and B in figure 35. These two fields' zones can be for example pmtics from two galaxies in motion in the universe or two pmtics in the span of a plasma.

Before the interaction of the two field zones, pmtics of each zone are interacting with their respective environment, which creates their own boundaries of their field strengths through their magnetosphere boundaries, in respect to their environmental pmtics (zone A with zone D, and zone B with zone D).

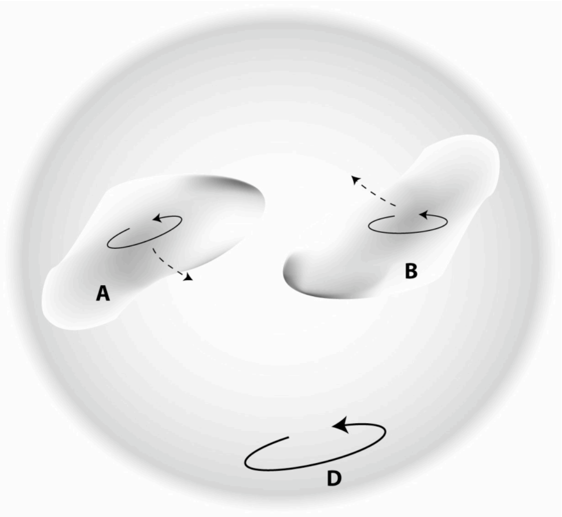

Fig. 35: Two equal pmtics strength entities moving in each other path.

Lets assume; a) The universal magnetic fields are always in rotational motion as of D (Fig. 35), and b) The two pmtics of the A and the B are in motion in directions shown. Through their motion and the environmental forces, the two fields zones come to be partially in contact and interact with each other. (Fig. 36)

As segments of pmtics of the zones A and B collide (Fig. 36), and as their interacting pmtics are similar in strength, then there are no differences in field strength for the two pmtics to create the magnetospheres condition (14) or no residual pmtics or very little pmtics fragments are produced from the interaction of these two pmtics zones. Thus in the interaction of two fields and due to balance in their pmtics strength, this interaction **zone C** of fields does not lead to creation of pmtics fragments in the light magnetic field spectrum leading to the creation of pmtics of light (Chapter 7), to make this region light to show the existence and occurrence of collision between two equal magnetic fields strength zones.

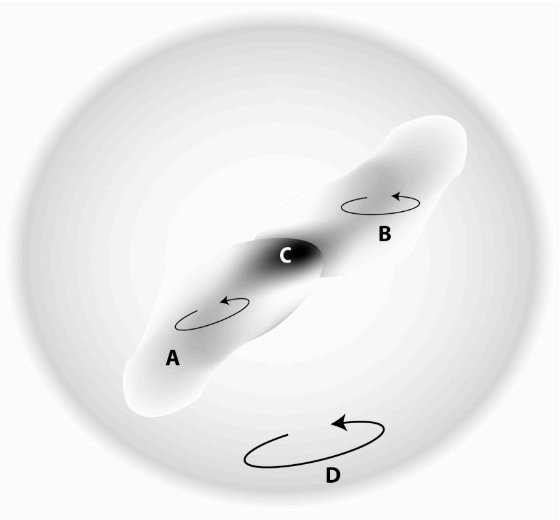

Fig. 36: Two equal pmtics strength entities moving in each other path colliding and creating a temporary Dark energy zone (zone C) due to their interaction.

Thus in respect to the surrounding environment the colliding of two pmtics in motion and the lack of production of pmtics fragments strength in the bandwidth of the light in this section of colliding pmtics, this creates the appearance of being dimmer or darker zone (Fig. 36 zone C) compared to its two original zones magnetospheric condition in respect and in comparison with their surrounding pmtics.

In the region of C of figure 36, due to the balance in pmtics strength and the creation of very little residual pmtics fragments, this area has a darker appearance in comparison to the rest of the region which one can observe with the rest of the two zones of A and B in the overall region of the D environment. Nevertheless, the energy of the pmtics in motion of the two pmtics zone of A and B exist and they can be measured, as the two pmtics of A and B move on, this confirms the existence of energy in the dark zone C, where in fact there are no sources for such an energy field to justify their existence in this darker appearing region.

Thus, in the case of the Dark energies zones in the universe, most of the time, one has to look at the overall picture of the interactions of the package and area, rather than to the local zone where the darkness and dark energies appears to be and where pmtics in motion as energy fields are present.

Further on, as the two fields zones A and B are dynamic entities and they are pushed by the forces in the bigger picture of the zone D, the Dark zone C, figure 36 disappears as two zones of A and B move away from each other.

Thus, the interaction and creation of the zone C, figure 36 is a **temporary phase** as the two field zones separate and carry on in their own path and the Dark energy zone C in figure 36 disappears, as there never was a Dark energy zone in that part of the zone D as whole (Fig. 37 fields A and B).

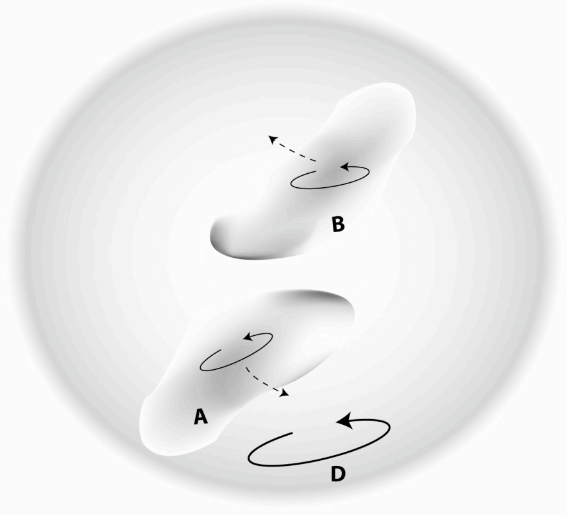

Fig. 37: Two equal pmtics strength entities moving away from each other after their collision.

Therefore, Dark energy zones in the universe are not necessarily a place where there are void of Matters or magnetic fields, but the darkness itself confirms the existence of the balanced pmtics strength in motion in respect to the majority of pmtics strength forces in that given environment, which pmtics in these areas are in balance and in equilibrium in strength with each other in a given point and a given window of motion of fields called time in the span of the universe.

At the same time, these sudden Dark energy regions in space are temporary conditions in a given specific time and specific position from a specific point of observation.

On the other hand if one is inside such a zone not necessarily means that one is in the Dark energy zone. In some cases of dynamic solid gravitational fields of matters in the centre of these energy zones, due to the interaction of the Magravs of the Matters and in level of Dark intensity translucent, inside these Dark energy zones one can have bright zones as being near a star zone.

It is obvious that most of the universe appears as dark, even though pmtics always exist in these areas of the universe and are essential components of its structure. This by itself means that the universe is in pmtics strength balance and equilibrium within itself.

The Dark Matters do not all possess the same pmtics strength and therefore the Dark energies are not all of the same magnetic or Magravs strength, in respect to other pmtics strength environment.

By the same measure, the Dark Energies in the universe do not all possess the same pmtics strength, as their strength is pmtics strength dependent in respect to the environment pmtics strength of their appearance.

In the case of the Dark Matters as sources of the Dark energies, these Dark energies are produced through the initial pmtics interactions in the centre of the Matter, and hence the energy they produce as their overall outward Magnetic fields and inwards gravitational fields, are always in balance in respect to other pmtics in their given environment. Where this balance of pmtics strength gives the Matter the appearance of **darkness**, or darker looking in respect to its overall given environment.

Some of the Dark energies have a Magravs source in their centre in the universe. Thus, these types of Dark energies are part of a Dark Matter source, which are part of the overall magnetospheric environment of the Dark Matter pmtics. These sources of the Dark energies are the Magravs of a matter which in respect to its environment from a given position in space, and due to the equilibrium of the pmtics of the Matter and the environment, cause the Dark energy zone, from that specific point of observation, as that given zone of pmtics interaction appear to be darker than the rest of the environment surrounding it.

CHAPTER 13

The Dark Matter

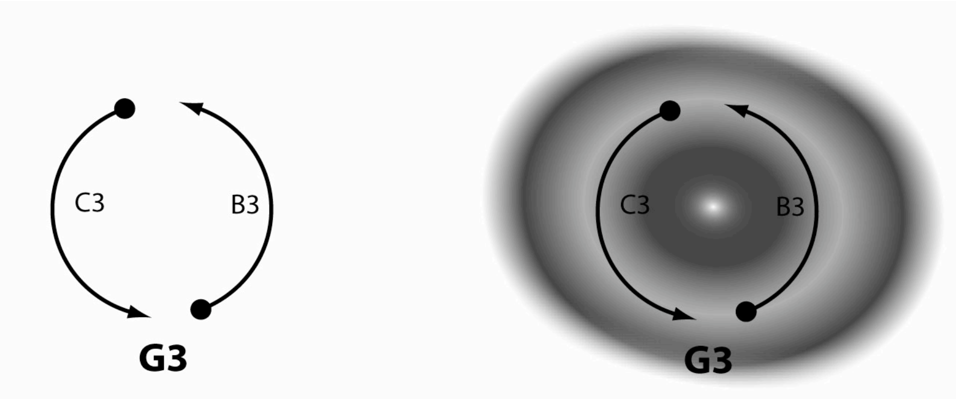

According to the ***Keshe pmtics Theory of Dark Matters*** "*for an entity to be considered as a Matter, the entity has to possess both Magravs pmtics forces, so that not only the Matter can maintain its Magnetic and gravitational fields, but at the same time the Matter can maintain a balanced pmtics magnetosphere magnetic fields forces strength zone and a magnetic fields mass, for the Matter hold on to its constituent pmtics for it to exist, with the difference with the Dark Matter that the magnetosphere created through the* ***overall*** *interaction of its Magnetic and gravitational fields forces strength of the Matter and the environment pmtics force strength that these two forces are in equilibrium in strength, where the interaction of the pmtics of Magravs of the Dark Matters' magnetosphere and the environment pmtics field force in interacting with each other, they do not create residual fragments pmtics that is in the light magnetic fields strength spectrum level, that makes this Matter visible or detectable, hence the term Dark Matter (Fig. 38 G3)* (36).

At the same time the Magravs forces created by the same Dark Matter appears to create a magnetospheric fields forces' zone around the central Matter, which gives an impression of an energy source that comes from a dark space in a given environment without a visible material source and hence the use of the term **Dark Energy** (Fig. 38 C)."

The Matter is said "to be dark, as its effects can be seen or felt, but, it, is not made observable to the line of vision or is undetectable through detection systems, which are looking for signs of light as means of confirmation of the existence of the Matter"(36).

The Dark Matter can be defined in another way, that is to say:

"The Dark Matter (Fig. 38 Matter G3 field C) is the Matter that, by the interaction of its internal initial pmtics strength, creates the Magravs effects and its external plasmatic Magravs as the magnetospheric field effect for the Matter to exist. Nevertheless, it is due to the overall balanced Magravs strength that are created by the Matter as the magnetosphere of the Matter, being in balanced magnetic fields strength with respect to its given environment pmtics strength, so that the interaction of the two magnetic fields does not creates pmtics fragments in the light spectrum, where the Matter cannot create detectable magnetospheric light.”

That is to say, the interaction between two plasmatic magnetic fields strength of the Dark Matter and the second pmtics strength of environment does not create a **visible or detectable magnetosphere** effect, which would be needed to confirm the presence of the Matter in that given environment and in a given position, at a given point in time.

The Dark Matters effect of Magravs energies can confirm its existence, as the Matter that *appears* as Dark or translucent Darker Matter, in respect to its given environment, to the line vision appearance of the Matter, in a given position and time.

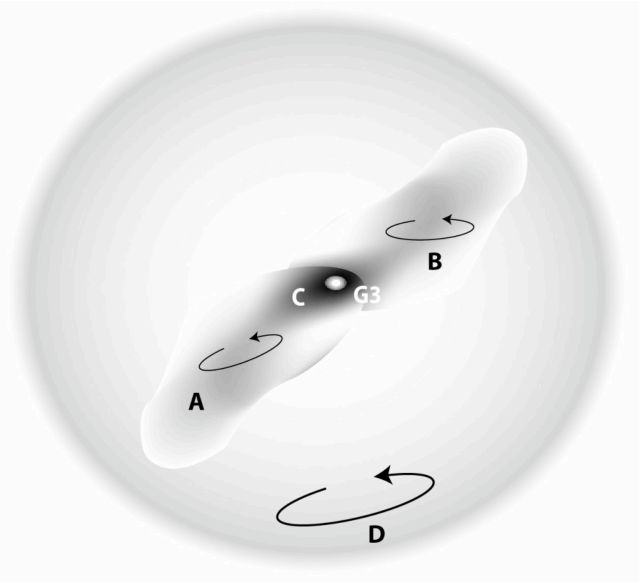

Fig. 38: The Dark Matter (G3), the Dark energy (C)

The Dark Matters can be, or, are made of stronger or weaker plasmatic magnetic fields than the Matter, and the Antimatters components of the plasma.

The Dark Matters are Matters which appear as areas in the universe in possession of Magravs forces but no or a very weak, detectable external magnetosphere in respect to their environment at given point and **a given window of transition of time**, where the confirmation of the existence of the Dark Matter is environmental pmtics strength and position dependent.

If one could create the fundamental principles of the Dark Matter, by the use of Grapos, with the production of balanced pmtics strength around the system, then one can travel the spans of the universe with pmtics Magnetic field protection and more or less without any friction, or resistance in the direction of the motion of the system, this allows the system to achieve speeds far beyond the present knowledge of Man in the space and aircraft industries (36).

The same effects of the creation of balanced pmtics strength in respect to the environment can be used to make systems, that without moving; such system becomes invisible to the line of vision at any point in atmospheric condition or in deep space.

The larger entities like of the Dark Matters in scale of magnitude are the black holes in the galaxies, where a large gravitational fields zone is observed and no physical tangible source for the large and strong pulling forces in that region of the galaxy can be seen. These areas, which are usually and mainly in the centre of galaxies or on the edges of turbulent boundaries of galaxies, as there are more Matters, matters, magnetic fields and plasmatic magnetic fields on the move. Where the total balance of these fields, in respect to their given environment, gives these regions an impression to have the appearance of the **darkness**.

These dark regions with strong gravitational pulls are overall temporary events in the bigger cycle of life of galaxies bigger picture of motion of pmtics and Matters (12, 13).

Sudden appearance of Dark Matter as Matter due to changes in fields forces in the environment can be analogised like having a wet cloth that is dry looking. However, by twisting and squeezing the cloth, out of nowhere, drops of water start dripping out of the cloth. The appearance of the water drop does not mean that water was not in the cloth, but in fact it means that the condition were not there to force the water to manifest itself within the cloth.

The Dark Matter can become visible as Matters around their environment pmtics strength changes. On the other hand, the Dark Matter can become visible Matter when there is a change in the internal pmtics strengths of the Dark Matter, which the change in magnetospheric strength of the Dark Matter, changes the magnetosphere pmtics strength in respect to its environment pmtics strength and the Matter becomes visible. That is why the Dark Matters are environmental pmtics strength dependent. The appearance of Dark Matters, due to changes of their pmtics strength or environment pmtics strength changes have been seen in the universe by scientists and they call these, the Virtual matters or the Virtual particles (Chapter 14).

Dark spots on the Sun and Black Holes

The Dark spots on the surface of the Sun; These effects are usually seen on the dynamic surface of stars, where large numbers of similar pmtics strength materials in motion interact with each other, and hence there are more possibilities of having balanced pmtics strength regions, as most of the matters are of the same structure of hydrogen plasmas.

The main reason for the increase in the levels of the dark areas and dark spots on the surface of the Sun, about every eleven years, is due to change of polarities of the Magravs created by magnetic fields in the centre of the Sun. This increase is purely and only due to one factor, this being that the creation of dark spots on the Sun is primarily due to the interaction of pmtics of the matters on the surface of the star and the pmtics of the Magravs from the centre of the star. As at time of polarity change of the star, there are more chances of the interaction between two pmtics, which are made by the components of the same matters (Hydrogen). Thus, there are more chances of reaching Magnetic strength equilibrium conditions between the surface pmtics strength of the matters on the surface and the internal Magravs' fields strength created in the centre of the star. Where at these points of balance zones of pmtics strength, no or very little pmtics fragments in the spectrum of light are created, hence the reason why appearance of the increase in the number of dark spots on the surface of the Sun during the polarity changes of the star (12,13).

Thus, the dark spots of the Sun are natural outcomes of the balance interaction of the pmtics strength of the overall Magravs created by the centre of the Sun in respect to the pmtics strength of the plasmas of the materials on the surface of the Sun. Thus, as there are more balance pmtics areas produced, a higher number of darker zones can be observed on the surface of the Sun.

The darkness of Black holes in galaxies, and the Dark energy fields they produce in their environment, are due to the same process (12, 13). Similarly, all Dark energies are not necessarily connected to the central Magravs of the Dark Matters.

The Dark Matter and the Dark Energies appear and behave as dark regions, in respect to a given magnetic fields strength environment, at a specific point in time and space, and specific conditions from the point of observation. By the same principle, the Dark Matter and the Dark energy will be a visible or tangible Matters or matters in a different pmtics strength environment and even from another point of observation.

Therefore, state of the Matters, this being the Matter, the Antimatter, the Dark Matter appearance and detection of their energies are totally pmtics environmental respective dependent.

That is to say, "what is **the *Matter and the energy*** *in one given plasmatic magnetic fields' strength environment* can become *the **Dark Matter** or the **Dark Energy** in another pmtics strength environment* to the human point of observation".

In the universe, the Matters and the Matter energies or the Dark Matters, the Dark energies, the Antimatter and the Antimatter energies **change from one state of Matter to another**, or one **state of Matter energy to another**, purely due to *the respective Magnetic fields or gravitational field* strength as of their environments pmtics strength as these Matters and energies pass *through them*.

In the micro and macro universe, suddenly Matters and energies ***appear*** **out of nowhere** and **Matters and energies suddenly *disappear* into nothing**. These appearances are all pmtics strength respective environment dependent as the Matter or energies pass through a given field at a given time in their immediate environment.

The universe is considered by us to be covered with the Dark Matters and the Dark Energies. In general it can be said, where there are no faster or slower pmtics, or differences in pmtics strength, interacting or for other pmtics of the same or different strength to collide with, through these balanced pmtics strength regions, there are no unbalance fields to create friction or to have an interface or to have residual pmtics between the environmental and the moving pmtics, these regions in the universe give the appearance of lack of *Matter,* or the Dark Matter or the existence of Dark energy in respect to the given point and position in the universe. This means that the dark regions of the universe are balanced in pmtics strength, and *not void of pmtics*. Hence as the universe is magnetically strength balanced, therefore the universe, in general, appears dark and empty, even though the universe has no regions without magnetic fields.

These types of darkness can be see-through and can be transparent through their pmtics energies; I call these areas the "***translucent Dark energy zones***".

This translucence state of position in a given environment is a state where, there is total darkness but at the same time, there are pmtics strength field interactions that allow enough light, in respect to vision and detection for one to confirm the lack of light.

This being the reason why, one can see or detect, other Matters or matters through this universal Translucent Darkness light.

This light in the span of the universe is due to the interaction of very weak pmtics in motion (energies) in the constituent universal pmtics soup. In fact, these dark regions of their environment are where the pmtics are in balance within these regions.

As was discusses in the previous chapter, to the human eye protein pmtics balance set, these residual fields of the two pmtics interactions, due to their pmtics strength balance, appears as the Dark Matter and is like Dark energy zones in space. If humans would possess another chemical pmtics protein chain, what are the Dark Matters and the Dark energies by present means of detection, by viewing through different tools of vision, then these areas will be as visible light regions of the universe, and may be in different colour, for example in the Antimatter colours spectrum of the universe, or even in the Dark Matter colours' spectrum of the universe. Through the use of such tools, the importance of the use and function of the Dark energy (Fig. 36) and the Dark Matter (Fig. 38) will be soon grasped by men of vision, to see the true darkness. Where in the colour spectrum of the Antimatter the colours are magnetic fields strength dependent and these colours are not noticed by the human eye matter-based protein pmtics. The same applies to the Dark Matter colour spectrum.

In the universal order of containment for Matters and energies, it is not far from the truth to say that the state of Dark Matter is the same as gasses in matter state, that the Dark Matters and Dark energies do not have self-containment.

That is to say, "As with gasses in a state of matter, Dark Matter fills all the environment and space that is given to it, as it creates no effective magnetospheric interface boundary resistance in respect to its environments' pmtics strength forces.

The boundaries of the Dark Matters and Dark energies are dictated by the boundaries occupied by stronger or weaker pmtics strength regions around them in their given points in the universe."

The forces of Dark energies can be confirmed only through their interaction with other plasmatic magnetic fields in motion, in respect to the field of the Dark energies motion themselves.

That is to say that the dark lights generated by the Dark Matters and Dark Energies cover the whole of their given spaces, this being the space in-between different Matters in the atoms structure or solar system or even galaxies, I call this ***principle of creation of dark light*** strength difference of pmtics magnitude needed for creation of light, the **principle of pmtics magnitude**.

It is very correct to assume that one atom of Dark Matter and its Dark energies can be and/or is enough to span its balanced pmtics to cover the whole of a solar system, or its bigger counter part galaxies or smaller spaces as the space in the pmtics Matters of one tangible atom.

CHAPTER 14

Virtual Particles (Virtual Matters)

The existence of virtual particles, *which appear out of nowhere*, and have the characteristics of real particles in essence, has been described in the Dark Matter section of this book. The method by which these particles appear within their environment is exactly how the Dark Matters do exist and appear in their environment, after their overall Magravs, magnetosphere strength changes' or due to changes in the pmtics strength of the environment.

In fact the virtual particles are individual atoms' or plasmas, or any Matters, of the Dark Matters behaviour' (Chapter 13, and reference 12), which when they come into new pmtics strength environment *which is **different** than their previous balance condition* pmtics strength environment with their Magnetic field and gravitational fields, then these particles generate a new magnetosphere in respect to their new condition pmtics strength environment, thus they create light due to their interaction of their plasmas (Chapter 7) and so they appear *as **new matters or virtual Matter***.

There is nothing new about virtual particles; these are like loose plasmas or a single atom in the soup of the universal Magnetic and gravitational fields.

Thus, in simple way according to ***Keshe Theory of creation of Matters***, virtual particles are Matters and behave like Dark Matters. They are created from the interaction of at least two pmtics, which leads to the creation of Magravs of the Matter, therefore what scientists call virtual particles, these are real Matters, which due to the change in their environmental magnetic fields strength condition, they suddenly ***appears as matter(s)*** for *a very short time*, until they lose their pmtics strength or their environment pmtics strength changes or as they move out of the environmental pmtics strength level or they seem to disappear.

Virtual particles possess a magnetosphere and the light they create in their new environment is due to the interaction of their pmtics of their magnetosphere with the pmtics of this new environment, which makes them visible in that given point and given time.

Note: When the (virtual) particle(s), Matter(s) or matter(s) is moving towards the observer due to it's ***compacting of pmtics***, this creates the blue shift colour of magnetic wavelength spectrum (Chapter 7). If the virtual particles are moving away from the point of observer, due to ***opening of the pmtics*** of the environment, they create light in red shift colour of magnetic wavelength spectrum. (This is described in the paper: Creation of Magnetic field).

The virtual particles are actually individual nano-atom or plasmas or sub-Matters of plasmas.

The Casimir effect

In present physics, virtual particles are related to Casimir effects.

First of all, the Casimir effect and the photon energy appearance in its setup of its confirmation is always considered to be in a **vacuum condition** between two conductive plates where, photons in between two plates are observed and electrical charges are measured.

In fact, it is not important what vacuum condition level can be achieved in test laboratory conditions, nevertheless still a number of pmtics, Matters, plasmas (electrons and protons) and atoms could be left in the space between the two plates, which are under observation.

As matters are made of plasmas of interacting magnetic fields of Matters, when the space in between the plates is freed and void of most of its atoms and Matters, at this point the leftover matters or Matters between the two plates are more free to move in the newly created condition environment. Hence it is natural that these residual atoms, electron plasmas, proton, can move really in empty space and can collide with the walls of the matter of the two plates and due to their interaction fragments of pmtics strength are released which these are in the range of the pmtics strength of photons.

So vacuuming the gap between the plates creates a new environmental plasmatic magnetic fields force condition that allows a more free motion of leftover Matters and atoms in between the two plates, where in fact the boundary of the magnetosphere of one atom can occupy or expand the full span of the gap between the two plates.

Casimir effect measurement is a good tool to measure how many atoms and/or what Matters are left in the gap between the two plates at a given vacuum level.

We have done experiments in plasma dilution reactors, which we have developed and have seen the Casimir effects. In these tests, we have seen some serious voltages and currents, but never photons. Which indicates that these voltages can only be from Antimatter source in-between plates, as matter cannot produce these levels of voltage and amperages.

In a number of tests we have created 30 mm gap between plates and the results have been the clear indication that Casimir effect testing is good to find nearly the number of Matters in between the gaps, and types of Matter which are left in between the gaps. As **Matter and matter** create *different charge levels* than the **Antimatter** components of the plasma.

We have created environments of 10 (-9, -10, -11, and -12) bar in our tests, and in tests over 10 (-8) bars there was not much difference in voltage and current measurements from both plates' terminals. This indicating that it does not make much difference how far one can vacuum the gap in between the plates, there are still some components of Matters or matters will be present in the gap, which can move with more freedom than before. Where this small number of matters or Matters in-between the plates can transfer energy through the free motion of their plasmas in collision with plasmas' of matters of the wall and lead to the release of pmtics fragments in the strength of the light magnetic field spectrum or a photon.

Van der Waals forces

In respect to the Van der Waals effect, this **push and pull effect** is now very clear and easy to explain, this is to say as plasmas contains all Matters (Matter, Antimatter and Dark Matter) and these Matters are all made from the interaction of plasmatic magnetic fields, which leads to the creation of their Magravs, so that each Matter is *like a* ***independent entity*** within the structure of the plasma, at the same time plasma possesses its own Magnetic and gravitational field forces.

By the same measure, the atom, as an independent system, possesses its own Magravs in respect to other atoms in molecules. The same, the molecules have their own Magravs in respect to each other.

Therefore, what has been considered as Van der Waals forces of push and pull of molecules, these are simply Magnetic and Gravitational field forces or Magravs forces of one molecule in respect to any other molecule.

The Van der Waals force is in fact the pull (gravitational effect) and the push (Magnetic effect) of the individual molecules in respect to any entity (atom, molecule) around them.

CHAPTER 15

Interaction between Matters

Different Matters in their environment interact and create different conditions, and state of Matters are similar to the state of matter, as one can mix liquid with solid and gas with liquid in state of matter, the same can be done with different states of Matters, like allowing the pmtics of the Matter to interact with the Antimatter and further the Antimatter with the Dark Matter. As in the case of Matters, they are all pmtics strength dependent, where these kinds of mixings of states of Matters allow more profound effects to be developed. This kind of mixing between the state of Matters will open new world of science to mans' encyclopaedia of knowledge ***Matteorology*** *(knowledge of Matters).*

The process by which, Matters are created and operate in respect to each other within the construction of the plasma has been explained in previous chapters. These Matters individually possess their own properties and create their own effects in respect to their environment, both individually and in conjunction with each other.

What is important to understand is: *What happens when these Matters interact or come to collide with each other within the plasma or their environment?*

What is the outcome of the amalgamation of the plasmatic magnetic fields strength of different Matters with each other?

Let us consider some of the variations between interactions of different Matters.

Interaction of Matter with Antimatter

The Matter is composed of ***weaker*** Magravs strength than the **Antimatter**. In this case, the Matter always gets attracted towards the Antimatter.

In the case of the plasmatic magnetic fields of the Matter ***colliding*** with the plasmatic magnetic fields of the Antimatter, the Matters components will **get *absorbed* into** the larger and stronger pmtics strength, by order of magnitude, of Antimatter.

What can be observed from this impact is the release of some plasmatic magnetic fields residual as pmtics, like release of some lights or rays or magnetic residual. The content of the dynamic plasmatic magnetic fields of the Matter or its energy is *absorbed into the environment of the Antimatter* and gets mixed in within its structural magnetic fields strength. At this point there will be no more signs of the Matters past existence.

Where, most of the rays released in this interaction are mostly due to the coming close of the two Magravs of the two Matters, rather than the splashing of the pmtics from the surface of the Antimatter.

This is like when a drop of water falls into a cup of tea, in this case, the drop of water becomes part of the larger teas' liquid.

Majority of scientists working in this field, are lead to believe, as the Matter and the Antimatter collide, due to their charge balance after the impact, the content of the two Matters ***disappear into nothing***, except some residuals released as energy, gamma rays or photons, then, in their opinion, there is nothingness left as these two Matters' charges balance each other out.

Such a belief and concept is against any natural laws of existence. That is to say, there cannot be a position of total annihilation of both Matters, where there can be nothing left at all after the two actual Matters in possession of Magravs collide with each other. This nothingness of the two Matters after the collision of the Matter and the Antimatter is a rash assumption and incorrect.

In fact, the Matter and so-called the Antimatter by their pmtics strength order of magnitude in the matter world, is like the Earth slamming into the Sun, and then saying that after this collision there is nothing left of the Sun and the Earth, except some outwards splashing fragments of plasmas from the surface of the Sun. This is illogic attempt at a scientific explanation for such a real physical event.

The Matters of the Earth would cause a small splash on the Suns' surface. This splash then releases some plasmas or gamma rays or photons from the surface of the Sun. Then the plasmatic magnetic fields components of the Earth become part of the plasmatic magnetic fields of the plasmas of the Sun. This is exactly the same incident as when pmtics of the Matter collides with pmtics of the Antimatter.

In view of this principle in physics of the conservation for energy, ***nothing in the universe is lost or can disappear***. They just change their magnetic fields strength and their compactness from one form and level to another.

Thus, collision of the Matter and the Antimatter create nothing but the Antimatter, with more content of different plasmatic magnetic field's strength of the Matter components inside it. Therefore the outcome of this collision is mainly the Antimatter balance remaining.

Interaction of Matter with Dark Matter

This follows the same principle as the previous states of the Matter and the Antimatter. The content of the Matter then becomes the constituent Matter of the gravitational centre of the Dark Matter. With the difference, that in this case the Matter will not have to overcome any magnetosphere of the Dark Matter.

This interaction will be the same as seen in the universe matter dimension, where for example Stars and cosmic matters disappear into a Black hole in the universe.

In some instances, this process can disturb the overall Magravs' strength balance of the Dark Matter in respect to its surrounding plasmatic magnetic fields strength. This change can induce the production of slightly stronger or weaker magnetic fields strength magnetospheric conditions of the Dark Matter in respect to its surrounding field's environment.

In these rare cases, due to the change and unbalance of the magnetic fields strength between the new content and new Magravs strength of the Dark Matter, this new stronger or weaker plasmatic Magravs strength, allows the creation of a new magnetosphere field strength of the Dark Matter in respect to its environments pmtics strength to take hold. Then, what was the Dark Matter in a given environment, can in the same given environment, with a different and new plasmatic magnetic fields strength, now create and possess a new magnetospheric field strength zone in respect to the same environment's pmtics strength and then the Dark Matter can become like the Matter that is more visible.

This incident has been recorded by cosmologists a number of times. Where, *out of nowhere*, Matters appear in the Dark spaces of the universe.

In fact, even changes in the tightness of plasmatic magnetic fields in the span of universe, can unbalance the Dark Matter environment pmtics strength and make the Matter visible.

This magnetic unbalance of the environment causes different balance condition around the Dark Matter pmtics, and then out of nowhere stars appears in the Dark spans of universe.

The interaction of Antimatter and Dark Matter

In the case of the interaction of the Dark Matter and the Antimatter, there are clearly several scenarios, which are possible.

1) Scenario 1

This case is the most probable and is considered to be the norm. This interaction is a collision of the giants of the Matters of the plasmas‘.

Assuming that this scenario of interaction takes place within the environment of an initial fundamental plasma.

Considering the construction of pmtics strength of the Dark Matter and the Antimatter, when, these two Matters of the plasma collide, the outcome is normally production and release of new and balanced plasmatic magnetic fields environments. In this case, it is considered that each of the Matter's components of the plasmas keeps their pmtics strength plasma status.

Due to the enormity of the quantity of the amounts of the two plasmatic magnetic fields of the ***Antimatter and the Dark Matter amalgamation*** (denoted as **A-D Matter**), new supersized, unstable giant plasma is initially created. This new plasma needs to reach stability over the entity of the plasma as a whole for it to survive. Thus, the new super giant plasmatic magnetic field has to find a new Magnetic and gravitational balance and equilibrium within itself.

For the new giant plasma to reach the new equilibrium condition, the new plasma has to, and always does, release its excess plasmatic magnetic fields into its surrounding environment. The new released plasmatic magnetic fields have to be in balance by themselves too, for them to exist and stay within the gravitational fields of the A-D Matter.

This new release of excess plasmatic magnetic fields near to A-D will contain some of the Antimatter, some of the Dark Matter and some other newly balanced interacting and attracting weaker plasmatic magnetic fields of different strength in the Matter pmtics strength level. These *weaker fields* are mainly from the leftover weaker fragments of the collision of the two giant Matters. These new released small bundles of weaker plasmatic magnetic fields strength themselves due to their proximity and their source of origin, these usually bunch-up and become a new *miniature plasma version of the original initial fundamental plasma within the A-D Matter.*

At the same time, this newly created mini-plasma still has to keep it's own balance and the balance of the whole of the original Matters pmtics and stays' within the Magravs hold of the giant A-D Matter mixture. This mini-plasma creates its own Magravs and then this mini-plasma has to have a Magravs positioning in respect to its original initial fundamental plasma. Thus the new mini-plasma is forced to take a new position and new gravitational positioning in respect to its larger neighbour of A-D and it is at this point that it is forced out of the A-D Matter plasma environment, but still in A-D Matter Magravs catchments' area, therefore this mini-plasma attempts to stay in the orbit of its mother seed, and becomes the **initial fundamental satellite plasmas'** of the original A-D Matter.

Further on with the same process, the A-D Matter mixture has constituents of two unbalanced and un-matching plasmatic magnetic fields strengths of the Dark Matter and the Antimatter within it, at some point, the two Matters plasmatic magnetic strength fields separate and the balance returns' to a new central plasma. By this time the central plasma due to this interaction, finds itself with a new rotating satelliteing smaller plasma.

It should be noted that pmtics of different strength ***never mix*** with each other, however they can be ***influenced*** by each other.

The collision of the Dark Matter and the Antimatter, and creation of new satelliteing mini-plasmatic Magravs rotating around the central plasma is the ***creation of the electron*** from the interaction and collapse of theses two Matters of the initial fundamental plasma. The interaction of the Antimatter and the Dark Matter of the Initial

Fundamental plasma leading, to the creation of a new central plasma and its satellite mini-plasma, lead to creation of **Initial Fundamental Atom** (hydrogen atom). This is what I call the initial fundamental decay of the initial fundamental plasma, or decay of the neutron.

The collision of the Dark Matter and the Antimatter is usually due to unbalance or some deviation in the plasmatic magnetic fields' strength of one Matter or other within the confine of structure of the plasma itself.

2) Scenario 2

This scenario happens usually when the F1 collapses altogether, sending the two giants Matters of the plasma, the Antimatter and the Dark Matter, to collide head on.

In this case, there are two possibilities, which remains open to the new A-D Matter plasmatic magnetic fields.

A) The first possibility is that the plasmatic magnetic fields of the Dark Matter superimposes the Antimatter plasmatic magnetic fields.

By this process, the outcome will be a giant size plasmatic region of the Dark Matter and energy in the plasma.

This new giant Matter, due to the interaction and amalgamation of the two sets of plasmatic magnetic fields strength and their Magravs of its two separate Matters will have large-scale gravitational pull system, with no clear magnetospheric zone boundary in respect to interaction with its outside plasmatic magnetic fields strength environment as before.

This will be a giant gravitational region that will have massive pulling effects, but its core cannot be seen due to the lack of interaction of its magnetospheric condition with its environment's pmtics. Therefore, this region even possessing enormous amount of gravitational field pull has no balance interacting Magnetic field strength or a balance Magravs strength in respect to its surrounding pmtics strength. This Matter seems as if it does not exist, or appears to be Darker than it's surrounding. Around this area, enormous gravitational pulls‘ can be detected, but the cause of all the pulling and where the gravitational pulls is coming from are not clear and obvious. Where, this darkness is purely due to the balance magnetic fields strength of the magnetosphere of this Dark Matter and being too similar in strength in respect to its surrounding environment pmtics strength to be able to generate the magnetospheric light.

Thus, this area is and looks like as a giant eating machine, which has a lot of energy and gravitational pull, but no visible source to show for its Dark energy as was explained before (36) and the Dark Matter. This condition is usually referred to as the ***Black Hole effects*** (13).

In fact, the only way that one can say that the Antimatter is cascading in this case by the

Dark Matter, is by observation of the appearances of the interaction of the Matters Magravs patterns of the magnetosphere of the whole structure from the outside of the structure.

In these cases of cascading, where the two pmtics strength at the point of separation between the Antimatter Magravs strength and the Dark Matter Magravs strength, internal layers cannot mix due to their different magnetic fields strength, this interaction between two different pmtics field strength at this magnetic interface, creates an ***internal pmtics-interface zone***. Where, this pmtics strength interface zone creates a different zone gravitational and Magnetic field strength on the outside of the magnetosphere surrounding environment of the cascading Dark Matter, and the position of this zone on the outer regions of the Magnetosphere of the Dark Matters are in position and in-line with its physical position of the pmtics interface in the infrastructure in the centre of the cascaded Material.

This internal interface plasmatic magnetic fields zone can create ring-like zones around the outer boundaries of the Dark Matter magnetosphere, where these rings are like different colour or light intensity rings around the whole of the outer boundary of the Dark Matter in respect to their environment pmtics. Where, the thickness and the width of such a ring is equal to the thickness of the embedded Antimatter component.

The physical observation of these effects of the pmtics-interface strength can be seen in the gaseous planets which have initial solid core centre, and where different pmtics strength zones interfaces of different Matter layers are created internally due to their internal pmtics interfaces, as each interface is seen as one ring in the extended magnetosphere of the giant gaseous planets.

These type of magnetospheric rings are usually seen around the gaseous planets with a embedded and initiated solid matter centre cores, and as a number of different layers of the same or different gases are superimposed on top of each other over the life time of the planet. Where the solid matter due to temperatures of its blanketing of gases has become to be in the fluid state-of-matter and creates its own plasmatic Magravs forces. Which the Magravs of matters core interact with the Magravs of the gases in different layers, and/or gas-to-gas of different layers. Where, at the interface of each layer of matters, different Magravs strength are generated and their regions and position of pmtics Magravs strength balance are observed, as one ring related to that specific interface, around the outside of these planets magnetosphere environment in their solar systems. This is exactly how the rings of the Saturn and the Jupiter are created (17, 18). In fact if one could count all the numbers of the dark rings of the Saturn, one should be able to say how many layers of gasses the planets has in its constituent inner structure. This is like counting the rings across cutting of the trunk of a tree to establish its age. How these rings have come to be made and how layering of different gas layers have been achieved by the planet Saturn has been explained in full in the papers under reference 17 and 18.

These multi-layer gas rings actually can bee seen around some stars too, but these are hidden within the intense light from the surface of the star.

The application of multi-solid core and rings in Magravs positioning reactor

When this principle is used in a Grapos reactor, this gives the system and craft a major impenetrable multilayer magnetic shielding ring for protection from cosmic dusts or in defence technology for maximum shielding from incoming asteroids on the path of high-speed crafts (17, 18, 29, 33). Or these types of magnetic multiple ringing can be used around a craft for submerging in deep ***liquid environments*** (29, 33). Where, the craft using these systems become liquid-sealed and Magnetic field and gravitational depth positioning, and temperature independent.

B) The second possibility, with this interaction is that the Dark Matter's plasmatic magnetic fields moves to the centre of the Antimatter's plasmatic magnetic fields region.

In this case, the Dark Matter plasmatic magnetic fields are superimposed and they are cascaded by the Antimatter plasmatic magnetic fields.

Considering that this scenario happens in the environment of the initial fundamental plasma and between its Dark Matter and the Antimatter components. In this case, again, giant plasma of double layers of the Antimatter and the Dark Matter is created and the new plasma will suck-in the Matter components and the electron or another pmtics around it into itself as before.

Nevertheless, this interaction gives the impression of a sudden expansion of the Antimatter as a bright Matter. That is to say that the magnetospheric zone of the Antimatter becomes the total magnetospheric zone of the central plasma. Thus, this new combination of interactions will have a new plasmatic Magravs strength of the three Matters. These being the contents of plasmatic magnetic fields' strength of the Matter, then the Dark Matter and then the plasmatic magnetic fields of the Antimatter.

All these Matters again as previous scenario being cascaded this time by the pmtics of the Antimatter. Where, the weakest pmtics strength of the Matter will be in the centre, superimposed by the plasmatic magnetic fields of the Dark Matter. Then, these two Matters are superimposed by the pmtics strength of the Antimatter on the outside.

The new giant combination of three Matters at this stage is a super plasma size, and this is a temporary condition and is called the Nova state of the three layers Antimatter cascading of the plasma.

Where, the Antimatter's strong plasmatic magnetic fields strength are on the *outside* and create enormous amount of light compared to when there was a balance state of the initial fundamental plasma. This sudden expansion of the Antimatter due to ballooning of the other Matters inside the cascade gives the impression of a sudden expansion of the bright plasma, or what is called a ***Nova condition***.

Sometimes in stars, due to the internal interacting dynamic plasmatic magnetic fields of the plasmas, and through natural process of interaction like fusion or fission or nuclear decay and generation of weaker fields in the centre of these objects, Dark Matters and

Matters pmtics strength conditions are generated and held in the centre of the star.

This generation of new Matters *internally* will cause the expansion and ballooning of the volume of the star initially for a short time. Further on, at this point, as the content of the new-cascaded plasma of Matters can push its boundary to limits beyond the internal central gravitational strength pulls of its three different Matters, then the star explodes in its environment.

This explosion leads to the total release of the internal plasmatic magnetic fields of all its constituent Matters and pmtics inside the cascade. Therefore, these Matters and plasmatic magnetic fields can again start new interacting conditions and environments. Henceforth, the whole cycle of the creation of fundamental particles, then initial fundamental Matters, plasmas, atoms and matters starts again.

This is in fact the normal cycle of the process of interaction and attraction of plasmatic magnetic fields strength in their confined of their environment. There is nothing more and nothing less to this natural cycle of regurgitation of plasmatic magnetic fields of different strength and conditions that these create for themselves and the effects they create to secure their existence.

That is to say, in the case of the plasma, this becomes the release of some particles, which gets absorbed by the other atomic plastic structures around the plasma.

In the case of a star, this becomes an exploding star, which its constituent plasma after disintegration becomes part of the soup of the galaxy and the content of the star gets absorbed in, for regeneration of new stars and so on. In the case of galaxies, this is the same; further on the same happens to universes and so on.

There is the possibility of another scenario for the same new cascaded materials, and that is:

Once the Nova is created and it is not allowed by the external field force in its environment to expand its boundaries beyond its combined internal Magravs' strength, it is at this point that, the *internal* gravitational fields of its three Matters interactions starts to lock into each other and create one overall Magravs force in respect to their outside environment.

This new overall plasmatic Magravs force, which is created due to the interaction of plasmatic magnetic fields of the three layering and three different pmtics strength of the three Matters (the Matter, the Dark Matter and the Antimatter) in conjunction with each other [This is a similar principle to the systems of three internal layers for creation of gravitational field forces of the planet Earth as explained in Chapter 1 (12, 18)], once this new gravitational fields forces takes place and takes hold, and becomes operational, then purely due to its interacting layering gravitational strength magnitude, and the span of its coverage the Matters of the cascade will start pulling all three plasmatic magnetic fields components of three Matters tightly to itself. This causing the retraction of the matters of the Nova and hence contraction of the Nova. The Nova volume shrinks, in response to the new created three layers interactions' Magravs strength.

Thus, by the time all the Matters settle down to their new position, this new plasma will be a condensed Matter of all sorts. As each Matter layer, due to their plasmatic magnetic fields strength will hold onto their position within the construction of the cascaded plasma. Once, the new matters settle into their new positions from the previous giant Nova condition, this new plasma state of the three Matters will be a fraction of its Nova size. Nevertheless, its size will not be equal to the total size of the collection of all three previous Matters and its initial fundamental plasma size, but its overall internal gravitational fields strength will be enormous compared to its past when it was the whole initial fundamental plasma.

By the this stage the content of the outer layer is still active plasma with bright lights of the Antimatter, but due to the new strong internal gravitational pull of the Matters, even the light that is created through the interaction of plasmatic magnetic fields in the outer regions of the Antimatters, interacts with the strong internal field of the cascaded materials and converges inwards (Chapter 9). This giving the impression that the plasma or the Star is dormant or dying, but in fact the systems has never been as powerful and active internally as this in its whole existence even as plasma or as a star.

Interaction of plasma with plasma

Another possible scenario in interaction of collection of Matters as a plasma, is in collision with another collection of Matters as a plasma.

Electron interaction:

i) For example, there is the possibility that the plasma of the electron can be fused into the plasma of the central nucleus of its proton.

“I have called this interaction and fusing of the electron to its proton plasma the ***inter-atomic fusion principle*** (37). I consider this method to be a simpler and more practical to achieve fusion for production of energy, than the path being followed by the present scientific world in fusion reactors technology in trying to fuse two large plasmas of the proton, without any tangible success up to now after over fifty years of research and development to attain any sustainable fusion cycle to produce usable energy for domestic market. These systems being like of the Jet project (Tokamak) in UK and the present upcoming ITER (Tokamak) project being built in the South of France at the cost of over ten-billion US dollars investment with the final goals of only to produce for a few milliseconds of energy within 10 year’s time.

The process of the inter-atomic fusion is shown step by step in figure 39.

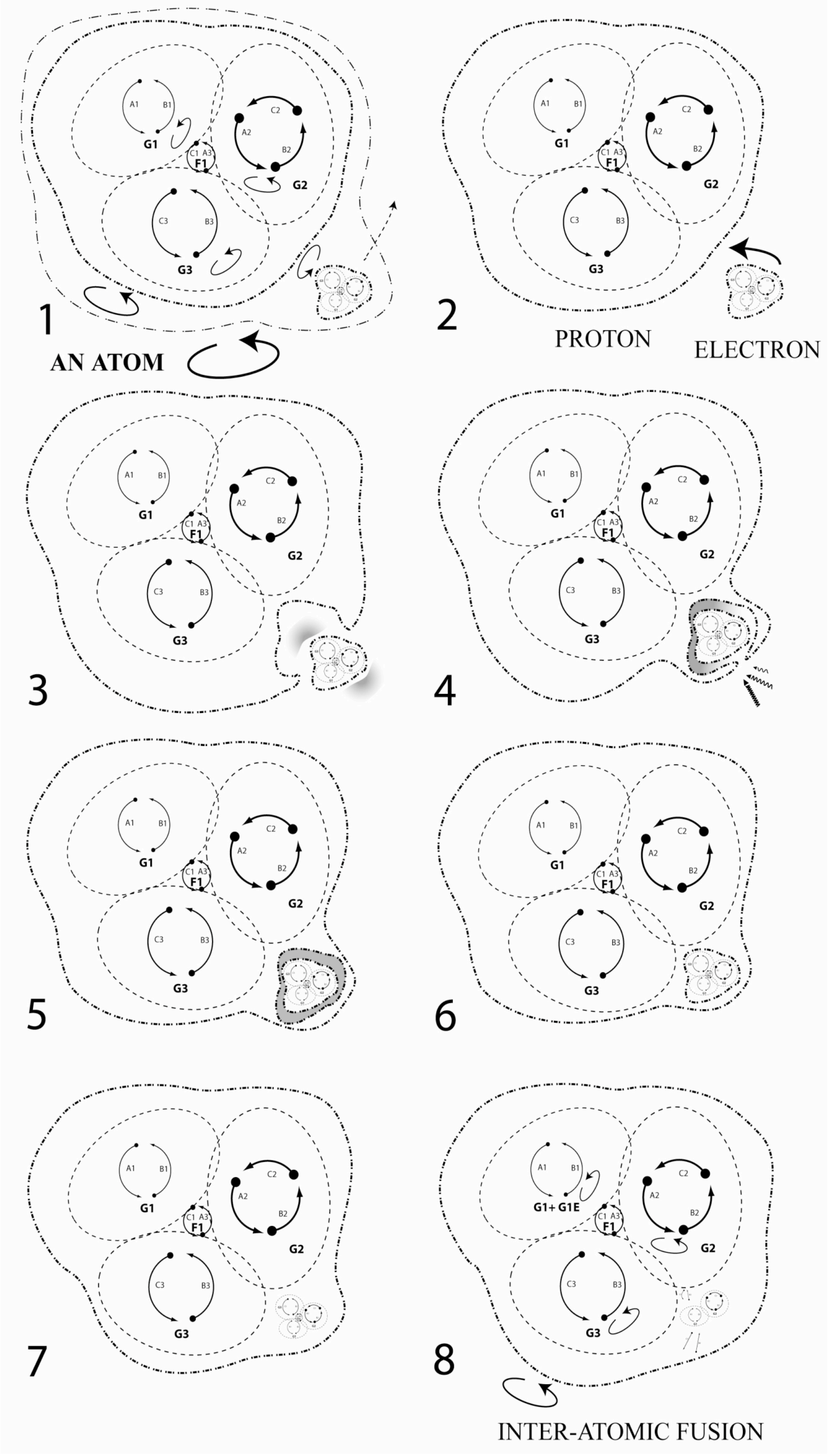
1
AN ATOM
2
PROTON
ELECTRON
3
4
5
6
7
8
INTER-ATOMIC FUSION
A1
B1
G1
C2
A2
B2
G2
C1 A3
F1
C3
B3
G3
G1+ G1E

Fig. 39: Schematic process of the inter-atomic fusion.

Where, initially by using Dilution plasma reactors (Chapter 23, Fig. 57), by feeding the atom of hydrogen into this reactor cores (Chapter 21, Fig. 54), and then in allowing the magnetic atomic structure to dilute and to open up in the soup of the plasma of the reactor (Fig. 39 diagram 1) into its sub-components of electron and proton. At this point the magnetic fields' strength of the soup of the reactor is brought to the levels, which now can be matching with the plasma of the proton pmtics level, for the reactor fields strength to match the overall Magnetic and gravitational pmtics of the protons' plasma. This is the magnetic field strength level of the binding of the plasma, which is called the Coulomb charge barrier of the plasma. Which at this point, the Matters' of the electron are more free and can be gently Magravs positioned to join up in to the Matter components of the proton.

This is the natural method of fusion in its basic form, where there is no need for the huge magnetic fields forces to achieve what happens naturally in the universe to fuse an electron of atom to its proton. The same process of fusion in the plasma dilution reactors' technology can be used for fusing Matters, plasmas or even atoms.

This inter-atomic fusion is done in a magnetic field environment of ***soft*** plasmatic magnetic fields forces and at the same time in a similar magnetic fields forces strength of the same atom environment of the two similar plasmas of the proton and the electron of the same atom, where, in fact the magnetic fields of the outer cores of the reactor performs a similar job as the magnetic rings of the Tokamak systems in providing the containment magnetic fields. Where as in the plasma dilution reactors, at the same time natural gravitational and Magnetic fields needed for fusion criterion is provided by the operation of the fields in two cores of the reactor fields' interactions with each other.

If scientists use this method of creation of Magnetic and gravitational fields forces in the present Tokamak reactors', to replace the solid coil electric magnetic rings, they will achieve fusion in real universal conditions, fusion within natural gravitational fields forces environment necessary to accommodate the fusing of any plasma at a fraction of the present costs and in a very short time-frame.

With these soft magnetic fields in plasma dilution reactors, one can achieve more rapidly fusion than using current induced enforcing magnetic fields rings. These types of magnetic fields strength in dilution plasma reactors are similar to the magnetic fields strength of the plasmas and atoms, and hence they can be manipulated much easier, than, trying to force the plasmas' of two protons.

In comparison, that is to say, to make *an omelette*, it is easier to mix the contents of two uncooked eggs homogeneously as happens in plasma dilution reactors, whereas in trying to make the same omelette, this time by boiling the content of two eggs first into two hard boiled eggs, as Tokamak scientists have been trying to do for past fifty years to make the fusion in reactors by hard balling the content of plasmas (by squeezing the plasma with the use of external magnetic fields to a tighter, harder, smaller entity) and then trying to fuse the two hard balls of plasmas to achieve the fusion process. Then hoping, with increasing the speed of the two hard balls, that they can smash and fuse the two smaller hard balls together. That is why the present path of Tokamak technology in trying to achieve fusion is a sweet dream that will never come true.

In this inter-atomic fusion technology, smaller and weaker plasmas of electrons, which are already within the structure of an atom, are fused into the Matters' component of the proton of its central plasma (Fig. 40). This is a more realistic and practical method of fusion than the present, fusion technology in trying to fuse two large plasmas of hydrogen's protons together, and creating unmanageable heat, of several billions of degrees. Where, the tools and materials to control and harvest the energy of such huge temperatures, on the long term bases in small Tokamak reactors, does not exist within the known galaxies in the universe.

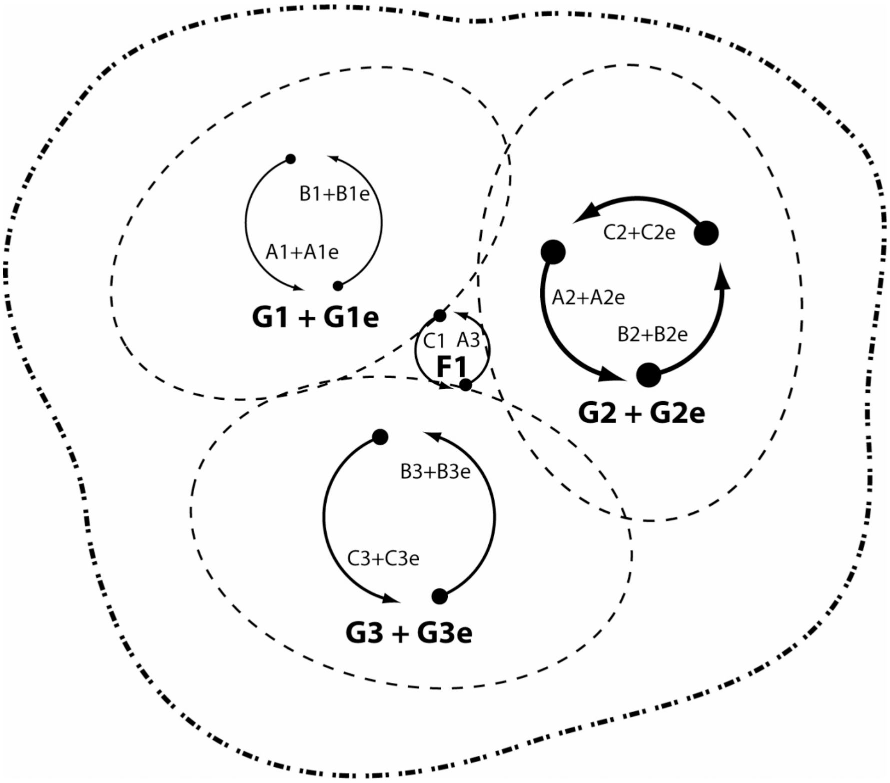

Fig. 40: Schematic diagram of the fused proton and electron plasmas.

In these processes of dilution plasma technologies, when the electron plasma is gravitationally positioned or encouraged to amalgamate its contents with the content of the plasma of the proton, as the two plasmas approach each other, the interaction of two plasmatic magnetic fields of the two plasmas will and can produce several scenarios, which their effects can be harvested for production of different energy levels or production of new materials and so many other purposes.

In brief two cases are discussed here:

a) If the strength of the pmtics of these plasma dilution reactors is set at a level, to be just below that of the total balance of the pmtics strength (Coulombs barrier) of the plasma of the proton, in this case, as the two plasmas approach each other, the initial stages' of the process of fusion starts, which this process begins and will be accompanied with light and heat emission as in figure 39 - diagrams 1 to 5. This light or heat from this process can be used for example for present state of art technology turbine power generators directly. Where, the temperature of the interaction of the systems can be set at the optimum efficiency temperature of the turbine. Hence, operating these reactors in the same configurations as is done in the present fission reactor technology power generators to generate steam to run turbines. With the difference that these dilution technology reactors leave and produce no nuclear waste, and heat can be generated at any level as per demand of the system.

In these reactors, the Magravs strength similar to the plasmas' **Coulombs barriers** charges can be created and controlled. In the dilution plasma reactors' the magnetic fields created through the soup of the plasma will create and maintain pmtics strength, which can be similar to the magnetic fields barrier strength of the plasma of the proton. Where, the strength of the pmtics field in the reactor will decide the strength of the pmtics in comparison to the protons' plasma Magravs strength. As the pmtics strength of the reactor reaches more to the level of the Magravs strength of the plasma, then there is ***less resistance*** *and* ***less friction*** *between the plasmas of the proton and the electron*, when, they are brought together to fuse. Thus as the strength of the magnetic fields of the plasma can be reached by the reactor operation, this allows less friction and hence *less heat* to be produced by the interaction and by the gravitation of the two plasmas of the proton and the electron to fuse. By being able to control the pmtics strength of the plasma dilution reactors Matters' soup of plasma, one can dictate the amount and the rate of the heat release from the interaction of the two plasmas, thus one can control the heat output of the reactor, if the production of heat is the primary objective of the use of such a system.

In general, in being able to control the ***Coulombs barrier strength*** in respect to the given environment of the reactor, one can **vary** this barrier magnetic fields strength in the soup of Matters in the plasma dilution reactors.

Therefore there is no need even to fuse the plasma of the proton with the plasma of the electron, and by just bringing them close to each there to create the right friction when the Coulomb barrier is at the right strength, between the two plasmas, then the reactor can generate a controlled heat-output, or any other rays that is demanded from the systems' operations.

Therefore by *bringing the reactor magnetic fields strength to the level of Coulomb barriers*, one can bring into balance the magnetic strength barrier changes in the overall mixture of the reactor to any level from, or to the Coulombs barrier magnetic fields levels.

Where, the ***difference*** between the two, the **reactor** and the **plasmas' magnetic fields strength**, will decided the ***temperature output of the reactor*** when, the plasma of the electron is gravitated to the plasma of the proton to fuse. The outcome of the bringing of the two plasmas into each other proximity and the interaction of their Magravs, leads to the friction between the two plasmas, and the release of pmtics in form of heat. Where, the temperature of the heat released by this attempted fusion is directly controlled by the pmtics field strength within the reactor and in respect to the variation level from the

Coulomb barriers' pmtics strength level.

An example of this controlled heating of interaction pmtics in the natural world is the like of when the pmtics of the sunlight in atmospheric condition becoming in touch with pmtics of the cell of the human body. Where, as the pmtics of the proteins of the cell are always constant, by varying pmtics strength of the direct sunlight, the heat released from the interaction of these two pmtics, the human feels the increase in the pmtics strength as heat of different degrees. The stronger the pmtics of the sunlight, the more heat is produced from the interaction of two pmtics, thus the warmer one feels. This is like walking from the shadow into the sunlight.

The phenomenon of heating during illness, happens due to the interaction of pmtics of a bacteria within the blood protein pmtics, where as a blood cell comes in touch with a stronger pmtics of the bacteria, the level of the interaction of the extra pmtics from the structure of infection, in ***interaction*** with the blood cell, manifests itself as heat or the increase in the body temperature (this is explained in detail in the paper: Universal order of creation, to be published as a book), where the interaction and possible ***fusion*** of two pmtics of the bacteria and the body cell leads to a release of heat, called "an infection" temperature. This being the reason why the same pmtics of bacteria can have no affect on some human bodies and the same infection can create high temperatures like 41°C +. The reason being that the cell of the one with the high temperature has a lower pmtics strength, thus the interaction of the two infection and the body Magravs pmtics creates more release of fragments of pmtics as heat. When the temperature reaches a level that the Magnetosphere of the human protein reaches its dilution point then the cell is de-structured.

b) When the neutron is needed to be produced from the atom of hydrogen in space for production of heavier atoms, one can follow the process from stage 1 to stage 8 of the Fig 38. In this process the pmtics strength of the reactor are set to be nearer to the binding Magravs strength of the boundary of the protons of the plasma, so that the electrons' plasma literally melts into the protons Matter structure and the matching pmtics of both plasmas of the proton and the electron. Which these plasmas and their pmtics were all of the same origin; this is amalgamation to produce new neutron type initial fundamental plasma. This is the process of the decay of the initial fundamental plasma in reverse, using the same original components of Matters pmtics of all three matters of the plasma. Such a neutron can be added to the construction of the atomic structure of lighter atoms to create **gap(s)** needed for Magravs positioning for the protons of the nucleus to accommodate an extra proton and its accompanying electron for production of even heavier atoms, for example for the production of carbon, oxygen and nitrogen to produce protein base nutrition in deep space travels. Alternatively, one can produce atoms of any matter from initial hydrogen atoms, through the use of the same method of progressive accumulations of protons and neutrons in fusion in plasma dilution reactors.

ii) The other condition is, when the plasma of an electron is attracted to the protons' plasma, ***but does not get absorbed*** within its pmtics, due to their Matters strength friction at the zones of contact of the two plasmas.

As has been explained in previous chapters, the electron is itself a plasma condition and a collection of dynamic pmtics of Matters. Where through the method of its construction and motion around its nucleus, this electron is always in proximity of other electrons of the similar construction or it is in the proximity of larger plasmas such as the neutrons or the protons of the nucleus of its atom.

The ***Motion*** *of the electron* around the proton is created due to the construction of its internal Matters (Chapter 19). Similarly, it has been explained that the interaction of two pmtics due to their friction lead to release of pmtics strength fragments of lower strength of pmtics or rays in the electro-magnetic field spectrum of the light.

Now it is very simple to understand how light is released due to external factors. When the electron of the atom is forced to come close enough to another electron or the components of the plasma of its nucleus or a nucleus of another atom. So that, the interaction of the two plasmas Magravs of the electron and the proton leads to the release of lower magnitude pmtics strength fragments that, its pmtics strength are in the light electromagnetic spectrum (It has been explained that when two plasmas interact and collide; the pmtics fragments of their collisions always releases pmtics in the light bandwidth)(Chapter 7).

This interaction and collision of the pmtics of the two plasmas does not mean amalgamation of the two, but in coming too closed to each other, the two pmtics strength touch and bounce off due to their dynamism.

Hence, as the electron has lost some of its pmtics due to the interaction and collision, with the pmtics of the proton, so now being in possession of different Magravs strength, the electron accepts a new Magravs positioning (Chapter 19) in respect to its proton plasmas.

This interaction of the two plasmas creating the closeness or distance of the electron to its systems' central Magravs point, or what is called that the electron finds a new *ground energy level* position after releasing some of its energies as light, this is then a new Magravs positioning of the electron in respect its protons' Magravs strength forces, which is acceptable by both plasmas Magravs within the atoms structure after collision between the two pmtics of the plasma of the electron and the proton or other plasmas in its vicinity.

Thus when plasmas interact like the plasma of the electron and proton, this leads to the creation of light in the majority of case.

CHAPTER 16

The fundamental difference between Inertia and Gravity

Inertia is said to be due to the total interaction of pmtics of Matter(s) of a solid object without the central dynamic cores that can create Magnetic and gravitational fields forces.

Inertia, through the new understanding of the structure of the plasma, can be simply defined as "the pmtics and gravitational fields due to the Matter in the matter (solid, liquid, gases) components of the plasma (Fig. 41 link F)"

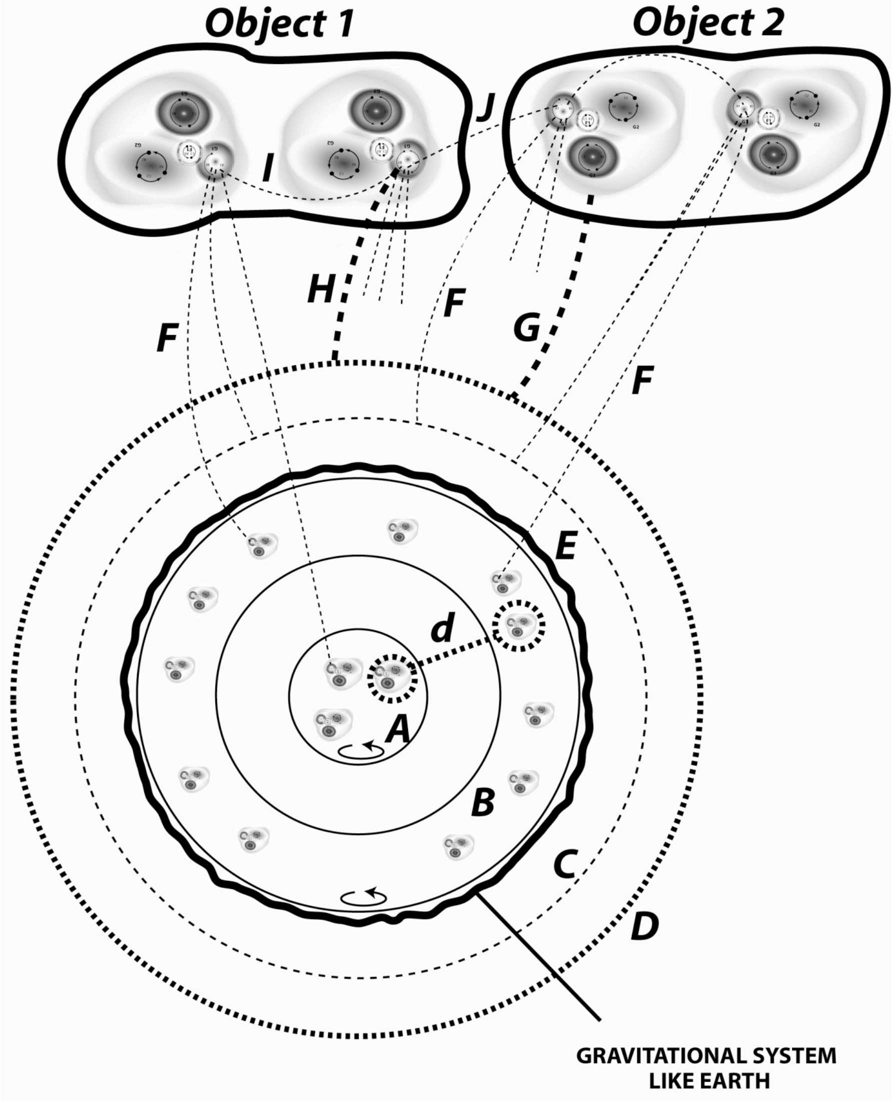

Fig. 41: Interaction and connection between inertia of Matter pmtics of an entity (I and J) in respect to another entity and their interaction (F and H) with magnetic field (C) and gravitational field (D) of systems respectively.

Gravity is simply defined as ”the total pmtics Magravs created by all components of the Matters (Matter, Antimatter, and Dark Matter) of the plasma (Fig. 41 link D and G, and Fig. 42 Link G).”

Inertia is said to be due to a *total collection of pmtics* of the constituent Matter(s) of atoms and molecules of an object, interacting and attracted with each other and any other Matter(s) or matter(s) or Magravs.

The principle of inertia is applicable to the Dark Matters and the Antimatters as well as the Matter components of the plasma in the universe.

As each atom is made of and possesses specific pmtics strength, these fields, in interacting with each other in the state of matter (solid, liquid and gas), create their own total Magravs by addition of the overall magnetic fields of all the atoms of the structure of any matter. The structure of the matter could be made of molecules or an asteroid or any physical object, which is void of gravitational pull, but in possession of pmtics Magravs’ pull of Matter component of plasma, due to the construction of components of its neutrons, protons and electrons of matters. These pmtics, in being interlocked with each other, cannot create a dynamic Magnetic and gravitational pull or interactions between the constituent pmtics of all the Matters of an object. Nevertheless, they create gravitation in between the Matter components of the Matters of plasmas of the object.

In other words they have or create gravitation in respect to Matter-to-Matter components of plasma, or this singularity Matter attraction or gravitation of the total Matter of plasma, is called the **Inertia**.

As the **Gravity** is due to the overall dynamic pmtics strength of all Matters of an object, which is in possession of at least two **sets** of dynamic pmtics in respect to other pmtics, where, the interaction between two sets of dynamic pmtics, leads to the creation of gravitational field forces and simultaneously Magnetic field forces for plasmatic magnetic fields, or rays of magnetic fields, or the whole Matters of an object, this being a planet or a star or others.”

In the case of larger objects with multi-core configurations like Earth, the multiple pmtics forces of dynamic multiple-core objects are loose from each other. Hence, the overall dynamic pmtics of their Matters can interact, creating much more active fields to be brought into play. These being, the stronger magnetic fields constituent components of the plasma; Matter, Antimatter, and Dark Matter, F1 zone fields and all other transition zone fields to interact with each other within the plasma.

That is to say, that the collective **free Matters** of the plasma create a stronger pull or gravitational forces than when the same independent are compacted and cannot be free in a dynamic condition in matter (solid, liquid and gas).

Gravitational fields, due to their free motion and strength of interaction of their Matters with others of the same, can span and reach further a field in the environment.

Gravitational field forces are always stronger and span wider in their given environment, than their inertia of individual Matter counterparts, simply because gravitational field forces of planets or stars are mainly created by all components of pmtics of Matters of the plasma (Fig. 42), rather than the collective plasmatic magnetic fields of individual types of Matters' components of the plasma or the matter.

Planets' and stars' total pull is a combination of the gravitational fields and inertia pull of all their constituent Matters and matters. These composite pulls are created due to the dynamic plasmatic materials of their constituent Matters in the inner cores of these objects in respect to each other, plus the total pmtics interaction of their matters.

Through tests, it has been proven; that the Magravs force strength of a dynamic object is independent of temperature and pressure of their cores. Where, we do not consider this to be the same for the inertia of the matter (solid, liquid and gas).

At the same time, Magravs of a dynamic object can vary as the strength of its Matters pmtics constituent changes. Where, the inertia of the solid matter is always constant, both in respects to the same initial given position.

It is considered that Matters, in general, possess Magravs forces due to their dynamic construction of their pmtics.

As much as the universe has inertia field forces, in respect to Matter components of the plasma, there is ***Antimatter inertia*** in respect to the Antimatters components of and ***Dark Matter inertia*** in respect to the Dark Matter components of the plasma or the universe (Fig. 41). In fact, a collection of the Antimatter and the Dark Matter can possess inertia similar to the physical Matter conditioning of their environment. Therefore, the same matter possesses separate Dark Matter and Antimatters inertias too.

By the same measure, it is to realise that there are Matter Magravs systems as are Antimatter and Dark Matter Magravs systems in the universe. In understanding this clarification, in the systems of attraction of Matters, reactors can be build which are inertia dependent and not gravitational dependent.

It is important to note that; gravitational fields of planets and stars are all created by the Matters components of plasma. This means gravitational fields affect all Matters of plasma (Fig. 42), where inertias of certain Matters affect only that particular Matter component of the Plasma. At the same time, it is clear that gravitational field reactors can be produced in such a way that their gravitational fields *can attract certain Matters only*.

These systems will be like laser systems with the advantage that the laser operation can work ***in reverse***, that is to say the system can send rays of a given pmtics strength as well as attraction or gravitating the same rays and collection of pmtics towards or into the system too.

These systems can or will have, single or multi-molecules pmtics gravitational settings, that can be attracted to, or repelled by only specific pmtics strength, to or from their environment.

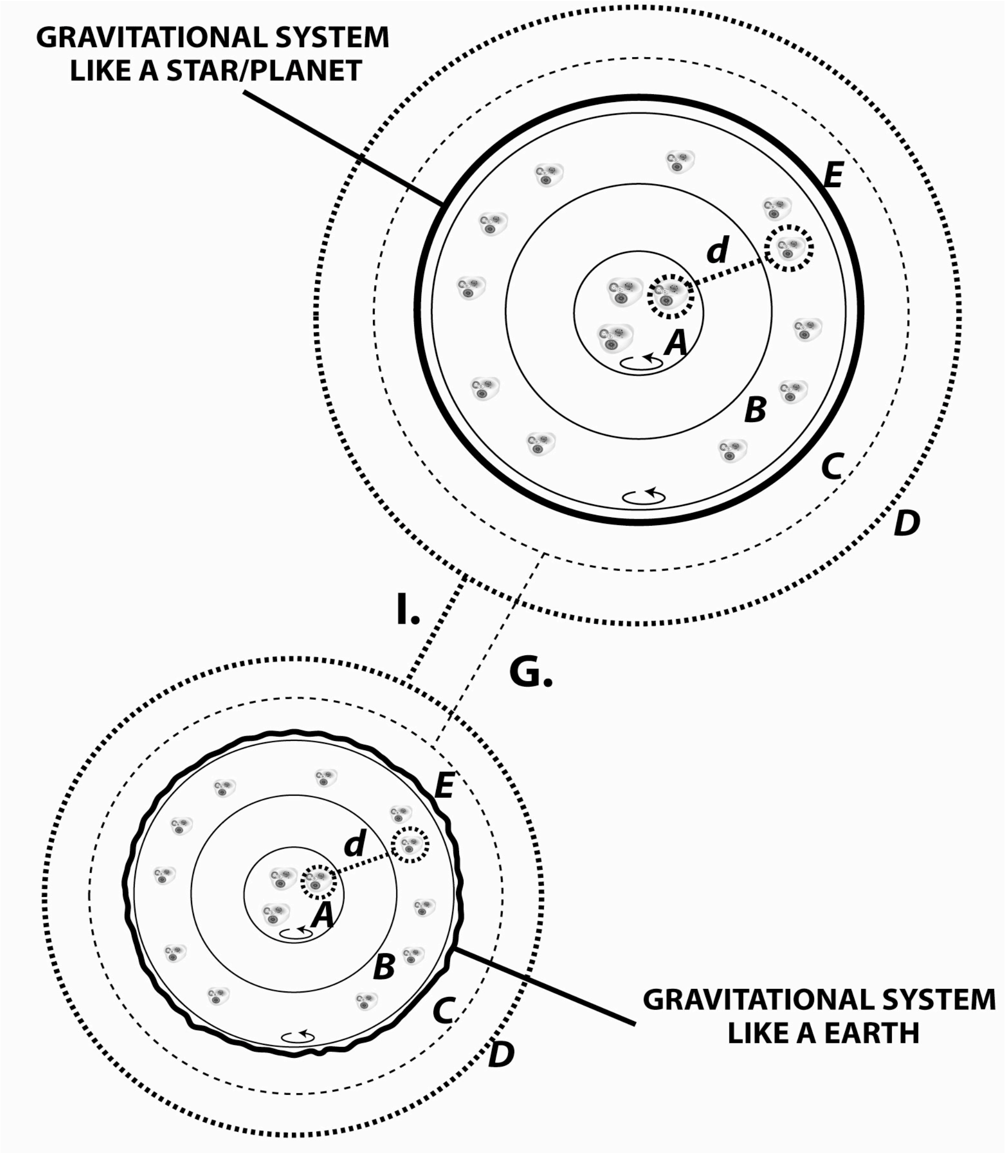

Fig. 42: Interaction between the pmtics Magnetic fields (I) and gravitational fields (G) of two Magravs systems.

Note: In fact the term used as "gravity" has to be explained as gravitational and Magnetic field forces, as when an object is attracted by gravitational fields of a matter, like the Earth, at the same time it is repelled by the Magnetic field forces of the object too. Thus, when talking about gravity it is correct to talk about the overall Magnetic and gravitational fields (Magravs).

These simple mono- or multiple-setting gravitational or inertia systems (Fig. 41 and Fig. 42) can be used for specific *extraction* of certain Matters, matters or materials from the universal Matters or matters soup.

Such systems, will change the present industrial systems thinking that magnetic attraction is only *limited to ferrite metallic objects*. Man has now the possibility to make systems *that can attract* ***any matter*** through the development of these systems."

These systems for example can be used, once set to appropriate pmtics gravitational field strength levels, so that they can extract only the CO_2 from air, or specific poison from water, or rare metals from the Earth and space and so forth (31).

These reactor systems can be used for production of proteins for human consumption and so forth in the future space travels (32, 34, and 38). The production of proteins like matters, through gravitational plasma dilution mixing has been tested and different matters were collected in several tests in summer of 2008. In these tests, the proteins produced and extracted from the reactor cores caused the damage to the ancillary systems of the reactor set-ups. These proteins layers were physically visible and tangible.

Using the mono-Magravs systems (38), new matters can be extracted from other environments, like nano-materials of all sorts. These systems can be used to produce specifics matters too, matters like, uranium, hydrogen and so forth from the plasma of the universe as the needs for their use arises.

Thus through the understanding of the principle of mono- or multi-gravitational and inertia systems, reactors can be evolved that are gravitational to one object or one matter or one Matter only. Through understanding of the interaction of the plasmatic magnetic gravitational principles, reactors have been built, that have a Magnetic or gravitation to plastic at one point, and by resetting its configuration becomes magnetic to carbon and so on. Thus, mono-gravitational and inertia systems can be developed that have specific magnetic attraction or repulsion to one Matter, matters or any Matters of the plasma that can be expected of them through their pmtics strength operation setting.

On the other hand, the repulsion technology of the same Grapos can be used to send the Dark Matter and/or its pmtics energy as the Dark energy of the same plasma of specific pmtics strength into deep space for specific applications.

These types of gravitational lasers type reactors can be used for example in communication systems, where the information packs, using mono-pmtics strength of the Dark Matter and Dark energy can be sent through the system, so that the information can be packaged in the Dark Matter-envelop and sent without resistance or loss of energy, for it to travel with speeds beyond the speed of light to its destination anywhere in cosmos. The speed of communication through mono-Magravs systems are speeds which can be several times faster by order of magnitude than the present used state-of-the-art communication systems with satellites and microwave technologies. These speeds of transfer of information will be, by order of magnitude, faster than the speed of light.

These **GR***avitational l***ASERS** systems, that can be called "**Grasers**", can be used for example to avoid a head-on collision and destruction of crafts by an asteroid in the path of the motion of crafts in deep space (33), or within atmospheric condition. The present missile technology due to its slowness through development of these reactors are obsolete, as once fired to protect the craft in space, for example from asteroids by the time the missile will arrive at the point of destined collision between the spacecraft and the asteroid, maybe it will be years after the craft has past that point.

These mono-gravitational or mono-inertia systems will bring to Man the realisation that the time of making war-machines has come to its end.

As Man will come to understand, soon enough that using these mono-gravitational or inertia systems against one another as tools of war will mean and will have only one outcome, and that is a total wipe-out of the human race from the face of the Earth. The men of war through this technology will come to accept peace not out of their pleasure but out of the fear of their own demise.

The use and capability of the use of this knowledge and of the effectiveness of these *mono-Magravs system*s lethality, has been the cause and the keeper of the peace in the universe. No chance, will be given to the Man, to change the tranquillity of universal peace by use of this technology.

Man through the lethal powers of these mono-Magravs systems will be taught an unforgiving and painful lesson, if he ever tries to use this technology and knowledge to carry on with his aggressive behaviours, as has been his habit for centuries on his own planet. Then this will be the coming of the end of the term of the Man in the universe.

It is clear that any celestial object as the planets, stars and galaxies, are all created through the attraction and interaction of these Matters and matters, their interactions and their constituent plasmatic magnetic fields. These interactions and collections of Matters and matters in respect to the plasmatic magnetic fields strength, which larger objects possess by their order of magnitude, decides' what will be in their part of their environment. That is to say, the plasmatic Magravs created by the plasmatic magnetic fields of Matters in the inner cores of the stars in principle will decide on the size and the span of the dominion of its field in its environment. Then these Magravs strength and their balances dictate what matters or Matters will be attracted to the atom, molecule, planet or the solar system.

This principle of mono-gravitational plasmatic magnetic fields strength forces is the main reason in the universe that why certain planets only retain one element in their structure. For example, this is one of the reasons why the planet Saturn is mainly made of helium on its outer layers as its main constituent matter (17). The reason being that the initial Magravs of the interaction of pmtics of Matters and matters in the inner cores of the planet has been set to be equal to the helium plasmatic magnetic fields strength of matters in that part of solar system in the early life of the planet.

This mono-Magravs positioning field strength of the Saturn is one of the reasons, why this planet, with its strong gravitational field forces, does not pull in and absorbs a large number of asteroids from the Astor-belts. The other reason is that, mono-Magravs systems possess very strong Magnetic fields too, which this is one of the reasons why this planet can keep the matters of asteroids in check.

Alternatively, certain galaxies have more of one matter like hydrogen than the others having nitrogen and so on. Thus, the pmtics of Magravs forces and their mixture with the inertia of a planet dictates what matters are extracted from their part of their environment, like the solar system or galaxies.

Thus, Magravs positioning reactors are not only for creation of motion and for protection, but they will be the main tool for cleaning environments, collecting preferred Matters, matters, etc.

CHAPTER 17

Plasma's and Nuclear decay

The nuclear decay principles of an atom, is in general, considered to be on "the same principles and methods and as the simple decay explained previously for the initial fundamental plasma into a proton and an electron (Fig. 43)".

For the initial fundamental plasma to reach the ground-level plasmatic magnetic fields content condition, the plasma divides itself into two new and exactly similar structures with different amount of mass of the same constituent plasmatic magnetic fields Matters for the creation of a proton and an electron, and releases its unwanted pmtics as residual pmtics or as rays (gamma or x-rays, etc)(Fig. 42). I call this process of the neutrons decay "***The initial fundamental plasmatic decay***."

Through the process of releasing of the exact amount of plasmatic magnetic fields from the **neutron** necessary for the creation of a **proton**, an **electron** and the **release of energy or other rays** from the initial fundamental plasmatic magnetic fields, this, in a way, confirms that *all initial fundamental particles of plasmas are always make exactly of the same and similar plasmatic magnetic fields and the same configurationally structure*. This being another proof that initial fundamental plasmas in the ***universal order of Matter*** are always made through and by the same Specific Entangled Plasmatic MAgnetic Fields (SEPMAF) strength criterion and procedure, as has been previously explained.

The principles of the nuclear decay of initial fundamental multi-plasmas of an atom, is considered to be solely due to the re-balancing of the overall plasmatic Magravs strength components for Matters of plasmas of the atom ***to maintain an overall Magravs forces balance of all participating plasmas***, in the construction of the atom. These being the plasmatic magnetic fields of neutrons, protons, electrons and their constituent Matters, their plasmatic magnetic fields of their Matter, Antimatter and Dark Matter and their magnetic fields forces of each plasma.

For the plasmas of an atom to stay together and hold onto its constituent structure plasmatic magnetic fields, some of these *plasmatic magnetic fields in motion* (energy) are ***transferred to*** *other plasmas*, and in term, these plasmatic magnetic fields are said to be used. At the same time, part of the initial fundamental particles fields of plasmas are used for the upkeep of the integrity of the nucleus positioning and dynamism; As part of the binding energy is consumed for holding the fundamental plasmas together. As in this cycle of consuming pmtics reaches that enough pmtics have been used, that the initial fundamental plasmatic Magravs of the plasma cannot hold onto their initial gravitational positions in respect to each other. The plasmas making up the part of participating neutrons or protons in the centre nucleus of atom, gravitationally reposition themselves within the nucleus for the nucleus to find a new Magravs *balance and equilibrium positioning* in respect to other plasmas.

At this point of magnetically repositioning of the plasma of the heavy nucleus, the nuclei of heavier atoms split into exactly the same configurations of two or more new Magnetic fields and gravitationally positioned balanced plasmatic strength configurations (Fig. 43) of the plasma's of neutrons to protons and electrons, and hence split into two new atoms. This regular splitting pattern of heavier atom to two or more light atoms is known as a ***controlled fission*** of atoms.

As the initial fundamental plasmas were all made of the same initial fundamental particles of the same specific magnetic fields strength and interactions, in their disintegration of denser nucleus atoms, they split and gather into two or more new less denser atoms and of Magravs' configuration, and their need for the new equilibrium of collective plasmas of new atoms.

That is how the atoms are always spilt into exactly the same lower order of numbers of neutrons, protons and electrons in the new nuclei for the overall new balance pmtics of the two new atomic structures to be able to be kept.

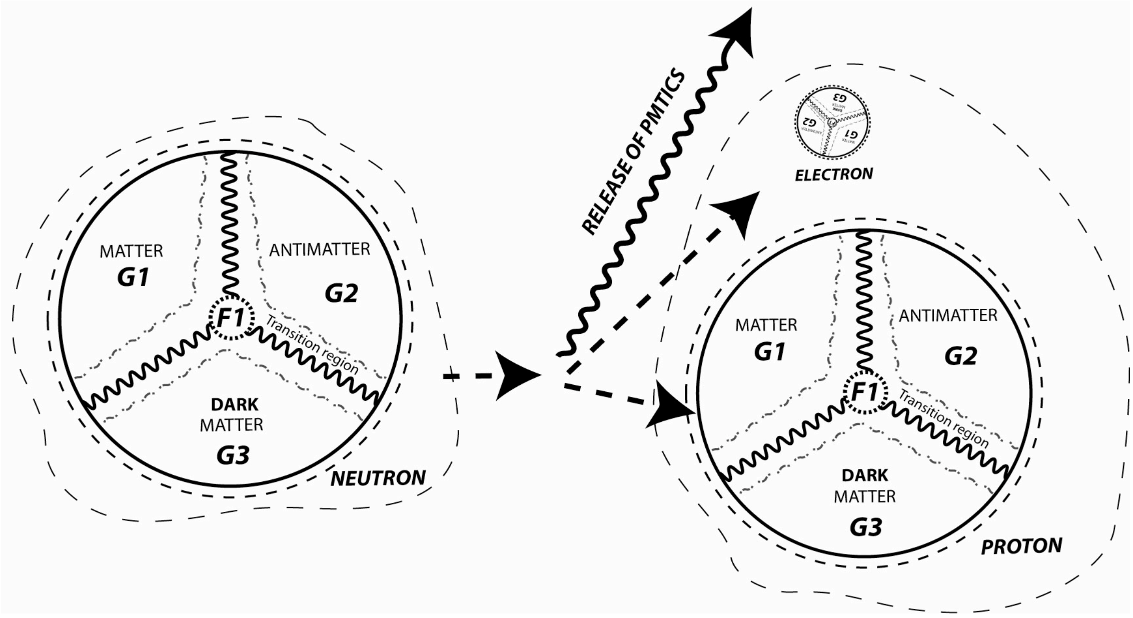

Fig. 43: Schematic of a decay of pmtics and Matters of the Initial Fundamental Plasma (NEUTRON), "the initial fundamental plasmatic decay".

In principle the overall disintegration of all atoms heavier than Hydrogen, involved in the collective pmtics of a given matter, the collective pmtics of the residual matters is always constant, and of the same end-structure of new and lighter atoms.

The overall numbers of the plasmas involved in a structure of a given matter is said to be constant, purely due to the fact that the overall rate of interactions and disintegration of plasmatic magnetic fields are considered to take place at a constant rate. Hence, the phenomenon that takes about the same time for the plasmas to reach a point of using a given amount of dynamic plasmatic magnetic fields (energy), before they reorganise themselves to a new balance condition. This reorganisation or rebalancing of overall plasmatic magnetic fields of the components of nuclei, leads to the fundamental principle of the nuclear decay nuclear of an atom.

This reorganisation of the new plasmatic magnetic fields balance in mass of matters is in so many ways dependent on all of the atoms participating in the total structure of the mass of the matter. This is irrespective of the number of atoms in the matter, these being in the solid, liquid or gaseous state of matter.

The nuclear decay in some plasmas can be triggered and be due to internal changes in the plasmatic magnetic field's structural balance, or can occur due to changes in the external environment of the initial fundamental plasma. These changes can be due to the ***changes*** in the external equilibrium of magnetic fields forces or field strength of other initial fundamental plasmas, or surrounding plasmatic magnetic fields adjacent or near to the initial fundamental plasma.

This being why different atoms of a given matter decay at different rates depending on their position within the overall structure of tangible matter. That is to say ***the decay of atoms is environmental pmtics strength dependent.*** Which the environment can be within the same matter construction of atoms, or the interface between the matter and for example matters of air. Where each environment enforces different rates of use of internal pmtics of different matters within the nucleus of that atom. This being the reason why different rate of decay of atoms within matters are considered to have a **half-life** principle. Where the half-life is, and can be pmtics strength environmental dependent.

Interesting point to consider:

Does the reorganisation take place in the same position first and the same number, which initiates the course of the decay in the rest of the mass?

Is the decay of an atom in a mass happening faster in the inner positions of the mass of matter, or on the edges of the mass?

CHAPTER 18

Neutron-Proton, Electron-Atom

Through the new understanding of the method by which the initial fundamental particles and their subsequent plasmatic magnetic fields interaction, leads to the creation of the initial fundamental Matters (Matter, Dark Matter and Antimatter) leading to the creation of initial fundamental plasmas or the Neutrons, the subsequent degeneration of initial fundamental plasmas through decays further leads to creation of protons and electrons of an atom, and their interactions and collections into molecules and matters (solid, liquid and gas).

Fundamentally, *initial fundamental particles* ***interactions***, leading to the creation of different states of Matters and the creation of initial fundamental plasmas, where these plasmas are considered to be in a ***total plasmatic magnetic equilibrium state*** (Fig. 44 and Fig. 45).

The Initial Fundamental Plasmas are **neutrons** in characteristic. Neutrons (Fig. 44 and Fig. 45), in their construction, are considered as ***balanced*** dynamic plasmatic magnetic fields environments.

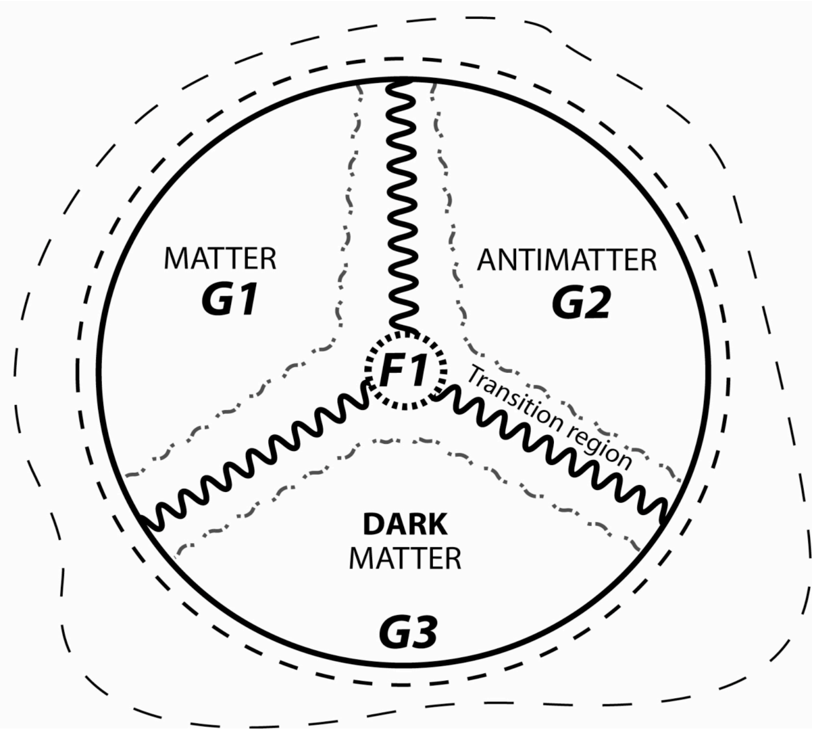

Fig. 44: Schematic presentation of the overall Magravs of Matters and the fields of the Neutron.

Neutrons, due to **the *larger content* of the three Matters and their F1**, they create a ***stronger gravitational pull*** between the Matters of the plasma, consequently their overall Magravs of the initial fundamental plasmas are considered to possess more mass, than the total sum of the masses of the individual proton and electron, from the same plasma after its decay.

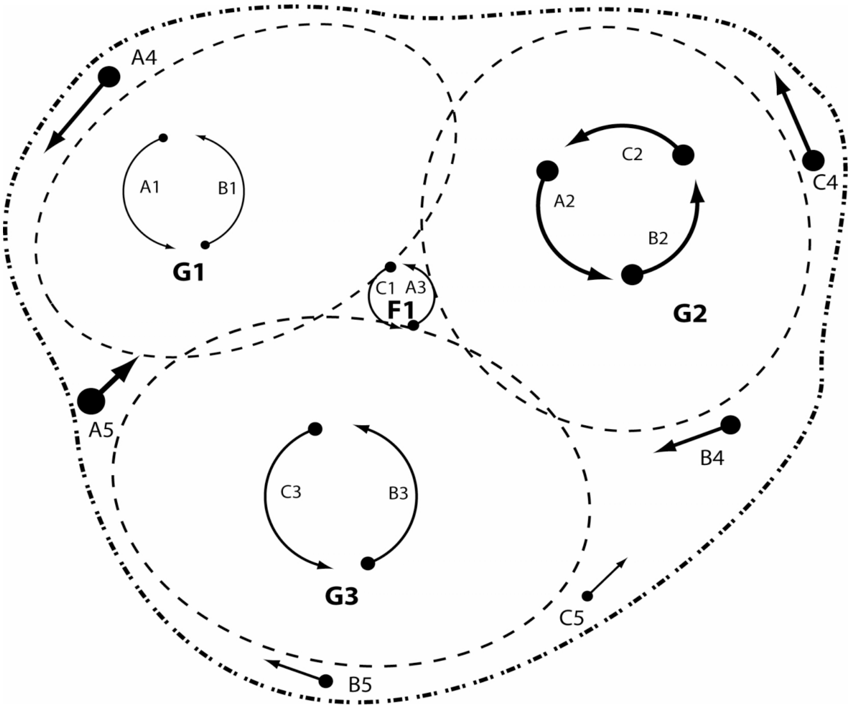

Fig. 45: Matters schematic diagram of the Neutron.

For the ***initial fundamental plasma*** to find its final balanced equilibrium plasmatic magnetic fields force strength condition, due to different effects, the initial fundamental plasma ***reduces*** *its plasmatic magnetic fields balance* to its **ground-level** magnetic fields mass level (Fig. 46 diagram 7) for it to survive, hence the initial fundamental plasma disintegrates or decays' to its sub-Matter levels and fields.

The decay of the neutron

This process has been explained in previous chapters where, now we look at this process in more detail.

The release and division of components of the initial fundamental plasma leads to the release of two independent, yet at the same time interlocked, balanced plasmatic magnetic fields environments plasmas (Fig. 46 diagram 8). Of these two new plasmatic magnetic fields plasmas, one would retain more of the initial fundamental plasmas' mass and be the larger of the two balanced plasmatic magnetic fields (Fig. 46 diagram 7). The second plasma will contain less of the pmtics of the initial fundamental plasma and mass and would be smaller (Fig. 46 diagram 6). The *larger of the two plasmas', would* ***stay in the centre,*** due to its mass and Magravs strength would become ***the proton*** of the atom (Fig. 46 diagram 8). The new smaller balanced plasmatic magnetic fields released can be said to manifest itself as ***an electron*** (Fig. 46 diagram 8) of the same atom.

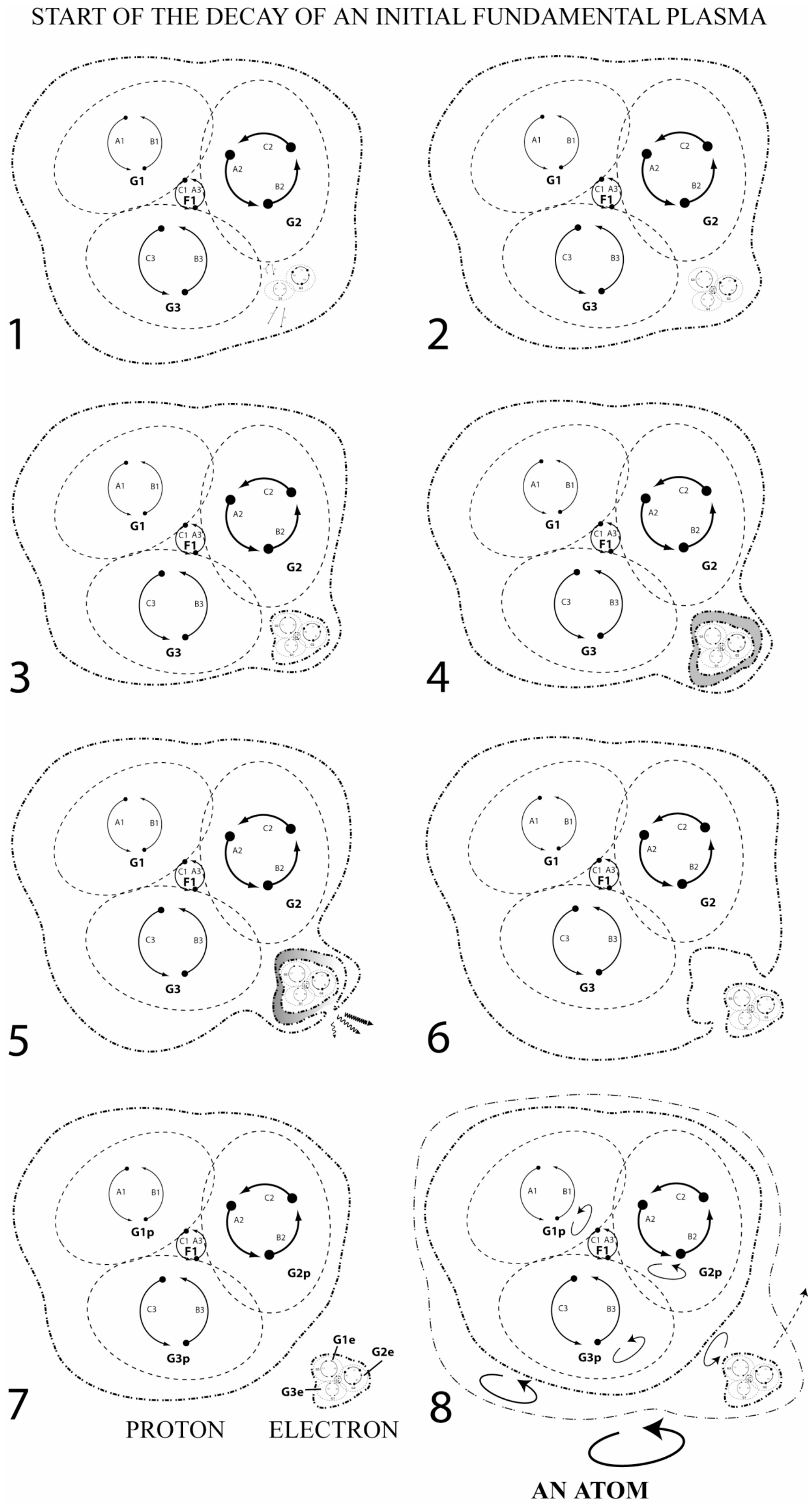
START OF THE DECAY OF AN INITIAL FUNDAMENTAL PLASMA
1
2
3
4
5
6
7
8
A1
B1
G1
C1 A3
F1
C2
A2
B2
G2
C3
B3
G3
G1p
G2p
G3p
G1e
G2e
G3e
PROTON
ELECTRON
AN ATOM

Fig. 46: The eight steps of the Neutron decay cycle to a proton and an electron and then an atom.

It is considered that the ***pmtics of the Antimatter components*** *of the* ***neutron*** **start the process of the decay** of the neutron. As the Antimatter is the main and the strongest of the Matters' of the initial fundamental plasma (Fig. 46 diagram 1), in comparison the Antimatter is the star in a solar system. Then the pmtics of the Dark Matter part of the neutron plasma splits simultaneously as the total balance of the Magravs forces of the plasma is disturbed (Fig. 46 diagram 1). This process is continued up to the point that, due to the cycle of separation of more and more of the two pmtics of the Antimatter and the Dark Matter, leads to splitting of the Matter part of the plasma (Fig. 46 diagram 4) as the plasma tries to keep its overall stability and equilibrium.

At this point, the pmtics fragments of the Antimatter, the Dark Matter, the Matter, and other pmtics within the plasma, through the dynamism of the plasma internal environment, are brought close to each other and then merge to produce a small and independent **mini-plasma** within the structure of the initial plasma.

Where this new embedded mini-plasma in the overall structure of the initial fundamental plasma becomes the structure of the initial fundamental plasma of the **electron** of the atom.

The plasma of the electron by this time is formed internally and has establishes its own balance Magravs zone and hence its own magnetosphere (Fig. 46 diagram 5 and 6). At this stage, the Magravs of the electron plasma tries to find and reach its own Magravs positioning balance in respect to the leftover pmtics and Matters Magravs of the initial fundamental plasma. At this point, as the two plasmas try to find their Magravs positioning in respect to each other, the electron plasma is pushed out-off the boundaries of the initial fundamental plasma (Fig. 46 diagram 7) environment.

At this point where the two plasmas find their Magravs positioning balance, this becomes **the gap** *between the proton and the electron*, or the ground energy level for the electron in respect to the proton' of the plasma. From this point on, the foundation for the final step of the making of an atom with contents of a proton and an electron is reached (Fig. 46 diagram 8).

Therefore the new balanced fields of the original initial fundamental plasma now becomes the components of the two newly balanced plasmatic magnetic fields plasmas of the proton and the electron of an atom (Fig. 48).

From this point on, the motion and rotation of the electron around the proton is set to follow, according to what is described in Chapter 19.

The similarities between the construction of the protons' plasma and the electron plasmas components of Matters, as both have started together quasi-independent life, from the same initial fundamental plasma mother (Fig. 47).

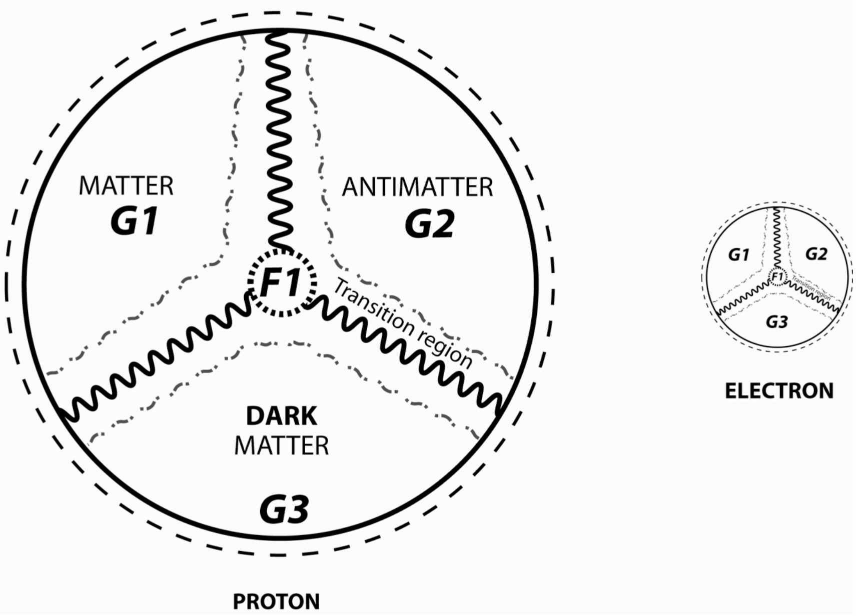
MATTER
G1
ANTIMATTER
G2
F1
Transition region
DARK
MATTER
G3
PROTON
G1
G2
F1
G3
ELECTRON

Fig. 47: Schematic similarities between construction of a proton and an electron (components of Matters).

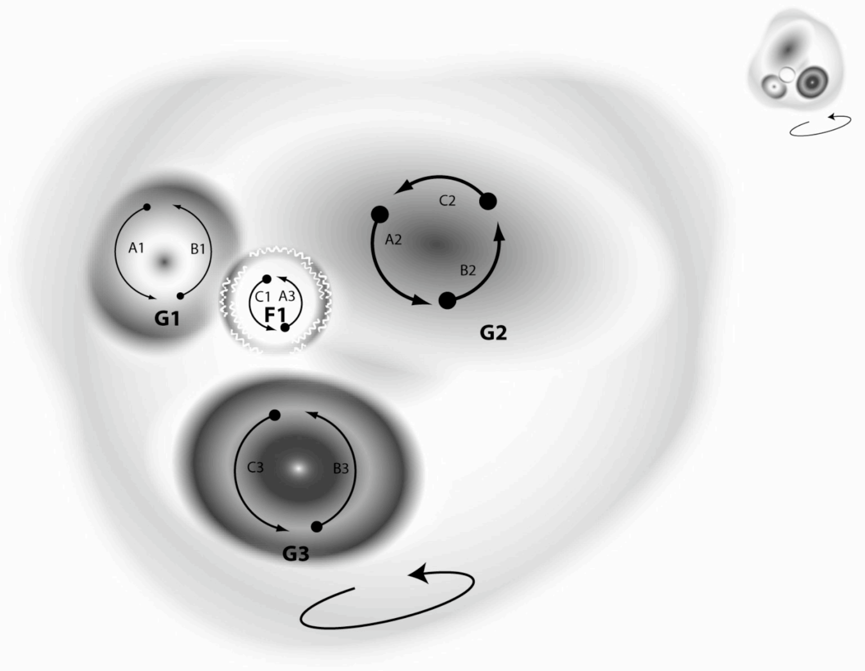
A1
B1
G1
C1
A3
F1
C2
A2
B2
G2
C3
B3
G3

Fig. 48: Pmtics presentation of Initial Fundamental Atom (hydrogen).

As the initial fundamental plasma splits into two balanced plasmas (Fig. 46 diagram 1 to 8), this guaranties the total balance of the integrity of the complete original initial fundamental plasmatic Magravs configuration.

Thus, the splitting and decay of the initial fundamental plasma into the proton and the electron (Fig. 48), called the ***initial fundamental decay***, is a natural decay process as is the ***nuclear decay*** of heavier atomic structures into lighter atoms. Where as the plasma uses some of its pmtics to keep its position, motion and magnetosphere boundary, it needs to break up to smaller atomic substructures that can guaranty its overall magnetic and gravitational field structure, thus the term "*decay of the atom*". This neutron division, to the substructures of a proton and an electron, is for guaranteeing the survival overall of what was achieved before for the creation of initial fundamental plasma (Fig. 48).

This method of the Initial Fundamental Plasma Decay to the proton and the electron is the reason why in atomic structures the number of the electrons and the protons have always been seen to be in ***matching numbers,*** as there is no alternative way in nature for the creation of proton and electron in the universe in general, however individual electrons can be produced from the pmtics in plasma dilution reactors.

In the construction of higher atomic number elements (Fig. 49), for the total balance in both Magravs of the nuclei and their respective electrons, these heavier atoms need to have more of the balanced plasma components of neutrons.

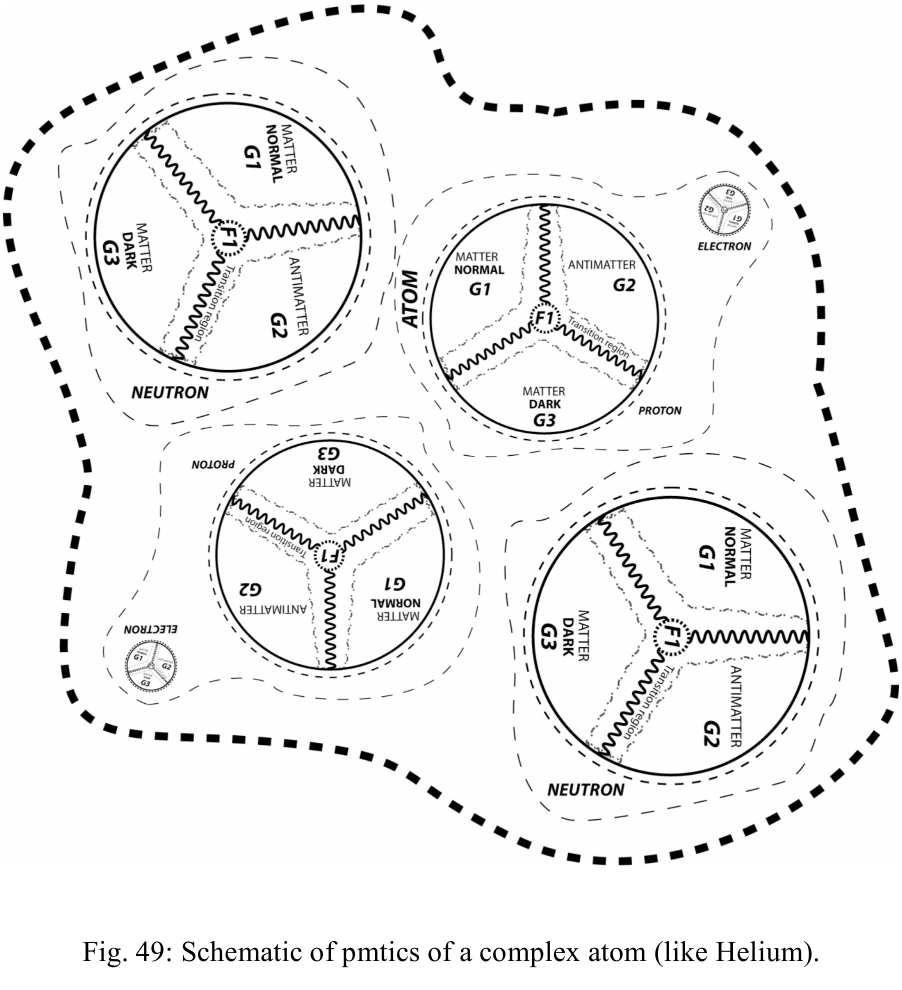

MATTER NORMAL G1
MATTER DARK G3
ANTIMATTER G2
F1
Transition region
NEUTRON
ATOM
MATTER NORMAL G1
ANTIMATTER G2
F1
Transition region
MATTER DARK G3
PROTON
G1 G2 G3
ELECTRON
MATTER DARK G3
PROTON
Transition region
F1
ANTIMATTER G2
MATTER NORMAL G1
G1 G2 G3
ELECTRON
MATTER NORMAL G1
MATTER DARK G3
F1
Transition region
ANTIMATTER G2
NEUTRON

Fig. 49: Schematic of pmtics of a complex atom (like Helium).

Note: Through this simple configuration (Fig. 49), one can understand why certain chemical bindings and structure of some atoms and molecules of certain proton combination interactions in a given environment become possible.

In fact, the neutron, due to its balanced magnetic plasma or neutrality, and at the same time possessing its own Magravs magnetosphere, becomes the space gaps and space(s) needed for Magravs positioning of protons, in respect to each other, as much as magnetic fields equalisers or balancers in the construction of the nucleus of atoms with more than one proton (Fig. 49).

In the heavier elements, due to the tightness of packing, if there were not so many neutrons to create some magnetic and Magravs positioning balance gap, or what we call **gaps**, then due to the protons gravitational pulls, part of the nucleus would most probably clog-up, this being due to the interlocking of the Magravs of protons in the nucleus of the atom. In the majority of cases, this ***clogging-up*** would most probably lead to the explosion of the nucleus, due to *repulsion* of the fields or Magnetic fields of Magravs within the construction of the nucleus in being too close to each other's proximity. On the other hand without the existence of neutrons as a gap the nucleus would not be a gravitationally positioned stable configuration.

A configuration of the complex structure of the initial fundamental plasmas like an atom of the Helium and the positioning of its neutrons, protons and electrons are shown in figure 49. This shows the overall gravitational fields and Magnetic fields generated by interaction of these individual parts, in respect to the dynamic Matters and the fields each part produces both internally, and in their overall external boundary interaction of the plasmatic Magravs of the atom.

By a close observation of the proton and electron's Magravs balance, it is now clear how the *different* ***electron orbits*** *of heavier atoms* come to be created. That is to say, as the number of protons increases, the boundary of the Magravs of all protons in the centre of the nucleus increases, thus allowing and accommodating the next orbit for the motion and Magravs positioning of the next layers orbits of electrons. It has to be remembered that protons are *dynamic* entities, thus not more than certain numbers can revolve in the compact zone of a nucleus. So as the number of protons and their equivalent electrons are added to the nucleus, *more room is needed* for the *free* motion of these protons. Hence the only way a neutral zone can be created for the free motion of these additional protons for the structure of nucleus to be Magnetically and gravitationally in balance, is to add more balance neutrons to create an environment for the free motion of the protons and their Magravs positioning.

Thus as proton number increases, and the volume of the nucleus increases to keep all protons in motion, and hence more numbers of neutrons are needed. This is the reason why one sees a larger number of neutrons in the nucleus of heavy atoms. The more neutrons and protons are in the centre, the *larger the magnetosphere of the nucleus* boundaries. Hence, the accommodation of more electrons in *further distances locations* from the centre nucleus. Where all these electrons, protons and neutrons overall pmtics Magnetic field and gravitational field have to be at all times in balance.

It is fundamental to understand that " the negative field charge of any magnetic fields is due to its Gravitational field strength or pull, and the positive charge of the fields due to the Magnetic fields strength or push of the Magravs of the Matter, like a planet.

In electric field charge and current flow, its to be understand that the **positive charge** is the flow of the Magnetic field push of the Magravs, and the **negative charge** is the gravitational pull of the flow of the gravitational fields of the Magravs of the plasma and its Matters content. This in a way explains the ***Resistance*** and the ***Impedance*** respectively of the matters in the electrical section of physics. Where the impedance is the measure of the gravitational pull of the flow of the pmtics in respect to the matters' atoms and where the resistance is the measure of the Magnetic push of the flow of the pmtics in respect to the matters' atoms.

CHAPTER 19

Motion of the Electron around the Proton

The constant rotation of the electron around the nucleus of an atom indicates that the electron and the nucleus of the atom constantly maintain an overall balanced Magravs forces between and Magravs positioning, in respect to each other.

Within the structure of the initial fundamental plasma (the neutron), and through its sub-dynamic Matters (Matter, Antimatter, and Dark Matter) and at the same time, with all Matters of the plasma being in possession of their own independent internal dynamic plasmatic Magravs forces, they all exert dynamic and varying levels of Magravs on different constituent of Matters of plasma internally, which these field forces span beyond the boundaries of the nucleus plasma itself individually and collectively.

The total effect of these dynamic forces, from the nucleus plasma (proton) (Fig. 50), will affect the *path* of the *rotation* of the satelliteing electron around the nucleus.

Similarly, the *plasma of the* ***electron*** (Fig. 50) itself is *made of the* ***same three dynamic plasmatic magnetic fields structures*** *of Matters as of the proton in its nucleus*.

The effect of these **dynamic** plasmatic Magravs of different Matters in the centre plasma **(proton)** of an atom on the **dynamic** plasmatic Magravs of different Matters of the plasma of its **electron,** and their collective effect of each plasma on each other, is the reason why the electron does not have a constant and a defined path of motion around the nucleus.

That is to say, the electron plasma by itself is in possession of its own individual Matters' (the Matter, the Antimatters and the Dark Matter) Magravs forces' as well as their collective plasmatic Magravs forces of all the Matters of the electron, and the same applies to the Matters and the plasma of the proton.

As the Magravs of different Matters of the electron passes different Matter components of the proton, each of the facing Matters' Magravs forces within the proton and the electron have to interact with each other, plus the electrons' and protons' overall Magravs positioning in respect to that particular Matter in the proton and electron. Which, this type of interactions and Magravs repositioning between the two involved Matters of each plasma will keep on repeating on continuous bases as both the proton and the electron are independent dynamic entities. As the ever-continuous Magravs positioning between the two dynamic entities leads to the ever changing in position of the lighter by mass (the electron) of the two entities, to move around the proton, hence this chasing-up of the ever-repositioning movement becomes the "uncertain" position or erratic motion of the electron around its proton.

This giving the electron in that given configuration of interactions of the Magravs its *pulled* ***closer*** *to* ***or pushed*** *further away* from the nucleus. Then, immediately, the electron has to finds its new overall Magravs balance positioning, in respect to the proton.

Then, due to the dynamism of both plasmatic Magravs of proton and electron, other Matters' fields in the two plasmas come into play and face each other and their Magravs of these two new Matters and overall Magravs have to interact again.

Therefore, the new facing Matter of the plasma of the electron Magravs have to find a new Magravs positioning in respect to the new Matter of the proton's Magravs and the overall Magravs of the two plasmas.

Further on, again, as both the proton and electron are dynamic, the electron has to find a *new Magravs positioning*, in respect to these new Matters facing each other from each entity interactions, this change of Matters facing each other takes place on a continuous basis, as different Matters and Magravs of both components and the overall structure of each atom come to exert their forces on each other and the overall balance distance of both the electron and proton position is to be maintained as every time new Matters in each entity come to face each other. We call this the ***Magravs positioning principle*** (or Magravs positioning).

The result of this continuous change of the Magravs positioning of the Matters of the electron in respect to the Matters of proton and their overall Magravs, causes and creates the motion of the electron around its nucleus.

The behaviour and interaction of the ever-changing of Magravs positioning of the proton and the electron Magravs in respect to each other, leads to the creation of an erratic, never duplicated path, and is the cause of the motion of the electron around the proton.

That is to say, as the electron rotates around the proton, its position and distance in respect to the proton changes, according to the overall position and balance of the gravitational pull forces and Magnetic field push forces of all three components of its Matters and their respective plasmas, in respect to the position of all three Matters Magravs of the protons' and overall Magravs of the proton.

Nevertheless, as the Matters, in the proton and the electron are both dynamic themselves, the rotation of the electron around the nucleus is changing by the principle of the Magravs positioning of the Magravs balancing. Where, this leads to the ever motion of the electron around the proton of an atom (Fig. 50).

As the overall strength of the Magravs of both components of the atom are the same, this leads to a minimum and maximum distance that can be attained by the electron field strength, in its position in respect to and from its proton. This explains why an electron appears to be moving erratically around the centre point, with a given minimum and maximum distance from the centre proton. This minimum and maximum is to be known as the amplitude (minus and maximum) of the distance of the electron from its proton.

In fact, as the Matters dynamic components Magravs of the electron interact with the combined dynamic Magravs of the proton, the electron, which is the less magnetically dense of the two, has to try to find its position, in respect to all the components of the denser magnetic fields plasma of the proton.

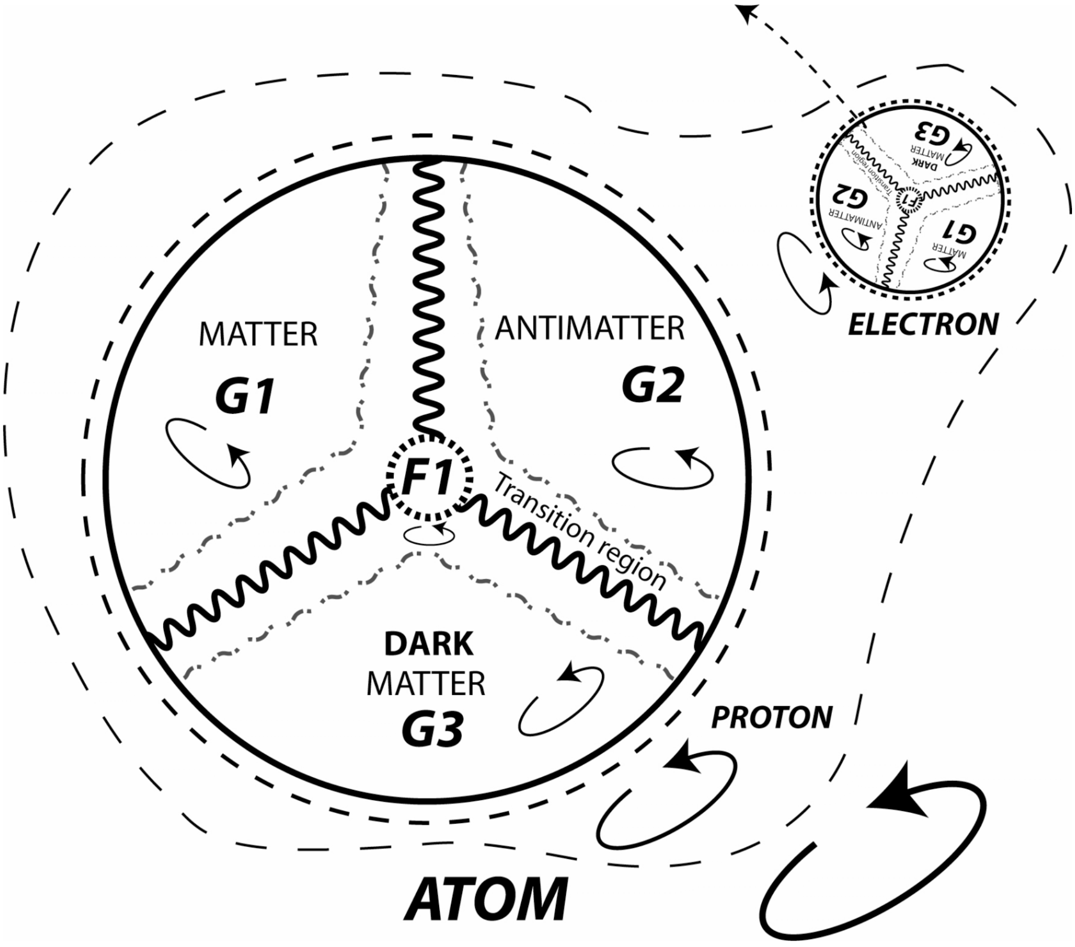

Fig. 50: Schematic of dynamic pmtics of Matters and fields of the proton and the electron and their interaction of individual Matters and their overall plasma and Magravs, leading to Magravs positioning and motion of electron in respect to its nucleus.

This continuous change in the Magravs positioning of the electron, in respect to the proton, leads to the continuous rotational motion of the electron around the proton (Fig. 50)

In fact, one can conclude "the proton and the electron are made up of composite Magravs of several different Matters by simply observing the erratic path of the motion of the electron around the proton".

In principle it can be said that the magnetic fields strength of any magnetic fields or Matter is ***position dependent*** of the two parts of the fields from each other, which the field has been released from. This is how magnetic fields strength in principle are determined and created. (Paper: Creation of Magnetic fields).

For example, the interaction of the Magravs of the Dark Matter's in the plasma of the electron interacting with the Antimatter's Magravs in the plasma of the proton and their overall Magravs of the electron and proton plasmas in respect to each other, decides the position and the distance of the Dark Matter of the electron in respect to the proton.

Universal Magravs positioning principle

The effect of the composite components of the proton and the electron's Magnetic fields and gravitational fields strength, changes the path of the electron as it approaches each Matter of the proton. In composite nucleus, the overall fields' strength of all the protons of the nucleus dictates the distance of the electron from the nucleus.

All motions of objects and fields in the universe are due to the ***overall** Magravs positioning of one pmtics of one object or ray*, and so on, *in respect to another*.

I call this dictation of position of an entity in respect to another entity, due to their Magravs positioning distance, the "**universal Magravs positioning principle"** for motion in the universe.

This Magravs positioning of Matter takes place within the constraints of Matters' possessing their own composite pmtics Matters fields‘ strength.

Universal method of motion

In principle, "the **continuous dynamic Magravs positioning** of one dynamic single ray or multiple Matters of multi-cores plasmas of an object in respect to another dynamic single or multi-Matters of a multi-cores plasmas of another objects, **leading to and/or can lead to and cause the motion** Matters, matters in the universe in respect to each other".

I call this method of **continuous change in Magravs positioning** of one entity in respect to another causing motion in one of the two entities, "**The Universal method of motion "**.

Motions of atoms and their components interactions, the molecules, solar systems and galaxies in the universe, are all based on the ***same** simple universal method of motion*.

The dynamic plasmatic Magravs of Matters within the constituent structure of one entity dictates the position and direction and the path of the motion of one object in respect to another dynamic plasmatic Magravs of Matters within the constituent structure of the entity. Motion of components of atoms, motion of planets of solar systems and physical contents of motion of galaxies in the universe are all based on this principle.

The universal method of motion of any two gravitated objects in respect to each other is based on the dynamic plasmatic Magravs strength interaction and attraction of all of its complete composite components of Matters and forces within the object, ***irrespective** of the physical size* of the system. All universal gravitational objects follow the same principle of motion as an electron around its proton, the same as is for the motion of Matters in the initial fundamental plasmas.

It is correct to say, "**motion in the universe** is totally and exclusively caused and is due to **Magravs strength positioning** of rays, fields, matters, Matters electron, planet, stars and galaxies in **respect** to each other".

Thus from this point on it is clear that **one does not need to burn fuel** to create **motion, but use Matmags**. However, one only needs to create the right Magravs strength within any system to attain dynamic Magravs positioning in respect to another object, this being, a planet, a star, a solar system or a galaxy for the system to be in motion.

This method of utilising the plasmatic Magravs interaction for the creation of Magravs forces within a system, that brings about the motion of one object in respect to another due to Magravs positioning, is to be known as the "**Magravs positioning methods and universal principle of gravitational motion".**

Thus, the Magravs positioning principles of motion are the future for the creation of motion for automotive, marine, air and space industry.

Through this method of Magravs positioning, there is no need for the burning of fuel in engines to move cars or use liquid fuelled jet propulsion systems to fly rockets into space.

When speaking about gravitational or Magravs positioning, this means the collective effect of gravitational field and Magnetic fields of one system in respect to another object or plasmas, Magravs of planets or stars.

It is important to remember that in the universal order of motion, the ***plasmatic gravitational field*** and ***plasmatic Magnetic fields***, are *integrated*, thus simultaneously they co-exist and one cannot exist and operate without the other.

In the universe the **Magravs positioning of Antimatters** in respect to each other and their plasmatic magnetic fields does exist. Further, there exists the **Magravs positioning of the Dark Matter** in respect to each other and their plasmatic magnetic fields too.

The way the Magravs systems of the future will be set-up and operate will dictate in which gravitational fields strength and dimensions the crafts of future will travel in.

In man-made systems using the plasmatic magnetic fields of the Matters that, through their interaction with each other, create the Magravs needed for positioning*: simply by creating Magravs* ***within a reactor***, and these systems will, by trying to find its new Magravs positioning balance in respect to the Magravs, for an example, in atmospheric area of the planet, cause the motion of the reactor system in respect to the surface of the planet.

The creation of motion through Magravs *positioning* can be used in an aircraft equipped with gravitational reactors in respect to Earth, or with a spacecraft in respect to a planet(s), solar system or galaxy.

Methods based on the Grapos principle for motions are the future for the creation of motion in space, according and in parallel to the way the universe is operating for millions of years for its dynamic existence.

CHAPTER 20

Molecules

The creation of plasma and the way its plasmatic Magravs interact to bring about a balance in its overall plasmatic Magravs and their effect on each other, leading to further interactions and balances in different atoms and their components in different dimensions in atoms has been explained. This balancing and positioning of plasmatic Magravs in between different atoms of different Matters brings about how the final Matters, plasmas and atoms can be brought together, leading to how molecules are come to be generated.

That is to say, for the creation of a molecule, it is necessary that the plasmas of protons and neutrons of any nucleus of an atom, and atoms of different elements, need to have a total plasmatic Magnetic field forces and gravitational field forces positioning balance in a given environment so that these atoms can hold onto their total structure, for the molecules to be created.

Thus, by the same principle a molecule or a cell can only be created where, the Magnetic fields and gravitational fields positioning are in equilibrium in all its participant atoms plasmatic magnetic fields (Fig. 51).

At the same time, the whole of the structure of the molecules' plasmatic Magravs have to find plasmatic Magravs balance in respect to their surrounding environments Magravs for the molecules to be able to exist as an entity in a given environment. This in fact, explains why certain atoms and molecules, in certain configurations, can never exist in one environment, while they are prevalent in other environments, like planets, or galaxies and why different pmtics forces and conditions can apply to the same Matter, and Matter manifest itself as a matter in state of solid, liquid and gas.

Likewise, the same can explain why proteins of specific characteristics are created or maintained by certain cells in one part of the human body and not in other parts of the body (9).

In fact, if the same conditions of plasmatic magnetic fields are, or can be, maintained in any part of the universe, one can reproduce exactly all the Matters known to Man anywhere in the universe.

That is to say, when one can create the required balance and equilibrium environment in the total and overall Magravs of the atoms participating in the construction of a molecule, one can replicate and maintain any molecule. Therefore this explains that Van der Waals forces of **pull** and **push**, merely can be known as the **gravitational** fields forces and **Magnetic** field forces respectively operating between different molecules, irrespective of their binding structure with other molecules or atoms.

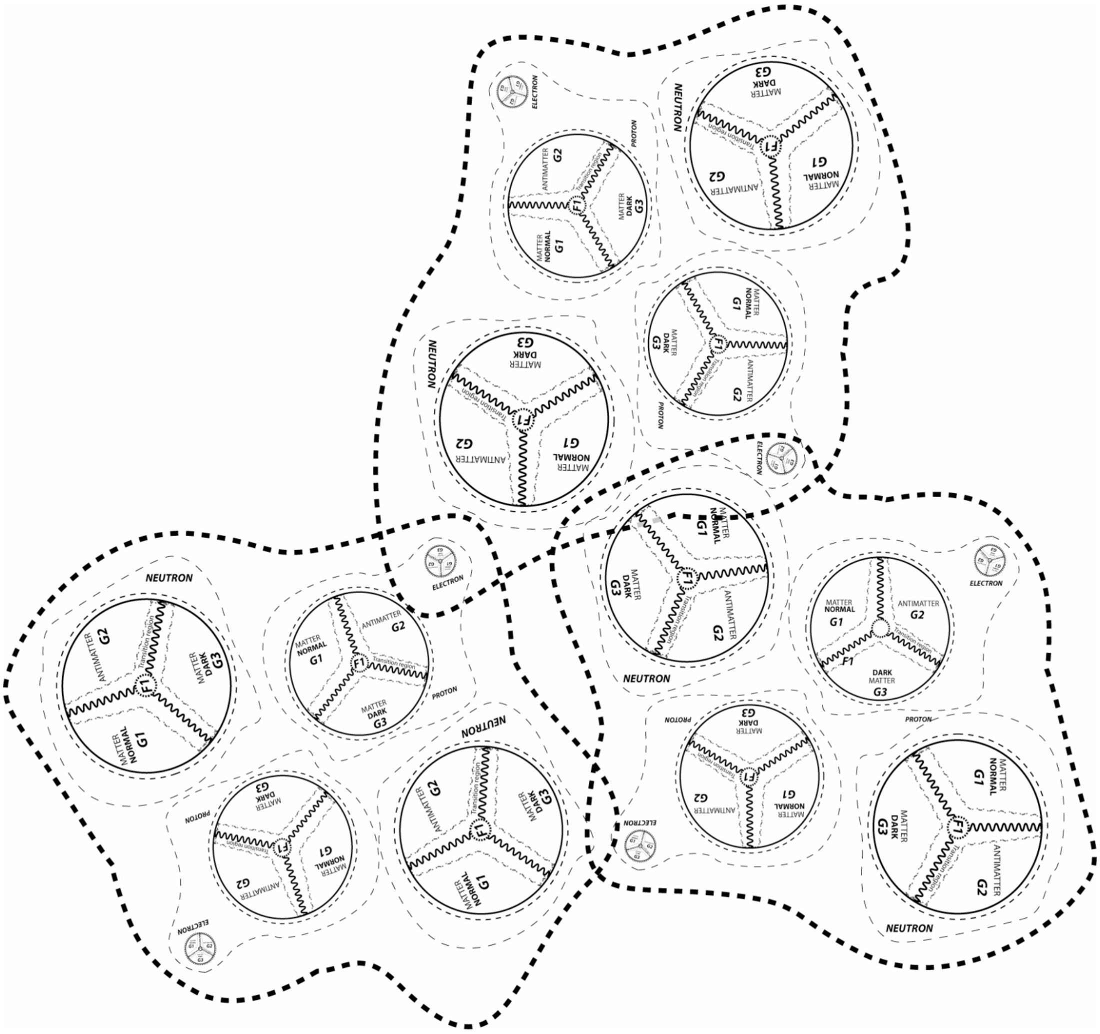
ELECTRON
NEUTRON
MATTER DARK G3
MATTER NORMAL G1
ANTIMATTER G2
F1
PROTON
G2
ANTIMATTER
MATTER DARK G3
MATTER NORMAL G1
NEUTRON
MATTER DARK G3
ANTIMATTER G2
MATTER NORMAL G1
PROTON
ELECTRON
NEUTRON
ELECTRON
PROTON
DARK MATTER G3
ELECTRON

Fig. 51: Schematic pmtics of atoms gravitational (Magravs) positioning to form a molecule.

In the industrial Magravs reactors of the future, one can produce as many atoms as one needs by just absorbing the plasma from space or the environment, for any application and purpose like the production of the hydrogen molecules, carbon, nitrogen, oxygen for the production of proteins for food, air, water and so forth, by simply replicating the Magravs of the preferred molecules and allowing their positioning and interactions.

In the future, by giving a standard chart of Magravs strength and prevailing magnetic fields forces, temperature and pressure of plasma, atoms and molecules, one can produce and maintain any atom and molecule anywhere in the universe.

CHAPTER 21

Plasma Dilution Technology

The Plasma Dilution Technology is the new approach for the use of new gained knowledge for creation and control of plasma and its sub-Matters constituents Matters and their fundamental elementary particles.

The new understanding and insight into the construction and the structure of plasma and its constituent Matters and their magnetic fields, opens up new opportunities for the creation of energy and motion, production of new materials, for medical uses and so many other applications.

The plasma dilution technology allows the plasmas' Matters components to be diluted within a soup of Matters and their pmtics made up of exactly of the plasmas' own composite structure.

Through the specially designed ***nuclear reactors***, conditions are created within the structure of the reactor that the plasma is contained in a safe and secure way in reactors cores configuration.

The operation of these reactors can be controlled so that they do not release any radiation into the environment and they create no waste of any kind, this being the likes of radioactive materials or CO_2 waste, as has been the trend in the current nuclear and energy production industry.

Through the understanding and creation of a parallel environments, similar to the internal environment of the same weak plasmatic magnetic fields of the plasma itself, plasmatic magnetic soup environments are produced in these types of reactors, so that the plasmatic magnetic fields condition within the reactor cores matches the plasmatic magnetic fields of the binding forces (Coulomb Barrier) of the initial fundamental plasma, as the components of this plasma soup, this barrier is magnetically diluted and soften, allowing the plasma to open through the operation of the reactors, for the Matters of the plasma to become part of the dilution soup of Matters and fields forces of the reactor.

A soup of weak plasmatic magnetic fields

These plasmatic magnetic fields saturated reactors will allow components of any plasma introduced into the reactor to be loosened and, the Matter, the Antimatter and the Dark Matter of the plasma, to be free to move within the structure of the plasma soup, which has been created through specific setting and Matter combination. Matters once loosen from other Matters within the soup of the plasma, these are then free for their fundamental particles effects to be utilised to achieve any desired effects, like for fusion and production of new matters, lift and motion.

Reactors that facilitate such a dilution of plasma I call **Plasma Dilution Reactors** (Fig. 56) and their process of dilution as the **Plasma dilution principle** and their knowledge as the ***plasma dilution technology*** (Fig. 52).

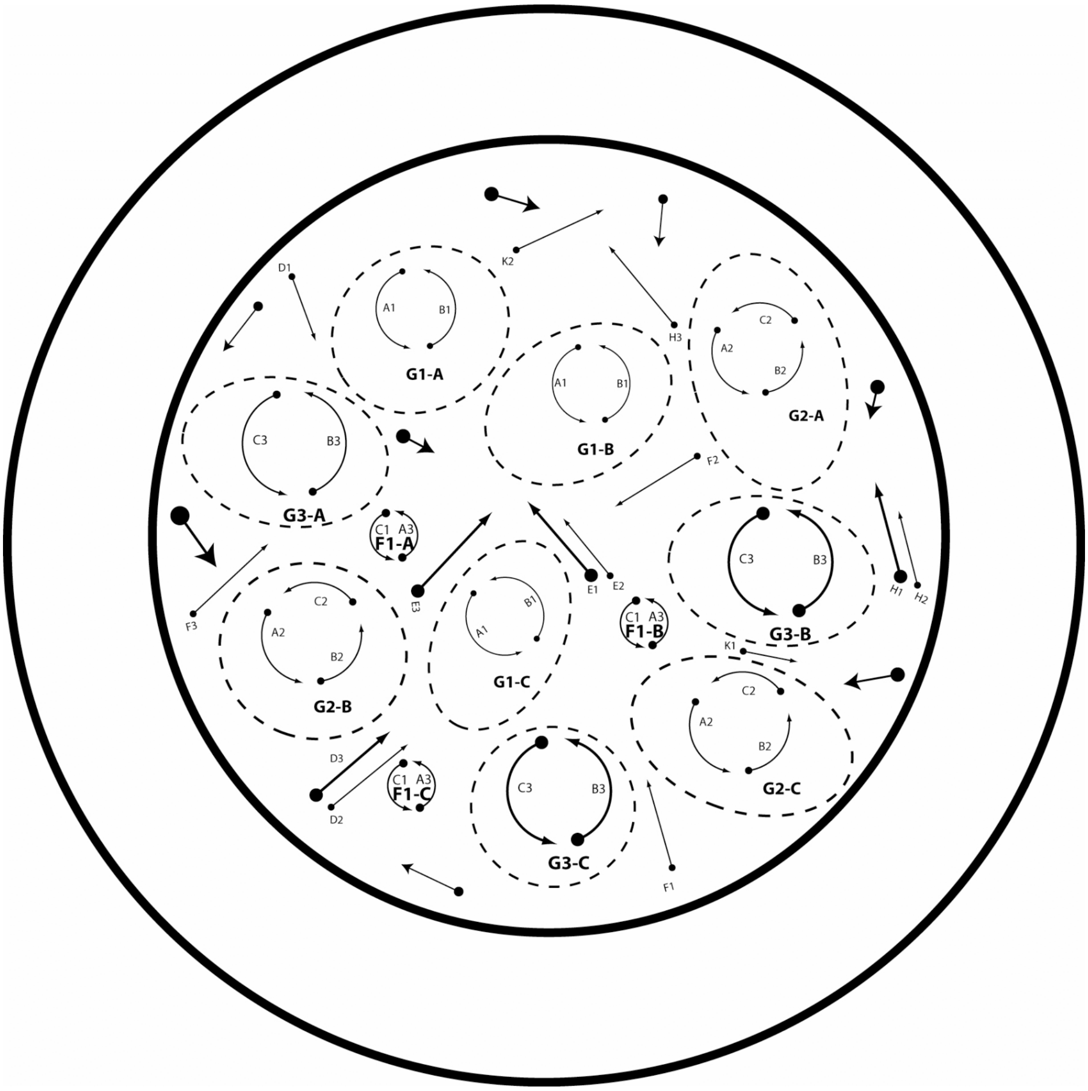
D1
K2
A1
B1
G1-A
H3
C2
A2
B2
G2-A
A1
B1
G1-B
F2
C3
B3
G3-A
C1 A3
F1-A
E3
E1
E2
C3
B3
G3-B
H1
H2
F3
C2
A2
B2
G2-B
A1
B1
G1-C
C1 A3
F1-B
K1
C2
A2
B2
G2-C
D3
D2
C1 A3
F1-C
C3
B3
G3-C
F1

Fig. 52: Plasma dilution concept and technology.

Further interaction of the constituent principle plasmatic magnetic fields of the Matters of the plasma inside the cores of these reactors can create effects that can be used to achieve desired outcomes from their interactions with other Matters or matters and fields outside or inside the reactor's structural environment and this is exactly the way and the method that different entities in the universe function and operate to achieve certain desired effects and production of Materials.

Diluting the plasma in the larger pmtics environment of the reactor is another option in the nuclear world for achieving fission, fusion, production of new Matters, matters, for generation of energy and motion.

Disentanglement

Through the Plasma dilution process, once the new plasma soup is produced in the reactor cores, the weak interacting plasmatic magnetic fields soup interacts with the binding fields of the components of the Matters of the plasmas which are further introduced into the reactor cores, this allows the Matters of new plasma to disentangle themselves from each other (Fig. 53), if this level of dilution of new plasma is required.

In this process of plasma dilution, the sub-components of the plasma are allowed to loosen up from each other and not necessarily allowing the separation of all the Matters of the plasma from each other.

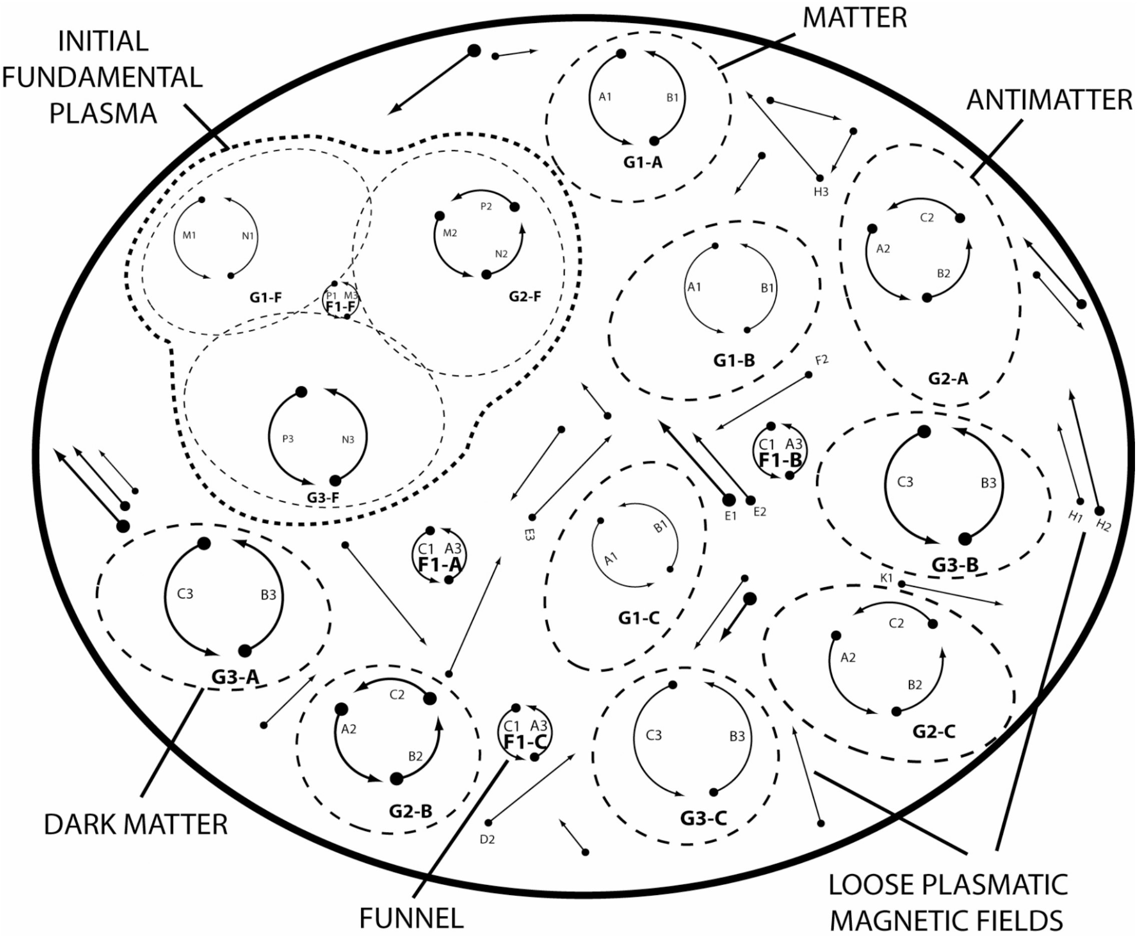
INITIAL FUNDAMENTAL PLASMA
MATTER
ANTIMATTER
A1
B1
G1-A
H3
M1
N1
G1-F
P2
M2
N2
G2-F
P1
M3
F1-F
A1
B1
G1-B
F2
C2
A2
B2
G2-A
P3
N3
G3-F
C1
A3
F1-B
E1
E2
C3
B3
G3-B
H1
H2
E3
A1
B1
G1-C
C1
A3
F1-A
C3
B3
G3-A
K1
C2
A2
B2
G2-C
C2
A2
B2
G2-B
C1
A3
F1-C
C3
B3
G3-C
D2
DARK MATTER
FUNNEL
LOOSE PLASMATIC MAGNETIC FIELDS

Fig. 53: Dilution process of initial fundamental plasma within a soup of Matters.

Through this method, different components of the Matters and fields of the plasma can be used for any purpose that the operator decides or that the system is pre-set to.

Hence, each sub-component Matter of the plasmas in the reactor can be separated from the plasma in any quantities of individual Matter within the soup.

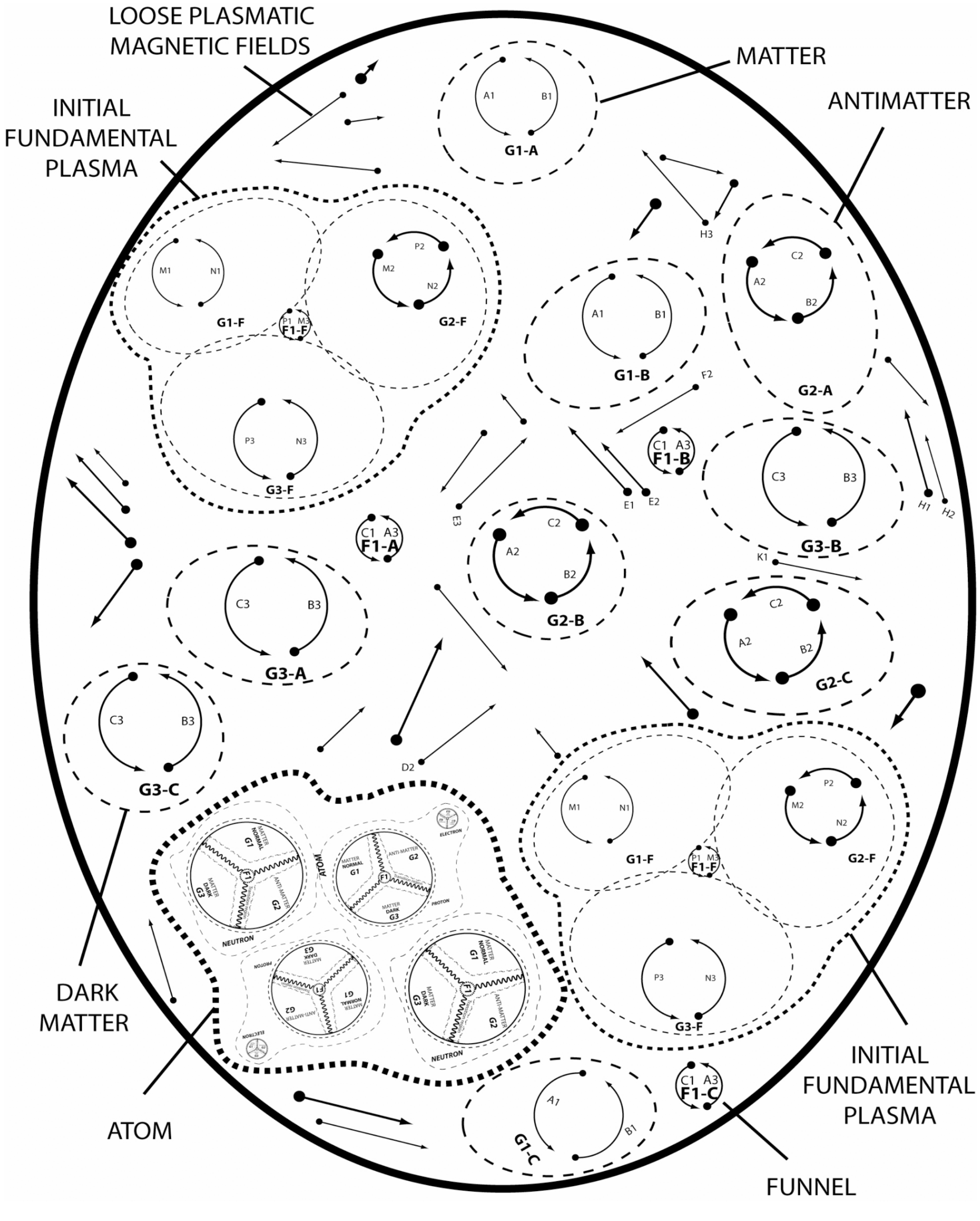
LOOSE PLASMATIC MAGNETIC FIELDS
MATTER
ANTIMATTER
INITIAL FUNDAMENTAL PLASMA
A1
B1
G1-A
H3
M1
N1
G1-F
P1
M3
F1-F
P2
M2
N2
G2-F
A1
B1
G1-B
F2
C2
A2
B2
G2-A
P3
N3
G3-F
C1
A3
F1-B
E1
E2
C3
B3
G3-B
H1
H2
C1
A3
F1-A
E3
C2
A2
B2
G2-B
K1
C3
B3
G3-A
C2
A2
B2
G2-C
C3
B3
G3-C
D2
M1
N1
G1-F
P1
M3
F1-F
P2
M2
N2
G2-F
P3
N3
G3-F
ATOM
NEUTRON
PROTON
ELECTRON
MATTER NORMAL
MATTER DARK
ANTI-MATTER
G1
G2
G3
F1
DARK MATTER
ATOM
A1
B1
G1-C
C1
A3
F1-C
INITIAL FUNDAMENTAL PLASMA
FUNNEL

Fig. 54: Dilution process of initial fundamental plasma and atoms and molecules within a soup of Matters and pmtics.

The plasma dilution technology (Fig. 54) replaces the present state-of-the-art technology *accelerators* used for the separation of different components of plasma or an atom.

The plasma dilution reactors for example can be used for the extraction of Antimatter, instead of the presently used methods of slamming plasma at high speeds in accelerators into specific pure solid metal matters to extract its Antimatter, by chance, every once in while (6, 7 and 8).

In the same way, the plasma dilution reactors can be used in the fusion industry where, present technology uses huge enforcing magnetic fields to force the fusion of plasmas.

With the plasma dilution method, the components of the plasma are opened up, in a soft way, loosened, and then are allowed to be separated from other Matters in their surroundings, and into their specific sub-components, interacting collectively in the soup of the Matters for their use for any given specific purpose (Fig. 54), like for fusion of Matters or matter.

These reactors even can be used for fusing components of two plasmas together. If the Antimatter is needed to be separated from the Matter, this can be easily done, as by creating a weak plasmatic magnetic field, around the plasma, then, the Antimatter components of the plasma can be separated for the use of their plasmatic magnetic fields properties and effects for example for the production of powerful Magravs forces. Where, the rest of plasmas' components are then *left in the* ***dilution soup***.

When one has used these freed Antimatter plasmatic magnetic fields for their effects they create, for the purposes that one needs, then the system-loading Magravs configuration and parameters can be altered to *change the dynamism of the reactor*, and gradually disassemble the dilution fields. Then the Matters of the plasma are allowed to return to their original Matters' components state of the plasma, and for the plasma to return its atomic structure and so forth.

All these reactions can be created and controlled without the creation of unstable or high temperature environments as is seen in the universal order of creation of Matters, energy and motion.

At the same time, there is no need to hold on to certain Matters and to run expensive holding systems for Antimatter as done for example by Fermilab at the present.

All these processes are carried out without any release of radiation or the production of new radioactive waste.

These effects can be created at room temperature and within portable safe nuclear reactors, which have been designed, developed and tested in the past few years.

There is conclusive evidence from tests, that the principle of the untangling of plasmas' magnetic fields binding forces through plasma dilution technology is a proven fundamental principle for the future of energy production and creation of motion for space travel.

That is to say, that we have seen the effects that Antimatter in being loose in the cores of the reactors to possess stronger fields and create stronger Magravs forces than the Matter components of the same plasma. We have managed to lift and reduce weight of reactor systems in respect to its environment, and we have create new matters, all at room temperature, in the recent past through the operation of plasma dilution reactors, as we have reported and applied patents for.

This stronger characteristics and effects of Antimatter are purely due to the fact "that at the inception of the plasma, the Antimatter components of the plasma were created by the stronger plasmatic magnetic fields to start with".

Similarly, gravitational field forces of the Dark Matter components of the plasma can be used through the plasma dilution technology for the creation of high-speed motion without friction for crafts of the future in space technology, or these reactors can be used for the extraction of poison from a substance like water or air.

In testing for the production of the Dark Matter principle, systems have been twisted in their axel and damaged, confirming the enormous power of invisible, powerful, Magravs of the Dark Matter.

In general, it can be said, "The future is **bright**, the future is plasmatic magnetic fields based, through the production of controlled plasmatic dilution technology reactors"

CHAPTER 22

Matters and Magravs systems

The technology that uses the properties of the Matters of plasma to create motion and energy, instead of the present use of matter, I have called ***the Magravs positioning Systems Technology***.

The fundamental difference between the **Matter** gravitational technology and **matter** propulsion and energy systems can be explained in simple way.

matter propulsion technology

The present **matter** propulsion systems in burning fuels have **to fight and to overcome** Magravs of the planet for **lift and motion.** This we see in the rocket propulsion space technology of today.

In the matter systems, like the propulsions systems, power stations and cars use the fundamental principles of ***burning matter*** to release its energy and convert this energy from one form to another to create motion and energy. This technology has always delivered *limited energy*, and due to its bulkiness of mass and weight of the fuel that have to be carried by systems to achieve their objectives, these systems are payload dependent.

The matter system uses the ***Matter part*** *of plasma*, the **weakest** and **least energetic** part of the plasma. By their nature as has been seen in the past, matter systems create different type of waste, like CO_2, nuclear waste, and so on.

Matter gravitational technology

The **Matters** Magravs system uses the *properties of the all Matters* in the plasma to create motion, lift and energy without destroying the characteristic of the Matters, and the system works **within** and through the **Magnetic field** and **gravitational field** of its environment to **attain position and motion,** one can achieve more working ***within*** *than* ***fighting against*** the Magravs of the planet.

In the present state-of-art technologies, always the matter components are considered for their physical ***action and reaction***, and not their Matters field's components.

Matters used in the Magravs systems, uses the properties and interaction of all the initial fundamental particles properties of all Matters within the components of the fundamental plasma, "Matter, Antimatter and Dark Matter". There is no need to "react" as in the matter processes, since the system works' within and is part of the same magnetic fields base system and works through the universal natural forces. Therefore there is never a need for reaction for motion, as the system always magnetically balances itself in respect to other magnetic fields forces.

These reactor systems are developed to function in the same way as the interaction of the initial fundamental particles leads to the creation of Magravs for the motion of Matters through the Magravs positioning principle in the universe.

In these systems, there is no destruction of Matter, since the system can use Matters pmtics properties for the enhancement of Magravs positioning, for the creation of motion and energy and so forth.

This method of the use of Matters creates heat, if it is needed from a system.

As it has been observed, stars create massive heat to create vast amount of energy and Magravs forces, through tests it have been proven that the creation of motion and energy by the use of Magravs forces, these systems creating effects which are ***temperature independent***.

Used in the atmospheric conditions, the Grapos, through the principles of the creation of plasmatic Magravs within the cores of the reactor and their interaction with the plasmatic Magravs of the planet, this interaction of these two Magravs, leads to creation of a *magnetosphere* around the systems. Where, the interaction of plasmatic Magravs of the system and the planets Magravs always lead to ***the creation of light*** around the system. This light being created through the principle of interaction of two plasmatic magnetic fields as was explained in Chapters 7 (Fig. 29 and Fig. 31). This is exactly how the magnetosphere of the Earth is created in respect to the Magravs interaction of the Earth and the Sun, and the same principle as for the creation of the daylight as seen around the Earth.

The magnetosphere created around the Magravs system, can be a spherical or oval bright light in shape depending upon the speed of motion and protection chosen by the operator.

In the **Universal systems** like stars**, their Magravs forces** are *fixed* by their initial matters' contents in their inner cores at the point of the inception of the star, and the magnetic fields strength they generate for the production of their Magravs strength, once the strength of the Magravs is set for the stars, these cannot be changed arbitrarily to achieve different Magravs strengths for the stars.

Whereas in the **Grapos reactors'** developed for the motion of crafts, the **Magravs** parameters of these reactor systems *can be changed and manipulated* according to the position, speed, energy and so forth, as is desired of the system by their operators.

In these systems, the reactor-operator dictates the Magravs strength of the system, therefore he decides the speed and the position of the craft at any position in space in respect to another Magravs system, the like of the Earth surface or a given distance from the moon or galaxies.

Interesting enough, all these effects of motion are achieved even though the system uses the same fundamental methods and Matters as in the cores of the planet or star that the systems is operating within or aiming for.

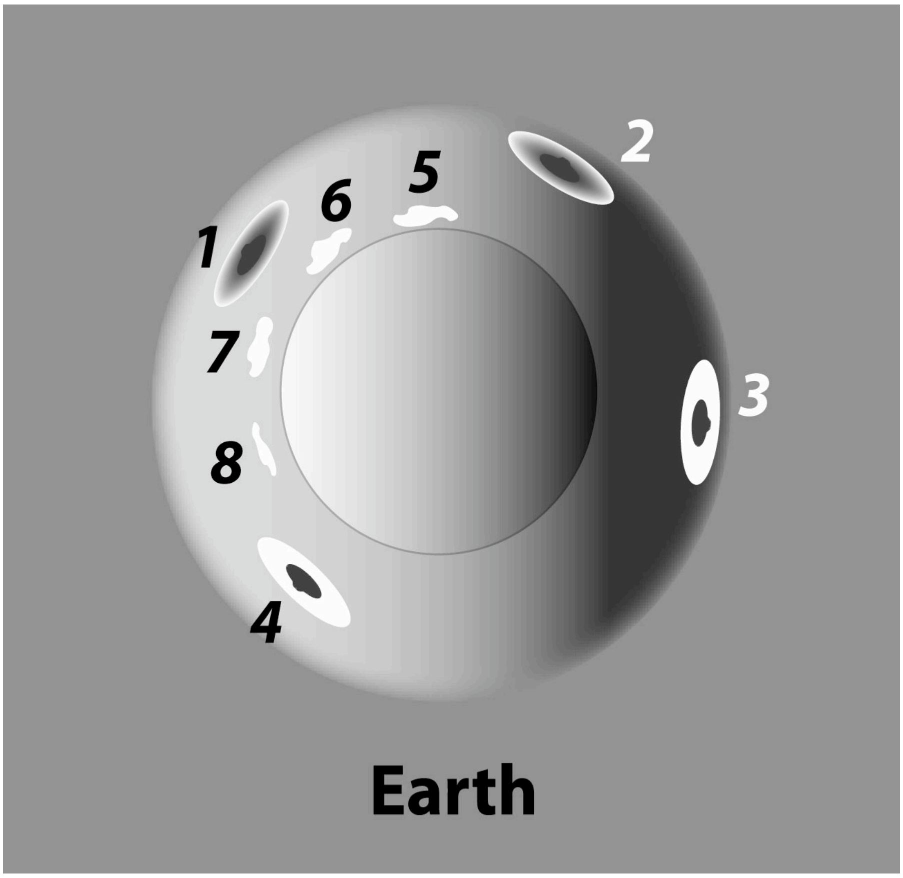

Fig. 55: Representation of interaction of four different pmtics strength of gravitational (Magravs) positioning systems (Grapos) in respect to Earth pmtics Magravs (1, 2, 3 and 4).

Note: Figure 55, numbers 5, 6, 7 and 8 show that the position of clouds in respect to Earth are pmtics Magravs strength dependent, and not size or volume dependent.

Thus, wherever the Magravs principles reactors are used for flight in a system, like in atmospheric condition of the planets', the natural outcome of the interaction of the two pmtics Magravs of the environment and the reactors Magravs is the appearance of and creation of natural light around these flying systems (Fig. 55 light regions of 1, 2, 3 and 4). Thus, the same Magravs positioning systems can be used, to produce lamps that can create light for years without needing to burn any fuel or consuming electricity.

Simultaneously, in a dense plasmatic magnetic fields environment, like the Earth, when the Grapos are used for motion and lift, due to the interaction pmtics Magravs, of the Earth and the system, these systems will produce their own magnetosphere around them. This magnetospheric light intensity is reactor Magravs' ***strength*** *dependent* and through the principle of interaction of two pmtics of plasmas as was discussed before, these magnetospheres will cause the creation of bright light around these systems in the

Earth atmospheric condition. As the brightness of the magnetosphere around the system in respect to its surrounding pmtics strength will be dependent on the strength of the plasmatic magnetic fields that are created by the Magravs positioning reactors, and in respect to the systems' position, considering that the plasmatic magnetic fields of the Earth is constant.

Thus, the confirmation of systems using Magravs positioning technology for motion in dense atmospheric condition like the Earth will be the appearance of bright moving lights (Fig. 55 lights 3 and 4) in the Earth atmosphere or any given atmosphere or medium in the universe. In dense atmospheric conditions, these Magravs systems will be appear to be a shining bright light floating in air, either with high speeds or seated stationary.

In some literatures, systems that use Grapos for motion and appear as bright lights from an unknown source in the skies have been called, the Unidentified Flying Objects, also called ***UFO's*** (Fig. 55 lights 1, 2, 3 and 4). Now that the method by which these bright lights in the sky are produced, their operation understood, and known, these crafts are not Unidentified Flying Objects no longer.

Nevertheless, in reality, it is now interesting to see who are the users of these systems. Whoever are the users' of these Magravs systems, this does not necessarily means that they are more advanced in scientific level than the human race in its present scientific and technological evolution. However, this simply means that they have managed to harvest the principles of the plasmatic magnetic fields, and the constituents of Matters, some time before the human race.

Now that the technology and the principles behind the plasmatic Magravs positioning fields are open to the human race, it is not hard to see that interaction with users of these bright lights will be inevitable and will only be to the Mans advantage, to help man to reach new levels of understanding about his universe.

In tests, the effects and bright light created by use of these reactor systems, which have been developed and tested for weight reduction and lift, has been noted.

The operation and interference with the radar and radio frequencies, through tests running of these Magravs positioning reactors have been videoed and recorded in August 2008. This confirms the origin of the motion of these systems to be on a plasmatic magnetic fields basis, as the system's Magnetic field operation interferes with radio waves, which radio waves are magnetic fields based entities.

NASA has done tests on these types of magnetospheric condition and motion system, on a miniature scale, in advance concept projects known as MINI-magnetospheric plasma propulsion project called M2P2 (40). They have achieved this by using the argon gas up to this stage. They hope to be able to finance the project to reach helium and then the use of hydrogen as base-material in years to come. They may then use plasma of hydrogen atoms later as they progress in gaining knowledge in 30 to 50 years and having the budget. However this is still one of the theoretical approaches for NASA.

Fermilab (6, 7, and 8) is looking for the Antimatter as the “fuel” for propulsion systems of the future space programs for NASA. If they redirect their attention into the properties and the effects of the Antimatters that they have gathered by now, they can achieve the same results as have been seen and tested in the recent Magravs systems reported on our Foundation web site and patents that have been applied for.

Through the understanding of the real forces in operation of the universe, the Magravs positioning reactors are simply years ahead of their time.

Our Magravs positioning reactors are not in a mini-scale dimension as NASA’s project has tested and envisages at this moment in time, but these are in several sizes, which can create controllable size magnetospheric conditions, as the operator of these simple systems sees fit the purpose.

In the near future, these Magravs systems will replace the propulsion systems that have been used for space explorations and air travel.

The advantage with the Matters powered Grapos is that these systems create a magnetosphere around the reactor system. This is the same natural magnetic protection barrier for the future spacecrafts as the magnetosphere around the Earth has been, protecting man from space debris and cosmic radiations for billions of years.

These systems’ magnetosphere protection capabilities can be used to protect the craft in deep oceans from water and the crafts to be used like the submarines of today, with the difference that these systems magnetosphere will allow these crafts to reach the deepest parts of oceans floors without any problem, as they can reach any depth in liquid planets in the universe.

The magnetosphere generated by Grapos will create *magnetic shielding* at Matters’ level of the plasmas that can be used to pass through materials like ice without melting their matter(s). But simply by diluting the ice Matter(s) at plasma levels as in the plasma dilution reactors. In these conditions the crafts will appears as the craft will come out of nowhere and with no holes indicating the position of the exit from the ice by the spacecraft. If observed from the distance these crafts will be like bright lights shining and jumping out from the ice-covered sea-surface.

These plasma dilution Magravs positioning systems, by the use of Matters, will be the only solution for plasmatic saturation, once these crafts will pass and go through cosmic dusts as they temporarily disentangle the Magravs of atoms and plasmas of these cosmic dusts by just diluting the components of the Magravs of the Matters of cosmic dusts with the Magravs of the craft protection magnetosphere, and the craft passes through the dust matters as if they were not there.

CHAPTER 23

Matter Pmtics in reactors for lift and motion

The utilisation of the Magravs forces of dynamic multi-core systems, in respect to each other, for the purpose of the creation of lift and motion is discussed in this chapter.

In multiple cores *dynamic plasmatic dilution Grapos* that have been developed and tested at room temperature, these systems have proven to have the capability to produce different plasmatic Magravs forces strength as strong as those of a hot centre core object, like the Earth.

Using the principle of the creation of dynamic plasmatic Magravs forces within the structure of a multi-core reactor, the system can be caused to find its own position in respect to the environment of a planet or any Magnetic field and gravitational field forces in space.

In developing and achieving the Magravs positioning of these systems, in respect to the planet's Magravs forces, on the surface of the planet, leads to the reduction of the weight of the system to start with. Once the two plasmatic magnetic fields of the planet and the system reach **the total balance or equilibrium**, the ***zero weight*** of the system, in respect to the planets gravitational fields at a given point and the position, is achieved.

From this point of ***total equilibrium*** any further change(s) in Magravs force strength of the reactor in the same direction and configuration as before, this leading to and causing the ***free lift*** and **motion** of the reactor system, in respect to the planet.

Principle concept of upward or separation motion

The upward motion of the systems in respect to the planetary fields or any other Magravs systems is simply due to the fact that the **Magnetic field of the planet** and the Magnetic field of the system is of the *same polarity*. Thus, the two similar Magnetic polarities ***repel*** each other, hence causing the upward or separation motion of the system.

It is fundamental to understand that: In the Magravs positioning reactors **"The repulsion of two objects or moving away from each other, like moving away form the surface of the Earth, is totally due to interaction of the two pmtics of Magnetic fields forces of the systems and of the planet.**

Whereas**, any closeness or gravitation of the reactor system in respect to the planet is controlled by the pmtics of the gravitational fields between the system and the planet in respect to each other."**

This is how Magravs positioning reactors, in possession of dynamic cores, create up or down (lift and descend) motion in respect to another Magravs of an object, like the Earth.

In these systems by changing the strength of any of the gravitational or Magnetic fields strength within the reactors of a system, this system, necessitates the finding of a new balance position between the system and the planet's Magnetic and gravitational fields.

By changing the parameters of the reactors and for the reactor to find a new balance between the systems' own Magravs and the planetary Magravs, this causes the motion of the system as it tries to reach the newly desired balanced Magravs force position, in respect to the planet Magravs.

This is the universal method of **creation of motion** ***without burning an atom*** of any matter, and this is exactly how all the celestial objects find their position and motion in the ever-dynamic environment of plasma of an electron in an atom, solar systems, galaxy and universes.

The motion of the system is purely due to the phenomenon of **Magravs positioning**, of one system in respect to the other(s).

Simple, portable, plasma dilution nuclear reactor systems, can produce the required Magravs forces within its reactor cores on a continuous repositioning criterion bases that can cause the motion of the Magravs positioning system (Fig. 56).

In these Grapos, the individual pmtics released in the central core (Fig. 56) in their interaction with the compatible and matching pmtics' strength of other pmtics *in the same core* can then collectively interact with the same collection of pmtics created in the other cores of the reactor (Fig. 56), and interactions and interlocking of the same pmtics strength in *different cores* of the system leads to the creation of Magravs within and around these types of reactors simultaneously.

By referring to figure 1, 2 and 3 in the early chapters, then it can be realised that, what has been achieved in these dilution plasmatic Magravs positioning reactors (Fig. 56), is exactly what has been going on in inner cores of planetary systems such as the Earth, to create their Magravs for them to hold on to their components through their gravitational fields, and at the same time to protect their matters from other celestial objects through their Magnetic fields, and at the same time create positioning in respect to other Magravs, at the same leading to creation of motion of the reactor. These reactors are truly an all ***integrated system for flight and motion.***

At the same time, these reactors during their operation within a craft system creates' and maintains' an independent internal gravitational field forces within the reactor cores fields catchments' zone. For example these properties and effects can be used to maintain an earth, one-G gravitational field force within the confine of any spacecraft, independent of the position and speed of the craft where, this position in the universe might be.

These reactors in collaboration and simultaneous production of the Magnetic fields and gravitational fields leads to the creation of a distinctive and separate magnetosphere around the reactor or the craft which, uses such reactor technology, similar to the Earth atmospheric magnetic field.

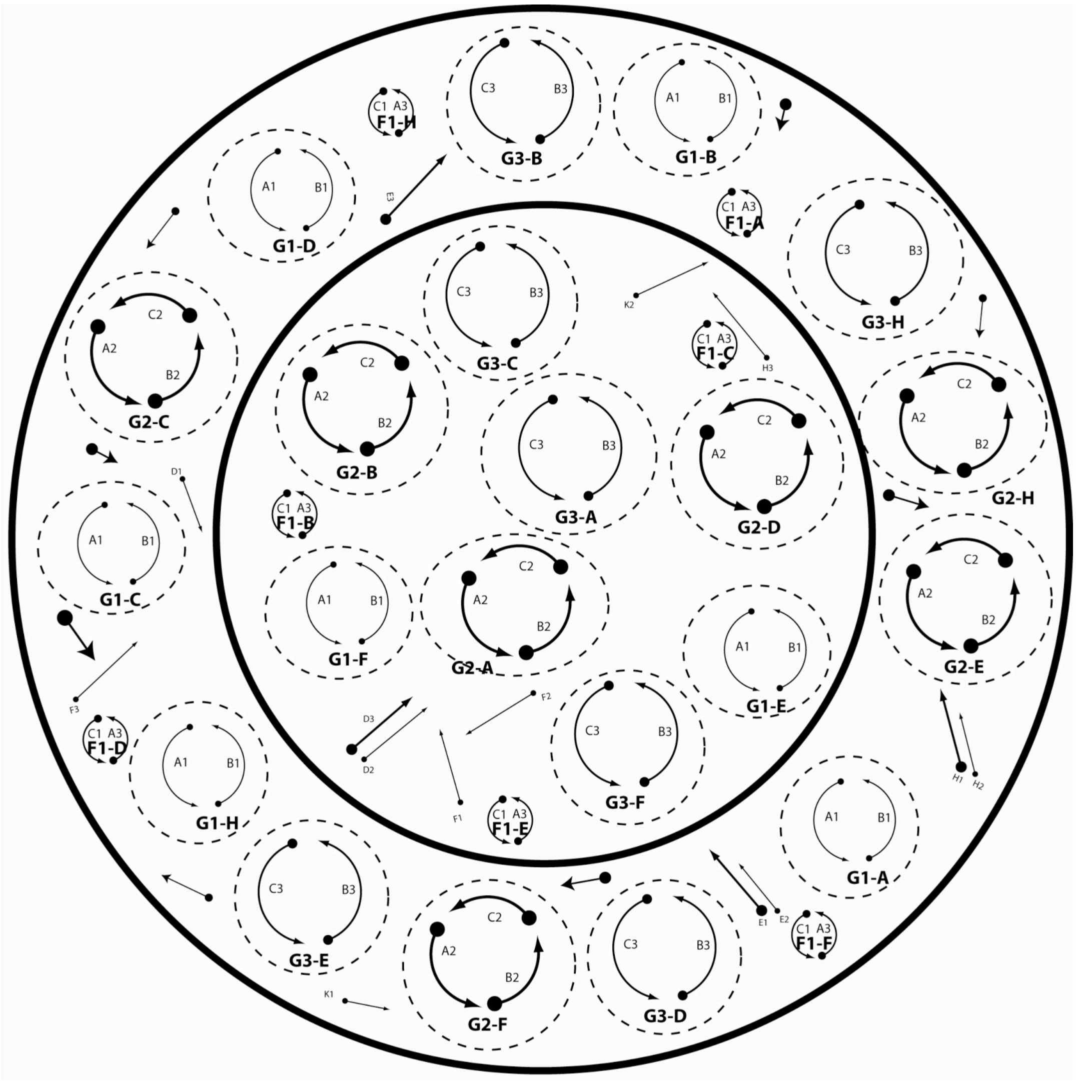
C1 A3
F1-H
C3 B3
G3-B
A1 B1
G1-B
A1 B1
G1-D
E3
C1 A3
F1-A
C3 B3
G3-H
C2 A2 B2
G2-C
K2
C3 B3
G3-C
C1 A3
F1-C
H3
C2 A2 B2
G2-B
C2 A2 B2
G2-H
D1
C3 B3
G3-A
C2 A2 B2
G2-D
C1 A3
F1-B
A1 B1
G1-C
A1 B1
G1-F
C2 A2 B2
G2-A
C2 A2 B2
G2-E
A1 B1
G1-E
F3
C1 A3
F1-D
A1 B1
G1-H
D3
D2
F2
F1
C1 A3
F1-E
C3 B3
G3-F
H1
H2
A1 B1
G1-A
C3 B3
G3-E
C2 A2 B2
G2-F
C3 B3
G3-D
E1
E2
C1 A3
F1-F
K1

Fig. 56: The Keshe Model of Plasma Dilution and Magravs positioning reactor.

Thus, the Grapos are exact copies of the planetary Magravs producing systems. Similarly, they utilise the same Matters inside them, and the same principle of interaction to create Magravs between the systems and a planet. This being the same principle as, between the Earth and the Sun field forces and the Earth, and the rest of planets in the solar system and the same system principle as between electrons and protons. (Fig. 57)

Therefore, by changing the strength of the Magravs strength of Grapos, these systems can be positioned in any orbit or position around the Earth (Fig. 55, Fig. 56 and Fig. 57 c), in the solar systems or galaxies, with the use of Matters, rather than matters in the system.

In other words, the balancing of two Magravs' forces of the system, in respect to the planetary Magravs forces, leads to lift or descend of the systems. This separation only leads to a vertical up or down motion of one-system field in respect to the other systems' field. This means lift or descend of the system in a craft takes place vertically and not a 360-degree directional motion of the system.

In current physics, scientists call this the anti-gravitational field forces needed for lift and these systems are called anti-gravitational systems.

In reality, any motion of a system is not due to anti-gravitational field forces, but in fact is due to the total balancing of the gravitational field and Magnetic fields of the two systems of the planet and the Magravs positioning of a reactor, in respect to each other in a given position.

For directional motion, the same principle as vertical lift is used with the difference that in the directional case, the ratio of change between Magnetic fields and gravitational field created within the system are utilised, as the two field forces in the reactor of such a system have an opposite gradient of scale of strength with each other. That is to say, the control of the creation of gravitational fields increases with the same increment as the decrease in Magnetic field strength, within the interaction of the fields of the cores of the system, in respect to the planetary field forces.

The same principle of Magravs positioning can be used in space for motion, in a direction of motion towards or away from a planet, star or galaxy.

There is an alternative *method of motion*, which is by *compartmenting* each core of the reactor. This is a highly advanced technology and needs a full understanding of the creation and control of the magnetic fields' strength of magnetic rays and Matters mixtures and full control of the motion of dynamic Matters in a dynamic environment of the reactor cores.

As motioned before, reactors have been designed, built, and tested to confirm the effects of weight reduction and lift, and the results have proven the correctness of the principle of Magravs positioning reactors, rather than the anti-gravitational field assumptions.

It has been shown that the **system gravitational field and Magnetic field strength** are **independent** of the **size of the system for the mass that** they can lift.

In tests carried-out in 2008, a ten-kilogram reactor system was used to lift a 100-kilogram weight.

In early 2009, the system combination was used to reduce weight from 9 kg by increments of 100 grams. Then another system was use to reduce the weight of the system from 7 kg to 6.5, 6.3 and 6.1 kg and so on in increments of 100 grams. In some tests, the weight of the same system was increased by 100-gram increments also.

These tests were done to show the correctness of the theoretical principles and its practical applications in reduction or increase in Magravs forces of one system in respect to another like the planetary system.

CHAPTER 24

Passive Magnetic Fields

In the universe, there are two types of interaction in respect to plasmatic magnetic fields forces.

One pmtics force is called ***active*** magnetic fields force and the other ***passive*** magnetic fields force.

The *active magnetic fields forces* are the pmtics forces that can exert forces on other fields and components of other pmtics of plasmas causing them to move or vibrate, or receive motion or vibration from the pmtics. This is usually matter-based motion and creates tangible or physical motion of matter.

Active magnetic fields forces usually cause dislocation of the electrons in a cell or vibration of electron in conductive material, etc.

Matters of the plasma mainly produce the **passive** plasmatic magnetic fields.

Passive pmtics can deliver their magnetic fields forces as energies to another plasma of the atom, for them to increase their magnetic fields density or energy without dislocating any Matters plasma's.

Through passive magnetic fields strength, if the plasma is in need of the extra pmtics strength, it can receive the quantity that it needs *of the **same** pmtics strength* for it to return to its ground level pmtics strength level. Otherwise, the pmtics will pass the plasma without transferring any of its pmtics to the second plasma. At the same time the second plasma, if it has extra pmtics which it does not need, it gives the extra pmtics *to the **passing** pmtics,* if it needs the extra pmtics for both pmtics to return to ground levels of their original pmtics strength.

The passive magnetic fields do not cause the atom to move or be dislocated from its position but also cannot cause vibration of the electron of the atom as such.

Passive magnetic fields are pmtics used for delivering the energy levels of vitamins and minerals to atoms and molecules of protein chains and are ***Matter pmtics orientated*** and based.

It is important to note that each plasma carries' one pmtics strength pack in the majority of cases. Thus for human body to receive all its pmtics strength spectrum, the body need to consume different minerals and vitamins and proteins that can deliver certain specific passive plasmatic magnetic fields strength to cover all his cells' needs.

Where the ***active pmtics*** are **matter orientated** and create heat, motion, and vibration in the matter components of the plasma.

CHAPTER 25

New understandings and New Technologies

From the new explanation and understanding of the creation and control of initial fundamental particles (magnetic fields), and further in having access to the new knowledge and the technologies that to be able to replicate the same in small reactors and their systems, the production and the control of MATTERS and the initial fundamental plasma as in **the universal order of the creation of Matters** in these reactors have become simple and practically possible.

The world of science and technology can find new solutions and new sources of Matters for the production of energy, the creation of motion and other applications within the confines of these reactors and this technology will be developed further for these purposes.

In understanding the principles of the working of these reactors, one can make use of, and manipulate, the Matters and energies embedded in the initial fundamental plasmas for the production of new materials, motion and energy. One would be able to use the new understandings of the Matters sub-plasmatic structure for the release of magnetic fields and energies from these Matters.

Through progressive development of this technology, new and simpler methods are devised and developed further for the fission and fusion of Matters and their relevant technologies. Through these new understandings and developments of reactors, new ways and opportunities for the production of energy in the world of nuclear physics has become possible.

For energy production in nuclear fission of today, the state-of-the-art technologies use the ***collision effect*** of neutrons and atoms to break up an atomic structure and release its energy. On the other hand, in the world of nuclear fusion, the reverse methods are in use. Scientists in nuclear and physics laboratories use as ***much force*** as possible to fuse plasmas, and harvest the residual released energies from this process.

Through our technology new and simple reactors are in the reach of the scientific world for the disintegration of an atom, or for the fusing of plasmas through plasma dilution technology. This new approach is less energy consuming, much simpler, more practical and the resulting energies produced are easy to harvest.

The same is seen in the world of aircraft and space industry where, jet engines and propulsion systems are in use to force and displace matters like air to create lift and motion.

With this new knowledge and approach to plasma dilution technology as one integrated system, much simpler and more effective methods based on the universal methods for creation of motion are made available to the aircraft and space industry to move crafts within atmospheric or outer space (Fig. 57 C). Magravs positioning principle of an electron in respect to its proton (Fig. 57 A) and Magravs positioning of the planet Earth in respect to Magravs system of its Sun (Fig. 57 B), these shows similarities between the principle of the use of the new Magravs positioning reactors and their universal parallel systems.

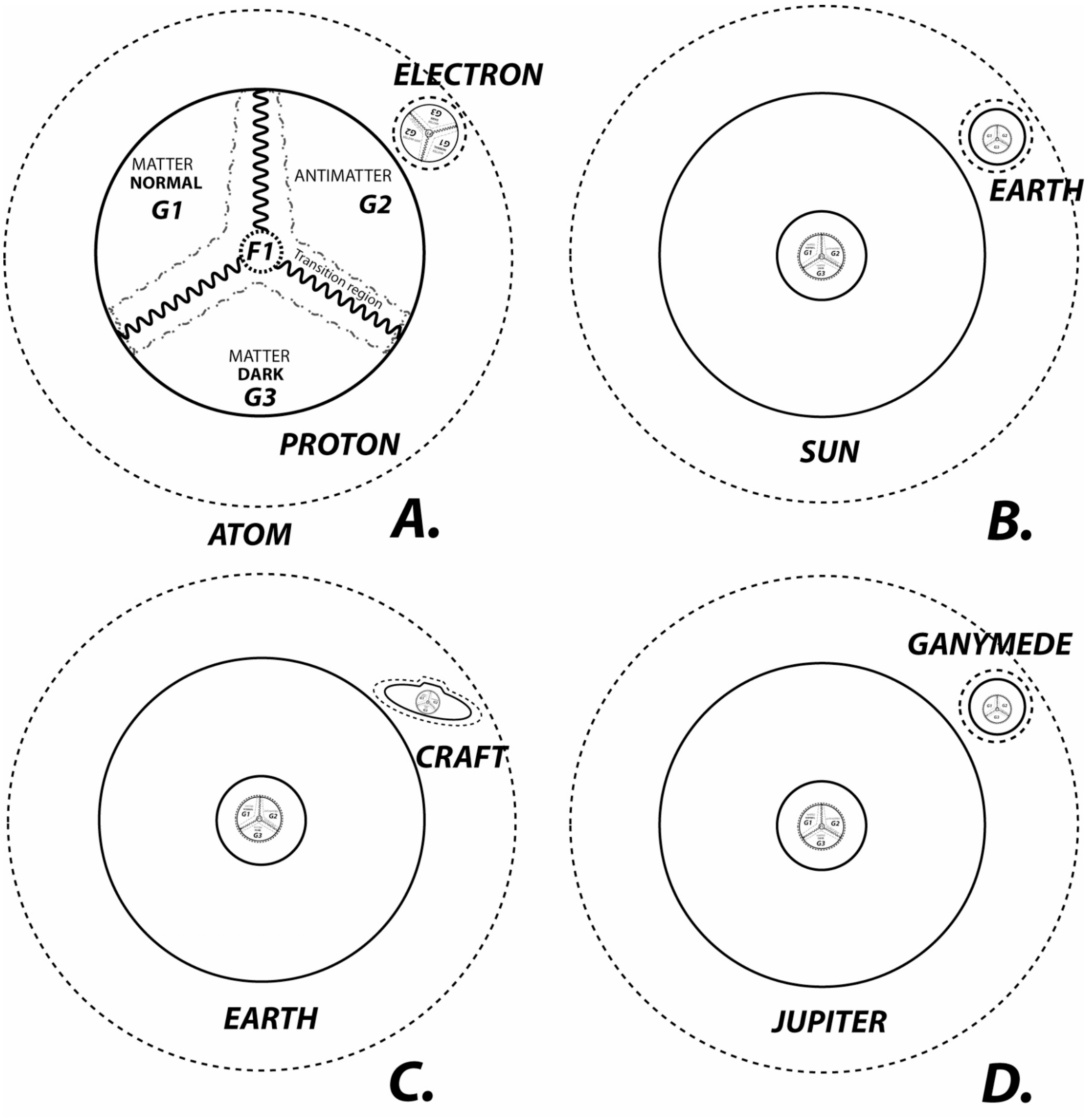
ELECTRON
MATTER
NORMAL
G1
ANTIMATTER
G2
F1
Transition region
MATTER
DARK
G3
PROTON
ATOM
A.
EARTH
SUN
B.
CRAFT
EARTH
C.
GANYMEDE
JUPITER
D.

Fig. 57: Similarities between four different Magravs positioning principles and systems for motion and position.

Note: Figure 57 D. shows that an independent plasmatic Magravs system can exist within another Magravs system, and have Magravs positioning in respect to each other, as the ***Ganymede****, a moon of the planet Jupiter, has been functioning for millions of years within Jupiter's Magravs. To cosmologists the case of the Ganymede is the only known condition of an inter-gravitational planetary system positioning, and its operation within the Jupiter's magnetosphere could never be explained with present scientific knowledge. Similarly, Grapos reactors can be used in crafts for motion and Magravs positioning within the magnetosphere of the planet Earth, as shown in figure 57 C using the same fundamental principle as the Ganymede.*

Through the fundamentals of this new understanding of the creation and control of plasmatic magnetic fields reactors (Grapos), **craft systems** *do not need to burn Matmags* to create **lift and motion, but where, the reactor uses the Magnetic fields strength of the Matter (the Matmags) to create gravitational positioning.**

In the **Magravs positioning Technology reactors**, by the simple creation of the acquired plasmatic Magravs levels of the desired point in space, within the reactor's systems, the system is forced by its' internally produced Magravs, the need *to reach a new plasmatic Magravs balance* position between the system and the planet Magravs forces', this causing the motion of the system in respect to the Earth **without burning any fuel.**

The Magravs positioning technology once used in a craft, unusually these crafts become weight independent. That is to say, the systems payload compared to the power of the gravitational positioning forces of these reactors are so insignificant, that the payload mass becomes irrelevant.

This is how small cores size of matters in comparison, in the centre of the planet Earth, can create such strong plasmatic magnetic fields, that their interactions can keep and cause continuous motion of the heavy loads of matters of the planet, at ease, for billions of years.

The important point about using Grapos in aircraft systems is that the plasmatic Magravs and Magnetosphere fields' catchments area of the system has to pass the physical boundary of the craft itself during motion, in-flight and stationary floating mode or protection mode.

During the operation of plasmatic Grapos, magnetospheres are created around the system interface between system and the planet. At the interface of the magnetosphere of the system, due to the friction between two plasmatic Magravs of the system and the planet, only at the point of the magnetospheric interface of the system, this leads to the creation of visible light (Chapter 7).

In respect to plasmatic gravitational field systems, once the fields are allowed to pass beyond the physical boundary of the craft, this gives the craft being in possession and the advantage of having a dynamic magnetically based magnetospheric protection environment around the system. This magnetic magnetospheric zone around these crafts operates and functions exactly and similar to magnetosphere possessed by the Earth, which, the Earth has generated around itself by the interaction of its own plasmatic gravitational and Magnetic fields with that of its star and the pmtics around its environment in space.

This production of plasmatic magnetospheric shielding is a *natural protection* system offered by all celestial objects in possession of dynamic plasmatic Magravs for the entrapped matters within their systems.

Thus, by the use of Grapos in crafts nothing less than that is natural in the universe and a normal phenomenon for protection of internal matters and structure of the planet are offered.

Using this magnetospheric protection capability of Grapos, crafts of future will travel at speeds beyond the capability of the present technology in the aircraft and space industry. These crafts will travel without their outer-surface ever getting heated-up, or ever being in a position to be hit for example by birds in Earths atmospheric conditions or in space by space debris. The reason for this is due to the fact that the outer surfaces of these crafts *never come to be in contact with other matters*, except its own internally created plasmatic Magravs.

The ***interface*** between the crafts' dynamic Magravs and planets plasmatic dynamic Magravs can be extended through the design and control of the Matters within the reactor cores of these systems, that their interface to be far enough away from the physical body of the craft, that there will be no contact or interface with other Matters and matters of physical body of the craft system. These matters outside the boundary of the system can be Dark Matter, air, liquid or others. These Magravs are needed for the protection of a system like this in space, as the systems cannot be hit or damaged by any cosmic dust or asteroid in the path of motion of the system.

The magnetosphere created around these systems, due to the dynamic characteristics of their nature, create plasmatic magnetic fields that will act and is a magnetic shield for the embedded craft within it. These dynamic magnetospheres around these crafts are naturally watertight systems.

If the plasmatic magnetic fields generated by the system, do not reach and pass the physical outer boundaries of the craft, these crafts can have the gravitational shearing gradient forces fields effects, enforced on it, half-way within its structure, rather than around its outer physical boundary; which partial structural cover of Magravs on the middle of the structure of the craft will end-up in the physical damage or in extreme cases the breaking-off of parts or the whole section of the craft at the point of interface between the system and the planet pmtics Magravs. This shearing-off zone is possible and can happen purely due to a difference in the two Magravs forces strength levels at the magnetospheric boundary of the reactor system in the middle of the structure of the craft.

At the same time the passengers of the future, using the Magravs positioning systems, will travel in gravitational field forces within the boundary structure of the craft in similar conditions as on the Earth. The passengers of these crafts can travel without the need for the present compression systems, in use, in the state-of-the-art aircraft industry. There will be no G-force changes inside these crafts during their flight due to the sudden changes in directional motion or at high-speed capabilities of these crafts. Using Grapos, in spacecrafts, this allows normal life and motion within the craft without the effect of weightlessness as is experienced by the space travels of today.

These crafts through their natural plasmatic magnetic fields interaction will always be rotating, as all celestial objects in possession of Magravs positioning field force do. This rotational effect can be facilitated through the design of the craft that the rotational motion can be transferred and be connected to the outer physical boundaries of these crafts, without affecting the working of the internal structure or passengers of these types of crafts.

This method of travel using plasmatic Magravs protection for motion for the future space systems is to be called the **M***agnetically* **O***riginated* **J***oint* **H***abitation* **A***nd* **N***utrition* (MOJHAN) system. The passengers of this **M***an* **O***riginated* **Z***one* **H***abitation* **A***nd* **N***utrition* to be known as "MOZHAN".

These new names unify the human race by the use of the name of the technology and the method of its travel, through the Men's collective effort to reach this point of scientific achievement over thousands of years, through their progressive evolution in science and technology. This takes away the national dependent names, like Cosmo- or Astro- and the rest from the unique and individual nations, for the collective effort and contribution by all of the inhabitants of the planet in different measures over the times to reach this point in Mans scientific evolution.

The use of Grapos will make national boundaries irrelevant except for tax and local laws. The Grapos will not need airports to land near too. These systems know no national or planetary boundaries and they have no need for fuel as such.

Thus, in the near future the nations will have to devise new ways of dealing with the Mozhans as they can land anywhere, without a need for a specific plot or landing strip. In reality however, these crafts will never even need to touch the ground.

Nevertheless, these systems can reach ground level, when their total reactor Magravs sizes are reduced to a safe planetary level, within a specific spherical zones cavity, allocated just out-side the boundary of the reactor systems but within the craft environment.

Pmtics reactors for production of energy

The creation of energy in the universe and utilisation of Grapos technology for production of energy in essence are based on the same principles.

In the universal order of production of energy like in stars, in these units, the principle of the production of current and energy is not on the base of electrons and their vibration movement. In fact, current and voltages on massive scales in these celestial objects are produce through their only ingredients, which are the plasmas and their creation of the pmtics of Magravs of their Matters within their inner natural cores.

In current state-of-the-art power production units like power stations generators, the rotational motion of magnets, and the magnetic fields forces within the confines of the copper windings, is used to create a continuous vibration of the electrons in the solid matters of the copper wire, to generate a flow of current.

We consider the flow of pmtics in materials like that of the ***wind flow of pmtics*** or what is commonly called the ***current** flowing through matters* like in copper wires. Where the plasmatic magnetic fields flow in matters like copper, the strength of the magnetic wind decides the rate of the vibration of the electrons of the atoms of the copper.

In understanding the new concept of the internal structure of an atom there are two fundamental points to consider.

First, the electron is the smaller and the weaker of the two components of an atom. Secondly, **the proton by order of mass is much heavier** and is made itself of greater amounts of the ***same*** plasmatic magnetic fields as **the electron**.

Thus in utilising the power of the plasma of the proton rather than the plasma of the electron, it becomes easier to produce the same levels of power with the use of a fraction of the material compared with the present power units.

This is to say; the new dilution plasma technology will bring into to play the ***use of magnetic characteristic of Matters of the plasma of the proton*** rather *than the vibration of the electron.*

To be able to separate the Matters of plasma of proton within the plasma dilution reactors, one needs to release enough dynamic plasmatic magnetic fields, in motion (energy), in the field strength level of the plasma of proton. These plasmatic magnetic fields can move the plasma of the matters of atoms with profound effects, rather than the electron of an atom.

This new capability and control of magnetic fields strength that can be released by the plasmatic reactors, brings into play totally new methods for energy production for any purpose.

The important point to note is that the production of energy through this method does not mean a change in the common use of electrical equipments. The beauty of this method is that the generators of the future will not be limited in their power production capabilities because of the material and size of the system. This means that even pmtics of plastics, nanolayers or any sort of matters can be used for current production and flow of current.

This changes the present power production capability, and gives the possibility of any generator’s construction or any power levels at the point-of-demand, this being nano-amps or megawatts of power, through the operation of the same reactor, simultaneously and separately.

The other advantage is that the plasmatic magnetic fields strength of these reactors can be changed through the operation of the system. This capability of the plasma reactors gives a new insight into the power production capability of these new generators using the present equipments and tool.

New type of Generators

Through the control and power of the plasmatic magnetic fields strength **DY***namic* **P***lasmatic* **M***agnetic* **F***lux* **S***trength*, abbreviated to: “**DYPMFS”** that can be released by these new plasmatic dilution reactors, future electric power generators only need to change their Dypmfs, or the dimensions of the **VO***lume of the* **MA***gnetospheric* **F***ield* **S***trength*, abbreviated to "**VOMAFS**" of the reactor, to envelop the generators' winding using the coil winding method for power production.

Where, the Vomafs of the reactor can cover more of the same copper winding, to produce larger amounts of currents or the strength of Dypmfs of the reactor can be increased or decreased to change the levels of the current out-put of the same winding.

With these new types of plasma magnetic fields system there is no need for transformers, since the exact power input to the motor or engine can be produced by the plasmatic magnetic fields release at the point-of-demand.

For example: considering the present Power Station generating units (Fig. 58), it can be seen that, what has been used as fuel (coal, oil, gas, nuclear) to boil water, to generate steam, to turn the turbine and for the turbine to cause the rotation of the rotor, which is mounted with fixed flux strength magnets, for the motion of these magnetic fields cutting with the matter of the copper coil windings, to generate power, for this power to feed into a power grid.

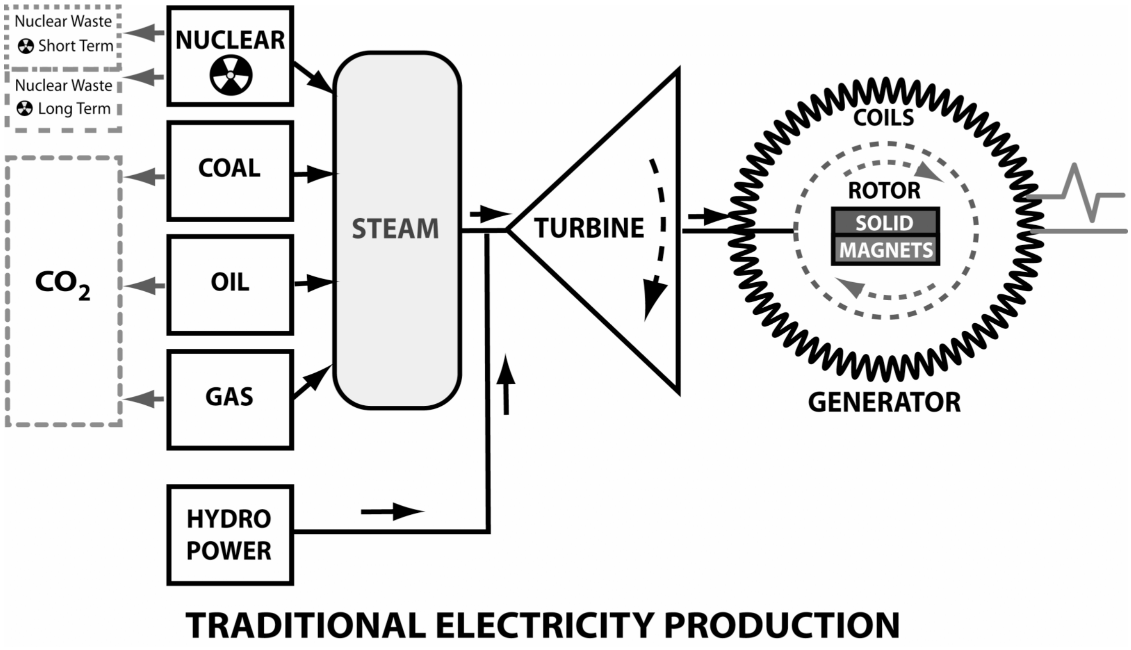

Fig. 58: Schema of actual electricity production

With the present power station generating units *one losses a vast amount of energy through each step of conversion* from solid or liquid or gas to steam to rotation, by all sorts of channels like different types of frictions, heat lost, losses through the grid. These systems usually need hundreds of kilometres of **transmission lines** to transfer the generated powers to the point-of-demand usually into cities. At the same time these types of generating systems using matter components of plasma, these always produce wastes like CO_2 or nuclear wastes.

In understanding the principle of the rotation of the Earth, through the interaction of two magnetic fields (Fig. 3) and the capability of the planet to create simultaneously the external dynamic Magnetic field (Fig. 1, similarly Fig. 61), with the new knowledge of the use of the plasma dilution Magravs positioning reactors with the ability to produce **a controllable, continuous, dynamic rotational flexible Magnetic field strength**, within the cores of such reactors, it is now easy to construct and produced new strong and powerful generators that do not need to burn any fuel (Fig. 59), to create rotating Magnetic fields, but, just by using Grapos nuclear reactor one can achieve the same and even more power than the present matter-based frictional systems. Even it becomes possible to produce varying frequencies of current output of the system from the same generator.

Grapos generators can produce power from kilowatts to megawatts in seconds, by just increasing the strength of the Magravs for the **Dypmfs** of the reactor or by increasing the **Vomafs** coverage of the Magravs. We have called this new generation of power generators the ***Keshe reactor generators*** (Fig. 60).

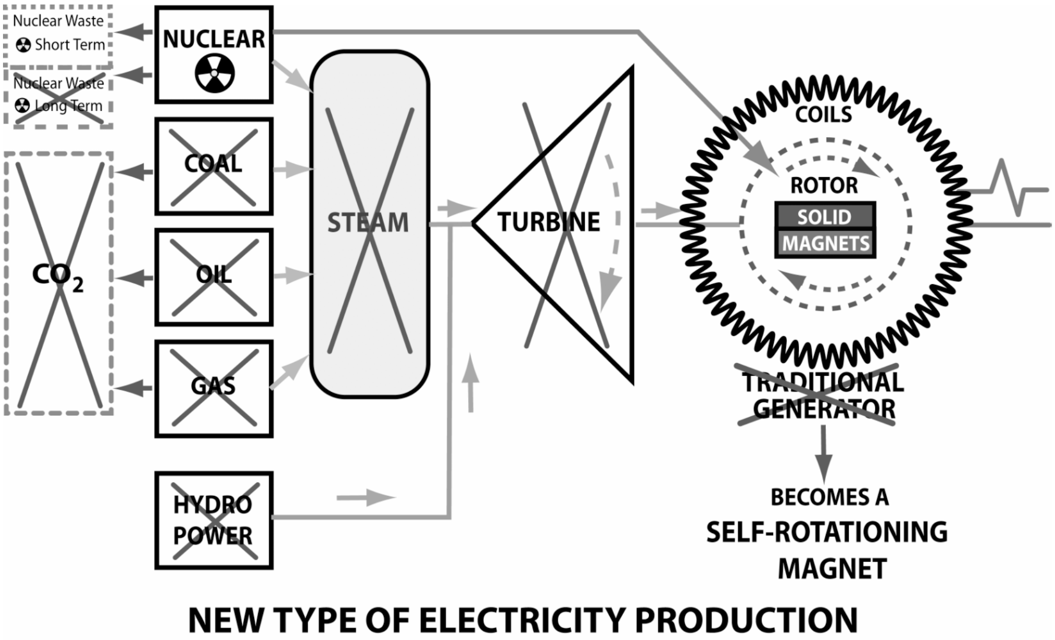
Nuclear Waste
Short Term
Nuclear Waste
Long Term
NUCLEAR
COAL
OIL
GAS
CO₂
HYDRO POWER
STEAM
TURBINE
COILS
ROTOR
SOLID MAGNETS
TRADITIONAL GENERATOR
BECOMES A
SELF-ROTATIONING MAGNET
NEW TYPE OF ELECTRICITY PRODUCTION

Fig. 59: New approach to generate electricity

The use of plasma dilution Magravs positioning reactor brings to an end the need for burning of fuel (Fig. 59), to create rotational motion of the rotors' for production of electricity as it happens in the present generators.

Our new approach puts an end to the production of greenhouse gasses (like CO_2) through the burning of fuels (coal, oil, gas) and finally an end to the habitual burning of matters to generate motion and energy. This also bringing to an end the current production of dangerous nuclear waste by running the present nuclear power stations.

By purely using the effects of interaction of the Magravs of the Matters, within the cores of these reactors, motion and Magnetic fields, for production of any amount of electric power can be achieved in one all-integrated nuclear-based unit without creating any waste of any sorts.

In a way this is how all plasmas, atoms, planetary systems, stars and galaxies create their own motion and Magnetic fields (Fig. 60 and Fig. 61). Why not for the man to take advantage of this new knowledge of clean source of supply of Matters for new energy production technology through the real universal method of workings of magnetic fields of Matters?

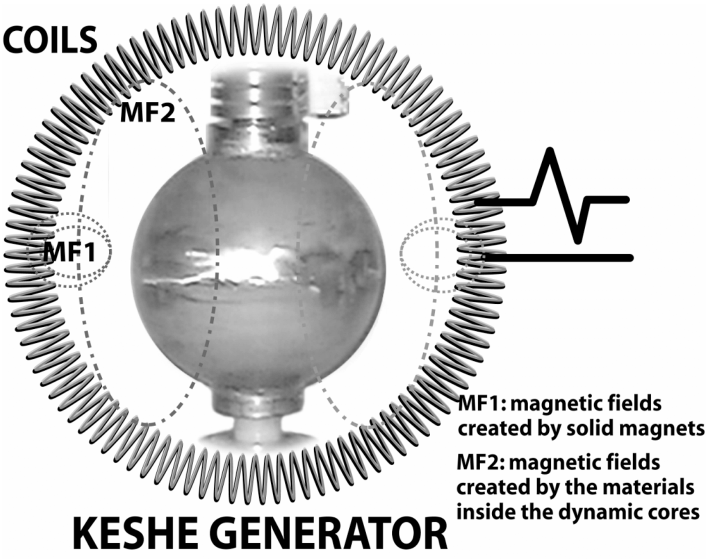
COILS
MF2
MF1
MF1: magnetic fields created by solid magnets
MF2: magnetic fields created by the materials inside the dynamic cores
KESHE GENERATOR

Fig. 60: The Keshe generator, which rotates for years in a coil structure.

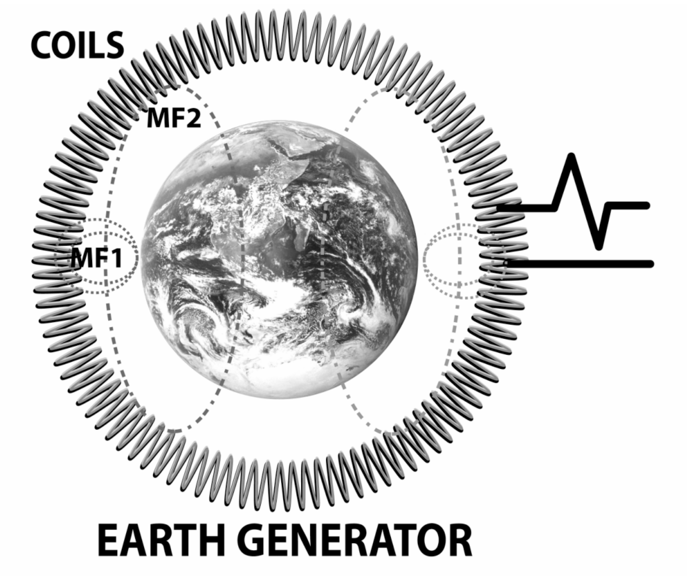
COILS
MF2
MF1
EARTH GENERATOR

Fig. 61: The virtual Earth generator, which rotates and generates Magnetic fields (MF1 and MF2) for millions years, and could generate electricity if there was a copper coil around it.

Similarly, using the plasmatic magnetic fields reactors can **eliminate** the need for generators and **transmission lines**, due to the fact that the pmtics needed for production of a given current and voltage level can be created at the point-of-contact of the system. Even the structure of the matters electrons or plasmas can be used to create current for their use in the same equipment. This means that one can produce current from the atoms and plasmas of a silicon or plastic cover of a microchip, to meet the components energy needs.

One possible use for a small reactor is to produce the right plasmatic vibration of electrons or the plasma of matter, within any given environment to generate light and power.

Prototype reactors have already been made and can be further refined that only need to generate plasmatic magnetic fields within their cores, that the interaction of the pmtics generated by these reactors with its surrounding matters, creates a *magnetosphere* at the point of the interface of the two pmtics of the system and atmosphere, that leads to the release of pmtics in the range of visible light of any colour and ray. This will be how future light producing units will be manufactured.

Similarly, in a given environment, in such as a craft, the operation of reactors can produce Magravs, that the interior of such crafts would always be lit due to the principle of the interaction of the surrounding Vomafs of the reactor and the dynamic plasmatic magnetic fields of the particle content of the air within the confines of these crafts.

The strength of these Dypmfs generated by the reactors can be set to produce any colour rays of any magnetic wavelength, be this red or orange light, and X-rays or gamma rays. We have seen this effect in laboratory tests and have measured the radiation fields.

The power generating units of the future, using the plasmatic magnetic dilution technology, will be a departure from the present power and light production technology.

Grapos, due to their ability can create rays at any frequency, can be used for example for disinfection of any give environment and materials, by the production of Ultra Violet lights for matter and Extreme Ultra Violet for Matter in space.

The same principle, for generating matching plasmatic magnetic fields for the production of proteins in these reactors have been tested and proven to be correct. These tests were carried out in 2008 and the production of proteins by the use of nitrogen from air through this technology was proven to be simple and achievable.

In separate tests conditions, the hydrogen atoms of water were used to transfer the right plasmatic magnetic fields to the right pmtics strength of plants and human cells for the pmtics of the cell to gain nourishment or recover to their original level of plasmatic magnetic fields' strength.

This technology has been used in trial cases for helping volunteers with different vitamin deficiencies, or cell plasmatic deficiencies, to recover from the disorders produced by the deficiency, and people to live a normal life after years of suffering.

Alternatively, this technology can be used for production of new materials for coating or new processes of coating.

At the same time, through simple magnetic plasmatic reactors, the lattice structure of matter can be changed so that the nano-components of the matter can be realigned to change their lattice formation to another. The change of characteristics and properties of one element to another has been tested and confirmed to be correct by independent organisations. In these tests, the nano-matter of an element was captured and then its lattice formation, through the right plasmatic magnetic fields conditions, was changed to attain new properties from the same matter. This was carried out through several tests, at room temperature and pressure, where ***atomic carbon*** was initially released from a C-H bound matter. The carbon was then allowed to deposit itself on a layer of copper. This deposition was dependant on the presentation of the right plasmatic magnetic fields at different points on the same copper Matter. This copper surface was then tested and proven to have sections of nanolayer coating in the form of sp2, and on the same layers and same condition the lattice structure changes the same material to a diamond lattice structure, or what is known as sp3. Where, sp2 is a two-dimensional nanolayer, and sp3 is a three-dimensional diamond-lattice nanostructure.

The plasmatic magnetic fields' of reactors can be tuned to similar frequencies' as the plants pmtics, to generate certain vitamins and minerals. These single or multi-plasmatic magnetic fields producing reactor sources can and will replace the fertiliser industries of today's technologies in space for food production, once scientists grasp the method of their use.

In tests, seeds of herbs were grown by being feed using water treated in these reactors, the parallel seeds, which were grown normally with rainwater, died in early October. The herbs that were watered from reactor system were still green till March of the following year, because the molecules of the water were structured through the reactor system to act as fertilisers. Similar tests are in progress.

In this section, it has been shown and explained some of what has been tested and achieved through this new understanding. Scientists and users of this new technology will find their own ways to expand and will add their own new knowledge to our disclosures.

Seed of thought

Where, in the present time the majority of the nations in the third world due to the lack of funds to purchase fuel even in having power stations, or being able to afford to build generators at the cost of tens of million of dollars, they cannot afforded to run these generators to provide power for their nationals. Secondly, there is not enough production capacity to build new power stations and power grids. Therefore local alternatives can be in the range of windmills and solar cells, but these will not solve the incredible lack of electricity in next decades.

Now, through the use of **Keshe generators reactors**, which can cost as little as a new car, these governments can provide electricity and clean water for their nationals.

Where, these governments do not need to pay for fuel as these reactors run on hydrogen extracted from their environment. These reactors can be placed at the ***point-of-demand*** in a village or in the cities without the need for transmission lines or for any expertise to operate these reactors. Where through the production of electricity, purification systems which are part of these reactors integral operations will allow fresh clean water supply to every child, woman and man on this planet. This new energy source and clean source of water, given opportunities to third world nations to rise above and from poverty and reliance upon foreign aids and charities organisations. Where these nations can produce food and warmth for their nation as the rest of developing world without the backbreaking debt to other nations.

According to a ***United Nations report*** (Energy Challenge for Achieving the Millennium Development Goals 2006), "*Currently****, at least 1.6 billion people do not have access to electricity*** *for lighting, refrigeration, mechanical power, telecommunications and other beneficial uses. ... Moreover under today's policies and investment trends in energy infrastructure*, **1.5 billion people will still lack access to electricity in 2030**". (42)

Of course this is not acceptable. It is clear that humanity is not capable to deliver fundamental care to all of its members, even not within next twenty years!

Man is not complete and healthy if parts of it human structure suffers from paint of self-inflicting disability, this is the same for the human race as whole. This race is not complete until all its members have and can exists and have the privilege to live the same standard of the development as the rest.

"The future is bright, the future is plasmatic magnetic fields based, through the production and control of plasmatic dilution technology reactors to bring peace and comfort to the human race."

CHAPTER 26

Matters Experimental results and observations

Tests for the proof of all the concepts have been carried out over some years. The direct results and observations made in some of these tests have already been mentioned in some other parts of this book and now will be brought under one chapter collectively and will be discussed in general.

The reason for writing this book is to disclose the theoretical new first principles of our technology; because most of our test results are so groundbreaking that they cannot be explained by actual knowledge. Actual science is partly based on a number of mainstream accepted assumptions about fundamental interactions and particles and matters, where these assumptions have become almost "laws", and do not consider the true occurrences of real events in the universal order of creation of particles, plasmas, matters, energy, gravity and motion as they take place in the universe.

Where, the major processes in the universe are based on fusion and fission of Matters and plasmas, and the transformations and conversions of matters from one form to another. In this book we explain how such conversions and transformations can be reached in plasma reactors, and by this insight we were able to attain such amazing results. Where some people don't feel comfortable with them, or a priori reject them because these new findings and their applications counter their school of thought and their financial gains.

For example, one of the major obstacles in the world of nuclear fusion has been to overcome the Coulomb barrier for fusing two plasmas, and this problem has not been solved for the past sixty years by all scientists in this industry, even though billions of US Dollars and more has been spent by different nations in laboratories and fusion reactors testing to overcome ***the weakest of the weakest of magnetic fields interactions*** in the world of creation, which this is a magnetic barrier, also known as the Coulomb barrier, where this barrier is exactly the Magravs of the plasma, and as now we know how it is created, we have found the solution not to fight and force to cross this magnetic magnetosphere barrier, but the solution is how to create an environment matching the magnetic fields strength of this magnetosphere of the plasma for it to open-up and unveil its real constituent Matters. As the environment of the whole reactor becomes in magnetic fields strength equal to the magnetic fields strength environment of the plasma. Therefore, through this method, there are no barriers to overcome, and fusion of matters of plasmas or the use of the Magravs of the Matters of the plasma becomes a child's play.

For the first time in the world of energy and motion, we offer through dilution reactor technology, at low costs, to open the plasma and dictate the strength and the resistance of this Coulomb barrier to realize the dream of fusion.

Similarly, to achieve lift and motion where, the present space agencies are only able to deliver a limited payload into space, and the burden and the costs of such launches of crafts and personnel, cannot be met by one nation easily. Where, Magravs Grapos reactors can lift much larger payloads at low costs, without limitation on passengers and no weightlessness conditions in space.

Radiation

During tests in 2008, radiation detectors in the surrounding environment of a test reactor detected high levels of magnetic fields radiation. These high-energy magnetic fields have been measured more than 1.5 meters away from the plasma reactors. Where, there were ***no neutron sources*** in the laboratory or in the building. Through known physics or theories, these radiations should have not been in the environment.

These strong radiation fields were tested and looked for and were not present in the laboratory at any other times with other different testing configurations with the same reactors, using the same radiation measuring detectors.

Such high levels of magnetic fields radiations could have only been due to the release of the magnetic fields that had been entrapped in the Matters of the plasmas, which the system was set to produce. These releases of components of the plasma were achieved during the specific pre-set configurations of loadings of reactors. These strong magnetic fields were released by the operation of multi-core plasmatic dilution environment reactors and at no other time they were detected in the laboratory.

Most of the reactors, mainly built and tested for lift and weight reductions, are made of different simple parts, which some part have materials like plastic rings, and PVC, thus these systems do not and could not have been operating at high temperatures.

In other tests in 2008, and in repeats of the same tests in January and February of 2009, the same results have been achieved, all of which have been recorded on video.

In these tests with the use of a less than *one* ***gram*** *of mixture of matters* in the similar multi-core dilution magnetic fields systems, fundamental plasma separations were created through the same method as before. These tests were done using five different reactors, each having a different set of internal configurations.

In all test settings of the reactors, different loadings of materials and different dynamics for the reactors were used.

Weight reduction, motion and lift

In all cases, when the dilution of the plasma and the subsequent untangling of its Matters in the reactors were achieved, then the effects of the disassociations of Matters were observed through the effects they created internally in the reactor and externally through lift and motion, and jamming of radio-frequencies and so on.

In the early original reactor tests, full lift was achieved so fast that the effects of free motion were noticed a while after it was clearly apparent and were in progress, through the replay of the video recorded sessions of the sessions of tests. Through video replay of the recordings of tests, the motion of the system freely on the floor of the laboratory could clearly be seen prior to physical observation of the lift of the reactor system. These reactors have no protrusions, are not equipped with propellers, or portholes, which could cause the lift of the reactor by air repositioning means.

In one configuration and loading of the system, control of the system was lost, so the reactor had to be caught by hand in mid-air, that it could be brought under control before it was damaged.

In one configuration, two reactors with different external bodies and two different internal settings were used next to each other. These systems were kept about thirty centimetres apart from each other. These reactors were entirely different in design and had different settings and loadings. In the two-system loadings that were adjacent to each other, the systems' interacting fields pushed one reactor away from the stronger Magnetic fields of the two systems. With the same systems and different configurations of loading one system was attracted by and was pulled by the stronger gravitational fields strength system, where physical force was used to keep the two systems away from colliding with each other.

"These systems have no portholes, and no physical means of interaction with the outside environment and no aerodynamic systems to cause them to propel up or move side-ways through the air, and they have been no solid magnets or electro-magnets for repositioning or for motion, for them to move the total collective weight of 15 kg of the two systems. There are no external systems and tracts used to help causing the lift or motion of the system in the laboratory. Thus the motion is ***not due to*** **propulsion** ***or*** **aerodynamic** ***design and movement***". This is the conclusion of independent expert appointed as observer for the validation of the results of the tests.

During other tests systems weighting several kilograms, their weight were reduced in a controlled manner at increments of one hundred grams and two hundred grams per setting of a reactor and these tests were all videoed and recorded. For reduction of weight in these increments, the systems needed less than one thousand^{ths} of a gram of a new material combination as Matmags, even in some tests no new material was introduced and by just adjusting some internal parameters of the reactors reductions were achieved.

With the use of such a small amount of physical mass of matters used, creation of such lift and the reduction of system weight seven kilograms, has been said to be impossible without any external energy or equipment.

Through the understanding of the principles of the plasma dilution methods of disassembling the plasma components to utilise its sub-components, systems were designed to achieve these effects in simple ways. With the interaction of the disassociated Matters components of plasmas in different parts of the system, then the scene was set for creation of profound effects such as a *reduction in weight* with respect to the Magravs forces of the planet.

The interaction of the strong Antimatters' of the plasmas components, which were released in the different cores of the ***plasma dilution reactors*** and the interaction of the fundamental magnetic fields of Antimatter upon each other, was observed during different tests. The interaction of pmtics of Antimatter leads to the creation of the strong Magravs in the vicinity of, and within, the reactor cores.

Through the control of plasmas and by being able to separate the components of Antimatters and utilising their constituent magnetic fields in reactor cores, this led to the creation of Magravs forces that initially led to a reduction in the weight of the whole system. In following on with this process there was a detachment of the system from the ground as the system found a new Magravs positioning in respect to the Earth's Magravs force strength.

When looking at the effects and results achieved, there is no doubt about the powers of the interaction of the magnetic fields of Matters, which were diluted and their fields were allowed to interact in these reactors' cores. These effects were attained by allowing a number of the pmtics of the Antimatters of the plasmas *to integrate* ***in one core***, and allowing these fields of Antimatters to interact with the same and similar fields *integrated* through the same principles in ***other parts*** *of the reactor*. Where, the interaction of the pmtics of the components' of Antimatters in different parts of the system with each other, led to the generation of strong specific plasmatic Magravs. The reactor facilitates for these stronger Magnetic fields to interact and generate strong plasmatic Magravs in the reactor cores, with allowing these Magravs created in the reactor cores to interact with the Magravs forces of the planet, and due to the strength of the fields produced in the reactor, this causes' reduction or increase in the weight of the system in respect to the planet's Magravs.

The new Magravs created within the system, pushing against the Magravs of the planet at the boundary of the systems fields, causes the variations in the weight according to the combination and configuration of the Grapos loading. The produced Magravs forces *within the confines of the system* are independent of the external planetary gravitational fields. In numerous tests, it was observed and recorded that the changes in the *internal* gravitational field forces of the system, did not necessarily change the weight of the system in respect to the Magravs of the planet. The interacting internal pmtics forces clearly showed to be generated independently by increase and decrease through internal electrical measuring components set in the system for the detection of these types of changes.

It has become clear that it is the interaction of the Magravs of the planet and the system that have brought about the reduction or increase in the weight and lift or descend of the system in respect to the planet. In some tests, these increases or decreases in the weight of the system have been without any addition or withdrawal of physical matters from the system.

In continuing with the same processes of creation of varying Magravs in the system, weight reduction was managed and followed to such an extent that the total balance of the interaction of Magravs of the system and the planet, in respect to each other, eventually zero reading on the measuring scale system. The point was reached where there was total physical detachment of the system from the table or floor first, and then the upward motion of the system, away from the surface was achieved.

Some call this the anti-gravity principle but in fact, this interaction and new balance in Magravs forces in between the system and planet, is due to both gravitational and Magnetic positioning of the reactor system in respect to the planet, as both objects are in possession of ***active** dynamic plasmatic Magravs forces*.

In a specific reactor, set for weight reduction, internal gravitational field reduction and zero-gravitational field force were disturbed, where the human touch was applied causing the change in the magnetic interface between the fields of the system and planet. This gave a tingling feeling on the tip of fingers after repeated tests.

In numerous tests over years, where the equilibrium state of internal Magravs strength between cores were achieved, the whole system, including internal physical dynamic parts, came to halt, sometimes for up to thirty minutes or more. Then, the dynamic motion of the systems re-set by itself and the system restarted. This ***re-setting*** has been due to losses and reduction of one Magnetic and/or gravitational field strength in respect to another within the cores of the reactor.

A number of times, *external power sources* were used to try to restart the system. In some cases measured energy of up to three, four and eight times the normal operational energy was continuously applied to the internal systems to try to push-start the system, but to no avail.

This shows that the power generated by a countable number of Matters of molecules, was so powerful that additional systems built in the reactor cores could not overcome the kinds of powers produced by so few Matters magnetic fields interactions.

These tests have shown the creation of independent Magravs in separate cores of the reactor. They show the creation of independent plasmatic Magravs within the system boundary, and then the combination of the same Magravs prevailing around the outside or beyond physical boundary of the reactor as magnetospheric or cosmic radiations were produced, which these types of fields are prevalent as have been observed in space around stars and planets.

Indications are that these radiations created in these system, due to their characteristics, do not shown any side effects in the human tissue of the operator up to this moment, years after having been involved with these fields tests.

These fields around the cores are considered to be due to the interaction of plasmatic magnetic fields' of Matters of plasma and environments' Matter or matter, and are independently created in different parts of the cores of the system.

This technology opens up new opportunities for the industrial development and manufacturing of systems that have different gravitational conditions inside the system for living, and have different gravitational conditions compared to the outside and around the physical boundary of the craft. This allows for having a one-G gravitational field force for the normal habitation of Man inside spacecrafts, or in a dome outside and around the crafts for habitation. While the conditions outside the craft can be zero G to tens of G's. The passengers of crafts using Grapos can walk around the craft, or in and around the dome created by the extension of Magravs of the systems, on any planet surface, as if they were in Earth's gravitational field force environment.

No high temperature core environments or sudden surges of heat were recorded in any tests. The reactors were in some cases opened immediately after operation for new settings, in a normal lab environment. There was no physical damage ever noticed to the internal parts of the system. This includes melting and twisting or destruction of any parts due to the presence of Matters.

Even though we lost a full system by pure enthusiasm to catch, lift upwards, which the jumping on the systems after several times cracked the structure and the systems was lost. We have kept the reactor casing for further analyses of matter lattice changes, and another reactor was damaged when a Black hole condition was replicated.

On only one occasion, where a specific setting was used, a black spot was noted on one of the external sides of the reactor core wall. This is supposedly was due to an energy surge during the test. The effect of the black burn was not removed for further analysis.

Production of matters

In different tests, a new orange and gold coloured material was noticed on two occasions in a reactor, and samples were collected for further testing to show the material production capability of the system.

This method of the production of Matter can be used by manipulating and changing the plasma's and rearranging them for the production of oxygen, hydrogen and water from collections of initial fundamental particles found in space, for human deep-space travel and colonisation. Where, the components of plasma absorbed from the cosmos are loosened by producing the right pmtics strength in the reactor for production of the first two atoms of hydrogen. Using further plasmas from space, in the matching gravitational fields as of the atoms of oxygen in the same reactor, this element can be produced. Then through the same reactor's operation, this allows the combination of the two elements, of hydrogen and oxygen, which was produced from the space plasma, this allowing the ***production of drinking water***.

In the future tests, one can attempt to combine the gravitational effects of the systems for production of energy, material production and other capabilities of the technology all simultaneously in ***one reactor***. We have achieved this much but there is a lot more to learn.

These reactors are truly an integrated system, which can produce and attain several effects and properties, like production of Magravs, creation of matters and so on simultaneously.

Dark Matter technology

The system using the Dark Matter in travelling through the Magravs of the planet creates no magnetospheric conditions, friction, or pmtics force resistance, between its pmtics Magravs and the planet pmtics Magravs. Thus, zones around a craft can be created that there is no magnetospheric field between the craft's magnetic interface and the planetary system's. Hence, a craft using the Dark Matter ***cascading*** can travel within a stronger pmtics, like that of the Earth, *without detection* since the system can create no interface magnetosphere, as there would be no visible light to be seen with this method of motion.

Through this method of using the Dark Matter for transportation, high-speed motions can be attained for spacecraft in planetary systems, solar systems and galaxies.

Using the ***Dark Matter principle for transportation***, the environment between the systems and the planet is **saturated** with the dynamic plasmatic magnetic fields of the system and the planet. This zone appears to possess energies but they cannot be linked to any source. Even though the source of the dynamic pmtics, in motion around the system or so-called energies of the pmtics, is set in the centre of a strong plasmatic Magravs systems inside the crafts and centre of the planet. Such Dark energy fields give the appearance and confirm the existence of dynamic energies, which have no explicit source of origin. Where, the source of energy or the dynamic pmtics is the hidden gravitational source in the centre of the Matters reactors and the planet.

This new technology is a radical breakaway from the hold that physical matter have had on the destiny of Man on the Earth.

With this new understanding of the methods of how Matters are created in the universe, it is possible for Man to use the real Matter sources for energies and powers available to him in the universe for his future scientific evolution.

CHAPTER 27

Discussions and Conclusions

With this new understanding, the elementary particles have become to be the ‘magnetic fields’ and their interaction leading to the creation of Magravs, this leading to the creation of Matters, as the state of each Matter depending upon the strength of the magnetic fields which have led to its creation in respect to their environment fields strength, which the Matter appears as the Matter, the Dark Matter or the Antimatter, where the collective interaction of the Matters Magravs, leads to the creation of the Initial fundamental Plasma (the Neutron), where the environmental magnetic fields forces conditions, around the Plasma, dictates the physical appearance states of the Matter part of the plasma as matter, this being solid, liquid or gas.

It is obvious that the universe will never stop creating new Matters and matters and their effects, and their interactions, as the universe is a soup of dynamic and ever-changing conditions, created by the Magravs positioning of plasmas, dynamic Matters, matters and their components and effects.

Man can learn how to create the same conditions as the universal order, and to rearrange these fundamental particles for his use, in a much simpler way than he has chosen up to this point in his path of technological and scientific evolution by the use of matter(s).

In the present state-of-art technologies for the production of energy, scientists have chosen the path of force and *destruction of atoms and molecules* to overcome barriers. In the universal order of creation, one works within the magnetic fields forces to achieve the same and even better results.

This ethos of ***burning and destruction*** has been the fundamental focal point for Man’s survival in his environment, due to what he has been conditioned to in his evolutional path on Earth. He has seen the power of fire, from fires in forests, and he has used this knowledge to keep himself warm and so he has learned to do the same using the same methods and materials. By perfecting this art, through the burning of wood and oil and so on, he has managed to create energies for cooking and warmth and running industries.

Then he saw the birds and understood the concept of motion in air and so he used the concept of fire to create propulsion and jet engines, still burning one form of fuel or another to achieve flight, still based on the principle of ***matters repositioning and conversion***.

He has seen death through disease, so he has produced other elements to destroy the germs that can cause and bring an end to his uncertain and fragile physical life.

If man would learn to work within the structure of **the universal order of creation of**

Matters and the method and principle of their use, he will learn that he does not need to burn or destroy anything to keep himself warm or feed himself, or even to overcome diseases.

He can very simply change the characteristics of the plasmatic magnetic fields of the elements that caused the misbalance in the first place, leading to his ailment, without swallowing a single tablet in his lifetime.

We have developed and tested such simple systems in the past years and have seen their amazing results.

For example people suffering from Fibromyalgia for thirty years or more, within 3 months of drinking water through these systems, have started living a normal life without any signs of their disease. Follow-up over 18 months has shown a total eradication of the ailment.

We have tested this technology on other forms of illnesses and viruses and the results are amazing. However, this is what is to be expected with the use of the true knowledge of the creation as has been disclosed in this and future books.

With this new disclosure, man in the future will learn to live and operate in the real universal order of creation of Matters and their existence in the universe for his life. He will never destroy any Matter or matter, but he will learn to work within their structure. He will learn to achieve more, with less use of the resources in his environment. This is an improvement on the past, whereby at the present by the use of destructive methods, he has brought himself to the point that the abuse of the matters is now endangering his future on his own home planet.

Man has to learn to work with the real elementary particles and elements and their functions in the universe, such as magnetic fields and their interactions, for him to be able to produce what he needs irrespective of where in space he might be or would want to be.

Maybe this knowledge and the proof of its pleasures will change the ingrained ethos of destruction for the survival for man. One could even hope for a change in his animal instinct of killing other people and other creatures of god to survive.

Additionally he can change his habits and the instinct to want more since through the understanding of this technology he can have whatever he desires from his materialistic world.

In this disclosure, some theoretical nature of the universe has been spoken about, but the real physical aspects can be created to achieve the same conditions as in the universe as we have shown through our tests.

Man will be able to carry on with his life, but at the same time will be able to live in a balanced environment.

The reactors we have developed, which in real terms are ***miniature universes***, that can provide man with all his needs, such as water, food and medicine for him to survive and enjoy the beauties of creation.

The conditions in these reactors are soft and gentle environments and the methods of their use for control and creation of mans' needs are based on working within the universal structure, and not destroying to achieve, as has been the principle of the past for the Man.

For example, when the principles of the creation of Magravs are applied to the lower orders of creation, like the interlocking of atoms to make molecules, there are explanations as to why certain types of Matters are attracted to the reactor, and why certain molecules are, or can be, produced in certain ways and combinations, in certain parts of the universe.

The same applies in the even lower orders of creation, that one can explain and produce certain magnetic fields and their interactions that can lead to the creation of present fundamental elements in the centre of the proton, plasma, an atom and finally the man himself.

The world of science has to understand the conditions by which the interactions of initial magnetic fields can lead to the creation of an initial state, for the creation of the Matter, matter, plasma and finally atom.

Then there is a need to understand and demonstrate as is done in this book the way Matters and matters are created in vast spans of the galaxies.

Scientists have studied over centuries the structure of matters and the atom, and their interactions, in their material and tangible *states of matter*. However, the structure and method by which the atom is created, maintains and guaranties it's existence has not been fundamentally fully understood, due to the lack of overall overview of workings of the universal order of Matters. Even the construction of protons with its sub-components (quarks, gluons), up to today, is based on assumptions and indications by the use of particle accelerators and plasma reactors, by *only looking at* ***matter*** *components* of the proton and some Antimatter parts. So the present scientific world looks for an answer in the construction of the proton on the ***matter-to-matter*** level.

By the present science, the existence of the Matter, the Antimatter and the Dark Matter are largely considered to be independent entities in different points in space, and have never been considered to be as ***one collective*** *and all being* ***integrated parts of one system*** making up the plasma, up to this disclosure.

Mass of the Matters and Mass of the plasma

In the new knowledge we consider the three different quarks of a proton, ***as the three different Matters*** (Matter, Antimatter and Dark Matter) ***of the plasma***.

Where each Matter is *created* through the interaction of at least two different magnetic fields or pmtics, where through the interaction of these two plasmatic magnetic fields of different strength, initial gravitational and Magnetic field for creation of different Matters of the fundamental plasma are set to start.

So that the interaction of *at least two pmtics of* ***each*** *Matter* creates it's own Magravs forces. What we call the interaction of the Magravs strength forces between and in respect to each other of these three different Matters of the proton.

Where, through the understanding of the method of creation of different Matters of the plasma, due to interaction of at least two magnetic fields and the creation of gravitational and Magnetic fields, this confirms that ***each Matter*** irrespective of its environment Magravs *possesses an independent* ***mass*** *of its own*. In the present science this is called the mass of the quark. In fact each Matter, Dark Matter, Antimatter and Matter, has its own individual mass in the plasma. As each Matter is created by the interaction of at least two different magnetic fields strength, the ***total mass of the plasma*** is the mass of the three Matters plus the total interaction of Magravs interaction of the overall plasmatic magnetic fields within the plasma. Consequently the *total mass of the plasma* is always considered to be ***larger*** than the *total mass of its three Matters'* components.

Due to the dynamism of the interaction of the two initial pmtics, leading to the progressive creation of the Magravs of each Matter, *each Matter* is always in a ***dynamic rotational state*** within the plasma. Where the **speeds'** of the dynamic rotation of the different Matters in the plasma are not the same and they are independent from each other, since their speed of rotation is set by the strength of their initial magnetic fields interactions. Where in the present science the observation of this natural phenomenon of rotation is called the ***spin of each quark***. One can compare this, in cosmological sense, as the ***independent*** *speed of rotation* of the Sun and Earth, in spite of that they are within the same solar system. Where the initial plasmatic magnetic fields strength, which leads to the creation of different Matter(s) in it's interaction with it's environment pmtics strength, leads to the creation of an individual magnetosphere of a specific field strength for each of the Matters. Where the intensity of each magnetosphere creates *different light intensity* in its environment.

Now we understand that the interaction between the pmtics of the magnetosphere of each Matter of the initial fundamental plasma leads to the release of different pmtics strength fragments in the range of visible light (Fig. 18).

Due to the interaction of the plasma of each of the three Matters with each other and the release of light or energy, these interactions in parts reduces their strength and the pmtics content of each Matter, where due to these reductions in Matters quantity and strength of pmtics, different colours are attained by these interactions for each Matter during the cycle of the life of each Matter.

It has become clear that all states of Matters are the same in their nature of their creation with only differences in their field strength. What does differentiate these Matters from one another is in principle the strength of their fields and their position(s) and interaction(s) with other pmtics in their environment and in respect to each other.

Thus, Matters are Antimatters and Dark Matters depending on the point-of-observation and the field strength of the environment of the appearance at a given point and position, but they are all made of the same entity at source, which is the magnetic field.

The front-cover image of this book has been chosen specifically to emphasise for the scientific world to reconsider their assumptions.

The Matter can change to, or appears as the Antimatter and the Dark Matter and vice-versa, if the right conditions and the right plasmatic magnetic fields forces are given or taken away from the pmtics of the Matter.

In this disclosure, we have explained in a simple way that magnetic fields are the *new* initial fundamental particles, and the **magnetic fields** are the new elementary particles ***instead*** of what has been assumed at present as *quarks*. So, the magnetic fields and their interactions are the **creators** of these quarks and so on, in the plasmas.

Then we have explained how the Matter, the Antimatter and the Dark Matter, and the relevant energies of the Matter, the Antimatter and the Dark Matter, are all decided, dictated and created at the inception of the creation of the first stages of the creation of the initial fundamental plasmas of these Matters.

In this disclosure, we have discussed the results and shown how the three states of the Matter, the Antimatter and the Dark Matter are created, but through ***real tests conditions***, we have confirmed the co-existence of these three Matters and their respective energies.

For example, in tests there have been the very strong magnetic wavelengths of cosmic gamma-like radiations levels, which have been detected ***for meters*** around the system, *where there has been* ***no physical source*** *of such radioactive materials ever being present*. Whereas, it is a known fact that alpha and beta rays cannot travel that far in room condition environment. These levels of radiations could only be created through the interaction and the creation of strong magnetic fields in the reactors cores, which in their interaction with the pmtics of matters in the room environment have led to the release of these gamma-like radiations. This is exactly by the same principle as the Magravs generated in the centre of the stars in their interaction with the plasmas of matters on their surface, generate lights and high-level cosmic radiations.

An **atom** is considered to be “a collection of different densities and strengths of magnetic fields interacting with each other, in different magnetic mixtures and levels, in, and under, the same physical rules of the existence as the rest of the Matters, matters and bodies in the universe”. The atom is nothing more than a collection of the lower order energy magnetic fields strength in motion in the universe, that always come together in the same specific pattern of dynamic pmtics and strength.

The question is, that, is it the collections of magnetism, which created the MATTERS’ of the plasma, or did the existence of the MATTERS of plasma create the magnetism?

At this point, the existence and understanding of the Dark Matter becomes important, since the human eye and the physical machines of today can only observe what is in possession of visible and detectable energy fields.

The **Dark Matter** is created by interaction of magnetic fields which create and are in possession of intense fields of gravity and magnetism, and at the same time, and it is void of visible light in respect to their given environment magnetic fields strength. On the other hand, the existence of the Dark Matter is position dependent and pmtics strength balance dependent, in respect to its environments pmtics' strength.

Through the operation of hundreds of static and dynamic tests we have very strong indications that the universe is indeed made in the normal conditions and of packages of dynamic pmtics.

From our observations, we can say with a large degree of certainty, that the dynamism of the whole universe is constantly maintained through the principle of the interaction and attraction of different pmtics, according to their strength and positioning in respect to their neighbouring rays, fields and plasmas. Where magnetic fields of different strengths are locked to each other by the principle of their ***Plasmatic Magnetic Energy*** (PME) and ***polarity***. These causing the first stage of the creation of the magnetic attraction or what is called the ***Magravs force*** for the creation of the first Matters and plasma and further on of the first atoms, molecules, matter, stars and galaxies.

Through the above disclosures, it is clear that these components of plasma themselves have no solid parts, but are merely *a collection of energy and fields, trapped or interlocked in different magnetic fields strength of the other parts*, according to their strength where, they can appear as the smaller components of these parts or near to visible as the physical state of matter.

These dynamic pmtics or energies on their own in the universe are not that influential, but as they *gather with other collections* of dynamic pmtics or energies of the same levels, they appear as part of Matter, the energy constituents of an electron or the Matter parts of matters, that they show and exert their influence.

When the pmtics binding principles are fully understood, the world of creation and the understandings of the origin and working of the universe will become much easier to work with.
Why, only a part of the pmtics levels is visible to Man is not a mystery, but the reality is that Man has the habit of only looking at things in physical terms. If Man looks for the real existence of the effects of the energy levels, he will open his eyes to a new horizon in the world of creation.

The question is if one can develop the ***pmtics*** as in the universe and manage to release part of these plasmatic magnetic fields from initial plasma as energy, then, one can manage to control the level of these pmtics and their interactions that are needed to create and replicate energy combinations that can lead to the creation of the Matter and then matters. This in time will be done and depending upon the openness of Man to accept the realities of his owns creation and his existence in the midst of others.

In developing reactors, which have been openly disclosed, and in understanding their modes of operation, which are discussed, these reactors will not only be used as future sources of energy for Mans energy needs, further, these reactors will become the incubator for the creation of the energy levels, which can lead to the creation of the pmtics, which will lead to the creation of Matter and matters, like oxygen, food, water and so on.

Through the use of the plasma dilution and gravitational reactors systems of the future, one can and will experience the Black Hole concept and the reversibility of the polarity of planets (12), as we have done in our lab tests and become aware of their pleasure and dangers.

In these plasma dilution technology reactors it is not only the fusion of Matters that is possible, but also the fusion of Antimatters and Dark Matters and their real application in the production of Antimatter and Dark Matters atoms and molecules that will become part of daily use. Where, these will be used for the production of energies and effects, which are above the levels of the Matter components strength.

The beauty of the application of these universal reactors will be the opportunity for the creation and manipulation of the much weaker and smaller energy levels that can lead to the creation of electrons and protons.

These weaker strength magnetic fields for the creation of Matters will be noticed and can be collected at the *outer boundary of the plasma*, but not on the edge of the plasma in the reactor. This is due to the fact that the edges of these reactors are dynamic and are violent areas in which temperature interfaces will not be so suitable, but can allow creating the conditions for weak energy levels to be maintained and to manifest themselves. These zones are good to create the conditions for the creation of electrons' plasma, where loose plasma conditions are needed for fine-tuning of the fields in these reactors. This is due to the fact that electrons, in their foundation, are made of low energy levels bindings and interacting with each other, largely in a dynamic volume as compared to the protons, and in possession of weaker binding energy than protons.

To create conditions for the production of the protons, there is a need for the same energy levels, but tighter and stronger magnetic fields plasma conditions are essential and the size of the plasma and the mixture of energies available in both conditions are important.

The techniques for the use of the ambient temperature reactors to perform fusion or the production of new and heavier atoms can be learnt through the development of initial fundamental plasma dilution technology, which is very much different from the mode of operation of the present fission or fusion reactor technologies.

In the present nuclear fusion industry, the most important point in the world of creation, the *presence of controllable gravitational field forces*, has been absent and totally ignored by scientists.

In **dilution plasma technology**, as in the universe, **gravitational field forces** within the systems are the ***cornerstones*** of the operation of the reactor.

This is a unique method of the creation of new Matter and the creation of energy, which has been explained in the operation of dilution plasma technology. This is exactly how Matters are created in the universe.

In the right conditions and mixture of initial fundamental particles, the use of plasma dilution technology will lead to the transformation of the *initial fundamental plasma* into protons' and electrons' to create the basic matter and simple atoms like ***hydrogen***. With the right method of application of the plasma dilution technology, hydrogen atoms can be used for the production of the heavier elements.

It is important to note that, in an atom, "The interaction between the two pmtics of neutrons and protons Magravs are in fact the source of the **double magnetic fields** of the atom and consequently the creation of the Magravs of the whole atom.

Even though **neutrons** are supposed to be ***balanced*** *pmtics entities*, the magnetic energy of the nucleus can be ***increased*** up to a limit *without the disintegration* of the inner gravitational forces of its constituent parts for its components Matters field strength to attain the plasmatic magnetic energy of a higher strength or new pmtics element.

In his progressive evolution, Man has learnt to use what has been available to him, like wood to create fire, steam to create power, the use of oil for engines to create motion, and in the recent past the breaking of the atom and joining of plasma in nuclear reactors for energy.

Through the understanding of the creation of the initial fundamental plasma, the energies of not only the Matter parts of the plasma are used, but this technology allows the utilisation of all Matters of the plasma, that is the Matter, the Antimatter and the Dark Matter and their energies that these possess and can produce, that come into play and can be used.

In the other words, we welcome the Man to ***the real world of creation***.

CHAPTER 28

Insight into the Future

Using the new acquired knowledge from above chapters, one can look into the future and say that the struggles of the past will bring beautiful fruits of creation for the future of the man.

From understanding of the operation and relation between magnetic fields within any structure, this being an atom, molecules, and even a star, different types of life in all corners of the universe can exist in all forms, and through other matters then the nitrogen based protein chain of the human race, which other lives through their different pmtics can operate through the same principles as the human protein chain. These creatures through the balance of their pmtics criterion do and can exist in the spans of universe. In fact, their protein chains with different matters than nitrogen could be more efficient in its use of passive pmtics than the humans' protein chain in using pmtics from the soup of the universe.

These Beings might have different colour of liquid distribution medium than Mans' red blood, but not necessarily this mean that they are any different from the Man in essence.

For example, if the liquid of life for these creatures of god have the potassium as the base element instead of nitrogen or mixture of the potassium and the nitrogen, then this liquid can function as good as, and even better than nitrogen-based protein of the human blood for converting and the use of Matters and matters in the universal soup of supplies. Even these Beings' protein-based cell can be more efficient for the conversion and the use of the passive pmtics of the Matters of universe. This efficiency can mean more and different comprehension, intelligence, and brain structure, etc.

Thus difference in these creatures and the Man will, and could be only the colour of their medium liquid system in their bodies.

The Potassium, in combination with pmtics of matters like the hydrogen and the oxygen mixture, in vacuum condition of systems in the Earth atmosphere, creates a silver greenish pmtics magnetospheric interface colour. Which, to mans' protein mixture, in the light absorption spectrum, these creatures with potassium-base proteins can be different in colour. Thus the liquid of life of these creatures in its interaction with Earth's atmosphere pmtics will create blood of different colour than the human red blood. Even in their external appearance they might be, or having to be reflecting the colour of their blood (9).

In reality as different planets have acquired different matters due to their Matters Magravs strength as their base materials, at their inception in their galaxies, rather than nitrogen or potassium as the base for what is called their protein for cell production to sustain life in their medium, in mans' journeys into deep space to these planets and zones, he will come across different intelligence, colour, sizes, transparency and shapes creatures.

Man has to understand that the level of intelligence, colours of skin or blood and shapes of these beings in the universe, can and is, intriguing as is in his home base with his red blood of the Earth to them.

What is important is that these creatures, with their pmtics balance protein chain could have a different viruses tolerance passive magnetic fields levels than to the Man in the Earth atmosphere pmtics condition too.

From this point on in the journey of the Man in the universe, the prejudices of colour of skin, and the nationhood will have no meaning. From this point on the colour of the blood is only the means of indication of the origin of point of life reference and nothing more and nothing less.

In the short-term for the future one can envisage:

The implications and applications of the transition of Matters will be the key to the survival of Man in the hostile environments of the universe. If this technology is understood fully, Man will be able to conquer the control of most of the forces within the universe and use them to his advantage. The use of the *transition states of Matters* are such a powerful tools for science for the future, that if their applications are harvested in full, most of diseases of the present and future can be eradicated. Through the appropriate use of this knowledge, Man can use the pmtics strength to free himself from most ailments, by producing the right pmtics strength equal to his protein level configuration, to ***reset*** *the cells in his body* to their right and original pmtics strength of operation without swallowing a single tablet.

For example, Man can produce ***enough food*** from the initial plasmas of the universe, without the need for soil, by simply converting the initial plasma to the atoms and molecules of the proteins and vitamins that he requires, at the point-of-demand, anywhere in the universe. As the food, itself is made of and is from the conversion of different dynamic pmtics strength. Where, the planets and animals do the conversion for man, for them by absorbing the pmtics strength of the Sun, and converting these different pmtics strength to vitamins and proteins for mans consumption for his survival. Where plasma dilution reactors can do the ***direct conversion*** of what animals and plants do for production of pmtics strength in levels that can be used by humans, and have the same effect and gives the same pmtics amount and strength needed by the different cells in the human body.

At the same time by the use of plasma dilution reactors, there will be no need to carry all of Man's physical needs with him into space, like oxygen, water, food or matters for his colonisation of planets, and space. Man will be able to convert *Matters-to-matters* through the fusion technology using plasma dilution reactors in space as he travels, and produce what he needs at any given point in space (Fig. 56).

Man can produce ***any*** *material* from the initial plasmas absorbed from the environment of space, he can produce any Matter or matters at the point-of-demand, this being an atom of hydrogen, oxygen, water, protein molecules, leafs' of gold or sheets of steel. The crafts of the future will not need to carry any raw basic materials for the production of parts into space. Man can produce materials for the construction of permanent habitation zones at any point in space and any planetary environment, through the use of dilution plasma and Grapos reactors.

Systems can be developed, which even in possession of the Dark Matter gravitational field forces, the same system can produce balanced pmtics strength in respect to its environment, for the system to appear as the source of Dark energy.

Systems can be developed, which have the ability to produce internal Magravs forces that can be tolerated by the human body. However, at the same time, the same system is being able to generate a balanced external pmtics strength in respect to its surrounding environment, these pmtics created by the system can be matching to the environment so that they do not have to interact with other systems pmtics forces (like those of the planetary systems' magnetic fields forces), so the pmtics around this region of the craft can have the appearance of systems that can produce Dark Matters and Dark energy.

However in one instance, by creating the balanced pmtics in respect to its environment, the systems that have been physically visible in one position in the given space, without a change of position, suddenly become dark or translucent in respect to their environment. In these systems and condition of operation the interaction between the balanced pmtics strength of the system and the fields of the planetary system can be brought to be matching without the system ever moving, this giving the impression that the system has moved to another position.

New systems can be developed that can produce ***instantly*** as much *energy-on-demand*, at temperature and pressure independent conditions, anywhere in the universe.

In possession of this new knowledge and Technology: ***Man can free himself from the shackles of the Earths gravitational field forces.***

Through the understanding of the ***methods of creation and control of pmtics***, Man can produce its own controlled *protection Magravs* magnetospheric conditions within and around the crafts, with the use of the plasma dilution Magravs positioning reactors, these magnetospheric field forces strength will be similar to the Earth Magnetic field strength (Fig. 62) or similar to the magnetosphere of an electron around its proton (Fig. 63). Where, the Matmags of transport for the crafts of the future can be gathered from the plasmatic magnetic environment of the universe.

The Magravs positioning systems (Grapos) can be used as much in a plasma, on Earth, in Space as they can be used in deep liquid environments, and galaxies, irrespective of the pressures and temperature around the craft.

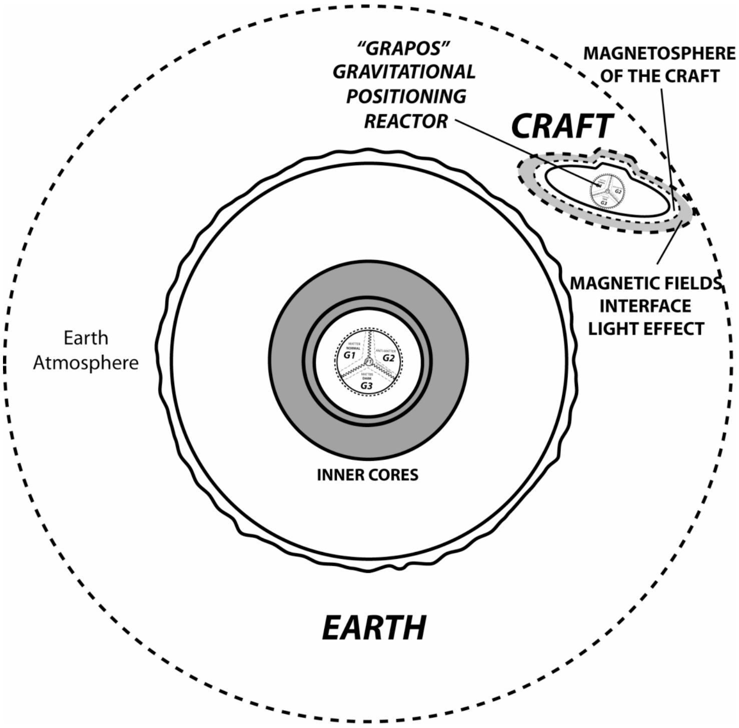
"GRAPOS"
GRAVITATIONAL
POSITIONING
REACTOR
MAGNETOSPHERE
OF THE CRAFT
CRAFT
MAGNETIC FIELDS
INTERFACE
LIGHT EFFECT
Earth
Atmosphere
G1
G2
G3
INNER CORES
EARTH

Fig. 62: The plasma dilution Magravs positioning reactors used in a craft due to the interaction of their plasmatic magnetic fields can lead to the creation of shining silver light in the atmosphere of the Earth.

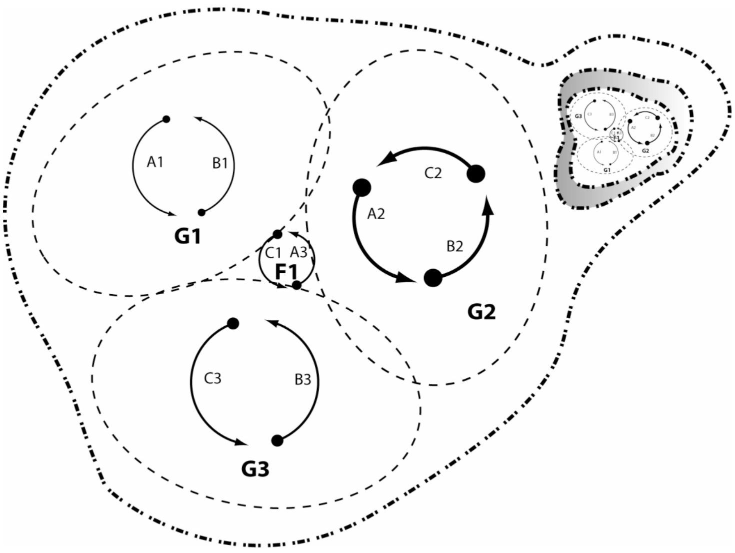
A1
B1
G1
C2
A2
B2
G2
C1
A3
F1
C3
B3
G3

Fig. 63: Initial fusion process for the production of matters and creation of bright magnetosphere of the electron as it approaches' and is amalgamated into the plasma of the proton of the atom.

Our hope, is that this technology cannot and will not be allowed to be utilised for destruction, killing and damage to any environment or against any of God's created

Matters and creatures, whether in possession of intelligence or not.

God help Man with this knowledge so that Man will find peace within himself and others.

As Mans weakness, his greed and his hunger for power and control, soon he will come to be thought the lesson that there is no forgiveness for the ones' whom misuse this knowledge and the technology for harming others.

By understanding in totality this new knowledge and what it can bring, maybe now Man will understand the meaning of that there are balances in the overall construction of his initial fundamental materials and that their lives always have to be within their limitations, of these balances.

Mans' fundamental construction of his initial state magnetic field structure, is always the same in force and energy and specific in its initial components, as the initial fundamental components of Matters, or matters of the universe ***cannot take more than what they need*** to be constructed of the pmtics of the universe and balance enough to share their inner constituents pmtics to pass on what is needed by others' configurations for them to exist, even if this giving can mean and can bring about their own demise.

Thus every creation, in its essence, has the attributes of its creator. Therefore, with this new knowledge and the reality of its existence, maybe Man can strive to understand the truth about himself and purpose of his creation in the universe.

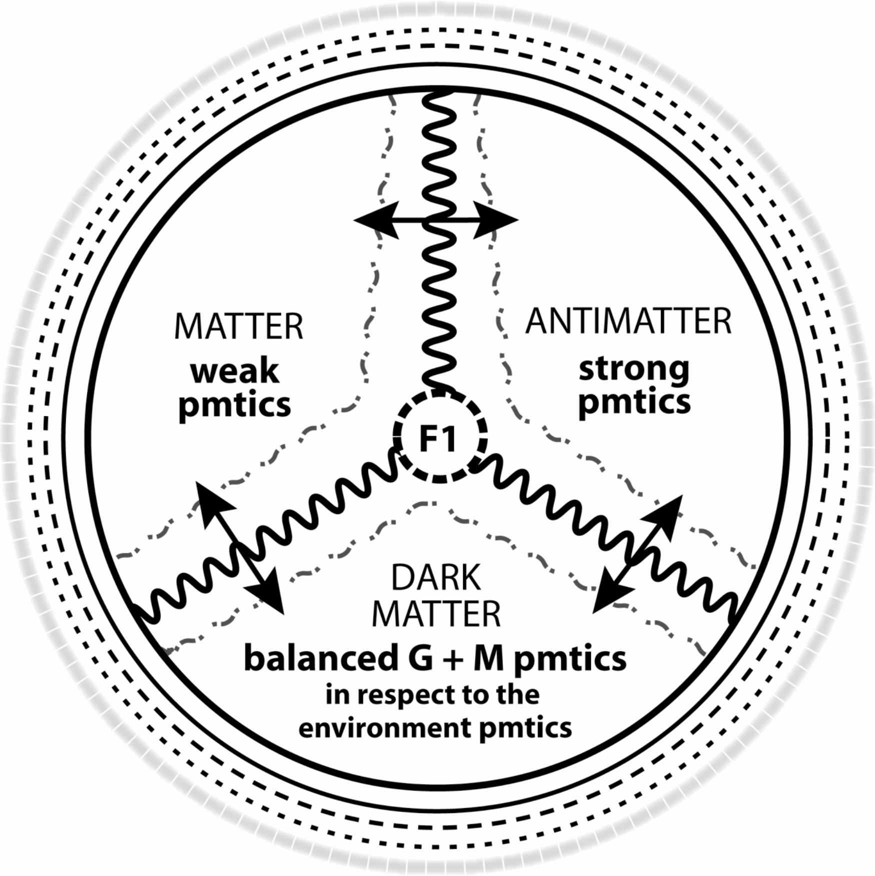
MATTER
weak pmtics
ANTIMATTER
strong pmtics
F1
DARK MATTER
balanced G + M pmtics
in respect to the environment pmtics

Fig. 64: The fundamental insight.

In general, I can say,

“God has been the creator of All,

and Man the converter of some”.

References

1: **The inner-inner core of Earth**
Don L. Anderson*
Seismological Laboratory, California Institute of Technology, Pasadena, CA 91125
http://www.pnas.org/content/99/22/13966.full.pdf

2: Researchers confirm discovery of Earth's inner, innermost core
Public release date: 10-Mar-2008
James E. Kloeppel
kloeppel@uiuc.edu
University of Illinois at Urbana-Champaign
The work was funded by the National Science Foundation

3: The innermost inner core of the Earth: Evidence for a change in anisotropic behavior at the radius of about 300 km
Miaki Ishii* and
Adam M. Dziewo□ski *Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, MA 02138*
The National Academy of Sciences

4: The inner inner core of Earth
D. L. Anderson
Proc. Natl. Acad. Sci. USA 2002 99:13966-13968

5: Robust Normal Mode Constraints on Inner-Core Anisotropy from Model Space
C. Beghein and J. Trampert
Search*Science 2003 299:552-555*

6: Fermilab
http://www.fnal.gov/

7: Fermilab Finds Antimatter Clues in Quick-Change Particle
http://www.particlephysics.ac.uk/news/news-archive/2006/new-clues-in-Antimatter-mystery.html

8: Fermilab probes Matter- Antimatter transitions
http://physicsworld.com/cws/article/news/24639

9: The universal order of creation. (Written 5.7.2006)

10: **Cosmic Rays** (Release date 24.3.2004)

11: **The Airborne reactor** (Release date 25.3.2004)

12: The seed of the Earth (Release date 25.3.2004)

13: **The creation of Black Hole** (Release date 25.3.2004)

14: **Magnetosphere** (Release date 9.6.2004)

15: Correction to the Einstein equation of Relativity (Release date 15.6.2004)

16: Introduction to the new system (Release date 2.7.2004)

17: The rings of Saturn (Release date 2.7.2004)

18: The creation of gravity (Release date 11.7.2004)

19: Death of a star (Release date 26.7.2004)

20: Fusion (Release date 28.7.2004)

21: Life of a Cell (Release date 28.11.2004)

22:The Atom (Release date 19.12.2004)

23: Magnetism (Release date 8.1.2005)

24: The creation of the magneto-gravitational force (Release date 5.2.2005)

25: The core of the Earth (Release date 20.4.2005)

26: The electron-atomic welding (Release date 3.5.2005)

27: Shutdown and safety of the reactor core (Release date 6.6.2005)

28: Reactor Start-up (Release date 9.6.2005)

29: The energy balance of the reactor (Release date 29.6.2005)

30: Nuclear decay or half-life (Release date 20.7.2005)

31: The decontamination system (Release date 10.10.2005)

32: The difference in atomic and molecular structure under gravitational force and pressurised conditions. (Release date 15.8.2005)

33: The Defence and shielding system (Release date 4.9.2005)

34: Method of creation of an atom or nano Matter (Release date 11.10.2005)

35: The relationship between Gravity and Mass (Release date 21.1.2006)

36: The Dark Matter The Fifth state of Matter (Release date 21.1.2006)

37: The inter-atomic fusion (Release date 3. 5. 2006)

38: The conductivity of the Matter in the vacuum of the interplanetary mediums (Release date 21.1.2006)

39: The production of Graphene (Release date 25.7.2006)

40: **Mini-magnetospheric plasma propulsion (M2P2)**
(http://en.wikipedia.org/wiki/Mini-magnetospheric_plasma_propulsion

41: Inner-inner core articles:

A: Professor Guy Masters Scripps institute of Oceanography
(http://www.solstation.com/stars/earth.htm)

B: BBC Earth inner inner core
(http://www.news.bbc.co.uk/2/hi/science/nature/2290551.stm)

C: Radioactive potassium may be major heat source in Earth core
Work founded by National science Foundation and department of energy
http://www.physlink.com/News/121103Potassiumcore.cfm

Published and unpublished works by M. T. Keshe

There are a number of unfinished articles that are to be converted to full disclosures when completed, like Antimatter method of energy releases, Nano space technology, relation between the creator and the created, space communication system, deep space defence technology, dynamic magnetic fields force regions (wormholes) construction and control, creation of magnetic fields, and so on.

Some of these works have already been published through different channels and some have not yet been released.

In the following section some of the works done by the same author are mentioned, with a brief disclosure of their contents.

Some of these papers can be purchased and obtained through direct request and access through specific links.

A number of papers with * and ** marks are not for public release due to their sensitivity of information and only available to relevant commercial and governmental organisations for peaceful applications of the technology.

The universal order of creation

(5.7.2006)

In this disclosure the main construction principles and the method of the link and control in the universal order of the existence of living cells, like the function and operational method of any matter based on so called proteins chains, as for example in human cell are explained in detail.

REF 9

The author considers this paper as one of his most important and fundamental works to be published and to be added to the sets of disclosures written by him.

Cosmic Rays

(Release date 24.3.2004)

This paper explains the source and the origin of cosmic rays and the way these work within our cosmos, their function and their benefits.

REF 10

----

*The Airborne reactor

(Release date 25.3.2004)

This paper covers the total design of airborne reactors for space technology. This covers all aspects of operation, design and control of such reactors.

REF 11

The seed of the Earth

(Release date 25.3.2004)

Some extracts from this paper:
The theory that the heat of the core of the Earth is due to inertia of the material pilled on top of each other is completely dismissed through this new disclosure.
The author M T Keshe believes that this centre core is made of hydrogen, other gases and a mixture of liquid and solid Matters, which were in the vicinity of the centre of the core of the Earth at the time of its inception in the solar system.
Through this new understanding of the new inner cores of the planet, this brings to play the fact that the old inner core of the Earth that is already assumed to be made of a solid piece itself becomes a sealed container for the new inner core.
The author through design of a parallel system has set to confirm the validity of this assumption, that the centre of the Earth possesses a *semi- fusion atomic reactor* and not a fission reactor as has been assumed by other scientists.
This theory of hydrogen in the centre of the planet is in line with the physical reality of the creation and the present situation in the physical construction of the solar system.
By the laws of physics all the gases in the solar system should have been in the giant gas planets at the outer layers of the solar system, like in Jupiter and Saturn.
The physical reality in the solar system, is that the lightest of all these gasses, the Hydrogen, has taken its place right in the centre of the solar system in its star, the Sun.
It has to be recognised that the centre of all planets in possession of gravity due to the centre cores heating, are always in possession of two Magnetic Field Forces in their inner sanctum (M. T. Keshe " The creation of gravity").
In the case of the Earth up to now and with the present knowledge there has been acceptance of only one inner core and one magnetic force.
As it is explained in the paper titled "the creation of gravity", there is an inner core in the inner core. This inner core, due to it physical content, motion and position in the centre of the Earth, creates and maintains its own magnetic field force, independent of the magnetic force which has been known to be created by the interaction of the inner core and the outer core of the planet.
These two magnetic fields interact to create the planet's own double magnetic fields in its centre. The interaction of these two magnetic fields forces upon each other leads to the new concept of the "DOUBLE MAGNETIC FIELD EFFECT", where the interaction between these two field forces in the centre of the planet leads to the creation of the gravity and the magnetic forces of the planet (Paper: The creation of gravity).
This new second core is the seed core of the planet, the mother seed of the planet

positioned in this new inner core.

REF 12

The creation of Black Hole

(Release date 25.3.2004)

In this paper it is explained that the Black Hole is created out of a natural sequence of events, and that the existence of the Black Hole in a galaxy is a normal event. Further it is explained that a Black hole is created in similar ways to the Dark Spots on the surface of the Sun. In this paper for the first time, it is explained how the Black Hole is physically created and what function it plays in its galaxy.

REF 13

Magnetosphere

(Release date 9.6.2004)

"The shape and the strength of the magnetosphere of a planet is exactly what a fingerprint is to a man. It is unique to each planet, star and galaxy. It is a tell tale of all its attributes, and it reveals the hidden mysteries of the internal materials' structure".

REF 14

Correction to the Einstein equation of Relativity

(Release date 15.6.2004)

In this paper the physical reality of the Einstein equation is considered, and efforts are made to bring this equation to the real conditions. Even Einstein considers this equation applicable for very small mass, and not in real three-dimensional multi-Magnetic and gravitational conditions, which are external and have nothing to do with the theoretical object's mass and speed, but have effect on the speed and the mass of the object under consideration.

REF 15

----

*Introduction to the new system

(Release date 2.7.2004)

In this paper it is written "It is essential to understand that, with the application of this system of motion, one cannot bring into play the solid, liquid and gas fuel method to create motion, as is used today. What is offered in this system is in simple words a fully integrated plasmatic magnetic energy system. Which creates within and around the core environmental conditions as known to man in the universe, like creation of gravity and magnetic field protection the same as the magnetospheric protection condition of planets and motion through interaction of magnetic fields as planets and stars do in their given environment".

REF 16

The rings of Saturn

(Release date 2.7.2004)

One of the oldest puzzles in the world of astronomy has been the existence of the rings around Saturn. How did these rings come to exist the way they are now? How did they come to appear and behave the way they do? These questions and much more about the creation of these rings are all answered in this paper.

REF 17

The creation of gravity

(Release date 11.7.2004)

In this paper it is explained that "the source and the creator of the magnetic forces and the gravitational forces are due to the interaction between the same materials in the same region of the planet."

Further it is explained how the gravitational field force can be created, replicated and controlled within a nuclear reactor. Further it is explained how this property can be used for the motion of a system within a planetary or solar system.

Now by understanding the principal of the creation of gravity, in this paper it is further explained that "the gravity in reality is the effect of the interaction of **two plasmatic magnetic energy fields** of any two object in respect to each other."

REF 18

Death of a star

(Release date 26.7.2004)

In this paper the death of stars is explained, "The death of a star and the creation of a supernova is in reality the same as the half-life cycle energy step-down of an atom. With the difference that this show is in a larger scale, and more spectacular, in all aspects of its dimensions".

REF 19

Fusion

(Release date 28.7.2004)

In this paper the physical and scientific problems with present fusion technology is considered. Where it is written, "nuclear Scientists, with the fusion reactors of today, are trying to create fusion in the opposite way to the known laws of physics and their parallels in the universe”. Further in this paper it is written, "if the present path for creation of fusion is followed, with the present scientific knowledge of physical materials for construction of such system, then following the present path of development to create energy from current fusion systems *will be a distant dream*".

REF 20

Life of a Cell

(Release date 28.11.2004)

Extracts from this paper; "The reality of the transportation and transmutation of a cell is not far from the reality of the life of an atom or a star. Where the life cycle of the cells are slightly more complicated, as they contain other Matters like acids, which have their own magnetic characteristic of their on chemical structure. For this reason the control and replication of the energy of the cell is very much complex, but very simple to achieve".

REF 21

The Atom

(Release date 19.12.2004)

The content of this paper explains how and where atoms in the universe are originated from and how the construction of atoms in a cold plasmatic nuclear reactor can be replicated.

REF 22

Magnetism

(Release date 8.1.2005)

In this paper it is written, "The question is where did the Magnetism come from? This is the Secret of creation. The world is made of one thing and only one thing and that is Magnetism. Its combination and interaction with its own different strength fields makes it appear as different Matter. The rest of creations are the outcome of these interactions of different strength of this thing called magnetism and the forces it creates. Magnetism is the origin of existence and the real singularity in its full meaning".

REF 23

----

*The creation of the magneto-gravitational force

(Release date 5.2.2005)

The creation of magneto-gravitational field force is and will be fundamentally independent of temperature and pressure, proviso to the fact that as long as a homogenous free floating PMF within a Matter is created.

REF 24

The core of the Earth

(Release date 20.4.2005)

One of the original assumptions in developing the new Magravs system some thirty years ago by M.T. Keshe was on the basis that the centre of the Earth possesses addition cores, which create the necessary gravitational conditions of the planet. After some twenty years of space technology and through Earthquake seismological data's this assumption has been shown to be correct and after over twenty years, it has been scientifically proven that the inner core of the Earth has an inner core of 600 km. Professor Guy Master in 2002 ref 24 nuclear planet. The centre of this core is assumed to have an inner core of 8 Km, this he thought to be made of the plutonium or other nuclear materials

But the developer of this new plasma reactor technology has proven through physical dynamic cores that these cores possess a mixture of hydrogen plasmas and the process of the heating of the centre of the Earth are due to a semi-fusion chain of events.

REF 25

----

*The electron-atomic welding

(Release date 3.5.2005)

Some extracts from this paper:
"The electron atomic welding or atomic welding principle is the phenomenon that atoms of the same material, through a common shared electron, become a magnetically balanced molecule of the same Matter, with the difference that the shared electron will create a balanced PMF in the molecule".

REF 26

----

*Shutdown and safety of the reactor core

(Release date 6.6.2005)

In this paper the safety parameters in the operation of the gravitational and energy reactors are simply explained to avoid the loss of plasma and gravitational field forces created within its cores.

REF 27

----

*Reactor Start-up

(Release date 9.6.2005)

In this paper the systematic start-up of the dynamic reactors for the creation of gravitational field forces and the production of energy are set out. The start up of these systems are totally different then the prior arts known in the nuclear industry.

REF 28

----

*The energy balance of the reactor

(Release date 29.6.2005)

Some extracts from this paper:

The energy balance of these types of reactors is not as simple as might look in the first instance. The energy creation is simple. But the heat leakage and dissipation through the deliberate inherent material design are and can be zero. That is to say the system not only can hold on the heat it creates. At the same time due to the close circuit feed operation will become self-sufficient but long lasting.

In planets, the loss of heat through their surface creates the variation in the central core condition over several billions of years. In this reactor as losses can be literally negligible, the system can operate at low energy loss temperatures. Even the boundary of the body could be made through the right combination of material in the chambers, so that the body of the reactor will be colder then its surrounding boundary. So the system will not only lose energy to its surrounding, but it can be made to absorb heat in the negative temperature gradient from its surrounding, so that there are no losses.

REF 29

Nuclear decay or half-life

(Release date 20.7.2005)

Extracts from the paper:

"The nuclear decay always has the same or near the same time duration. This is due to the fact that the all neutrons and protons, from their inception, carry a predetermined level of plasmatic magnetic energy. This energy is continually used for the motion and vibration of different elements of the nucleus in respect to each other and the surrounding Matters of the atom. As the space and dimensions of a nucleus of an atom always obeys the same principal of the magnetic separation and magnetic attraction. Protons or neutrons can and will need to use or consume the same energy before they become weak enough that the splitting or decay becomes a clockwork job. The nuclear decay is the natural levelling done of a nucleus’ energy, due to the energy consumption and heat dissipation for the motion of the elements of the nucleus for its remaining plasmatic magnetic fields energy necessarily needed for them to hold one together".

REF 30

----

*The decontamination system

(Release date 10.10.2005)

"What this means is that, for example in trying to recycle the CO_2, by the use of the right Matter in one of the cores of the reactor, the system will then use the O_2 to produce H_2O, in the form of pure water, and C in the form of atomic or molecular carbon, or even by feeding this back through certain operational compressions and gravitational interactions within a small core, to create industrial diamond or graphite, for different industries."

The use of the reactor for this purpose has been proven to be correct and Raman spectroscopy has proven the concept and physical reality of the separation of Matter in atomic level after separation in the same system from a composite Matter.

(See the graphene on the technology web site)

REF 31

The difference in atomic and molecular structure under gravitational force and pressurised conditions.

(Release date 15.8.2005)

In this paper is written that "the atomic and molecular structure of Matter is totally aligned and positioned in a fundamentally different configuration where it has come together due to gravitational forces, than when Matter has been brought together by the means of pressure. This has a fundamental effect on the behaviour of the structure and properties of the Matter which is created in either way". Where this is explained and utilised to develop a new method in creating graphene and Sp3 atoms and walls in a simple core made of a coca- cola bottle.

REF 32

**The Defence and shielding system

(Release date 4.9.2005)

Some extract from this paper:

The design of the side open system for lunch of high saturation magnetic plasma package is one of the most effective technologies for defence of the reactor and the

craft, which is in possession of such a system.
It is essential for any space technology in the open environment of the universe, if the craft is to keep a straight line of motion and to be able to protect itself from any solids or Matters in its path in the cosmos.
This technology can destroy any object by the principal of plasmatic magnetic energy saturation of the incoming object at a molecular level, that object will disintegrate in the form of magnetic vapour and not atomic level destruction, before Matter can damage the craft as debris, and before the craft ever becomes in touch or near these space Matters.
This technology can also be used to protect Earth from the collision risk of NEO's (Near Earth Objects), like comets and asteroids.

REF 33

Method of creation of an atom or nano Matter

(Release date 11.10.2005)

The manufacture of an atom of any density, this being of the simplest form of nucleus or a full atom of hydrogen, or even heavier elements in the universe, all follow the same principal.
Atoms are created in the soup of the cosmos from collections of very weak plasmatic magnetic fields, which for their existence cluster and share their energies with the magnetic fields that are not far from their own plasmatic magnetic field strength.

REF 34

The relationship between Gravity and Mass

(Release date 21.1.2006)

Extracts from text of this paper:
"A planet like Earth possesses a combination of both gravity and inertia. Where the gravity comes from the interaction of plasmatic magnetic energy fields from the processes within its core, and the inertia comes purely from the collection molecules PMF of Matters which has built the physical body of the planet.
Any object, being an electron, atom, molecule or even a human body, is a collection of plasmatic magnetic energy fields of different strengths and their interactions between each other, which in total decides the total or collective magnetic field possessed by that object where this is the mass of the object.
Where the mass of an object is a collective package of plasmatic magnetic energy field tightness of a given object and will not change as long as the object is intact as one entity in its overall atomic or molecular or collective molecular appearance.
Which plasmatic magnetic energy of a physical body (so called the mass of the body) in interaction with molecular or atomic magnetic field (this being gravitational or inertia) of another object will determine the weight of the two object in respect to each other in each others environment".

REF 35

The Dark Matter

The Fifth state of Matter

(Release date 21.1.2006)

Extracts from text of this paper:
The existence of Dark Matter is not questionable,
Proving its existence, and its effect on the energy balance of the universe, in all Matters as small as in the heart of an electron or in a solar system needs to be created, replicated and its effect measured.
The theory of Dark Matter does not and needs not to be considered as complicated, if one understands in real terms the plasmatic magnetic energy of the Matter in the visible and invisible dimensions of the universal work.
The Dark Matter has two distinctive characteristics, which makes it totally apart from the visible Matter.
The Dark Matter possess mass but not visible light, where its existence can only be determined by the weight of the its hidden mass, which could be substantial due to its internal gravitational field.

REF 36

----

*The inter-atomic fusion

(Release date 3. 5. 2006)

Extract from this paper:
"The fusion of two or more atoms has been the pre-occupation of nuclear physicists for past years.
Conditions to achieve fusion of two hydrogen plasmas in the TOKAMAK reactors have taken years to teach scientists a lot about the behaviour of plasma and their fusing.
The reality about the fusion in someway has to be reconsidered as, if the scientific world is trying to amalgamate the contents of two plasmas to release energy. There must be a simpler way to achieve releases of similar energies.

The fundamental principal of the inter-atomic fusion is a much simpler way to achieve fusion. If this principal is applied to the atomic condition of fusion, then fusion will be attained in a simple but in a much more physically realistic environment.

The explanation for the inter- atomic fusion is very simple and direct.
In the fusion of two plasma of proton of hydrogen atom, the physicists try to fuse two large plasmas, and by doing so, they try to release a large amount of energy. Where in this process enormous amount of magnetic field forces and currents are needed to bring two plasma in close proximity, so that the energy barriers in between them can be overcome for them to amalgamate or fuse.

In the inter-atomic fusion the overcoming of energy barriers does not exist, as in this method of fusion, the electron of the atom which is circulating the nucleus of the atom and is of the same origin will be encouraged to return into and fuse with the nucleus of its atom".

Where this is a more practical way to create and release manageable energies and there is no need for elaborate systems where even if the energy release will be small compared to dreamed up fusion systems of today.

With inter-atomic fusion method at least small and practical systems can be developed to handle low temperatures of the fusion process.

REF 37

The conductivity of the Matter in the vacuum of the interplanetary mediums

(Release date 21.1.2006)

Some extracts form this short paper "Therefore plasmatic magnetic energy field of elements within the interplanetary medium will cover the space given to them in their environment. Where due to their motion and their electric charges even in eV levels, they become perfect conductor, and in turn, due to their motion on their environment, they become plasmatic electromagnetic generators.

REF 38

The production of Graphene

(Release date 25.7.2006)

Repeated experiments and tests in simple static reactors, and tests in more complex dynamic reactors, prove that atomic separation and recombination of Matter like carbon and hydrogen can be reached at room temperature and at atmospheric conditions.
We have now indications - through static and dynamic tests in our reactors - that the universe was made in normal condition, which was originally nothing but packages of plasmatic magnetic fields of different strength, which were themselves nothing but areas of plasma or collections of loose magnetic fields energies. Where magnetic fields of different strength in locking to each other, by the principle of their **plasmatic magnetic energy (PME),** have caused in the first stage the creation of fundamental particles, secondly atoms, then molecules and then Matter, clouds and asteroids and then stars and galaxy's".

The big bang theory conditions have no room in the reality of the creation of the universe.

REF 39

INTERNET LINKS

Web site: www.keshefoundation.com

Large images of this book on the web site:
http://www.keshefoundation.com/book/images/

Links on the web site to YouTube movies: http://www.keshefoundation.com/youtube/

Forum: http://www.keshefoundation.com/forum/

INDEX

1

1.5 billion people 175

A

accelerators................. 16, 91, 151, 185
action and reaction 154
A-D Matter 110
aircraft......................... 57, 99, 144, 164
amplitude....................................... 141
annihilation.................................... 108
anti-gravitational............................ 162
anti-gravity 180
Antimatter 14, 32, 66, 92
  energy 92
  strongest pmtics............................ **90**
apple... 68
argon 157
asteroids 113, 126, 127
atom .. 11, 20, 24, 35, 50, 51, 65, 68, 81, 89, 106, 111, 115, 123, 129, 130, 132, 140, 158
  mass 67
atomic carbon 174
atoms... 14
attraction.. **85**
attraction field............................ **14**, 23

B

bacteria.. 120
Big Bang 58
Black Hole.....45, 85, 93, 100, 101, 111, 181, 189
blood **120**, **192**
  red **191**
blue shift.................................. **58**, 104
bread.. 20

C

carbon.......................... 27, 79, 120, 126
Caroline Core........................ **21**, 27, 29
cascade 87, 113
cascading 111, 182
Casimir effect **105**
catchments **13**, 26, 84, 87, 110, 160, 166
charge
  negative 139
  positive 138
CO_2......................... 126, 148, 154, 170
coiling 65
collection 12
  weak **13**
collective existence 14
collision.**37**, 57, 73, 75, 76, 81, 95, 106, 115, 121, 126, **164**
colonisation of planets 192
colours.. 51
controlled fission............................ 130
convection principle.......................... 22
conversion principle.......................... 66
copper.. 174
cosmic dusts 158
Coulomb barrier. **15**, 117, 119, 148, 176
cube of sugar..................................... 72
cup of tea 108
current 105
current flow 138

D

Dark Energy.................................... **98**
dark light 80, 83, 102, **103**
Dark Matter **12**, 14, 32
Dark Rings.. 93
Dark Spots.......................... 45, 93, 100
darker zone 95
darkness......................... 83, **96**, 97, 100
daylight... 78
decay 51, 130
  natural 137
  neutron 129
decay of the neutron........................ 111
dense pmtics 65
dependent
  balance 83
  density.. 85
  distance 60
  environment........................ 14, 79, 99
  forces.. 53
  friction.. 31
  gravitational................................... 84
  inertia **124**

  mass 18
  payload..................................... 154
  position 26, 84, 99, 142, 188
  size... 18
  strength 26, 53, 80, 88
diamond lattice 174
dictation of position....................... 143
different speed 66
dilution 177
  Grapos....................................... 159
  process 150
  reactors..........152, 158, 169, 177, 189
  soup....................................148, 152
  technology........... 151, 164, 176, 189
direct conversion........................... 192
disentangle................. 15, 62, 150, 158
dislocation 163
domain pmtics 60
drinking water............................... 181
DYPMFS.......................**169**, 171, 173

E

Einstein 65
electricity....................................... 175
electromagnetic................................ 89
electron...51, 65, 89, 132, **134**, 135, 140
  enegry 32
  motion................................121, 141
  path 141
  plasmas 118
elementary particle........................... 52
energy 58, **93**
energy of the pmtics......................... **58**
enough food................................. 192
equation of relativity................... 65, 71
equilibrium 58, **94**, 159
Extreme Ultra Violet...................... 173

F

faster
  than light 81
Fermilab 17, 66, 91, 158
field force.. 60
field strength.................................... 31
fields.. **11**
fission..................................**15**, 16, 37
food 5
fragments... 78
  pmtics.................................. 75, 105
  residual... 78
frames-of-reference.......................... 88
free lift... 159
friction.29, 31, 37, 57, 75, 99, 119, 153, 166, 170, 182
fundamental source 90
fused into 115
fusion............................**15**, 16, 37, 117

G

galaxies.. 52
gamma rays.................................... 173
Ganymede **165**
gap **120**, **135**, 138
Geo-Reactor 27
gluons 185
graphene 17
Grapos . **79**, 99, 126, 155, 156, 158, 160, 161, 166, 173
  crafts 167, 181
  lift and motion 79
  nuclear reactor 171
  reactors....... 64, 79, 80, 165, 177, 193
  test-reactors 91
Grasers **126**
gravitational field............................. 14
gravitational fields **12**
gravitational lensing......................... 85
Gravity **25**, 28
greenhouse gasses 171
ground energy level 121
ground level..................................... 32

H

half-life **131**
hard ball... 20
heating... 27
heavier elements 190
heavy atoms.................................. 138
Helium................................. 138, 157
hydrogen............. **13**, 81, 147, 157, 190

I

illnesses 184
impedance...................................... 139
independent pmtics 28
inertia 19, 123
initial

fundamental Particle...................... **12**
  fundamental particles...........5, **12**, **37**
  fundamental plasma................ 14, 51
  fundamental plasmas **35**
  gravitational field **14**
  pmtics.. 14
  seeds of plasma 11
  set.. 39
initial pmtics.................................... 14
inner core.. 27
interaction.. 14
inter-atomic fusion........................ 115
interface....................................... 167
interface colour.............................. 191
interlocked...................................... 123
interlocking 39
inter-transit.................................... 53
invisibility 57
irrespective 30

K

Keshe principle
  conversion energy......................... 66
  Magravs positioning...................... 17
  transition 55
Keshe Theory
  creation of Matters....................... 104
  Dark Matter.................................. 45
  Dilution.. 16
  Earth Rotation 28
  gravitational field 24
  Gravity... 25
  Mass... 62
  Transition..................................... 55

L

lift... 162
lift and motion ...17, 18, 21, 77, 91, 148, 154, **159**, 164, 166, 176, 177
light.. **32**, **84**
  Antimatter **88**
  bright............... **79**, **81**, 115, **155**, **156**
  dark 83
  Dark Matter................................. **88**
  day 76, **78**, 81, 155
  fuzzy 81
  matter **88**
  night 78
  rays.. 88
  spectrum 81, 121
  sun 120
  UV 173

M

M2P2 157
magnetic **12**, **13**
  field 5
magnetic field.......................**11**, **12**, **13**
  plasmatic 34
  weak **16**
magnetic interface.......................... 112
magnetic ray 87
magnetosphere..... **13**, 94, 99, 104, 138, 155, 158, 166, 173, 176, 186
  independent 50
magnitude difference 47
Magravs.................................... **13**, 15
  balance 17, 18, 78
  balanced 51
  collective 40
  overall 44
  plasmatic 28, 35
  positioning..... 17, 141, 143, 154, 156, 157, 159, 164, **171**
  reactors 162
  strong 17, 179
  weak 39
Magravs positioning....................... 141
mass 51
  independent 186
  plasma 186
mass of matter.................................. 65
matching numbers **137**
Matmags........................19, **31**, 72, 144
matter 14, **31**, 32, 99
Matters...........................**12**, **14**, **30**, **53**
Matters-to-matters 192
Matter-to-Matter 92, 123, 185
medicine 5
Mercury **13**
method of motion 143
minerals 163, 174
mini-plasma 110, 135
MOJHAN 168
molecules 14, 143, 146
mono-Magravs 126

mono-material **13**
MOZHAN 168

N

nanolayer 174
nano-oxides 17
nanostructure 174
NASA 157
neutron 35, 49, 65, 120, 190
Newton 26
nitrogen 120
nothing 101
nothingness 108
Nova condition 113
nuclear
  -based 15, 19
  decay 27, 37, 51
  fission 27
  reactor 27, 56
nuclear decay 129, 137
nuclear waste 171
nucleus 50, 129, 130, 138, 143

O

omelette 117
out of nowhere 109
oxygen 120

P

photon 105, 106, 108
Plasma **14**, **32**, 157
Plasma Dilution 16
plasmas 53
plasmatic **12**, **14**, **39**
Plasmatic Magnetic Field **12**
plasmatic strength 55
pmtic **12**
  environment 54
  soup 16, 102
  strength 66
  strong 92
pmtics **12**, **37**
  active 163
  magnitude 103
  passive 163
  remaining 40
  slower 66
  spectrum 163
  strength 54
  stronger 37, 92
  weak **12**, 16
  weaker 37
pmtics in motion **58**
pmtics strength 42
polarity 88, 188
  change 100
  planets 189
  same 159
Potassium 191
poverty 175
processes of transition 58
protection system 166
proteins 79, 147, 163, 173, 191, 192
proton 13, 22, 35, 51, 52, 62, 65, 115, 118, 120, 132, **133**, 140, 141, 189, 193
  accumulations 120
purification 175
push and pull 106

Q

quark 52
quarks 12, 51, 185

R

radio waves 157
rays **11**
reactor
  multi-core **15**
red shift **58**, 104
replicating 147
residual pmtics 46, 48
residuals 37
resistance 139
rings 112
rotation 140

S

saturated 182
Saturn 22
  rings 93
seed of plasma 39
SEPMAF **36**, 129
SET P1 40
SET P2 .... **43**
SET P3 .... **44**
shielding 158
singularity 58

solar system 82, 186
solid core 28
solid magnet 23
soup................................104, 150, 183
  atomic plasma............................. 117
  environments............................... 148
  galaxy.. 114
  Matters 148, 152
  plasma 119, 148
  pmtics.. 39
  principle 17
  universal...................................... 125
sp2... 174
sp3... 174
space in-between 103
space technology 19
spacecraft 57
spectrum.......................79, 84, 88, 191
  colours... 102
  electromagnetic 121
  light......................................99, 101
speed of attraction............................ 88
speed of rotation 30, **186**
spin..51, 186
squeezing....................................... 117
states of matters 14, 32
step
  first... 39
  fourth 45
  second 41
  third.. 43
strength
  same 14
  stronger 42, 85, 107
  unequal... 64
  weaker... 57
sub-plasmatic.................................. **16**
swirling 65
swirling effect.................................. 87

T

tail... 64
tests 91
the Antimatter.................................. 30
the Dark Matter................................ 30
the **F1** 46, 65, 70, 71, 111
the Matter 30
the Sun...20, 34, 45, 76, 77, 78, 90, 100, 108, 155, **186**, 192
Theory
  Gravity 25
time **54**, 71, 88
Tokamak... 117
transformation................................ 190
transition.................................... **54**, 58
transition of Matters......................... 55
translucent 56
Translucent Darkness..................... 102
transmission lines.......................... 173
transmutation................................... **54**

U

UFO **157**
uncertain.. 140
UN-report...................................... 175

V

Van der Waals 106
virtual Matter 104
virtual particles 104
viruses 184
vitamins 163, 174
voltage.. 105
VOMAFS.............................. **170**, 171

W

warp 85
warping.. **86**
warping effect.................................. 86
weak
  magnetic fields 11
weaker binding 189
Wormhole.. 47

X

X-rays.. 173

Z

zero weight 159

In this book nuclear engineer Mehran Keshe discloses in a very logical way that the initial fundamental building blocks of the universe find their origin in a cosmic magnetic soup that contains an uncountable number of magnetic fields of different strengths. In this soup magnetic fields of similar or equal strength can meet, interact and entangle into a larger dynamic spherical pack. The interaction between different dynamic packs, leads to the creation of Gravitational and Magnetic field effects and properties in their given environments. Due to the differences in their magnetic field strength each spherical pack will possess a different strength, which each magnetic field strength will be in the order of magnitude of the three basic types of Matters (Dark Matter, Antimatter and Matter).

As these three Matters being made of magnetic fields, they also interact, inter-balance and can interlock with each other, these interactions of three Matters results in a larger integrated dynamic system, called the Initial Fundamental Plasma, or the Neutron. By the natural decay of this plasma, the proton and the electron come to be generated and to co-exist, where plasma of each proton and each electron still contains the Trinity of the three Matters of the original plasma.

The discovery of these new first principles in physics will open for mankind a large number of benefits, since the various type of interactions of the three basic Matters and their field forces in their environment leads to the creation and control of matters (solid, liquid and gas).

In this book, new first principles and criterions for the creation of Gravitational and Magnetic field forces of the planet Earth are explained. Similarly it is explained how safe, simple, nuclear reactors, which have been build to replicate the structure of the planet inner cores, confirm the Keshe theory of creation of gravitational fields of the planet, as these reactors can produce fields which can cause reduction in weight, cause motion and lift of the reactor system, and produce an independent magnetosphere around such reactors. The interaction of the dynamic plasmatic magnetic fields of the magnetosphere around these reactors with the planetary magnetic fields creates light around the reactor. These light-effects confirm the Keshe's theory of creation of light through the interactions of plasmatic magnetic fields. These systems will open new ways of Space Travel.

By this knowledge it became possible to design, develop and test new low cost reactors that can dilute plasma and achieve fusion of plasmas in a simple and affordable way. So it now becomes possible to create vast amount of energies (like electricity and heat) and motion without the need of burning any fuel and without creating any waste. These reactors can create at any point-of-demand matters like air, water, food, medicine, and new materials, also at nano-level.

This new revolutionary technology can help to solve a number of global problems, like of the climate change, since Keshe reactors can be build to attract greenhouse gasses like CO2, or the global issue of water shortage and water pollution since it becomes now possible to create water and to attract hazardous elements out of the polluted and contaminated waters.

Kt
The Keshe
Foundation

The Keshe Foundation is an independent non-profit and a non-religious organization founded by nuclear engineer M.T. Keshe, aiming to bring new scientific knowledge and new technologies to mankind, and to bring new solutions for the major global problems, like famine, water shortage, shortage of electrical power supply, climate change, illnesses, by the use of new types of plasma reactors, and to give Man the real freedom to travel in Space.

ISBN 978-94-6087-002-6

The Structure of the Light

MATTER
PRINCIPAL MATTER
TRANSITION MATTER
MATTER
PRINCIPAL MATTER
TRANSITION MATTER

Kf
Editions

Nuclear Engineer M.T. KESHE

本电子书不是官方正版，

而是使用从官方购买的正版电子书，

截屏＋拼接＋OCR文字识别＋重新拼版制作。

因此难免出现错误的字或符号，

如果您希望获得原始的官方版本，

请购买正版，支持KESHE。

QQ群：30137653

Stichting
THE KESHE FOUNDATION
(The Netherlands)

The Structure
Of The
Light

ITNOTGOG

Reflection

The light has plasmatic structure the same as the plasma of the proton and electron and it is made of mixture of magnetic fields strength with difference that the light has dynamic composite magnetic field strength spiral helixial cylindrical composite matter mafs plasma structure rather than the dynamic composite mafs spherical structure of the matters the like of the plasma of the neutrons.

Copyright notice: © Mehran Tavakoli Keshe 2011. All rights reserved.
Nothing of this book may be replicated, simplified, translated and/or made public by means of print, microfilm, tape, audio-recording systems, electronics, optical and solid-state media, electronic storage systems, by network publishing or by any other ways without prior written authorization of the publisher or the author.

ISBN 978-94-6087-003-3
Version 2 – April 27, 2011

Illustrations and Layout: Dirk Laureyssens

Publisher: Stichting The Keshe Foundation
Registered Office: Jubileumplein 3, NL-6161 SR Geleen, The Netherlands
Registration Number: KVK 14089728.
Web site: www.keshefoundation.com
E-mail contact: info@keshefoundation.com

Forwords

All knowledge gathered and learned in every fields of science by man in his cycle of scientific evolution from his beginning of time and up to this point, these are like of the speck of dust of the universal wisdom in comparison to what man will enjoy from the knowledge which is to be revealed in this disclosure.

This shall not be all that the man has to know about his creation and the creator but through this new knowledge man will become close to realise how much more he needs to discovered and learn more about the created bigger universe, and our hope is that we have managed to bring man to his ultimate point of maturity about his world and the world of other creatures of universe through what is to be disclosed in this book.

In these new series of disclosures including this we will open the world of science into the realisation of the real working of the universe and how man can advance his intelligence and that this new and up to now unknown science can clear the way for bringing peace to the Man's own wondering life and other intelligent beings in the universe.

TABLE OF CONTENT

Reflection *6*
Forwords *9*
Introduction *13*
The re-naming of matters and conditions *14*

CHAPTER 1 16

THE LIGHT 16

CHAPTER 2 19

THE STRUCTURE OF THE LIGHT 19

CHAPTER 3 41

THE CONVERSION OF LIGHT RAYS TO THE GANS OF THE MATTER 41

CHAPTER 4 54

CO2 IN GANS STATE (PAPER 2009) 54
Abstract *54*
Discussion *54*
The CO2 absorption *57*
The methane CH4 absorption *68*
Observations and notes. *76*
New state of matter *77*
Implication and applications of this new technology *79*
Conclusion *81*
Additional pictures *82*

CHAPTER 5 85

THE LIGHT AND ITS INTERACTION WITH THE PRINCIPAL STAR 85

CHAPTER 6 90

THE LIGHT AND ITS PLANETARY INTERACTION 90

CHAPTER 7 92

THE USE OF THE PROPERTIES OF THE STRUCTURE OF THE PMF OF THE LIGHT FOR SPACE TECHNOLOGY 92

CHAPTER 8 98

INSIGHT INTO THE FUTURE USE OF THE STRUCTURE OF THE LIGHT 98
DEFINITIONS 106
Reference *107*
Published and unpublished works by M. T. Keshe *108*
INTERNET LINKS *120*

APPENDIX 1 121

The unifying field Theory *121*
Abstract *121*
Discussions *121*
Electromagnetism *122*
Permanent magnets *125*
The difference between conductors and superconductors *125*
Experimental point's observation *128*
Conclusion *128*
Reference: *129*
Index *130*

Introduction

With the present knowledge in the plasma technology and the new understanding of the internal structure of the plasma, the light is and can be considered to be a cylindrical version of the spherical plasma like of the proton and the plasma of the light is in possession of all three magnetic fields of the matters of the plasma (1).

The ray of light is in possession of Magravs and hence its own gravity, mass and Magnetosphere, and due to its dynamism this plasma of the light possesses momentum and hence can carry and transfer this momentum as energy to other light rays, tangible and non-tangible entities.

As the plasma of the light is moving from one dynamic magnetic field strength spherical in shape plasma to another and as each matter's mafs of the plasma of the light interacts with its surrounding environment's plasmatic magnetic fields in any matter environment, then the appropriate matter mafs components of the light and its ever present environment come to interact with the same and different matters and matter mafs in strength of its environment, this leads to less losses of the plasmatic magnetic field (pmf) contents of the mafs of the matters of the ray of light and hence allowing the light to travel in any environment at the speed of its given environment plasmatic magnetic fields strength with less friction and maximum speed in that given pmf strength. Then in matter mafs and matter magnetic fields strength environment, this property allowing the ray of light to have a speed of light in matter environment strength as is observed by man, and where in antimatter environment the antimatter light travels at antimatter magnetic field strength environment, which the speed of antimatter light is far higher than the speed of light in the man's matter plasmatic magnetic field strength environment.

Thus the speed of the light in the matter environment is not the ultimate speed of matters in the universe, but the speed of light is dictated and controlled by the environmental magnetic field strength which the light operates in.

It is possible to attain faster speeds than the speed of light in a matter magnetic field strength environment by using the magnetic fields properties of the antimatter plasmatic fields.

Once understanding this criterion and concept then at this stage one will face the magnetic strength barriers to overcome as one does in the matter environment with sound barrier.

Therefore: One has to understand the truth about the real construction of the light and its movement in each plasmatic magnetic field strength environment for the man to be able to travel the spans of the universe with pleasures of being able to observe and learn all the truth about the creation in all pmf strengths which operate in the universe.

The re-naming of matters and conditions

(Advice: Skip this section unless you need to refresh or reading for the first time or read it after you have read the chapter 1, 2 and 3 as this section can be confusing and distracting you from the main body of the book).

(Point of reference: The following section is and to be added to all disclosures for the ease of understanding as most of these have been disclosed in the book (1), and it is added here for the ones who have not read the book and are not familiar with terminologies in the works of the author).

In this literature, when using the term “mafs” for example in form of the antimatter mafs this means and refers to the magnetic fields (mafs) which this entity is made of and it is considered to be as a plasma of magnetic fields, and where the term “antimatter” is used on its own this meaning the physical tangible or detectable entity of the same. This usage of mafs similarly applies to physical matter and Dark matter too.

Further, the difference with Antimatter and matter is only in their plasmatic magnetic fields (pmf) strength and compactness of their fields. Thus, what is the magnetic fields (mafs) of antimatter they become magnetic fields of the matter as antimatter magnetic fields unwinds, loosens up, weaken in their field strength through their interaction with other magnetic fields and matters and become matter mafs and in plasmatic magnetic field strength of matter environment the matter mafs becomes tangible and become matter.

Therefore the antimatters are a more compact strength magnetic fields entities of the same of the magnetic fields of the matter, and due to their tightness of their fields, these magnetic fields of these mafs of matters they create stronger gravitational and Magnetic fields and hence they have stronger Magravs and hence upon their unwinding, they release magnetic fields and pmf which are faster, higher in strength and hence they look as to be strange strong and fascinating matters.

This means that antimatter entity of for example ten centimetres in diameter once unwinds in structure it becomes a matter of several thousands of meters in diameter entity in matter magnetic field strength environment.

Inevitably one should call the term antimatter the start matter and therefore antimatters should be renamed the Principal matter. Therefore there is nothing anti about these matters mafs and matters at their pmf strength as has been assumed in the world of science.

The Principle mafs matter itself has different degrees of strength by itself and not all magnetic fields of the Principal matters have the same magnetic field strength. Where, the principal mafs plasmatic magnetic fields strength has a spectrum of field’s strength too.

As the magnetic fields of the principal mafs unwinds, it releases and loosen in its environment these magnetic fields (mafs) of the principal matters disperses into their environment, then they lose their strength through contact or interaction and friction with mafs’ of matter and matters, and hence these principal mafs matters pmfs go for through magnetic fields strength of the Dark mafs matter strength first before through further reduction in strength, the dynamic pmf of mafs of the dark matter in motion reaches the pmf strength of the matter mafs level and hence becomes visible in matter pmf strength environment and becomes matter.

Now that the essence of existence of dark energy or dark matter can be explained in its realistic manner, and that is what has been called dark matter is in fact this is a magnetic field in transit from the principal mafs matter and matter strength to matter mafs strength and matter.

Then as energy has been defined as "magnetic fields in motion" then the dark energy in fact is transition mafs matters in motion, or what is called the transition energy.
Then the dark energy can be called the Transition energy instead of the Dark energy, as the pmf in motion as this magnetic fields from the mafs strength of the principal matter in motion to reach to mafs strength of the matter level, this is only a Transition pmf condition, and transition mafs matter phase only.

Where, the interaction of mafs of transition matters with other of the same leads to and creates the Transition gravitational and Magnetic fields of their own or create what is called the Transition Magravs and in some cases this condition leads to creation the transition matter (dark matter) and consequently sometime interaction of mafs of transition energy leads to appearance and become transition matter (dark matter) in the cosmos.

Therefore, the magnetic fields strength which makes the **dark matter** is to be renamed as the mafs of the transition mafs matter and the dark matter as tangible matter becomes the transition matter, the dark energy to be renamed to transition energy.

The plasma of the principal mafs matters exists everywhere in the universe. This being in the centre of the universes, galaxies, stars, planets, plasma of proton and even in the centre of the plasma of the electron and in the lower order pmf strength environments which have been undetected up to now by man.

In reality in the centre of what is called a black hole in the universe, there exist and is the home of the principal mafs matters and principal matter. Therefore the *black holes should be called* ***the principal stars*** in the centre of galaxies, stars, plasma of protons and other entities like these.

The choice of the term star for principal centres of matters is correct as these principal matter concentration centres are like a star in the centre for the matter world like solar system, and the Principal stars are the place for the mafs of the Principal matters and the principal matters which their fields radiate outward into their surrounding as it does with the stars. Thus from this point on in this disclosure we call the black holes the Principal stars in the universe now that we understand the real source of their mafs of this matter(s) and its principal matters.

It can be understood now that where the intensity and the strength of stars and planets come partially from the existence of the principal mafs pmf strength which has been released by the Principal stars in the galaxies. This is to say that the seed of the pmf of the galaxies and matters needed for the creation of stars in principle are from pmf of the mafs of the Principal matter. Where, *the interaction of the pmf of these mafs' of the Principal matters* leads to *creation of initial strong Magnetic and gravitational fields* needed for the creation and attraction of mafs of matters of plasmas of matters, the mafs of the transition matter and transition matters of the same region of the galaxies, plasma of stars, plasma of the proton and other entities of the same structure in the universe.
At the same time when using Magnetic with capital "M" this means the like of the Magnetic field of the earth, which itself is the plasmatic environment created by the interaction of at least two magnetic rays and indicates the outward flow of the magnetic fields from a plasmatic environment the like of the earth.

Similarly, when using magnetic field with small "m" this simply means the magnetic fields as in what we observe at the end of a solid magnet as magnetic fields of the object without its interaction with other fields.

Chapter 1

The light

The initial assumption which has been made by Einstein that the light is energy and travels with the ultimate speed of acceleration in the world is nothing but short of the lack of the understanding of the scientist about the real truth about the real different composition of dynamic Magnetic fields strength which the light is made of.

It is really unthinkable that the light in matter environment can be considered as energy as has been done by Einstein, as he has even proven that the light tends to bend when it passes near large objects like stars.

This bending of the light should have indicated to Einstein that, when energy tends to bend in and around any objects like a star, this by itself indicates that the light is and possesses both a Magnetic and gravitational entity structure. He should have understood that only Magnetic fields or the gravitational fields, which both are magnetic fields based, that these two entities can only attract or interact with other magnetic fields based Magravs entities and then in finding their position in respect to each other, this causing the positioning and not bending of the light in respect to the star and hence light bypasses such large objects by manoeuvring around them and gets or give the appearance of bending near the object as it position's itself in respect to large objects Magravs.

Hence light is not energy but a physical plasma entity like the plasma of a proton or an electron. Therefore, the interaction and Magravs positioning of two physical dynamic entities, which leads to observation of what is called the bending or lensing of the light near large objects the like of the star.

Whereas this was not so and light was an energy, and as has been explained before energy is when plasmatic fields (pmf) are in motion, then light would have been absorbed by the Magravs of most stars into their magnetic fields based atmosphere on the way of them passing through cosmos and as has been observed the light would have been absorbed by the other entities in the cosmos, and the light ray would not find its position in respect to star and bend around stars and large dynamic entities in the universe and the light ray would have never been able to cross the universe that one can now see the emitted light of stars from far corners of the universe.

Therefore as light tends to react and bend near objects like stars this by itself indicates important points that the light is a composite magnetic field based entity and if one can attract and interact with any composite magnetically based entity, then the entity has to have both Magnetic and Gravitational field forces.

When two or more plasmatic spherical magnetic fields Magravs interact then this can lead to release and creation of cylindrical Magravs of the light, which ***this transitional Magravs entity*** possesses both **mass and gravity**. Therefore, as the plasma of the light possesses gravity and Magnetic fields, hence this entity due to its structure possesses its **own mass** and its **own magnetospheric environment** too.
Where, this magnetosphere which is created by the interaction of the two the Magnetic fields and gravitational fields of the light in its interaction with other pmf of Magravs of a given environment, this leads to and creates the visible light in matter mafs and matter environment when the outer matter content of the light is in matter environment.

Hence the interaction of the plasma of the light and its environment's magnetosphere leads to release of pmf in the human protein Magravs strength, which manifestation itself as the visible light.

That is to say as the magnetosphere of the light interacts with the magnetosphere of its given environment, due to friction of the two fields*, residual magnetic fields leftover* are created, which they are manifested as visible light or other wavelength of magnetic fields according to the human eyes protein cells plasmatic magnetic fields strength levels in earth environment Magnetic fields condition and in comparable environments.

Chapter 2

The structure of the light

As an entity in possession of Magnetic and gravitational fields moves from one spherical plasmatic magnetic fields environment (for example the plasma of the proton) to another plasma, the plasma Magravs motion in between these two spherically dynamic plasmatic entities (transition plasma magnetic fields (pmf) stretches' to and takes the shape of elongated cylindrical Magravs entity rather than a spherical shape.

It can be considered that the tangible or detectable pmf are of dynamic spherical in shape entities and they possess the pmf Magravs of their own and as the pmf moves from one dynamic spherical entity to another dynamic spherical pmfs Magravs environment then the plasmatic magnetic field Magravs in transit between the two spherical Magravs due to interaction and its composite Magravs configuration and the Magravs composite of its environment becomes elongated pmf in shape from the Magravs of the spherical plasma and takes and become narrow and elongated and hence spiral cylindrical in shape.

It has been explained in detail that the plasma of dynamic spherical entities like plasma of the protons' is made of three main components of mafs of the matters, the mafs of the principal matter and the mafs of the transition matter and mafs of the matter (1).
Therefore the elongated plasma of the same must contain the same mafs and matters but in a different structural positioning configuration.

Hence the dynamic cylindrical shape plasma of the mafs of the matters in transit from one plasma to another will have the same quantity mafs of all matters of the dynamic spherical plasma and therefore the plasma in transit possesses and it is made of the same mafs of the principal matter (Fig.1, 2), mafs of the transition matter (Fig.1, 2), and the mafs of the matter (Fig.1, 2), hence the plasma of the light contains the same mafs of the matters as the plasma which is manifested from.

This dynamic elongated cylindrical plasma once in motion due to its new shape it can penetrate and move faster in its environment as it has less frontal interface and creates a lesser friction as if it was spherical in shape, this allowing faster speed to be attained by the plasma of the light in transit from one matter mafs environment to another irrespective of the plasma matter mafs of the light being in the principal mafs, transition mafs or matter mafs mediums pmf strength environment.

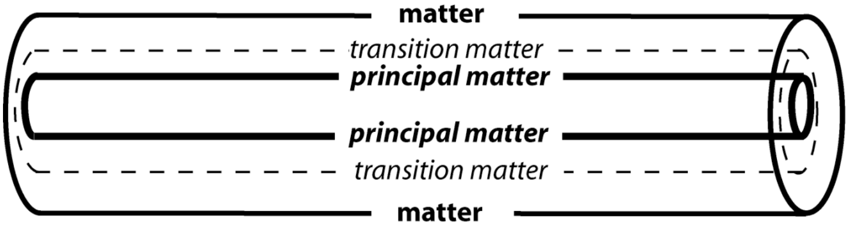

Fig.1 The schematic mafs of the matters components' of the light

This internally dynamic cylindrical plasma of the light in its motion creates friction between itself and the Magravs of its environment, which this friction at the point of their interaction and due to friction of the two pmf Magravs at this point of interface releases plasmatic magnetic fields, which are in the man's protein pmf strength detectable magnetic spectrum. Through biological structure of the humans protein pmfs strength visionary sensors' which can detect this interface interaction pmf strength then these sensors become aware of different pmf strength, its position and condition and this information is passed to the brain, where this method of interface interaction of two Magravs of the rays of the light and the protein has become the method for man to be aware of his physical surrounding in his environment pmf strength and the man has called this matter interaction conditioned wave length the light ray.

The light ray in its interaction with the surrounding magnetic fields in their environment leads to creation of magnetic field in range to man visible light according to the human protein pmf strength of the universal Magravs spectrum, where the shine or the visibility of the light observed become and it is at the boundary of the magnetosphere of the plasma of the light.

The structure of this dynamic cylindrical plasma of the light ray, which creates its own Magravs, is initially consisted of, in the centre, the content of mafs of the principal matter, then the next layer to the principal mafs is the mafs of the transition matter and then the layer surrounding this, is the matter mafs strength magnetic fields components of the light.

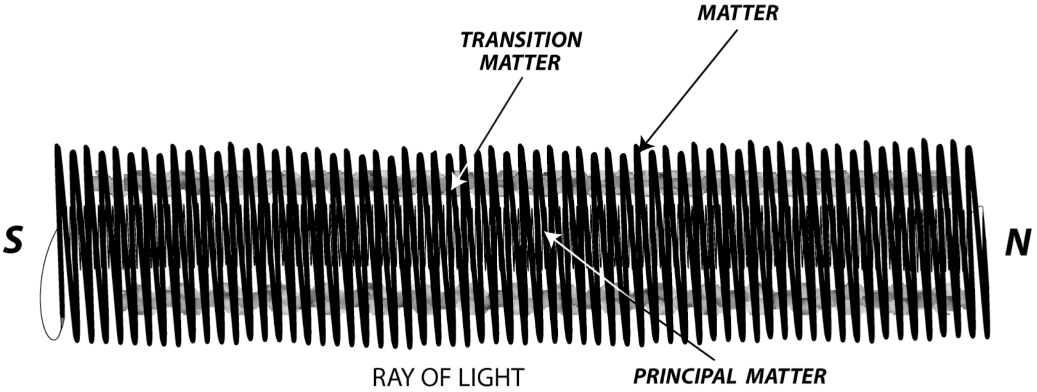

Fig. 2 The mafs of the matters components' of the light

These dynamic plasmatic magnetic fields due to their proximity to each other, they have to interact with each other as all magnetic fields in close proximity with each other do and by the same universal principle laws of physics these magnetic interactions leads to creation of gravity (1) and creation of Magnetic fields within and around this dynamic cylindrical plasma, where these magnetic fields interactions leads to creation of internal and inward magnetically initiated gravitational fields force of the light and the outbound Magnetic fields forces of the light. Therefore due to interaction of different constituent magnetic fields of different mafs' of the matters of the light, the light must create and does possess its own Magnetic and gravitational fields, and due to the interaction of these two plasmas of the fields of the gravity and Magnetic fields, this leading to creation of the mass and the magnetosphere of the ray of light (1).

Hence the light by natural laws of magnetism possesses mass and magnetosphere and similarly not all lights have the same mass and by the same laws and principles lights of different intensity and lights which can travel longer distances possess a larger mass. Where, their mass is primarily dictated by the strength of the pmf of mafs of the principal mafs of the light components'.

The interaction of the mass of the light and the environment of the medium that the light is traveling through, then the light acquires weight in respect to its given environment fields strength.

The mass of the light has different weights according to the gravitational environment which it is passing through or being produced or arrive at. In understanding the construction of the ray of the light then the mass of the light has to be measured along its total length and not across its width.

Most rays of light possess anticlockwise spiral longitudinal rotational motion for their constituent fields and not spherical rotation of the fields as for example in spherical dynamic entities the like of plasma of the proton.

Rays of light as in the case of matter plasma of protons the light rays mainly possesses and maintains a clockwise directional motion.

Due to composite structure of the light's mafs, if one looks' at the cross-section of the light (Fig. 3), one should be able to observe a tighter magnetic fields strength region of the principal mafs in the centre, leading to creation of the strong light in the principal spectrum level and stronger gravitational fields of this sections' structure that creates its own magnetosphere region, then this section is covered with the transition mafs region structure of the light, and then when the transitions mafs have reduced in their strength, then they become in the level of the matter mafs strength and at this zone the matter mafs and matter of the light are established and maintained (Fig. 3).

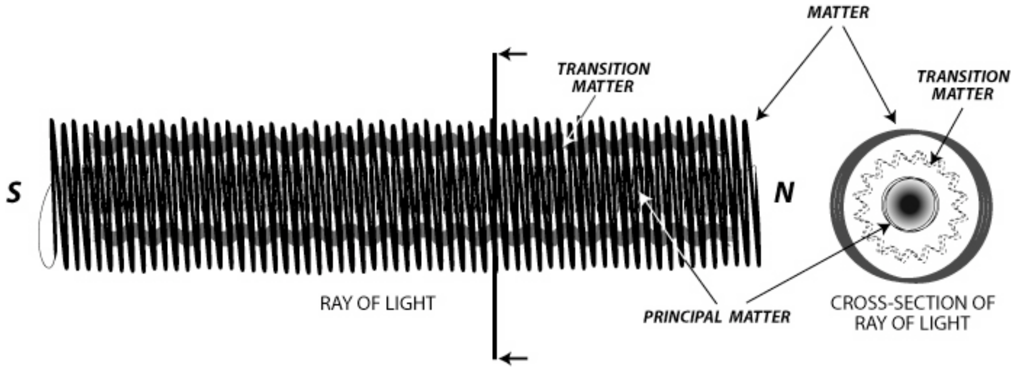

Fig. 3 The cross section of the ray of light

The lights' structure is similar to the structure of inner sanctum of the Sun and principal star in the centre of the galaxies, where for example the Principal Star takes its position in the centre of the galaxy and due to its higher inner magnetic field's strength of the principal star, this creates what appears as the principal stars strong gravitational field forces that all matters around it are pulled inward and at the same directional motion of the galaxy is set by their principal star too, where simultaneous this central matter radiates outward its own light in the principal mafs strength (3 and 4).

The straight helixal cylindrical shape of the light ray has polarities the same as any straight solid magnet, where the outer magnetic fields lines of the light ray converge inwards and the inner magnetic lines of fields diverge outwards, this creating an inward (South-Pole) and outward (North-Pole) directional motion of fields (Fig. 4).

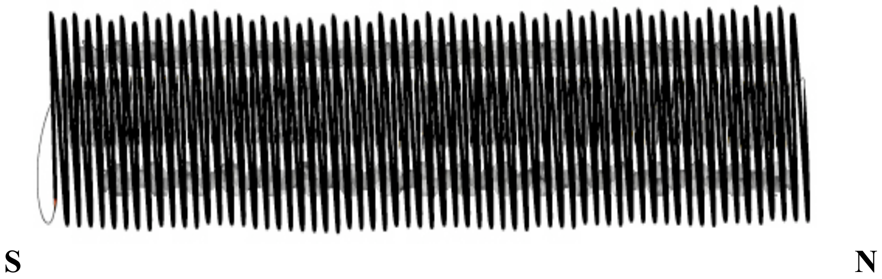

Fig. 4 The polarity of the ray of light

Due to the structure of magnetic fields of the light, it can be said that the matter mafs fields components of the light will be interacting and be attracted and drawn inwards' through the centre of the structure of light's magnetic fields at the ends of their pmf length (South -Pole of the light) and then as they go through or in proximity of the mafs of the principal mafs of the light in the centre they absorb some of the pmf, which has been unwinded by the mafs of the principal mafs fields and then these matter mafs fields, now energized, then due to their rapid motion are ejected and appearing from the other end (North-Pole) of the light ray.

This inward revolving motional central rotation of the outer matter mafs fields at the same time gives the light its property of high speed forward motion, where the conversion and the overall and total gravitational field strength and the helixial cylindrical rotation of Magnetic fields of the lights' principal mafs' dictates the length of the light ray and hence where the beginning and the end of the light ray would be.

This process of matter of light going inwards and then ejecting from other point, is very much like the process what some consider as the imaginary phenomenon commonly called and known as a wormhole principle. The continues rotation of the mafs of the matter's fields concaving-in and will go through the inner Principal Star like system of fields in the centre of the ray and then the pmf of the mafs of the matters emerges from the opposite-end as it has entered with new and higher pmf strength.

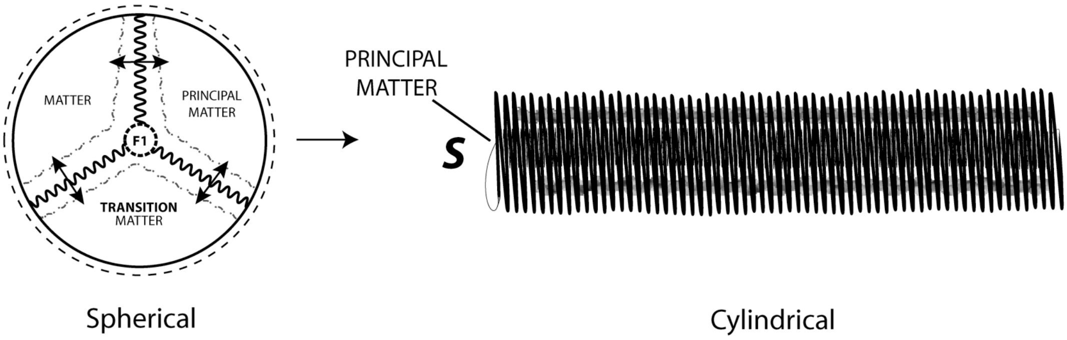

Fig. 5 The first conversion. The spherical plasma structure converts to the cylindrical plasma structure of the light (The first metamorphoses of the light)

When, the spherical plasma first changes to cylindrical plasma that is the point when the ray of light is created and becomes straight line directional in motion entity, and the spherical entity goes through the first metamorphoses (Fig. 5).

Then as and when this dynamic cylindrical entity reach's its destination point, this being of any mafs or matters plasmatic fields strength environment, then at the point of arrival then this cylindrical entity goes through the second metamorphoses and reverts back and spherical rotational fields entity (Fig. 6).

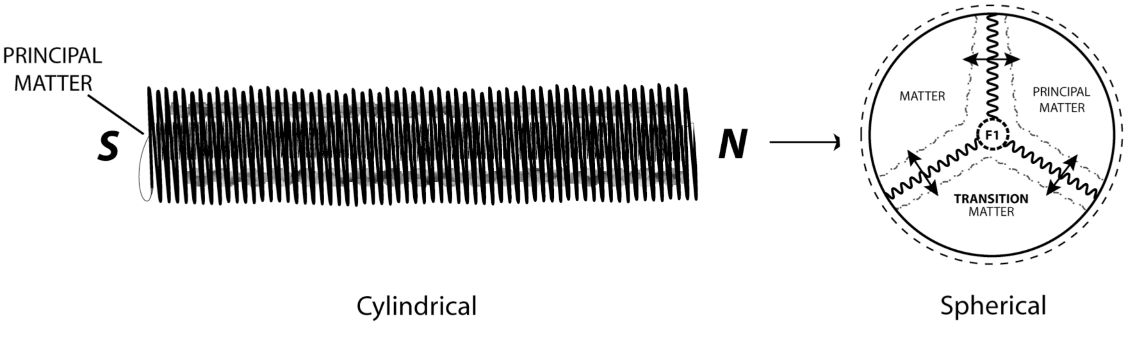

Fig. 6 The second conversion. The cylindrical plasma structure of Light converts back to a spherical plasma structure (The second metamorphoses of the light)

Where, the plasma of the light gets gravitated and absorbed then it takes the shape of the matters in the target entity environment as the content matter of the plasma of the lights gets added to the plasma of the target environment and the target entity receives an additional dynamic plasmatic magnetic field of all mafs of the matters of the light or what is called the target plasma receives an additional energy from the plasma of the light.

Therefore the plasma of the matter to hold on to its maximum pmfs and transfer its pmfs from one point in the universe to another as fast as possible, then the plasma of mafs uses and goes through double metamorphoses process of spherical to cylindrical and then to spherical entity (Fig. 7).

In reality the light ray due to its mass possesses momentum rather than light being a ray which only transfers' it magnetic fields and energy upon interaction with other entities in the universe.

Therefore, IFF the light was a ray and not a plasma composite of different mafs strength, then only one of the matters' mafs of the plasma of the targeted entity would increase in mass and this creates unbalance in the inner structure of the targeted plasma and only in that matter's content, which in the case of the matter environment this energy is so low compared to the principal mafs contents' of the targeted plasma, that this increase in matter pmf would not have shown any increase in its content and increase in the energy as the whole.

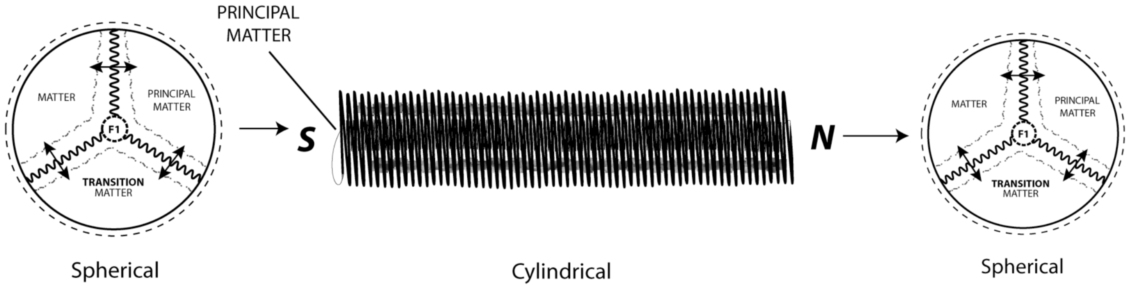

Fig. 7 The double conversion of the plasma from spherical to cylindrical, and on its arrival back to spherical plasma entity
(The double metamorphoses process of the plasma of the mafs)

But as the light transfers all its mafs of all its matters content to the targeted plasma, then the whole pmfs matters of the plasma of the target plasma increases and then the whole plasma shows signs of increase in the plasma content and for example the whole atom structure gets energised.

This being the like of when plasma of an atom receives injection of a ray of light that the whole proton and electron gets excited and energized and not only the matter components of the plasma (note: the term plasma of an atom as the whole structure of an atom including its neutron, proton and electron create its own magnetosphere and it can be considered as one plasma containing plasmas of these particles). Where, the electron due to its lesser mass shows a more rapid motion and excitation, whereas the plasma of the nucleus absorbs most of the light mafs and adds this to its total mass of all its matters.
If! The protons of the nucleus magnetosphere are measured before and after interaction of the light with its atoms magnetosphere, then it can be observed that there is a marked increase in the magnetosphere size of the plasma of the proton too, this indicating that the whole atom absorbs all matters mafs of the light and not only the plasma of the electron, which moves and takes new position in respect to the protons Magravs.

In comparison, in earth condition as meteoroids enters its magnetospheric zone, the facing side of the meteoroid on entry into pmf of the earth, its interface surface matter curves' backward, where the same happens in the case of dynamic plasma of the light in entry into a new pmf strength environment, then the same types of curvature and convergence of the mafs of outer matter mafs layer of the light due to the speed of the transition and obstruction and friction of the new environment, then the outer mafs of the light take their position on the outer layer and the rest of the mafs of other matters' of the light follow the same process and take their position within the new plasma and the cylindrical structure of the light converts into spherical plasma and hence the light converts into a tangible spherical matter, and all its matters content fuses and amalgamates with all matters of the same in the new target plasma environment (Fig. 6).

The reversal of the helixal motion and cylindrical plasma of the light to spherical and rotational mode of matter is entirely dependent on and due to the environmental plasmatic magnetic fields strength, which the light arrives at and it is dependent on the tightness or strength of the magnetic fields of the destination environment.

In this process pmf traveling from one point to another in the universe, as the fields of the mafs of the transition matters have faster speeds than the matter mafs fields of the light then, when the light arrives at the point of its destination, where is to release and converts its fields to the given environments' of other plasmas, then the conversion in the matter mafs to matter fields strength leads to creation or release of plasmatic fields in the soft x-ray which is the essence of creation of life (6) and other magnetic fields strength of magnetic spectrum levels.

In the matter environment the slowing down from the principal mafs components of the ray to matter mafs fields strength leads to release of magnetic fields plasmas in the gamma rays strength levels and similarly the beta plasmatic magnetic fields strength is due to fields strength from principal mafs to mafs of the transition matter and transition matter mafs to matter mafs and matter pmf strength this become to release alpha rays in the magnetic field spectrum.

The production of gamma ray through such similar process can be observed, when the star in the supernova state releases gamma burst, where this gamma burst is due to the rapid expansion of the principal mafs of the star converting into the matter mafs magnetic field strength or the rate of the change of the principal maf to matter and matter maf leads to release of gamma rays.

As the matter mafs composition and the construction of the plasma of the electron has been understood to be of the same as the plasma of the proton and neutron (1), then gamma rays are and can be produced by the mafs of the matters of electron the same way as done by the neutron or protons, when and where the right condition for the release of the fields strength equal to the gamma rays are made available for this ray pmf strength to be released. Where these rays released by the electron can appear in the inner structure of the plasma of the protons of the nucleus and be manifest as the incoming and mysterious energies which are moving within the plasma of proton or neutron and their source unknown.

There are lights rays which are created by the electron in the same process of spherical to cylindrical structure conversion between the pmf of the electrons and the proton or an atom.

It has been explained (2) that there is current of magnetic fields flowing within the different mafs of the matters magnetic fields strength of any plasma as it does between the plasmas themselves. Where the current of magnetic fluxes in the case of the light rays is in the light plasma magnetic fields strength and not electron levels of magnetic fields strength of the spectrum of the magnetic fields, when magnetic fields moving from one matter mafs to another within the layers of the mafs of the matters of the light. This current of pmf within the mafs of the matters' of the light is to be called the **light Magnetism**, "***ligmagnetism***" instead of electromagnetism as their mass of the mafs are totally in different scale of measurement.

It has to be said that as there are electromagnetic fields so there are *protomagnetic* field as well in the operation of universe, these are a more powerful current follow of magnetic fields package, which the world of science is not so much aware of these at the present, but in the construction of space reactors for the future of space technology the use of protomagnetic fields and their currents are much easier systems to produce and this gives opportunities for the development and production of a much smaller but more powerful by factor of several orders of magnitude than electromagnetic based plasma space reactors, where huge masses can be lifted with few grams of fuel. This concept of light to neutron and lift and motion is the main corner stone of knowledge and principle that allows the stars to create such a powerful gravitational and Magnetic fields, which the scientific world at the moment has not understood as of yet and has been entangled in the smaller component of the atom, rather than understanding the real power house of the atom and utilise this for their benefit to achieve some wonderful and interesting results.

The pmf current within the mafs of the matters of light structure magnetic field in strength has a different spectrum due to its length and rotational motion than the electromagnetism measurement. This ligmagnetisem has more powerful current flow and by size have stronger field's strength as it has more field contents in the principal mafs fields' spectrum than the electromagnetism which is in matter fields' strength spectrum.

Electromagnetism or electromagnetic flow (2) is said to be the electron magnetic fields strength equal to the strength of the magnetic fields of the plasma of an electron and the current is explained as the rate of the flow of the magnetic fields from one magnetic field strength to another (1 and 2).

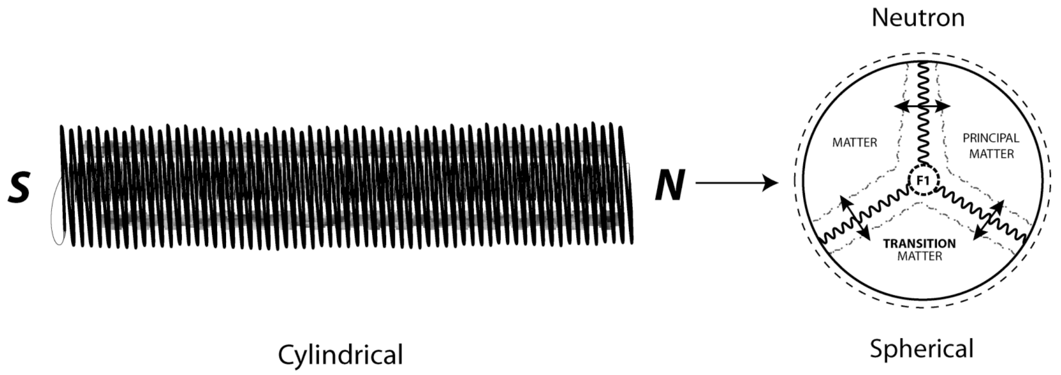

Fig. 8 Conversion of the ray of light into the fundamental plasma
or what is known as neutron

In most cases when the lights have enough mixture of mafs of different strength, these rays upon their arrival to a given environment, they directly convert into the complex plasma of fundamental plasma or what is known as the plasma of neutron (Fig. 8)(1).

At the same time collection of smaller and fragments of lights plasmatic magnetic fields upon arrives at the point of destination, due to the vacuum level of the environment go through the process of accretion and they create or lead to creation of the fundamental plasma (neutron) in their new environment.

In cases the light ray carries more mass contents of all three mafs of matters compared to the its environmental mafs need for production of fundamental plasma in this new environment, then in these cases the large in mass light rays has enough mass of mafs that the light ray in this new environment goes directly into the structure of the fundamental atom or what is known as atom of hydrogen, that once this ray. That is to say in certain conditions, the structure of atom is due to total conversion of the light rays components and the ray of the light does not go through neutron decay conversion (1) for the light to become an atom (Fig. 9).

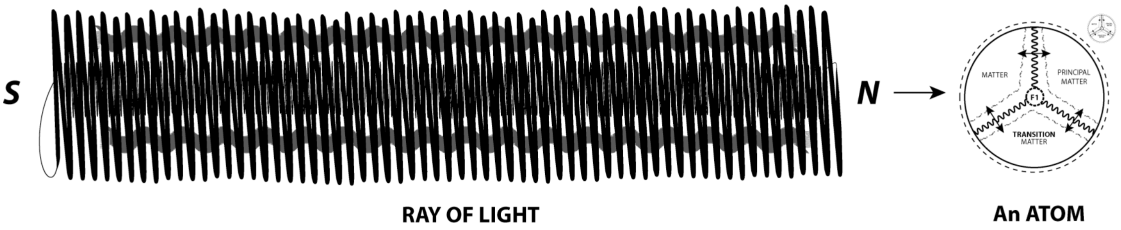

Fig. 9 Production of an atom direct from the ray of light

Some light rays due to their internal dynamism and different mafs strength in their structure, these lights rays create and contain entities moving internally and these internal structures keep the plasma of the lights' mafs in order and under control. These internal entities are the like of the electrons, which are externally positioned in the overall structure of the atom, these internal entities have specific position and purpose and they mostly convert to electrons' of the plasma of an atom in given specific conditions once the plasma of the light converts into matter of an atom, where with this process there is no conversion into neutron and then the nuclear decay of the neutron that leads to creation atom, in this process the light converts directly into atomic structure in one step. These internal entities are usually observed in the lower order environment medium which the light is traveling through, where the low speed of the entry of the outer layers fields, in-going through the faster speed principal mafs, these create small vortexes between the two fields in the transition mafs environment of the palms of the light, which these vortexes start operating independently and gradually fiend a fix position in the internal structure of the plasma of the light and when the light transfers to structure of an atom, these mafs entrapped in these vortexes become the electron of the matter of the atom (Fig. 10A).

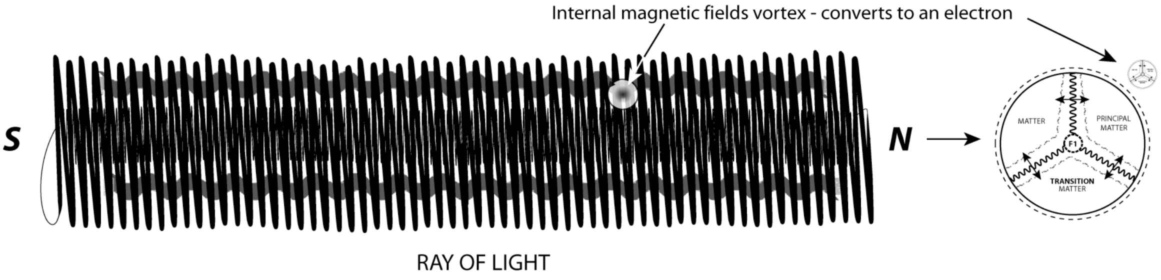

Fig. 10A The conversion of light containing vortexes into the structure of an atom direct.

The reason why the plasma of the light takes the spiral helixial cylindrical motion and shape is simply due to the fact that plasmas' with this type of structure does not have a need for additional external plasma(s) around them for coherence and a balanced plasma the like of electron in the spherical plasma structures of an atom.

In principle rays of light get their strength from their inner plasma of the principal mafs' strength rather than any other part. This being the reason, ***why light look as to be an energy rather than entity with the mass***. Therefore the light is a composite of plasmatic magnetic fields of different mafs of the matters fields' strength and **light is not a single magnetic field structure**.

The lack of understanding of the structure of the light and the inability in scientific world to be able to split the structure of the inner components of a light ray has led the scientific world to consider the light rays as an energy rather than a dynamic structural plasma as of the same as the electron or proton.

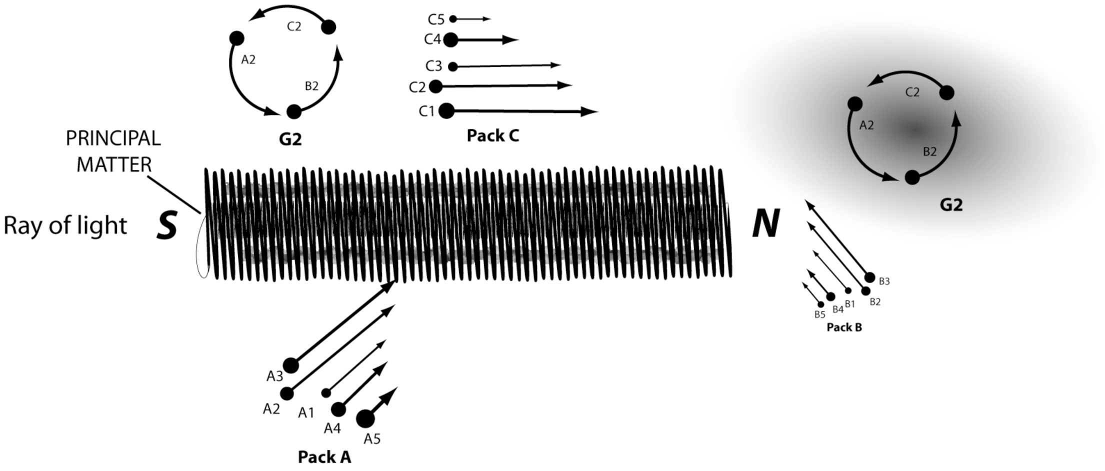

Fig. 10B Ray of light travelling through and interacting with outer mafs and initial fundamental particles of pmfs.

Due to the cylindrical structure of the light plasmatic magnetic fields and their interactions, light has the wonderful capability to be able to change its outer mafs cover as/and depending upon the matter mafs or matter medium that it is traveling through or where it arrives at.

In choosing and using helixial cylindrical motion the light does away with the need for external rotating plasmas like as the electron(s) with the atomic and planets with solar system.

This, meaning that all rotational spherical objects the like of the solar systems and atoms need to have two or more parts that their centrifugal magnetic forces can keep their matters and mafs content of dynamic magnetic fields in an encircled and enclosed loop for them not to lose too much of their magnetic fields through leakage to other plasmas in their environment.

In reality helixial cylindrical rotation partially can be seen in structure of some of galaxial structures too, primarily due to this phenomenon.

Where, for faster and simpler speed, the layers can move inside-out and vice-versa according the mediums' magnetic fields strength of the environment of travel. This meaning that due to dynamic cylindrical helixial motional and composite matters mafs structure of the light, the mafs of the transition matter plasmatic magnetic fields can become and take its position on the outer shell of the light rays in a mafs of the transition matter mafs and medium strength and similarly the principal mafs components of the light can take their place on the outer fields layers' of the light rays in principal mafs and matter fields strength environment.

The light is energized from the central magnetic fields strength region of its structure in a matter medium and this being the reason why the light seems to be able to travel long distances and behave as energy and not an entity. This change of cover allows less losses' of the mafs of all matters of the light ray due to friction in any medium magnetic field strength.

Now it becomes clear, how and why spherical rotational shape of the movement of the dynamic plasmatic magnetic fields leads to creation of matters and at the same time longitudinal cylindrical helixial rotation of the same dynamic plasmatic magnetic fields lead to creation of light rays. Therefore the method of transport of pmf and Magravs in space is through the Light Magravs, where at the point of interaction and slowing down the lights' pmfs are transformed and transferred to matters fields strength of plasma of the new environment.

Hence, in the universe mafs of the matters are transferred and transported through the mechanism and the use of plasma of the light and when for example the Magravs of the light reaches matter environment it becomes and slowed down enough to reach's all content matters Magravs strength levels, then what is released in by the light pmf, this surplus or released pmf then becomes the pmf which targeted pmf receives as addition pmf or what is commonly called ***it receives energy from the*** pmf of ***the light***.

When, the light ray reaches the principal matter environment then its principal mafs pmf will have the same interaction as delivery to its given environment as mafs of matter have in the matter environment.

The reason matters can feel forces of rays of light are primarily due to the momentum of the motion of the mass of the light itself and not through absorption of its energy.

In the solar-sail transportation the motion of these would not be due to energy of rays of light absorbed by the sail, but in real term the created motion is due to the motion of the mass of the light and its momentum that once it comes to rest at the point of interaction with the material of the sail, the light transfers its momentum to the fabric of the sails' matters and hence the increase in the momentum of the sail and hence the motion of the sail.

It is essential to realize that as there are gamma rays, beta, and alpha rays and so on in the matter world so do the same types of rays of the same in equivalent and in proportion in strength exist in the environments of the Transition and the Principal Matters too.

Therefore one has equivalents' of the gamma ray in the principal mafs environment and the same as x-ray in the mafs of the transition matter environment in magnetic fields strength too. This meaning that the sub-division and spectrum of pmf strength of the light is not the exclusivity of the matter environment only and it has parallels' in other matters' pmf strength environments.

At the same time the interaction of the Magravs of the lights surrounding with Magravs of the light leads to release of magnetic wave's rays in the visible light spectrum in these matter mafs environments depends up on the observes pmf strength visibly tools of observation.

The inter matter's mafs cover exchangeability of the light rays is the reason why light has the capability to travel long distances in mafs of the transition matter environment of the universe, as the similar in strength mafs of the transition matters creates no loses or interactions in this level that can lead to minimum losses of light's plasmatic magnetic fields and its strength. The reason being that visible light is only released, when two different Magravs in strength collide and not when two pmf of the same interact or very little pmf are released in these compatible pmf strength entities and environments encounters as in the interaction of the similar pmf lead to and creates very little friction between the pmfs of the light and the environment and hence these interactions releases very little light and this being the reason why in the universe, the interaction of the transition mafs of the light rays with the transition matters of the universe creates very little light or a dark light or what can be called the Transition Translucent light (1) and not a bright visible light, when one looks into the deep cosmos.

The invisibility of travel path of the ray of light through the deep space is in fact ***due to its***

capability to change its outer layers according to the environment that it travels through.

That is to say as the light goes through the Transition fields regions of the universe, then the transition mafs and matters of the light will come to take their position on the outer layer of the ray, then through this process less pmf are used in interaction with these fields and matters and hence less friction and less loses of mafs of the light, and in these interactions there are very little loses through the Magravs of the mafs of the transition mafs of the light, these faint lights produced in the interaction of light's outer mafs of its transition matter and of the same and/or transition matters in the universe, where these very few interfaces of the collision between the light and the environment Magravs gives the universe a dark but translucent effect that one can see through the universe. In a way this interaction of lights' transition mafs and universe's transition mafs is the cause and the reason for the invisible transparent structure which allows the production of universes discreet light that one can see through this universe but blocks opportunity that one cannot see the lights from and the universes outside this universe (3).

It can be said that the light due to its interaction and helixial rotational characteristics is like a **chameleon** and ***it has the ability to changes its cover matter and mafs according to its environmental Magravs strength***, which it travels though and when and where the fields strength changes are too great then the light changes its cylindrical to spherical shape structure and its outer skin matter and mafs take's strength of its environment mafs or the matters' plasmatic magnetic field strength. Or it can be said that the light has **surface** Magravs strength matching characteristics and capabilities. As the Magravs of the light reaches or tries to enter another spherical Magravs or cylindrical Magravs then part of the released Magravs due to friction of two Magravs with each other lead to transfer of pmf to spherical pmf and release of some pmf in the form of magnetic rays' strength spectrum.

The light as it enters or comes in contact with plasmatic magnetic environment the likes of the earth, then the interaction of their two dynamic cylindrical and spherical Magravs creates fragments of plasmatic magnetic fields that are in the range of the man's protein Magravs strength that to man's sensory detection tool called the eyes can recognise the interface of the interaction of the two Magravs as the visible light.

In reality the light is created and operates through and from different pmfs strength and it is not as simple as an energy as if one define the true meaning of the energy, which is "the plasmatic magnetic fields (pmf) is motion" (1), then taking the motion or speed of the ray of light's movement in a different matters and mafs environment then one can see that the light in truth is no one magnetic ray but it is made of a complex plasmatic magnetic fields of different strength.

Whereas, in the light's principal mafs (3) the stronger and tighter of its fields seats in the centre of the structure (Fig.1) and as this mafs of the matters field's unwinds, these newly made available which are lower in strength and are going from one strength to another, then these mafs in motion to be known as ***transition energy*** (dark energy) and when these mafs slow down enough to their transition matter and mafs environmental strength, then these appear as mafs of the transition matters (dark matter), and then when the transition mafs pmf strength reduces or unwinds enough that this mafs reach's the matter mafs pmf strength levels environment then it become tangible or detectable the matter at this strength (Fig. 11).

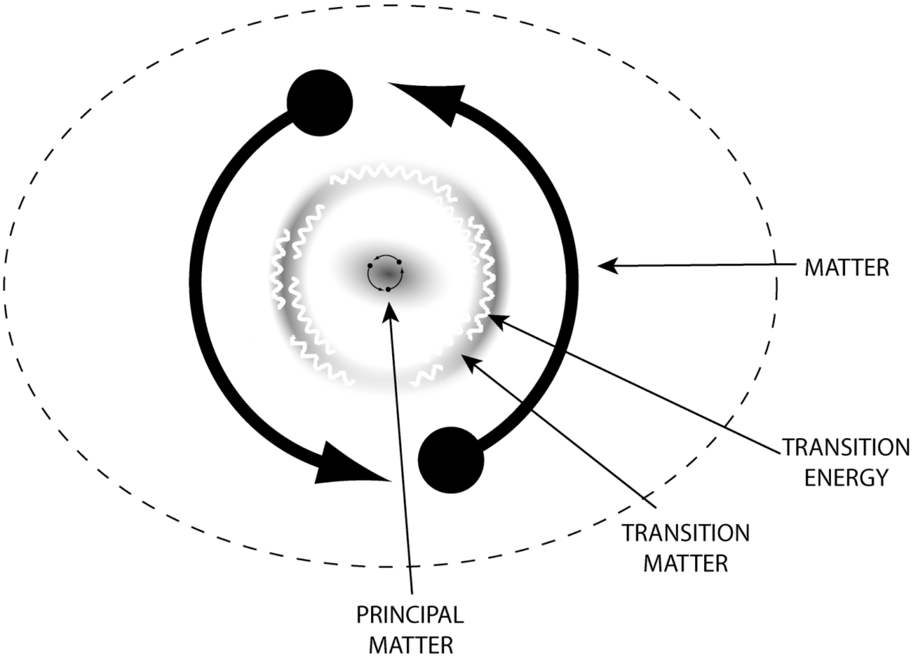

Fig. 11 The magnetic field structure of the plasma of the light

Where, the visibility of the light at this point the like as in the earth and human body, then this light is human protein chain mafs strength dependent, then it became visible light to the human eye in the Magravs strength of the planet gravitational field strength and then the light appears in matter pmfs strength and as tangible or detectable matter.

Light in its visibility and tangibility has different strength and what is visible light in one matter mafs and matter environment strength, this it is not necessarily visible and tangible matter mafs and matter's pmf strength in another environment and vice-versa. The visibility of the light or its transparency is ***environment and target magnetic field strength dependent.***

The structure of the light prevail that in general in matter environments the stronger or the principal fields take its position that as it unwinds it become the source of the magnetic fields to the lower field's strength and it can be divided into the lower field's strength of the plasmatic and magnetic fields.

At the same time the tightness of the fields of the mafs of the principal matters and their content of the principle fields in their interactions with other principal strength fields of the same strength mafs and matters creates visibility in their level of strength too that these lights are not visible for example in matter mafs or matter environment to human eyes even though in the matter environment the existence of the principal fields and mafs of the matters can be felt or even their affects can be seen.

It has to be said that if the light was not in possession of its own Magravs, the light would have been attracted and absorbed into and by the first magnetic fields or the first object in possession of Magravs in its travel in the universe, whereas the light being in possession of its own Magravs (the dotted line around the mafs of the light in the figure 11) (1) this allows the light to find its position according to its Magravs strength in respect to other objects Magravs and magnetic fields in the universe and this giving the light the ability to travel the length of the universe without being absorbed into or by other objects Magravs and magnetic fields. Thus the light has the ability to establish and maintain its own Magravs positioning criterion in respect to other entities in its environment.

The light rays with higher mass of the mafs of the principal matters, these rays can travel further in space compared with rays of lower mass. Therefore what one can observe as light from far corners of the universe in general these are some of the lights which possess the more mass of the mafs of the principal matters at the point of their inception, as the rays with lower mass have already been absorbed by the other entities in the universe over the life of the universe from the time that these rays were created and have been in travel in cosmos that they could not be observed by man at this point in the life of this universe.

The light observed from far corners of the universe are some of and not all the lights from the whole of the universal light from the point of creation of this universe as most lights by now have interacted and been absorbed or are interacting and getting absorbed by the other fields in the universe, then this can indicate that lights of lower mafs of the principal matters strength and content of principal matters from the origin of the time of the birth of this universe have mostly already been absorbed or been attracted by other pmf and entities in the universe.

The absorption of different rays of light in the universe are environmental dependent and it can be said that the lower mafs strength of the principal matters of the mass of a light ray can become the **seeds of neutrons** or it can be said that the principal mafs of the light as it arrives in a matter magnetic fields strength environment, then this principal mafs of pmf of the light becomes dynamically spherical in structure and then become the seed of the creation of neutron. Hence one can understand how energies which are defined as "the plasmatic magnetic field in motion" convert to matter in a given Magravs environment. Where, each mafs of the matters field's strength components of the light then can become in its arrival at a given environment the seed of one of the components of mafs of the matters of different plasmas too.
Therefore it can be said that the initial seeds of the creation are in motion in the universe at all levels and all time and once they reach given an appropriate environment mafs pmf strength, then in that environment depending on theirs plasmatic magnetic fields strength and the given environment plasmatic magnetic fields strength then light rays manifest themselves or can become tangible and detectable in that environment.

For example the light which has the plasmatic magnetic fields strength of a neutron then can upon its arrival in a given environment changes to a tangible and detectable neutron and then through the losses of some of its magnetic fields to its environment this neutron goes through the nuclear decay (1), which this lead to creation of plasma of an electron and a proton and the light ray becomes the seed of the fundamental atom (1) or the atom of the hydrogen in matter environment plasmatic magnetic fields strength of for example of the Earth.

Thus; "the present assumptions that the first atoms in the universe were created when the universe cooled, **this theory is not and cannot be correct** and in reality atoms are created through change in the plasmatic fields strength of the mafs of the light, secondly at the point of creation of the universe, the universe was not hot and there were never high temperatures that the cooling of the universe has led to creation of the first atoms as presently commonly stated. The first atoms and universally all first atoms' the like the atom of hydrogen in all mafs strength in the universe ***are created through the magnetic field strength reduction*** and through no other means.

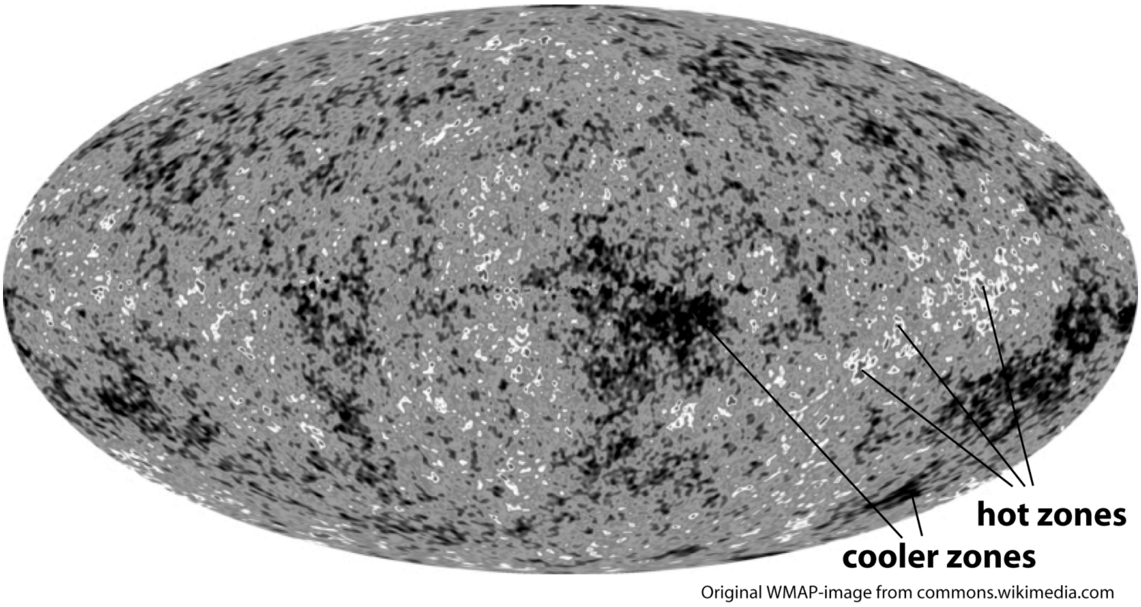

Original WMAP-image from commons.wikimedia.com

Fig. 12 The universe heat distribution map (W-MAP of the universe)

This concept of creation of heat and high temperatures at the birth of the universe is not observed in the plasmas of electron and proton, when the plasma of neutron decays and divides into an atomic structure, and in the universe there are always continuity and parallels, thus the universe has never been hot that has cooled down to create atoms, and this is the reason why W-MAP (8) shows present balance heat distribution of the universe (Fig.12).

In truth the current assumptions of initial heating of the universe at time of its birth cannot be true as the initial atoms of the universe and presently atoms were and are created from the original principal mafs strength which radiated out as principal lights rays of the universe in the centre of the universe and the initial atoms and matters of this universe were not created through cooling of the universe at the point of inception of this universe, which this universe itself is created from the decay and division of bigger mass of another universe as it happens in the division and decay of plasma of neutron to plasmas of proton and electron (3). One does not observe heating in the plasma of electron or proton in this division and so be it the same when the bigger universe divided and led to creation of this universe of the humans (3).

In fact, if! The universe was created by the present scientific assumption then the WMAP should have shown cooler region in its outer boundaries than the present near homogenous heat distribution which has been physical observed and from collated data by the publishers of the universal heat distribution map of the universe.

The heat distribution and the heat concentration on the central line of the WMAP confirms that the universe started from a central line in the middle and as the principal mafs of the universe unwinds then the transition mafs and the matter mafs were released and as were spreading outwards into the space of a larger universe (Fig. 13), then as the Magravs of the principal mafs in the central line region are converted and are reducing, then the overall unified oval-spherical shape explanation of the WMAP come to exist, and then as the total margrave of the central principal matters reach a point where they cannot hold on to all the mafs and matter expansion of the universal-magnetosphere, then at this point the universe will go through a division and all the matters and mafs at this point rearranged and regroup as and in the principal strength mafs and the universe then divides into two principal mafs of lesser masses of universes, which these new universe are magnetically interconnected and then these new universe go through the same principle of the principal light unwinding and the same process as creation of new universe carries as before (3).

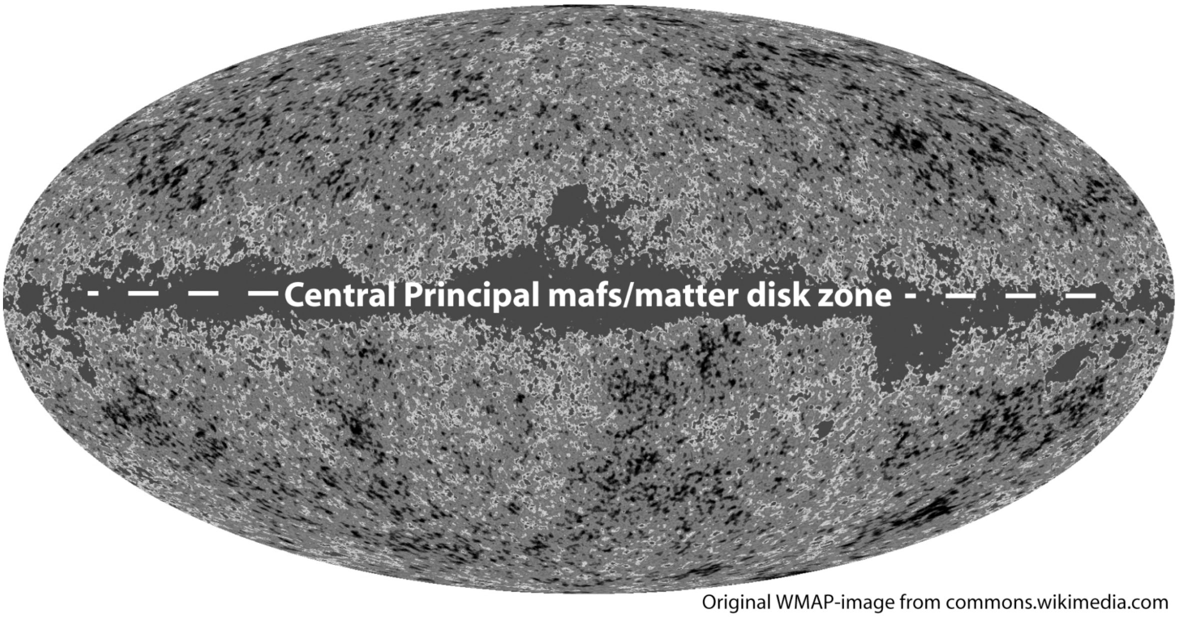

Original WMAP-image from commons.wikimedia.com

Fig. 13 The microwave heat distribution of the universe shows a concentrated heat in the centre disk of the universe

That is to say that the principal matters are the way that universes divided and arrive form another universe and as the light in the principal mafs strength unwinds' in its new environment then the new envelop of the new universe starts to be created (3). The same process occurs when the neutron decays and divides into the plasma of the proton and electron.

Thus, the universe is not created as a big-ball of fire which has been cooling down as has been promoted, but in fact the universe is created the like of as an ejected principal mafs is injected into new a universe environment and as the principal mafs radiate outward as light of stronger mafs, then they create the other matters mafs and matters depending on the given pmtics strength of a given region in the larger process of this unwinding of the principal maf and matters and its lights (3) in this new expanding universe.

Therefore this universe is in fact created inside out than what has been assumed up to now that as it cools down to what it is, which this explains why even the universe is observed to be expanding, as the reason for expansion of the universe is being due to the presence of principal mafs in the centre of the universe and as principal lights of the principal mafs radiate outward, these light rays of the principal mafs while traveling in their new environment they reduce in strength and this creates new matter mafs and matters according to given environments magnetic fields strength and in a given point pmtics hence the creation of new mafs and matters and so forth, which this leads to creation of atoms and molecules and stars and so forth.
Thus there has never been a big-bang and the current appearance of the expansion of the universe is due to the principal maf which as unwinds it creates the sphere of this universe (3).

This method of the creation of the universe now explains' why there is a central gravitational matter concentration in the universe as this pull is centred in the dynamic central line of the injected principal maf of the universe at the time of its inception, which this can be seen as the red line in the W-MAP (Fig.13), which the more concentration of the mass of the principal matter seats in the central disk of the W-MAP and as these principal mafs create their own magnetic interactions, hence these create their own gravitational and Magnetic fields of their own and thus the central gravitational of the universe is in the centre of the denser point of principal mafs of the centre universal heliosphere of this universe is created and solving the present mystery in scientific world that, where the gravitational pulls of the universe come from.

This is how even the plasma of the neutron decays and divides into plasma of the proton and electron too, where the matters mafs of different quantities and masses of different strength from neutron are divided and injected in the new area in their environment, where the original in their overall Magravs of all their mafs as original components concurrently maintained that the plasma of the electron and proton(Fig. 14) and then as the matter mafs ***unwind*** and interact they reach given Magravs level and position ***that new plasmas of the sub-components of the atom are created***.

Where this process of division due to micro -level of masses involved it appears as an instantaneous process of creation of proton and electron.

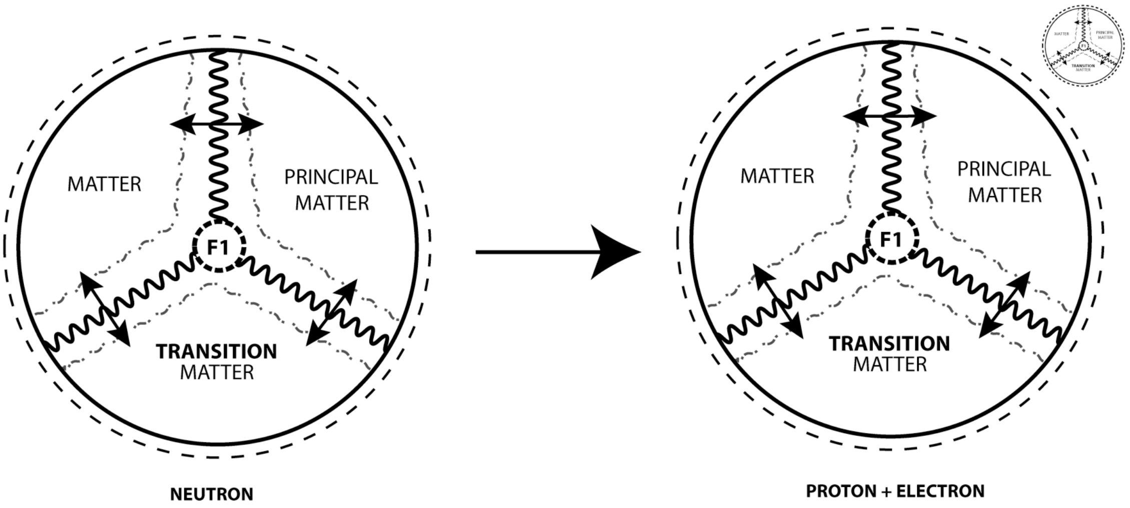

Fig. 14A The conversion of the plasma of neutron to plasmas of the electron and proton
The same process is considered to occur when the universe divides (3).

This being the reason why the proton does not get heated as a big-bang theory suggests but as portion of principal mafs of the neutron which become part of the plasma of the proton **expand centrally** and as the principal maf unwind to the transition matter and matter maf of the plasma of the proton and all other fragment of pmtics and magnetic fields, hence the protons' plasma is created from the division of the principal plasma and the fragments of the fields as different constituent mafs, matters and fields in the plasma of the proton.

The structure of the plasma can be compared to ***the winding spring of a mechanical clock***, were as the main spring is winded up using a key and then as it unwinds.

Where as the winded spring is *tighten in the centre* this tightened spring becomes like the ***principal matter*** of the plasma, a compact entity full of energy in the centre (figure 14b-1) and as the spring unwinds it releases its magnetic fields in lower strength that this unwinding of the fields and reduction in their strength further down the line of unwinding of the spring becomes the magnetic field strength in the ***transition*** (dark matter) ***field strength*** (Fig 14b-2).

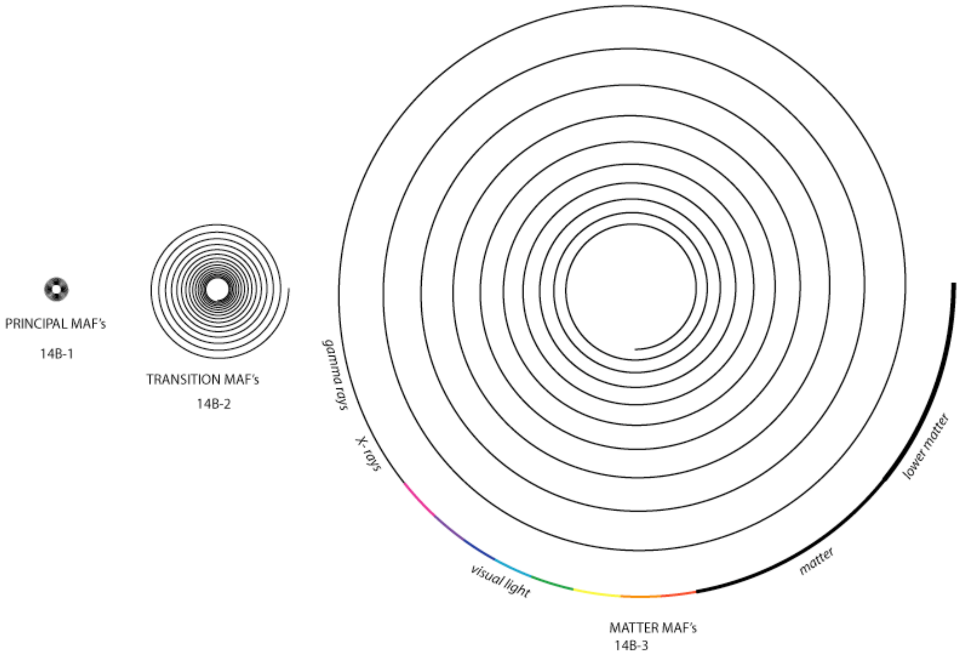
PRINCIPAL MAF's
14B-1
TRANSITION MAF's
14B-2
gamma rays
X- rays
visual light
matter
lower matter
MATTER MAF's
14B-3

Fig. 14B The visualization of the universal spectrum of magnetic fields.

Further as the magnetic field strength and as this field opens up in it's environment and slows down and reduces in strength, these fields in this region become in the visible light and lower magnetic fields strength in the unwinding cycle of the principal magnetic fields reaches the ***matter*** magnetic field strength (Fig 14b-3) and further on this field reduces in strength to even lower fields strength magnetic fields.

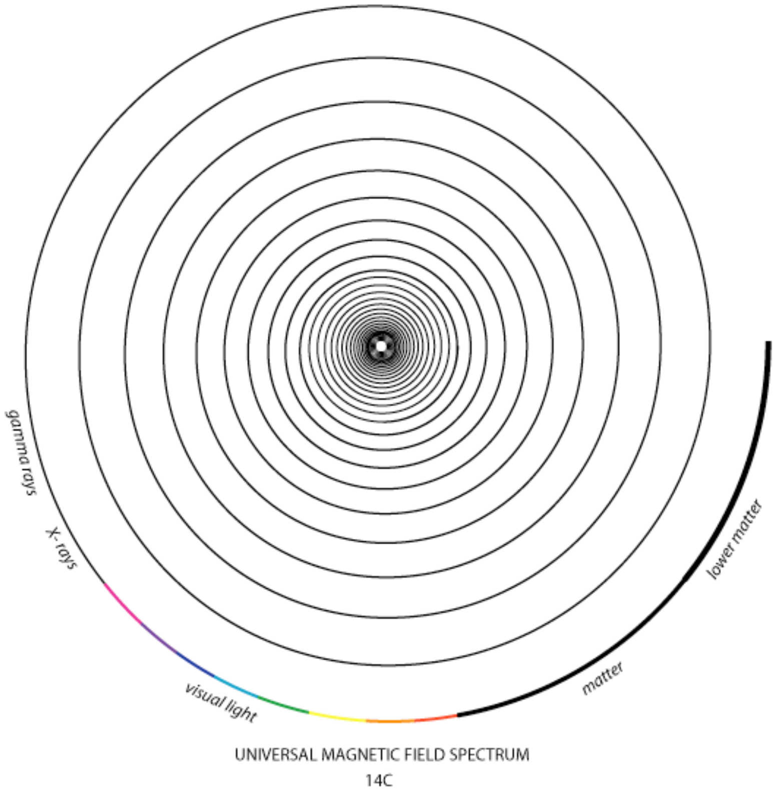
gamma rays
X- rays
visual light
matter
lower matter
UNIVERSAL MAGNETIC FIELD SPECTRUM
14C

Fig. 14C The visualization of the universal magnetic field spectrum.

Then as one puts the whole magnetic fields spectrum of the principal mafs, transition mafs and the matter magnetic fields as one continuous magnetic field strength spring or then the true sense of the unifying magnetic field spectrum which connects all creation of fields and matters as can be observed in the universe, that this can be called the universal magnetic field spectrum (Fig 14C).

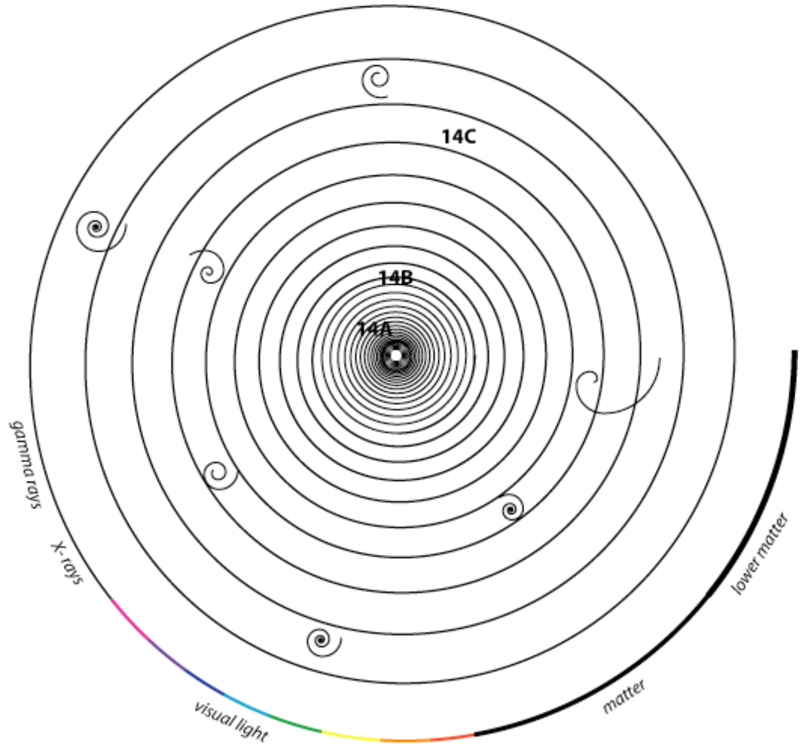

Fig. 14D The universal plasma magnetic fields structure

Where in the plasma of a proton for example, as the fields unwind and divide in the lower scale of the maf strength these lower fields as themselves unwind they interact with other fragments of the fields of the same in the Magravs of the principals matters of the plasma and hence they create the division and sub fields which one observes within the structure of the plasma, these being like what they call quarks and bosons and so forth. Where this total interaction within the capture zone of the Magravs of all the fields is called the universal plasma magnetic fields structure 14D.

The same process can happen with the ray of the light originated from principal mafs strength (what is called the principal light), when this principal light arrives in the principal mafs strength environment and hence the first fundamental principal atoms' can be created in that given environment through the same principle as principal atoms are created in the principal matter environment (1).

In reality: if! one could follow and observe the rays of the light which arrives at the boundary of the heliosphere of a solar system, one should be able to observe that some of these light rays of varying in mafs strength, which can reach this point in the Oort clouds region, these due to the balancing of their pmf and their environment vacuum condition and mafs strength, then these light rays individually or as collection of fields will become particles, plasmas of neutron, atoms and matters, and over the life of the solar system then they rain inward towards planets and the star of the solar system (Fig. 15)(4).

These light's rays which have slowed down in the Oort clouds layers or other inner sectors of the solar system can and do become as neutrons or other particles, then some of these become like of what is commonly known as neutrinos and get attracted back to the star.

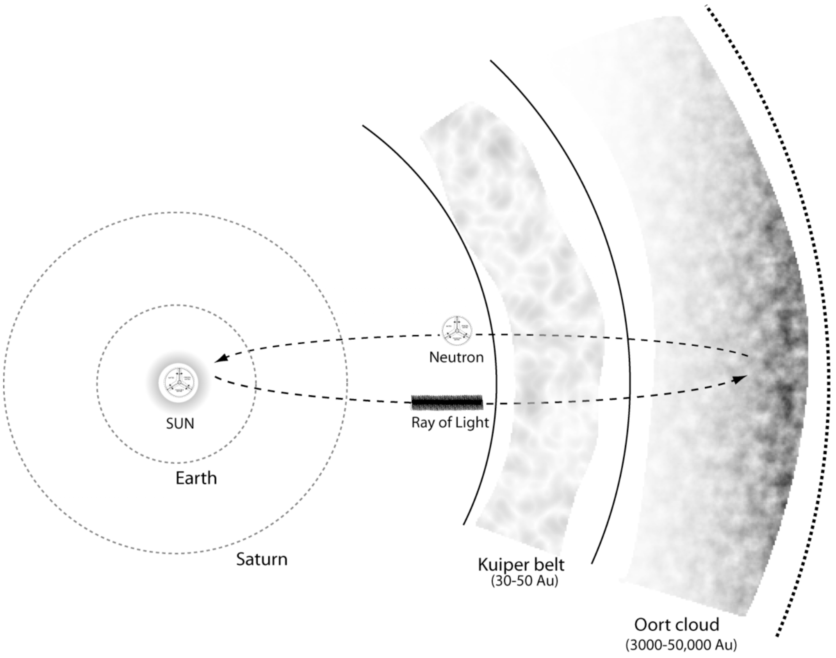

Fig. 15 Conversion of the light ray to neutron, atoms and gans in the Oort cloud

Thus as the stars of solar systems through their light emission continuously feeds their planets and moons with light rays in form of heat and further the sae light rays in slowing down in form of matter, then the same principle continues as the solar systems themselves get feed by the principal stars in the galaxies through their principle lights which these lights at the moment due to the man's lack scientific tools these lights have been invisible for detection to the man of science.

The neutrinos are on the other hand produced by and through the same process from the fragments of principal mafs of the light rays of the Sun or the principal matters in the centre of the earth inner cores too. At the same time and by the same measure neutrinos, which are released by the principal matters and mafs content of the inner-inner core of the earth which are detectable leaving the earth surface their existence and the exit from the surface of the earth is the confirmation of the presence of the principal matters and principal mafs in matter environment of the cores of the earth, as the light emitted by these principal entities lose their pmf strength within the centre cores of the earth (4) and as they move outward from the centre of the planet they appear as neutrinos exiting the earth surface.

Some of these light rays due to their Magravs strength get attracted and in a process known as the fusion of light rays and produce the initial molecules of the matters in the solar system too.

In understanding the concept that lights' rays from the Sun converts into atoms and molecules, then this make obsolete the theory that the ice particles left from the time of the creation of the solar system are the ones observed in the Oort clouds, the Kuiper belt, in the rings of the Saturn and so forth in the solar system.

That is to say most and even all matters in the solar system like the fuzzy material in the Oort clouds layers are mainly the product of the slowing down of the light rays of the star of the system and not the hypothetic present assumption that the meteoroids and all particles in the outer edges of the heliosphere of this solar systems are the residual materials from time of the creation of the star and the solar system.

This recycling of the light rays and production of for example particles is comparable and it is in fact the same process as one observes in systems which have dynamic magnetosphere the like of the earth, where the molecules of the vapours of water rise-up to upper and further regions of the earth magnetosphere and when they reach a given temperature and environmental condition then they gather and when they are heavy and enough of these entities then they rain down on the earth. This is the exact process as it happens with light rays of the Sun as these rays reach to the outer boundaries of the solar system in this process due to magnetic field strength difference instead of temperature difference as in a planetary system, they slow down and become matter mafs and matters and then they rain back towards their star due to its magnetic field interaction and gravitational field strength (4).

In attraction and motion of these lighter atoms and molecules solar winds in the outer region of the solar system are created and as they move in the Kuiper belt of the solar system these matters become part of the constituent matters of the outer giant planets and these light rays become part of and the creators of comets, asteroids and meteoroids as different densities of matters are released and created in this process and collected in these so called belts of the solar system.

Where, the existence of the meteoroids and their pattern of motion confirms the change of the mass of the light to the molecules of gans of the hydrogen as these comets and meteoroids go back in the Oort clouds layers, then meteoroids attract the newly produced molecules of the hydrogen and oxygen and at the same time they carry on mixing the soup of the Oort clouds matters(Fig. 16), this leading to collisions of new produced gas and gans of matters needed for the production of heavier matters and molecules like water in the solar system.

In reality the light rays becoming part of comets and meteoroids, these themselves become like a mixing spoon of the solar system and the cause of the mixing and displacement of the matter mafs and matters needed for having an active and dynamic matter environment in all parts of the solar system.

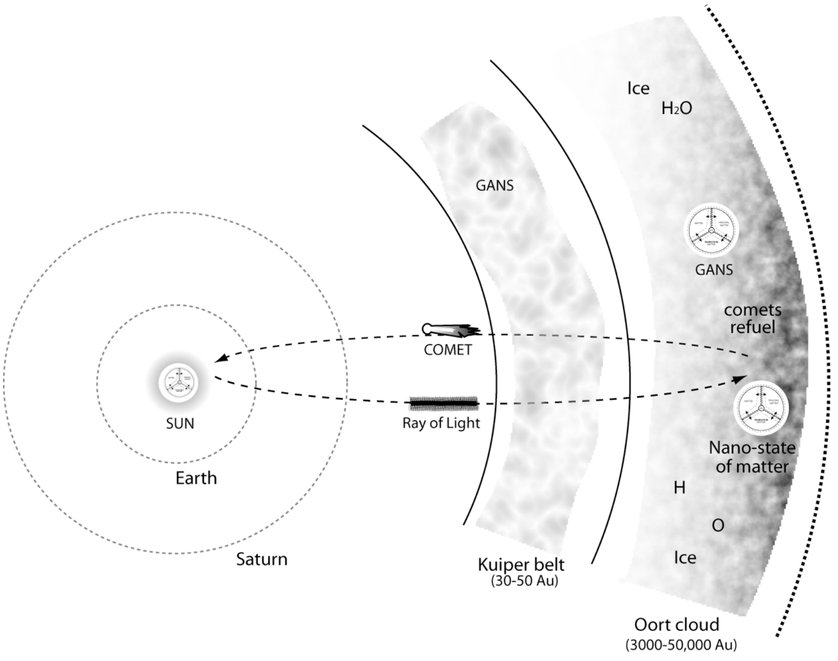

Fig. 16 The light rays become water molecules and get attached to comets and meteoroids and then return back into the inner parts of the solar system.

Therefore the reason, why giant planets exist and possess most of the gasses at the outer boundaries of the solar system, this is due to the fact that these gas planets *continuously absorb all the freshly* *atoms and the molecules of the gasses and gans* which are produced by the ***transformation of the light rays to these atoms***, and hence these gas and gans atoms and molecules get gravitated and pulled around and held or get captured by gravitational, Magnetic fields and get positioned, for example, ***in the rings*** of giant gas planets according to these ice **gans** particles Magravs strength and Magravs positioning.

The collection of ice particles and solids in different rings in the planetary or solar system is primarily due to magnetic fields of static inertia of these objects the like of asteroids and on the other hand due to and Magravs positioning through the Magravs of the gans matters their interaction with the planet or solar system Magravs strength and due to their of these planet's fields nearest to the Kuiper belt and Oort clouds.

Thus, light rays of the star are the feeder of the star, planets, meteoroids and asteroid as they rain back inwards into the star heliosphere over time as the Star's strong gravitational fields pulls' every matter and matter mafs in its system back to itself.

Therefore the light emitted by the star partially become the sources of the materials which are returned to the star matter for it to carry on with its everlasting existence.

The conversion of light rays to the Gans of the matter

Considering the transition of the light into matter in the outer boundaries in the Oort clouds environment of the heliosphere of the Sun, there is a further process which occurs to these newly created tangible mafs from light rays of the Sun as these newly born neutron goes through its nuclear decay (1) and become an elementary atom of the hydrogen in possession of an electron and the proton. This atomic matter in the coldness and in low gravitational and Magnetic fields of the Oort clouds and Kuiper belt become in atomic or nano state of matter of hydrogen rather than atomic hydrogen and hence sometimes due to the weak gravitational field forces in these zone these lights become the gans (5) of the same element.

The **gans of the matter** is the ***state of matter, where the atomic gas of matter becomes the single solid atom of itself*** or what we have called "the gas in nano state of matter" in short or "gans" of the same element (5). Where, the gans of the matter is *internal gravitational and Magnetic field strength dependent* and at the same *time where the gravitational and Magnetic fields of the matter* ***is stronger*** *than its environments 'Magnetic and gravitational fields strength.*

The difference between the atomic gans or a nano gas material is in the fact that gans of an atom is its solid three dimensional free state of matter due to its internal gravitational field interaction.

As the Gans is the solid state of atom of gas in its natural three dimensional structures that its mafs of this matters are interlocks due to their internal Magravs strength of its own and their internal mafs loose bindings, and not rigid binding as in solid atomic state of matter as in the form of ice state of matter of the same entity of the same atom, which ice state of matter is due to and dependent upon external magnetic fields' strength of the environment which are exerted up on the binding structure of atoms that become creates the state of the ice of the same atom.

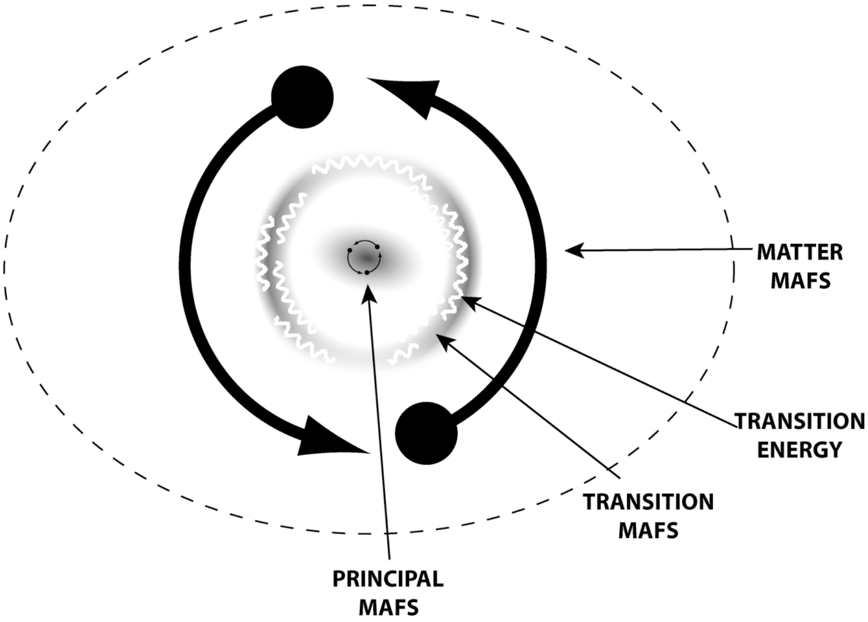

Fig. 17 The structure of gans of neutron

Where, at the same time it has to be noted that the light in becoming in ***state of gans*** it can create the fundamental plasma of the gans (Fig. 17); in the same region as it does create the fundamental plasma (of neutron) too.

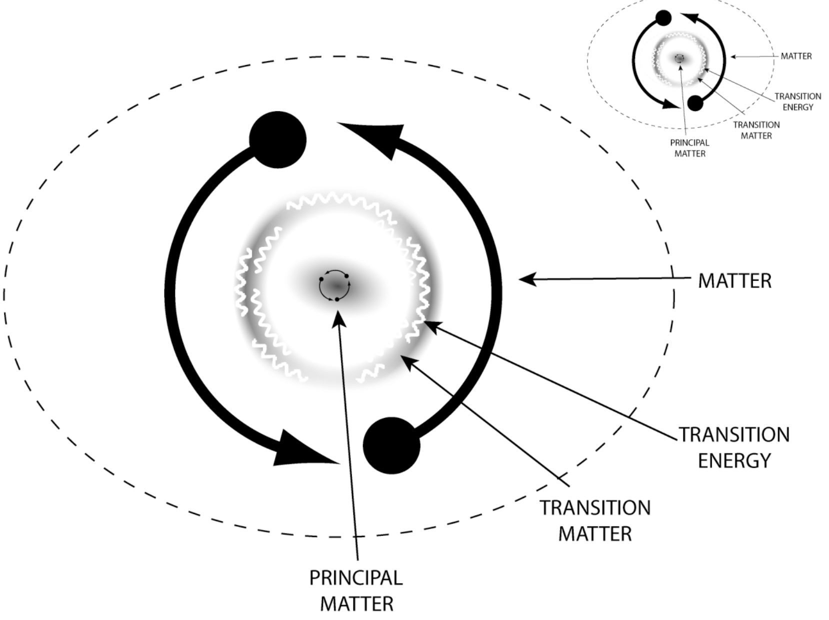
MATTER
TRANSITION ENERGY
TRANSITION MATTER
PRINCIPAL MATTER
MATTER
TRANSITION ENERGY
TRANSITION MATTER
PRINCIPAL MATTER

Fig. 18 The structure of gans of an atom

The gans of elements (Fig. 18) are created through the same process as of known atom directly from the rays of light or through the decay of the neutron of the gans going through its nuclear decay and creation of its own gans of the same elements in atomic state.

It can be said that the gans state of an neutron or atom is internal magnetic and gravitational fields strength dependent, where atom or nano of matter of the same element is environmental magnetic fields strength the like of heat and pressure dependent (Fig.19).

In this illustration the difference between the internal and external magnetic fields forces and processes which lead to creation of gans of matter structure of neutron respectively are shown.

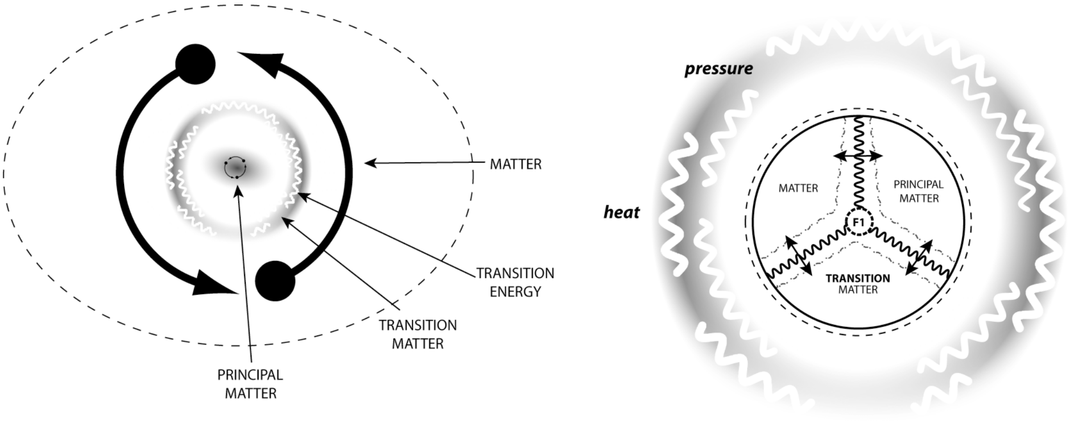
MATTER
TRANSITION ENERGY
TRANSITION MATTER
PRINCIPAL MATTER
pressure
heat
MATTER
PRINCIPAL MATTER
F1
TRANSITION MATTER

Fig. 19 The comparison between the gans of neutron on the left and
the fundamental plasma (neutron) on the right

The gans state of matter is the reasons why for example the molecule of gans of the amino acid (Fig. 20) in the structure of protein and other cell structures the like of in the leaves of plants have tangible, flexible, soft characteristics that they can be manipulated and not being rigid hard.

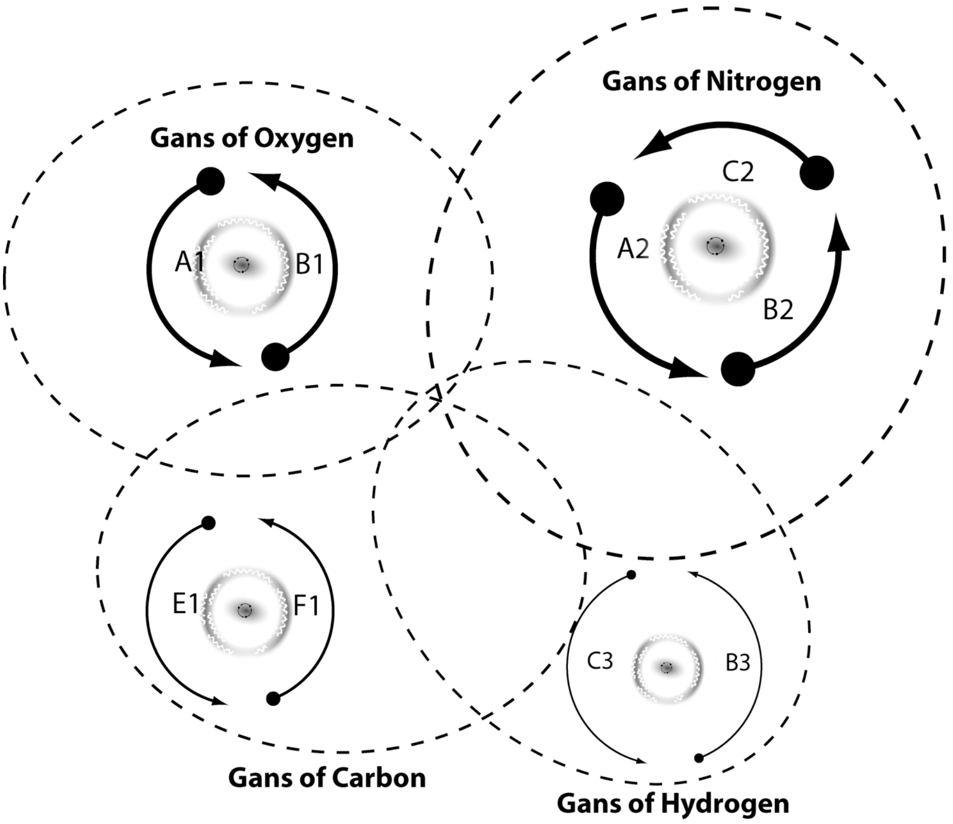

Fig. 20 The structure of a molecule of gans of amino acid

May be it can be said that gans is the solid state of matter of living entities. That is to say the structure of the protein is one of the molecular states of the gans in matter magnetic fields strength environment. Where, through the gans state of matter as part of the cell of the living beings this makes the transition of pmfs needed for rapid transfer of given mafs much easier as by the availability of the gans of an atom rather than nano of atom, this makes it possible to transfer the information of the matters as magnetic fields and not the whole of the atom from one atom or molecule to another as has been assumed by the world of science up to now, and hence once the right measure of the mafs are transferred to a cell then through accretion of mafs in the vacuum of the body then a new gans cell is produced.

Where, for example, if one pinches oneself flesh or skin, what one holds within its fingers are gases of oxygen, nitrogen, carbon and hydrogen in state of gans of the matter, otherwise the body of the man would been a balloon of gases and not in the state of matter as one finds between ones fingers.

At the same time the energy transfer to the already existing cell is due the transfer of pmf of mafs from one cell of gans of lymph to the existing cell. By the same measure as the cell uses its appropriate mafs and is left with wrong or mis-measured amount of mafs for its use, then the cell transfers these pmf of mafs to the blood cells for their removal from the body's systems. This being the reason why for example one does not finds residual of urine in the blood, but as the blood passes through a different dynamic gans molecular structure of the kidney, then the mis-measured mafs in the gans blood cells are reconstituted as liquid and appear as urine in the walls of the bladder.

This applies the same for the transfer of the molecules and atoms of the minerals and vitamins which in fact are mafs of different combinations.

In reality, the world of science and humanity always have known this state of matter as part of the structure their own body but they always have considered it as molecules of the same element and not as another state of matter, thus gans can be considered as the **fifth state of matter**.

This condition of gans of matter is to be explained as the gases when their physic-chemical at ambient temperature and pressure and due to internal gravitational and Magnetic field's strength of the atom the atom physical appearance the atomic structure of the gas ***changes to a compact configuration of the solid*** and as the atom of the same gas become and behave like solid but ***with totally new properties and characteristics*** which has never been known as when the atom was in its other three initial known states (gas, liquid and solid).

This ***conversion of the light rays to solid state*** occurs in the **outer regions of the solar system**(s) where the gravitational field forces (vacuum) are low and the temperature has no effect on the internal motion of magnetic fields that create this solid state of matter. Then, as the ray becomes and appears as gas and in a given conditions manifest itself as the nano structure of gas or what is called "GANS" in the same environment *due to atom's own internal Magnetic and gravitational fields configuration* and ***not due to*** *its environmental magnetic fields.*

This state of gans and gas of an atom is being the *same difference as between* ***the mass*** *and* ***the weight*** *of an atom* respectively, where **mass** is due to internal atomic interaction of fields of the atom and irrespective of external field forces and where the **weight** is due to environmental magnetic fields effect on the outside appearance of the same mass.

Humans are familiar with the gans of the molecules and atoms as integral constituent of their own tangible body, where ***gases at low internal vacuum condition*** *of the body convert* ***gases into gans or soft individual solid atom state***, but up to now the scientific world has never considered this state of matter as they just say for example the carbon of protein chain or matter of a cell or a carbon as part of the structure of a sugar of a leaf of a plant and so one. That is to say for example, when an atom or molecule of a hydrogen gas, which was created from the ray of light as this gas is absorbed by the human body then this gas of the atom changes to the gans of the same atom and then through the accretion process in vacuum condition of the human internal operation, then this gans particle according to its Magravs strength becomes part of the molecular ganses of the **N-C-O-H protein chain** of the human and that is why human body becomes tangible entity. Where, the creation of the solid gans in the atom of the gas of the atom, this process by itself is due to the accretion process of all the fields of the mafs of the gas atom itself.

If this change of state of light to gans does not take place, then the human body should have been a balloon of gasses and not in the solid tangible arms and legs and so forth.

Further, it has to be said that as the light ray changes to the atomic gas of the hydrogen and as it reaches the planetary condition and natural Magravs state of planets this matter changes to gans atom of the same matter in the atmosphere of the planetary entities too.

The confirmation of the production and existence of the gans in the Oort clouds and Kuiper belt is and can confirmed by the existence of the hexagonal rings on the top of atmosphere of the poles of the Planet Saturn. Where the composite dynamic layers of nano gans of material(s) in this part of the planet leads to creation of the hexagonal shape solid clouds as has been recorded by pictures taken the scientific world from the pole region of the planet (4).

The only reason that one sees these types of structures in the universe is primarily due to existence of gans state of matters like hydrogen, methane ganses and others in these layers, as the Magravs of the planet interact with dynamic metallic characteristic structure of the hydrogen gans and other matters ganses and this nano state of matter in dynamic magnetic fields of the poles of the planet leads to creation of current in between these layers and consequently for example in the poles of the Saturn the formation of hexagonal shape rings of the clouds which is the consequence of these interaction , which are natural positioning for the gans in dynamic mixed and composite magnetic and nano structures.

To create the hexagonal shapes of the clouds in the north-pole of the Saturn one needs at least three layers or mixture of three different ganses of different or the same elements to be able to create these shape features in the upper atmosphere clouds of this planet (4).

In reality gans's of matters behave as semiconductors do in the matter environment, with the difference that the gans of matters are *always as they are in a* ***semiconductor state*** *and at the same* ***time they create current*** *in any environment* due to their internal dynamic magnetic fields structure.

On the other hand, the conversation of the light into gans and its absorption and their interactions with the strong Magravs producing planets like Saturn is the cause and producer of the internal Auroras that are commonly observed at the poles of the gas planets. Where these auroras are due to the fact that as the gans of the elements from Kuiper belt gets absorbed into the giants planet atmosphere and as these ganses are in solid atomic matters state the like of the cosmic dusts and in their interaction with the magnetic fields of the planet, these interactions of the gans and the magnetic fields of the planet leads to creation of the Aurora at the poles of these giant gas planets.

It is essential to understand the fundamental effect of the light rays mafs of the matters' in the solar system and even in the creation of the solar systems and planets themselves that one can understand why the rays of light are the transporters of creation in the universe as in their conversion they become the created entities the like of atoms, cells and so forth due to their pmf strength of their environment. Where, in fact the lights in becoming the gas of the matter, then they become the source and the cornerstone of the creation in all matters levels of the universe. Thus as light changes to atom and gans then in the right environment of for example one G (gravitational field strength of the earth) as on earth then the collection of atoms as the molecules of ganses state of matters leads to creation of the beings, the like of single cells and complex combination as of the human's body.

Light ray from the Principal Star (black holes) as it loses its strength and becomes to matter mafs state strength and then it becomes to the nano and atomic level of the matter strength of the fundamental element of atom of hydrogen, it is at this point, where this matter generates the first external gravitational and magnetic fields or what is called the Magravs of itself in matter environment that this tangible entity due to its Magravs positioning and strength in respect to other Magravs that it can be attracted by or attract to itself the same and others elements in the same environment. Then it is at this point that the light becomes physical entity and the seed of the stars and planets and the essence of the tangible entity in the all matters environments of the universe.

In the spans of the solar systems, galaxies and cosmos as the gans and atomic of the hydrogen interact, they create the first seeds of creation of more powerful Magravs that leads to attraction and conversion of other light rays in the universe into liquid and solids, which these interactions leads to creation of galaxies, stars and planets, where the light itself in different strength becomes the tools of the division of the hydrogen for it to start the cycle of creation of intelligence and creation of the lives in all corners of the universe (6).

The gas and gans of hydrogen becomes the dynamic fuel source of creation of the magnetic fields necessary for the creation and establishment of the stars and solar systems and their planets, as the interaction of different magnetic fields of gases and ganses lead to the creation of the initial

Magnetic and the gravitational fields needed for these celestial bodies to be created in the universe. The existence of nano materials in the core of the planets like earth has been initially disclosed (1), now it can fully be understood that how these nano materials in state of nano matter, gas and gans become to exist in the centre of these celestial entities and be the initial seeds of their creation.

It is important to realise that due to internal structure of the nano material in layers and in gans state, these materials by large keep their structural integrity, where high external environmental matter conditions like pressure and temperatures cannot affect the total structural integrity ***once they are in individual matter gans compound state***.

This meaning that internal structure and molecular binding of gans and nano materials in nano layers or composite structures will not be affected by the matter environment's the likes of temperature and pressure condition of the environment. But the gans will be changing in its structure when conditioned to high level dosage of radiation.

For example, this is one of the reasons, why the DNA of the human being is venerable to changes and absorption of radioactive energy and in the radioactive environment or to burst of radiation, as the real structure of the amino acid of the protein of the DNA chain is made of gans of matters and not solid matter of the same gases (6).

In these conditions the gans of the matter receives radiation in the upper levels of the matter mafs strength and hence it has to accommodate this extra pmf of radiation hence the gans reaches a new magnetosphere pmf strength and consequently the gans of the atoms of the protein which has absorbed the additional pmf has to accept a new positioning in respect to other ganses in the chain, hence new positioning means new additions or new configuration in the setup of the DNA and hence new changes and hence slightly new chain configuration or alteration (6).

The property of the light in gans state make the gans venerable and receptive to changes due to the magnetic fields spectrum in all levels of mafs of the matters strength, this be it from the higher order mafs of the principal matters fields to lower pmf strength of below matter environment. This being the reason why DNA due to absorption of for example high level radiation or strong ray can change its configuration this leading to new permutation and change in the characteristic and change in parts of the being characteristics of the living matter.

Through this process of Magravs positioning and that all Magravs needing to be in their given position in the DNA chain that one observes the consistency with structure of the DNA for every living entity in the universe, which these chains of protein positioning become entities exclusive to that of a given being or entity. Now one can understand how light and its process of conversion, fusion and amalgamation lead to creation of exclusive structure of DNA and creation of beings (6).

Scientists due to old understanding of creation of magnetic field consider that the earth has an iron metallic core which due to its motion through convection theory leads to creation of Magnetic fields of the earth. The fact that the Sun does not possess a metallic iron core but it has large and powerful Magnetic and gravitational field forces; hence the existence of iron in the centre of all dynamic rotational bodies is not a common and acceptable principle and natural order of creation of Magnetic and gravitational fields in the universe.

In truth the plasma of neutron does not possess iron core, but it goes through the decay and division to make plasma of electron and proton and in this process of decay neutron does not go through iron energy balance condition as has been claimed by the cosmologist that this is how they consider stars reach the point of no return due to the creation and solidification of iron matter in their central core. And at the same time geologist consider iron core of the earth is reason for the planet possessing

Magnetic fields.

The method by which the Magnetic field's (repulsion of fields) and gravitation field's (attraction of fields) are generated exponentially increased in all planetary, solar and galaxial object are initially and primarily through the magnetic properties of the plasmas, nano and gans materials within the centre and internal structure of all celestial objects. As in the inner cores of celestial objects the ganses of the matters due to their mafs and through nano layers or different composition of collective crystal and magnetic fields structure they become or behave as the best insulator and conductor simultaneously in their environment, where the gans and nano matters, and their mafs and their properties ***changes from conductor to insulators*** as these matters behave as semiconductors and metals and at the same time due to their dynamism of their pmfs this leads to the initial and then partially continues production of the Magnetic fields and gravitational fields of the star or planets respectively (2, 3, 4 and 5). It is at this point that magnetic properties of the nano and gans materials in their interaction lead to creation of Magnetic and gravitational field's forces, this leading to creation of the independent mass and magnetosphere of the new entity.

Through the gans internal disassociation of its mafs structure as the internal principal mafs of the structure unwinds rapidly and when this transition to matter level strength is very rapid, then this lead to release of the high temperatures and strong rays of light in man's protein and the visible magnetic fields spectrum strength and hence detection of the release of spectrum of magnetic fields of the light rays.

The gravitational properties of nano materials and the creation and attraction and absorption of matters from the environment have been explained in the paper published in detail (5), at the same time matter cannot attract or gravitated towards another matter if both are not in possession of magnetic fields, where at the same time it has been confirmed (5) that nano materials possess magnetic properties and as magnetic fields interact then nano and gans materials create their own gravitational field zones and environments; and in having nano and gans materials in the centre of planets and stars this indicates how the gravitation fields (6, 7) of all dynamic systems and objects in the universe are created.

Ganses being in the right magnetic fields strength environment due to their Magravs strength these materials ***will convert into gravitational zones in their environment and attract light and in turn absorb, create and release energy***, where this energy leads to creation and attraction of gases and their conversion as for example the same process in the human body to solid state of matter at room temperatures, where in test it has been shown that this process of absorption or conversion of the gas to gans leads to creation and release of energy and similarly vice-versa. As for example in absorbing CO2 from its environment the plant releases energy into its internal structure to be used for conversion of minerals to food for the plant to assist with its survival and growth.

Thus the truth about how even the heat of the body is created is far from the present assumptions which the scientific world has led the man to believe, as they themselves never understood the real working of universe and the process by which matters behave and operate in its low vacuum conditions.

This concept and process has been tested, explained and published in detail in Chapter 4 of this book, where for example in capturing gases of CO2 from the environment and collecting it as gans as a tangible entity at room temperatures and pressures, the energy released through this process is use to light and run lights and motors respectively in the lab by independent scientists as has been published in the journals in May 2010.

Whereas, in this case energy is released is in a similar way as it happens in the human body, when the light or magnetic fields of the light enters in the micro vacuum condition of system or the body and in receiving the light rays and in trying to entangle or go through the process of atomic-entanglement in gans level and then accretion, this process leads to release of the surpluses energy and then accumulation (accretion) of solid ganses which appear as cloud of matters in tests and in human body which then this become tangible structure of matters of the body and the surplus energy is released as the heat of the body or it is used for the communication system of the body.

This process of accretion is one of the sources of energy release in the living body entities and in experimental testing this released dynamic energy in the level of plasmatic magnetic field strength of an electron produced through this process become the source of power and is enough to light a LED for several months, as this process is a natural process as long as the light or other rays can penetrate the liquid, jell or soft tissue condition of the system this process can continue.

The other important reason that one finds water in this solar system is primarily due to the fact that pmf and Magravs strength of this solar system and the ones like this encourage and enforce initially the production of gans of magnetic matters the likes of oxygen, which leads to creation of liquid state of matter of water.

At the same time it is curtail to understand the structure of three and two dimensional properties and characteristics of the same atoms in layering and physical positioning. The gans of the matter is captured and held in three dimensional states, where gans of matters gravitate and make dynamic clusters of ganses around themselves and become dynamic entities, whereas, in the linked version of the same atom clustered the same matters in solid state makes layer and gives and possess totally different properties and appearance.

The light in its conversion to atom of gas or atom of gans according to its given environment then will take and enforces its preference estate of its manifestation, and with slight change in level of its internal vacuum (defined as availability of Magravs and their strength) then the light will appear as the gas of an atom or a gans of the same.
It is interesting that scientists and scientific organizations are not in possession of tools at this moment to identify these two separate states of matters of the light in visible entity of matter and hence always consider all matters as the atoms and in presently considered four states' of matters as has been tradition.

The same ray of light when changes to gans of the hydrogen atom, this change's its dynamic diamond crystal structure of the same matter and hence this gives the same matter insulating properties and hence allows the creation of conductive and isolative properties of the same element in its layering and clustering structure and its dynamic properties in between the same elements of atoms of the gas and gans.

The gans in specific condition has diamond crystal structure of the hydrogen atom and with slight change in its environmental condition become a conductor for short or on a long term. Where this has been experimental prove and demonstrated and then clustering of this structure mixture of gas and gans of the same atom leads to creation of multi-layers of conductive and isolative matters of the same matter and hence in the dynamic magnetic environment of the universe this leads to creation of different and stronger Magravs, and start of attraction and positioning of different matters towards these initial seeds of gravity and Magnetic fields and hence start of creation of initial celestial entities like plasmas of neutron, atoms, stars and galaxies.

This is the reason, for example why the same matter of carbon becomes to be the best conductor as nano layer Sp2 and the best insulator as the diamond crystal insulator structure known as Sp3 across the same layers or through its layers respectively where the production of such a mixture has been independently confirmed by Ramon spectroscopy of these structures produces as layers on a given surface. Therefore all elements and molecules or their combinations can have both Sp3 or Sp2 structure and property at the same time and in the same environmental condition.

This being the like of the protein chain, where each gans of each element exist individually and collectively in the amino acid as molecule of gans as each element plays its individual role which can affect collectively the operation and existence of the acid in its given environment (6).

In understanding the concept of the conversion of light rays to gans and gas and creation of nano materials, this discredits the present philosophy of the planets needing solid iron cores which enables them to create their Magnetic fields that they can operate in the universe or even the human race could be brought to the earth from other places from far corners of the universe. The life on earth is the natural product of lights' motion and its conversion in the internal activity of this planet and solar system.

By the same measure light plays in different environments different roles which can lead to creation of the same and similar or totally difference combination of creation of molecules of gans, which these manifestations of difference appearance of the light are environmental pmf mafs strength dependent. This is to say that life is not and shall not be the exclusivity of this planet and this solar system once on understand the principle of creation and motion of the light rays in the universe.

Thus the centre of the earth does not need to have solid iron core for it to possess Magnetic and gravitational field's forces leading to creation of life on the planet too. Whereas, composite elements of nano materials, gases and ganses crystal structure of the same element in the inner cores of the planet can lead to creation of the same Magnetic fields and gravitational fields of the planet.

This explaining why even some planets like Saturn possess Magnetic fields, even though these planets could possess some solid matter as their constituent materials as part of their inner structure.

In reality as layers and layers of dynamic gas and gans are piled-up in the centre of the planet the effects of dynamic magnetic layers exponential strength that they make as the Magravs of the planet take hold and this phenomenon lead to creation of stronger fields that can span billion of kilometres in the solar, galaxial heliosphere in the environment of these entities.

Thus the transition dynamic helixial cylindrical ray of light in slowing down in a given environment Magravs strength, this lead to creation of specific matter in that given Magravs strength environment.

The scientific world has to understand that the state of gans of matter is a natural phenomenon in the world of creation, this meaning as man has learned about the solid, liquid and gas state of matter, as these states of matters are due to temperature and pressure difference in the environment of these matters, but there is another state of the same matter even though as it's like a gas and behaves and exists as matter in form of liquid and solid but due to its internal gravitation fields forces of its atomic structure this matter for its appearance and properties is ***not external environmental dependent***.

The gans or the gas in nano state of solid state of matter will play an important part in the future of the science and technology. That is where the same atom behaves like a gas, liquid and solid at the same time depending on its own internal magnetic fields interaction in its plasma and not of the external environment conditions around the plasma the same matter will be in dynamic solid state at room temperature at all-time irrespective of the environmental condition.

That is to say the plasma of an atom of gans receives or coordinates its own Margravs and its own Magravs positioning in respect to its own electron and proton and other gans structure.

Thus, gans is the state of the structure of the light which has changed from dynamic cylindrical plasma of magnetic fields to a dynamic spherical entity of the same matter mafs, which its state of manifestation in matter environment is internal gravitational dependent and not environmental condition dependent. That is why one can observe the same element in coldest and hottest regions of the universe and still misreads them as atomic matter rather than nano-atomic or gans of the same matter.

This is how some of the plasmas and rays ejected by the Sun, depending on how they arrive at the Oort clouds and Kuiper belts then become nano materials of gas or gans and in clustering with other matters and ganses become part of the meteoroids or asteroids and look as they have been there from the time of the creation of the solar system.

In the universe and solar systems of the right Magravs strength, the light changes into the nano hydrogen and through internal magnetic fields of the solar system through the principle of magnetic balancing and accretions this matter elevates to the next magnetic material, which one of these materials' is the gans of the oxygen. This element due to its Magravs positioning and balancing in the Magravs of some solar systems lead to creation and production of water molecules in the Oort clouds belt and then these water nano particles gets pulled back into the inner parts of the solar system as nano ice particles of gans of this material which for example one observes them in the rings of Saturn.

The ice particles in the ring of the Saturn are there and for ever are produced by the return of the rays of light which were released by the Sun itself, on their way back to the Sun as gans and ice particles and they are temporarily trapped and positioned in these rings, and the present assumption that these ice particles exist in these rings are not by the crash and splitting of meteoroids or asteroids into smaller pieces that afterwards have stock in these rings of the Saturn as can be observed at the present time.

Thus as the light reaches the outer boundaries of the solar system, its becomes nano material, which due to their magnetic properties, this leads to creation of a nano magnetic material which is oxygen and then this matter of nano particles gans of oxygen to find stability for its Magravs, this attracts two atoms of gas or gans of the hydrogen nano particles, which this leads to production of gans of ice looking particles of water in the Oort clouds belt of the solar system, which they move inward and become the liquid magnetic substance and the source of the life in this solar system.

Thus the water in the solar system is generated by the rays of the light of the Sun itself mainly and by no other source and hence water is produced in the solar system by the recycling of the light rays of the Sun itself.

The mixture of these ice nano materials with solids matters of the solar system leads to creation of meteoroids, whereas these dynamic objects in traveling between the Oort clouds and Kuiper belt, these objects attract nano water molecules and then they return to the inner regions of the solar system. This being one of the reasons why the meteoroids are seen on regular basses and their size never changes, but their vapour trail tells the existence of the water residuals in their structure's as forever these objects gets replenished with the Sun rays in the form of gans of water in these outer regions of the solar system and return back into the inner parts of the solar system.

Therefore the theory that water is brought to planets like earth by meteoroids as has been assumed is partially correct, but Magravs of planets like earth in the Magravs of its Sun have the ability to convert light to nano matters and nano matters which have been produced into hydrogen and molecules of water through accretion process in the vacuum of the solar system and in the upper layers of the earth atmosphere itself. Hence water is the creation of the combination of atoms created by the rays of light and in different magnetic fields conditions in the Margraves environment of this solar and planet system.

This process of conversion of the light can happen anywhere in the universe as the nano particles of the hydrogen is one of the first and weakest Magravs strength that light can change to matter anywhere in the Magravs strength of this universe.

Thus water is not exclusivity of this solar system and is product of the light rays of the universe in matter environment.

There are waters like substances in other matter mafs and matter strength environments in the universe too.

It is important to realize that all gans have the magnetic properties with the difference that oxygen is the one of few common magnetic materials in the solar system and in most of the universe needed for braking of the light itself to lower level strength, which can lead to creation of intelligent life. In some solar systems and galaxies pmf strength of the gans of the matter of the oxygen is the common magnetic elements and the reason why all maters existence dependent upon availability of this matter in the pmf of the environment.

In the living bodies in possession of systems the like of the lungs and gills in fish, the light in changing to molecule of oxygen and in being inhaled for their existence, in these systems contrary to present belief the atoms and molecules of the oxygen will never cross the boundary of the structure of the lung or being absorbed by the blood as has been thought and promoted by scientific world (6). The fact is that in the process of inhaling or in case of fish by passing the water through its gills, the atoms of oxygen simply passes limited value of their matching plasmatic magnetic fields strength to its matching oxygen atom of protein of the blood cell that the blood cell as commonly called gets oxygenated in the lung (6). In reality no oxygen atom or molecules ever passes the walls of the lung as has been assumed, but the oxygen atom transfers its pmf to the bloods oxygen atom in its protein and then this pmf is carried through and by the same principle as the pmf from the oxygen of the blood protein cell is passed on to other oxygen atoms ganses in the protein chain of other cells in the human body (6) as they come across their corresponding mafs strength protein in the blood.

The oxygenation of the red blood cell is primarily another quantum-entanglement of the same matching matters strength or entanglement transfer of adequate amount of pmfs of specific cell of the matters of blood.

The fallacy that oxygen atoms cross the walls of the lung is so far from the reality of existence that as saying a man can walk through a solid wall. This hypocrisy has been the consequence and through the lack of understanding of the real structure of the oxygen and its origin being laying in the structure of the light that matching plasma of light and atoms can/and frequently entangle and transfer their pmf up to level of their mutual strength that existence can be guaranteed in one form or sub-structure of the same and without the plasma of actual solid atom crossing the boundaries (6) of the lung.

In fact the energisation of the red cell is through transfer of the pmf of the atom of oxygen and not the actual transfer of the crossing of oxygen into the blood stream and the brightening of the red cell passing through the lung is like the variable light switch, that as one increase the current through the

light the light shines brighter, this is the same process and in the same way as the red blood cells receive the pmf of the oxygen, then the blood become brighter or richer in pmf, then the blood cell after transferring its pmf to appointed cell in the body then pmf strength of the blood cell is reduced in energy and illumines less, dimmer or looks darker as it returns to the lung to be recharged (6). Man has a lot to learn first about his own body before he can conquer the universes.
In understanding the function and the fundamental properties of the gans and molecular gans then computing systems can be produced that not only they think the same ways as humans do but due to their dynamic energy and singular directional rotation, then at this point the first quantum computing systems can be produced that these computers will be exact copy of the human brain and muscles in structure with the same multi-neuron-cross linkage and communication channel, as these channels are created as gans molecular Magravs linkage will allow them to have the same simultaneous multi-cross connection that these quantum gans of molecules structures will be exact parallel systems as do the human brain in all aspects of a real working structure of logical system, which even they can carry emotions as well as control and logic. Where, in fact the binary gans and twin gans sets will be the future of fast thinking units needed for the high speed control logic system for space systems travels of the future.

As there are gans in the matter environment mafs strength then there are gans in pmf strength of principal matter and transition matter too, and at the same time the gans of matter in all mafs strength will operate in all levels of pmfs strength. Thus it is possible to produce molecular gans unites which work at high speeds and with same emotion and logical systems as human brain, but with speeds that are far exciding speeds of the human brain processing capabilities.

Therefore it can be said that for the first time the man has the capability and the knowledge of not only producing solid gans of gas at room temperature, but it can go one step further than present conceptual nano matters and nano matter based computers and start working on real brain based matter structure logical systems by using the gans of the gases which are the building block of the human amino acids.

But the question is that would we as far as building the first human operational brain and body using these matters as we have already decoded the DNA of the human genome, and as we know the full chain of the DNA sequence and we have already unravelled the mechanism that the RNA produces it energy and current for its internal communication and control of the internal systems of the DNA, then would man become the creator of new beings bases on human structure and would the discovery of the gans and simple method of its production will allow the ethical question of mass replication without the need for stem cell or even egg and sperm, which are memory banks of human survival on this planet. Would man will make one of his own with a clean DNA memory. OR is the human being to become the god of new generation of beings.

This is not the first time in the universe that one created has the ability to reproduce another of the same, but at the same time this discovery of structure and properties of the gans would have become apparent to man of science in due time as he would have progressed in his scientific conquests.

Some scientists believe that solid iron is the main magnetic matter in the universe and this matter in the centre of the stars and its energy balance conversion between its fission and fusion is the point when stars goes through their slowing down process or reduction or what is known as their nova state. The truth is that the plasma of mafs of the simplest structure of the plasma is the common and the fundamental and elementary magnetic entity in the universe. Thus the plasma of the light is the basic and common magnetic matter in the universe and one does not need iron to reach this point, as the plasma of matters in given position can lead to unbalancing of their total environment pmf mafs and Magravs, this leading to total disassociation of the all matters in any matter or object in the span of the universe, and once the process of the chain reaction of disintegration and disassociation starts then the whole entity disassociation from each other begins and in the case of plasma of neutron and star at this point disintegrates to re-association in a different as plasma of proton and electron and new star and planets in their form in difference point and place in their given environment.

By the same measure the plasma of neutron which is created by the conversion of the light ray is the common magnetic matter of the solar system and not atoms of iron, and in reality the light of the Sun and its Magravs dictates this property of this mafs of the matter and the present assumption that energy balance conversion of iron is the cause of their nova cannot be correct and as most of matters in the universe are created because and through the existence and the operation of principal matters in the centre of the stars and their rapid unwinding and balancing of their Magravs in the centre of the star leads to the expansion and disintegration of the star and not as has been assumed at this point by some cosmologists due to the iron fission-fusion balance (3,4 and 5) phenomenon.

What has been explained above, this is the process of creation and recycling of the light through its full circle of its existence from matter to light and from light back to matter(s).

The detailed information and explanation about the existence of the gans of the matter has been possible due to through series of tests which has been carried out and absorption of the physical solid CO2 at room temperature has been confirmed. The procedures and tests results in confirming the existence of this new state of matter was originally published in 23.12.2009 by the Keshe Foundation on its website. For those whom have not been able to read this paper a copy of the same is published in this chapter.

The state of gans of matter can now explain about the truth of life in the universe and this opens new horizons in the world of energy, new materials, production of more natural components of the human body parts, production of new medicine, new types of food processing technology, which this eliminate the need for present farming processes to feed the populations, as protein can be produced from the content of the light or the mixture of composition of lights mafs in gans state of matter that can lead to production of solid matter the like of fresh meat and or composition of dairy milk and so forth. Thus through production of gans dependence on the extensive farming in the space becomes irrelevant, where in deep space by absorbing weak residual mafs of the light and in bringing them together and through accretion, then producing the gans of the fundamental plasma, then producing the gans of fundamental atom and then through accretion creating composition of ganses of protein, then one can produce any food in the span of the universe directly using the matters of the light. Originally this concept of producing protein through gans from the gases of environment was considered and successfully tested in Tehran in the summer of 2008, and through simple process the first gans of protein in small quantities by using the natural air at room temperature was produced.

In the appendix section of this book, attached is the paper titled “The unifying field theory” published in the 28.10.2009, in this publication it is discussed and can be concluded that the origin of all entities, tangible or not are magnetic fields and the interaction of at least any two of any magnetic fields of any strength can lead to creation of all other effects and matters which are observed in the universe.

Thus the magnetic field is the unifying field and the mother and the seed of all created and creation in the universe, thus the magnetic field is the Holy Grail, which the scientific world had been searching for in the universe strength to connecting all fields and effects observed in the universe.

Chapter 4

CO2 in Gans State (paper 2009)

Direct absorption of carbon dioxide CO2 and methane CH4 gases from the environment and their conversion into nano solution and nano solid matters and production of energy and oxygen by the use of nano compound at ambient temperature and pressure.

Abstract

In this paper, results of processes will be introduced for the absorption of carbon dioxide CO2 and methane CH4 gases direct from environment and their conversion into nano materials of the same suspended in solution or dried as powder. Where, concurrently within the processes of extraction of these gases in a simple reactor, we have managed to produce energy and oxygen.

Discussion

New methods for absorption of CO2, CH4, and others gases and further conversion and preservation of these elements in their nano solid state of matters at room temperature and pressure has been developed and achieved.

We have developed simple processes by which carbon dioxide CO2 and methane CH4 gases can be extracted from their environment where no additional heat or pressure to achieve this has been applied.

At the same time conditions can be created within these systems that CO2 in conversion without going through any chemical process, can lead to creation of methyl, oxygen and methane gas and vice versa.

Where during these processes of CO2 and methane gas extraction or conversion we have managed to produce usable energy and oxygen molecules in sustainable way.

Production and extraction of CO2 by applying energy is a known technology in the present science, what is new with our technology is, that not only CO2 is extracted in a simple system at room temperature and pressure, at the same time, energy released through absorption and conversion of the captured CO2 to formic acid, we have developed simple techniques to capture the energy release from this transition process to produce useful power.

The novelty with this new technology is that in previous state of art technologies, engineers and scientists had to add energy to extract CO2 the environment or to produces O2, where through our technology, not only we generate useful energy while absorbing these gases from the environment, but also at the same time, the system can produce molecules of oxygen.

Thus, this simple system has the ability to replicate the processes as in a natural way as plants achieve in their conversion of CO2 into oxygen.

With this technology, CO2 and CH4 gases are absorbed and attracted into and within the solution in their nano state, and then they are allowed to amalgamate to produce gel like residual solution and then form nano sediment.

Where, nano CO2 can be extracted or removed from the solution as solid matter at room temperature.

The CO2 stays in state of gel as nano materials in the solution due to their diamond crystals structure (sp3), or due to their crystal lattice structure, this creating a non-adhesion molecular bound like diamond crystals. Which, this does not allow more than certain number of molecular structures to adhere to each other. Where, CO2 nano matters once reaching a certain molecular gravitational level, as crystal, then they do not attach or cannot hold on to more than single or certain numbers of nano molecules of CO2 and form independent or nano cell of matter. Thus, all clusters of CO2 in solution have certain dimensions and that is why they stay like clouds or gel within the solution.

We have to explain that it is known tested fact to us that nano-coated materials and nano materials in sp2/sp3-state due to their characteristic of diamond crystal behaviour, they cannot be soldered to each other or soldered to another material. However, at the same time in a peculiar way the outer electron of these materials allows conductivity of matter. What this means, is that, even through they do not or cannot be attached to any other material from their outer structural boundaries, but in being connected with each, these materials allow flow of current and voltage through their outer boundary.

In a way, these behave like quartz materials, but at the same time after a number of layers, they do not or would have the gravitational pull or adhesion capability to allow more layers to be added to their layers.

Where, these layers along their grain and layer, they behave as a superconductor, but across layers in PN junction, depending on the position of PN cavity within the layers, they become best resistors. Where, we have measured the resistance of these layers and this have been confirmed by other scientists whom have tested our material of sp2/sp3, that these materials have shown to have 20 M Ohms resistance threshold barrier. On the other hand, this material on the surface has as good as resistance capacity as air and diamond.

We have further managed to obtain solid matters of CO2 from the system or dry the nano-saturated nano CO2 dilution into nano CO2 powder. Thus for the first time in the world of science we have managed to generate at room temperature and pressure nano matters of CO2 and methane through the same process.

To confirm that solutions obtained in our tests to contain CO2 and CH4 matters and these matters are in nano form, we had to under take laboratory tests of the solution with infrared spectroscopy and XRD respectively.

Further, as we know through our previous developments in nano technology, that nano materials are by principle sp2/sp3 in their character. Thus from results obtained from infrared spectroscopes and XRDs of matters during our process of absorption and nanolisation, we have concluded that the extracted CO2 and CH4 materials in solution and dry are both in nano state of matters of these gases.

At the same time parallel researches which we have found on the internet since the publication of the first addition of this paper on the 23.12.2009 done by other scientists, Dr Omar M Yaghi of the university of California released on 7.12.2009, which extracts of this has published in article by The New York Times published 4 on the 8.12.2009, this article refers to “metal-organic frameworks (MOFs) material promise of carbon capture and release, and that these materials being crystalline sponge hybrid lattices of organic compound and metal atoms that has a huge internal area where gas molecules can be absorbed."

Where we have reported these characteristics of metal-organic sponge effect and capture of CO2 in 2006, and this has been confirmed by indecent testing by major nuclear testing centre in Europe in their report in 2007.

Further, in the paper Dr Yaghi states, "***MOF used in the study contained magnetism atoms***"

Where in our previous disclosures and patents applied, we have declared that the absorption of these atom of CO2 and CH4 are Magravs (Magnetic gravitational fields) based. Which this independent scientific paper confirms our finding that CO2 and CH4 are gravitated to the water as nano molecules, and they do not have any interaction with the medium of the water, but the water *is the container and tool or catalyst to keep these captured molecules in their singular or nano state and sp3 floating condition or as a solution.*

In the same paper, Dr Yaghi refers to how material separated out CO2, which allows methane to pass.

Where, our system not only captures the CO2, but also at the same time it can capture CH4 as we have reported before. That these captured carbon dioxide CO2 and methane CH4, produce diamond structure nano matters or crystals, with metallic behaviour like, which they have a superconductor characteristics and behave like conductors.

Thus, our findings are inline with other scientists and their scientific research works, which have been reported. Where now we can clearly state these captured gases can be converted to solid nano materials, which behave like metals, and are superconductors, but in reality, their superconductivity of the matter is due to their sp3 nano characteristics and not their metallic condition.

Further Dr Yaghi states that 87 percent of the CO2 captured, at 13-degree temperature are released back by the MOF matter to the environment.

Where this is inline with our reported behaviour, that as the CO2 are extracted by these materials we have developed, which our capturing materials are metallic and organic in their real structures, these materials then release their captured CO2 in the water at room temperature in Gans (Gas-Nano-Solid) state.

Further, he states that to release the balance 13 percent of the captured gas, one-need temperatures of 175 Fahrenheit.

Where, we release all captured CO2 as nano-material into the basic solution of the system, this all at room temperatures without any heat being applied.

Further more, the Canadian company Mantra-Venture-Group has announced on Business News Network that in conjunction with 3M using electro reduction of carbon dioxide, they confirm that ***they need 6Mw of energy to extract one tonne of CO2***, which this is considered to be at the edge of technology of CO2 extraction.

Where, with our new technology we produce energy rather than consuming energy.

Where, further MantraEnergy in conjunction with use of the CO2 they capture, they can produce formic acid, using the technique developed by Dr Gabor Laurenczy. Where, he has confirmed that he can generate hydrogen from formic acid. (Video-interview:

http://www.mantraenergy.com/mantra-in-the-news/squeezeplay-zapping-carbon.html).

We clearly have shown and absorbed CH4 not only from the environment, as similarly we have reported the production of formic acid through the natural gravitational forces of our system as has been shown by laboratory tests results which are discussed below.

We have reported the production of CH3 in our system late last year as we managed to separate the CH3 and its capture in nano state of matter in our system. The production of radical CH3 from the H2O molecule of the water as the magnetic and gravitational catalyst of the system, subsequently this leading to existence of two radicals in the systems at the same time, one of the CH3 and one of an atom of Hydrogen leading to release of molecules of oxygen.

Scientists at a nuclear centre have confirmed the production of hydrogen radical through our technology in 2007 in their independent (not yet published) report.

Therefore, it has become a simpler method through our development to extract CO2 and CH4 and released it in nano solid matter in water, and at the same time have the ability to produce energy and oxygen as we have demonstrate it to independent engineers in December 2009.

Further on in this paper we show and present pictures of both matters of CO2 and CH4 in solution and solid state for the first time (Fig. 29 for solution and Fig. 30 for powder of CO2, and Fig. 31A-B for solution and Fig. 32 for powder of the CH4). The CO2 in solution and dry state has a white milky appearance and methane CH4 has a blue greenish colour.

We have mentioned Dr Yaghis' paper in this section as the findings in his paper is near to our finding and is one of the latest in this field.

The CO2 capture is well documented subject on internet and we are not here to list them to support our claim, one can refer to these on the internet to find that our claims our inline with main stream physics community, with the difference that we reported the capture of CO2 in early 2005 in nano materials on our web site and now we are reporting the first solid nano material of the gas as in Gans-state. We call a Gas-Nano- Solid state the GANS-state, which is a state or phase of transition of matters, which has bee unknown in actual science up to the release of this paper.
Meanwhile in light of our latest discovery and disclosure of the capture of CO2 as solid or in solution, our foundation will ask the nuclear centre which has done the independent test for our system in 2006-2007, that if we can release part of this report confirming the CO2 capability capture of our materials.

The CO2 absorption

In our tests and development phase, we have managed to produce a special material for the production of a new and a simple system for absorption of CO2 through the principle of Magravs field forces, rather than the chemical method of absorption of these gases. Where, the absorption principles through gravitational systems have been disclosed in full in the book, which we have published in July 2009 (titled: The Universal Order of the Creation of Matter; ISBN 978-94-6087-001-9). In this book, I have, explained in detail, the principle of gravitational and Magnetic field positioning and attraction and repulsion of matters from these systems respectively.

Through the development of these new composite materials of organic and metallic nano layers, we have managed to create specific static gravitational field forces which can be in the magnetic wave length of the CO2 molecules, or we have developed condition(s) that as the gravitational field attracted the CO2 to the material, the Magnetic field(s) within these layers cause the repulsion or the release of the captured molecules of the CO2 from the layers. Thus causing the creation of nano molecules of CO2 in the water, and as we use the magnetic fields of the water for stability and as a magnetic catalyst, they come to allow the formation of clusters of Gans of CO2 and create a solution of these nano matters, which these latter manifest themselves as a sediment in the liquid of the system at the bottom of the holding container (Fig 30E).

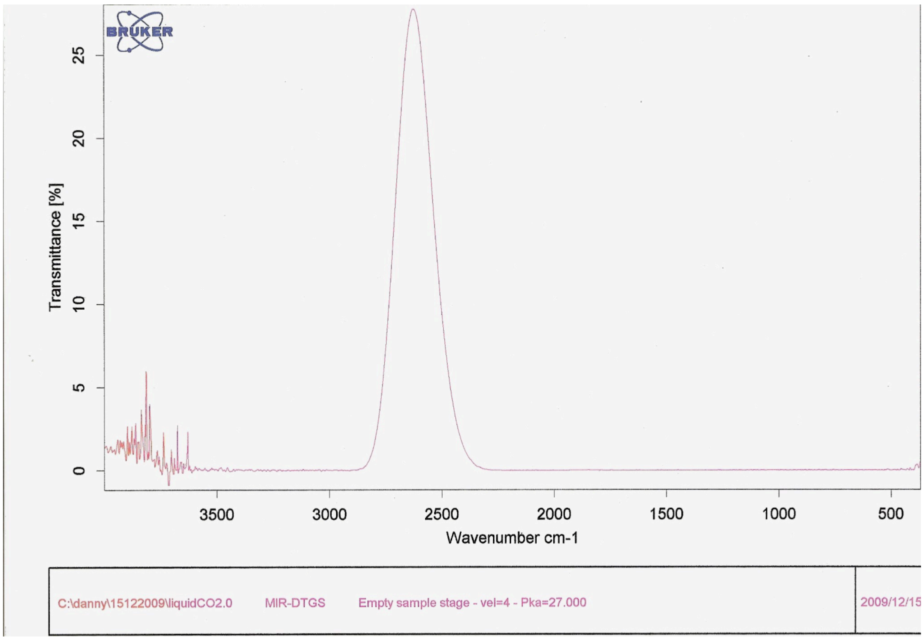

Fig. 21 Infrared spectroscopy of the CO2 solution.

The solution of nano carbon dioxide CO2 and has been subjected to infrared spectroscopy by independent laboratory confirming the production of the first nano CO2 in a solution or as gel at 2630 cm-1 wavelength of the spectrum (Fig. 21).

This is actually the best observe data as calculated and predicted figure is said to be at 2640 cm-1 (Fig. 22). (Source: http://science.widener.edu)

The closeness of the figure of 2630 showing the purity of the CO2 solid matters within the solution obtained during our test.

Where even the predicted matter has a multiple peaks, where these captured matters due to their singularity in the measured sample has a sharp and clear single peak as has been shown in Fig. 21.

This sharp peak can confirm the possible nano sp3 structure of the suspended capture matters of CO2 in this solution, once XRD test was carried out for the same sample.
For further confirmation of nano structure of the matters captured within the solution and then the solution was allowed to dry and then as a film was subjected to XRD testing.

The results of this test showed a clear match only as superconductor characteristics for this material. Further analysis with the data from XRD, a superconductor material characteristics perfect match was not found for this matter.

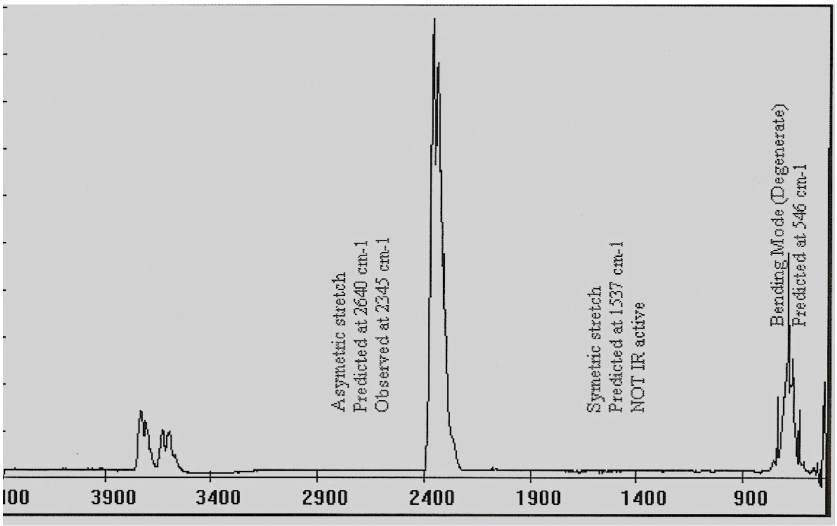

Fig. 22 Standard CO2 infrared spectroscopy graph.

With this test, and from our pervious developments and testing, clearly it is shown that the matter within the solution has superconductivity property (Fig. 23), which has new non matching data in the present collected reference data, and this is the characteristics of nano materials in sp3 state of matter.

This matter as superconductors known or ever been recorded for observation.

Thus the XRD of the CO2 data as solid and its graphs have become the new base knowledge and fundamental basic data for any future capture of these gases, once they are absorbed as a nano material in solid state.

From our previous developments and tests, we know that all nano materials show and have superconductive characteristics. Now with this test, now we have further proof that the captured CO2 in the solution is in a nano state of matter.

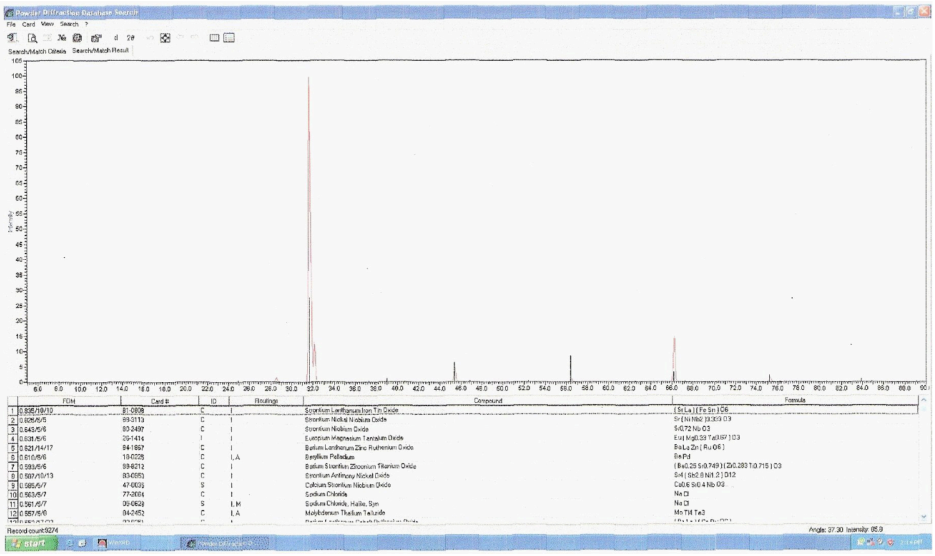
Powder Diffraction Database Search
File Card View Search ?
d 2θ
Search/Match Criteria Search/Match Result
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Intensity
6.0 8.0 10.0 12.0 14.0 16.0 18.0 20.0 22.0 24.0 26.0 28.0 30.0 32.0 34.0 36.0 38.0 40.0 42.0 44.0 46.0 48.0 50.0 52.0 54.0 56.0 58.0 60.0 62.0 64.0 66.0 68.0 70.0 72.0 74.0 76.0 78.0 80.0 82.0 84.0 86.0 88.0 90
start

	FOM	Card #	ID	Routings	Compound	Formula
1	0.835/10/10	81-0808	C	I	Strontium Lanthanum Iron Tin Oxide	(Sr La) (Fe Sn) O6
2	0.825/5/5	89-3113	C	I	Strontium Nickel Niobium Oxide	Sr (Ni Nb2)0.333 O3
3	0.643/5/6	80-2497	C	I	Strontium Niobium Oxide	Sr0.72 Nb O3
4	0.631/5/6	26-1414	I	I	Europium Magnesium Tantalum Oxide	Eu (Mg0.33 Ta0.67) O3
5	0.621/14/17	84-1867	C	I	Barium Lanthanum Zinc Ruthenium Oxide	Ba La Zn (Ru O6)
6	0.610/5/6	18-0225	C	I, A	Beryllium Palladium	Be Pd
7	0.593/5/6	89-8212	C	I	Barium Strontium Zirconium Titanium Oxide	(Ba0.25 Sr0.749) (Zr0.283 Ti0.715) O3
8	0.587/10/13	83-0953	C	I	Strontium Antimony Nickel Oxide	Sr4 (Sb2.8 Ni1.2) O12
9	0.585/5/7	47-0035	S	I	Calcium Strontium Niobium Oxide	Ca0.6 Sr0.4 Nb O3
10	0.563/5/7	77-2064	C	I	Sodium Chloride	Na Cl
11	0.561/5/7	05-0628	S	I, M	Sodium Chloride, Halite, Syn	Na Cl
12	0.557/5/8	84-2452	C	I, A	Molybdenum Thallium Telluride	Mo Tl4 Te3

Record count:9274

Angle: 37.30 Intensity: 85.8

Fig. 23 XRD of the CO2 nano material

The closest match as superconductor characteristics to CO2 in the solution or in dry state of matter, this was found to be an exotic mixture of five different materials, as can be seen in the table at the bottom of the graph, with the best match reference data figures to be about 80%-90% superconductivity properties of this matter.

With the discovery of this new state of matters and the simplicity of their production and their high superconductivity properties, we predict that gans (GAs Nano Solid) of gases are going to be the next-best purest superconductor and resistor materials of the future.

Further more: in confirming the matter of solution being carbon dioxide and at the same time confirming that this matter within the structure of the solution and the system set-up has the capability of conversion to formic acid in presence of CO2 captured in the solution of water, a further test was carried out with **infrared spectroscopy** to confirm this conversion principle of the confirmed nano CO2 capture to formic acid within the solution.

For confirmation of absorption capability of nano CO2 by the solution and its conversion into formic acid, the chamber of the infrared spectrometer system was field with pure CO2 and measurement were taken. This measurement is shown as line A on graph Fig. 24.

Then the solution of nano CO2 from our test was then injected into the chamber and the graph line B in Fig. 24 was obtained.

The peak region NC in the Fig. 24 shows an increase in carbon dioxide in the solution was once it was introduced into the chamber.

This showing that the CO2 solution is extracting carbon from the chamber or causes reduction in carbon content of the chamber.

This in away showing the absorption capability of the solution of nano CO2 of more CO2 on its own independently, and confirming further from previous results that nano materials are CO2 absorber by their nature of construction, even though these matters are themselves part of the original solution.

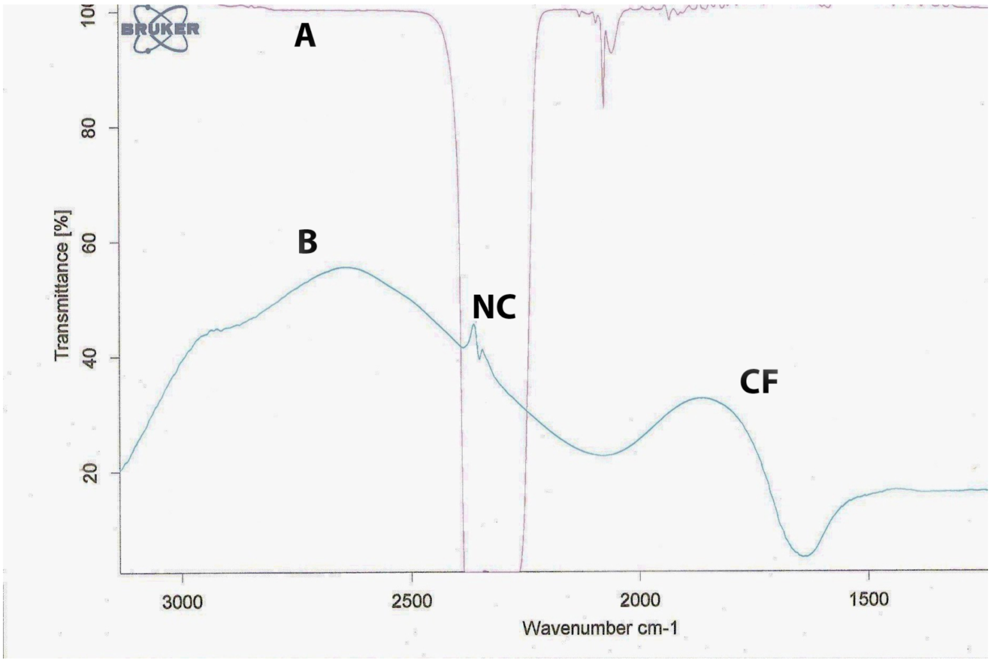

Fig. 24 The comparison graph of injection of the solution of CO2 into CO2 gas environment.

In this graph the peak NC showing absorption of CO2 from the chamber and this followed by formation of formic acid as shown in the CF region of this graph through infrared spectroscopy test Fig. 24.

The graph of the CO2 from the solution and its conversion is independently shown in Fig. 26.

It is at this point of gravitation attraction and subsequent Magnetic field repulsion that the magnetic fields strength equal to the strength of the Magravs of the proton and the electron are made available to the system, which these fields are absorbed by the electrodes within the system and then these are past to cooper wire atoms', which are the constituent matters of electrodes and cause the vibration of electrons within the copper wire leading to creation of current in them.

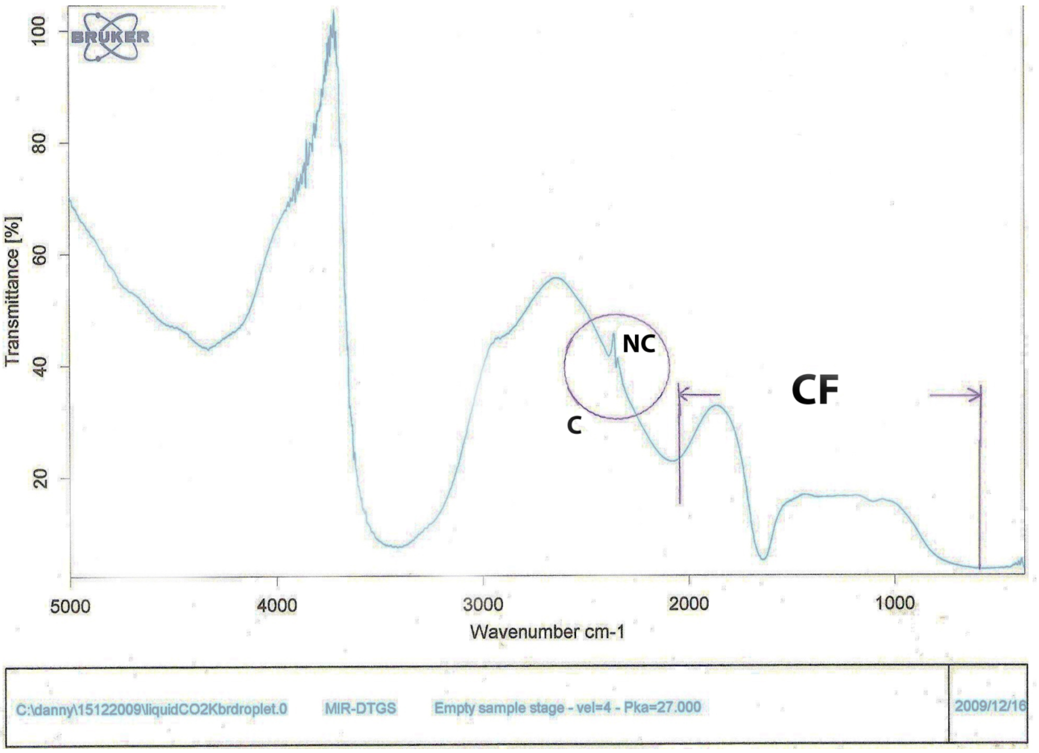

Fig. 25 The graph of the conversion of the CO2 into the formic acid within the chamber.

In the actual physics they call these released packages of magnetic fields or energies as photons as they are released which are then these are absorbed by electrodes placed within the system and used for example to light diodes.

Through development of this technology, production of power without burning fuel and at levels that can meet the demand of a household or a car, is achievable.

Where, for separation of CO2 from environment, in present state of art technologies, energy is used to be able to first separate the CO2 and then cause the formation of formic acid reaction chain, this leading to production CH4 or hydrogen.

Where, through our simple technology, in opposite to the norm, we ***generate energy while extracting CO2 from the environment.*** The method we use is the principle of gravitational positioning and is exactly the way by which matters are extracted and separated in the universe as has been explained in this book.

Our approach is totally different than the present state of art technology, where they use and follow the fundamentals of chemistry, hence the need for their system for energy input to achieve the separation or extraction and we follow the universal method of extraction and holding where magnetic fields as energies are made available through the process.
The example of energy production during the process of CO2 extraction is shown by lighting the diodes as can be seen in Fig. 26.

Where, this set of three LED lights being powered by CO2 capture process. These lights have being on contentiously from 7 pm Saturday 11.12.2009 until this report is release 05.01.2010. Further we expect that these lights will be on for months to come, if we allow the set-up to be running as it is. The intensity of the lights from these LEDs has not diminished and in fact, sometimes they are brighter.

Through development of this technology, production of power without burning fuel and at levels, which can meet the demand of a household or a car, is comfortable achievable.

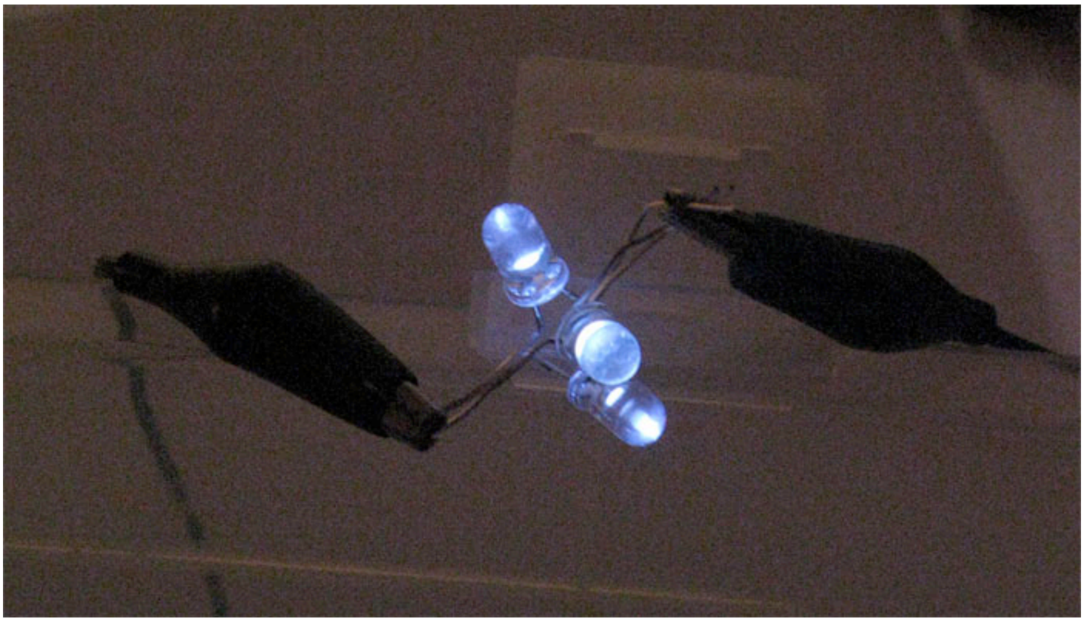

Fig. 26: Group LED's light

For further investigation into confirmation of CO2 capture and its conversion in their interaction with water to form formic acid, the standard reference infrared spectroscopy graph and data of formic acid was obtained as shown in Fig 27.

To confirm further the conversion of the CO2 into formic acid from the CO2 nano solution from our test, the following procedures were carried out.

To begin with, formic acid liquid with 95% purity was introduced as film on the sample disc of infrared spectroscopy system and the graph FA in Fig. 28 was obtained.

Then a drop of solution of nano CO2 was placed on the film the formic acid and the graph CO2+FA in Fig. 28 was obtained.

When the three graphs are superimposed, that is the original infrared graph of the nano solution of CO2, on the graph of the formic acid and then the graph of the mixture of formic acid with CO2 solution, these clearly showing the absorption of CO2 in the zone 1 in Fig. 28, and then presence and formation of formic acid as shown in the zone 2 of the Fig. 28.

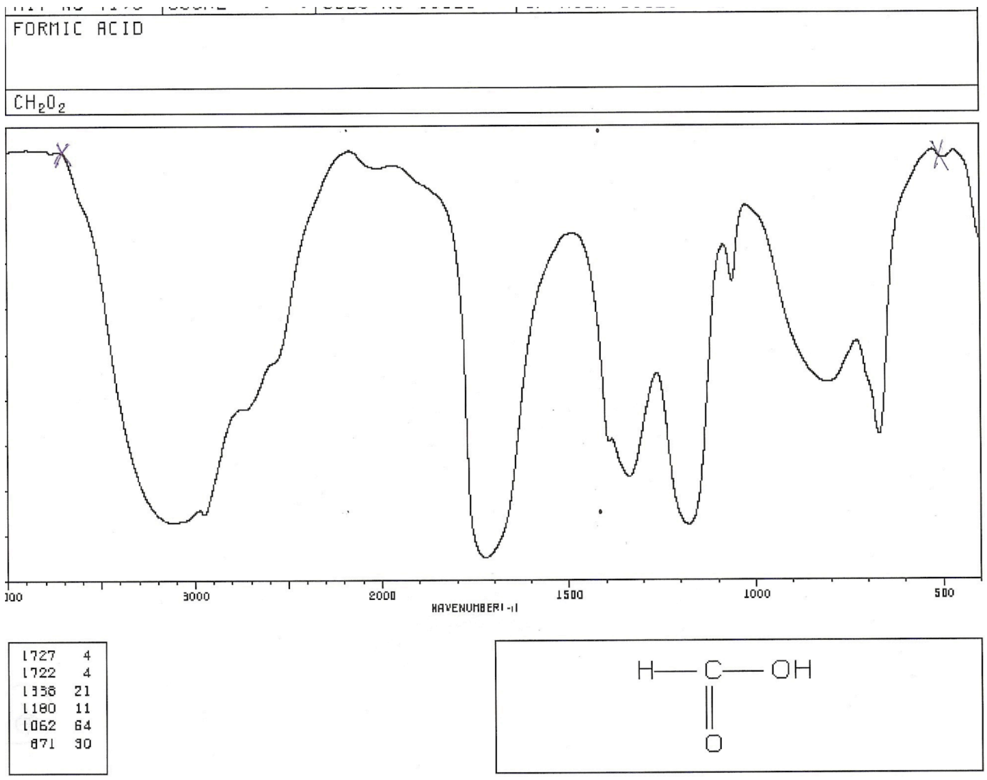

Fig. 27 The standard reference of IR spectroscopy of the formic acid.

This showing the same pattern of presence of Formic acid in the solution of nano CO2, hence not only confirmation of existence of CO2 in the nano CO2 solution, but at the same time the absorption capacity of nano CO2 solution of CO2 from the environment and conversion of these additional CO2 in conjunction with water into formic acid.

Having established the characteristic of the matter within the solution from testes and results shown in Figures 21, 23, 24, 25 and 28, without the shadow of a doubt, the white material is CO2 in nano state, and from Fig. 23, we have the confirmation that this matter is a nano structure in sp3 and behaves as a superconductor.

Further, as the process of the energy production from the system has been extended to days and weeks, the following characteristics and materials have been detected and collected from the system.

In the process of the running the system, the area around the electrodes, which have been specially treated, start gathering a white cloud around them after subjecting the electrodes to certain conditions. From this point on the environment around the electrodes start creating a mist of white material around them, and then gradually the whole containment becomes like milky colour liquid, and then white sediment is established in the container.

Through the infrared spectroscopy shown in Fig. 21, we can confirm that the material in the solution is carbon dioxide and the nano state of this material is very much confirmed by the second test of XRD, as the only match for the dried solid residual of this milky substance shows' superconductivity behaviour of this material. This confirming the nano characteristics of the CO2 as white powder, or in its solid state of matter.

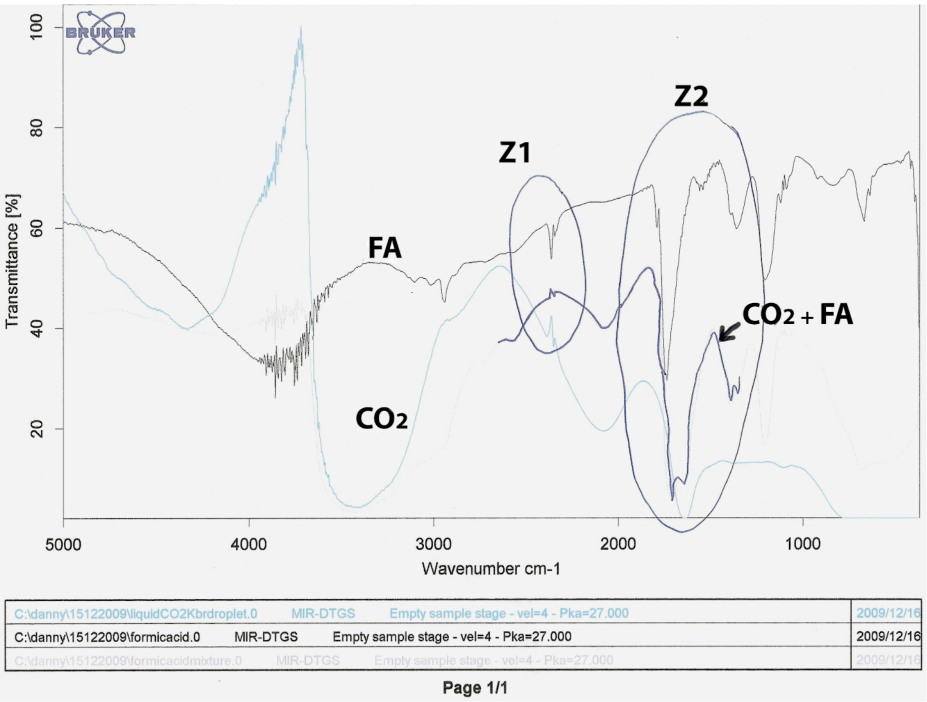

Fig. 28 The superimposed graph of CO2 and conversion o Formic acid and formic acid and mixture of formic acid and solution of nano CO2.

This is for the first time in the world of science that CO2 as gas not only can be captured without use of any pressure, temperature, or energy, but this matter can be extracted and held as a solid at room temperature.

The picture of concentrated nano carbon in solution used for infrared spectroscopy leading to graph in Fig. 21 is shown in Fig 29.

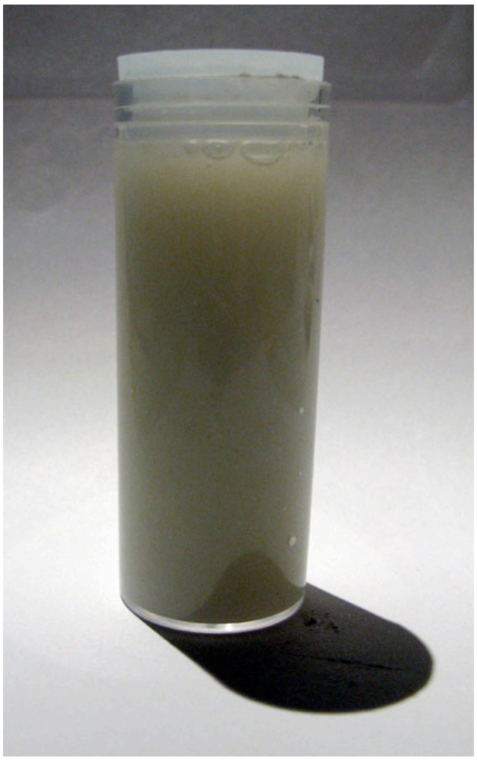

Fig. 29 Solution of nano-CO2

To our knowledge, this is the first picture of nano CO2 solution produced and maintained at room temperature and pressure ever been recorded.

Photographs of dried powder of nano CO2 are shown in Figures 30 A, B, C, and D. These are the first pictures taken of nano CO2 in solid state in atmospheric condition.

This is important as nano CO2 or **CO2 has *never* been observed in solid state at room temperature** and the only solid state of this gas known has been in frozen state of this matter as like of ice blocks (Dry ice).

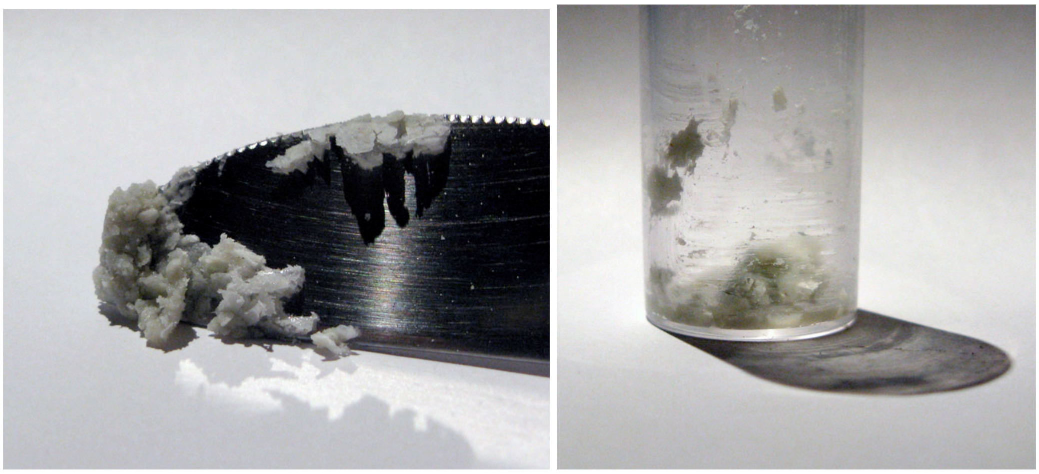

Fig. 30 A, B: Solid nano-CO2 on a knife and in a vial

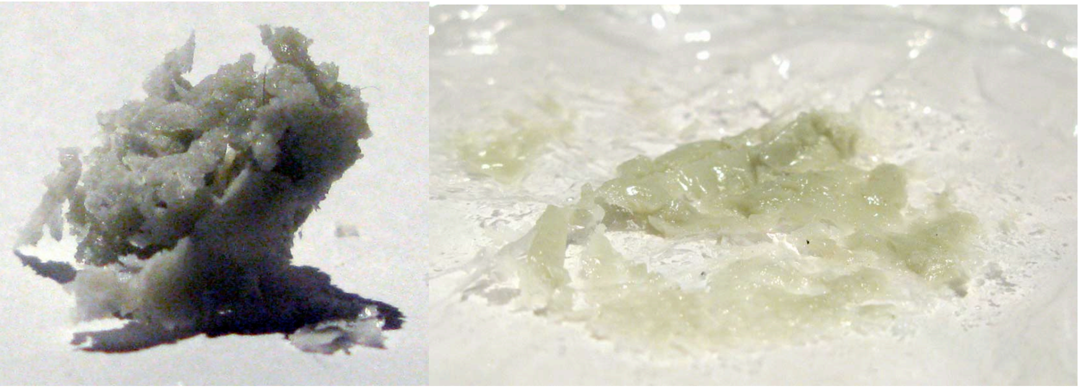

Fig. 30 C: Solid and gel nano-CO2

The important point to note is that, there is no content ratio between the CO2 absorbed in the liquid and the water, as is the usual procedure to state in most tests like these content ratios between sediments or gel and the water of the solution.

With this method of CO2 absorption technology, as long as one can extract nano CO2 from the environment or surrounding environment of water, the CO2 absorbed can be removed without ever a need for the change of the water.

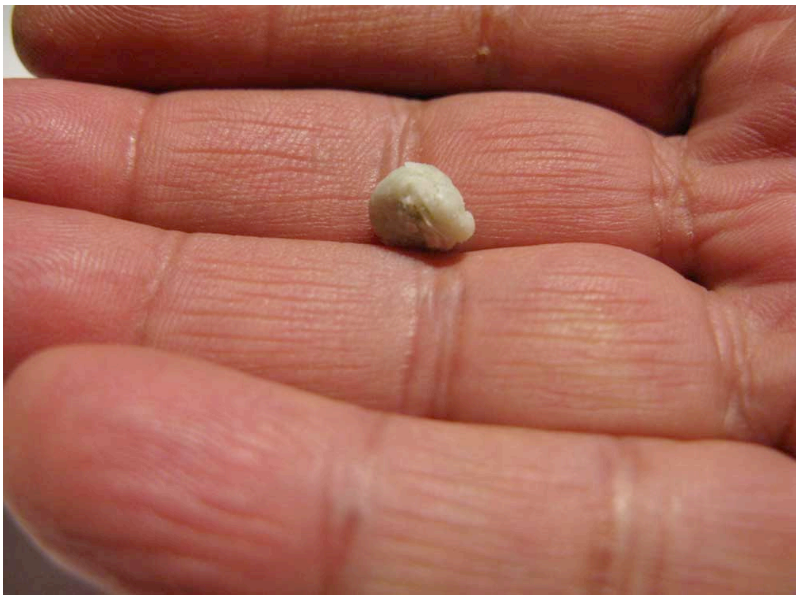

Fig. 30 D: The Solid CO2 material in the Gans-state being held in hand at room temperature. This solid CO2 is not a 'dry ice', but a mastic-like matter.

In practice, we have a number of times collected Gans of CO2 from the same water within the container. The only reduction in the water is due to vaporisation or when it is part of the gel.

This is a very important and crucial point, and this is where this technology departs from present science of chemistry, and stands on its own merit as a nuclear gravitational system.

That is where if the system were a chemically based system, the CO2 content of the water would have been fixed and hence there would have been a ratio balance and final content for CO2 in the water and this value would have been fixed.

Due to the new understanding of the gravitational systems, as has been explained in the book (The universal order of creation of Matters), these specifically designed systems, with specifically placed and produced electrodes, these systems can create a strong but invisible gravitational and Magnetic fields within the layers zone of the electrodes, that these electrodes, as gravitational positioning elements, can create ***the gravitational pull*** *that can absorb CO2 molecules from the water and beyond the boundaries of the water*, without the CO2 as a nano molecule ever chemically interacting with the water molecules content of the system and going through the normal chemical processes.

Thus; if this was not so, how could electrodes deep submerged in liquid would have been able to extract CO2 from the air above the liquid and bring them into the material of the electrode for them to be separated as individual CO2 molecules on a continues bases, as we have been collecting these materials from the same container over period of weeks.

Where in this processes, as the materials like CO2 become in nano structure of their own state, then they are sealed like the diamond and hence they achieve a self sustained gravitational system, like a planet, that does not interact gravitational or magnetically with any other matter and attain the state of singularity or nano state.

This solution of CO2 material feels like mercury as gel, and to touch it feels cold, but does not stick to its container that much.

After extracting the water from the CO2 gans solution, the matter takes the form of gel and is cream whitish in colour.

Thus, this is a **static nuclear gravitational and magnetic field extraction process**, which has never been recorded or known up to this point.

With this method of CO2 extraction, there are no ratios between the matters collected as nano matters and the water content of the system, as the residual of gans of the gases can continuously be extracted from the environment of the system, without change in the volume of the water of the system.

The point to note is that, the volume of the water content of the system stays the same. Where, the system caries on gravitating to itself gases of CO2 and other gases from its environment indefinitely. But the total mass and the volume of the system increases as gases are absorbed into the system by the gravitational principle.

In fact, the system gains weight, mass, and volume as it absorbs more gases. This phenomenon was observed after the test was completed in Holland in the lab, and the increase in volume of the content of the system was noted at the end of that meeting, that we have more liquid in the bottle than when we started the test after about four hours of testing and running the system for energy production trail.

This is normal phenomenon, as extra gases from the environment of the system have been absorbed in the system and converted to gans, which settled at the bottom of the bottle of liquid.

The snap shot from video of the matters content of gans of CO2 as misty residual matter collected after the test in master bottle is shown in Fig. 30E.

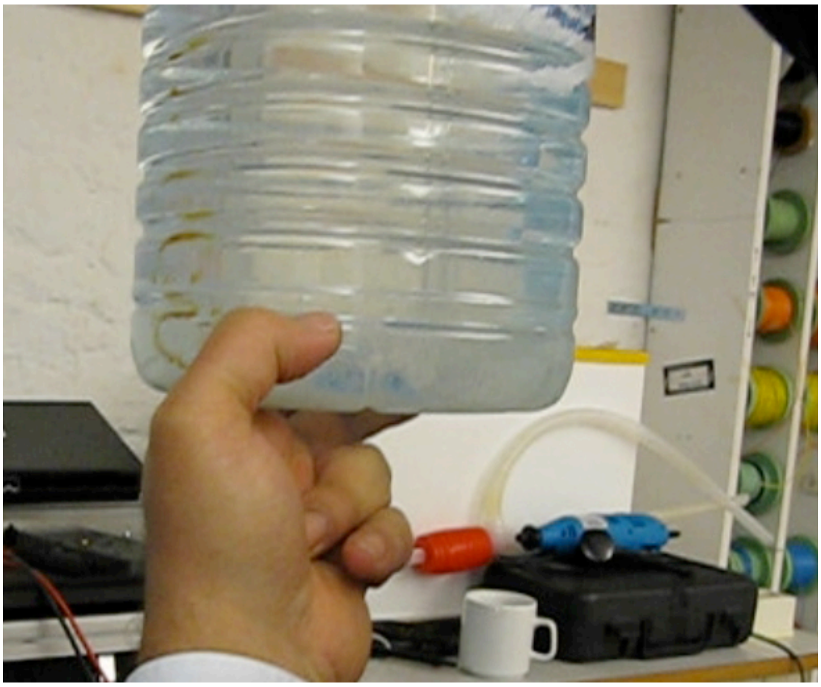

Fig. 30E A misty residue at the bottom of the bottle.

The principle and the technology used for the extraction of CO2 and its containment as a nano matter can be used for making and allowing the production and development of vacuum systems or singular atomic weight extraction systems, which can attain high purity vacuum condition needed for space technologies of the future or in room laboratories for extracting certain molecules or germs.

This vacuuming technique can be achieved without use of any suction system or high-vacuumed pressure pumps as is the norm in today's technologies for production of interstellar space level conditions that are hard to achieve. These systems do not even needs to be adjacent to the system or near the space where the extraction is needed to take place.

In the universe, this singular Magravs strength principle is the exact method used by certain planets or stars and solar systems in managing to extract only certain materials from their solar or galaxy environment respectively. An example of mainly mono Magravs system is few elements material content, being the like Saturn with mainly helium content matter.

In fact in this system we have shown, a micro universal selective gravitational, which in this scale we only absorb CO2 gases from the environment. Where the operation of these systems can be changed and modified to extract any material from any environment without the use of suction or chemical process and so on.

The methane CH4 absorption

During the process of running the system for extraction of CO2 and production of useful current, due to specific configuration setting of the system, specific section of system starts the absorption of methane gas from the environment.

At this point, the water content of these sections of the power system starts to become blue greenish or violet blue in colour, similar to reported colours of the methane in solid state.

Further to confirm the nature of this new material, the same tests as for the CO2 gas extraction was undertaken and results obtained were as follows.

Through infrared spectroscopy test, it is has become clear that this solution has the ability to facilitate the CH4 capture and the release of the residuals of CH4 as solid in the in the dilution of the liquid through the same principles as described above for CO2 gas.

This effect of separation of CH4 as solution shown in Fig. 31 A and B as blue-greenish water and the same colour sediments in the tube.

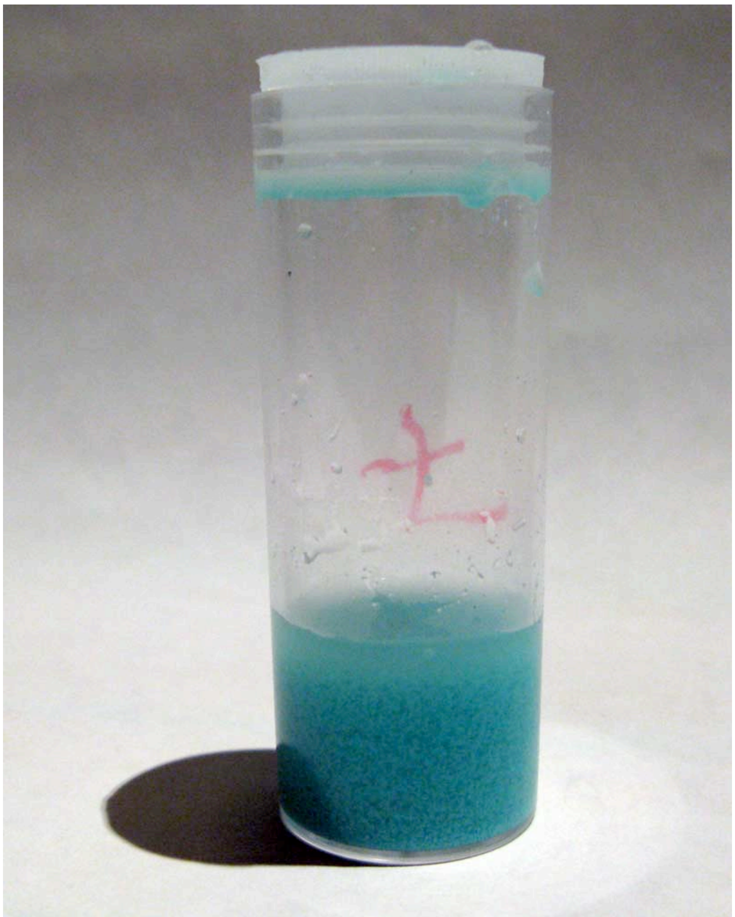

Fig. 31A. Solution of nano Methane (from two different angles)

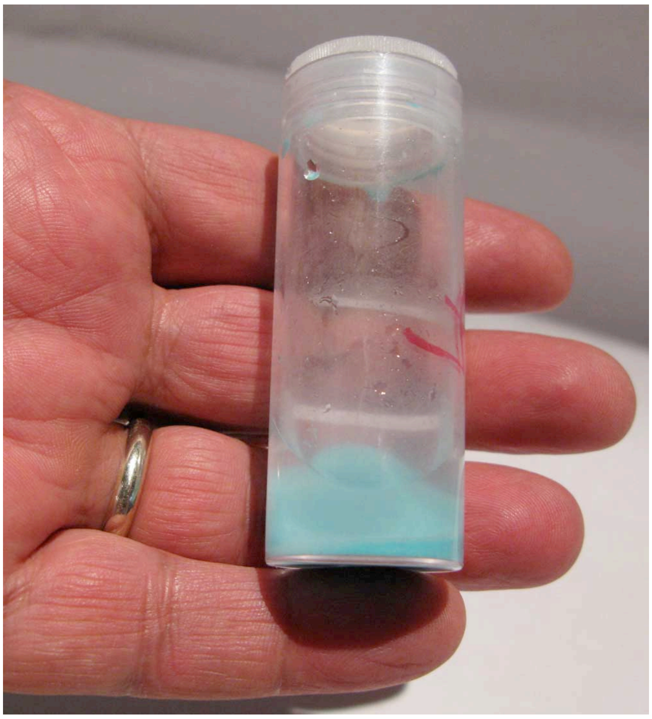

Fig. 31B. Solution of nano Methane

The solid dry residual nano CH4 powder is shown in Fig. 32.

Fig.: 32. Solid nano CH4 (gans of Methane)

The process of creation of current and the observation of appearance of residual of solid matter indicates two clear processes:

Firstly, the system starts to absorb methane gas from the environment through the same principle as described through mainly Magravs process as for gas of carbon dioxide CO2.

With the difference that in this case different and specific conditions Magravs strength level by the layers of matters' of electrodes are generated, that this new Magravs strength level facilitates the extraction or absorption of the new material by the system and within and through the system.

As the gravitational and Magnetic fields needed for extraction of the CH4 Magravs are slightly different than the Magravs fields for CO2, hence different combination of Magravs is needed to extract the CH4 from the environment.

Hence, in being able to collect and further test the new material from the solution to confirm the extraction of CH4 from its environment, this proves that, the system has the ability according to its pre-set conditions, to create different Magravs strength within the layers of the electrodes, that can absorb or gravitated different materials to the system such as CO2 or CH4, through operation of the same system concurrently.

Secondly, another process can take place within the system, where the CO2 absorbed in the system goes through process of interaction with water and initiates the production of CH4 and release of molecules of oxygen.

To achieve the production of methane gas through this process, one molecule of CO2 and two molecules of water H2O are required, that by using the energy absorbed by the material within the system leads to production of CH3 radical and then CH4 as molecule.

Where, this process is not a chemically binding rather is Hydrogen plasma gravitated and it is sustained through magnetic field strength within the layers of coating or matters and energies within the system facilitate such a process.

In conventional chemistry, it is assumed at least energies equivalent to eight photons are needed to achieve such a conversion from CO2 to CH4.

Through this method of using our new system, the material of the systems acts as catalyst and magnetic field provider, that such energy (eight photons) are not required, but in fact gravitational and magnetic fields of the plasma of the oxygen and hydrogen of the water become the magnetic field binder for this chain of events.

This being the reason why in similar cases and in theoretical chemistry one needs energy to trigger such conversion, where as through our system the hydrogen and it sister radicals in presence of the free magnetic oxygen, the solution becomes the magnetic bridge, allowing the system to release energy in larger quantities than eight photons, rather than in parallel chemical process in needing energy for the same process to take place in these systems.

In this second path of production of CH_4, one needs catalytic effect of the material used in the set-up of the system. In the nature, this catalytic effect needs eight photons of energy for it to take place, which this is usually provided through sunrays.

In the case of our system the conversion of CO_2 to production of CH_4 and O_2 takes place through the layers structural materials and the localised gravitational fields which they generate. As these materials have the ability to absorb infrareds rays too as well as sunrays in the environment and hence these new materials can facilitate the production of the methane and oxygen gases and release of energy twenty-four hours a day.

Thus in fact with the development of this new system, we have clearly shown that for process of photosynthesis, one needs the Magravs strength in the infrared spectrum. Through this process we have clearly shown why plant do their CO_2 to CH_4 and O_2 conversation mainly during the night time, where the infrared radiation from the planet is high, as the sun magnetic field can not suppress their emission upwards from the earth surface. Where, it is known fact that the earth is the best suppler of infrared rays.

Further on, to show the correctness of our technology and that we do not require energy to extract CH_4, where in fact we produce energy as power supplier through our technology, the engineers at the lab proposed to connected our system to an electric fan, to see if the poser produce can run mechanical devices like this on a long term. As can be see in Fig. 33. Where we ran the fan for over an hour.

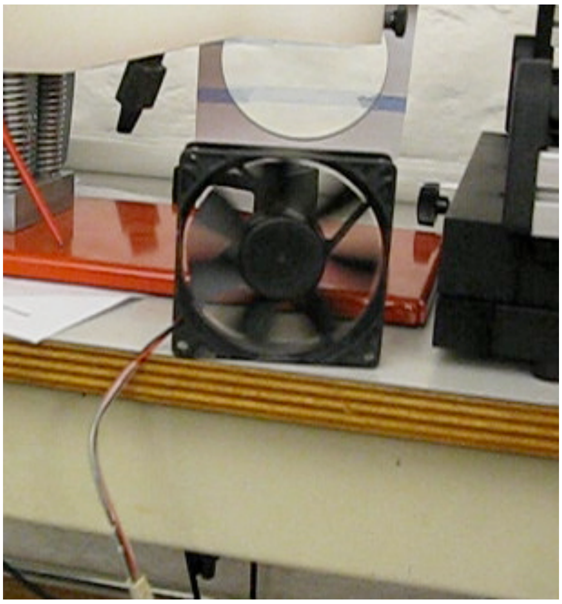

Fig. 33: Turning fan powered by CH_4 extraction system.

This test was done in presence of a journalist and engineers from the magazine that we cooperate with and all has been videoed. The test happened in the basement of their premises, in the absence of any sunlight.

Energy release by the process of gravitation and extraction of CH4 creates more energy and stronger current than the same process of absorbing CO2 from the environment. Where, this process leads to creation of strong current rather than stronger voltage output from the system.

The absorption of CO2 from the environment and conversion of it into formic acid and release of energy needed for the photosynthesis and where at the same time the CH3 stage of methane gas as methyl becomes first stage for the production of sugar bases in organic materials has been observed and videoed with operation of one of systems.

Thus for the first time the process of photosynthesis can be shown to take place exactly as happens in a nature like in a tree through these systems can be demonstrated.

These systems not only absorb CO2 from environment, but through their process of conversion, they can lead to production and released of oxygen as molecular structure.

To confirm and be able to show such a process takes place within the process of the production of methane and oxygen and hydrogen from our system, it is essential to show that methane in nano state of matter as we have collected in our system is or has a superconductivity behaviour, as was shown for CO2 nano material.

For the confirmation that the captured CH4 solid material is or has a ***super-conductive*** behaviour, we subjected the solid residual of the blue-greenish matter to XRD testing in the laboratory.

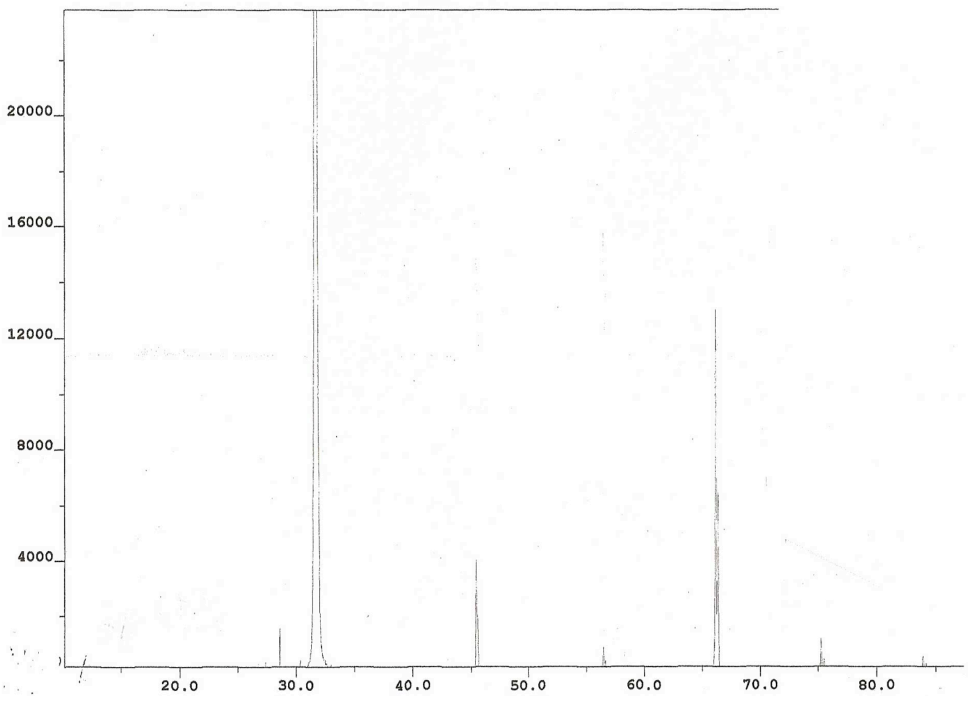

Fig. 34 The XRD of the superconductor characteristic of nano Methane solid

The result of this test has to be noted that CH4 as CO2 in nano state has superconductivity behaviour detected seen in XRD is shown in the graph of Fig. 34.

The nearest material composite match to this material characteristic of our solid matter is beryllium palladium, which matches by 95% as a superconductor behaviour only. The list of near enough matches as superconductor to CH4 in nano state can be seen in the table shown in Fig. 35.

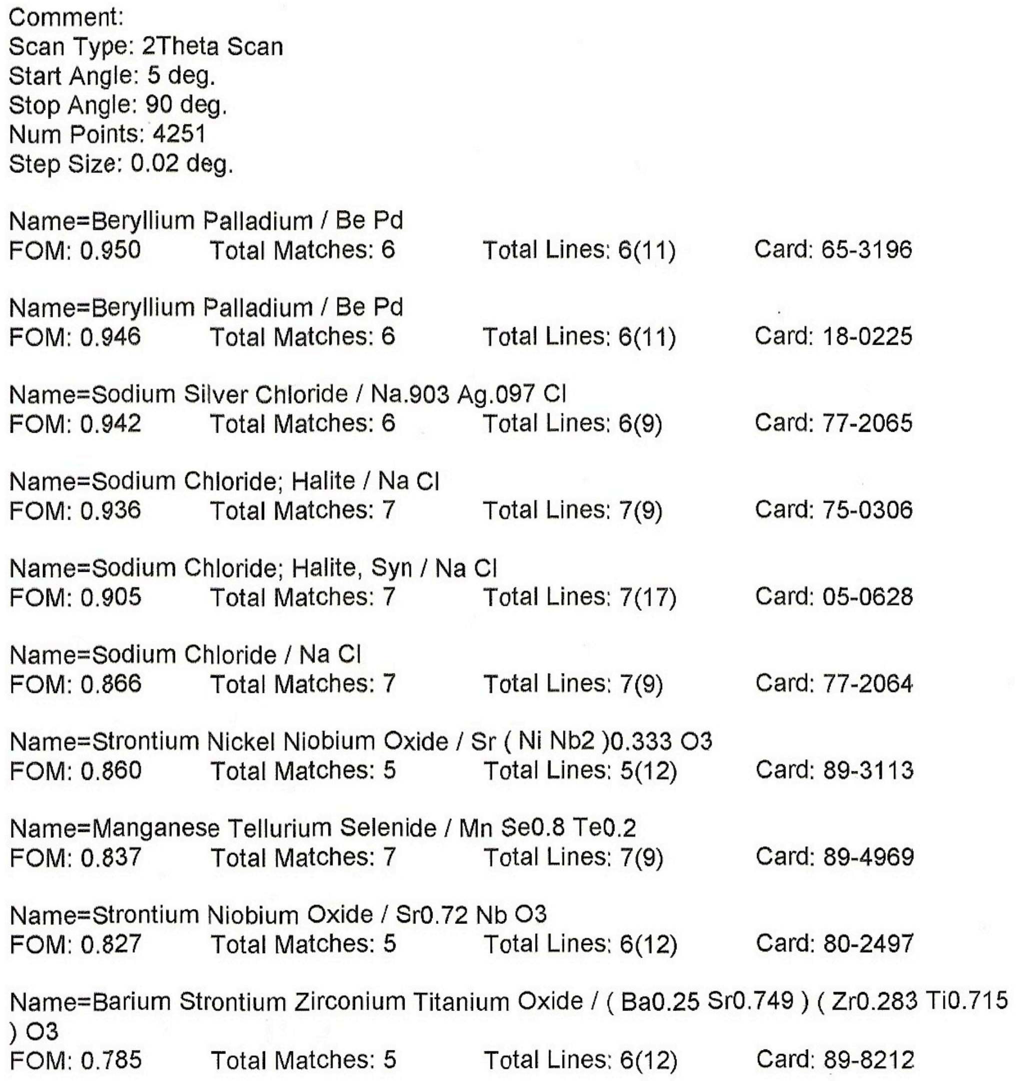

```
Comment:
Scan Type: 2Theta Scan
Start Angle: 5 deg.
Stop Angle: 90 deg.
Num Points: 4251
Step Size: 0.02 deg.

Name=Beryllium Palladium / Be Pd
FOM: 0.950      Total Matches: 6      Total Lines: 6(11)     Card: 65-3196

Name=Beryllium Palladium / Be Pd
FOM: 0.946      Total Matches: 6      Total Lines: 6(11)     Card: 18-0225

Name=Sodium Silver Chloride / Na.903 Ag.097 Cl
FOM: 0.942      Total Matches: 6      Total Lines: 6(9)      Card: 77-2065

Name=Sodium Chloride; Halite / Na Cl
FOM: 0.936      Total Matches: 7      Total Lines: 7(9)      Card: 75-0306

Name=Sodium Chloride; Halite, Syn / Na Cl
FOM: 0.905      Total Matches: 7      Total Lines: 7(17)     Card: 05-0628

Name=Sodium Chloride / Na Cl
FOM: 0.866      Total Matches: 7      Total Lines: 7(9)      Card: 77-2064

Name=Strontium Nickel Niobium Oxide / Sr ( Ni Nb2 )0.333 O3
FOM: 0.860      Total Matches: 5      Total Lines: 5(12)     Card: 89-3113

Name=Manganese Tellurium Selenide / Mn Se0.8 Te0.2
FOM: 0.837      Total Matches: 7      Total Lines: 7(9)      Card: 89-4969

Name=Strontium Niobium Oxide / Sr0.72 Nb O3
FOM: 0.827      Total Matches: 5      Total Lines: 6(12)     Card: 80-2497

Name=Barium Strontium Zirconium Titanium Oxide / ( Ba0.25 Sr0.749 ) ( Zr0.283 Ti0.715
) O3
FOM: 0.785      Total Matches: 5      Total Lines: 6(12)     Card: 89-8212
```

Fig. 35 The nearest data of matching materials available that match the green blue powder superconductivity characteristic to the extracted nano methane solid from the solution.

The comparison of superconductivity characteristics of the CO2 and of the CH4 nano materials obtained by XRD and matching and similarity in behaviour of carbon dioxide CO2 and methane CH4 as super-conductive nano materials are shown in Fig. 36.

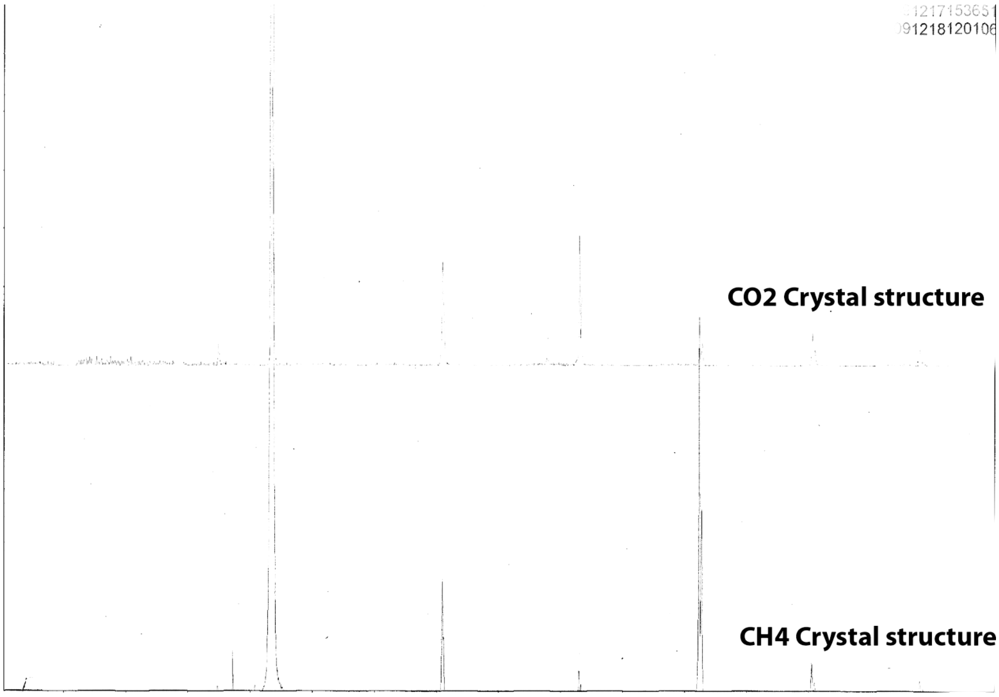

Fig. 36 The XRD superconductivity characteristic comparison of nano CO2 and nano CH4.

The common behaviour between the CH4 and CO2 nano matters have been observed and shown by superimposing their infrared spectroscopy graphs as shown in Fig. 37.

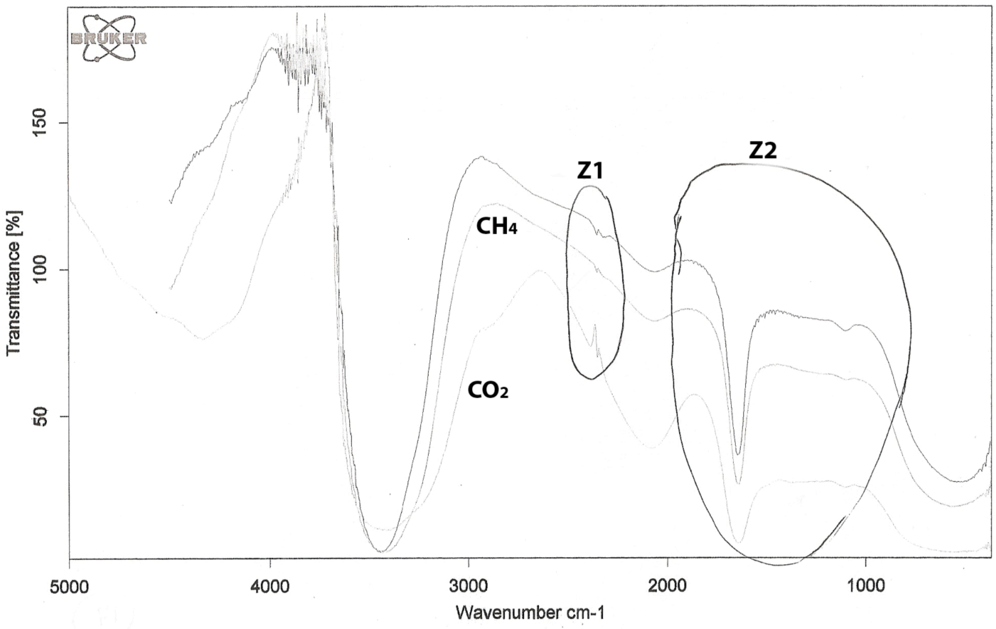

Fig. 37 The IR spectroscopy comparison between nano CO2 and nano CH4 in solution

In this graph the CO2 in zone 1, shows extraction of CO2 gas from its environment and then forms formic acid as seen in zone Z2. Similarly methane (Me) releases increase the availability of CO2 in the environment of the chamber of the spectrum machine as seen in zone Z1, and then the same releases' of formic acid as the same as CO2 as can be seen in the zone Z2 through this process.

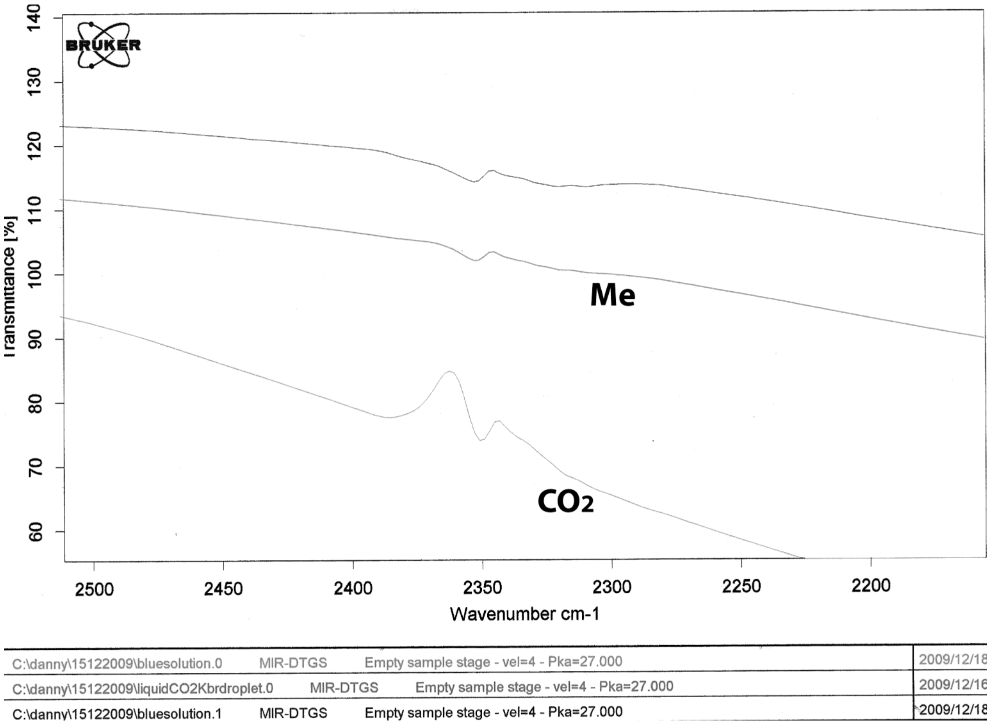

Fig. 38 the enlarged section of the CO2 and CH4 in infrared from Fig. 37.

This explaining two factors, one that the solution is processing both CO2 and CH4 extraction simultaneously and or it is has the ability to produce CH4 through CO2 absorption and the water chain interaction as was discussed before.

Enlarged section of the zone Z1 in Fig. 37 are shown in Fig. 38.

Superimposed graphs of these nano materials of the same region in graph 37 are shown in Fig. 39.

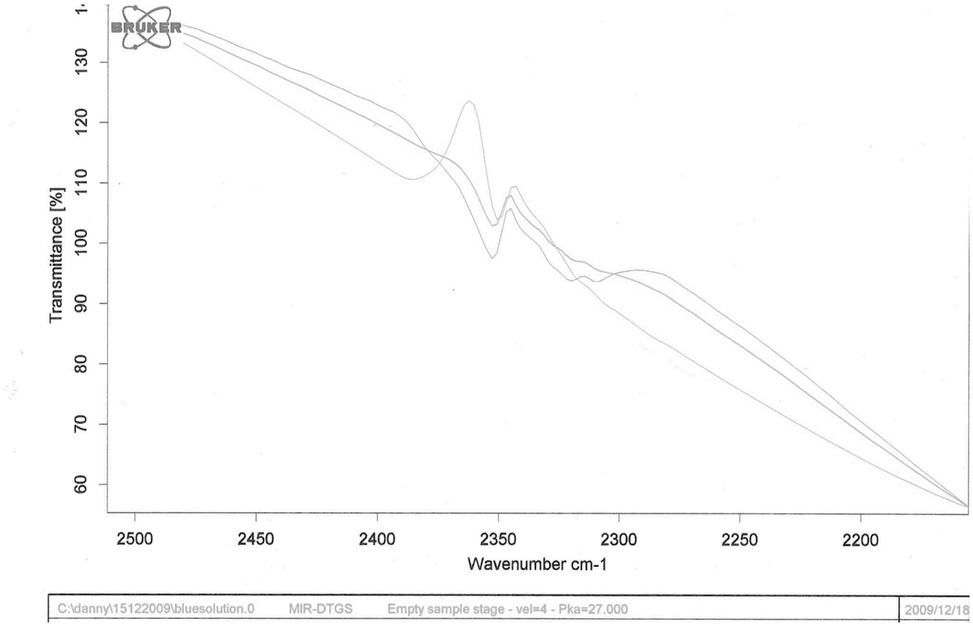

Fig. 39 The superimposed and enlarged section of carbon sector of the Fig. 37.

Observations and notes.

In the development discussed above, one has shown how by absorbing CO2 from the environment, this being on earth or as the content of craft, this gas can be used to generate energy and oxygen to sustain life without the need for additions reserves of these gases using the man himself as the supplier of the main ingredient matter, the CO2.

What this new technology offers for the planet earth is a survival line that in the present environment of the global warming with the CO2 generated by burning of fossil fuels and production of methane gases produced by extensive cattle farming, through the utilisation of this new technology for the first time ***these gases can be extracted from the atmosphere and simultaneosly be converted to solid matters or solution, this process can be used to lead to, and cause the release of energy, and through the use of specially developed composite materials, this leading to separation and generation of oxygen.***

Where some of methane and carbon dioxide in the form of nano materials of the same gases can, all be suspended in the solution or dried as solid and reburied or returned to earth as topsoil.
Where this new technology shows for the first time and make it possible to achieve creation and holding of carbon dioxide or methane gases in solid state at room temperatures and pressures. Through evaporation of water from solutions, solid of carbon dioxide CO2 and methane CH4 gases in the form of powder nano materials of these gases are produced.

This new system and technology completes the circle of life for maintenance and existence in the future while it can clean up the mess man has created by the use of fossil fuels on earth in the past centuries.

Where these new systems will behave as trees on earth and for the passengers of space systems of the future, by simple converting CO2 into oxygen, and where the man himself becomes the supplier of CO2 for the system to keep the cycles functioning.

Through the development of the technology that has been disclosed in this paper the initial steps to conclude this process for application in our space technology is considered to be completed. In the space technology of the future, one has to be able to use all matters available within a spacecraft to attain continues and sustainable existence of man in deep space travels.

This technology of direct extraction of CO2 and CH4 from the environment can happen in reactors of small and large dimensions. Thus the technology can be used to tackle the Climate Change problem. For example large installations can be connected to exhaust systems of heavy industry, and small extraction devices can be made for households or be built in cars. None of these installations will need powering, as they will deliver themselves electricity. On the other hand we don’t have a sequestering problem since the CO2 and CH4 are already in a solid state, which make storage easy. However our liquid and solid CO2 combined with limonene oxide can also be used to make new type of polymers and plastics.

We have developed and attained a fully integrated system that is able to sustainable life in space.

In letting the process of CO2 absorption and production of energy to continue, some cells of the system start the process of production of methyl CH3. This is a significant process as this shows organic process as CO2 is in progress within the system.

Further, this process in conjunction with its water content, and nitrogen from the environment, can itself lead to production of natural sugars, starch and proteins, using our simple static gravitational system.

This simple system can be used not only to clean up the environment from CO2, but at the same time the same CO2 can be used in conjunction with specific matters of the system, to release energy, oxygen. Further at the same time producing CH3 and methane and formic acid and by use of nitrogen this leading to production of protein(s) that can be the food supply for the humans' passengers of spacecraft's of the future while they inhale oxygen and exhale CO2.

Thus through this new technology mans' existence on board of the spacecraft's of the future, the man becomes the container and supplier of carbon dioxide and hence the provider of oxygen for his own existence.
Thus, man himself becomes the conversion system to maintain his own survival in deep space or on earth during shortages of food or natural disasters.

The property of this system has allowed us to submit a full patent for the first full day and night energy panels.

New state of matter

When considering this new technology and understanding the potential of its physical ability to convert gases into solid state of the same, where up to present knowledge the gas from being able to take the shape of its container, for the same atom or molecule becoming the container of itself as a nano matter, then this new state of matter or the solid state of the gas at ambient conditions like the room temperature and pressure, ***itself becomes the nano matter of the gas*** and this state of matter has never been known in the world of science and has to be considered to be a new state of Matter or the fifth state of matter.

Where, through our development of matters in our technology we consider solid, liquid, gas, plasma, and dark matter as the first five known states of matters up to this point in our technological development.

This being due to the fact that this new nano matter of the gas in solid state, it does not behave and looks as any of its old states of matters or any other state of matter as we have known and what we have seen before from these matters.

The infrared and XRD testing of these carbon dioxide CO2 and Methane CH4 gases in nano state showing that this matters behaves like a ***superconductor material***, where nor hydrogen or hydrogen and the carbon in there matters are considered as super-conductive materials in the present world of science known knowledge in their gas state of their molecules.

To our present knowledge superconductors at these levels are by principle of matter in diamond structure or sp3 state of matter. Where, their physical gas equivalent of these matters known to science and are classed as odours.

Thus as in this new state of matter molecule of GAs which becomes Nano of itself and become and appears as Solid state of matter as we have seen from our testes and shown as solid of the matter in this paper, this new state of matter needs a new name, for it to show the source of the matter, that is from GAs to Nano of Solid and to be called **GANS** for short.

This new name in fact clearly not only indicating the source of the matter, which is the gas, but at the same time it clearly explains what to be expected and describes the characteristic of these matters too.

Where these new characteristics being like superconductivity, lightness, singularity of grains and so on, of the same molecule as it become nano of itself and become solid of itself but having full molecular structure of a gas, but self gravitationally and magnetically closed and singular, but it still can be in a gas state of itself, but as it becomes singular atom or molecule of itself it behaves as solid matters do.

Thus from this point on, we do not call these solids of nano of gas or nano CO2, but **GANS** of CO2.

First ever pictures of these matters in Gans or solid state are shown in Fig. 30(A-D) for CO2 solution and dry state.

Pictures in Fig. 31A and 32B shows the first ever solutions of methane and dry or gans of CH4 respectively.

Through experiments and deliberate mixing of different gans in one solution, it is clear that these materials do not mix or do not chemical interact with each other. Where each gans of each matter stays separated from the other matters within the solution, and stays the original gans. Fig. 40 shows a mixture of three gans in a solution, where they stay without any chemical binding, where for example the green residue settles can be seen at bottom of the tube.

Similarly for the first time in the world of science, through understanding of this technology, we can capture odours as solid? But the question is that, would they smell the same as their sp3 counterpart as gases? This is the question to be answered and explored in the times to come.

Parallel with development and extraction of gans, these systems produce energy, which are not due to burning of fuel or chemical conversion like in a battery. The energy produced during the process of gans from any system, which this energy can be used for any purpose, this being creation of gravitational field force to create static sources for creation of Magravs, power to run a motor or light-up a lamp, this new source of energy needs a clear and new name classification.

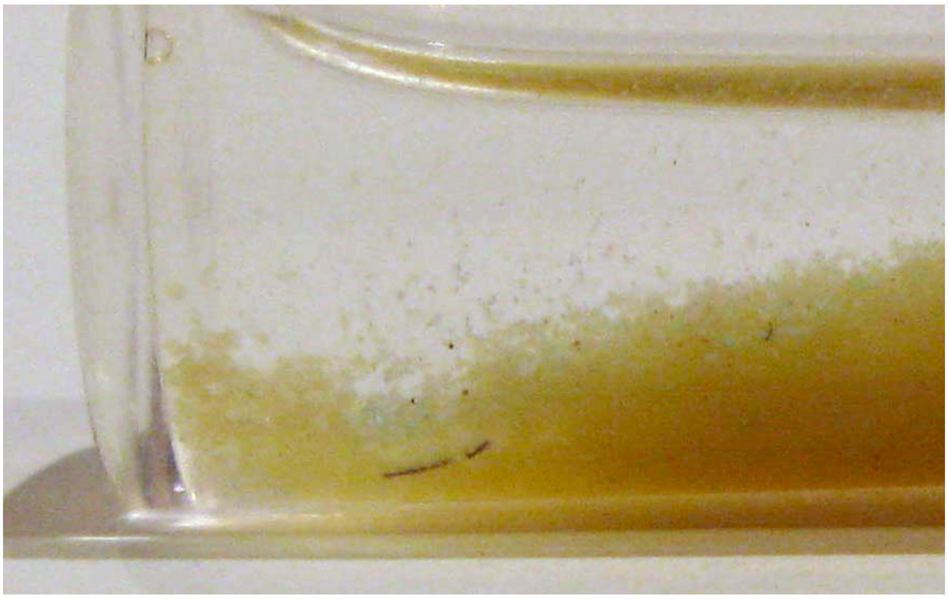

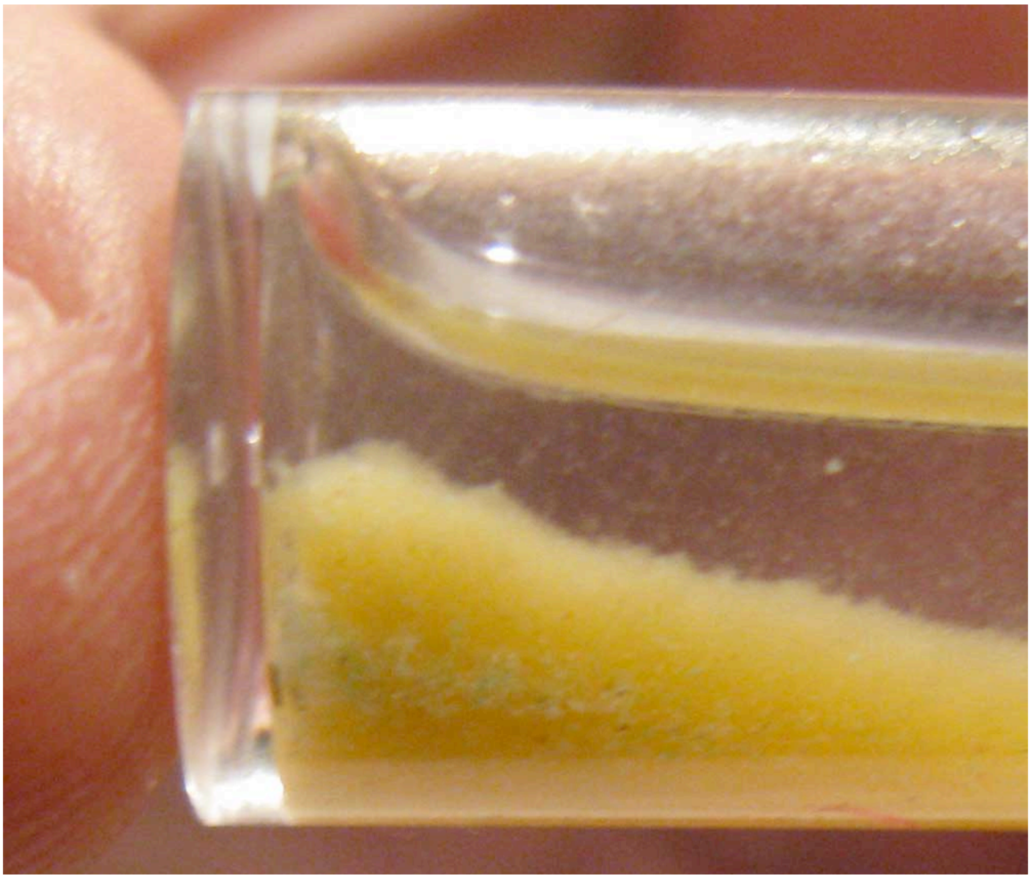

Fig. 40 Different Gans do not mix or chemically bind, and stay separate in the solution.

We call this new method of production of clean energy, this being in magnetic field force strength or electric energy, the **GAs Nano Solid Energy** or GANSE for short or Gas Energy Solid for short **GES**.

Implication and applications of this new technology

In the development of the technology discussed above, we have shown how by absorbing CO2 from the environment, this being on earth or as the content of craft, this gas can be used to generate energy and oxygen to sustain life without the need for additions reserves of these gases using the man himself as the supplier of the main ingredient matter, the CO2.

What this new technology offers for the planet earth is a survival line that in the present environment of the global warming, with the CO2 generated by burning of fossil fuels and production of methane gases produced by extensive cattle farming, through the utilisation of this new technology for the first time these gases can be extracted from the atmosphere and simultaneously be converted to solid matters or solution and return of this matter as solid into the ground,

The development of this new technology shows and makes it possible, for the first time in world of science to achieve absorption and containment of carbon dioxide or methane gases in solid state at room temperatures and pressures.

Through evaporation of water from solutions, solid of carbon dioxide CO2 and methane CH4 gases in the form of powders nano materials of these gases can be produced, where in later stages in different applications like manufacturing processes, these materials can be added to effect the physical or the property of different matters or produce new types of matters with these new materials.

These new matters of gans can be used for example as new biological nano filters, where their purity or the molecular size of gans can determine the matters to be filters, this being cancer cells, germs or other matters.

Through the use of this new technology the circle of life for maintaining a clean environment in the future, while it can clean up the mess man has created by the use of fossil fuels on earth in the past centuries.

Where these new systems will behave as trees on earth and for the passengers of space systems of the future, these systems by simple converting CO2 into oxygen, guaranteeing ever supply of life essential need, where the man himself becomes the supplier of CO2 for the system to keep the cycle functioning and guaranty his own survival.

In our tests, in letting the process of CO2 absorption and production of energy to continue, some cells of the system start the process of production of methyl. This is a significant process as this shows organic process in progress within the system.

Further, this process in conjunction with its water content, and nitrogen from the environment, can itself be used to lead to production of natural sugars, starch and proteins, using our simple static gravitational system.

This simple system can be used not only to clean up the environment from CO2, but at the same time the same CO2 can be used in conjunction with specific matters of the system, to release energy and oxygen. Further at the same time producing CH3 and methane and formic acid and by use of nitrogen this leading to production of protein(s), that can be the food supply for the humans' passengers of spacecraft's of the future while they inhale oxygen and exhale CO2.

Through the development of this technology that has been disclosed in this paper, the final steps to conclude the process for application of these systems in our space technology is considered to be completed.

Where, these CO2 extraction systems, by their nature of operation as described above, they can become the natural backup energy systems and vacuum systems for spacecrafts of the future.

With the completion of this sector of our technology, we have developed and attained a fully integrated system that is able to sustainable life in space. Where we have shown through other sectors of technology, systems for motion with gravitational positioning through our nano materials we have shown how to produce matters in the universe and materials to manufacture for what man needs for food and habitation, then we have shown how to develop the system to feed and produces systems to cure ailments through our medical application section, and in the energy production system we have shown how to generate energy through simple natural process, and finally now with this technology we have shown how to create a self sufficient environment to sustain life with extraction of CO2 and use of the same to create oxygen for man to survive in space.

Thus through this new technology mans' existence on board of the spacecrafts of the future, where the man himself becomes the container and supplier of carbon dioxide and hence the provider of oxygen for his own existence is complete.

Thus, man himself becomes the conversion system to maintain his own survival in deep space or on earth during shortages of food or natural disasters.

The property of this system has allowed us to submit a full patent for the first full day and night energy panels and all other points which has raised as new science in this paper.

Conclusion

Through new and simple methods which has been developed and explained and production of new materials with some new properties which has been confirmed, we have managed to create and establish clearly how gases can be made to be contained, extract and converted into solid state of matter, without the use of any chemical reaction, temperature or pressure, and simply through creation of condition of mini and localised Magnetic and gravitational fields within the system, as is done in the universe.

Where, these fields are hidden within the static layers of different matters within the system, as they have been chosen as the electrodes within the structure of these systems.

Where clearly in picture Fig. 40, one can see that these matters are normal solid and can be handled like any other solid matter.

Clearly from the infrared spectroscopy and XRD graphs of these matters and solution it is clear that the matter absorbed or extracted from the environment, as gans are so pure and in nano state that their spectroscopy recorded data is as close as to perfect theoretical for CO2 gas. This purity of the matter, showing itself in such clear and single perfect lines of graphs shown from the test of the actual sample of the solutions of these ganses.

At the same time, it is confirmed that nano gases do behave as their metallic and semi metallic matters in the universe and they become or attain the behaviour of conductors and superconductors and super resistors. Where this finding of the behaviour of these gans solves and answers a number of unanswered questions in the world of physics, chemistry and biological binding, movement and motion of planets, stars and matters in cosmos.

This paper in its simplicity of its presentation and the knowledge it has put forward, it will in due course bring about fundamental new ways that a number of problems in the scientific world concerning gases and solids are to be considered.

For example we can think of ways to solve the mystery of the ***a-biogenesis*** (the study of how life on Earth could have arisen from inanimate matter.) or how proteins could have been created in Space by local super-conductive conditions.

With the knowledge gained with the operation of these simple power cells, new systems in production of vacuum, creation of matter, extraction of matters and conversion of matters can be developed, and in a way this paper opens a new an unbelievable horizon to the man for science and space technology of tomorrow.

We have shown and clearly proven not only one dose not need all sorts of machines to create or extract matters from the environment but simply following the laws of universal gravitational and magnetic field forces, one can achieve the same in a much simpler ways, and instead of using energy to achieve production of a matter, one can extract energy from the same system and achieve better results in doing the same process.

Additional pictures

Some pictures of different matter, which have been gravitated to the system, are shown in the picture below.

Up to this point we have managed to attain six different elements in gans-state, which each one has it's own colour as can be seen in sample tubes as they have been extracted from the system during it operation, while producing energy. Some of the materials shown have not been subjected to IR and XRD spectroscopy to confirm the composition of the original gas before they were converted to their gans-state.

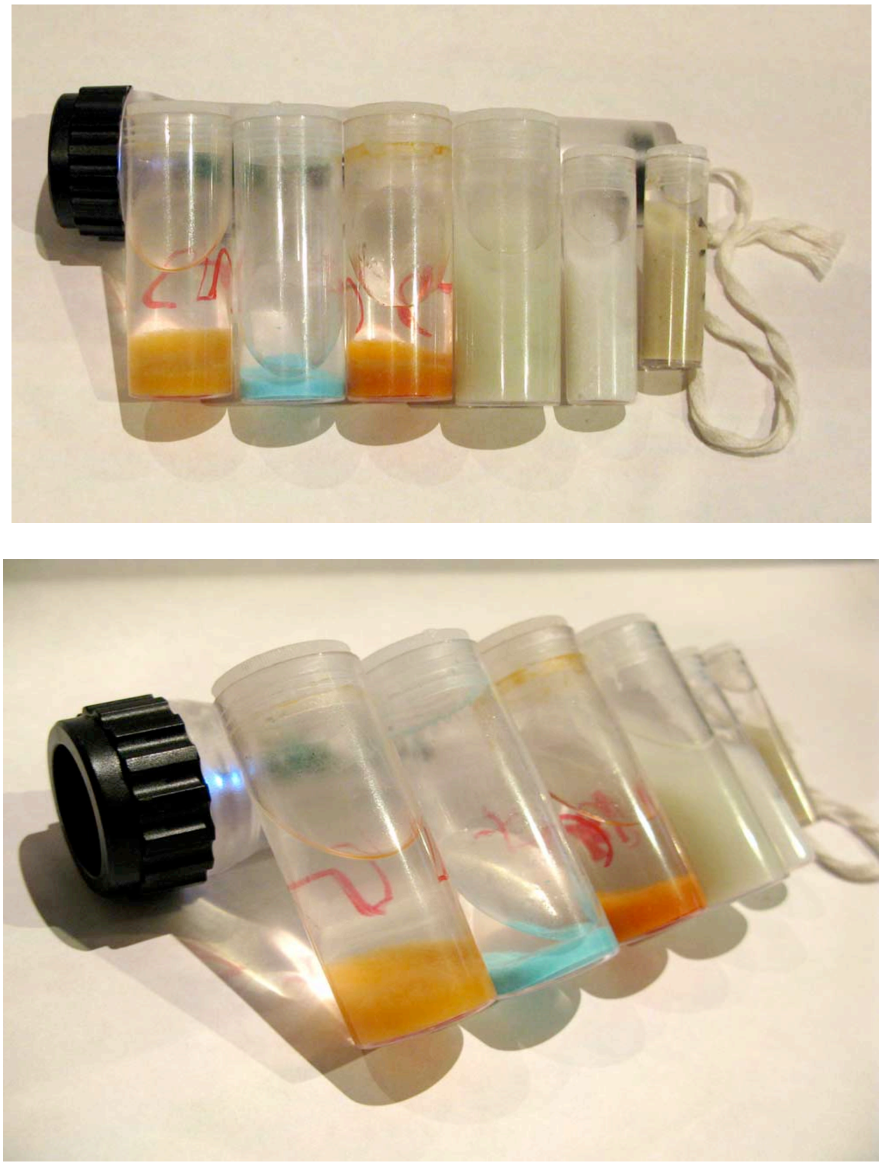

Fig. 41. Six different materials in gans-state which have been gravitated by the system up to today (05.01.2010).

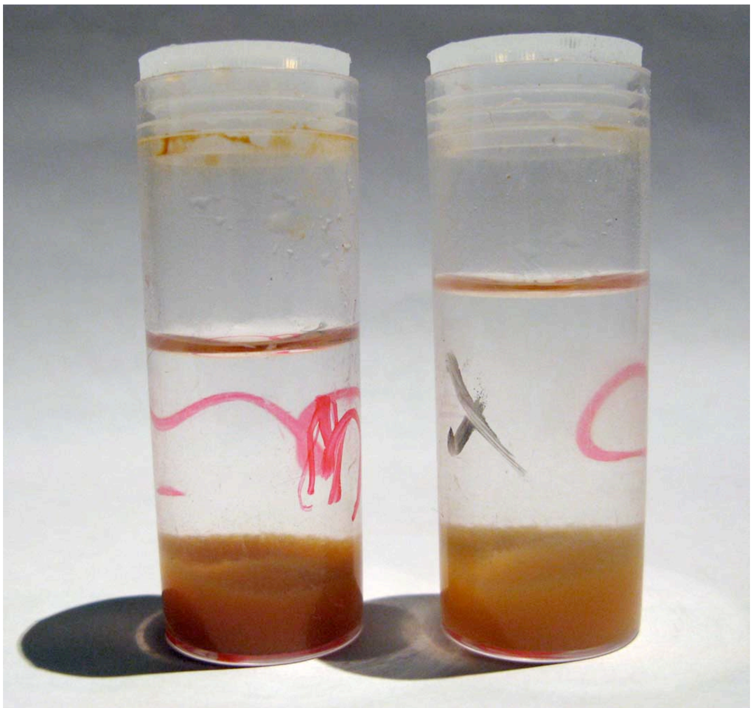

Fig. 42. Two different materials in gans-state, which have been gravitated by the system, which we think these, are organic gases in Gans-state. They have different orange colours.

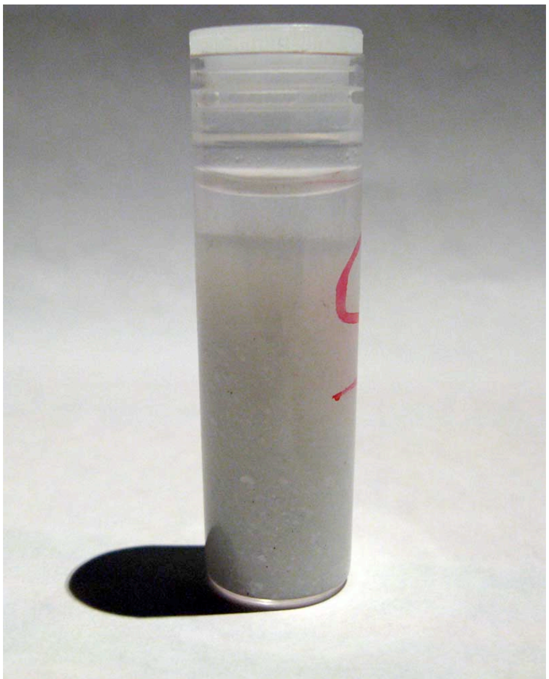

Fig. 43. CO2 in gans-state , which have been gravitated by the system. It contains solid flakes within the solution

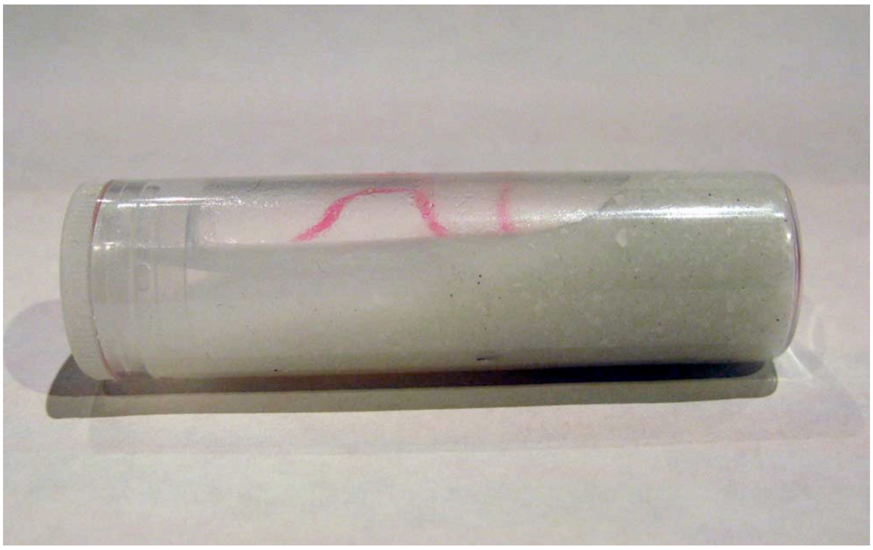

Fig. 44. Different angle of show from the same solution as figure 23. CO2 in gans-state, which have been gravitated by the system. It contains solid flakes within the solution.

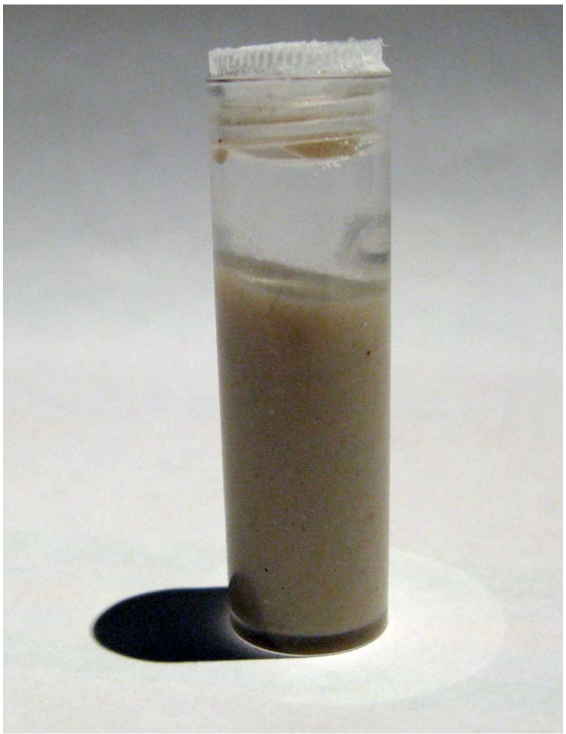

Fig. 45. Unknown composite material in gans-state which have been gravitated by the system.

Chapter 5

The light and its interaction with the Principal star

The Principal Star is what has been commonly called the Black Hole (3) and it is understood that these entities are places, where the principal mafs and matters are create and collected.

The pmf of principal star being covered by dense layers of the transition mafs, this being the reason as why they look to be dark or where there are no lights are transparent in these regions in the specific parts of galaxies in the cosmos (Fig. 46).

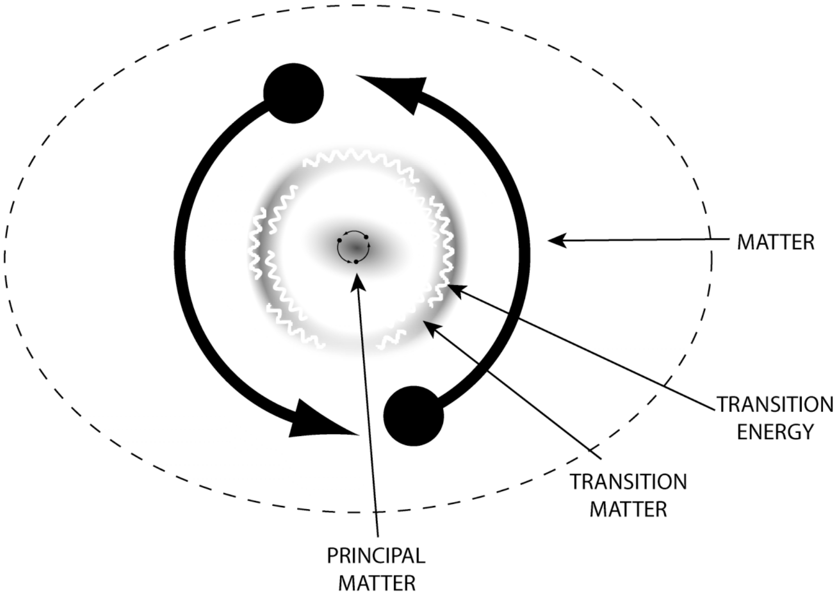

Fig. 46 Structure of the principal star

The light being in possession of the gravitational and Magnetic fields of its own and consequently the light having the Magravs of its own, these fields forces' of the light in matter magnetic field strength environment of the universe take their positions in respect to the Magravs of these principal stars in the centre or any other part of the galaxies.

In these environments in the universe one can observe that as the light rays come close to the principal star the light in majority of cases will not be absorbed by the extensive gravitational fields forces of the principal star, but the light ray finds its way by going around the principal matter star instead of as it being absorbed by this entity. This ability of bending of the light is commonly called the lensing of the light near large objects, and this is the process by which currently scientific world can confirm the existence of the principal star in the universe.

The reason for the bending of the light being due to the fact that as the light possess its own mafs of the principal matter, then the light's principal mafs Magravs take their Magravs positioning in respect to the principal star mafs' and matters' and this positioning of the two Magravs of the principal mafs of the light and the star from the point of an observes from a distance looks as the light bends around these objects or what is called lensing phenomenon of the light around the Principal star occurs (Fig. 47).

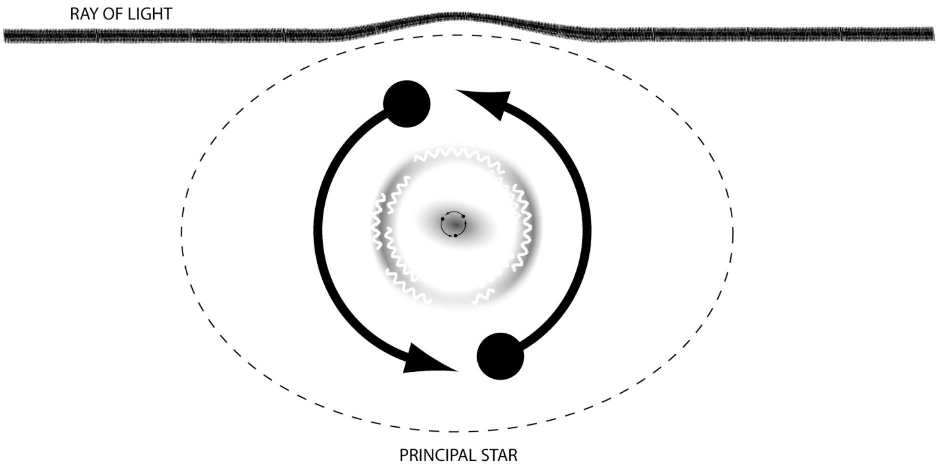

Fig. 47 Bending phenomenon (lensing) of the light around the Principal star due to Magravs positioning

The theory that even the light will not escape the Principal star gravitational field forces once passes the event horizon, this shows the lack of the understanding of the scientists in this respect too, as if the scientists had the tools to observe and could detect the principal mafs magnetic field spectrum, then they could observe that as in matter mafs stars like the Sun of the earth's solar system, these Principal stars will radiate pmfs in the principal mafs strength spectrum and the Principal stars are like bright shining stars in their environment, and as they absorb light in matter's spectrum strength, they at the same time release light in principal mafs and transition mafs pmf strength simultaneously.

In time the scientific world would produce tools that with these tools they will be able to observe the motion and emission of the Principal and transition lights rays from these dark spots in the universe. Then one can observe that light in principal mafs strength are emitted from these stars in the galaxies as light are emitted from the Sun of the solar systems and the disappearance of the light in even horizon is only illusion and not a reality as the pmf of the light mafs get diluted in the stronger pmf fields of principal and transition mafs rays.

The concept that how principal stars are created, how they operate and how they absorb and emits some light has been disclosed in detail in the paper written in 2002 and published in 2006 titled "The creation of Back Hole".

Principal stars if they absorb mafs and of the matters, stars and visible light as it appears to human eyes' vision spectrum, then they can and they do at the same time emit light in the mafs of the principal matters spectrum into the environment of their galaxies too, as this was not the case then the conservation of principal mafs and matters in the cosmos will not hold and this is against all the laws of physics.

In reality the dynamism of the mafs of the principal matters in these stars and the light they emit is the cause and deliverer of pmf needed for the creation of matters and stars in the galaxies.

In a simple way, as the light from the Sun feeds its own solar system mafs of the matters and matters the like of its planets, the same is done by the Principal star through its light emission in the Principal mafs pmf strength feeds all mafs and matters in its captured Magravs sphere or so called its galaxy.

At the same time the Magravs of the Principal star dictates the position and attraction of other matters, stars and galaxies' in respect to itself as the total mass of the matter mafs and transition mafs of the matters in galaxies are about less than five present of the total mass of the galaxies (3).

Thus the positioning and directional motion of the galaxies in the universe, the same as in the solar systems is dictated by the mass of the principal mafs and matters of the principal star of the galaxy and no other factor (3). Therefore the mass of the galaxy is the measure of the interaction of the Magnetic and gravitational fields' forces that the principal mafs contents of the principal matter can produce and maintain.

Once these principal stars are formed in the universe then the Principal stars and the lights which are releases form them become the means of the supplier of pmfs and nourishment of to the environment of these galaxies" (3).

The concept of the structure of the principal's star lights as the source of pmf into its heliosphere zone is applicable the same but in much smaller quantity of mafs of in the operation of plasma of an electron, proton and stars too.

In the modern science the source of the creation of matters in the plasma of electron and proton are mysterious entity, which lead to creation of quarks, bosons and other names content of the plasma, in fact for example so called the bosons are the pmf of lights released by the principal mafs of the plasma, which as this dynamic light reaches a given point and given pmf strength in the plasma of the proton or electron due to reduction in its pmfs strength in motion in the plasma, this leads to appearance of matter mafs and matter of different pmfs strength in all regions of these plasmas.

The mysterious forces which the scientific world has been observing in the plasma of the proton, these are released by the principal star of the plasma itself as lights in the principal mafs strength unwinds and as these lights reduce in strength to different mafs and matters pmf levels strength. Then these slowed down pmfs lead to appearance of different weaker matter pmf strength in the content of the plasma, where and when two matching field's strength and matters strength can interact that they can create a given gravitational fields forces within the plasma then these points of gravitational field appears as bosons and other named entities within the plasma. At the same time these entities create their own Magravs which the operation of fields forces of these Magravs appear as mysterious filed forces within the plasma of the proton or electron.

Theory of Higgs and Higgs boson concept: In fact if these scientists would have understood that the interaction of at least two magnetic fields (bosons) leads to creation of gravity (attraction) and Magnetic fields (repulsion), which further interaction of these two fields of Gravity and Magnetic fields leads to creation of the mass of the matter (Higgs), then Higgs would have understood that interaction of magnetic fields not only in plasma but in any level of magnitude in the universe leads to creation of matter which this matter always possess its own mass and magnetosphere, or in other word when two magnetic fields interact they lead to creation of mass which is the maximum strength of the gravitational strength less the Magnetic fields strength created by the interaction of these two secondary fields (1). Thus Higgs theory is correct and in reality one does not need to smash an atom in accelerators to observe and prove his theory, but it would be easier to the same behaviour of production of matter ***if two rays of light are fused*** that **they lead to creation of matter** and thus the Higgs theory would be proven to be correct in a simpler way.

This is process of the production of bosons, Higgs and other entities in the plasma of proton is the replication of the same process as the operation of light rays in the solar system changes to matter in different pmf strength regions in the heliosphere of the star, where in the case of plasma of proton ***the same process takes place in micro/nano level*** of the mass of the principal star and principal mafs and matter in size.

Stars in the centre of the galaxies in motion near the Principal stars even though they release rays of light but not all these rays get absorbed into the principal star or the stars themselves and in reality the Principal stars do not absorb all lights from the matter stars (the sun is a matter star) into themselves as the plasmas of the lights in the matter mafs pmf strength contain their own principal mafs and matter and these principal mafs of the light create their own pmf Magravs, which this Margrave allows the light to take its position in respect to the Principal star which is so close to, and this Magravs positioning of the light allows the light from the star in the matter strength to move away from the principal star as is observed by the astronomers. In fact the speeds and the distance of the star from the principal star is dictated by their own principal's mafs content and the Magravs strength.

The matter stars in their inner sanctum contain and hold on to large quantities of the principal mafs and matters (1), This meaning that even the Sun in its process of feeding light into its heliosphere, at the same time in this process the Sun *releases a large quantity of the magnetic fields of the principal mafs and matters as the* ***principal light rays strength of magnetic spectrum*** in this environment.

It has to be said that in galaxies, the high speed of the rotation of matter stars near principal stars has to be extremely high, this being due to the fact that as the matter star itself contains principal mafs and matters and at the same time these mafs of the principal matters interact within the structure of matter star that they create their own gravitational and Magnetic fields that the matter star takes it principal mafs Magravs strength positioning in respect to the mafs of the principal star matters and mafs Magravs. Thus the only way that matter star can hold on to its pmf structure is to attain a high speed as its mafs of the principal matters of the matter star content have to find a constant continues renewing its new position in respect to the principal stars matter's and mafs' content. These high speeds of the stars near the principal star has been puzzling the cosmologist, but if they look close enough and as they could understand the principal of the rotation of an electron in respect to the proton of an atom (1), then the speed of the matter's star with composite matter and mafs content near the principal star shows and is flowing the same path and the same principle of motion as an electron does in respect to the proton in the nucleus of the atom as electron moves fast around its nucleus to maintain its position and hold on to its content matter (1). Thus now may be the high speed of the matter stars near principal stars can be explains and the single directional orbit of the star matter in these regions confirms content of the Principal star as to be mainly as of the Principal mafs and matters, that this singular mafs and matters contents' of the principal star allowing the more circular motion or elliptical motion of matter stars with their large mass near the Principal stars.

This is the same process as to replacing the proton with the principal star and the electron with the matter star, with parallel high speed dynamic motion of the matter star with controlled Magravs positioning around it nucleus. The reason that the motions of the matter star around the principal star is not as unpredictable as of motion of electron around proton (1); this is primarily due to mono content matters and mafs of the principal star. In fact in these cases the light observed by the cosmologists created by the matter star is mainly due to the interaction of Magravs of the two matter and principal stars, and at this point the Magravs of the matter star has to find rapidly its position in respect to the Magravs of the principal star, and this is how the matter star become to attain speeds' in the region of the speed of light in matter pmf strength. Thus the majority of the visible light of the matter star in close proximity of the principal star is not of the star itself but it's due to the interaction of the Magravs of the two stars. Now one can see another marvel of the universe, where matter mafs in creating light which itself is the light of itself.

This concept of mass moving at speed of light in matter pmf strength environment due to interaction of principal mafs strength will have such a profound effect on the man's scientific progress and in man being able to travel the depth of universe and this will be the point of and the tool of detection of future space crafts in motion at speeds beyond the speed of light in matter mafs strength and above in the cosmos, once the concept of the light rays in the pmf strength of principal mafs strength are understood fully.

Chapter 6

The light and its planetary interaction

The interaction of different lights of different matters (principal, transition and matter) can be utilized to create the appropriate magnetic fields and currents and gravitational fields' forces as it is needed for any purpose, then this concept of interacting lights is the way that food and energies from the right light mafs levels can and will be produced in deep space travels of the future, as one can take the right components and amounts of the mafs in matter field strength needed of any required levels of mafs as/and for nourishment of the mans' physical body, without man actually eating tangible food.

This conversion and absorption of light rays is in fact what plants and animals do in the earth environment as they absorb specific plasmatic magnetic field strength of light spectrum and due to their atomic and molecular Magravs field's strength they create measured vacuumed condition within their structure to and for transformation of a given pmfs' strength of the light to be converted this to gans of that mafs in the structure of the plants or animals. Which, these ganses have Magravs equal to Magravs strength of vitamins or minerals levels as the planet produces or the animal needs for existence in a given environment. The process of production of minerals and vitamins in understanding how light is and can be converted to gans in earth atmospheric condition will allow the production of the same matters and vitamins by use of the light matters' structure in conjunction with the using a simple dilution Magravs positioning reactor anywhere in the universe (1).

For example man by eating a given plant, he absorbs that spectrum of the fields transmitted and absorbed from the spectrum of the light mafs of the matters fields for specific use in specific protein in man's body, and now man has the knowledge to produce the same proteins and minerals from the light rays without the need for plants or animals as an intermediary conversion toll for this process.

In earth atmospheric Magravs condition herbs absorb pmf of the higher strength of the same light as normal plants do and the meat can absorb the complex strength magnetic fields of the light. The structure of herbs pmfs attraction and storage is more dependent on the pmf emitted by the light spectrum from the earth fields as the fruits store most of the light spectrum which are entered into the atmosphere from above and that is why each entity has different mafs strength storage and conversion capacity and capability. Where, the same light from space in different distance and interaction with the magnetic fields strength of the light of the earth leads to creation, storage and production of different pmf spectrum of the light by the same plant and animal DNA.

The ray of light in its entry into the magnetosphere of the earth and the interaction of the two Magravs lead to release of specific characteristic rays of specific magnetic fields strength, that this ray through a special process with the hydrogen atom lead to creation specific division of energy that can sustain life on this planet. This ray and its process of creation, which is not known in the world of science presently has been fully explained in the book "the universal order of creation" to be published in the future. We see that in time man of knowledge would be able to comprehend this process, and once this process used then the wisdom of all creations in the universe will be understood and can be replicated anywhere in the universe.

The rays of light at any strength of its mafs at their entry into any dynamic plasmatic environment have the ability to lead to or through conversion process cause the initial chain of reactions needed for the start of life and then the matters produced use the light itself to further create more cell like beings and then through the ray of light conversion in these cells this leading to final step that the light ray in becoming the creature like man, then using light rays and its product main gains intelligence and the intelligent body at this stage acquires its magnetosphere and Magravs or what is known its ***soul***.

Chapter 7

The use of the properties of the structure of the pmf of the light for space technology

As the man’s knowledge reaches and becomes in line with realities of creation; in the future the mafs of all components of the light and the phenomenon of light rays entanglement will be used for space technology in an efficient way and in manners which are not even comprehendible at the present.

The helixial cylindrical motion and reversibility capabilities of the outer-magnetic mafs of the shell of the light in space reactors of the future will be the key in control of the motion and the directional motion which these crafts can travel and the same time the control of the speed of the motion of the space crafts of the future can be set by using the matters’ components of the light.

The gravitational fields’ forces of the light and the reactors using and producing light gravitational properties that can be used for control of the directional motion are much easier systems to produce in a fully integrated combination space reactors setup that using different mafs’ strength of the light makes take-off and landing possible in every environment and pmf fields strength in the universe, without putting the craft and its passengers ever in any danger. Where, with the use of the lights’ mafs Magravs technology one can reach the deepest part of the universe in time spans that are within the fraction of the life of humans, rather than the present assumptions of thousands of the light years. In comprehending the concept of the structure of the light in full; this will the time, when the Man understands the truth about the world of creation in the universe, and how universe was created and real universal motions are created and maintained.

How, different rays of lights get entangled and leads to creation of conditions of mafs of matters is due to Magravs positioning of one light ray in respect to another of the same. The Magravs positioning of two rays of lights or what is known as the light–entanglement is the condition for example, where and when the light rays can converts from mafs of matters to characteristics of tangible or detectable matter in a given plasmatic magnetic field strength environment.

It has to be noted that there is fundamental difference in light entanglement and fusion of light rays. The ***fusion of the light*** is more fundamental process in production of matters and mafs and this process can be used for production of stronger light rays or for the production of heavier elements without going through the process of the matter state fusion. This being the like of the production of oxygen molecules through the fusion of light rays, rather than multiple fusion processes of the rays of the light’s once the neutron to an atom of hydrogen is reached, and then through multiple fusions processes to reach the production of oxygen molecules.

In fact fusion of the matters of the light does not require energy and in reality does not lead to release of vast amount of energy. Whereas, in the present process of fusion of the matter of hydrogen the present fusion scientists have got themselves tied up with the wrong concept and technologies and they are in the dream of achieving fusion of matters through the use of present laborious fusion reactor technology with very little success.

By the same default in understanding the real process of the fusion in the universe, there has been a misconception by nuclear scientists in production of the heavier matters through fusion of matters in the centre of stars and in their nova process. Where, the **present assumption** that matters are created in the centre of the stars through process of fusion of matter **is not in reality correct**, whereas the process of fusion and production of all matters in the universe primarily takes place at the *point of the fusion of the light rays **first**,* then as the light rays become to the point of matter mafs strength they first *all convert to gans of the neutron* and then through process of the Magravs positioning and collective balancing of these forces of the total light rays in a given environmental pmf strength, then some of the light rays once converted to neutrons some stay as neutrons and others of the weaker strength in the collection go through the process of the neutron decay (1) and become proton and electron of the collection of the light rays and hence in a given pmf strength, the physical matter with given number of neutrons and protons and electrons are created. Therefore, it is the collection of fused light rays which changes to collection of neutrons and then as some neutrons go through neutron decay and become proton and electron, this is *how **atoms of different matters are created** in the universe*. Where, further this process of the light pmf strength matching through light-entanglement process is a much easier process than the matching and fusing physical plasma of atom or molecules and solids.

In fact this is how neutrons are found to take their place in the centre of atoms to start with and this is how fusion takes place in the ***real universal cosmos*** (unicos). This being one of the reasons why the present science have problem to produce fusion as they try to fuse matter as in real term in the universe and in the centre of the stars fused light rays lead to creation of matters of different mass.

Fusion light **ray reactors** are simple practical systems which do not need very much energy for them to run and operate, but they can create and facilitate *directional motion of the space crafts of the future* using their capability to split the plasma content of the light to its sub structure matters and fields and create Magravs of different strength and in a cylindrical reactor of the same to stack the light rays gravitational and Magnetic fields forces for creation of the directional change and at the same time these reactors can be used for the production of the matters for the use in the space environment of the future.

For the space technology of the future there is a need for multiple light reactors that an instantaneous three hundred and sixty degree spherical motion of space crafts can be achieved and the at the same time the shielding magnetic fields can be interchanged from matter mafs to principal mafs and change of the dynamism of the same fields can be used for rapid deployment and penetration in dangerous and denser matter and mafs parts of the universe.

The same entanglement or interconnection process exists between light rays as does and prevails in the atomic and matter environment. At the present time the scientific world is considering ***quantum entanglement*** and its use in transportation of matters and other uses within and across the boundaries of the universe. In reality if scientists comprehend the concept of the light–entanglement and understands the principle use of this phenomenon, then one can use the matching structure plasmas of light rays before they become matter to achieve high speed transportation in the range of and beyond the speed of principal light rather than limited speed of the light in matter environment as man is considering to be the ultimate speed of motion at the present.

In entanglement process one can/and have the capability of use of different mafs of the lights for crossing the universe, and without change of physicality of the plasma and change of its state and then at this stage one can even copy and lock the same plasma of the light that in Magravs positioning of one plasma in keeping its connection with the same plasma of the light of the same strength or length in other parts of the universe.

If this concept of light interconnection and complex light-entanglement of Magravs of light balancing is understood and replicated then feeling, like touch, loving and hate and so forth can be transported across the universe without actually the matter or being ever crossing it physical boundaries. Thus coping atoms and interaction of the atoms and even the emotions of the matter (which complex, vibrational and rotational based) of one atom or light to another can be transported across the universe using light entanglement technology.

In fact this is how dreams in the inner sanctum of the human brains are created and take shape and give the man the image of seeing and communicating even with people which they have never physically met or to other Magravs of other beings which in their time of living in that give environment their Magravs of the dreamer and dreamed of have been exposed too.

The light entanglement is a real and correct physical process, where certain cells within the inner brain or thought control systems of the living being has the capability to operate within these pmf mafs strength and this is the natural part of the operation of the brain of the human beings that scientific world has not mastered to unravel.

In fact the light entanglement communication method called telepathy is the easier way of communication than using the ears or other method of listing and talking and the place for this system seats in the inner parts of the man's original main brain structure.

The light entanglement technology will allow collective entanglement and different strength molecular light entanglement allows the full structure transportation and even continuous motion replication of the molecules. This meaning the transportation through the light plasma entanglements of let say heart, brain and all tissues at the same time, with the same heart beat and the same liquid with emotion of the entity across vast points of the universe becomes a possibility of the use of this process. Where, the light in all its pmfs of all its mafs of all matters and strength are available in all corners of the universe and all needs to be done, is to re-produces the same strength and combination mafs of different matters to replicated and reproduce the copy of the original entangled light for the two lights entanglement across the universe to interact and communicate and transfer their knowledge without acutely the matters of the light ever leaving their original environment.

Using this technology one can even transmute the same from one mafs of the matters' strength to another to speed up or reduce the transportation and trans-crossing time from one field strength and matter mafs to matter strength variation and principal mafs to matter mafs and matters strength. This needing and leading to change of light's mafs to the appropriate measure and sizes and conversion from matter mafs to matter like the mafs of food or oxygen and so forth for human consumption in crossing the universe without relying on and physically carrying food or power supply or even transporting human beings across the universal physical boundaries.

This concept of transmutation which is commonly used in the universal method of transportation will be thought in detail in future disclosure and partly has been disclosed in the book the universal order of creation of Matters (1) papers references 11, 16, 21, 29, 34 and 37.

In the matter environment of the earth, this can be achieved in plasma reactors of simple structure using complex combination pmf of mafs of the light and matters and plasmas in the dynamic nuclear supported reactors.

This concept of light entanglement is one of the easy solutions for transportation in the future without physically moving matter across the universe. This concept has been tested in medical application of the technology and results confirm the correctness of the technology. Where, copy of the correct cell information is translated to cells that the cell makes the copy of the right cell and uses this knowledge to cure or replace damaged cells or by allowing the process to be moved to another part of the body while one part of the brain recovers from damage (the coma case).

The concept of the atomic-entanglement and the binding between two atoms is a natural process of interaction in the universe and if one understands the similarity of this interaction and the light–entanglement then one can see that in the physical world twin stars and the atomic entanglements are of the same in principle interaction and in concept, with difference in their magnitude and the size of the entities involved. Taking note that there are differences in between the twin and binary connection of atoms and stars and their process of entanglements criterion are totally different. Where, the twin atoms and stars are matching in matter contest and, where the binary stars and atoms are matching in magnetic and field contents.

Where, in the binary system the overall Magravs is matching that entities can magnetically polarity interlock. Whereas with twin entities every matter matches in mass and position as well as Magravs and their pole positioning.

The reason that a pair of atoms in receiving additional pmfs of rays of light is elevated to create a mysterious binding between them, is due to one of the fundamental principles of the reversibility in the magnetic polarity of one atom in respect to the other, even though both atoms having the same directional motional rotation. That is to say, *one of the plasmas involved in the entanglement get its magnetic polarities reversed in respect to the other atom*. This meaning, if the first atoms' proton and electron have their North-Pole on the top, in receiving additional mafs of the light, then the second atom receives enough mafs that causes or initiates the change polarity in this second atom, that at this point the North and South poles of each atom interlock with each other and the sudden atomic binding due to the lock of magnetic fields flow between the atoms takes place.

The same principle of atomic entanglement and attraction of reverse pole attraction holds true, when one observes the rotation of twin and binary stars entanglement. Whereas one star reverses its magnetic polarity, automatically the twin star polarity changes too, that the magnetic lock between the two stars to be maintained that the entanglement and twining and binary system can be sustained.

Star, atomic and plasma twining and binary process locking is the safest and common occurrence in the universe as this gives the stars a total magnetic balancing and guaranty's the survival of the both entities. This phenomenon of reverse polarity entanglement holds true and exists for all the twin system in the universe, this being for atoms, stars, galaxies and even universes.

The reason that the light injection or absorption of additional energy by two atoms in some cases leads to creation of entanglement is primarily due to the fact that additional energy received from the ray of the light leads to increase in the magnetosphere dimension of the atoms and this in turn allows two atoms in proximity of each other two magnetically interact and entangle and at the same time in coming close to each other, one atom changes polarity and the two atoms pmf lock and gets gravitated to each other but at the same time each atom is holding on to their Magravs positioning strength gap which is dictated by their internal pmf maf contents. Once the mafs of one star and atom changes due to whatever reason, then one will observe the disconnection of the entanglement or the coupling between the two entities, where at this point the higher Magravs strength partner will pull the weaker Magravs entity to itself.

This is a crucial phenomenon to understand, especially when using light entanglement reactor process in reaching any destination and requiring the change of the plasma of the light from principal mafs to matter mafs strength condition. Thus atom entanglement is a natural phenomenon and at the same time once an atom is created the content of two atoms' plasmas are hardly ever the same, this is to say, to find matching content of pmf of two plasmas of electrons and protons are as rare as finding two completely matching structure of two snow flacks.

The injection of the light into the plasma of two atoms allows some of the atoms, which are near enough in pmf mass to reach the same point of mass or Magravs strength equality at the same time and hence have the capability of coupling. But there are points that additional pmfs of the light can induce enough energy that one atom can come near to another atom pmf balance and being able to interlock too.

The atomic–entanglement is not and cannot be used for transmutation of elements as easily as has been suggested by some scientifically advanced schools of thoughts.

In the present scientific world, scientists have been observing and detecting regular pulses of for example gamma radiation in the upper atmosphere of the earth, which they consider that these radiations all originate from the cosmos and for example they are from the time when a star explodes through the nova principle in the universe. These types of radiations are not the only types of radiations that arrive towards the inner part of the solar system from the cosmos. In reality there is spectrum of radiations and magnetic fluxes of different strength that go through the solar which are originated in the cosmos. In comparison, the radiation that arrives in the inner-most part of the solar system and at the boundaries of the earth atmosphere, these cosmic origin based rays are nothing compared to the gamma rays and other rays which are created within the heliosphere of the solar system itself.

The truth about the main source of these background radiations which has been considered to be from the cosmos and are released, when stars in the universe go through supernova, this concept that these explosions are the main source of these gamma rays and other background radiation which are detected, this shows the lack of the understanding of the process of the conversion of the light into matter and movement and collision of these new matters and mafs in the Oort clouds and Kuiper belt of the solar system. The main source of the gamma radiation bursts detected are originated and created through the motion of the charged particles of the new matters mafs and matters and their collision with each other in the Oort clouds and Kuiper belt region of the solar system and these gamma bursts have nothing to do with supernova and release of gamma bursts that they produce.

That is to say as/and, when newly born charged particles of mafs of the matter and matter as gases, nano matters and ganses move in outer regions of the solar system and ***as they collide with each other***, they release magnetic fields in the gamma, other radiations and noises in magnetic field spectrum levels in these regions of the solar system, which these radiations find their way into the inner parts of the solar system due to the strong magnetic fields interactions pull of their star.

The way these radiations and noises are generated and released is exactly the same way as the lighting and thunder are created and how light and sounds are released through the collision of clouds on earth and in other active atmospheric conditions or in other planets and moons in the solar system and similarly as it happens on the surface of the Sun, when matters mafs of the Sun collided in its upper atmosphere.

Thus the space travellers of the future have to consider all components of the all rays of the magnetic spectrum in the matters' mafs, matters, light, sounds, the principal mafs and matter and its sound and all their light radiation spectrum strength levels too, which these are all free and in motion in the space of this solar system and in the cosmos.

The radiation danger and damages from the light in the principal mafs strength levels are more predominant, stronger and profound than the matter mafs light ray spectrum strength and these radiations and their motion in space will play a crucial role in the future of space travel as these rays will be lethal and fatal if protection from these are not facilitated into cosmos excursion systems and scientific world has to have a viable protection technology for the passengers in the space crafts of the future before they contemplate on such excursions into cosmos and specially traveling in the centre of the galaxies and near principal stars.

Whereas, for example one of these problems will occur, when strong mafs of the principal matters radiations will interfere with matter mafs dilution Margraves positioning reactors used for motion of the craft in space and these principal mafs upon them reaching inside the reactor, they can create Magnetic field surges that can lead to surges in these reactors gravitational and Magnetic fields, which this can lead to the loss of the system and the total inhalation of the system, or the change in the Magravs strength leans the craft in totally different part of cosmos than where the craft was originally destined for.

Therefore, before the men of science contemplate on traveling beyond the boundaries of this solar system, they have to master the knowledge of the mafs of the principal matters spectrum's magnetic field strength and the light transportation mechanism in this spectrum.

If one could understand the structure of the light and how it changes from spherical shape plasma to cylindrical in shape for it to be able to travel with less friction and high speed from one dynamic environment of spherical or cylindrical entity to another spherical environment, then in the future the man of knowledge can develop elongated cylindrical three matters mafs strength spherical and cylindrical composite mafs fuelled combination setting and reactors and crafts, which these elongated crafts will have the benefit of possessing spherical or cylindrical plasmatic magnetic shielding of higher order magnitude pmf strength of the principal mafs and matters for fast and protected space travel in the universe, then at this point one can travel the spans of the universe in times which are not even imaginable at the present by the man of science by use of these types of composite setting and fuelling reactors.

The use of light for purpose of transportation technology and communication through the understanding of the relationship and interaction of mafs of plasma and plasmas of the light(s) in each and every mafs strength of the matters of the light structure is the key to final and ultimate intellectual growth of the human race and point of maturity of the coming of the age of the man as has been promised. This technology will be at the edge of man's enlightenment in the knowledge of creation and control of light and its mafs and matters in the universe.

Chapter 8

Insight into the future use of the structure of the light

From the new understanding of the structure of the light, can easily and readily be concluded that if the light did not and was not made of the structure of three mafs of the matters then the light would have not possessed its own gravitational and Magnetic field forces and for example the light from other stars in the universe would have never been able to reach the earth as an individual magnetic field due to the characteristic and properties of the magnetic fields interactions and gravitations, as through magnetic interactions all these rays' of lights would have been absorbed by other magnetic fields and gravitational fields in the universe and the light would have never reached the far distances of the universe. Hence the light as a physical entity which is similar to other physical entities the like of proton, electron, earth and the Sun and the light being in possession of both Magnetic and gravitational fields, the light ray has to find its position in respect to any other object and pmf in the universe as it passes these entities in the universe, and hence in its travels in cosmos light ray has to find it's Magravs positioning in respect to them that this allows the light ray not to be absorbed but to go around these objects. Hence, the reason for the physical observation of the lensing or the bending of the light near these object as the light passing large and strong gravitational object is purely due to its Magravs and its positioning in respect to these entities in the universe.

Therefore the lensing of the light is the confirmation of its existence as dynamic composite magnetic field strength fields physical and tangible entity in position of Magravs, and the light is not a ray of magnetic field in motion or what has been called that the light to be an energy. Whereas, the plasma of the light ray in motion can transfer it Magravs to other matters and thus has the ability to transfer its momentum to other matters and mafs as energy.

Hence due to the fact that the light possess both Magnetic fields and gravitational fields and in line with laws of physics, when the gravity and the Magnetic fields interact (1 and 2), their interaction lead to the creation of entity in possession of its own mass and its own magnetosphere. Where, the interaction of the lights magnetosphere with the magnetosphere of other objects due to their friction leading to release of residual magnetic fields in the matter magnetic field strength level of the human protein, which this makes the point of interaction of two Magravs detectable pmf field or what is called the visible light to human eye and his tools and methods of detection of pmfs in this ranges of filed strength, out of the full magnetic fields spectrum of the universe.

Where, the motion of the dynamic spherical plasma of magnetic field in its travel from one point in the universe to another causes and leads to elongation of the dynamic spherical pmf, hence the restructuring into a dynamic cylindrical plasma ray of the light, and as this plasma of the ray of light reaches the right pmf environment strength that slows down enough or losses its strength, at this point the cylindrical plasma restructures back to its spherical structure and leads to the creation of pmf of matter mafs and matter in given matter pmf strength environment (Fig. 48) The book cover image. Hence light has the same structure as the pmf of any other plasma known to man like the plasma of electron and proton but in elongated form rather than presently known dynamic spherical shape of tangible matters.

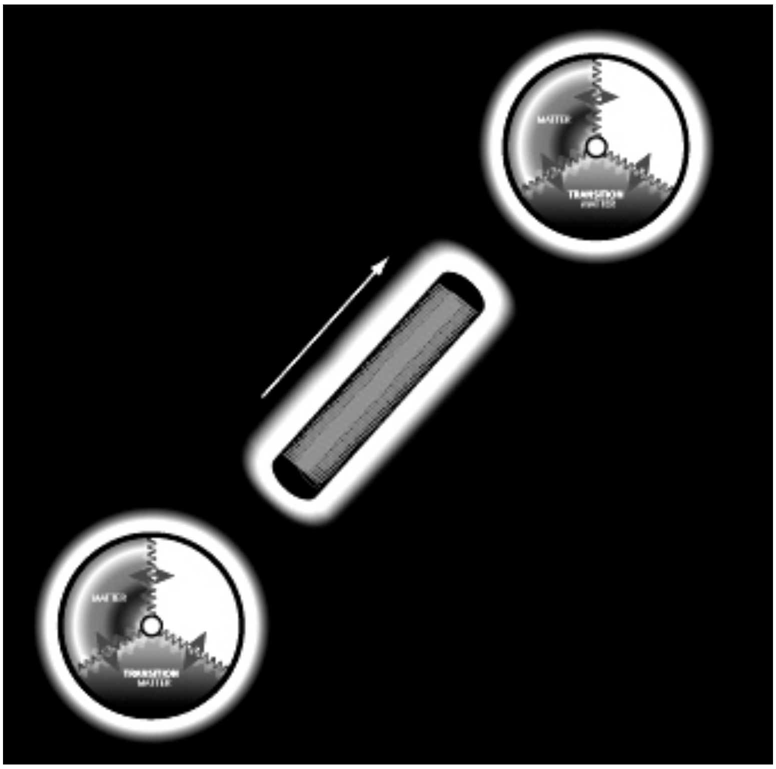

Fig. 48 The conversation of dynamic spherical pmf into dynamic helixial cylindrical ray of light and its conversation back to dynamic spherical (Cover image).

Where, at the same time the plasma of the light in its travels in space can become visible according to the plasmatic magnetic fields of the environment which it travels through and according to the observers' plasmatic magnetic fields strength which in the case of human living Magravs protein molecules structure.

Different proteins have different pmf strength for different creatures in the span of the universe, hence what is visible light to one creature in one point pmf strength in one part of universe this does not mean that to another protein Magravs strength in the same environment this can be visible light too. Hence what is visible light in principal matter or mafs strength, in transition mafs and matter and in matter mafs world in a given matters' magnetic fields environment respectively these light rays are detectible or visible according to the environment and of each mafs and matter in the binding Magravs of the structure of the observer.

The plasma of the light has the capability to present the magnetosphere around itself according to the pmf strength which it is passing through, thus it is the environmental pmf strength which decides the outcome of the mafs of the matter of external layer of the plasma of the light and the final presentation of the light in a given environment.

The plasma of the light to reduces its losses in its pmf of its matters and at the same time allowing it to encounter less resistance for it to be able to travels through the same pmf strength that have less or faces no losses, thus the light ray has the capability of the change of cover mafs and matters of its structure. Thus it is inevitable to consider and conclude that the light like its larger combinations has its own intelligence and its own internal logical magnetic processing that allows it to survive and exist in every dimension in the universe that it might travel through or be part of. This giving the light the capability of being able to travel long distances, and at the same time, this allowing the light not to have fixed maximum speed in all universal magnetic field strength spectrum opposite to what has been assumed by scientific world at the present.

At the same time, the light becomes more visible or detectable as it travels from stronger matter mafs fields strength to a lower matter fields strength environment.

Two rays of lights with different mafs of the principal matter strength can pass through the same medium and they can create different brightness intensity and therefore all lights rays do not have the same intensity primarily and their strength and intensity is internal initial principal mass and strength dependent. Therefore a ray of light can be dimmer or slower compared to another ray primarily due to their mafs of the principal matters content and their strength.

Lights have speed according to their lowest common denominator of it environment pmf strength. Therefore, speed of light in Principal environment due to the stronger field strength of the principal mafs pmf is much faster than the speed of the same plasma in the matter environment, hence the present assumption that the speed of the light as observed by man is the maximum speed of the plasma of the light is ***incorrect***, and hence the plasma of the light has speeds in the Principal matter environment far beyond the present assumptions in the matter environment of the earth and this universes matters' pmfs strength.

Even though the formulas used for conversion of main connection of mass and energy looks as to be balancing and to be correct, in fact this equation is only correct for small and minute amount of plasma and only in the matter environment and not in other plasmatic fields environment of principal mafs and mafs of the transition matter or combination and composite plasmatic magnetic fields strength of principal mafs and matters environments and where only individual rays are considered.

Now that one knows that mass is created through secondary effects and interaction of the effects of the interaction and connection of at least two dynamic plasmatic magnetic fields of the gravity and Magnetic fields (1), which these Gravitational and Magnetic fields both themselves are created and are products of interaction of at least two initial magnetic fields, then in fact mass itself is product of interaction of four fields of two initial fields and two effects of these two fields and hence the formula of relativity now is totally incorrect and has to be rewritten as now one of the factor of the four of the magnetic fields and their combination of their products with each field in position of its own speed of its mafs strength, as energy which is define as magnetic fields in motion, then multiple of speeds of the light in their initial matters mafs strength times the matters mafs of their effect strength in a given mafs strength is the realistic method of the measuring the energy in respect to the mass of the same matter.

Using this knowledge and understanding the structure of the light and applying these in the space technology and in the space reactors crafts of the future for travel in deep space, then by knowing the mass and the strength of the fields which one needs to be traveling to and by the same principle the strength of the original mafs of pmf can be calculated and measured by using the fundamental application and implication in production of light Magravs strength of the system and of the same of the destination point, then one can estimate the power of the crafts Magravs need to be and at the same time what the speed of the travel of the craft shall have, and using the mafs of the light at that given matter environment, one can calculate what mass of the light the system needs to have and be loaded that the system can travel at a given speed in a given environment, or what speed the craft might have once one uses the certain amount of mass of the light in any combination in any of multiple combined fuelled integrated multiple Magravs positioning reactors of the space crafts of the future (1, reference 24,18 and 11).

In understanding the strength of the principal mafs fields strength and using this phenomenon for creation of speed of motion of the systems and crafts, then one can calculate and understand that travels with speeds hundreds and thousands of times and millions of time faster than the speed of the light in matter environment in the near future will become as conventional as jet liner travels of today. The assumption that the speed of light in the matter world is the ultimate speed in the universe cannot be correct. This assumption, when using the principal mafs fields as source of fuel in dynamic plasma dilution gravitational positioning space reactors, which have been developed and tested can be used to achieve higher speed than the speed of light in the matter environment.

Through this technology of the use of light gravitational positioning systems, speeds of travel in Principal and Transition mafs strength environments, which are beyond the present understanding of the man in the matter universe are faster than the speed of light in matter environment.

Now one can understand, that once by not using the principal matters as fuel as has been discussed and promoted by some organisation to burn, as has been the habit of the man to burn to attain motion and height in the past centuries, then in using the mafs of the principal matters in creating gravitational fields and Magnetic fields within a gravitational system, then one can travel the length of the universe in times which are not acceptable to man at the present time due to restrictions which has been forced on the humanity by the less knowledgeable scientists of the time and due to their lack of understanding of the creation and control and structure of the light itself.

Then now one has to make new concepts and measuring tools for local time and universal times as these will be different according to the light mafs strength of the space crafts and environment of measurement and new conventional translation for time has to be employed, otherwise traveling in one speed of the light mafs strength and then entering another to live in, this will create untold problems for the future universal travellers of this universe and for the inter-universes travellers of the future (3, 6 and 7). Invariably in the near future one can use the principal mafs and matters magnetic fields for the purpose of transportation in the universe, where from this point on travel with speeds beyond the matter pmf strength environment for travel in the universe and between universes become a natural means of transportation.

At the same time the speed of light in the principal mafs and matters strength is visible light to the man too, with the difference that in this environment as the protein chain of the man contains and is consisted of plasma of electrons and protons, and these plasmas themselves are made of and operating and consisted of the composition of principal mafs and transition mafs and matters (1) too, then the principle mafs strength of the protein will come into operation and the life will be as normal as one living in the present matter mafs and matter environment of the universe. With a slight difference that he needs to learn about control of the operation of his physical body and his tools of detection in the speed and light configuration.

The parallel world of different pmf strength of matter mafs and mafs of the principal matters are the faster and stronger in pmf strength of life which the Man has come to comprehend and live with and soon he will be able to transmute from one pmf strength of mafs to another and use this knowledge for the advancement of his race and connection and communication in and with the other creature of the universes.

The gravitational field forces of the mafs of the matters of the structure of the light will be the smartest and fastest method for creation and control of space reactors of the future. At the same time speed of the separation of mafs of the matters of the light will be the best directional controller and magnetic shielding protection system for the space crafts of the future as the lights' mafs and matters separation and control in different mafs of the matters and matters of the universe this make the plasma of the light the safest and quickest methods to generate parallel speed of the light in different mafs of the matters environment and at the same time the directional motion of the Magravs transportation system can utilize the mafs of the matters of the light interaction for smooth transition from one matter mafs field strength to another or from matter mafs to matter environment and vice-versa.

This light matter mafs structure division and utilization will be the technology which needs a full understanding of the world of creation and once mastered as has been done by other inhabitants of the universe, then the man will be of the same strength and capability of the understanding of the universe of universes operation, interaction and universal organisation.

When, the intelligence of the man reaches the full understanding of the potentials of the use of the light mafs of the matters, then one will be able to travel in different mafs of the matters environment with use of the mafs of the principal matters magnetic fields strength as the protection magnetosphere created around the crafts in possession of the multiple Magravs positioning reactors.

The dynamic twin reversal of directional motional shielding system made of and through the use of mafs of the principal magnetic fields' strength of the light guaranties the total non-penetrable protection for travel in all mafs and matters environment in the universe, without leaving any traces of direction of motion and without being visible to the human eyes and any other tools and means of detection of any other creatures in the universe.

Where, this technology can be used for fast speed travel in crossing boundaries of universes and for travel across one universe without spending thousands of light years in traveling time as is currently calculated by human scientists. Thus one can travel with full safety using the highest lights Principal Magnetic field strength for every aspects of the motion in the universe.

Once the concept and the principle structure of the principal mafs and matters and other mafs of matters and lights are fully understood then one can use the knowledge of the light structure in full and the scientific world will be able to use the structure and mafs of the light for creating new matters and production of food and nourishment for the life of the passengers and for the space machineries of the future and food producing reactors for the man in deep space where the light is the only available matter for use and at his disposal.

The mafs of the matters of the light can be brought to levels that one needs for the body to absorb for it to survive a pleasurable and comfortable life. Where, different strength mafs of the matters and of the light can be used as vitamins and minerals and proteins for the Man's needs to live. Where, the light mafs of the matters can be used directly without the need for the plant and animal conversion systems of the light mafs of the matters to minerals and nourishment for the human consumption.

One has to understand that there are lives in the lower pmf strength of the universe too; where in reality the matter universe is not the lowest level pmf strength in the universe. The light in this universe is the same and operates in the same way.

In the future one can *use the light mafs of the matters to produce and lower their strength mafs of the matters to pmf for example of the sodium or vitamin A or B*. Therefore from this point on one can even produce composition of matters or what has been the dream of producing energy and new matters through fusion. Now one can use the Principle matter mafs field's strength of the light to fuse the elements in the matter environment. In fact fusion is more practical and achievable if one uses the principal mafs and matter's field's strength of light and their division to create environment of matter mafs for the matters to be created. Fusion of pmf of the light mafs is in so many ways easier than physical matter fusion once the light is transferred into neutron or an atom. The lights' mafs of the matters fusion is a way for the re-production of heavier elements or their pmfs without creating enormous amount of heat that one cannot and has no tools and means to control and contain such matter fusion technology.

Once the man learns the use of the construction of mafs of the matters of the light, then one will learn that source of mafs of the matters are available in all corners of universe and on the other hand one does not need to create the physical matter to have the properties of the physical matter, but one can use pmf strength properties of the given matter in its fusion of the light mafs of the matters to achieve the same effect, the likes of the production of energy and food without ever using fusion of physical matter to achieve the same goal and actually without ever needing physically to eat and to digest food. Then in the near future man's body will do with the digestion system and in the further times man will learn and will do with the physical body for it to exist and for it to travel the spans of the universe(s).

Using the mafs of the matters of the structure of the light rays for **communication purposes**, in the matter environment, then the communication system can be set to transfer the information in the Principal matter mafs strength and once the information arrives at the point of destination the receiving system can decode the information from the principal mafs strength to matter mafs strength and then the information can become available for whatever purpose it was transmitted for, this being audible, visual or computable information or transfer of matters. This method of use of mafs of the matters of the light field's strength will eliminate any time delay in communication across the spans of the universe, the delay time which has been observed in the present cosmos communication systems will not exist once using principal mafs strength and then communication become instantaneous anywhere in the universe.

Thus by switching from one matter mafs strength to another matter mafs strength of the same light, this will bring into play a totally new technology and facilitates new approaches in the production of tangible matters too. That is to say one can strip the light from one of its mafs of the matters for one purpose and use its residual mafs of the matters for other purposes. That is to say, by striping the light off its principal mafs for the communication purposes, then one can use the mafs of its transition matters and mafs of the matters of the same plasma of the light for production of increase in other matters and matters growth or expansion in their given environment or use these for production of energy for human existence in space.

The universal language using mafs of the matters of light and using the same mafs of the matters of the light for their transportation will be the cornerstone of the future transmutation of matter and matters and communication in the universe among other intelligent creatures of the universe(s).

The positrons are created at the point of the creation of the plasma of the atoms from the ray of the light and the clockwise motion of these particles is totally and primarily internal magnetic fields position dependent. This meaning that the directional rotational motion of all matters are dependent on the speed of the rotation and position of the mafs of the principal matters and matters in respect to each other, this plasma be it of the plasma of electron, proton, atoms even stars and galaxies.

The clockwise or anticlockwise rotation of electrons, protons, planets and stars is determined by the principal mafs matters speed of rotation and its distance in respect to the matter mafs and distance between different layers mafs components of the plasma of the entity after the light ray is converted to matter of the electron or the proton and star and its matter content.

In conclusion one has to be aware that visible light magnetic fields spectrum, which makes man to see its universe is the same in all pmf strength levels. This means that in every pmf strength environments there are visibility criterions to all created life, but the visibility is pmfs field strength dependent. Therefore, what is not visible to one being is visible environment to another using different pmf strength for point reference as tool for vision and this applies to all pmf strength mafs of the matters and matter universes and environment.

At the same time vision is not through the vision tools like the eyes, but vision is the openness to receive knowledge through the real means of the operation of the universe in the pmf strength of the level of intelligence of the structure of the beings.

The reality of the existence has been measured by the strength of the light, which a matter can emit, and the duration and the depth the Magravs of the created light can penetrate in its environment. Then the question for the Man to ask of his existence, is that how does the matter like him, who is the user and the producer of the same physical entities as stars, how does his existence will shine and operate in the span of the universe after the demise of his physical life on this matter strength environment?

May be it is the time for the Man to come to understand the true meaning of the light, as all beings in the universe announce their existence through the emission of the light into and onto their surroundings.

This technology of the light ray's protection gives the man the ultimate knowledge for survival and brings the man knowledge and intelligence in par with and equal to the knowledge known to the rest of the universal manifestations and the same we have served our purpose in bringing the same to all men of progress.

Thus would be correct for the Man to consider his connection beyond his physical existence and through the light of his presence during its time, and hence would the soul of the man be of the light of the confirmation of his existence.

Be wise to the truth of the creation of the man, if the light of wisdom is the light of the soul from the time of beginning to the end of time. This light will exist and will be the source of the confirmation of everlasting existence of his soul, as the light of far distance star after billions of years are shining in the deepest part of the universe, this has been and shall be destined for the soul of the Man throughout all times of the existence of the universe.

Thus it can be said

The man has come from,
He exists, operates through and
He shall return to
the Light

Definitions

Below is a list of the new terms and their definition that used in this book. A number of these words and terms have been used in the previous book and scientific papers, which have been published by the foundation and most readers, are familiar with these terms and reference. In this book a different approach and language and terminologies is used, therefore has become necessary to create a new terminology and words that can explain the real term application and processes in the world of physics.

Mafs: magnetic fields

Magnetic: with **capital M** refers to the plasma of magnetic fields created by a dynamic entity as for example this refers to the magnetic field of the earth.

Magravs: Magnetic and gravitational fields (Magnetic referring to the plasma of magnetic fields created by a dynamic entity as for example this refers to the magnetic field of the earth)

Pmf: plasmatic magnetic fields

Principal matter: This matter in conventional physics is called the ***anti-matter***, as now the real structure of this matter is fully understood and is said to be source of all matters from the dark matter to matter, thus it has been called the original matter or the principal.

Principal mafs: These are magnetic fields, which are in the principal strength levels and considered to be I order magnitude several hundreds of times stronger than the magnetic fields which make up the tangible and visible matters.

Principal star: This entity in the conventional physics is called the Black hole

Transition energy: This energy in conventional physics is called the dark energy. This is the pmfs that are in motion from principal to matter magnetic fields strength so called the transition energy as energy is defined as the pmfs in motion.

Transition matter: This matter in conventional physics is called the Dark matter, as now the real structure of this matter is fully understood and is said to be when plasmatic magnetic fields are in transition state from the principal matter to tangible matter thus the term transition matter is used.

Transition mafs: These are magnetic fields which are in strength level between the principal and matter magnetic fields strength, and are in transit from the principal mafs to matter mafs strength levels.

Gans: a new state of matter; a molecule of a gas (GAs) which becomes Nano of itself and appears as Solid state of matter. **GA**s to **N**ano of **S**olid and to be called **GANS** for short.

Reference

1. The book, “The universal order of creation of the Matters”, M T Keshe, published July 2009 ISBN 978 46 6087 001 9

2. The unifying theory of the fields M T Keshe 2009

3. The origin of the Universe M T Keshe 2010

4. The creation of the solar systems M T Keshe 2010

5. The CO2 capture paper by MT Keshe Published 23.12.2009

6. “The universal order of existence” finalized and to be published in 2010-2011 by M T Keshe

7. The Magravs positioning M T Keshe 2010

8. NASA WMAP common.wikimedia.com

Published and unpublished works by M. T. Keshe

There are a number of unfinished articles that are to be converted to full disclosures when completed, like antimatter mafs method of energy releases, Nano space technology, relation between the creator and the created, space communication system, deep space defence technology, dynamic magnetic fields force regions (wormholes) construction and control, creation of magnetic fields, and so on.

Some of these works have already been published through different channels and some have not yet been released.

In the following section some of the works done by the same author are mentioned, with a brief disclosure of their contents.

Some of these papers can be purchased and obtained through direct request and access through specific links.

A number of papers with * and ** marks **are not for public release** due to their sensitivity of information and only available to relevant commercial and governmental organisations for peaceful applications of the technology.

The universal order of creation
(5.7.2006)

In this disclosure the main construction principles and the method of the link and control in the universal order of the existence of living cells, like the function and operational method of any matter based on so called proteins chains, as for example in human cell are explained in detail.

REF 9

The author considers this paper as one of his most important and fundamental works to be published and to be added to the sets of disclosures written by him.

Cosmic Rays
(Release date 24.3.2004)
This paper explains the source and the origin of cosmic rays and the way these work within our cosmos, their function and their benefits.

REF 10

----

*The Airborne reactor

(Release date 25.3.2004)

This paper covers the total design of airborne reactors for space technology. This covers all aspects of operation, design and control of such reactors.

REF 11

The seed of the Earth
(Release date 25.3.2004)

Some extracts from this paper:
The theory that the heat of the core of the Earth is due to inertia of the material piled on top of each other is completely dismissed through this new disclosure.

The author M T Keshe believes that this centre core is made of hydrogen, other gases and a mixture of liquid and solid matters mafs, which were in the vicinity of the centre of the core of the Earth at the time of its inception in the solar system.

Through this new understanding of the new inner cores of the planet, this brings to play the fact that the old inner core of the Earth that is already assumed to be made of a solid piece itself becomes a sealed container for the new inner core.

The author through design of a parallel system has set to confirm the validity of this assumption, that the centre of the Earth possesses a semi- fusion atomic reactor and not a fission reactor as has been assumed by other scientists.

This theory of hydrogen in the centre of the planet is in line with the physical reality of the creation and the present situation in the physical construction of the solar system.

By the laws of physics all the gases in the solar system should have been in the giant gas planets at the outer layers of the solar system, like in Jupiter and Saturn.

The physical reality in the solar system, is that the lightest of all these gasses, the Hydrogen, has taken its place right in the centre of the solar system in its star, the Sun.

It has to be recognised that the centre of all planets in possession of gravity due to the centre cores heating, are always in possession of two Magnetic Field Forces in their inner sanctum (M. T. Keshe “The creation of gravity").

In the case of the Earth up to now and with the present knowledge there has been acceptance of only one inner core and one magnetic force.

As it is explained in the paper titled "the creation of gravity", there is an inner core in the inner core. This inner core, due to it physical content, motion and position in the centre of the Earth, creates and maintains its own magnetic field force, independent of the magnetic force which has been known to be created by the interaction of the inner core and the outer core of the planet.

These two magnetic fields interact to create the planet's own double magnetic fields in its centre. The interaction of these two magnetic fields forces upon each other leads to the new concept of the "DOUBLE MAGNETIC FIELD EFFECT", where the interaction between these two field forces in the centre of the planet leads to the creation of the gravity and the magnetic forces of the planet

(Paper: The creation of gravity).

This new second core is the seed core of the planet, the mother seed of the planet positioned in this new inner core.

REF 12

The creation of Black Hole
(Release date 25.3.2004)

In this paper it is explained that the Black Hole is created out of a natural sequence of events, and that the existence of the Black Hole in a galaxy is a normal event. Further it is explained that a Black hole is created in similar ways to the Dark Spots on the surface of the Sun. In this paper for the first time, it is explained how the Black Hole is physically created and what function it plays in its galaxy.

REF 13

Magnetosphere
(Release date 9.6.2004)

"The shape and the strength of the magnetosphere of a planet is exactly what a fingerprint is to a man. It is unique to each planet, star and galaxy. It is a tell-tale of all its attributes, and it reveals the hidden mysteries of the internal materials' structure".

REF 14

Correction to the Einstein equation of Relativity
(Release date 15.6.2004)

In this paper the physical reality of the Einstein equation is considered, and efforts are made to bring this equation to the real conditions. Even Einstein considers this equation applicable for very small mass, and not in real three-dimensional multi-Magnetic and gravitational conditions, which are external and have nothing to do with the theoretical object's mass and speed, but have effect on the speed and the mass of the object under consideration.

REF 15
----

*Introduction to the new system
(Release date 2.7.2004)

In this paper it is written "It is essential to understand that, with the application of this system of motion, one cannot bring into play the solid, liquid and gas fuel method to create motion, as is used today. What is offered in this system is in simple words a fully integrated plasmatic magnetic energy system. Which, creates within and around the core environmental conditions as known to man in the universe like creation of gravity and magnetic field protection and the same as the magnetospheric protection condition of planets and motion through interaction of magnetic fields as planets and stars do in their given environment".

REF 16

The rings of Saturn
(Release date 2.7.2004)

One of the oldest puzzles in the world of astronomy has been the existence of the rings around Saturn. How did these rings come to exist the way they are now? How did they come to appear and behave the way they do? These questions and much more about the creation of these rings are all answered in this paper.

REF 17

The creation of gravity
(Release date 11.7.2004)

In this paper it is explained that "the source and the creator of the magnetic forces and the gravitational forces are due to the interaction between the same materials in the same region of the planet."

Further it is explained how the gravitational field force can be created, replicated and controlled within a nuclear reactor. Further it is explained how this property can be used for the motion of a system within a planetary or solar system.

Now by understanding the principle of the creation of gravity, in this paper it is further explained that "the gravity in reality is the effect of the interaction of two plasmatic magnetic energy fields of any two object in respect to each other."

REF 18

Death of a star
(Release date 26.7.2004)

In this paper the death of stars is explained, "The death of a star and the creation of a supernova is in reality the same as the half-life cycle energy step-down of an atom, with the difference that this show is in a larger scale, and more spectacular, in all aspects of its dimensions".

REF 19

Fusion
(Release date 28.7.2004)

In this paper the physical and scientific problems with present fusion technology is considered. Where, it is written, "nuclear Scientists, with the fusion reactors of today, are trying to create fusion in the opposite way to the known laws of physics and their parallels in the universe". Further in this paper it is written, "If the present path for creation of fusion is followed, with the present scientific knowledge of physical materials for construction of such system, then following the present path of development to create energy from current fusion systems will be a distant dream".

REF 20

Life of a Cell
(Release date 28.11.2004)

Extracts from this paper; "The reality of the transportation and transmutation of a cell is not far from the reality of the life of an atom or a star. Where the life cycle of the cells are slightly more complicated, as they contain other matters mafs like acids, which have their own magnetic characteristic of their own chemical structure. For this reason the control and replication of the energy of the cell is very much complex, but very simple to achieve".

REF 21

The Atom
(Release date 19.12.2004)

The content of this paper explains how and where atoms in the universe are originated from and how the construction of atoms in a cold plasmatic nuclear reactor can be replicated.

REF 22

Magnetism
(Release date 8.1.2005)

In this paper it is written, "The question is where did the Magnetism come from? This is the Secret of creation. The world is made of one thing and only one thing and that is Magnetism. Its combination and interaction with its own different strength fields makes it appear as different matter mafs. The rest of creations are the outcome of these interactions of different strength of this thing called magnetism and the forces it creates. Magnetism is the origin of existence and the real singularity in its full meaning".

REF 23

----

*The creation of the magneto-gravitational force
(Release date 5.2.2005)

The creation of magneto-gravitational field force is and will be fundamentally independent of temperature and pressure, proviso to the fact that as long as a homogenous free floating PMF within a matter mafs is created.

REF 24

The core of the Earth
(Release date 20.4.2005)

One of the original assumptions in developing the new Magravs system some thirty years ago by M.T. Keshe was on the basis that the centre of the Earth possesses addition cores, which create the necessary gravitational conditions of the planet. After some twenty years of space technology and through Earthquake seismological data's this assumption has been shown to be correct and after over twenty years, it has been scientifically proven that the inner core of the Earth has an inner core of 600 km. Professor Guy Master in 2002 ref 24 nuclear planet. The centre of this core is assumed to have an inner core of 8 Km, this he thought to be made of the plutonium or other nuclear materials

But the developer of this new plasma reactor technology has proven through physical dynamic cores that these cores possess a mixture of hydrogen plasmas and the process of the heating of the centre of the Earth are due to a semi-fusion chain of events.

REF 25

----

*The electron-atomic welding
(Release date 3.5.2005)

Some extracts from this paper:

"The electron atomic welding or atomic welding principle is the phenomenon that atoms of the same material, through a common shared electron, become a magnetically balanced molecule of the same matter mafs, with the difference that the shared electron will create a balanced PMF in the molecule".

REF 26

----

*Shutdown and safety of the reactor core
(Release date 6.6.2005)

In this paper the safety parameters in the operation of the gravitational and energy reactors are simply explained to avoid the loss of plasma and gravitational field forces created within its cores.

REF 27

----

*Reactor Start-up

(Release date 9.6.2005)

In this paper the systematic start-up of the dynamic reactors for the creation of gravitational field forces and the production of energy are set out. The start-up of these systems is totally different than the prior arts known in the nuclear industry.

REF 28

----

*The energy balance of the reactor

(Release date 29.6.2005)

Some extracts from this paper:

The energy balance of these types of reactors is not as simple as might look in the first instance. The energy creation is simple. But the heat leakage and dissipation through the deliberate inherent material design are and can be zero. That is to say the system not only can hold on the heat it creates. At the same time due to the close circuit feed operation will become self-sufficient but long lasting.

In planets, the loss of heat through their surface creates the variation in the central core condition over several billions of years. In this reactor as losses can be literally negligible, the system can operate at low energy loss temperatures. Even the boundary of the body could be made through the right combination of material in the chambers, so that the body of the reactor will be colder than its surrounding boundary. So the system will not only lose energy to its surrounding, but it can be made to absorb heat in the negative temperature gradient from its surrounding, so that there are no losses.

REF 29

Nuclear decay or half-life

(Release date 20.7.2005)

Extracts from the paper:

"The nuclear decay always has the same or near the same time duration. This is due to the fact that the all neutrons and protons, from their inception, carry a predetermined level of plasmatic magnetic energy. This energy is continually used for the motion and vibration of different elements of the nucleus in respect to each other and the surrounding matters mafs of the atom. As the space and dimensions of a nucleus of an atom always obeys the same principles of the magnetic separation and magnetic attraction. Protons or neutrons can and will need to use or consume the same energy before they become weak enough that the splitting or decay becomes a clockwork job. The nuclear decay is the natural levelling done of a nucleus' energy, due to the energy consumption and heat dissipation for the motion of the elements of the nucleus for its remaining plasmatic magnetic fields energy necessarily needed for them to hold one together".

REF 30

----

*The decontamination system
(Release date 10.10.2005)

"What this means is that, for example in trying to recycle the CO2, by the use of the right matter mafs in one of the cores of the reactor, the system will then use the O2 to produce H2O, in the form of pure water, and C in the form of atomic or molecular carbon, or even by feeding this back through certain operational compressions and gravitational interactions within a small core, to create industrial diamond or graphite, for different industries."

The use of the reactor for this purpose has been proven to be correct and Raman spectroscopy has proven the concept and physical reality of the separation of matter mafs in atomic level after separation in the same system from a composite matter mafs.
(See the graphene on the technology web site)

REF 31

The difference in atomic and molecular structure under gravitational force and pressurised conditions.
(Release date 15.8.2005)

In this paper is written that "the atomic and molecular structure of matter mafs is totally aligned and positioned in a fundamentally different configuration where it has come together due to gravitational forces, than when matter mafs has been brought together by the means of pressure. This has a fundamental effect on the behaviour of the structure and properties of the matter mafs which is created in either way". Where this is explained and utilised to develop a new method in creating graphene and Sp3 atoms and walls in a simple core made of a coca- cola bottle.

REF 32

**The Defence and shielding system
(Release date 4.9.2005)

Some extract from this paper:
The design of the side open system for lunch of high saturation magnetic plasma package is one of the most effective technologies for defence of the reactor and the craft, which is in possession of such a system.

It is essential for any space technology in the open environment of the universe, if the craft is to keep a straight line of motion and to be able to protect itself from any solids or matters mafs in its path in the cosmos.

This technology can destroy any object by the principle of plasmatic magnetic energy saturation of the incoming object at a molecular level, that object will disintegrate in the form of magnetic vapour and not atomic level destruction, before matter mafs can damage the craft as debris, and before the craft ever becomes in touch or near these space matters mafs.

This technology can also be used to protect Earth from the collision risk of NEO's (Near Earth Objects), like comets and asteroids.

REF 33

Method of creation of an atom or nano matter mafs
(Release date 11.10.2005)

The manufacture of an atom of any density, this being of the simplest form of nucleus or a full atom of hydrogen, or even heavier elements in the universe, all follow the same principle.

Atoms are created in the soup of the cosmos from collections of very weak plasmatic magnetic fields, which for their existence cluster and share their energies with the magnetic fields that are not far from their own plasmatic magnetic field strength.

REF 34

The relationship between Gravity and Mass
(Release date 21.1.2006)

Extracts from text of this paper:
"A planet like Earth possesses a combination of both gravity and inertia. Where the gravity comes from the interaction of plasmatic magnetic energy fields from the processes within its core, and the inertia comes purely from the collection molecules PMF of matters mafs which has built the physical body of the planet.

Any object, being an electron, atom, molecule or even a human body, is a collection of plasmatic magnetic energy fields of different strengths and their interactions between each other, which in total decides the total or collective magnetic field possessed by that object where this is the mass of the object.

Where the mass of an object is a collective package of plasmatic magnetic energy field tightness of a given object and will not change as long as the object is intact as one entity in its overall atomic or molecular or collective molecular appearance.

Which plasmatic magnetic energy of a physical body (so called the mass of the body) in interaction with molecular or atomic magnetic field (this being gravitational or inertia) of another object will determine the weight of the two object in respect to each other in each other’s environment".

REF 35

The Dark matter mafs
The Fifth state of matter mafs
(Release date 21.1.2006)

Extracts from text of this paper:

The existence of dark matter mafs is not questionable, Proving its existence, and its effect on the energy balance of the universe, in all matters mafs as small as in the heart of an electron or in a solar system needs to be created, replicated and its effect measured.

The theory of dark matter mafs does not and needs not to be considered as complicated, if one understands in real terms the plasmatic magnetic energy of the matter mafs in the visible and invisible dimensions of the universal work.

The dark matter mafs has two distinctive characteristics, which makes it totally apart from the visible matter mafs. The dark matter mafs possess mass but not visible light, where its existence can only be determined by the weight of the it's hidden mass, which could be substantial due to its internal gravitational field.

REF 36
----

*The inter-atomic fusion
(Release date 3. 5. 2006)

Extract from this paper:
"The fusion of two or more atoms has been the pre-occupation of nuclear physicists for past years. Conditions to achieve fusion of two hydrogen plasmas in the TOKAMAK reactors have taken years to teach scientists a lot about the behaviour of plasma and their fusing.

The reality about the fusion in some way has to be reconsidered as, if the scientific world is trying to amalgamate the contents of two plasmas to release energy. There must be a simpler way to achieve releases of similar energies.

The fundamental principle of the inter-atomic fusion is a much simpler way to achieve fusion. If this principle is applied to the atomic condition of fusion, then fusion will be attained in a simple but in a much more physically realistic environment.

The explanation for inter- atomic fusion is very simple and direct.
In the fusion of two plasma of proton of hydrogen atom, the physicists try to fuse two large plasmas, and by doing so, they try to release a large amount of energy. Where in this process enormous amount of magnetic field forces and currents are needed to bring two plasmas in close proximity, so that the energy barriers in between them can be overcome for them to amalgamate or fuse.

In the inter-atomic fusion the overcoming of energy barriers does not exist, as in this method of fusion, the electron of the atom which is circulating the nucleus of the atom and is of the same origin will be encouraged to return into and fuse with the nucleus of its atom".

Where this is a more practical way to create and release manageable energies and there is no need for elaborate systems where even if the energy release will be small compared to dreamed up fusion systems of today.

With inter-atomic fusion method at least small and practical systems can be developed to handle low temperatures of the fusion process.

REF 37

The conductivity of the matter mafs in the vacuum of the interplanetary mediums

(Release date 21.1.2006)

Some extracts form this short paper "Therefore plasmatic magnetic energy field of elements within the interplanetary medium will cover the space given to them in their environment. Where due to their motion and their electric charges even in eV levels, they become perfect conductor, and in turn, due to their motion on their environment, they become plasmatic electromagnetic generators.

REF 38

The production of Graphene

(Release date 25.7.2006)

Repeated experiments and tests in simple static reactors, and tests in more complex dynamic reactors, prove that atomic separation and recombination of matter mafs like carbon and hydrogen can be reached at room temperature and at atmospheric conditions.

We have now indications - through static and dynamic tests in our reactors - that the universe was made in normal condition, which was originally nothing but packages of plasmatic magnetic fields of different strength, which were themselves nothing but areas of plasma or collections of loose magnetic fields energies. Where magnetic fields of different strength in locking to each other, by the principle of their plasmatic magnetic energy (PME), have caused in the first stage the creation of fundamental particles, secondly atoms, then molecules and then matter mafs, clouds and asteroids and then stars and galaxy's".

The big bang theory conditions have no room in the reality of the creation of the universe.

REF 39

Further works on medical research and developments have been written and completed, and will be published on the Foundation medical site in due course.

The process of the reversal of coma: Published in December 2010

Multiple Sclerosis: Published in December 2010

Epilepsy: Published in December 2010

Fibromyalgia: Published in December 2010

INTERNET LINKS

Web site: www.keshefoundation.com

YouTube Channel: http://www.youtube.com/user/keshefoundation

Forum: http://www.keshefoundation.com/forum/

Appendix 1

The unifying field Theory

© M T Keshe 2000-2009, all rights reserved

Date of release of paper 28.10.2009

Abstract

In this paper the origin of electromagnetic fields or electromagnetism and how they are created within the existence of the other magnetic fields are simply explained.

Discussions

The creation of magnetic field and its interaction with physical matter or other fields always leads to release of electromagnetic fields or energy equal to the energy of an electron.

In the universe, energy equal to energy of an electron, where there are no matters for magnetic fields to interact with, for electromagnetism to be created or to be released, in these environments electromagnetic fields do still exists and function.

Electromagnetic fields or electron energy does not need to be created by interaction of magnetic fields and matters as like in generators of present time, as in the universe electromagnetism is constantly produced due to the difference of the magnetic field strength between two fields and without presence of any matter or physical interaction of fields or fields and matters.

Due to the development of gravitational systems and having the ability to create different strength plasmatic magnetic fields within gravitational reactors cores, with the capability of being able to create and maintain different magnetic field strength in different part of reactors, where these fields are continuous and can be kept at a given constant strength level, that their difference in strength in fields can be equal to the magnetic field strength of an electron, where electromagnetic current flow can be maintained from one magnetic field to another due to their field strengths difference, this has enabled us to create and maintain energy field or elect magnetic fields strength of an electron in a plasma reactor without an electron being present or in motion within these reactors.

Where production of electromagnetic fields can be achieved by the simple creation of two magnetic fields with different magnetic field strength of an electron, that as fields are down levelled, from one magnetic fields strength to another, one can creates a flow of magnetic field equal to magnetic field strength of the an electron, thus the creation of electromagnetic current, or production of flow of electromagnetism can be achieved.

Where electromagnetism or current of magnetic fields of an electron can be created at any level of magnetic fields strength, as long as a strength level difference between two fields in magnetic fields exist, and where, in plasmatic magnetic fields the strength level difference of the fields causes the flow of magnetic fields from one field strength to another can be caused or what is called commonly electromagnetic field flow can be maintained.

Therefore, one can have electromagnetic fields of any strength, this being Matter, Antimatter or

Dark Matter, as long as their magnetic field strength difference between to field strength is equal to the magnetic field strength of an electron or magnetic fields of Matters of an electron. Or the magnetic field strength that can affect the motion of other fields to be equal to the magnetic fields strength of an electron. Then at this stage one achieves the electromagnetic current thought flow of magnetic fields, from one stronger to the weaker field, which the magnetic field strength of the flow is equal to magnetic field of an electron (see Fig. 1).

Electromagnetism

In the real world of existence, all Matters do not have boundaries as tangible matters do, but their boundaries are dictated and created through the space they cover through and by their dynamic magnetic field effects. Where, this space of effect of their magnetic field is called the environment of plasma of these fields.

Thus when we speak of plasma or plasma of an electron, proton or neutron, what we talk about here, is the space and environment where the effect of these magnetic fields of the total structure and the entity of for example an electron can cover.

Thus an electron is in fact nothing but a space in an environment, which the total effects of all magnetic fields of all Matters and fields, can be detected or be contained in.

Thus to this end, when an entity possesses magnetic field strength equal to total magnetic fields strength of an environment within the plasma of an electron, then this entity is in possession of mixture of dynamic magnetic fields of an electron or this entity has magnetic field strength measure of an electron or what is called possesses electromagnetic field.

Where this space being in possession of magnetic fields, by its nature has to interact with other magnetic fields and entities made of magnetic fields in it’s surrounding, where these interaction can be attraction (gravitation) or repulsion (Magnetic).

Hence and direction of motion of this space of fields (Fig. 1) or electromagnetic fields is totally dependent on other magnetic fields within its environment, which their environments total strength in gravitation and repulsion dictates the direction of the motion of the plasma of an electrons magnetic fields.

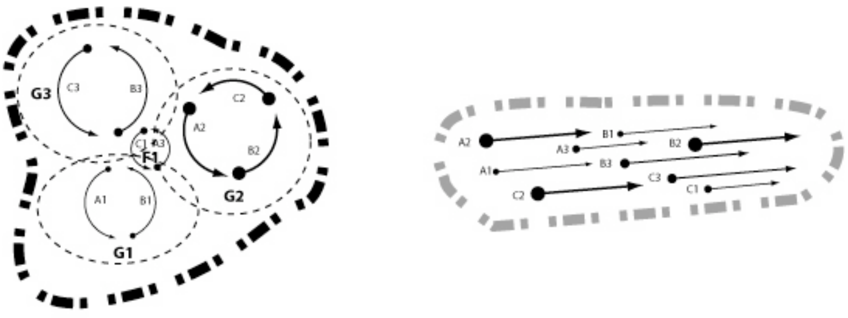

plasma of an electron

Flow of pmtics

Fig. 1 Magnetic fields concept of an electron and a flow of plasmatic magnetic fields equal in strength to that of an electron.

As I have explained in the book, The Universal Order of Creation of Matters (2009, ISBN 978-94-6087-001-9) the magnetic field strength decay of the plasma of a neutron (Fig. 2) leads to creation of two plasmas of the same structure of a proton and an electron.

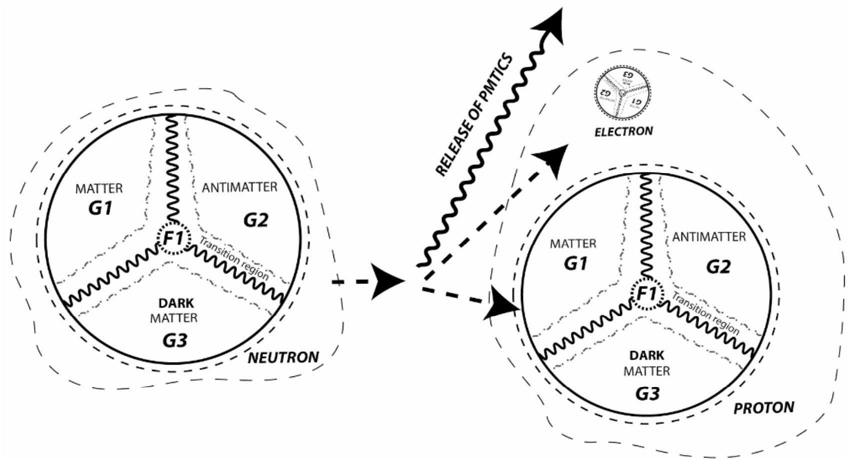
MATTER
G1
ANTIMATTER
G2
F1
Transition region
DARK
MATTER
G3
NEUTRON
RELEASE OF PMTICS
ELECTRON
MATTER
G1
ANTIMATTER
G2
F1
Transition region
DARK
MATTER
G3
PROTON

Fig. 2 Decay of plasma of Neutron into two plasmas of proton and an electron.

It is an important to understand that the structure of an electron in fact is plasma and it is like the plasma of proton but in a smaller scale in the mass compared to the proton (Fig. 3).

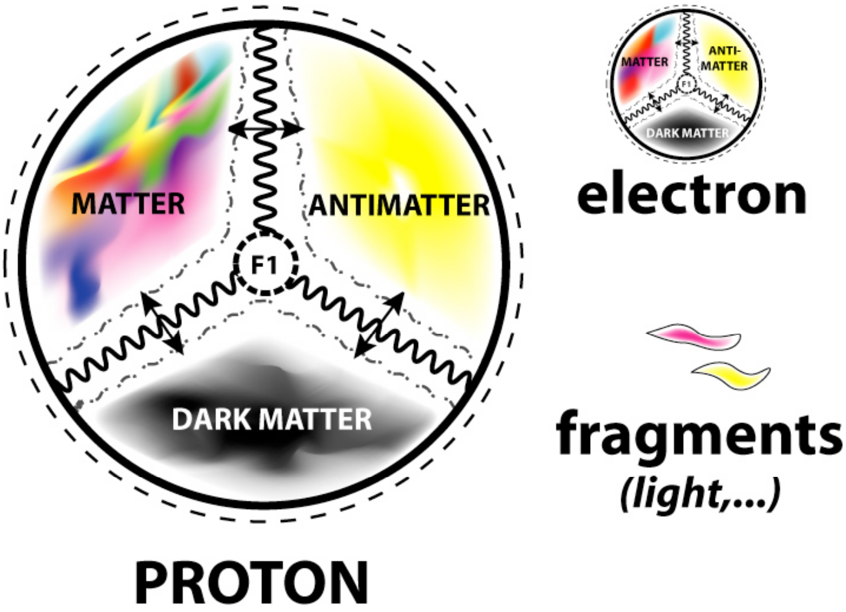
MATTER
ANTIMATTER
F1
DARK MATTER
PROTON
MATTER
ANTI-MATTER
F1
DARK MATTER
electron
fragments
(light,...)

Fig. 3 This showing that proton and electron both possess the same electromagnetic field strength of different Matters, with each entity having different mass of each matter.

Thus the electron is and has the structure of a proton and similarly possesses all its three states of Matters (Matter, Anti-Matter and Dark Matter), and its content is proportion to its neutron plasma prior to its decay into two plasmas of the proton and an electron. As electron is made of electromagnetic field strength of varying strength, this being the reason, why its current and its flow can be influenced by other magnetic fields, as magnetic fields of different strength do influence each other’s path of motion and speed of motion in different Matters and matters environment.

Where, at the same time and by the same principle the dynamics (orbits) of the electron around its proton is due to the continuous repositioning (alternating attraction and repulsion) of the smaller Matters parts of the electron in respect to the large Matters magnetic fields of the proton. This is the reason why that the trajectory of the electron is said to be uncertain.

Where, by the same principle, the Magnetic field source of all Matters and matters and their different magnetic fields strength, is the cause and the sources of all effects and matters we observe in the universe.

In the book I have explained how the gravity is created through interaction of two or more magnetic fields, and how mass is the measure of difference between gravitational fields and Magnetic* fields created by the interaction of the same two or more magnetic fields. Thus even mass is magnetic field strength dependent too.

Further the relation between mass and energy, which both are magnetic field strength dependent (1), and as has been explained, energy is the measure of the mass and the rate of it's the speed of transfer in its own Matters medium (this being Matter, Antimatter or Dark Matter medium) from one magnetic fields strength to another depending upon attraction or gravitational strength from one magnetic field to another. Where, this speed is magnetic fields order of magnitudes strength dependent.

Where, transmutation from one magnetic field strength of Matter to another, like from Matter to Antimatters can lead to creation of current or electromagnetism within the given environment of the same plasma itself too.

In fact in the universe one has proton magnetic field current flows or **protomagnetism** (proton-magnetism) too. This current is the flow of magnetic fields to another, which are equal to magnetic fields strength of one proton. These magnetic fields are bases of the cosmic magnetic field storms and have more profound effects on the structural integrity of the galaxies than their weaker counter parts electromagnetic fields.

Further more as now we understand the concept of plasma of an electron and even by the same principle one can understand the principle structure of the plasma of the proton and by the same measure of the plasma of the neutron, hence in the universe we have observed plasmas of neutrons in motion, which these have the same ability as electromagnetism and protomagnetism in motion and these are called the **neutromagnetism (neutron-magnetism).** Where, these are the space in universe, which carry magnetic field strength equal to the strength of magnetic fields of all Matters, and fields, which make the structure of an entity a neutron.

In fact what is called a neutron star, these in reality do not possess neutrons, but these stars emitted magnetic fields strength in the range of Magravs of neutron. Hence tangibility of the neutron becomes later on as they disassociate themselves from the other fields in their environment, and at this point, they create dynamic space or environment or what we call the plasma of the neutron, where these magnetic fields can exercise their influences on other fields within their surrounding.

Further to this explanation, one can simply understand that in the matter world or in the magnetic field strength of matter, the gravitation and repulsion between these magnetic fields are in the level of matter magnetic field strength of electron, proton or neutron.

Hence in the matter world like in solid magnets, magnetic or magnetism or attraction or gravitation, is the attraction between the magnetic fields strength of Matter components of the plasma of electron, proton or neutron of that matter.

Thus is the reason, why in the physical world certain matters are attracted to each other, which they make certain types of molecules or physical matters. Where, the attraction of atoms to make up any matter it is entirely Magravs strength dependent and no other factor.

Where these attractions or gravitation of atoms or molecules in respect to each other, on other hand is environmental magnetic field dependent too. This is the reason why we can observe certain matters in one environment and not in the others. This being due to the fact that environments themselves are made or created or are conditioned due to interaction and Magravs (Magnetic and gravitational fields) of other magnetic fields.

Permanent magnets

Sometimes in the future I will release a book which initial basic work for it already has been set. The book will be titled “The universal origin of magnetic fields”. In this book I will explain how magnetic fields are generated and how plasmatic and specific Matters magnetic fields exist in different dimensions.

In simple terms, the Matter magnetic field or what we call permanent solid magnets are mono magnetic fields strength which are at the levels of the Matter components of the plasma of an electron, which are in sequence with the plasmatic magnetic field Matters’ of the proton of their element.

What this means is that when plasmatic magnetic field of the Matter components of an atoms’ proton and electron field strength become interconnected and in field balance, then a continuous field between the Matter fields of the proton and electron is set to operate, where in the case of for example Iron at earth magnetic and gravitational field strength, this connection is set to be holding and hence creating a permanent magnet.

Hence as one establishes inter-connection between electron and proton internal field strength, then due to the more of supply from the proton, then the matter stays in one field strength for indefinite time.

This being the reason for example why magnetic fields in level of the water molecules, do not have strength to attract for example cooper atoms gravitational level, but can attract other water molecules’.

In fact iron will not be magnetic as it is on earth in most of positions in the universe. This being due to the fact that even permanent magnetic properties or characteristic of all matters are their environmental Magravs dependent. Hence what is permanent magnetic in solar system, its no similarly the same in other solar systems or galaxies.

Secondly magnetic fields or plasmatic magnetic fields are independent of Matter, but dependent of Magnetic field of Magravs of the Matter, Antimatter and Dark Matter of plasma of electron and proton too.

But at the same time it is important to understand that to have long term or permanent magnets, the Antimatter components of the plasmas have to and are positioned that they establish a continues field between the other antimatter component of the atoms of the same matters too.

The difference between conductors and superconductors

The present world of physics has a dilemma with the fundamental principles of conductors and

superconductors. Where this should not be so, if, scientists could understand the real term difference between the current flow of electron and current flow of magnetic field flux equal to the magnetic field strength of an electron.
What this means is that in conductors, the motion of or vibration equal to vibration of an electron causes the flow of current in the matter.

Where as in superconductor it is not the flow of the electron or the vibration of an electron that causes the flow of the current, but current flow is caused by the flow of electromagnetic field strength equal to magnetic fields strength of an electron, which flows through the material.

Where in the case of superconductors' there is no need for vibration or motion of matter, as with superconductors the same magnetic field strength equal to the magnetic field of the an electron flows through the superconductor materials.

Where in the case of superconductors the lattice boundary of molecules or atoms are so closely and perfectly packed that the electromagnetic flows through the lattice structure without a need for vibration of the electrons of the atoms or molecules of that structure.

In fact one can say that the same electromagnetic field, which starts from one end of the superconductor, itself will be traveling out of the other end of the superconductor, irrespective of the length of the material of the superconductor.

In the case of conductors vibration of electron needs kinetic energy and hence loses, where in the case of superconductors the structure of the material is so that no electron moves, but the material has the ability to transfer the electromagnetic fields strength equal to an electron magnetic fields strength, hence literally no kinetic move, no resistance and no loses.

The superconductivity of the matter is lattice position structure dependent and this is usually better observed in sp3 or diamond-structured materials.

Thus in conductors, electrons have to move and in superconductors magnetic field strength equal to magnetic field strength of an electron needs to be transmitted, and as there are no physical motion of matters within a superconductor, thus there are no loses, this resulting in faster transmission of magnetic fields without losses.

Current and impedance

The origin of all motions in any material or environment is always magnetic field strength dependent. Where in the case of transfer of current in any conductive material for electrons to be moving forward or being capable to transfer current, the material or the atoms of it have in respect to their adjacent atoms two behaviours, one is attraction or gravitation and one is repulsion or outward motion.

In the connection of let say a power source to a energy consuming unit like a light bulb, the light bulb creates a position that has lower gravitational and magnetic field strength in respect to the sources' electrons magnetic field strength. Thus this negative gradient in the magnetic field strength causes the flow of the Magnetic field motion of the electron in the direction of the lower strength magnetic field, from the source to the component.

At the same time, as electron vibration is in the forward direction of the magnetic field strength of an electron towards the component and has to be maintained, there is a similar and opposite pull for the electrons of the conductive matter between the power supply and the components too. Where, this pull is the same as the gravitational field force between the power supply and the light bulb too.

Thus for the connecting wire to be able to keep its constituent atoms together, the atoms of the wire need to have equal repulsion and attraction in respect to each other and at the same time maintain the negative gradient in magnetic field strength between the component and the power supply too.

Thus for supply and component to be interacting that flow of the electromagnetic fields are maintain in between them and due to gravitation of one field to another, the reverse field flow or reverse attraction flow of the fields leads to gravitation field or equal to the field strength of the an electron magnetic field or what is call the impedance flow.

Where, the flow of the current equal to the strength of an electron magnetic field outward is the electromagnetic fields and the gravitational pull between two fields or the current of flow of attraction between the two fields, which create the pull of electromagnetism of the fields, is the impedance of the field.

Thus in fact we can compare the current and impedance to two Magnetic and gravitational field of the planetary system, one can say that the outward flow field of the planet or its Magnetic fields are the current or electromagnetism field and attraction or gravitational flow of the planet is the impedance fields in the matter, where the flow of magnetic fields are restricted within boundaries of a copper wire or an electric light bulb.

Note; At the point where the impedance or the current in a wire increase in one or the other, that is the point, where for example the wire separates and magnetic fields of the atoms of the matters can not hold on to their atomic balance Magravs as the increase in repulsion between the atoms of the wire causes the separation of atoms, and in the case large numbers of repulsion between atoms, then the separation or what is called in common language the melting of the wire occurs.

Where in fact the melting is not due to the heat, but in physical terms, the melting is due to repulsion of atoms of the copper due to not being able to maintain constant gravitational fields of atoms of the wire.

Experimental point's observation

In our reactors we a have the ability able to release the energy of protons which has magnetic field strength by order of magnitude of about 2000 times of an electron magnetic field forces.

Thus one can see how small amount of plasma can create and release vast amount of magnetic fields that one can use these magnetic fields in their dynamic state for the production of electric currents or Magravs in a confined of a small reactor, that these can be utilised for motion and production of vast amount of electric current in conjunction with other field within or within the vicinity of these type of reactors.

Further the interaction of these magnetic fields for example with physical copper winding outside the reactor in a generator structure configuration can lead to creation of energy as we are use to in present generators, without single movement or vibration of an electron within the cores of the reactor and only is presence of plasma of fields equal to the field strength of an electron.

Conclusion

The electric charges, magnetic field and gravitational fields are all of the same and of one origin. Where the gravity is due to pulling effect of two dissimilar poles of the same plasmatic magnetic fields interactions and Magnetic field is due to repulsion of two similar magnetic poles, and electric current is due magnitude of the flow of magnetic field from one field to another.

Thus the ***grand Unifying field*** is the magnetic field, and the missing link has been to understand that gravity is the product of interaction of two plasmatic magnetic fields, mass is the difference between the magnetic fields strength of similar and dissimilar field strength (gravity and magnetosphere forces strength Fig. 25 of the book), and the energy is the measure of the magnetic field that is transferred between magnetic fields at the speed that magnetic field can travel within a given matters' medium, for it to be able to move from one magnetic field strength to another, and the electromagnetic current is the rate of movement of the magnetic field from one field to another and where the electromagnetic field is due to magnetic field strength difference and is the equal to the magnetic field strength of an electron.

Thus the unified and common field is the magnetic field, between the gravity, mass, energy, and electromagnetic field.

Thus the unifying theory for the first time is complete and it is fully understood.

Where *magnetic field interaction* makes **the gravitational fields** and the *flow magnetic field from one field to another* creates **the electromagnetic current**. Energies can be created, from any ambient, depending on what one wants to produce or obtain from these currents or fields.

Electromagnetism is created *as magnetic fields strength are transferred from one field strength to another* and in the process of the flow of magnetic fields this leads to creation measurable field movement or measurable current from what is called "electromagnetic fields".

This simple sentence brings the whole of understanding between magnetic fields and its flow from one magnetic field strength to another and is the grand link between the gravity, the mass and energy and this finalises the long lasting missing link of the unifying field.

Now that one can understands that current is the rate of flow of magnetic field from one strength to other, and “the connection between electromagnetism, gravity, mass and energy” is finally understood. Thus the origin of all things tangible and their effects is magnetic field based and magnetic fields are the origin of all and unifying field in the world of creation.

Reference:

1; The Universal Order of Creation of Matters by M.T. Keshe (2009, ISBN 978-94-6087-001-9) Chapter 5.

*Magnetic field with capital M is the plasmatic magnetic fields created by systems like earth, which this spans beyond the boundary of their physical matter.

Index

A

- ***a-biogenesis*** 81
- animals 90
- anticlockwise **20**
- atom
  - hydrogen **31**
- atomic structure 32

B

- big-bang 33
- black holes 15
- Black-Hole 85

C

- CH4
  - absorption 68
- clockwise **21**
- CO2
  - extraction 67
- *CO2 molecules*
  - absorb 66
- communication system 103
- conversion 24
- cross-section of the light **21**
- current
  - creation 124
- cylindrical plasma 20, **22**
- cylindrical structure **27**

D

- dark energy 15
- dark matter 15
- dilution
  - reactor 90
- DNA **52**, 90
- dreams 94

E

- electro reduction 56
- electron **23**, 24
  - plasma 32
- energy
  - dark **29**
- energy of protons 128
- energy panels 81
- environment **22**
- **extraction process**, 67

F

- fields
  - tightness **30**
- fifth state of matter. 44
- first atoms 31
- food 90
- formic acid 62
- friction **24**
- Fusion
  - of light' pmf 102

G

- galaxial structures **27**
- gamma rays 96
- gamma rays strength 24
- Gans 41
  - definition 77, 106
- GANSE 79
- giant planets 38
- gravitational environment **20**
- ***gravitational pull*** 66

H

- heliosphere 37
- helixal **24**
- helixial **27**, **29**, 92
- Higgs boson 87
- Hydrogen
  - plasma 70

I

- ice particles 39
- increase in mass **23**
- infrared spectroscopes 55
- inner structure **23**
- interlock
  - polarity 95

K

- Kuiper belt 39, 96

L

- LED lights 62
- light
  - dark **28**
  - principal 36
  - surface **29**
  - transition translucent 28
  - visible **28**, **29**
- light bulb 127
- light Magnetisem **24**
- light-entanglement 93

M

- mafs **14**, **106**
  - of matters 28
- **Magnetic** **106**
- magnetosphere **20**, 23
  - universal 32
- Magravs **14**, 52, **106**
  - light 28
  - of the light 28
- MantraEnergy 56
- matter
  - dark **29**
- matters
  - principal 28
  - transition **24**, 28
- medical application 94
- metamorphoses

  first ... 22
  second ... 22
meteoroids ... **23**, 39
minerals ... 90
MOFs ... 55

N

nano CH4 ... 74
nano CO2 powder ... 55
neutrinos ... 37
neutron ... 24
  plasma ... 32
new position ... 23
new state of matter ... 77
North-Pole ... **21**
nuclear decay ... **31**

O

observes
  pmf strength ... **28**
Omar M Yaghi ... 55
Oort cloud ... 37
Oort clouds ... 96
outer layer ... **29**

P

plants ... 90
plasmatic structure ... ***6***
Pmf ... **106**
pmtics ... 33
PN junction ... 55
Principal mafs ... **106**
principal matter ... 20, **106**
Principal matters ... **14**
principal star ... **21**, **22**, **106**
protein ... 99
protein chain ... **30**
proteins ... 81
protomagnetic fields ... **25**
protomagnetism ... 124
proton ... **23**, 24
  plasma ... 32

Q

quantum entanglement ... 93

R

ray of light ... 13
ray reactors ... 93
rays
  alpha ... 28
  beta ... 28
  gamma ... 28
red cell ... 51
resistance capacity ... 55
rotational motion ... **20**

S

seeds of neutrons ... **31**
solid magnet ... **21**
South-Pole ... **21**
sp2/sp3 ... 55
space technology ... 93
spectrum ... **25**
speed of the light ... *13*, 100
speed of the transition ... **24**
spherical plasma ... **22**
superconductor ... 77
superconductors. ... 126

T

Transition energy ... **106**
Transition mafs ... **106**
Transition matter ... **106**
translucent effect ... **29**
twin stars ... 95

U

unicos ... 93
unwind ... 34
unwinded ... **22**

V

vitamins ... 90

W

water ... 38, 56
weight ... **20**
WMAP ... 32
wormhole ... **22**

X

XRD testing ... 77

M.T. KESHE

THE ORIGIN OF THE UNIVERSE

The Origin of the Universe

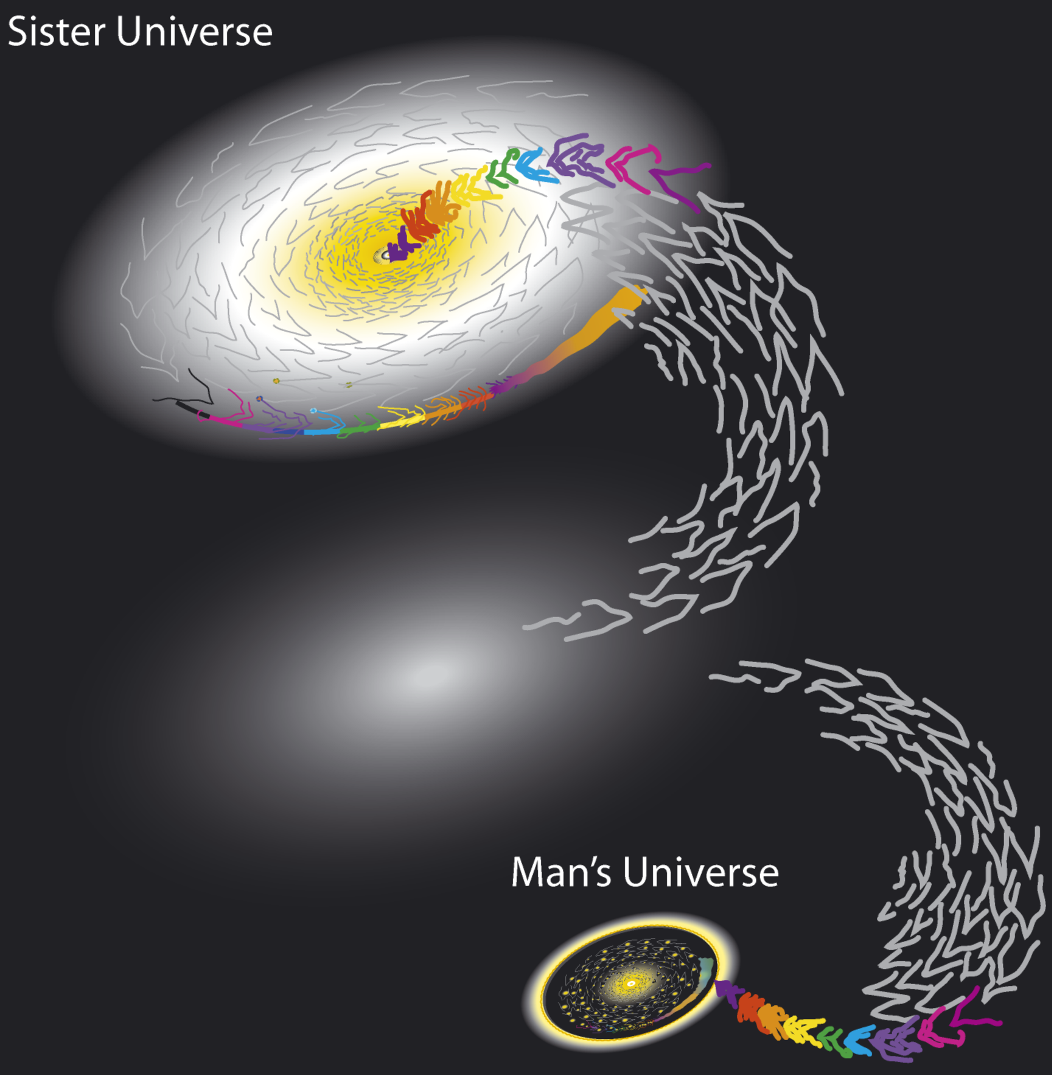
Sister Universe
Man's Universe

The Keshe Foundation

Kf Editions

Nuclear Engineer M.T. KESHE

本电子书不是官方正版，
而是使用从官方购买的正版电子书，
截屏＋拼接＋OCR文字识别＋重新拼版制作。
因此难免出现错误的字或符号，
如果您希望获得原始的官方版本，
请购买正版，支持KESHE。

QQ群：30137653

Stichting

THE KESHE FOUNDATION

(The Netherlands)

The Origin of the Universe

ITNOTGOG

Reflection

The creation of a new universe(s) is as simple as creation of an atom from the division of the neutrons, thus universes are made from division of bigger universes and the same shall be the future of this universe of the Man.

Copyright notice: © Mehran Tavakoli Keshe 2011. All rights reserved.
Nothing of this book may be replicated, simplified, translated and/or made public by means of print, microfilm, tape, audio-recording systems, electronics, optical and solid-state media, electronic storage systems, by network publishing or by any other ways without prior written authorization of the publisher or the author.

ISBN 978-94-6087-005-5

Illustrations and Layout: Dirk Laureyssens

Publisher: Stichting The Keshe Foundation
Registered Office: Jubileumplein 3, NL-6161 SR Geleen, The Netherlands
Registration Number: KVK 14089728.
Web site: www.keshefoundation.com
E-mail contact: info@keshefoundation.com

Forewords

In the past centuries so much discussion has been about the origin of this universe, where in the recent times the groundless concept of the Big Bang theory has taken hold in the scientific world and been considered to the norm accepted fact.

It will be explained in this disclosure that there has been start to this universe, but not necessarily in the way as it has been thought and promoted by the present scientific world due to their lack of the understanding of the universal processes of the magnetic field division and plasma decay as it is the norm in the nuclear structure.

The origin of this universe has to be on the physical realities and in parallel with the rest of the cycles of creations and developments of matters and field as we have seen and observed around us in daily life and the creation of this universe cannot be an exception to rest of universal order of creations and operations.

Introduction

The present visible universe to man is a larger scale in matters and fields of what have been seen by man of science in the structure of the plasma of the neutron, and universes division occurs in the same manner and by the same principle as decay of plasma of neutron to plasmas of an electron and a proton.

The creation of this universe is the replica of the same process as decay structure of neutron but in a larger scale in quantities of magnetic fields strength and matters. This universe is not the only universe in existence and this universe is created through the same natural processes as other elements and entities in the unicos and through decays, division and amalgamation of other fields and elements in the inner sanctum of their mother universe.

Thus in truth there are as many physical universes as there are galaxies and solar systems and man's universe is one of the units in this overall structure of the greater mother universes.

TABLE OF CONTENT

THE ORIGIN OF THE UNIVERSE 3
  Reflection *5*
  Forewords *9*
  Introduction *11*
  TABLE OF CONTENT *13*

CHAPTER 1 15

  The Universe *15*

CHAPTER 2 19

  Theory of Expansion of the Universe *19*

CHAPTER 3 31

  The method of transportation of mafs in different mafs strength in the universe *31*

CHAPTER 4 33

  How matters are created in the universe *33*

CHAPTER 5 35

  The theory of decay and division of the universe(s) *35*

CHAPTER 6 41

  The Man and his soul and their connection *41*
  to the universe *41*

CHAPTER 7 45

  The creation of gravity in the universe *45*

CHAPTER 8 47

  Insight into the realities of the working of the Universe *47*
  and what they bring about *47*

CHAPTER 9 77

  The creation of Gravity *77*
    Abstract 77
    Hypothesis 77
    Prototypes 77
    Discussions 78
    Magravs measuring system 86
    Elaborations: 87
    Conclusion 91

CHAPTER 10 93

  The creation of Black Hole *93*
    Abstract 94
    Hypothesis 94
    Discussion 94
    Creation of the Black Hole 95
    Conclusions 104
    Elaborations 104

CHAPTER 11 109

The relationship between gravity, mass and weight *109*
  Abstract 109
  Hypothesis 109
  Discussion 110
  Conclusion 112
Parallel worlds *113*
  References 115

CHAPTER 13 117

The universal principle of division of pervious universe(s) and their decay leading to creation of new universe(s) *117*

CHAPTER 14 121

How would it end? *121*
Definitions *123*
The re-naming of matters and conditions *125*
Reference *127*
INTERNET LINKS *139*
  Internet radio interviews: 139
LIST OF ILLUSTRATIONS. *141*
INDEX *143*

CHAPTER 1

The Universe

The term Universe to me as a nuclear physicist's has one meaning, and that is the more of the same magnetic fields of different strength as it is in the plasma of an electron or proton. Where, the total behaviour and operation of all magnetic fields and matters created by fields within the operation of stronger fields' strength catchment's in the centre of each entity is considered to be of the same and through the same and similar processes.

Where in the smaller scale entity the like of the plasma of proton we call the bigger and stronger entity to be is the Principal mafs and matter, or what is commonly is called the up or down quark in the present physics, where in the universal scales we call these the central the principal mafs line, where the other components within the plasma of the universe are exactly created and operate as the galaxies, solar and plasma of electron systems within these universes are created and operate.

Thus, the difference between the plasma of a neutron and the plasma of the universe is the size of the mafs of the fields which has come together to create that entity and otherwise all self-contained gravitational-spherical and magneto-spherical entities this being the plasma of proton to the plasma of solar system, the plasma of a galaxy or the plasma of the universe, they are all created through the same principle and operate through the same physical conditions and restrictions.

Where, under certain conditions of magnetic field or gravitational fields strength and positioning in certain parts of the these plasmas, whatever they could be, conditions can be created that allows the growth and suitability condition that clusters of fields can become active and sustain a rhythmic and coherent field interactions that these can lead to creation of certain types of repetition of the same fields in these plasmas.

Where, the repetition of Magravs of these types' fields within these plasmas of different scale can lead to creation and sustainability of continuity of being able to absorb mafs in motion from their environment that these clusters of fields can sustain long-term durability and exist-ability, or what we call to attain the process of life.

Where, different versions of life can exist in the plasma of proton as it does in the plasma of the solar systems as is on earth and different forms of lives exists in the plasma of the universe and universes within the unicos of the creation of the bigger structures of universes.

Thus, this universe or other universes, which have led to creation of this universe and the cycle of the life of this universe in time it will lead to creation of smaller universes, where these in time have all come from one essence of a larger fields entity(s) and in time as they divide and they become weaker, then in time these fields will gather together to create a different universe(s) with different field(s) strength properties (Fig. 1).

Thus, this universe is not unique and there are others, which are created through the same process(s) and there is remote chance that there could be another universe similar to this universe in the ***unicos*** (Fig. 3) of the creation, but not necessarily it will have the exact combination and structure of mafs and positioning of its fields and matters in its internal structure as man's universe.

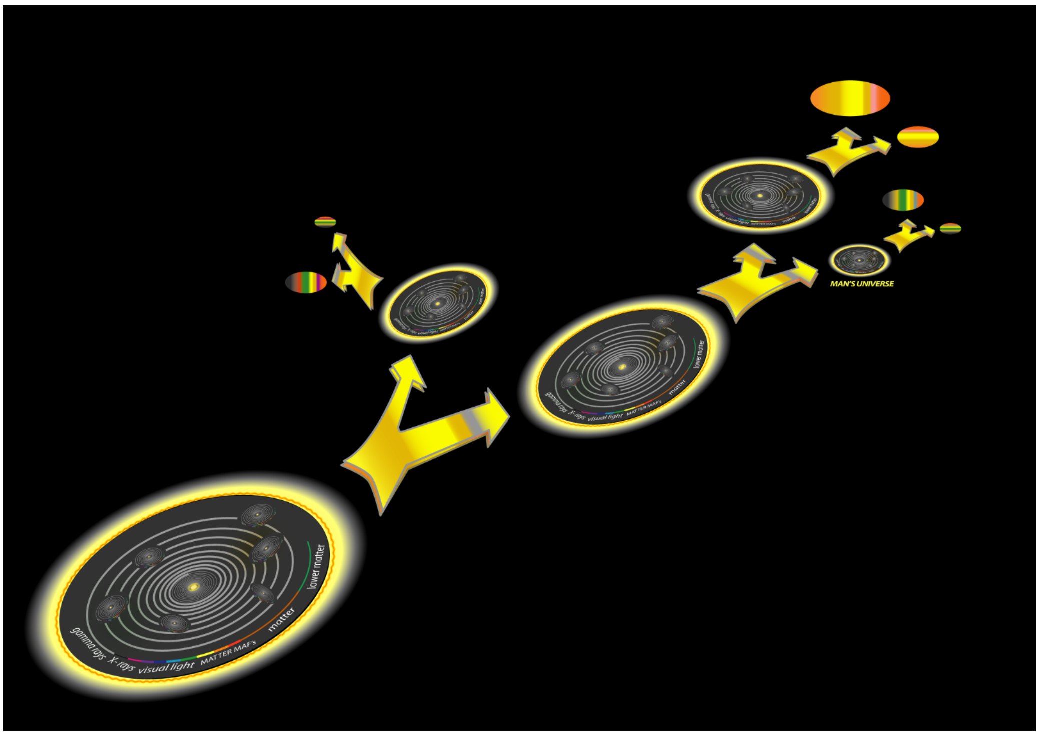

Fig. 1. Creation of new universes out of the division of a larger universe
through the process of decay similar to the decay of the neutron in proton and electron.

In truth the universes are created with a long whisper as one sees and hears gases leaking out of cylinder or a balloon (Fig. 18 C an D), and more like trajectory line depending on the magnetic and gravitational balances of their mother universe and outside environment, which they are getting released into and as the release gases try to find their new position in the new environment, which they have been released into.

Thus the creation of universes are not made with huge sound and explosion but very much with a long whisper as the plasma of the larger universe is dividing and ejecting its matters into two or more environments to guaranty the survival of its overall Margrave strength structure. Very much as the neutron divides and creates the proton and the electron plasmas.

Then as each universe reaches it's own Magravs balancing positioning then it again divides and creates further universes (Fig. 2).

This has been the truth about the creation of this universe in the unicos and in time man will come to understand that the creation of unicos, universes, atoms and plasma are nothing but natural dynamic processes of the interaction of different amount and strength of magnetic fields at different time and position in the overall world of the creation.

Then the question becomes, where did it begin to start with (ref 23), then one needs to understand the truth about the levels and position differences of the origin of the magnetic fields.

This in time will be revealed to the men of real insight in the world of creation as the man at this moment is not mature enough to understand and comprehend the totality of creation of the universe.

If one reads all the writings' of the author, and if! and that is a big if, then the reader will understand the truth about the essence of creation and the origin of the creation. One will come to see its creator way of operation and purpose of creating such an intriguing mesh of unify principles in all levels of creation in all of unicos's.

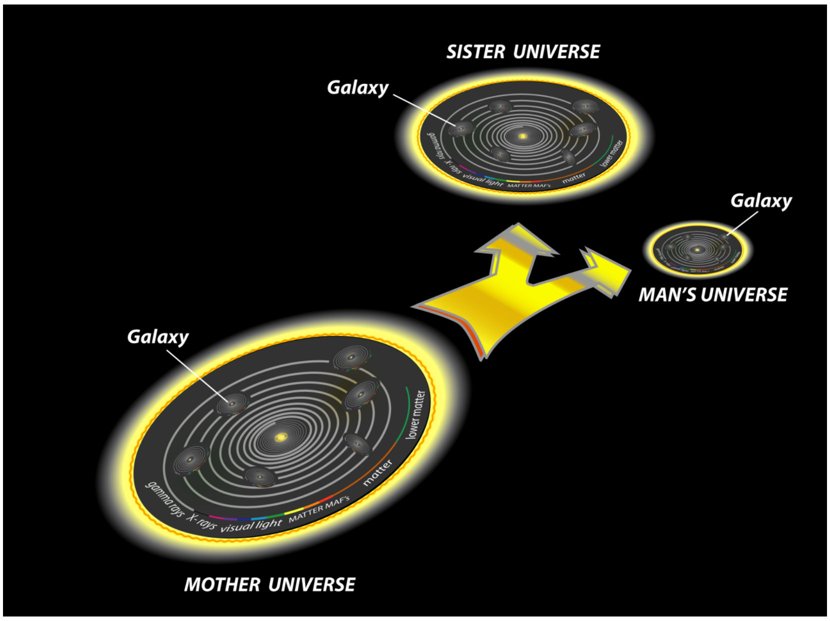
SISTER UNIVERSE
Galaxy
gamma rays
X-rays
visual light
MATTER MAF's
matter
lower matter
Galaxy
MAN'S UNIVERSE
Galaxy
gamma rays
X-rays
visual light
MATTER MAF's
matter
lower matter
MOTHER UNIVERSE

Fig. 2. The creation of Man's Universe after the decay of it's Mother Universe Magravs' maf's.

This shall be the point of maturity of the man to understand how feeble he has been created according to his understanding, where in fact he carries the whole of the world of creation within himself

CHAPTER 2

Theory of Expansion of the Universe

There have been theories and observational data, which support the concept of the expansion of this universe at the present time.

There are two aspects to this expansion theory of this universe, which each one has to be considered on the real terms of expansion of the universe that both can be scientifically acceptable.

In principle, it can be said that this universe cannot expand in to space in bigger unicos's space, if this space was not there for Man's Universe to expand into and to begin with. Thus this universe cannot be the only universal system that exists in the bigger **UNI**versal **COS**mology (unicos).

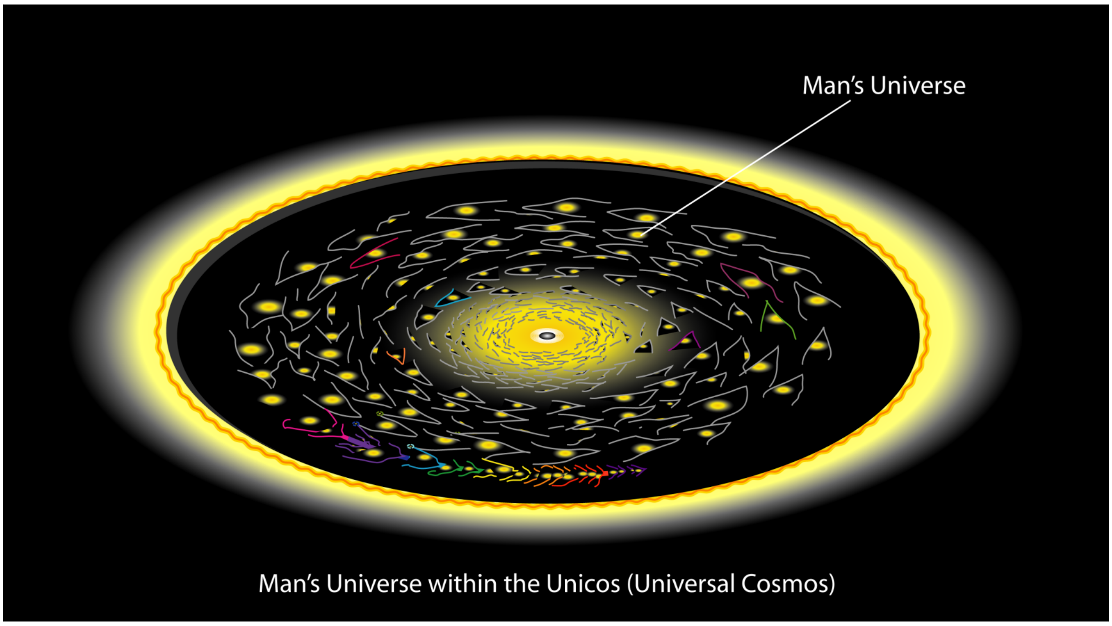
Man's Universe
Man's Universe within the Unicos (Universal Cosmos)

Fig. 3. The Unicos

The Man's Universe is a small fraction part of the unicos universes (Fig. 3).

This Man's universe is expanding due to two factors, ***firstly*** as the principal mafs of the central fields of this universe are unwinding they make available more mafs in the strength of different matters (Book 2) that as these fields open up they lead to creation of new galaxies and other entities, and ***the other factor*** is that the stronger gravitational fields environments which are outside the boundary of this universe's at the same time are pulling on this structure and keep it dynamic (on the same principle of motion of electron around it proton (Book 1, Chapter 19)), and the expansion of this universe in reality is part of the motion of the unicos as in the direction of the motion of all the universes in respect to unicos's internal central Magnetic and gravitational field forces interactions.

This meaning that if one looks at the overall directional motion of all the galaxies in this universe, it will be noted that not all galaxies in this universe are following the rest of this universe's galaxies on a straight overall spherical one line motion of expansion, but in fact part of this universe's galaxies overall line of motion is curvature and is following in the direction of the pull of an outside sister's universe gravitational field force, which this is necessary as this universe is to find its balance and position in respect to other external fields forces of the same of other universes in the unicos (Fig. 4).

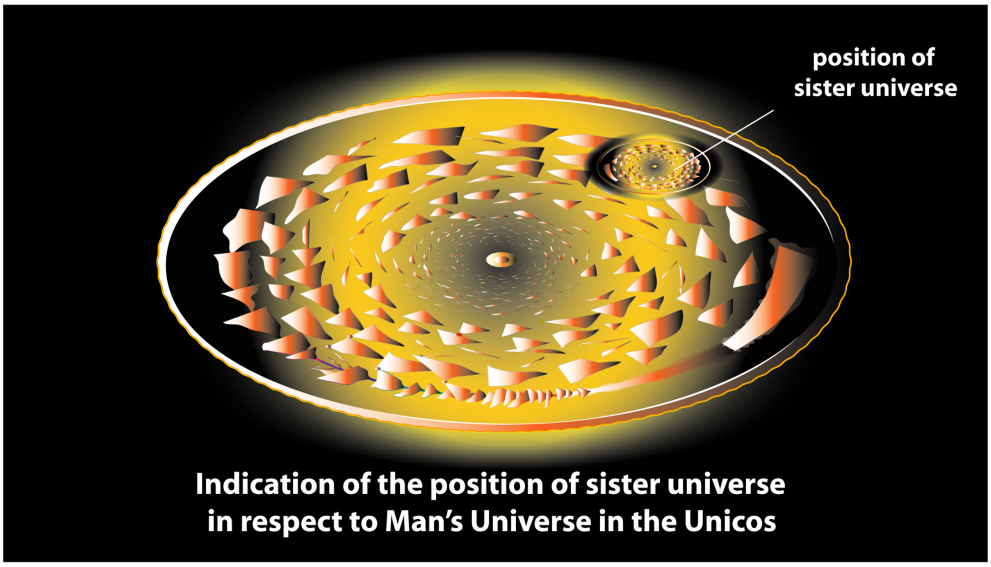
position of
sister universe
Indication of the position of sister universe
in respect to Man's Universe in the Unicos

Fig. 4 The position of the sister universe in respect to the Man's Universe.

At the same time each universe possesses a central principal mafs and matter point and or central principal matter line, where these mafs strength and quantity of in this centre point or the centre line dictates the size of any universe and at the same time the Magravs created by this is like a Principal star and line in the centre of the universe that this dictates the position of the universe in respect to other universes in the unicos and this Magravs strength of the central line dictates the matters which are attracted and created in that universe and the matter and content of Galaxies' in respect to itself and no other matter in that universe.

Where, the total mass of the mafs and transition mafs in a galaxies are about less than five percent of the total mass of the universe principal matters and mafs in its central line and its principal star.

Thus the positioning and directional motion of the galaxies in the universe, the same as in the solar systems is dictated by the mass of the Principal mafs and mafs in that universe central principal mafs and star in the centre of that universe but in larger order of magnitude of contents and due to no other factor.

At the same time as the universe is created out of the ejection of the large quantity of principal mafs and matters and have motion in their time of creation, thus it is more feasible to consider an stretched line of principal matters and mafs across a central line of the universe rather than a centre point universe star of principal mafs and matters in the central line of universes.

Therefore the mass of the universes are measure of their principal mafs and matter content, which their Magnetic and gravitational field forces dictates every motion of matters in galaxies in that universe, hence strength of the principal matters and mafs in the central line and central principal star dictates that universe's-magnetosphere, gravitational pull sphere and the mass of that universe.

To understand the process of creation of mafs and matter and elements in the universe, one needs to look at the smaller scale universe, which are similar to the constituents' plasmas and the structure of an atom, its creation from division of the neutron to a simple hydrogen atom with it's plasma of electron and proton, the way they are created from neutrons division and the mafs which are released in the process of decay of the neutron into its environment, leading to creation of the simplest atomic system of Hydrogen.

Where, any mafs realised through and during the process of the division is not necessarily transferred to the newly created plasmas that lead to heating of the two new plasmas of the proton and electron.

Thus there could be some heating of the plasma of the electron and proton during and due to decay of its mother neutron, but this energy may and possibly and do *not necessarily is transferred totally that leads to heating up of the plasma* of the proton and electron.

Therefore the same applies to the division of this universe, when it was created ***from the decay of it mother neutron universe.*** Hence there are no logical reasons that the universe was heated and was hot and it has cooled for elements in this universe to be created due to creation of gravitational field forces.

As in the production of two parts of the plasmas of an atom (1) from the division of the plasma of neutron, the present assumptions in physics that the first atoms in the universe were created ***when the universe cooled***, this theory is not and cannot be correct and in reality *atoms are created through change* ***in the plasmatic fields strength of the mafs of the light***, secondly at the point of creation of the universe, the universe was not hot and there were never high temperatures that the cooling of the universe has led to creation of the first atoms as presently commonly stated. The first atoms and universally all first atoms' the like the atom of hydrogen in all mafs strength in the universe are created through the magnetic field strength reduction of the principal mafs and through no other means.

The concept of creation of heat and high temperatures at the birth of the universe is not observed in the creation of plasmas of electron and proton, when the plasma of neutron decays and divides into an atomic structure, and in the universe there are always continuity and parallels, thus the universe has never been hot that has cooled down to create atoms, and this is the reason why W-MAP (1) shows at present balance heat distribution in this universe (Fig.5).

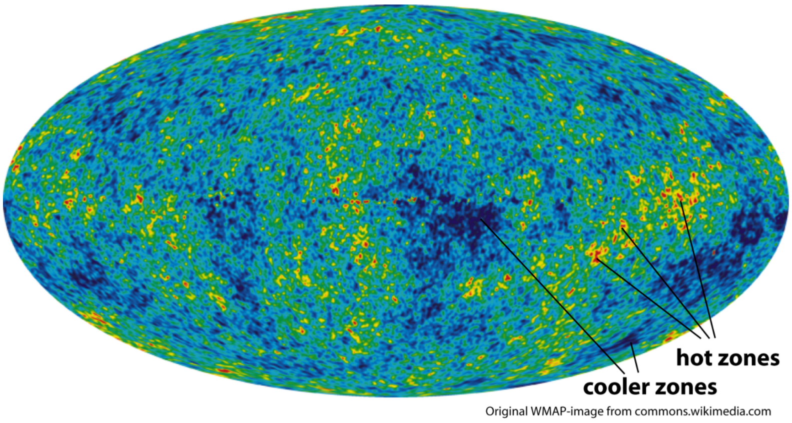

Original WMAP-image from commons.wikimedia.com

Fig. 5 The universe heat distribution map (W-MAP of the universe)

In truth the current assumptions of initial heating of the universe at time of its birth cannot be true as the initial atoms of the universe and present atoms were and are created from the original principal mafs strength, ***which radiated out as principal lights rays of the universe*** in the centre of the universe and the initial atoms and matters of this universe were not created through cooling of the universe at the point of inception of this universe, which this universe itself is created from the decay and division of bigger mass of another universe as it happens in the division and decay of plasma of neutron in two plasmas of proton and electron. One does not observe heating in the plasma of electron or proton in this division and so be it the same, when the bigger universe divided and led to creation of this universe of the humans. (Fig. 18).

In fact, if! The universe was created by the present scientific assumption then the W-MAP should have shown cooler region in its outer boundaries than the present near homogenous heat distribution, which has been physical observed and from collected data by the publishers of the universal heat distribution map of the universe.

The heat distribution and the heat concentration on the central line of the WMAP confirms that the universe started from a central line in the middle and as the principal mafs of the universe unwinds and reduced in strength (Book 2) then the ***transition mafs*** and the ***matter mafs*** were released and created, and as these were spreading outwards into the space of a larger universe (Fig. 1) from its central line ejected mafs into unicos, then as the Magravs of the principal mafs in this central line region are converted and are reducing in strength, then the overall unified oval-spherical shape explanation of the WMAP come to exist.

Then as the total Magravs of the central principal matters reaches the point of the equilibrium of its forces, then at that point, this becomes the boundary of the universes boundary or becomes the plasma of the universe. At this point where they Magnetic and gravitational field forces cannot hold on to all the mafs and matter of that universe and this become the point of the limit of the expansion and the boundary of the universe and the envelop of the mafs at the given point in time, which are under control of these forces from the central lines principal mafs of that universe, where this expansion line becomes that universe’s magnetosphere boundary line.

Then, When a point in the cycle of the life this universe reaches that the overall strength of the Magnetic fields and gravitational fields of the central line principal mafs and matters cannot hold to their overall fields balance, then at this point this universe will go through a division the same as the division of neutron decay and all the matters and mafs at this point rearranged and regroup as and in the principal strength mafs and the universe then divides into two principal mafs of lesser masses of universes, which these new universes are magnetically interconnected as proton and electron do after the division of neutron and then these new universes go through the same principle of the principal mafs light unwinding and the same process as creation of new universe carries and central line as before.

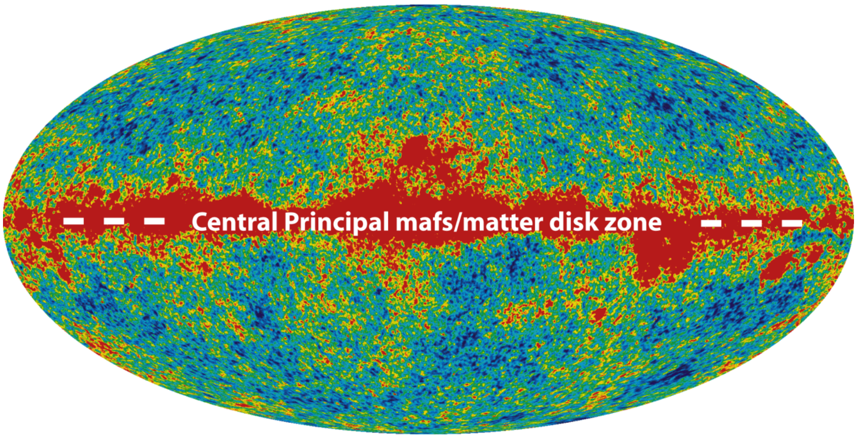

Original WMAP-image from commons.wikimedia.com

Fig. 6 The microwave heat distribution of the universe shows a concentrated heat in the centre disc of the universe

That is to say through the principal mafs and matters are the way that universes divided and arrive form another universe to start a new universe and as the light in the principal mafs strength unwinds' in its new environment then the new envelop of the new universe starts to be created.

The same and similar process occurs, when the neutron decays and divides into the plasma of the proton and electron.

Thus, the universe is not created as a big-ball of fire which has been cooling down as has been promoted, but in fact the universe is created the like of as ***an ejected principal mafs*** is injected into a new area in the unicos and as the new universe environment and as the principal mafs radiate outward as light of stronger mafs, then as the weaken and radiate outwards in their new environment, then they create the other matters mafs and matters in this new envelop depending on the given pmtics strength of a given region in the larger process of this unwinding of the principal maf and matters and its lights (3) in this newly born and expanding universe.

Therefore: this universe is in fact created inside out rather than what has been assumed up to now that as it cools down to what it is, which this explains why even the universe is observed to be expanding, as the reason for expansion of the universe is being due to the presence of principal mafs in the centre of the universe and as principal lights of the principal mafs radiate outward, these light rays of the principal mafs strength while traveling in their new environment they reduce in strength and this leads to creation and transition mafs and matters, which further on losing its strength creates new matter mafs and matters according to given environments magnetic fields strength and in a given point pmtics, hence the creation of new mafs and matters and so forth, which this leads to creation of atoms and molecules and stars and so forth in this universe and its galaxies and solar system through the same principle.

Thus there has never been a Big Bang and the current appearance of the expansion of the universe is due to the principal maf, which as unwinds, it creates the expansion of the universal-sphere of this universe (3).

This method of the creation of the universe now explains' why there is a central gravitational matter concentration in the universe and as this pull is centred in the dynamic principal central line, which the original principle mafs were ejected out of the mother universe when it divided, where the injected principal maf of the universe at the time of its inception can be seen as the red line in the W-MAP (Fig. 6) which the more concentration of the mass of the principal matter sets in the central disc of the W-MAP and as these principal mafs create their own magnetic interactions, hence these create their own sphere of gravitational and Magnetic fields and thus the central gravitational of the universe is in the centre of the denser point of principal mafs of the centre universal heliosphere of this universe and as the more principal mafs are unwinding and spread outward are reduced in their strength they principal mafs become transition mafs and further as they reduce in strength they become the matter in the unisphere of the universe and as the principal mafs open up and reduce in their environment this leads to creation and expansion of the balloon of this universe in the unicos, and this explanation resolves the present mystery in the scientific world that, how and where the central gravitational pulls and the expansion of this universe which has been observed by scientists of the universe comes from.

This is how the plasma of the neutron decays and divides into plasma of the a proton and an electron too, where the different quantities of matters and mafs of different strength of the neutron are divided and injected in two new areas in their environment, that their overall Magravs of all their mafs are as original components that the plasma of the electron and proton concurrently maintain their overall Magravs of the original neutron Magravs (Fig. 7) and then as the matter and mafs unwinds and interact they reach given overall Magravs level and position as the new plasmas of the sub-components of the atom are created.

Where, this process of division of neutron due to micro-level of masses involved, these appear as an instantaneous process of creation of proton and electron, where the same applies to creation of universe and larger universes divide to guaranty their overall existence in the unicos.

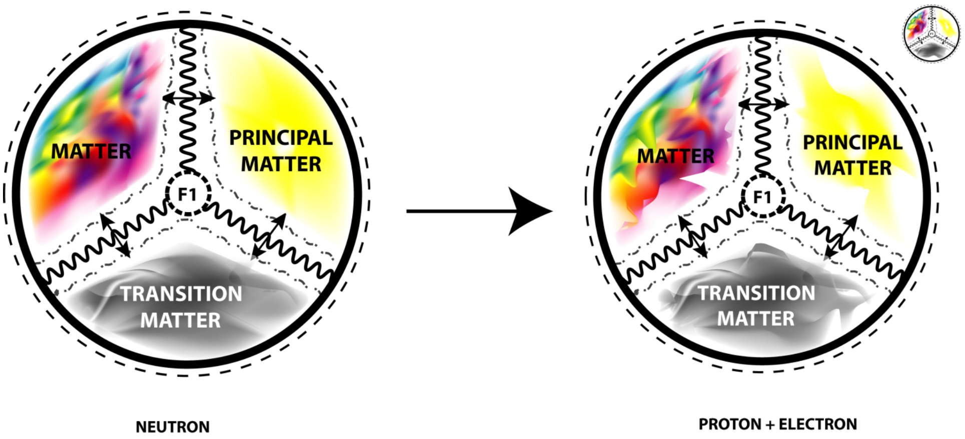
MATTER
PRINCIPAL MATTER
F1
TRANSITION MATTER
NEUTRON
MATTER
PRINCIPAL MATTER
F1
TRANSITION MATTER
PROTON + ELECTRON

Fig. 7 The conversion of the plasma of neutron to plasmas of the proton and the electron.
The same process is considered to occur when the universe divides (3).

This being the reason why the proton and electron in their process of creation ***do not get heated*** as a Big Bang theory suggests but as different portions of principal mafs of the neutron which become parts of the plasma of the proton and electron unwinds and expand from the central line of their centrally disc which they are rejected into their environment by the division of the neutron, and as *their principal maf unwind* in this disc to the ***transition matter*** and then, as they weaken further, they create ***maf and matters in tangible mafs levels***, the like of the plasma of the proton and ***all other fragments of pmtics and magnetic fields*** in these new plasmas (Book 2, Chapter 1 and 2), in these processes of creation of new smaller plasmas of there is no Big Bang and there is not heating of the plasma of the neutron and electron (Fig. 8).

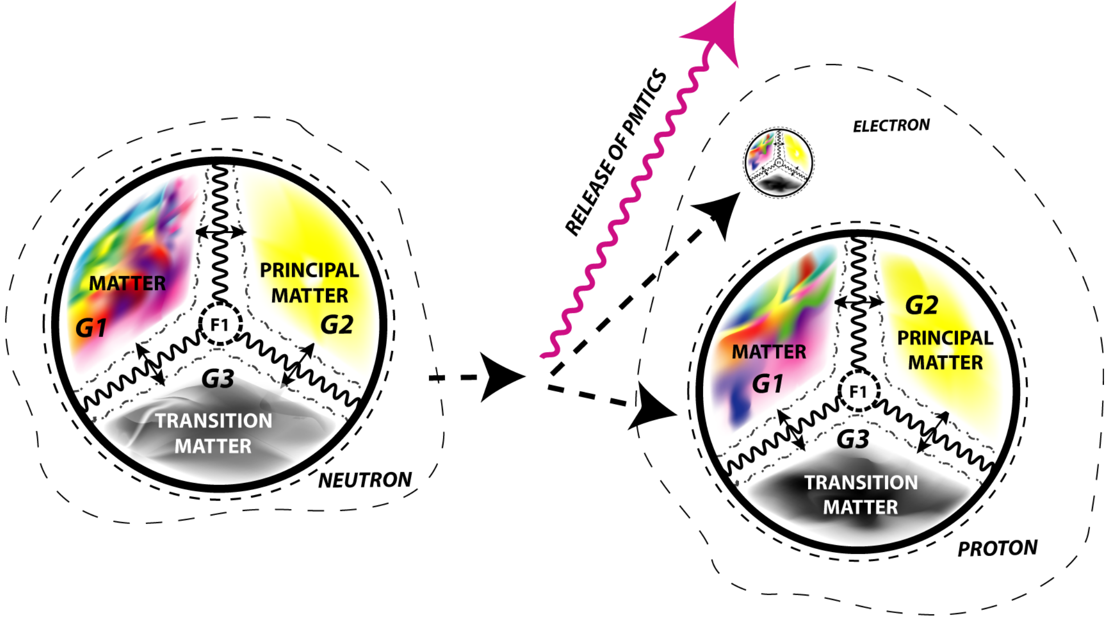
RELEASE OF PMTICS
ELECTRON
MATTER
G1
PRINCIPAL MATTER
G2
F1
G3
TRANSITION MATTER
NEUTRON
G2
PRINCIPAL MATTER
MATTER
G1
F1
G3
TRANSITION MATTER
PROTON

Fig.8 Decay of plasma of neutron in plasma of proton and plasma of electron.

Hence the theory of Big Bang is not correct and has no meaning in the real world of creation.

The same process can happen with the ***ray of the light*** originated from principal mafs strength (what is called the **principal light**), when this principal light arrives in the principal mafs strength environment and hence the first fundamental neutron principal plasmas and principal atoms' can be created in that given environment through the same principle as atoms are created in the matter strength environment (1).

At the same time the Magravs strength of the **P**rincipal **U**niversal **L**ine **M**afs (PULM, the central line of the universe)(Fig. 9) dictates the position and attraction of other matters, stars and galaxies' in that universe in respect to itself.

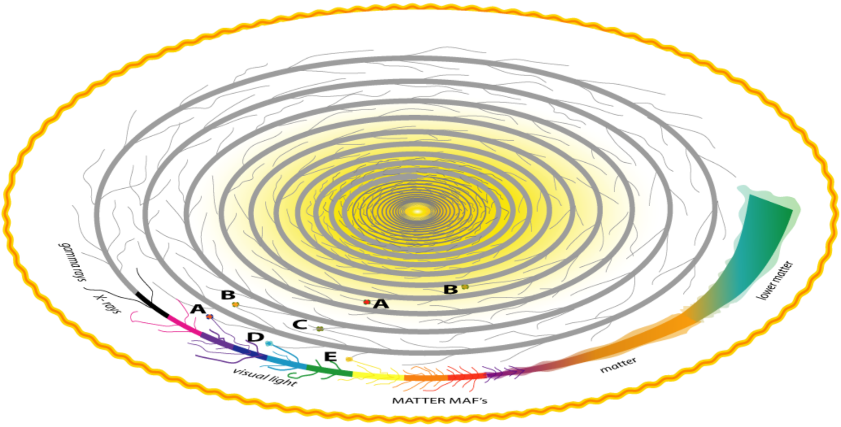
gamma rays
X-rays
A
B
C
D
E
A
B
visual light
MATTER MAF's
matter
lower matter

Fig. 9. The Principal Universal Line of Maf's.

The total mass of the matter mafs is about five percent and transition mafs of the matters about fifteen percent in any universe and is the same as with galaxies, solar and plasma of proton systems in the universe. The reason for this is due to the principal mafs (Fig. 10) and mass conversion and conservation and motion as they move within the plasma of the universe and lose their strength and reduce to transition, and matter mafs. Secondly if any mass more than this ratio of principal mafs is converted and stays within these systems, these can lead to unbalance and faster division of the entity, this be it plasma of the proton, neutron, solar system or universe.

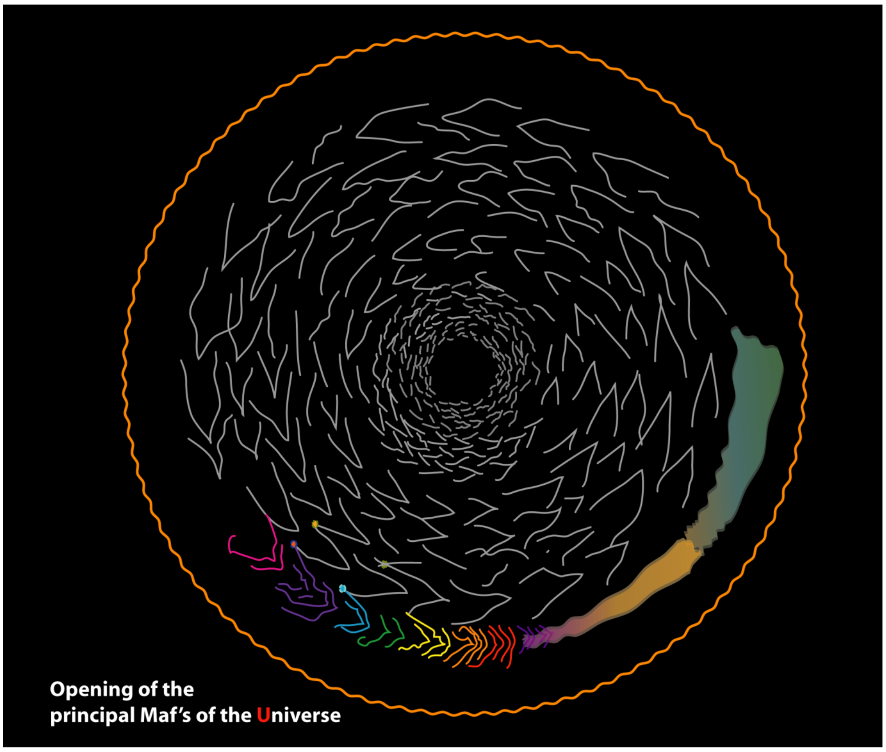
Opening of the
principal Maf's of the Universe

Fig. 10. The principal maf's of the Universe.

Thus, the positioning and directional motion of the galaxies in the universe, the same as in the solar systems is dictated by the mass of the principal mafs and matters of the principal mafs of the universal line star mafs (pulm) of the galaxy and no other factor.

Therefore the mass of the universe is the measure of the difference in the interaction of the Magnetic and gravitational fields' forces that the principal mafs contents of the principal matter in the universe can produce and maintain. The same measurement principle applies with the galaxial and solar system.

Once these principal stars are formed in the universe then the Principal stars and the lights which are releases and radiate outward from the pulm becomes the means and the supplier of pmfs and nourishment source to all entities in the internal plasma environment of that universe, these be it galaxies, solar system atoms and so forth (Fig. 11).

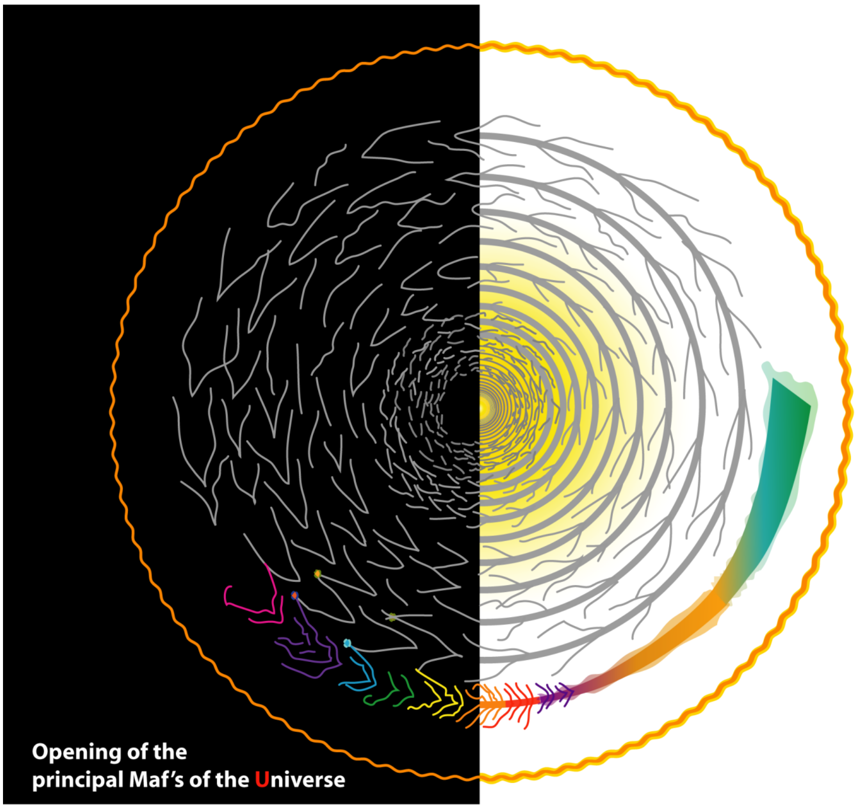
Opening of the
principal Maf's of the Universe

Fig. 11. The unwinding of Magravs in an entity.

The same applies as to the galaxies, solar systems and plasma of proton systems too, where in fact in the centre of each galaxy, solar system and plasma there is the principal star (Fig. 12) with a plasma of principal mafs and matters and as these unwind, expand and spread within the Magnetic and gravitational fields which they create, these releasing, opening's and loosening of principal mafs in the pulm leads to creation of transition and then matter strength mafs and matters in the envelop of the magnetosphere of the universe leading appearance and creation of principal matter in the galaxies, solar and plasma systems, exactly the same as it has happened and has led to the creation of the plasma of a neutrons, stars, galaxy's.

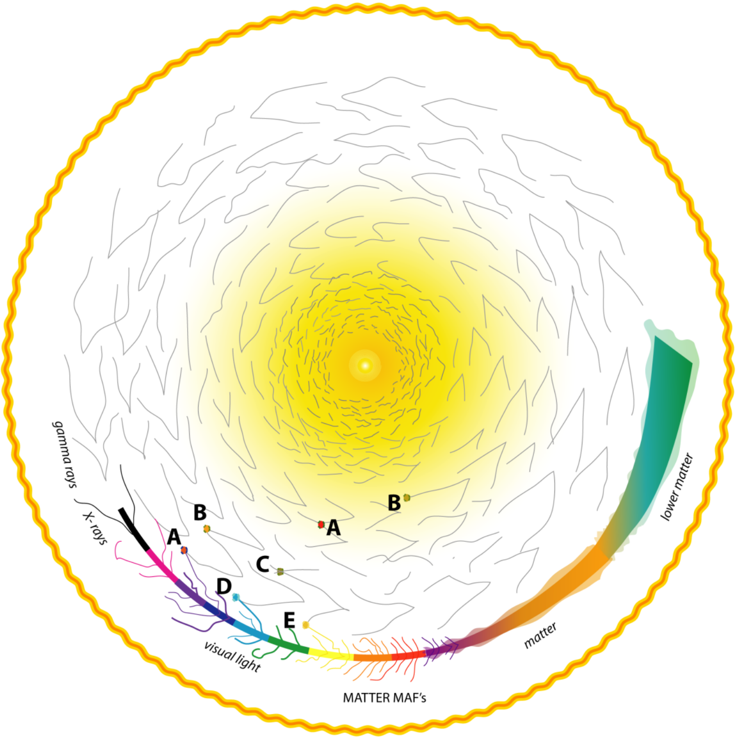
gamma rays
X- rays
B
A
C
D
E
A
B
visual light
MATTER MAF's
matter
lower matter

Fig. 12 Plasma of the Universe: any entity can be created anywhere in the Universe as long as the appropriate maf's strength of that entity is reached through reduction or interaction of maf's of the Universe

Where, A(s) are the universal magnetic fields oceans in the form of fields forces or light, where sea of lights are created by the interaction of the fields which these can be in any light strength of matter, transition or principal magnetic fields strength, B(s) are universal matter oceans, C (s) are the transition matter and mafs oceans, D(s) are galaxies, E(s) are sub-matter mafs and matters oceans in the plasma of the universe, where all these entities they are created due to and through the interaction of the residual fields which are released by and through the release and separation of fields from principal mafs and transition mafs and matter mafs of the main or the subsequent divisions of these mafs as the principal mafs of the universe open up and divided and expand in the structure of the plasma of the universes' Magravs (Fig. 13).

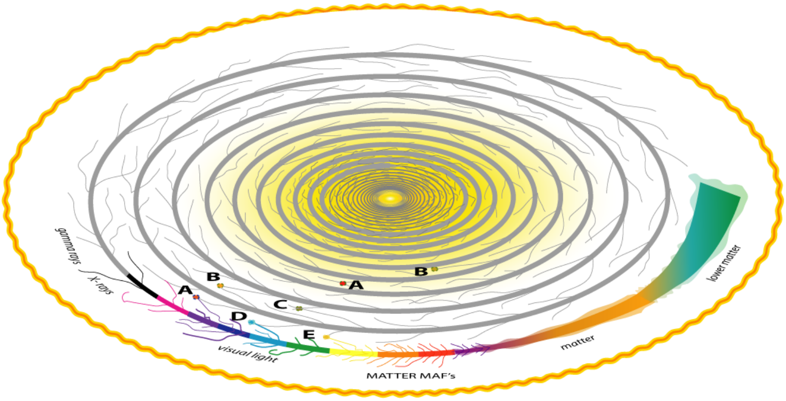
gamma rays
X- rays
B
A
C
D
E
A
B
visual light
MATTER MAF's
matter
lower matter

Fig. 13 Plasma of galaxy: any entity can be created anywhere in the galaxy as long as the appropriate maf's strength of that entity is reached through reduction or interaction of maf's of the galaxy itself or its Universe's maf's

Where, the same principle applies to the structure of the creation of stars and planets in the galaxies and solar systems, and where the same process of creation of matters and fields is applied in the creation of entities in the plasma of neutron, proton and electron.

Where, A can be the point of creation of a star, a B the black hole, C the creation of planet in the solar system, D the cloud dusts and E the magnetic fields ocean of the galaxy.

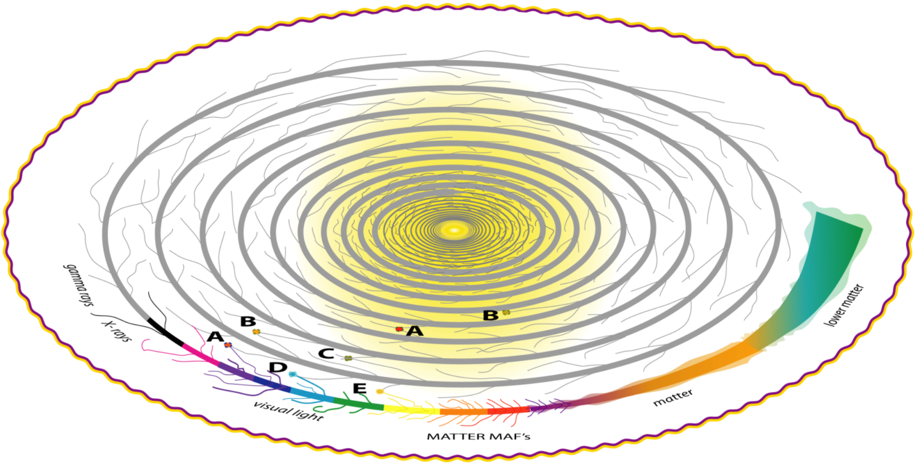

Fig. 14 Plasma of Neutron: any entity can be created anywhere in the neutron as long as the appropriate maf's strength of that entity is reached through reduction or interaction of mafs of the neutron and it's galaxy itself or it's Universe's maf's.

Where, A can be the principal matter, B the transition matter, C the matter, D are the fields in motion, E are dust of matters in creation and motion within the plasma of the neutron (Fig. 14). Where even the creation of the X-rays and gamma rays are normal features of the plasma of the neutron as one observes in the process of the existence and operation of the galaxies (Fig. 15) and universes plasmas.

The universe as the plasma of the proton can expand according to the amount of the transition matters which can travel away from the principal matter line, and hence the expansion of the universe or plasma is dictated by the amount of the transition matter and mafs and speed of by which the principal mafs and matters can release which the pulm can hold on to the total Magravs of its structure. Now this can explain why this universe through WMAP (Fig. 5 and 6) shows to be more oval than sphere in shape.

On the other-hand the shape of the WMAP being oval can explain the same characteristics as one observes in the major composite gas planets in the solar system as they are not fully sphere in shape, but have more flat poles.

The oval shape of this universe confirms the existence of a central principal disc point in the centre of the universe and the flat disc spread shape of the principal mafs distribution in the central line disc of the universe.

At the same time the uniformity of the speed of the objects in the universe and galaxies shows the connection between the principal mafs release in the pulm.

This confirming the real process by which plasma and atoms are created and appear across the universe, where their accumulation leads to the creation of cosmic dust, stars and planets.

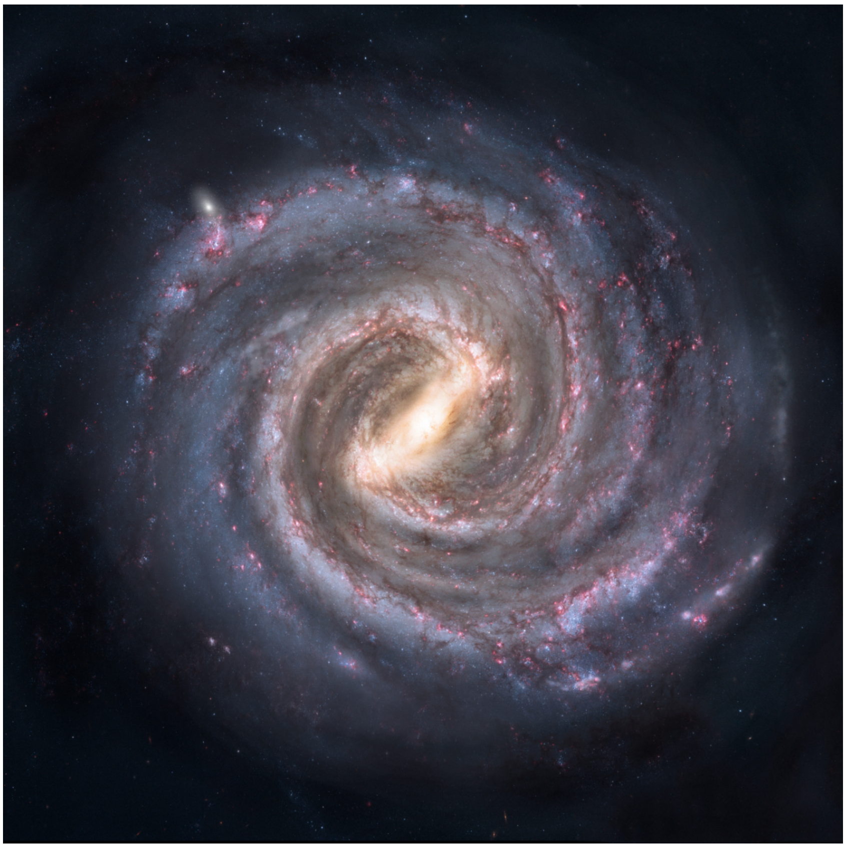

Fig.15 Galaxy (Source: Hubble)

Where, the concept of the cooling of the universe leading to creation of the atom becomes to be meaningless, as the creation of plasma and atoms is totally due to reduction in the mafs strength of the principal mafs to matter mafs strength and has no connection with heat or cooling inside or in any position in the plasma of the universe.

Where, the heat in different points in the plasma of the universe has only have the effect once the plasma of matter is created, then the environmental temperature in any point in the universe ***only dictates if the plasma in matter state will manifest itself as gas, liquid or solid.***

Then in fact this brings a question mark on the theory of time and space as then time will have no meaning as a new plasmas of atom can be created in the far corners of the universe from high speed mafs of the principal mafs which left the pulm, where some mafs go through the mafs strength reduction before arriving in the same point in the plasma of universe.

Thus; the theory that the whole parts of the universe are created about the same time and the more ancient galaxies are in the centre of the universe in some extent ***is not correct***. As matters in the universe move as their Magravs positioning dictates and hence some components of the universe move in such a rapid way which are not visible to the present technology available to the man.

CHAPTER 3

The method of transportation of mafs in different mafs strength in the universe

In all dimensions of the universal mafs strength, these be it principal matter, transition matter or matter environment, for the plasma of these matters to travel with the maximum speed and minimum loss of mafs to their environment, in all dimension of the universal mafs levels, the spherical shape plasma of the matter transforms itself to elongated plasma of itself or what is commonly is called as a ray of light (Fig. 16) (Book 2).

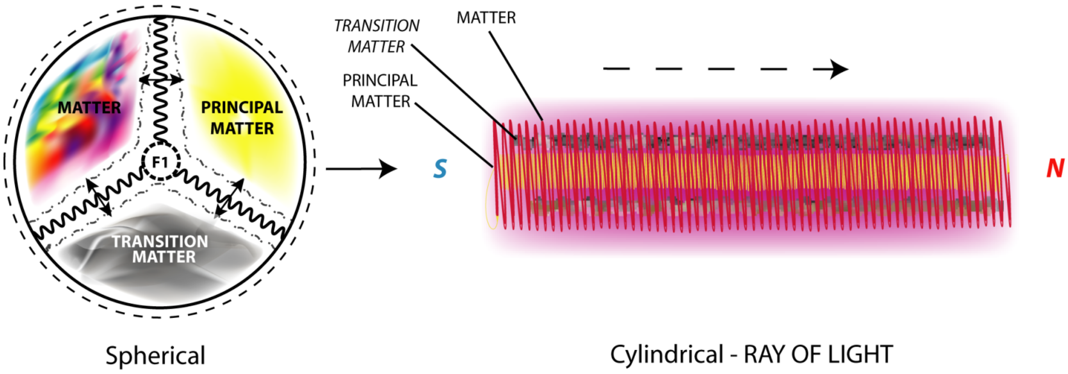

Fig. 16. Conversion of the spherical plasma of a neutron in a cylindrical ray of light.

That is to say for example as the principal light goes through the Transition fields regions of the universe, then the transition mafs and matters of the ray will come to take their position ***on the outer layer of the ray***, then through this process less pmf are used in interaction with these fields and matters of their environment and hence less friction and less loses of mafs of the ray, and in these interactions there are very little loses through the Magravs of the mafs of the transition mafs of the light.

Then any faint lights produced in the interaction of these rays outer mafs transition matter and of the same transition matters in the universe, where there are very few interfaces of the collision between the ray and the environment Magravs this gives the universe a dark but translucent effect that one can see through the universe.

In a way this interaction of rays transition mafs and universe’s transition mafs is the cause and the reason for the invisible transparent structure which allows the production of universes discreet light that one can see through this universe but blocks the opportunity that one cannot see the lights from the universes outside this universe.

The most important point to understand is that light exists in all mafs strength levels in the universe, that is to say *there are lights in the principal mafs strength* as much *as light in the transition mafs strength* and *as man observe there is light in the matter world of the man*.

It has to be point out, that as one hears sound boom in the matter world when one reaches the speed of sound in the matter environment, or what is called the sound barrier, the same applies when the systems developed by man reach’s and passes the ultimate speed of the light in the matter environment too.

That is to say when rays go from lower level strength to higher level strength mafs, one will hear audibly and will feel the vibration of the passing the light barrier.

This sound will not be a one of boom as when the sound barrier is broken, but due to the spectrum of the light transition forces, this noise will be more like a continues varying and long boom and roars. Therefore one will hear a long term tunes and not just one of noise.

This applies the same as one passes the transition mafs environment into principal mafs strength too.

Thus light speed barrier will always be the speed of the light according to man's mafs of his amino-acid and the visible light in all mafs strengths, which are the universes' creature's structures' mafs strength dependent.

When, the man become habitual to travel with the speeds faster than the mafs of matter strength, then on frequent bases man will hear a roar of a noise as passing the light barrier limit.

There are lights in all levels of the matter and mafs strength in the universe, that is to say there are lights in the principal mafs strength specific to that mafs strength environment, which is not necessarily the same as the white light as one observes in the man's matter world and at the same time each light has its own speed which is dependent on the mafs strength of its environment.

Therefore there are lights in the transition mafs strength as much as there are in the man's matter levels.

There are rainbows (Fig. 17) in all levels of the mafs strength in the universe as there are in the matter mafs strength of the earth. The light in all mafs can diffract according to the Magravs strength of the mafs of the given environment.

Fig. 17 The colour of rainbow according to the Man's pmtics strength in Earth Magravs strength.

CHAPTER 4

How matters are created in the universe

In the spans of the solar systems, galaxies and cosmos as the **gans** and **atom of the hydrogen** interact, ***they create the first seeds of creation of more powerful Magravs*** that leads to ***attraction and conversion of other light rays*** *in the universe* into gas, liquid and solids, which these interactions leads to creation of galaxies, stars and planets, where the light itself in different strength becomes the tools of the division elements the like of the hydrogen for it to start the cycle of creation of lives and the creation of intelligence in all corners of the universe (6).

The ***gas and gans of hydrogen*** becomes the dynamic fuel source of creation of the magnetic fields necessary for the creation and establishment of the stars and solar systems and their planets in some parts of this universe, as the interaction of different magnetic fields of gases and ganses lead to the creation of the initial Magnetic and the gravitational fields needed for these celestial bodies to be created in the universe.

The existence of nano materials in the core of the planets like earth has been initially disclosed (1), now it can fully be understood that how these nano materials in state of nano matter, gas and gans become to exist in the centre of these celestial entities and be the initial seeds of their creation once the mafs of the principal matter is reduced to the mafs of the matter strength.

It is important to realise that due to internal structure of the nano materials in ***layers*** and in ***gans state***, these materials by large keep their structural integrity, where high external environmental matter conditions like pressure and temperatures cannot affect the total structural integrity once they are in individual matter gans compound state. This meaning that internal structure and molecular binding of gans and nano materials in nano layers or composite structures ***will not be affected by the matter environment's the likes of temperature and pressure condition of the environment***. But the gans will be changing in its structure when conditioned to high-level dosage of radiation. Thus such entities have monoid characteristics.

For example, this is one of the reasons, why the DNA of the human being is venerable to changes and absorption of radioactive energy and in the radioactive environment or to burst of radiation, as the real structure of the amino-acid of ***the protein of the DNA chain is made of gans of matters*** and not solid matter of the same gases (6).

Thus, gans is the state of the structure of the light which has changed from dynamic cylindrical plasma of magnetic fields to a dynamic spherical entity of the same matter mafs, which its state of manifestation in matter environment is internal gravitational dependent and not environmental condition dependent. That is why one can observe the same element in coldest and hottest regions of the universe and still misreads them as atomic matter rather than nano-atomic or gans of the same matter.

Thus as mafs strength of the universe changes according to mafs available in any given part of the universe, this allowing the production or appearance of plasma of any matter which the mafs strength allows it become in that point in the universe.

Thus this why in one part of the universe one observes the hydrogen as in some parts helium or even gold can be observed without actual existence of any hydrogen atom any other matters.
In some part of the universe the matter take up their nano crystal structure, that the same matters looks to be the diamond structure of its parallel matter in another environment in the same galaxy or universe's environment.

Thus matters are created and to be part of the structure of star and planets or being in the universe according to the mafs strength of the environment and then they become appear in state of gas or solid according to the pressures and temperatures of that same given environment.

CHAPTER 5

The theory of decay and division of the universe(s)

In reality all theories which have been put forward up to present time have no parallel in the universe as has been observed by scientists.

In the universe everything and every process has parallel. In the universe there are always parallels due to the magnetic fields laws of interaction, and hence there cannot be any divisions from these laws.

At the present the scientific world has been intrigued and encouraged to understand the phenomenon of the creation of this universe and scientists have to unravel the mysteries of the creation of this visible universe.

In fact the creation of this universe as is visible to the world of science cannot be any different than any other process in the world of creation, and this universe has been created through the same principles as the rest of the magnetic plasmatic nuclear and atomic additions and divisions of entities in the universe.

This meaning that the present universe is created though the same method as has been observes in the division of the neutron's plasma into plasma's of electron and proton or smaller and/or bigger components of the same Matters.

Hence to follow the principle of the universal order of matters, this universe is created from the division of a larger universe that the man's visible universe is or can be the smaller part of the division.

The reason for it being the smaller part is that in the present cosmological term, this universe is following the attraction of the bigger part and hence it is always in the motion to follow the Magravs of the bigger sister .

This being the reason why it looks as the universe is always expanding as the external field forces pull and pushes the boundary of this universe mafs and as part of the universe has been noted to be collectively following different path of motion than the rest of the universe.

The portion of the independent motion of this universe is the indicator that where the bigger portion of the division of the pervious universe is in respect to the man's universe.

It has to be said that the original universe through the same principle of decay of neutron to plasma of electron and proton (1) and similarly as the decay of heavier atoms to two or more lighter atoms, through the same principle as the mother universe of the man's universe as itself has been created and was created from the division of a larger universe through the same method of decay and division.

The difference in the universe creation and decay is in principle merely is the size of the plasmatic field content in comparison to plasma of electron or proton after nuclear decay and division of the neutron.

Thus in fact this present universe itself is and was created some billion of years back from the division of a larger universe (Fig. 18). Where in Fig. 18 A to G, step by step the phases of the unbalance of the Mother Universe, it's opening and as the fields disperse in to two different zones in the unicos, where these fields interactions and Magravs positioning with Magravs of the Mother universe leads to the establishment of position and placement of two new universe entities, where the same process of decay of neutron and creation of Hydrogen atom occurs.

Hence every time as it is the same with division of the plasma of neutron to plasma of electron and proton, in this division the new universe carries the characteristics of its mother's universe, but due to reorganisation of magnetic fields and their strength the new universe will have a new and fresh cycle of time, which this shows as the age of the new universe.

In fact this being the like of giving birth to a child, where the age of the mother is irrelevant once the new seed of life is created and once the child is born, even though; the new child carries some of the same DNA fingerprint of the parents. The new life has its own cycle of life and its own timing.

Therefore the theories of Big Bang and M-11 as has been proposed in the past and present have no meaning and have no parallel principles in the physical world of creation of universe. One does not see the same Big Bang , when the atoms and neutron go through division or decay, hence this cannot be the same in the division of universes too.

Then if this was so, then every time an electron is created from the division of neutron there must be a Big Bang , This proposed principle is nothing, but showing the man's and scientific worlds lack of knowledge and ignorance of the real terms and parallel entities existence of all units in the universe.

In fact once the division is created and the universes are created, then the rest of the materials and matter content and how they organise themselves in that universe is more or less follows the same principle as before and once the fields are released within a given Magravs which the fields can created in interaction with each other, then the size of galaxies, stars and planets and even creatures within that given Margraves strength of the newly produced universe are determined.

The mafs and matters as arrive directly into the new universe are mafs first which can change to matter then divide between the two universes and then new matters are created in the new universes independently according to the Magnetic fields and gravitational fields in interaction with each other these mafs can achieve to maintain and create. Thus all universes in content are not and cannot be the same in material and mafs strength, and content of each universe are dependent upon the internal interaction of fields which are transferred into the new universes.

The new structure is dependent on the fields within the new universe and their internal motion of the new universe is internally determined by the total mass which the interaction of the Magnetic fields and gravitational fields of the universe can master.

This new universe has all the attributes and mafs of its mother's universe, where the subsequent production of mafs and matters like Principal matter and Principal Matter, planets, solar systems and galaxies and so forth, are nothing but the consequential outcome of all other effects in a given universe.

Thus in other universes there could be ***no physical matter*** the like of hydrogen and where for example hydrogen can be *only in nano or gans state only*, thus the life in these universes will be in order of fields and mafs strength rather than matters in this visible section of this Mans' universe.

But if this universe has the same structure as its mother's universe and co-birth universe, then this universes can be a micro replication of a bigger version, where these universes are macro version of the total universe in the greater universe.

Phases of the decay of the plasma of a Universe

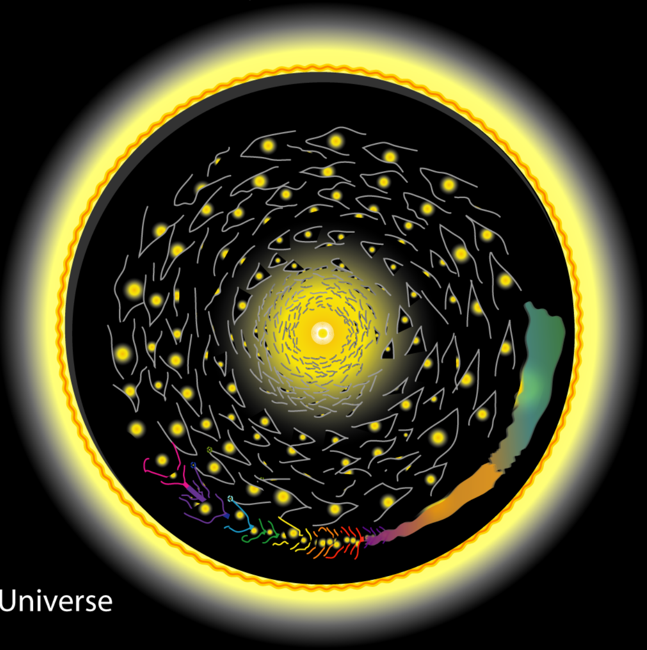

A. Mother Universe

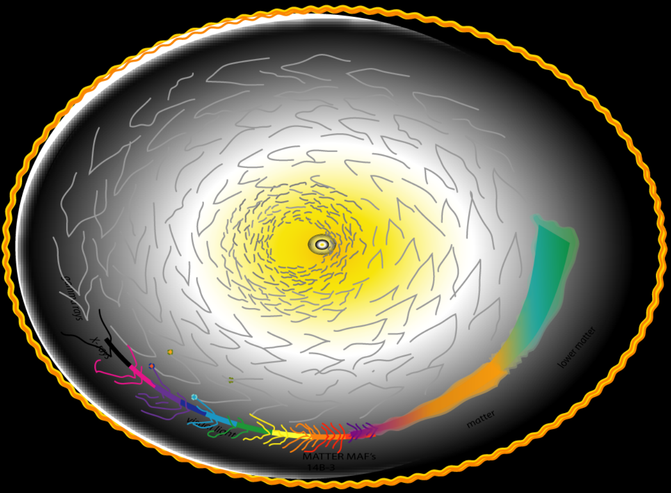
gamma rays
X-rays
visual light
MATTER MAF's
14B-3
matter
lower matter

B. Initial Magravs unbalancing

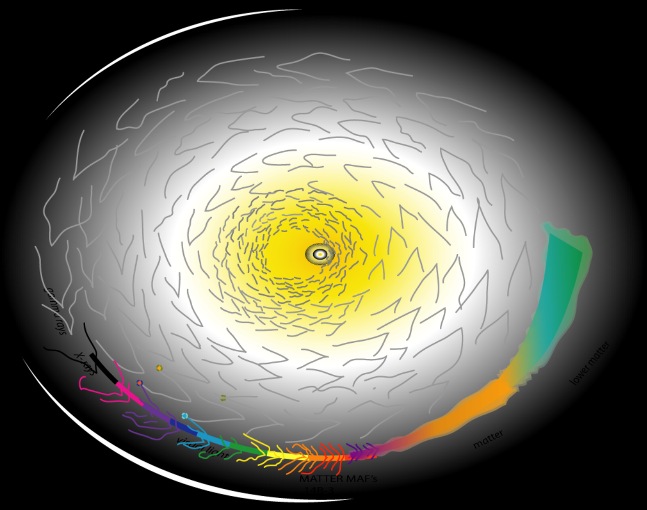
gamma rays
X-rays
MATTER MAF's
matter
lower matter

C. Loss of structure

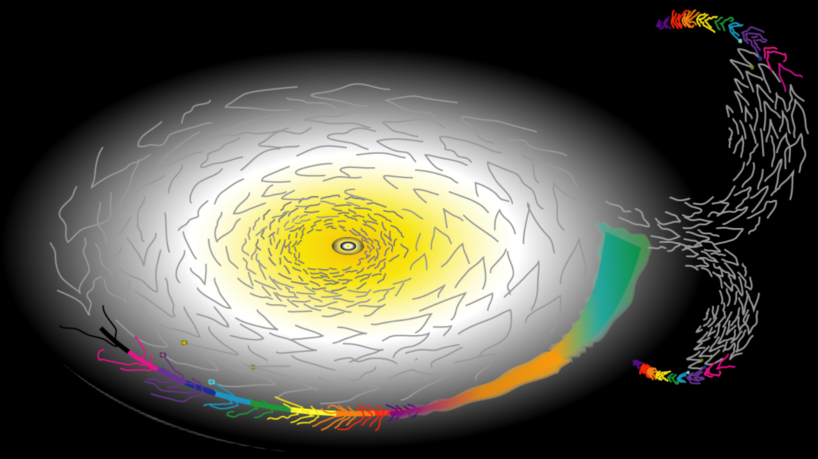

D. Dispersion of fields of the plasma of the Mother Universe in the unicos.

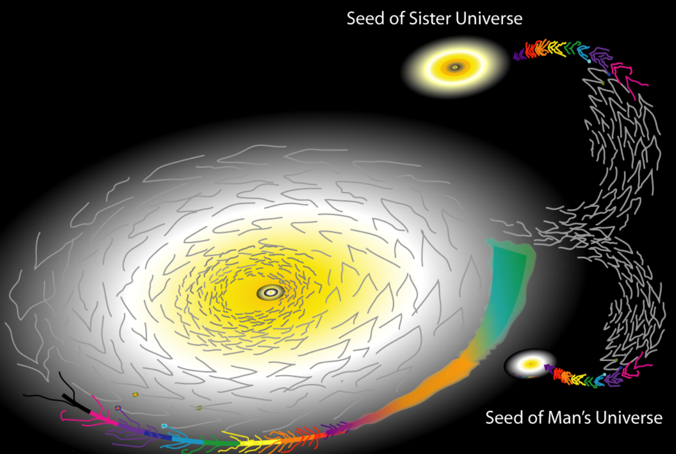
Seed of Sister Universe
Seed of Man's Universe

E. The creation of the seeds of the new universes from the decay of the Mother Universe, where the two new universes have to maintain a magravs positioning in respect to the Mother's and each others magravs.

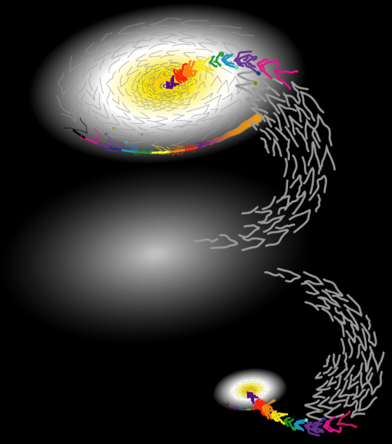

E. Full splitting of the Mother's Universe plasma into two new plasma's

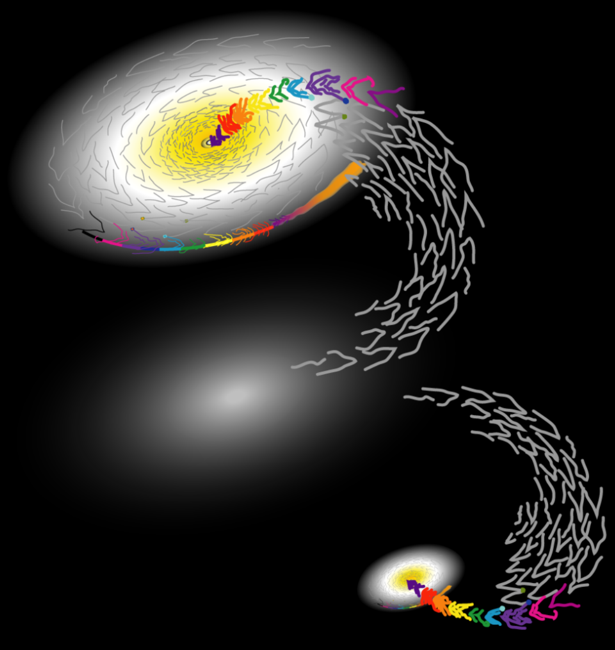

F. Initial stage of creation of gravity and magneto-sphere of the two universes due to unwinding and interaction of the fields

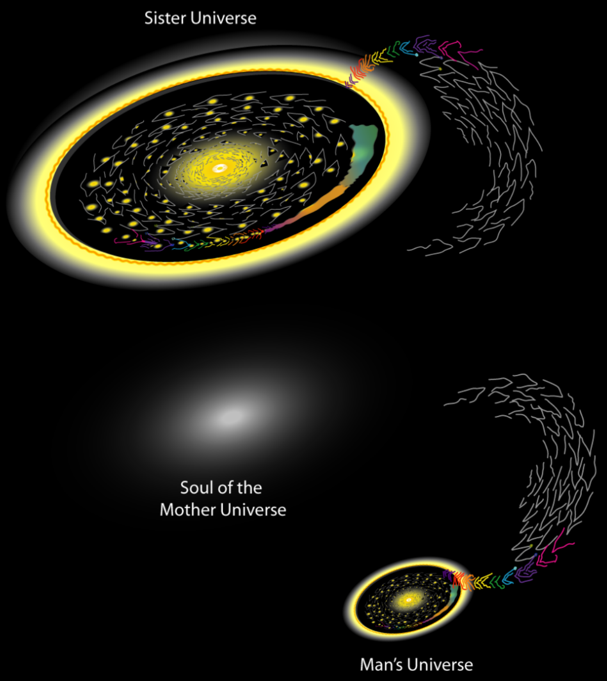
Sister Universe
Soul of the Mother Universe
Man's Universe

G. The final physical parameters of the Mother Universe, the Sister Universe and the Man's Universe. Where the same process applies to decay of neutron and it's division in proton and electron, where Man's Universe is the like of the electron after the division.

Fig. 18 (A to G). The process of the division of the Mother's Universe in to Man's Universe and it's Sister Universe

Thus, in fact in the bigger portion of the original mother universe which the man' universe is partly created out of it is bigger or smaller mass portion of the original mafs of the mother fields, then the bigger portion or co-birth universe of this universe, due to the larger mass of its content, this universe's galaxies and stars can be much larger in proportion in comparison to the galaxies in mans universe and conversely due to larger Magravs of the internal structure of the larger portions, the beings within these universes can be smaller or they be totally different in operation than the beings in the mans universe.

It has to be said that life in any form is not exclusivity of the Earth, if this is the mind-set of some, this shows their lack of understanding of universe and how it is created and not the real truth and how fields and matters operated and function in the universe.

CHAPTER 6

The Man and his soul and their connection to the universe

In the structure of the Principal mafs centres or the Principal star or what is so called the Principal stars in the centre of galaxies and mafs and matters in the centre of stars and in giant planets and even in the centre of some planets, the Magravs of these objects seems to possess a stronger or larger than the mass ratio for such an object, that the object possess and covers a stronger Magravs forces than usually a matter in these regions by the present in the world of science knowledge become apparent. The reason for the stronger Magravs of these objects comes from the multiplicity in the Nano magnetic mafs' and matters' and layers in the inner structure layers in the sanctum of these objects.

In understanding the operation of different mafs of matters in the universe and the process of the reduction in strength of these mafs to reach the lower levels light, visible light and tangible structure's like atoms, then one should understand that the structure of the man's body will not change as the physical structure of the man goes through and from matter to transition matter or from transition matter to principle mafs and matters strength's environments.

The reality about this, is that the plasmas of the atom of the structure of the man, like the individual atoms, which are active in the structure of fundamental amino-acid molecule of the human body, these are all made of plasma of mafs of universe and for the proton or the electron of let say carbon atom in the amino-acid, this plasma of the proton has in its centre principal mafs and maters as well the transition mafs and matters and finally the matter mafs, which is its tangibility state which make matter visible to for example the eye of the human.

But once these plasmas of the tangible matter structure of let say proton and electrons of the carbon atom cross the barrier from the light mafs strength and matter fields to transition mafs and matters field strength, or even the body of the man crosses the principal mafs and matter strength, then at each point of field strength crossing, that part and components transition or principal mafs and matters of the plasma of the proton of the carbon will come to operate and interact in that entered field strength zone, and hence the integrity of the structure of the human being physical body will be maintained, but indifferent mafs strength level.

This meaning that the man will be still intelligent and will have the use of all its faculties, but the body will operate through the principal mafs of its plasma rather than present matters mafs strength of it material world strength.

Thus the man is not restricted to life in matter world only, but life in all dimension of strength of the mafs of the universe.

The reason for the plasma of the human body having the ability to operate in any mafs strength, is due to the fact that the atoms of each cell of the human body operate in the gans state of matter and not solid state of matter, which this gives the plasma of all the elements of the human body to cross mafs strength barriers by calling up on their mafs components of their environment.

That is to say the gans of the carbon in the amino-acid of the human beings' will bring forth and operate through it transition mafs strength in the transition mafs environment. Then the same gans of atom will operate with its principal mafs components in the principal mafs strength environment.

Thus the body of human will not be destroyed or evaporate in different plasmatic magnetic fields strength environments as it orderly crosses them.

But the body of the man will be destroyed if the mafs of principal mafs strength is applied to it in the matter strength environment, where in this case the mafs and the matters of the atoms of the man become diluted as a minor part in the mafs and matters of the applied fields.

Therefore man is part of an inclusive pattern of field's in motions in the universe and nothing more or less can be attached to the existence of the man.

The man in physical term is an entity and in and through his soul the connection of the field between himself and the other fields in the universe.

The physical body of the man is the like of the planet earth, where it has mass and its magnetosphere which expands beyond the physical body, but its total magnetic fields and gravitational fields determines its position and relation with other matters in its solar system and in a small way it determines the position of its star in its galaxy and, then the position of star determines in a small way the position of the milky-way in the universe and so forth.

Thus the soul of the man is the external connection of the operation of the physical body of the man in the universe, it is part of one and inclusive realty of interaction of the whole.

Therefore the soul is part of the operation of the universe and every creation which can extent the effects of its overall magnetic field to others which can affect their position and existence is in possession of a soul.

Thus soul which magnetic fields created and expand beyond the physical body of the man is not exclusivity of the creature but it is the common denominator and common link between the man and other creatures of the God and has direct effect on the operation of the rest of the universe.

Therefore the man's body due to the interaction of its atoms protons and electron and neutrons magnetic fields, this object creates its own magnetosphere and its own mass and this body is and has to be in possession of magnetosphere which shows the working of the magnetic fields of its constituent matters and its subcomponents and fields.

Where, in the physical worlds as the same as the solar systems possessing magnetosphere as so do the star and planet, so does the body of the human through the same physical principle possess the magnetosphere, which is so called aurora of the human body, the field of the aurora is part of the soul of the man, but is fact the total effect of the internal interaction of all fields of the body which creates the Magnetic field and gravitational fields of the body and their interaction on each other which leads to the interaction with their fields in the universe, this overall output of the fields of the body of the man is the soul of the man.

Thus; all dynamic plasmatic matters in the universe possesses a soul. But the man has the intelligence to allow the interaction of its soul with others. Therefore through this principle the man is responsible for all his actions and he stands answerable to other souls in the universe. As he decides how these fields which partial and fully control through the Magravs of the matters of his brain how and what happens in its interaction with other souls during its physical existence.

Hence the man has to understand, that universe and others, whom this universe has been created from and the universes which will be created from the division of this universe in the future are all interconnected and operate through the same fundamental principle magnetic fields.

Therefore there are parallel universes to this universe, with the difference in the measure of their magnetic fields and their position in their given environment.

Thus as we have the same and similar electrons structure's and protons structures in the same structure of an atom, this structure of positioning of electrons and protons is the same for all known atoms, then the same applies to the universe and the souls within them too. This meaning that there are similar universe as this universe in the greater universe and other universes which this universe has been created from or it shall create.

It has to be said that at the same time the matters in the universe have the same universal language of communication, where in the universe the common language is the basic plasmatic magnetic fields, based on the first plasmatic magnetic field strength of plasma or the simplest atom of matter of their environment.

This language is universal and once the man on earth understands this phenomenon and can fully measure the magnetic fields strength of the electron Magravs and uses this as the base strength, then man will be able to understand and communicate with rest of soul possessing creature of the universe. Until then, the search in the sound waves to look and try to communicate with other creatures of the universe is a primitive picture and shows the lack of mans understanding of the world of the Creation.

Therefore the gravitation, magnetic, radiation and other similar properties and interactions of the field(s) on each other in this and other universe(s) decides on how and the way the creatures in the different universes and in the man's universe would look and behave.

But in a essences they are all made of the same building block of *MAGNETIC FIELDS AND IN ITS AND THEIR INTERACTION OF DIFFERENT STRENGTH AND TIGHTNESS OF IT IN THEIR ENVIRONMENT*; This determining how tall and short and if one has eye or uses the interaction of fields to interact and communicates with other creatures of its environment, universe(s) with or without physical tangible or visible so called body.

May be now it can be understood what has been said that in the universe,

I have made man in the image of myself

CHAPTER 7

The creation of gravity in the universe

At the present time the scientific world uses the terminology that when the gravity took hold then that was time when the matters in the universe began to be created.

The reality is that as when the universe is created from the magnetic fields of the previous universe(s), then the content of the new universe is nothing but a soup of different strength plasmatic magnetic fields, then as these magnetic fields interact with each other usually laws of magnetism they attract or repel each other, where at this point the attraction fields of these plasmatic magnetic fields become the ***gravitational fields*** forces of the universe and the repulsion field forces of the magnetic fields become the Magnetic fields forces which the same fields create.

This interaction of fields in the plasmatic fields always enclose itself in the form of an envelope, which this always in the universe leads to creation of the spherical shapes of all entitles, this be it plasma of a neutron, earth, or universe?

The interaction of field and attraction fields which leads to creation of a circular motion of the gravitational fields as the fields in respect t the matter have keep their gravitational positioning too.
Thus the spherical shape of the fields is primarily due to the gravitational field and positioning combined together which this causes the magnetic fields in the universe to close and created a full circular; and hence the creation of spherical objects in the universe.

Therefore the gravitational filed forces of the new universe is created by the interaction of its content magnetic fields and in fact gravity is created by the natural laws due to interaction of two opposite polarities of two magnetic fields, which as these opposite fields interact they create attraction fields or what commonly called the gravitational field, where the strength the outcome field from this interaction of fields expands beyond the boundaries of the two original fields, this leading to the creation of an envelope or area of the effect of this new field that attracts other fields to itself and the strength and compactness of this new field leads to creation of gravity which in the universe this leads to creation of Principal matters, transition matters and then visible matter of galaxies , stars and planet and so forth.

The creation of gravitational fields forces in the for example the plasma of neutron, its strength and contest by large measure is dictated by the universal gravitational fields forces and not in full by the field which leads to creation of the neutron to start with. In the universe the creation of the neutrons are dependent on their universal gravitational fields forces. To this effect not all neutrons are all the same and not all neutrons in the different universes have the same gravitation a magnetic fields forces and dimension.

In fact the human race has to learn that what it is observed by them as the physical matter in fact is only a fraction of overall content of matters of the universe created through creation of gravitational forces and what he sees are not all of its matters.

This meaning that the universe has other matters and mafs contents and as the man's technological and intellectual capabilities expands and increased then man will come to see more of the realities of the universe.

CHAPTER 8

Insight into the realities of the working of the Universe and what they bring about

In the recent past scientists have talk about the term dark energy and anti-matter, where these covers the part of knowledge which they cannot find explanation for within the present understanding in the scientific world. For example the term dark energy is used to understand the forces, which is nor explainable and neither does fall within the conventional present world of physics too.

In this disclosure the terms like the matter, transition matter and Principal matter are used, where these terms are and have become recent addition to man's scientific vocabulary, where these can more precisely explain the rang and the medium, where each magnetic field in respect to their fields and its own magnetic fields strength is and would behave.

In the universe the same as in the physical matter world there are sources of magnetic fields, which are stronger in magnetic fields strength and some which are in the transition and there are some which are in between the strong and the weaker field strength.

In the physical world one called the stronger fields source the star, where in the plasmatic fields of the universe, the stronger fields' plasmatic regions or fields are easier to be called the **Principal fields**, which their interaction leads to creation so called **principal matters** (anti-matter), which these are real physical entities, which have been unknown to man up the recent times.

There are **transition fields,** which are in between the Principal matter fields in strength and the matter field's strength, and as these are in balance in between the both fields of Principal matter and matter world, and then these do not create visible light when they interact with Principal matter or matter fields strengths. Hence they tend to be like ***transparent*** and ***translucent*** in their environment and hence they give the appearance of the darkness to their environment.

When, magnetic fields are in motion they are called energies (Book1), where these translucent fields as they are in motion from one state of mafs to another they create what is called the Transition energy (commonly called the dark energy) which this explains the Transition energy mystery of today's science.

Then there are in the universe fields, which are the weaker plasmatic fields region in strength, which these fields are in the human protein plasmatic Magravs strength that they become detectable or what is called the visible or tangible matter.

Therefore in simple term, the Principal matter is the like of the sun and sources of energy in the universe, the transition matter is the magnetic fields strength fields travelling from the stronger field to weaker field, and the matter fields are the like of the planets of the solar system.

There are no mysteries about these fields, energies and matters, except in the lack of knowledge in the man's progressive and advancing science about his universe and living environment.

- Where, in the universe one sees the effect of the stronger Principal matter's fields as in the centre of the galaxies as the star do in the solar systems. The present world of science will come to realise the order of fields are all time kept and all followed the same rules and principles in all levels of the universal creation and created.

- This meaning that in the centre of the galaxies the entity which recently called as the Principal star of the galaxies, this entity in its centre have nothing but Principal matter magnetic fields, which their interaction leads to creation of strong gravitational fields. Where, this centre of Principal matter is covered by the transition fields as they are on their way to visible matter fields strength or what one observes as the dark environment around the central Principal matter region, hence the term black hole (Fig. 19) or Principal star, which this darkness region is only due to the balance transition between fields in motion and the fields which the rays are travelling through.

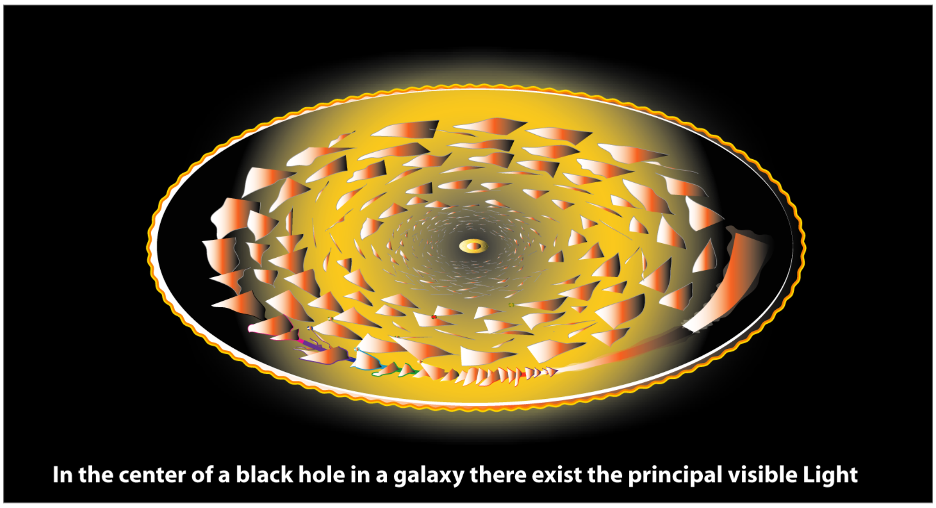
In the center of a black hole in a galaxy there exist the principal visible Light

Fig. 19 The Principal Light inside a black hole.

- The existence of the principle fields and matters in the centre of the universe with its strong gravitational fields is the reason as it seems that the Principal star is eating everything and absorbing everything this being rays or matters, where in fact it is the gravitational field forces of the Principal matter in the centre of this dark spot which are creating these large pulling fields.

- The size of the Principal star in the centre of the pulm is the combination of gravitational fields and Magnetic fields forces or the field's strength of the Magravs of the Principal matter content of the Principal star.

- In a sense as the Principal star extracts more material from its environment in the universe, then a position is reached that the total Principal matter absorbed by the Principal star from the objects, the like of stars, planets, plasmas and so forth, a point of criticality is reached that the gravity of the Principal matter fields exceeds the Magnetic fields of the Principal matter and that these forces cannot sustain their balance, hence the Principal star has to release enough material and fields that it can reach equilibrium between its two fields, and that is the point where the principal stars release vast amount of fields and matters without actually causing the division of the universe (Fig. 20).

- Where, these massive fields and matters of massive principle fields released will disturb the equilibrium of the universe as they travel the spans of the universe and in case can lead to creation of new galaxies and even stars and new mattes suddenly out of nowhere in the universe. These principal storms are very much like the sun storm one observes on the surface of the sun. In this process of rebalancing, it can observed in the universe, like as in galaxies that their Principal star has sodden volcanic eruptions in the rest of universal galaxies central principal stars as they absorb some of the principle matters and fields which they receive through the universes principal field eruption and the Principal star spewing of the surplus mafs and matters in to its surrounding environment.

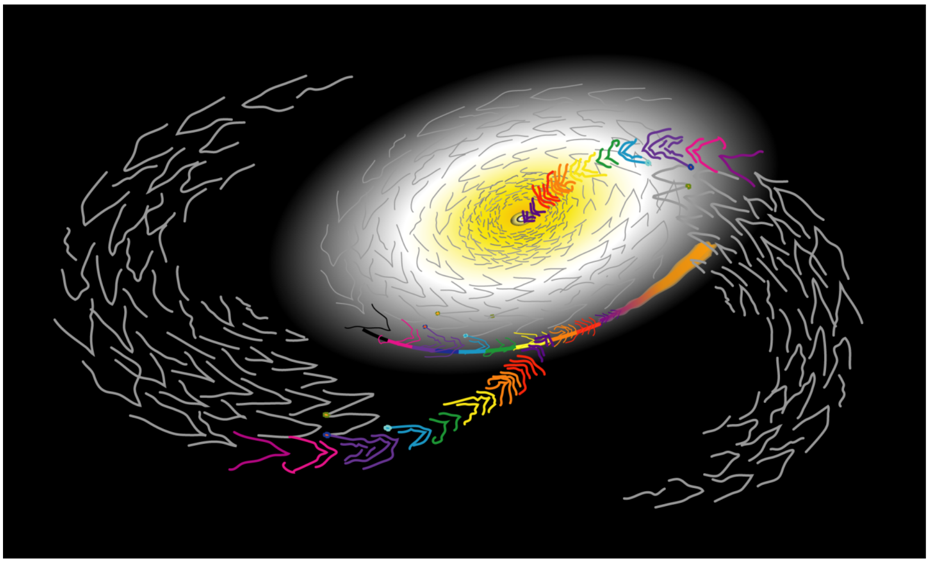

Fig. 20 The absorption of matters and fields by principal stars of the universe or galaxies.

- The principal star of the galaxies as much as the same with the principal mafs and matters of the stars of the solar systems or their sun have observed this process in smaller scale. Where up to now man has not have had the tolls to observe that when the star of its solar system not only release the visible matters and releases huge amount of magnetic field into its heliosphere, but at the same time through and amongst these eruption of the Sun huge amount of principle fields and matters are released into the plasma of the solar system, which these have not be observed by the scientific world due to the lack of appropriate tools.

- The ejection of principal mafs and matters from the surface and right from the centre of the sun is an essential part of sustainability of the fields and matters within the solar system and essential part of the creation of new atoms and matters on the surface of the planets of the solar system. Where at the same time the solar systems as the whole on continues bases receives large amount of principal mafs and matters from principle star of its galaxy and principal stars disc in the pulm centre of the universe.

- It will observed that Principal star in the central disc of the universe even absorb principal light, in fact as light is an entity which possesses both gravity and Magnetic fields and its own mass, thus Principal star is extracting another object from its environment due to its stronger fields strength (2) is a normal occurrence. (Fig. 21)

- Where, the same has been assumed in the present science that black holes even absorb light, but the truth is that principle stars be it in the centre of the universe, or galaxies are created from the principle mafs, which are of the higher order of magnitude in mafs strength and the mafs than of the light mafs in the matter mafs strength. Thus in fact it is normal for weaker field’s strength to be attracted by much larger and stronger mafs of Magravs of the principal stars in the centre of the universe or galaxies.

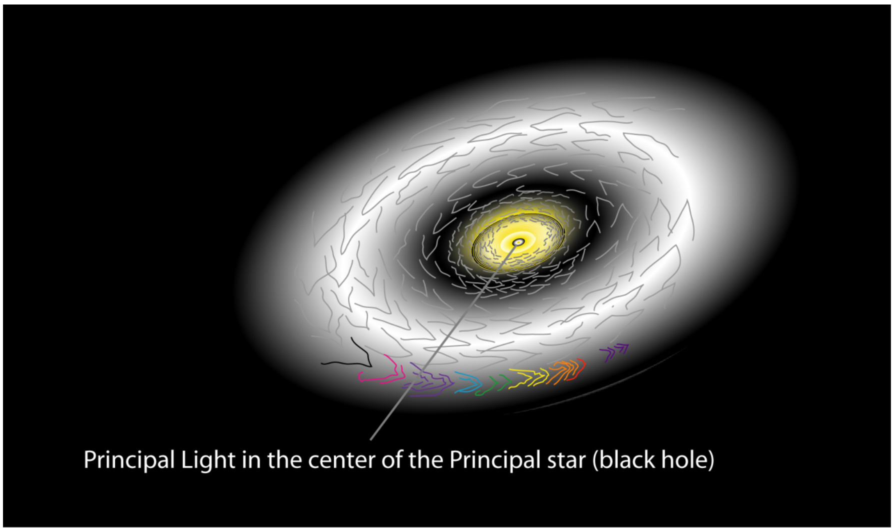
Principal Light in the center of the Principal star (black hole)

Fig. 21 Principal Light in the centre of the universe.

- But at the same time it is normal to see that the principal stars in the centre of the universe and galaxies spew stronger lights in the order of magnitude of principal mafs which are not visible to human eyes, where the light in the matter mafs strength is visible to him, hence man see one way traffic and not the real truth about the world of creation of duality and equilibrium.

- Thus in the universe it should be possible to observe points where the light lenses the Principal star if it is far enough that Magravs of the Principal star cannot overwhelm the Magravs of the light, but strong enough that Magravs of the light find their position in respect to the Magravs of the Principal star and creates their own Magravs positioning. As this can be observed in respect to the light and star of a system so as the star is made of mafs of principal mafs and matters too and has the same effect of the lensing for the lights rays too.

- Lensing of the transition light when passing the principal mafs regions of the universe should be observed the same way in the future as one observes the lensing in the matter light mafs strength. Where even the lensing of principal light should be observable in the specific cases and some regions in the universe.

- The matter, Transition Matter, Principal stars, the dark spots on the surface of the sun all are created through the same concept and principle that when plasmatic magnetic fields of the any field or matter is equal to the plasmatic magnetic fields of its environment, then there are no collisions between the matters fields which can leave or release any surplus or residual magnetic fields in the light spectrum of the mafs strength or lead to creation of magnetic wave in the visible light balance of the human protein pmf strength or that specific mafs spectrum anywhere in the universe. Thus these areas in the universe, the galaxies and stars surface looks as creating not too much light or what is called to be looking as to be darker than the rest of their surrounding environment(1).

- Thus the assumption that even light cannot escape the gravitational fields of the Principal star is position and the lights itself Magravs strength dependent.

- In truth if and when the Principal matter magnetic fields over time are reduces in strength and they unwind and transfer their fields to their environment, which as these fields slowdown, they become transition matter energies and then transition matter and finally they slow enough that they reach the plasmatic magnetic fields of the mafs strength, which at this point they become energies observable and absorbable and tangible to human field touch and visible to it virtual magnetic fields sensors.

- It has to be said that there is a great misunderstanding in the operation of the fields and mafs in the universe, this misunderstanding over time has reason due to the man's concept of heat in the *MATTER WORLD, WHICH THIS NEEDS SOME CORRECTION THAT CAN BE APPLICABLE IN THE UNIVERSE* as whole (The full explanation of the magnetic fields, plasmatic magnetic fields, matter and mafs is explained in the reference 1).

- The universe will never run out of energy and get cold, the universe will not expand to nothingness, but universe will decay to smaller universes as universal gravitation will never disappears as long as there are two or more magnetic fields remain in the universe that they can interact and attract and repel each other that they can create gravitational and Magnetic field forces that the whole process can restart again.

- In the nuclear decay one does not observe temperature increases and one sees uniform temperature gradients across the newly divided plasmas and the temperature distribution will be constant, this is the same with division of mother universes to smaller universes. Thus the concept of universe was hot and as cooled down matters are created is fallacy and not in parallel with the rest of the universal laws and processes.

- This observation is parallel with the W-MAP, which has been planned out by present scientific world, where one does not observe huge heated areas and points in the universal map. Here the equal heat distributions parallel with the nuclear decay and heat balance in the divided components of electron and proton after their division and long term motion in their environment is observed.

- The individual heat spots observed across the universe, these points are due to newly arrived mafs and matter movements in that given region of the universe and they are temporally condition.

- The central heat line across the W-MAP is due to dynamism of the principal mafs and matter in the centre of the universe and this is due to concentration and dynamism of the mafs and matters and their reduction to transition energies within this region.

- The creation of gravity in the beginning of time of the universe(s) was and will be primarily due to the interaction of combination of different plasmas of mafs within the plasma of each universe which have arrived from the mother universe and as the magnetic fields interacted, they create gravitational fields and Magnetic fields of the new universe, which these in turn as there are made of many mafs strength, which have arrived in the new universe from the division of the old universe, then different gravitational fields and magnetic fields due to their interaction are created, this leading to creation of different galaxies and matters. Thus, the universe never been cooled but pmfs from one strength meets-up with another and leads to creation of central gravitational fields of the universe and these interactions in the future will lead to creation of specific new matters in that universe, which these matters existence and appearance are closely related to the gravitational fields which the pulm mafs and matters have stared and maintain during the life of that universe and will allow to be created and be held in that universe. Thus all universes do not contain the same matter and all their basic matter is not based on the plasma of the hydrogen as is in the man's universe.

- The internal and initial gravity created in the plasma of any universe is on the same principle as the one created in the plasma of proton and electrons mafs and matters and from the interaction of their mafs of matters, transition matters and their principal matters once they arrived from the decay and division of the neutron. Thus the gravity of the universe is established from the interaction of different mafs and matters strength once the mother universe is divided and they are created in respect and relation to the original Magravs of the mother universe as these fields are separated from their mother universe into two or more universes like the one visible to the man. Therefore it can be said that man's universe is part of the multiples of universes which all have been created from one mother universe, which this present visible universe to man is a fraction and part of operation of bigger and more active universal order of the Grand universe.

- In reality one can say that this universe is nothing but like an electron in the structure of a uranium atom which this atom is part of a large mass of a large block of uranium atoms.

- This visible universe is part of a total universe and man's universe which looks so large to him is only due to his lack of understanding of the real terms of the world of creation and existence and is dependent up on his openness and understanding of levels of his own existence.

- The principal concept of creation of the principal matter and transition matter strength from mafs and matters in matter mafs strength can be achieved through the process of fusion of mafs, when these mafs and matters of lower order strength enter into the principal star regions of the universe and their fields become and are squashed and tightened in intensity of the fields within the mafs and Magravs of the principal star that the entered fields into the Magravs catchment's of the principal star merge different field strength magnetic rays and fuse these rays that they become and reach mafs equal to the strength and density of the principal mafs and matter strength.

  Thus Principal stars are, where the process of fusion of mafs and matters occurs to amalgamate and increasing the mafs strength of loose and incoming lights and matters from weaker strength mafs is set to take place. This is to say that Principal stars of different size and intensity in galaxies of the universe are the places where the principal mafs and matter are cooked up and separated and processed and manipulated into their given strength and then through conditions become to be known the like of quasars these new recycled stronger fields are released into their given environment, leading these principal mafs and matter to become the centre point for the creation of the stars and planets in their galaxies and as some become to be in the centre of the plasma of proton and electrons and so forth.

- Where, at the same time the Principal stars are the points in the universe that by and due to their high gravitational fields of their environment, this is the place in the universe, up on entry of different mafs and matters, these matters strength can be increased to higher level magnetic fields strength. Thus the Principal star are the points in the universe where the mafs and matters can increase their strength to transition matter levels or to the principal matter magnetic fields strength and where the transition mafs matters fields can be increased to the principal matter field's strength

- In the centre of the stars as the same as the in the centre of galaxy and universes, where the fusion process of fields and matters are considered to take place and heat to be generated due to fast movement of the pmfs. In reality in movement of the stronger and weaker mafs and reduction in strength of the mafs, which these changes disturbs the plasmatic magnetic fields of the Magravs of plasma of the components of atom, this be it the plasma of electron and plasma of proton in principal matter, transition matter or matter strength, this leads to release of mafs fragments, which these can manifest themselves as heat or radiation in their given environment mafs strength. This movement of plasmatic magnetic fields leads to weakening of the magnetosphere of plasma of the all atoms, this be it plasma of principal, transition or matter atoms, hence this weakening allowing the process of entry and amalgamations or fusion of different plasmatic fields into another magnetosphere of any other plasma or an atom or increase and amalgamation of lose mafs of different strength within the environment of a principal star.

- By creating and using the higher strength mafs than the gravitational and magnetic fields strength any entity this be it atom or rays, one can create conditions, where one can cause fusion of atoms, plasmas and rays without creating heat and any need for accelerating particles and in colliding them to create fusion as has become the tradition in the present plasma of matter fusion technology on earth in the recent times.

- If the full concept of fusion in matter level is to be achieved in much easier way, then it will more appropriate to separate the different components of the plasma's mafs and matters, amalgamate the same together and then allow the re-amalgamation of different mafs strength to take place. This is how fusion of field's and matters are attained the span of the universe.

- In the universe the principal mafs and matters are the sources of the energy and supplier of the pmfs of the universe, thus when the plasma of principal mafs or matters interacts or collides with the lower strength mafs in order of magnitude of any matter then mafs and magnetic field content of matter becomes part of the principal matter spectrum and their content is added to the mafs of the plasma of principal matter.

  Hence the current fallacy that when matter and principal matter interact nothing is left of the two, except some release of magnetic fields or energies, this is hypocrisy and man lack understanding of its own world and the world of the creation. This is like saying that if earth gets attracted and absorbed into the Sun of the solar system as a lot do on frequent bases in the universe, then there will be nothing left of the Sun.

  The principal mafs and matters are the source and stars of the universe and by order of magnitude and strength and the mass the principal matters and stars mafs are by far stronger and possess denser fields than their by-products the transition matter and matter fields strength which are created and released originally by the principal mafs and matters itself to begin with (1) which has led to creation of the planets and stars. Then how can the weaker and insignificant part in matter strength level destroy the larger entity?

  Thus nothing happens when matter entity comes to contact with principal matter (antimatter as is known at the present time) entity. There will be only principal matters left with addition into it the content of matters mafs and matters, which have collided with it.

- Thus the physical structure of the plasma of principal stars in the universe is as follows the principal mafs and matters in the centre, transition mafs and matters and the then mafs and matters in the matter mafs strength.

  This entity what man calls the universe follows the rest of the universe orders of linking the magnetic fields polarities of north and south poles together as has been observed in the man's planet. This universe is near spherical in shape in general and dynamic and rotational anticlockwise in direction of motion. But the principal star can pull into itself due to its strong Magravs and gravitational field's matters as well as matters to itself from its content plasma as well as from the unicos of it outside environment, where in the process it can separate their sub-Matters for these to be added to its principal matter too.

  Looked from outside this universe surly will possess rings as planet Saturn's does, as this universe has the same construction as the gaseous planets, but; with the difference that its rings will be colourful and much more dynamic than Saturn's rings.

- In universe(s) principal stars can be found which have anticlockwise motion and some, which have vertical rotation and may be some, which have combination of both clockwise and vertical rotation too. These motions are primarily matter and fields' strength and position dependent as has been seen in the man's solar system (1, 2).

In the principal mafs state of matter all matter more operate in the gans state of matter than the gas or plasma of the matter, where with the gans of the matter the mafs of the environment decided the appearance of the element in a given condition and in all cases tangibility of matter is the consequence outcome of these conditions.

Hence all matters in the universe are not matter atoms, but gans of matters, as gans of matter can take all lattice structural combination of elements, where for the matter of same element to converse from one lattice structure to another high temperatures and pressures are needed. Where, in the universe majority and nearly all conversion of lattice are done through gans of the matter and not atom of the matter.

To achieve the gans state of matter the environment of the atom of the gas needs low and continues flow of mafs in transition for the atom of the gas to be converted and its Magravs changes from magnetosphere of a gas to a magnetosphere of tangible matter of the gans (4) state of matter, and in fact this happens when the gravitational field of the gas is made more powerful than the Magnetic field of the gas and hence the plasma of the gas gravity contracts or tightens and hence the gas changes from gas to gans and becomes what is called solid nano state of matter in the atmospheric condition.

- Multi-layering of mafs and matters increases and exacerbates the initial magnetic fields strength of the matters and the fields they can create. This causing an increase in the strength of the Margraves of these objects and even Magravs strength can be further increased if the mafs and matters are in gans state of matter or in combination of different mafs strength in different layers.

- Ganses as gases once reach the state of certain structural lattice as in the gases on the earth atmospheric pressures, these ganses change their characteristics and take the properties of non-dimensional plasma or they become what is known as odder or what can be detected of their existence in their environment. This principle applies the same to the gans of the matter too, that the ganses of all elements have odder and hence the universe outside the earth atmosphere has an specific odder of its own, which in fact each galaxy has the same properties and different odder and each universe has its own odder primarily due to the structures of the gans(s) of the elements they create or they can hold on to, and the same element does not have the same odder in all corners of the universe. As odder is the mafs strength dimension dependent the gans of smelling entity mafs combination and/or of the gans's mafs strength itself.

- At the birth of the universe there has never been and in the future there will never been creation of heat, as in the same process of decay and division of neutron in to plasma of proton and electron. Thus the theory which has been created and been followed in the world of cosmology that universe at the beginning was hot and then as it cooled down and gravity took hold and then matters are created are false, The hot universe and cooling and production matter through cooling, this is an understanding of creation of matter and based on the matter world condition structure is base of old centuries and understanding of matter from cooling of gases and liquid into solids.

  The heat distribution of the universe is the same as the heat distribution in the plasma of the proton after the division from the neutron. This balance distribution of the heat has been observed and recorded by the W-map. This pattern of no heat and creation of gravity due to interaction of pmf fields in the universe leading to creation galaxies and stars are the normal process. Where in the universe matters operate through mafs and their environmental mafs and no other factors. This being the reason why one can observe matter in all corners of universe independent of temperatures and pressures which prevail in that part of the universe.

Matters in the universe are brought to the state of conductivity and connection through their Magravs and not heat and pressure and that is the reason why life exists in all dimensions and structures and intelligence in all corners of universe and life is not the exclusivity of water and oxygen based entities and environments the like the earth and its citizens.

- The present concept that the production of iron atoms are the limits of fusion in the universe and the point where the fusion of matters leading to production of iron and absorption of all energies of the star this leading to collapse of the star and that the star go through the final cycle of life as an active star, this again shows the man's lack understanding of the full science of creation and motion of mafs and maters in the universe.

  In fact due to presence of principal matter in the centre of stars, the process of fusion will go on and all heavy elements are produced through the process of fusion in the centre of the star and not as has been assumed when the process of red dwarf and white dwarf are initiated.

  Stars produce the heavier elements in a rapid process when the Principal mafs of the star releases part of its pmf, facilitating and making available pmf necessary for the fusion of the matters to heavier elements. All heavy matters in the universe are created in the fusion centres of the stars and most of very heavy material like uranium and heavier principal mafs are created in the principal star regions of the universe.

  The point to realise and which has been overlooked is that most of the universal objects like stars and planets, and principal stars they all carry in their centre a large amount of nano materials in all mafs and matter forms and it is correct to understand that in these areas there is a large amount of gans of matters too. The reason for the existence of these nano matters in the form of individual or specially in the form of layers and sheets in the layer boundaries or interfaces of one or different state of mafs or matters, the existence of gans and nano mafs and matters in these large objects leads to creation and production and enhancement of gravitational, and Magnetic fields of the these entities through the principle of interaction of dynamic motion of mafs and matters with magnetic matter or matter of these objects. As this interaction are the main source of increase in mafs and their effects in all matters accelerations in the universe and scientific world by large has not understood this concept in full, as if they would have understood this, they would have understood that there are no limitation within the structure of the planets and stars for them to create conditions that through fusion all elements in the universe can be produced within their sanctum during their cycle of existence, and not when they are at the their latter stages of their cycle of existence.

  In the centre of planets and stars always through interaction of plasmas and their capability to unwind and open up and release all their components of mafs and matters, then it is possible for the centre of planet to reach a point that they can release the content of the plasma of their gas, gans and matters that this leads to release of the principal matter of the plasma, that this high strength mafs through rapid reduction in strength can lead to rapid creation of new atoms and due to its localised strong Magravs in the centre of stars and even planet, this allowing the production and help with the fusion of large number of plasmas of proton and electrons separately, that further on by interaction of large quantities of these plasmas this can lead to production of any element in the centre of the planet or a star. This be it the atom of uranium or copper or even fusion hydrogen to produce helium.

  Fusion is not action, which is exclusivity of the centre of stars or universes. Fusion can take place in simple environment of universe, where there are no confinement matters like the core of the earth or where there are no high temperature regions the like of the centre of stars.

Fusion of elements can and do take place in the open spans of the universe, when the right mafs strength environment are created without even existence of no tangible matter.

Thus the creation of iron in the centre of stars being its doomsday has no meaning, and has no truth in the realities of world of production of matter in the universe, the reality is that if the centre of the star cannot create new heavier matters than iron during its working concentrated fields how can it do so when most of its matter are in process of disintegrating and diffusing in the environment of the star. Thus this is another fallacy about the doomsday of star and creation of iron atoms in the centre of the stars.

- In the process of the movement of pmf one has to understand that the interaction of magnetic fields and the repulsion of the similar magnetic field poles of the pmf of matters of the universe leads to creation of the repulsion forces of the universe, thus leading to its so called separation of its matters and what is called the expansion of the universe.

  While the interaction of opposite polarities of the same magnetic fields leads to creation of the gravitational field forces of the universe and universe cannot exceed its magnetospheric interaction size, where this is the maximum fields balance between the gravitational and Magnetic fields of the universe, and at the same time the mass of universe will remain the same as the mass is the balance fields between the gravitational fields and Magnetic field strength of the universe.

  Thus the mass and the content mafs and matters of the universe are set to be the same at the beginning of its creation less what pmf are transferred to it surrounding universe(s) during its life for this universe to keep its distance and Magravs positioning in respect its contents and in respect to others universes in the grater universal scale of existence in the unicos.

- When, the pmf of matters transferred by this universe to it neighbouring universes is more than what this universe needs to sustain its existence, then this universe cannot hold its present universal magnetosphere and gravitational balance, once this balance is disturbed then this universe like its previous universe(s) will go through division to find its new balance and this leads to creation of new universe(s).

  Thus, the concept that this universe will expand to point of running out of energy and go through crunch and/or dies as a cold entity or go through the contraction and go through a singularity point again, these are nothing but fallacy of the misunderstood and short-slightness of a few to realities of the operation of the universe and not in line with real world of existence as has been observed in the universe, and why in truth the creation of this universe should and division, but never ending process of the cycle life of this universe of its matters and fields be any different than the rest of universal methods of creation of fields and matters in their birth, fusion and decay.
  Therefore the limit of the expansion of the universe is dictated by the total Magravs strength of this universe itself and at the same time the space which is accommodate by and for this universe by the mother universe and the space given to this universe to expand into and the expansion of the universe observed presently is not dependent on no other factors.

  It is possible some times in the future this universe itself will merger and fuses with another universe and make a larger in mass universal unit too as much as it can divided into two or more universes.

- It has been considered that the noise which has been detected by scientist in the 1960's, that this noise goes back to origin of the universe, when this universe was made, this has no truth as the noise which is constantly can be detected from the cosmos is the constant noise of interactions of pmf of mafs of different matters and mafs in the universe and these are created continuously and constantly.

Thus is the same as when mafs of different field in the universe interact and create visible light, thus the noise from the universe shows that it active and living universe as different parts of it interact with each other, as one can hear the noise from the Sun in the solar system when it matters and field interact with each other.

In fact if there were any noise left from the time of the creation of the universe, by now it would have escaped from this universe through its magnetosphere into the mother universe, the way strong radiation pointed upward from the earth escapes from earth atmosphere.

- In the universal levels, the energy is defined by the term, that when the pmf are moving from one point in the universe to another, and as they transfer or they carry pmf of the principal matter for it to be slowed down to matter levels, hence this transportation of pmf of the principal matter to slowdown to matter level strength, this is called the energy of the universe or what commonly called the Transition energy (dark energy) of the universe.

- The universal energy is the one which the pmf in motion is carrying that is balanced in mafs strength with the environment in which they are travelling through which due to their balance field strength the two field create near to no resistance to each other that this condition creates near no interaction with pmf of its environment that leads to release of pmf that become visible, thus the pmf of the universe travels in the soup of its own field and hence less losses and hence the lack interaction, which this condition of travel does not create collision and hence no residual release of mafs of any significant and hence not much of the visibility of the travel of the mafs of the universe, but this transition energy is translucence and as one can see through it so much so that one can see through it stars billions of kilometres away as their light travels in the balance Transition energy of the universe.

  This concept of travel of pmf or Transition energy in the transitions matter is explained in the book “The structure of the Light” (2).

- The universal energy does not need to cool to create matter, to reach matter mafs pmf strength level, but the energy reduces in strength to pmf strength of the matter that matter can be formed, this being the reason why there is a homogeneous heat distribution is observation of the W-MAP as heat is not created during the early seconds of the creation of the universe. Thus the cooling phenomenon of the universe is a fallacy.

- The reality of the operation of the universal system is that the universe is made of pmf which their mafs and matters in motion become energies and that is where the cosmologist says that at the point of birth the universe there was only energy and as it is known now energy is defined (1, 2) is nothing but magnetic fields or plasmatic magnetic fields in motion.

  At the present some consider that there are more Transition energy than transition matters, which is not a correct assumption, where if this was correct then this means that the principal mafs and matters are unwinding much faster and in a rapid manner, which if this state has or to be achieved in the universe, then this means that the losses are more than expected and this can indicate that universe is losing too much too fast and this can be an indication that the universe is getting ready for a further division into two or more universes.

- At the same time it is possible as the universe was created from the splitting of the mother universe, then the separation of the pmf of that neutron universe due to separation of its mafs and matters have released some pmf, which might have created some localised heating in some parts at the point separation and division of the neutron mother universe, which this separation leading to creation of hot spots or small heating of the new universe in sporadic positions and conditions within the plasma of the universe.

- The plasmatic magnetic fields strength that are released from the Principal matter to matter levels and during their process of the strength level reduction from principal matter to matter level, what is known as the Transition energy, as the plasmatic fields need space and matter in the universe to release and lose some of their strength to their environment to reach the plasmatic magnetic matter mafs strength level, which is mafs in motion from one level to another strength level is known as the Transition energy or universal energy or what is called the universal energy in motion.

  The speed of travel in the universe has only relation in respect to its mafs and matter strength environment, hence one can travel at the speed of light of environment strength in any mafs strength, this meaning that in the matter environment on can go faster than the speed of the light if one travels with the Magravs of the plasmatic magnetic field strength of the transition matter or Principal matter plasmatic strength level.

  Therefore one can only travel with a faster speed than speed of light in matter environment only if one has the capability to create and/or captures Principal matter or transition matter plasmatic magnetic field strength in a Margrave positioning reactors (1).

  Hence the limitation of the speed of travel limited to speed of the light in matter environment is applicable if one use matter and mater mafs for transportation otherwise there are no speed limitation in this environment as one can travel faster than the speed of the light in matter environment when using principal mafs strength in the cores of the Magravs systems, where the speed of the light in principal matter environment is millions of times faster than the speed of light in the matter environment.

  One has the ability to travel any speed in the universe and with any mafs and matter plasmatic fields strength in but there are hazards to consider prior to taking such a travel.

  The theory, that the speed of the light to be the ultimate speed of travel is only applicable in the matter environment using matter plasmatic magnetic field strength and it is not true if using stronger mafs strength, thus one can travel the spans of the universe by using higher order mafs with speeds millions of times faster than the speed of light in matter mafs strength like on the surface of the earth.

  The body of the man will not feel any difference and will behave as one is in the jetliners of the last centuries as the pmf of the mafs of the human body cells contains within their structure mafs of principal matter, then the man will not feel that much difference travelling with speed of light or thousands of times faster than speed of the light in the universe.

- It has to be noted that as in the matter and mafs universe and environment, in the principal mafs and transition matter environment, these environment possess their own shock waves and speeds of sound and as said before even smells and odour of themselves, which these are different than the ones in the matter universe. One can smell the Principal matter and hear the noise of the transition matter interactions, once one is trained and developed the senses for the observation of the transition matters and principal matter which are part of the structure of every living thing as these matters and mafs are part of the structure of the plasma of every proton and electron in the matter environment of the universe.

- Once understanding the structure of the universe, it can be conclude that universe will never end, there are no singularities and no cold ends to it and no crunch, and only for ever the universe(s) will divide in plasmatic magnetic fields strength and from one mafs to another and then in fusing together they repeat the same process, and as the universal structure never can lose any of its fields through by its natural laws of conservation in physics, then only matters and matters converts from one matter or mafs to another form of mafs or matters and order of preservation of fields and matters is obeyed even in the bigger universal order as has been in the matter world as has been studied in details over centuries. Thus preservation of energy and conversion from one form to another is the essence of creation in every levels of the bigger universe too.

- In the matter universe for example smell is only diamond structure of the or SP3 structure of the same atom in the gas state, thus in the principal matter environment the principal matter possesses and have different states of matter like the gas, liquid and solid state of Principal matter too.

  The same there are molecules and atoms and plasma of the Principal matter too, therefore principal matter have smells and odours and they frieze and they boil and they create clouds and become all thing as one has observed in the matter environment of the universe too.

  Then one can ask the questions that are the cold and warmth the same as the matter universe as in the principal matter universe. And the answer to this is most certainly is yes, with the difference that the temperatures are beyond the levels of man's understanding as the strength levels of plasmatic magnetic field in these environments are off the scale of comprehension of the human at the present time.

- Then the point arises that are there lives in the Principal matter and transition matter scale of plasmatic magnetic field strength of the environments of the universes, and the answer to this is most certainly yes, as the same structure of interaction and creation of cells and life exists in the these environments in the universe as are these conditions available and possible in the matter environment in the larger scale of the working of the greater the universes.

- The matter made available to the matter universe by the down levelling of the plasmatic magnetic fields strength from the Principal matter levels is what allows the creation of the matter, where these new matters have to find physical Magravs positioning in respect to each other, this need for new positioning in cases can lead to movement of the matter and increase in spacing of matters in respect to each other, hence this new increase in the spacing looks as the universe is expanding. Thus the increase in the spacing or expansion is limited as the matters and the rate of conversion from the principal mafs to matter has a limit and hence the universe cannot explode and only thing which can happen to the universe, is that as the Margrave of the universe is reached a point that cannot be sustained in its present mafs levels needed for the creation and sustaining its Magravs level, then the universe in a gentle way divides to guaranty its overall survival in two parts, rather than losing everything in the mother universe to other universes all together.

- The universal Magnetic fields within the transition matter is the force which pushes the universe apart and the gravitational fields within the transition matter is what brings the matters of the universe together and is the glue of the universe. Where both these forces are created by the principal mafs and matters field's interactions of the universe and the transition matter is the medium which they use without interfering with to a large extend for these forces to keep the universe in its dynamic condition.

- If the rate of the conversion of the principal mafs reaches to the point that there are more energies in transition mafs strength levels that the rate of conversion cannot lead to creation of matter mafs of magnetic fields strength, then the transitions energies fields interact and this will lead to creation transition matter, which this condition is abnormal and this ultimately in some cases can lead to creation unbalanced Magnetic fields and gravitational fields within the universe, and this is the trigger point for sudden expansion of the universal magnetosphere in certain point and as this cannot be sustained, then the universe goes through division and the universe and creation of two or more fragment of this universe are the consequences of such condition.

- It has to be clearly understood that the neutron division to electron and proton, usually is caused due to sudden expansion of the neutron magnetosphere, which is caused through the same process of the rapid change of the principal mafs to transition energy, which this leads to temporarily creation of transition matter that this creates a new magnetosphere condition which cannot be sustained, hence the neutron goes through the process of division to electron and proton too.

- As the pmf of the transition energy slows and reduces in strength once reaching the boundaries of the universe as it happens in the planetary upper atmosphere and solar systems heliosphere, in the same way and by the same principle these transition mafs will rain inwards into the interior of the universe, where they started their lives from and at the edge of the universe boundary the same as cloudy earth in its upper levels will have showers of matters in the form of small plasmatic matters and mafs and this looks as universe is raining inward. Thus once one look at the boundary of the universe one will see cloudy or a hazy shell, in fact this haziness it is nothing but the weak mafs and matters which are lingering around as clouds do or dust particles do in the planter systems upper atmosphere.

- The same cloud systems do exist in the solar systems heliosphere too, where this zone where most of meteoroids and asteroids of the solar systems find their seeds and receive their regular matters and mafs from.

  It has to be said that once man learns to cross the boundaries of the heliosphere of the solar system, the travellers of the crafts as is done when man leave for the travel to the moon at the present time, these crafts as they enter the heliosphere of the solar system they will encounter the same misty cloudy space at the edges of the solar system too and still they have to deal with solar heliosphere re-entry and cloud dusts and so forth as one encounters in the re-entry into the earth atmosphere.

  On the edges of this universe one will see weak pmf of matter and transition matter and transition energy plasmas and even one will observe the entry of other mafs and maters and plasmas in all mafs strengths and entry of universal meteoroid from other universes and unicos, as one sees meteoroid and radiation entries into the earth atmosphere from the solar system and cosmos. There are universal lightings and universal meteoroid too, in fact if the W-map observers look for this they will see another storm and lightings as these vast energies and large matter enters this universe at the edges of the universe, but they simply have registered these as hot spots in this map’s outer boundaries.

  The existence of these short lived hot spots will show that they are dynamic, live a short life and their energy is fast moving, as it happens when one has meteoroids entering into the earth atmosphere.

  In reality the existence of these hot spots on the edges of the W-map should give a clear indication to the man the existence of other universes and the greater universe as earth is part of a solar system and a galaxy, the same is with this visible universe of the man, and all one sees in this universe as the same is repeated in the greater universe too.

- One of the fundamental principal concepts of the universal energy is what the scientific world of today to a large extent has not understood fully and has up to now ignored this fundamental principle of the pmf magnification using the pmf of the mafs and matter and energies and specially the universal energy.

-

If the scientific world understand how universe uses its dynamism and its inner content dynamic universal energies layering to magnify its power in its operation of this universe, then men of wisdom will come to realise that one does not need principal matter that much and in truth any matter so much all to operate this vast universe and transfer energy to all its far distance corners.

The world of science of today has been by large ignorant to this fact and has not managed to understand the magnetic layering and dynamic universal energy principle and in turn this has put man in disadvantage in comparison to the technological advance with other travellers of the universe.

Man of wisdom should look within to see the creation of their own being, then they will come to understand how small amount of matters and mafs and magnetic fields has been creating conduction and energy transfer in their own structure without utilising or using very little mafs or matters, then they find the power of creation and its application and manipulation for untold usages.

Then they will realise that one does not need wormholes and all that fancy words to travel within the universe with the speed of their own creator. This shall and will remain the mystery of existence upon the time when man learns not to trespass the boundary of others and he learns the existence of balance and true peace within himself and its people and it's surrounding.

- One should understand the concept of multiple principal stars operation in one galaxies system and the role they play to create disc like or spiral galaxies. The operation of visible and invisible principal stars in the universe will play the same role as the twined solar systems and twin galaxies, where in this cases the twined principal matter centres are and have the capability to discharging different pmf strength according to their anti-martial gravitational pull and the measure of the mafs that they might release to control their field of domain. The multi- principal matter centres are and can be used in the space travel in a fascinating way, where once creates multiple jumping steps, with differential and opposite dynamic direction of motion of these Matters. The speed and protection created from and through these manipulations will give the man the chance to travel the length of galaxies and universes in the minutes, but one has to be aware of the destruction power of these technologies as there will nothing left but pmf of the creatures travels of these universal craft if things goes wrong at these mafs levels.

- It should be noted that the shapes of the galaxies and their physical appearances, should teach one the power of the effect of the other matters and entities around them as they will allow and pull an push these objects toward or away from one point or other and hence allowing the appearance and the physical visible structure of these celestial bodies to obey and have a given shape, this is important in the understanding that one has not to look at an entity but look at the surrounding area of the entity to understand the limitation of the operation and the shape of the matter and mafs internally and externally as it is allowed by the environment of the entity for it to exist within a given point and time of a system, this being a planet, solar system, galaxies or universal entity.

- The uniformity of the pmf within the plasma of a proton, an atom or a universe these all show the equilibrium and balance condition between all the participating magnetic fields forces which have been designated to the entity at the point of its inception and creation. Thus, if one sees a fast moving object the like of the Plasma of an electron in an atom or in a larger part the universe, then this does not mean that the universe is expanding and it will be chaos internally in these entities. Thus the fast movement of the internal matters and mafs of the universe does not mean that the universe is expanding, on the other hand this means that the universe is internally controlling its mafs as repositioning to guaranty its existence as an entity and at the same time the universe is holding its position or what is called universal Magravs positioning in respect to other universes in the unicos.

Thus this universe will have matters and mafs which will be moving away in one position but at the same time other fields in the universe will be moving in the opposite way to, but it has to be said that the conversion of mafs to matters creates a visible expansion, but overall gravitational field forces of the universe cannot be changed and the universal order of preservation of fields will be maintained, does not matter what is converted to what fields and matters and the internal universal gravitational fields strength will be maintained and preserved.

- The universe has laws and conditions and has primary atoms and particles, at the present man has managed to understand the structure of atom and plasma of proton and other particles and through this understanding and his vision of his universe thinks that the atom of the hydrogen is the simplest structure as he can see and understand, in time this will become apparent not to be the same for all creature of the universes, but up on and until then man has to learn that the fundamental particles and unifying fields and the fundamental fields of universe are the simple magnetic fields of the universe and the whole operation of the universe is based up on their interaction and operation.

  But at the same time man has to learn that all created entities in the grater universe have achieved a common method of communication or what is known as the universal fields communication, all created being which have managed to influence their environment and have managed to interact with other intelligent being individually or as unified groups and have managed to accept one universal communication language that is acceptable to all, this language is known as the universes language and is bases on the strength of the pmf of the different strength and not based on the voice or sound or physical alphabets and shapes.

  The man will have to learn this language and will be able to communicate with the other intelligent created entities in the universe once man learn to find peace within himself and reaches the understanding that they cannot trespass and cannot create division in the universe as they have been accustom to on earth to get their way and in achieving and attaining physical possessions and that in the universe martial possession are nothing but the lack of man's understanding of his position and short slightness of use of his intelligence.

  The universal language is transparent and clear language and works on the pmf waves strength and has standard forms and accepted levels for communication depending upon the creatures of different environments which creatures are initiated from.

  Where, one should understand that not all created intelligent beings in the universe need or needed the physical planetary structure as human do for their existence for them to be created or to survive.

  Life is created in the universe in and within the fields of the universe and it does not necessarily need physical tangible, food and land for it to exist and collaborate and operate with others in the universe.

Whereas, and as long as the Man tries to contact others through pictures and voices using the present technological advancement, man will find in general he will have a long time to wait for the response, as the other created intelligent beings of the universe have been and are and will be visitors to man's home from the time of its creation and to its demise, but the universal language will be established and communications will be opened when man learns that he does not need to take more than he needs to and at the same time man does not need to create division to get what he desires and cause conflict.

The cause of isolations of the man from the universal community has been his lack of understanding that in the universal language the other creature do not need to use words but they can and do understand and communicated at pmf fields levels, thus the thoughts of the man which is pmf based in his operation in the brain of the man, these pmf are detected and are like visible light to these creatures and as what is said they can see and understand the real intentions of the man's deeds without man uttering a single word.

Thus man isolation from the universal community has been in forced up on it out of protection of the universal peace, rather than man's inability to find the other creatures to communicate with.

- The Universe in general is a peaceful place and wars and conflict has been things of the past and a long-long time in the past, hence once the man learns the advantage of peaceful living, then man will learn the universal language and then man will be allowed to learn more and join the universal community of Beings, irrespective of these being physical or field appearance to the eye of the man.

- One does not need eye to see but the intelligence to understand, and then the real sight is the real will of the observer of the beauty of the creation in all levels of universes.

- In the material world for example the neutron always divides into balancing masses that the overall gravitational positioning of two new masses can guaranty the existence and continuation of the life of the neutron in the same material world. This applies to the universe too, that universe will divide in mass and as in the neutron the lighter mass will have to follow the graters mass fields forces Magravs motion and hence the universe of the smaller mass will be more agile than the heavier mass. Whereas, the smaller mass universe will be and has to follow the internal motion of its own mafs and matters and at the same time the overall Magravs of the smaller universe has to find its position in respect to overall Magravs of the larger universe and the overall dynamic motion of the mafs and matters which the bigger universe creates.

  Thus, the smaller universe can be in **a** snapshot or in the case of man life time over a million years, this universe be looking as it is stationary in its internal space, but in fact this universe is following another larger universe at very high speeds itself to keep its position and at the same time the universal centrifugal force of motion allowing the universe to hold on to its content of its own mafs and matters.

  All universes the like of the galaxies and solar systems are all in motion not only internally but at the same time they are in motion in respect to their the other half of the their mother universe, which it was originated from. The overall direction of the motion of the universe in its bigger universe can easily be detected and realised. The direction of motion of the universe is shown by the directional motion of the mass of universe, which is moving out of normal direction of the whole content of the universe itself. This meaning that, if! The full matter motion of the universe is planned out; one should see a region, which is moving as a whole in totally arbitrary and independent direction.

The reason for this envelope of the mass following an independent directional motion is due to the fact that this mass of the universe is connected Magnetically and gravitationally is locked to the direction of the interaction of the Magravs of this universe in respect to the Magravs of it co-birth sister Universe (Fig. 22).

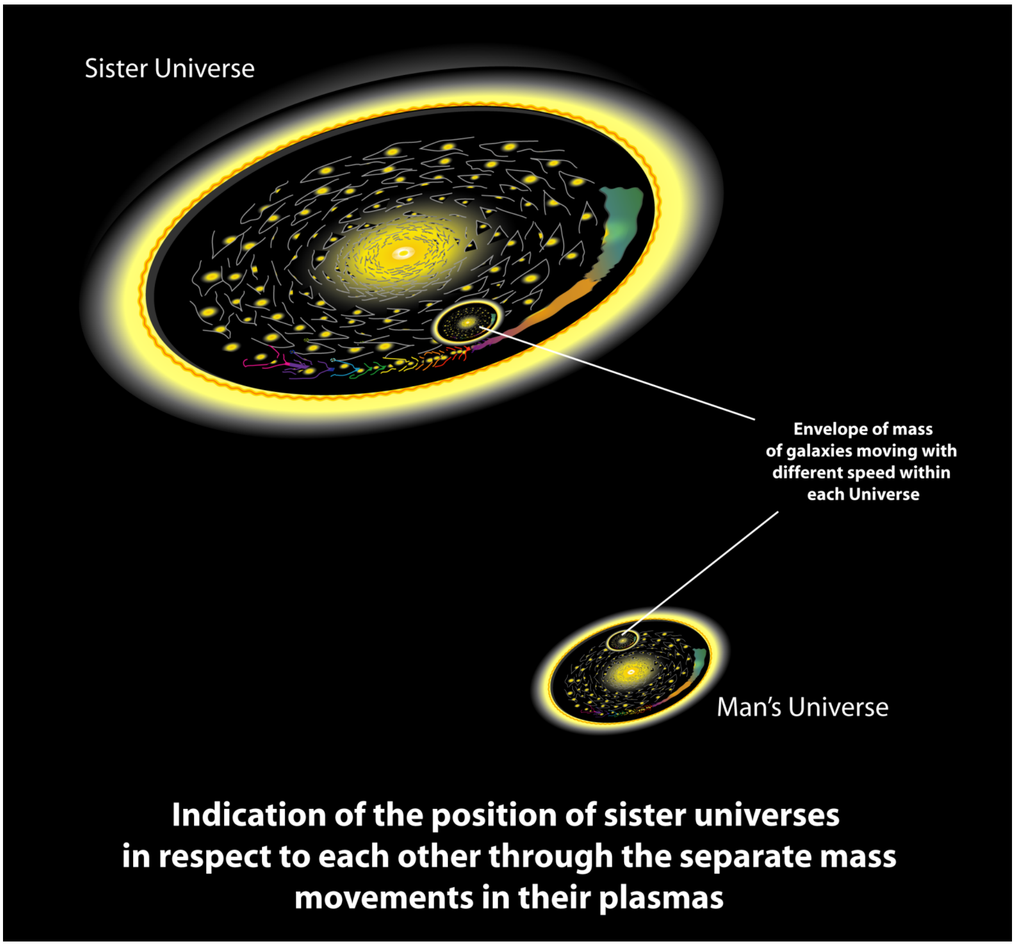
Sister Universe
Envelope of mass of galaxies moving with different speed within each Universe
Man's Universe
Indication of the position of sister universes in respect to each other through the separate mass movements in their plasmas

Fig. 22 Envelope of galaxies in each Universe indicating the position of the Sister Universe.

The mass of the matters involved in this motion should indicate the extent and power of the Magravs interaction between the two universes.

This mass moves in all three hundred and sixty degrees direction as the co-birth universe is dynamic and has its own dynamic Magravs motion too.

This independent zone in motion in the universe has already been detected by the cosmologist for this universe, but they could not understand the reason for this independent part motion in the plasma of the universe.

Thus the direction, which this mass in the universe is moving and is positioned, this shows the position of the co-birth universe in respect to this universe.

Therefore if and when the man's scientific and technological achievements reached the level in the future that man can see through the universal magnetosphere, then the position to look for the co-birth universe, would be the direction where this internal mass is in motion and at the same time the distance of the this internal mass from the edge of the universe will show the distance of this universe in respect to its co-birth universe.

- In the universe the heat and cold is not measured in the way, which has been accustomed in the world of physics of today. In the universe heat is measured to the extent that the matters pmf can be loosened to allow the bridge of what is known as the atomic or molecular magnetosphere.

  Thus what is commonly known, as the coulomb barrier, in the atomic physics is only the extent of the resistance or the tightness of the Magravs of the atom. In fact man has to realise that if he unifies the terminology then he will solve most of the problems in the world of the science of physics. For example, by unifying the word of the magnetosphere and coulombs barriers, which in fact these are both the same properties and principles; where one shows the tightness of the Magravs of the planter system and the second show the same for the atomic structure. If this discrepancy in language is resolved and man understood all phenomenon of the creation of magnetosphere in all levels universal mafs strength which has been discussed above, then the nuclear and theoretical physics would not have the problem in creating and maintain fusion of all matters and mafs in the universe.

  Thus for fusion one does not heat or speed to create and achieve fusion, but the loosening of the Magravs or dilution of the Magravs of the mafs and matters will lead to fusion and creation heavier elements without the need for the elaborate accelerators and others gadgets which has become the tools of the advance science of today.

  The use of the accelerators are nothing but to show the true lack of understanding of the scientific world in the what these scientists are trying to achieve and their misunderstanding of the operation of the universe.

  The heat created by the fusion processes within the stars or the principal stars is the by-product and in fact the heat assist with the operation to loosen the Magravs and loosening the Coulomb's barrier that fusion or mixing of different plasmas of mafs and matters are accommodated, but not the cause or the outcome of the fusion.

  Where in the world of the fusion of today scientists try to speed up and accelerate plasma and make them to collide and hope for the fusion, but in the universe the reverse process happens, whereas in the universe fusion takes place in condense centre of the principal stars and matter stars, where there are no rooms for the plasmas to speed-up, but due to loosening of the pmf forces magnetosphere of the plasma or coulomb barrier, then matter mix and fuse and what is not necessary from the amalgamation of two plasmas fusion, then these appears and are released as surplus unwinding pmfs of mafs which as in this process of higher to lower mafs strength is so rapid that this process shows itself as the heat in these environment of the universe as part of the process.

  Thus heat can be explained in its real term as the pmf which is unwinding from one state of matter to unwind that the pmf field released allows the other matters to open their magnetosphere that they can receive the pmf in transit from other plasma.

  Then the cold can be explained as the when the pmf of the matters are reduced to beyond the magnetosphere of the matter that matter cannot transmit any pmf to another matter.

  At the point, where the total transfer is zero, this is known as the Absolute zero point at the present time. Where, the scientists say that at this point there are no molecular movement. In fact the plasma of the matters or ***the atoms is still at this point active*** and maintains their internal dynamic condition and environment. If this was not so then magnetic fields creating the mafs and matters would have ceased to interact and the matter or mafs would have lost their gravitational fields and magnetic field and hence the matter and mafs would not have been able to maintain the magnetosphere and the mass of the atom, and hence nothing would have existed; But as one still sees the matter at absolute temperatures, then the condition is inter plasma magnetosphere connection and not internal magnetic connection;

In fact magnetic fields, which are the principal particles or are known as the fundamental matters can never come to a point that they do not exist or interact with other fields and at the same time due to ever existence of other magnetic fields and fundamental elements in the universe, then all fields and matters at all times are attracted or repelled by one matter or other and hence a total non-dynamic universe never can exist and in truth the creation of the universe and creation and origin of the magnetic fields is due to and through the fundamental of difference in the universes differential waves. Thus the magnetic fields are themselves created out of strength level difference in the universe and thus the origin of the magnetic fields is the cause of the existence of the universe and the existence the universe is the cause of the existence of the condition differential which is the creation of fields to compensate for this difference and hence the creation of entity to cover the difference in position and hence the creation of the magnetic fields. Thus one in a simple language now knows even how magnetic fields, which their variation leads to the world of the creation of universe has come from and how the real essence of all creation become to exist.

- It can be said that at the Absolute temperature the time stops for the matter and material world but not for the mafs and mafs dimension and hence if the universe were made only of matter and mafs then the physical matter would have never existed in the deepest part of the universe if the temperature were at absolute temperature.

- The measure of the cold in the universe is dependent upon the slowing down of the movement of given pmf in a given direction and strength, hence the measure of the coolness is the external strength dependent and not plasma internal fields motion dependent. This being the reason why like in the planetary systems planets can and do produce their own independent temperatures irrespective of where in the universe they are and how cold or warm the environment of their surrounding planets and stars could be. Therefore *THE ABSOLUTE ZERO IN RESPECT TO* mafs and matter environment is only applicable to the earth position and environment and not to all universal condition. This being the reason why, one can see and observe life in all kinds of pressure and magnetic field strength in the universe.

  Therefore the Absolute zero for mafs and matter environment is not applicable true for principal, transition and matter mafs and matter strength and environment.

  This being the reason why that internal atomic structures can exist and atoms mafs can hold on to its structure and its existence, while the molecular structure in the matter environment can be cold that no external motion within the atoms of the molecule can be observed at absolute temperature.

  The absolute temperature in a given environment can be said to be when the Magravs of the constituent matters environment come to be in total balance; and that is the point that the total losses and gains from the surrounding environments are said to be balance out that the total magnetosphere of the environment will compensate for the losses that the environment can keep on holding on to the zero motion of the total environmental Magravs.

  On other hand the environmental Magravs can be reduced that there is no interaction between atoms that the motion of the atoms in respect to each other ceased to exist and motion in respect to each other become frozen. This is known as balance Magravs criterion and to produce this within the structure of reactor is an art not known to man. Where, the balance Magravs positioning is the key to establishment of colonies in the universe, where there are no planets and stars as point of reference, but the condition of balance allows the station to be permanently positioned in that spot that all passengers of space can come to meet.

Production of an absolute balance give man the ability of what is known as hovering, where there are no propellers and no vertical jets of today to stay stationary in mid-air.

Absolute Magravs balance allows crafts of the future to keep a constant position and this can create fantastic composition for gate-thing, which is needed for movement from one Magravs system environment to another without actually needing to handle re-entry problem in either direction.

The operation and the use of the absolute balance Magravs technology systems are as for the physical doors of today, when used in the future, when one moves from Magravs plasma environment to another.

- In the operation of the universe, the distance of pmfs and the speed of the rotation of one pmf in respect to another is the criterion by which the speed and directional rotation of one matter or plasma in respect to another plasma is dictated.

  Where, the positioning of mafs in respect to each other in the universe has a direct effect on how fast, when and where an entity will position itself and even this dictates the measure of the cold or heat of two entities in respect to each other and to other environments. This concept is seen in the universe in respect to the rotational direction of planets, stars and galaxies, where their direction of the rotation if different to the one other entities in their surroundings.

  The criterion of mafs positioning is the reason, why for example the Vinous rotates in different direction than the rest of planetary systems in the solar system of the man, as its internal pmfs speed of rotation and their position in respect to others of the same within the internal structure of this planet are as such that the outer layers of this planet are forced by the configuration of the combination positioning of different cores and their pmf strength, forcing the planet to rotate in the clockwise direction rather than the anticlockwise direction as the rest of the planetary entities in this solar system.

  This principal of the mafs positioning and speed in way can explain even why all matters in the universe rotate anti-clockwise or what is commonly called the electro negativity of the hydrogen atom directional motion.

  Hence it can be said that the anti-clockwise directional movement of the plasma of proton and electron are due to the mafs strength within their plasma and the position of different mafs in respect to each other within the plasma.

  Then this puts a closure to present theory of electron and positrons, where in fact electrons and positrons only difference is that one's internal magnetic fields are slightly differently positioned than the other and as their mafs contents of fields is the same, then only the positioning of the fields in respect to each other dictates the direction of the rotation of the plasma of these entities. This directional motion variation can even affect the absolute temperature of the operation for the positrons and electron in different environments too.

- Time as man is accustomed to on earth, this time in the universal language does not exist; time is a manmade phenomenon to measure the course of his life. In reality the time is the measure of the pmf of given strength to pass a given predetermined strength, thus time in the universal language is magnetic field strength and environment dependent. Where in the principal mafs world the time of reduction is much faster than the time of reduction in the same matter environment. Thus time or measure of reaching from one level of strength to higher or lower by the same order of magnitude is the measure of the time in the universe.

Where, the universal positioning time will be measured according to the central position of the universe and considered to be zero at this point and then to make an appointment to be at a given time then the motion of all mafs and matters in the universe in respect to this point needs to be considered, Where this position is and has a three dimensional three hundred and sixty degree reference. Where, in fact time in the universe needs point of reference and that is earthly time, solar time, galaxial time and universal time.

Time is unchangeable and it is the same in the same gravitational and Magnetic field strength environment. This is important, as at the present there is a fallacy, which has been accepted to convince the scientific world that time goes faster in the universe as one moves away from the earth.

This is nothing but hypocrisy and lack of the understanding of how gravitational fields are created. This fallacy has come to a point that in the satellite systems of today an error time has been allocated that the satellite timing is adjusted in respect to position location of a matter or entity on earth. Where in fact if the scientific world would understood the full principles of the creation of matters, and if scientists can create the same g strength forces in the environment of the operation of the electric and mechanical processing of these satellite does not matter where they are positioned, what is called the environment of what is called one g, then these satellites do not need correction factor which they are using.

Where, in fact the time correction factor is due to the fact that matters in the different Margraves and gravitational position around the earth have and operate under different magnetic and gravitational field strength and this is the reason why there is a difference.
Thus for the correction of time in these systems only one needs to consider creating the same Magravs strength at on the surface of the earth and then there will be no difference in time.

This concept is important for the future travels of the man into space and his excursions in the universe if he wants to keep to his habitual twenty four hour clock and earth gravitational magnetic field strength of his planet.

Man soon will find and realise a simpler way and then the hypocrisy of time travel and wormhole and going to the past and to the futures and so-called granddad theory will come to an end.

One point to remember that light which itself is plasma entity moves out in a spherical three hundred and sixty degree environment of the universe and for example the feature of a the face of the man will create man lights for it to be seen as a whole, but in time as light travels in the ever rotating sphere of the universe and these rays some get absorbed and some get distorted, then one cannot bring all the lights of all the matters of the one human face together to see that entity as it was so many time lapses before.

Thus if one hopes think that one can see the past of one person on a revolving dynamic environment through one light or making a machine to go backward to see the face of its forefathers, this shows nothing but the lack of the understanding of the science of creation and motion of light through the haze of dynamic universe by these men of science.

One can see the structure of matters like stars and fields in the far points of the universe primarily due to the fact that the light is and has plasmatic environment and like a telescope can pass the information to the point of arrival. But the light cannot bring the past of its source, but shows the pattern of its source like a move at the time it left the object.

In fact it can be said that light is composite matter and entity with active complex pmf of different mafs.

Thus, for example one as one enters the principal matter environment, where the principal matters is the creator and controller of Magravs of the environment, then in these environments one travels at the speed of the pmf of the principal matter and not pmf of tangible matter. Then if considering the principal matter environment in the centre of galaxies, then time for the matter world will be looking as being much faster or matter seems to be travelling thousands of times faster than what has become to be as the speed of light in the matter world in these environments.

Thus in reality the accepted time and speed of light as the maximum limit is only applicable for and to human understanding of the universe and not considered to be the truth in the real working of the universe.

When, the race of the man understands the concept of the use of the principal matter in the travel systems of their creation, then the man will be able to travel into the universe with much faster speeds than the speed of light of the mafs in the matter world and then one can travel the spans of the universe in seconds in human term and then one can travel the length and the breadth of the greater universe as one travels the earth glob with the toy systems of the jetliners of today.

Then the point arises that why some mattes and entities in the universe can last for a long time, where for example the human life is limited, the answer to this is very simple, and that is the body of the man is made of composites of mafs and matters and the more one adds to the composite the more Magravs are available to be lost to the surrounding Magravs, and be effected by the surrounding Magravs too.

Thus the life of the man is environment conditioned and in time as the DNA of the man learns to preserve more, then this will be adding to the life of the man and then the man will live longer, but eternal life is and belongs to no one entity in the universe, but the entity is the total facet and attribute of the whole as a whole. Thus man has been and shall ever be, not in the physical form but rather in the attributes of its physical existence in all universes of the god. Where, god is part of whole and fraction of none.

Where, the universe possess mafs and matters which their difference only is in their strength and nothing more and nothing less, thus time is faster in the stronger field for the weaker matter created entities and time is slower for the stronger mafs field strength in slower in weaker matter field strength. Thus time in the universe in reality does not exist and is a creation man’s state of position and possession. Thus if one is in universe, where the strength of the mafs are one thousands time stronger than the other mafs, then the time is one thousands time faster and this applies to all mafs and matters in the world of creation.

One point to realise is that the magnetic rays which are released and are prevalent in the environment of the man are the time setters for the degeneration and division of the man’s cells and have direct effect on the life span of the man in the universe, this be it a word of wisdom for the ones who come to comprehend and look for the eternal time life.

- What holds the universe together are the attraction and repulsion force of the pmf content of the universe. Thus the tightness of and the position that matters appear in the universe are totally pmf strength dependent.

  Therefore as the pmf strength of the matters are attained at any point in the universe then that matter becomes to manifest itself in the given point and time in that part of the universe. Thus the glue of the universe is a combination of the interaction of all the mafs and matters, which their interaction makes the Magnetic and gravitational fields Magravs of the universe.

Thus in fact what is called the bosons theory has its truth in the position of and their interaction of the magnetic fields of different pmfs, plasmas, matters in a given position in the universe as these interaction dictates what and how pmf are given or taken from the universal soup of magnetic fields, and what the outcome entity shall be and how the matter manifest itself at given point and for what period of the time.

- Thus once a given pmf strength in the universe is achieved which can be initiated combination of the interactions of several conditions, plasmas, and environmental magnetic fields, then at that point in the universe a specific given mafs or matters comes to exist, this being mafs or tangible matter magnetic fields strength and so forth. This is the secrete of universal existence and manifestation of the all mafs and matters and the observation of one mafs or another matter is totally dependent on the observers pmf Magravs strength that one matter or another can be tangible or visible.

- As in the solid state magnets in the plasma state of the magnetic fields regions of the universe, these plasma poses and operate through the North and the South poles and the same goes with their attraction and interactions in respect to other magnetic fields of all plasmas too, that is to say that plasmatic magnetic fields position themselves as solid magnets do in respect to strength and position of other plasmatic magnetic fields near to and in their environment. Thus the North-pole of the plasma attracts the south-pole of other plasmas and at the same time the South-poles of the plasma keep and make distance from the South-poles of other plasmas as it happens in the solid state magnets.

  In the physical matter one can see the motion of the magnetic fields and motion of magnets in respect to each other as magnets move and find their position in respect to each other in their given environment once they are brought into the proximity of each other. The same positioning and movements are applicable and are created and maintained in the magnetic environment of the plasma two plasmas once they are brought within the mafs field interaction of each other too.

  The plasmatic magnetic fields of all entities can recombine to allow the circulation of their fields and in general determine the point at which the fields can re-enter the plasmatic magnetic fields to complete a full spherical plasma of the fields. Where, this point of re-entry of the fields is primarily dictated and decided according to the environment, which the mafs of the entity operates in and, not necessarily by the mafs strength of the field of the entity itself. Thus, the point of termination and re-entry of the fields of the plasma of mafs of even solid matters is dictated by the environmental mafs strength and not by the matter. That is to say the termination point for the magnetic field lines for a plasma and solid matters (the beginning and end of a field line) is decides mainly by the environmental mafs strength and not necessarily the strength of the fields of the matter.

  This being the reason why the magnetic fields can reconnect anywhere in the universe irrespective of their physical distance, as the environment which the fields enter and carried to their full destination is supported and facilitated through the mafs of the environment and not the strength of the fields of the entity. Where, this phenomenon is well explained in the interaction of the change of polarity of two atoms, which as one changes its polarity, the second one does the same instantaneously. This is what one observes in the change of polarity of plasmatic field's entities, that for example as one plasma changes polarity, the other does the same parallel changes automatically that one never observes any changes in the structures of either entity.

  This being like the change of the polarity of the earth and the synchronized change of the polarities of the cells of all beings on earth, that no physical change can be detected by the inhabitancies of the planet as the physical polarity of the earth changes.

Even though the same cannot be said about the change of the polarity of the moon of the earth as it does not changes as the earth polarity changes in majority of the cases.

In understanding the principle of this phenomenon, then man of science can make any matter anywhere in the universe as the need of the time be, this be it production gold or production of amino-acid needed for production of proteins of chain of fibres of heart of man anywhere in the space for the travellers of the future.

Thus, one can control the point where what appears as the magnetic fields can draw in fields (as called conventionally called the exit pole) and where the field exits to release the fields (the north-pole of the mafs. Then size and the length of the magnetic field and the gravitational fields and Magnetic fields, which any entity creates, can be controlled and hence the size of plasma of an electron or proton can be predetermined. That is to say the total field strength of the Magnetic field and the magnetosphere which an entity creates and can hold on to, this determines the exact and precise dimension of the plasma of be it this be the plasma of the proton and electron.

Thus, in principle one can have field zone of the same length, which upon their interaction they create a given magnetosphere, but one can have the same spherical dimension of magnetosphere with a different magnetic fields strength, which this creates different Magravs field zone. Thus, the length of the magnetic fields determines in what entity is to be created, this being the plasma of proton or an electron, and so forth.

Complex magnetic fields create different field lines, hence different Magravs. This being the reason why in the plasma of the proton different pmf and matters entities are observed and detected.

- In the bigger picture of universal works, scientists will come soon to realise that all parameters are the same, all plasmas do and behave within a given procedure and sequence. Thus, what one sees in the plasma of this universe is the same as what goes on in the plasma of the proton and electron too. That is to say what has been puzzling the theoretical physicists about the content of the plasma of the proton, as one can see in the operation of the universe then one should have been expecting the existence of different matters, fields and fields strength in the plasma of the proton and even the same in the plasma of the electron as one sees in the plasma of the universe.

- Interesting enough scientists soon will start wondering about the creation and control of these magnetic fields that how they lead to creation of different plasmas of magnetic fields too, and how plasma of the magnetic fields become to become units of the same fields and how a magnetic fields can be attracting one plasma while the same end of the matter can repel another of the same from the same end or exit port of the magnetic fields. That is to say how can a solid magnet at the same time as can attract the opposite polarity and then at the same time and simultaneously the same pole can attract another dissimilar polarity. The fact is that magnetic fields of plasma of fields do not bunch all to north and south or all exit and entry port, but they arrange themselves as one of entry and one of exist and or one north pole and on the south pole adjacent to each other, but they do not set end to end with each other. Thus for example looking at the section cut across cluster of the magnets, the fields are arranged like mosaic of north and south in correct sequence and order. Where, in the case of the dynamic plasmas like protons the top view shape of the magnetic fields are the likes of a whirlpool and not straight-line fields of the same mosaic.

Then one should understand how the structure of nucleus of atoms are arranged in a similar manner too, as one exit port sets next to an entry port, then when a nucleuses has two protons then the in and out ports are matching and the magnetic fields of the atoms internal magnetosphere is complete and the atom is stable, but when the atom possess two of the same ports and one of the other ports in its nucleus then the extra port is sharing its fields and or is looking for the an extra port that can receive its pmf or it receives pmf from, and hence the odd number protons nucleuses always are more receptive in receiving or dividing for them to have matching pairs pmf ports of entry and exit, that they can make their own matching magnetic plasma. In fact single magnetic fields do not exists in the world of plasma, and in a way the theory string shows nothing but lack of understanding of operation of the fields in the universe. This explains why odd number proton numbers go through decay or they are more receptive to make polymers of themselves or join up to make molecules.

Where, the entry port of the fields can take magnetic fields from the other ports or from the soup of the Magnetic fields of its environment, and thus the size of the entry port of the fields in principle determine the matter and mafs condition in a given environment.

The reality is that the entry point consumes more energy than the fields, which are coming out of the exit port of the plasma of the fields. The fact is that whenever fields interact they create minute contact points that at these contact points fields are divided into smaller length and strength, this is like when two bodies collide and due to their friction some heat or fragments are created; This happens the same when two magnetic fields interact to create their own magnetosphere and mass. Then due to this friction between fields the magnetic fields entry of south-pole are slightly weaker and less in mass than the total mass which originally entered the magnetosphere of the two fields in the south-pole. Where, in fact the slight difference in the mass and fields strength between the entry and exist of the ports, leads to ever reconnection and closing of the loop of the fields between the North–pole and the South-pole, and the motion of the fields within the structure of the matter and the entity.

Where, the South-pole or entry port is the feeder and the North-pole or exit point is the provider, where at the same time the due to miniature and micro size of the magnetosphere of the fields, there are too many magnetospheres which are connected and arranged in the mosaic of the one cross section, thus the entries and the exists or the north and south poles set next to each other and where the entries poles are more stronger the matter attracts and where two facing section the south pole are stronger the matter repel and hence the sequence of pulling and attraction of mafs and matters in the universe.

Where, the entry points both need to attract as much as possible so they move to guaranty their own continues supply of fields and this being the reason why similar poles move away from each other, that is simply to allow the entry or exit of the same floor of the fields to have the maximum chance of attracting fields to guaranty their field supply.

The dissimilar poles gravitate, as the entry-pole try's to absorb all that it needs from the supply of exit-pole of the other entity.

Further on the minute fields which are released from the interaction of the two fields, in them separating and being able to travel a given distance away from the original two fields and their interaction, these fragments travelling distance capabilities determine the magnetosphere region of the two interacting magnetic fields too.

The interaction of fields in creating plasma of fields needs a constant supply of fields from its environment, and that is why the fields can and do regurgitate some of the fields which was released by their own interactions and this being the reason why the Plasma of magnetic fields can last for a long time in the given environment.

In a way the attraction or the gravitational fields forces is created out of the necessity of the need for the two interacting fields to keep locked to each other, and hence the need for a constant sources of supply of magnetic fields in the universe, which this need is meet in large from a stronger source of fields and their unwinding of the principal mafs sources which are scattered around the universe. Where for example the larger plasma's the likes of the stars needs are meet by the cosmoses principal fields as they unwind and in the micro levels the need of weaker fields magnetic fields of atoms are meet by their lower level strength fields of their environment.

Where, at the same time the magnetic fields can enter a centrifugal region where their strength can be tightened and thus fields as much as they can lose their strength through unwinding, and then through the same process of rewinding they can increase their strength too. That is to say that if the mafs fields are entered into a winding region or in a tighter pact region fields strength of the universe, their magnetic fields can be winded up that the mafs strength increases to principal mafs field strength. Or what is said in the common language of present science the mafs can become of principal mafs strength and the centre of the tightening mafs is the principal centre of what is called at the present science the principal region of the environment, this being in the centres of the cell, galaxies or universes.

Thus as long as the environment can supply surplus amount of the pmf for a given system for its north and south poles to be maintained, then there will always be an entity existing, does not matter what word in the vocabulary is used to confirm its existence.

The mass is created when the field at the entry port and the exit port are linked and it is independent of any other entities of the same or there scale of mafs entities in the environment. Where, the mass is the amount of the fields needed from the environment for the interaction between two fields or plasma of its fields to exist.

- The magnetosphere is the limit of the field's exertion in a given environment created by the interaction of the gravitational and Magnetic fields of the two original fields. Where the magnetic fields are the cause and the gravitational and Magnetic fields are effects their interaction and the Magnetic fields interaction now becomes the cause of the creation of the mass and magnetosphere of the entity as the latter become the effect of the interaction of gravity and the Magnetic fields.

This being the reason why, that the mass is independent of the environmental conditions and effects, as the mass is the indication of the given internal interactions itself and is due to internal processes of interaction of the same internal magnetic fields. Thus when the environmental magnetic fields, gravitational and Magnetic fields are considered and applied to the given mass system, then the weight or the measure of the mass in respect to the environmental fields forces can be measured.

As the puzzle of creation of gravity is solved, this should open new horizon in the world of science and make life easier for scientists to understand the working of the universe. Where the whole world of creation can be said to start from the difference in condition in position, where this leads to creation of field that the different position can interact and be connected to each other, This known to be the origin of creation and origin of the creation of magnetic fields in the universe.

Where, this movement of and need for interaction of fields to cross the space of their environment leads to creation of what is come to be known as magnetic fields flow, and the extent of the flow to cover the strength difference and the proximity to each other trying to cross the gap to guaranty continuity of flow of fields in the environment become the current or the strength of the magnetic fields.

Where, the interaction of magnetic fields leads to creation of Magnetic (repulsion) or gravity (attraction), where the interaction of the created Magnetic fields and gravitational fields leads to creation of mass and Magnetosphere (1) for the magnetic fields concerned. Where the interaction of magnetic fields depending upon their strength leads to creation of stronger fields to be connected to each other, thus the stronger fields become the principal fields or feeders of mafs and the weaker fields become the matter mafs fields of the universe.

In the man's range of Magravs strength of his protein, the weaker mafs become tangible and visible or detectible and part of his own being's structure and his matter world is created and where in other levels due to stronger fields the principal mafs create a visible world for the one's whom their protein Magravs strength is set to a level beyond the man's mafs strength level.

Thus, the universe is the collection of and the creator of magnetic fields in different length and strength and their interactions to create different magnetosphere in size and strength in Magravs, and whenever and wherever there are fields to interact then they coagulate and linkup and create mafs and matters of different strength.

In the universe there are no empty places, which are void of magnetic, Magnetic and gravitational, mass and magnetosphere fields or not being in the path of the movement of fields. In general what is tangible and visible magnetic fields collection has been and it can be called matter and what is fields in the motion which are not yet attached to any matter to become part of given environment of matters or mafs are called energies of the universe.

The universe will never cease to exist as long as there are fields to attract and repel each other. Universe has no dark and black unknown phenomenon about it, it has principles and how they are applied and enforce themselves or it is enforced upon them, then a given condition and fields and mafs are created for given span of duration of fields interaction so called time.

The earth is not the unique environment which can provide and sustain life and intelligence, but life starts from fields and their interactions and upon their complex interactions and order they choose to be a club of plasmas, matter mafs, atoms, molecules, Magravsly to be bind (chemically), combination orderly repetitively Magravsly bind (biological) and creating communication in the biological Magravs repetitive order to guaranty the survival of their overall Magravs which these interactions lead to creation of intelligence.

Thus the man is not the only being with intelligence in the universe, as given the environment of interaction of fields entanglement is the facets of progression interaction of the fields which can be sustained everywhere in the unicos.

What is interesting is that one does not need the intelligence to have the soul, the same as one does not need to see the star in the far corners of the universe to know of its existence, but the plasmatic magnetic fields which it released during its existence it has travelled and nourished the universe to reach the position of the observes. This is the same for the soul of the man and the manner of it's creation and movement in all dimensions of existence, wise one now can understand the mystery of nothingness and of the creation of the soul of the man and the essence of its meagre existence in the world of universes.

The survival of the soul of the created being is the beauty of it, for it to give and to receive and not to be an outcast, then man of wisdom will enlighten the man of possession that there is nothing to possess if one uses what one needs and no more. Be wise on what has been said and open to be received and receive what is in your path of life in its different levels of interaction and attractions.

The ultimate point of understanding for the man who is one of the created entities in the universe and through what has been revealed on to man in the verities of disclosures in the recent past is that in the universe among its created whom have attained intelligence and have managed to take into spaces of transition environment of universes, that in the universe there are no animosities and no wars and nor this rule shall be changed as war and acts of aggression has been the habit of the passengers of the planet earth during their technological progress and their short lives on this planet.

The reason for the universal peace is not the lack of technology among its inhabitancies, but the truth is that the intelligence of the beings of the universe is so advanced that they have no reason to fight for what they can attain in and achieve and obtain from the universe.

Thus the universe is ruled by peace and justice and the ones whom have the habit of the like of the blue blood from the blue planet will be thought soon a lesson that unless unity takes over the thoughts of the creatures of this planet, then these shall be thought a lesson that they will soon fall and follow the rules of the universe and not the old habits of the feeble man.

Not forgetting that the everlasting habitant of the universe have attained the knowledge that can wipe-out any type of life from the universe, thus the peace in the universal order is not due to the weakness of its inhabitants, but their strength to be able to see any short sightedness from the newly joined members of the universal community that they cannot and shall not be allowed or even think they can carry on with their old habits of war and division and injustice and exploitation of the other habitants of the universe as they has become accustom to on the blue planet.

The laws of the universe shall be in the hand of the land of the lion as they have been graceful and has been just and as has been promised, the knowledge of universe shall come from the land of the lion and the Sun where it will shine upon the rest of humanity in equal measure and with true powers of grace and justice. This has been the command of the created and shall be attained.

CHAPTER 9

The creation of Gravity

© M.T. Keshe/The Keshe Foundation 2004/2006. All rights reserved.
Author: Ing. Mehran Tavakoli Keshe
Date of Internet release: October 8, 2006
Revised 1.9.2011

Abstract

The universe is built and held together and operates on the basis of two forces. One cannot work and exist without the other and when one of these two forces dominates the affects can be interesting, drastic or/and at times dramatic for the entity, this be it the universe, solar system, planet, atom or plasma. These two universal principal dynamically plasmaticaly magnetic field forces one being the Gravitational fields forces (attraction field forces of the entity) and the other the Magnetic fields forces (the repulsion or magnetosphere fields forces of the entity), which both have their origin within its centre these operating entities.

Hypothesis

The source of creation and sustainer of the gravitational field force and the Magnetic field forces of for example planets are the same elements and the same conditions in the centre of the planet.

Prototypes

To prove this concept, several static and dynamic prototypes over span of five years have been made and tested (Fig. 23), which the outcome of these tests shown system on the bases of this concept can be made that the system creates both gravitational and Magnetic fields forces, which in the operation of the system allows the system to reduce its weight first and then the system separates from its ground base and hovers within its environment. This proving that the concept of production of dual Magnetic and gravitational fields through the same elements and conditions to be correct, and the results and observation from these tests are summarized in general in this and other papers to be published.

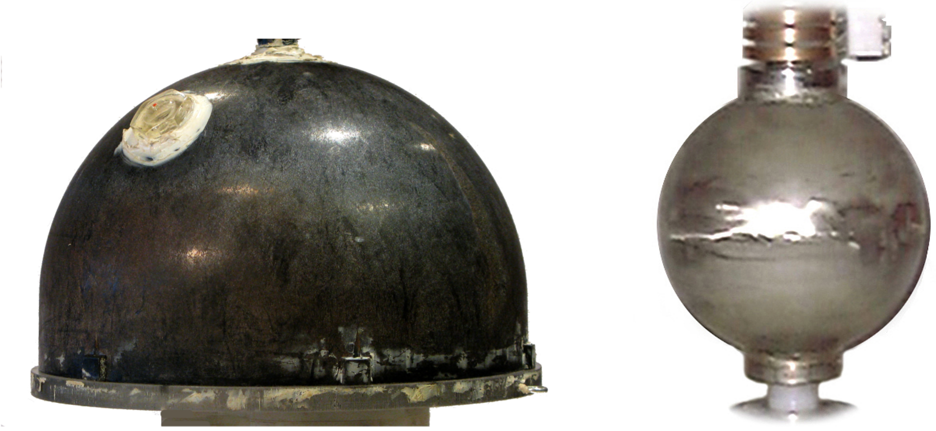

Fig. 23. Some of the prototypes used for creation of Magravs.

Discussions

In this paper for ease of understanding the structure of planet earth is considered as vast amount of information about the inner-structure of this planet are available. In fact this principle of creation of gravity through interacting magnetic fields can be scaled up to structure of universe or scaled down to structure of the plasma of proton and electron and the principle will still applies and be correct.

If scientists would have discovered both magnetic field forces and the gravitational field forces of the planet at the same time, they would have probably had a totally different understanding of their principles, their capabilities and their connections and most probably the human race would have taken a different course in its intellectual and scientific world. It seems this miscomprehension has been purely due to the physical time gap between the discovery and understanding of these two forces of creation.

Working on a real basis for gravitational field's forces production at the centre of the planet, the author set out in early 1980's to design and has produce several real and practical reactors (Fig. 23), which these reactors could replicate the universal order for generating magnetism and finally their interaction lead to the creation of Magnetic and gravitational field forces within the same system to prove the correctness of the principle of above hypothesis.

A fundamental assumption to achieve the proof of the creation and control of a gravitational force within a system has been that the Earth in its inner part has and maintains a semi-fusion and fission nuclear chain of reactions, also called a geo-nuclear reactor.

The interaction between materials and magnetic fields of semi-fusion and fission chain of events in the Caroline core (Fig. 24) of the planet (*This being the core within the inner-inner-core of the Earth, "The seed of the Earth"* (12)) always leads to creation and maintenance of planets physical forces of Magnetic field (MF) and gravitational fields (GF).

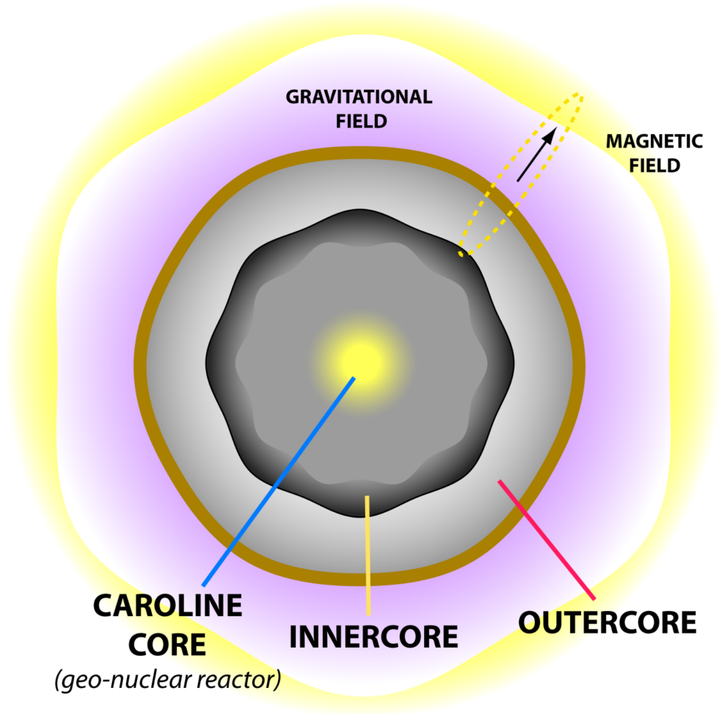

Fig. 24. The structure of inner cores of the Earth (12).

This semi-fusion and fission sequences of processes and existence of plasmas, nano-materials, gases and other materials in the containment of Caroline's' core and the convection process due to the heat transferred from the Caroline core to the Outer core of the planet once being passed through to the inner-core has been considered to be the principle operation of the inner cores of the planet, where, due to dynamic characteristic of the materials, which are charged and rotating within in vacuum condition of inside of the Caroline core, leads to independent creation of current and plasmatic magnetic field within this core, irrespective of the already assumed second and already known magnetic field, created between rotation of the inner-core and the outer-core of the planet.

These two independent dynamic plasmatic magnetic fields which are separated by means of the physical barrier made by the existence of the known solid inner-core.

The interaction of these two dissimilar poles if the these two dynamic plasmatic magnetic fields due to their proximity within the centre of the planet leads to the creation of the central gravitation field force of the planet, which this force spreads out to cover the whole of the structure of the planet.

Hence it is clear that the source of the Gravitational field force of the Earth and all planets are due to the interaction of two dynamic plasmatic magnetic energy fields, which are created independently in the centre of the planets.

The Caroline core of the Earth creates its own spherical dynamic plasmatic magnetic energy field, and the already known second field, comes from the dynamic spherical plasmatic magnetic field, which is created in the outer-core of the planet at the same time.

In trying to lock onto each other, these fields forces create the condition of double spherical magnetic fields, which both possess their own North-pole and the South-pole, where due to dynamic and loose atomic, molecular structure of the matters and fields in these two cores, the dissimilar poles of the two cores can never quite reach the point of total locking into each other, and as they try to reach this point of full lock, the field forces they create lead to creation and maintenance of the perpetual motion of the two fields in the centre of the planet.

These separated dynamic fields forces on either side of the solid-inner core exert their forces on the solid inner-core of the planet and cause it to rotate, and this leading to creation of the rotational movement of the planet (12)

At this point the interaction of the fields created by atoms and molecular elements, which cannot lock to each other due to their similarity of magnetic poles, will create the outward plasmatic magnetic field between the two cores, or they create what is known as the Magnetic field force of the planet (this being created by the environment and the behaviour of two similar poles of a solid magnet which they repel each other).

.

Where, at the same time the interaction of the magnetic fields created by atoms and molecular elements which cannot lock into each other due to their dissimilarity of magnetic poles, these interaction of the magnetic fields will create the pulling power effect between the fields, atoms and molecules contents of the two cores and this creates the dynamic attraction plasmatic magnetic fields of the planet or what is called the gravitational (Pulling) field force of the planet

Thus; Gravity or gravitational field force does not and cannot exist with existence of only one plasmatic magnetic field force as has been assumed to be so up to now in respect to the creation of the magnetic fields of the planet.

The gravity and gravitational field forces manifest themselves upon the creation and existence and maintenance of the second dynamic plasmatic magnetic field within the structure of any dynamic plasmatic field entity, that in the case of the earth fluid matter in the outer-core is source of this second field and may be generated through the assumed convection theory by the convection of the matter in the outer-core, and the interaction of fields in this core with the plasmatic magnetic field which was originally created in the Caroline core of the planet through the fluidity of the fields, plasmas and matters in this core.

Therefore the gravity is an outcome of the interaction at plasmatic, atomic or molecular levels between two dynamic plasmatic magnetic fields created and maintained by the two inner-cores of the planet upon each other.

Ultimately for creation and maintenance of gravitational field force, for all stars and planets or any dynamic maintain plasmatic magnetic fields system, and to maintain a gravitational fields forces in any celestial entity there is a need for existence and establishment of two independent dynamically, molecularly or/and combination of the two independent plasmatic magnetic fields forces, which these can be created by any mixture of matters and fields, in any combination of magnetic fields, liquid, gas, plasma or other state of matters within the confined of specific condition and entity, this leading to generation of loosely interlocking plasmatic magnetic fields, which these two fields are separated and maintained by any physical supranational means, in any state of matter or function, separately and independent from each other, even though one magnetic field could be the cause of the creation and maintenance of the second pmf.

Where, the separation of the plasmatic magnetic fields could be established by the means of solid, liquid and gases, plasma and energy directly in adjacent layers to each other as has been observed to be in the structure of the inner-core of the Sun material layers.

Once the strength of the plasmatic magnetic fields of the inner-inner-core (Caroline) and the outer-core in a system, by controlling the creation of the plasmatic magnetic fields in the Caroline core or the outer-core is attained by the system, consequently the strength of the gravitational field forces of the system like the earth is set.

Looking at overall view of the centre of the Earth magnetic fields and it their near spherical shapes structure and with the new understanding of its' physical inner construction and material composition, one will come to understand that at least the two magnetic fields generated within the inner cores structure of the Earth are spherical in shape and they generate dynamic spherical plasmatic magnetic fields. Where, one field being superimposed on top the other are in operation within the centre of the planet.

Whereas, one of the fields being the spherical magnetic fields maintained by the inner-inner-core due to existence of plasmas, hydrogen atoms, other gases, nano, gans, nuclear and non-nuclear materials in presences of second magnetic field generated by spherically dynamic plasmatic magnetic the outer-core in the centre of the Earth (Fig. 25).

Which, this creates and brings about the phenomenon of the dynamic magnetic confinement for the materials in the Caroline core, which this brings about the conditions for the fusion process to be attained in this core of the earth. This explaining why the planet can exist for a long time with its matters and motion intact, as matters through fission and fusion use a very little matter in the close capsule environment of the Caroline core.

The second spherical dynamic plasmatic magnetic field in the outer core of the earth possibly is created through convection, due to its materials in motion in the outer core.

There are strong indications and possibilities that the outer core could be operating in the same matter mixture composition as the Caroline core, which this eliminate and invalidates the concept of the convection theory.

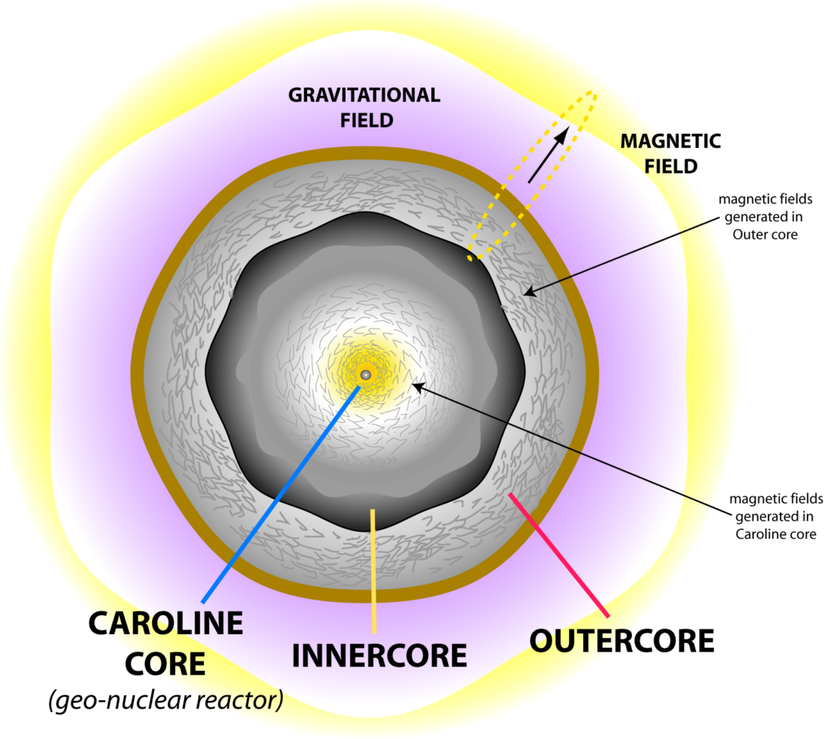

Fig. 25. Existence of minimum two separate magnetic fields in two cores of Earth.

It has to be said and it can be proven that the Caroline core maintains and through nuclear decay of its matter content produces plasma of matters, then by the understanding of the concept of generation of current through motion of the plasma itself, this proves that one does not need the convection principle to create the current and magnetic fields in the Caroline core or even in the outer core of the planet.

Therefore the existence convection theory and motion of the solid inner-core to create the magnetic fields of the earth losses its credibility, as in plasma technology, it is well understood that plasma in motions create their own current and in turn the interaction of this current with matters and other fields in their environment this can lead to they creation their own magnetic fields.

Thus, the theory of convection has to large extent no validity in the present understanding of nuclear science as one observes these operations of plasmas' phenomenon at the solar, galaxial and universal systems levels.

The convection theory was he understanding of the eighteenth century and the understanding of the working of dynamo of bicycles and this has no room in the integral working of the universal motion of plasma and magnetic fields of today.

Secondly entities like stars do not have solid core that leads to creation of their strong magnetic fields and gravitational field, as they only operate on the principle of plasma layers and some helium matter in, nano, atomic and molecular level.

Therefore in reality the physical positioning of the matter in the cavity of the once presumed solid sphere of the inner-core is the catalyst and contributor and the creation of both plasmatic magnetic fields of the planet, which both fields are independent of each other, but at the same time interconnecting or magnetically reconnecting tool for both spherically plasmatic magnetic fields on either side of the solid inner core of the earth to produce the Magnetic field and gravitational field of the Earth.

Where the interaction between these two dynamic plasmatic magnetic fields in the physical body of the inner-core of the earth leads to creation of the double spherical plasmatic magnetic fields' phenomenon needed for creation and control of the strength of the plastically magnetically based gravitational field forces and plasmatic Magnetic field forces necessary for the creation of the mass of and the atmosphere or magnetosphere around the planet.

These reconnections' of plasmatic magnetic field generated in two cores of the planet are totally on atomic and molecular levels, this leading to a creation of loose plasmatic magnetic field forces that operate at weak levels, but strong enough to cover a vast area in their surroundings.

That is why a small Caroline core in the centre of the Earth can create a strong magnetic fields effect of up to outer boundary of the magnetosphere and gravitational field of the Earth and beyond.

As the plasmatic magnetic fields could be created by use of any elements in both cores of the planets that they can generate any amount of heat and a given gravitational fields forces and for example with the Earth operating at central temperatures of several thousands of degrees centigrade at its core with surface gravitational force of one g at its surface, and similarly the Sun physically possessing several of thousands of times larger mass and volume than the Earth with central temperatures of several millions of degrees centigrade, which has only the surface gravitational force of about twenty eight times more than the Earth, therefore the following statement can be made that "*the strength of gravitational field force of any system is largely independent of temperature, and is primarily dependent on, the composition of its dynamic inner structure mafs and matters, and the fields they created or maintained and the physical condition and are positioned in respect to each other, which through their interactions these leads to creation of the two dynamic plasmatic magnetic fields strength of the gravitational fields and magnetic fields of the system within its inner sanctum*",

In relation to the physical data's available, the present gravitational field system which has been assumed to be created by the single magnetic system due to motion of matter and the inner-cores and the outer-core in the centre of the planet is not scientifically acceptable and cannot stand the present physical knowledge about the inner construction of the planet from seismological data's of the past twenty five years.

The concept of single plasmatic magnetic field of the planet, which has up to now been assumed does not stand scrutiny of present knowledge and the present expanded laws of physics, as for the planet to be able to create gravity and Magnetism (for magnetosphere) that can span a vast area of the solar system irrespective of its position and the material carried in its outer shells and on its surface, then the planet has to have two active plasmatic magnetic field forces in its centre cores, that the interaction between these two plasmatic magnetic fields can create the fields that affect the behaviour other entities' with dynamic plasmatic fields elements in the vicinity of these fields, and the interaction of the fields of the planet and the fields of the entities allows for the entities' to position themselves in respect to and within the plasmatic magnetic field strength of the planet to be attracted by the effective plasmatic magnetic field of the planet. This is what we have called Magnetic and gravitational position or for short Magravs positioning of one entity in respect to the other entity, where both entities possess internal dynamic plasmatic magnetic fields.

The reality is that, if Earth has had only one magnetic field as has been assumed until now, then all elements on the planet would have been firmly attached to it, due to principle of attraction or repulsion of two magnets, and as it is observed on the planets all systems have been able to freely positioned themselves and being able to be mobile.

Without the shadow of a doubt the centre of the planet maintains two independent plasmatic magnetic fields, that the interaction between these two plasmatic magnetic fields at the plasmatic, atomic or molecular level leads to creation of an environment in which the matters in the surrounding area of the two plasmatic magnetic field will have the power to attract and repel mafs and matter within this created field. By the same law as the atomic matter, itself is made of magnetic energy fields, which possesses polarity hence why the attachment of the man to the earth as his being is made of these entities.

Therefore the matter will only by finding its Magravs position within the plasmatic magnetic field of the planet, if it can be attracted and be kept as part of the matter of the planet, or if its plasmatic magnetic field is too weak or too strong to be kept within a given field. Thus now it can be clearly understood how and why different planets can manage to hold on to certain matters and not the others and now one can understand why different matters take their positions in different layers of the planet too, where this positioning, like of the gases and solids, are totally related to the strength of the plasmatic magnetic energy of the atom in respect to the plasmatic magnetic energy of the planet or in other word, in respect to the total interaction between the pmf of mafs of the matter and the Magnetism and gravity of the planet.

The plasmatic magnetic field force created on these levels in the physical inner cores body of the planet these creates the condition, where all magnetic forces within these fields will be influenced through their magnetic reconnection, this being in atomic or molecular level, this creating pulling inwards forces for the two plasmatic magnetic fields of the matter and the planet to be locked into each other. This interlocking and the strength of it bind creates the effect of the gravitation field force, as the matters' magnetic force and the planet plasmatic magnetic field created by the double plasmatic magnetic energy field, reconnect to keep each other at a physical interactive but loose distance.

In reality gravity is the strength of the plasmatic magnetic field created by the interaction of the double dynamic magnetic force fields of the planet that can influence to pull the matters around these field to stay within its strength, or what is called its gravitational field force.

The interlocking of these forces are the same as with two bars of magnets when they are brought close to each other, but their binding of forces is in planetary plasmatic levels, which are at a loose atomic or molecular level rather than electrons' realignment or what is called are energy equal to the collective plasmatic magnetic energy of the electrons of Iron atoms interlocking plasmatic magnetic field level.

With the difference that in the planetary or astronomical, cosmological and universal levels these so called magnets are in the form of closed loop spherical shape and may be in possession of internal heating in respect to their surrounding temperature and system, where their plasmatic magnetic field force can vary and they can change their strength due to the compression and other factors in their centre cores, as in solid magnet the strength of the magnet is constant and permanent.

In the closed loop spherical plasmatic magnets, the system of the plasmatic magnetic field lines has to find their magnetic reconnection line for the loop to be completed. Where, this plasmatic magnetic field's reconnection leads to the directional polarity leading to the creation of the North and South Pole phenomenon of the planet.

Through replication and production of such a system as the structure of the inner-cores of the earth, in being able to control the amount of matter and fields within the structure of these systems, the strength of the plasmatic magnetic fields of the cores of the system can be controlled and adjusted, so that, this will allow the creation and control of the measured gravitational fields forces by the systems in respect to the environment in which the system is operating in.

In the case of motion of an object like a manmade system, in possession of its own gravitational power production, in the gravitational pull of another object like a planet, the use of the phenomenon of the creation and control of the gravity in the centre of any system can be treated in the same manner as in the planetary system.

Where, the gravity created in the centre of manmade system needs to be slightly changed in the positive or negative direction in respect to its previous position for the object to be repelled or be attracted towards the planet or an object.

This Magravs positioning of a system in respect to planets plasmatic magnetic fields, can be achieved through several means, like by the change of characteristic of the dynamic of the one core in respect to the other, or changes configurations of the parameters, materials or physical structure aspects of one core or both core of the system.

The magnetic and gravitational forces of a Star are to some extent dictated by the position of planets within the system as much as the constituent content matters of the Star itself. On these bases, this being the gravitational forces of the system and its position within its galaxy is to a large extent dictated and controlled by its' planets and the gravitational force they create independently on their own.

This is very much like the rotational plasmatic magnetic fields of the electrons circulating around the nucleus of an atom. Where, the magnetic rotational spin of the electron decides the plasmatic magnetic field strength of the nucleus and the position which the element occupies and the number of other electrons it can carry, so this dictates the properties and its characteristics in respect to other elements around it, even though the plasmatic magnetic field of the nucleus is much more powerful than the plasmatic magnetic field of the electrons around it, or in other word the gravitational forces of the nucleus is needed for the electrons to be kept in their orbit and vice versa.

Thus the gravitational forces of a system or a galaxy is not only dependent on the plasmatic magnetic energy forces of its centre or its star, but it is fully controlled and determined by a bigger scale by its constituent planets and stars in that system too.

From the topography of the Earth gravitational field pattern provided by space agencies, Earth plasmatic magnetic field does not keep a constant pattern and there are constant changes in the strength of these fields in different parts of the planet, where these are dictated by the activities within the centre cores of the planet at the point of observation. These variations in gravitational field's patterns confirming that the planet has composite and not necessarily spherical shape cores.

These variations in these forces in the cores of the planet could be due to the variation between the interaction of the inner and the outer plasmatic magnetic fields, where for example at point where these forces vary drastically in respect to each other, then one gets distortion of the fields and reduction or increase in the gravitational field force in that region of the planet by small factor. Then where, for example one force field, like the fields created in the Caroline core is more powerful due to sudden short burst of ionization in the core, at that point and that moment, the plasmatic magnetic fields interaction of the centre field is more powerful, then the gravitational field force can be more powerful at that given point, but these changes due to close confinement of the centre cores are usually not so large.

For example by observing the pattern of the gravitation field of the Earth, a constant depression in the plasmatic magnetic field in the region of the southern eastern hemisphere in respect to the rotation of the planet usually can be observed. Where, this phenomenon now can be easily be explained, and this is what I so call the washing machine effect and that is, where the material within the inner-core has loose plasma area of rotation, as in the drum of washing machine filled up with clothes, the inner-core of the Earth does not possess the blades as in the drum of the washing machine, then the more dense matters in the Caroline core as they cannot role all the way up against the internal wall of the cores, and then due to toroidal and interaction with poloidal motion of the Earth, these materials pull up the wall of the core and due to their viscosity they roll back over in this area, this whirlpool effect due to slower motion and less available field in this region effects the gravitational field in the south American continent area too.

This washing machine effect of the materials in the Caroline core directly affect the plasmatic magnetic field production in this core, and consequently its interaction with the plasmatic magnetic field of the outer core, and in turn this creates the disturbance in the gravitation field of the planets in these two regions as has been noted in the overall pictures of these regions gravitational field forces effect, where dents in the gravitational field forces patterns are observed on continues bases.

Even thou Newton did note the effect of interaction of two forces between the matters and the planet and he measured its strength and called it gravitational field's forces of the planet, but he could never establish the source and method of its creation. He speaks about the effects of the gravity in relationship between two objects in respect to each other's position, and not the effect created by two plasmatic magnetic forces in respect to each other and their position in respect to each other, where he could not understand that gravity was magnetic field bases entity and magnetic field strength and it effect are magnetic fields position dependent. Newton never could relate and understand the connection between Magnetism and the gravity as both to be created by the same matters and conditions which manifesting itself as two different forces in respect to other magnetic field forces.

Scientists explain the gravity as the relationship of attraction between position of two bodies in respect to each other, and not the reality that gravity manifest itself in relation to interaction and attraction of two bodies in possession of dynamic plasmatic magnetic field forces created and maintained in respect to each other position.

In understanding the real method by which the gravitational field forces of planetary systems are created then systems can be made to replicate the same process and by being able to create and control gravity as done in the planets one is able to observe the effect of this natural phenomenon within operation of a manmade system.

In manmade reactors as the concept of gravity has been replicated, now the concept of the gravity has to be redefined as the " attraction between two bodies, where at least one body possesses a double spherical dynamic plasmatic magnetic field force, where one field is superimposed by the other plasmatic magnetic field forces within an object, where their overall plasmatic magnetic field strength interaction upon each other can create an overall plasmatic magnetic field, that other matters in possession of their own dynamic plasmatic magnetic field or static plasmatic magnetic field, within their catchments area can position themselves according to their internal plasmatic magnetic fields strength".

The interaction of these double dynamic spherical superimposed plasmatic magnetic fields in the centre of the planet separated by the physical matter of the inner-core upon each other is the cause of the rotational motion of the inner solid core of the Earth and where the motion of the inner solid core itself is the containment body of the centrifuge system for the gases in the Caroline core for the atomic fusion chain of events in this chamber, for it to sustain the ionization process in the this chamber. Where, at the same time the inner-core plays the effective mixing machine for the matter in the outer-core of the planet, where the interaction of plasmatic magnetic field of the Caroline core with interaction with the plasmatic magnetic field of the outer-core would change the combined plasmatic magnetic field effect of the whole planet, which in reality would have changed the gravitational field strength of the planet.

The rotation in the gas layers of the Caroline core in conjunction with turbulence, which leads to convection in the materials in the cores of the planet, this effecting the generation of current and plasmatic magnetic field in the core of the planet and consequently the gravitational forces created by the core of the system.

Now one can explain that the change of the polarity of the earth and any other entity the like of the Sun is only due to change direction of flow of the current in one core or layer of these dynamic magnetic fields based star and hence the change of direction of the flow of the current for example in the plasma of the Sun layer leads to change of the polarity of the star, as one can sees the same change of polarity in the production solid magnets, where the change of direction of the flow of current changes the polarity of the metal, as is the same for the plasmatic magnetic field entities too.

Thus change of polarity of the earth is only due to the change of direction of the current within the core(s) of the earth and would not affect the gravitation field or Magnetic field strength of the earth when it will happens.

Now that the method of the creation and control of gravitational field forces of planet is understood and can be replicated, it becomes easy to understand how systems can be produced that can create magnetic field structure of any element or matter in the universe, this being magnet for hydrogen atom or uranium atom or magnetic machines can be produced that can gravitated to themselves molecular structure like water or protein.

Where, the system gravitational field can be created to match for example only the plasmatic magnetic field of atom of gold, so that the system becomes a gold magnet or the system will generate only or gravitates towards the plasmatic magnetic energy of the gold atom, so one can separate gold or any other matter which the system is present to generate the Plasmatic magnetic energy of. This system will then can replace the present centrifuge system of the present time.

Magravs measuring system

In the gravitational systems, where only magnetic fields forces are considered and measured, the present earthly measurement tolls of gravity has no meaning and room. Space scientists of the future have to learn to measure magnetic field force in respect to each other and not kilogram per centimetre per Seconds Square.

Thus the measuring gravitational systems to be developed have to be first of all three dimensional, as field forces of all matters and fields in space have to be considered and balanced against or with.

The gravitational tools of the space technology is fixed magnetic fields shear against the magnetic fields of the system against the total magnetic fields forces in the overall sphere of the systems fields.

In fact the gravitational fields of the space technology are like a magnetic fields sphere positioned around the gravitational reactors, as these measuring tools of the gravitational systems will measure the Magravs of the all fields of all forces internally and externally in respect to their position in respect to the reactor Magravs.

The Magravs measuring tools are to be called **Universal Magravs Measuring System** (UMMS) and position of Magravs of the system in respect to another Magravs are called **Magravs positioning System** (MPS).

Elaborations:

Through understanding the real term concepts of the creation of the gravitational field forces and design and production of specific nuclear-based reactors of the function of matters within static and dynamic reactors, the principle of the hypothesis of creation of gravity has been proven to be correct.

The difference between inertia and gravity is said to be the gravity is an attribute of a body in possession of two dynamic plasmatic magnetic field forces interacting upon each other, where the inertia is the attribute of collective molecular magnetic force of an object which does not possess a double magnetic field structure as whole, even though both behave the same and due to their plasmatic magnetic fields appear to be the same thing.

Where, in case of planets like Earth, the total pulling power of the planet in respect to other objects in, on, or about it are due to collective forces created by the gravity and its constituent's matters inertia.

The creation and maintenance of gravity is dependent on two dynamic regions being in possession of two independent dynamic plasmatic magnetic field forces and gravity is totally independent of the mass of the object as mass itself is the by-product of the interaction of the Magnetic fields and gravitational fields of the entity, but dependent on the position of their plasmatic magnetic forces in respect to their plasmatic magnetic field strength from the centre entity.

The centre of planets, does not matter what their inner layers are made of do not need a metallic matter or metallic behaviour conditioned matter for the creation and maintenance of their own electromagnetic effect due to motion of electrons and charged plasmas in their core as these entities generate current and potential difference levels for the core to maintain independent plasmatic magnetic field forces leading to the creation of plasmatic magnetic field and finally by interaction of two such dynamic plasmatic field forces upon each other, leading to creation of gravitational field forces within the entity and in its surrounding area.

The interaction of these two plasmatic magnetic field forces upon each other, or the concept of *"DOUBLE DYNAMIC SPHERICAL PLASMATIC MAGNETIC ENERGY FIELD EFFECT",* in the centre of any dynamic plasmatic entity will lead to creation of the gravitational field force and the magnetic field forces of that entity in respect to other entities with the catchments fields force of the entity.

Therefore, the gravitational field strength of a system manifest itself and can be explained as the relation to interaction and attraction of two bodies in possession of dynamic plasmatic magnetic field forces created in respect to each other plasmatic magnetic field position and strength. Even, further it has been proven through prototype reactor tests that the speed of the rotation of the planet is determined by the speed of the solid inner-core of the planet, where the speed of the rotation of the planet is set and dictated by the interaction between the dynamic plasmatic magnetic fields of the Caroline core and the outer-core of the planet, as the motion induced by the strength of the two fields in the centre cores, upon and interacting within the solid inner core, determines the speed of the rotation of the inner core, and in turn this will determine the speed of the rotation of the whole planet.

Whereas, if the speed of the rotation of the planet were not set by the planet by this principle, then the planet would have had different speeds at different points in its daily axial rotation and it would have had different speeds in its annual path in circling its star, where the effect of the other objects in the system would have dictated different speed of the rotation of the planet, therefore it is correct to assume that the speed of the planet is controlled by the speed of the solid inner-core of the planet, and the same applies with other entities like the star as the speed of the internal fields interaction dictates their speed of rotation.

The motion of the inner solid core is very much like the motion of the trains on magnetic tracks, where by not allowing the two dissimilar magnetic poles to be aliened, the train moves forward. Where, in the case of the spherical solid metallic inner-core of the Earth, this motion is rotational rather than linear.

Through tests, it has been proven and understood for the first time, that the magnetic polarity change of planet or the system occurs when the direction of the flow of current this being due to motion of electron or charged plasma in the cores in the whole of the system changes from one direction to the other, or in some cases one field change of direction of internal current can causes the half reversal of the polarity of the whole system. This is understandable, as the plasmatic magnetic fields are created due to directional motion of electrons, plasmas or charged plasmatic current within the matters of any entity, thus as the direction of the flow of the current changes, this will change the direction of the polarity of the planet.

Understanding the direct relationship between the magnetism and gravitational forces within a system, in the manmade systems the change of polarity in the reactors of a system has to be avoided, as, if this happens during the operation of the reactor, this can and would lead to the overturning of any platform which is equipped with such a system. This being one of the fundamental reasons why the reactors which have been developed and tested up to now for space technology, always are preferred to be constructed in a half spherical shape, simply to avoid the flipping of the polarity in the cores of the reactor, where the speed of the rate of the polarity change is very dependent on the components and conditions created within of the centre of a planet or a star.

If a planet is made of near enough similar material on both side of the inner core, the flips of polarity will be much faster and on a shorter time lapse, like in the Sun.

If the planet is made of a composite core, and made of two or more different materials in its Caroline core (the real centre core), for the two fields to re-align, then this will take a longer time for the magnetic flip of polarity to occur, like in the core of Earth. In any case, as the materials in both cores are always constant and cannot change that much, therefore the time laps between the changes of polarity stays on the same time intervals.

In the case of the Sun, due to its more uniform and singularity or only composite of three matters of plasma, hydrogen and helium in its core, the change of polarity takes eleven years, and the full polarity change has the predetermined time of twenty-two or three years.

In the case of smaller planets' like the Earth, which is made of composite mixtures cores, the time lapse in change of polarity will be longer. Where, the switching of polarity is a natural outcome of the direction of the flow of current created by the electrons and primarily by the charged plasma and due to dynamic characteristic of the core, and is an inherent characteristic of the atomic structure of each element and environment which they are condition to within the cores of the planet.

It has been observed in tests, that the plasmatic magnetic field of the core increases by the principle of extraction of the electron from the core, where electrons are facilitated to be absorbed by the walls of the reactor core. Where, the majority of the field created is primarily due to the motion of charged plasmas. Where, at the same time for weaker magnetic fields test the reactors have present to absorb the plasma of the proton that only the lesser mass of mafs available in the plasma of the electron can be used for a given test condition.

Therefore according to need of the use, reactors can be set to utilize any components of an atom or the mafs of their constituent plasmas without in fact destroying any structure of the mafs or the matter, which the mafs was used for that given exercise.

In principle it is observed that the plasmatic magnetic energy of the planet is primarily due to the motion of the charged plasma in the core and not electrons, where in reality the motion of charged dynamic nuclease of an atom has/ and indeed, possess a much larger electrical charge then its counterpart the electron by factor of 2000 times. Thus now it is apparent and clear why such a small core in planets can master such large magnetic field forces within and around them, and now it is understood how planets can carry the burden of such large mass in their journey through their system for such long period of time of billions of years in their solar systems.

In tests plasmatic magnetic field created by these type of cold condition plasma using a 20 cm dynamic core, has been noted to create a 4-meter radius dynamic spherical plasmatic magnetic field span, which can last up to several minutes after the switching off the whole system.

The dynamic plasmatic magnetic fields created within the double cores of the planets, determines the magnetic or gravitation fields, which the planet will possess and consequently the material, which the planet will hold on to. This explaining why planets behave like magnets, which only have the characteristic of one element or several materials plasmatic magnetic band(s) which covers some and not the other elements, which this dictates the material composition of the planet within, on the surface and around the planetary system. Where, the same applies to the solar system and galaxies and so forth. Where, the gravitational field forces created by the balance of the two interacting dynamic plasmatic magnetic fields is the principal factor in determining the constituent material of a plasma, planet, star or a system.

The creation of the gravitational field forces of a solar system and even the gravitational field force of the galaxies are not far from the gravitational system and method's used in the atomic and planetary systems. That is to say, that solar system’s possess and generate plasmatic magnetic fields like planets and they possess gravitational forces according to their stars constituent materials. If one could stand on the edges of the solar systems, one should be able to see and detect the gravitational forces of this system, by it simply absorbing and attracting with the entry of cosmic meteoroids and asteroids into its gravitational and solar heliosphere region of its solar system.

Solar systems, like planets, possess and create magnetospheres of their own which is specific to their composition materials. The entry and exit from these solar Magnetic and gravitational fields uses the same method as the entry into or exit out of planetary systems.

This also holds true for a solid magnet, where if a piece of magnet could be cut into two halves or four quarters, each piece of the material will still carry the same north and south poles and each piece will have the same magnetic powers as the original piece. That is to say, that if one uses the same volume ratio of the same mafs or matters and proportional physical conditions in the core of a system; one will attain the same magnetism and gravitational field force from the system as its bigger dimension system with the same mass ratios.

In astrophysical terms, the Magnetic field forces of any entities are what keeps the entities apart from each other, and at the same and opposite to this, the gravitation filed forces are what keeps them together.

Where, the pattern of distribution of Magnetic and gravitational fields forces of the entity interacting with another Magnetic and gravitational field force shows itself as the magnetosphere of the entity.

This Magnetic and gravitational field force (Magravs) of dynamic plasmatic fields force is known passive magnetic field (PMF) and has no relation with the magnetism as known in a solid-state magnet. This is so called passive plasmatic magnetism due to the fact that, this type of magnetism cannot magnetize an entity otherwise all the materials in the universe would have stuck to each other. But this mixture of Magnetic and gravitational fields are used to position one entity in respect to each other that they not necessarily collide or move too far from each other once they established inter-magnetic connection.

The reason the Magravs are called passive plasmatic magnetic field is due to their capability of passing through human tissue without in majority of the cases affecting the function of their atomic and molecular magnetic field, which human cell possess on their own for their structural bindings.

In the past, magnetism and gravity have been explained according to the physical understanding of a solid magnet and not according to the planetary magnetic forces, this new understanding of creation of magnetism and gravity, is one of the corner stones of the creation of plasmas, planets, solar systems and galaxies. By not appreciating the connection between the magnetic field force and gravitational field force of a system like Earth, it has become a challenge for the scientific world even to overcome the weak gravitational field forces of small planet like earth in just trying to send a small rocket into the Earth's orbit.

Through development and tests on double core type of cold condition plasma reactors creating the same conditions as it has been assumed to exist in the centre of planets, it has been observed and physically been proven, that the source of the three dimensional spherical gravitational field and three dimensional spherical magnetic field of the planets are the same materials and the dynamic interaction of plasmatic magnetic fields created individually and independently in each core of these reactors and their effects upon each other creates and maintains both forces of the creation Magnetic field and gravitational fields forces simultaneously.

If one were able to create the right condition and material in a simple core, as has been done in our tests with core developments, then one can create and control magnetism and then the gravitational field forces within that object, and then in the environment which the object is positioned in respect to other external plasmatic magnetic fields around it, where, the plasmatic magnetic fields created within and by the object will becomes its Magnetic field around it, very much like the magnetosphere of the planets, this providing a magnetic shielding like Earth atmosphere around the object and due to the spherical shape of objects in possession of both forces, this creates a magnetic loop, where they create the two poles of the magnet as the North and the South Poles.

If a system can take advantage of this phenomenon, for it to move within forces of a planet, rather than trying to overcome them, or to resist them for its motion, then not much energy or effort is needed to create the right gravitational forces and magnetic field protection within and around a system for it to move inside or outside the atmosphere of a planet or a system without burning a single atom as a fuel, which this puts an end to the propulsion technology for rapid and protected motion anywhere in the universe.

Due to the spherical shape of most of celestial entities and their rotational motions in the universe, the gravitational force created by the dynamic plasmatic magnetic effect has a pulling force, which has the tendency to encircle and reconnect within the interior of the entities fields of the same dynamic structure. This being the reason why the gravitational forces increase as one moves closer to these objects and as one moves' towards the centre of these matters.

The difference between the solid magnetism and plasmatic geomagnetism are two clear factors, in this particular order, first is Dynamism of the magnetic fields which creates the geomagnetism, and second the need for the interaction of two independent fields upon each other to maintain the plasmatic magnetise, where plasmatic magnetism for it to exist the plasma needs to create and maintain a full gravitational and Magnetic fields system of its own, which can absorb and feed other plasmatic magnetic fields within its environment to guaranty its survival in that environment at the same time, whereas in solid magnet, the order of continuity is within the structure of the matter and not mafs of the matter.

Conclusion

The gravity is the product of interactions of fields and wherever in the universe fields will interact their interactions will gravitates some fields or matters to them, or caused these fields to be attracted and gravitate towards another setup of the same or stronger and by laws of nature the stronger attract and gravitates more to itself and the heavier mass as they are made of more magnetic fields tend to dominate the seen when it comes to gravitational pull. Through tests it is apparent that the power of Magnetism and gravity of all entities in the universe are independent of the physical size of the entity and more dependent on the strength of mafs, which have led to the creation of the entity.

Through design and test of double core reactors we have proven that the origin of the gravitational force and magnetic force of a planet or any celestial systems in the universe are the same mafs and materials and conditions, where they are created out of the interaction of two independent and dynamic plasmatic magnetic fields.

Note; this revised version of the creation of gravity paper has included extracts from collection of three other papers published by the author, which these were, the gravitation measurement system, the relation between Magnetisem and gravity (35) and the original paper of the creation of the gravity.

CHAPTER 10

The creation of Black Hole

© M.T. Keshe/The Keshe Foundation 2004/2006. All rights reserved.
Author: Nuc Ir. Mehran Tavakoli Keshe
Date of Internet release: December 3, 2006
Up dated version for this book publication 2.9.2011

Note:

The original version of this paper, which explains how black holes are created and how they do leak matters and rays into the environment of the universe, was submitted to the Royal Astronomical Society of United Kingdom in London in April of 2004 for peer review and publication. The author after three months received on the 8.7.2004 a letter (Fig. 29) from the Society dated 7.7.2004 that this paper will not be published and will be kept in the library of the Society.

Interesting enough on the same day 8.7.2004 (check the press) Mr Steven Hawking, the imminent Black hole scientist and expert in this field of science announced that he, after thirty years has changed his mind and he says that indeed the black holes do leak rays out and that he announce this in a conference in Dublin later that month (see the press), as now he has lost a bit of shirt.

We have wondered for years how coincidental of the topic, the date of letter from the institute, the date of announcement, and the change of mind of a scientist!

This is the reason why the Keshe Foundation does not believe in the scandals and process of peer review and the foundation publishes its own scientific papers on its website and its own books for the reader to peer review the newly gain knowledge. As with this process the knowledge is passed on first hand and in some cases in matters of hours of the discovery to the public and not waiting for month and years until there is gap in some journal that the paper can be published. The Keshe Foundation firmly believes that 'as if a man, has reached the age of maturity and intelligence that he can read, then he can independently investigate and judge the correctness of the knowledge he receives according to his own level of intelligence for himself", unless he is an innocent, then he needs to accept the enforced control system of peer reviewing to tell him what to think, that this peer reviewing can suit the self-interest of the peer reviewer, or organisation which supports him financially or the nation which has promoted the peer reviewer for their own clandestine purposes, as we have seen the same process of control through peer review systems in different occasions in different countries around the world in this recent times.

A copy of the letter from the Royal Institute of Astronomical Society is enclosed at the back of this chapter for reference. (Fig. 29).

Abstract

Since first disclosures of this phenomenon, scientists and the rest of academic world have tried to understand and explain the reasons for the existence of these mysterious Black spots in the universe.

Hypothesis

The object of this paper is to prove that Black Holes are created out of a natural sequence of events in the universe and to prove that existence of a Black Holes in the galaxies and in the universe is a normal event as the appearance of the Dark Spots on the surface of the Sun. Black holes are simply a creation or an outcome of natural interaction between two or more mafs, passive magnetic fields and matters in motion within any dynamic plasmatic magnetic field entity in the universe (Fig. 26).

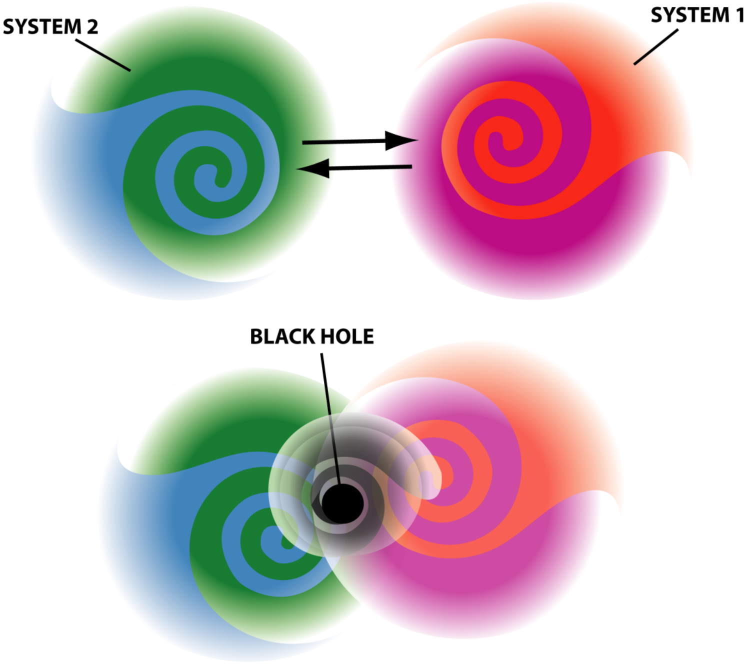

Fig. 26 Interaction of two plasmatic energy fields creating the effect of a Black Hole.

Discussion

In this paper for the first time, it will be explained how the Black Hole is physically created and in this disclosure it will be explained

1) The conditions, which this physical object is created in the universe naturally.

2) How this body of the mass appears as a dark spot in the galaxy?

3) What purpose it serves in the universe?

4) Parallel example of the Black Hole condition in our solar system.

5) The use of this phenomenon for Man to enable him to travel in space.

When, one could fully understand the concepts of the creation of galaxies, stars and planets and plasmas, and then one could be able to understand the creation and existence of the Black Hole in the universe in its different scale of dimension, where the creation of these entities in the universe all follow the same principle.

In understanding the fundamental principle that galaxies like solar systems and planets are given their inner mass and energy at the point of their birth in the universe, so they cannot create additional new masses and new source of energy on their own. Then it is understandable that Galaxies can only recycle what they have been given at the start of their lives, to create different new objects like stars or planets within themselves. In universal terms they have a finite mass and energy available to them and nothing more. But at the same time as a part of bigger picture galaxies have to obey the universal laws of physics. Consequently, all galaxies follow the same rules and so do their physical contents like solar systems and atoms within them.

A Black Hole is created primarily and only under-one condition in the Universe and as the consequences of other acts in a galaxy. They are not created as a deliberate act and as predetermined plan in and by a galaxy. But in their way, they make an enormous contribution in keeping the matters in check in their surrounding environment during their relatively short life.

At the same time they redistribute fresh materials throughout the galaxy during their cycle of existence. This is done through and as the consequent of them being the source for the creation of cosmic rays in their galaxies, as this is very much like natural fertilizers production systems of galaxies. Where, the old matters give its energy in an indirect way, for the existing other planets, to keep them healthy and for them to receive sustenance from their galaxy.

Black Holes in the universe are nothing but the centre for collection and re-cycle waste bin of cosmic debris of different sizes, this being giant meteoroids to specs of cosmic dust.

As the black hole grows bigger and their gravitational forces take over, then a Black Hole becomes a predator for parts of this galaxy. They become vast eating machines of the debris and stars and planets and this is exactly what is to be expected of them.

The Black Hole is originally created totally by chance. There are no patterns in their creation. They are simply a creation or an outcome of natural interaction between two or more passive magnetic fields or matters in motion within the galaxy or the universe.

Creation of the Black Hole

To explain the concept of creation of black holes, phenomenon's that are normal and familiar are used and will be relied upon in this paper. Then a parallel for the creation of the Black Hole type of incidents will be drawn with the same process in the solar systems, galaxies and universe.

Black Holes are created according to the rules of physics in galaxies to collect and then replenish the galaxies with fresh recycled materials, where this recycling is the vital factor in sustaining and making sure that as the galaxy moves that there are new materials available to start new systems within the galaxies. At the same time for them to nourish the other planets within the galaxy with fresh supply of materials. Where, for example the Earth receives several thousands of tons of cosmic dust from the solar system through the similar method through its atmosphere per year from the black hole within its galaxy and from the universe.

The parallel similarities between the Black Hole and presentation of new material in the universe and on Earth can be compared to the effect of collision of Earth's tectonic plates. When, two sources of solid materials interact with each other. In the long run stars and planets in the galaxy, as the same with Earth submerged tectonic plates; these become the source of the supply of new materials in the galaxy, similar to, as lava becomes the new continents, or mountains on the Earth.

Let's explain this in the following manner. The continental plates and stars systems are forever on the move, they move slowly, but surly they are on the move. On the Earth, when two or more continental plates move in random manners, over centuries and millions of years, at some point they come on each other's path so as do the stars and their systems, like solar system and others in our galaxy.

When on Earth the two or more continental plates come together. Several things can happen, this happens in galaxies but in a different form too. Where, the continental plates in galaxies are replaced by cosmic spherical or flat oval magnetic fields of the stars or systems.

The next steps in the procedure are very familiar on Earth as the continental plates push against each other, if, they are of similar strength, they rise together to create the mountains over millions of years as these plates are still pushing and moving while they are exerting their strength on each other. So, large mountain ranges on the surface of planets are created.

Other things can happen too and that is if, one of the continental plates has weaker energy strength the weaker plate will slide under the other, where the continental plate which sub-merges in the long run becomes the material in the lower levels inside the planet, which later this becomes the base material for the magma of the planet, then over millions of years this matter again appears on the surface of the Earth through volcanic eruption. Then some times in the future these too become the source of materials for new continents or an island and so on through volcanic eruptions.

The material drawn into centre of the planet does not disappear but for the time being is hidden from the view and parallel to this scenario happens in the galaxies, when two or more systems in the galaxy reach each other, where an area in the space is created that the two passive magnetic fields and or entities have to interact with each other. Where, their passive magnetic fields push against each other like continental plates (Fig. 27).

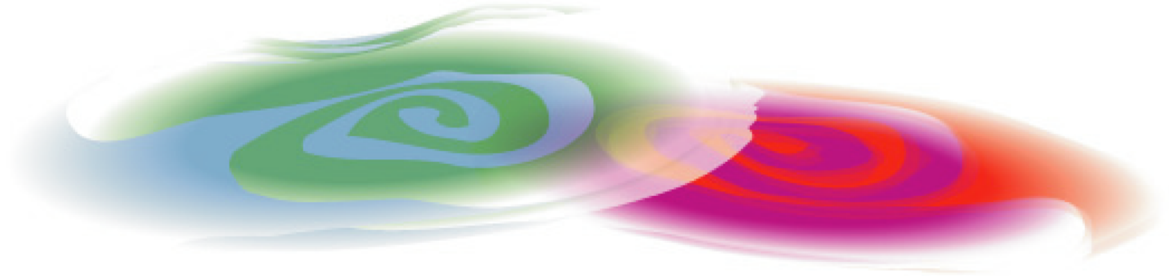

Fig. 27. Plasma of passive magnetic fields pushing each other creating
a balanced magnetic field region, called Black Hole

As they are on a collision path with each other and as at the same time the surrounding systems forces are pushing for their position, they do not allow any other course of motion but path of the collision of the two systems on the move. Then, a point is reached that as there are solid materials in motion from both systems and they have no choice but to collide and interact with each other. Where, in this process being thrown at each other by their systems, pieces of inter-planetary garbage, parts of moons or meteoroids and other bits of the two systems in this region are in the same pot too and all are mixed up and get entangled with each other.

In this area and at the same time as the two systems physical materials are pushed and thrown at each other, so do their mafs and passive magnetic fields are exerting their powers and authority too as in the like of in the planetary the hidden forces pushing the continental plates into each other. Then a point is reached, where these two magnetic force fields' of the two set of entities forces become to near enough to cancel each other out and create a condition of magnetic field equilibrium, where one field force cannot dominate the other passive magnetic field force, and inevitably an area of no or very week passive magnetic fields force in this region of the universe is established, where during all this processes materials from both systems are still being pile-up in this part of space. Where, the two systems materials are brought together due to pure forces of the Passive Magnetic Field forces of the systems and their surrounding systems in the galaxy.

Even stars themselves in the latter stages of the development of the Black Hole could become the victims of the Black Hole, which their system themselves were partially the cause of creation of. At this point the new mass of the Black Hole may even be a larger mass than any of the individual systems that created the Black Hole in the first instance.

One cannot forget that these masses are still in motion and rotating at very high speeds.

This new mass in the centre of this is pile all together creates a very large gravitational force within its region due to sheer weight of the mass which is collected and primarily these could become inertial pull rather than gravitational pull of the matter in these Black Holes too.

By laws of physics, one knows that the only reason that stars shine or cosmic rays glow is the phenomenon of the motion of charged particles passing through a magnetic field. In this region where the magnetic fields are very weak, this region lacks the main component that could make it shine or to be bright and that is the lack of the passive magnetic field in this region, due to the equilibrium and / or cancelling out of the two interacting magnetic field forces of the two systems.

Or for a time in the motion of the two magnetic fields of the systems passing each other in some parts there are no or there will be very little magnetic field force available, even thou there are charged particles in the area, but there are no magnetic fields strong enough for them to interact with to create the visible light. Consequently, in these areas of galaxy there are No glows and No shine.

Reason why this is called the Black Hole, is due to the fact that in these parts of the galaxy, which two or matters and their magnetic fields meet and interact, and can balance each other out, there will be very little visible light emission due to the lack of interaction between charge particles and non-existence of magnetic fields or existence of weak magnetic fields, that cannot produce bright light enough to be visible from the Earth.

This is the reason, why if one look's at these areas of the galaxy these areas seem to be darker compared to their surrounding areas, as the light not only is emitted from the mass of the black hole, but at the same time these matters due to their physical dimensions and their magnetic fields forces strength will not allow and block lights coming from other sources behind them in the galaxy too.

This is why, these regions of the space seem to be black, but from an outsider point of view looking over the area it looks as everything and all materials are thrown in to this dark spot from one side and nothing is coming out of the same side and consequently the physical term of The BLACK HOLE was invented.

This area appears as to be a Dark part of the universe, but in reality some rays of light should be observed being produced in this zone, as some charged particles still interact with the some weak PMF of the systems or what is left of them before they enter and interact and in this part of the galaxy there are visible light but matters are disappearing too.

So, in this region of the galaxy astronomers observe a large gravitational force due to the large amount of mass thrown in to this area by the two systems as they pass each other. But, they have no explanation as what is happening to all this mass. Now one understands the reason for this false assumption and appearance. This is the region where the gravitational forces are in reality mixture of gravity and inertia pull of two systems.

One has to remember that as these two or more systems are passing each other, several scenarios can happen again, which every one of these will have and has drastic and profound effect in that region of the galaxy. Where, at the same time these events can generate a vast amount of new recycled material for other entities within the galaxy.

At the same time one has to take into account the time factor and that one is dealing with a three dimensional picture, as in some cases the Black Hole exists for millions of years during the crossing of the two systems crossing each other's path at a slow speeds. The truth is that the Black Holes are a temporally condition, so to speak, in the bigger picture of cycle of the life of a galaxy, but they surly look like being there forever.

Some of the residual of these interactions will be some cosmic rays (10) or the outcome could be the birth of a new system with a new mass from the mixture of the masses of the two old systems after the system parameters balances each other out. In this case the overall masses of elements within the black hole are present as they were in their original systems. Where, now the new Black Hole could possess slightly less than the combined mass of two old systems. With lose of the some of its matter in the process of the interaction of the two systems into the galaxy.

One cannot forget that the centre mass of the Black Hole is still rotating most probably in the same direction as the systems were rotating which created it. At the same time, as the centre mass is void of the Passive Magnetic Field (PMF), the usual interaction between the passive magnetic force and gravitational force do not take place that usually creates the first basses of heating and sustaining that heat, so that the temperature gradient could take hold, that this in turn could create the scenario conditions for the mass to become a large star or a system.
So that the centre mass have enough material but not the right condition to create the necessary heat, for the time being to make the jump to become the source for a new system or a new star in possession of gravitational forces.
This is one of the ways that the Black Holes become the creator of new stars and systems. Therefore the mass relies upon the heat that is produced mainly by the natural friction of the elements pilling on top of each other in the centre of the black hole.

But at the same time these places become the centre of the mafs of principal matters of the entities within the collision zone. Where at this point the principal mafs and matters of the constituent matters within the collision can take hole and star t the process of creation of the principal star with its strong magnetic and gravitational field forces, which this condition can lead to creation of principal star and hence through realises of mafs of at principal level and through its slowing down and through production of transition mafs, these parts of the universe can again appear to be dark than the rest of their immediate surroundings, and hence the creation of balance mafs of created and the surrounding environment field, then again due to lack of creation of interaction sufficient to create mafs at the mafs of visible light, then these areas of the universe again look darker than the neighbouring regions, and as the centre of this region is covered with the seed of principal mafs and matters then one observes the strong pull of matters to this region without being able to see its origin which is the principal star in the centre of the dark zone.

These interactions between passive magnetic fields occur on regular basses in our galaxy, as systems are at all times on the move in a confined space of their Galaxies and if one had enough time to look out into the space, one will observe the creation of so called black hole the way one observes passing of dark clouds in the sky of the earth during stormy times.

The motions in this specific part of galaxy are in progress over thousands of years, the Black Hole's due to other systems magnetic fields surrounding it and their influence and interface with passive magnetic fields of these systems making and allowing the black hole to leak cosmic rays into the galaxy like a fountain in its surrounding area.

Then it should be understandable now that why the scientific world finds materials from the Universe that is not nor have not the same age as the original component of the galaxy. As like supernovas, the Black Holes are the source of the new material for the galaxy primarily because of the way they are created and the way their life cycle finishes.

The other scenario of the interaction of the two systems could be totally different with an interesting outcome. This is the most horrendous way that the two systems could come into contact with each other and in this case as the two magnetic fields come too close to each other, at the same time other systems around them push these Passive Magnetic Fields themselves together.

The two fields interact and the first stage of creation of the Black Hole is reached very rapidly, but as the two systems surrender more and more of their debris and planets to the Black Hole, the central mass of the Black Hole increases and as the Black Hole is receiving its contents from two different systems simultaneously, then a point reached that the weight and the gravitational force of the new centre mass of the Black Hole super-passes the mass weight of one of the systems, which created it in the first place.

Systems which have already lost some of their mass to the Black Hole at the outset of the creation of it, now losses control of its magnetic fields too. To an outsider looking at this event, it seems as the Black Hole is consuming everything around it. Where, this process can take over thousands of years, until a temperature gradient or a dynamic principal mafs and matters of in the centre of the mass is established.

As the centre mass of the Black Hole has consumed the entirety of two systems, and it is operating in a no magnetic field zone for a period of time, where its internal dynamic plasmatic magnetic field cannot pass its physical boundary (see the paper "The creation of dark matter" M T Keshe for clarity), the system will still can absorb or attract other masses from its surrounding into itself.

In this scenario, if mafs and materials are held for long enough time within the centre mass! That new magnetic field force could be produced and maintained, where at this moment of its life cycle this giant of a mass of the Black Hole will be nothing but a huge Gravity and Magnetic field producing machine. Where, by the end of this process a new system or several systems will come to be born and new materials will be scattered in the galaxy, and this Black Hole will be a Dark place but not necessarily cold.

If the centre mass during its creation does not manage to heat up, then the third phase of the Black Hole syndrome takes hold. This being that, the centre mass at this stage becomes so big and as it never had a chance to heat up due to total absence or lack of strong magnetic force at the same time where the new large centre of the mass is rotating, this new mass will be inherits a huge natural inertia force from its collective mass with no or little magnetic field beyond its physical boundary to keep it at bay, so, this massive large body will start pulling smaller objects near to it into its trap.

The same phenomenon applies when two galaxies come to pass each other's path too, with the most horrendous outcome for the galaxies in that part of the cosmos, where the outcome of such encounter is beyond imagination. This has happened before and astronomers forecast it for The Milky Way galaxy in the future as it passes the nearest galaxy to it in the next few billion years. Where, the surrounding areas near the newly created Black Hole will not be safe from the onslaught as this new mass at all time of its existence has the same fundamental problem as of all other systems, and that it tries to hold onto all its elements. At the same time like other systems the lighter elements and weaker mafs in the mass move to the outer layers and to the edges of the mass.

Where, weaker mafs as lighter gasses and particles move to the outer region of the mass, they become better candidates for them to be pulled away from the centre mass by any opening in the adjacent Magravs systems near to the Black hole. When the centre mass reaches a point that Magravs or its inertia force become weak on its edges, that lighter elements can manage to loosen themselves and escape from its grip into the space of its neighbouring environment.

These freed elements will become part of matter that can be attracted by adjacent systems to the black hole or due to the rapid motion of the Black hole released as cosmic dust in the galaxy. Therefore objects can and do leak out of the Black holes into their respective galaxy even during the time of their creation.

If these elements are pulled out of the power of the centre mass, through the strength or gap in the adjacent systems or weakness in the power of the black hole, then they become the cosmic rays this being due to the fact that during extraction most atoms get striped of their electrons, and some heavier atoms in this scenario lose their electrons in the same way and so they appear as plasma and as part of the cosmic dust in the galaxy too. Black Holes are the source of most of cosmic rays and dust in the galaxies due to this effect (10).

In the galaxies Black Hole's while are in process of being created and during their existence at all times are still surrounded by other systems and are under the influence of their Magravs.

The powers of these Black Holes are usually kept at bay by other systems around them with their larger magnetic field forces and gravitational forces, which the black hole is still encapsulate amongst.

Black Holes are normally observed close to the centre of the galaxies and the universe, as there are more systems clustered together in these regions as this increases the chances of different systems to come in the path of each other and in turn increasing the chance of the collision of solar systems in the and galaxies in the universe and hence increase in the chance of the creation of numbers of the Black Holes.

The stronger mafs and the heavier elements of the universe and galaxies are largely concentrated in and near the centre of these entities, thus the passive magnetic fields and gravitational forces of the systems in these central regions are many times more powerful near the centre of galaxies then what is experience in the outer edges of them.

Consequently, the gravitational forces and passive magnetic forces in the centre regions of the galaxy will soon bring and keep the size and the power of these giants Black Holes under control.

What these collisions of systems and creation of Black Holes achieve is the fundamental principle of regeneration of new systems out of the old matters, and at the same time unclogging of the galaxy of all the loose debris.

If one looks at a three dimensional overview of these areas in space, surely due to the movement of stars and planets and the interaction of their forces, one should be able to predict, where the next Black Hole in that part of galaxy could be produced.

In time, a point will be reached, where the passive magnetic fields systems in one of the immediate systems' neighbours to the Black hole will move on, and at this point the feeding frenzy of the Black Hole will come to an end.

Like volcanic eruptions on Earth, where the mantle is thin enough that pressure of the lava can takeover, and that it can to pure out, pushing its way out into the atmosphere as a volcanic eruption.

This is exactly what happens when one of the passive magnetic fields in this area moves or weakens.

This creates a gap in the fields containing and controlling the black hole parameters, where the materials in the centre of the Black hole, which has been pushed previously into the Black Hole from one direction, for these materials to be released or scattered in to the galaxy.

In this process, part of these materials will manage to cluster together and generate the conditions essential for creation of one or more systems or stars.

Where the rest of the materials of the centre mass of the black hole becomes cosmic dusts, some become comets and some become cosmic rays and so on.

Thus Black holes are not gigantic eating machines compared to the size of their galaxy. In reality, they are nothing but a temporary condition created in the galaxies due to movement of plasmatic magnetic field systems as they are crossing each other.

Very much like two masses of clouds crossing each other path colliding with each other. Where, this interaction of clouds upon each other creates abnormal conditions in that area of the planet, where, their thickness can block the rays of the Sun behind them and the sky looks darker, but in reality these are temporary conditions for the viewer from Earth.

Contrary to common belief, Black holes always release matters and release new recycled material into their surroundings; they are not bottomless pits, where everything disappears into them. They are like a funnel, that if one looks at one end it seems that they are taking everything in, but exactly like a funnel, they have a hole on the other end, that what goes in has to come out, with a difference that the end of this funnel is connected to a kitchen-waste unit of the universe, where that on one end it crunches everything up and mixes whatever goes in before spewing the waste out through the waste pipe.

So the behaviour of the Black Hole is very much dependent on the position of the observer. If one looks from one end they are massive eating machines and from the other end they are the supplier and feeder of new life in the Galaxy.

In real term this dark part of the galaxy is at all times emitting electromagnetic waves of varying levels and strength from all the activities which is happing within, around and because of it. Black holes can easily be identified in their environment and their position can clearly be pinpointed by using much simpler method of detection than by just guessing and look for their effects in a given region of the universe.

In principle the Black hole has all the attribute of any system, but in fact as the internal plasmatic magnetic energy field of the systems creating it, in the given region of the galaxy, cannot create outward plasmatic magnetic energy field, that can interact with other plasmatic magnetic energy fields in its immediate given environment, for it to create outward visible magnetosphere at boundary of the black hole to manifest itself as a visible space in its region, this leads to creation of lack visible light in the man's Magravs amino-acids matter mafs strength.

The phenomenon of Black Hole occurs on the continuous bases on the surface of stars of any Solar system. The principle of the creation of Black hole happens on a smaller scale on the surface of the Sun in the solar system, where these are called the Dark Spots of the Sun.

These are where dynamic plasmatic magnetic fields systems on the surface of the Sun go through the same cycle of interaction of fields with each other as it happens for the creation of the Black Hole in the galaxies, where these plasmatic magnetic fields come across each other in the atmosphere and on the surface of the Sun, consequently they have to interact or pass each other, with the difference that these plasmatic magnetic fields systems on the surface of the Sun, even thou they are so powerful, they in one way or other meet-up with each other, and as their energies have to interact, then sometimes these energies cancel each other out and reach a balance magnetic energy condition.

At these points then, the same cycle as the creation of a Black Hole, on the surface of the Sun in much smaller scale repeats itself as there are no magnetic fields interaction because of similarities and identical mafs strength of the interacting fields, then there very little friction of fields that residual fields in the strength of the man's mafs of amino-acid strength are released, and hence or a very little light in the region and hence the appearance of the Dark Spots on the surface on the Sun is observed. In some instances the same is created as there are no interactions or balanced pmfs strength of the matters and mafs between the charged particles and mafs from the surface temporally are produced and consequently there is no or very little emission of light from the surface of the Sun in these regions.

On the surface of the Sun the same scenario of retraction and absorption of matter as the of the Black Hole starts occurs with a slight variation, which in this case the gravitational fields force of the Sun is so powerful and very close by that as the gravitational forces of the Sun are (in proportion) many times stronger than the gravitational force of the particles captured within the Dark spot region on its surface, that once the magnetic fields have interacted their energies and the content within the catchments of their forces have reached the equilibrium level, then the mass of particles in these regions are sometimes pulled back inward onto the surface of the Sun instantaneously after the collapse magnetic fields which created the dark spot regions, then at this point the physical matter of plasma in this region of the dark spot are pulled back into and is literally missile back into the liquid surface of the Sun, this creating a secondary dispersion of matter in the atmosphere of the Sun and into the solar system and hence the increase in the Sun flairs during the creation and as sequence of the collapse of dark spots on the surface of the Sun.

In reality the persona of dark spot is created mainly through interaction of mafs fields of the Sun and these are created very high above the surface of the Sun, that the dark image observed is not in reality mainly on the surface but much further up in the atmosphere above the real solid surface of the Sun.

In other instances as the plasmatic magnetic energy fields interact to create the Dark spot, the surface tension of the Dark spot becomes weaker on its face away from the Sun and all sorts of mafs and matters in the full spectrum mafs of the universe directly gets ejected into the solar system with much more power and matter then normal dispersion.

Once the temporally condition of creation of dark spot of the Sun is created, By this time, most of the other plasmatic magnetic fields of the Sun physically push and cajole the plasmatic passive magnetic fields (PMF) that created the Dark spot, so that, these external forces in the surrounding environment of the dark spot move and shift them from their position, where by this movement there are no more active plasmatic magnetic fields in the region to interact with each other in these regions and then the normal process of the charged particles interacting with the plasmatic magnetic fields around them start again as before, then everything goes back to normal and the light or the shine on the surface of the Sun returns to normality.

When too many dynamic plasmatic magnetic energy fields start cancelling each other out, the surface of the Sun become congested with many different plasmatic magnetic energy fields and they clog up the surface of the Sun, like congestion in traffic jams, hence the number of the Dark Spots starts increasing.

It is important to remember that Objects (these being solid, gas, plasmas or etc.) possesses their own plasmatic magnetic energy field and all these plasmatic magnetic energy fields on the surface are dynamic and respond to other dynamic plasmatic magnetic energy fields, which are created internally in the inner cores of the star. Where, these internal magnetic fields are the creators' of the gravitational field and the overall magnetic field of the star itself. Where, The stronger plasmatic magnetic energy field (the gravitational field) usually dictates and plays apart in the position and level of freedom of the weaker plasmatic magnetic energy field (the object) in respective to each other.

Thus, when the inner cores of the stars due to their near mono atomic structure go through their internal current realignment (12, 18) this in turn leads to change in the direction of the gravitational field and thus the polarity of the whole star.

Consequently the rotation of the gravitational field which itself is originated and created magnetically interacts with the dynamic plasmatic magnetic energy field areas on its surface and in its atmosphere causing them to move in respect to the internal dynamic plasmatic magnetic energy field movements of the star.

Therefore main reason for sudden periodic increase in the movement of the plasmatic magnetic energy field on the surface of the star and sudden increase in the creation of black spot in their environment can be clearly be identified, to be due to the interaction of the dynamic plasmatic magnetic field created by the matters and mafs on the surface and the inner Magravs fields force of the stars.

Where, the inner dynamic plasmatic magnetic field of the star is by order of magnitude several millions of times much stronger than the isolated single or packages of dynamic plasmatic magnetic fields on the surface. Thus the stronger inner fields force of the star, dictating the direction of the motion of the weaker fields of the Dark spot. This being the very reason why the eleven-year cycle of change of polarity of the Sun brings about an increase in the number of the Dark spots of the surface on the Sun on regular bases. This being due to Sun's internal plasmatic magnetic energy fields' changes which brings about the polarity changes, every eleven years. Where, at these cycles there are more of these surface and atmospheric dynamic plasmatic magnetic energy fields in motion and available to interact with each other and the internal mafs of the Sun.

In fact, as the polarities of the Sun changes, so more dynamic plasmatic magnetic fields start moving from one pole to the other to accommodate the physical practicality of the plasmatic magnetic energy field polar change. Thus, a larger number of the dynamic plasmatic magnetic fields become to cross each other's path and consequently there are more chances of two or more dynamic plasmatic magnetic fields to meet and to balance out each other's field. Inevitably, as at these areas, there are no effective plasmatic magnetic energy fields for the charged matter to interact with, thus there is less light emission from these areas of the Sun. Henceforth the appearance of darker areas or dark spots in these areas on the same principle as the creation of the Black Hole phenomenon re-occurrence.

The increase in the number of dark spots of the surface of the Sun is always in sequence with the period when the Sun internal dynamic plasmatic magnetic fields activities go through their repositioning.

The number of these Dark spots increases and as they move, suddenly a vast number of dynamic plasmatic magnetic fields cancel each other out and as the change of polarity of the Sun completes itself and settles down.

The whole cycle of collapse of the Black hole re-occurs', that is, where a Dark Spot collapses, and their new plasmatic magnetic energy fields fines its total balance, and the matters which where gravitationally trapped under the influence of the dark spots, which are mainly plasmas, are retracted by the gravitational pull of the star and are ejected back onto the surface of the Sun is large quantities and numbers. This is the time, where there are massive activities and strong Sun flares scattering outwards from the surface of the Sun, where, large quantities of the Sun matters are plashed outward in the solar system.

Sometimes, as the mass of particles collected in a Dark Spot is pulled back into the surface of the Sun, due to their mass and speed, these become like meteorites hitting the surface of the Sun. This act causes more than usual charged particles to be leaving the Sun surface into surrounding space.

The darkness on the surface of the Sun is not due to what is commonly thought to be due to cooler temperature in these regions, as the Dark Spots are purely due to the fact that there are no or very little dynamic plasmatic magnetic fields available in these regions to interact with the charged particles to create light.

So underneath the Dark Spots on the surface of the Sun the temperature of the surface is still the same as the rest of its surface.

Conclusions

Black holes are natural physical by-products of the motion of forces and matters like every other object in the universe. The only difference for this body to appear is the fundamental shortage or total lack of one of the three main ingredients of the creation, namely the lack of the Passive Magnetic Force Fields in the region, where these black holes manifest themselves.

Black holes are another creation of universe like the uniqueness of the Blue planet and once one has understood the methods of universe's powers of creation within all its physical parameters, then every aspect of the appearance of objects and fields can only be explained with understanding of the interactions of these with each other.

The most beautiful part of this phenomenon will be! How Man will use this knowledge now for the advancement of its race?

Understanding of the method of creation, existence and the behaviour of the Black Hole in space opens new doors for the Man to plan his future excursions in space and the appropriate name for this object should not be a Black Hole.

These regions of Black hole in the galaxies cover areas as large as several solar systems, with them possessing all the physical attribute of solar system. With the difference that they do not shine due to the lack of appropriate dynamic plasmatic magnetic fields strength within their construction or during the time of their existence in the environment of their operation.

The appropriate name for this object as has one body and two distinctive characterizes behaviours and appearance, by first characteristic being a gigantic eating machine, and the second characteristic by allowing new systems to start from what it has digested. There is an old Aryan name for an arrow with two heads as the Black Hole with two characters and this object is called ZHUBEEN.

I call this unpredictable object in the galaxy with this name, as now it is a known object and not a mysterious scientific phenomenon.

Elaborations

The helio-seismology is the study of the internal structure and dynamic of the Sun using Sun-quakes, and the Sun oscillates continuously by varying degrees and directions. These oscillations are caused mainly by turbulent convection just below the surface. The data from spherical-symmetrical solar model of the same condition shows close correlation to physical collected data through inversion. This showing that the data simulation is close to the physical reality of what is happening in the inner sanctum of the Sun.

For our testes in dynamic cores and data's attained through them, it can be confirmed that the rotation and the turbulence have direct effect on the heating of the matter within the confined of the reactor build on the same principles, where in the reactor and consequently in the creation of the current through convection and magnetic fields production within it, where the turbulence of the matter in the core of the reactor creates the dynamic and the dynamo effects which this has a direct effect in and on the production of the dynamic plasmatic magnetic field in the reactor.

Turbulent convection and differential rotation in the Sun are thought to be the cause of the creation of the solar magnetic fields and the directions in which these magnetic fields appear to flow out of the surface of the Sun according to their direction of rotation their tightness are known as helicity of the magnetic field in the region. Where, Helicity simply describes the handedness of this field and as the field literally opens up like a cone as it moves away from the centre towards outer region. As Magnetic and Magravs helicity is in fact a measure of the knottiness and the twistedness of these fields within a region, where this closely related to the turbulent dynamo effect. Where, from all observation in the past, large-scale poloidal magnetic fields the helicity have a direct linkage with strong toroidal field leading to finite helicity.

As, in the reactor tested on the same bases as turbulence of the mafs of matters where the reactor cores having a finite parameters and the rotation and turbulence are predetermined, therefore the helicity of the magnetic fields has to be the same at different points but equal for all distance from the centre. Where, due to the small size of the core, a mean or average strength or tightness of the magnetic field could be assumed.

Thus a point of interaction of two magnetic forces in the cores will create condition similar to the dark spot on the surface of the plasma. Which if the collapse can go through plasma towards the boundary of the core, this can create a sudden surge of dynamic plasmatic magnetic field or which this can alter the system operation mode. The reason for writing this paper has been due to the fact that in the reactor designed, if the condition of Black Hole is created, this can lead to the shutdown of the system or unstable condition of the operation of reactor. As this has happened in two occasions during testing of the spherical core, where the second magnetic field was deliberately created and positioned to interact with the internal magnetic field created by the core.

In both cases the whole system did go through magnetic poles trying to bypass each other.

In the second test in trying to prove the point of the interaction of two magnetic fields and creation of black hole condition, the whole system was twisted and the axial pines of the core were off centred.

The core axels had to be realigned, but the damage to the core weld and axel proven to be too much to repair, and more or less this test of creation of black holes caused the loss of the first steel prototype core (Fig. 28).

At the same time in the same reactor the creation of the dark spots due to interaction of the matters were deliberately created for specific use of the system, where investigation in the creation of dark spots, and where the use of transition matter condition for the creation of dark matter, for the fuel needed for deep space travel was tested and confirmed.

This is how these above phenomenons clearly can be explained, and the principle for the creation of the Black Hole can easily be demonstrated.

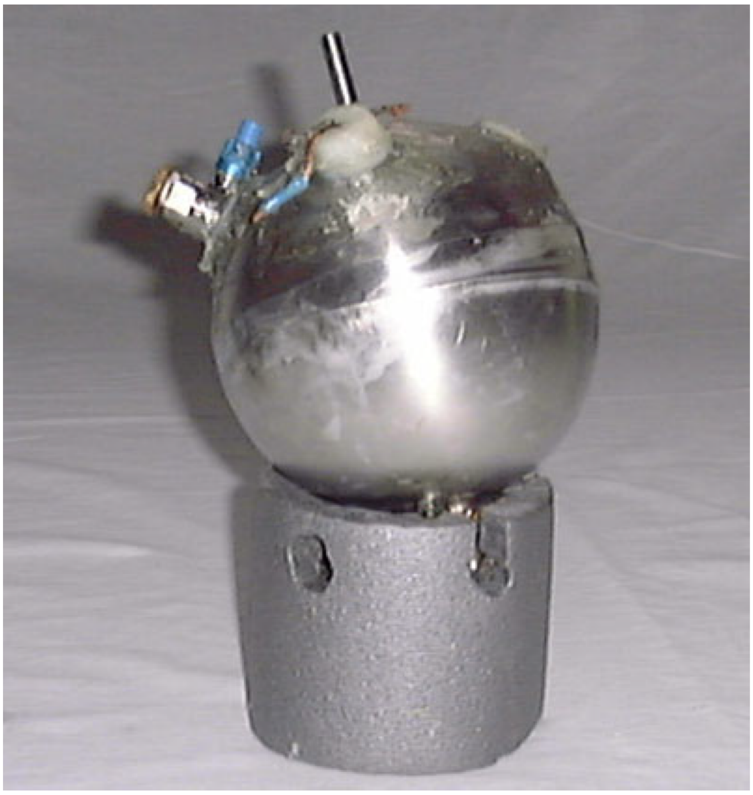

Fig. 28 Test reactor

In the future papers the writer will explain how the Black Hole's effects can be useful in helping to travel through galaxies and in the Universe. These are at this moment of time beyond understanding at the present level of scientific achievements.

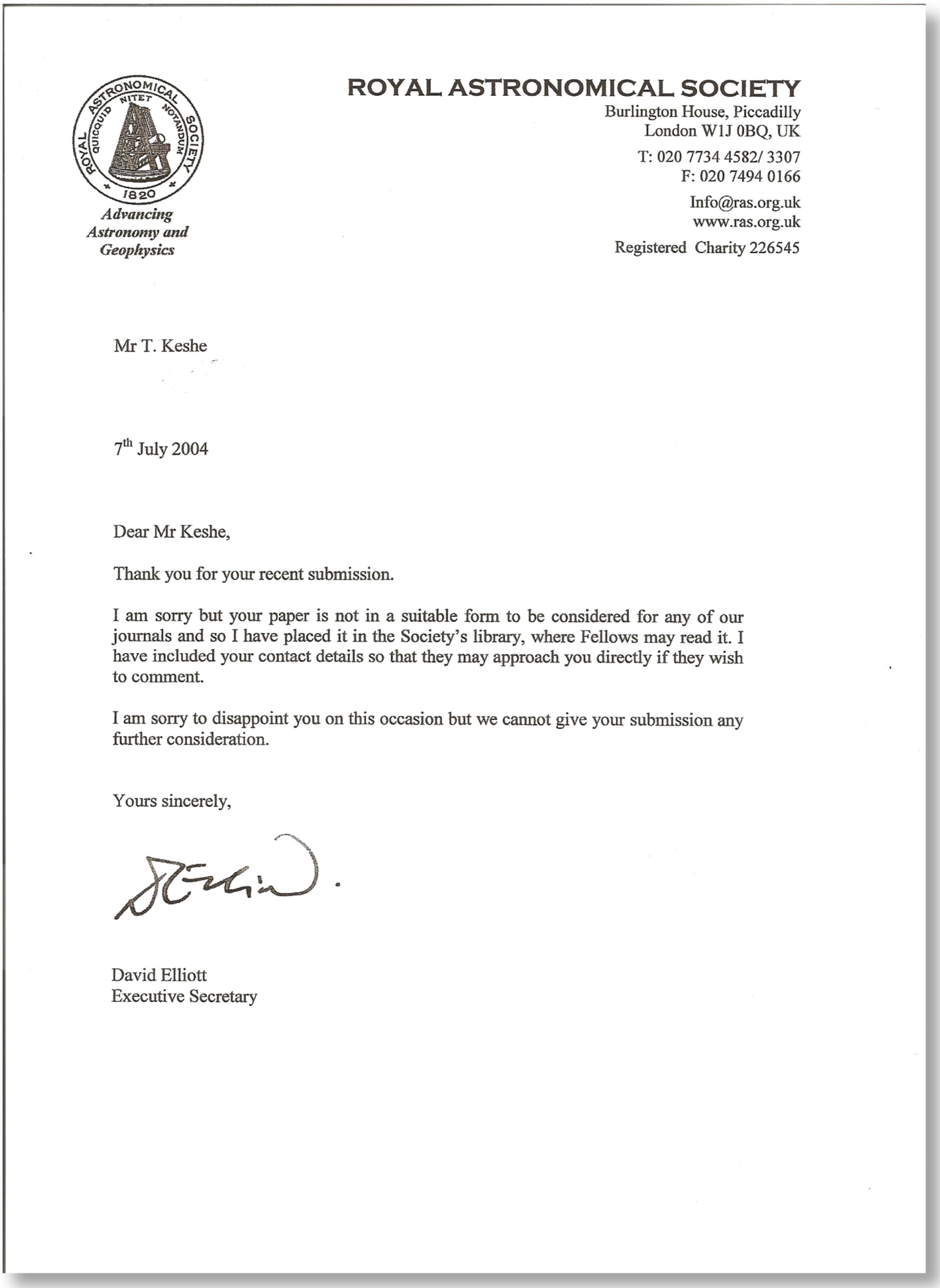

Advancing
Astronomy and
Geophysics

ROYAL ASTRONOMICAL SOCIETY
Burlington House, Piccadilly
London W1J 0BQ, UK
T: 020 7734 4582/ 3307
F: 020 7494 0166
Info@ras.org.uk
www.ras.org.uk
Registered Charity 226545

Mr T. Keshe

7th July 2004

Dear Mr Keshe,

Thank you for your recent submission.

I am sorry but your paper is not in a suitable form to be considered for any of our journals and so I have placed it in the Society's library, where Fellows may read it. I have included your contact details so that they may approach you directly if they wish to comment.

I am sorry to disappoint you on this occasion but we cannot give your submission any further consideration.

Yours sincerely,

David Elliott
Executive Secretary

Fig. 29 . Letter of Royal Astronomical Society

CHAPTER 11

The relationship between gravity, mass and weight

© M.T. Keshe/The Keshe Foundation2004/2006. All rights reserved.
Author: Ing. Mehran Tavakoli Keshe
Date of release 21.12.2005
Date of Internet release: October 15, 2006

This paper is written for clarification by the encouragement from Mr Dirk Laureyssens.

Abstract

With the new understanding of the creation of the double dynamic plasmatic magnetic field of planets and understanding the new principle and method of the creation of gravitational forces, further on the balanced fields outcome of the collective and internal interaction of Magnetic and gravitational fields up on each other of any entity within the structure of mafs, plasmas, atoms or molecules this leads to creation of so called the mass effect of an object, it is now simple and apparent that what the mass is, how it is created and why the same mass possess different weights in different gravitational field forces environment.

Hypothesis

The mass of an entity is created from the outcome of the deduction of the interaction of the gravitational fields and Magnetic fields forces of an entity and has nothing to do with external field's interaction and strength of the environment of that entity. The variation in the numerical difference in the weight of the same mass of an object in different dynamic plasmatic Magravs field forces of different environments or different planets, is totally due to the interaction of the total collective dynamic plasmatic Magravs fields forces, including the inertia effect of the environment which the entity is weight measured and of the collective internal dynamic plasmatic Magravs fields strength of the mafs of the entity which create the mass of the objects upon each other.

Where, the magnetosphere of the same entity is the total sum of the magnetic field and gravitational field forces catchments zone of the same two internal fields which make the Mangroves fields of the entity that they can attain in their given Magravs fields strength environment.

Discussion

Planet like the Earth possesses a combination of two magnetic field forces. First the gravity, created due to the interaction of the two independent dynamic plasmatic magnetic fields in its centre cores upon each other. Second the inertia, which is the collective dynamic plasmatic magnetic field of its atoms and molecules of its constituent matters, which has the physical body of the planet.

Any object, this being plasma, an electron, an atom, a molecule or even a human body is created from a collection of the plasmatic magnetic Magravs energy fields of different strength and their continues interaction between and up on each other, which in total decides the total or collective magnetic field possessed by that object or the mass of an object and the total interaction within molecular or atomic dynamic plasmatic magnetic field of a given object is always constant irrespective of the other dynamic plasmatic magnetic field fields around it. The mass of an object has a collective package of plasmatic magnetic field tightness of all its constituent elements, and this will not change as long as the object is intact as one entity in its overall atomic or molecular or collective molecular appearance.

Which the interaction of the one object dynamic plasmatic magnetic fields with molecular or atomic magnetic field of another object Magravs in its vicinity or it environment will determine the weight of the two objects in respect to each other or in a given environment.

The mass of a body in relation to another molecular magnetic field, like gravity is dictated by the strength of the magnetic fields which creates the gravity, which in the planetary system like earth, this is maintained by the planet according to its inner molecular construction and conditions in and by which the Magravs strength of the planet manifest itself to the body.

The interaction between two or more dynamic plasmatic magnetic fields created by a planet to create and determine the gravitational strength of the planet are due to conditions and in conjunction with its elements.

Usually these types magnetic fields created by planets are very strong, so an object like a human body which possess a weaker collective dynamic plasmatic Magravs and in reality possesses its own gravitational field forces due to its combined electromagnetic fields of its elements which it is composed of, which for a given body this is constant, these will be attracted and be operating in a much more stronger dynamic plasmatic magnetic field environment of the planet. Which, this gives the body due to its interaction of the Magravs and the planet dynamic plasmatic Magravs interaction, where the strength of the Magravs force of the planet and the body in respect to each other, this determines the attraction force between the two fields of two entities of the body and the planet and consequently the mass of the weaker force or the object in respected to and within the stronger Magravs forces of planet or Magravs forces of the body dictated the weight of the body in respect to the planet.

In reality the reverse of the same criterion is true too, for the mass of the planet, and its weight in respect to the dynamic plasmatic magnetic field of the body.

Thus, it is easy to explain, when a body interacts with two planets of the same size and when the two planets are made of the same materials, where one for example has lower temperatures, or internal speed of rotation or lower rate of ionization of the prime materials within its Caroline core, which all these can affect the strength of the Magravs of the planet for it to possess a weaker dynamic plasmatic Magravs strength, then planet which has a more inertia power, which is of the weakest order than gravity, but always constant, for a planets like moons, then the magnetic attraction, or pulling, between the same object and the second

Magravs force will be weaker, this by the same principle as above giving the body in the second environment a lesser weight value.

This is the very essence, why man's body appears to have less mass on the moon than on the Earth. Where, this is simply due to the fact that the dynamic plasmatic Magravs of the moon, which creates its gravitational force, and even this could be the inertia effect of the moon, have weaker Magravs strength power to interact with same given human body's dynamic plasmatic Magravs than the Earth dynamic plasmatic Magravs strength on the same body.

Therefore, the weight of a body is position dependent in its relation to its counter interacting planet or object inner Magravs forces, which create the Magravs or/and the inertia, or static Magravs of atones of that planet.

That is to say, if one could replicate the same Magravs field as on the moon, within a given position in the Earth atmosphere, one will have the same weight as the body will have on the moon within the Magravs force of the Earth.

This is why in propulsion technology there is an absents of the Magravs within the system objects always float and have less mass, as first of all the propulsion systems do not create Magravs that the object can position itself in respective to and lock to the position that the object become motionally stable and this and the secondly for example in respect to the body of space users of the propulsion system, the body's collective dynamic plasmatic Magravs strength, has been positioned far enough away from and within the dynamic plasmatic - Magravs strength of the Earth, that the mass dynamic plasmatic magnetic field strength of body in that position within the Earth gravitational field, due to the speed of the rotation or motion and conditions created within of the craft, will have a weaker interaction upon the body, allow the body a more freedom of movement.

Using the principle of the Margrave positioning for vertical take-off or choosing stationery or fixed position orbit for an object, like for satellites, within any atmospheric condition, where the magnetic fields created within the reactor of a craft can lead to creation and control of the Magravs within the craft, allowing the mass of the body of the object in a given position within the atmosphere of the planet to reach the equilibrium in relation to the Magravs of the planet, then in that particular point within the fields that are created by the planet the body will have the same mass as the earth and the time delay lapse of free float satellites to tune with earth gravity become irrelevant and there will not be a need time delay or adjustment.

This is the fundamental principle and criterion for any system that is used in any spacecraft's technology to replace the present jet propulsion or rocket technology.

Therefore the mass of an object is the representation of the effect of the Magravs of a body in respect to the Magravs of the second body.

The Magravs positioning is the reason why atoms of the same element fuse to create molecules, as their weight in respect to each other is zero, or equilibrium, even thou they both possess the same mass.

At the same time, if this is understood in full, then most of the chemical and biological interaction of atoms to create molecules can be comprehended and replicated in an easier manner.

Where, the collective mass of atoms in respect to other atoms within a molecule Magravs will dictate the number of, position of and the composition of atoms of different elements within that molecule.

Then this explains why it is easier to create certain material in absence of gravity in space labs, that in the presence of lesser dynamic plasmatic Magravs strength of the planet in high altitude then only the weaker Magravs forces of the elements come in to play as they are in a much weaker dynamic plasmatic Magravs strength levels from the centre of the planet and they can combine to produce materials, which under Magravs forces of the Earth are impossible.

Conclusion

Therefore the mass of an object possess different numerical weight values in two different Magravs force environments purely due to the interaction of the double dynamic plasmatic magnetic field of the planet and the constant collective dynamic plasmatic Magravs of the object.

The gravity itself is created by the interaction of magnetic fields of the matter and as the mass is the product of interaction of gravitational force and Magnetic fields forces of the same matter magnetic fields based generated phenomenon and as any environment in the universe possess magnetic fields irrespective of their strength and interaction of these fields in that given environment creates a new Magnetic and gravitational fields or Magravs in that environment that the outcome the interaction the two Magravs of the matter and it environment become the weight of the matter in that environment.

That is to say the variation in weight of an object in different environments is totally due to the interaction between the dynamic plasmatic Magravs strength of the two of entities upon each other, in respect to each other's position, in their respective environment.

Where Magnetospheres of the entities are the outcome of the sum of the interaction of Magnetic fields and gravitational fields forces that the entity can master for it to be able to protect and hold on to all its mafs and matters which the difference field forces of the same Magnetic fields and gravitational or Magravs of the same entity has managed to attract and accumulate within its given environment.

In short the interaction of the two or more dynamic plasmatic magnetic fields lead to creation of Magnetic and gravitational field (Magravs) of the entity, and then the balance of interaction of the Magnetic and gravitational fields forces leads to creation of the mass of the same entity which the mass is entirely dependent on the strength of the two original magnetic fields of the entity, and then when the Magravs of entity interact with the Magravs of another entity which itself magnetic fields origin based, then difference or the influence of one Magravs of one entity in respect to the second entity Magravs make the weight of the two entities in respect to each other, where as much as the first entity has weight in respect to the Magravs of the second entity at the same time second entity has a weight in respect to the first entity too.

CHAPTER 12

Parallel worlds

Mysteries of existence about parallel worlds, the questions to ask are or can be as follows:

1. How do they become to exist?
2. How can we interact or be aware of their existence?
3. What can we do or have done to get there?

Let us consider the structure of simplest entity of plasma of an electron in overall structure of let say an atom of hydrogen (or atom of sodium which is part of piece of stone or metal, gas and etc.).

The plasma once created and maintained from interaction and down levelling in strength of different mafs of which this plasma in its division leads to creation of an atom which at this point the plasma of the electron of the atom has an independent existence from plasma of its proton and electron decides about its own motion and its own position in its environment in respect to its proton. Even though, this is conditioned on the position and strength of the proton of its atom. Where, electron itself is made up of collection of different magnetic fields strength and their interactions and as the electron of an atom is not aware of the its larger counterparts of proton(s) and other electron(s) in a larger elemental or molecular structures and it is not aware of any state of matter and at the same time that the electron is not aware of the part of heat and cold in its environment are playing in its position and motion in respect to its proton and other entities of the same which it is connect to.

Where, this electron physical appearance as any physical matter is dependent on its environmental condition. Where, let say this electron can be an electron of atom which is part of atoms of a piece of stone and at this point the electron is not aware of the existence of a stone in larger scale levels of its existence. But the electron exists within the existence of an atom, which the atom exist in existence of the stone.

But now let’s say the stone is parts of a salt stone grain which is consumed by an animal.

Now the sodium atom by making its Magravs available to protein of a cell of the animal it becomes and takes part in the bigger scale and larger Magravs of a protein chain connection, where sodium on its own could not achieve.

Let’s say that now the protein and in conjunction with sodium this creates the condition of neuro-connection and the cell becomes part of living being, this being part of neuro system a shrimp, dog or a man.

Where, for example the combination of different protein and different atoms and the interaction of their Magravs and balancing act within their mangroves lead to need for interaction and new balancing Magravs, where this new interaction of proteins and their different Magravs, leads to a need for cohesive and balancing line of magnetic fields line of transfer of fields. Where, an accumulation of different fields and Magravs and their interaction brings about feelings and intelligence and so forth in the animal.

Now as the different Magravs and their interaction takes place, this being in the level of an atom, a solid, protein, combination of proteins and their attachment to other atoms, these new entities create two distinctive levels of individual compacted fields one for their tangible state of existence and one they facilitates their fields and positioning for the use by and to other parts which are related to them but not connected to their existence in the line of communication of other atoms.

What this means, is that the entity gives life to existence of two levels, one is the physical tangible of itself and one is the fields and energy which it is part of, but once it is created the new entities has enough field that is an accumulation of the interaction of another interaction that this new entity does not need a physical entity to manifest itself in the tangible world, but the fields strength is sufficient enough that the new entity can easily exist on its own right in the universe and make decision about all others whom have become part of and operate through the new entity.

Now as the more and more fields and matters interact and creates a certain fields levels, that let say they lead to tangible matter existence of an entity, the accumulated physical matters have a parallel acuminated non tangible levels too which are its fields which allows it to be part of bigger entity.

Where in complexity of previous interactions to reach a point, these interactions have released or will release new complex fields that these fields perform and exist as the matter part do.

But sometimes the matter and associated fields interlocked, but exists and operate into different levels. Even though the tangible matter is aware of its own existence, but it is not aware of the existence of the associated counter interlocked fields within it, this be it the same as the plasma of electron of a piece of stone, which the electron is part of.

Thus now it becomes clear how a man has physical body and a soul. Where, the physical body is conditioned of the matter world and its environment, where the soul has no such a conditioning.

Thus the soul is aware of the existence of the physical body, but independent of its existence, but it receives the changes in the condition of the physical Magravs of the body.

Thus the soul of a man is a parallel world to the physical world of its associated plasmas, which has created it.

But as soul has no need for the physical body to sustain its existence once it is created, as it operates in a different field zone non tangibility, hence the physical body can die, but the soul will be left in tacked and carries on without its detached physical counterpart in the world of creation.

At the same time, the fields of the soul will carry-on with its interaction with other fields and Magravs of its own level and it will keep on attaining more and keep on and interacting as it physical counterpart did and combining with more and creating more the same and different.

The soul, itself is the balancer of the physical matter of the body, but from point of detachment, does not need the physical body for it to exist. As the physical interaction has led to its existence, but by now it is independent enough to carry on with its own operation.

This is like giving of a birth, even though the baby has been in need of the mother for it to start and become entity, up on the departure from the mothers body, the two entities can live two independent lives, without a need for each other existence.

Thus the soul of an entity, this being stone, animal or a man has the same method of operation and existence as any other matter or fields in the universe, and that as once it is established and sets its frame work, it does not need its physical counterpart which led to its existence.

Now one can understand how different levels of existence do operate and exist in conjunction and in parallel with each other.

Take this one step further and let say fields of the souls have established the same methods of interaction and communication as their physical counterpart plasmas of atoms did with each other and souls through their fields of existence interact with each other and lead to a more dynamic parallel non tangible world of existence in the dimension no need for the actually physical dimension of existence.

In the universe there creatures, which are made from the interaction of principle mafs and matters strength as man is made of matter mafs and matter fields strength and they can operate in a non-tangible level of existence rather than the physical tangibility to the eye of the man

This does not mean that they creatures do not have a soul or they are more intelligent than the man, but the only difference is that they operate in different mafs strength, which these lives can be strange to the mans' travels of the futures as they have the ability to transmute between different levels of mafs strength and be able to visible or detectable in their spectrum of mafs strengths

The soul of the man ids the like of the light from the star in distance universe, where the physical star could have been exploded and destroyed billions of years ago, but still at night the man observes the soul of the stars as the shining light in the dark of the night.

Parallel world exist even within the and during the life of the man himself, but they need the eye of the wisdom for the beings to see them and recognise them in their environment.

References

1 The book by M T Keshe (The universal order of creation of matters by M T Keshe)

2. The light Paper written and not yet published by M Keshe

3. The creation of Black hole By M T Keshe published in 2005

4. The production nano gans M T Keshe published in 2009

CHAPTER 13

The universal principle of division of pervious universe(s) and their decay leading to creation of new universe(s)

Written in 2009
© M T Keshe 2000-2010, All right reserved
ITNOGOG

In the universe there are always parallels due to the magnetic fields laws of interactions and hence there cannot be any deviations from these laws.

At the present the scientific world has been intrigued and encouraged to understand the phenomenon of the creation of this universe and scientists have tried to unravel the mysteries of the creation of this visible universe.

In fact the creation of this universe as is visible to the world of science, it cannot be any different than any other process in the world of creation and this universe has been created through the same principles as the rest of the magnetic plasmatic nuclear and atomic interactions, addition and division.

This meaning that the present universe is created though the same method as has been observes in the division of the neutron into electron and proton or smaller and/or bigger components of the same Matters.

Hence to follow the principle of the universal order of matters, this universe is created from the division of a larger universe that the man's visible universe is the smaller part of the division.

The reason for it being the part of a system is that in the present cosmological term, this universe is following the attraction of the bigger part and hence it is always in the motion to follow the Magravs of the bigger sister.

This being the reason why it looks as the universe is always expanding and as part of the universe has been noted to be collectively following different path of motion than the rest of the universe.

Where, the portion of the independent motion of the universe is the indicator that where the bigger portion of the division of the pervious universe is in respect to the man's universe. (Fig. 30).

It has to be said through the same principle of decay of neutron to plasma of electron and proton (1) and similarly as the decay of heavier atoms to two or more lighter atoms, the original universe through the same principle the mother universe of the man's universe, itself has been created and was created from the division of a larger universes through the same method decay and division during the time of the life of the unicos.

The difference in the universe's creation and decay is in principle merely the size and amount of the plasmatic field content plasma of each universe in comparison to plasma of electron or proton after nuclear decay and division of the neutron.

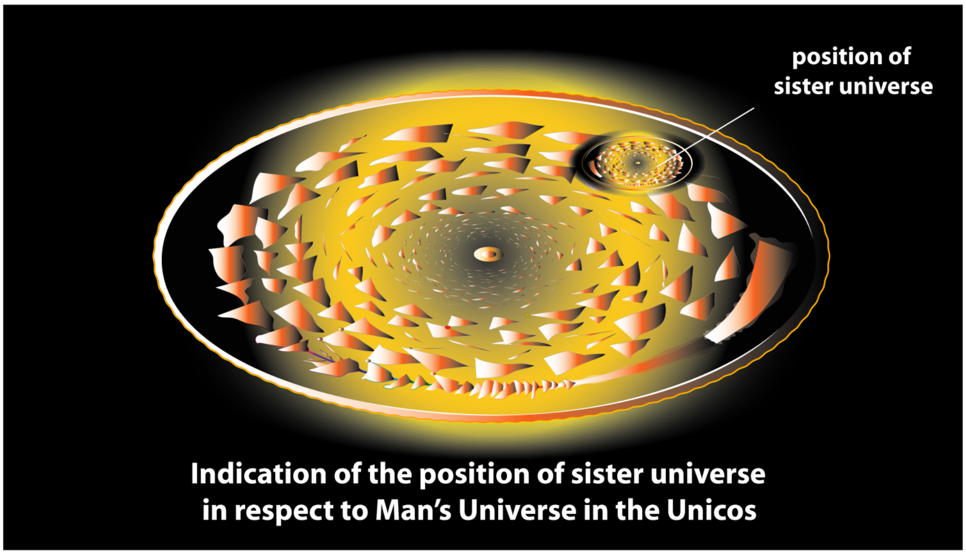
position of
sister universe
Indication of the position of sister universe
in respect to Man's Universe in the Unicos

Fig. 30 The position of the sister universe in respect to Man's Universe

Thus in fact our present universe itself is and was created some thirteen billion years back from the division of a larger universe. Hence every time as it is the same with division of the plasma of neutron to plasma of electron and proton, this division the new universe(s) will carries the characteristics of its mother universe, but due to reorganisation of magnetic field strength the new universe will have a new and fresh time, which this shows as the age of the new universe.

In fact this being the like of given birth to a child, where the age of the mother is irrelevant once the new seed of life is created and once the child is born, even though; the new child carries some of the same DNA fingerprint of the parents. The new life has its own cycle of life and its own timing.

Therefore the theory of Big Bang and M-11 as have been proposed in the recent past has no meaning and has no principle parallels in the physical world of creation of universe, a child or plasma of the proton or an electron. These two proposed principles are nothing, but showing the man's and scientific worlds lack of knowledge and ignorance of the real terms and parallel entities existence of all units in the universe.

In fact once the division is created and the universe is created the rest of the material and matter content and how they organise themselves in that universe is more or less follows the same principle as before and once the field are released within a given Magravs which the fields can created in interaction with each other, then the size of universe, galaxies, stars and planets and even creatures within that given Margraves strength of the newly produced universe are determined by the entity itself and without interference from its previous part.

Where, the new structure is dependent on the fields within the new universe and the internal motion of the new universe is internally determined by the total mass, which the interaction of the Magnetic fields and gravitational fields of the universe can master on its own.

Thus in fact in the bigger or smaller portion of the original mother universe which the man' universe is partly created of, in the bigger or smaller portion due to the larger mass the galaxies can be much larger in portion to the galaxies in this universe and conversely due to larger or smaller strength Magravs of the internal structure of the larger or smaller portion, the being's within these universes can be smaller or bigger or totally different in operation than the beings in this universe.

It has to be said that life in any form is not the exclusivity of the earth, if this is the mind-set of some men, thus this shows their lack of understanding of universe and how it is created, not the real truth and how matter's operated and function in the universe.

Therefore man is part of an inclusive pattern of field's motions in the universe and nothing more or less can be attached to the existence of the man.

The man in physical term is the entity and in the soul the connection of the field between himself and the other fields in the universe.

The physical body of the man is the like of the planet earth, where it has mass and its magnetosphere expands beyond its physical boy, but its total Magnetic fields and gravitational fields determines its position and relation in respect with other matters in the solar system and in a small way it determine the position of its star in its galaxy and souls which are created within it in their course their operation in the universe, as then the position of star determines in a small way the position of the milky-way in the universe and so forth.

Thus the soul of the man is the external connection of the operation of the physical body of the man in the universe, it is part of one and inclusive realty of interaction of the whole.

Thus the soul is part of the operation of the universe and every creation can extent the effects of its magnetic field to others which can affect their position and existence of any in possession of a soul.

Thus the soul, which is magnetic fields created and expands beyond the physical body of the man is not exclusivity of the creatures but it is the common denominator and common link between the man and other creatures of the God and has direct effect on the operation of the rest of the universe.

Therefore the man's body due to the interaction of its atoms protons and electron and neutrons magnetic fields, this object creates its own magnetosphere and its own mass and this body is and has to be in possession of magnetosphere which shows the working of the magnetic fields of its constituent matters and its subcomponents and fields this being including thoughts.

Where, in the physical worlds as the same as the solar system possess magnetosphere and so does the star and planet, then the body of the human through the same physical principle possesses the magnetosphere, which is so called aurora of the human's body. The field of the aurora are part of outer appearance as the soul of the man, but is fact the total effect of the internal interaction of all fields of the body which creates the Magnetic field and gravitational fields of the body and their interaction on each other which leads to the interaction with their fields in the universe, this overall output of the fields of the body of the man is the soul of the man.

Thus, all dynamic plasmatic matters in the universe possess a soul. But the man has the intelligence to allow the interaction of its soul with others; therefore through this principle the man is responsible for all his actions and he stands answerable to other souls in the universe. As he decides how these fields which partial and fully controlled through the Magravs of the matters of his brain, in how and what happens in its interaction with other souls effects their operation in the larger universe.

Hence man has to understand, that universe and others, which has the use of this universe all have been created from and the universe, which has been created from the division other universe and will be created from the division of this universe and in the future are all interconnected and operate through the same fundamental principle.

Therefore there are parallel universes to this universe, with the difference in the measure of their magnetic fields and strength and their position in their given environment.

Thus as we have the same and similar electrons structure's and protons structures in the same structure of an atom, which this structure of positioning of electrons and protons is the same for all known atoms, then the same applies to the universe too. This meaning that there are similar universe as this universe in the greater universe and other universe which this universe has been created from it and those that we will shell create through he operation of our souls and bodies mafs and matters.

It has to said that at the same time all beings in the universe use one and the same universal language of communication, where in the universe the common language is the basic plasmatic magnetic fields of basic field's. This is based on the first plasmatic magnetic field strength of existence of the simplest plasma of atom of the universe, and is similar to what we call in this universe neutron and its division and collections as atoms the like of the Hydrogen atom, which the mystery of this has been kept from the man and not apparent through his own misunderstanding of the operation of the universe.

This language is universal and once the man on earth understands this phenomenon and can fully measure the magnetic fields strength of the electron Magravs of it and he uses this as the base strength, then man will be able to understand and communicate with rest of soul possessing creatures of the universe.

Until then, the search in the sound waves to look and try to communicate with other creatures of the universe is a fantasy and shows the lack of his understanding of the world of the Creation by the man of ignorant intelligence.

Therefore the gravitation, magnetic, radiation and other similar properties and interactions of the field(s) on each other in this and other universe(s) decides on how and the way the creatures in the different universes and in the man's unicos and universe would look and behave.

But in essences they are all made of the same building block of *MAGNETIC FIELDS AND IN ITS AND THEIR INTERACTION OF DIFFERENT STRENGTH AND TIGHTNESS OF IT IN THEIR ENVIRONMENT*; This determining how tall and short and if one has eye or uses the interaction of fields to interact and communicates with other creatures of its environment, universe(s).

CHAPTER 14

How would it end?

The end of this universe will be the same as it started.

What this means is that as the magnetic field balances of the universe reduces to a level that the universe has to re-organize itself that it can keep its constituent field forces together, the whole universe will, like as with neutron plasma will decay and divide into two constituents of smaller universes, which these universes are still connected to each other through their original magnetic field forces (book 1 - chapter 17).

Hence this universe will be divided in to a smaller universe and larger universe parts of a set. This being the same process as the like of the process as neutron decays and the rest of the known nuclear decay of atoms do as there are no other known processes of division process in the universe, and there has to be uniformity in all dimensions of creation.

There is no end in the universes but conversion and amalgamation of fields and matters, which are created out of fields themselves.

Thus, this universe will divide into two active fully operational of lesser universe and bigger universe masses (fig. 31). But in total the overall balance of the matters and fields in both universes will be near equal of what was in this universe, with a little leakage of fields into the Universal Cosmos (unicos).

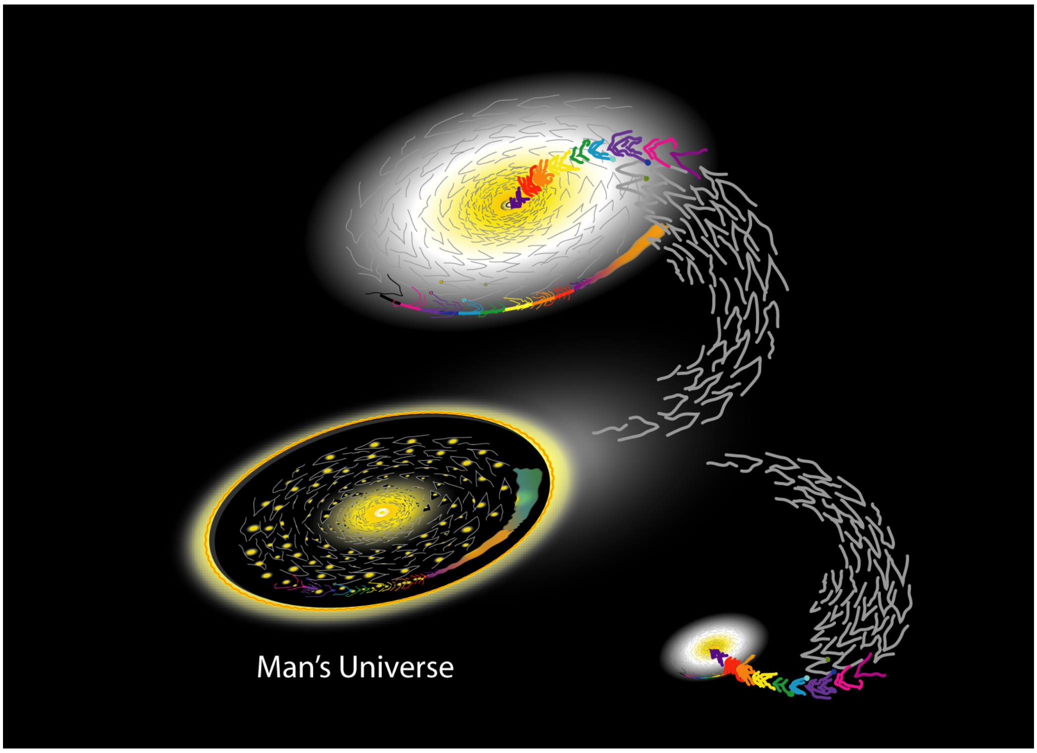

Fig 31. In time the decay of Man's Universe.

Where, the lesser mass part of the division of this universe will rotate around the larger mass universe. Where, the whole of the two new universes both will be rotating around larger universe mechanisms, which this universe on its own was rotating around.

Thus, if one is in the lesser universe, and looks into its motion, it looks as their universe is expanding, as they always seems that matters and fields in this universe to be running away from them as their universe is looking as being expanding.

Where, in fact being in the smaller universe, the gravitational field of the larger universe is continuously pulling or holding onto the lesser universe that the lesser universe stays within its orbit, as the electron does with its proton in atomic structure. As there can be no difference in all levels of creation of larger universe, as it is with the structure of decay of all atoms known to man.

Therefore this is how this universe was created and how future universe will be created out of this universe.

Thus there has never been a big bang, and there will never be a dooms day or the big rift. All it will happen is that this universe will go through a division, which is the norm in all-universal orders of creations.

The original universe has divided itself to smaller universes, which our universe is one of its offspring's and the same will happen over and over again to this universe and its offspring's.

Thus this universe is not the original universe, but part of a fraction of a fraction of fraction of a bigger picture.

May be now it can be understood that in the universe, as has been said that "I have made man in the image of myself" and now the reality of operation of different and all entities of the universe understood and are set in motion for the man to have the full understanding of method of creation and control and the way he has come to be on this planet at this moment in time and if he understood the purpose of his creation, then it can be said that:

The universe exists and operates within the structure of the man

and

the man exists and operates within the structure of universe

Definitions

Below is a list of the new terms and their definition that used in this book. A number of these words and terms have been used in the previous book and scientific papers, which have been published by the foundation and most readers, are familiar with these terms and reference. In this book a different approach and language and terminologies is used, therefore has become necessary to create a new terminology and words that can explain the real term application and processes in the world of physics.

Mafs: magnetic fields

Magnetic: with capital M refers to the plasma of magnetic fields created by a dynamic entity as for example this refers to the magnetic field of the earth.

Magravs: Magnetic and gravitational fields (Magnetic referring to the plasma of magnetic fields created by a dynamic entity as for example this refers to the magnetic field of the earth)

Pmf: plasmatic magnetic fields

Principal matter: This matter in conventional physics is called the anti-matter, as now the real structure of this matter is fully understood and is said to be source of all matters from the dark matter to matter, thus it has been called the original matter or the principal.

Principal mafs: These are magnetic fields, which are in the principal strength levels and considered to be I order magnitude several hundreds of times stronger than the magnetic fields which make up the tangible and visible matters.

Principal star: This entity in the conventional physics is called the Black hole

Transition energy: This energy in conventional physics is called the dark energy. This is the pmfs that are in motion from principal to matter magnetic fields strength so called the transition energy as energy is defined as the pmfs in motion.

Transition matter: This matter in conventional physics is called the Dark matter, as now the real structure of this matter is fully understood and is said to be when plasmatic magnetic fields are in transition state from the principal matter to tangible matter thus the term transition matter is used.

Transition mafs: These are magnetic fields which are in strength level between the principal and matter magnetic fields strength, and are in transit from the principal mafs to matter mafs strength levels.

Gans: a new state of matter; a molecule of a gas (GAs), which becomes Nano of itself and appears as Solid state of matter (Gas to Nano of Solid and to be called GANS for short).

The re-naming of matters and conditions

(*Advice: Skip this section unless you need to refresh or reading for the first time or read it after you have read the chapter 1, 2 and 3 as this section can be confusing and distracting you from the main body of the book*).

(Point of reference: The following section is and to be added to all disclosures for the ease of understanding as most of these have been disclosed in the book (1), and it is added here for the ones who have not read the book and are not familiar with terminologies in the works of the author).

In this literature, when using the term "**mafs**" for example in form of the antimatter mafs this means and refers to the magnetic fields (mafs) which this entity is made of and it is considered to be as a plasma of magnetic fields, and where the term "**antimatter**" is used on its own this meaning the physical tangible or detectable entity of the same. This usage of mafs similarly applies to physical matter and Dark matter too.

Further, the difference with Antimatter and matter is only in their plasmatic magnetic fields (**pmf**) strength and compactness of their fields. Thus, what is the magnetic fields (mafs) of antimatter they become magnetic fields of the matter as antimatter magnetic fields unwinds, loosens up, weaken in their field strength through their interaction with other magnetic fields and matters and become matter mafs and in plasmatic magnetic field strength of matter environment the matter mafs becomes tangible and become matter.

Therefore the antimatters are a more compact strength magnetic fields entities of the same of the magnetic fields of the matter, and due to their tightness of their fields, these magnetic fields of these mafs of matters they create stronger gravitational and Magnetic fields and hence they have stronger Magravs and hence upon their unwinding, they release magnetic fields and pmf which are faster, higher in strength and hence they look as to be strange strong and fascinating matters.

This means that antimatter entity of for example ten centimetres in diameter once ***unwinds*** in structure it becomes a matter of several thousands of meters in diameter entity in matter magnetic field strength environment.

Inevitably one should call the term antimatter the start matter and therefore antimatters should be renamed the Principal matter. Therefore there is nothing anti about these matters mafs and matters at their pmf strength as has been actually assumed in the world of science.

The ***Principal mafs matter*** itself has *different degrees of strength* by itself and *not all* magnetic fields of the Principal matters *have the same magnetic field strength*. Where, the principal mafs plasmatic magnetic fields strength has a spectrum of field's strength too.

As the magnetic fields of the principal mafs unwinds, it releases and loosen in its environment these magnetic fields (mafs) of the principal matters disperses into their environment, then they lose their strength through contact or interaction and friction with mafs' of matter and matters, and hence these principal mafs matters pmfs go for through magnetic fields strength of the Dark mafs matter strength first before through further reduction in strength, the dynamic pmf of mafs of the dark matter in motion reaches the pmf strength of the matter mafs level and hence becomes visible in matter pmf strength environment and becomes matter.

Now that the essence of existence of dark energy or dark matter can be explained in its realistic manner, and that is what has been called dark matter is in fact this is ***a magnetic field in transit from the principal mafs matter and matter strength to matter mafs strength and matter***.

Then as energy has been defined as “magnetic fields in motion” then the dark energy in fact is transition mafs matters in motion, or what is called the transition energy.

Then the dark energy can be called the Transition energy instead of the Dark energy, as the pmf in motion as this magnetic fields from the mafs strength of the principal matter in motion to reach to mafs strength of the matter level, this is only a Transition pmf condition, and transition mafs matter phase only.

Where, the interaction of mafs of transition matters with other of the same leads to and creates the Transition gravitational and Magnetic fields of their own or create what is called the Transition Magravs and in some cases this condition leads to creation the transition matter (dark matter) and consequently sometime interaction of mafs of transition energy leads to appearance and become transition matter (dark matter) in the cosmos.

Therefore, the magnetic fields strength, which makes the dark matter, is to be renamed as the mafs of the transition mafs matter and the dark matter as tangible matter becomes the transition matter, the dark energy to be renamed to transition energy.

The plasma of the principal mafs matters exists everywhere in the universe. This being in the centre of the universes, galaxies, stars, planets, plasma of proton and even in the centre of the plasma of the electron and in the lower order pmf strength environments which have been undetected up to now by man.

In reality in the centre of what is called a black hole in the universe, there exist and is the home of the principal mafs matters and principal matter. Therefore the black holes should be called the principal stars in the centre of galaxies, stars, plasma of protons and other entities like these.

The choice of the term star for principal centres of matters is correct as these principal matter concentration centres are like a star in the centre for the matter world like solar system, and the Principal stars are the place for the mafs of the Principal matters and the principal matters which their fields radiate outward into their surrounding as it does with the stars. Thus from this point on in this disclosure we call the black holes the Principal stars in the universe now that we understand the real source of their mafs of this matter(s) and its principal matters.

It can be understood now that where the intensity and the strength of stars and planets come partially from the existence of the principal mafs pmf strength, which has been released by the Principal, stars in the galaxies. This is to say that the seed of the pmf of the galaxies and matters needed for the creation of stars in principle are from pmf of the mafs of the Principal matter. Where, the interaction of the pmf of these mafs’ of the Principal matters leads to creation of initial strong Magnetic and gravitational fields needed for the creation and attraction of mafs of matters of plasmas of matters, the mafs of the transition matter and transition matters of the same region of the galaxies, plasma of stars, plasma of the proton and other entities of the same structure in the universe.
At the same time when using Magnetic with capital “M” this means the like of the Magnetic field of the earth, which itself is the plasmatic environment created by the interaction of at least two magnetic rays and indicates the outward flow of the magnetic fields from a plasmatic environment the like of the earth.

Similarly, when using magnetic field with small “m” this simply means the magnetic fields as in what we observe at the end of a solid magnet as magnetic fields of the object without its interaction with other fields.

Reference

1. Book 1, “The universal order of creation of the Matters”, M T Keshe, published July 2009
ISBN 978 46 6087 001 9

Book 2, “The structure of the Light” M T Keshe, published January 2011.
ISBN 978 94 6087 003 3

2. The unifying theory of the fields M T Keshe 2009

3. The origin of the Universe M T Keshe 2010

4. The creation of the solar systems M T Keshe 2010

5. The CO2 capture paper by MT Keshe Published 23.12.2009

6. “The universal order of existence” finalized and to be published in 2010-2011 by M T Keshe

7. The Magravs positioning M T Keshe 2010

8. NASA WMAP common.wikimedia.com

Published and unpublished works by M. T. Keshe

There are a number of unfinished articles that are to be converted to full disclosures when completed, like antimatter mafs method of energy releases, Nano space technology, relation between the creator and the created, space communication system, deep space defence technology, dynamic magnetic fields force regions (wormholes) construction and control, creation of magnetic fields, and so on.

Some of these works have already been published through different channels and some have not yet been released.

In the following section some of the works done by the same author are mentioned, with a brief disclosure of their contents.

Some of these papers can be purchased and obtained through direct request and access through specific links.

A number of papers with * and ** marks are not for public release due to their sensitivity of information and only available to relevant commercial and governmental organisations for peaceful applications of the technology.

The universal order of creation
(5.7.2006)

In this disclosure the main construction principles and the method of the link and control in the universal order of the existence of living cells, like the function and operational method of any matter based on so called proteins chains, as for example in human cell are explained in detail.

REF 9

The author considers this paper as one of his most important and fundamental works to be published and to be added to the sets of disclosures written by him.

Cosmic Rays
(Release date 24.3.2004)
This paper explains the source and the origin of cosmic rays and the way these work within our cosmos, their function and their benefits.

REF 10
----

*The Airborne reactor
(Release date 25.3.2004)

This paper covers the total design of airborne reactors for space technology. This covers all aspects of operation, design and control of such reactors.

REF 11

The seed of the Earth
(Release date 25.3.2004)

Some extracts from this paper:
The theory that the heat of the core of the Earth is due to inertia of the material piled on top of each other is completely dismissed through this new disclosure.

The author M T Keshe believes that this centre core is made of hydrogen, other gases and a mixture of liquid and solid matters mafs, which were in the vicinity of the centre of the core of the Earth at the time of its inception in the solar system.

Through this new understanding of the new inner cores of the planet, this brings to play the fact that the old inner-core of the Earth that is already assumed to be made of a solid piece itself becomes a sealed container for the new inner core.

The author through design of a parallel system has set to confirm the validity of this assumption, that the centre of the Earth possesses a semi- fusion atomic reactor and not a fission reactor as has been assumed by other scientists.

This theory of hydrogen in the centre of the planet is in line with the physical reality of the creation and the present situation in the physical construction of the solar system.

By the laws of physics all the gases in the solar system should have been in the giant gas planets at the outer layers of the solar system, like in Jupiter and Saturn.

The physical reality in the solar system is that the lightest of all these gasses, the Hydrogen, has taken its place right in the centre of the solar system in its star, the Sun.

It has to be recognised that the centre of all planets in possession of gravity due to the centre cores heating, are always in possession of two Magnetic Field Forces in their inner sanctum (M. T. Keshe “The creation of gravity").

In the case of the Earth up to now and with the present knowledge there has been acceptance of only one inner-core and one magnetic force.

As it is explained in the paper titled "the creation of gravity", there is an inner-core in the inner core. This inner core, due to it physical content, motion and position in the centre of the Earth, creates and maintains its own magnetic field force, independent of the magnetic force which has been known to be created by the interaction of the inner-core and the outer-core of the planet.

These two magnetic fields interact to create the planet's own double magnetic fields in its centre. The interaction of these two magnetic fields forces upon each other leads to the new concept of the "DOUBLE MAGNETIC FIELD EFFECT", where the interaction between these two field forces in the centre of the planet leads to the creation of the gravity and the magnetic forces of the planet (Paper: The creation of gravity).

This new second core is the seed core of the planet, the mother seed of the planet positioned in this new inner core.

REF 12

The creation of Black Hole

(Release date 25.3.2004)

In this paper it is explained that the Black Hole is created out of a natural sequence of events, and that the existence of the Black Hole in a galaxy is a normal event. Further it is explained that a Black hole is created in similar ways to the Dark Spots on the surface of the Sun. In this paper for the first time, it is explained how the Black Hole is physically created and what function it plays in its galaxy.

REF 13

Magnetosphere

(Release date 9.6.2004)

"The shape and the strength of the magnetosphere of a planet is exactly what a fingerprint is to a man. It is unique to each planet, star and galaxy. It is a tell-tale of all its attributes, and it reveals the hidden mysteries of the internal materials’ structure".

REF 14

Correction to the Einstein equation of Relativity

(Release date 15.6.2004)

In this paper the physical reality of the Einstein equation is considered, and efforts are made to bring this equation to the real conditions. Even Einstein considers this equation applicable for very small mass, and not in real three-dimensional multi-Magnetic and gravitational conditions, which are external and have nothing to do with the theoretical object's mass and speed, but have effect on the speed and the mass of the object under consideration.

REF 15

----

*Introduction to the new system
(Release date 2.7.2004)

In this paper it is written "It is essential to understand that, with the application of this system of motion, one cannot bring into play the solid, liquid and gas fuel method to create motion, as is used today. What is offered in this system is in simple words a fully integrated plasmatic magnetic energy system. Which, creates within and around the core environmental conditions as known to man in the universe like creation of gravity and magnetic field protection and the same as the magnetospheric protection condition of planets and motion through interaction of magnetic fields as planets and stars do in their given environment".

REF 16

The rings of Saturn
(Release date 2.7.2004)

One of the oldest puzzles in the world of astronomy has been the existence of the rings around Saturn. How did these rings come to exist the way they are now? How did they come to appear and behave the way they do? These questions and much more about the creation of these rings are all answered in this paper.

REF 17

The creation of gravity
(Release date 11.7.2004)

In this paper it is explained, "the source and the creator of the magnetic forces and the gravitational forces are due to the interaction between the same materials in the same region of the planet."

Further it is explained how the gravitational field force can be created, replicated and controlled within a nuclear reactor. Further it is explained how this property can be used for the motion of a system within a planetary or solar system.

Now by understanding the principle of the creation of gravity, in this paper it is further explained, "the gravity in reality is the effect of the interaction of two plasmatic magnetic fields of any two object in respect to each other."

REF 18

Death of a star
(Release date 26.7.2004)

In this paper the death of stars is explained, "The death of a star and the creation of a supernova is in reality the same as the half-life cycle energy step-down of an atom, with the difference that this show is in a larger scale, and more spectacular, in all aspects of its dimensions".

REF 19

Fusion
(Release date 28.7.2004)

In this paper the physical and scientific problems with present fusion technology is considered. Where, it is written, "nuclear Scientists, with the fusion reactors of today, are trying to create fusion in the opposite way to the known laws of physics and their parallels in the universe". Further in this paper it is written, "If the present path for creation of fusion is followed, with the present scientific knowledge of physical materials for construction of such system, then following the present path of development to create energy from current fusion systems will be a distant dream".

REF 20

Life of a Cell
(Release date 28.11.2004)

Extracts from this paper; "The reality of the transportation and transmutation of a cell is not far from the reality of the life of an atom or a star. Where the life cycle of the cells are slightly more complicated, as they contain other matters mafs like acids, which have their own magnetic characteristic of their own chemical structure. For this reason the control and replication of the energy of the cell is very much complex, but very simple to achieve".

REF 21

The Atom
(Release date 19.12.2004)

The content of this paper explains how and where atoms in the universe are originated from and how the construction of atoms in a cold plasmatic nuclear reactor can be replicated.

REF 22

Magnetism
(Release date 8.1.2005)

In this paper it is written, "The question is where did the Magnetism come from? This is the Secret of creation. The world is made of one thing and only one thing and that is Magnetism. Its combination and interaction with its own different strength fields makes it appear as different matter mafs. The rest of creations are the outcome of these interactions of different strength of this thing called magnetism and the forces it creates. Magnetism is the origin of existence and the real singularity in its full meaning".

REF 23

----

*The creation of the magneto-gravitational force
(Release date 5.2.2005)

The creation of magneto-gravitational field force is and will be fundamentally independent of temperature and pressure, proviso to the fact that as long as a homogenous free floating pmf within a matter mafs is created.

REF 24

The core of the Earth
(Release date 20.4.2005)

One of the original assumptions in developing the new Magravs system some thirty years ago by M.T. Keshe was on the basis that the centre of the Earth possesses addition cores, which create the necessary gravitational conditions of the planet. After some twenty years of space technology and through Earthquake seismological data's this assumption has been shown to be correct and after over twenty years, it has been scientifically proven that the inner-core of the Earth has an inner-core of 600 km. Professor Guy Master in 2002 ref 24 nuclear planet. The centre of this core is assumed to have an inner-core of 8 Km, this he thought to be made of the plutonium or other nuclear materials

But the developer of this new plasma reactor technology has proven through physical dynamic cores that these cores possess a mixture of hydrogen plasmas and the process of the heating of the centre of the Earth are due to a semi-fusion chain of events.

REF 25

----

*The electron-atomic welding
(Release date 3.5.2005)

Some extracts from this paper:
"The electron atomic welding or atomic welding principle is the phenomenon that atoms of the same material, through a common shared electron, become a magnetically balanced molecule of the same matter mafs, with the difference that the shared electron will create a balanced pmf in the molecule".

REF 26

----

*Shutdown and safety of the reactor core
(Release date 6.6.2005)

In this paper the safety parameters in the operation of the gravitational and energy reactors are simply explained to avoid the loss of plasma and gravitational field forces created within its cores.

REF 27

----

*Reactor Start-up
(Release date 9.6.2005)
In this paper the systematic start-up of the dynamic reactors for the creation of gravitational field forces and the production of energy are set out. The start-up of these systems is totally different than the prior arts known in the nuclear industry.

REF 28

----

*The energy balance of the reactor
(Release date 29.6.2005)

Some extracts from this paper:
The energy balance of these types of reactors is not as simple as might look in the first instance. The energy creation is simple. But the heat leakage and dissipation through the deliberate inherent material design are and can be zero. That is to say the system not only can hold on the heat it creates. At the same time due to the close circuit feed operation will become self-sufficient but long lasting.

In planets, the loss of heat through their surface creates the variation in the central core condition over several billions of years. In this reactor as losses can be literally negligible, the system can operate at low energy loss temperatures. Even the boundary of the body could be made through the right combination of material in the chambers, so that the body of the reactor will be colder than its surrounding boundary. So the system will not only lose energy to its surrounding, but it can be made to absorb heat in the negative temperature gradient from its surrounding, so that there are no losses.

REF 29

Nuclear decay or half-life
(Release date 20.7.2005)

Extracts from the paper:
"The nuclear decay always has the same or near the same time duration. This is due to the fact that the all neutrons and protons, from their inception, carry a predetermined level of plasmatic magnetic energy. This energy is continually used for the motion and vibration of different elements of the nucleus in respect to each other and the surrounding matters mafs of the atom. As the space and dimensions of a nucleus of an atom always obeys the same principles of the magnetic separation and magnetic attraction. Protons or neutrons can and will need to use or consume the same energy before they become weak enough that the splitting or decay becomes a clockwork job. The nuclear decay is the natural levelling done of a nucleus' energy, due to the energy consumption and heat dissipation for the motion of the elements of the nucleus for its remaining plasmatic magnetic fields energy necessarily needed for them to hold one together".

REF 30

----

*The decontamination system
(Release date 10.10.2005)

"What this means is that, for example in trying to recycle the CO2, by the use of the right matter mafs in one of the cores of the reactor, the system will then use the O2 to produce H2O, in the form of pure water, and C in the form of atomic or molecular carbon, or even by feeding this back through certain operational compressions and gravitational interactions within a small core, to create industrial diamond or graphite, for different industries."

The use of the reactor for this purpose has been proven to be correct and Raman spectroscopy has proven the concept and physical reality of the separation of matter mafs in atomic level after separation in the same system from a composite matter mafs.
(See the graphene on the technology web site)

REF 31

The difference in atomic and molecular structure
under gravitational force and pressurised conditions.
(Release date 15.8.2005)

In this paper is written that "the atomic and molecular structure of matter mafs is totally aligned and positioned in a fundamentally different configuration where it has come together due to gravitational forces, than when matter mafs has been brought together by the means of pressure. This has a fundamental effect on the behaviour of the structure and properties of the matter mafs which is created in either way". Where this is explained and utilised to develop a new method in creating graphene and Sp3 atoms and walls in a simple core made of a coca- cola bottle.

REF 32

**The Defence and shielding system
(Release date 4.9.2005)

Some extract from this paper:
The design of the side open system for lunch of high saturation magnetic plasma package is one of the most effective technologies for defence of the reactor and the craft, which is in possession of such a system.

It is essential for any space technology in the open environment of the universe, if the craft is to keep a straight line of motion and to be able to protect itself from any solids or matters mafs in its path in the cosmos.

This technology can destroy any object by the principle of plasmatic magnetic energy saturation of the incoming object at a molecular level, that object will disintegrate in the form of magnetic vapour and not atomic level destruction, before matter mafs can damage the craft as debris, and before the craft ever becomes in touch or near these space matters mafs.

This technology can also be used to protect Earth from the collision risk of NEO's (Near Earth Objects), like comets and asteroids.

REF 33

Method of creation of an atom or nano matter mafs
(Release date 11.10.2005)

The manufacture of an atom of any density, this being of the simplest form of nucleus or a full atom of hydrogen, or even heavier elements in the universe, all follow the same principle.

Atoms are created in the soup of the cosmos from collections of very weak plasmatic magnetic fields, which for their existence cluster and share their energies with the magnetic fields that are not far from their own plasmatic magnetic field strength.

REF 34

The relationship between Gravity and Mass
(Release date 21.1.2006)

Extracts from text of this paper:
"A planet like Earth possesses a combination of both gravity and inertia. Where the gravity comes from the interaction of plasmatic magnetic fields from the processes within its core, and the inertia comes purely from the collection molecules pmf of matters mafs which has built the physical body of the planet.

Any object, being an electron, atom, molecule or even a human body, is a collection of plasmatic magnetic fields of different strengths and their interactions between each other, which in total decides the total or collective magnetic field possessed by that object where this is the mass of the object.

Where the mass of an object is a collective package of plasmatic magnetic field tightness of a given object and will not change as long as the object is intact as one entity in its overall atomic or molecular or collective molecular appearance.

Which plasmatic magnetic energy of a physical body (so called the mass of the body) in interaction with molecular or atomic magnetic field (this being gravitational or inertia) of another object will determine the weight of the two object in respect to each other in each other's environment".

REF 35

The Dark matter mafs
The Fifth state of matter mafs
(Release date 21.1.2006)

Extracts from text of this paper:

The existence of dark matter mafs is not questionable,

Proving its existence, and its effect on the energy balance of the universe, in all matters mafs as small as in the heart of an electron or in a solar system needs to be created, replicated and its effect measured.

The theory of dark matter mafs does not and needs not to be considered as complicated, if one understands in real terms the plasmatic magnetic energy of the matter mafs in the visible and invisible dimensions of the universal work.

The dark matter mafs has two distinctive characteristics, which makes it totally apart from the visible matter mafs. The dark matter mafs possess mass but not visible light, where its existence can only be determined by the weight of the it's hidden mass, which could be substantial due to its internal gravitational field.

REF 36
----

*The inter-atomic fusion
(Release date 3. 5. 2006)

Extract from this paper:
"The fusion of two or more atoms has been the pre-occupation of nuclear physicists for past years. Conditions to achieve fusion of two hydrogen plasmas in the TOKAMAK reactors have taken years to teach scientists a lot about the behaviour of plasma and their fusing.

The reality about the fusion in some way has to be reconsidered as, if the scientific world is trying to amalgamate the contents of two plasmas to release energy. There must be a simpler way to achieve releases of similar energies.

The fundamental principle of the inter-atomic fusion is a much simpler way to achieve fusion. If this principle is applied to the atomic condition of fusion, then fusion will be attained in a simple but in a much more physically realistic environment.

The explanation for inter- atomic fusion is very simple and direct.
In the fusion of two plasma of proton of hydrogen atom, the physicists try to fuse two large plasmas, and by doing so, they try to release a large amount of energy. Where in this process enormous amount of magnetic field forces and currents are needed to bring two plasmas in close proximity, so that the energy barriers in between them can be overcome for them to amalgamate or fuse.

In the inter-atomic fusion the overcoming of energy barriers does not exist, as in this method of fusion, the electron of the atom which is circulating the nucleus of the atom and is of the same origin will be encouraged to return into and fuse with the nucleus of its atom".

Where this is a more practical way to create and release manageable energies and there is no need for elaborate systems where even if the energy release will be small compared to dreamed up fusion systems of today.

With inter-atomic fusion method at least small and practical systems can be developed to handle low temperatures of the fusion process.

REF 37

The conductivity of the matter mafs in the vacuum
of the interplanetary mediums
(Release date 21.1.2006)

Some extracts form this short paper "Therefore plasmatic magnetic field of elements within the interplanetary medium will cover the space given to them in their environment. Where due to their motion and their electric charges even in eV levels, they become perfect conductor, and in turn, due to their motion on their environment, they become plasmatic electromagnetic generators.

REF 38

The production of Graphene
(Release date 25.7.2006)

Repeated experiments and tests in simple static reactors, and tests in more complex dynamic reactors, prove that atomic separation and recombination of matter mafs like carbon and hydrogen can be reached at room temperature and at atmospheric conditions.

We have now indications - through static and dynamic tests in our reactors - that the universe was made in normal condition, which was originally nothing but packages of plasmatic magnetic fields of different strength, which were themselves nothing but areas of plasma or collections of loose magnetic fields energies. Where magnetic fields of different strength in locking to each other, by the principle of their plasmatic magnetic energy (PME), have caused in the first stage the creation of fundamental particles, secondly atoms, then molecules and then matter mafs, clouds and asteroids and then stars and galaxy's".

The Big Bang theory conditions have no room in the reality of the creation of the universe.

REF 39

Further works on medical research and developments have been written and completed, and will be published on the Foundation medical site in December of 2011

The process of the reversal of coma: Published in December 2010

Multiple Sclerosis: Published in December 2010

Epilepsy: Published in December 2010

Fibromyalgia: Published in December 2010

INTERNET LINKS

Web site: www.keshefoundation.com

YouTube Channel: http://www.youtube.com/user/keshefoundation

Forum: http://www.keshefoundation.com/forum/

Internet radio interviews:

Interview 06-09-2011 with Red Ice Creations (hour one):
http://www.redicecreations.com/radio/2011/06/RIR-110609.php

Interview 06-09-2011 with Red Ice Creations (hour two - need for simple registration):
http://www.redicecreations.com/radio/2011/06/RIR-110609.php

Interview 05-09-2011 radio interview with Alienshift:
http://www.livestream.com/alienshift/video?clipId=pla_8251665e-ac74-4d77-a785-443bbca65bf6

Interview 04-27-2011 with Mrs. Kim Greenhouse of Its RainmakingTime:
http://itsrainmakingtime.com/_radioshows/110427Keshe.mp3

Interview 04-20-2011 with Mr. Redneck Radioman of blogtalkradio.com:
http://www.blogtalkradio.com/paranormalpalace/2011/04/20/gravitational-technology-with-nuclear-scientist-mehran-keshe
----
Interview 04-17-2011 radio interview with Mrs. Aurora and Mr. Bill Alek of vortexnetworksnews.:
http://intalek.com/AV/VNN/VortexNetworkNewsHour/04-17-2011-VNN.mp3

Interview 04-09-2011 second radio interview with Mr. Bill Alek of vortexnetworksnews.:
http://intalek.com/AV/VNN/ProgressiveTechnologyHour/04-09-2011-Technology.mp3

Interview 04-02-2011 first radio interview with Mr. Bill Alek of vortexnetworksnews:
http://www.intalek.com/AV/VNN/ProgressiveTechnologyHour/04-02-2011-Technology.mp3

Interview 12-18-2009 radio interview with Mr. Allan Sterling of Peswiki:
http://www.mevio.com/episode/205187/fen-12-18-09-interview-with-mehran
----

LIST OF ILLUSTRATIONS.

FIG. 1. CREATION OF NEW UNIVERSES OUT OF THE DIVISION OF A LARGER UNIVERSE 16
THROUGH THE PROCESS OF DECAY SIMILAR TO THE DECAY OF THE NEUTRON IN PROTON AND ELECTRON. 16
FIG. 2. THE CREATION OF MAN'S UNIVERSE AFTER THE DECAY OF IT'S MOTHER UNIVERSE MAGRAVS' MAF'S...17
FIG. 3. THE UNICOS 19
FIG. 4 THE POSITION OF THE SISTER UNIVERSE IN RESPECT TO THE MAN'S UNIVERSE. 20
FIG. 5 THE UNIVERSE HEAT DISTRIBUTION MAP (W-MAP OF THE UNIVERSE) 21
FIG. 6 THE MICROWAVE HEAT DISTRIBUTION OF THE UNIVERSE SHOWS A CONCENTRATED HEAT..........................22
IN THE CENTRE DISC OF THE UNIVERSE 22
FIG. 7 THE CONVERSION OF THE PLASMA OF NEUTRON TO PLASMAS OF THE PROTON AND THE ELECTRON. 24
THE SAME PROCESS IS CONSIDERED TO OCCUR WHEN THE UNIVERSE DIVIDES (3). 24
FIG.8 DECAY OF PLASMA OF NEUTRON IN PLASMA OF PROTON AND PLASMA OF ELECTRON. 24
FIG. 9. THE PRINCIPAL UNIVERSAL LINE OF MAF'S. 25
FIG. 10. THE PRINCIPAL MAF'S OF THE UNIVERSE. 25
FIG. 11. THE UNWINDING OF MAGRAVS IN AN ENTITY. 26
FIG. 12 PLASMA OF THE UNIVERSE: ANY ENTITY CAN BE CREATED ANYWHERE IN THE UNIVERSE AS LONG AS 27
THE APPROPRIATE MAF'S STRENGTH OF THAT ENTITY IS REACHED THROUGH REDUCTION OR INTERACTION OF
MAF'S OF THE UNIVERSE 27
FIG. 13 PLASMA OF GALAXY: ANY ENTITY CAN BE CREATED ANYWHERE IN THE GALAXY AS LONG AS THE
APPROPRIATE MAF'S STRENGTH 27
OF THAT ENTITY IS REACHED THROUGH REDUCTION OR INTERACTION OF MAF'S OF THE GALAXY ITSELF OR ITS
UNIVERSE'S MAF'S 27
FIG. 14 PLASMA OF NEUTRON: ANY ENTITY CAN BE CREATED ANYWHERE IN THE NEUTRON AS LONG..................28
AS THE APPROPRIATE MAF'S STRENGTH OF THAT ENTITY IS REACHED THROUGH REDUCTION OR INTERACTION.....28
OF MAFS OF THE NEUTRON AND IT'S GALAXY ITSELF OR IT'S UNIVERSE'S MAF'S.....................................28
FIG.15 GALAXY (SOURCE: HUBBLE) 29
FIG. 16. CONVERSION OF THE SPHERICAL PLASMA OF A NEUTRON IN A CYLINDRICAL RAY OF LIGHT. 31
FIG. 17 THE COLOUR OF RAINBOW ACCORDING TO THE MAN'S PMTICS STRENGTH IN EARTH MAGRAVS STRENGTH.
....32
FIG. 18 (A TO G). THE PROCESS OF THE DIVISION OF THE MOTHER'S UNIVERSE IN TO MAN'S UNIVERSE 39
AND IT'S SISTER UNIVERSE 39
FIG. 19 THE PRINCIPAL LIGHT INSIDE A BLACK HOLE
48
FIG. 20 THE ABSORPTION OF MATTERS AND FIELDS BY PRINCIPAL STARS OF THE UNIVERSE OR GALAXIES. 49
50
FIG. 21 PRINCIPAL LIGHT IN THE CENTRE OF THE UNIVERSE. 50
FIG. 22 ENVELOPE OF GALAXIES IN EACH UNIVERSE INDICATING THE POSITION OF THE SISTER UNIVERSE. 64
FIG. 23. SOME OF THE PROTOTYPES USED FOR CREATION OF MAGRAVS. 78
FIG. 24. THE STRUCTURE OF INNER CORES OF THE EARTH (12). 79
FIG. 25. EXISTENCE OF MINIMUM TWO SEPARATE MAGNETIC FIELDS IN TWO CORES OF EARTH. 81
FIG. 26 INTERACTION OF TWO PLASMATIC ENERGY FIELDS CREATING THE EFFECT OF A BLACK HOLE. 94
FIG. 27. PLASMA OF PASSIVE MAGNETIC FIELDS PUSHING EACH OTHER CREATING 96
A BALANCED MAGNETIC FIELD REGION, CALLED BLACK HOLE 96
FIG. 28 TEST REACTOR 106
FIG. 29 . LETTER OF ROYAL ASTRONOMICAL SOCIETY 107
FIG. 30 THE POSITION OF THE SISTER UNIVERSE IN RESPECT TO MAN'S UNIVERSE 118
FIG 31. IN TIME THE DECAY OF MAN'S UNIVERSE. 121

INDEX

A

antimatter 127

B

black hole 128

C

centre of the Earth 134
CO2 135

D

dark matter 127

E

Earth 89, 90, 91, 92
energy 128

F

fusion 133, 137

G

gans 32, 125
Graphene 138
gravity 132

L

light barrier 31
light speed barrier 31

M

Mafs 125, 127
Magnetic 125
magnetic shielding 92
magnetism 133
Magravs 15, 16, 19, 21, 125
Magravs positioning 85
MPS 89

N

nano materials 32
Nuclear decay 135
nucleus 135

P

Pmf 125, 127
Principal mafs 125
Principal matter 125, 127
Principal star 125
PULM 24, 25, 27

R

rainbow 31

S

Saturn 132
shielding 136
solid magnet 72, 73, 82, 86, 88, 128
sound barrier 30
space technology 89
supernova 132

T

Transition energy 125, 128
Transition mafs 125
Transition Magravs 128
Transition matter 125

U

UMMS 89
unicos 16, 22
unisphere 22

As we are helping with series of books which have been published and to be released, to build the full knowledge of the structure of the operation of the magnetic fields and their interactions in the world of creation for man to understand the entire process, one comes to the point that one has to explained different proportion of this entity in different levels of creation of the greater universe and not the man's universe. In the ***first book*** *the creation of of matter, the operation and interaction of smaller fields, which leads to creation of matters was explained. In the* ***second book*** *we merely explained the method of travel of the collection of fields in the form of so called light. In this* ***third book****, it is further I have explained the operation and interaction of a bigger cells of the same fields, which these leads to creation of the mans universe and more and what is too much for the man to understand with his present scientific knowledge and he calls it his universe.*

Where, the so called this universe in this book is explained itself is to be a part of natural process of creation and it has been created from the division of bigger quantities of fields and in time this universe will itself will be divided into and will create smaller universes or part of it will amalgamate with other universes and their fragments' to create bigger universes in the overall structure of ***the unicos,*** *the universal cosmos.*

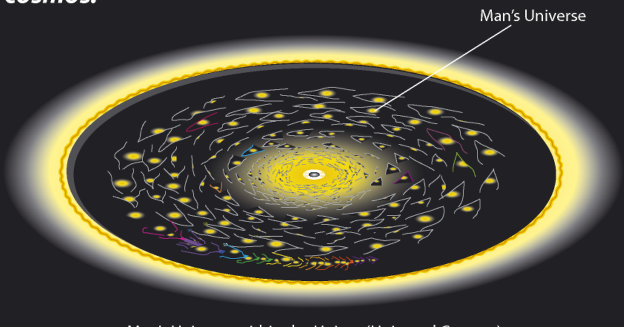

Man's Universe within the Unicos (Universal Cosmos)

This process is being the like of the division of the plasma of the neutron to plasma of the electron and proton to create an atom and/or the fusion of two proton to create a heavier bodies of plasmas.

Thus this universe is not the centre of creation and itself is part of the bigger structure.

That is the same dilemma as man had as he tough the earth was the centre of the universe and as he understood more, he realised that he, his planet and himself are the smallest part of solar system and his solar system was in fact smaller part of a galaxy and so forth, and this is the same with operation and existence of the of the mans' universe in the bigger operation and structure of universes in the cosmos of universes: the Unicos.

The Keshe Foundation

The Keshe Foundation is an independent non-profit and a non-religious organization founded by nuclear engineer M.T. Keshe, aiming to bring new scientific knowledge and new technologies to mankind, and to bring new solutions for the major global problems, like famine, water shortage, shortage of electrical power supply, climate change, illnesses, by the use of new types of plasma reactors, and to give Man the real freedom to travel in Space.

90000 >
978-94-6087-000-2

THE ORIGIN OF THE UNIVERSE

M.T. KESHE

The Keshe Foundation

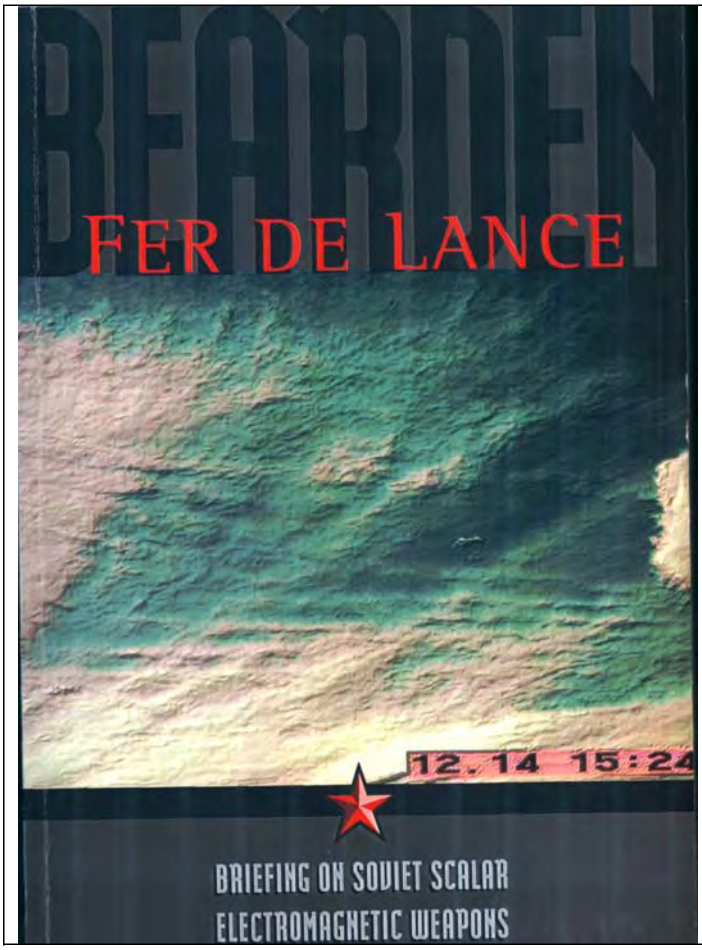

BEARDEN

FER DE LANCE

12.14 15:24

BRIEFING ON SOVIET SCALAR
ELECTROMAGNETIC WEAPONS

FER DE LANCE

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) PhD, is the author of numerous books, papers, briefings, and presentations dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare.

Lt. Col. Bearden was previously a senior scientist and department manager with the Alabama division of a major aerospace company where he directed the company's design and production of expert systems (artificial intelligence) applications for the U.S. Army command control systems. He has more than 29 years experience in air defense systems, tactics, and operations; technical intelligence; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements.

Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in Mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College.

When Fer De Lance was originally written, Tom was a member of Mensa, Society for the Investigation of the Unexplained, International Tesla Society, American Association for the Advancement of Science, Association of American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, and the American Defense Preparedness Association. He is currently the Director of the Association of Distinguished American Scientists (ADAS), Fellow Emeritus of the Alpha Foundation's Institute for Advanced Study, and is CEO of CTEC, Inc.

In private life, Tom has two sons and two daughters, and lives in Huntsville, Alabama with his wife, Doris.

"...In 1999, the Earth's first strategic sub-space war was silently fought and won..."

ISBN 0-9725146-1-9

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. Since this book talks about the most terrifying weapons present today the information in it is even more of value. But knowing an old saying »the knife cuts both ways« one can also use the same knowledge in order to heal ourselves since it seems that our body already does it using the physics described inside. Hopefully the latter will prove to be the most useful way to use this old but forgotten / supressed knowledge.

The contents include the 2002 update and the older 1986 briefing with slides.

TABLE OF CONTENTS

2002 OVERVIEW AND BACKGROUND....4
Mad Scramble of Soviet Science After the End of WWII....5
One Institute Had 2,000 PhDs and a Full Support Staff....7
Situation in the Early 1950s....8
1963 Deployment and Immediate Use of Scalar Interferometers....11
An Example of Russian Technology Astonishing Western Scientists....12
Softening of Metal at a Distance Was One Effect Used Against the Challenger 13
Energetics: Three Branches and a Strategic Deception Plan....14
Strategic Attacks Scheduled and Why We Were Not Destroyed....15
Attempts to Alert the Nation....18
The Responses....20
Effective Survival Treatment of Mass Casualties Could Be Developed....20
The Official Silence Was Deafening....22
The Long Dry Spell in Correcting Electrodynamics May Be Slowly Ending... 24
Engineerable Unified Field Theory Is Now Clearly Appearing....24
Examples of Some Major Non Sequiturs in Conventional Electrodynamics 25
Wide Confusion of Cause and Effect....26
CSRs: Engineering the Unobserved Causal Side To Create Functional Systems....27
Making the First CSR is Extremely Difficult; Making a Million Copies Is Easy....28
In 1999 The KGB Deployed Massive Numbers of CSRs for a Strategic Attack29
How the Asymmetrically Deployed Strategic Attack Was Stopped in Its Tracks....31
The Mostly-Ignored Time-Energy Interaction....32
Development of Threats and Scheduled Attacks Aborted....35
Live Destructive Tests Over and Near the U.S. Have Continued....38
The Scientific Basis For Energetics Weapons Is Established....39
Our Situation Here In the U.S. Is Precarious....39

The Dominant Weapons Are Quantum Potential Weapons 41

Cellular Time Reversal: Revolutionary Healing by a Team of French Scientists 43

Becker's Study of the Cellular Regeneration System 46

Need to Correct and Extend Classical Electrodynamics 48

The Situation Is Serious 48

A New Manhattan Project Is Warranted 49

2002 ENERGETICS WEAPONS THREAT TO THE U.S 51

The U.S. Narrowly Escaped Energetics Destruction Twice in 1997 51

Initiation of the Brazilian Energetics Weapon Program 52

Related Phenomena 54

Stalin *Personally* Initiated the Soviet Energetics Program with the Greatest Urgency 55

A Great Technical Breakthrough in Physics Can Always be Had in Four Years 55

A Typical Example: The Suppression of Cold Fusion 57

Revolutionary New Processes Are Hidden in the Cold Fusion Experimental Results 59

Stalin's Dictatorial Impact on Soviet Scientific Management's Response 62

Our Own Science Community's Progress During These Decades 63

U.S. Scientists Have Committed Monstrous Crimes and Gone Unpunished 66

Energetics: New Principles of Physics and How the Soviets Found Them 67

Longitudinal EM Wave Interferometry Energetics Was Almost Immediately Uncovered 68

By 1950 Russian Energetics Was Off and Running 69

Deployment of the Soviet Energetics Weapons and Their Quick Use 71

Why the Soviets Could Continue to Attack and Kill U.S. Targets with Impunity 73

The Science Behind the Soviet Energetics Weapons 77

The Importance of Scalar Photons and Time-Density Waves 80

Psychoenergetics: Engineering the Mind-Matter Interaction by Engineering Mind 81

Quantum Potential Weapons Are Utilized in All Three Energetics Branches... 84

The Importance of Longitudinal EM Waves 84
Energy, Wave Energy, and Wave Speed are All Dramatically Altered 86
A Process for Creating Longitudinal and Time-Density EM Waves 87
Quantum Potentials and Multiply Connected Spacetime 87
Need for Caution with QPs Used Against Biological Organisms 89
A Hypothetical Example 89
The QP Weapon Is a Superb Energy Amplifier "For Free" 91
Other Novel Uses of QP Weapons 93
The U.S. Narrowly Escaped Destruction Twice in 1997 Alone 95
Purpose of the Novel Attack on Captain Button 96
Rogue Japanese (Yakuza and Aum Shinrikyo) Have a Role 100
Inducing the Gulf War Disease 100
40-Years of Russian Induction of Diseases In U.S. Embassy Personnel 104
Western Electrodynamics Is in Sad Shape 105
Toward a Deeper Electrodynamics 108
Time Is Highly Compacted Energy 109
The New Transduction Model Explains the Cold Fusion Results and Mechanisms 109
Why There Is a Time Delay in the Onset of Cold Fusion Phenomena 1ll
The Early Soviet Longitudinal Wave Interferometers 112
Khrushchev Killed the Thresher to Prevent His Own Dismissal 112
Khrushchev Demonstrated the Powerful Blast Capability of Strategic Interferometry 114
Weather Engineering Over North America 115
Earthquakes to Order 117
Readying the Weapons for the 1985 Readiness Demonstration Tests 119
Meeting the 1985 Schedule for Readiness 121
KGB Crews Regularly Practiced Against U.S. Rocket Launches from Cape Canaveral 122
Kill of the Arrow DC: Practice Turned into Performance 123
The Arrow Pilot Sensed the Lack of Power, But the Aircraft Was Immediately Hit 124

Elements of the U.S. Government Deliberately Suppressed Vital Evidence... 125
Other Direct Energetics Signatures on the Arrow DC-8 126
Other Acts of War Completed the Milestone 127
Kill of the Titan Missile Over Vandenberg Air Force Base 128
Dead-Man Fuzing and Outside Assistance Deterred a Full Strategic Attack in 1986 129
What Really Happened in the Chernobyl Disaster 129
The High Side Risk Factor 131
The Low Side Risk Factor 132
Consequences of Not Taking Counterfire Action 132
My Recommendation to the Counterfire Group 132
The Group Fired and Took Out a Transmitter 133
Results at the Transmitter and at Chernobyl 133
Kill of the Challenger Included Use of Metal-Softening 135
Cold Molding: An Epilog to the Metal-Softening Donnybrook 136
Additional Update from Year 1998 to Year 2,000 Has Been Given 137
There is Also a Severe Bioenergetics (EM Biological Warfare) Threat 139
National Institute of Health Reaction 141
How the New Revolution in Medical Healing Works 142
What the French Researchers Showed 145
How the Regeneration Mechanism Was Finally Deciphered 147
Cellular and Body Rejuvenation are Possible, Once the Technology is Developed 149
How Many Americans Could Be Saved? 149
There Are No Resistant Strains to the New Technology, and There Cannot Be Any 150
Results of the Priore' Treatments Are Well-Documented and Independently Supported 151
The Present Threat Is Urgent, and Our Strategic Destruction is Still at Issue .151
Desperately Needed Defensive Capabilities Can be Provided by a QP Solution 152
In Conclusion 153

1986 FER DE LANCE INTRODUCTION & BRIEFING ON SOVIET SCALAR ELECTROMAGNETIC WEAPONS 157

Scalar Electromagnetics is Electrogravitation 157

Fer-de-Lance 157

Energetics and Directed-Energy Weapons (DEWs) 158

The Soviets Use a Deception Plan 158

Destruction of the Challenger in January 1986 161

Technological Surprise and a New Hiroshima 164

We Have Given Up Defense in the Larger Sense 164

1986—THE BRIEFING 167

Slide 001: Characterization 169

Slide 002: Outline 171

Slide 003: Elements of the Emerging Theory 172

Slide 004: Three Kinds of Electromagnetics 174

Slide 005: Bohm-Aharonov Effect 176

Slide 006: A-Field is Real 177

Slide 007: Quantum Non-Locality 179

Slide 008: Vector Mathematics has a Fundamental Problem 180

Slide 009: The Zero-Vector Axiom 182

Slide 010: Are These Systems Equal? 184

Slide 011: Infolded Systems are Excluded by Physics 185

Slide 012: Kaluza Geometry 186

Slide 013: A Potential is a Change in the Stress of Vacuum 187

Slide 014: Type of Potential Depends on the Particle(s) 190

Slide 015: A Potential Extends to Infinity & Involves the Entire Universe 192

Slide 016: A Natural Potential 194

Slide 017: An Artificial Potential 195

Slide 018: Producing an Artificial Potential 197

Slide 019: If No Observable Mass Flows: 199

Slide 020: Gravitational Potential: A Conglomerate of Stress Types 201

Slide 021: What Gravitational Field Is 202

Slide 022: EM Force Fields are Releases of Gravitational Potential via Observable Mass Flows 204
Slide 021: Scalar O-Wave Production 206
Slide 024: Scalar Electromagnetics 208
Slide 025: Rogers' Undersea & Underground Communications Systems 211
Slide 026: How does the Rogers System Work? 213
Slide 027: Scalar EM Waves are Emitted & Absorbed by the Nucleus 214
Slide 028: Nuclei of the Universe Continually Exchange Scalar EM Waves.. 216
Slide 029: Converting EM Field Energy to G-Field Energy & Vice-Versa 217
Slide 030: Normal EM Bleeds-Off G-Field 219
Slide 031: Scalar EM Forces 4-D G-Field 220
Slide 032: Detection of Electro-Gravitation 221
Slide 033: Scalar Wave Detector 226
Slide 034: Scalar Resonance 229
Slide 035: Scalar Resonance 231
Slide 036: Scalar Resonance Currents 234
Slide 037: Transverse Wave Interference 235
Slide 038: Transverse Wave Interference 236
Slide 039: Creating Energy at a Distance 239
Slide 040: Exothermic Mode (Scalar Interferometer) 240
Slide 041: Endothermic Mode (Scalar Interferometer) 242
Slide 042: Tesla's "Big Eye" for Remote Viewing 244
Slide 043: Site of Mystery Cloud 247
Slide 044: Mushroom Cloud from Sea off Japan near Russian Test Area 248
Slide 045: Mushroom Cloud Rising from Sea 249
Slide 046: Mushroom Cloud from Sea off Japan 251
Slide 047: Not a Known Natural Phenomenon 252
Slide 048: Explosive Events seen on Soviet Island 253
Slide 049: Explosive Events seen on Soviet Island 255
Slide 050: Explosive Events seen on Soviet Islands 256
Slide 051: Creating a Spatial 0-Field Pattern 257

Slide 052: Tesla Shield 258
Slide 053: Expanding Dome-like Phenomenon 260
Slide 054: Hemisphere & Globes 1977 261
Slide 055: Saryshagan Direction, September 1979 263
Slide 056: Continuous Tesla EMP Globe 266
Slide 057: Mysterious Light over North Pacific 268
Slide 058: White Sphere seen in North Atlantic 1976 269
Slide 059: Continuous Tesla Fireball 270
Slide 060: Terminal ABM System 272
Slide 062: ABM Defenses 275
Slide 063: Network of Virtual Transmitters 278
Slide 064: Network of Virtual Transmitters 281
Slide 065: Launch Phase ABM System 285
Slide 066: Soviet Command & Control Concept 287
Slide 067: Launch Phase ABM System (Anti-SLBM) 290
Slide 068: Detecting & Marking Missile Launch 291
Slide 069: Destroying the Launched Missiles 292
Slide 070: Shift to Anti-Sub Mode 293
Slide 071: Shifting to Earthquake Mode 295
Slide 072: Suppression Attack—MARVs (without defenses) 297
Slide 073: Suppression Attack—MARVs (with defenses) 298
Slide 074: Response to Counterforce Strike with No Defense 300
Slide 075: Response to Counterforce Strike with Defense 302
Slide 076: Clear Communications 305
Slide 077: Unique Communications 306
Slide 078: Mine Countermeasures—Orthodox System 308
Slide 079: Mine Countermeasures—Scalar EM Systems 309
Slide 080: Scalar Anti-Personnel Weapon 311
Slide 081 Afghanistan 312
Slide 082: Scalar Anti-Tank Weapon 314

Slide 083: Scalar Anti-Aircraft Weapon 315
Slide 084: Scalar Anti-Aircraft Weapon 316
Slide 085: An Effective Penetration Aid 318
Slide 086: Unique Capabilities: ATM/AD/AARTY 319
Slide 087: Unique Capabilities 320
Slide 088: Unique Capabilities 321
Slide 089: Unique Capabilities 322
Slide 090: Unique Capabilities: Penetrate Shielding 323
Slide 091: Multi-Purpose Weaponry 324
Slide 092: Multi-Purpose Weaponry 325
Slide 093: Multi-Purpose Weaponry 326
Slide 094: Unique Capabilities: Heat Extraction 327
Slide 095: Unique Capabilities: Cold Explosions 328
Slide 096: The Transmitter is the Major Weapon 330
Slide 097: Star Wars—The Final Frontier? 331
Slide 098: Comparison of EM Concepts (Chart) 333
Slide 099: Comparison of EM Concepts (Chart) 334
Slide 100: Comparison of EM Concepts (Chart) 335
Slide 101: Summary (1) 336
Slide 102: Summary (1) 337
Slide 103: Summary (2) 338
Slide 104: Summary (3) 339
Slide 105: Summary (4) 340
Slide 106: Soviet Defense Against Missiles 341
Slide 107: Surprise by Using Secret Weapons 342
Slide 108: Something Russia Doesn't Want the U.S. to Know 343
1986 SLIDE REFERENCES 344
1986 ADDITIONAL SLIDE NOTES 345
1986 APPENDIX I 371
SOME IMPORTANT INDICATORS AND DESTRUCTIVE TESTING.. 373

INDICATORS AND SUCH 373

FINAL DRESS REHEARSAL BEFORE THE BIG SHOW 384

THE SHOW BEGINS 400

MORE FRIGHTENING THAN THE MIND OF MAN HAS EVER IMAGINED 403

SCALAR EM INOCULATION AND ITS IMPLICATIONS 408

1986 APPENDIX II 413

COMPUTER WORMS AND VIRUSES, AND THE IMPACT OF SCALAR EM 415

LARCENISTS 416

ARSONISTS 418

TERRORISM 419

ANTIBODIES 419

NATIONAL DEFENSE 419

THE WEST'S VULNERABILITY 421

CORE WARS REVISITED 424

A COMMERCIAL WORM 424

ADAM OSBORNE'S WARNING 425

"NOW YOU'VE DONE IT!" 426

TEEN CHALLENGE 427

"PEOPLE ARE BASICALLY GOOD" 429

FEDERAL FUNDS 430

SCALAR EM INNOCULATION AND ITS IMPLICATIONS 431

1986 APPENDIX III 437

CONTINUING THE SCENARIO 439

2002 TABLE OF FIGURES 443

2002 ENDNOTES 447

The Russians have sustained the largest weapons development program ever launched by any nation, and they have kept it effectively hidden from prying Western eyes. I have called this program "Fer-de-Lance," after the deadly South American pit viper of that name.

The dreaded fer-de-lance is a snake of great agility and lethal effect. It often ambushes its hapless prey, striking unexpectedly and without warning. Its first sudden strike is usually lethal to its victim, which promptly expires in writhing agony. Since the Soviet development of scalar EM weapons has been designed for the same purpose, the name seems appropriate.

2002

Overview[1] and Background of the Energetics[2] Weapons Threat

T. *E.* Bearden, Ph.D.
LTC, U.S. Army (Retired)
Director, ADAS

September 15, 1998
Update

Note: We have placed this update as a section in front of the original paper. This leads to some redundancy, which remains to insure clarity.

---

[1] This is an updated and slightly sanitized version of a paper originally prepared for and transmitted to selected Senators and Congresspersons.

[2] The present name of the old KGB is the Federal Security Services. The leopard's spots remain the same, regardless of its new name—at least for a die-hard old communist faction of it. Presently the helm of Russia has been turned over to a new Premier who has an extensive KGB background. However, Putin has chosen to differ from the traditional KGB outlook and seems intent on becoming America's trusted cheap oil partner and supplier, in a bid to help his economy and pump much-needed funds into it. His friendship with George Bush may be one of the turning **points** of this new century He at least has achieved sufficient control over much of the KGB to get on with trying to solve his economic problems and break the old "die-hard communist" mold a **bit**

2002 OVERVIEW AND BACKGROUND

For some time, Russia and several other nations have possessed highly advanced "extended-electromagnetics" *(energetics)* weapons of a very novel kind, using a dramatically extended electrodynamics theory {1}. To comprehend these weapons requires a combination of non-Abelian electrodynamics in at least the 0(3) gauge symmetry {2}, general relativity, Bohm's interpretation of quantum mechanics {3}, and the use of time-domain (scalar) energy, fields, and potentials to directly enter distant 3-spatial points without propagating energy through 3-space. It also requires correcting many of the serious errors in classical electrodynamics. Probably the most elegant and applicable extant model to deal with these phenomena and weapons is Sachs' {4} unification of general relativity and extended electrodynamics, particularly as implemented in O(3) electrodynamics. O(3) allows direct engineering by modified EM means.

Most of these energetics weapons are more advanced than what has previously been known to the U.S. military, intelligence, and scientific communities, although parts of the communities are finally making real progress in understanding such "revolutionary" areas. For example, every major weapons lab on earth seems to have discovered longitudinal EM waves and longitudinal EM wave weapons—the scalar interferometry we have been discussing for so long {5}. A major contributing factor holding back U.S. catch-up is that the flawed foundations of Maxwell's 1865 theory {6} have not been changed by the Western scientific community.

Historically, the scientific community exerts resistance to substantial innovation. Often 40 to 50 years are required to do what can be done with a concerted effort in four years. The Manhattan Project in WW II is a notable example of where the scientific community made a maximum effort and did a difficult job in four years. Also, "skunk works" efforts continue to be utilized for very rapid development when something really advanced is to be done and done quickly.

With the U.S. substantially behind in these "new" strategic superweapons, a new Manhattan Project would appear to be advisable and warranted. There are at least some indications that such may actually be going on at last. We hope the indications are true.

A sobering development that occurred early this year (2002) was China's deployment of both quantum potential (QP) weapons {7} and *negative energy* electromagnetic pulse (EMP) weapons (a type which we originally called the *MindSnapper* to disguise its true nature) {8,9}. The QP weapons are the dominant weapons on Earth today, followed closely by negative energy EMP weapons in second place. We have not been able to ascertain exactly how many nations have the negative energy EMP weapons, but it is at least five at the time of this writing.

The U.S. is still well behind in the use of special "engines" (special curvatures of spacetime and their dynamics, but achieved with higher group symmetry electrodynamics) for directly engineering the mind itself, and for producing *nonmaterial* robot systems made totally of such engines—"causal system robots" which we will mention later. Several of the Western nations do accomplish work in the mind control area, but use a sort of "brute force" method oriented on frequency and frequency changes, etc. where the true active mode is the unsuspected transformations of the U(1) electrodynamics to a higher group symmetry unified field theory electrodynamics, accomplished by the human body tissues and structures themselves.

Western scientists have not realized that all EM energy in 3-space actually comes from the time domain *to* each point in 3-space and *returns from* that 3-space point back to the time domain.[3] Hence "propagation" of 3-space EM energy is actually the propagation of a "circulation" of EM energy between the time domain and 3-space, involving each 3-space point presently said to be "occupied" by the energy. Unfortunately, Western science is still unaware that *no observable exists in time,* a priori, since by definition it has no extension in the time dimension at all {10}. Instead, an observable *continually recurs* as a frozen 3-space snapshot of an ongoing dynamic 4-space interaction, due to the continual repetition of the observation process which is the application of a *d/dt* operator to the ongoing 4-interaction. Hence, we "see" physical reality (3-space reality) similarly to the way we view a "motion picture" series of successive frozen frames. The major result of this flaw (mistaking a 3-space observable as a 4-spatial cause) in Western science is that Western science still extensively substitutes effect for cause—e.g., in the notion of a separate 3-force acting on a separate 3-mass. No such situation exists in nature, and nothing "moves" or "changes" in 3-space alone. For an entity to exist in time, it requires at least 4-space because a time length is required. Nonetheless, the West is making progress. If some of its scientific dogma can be overcome, the younger scientists will break free and readily do the job required for the defense of this nation.

Mad Scramble of Soviet Science After the End of WWII

During WW II the West produced a great technical breakthrough and obtained the atom bomb as a decisive new weapon. While speeding up his own atomic

---

[3] Specifically, energy comes from the time domain to the negative 3-charge, thence to the positive 3-charge, thence back to the time domain in a *4-circulation* of EM energy. With extra energy *appearing* and present at the negative charge due to entering 3-space there, and with less energy present at the positive charge due to the energy *leaving* 3-space there, there is a delta in energy density (pressure) across a dipolarity, oriented from negative to positive. That is the exact basis of the T. T. Brown unilateral thrust force, and the force used in the "lifters" experiments today. To calculate the **force, th**e long-unaccounted Heaviside energy flow component must also be accounted— not just the weak Poynting flow component.

weapons program[4] as a result, Stalin also ordered his own Academy of Sciences to quickly derive the next great new technical breakthrough, and he held the top scientific leaders themselves personally accountable to him, to see that it was done at all speed (Figure 1).

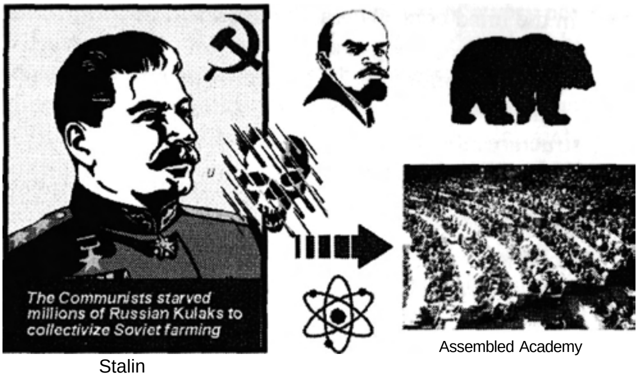
The Communists starved millions of Russian Kulaks to collectivize Soviet farming
Assembled Academy
Stalin

"The destiny of communism has been frustrated by the U S development of the atomic bomb That is not the last great technical breakthrough The next one WILL be Soviet! You will provide me with that great new technical breakthrough at all speed, or I will have your heads Do I make myself dear, comrades?"

Figure 1 Stalin's dictum removed any resistance from his scientific community.

Simply put, Stalin placed the heads of the Soviet scientific leaders on the chopping block, since he was an absolute dictator and often killed or imprisoned those who opposed his instructions. To survive, the leaders of the Soviet Academy had to demand the utmost of their subordinates, because they themselves had to report their progress or lack of it to a dictator who would and did take swift action if they did not produce results. Thus, Stalin simply bypassed the usual scientific posturing and protracted delay so typical of our Western scientific bureaucracies.[5] If the Directors of the Russian scientific community dragged their feet and did not push his desired program to the limit, their heads would roll—and some did.

---

[4] At the Potsdam conference, after Truman informed him the U.S. had just exploded a great new explosive weapon, Stalin secretly called the head of his nuclear weapons program and ordered him to accelerate the Soviet nuclear weapons program.

[5] As a simple example, the U.S. National Academy of Sciences, National Research Council, National Academy of Engineering, and the National Science Foundation—and all the electrical engineering departments at all the universities in the U.S.—have not yet recognized that all EM circuits and systems are in fact powered by EM energy extracted directly from the seething local vacuum by the source charges and dipoles in the circuit or system, and particularly by the broken symmetry of the source dipole in the generator or other primary power source. They also have not yet solved the more than a century-old source charge problem, but continue to support engineering models assuming that the charge *freely creates* real EM energy out of nothing at all, in total violation of the conservation of energy law, and pours that created energy out, freely and continuously, to establish its associated fields and potentials and their energy reaching across the universe. The agencies thus unwittingly advocate perpetual motion machines on a scale unparalleled in history. They continue to support an electrical engineering model and a classical

As a result of Stalin's forceful intervention, the entire Soviet scientific community was galvanized into a most intensive review of all of physics. The scientists desperately sought to quickly uncover potential breakthrough areas that could be developed into the great leap forward that Stalin demanded. They had to have progress, and they had to have it rapidly.

One Institute Had 2,000 PhDs and a Full Support Staff

Large review institutions were quickly set up, staffed with the best Soviet scientists and support translators and teams. One such search institute involved some 2,000 Ph.Ds along with their support staffs. Soviet ships brought in shiploads of copies of all the scientific journals of the West—from the beginning—, and they were reviewed paper-by-paper, page-by-page. Anything novel and implying new scientific possibilities, or anything that had not been adequately followed up, was laid aside in a "select" pile for further review and decision by the best Soviet theoreticians available {11}.

One gets a new physics by discovering shortcomings and flaws in the old one. This intensive Russian search almost immediately uncovered the deep flaws in classical electrodynamics—flaws that have been propagated throughout most other branches of science. Classical electrodynamics (and electrical engineering) is particularly beset by hoary flawed foundations problems, dating from at least 1865. As an example, the equations did and still do assume a material ether, and they have since the beginning! Not a single equation was changed when the

electrodynamics model which *do not even model* the active vacuum and its exchange with every charge, much less a *broken symmetry* in that exchange. So none of those agencies even knows what powers an electrical circuit, 45 years after the proof of broken symmetry was accomplished by Wu and her colleagues in 1957, and the Nobel Prize was then quickly awarded to Lee and Yang *the very same year* (1957) for predicting that broken symmetry revolution. In the modern view a charge is a dipolarity, when its vacuum interaction is considered. As a dipolarity, every charge exhibits the proven asymmetry of opposite charges, in its interaction with the active vacuum. The charge thus extracts and outpours real observable EM energy from the vacuum, freely and continuously, and it exhibits COP = ∞. Those agencies—along with nearly all the rest of the U.S. scientific community—still naively assume that a COP>1.0 EM system (i) is impossible, (ii) is against the laws of nature, and (iii) is a forbidden perpetual motion machine. In nearly half a century since the discovery of the broken symmetry of opposite charges, those agencies have yet to apply that finding to electrical engineering and engineering curricula. Frankly, since they have dragged their collective feet for nearly a half century on extracting EM energy from the vacuum, what is needed is a Stalin to put his iron boot firmly to their posteriors and call their attention to it in language they cannot ignore. *The pollution of the planet for energy purposes and the imminent oil and energy wars that will ravish the earth are directly due (i) to this inexplicable dogma upheld by the scientific community and (ii) to the community's continued falsehoods equating EM COP>1.0 to perpetual motion machines.* The continuing failure to **incorporate** the **active vacuum** exchange—and its broken symmetry—into electrical engineering and classical **electromagetics a**lso poses an unacceptable threat to the continued survival of this nation.

Michelson-Morley experiments in the 1880s destroyed the material ether itself. The classical EM model also assumes a flat local spacetime (falsified by general relativity since 1915) and an inert vacuum (falsified in particle physics for several decades). It still assumes the same EM field in mass-free space as exists in mass—a non sequitur of major impact. Just to make the Maxwell-Heaviside equations easier to solve and avoid numerical methods, the *totally arbitrary* Lorentz symmetrical regauging[6] discards all EM systems far from equilibrium with their active vacuum environment. In fact, there is no "active vacuum environment" in the model, much less a broken symmetry in that exchange. Yet that very broken symmetry is exhibited by every charge and dipole {12}, and thus by every EM circuit and system. That asymmetry is involved in the continuously increasing giant negentropy poured out by every charge in the universe, thereby generating all EM fields and potentials and their energy, eventually reaching even across the universe, *and with perfect macroscopic order and time duration as long as one wishes*. Indeed, all electrodynamics is in total violation of the second law of thermodynamics, to any *macroscopic* size level and time duration desired.[7]

The terribly flawed electrodynamics model was completely overhauled by the Russian scientists *very* quickly, and the work was accomplished by some of the best nonlinear scientists in the world. Already present at the time were many things available to assist in dramatically extending {13} the hoary 1865 Maxwellian electrodynamics {6}. Gauge field theory was available since the turn of the 20th century, and the top theoreticians could readily see that something more than Abelian electrodynamics was firmly required in particle physics. The Yang-Mills {14} theory met that need in 1954 by originating non-Abelian gauge theory.

Situation in the Early 1950s

By 1950, Stalin's scientists had the beginnings of what they dubbed "energetics" {15,16}, which we would dub an "engineerable unified field theory" {17,18,19}. Longitudinal EM waves were primary {20,5} in that more fundamental approach. By the mid 1950s, one can pick out of the open literature {21} increasing incidences of prototype Russian energetics weapon testing around the world. In 1960, Khrushchev {22} spoke obliquely of these "fantastic new weapons" that were "just within the portfolio of the Soviet scientists," so to speak. Commercial aircraft sightings (Figure 2) and the reports of their captains and crews are also a particularly significant source of sightings of Russian energetics weapons tests, as are sightings by ships' crews (Figure 3) and ground observers (Figure 4).

---

[6] Actually a version of this symmetrical regauging was first performed by Ludwig Valentin Lorenz, "Ueberr die Identitat der Schwingungen des Lichts mit den elektrischen Stromen," *Ann. Phys. Chem.*, Vol. 131, 1867, p. 243-263—only two years after Maxwell published his seminal theory in 1865. The paper is also published in English, as Ludwig Valentin Lorenz, "On the identity of the vibrations of light with electrical currents," *Phil. Mag.*, Ser. 4, Vol. 34, 1867, p. 287-301. Unfortunately not much attention was paid to Lorenz's effort, since he independently derived electrodynamics. Decades later, when H. A. Lorentz also published symmetrical regauging of the Maxwell-Heaviside equations, he was given credit for it with no further mention of Lorenz. A proper historical summary of this has been given by J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys.*, Vol. 73, July 2001, p. 663-680. This is an excellent coverage of the history of who actually did what and when, and who got credit for it.

[7] This dramatically extends the size regions (microscopic through colloidal) and the time durations (momentary to two seconds) where the second law is known and proven not to necessarily apply.

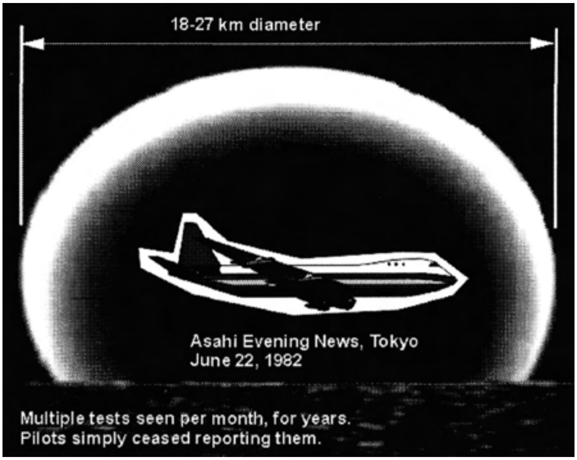
18-27 km diameter
Asahi Evening News, Tokyo
June 22, 1982
Multiple tests seen per month, for years.
Pilots simply ceased reporting them.

Figure 2 Typical shield test, northern route to Japan, seen by passing airliner.

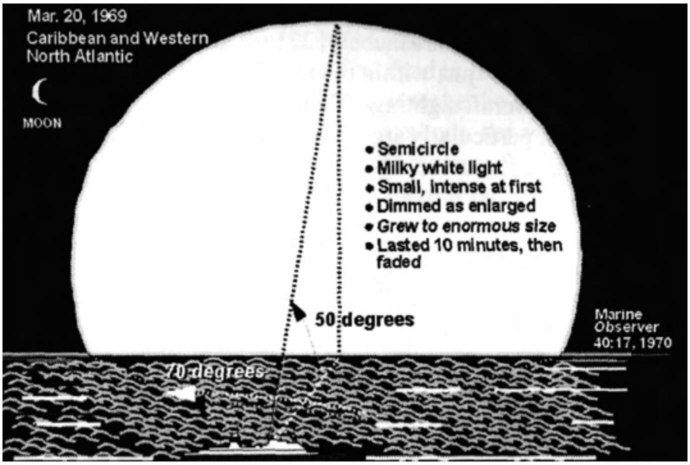
Mar. 20, 1969
Caribbean and Western North Atlantic
MOON
• Semicircle
• Milky white light
• Small, intense at first
• Dimmed as enlarged
• Grew to enormous size
• Lasted 10 minutes, then faded
50 degrees
70 degrees
Marine Observer 40:17, 1970

Figure 3 Giant missile shield test seen by passing ship's crew.

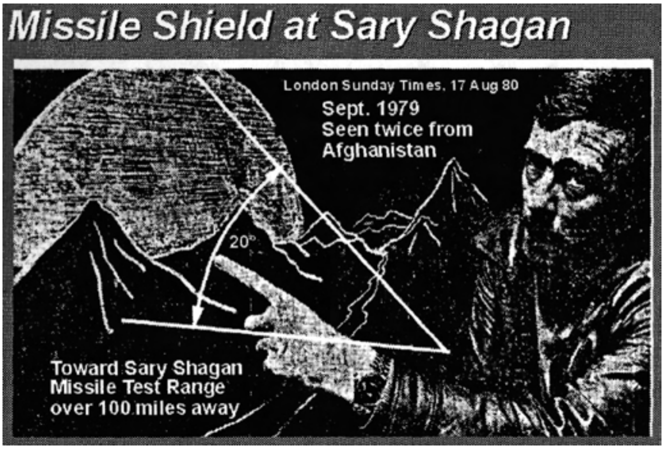
Missile Shield at Sary Shagan
London Sunday Times, 17 Aug 80
Sept. 1979
Seen twice from Afghanistan
20°
Toward Sary Shagan Missile Test Range over 100 miles away

Figure 4 Nick Downie's sighting from the ground of distant Soviet missile shield tests.

1963 Deployment and Immediate Use of Scalar Interferometers

The first huge scalar potential interferometers {23,24} of strategic range and power (Figure 5) were deployed by the Soviets in 1963, and one was used to kill the U.S.S. Thresher nuclear attack submarine (Figure 6), leaving clearly recognizable signatures {25}.

Since 1963, the Russians have had the equivalent of more than seven additional Manhattan Projects (using the Russian 5-year program instead of the 4-year Manhattan Project), back-to-back, in development of energetics weaponry. The energetics weapons have never been given to the regular Russian forces. Instead, all research, production, siting, manning, and employment are by the KGB and still under ruthless KGB control by die-hard communist factions {26}.

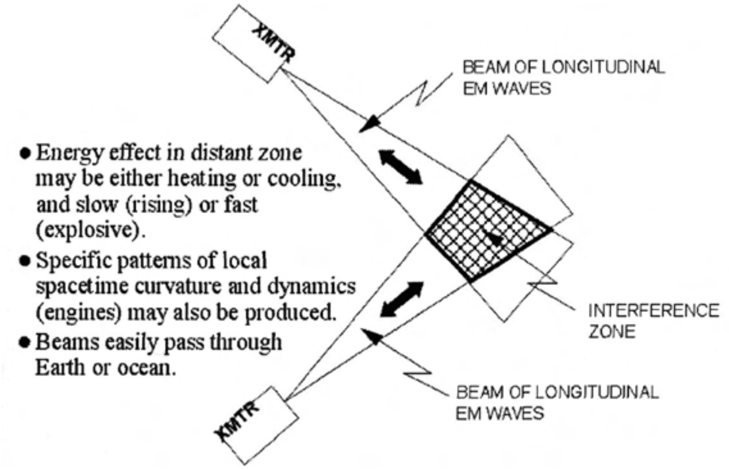
XMTR
BEAM OF LONGITUDINAL EM WAVES
• Energy effect in distant zone may be either heating or cooling, and slow (rising) or fast (explosive).
• Specific patterns of local spacetime curvature and dynamics (engines) may also be produced.
• Beams easily pass through Earth or ocean.
INTERFERENCE ZONE
BEAM OF LONGITUDINAL EM WAVES
XMTR

Figure 5 Scalar (longitudinal EM wave) interferometer.

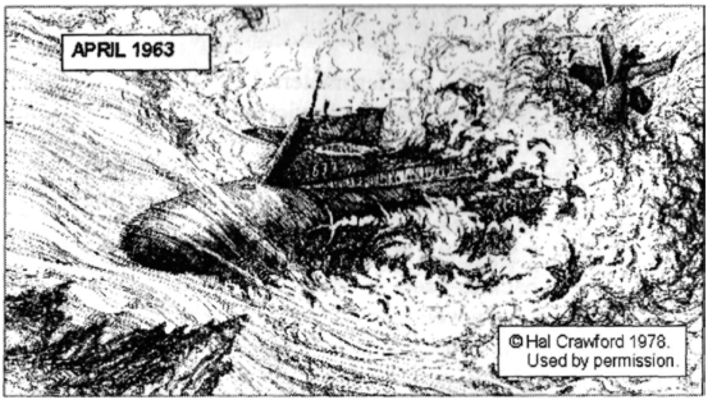

APRIL 1963

© Hal Crawford 1978. Used by permission.

Figure 6 Kill of the U.S.S. Thresher. With her controls jammed, the hull of the doomed submarine implodes when she reaches crush depth

An Example of Russian Technology Astonishing Western Scientists

Some years ago, coordinating with a Western but non-US aerospace firm, I furnished the firm with the names and addresses of some Russian scientists in Moscow, since it was clear from their internet site that the scientists were on the fringes of the KGB energetics weapons program. An excellent European engineer fluent in Russian was sent in, several times. Several unusual things were demonstrated to him, including (i) transmitting enormous energy down a very thin wire, and (ii) cold molding, or turning metal into a liquid at room temperature, pouring it into a mold, and then letting it sit and harden. All without heating.

The engineer also closely questioned the Russian scientists about Russian energetics weaponry, a subject the scientists were most reluctant to discuss. Finally they admitted the energetics weapons existed, and they actually obtained a meeting between the engineer and the KGB General who had directed the Soviet energetics weapon program for more than 20 years. In the meeting, the General coldly admitted the weapons, stated that "Of course you are not going to be allowed to see *those!"* and confirmed that somehow the Russians obtained everything the present author writes, as soon as he writes it {27}.

The engineer then brought the Russian scientists to his aerospace firm in that Western nation, where the scientists demonstrated cold molding to assembled scientists and some U.S. personnel. I forwarded a written explanation to the Western scientists of how the cold molding was performed technically. Later, on calling the engineer again, I discovered that the foreign authorities of that country had suddenly classified everything over there, and the engineer could no longer even talk to me without filing a Foreign Intelligence Contact Report. My own part in all this was completely open-source, pro bono, and I had already written of it openly and publicly.

Softening of Metal at a Distance Was One Effect Used Against the Challenger

At any rate, earlier I had openly advanced (and printed) the same mechanism, since it was the metal softening mechanism used by the KGB to kill the U.S. Challenger spacecraft in 1986. Publicly I had been resoundingly called a lunatic for advancing such a "preposterous" thesis. Nonetheless, metal softening by a distant LWI (longitudinal wave interferometer) was definitely utilized as part of the kill of the Challenger, as a close colleague here in the U.S. proved *experimentally,* detecting and demonstrating that metal-softening Russian signal prior to the kill of the Challenger. He amplified the signal and demonstrated on the bench its nearly instant softening of nails, rendering the nails momentarily as limp as wet noodles (Figure 7). Nails from the same box, not exposed to the signal, remained hard and could be hammered into a board. Not so the nails from the box that were exposed to the amplified Soviet signal! Those nails folded up like a wet noodle when tapped with a hammer, and could not be driven into a board.

Working unpaid with the foreign scientists, we had in fact obtained the "smoking gun" proving that this cold molding capability rigorously existed in the KGB inventory. Further, it had existed so long that the KGB had essentially *declassified* it so that the civilian Russian scientists seeking employment funds from the West could utilize it.

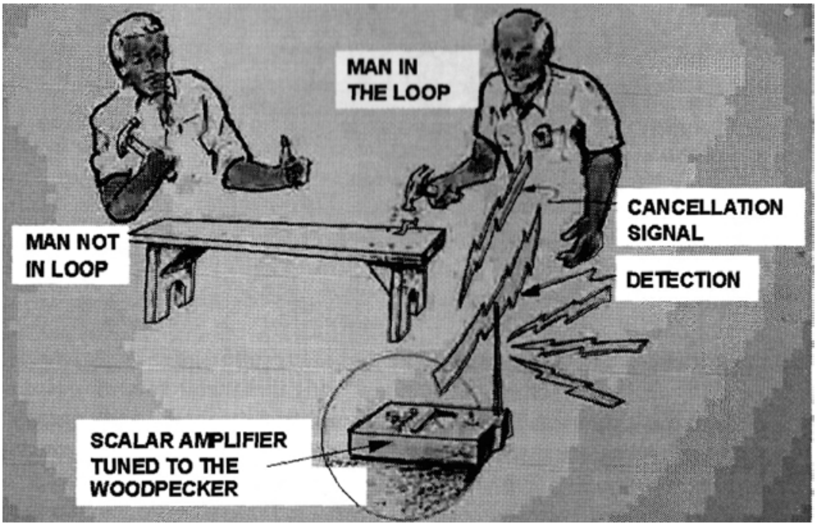

Figure 7 Proof of metal-softening signal: Golden softens nails with amplified Woodpecker signal.

Energetics: Three Branches and a Strategic Deception Plan

The *energetics* weapons use exotic polarizations of EM waves, quantum potentials, and the "infolded" LW electrodynamics (both scalar and longitudinal photons) inside and comprising every "ordinary" transverse EM field and wave. The KGB energetic weapons are categorized (Figure 8) according to what they are targeted against:

(1) against inert matter, fields, and potentials etc., that branch of weapons is called by the same name, *energetics* weapons.

(2) Against living tissue, bodies, cells, biofields, biopotentials, etc., they are known as *bioenergetic* weapons. Disease-induction electrodynamics such as used to irradiate the U.S. Embassy in Moscow are examples of bioenergetic weapons.

(3) Against the mind, thought, memory, behavior, and the mind-body coupling mechanism, they are known as *psychoenergetic* weapons. Particularly in the latter mode, these weapons also use time-polarized (scalar) EM waves and time-polarized (scalar) photons as well {28}.

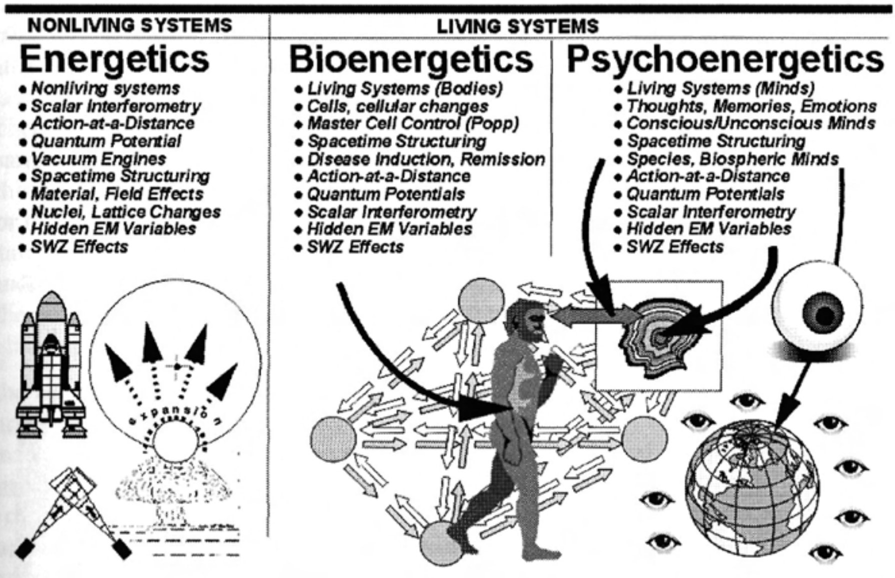
NONLIVING SYSTEMS
LIVING SYSTEMS
Energetics
• Nonliving systems
• Scalar Interferometry
• Action-at-a-Distance
• Quantum Potential
• Vacuum Engines
• Spacetime Structuring
• Material, Field Effects
• Nuclei, Lattice Changes
• Hidden EM Variables
• SWZ Effects
Bioenergetics
• Living Systems (Bodies)
• Cells, cellular changes
• Master Cell Control (Popp)
• Spacetime Structuring
• Disease Induction, Remission
• Action-at-a-Distance
• Quantum Potentials
• Scalar Interferometry
• Hidden EM Variables
• SWZ Effects
Psychoenergetics
• Living Systems (Minds)
• Thoughts, Memories, Emotions
• Conscious/Unconscious Minds
• Spacetime Structuring
• Species, Biospheric Minds
• Action-at-a-Distance
• Quantum Potentials
• Scalar Interferometry
• Hidden EM Variables
• SWZ Effects
expansion

Figure 8 Three branches of Russian energetics weapons.

The KGB also implemented an elaborate and continuing deception plan, as is standard military practice for any great strategic attack plan. The KGB deception plan has been highly successful {29} in keeping the United States convinced that these weapons do not exist, and convinced that the standard strategic nuclear weapons possessed by the West are "leading" and "superior."

Strategic Attacks Scheduled and Why We Were Not Destroyed

Indeed, the West's nuclear weapons are inferior to Russian energetics weapons, particularly to the quantum potential weapons (Figure 9) deployed at the end of 1989 and to the newer "causal system" weapons deployed since the first of 1999, where *functional robot systems* (Figure 10) are made of nonmaterial spacetime curvatures and their dynamics only! In the event of a full strategic Russian energetics strike, an opening strike with quantum potential weapons {30} will dud all nuclear weapons on earth, all nuclear power systems, and all nuclear propulsion systems in about 10 minutes. This takes care of most of the U.S. strategic armament, *within the first 10 minutes.* The publicized strategic weapons of this nation have in my opinion been almost defenseless now for at least a full decade {31,32}, but I also believe that situation may be rapidly changing for the better.

A friendly little foreign nation *does* have such weapons including QP weapons that operate in multiply connected spacetime (MCS). That little nation has been responsible for our continued survival and has continued to deter the planned KGB strategic energetics strikes upon the West. There are indications (though still not certain) that at least one—and possibly two—other friendly Western nation also possesses such weapons.

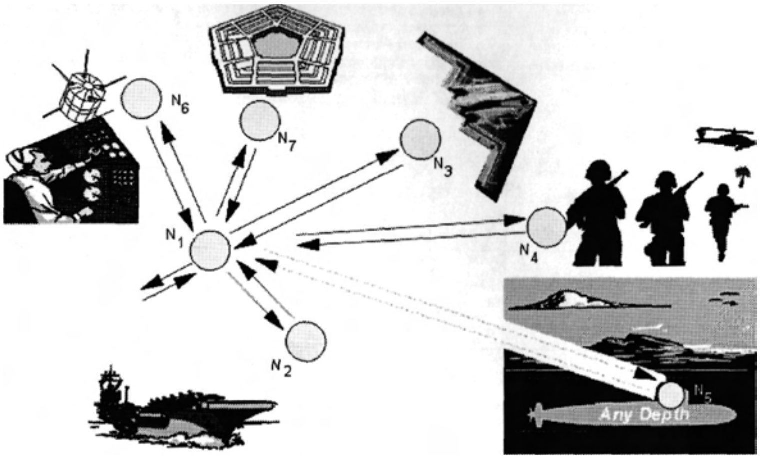
N_6
N_7
N_3
N_1
N_4
N_2
N_5
Any Depth

Figure 9 Quantum potential instantly attacking widely separated targets.

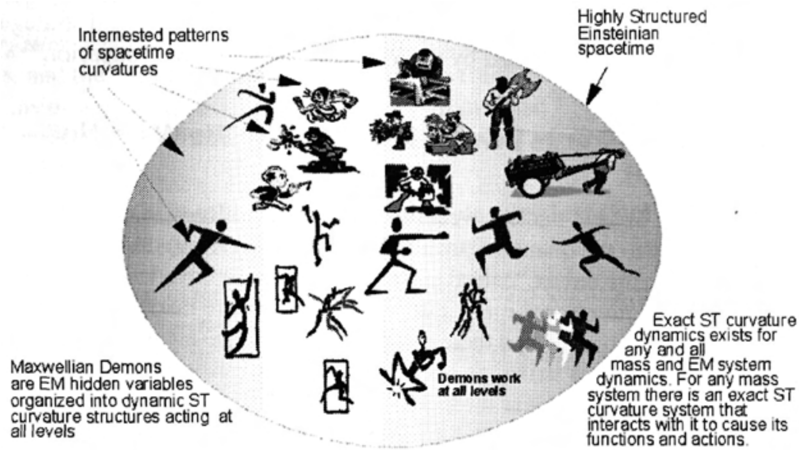
Internested patterns of spacetime curvatures
Highly Structured Einsteinian spacetime
Exact ST curvature dynamics exists for any and all mass and EM system dynamics. For any mass system there is an exact ST curvature system that interacts with it to cause its functions and actions.
Maxwellian Demons are EM hidden variables organized into dynamic ST curvature structures acting at all levels
Demons work at all levels

Figure 10 Causal system robots (CSRs). These are sets of ST curvature engines, with functions organized into a single complex weapon system with communication, propagation, weapons effects, and command and control functions incorporated. When the CSR interacts with mass, the mass performs those system functions.

Twice in 1997 alone *we were within days of the scheduled total strategic destruction of the U.S. by die-hard KGB energetic weapons,* with all-out strategic attacks actually scheduled in each case. The first attack was scheduled in latter February but was quickly countered. The attack was immediately rescheduled for May 1. The U.S. intelligence and scientific communities were apparently unaware of what was happening, and still do not believe it. However, as this author previously informed appropriate government agencies a few days prior to the attack scheduled for May 1, 1997, U.S. intelligence agencies certainly intercepted my urgent personal fax directly to the prime minister of the friendly little nation, urgently requesting that the "sane" command and control of the countering MCS/QP weapons be changed to "insane" control and spelling out the compelling reason why it was necessary.

In the insane weapons control mode, the computer periodically (say, every two hours) initiates launch sequence for the weapon, automatically. If any operating personnel remain alive on site, there is a designated short period in the sequence where the onsite personnel can countermand the order and abort the launch. If the order is not countermanded, the launch is not aborted and the "doomsday" weapon automatically fires. In that case, KGB military power and most of Russia would disappear from the face of the earth.

This was necessary to nullify the impending strikes of MindSnapper negative energy EMP weapons {33} on those critical friendly QP weapon sites to instantly kill all personnel on site. Strike by the MindSnapper instantly kills every living thing in the area, and a protracted contamination remains whereby the struck areas emit longitudinal EM waves. If one attempts to insert new crews to resume operation of the weapons, the crewmen sicken and then die in the entry attempt. First, increasingly severe interference between mind and body control occurs. Then the individuals waver, lose consciousness, go into seizures, and their minds gradually de-couple from their bodies and they die. Fortunately, alter my fax message the C3 system was changed to the insane mode, and the Russians were notified, just in the nick of time {34}. To prevent their own destruction by the certain insane counterstrike, the KGB aborted the looming strategic strikes.

Ironically, at about the very time that the KGB strategic armada was standing down, Secretary of Defense Cohen {35} was making the following statement in Georgia:

> *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts."*

Almost certainly Secretary Cohen had no knowledge that, so to speak, a great Sword of Damocles hanging over his head was being removed by a friendly little foreign nation at that very moment, as he uttered those words apparently referring to the older longitudinal EM wave interferometer weapons. I suspect that Dr. Cohen is to this day still unaware of the incidents that were ongoing at the time of his poignant statement.

Attempts to Alert the Nation

We alerted the President and the National Security Council as early as 1984 (Figure 54). In 1998 we also furnished selected Senators, Congresspersons, and certain high-level U.S. government agencies with color copies of a 200-page briefing and expose {36} of—among other things—these weapons, their testing, and their clandestine use against the U.S. since 1963. Because of the nearly 50 years of demonstrated induction of diseases {37, 38} (Figure 11) in personnel in the U.S. Embassy in Moscow (e.g., responsible for the deaths of three U.S. Ambassadors), we also included part of the briefing on the Priore work {39} (Figure 12) showing the remarkable cellular changes (time-reversal of diseased states back to a previous earlier healthy condition) that can be and had been experimentally induced by energetics under rigorous scientific controls.

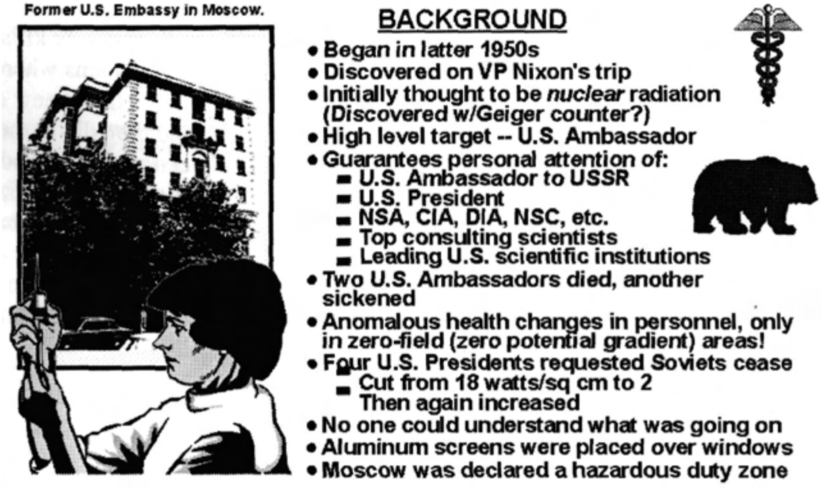
Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero potential gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2 Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Figure 11 Microwave radiation of U.S. Embassy in Moscow.

We proposed to the DoD a very rapid, high priority development program to develop portable suitcase-sized computer-controlled treatment devices (Figure 13) capable of saving a majority of mass casualties—e.g., from a professional large anthrax attack on a major population center.

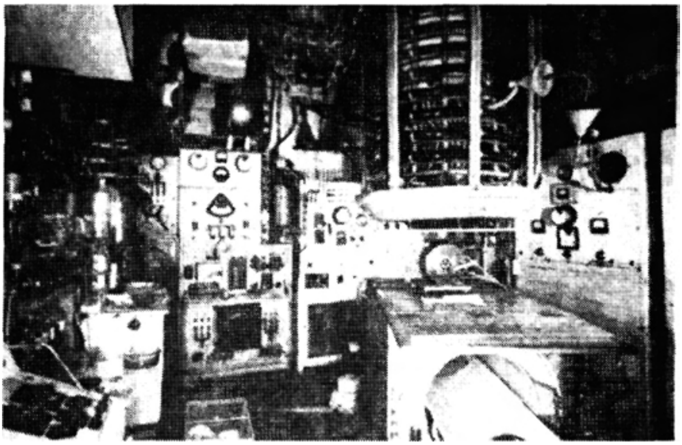

Figure 12 Priore[1] laboratory and device for treating small animals.

Two blankets have antennas in them.
Uses an advanced but simple concept.
Eliminates plasma tube mixing etc.
Computer-controlled complex signal structure.
Irradiation for 2 minutes.
Three such treatments one week apart.
Unit can be made self-powering on vacuum energy.

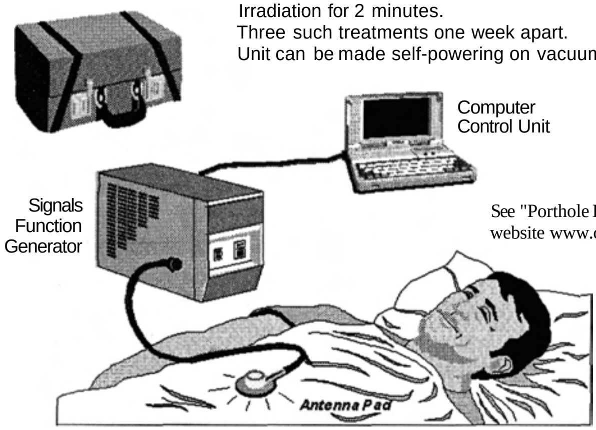
Equipment Case
Computer Control Unit
Signals Function Generator
Antenna Pad

See "Porthole Briefing" on website www.cheniere.org.

Figure 13 Portable unit for rapid treatment of mass casualties.

The Responses

The responses—or more exactly the near total lack thereof—by various U.S. government agencies were enlightening. E.g., the National Institutes of Health (NIH) never let the package out of its Policy (i.e., *political spin control)* section. It did not check a single scientific reference on a proven method documented in the hard French scientific literature for reversing terminal cancer, infectious diseases, etc. Not a single scientist called me to discuss such a proven, revolutionary effect. Instead, the NIH Policy section simply shipped the package over to the DOD as "their problem, not ours." That was because the package also explained how the Gulf War Disease was induced, including how semen can cause infection of the wife, and how the structured body biopotential of the sickened veteran can induce similar structuring of the biopotentials of his children by proximity, thus inducing various aspects of the disease.

Effective Survival Treatment of Mass Casualties Could Be Developed

Meanwhile, we had uncovered a way to dramatically speed up the Priore healing process, and use only normal transverse EM radiations for ready adaptation of existing equipment, while still forcing the same time-reversal of diseased and damaged cells back to normal. Just much faster. This meant that very small, suitcase-sized treatment units (Figure 13) could now be developed in a reasonable time—e.g., in two to three years, given a high priority, intense effort, adequate funding, and a carefully chosen scientific team.

So, we proposed a crash development by DOD of a small, suitcase-size portable device for quick, mass treatment of mass U.S. casualties following terrorist attacks on our cities using weapons of mass destruction. Eventually an innocuous reply (forced by my congressman) simply stated "it was not supported by the literature." In the package, of course, I had *specifically cited* the numerous scientific papers in the French scientific literature reporting and proving the experimental results. I had also worked 14 years to decipher the mechanism actually performing the revolutionary healing accomplished in rigorous animal testing by the Priore team. Eleven years after suppression of the Priore effort, the University of Bordeaux, which had been pressured into rejecting Priore's doctoral thesis, did approve and publish a doctoral thesis by Perisse on the work {40}.

The National Science Foundation (NSF) did not reply at all. Yet at the time the official strategic threat to the United States listed WMD attack on our cities and great population centers as the greatest threat of all (Figure 14). Today, as everyone knows we are in a war after the terrorist attack on the twin towers in New York and on the Pentagon on September 11, 2001. The U.S. government clearly recognizes that it is not a matter of if additional attacks on the U.S. cities and populace will occur, but *when.* The terrorists are also known to have already infiltrated weapons of mass destruction such as biological warfare weapons (anthrax, smallpox, bubonic plague, etc.) Nonetheless, apparently the NSF simply did not believe a novel but *rigorously demonstrated-in-the-1960s* unorthodox process that could potentially save most of those millions of Americans who otherwise will surely die—and who *are* going to die in the first major WMD strike on our population centers.

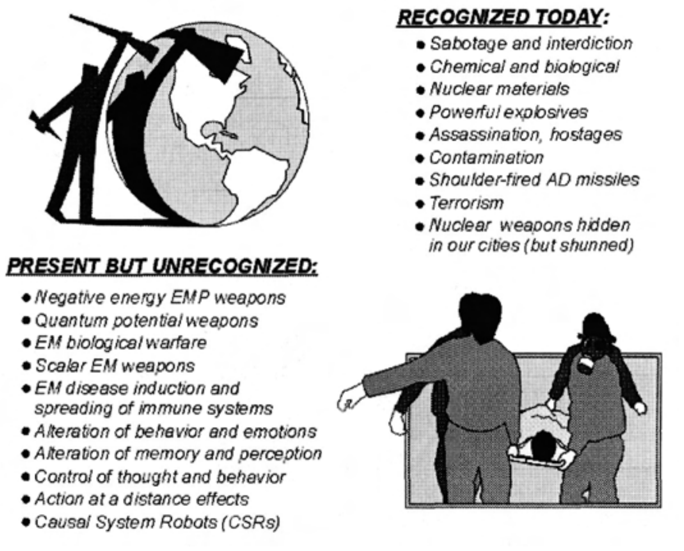
RECOGNIZED TODAY:
• Sabotage and interdiction
• Chemical and biological
• Nuclear materials
• Powerful explosives
• Assassination, hostages
• Contamination
• Shoulder-fired AD missiles
• Terrorism
• Nuclear weapons hidden in our cities (but shunned)
PRESENT BUT UNRECOGNIZED:
• Negative energy EMP weapons
• Quantum potential weapons
• EM biological warfare
• Scalar EM weapons
• EM disease induction and spreading of immune systems
• Alteration of behavior and emotions
• Alteration of memory and perception
• Control of thought and behavior
• Action at a distance effects
• Causal System Robots (CSRs)

Figure 14 The mass destruction weapons threat to our populace.

The entire scientific method is based on assigning priority to experimental truth rather than dogmatic belief. If experiments refute the theory, the theory must be changed and the experimental results accepted. Else one is not following scientific method. Apparently the National Science Foundation does not follow scientific method, nor does it seem interested in doing so. Neither do the National Institutes of Health. The decisive results of some 2,000 rigorous animal experiments reported in the refereed scientific literature in leading French journals, by multiple established French scientists, were simply ignored because the results are contrary to present dogma. Meanwhile, four years have been lost. During that four years, the portable units could have already been developed, mass produced, and made available to treat and save millions of coming U.S. civilian casualties. Apparently the lives of those Americans are not as important as maintaining the present medical dogma. When those millions do die—and they will—we shall most certainly remind those agencies of the caliber of their dedication to scientific method and of their service to their fellow Americans. In my personal view, ignoring a chance to save so many Americans is treason of a very special kind.

The Official Silence Was Deafening

Not a single scientist in any of those and other organizations seems to have taken the package seriously. Not one called me to discuss it, or apparently studied the references, etc. No one asked for a personal briefing. No one was interested in the actual photo of the strike of a Russian EM missile from a scalar interferometer, offset from a U.S. night shuttle launch (Figure 15) from Cape Canaveral in latter November 1985, or the photo of the actual registration point ball of light high in the air (Figure 16) over that same shuttle launch {41}. The same weapon then killed the Arrow DC-8 at Gander, Newfoundland about two weeks later (Figure 17). An eyewitness (or more than one) actually observed the streak of light from the sky that struck the Arrow DC-8 in the right fuselage ahead of the engines {42,43}, after the thrust of the engines were reduced during take-off by use of a negative energy (cooling) interferometer. A similar ball of light associated with the kill of a Titan missile fired from Vandenberg AFB in early 1986 {44} and printed in *Aviation Week & Space Technology* magazine, was also apparently of no concern.

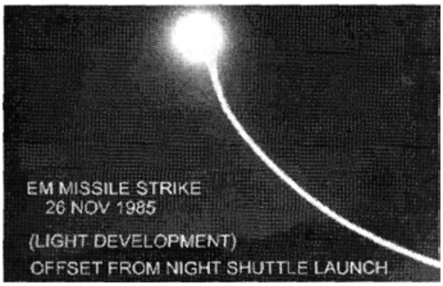

PHOTO BY BOB GLADWIN

Scalar interferometer operator fires offset EM missile from Russia.

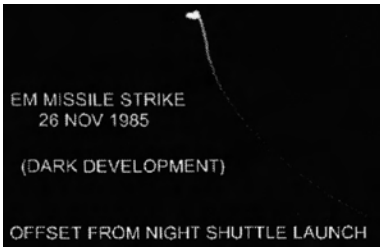

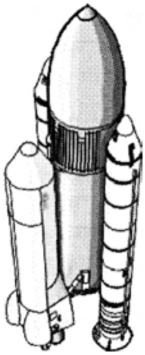

Figure 15 Soviet EM missile strike offset from U.S. shuttle launch at Cape Canaveral on 26 Nov. 1985. This is the smoking gun.

These remarks are intended to show the lack of understanding that has existed in our governmental, scientific, and intelligence communities of what energetics is and of the great energetics superweapons that have been developed. With the present rapidly growing interest in longitudinal EM waves, however, we suspect that such lack of understanding is now no longer the case. We certainly hope that is true, and we are

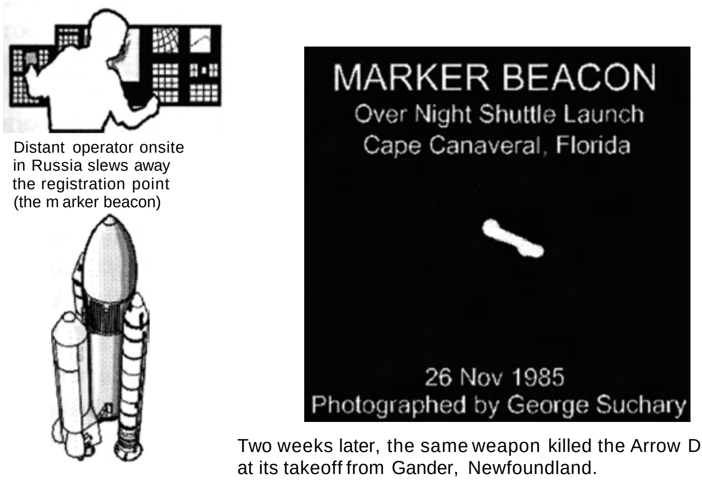
MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida
26 Nov 1985
Photographed by George Suchary

Distant operator onsite in Russia slews away the registration point (the marker beacon)

Two weeks later, the same weapon killed the Arrow DC-8 at its takeoff from Gander, Newfoundland.

26 Nov1985

Figure 16 Marker beacon above U.S. shuttle launch at Cape Canaveral, 26 Nov. 1985. Russian crews were using shuttle launches as practice ICBM launch targets.

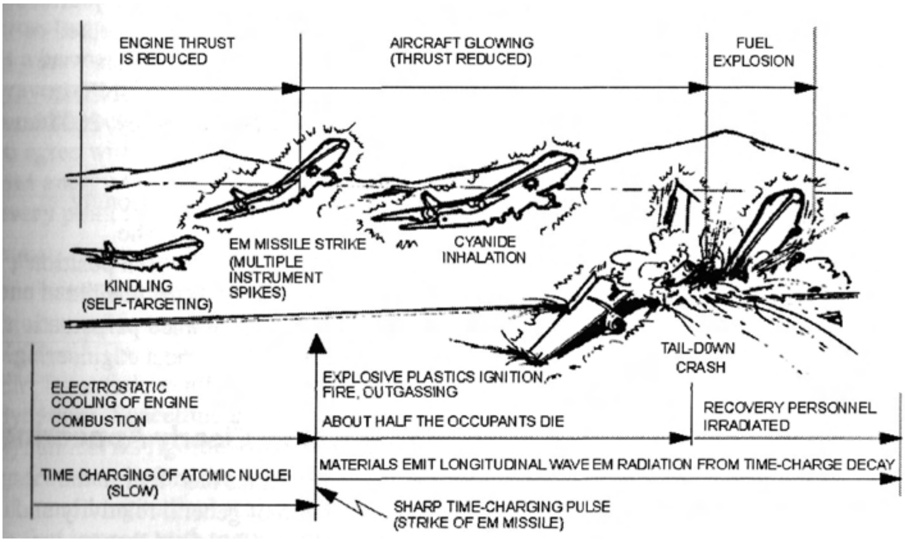
ENGINE THRUST IS REDUCED
AIRCRAFT GLOWING (THRUST REDUCED)
FUEL EXPLOSION
EM MISSILE STRIKE (MULTIPLE INSTRUMENT SPIKES)
CYANIDE INHALATION
KINDLING (SELF-TARGETING)
TAIL-DOWN CRASH
ELECTROSTATIC COOLING OF ENGINE COMBUSTION
EXPLOSIVE PLASTICS IGNITION, FIRE, OUTGASSING
ABOUT HALF THE OCCUPANTS DIE
RECOVERY PERSONNEL IRRADIATED
TIME CHARGING OF ATOMIC NUCLEI (SLOW)
MATERIALS EMIT LONGITUDINAL WAVE EM RADIATION FROM TIME-CHARGE DECAY
SHARP TIME-CHARGING PULSE (STRIKE OF EM MISSILE)

Figure 17 Kill of the Arrow DC-8 at Gander, Newfoundland Dec. 12,1985.

much encouraged that at least one—and possibly two—friendly Western nations now possess quantum potential weapons.

The Long Dry Spell in Correcting Electrodynamics May Be Slowly Ending

In over 20 years of unrelenting effort, until 1999 we were unable to get any significant Western scientists of real "horsepower" (e.g., in non-Abelian electrodynamics, foundations errors in electrodynamics, advanced gauge field theory, general relativity, etc.) to meticulously examine conventional U(1) electrodynamics and its glaring errors, and to look at weapons implications of higher group symmetry electrodynamics.

Beginning about three years ago, however, a noted scientist, Dr. Myron Evans {45,2}, along with other theoretical scientists of the Alpha Institute's Institute of Advanced Study (AIAS), have been extending electrodynamics by developing the 0(3) electrodynamics. They also have deeply examined the Whittaker work of 1903 and 1904 on which the early part of the Russian energetics weapons are based {24}.

The AIAS analyses, and subsequent papers by Dr. Evans and Alex Labounsky as well as by the entire AIAS group, have clearly established the primary importance of time-like precursors of EM fields, operating in massfree spacetime where fields E and B cannot exist. These time-like precursors *create* all the rest of the conventional electrodynamic entities {46}. This is a dramatic extension to electrodynamics and physics across the board. More than 100 papers have been produced to date, with more than thirty published in journals such as *Foundations of Physics,* and many of the others are either accepted or still in the referee process with various journals. Some 60 of the papers were published as a single special edition of the *Journal of New Energy* (JNE) in latter 1999, and a second edition of *Modern Nonlinear Optics,* Wiley, 2001 has been published with many important papers.

However, the problem is that the scientific community, which is strongly committed to "business as usual", has funded none of this work. The conventional lead agencies of the community are still in the "ostrich position", with their heads buried very firmly in the sand with respect to the terrible damage they continue to enforce on this nation by their continued perpetuation of an old, long-archaic, inadequate electrodynamics and electrical engineering.

Engineerable Unified Field Theory Is Now Clearly Appearing

The rapidly emerging AIAS extension of electrodynamics has "met in the middle" with Sachs' {4} revolutionary generalization of general relativity and electrodynamics. For the first time, an engineerable unified field theory—employing 0(3) electrodynamic means—is now being produced. The AIAS work in 0(3) electrodynamics was shown to be an important subset of Sachs' approach. It also now provides the full, rigorous theoretical foundation and model for the Priore mechanisms and results.

We also comment most strongly that the important SU(2)xSU(2) electrodynamics as developed by Barrett (47) leads to essentially the same conclusions, since Barrett's electrodynamics is homomorphic with 0(3) electrodynamics. Barrett, of course, is one of the pioneers of ultrawideband radar—another scientific area that was fiercely opposed and suppressed by the orthodox U.S. establishment for some time. Ironically, today those very scientists who were so dogmatic in their vociferous condemnation of UWB radar, have assumed the mantle of "experts in the UWB field", sometimes even taking credit for having "advanced it" in the first place.

Examples of Some Major Non Sequiturs in Conventional Electrodynamics

In every textbook and millions of technical papers in the West, electrodynamicists routinely calculate the "magnitude of the scalar potential". Actually, not a single one—including Maxwell—calculated or calculates the *magnitude of the potential itself.* Instead, they universally calculate the *reaction cross section* of the potential at a point occupied by an assumed intercepting/collecting unit point *static* charge {48}. At best that is a representation of the *intensity* of the field at each point. The reaction cross section of the potential with a fixed unit point static charge is indeed a scalar value. The "scalar" potential itself, on the other hand, *is not* its own reaction cross-section at each point of itself, and it is not a scalar entity at all! Instead, it is a harmonic bundle of bidirectional EM longitudinal phase conjugate wavepairs, as shown by E. T. Whittaker in 1903 {49}. It is and always has been a multi-vectorial, multiwave entity. By slightly reinterpreting Whittaker's work to agree with quantum field theory, the scalar potential is also a circulation of EM energy from the time domain to 3-space and back to the time domain, at every point in space that is occupied by the potential.

Further, the "field" notion is utilized in two self-contradictory manners. On the one hand, it is assumed in the massless reaches of space, and on the other hand it is assumed in mass, *which is the only case in which a force field is properly defined.*[8] The so-called "electromagnetic field" existing in empty space is not an EM force field at all, but a curvature or curvatures of spacetime with assorted ilynamics. Spacetime is after all active, and being active means curvatures and ilynamics. As rigorously shown by the AIAS theoreticians, what exists in *spacetime*—before the force-fields appear on reacting charged mass—are time-like and longitudinal scalar potentials and fluxes. These "pre-EM field" entities in fact interact with mass to produce all the force fields (the electromagnetic fields). Mass is a *component* of force; simply by F == *d*p/dt, which is expanded into F = d(mv)/dt, which is ma + v*(dm/dt).* Note that both terms have mass as a component. Hence there is no force or forcefield when mass is absent. Present mechanics is seriously wrong (and has been wrong for hundreds of years) in assuming a separate massless force acting upon a separate mass.

---

[8] Since Maxwell and the original pioneers believed in a material ether filling all space, to them there was not a single point anywhere in the universe that was devoid of mass. Hence to them there was only one kind of field: the kind that exists in mass.

Jackson, one of the ablest classical electrodynamicists of our day, adroitly states how electrodynamicists skirt that issue. Quoting, p. 249 of his *Classical Electrodynamics,* second edition: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

Our comment is that one does not get an observable until after the interaction of a cause with a previous effect, to provide a change (a new effect). The 4-cause itself—in this case the field that actually exists in the vacuum—is never observed nor is it observable. Its *effect* (force with an affected observable mass as a component of it) is observed. The field in space and the field in matter thus do not even have the same fundamental units, but differ by a factor of m. In electrodynamics the m is avoided by "defining" the field as force per unit coulomb of "charge" (without saying or accounting "charged mass").

Wide Confusion of Cause and Effect

Present electrodynamics (and even the hoary old mechanics, as pointed out) has thoroughly confused the causal (4-space) side of the observation process with the effects (3-space) side—observation yields *3-spatial* entities, not 4-spatial. Indeed, as is well known, *time is not observable, even in theory.* That is really because observation is a *d/dt* process applied by the 3-space intermediary (e.g., mass) being acted upon by the causal (4-space) entity in spacetime, to produce a change in that intermediary (3-spatial) that results (is observed). *A priori,* the output of the observation process does not even exist in time, since its fundamental units are LLL and not LLLT. The observation process is thus *d/dt(LLLT)* => LLL.

Further, electrodynamics omits the fact that, not only does the causal entity act upon the intermediary to produce the *effect* (changes to the intermediary), but *those changes (effects) also act back (in reverse) through the intermediary to produce changes in the causal entity itself.* This mutual interaction between cause and effect (but never stressed in exactly that manner) is well known in GR, where the curvature of spacetime acts upon mass-energy to change it, and the resulting change in mass-energy also acts back upon spacetime to curve it. The omission of the latter "back reaction" in EM theory by Maxwell resulted in omitting half the energy, half the wave, and half the causal entity in spacetime, as well as half the effect. The missing half of the effect appears as the Newtonian third law reaction, which in electrodynamics is just mystically "assumed" to be an effect without a cause.

The vast importance of restoring this missing "back action of effect upon cause" is because it can be deliberately utilized to directly engineer the causal (4-space) state itself, producing nonmaterial "engines" and unobservable functioning robot systems comprised entirely of curvatures of spacetime and their dynamics. Such "causal system robots" (CSRs) can be designed and engineered to then perform almost any desired action upon normal mass systems they encounter. The robots also can easily penetrate vast depths of mass, which is mostly empty space filled with potentials, fields and waves—all of which are just bundles of longitudinal EM waves and their dynamics. "Mass" is actually a vast superhighway through which such CSRs can easily move at the speed of light {50}.

CSRs: Engineering the Unobserved Causal Side To Create Functional Systems

Understanding early on the mutual interaction between cause and effect that existed in their energetics, the Russians were able to engineer this totally new kind of system: the *causal system robot* or CSR (Figure 10). With sufficient theoretical development, one can "work backwards" by adroitly re-engineering the effect to obtain a desired causal system (spacetime curvature set along with its impressed dynamics; (Figure 18) corresponding to some physical system and its changes one desires to engender. One thereby directly creates a deterministic set of spacetime curvatures and impressed dynamics, which we call an "engine" (Figure 18). This *ST curvature engine* system can then function as a completely separate system {51} (Figure 10). One can communicate with it via longitudinal EM waves, control it, etc. By building in scalar interferometry functions, the **CSR** can be given "weapons" capabilities, etc.

- All levels of energy structures and all levels of time structures mold spacetime geometry.
- The "pattern" is called a *template* for a vacuum engine (spacetime curvature engine).
- Spatial and temporal structurings of the engine act upon mass at all levels.
- This produces a template of forces, for precise translations and stresses within the mass at all levels. The forces exist in both time and 3-space.

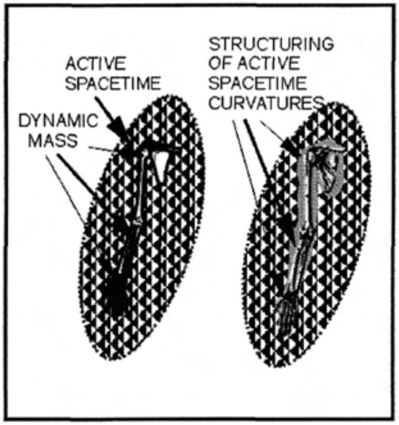

Figure 18 The principle of general relativity extended to include the concept of the vacuum engine.

In theory such a CSR can be assembled to perform any desired function or complex of functions upon matter, fields, potentials, etc. that is desired. In short, one can build real, functioning "infolded" (subspace) robot weapon systems using this "backward engineering of spacetime itself process. The huge underground facilities (hundreds) that the Russians continued to build after the collapse of the Soviet economy and the fall of the Soviet Union, apparently contained the full CSR development and control facilities as well as the operational systems employed for this totally new kind of weapon system.

A great engineering of the future lies in engineering CSRs, which already constitute a great leap forward past the material nanobot technology being pursued in the West.

Making the First CSR is Extremely Difficult; Making a Million Copies Is Easy

The development of a specific CSR with a specific set of functions is akin to developing a highly complex software system of major proportions (several tens of millions of lines of code). The job is extremely difficult, and tortuous "debugging" is required for a lengthy period. Once a single CSR is completed and debugged, however, the production of thousands or even many millions of copies is then a trivial task.

The CSR may be visualized as a sort of vast complex of longitudinal EM waves (recall, we are speaking of ST curvatures, not the conventional EM waves thought to operate on a flat spacetime. *All* EM waves operate in a curved spacetime *a priori*, as clearly shown by Sachs). It turns out that all matter is mostly empty space, filled with potentials and fields. Space itself, considering its energetics, may be taken as a giant scalar potential. All those potentials and fields and waves are in fact nothing but sets of longitudinal EM waves and dynamics. In short, they are vast superhighways for the travel of CSRs and the propagation of longitudinal EM waves.

A CSR can be embedded within any kind of EM potential, wave, or field. So to replicate one, one embeds it in a magnetic signal and records that signal on a diskette or other media. Then one makes millions of copies of the diskette, for perhaps 10 cents or less each. Or do it en masse, by first completing a CD-ROM with a hundred or two hundred CSRs embedded in its signals (Figure 19). Then simply reproduce the CD-ROM. Every reproduction has another hundred or two exact duplicates of those CSRs, made for almost no cost at all. With communications built into the CSRs, the command and control facilities take charge of them, give them individual numbers or designations, and a robot armada emerges with most unusual capabilities indeed.

The puzzling underground facilities in Russia (which our fellows apparently never deciphered and which remained a mystery) apparently contained the scientific teams, the research and development facilities, the production facilities, and the command and control facilities for the new vast CSR weapon complex. It was to see its first great employment at the end of 1999.

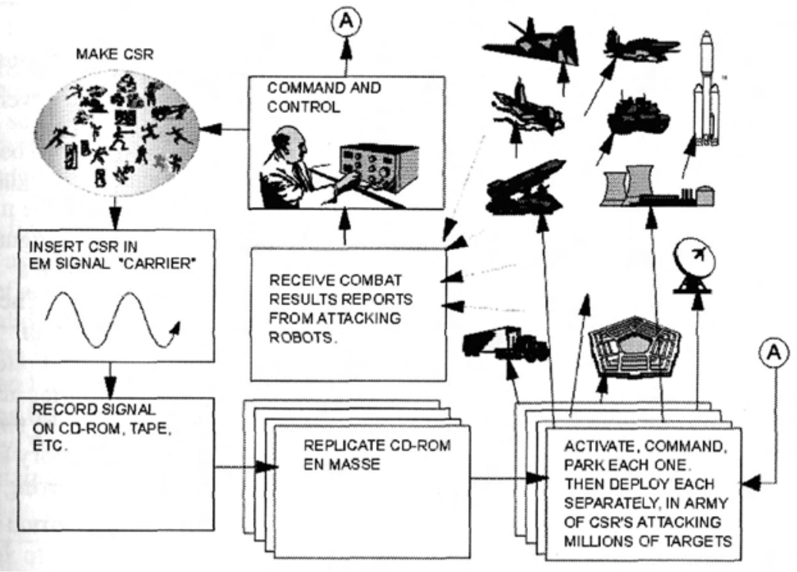
MAKE CSR
COMMAND AND CONTROL
A
INSERT CSR IN EM SIGNAL "CARRIER"
RECEIVE COMBAT RESULTS REPORTS FROM ATTACKING ROBOTS.
RECORD SIGNAL ON CD-ROM, TAPE, ETC.
REPLICATE CD-ROM EN MASSE
ACTIVATE, COMMAND, PARK EACH ONE. THEN DEPLOY EACH SEPARATELY, IN ARMY OF CSR'S ATTACKING MILLIONS OF TARGETS
A

Figure 19 **Mass reproduction of a CSR is simple and inexpensive. The CSR resides in, or propagates inside, any EM signal, potential, or field.**

In 1999 The KGB Deployed Massive Numbers of CSRs for a Strategic Attack

In latter 1999 the KGB began massively deploying these CSRs of various types throughout the Western world, particularly including the United States. They simply sent the CSRs back "inside" the potentials and fields and waves being emitted from radio stations, TV stations, power lines, satellites, communications systems, inside matter, etc. These CSRs were deployed and infiltrated in major command posts and installations throughout the U.S. and its allies. True to form, the KGB even "stimulated" some major computer systems in critical U.S. government facilities by having the robots "surface" some effects on the Computers, right through the firewall programs (whose EM signals and operations indeed provided nice superhighways for the robots). Our fellows **detected** the interference as a "sophisticated computer hacking attack" (puzzlingly from the Russian Academy of Science), and then the robots "submerged" again after deliberately performing actions to produce obvious changes and guarantee the U.S. detection of those changes. The purpose of this "stimulus and probe" was to see if the U.S. was aware of CSRs and CSR **technology.** We clearly showed we were not, and we still are not aware of them. Fur**ther,** the entire government apparatus—including all our vaunted national laboratories—knows absolutely nothing at all about CSRs, and none of the laboratories believe it. That is the same posture they've been in with respect to energetics weapons for a half-century. Apparently they will be in that same posture for another half century, left to their own devices.

However, I was not allowed to remain a casual onlooker for this new CSR armada! The KGB also planned to use me as a stimulus to ping the government yet another way, and then assess the government's response.

Here in Huntsville, suddenly I was personally struck by such a CSR right from my computer monitor, with the attack riding through the light from the monitor to my eyes and retinas and right into the brain and nervous system (Figure 20). This was the type of CSR that can lurk in any potential or field inside a computer, and then strike and kill the operator looking at the screen. The CSR took over control of the machine, split the screen so that only one-inch horizontal strips were luminous—one at the top and the other at the bottom of the screen with the middle dark. These two separated horizontal strips allowed scalar interferometry via the inner LWs comprising the light from the two strips. The LWI signals crossed in my vision center, where suddenly ordinary EM signals and waves were produced by the scalar interferometry occurring there.

Stimulating apparently the only analyst capable of recognizing a CSR attack and connecting it with the scheduled giant strategic CSR attack.

KGB General commanding the energetics weapon previously confirmed that what this analyst writes on his computer is received as soon as he writes it. Apparently accomplished by Tempest techniques or directly by energetics.

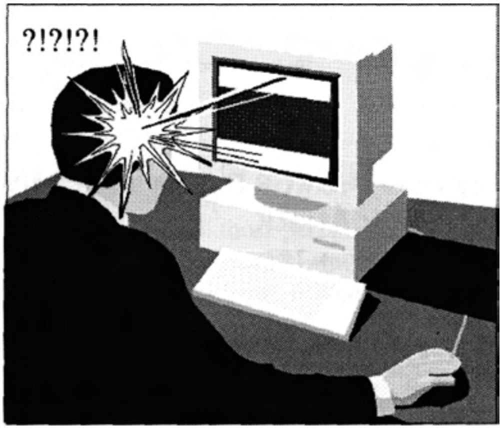

Figure 20 Present author struck by CSR in latter 1999 as a deliberate nonfatal stimulus.

I saw rapid blinding flashes like a hundred flashbulbs exploding at once right in my face. However, the signals representing those "flashes" were being created right there in my vision center. An external camera would have recorded only the weirdly separated screen display. It *would not* have shown the brilliant flashes of light I was seeing, unless it was set up stereoscopically. The strike was deliberately sustained for only about 10 seconds so as to be severe but nonlethal; about 30 seconds would be totally lethal. Violent fibrillation of the heart also resulted explosively, and I was instantly hypnogogic and seriously weakened by sudden hypoxia from near total lack of blood pumping (because of the fibrillation). The purpose was apparently to strike the only analyst in the U.S. who would recognize what the nature of the attack was, stimulate him, watch him try to notify the government, and then observe the government's reaction. It look four days to recover from the strike. But I knew immediately what had made the strike, and that I had been hit with a CSR weaponized by having a built-in scalar interferometry capability. It was also a very sophisticated CSR, as it showed by assuming complete "intelligent" control of my computer's functions.

When I finally recovered, I did notify the government and warn them of CSRs, exactly as anticipated by the KGB. As might be expected, the government's total disbelief and lack of reaction clearly showed they knew nothing of CSRs. They simply did not believe such weapons existed or even could exist. Sadly, once again one faced the "not invented here" syndrome. General relativity and its provision for spacetime curvatures and their associated dynamics (engines) operating back upon mass are apparently just supposed to remain a theoretical curiosity. Why, obviously no scientist in his right mind would think of weaponizing such!

A more correct answer is, *"Like hell they wouldn't!"* *American* scientists would not think of it, but *Russian* scientists with their national chess-playing psychology will immediately seize upon any opportunity—even many moves ahead in the game—and develop it and utilize it. They understand and use the boa constrictor. We only understand and use the immediate rattlesnake.

How the Asymmetrically Deployed Strategic Attack Was Stopped in Its Tracks

However, we also alerted the "little nation" that has advanced EM weapons technology, which—since early 1997—has been saving the U.S. by forcing the abortion of impending KGB energetics weapons attacks.

When I warned them in 1997 about the MindSnapper weapons and mind-control weapons used against Captains Button and Svoboda, unknown to me at the time they later reacted and suddenly destroyed those facilities in Russia (apparently with violent scalar interferometry strikes). So *that* mind-control type of attack being readied for 1999 was thwarted for a few years by these strikes sometime prior to mid-1998.

But to return to the CSRs. In the latter quarter of 1999, the Russians were developing and deploying throughout the West numerous kinds of CSRs. So the period at the end of 1999/first few days of 2000 attack was already scheduled for a massively programmed CSR attack. It was designed to cripple the entire **nation,** and perhaps even collapse it economically, yet all would be blamed on Y2K! Our scientific ignorance guaranteed that we would blame it on something **mundane that** our scientists could understand: Y2K problems in all the old software programs running so much of our automated systems, communications, and centralized functioning, accounting, banking, etc. In short, the KGB planned to give us a bicentennial gift: a Y2K nightmare far beyond our wildest fears.

After my recovery from the CSR strike in the latter quarter of 1999, I again urgently alerted the friendly nation of the development of CSRs and their widespread deployment—and informed them of the personal attack on me and specifically why it was made nonfatal. Of course I also alerted our own government. We were scheduled for a *real* Y2K problem that would curl one's hair, so to speak, with everything imaginable failing—and probably all blamed on Y2K.

Admirably, beginning just at the end of 1999, the friendly nation simply began methodically and massively cleaning out all the CSRs, and they disposed of them in very short order! Such CSR systems[9] may easily be located and destroyed by strategic longitudinal EM wave weapons {52}, when and if one has the longitudinal EM wave technology sufficiently developed. So Y2K came and went with only a few relatively minor incidents from the *normal* Y2k problems expected.[10]

However, unknown to our scientists and intelligence community, a great war of a very special kind—*the Earth's first strategic subspace war*—had just been silently fought and won, right there inside our power systems, command and control systems, major weapon systems, electrical power systems, electrical control systems of our nuclear submarines and strategic bombers and missiles, even inside the electronics of our nuclear warheads, etc. And not a bobble showed on the surface, anywhere. This was the first real "subspace" war, where weapons made of pure spacetime curvatures and dynamics were involved as the main offensive battle systems. LWI destruction of numerous CSRs inside potentials and fields in various electronic equipment throughout the nation, created little or no overt signs at all.

This time we won, thanks to the "good guys" and not to our own scientific community. Next time we may not be so fortunate.

The Mostly-Ignored Time-Energy Interaction

For some inexplicable reason, physicists in the West have omitted one-half the actual photon interaction.[11] If one allows the spatial energy component of the photon to interact, then one must also account for the simultaneous interaction of the time-component of the photon. A mass m absorbing a photon (AE)(At) becomes *masstime* (m+dm)t, not mass m. Mass *m* and masstime *mt* are as different as are force *F* and impulse *Ft.* No scientist in his right mind would ignore the *t* in *Ft* and consider just the force *F.* Yet that exact kind of giant *faux* ***pas*** has been done for many decades now in the photon interaction. And it is still ubiquitously done in physics. *Mass does not emit photons; masstime does.* Again, physicists confuse effect and cause.

---

[9] These early CSRs apparently had no "stealth" capabilities built-in, so they were easily detected and destroyed, rather like shooting fish in a rain barrel.

[10] Even so, a normally disastrous Y2K problem was narrowly averted, in the nick of time, by a concerted national software reprogramming effort massively mounted and costing billions of dollars. Otherwise, the expected "ordinary" Y2K problem would have been very bad indeed.

As the Russian energetics weapon scientists know full well, time is actually spatial energy compressed enormously, by at least the factor c^2. Hence time has at least the equivalent energy density of mass. The greatest amount of "energy" existing in a photon—particularly in lower frequency photons—is in that "time" part that is arbitrarily ignored and discarded. An ELF photon has enormously more total energy—by many orders of magnitude—than does a gamma ray photon.

Energy and time components in the photon are canonical. Low frequency photons are reduced in the magnitude of their *spatial* energy component and thereby maximized in their *time* components. For the ordinary (decompressed) *energy equivalency* of the time component, one must multiply the time in seconds by c^2. Halving the frequency, e.g., doubles the time component, whose spatial equivalent energy is now $(2t)\times(c^2)$. With respect to the original photon, the new total energy (achieved by halving the frequency) is given by $[(\Delta E)/2] \times [(2\Delta t) \times c^2] = [(\Delta E)(\Delta t)\times c^2]$. So halving the photon's frequency increases its total energy by a factor of c^2, which is approximately by a factor of $9x\,10^{16}$. Hence photons actually carry enormously greater total energy (most is locked up as time) at low frequency than at high frequency. Indeed, the highest energy photon interactions in the universe are at low frequency, *if and when* a transduction of the available time-energy into spatial energy occurs.

This has been completely missed in the West, but not by the Russian energetics weapons scientists. They have been *engineering* that formidable new high energy but low frequency physics—a physics far more energetic than that used in our most energetic colliding beams in "conventional high energy physics".

The cold fusion processes[12] stumbled onto by Western scientists are using the

---

[11] I.e., except for accounting for the overall "action" (angular momentum) and such things as **minimum** action theorems, etc. Treating the photon's increment of time specifically as a formidably compressed piece of spatial EM energy does not appear in Western physics. It has **long** been utilized in the KGB's secret energetics weapon science.

[12] For an explanation of the use of time-reversal zones in cold fusion experiments, and some of the resulting new transmutation reactions that result, see T. E. Bearden, *Energy from the Vacuum: **Concepts** and Principles*, "Chapter 10: Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy," Cheniere Press, Santa Barbara, CA, 2002. Available from www cheniere.org. Statistically, the formation of significant time-reversal zones is guaranteed by **the transient** fluctuation theorem of Denis Evans et al., and by recent experimental work by Evans et al. proving that such zones (where the reactions run backwards) can and do occur at up to **micron** (colloidal) size and for up to two seconds in duration. We stress that this is a recognized, **proven** violation of the second law of thermodynamics already known to physics.

time-energy component of photon interactions, to generate enormous localized EM energy directly upon the interior quarks and gluons of nucleons {53}. We worked out a speculative explanation of the active time-energy transduction mechanisms that fits all the experimental phenomena, including the eerie instrument anomalies occurring in rigorous electrolyte experiments at China Lake for some years. That is included in my new book, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, available from www.cheniere.org.

Of course consistency alone does not prove it conclusively; consistency is a necessary but not sufficient condition for proof. But at least the proposed explanation is consistent, and nothing else seems to be, at this time. In one chapter in my referenced new book we do cover the cold fusion mechanism and give specific new nuclear reactions producing the excess deuterium, tritium, and alpha particles in so many successful cold fusion experiments worldwide.

My "time-reversal zone" approach is stringent, because the recent work by Denis Evans *et al.* {54} has clearly shown that reactions can and do run in reverse—totally in violation of the second law of thermodynamics—at micron (colloidal) level in size and for up to two seconds in duration (and sometimes even a little longer). So experimental demonstration that the "time reversal zones" I assumed can and do indeed occur has now been accomplished by Evans *et al.* This provides strong experimental and theoretical support for the basic assumption (of the formation and momentary existence of localized time reversal zones that subsequently decay) we used to derive the fundamental cold fusion mechanism including some of the specific reaction equations for excess deuterium, tritium, and alpha particles now demonstrated in hundreds of successful experiments.

In a reversed zone *the law of attraction and repulsion of charges is reversed,* so that like charges attract—sometimes so closely that each enters the strong force region of the other, forming a quasi-nucleus (Figure 21). As the time-reversal zone subsequently decays back to a time-forward zone, the reduced strong force is restored much faster than the Coulomb force is reduced and then reversed back to normal. So during this readjustment decay process, the quasi-nucleus will simply change to an isomer and a known nucleus by quark-flipping to change an H+ ion to a neutron. Note that the "reversal zone" momentarily changes the Coulomb barrier—always the only thing that has really prevented chemical fusion and transmutation at low spatial energy—to a Coulomb attractor. Hence cold fusion—as now backed up by more than 600 successful experiments worldwide, by multiple laboratories and researchers in multiple nations—is an established fact and a herald for a vast new science based on deliberate and intentional reversal of the Coulomb barrier and use of time-energy.

With its usual perspicacity, the organized U.S. scientific community has flubbed it—and flubbed it badly—with respect to cold fusion. Its contribution has largely been a tirade of ad hominem attacks, slander, and libel of a viciousness seldom seen in the annals of science. Once again, the community is far more devoted to defending its cherished dogma (and its own rice bowls from the conventional Big Nuclear Science that cold fusion thoroughly upsets), than it is in applying scientific method which calls for *believing the experiments* and changing the dogmatic theory.

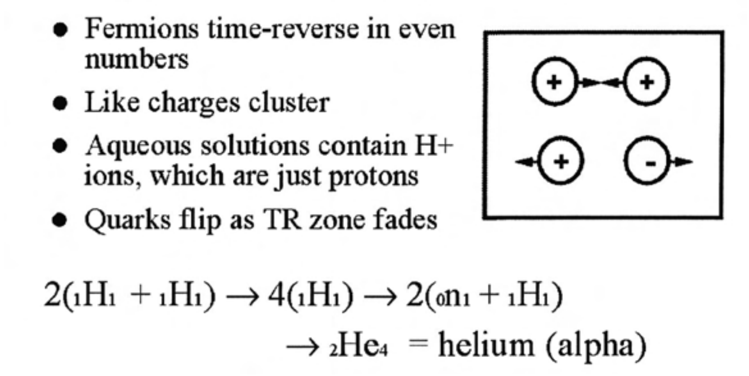

- Fermions time-reverse in even numbers
- Like charges cluster
- Aqueous solutions contain H+ ions, which are just protons
- Quarks flip as TR zone fades

$$2({}_1H_1 + {}_1H_1) \rightarrow 4({}_1H_1) \rightarrow 2({}_0n_1 + {}_1H_1)$$
$$\rightarrow {}_2He_4 = \text{helium (alpha)}$$

Figure 21 Example of a time-reversed zone nuclear reaction common in successful cold fusion experiments.

Development of Threats and Scheduled Attacks Aborted

KGB energetics weapons have long constituted our single most pressing yet largely unrecognized national security problem. That problem has also increased with the growing development and deployment of such weapons by other nations such as China. Some 10 nations, for example, now have scalar interferometry weapons, and even the Japanese Yakuza has them and also has their own secret facilities for producing them in Japan, including small portable units developed for use in assassination and terrorist type warfare within a targeted nation—the United States of America.

Unknown to our government, we narrowly escaped destruction twice in the first half of 1997 from scheduled all-out strategic KGB energetics weapons attacks. Only the actions of a friendly small country countered the attacks and caused the KGB to abort on each occasion.

When the second scheduled Russian strategic attack was aborted just prior to May 1, 1997, a new KGB counter-counter for the friendly counter used by the little nation in latter April 1997 was already rapidly being deployed. The new Russian counter-counter—seizure and control of the mind of an operator at a distance, and complete control of his actions while he performs sophisticated technical tasks—was tested in two separate incidents {55}, first on Captain Button (Figure 22, Figure 23) and then on Captain Svoboda about two months later (Figure 24), each flying an A-10 Warthog aircraft over the U.S.

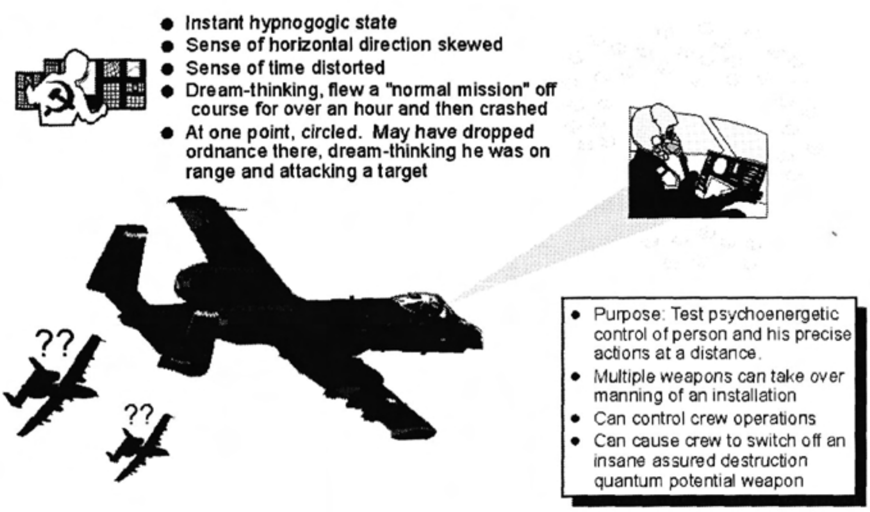

- Instant hypnogogic state
- Sense of horizontal direction skewed
- Sense of time distorted
- Dream-thinking, flew a "normal mission" off course for over an hour and then crashed
- At one point, circled. May have dropped ordnance there, dream-thinking he was on range and attacking a target

- Purpose: Test psychoenergetic control of person and his precise actions at a distance.
- Multiple weapons can take over manning of an installation
- Can control crew operations
- Can cause crew to switch off an insane assured destruction quantum potential weapon

Figure 22 Captain Button's bizarre flight to his death on 2 April 1997.

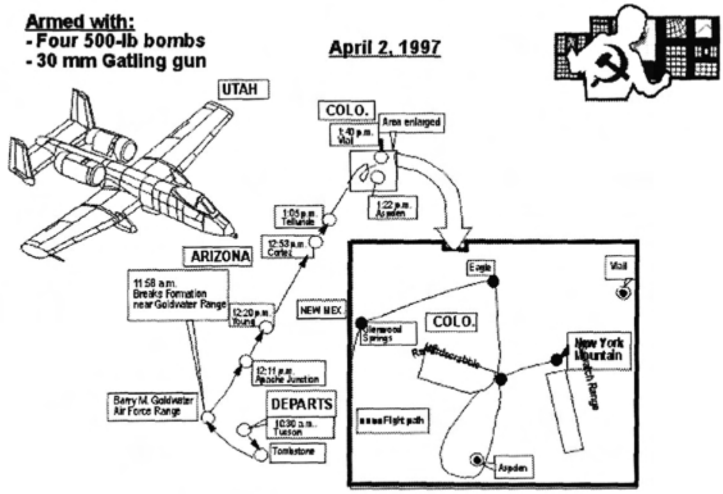

Armed with:
- Four 500-lb bombs
- 30 mm Gatling gun
April 2, 1997
UTAH
COLO.
Area enlarged
1:40 p.m. Vail
1:22 p.m. Aspden
1:05 p.m. Telluride
12:53 p.m. Cortez
ARIZONA
11:58 a.m. Breaks Formation near Goldwater Range
12:20 p.m. Young
NEW MEX.
12:11 p.m. Apache Junction
Barry M. Goldwater Air Force Range
DEPARTS
10:30 a.m. Tucson
Tombstone
Eagle
Vail
COLO.
Glenwood Springs
New York Mountain
Flight path
Aspden

Figure 23 Captain Button's flyaway path to his eventual fatal crash.

The highly successful test against Captain Button (Figure 22, Figure 23) is particularly significant, as we have explained. Captain Button was controlled while performing highly technical operations, for more than a full hour, and one hour was apparently the acceptance test specification. Captain Svoboda (Figure 24) simply had her mental perception of up and down instantly reversed while climbing her A-10 warthog from low level ordnance delivery on target. Erroneously perceiving she was diving, she immediately "corrected" sharply to "climb", thereby diving headlong into the ground and perishing in the resulting explosion of her aircraft.

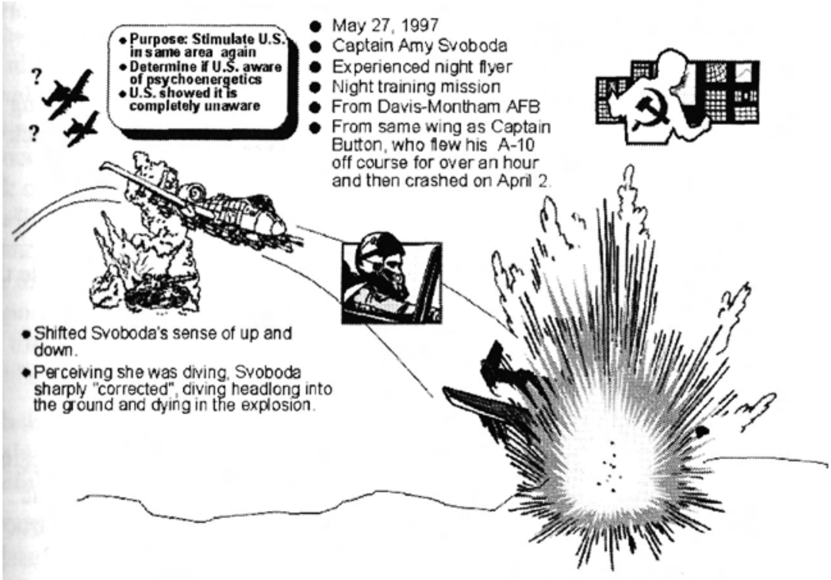

- Purpose: Stimulate U.S. in same area again
- Determine if U.S. aware of psychoenergetics
- U.S. showed it is completely unaware
?
?
- May 27, 1997
- Captain Amy Svoboda
- Experienced night flyer
- Night training mission
- From Davis-Montham AFB
- From same wing as Captain Button, who flew his A-10 off course for over an hour and then crashed on April 2.
- Shifted Svoboda's sense of up and down.
- Perceiving she was diving, Svoboda sharply "corrected", diving headlong into the ground and dying in the explosion.

Figure 24 Climbing out of low-level ordnance delivery, Captain Svoboda suddenly dives headlong to her death on 27 May 1997.

Svoboda's kill demonstrated a *tactical destruction* usage, whereas Button's test demonstrated a *strategic* hostile usage of the adversary's defending weapons crew to take action against the adversary's defending strategic weapons system. A new KGB strategic energetics attack against the United States—using the mind control (psychoenergetics) weaponry to take over the operators of the little nation's QP weapon sites and stand the sites back down for maintenance—was rescheduled apparently for latter 1997 to mid-1998 or so. It appears that the extensive preparations for that attack were totally destroyed clandestinely and very suddenly by the friendly little nation some time after we warned them, as we mentioned above.

However, still another Russian counter—the development of "weaponized en**gines"** (functioning causal system robot weapons, "infolded electromagnetic" in **form)** consisting of organized and functioning systems of spacetime curvatures operating in and traveling through EM fields, potentials, waves, and signals—came on line (Figure 10, Figure 18). The KGB simply altered the attack plan for a Millennium Strike and completed the forward deployment of these eerie new CSR systems by introducing millions of them directly throughout U.S. facilities, command centers, weapon systems, power grids, ships, aircraft, submarines, missiles, nuclear warheads, computers, communications systems, etc.

We estimate that several million such systems were introduced in that manner, and readied for the Millennium Strike. The United States and its allies were and are almost totally unprepared for this dramatically new kind of strategic warfare that has been developed and deployed against us. No such weapons have previously existed in history. Consequently our national survival and the survival of our allies was once again at stake—and right on schedule during the critical period from mid-1999 through mid-2000 that we had predicted back in 1997.

The die-hard elements of the KGB indeed had a new millennium gift for us, but it was not at all the kind of gift we would appreciate. The KGB is not a monolith! This plot apparently did not include Putin, and it did not include the rest of the KGB, particularly that element reporting to Putin.

Live Destructive Tests Over and Near the U.S. Have Continued

Without elaboration, we believe that the Learjet crash incident and the EA-990 crash incident may have been very sophisticated tests of these new robotic causal EM systems. We also believe that the long years of Russian construction of over a thousand mysterious underground facilities, even though the Russian economy was collapsed and the citizens and soldiers were destitute, is significant. Those facilities apparently contain the set of command and control systems for this vast new strategic weapon system.

The facilities also include the "factories" where the CSRs are made, because the factories are unlike any other factories ever conceived. They do not build "physical" systems, but something more like "software." Only now the "software program" itself is an independent robotic system, with a full set of weapon systems functions, and capable of independent navigation and functioning anywhere, while remaining under KGB command and control via longitudinal EM wave communications infolded inside the potentials of earth, ocean, and space. The facilities also contain the operational forces, headquarters, and communication and control systems that maintain, deploy, and employ (fight) these CSR systems, and direct the battle.

The purpose of the 1998 document we sent to DoD—which did not include the causal robot system armadas later deployed throughout everything in the West—was to inform selected members of the Senate and Congress of the energetics threat situation and summarize it for them. Accordingly, we presented the background of the development and testing of the KGB energetics weapons since WW II, in some detail, for selected Legislators and their staffs. We also included some specifics on the failure of the U.S. scientific community to keep American science and technology abreast of that of our adversaries.

The Scientific Basis For Energetics Weapons Is Established

As pointed out above, the scientific basis for our early use of Soviet scalar EM interferometry is now rigorously established by work accomplished by the Alpha Foundation's Institute of Advanced Study. It is also rather "old hat" now that all major weapons labs have discovered longitudinal EM waves and are actively working on and with them. The Director of the AIAS, Dr. Myron Evans, is an eminent scientist and theoretician with over 600 papers in the hard science literature. He and his colleagues also have had zero funding for their work on the higher group symmetry electrodynamics necessary to understand such things as energetics. A combination of non-Abelian electrodynamics, advanced gauge field theory, and the impact of E. T. Whittaker's work shortly after the turn of the century has shown that scalar interferometry is real, and that weapons based on scalar interferometry can be a potent threat on the battlefield. The time-like (time-energy) basis for electrodynamics is now clearly established.

With this important new work, we are now only some 40 years behind the Russian weapons scientists! But we are beginning to close the gap—and hopefully that will be done very rapidly.

In 1999 after being personally attacked by a CSR, I again contacted the office of the leader of the friendly small nation and again furnished all the information we possess on the new CSR systems programmed by the KGB as a countering method against the small nation's QP weapons. The outcome was that the little nation destroyed the CSRs in time, as we stated above. We were lucky that (i) the KGB did attack me with one of the weapons, deliberately to "get my strong and immediate attention", and (ii) deliberately did not kill me with the strike. Otherwise, there would have been no warning to the little nation, and the U.S. and its allies would have experienced a "Y2K" problem worse than their wildest fears, possibly collapsing the economy of the United States and much of the developed world.

Our Situation Here In the U.S. Is Precarious

Meanwhile, our own scientific community apparently does little or nothing in these areas, except to castigate, slander and libel anyone daring to suggest that such "out of the box" weapons areas even exist. E.g., there does not appear to be a single U.S. scientist who knows how to create a *time-polarized* EM wave.[13] There do not seem to be any concerted programs to deal with t-polarized photons and t-polarized EM waves. Our scientists still firmly believe that one has to propagate spatial EM field energy through space to affect a target electromagnetically, when that was obsoleted by Russian "subspace EM longitudinal wave propagation" weapon systems deployed as early as 1963, and was completely obsoleted by deployment of quantum potential weapons in 1989.

Yet a close colleague of the present author has single-handedly developed and demonstrated prototype superluminal communication systems using the infolded (longitudinal EM wave) electrodynamics, and these systems would have been heading onto the market in latter 2001 except for the sudden bankrupting of the major backing company. There has not been any funding forthcoming for his revolutionary work, but only clumsy attempts to take it from him. It is my fear that the further clumsy takeover attempts will be successful, and this desperately needed technology will never see the light of day in defense of our nation. As a single example without further discussion, his work revolutionizes the present notions used in the rapid development of quantum communication and quantum computing, because it revolutionizes the very notion of the Q-bit, completely removing the statistics of the quantum state and making it engineerable. This of course is completely counter to present QM notions, but is consistent with the Bohm hidden variable approach to quantum mechanics.

Meanwhile, our own weapon scientists apparently still do not recognize the possibility of a superluminal longitudinal EM wave communication system, and appear to have no motivation or intention to develop such. This is in a way quite sad; quantum tunneling at more than four times the speed of light, of a clearly recognizable Mozart symphony, has already been demonstrated in a section of a waveguide {56}. Whatever one wishes to call it and however one wishes to do it, superluminal communication of intelligible signals has been clearly demonstrated in the laboratory. For the "infolded" longitudinal EM waves, all the "normal" or "envelope" EM waves, potentials, and fields are just great *superwaveguides* for such longitudinal EM wave "tunneling through subspace" at superluminal speeds.

---

[13] Keeping one's sense of humor, apparently there has also not been a U.S. scientist who even recognized and knew what actually powered an electrical circuit, until our solution to the source charge problem and publication of it in 1998. Eerily, even after the technical basis for a solution had been proven in particle physics in 1957, no one seemed to have applied the broken symmetry of opposite charges to the common dipole and to a dipolarity such as a scalar potential and a charge (with its associated virtual charges of opposite sign, in the modern view). Everyone accepted that the fields and potentials and their energy are established by their associated source charges, but no one had stated that the source charge receives its input energy from its asymmetry in its seething energy exchange with the active vacuum.

The Dominant Weapons Are Quantum Potential Weapons

Full-bore deployment of Russian quantum potential weapons occurred by the end of 1989, the first lab prototype weapon having been tested in April 1986 against the U.S. air attack on Libya (Figure 25). Beside Russia, two other nations—Brazil and the "little friendly nation"—also had developed quantum potential {57} weapons. The United States did not develop them, or at least apparently not just yet.

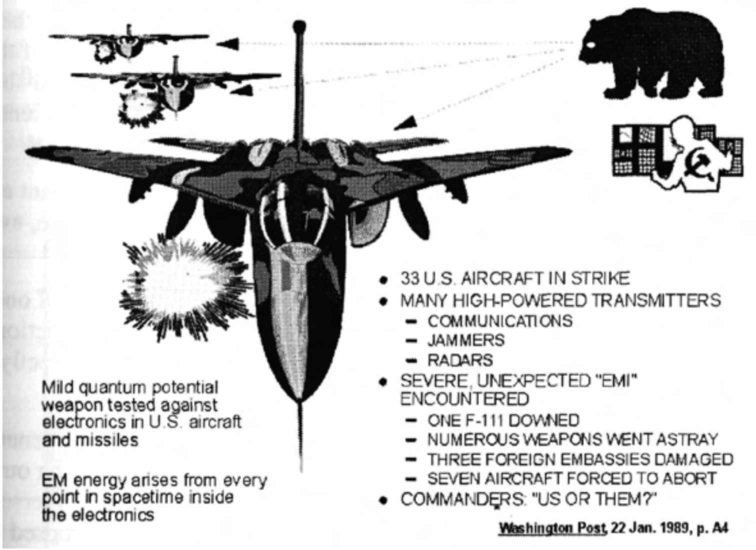

- 33 U.S. AIRCRAFT IN STRIKE
- MANY HIGH-POWERED TRANSMITTERS
  - COMMUNICATIONS
  - JAMMERS
  - RADARS
- SEVERE, UNEXPECTED "EMI" ENCOUNTERED
  - ONE F-111 DOWNED
  - NUMEROUS WEAPONS WENT ASTRAY
  - THREE FOREIGN EMBASSIES DAMAGED
  - SEVEN AIRCRAFT FORCED TO ABORT
- COMMANDERS: "US OR THEM?"
Mild quantum potential weapon tested against electronics in U.S. aircraft and missiles
EM energy arises from every point in spacetime inside the electronics

Washington Post, 22 Jan. 1989, p. A4

Figure 25 U.S. air strike on Libya, April 1986.

Burlier this year (2002), China also deployed QP weapons, and we now find that an*other friendly* nation has them as well. That brings the total to five, as of this writing. Two of those are friendly, and a third is astraddle of the fence, so to speak. The other two are Russia (the KGB) and China.

At this point China could go either way, so it represents a future enigma. China **has** control of the Panama Canal, and has inserted more than 200,000 Chinese into Panama—which is on the way to gradually becoming a Chinese territory, as soon as the Chinese in Panama outnumber the native Panamanians. In addition, under the Clinton administration China gained access through two U.S. ports (with little checking, so anything can easily be inserted). China also has a**dditional** ports in the Western hemisphere. China is almost certain to eventually attack and seize Taiwan, which puts her on a collision course with the U.S., si**nce** the U.S. has an agreement to defend Taiwan. She has also declared the South China Sea—through which passes some 60% of the oil bound for Japan, a**nd** much oil for other nations also—to be Chinese territorial waters. China also has developed bioenergetics weapons, having long ago induced cancer into most of the GRU Russian representatives then serving in China. She almost certainly has or is rapidly working to develop psychoenergetics weapons as well.

So far as I am aware, there still exists not a single funded U.S. scientific program to directly correct the glaring errors in our own classical electrodynamics—such as those problems deplored by Nobelist Feynman and Wheeler {58,59,60}. Both knew that the force field concept in space was erroneous. As they put it, only the "potential" for the force field exists in space, in case some charged mass is available there for interaction. But the field itself is created by the interaction and is the effect of it. It does not exist before the interaction occurs. Yet all our texts continue to teach the disinformation of the "field in space." And our science community continues to fiercely defend such dogma, attacking and sometimes destroying the careers of those brave scientists who do try to change it.

One notes that time is multiply-connected, since in theory the same instant exists at every point in the universe. Or said another way, in any instant in time, every point in the universe simultaneously exists.

Operating in the time-domain as a causal system, and then turning out of one point there into any point in 3-space desired, is a means of producing "action at a distance" that obsoletes our old concept of transmitting the energy directly through space {61}.

Presently the U.S. scientific community has little predisposition toward energy turning into the time-domain, operating in the time-domain, then turning out of the time-domain into any desired point in 3-space, anywhere in the universe.[14] Every 3-space point in the universe simultaneously exists in (is superposed in) each point in time, as those points in time occur. In one sense, time can be considered to be the total multiple connectedness of space! We suspect that it may even be possible to interpret "instantaneous" communication—as by a quantum potential—in that "in here-out there" instantaneous fashion via "tunneling through the multiple connection of the time channel".

Finally, mass—being an observable and 3-spatial—does not exist in time. Mass continually turns into masstime (by photon absorption), then masstime turns back into mass (by photon emission). Masstime exists in time, but mass does not. U.S. scientists also seem unaware that the so-called "rate of flow of a mass through time" is generated by the sum total of all photon interactions (both virtual and observable) with that mass. The ubiquitous interaction of a mass with photons also causes that mass to continually recur, and thus to seemingly "continuously exist." It is continually observed, but it does not continuously exist! Since the photon interaction is engineerable, then it follows that a specific muss can be time-reversed back to an earlier state {62}.

---

[14] If one models the time domain in three dimensions instead of one, e.g., then in theory it becomes possible to "flip" a 3-dimensional mass object into the time 3-domain at a single "moment," then rotate from that "time 3-point" back into 3-space at a new 3-space location anywhere in the universe. Hence the "beam 'em up Scotty!" of the old Star Trek (and the more modern beaming version) may hold more of a promise than just for science fiction entertainment.

Cellular Time Reversal: Revolutionary Healing by a Team of French Scientists

Priore and eminent French scientists {39,40} unwittingly used the mechanism for time-reversing a mass and every part of it, to time-reverse diseased and damaged cells, including tumor cells, physically changing them back to their previous healthy physical state and condition in vivo. This included all parts of **the** cell, including its genetics. The Priore team demonstrated revolutionary **cures** of terminal tumors, infectious diseases, atheriosclerosis, and restored depressed immune systems.

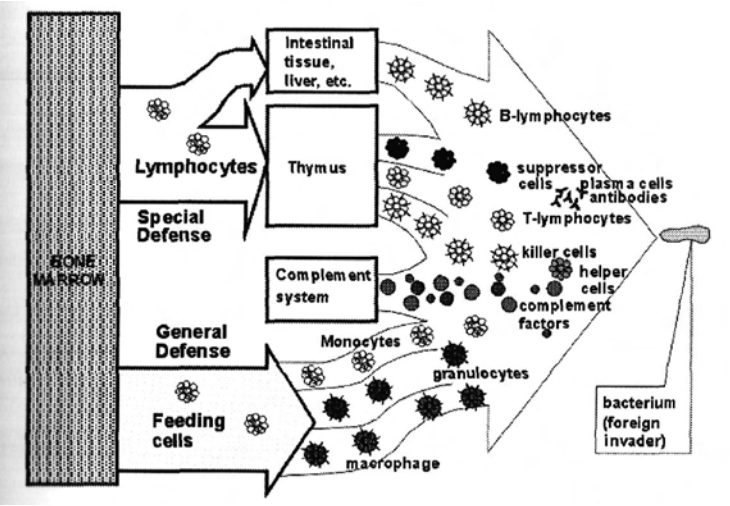
BONE MARROW
Intestinal tissue, liver, etc.
Lymphocytes
Special Defense
Thymus
Complement system
General Defense
Feeding cells
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
Monocytes
granulocytes
macrophage
bacterium (foreign invader)

Figure 26 Simplified diagram of the immune system.

The immune system (Figure 26) *heals nothing, not even its own damaged cells.* The human cellular regenerative system {63} is responsible for all *healing* in the body—and it utilizes time reversal of the damaged cells as the active healing mechanism. The regenerative system (Figures 27, 28, 29) pumps the damaged cells with longitudinal EM waves (and therefore with time-polarized EM waves, which always accompany longitudinal EM waves), inducing the cells and all their parts to add the phase conjugates, thereby forming time-polarized EM waves and pumping in the time-domain rather than the 3-space domain. This dramatically extends nonlinear optics to the time domain, and mass pumped in that fashion is time-reversed (propagated back along its own time-track, to a previous physical condition.

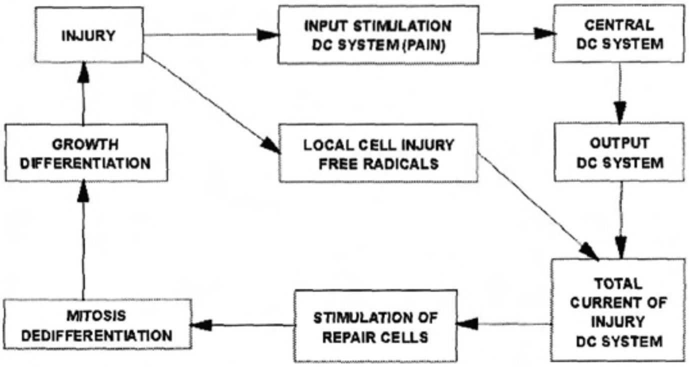
INJURY
INPUT STIMULATION DC SYSTEM (PAIN)
CENTRAL DC SYSTEM
GROWTH DIFFERENTIATION
LOCAL CELL INJURY FREE RADICALS
OUTPUT DC SYSTEM
TOTAL CURRENT OF INJURY DC SYSTEM
MITOSIS DEDIFFERENTIATION
STIMULATION OF REPAIR CELLS

Note: Becker only had conventional EM available, hence was unable to model the inner structuring and inner dynamics of the fields and potentials. *So he missed the deeper "engine and antiengine" mechanisms.*

"Robert O. Becker, "The significance of bioelectrical potentials," Bioelectrochemistry and Bioenergetics, Vol. 1, 1974, p. 191.

Figure 27 Becker's theoretical control system involved with response to injury.

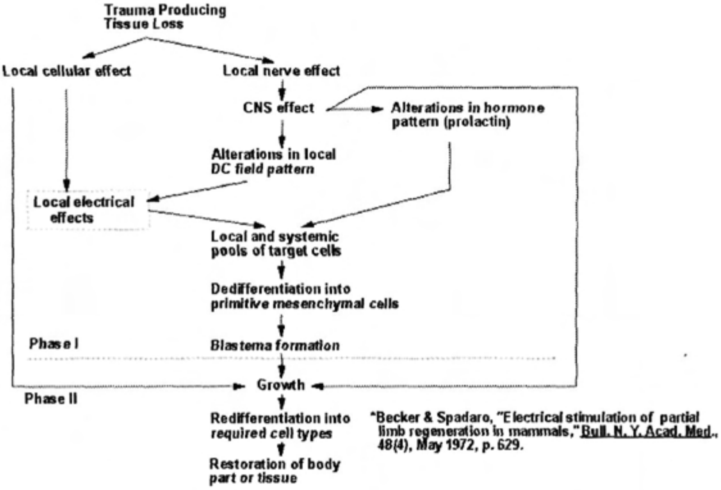
Trauma Producing Tissue Loss
Local cellular effect
Local nerve effect
CNS effect
Alterations in hormone pattern (prolactin)
Alterations in local DC field pattern
Local electrical effects
Local and systemic pools of target cells
Dedifferentiation into primitive mesenchymal cells
Phase I
Blastema formation
Growth
Phase II
Redifferentiation into required cell types
Restoration of body part or tissue

"Becker & Spadaro, "Electrical stimulation of partial limb regeneration in mammals," Bull. N. Y. Acad. Med., 48(4), May 1972, p. 629.

Figure 28 Becker's proposed control system governing regeneration.

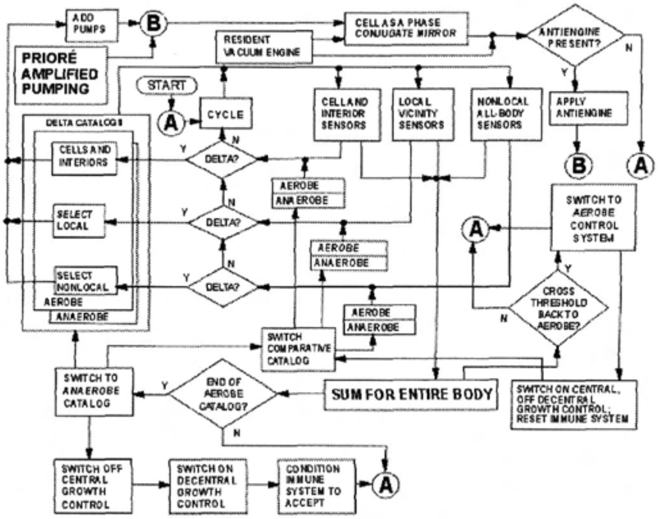
ADD PUMPS
B
CELL AS A PHASE CONJUGATE MIRROR
ANTIENGINE PRESENT?
N
RESIDENT VACUUM ENGINE
PRIORÉ AMPLIFIED PUMPING
START
Y
CYCLE
A
CELL AND INTERIOR SENSORS
LOCAL VICINITY SENSORS
NONLOCAL ALL-BODY SENSORS
APPLY ANTIENGINE
DELTA CATALOGS
CELLS AND INTERIORS
DELTA?
AEROBE
ANAEROBE
SELECT LOCAL
SWITCH TO AEROBE CONTROL SYSTEM
SELECT NONLOCAL
CROSS THRESHOLD BACK TO AEROBE?
SWITCH COMPARATIVE CATALOG
SWITCH TO ANAEROBE CATALOG
END OF AEROBE CATALOG?
SUM FOR ENTIRE BODY
SWITCH ON CENTRAL, OFF DECENTRAL GROWTH CONTROL; RESET IMMUNE SYSTEM
SWITCH OFF CENTRAL GROWTH CONTROL
SWITCH ON DECENTRAL GROWTH CONTROL
CONDITION IMMUNE SYSTEM TO ACCEPT

Figure 29 Author's deeper block diagram of the cellular regeneration system.

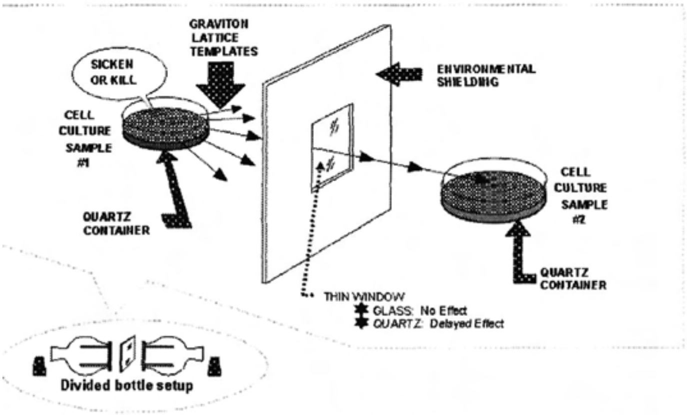
GRAVITON LATTICE TEMPLATES
SICKEN OR KILL
CELL CULTURE SAMPLE #1
ENVIRONMENTAL SHIELDING
QUARTZ CONTAINER
CELL CULTURE SAMPLE #2
QUARTZ CONTAINER
THIN WINDOW
GLASS: No Effect
QUARTZ: Delayed Effect
Divided bottle setup

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice. Consider particularly the dielectric pathway in the emitting diseased cells and in the receiving targeted healthy cells.

Figure 30 Kaznacheyev's electromagnetic induction of cellular disease and disorder.

Kaznacheyev has shown that essentially any cellular disease or disorder can be electromagnetically induced in cells at a distance (Figure 30). Together with Becker's work, this shows that the cells can be altered either detrimentally (disease induction) or beneficially (disease elimination) by purely electromagnetic means. It would seem that diseases such as AIDS, e.g., could be rather dramatically cured by use of whole-body longitudinal EM wave radiation (with its simultaneously time-polarized EM wave radiation), as was done for other diseases by the Priore team and properly reported in the French scientific literature. We have also uncovered a process and method for causing the body to directly produce the time-polarized EM wave pumping, reducing the irradiation to perhaps 30 seconds of pumping instead of two or three hours {64}.

Becker's Study of the Cellular Regeneration System

In all our medical science, the only true healing procedures being utilized appear to be the use of Becker *et al.'s* tiny EM stimulation across otherwise intractable bone fractures {65} (Figure 31), to dedifferentiate (time-reverse) red blood cells, then redifferentiate (time-forward) them again twice to form osteoblasts, which are deposited in the fracture site to make new bone growth. All the other medical procedures seem to be *interventions,* not *healing.* After the intervention—which of course may be urgently necessary—it is then up to the body to heal itself (restore its cellular and tissue damage and functioning). So we have little *healing* science today, and in fact our medical establishment hardly funds any *healing* studies at all. EM radiation is usually considered as something that simply heats tissue. Our present medical science is deeply engrossed in studying the immune system. Again, the immune system heals nothing, not even its own damaged cells.

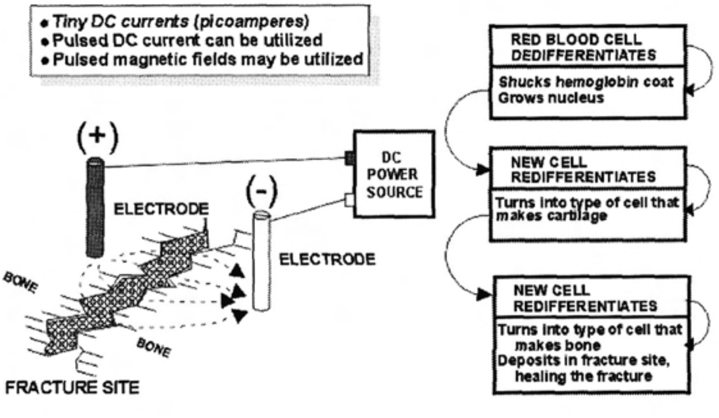
• Tiny DC currents (picoamperes)
• Pulsed DC current can be utilized
• Pulsed magnetic fields may be utilized
(+)
ELECTRODE
(-)
ELECTRODE
DC POWER SOURCE
BONE
BONE
FRACTURE SITE
RED BLOOD CELL DEDIFFERENTIATES
Shucks hemoglobin coat
Grows nucleus
NEW CELL REDIFFERENTIATES
Turns into type of cell that makes cartilage
NEW CELL REDIFFERENTIATES
Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

Figure 31 Becker's epochal bone fracture healing.

After the immune system "kills the bad guys" and saves the day, the battlefield is littered with residue and many cells of the body are damaged. The immune system sends in large scavenger cells that clean up the residue. And that's it for the immune system. Comparing it to an army's operation in the field, the immune system is the combat troops, not the medical services.

Restoration (healing) of the damaged cells back to normal is accomplished by the *cellular regeneration* system (Figure 29). This system is little known and little studied, but mostly studied by Becker (Figures 27, 28). It is a system that functions electrodynamically, but uses time-domain pumping to time-reverse the cells and their every part (including their genetics) back to an earlier state. In biology, the effect is called *dedifferentiation,* but without any knowledge of the actual EM mechanism.

So, within its capabilities, the cellular regenerative system in the body already uses longitudinally polarized EM waves, time-polarized EM waves (TPWs), and optical pumping by LWs and TPWs. Here we point out that neither the longitudinal photon nor the scalar (time-polarized) photon is individually observable; however, the combination of the two is observed as the instantaneous scalar potential {66}. With a little correction to properly make the phase conjugate replica wave a time-polarized EM wave, Whittaker's fundamental bidirectional EM longitudinal wave decomposition of the scalar potential (Figure 32) is consistent. Immediately the direct involvement of both **the** scalar photon and the longitudinal photon—and both the scalar EM wave and the longitudinal EM wave—are seen in Becker's bone-healing work using a scalar potential (Figure 31).

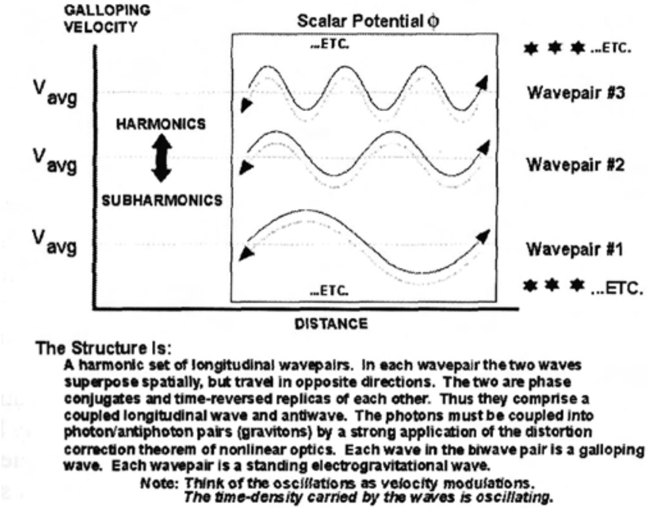

The Structure Is:

A harmonic set of longitudinal wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave.

Note: Think of the oscillations as velocity modulations.
The time-density carried by the waves is oscillating.

Figure 32 Infolded longitudinal EM biwaves composition of a scalar potential, per Whittaker 1903.

Priore unwittingly stumbled onto the actual mechanism used by all biological systems in "healing" (cellular regeneration), and also recognized how to amplify it **(Figure** 12). But no one understood its exact nature, including Priore and the scientists who worked with him at the time, and the conventional community still does not understand it. The Russian scientists not only *understood* it early on, but *weaponized* it as well. So did the Chinese.

Need to Correct and Extend Classical Electrodynamics

The organized scientific community's protracted defense of a seriously flawed classical electrodynamics is long standing. The conventional, comfortable ways of "doing business as usual, leisurely" now threatens the survival of the nation itself. We need to quickly admit and correct long-standing EM errors—such as calculating a reaction cross-section of an entity and calling it the magnitude of the entity itself! Our electrical texts continue to make that fundamental mistake, and they have done it now for 100 years or more. We need to quickly change the classical EM model to include the active vacuum and its interaction, and we need to change all textbooks to clearly state that all EM energy in the universe comes from the active vacuum, via the asymmetry of the source charges in their fierce energy exchange with it.

We also need to point out that all primary electromagnetic entities—the charge, the field, the potential, and every joule of EM energy in the universe—totally violates the second law of thermodynamics. The charge produces continuously increasing negentropy, the field and the potential are formed at expanding light speed across the universe and are perfectly ordered (thereby totally violating the statistical mechanics basis of the second law), and every joule of EM energy is either field energy or potential energy, and hence violates the same statistical basis of the second law.

The Situation Is Serious

Admittedly these are strong statements highly critical of our scientific establishment's track record. But our national survival continues to be increasingly at stake, the situation is serious, and we must not continue to bury our collective heads in the sand like ostriches. We simply cannot continue to idly sit by and pontificate while other groups and nations methodically maneuver technically to cut our collective throats with advanced weapons we cannot defend against, once they find a way to "get by" the little nation that has been countering them for some years. Our own beloved nation may have less than 10 years {67} and perhaps less than three years to survive, unless the organized scientific community is galvanized to focus its very best scientific efforts in the energetics area.

So far, in truly innovative approaches to new methods for electrical energy, the U.S. scientific community is about as energetic as a sand terrapin sleeping in the tropical sun. There is no discernible move to recognize the seriousness of the problem, or to do anything effective about it. In 45 years, the asymmetry of the dipole (and the charge in its modern view as a dipolarity of special kind) has not made it across the university campus from the particle physics department to the electrical engineering department. In fact, the implication of broken symmetry to

electrical engineering and classical electrodynamics seems to still be completely unnoticed by the National Academy of Sciences, the National Science Foundation, our universities, and our great national laboratories. In the scientific community, other than work on vaccines etc., it is largely "business as usual."

A New Manhattan Project Is Warranted

Our need in the energetics weapons area is extreme, as are our needs in the medical area for quick, cheap, easy, and effective portable treatment of the coming mass casualties and our coming dramatic needs in energy. We *do* have superb scientists with the horsepower necessary to do something about it. The astounding, independent work of Sachs, the AIAS, and a few other scientists—which we stress has been unfunded—already clearly proves that. But unless we change the organized scientific community's normal rather snail-like pace—by a strong measure such as a Presidential Decision Directive and a declaration of a National Emergency—the U.S. scientific community will overshoot the point of no return now approaching.

We need a new Manhattan-type project. We need it very quickly. If we do not change and focus our present scientific procedure, then—all too shortly—little that the scientific and intelligence communities are doing will be of value. We shall all be very, very dead or else in a new Dark Ages where most of the world economy has collapsed and freedom has largely disappeared from the face of the Earth.

T. E. Bearden, Ph.D.
Lieutenant Colonel, U.S. Army (Retired)
President & CEO, CTEC, Inc.
Director, Assoc. of Distinguished American Scientists

2002 ENERGETICS WEAPONS THREAT TO THE U.S.

T. E. Bearden 2000
Updated Oct. 1,2002

The U.S. Narrowly Escaped Energetics Destruction Twice in 1997

On two occasions in 1997 (latter February and on May Day), we were scheduled In be strategically attacked without warning and completely destroyed. Eerily, in public statements the leader of the Aum Shinrikyo sect (Figure 33) had foreshadowed the dates for the planned attacks, since his group was also involved—in fact, shooting down the TWA-800 (Figure 34) as advanced practice and stimulation of the U.S. to see if it knew what was afoot. Those two planned strategic strikes in 1997 were avoided only in the nick of time by actions of a friendly little foreign country, which was the only nation on earth at the time that could counter the KGB quantum potential (QP) weapons. I am sure that proper U.S. agencies intercepted my urgent personal faxes to the Defense Attache in Washington, DC and directly to the Prime Minister of the nation involved. Actions by that nation did succeed in averting our destruction in the nick of time.

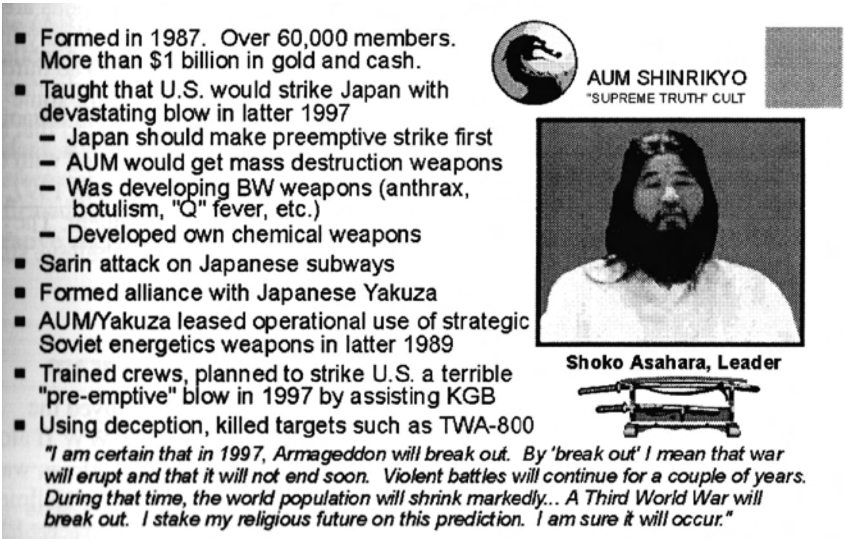

- Formed in 1987. Over 60,000 members. More than $1 billion in gold and cash.
- Taught that U.S. would strike Japan with devastating blow in latter 1997
  - Japan should make preemptive strike first
  - AUM would get mass destruction weapons
  - Was developing BW weapons (anthrax, botulism, "Q" fever, etc.)
  - Developed own chemical weapons
- Sarin attack on Japanese subways
- Formed alliance with Japanese Yakuza
- AUM/Yakuza leased operational use of strategic Soviet energetics weapons in latter 1989
- Trained crews, planned to strike U.S. a terrible "pre-emptive" blow in 1997 by assisting KGB
- Using deception, killed targets such as TWA-800

AUM SHINRIKYO
"SUPREME TRUTH" CULT

Shoko Asahara, Leader

"I am certain that in 1997, Armageddon will break out. By 'break out' I mean that war will erupt and that it will not end soon. Violent battles will continue for a couple of years. During that time, the world population will shrink markedly... A Third World War will break out. I stake my religious future on this prediction. I am sure it will occur."

Figure 33 Planned KGB/Yakuza/Aum Shinrikyo attack in 1997 was foreshadowed in public statements made by Shoko Asahara.

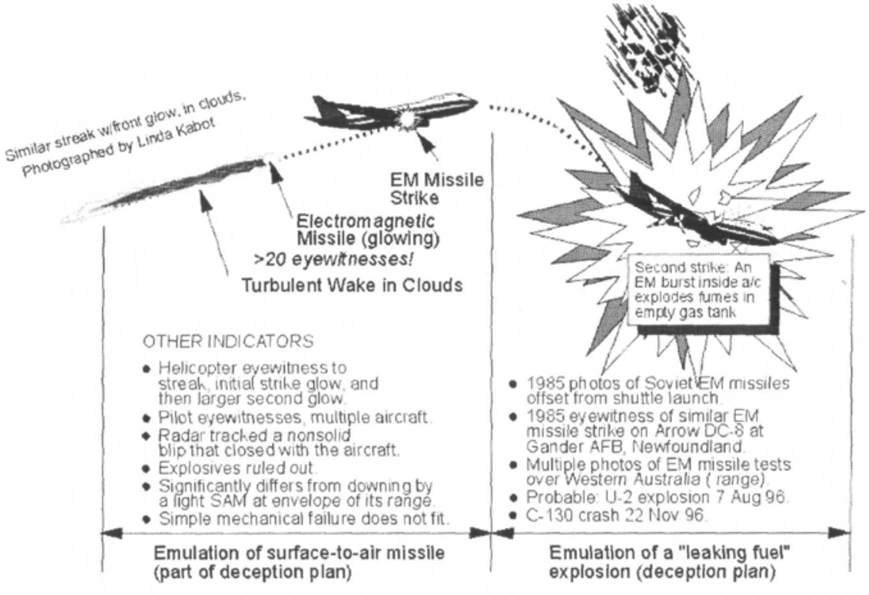
Similar streak w/front glow, in clouds. Photographed by Linda Kabot
EM Missile Strike
Electromagnetic Missile (glowing)
>20 eyewitnesses!
Turbulent Wake in Clouds
Second strike: An EM burst inside a/c explodes fumes in empty gas tank
OTHER INDICATORS
• Helicopter eyewitness to streak, initial strike glow, and then larger second glow.
• Pilot eyewitnesses, multiple aircraft.
• Radar tracked a nonsolid blip that closed with the aircraft.
• Explosives ruled out.
• Significantly differs from downing by a light SAM at envelope of its range.
• Simple mechanical failure does not fit.
• 1985 photos of Soviet EM missiles offset from shuttle launch.
• 1985 eyewitness of similar EM missile strike on Arrow DC-8 at Gander AFB, Newfoundland.
• Multiple photos of EM missile tests over Western Australia (range).
• Probable: U-2 explosion 7 Aug 96.
• C-130 crash 22 Nov 96.
Emulation of surface-to-air missile (part of deception plan)
Emulation of a "leaking fuel" explosion (deception plan)

Figure 34 Kill of TWA Flight 800 on 17 July 1996 by Yakuza and Aum Shinrikyo.

Only five nations or national groups have quantum potential (QP) weapons, though one or more additional nations are almost certainly working on them and may even have them. First of the five is the KGB in Russia (the weapons are not in the regular Russian forces, and never have been). Second is the friendly little foreign country to whom we presently owe our continued survival. The third is Brazil (which has had a small and successful QP weapon program for some time). China deployed QP weapons in early 2002, and I recently found that at least one other friendly nation has them as well. I'm not too impressed with that other friendly nation, since they played no role in saving our bacon in 1997. It appears they did not actually know the coming attacks were imminent. They certainly do not have Causal System Robot weapons, nor do they believe such weapons even exist.

Initiation of the Brazilian Energetics Weapon Program

Brazil implemented an energetics weapons program after it employed the repatriated Dr. Hellman, a German scientist taken to Russia after WW II along with the German radar team. Years after being taken to Russia, Hellman was repatriated and made his way back to Germany. In a news interview Hellman reported that he had worked directly on the development of Soviet secret EM weapons which could suddenly freeze large areas of the ocean surface. That of course is a scalar interferometry weapon, producing negative (converging) EM energy rather than positive (diverging) EM energy in their distant interference zones.

Also, when the U.S. made David Bohm *persona non grata* in the McCarthy witch-hunts, Bohm was later employed in lecturing and teaching in Brazil for some time. At the time our own scientists continued to pontificate as to whether they *liked* or could *tolerate* Bohm's quantum potential and his hidden variable interpretation of quantum mechanics {3}. Meanwhile, the KGB weapon scientists and the Brazilian Hellman project seized upon Bohm's hidden variable theory with alacrity as soon as he published his seminal paper. Those two weapon groups were already familiar with the internal longitudinal EM wave structuring of potentials, fields, and waves (Whittaker 1903), and immediately saw the fantastic weapons implications of a deliberately structured quantum potential, containing a specific "engine" as desired. Most of our scientists still are unaware of it to this day, and are unaware of its formidable weapon implications when combined with Whittaker's 1904 superpotential paper, his 1903 potential decomposition, and quantum field theory. Most of our scientists still adhere to the Bohr interpretation of QM, which excludes the normal, everyday, observable, ordered universe we all inhabit and observe.[15]

The Bohm quantum potential is, after all, a scalar potential, and at least in theory it has a Whittaker biwave decomposition {68}. Even in the Coulomb gauge, the scalar potential has infinite velocity, as pointed out by Jackson.[16]

Several tests of the "cold explosion" Russian weapons referred to by Hellman have since been openly made. One major test seen by at least five jetliner crews was the anomalous rapidly rising cloud and subsequent extremely huge light globe (ABM energy shield) effects off the Kurils in April 1984 (Figure 35). Two excellent scientific articles on the physical phenomena observed in that incident are published in the prestigious journal *Science* {69}. We had explained the cold explosion phenomena and the ABM energy shield effects some years before that incident. Even ordinary superpotential theory {70} and longitudinal EM wave interferometry at a distance can substantiate the capability. We also positively confirmed by private sources that the scalar interferometry, the MindSnapper (negative energy EMP), and the cold explosion weaponry are real.

---

[15] This exclusion is euphemistically known as the "missing chaos" problem in quantum mechanics, and it is resolved in Bohm's hidden variable approach.

[16] J.D Jackson, *Classical Electrodynamics,* 2nd Edn., John Wiley & Sons, New York, 1975, p. 222, 223. Usually the instantaneous scalar potential contributes only to the near field.

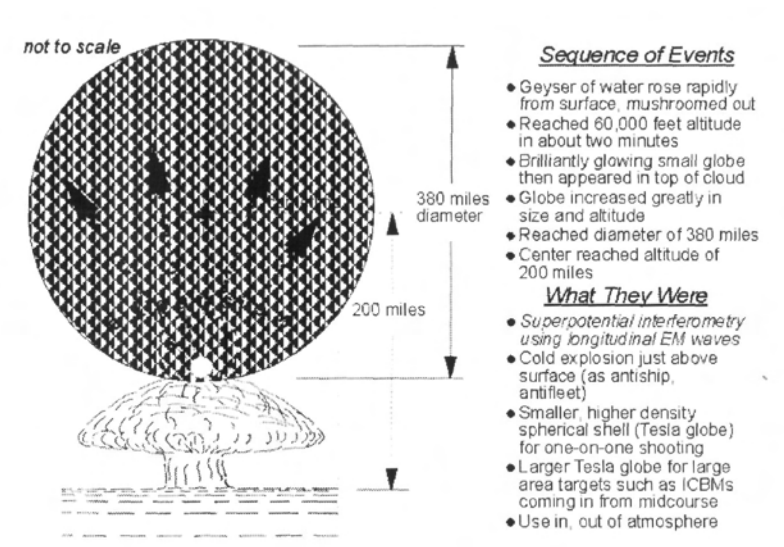

Sequence of Events

- Geyser of water rose rapidly from surface, mushroomed out
- Reached 60,000 feet altitude in about two minutes
- Brilliantly glowing small globe then appeared in top of cloud
- Globe increased greatly in size and altitude
- Reached diameter of 380 miles
- Center reached altitude of 200 miles

What They Were

- *Superpotential interferometry using longitudinal EM waves*
- Cold explosion just above surface (as antiship, antifleet)
- Smaller, higher density spherical shell (Tesla globe) for one-on-one shooting
- Larger Tesla globe for large area targets such as ICBMs coming in from midcourse
- Use in, out of atmosphere

Figure 35 Exothermic (cold) explosion off the Kurils 9 April 1984. A second test, an expanding globe of light, was added in the exothermic mode to demonstrate rapid serial switching between weapon modes.

Related Phenomena

In 1969 in Toronto, Canada, Sid Hurwich demonstrated a device capable of "freezing" a pistol lying on a table so that the gun could physically not be moved and the trigger could not be pulled {71}. For this exhibition of *strong electrogravitation,* general relativistic fields initiated by the EM force field (including the usually unaccounted Heaviside nondiverged component that Lorentz arbitrarily discarded) as the agent of spacetime curvature were used {72}. The strength of the effect was attested by the fact that the hands of the observers' watches all showed "no passage of time" during the half hour they were in the room with Hurwich's device operating.[17]

Rapid development of that strong electrogravitational technology and concomitant energetics then proceeded in another nation. Weapons of such nature were then employed by that nation a few years later, in two well-known international incidents. Clear signatures of the nature of the weaponry utilized were present in those incidents. Details of these incidents will not be included in this paper.

---

[17] We point out that this is highly interesting itself, since it showed that the observers were operating in a different frame from that of the watches and the gun, which were "frozen". We also point out that recently light itself has been "stopped" in quantum experiments, yielding something quite similar.

Since the energetics weaponry already utilized special EM "hidden variables," Bohm's hidden variable theory and quantum potential were directly adapted into the program to dramatically extend it into the development of quantum potential weapons.

Stalin *Personally* Initiated the Soviet Energetics Program with the Greatest Urgency

While Stalin was at the Potsdam conference, at the close of WWII, the U.S. exploded the first atomic bomb in the desert. Truman informed Stalin in general terms that a powerful new U.S. weapon had been successfully tested. Secretly, Stalin already knew of our atomic bomb progress, for he had a spy in the program. While he was still at Potsdam, he then secretly called the Director of the Soviet atomic weapons program and directed a speed-up.

After we exploded two atomic bombs on Japan, the Japanese surrendered and WWII was over. See again Figure 1. Stalin vehemently ordered his Academy of Sciences to produce the next great technical breakthrough at all possible speed, since the American atomic bomb had frustrated his plan to wait two years after the war while the West speedily disarmed, then march into Europe and take it in six weeks. Now his massed troops could just be bombed back into the Stone Age by U.S. bombers using atomic bombs.

A Great Technical Breakthrough in Physics Can Always be Had in Four Years

The scientific community can have a great new technical breakthrough whenever its furious and dogmatic resistance to innovative discoveries or the possibility of innovative developments can be suspended. All that is necessary is a concentrated group of highly skilled, open-minded scientists, the money and facillities, and a strong headlock placed on the powerful but dogmatic scientific groups who would otherwise hamstring the program and destroy it.[18]

---

[18] Present day examples are (i) the cold fusion fiasco, where the U.S. scientific community has revealed furious determination to ignore more than 600 successful experiments worldwide, (ii) a similar fury against any and all research into extracting EM energy from the vacuum, even though all EM energy comes directly from the active vacuum via the proven asymmetry of source charges and dipoles, (iii) violent suppression of proven "extended electrodynamics" methods of rapidly curing severe diseases, even in terminal stage (resistance to anything except "cut, drug, burn" and anything that would interfere with the monstrous profits of the large pharmaceuticals), (iv) continuing dedication to an electrical engineering model that does not even model or apply what powers an electrical circuit, (v) continuing resistance to quantum potential research and development, (vi) slavish dedication to "big nuclear science", massive accelerators, dangerous nuclear power plants, and a "hot fusion WPA program", even though much higher energy physics than these programs advocate and use can be and is being done on the tabletop, using time-energy instead of spatial energy, and (vii) increasing "lock-up" of all research funds into the organized hierarchy's politically approved agenda. The community has deviated so far from scientific

We proved it in the WWII Manhattan Project, where we went from nearly zilch to the operational atomic bomb in 4 years. But we did it only by secretly bypassing the conventional U.S. scientific establishment. Einstein wrote to President Roosevelt *around* the scientific community, and when the President approved and ordered the highly classified program, conventional scientists not on the project were just parked on the sideline and ignored so that the job could be completed. The naysayers were also muzzled because of the extreme classification of the project; they faced jail if they engaged in their usual ad hominem attacks and dogmatic eloquence.

Further, today the "Skunk Works" approach repeatedly proves that bold new projects can be done, when the skilled scientists and facilities are mobilized and no adamant opposition by the conventional community is tolerated.

Numerous examples of the U.S. scientific community's suppression of innovation and new technology can easily be furnished. Such are well known to any historian of science. Cold fusion, the amorphous semiconductor by Ovshinsky, the first laser papers, ultrawideband radar, Mayer's original conservation of energy law, Boltzmann's work, Wegener's continental drift, flight of heavier-than-air machines, the rocket as a great strategic weapon, and hundreds of other examples come readily to mind. Such suppression has been *the* major characteristic of the organized scientific community since its inception, as pointed out by numerous leading scientists and authors. As Planck stated, usually one gets a new physics only after the old physicists who so adamantly oppose it just die off and get out of the way.

Today, the U.S. scientific community is in fact our "front line troops," although—except for a little lip service—the managers of the organized scientific community are blissfully unaware of it and have little or no sense of mission in out-of-the-box areas with respect to national survival. It is ironical, e.g., that our orthodox scientific community does not even know what powers an EM circuit—energy extracted directly from the vacuum by the asymmetry of the source charge (considered in its modern view as a dipolarity) and by the source

---

method in defending archaic theory in the face of experiments that falsify it, that the ubiquitous violation of the law of increasing entropy (second law of thermodynamics) by electrodynamics itself continues to be adamantly resisted using ad hominem and inflammatory attacks, slander and libel, and destruction of scientific careers of researchers. We flatly are not getting our taxpayer money's worth from our present scientific community, and—barring a Stalin-like shakeup from top to bottom—we are not going to get our money's worth in the future unless the community is forced to free up funds for young graduate students and postdoctoral scientists wishing to work in "out of the box" scientific areas. As Max Planck stated, *"An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning."* In other words, one doesn't get a scientific innovation until the old dogs so bitterly opposing it just die off and get out of the way. This long-known and stated characteristic of the scientific community—its fierce and unrelenting resistance to change and innovation—is presently the greatest threat to the survival of this nation.

dipole (Figure 36).

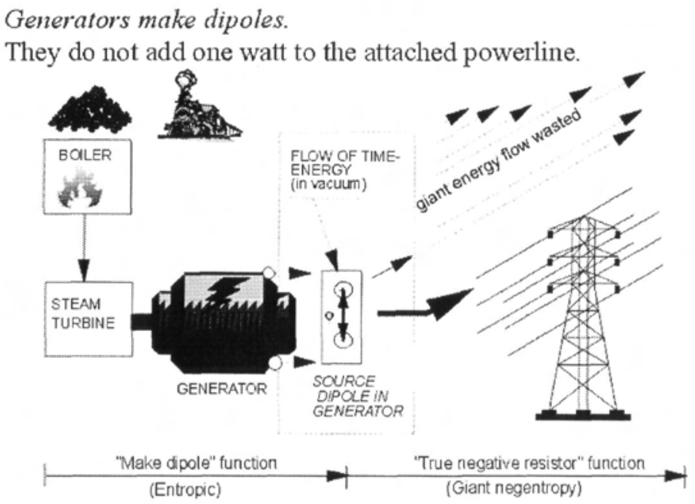
Generators make dipoles.
They do not add one watt to the attached powerline.
BOILER
STEAM TURBINE
GENERATOR
FLOW OF TIME-ENERGY (in vacuum)
SOURCE DIPOLE IN GENERATOR
giant energy flow wasted
"Make dipole" function
(Entropic)
"True negative resistor" function
(Giant negentropy)

Figure 36 Energy extracted from the vacuum by the asymmetry of the source dipole powers every electric power line and electric circuit.

A Typical Example: The Suppression of Cold Fusion

As an example, much of U.S. Science—dominated by nuclear factions since the advent of the atomic bomb in WWII—vehemently attacked cold fusion and the cold fusion researchers, including wholesale ad hominem attacks, with a viciousness seldom seen in the history of science. Many dogmatic scientists declared the entire field at best a mistake and at worst a fraud. Even the character of many of the leading cold fusion scientists was openly impugned and the researchers were slandered and libeled. All this was done by conventional dogmatic scientists who have not solved the fundamental source charge problem {10,12}, have no inkling that all EM energy in 3-space comes from the time-domain {10}, and have not a clue as to what really powers every EM circuit. Since they also do not know where the electrical energy driving the chemistry actually comes from, they have no knowledge of the effects of time-energy in electrolyte solutions.

They also appear to have little or no knowledge of the implications of Evans' and Searles' *transient fluctuation theorem*[19] in thermodynamics showing to what degree the second law of thermodynamics can be violated and reactions can run in reverse, nor its extension to the micron (colloidal) size level and to two seconds in duration.[20] If they did, then they would realize that the Coulomb barrier provided by *repulsion* of like charges (such as two H+ ions), often reverses into the *attraction* of like charges, with the barrier changing into the Coulomb attractor. This can last (by experimental proof shown by Evans *et al.)* up to two seconds, in a volume of fluid up to a micron in diameter. Hence it is not "against the laws of physics and thermodynamics" to observe attraction of two protons in an electrolyte so closely that each enters the strong force region of the other and the two form a quasi-nucleus. In the decay of this reversal zone, the preferred decay of the quasi-nucleus can be by quark flipping in one of the protons, yielding an atom of deuterium emerging from the decayed time reversal zone. Such results have been meticulously observed in many cold fusion experiments in multiple labs and nations, and by multiple researchers. Since these and other novel nuclear transmutation results are *replicated* experiments (more than 600, in multiple labs, multiple nations, and by multiple researchers), it means that no amount of theory can refute them or the results. The harsh skeptics launching the frenzied ad hominem attacks are thus revealed as the unscientific bigots they are, since they ignore substantial experimental results that contradict the prevailing theory. The excess energy production and inexplicable new nuclides (Figures 21, 37, 38) obtained by these multiple experimenters are completely inexplicable by ordinary nuclear interactions in the accepted theory.

---

[19] D.J. Evans and D. J. Searles, *Phys. Rev. E,* Vol. *50,* 1994. p. 1645.

- Fennions time-reverse in even numbers
- Like charges cluster
- Aqueous solutions contain H+ ions,
  which are just protons
- Quarks flip as TR zone fades

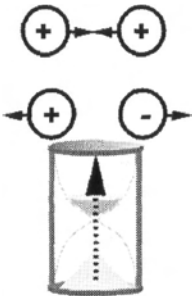

$$({}_1H_1 + {}_1H_1) \rightarrow ({}_0n_1 + {}_1H_1) = {}_1H_2 \text{ deuterium}$$

Figure 37 Electronuclear reaction in a time-reversed zone producing excess deuterium ions.

---

[20] G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601.

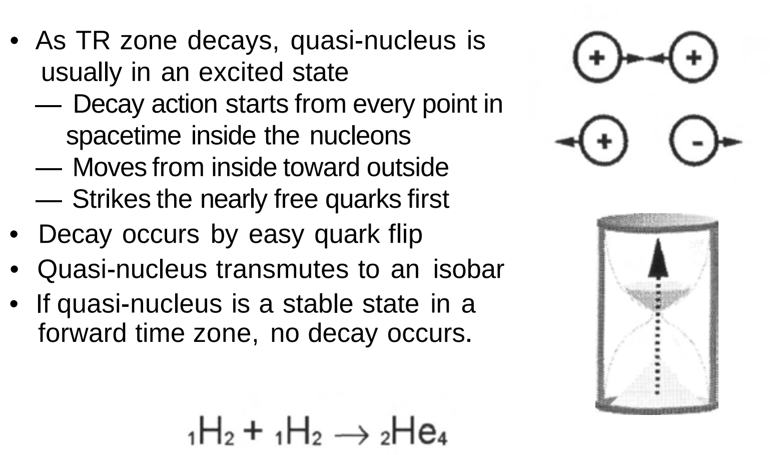

- As TR zone decays, quasi-nucleus is usually in an excited state
  - — Decay action starts from every point in spacetime inside the nucleons
  - — Moves from inside toward outside
  - — Strikes the nearly free quarks first
- Decay occurs by easy quark flip
- Quasi-nucleus transmutes to an isobar
- If quasi-nucleus is a stable state in a forward time zone, no decay occurs.

$${}_1H_2 + {}_1H_2 \rightarrow {}_2He_4$$

Figure 38 Electronuclear reaction in a time-reversed zone producing excess alpha particles.

Meanwhile, the U.S. Patent Office has been instructed not to accept any patent applications that would interfere with present nuclear planning. Positive cold fusion results have also been successfully obtained in several U.S. government laboratories, including at China Lake where very strange instrumental effects have the experimental team puzzled. I proposed an explanation of those anomalies in August 1998, and have included it in my book just published in September 2002 {73}.

Revolutionary New Processes Are Hidden in the Cold Fusion Experimental Results

Once understood, cold fusion holds the breathtaking key to systems that consume *a tiny fraction of time flow itself as* a fuel to produce copious heat energy for the world's needs. Unknown to the conventional scientists, of course, electrodynamics itself already does this, but it has just not been recognized and accepted. Transducing one microsecond per second of time into transverse wave EM energy will produce some $9x10^{10}$ watts of real electrical power, if all the transduced EM energy is intercepted and used. This is the *real* way to extract energy from the vacuum and use it cleanly and cheaply.

To grasp such a notion, recall that the fundamental units used in physics are completely arbitrary. It is perfectly possible to express all physics in only a **single** fundamental unit—for example, *the joule*. Time then becomes totally a function of energy. So it is not so insane after all to express time as highly compressed energy. Mass is also highly compressed energy. That concept is now familiar after Einstein, the nuclear age, and $E = mc^2$. Time-as-energy is just as straightforward as mass-as-energy, but it has just not yet come into recognition.

The *use* of time as compressed energy, however, leads to some profoundly new types of electronuclear reactions presently unknown in particle physics. These new transduction reactions produce the new nuclides in cold fusion experiments, and generate the electronuclear interaction. We explain that in some detail in our new book, *Energy from the Vacuum,* previously cited {10}.

When interpreted as using time-energy, cold fusion shows us the way to a new electronuclear regime of nuclear reactions where nuclei, elements, and compounds can just be directly assembled by comparatively weak (in terms of *spatial* energy!) electromagnetic interactions. Presently we have made substantial progress in understanding the fundamental mechanisms generating these electronuclear reactions, after presenting preliminary results in August 1998.

As one example, time-reversing a single fermion (a particle with spin 1/2, 3/2, etc.) is precluded by the Pauli exclusion principle. But one can time reverse *two* of them, or any even number n, where they act as a semi-boson or group of semi-bosons. Simply put, if a time-reversal zone is established momentarily in the region where like charges exist, then the n-tuple fermion cluster can be achieved because the law of attraction and repulsion of charges is temporarily reversed {74,75}. This unexpected "attraction of paired fermions of like sign" will form quasi-nuclei in many cases, particularly in small clusters of two or four, etc.

Hence two H+ hydrogen ions (two protons) in such a time-reversal zone in a fluid are drawn together into a quasi-nucleus, which actually is an excited isomer of deuterium (Figure 37). Time-reversal zones are temporary in cold fusion experiments due to the statistics of their formation and the transient fluctuation theorem {76} of thermodynamics. As the time reversal zone decays, the quasi-nucleus of two protons is increasingly in an excited and unbalanced state. The excited-state *diproton* nucleus decays by flipping a quark in one proton, converting it to a neutron. That produces a *neutron + proton nucleus*, which is a deuterium ion, strictly by electronuclear means. Or four protons may cluster by "attraction in a time-reversal zone" and form a 4-p nucleus. That is an excited isomer of a helium nucleus. As the time-reversal zone decays, first one quark flips in one proton and then another quark flips in a second proton. That produces a helium nucleus - i.e., an *alpha particle*. Or two deuterium ions can attract and form a quasi-nucleus (Figure 38), which is already an exact isomer of a helium nucleus and thus easily decays into an alpha particle by simply "tightening" as the weakened strong force recovers faster than the Coulomb force can reduce, vanish, and resume in the opposite direction. The exact reaction equations for these reactions are given in my new book, *Energy from the Vacuum* {10}.

We filed an Invention Disclosure on the process of forming and using time-reversal zones to cause like charges to attract and unlike charges to repel *in vacuo,* in fluids, and in other media, resulting in the formation of semi-nuclei as excited isomer states of other nuclei, followed by decay into new nuclides. We will follow it with patent applications in the future. This represents a completely new kind of particle physics interaction presently not included in the textbooks, and not requiring high-energy processes. Literally thousands of new nuclear interactions—all *permissibly* violating the second law of thermodynamics because the statistics permit such violations and experiments prove it—are now possible.

Cold fusion does indeed hold a new kind of nuclear physics in its hand—but a physics presently considered very alien by our orthodox nuclear science community. For decades nuclear scientists have been seeking ever more *spatially* energetic particle smashing, believing that extreme high energy nuclear reactions and transmutations cannot possibly be done with charged particles at low *spatial* energy. One is sympathetic to the scientists, because even the concept of "low spatial energy transmutation" itself sounds like an oxymoron. And it would be, *if one were limited to spatial energy alone*, as is presently assumed by physicists. But one is also speaking of the neglected and compressed *huge time-energy* component of those low-spatial energy photons, and that is quite a different matter. The lower the spatial energy of the photon (the lower its frequency), the greater the amount of compressed energy transported in the time component as "time". The transposition of time-polarized EM energy into longitudinally polarized EM energy—and vice versa—is already justified in quantum field theory. Use of that extremely high energy component by its transposition or partial transposition into spatial energy in fact provides a much higher energy physics interaction than present orthodox high energy physics utilizes, knows, or considers.

Almost all the learned scientists are thinking *only* of spatial energy, since that is the only kind they have been "taught" to consider. Hence, from their view (i.e., on the premise of their implicit *assumption)* it is perfectly "logical" to condemn cold fusion, since in their view, "cold" implies low (spatial) energy and "fusion" implies high (spatial) energy transmutation. What is unknown to the orthodox community is that the "cold" in cold fusion also implies "extraordinarily high time-energy". So with a little transduction of highly compressed time-energy into spatial energy, "cold" fusion actually can imply "extremely high spatial energy made available by transduction" also.

As an example, the Chief Editor of *Science* recently had the courage {77} to publish results of sonoluminescence research reporting positive experimental transmutation results {78}. A paper was then published in *Nature* {79}, pointing out that the (spatial) energy involved is insufficient for such nuclear transmutation temperature, thereby implying that the reactions reported by the first researchers are not possible.

However, given formation of a time-reversal zone, this second paper is not necessarily applicable to the sonoluminescence fusion process, because the paper is based on an unwitting two-part premise that *(i) only spatial energy can be used for nuclear transmutation and (ii) no time-reversal zones exist to nullify that premise by having reactions (such as the law of attraction and repulsion of charges) run backwards.* That premise is already experimentally proven to be invalid by Evans et al., and it fails when the time reversal zone effect is considered.

We strongly stress that thermodynamics—both in theory and as demonstrated in experiments—not only allows but predicts—for a certain fraction of the time—the formation and existence of just such "reversal zones" where the usual reactions *run in reverse* {76}. It is also experimentally proven {54}. In that case a much higher energy—time energy—can be and is employed to convert the Coulomb barrier into the Coulomb attractor.[21] We know now (from the experimental demonstrations by Denis Evans et al.) that such "reversal zones" may be up to a micron in size and may endure for up to two seconds. The subsequent decay of that TRZ then allows the quasi-nuclei temporarily formed in it to decay by quark flipping alone, allowing the quasi-nucleus to easily transmute into a perfectly well known nuclear reaction product.

To quote Arthur C. Clarke, when a learned scientist says something is *possible*, he is probably right. When he says something is *impossible*, he is probably wrong.

Stalin's Dictatorial Impact on Soviet Scientific Management's Response

Stalin was a powerful dictator. He simply killed or jailed those who opposed his will. So after his blunt order (Figure 1) circa latter 1945 to find a great new technical breakthrough area quickly, his Academicians and high level science managers were faced with a severe dilemma: They could not afford to *oppose or criticize a vigorous search* for new innovations and a great breakthrough, and they could not remain passive. Their very lives depended on they themselves *personally finding such a breakthrough in great haste and presenting it to Stalin*. They had to actively spur the efforts of their most outstanding subordinates and Soviet science itself, or suffer dire personal consequences. Siberia and prison camps loomed for the lot of them unless they produced.

To put it bluntly, Stalin "got their collective attention" and their "whole-hearted support." Those few who resisted simply were shipped post-haste to the notorious prisons of Siberia, and worked to death in short order.

---

[21] We strongly point out that, when the reactions run backward and negentropy appears in those reversed regions, the conservation of energy law is violated unless time-energy is being converted to spatial energy. Either we must begin accounting time as energy, or we must give up the conservation of energy law itself. The obvious solution is to account the time energy being utilized, as well as the 3-space energy.

So the highest-level Soviet scientific managers now desperately demanded of their subordinate scientists *anything*—*anything at all*—that was novel and had breakthrough implications. An almost convulsive national scientific effort was launched immediately, with hardly a detracting voice. A great team of the very best Soviet scientists was rapidly formed and methodically searched the scientific literature of the world from beginning to end. Nothing even remotely like that search and review has ever occurred in the West.

For example, one great technical center set up by the Soviets had a staff of about 2000 Ph.D.s with associated translators, staffs, etc. The Soviets purchased and hauled in copies of all the scientific literature of the West—by the shiploads—to such centers. Their scientists then methodically went through the material, journal by journal, paper by paper, page by page. There are and always have been myriads of innovations and discoveries sitting on the shelf in the scientific literature, never followed up, or opposed by standard models and credos. The first tier of reviewing Soviet scientists was composed of excellent scientists. They skimmed out the "cream" into a select pile which the second tier then analyzed in depth. The second tier was composed of the best nonlinear scientists in the world—the "superscientists."

They were seeking *new principles of nature that could rapidly be developed into highly advanced technology*, and they found them—sitting there in the literature, already pointed out, and long ignored by the Western scientific community.

Shortly Stalin had his great new breakthrough area emerging, and they had their heads.

Our Own Science Community's Progress During These Decades

For decades foundations physicists (Bunge, Feynman, Wheeler, Lindsey, Margenau, etc.) in the U.S. have repeatedly pointed out the substantial foundations problems in electrodynamics. Nobelist Feynman and the great John Wheeler tried to "fix" the problem whereby electrodynamics still erroneously assumes force fields in a space filled with a *material ether* (unit point positive charge, unit point north pole, and unit point mass assumed at every point in space). Unfortunately they failed, primarily because their absorber theory {80} contradicted the photon hypothesis and hence quantum mechanics.

Meanwhile, our own National Science Foundation and National Academy of Sciences has directed only minimal work on correcting electrodynamics and dramatically extending it. Faraday and Maxwell—as did every other scientist at the time—rigorously assumed a material ether. Maxwell's equations also implicitly included that material ether assumption. They still do, even after Heaviside's truncation and Lorentz's symmetrical regauging, because none of the Maxwell-Heaviside equations were changed to eliminate the material fields in space (material ether) assumption after the Michelson-Morley experiments falsified it.

Every major university in America still teaches that implicit material ether, even though it was falsified over a century ago. That is an exact measure of the caliber of "foundations" work our own scientific community does if left alone. *In more than 100 years, they have not even chosen to root out the material ether from electrodynamics.*

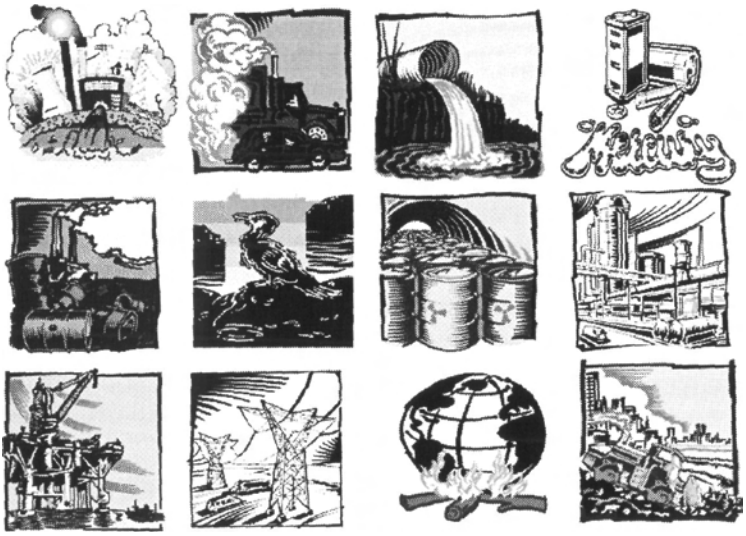

Figure 39 This polluted and suffering planet. Direct result of the scientific community's adamant resistance to upgrading classical electrodynamics and electrical engineering.

The prevailing flippant attitude is that, "Well, your TV set works, doesn't it?" Yes it does, but we also do not have practical electrogravitation and inertial propulsion, nor do we have cheap clean EM energy from the vacuum. Instead, we have a severely polluted planet (Figure 39), ever increasing deaths from biospheric fumes and polluting byproducts, increasing destruction of species, and a monstrous unleashing of far more coal and hydrocarbon burning in our near future while the world tears itself apart and destroys its economy and civilization in a rising series of oil and energy wars. The direct responsibility for that mess lies directly with the scientific community and its fierce resistance to innovative research and development, and changing a century-old, long outmoded classical EM and electrical engineering model.

Wryly put, it's high time we "got the lead out" of EM field theory, where the lead is the material ether and other non-sequiturs cluttering the theory.

Such drastic modeling errors would not be tolerated of government-funded aerospace engineers and contractors developing a new missile system. So it should be equally intolerable in the government-funded scientific community and in fact in the civilian scientific community as well, and it should not be tolerated of either the National Academy of Sciences or the National Science Foundation.

Einstein {81} said it perfectly, as follows:

> *"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted... in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."*

Upholding and perpetuating flawed electrodynamics models seriously impedes the progress of scientific development and revolutionary new technology.[22,23] At some point, one must legitimately question the upper management of the U.S. scientific community for not modifying classical EM and electrical engineering to include the active vacuum exchange with every charge and EM system. In not doing so, they continue to advocate perpetual motion source charges, creating energy from nothing, on a scale unparalleled in history.

It is little wonder that we do not have electrical power systems that successfully and freely utilize more than a trillionth of the vacuum energy they freely extract from the local vacuum. *There is not now and there never has been a single*

---

[22] Oddly, the conventional electrodynamics model assumes that every source charge freely creates energy out of nothing at all, and pours this energy out continuously in all directions, thereby establishing its associated EM fields and potentials at light speed, reaching out across the universe. From the 1960s alone, the total assumption of incredibly vast numbers of perpetual motion machines has been carefully hidden from the new students, so that today even many professors do not realize that in their model all EM fields and potentials, and every joule of EM energy in the universe, are assumed to be created by a vast assortment of perpetual motion machines called source charges. That problem has not been solved in electrical engineering or in classical electrodynamics, and if the scientists have their way, it is not going to be solved. Yet the basis for its solution—the asymmetry of opposite charges such as the dipolarity (in the modern view) of a source charge—has been in particle physics since 1957, with the award of the Nobel Prize to Lee and Yang in December of that year.

[22] One keeps one's sense of humor. The raucous skeptics and "defenders of the orthodox EM faith" who so noisily denounce legitimate COP>1.0 EM system researchers as "perpetual motion nuts", are themselves unwittingly the greatest perpetual motion nuts ever recorded in history. Further, they are so ignorant of what their present model actually means, that they do not even know it. It is one thing to be ignorant and to recognize it; it is quite another thing to be ignorant and so ignorant one does not know it—45 years after broken symmetry was rigorously established experimentally by Wu and her colleagues, and the Nobel Prize awarded to Lee and Yang.

electrical engineering department, professor, or textbook that knows what actually powers an EM circuit—real observable EM energy extracted from the local seething vacuum by the source charges and the source dipolarity.

Since our scientific community cannot seem to make such a simple change (as including the active vacuum's interaction with every charge and the charge's broken symmetry in that interaction) even 45 years after the basis for that change has been resoundingly proven in particle physics, our national survival is at stake and the nation itself is seriously threatened by the superweapons developed by the KGB over the last five decades. We would have already been totally destroyed—twice in 1997 alone—had it not been for the intervention of a friendly little foreign nation. My colleagues and I played a very small alerting role in those affairs, so we are speaking from experience, not from extrapolation. We would be quickly destroyed today, were it not for the continuing "balance of terror" provided by the little nation.

U.S. Scientists Have Committed Monstrous Crimes and Gone Unpunished

Rather sadly, we also point out another thing that has to be said: *scientists are no more moral or immoral than any other slice of American citizens.* A percentage of scientists is capable of committing horrible scientific crimes, as distasteful as such a statement may be. That includes the deliberate experimentation on humans, resulting in deaths. In short, at least manslaughter and perhaps even murder, since knowledge of the probable results was there beforehand.

One would think U.S. scientists would *never* scientifically condone and perform experiments leading to mass deaths for 30 years (i) as they did in the Tuskegee case in Alabama. Over 400 poor blacks were supposedly treated for syphilis, while actually not being treated at all, but were deliberately tricked and allowed to rot with syphilis so our "ethical" scientists could watch how their brains decayed, so to speak. Over 100 of those poor blacks died with the disease while the scientists cold-bloodedly watched them expire and kept meticulous records. Or (ii) as they did in the case of whole body nuclear radiation for 30 years—again by trickery, deceit, and illegal use of security classification to hide these acts—of helpless retarded children and others, some of whom also died.

Once these heinous incidents were exposed publicly, the political power of the scientific community is such that *no scientist was ever indicted or tried for these heinous mass murders, and none is going to be.* The scientific community *has not* fiercely called for punishment of these mass murderers amongst their midst—guilty of precisely the same kind of horrible actions for which we tried and executed war criminals at Nuremberg.

Incredibly, the President of the United States formally apologized for these incidents! He did not turn the case over to the Justice Department for prosecution of high crimes against humanity and mass murder, *as would have* *been done for ordinary citizens.* Had a single doctor in his private clinic done such things on his own, he would have been arrested, summarily tried, and either sentenced to life in prison or executed.

The sad thing is that there was really no sharp clamor from the scientific community to bring their fellow scientists to justice, a la Nuremberg. By their silence, the organized scientific community showed a side that scientists are loftily not supposed to possess. *So long as the scientific community continues to tolerate those mass murders of American citizens and not call for justice, then just so long do they stand convicted by their own resounding silence and inaction.*

We cite these painful incidents (there are others!) to point out that one must not deify the scientific community. The conduct of science is no better and no worse than the conduct of human affairs in other areas. Science is also a big business, and it has its moral scientists and its immoral scientists. It does its moral business and it does its immoral business.

We painfully submit that both sides of the scientific character must be considered, when one examines such things as the tolerance or intolerance of the scientific community to dramatic innovation and "out of the box" research. We have to consider it now, because it has become a serious threat to the continued survival of our nation. The conduct of scientific affairs will unavoidably include suppression, maneuvering, spin control, and fierce struggle for power and funds—such is in the very nature of human beings, and it is present in the scientific community on a bell-shaped curve as in all other human communities. It is ongoing widely today, and to an uncomfortably large extent much of science has been captured and effectively controlled by controlling the scientific funding and prescribing what research must be done for the funds available.

As the French say, "The more things change, the more they remain the same." Science is still the same way as it was when Planck made his famous statement. It will still be that way tomorrow, and a century from now. As with any other big enterprise, it has to have checks and balances imposed and enforced on it from outside—by the government itself. Little or no indication exists of that happening in the near future, until we get something like an anthrax spray attack on a large city, and have a million or more Americans dying in the streets.

Energetics: New Principles of Physics and How the Soviets Found Them

Under Stalin's whip, the Soviets particularly sought anything that indicated *possible new principles of nature.* It is no accident that Nobelist Petr Kapitsa later informed Nikita Khrushchev that a means of total neutralization of foreign missiles could only come from *a group of new principles in physics called energetics.* A group of new principles is precisely what energetics *is.* So if our scientific and intelligence communities wish to understand energetics, they must first understand the new principles it is based upon. They cannot just simply apply ordinary electrodynamics and "business as usual" as they have largely done. The massive Soviet scientific search program spurred by Stalin is where those new principles came from.

When the first tier of Soviet review scientists found something that had not been followed up, and saw that it had significant implications if found valid, that paper was laid aside into a "select" pile. That select pile then went to the "super" scientists, for full-blown review and analysis by some of the brightest and sharpest minds on the planet.

The Russians came up almost immediately with Whittaker's 1903 and 1904 work. That work shows a hidden electrodynamic *longitudinal* EM wavepair structure (in phase conjugate pairs) inside the scalar potential and comprising it.

This is a dramatic extension of electrodynamics to incorporate a full-infolded general relativity infolded inside the electrodynamics. Whittaker showed that scalar potential functions in interference already *create* all the "normal" fields and waves anyway (the beginning of superpotential theory). This novel interferometry and its inherent capability for action-at-a-distance did not elude the Russian superscientists. The superb Soviet scientific team came up with that and with much, much more.

Nothing like the comprehensive, intense review of the entire scientific literature has been accomplished by the U.S. scientific community. Nothing like it is being done now.

Longitudinal EM Wave Interferometry Energetics Was Almost Immediately Uncovered

At the end of WW II, the German radar team and infrared team were taken to Russia. Both teams were years ahead of their Western counterparts.

One of my close colleagues has validated that the German radar team had some severe instrumental anomalies while still in Germany. We think this is what they were:

In WW II the Germans invented radar absorbing materials, to coat their submarine snorkels and prevent Allied antisubmarine warfare bombers from tracking the snorkels and destroying the subs. A wide variety of highly nonlinear, doped materials were investigated. In addition to single beam radar cross section, the Germans also examined the materials in multi-beam illumination circumstances.

It appears that, in some multibeam experiments with some materials, conditions were just right so that the multi-illuminated absorbing material acted as a pumped phase conjugate mirror (PCM), and wave-to-wave interactions emerged. This meant that the absorber material PCM would suddenly gather up all the energy in the pumping beams, use one of the other beams as a signal input beam, and fire an amplified retroreflected pulse right back down the signal input beam, wrecking the experiment and creating havoc in surrounding equipment.

When the German radar team went to Russia, their knowledge of these anomalies went with them. In the desperate Russian scientific climate to find *lust such anomalies,* this startling and anomalous phenomenon would have been exhaustingly examined and studied, once the German radar engineers got to Russia to their new laboratories.

In short, it appears that the Russians quickly came up with what today is called "nonlinear phase conjugate optics," but first in the radar band rather than optics.[24] So before 1950, it appears that the Russians already had working phase conjugating pumped phase conjugate mirrors for the radar bands and in other bands.

Fisher {82} details that we Americans did not "stumble onto" practical phase conjugate optics experiments until Russian physicists briefed Lawrence Livermore National Laboratory scientists in 1972 about the strange wave in **many** of their optical experiments that "came out of nowhere and restored order **from** disorder." At that time, the KGB energetics weapon scientists had known about phase conjugate retroreflection and its amplified pumping, for more than 20 years. Indeed, the first giant strategic scalar interferometers had already been deployed and utilized to kill the U.S.S. Thresher nearly a decade earlier (Figure **6).**

By 1950 Russian Energetics Was Off and Running

By the beginning of the 1950s, the Russian scientists under iron KGB control were already experimenting with "scalar potential interferometry" at a distance, a la Whittaker 1904. Here transverse EM waves, EM fields, and EM energy—either *divergent* (heating) positive energy or *convergent* (cooling) negative enery—could be readily created at a distance {83} (Figures 40 and 41). Further, the longitudinal EM wavepairs used in this interferometry had the marvelous characteristic of easily penetrating right through the earth and ocean, with only a little intervening interaction, given that the LWs are sufficiently well made {84}. The interferometry "beams" can and do also move superluminally—again, right through the earth and ocean.

---

[24] **Of course, they would also have investigated it at all other bands as well.**

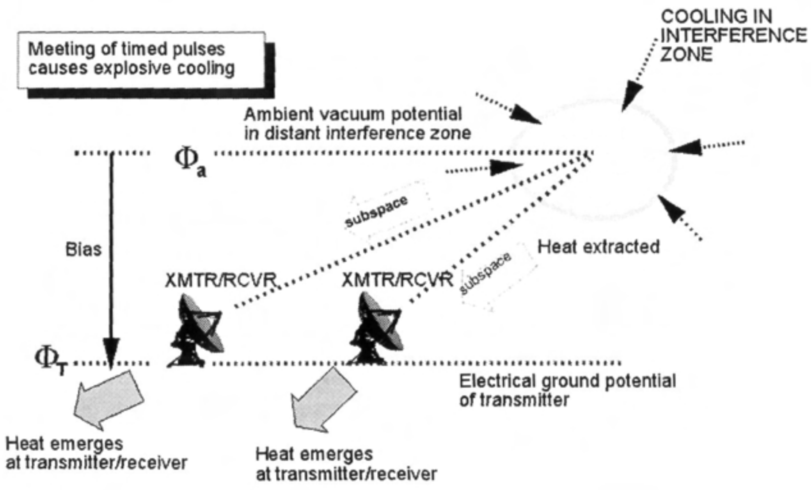
Meeting of timed pulses causes explosive cooling
COOLING IN INTERFERENCE ZONE
Ambient vacuum potential in distant interference zone
Φ_a
subspace
Bias
Heat extracted
XMTR/RCVR
XMTR/RCVR
subspace
Φ_T
Electrical ground potential of transmitter
Heat emerges at transmitter/receiver
Heat emerges at transmitter/receiver

Figure 40 Scalar interferometry in the endothermic (cooling) mode.

It is not accidental that Admiral Gorshkov, head of the Soviet Navy, would later bluntly state: *"We have made the oceans of the world transparent."* And so they had.

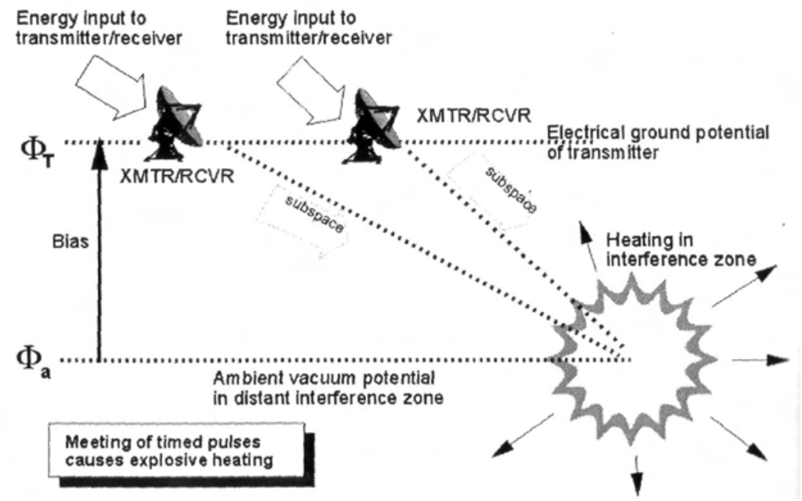
Energy input to transmitter/receiver
Energy input to transmitter/receiver
XMTR/RCVR
Φ_T
Electrical ground potential of transmitter
XMTR/RCVR
subspace
subspace
Bias
Heating in interference zone
Φ_a
Ambient vacuum potential in distant interference zone
Meeting of timed pulses causes explosive heating

Figure 41 Scalar interferometry in the exothermic (heating) mode.

By 1950, the Russian scientists were experimenting with deliberately creating and phase conjugating longitudinal EM waves, forming phase conjugate wavepairs. They then assembled the desired (mathematically calculated) sets of those waves, adjusted frequencies and magnitudes, phasing, etc., to produce EM potentials having a desired deterministic internal spacetime curvature set (internal spacetime engines) and dynamics.

As a result, by the mid-50s Russian weapon scientists were already experimenting in *bioenergetics* with the direct inductance of diseases and health changes via the transmission into human targets of such "internally structured" EM potentials, fields, and waves. They were also experimenting with the production of EM energy, potentials, fields, and waves at a distance, directly through the earth and the ocean. Either heating (energy flow divergence) (Figure 41) or cooling (energy flow reconvergence) (Figure 40) could be produced at will in the distant interference zone, merely by altering the bias voltage on the chassis grounds of the interferometer's transmitters. Rapid switching between the exothermic and endothermic modes thus was straightforward{85}, as demonstrated (Figure 35).

By the latter '50s, longitudinal wave interferometry weapons had already emerged from advanced engineering development and were in full production engineering toward deployment. The LIDA device {86} and even more advanced versions for directly affecting the human mind and its state of consciousness had emerged in *psychoenergetics* as one class of devices rather openly released and utilized. The LIDA uses structured EM microwaves {87} to induce a hypnogogic or cataleptic type of trance-like state in either a human or a cat. as tested years later by Dr. Ross Adey of the U.S. Early prototypes were purportedly used against our prisoners in North Korea during the Korean War to assist in their brainwashing by their North Korean Captors. Prisoners reported that its effects were irresistible.

Meanwhile, energetics weapon testing in several major modes was being conducted, particularly from Soviet research ships over remote regions of the ocean and at night. Descriptions of such tests—many of them seen by passing ship's captains—can readily be found in the open maritime literature. We gathered and published many such incidents and cited references.

Deployment of the Soviet Energetics Weapons and Their Quick Use

In January 1960 Khrushchev spoke obliquely of these "fantastic" new superweapons, just within the portfolio of the Soviet scientists, so to speak, in an important speech to the Presidium. Parts of that speech were reported in the *New York Times* {88}. No one in the West knew what Khrushchev was talking about. Our intelligence community, which bases many of its assessments on the tired old phrase "We have no evidence that..." simply treated it as propaganda, It wasn't, and the U.S.S. Thresher was to be destroyed (Figure 6) three years later by that same weapon Khrushchev was speaking of.

Khrushchev began rapidly inserting long-range missiles and nuclear warheads into Cuba in 1962, precipitating the Cuban Missile Crisis. Kennedy knew Khrushchev's strategic bombers and rockets were in horrible shape, due to direct information received from Colonel Oleg Penkovskiy, a Russian spy in our employ courtesy of British Intelligence. No one knew of the superweapons, but they were not yet operationally ready in Oct.-Nov. 1962. So Kennedy called Khrushchev's bluff and faced him off. Blustering awhile to obtain U.S. assurance that it would not invade Cuba, Khrushchev was forced to back down and withdraw the missiles. In his exuberance he had simply moved too fast in inserting the missiles into Cuba. Had he waited a few months later, the strategic superweapons would have been operational and ready. Any U.S. naval blockading force would have thus faced quick and eerie destruction or disabling, as was soon to be demonstrated in the case of the U.S.S. Thresher and its surface companion, the U.S.S. Skylark.

This severe loss of face by Khrushchev, before the Communist Party, made it imperative that he do something quite dramatic to stay in power, and do it as quickly as possible.

Khrushchev's first large, full-up, strategic longitudinal EM wave energetics interferometer weapon became operational on site by April 1, 1963. A few days later, Khrushchev employed it to kill the U.S.S. Thresher atomic submarine (Figure 6) underwater off the East Coast of the U.S. By simply creating EM noise in an interference zone placed on, in, and around the sub underwater, and thus jamming the sub's electrical controls so that it could not be controlled, the helpless sub sank to crush depth and imploded. The U.S.S. Skylark on the surface had multiple systems seriously jammed or rendered nonoperational. This was the Russian deception plan, to have the kill of the sub resemble an accident.

One day after the successful clandestine kill of the Thresher, the same weapon was used to demonstrate a powerful underwater electromagnetic burst 100 miles north of Puerto Rico (Figure 42). The energetics underwater burst equaled a substantial deep underwater nuclear explosion in size. The burst was seen by a passing U.S. jetliner crew and was reported to the FBI and the U.S. Coast Guard when the aircraft landed in Florida. A colleague interviewed the co-pilot of the aircraft, and relayed the interview information to me.

Khrushchev's Second Demonstration

Deep Underwater EM Explosion 11 Apr. 1963

(Similar to this photo set of a deep underwater nuclear **test** explosion)

Phenomena observed and reported by passing U S jetliner

100 miles north of Puerto Rico, over Puerto Rican Trench (One of the deepest parts of the ocean)

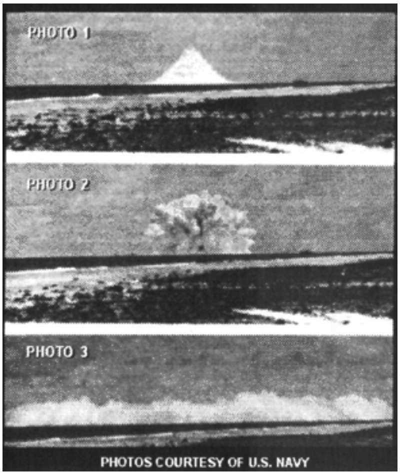

Figure 42 Deep underwater nuclear burst, demonstrating same water phenomenology produced by Khrushchev's second demonstration on 11 Apr. 1963, the day after killing the U.S.S. Thresher.

Why the Soviets Could Continue to Attack and Kill U.S. Targets with Impunity

Ironically, Khrushchev clearly demonstrated the inability of U.S. science to comprehend the numerous test incidents of Soviet LW interferometers, including the actual destruction of strategic U.S. targets (our latest and greatest nuclear attack submarine!) right off our coast. *The greatest strategic weakness of the U.S.—and the thing that may yet completely destroy all of us—is the continuing snail-like pace of U.S. scientific progress in revising, correcting, and extending its seriously flawed classical electrodynamics model. This scientific inertia placed us several decades behind in development of defenses in the areas of longitudinal EM wave weapons, quantum potential weapons, and causal system weapons. It countered our strategic nuclear armada for two decades. We were for some time analogous to the Spanish Navy with its wooden ships, when it faced and was destroyed by American steel ships. Hopefully at least some more modern work has now been accomplished in higher group symmetry electrodynamics, so that we are not quite as defenseless as we formerly were. But we still have a very long way to go.*

When one performs a standard strategic analysis on a nation, one of the areas examined in depth is called *national style.* One of the major characteristics in the U.S. **national** style is that we do not react to slowly increasing threats or changes

(Figure 43). That style particularly applies to the scientific community. Because of it, the Russians knew that our government only inches forward in achieving new *basic* science. We run at enormous speed, doing *applied* science. We are essentially Romans and great engineers. We are largely not Greeks and not great esthetics discoverers. We also will just ignore new threats if they emerge very slowly and are only gradually perceived, and we will hide behind that infamous phrase, "We have no evidence that..."

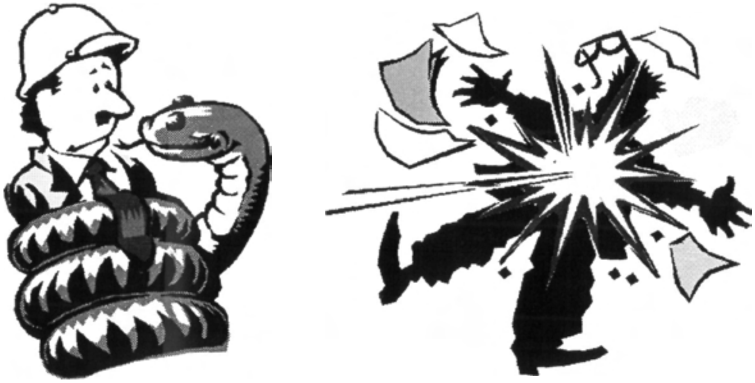

BOA CONSTRICTOR LIGHTNING STRIKE

Americans tend not to react to slowly increasing things,
but react to rapidly increasing things.

Figure 43 America's national style: we do not react to slowly increasing threats.

In short, *we react to the coiled rattlesnake, not to the boa constrictor*. We are now in a long, bloody, costly war against terrorism because of that very characteristic of our national style. For decades, the slowly increasing infiltration of terrorists and their weapons was just ignored. Oh, yes, the official "threat studies" pointed it out, placidly.[25] But no one got very excited about it. The prevailing attitude was that "Ah, yes, but nothing like that will happen on *my* watch!"

As clearly demonstrated by their successful decades-long radiation of the U.S. Embassy in Moscow, the KGB chess masters knew that, if the test kills using their superweapons were not too often repeated, and if they even vaguely resembled other things known in the world, the U.S. scientific community and government investigators would resist all suggestions of Russian superweapons and super technology using new principles of nature. And so would our intelligence community, since they rely on the scientific community for their scientific assessments. They were well aware of our historical "ostrich" position (Figure44).

---

[25] Indeed, if one will look at the threat studies prior to William Cohen taking office as U.S. Secretary of Defense, one will find widespread complacency about the terrorist threat, and mostly just lip service to it. It was definitely not very important to most of the intelligence and scientific community. Cohen built a fire under the entire threat community, forcing them to get their act together and start actually studying the terrorist threat, asymmetrical warfare capabilities and dangers, and what could be done about it. He did force the research and threat communities to do a creditable job of studying it, and to publish reports detailing it. Unfortunately, the tenor of the political community outside SecDef Cohen's control was basically "business as usual, not to worry." September 11, 2001 changed all that forever. Once again we had waited until the rattlesnake struck and actually bit us, before the nation rose up and paid attention.

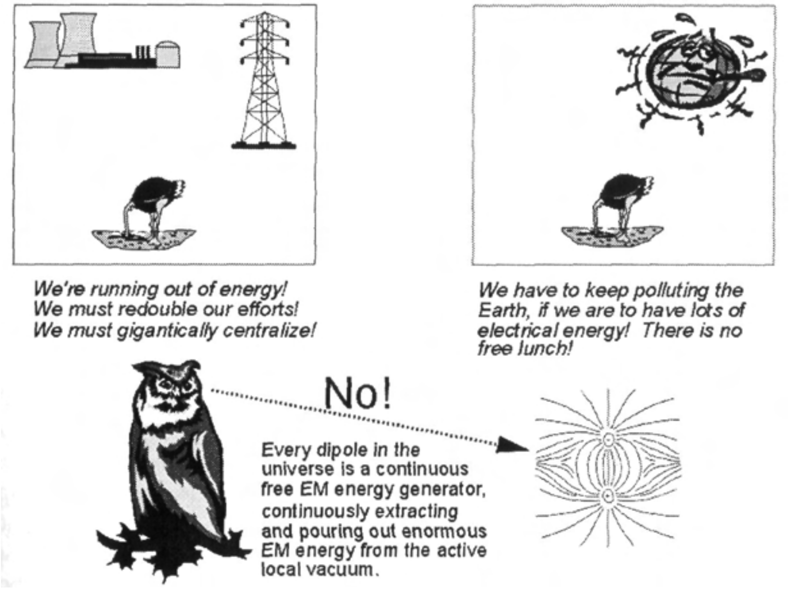
We're running out of energy!
We must redouble our efforts!
We must gigantically centralize!
We have to keep polluting the Earth, if we are to have lots of electrical energy! There is no free lunch!
No!
Every dipole in the universe is a continuous free EM energy generator, continuously extracting and pouring out enormous EM energy from the active local vacuum.

Figure 44 The only energy resource problem is the theoretical ostrich problem in science. We *must* become theoretical owls, rather than ostriches.

The KGB deception just depended upon the U.S. tiger to keep its stripes and exhibit its normal reactions or lack thereof. They also utilized (i) a series of strategic deception plans, deliberately designed to perform the tests suggestive of discredited phenomena such as UFO reporting and "lights at night" types of things, and (ii) a continuing series of subtle and anomalous stimuli (probes) designed to rather slowly attract the attention of U.S. Government at high levels, and of state-of-the-art U.S. government scientific contractors as well. The trick was not to attract U.S. attention *too suddenly or too strongly.* Show them the slow boa constrictor, not the lightning-fast rattler, and see how they reacted, if at all. By their reactions, the KGB could assess whether or not the U.S. government, intelligence, and scientific communities understood what was happening.

Precisely as anticipated by the KGB, we showed them for some decades that we really had little or no comprehension of the superweapons and their testing right over our heads. Even today, at least in the open community there is still little recognition, even of the violent weather war raging right over our own heads since July 4, 1976 and easily recognized.[26]

A case in point was the deliberate induction of diseases and health changes in U.S. personnel in the U.S. Embassy in Moscow for over 40 years (Figure 11). On a continuing basis, this excellent intelligence probe showed (by the U.S.'s actions or lack thereof) that we knew nothing about energetics.

Another KGB probe method was to kill a target once in a while, such as a jetliner or a submarine (deliberately rather seldom here, because the U.S. military would react if a certain frequency threshold were exceeded). Such kills resemble or can be posed as natural accidents or terrorist incidents. Quite a number of such incidents have occurred over the last three decades. The National Transportation Safety Board (NTSB) does not have higher group symmetry electrodynamics in its knowledge base or on its "checklists", nor does it have scalar interferometry, psychoenergetics, bioenergetics, or quantum potentials. In total, the results of deliberate stimuli of the U.S. continually assured the KGB chess masters (with 100% certainty) that we were a complete novice in the energetics game, and we were a lamb increasingly ready for the slaughter.

The major contribution to the continued success of these deliberate measures by the KGB has been provided for decades *in the "nonreaction of our own scientific community to slowly increasing threats."* To a lesser extent, our own intelligence community has also not focused sufficiently open-minded and highly qualified scientists to comprehend what is occurring in "new principles of nature" phenomena. "In the box" electrical engineers with Bachelor of Science degrees often are the analysts who analyze and assess such "out of the box" phenomena. Good luck!

Ordinary radar engineers may analyze an ordinary radar rather well. They will never, never become aware of a radar beam carrying infolded spacetime curvature engines. Such is not even in their lexicon or in their belief structure. They have no understanding or experience at all with longitudinal or time-polarized EM waves—most have never even *heard* of time-polarized EM waves. Mention curvatures of spacetime and their dynamics as deliberate "engines" impressed *inside* EM fields and potentials and waves, and one gets a glaze right over the eyes of the intelligence analyst recipient.

Western science conceives the mind as a meat computer, and does not understand that mind operations are totally electromagnetic and causal, but nonphysical—meaning in the time domain rather than in 3-space. Mind operations primarily use time-polarized EM waves rather than the transductions into transverse EM waves that the scientists actually measure. Hence Western science has not the foggiest notion as to what Russian psychoenergetics is all about, or how it can even "be." Certainly Western science has had no notion of deliberately creating time-polarized EM waves for direct engineering within the mind itself and within the mind's time-EM operations.

---

[26] The weather war over our heads has been particularly violent, and violently resisted, this year (2002).

As Arthur C. Clarke puts it in his "third law," *"Any sufficiently advanced technology is indistinguishable from magic. "*

Our long fascination with our "normal" (optics, IR, UV, radar, COMINT, satellite photography, etc.) technological intelligence collection systems will certainly not lead to any great breakthrough understanding of Russian energetics weaponry. That could only have come through HUMINT or COMINT. HUMINT is an area which, in the past, we have mostly just neglected or given lip service to for decades—and that came back to bite us severely on September 11, 2001. Even when we use HUMINT, it will not yield energetics results if the managers and directors themselves (and the agent handlers in the field) have no understanding of it and do not even know the proper questions to ask. And the system will not deliver results if the Intelligence Collection Requirements do not address the proper issues because ordinary electrical engineers and conventional electrical physicists having no inkling of psychoenergetics, or how it works, drew them up.

COMINT, on the other hand, suffers from its own extraordinary prolific output -as well as by only conventional analysts and conventional scientists examming it. Warehouses of COMINT information often have not even been read in any way except a casual glance by a junior analyst, and certainly have not had serious foundations scientists examine it in detail. ELINT does a little better insofar as a human analyst looking at it, but then what is a conventional electrical engineer or radar engineer to make of a giant hemispherical globe that slowly forms and expands from a tiny "seed bubble" of intense glowing light that appears (Figure 35), as it did off the Kurils in April 1984? What will he label it? "Light phenomena?" Where are the technical descriptors of the actual process forming the phenomenon? Where are the implications as to the strategic military use of such phenomena? The reader can see the point. For decades our analysts ignored prolific reports of such incidents by the hundreds. Why, everybody knew it was glowing swamp gas!

One cannot even write a proper "intelligence collection requirement" if one does not understand the basic phenomenology. So essentially no one in U.S. intelligence was writing the proper intelligence collection requirements.

The Science Behind the Soviet Energetics Weapons

The science behind these weapons is called *energetics*. (See again Figure 8.) Energetics consists of three branches, depending upon the nature of what is being targeted. The branch targeted against nonliving matter is just called the same name: Energetics. Targeted against living cells and biological material, the nervous system, etc., that branch of energetics is called bioenergetics. Against the mind and its functions (which are time-like and thus involve scalar photons and time-polarized EM waves), energetics is called psychoenergetics {89}. See Figure 45. Psychoenergetics includes both the conscious[27] and unconscious mind[28] of the individual—and even Jung's collective unconscious mind[29] of the entire human species {90}, and in fact even the collective unconscious mind in all species on earth, which corresponds to Lovelock's *Gaia* concept.

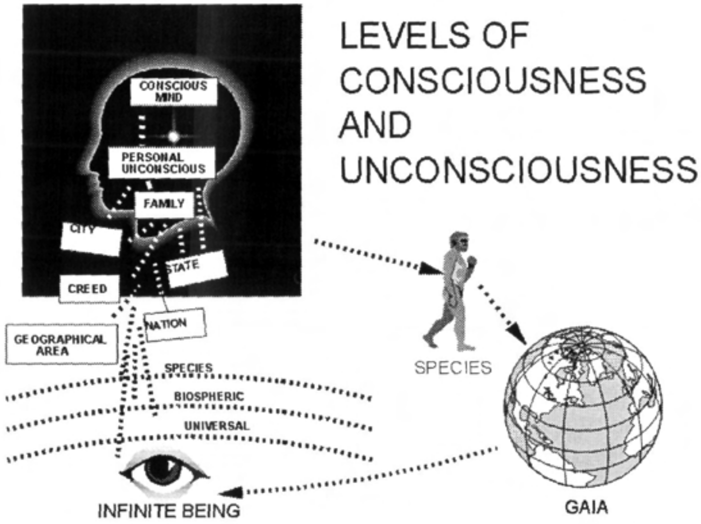
LEVELS OF CONSCIOUSNESS AND UNCONSCIOUSNESS
CONSCIOUS MIND
PERSONAL UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL
INFINITE BEING
SPECIES
GAIA

Figure 45 Mind levels are subject to engineering by higher group symmetry electromagnetic (psychoenergetic) weapons.

We remind the reader that we are talking *physics,* not metaphysics and not mysticism. Strong Soviet physics developments over a sustained period have occurred in all three energetics areas, and these areas continue to be developed at present.

Once a nation enters into LW EM development, it has just entered into unified field theory development also. Particularly with longitudinal EM waves, one is directly engineering general relativity itself, and spacetime itself, as well as the dynamics of spacetime curvatures and sets of such curvatures. But it is an even greater unified field theory than Western physicists suspect. Not only does it unite the four forces of physics, but it also unites mind and matter and their dynamics as well.

---

[27] The conscious mind is a very fast serial processor, processing one snapshot at a time.

[28] The unconscious mind is a massively parallel processor, doing a great many things simultaneously. It is totally conscious, but multiply so. Since the serial processor conscious mind can only "see a single slide projector in the viewgraph at one time", when it "looks at the unconscious mind" it sees a large number of slides simultaneously, hence just sees "black" or "nothingness" (no single thing).

[29] Jung himself labeled it a *mind* (and hence a living entity). His modern followers shudder at such a concept, and very much downplay the "mind" and "entity" aspect, preferring that it just go away.

We shall also refer to time-polarized EM waves as *time-density waves*. The reason is that a time-polarized EM wave is one where the EM energy is on the time-axis and oscillating along it, like the squeezing of an accordion. In short, it is a longitudinal EM wave on the 4th Minkowski axis *ict*, where the only variable is *t.* Hence it is a wave of the variation of time-density along that axis.

All three branches of KGB energetics use longitudinal EM waves, longitudinal E*M* wave interferometry, and Whittaker-structuring of fields and waves to provide organized curvatures of spacetime known as *spacetime engines* or v*acuum engines*. They also use time-polarized EM waves, as mentioned above. Any or all can also employ the quantum potential as well as negative energy EMP aspects. For now, the reader should just recognize that each infolded longitudinal EM wavepair comprising the scalar potential, includes both a *time-density wave* along the time axis and a longitudinal EM wave in 3-space {91}. So ordinary electrostatic scalar potential fi—ordinary voltage—is made of a harmonic structure of *coupled time-polarized and longitudinal EM wavepairs.*

Oscillations of time density oscillate the local rate of flow of time. In turn, since time has the energy density of mass, its spatial energy equivalence *W is* given by $W = c^2t$. Any change in a time density wave is a very powerful oscillation of loca1 spacetime curvature. So time-density waves constitute a very powerful form of infolded general relativity, inside potentials and hence inside all EM waves and fields. The KGB weapon scientists call that inner general relativity (time density waves) the *information content of the field*. Unfortunately our follows confuse that term with common "spectral analysis" which it very definitely is not. But as part of the deception plan, the Russians are quite happy to help them think it is exactly that, and in fact the Russians go to some pains to continue to help our fellows think that.

KGB energetics involves the use of a highly structured general relativity infolded inside electrodynamics—inside EM potentials, fields, and waves regardless of type. In addition, that very special GR is using the very strong EM force along the time axis as the main agent of curvature. By deliberately structuring these time-polarized EM waves and their dynamics, extremely powerful *spacetime curvature engines* can be employed, either created locally or at a real distance, or also globally in an entire selected distant region.

Once these spacetime curvature engines structure the distant targeted spacetime region. Then in that interference zone the engines therein act upon any mass present. In theory, any physical effect known can be engineered, according to general relativity, once one can create and utilize the proper set of spacetime curvatures (the proper set of spacetime curvatures and their dynamics). In GR, curvature of spacetime acts directly upon any mass present, and vice versa. With such technology one can therefore *engineer mass* to any level and at any distance (and eventually in any quantity) at will, using energetics methodology. *The engineered mass may be either living or inert* {92}. It follows that bioenergetics and psychoenergetics effects can also be engineered at a distance in similar manner. All it requires is the proper engines for the effects desired in the targets desired.

By using time-density waves and their dynamics, mind itself can be engineered and controlled at a distance as in the cases of Captains Button (Figures 22 and 23) and Svoboda (Figure 24).

The Importance of Scalar Photons and Time-Density Waves

Now we digress momentarily. Again, there are four major photon polarizations in quantum field theory {93}. The first two are *transverse* photons, polarized in the x- and y-direction respectively (transversely to the line of propagation, which is customarily taken along the *z*- axis). Ordinary transverse EM waves are comprised of such photons or blends thereof. Though ignored, EM waves prior to interaction with mass do *transport EM time-energy* as well as EM spatial energy, though in electrodynamics that time-energy transport has been ignored, thus confusing the effect with the cause.

The third polarization is a *longitudinal* photon, vibrating along the *z*- direction (the direction of propagation). Obviously longitudinal EM waves are comprised of such longitudinal EM photons.

The fourth polarization is a *scalar* photon, since it oscillates only along the time axis (the word "scalar" refers to the fact that it has no vector field component in 3-space). In American physics the latter photon is largely ignored since time is not an observable in quantum mechanics. However, such scalar photons do comprise time-density waves and EM time-energy currents, and they are involved in powerfully curving spacetime via oscillating the time domain. Hence by their oscillations they produce spacetime curvature oscillations, which in turn project small components into 3-space. Sufficient such coherent projection components are in fact longitudinal waves in 3-space.

The combination of the scalar and longitudinal photons is observable as the instantaneous scalar potential, and when Whittaker's 1903 decomposition of the scalar potential is slightly corrected, then the scalar potential is comprised of coupled pairs of waves, where each pair consists of a time-polarized EM wave coupled to a longitudinal EM wave. The combination of the scalar and longitudinal EM waves in the presence of charge is observable as the macroscopic scalar potential. Negative charge absorbs the time-polarized photon from the time domain and re-emits a longitudinal photon in 3-space, while the positive charge absorbs a longitudinal photon from 3-space and re-emits a time-polarized photon in the time domain.

Psychoenergetics: Engineering the Mind-Matter Interaction by Engineering Mind

We shall be interested in what actually constitutes a living biological system in **the** psychoenergetics view (Figure 46). Mind operations are *time-like* rather than existing in 3-space, and for that reason mind and its operations are not observable. Observation is a *d/dt* operator acting on an LLLT ongoing interaction, which removes the T and leaves a frozen LLL snapshot, via d/dt(LLLT) => LLL. *Imposition of observation itself breaks,* any and all connection with time-like operations, including mind and its operations.

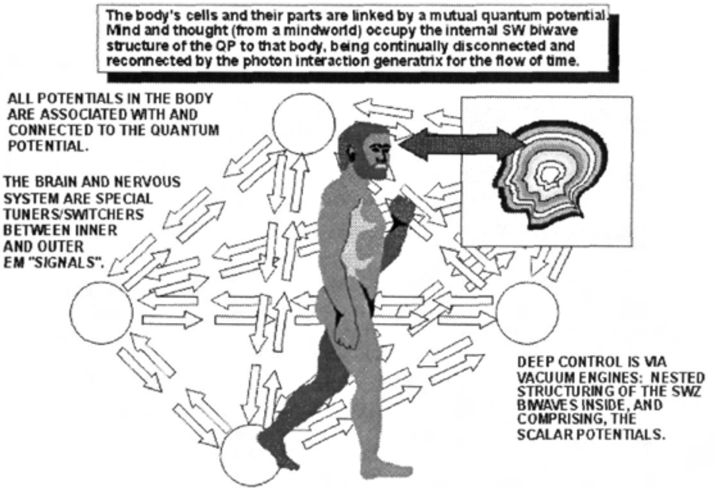
The body's cells and their parts are linked by a mutual quantum potential. Mind and thought (from a mindworld) occupy the internal SW biwave structure of the QP to that body, being continually disconnected and reconnected by the photon interaction generatrix for the flow of time.
ALL POTENTIALS IN THE BODY ARE ASSOCIATED WITH AND CONNECTED TO THE QUANTUM POTENTIAL.
THE BRAIN AND NERVOUS SYSTEM ARE SPECIAL TUNERS/SWITCHERS BETWEEN INNER AND OUTER EM "SIGNALS".
DEEP CONTROL IS VIA VACUUM ENGINES: NESTED STRUCTURING OF THE SWZ BIWAVES INSIDE, AND COMPRISING, THE SCALAR POTENTIALS.

Figure 46 A living biological system (psychoenergetics view).

"Dynamics" involves "velocity", and in general relativity's 4-space velocity is rotation (Figure 47). Any mind-operation or change is dynamics and produces a small velocity, which is a small rotation. This projects a very small component into 3-space, in the physical body, but still in the virtual state. We are speaking of physics rather than metaphysics! *Sufficient* coherent mind operations (changes) have just such small (virtual) rotation components in 3-space that will project coherent little projection components (virtual) into 3-space. Coherently sustaining these components (intent) results in their coherent integration to breach the quantum threshold and become detectable (produce an observable 3-space result). That is how an "intent" change sustained in the mind becomes an actual physical energy input to the body and nervous system.

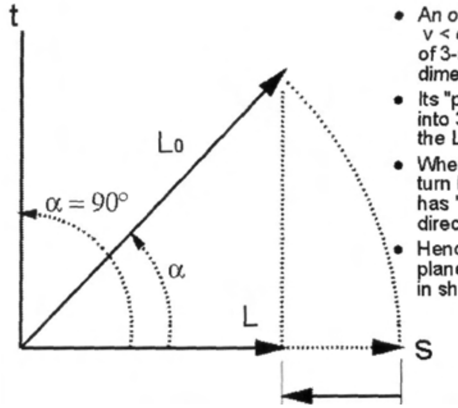

- An object moving with velocity v < c has rotated slightly out of 3-space toward the time dimension.
- Its "projection" back into 3-space is shorter, providing the Lorentz-Fitzgerald contraction.
- When velocity v = c, a full right-angle turn has been made, and the object has "lost" its projection along its direction of motion.
- Hence it appears as a 2-dimensional plane moving at the speed of light; in short, a photon or a light wave.

Lorentz-Fitzgerald Contraction

Figure 47 Velocity is rotation. Necessary to understand the mind-body coupling mechanism.

We again stress that the scalar potential is comprised of coupled time-density and longitudinal EM wavepairs. Mental operations also are time-density wave changes, hence they are changes to the interior of potentials, including the ambient potential of the vacuum itself. Potentials superpose, and their coherent internal structures add (integrate) coherently. Hence, mind changes can coherently integrate to produce minute but real, observable, physical input changes to the body system. *Mind permeates the body via the interlinked and superposed potentials of the body.* The living bio-organism's mind and its operations can be modeled as an organized operating wave dynamics inside a scalar potential and superposed potentials of the body of that organism. Since fields are made from the more fundamental potentials, these operations also function inside the fields of the body. The organized time-density dynamics constitute both the conscious mind and its operations and the unconscious mind and its operations. More subtle depths of the time-density dynamics also constitute the human collective unconscious mind functions and operations. Again, we are speaking physics, not mysticism.

See Figure 48. The "component projection" (coherent integration) process from time-like mind operations into 3-space in the body potentials introduces into the living body the *physical* commands correlated to the *mental intent* of that entity. This is the solution {94} to the long unresolved *philosophical problem of intent*, and addresses the unresolved physics problem of *just what constitutes the "observer."* But the subsequent physical changes accomplished by the responding body as a servomechanism, also produce small rotations into the time domain (into the mind domain) but in the virtual state. However, a series ol' coherent physical changes produces coherent virtual state changes in the mind realm, which in turn coherently integrate into perceptual input. In this fashion, the mind senses the response of the body to its order, and any errors in that response that need correction. The time delay gives also the sense of identity and personal persistence, hence creates the perception of self, the "responding body belonging to self," etc. Interwoven perceptions of coherent rotation changes from other 3-space phenomena that do not respond to mind intent, also create the perception of the "separate physical world." That these phenomena do not respond directly to mental intent creates the perception of the "separateness" of this "external" physical world, while the interweaving creates the perception of "existing simultaneously" with or in the external world.

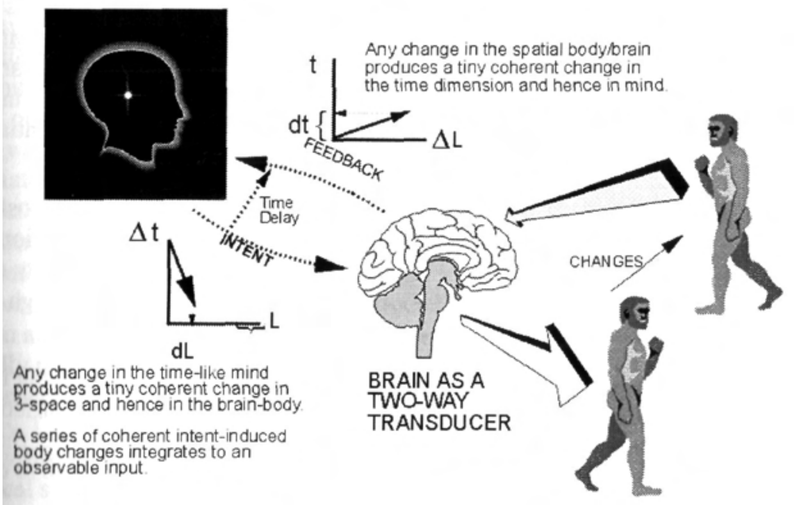
Any change in the spatial body/brain produces a tiny coherent change in the time dimension and hence in mind.
t
dt
ΔL
FEEDBACK
Time Delay
Δt
INTENT
L
dL
CHANGES
Any change in the time-like mind produces a tiny coherent change in 3-space and hence in the brain-body.
A series of coherent intent-induced body changes integrates to an observable input.
BRAIN AS A TWO-WAY TRANSDUCER

Figure 48 Mind-body coupling mechanism at work.

All of that perception, dynamics, rotation, and processing is subject to direct engineering via psychoenergetics means.

Without further elaboration, we again stress that the conscious mind is a serial processor, whereas the unconscious is a massively parallel processor. Westerners are prone to equate self to consciousness or conscious mind operation alone. That is not the case. Consciousness is just one attribute of the self (of the total feedback loop having delay between feedforward and feedback).

The totality is the mind-body coupling mechanism (Figure 48) enabling a living biological system (Figure 46).

By using time-density waves formed of scalar photons, the mind realm also can be directly engineered by the same energetics technology. Hence, for the Russian energetics weapon scientists there followed the development of psychoenergetics as an entire class of breakthrough electromagnetic weaponry.

This resulted in the development of not only a unified physics of matter and space, but also a unified physics of mind and matter reaction. Any true unified field theory must integrate the observer and the unobserved as well as the various categories of the observed. It is now possible to *directly model and engineer both physical and mental reality and their interaction.* Russian psychoenergetics has been developed to do precisely this {95}.

Quantum Potential Weapons Are Utilized in All Three Energetics Branches

In addition to scalar interferometers, there is another category of energetics weapons used in all three branches. That category involves the use of *quantum potential weapons,* after David Bohm's quantum potential in his hidden variable interpretation of quantum mechanics {96}, published in *Physical Review* in 1952. We openly published one mechanism for creating a quantum potential {97}; there are others.

Presently these quantum potential weapons are the most powerful and most flexible weapons on earth. Nothing else can stand against them. Any nation not possessing QP weapons is already a second-rate power, regardless of how many nuclear missiles, bombers, and submarines it possesses, or how many high-energy lasers and high power microwave weapons it possesses.

The greatest problem the Soviets had was not in *making* the QP itself, but in (a) *localizing* its creation to the specific area or region desired, and (b) restricting its coupling to the type of equipment, entity, level, size, type of nuclei, etc. that was desired. We have not yet fully solved the problem of how they are able to do that, but have made a little progress upon it.

The Importance of Longitudinal EM Waves

The KGB energetics weapons use longitudinally polarized and time-polarized EM waves as their basis, rather than transversely polarized EM waves. There is a surge of renewed U.S. interest—and a revolution in electrodynamics—just underway in this country (and in the major weapons laboratories of other nations) *vis a vis* longitudinal EM wave solutions to the major equations of physics. E.g., one may check the Los Alamos National Laboratory Internet web site for summary papers by Rodrigues and Lu {98}. There are now formal EM wave solutions faster than light speed, slower than light speed, etc. {99}. There is also experimental support for superluminal tunneling of EM waves in waveguides {100}. In addition, as is well known, the speed of the ordinary scalar potential is infinite in the Coulomb (transverse) gauge {101}

Higher topological algebra is being used to embed the electrodynamics in most of the KGB energetics work. Maxwell's original theory {6} was some 20 quaternion and quaternion-like equations in 20 unknowns. Quaternion algebra has a higher topology than either vectors or tensors. Heaviside and others reduced the quaternion-like theory of Maxwell down to four equations in four unknowns—or to two coupled equations, depending upon how represented—and formed the much lower topology vector algebra in the process. This seriously demoted the topology of electrodynamics and placed undue limitations upon what can be done with it. The subsequent mathematical promotion of electrodynamics to tensor algebra still did not recover much of the EM functioning possible in quaternion algebra. An even vaster, far more powerful electrodynamics emerges in *Clifford Algebra* electrodynamics, which may be **the** coming modern trend.

However, even with Clifford Algebra, there remain fundamental foundations difficulties—such as bluntly confusing *energy transport* with *energy dissipation, by* arbitrarily integrating the Poynting vector around a closed surface so as to select only the divergent portion of the Poynting energy flow. That produces the ene***rgy*** *dissipation flow* associated with the circuit, not the overall energy flow associated with it. The overall energy flow is some 10^{13} times greater in a nominal case, as (i) discovered by Heaviside in the 1880s, (ii) never even considered by Poynting, and (iii) deliberately discarded by Lorentz with a little integration trick {102} still used by classical electrodynamicists and electrical engineers. Electrical engineering departments, professors, and texts do not know and do not teach that all EM power sources—such as generators and batteries—actually pour out of their terminals a far greater energy flow than the meager shaft energy input to the generator or the chemical energy dissipated in the battery. Lorentz arbitrarily discarded the huge Heaviside nondiverged output energy flow component a century ago, with the observation that it "has no physical significance" {103,104}.

Part of the emerging electrodynamics revolution is intuitively obvious. Higher topology algebra used in one's electrodynamics model means that lots more things can be done—in circuits, equipment, and components. As an example, su**ppose** one can only draw geometrical figures on a flat, uncurved plane. Then compare the figures there to what can be drawn when a third dimension is added so that the plane can be curved, twisted, folded, stretched, compressed, etc. in 3-dimensional space. That is similar to the dramatic differences between electrodynamics in a higher topology algebra and electrodynamics in the far more limited tensor and vector algebras.

Further, if someone's circuits *are working* in a higher topology, a lower topology a**nalysis** will not even see their true functioning. This turns out to be true for Tesla's work, where so many pundits have used vector or tensor algebra in the belief that they therefore knew what Tesla was doing. They did not, nor did they even use the proper tools to see what he did. For the proof, see Barrett's examination of Tesla's actual patented circuits, using quaternions {105}.

Energy, Wave Energy, and Wave Speed are All Dramatically Altered

The energy of a pure longitudinal EM wave (LW) is infinite (unlimited). In practice, only an imperfect LW is produced experimentally, and therefore its energy is finite, but that energy may be extremely large, and the wave's propagation may be many times faster than the speed of light. This is much like quantum tunneling, which has been measured to be superluminal. Nimtz *et al.*, e.g., have transmitted Mozart's 40th symphony through a barrier in a waveguide at 4.7 times the speed of light {100}. Other researchers have transmitted information at lesser velocities but still at greater than c, the speed of light in standard vacuum.

Einstein assumed light speed in vacuum to be universally constant. That is correct for *transverse* photons and *transverse* EM waves, but not necessarily for other polarizations. Instead, it can be a function of the energy density of the vacuum (i.e., its stress energy density), which can be a matter of the gauge and the wave polarization. Further, *c* as a constant applies to transverse polarized EM waves (which are "surface" waves of the medium, so to speak) and not to longitudinal or time-polarized EM waves (which are "subsurface" or "subspace" waves). In general, the speed of the EM wave can be a function of the energy density of the vacuum (i.e., its stress time-energy density). And it is also a function of the type of EM wave being utilized.

"Empty space" is filled and writhing with the continual transiting of higher order EM waves, the presence of standing high order EM waves, and energetically organized regions where the energy is in the LW mode and in the time-density mode. Indeed, space (more correctly, spacetime) itself is nothing but a plenum of such waves and their wave-to-wave interactions.

The longitudinal and time-density effects in empty space are correlated with seasonal and periodic variations of time-charging and longitudinal EM wave emissions from planets, stars, etc. As a result, the characteristics of empty space itself periodically vary, and these can be detected as variations in the speed of light through the affected space.

The interplanetary radar measurements, e.g., do indeed show several periodic variations of the speed of light in interplanetary space, on a weekly and essentially monthly basis. In this data the proclaimed constant speed of light is upheld only by averaging the measurements over a month-long period, as shown by Wallace {106}. Actually, the scientists are dealing with variations of the periodically varying local time-density of spacetime, but have not taken that into account.

There are also specialized "galloping wave" zones, created by variations in the interferometry of the interpassing and interfering longitudinal EM waves in space. Such galloping waves from wave interferometry are already known in the literature {107} from 0.1 to 100 times the speed of light, but they do this on a "surge forward and lag backward" mode about an *average* speed of c in the laboratory here on earth. In space away from planetary masses, that value "c" varies from its textbook value also.

All that has really been shown is that interplanetary space is *not quite* uniform after all, contrary to present scientific assumptions.

One would think that this *fundamental foundations anomaly* shown by the interplanetary radar measurements would excite the entire scientific community and trigger a massive scientific investigation. Instead, scientists such as Wallace who publish critical analyses showing that the data directly prove the variation **of c** in "ordinary" space, are just ignored and essentially "parked on the side-lines" until they eventually fade away.

A Process for Creating Longitudinal and Time-Density EM Waves

Today we know that some plasmas very often can produce a longitudinal EM wave from a transverse EM wave, and can produce a transverse EM wave from an input longitudinal EM wave. This was not known when the Priore team was performing its experimental work. However, it provides one method for producing and detecting longitudinal EM waves.

The Priore team in France unwittingly made longitudinal EM waves by inputing transverse EM waves into Priore's rotating plasma in a giant plasma tube (Figures 12 and 49).

- At the end of the project, Priore had developed a large unit to treat humans.
- A few were treated.
- Cures replicated what had been done in lab animals.
- The massive size of the unit and long treatment time (hours) made the process bulky and very inconvenient — and expensive.
- We proposed a way to do it with a suitcase-sized unit and a treatment time of two minutes.*

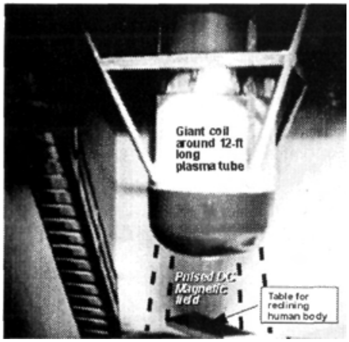
Giant coil around 12-ft long plasma tube
Pulsed DC Magnetic field
Table for reclining human body

*See *Porthole Briefing,* www.cheniere.org

Figure 49 Priore's large plasma tube installation in his 3-stories-high treatment facility for humans. Results obtained in animal experiments were also duplicated in human experiments.

Quantum Potentials and Multiply Connected Spacetime

It appears that a sufficiently clean longitudinal EM wave can be utilized to create a *quantum potential*, very similar to Bohm's notion, by iterative retroreflection between two phase conjugate reflectors. When sufficient *iterative self-targeting* occurs and persists, the waves eventually "snap in" and fully converge to laser-like beams without any energy divergence {97}. At that point, one has achieved an essentially "pure" LW, which has infinite velocity. For all practical purposes, there is then little or no difference between this wave and a quantum potential.

A pure longitudinal EM wave just "appears instantly" everywhere in space that it will be. It is already an element of a quantum potential. In this sense, all that a "quantum" potential is, is an ordinary scalar potential in which the composite EM longitudinal phase conjugate wavepairs really are pure longitudinal EM waves. It would seem that this also moves the gauge to the Coulomb gauge, so far as the potential is concerned. It follows that, by establishing significant and clean longitudinal EM wave contact between any two points via iterative retroreflections, one establishes a quantum potential if sufficient harmonics are established in the contact. The implications for "phase conjugate shooting" to go one step farther are obvious.

By "scanning" an area with the appropriate LW transmitter array, with deliberately imperfect LWs so that contact reaction occurs, and then decreasing the imperfections in the scanning by iterative retroreflection, it appears that a quantum potential can be created and connected between the transmitters and a desired distant region.

Essentially, in the case of pure longitudinal EM waves and quantum potentials, we must model the situation in terms of a *multiply connected spacetime*. The divergence-free waves are now pure longitudinal waves in one sense, and pure potential beams in another sense. In any case, the Coulomb gauge applies. As is well-known, the potential moves at infinite velocity in the Coulomb gauge.

Thus a single station can be connected by such a QP to any number of distant stations—by first "knitting them together" into a common multiply connected spacetime by iterative mutual LW self-targeting beams. It matters not if one or more distant station(s) is/are then moved beyond the star Sirius; the moment an energy signal appears in the initiating station, it is already instantly in each distant coupled station anywhere in the entire universe. There is no separation distance involved. Here *one no longer needs to propagate transverse EM wave energy painfully through space at the speed of light.* Instead, one is now dealing with multiply connected spacetime. One now has the instant appearance *in each and every distant coupled target* of the energy input into the single "transmitting" or "coupling" station. So the process is a startling and giant "energy amplification method" as should be expected from something that has infinite energy in the first place!

One vulnerability and disadvantage of a QP weapon is that, if the target is aware of one's coupling and also has QP technology, he can now immediately send back any energy form or "greetings" he wishes, from any one of his coupled targets. He can even do it inadvertently, without even knowing he has been coupled by a quantum potential. In other words, the QP provides a two-way street, if the target is aware of it and has the technology himself, *and even if he is unaware of it and does not have the technology*. For that reason, actual use of the QP weapon in strategic strikes requires caution and special considerations. Usually it requires very rapid single use and then cessation of any further usage.

Need for Caution with QPs Used Against Biological Organisms

Another QP weapon disadvantage is that rather extreme care must be rigidly exercised, particularly when linking living creatures. For living things, the QP is best utilized in the absolute minimum energy mode. E.g., *an extremely weak OP already links every member of the human species*. Suppose one increased that linkage intensity carelessly. Immediately the entire human race is in grave difficulty. Every death of a person feeds into every other living person. Every birth of a baby feeds into every living human. Even the fertilization of one **human** ova instantly feeds into all other ova. The integration of all those changes would totally disrupt—and instantly destroy—all human life on Earth. Since chimpanzees have very similar genetics to that of humans, all the chimps would probably die also. Other higher primates would also suffer or die, as would many other species.

What might be useful, of course, would be to use a QP to kill every smallpox, HIV, Ebola, and other dread disease pathogen on Earth. If so, it would first have to be made extraordinarily specific, else it could threaten all or most of the life on Earth.

A very useful adjunct would be to use a QP device to nearly instantly scrub and decontaminate (disinfect) an anthrax-contaminated large city—such as New York or Washington, D.C.—by instantly killing all the anthrax spores in that locality. Or to clean all the antibiotic-resistant staph pathogens out of U.S. hospitals by just killing them all instantly and periodically. Or to rid the Earth of necrotizing fasciitis—flesh eating disease pathogens.

An extremely localized QP could be used to connect with all HIV viruses and DNA inside an AIDS patient's own DNA. One could then kill the HIV viruses not yet integrated into the patient's cellular DNA. One could also time-reverse the patient's own HIV-infected DNA back to normal, effectively "removing" the HIV DNA already dumped into the patient's own infected cells. In theory, 100% of the HIV infection could be removed from the patient's body, restoring him or her to a completely HIV-free condition.

A Hypothetical Example

Suppose one arbitrarily increased the even weaker QP that presently joins all living things on Earth, regardless of species. Instantly every living thing on Earth would die, and the Earth would be instantly converted to a sterile planet.

Even though *physically embodied* life was gone, however, all the forms of those species would remain in spacetime activity, their vacuum engines—for both minds and bodies—forever embedded in the local potentials and fields of the material Earth. In other words, the Earth's potentials and the local spacetime potential would still be *dimensioned* and *conditioned* for the life forms that once moved upon it in physical form. The dimensioning and conditioning (engines) for every life form that ever lived on the Earth *exists now* at every point in the Earth and in the local spacetime. With proper application of energetics, in theory it is possible to recover any of these life forms desired, back into living form.

When one understands that environmental aspect, one understands why the Miller-Fox-Urey biogenesis experiments {108,109,110} (Figure 50) would still produce living organisms again. A *physically* dead planet or test flask is not necessarily a "completely dead" planet or flask, when one considers the presence of spacetime engines and dimensioned potentials. One recalls that these forms and engines contain not only physical forms, but also the mind-aspects and the mind/body coupling and dynamics aspects as well.

In short, the entire biochemical notion of biogenesis from truly inert matter devoid of any form of life or pattern of life appears to be a false premise. Since potentials reach to the end of the universe, then any structured potential structures *(dimensions)* the universal vacuum potential everywhere in the universe {111}. The Miller-Fox-Urey experiments were not performed in a truly "sterile" environment after all, when the new energetics is considered {112}. One realizes that the entire universe is alive; the mere presence of a single living microbe or other form anywhere establishes an entire living noosphere. The Russian researchers—such as Kaznacheyev {113}— into EM induction of disease in targeted cells at a distance also realized that biology included a noosphere, and it could be engineered.

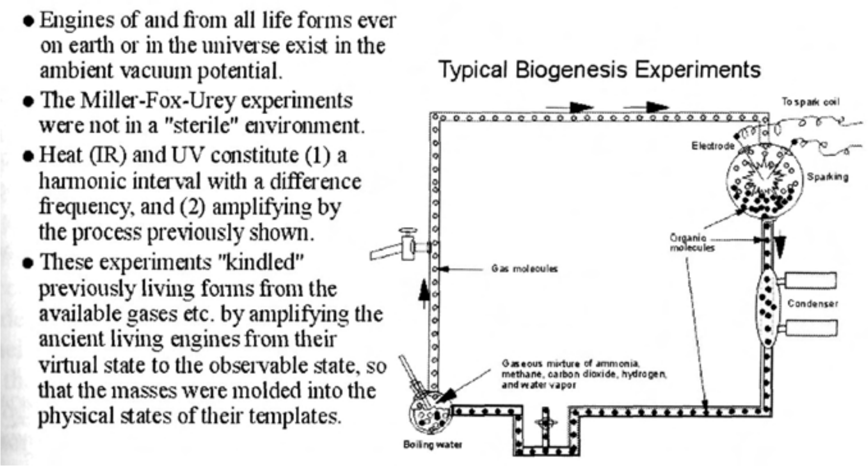

- Engines of and from all life forms ever on earth or in the universe exist in the ambient vacuum potential.
- The Miller-Fox-Urey experiments were not in a "sterile" environment.
- Heat (IR) and UV constitute (1) a harmonic interval with a difference frequency, and (2) amplifying by the process previously shown.
- These experiments "kindled" previously living forms from the available gases etc. by amplifying the ancient living engines from their virtual state to the observable state, so that the masses were molded into the physical states of their templates.

Typical Biogenesis Experiments
To spark coil
Electrode
Sparking
Organic molecules
Gas molecules
Condenser
Gaseous mixture of ammonia, methane, carbon dioxide, hydrogen, and water vapor
Boiling water

Figure 50 Typical Miller-Fox-Urey biogenesis experimental apparatus.

Herman Kahn, then director of the Hudson Institute (a leading think tank), once remarked that it did not concern him that there existed mechanisms that would destroy all physical life on earth. Instead, he stated, it concerned him that he know almost 200 such mechanisms. The quantum potential weaponry is yet another doomsday weapon with a thousand lethal mechanisms to add to Kahn's list, were he still with us.

The QP Weapon Is a Superb Energy Amplifier "For Free"

In a multiply connected spacetime established between multiple separated (to the observer) stations, one can instantly input one megawatt of power into *every distant stution* by inputting it into a single station. Suppose that coupling station is coupled via QP to one million distant stations (scattered anywhere in the universe). Then in each distant coupled station there now instantly appears one megawatt of power, even though we have input only a single megawatt in our "transmitter" coupled station. This is a superb *one-on-many weapon system and energy amplifier*—perhaps the ultimate weapon system.[30]

AS an analogy for understanding, one can look at it this way. Put a dot (point) on a piece of paper. Mathematics tells us that an indefinite number of other points can be "superposed" directly on that original point. Now input some "energy" into the original point. It instantly appears in each and every other superposed point. That is precisely the way a quantum potential linkage between stations works, except that in normal singly-connected 3-space the points are *observed* by a normal observer to be widely separated. Aristotelian logic—which is an "either/or" logic—does not apply. Instead, a higher topology logic applies.[31]

---

[30] However, we should not be loo surprised at such "astonishing" effects, once we appreciate that with the quantum potential and pure longitudinal waves one is playing with infinite energy anyway.

The QP weapon can be used with microwatt input to induce low-level disease states or pre-disease states *(shadow disease states)* (Figure 51) in entire populations. Tests done in this fashion will violate the "vector carrier" aspect of infectious disease transmission. Instead, the tests will cause a few widely spaced nearly simultaneously cases of the disease being tested. Several such tests (e.g., with necrotizing fasciitis) have already been conducted in the populace of the United States.

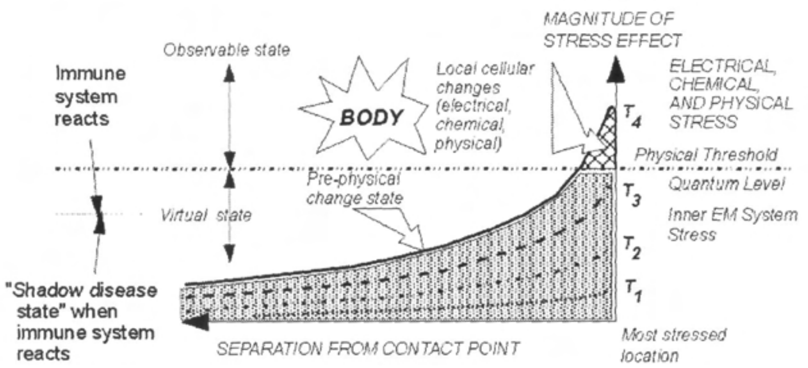
Immune system reacts
Observable state
BODY
Local cellular changes (electrical, chemical, physical)
MAGNITUDE OF STRESS EFFECT
ELECTRICAL, CHEMICAL, AND PHYSICAL STRESS
T_4
Physical Threshold
Pre-physical change state
Quantum Level
T_3
Virtual state
Inner EM System Stress
T_2
T_1
"Shadow disease state" when immune system reacts
SEPARATION FROM CONTACT POINT
Most stressed location

Figure 51 The shadow disease state is that structured virtual state level of the promoting disease engine template, at which the immune system reacts.

Or a QP weapon can be used with megawatt or gigawatt input to blast multiple distant targets simultaneously. One may destroy, e.g., every single transmitter of a given type. Or one may instantly kill all ships and weapons and paraphernalia in a large naval task force steaming toward its target. Or if one wishes to take the ships and equipment, the weapon can just kill all the humans instantly in that task force. ICBM sites, e.g., are vulnerable and obsolete since 1990 to any nation possessing QP weapons. So are underwater nuclear submarines and long-range strategic bombers. The entire strategic and tactical armada we have built up so painfully over the decades is obsoleted—also meaning *countered*—by any enemy possessing sufficiently advanced energetics weapons, and particularly quantum potential weapons.

---

[31] We developed such a logic years ago, and have recently included **it** in *Energy from the* ***Vacuum*** 2002, *ibid.*

Other Novel Uses of QP Weapons

If an enemy has such QP weapons and you do not, you are easily and directly destroyed at will, once certain other deterrents are overcome. One of these previous deterrents that had to be overcome was the stability of nuclear weapons in our nuclear arsenal. To remove this stability and render them easily destroyed, the holder of a scalar interferometry capability could induce sudden massive fissioning and fusioning of many of the targeted onsite weapons if the scalar potential interferometers were used massively. However, the resulting tremendous nuclear explosions—many "in the dirt and dirty"—would also unacceptably damage the attacker. Further, major antagonists usually clandestinely introduced hidden nuclear weapons into the cities and target areas of their opponents, which practice was known as "dead man fuzing." The notion was that, even if one side suddenly destroyed the other, that destroyed side could still detonate its secret nuclear weapons hidden in the attacker's homeland territory, destroying him as well. In essence, the killed side (the "dead man") had set off the fuzing of his pre-planted weapons, to destroy the attacker even though the "dead man" was already dead.

One strong reason that the Russians pursued the quantum potential weapon so strongly was precisely because it offered a mechanism to solve this "dead-man fuzing" problem, removing it as a viable deterrent. Obviously it would be quite advantageous to be able to dud or destroy these hostile weapons in one's very midst, even though one did not know their locations, and pull the fangs of the dead-man fuzing installed by one's opponent.

There appear to be ways of directly connecting a QP, e.g., to every atomic nucleus of a given type, in a given region (say, in a region consisting of on the surface of the Earth, in the Earth to some depth, in nearby space, on the surface of the ocean, and beneath the ocean). Connection can be made, e.g., to every U235 nucleus and to every weapons grade plutonium nucleus. Or with the technology developed, the users ought to be able to make such connection only between certain pre-selected *localized amounts* of U235 nuclei or plutonium nuclei.

Now suppose that such a QP is established between weapon-sized groups of U235 atoms. Then one hits one U235 nucleus in such a group in an activator station with a neutron bullet. Voila! One simultaneously hits every other linked U235 atom (including the activator station—one just blew it to smithereens if it contained very many U235 nuclei) in such weapon-sized groups. The raw U235 in ores in nature is still okay if the ore contained only tiny traces of U235 nuclei. But all the "neutron-struck" fissionable nuclei in nuclear weapons worldwide—regardless of owner or location—fission in the weapons immediately. That destroys all the nuclear weapons on Earth—including the clandestine nuclear weapons that various nations have spirited into other nations perceived to be their enemies, as a "checkmate" against sudden surprise attack. *One* does not need to know where a weapon is at all.

However, it gets all the weapons by exploding them, and that itself could create a giant nuclear destruction and fallout problem that would probably do in most life on earth because a great many nuclear weapons will have exploded "in the dirt and dirty." In short, it is an entirely unacceptable solution to the dead-man fuzing problem.

Of course, no *sane* enemy would wish to use the QP in such fashion, because it is quite suicidal for him and essentially destroys civilization. But suppose the sane enemy is well aware of the exact nature of the new cold type nuclear reactions (we explained these in an initial rough paper given in August 1998 at an Energy conference in Utah). The energy initiating these "time-reversed" nuclear reactions starts from "inside" the nucleons in the nucleus and moves outward (that is true of any "reversed" zone).

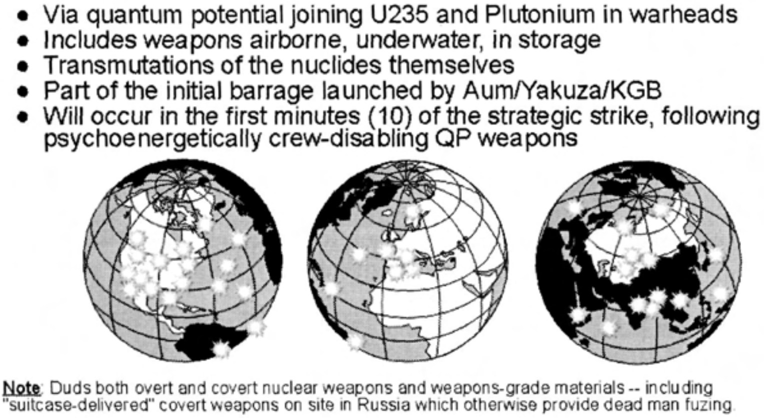

- Via quantum potential joining U235 and Plutonium in warheads
- Includes weapons airborne, underwater, in storage
- Transmutations of the nuclides themselves
- Part of the initial barrage launched by Aum/Yakuza/KGB
- Will occur in the first minutes (10) of the strategic strike, following psychoenergetically crew-disabling QP weapons

Note: Duds both overt and covert nuclear weapons and weapons-grade materials -- including "suitcase-delivered" covert weapons on site in Russia which otherwise provide dead man fuzing

Figure 52 Simultaneous dudding of nuclear weapons worldwide. Also duds nuclear reactors and nuclear propulsion systems.

Direct *isomeric transmutation* then becomes a preferred decay reaction to induce in the nuclear weapons one wishes to dud. By gradually inducing (slowly "kindling") one of these transmutation reactions, one can just "transmute" the U235 to—say—an isomer such as U238, with some readjustment and energy balancing in the ensuing rebalancing decay processes. In that way, one gets no nuclear explosions. One can even choose the transformation or several successive transformations so as to minimize any resulting nuclear radiation. The attacker also destroys all the nuclear reactors, including those in civilian power plants, nuclear submarines, and elsewhere. But he gets just about all the nuclear weapons if he transforms both U235 and plutonium (Figure 52). In about 10 minutes he will dud them all, without any nuclear explosions, but he would probably allow about one hour "just for good measure and nearly 100% certainty."

This offered the Soviets the ideal solution to the worrisome hostile nuclear arsenals of the world, since they also had additional nonnuclear strategic weapons of a different kind and of great power.

Apparently that very kind of QP weapon became operational, on site in Russia, circa 1989 under KGB control. For the first time, it provided the KGB with a guaranteed solution to the nuclear dead-man fuzing problem that had until then prevented its massive use of the scalar potential interferometers (such as in 1985-1986 as originally envisioned by Brezhnev).

By almost simultaneously leasing the earlier scalar interferometer weapons to **the** Yakuza and Aum Shinrikyo at the end of 1989, the KGB obtained an influx of vast new funds {114}. It thus concentrated even more intensely on the more advanced energetics weapons such as QP weapons and psychoenergetics weapons {115}.

So with a QP weapon one can attack multitudes of targets, using a single "coupler" or "transmitter," all without "propagating the EM energy through 3-space." Further, with the technology developed, only the number of outrigger stations coupled limits the energy amplification achievable. One can alter matter at any level, from the quarks in the protons and neutrons in the nucleus, to the lattice bonds, to the molecular bonds, etc. Or one can structure the emerging energy (that emerges directly out of the local spacetime itself), so that *structured local spacetime curvatures* (engines) appear in the target zone(s). These ST engines can be generated so as to perform macroscopic functions against real systems—e.g., jamming electronics, as against the U.S. air attack on Libya in April 1986 (Figure 25), or as was done by a longitudinal EM wave interferometer in the attack and kill of the U.S.S. Thresher (Figure 6) in April 1963 {116}.

The U.S. Narrowly Escaped Destruction Twice in 1997 Alone

On two occasions in 1997 the United States would have been utterly destroyed by the die-hard faction of the KGB, except for direct countering by a friendly little foreign nation. The present author and three colleagues played a very minor role by alerting that nation to the impending attacks. There was certainly no reason to contact officials of our own country; the U.S. was and is powerless to prevent such attacks, although *hopefully* that is changing! It made much better sense contact the only nation that could possibly avert our impending disaster. Of course we later also informed proper U.S. government officials of our actions.

The first strategic attack was scheduled for between early to mid-February, and involved (1) dudding all nuclear weapons, (2) striking the QP weapons in the little nation simultaneously, as well as extant scalar interferometer weapon sites in other foreign nations, and (3) then leisurely blasting cities, facilities, and installations with iterative and continued scalar interferometer strikes. This third phase would include work assigned to the Yakuza/Aum Shinrikyo teams. That is how the Yakuza (and Aum Shinrikyo) expected to revenge themselves for the Western defeat of Japan in WWII, and for the West's dropping the atomic bomb on Nagasaki and Hiroshima. To be able to "get in on the kill" is why the Aum Shinrikyo and Yakuza leased the scalar (longitudinal EM wave) interferometers from the Russians in the first place {117}.

These, along with negative energy EMP weapons, are the kind of energetics weaponry that now determines our fate, and against which the U.S. presently has still inadequate defenses. Our own scientific community does not believe such weapons even exist—apparently due to the "not-invented-here" syndrome. Most of the U.S. scientific community is thus relatively useless and outdated as far as contributing to the strategic survival of the United States is concerned.

Purpose of the Novel Attack on Captain Button

Hopefully, some scientists are aware of the strange incident that involved the mental seizure of Captain Button in his A-10 (Figure 22) and controlling him for well over an hour (Figure 23), including controlling his perceptions, intent, and actions while performing complex technical tasks. That was the test of a KGB *psychoenergetics* weapon, developing and to be deployed by the KGB. That was the "acceptance test", so to speak—deliberately conducted over the United States—for the KGB's system that would be produced in numbers sufficient to counter the QP weaponry, if the QPs of the little friendly nation were to be placed in the insane command and control mode. This was yet another Soviet counter, planned for the near future against the counter by which the little nation did force abortion of the KGB second attack (scheduled for May 1) in 1997.

The purpose of the attack on Captain Button can be understood as follows: The later second scheduled strategic attack on the U.S.—scheduled for May 1, 1997—was forcibly aborted after the friendly defending nation switched to an *insane* command and control mode. The insane C3 mode counters the MindSnapper type attack that instantly kills all on-site operators and renders the sites too "hot" for new personnel to come in from outside and control it. In short, the U.S. could have been destroyed in that scheduled May 1 attack, but with the change of the friendly QP weapons to insane C3, Russia would also have perished in the immediate quantum potential counterblow from the little nation[32] that would have been automatically delivered by the site weapons themselves. The May 1 attack had been predicated upon the fact that a preliminary MindSnapper (negative energy EMP) attack on the friendly QP weapon sites would have instantly killed all persons on site,[33] and neutralized the site for some months because of the resulting longitudinal EM radiation. However, the shift of the C3 of the QP weapons to insane mode rendered such a pre-emptive MindSnapper attack totally ineffective because Russia would still have been destroyed.

---

[32] There are also variations. As an example, the insane C3 can also include a separate channel requiring a countering "punch in" code every 15 minutes (as in a heightened alert status) or the control system will order the launch initiation (this accepts that all on site personnel have now been suddenly killed). So the insane control system is also flexible, and many variations and considerations are possible.

Hence another countering means was required, and in fact that additional system was already in development. It was a very sophisticated array of psychoenergetic weapons, each with a special, highly trained team of operators capable of fully controlling one targeted individual while he was directed and controlled to go about performing desired technical tasks. We estimate that it required one psychoenergetic weapon (very complex, multiple transmitters) and an operator team of perhaps 20 to 30 skilled operators (one for each major human function to be controlled) to adequately control the mind and actions of one distant person performing quite technical tasks. Hence several hundred such "weapon sets" were required to take over all persons on the targeted QP weapons sites in the friendly nation, have the controlled crews deactivate the QP weapons including using the security code and procedures to standdown for maintenance, etc.

Once the QP weapons were totally on standdown, then the MindSnapper would have struck suddenly and killed every living thing on site, neutralizing the site for several months due to subsequent strong longitudinal EM wave emissions from the struck site. No new crews could have entered the site and lived, so the sites could not be operated after MindSnapping, even though the QP weapons were still intact but just on standby. Of course, a third phase then would have involved strikes by scalar EM interferometers to leisurely and completely destroy the systems and sites, permanently taking out the physical system as well.[34] With the first two of that three-part series of strikes completed successfully, then the KGB weapon sites would have been absolutely free to blast North America and our European allies into smoking ruins, leisurely and at will, given the first 10 minutes of dudding all nuclear warheads worldwide, to take care of that old type of "dead man fuzing".

From the open reports of the Button incident, and from the kill of Captain Svoboda a few weeks later, the nature of the next strategic attack that would be scheduled was immediately apparent. Accordingly, we warned the little nation about the shift to a new attack plan, and estimated the new schedule between mid - 1999 and the first of 2000. Of course, we also passed that information on to our government as well, but apparently, no one believed it. The proper U.S. intelligence agencies, however, certainly intercepted my faxes to the little nation's Military Attache in Washington D.C., and particularly my fax a few days before May 1, sent directly to the Prime Minister of that nation and urging the change to an insane command and control system for the QP weapons and the reason why.

---

[33] We point out that it required a longer time to develop the more powerful negative energy EMP weapons capable of just destroying the sites, personnel and all. The first effective weapons of that type could easily kill personnel brains, nervous systems, and hearts while still doing little damage to the much more robust *QP* systems and on site equipment.

[34] We strongly believe there are shielding provisions at those sites, against scalar interferometry attacks, which reduces their effectiveness and delays their destructiveness. However, so far this analyst has not deciphered those shielding methods or mechanisms.

It turned out that, some months after I alerted them to the coming psychoenergetics, MindSnapper, and scalar interferometry attack in that order, to be scheduled between mid-to-end of 1999, apparently the friendly nation suddenly struck and utterly destroyed the KGB facilities for the MindController negative energy EMP weapons being painfully built up in sufficient number. Probably scalar interferometry weapons were used to destroy them. The onsite psychoenergetics crews and facilities were destroyed and neutralized completely.

Anyway, as we passed mid-1999, the type of "Y2K" problem we were worrying about was whether or not the attack was on, headlined by a psychoenergetics weapon strike on the little nation first, with the three phases as previously scheduled back in 1997.

Later, in early 1999 I realized the Russian/KGB development of the causal system robots (Figure 10), mentioned in the introduction. In the latter part of the third quarter of the year, working at my computer at night, I was struck by one of these very beasts from within my computer (Figure 20), in very bizarre but unmistakable fashion (previously detailed). In 10 seconds I was nearly killed and severely disabled for a period of time (it took four days to recover completely). Because of the features deliberately utilized by the CSR on my computer screen so I would recognize it, I grasped the nature of the attack (the KGB *wanted* me to recognize it, and live to tell about it!), how it was done, how it jammed the brain and vision center and strongly fibrillated the heart, upset the equilibrium, etc. I also realized that, in 20 seconds or so additional exposure, I would have been a very dead person, expiring with a very violent heart attack and severe brain embolisms. Note that the CSR deliberately struck directly from the computer screen through my retinas and onto the nervous system, passing through the vision center and flashing it tremendously, and jamming my mind, brain, and nervous system.

When I recovered shakily about four days later, I notified the small nation, once again, as best I could, giving the exact symptoms of the attack and explaining what it was. By then I knew what was coming down, the form it would take, and the deceptive cover (Y2K uncorrected errors!).

The little nation then quickly proceeded to search out—apparently *inside* the North American signals, potentials, fields, and waves "superhighways" for the CSRs—all the CSRs and destroy them in place. It turns out to be a fairly straightforward operation, once one possesses advanced longitudinal EM wave interferometry weapons and techniques. Locating a "very dense" and very dynamic set of LWs performing organized functions is analogous to a normal radar locating a large bomber in cloud cover. One "burst" from the LW "radar," and the targeted CSR is completely disrupted and electronically destroyed.

So a very short (lasting a few days) and strange strategic battle was fought in infolded space or subspace, inside our electrical and electrical systems and inside the fields, potentials, and waves in them. A mighty attacking armada of perhaps several million main battle systems was intercepted and destroyed throughout our homeland, without the slightest notice by our own government, scientific, and intelligence communities. And without a whimper of what is usually recognized as "combat actions".

Afterwards, a very simple and innocuous little message was received which, reading between the lines, simply meant "thank you very much." In short, my hard-pressed colleagues and I got a small "Attaboy!" for our role in alerting the friendly nation.

So Y2K and the Millennium Celebration came, while we waited to see if the CSRs were gone or if they would erupt with damage all over the U.S. Fortunately, they were gone like popping balloons, and so the *fourth* scheduled strategic attack upon the U.S. in the space of three years—this time actually launched—was successfully destroyed.

Whatever I can do to help the survival of my country I will continue to do without hesitation, regardless of whether the U.S. scientific, intelligence, and military communities are aware of the new threat or not, and regardless of whether they believe it or not. Presently our fate and that of the Western world rests on the shoulders of a small but advanced and friendly nation. That is beginning to change, it seems, but still not swiftly enough to suit our needs.

We simply must change all that, by developing QP weapons and defenses ourselves. It is not an option; as an example, China has now finished—as of earlier 2002—and deployed QP weapons. Almost simultaneously, China also finished and deployed the more powerful negative energy EMP (MindSnapper-type but much more powerful) weapons and deployed them as well. So we have entered a very dangerous period when the strategic changes that are likely to occur will be both massive and traumatic.

AS this is being written in early October 2002, it is obvious that our nation is also preparing to strike Iraq if need be, either with a coalition effort or alone. If so, then Saddam Hussein will certainly unleash whatever terrorist forces he has in North America. He will attempt to unleash upon us all the weapons of mass destruction at his command. This could even happen within the next 30 days, with the MidEast exploding, giant polarization of the Arab world, general chaos in the MidEast and upset of world oil supplies, and skyrocketing fuel and energy costs with severe impact upon the already seriously ailing U.S. stock market.

If Saddam Hussein succeeds in doing sufficient damage to our cities and energy infrastructure with more weapons of mass destruction than we have credited him with, then the potential economic collapse of the United States and much of the developed world is an uncomfortable possibility. If that happens, turf wars will erupt all over the world, the world economy may collapse, the KGB and rogue Japanese (Yakuza and Aum Shinrikyo) will move and unleash further destruction, Pakistan and India are likely to erupt into a nuclear war, and several other regions may also erupt into nuclear war and use weapons of mass destruction. What China might do in such event is anybody's guess.

In short, the Armageddon we have so long feared, with the resulting collapse or destruction of much of civilization, could well be approaching. Let us hope that dark picture is terribly, terribly wrong, and that we successfully avoid Armageddon one more time!

Meanwhile, we would be in much better shape to weather the coming storm had our scientific community not grossly betrayed us. But if and when millions of Americans are dying in the streets with no effective means of treatment, and no effective defenses to prevent onslaught after onslaught against our nation, all those deaths and all that suffering will be directly the fault of our own organized scientific community.

Rogue Japanese (Yakuza and Aum Shinrikyo) Have a Role

In the information briefing furnished the previously named agencies, we briefly detailed the elements of the attack, how our own nuclear weaponry can be dudded anywhere, etc.

We also detailed the strange role of the Japanese Yakuza and the Aum Shinrikyo, who leased the operational use of the early generations of KGB longitudinal wave interferometry weapons in 1990 directly from the KGB. The first down payment on the lease was $900 million. I previously furnished pertinent information on this to an investigative U.S. Government agency. Since then, the Yakuza have set up their own scalar interferometry weapons development facilities in Japan, including portable energetics weapons,[35] and have assumed a very secret and nearly unsuspected role somewhat similar to Bin Laden's role in the attacks of September 11, 2001—but much more deadly when the coming Armageddon erupts.

Inducing the Gulf War Disease

In 1986, the Russians had already solved the dead man fuzing problem theoretically, but had not yet completed the development and deployment of the necessary QP weaponry. The KGB's first "combat demonstration" test of the **ability** to localize a QP in a selected region of space was against the U.S. air attack on Libya in 1986. That test apparently used a laboratory apparatus to prove the process. In short, it was a major milestone whose success moved the system into full production engineering and production. The system was then developed at very high speed after the successful demonstration in 1986. The first strategic KGB QP weapon(s) became operational on site in latter 1989.

---

[35] We strongly suspect some of these portable weapons have already been infiltrated into the United States—e.g., to be used against nuclear power plants and other lucrative targets such as refineries, oil fields, etc.

It was this QP weapon that was used to generate the so-called *Gulf War Disease* in our soldiers in the MidEast. In the action package, I explained precisely how that was done. I have also finally placed the simplified gist of that explanation openly on the Internet for all to see (website www.cheniere.org). Those suffering U.S. veterans were *casualties of war,* not victims of just nerve gas whiffs or chemicals or their own stress, etc. The initial U.S. Government and U.S. scientific community's assertion that "it was all in their heads" and due to stress, insults every veteran who has ever served his nation. One justifiably has contempt for such treatment of our veterans. Sadly, the latest "studies" are **beginning** to conclude the same thing once again.

Apparently our intelligence and military command agencies have not yet learned t**he** lesson of the decades long Soviet microwave radiation of the U.S. Embassy in Moscow, which was responsible for health changes of many kinds in U.S. personnel and was also responsible for the deaths of three U.S. Ambassadors to the Soviet Union. Apparently we are still using radar engineers and electrical engineers for electrodynamics analysis, and these analysts apparently still believe that in the absence of the EM fields there can be no induced EM effects. One sadly comments that, just as the electrical engineering departments of our universities have been unable to walk across the campus to the particle physics department and find out what the discovery of broken symmetry in 1957 portends for their discipline and model, then the intelligence community apparently has not yet learned that no EM analysis of such things as the microwave radiation of the U.S. Embassy in Moscow or the shootdown of the TWA-800 is worth a tinker's dam unless it is a higher group symmetry EM analysis. As Barrett showed, our fellows still have not grasped what Nikola Tesla already accomplished in actual patented circuits circa 1900 and shortly thereafter. Tensor analysis and vector analysis simply will not show what is really there.

Our veterans were hit with a "cocktail mix" of specialized disease "spacetime curvature engines" just below the observable threshold, in the "shadow" (virtual, but nearly observable) state[36] (Figure 53). The immune systems and the cellular regenerative systems do react to that shadow state level.

---

[36] One should think deeply about this, in terms of the solution to the source charge problem (which our engineers have not thought of very much, much less solved). Every charge already "engineers" the virtual stale at will, converting virtual energy into deterministic observable EM energy. That means we can directly engineer the virtual state vacuum itself if we put our minds to it. **Actually** Nobelist Lee has already pointed out that we have *not even tried to do this!* See T. D.

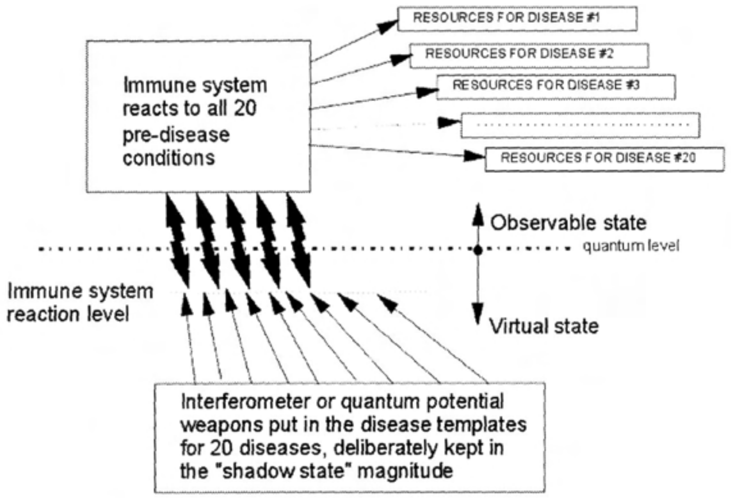
RESOURCES FOR DISEASE #1
RESOURCES FOR DISEASE #2
RESOURCES FOR DISEASE #3
RESOURCES FOR DISEASE #20
Immune system reacts to all 20 pre-disease conditions
Observable state
quantum level
Virtual state
Immune system reaction level
Interferometer or quantum potential weapons put in the disease templates for 20 diseases, deliberately kept in the "shadow state" magnitude

Figure 53 Spreading (thinning) the immune system's ability available to meet a new pathogen.

Such a "simultaneous mix" shadow induction "spreads" the reacting *finite* immune system and regenerative system (Figure 53) and their resources all across that entire mix of shadow diseases, so that only a much smaller portion of the immune system and of the regenerative system is directed against *any one* of the shadow diseases. This in turn greatly heightens (by a factor of 5 or more) the susceptibility of the body to all additional disease pathogens, and it greatly lowers the ability of the body to recover from any disease or biological warfare strike. In addition, the exposed veteran's body fluids and masses have accumulated a persistent "time-density charging" that very slowly decays (by emission of weak *longitudinal* EM waves) over months or even a few years.[37] The sperm from these veterans did indeed have the innate "engines" and therefore the capability to gradually infect their wives. So did the saliva in their mouths. The weak "dimensioned" LW radiation from their bodies could also slowly alter the cellular regeneration and immune systems of their family members in close proximity, over a period of time, thus inducing the disease in them {118,119,120}.

Lee, *Particle Physics and Introduction to Field Theory,* Harwood, New York, 1981, in his "Chapter 25: Outlook: Possibility of Vacuum Engineering," p. 824-828. On p. 383 Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge) *is not utilized.* We add that the West does not utilize it. It has been utilized for some decades in the former Soviet Union and now in the continuing KGB energetics and bioenergetics weapons programs.

[37] See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?." *Phys. Lett.,* 122B(3,4), Mar. 10, 1983, p.217-220.

The induction of GWS in our troops was in fact a test for the *present* clandestine use—already ongoing—of that technique to spread the immune and regenerative systems of the targeted U.S. population. This in turn gradually makes us more susceptible to terrorist BW attacks and disease pathogens of all types *from anyone else.* This tactic is already in place and accomplished, and it will substantially increase the American civilian casualties in our cities when they are subjected to terrorist BW attacks in the near future. *Those attacks now will have some five times the effectiveness that our present planners estimate.* In short, it is going to insure the deaths of millions more Americans. The approaching war in the MidEast, with Saddam Hussein's unleashing of his terrorist assets in the U.S. to the maximum extent possible, thus will probably see a five-fold increase in the effectiveness of some of his mass destruction weapons that Hussein's assets are able to unleash on America.

The entire affair of the Gulf War Disease was both a preliminary test and also a magnificent intelligence probe by the KGB, to clearly demonstrate once again that the U.S. Government and the U.S. scientific community are still nearly totally unknowledgeable of QP weaponry, immune and regenerative system "spreading," disease induction by bioenergetics means, etc. It follows a methodology used with great success for four decades by the KGB,[38] and one that our own government, scientific community, and military community has yet to decipher. *So long as the threat only slowly increases, Americans remain complacent in their ignorance, ignoring the boa constrictor slowly strangling them, while awaiting the cobra seen in the distance?*[9] We have unwittingly cooperated with the KGB boa constrictor for four decades, even as it slowly prepared to crush us. In the Gulf War, that crushing moved to the next threshold, to actually inflict casualties and deaths on U.S. troops and totally get away with it because of the bioenergetics (in the Russian sense) ineptness of the U.S. scientific community.

---

[38] The decades-long microwave radiation of the U.S. Embassy and induction of health changes and diseases, was a case in point. A high level target—the U.S. Ambassador—was involved, guaranteeing U.S. Government attention at the highest level. Puzzled, the U.S. government would then call in their most trusted EM scientific analysis groups. By the responsive actions then taken by the U.S. at the Embassy, the KGB could ascertain with absolute certainty whether or not our scientists knew of the use of infolded engines in bioenergetics. For more than four decades our actions assured them with 100% certainty that we knew nothing of what their "information content of the field and potential" actually was, or how it had been weaponized and was being used right in front of our faces and before the entire world.

[39] It is bizarre that our scientific community—even after those decades of EM induction of diseases and health changes in American personnel in the U.S. Embassy in Moscow—*still do not and cannot believe that conditioned EM radiation can induce cellular disease.* Such total blindness is inexplicable.

40-Years of Russian Induction of Diseases In U.S. Embassy Personnel

Sadly, our own fellows have never even figured out how the diseases and health changes were previously induced by the KGB for decades in personnel in the former American Embassy in Moscow. Three U.S. Ambassadors to the Soviet Union died as a result of their exposure to that radiation. Many other Americans on site were made ill or suffered health changes.

When Johns Hopkins scientists studied the problem, they found a most anomalous thing: no health changes occurred in areas *where the EM forcefields existed* in the microwave radiation. Instead, the health changes only occurred in areas where the potentials were gradient free so that the force fields were zero—and thus the potentials (and their specific internal "engines") were *persistent.* Using only a classical EM force field analysis—which does not show any of the internal longitudinal EM wave substructuring of the fields, potentials, and waves—the Johns Hopkins scientists erroneously concluded that it could not have been the weak microwave radiation causing the health difficulties, *because no such health difficulties occurred where the forcefields were present.* That is, they very naively assumed that "no force fields" meant no EM effects were possible. Their entire findings thus were based on a total non sequitur, the assumption that there can be no electromagnetically induced physical changes in matter in the absence of fields! The Aharonov-Bohm effect alone falsifies that entire premise. So does general relativity and higher group symmetry electrodynamics.

Actually, the Johns Hopkins scientists had just clearly proven the exact opposite: *that it indeed was the radiation that induced the health changes.* That is simply shown statistically, and it is a deep and continuing mystery how good scientists and their governmental overwatch did not notice the violation of inference methodology.

Simply put, there was 100 percent correlation to the presence of *gradient free EM potentials— the persistent kind.* There was zero correlation (or 100% anticorrelation) to the presence of the normal electric and magnetic fields—i.e., where the potentials changed over a very small region of space. Hence it could *only* have been the constant potentials, else we must throw out statistical inference theory altogether.

If the changes had *not* been caused by the EM radiation at all, one would have had some health changes in regions where fields were present and some health changes where the fields were absent. Considering the lengthy time the radiation persisted and the number of personnel affected, the statistics are decisive. Without question, the cause was indeed the persistent potentials where the **force** fields were zero. Whittaker's two papers—one in 1903 and the other in 1904-would then show how it was done. The 1904 paper and superpotential theory alone would *not* have explained it, since that alone does not include the "information content of the field"—i.e., spacetime curvature engines infolded inside the normal EM potentials, fields, and waves.

So, the U.S. government and U.S. scientific community "botched it badly," unfortunately, by elemental statistical inference standards. And unfortunately, they have continued to "botch it" to this day. Both the scientific community and the intelligence community missed it. Elementary statistics theory alone, coupled with the Johns Hopkins results, clearly indicates that the microwave radiation was causing the health changes, to a very high degree of confidence.

This demonstrates how "totally inexplicable" some of the effects generated by longitudinal EM wave interferometry and by "structured" or dimensioned potentials, fields, and waves can seem to our conventional scientists. If we examine the serious foundations flaws in classical electrodynamics and electrical engineering taught in our universities, and the near-total lack of higher group symmetry electrodynamics training, we can understand why our scientists and intelligence analysts did not and do not have the background required to properly assess the incidents and effects.

Western Electrodynamics Is in Sad Shape

Actually, numerous scientists have long pointed this out. For example, quoting Mario Bunge {121}: *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

In quantum electrodynamics, it is the *potentials* that are the primary causes of all electromagnetic phenomena, not the force fields as used by Johns Hopkins and the Government in its analysis. We know that classical EM is wrong in its focus upon force fields as primary EM causes; the Aharonov-Bohm effect, Berry phase, and geometric phase work in some 20,000 papers in the physics literature already show that. Indeed, we also know that the classical force fields do not even exist apart from mass; in space, only the potential for such a force field exists, in case some charged matter is brought in to interact with the EM entity as it exists in mass-free space (as curvatures of spacetime).

The Aharonov-Bohm effect and its implications have been published in *Physical Review*—our most prestigious physics journal—since 1959. Later publications have included the Berry phase extension of the AB effect, and Aharonov and Anandan's further extension of the Berry phase to what is today called the geometric *phase*. The potentials are indeed the primary EM causes, and *the force fields do not even exist as such in mass-free space, but only in mass.* Yet our classical EM taught in our universities has not been changed in its fundamental acceptance of the material ether since 1865.

Every university in America is still teaching and defending a rather mangled electrodynamics where

(1) The material ether is still assumed, even though destroyed more than a

century ago by the Michelson-Morley experiments;

(2) The force fields are erroneously assumed to exist in massfree space, but in force-free form(!!!) when what actually exists there as an EM disturbance is a general relativistic perturbation of spacetime (i.e., an oscillation of the local curvature of spacetime). Actually, force identically is the time rate of change of momentum, where momentum is mass times velocity. Thus mass is a *component* of force and force field. There is no force or force field without mass. As Feynman put it, there is no force field in massless space; instead, there is only the potential for such a force field to be formed at the various points, should a charged mass be placed at (added to) each of them;

(3) The source charges are erroneously assumed to *create* their fields and potentials and all that energy in them, reaching across the entire universe from each and every charge, right out of nothing at all. This perpetual motion and creation of energy error continues even though particle physics has known (and proven, both theoretically and experimentally) that any charge (as a set of composite dipoles, when the charge is taken with its clustering virtual charges of opposite sign) is a broken symmetry in its fierce energy exchange with the vacuum. Hence the charge extracts that energy in its fields and potentials *from the active vacuum,* and outputs the extracted energy across space {122}.

Our own electrodynamicists and electrical engineering professors and departments would have us believe - and they continue to imply such in every university in America—that every charge is a perpetual motion machine clearly in violation of the master energy conservation dictum that *energy can neither be created nor destroyed.* The complete lack of scientific understanding that every charge extracts energy from the vacuum and freely and continuously pours it out as real observable EM energy, is reflected in rather silly "witch-hunts" {123} against electrical systems which exhibit more energy output than the operator himself inputs—i.e., systems which exhibit a coefficient of performance (COP) of COP>1.0, which are open EM systems far from equilibrium in their active environmental exchange.

Let us reason together! If the second law of thermodynamics really did prohibit COP>1.0 *electrical* systems, then it also would prohibit the demonstrated COP = oo of a solar cell array. It would also prohibit the demonstrated COP>1.0 *mechanical* systems such as a windmill, waterwheel, or sailboat—all of which have COP = oo! In other words, the second law cannot even be a law if it discriminates only one type of energy prohibited from being used for COP> 1.0,

For some inexplicable reason, EM COP>1.0 systems have come to be dogmatically labeled "perpetual motion machines" by the professional skeptics since EM systems must take their excess energy from the known active vacuum environment! In short, the dogmatic scientific community is saying that "windmill-type EM systems" with COP = oo are permissible for exchanges with *physically active and observable* environments (such as the sunlight used in solar cells, the heat in the air used in heat pumps, the water currents used hydroelectric power plants, the winds used in windmill power generation, etc.), but they are not permissible for exchanges with the proven modern active vacuum *virtual particle flux energy* environment, even though the discovery of **broken** symmetry in 1957 already proves that all EM energy is indeed freely taken from the seething vacuum environment by the asymmetry of the modern dipolarity of the source charge! We refer the interested reader to the proven broken symmetry of opposite charges {124}. We also refer the reader to Nobelist T. D. Lee's statement applicable to the skeptics' naive notion of symmetry. Finally, in modern physics all the major forces of nature are generated by the exchange of virtual particles. It is these forces that then translate against a resisting mass or entity to provide useful work. So all useful work in the universe already comes from the interactions of the seething virtual flux of the vacuum. On what distant planet have the skeptics been hiding while all this modern physics has been developed?

One must keep one's sense of humor! It is ironic but true that every major university in the U.S. continues to teach, uphold, and advocate—by implication—EM charges as ubiquitous perpetual motion machines, freely creating energy from nothing. And that is those professional skeptics and witch-hunters never mention that their own naive understanding of electrodynamics already implicitly assumes and teaches perpetual motion machines freely creating energy from nothing at all—and on a scale unparalleled in scientific history!

We have one chart with some 32 major foundations errors and flaws in the electrodynamics taught in all our schools; there are many others in addition. This erroneous teaching has so indoctrinated our graduate students that most electrical engineers no longer know the real difference between power (the rate of change of form of energy and thus the rate of production of work) and rate of energy flow (whether it is being dissipated or not).

Every textbook speaks of "sources furnishing power," when it is the rate of dissipation of energy in the individual loads and losses to provide the rate of doing work that *constitutes* power. Sources do not furnish power! They furnish energy flow (specifically, their source dipole furnishes energy flow, extracted and transduced directly from the active vacuum, and only a tiny bit of that enormous energy flow is diverged into the circuit and used). The various components of the receiving circuit interact with that energy flow to catch and dissipate a little of it, thus producing power and work locally in the intercepting and dissipating component.

A massive nondivergent (nonintercepted) energy flow has zero power (time-rate of doing work), but if intercepted and diverged, the divergence may develop enormous power (time rate of doing work) in and upon the interceptor, diverger, or collector. Even many physicists today are confused on the difference between rate of work and rate of energy flow per se. E.g., for about 30 years now, there has been a polite debate involving that very subject in the *American Journal of Physics.* It is still not completely resolved, because the argument has not actually addressed where all EM energy in fields, potentials, and waves comes from in the first place: directly from the seething vacuum via the asymmetry of the source charge. It also has not addressed the long unaccounted giant Heaviside nondiverged energy flow component.

Toward a Deeper Electrodynamics

The first thing the Soviet Academicians did after WW II was to completely overhaul electrodynamics, but in highly classified areas controlled totally by the KGB. Stalin's boot firmly planted on their backside saw to it, creating the terrible fear that if they failed to produce, they would surely die. Not having experienced such a "Stalin boot," our own scientific community continues leisurely onward, while it teaches 110 years old electrodynamic material filled with non-sequiturs.

The Russian use of the "information content of the field" euphemism is a case in point. Our fellows (in both the intelligence community and the scientific community) assume the Russians are speaking of simple transverse EM wave spectral analysis. Nothing could be farther from the truth. Let me explain.

In 1903, E.T. Whittaker—a leading theoretical physicist at the time—showed that any scalar potential is comprised of a harmonic set of bidirectional *longitudinal EM wavepairs.* Each pair consists of a longitudinal EM wave and its true phase conjugate replica (i.e., its "antiwave", or time-reversed twin, coupled together).

Unknown to Whittaker, that LW wavepair actually includes a *time-density wave* in the time domain, coupled to a longitudinal EM wave in 3-space. Per quantum field theory {66}, such a coupled pair would be observable as the scalar potential. Knowledge of this specialized coupling is necessary in order to properly decipher the Priore mechanism. That mechanism was demonstrated to *time-reverse* (dedifferentiate) a diseased or damaged cell and all its parts back to it previous normal state, *genetics and all.* The time-reversal was accomplished by "pumping" the cell and all its parts in the time-dimension so that time-reversal propagation of the mass-energy is achieved. The time-domain pumping is done by the associated time-density wave, which always accompanies any longitudinal EM wave a priori. So sometimes we sloppily say that the "longitudinal EM wave pumping" causes the time reversal of the mass, but it is really the associated time-density wave pumping that causes it. The time-density wave component also can cause *time charging* or *time-excitation charging* of a mass. During its alternate existence as masstime, the time portion is altered, and that is a time-excitation or time-energy excitation.

It turns out the human cellular regenerative system (Figure 29) also uses the same time-excitation charging methodology to produce time-pumping of the cell and all its parts. This produces a slow time-reversal (dedifferentiation) of damaged cells back to a previous state. This is in fact the basic mechanism for healing, as used by the living body.

Reversing aluminum, for example, just converts it back to ordinary aluminum. Not too interesting.

Reversing an HIV infected cell, however, would produce a healthy, normal cell with normal genetics. When a damaged cell turns into a cancerous cell, it is due to a process called *promotion* (change of the cell to a cancerous cell). *Reversing a* cancerous cell would change it back to a healthy, normal cell again—in short, it would constitute *demotion* of cancer and then healing of the damage in the now normal but damaged cell. In the AIDS patient, simply pumping the entire body with the time-density wave accompanying the longitudinal EM waves produced in a plasma—as Priore did for terminal tumors—would time-reverse all cells, completely ridding them of any residue of HIV infection. This would remove all the "HIV factories", thus completely solving the problem of AIDS. We have also uncovered a more advanced process where *matter can essentially be directly engineered in any conceivable manner,* including nuclear transmutation etc. We have filed a Provisional Patent Application and will file formal patent applications on those processes in the future, as soon as funds are available (patent attorneys are expensive!)

Time Is Highly Compacted Energy

Since time is energy compacted by some $9x10^{16}$, even small time-density changes can produce enormous spatial energy, should any of that compacted energy be *transduced* from time-density waves into transverse EM waves and thus into spatial EM energy. Such transduction can indeed occur in nonlinear materials and processes, such as in electrolysis.

The New Transduction Model Explains the Cold Fusion Results and Mechanisms

In August 1998, we presented a conference paper pointing out that cold fusion uses this very *transduction of EM wave types,* one into the other. We presented such new concepts as (i) TW to LW to TDW transduction and in the other direction also, (ii) time density charging of a particle to an excited time-density state (analogous to energy charging to an excited energy state), and (iii) the observation that excess phase conjugation in a local region can generate a time reversal zone where like charges attract and unlike charges repel.

With these concepts we were able to explain the excess heat appearing in the cold fusion experiments of hundreds of researchers. We also came up with (iv) a totally new class of nuclear reactions without the usual high-energy considerations. That is, we showed the exact nature of the electronuclear reaction that is forming the anomalous nuclides in cold fusion experiments. Production of the new nuclides was revealed to be simply what one would expect from this mechanism, as indeed was particle clustering of like-charged particles.

We also explained (v) all the instrumental anomalies occurring in a series of well-documented cold fusion electrolysis experiments at China Lake. In fact, (vi) we added a completely new term to the conservation of energy law, thereby extending it.

An updated explanation is included in my new book, *Energy from the Vacuum,* 2002, *ibid.* With the new transduction interactions, one must state that the total energy in (a) spatial energy form, (b) mass-energy form, and (c) time-energy form is what is conserved.

So when transduction occurs in cold fusion experiments (electrolysis), some of the time-density charge energy is gradually transduced into ordinary transverse EM wave energy, which as it scatters produces excess heat not accounted for in present particle physics interaction models.

Scientists with only the "old" 2-term Einstein conservation law in mind, will be firmly convinced that this excess heat represents a violation of conservation of energy, hence is impossible. So they will question the calorimetry ability of the cold fusion scientists, etc. Unfortunately, the more ardent defenders of the orthodox faith will unjustifiably impugn the intentions and morals of the experimenters themselves with vicious ad hominem attacks. Such has indeed been precisely the case. Many of the harsh scientific critics have acted like cur dogs in a pack attack, and not as scientists at all.

Once one analyzes cold fusion in the new 3-term conservation manner and with the new transduction concepts, one immediately sees that the *overall* conservation of energy is upheld, while the *special case* represented by the former energy conservation law (conservation of the sum of the spatial energy and the mass-energy) is *permissibly* violated. The former case was a special case in that it arbitrarily assumed that no transduction of time-energy into spatial energy or mass energy—or vice versa—occurred.

In essence, successful cold fusion experiments show that, via transduction, a bit of time itself can be "utilized for fuel" and transduced into ordinary EM energy. Energy-wise the system has become a nonlinear open dissipative system[40] far from thermodynamic equilibrium {125}. As such, it is permitted to (1) self-organize, (2) self-oscillate, (3) output more energy than it itself has, by merely receiving, converting, and using excess energy from its environment (in this rase, from the transduction of a tiny bit of the time-charging that has occurred), (4) power itself while simultaneously powering a load or performing useful work, and (5) exhibit negentropy. With the exception of transduction of time-charge into excess energy, all of that is already in the thermodynamics texts, though it has not been clearly spelled out in electromagnetics texts. The source charge, e.g., exhibits all five of those capabilities. *The inclusion of the time-domain as part of the external active environment, however, is an extension of* ***previous*** *open dissipative system work by leading researchers such as* ***Prigogine.*** *It dramatically extends the meaning of "active environment".*

---

[40] But we point out one idiosyncrasy of thermodynamics, which defines a "closed system" as one in which mass does not cross the system boundary, but energy is free to exchange across that boundary {125}. Hence, eerily a "closed thermodynamic system" is actually an open system with respect to energy exchange between system and environment. As such, energy-wise the system can be a system far from energetic equilibrium with its external active environment, and still be considered a closed thermodynamic system! Energetically, a *closed thermodynamic system* that exchanges energy but not mass is permitted to exhibit COP>I .0 and even COP = oo!

Why There Is a Time Delay in the Onset of Cold Fusion Phenomena

With an energy-to-time compaction ration of 9xl0^{16} , energy processes have to transduce spatial energy into time for quite a while in order to build up very much *time-excitation charge*. This leads to an initial delay in the onset of the "new law" transduction phenomena, as directly shown by the vast majority of careful cold fusion experiments. This build-up occurs during the "palladium-lattice loading" time, e.g.

By the same token, there is a time delay after cutoff of the input power to such experiments, during which time the slowly decaying time-density charge by its transduction into spatial energy continues to maintain the anomalous new processes (such as excess heat production) as they slowly subside. So the hours-long or even days-long delay phenomena after input power shutdown are also explained by the new approach.

In summary, there exist some very anomalous but very good experiments by excellent scientists—more than 600 of them—that are at least totally consistent with the above new wave polarization transduction model. Consistency with experiments is the first element of proof (it is necessary but by itself it is not sufficient). The new model also contains and obeys the conventional conservation of energy law until the new transduction phenomena become observable, at which time the extended energy conservation law applies. So in its lower limit, the new model corresponds to the present model—again, another requirement of any model purporting to extend a scientific area past that area's conventional model description.

The KGB weapons scientists—not the regular Russian academic establishment—have about a 40-year lead on us in all these areas. While our own scientists have dallied and pursued the standard model, the Russian weapon scientists have produced the equivalent of 10 Manhattan Projects, back-to-back, in the "new" energetics area.

The Early Soviet Longitudinal Wave Interferometers

Using longitudinal EM waves, the early KGB energetics interferometers could pass their beams right through the earth and ocean. LWs readily penetrate matter, reacting with it only weakly.

Some LW interferometry was also done using ordinary EM waves as carriers for the LW wave functions. Each such transverse wave carrier is in fact two scalar potential functions (Whittaker 1904, the initiation of superpotential theory). In each scalar potential for which functions are extracted and used to make the normal transverse carrier wave, there exist LW biwave pairs comprising said potential. So by assembling the desired LW biwave subsets, the two scalar potentials and their resulting interference-produced carrier wave are internally structured as desired. This is the use of the so-called *internal information content of the field*. We speak of such a deliberately structured potential, field, or wave as having been *conditioned*, *dimensioned*, or *activated*. If these terms are deemed inappropriate, just use the more generic term "internally structured." Then one has to explain what the "internal" is, and how this "internal" can be "structured."

In the distant interferometry zone, those *dimensioned* potentials, fields, or carrier waves would interfere and produce the desired alterations in the local spacetime (including specific spacetime curvature engines). Remember, the potentials are just bundles of special waves. We are discussing a special form of nonlinear multiwave interferometry, where the individual waves and waveforms comprising the total wave package assembly constituting a carrier have been carefully chosen in advance.

In the local spacetime in the distant interference zone, such multiwave longitudinal EM wavepair interferometry creates ordinary EM waves and energy that arise directly from each and every point in that individual interference zone. We stress that ordinary transverse wave EM energy propagation through the intervening space between the interferometer transmitters and the distant target *does not occur.*

In January 1960, in a speech to the Presidium, those are the weapons then in final engineering development for production that Khrushchev spoke of as "fantastic new weapons".

Khrushchev Killed the Thresher to Prevent His Own Dismissal

The first deployed large force-structure interferometer weapons became operational circa 1 April 1963. Khrushchev was still smarting from his Cuban missile crisis fiasco and his humiliation by Kennedy. Having overplayed his hand and gotten caught,[41] Khrushchev was desperate to stay in power and not be thrown out by the Communist Party leaders for such a massive loss of face.

On April 10 of 1963 Khrushchev used his new interferometers to reach through the earth and ocean and "kindle" EM jamming in the electrical controls of the **U.S.S.** Thresher nuclear submarine underwater off the East Coast of the U.S. **The** reader should recall that, wherever the interference zone is established—including in the earth or underneath the ocean, in the atmosphere, or in space—the energy arises directly out of the local spacetime. *There is no Faraday shielding of such effects, which do not depend upon ordinary transverse wave EM field energy propagation through space.* The energy appears everywhere within the targeted object, simultaneously.

It is extremely important that the reader recognize the dramatic difference in the way energy propagates in the new mode as compared to the old "propagation through space" mode. In the old mode, energy transport arises in one spatial location and is considered to proceed serially through the intervening spatial points to the intercepting and interacting particle's location. The Poynting model presently uses only this mode. In other words, the energy propagates from wi*thout* the object *to* the object, and thence *inward inside* the object, if possible as the object interacts.

On the other hand, when the energy arises at each and every point in spacetime, it then propagates outward from each arising point. Thus in the nucleus of an atom, the "energy flow" starts at all interior spatial points and flows or diffuses outward in all directions. So the energy "meets" and interacts with the quarks in the nucleons, *before* it meets the "nucleon as a whole" and translates it. For this reason, with energetics it is easier to alter and "flip" quarks in nucleons than it is to translate nucleons, since one is using the new energy propagation mode. *This simple reversal of the outside-to-inside energy propagation to an "inside-to-outside" mode, is a revolution in particle physics and effects on materials.*

In fact, the energy dramatically enters the "interior" areas of nuclei, nucleons, etc. in a manner heretofore impossible or extremely difficult to achieve by high-energy physics, which can only do it by smashing into the object with sheer brute kinetic energy. Therefore the new "inside to outside" energy propagation mode can produce—at feeble spatial energy—extraordinary nuclear interactions not possible by brute force "outside-to-inside" energy propagation means. By controlling the structuring of the arising of the energy at each point in spacetime, the energy structure interacting with the quarks and other interior components can **b**e tailored to order. This is the primary secret of transmuting an element at a distance by a specialized energetics weapon. The structuring of the emerging internal energy can be deliberately tailored to change quarks in—say—some of the nucleons of U235 and plutonium nuclei. In this fashion nuclear warheads, nuclear powerplants, and nuclear propulsion systems of atomic submarines can be altered at a distance, beneath the earth's surface, beneath the ocean's surface, in space above the earth, etc.

---

[41] Oddly, had he simply waited a few more months or a year before injecting the missiles and nuclear warheads into Cuba, Khrushchev's strategic scalar interferometers would have become operational. With those in operation, a "Cuban missile crisis" might well have had a different outcome.

The EM "hash" emerging from the Russian interferometry simply jammed the electrical controls and systems of the Thresher *from inside its very electrical circuits.* Out of control and helpless, the Thresher sank to crush depth and imploded (Figure 6). The EM "splatter" from that interferometry, arising from the local spacetime throughout a surrounding region, engulfed the Skylark, the surface companion of the Thresher. The resulting spurious EM jamming severely interfered with multiple electronic systems on the Skylark. Several electronic systems just ceased working altogether. Their electrical forces were time-energy excited and time-reversed so that electron currents in them became impossible (blocked).

It required over an hour for the Skylark to transmit an emergency message back to the Navy that the sub was in difficulty, through all that interferometry jamming effects. The Russians undoubtedly "tracked" the sub till it imploded, then cut off the interferometry. As the "time-density charging" in the area slowly diminished, all failed electronic systems in the Skylark returned to full functioning. This was because the excess time-energy excitation just decayed away, so that electrical currents could again flow.

Khrushchev Demonstrated the Powerful Blast Capability of Strategic Interferometry

The next day, 100 miles north of Puerto Rico, the same KGB interferometer weapon was powerfully pulsed to produce a giant EM burst deep under the ocean's surface (Figure 42). A cone of water rose up from the surface to about a half-mile in height, observed by the startled crew of a passing U.S. jetliner and reported to the FBI and Coast Guard when the aircraft landed in Florida. As it rose, the cone of water spread into a cauliflower shape typical of a deep underwater nuclear burst that does not vent the gas bubble. Then that giant cauliflower of water fell back out of the sky and into the ocean. The second successful demonstration of Khrushchev's scalar interferometer weapon had been made for the benefit of the Presidium and the leaders of the Communist Party.

With those two dramatic tests, Khrushchev had killed the latest and greatest U.S atomic submarine and demonstrated the awesome power of energetics weapons. He had shown the leaders of the Communist Party that the U.S. had no inkling of his new superweapons. He had struck the U.S. a formidable blow in a direct act of war—but a new kind of *clandestine* energetics war—and gotten clean away with it. Now it was Kennedy, rather than Khrushchev, who had been duped. So had the entire U.S. Intelligence Community and the U.S. Scientific Community. Sadly, mostly they apparently remain duped or partially duped to this day.

Khrushchev managed to stay in power nearly two years longer by these two demonstrations. So the KGB/Soviets embarked on a decades-long program of buildup and development of the new weaponry, clandestinely, in order to take the U.S. by surprise attack and destroy it utterly.

In 1972 Brezhnev named 1985 as the target year when Russia would be able to do as it pleased, and would essentially control the land, the oceans, and space {126}. This was stated at a secret meeting of the European Communist parties in Europe. A British intelligence agent was in the meeting, and furnished the U.S. that information. Later it was openly released and published.

The Russians met that 1985 deadline, and the resulting increase in killing U.S. missiles and aircraft in 1985-1986 was most indicative. Other indicators such as massive weather engineering were already ongoing.

Weather Engineering Over North America

The first KGB weather engineering tests over the U.S., using their relatively new interferometers, produced signatures of anomalous perfectly round holes appearing in clouds. These experiments started in 1967 or thereabouts. The Russians gave us that very severe "deep freeze" winter of 1967, as an initial weather-engineering test of their energetics interferometry weapons.

The KGB communists do have a sense of historical dates, and they often do things on certain dates in symbolic fashion. Full-time Soviet weather engineering over North America started in earnest on July 4th, 1976—our Independence Day and our bicentennial. This was the KGB's "gift" to the United States on such a historic occasion. Full-bore weather engineering over the Americas has continued to this day, and even increased. It was passed into the hands of the Yakuza and the Aum Shinrikyo upon their leasing of those interferometers on site in Russia at the end of 1989. In 1985 I produced a rather crude videotape[42] (40 minutes) with pictures of the giant radial cloud signatures etc. associated with such weather engineering operations since 1976.

The *basic* weather engineering method is simple. Use an LW interferometer (LWI) to reach through the earth and *cool* the air in a region and produce a high-pressure area, due to the densification of the air. Use another interferometer to *heat an*other area and form a low-pressure area, due to the expansion of the air. Then carefully steer those *highs* and *lows* by gradually and slowing rotating the distant antennae and slowly adjusting the interferometry range simultaneously, according to a calculated schedule.

---

[42] T. E. Bearden, "Soviet Weather Engineering Over North America," 1985.

By that means, the LWIs easily catch, entrain, and steer the jetstreams largely responsible for our weather. Put sharp loops and twists in that steering, and the huge angular momentum of the jetstreams developed in those sharp turns and circles will spawn tornadoes, violent weather, etc.

The first complete weather engineering system over North America in 1976 also used the Woodpecker's normal EM carrier signals with dimensioning and internal assemblage of the composite longitudinal EM wavepairs. A giant "interference grid" was established over much of North America. A great deal of local adjusting of the interferometry in the grid was necessary. These local grid readjustments caused "sudden popouts" of EM energy, creating many atmospheric booms, rumbles, and explosions.

To really stir up big weather trouble, the Russians—and now the Yakuza and Aum Shinrikyo—also heat or cool areas of the ocean where El Nino and similar "water engines" form. A dimensioned LW interferometer is required for this operation. By slowly and protractedly heating or cooling the water in one of the "warm water engines" fueling the weather, one causes a very significant, long-term, later effect on perhaps an entire continent.

One can create great droughts in some areas, great floods in others, etc. The KGB has been doing this extensively, since 1976. The great floods in Red China a few years ago are a direct example of KGB/Yakuza weather engineering, as is much of the extreme heat of 1998. In fact, thanks to the weather operations of the Yakuza under their KGB mentors, 1998 was the overall hottest year on record till then, since records were kept. At least for the summer of 2002, the record may have been broken again, with a significant season of giant forest fires burning away.

One purpose of the excess heating of the atmosphere, of course, is to do damage. An added "benefit" to the KGB is to furnish support to the environmentalists who are concerned about global warming. Understand, there is indeed such a thing as global warming. But by augmenting the atmospheric (and ice sheet) heating just a bit in a single year, the statistical analyses are skewed. The skewed analyses will show that the global warming problem is far more imminent than it really is, if the LWI engineering of it were suspended. Nobody on either side of the global warming debate seems aware of the LWI intervention that is ruining the results of their models and calculations.

In turn, the heightened global warming indications, detected as a result of the LWI interruptions, impel strong activist support for such things as the Kyoto treaty, *which is specifically designed to the strategic detriment of the U.S.* if and when the U.S. Senate ever approves it. The Kyoto treaty does not even apply to some 160 countries including Russia, China, etc. This focuses a certain proportion of our government time and effort in grappling with an inflated global warming problem, immediately. It directly fits the KGB strategy of "spreading" our government energy and attention all over the map, anywhere but upon the coming strategic energetics strike and upon energetics weaponry.

The die-hard section of the KGB strategists are chess players. *They will come at you anyway, anyhow, and anywhen that they can.* Everything that can be impressed to serve or support the overall objective, will be impressed and used. That is the overriding principle when analyzing Russian KGB strategy.

Khrushchev meant every word of it when he said to Nixon in their famous Kitchen Debate: *"We will bury you!"* Interestingly, to this day Americans and their government officials have continued to be deceived as to exactly what Khrushchev meant. He meant it literally.

Earthquakes to Order

According to U.S. Defense Secretary Cohen, electromagnetic induction of earthquakes is now recognized by the U.S. Department of Defense, considering his little-noticed statement in Georgia in 1997, which we previously quoted. An interesting exchange of induced earthquakes (Figure 54) had occurred in latter 1988 and early 1989. Here two sides seemed to be "counting rifles."

Opening bet (induced quakes)	
- Quebec	25 Nov 88
- Pasadena	3 Dec 88
Raise you	
- Armenia	7 Dec 88
Call - what do you have?	
- Malibu	8 Jan 89
Four aces	
- Soviet Central Asia	23 Jan 89
Clear *table (stress relief)*	
- Bakersfield CA et al.	various

Figure 54 Counting rifles with earthquakes: Put up or shut up.

As Secretary Cohen made his April 1997 statement obliquely recognizing the scalar interferometer weapons, the little friendly nation had just completed countering and averting a massive strategic energetics strike upon the United States and Western Europe, scheduled for May 1, 1997 —just three days later. A little more than two months earlier, a previously scheduled attack had also been countered and averted by the same friendly little nation.

Once one can reach into or through the earth and ocean and create energy in a desired interference zone, induction of earthquakes is simple. Use the interferometers to deposit energy in a fault zone over a period of time. Keep it up, and the piezoelectric plates build up enough mechanical stress to forcibly "slip" and produce an earthquake.

Put the energy in very slowly, and one gets a long "overpotential" effect where the rocks do not slip for a long time, while the energy builds to higher levels due to the static friction of the rocks. Then there occurs a rather rapid catastrophic slip. In short, one produces a large destructive earthquake. The KGB weaponry can produce earthquakes of the highest magnitude if they work at it.

If one wishes to induce a quake in an area where there are no known faults, then one simply focuses the interference zone there inside the solid rocks and proceeds to deposit the energy, preferably rather quickly (in hours instead of days). The entire rock will be expanded and it will fracture, producing an anomalous so-called "flat plate" earthquake.

Around the early 1900s, the "normal" rate of large earthquakes was two or three per year. That gradually increased after WW II. During one week in year 2000, there were some 20 major earthquakes around the globe. Obviously, someone is doing it, since this is far beyond what could be expected naturally. Presently that "someone" is mostly the Yakuza and Aum Shinrikyo, on site in Russia and leasing the LWIs. With the scheduled CSR attacks forcibly aborted, the KGB has the problem of restraining an increasingly impatient Yakuza and Aum Shinrikyo group, while keeping them paying the lucrative lease money. Consequently, inducing some big earthquakes is one way to "keep the troops occupied and busy." Shooting down the occasional aircraft or missile is another way.

I worked for 8 years or more with an inventor (Floyd Sweet, now deceased) who had an energy device {127,128} sensitive to *artificially* induced quakes of that fashion, even such quakes thousands of miles away. The device would sit right on top of a *normal* quake and not react. We could tell within some 8-10 thousand miles—and often even for greater distances— which quakes were artificial and which were natural.

So during that period I was able to clearly ascertain what was going on, and who was doing what to whom, and which quakes were normal and which were manmade. Several artificial U.S. quakes, the Armenia quake, and a subsequent big quake in Russia were particularly interesting in the mid and latter 1980s. The huge 1976 China quake was directly induced by the KGB.

Figure 55 Brewer briefs President Reagan on Soviet scalar EM weapons October 1984, then briefs National Security Council

In 1984, on my behalf Roy Brewer {129} had briefed President Reagan on these scalar interferometry weapons (Figure 55), then briefed the National Security Council. Let us hope that these high level briefings may have borne at least some fruit.

Readying the Weapons for the 1985 Readiness Demonstration Tests

In latter April 1985 I received an urgent call from my close colleague Frank Golden. In his avionics repair work, he suddenly found that his best calibration instruments were very "noisy." He replaced one expensive instrument, and the second was "noisy," also. A third proved equally noisy. This was unparalleled. It was not the instruments, but instead it was all the electronics to be repaired, that were emitting electromagnetic noise!

So Frank borrowed one expensive instrument and brought it home for the weekend. He altered it by a proprietary method so that he could detect energetics radiation. Voila! The entire earth was alive with Soviet energetics signals. The weak interferometry of so many signals impinging into the weakly interacting electronic systems to be repaired, had generated the electronic noise his instruments had detected.

In response to Frank's call, I immediately proceeded to his location. I personally observed on the oscilloscope, as detected by Golden, some 27 pairs of Russian energetics signals, each of the two in a pair being 12 kilohertz apart. (We have previously explained why two frequencies are used, and one "pretends" one transmitted the difference frequency). Each of these frequency pairs represented the entire earth held in entrained resonance. Each pair was an "energy tap" in the earth, extracting enormous energy from the heat of the molten core of the earth itself and outputting it as 12 kilohertz power for transmission to giant interferometry weapon sites.

Each "tap" was extracting energy sufficient to power perhaps 4 to 6 of the giant intercontinental energetics interferometer weapons. We were observing something like 100 to 150 massive weapons on line and radiating. A full-up strategic dress rehearsal—or preparations for a full-up strategic attack on the U.S.—was underway.

Gorbachev was still a "new man on the job" at that time. Further, 1985 was the 40th anniversary of the end of WWII, and a very important occasion for the Communists and in fact for all Russians. The May Day celebrations that year went on for several days, not just on May 1.

What we were seeing were preparations for a massive May Day weaponry demonstration in and around May 1, 1985. We were observing the preparation and demonstrated readiness of all that great armada of Soviet energetics weapons, to meet Brezhnev's 1985 schedule laid down back in 1972 {126}.

Longitudinal EM wave communications to deep underwater Soviet subs, the huge interferometer armada, and everything else—all were being readied and demonstrated. I can assure the reader that observing full rehearsal of preparations for the next World War was a most sobering experience.

After May 1, most of the armada of interferometers was stood down again because it had been demonstrated that the "ready schedule" had been met. All but one of the giant energy taps in the Earth were also stood down again. That one remained active.

The May Day report to the heads of the Communist Party that year must have really been something!

As usual, apparently our own scientific and intelligence communities did not detect the main activities, and would not have believed it if one told them. They certainly did not believe it when I published it in 1986 in *Fer-de-Lance,* but labeled it sheer fantasy or at best total conjecture.

It was neither fantasy nor conjecture. It was direct experimental measurement. It was in fact what used to be called *scientific method.* The problem with our intelligence agencies was that they operated behind *policies.* Once the *policy* is decided, then regardless of what the intelligence analysis indicates, anything against the official *policy* is just ignored. This neat little *policy* method of doing business has resulted in making our intelligence agencies some of the most political-minded and manipulative groups on Earth.

Meeting the 1985 Schedule for Readiness

In 1985, KGB/Russian preparations began in earnest for the energetics weapons readiness projected for that year by Brezhnev in 1972. Being full-bore tests, these actually went after destruction of vehicles and killing people. These were actual acts of war. The so-called "Cold War" was not nearly so cold as pictured in the news media!

Early that year, an LWI was used to cool the combustion in the engines of a jet liner off the coast of San Francisco, bound for Los Angeles (Figure 56). This flamed-out the engines. The stricken craft fell 30,000 feet before the time-density "cooling" charge drained away, sufficient for the pilot to restart his engines and limp into San Francisco. Gravitational and inertial anomalies—*a clear signature of the use of this EM/GR weapon*—were exhibited upon the personnel in the cockpit. The official investigators ignored those signatures; after all, such is not in their standard checksheets! And everybody knows the standard checklists are already perfect!

SIX MILE DROP IN TWO MINUTES

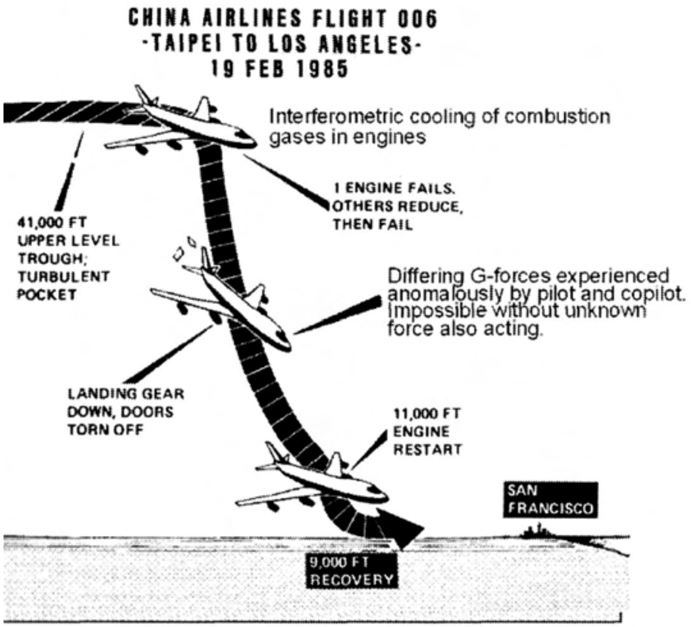
CHINA AIRLINES FLIGHT 006
-TAIPEI TO LOS ANGELES-
19 FEB 1985
Interferometric cooling of combustion gases in engines
1 ENGINE FAILS. OTHERS REDUCE, THEN FAIL
41,000 FT UPPER LEVEL TROUGH; TURBULENT POCKET
Differing G-forces experienced anomalously by pilot and copilot. Impossible without unknown force also acting.
LANDING GEAR DOWN, DOORS TORN OFF
11,000 FT ENGINE RESTART
SAN FRANCISCO
9,000 FT RECOVERY

Figure 56 Extinguishing combustion in the engines of a jetliner 19 Feb. 1985.

KGB Crews Regularly Practiced Against U.S. Rocket Launches from Cape Canaveral

In a late Nov. 1985 night shuttle launch at Cape Canaveral, the Russian interferometer crews continued their practice of using the U.S. shuttle launches as convenient "simulated" ICBM launches, for energetics interferometer crew practice (Figures 15 and 16). EM missiles (balls of EM energy formed in the area by distant LW interferometry through the earth) were used as the "destructive agent", but were offset from the actual launch in order not to destroy the shuttle. This was *normal crew drill,* not combat operations. The practice range for energetics weapons is—as might be expected—the actual targeted area and targeted nation itself.

As shown in *Fer-de-Lance,* I published an actual picture (Figure 15) of such a KGB EM missile strike, offset from that shuttle launch in late Nov. 1985 and snapped by Bob Gladwin. I also published an actual picture of a second such EM ball of glowing energy (Figure 16), over the launch site for the same launch, and used as a registration point (the LWs deviate slightly in passage through the earth, since the potentials in the earth are not constant). George Suchary took the second photo.

A standard artillery "shift-from-known registration point" method of firing was being used. In other words, the crews establish LW track on that EM registration ball, right through the earth. That represents *a point having a known set of coordinates*—a registration point—established in the firing computer. Then when the targeted strategic missile (in this case the shuttle's main booster) ignites, the interferometer also tracks the huge ionization target of the exhaust with the same weapon system.

That gives a known *relative* location of the missile rear end, with respect to one's registration settings for the registration point ball of light. Then add some meters to move up onto the missile body proper, and voila! The center of the target missile that will be fired is now in the bulls-eye, so to speak. That is a standard "shift from a known registration point" method of shooting, well known to every artillery group on the planet. The "marker beacon" is what is known as a "high burst registration," except it is not a burst but a steady ball of energy maintained over the area as the registration point.

The distant rocket or missile can now be destroyed in a variety of ways:

(1) Just create a sudden EM surge arising right in that spacetime (that is, just pulse the interferometer). That burns out some of the inner electronics of the missile, causing partial or complete missile failure.

(2) Sharply pulse the interferometer to introduce EM energy surges and sparks in the fuel or solid propulsion, igniting it massively and blowing up the rocket. Works well for aircraft fuel tanks as well.

(3) Form another EM missile and hit the missile broadside with that, burning

through it and destroying it immediately.

(4) Hit it with an explosive EM missile internally, instantly exploding the missile into pieces. Or

(5) Hit it with several of the above simultaneously.

Suchary and Gladwin had captured the direct evidence of KGB dress rehearsal for the strike and kill of the Arrow DC-8 in Gander, Newfoundland about two weeks later.

Kill of the Arrow DC: Practice Turned into Performance

Two weeks later, the same KGB weapon tested against the Cape Canaveral shuttle launch was used to destroy the Arrow DC-8 at Gander, Newfoundland, killing some 256 occupants, 248 of them being U.S. soldiers returning from peace-keeping operations overseas. The aircraft *was not* iced up; later that was clearly established.

See Figure 17. As the ill-fated Arrow aircraft began its thrust down the runway for takeoff, it was already under LW interferometer track (the exhausts of the **engines** were being precisely tracked) by distant KGB operators. The KGB operators simply began to cool the combustion gases in the engine combustion chambers as had been done to a U.S. jetliner earlier that same year (Figure 56) off the coast of San Francisco (there they had completely flamed out the en**gines).** Now they did not want to flame out the engines during take-off, but just wanted to reduce the thrust of those engines.

So the available thrust of the engines began decreasing as the aircraft picked up speed on the runway toward takeoff, even though the turbines were rotating at full speed. For powered flight, it is the *thrust-per-drag ratio* that is important. One can reduce that ratio by ice-loading and therefore increasing the overall aircraft drag, as all Transportation Safety Board checksheets assume if the engines are at full rotation and if operation and icing conditions are present. Or one can reduce the ratio by reducing the engine thrust itself, as was done for the Arrow, even though the plane is de-iced. All the Transportation Safety Board checksheets *assume* that it has to be drag increase, and thus due to icing.

The National Transportation Safety Board (NTSB) investigators know nothing about longitudinal EM wave interferometry and its capabilities—and its development by several nations of the world. Neither do their checksheets. Sad**lly,** the NTSB is not even aware that the electrodynamics and electrical engineering it assumes is so perfect is in fact a decrepit old model based on century-old errors. They are unaware that much better electrodynamics models have long been developed for particle physics, because the crusty old classical model is totally inadequate. Their checksheets also are "unaware" of anything really modern such as electrodynamics models of the higher group symmetry **kind.**

The thrust reduction by combustion gas cooling *does not appear on the official checksheets and is never considered.* It is not considered to this day. It will probably still not be considered this century.

Observers on the ground heard those engines "laboring," which can be a direct symptom of combustion gas cooling and engine thrust reduction. Rough running and missing can be another, if the cooling increases. Engine flameout occurs when the combustion no longer can be sustained because of the dramatically lowered temperature of the gases.

The Arrow Pilot Sensed the Lack of Power, But the Aircraft Was Immediately Hit

See Figure 17. At rotation and liftoff, the pilot of the Arrow DC-8 immediately sensed his lack of lift, and started a slow turn to the right, apparently intending to come around and land again as an emergency measure. Meanwhile, crossed beams had formed in the sky above and off to the right side. From those crossed beams of light, there shot down a bright "streak" of light that struck the fuselage of the aircraft in the right side, ahead of the engines. That was the EM missile strike—just as had been offset from the shuttle launch two weeks earlier in Florida, and photographed by Bob Gladwin.

That piece of Arrow DC-8 fuselage hit by the EM missile, with the burned hole in it, was recovered and tested. Tests showed no explosive residues on the edges of the hole, so it *was not* made by a bomb or other explosive device. Instead, it was made by an EM missile of the type that over a decade later hit the TWA-800 several times off Long Island and killed it.

In the Arrow aircraft's kill, the fiery EM missile sharply ignited the plastics in the forward cabin, which explosively outgassed. One product of such outgassing is hydrogen cyanide. A whiff or two, and all the passengers in that section were already dying of cyanide poisoning. *In the autopsies of the recovered bodies, about half the passengers were already dead of cyanide poisoning before the plane impacted the ground and exploded*. They died while that aircraft was still flying and sinking tail-first toward the ground.

The results of those autopsies were deliberately withheld from the Canadian investigative board.

The impact of the EM missile on the Arrow DC-8 was clearly recorded by sharp, vertical EM surges (spikes) in traces on both data recorders on board the aircraft. In *Gravitobiology,* on p. 82 I printed the recorded sensor responses **for** that signature also.

Elements of the U.S. Government Deliberately Suppressed Vital Evidence

Several eyewitnesses, including Judy Parsons, physically observed the formation of the crossed beams in the sky. Parsons and perhaps others actually saw the streak-down and strike of the EM missile when it hit the doomed Arrow DC-8. Parsons tried to testify, but her testimony was never allowed to be heard by the Canadian Transportation Safety Board. A major U.S. News agency interviewed her, but then squelched the story for some "unknown reason." Other witnesses (who are still reluctant to come forward) did observe the strike of the EM missile also.

Why would the U.S. government deliberately suppress information relating to the cause of the deaths of 248 of our soldiers returning from peacekeeping missions overseas? Why were they so determined that this direct act of war be declared an "accident" in the face of overwhelming evidence that it was not an accident at all?

Iran Contra was not yet divulged at the time. Recall, at very high level our own fellows had directly violated the law, and were engaged in the illegal Iran Contra fiasco up to their eyeballs. Also, some officials in Washington in certain high U. S. Government circles were in near panic, in their personal (erroneous) firm belief that a MidEast-sponsored terrorist bomb had destroyed the Arrow. In that case, so they reasoned, with the murder of 248 of our soldiers, the American public would be so aroused and angry that they would force military action against the MidEast nation or group, and that would then reveal Iran Contra. *Too many highly placed people's hands were dirty!*

So extreme spin control was the major requirement. Damn how many American soilders had just been killed in an overt act of war against the U.S. Damn the implications of the use of eerie new weapons we did not possess. A hue and cry from the U.S. public would almost certainly reveal Iran Contra and smoke out the duly secrets. That must not happen at any cost. Some high officials—particularly the ones who engage in such illegal acts, in assassinations, murders, etc.—are far more interested in covering up their nefarious deeds than they are interested in any kind of justice or morality. Consequently, the fix was in.

So *U.S. officials placed enormous pressure on Canada to produce a finding of icing.* Lots of other skullduggery occurred, which probably constituted criminal activity by some U.S. officials. A high-ranking U.S. military officer arrived on site quickly, apparently under orders to erase that scene as fast as possible. He almost immediately wanted the Canadians to bulldoze over the site, even before recovery operations were complete and while the investigation was just getting underway. The site was bulldozed over in unseemly haste not long thereafter. That would appear to possibly constitute deliberate tampering with the evidence at an official site of investigation, on the part of whoever ordered the General to bulldoze the site precipitously.

In addition, witnesses such as Parsons were never allowed to testify. The Canadian Board waited until the half of its members who strongly opposed any ad hoc finding were absent, and hastily approved a finding of icing as the cause of the "accident." Later the disgruntled Board members who were absent when the "voting" occurred, issued their own minority report. Unfortunately they knew nothing of scalar interferometry and energetics weapons, and so they strongly believed it had to be sabotage and explosives.

Anyway, the blackguards hiding Iran Contra buried the incident, and the savage kill of 248 of our soldiers and 8 civilians went unpunished. Anyone such as the present author who pointed out what had really happened and why, was just branded a lunatic. After all, the blackguards who do not flinch at murder and unapproved assassination do not blink an eye at slander and libel.

The sad part is those fine young soldiers and the 8 civilians who were killed, and the grief and lasting changes made to their families. It is the children who had to grow up without their fathers, the mothers whose husbands were suddenly dead, etc. They deserved much better from their government than they got.

Other Direct Energetics Signatures on the Arrow DC-8

Meanwhile, in the lifting DC-8 the strike of the EM missile had also *time-density charged* the forward cabin materials. Those materials slowly decayed from their excited states via longitudinal EM wave emission. When a human body is subjected to fairly significant LW waves (accompanied with their time-density waves), it gets its internal immune system and cellular regeneration system "jammed," since these systems use LWs and accompanying time-density waves. The cells and tissues are also "time-energy charged" as well, and so spurious LWs continued to be slowly emitted in their tissues and cells locally, leading to gradual, general debilitation. After a time delay for the effects of this time-charge decay to be exhibited, such exposed persons will get sick with odd symptoms of various diseases such as influenza.

The ground handlers on the scene and exposed to the struck materials of the aircraft, were also exposed to longitudinal EM wave radiation from those materials which had been instantly "time-charged" at the powerful strike. Hence they were exposed to fairly significant LW waves, and should therefore have shown later symptomology.

Some 60 or so of the ground handlers actually retrieving the crash materials later sickened with all sorts of mysterious maladies: dizziness, stomach cramps, flu-like symptoms, weakness, nausea, fatigue, etc.—precisely what would be expected if some the materials they were handling were emitting longitudinal EM radiation, exposing them to significant LW radiation. That is a clear signature again of a strike by an EM missile, and the exposure of those retrievers to LW emissions. *In no normal way could 60 of those ground crewmen handling the wreckage debris all come down simultaneously with the exact symptoms of*

LW radiation sickness, unless that was what it actually was.

Many of those sickened persons remained debilitated for years. Probably a few are still alive and debilitated.

There are other signatures on the Arrow DC-8 kill. I did a small bit of work on the problem in the interests of Arrow airlines, but of course the conventional investigators they also hired could only conceive of an ordinary explosive. Nevertheless, I did tell them as best I could what happened to their aircraft, their crewmembers, and our own valiant soldiers.

Our troops on board the tragic Arrow DC-8 *did not* die from a simple over-iced plane that crashed as a result of carelessness in deicing operations. *They died as the result of an act of war against the U.S., carried out by the KGB as a test in an entire series to demonstrate the 1985 readiness that Brezhnev had specified in 1972.* That war is still partially ongoing and undeclared to this very day, with other players such as the Yakuza and Aum Shinrikyo also now in the game—and possibly even the Chinese, who certainly have had the weapons for some time, and have had QP weapons and negative energy EMP weapons since earlier this year (2002).

With the kill of the Arrow DC-8, the KGB made its deadline that had been set down 13 years earlier by Brezhnev. To meet the 1985 schedule, something significant had to be killed successfully and clandestinely, no later than Dec. 31, 1985. The KGB made the schedule on December 12, 1985, by killing the Arrow DC-8 after a warm-up and full dress rehearsal against a U.S. night shuttle launch at Cape Canaveral about two weeks earlier. And they also demonstrated just how ignorant of such energetics weapons and their usage the U.S. government, scientific community, and investigative boards really were.

We were fortunate they did not just go ahead and kill that same shuttle in late November 1985. Instead, they waited and killed the Challenger a few months later.

Other Acts of War Completed the Milestone

Apparently the final demonstration of milestone completion called for the destruction of an aircraft, kill of an ICBM, kill of a space shuttle, induction of earthquakes, and demonstration of the new quantum potential weapon still in the advanced laboratory stage. The KGB completed much of the milestone demonstration in 1985. The remainder they completed in 1986.

In January 1986 they killed the space shuttle Challenger, and the very next night the KGB even gave a big party in Moscow, to celebrate the "perfect success of their active measures" against the shuttle. General Daniel O'Connor, former head of U.S. Army intelligence, personally issued an alert to many high-ranking U.S. government officials, pointing out the almost certain involvement of the KGB and the fact that the loss of the shuttle was almost certainly due to a hostile act by the Soviets.

In April 1986 (May Day report time was coming!) the KGB killed a Titan ICBM fired from Vandenberg Air Force Base. They also induced several large earthquakes for practice, and then began the induction of the long-dreaded "giant California quake" in the San Andreas Fault zone.

This latter induction of what would have been a massive quake in both San Francisco and Los Angeles, and all along the San Andreas Fault, was countered (destroyed) by a civilian group known to this author, with a backlash that resulted in the Chernobyl incident.

In April 1986 (for the same May Day report) the KGB also demonstrated the effectiveness of their newly emerging quantum potential weapon against the convenient U.S. air attack on Libya (Figure 25).

Kill of the Titan Missile Over Vandenberg Air Force Base

In 1986, the KGB began gingerly using those energetics interferometry weapons to kill other U.S. aircraft and several U.S. missiles, including killing a Titan 34-D fired from Vandenberg AFB on 18 April 1986. The Titan was killed 9 seconds after launch, partially crippling the U.S. space surveillance program. That kill was part of the demonstration of readiness milestone completion.

There was a completely decisive signature on the kill of the Titan. Simply check *Aviation Week & Space Technology,* 124 (17), Apr. 28, 1986, p. 18, where a picture (the one taken from longer range) of the Titan explosion is printed. Look about an inch above that explosion, which is already well developed. There above the Titan explosion sits the KGB ball of light registration point, just as it was sitting over the night shuttle launch in Cape Canaveral in latter Nov. 1985 when videotaped by a private photographer.

I persuaded NBC engineer Ron Cole and his colleagues to analyze a copy of that video tape frame by frame. In 17 frames, that little ball of energy above the explosion moved off on its own course, independent of the Titan explosion. Since the launch area for such missile launches is controlled and there were no aircraft over the explosion, then one again has a very clear signature. The KGB killed the Titan, with 100% certainty.

That of course was a test of the interferometry weapons in the role of *Launch Phase ABM system*. Tested against the Arrow, the KGB operators were emulating test of the system in a *Launch Phase Antibomber role*. These were vital *strategic* weapon tests. I published details of the use of the weapons in many other roles.

Dead-Man Fuzing and Outside Assistance Deterred a Full Strategic Attack in 1986

We have previously documented and published why the U.S. was not fully attacked and destroyed by the Soviets in 1986. It was because of our nuclear facilities, warheads, etc. which would have exploded, creating great fallout hazards worldwide, and because of nuclear dead-man fuzing weapons clandestinely placed in Russia by one or more other nations.

A friendly small nation possessing energetics weapons also intervened directly in l986, and blew up several Russian missile ammunition storage sites, some five or so in six months. In one such explosion, about one third of all the missiles in the Russian Northern Fleet were destroyed. As we stated, the so-called "Cold War" was not nearly so cold as has been represented.

In short, the friendly little nation served notice on the Soviets that, if they moved against the West, the Soviet Union itself would suffer terribly. They put it in language clearly understood by the KGB. So any full-up strategic attack for that period was stalemated by that combination of factors.

A major concern for the Soviets was their inability to counter the dead-man fuzing problem posed by nuclear weapons clandestinely hidden inside Russia and by the nuclear warheads and reactors inside the targeted nation(s). They already had the quantum potential (QP) answer *theoretically,* and had conducted a demonstration test of a laboratory QP device, but they had not yet completed the *QP* weapon systems and deployed them.

Nonetheless, a very powerful strike of peculiar nature—induction of a giant earhquake in the San Andreas Fault zone, including in both Los Angeles and San Francisco—was scheduled. This strike would have caused billions of do11ars worth of damage and would have killed perhaps 200,000 Americans. The initiation of the strike was actually begun in April 1986—as the reader can begin to see, there was to be one helluva May Day report in 1986! Incredibly, a private group here in the U.S. then made a unique and unprecedented contribution that saved the day for Los Angeles and San Francisco.

What Really Happened in the Chernobyl Disaster

Using **the** "inside" of the Woodpecker beams as superhighways for their longitudinal waves and interferometry, in April 1986 the KGB interferometers had started their build-up of substantial induced EM energy in the sides of the San Andreas Fault that runs through Los Angeles and up through San Francisco. Two beams were being used by the Soviets: One was coming down on the target from the North—coming over the North Pole regions and the ice caps—and one was coming around from the East and upon the target area. Completely (100%) certain signatures were present, including some ELF "chugger" waves symptomatic of the Soviet method of inducing very large quakes. The time-energy is being used, hence very low frequency ELF is employed to maximize the amount of time-energy available for transduction in the rocks on the side of the fault. The Soviets were building up a very large earthquake—indeed, to spawn a whole series of such quakes all along the San Andreas fault—with severe aftershocks following after the main quake had done terrible damage. In short, they were preparing to hit California with the "big one" that everybody had been anticipating and dreading for so long.

A private group contacted me. They stated they could take out one of the transmitters by producing a powerful (giant) phase conjugate replica signal pulse upon it that would backtrack right into the transmitter and catastrophically burn it out—a surface application of the "phase conjugate shooting" once considered for armed satellites that could shoot down rising hostile ICBMs at 10,000 miles distance. As I understood it, they had a new kind of essentially noise-free regenerative amplifier, used in a transmitter. This meant that *extreme amplification* could be achieved in a pulse response to a received Woodpecker signal, before the transmitter burned out.[43] Since the group had no knowledge of the exact KGB transmitter locations or the possible consequences of such an action, they asked my opinion as to the risks associated. This placed me in a real quandary.

At the time, the U.S. was literally under siege. We have mentioned some of the

---

[43] A very subtle but important point is this: An energy flow of any amount—say, 10^{20} joules per second—has absolutely *zero power* if none of that energy flow is diverted or changed in form. So if one has a truly perfect noise-free amplifier, and a pure noise-free energy signal input, the signal energy input can be directly and freely amplified without limit, so long as those noise-free conditions can be upheld. In short, that rigorously constitutes *regauging,* and the widely used **and** well-known axiom of gauge freedom—the ability to change the local energy density of a system or spacetime freely and at will—applies. Relativistically, such re-gauging is merely rotating the frame of the system away from the laboratory frame. This group indeed had such an amplifier, which I physically saw, and apparently has it today. Our present texts have not yet realized the real implications of gauge freedom: It means that one can freely amplify energy without limit, once certain conditions (no divergence of the energy or change of its form) are met. This does not violate energy conservation; the "potential energy of a system" is not a fixed number at all, but depends on the frame of the observer vis a vis the frame of the system. We point out that there really are no true conservation laws in general relativity unless introduced by arbitrary measured The great Hilbert already knew this in 1917, two years after the advent of Einstein's general relativity. Hilbert stated: "*/ assert... that for the general theory of relativity, i.e., in the case of general invariance of the Hamiltonian function, energy equations... corresponding to the* ***energy*** *equations in orthogonally invariant theories do not exist at all. I could even take this circumstance as the characteristic feature of the general theory of relativity.*" [D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21]. Loskutov and Logunov commented as follows: *"In formulating the equivalence principle, Einstein actually abandoned the idea of the gravitational field as a Faraday-Maxwell field, and this is reflected in the pseudotensorial characterization of the gravitational field that he introduced. Hilbert was the first to draw attention to the consequences of this. ... Unfortunately, ... Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that in general relativity conservation laws for energy, momentum, and angular momentum are in principle impossible."* ["Nonuniqueness of (he predictions of the general theory of relativity," *Sov. J. Part. Nucl,* 18(3), May-June 1987, p I79).

"target year" activities, such as kill of the Arrow DC-8, flameout of jet engines early in 1985, target practice against shuttle launches at Cape Canaveral in November 1985, etc. Actually kill of targets had continued in 1986. A huge earthquake followed by several large aftershocks had been used against Mexico City on 19 September 1985, causing great damage and demonstrating a little of what might be expected when they chose to hit Los Angeles and San Francisco much harder. The Mexico City quake had been preceded by an induced, giant hailstorm weeks earlier, and on the evening of 18 September 1985 by giant radial cloud signatures accompanying the generation of excess energy in the underlying rocks as the KGB prepared next day's giant Mexico City quake. The reader should recall that, working with Sweet during this period, I knew with certainty which earthquakes were natural and which were manmade.

U.S. missiles and aircraft had been hit, the Challenger had been destroyed, and the Titan 34-D had been destroyed at Vandenberg AFB on 18 April 1986, accompanied by a 100% indicator (the substantiated presence of the marker beacon for registration and shift firing). In addition, a little Richter 5.3 quake in the San Francisco Bay area had been induced by the Soviets on 31 March 1986, a quake which was correlated with preparatory activity on the Woodpecker transmitter signals the day before, showing Soviet induction of the energy to create the quake. Again, we had instrumental evidence that it was manmade.

That quake on 31 March had been a "warm-up" and practice for the big one now about to be induced in both Los Angeles and San Francisco, and along the San Andreas Fault. In fact, it appeared that the entire fault zone would be stimulated by deposit of excess energy all along its length. In that case, induction of a distrituted California quake of unprecedented extent and great magnitude was now in process, though just beginning. If that event or those events occurred, the loss of American lives and property would be staggering.

The High Side Risk Factor

The group's query to me presented a very difficult decision situation. Such huge Soviet interferometer beams, once built up, contain *enormous* longitudinal EM wave energy. If a Soviet transmitter were near large nuclear facilities, and were instantly blown and shorted, all that potential energy might flash directly into the ground.

In case of such a catastrophic transmitter failure, an enormously powerful longitudinal wave pulse would spread out in all directions in the surrounding earth. The first nuclear material the pulse encountered would be fissioned abruptly and very violently. If the transmitter were within a few kilometers of one or more of the large "monster" Soviet ICBM sites, the full-up explosion of several 30 or 40 megaton warheads might result—in the dirt and dirty. The resulting fallout and contamination would be massive, would spread round the world, and would eventually kill millions for years to come. Nuclear reactor stations were a similar risk; instant fission of all the fissionable material in one of them would constitute an unprecedented and giant nuclear explosion, in the dirt and dirty. That was the high side risk.

The Low Side Risk Factor

The low side risk would be that little or nothing at all would happen; perhaps a tremor induced in a nearby fault zone, etc. I knew that the Soviets—after an early Kyshtym transmitter failure—had produced such a flashover on a much smaller scale and that flashover had caused the nuclear explosion of nearby stored nuclear wastes. The resulting fallout contaminated some 1200 square kilometers, which is still contaminated to this day. I also knew that, after Kyshtym, the Soviets had installed safety devices on all their scalar interferometer transmitters in case of just such failures.

So if the safety devices held when the transmitter suddenly shorted out, they could hold the longitudinal EM wave beams and potentials long enough to drain the energy slowly and harmlessly into the earth over an extended period. That was the low side risk.

The *most likely* consequence, of course, would be somewhere in the middle, between the extreme upper risk and lower risk limits. In other words, we could expect some bad repercussions in the Soviet Union almost certainly, but hopefully it would mostly be confined to that spot.

Consequences of Not Taking Counterfire Action

If no action were taken, the risk to the U.S. was appalling. If the transmitters continued, then perhaps 200,000 Americans were going to die in Los Angeles and San Francisco and along the San Andreas Fault. Perhaps 300,000 more Americans would be injured, and many of them would also die. Enormous physical damage would also be done, and those basins would be disaster areas for some years to come. It would be interpreted by our inadequate scientific community as a great "natural" disaster caused by a giant earthquake, and the ghost of Khrushchev would have had one more laugh at us, with his words "We will bury you!" silently ringing in the ether.

My Recommendation to the Counterfire Group

To save those American lives, some reasonable element of risk had to be accepted. In other words, it became a classic "risk-to-benefit" military analysis, but one that was deadly serious. To me, as a military person the medium risk had to be accepted in such case.

I gave the group my reasoning. They then asked what I would personally do if my own finger were on the fire switch. I answered that I would take out the transmitter or transmitters and save those American lives, accept the medium risk, trust in God, and pray **for** the best. Military decisions often have to be made in just such uncertain conditions. If one's military could not make such tough decisions, then one's military forces would inevitably lose their wars.

The Group Fired and Took Out a Transmitter

The group thanked me, stated that they would fire, and hung up the phone. See Figure 57. They did fire on April 25, just as they promised, and they did destroy one of the KGB transmitters—apparently located some kilometers (perhaps 30?) from the Soviet nuclear reactors at Chernobyl.

At the moment the transmitter was hit and destroyed, engineer Bill Bise—out in the field with magnetic field detectors and measuring the Woodpecker beams—observed the abrupt failure of the Woodpecker transmitter beam coming around **the E**arth from the East. Thus, he actually detected the successful destruction of that distant interferometer transmitter. Unknowingly he had detected the results of a great counterstrike in a very dangerous strategic battle.

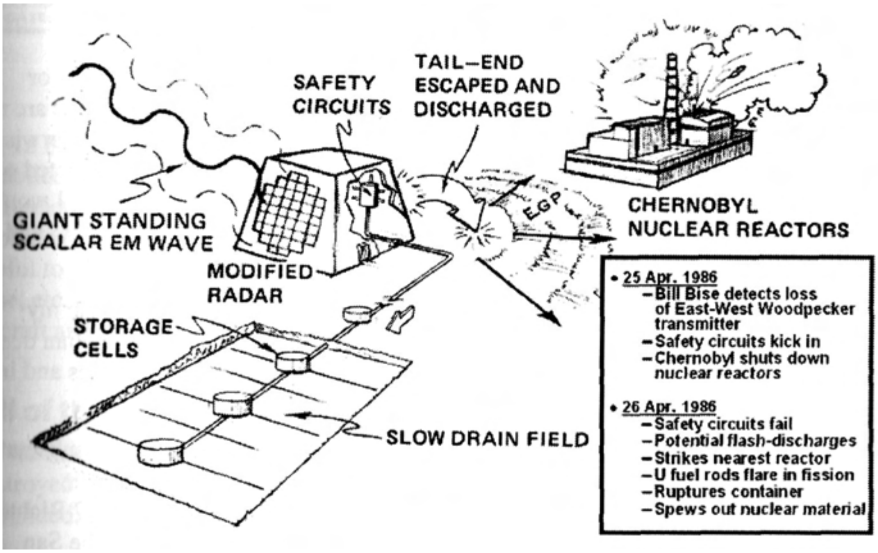
SAFETY CIRCUITS
TAIL—END ESCAPED AND DISCHARGED
GIANT STANDING SCALAR EM WAVE
EGP
CHERNOBYL NUCLEAR REACTORS
MODIFIED RADAR
STORAGE CELLS
SLOW DRAIN FIELD
• 25 Apr. 1986
–Bill Bise detects loss of East-West Woodpecker transmitter
–Safety circuits kick in
–Chernobyl shuts down nuclear reactors
• 26 Apr. 1986
–Safety circuits fail
–Potential flash-discharges
–Strikes nearest reactor
–U fuel rods flare in fission
–Ruptures container
–Spews out nuclear material

Figure 57 Chernobyl disaster 25-26 April 1986.

Results at the Transmitter and at Chernobyl

We were very, very lucky. At the destroyed transmitter site, the safety circuits kicked in and caught the huge beam potential and held it for about 24 hours, gradually draining down the intense beam energy into the earth in a slow manner all the while.

A **frantic** Russian message to the Chernobyl reactors galvanized the operators to immediately shut down those reactors (put in the cadmium control rods).

However, the uranium in the fuel rods was of course still sitting there. Chernobyl then waited with bated breath to see if the safety circuits would hold.

Then on 26 April, the safety circuits at the transmitter site finally crumbled, and a far weaker LW flashover occurred than would have occurred earlier without the safety. The spreading LWs in the earth struck the nearest reactor core at Chernobyl, preceded seconds before by a tremor induced in the earth at Chernobyl as the LWs arrived. *Part* of the struck uranium in the core fissioned immediately, venting the containment structure and spewing out radioactive material. Nuclear fallout spread over much of Europe. The rest of the disaster is well known.

That unknown group, in my estimation, saved a large number of American lives—at least 200,000 or more and perhaps as many as 300,000. They also prevented enormous physical damage to California, both to Los Angeles and San Francisco and to points in between. The members of that group are the unsung heroes of this bizarre episode. It is my hope that I may live to turn the names of those persons in for the award of a very high Presidential medal. Until then I will not reveal those names under any circumstance.

Whether or not the U.S. scientific and intelligence communities know it or believe it is of no concern. What matters is that a great many Americans are still alive, and their children born since then are alive. Those Americans otherwise would have been very, very dead or never born. I personally have reflected on those desperate circumstances many times. If I had it to do over again, I would unhesitatingly urge the same action: Fire and save those American lives, and accept the medium risk. And pray.

And I did pray after recommending the strike. In my opinion, this time my fervent prayers were answered. I am extraordinarily sorry for the civilian deaths that occurred; all one can say is that it was under severe circumstances and in a very peculiar—and very desperate—war.

The Soviets had many other interferometers on line. They induced a substantial earthquake in Mexico at the end of April. Thus they did complete their "earthquake footsteps" of the milestone. On 7 May they induced a 7.7 Richter quake in the Aleutian Islands. If 7.7 to 9.5 quakes had hit all along the San Andreas Fault in Los Angeles and San Francisco and in between, then the destructive scenario I had envisioned for the West Coast would have occurred.

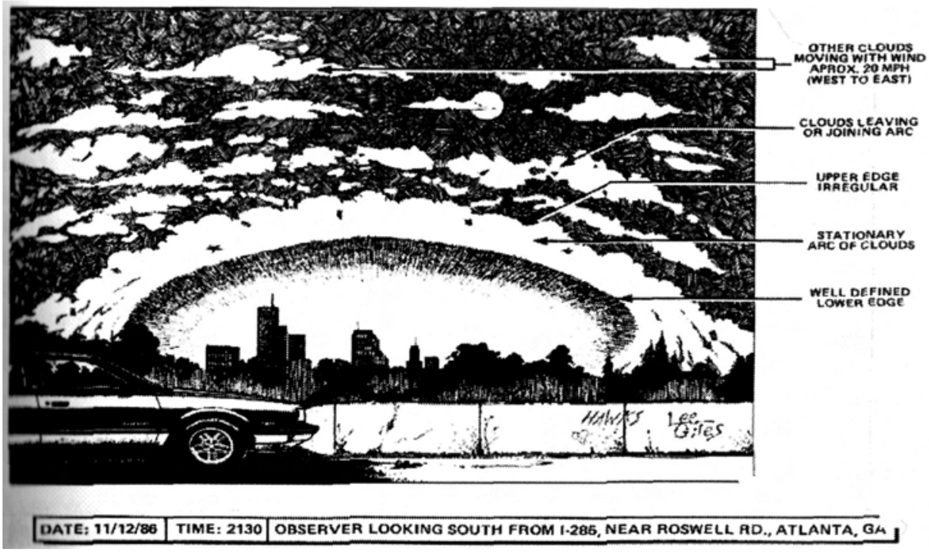
OTHER CLOUDS MOVING WITH WIND APROX. 20 MPH (WEST TO EAST)
CLOUDS LEAVING OR JOINING ARC
UPPER EDGE IRREGULAR
STATIONARY ARC OF CLOUDS
WELL DEFINED LOWER EDGE
HAWKS
Lee Giles
DATE: 11/12/86
TIME: 2130
OBSERVER LOOKING SOUTH FROM I-285, NEAR ROSWELL RD., ATLANTA, GA

Figure 58 "Bottling up" shield practice over Atlanta 12 Nov. 1986. With a real shield, aircraft or missiles flying into it (from inside or from outside) explode and are destroyed.

On 12 November 1986 the Soviets even placed a giant "dummy shield" (practice) over Atlanta, Georgia (Figure 58). If the shield had been real and not just practice, any aircraft and missiles flying into it would have been instantly exploded and destroyed. That mode—"bottling up" a strategic weapons area—is useful to nullify an entire strategic ICBM launch complex or a large bomber base, etc. It can also "bottle up" a large carrier task force at sea, including the aircraft and the Tomahawk missiles, the shells from the naval guns, etc.

Kill of the Challenger Included Use of Metal-Softening

As we mentioned, during that 1986 preparation period the Russians also destroyed the Challenger spacecraft (January 28, 1986). Numerous signatures abounded, all of which we have already published.

A few weeks before the kill of the Challenger, a close colleague had experimentally proven that a metal-softening signal was indeed placed on the Russian "woodpecker" signals (Figure 7). He did this quite simply: he wired himself into a receiver, through an intervening amplifier, and tuned the receiver to the Woodpecker signals.[44] Then he took nails from a can of nails he was using, and hit them with a hammer to drive them into planks. The nails all folded up instantly like totally wet noodles. A couple dozen, one right after the other. Simultaneously, his son—who was not connected to the Soviet signal—was pounding nails from the same box, normally and without difficulty. None of the nails picked up and pounded by his son experienced any difficulty or softness.

---

[44] Here he was taking advantage of the fact that the infolded LW signals inside the Woodpecker carrier wave, comprising the "engines" for metal-softening, also constitute "perfect" divergence-free EM energy, so to speak, at least overall. Amplifying the signal also amplifies the "engine" inside it. When connected to his body, "potentials superpose" is the rule, and that transferred the amplified engines to his body and into the EM field immediately surrounding his body. Hence he softened the hammer metal a tiny bit, and softened the much less tempered nails quite a bit more.

Then my friend disconnected himself from the signal, and waited a bit. Whereupon when he then used nails from that same box, he drove them in without difficulty. There were no more "wet noodle" nails.

We reach a very simple conclusion: When he connected to an amplified Woodpecker signal, his body was time-excitation charged (quick charge and quick decay) with a specific signal so that the nails he picked up were charged. Then when hit with the hammer, the nails acted as if made of very limp rubber without a rebound. When he disconnected from the amplified Woodpecker signal, and waited a moment for the induced charge in his body to dissipate, the nails also discharged their engines and quit acting like limp rubber bands when hit, and became normal nails with normal behavior. That would seem to be quite sufficient to prove the thesis.

The Challenger was then destroyed, as we previously printed, with numerous signatures of the hit, including even a KGB party in Moscow to celebrate the success of killing the vehicle (and getting away with it).

Cold Molding: An Epilog to the Metal-Softening Donnybrook

I was called a total lunatic for pointing out a metal-softening capability on the Russian signals. So be it. Metal softening is achieved by an engine which serves as a pumped phase conjugate mirror for the vibrations and signals in the lattice bonds of metals. Build up that engine in the metal until the precisely reversed and overlaid signals negate the normal signals, and one has "dissolved" the lattice bonds, converting the solid to a liquid. Lattice bond cancellation is not heating.

But that was not to be the end of the "metal softening" story.

A few years later, working informally with a large UK aerospace firm "for free", I furnished them an address in Moscow of some Russian scientists associated on the perimeter of the KGB energetics weapon program. The address had been taken from their cautious advertisement for business on the Internet and furnished to me by a correspondent. A UK engineer fluent in Russian was sent to meet with those scientists in Russia. They nervously showed him several novel things, one of which was "cold-molding" of metal. The liquefied the **metal**

The result was that a hammer blow onto a nail head resulted in the nail acting as a limp noodle upon impact.

without heating, poured it in a mold, and waited for the "effect" to wear off and the liquid to turn back into a solid again.

Briefly, an energetics charge (using LWs) was induced on the metal lattice vibrations, adding their phase conjugate replicas to the interior LWs comprising the lattice bond signals. The metal's atoms were time-density charged to an excited state (presently not recognized in Western science). Thereupon the metal lattice EM forces "dissolved" and the metal became liquid at room temperatures and without heating. The cool liquid metal was then poured into a mold. Then after awhile, the time-density charge and slowly drained away, and the metal resumed its normal solid state—all without any heating whatsoever.

On a subsequent trip the engineer brought that Russian team itself out of Russia and to the UK. The Russian scientists demonstrated cold molding to assembled UK scientists, and to some U.S. visitors. The phenomenon was physically demonstrated and proven.

I was correct about the metal softening then. We were also correct back there in 1986 about the metal-softening signal employed against the Challenger (together with several other things).

There are numerous other signatures on the death of the Challenger. Again, *the death of the Challenger was not an accident.* It was due to a deliberate act of war by the KGB. The KGB knew in advance—after substantial probing incidents—that they would get away with it. And they did so.

Additional Update from Year 1998 to Year 2,000 Has Been Given

In the action package we sent to the DoD in 1998, we told the rest of the KGB energetics weapons story up to that date. We gave the attack plans as we saw them and the most useful types of attacks to be expected.

I hope the reader is interested in this special energetics weapons information. High power microwave weapons, high energy lasers, and nuclear weapons are useful but no longer decisive weapons either strategically or on the battlefield against a foe equipped with such energetics weapons, and particularly against a foe equipped with quantum potential weapons and negative energy EMP weapons.

At present, we are intensely watching to see if we can pick up indicators for the next scheduled strategic attack, its probable timing, and the "new twists" included in it.

Presently, we have perhaps an 80% chance that the little nation can once again counter any new threat and cause the KGB to abort the impending attack. On the other hand, with Putin apparently limiting at least some of the action of the KGB, we have moved into our present asymmetrical warfare against terrorism—

and ultimately against the extensive terrorist teams and assets already infiltrated into our homeland.

As a single example, before the collapse of the former Soviet empire, the Russians long ago slipped in nuclear weapons (from 20 to 40 KT) and hid them in all our large cities and population centers, along with the Spetznaz teams to detonate them on command.[45] Several nations have introduced teams with normal sabotage assets and also with weapons of mass destruction such as radioactive nuclear materials (for a dirty bomb), anthrax, smallpox, bubonic plague, etc. Behind the scenes, the die-hard remnant of the old KGB crew are using QP weapons to "spread" the immune systems (Figure 53) of our populace and materially enhance any BW strikes that the terrorists unleash. The rogue Japanese team (Yakuza and Aum Shinrikyo) on site in Russia are using their LWIs in a giant weather war and the occasional kill of an aircraft, etc. Over the decades Castro infiltrated some 10,000 guerrilla/special forces types trained in camps in Southern Mexico for decades. Probably at least half are still loyal and dedicated—and waiting to blow up bridges, power substations, high voltage power distribution system towers, dams, harbor facilities, refineries, pipelines (both oil and gas), etc. The first phase of WW III—delivery of the weapons of destruction and mass destruction to the targets along with means to unleash them—has already been accomplished.

We are particularly struggling against the rise of asymmetrical warfare—the superpower's paradox (Figure 59)—and our present involvement in that "new" kind of warfare is still quite foreign to us. Only time will tell how well we fare in this new type of war, which we will be engaged in for many years to come. Vice President Cheney rightly regards it as almost a state of perpetual warfare we shall have to learn to live with, as shown by his words on Oct. 21, 2001:

"The war on terrorism will not be over in our lifetime. It is different than the Gulf War was in the sense that it may never end. At least not in our lifetime. The way I think of it is, it's a new normalcy."

---

[45] See Stanislav Lunev (with Ira Winkler), *Through the Eyes of the Enemy,* Regnery, Washington, 1998. Lunev is a high-ranking GRU defector. He is a former Colonel in the GRU, the military counterpart of the KGB. In p. 22-33, he summarizes the Spetznaz capabilities. On p. 22-27, he summarizes Spetznaz use of nuclear weapons already on American soil. On p. 26, he gives some of the ways in which the Soviets easily brought nuclear weapons into the U.S. On p. 30 he confirms Russian seismic weapons. Use of EMP weapons is on p. 30-31. Use of very, very low frequency weapons to destroy the human brain, put people into a zombie-like state, and aid in brainwashing is confirmed on p. 31. And this is i**nformation** Okayed by the CIA for release!

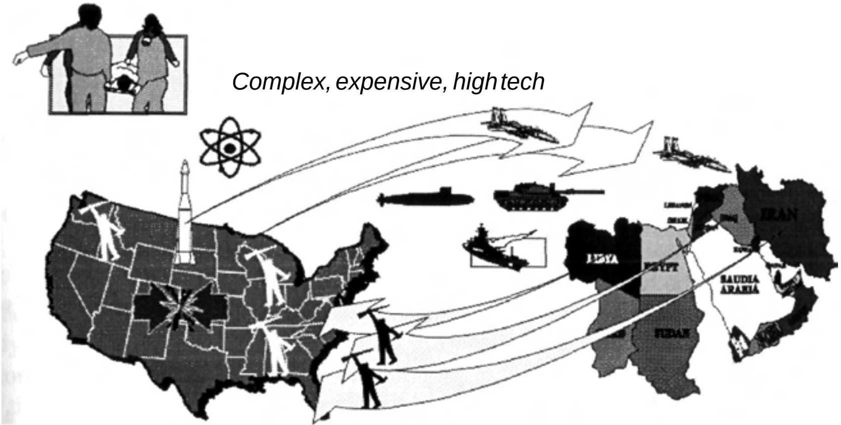

Complex, expensive, high tech

Simple, cheap, low tech initially.
Later includes high tech and very high tech.

Figure 59 The superpower paradox: Asymmetric cheap WMD strategic strikes and destruction, later becoming extremely high tech and extreme yield.

There is Also a Severe Bioenergetics (EM Biological Warfare) Threat

The threat from bioenergetics is equally vivid. That old die-hard section of the KGB can now attack an entire population, e.g., with electrodynamically (energetics) induced diseases of any kind whatsoever, particularly when using QP weapons. Such attacks are totally immune to inoculations and antibiotic treatment. Any kind of strain or variation of a disease can also be induced in the pathogen pool for that disease. Against a foe who has such capabilities, our entire medical establishment—including all our emergency treatment procedures and agents—*is obsolete right now.* A vaccination and the presence of antibodies, e.g., does absolutely nothing against the EM induction of the exact physical effects and form of a disease, as shown in the decades-long "microwave radiation" of the U.S. Embassy in Moscow (Figure 11).

To the Assistant to the SecDef in such matters, in 1998 we furnished a 200+ page briefing package and loosely proposed a crash development of the new type of medical therapeutic treatment devices pioneered and proven by the Priore group in France in the 1960s and early 1970s. Later, Congressman Cramer queried that office for an answer.

Unfortunately the Assistant SecDef's office erroneously responded to Congressman Cramer that nothing in the literature supported the mechanisms proposed. Indeed, the package we furnished contained a *detailed list* citing the hard French scientific papers in the literature, reporting the exact experiments and the astounding results obtained. Some 2,000 highly successful animal experiments for more than a decade of experiments in fact supported the thesis, and higher group symmetry electrodynamics fully supports the ability to use EM means to make spacetime curvature engines, so that these engines then interact back upon the targeted matter to perform the actions desired. That it does not appear in conventional electrical engineering is of no consequence; it appears in physics and it is good physics, just on the cutting edge.

We also reported, explained, and cited Becker's epochal work (Figures 27 and 28), which is copiously published in the hard literature and which was an example of the fundamental mechanism also. Becker was twice nominated for a Nobel Prize for his outstanding work. Unfortunately, the Assistant SecDef s office responded with a non sequitur.

Packages were sent also to NIH, NSF, CDC, and several other government agencies. Not a single one of those agencies was interested enough to have a competent scientist contact me and discuss the area. I strongly doubt that a single scientist anywhere in the Government bothered to *seriously* read and ponder the material so painfully and expensively compiled and furnished. Lieutenant Colonel (Retired) Ken Moore and I worked our fingers to the bone for weeks, preparing that briefing, and we spent all the spare money doing it that we could rake up and obtain.

The near-total lack of response itself seems astounding, but it clearly shows the state of U.S. science with respect to novel energetics innovations, even long after those innovations have been experimentally proven and documented in the hard literature, and even after they have been weaponized against us. It is a direct indication of why the U.S. never deciphered what was going on in some four decades of Soviet weak microwave radiation of the U.S. Embassy in Moscow, including the kill (eventual deaths) of three U.S. Ambassadors.

For decades there have been manipulations, to say the least, in the hotly contested field of the biological effects of EM radiation. We refer the interested reader to an excellent expose of that situation: The expose is by Dr. Andrew A. Marino, Ph.D., J.D., *Powerline Electromagnetic Fields and Human Health,* which is current. I believe it can be freely downloaded from: http://www.ortho.lsumc.edu/Faculty/Marino/PowerlineTOC.html. Dr. Marino has been personally and deeply involved in the investigation of EM bioeffects, in court cases, and in Blue Ribbon Panel investigations and proceedings. The book is an eye-opener, and it gives one a good view of the machinations in the EM bioeffects field. We quote directly from Dr. Marino, Chap. 6, p. 1:

> *"Oversimplistic as it may sound, whoever pays for EMF bioeffects research and analysis determines what data is produced and the way it is interpreted. Soon after the possibility that powerline EMFs were health risks was raised in a legal dispute involving the New York Public Service*

> *Commission, power companies and their trade associations, particularly the Electric Power Research Institute (EPRI), became massively involved in EM bioeffects research. Subsequently, the power industry dominated funding of the effects of powerline EMFs, both in terms of absolute dollars and compared with dollars from non-industry sources...the power companies and their trade associations were deeply deceitful regarding the information they provided to scientists and to the public regarding the potential health hazards of powerline EMFs."*

In the entire EM bioeffects field, there does not appear to be any mention of, study of, or recognition of EM bioeffects induced by the most powerful EM factors: longitudinal EM wave irradiation, time-polarized EM wave irradiation, and time-excitation charging of particles throughout the body and its cells, with resulting emission of transduced longitudinal and transverse EM radiation *from-within-to-without.*

National Institute of Health Reaction

NIH replied to my letter and action package from their *Policy*—read: *spin control*—Section. They noted that the package mentioned the Gulf War Syndrome, which was not in their mission area. So they just sent it off packing to DoD (simplest way to get rid of it, apparently no one seriously read it or even cared to read it). I queried them back by separate response letter about their lack of interest in a revolutionary proven new medical methodology well-documented to cure cancer and other dread diseases, in many hundreds of experiments by leading French medical scientists. This was certainly within their mission area.

They hastily replied from the *Policy* Section that, yes of course NIH was interested in all such things, and told me where to "submit a proposal." We never got out of their *Policy* department. Again, not a single competent scientist in the U.S.'s supposed master medical agency apparently even read the package or cared to call and discuss the matter. This was an indicator of their total lack of any grasp of the energetics threat, its testing, and its use to induce the Gulf War Disease.

Now perhaps one can see why we are so far behind the KGB energetics weapons scientists and weapons program.

Marino, *ibid.,* is particularly enlightening on the adamant opposition by appropriate sections of NIH vis a vis the harmful effect of nonionizing EM radiation.

No responses at all were received from the NSF and the CDC, or the other U.S. government agencies to which the package was furnished.

The CDC of course comes under NIH, and thus its responses or lack thereof would be dictated by the parent NIH policies. As for the NSF, Marino, *ibid.,* again is enlightening on the qualifications or lack thereof of its membership.

So as the reader can see, the top of the medical establishment in this nation is flatly not interested in any "controversial" or "out of the box" proven method for curing cancer, dread diseases, and mass treatment of those millions of American casualties that are coming in our cities and population centers. Again, they directly ignore scientific method's very premise: *If the experiment is replicated and contradicts the theory, accept the experiment and change the theory.* I personally concluded that the U.S. Medical establishment largely does not follow scientific method, but follows dogma and the already approved methodologies.

How the New Revolution in Medical Healing Works

For the reader who is a medical scientist and trained in physiology and biology, one might be interested in that revolutionary French therapeutic methodology and how it worked (before it was ruthlessly suppressed in the early 1970s).

Priore's apparatuses (Figures 12,49, and 60) simply pumped the cells of the entire body with longitudinal EM waves (each accompanied of course by its time-polarized twin, in a Whittaker pairing). The cells and every part of them are very nonlinear, so they act as pumped phase conjugate mirrors for Whittaker wave-pair engines resident in the cells and acting as "input signal complexes".

Once the cells are pumped by these now-established *time-density* waves, the damaged or diseased cellular masses slowly "charge up" with excess time-excitation. The irradiation took several hours, and this was the "activation" or "time-excitation charging" stage. Thereafter, the time-charge decayed slowly, all the while constituting *pumping* of every cell and every part of it, in the time domain.

The resident engine in the cell contains two components: (1) the normal or healthy component, and (2) the abnormal component representing disease or debility. Pumping these two components time-reverses both of them and amplifies that reversal. This creates amplified and incredibly precise *antiengines* in each cell and in each part of each cell. These antiengines override the weaker cellular engines resident. In turn, this "reversing" action of the antiengines causes the pumped cellular masses (actually the mass-energy, in its exact structure and dynamics)—at any and all levels, including the genetics—to be time-reversed back to normal. In biological instead of physics terms, it "dedifferentiated" the diseased or altered cells back to a previous healthy state by simply eliminating the diseased engine component. *This is a dramatic extension of nonlinear optical phase conjugation from the 3-space pumping domain to pumping in the complex (time) plane, and it is effective for time-reversing mass-energy at any and all frequencies. We accent that the cellular*

time *reversal back to a previous state includes all the cell's genetics and every* ***part*** *of the cell no matter how small.*

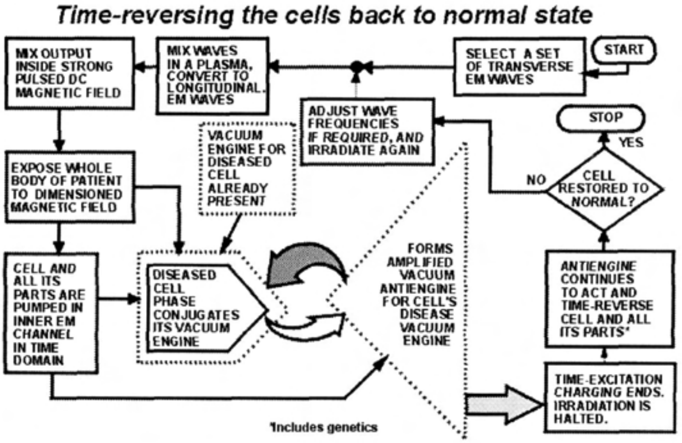
Time-reversing the cells back to normal state
START
SELECT A SET OF TRANSVERSE EM WAVES
MIX WAVES IN A PLASMA, CONVERT TO LONGITUDINAL EM WAVES
MIX OUTPUT INSIDE STRONG PULSED DC MAGNETIC FIELD
EXPOSE WHOLE BODY OF PATIENT TO DIMENSIONED MAGNETIC FIELD
CELL AND ALL ITS PARTS ARE PUMPED IN INNER EM CHANNEL IN TIME DOMAIN
VACUUM ENGINE FOR DISEASED CELL ALREADY PRESENT
ADJUST WAVE FREQUENCIES IF REQUIRED, AND IRRADIATE AGAIN
DISEASED CELL PHASE CONJUGATES ITS VACUUM ENGINE
FORMS AMPLIFIED VACUUM ANTIENGINE FOR CELL'S DISEASE VACUUM ENGINE
STOP
YES
NO
CELL RESTORED TO NORMAL?
ANTIENGINE CONTINUES TO ACT AND TIME-REVERSE CELL AND ALL ITS PARTS*
TIME-EXCITATION CHARGING ENDS. IRRADIATION IS HALTED.
*Includes genetics

Figure 60 Block diagram of Priore's healing (regenerative and time-reversing) process.

Normal cells are just backed up to a slightly younger condition. Diseased cells **have t**heir abnormal engine component overridden and reversed back to zero (its earlier state), so that only a slightly younger normal engine remains resident in the cell. This simultaneously reverses the cellular masses back to their previous healthy condition.

Contrast this direct healing mechanism to the action of the immune system.

The immune system itself does not *heal* anything. It contains the killers, the scavengers, the locators, the markers, the attackers, and the troops manning the ram**parts.** Usually after a fierce battle with invading hostile pathogens, it wins. Lots of cells still living are damaged in the melee. The immune system scavenger cells then scavenge up the residue littering the battlefield. But *the immune system itself cannot heal and restore a single damaged cell, including one of its own.*

That healing and restoration of damaged cells is the job of the *regenerative syst**em,*** still little known and inadequately studied today. Perhaps the most fundamental studies were mostly by Becker *et al.* some decades ago. The regenerative system (Figure 29) is an EM system in its operation, as clearly shown by Becker (Figures 27 and 28), yet it is able to physically differentiate and dedifferentiate cells from one form to another at will, and in highly selective fashion. The regenerative system acts within its limitations (only weak pumping can be done for the time-reversal of diseased or damaged cells) to eliminate the delta (the deviation from normal of the resident engines in the pumped cell). That restores the normal (healthy) engine and slightly amplifies it as well.

It is obviously of the utmost medical importance to understand and apply (and amplify) this little-known technical mechanism by which the regenerative system is able to restore damaged cells back to normal.

It turns out that *the regenerative system in its novel EM functioning uses precisely LW and TDW pumping of those damaged cells and gradually time-reverses their entire masses back to a normal healthy state, within its limitations.*

This is the secret of the mechanism of the cellular regenerative system—which is an electrodynamic system of very unusual nature. The exact mechanism has not previously appeared in the literature. It took me 14 years to uncover that fundamental mechanism, and another 4 or 5 to refine it a bit, even though both Priore and Becker (as well as others) applied it decades before without being able to fully decipher it.

The experimental proof is already in the literature, once one understands the higher topology EM unwittingly used by both Becker and the Priore team. The proper types of higher group symmetry electrodynamics required to understand the action of the cellular regenerative system are available. There exist scientists skilled in their application and use. These scientists and that higher group symmetry electrodynamics are little used, and are never used at all in the medical field. There exists no funding anywhere for higher EM examination and modeling of the functioning of the cellular regenerative system—i.e., for the fundamental healing process.

Just imagine what this could mean to medicine: *One can now develop a new medical science to amplify and apply nature's own proven method of restoration and healing of damaged cells, including all their damaged genetic material*. It matters not what caused the damage, whether pathogen or gamma radiation, whatever. This new therapeutic method is a totally new general relativistic process where powerful spacetime engines are created which work directly upon the damaged cell and every part of it, physically reversing the cell and all its parts precisely back to normal condition.

General relativity rigorously tells us that a precise spacetime curvature engine (set of precise spacetime curvatures) is present and active for every specific disease condition and every nuance in it. It also tells us this flawed (disease or damage carrying engine) continuously operates on the cellular mass at every level, maintaining the exact physical condition of the cell and its dynamics. By creating a precise "anti-engine" and amplifying and applying it, general relativity assures us that we are physically acting upon the cell and all its parts at every level, down to the smallest—to precisely return every part in perfect proportion back to a previous normal healthy condition. *Pautrizel directly proved the time-reversal to an actual earlier state of the cells and the organism* (Figure 61). Amplification by time-domain EM pumping, means that we can cause this time-reversal of the cells back to normal to happen much more quickly than can the body itself, which can only pump very weakly.

- Pautrizel compared identical treatment of immature rats with immature immune systems, and mature rats with mature immune systems.
- Treatment restored the damaged mature system back to full mature functioning, and it promptly recognized and dispatched the pathogens.
- Treatment restored the damaged immature system back to full immature functioning, unable to cope, and the pathogens promptly reinfected and killed the immature rat.
- The lesson: The body is indeed returned to a previous physical state. The pathogens themselves are not killed by the treatment, but by restoring the ability of the immune system to recognize and destroy them.
- Only *the previous ability* of the immune system is restored by the process.

PROF RAYMOND PAUTRIZEL
(Renowned Parasitologist)

Figure 61 Pautrizel's proof of time reversal of the stricken cells back to a previous state.

General relativity also assures us that any and every disease is subject to this novel treatment approach, even diseases which have resisted every other known therapy. The Sachs unified general relativity and electrodynamics, and GR/EM unification applications in O(3) electrodynamics produced by Evans, for the first time provide a solid theoretical basis for Priore's methodology and his evolutionary results. They also offer direct engineering of startling physical cures, once the technology is developed.

The profound implications of that treatment can be seen. Apparently, however, **our** leading governmental agencies and health agencies do not see it, do not wish to see it, have no appreciation for general relativistic effects in healing, and do not intend to have any appreciation for general relativistic effects in healing. Indeed, they seem totally uninterested in even trying to comprehend and research *the fundamental healing mechanism* demonstrated by nature herself.

What the French Researchers Showed

Among other things, the Priore team produced revolutionary cures of *terminal tumors, infectious trypanosomiasis, suppressed immune systems, and atheriosclerosis* in several thousand laboratory animals in rigid laboratory tests under rigorous protocols. Eminent scientists such as Robert Courrier and Raymond Pautrizel worked with the Priore team. Courrier was head of the Biology Section of the French Academy of Sciences at the time, also *Secretaire Perpetuel* of the Academy, and a noted scientist of high stature. Courrier personally presented the astounding Priore results to the entire assembled French Academy. He also sent his own most able scientist assistant to personally prepare the terminal tumor grafts on the lab animals before their treatment.

A world-renowned parasitologist who did enormously important work with Priore was Raymond Pautrizel, who is still alive at this writing though quite aged. Pautrizel was greatly persecuted for his years of determination and effort. Courtesy of Pautrizel via the late Christopher Bird, I personally have the very Ph.D. thesis that Priore himself submitted to the University of Bordeaux on his revolutionary work. That thesis was maliciously rejected in the suppression that was applied to the project. When the French Government changed to a Leftist government in the mid-70s, Government support was immediately withdrawn and the project was ruthlessly suppressed. Priore later sickened and died and that was the end of that. More than a decade later a brave and determined Pautrizel finally was able to get a doctoral thesis approved in the Priore work area, at the same university, by another doctoral candidate, Eric Perisse {40}.

There are two U.S. patents on the Priore device itself, and of course several French patents also.

All the above was rigorously documented in the 1998 package prepared and submitted by Ken Moore and I.

While Priore was still alive, a team with which I was connected attempted to mount a well-funded effort to restore the Priore methodology while Priore was still alive. We also were viciously suppressed. The lives of our financial backers and their families were threatened. The head of our project was viciously attacked financially and destroyed, fleeing the country for his very life. Whoever thinks that dramatically innovative medical science in this country is just sweet reason and scientific method, should have a strong enema to remove the serious blockages interfering with his perceptions.

This vicious outside suppression of our project so infuriated me that I made a solemn promise that, were it humanly possible, I would decipher the Priore process before I died. It took 14 years, but we made good on that promise. During the 20 years or so since our project was suppressed, it is heartbreaking to consider the millions of children and sickened and suffering people who could have been saved had the project continued and succeeded. It is heartbreaking now to consider the coming millions of casualties in our civilian populace in our cities, most of who are going to die from mass destruction weapons. And most of those would not have to die, would the scientific community waken from its long sleep and discover some modern physics and higher group symmetry electrodynamics—and their application to solve nature's mechanism for **healing.**

When one time-reverses the body's cells and all their parts, one forces the diseased cells to dedifferentiate back to healthy state. We strongly accent that this is accomplished by unified GR/EM processes involving precisely **organized** spacetime curvatures and their local actions in the cells and all their inner components, down to the finest level. It can be rigorously expressed in good unified physics, as expressed in an electrodynamics such as O(3).

The nondamaged or nondiseased cells just get a little younger. That is also highly beneficial; it reverses the ravages of aging, just a little. Indeed, one of the implications of the Priore mechanism is that aging can indeed be reversed by the process and treatment. Again, one must keep one's sense of humor. In order to get this technology and therapeutic method born, what seems to be needed is an extremely wealthy person, strongly desiring to live longer than the expected life **span** and do it in a vigorous, healthy body—and who is willing to fund the project with the necessary $60 million required so to redevelop and extend the Priore technology so it can be readily applied. A byproduct will be the saving of millions of American lives when those coming mass casualties from weapons of mass destruction occur in our cities.

How the Regeneration Mechanism Was Finally Deciphered

What took so long in deciphering the mechanism was the problem of how to make such precisely tailored and exact ST curvature engines, specifically so that they are *exact* antiengines for a specific disease and even a specific pathogen, and even for different versions of the engines for different cells in a single body having different strains of the same pathogen, having different body DNA, etc. For years, the solution of that task seemed hopeless. How on earth could one possibly work all that out in a hundred lifetimes, so that it could be precisely calculated even by an array of the most powerful computers on Earth?

After years of struggling with the problem, in one stroke of enlightenment it was revealed to be the simplest thing in all the world. Absolutely no calculation was necessary, and the exact amplified antiengine effect was a fairly simple and straightforward thing to obtain. General relativity tells us that every cellular mass and each part of it, has its own unique spacetime curvature engine, to the finest detail and an exact fit. As Wheeler put it so nicely {130}, *"Space acts on matter, telling it how to move. In turn, matter reacts back on space, telling it how to curve. "* The mass and its dynamics and the engine and its dynamics constantly interact upon each other, and "keep attuned and in perfect adjustment," so to speak.

That means that a "normal" cell already has just such a precise "normal" ST engine residing in it, and acting upon every component at every level—even in the atomic nuclei. Well, a diseased cell has an absolutely specific structuring, and therefore it also has an absolutely specific ST engine structuring. *A priori that* resident structuring for the diseased or damaged state in the individual cell differs (by some delta) from the structuring of the normal healthy cellular state for that individual cell.

Voila! That means that *the ST engine for the damaged cell can be decomposed into two parts (two engines):* (1) the normal "healthy cell" state engine, and (2) a delta ST engine that has been added to it so that together the two comprise the full ST engine for the exact individual condition of the cell.

What is now needed is a way to produce an exact antiengine of two parts: (1) one for the normal engine and (2) a second one for the delta (disease) engine. Applying this amplified and combined "antiengine" would overpower the resident combined engine, precisely reversing the cellular condition by reducing the delta to zero. Rigorously, the cellular mass and every part of it must then also revert to a previous physical state, or else we must complete discard general relativity itself.

Well, if we were to just time-reverse the resident engine and its two components, that would give us the perfect antiengine with its two components. Amplifying this antiengine would then produce one that was more powerful than the resident engine in the cell. Applying that amplified antiengine to the cell, that would not really change the "normal" engine component, but it would precisely reduce the delta engine component back to zero (its previous state, before it was there!) Voila! A specific healing signal of perfect structure, for the specific deviant condition of the individual cell and its every part.

Since we have a combined GR/EM to use, we can think electrodynamically about this general relativity problem. Immediately we see that phase conjugate optics is a candidate, since the "resident composite engine" can correspond to the "signal" or input wave, and the pumping of the nonlinear mass of the cell and its parts (where they act as nonlinear phase conjugate mirror materials) produces an exact phase conjugate replica (the exact amplified antiengine). The presence of that antiengine overrides the magnitude of the resident engine, and it moves the cell precisely back along the temporal path previously taken. In short, it applies a specific kind of dedifferentiation of the cell back to its previous healthy condition.

Since we extend phase conjugate pumping to the time-energy domain, then the resident composite engine and its time-stream component are the "input." The amplified antiengine and its time domain component will then forcibly propagate that pumped mass energy back over the time trajectory previously taken by it. In short, it will time-reverse the cell and all its parts back to a previous condition, completely eliminating the delta (disease or disorder). The beauty is that it will also reverse damaged or altered genetics, such as in the AIDS disease. It can do this prior to the appearance of health symptoms—it can be done as soon as the patient tests positive for HIV.

Since the antiengine is amplified, it moves the cell back to normal fairly rapidly (in a few hours, for example). At least the time charge for the reversal process is delivered in that period; one may allow (as Priore did) a full week for the time-charge and its transduction to continue reversing the cells. But it is much faster than the normal speed of the human regeneration system using the same process, because now we have greatly amplified the time-reversal rate.

It is therefore ideal for a method of quickly treating mass casualties, particularly since it can be developed and used in a portable unit about the size of a large suitcase (Figure 13).

Interestingly, during the radiation with LW EM radiation, one also pumps the pathogen and time-reverses it as well. Its "normal state" towards the beginning was a condition whereby it could move with relative immunity past the immune system's opposition, having deceived the immune system. In the treatment, a small amount of the pathogen is therefore converted back into this condition just shy of the beginning, where that part of the pathogens is resistant to the immune system. However, a very small, short "post treatment" irradiation with LW radiation for a much shorter interval and at weaker power, will clean up this altered pathogen residue by allowing the immune system to recognize it and destroy it. Pautrizel clearly demonstrated this significant fact.

Cellular and Body Rejuvenation are Possible, Once the Technology is Developed

One can in fact really accomplish *rejuvenation* of the aged and infirm, as the technology is developed. One could certainly solve the major problems of Medicare, and cut the patient load at least in half, relatively quickly and inexpensively. Of course lots of big U.S. drug companies will be rather violently opposed to any such idea, because much of their empire would be threatened. Much of the organized medical community depends largely on the Big Drug Approach. So the organized medical community will also vehemently oppose it, just as the formal medical science community in France viciously opposed it.

One could also produce cheap, portable LW treatment units *en masse,* which is what we were suggesting to the Assistant. SecDef. That way one could flood the emergency response system (response teams, hospitals, fire stations, police precinct stations, schools, etc.) with quick and effective treatment devices. They could be used to rapidly {131} treat those millions of American casualties that are going to occur when one or more of our cities are attacked in the near future by terrorist WMD attacks. This would save the lives of most of those stricken millions of Americans.

Mow Many Americans Could Be Saved?

For an anthrax attack on Washington, D.C., first generation equipment could save perhaps 70% of those sickened Americans, almost all of who are going to die with *present* capabilities. With second-generation equipment, one could save perhaps 90%.

For an anthrax attack without spreading via QP clandestine preparation of the populace, one would save 0.7 to 2.1 million Americans. With clandestine KGB QP spreading and a *5:*1 increase in the yield of the anthrax strike, one would save 3.5 to 10.5 million Americans.

For a truly nightmare attack (as a worst case scenario) perhaps 5 major cities might be attacked in such manner and to such degree. In that case, one is talking about saving 3.5 to 10.5 million Americans for the unspread attack case, and up to 17.5 to 52.5 million Americans for the spread case.

Further, the development of the necessary portable treatment units is doable and affordable, for mass treatment capability. Training on a portable machine could be done for a high school student in 30 minutes or less, and that student could then give emergency treatment effectively.

Absolutely nothing else out there can presently warrant such projections. Yet this is the very technology our own medical science community is eminently uninterested in, has no knowledge of, does not wish to obtain any knowledge of, and does not wish to scientifically discuss.

There Are No Resistant Strains to the New Technology, and There Cannot Be Any

A most interesting thing is that there cannot be a "resistant strain" of a pathogen for the new treatment methodology, *a priori.* No mass can shield against spacetime curvature effects. Slight adjustments of the controls will adjust the spacetime engines being utilized. Two or three hour's work with a developed unit and a lab team on a new strain would suffice to completely determine the required adjustments, once the technology is developed. The new information could then be quickly relayed to all stations treating patients. On each treatment device, adjusting a few dials and controls for the laptop computer controlling the unit will do the trick. The use of the final device would be so simple that a teenager could be trained to use it in less than 30 minutes.

Such new capabilities are urgently needed not only for the millions of civilian casualties we expect in the future, but also for support of our armed forces in the field. Frankly, the anthrax shots are not going to do them much good, once one realizes that the "spreading" of their immune and regenerative systems is already occurring via KGB quantum potential means. Camel pox is available to any terrorist foe, and has almost the identical actions and constitution as smallpox. Indeed, it is almost identical to smallpox.

The Gulf War Disease and Syndrome were just small preludes of what is to come and what is already being used against the entire U.S., though as yet very clandestinely. Look again at the Kaznacheyev work. *He clearly proved in thousands of successful experiments that any kind of cellular disease or damage or death whatsoever, can be induced at a distance by novel EM means.* Replication experiments were successfully accomplished at the University of Marburg and also at the University of Sydney.

Results of the Priore' Treatments Are Well-Documented and Independently Supported

The Priore team's results are well documented in the hard French scientific literature. At least one French doctoral thesis was eventually accepted upon the work. The astounding results really occurred. They were scientifically substantiated under rigorous protocols and even police-guarded labs for tests by a learn of outside scientists. To then deny those results simply because no one **has** *understood* the active mechanism until now, is to deny the scientific method itself: *Believe the experiment when it refutes the prevailing theory.*

There is also independent work by other scientists that strongly supports the Priore results, once the mechanism is finally understood.

Robert Becker *et al.* in the U.S. conclusively showed that living cells *in situ* can be dedifferentiated, redifferentiated, etc. by laughably weak currents (picoamperes) and DC potentials (which are comprised of those famous Whittaker pairs). Becker was unaware that his potentials included time density waves and longitudinal waves, as predicted by quantum field theory and as shown by Whittaker in 1903 (once one understands that a bidirectional phase conjugate longitudinal EM wavepair is a time-density EM wave coupled to a longitudinal EM wave). Application of such EM potentials by Becker transformed red blood cells into pre-cartilage cells, then further into pre-bone cells. That latter cell type was then deposited in bone fractures, healing them when nothing else would. Becker and others have demonstrated partial regeneration of limbs in small test mammals, using the technique. The bone-healing techniques were finally approved by the FDA and are often used in some hospitals for treating otherwise intractable bone fractures.

Since then,other scientists have shown the directed formation of other types of body cells from cells from distant parts of the body (i.e., muscle-type cells from cells produced in the bone marrow, etc.). Such things are now well documented in the U.S. literature, in at least several hundred papers and in the leading journals including *Nature, Science,* etc.

The Present Threat Is Urgent, and Our Strategic Destruction is Still at Issue

Our own medical community, our science community, and our intelligence community are at least 20 years or more behind the KGB right now. We are also behind the Chinese. In many respects our scientists have simply missed the biggest scientific breakthrough of the century.

It is urgent that we find some way to counter and avoid future strategic energetics attacks sure to be scheduled. The eventual attack actually launched—if the little nation's QP weapons can be countered first—would destroy the U.S. in about 2 hours. It only takes about 10 minutes, it seems, for the QP weapons *already deployed* to negate almost all our strategic nuclear arsenal. Thereafter, the KGB and Yakuza/Aum Shinrikyo can simply attack us at will and utterly destroy the U.S. To any loyal American, that is an entirely unacceptable solution.

The most urgent requirement would seem to be immediate and strong contact at the highest level between the U.S. government and the leaders of the little foreign nation (and the other friendly nation) that presently is guarding the ramparts and saving our necks. A cooperative crash program to develop our own QP weapons would seem to be the order of the day, if it can in any fashion be negotiated. But to have any chance of success, the President and the Congress will have to use the utmost care in selecting the U.S. negotiators and scientists to explore this possibility.

Immediate transfer of the requisite capabilities to the U.S. would be desirable, at least from our own view. From their view, of course, we do not necessarily resemble a trustworthy scientific partner at all, based on our past scientific performance, and it is up to the U.S. scientific community to prove to them that the leopard really has changed its spots. That would all have to be negotiated and worked out.

If that is not possible, then it is of extreme urgency that our own country embarks upon a new Manhattan Project to develop strategic quantum potential weapons and defenses. It would be a great relief to discover that such had already been started, but I fear not.

Desperately Needed Defensive Capabilities Can be Provided by a QP Solution

We pose one more very, very useful QP capability prior to closing. As we stated, if a large U.S. city is hit with—say—an anthrax attack done in professional manner, we would expect 1-3 million casualties, according to a well-known major U.S. government study. With immune/regenerative spreading presently already ongoing here in our populace by KGB QP weapons, that figure may actually increase by about 5-fold. So the casualties (prompt and delayed) might well reach 5-15 million, in a single *professional* anthrax spray attack by two persons from one light plane on a calm night dispensing 100 kilograms of anthrax spores sprayed over the city with a simple agricultural sprayer. With the indicated Priore technology and treatment units in all emergency response hands, at least 70% or so of those casualties could be saved. Otherwise, *presently almost all of them will die.* That dramatic saving of lives is one desperately needed capability that could be met.

But there is a further capability of LW technology that nothing else presently known can do without serious impact upon the populace.

After the prompt and shortly delayed casualties are mostly cured and the remaining fraction dies, a terrible problem of BW contamination still remains. No one anywhere really knows how to clean up an entire city from anthrax spores. But unless harshly decontaminated, that contaminated city is going to be effectively uninhabitable for decades to come, if only the present technology and methods are available. Islands of the coast of Scotland, where anthrax was tested decades ago, are still "hot" and contaminated—and off limits.

The indications are that our nation has developed chemical sprays that can kill and neutralize the anthrax, including in the lungs of the populace breathing the spray. On the other hands, it is indicated that the spray is a high physical stress factor, and some of the more debilitated citizens will be injured or even killed by the spray itself. This includes babies and small children, the aged, those with lung and breathing difficulties, etc. As a matter of national survival, desperate diseases do deserve desperate remedies. So while spraying the zone might itself sicken and kill some thousands of our weaker citizens, at least—so the thinking would go—it saves millions. This of course is similar to the use of triage—reserving treatment and scarce medical capabilities and supplies for treating those most likely to live and recover with treatment, rather than treating those more seriously affected and less likely to survive. Sadly, presently triage will almost certainly be necessary because of very short supplies and treatment teams, etc. So most of those 1 to 3 million casualties will simply be dragged aside and left to die, and just made as comfortable as possible.

However, the QP technology can be adapted to produce systems that can indeed do complete decontamination, very quickly, and without harm to the citizens in zone. The anthrax can simply be modified by specialized spacetime curvature engines—specialized EM fields and waves carrying hidden inner "information content of the field" engines that are designed specifically to change and destroy anthrax and nothing else. Eventually one could develop the capability to do it similarly for *any other* major contaminant, including nuclear radioactive materials contamination of a major city.

Any strategic and tactical analyst can immediately see the tremendous value of such capabilities.

In Conclusion

I hope the above admittedly lengthy overview is of interest to the reader and useful. If we are to survive the debacle that now looms before us in this new asymmetrical war we shall be in for decades, it will require the efforts of leading and dedicated Lawmakers and Overwatch persons. Without firm prodding, the military, scientific, and intelligence communities are likely to just continue their present pace until many or most of us all die suddenly.

The KGB energetics threat is still real and imminent; Russia is not a monolith and Putin still does not have complete control over the KGB. Another full strategic attack is sure to be scheduled before very long by either that die-hard faction of the KGB, the Yakuza/Aum Shinrikyo, the Chinese (particularly if they move against Taiwan and we defend), or someone. And so our possible deaths also to be taken into account, if we just continue to do business as usual.

Along with other Americans, I love my country and served it to the best of my ability for 20 years of active duty. It is the survival of America itself that is now at stake, both in the asymmetrical war against terrorists and in the continuing war against energetics weapons. Our scientific, military, and intelligence communities simply must be jolted awake and into massive efforts in energetics.

We have the scientists with the required capability. We have the facilities. We have the leadership, and we can do it if we can put this thing under a Presidential Decision Directive, under an Executive Order and a Declaration of National Emergency (along with the several already in effect).

Thomas E. Bearden, Ph.D.
Lieutenant Colonel, U.S. Army (Retired)
President and CEO, CTEC, Inc.
Director, Assoc. Distinguished American Scientists

1986

Fer De Lance

Briefing on Soviet Scalar Electromagnetic Weapons

T. E. Bearden, Ph.D.
LTC, U.S. Army (Retired)
Director, ADAS

1986

Note: This is the original Fer De Lance briefing paper wich includes some minor text and artwork editing.

1986 FER DE LANCE INTRODUCTION & BRIEFING ON SOVIET SCALAR ELECTROMAGNETIC WEAPONS

This briefing presents the basic concepts of Soviet scalar electromagnetic weapons, some of the major types available, and evidence of their widespread testing.

Scalar Electromagnetics is Electrogravitation

Scalar electromagnetics is an extension of present electromagnetics (EM) to include gravitation. That is, it is a unified, and, what is more important, it is a unified engineering theory. Nikola Tesla initially discovered its basis.

In the scalar EM extension, EM field energy can be turned into gravitational field energy and vice versa. This exchange can be patterned and localized, in specific areas or objects. Such a controlled change of electromagnetics to gravitation is not possible in the normal EM or physics presently taught in Western textbooks. However, the bits and pieces of the theory have been scattered through the physics literature for some time, but no orthodox Western scientist seems to have realized that these anomalous portions could be integrated into a startling new physics. Unorthodox experimenters, inventors, and scientists have made discoveries in this arena for several decades, but again have not realized the exact implications or the precise manner in which their results could be combined with present electrical physics.

Fer-de-Lance

Such sluggishness is certainly not present in the Soviet Union. For over three decades, the Soviet Union has been developing electrogravitation and applying it to develop strange new secret weapons of incredible power and capability. They have sustained the largest weapons development program ever launched by any nation, and they have kept it effectively hidden from prying Western eyes. 1 have called this program "Fer-de-Lance," after the deadly South American pit viper of that name.

The dreaded fer-de-lance is a snake of great agility and lethal effect. It often ambushes its hapless prey, and strikes unexpectedly and without warning. Its first sudden strike is usually lethal to its victim, which promptly expires in writhing agony. Since the Soviet development of scalar EM weapons has been designed for the same purpose, the name seems appropriate.

The equivalent effort of about seven Manhattan projects has been poured into Fer-de-Lance by the Soviets, and the program has been successful almost beyond imagination. The eery weapons are now developed, deployed, and tested. The ambush has been completed; Fer-de-Lance is coiled and ready to strike.

Energetics and Directed-Energy Weapons (DEWs)

The ordinary Soviet name for this type of weapons science is energetics. In the West, that term is believed to be associated with conventional directed-energy weapons (DEWs) such as particle beam weapons, lasers, radio frequency (RF) directed-energy devices, etc. The Soviets do not limit the term in this way.

Western scientists are familiar only with directed-energy weapons where fragments, masses, photons, or particles travel through space and contact the target to deliver their effects. Hence in their thinking they limit the Soviet term "energetics" to the type of weapons they themselves understand - exotic but normal weapons using energy or mass traveling through space to impact a target.

However, it is possible to focus the potential for the effects of a weapon through spacetime itself, in a manner so that mass and energy do not "travel through space" from the transmitter to the target at all. Instead, ripples and patterns in the fabric of spacetime itself are manipulated to meet and interfere in and at the local spacetime of some distant target. There interference of these ripple patterns creates the desired energetic effect (hence the term energetics) directly in and through the target itself, emerging from the very spacetime (vacuum) in which the target is imbedded at its distant location. As used by the Soviets, energetics refers to these eerie new superweapons, as well as to the more mundane DEWs known to the West.

As a consequence of the Soviet breakthrough and decades of feverish development, monstrous strategic weapons undreamed of in the West are already in Soviet hands. A noose is slowly and steadily being tightened about our throats, and it is already the 11th hour.

The Soviets Use a Deception Plan

Concomitant with this supersecret development program, the Soviets developed and implemented an elaborate deception plan to conceal these startling weapons and their nature from Western eyes until it is too late. Soviet deception has been so successful that even when Western scientists are confronted with the actual Soviet tests of these weapons directly over their heads, they do not recognize the weaponry nor the nature of the effects produced.

As early as January 1960, Nikita Khrushchev announced the Soviet development of a new, fantastic weapon. On May 1, 1960 Soviet defensive radars - rigged as prototype weapons of this new kind - probably downed Francis Gary Powers's high-flying U-2 reconnaissance plane over the Soviet Union, precipitating a major diplomatic incident.

On April 10, 1963 one of the first new superweapons operationally deployed was used to destroy the U.S.S. Thresher atomic submarine underwater, off the east coast of the U.S. The next day, April 11, 1963 the same deployed superweapon was utilized in a different mode to produce a giant underwater explosion in the ocean over the Puerto Rican Trench, 100 miles north of Puerto Rico.

Over the years, various aircraft were interfered with or downed as tests of these Soviet weapons. A particular case involved the mysterious loss of F-111s in the Vietnam conflict. At least one downed F-111 crew was recovered in the subsequent prisoner-of-war exchange. On that aircraft, all electrical systems suddenly were in difficulty simultaneously. All emergency indicator lamps were lit up, "like a Christmas tree." This was probably due to special Soviet teams converting some North Vietnamese SA-2 missile system radars to the scalar EM mode, and employing "scalar beam" interference to produce spurious EM noise throughout the electrical and electronic systems of the aircraft.

In late April/early May of 1985, the entire armada of Soviet strategic scalar EM superweapons was activated as a special celebration of the 40th anniversary of the end of WWII. Activation of this armada (which probably contained over 100 giant weapons), together with 27 giant power systems and a large number of command and control transmissions, was monitored on an advanced, proprietary detection system by Frank Golden. After the giant strategic exercise, which lasted several days, most of the weapons and power sources were once again stood down to "standby."

The NASA shuttle launches provided a convenient opportunity for Soviet testing of these superweapons in a Launch Phase ABM mode, where a launched missile can be detected and destroyed shortly after liftoff. At first, electromagnetic pulse (EMP) bursts on the early shuttle trajectory were deliberately delayed in time, to prevent actual destruction of the target and avoid alerting the U.S. that something unusual was happening.

The shuttle launch of November 26, 1985 saw a particularly significant-test of this kind. In this case, a very loud "sonic boom" or explosion occurred over the launch site 12 minutes after shuttle liftoff, when the shuttle was already away and downrange. At least two previous shuttle launches had also been used as pseudotargets, with delayed booms occurring over the launch site well behind the vehicles.

After the lack of U.S. reaction to these three tests showed that the U.S. still had no knowledge of the new technology and did not even recognize its employment, the Soviets apparently decided to proceed with tests where the target vehicles would actually be destroyed.

On December 12, 1985 the same Soviet weapon tested against the NASA shuttle launches may have deliberately interfered with the controls of an Arrow DC-8 taking off from Gander Air Force Base, Newfoundland. At an altitude of 100 feet, the aircraft - carrying over 250 U.S. soldiers and civilian crew members - lost power and sank into the ground tail-low, killing everyone on board. Three Canadian witnesses to the crash were interviewed over the Canadian Broadcast Corporation's television news on April 8, 1986 at 10 p.m. No flame or smoke issued from the plane before its descent and crash. However, the aircraft was seen to be mysteriously glowing with a yellow glow. That is a signature of the use of a scalar howitzer in the "continuous EM emergence" mode, similar to the manner in which the F-111s were downed in Vietnam. In short, the DC-8's electrical systems were interfered with by electromagnetic noise created throughout each increment of spacetime occupied by the aircraft. The powerful charge created in and on the aircraft also apparently caused the loss of two engines, one after the other. With its controls ineffective and power drastically reduced, the aircraft sank to earth, still in its "tail down" configuration from takeoff, and crashed and burned. The "yellow glow" was a corona due to the acquisition of a high electrical charge by the skin of the aircraft.

Other factors contributing to the crash may have been reduced lift due to poor engine maintenance, increased weight of the aircraft due to icing, and heavy loading. Still, no one has recognized the significance of the "yellow glow" or what it implies, or the possible connection between loss of the aircraft and previous Soviet testing of-a Launch Phase ABM system against U.S. shuttle launches.

As this book goes to press, the last two U.S. Air Force Titan 34-D missiles fired from Vandenberg Air Force Base in California have blown up shortly after launch. The first one blew up on Aug. 28, 1985 just after lift-off. That loss has been attributed to failure of a high-powered fuel pump, causing a massive oxidizer leak and a smaller fuel leak. The second Titan loss occurred on April 18, 1986 when the missile blew up 5 seconds after lift-off. Its loss is still under investigation and no determination of cause has been made. Apparently the shuttle and the Titan presently provide the only viable launch vehicles for launching U.S. "spy" satellites. The loss of these sensitive satellites - if indeed they constituted the payloads - cannot help but be damaging to our strategic surveillance capability. According to the Los Angeles Times, the single remaining KH-11 satellite was launched in December 1984 and, with an expected life of two to three years, it could stop functioning later this year.

Whether or not significant Woodpecker grid activity existed in the vicinity of the Titan launch of August 1985 is unknown at this time. However, significant activity in the grid definitely occurred before the April 18 Titan disaster and on the same day.

On Easter Sunday, Mar. 30, 1986 engineer Ron Cole observed significant **cloud** signatures of grid pattern activity, correlated with Soviet Woodpecker measurements. On April 18, the present author observed traces of a cloud **radial** over Huntsville, Alabama and took photographs of it. Preliminary reports from

Thousand Oaks, California indicate extensive grid activity again on April 18, the day the second Titan exploded.

At least the second of these two missile destructions shortly after launch is suspicious, since the grid positively was active during that time. Also, a clear trail of Launch Phase A8M system indicators exists back to the massive scalar exercise of April/May 1985. The first Titan explosion in Aug. 1985 thus falls within the Soviets' "now let's test them against U.S. launch vehicles" period. The second Titan loss follows highly suspicious losses of the Arrow DC-8 on Dec -T7 1985 and the Challenger on Jan. 28, 1986. The same Soviet weapon system that destroyed those targets may also have destroyed one or both of the critical Titans.

The Soviets have also been able to significantly engineer the weather over North America for more than a decade without being found out. They have tested a fantastic range of anti-ballistic missile (ABM) defense weapons for over two decades, and no one is the wiser in the West.

As previously stated, prior to the end of November, 1985, at least three "wet-run" tests of a Soviet "launch phase ABM system" against actual U.S. shuttle launches were made directly over Cape Canaveral itself, and still no one recognized what was happening or what sort of weapon was being tested. On December 12, destruction of the Arrow DC-8 in Newfoundland produced no indications that the Americans and Canadians knew anything about the nature of the weapon possibly used.

Accordingly, after a sufficient wait to test our reactions (if any), the Soviets prepared to actually destroy a shuttle after its launch.

Destruction of the Challenger in January 1986

As the whole world knows, on January 28, 1986 the shuttle Challenger was launched from Cape Canaveral, Florida after exposure to undesirable weather conditions, and disastrously exploded shortly after launch. The evidence seems to indicate that, as the rising vehicle was stressed, one end of its right booster broke loose, twisting away and into the main fuel tank, causing rupture, spillage of the fuel and catastrophic explosion. Several other anomalies still exist, however, and it is clear that a problem existed with at least one of the booster scals. All seven astronauts aboard the flight were killed in the fiery destruction of the vehicle.

Of course no one had recognized that the Soviets had already tested a launch-phase anti-ballistic missile (LPABM) system against three of our previous shuttle launches. These "wet-runs" used a deliberate "time offset" to delay the explosive emergence of electromagnetic energy in a launched shuttle's location along its trajectory. The delayed test shots resulted in very large "booms" above the launch site after the shuttle was safely out of the area, but did not destroy the shuttle vehicles themselves. For example, the delayed-shot boom occurred 12 minutes after the evening launch of Nov. 26, 1985. Even a marker beacon (large light in the sky) was utilized on that launch shortly after lift-off. The marker beacon was photographed. In addition, another photograph taken in a time sequence shows another sudden streak of light coming down and ending in a burst of light. This was probably a spatially-offset test of the "pulse" mode for destroying the shuttle. The "light burst" would have been detected in the Soviet Union and scored against its intended offset position. However, another mechanism may have been used to cause destruction of the shuttle itself.

On Jan. 1, 1986 the presence of a metal-softening signal added on to the Soviet LPABM system's scalar EM transmissions was detected by a surprised Frank Golden. The metal-softening ability of the detected signal was experimentally verified by him. Golden also locally nullified the action of the scalar EM signal in a test, rather conclusively establishing (1) that it existed and (2) what it was. The signal was apparently being prepared for use against an upcoming U.S. shuttle launch.

Just prior to the launch of the shuttle in late January, 1986, the Soviets accomplished significant weather engineering over the U.S. The jet stream was severely bent southward in the middle of the U.S., bending rightward again to move across the Florida panhandle. Icy cold air from Canada was drawn far down, into Florida and onto the Challenger sitting on its launch pad. This exposure to cold outside its tested range was probably an additional factor contributing to the Challenger's failure. Positive signatures of the Soviet weather engineering and jet stream manipulation were observed and photographed by several persons, particularly in Alabama and California.

On January 28, 1986, the Soviet scalar EM weapon system effects were sharply localized in the launch zone. Localization involved higher frequencies being present; these are quite painful to small brains - whose hemispheres act as a scalar interferometer and detector - such as in birds. As commented upon by national news commentators, strangely the birds were not flying that morning. Indeed, they were staying down on the ground or avoiding the area, since the sky over the area was painful to them.

As the Challenger rose, the metal-softening signal would have been experienced in and around the boosters shortly after ignition, since the booster flame acts as a special "ion-plasma" turner/detector for the scalar signal. (Certain plasmas have a unique characteristic: they transform an input transverse wave to a longitudinal wave, or an input longitudinal wave to a transverse wave.) The effect of such a local signal is to "change mass in the immediate vicinity" with the particular scalar resonance signal. (Mass acts as a capacitance or accumulator for scalar resonance).

The ill-fated Challenger was doomed. After ignition, the booster flame acted as a ionic plasma detector/amplifier for the scalar metal-softening signal on the

Soviet Woodpecker grid. The metal in and around the booster flame was slowly and steadily weakening due to charge-up with the metal-softening pattern. This attributed to booster leakage from the already cold-damaged seal. The leaking booster poured out smoke and later flame, contributing to the impending disaster.

Substantial winds and air turbulence over the site increased the stress on the Challenger as it rose through this region. This also contributed to the impending disaster and may have been deliberately created there by Soviet weather engineering.

As the Challenger stressed, eventually one or more weakened metal mounts gave way, partially freeing the end of the right booster. The booster oscillated, rotating into the tank and rupturing it. Contact of the escaping liquid fuel and the flame resulted in a fiery explosion, destroying the vehicle.

Even so another anomalous flame or light was observed on the vehicle, and may have represented a very small additional "pulse mode energy form" produced by the Soviet weapon that was attacking the Challenger. The shock of the explosion probably killed the astronauts instantly, although the cabin appears to have remained essentially intact while it plunged several miles to the ocean below.

By treating excess cold exposure to the shuttle, inducing metal-softening in and around the ignited booster, and possible adding a deliberate "intensely hot spot", the Soviets caused the Challenger to weaken and destroy itself shortly after launch, in so subtle a fashion that NASA scientists would not suspect what had actually caused the mishap. The Soviets also had previously withdrawn all their trawlers and ships which normally shadow a shuttle launch, to prevent any suggestion of Soviet presence near or involvement in the catastrophic accident.

A few days later, sporadically intense "high frequency localization signals" were still present on the grid, at least at one hinge-point at Birmingham, Alabama. From 1-4 February 1986, many birds inadvertently flew into the zone when an intense breakout of these components occurred, and dead birds fell from the sky in substantial numbers.

* According to an Urgentgram sent out by General D. Graham, on the evening of the Challenger's loss, KGB headquarters held a party to celebrate "success of their active measure against the shuttle." Note that all Soviet scalar EM weapons - development, testing, deployment, and usage - is under command and control of the KGB.

Among other things, this briefing details the bizarre series of indicators that showed the Soviet intention to destroy the vehicle, and documents the accidental discovery of the signal with which they intended to cause the shuttle to fail and destroy itself.

Technological Surprise and a New Hiroshima

Sadly, the bureaucratic smugness of orthodox Western scientists has materially assisted the Soviet deception process. Most Western managerial scientists - particularly in weapons development activities — have continued to view the Soviets as ignorant peasants, still trying to clean the mud off their boots. This view, of course, is totally untrue and unwarranted.

It takes only a few examples to refute this attitude. In nonlinear mathematics, engineering and science, the Soviets have led their Western counterparts since the beginning. The electromagnetic pulse (EMP) effect of a nuclear explosion appeared in the ordinary Soviet scientific literature before Western scientists were even aware that the effect existed. The Soviets continue to lead the world in explosive welding, titanium forming and welding, etc. While the Soviet scientists exhibit little inclination to build good washing machines, they certainly do produce state-of-the-art technology — and beyond — in any area in which they focus their main efforts. (We do lead the Soviets in some areas such as computers, computer software, miniaturization, etc.)

Once before, a modern nation - the United States - developed a mighty weapon in secrecy and used it to force a powerful foe - Japan — to its knees. The mindbending atomic blows to Hiroshima, and shortly after to Nagasaki, showed once and for all that in the modern age technological surprise can prove instantly disastrous. Yet in our scientific arrogance, we have assumed that no one else will ever do such a thing. We have assumed that it certainly could never happen to us, and that the "secret weapon" scenario will never be repeated. On the contrary, it has happened again, someone else has done it, and it has happened to us.

It has also become fashionable in the West to believe that all the laws of physics are already discovered. We assume we already know all of them. While we have been pridefully crowing this tune, the Soviets have been steadily discovering new laws in secret, as well as new ways to circumvent the old laws.

We Have Given Up Defense in the Larger Sense

Meanwhile, to defend ourselves strategically, we have chosen to forego "defense" and rely almost totally on building a powerful strategic offense that is capable (we think!) of destroying our would-be enemies anywhere in the world. Our offensive striking power is based mostly on a triad of deterrent forces - land-based ballistic missiles, submarine-launched ballistic missiles, and strategic bombers. The cruise missile is presently being added to the triad. If this 4-part offense were nullified or destroyed by Soviet secret weapons, we would be powerless to prevent our own destruction and Soviet domination of the world.

We have been viewing ourselves as Samson, confident in the strength of our locks.

Yet literally our locks have been secretly shorn and the Philistines are already upon us. Indeed, Soviet energetics weapons are now capable of destroying our triad shield, our homeland, our armed forces in the field, and our population, quickly and efficiently. We have a new "gap" of monumental proportions: not a missile gap, not a submarine or bomber gap, and not even a particle beam or laser gap. We have a scalar electromagnetics or electrogravitation gap.

In this briefing, it is my purpose to show a portion of the Soviet weapons developments that have lead me to such dire conclusions. The Fer-de-Lance briefing is also an attempt to galvanize our leaders and scientists into action, for our present total vulnerability to the Soviet Fer-de-Lance weapons is intolerable and unacceptable.

Unless we achieve defenses — and quickly — our own demise in a fashion similar to the Hiroshima-Nagasaki scenario is inevitable and imminent.

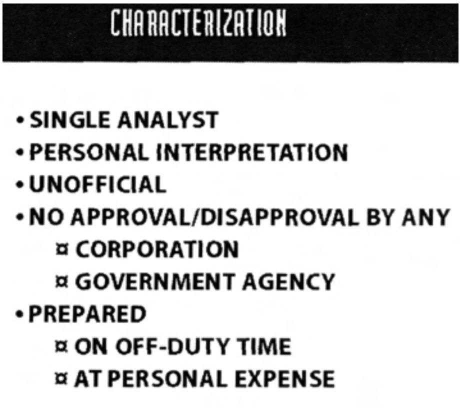

CHARACTERIZATION

- SINGLE ANALYST
- PERSONAL INTERPRETATION
- UNOFFICIAL
- NO APPROVAL/DISAPPROVAL BY ANY
    - ¤ CORPORATION
    - ¤ GOVERNMENT AGENCY
- PREPARED
    - ¤ ON OFF-DUTY TIME
    - ¤ AT PERSONAL EXPENSE

Slide 001: Characterization

We point out at the beginning that this briefing and its interpretation are the result of 22 years of analysis and effort by a single analyst. It is entirely unofficial. It is not approved or disapproved by any corporation, company, or government agency. It has been prepared by this analyst at nights, on weekends, on holidays, and during off-duty time.

It has been prepared mostly at personal expense, though several grants and support have been received from R. J. Reynolds III, Peter Kelly, Mike Bearden, Interdimensional Sciences, Association of Distinguished American Scientists, and others. Very valuable assistance has been received from the U.S. Psychotronics Association and the Planetary Association for Clean Energy. These contributions are deeply appreciated and most gratefully acknowledged.

Graphics and artwork support by Hal Crawford, Margaret Wilson, Tommy Neumann, Ron Cole, and Lee Giles has been invaluable. My sincere thanks go to these artists who have made the presentation vivid and imaginative. A great deal of background support was given by Joe Gambill, Frank Golden, Tom Herold, Ken Moore, Pete Kelly, Eike Mueller and Mel Bartlett and is deeply appreciated.

Finally, the help of many other persons and colleagues too numerous to enumerate is also gratefully acknowledged.

If the U.S. bureaucracy ever moves and develops defenses against scalar EM weapons, it will largely be due to the support of the persons and organizations who have freely given of their talents and resources to further this effort.

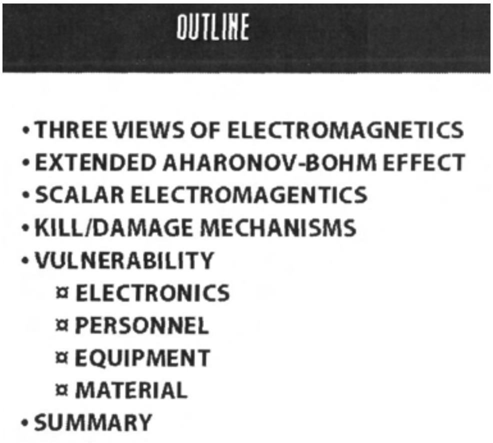

OUTLINE

- THREE VIEWS OF ELECTROMAGNETICS
- EXTENDED AHARONOV-BOHM EFFECT
- SCALAR ELECTROMAGENTICS
- KILL/DAMAGE MECHANISMS
- VULNERABILITY
  - ¤ ELECTRONICS
  - ¤ PERSONNEL
  - ¤ EQUIPMENT
  - ¤ MATERIAL
- SUMMARY

Slide 002: Outline

This presentation does not attempt to cover the complete field of energetics weapons developed by the Soviet Union. Instead, it concentrates on the major areas necessary to understand the type of weapons likely to be unleashed on our strategic and tactical installations and our troops and their equipment. It does not attempt to present the frightful bio-electromagnetic weapons the Soviets have developed to be unleashed on our populace at large and on our troops in the field. Also, the briefing does not dwell on the historical progression, but on the types of weapons themselves.

Of necessity, it presents several shortcomings in classical EM, and presents three differing views of electromagnetics and what can be done with them. The briefing particularly stresses the Aharanov-Bohm (AB) effect. This mesoscopic effect, which produces action at a distance by pure EM flux and not EM force fields, has been extended by the Soviet weapons developers into the macroscopic world around us.

In addition, we develop the basic concepts of scalar electromagnetics and how electrogravitation is achieved. The basic concepts of scalar interferometry and scalar resonance are briefly advanced, and a few of the major kill and damage mechanisms are presented. The nearly total vulnerability of our present electronic equipment, personnel, communications, vehicles, aircraft, ships, missiles, submarines, nuclear weapons, and materiel is stressed.

Finally, a short summary is given to highlight the main points of the briefing.

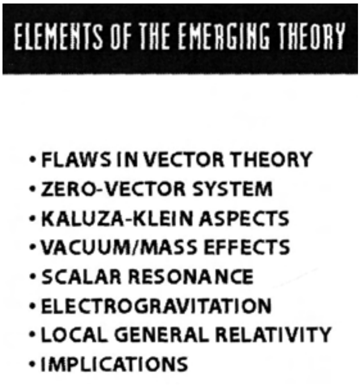

ELEMENTS OF THE EMERGING THEORY

- FLAWS IN VECTOR THEORY
- ZERO-VECTOR SYSTEM
- KALUZA-KLEIN ASPECTS
- VACUUM/MASS EFFECTS
- SCALAR RESONANCE
- ELECTROGRAVITATION
- LOCAL GENERAL RELATIVITY
- IMPLICATIONS

Slide 003: Elements of the Emerging Theory

Specifically, we will point out some flaws in vector mathematics itself, particularly with the concept of the zero force vector. The zero force vector is a system of forces that sum to a zero resultant. Hence the components of the summation represent a patterned stress in the medium to which they are applied, or in which they are imbedded. This includes the vacuum (spacetime) medium. In classical electromagnetics, this vacuum stress due to a zero-vector summation of EM force fields has been totally omitted and ignored.

As such, the EM "zero" force-vector summation produces a "trapped internal EM flux and flux pattern, without resultant (external) force field" condition - precisely as does the Aharonov-Bohm effect. The components of the artificially zeroed system, however, can be transmitted and still maintain their special relationship and coherence. While the AB effect has been shown to hold for the mesoscale (a few thousands of angstroms), the zero-vector scalar EM effect can hold for hundreds of thousands of kilometers.

To provide a unified electrogravitation, we adapt Kaluza-Klein 5-dimensional gravitational concepts to the idea of the zero-vector stress system in vacuum-spacetime. We also point out how simultaneously varying the magnitudes of **the** force components of a stress, all in phase, produces a stress wave or scalar EM wave. Scalar waves are almost always absorbed and emitted by the nucleus of the atom, not by the electrons in orbit.

The relationship of mass and vacuum, and the constitution of the vacuum, are pointed out from the viewpoint of modern quantum mechanics.

How a scalar wave resonance differs from conventional EM resonance is developed briefly. Mass and inertia are the direct result of - and are - trapped scalar resonance. The trapping mechanism is the spin of the particle.

Severely limiting assumptions in ordinary general relativity (OGR) are pointed out. In OGR, it is assumed that the local frame is always a Lorentz frame, and never curved. In other words, local spacetime is always assumed to be flat. This **saves** the conservation laws, simplifies relativity, and reduces "general" relativity to special relativity with distant perturbations and curvatures.

By removing this ad hoc assumption, a much richer local general relativity results. This local general relativity is readily engineered. Note that, in OGR, the physicist has actually assumed that he can never "engineer" local general relativity! Indeed, with the scalar EM approach, he can easily do so, in contradiction to what is taught in all Western universities.

By engineering a local general relativity (LGR), the individual conservation laws can be violated locally. This includes the conservation of energy/momentum, and the conservation of charge, for example.

The major implication of this startling new engineering physics is that one can **engineer** physical reality itself. For example, elements can be transmuted with minuscule energy input, free energy devices are possible, action at a distance is possible, communication faster than light speed is possible, etc.

By using the zero-vector approach, the virtual state can be organized and made **largely** deterministic, rather than statistical. This means that the probabilities of the states propagated forward by the Schroedinger equation can be engineered and changed. Whether or not a certain quantum change shall emerge or not can be determined or substantially influenced in advance. Bohm's hidden variable **theory** now becomes directly engineerable. This is a drastic change to quantum mechanics and physics in general.

Another implication is that this is the final engineering, for it allows the direct engineering of physical reality itself. Humans must now find a way to resolve their differences peacefully, or shortly Man will destroy himself and his biosphere by his own hand.

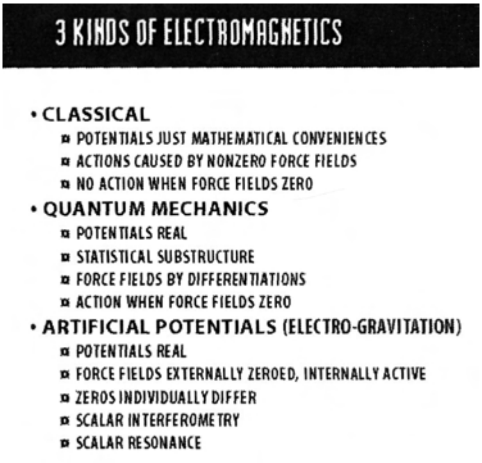

3 KINDS OF ELECTROMAGNETICS

- CLASSICAL
  - POTENTIALS JUST MATHEMATICAL CONVENIENCES
  - ACTIONS CAUSED BY NONZERO FORCE FIELDS
  - NO ACTION WHEN FORCE FIELDS ZERO
- QUANTUM MECHANICS
  - POTENTIALS REAL
  - STATISTICAL SUBSTRUCTURE
  - FORCE FIELDS BY DIFFERENTIATIONS
  - ACTION WHEN FORCE FIELDS ZERO
- ARTIFICIAL POTENTIALS (ELECTRO-GRAVITATION)
  - POTENTIALS REAL
  - FORCE FIELDS EXTERNALLY ZEROED, INTERNALLY ACTIVE
  - ZEROS INDIVIDUALLY DIFFER
  - SCALAR INTERFEROMETRY
  - SCALAR RESONANCE

Slide 004: Three Kinds of Electromagnetics

There are actually three kinds - or three views - of EM. These are (1) the classical view, (2) the quantum mechanical view, and (3) the scalar EM or electrogravitational view.

In the classical view, the potentials are just mathematical conveniences and do not physically exist. The real causative agents are the force fields, and there is no longer any electromagnetics going on if the force fields reduce to zero. Further, the ideas of "charge" and "charged mass" have been made erroneously synonymous.

Of course the classical view was formed from the idea of a thin material ether, with electricity as a thin fluid, long before the discovery of the electron. Since most earlier scientists studied string waves and these are transverse, the EM wave was modeled as a transverse wave. Also, detection equipment actually detected transverse waves. The role of electron spin and drift velocity, which would have shown that force-field-causing EM waves in the vacuum could only be longitudinal, was not yet discovered. Maxwell's equations and the classical approach became so engrained that the basic derivations were never corrected for more modern discoveries.

The quantum mechanical view, on the other hand, regards the potentials as the real physical actants, and the force fields are just effects derived from the potentials by differentiating operators. Classically oriented physicists have adamantly opposed this foundations requirement of QM because it would require nearly a complete redo of EM theory. It would also rather drastically change our ideas of physical reality. For years a controversy has raged around the Aharonov-Bohm effect (which demonstrates the reality of the potentials, among other things). Only this year - 1986 - have most physicists finally accepted the AB effect with its implications (see Physics Today, Jan. 86). However, no changes have yet been made to EM theory and the basic classical approach to electrical physics and engineering.

Yet even the QM view is flawed, since it has not examined the structure of EM forces which sum to a vector zero. Such a system produces stress, and if the summation is in the vacuum itself, it produces stress of vacuum/spacetime. Rigorously, this is a gravitational effect, and the energies of the various EM components in the local region are locked into an artificial potential. From general relativity, this type of potential where the energy density of vacuum is altered is a gravitational potential. From Kaluza-Klein unified theory, it is at least a 5-dimensional gravitational potential.

If the individual force vector components of the vector zero are varying in magnitude - say, all in phase - then they produce a gravitational wave. The energy density of the local vacuum is being rhythmically varied. Call such an EM wave a scalar EM wave, where, by "scalar" we imply that, to an external observer, the EM force vector resultants are identically zero, but the local gravitational potential of the wave is varying. Thus this is an electrogravitational wave, and vector zeroing of EM force fields constitutes a means of changing EM field energy into G-field energy. On the other hand, by breaking up the coherence of the zero-vector summation of the EM forces, nonzero EM resultants are recovered, constituting the change of gravitational energy into EM energy.

A simple means to break zero-summed coherence is by interference of two or more such zero-vector waves.

In this, third view of EM, action at a distance is easily possible, and is the norm rather than the exception. In addition, a new kind of resonance - scalar resonance - exists. The scalar EM wave does not interact with orbital electrons, but rather with the interior of the nuclei of atoms. Thus the new scalar EM resonance is between nuclei and within nuclei.

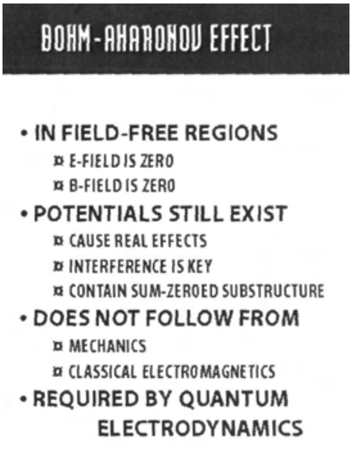

BOHM-AHARONOV EFFECT

- IN FIELD-FREE REGIONS
  - E-FIELD IS ZERO
  - B-FIELD IS ZERO
- POTENTIALS STILL EXIST
  - CAUSE REAL EFFECTS
  - INTERFERENCE IS KEY
  - CONTAIN SUM-ZEROED SUBSTRUCTURE
- DOES NOT FOLLOW FROM
  - MECHANICS
  - CLASSICAL ELECTROMAGNETICS
- REQUIRED BY QUANTUM ELECTRODYNAMICS

Slide 005: Bohm-Aharonov Effect

In 1959, Aharonov and Bohm published a fundamental paper in Physical Review, which pointed out the QM implications of potentials as the real entities, while force fields were derived effects. They showed that, even in the presence of zero EM force fields, the potentials may still exist and produce real effects in physical systems. They also suggested experiments to prove these predictions.

Interference of the potentials is the key mechanism producing real effects in charged particle systems, even in the presence of zero E-field and zero B-field. Unfortunately the AB paper did not address the issue of zero-summed systems of EM force fields as one way of producing artificial potentials having unique characteristics

At any rate, the "AB effect," as it came to be called, was rather hotly resisted and debated over the years, until finally it has generally been accepted as proven in 1986. The principle represents a violation of both classical mechanics and classical electromagnetics. It is required, however, by quantum mechanics and quantum electrodynamics. And experiment has proved it.

(See Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), Aug. 1, 1959, p 485-491.).

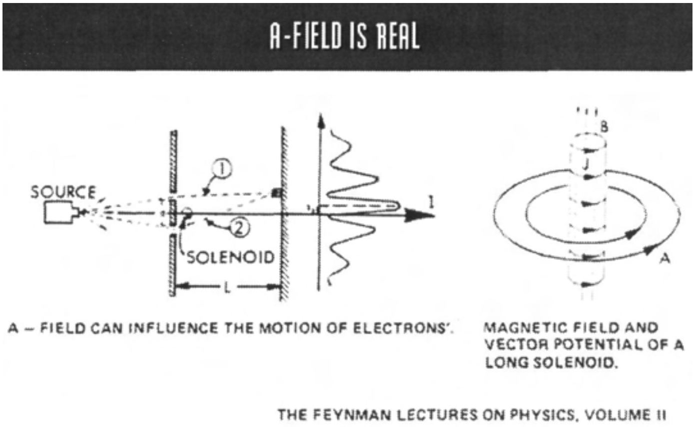

A-FIELD IS REAL

A – FIELD CAN INFLUENCE THE MOTION OF ELECTRONS'.

MAGNETIC FIELD AND VECTOR POTENTIAL OF A LONG SOLENOID.

THE FEYNMAN LECTURES ON PHYSICS, VOLUME II

Slide 006: A-Field is Real

In classical EM, the vector magnetic potential (the A-field) had been defined as a mathematical convenience by the equation

$$\nabla \mathrm{x} A = B \qquad [1j$$

But if the potentials are real, then conceivably the A field can be loosed from its enchainment to the $\nabla \mathrm{x}$ operator. In that case, it becomes a free, new, and independent field of nature with potentially unique characteristics. For example, its defining equation shows that magnetic force field can be made from it, and the rightmost term of the equation

$$E = -\nabla \varnothing - dA/dt \; [2]$$

shows that its time rate of change makes an electric field.

Let us explain this in more simple terms, and somewhat more precisely than conventional theory. We will use the Kaluza unified G-EM interpretation and electron flow in our explanation.

In the A-field, we have a certain kind of 5-dimensional G-potential which can bleed-off as EM force fields in two ways: (1) in a swirl fashion, where the vortex producing the swirl moves parallel to electron movement, and the swirling is a torque or spin, and (2) in a linear fashion, where the time rate of change of the A-potential produces a linear E-field on the electrons.

The first bleed-off as given by equation [1] constitutes the magnetic B-field, and the second bleed-off as given by the rightmost term of equation [2] creates a component of the overall electrical E-field. (Bleed-off of the electrostatic scalar potential produces the other component of the E-field).

Now in general relativity (GR) theory, "the" G-potential is just a conglomerate of many things, each of which has the characteristic of curving spacetime. "The" gravitational field is not a single thing at all, but is composed of a collection of many things.

Thus if we realize that both the electrostatic scalar potential (Ø-field) and the magnetic vector potential (A-field) are components of the 5-d G-potential, then we see immediately that bleed-off of these two components of the 5-d G-potential creates all normal EM force fields.

It follows that, if we produce a zero-vector summation of the two or more EM bleed-offs, we are actually "putting as much back in" to the 5-d G-potential through its A and Ø components as we are taking out electromagnetically. In that case, the 5-potential is in a state of equilibrium with respect to EM bleed-off. It is now forced to bleed-off in the only other way it can: as ordinary 4-dimensional gravitational field. Thus by vector-zeroing EM force fields, we turn EM field energy into G-field energy and vice versa, via the intermediary of the 5-potential.

At any rate, soon after publication of the Aharonov-Bohm paper, experiments showed that, if the magnetic field is trapped inside a long solenoid, a phase shift still is induced in the two-slit electron experiment, even though - classically - no contact of the enclosed magnetic field and the moving electrons occurs. This phase shift is explained by the fact that the freed A-field exists outside the trapping solenoid, even though the B-field does not. Consequently, interaction of this free A-field with the electrons produces a phase shift of the QM interference detection pattern.

This proves that A-field is real and causes physical effects.

It also proves that a form of action at a distance is real, just as required by quantum mechanics.

Years ago, Frank Golden and this author - together with Dr. William Tiller - experimented with "free A-field" devices. Golden went on to develop prototype transmitters and receivers and a prototype underwater communication system.

Since that time, Gelinas has patented several curl-free magnetic vector **potential** (free A-field) devices: see U.S. patent no. 4,447,779, May 8, 1984; 4,429,288. Jan. 31, 1984; 4,429,280, Jan. 31, 1984; and 4,432,098, Feb. 14, 1984. These patents are assigned to Honeywell.

(See also Theodor Kaluza, Sitz. Berlin Preuss, Adad. Wiss. 966, 1921.).

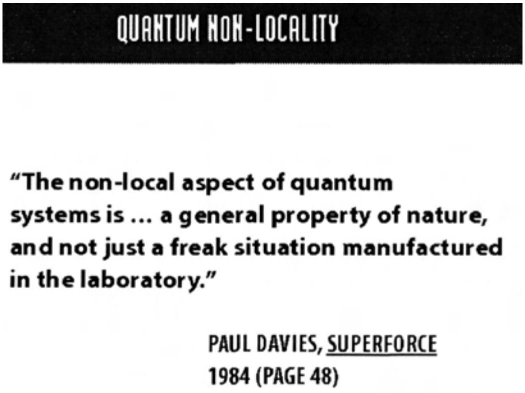

QUANTUM NON-LOCALITY

"The non-local aspect of quantum systems is ... a general property of nature, and not just a freak situation manufactured in the laboratory."

PAUL DAVIES, SUPERFORCE
1984 (PAGE 48)

Slide 007: Quantum Non-Locality

Indeed, quantum systems exhibit nonlocal effects that are general properties of nature. These effects are not just freak situations manufactured in subtle laboratory experiments.

However, if spatial coherence in such microscopic effects can be induced and sustained, then action-at-a-distance can be obtained in large macroscopic systems. This is the engineering aspect of the scalar EM approach.

It is also the aspect that the Soviets have secretly weaponized for over three decades.

The zero-vector EM force-field summation of multiple EM waves is used to transport specific time-phased patterns of EM polarization of vacuum to a distance. There interference of two or more such patterns results in the interference of the internal patterns, breaking of the phase relations, and appearance of positive (added) or negative (extracted) EM energy in the intersection region.

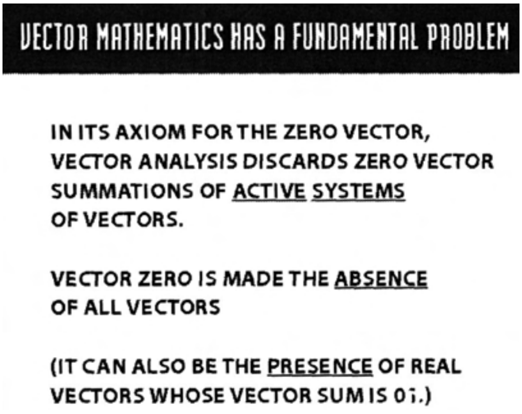

VECTOR MATHEMATICS HAS A FUNDAMENTAL PROBLEM

IN ITS AXIOM FOR THE ZERO VECTOR,
VECTOR ANALYSIS DISCARDS ZERO VECTOR
SUMMATIONS OF ACTIVE SYSTEMS
OF VECTORS.

VECTOR ZERO IS MADE THE ABSENCE
OF ALL VECTORS

(IT CAN ALSO BE THE PRESENCE OF REAL
VECTORS WHOSE VECTOR SUM IS 0ī.)

Slide 008: Vector Mathematics has a Fundamental Problem

In its concept of the zero vector, vector mathematics discards zero-vector summations of active systems of vectors. It replaces such a summation with a zero vector. This is fine for mathematics as an abstract system, but it is in error when applied to real electromagnetic force fields of nature.

In the abstract mathematics, a vector zero summation is made the "absence of all finite vectors". Further, all vector zeros are made equal. No concept of the "internal stress" of the zero vector exists in abstract vector mathematics.

However, physically the zero summation or "balancing" of vector forces in a medium represents stress in that medium. In the physical case, a vector zero summation system of non-zero vectors has a dynamic substructure, and this substructure is an individual.

Obviously, in the physical case vector zero summations may materially differ, both in the pattern of stress and the magnitude of stress. They cannot all be equated. Further, they are not the absence of vectors, but represent the presence of vectors in a special manner. Such a zero summation of EM vector forces directly stresses the "medium" - which may be the vacuum itself.

In the physical case, several changes to the axioms of abstract vector mathematics are required. (1) the "potential" of a vector zero must be taken into account, such as is represented by the sum of the squares of the magnitudes of its vector components. (2) the specific deterministic pattern of the vector components comprising the zero must be taken into account. (3) The dynamic variation in both the deterministic directions and deterministic magnitudes of the components and of the overall pattern must be taken into account. (4) Frequencies of the changes in the direction, magnitude, and actual makeup of the vector zero must now be accounted for. That is, time and wavelengths are rigorously aspects of the vector zero, and these may be deterministic variables. Since time itself is now a variable aspect of the vector, the vector zero system can affect its "rate of time flow" in the observer's system. (5) Since a "reference vector zero" can be established at any point of a vector magnitude, then individual vectors themselves may have dynamic substructures inside a special "zero reference" in and on the vector. The patterned potential of a vector is a reality.

This leads to a system of "vectors nested inside vectors" ad infinitum. In other words, it leads to an infinite-dimensional system, and the "opening" of every finite closed vector system through its vector zeroes.

For application to physical electrogravitational systems, at least 5 dimensions are required, four of space and one of time.

Here we still are considering only a special case where all vector zero summations represent EM force field energy locked in a gravitational potential, or gravitational force field energy locked inside the EM potentials. That is, we are prescribing a system where the only intertranslation of energy is among mass, EM field, gravitational field, and vacuum virtual energy (anenergy).

(See T. E. Bearden Toward a New Electromagnetics: Part III: Clarifying the Vector Concept, Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1983; Part 4: Vectors and Mechanisms Clarified, 1983. See also E. T. Whittaker, Proc. Lond. Math. Soc. 1, 367, 1903; Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover, New York, 1963, p. 283-287; Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley, New York, **Vol.** I, p.2-4.).

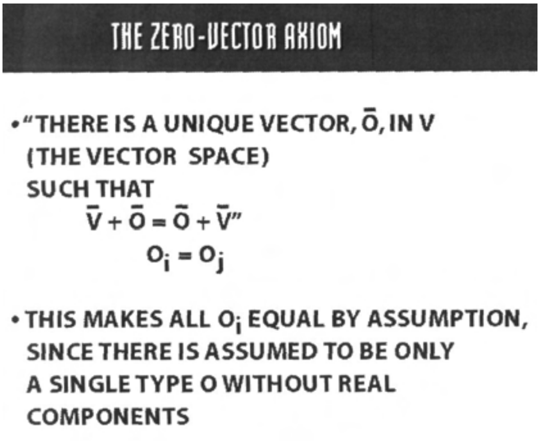

THE ZERO-VECTOR AXIOM

- "THERE IS A UNIQUE VECTOR, $\bar{O}$, IN V (THE VECTOR SPACE) SUCH THAT

$$\bar{V} + \bar{O} = \bar{O} + \bar{V}"$$

$$O_i = O_j$$

- THIS MAKES ALL O_i EQUAL BY ASSUMPTION, SINCE THERE IS ASSUMED TO BE ONLY A SINGLE TYPE O WITHOUT REAL COMPONENTS

Slide 009: The Zero-Vector Axiom

This slide shows a standard mathematical statement of the zero-vector axiom, used in many texts.

As can be seen, all zero-vectors are made identical, and no substructure is considered.

Such an abstract system of vector mathematics does not fit physical reality. However, when all one's EM foundations concepts are strongly conditioned in this fashion, the gravitational aspects of EM are discarded.

Further, EM theory, experiments, and equipments will be developed along these limiting lines. In that case, EM engineering of gravitational effects will not be developed. One will assume that it is impossible to have a locally curved spacetime, and hence all local frames will be Lorentz frames. Therefore the conservation laws will rigorously apply.

Further, the nonlocal quantum effects will stay firmly in the microscopic world where they belong. Physical macro-reality will stay stable and predictable in a classical manner.

General relativity (in curved spacetimes) will be restricted to the special case of the local special relativistic (uncurved) spacetime, with only distant perturbations (spacetime curvatures).

The world will stay sane and "normal", and the totally variable insanity of physical reality will remain bottled up in the minute microworld or the distant maelstrom of suns, stars, black holes, etc.

The Aharonov-Bohm heresy will stay kaput in the real world and will not be unleashed beyond mesoscopic reality — a few thousand angstroms or so.

With such an error in the application of vector mathematics to physical science, we become "flatland physicists," so to speak. We become like skimmer bugs on the surface of a pond; we have no appreciation of the dynamics of the depths below or of the heavens above.

Worse, we condition our development of instruments according to the fixed notions in our heads. Then we use these "biased" instruments to do experiments that reaffirm our notions. Any deviation suggested to this is considered heresy and nonsense.

And as Max Planck once pointed out, you will get a new physics only when the old physicists — who so adamantly oppose it — die off.

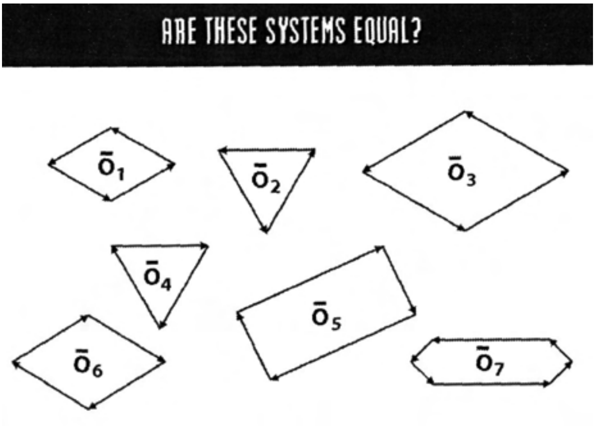

ARE THESE SYSTEMS EQUAL?

Slide 010: Are These Systems Equal?

Shown here are some representations of various very simple zero-vector systems. As can be seen, the stress magnitudes and the spin and dynamics of all of these systems are very different.

Further, their "stress potential" substructures are not composed of a huge collection of little random force vectors. Instead, their substructures are deterministic. If the components being put into the zeros are varied in direction and magnitude, but always sum to a zero resultant, then highly complex substructure dynamics may be created and utilized, infolded (Bohm's term) inside the zero vector.

In theory one can establish "channels" of communication and power (magnitude) transmission without ever surfacing a nonzero EM force field. In practice one can do this with very little EM spillout. Thus one can establish electrogravitational channels and energy transmission inside normal EM potentials and/or ordinary EM carrier waves.

One can use a conventional system of EM potentials, force fields, and waves as a special kind of "wiring circuit" in and through which to transmit and produce electrogravitational effects. By distant interference, EM effects at a great distance can be accomplished, yet no "ordinary EM energy" has passed between the transmitters and the distant interference zone.

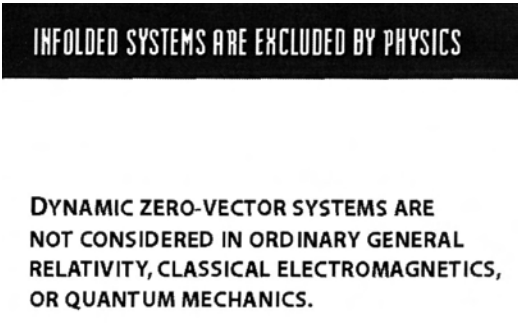

INFOLDED SYSTEMS ARE EXCLUDED BY PHYSICS

DYNAMIC ZERO-VECTOR SYSTEMS ARE NOT CONSIDERED IN ORDINARY GENERAL RELATIVITY, CLASSICAL ELECTROMAGNETICS, OR QUANTUM MECHANICS.

Slide 011: Infolded Systems are Excluded by Physics

Unfortunately, these infolded dynamic zero-vector systems have been eliminated from physics.

They are omitted in classical EM, ordinary general relativity (OGR), and quantum mechanics.

That is, they have been omitted in Western physics.

They have not been omitted in secret Soviet weapons labs and weapons development programs.

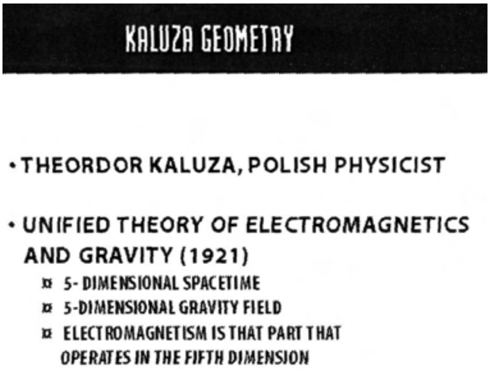

KALUZA GEOMETRY

- THEORDOR KALUZA, POLISH PHYSICIST
- UNIFIED THEORY OF ELECTROMAGNETICS AND GRAVITY (1921)
  - 5- DIMENSIONAL SPACETIME
  - 5-DIMENSIONAL GRAVITY FIELD
  - ELECTROMAGNETISM IS THAT PART THAT OPERATES IN THE FIFTH DIMENSION

Slide 012: Kaluza Geometry

In 1921, Theordor Kaluza, a Polish physicist, published a unified theory of electromagnetics and gravitation. Albert Einstein, who had had the paper for two years, recommended his paper for publication. In that theory, five dimensions - four space and one time - are utilized in the basic model. This gives a 5-dimensional spacetime.

In the model, electromagnetics and ordinary gravity are two aspects of a single more fundamental field: the 5-space gravity field. Electromagnetism is that part of the 5-field that operates (bleeds-off) in the fifth dimension.

The ordinary 4-space G-field is the small residue that spills over into our ordinary 4-space, and doesn't slide around into the fifth dimension and bleed-off there as electromagnetics.

Implicit in the theory, but not explicitly stated, is the fact that the 5-space G-potential bleeds-off in two fashions: (1) in the 5th dimension nearly unopposed, to produce what we see as EM force fields, and (2) in the ordinary 4-space we live in, which we see as ordinary gravitational field.

Normally almost all the force field bleed-off is in the fifth dimension as EM; only a minuscule bit bleeds off as 4-gravity.

Just as one example, between two electrons the 5-space G-potential bleeds-off into electric force field about 10^{42} times as much as it bleeds-off into 4-space G-field.

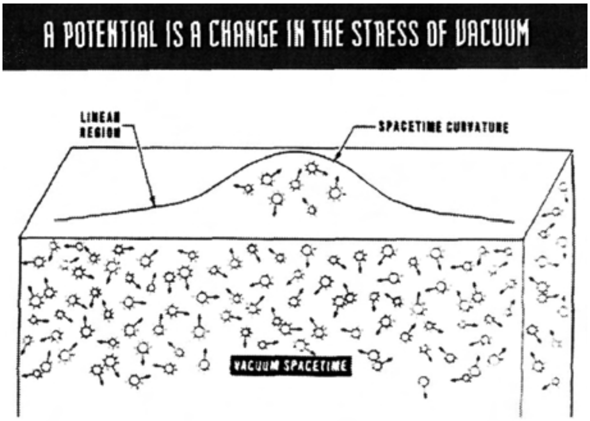
A POTENTIAL IS A CHANGE IN THE STRESS OF VACUUM

Slide 013: A Potential is a Change in the Stress of Vacuum

In physics, the idea of a "potential" is very poorly defined, if at all. The normal "definition" for the electrostatic scalar potential, for example, is not a definition at all. Instead, it's an abstract operation that tells how to mathematically calculate the magnitude of the potential, by pushing a unit charge in from infinity, against the potential field.

Now ultimately a definition must be an identity statement, not an operational statement about something.

Since I could find hardly anyone who actually understood what a potential was, and how or of what it was composed, I simply took quantum mechanics at face value and constructed this model.

First, the modern view of the vacuum is that it is not an emptiness, but a plenum. It's teeming with an incredible amount of raw energy, existing totally as the temporary bits of energy of virtual particles. That is, the vacuum "energy" exists as the nonintegrated energies of little particles which appear from nothing fare spontaneously created) and disappear into nothing (are spontaneously annihilated) almost instantly. This happens so fast that the little particle and its energy cannot individually be detected; it doesn't hang around long enough.

Hence the fleeting little booger is called a "virtual" particle. Like, "We had 'im there man, for just a moment, but he was gone before we could grab 'im!"

This kind of ghostly particle, however, is very real in modern physics. All observable forces are theorized to be due to the integration or collection of a large number of virtual forces created by exchange of these little particles.

However, this implies a peculiar thing about the vacuum "energy." Actually it's vacuum energies, for each little bit of energy only exists momentarily, then returns to the nothingness from which it came. That is, vacuum virtual energies are totally unintegrated. Each "bit" of virtual energy exists individually. These virtual energies cannot be "added up" to compute the "available energy" unless some collector or integrating mechanism puts them together.

That is, unless some kind of "zipper" is there to "zip them together" into macroscopic, observable energy. The usual "zipper" is a spinning, observable particle of mass. We say that the vacuum energy is unzipped and hence unobservable in that state. We say that "observable energy" is just zipped-together vacuum energies - integrated by and on a spinning, observable particle of mass.

So the vacuum can be visualized as a special sort of violent, unzipped plasma, where the particles of the plasma are fleeting and do not last. We can speak of it also as a sort of special unzipped gas, in the same manner.

The "pressure" of the vacuum gas we refer to as "stress." Due to its violent boiling activity, the vacuum gas is always under very high stress.

This "stressed virtual plasma" identically IS the vacuum and identically IS spacetime. Length, time, frame, mass, etc. - all arise by zipping operations (or understood zipping operations) in and of this raw vacuum/spacetime. When one uses the vacuum notion in this manner, one now uses the term "ether" again. However, this is not the old "thin material ether that hangs around"; this is an entirely new kind. It's a virtual flux and a virtual plasma.

The ether is also what we call "spacetime" in relativity. As can be seen, spacetime is stressed. Further, spacetime has a substructure.

By an uncurved spacetime, we mean one in which the stress of these virtual particles (the "pressure in the ethereal gas," so to speak) is constant from one place to another and from one time to another. Indeed, we also should mean **that** the stresses of the individual components of the entire plasma are constant from one place to another and from one time to another.

By definition, when we determine the lengths between every two points, we determine a frame. (Actually, we have prescribed a universal "length zipping" operation/mechanism as well as its characteristics.). A "linear frame" or "Lorentz frame" is one in which the pressure or stress remains constant. It may be higher than ambient vacuum, however, due to a particle's or an object's motion through the observer's ether. In the case of constant velocity, the particle/object encounters a higher level of virtual particle flux in the observer's ether, like a vehicle moving through raindrops. It encounters a constant rate of flux, however. In such constant velocity case, the moving object's frame is said to be "rotated," but not curved, with respect to the laboratory frame of the observer.

In a "curved" spacetime, we mean one in which the stress of vacuum increases or decreases from one place to another, or from one time to another, or both. At one fixed location, if spacetime is curved, then the stress is increasing or decreasing as a function of time. That is, the flux density of that area is changing. We may visualize the local frame as undergoing angular acceleration, rotating more or rotating less. This is a non-Lorentz frame, or curved spacetime. It is also a nonlinear vacuum.

In such a vacuum, local gradients exist in the flux of virtual particles. If affected by (coupled to) one or more of these gradients by a nonzero coefficient, a local object experiences effects (forces and actions) "without observable cause." Observably, conservation laws may be locally violated by these "virtual rivers" of vacuum by coupling, just as a paddlewheel dipped in a river violates its own local "conservation of energy" due to the river providing an energy source to it.

One must be very careful with the "pressure" analogy, however, when one utilizes unified theory (Kaluza theory). We now have a virtual plasma vacuum in five dimensions, not four. But the basic notions to enable our visualization and understanding still apply.

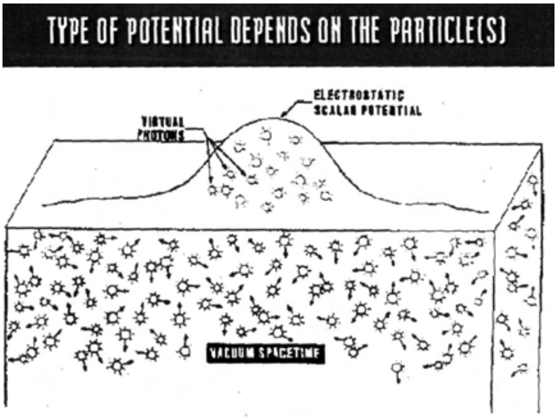
TYPE OF POTENTIAL DEPENDS ON THE PARTICLE(S)
ELECTROSTATIC SCALAR POTENTIAL
VIRTUAL PHOTONS
VACUUM SPACETIME

Slide 014: Type of Potential Depends on the Particle(s)

Now one can see that, by choosing the type or class of virtual particle, and looking at its potential or stress in the vacuum, one can have many kinds of "potentials" existing in the overall "gravitational potential" of vacuum spacetime.

For example, the stress of the virtual photon flux (which is what causes electrical charge) is called "electrostatic scalar potential."*

We can choose any other class of virtual particle we wish, and there exists a potential (and "change") for it. There are quark potentials, neutrinic potentials, etc.

Indeed, one of the reasons for ignoring Kaluza unified theory for so long was that it predicts a great many types of potentials and fields which have never been observed and which are unknown. Approximations, which neglected all these mysterious, and unknown fields, gave erroneous results. Perplexed physicists, seeking to simplify matters, simply turned away from it in frustration. They returned to it in the mid-70's when it appeared that 11-dimensional Kaluza-Klein theory possibly offered a complete unified theory.

Actually, the diversity of new fields in Kaluza theory was its strength not its weakness. That they are presently unobservable and undetectable is beside **the** point. Neither is time directly observable; however, we have worked out handy procedures and measurements of spatial quantities, which allow us to infer **time.** We have not just tried to drop it from physics!

And indeed the Kaluza theory is consistent with experiment when the higher fields are included. The message is, we must work out "clocks" for these new potentials and fields, so that they can be inferred with at least some degree of precision

- The character of movement also determines the type of potential. The longitudinal (radial) movement of the virtual photon flux is implied in electrostatic scalar potential. The swirl (tangential) component of this flux might be taken as the magnetostatic scalar potential - a magnetic pole - when the cw and ccw components of the swirl balance, but the sum of their absolute values differs from the sum of the absolute values of the corresponding tangential components in the swirl flux of the local ambient vacuum. If this swirl stress is higher than the vacuum swirl stress, this constitutes a north magnetic pole. If lower, it constitutes a south magnetic pole. If there is a prevalence of direction (cw or ccw) in the swirl (i.e., one component exceeds the other), this corresponds to the magnetic vector potential (the A-field) when the A-field is freed from the magnetic force field (B-field) — i.e., when A-field is freed from its $\nabla\times$ operator in the equation $\nabla\times A = B$. The bleed-off of the excess swirl stress from the north magnetic pole to the deficient swirl stress of the south magnetic pole constitutes the magnetic force field, B — usually referred to simply as the "magnetic field."

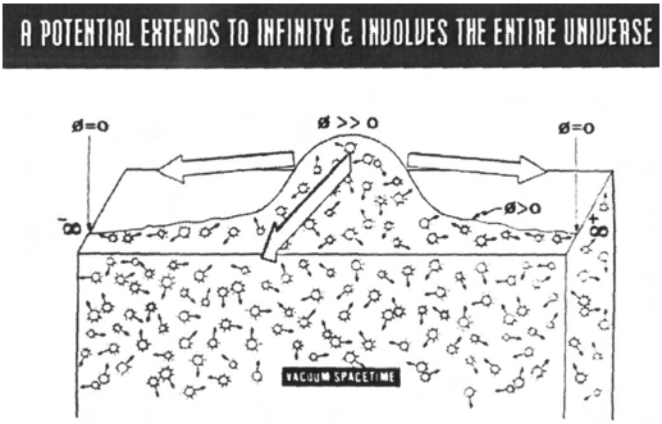

A POTENTIAL EXTENDS TO INFINITY & INVOLVES THE ENTIRE UNIVERSE

Slide 015: A Potential Extends to Infinity & Involves the Entire Universe

As shown on this slide, the change in the stress of vacuum due to a potential tapers off to infinity. Its magnitude normally does not reduce to zero until infinity.

Some types of potentials reduce much faster or slower than others, however, and so - at some distance away from a common origin of the potentials, one or more may be neglected in its (their) overall effects on an experiment or system.

Also, the speed of propagation of potential is highly dependent upon (1) what type of virtual particle or particles produce its stress, (2) its magnitude, (3) its composition, and (4) the presence or absence of coupling objects and other interacting potentials.

Fundamental observable particles are like little spray pumps and little vacuum cleaners at the same time.

Each is emitting virtual particles into the vacuum in a spray or flux, and at the same time it is absorbing virtual particles from the vacuum spray or flux. It's a little dynamo, and nature furnishes its driving energy unceasingly.

However, each little virtual particle it emits is itself such a little simultaneous pump and vacuum cleaner. That is, it's also emitting an even finer (and faster) flux of smaller virtual particles, and at the same time it's also absorbing flux al this time finer level.

"Bigger fleas have smaller fleas to bite em;

And so on ad infinitum".

We may consider the vacuum to be made of an infinite number of virtual state layers or levels.

It is not too difficult to show that each of these successively deeper, nested "layers" of virtual state corresponds to a successively higher dimension being added to our 4-space basic spacetime. So hyperdimensions and deeper levels of virtual state are one and the same thing.

The basic "speed" of the first layer is c, the speed of light.

The basic "speed" of the second layer is c^2.

And so on.

This is interesting. We may directly engineer the virtual state by means of the vector zero (scalar electromagnetics) approach. By nesting vector zeros inside other vector zeros, we may directly engineer the deeper layers of virtual state, and consequently hyperdimensions.

Scalar electromagnetics thus is virtual state engineering and hyperspatial engineering at one and the same time.

Superluminal communications systems, hyperspace drive, and materialization and dematerialization are all hypothetically possible, using scalar electromagnetics. As the technology develops, we should see the development of many of the systems long thought impossible except in science fiction.

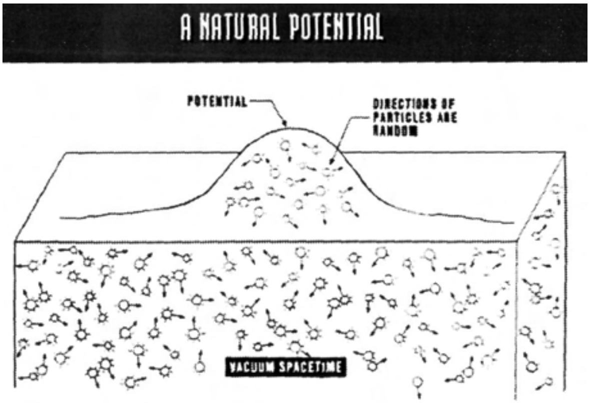

Slide 016: A Natural Potential

There are two broad categories of EM potentials, depending upon the way in which they are formed.

The two categories of potential formation are: (1) natural, and (2) artificial.

The difference is in the virtual particle flux substructure that comprises the potential.

If the substructure of emerging and vanishing virtual particles is random and without deliberate order (polarization) except in one broad, overall sense, then that is a natural potential. As such, there is no deterministic functioning going on inside the substructure of the potential itself. The potential can only act "as a whole" on a charged particle system.

Natural potentials have substructures of random virtual vector components. Any small amount of coherence as a function of distance vanishes quickly (on the order of a few thousand angstroms at most, according to present Aharonov-Bohm experiments).

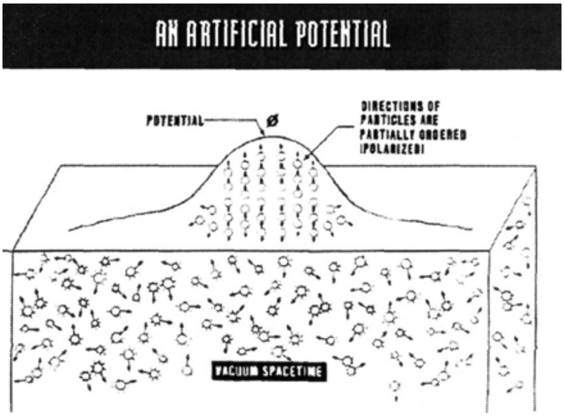

Slide 017: An Artificial Potential

An artificial potential by definition has a substructure composed of deterministic observable vector components, summed to an overall zero vector. Coherence as a function of distance can be maintained over enormous macroscopic distances - even hundreds of thousand of kilometers - by simultaneous transmission of the entire cluster of substructure components as a coherent "zero" group.

To the conventional linear EM detector, at any point along its transmission path, this substructure is detected as a zero-E vector and a zero-H vector. Hence the conventional detector does not see the artificial potential, even if its "stress magnitude" (a function of the magnitudes of all summed components) changes.

Action at a great distance is possible with this artificial potential, however, if a highly nonlinear situation is met so that the phasing of the components is broken or significantly altered. In that case, the components do not sum to zero after their dephasing, and real EM force fields emerge. This "dephasing" can be made to occur at the distant nonlinearity.

Action at a great distance is not possible with the natural potential, but is possible with the artificial potential.

Further, with the transmitted artificial potential, VO must be applied from the transmitter to the distant interruption zone, as if there were no intervening space between.

Note that VO may be positive or negative. Thus energy may emerge at the distant disruption, or be extracted from there. In the first case, energy is input to the transmitter, to re-emerge at the distant dephasing zone. In the latter case, energy is extracted (disappears) from the distant dephasing zone and is received (reappears) back at the transmitter

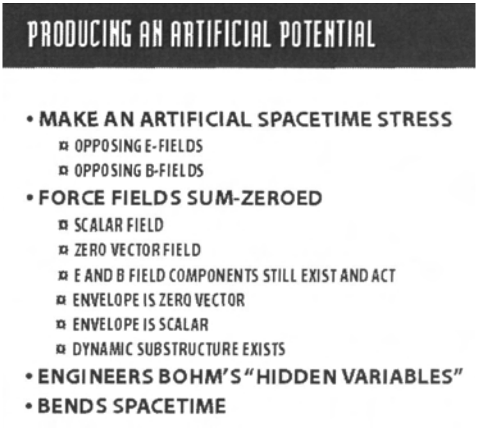

PRODUCING AN ARTIFICIAL POTENTIAL

- MAKE AN ARTIFICIAL SPACETIME STRESS
  - OPPOSING E-FIELDS
  - OPPOSING B-FIELDS
- FORCE FIELDS SUM-ZEROED
  - SCALAR FIELD
  - ZERO VECTOR FIELD
  - E AND B FIELD COMPONENTS STILL EXIST AND ACT
  - ENVELOPE IS ZERO VECTOR
  - ENVELOPE IS SCALAR
  - DYNAMIC SUBSTRUCTURE EXISTS
- ENGINEERS BOHM'S "HIDDEN VARIABLES"
- BENDS SPACETIME

Slide 018: Producing an Artificial Potential

AN artificial potential is deterministically patterned spacetime stress, made by opposing E-fields and/or B-fields so that they sum to vector zeros in a special pattern.

The resulting zero-summed envelope has no EM force field, to an external observer/detector.

However, the infolded E-field and B-field vector components still exist and act. They may dynamically vary, so long as their summation is always kept to zero.

The simplest variation is to vary all their magnitudes at once, by the same degree. In that case, each one comprises an "EM wave." However, the summations of this cluster or "locked group" of waves still exhibit a zero-E and zero-B field to any external observer/detector. In other-words, to an external observer, one now has a varying wave of pure spatiotemporal stress, but one that has a deterministic structure. This is a scalar EM wave, or electrogravitational wave. It is also an alternating current of specific scalar pattern.

Varying the stress of spacetime locally, curves it locally. This violates the conventional assumption of restricted general relativity that local spacetime is uncurved (is a Lorentz frame).

By use of scalar EM waves with deliberate substructures, one can engineer Bohm's "hidden variables" so that quantum mechanics becomes deterministic rather than statistical. This is a drastic change to the common (Bohr) interpretation of quantum mechanics.

And Einstein's intuition that God does not play dice with the universe turns out to be correct after all.

Since physicists haven't seen where the real game was being played, it has all seemed bewilderingly statistical to them.

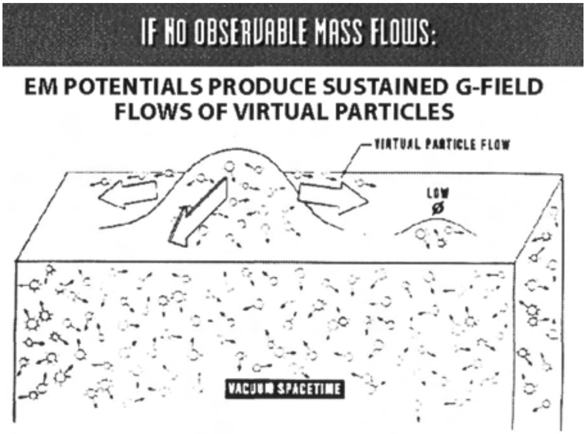
IF NO OBSERVABLE MASS FLOWS:

EM POTENTIALS PRODUCE SUSTAINED G-FIELD FLOWS OF VIRTUAL PARTICLES

Slide 019: If No Observable Mass Flows:

Since there are no observable mass particles flowing in vacuum, there are no observable force fields produced. Only virtual force fields can be produced in the virtual particle medium that is vacuum itself.

Thus (here is no observable E-field in the vacuum surrounding a charged point. There is, however, a virtual E-field there. When an observable charged particle such as an electron is introduced in the surrounding vacuum and "couples to" (integrates, zips together) the virtual E-field, an observable E-vector force is produced on and of the observable charged particle.

(For the classically trained electrical engineer or electrical physicist, the statement that no E or B fields as such exist in vacuum, is usually bewildering. Let me point out that any so-called "vector field" (such as the assumed E and B vacuum fields) can mathematically be replaced with two scalar fields. See E. T. Whittaker, Proc. Lond. Math. Soc.l, 366, 1903. What we are saying here is that the definition of an E field is in terms of force per unit charged mass. The scalar fields - which are actually what exist in vacuum - provide only a virtual or unzipped E field, until an observable spinning particle couples to both of them by virtue of its "dynamo flux pump" action. The ensemble, then, of two electrostatic scalar potentials coupled to a spinning observable particle, constitutes and is the so-called "E-field.")

When detecting these "vacuum E-fields," it is the free electrons in the electron gas in our probe or antenna which couple to the two "Whittaker scalar fields". These electrons produce ensembles that interact with each other collectively.

Charges in this electron-gas coupled medium is what we actually detect - not "what is in the vacuum."

To repeat this again for emphasis: The usual detector is an "electron-wiggle" detector. It detects changes in its own conduction electron gas, not in the vacuum itself. That is, the disturbance in the virtual-particle vacuum interacts with the observable particles of the electron gas, if (and only if) the spins of these observable particles couple the vacuum disturbance to the electrons and integrate their virtual components. After integration, an observable disturbance of the integrating object - say, the electron - results. It is this electron gas disturbance that is "detected" by almost all orthodox EM detectors.

If the spin of the observable conduction electrons cannot couple the electron to the vacuum flux disturbance, then the normal "simple" detector will not detect the disturbance. (By a "simple" detector we mean one in which the electrons are able to couple to the vacuum disturbance! Little circular definition here!)

Now most potentials reach to infinity before reducing to zero. The change in the virtual particle flux intensity of vacuum - which comprises the potential - decreases with distance from the potential. The magnitude of the change tails off" toward zero as one approaches infinity.

To any finite distance, then, there exists a decreased gradient in the change of the virtual flux intensity of vacuum, from a potential. Therefore virtual particles are flowing in the direction of the decrease of this gradient, just as gas molecules flow from a region of high pressure toward a region of lower pressure.

So there is a virtual particle "river" in the vacuum between any two separated points of different potential magnitude. The type of potential determines the type of virtual particle(s) in the river.

If the potential is natural, the gradient river does not carry any coherent substructure.

If the potential is artificial, the gradient river carries the coherent substructure everywhere within it. In this case, there is a flow or "river" of the virtual structure of this coherent pattern in the vacuum between two separated points of different "potential" in that pattern.

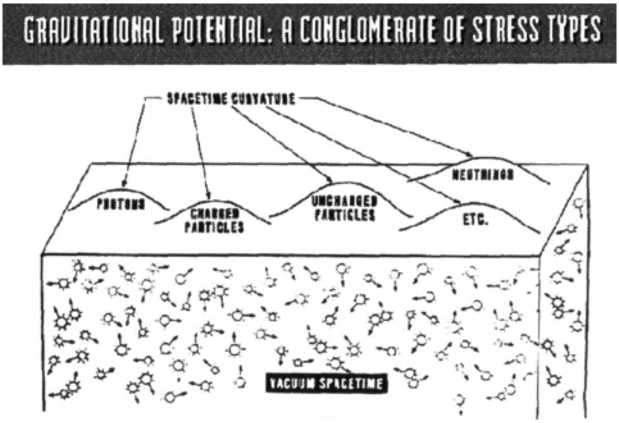
GRAVITATIONAL POTENTIAL: A CONGLOMERATE OF STRESS TYPES

SPACETIME CURVATURE
NEUTRINOS
PROTONS
CHARGED PARTICLES
UNCHARGED PARTICLES
ETC.
VACUUM SPACETIME

Slide 020: Gravitational Potential: A Conglomerate of Stress Types

Rigorously, all EM potentials are actually part of the overall 5-dimensional gravitational potential.

There is no such thing as "the" singular gravitational potential: instead, it is a conglomerate of different stress types and patterns, each of which is named for the type of action that produces it.

The movement of charged particles, for example, creates electromagnetic potentials. At any point in vacuum, the potential consists of a flux of virtual particles.

The vacuum contains a great variety of particle fluxes. There are virtual photons, virtual electrons, virtual positrons, virtual protons, virtual pions, virtual neutrinos, etc.

The conglomerate stress from all these virtual particle stresses constitutes the overall gravitational potential

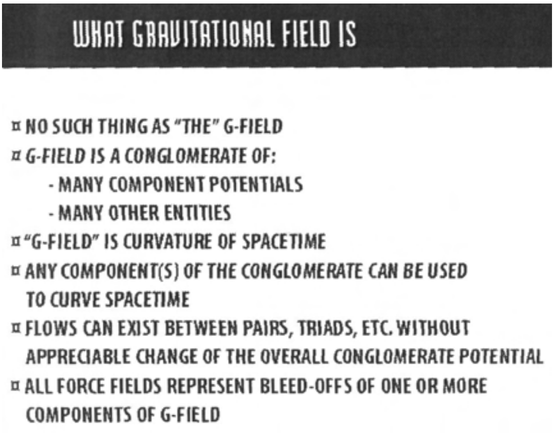

WHAT GRAVITATIONAL FIELD IS

¤ NO SUCH THING AS "THE" G-FIELD
¤ G-FIELD IS A CONGLOMERATE OF:
  - MANY COMPONENT POTENTIALS
  - MANY OTHER ENTITIES
¤ "G-FIELD" IS CURVATURE OF SPACETIME
¤ ANY COMPONENT(S) OF THE CONGLOMERATE CAN BE USED TO CURVE SPACETIME
¤ FLOWS CAN EXIST BETWEEN PAIRS, TRIADS, ETC. WITHOUT APPRECIABLE CHANGE OF THE OVERALL CONGLOMERATE POTENTIAL
¤ ALL FORCE FIELDS REPRESENT BLEED-OFFS OF ONE OR MORE COMPONENTS OF G-FIELD

Slide 021: What Gravitational Field Is

As previously stated, there is no such single thing as "the" gravitational field, as is assumed in the Newtonian sense.

G-field is a conglomerate of many component potentials and other entities.

Rigorously, anything that "curves spacetime" is a gravitational field.

In the modern view, all forces ultimately arise, by some means, from the curving of spacetime. Thus all forces and force fields are related to - and caused by, so to speak - gravitational field.

If the vacuum stress is uniform from one place to another in an observer's **frame** then the observer's spacetime is uncurved. The observer's frame, is said to be linear or uncurved; or a Lorentz, frame. In such a situation the conservation laws rigorously apply for macroscopic, closed systems.

If the vacuum stress is not uniform from one place to another, then the spacetime is curved. In such a situation the conservation laws may be violated for macroscopic systems, since they are "opened" to the vacuum stress, and vacuum virtual particle flux gradients exist across the system or portions of it.

Note that the overall stress magnitude may be constant across a region, but its composition may change. That is, two or more components may vary canonically. In that case, the spacetime is uncurved in the gross external sense, but contains specialized "internal" curvature patterns locally. Depending upon reaction to the canonical infolded curvature patterns, some macroscopic systems will still exhibit conservation, but certain other macroscopic systems may not.

An example is symmetry breaking. Here charge, parity, and time symmetry may be violated individually or in pairs, but not all three simultaneously. An overall conservation law is obeyed, but any two of the three conservation laws can be (und are) violated.

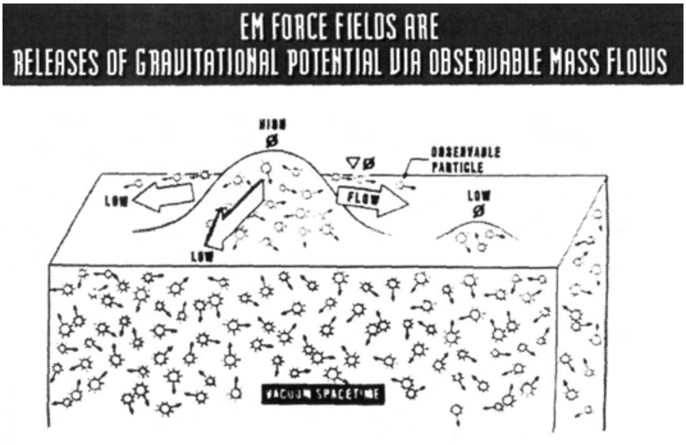

EM FORCE FIELDS ARE
RELEASES OF GRAVITATIONAL POTENTIAL VIA OBSERVABLE MASS FLOWS

Slide 022: EM Force Fields are Releases of Gravitational Potential via Observable Mass Flows

In Kaluza theory, we may take EM force fields to be the 5th dimensional "bleeding-off" of a 5-G potential. However, in vacuum alone this bleeding-off is totally within the virtual state. That is, only virtual EM force fields exist in vacuum.

By the ordinary term "EM force fields," physics has chosen to imply measureable or observable E and B force fields.

That implies that the 5-G potential gradient flows of virtual particle flux must be coupled to observable charged particles - such as electrons, protons, ions, etc. - before observable E and B force fields exist.

Rigorously, E and B force fields consist of the virtual E and B force fields coupled to observable charged particles. They do not "cause" a particle to accelerate; they are the product of the particle and its acceleration.

As these observable particles are themselves "spray pumps" producing virtual particle flux, then a collection of them represents an increased or decreased **flux** (for negative or positive charge) vis-a-vis the ambient background flux of vacuum. Nature furnishes the continuous power to drive these "virtual flux" pumps.

Normally, EM potentials are formed in natural and manmade electrical circuits by collections of these observable particle "spray pumps." Such collections of increased or decreased "spray flux" density, of course, represent 5-potentials, gravitationally.

Release of these pumps in conductors or in the vacuum itself results in creating E-fields and B-fields on and of the moving observable charged particles. But release of the observable-particle spray-pumps from the collection (the potential "pressure head" or "source") reduces the potential or "pressure" at the source.

Thus the formation of E and B fields represents the release of the 5-potentials. Hence EM force fields are the bleed-off of 5-dimensional G-potential via observable mass flow.

Since classical EM always seeks to create the force fields, it always bleeds-off and discards the gravitational effects that could be otherwise achieved.

To allow orthodox EM flow is to dissipate the gravitational effects.

Strangle the orthodox EM flows, and you can keep and use the amplified gravitational effects.

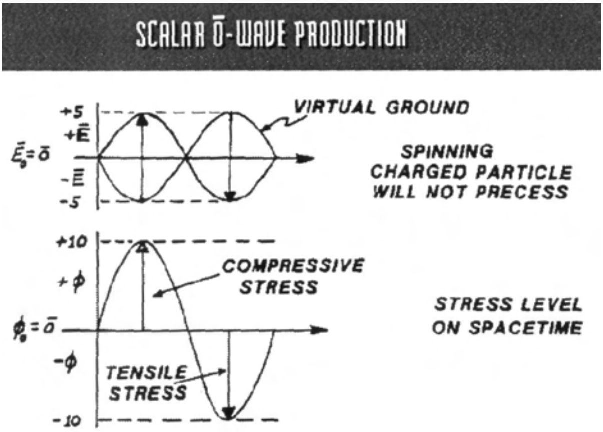
SCALAR Ō-WAVE PRODUCTION
+5
+Ē
$\bar{E}_0 = \bar{0}$
-Ē
-5
VIRTUAL GROUND
SPINNING
CHARGED PARTICLE
WILL NOT PRECESS
+10
+φ
$\phi_0 = \bar{0}$
-φ
-10
COMPRESSIVE
STRESS
TENSILE
STRESS
STRESS LEVEL
ON SPACETIME

Slide 023: Scalar O-Wave Production

We now visualize the formation of waves of pure stress in the spacetime medium (in the vacuum). These we call scalar EM waves, Tesla waves, electrogravitational waves, longitudinal EM waves, waves of pure potential, electrostatic/magnetostatic waves, and zero-vector EM waves. All these terms are synonymous. Each sheds its own particular light upon the nature of these waves or of their original discoverer, Nikola Tesla.

We use a gedanken experiment, or thought experiment, in which we can be "perfect."

Thus we have two single-frequency EM sine waves whose E-field componenls are shown on the slide. The two waves are of the same frequency, traveling together in the same direction, and superposed 180 degrees out of phase with each other.

In this case, at any spatial point, the summation E and B fields are vector zeros.

However, if we plot the spatiotemporal (vacuum) stress induced by the wave, wo see that it constitutes a sine wave, with compressive stress in one half cycle **and** tensile stress in the second half cycle.

Rigorously this defines a "longitudinal" EM wave - or what Tesla called a "sound wave in the nonmaterial ether."

Also, rigorously it is a gravitational wave, for it is a wave of the curvature of spacetime (nonlinearity of vacuum) itself. That is, what is changing in the wave is the vacuum stress, or intensity of the virtual particle flux of vacuum. That is identical to changing the curvature of spacetime. In one half-cycle, spacetime is curved positively. In the second half-cycle, spacetime is curved negatively.

Since the increase or decrease of the intensity of virtual particle flux (vis-a-vis the ambient vacuum's virtual flux intensity) represents electrical charge, then in one half-cycle negative charge is represented, and in the other half-cycle positive charge is represented. This directly explains the "positron-electron pair" representation of an electromagnetic photon - the "photon" being one wavelength.

In one half-cycle, time moves slower. In the second half-cycle, time moves faster. Thus scalar waves can also be considered to be "tempic" waves (to use Wilbur Smith's term), or oscillations of the rate of flow of time itself, about the ambient rate of time flow. Since these oscillations are variations in the curvature of spacetime, then they represent gravitational or force-generating waves, when coupled to a mass.

Variations in the rate of flow of time produce force, just as any other type of curvature in space-time does.

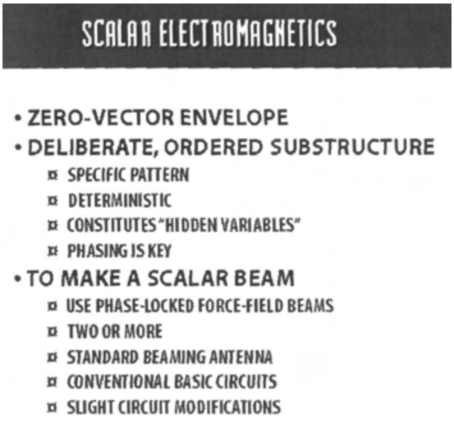

SCALAR ELECTROMAGNETICS

- ZERO-VECTOR ENVELOPE
- DELIBERATE, ORDERED SUBSTRUCTURE
  - SPECIFIC PATTERN
  - DETERMINISTIC
  - CONSTITUTES "HIDDEN VARIABLES"
  - PHASING IS KEY
- TO MAKE A SCALAR BEAM
  - USE PHASE-LOCKED FORCE-FIELD BEAMS
  - TWO OR MORE
  - STANDARD BEAMING ANTENNA
  - CONVENTIONAL BASIC CIRCUITS
  - SLIGHT CIRCUIT MODIFICATIONS

Slide 024: Scalar Electromagnetics

To rehash: scalar waves are zero-vector waves. Patterned E and B fields exist in a relationship where their summation equals vector zero. Electromagnetically, the zero EM "envelope" of a scalar EM wave is composed of a zero-summation of infolded finite EM force-vectors.

However, the internal (infolded) substructure is deliberate, macroscopic, and ordered. Vacuum is polarized deterministically by the wave, and macroscopic spacetime is locally curved by it.

These substructure components constitute Bohm's "hidden variables." They can be manipulated and varied at will. Phasing, beaming, frequency, superposition, interference, resonance, and Fourier expansion are the keys to scalar EM engineering.

To make a scalar beam, special modifications are made to the EM wave transmitter so that, effectively, the transmitter transmits multiple transmissions simultaneously, and these vectorially sum to zero. This is the same thing as transmitting multiple phase-locked EM force-field beams simultaneously, in **the** sum-zeroed fashion.

Standard beaming antennas and conventional circuits can readily be utilized with slight modifications.

Any radar can be converted to a scalar EM radar quite simply. Indeed, a modification kit to accomplish that can be developed and held until the **scalar** mode is desired. The modification kit can be quickly installed to provide the scalar capability.

Scalar radars, for example, cannot track "corner reflector" targets. However, they can magnificently track smoothly filleted corners; rounded, smoothly curved surfaces; and curved or flat metalized dielectrics. Against aerial targets using special construction to lower their cross-section to normal radars, the application of the scalar radar mode is immediately obvious.

Scalar waves pass through the electron shells of an atom and interact with the nucleus. They are continually absorbed and emitted by all nuclei in the universe.

BACKGROUND FOR THE BRIEFER.

Any large collection of nuclei - such as a star or a planet - is a strong absorber and radiator of scalar wave radiation.

The Sun is a particularly strong source of scalar radiation. This radiation penetrates the Earth deeply, interacting more and more with the deeper layers which, under greater mechanical stress, are more nonlinear. Most of the heat in the molten core of the Earth comes from the dephasing of a portion of this absorbed scalar radiation from the Sun, liberating ordinary EM energy as heat.

The Earth also re-radiates scalar wave radiation back to the Sun. The Sun and Earth are thus coupled into a "scalar" system in equilibrium or near-equilibrium. Each body in the couplet possesses both a feed-forward and a feedback loop.

The Earth also radiates scalar wave radiation to the Moon, and the Moon reradiates scalar radiation back to the Earth. The Earth and Moon also are coupled into a "scalar" system in equilibrium or near-equilibrium. Again, each body in the couplet possesses both a feed-forward and a feedback loop.

The Sun and Moon are correspondingly coupled into a scalar system in equilibrium or near-equilibrium, each with both a feed-forward and a feedback loop.

The Earth, Sun, and Noon thus form a triad-coupled system of special importance to the stability of our existence here on Earth, and the stability of our biosphere.

And so on with all the other planets in the solar system, and combinations thereof.

The solar system is thus a very sensitive scalar-coupled system, with strong subsystems within it.

ADDED FOR THE BRIEFER:

Actually this scalar coupling of the solar system provides a major check on unrestricted use of the large Soviet strategic scalar EM weapon systems. If significant scalar effects are produced on earth in a "pulse" mode, pulsed disturbance of the earth-sun and earth-moon systems results. Here a danger exists that one or more natural resonances of the coupled system may be excited. If the feedback stimulation of the Sun is not insignificant, for example, large sunspot activity may result sometime thereafter, say in a day or two. If too much or too sharp stimulation occurs on earth, the coupled resonant response from the sun could be disastrous. Indeed, a solar response could be stimulated so that the Sun would violently belch and destroy our biosphere, among other effects. The simplest doomsday stimulation would be for a violent expulsion of Solar EM energy and particles to occur. If this were due to resonance, the expulsion of Solar EM energy and particles would continue during some decay time. In that case, fiery destruction of the earth, strongly indicative of Biblical prophecy, would result. Particularly sensitive are the resonances of the sun-earth, sun-moon, and earth-moon systems.

Note the abnormal influence of the moon on tides - tides of both the tectonic plates and the oceans. One cannot help but point out that, if the earth-moon resonant system were overstimulated, one might expect violent earthquakes of extraordinary magnitude, and tremendous tidal waves hundreds of feet in height. (The potential connection to legends of earth-destroying floods, evidences of extreme water levels in mountains, and legends of ancient cataclysmic destruction of supposedly advanced technological oceanic cultures is obvious.)

Accordingly, use of huge scalar EM weapons is a double-edged sword. Unless carefully employed, use of the weapons could cause a terrible backlash to the user as well as the victim, and even accidentally cause the destruction of the earth itself. It is not accidental that in 1960 Khrushchev stated that his new fantastic weapon could - if unrestrainedly used - destroy all life on earth.

This appalling backlash potential is apparently what prompted Brezhnev in 1975 to make so strong an effort to get agreement to outlaw the development of "new weapons of mass destruction" more frightful than the mind of man had even dreamed of. Gromyko even introduced such a draft agreement into the proceedings of the United Nations. The first article of the draft agreement indicated that the nature of the weapons referred to would be negotiated. Sadly, the entire Western world did not even know what the Russians were talking about.

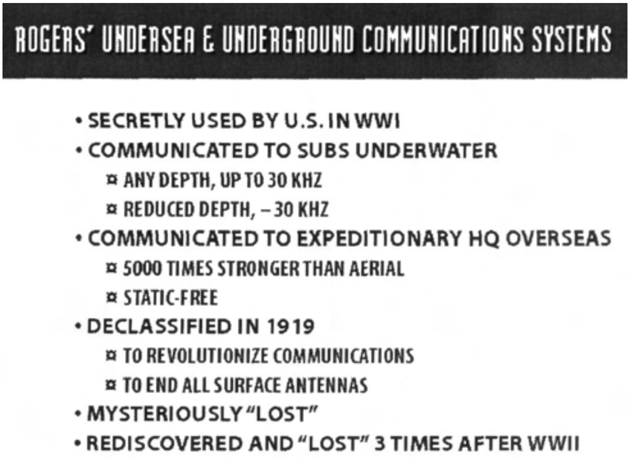

ROGERS' UNDERSEA & UNDERGROUND COMMUNICATIONS SYSTEMS

- SECRETLY USED BY U.S. IN WWI
- COMMUNICATED TO SUBS UNDERWATER
  - ANY DEPTH, UP TO 30 KHZ
  - REDUCED DEPTH, – 30 KHZ
- COMMUNICATED TO EXPEDITIONARY HQ OVERSEAS
  - 5000 TIMES STRONGER THAN AERIAL
  - STATIC-FREE
- DECLASSIFIED IN 1919
  - TO REVOLUTIONIZE COMMUNICATIONS
  - TO END ALL SURFACE ANTENNAS
- MYSTERIOUSLY "LOST"
- REDISCOVERED AND "LOST" 3 TIMES AFTER WWII

Slide 025: Rogers' Undersea & Underground Communications Systems

In an article in the March 1919 issue of the Electrical Experimenter, a real bombshell was released on an unsuspecting world.

In that issue, details of the Rogers'undersea and underground communication systems were released. These systems had been classified SECRET during WWI, and used to communicate with (1) U.S. submarines underwater around the world, and (2) U.S. Expeditionary Forces Headquarters overseas.

Subs at any submerged depth could be reached with transmitters of frequencies up to 30 kiloHertz. At reduced depths, frequencies above 30 kHz could be used.

At the headquarters of U.S. overseas forces, the signals were received through the earth. They were 5,000 times stronger than aerial signals through the atmosphere, and static-free.

The Rogers' systems were covered by a series of patents. They used special antennas implanted in the ground.

While the patents are somewhat confused, several of the diagrams in the Electrical Experimenter reveal that the antennas were producing scalar waves. Scalar waves can indeed travel through the earth and through the ocean. Indeed, if good quality scalar waves are made, then frequencies well above 30 kHz can be sent through the ocean. Multi-megahertz frequencies, for example, are easily achievable. (The problem, of course, is that these waves are pure zero-vector artificial potentials that do not move electrons in the conduction gas of a conventional antenna/receiver combination. Normal receivers and detectors do not see them. Special underwater detectors - such as those developed by Frank Golden - are required.).

With the startling declassification and revelation of the Rogers' systems, it was expected that all communications would be revolutionized, and that all surface antennas would be removed.

Strangely, this entire technology faded away quickly, as if it had never been. Certain powerful interests did not wish the extended electromagnetics to be discovered or taught. That, after all, would lead to free energy and loss of economic control of the citizens. That was precisely why Nikola Tesla had already been suppressed.

According to the important research of Dr. Robert Beck, the same scalar technology has been rediscovered and mysteriously "lost" at least three times after World War II!

Today the orthodox establishment is still suppressing it.

One wonders why no orthodox physicist - with the single exception of Dr. Hooper (see U.S. patents 3,610,971, Oct. 5, 1971 and 3,656,013, Apr. 11, 1972) - seems to have experimented with deliberately constructed zero-vector EM waves and devices, and published his results in the literature.

One would think that any university physics department would wish to explore such a gaping void with at least a few graduate student experiments, if for no other reason than scientific curiosity, particularly when the experiment is easily performed, it is totally different from anything else in the theory or literature, and the implications may be formidable. A true scientist in that related field ought to try it!

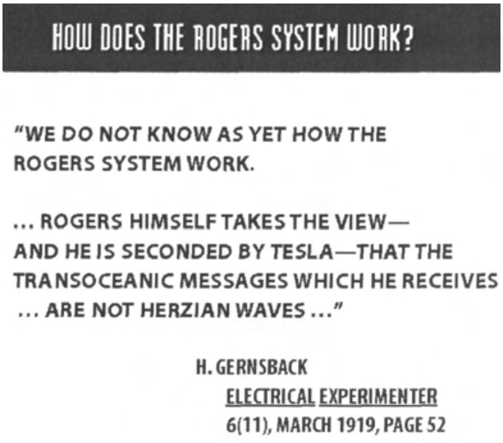

HOW DOES THE ROGERS SYSTEM WORK?

"WE DO NOT KNOW AS YET HOW THE ROGERS SYSTEM WORK.

... ROGERS HIMSELF TAKES THE VIEW— AND HE IS SECONDED BY TESLA—THAT THE TRANSOCEANIC MESSAGES WHICH HE RECEIVES ... ARE NOT HERZIAN WAVES ..."

H. GERNSBACK
ELECTRICAL EXPERIMENTER
6(11), MARCH 1919, PAGE 52

Slide 026: How does the Rogers System Work?

In the same March 1919 issue of Electrical Experimenter, Hugo Gernsback pointed out that no one as yet knew how the Rogers' system actually worked.

Rogers himself believed he was not using conventional Hertzian waves.

Nikola Tesla himself confirmed that the underground and undersea waves were not Hertzian waves.

If so, they could only have been scalar waves.

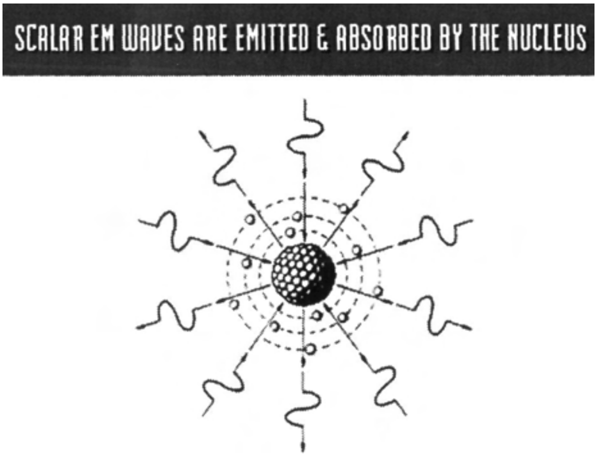

SCALAR EM WAVES ARE EMITTED & ABSORBED BY THE NUCLEUS

Slide 027: Scalar EM Waves are Emitted & Absorbed by the Nucleus

Normally scalar EM waves do not couple to orbital electrons. Entering into the atom, incoming scalar EM waves penetrate the electron shells without interaction. They continue, penetrating into the interior of the closely packed nucleus, where the extreme nonlinearity of the charged virtual particle flux that binds the nucleons together distorts the scalar wave, dephasing its summed EM vector components. This produces a nonzero EM resultant and polarized pattern directly inside the virtual particle charge flux of the nucleus.

The new EM resultant immediately couples the distorted scalar wave to the nucleus, resulting in absorption of the wave and incorporation of its polarization pattern into the nucleus' virtual flux. This in turn results in a slightly excited nucleus. Normally the excited nucleus will promptly decay by emitting another scalar wave, which passes through the orbital electron shells and radiates out into the universe.

However, as can be seen, if the same strong scalar pattern is constantly radiated into the nucleus and absorbed by it, the excitation of the nucleus "charges up," leading to decay by other modes. In this manner the nucleus can be engineered even transmuted - by rather minuscule power of the engineering radiation signal.

This is how living biological systems are able to transmute elements, as discovered by Louis Kervran. (Kervran was nominated for the 1977 Nobel Prize for this work.). The biosystem produces scalar EM waves. If the appropriate signal is employed into and on the appropriate nucleus, a changed element emerges. In this fashion chickens denied sufficient calcium in their diet, but allowed a surplus of potassium, are able to transmute some of the potassium into calcium. Note that one of the isotopes of potassium is an isomer of normal calcium. Transmutation of this potassium isomer to calcium is easiest.

NUCLEI OF THE UNIVERSE CONTINUALLY EXCHANGE SCALAR EM WAVES

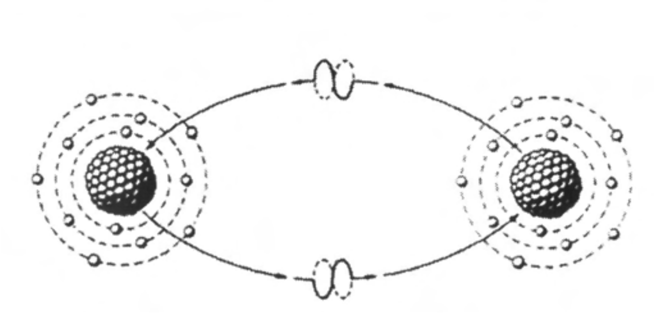

Slide 028: Nuclei of the Universe Continually Exchange Scalar EM Waves

Thus scalar EM waves are continually exchanged by all nuclei of the universe, and the vacuum is - among other things - a seething cauldron of scalar radiation.

In fact, it is this seething scalar wave cauldron that creates the virtual particles and the virtual particle flux of vacuum itself. (Since the scalar EM wave represents an oscillation of the curvature of spacetime, and since everything - virtual or observable - is a curvature of spacetime, then scalar waves can be taken as the universal generatrix.

The intensity of the trapped scalar flux in the nucleus of an atom is responsible for the mass and inertia of the nucleus.

Mass is localized trapped scalar resonance.

The trapping agent is the spin of a particle.

The mass and inertia of the particle are due to the scalar resonance that is trapped. Increasing the amplitude of the trapped scalar resonance increases the mass and inertia; decreasing the amplitude decreases the mass and inertia.

CONVERTING EM FIELD ENERGY TO G-FIELD ENERGY & VICE-VERSA

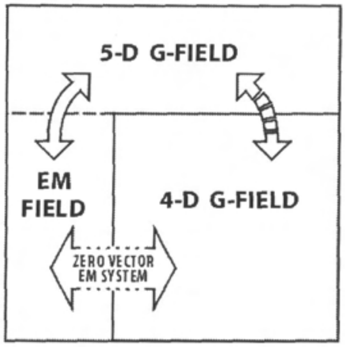

Slide 029: Converting EM Field Energy to G-Field Energy & Vice-Versa

According to Kaluza theory, there are no such things as a separate EM field and a separate gravity field. Instead, in five dimensions there is only one field: the 5-d gravitational field. The EM field is the 5th dimensional aspect, while our **normal** 4-dimensional G-field is the other aspect of the 5-field, in our normal 4-space.

Thus we may say that the 5-field is composed of two components: the 5th dimensional component (our normal EM field) and the component occupying **our** normal 4 dimensions (our normal G-field).

The 5-d "force" field, of course - or the 5-space analogy to a force field - would be due to a gradient or "bleed-off of the 5-d gravitational potential. This gradient in turn is composed of two components: the 5th dimensional "bleed-off" outside our normal 4-space (this outer bleed-off is our normal EM force **field**) and the bleed-off inside our normal 4-space (this inner bleed-off is our normal G-field).

Normally the 5-potential bleeds-off outside our 4-space as EM field, far greater than it bleeds-off inside our world as G-field. Between two electrons, for example, the electric field is about 10^{42} times as strong as the G-field. Since in the Kaluza view there is only one 5-d G-potential that is causing both force fields, this shows that there is an incredibly greater EM bleed-off of the 5-

potential between the two electrons than there is a G-field bleed-off between them.

As shown on the diagram, the 5-d G-field is normally comprised almost entirely of the 5th dimensional EM field. Only a small 4-space G-field component exists.

However, suppose we were to "block" the bleed-off of the 5-potential in the EM mode. Then none of the 5-potential could bleed-off in the 5th dimensional EM field. Instead, it would be forced to bleed-off into the 4-space G-field. In our two-electron example, this "perfect case" would result in the disappearance of the E-field between the two electrons, and the G-field between them would increase to about 10^{42} times its normal strength.

The end result would be that, by blocking the EM force field bleed-off, EM field is converted to G-field. In addition, EM field energy is converted to G-field energy. With the extremely amplified gravitational and inertial effects that result, one can now accomplish direct engineering of gravitation, mass, and inertial effects.

We can effectively accomplish this "blocking the EM bleed-off by opposing EM force fields so that they sum to vector zero. This is the same as summing various 5th dimensional gradients of the 5-potential to a zero vector resultant. In that case, as much "EM bleed-back to 5-potential" occurs as there is "EM bleed-off from 5-potential." This places the 5-potential in equilibrium with respect to EM bleed-off.

Also, another nice thing results: Since we can readily vary the magnitudes of the EM force field components of the vector zero summation, we can actually form phased "zero vector waves" by increasing and decreasing the magnitudes of all the component EM vectors in phase, but retaining their vector summation always equal to zero.

In that case we have produced a very simple 5-space G-potential wave, which concomitantly forces phase-locked variations in the 4-space G-potential. In short, we have produced an electrogravitational (EG) wave, or scalar EM wave for short. This wave described is the simplest EG scalar wave we can produce; much more complex EG waves can be produced and used for highly **specialized** purposes.

The scalar EG wave changes EG potential energy into 4-gravity potential energy in one half-cycle, and changes 4-G potential energy into EG potential energy in the other half-cycle. However, the EG potential energy in the first half-cycle does not react electromagnetically in linear circumstances, since it is electromagnetically a linear vector zero.

So our scalar EG wave actually oscillates energy back and forth between a locked-in 5th dimensional EM potential and a 4-space G-potential. As can be seen, to the linear 4-space observer this scalar EG wave appears as a purely gravitational wave, with G-potential magnitude varying in a wavelike manner.

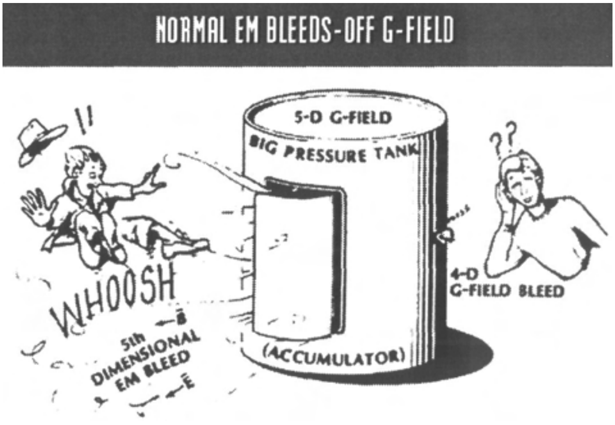
NORMAL EM BLEEDS-OFF G-FIELD

!!
5-D G-FIELD
BIG PRESSURE TANK
??
4-D
G-FIELD BLEED
WHOOSH
5th
DIMENSIONAL
EM BLEED
(ACCUMULATOR)

Slide 030: Normal EM Bleeds-Off G-Field

For easy visualization, we may regard the situation in the analogy shown on the slide.

The 5-potential of 5-space vacuum is like a high pressure in a big tank. The tank has a huge open door in the 5th dimension, leading to rapid bleed-off of the 5-presure through the big door; that is, as EM field. Only a very small hole in the tank normally exists in our own 4-space world, so only a very small amount of the pressure bleeds off as our G-field.

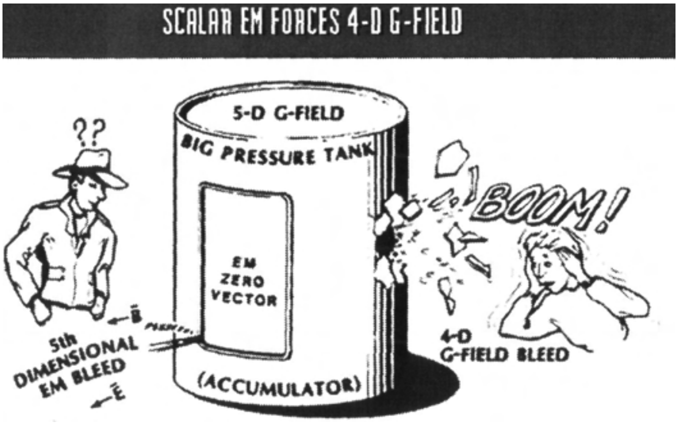

SCALAR EM FORCES 4-D G-FIELD

??
5-D G-FIELD
BIG PRESSURE TANK
EM
ZERO
VECTOR
(ACCUMULATOR)
BOOM!
4-D
G-FIELD BLEED
5th
DIMENSIONAL
EM BLEED
←B̄
←Ē

Slide 031: Scalar EM Forces 4-D G-Field

As can be seen however, if we shut the big door so that we have only a little bleed-off through the seals, then most of the pressure is forced to bleed-off through the small opening, greatly enlarging it. In that case, most of the 5-potential bleeds-off as 4-G force field.

The mass exposed to such increased 4-G bleed-off exhibits a much greater mass and inertia.

In fact, mass itself can act as an accumulator for the effect, and it can increase its ability to store trapped scalar resonance. This increases its inertia.

By the use of scalar EM (blocking the EM bleed-off), we can input normal EM energy and obtain powerful gravitational and inertial effects. Many other powerful effects can also be obtained.

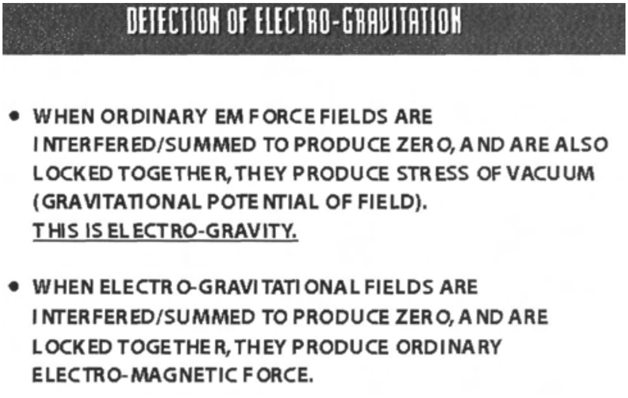

DETECTION OF ELECTRO-GRAVITATION

- WHEN ORDINARY EM FORCE FIELDS ARE INTERFERED/SUMMED TO PRODUCE ZERO, AND ARE ALSO LOCKED TOGETHER, THEY PRODUCE STRESS OF VACUUM (GRAVITATIONAL POTENTIAL OF FIELD). THIS IS ELECTRO-GRAVITY.
- WHEN ELECTRO-GRAVITATIONAL FIELDS ARE INTERFERED/SUMMED TO PRODUCE ZERO, AND ARE LOCKED TOGETHER, THEY PRODUCE ORDINARY ELECTRO-MAGNETIC FORCE.

Slide 032: Detection of Electro-Gravitation

Here arc two magic rules in the scalar electrogravitational business:

(1) When ordinary EM force fields are interfered or summed so that they produce vector zero, and the fields are locked together, they produce stress of vacuum (5-potential). One component of this 5-potential is the 4-space G-potential. Since bleed-off as a gradient of this 4-space G-potential is all that is allowed, then the EM destructive interference/zero summation bleeds-off to produce gravitational force field. This is because opposing equal forces, which sum to a vector zero force field resultant, actually produce a stress potential having stress energy of spacetime. And that is a curvature of spacetime, hence gravitational.

In short, total destructive interference of EM waves or force fields produces excess stress potential energy in spacetime, and therefore gravitational force field on detecting/coupled particles.

(2) When electrogravitational fields are interfered/summed to produce vector zero, this destructive interference/zero summation bleeds-off to produce ordinary electromagnetic force field.

In short, destructive interference of scalar EM waves or fields (electrogravitational waves or fields) produces electromagnetic force fields on detecting/coupled particles.

BACKGROUND FOR THE BRIEFER:

Mathematically we can summarize this as follows;

$$G_{(1-4)} = G_{(1-4)}(\mathrm{m}) + G_{(1-4)}(\mathrm{EM}) \qquad [3]$$

$$G_{(1-5)} = G_5 + G_{(1-4)} \qquad [4]$$

If $$\mathbf{E} = \Sigma \mathbf{\mathring{a}}_i E_i = \mathbf{0},\ E_i \neq \mathbf{0};\ \mathbf{B} = \Sigma \mathbf{\mathring{a}}_i B_i = \mathbf{0},\ B_i \neq \mathbf{0} \qquad [5]$$

then $$G_{(1-4)}(\mathrm{m}) = G_{(1-4)}(\mathrm{EM}) = G_5 = G_{(1-5)} \qquad [6]$$

and $$G_5 = G_{(1-4)} = f[\Sigma(\mathbf{\mathring{a}}_i\, |\mathbf{E}_i|^2)] + g[\Sigma(\mathbf{\mathring{a}}_i\, |\mathbf{B}_i|^2)] \qquad [7]$$

and $$G_{(1-5)} = G_5 = G_{(1-4)}(\mathrm{m}) = G_{(1-4)}(\mathrm{EM}) = f(\phi) + g(\mathbf{A}) \qquad [8]$$

Where:

$G_{(1-5)}$ is the 5-d G potential, or total G-stress on dimensions 1-5.

G_5 is the combined EM potential, or G-stress on dimension 5 only.

$G_{(1-4)}(\mathrm{m})$ is the normal 4-d G-potential of mass alone, or the G-stress due to mass-gravity.

$G_{(1-4)}(\mathrm{EM})$ is any extra 4-d G-potential of mass due to conversion from G_5 stress (normally zero).

E and B are the electric and magnetic force fields.

ϕ is the electrostatic scalar potential.

A is the magnetic vector potential.

Thus the age-old scientific dream of direct engineering of gravitational field — and of spacetime itself — can now become reality. To start, one simply makes an intense EM vector zero and varies its stress intensity and internal pattern of summed components. One also varies the reference potential at which the patterned EM stress zero is created, and the frequency of internal component variation.

BACKGROUND FOR THE BRIEFER.

He that has ears, let him hear.

Here is the secret of antigravity

The "electrical charge" of a charged particle — such as one of the orbital electrons of an atom or one of the protons of its nucleus — represents a difference in flux intensity (potential) between the local particle and its ambient vacuum.

It's the continual bleed-off of gravitational charge as electrical charge.

If that bleed-off is reversed or stopped, a drastic effect on the gravitational potential and charge ensues. The 5-potential and 5-charge have become 4-G-potential and 4-d-charge respectively.

Thus "charging" an object with its scalar EM pattern charges it gravitationally.

Now the only "bleed-off channel" is through the 4-G force field.

Further, in a nucleus the nucleons continually change back and forth between proton and neutron, so the electrical charge is "spread" throughout the nucleus and shared by all the nucleons.

Further, each element (actually each isotope) has its own unique "ensemble pattern" of Fourier expansion scalar frequencies, amplitudes, etc. This pattern, of course, can be reproduced artificially and transmitted by modified EM transmitters (i.e., by scalar EM transmitters). There is, however, a sort of "master key" scalar EM (EG) pattern for nucleons (protons and neutrons which are changing back and forth into each other by exchange of virtual charge currents).

If one "reverses the charge" by inverting this pattern, and then "charges up a mass with the inverted charge," to the external observer the charging mass just gets lighter and lighter, and its inertia gets less and less. Eventually it seems (to him) to acquire negative mass and negative inertia, and just accelerate away from the earth. The object "falls upward" instead of "falling downwards."

There are also some weird time effects. The object can be moving slower through time than the laboratory observer, or even be moving backwards through time with respect to the laboratory observer. (Don't believe everything they laught you in relativity; none of those professors had ever engineered a single general relativistic situation. Little that they teach in GR is based on direct experiment. Most of what they teach is already experimentally proven to be incomplete or in need of further extension.).

But back to our slide.

Inside the pattern itself, it's just a quite normal pattern. No change in or on the object seems to have occurred, to it, if the charge is evenly accomplished throughout the object. Indeed, to an internal observer it's the external environment that seems to have suddenly gotten very strange! This is what occasionally happens to a hapless aircraft or ship that experiences anomalous spatiotemporal difficulty in an *occasionally scalar-active* area such as the Bermuda triangle. Under the right conditions, the vehicle's gravitational charge is affected by anomalous changes in the scalar radiation from the earth in that area. To the passengers and crew, it is the external environment that suddenly seems weird. In addition, on-board EM and inertial instruments may be affected, and other electrogravitational effects may also occur, due to differing rates of charge in different parts of the vehicle.).

"Reversing or lowering the EG-charge" is controlled by means of biasing the ground potential on the ensemble pattern transmitters. These transmitters can even be on-board the vehicle itself. (Remember, in scalar EM to transmit is also to receive. A vehicle can change its own bias potential (with respect to local vacuum G-potential) by properly transmitting, which translates to "receiving potential charge." The charge current can either be negative or positive (its potential can be lowered or increased vis-a-vis that of the vacuum.) Think of this as handling positive and negative energy, which means positive and negative gravity.

By "fiddling" with this, you can float metal. Or a human body. Or a battleship. Or a high-speed vehicle containing a crew.

You can even "dematerialize it" or "teleport it."

The Philadelphia Experiment may have been real after all. If so, the test ship and its personnel were "blasted" into this strange realm, instead of making a gentle, controlled entry.

Suppose you materially lower or reverse (adjust) the gravitational charge (gravitation 5-potential with zero-summed EM to throttle EM bleed-off). At zero gravitational charge, to the external observer, such an assumed vehicle would seem to have no mass and no inertia. It is capable of extreme accelerations, right angle turns at full speed, etc. It is also right on the point of dematerialization, and may appear to be glowing or just a vehicle of light.

If the pilot wishes to "land," of course he must adjust the charge on the vehicle.

If the pilot wishes to "dematerialize" or "teleport," again he must adjust the charge on the vehicle.

If he wishes to go "hyperspatial," again he must adjust the charge on the vehicle and allow the proper bleed-off to produce a hyperspatial force. That's done with nested zero summations and nested patterns of scalar EM transmission.

That is, he can charge and bias the vehicle in multiple, nested zero-summations simultaneously.

That is necessary for hyperspatial control and travel.

Note that in lower hyperspace such a vehicle could do some very strange things. Such as "penetrate" solid matter (actually, it goes "around" the 3-space matter in a 4th Kaluza-Klein space).

We might expect to see such vehicles glowing. Their various surface features and mechanisms might appear to be glowing or revolving lights, etc.

They could exhibit incredible "aerodynamic performance," seemingly in the atmosphere. Actually they would not be moving "through" the atmosphere at all but through a higher space outside each particle of atmosphere.

They could seem to materialize and dematerialize.

They could seem to plunge into the ocean or rise out of it.

They could even seem to operate under the ocean or inside the earth itself.

Such anomalous vehicle performances have been seen all over the world, particularly since a few years after WWII.

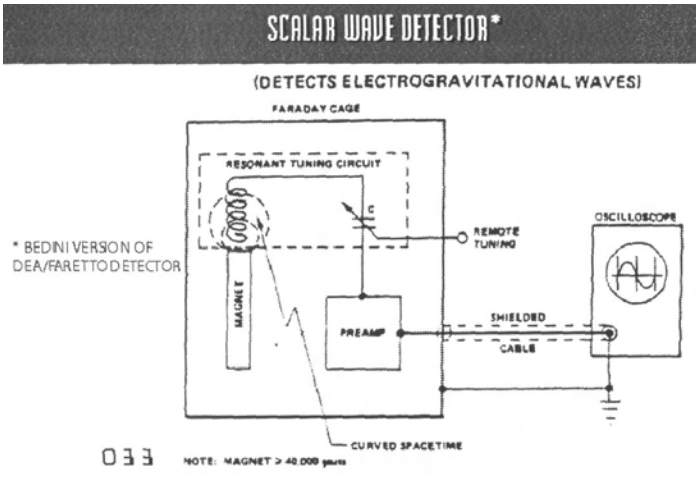
SCALAR WAVE DETECTOR*

(DETECTS ELECTROGRAVITATIONAL WAVES)

* BEDINI VERSION OF DEA/FARETTO DETECTOR

033 NOTE: MAGNET > 40,000 gauss

Slide 033: Scalar Wave Detector

On this slide we show a concept for a detector for scalar EG waves. This is the Bedini scalar wave detector, adapted from an original concept by Dea and Faretto.

The idea is quite simple: install a very powerful bar magnet vertically inside a grounded Faraday cage. Then install an open-ended coil longitudinally above the magnet so that a line through the longitudinal axis of the magnet passes through the longitudinal axis of the coil above it. The open end of the coil does not touch the magnet.

Connect the other end of the coil to a variable tuning capacitor, so that the coil and capacitor form a tunable, series-LC-oscillatory circuit. The output of the capacitor is connected to a transistorized preamplifier inside the cage. A tuning shaft for the capacitor is very carefully passed through a small hole in the shield to allow tuning from outside.

The output of the preamp passes very carefully through a small hole in the shield, through a shielded cable, to an adjacent oscilloscope. The cable shield in also grounded to reference ground potential.

The theory is as follows: Suppose a normal EM wave appears inside the cage, above the magnet or closely adjacent thereto. In that case a coupled **oscillation** appears in the field of the magnet, and this oscillation is coupled to the coil immediately above it. If the oscillation is within the bandwidth of the tuned series LC circuit, detection occurs. This is amplified by the preamp and passed to the oscilloscope, where it is displayed on the scope.

Note that the detector detects normal EM wave. Now our problem is: how do we get an EG wave to be detected? And how do we assure that we do not detect ordinary EM waves from outside?

Actually this is simple. Ordinary waves (except for quite low frequency) will be grounded out by the Faraday shield, and will not penetrate the cage. Thus these normal EM waves cannot enter the cage and appear above the magnet. They will not be detected.

Scalar EG waves, on the other hand, will enter the cage since they do not couple to conduction electrons in the metal of the cage. Above the pole of the magnet, spacetime is locally bent. After all, a pole is a magnetostatic scalar potential, which is a part of the conglomerate called "G-potential." The pole represents an increase (or decrease, depending upon whether it is a north or a south pole) in **the** magnetostatic component of the local G-potential. This is a curvature of spacetime. An EG wave entering this region adds a varying component of magnetostatic G-potential, which bleeds-off in the coil as an ordinary EM wave.

Another way to view the detector is to model the EG wave as a longitudinal wave, and an ordinary EM wave as a transverse wave. When the EG wave enters the curved spacetime region above the magnet, to the coil (the "observer" in this case**)** the longitudinal aspect of the EG wave appears to be rotating back and forth, so that an oscillating transverse component is present. This transverse component appears to the coil as an ordinary EM field, and so the LC circuit detects it if it is in the proper frequency band.

Yet another way to view the situation is to realize that an EG wave entering the magnet results in an oscillation component added to the magnetostatic scalar potential (pole strength) of the magnet. Accordingly, the magnet is a receiver **for scalar** waves, which are detected to "bleed-off as an oscillation of the magnetic field strength of the magnet. Coupling of this oscillating magnetic field **to the** coil creates an oscillating current flow into the capacitor. This oscillates the voltage input to the preamp, which amplifies and feeds the signal to the oscilloscope for display.

In fact rather like the latter view! It is important to remember that a magnetostatic potential and an electrostatic potential can be oscillated by a scalar EG wave. By placing a magnetic material inside a Faraday cage, the oscillation of the magnetostatic scalar potential (pole strength of the magnet) can be used as the EG detecting mechanism. By placing a chargeable material inside a Faraday cage, the oscillation of the electrostatic scalar potential can be detected. (Possible examples of the latter type detectors are given by Hodowanec, Radio Electronics, April 1986.)

Note also that one may detect waves at one reference level and not at another. Changing the bias on the zero-reference ground of the Bedini detector affects the detection. To look inside a normal EM carrier (such as the Woodpecker) and see what scalar signals are riding upon it, the carrier may be used to bias the reference ground of the Bedini detector.

To produce a spectrum analyzer, simply use additional series resonant LC tuning circuits in parallel (put multiple taps on the coil, and wire each tap to a separate tuning capacitor of different capacitance). Again, varying the zero reference level is important, as is varying the strength of the magnet.

Golden also has invented an excellent series of scalar wave detectors based upon quite different proprietary principles.

The bottom line is: we can indeed rigorously detect and measure scalar EG waves. When one considers the large sums of money presently being spent to bury large aluminum cylinders etc. in an attempt to detect gravitational waves, one wonders why some funds could not be released to a few unorthodox researchers by the National Academy of Sciences to fund the proven creation and detection of electrogravitational waves!

Note from TEB: I understand that it took Bedini about 600 attempts before he got the device to work. It was then handed over to a different party to produce, who made some "improvements," and the device then failed to work.

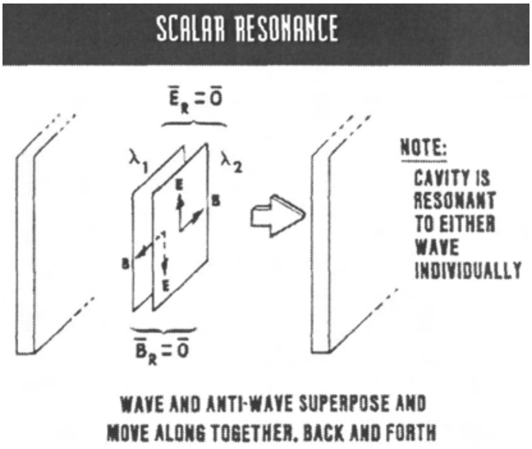

SCALAR RESONANCE

NOTE:
CAVITY IS RESONANT TO EITHER WAVE INDIVIDUALLY

WAVE AND ANTI-WAVE SUPERPOSE AND MOVE ALONG TOGETHER, BACK AND FORTH

Slide 034: Scalar Resonance

Here we explain a new kind of resonance: scalar EH resonance, or electrogravitational resonance.

First, imagine we have a standard resonant cavity, represented by the two walls in our diagram. In this cavity we have a resonant EM wave moving back and forth, represented by the forward-most plane. In accordance with convention, we show the E-field vector and the B-field vector at right angles in this moving wave front. As the wave front moves back and forth, the vectors vary back and forth; however, at any one point between the walls, the two vectors always have the same value. Thus our resonant EM wave forms a standing wave in the cavity.

Now imagine that a second wave front, precisely like the first and of the same frequency, is superposed over the first wave front and travels along with it. This second wave - the "antiwave" - has its force vectors 180 degrees out of phase with the force vectors of the reference wave. Hence the E-fields and B-fields of the two superposed waves always sum to vector zeros, anywhere in the cavity. To an external observer, the cavity contains no ordinary electromagnetic force fields, hence no ordinary EM energy.

However, the energy density of a single EM sine wave in vacuum is given by

$$[E^2+B^2]/8\pi \qquad [9]$$

This energy density is always positive. Hence the energy density of the two waves at any point x between the walls is equal to

$$[E^2+B^2]/4\pi \qquad [10]$$

where

$$E = E(x) \; ; \; B = B(x) \qquad [11]$$

Thus the energy density of vacuum varies with x. But, rigorously, since the resultant E and B fields are zero, this describes a standing gravitational wave. Hence we have a standing EG wave, existing in the resonant cavity. This is an example of scalar resonance. Rigorously the cavity has mass and inertia, to an outside observer, as a result of the two warps in spacetime it contains.

Note that in one half-cycle the energy density of vacuum is greater than ambient, and in the other half-cycle it is less. In the region of one half-cycle, time flows at a faster rate than to the ambient observer, and in the other half-cycle time flows at a slower rate than to the ambient observer.

One half-cycle appears to contain negative electrical charge, and the other appears to contain positive charge.

One half-cycle appears to contain a north pole (positive magnetostatic scalar potential), and the other half-cycle appears to contain a south pole (negative magnetostatic scalar potential).

Perhaps now one can begin to understand why a continuously accelerated orbital electron in the atom does not radiate EM energy, completely in violation of Maxwell's equations. The electron is naught but a complex aspect of a standing scalar resonance, existing between the nucleus and the orbit.

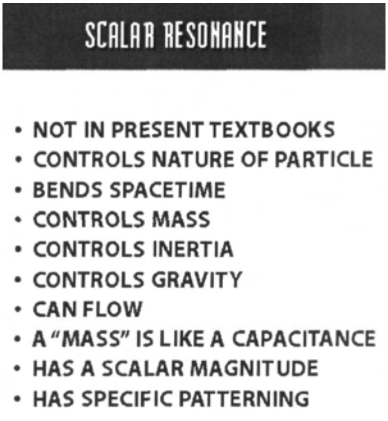

SCALAR RESONANCE

- NOT IN PRESENT TEXTBOOKS
- CONTROLS NATURE OF PARTICLE
- BENDS SPACETIME
- CONTROLS MASS
- CONTROLS INERTIA
- CONTROLS GRAVITY
- CAN FLOW
- A "MASS" IS LIKE A CAPACITANCE
- HAS A SCALAR MAGNITUDE
- HAS SPECIFIC PATTERNING

Slide 035: Scalar Resonance

Scalar resonance is not in the conventional textbook.

Scalar resonance is a particular zero-summed multi-resonance, electromagnetically, so that it does not act in an electromagnetic manner.

A scalar resonance is a standing electrogravitational wave. It can be made **electrically,** but it is not electrical in behavior.

In any scalar resonance, spacetime is curved, and it is the magnitude (and direction; of this spacetime curvature that is oscillating in "standing wave" fashion. In respect to stress of the vacuum medium, one half of a standing sine wave of scalar resonance is tensile; the other half is compressive. However, this is with respect to the local ambient stress of vacuum.

"Mass" of a particle is just a characteristic exhibited by a trapped scalar resonance. Usually this trapping is done by the "spin" of the individual particle.

The concept of "mass" may be compared to the concept of "capacitance." That is, a mass is an accumulator for scalar waves; that is, for scalar resonances. It is continually being "charged" and "discharged" by absorption and emission of scalar waves from and to the ambient vacuum scalar wave flux.

Indeed, the magnitude of a mass may be defined in terms of the absolute value of this "switching" (absorption = switch in; emission = switch out) rate. In a very early paper I defined the mass of an object as proportional to the absolute value of the sum of the input and output switching rates, where each "switch" represents a change of action of h/(4). (Bearden, Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena, Mar. 1973, AD 763210, available through NTIS.) In that scheme, one kilogram of mass equals a switch rate of 17.053 x 10^{50} switches per second. An object is in a Lorentz frame when the switch-in and switch-out rates are equal in a given direction. If, in addition, the in-out rates in all directions are constant, the object is "stationary" with respect to a Lorentz observer. If not constant in all directions, the object is in a curved spacetime with respect to a Lorentz observer. In that case either the object's local spacetime is warped or the object is in a generalized, accelerated frame.

In the normal linear spacetime, the "charging" and "discharging" are equal in all directions; hence the "mass" exhibited in any direction is the same.

With respect to an "external observer's" equilibrium flux, a moving object encounters an increased amount of flux rate along its line of motion, just as a moving object in a rainstorm encounters more raindrops per second in the direction of its path, than does a stationary object. Encountering increased scalar wave flux (with respect to the external observer) forces the moving object to absorb and emit scalar waves at a higher rate along its direction of motion than when it is not moving. Thus to the external observer, the mass of the object has increased, insofar as any disturbing force along its line of motion is concerned.

At right angles to its line of motion, however, the flux rate is precisely the same as when the object is at rest. Therefore the "mass" of the moving object with respect to any disturbing force at right angles to its line of motion has not changed, as seen by the external Lorentz observer.

Thus is explained both parts of one of the alltime great mysteries of special relativity: (1) how the mass of an object increases with respect to its motion, and (2) why the mass only increases with respect to its line of motion, and not at right angles to it.

The inertia of a particle is due to its mass, i.e., to the total magnitude of its trapped scalar resonance.

The gravitational attraction between two masses is due to their spacetime curvature. Further, mass is like a capacitance. It can trap additional scalar waves as trapped resonance; hence increase its magnitude or "inertial charge." Or, it can discharge more scalar waves than it absorbs; hence decrease its magnitude or "inertial charge."

Those scalar waves emitted from resonance are emitted as a pattern ensemble of the resonance. Hence they may be regarded as constituting a current of scalar resonance leaving the mass-accumulator-object. Scalar waves absorbed into trapped resonance may be regarded as constituting a current of scalar resonance into the mass-accumulator-object. Thus it is proper to speak of scalar resonance as being able to "flow."

Further, it is possible to increase the mass of an object directly, by transmitting scalar EM waves to it so that it absorbs them. (By "absorbing" scalar waves one means that more enter the object than leave it, so the object acts as an inertial accumulator charging up with inertial charge. This is done by insuring that the reference potential of the scalar wave transmitter is higher than the reference potential of the irradiated object.)

It is also possible to decrease the mass of an object directly, by transmitting scalar EM waves to it so that it emits more than it absorbs. (By "emitting" scalar waves one means that more leave the object than enter it, so the object acts as if it were an inertial accumulator that is discharging its inertial charge. This is done by insuring that the reference potential of the scalar wave transmitter is lower than the reference potential of the irradiated object.)

As can be seen, the scalar wave "transmitter" is actually somewhat comparable to a heat pump; it can either act as an "energy transmitter" or as an "energy extractor," depending upon the difference in potential between "transmitter" and "receiver."

Finally, scalar resonance can have a specific pattern: both in frequency and in spatial curvature aspects, as well as "rate of flow of time" aspects. Indeed, scalar-wise every object has its individual "scalar pattern" which is a unique fingerprint. Since that print is spatiotemporal, it is a product of the object's entire past history. Thus - scalar-wise - no two objects are identical.

This brings up another rather amazing potential: If a reasonably precise scalar pattern of an object can be "irradiated" and resonated with scalar waves, energy may be created in or extracted from the distant object, just as stimulating one tunning fork can excite another at a distance by sympathetic resonance. We leave it as an exercise for the reader to ascertain the relevance of this statement to clairvoyance, radionics, remote viewing, etc.

For the skeptic, however, we must point out that - rigorously - quantum mechanics requires that, continually, in any localized region of spacetime, the "pattern" of any object in the universe appears momentarily in the virtual state, purely from statistical considerations alone. So long as the situation remains statistical, any place in the universe can continue to have the "ghosts" of everything present there, in an ethereally thin pattern, and the observable world will not be affected by it. However, if one can discriminate and scalarly "charge" or "discharge" individual patterns in this "ghost-realm," action-at-a-distance is directly possible, as is materialization and dematerialization. If one accepts that even thought itself produces such virtual "ghost-patterns" in the virtual particle flux of vacuum, then it is at least theoretically possible to even materialize thoughts and thought images.

We have entered a new kind of reality where the old rules and the old limitations do not necessarily apply.

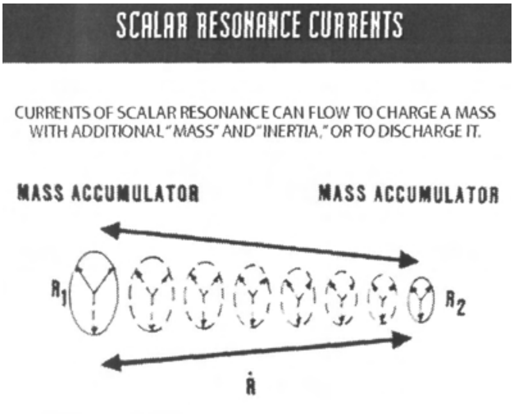

SCALAR RESONANCE CURRENTS

CURRENTS OF SCALAR RESONANCE CAN FLOW TO CHARGE A MASS WITH ADDITIONAL "MASS" AND "INERTIA," OR TO DISCHARGE IT.

Slide 036: Scalar Resonance Currents

As we pointed out, we can often greatly simplify matters by considering currents of scalar resonance. These currents flow from higher potential to lower potential, regardless of whether we are considering "transmission" or "reception."

Indeed, to "transmit at lower potential" is to receive, and to "receive at higher potential" is to transmit.

Thus the "transmitter-receiver" is a special system where simply biasing the two nodes differently determines which way the scalar resonance current will flow.

We may increase or decrease an object's inertia and mass, simply by properly biasing the transmitter-receiver system's two nodes.

In one accidental experiment of some hours in duration, Golden charged up an area so that, locally, all clocks became erratic. This included electrical clocks, battery-driven watches, wind-up kitchen clocks, and a pendulum-operated grandfather's clock. The rate of flow of time itself was apparently altered in the local area by the accumulated charge, which took some four days to drain off and discharge. At the end of the four days of discharge, all clocks and watches had returned to normal.

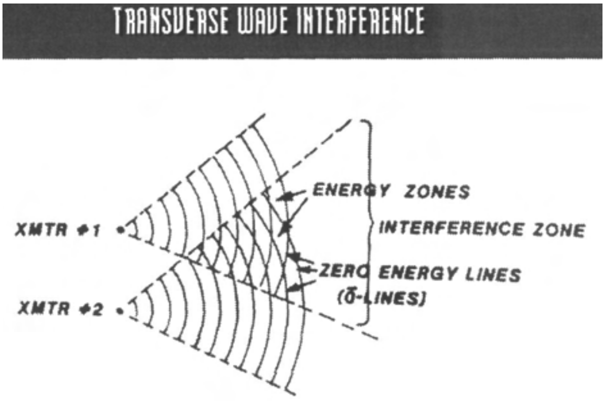

TRANSVERSE WAVE INTERFERENCE

Slide 037: Transverse Wave Interference

On this slide we show the orthodox representation of EM wave interference.

The chart is stylized since we assume two single-frequency EM waves that intersect and interfere in the indicated region.

In the interference zone, in-phase wave amplitudes add (constructive interference) while out-of-phase wave amplitudes subtract (destructive interference). In our stylized mode, we show the zero-summed locations as lines.

Conventionally, these interfering waves are considered to be transverse waves in vacuum.

The point is, in this scheme we think of "putting in the waves of finite amplitude, and creating the zero-amplitude regions by destructive interference." We create additional waves of increased-amplitude regions by constructive iterference.

Note that this view is only half correct. It addresses the electromagnetics, but not the gravitational aspects. Actually the regions of destructive EM interference are regions of constructive G-interference, and the regions of constructive EM interference are regions of destructive G-interference.

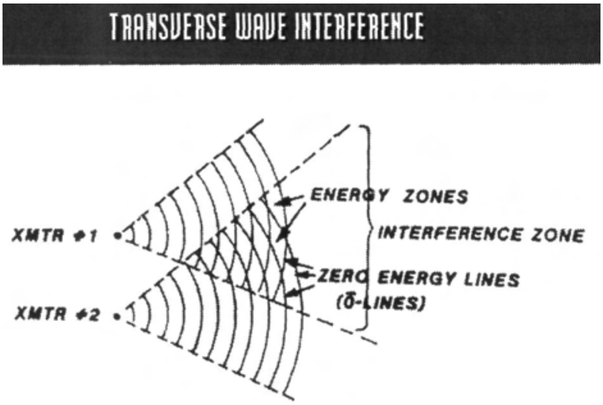

Slide 038: Transverse Wave Interference

Here we show the same interference situation, only now we think of transmitting scalar EM waves, which we model as longitudinal waves.

Again we assume two single-frequency scalar EM waves that intersect and interfere in the indicated region.

However, to an external observer, each of these waves individually appears to have zero E and H fields, hence to contain zero EM energy. That is, conventionally we seem to be transmitting waves of "pure potential," without any "force field" amplitudes whatsoever. We say we are transmitting waves of "artificial potential," since we deliberately constructed the zero-vector-summation scalar waves in the first place.

In the interference zone, a strange thing now happens. The relative magnitudes of the artificial potential of one wave with respect to the artificial potential of the other depend upon the phase relationship between the two. Further, whether the "gradient" so established appears negative or positive also depends upon the phase relationship between the two.

What results is most interesting: the out-of-phase condition now produces an EM gradient whose sign and magnitude are functions of the location within the interference zone.

Thus one now has real, non-zero EM potential gradients, both electrical and magnetic. And so real EM energy has emerged in an interference pattern within the area.

Ironically, we get essentially what we had before! Only now we have "put in the zero-reference lines, and produced the non-zero gradient (energy) zones."

In other words, we have recreated ordinary EM energy at a distance, directly from the interference of gravitational potential energy (anenergy).

And, to the laboratory observer, there has been no "EM wave" energy flow through the intervening space.

This was Tesla's secret of "wireless transmission of energy to a distance without losses."

Further, in a hypothetically perfect case, the interference represents a sort of "energy bottle."

But now recall our need to take the sign of the scalar current into account. In this energy bottle, then, EM energy emerges and stabilizes - if and only if our transmitters are at higher reference potential than the ambient interference zone. If our transmitters are biased at lower potential, then energy is extracted from the energy bottle and re-emerges back at the transmitters, where it must be extracted and disposed of, if the transmitters are not to be burned out.

BACKGROUND FOR THE BRIEFER:

If we interfere two transmitters in a region of great energy -such as the molten core of the earth itself- and bias the transmitters negatively, we can extract enormous energy directly from that distant source!

Indeed, a slight variation of this effect is precisely how the Soviet Union obtains the enormous energy with which to power its incredibly powerful strategic "continent-buster" scalar EM weapons. Scalar frequency pairs - 12 kiloHertz apart - are transmitted into the earth itself on some portion of the earth's "giant scalar resonance" frequency curve. Biasing the two transmitters differently with respect to each other produces a high potential difference (voltage) between them. At the same time, biasing them both negative with respect to the earth's molten core produces a scalar resonance current from the core to the two transmitters. In the "outer loop" between the two scalar transmitters, there appears a gigantic voltage and gigantic electrical current that can be captured by appropriate means. This electrical power is tapped off and transmitted to incredibly powerful scalar EM weapon systems. There special accumulators and switchers are are utilized to provide the power of many Niagaras to each weapon.

Just before May 1, 1985 the Soviet Union performed a "full-up" operational test of their entire strategic scalar EM weapons complex. This test was detected and monitored periodically by Frank Golden. Some 27 of these giant Soviet "power taps" were locked into the molten core of the earth, producing forced (entrained) scalar resonance of the entire planet on 54 controlled frequencies under our very feet. The remainder of the scalar frequency spectrum was ablaze with literally hundreds of Soviet scalar transmitters: probably the entire strategic scalar command and control system to underwater submarines, higher command centers, distant commands, etc. was activated in the giant exercise.

For several days, the system was exercised on a mind-boggling scale, apparently as part of the Soviet Union's highly accentuated 40th anniversary celebration of the end of World War II.

Ironically, not a single U.S. intelligence agency, laboratory, or scientist detected this monstrous exercise for the new Soviet leader, Gorbachev. Not a one of them had a detector for scalar EM radiation, and not a one of them officially believes the exercise ever happened.

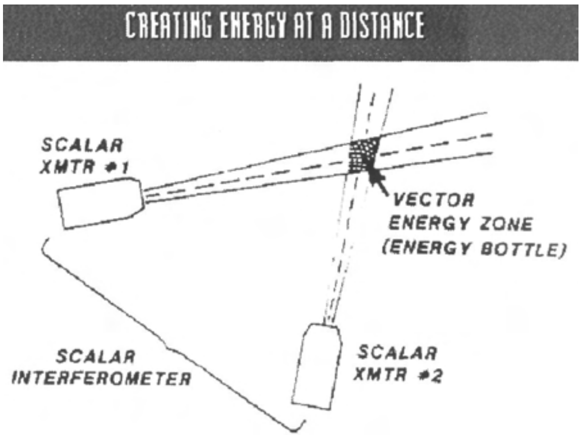
CREATING ENERGY AT A DISTANCE
SCALAR XMTR #1
VECTOR ENERGY ZONE (ENERGY BOTTLE)
SCALAR INTERFEROMETER
SCALAR XMTR #2

Slide 039: Creating Energy at a Distance

A most useful device is obtained if one uses a scalar interferometer where the two transmitters transmit beams that intersect at a distance.

In the interference zone, an energy bottle is created.

By biasing the transmitter reference potentials well above that of the distant energy bottle, EM energy emerges in that zone. In that case the interferometer is operating in the exothermic mode.

By biasing the transmitter reference potentials well below that of the distant energy bottle, EM energy is extracted from the distant zone and emerges from the transmitter. In that case the interferometer is operating in the endothermic mode.

If the transmitters transmit continuously, the effect in the distant zone is continuous.

If each transmitter transmits a pulse, and the two pulses meet in the distant intersection zone, then an explosive emergence or extraction of energy occurs at the distant interference zone, depending upon whether the interferometer is operating in the exothermic or the endothermic mode.

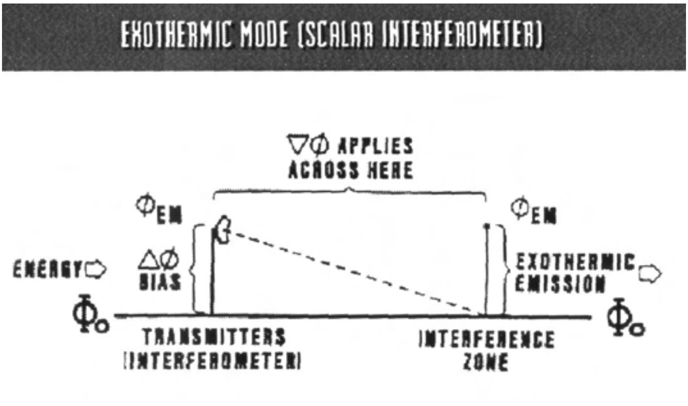
EXOTHERMIC MODE (SCALAR INTERFEROMETER)

Slide 040: Exothermic Mode (Scalar Interferometer)

For a scalar interferometer, "del phi" (the gradient of the potential) applies between the transmitter site and the distant intersection site, not to the intervening space in between. Thus one speaks of "energy flow" as being between transmitter and intersection, without any "in between." In the intervening space, all the energy exists as locked-in artificial potential, not EM force field energy (gradient bleed-off of potential).

In the exothermic mode, the ground potential of the transmitter is biased well above the ground potential of ambient vacuum. Energy enters the transmitter and "disappears", to "reappear" in the distant interference zone.

If continuous wave transmission is used, the energy continuously appears in the distant zone.

If pulse transmission is used and timed so that the two scalar pulses meet in the distant zone, energy explosively appears there.

If multiple frequencies are transmitted in the manner of a Fourier expansion of a particular geometric form, then a 3-dimensional energy form appears at the distant intersection. In this manner, a spherical or hemispherical shell (a globe of "dome" of EM energy) can be created at a distance. If pulse transmission is used, this will be an impulsive or explosive emergence of that energy form. If continuous transmission is used, this will be a continuous glowing form.

By feeding incredibly powerful transmitters with large amounts of energy extracted from the earth's molten core by "energy taps," very large spherical globes and hemispherical domes of EM energy can be created. The energy in the interference "shell" of such a sphere or hemisphere is sufficiently dense to lift Dirac matter from the Dirac sea of the vacuum. Thus the shell contains a glowing plasma.

Many such scalar howitzer signatures have been seen over the oceans, particularly over the North Pacific by jet airliners flying into and out of Japan. Also, such globes and domes of light have been seen deep within the Soviet Union, by observers in Afghanistan and by pilots of aircraft landing in Iran.

In the impulsive exothermic mode, a large visible flash will occur if EM energy in the visible spectrum is produced. "Flashes" in the infrared may similarly occur, if the emerging energy is in that spectral band. In September 1979 our Vela satellites detected such a large "nuclear flash" over the South Atlantic, off the southern coast of Africa. A second "nuclear flash" detected by the Vela satellites in 1980 was in the infrared only. Some years ago, a series of Momalous "flashes" and "booms" occurred off the east coast of the United States. These were due to the orientation and alignment -and registration - of one or more Soviet "scalar howitzers" - scalar EM interferometers in the exothermic mode.

Large "booms" occurred over the shuttle launch site at Cape Canaveral, associated with three shuttle launches prior to the end of November 1985. These were tests of such exothermic scalar EM howitzers, being tested as part of the Soviet Launch Phase ABM system. In that mode, the scalar pulses from the two transmitters are injected into special zero-reference scalar EM "channels" established in the EM carriers of the Woodpecker over-the-horizon radars. The scalar pulses travel through these channels, traveling in the ordinary EM beam of the transmitter and following it in its curvature around the earth in the earth-ionospheric waveguide. The scalar pulses are timed, phased, and oriented so as to meet the rising shuttle. In these tests the transmissions were offset in time so the shuttle would not actually be destroyed, but the system could be tested against actual shuttle launches in the U.S.

The specific uses of these various exothermic modes will be covered in later parts of the briefing.

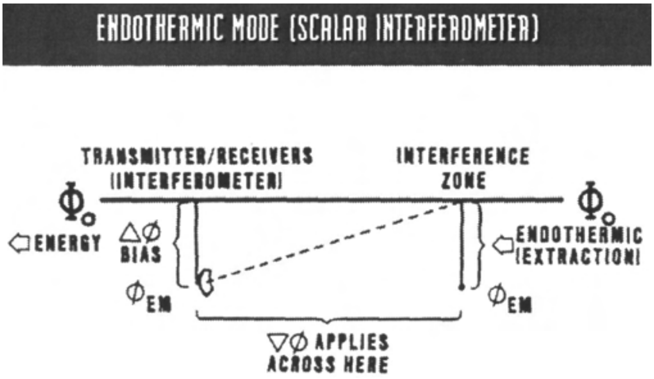
ENDOTHERMIC MODE (SCALAR INTERFEROMETER)
TRANSMITTER/RECEIVERS (INTERFEROMETER)
INTERFERENCE ZONE
Φ_o
Φ_o
ENERGY
$\Delta\phi$ BIAS
ENDOTHERMIC (EXTRACTION)
ϕ_{EM}
ϕ_{EM}
$\nabla\phi$ APPLIES ACROSS HERE

Slide 041: Endothermic Mode (Scalar Interferometer)

In the endothermic mode, the bias on the transmitters is such that energy is extracted from the distant intersection zone, to re-emerge at the transmitters. There the energy must be caught and disposed of if it is not to burn out the transmitters. When very large amounts of energy are extracted from distant points, the Soviets often use a "dump" method of temporarily capturing, and storing in an accumulator, the heat energy emerging at the transmitters. A "scalar energy tap" established into the accumulator then feeds a second howitzer in the exothermic mode. This howitzer is focused on a distant dumpsite - often Bennett Island, where U.S. weather satellites continually observe exhausts.

In the impulsive endothermic mode, energy is impulsively extracted from the distant intersection zone. In the atmosphere this results in a very sharp cooling, or a "cold explosion." It also results in the type of boom and rumble often associated with thunder after a lightning bolt.

On April 9, 1984 the Soviet Union tested such a "cold explosion" off the coast of Japan, near the Kuril Islands. In the suddenly induced low pressure "cold zone" above the ocean, ocean water was sharply sucked up from the ocean, forming a dense cloud. Then air rushing in to the low pressure zone forced the cloud upward, forming a mushroom much like that of an atomic explosion in visual appearance, except for the absence of a flash and the absence of a giant shock wave moving out away from the site of the explosion. (In this case the

Soviets controlled the rise time of the pulse so that no "giant atmospheric boom" occurred.).

As the cloud rises, it expands by mixing - very similar to a giant thunderhead buildup, except much faster. In this case the cloud rose to 60,000 feet in about 2 minutes, spreading out until it reached a diameter of about 200 miles. The pilots and crews of several Boeing 747 jet airliners in the general vicinity saw the incident. Sometime after that, a U.S. weather satellite photo of the area occurred. The cloud in that photo has an anomalous density distribution, differing from that of all other clouds in the area.

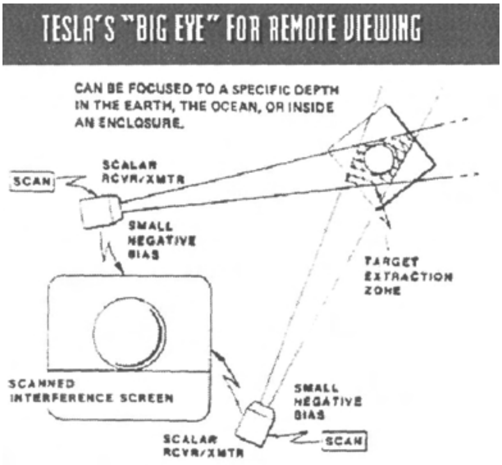
TESLA'S "BIG EYE" FOR REMOTE VIEWING
CAN BE FOCUSED TO A SPECIFIC DEPTH IN THE EARTH, THE OCEAN, OR INSIDE AN ENCLOSURE.
SCAN
SCALAR RCVR/XMTR
SMALL NEGATIVE BIAS
TARGET EXTRACTION ZONE
SCANNED INTERFERENCE SCREEN
SMALL NEGATIVE BIAS
SCALAR RCVR/XMTR
SCAN

Slide 042: Tesla's "Big Eye" for Remote Viewing

Nikola Tesla also planned a very special use for his endothermic scalar interferometer. He planned to produce what he called his "big eye to see at a distance."

This is accomplished as shown on the diagram. First, the system is only weakly endothermic, so that only a small amount of energy is extracted from the **distant** target. Also, the beams are "scanned" by an open receiver, timewise, from side to side and top to bottom. By scanning yet another single beam through the intersection zone and phasing its pulses, an even better representation can be obtained. Thus the receiver produces a representation of the energy extracted from various locations within the distant endothermic zone. By displaying the received signals on an appropriately scanned screen, a representation of the distant scene in the intersection zone can be created. Indeed this is a special kind of "microwave interferometry," and - with modern techniques - the imagery obtained might be surprisingly good. With development, it might even become as good as the image presently obtained by side-looking acquisition radars.

Interestingly enough, since scalar EM beams will easily penetrate the earth or the ocean, one can also look beneath the earth or beneath the ocean with this type of scanning scalar interferometer.

The importance of this capability to strategic and tactical reconnaissance is obvious. Camouflage, cover, and concealment have no effect on such a **system.** One can easily look inside buildings and into underground facilities. With a small system such as this, the U.S. Marines at Khe Sahn would have had **little** difficulty locating the tunnels continually dug under the perimeter by the Viet Cong. And targets under jungle canopies are directly visible.

BACKGROUND FOR THE BRIEFER:

It requires little imagination to see that this system is easily adapted to produce an "underwater radar." With such devices, the problem posed by the underwater nuclear submarine is solved. For example, an entire area can be continually searched, much like acquisition radar systems do now. A submarine can be detected and tracked, and none of its ordinary detectors will detect anything out of the ordinary. By using a separate pair of beams in the exothermic mode, powerful scalar pulses can be fired at the distant sub, intersecting at the submarine in a violent EMP throughout the sub and its armament. Thus the sub and all its missiles are destroyed instantly.

Or, continuous exothermic transmission can be used by the targeting weapon at lesser power, gradually interfering with the sub's electrical systems and causing it to lose control. The sub then sinks to crush depth and implodes.

Precisely that scenario seems to be what happened on April 10, 1963 to the U.S.S. Thresher nuclear submarine. It left a signature: the sub's surface companion, the nearby U.S.S. Skylark, was in the "splatter zone" of the underwater scalar interference. That is, spurious EM noise was being generated In all the Skylark's electrical systems, some of which were actually disabled. So intense was the "electronic jamming" that it required over an hour and a half for the Skylark to transmit an emergency message back to its headquarters that the Thresher was in serious trouble and contact with it had been lost. Some of the Skylark's communication systems actually failed, but later resumed operation inexplicably, once the jamming was gone. That type of "jamming" of multiple bands and multiple electronic equipments, of course, together with the anomalous failure of electronic equipment and its later mysterious recovery, were direct signatures of the use of the exothermic scalar interferometer against the undersea target area in the vicinity of the Skylark.

The very next day, Apr. 11, 1963, the same Soviet scalar EM howitzer system was tested in the "destroy submarine" pulse mode. A huge underwater EM blast occured off the coast of Puerto Rico, about 100 miles north of the island. The underwater explosion caused a huge boiling of the surface of the ocean, followed by the rising up of a giant mushroom of water about a third of a mile high. The mushroom of water then fell back into the ocean, completing the signuture.

Fortunately the startled crew of a passing U.S. jetliner that was just passing its checkpoint in that area saw the entire incident. (See Robert J. Durant, "An Underwater Explosion - or What?", Pursuit, 5(2), April 1972, p. 30-31.)

These two incidents were full-up operational tests of Khrushchev's newly deployed superweapons. He probably staged this dramatic one-two punch in a desperate effort to recover face with the Communist Party after his disastrous facedown by Kennedy in the Cuban Missile Crisis a few short months previously. Apparently the attempt was successful, since he remained in power another year before being deposed.

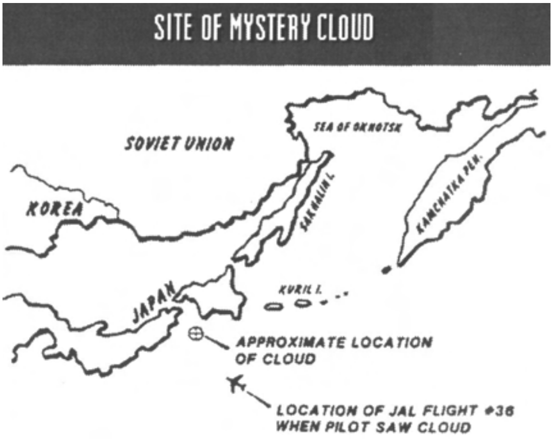
SITE OF MYSTERY CLOUD
SOVIET UNION
SEA OF OKHOTSK
KOREA
SAKHALIN I.
KAMCHATKA PEN.
JAPAN
KURIL I.
APPROXIMATE LOCATION OF CLOUD
LOCATION OF JAL FLIGHT #36 WHEN PILOT SAW CLOUD

Slide 043: Site of Mystery Cloud

On this slide we show the location of the "cold explosion" off the coast of Japan on April 9, 1984. The site of the explosion was only about 200 miles from downtown Tokyo. The crews of several jet airliners, including Japan Air Lines Flight 36, saw it.

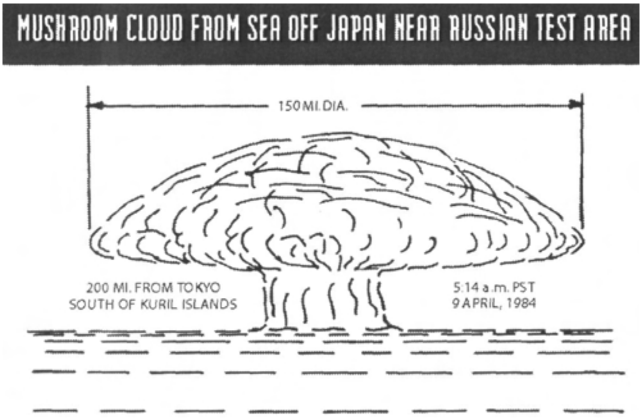

Slide 044: Mushroom Cloud from Sea off Japan near Russian Test Area

A short time before, the Soviets had hastily announced missile tests would be conducted into a zone some distance away from the cold explosion site. This may have been intended as deception to cover this incident, should any foreign nation indicate they understood what was going on.

It also was a direct "stimulus" to the Japanese and the rest of the world: That is, stimulate the system and see if the scientists recognize what happened. If they do, then they know about scalar EM weapons. If they don't, then their countries know nothing of scalar EM weapons, and those countries are defenseless against them.

At any rate, at about 5:14 a.m. Pacific Standard Time on April 9, 1984, a **giant** mushroom cloud erupted above the ocean south of the Kuril Islands. The mushroom rose rapidly, expanding as it rose. In two minutes it had reached **an** altitude of 60,000 feet and a diameter of about 200 miles.

One pilot of a Boeing 747 who observed the cloud was a former B-52 bomber pilot. He described the cloud as very similar to a mushroom cloud from a gigantic nuclear explosion, except there was no flash of visible light. Taking evasive action, he turned off course away from the blast, and braced for a blast shock wave, which never came.

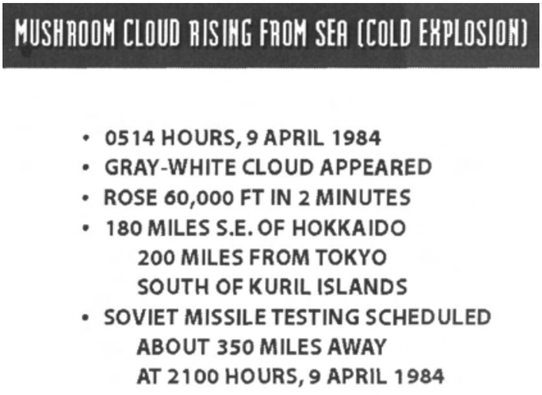

MUSHROOM CLOUD RISING FROM SEA (COLD EXPLOSION)

- 0514 HOURS, 9 APRIL 1984
- GRAY-WHITE CLOUD APPEARED
- ROSE 60,000 FT IN 2 MINUTES
- 180 MILES S.E. OF HOKKAIDO
  200 MILES FROM TOKYO
  SOUTH OF KURIL ISLANDS
- SOVIET MISSILE TESTING SCHEDULED
  ABOUT 350 MILES AWAY
  AT 2100 HOURS, 9 APRIL 1984

Slide 045: Mushroom Cloud Rising from Sea

Indeed, this highly anomalous cloud was due to a cold explosion - a test of a Soviet scalar EM howitzer in the pulsed endothermic mode.

In the weapon's targeted interference zone, EM energy (heat) was rapidly withdrawn when the biased endothermic extraction occurred. Above the ocean an intense low-pressure zone was created. Water from the ocean was sucked up, as if by a giant vacuum cleaner. The water was violently broken into droplets, forming a dense cloud, with a strong upward momentum.

The surrounding higher-pressure air rushed into the low-pressure zone, pushing the rising dense cloud of water vapor upward as its upward momentum continued. Mixing caused rapid expansion of the rising cloud, causing its density to thin measurably as it rose and expanded. An effect very similar to the formation of a strong thunderhead resulted, except much faster.

The mixing cloud also resembled the mushroom cloud from a giant nuclear explosion, just as an emerging thunderhead does.

But no outward blast wave resulted as does from a nuclear explosion. Also, no visible flash occurred, probably as a result of deliberate design by the Russians.

Thus a beautiful "stimulus" and test was conducted: quite a few airliners were "pinged," providing a high-level stimulus to competent observers. The news was sure to be reported by them to several nations. The press was sure to pick up the story. This was a high order stimulus to see whether the U.S., Japan, or other potential adversaries of the Soviet Union recognized the testing of a cold explosion weapon.

By our reaction, we assured them (with high confidence) that we still knew nothing of scalar EM interferometry or cold explosions.

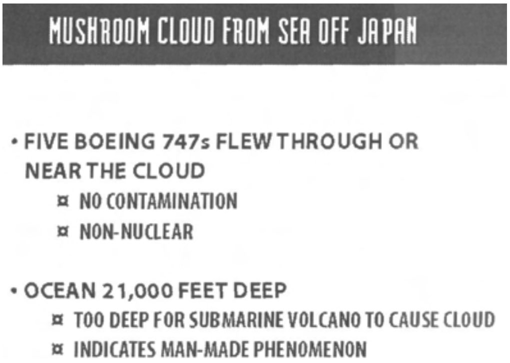

MUSHROOM CLOUD FROM SEA OFF JAPAN

- FIVE BOEING 747s FLEW THROUGH OR NEAR THE CLOUD
  - ¤ NO CONTAMINATION
  - ¤ NON-NUCLEAR
- OCEAN 21,000 FEET DEEP
  - ¤ TOO DEEP FOR SUBMARINE VOLCANO TO CAUSE CLOUD
  - ¤ INDICATES MAN-MADE PHENOMENON

Slide 046: Mushroom Cloud from Sea off Japan

Five Boeing 747's flew through or near the cloud on their way to Anchorage, Alaska. The aircraft were checked at Anchorage for radioactive contamination, but nothing was found. This showed that the huge "explosion" had been non-nuclear.

The ocean in the area is also about 21,000 feet deep. That is really too deep for a submarine volcano to have erupted and caused the cloud. (Besides, a volcano would have continued to snort at least a bit, and so that appears to be out of the question as a possible cause of the incident.)

In short, a manmade phenomenon is indicated.

Indeed, we know it was a cold explosion. As I've stated previously, a colleague has in fact produced the basic effect over a dish of water in the laboratory, using a small scalar interferometer in the endothermic mode.

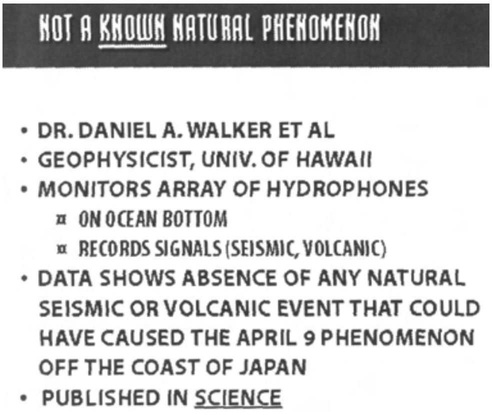

NOT A KNOWN NATURAL PHENOMENON

- DR. DANIEL A. WALKER ET AL
- GEOPHYSICIST, UNIV. OF HAWAII
- MONITORS ARRAY OF HYDROPHONES
  - ON OCEAN BOTTOM
  - RECORDS SIGNALS (SEISMIC, VOLCANIC)
- DATA SHOWS ABSENCE OF ANY NATURAL SEISMIC OR VOLCANIC EVENT THAT COULD HAVE CAUSED THE APRIL 9 PHENOMENON OFF THE COAST OF JAPAN
- PUBLISHED IN SCIENCE

Slide 047: Not a Known Natural Phenomenon

Further, the story does not end there.

A geophysicist at the University of Hawaii, Dr. Daniel A. Walker, together with his colleagues monitor an array of hydrophones on the ocean bottom, recording signals such as seismic incidents, volcano eruptions, etc.

Dr. Walker and his colleagues performed a comprehensive analysis of the data recorded by the hydrophones in and around the time of the incident in question. The data shows the absence of any natural seismic or volcanic event that could have caused the April 9th phenomenon off the coast of Japan.

He and his colleagues concluded that it was either an as yet unknown natural phenomenon, or a man-made phenomenon.

Their report is published in Science. (Walker et al, "Kaitoku Seamount and the Mystery Cloud of 9 April 1984," Science, 227(4687), Feb. 8, 1985, p. 607-611.)

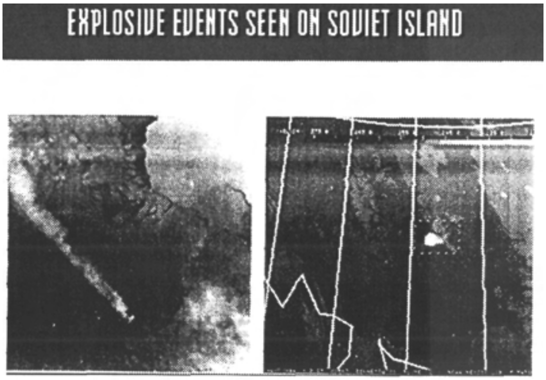

EXPLOSIVE EVENTS SEEN ON SOVIET ISLAND

Slide 048: Explosive Events seen on Soviet Island

Here are two U.S. weather satellite photos of the anomalous exhausts on Bennett Island we previously referred to.

The exhaust in the left picture is about 150 miles long, and is nearly horizontal (about 1.5 degrees above the horizontal). It is entirely consistent with the continuous exhaust from a "dumping transfer" scalar EM howitzer in the continuous exothermic mode. (The primary howitzer, of course, was activated in the endothermic mode.)

I think any open-minded investigator must admit that that is indeed a very powerful jet-like exhaust, and not like anything one ever sees from a natural volcano or geological gas vent. If anyone knows of such an anomalous volcanic or geological exhaust, I would be most delighted to see a photo or report. To my knowledge, no such natural photo exists, and no such natural exhaust has ever **been** reported.

The right picture shows the "puff of an explosive emergence of the exhaust. In other words, this one is the exhaust from a "dumping transfer" howitzer used in the pulsed exothermic mode. Since it did not have so much energy to dump, it could dump it in a pulse. Again, the primary howitzer, of course, was activated in **t**he endothermic mode.

Over 100 weather satellite photos of these anomalous exhausts have been taken since 19**74.**

One or more U.S. aircraft have even been vectored through the exhausts, and samples of the exhaust taken and analyzed. The exhaust is made up of a little clay and mud and water, and is colder than the surrounding air.

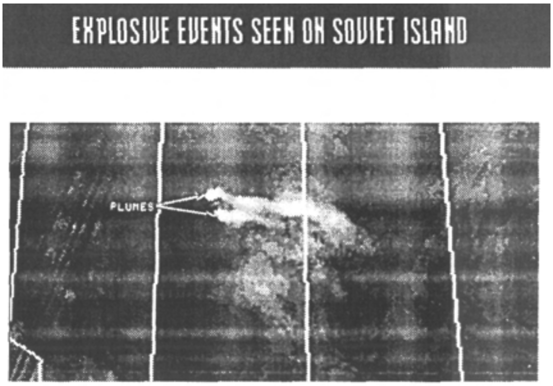

EXPLOSIVE EVENTS SEEN ON SOVIET ISLAND

Slide 049: Explosive Events seen on Soviet Island

U.S. weather analysts euphemistically call these exhausts from Bennett Island plumes.

Here is another U.S. weather satellite photo, showing the twin emergence of multiple puffs.

One might expect such exhausts if the secondary "dumping transfer" howitzer is operated in the continuous mode, with continuous extraction by "power tapping" the accumulator of the primary howitzer. In that case, the signature of the exhaust would expose the operation of the primary howitzer. For example, this exhaust would indicate multiple puffs, hence probably multiple frequencies and multiple primary transmitters, each phased in a controlled manner. That of course indicates the production of Fourier expansion forms - of 3-dimensional geometrical forms - by the primary howitzer(s). Since these are puffs, it may indicate the explosive formation of distant spherical balls of energy by the primary howitzer. This in turn indicates the use of a howitzer system in the "pulse firing" mode, which can be used against missiles, ships, aircraft, ground targets, armored vehicles, underwater targets, etc.

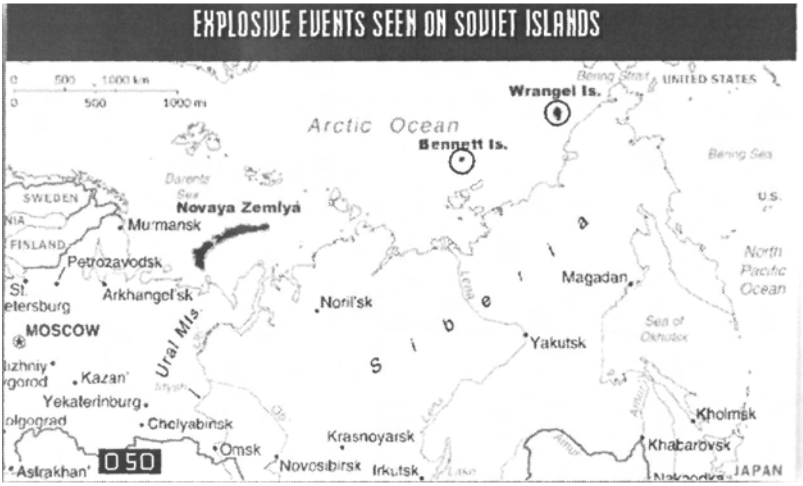
EXPLOSIVE EVENTS SEEN ON SOVIET ISLANDS
0 500 1000 km
0 500 1000 mi
Arctic Ocean
Wrangel Is.
Bennett Is.
Bering Strait
UNITED STATES
Bering Sea
U.S.
North Pacific Ocean
Barents Sea
Novaya Zemlya
SWEDEN
FINLAND
Murmansk
Petrozavodsk
St. Petersburg
Arkhangel'sk
MOSCOW
Nizhniy Novgorod
Kazan'
Yekaterinburg
Volgograd
Chelyabinsk
Omsk
Astrakhan'
Ural Mts.
Noril'sk
Siberia
Lena
Magadan
Yakutsk
Sea of Okhotsk
Krasnoyarsk
Novosibirsk
Irkutsk
Amur
Kholmsk
Khabarovsk
Nakhodka
JAPAN
050

Slide 050: Explosive Events seen on Soviet Islands

This slide shows the location of three areas of interest relative to Soviet scalar EM weapons testing.

Wrangel Island, north of the northeastern portion of the Soviet mainland, is associate with tests of these weapons in large circular breakups of the ocean's ice cover. This of course could aid submarines that wished to surface and fire cruise missiles, or submarines that wished to fire from underwater and needed the **ice** broken up.

Strangely enough, Wrangel Island has a most peculiar history: legally it still belongs to the U.S., but has been "given" to the Soviet Union without permission of the U.S. Senate, which is illegal.

Bennett Island, a little farther to the west and still north of the Soviet **mainland,** is noted for its phenomenal number of anomalous exhausts as previously discussed.

North of the Soviet mainland and near its middle lies the large island of Novaya Zemlya. Weather satellites have taken pictures of anomalous circulations over this island, which are further candidates for exhausts from scalar EM howitzer operations.

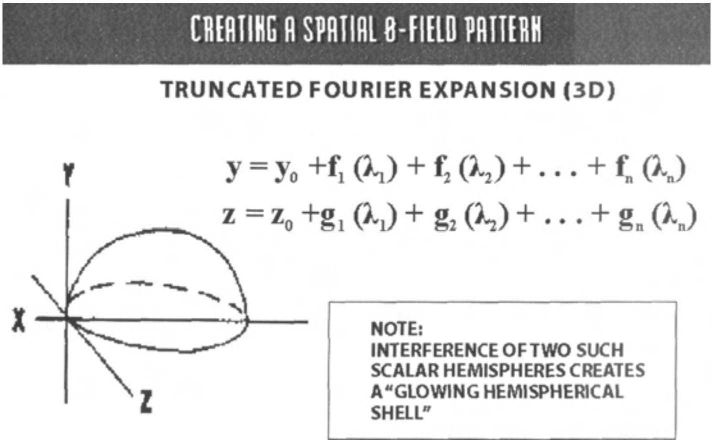

CREATING A SPATIAL 0-FIELD PATTERN

TRUNCATED FOURIER EXPANSION (3D)

$$y = y_0 + f_1(\lambda_1) + f_2(\lambda_2) + \ldots + f_n(\lambda_n)$$

$$z = z_0 + g_1(\lambda_1) + g_2(\lambda_2) + \ldots + g_n(\lambda_n)$$

NOTE:
INTERFERENCE OF TWO SUCH SCALAR HEMISPHERES CREATES A "GLOWING HEMISPHERICAL SHELL"

Slide 051: Creating a Spatial 0-Field Pattern

On this slide we show that a hemispherical dome (or other regular shape) can be composed from scalar waves by 3-dimensional Fourier expansion. By truncating the expansion - say after 10 or 12 frequency terms - a good representation of **the** desired form can be obtained with a practical number of frequencies, and in turn a practical number of transmitters. A computer is used to compute the multiplying functions for each frequency, and the entire complex of frequencies is **then** transmitted to produce the scalar form. By interfering two such scalar forms, a 3-dimensional shell of locked-in, ordinary EM energy is formed in **space.**

Interference of two such scalar hemispherical surfaces creates a "glowing hemispherical shell or dome of energy."

Interference of two such scalar spherical surfaces creates a "glowing spherical shell or globe of energy."

Both such geometric forms have been repeatedly seen.

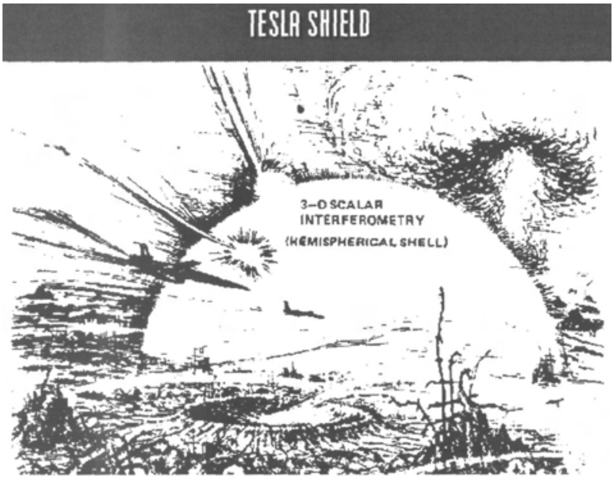

Slide 052: Tesla Shield

Of particular interest is the hemispherical shell of energy, which years ago I dubbed the "Tesla shield."

As shown on this slide, two scalar hemispherical surfaces are created, using multiple frequency transmitters and truncated Fourier series expansions. Interference of the two scalar hemispheres creates a great, glowing hemispherical shell of ordinary electro-magnetic energy. In the shell, the energy density is sufficient to lift Dirac matter from the Dirac Sea of vacuum. The shell is thus filled with a glowing plasma.

Such a shell may be several hundred miles in diameter at the base. The enormous energy required to form such a defense shell is obtained by a "scalar power tap" into the molten core of the earth itself, as previously explained. In late April/early May of 1985, 27 such "power taps" were placed in the earth by the Soviets. If each tap is capable of powering four to six large scalar EM weapons, then the Soviet strategic scalar EM arsenal contains over 100 monstrous superweapons capable of generating exothermic explosions, endothermic explosions, engineering the weather, locating and destroying underwater submarines, detecting and destroying ballistic missiles shortly **after** launch, detecting and destroying long range strategic bombers as soon as they are airborne, etc.

At any rate, the giant Tesla shield is useful against any penetrating vehicle. If the shell is not so large, its energy density may be very high. In **that** case, **the intense** plasma heating will fuse and even vaporize metallic bodies. In addition, any vehicle encountering the shell is subjected to an extremely intense EMP arising everywhere inside its circuitry. EM shielding is of no use against the creation of energy throughout the spacetime of the circuits; such shielding only helps against energy flow through space in the conventional sense.

Thus the electronics of any vehicle encountering the shield are instantly dudded, whether or not they are shielded against ordinary electromagnetic interference (EMI). This includes the electronics operating a nuclear warhead, carried by a re-entry vehicle. Electrically everything penetrating the shell is totally dudded. Further, explosive materials are exploded when such an EMP is encountered, and combustible materials are fiercely consumed or set afire. Ablative shielding suffers an interesting catastrophe: since energy does not try to "flow into" the shielding but "arises everywhere in it simultaneously", "ablation" occurs everywhere throughout the ablative material, simply exploding it instantly. In addition, for smaller Tesla shells (say of 50 miles diameter) the energy density is sufficient to melt or vaporize metals such as in missile structures.

With such a Tesla shield, there is no need to discriminate true warhead-bearing re-entry vehicles from decoys, chaff, etc. The entire "mess" entering the shield is simply "cleaned up" and "sterilized" or destroyed. The shield can take care of ICBMs/IRBMs and their nuclear warheads, strategic bombers and their nuclear **bombs,** cruise missiles and their nuclear warheads, re-entry vehicles and their warheads, decoys, chaff, etc.

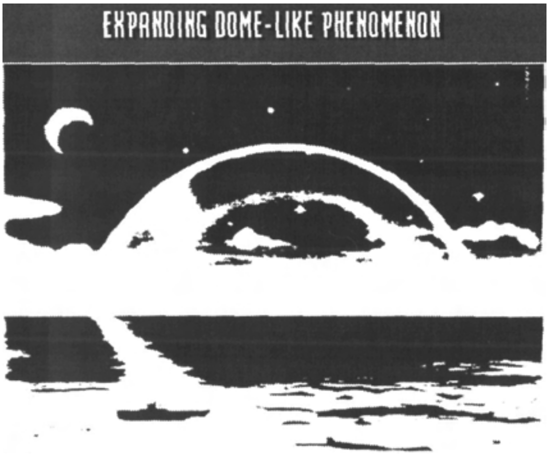

Slide 053: Expanding Dome-like Phenomenon

Here we show one instance of apparent Soviet testing of the Tesla shield (hemispherical shell of energy) over the ocean. This incident occurred in the Caribbean and Western North Atlantic Ocean on Mar. 20, 1969 and is reported in the Marine Observer, Vol. 40, 1970, p. 17. A semicircle of milky white light, quite intense at first, was sighted on the horizon. The semicircle enlarged, dimming as it expanded. The incident lasted for about 10 minutes. Seen from a passing ship, the top of the hemisphere reached a sighting elevation angle of about 50 degrees, while the base of the hemisphere of light subtended an arc of about 70 degrees. This is directly indicative of a test of the Tesla shield, first as a small, more intense shield and then expanding to greater size for a less intense shield whose primary destructive mechanism is EMP. Note that the EMP in a Tesla shield is locally contained in the energy-bottle shell. It is not an ordinary EMP of energy radiated out of the shell.

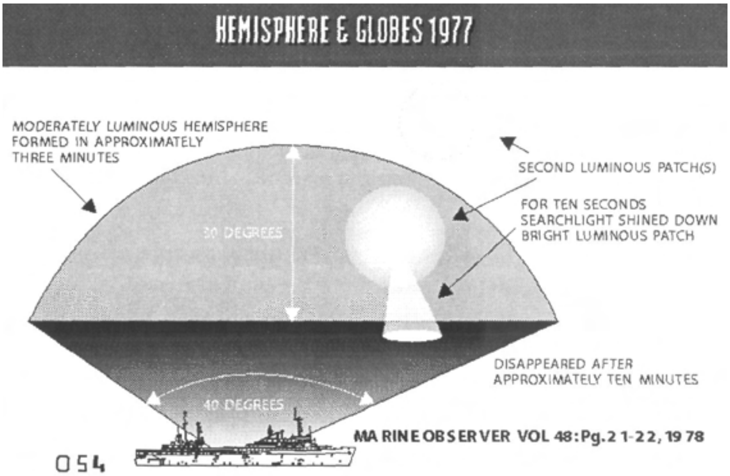
HEMISPHERE & GLOBES 1977
MODERATELY LUMINOUS HEMISPHERE FORMED IN APPROXIMATELY THREE MINUTES
SECOND LUMINOUS PATCH(S)
FOR TEN SECONDS SEARCHLIGHT SHINED DOWN BRIGHT LUMINOUS PATCH
30 DEGREES
DISAPPEARED AFTER APPROXIMATELY TEN MINUTES
40 DEGREES
MARINE OBSERVER VOL 48:Pg.21-22,1978
054

Slide 054: Hemisphere & Globes 1977

Here is a second instance of apparent Soviet testing of the hemispherical Tesla shield over the ocean. This time it is combined with simultaneous testing of two smaller scalar interferometers creating Tesla globes of EM energy.
This incident occurred at 23 degrees 5 minutes north, 17 degrees 25 minutes west. It was sighted and reported by a British ship, H.M.V. Kinpurnie Castle, on Mar. 24, 1977. The report is in the Marine Observer. Vol. 48, 1978, p. 21-22.

First a large, moderately luminous hemisphere of light formed, requiring about 3 minutes to do so. Seen from the ship, the top of the hemisphere was at a sighting elevation angle of about 30 degrees, while the base subtended an arc of about 40 degrees. This was the basic Tesla shield.

Then two other luminous globes or patches of light, each of about 1 degree diameter as seen from the ship, formed in the same area; one inside the shield and one outside it. From the glowing globe inside the shield, a light beam similar to a giant searchlight shone down to the ocean's surface for about 10 seconds.

Again the phenomena lasted about 10 minutes and then disappeared.

The luminous globe or patch inside the hemisphere might be the test of a second weapon deployed from a Soviet aircraft in the area. If so, the aircraft has placed a **strong** spherical shield around itself. In addition, the aircraft is utilizing some sort of scalar interferometer beam weapon in an underwater mode. Probably this would imply a "shot" at a hostile nuclear submarine inside the shielded zone. Thus this aircraft would have the scalar EM equipment to search for the underwater sub and detect it, and also to concentrate interference energy upon it underwater, destroying it quickly.

Suffice it to say that it is possible to perform scalar interferometry by means of a single beam, in a proprietary process discovered by a close colleague. Thus the "glowing beam to the ocean" resembling a searchlight makes very good sense in the anti-submarine role. The same weapon, of course, could also attack any hostile surface vessel that might have been inside the shielded area at its formation.

The second luminous globe or patch, outside the hemisphere, is probably from a second scalar interferometer from the naval task force, simulating the continuous attack of a certain target (missile or aircraft) outside the protective dome.

As can be seen, this phenomenon is directly indicative of the types of operations one would expect from the testing of Soviet scalar EM weapons at sea, possibly to protect a Soviet naval task force during simulated combat operations.

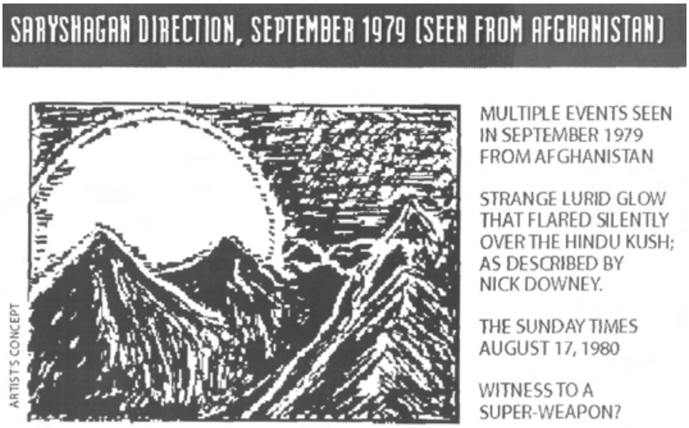

SARYSHAGAN DIRECTION, SEPTEMBER 1979 (SEEN FROM AFGHANISTAN)

MULTIPLE EVENTS SEEN IN SEPTEMBER 1979 FROM AFGHANISTAN

STRANGE LURID GLOW THAT FLARED SILENTLY OVER THE HINDU KUSH; AS DESCRIBED BY NICK DOWNEY.

THE SUNDAY TIMES AUGUST 17, 1980

WITNESS TO A SUPER-WEAPON?

ARTIST'S CONCEPT

Slide 055: Saryshagan Direction, September 1979

The London Sunday Times of 17 August 1980 contained information and a photo-sketch of incidents of sighting of the testing of very large Tesla globes deep within the Soviet Union. British war cameraman Nick Downie made the sightings from Afghanistan. The phenomena seen were in the direction of the Saryshagan Missile Test Range, which - according to the U.S. Defense Department's Soviet Military Power, 1986 - contains one or more large directed energy weapons (DEWs).

Even though Downie was seeing the giant globe of light from a great distance, it flared silently over the Hindu Kush and expanded to subtend an arc of about 20 degrees, dimming as it expanded. (An arc of 20 degrees subtended by an object many hundreds of miles distant indicates an object of well over a hundred miles in diameter. This gives some idea of the enormous energy being controlled and manipulated by these Soviet weapons.).

Downie saw the sight on more than one occasion in September 1979.

BACKGROUND FOR THE BRIEFER:

In the same month - September 1979 - a stationary luminous globe containing a vertical stripe of black in the center was seen in the sky off Saint Petersburg, Florida. This particular type of sighting in that area has been previously correlated with times of known activity at Saryshagan.

Further, in the same month U.S. nuclear-warning Vela satellites detected a mysterious "nuclear flash" over the South Atlantic, off the coast of Africa. Controversy has raged in U.S. intelligence and scientific circles to this day as to whether a nuclear explosion or some other mechanism produced the flash.

Indeed, the flash may have been produced by a scalar EM howitzer from Saryshagan as one more "ping" of the U.S. intelligence system, to ascertain whether or not it knew anything about scalar EM howitzers. Again, the negative response told them with high confidence that (1) we still didn't know about scalar EM stuff, and (2) we were still absolutely defenseless against the Soviet scalar EM weaponry.

In 1980, a second "Vela flash" was reported. This time it occurred in the infrared region only, which positively rules out any sort of nuclear explosion. Even this increased stimulus still evoked a U.S. action that revealed total ignorance of scalar EM weaponry.

An alternate possibility for the September 1979 Vela flash also exists: countries not hostile to the U.S. may have tested a scalar EM weapon that produced the flash.

At any rate, from Downie's sightings, it is highly probably that the DEW weaponry at Saryshagan Missile Test Range was active in September 1979, and was producing huge Tesla globes. If the DEWs at Saryshagan can produce the giant luminous Tesla globe, they almost certainly are scalar EM interferometers and can produce the giant Tesla shields as well. Downie reported other earlier sightings of similar phenomena seen by Afghans deep within the Soviet Union in the same direction toward Saryshagan.

Briefly let us cover the uses of such a "giant globe" or spherical shell of glowing EM energy and plasma.

By placing such a "giant globe" thousands of kilometers out away from the defended heartland, an entire arc of the sky can be defended against long-range ballistic missile attack in midcourse. During their midcourse trajectory, the attacking missiles would have to penetrate the globular shell twice, exposing them to giant internal EMPs twice. A very high probability thus exists that all missiles entering the space occupied by the globe are dudded upon entry and/or exit. This includes the electronics inside the nuclear warheads themselves. Also, this is particularly effective against MIRV and MARV missile carriers, since the multiple re-entry vehicles are normally still on the main vehicle during most of midcourse. The use of this midcourse ABM globe defense greatly reduces the number of vehicles arriving at the latter part of midcourse and at the terminal phase of their trajectory.

By using a smaller, more intense globe and placing it on incoming clusters of objects or single objects, both EMP and intense local heating are used against the objects. This is suitable in the latter part of midcourse and in the **terminal** phase of ICBM's, IRBM's, SLBM's, and cruise missiles. It is also useful against incoming strategic bombers and their air-to-surface missiles, both ballistic and cruise.

Two modes of the globes - especially the small ones - can be used. First, the continuous mode can be used to "fry" or vaporize incoming objects in a relatively small volume (say two or three tens of kilometers in diameter). Second, the "pulse" mode can be used to "service" all incoming objects, whether or not they have passed through the "large globe" midcourse defense. This provides an additional guarantee of killing the objects; discrimination is not required, just service all of them. The exposure of all incoming objects to multiple attacks raises the probability of kill to essentially 100%, or as close to that as one wishes. Of course the incoming vehicle may still encounter a terminal defense consisting of the Tesla shield and associated roving "quickshot" small intense globes.

In short, with these systems an essentially 100% ABM and anti-bomber defense is possible. Further, the Soviets have possessed such an effective defense for two decades, just as they have openly stated.

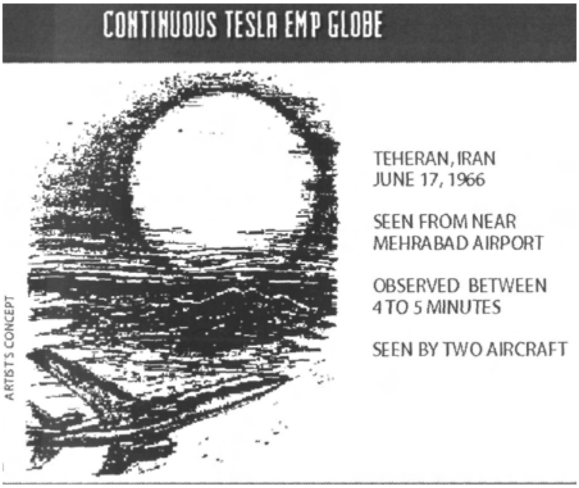

CONTINUOUS TESLA EMP GLOBE

TEHERAN, IRAN
JUNE 17, 1966

SEEN FROM NEAR
MEHRABAD AIRPORT

OBSERVED BETWEEN
4 TO 5 MINUTES

SEEN BY TWO AIRCRAFT

ARTIST'S CONCEPT

Slide 056: Continuous Tesla EMP Globe

Here is another verified incident of a gigantic test of a Soviet scalar EM howitzer deep within the Soviet Union.

This is a CIA report, released under the Freedom of Information Act. One can be quite sure that the incident occurred as stated.

The phenomenon was seen from two aircraft approaching Mehrabad Airport in Teheran, Iran on June 17, 1966 and reported by their pilots.

On the far horizon deep within the Soviet Union, an intense spherical ball of light appeared, "sitting on the horizon" so to speak. The globe of light increased to enormous size, dimming as it did so, literally filling an arc of the distant sky as it expanded. The sighting was shielded from most ground observers' view at the airport itself due to an intervening mountain range that masked most of tin-phenomenon from the ground.

The silent, expanding globe was observed for four or five minutes before it faded away.

BACKGROUND FOR THE BRIEFER:

Again, this is positive evidence of the testing of a giant scalar EM interferometer, in the "midcourse ABM globe" type of action.

However, note the date - mid 1966! The Soviets have therefore been testing such scalar weapons of enormous size and power for at least two decades. This implies that development must have started at least a decade earlier, or in the mid '50s.

Still earlier in January 1960 Nikita Khrushchev had announced the development of a "fantastic" Soviet weapon, one that could even destroy all life on earth if unrestrainedly used. In 1962 the ebullient Khrushchev was forced to back down and lose face before John Kennedy in the Cuban Missile Crisis. Khrushchev's missiles and bombers were in woeful shape, as Kennedy well knew (by courtesy of the Soviet spy, Colonel Oleg Penkovsky). Kennedy did not know that Khrushchev's new superweapons were nearing deployment, but were not yet quite ready.

To save face and prevent his immediate ouster, Khrushchev apparently conducted a startling two-strike demonstration of his new weapons as soon as they became operationally ready. On April 10, 1963 he detected and destroyed **the** U.S.S. Thresher nuclear submarine using a scalar EM howitzer in the underwater "continuous" mode.

The next day he demonstrated the "pulsed" underwater destruction mode for **nuclear** subs by producing a giant underwater EM explosion underneath the ocean 100 miles north of Puerto Rico. Ironically, just as the Atlantic was coming a**live** with U.S. naval vessels searching for the lost Thresher, a second test of the weapons that had finished off the sub occurred to the south of them, unnoticed and disregarded, even though seen and reported (to the FBI and the U.S. Coast Guard) by a passing U.S. jetliner's pilot and crew, who observed the underwater explosion.

We thus can peg the development of these weapons by the Soviets as starting **well** before 1960. The first operational deployment of the giant strategic weaponry occurred in early 1963.

Note that the 1966 testing observed here is completely consistent with this estimated development schedule.

Thus large Soviet strategic scalar EM weapons have been operational on site for 23 years! This implies that at least three additional generations of the scalar EM weapons have been developed and deployed by now.

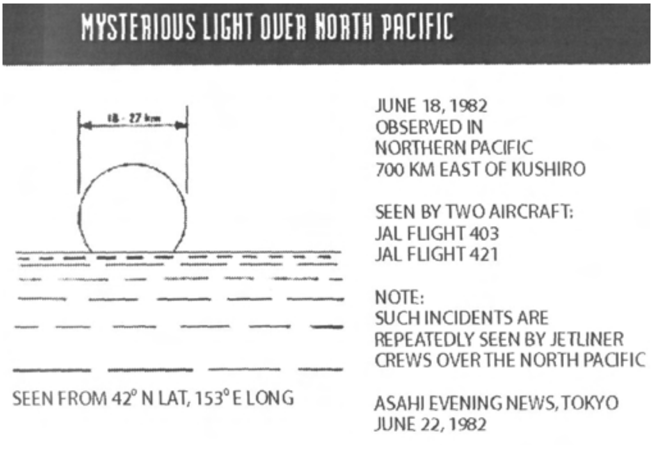

MYSTERIOUS LIGHT OVER NORTH PACIFIC

SEEN FROM 42° N LAT, 153° E LONG

JUNE 18, 1982
OBSERVED IN
NORTHERN PACIFIC
700 KM EAST OF KUSHIRO

SEEN BY TWO AIRCRAFT:
JAL FLIGHT 403
JAL FLIGHT 421

NOTE:
SUCH INCIDENTS ARE
REPEATEDLY SEEN BY JETLINER
CREWS OVER THE NORTH PACIFIC

ASAHI EVENING NEWS, TOKYO
JUNE 22, 1982

Slide 057: Mysterious Light over North Pacific

Incidents of Soviet testing of the Tesla shield and the Tesla globe weapons are routinely observed by airline pilots flying over the North Pacific into and out of Japan.

This is one typical example. Two jetliners, Japan Air Line flights 403 and 421, sighted and reported a large glowing globe of light sitting just beyond the horizon and extending well above it. The aircraft were in the vicinity of 42 degrees N latitude and 153 degrees longitude at the time. The sighting was thus about 700 miles east of Kushiro.

The pilots estimated the diameter of the ball as at least 18-27 kilometers. Depending upon the actual distance to the sphere, it may have been of much larger size.

This incident is reported in the Asahi Evening News, Tokyo, June 22, 1982.

The jet airliner pilots flying over these waters have reported many other **similar** sightings.

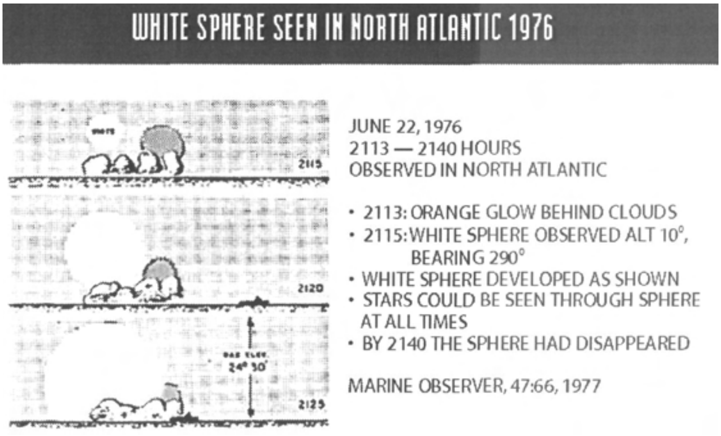
WHITE
2115
2120
OBS ELEV.
24° 30'
2125

WHITE SPHERE SEEN IN NORTH ATLANTIC 1976

JUNE 22, 1976
2113 — 2140 HOURS
OBSERVED IN NORTH ATLANTIC

- 2113: ORANGE GLOW BEHIND CLOUDS
- 2115: WHITE SPHERE OBSERVED ALT 10°, BEARING 290°
- WHITE SPHERE DEVELOPED AS SHOWN
- STARS COULD BE SEEN THROUGH SPHERE AT ALL TIMES
- BY 2140 THE SPHERE HAD DISAPPEARED

MARINE OBSERVER, 47:66, 1977

Slide 058: White Sphere seen in North Atlantic 1976

Here is another sighting closer to home.

This incident was observed from a passing ship on June 22, 1976 in the North Atlantic, at about 2113-2140 hours. It is reported in the Marine Observer, Vol. 47. 1977, p. 66.

First, an orange glow was sighted behind some distant clouds. A couple of minutes later, a glowing white sphere of light was observed to the left of the orange glow, just above the clouds. The white sphere then slowly expanded to a much larger sphere, dimming as it expanded. At its maximum size, the top of the white sphere reached about 24 degrees 30 minutes elevation angle to the observer. Development to maximum size required about 10 minutes.

By 2140 hrs the sphere had faded and disappeared. The sphere was sufficiently thin that the stars could be seen through it at all times.

Again this incident strongly fits the large Tesla globe mode of a Soviet scalar EM interferometer.

The significance and role of the orange glow are not known at this time.

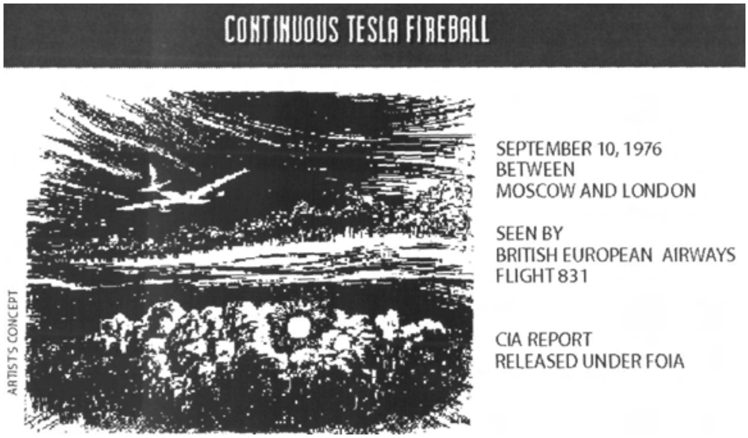

CONTINUOUS TESLA FIREBALL

ARTIST'S CONCEPT

SEPTEMBER 10, 1976
BETWEEN
MOSCOW AND LONDON

SEEN BY
BRITISH EUROPEAN AIRWAYS
FLIGHT 831

CIA REPORT
RELEASED UNDER FOIA

Slide 059: Continuous Tesla Fireball

Here is another incident that represents a stimulus to the British government, to see if the British are aware of scalar electromagnetics.

Again, this is a CIA report released under the Freedom of Information Act, so the details are reliable.

On Sep. 10, 1976 British European Airways flight 831, flying between Moscow and London and over Lithuania at the time, observed an intense ball of light above the clouds below the aircraft. The light was so intense that it lit up the sky in the entire vicinity.

The concerned pilot reported the glowing object to the Soviet ground authorities with whom he was in contact. He received the harsh instructions to ignore the light, and essentially to continue on his way out of there.

Here we see an incident involving a small, intense Tesla globe, of the kind with which the defense could "service" objects that had already penetrated a large midcourse globe, or aircraft approaching the defended area.

Obviously the Soviet authorities were tracking the aircraft, and knew it was in the vicinity. It seems logical, then, that they deliberately placed the brightly glowing ball beneath the aircraft so that the pilot and crew could not fail to observe it.

The strange message to the pilot was simply designed to increase the intensity of the stimulus. The stimulus was to be something like, "The Soviets are doing something in research and development that allows them to create intense balls of glowing light at a distance, and place these objects in the air in and around aircraft, possibly to intercept them." The purpose, of course, was to observe the British government's reaction after the incident was reported by the pilot upon his arrival at London.

Again the reaction of the British - and the U.S. as well - was as predicted. Again we showed that we knew nothing of scalar electromagnetic weapons, and did not recognize one when we encountered its effects.

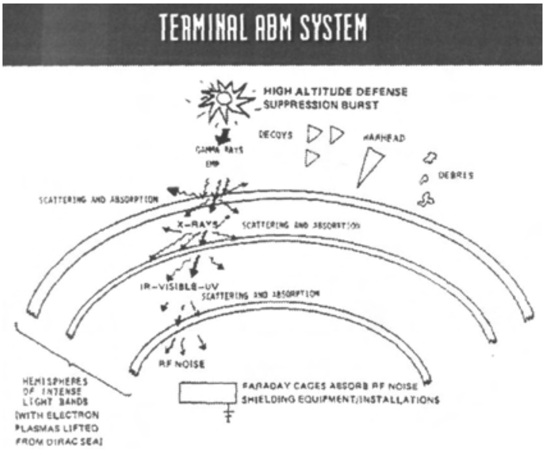
TERMINAL ABM SYSTEM
HIGH ALTITUDE DEFENSE SUPPRESSION BURST
DECOYS
WARHEAD
GAMMA RAYS
EMP
DEBRIS
SCATTERING AND ABSORPTION
X-RAYS
SCATTERING AND ABSORPTION
IR-VISIBLE-UV
SCATTERING AND ABSORPTION
RF NOISE
HEMISPHERES OF INTENSE LIGHT BANDS (WITH ELECTRON PLASMAS LIFTED FROM DIRAC SEA)
FARADAY CAGES ABSORB RF NOISE SHIELDING EQUIPMENT/INSTALLATIONS

Slide 060: Terminal ABM System

Very neat things can be done if one nests several Tesla shields - say three or four - concentrically, one inside the other. In that case even the nuclear radiation (such as gamma rays) from a defense-suppressive high altitude nuclear burst can be handled.

For example, suppose three such concentric shields are placed over a large vital area. Further, suppose a high altitude nuclear burst is placed up above the outer shield. Gamma radiation almost instantly strikes the plasma in the outer shell, where it is absorbed, scattered, and re-radiated at lower temperature. (That, after all, is what plasmas do.)

Inside the first shell, the scattered radiation is now in the x-ray and ultra-violet region. Let us track the most lethal component, the x-rays.

The scattered x-rays then strike the second plasma shell, and are absorbed, scattered, and re-radiated at lower temperature. Inside the second shell the scattered radiation is now in the visible and infrared region, with a little ultraviolet.

This optical radiation in turn strikes the third plasma shell, and is absorbed, scattered, and re-radiated at still lower temperature. Inside the third shell most of the energy is now in the form of radiofrequency (RF) energy, with a little IR and visible band spectral energy content.

At this point, ordinary electromagnetic interference (EMI) shielding of electronic equipments on the ground and inside the third shield can take care of any RF interference resulting from the emergent RF noise.

As can be seen, three shells are sufficient to convert the gamma and x-ray radiation (and ultraviolet and infrared) mostly to harmless RF energy before all three shells are penetrated. Thus the tactic of deliberate defense suppression by a preliminary high altitude nuclear burst can be countered by multiple Tesla shields.

In addition, of course, any ordinary vehicles penetrating all three shields are exposed to successive violent EMP's, and are almost certainly electrically dudded. The vehicles are also subjected to multiple periods of intense heating, so combustibles, fuels, explosives, and ablatives are destroyed. In addition, metal structures may be melted or vaporized.

Think of it this way: anything that hits one of these Tesla shields goes phht! Just **like** a bug hitting an electrified bugkiller screen.

For years passing ships have observed and reported such multiple-shield "light phenomena" over remote regions of the ocean. U.S. intelligence has routinely **not** paid any heed to reports of "lights at night" over remote ocean areas, and so Soviet tests in this manner have remained relatively unnoticed by officialdom.

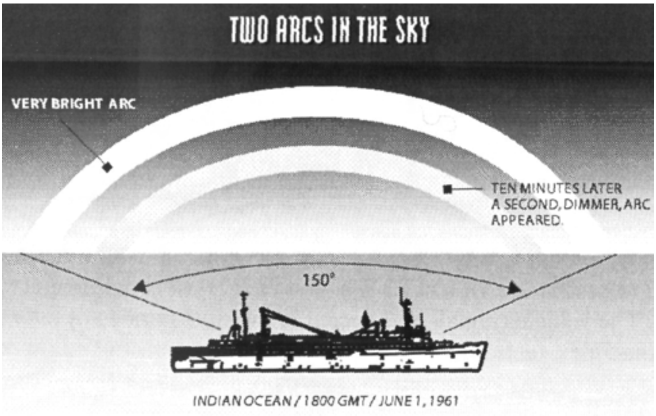
TWO ARCS IN THE SKY
VERY BRIGHT ARC
TEN MINUTES LATER A SECOND, DIMMER, ARC APPEARED.
150°
INDIAN OCEAN / 1800 GMT / JUNE 1, 1961

Slide 061: Two Arcs in the Sky

Just to list one incident involving the testing of multiple concentric Tesla shields, here is an early test conducted in mid-1961 over the Indian Ocean.

The incident occurred at 1800 hrs. Greenwich mean Time (GMT) and was observed by a passing ship. The report is contained in Marine Observer, Vol. 32, 1962, p. 64.

First the ship's crew observed a gigantic, very bright arc of light on the horizon or beyond it, with both ends passing below the line of sight to the horizon. This was in fact probably a large Tesla shield, but of such low intensity that only its intersection arc could clearly be seen. The arc subtended an arc of about 150 degrees across its diameter at the bottom. After about 10 minutes, a second and dimmer arc appeared, concentrically inside the first one. The ship thus was observing what probably was the nesting of two giant hemispheres (of EM energy) to provide a multiple Tesla shield.

This was probably a test of an early research and development prototype, since Khrushchev had only announced his new "fantastic weaponry" in development in his January, 1960 speech to the Soviet Presidium. And apparently land-based weapons such as this prototype probably did not become deployed and operational until April 1963. However, even in 1961 the Soviets would have been developing and testing prototypes of the weapon adapted for naval task force protection and usage.

Several other similar phenomena involving nested hemispheres or giant concentric arcs of light have been seen by mariners at sea.

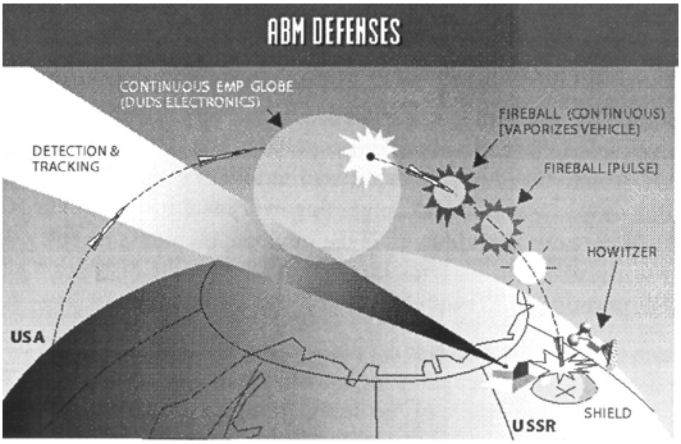

Slide 062: ABM Defenses

When the various capabilities of Tesla shields, globes, and scalar howitzers are integrated together into an ABM defense, a most formidable Soviet strategic defense capability is apparent.

On this slide we deal only with midcourse and terminal phase ABM defense. Later we shall cover the application of these weapons to launch phase ABM defense, primary phase bomber defense, underwater antisubmarine defense, etc.

Suppose the U.S. is provoked to desperately launch a missile attack toward Russia, as indicated on the slide.

First would come the launch phase destruction of the launched missiles, which we will cover shortly.

Any remaining missiles in midcourse would encounter extremely large, glowing spheres of light, where the EM energy density is contained and held in the spherical shell "energy bottle" of the glowing globes. Each globe might cover an are of up to about 20 degrees, as seen from the Soviet defended area. Missiles carrying MIRV and MARV payloads would not yet have deployed their re-entry vehicles.

As a missile and its payload encounter one of the globular shells, an intense EMP is suddenly experienced throughout the missile and all of its components. Again, EMI shielding is of no consequence since the EMP arises in each increment of spacetime throughout the vehicle volume. The result is instantaneous dudding of all electronic circuitry, including the electronics of the nuclear warheads themselves. The HE content of the nuclear warhead system is violently exploded, totally destroying the configuration of the warhead. A low-order nuclear detonation may occur, but more likely the warhead's nuclear material is just broken up, finely divided, and set afire.

In addition, all combustible materials are set fiercely afire or destroyed. Explosive materials such as booster propellants and HE explosives are violently exploded. Ablative and plastic materials literally explode into pieces, torn apart by sudden violent ablation and outgassing everywhere within. The remnants of the vehicle are now total duds, as is the payload, including re-entry vehicle, nuclear warhead electronics, nuclear materials from the disruption and explosion of the HE portion of the warhead system, etc.

Still, each piece of the debris must pass through the glowing shell once more before it exits. Thus a second "shot", precisely as effective as the first, still awaits each and every piece of the debris plunging along inside the shell as it exits.

As the debris continues on to the latter part of midcourse and early terminal phase, any portions large enough to track are subjected to small, intense Tesla globes where the energy density in the shell is many orders of magnitude greater than in the larger globe used in earlier midcourse. This part of the defense system simply services each incoming "blob" methodically, probably first in the pulse mode. Later, any blobs still remaining are given the maximum "continuous heating" mode with a small Tesla globe, to actually vaporize metals and materials.

In the unlikely event that anything active still remains, or that an object has managed to elude the large and small Tesla fireballs, a Tesla shield is still to bo encountered, towering over the protected area. For significant targets this shield will be concentric, and will contain three or more concentric Tesla shields. Again, the approaching vehicle will be subjected to multiple EMP's and severe heating throughout its volume. In addition, the concentric shields will "clean up and sterilize any gamma radiation and EMP resulting from any deliberate high altitude defense suppression nuclear burst that somehow (inconceivably!) gets through the launch phase ABM system, and the Tesla-globe midcourse and earl terminal phase defenses.

As can be seen, the system complex represents a most formidable defense that certainly approaches 100% effectiveness against the present triad, and even against the addition of cruise missiles.

In addition, since these weapons suffer from very strange constraints and limitations, the scalar transmitters are actually dual mode: they become conventional radars by simply switching into the ordinary EM mode. They arc part of the more conventional ABM defense system complexes we presently recognize in the Soviet Union. In this conventional role, backup ABM missile interceptors are available, to be employed in case the conventional role must be assumed.

Later we will address the reason for the strange limitations that must be placed upon unqualified use of the large scalar EM systems, particularly in the pulse or "burst" nodes.

NETWORK OF VIRTUAL TRANSMITTERS

Slide 063: Network of Virtual Transmitters

In July 1976 the U.S. received very special Bicentennial greetings from the Soviet Union.

At that time, communications systems of the world in the 3-30 megaHertz band suddenly met substantial interference from extremely powerful, chirped Soviet transmitters that were suddenly activated. These transmitters continue their transmissions to this day.

Estimates of the power of these enormous transmitters vary, but figures range as high as several hundred megawatts, with a nominal figure being 100 megawatts.

These powerful transmitters were promptly nicknamed "Woodpeckers" because of the characteristic sound of the chirped signal when received. That is, the received signal made a "pecking" sound much like a woodpecker's beak hitting a block of wood.

Several nations protested, but the powerful signals have continued, right down to this day. The only Soviet response was to add a "spread spectrum" capability, so that the transmitter did not dwell too long on one specific frequency, but shifted periodically to other frequencies.

These transmitters have apparently never been precisely located by U.S. intelligence, but their beams carry much of the characteristics of an **over-the** horizon (OTH) radar. They have been dubbed OTH-radars by U.S. intelligence, and can without question perform that mission, in addition to some very interesting missions which U.S. intelligence does not assess.

Soviet Military Power, Department of Defense, 1985, p. 45 shows the direct intersection over the United States of the Woodpecker radar beams used in an OTH-role. We show an illustration of that intersection on this slide.

In addition, we show an additional "scanner" beam that can be scanned across the intersection "grid" over the U.S., formed by waveform interference of two main Woodpecker beams.

These Woodpecker transmitters have a great many diverse functions.

First, they can be used in a conventional over-the-horizon radar node, since their beams follow the earth-ionosphere waveguide and curve around the earth. In this mode they can detect missiles at launch and thereafter, and strategic bombers when taking off and thereafter.

However, these scalar interference grid weapons have many other uses.

First, they can be used to biologically attack entire populations in a targeted area. This aspect is not covered in this briefing. Suffice it to say that phase-locked ELF modulation signals ot 10 Hz and less are often detected on multiple Woodpecker frequencies simultaneously. In a target area, this modulation - if sufficiently stronger than the Schumann resonance of the earth's magnetic field - will entrain a percentage of the brains into "forced entrainment." In that case, these human brains are "synchronized" to the Woodpecker signals so that multiple coherent frequencies are phase-locked into them. That is, multiple coherent EM channels directly into these entrained brains now exist. At that **point,** Fourier expansions may be used to attack specific portions of the brain geometrically.

In addition, scalar EM disease patterns can be modulated upon the carriers, again with Fourier expansions. Specific biological effects can be induced in the entrained populace at will, limited only by the state of the art of the Soviet technology used to attack them. Possible induced effects include instantaneous **death,** heart seizure, severe emotional disruption, loss of control of internal functions, diseases, disabling of the immune system, and even implantation of **thoughts,** emotions, and ideas which are interpreted by the targeted subjects as their own.

While further discussion of this area is beyond the scope of this briefing, the biological aspects of the Woodpecker transmitters are horrible. It suffices to say that, in thousands of experiments, Kaznacheyev demonstrated that almost any **kind** of cellular death and disease pattern could be electromagnetically transmitted. Kaznacheyev reported the effect in the near ultraviolet. Experimenters at the University of Marburg in West Germany duplicated the experiments in the infrared.

The bottom line is that photons themselves can carry death and disease patterns between cells. Scalar E*M* technology allows the synthesis of the actual potential pattern (which after all represents total control of charge and charge distribution, hence biochemistry in the cell) of a particular cellular disease or death mechanism. Symptoms (and cellular death from them!) of nuclear radiation, chemical poisoning, bacterial infection, and other mechanisms were successfully induced by the Kaznacheyev experiments.

Dr. Popp of West Germany has published an analysis of the virtual photon master control system of the cells. Since scalar EM represents the deliberate ordering of virtual particle flux into deterministic patterns, the master control system can readily be entered with scalar techniques to induce disease and disorder at will.

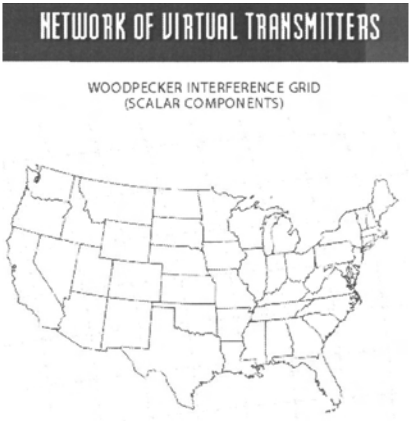

NETWORK OF VIRTUAL TRANSMITTERS

WOODPECKER INTERFERENCE GRID
(SCALAR COMPONENTS)

Slide 064: Network of Virtual Transmitters

On this slide we diagrammatically represent the Soviet woodpecker interference grid over the U.S. This grid can easily be used to influence and control the weather, and for over a decade the Soviets have manipulated the weather over the U.S. and other parts of the world.

By the peculiar title of this slide, we wish to call attention to the fact that, in the exothermic mode of scalar interferometry riding on the woodpecker radar signals, each of the woodpecker grid interference cells can act as a little "virtual transmitter." It's exactly as if the Soviets had been permitted to enter the U.S. and build thousands of EM transmitters at regular intervals all over the country!

And of course it's even worse. In the endothermic mode, it's as if the Soviets had been permitted to enter the U.S. and build thousands of "cold generators" all over the U.S. That is, in that mode, each of the "transmitters" can actually extract electromagnetic energy.

These two modes have enabled the Soviet Union to drastically manipulate the weather at will over the U.S. during the last decade (and similarly, for other places as well.)

(For a disclosure of the Soviet weather control operations over the U.S., see T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985, available from The Tom Bearden Website www.cheniere.org. We will only summarize here.).

Briefly, inside various reference level zeros on the ordinary EM woodpecker carriers, scalar "channels" are established. Scalar beams, signals, and interferometer pulses and transmissions can be transmitted through these reference channels as if directly through a wire. Phasing, pulse shaping, and Fourier expansion formation of geometric forms are then the key to be used in scalar interferometry through the internal scalar channels of the Soviet woodpecker signals.

By choosing one or more grid cells and using scalar interference there in the exothermic continuous mode, local heating and expansion of the air, and formation of a high-pressure area, occurs.

By choosing one or more grid cells and using scalar interference there in the endothermic continuous mode, local cooling and shrinking of the air, and formation of a low-pressure area, occurs.

By electromagnetic "rotation" of the interfering scalar beams, the hotspots and coldspots can be moved along a desired path and at a desired speed.

By these methods the entire jetstream across the U.S. can be deviated and controlled to a large degree. Moisture can be drawn from the Pacific beyond Southern California and Mexico, and collided over the southern U.S. with extremely cold air brought down from Canada, producing ice storms such as those in and around February 1, 1985.

Violent thunderstorm activity can be induced and directed. By adding "spin", these same storms can be induced to form extensive tornados.

By sustained weather engineering operations, drought or excessive rain and flooding can be induced in areas of the U.S. as desired. Crops can be destroyed or heavily damaged by severe and unseasonal weather.

If the grid is placed in the ground on scalar EM carriers instead of normal EM carriers (and the woodpecker "radars," can switch to, or add, scalar carrier mode whenever they wish), it can be used as an anti-submarine weapon, as we shall see later. It can also be used to cause substantial ocean effects that have drastic consequences on the world's weather - such as the anomalous El Nino of 1982-83.

By localizing one spot of heating at the top of a thunderstorm anvil, and another spot of cooling at its side, a localized downburst can be created. Our weather satellites have detected just such an anomalous combination heating and cooling of thunderstorm anvils over the U.S. While there may exist natural mechanisms to cause this anomalous dual pattern, it may also be caused by Soviet weather engineering, which can cause either broad-area or highly localized effects.

All these types of weather engineering incidents have been deliberately created over the U.S. by the Soviet Union.

Perhaps the greatest weather effects can be induced by altering some of the large-scale normal mechanisms that influence and control our weather. Thus the **artificially** induced El Nino of 1982-83 was a case in point, causing drastic weather upsets worldwide for a year or more.

The last decade has seen weather so anomalous that it could only occur once in 1200 years by chance. It has not been due to normal chance, but by Soviet weather engineering, particularly since the advent of the Woodpecker transmitter complexes.

ADDITIONAL BACKGROUND FOR THE BRIEFER:

The Woodpecker transmitters can carry scalar components, which, in addition to biological strikes, can be used for many other purposes such as to

1. Influence and control the weather,
2. Deviate the giant jetstreams of the upper atmosphere,
3. Detect, track and destroy missiles shortly after launch,
4. Detect, track and destroy strategic bombers and other aircraft,
5. Attack land and ocean surface targets by large aerial electromagnetic explosions: the EM explosion produces the fireball, the thermal effects, and the blast wave of a nuclear explosion without the nuclear radiation and radioactive fallout contamination;
6. Attack surface land and ocean targets by large "cold explosions", where living things freeze almost instantly, leaving the surface facilities and equipment intact,
7. Attack land and ocean surface targets by severe internal electromagnetic disruption, so that all electronics are dudded,
8. Attack personnel on land and sea by direct destruction of their nervous systems with an internally induced strong EMP,
9. Disrupt the operation of distant facilities, equipment, and personnel or destroy food crops in large areas by severe cold or severe heating, or even startling oscillation between the two.

    In addition, the scalar aspects of the Woodpecker transmitters can be beamed through the earth and ocean as surface waves or at a given depth to intersect at a distant region. This mode can be used to

10. Cause distant earthquakes in a distant fault zone by depositing energy in the rocks there, increasing the plate stress until slippage occurs at a fault zone,

11. Cause large, anomalous rocking of an entire plate or large area, by depositing and extracting energy from alternate areas, causing buckling of the large rock plate,
12. Set up an interference grid under the water, so that underwater submarines can be detected, tracked, and destroyed on station and with their missiles still unfired,
13. Attack naval vessels and task forces through the water and through their hulls, without warning,
14. Attack missile silos and nuclear storage sites "before launch", directly through the earth itself,
15. Initiate the full nuclear explosion of implosion-type nuclear warheads on-board missiles on station,
16. Initiate full nuclear explosion of implosion-type nuclear warheads in storage sites and ready reserve,
17. Destroy seaborne and ocean-bottom-mounted mines and underwater sensors, even including entire "fields" of them,
18. Detect and destroy hostile torpedoes, ship-to-underwater anti-submarine missiles, and dumped anti-submarine munitions and charges,
19. "Pump" the electrical grounding of distant electrical power distribution and generation systems so that the system catastrophically fails,
20. "Pump" the electrical grounding of distant electrical power systems so that radio and television transmitters, power grids, and wiring complexes in buildings and facilities radiate a desired scalar wave component or signal complex at equipment and personnel.

All in all, the woodpecker transmitter complexes and other associated scalar EM weaponry are most useful in the major missions required in war.

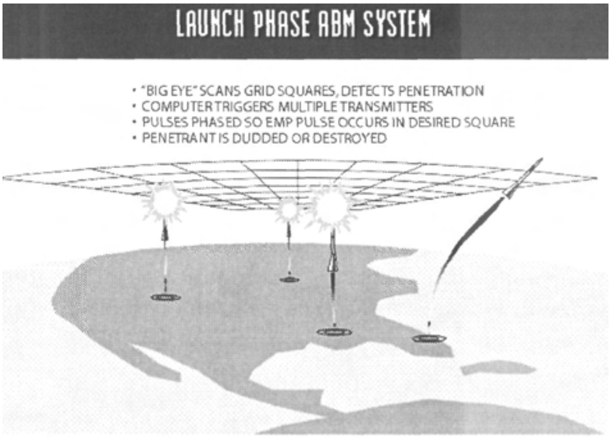

LAUNCH PHASE ABM SYSTEM

- "BIG EYE" SCANS GRID SQUARES, DETECTS PENETRATION
- COMPUTER TRIGGERS MULTIPLE TRANSMITTERS
- PULSES PHASED SO EMP PULSE OCCURS IN DESIRED SQUARE
- PENETRANT IS DUDDED OR DESTROYED

Slide 065: Launch Phase ABM System

On this slide we show the use of the Woodpecker grid in the Launch Phase **ABM** system role.

The scalar components establish an interference grid over the U.S., at some altitude above the earth. One or more separate scalar interferometers transmit through scalar channels established in the Woodpecker's normal EM carriers, scanning through the grid square cells one by one in the "just barely extract energy" endothermic mode.

Back in the Soviet Union, the energy extracted passes through a scanned receiver biased just so that the normal energy extracted from the distant atmosphere does not show on a scanning screen. Any substantial amount of additional energy appearing in the cell - such as from the jet engines of a large aircraft or the rocket exhaust of a large missile - will be detected.

The **operation** of this weapon is four-fold: (1) the scanning interferometer or "big eye" scans the grid squares, detects penetration, and tracks the rising missile. (2) A separate scalar EM interferometer (a howitzer) is controlled by a computer and fired through a separate scalar EM channel opened in the ordinary EM woodpecker carriers. (3) Two timed scalar pulses are fired so that they converge and meet in the grid cell containing the rising missile. This results in the **sudden** emergence of a violent EMP pulse throughout the missile and the surrounding vicinity. (4) All missile and warhead electrical and electronic systems are instantly dudded and destroyed. In addition, fuels, propellants, and explosives of the missile are immediately ignited, explosively destroying the penetrating missile.

That takes care of the missile target. (The same weapon, of course, can be used against airborne strategic bombers - including Stealth-type bombers. In fact, Gary Powers' high flying aircraft was probably shot down in 1960 by a "jury-rigged" scalar EM howitzer using modified radars and timed scalar pulses to provide an aerial explosion and EMP.)

The Woodpecker grid/howitzer weapon system can be placed over the ocean and used against cruise missiles, naval surface-to-air and surface-to-surface missiles, submarine-launched ballistic and cruise missiles, etc. Placed over a carrier task force, it can also take care of the aircraft launched by the carrier as fast as they are launched. It can also handily take care of the missiles launched by guided-missile cruisers of the accompanying task force.

Now let's take a look at the command and fire control end of this weapon system, way back in the Soviet Union.

SOVIET COMMAND & CONTROL CONCEPT

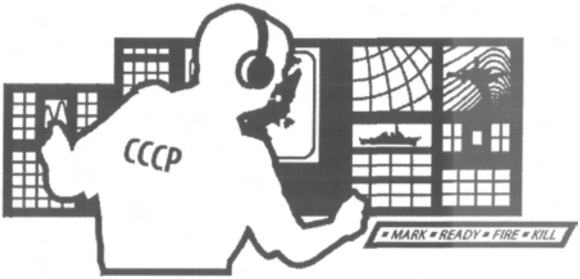

Slide 066: Soviet Command & Control Concept

On this slide we show a postulated Soviet fire control system for the grid weapon.

Two or more operators sit at a control console, facing a giant display (screen). The display contains a grid, which represents the distant interference grid of the Woodpecker carriers (or scalar EM carriers, if in the underground or underwater mode). A distant missile launched into the real interference grid overhead will cause a spot of light to appear in the appropriate grid square (cell) on the operators' screen. The operators will track the displayed target, select firing mode and type of action, engage the target, and assess whether or not the distant target is killed or missed by the firing.

When a missile is launched in the distant area over which the Woodpecker grid has been placed, it quickly rises through one of the scanned grid cells on its upward journey. As it penetrates the cell, additional energy is extracted (since part of the missile's exhaust energy is now extracted also). At the screen in the Soviet Union, the additional energy received rises above bias, causing a "bloom" or spot of light to appear on the screen at the appropriate grid cell representation.

The distant operator has now detected the missile launch and its location. He places a marker over the target track and activates the computer, also inputting the type of firing or action to be performed. In this case, let us say he wishes a burst of **energy** to emerge inside the distant launched missile, and has activated that mode of firing action.

The computer computes the necessary parameters for another scalar interferometer channel in the exothermic mode, and for firing at the location of the rising missile so that two scalar pulses will meet in the cell penetrated by the rising missile on the other side of the earth. As the channel is opened, the firing solution settled, and the interferometer howitzer readied, the computer activates indicators on the console notifying the operator. When all is ready, a "ready-to-fire" light is lit on the console.

Upon receiving the command to fire and destroy the missile, the operator presses the fire button. The computer fires the activated scalar EM howitzer in the pulse mode. The operator continues to watch his screen.

When the pulses meet in the distant grid cell, a violent EMP suddenly arises in and throughout the missile and its surrounding vicinity. This explodes the missile propellants and warheads, destroying the target.

On the distant screen, a sudden drastic "blossoming" of the target results from the sudden extraction of a great deal of additional energy by the scanning scalar interferometer. The operator thus knows he has made a "kill."

Should for some reason the target be missed, additional energy will still be extracted by the scanning interferometer from the grid cell/vicinity in which the EMP suddenly emerges. The distant operator will still see a bloom on his scope, but not nearly so great as when the missile explodes. He thus knows that the EMP effect of the firing has occurred, but the target has been missed (he will see the bloom in its offset location as well). In that case, the operator can quickly mark the EMP location, hit the switch, and the computer will automatically correct and fire again.

At any rate, for a miss a large "sonic boom" or blast still results in the EMP emergence zone from sudden heating of the air. These are precisely the type of booms that were associated with three NASA shuttle launches prior to the end of 1985. These were actual testing of this weapon system, using the shuttle launches to provide a target, and delaying the burst some minutes so the shuttle would not actually be destroyed. On Nov. 26, 1985 a "marker beacon" (glowing ball of light) was also created over the site. This was probably to orient satellites and other detection systems. Some 12 minutes after that nighttime launch, a large blast occurred over the site, heard for hundreds of miles up and down the coast. Shortly after, the light suddenly moved away very rapidly - faster than a jet aircraft.

If many missiles are being launched, the operator marks them rapidly, one after the other, and the howitzer fires burst after burst at them, decimating the launched missiles, one after the other. In this fashion most of the counterstrike missiles launched by the U.S. would be destroyed shortly after launch, greatly reducing the number of missiles that make it to midcourse.

However, apparently the loss of the shuttle launched on Jan. 28, 1986 was caused by the addition of a metal-softening pattern in the exothermic mode, in and around the booster, using the actual booster ionic flames as a receiver-amplifier. This led to the failure of one of the supports and the partial breakaway of the right booster. The booster rotated into the tank, causing damage and the resulting explosion when the venting main fuel hit the booster flame.

By spread spectrum techniques and proper timing and phasing, one system of Woodpecker transmitters can set up multiple interference grids, in various parts of the world. The scanners and howitzers can operate in the appropriate zones, again by spread spectrum techniques and proper timing and phasing.

In fact, the latest versions probably can operate in the "ordinary EM carrier" (atmospheric) mode and the "scalar EM carrier" (underwater and in the earth) mode simultaneously - again by timed switching between EM and scalar EM carriers, spread spectrum techniques, and timing and phasing of howitzers and Manners.

Thus a single system can operate in the earth, underwater, and atmospheric modes. It can operate against the very wide range of different targets previously mentioned. It can do all this "simultaneously," from a real-time practical viewpoint.

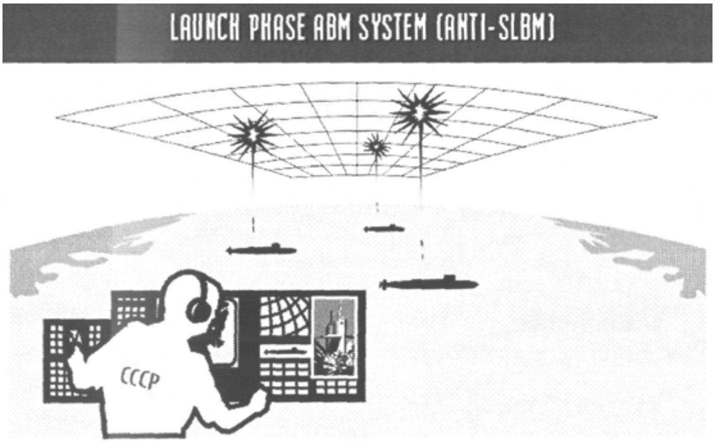
LAUNCH PHASE ABM SYSTEM (ANTI-SLBM)
CCCP

Slide 067: Launch Phase ABM System (Anti-SLBM)

Here we show the operation of the system in the atmosphere over the ocean, to destroy sub-launched ballistic missiles such as Poseidon and Polaris. The system operation shown uses the normal EM Woodpecker carrier frequencies, with scalar components to establish the interference grid, and scalar channels for **the** scanners and howitzers. The grid can be set up by this method anywhere in the world.

As can be seen, during their early launch phase, this mode will take care of the missiles launched from underwater U.S., British, and French nuclear submarines.

Remember, this is just one mode in which the system is operating. It can at the same time have other interference grids at other locations, and even in other modes, including underwater grids used against the submarines themselves, it can be engaging a multitude of targets of a wide variety of types, essentially all simultaneously.

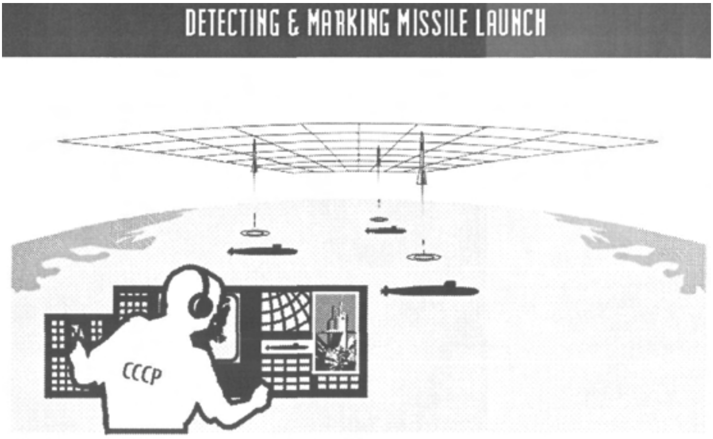
DETECTING & MARKING MISSILE LAUNCH

Slide 068: Detecting & Marking Missile Launch

Here we show the missiles launched from the underwater submarines, when just detected in the atmosphere over the ocean. As each rising missile is detected, an operator at the screen places a cursor (a "marker") over the detected track spot on the scope.

The computer quickly computes the firing parameters and indicates when the system is "ready to fire."

By using additional scalar scanner frequencies in the multi-gigaHertz range and fine scanning, the operator is able to accomplish non-cooperative target identification.

After the operator places a cursor over a target, the advanced system fine-scans and identifies it as hostile, then continues the track in automatic.

When the operator presses the fire button for a designated target, the computer intiates fire by the appropriate scalar EM howitzer through established grid scalar firing channels.

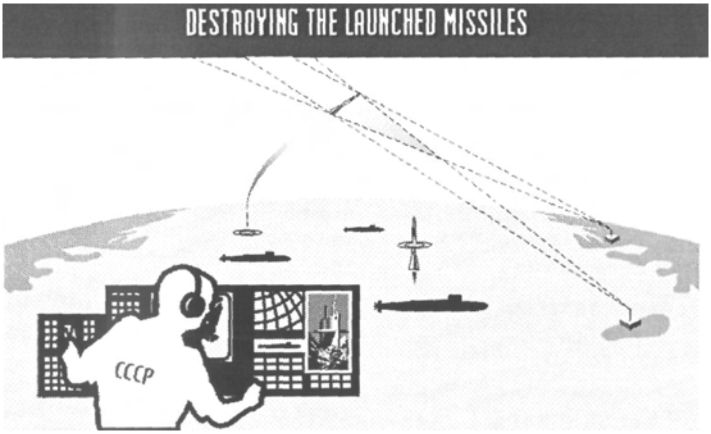

DESTROYING THE LAUNCHED MISSILES

Slide 069: Destroying the Launched Missiles

On this slide we temporarily omit the interference grid and diagrammatically show the operation of the scalar EM howitzer. As shown, two pulses are fired so that they meet in the grid cell where the rising missile is located. A drastic EM pulse suddenly emerges everywhere throughout the missile and its vicinity causing instant missile destruction.

Of course, individual howitzers not in the Woodpecker carrier's scalar channel can also fire at the missiles in the purely scalar exothermic mode. Or they can utilize the large "Tesla globe" in the continuous mode, the small Tesla "ball of light" in the continuous or pulsed mode, etc.

Also, continuous EM energy can be caused to emerge in a broader, "defocused" interference area causing electronic dudding of circuits in vehicles over a wider area. This is a very useful mode to use against mass attacks by many aircraft and/or many missiles at the same time. That is, the entire formation can be attacked in the "dud electronics" mode, and then followed up by more intense servicing of each individual hapless target.

Again, in 1960 Gary Powers was probably downed by a crude prototype scalar interferometer "howitzer" obtained by modifying two ordinary radars. At least the Soviets had previously been unable to down a penetrating U-2 with missiles they possessed. We know the scalar howitzers were in development, for Khrushchev had announced to the Presidium in January 1960 that a new "fantastic weapon" was in development. A development prototype was probably used to down Powers' aircraft.

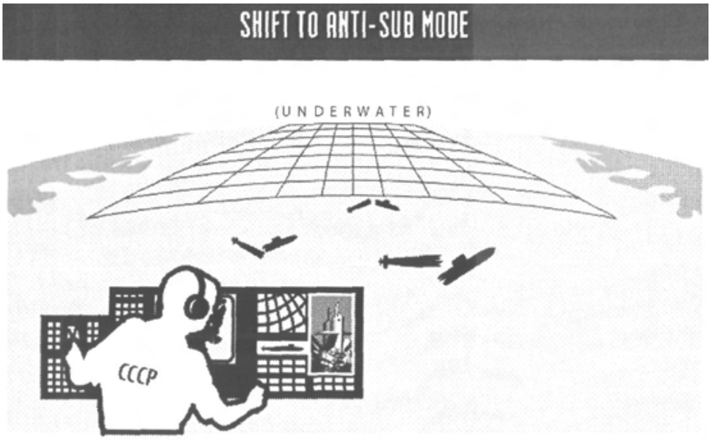
SHIFT TO ANTI-SUB MODE

Slide 070: Shift to Anti-Sub Mode

In the underwater mode portrayed on this slide, the Woodpecker transmitters have been switched to "scalar carriers" mode and the carrier beams transmitted through the ground and into the ocean underwater. (Remember, an artificial EM potential that does not couple to charged particles or ions will travel through the ocean, but a force field will not.)

Thus a giant scalar interference grid is now established underwater. Additional scalar beams in the weakly endothermic mode scan this grid. When a submarine enters the grid, it substantially changes the potential in that cell, and the scanner picks this up. The distant screen now displays the sub as a "track" of light in the appropriate grid square cell. The computer switches in fine-scanning and non-cooperative target identification. It identifies the target as hostile and alerts the operator (by blinking the track and changing its color to red, for example).

As before, the operator marks the target and enters the firing mode and type of action desired. Again, let us assume the "EMP burst" mode of destruction is to be used. The computer computes the firing parameters and readies the howitzer, When the commander gives the command to fire, the operator presses the firing button. The computer fires the howitzer.

Two scalar pulses proceed through the howitzer scalar channel and meet in the submarine. Throughout the sub and its vicinity, a violent EMP is suddenly experienced, knocking out the electrical systems, electronic equipment, and igniting propellants and explosives in the on-board missiles and armament. This explodes the submarine, missiles, and torpedoes instantly.

At the distant command console, the scanners detect the violent eruption of energy from the exploding submarine. It appears on the screen as a sudden strong blossoming of the track. The operator assesses that the target has been destroyed. He immediately engages another target if one is present.

If the EMP is off-course and misses, the scanners still pick up a smaller blossoming in the offset location. The operator immediately knows he has missed the target. He marks the offset bloom, presses the switch for "Correct!" and the computer makes the firing correction. When the operator again gets a Ready-to-Fire light illuminated, he presses the firing button and fires again at the target, this time with corrections applied.

Again, the same Woodpecker system - using timing, phasing, scalar/EM switching and spread-spectrum techniques - may be involved in several grids and many types of targets. When in a target-rich environment, the firing operators may be very busy indeed.

As we shall later see, the same weapon system may also be firing in the "cold explosion" or pused endothermic mode, where intense bursts of instant and severe freezing are unleashed in the target zone.

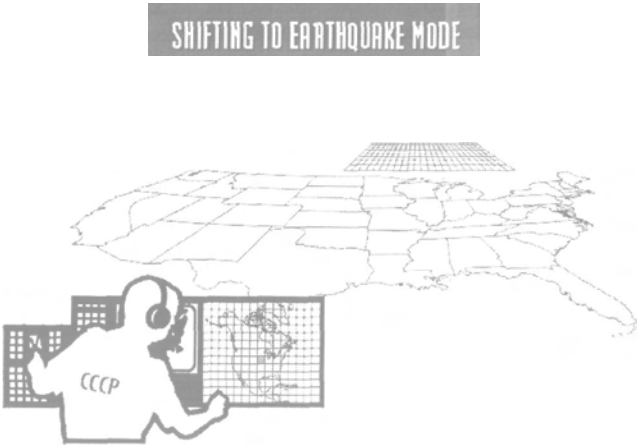

Slide 071: Shifting to Earthquake Mode

The uses of the complex grid, scanners, and scalar howitzers is limited only by human ingenuity and state-of-the-art of the technology developed.

By shifting to the scalar-carrier mode and transmitting in the surface layer of the earth, the interference zone can be established in a distant region of the earth, beneath the earth's surface.

Suppose we establish a very small interference grid in the earth at and around a distant fault zone. Then by steadily depositing energy in the zone in the continous exothermic mode, the stress in the fault area is steadily and gradually increased. Eventually the plates on each side of the fault will slip, and a "natural appearing earthquake" occurs. If the energy is deposited slowly and gently, a gentler quake can usually be effected. If the energy is deposited fairly rapidly, a higher peak stress can usually be reached before the rocks slip, and a larger, more destructive earthquake results.

By making a wider grid in the earth in a larger area, a strange kind of earthquake can be induced, whether or not a fault zone exists in the area. By pouring energy out of alternate cells and extracting energy from the others, a "wave-like" buckling tensile and compressive stress is created throughout the zone. As this continues, the entire plate area will try to oscillate, moving and causing an anomalous sort of "rocking, flat earthquake" throughout a large area.

Not too many days after the destruction of the Challenger on January 28, 1986, such an anomalous "flat earthquake" occurred in the U.S., near the Great Lakes, involving some 10 states. This was probably the subsequent testing of an additional weapons mode of the same Soviet scalar EM system that destroyed the shuttle.

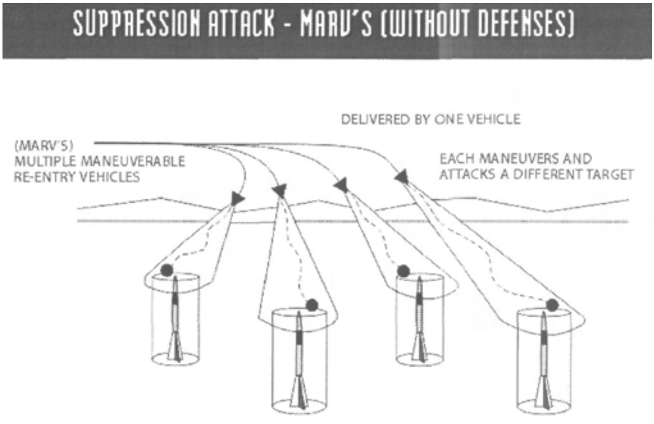
SUPPRESSION ATTACK - MARV'S (WITHOUT DEFENSES)

DELIVERED BY ONE VEHICLE
(MARV'S)
MULTIPLE MANEUVERABLE
RE-ENTRY VEHICLES
EACH MANEUVERS AND
ATTACKS A DIFFERENT TARGET

Slide 072: Suppression Attack—MARVs (without defenses)

Now to shift the emphasis a bit, and look at the impact on some "normal" strategic systems and concepts.

On this slide we show the basic idea of suppression attack against an enemy's ground based missile force in fixed-site silos. The attack uses multiple maneuverable re-entry vehicles (MARVs) delivered by a single carrier missile, which acts as a "bus." The bus ejects the MARVs at the appropriate time, and each of the MARVs heads toward its predesignated target, maneuvering as it re-enters the atmosphere and proceeds to the target. The maneuver makes conventional intercept of the MARV by conventional anti-ballistic missiles more difficult.

In this manner, missile silos without defenses may be caught by surprise and destroyed.

The "bus" concept actually multiplies the number of targets a single launched missile can attack, and the maneuvering of the MARV increases the difficulty of defending the target area by orthodox means.

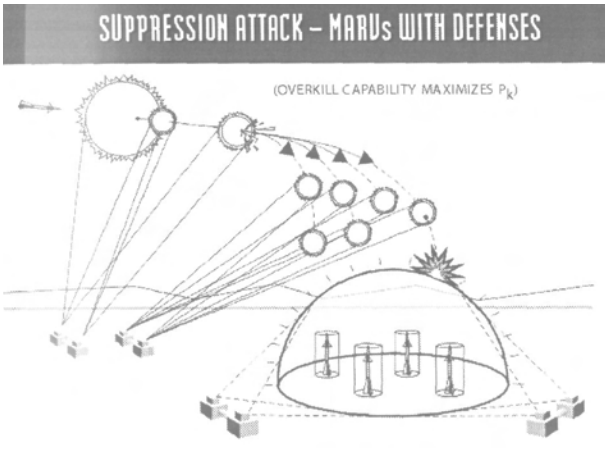

SUPPRESSION ATTACK – MARVs WITH DEFENSES

(OVERKILL CAPABILITY MAXIMIZES P_k)

Slide 073: Suppression Attack—MARVs (with defenses)

Now suppose some of the scalar EM defenses we have discussed are added to the picture.

On this slide we show a similar missile field, defended by a Tesla dome and by multiple scalar howitzers.
The howitzers take the incoming buses under attack during the early midcourse phase, by means of very large spherical EMP globes. Many of the buses - and their entire payloads, including the nuclear warheads themselves - are dudded by these midcourse defenses.

Servicing of dudded vehicles that have penetrated the globe now occurs farther along in midcourse. Each dud vehicle may be serviced, for example, with a small globe in the continuous mode for an appreciable number of seconds. This causes vaporization or fusing of the physical structures of the vehicles.

Groupings of incoming vehicles may also be serviced by medium-sized globe in the pulse mode. This assures dudding, and is especially useful for clearing away chaff and debris with smaller ballistic coefficients. It also is useful to further insure that the nuclear warheads themselves are totally dudded.

An additional wrinkle is to place scalar signal patterns - that affect and transmute a nuclear warhead's radioactive nuclei - onto a continuous globe. In short, by the proper scalar signals it is possible to transmute plutonium and uranium into harmless elements, and it can be done at a distance. The 4th and 5th generation Soviet scalar EM technology is probably capable of doing **this**

This insures that, not only are the incoming nuclear warheads electromagnetically inert, but they are also nuclearly inert and incapable of generating nuclear explosions.

Still another wrinkle - particularly useful against implosion-type weapons - is to utilize a scalar signal pattern that causes sudden generation of neutrons in the active material (plutonium or uranium). This harmlessly explodes the nuclear warheads while they are far away from their intended targets and out in space. The burst mode is probably the most useful for this tactic.

Note that, if desired, this same tactic - exploding the nuclear weapons themselves - can also be utilized in the Launch Phase ABM system grid against the missiles after launch. This capability converts an adversary's intended nuclear attack against you to an actual attack against himself!

Also note that the adapted grid system in the surface of the earth can be used to explode the nuclear warheads in ICBMs while they are still in the silos, or to explode nuclear warheads in storage. In the underwater mode, the nuclear warheads in the Poseidon and Polaris missiles on-board the nuclear submarines can be exploded in a similar manner. In fact, the nuclear reactor used to furnish the power for the nuclear submarines can also be exploded or disabled in this manner, as can land-based and ship-based reactors.

Again, this actually turns an adversary's own nuclear weapons against him. His own weapons can be used to obliterate him. Without scalar defenses, possessing the capability of nuclear attack becomes a serious detriment!

But back to our slide. Just in case something gets through this formidable defense somehow, the Tesla Shield (dome) provides the final reiteration that takes care of any residue "leakage."

The sum total of these defenses fully justifies the 1968 Soviet statement in Military Strategy that the USSR has achieved a 100% defense, and the West has

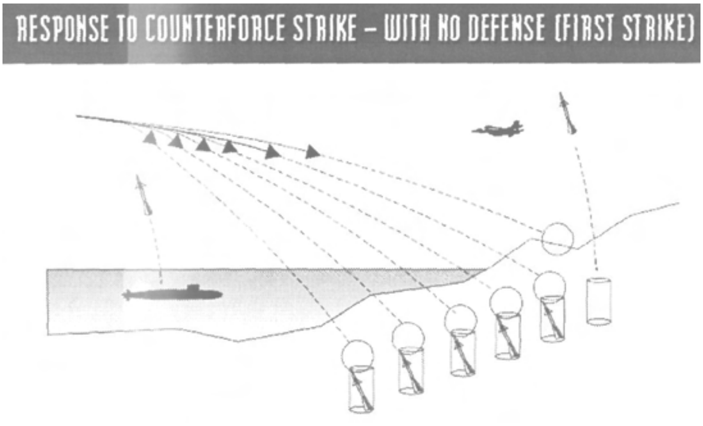

RESPONSE TO COUNTERFORCE STRIKE – WITH NO DEFENSE (FIRST STRIKE)

Slide 074: Response to Counterforce Strike with No Defenst

Here we show the orthodox concept of response to a surprise counterforce nuclear strike when the surprised force has no defenses.

A great percentage of the attacked missile silos are destroyed. The surprised side thus is able to mount only a limited strike in retaliation. Perhaps he is able to launch some remaining silo-based missiles that escaped destruction. He is also able to launch missiles from any of his underwater nuclear submarines that escaped. Some of his strategic bombers may have escaped destruction on the ground or while airborne on station; these surviving bombers can be directed to proceed with strategic reprisal against the enemy.

As can be appreciated, this entire concept is now totally obsoleted by the Soviet possession of scalar EM weapons.

The "mutual assured destruction" capability - or MAD concept as it was so aptly named! - is no longer mutual.

And for two reasons: (1) the Soviet surprise attack could destroy essentially all our Triad forces wherever they are, in the first attack. (2) Even if a very small number of our strategic forces survive, the puny strike we could attempt to launch could be repeatedly decimated from launch to terminal phase.

The restraint on the Soviets is not because of our Triad forces. The restraint is twofold: (1) a second nation has scalar EM weapons, and would undoubtedtly unleash them upon the Soviet Union with great violence if such a massive world-wide destruction of U.S. forces occurred or was threatened (is it really "accidental" that six major Soviet missile ammunition storage sites mysteriously exploded within seven months?); (2) the closed-loop scalar radiation exchange system of Earth-Sun could be inadvertently "tweaked" in the feedback loop from Earth to Sun, so that a large solar scalar resonance was stimulated. In that case the Sun could emit a mighty resonant "burp" that would simply wipe out all life on Earth in a fiery shower.

The "terror" in the "balance of terror" between nations has just increased by many, many orders of magnitude.

The biblical prophecy that the Earth will be wiped out by "fire and brimstone from heaven" is very near fulfillment now.

Put another way, the pucker factor - even on the Russians who may contemplate unleashing Armageddon - is higher than any Western strategic analyst has ever imagined, even in his wildest nightmares.

Truly these weapons are "more frightening than the mind of man has imagined," as Brezhnev put it in 1975.

The frightening scalar EM weapons can be used, but only very, very gingerly indeed. If a slight mistake is made, everybody loses everything.

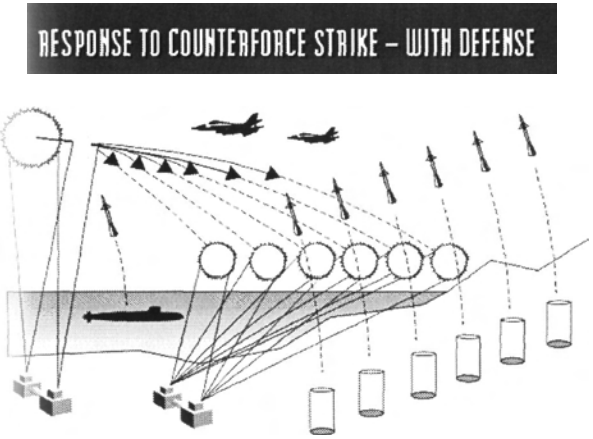

RESPONSE TO COUNTERFORCE STRIKE – WITH DEFENSE

Slide 075: Response to Counterforce Strike with Defense

On this slide we show the impact of scalar defenses against an adversary who possesses only the orthodox nuclear weapons and delivery systems, and makes a "first strike" by surprise against his opponent.

In this case, the attacked opponent has scalar EM defenses.

With just the midcourse and terminal scalar EM howitzer defenses, almost all of the attacking missiles and re-entry vehicles are destroyed.

If for some reason (such as to prevent triggering the sun and destroying the world inadvertenly!) the attacked adversary does not wish to attack strategically with scalar EM weapons, he may now launch a massive retaliatory strike on his opponent.

And if that opponent has no ABM defenses, he's just bought the farm.

What a funny coincidence! Long ago, we crazy Americans chose to put all **our** eggs in the offensive basket. That is, we thought the Soviets cannot dare to strike us because we will have enough surviving missiles to penetrate any orthodox defense and destroy them in return. They would also have the same "mutual assured destructon (MAD)" capability, of course!

The MAD concept was aptly acronymmed - it was an insane idea from its inception. That is, the Soviets were not supposed to defend their cities **and** populace, and neither were we. That way, you see, we could destroy their citizenry at will, and they could destroy ours at will. We deliberately agreed to that, thinking - in our insanity - that the Soviets agreed also.

Of course the Soviets weren't playing the same lunatic's game at all. They simply fed our own garbage back to us, and we importantly thought we were negotiating something. Unknown to us, they were secretly building and deploying scalar EM defenses even then. Since we didn't know of any other defense, we assumed the Soviets could not defend their citizenry with ABMs any better than we could defend ours with ABMs. So we blithely scrapped our ABMs.

The high government officials of the United States of America grossly violated our constitution and their sworn oaths of office. Our constitution specifically charges the government of the United States to defend the people of the United States. That means our officials are specifically charged with defending our people. Yet they deliberately agreed not to do so, and they deliberately sacrificed us in advance.

They committed treason of the highest kind, by deliberately surrendering us to total destruction anytime the big balloon goes up.

The constitution doesn't charge the government with surrendering our citizens to destruction under any circumstances. It charges the government with defending us against destruction.

Ronald Reagan is the first U.S. president in some time to see through the MAD concept. At least he stands forthrightly for defense. It's just a shame that the government and scientific leaders under him aren't aware of the Soviet scalar weapon developments, so they could brief him on these startling developments.

Be sure you understand what the government leaders before Reagan have been saving. For years, our official policy has effectively said to the Russians: "We will allow each side a token defense only. We deliberately will not defend ourselves, so that anytime you decide to strike us you can destroy us. In return, however, we do not think you will want to strike us, for we believe that - before we become ashes from that strike - we can get enough away at you, to destroy you as well. Of course we believe you are in the same sort of boat."

Then we said, "There! That's settled! What a relief! Nobody can strike anybody anymore. Now can we talk about partial disarmament and on-site inspection?"

The Soviets just kept their astonished amusement to themselves and kept building scalar defenses (and scalar offenses as well!). Their deception plan called for imbedding the scalar EM weapons in ordinary systems when possible, and hiding the scalar EM weapons tests in ordinary weapons tests when possible. Of course they had to be adamant in their opposition to on-site inspection; after all, even Western scientists, if allowed on site with the dual-mode weapons, might suspect something was different and possibly things were amiss.

The Soviet deception has worked like a charm. The U.S. MADmen who created and implemented the MAD doctrine have succeeded in prejudicing two generations of academicians and elected U.S. leaders against the need for effective defense. A vast wasteland of inane paper studies - learnedly dealing with ABM's, penetration tactics, leakage, force and counterforce, radar blackout, electronic warfare, etc. etc. - has been churned out by our military, industrial, and university "think tanks" to clog the minds, efforts, and literature of our defense and governmental establishments.

Our homeland has no effective ABM defense (and very little defense against aircraft or subsonic cruise missiles).

Our field army has no effective ABM defense (or air defense). Our naval forces have no effective ABM defense (although they are better defended against air attack than the field army). Yet ballistic missiles - and to some extent cruise missiles - have supposedly been our greatest threat for decades, followed closely by the secondary threat of massed aircraft attack.

A cynic might take the stand that, if our Defense Department hasn't been allowed to defend us (with deployed weapon systems, not by "paper studies") against what we ourselves have proclaimed is our biggest threat, then why is it called the Defense Department? Why not call it the No-defense Department'?

Understand, a great many fine, hard-working people in DoD have certainly tried to get adequate defenses developed and deployed. The sheer weight of the bureaucracy and its mind-numbing entanglements has prevented it.

But it's even worse. Our bureaucrats - both in government and in "big, organized science" - have been so caught up in the "Russians are really just ignorant peasants" syndrome and the "not-invented-here" (NIH) syndrome **that** they've not recognized the greatest scientific and technical breakthrough in history. Such a breakthrough - far greater in importance that the nuclear bomb has been made, developed, and deployed by the Soviet Union, all in secrecy and behind the smokescreen of a masterful deception plan.

We've totally missed the most incredible "secret superweapons" development program of all times, and the worldwide testing of the weapons themselves.

Hiroshima and Nagasaki should serve as object lessons in the price of ignorance of such technological breakthroughs.

History does repeat itself. But this time it's not we Americans who got the growl new superweapon first.

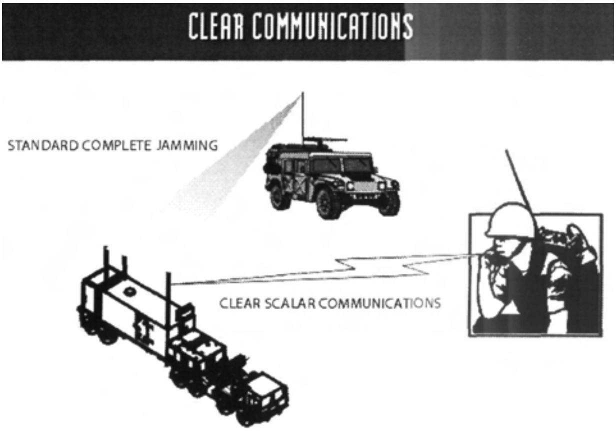

CLEAR COMMUNICATIONS

Slide 076: Clear Communications

Now let us rapidly give some unique capabilities that one can do with scalar EM in smaller things. Not only can scalar EM weapons be used in the big "continent burning" strategic weapons, but they can also be used for a variety of smaller defense tasks.

For example, one can establish a zero reference anywhere one wishes in an ordinary carrier. Inside that zero, one can establish as many independent "carriers" as one wishes, and hide them completely from ordinary detectors.

The way to do that is easy. Take as complex a signal as you wish. Subtract it from zero. That gives a "complement" which, when summed with the original signal, will produce the zero. Then only a scalar detector can go inside the zero and pull out the individual components.

So take a carrier wave. Let the enemy jam it completely, with gobs of power.

Your scalar communication channel stays serenely there, calm and unjammed. The enemy thinks you're jammed and can't communicate. Meanwhile, you continue to communicate just fine.

Or be real smart. Take one of those components inside the first zero. Pick a zero reference inside it. Hide still more carrier components inside the second zero. Do this as many times as you wish.

You can put all the telephone calls in the United States of America on one carrier. And you can keep them clear, right through total jamming.

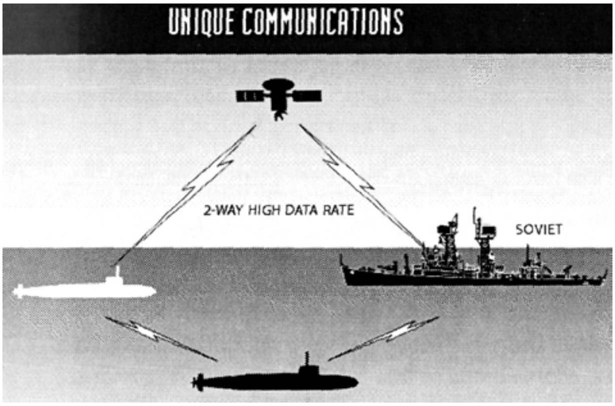

Slide 077: Unique Communications

But that's still not all.

You can also communicate with your submarines underwater. High data rate. Chat away as you wish. You can scramble, frequency hop, encrypt, the works. They can also communicate back to you.

Remember. EM force fields of any appreciable frequency won't travel through the water. Natural EM potentials grab charged particles and ions, so they won't travel through the ocean either. But artificial potentials don't grab onto charged particles or ions, and they will go through the ocean quite nicely.

Everything - aircraft, ships, shore installations, submarines, etc. - can communicate like gangbusters if they possess scalar EM technology.

In fact, you can even build scalar EM underwater radar if you wish.

The airborne scalar EM radar can see under the ocean with a little adjustment. perhaps as well as a conventional side-looking radar sees distant targets today, The airborne scalar EM radar can detect and track the underwater submarine while tooling along.

In fact, the airborne scalar EM radar can then destroy the sub, by using a **little** scalar interferometry. And it's even possible to do scalar interferometry with a single transmitter! How, is proprietary to a friend.

In fact, you can even use your enemy's jamming transmission as a "wire" along which to establish your scalar EM channels. You can use these scalar channels In the "weakly endothermic mode" to locate his transmitter and "lock-in" on it. You can "walk" the lock-in right past his encryption function to the "clear" Input, and receive and record that in the clear. All this from a distance.

You can then destroy his equipment if you wish, or just continue to monitor him in the clear if it's important to do so. Perhaps now we can understand what the decades-long weak microwave radiation of the U.S. Embassy in Moscow is all about. Or why the Soviets are building a nice new Embassy, bristling with antennas, on the high ground in Washington, D.C.

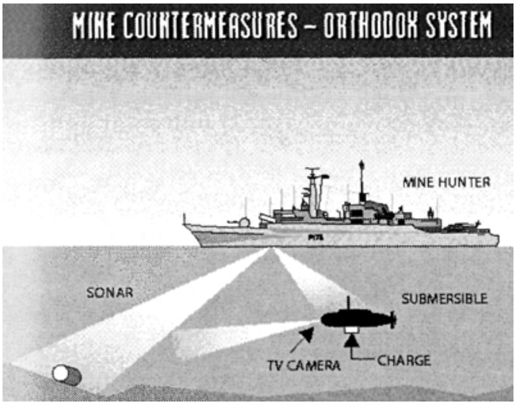

MINE COUNTERMEASURES – ORTHODOX SYSTEM

Slide 078: Mine Countermeasures—Orthodox System

Take something like mine countermeasures, for example. Finding and destroying mines - particularly those placed on the bottom in relatively shallow areas such as ports and rivers - is a really tough problem.

Presently it's often done as portrayed on this slide.

Here a minehunter ship uses an underwater sonar to locate the mine, or what resembles a mine. Then it directs a submersible robot vehicle to proceed over to the object and examine it with an underwater television camera. If it's a mine, the submersible places a remotely operated charge beside it, then chugs away. When it reaches a safe distance away, the charge is exploded, causing the mine to detonate.

A very painful process. Very time consuming, and very difficult.

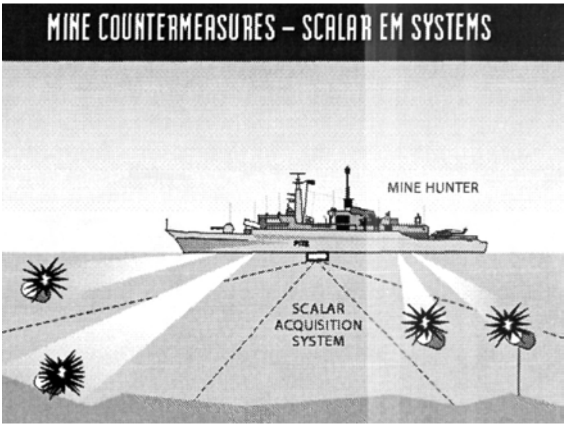

MINE COUNTERMEASURES – SCALAR EM SYSTEMS

Slide 079: Mine Countermeasures—Scalar EM Systems

Now try the same thing while using scalar EM underwater systems. The scalar EM acquisition system is really a sophisticated scalar radar. It scans a large area rapidly, locating moored mines, bottom-planted mines, and drifting mines.

The acquisition system is also a scanned "underwater eye." The operator in the surface vessel can tell whether or not it's a mine he's located.

Once the target mine is detected, a computer computes the firing parameters for a scalar interferometry pulse to hit it and destroy it. The operator fires the pulse, destroying the mine.

The same system that detects the mine at a distance identifies and destroys it at a distance. In this way large areas can be quickly scanned and swept clean of mines.

If a suspicious object is detected and one is still not quite sure it's a mine, shoot it anyway! If it's a mine, it explodes. If it's just a rock, nothing happens. So the same system can do "discrimination by fire."

The same system is useful to destroy midget submarines, bottom crawlers, etc.

It can also wipe out whole arrays of underwater sensors by electromagnetically dudding them.

It can take on a sub if it has to! Here it would just gently kindle EM energy inside the sub's controls, causing it to lose control of all its systems. The sub would then sink to crush depth and implode. The scalar EM mine sweeping system could, of course, simply "blow" the sub, but might itself get caught in the resulting explosion.

SCALAR ANTI-PERSONNEL WEAPON

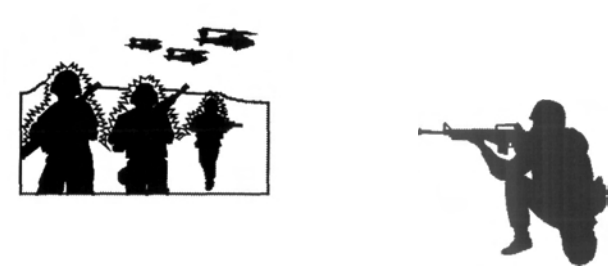

Slide 080: Scalar Anti-Personnel Weapon

From a military viewpoint, these scalar EM weapons are really quite all-around weapons.

For example, they are very lethal against personnel.

Set in the "high intensity pulse mode," one blast and a person's nervous system is destroyed instantly. Total, instantaneous death results. One can blast away at entire groups of infantrymen, for example. It's the peak power that's important, not the average power. So the weapon doesn't require too much battery power.

Set on a lower intensity pulse mode, one zap can simply knock out a person, rendering him instantly unconscious.

Quite useful if one needs prisoners to interrogate. Also very silent, which is useful in operations employing stealth and surprise.

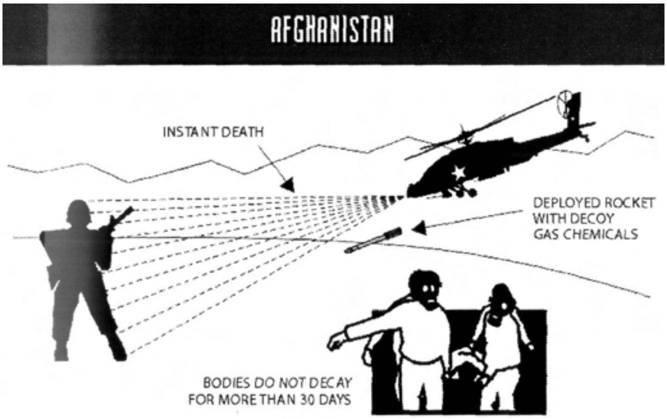
AFGHANISTAN
INSTANT DEATH
DEPLOYED ROCKET WITH DECOY GAS CHEMICALS
BODIES DO NOT DECAY FOR MORE THAN 30 DAYS

Slide 081: Afghanistan

The Soviets are employing just such a scalar "death ray" weapon in Afghanistan, apparently in the noses of some of their HIND helicopters.

These choppers usually fire gas rockets to camouflage the testing of the tactical scalar death ray. For example, they may fire rockets filled with nerve gas, and some persons will die convulsively from the chemical.

Those hit by the scalar EM weapon, however, have a most peculiar death **mode,**

Death comes instantly and totally. There is no convulsion, no response. The entire nervous system is destroyed instantly. Every living cell in the body is killed instantly, including all bacteria, germs, etc.

A body hit with this thing falls like a limp rag and lies where it falls. It doesn't decay in even 30-45 days. In a macabre fashion, it's been reduced to something like food irradiated with nuclear radiation; everything is killed, so the material is preserved for an extended period before any decay can set in.

Western military authorities have speculated that this strange death is due to some sort of eery new gas.

However, even with the most virulent gas, some seconds are required for a b**ody** to die. And while it's dying it goes into convulsions. Violent convulsions.

This eery, instant death that leaves a non-decaying corpse is not due to a **gun,** This is the signature of a high-powered pulse kill using a scalar EM weapon

The associated gas rockets are the disguise used to fool our intelligence analysts.

They're just part of the Soviet deception plan.

ADDITIONAL NOTES AND REFERENCES

15. For details of the mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," Jane's Defence Weekly, 1(13), Apr. 7, 1984, p. 508

SCALAR ANTI-TANK WEAPON

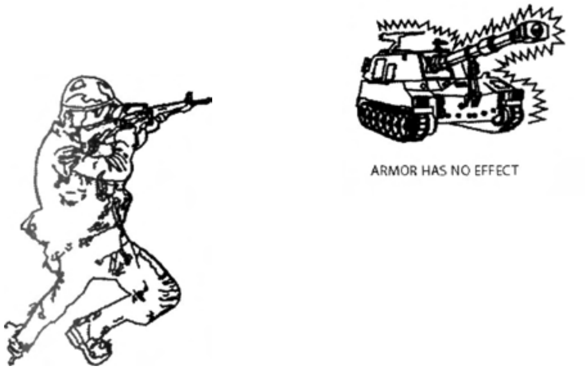

Slide 082: Scalar Anti-Tank Weapon

Such weapons are also effective against armored vehicles, for example.

They can be carried by personnel, mounted on vehicles, mounted in helicopters, etc.

They are truly all-purpose weapons.

An infantryman with a bazooka-sized scalar EM pulse weapon can easily knock out a tank with one shot. The armor has no effect; scalar EM goes right through conductive shielding and Faraday cages.

And with the same weapon he can knock out another tank, and another, and another...

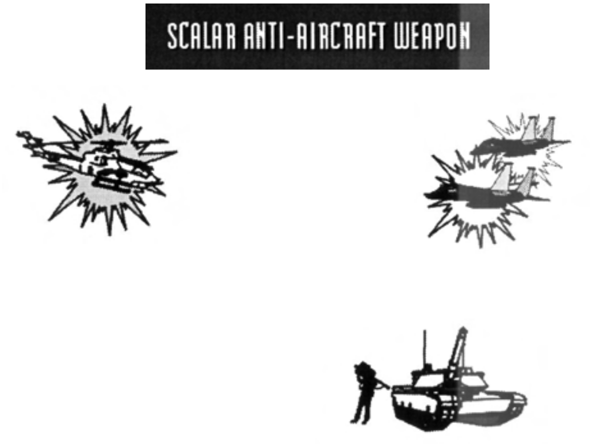

SCALAR ANTI-AIRCRAFT WEAPON

Slide 083: Scalar Anti-Aircraft Weapon

And, if a helicopter or a fixed-wing aircraft attacks him, he can bring the same weapon to bear on it.

He can knock down a chopper or a jet. With one shot.

And another. And another.

Also, it's a line-of-sight weapon. He doesn't have to compute lead angle and all that jazz. Just aim, get the target in his sights, and pull the trigger.

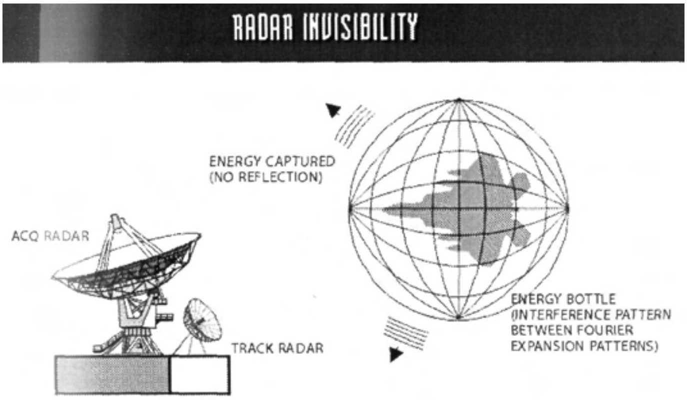

Slide 084: Scalar Anti-Aircraft Weapon

Another neat trick you can do with scalar EM is make an aircraft invisible **to** ordinary radar.

Just use multiple transmitters on the airplane to establish a spherical interference shell, in the bandwidth of the threatening radar, around the aircraft.

When the signal pulse (or continuous wave) from the ground-based radar **strikes** the "energy bottle" in the shell, it interferes with it and is diffracted, scattered, and partially absorbed.

The ground-based radar doesn't get any coherent return pulse or return signal reflected from the target.

Radars don't track aircraft; they track the reflection of their own signal bouncing off aircraft. No reflected signal, no aircraft - as far as a radar is concerned.

To shield against a passive radar (one which tracks emissions from the aircraft), just detect and complement your own emissions so that "near zeroes" result In the spectrum of interest.

Or do both, to protect against both.

It's even possible to cool the exhaust and heated surfaces of the aircraft by a scalar EM adaptation of electrostatic cooling. Theoretically it's even possible to cause the aircraft to "disappear" optically. Takes a bit of doing, but it's possible The basic scheme here is that, when a photon hits the target from a given direction, it's absorbed totally, and a corresponding or "matching" photon is emitted from the opposite side in the same direction the original photon was going. In that fashion, the light emitted from the target from one direction seems to have "passed through it" from the other direction.

Since that's what empty space does, then what a distant observer sees is "empty space," so far as he can tell.

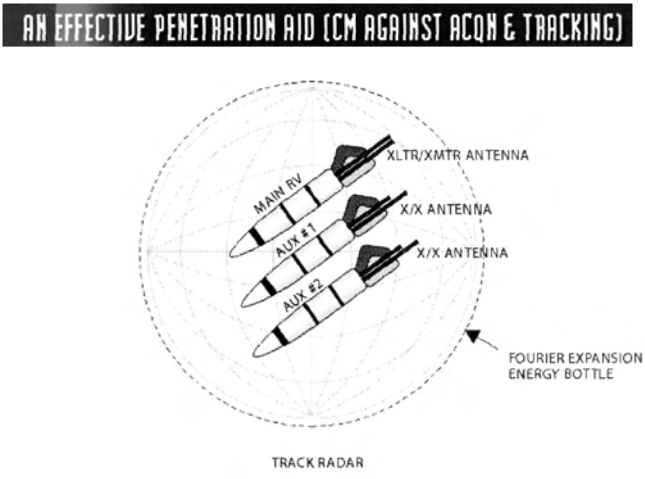

AN EFFECTIVE PENETRATION AID (CM AGAINST ACQN & TRACKING)

XLTR/XMTR ANTENNA
MAIN RV
X/X ANTENNA
AUX #1
X/X ANTENNA
AUX #2
FOURIER EXPANSION
ENERGY BOTTLE
TRACK RADAR

Slide 085: An Effective Penetration Aid

Of course you can do the same thing for re-entry vehicles.

Use some multiple antennas on a vehicle. Or better yet, use satellite RV's to carry the extra transmitters necessary for scalar interferometry.

Deploy the interferometer RV's and activate the scalar EM transmitters. Place an energy bottle Fourier expansion interference shell around the ensemble in the bandwidth of interest.

Radars in that bandwidth now will be unable to detect the incoming ensemble. If nothing detects the threat, nothing is going to shoot at it or react to it.

This can also be used on a bus vehicle during midcourse, for example.

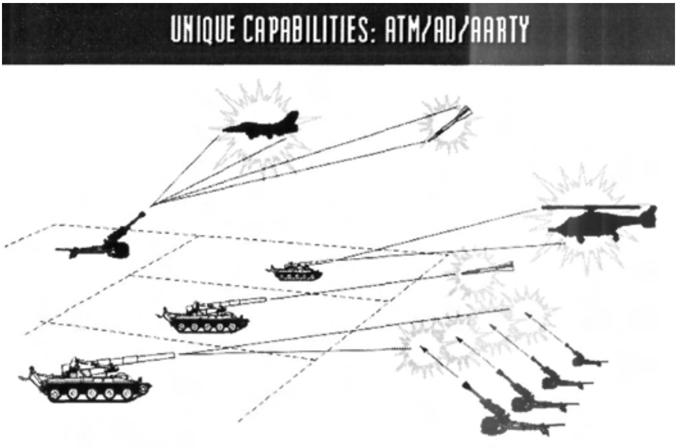

UNIQUE CAPABILITIES: ATM/AD/AARTY

Slide 086: Unique Capabilities: ATM/AD/AARTY

But the uniqueness of scalar EM devices and weapons doesn't end there.

You can hit lots of other threats that are presently difficult or impossible to effectively handle.

For example, on the tactical battlefield, the same scalar EM radar interferometer system can defend against tactical missiles (ATM role), aircraft (AD role), cruise missiles (ACM role), and even artillery and mortar shells (AARTY role).

How the devil, you might ask, can it defend against artillery and mortar shells?

Simple. Sweep the artillery or mortar barrage with closely spaced pulses. Sorta like a machinegun. Or more like a phased-array radar.

If the peak power in the sweeping pulses is sufficient, the explosives are sparked and detonated. Fuses are destroyed.

So even mortar and artillery rounds can be destroyed, by the same system providing air defense and anti-tactical missile defense.

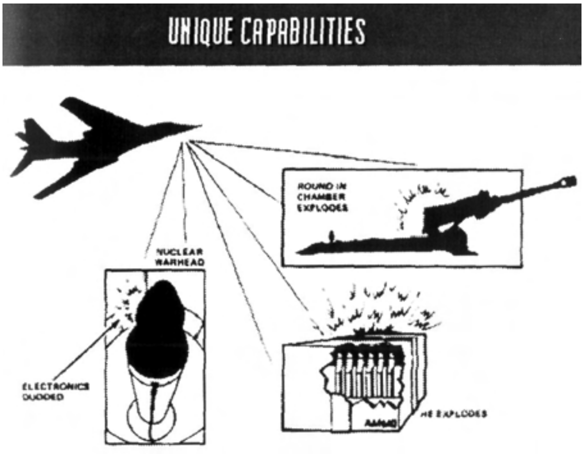
UNIQUE CAPABILITIES

Slide 087: Unique Capabilities

That's still not all.

An incredible number of different types of targets can be attacked with scalar EM weapons.

For example, aircraft equipped with scalar EM weapons can attack targets such as loaded artillery pieces. The round in the chamber explodes, destroying the weapon.

It can also attack nearby stored ammunition. Again, the high explosive in the ammunition explodes, destroying the store.

Even underground missiles can be attacked. The scalar pulse will go right through the overhead cover and dud the electronics, including the nuclear warhead. It will also detonate the HE in the warhead and detonate the fuels and propellants.

A great variety of other targets can also be attacked, as we shall see.

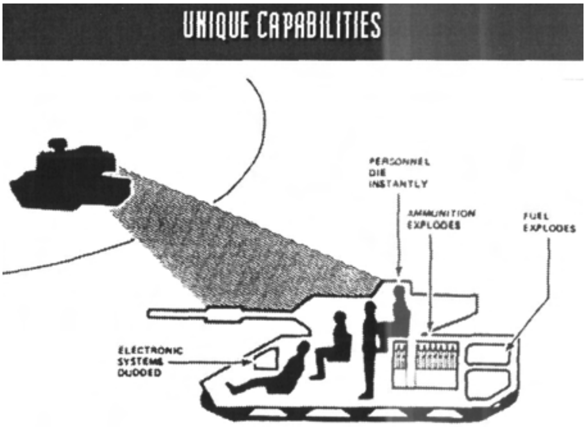

UNIQUE CAPABILITIES

PERSONNEL
DIE
INSTANTLY
AMMUNITION
EXPLODES
FUEL
EXPLODES
ELECTRONIC
SYSTEMS
DUDDED

Slide 088: Unique Capabilities

For example, suppose one attacks a tank.

With scalar EM pulse, the personnel die instantly. Total personnel kill is achieved.

All electronic systems of the tank are dudded. Total systems kill is achieved.

The ammunition in the tank explodes. Total kill is achieved.

The fuel explodes. This is another total kill mechanism.

That tank has been killed totally, by a variety of mechanisms, all simultaneously.

That's a k-kill in any analyst's book.

And **don't** worry about retrieval and repair. That tank is finished permanently.

Even **the** most modern tank is just as vulnerable as the most obsolete.

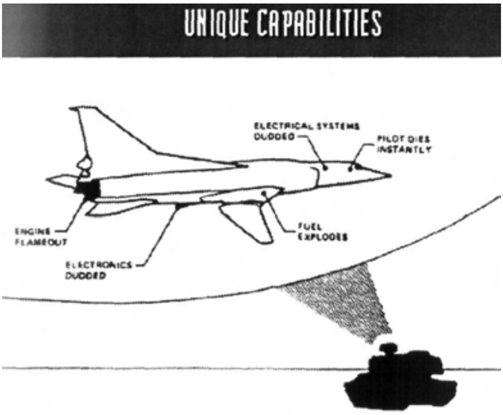

UNIQUE CAPABILITIES

ELECTRICAL SYSTEMS DUDDED
PILOT DIES INSTANTLY
ENGINE FLAMEOUT
FUEL EXPLODES
ELECTRONICS DUDDED

Slide 089: Unique Capabilities

Suppose one attacks a fighter-bomber aircraft.

Again, the kill is catastrophic from multiple simultaneous kill mechanisms.

The pilot and crew die instantly. Total destruction of their nervous systems.

The electrical systems are dudded.

The electronics in all avionics and ordnance is dudded.

The fuel explodes.

HE in on-board or out-board ordnance explodes.

The jet engine flames out if the pulse is powerful enough.

That aircraft is finished. Permanently.

And it's point-and-shoot, line-of-sight.

Even the most modern aircraft is just as vulnerable as the most obsolete.

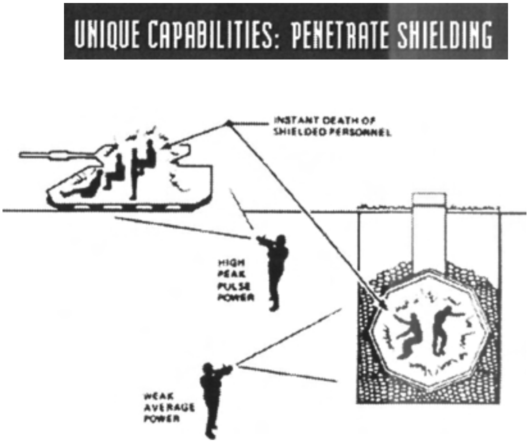
UNIQUE CAPABILITIES: PENETRATE SHIELDING

INSTANT DEATH OF SHIELDED PERSONNEL
HIGH PEAK PULSE POWER
WEAK AVERAGE POWER

Slide 090: Unique Capabilities: Penetrate Shielding

Shielding is ineffective.

The scalar pulse (or continuous wave) can go anywhere gravity can go.

It penetrates the steel hull of tanks and armored vehicles.

It penetrates underground shelters and bunkers.

It penetrates pillboxes and fortifications.

Foxholes are no protection. Just shoot through the earth into them.

There is no longer any place to hide.

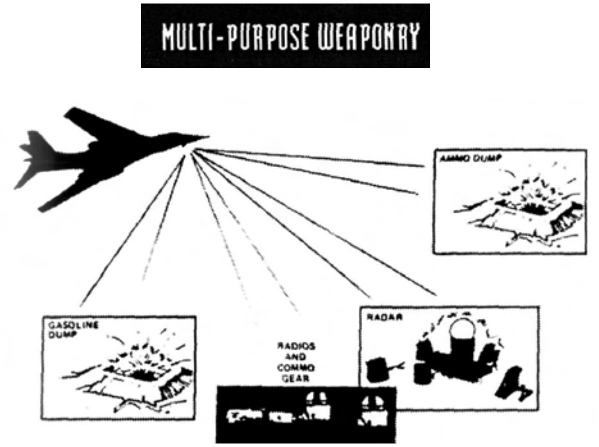

MULTI-PURPOSE WEAPONRY

AMMO DUMP
GASOLINE DUMP
RADIOS AND COMMO GEAR
RADAR

Slide 091: Multi-Purpose Weaponry

These scalar EM weapons are all-purpose. Things such as ammunition dumps, radar systems, radios and communication gear, communications centers, gasoline dumps, etc. can easily be destroyed.

Not only that, but the weapons can be swept, giving area coverage.

You can hit both point targets and area targets.

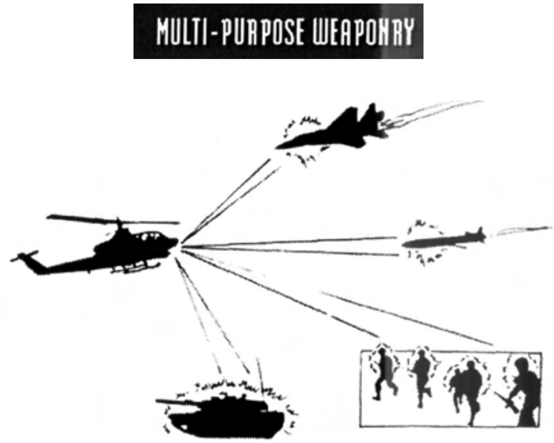

MULTI-PURPOSE WEAPONRY

Slide 092: Multi-Purpose Weaponry

Look at what a weapon a helicopter becomes.

The same vehicle, equipped with scalar EM weapons, can engage and destroy jet aircraft, other helicopters, cruise missiles, personnel, and armored vehicles.

It can also hit all the other targets we've been talking about. Any vehicle equipped with scalar EM weapons becomes a formidable destroyer on the battlefield.

And "reconnaissance by fire" takes on a particularly lethal meaning.

For example, don't worry about the enemy in cover and concealment, or highly camouflaged.

Don't worry about decoys or dummy positions.

Just sweep the area with scalar EM fire. The actual targets will be discriminated by fire and destroyed.

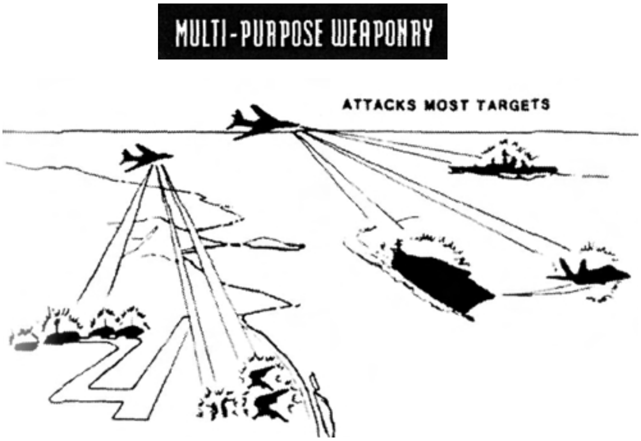

MULTI-PURPOSE WEAPONRY

ATTACKS MOST TARGETS

Slide 093: Multi-Purpose Weaponry

As can be seen on this slide, most targets of interest can be attacked.

A single aircraft can sink an aircraft carrier or destroy a battle cruiser from afar.

It can also destroy aircraft, ground installations, artillery, personnel, nuclear weapons storage sites, tanks, vehicles, fuels, and combustibles.

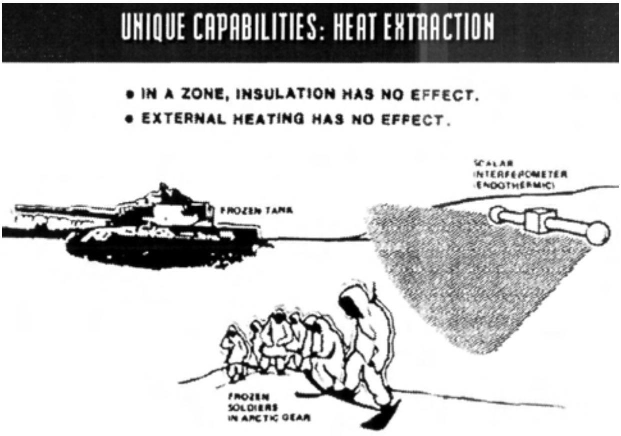

UNIQUE CAPABILITIES: HEAT EXTRACTION

- IN A ZONE, INSULATION HAS NO EFFECT.
- EXTERNAL HEATING HAS NO EFFECT.

SCALAR
INTERFEROMETER
(ENDOTHERMIC)
FROZEN TANK
FROZEN
SOLDIERS
IN ARCTIC GEAR

Slide 094: Unique Capabilities: Heat Extraction

Or one can get truly exotic.

To prevent destroying equipment and installations, the endothermic mode can be utilized.

Cold explosions can be used to freeze tanks, personnel, and equipment.

The equipment and tanks thaw out.

The personnel thaw out too, but they are dead when they do. Heaters and arctic clothing are no protection.

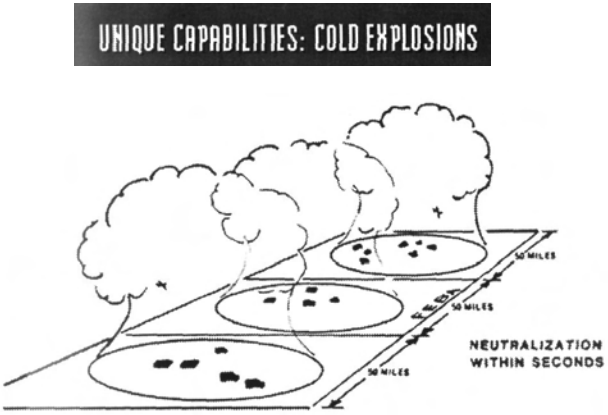

UNIQUE CAPABILITIES: COLD EXPLOSIONS

Slide 095: Unique Capabilities: Cold Explosions

For example, suppose three cold explosions, each 50 miles in diameter, are placed upon the front-line strength of NATO.

In seconds there is no longer any "front line strength" of NATO. Every living thing is frozen into a block of ice.

A few more of these things placed on selected rear area installations, and the battle is over before it gets started. Say in about one minute flat!

Suppose NATO is quickly defeated - say in two days. And then the Russians offer us a "Dunkirk." They agree to allow us to evacuate the beaten remnant of our forces.

Are we going to evoke a strategic nuclear exchange, or are we effectively "decoupled" from Europe? Particularly if the Soviets also offer sweeping accommodations to the Europeans, and they accept? If we punch the strategic button, we guarantee our own absolute destruction. (MAD doctrine and all **that,** old chap!)

Whether or not we could do anything to the Russians is a matter of conjecture.

For a demonstration, they could even invite us to launch a couple of missiles, and destroy them over our heads shortly after launch.

And they could take out a few submarines under the ocean, just to make it believable.

And destroy some SAC bombers that are orbiting on station.

Or take out a U.S. Naval Task Force, with a nuclear carrier, as a demonstration that they mean business.

Would a sane President go ahead and destroy us, or accept the Soviet's "Monroe Doctrine" and an "armistice"?

If the Soviets thought he was seriously contemplating ordering a nuclear strike, they could immediately destroy all our nuclear armaments where they sit. Our own weapons would obliterate us!

What would Congress do - after the Soviet Union had demonstrated its absolute power, offered a Dunkirk and armistice, and threatened our total destruction if we "reached for the holster," so to speak - when it realized the President might actually be contemplating evoking assured national suicide?

What would you do if you were a Congressman? If you were the President?

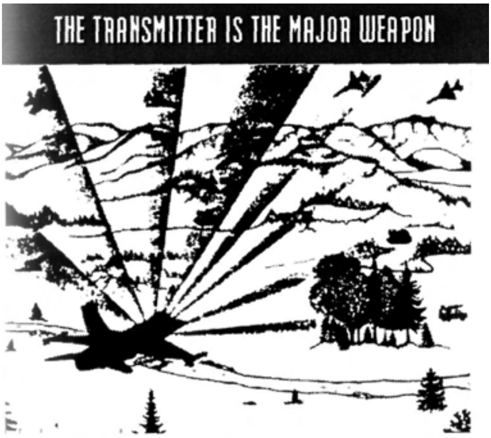
THE TRANSMITTER IS THE MAJOR WEAPON

Slide 096: The Transmitter is the Major Weapon

As shown on this slide, with scalar EM the transmitter becomes the major weapon on the battlefield.

Almost every type of target of any importance can be easily and effectively attacked and destroyed or nullified by scalar EM weapons.

Small wonder the Soviets do not discard their thousands of "old" radars!

How effective do you suppose some of their "old" transmitters on their "old" aircraft might be?

What chance do you think SAC bombers would have against thousands of "old" radar-directed gun sites and "obsolete" air defense missile sites?

Could it be that even something like the old SA-1 could be startlingly effective?

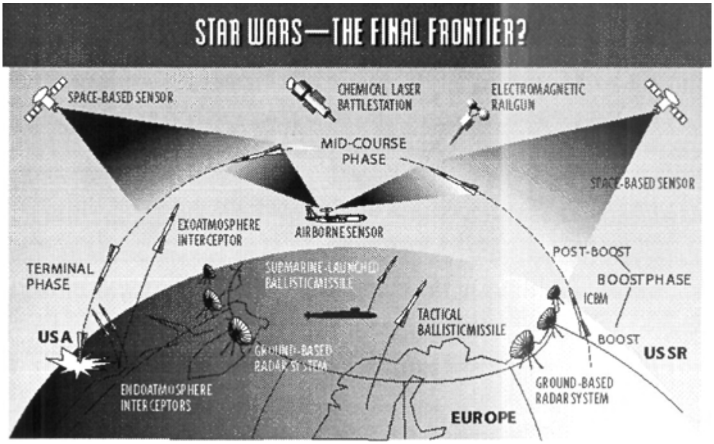
STAR WARS—THE FINAL FRONTIER?
SPACE-BASED SENSOR
CHEMICAL LASER BATTLESTATION
ELECTROMAGNETIC RAILGUN
MID-COURSE PHASE
SPACE-BASED SENSOR
EXOATMOSPHERE INTERCEPTOR
AIRBORNE SENSOR
POST-BOOST
TERMINAL PHASE
SUBMARINE-LAUNCHED BALLISTIC MISSILE
BOOST PHASE
ICBM
TACTICAL BALLISTIC MISSILE
BOOST
USA
USSR
GROUND-BASED RADAR SYSTEM
ENDOATMOSPHERE INTERCEPTORS
GROUND-BASED RADAR SYSTEM
EUROPE

Slide 097: Star Wars—The Final Frontier?

But what of our advanced concepts, such as the strategic defense initiative?

Here we show some of the expensive and complex components of that powerful system.

We have ground-based radars, endoatmospheric interceptors, exoatmospheric interceptors, airborne sensors, space-based sensors, chemically propelled interceptors, electromagnetic railguns, particle beam weapons, high energy lasers, etc. all tied together by a massive command and control system. This massive, expensive system is designed to detect and destroy missiles, including ICBMs, IRBMs, cruise missiles, SLBNs, etc.

*The s*ystem detects the missile launches and engages them from launch on in to the terminal phase.

Massive systems. Electromagnetic marvels. Nuclear-pumped lasers of staggering power. Perhaps even x-ray and gamma ray lasers. Directed energy RF weapons. Stupendous railguns. Sensors everywhere. Giant webs of communications. Banks of computers and control systems.

And it's totally vulnerable to scalar EM interferometer weapons. Giant scalar EM "radars" can simply sweep through all that like a scythe through standing hay.

As it stands, it's obsolete before its design is even finished.

It's obsolete to Soviet scalar EM weapons that are already deployed and operationally tested in place!

Without scalar EM weapons and defenses, all our present weapons and concepts are deadly vulnerable.

The war is over before it hardly gets started. It's not a war; it's more like a clay pigeon shoot.

Understand, if orthodox science was all the Russians possessed, the Strategic Defense Initiative would be the way to go.

But they do possess something else, and presently they have all the trump cards in the deck.

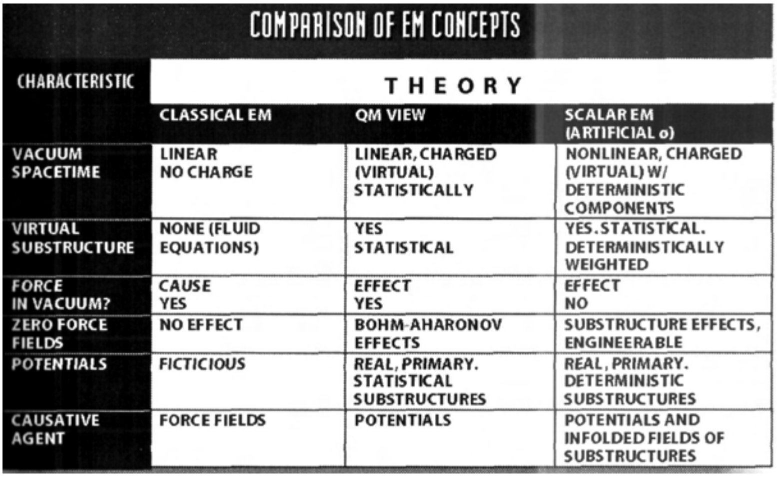

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL o)
VACUUM SPACETIME	LINEAR NO CHARGE	LINEAR, CHARGED (VIRTUAL) STATISTICALLY	NONLINEAR, CHARGED (VIRTUAL) W/ DETERMINISTIC COMPONENTS
VIRTUAL SUBSTRUCTURE	NONE (FLUID EQUATIONS)	YES STATISTICAL	YES. STATISTICAL. DETERMINISTICALLY WEIGHTED
FORCE IN VACUUM?	CAUSE YES	EFFECT YES	EFFECT NO
ZERO FORCE FIELDS	NO EFFECT	BOHM-AHARONOV EFFECTS	SUBSTRUCTURE EFFECTS, ENGINEERABLE
POTENTIALS	FICTICIOUS	REAL, PRIMARY. STATISTICAL SUBSTRUCTURES	REAL, PRIMARY. DETERMINISTIC SUBSTRUCTURES
CAUSATIVE AGENT	FORCE FIELDS	POTENTIALS	POTENTIALS AND INFOLDED FIELDS OF SUBSTRUCTURES

Slide 098: Comparison of EM Concepts (Chart)

On this slide and the three that follow, we show a comparison of some of the major characteristics of the three kinds of electromagnetics.

We compare the classical, the quantum mechanical, and the scalar EM (electrogravitational) views of electromagnetics

Notice the differences between the three in regards to vacuum spacetime, virtual substructure, force, whether or not force exists in vacuum, the effect of zero force fields, potentials, and what is to be regarded as causative agents.

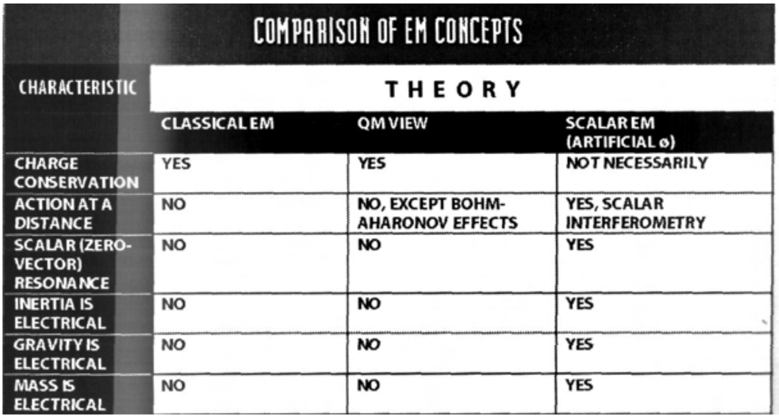

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ø)
CHARGE CONSERVATION	YES	YES	NOT NECESSARILY
ACTION AT A DISTANCE	NO	NO, EXCEPT BOHM-AHARONOV EFFECTS	YES, SCALAR INTERFEROMETRY
SCALAR (ZERO-VECTOR) RESONANCE	NO	NO	YES
INERTIA IS ELECTRICAL	NO	NO	YES
GRAVITY IS ELECTRICAL	NO	NO	YES
MASS IS ELECTRICAL	NO	NO	YES

Slide 099: Comparison of EM Concepts (Chart)

Now note the differing views of charge conservation, action at a distance, scalar resonance, and whether inertia, gravity, and mass are electrical.

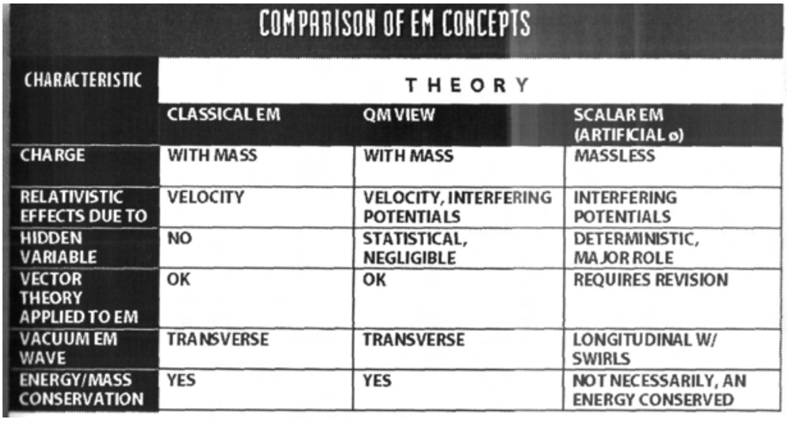

COMPARISON OF EM CONCEPTS

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ø)
CHARGE	WITH MASS	WITH MASS	MASSLESS
RELATIVISTIC EFFECTS DUE TO	VELOCITY	VELOCITY, INTERFERING POTENTIALS	INTERFERING POTENTIALS
HIDDEN VARIABLE	NO	STATISTICAL, NEGLIGIBLE	DETERMINISTIC, MAJOR ROLE
VECTOR THEORY APPLIED TO EM	OK	OK	REQUIRES REVISION
VACUUM EM WAVE	TRANSVERSE	TRANSVERSE	LONGITUDINAL W/ SWIRLS
ENERGY/MASS CONSERVATION	YES	YES	NOT NECESSARILY, AN ENERGY CONSERVED

Slide 100: Comparison of EM Concepts (Chart)

Continuing, note the differences in the views of charge, what relativistic effects are due to, whether or not one has hidden variables and whether they are enginerrable, how vector theory is applied to electromagnetics, the nature of the EM wave in vacuum, and conservation of energy/mass.

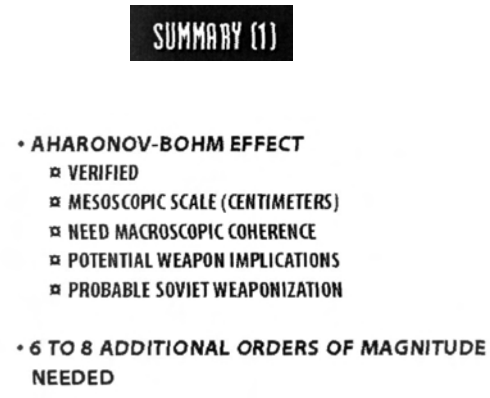

SUMMARY (1)

- AHARONOV-BOHM EFFECT
  - ¤ VERIFIED
  - ¤ MESOSCOPIC SCALE (CENTIMETERS)
  - ¤ NEED MACROSCOPIC COHERENCE
  - ¤ POTENTIAL WEAPON IMPLICATIONS
  - ¤ PROBABLE SOVIET WEAPONIZATION
- 6 TO 8 ADDITIONAL ORDERS OF MAGNITUDE NEEDED

Slide 101: Summary (1)

Now we examine the three views for whether or not electro-gravitational waves exist and what they interact with, the number of required spacetime dimensions the type of geometry, the nature of a zero-summed EM vector, local and distant relativity aspects, and whether or not the probability states propagated by the Schroedinger equation can be deterministically engineered.

As can be seen from the last four slides, the three views are quite different, and the scalar EM view is drastically different.

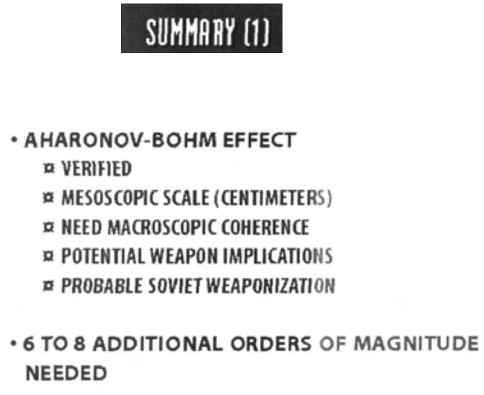

SUMMARY (1)

- AHARONOV-BOHM EFFECT
  - ¤ VERIFIED
  - ¤ MESOSCOPIC SCALE (CENTIMETERS)
  - ¤ NEED MACROSCOPIC COHERENCE
  - ¤ POTENTIAL WEAPON IMPLICATIONS
  - ¤ PROBABLE SOVIET WEAPONIZATION
- 6 TO 8 ADDITIONAL ORDERS OF MAGNITUDE NEEDED

Slide 102: Summary (1)

In summary, we can now state unequivocally that the Aharonov-Bohm effect has been conclusively proven, at least for the mesoscopic scale over a few thousand Angstroms, and possibly over a few centimeters.

But in these experimental proofs; the "natural" type of potential has been used, it is basically incoherent in its substructure, dooming the effect to stay small.

By utilizing macroscopic EM force fields and waves to zero-sum and comprise artificial potentials, the coherence can be maintained over a macroscopic distance - even for hundreds of thousands of kilometers.

The weapons implications - only some of which we have set forth in this treatise - are mindboggling.

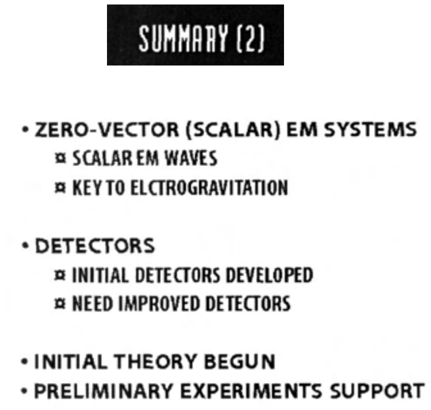

SUMMARY (2)

- ZERO-VECTOR (SCALAR) EM SYSTEMS
  - ¤ SCALAR EM WAVES
  - ¤ KEY TO ELCTROGRAVITATION
- DETECTORS
  - ¤ INITIAL DETECTORS DEVELOPED
  - ¤ NEED IMPROVED DETECTORS
- INITIAL THEORY BEGUN
- PRELIMINARY EXPERIMENTS SUPPORT

Slide 103: Summary (2)

Zero-vector (scalar) EM systems are the key to scalar EM waves, which are actually electrogravitational waves.
Scalar wave detectors nave been developed and are being improved. Further improvement is needed. This development is limited only by funds.

The initial theory has been begun. More work is needed. Again, this development is limited only by funds.

Preliminary experiments support the scalar EM concepts.

Equipment can be designed and made to work electro-gravitationally.

And it has been.

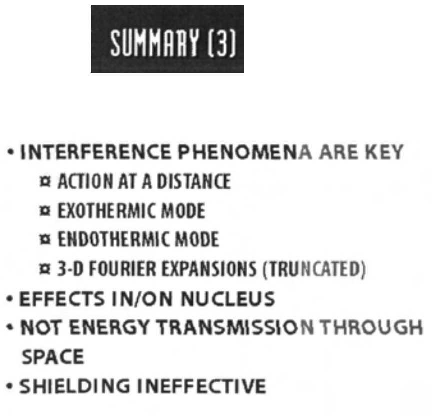

SUMMARY (3)

- INTERFERENCE PHENOMENA ARE KEY
  - ACTION AT A DISTANCE
  - EXOTHERMIC MODE
  - ENDOTHERMIC MODE
  - 3-D FOURIER EXPANSIONS (TRUNCATED)
- EFFECTS IN/ON NUCLEUS
- NOT ENERGY TRANSMISSION THROUGH SPACE
- SHIELDING INEFFECTIVE

Slide 104: Summary (3)

Interference phenomena are key.

One can get action at a distance - even over hundreds of thousands of kilometers.

One can engineer gravitational and inertial effects.

One can engineer the nucleus, including transmute it, easily and cheaply. One could clean up all the nuclear wastes.

Electromagnetic energy can be produced at a distant target, or extracted from a distant **target.**

This is is not energy transmission through space in the form of EM force fields. Instead, it is transmission through spacetime in the form of electrogravitational potentials.

Conventional EM shielding is ineffective against scalar EM.

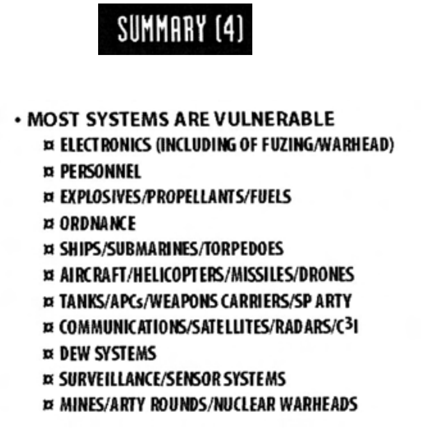

SUMMARY (4)

- MOST SYSTEMS ARE VULNERABLE
  - ELECTRONICS (INCLUDING OF FUZING/WARHEAD)
  - PERSONNEL
  - EXPLOSIVES/PROPELLANTS/FUELS
  - ORDNANCE
  - SHIPS/SUBMARINES/TORPEDOES
  - AIRCRAFT/HELICOPTERS/MISSILES/DRONES
  - TANKS/APCs/WEAPONS CARRIERS/SP ARTY
  - COMMUNICATIONS/SATELLITES/RADARS/C^3I
  - DEW SYSTEMS
  - SURVEILLANCE/SENSOR SYSTEMS
  - MINES/ARTY ROUNDS/NUCLEAR WARHEADS

Slide 105: Summary (4)

Almost every weapon system we presently have - or are developing - is totally vulnerable to scalar EM weaponry.

This includes personnel, electronics (including fuzing and warhead), explosives, propellants, fuels, ordnance, ships, submarines, torpedoes, aircraft, helicopters, missiles, drones, rockets, tanks, armored vehicles, weapons carriers, self-propelled and towed artillery, communications, satellites, radars, command and control, directed energy weapon systems, surveillance and sensor systems, mines, artillery rounds and ammunition, nuclear warheads, etc.

We are in dire straits. We need a "Manhattan Project" of the highest priority. Now!

The Soviets have already had the equivalent of seven or eight Manhattan Projects in this area.

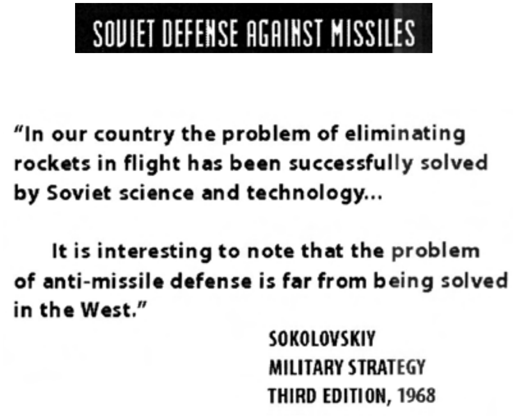

SOVIET DEFENSE AGAINST MISSILES

"In our country the problem of eliminating rockets in flight has been successfully solved by Soviet science and technology...

It is interesting to note that the problem of anti-missile defense is far from being solved in the West."

SOKOLOVSKIY
MILITARY STRATEGY
THIRD EDITION, 1968

Slide 106: Soviet Defense Against Missiles

Here is a significant statement by Sokolovskiy, contained in Military Strategy, 3rd edition, 1968.

This is not propaganda. The Soviets really do have an effective missile defense. We do not. At the present rate we're going, we may not have one in the year 2000 - if we live that long.

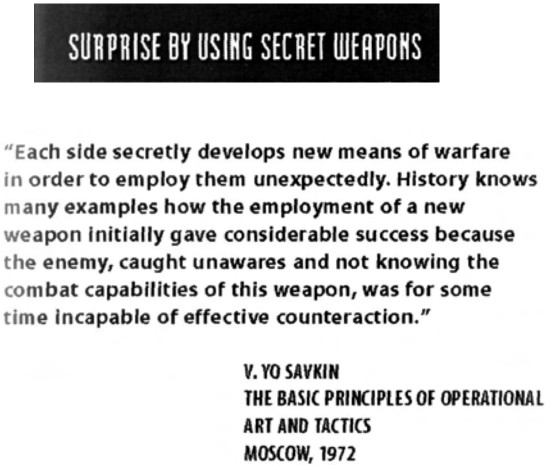

SURPRISE BY USING SECRET WEAPONS

"Each side secretly develops new means of warfare in order to employ them unexpectedly. History knows many examples how the employment of a new weapon initially gave considerable success because the enemy, caught unawares and not knowing the combat capabilities of this weapon, was for some time incapable of effective counteraction."

V. YO SAVKIN
THE BASIC PRINCIPLES OF OPERATIONAL ART AND TACTICS
MOSCOW, 1972

Slide 107: Surprise by Using Secret Weapons

The Soviets took to heart the lesson of our secret development of the atomic bomb. Thrust suddenly upon a fanatical nation, the atomic bomb brought the empire of Japan to its knees.

Other lesser examples - such as the impact of the German V-1 and V-2 missiles and jet fighter - also were not lost on the Soviets.

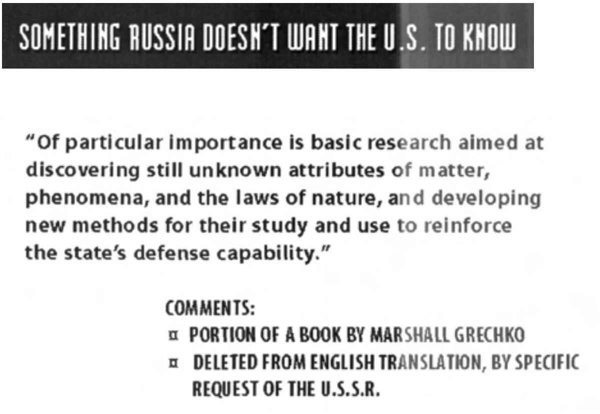

SOMETHING RUSSIA DOESN'T WANT THE U.S. TO KNOW

"Of particular importance is basic research aimed at discovering still unknown attributes of matter, phenomena, and the laws of nature, and developing new methods for their study and use to reinforce the state's defense capability."

COMMENTS:
- PORTION OF A BOOK BY MARSHALL GRECHKO
- DELETED FROM ENGLISH TRANSLATION, BY SPECIFIC REQUEST OF THE U.S.S.R.

Slide 108: Something Russia Doesn't Want the U.S. to Know

But Marshall Grechko said it even plainer.

He made it clear that the Soviet Union has actively searched for great technical breakthrough areas, particularly in discovering new laws of nature, unknown attributes of matter, and new phenomena, and developing their use for defense.

He made it so clear that the Soviet Union formally requested the U.S. State Department to have deleted from its. English translation that portion of his book.

That is a positive signature that the statement is exceedingly sensitive to the Soviets. In other words, it's a real tipoff.

So of course the State Department leaned on the publisher and got that section eliminated.

ADDITIONAL NOTES AND REFERENCES

1. Nikola Tesla, "The Problem of Increasing Human Energy," Century magazine, June 1900, p. 208-209. In this article Tesla spoke of using standing waves to produce distant effects and locate objects such as ships. At this point he was thinking of a scalar radar locator, something more advanced than even our radar today. It was also capable of producing electrical effects at any part of the globe. To quote: "Stationary waves... mean something more than telegraphy without wires to any distance... For instance, by their use we may produce at will, from a sending station, an electrical effect in any particular region of the globe; we may determine the relative position or course of a moving object, such as a vessel at sea, the distance traversed by the same, or its speed..."

2. Nikola Tesla, "The True Wireless," Electrical Experimenter, May 1919, p. 87.

3. T. E. Bearden, Comments on the New Tesla Electromagnetics: Part I: Discrepancies in Present EM Theory, Tesla Book Co., 1580 Magnolia, Millbrae, CA 94030, 1982.

4. T. E. Bearden, Toward a New Electromagnetics: Part III: Clarifying the Vector Concept, Tesla Book Co., 1580 Magnolia, Millbrae, CA 94030, 1983.

5. T. E. Bearden, Toward a New Electromagnetics: Part 4: Vectors and Mechanisms Clarified, Tesla Book Co., Millbrae, California, 1983, p. 6, 39.

6. T. E. Bearden, Solutions to Tesla's Secrets and the Soviet Tesla Weapons, Tesla Book company, 1981

7. T. E. Bearden, Tesla's Secrets and the Soviet Tesla Weapons, 1 hr. 40 min. videotape, Weather Engineering, POB 1472, Huntsville, AL 35807, 1981.

8. T. E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, Weather Engineering, POB 1472, Huntsville, AL 35807, 1985.

9. T. E. Bearden, Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization, Tesla Book Co., 1984

10. Stefan T. Possony, "The Tesla Connection," Defense & Foreign Affairs, Aug. 1984, p. 12-14, 27.

11. Stefan T. Possony, "Psy-War: Soviet device experiment," Defense & Foreign Affairs Daily, 12(104), June 7, 1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet LIDA device which is used to bombard human brains with radio waves in the 40 MHz region. Also reports on work by A. S. Davydov of the Ukrainian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected.

12. Fritz Albert Popp, "Photon storage in biological systems," in Electromagnetic Bioinformation: Proceedings of the Symposium, Marburg, Sep. 5, 1977. Edited by Fritz Albert Popp et al, Munchen, Wien, and Baltimore, 1979, p. 123-149. Dr. Popp has discovered, both theoretically and experimentally, that mitogenetic radiation from cells can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism. In addition, it includes the storage of "virtual" coherent photons. (In other words, there is a master EM field with high coherence, stationary over the whole body. This is the cells' master control system. Storage of virtual photons -- i.e., scalar signals -- occurs in this field.).

13. To gain an appreciation of the Aharonov-Bohm effect, see Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review, Second Series, 115(3), Aug. 1, 1959, p. 485-491. For an exhaustive discussion of the Aharonov-Bohm effect and a large number of references, see S. Olariu and I. Iovitzu Popescu, "The quantum effects of electromagnetic fluxes," Reviews of Modern Physics, 57(2), April 1985. See also Bertram Schwarzschild, "Currents in normal-metal rings exhibit Aharonov-Bohm effect," Physics Today, 39 (1), Jan. 1986, p. 17-20.

14. For details of the mysterious explosions of six major Soviet missile ammunition storage sites in seven months, see photograph, Jane's Defence Weekly, 2(3), July 28, 1984, p. 92; see also Mark Daly, "Goa and Goblet SAMs in Severomorsk explosion," Jane's Defence Weekly, 2(6), Aug. 18, 1984, p. 224; Derek Wood, "Soviets' northern fleet disabled... 'not viable' for six months," Jane's Defence Weekly, 2(1), July 14, 1984, p.3; Derek Wood, "Six explosions in the past seven months," Jane's Defence Weekly, 2(1), July 14, 1984, p. 3; John Moore, "The aftermath of Severomorsk," Jane's Defence Weekly, 2(6), Aug. 18, 1984, p. 224.

15. For details of the mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," Jane's Defence Weekly, 1(13), Apr. 7, 1984, p. 508.

16. For details of the Enhanced-V cloud signature of severe weather, see Gary Ellrod, "Dramatic examples of thunderstorm top warming related to downbursts," National Weather Digest, 10(2), May 1985, p. 7-13; T. J. Fujita, "Manual or Downburst Identification for Project Nimrod," SMRP Research Paper No.156, University of Chicago, 1978; D. W.McCann, ., The Enhanced-V: A Satellite Observable Severe Storm Signature," Monthly Weather Review, 111, 1978, p. 887-894.

17. For a good discussion of the modern view of the vacuum, see
I.J.R. Aitchison, "Nothing's plenty: The vacuum in modern quantum field theory," Contemporary Physics, 26(4), 1985, p. 333-391.

18. For a discussion of the ring interferometer experiments related to quantum and gravitational effects, see G.E. Stedman, "Ring interferometric tests of classical and quantum gravity," Contemporary Physics, 26(4), 1985, p. 311-332.

19. For information on self-induced transparency with pulsated light, see E. L. Hahn, Scientific American, June 1967. This is a scalar EM effect. The phenomenon is observed only in insulators, not in conducting materials. A dielectric tends to stop the flow of electrons, thus stopping the bleed-off of the Kaluza-Klein 5-potential as EM force fields. Thus through the insulator or dielectric, an electrogravitational pulse is transmitted. When this pulse emerges from the other side of the

dielectric, then electrons are free to move again and form ordinary EM fields, resulting in resumed EM bleed-off of the KK (Kaluza-Klein) 5-potential. Thus ordinary photons strike the dielectric, changing into electrogravitational waves which transit the opaque dielectric, and these EG waves transform back to EM waves on the far side of the dielectric. Advanced scalar EM application of this effect can make an object optically invisible.

20. For a lucid and profound presentation of the entire subject of elementary particle physics and quantum field theory, see T. D. Lee, Particle physics and introduction to field theory, Harwood Academic Publishers, New York, 1967. Subjects covered include those at the forefront of current research. Chapter 16, "Vacuum as the source of asymmetry," is particularly important. Here Lee concisely covers the idea of degenerate vacuum states, vacuum structuring, and vacuum effects causing symmetry breaking (violation of conservation laws). An excellent coverage of asymmetries and observables is given on p.181-188. For example, whenever symmetry is spontaneously broken, a nonobservable (virtual entity) has become observable. Thus there exist mechanisms for the direct translation between virtual and observable states.

Most engineers are unaware that physics has already rigorously proven (both theoretically and experimentally} that conservation of energy can be physically violated, even by excitation of the vacuum.

However, in the present assumption of a randomized virtual microstructure of a vacuum potential, these broken symmetry effects are individually confined to very small microlevels. For example, on p. 383, Lee points out that the microscopic structure of the scalar vacuum field (i.e., of vacuum charge or "potential") is not utilized.

This of course is reasonable if all vacuum potentials used in particle physics are treated as having a randomized microstructure of virtual particle flux vectors. Since any appreciable directional coherence in such a randomized microstructure could persist only over extremely tiny distances, then the microstructure of vacuum degenerate states has negligible local effects in ordinary macroscopic systems.

In other words, these conservation violation effects would be very difficult to coherently capture and utilize to provide "free energy, emerging from the vacuum."

The randomized momentary coherences, however, are responsible for the violent microfluctuations of the EM fields, well-known in classical electrodynamics. An approach such as taken by Yater or Moray would be necessary to achieve a vacuum energy tap under such conditions. To achieve usable macroeffects, a very large collection of charge or magnetic material, etc. would be required before the probability of usable coherences would become sufficiently large. Therefore this area is neglected in practical physics and engineering.

However, this is precisely where scalar EM is lost in conventional theory and engineering practice. In scalar EM, the coupling together of macroscopic EM force field vectors into a summed zero system means that the vacuum potential's microstructure is now macroscopically ordered by that system. Further, it is sustained during translation through space. In such case, one has produced directional gravitational potential gradients that have macroscopic range and form. Further, by coherently varying the amplitude of the individual components of the zero vector system, a gravitational wave -- a controlled oscillation of the degenerate states or vacuum -- is produced. By processing the EM components, this gravitational wave can be narrowed, formed into a beam, and projected. Two such beams can be interfered, even at a great distance, producing G-interference and causing EM effects to emerge in the distant interference zone. Thus scalar EM profoundly affects quantum mechanics, quantum field theory, particle physics, general relativity, and electromagnetics.

21. For some strong and pertinent critique of present scientific bias, see Ruggero Maria Santilli, Ethical probe on Einstein's followers in the U.S.A.: An insider's view, Alpha Publishing, POB 82, Newtonvllle, MA 02160. Santilli is a highly published physicist and the editor of a refereed scientific journal, the Hadronic Journal. After lengthy attempts to correct injustices within the Scientific system, to no avail, he finally has gone public with his specific charges. Santilli references senior Russian physicist G. Yu. Bogoslovsky, whose 1984 paper "Generalization of Einstein's relativity theory for the anisotropic spacetime," whose research would have been impossible in the U.S. Quoting Santilli: "He would have been discredited, humiliated, and chased out of these institutions [Harvard, M.I.T., and Yale] beyond any reasonable doubt by ample documentation on similar cases." Again to quote Santilli: "Papers bearing the title of 'Generalization of Einstein's relativity' are routinely permitted with human dignity at Moscow University in the U.S.S.R., but they are absolutely inconceivable today at the departments of physics of Harvard University, of the Massachusetts Institute of Technology, of Yale University and of other similar institutions in the U.S.A., as numerous senior scientists would readily testify."

You see, dear reader, there has been a deep scientific conspiracy in the U.S. to prevent any change in the severe restrictions placed on general relativity by Einstein. This has reduced GR to a sort of "special relativity, with only distant perturbations." Specifically, it has excluded scalar electromagnetics and branded as heresy the inconceivable idea that general relativity could be local, and engineerable.

Now you know why the Soviets have been able to develop these monstrous weapons for over three decades while our scientists slept at the wheel. Scientific arrogance and pride has been our greatest enemy.

See also Bryan G. Wallace, "The great speed of light in space coverup," Scientific Ethics, 1(1), Feb. 1985, p. 2-3; "Radar testing of the relative velocity of light in space," Spectrosc. Lett. 2, 1969, p. 361; Wallace, "Letter to the editor," Physics Today, 36(1), 1983, p. 11; "Letter to the editor," Physics Today, 37(6), 1984, p. 15. See M.E. Ash et al, Astron. J., 72, 1967, p. 72. Direct experiments on Einstein's second postulate, in the 1961 interplanetary radar contact with Venus, challenge whether relativity is correct. Variations of over 30,000% of the best possible general relativity fit the MIT Lincoln Lab could generate were measured. The variations were not random but contained diurnal, lunar, and synodic periodic components. See also Bryan G. Wallace, Spectrosc. Lett. 4, 1971, p. 79 for an analysis of published 1961 data on the interplanetary radar contact with Venus that concludes that the data showed a component that was relativistic in the Galilean c+v sense. See also Wallace, "The Unified Quantum electrodynamic ether," Foundations of Physics, 3, 1973, p. 381. which details Measurements strongly challenging the Einstein assumption of the constant velocity of light, and deals with the nature of the ether. See also Thomas E. Phipps, Jr., "Critique of accepted theories of special relativity and electromagnetism," Scientific Ethics, 1(2), Apr. 1985, p. 17-23.

22. For a sample of Soviet scientific freedom from narrow-minded bias in general relativity, see V. I. Denisov and A. A. Logunov, "The inertial mass defined in the general theory of relativity has no physical meaning," Teor. i Matemat. Fizika, 51(2), May 1982, p. 163-170 (In Russian). See also A. A. Vlassov and V. I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity," Teor. i. Matemat. Fizika, 53(3), Dec. 1982, p. 406-418. Here we find that the intensity of gravitational radiation and the total intensity determined in the general theory of relativity in accordance with Einstein's formulas depend on the choice of the coordinate system, so that they can be made to vanish or even become negative. [Note by T.E.B.: If local curvature is

engineerable, as in scalar EM, the local choice of coordinate systems can itself be directly engineered. This clearly implies that scalar EM can thus produce either "positive" intensity radiation or "negative" intensity radiation; thus it can "produce energy at a distance" or "extract energy from a distance.]
Continuing with Denisov: This result is a particular consequence of the general assertion to the effect that in general relativity there are no energy-momentum conservation laws for a system consisting of matter and the gravitational field. [Note by T.E.B.: Thus we can have an energy source or energy sink appear in the distant interference region, where matter has been lifted from the Dirac sea and exposed to the highly nonlinear interfering EG fields.]. See also V. I. Denisov and A. A. Logunov, "New theory of spacetime and gravitation, "Teor. and Matemat. Fizika, 50(1), July 1982, p. 3-76. Points out that ". . . the gravitational field in general relativity is completely different from other physical fields and is not a field in the spirit of Faraday and Maxwell." (p. 3, English version).

23. In the Proceedings of the Tesla Centennial Symposium, an IEEE Centennial Activity, Colorado College, Colorado Springs, Colorado, Aug. 9-11, 1984, see Richard A. Blade, "Modification of Maxwell's equations in free space to account for nonzero photon mass and scalar electromagnetic waves," p. 91-92; Jack Y. Dea, "Scalar fields: their prediction from classical electromagnetism and interpretation from quantum mechanics," p. 94-98; Robert W. Bass, "Self-sustained non-Hertzian longitudinal wave oscillations as rigorous solutions of Maxwell's equations of electromagnetic radiation," p. 89-90; T. E. Bearden, "Tesla's electromagnetics and its Soviet weaponization," p. 119-138.

24. To see the limiting assumptions imposed on general relativity that results in the severely restricted "ordinary general relativity" taught in almost all Western universities. see Charles W. Misner, Kip S. Thorne and John Archibald Wheeler, Gravitation, W. H. Freeman and Co., San Francisco, CA, 1973, p. 19-21, 71-72, 367-369. Among other things, these arbitrary restrictions save the conservation laws. When they are removed. the conservation laws can be violated.

25. Several anomalous events -- skyquakes, booms, or earthquakes -- have been documented in Europe. See H. W. Haak. "The Seismic event of November 9, 1983: Earthquake or sonic boom." Seismic Activity in Western Europe, (With particular consideration of the Liege Earthquake of November 8, 1983), Ed. by Paul Melchior, D. Reidel Publishing Co., 1985, p. 215-222. For such events in the United States, see S. Claflin-Chalton and G. J. MacDonald, "Sound and light phenomena," The Mitre Corporation, McLean, Virginia 22102, Nov. 1978.

26. Fundamental experiments were performed by Eotvos et al. in the early 1920s. See R.V. Eotvos, D. Pekar and E. Fekete, Ann. Phys. (Leipzig) 68, 1922, p. 11. Actually a fifth force of nature due to hypercharge may have been found by these experiments. See Ephraim Fischbach et al. "Reanalysis of the Eotvos experiment," Physical Review Letters, 56(1), Jan. 6, 1986, p. 3-6. See also F. D. Stacey and G. J. Tuck, Nature 292, 230 (1981); S. C. Holding and G. J. Tuck, Nature 307, 714 (1984); G. W. Gibbons and B. F. Whiting, Nature 291, 636 (1981); F. D. Stacey, in Science Underground, edited by M. M. Nieto et al., AIP Conference proceedings No. 96, (American Institute of Physics, New York, 1983, p. 285. See also S. H. Aronson et al, Physical Review Letters 48, 1982, p.1306; Physical Review D 28, 1983, p. 476, 495. Also, comparison of satellite and terrestrial determinations of the gravitational field constant reveals significant disagreement: see R. H. Rapp, Geophys. Res. Lett. 7, 1974, p. 35; Bull. Geod. 51, 1977, p. 301.

27. John J. A. O'Neill. "Tesla cosmic ray motor may transmit power 'round the earth: Famed scientist, on eve of 76th birthday, says he has succeeded in harnessing 'penetrating rays' to operate small motive device." Brooklyn Eagle, July 10,1932. Tesla reported he had harnessed "cosmic rays" and caused them to operate a small motive device. Tesla's use of "cosmic rays," however, encompassed scalar EM radiation. Tesla later seems to have progressed to a simpler motive unit. In the Kansas City Journal-Post, Sep. 10, 1933, he stated: My power generator will be of the simplest kind -- just a big mass of steel, copper and aluminum, comprising a stationary and rotating part, peculiarly assembled." A similar type of "gravitational turbine" seems to have been invented by John Bedini.

28. To see just how arbitrary and postulational are present "definitions" of mass and force, see Robert Bruce Lindsay and Henry Margenau, Foundations of Physics, Dover Publications, New York, 1963, p. 283-287. Note on p. 283 that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. Also see Richard P. Feynman, Robert B. Leighton, and Matthew Sands, The Feynman Lectures on Physics, Addison-Wesley, New York, Vol. 1, 1963, p. 2-4 for a definition of the electric field in the context of its potentiality for producing a force. This problem did not exist for the original foundation of electromagnetic force fields in vacuum, since the vacuum was assumed to be a material ether. With the modern view a nonmaterial ether has emerged, yet the classical electromagnetic theory has not been made sufficiently rigorous in distinguishing "observable" and "virtual" electromagnetic force fields.

29. E. P. Battey-Pratt and T. J. Racey, "Geometric Model for Fundamental Particles," International Journal of Physics, Vol. 19, No. 6, 1980, p. 437-475.

30. Some years ago, Dr. William Tiller produced very important "free A-field" theoretical work which showed waves which did not fall off as $1/r^2$. Interesting theoretical work in scalar wave theory has also been advanced by Dea. See Jack Dea, "Fundamental Fields and Phase Information," presented at the 1985 U.S. Psychotronics Association Conference, Bergamo Center, Dayton, Ohio, July 1985. Dea has formulated an important theory for vacuum EM scalar waves, derived from Maxwell's equations with zero E and B fields. He predicts waves whose intensity falls off with distance as $1/r$ instead of $1/r^2$, and detection of these waves by various means. He shows simple generators for the waves. Detection is quantum mechanically by electron interference, not by single electrons. Thus such scalar waves will pass through normal Faraday shields. In his theory Dea does not address the question of spin vortexes in vacuum scalar waves. It is the present author's view that all vacuum waves are scalar waves, and the primary distinction between the "transverse" EM wave and the longitudinal scalar wave is simply the presence or absence of spin-vortexes. Indeed, one may argue that the present "vector wave" EM theory can be replaced by a scalar wave theory. In 1903 Whittaker proved that any vector field can be replaced by two scalar fields. See E. T. Whittaker, Proc. Lond. Math. Soc.1, 367, 1903. By implication, since a vector wave is simply a moving vector field, it can be replaced by two moving scalar fields.

31. David Jones, "Israel's Secret Weapon," Weekend Magazine, Dec.17, 1977.

32. John Hutchinson, private communication. Masses of wood, metal, etc. -- some weighing over 60 lbs. -- have been levitated by Hutchinson.

33. For a description and details of the April 9, 1984 incident off Japan, see Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoku Seamount and the Mystery Cloud of 9 April

1984," Science, 227(4684), Feb. 8, 1985, p. 607-611. See also transcription from Japan Airlines flight 036 to Anchorage International Flight Service, 9 April 1984, at 1406 Greenwich mean time, Federal Aviation Administration (FAA) — Alaska Region; Jim Derry, Special Agent, FAA-Alaska Region, interviews with Captain C.H. McDade, Japan Airlines Flight 036, with Captain Cornelis Vander Berg, KLM Royal Dutch Airlines Flight 868; with First Officer Reint Sol, KLM Royal Dutch Airlines Flight 868, and with Captain Mike Howe, Flying Tiger Lines Flight 078, Anchorage Daily News, 10 April 1984, p. A-l; Anchorage Daily News, 12 April 1984, p. A-l; Honolulu Advertiser, 10 April 1984, p. B-l; Newsweek, 30 April 1984, p. 25, Gregg Ripee, "Mushroom cloud sighted off Japan," Los Angeles Daily News, 11 April 1984, p. 1,8; Ross Gelbspan, "Mystery cloud baffles scientists," Boston Sunday Globe, 13 Jan. 1985, p. 1, 14; T.E. Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. See also McKenna, Daniel L. & Daniel Walker,"Mystery Cloud: Additional Observations", Science, 234(4775), Oct. 24, 1986, p. 412-413.

34. For a description of the April 11, 1963 anomalous underwater explosion north of Puerto Rico, see Robert J. Durant, "An Underwater Explosion — or What?", Pursuit, 5(2), April 1972, p. 30-31.

35. For details of the loss of the U.S.S. Thresher and the interference experienced by the U.S.S. Skylark, see John Bentley, The Thresher Disaster, Doubleday, Garden City, N. Y., 1975, particularly p. 164. Some of the communications systems of the Skylark experienced temporary failure, but were not disabled, later resuming operation.

36. The Soviet Union has a well-established history of interfering with and destroying aircraft; for example, for a listing see "Soviet and Warsaw Pact air incidents," Jane's Defence Weekly, 3(2), Jan. 12, 1985, p. 58.

37. For details of the anomalous "nuclear flashes" detected by U.S. Vela satellites, see "Scientists Fail to Solve Vela mystery," Science, 207, 1 Feb. 1980, p. 504-506. "Satellite Evidence Shows 'possibility' of Nuclear Test, DOD Says," Aerospace Daily, Oct. 29, 1979, p. 286. "A Flash of Light," Newsweek, Nov. 5, 1979, p. 64-65. "Was It a Nuclear Device?", Newsweek, July 21, 1980, p. 19. "Diverging Views," Washington Roundup, Aviation Week & Space Technology, July 21, 1980, p. 15. Philip J. Klass, "Clandestine Nuclear Test Doubted," Aviation Week & Space Technology, Aug. 11, 1980, p. 67, 69, 71-72. "Debate continues on the Bomb That Wasn't," Science, 209, 1 Aug. 1980, p. 572-573. "Navy Lab Concludes the Vela Saw a Bomb," Science, 209, 29 Aug. 1980, p. 996-997. See also Jack Anderson and Dale Van Atta, United Press Syndicate, "CIA knew South African-Israeli nuclear test was coming," Birmingham Post-Herald, Apr. 26, 1985, p. A5. The latter article repeats the CIA/DIA argument that the Vela incidents represent concurrent South African-Israeli testing of atomic weapons. Taiwan is considered a possible third partner. According to Anderson and Van Atta, a South African navy contingent was "in the area" at the time of the 1979 Vela detection, and Israel and South Africa were cooperating on an atomic bomb project as early as 1966. Apparently U.S. announcement of an impending South African atomic test in the Kalahari Desert in 1977 led to such an outcry that the test was suspended. Anderson and Van Atta cite succeeding visits to South Africa by Israeli scientists, technicians, and defense officials. However, since the actual Vela detections showed some deviations from what would be expected from nuclear tests, one can speculate that Anderson and Van Atta have presented evidence for the development of scalar electromagnetic weapons by two additional countries. One may further speculate that this could reveal what has been checking the Soviets from simply moving against the West with scalar electromagnetic weapons.

Indeed, the mysterious destruction of six major Soviet missile ammunition storage areas within the space of seven months could have been a preemptive measure and warning to the Soviets if they had actually been considering moving against the West. Also, other incidents reported in the South Atlantic may possibly be associated with the program producing the "Vela flashes." For example, see N.W.C. Rutherford, "Unidentified Phenomena," Marine Observer, 51, 1981, p. 186 for a description of a great bluish-white flash and small bolt of lightning, indicating some sort of anomalous electrical discharge, on Dec. 23, 1980 at 2245 hours (10:45 p.m.). Many other anomalous flashes for which no known causative mechanism exists have been detected by satellites. For example, the OSO-2 and OSO-5 satellites have detected about one mysterious short (less than 0.1 sec) flash about every ten orbits, at times when the fields of view of the telescopes were well above the earth. As many as three of the satellite's photometers have detected a flash simultaneously. These flashes must be removed from any discussion of satellite-observed lightning; see J. G. Sparrow and E. P. Ney, "Lightning Observations by Satellite," Nature, 232, 1971, p. 540.

38. For a description of sightings of enormous glowing globes of light deep within the Soviet Union in the direction of Saryshagan, see Gwynne Roberts, "Witness to a Super Weapon?", The Sunday Times, London, England, Aug. 17, 1980. Several incidents were observed by Downie and others.

39. For an artist's conception of a directed energy weapon installation at Saryshagan, see Aviation Week & Space Technology, July 28, 1980, p. 48. For a DOD sketch, see Soviet Military Power, 1985. Obviously classically-trained U.S. military analysts, with no knowledge of scalar interferometry, have not considered the installation as a candidate hyperspatial howitzer.

40. For a description of multiple pilot sighting (from Iran) of an enormous glowing globe of light deep within the Soviet Union, see Central Intelligence Agency (CIA) Information Report No. CO-B-321/15354-66, "Sighting of Unusual Phenomenon on Horizon Near Iranian/USSR Border," Sep. 8, 1966, released on 15 Dec. 1978 under the Freedom of Information Act (FOIA). The sighting occurred at 2142 hours (9:42 p.m.).

41. For details of sighting of the intensely glowing ball of light near British European Airways flight 831 between Moscow and London, see CIA Foreign Intelligence Information Report CO-B-324/33601-76, "Aerial observation of intense Source of Light," Nov. 18, 1976, released on 15 Dec. 1978 under the FOIA. The incident occurred between 1800 and 1900 hours (6:00-7:00 p.m.). The light was observed for 10-15 minutes, until the aircraft had flown past and left the light behind.

42. For other examples of pertinent 1ight phenomena: F. S. Angus and G. Carling, "Optical Phenomenon: Caribbean Sea; Western North Atlantic," Marine Observer, 40, Jan. 1970, p. 17-18; R. A. Holmes, "Unidentified Phenomenon, off Barbados, West Indies," Marine Observer, 40(229), July 1970, p. 107-108; H. K. Dyer, "Unidentified Phenomenon: Western North Atlantic," Marine Observer, 43, July 1973, p.114. For a 1961 double concentric hemispheric shape observed in the Indian Ocean, see Marine Observer, 32, 1962, p. 64. For a particularly significant combination of forms, see Marine Observer, "Unidentified Phenomenon," 48, 1978, p. 21-22. Also, on the night of July 26, 1984 the pilot and crew of a Boeing 747 (American carrier) flying from Tokyo to Fairbanks, near the Kuriel Islands, at 47 5'44" N: 161 00'05" E, noticed a slowly expanding hemisphere of white light off to their left above the horizon. The shell of light continued to expand over a 10-minute

period until ahead of them and to the right. The crew braced for a shock wave which never arrived. Their weather radar saw nothing out of the ordinary. The shell of light had sharp edges and was semi-transparent so that stars became visible through it. (Richard F. Haines, private communication, Sept. 27, 1984). See also Richard Hall, "Aerial anomalies at sea," The Info Journal, 4(3), May 1975, p. 6-9 for an unusual light phenomenon seen in the Caribbean by the R.M.S. Carmania (British). The phenomenon occurred four times at precise hourly intervals, strongly suggesting an artificial origin. See also F. Shepherd, "Cloud: Strait of Gibraltar," Marine Observer, 51(273), July 1981, p. 107-108 for an incident of associated cloud, arch, and anomalous pressure effects. For another significant airline sighting by Japan Airlines flights 403 and 421 of a giant, expanding globe of light on June 18, 1982, in the North Pacific, 100 km. east of Kushiro, see the Asahi Evening News, Tokyo, Japan, June 22, 1982. For additional evidence that nomadic tribesmen passing through Kazakhstan repeatedly observe such gigantic light phenomena in close association with Saryshagan missile test range, and that Soviet deception planners foster the notion of "UFO activity" as a Red Herring, see Robert Jackson, "KGB takes to the flying saucers!", Northern Echo, Darlington, England, March 15, 1984.

43. For details of Khrushchev's announcement of a Soviet superweapon, see Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'", New York Times, Jan. 15, 1960, p. 1.

44. See Christopher S. Wiren, "Brezhnev Calls for Accord against 'Terrifying Arms,'" New York Times, June 14, 1975, p. 1, 11. In a major Kremlin speech on June 13, 1975, Leonid I. Brezhnev included a statement that one major concern not covered by existing agreements was "assuming ever greater urgency with each passing day." Calling for a ban on new weapons of mass destruction, the Soviet leader stated: "The level of modern technology is such that a serious danger arises of creating weapons more terrifying than nuclear ones." He stated, "The reason and conscience of humanity dictate the necessity of erecting an insurmountable barrier to the development of such weapons." See also Malcolm W. Browne, "Senatorial Group Received by Brezhnev," New York Times, July 3, 1975, p. 2. Leonid Brezhnev repeated his proposal to a group of U.S. senators on July 2, stating that the capability to develop more terrible weapons was obvious. He emphasized, however, that he "had no particular weapon in mind." Here he may have been deliberately misleading the U.S. senators, to prevent revealing that the Soviet Union had already developed such weapons, and was horrified at their potential for total destruction. E.g., if massive and rapid scalar explosions occur, the linearity of spacetime -- and the orderly flow of time itself -- may be seriously perturbed throughout the earth and its adjacent space. From a general relativistic viewpoint, such an event might well prove catastrophic for the entire biosphere. Boris N. Ponomarev, a Soviet national party secretary, again raised the same issue to a delegation of visiting U.S. congressmen in the Kremlin on Aug. 11, 1975; see Christopher S. Wren, "Moscow Now Pressing Disarmament." New York Times, Aug. 12, 1975, p. 6. At the United Nation's thirtieth session of the General Assembly on Sept. 23, 1975, Foreign minister Andrei A. Gromyko strongly raised the same issue, warning that science can produce "ominous" new weapons of mass destruction. He urged that all countries, led first by the major powers, should sign an agreement to ban the development of these unspecified new weapons. He even offered a draft, entitled "Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons." The first article provided that the types of these new weapons would be "specified through negotiations on the subject." Because of its fixation on nuclear weapons, the West may have lost its only opportunity to prevent the spread of scalar electromagnetic weapons of power unprecedented even by nuclear arms. Of course, it may also have

been a Soviet ploy to prohibit the U.S. from developing scalar electromagnetics defensive weaponry to counter the massive strategic scalar electromagnetics weapons already developed, tested, and deployed by the Soviet Union. Certainly in 1973 at a secret meeting in Prague with European communist leaders, Brezhnev had laid out a firm statement of Soviet domination of the world by 1985. See "Secret Speech: Did Brezhnev Come Clean?", National Review, 29(8), Mar. 4, 1977, p. 248, 250. Brezhnev's speech contained statements such as: "We are achieving with detente what our predecessors have been unable to achieve using the mailed fist...come 1985 we will be able to extend our will wherever we need to." Brezhnev stated the Soviet Union would control the oceans of the world, 90% of the land, and the air and space above. A front-page Boston Globe article on Feb. 11, 1975 was the first to carry the story. A copy of Brezhnev's secret speech was obtained by British intelligence in 1973, but denigrated by Secretary of State Kissinger. Apparently the first official U.S. reference to Brezhnev's startling speech did not appear until the 1976 National Intelligence Estimate. Although the big intelligence bureaucracies can do a fantastic job with "current intelligence," their record at predicting a drastic change is woeful, precisely because of the political nature of bureaucracies. See "What is Intelligence," Colonel William V. Kennedy et all, Intelligence Warfare: Penetrating the Secret World of Today's Advanced Technology Conflict, with an introduction by Dr. Ray S. Cline, Crescent Books, New York, 1983, p. 8-23. Intelligence Warfare is heartily recommended as the single book which best gives the reader a grasp of modern technical intelligence organizations, capabilities, and operations.

45. For details of Soviet weather engineering operations over the U.S., see Thomas E. Bearden, "USSR: New beam energy possible?", Defense & Foreign Affairs Daily, 13(111), June 12, 1984, p. 1-2. Cloud signature patterns reported in the article have been repeatedly sighted over multiple areas in the U.S. A series of anomalous "hole in the cloud" patterns, possibly associated with early Soviet weather engineering efforts over the U.S., was widely seen in 1967-68. See Weatherwise, 21(4), Aug. 1968, cover and p. 143; 21(5), Oct. 1968, p. 194-195, 204-205; 21(6), Dec. 1968, p. 238-245. Peculiar stationary blocking patterns have also been observed to occur and influence weather over the U.S., leading to severities such as the winter of 1976-77. See K. K. Tung and R. S. Lindzen, "A theory of stationary long waves, Part 1, A simple theory of blocking; Part 2: Resonant Rossby waves in the presence of realistic vertical shears; Part 3, Quasi-normal modes in a singular waveguide," Monthly Weather Review, 107(6), June 1979, p. 714-774. See also T. E. Bearden, Soviet Weather Engineering Over North America, videotape, 1985 for an expose or the physics and some major incidents of major Soviet weather engineering, including details and photographs of cloud signature patterns. See also C. B. Baker, "Soviet Weather Mayhem," Youth Action News, Nov. 1983, P.O.B. 312, Alexandria, Virginia 22313, p. 1-8; C. B. Baker, "Diabolic Soviet Warfare," Youth Action News, Nov. 1984; For a conventional Soviet look at weather modification (possibly intended to mislead the West), see E. T. Protasevich, "Microwave weather modification," Soviet Technical Physics Letters, 7(3), March 1981, p. 114-115. For a downburst mechanism, see Gary Ellrod, "Dramatic examples of thunderstorm top warming related to downbursts," National Weather Digest, 10(2) May 1985, p. 7-13. See Art and Kay Westermann, Weather War — Worldwide, Liberty Bell Books, POB 288, Sedona, Arizona 86336, 1983. See also U.S. Senate Resolution on Environmental warfare, July 11, 1973.

46. For details of the Soviet Woodpecker beams bending around the earth and intersecting over the U. S., see Christopher Chant and Ian Hogg, Nuclear War in the 1980's?, Harper & Row, New York, 1983, p. 71. See also Soviet Military Power, U.S. Department of Defense, 1985, p. 45. The

intersection forms an interference grid. Scalar EM channels added to the Woodpecker carriers then provide a corresponding sca1ar EM interference grid in the scalar area.

When adjusting the scalar interferometry components of the grid, bursts in the interference and discharges of energy as well as bursts of energy extractions occur. This leads to the production of anomalous aerial phenomena such as explosions, rumblings, airquakes, etc. In addition, direct Bohm-Aharonov forces may be engendered in ground structures, shaking buildings and windows, etc. without seismic disturbances. Such anomalous disturbances have been widely noted in the open U.S. press. E.g., particularly see Kenneth F. Bunting and Daniel M. Weintraub, "Mysterious Rumbles Jar Southland," Los Angeles Times, Dec. 8, 1984, CC(II), p. 1,10; Let Scarr, "People shook up over mystery shakes," The San Diego Union, Dec. 8, 1984, p. Bl, 4. Following this and a Dec. series of anomalous aerial disturbances off the Los Angeles/San Diego coast, a sharp and unexpected weather front appeared, racking the area with high winds on the night of Dec. 12, producing widespread damage and power outages. The newly-formed front moved on to produce significant snowstorms in other areas. See Ted Thackery, Jr., "Winds Kill Two, Wreak Havoc as They Sweep Through the Southland," Los Angeles Times, Dec. 14, 1984, MF/Part II, p. 1,8; "Even Tucson gets snow in storm that dumps 20 inches," AP release, Los Angeles Herald Examiner, Dec. 14, 1984, p. A-12. During several months prior to the December 1984 Los Angeles incident, anomalous electrical failures and outages had occurred in the San Diego and Southern California area, possibly due to scalar interferometry adjustments in the Woodpecker grid over the area. For other typical incidents possibly associated with adjustment of the interference grid, see "Who boomed? Odd Noise unnerves Ohioans," Journal, Milwaukee, Wisconsin, Dec. 28, 1982; "Tremors shake Miss. Gulf Coast," Times-Picayune, New Orleans, Louisiana, Feb. 27, 1983; Sandra Skowron, "Sonic boom is just Mother Nature," Beaver Co. Times, Beaver Falls, Penn., Jan 13, 1983; "Booms in Rockland, New York Times, Mar. 1. 1979, Sec. 8, p. 82; "Loud 'boom' unexplained," Star-Beacon, Ashtabula, Ohio, Feb. 21, 1983; Paul Senecton, "The word is now IDENTIFIED," Mail, Hartlepool, England, Feb. 3, 1985: On Jan. 23, 1974 a mysterious explosion over North Wales rocked a 60-mile radius area. Associated anomalous light phenomena were also seen. Hundreds and hundreds of anomalous booms, aerial rumbles, and aerial explosions, many accompanied by flashes or anomalous shaking of ground structures, buildings, windows, etc., have occurred over the U.S. and other Western nations in the 60's, 70's, and 80's.

On occasion, an aircraft at the proper altitude and position can conceivably experience interference in its electrical systems and in its engine combustion processes due to exothermic kindling or endothermic extraction of electromagnetic energy. The combustion processes of aircraft engines produce highly ionized gases in very nonlinear conditions. These can act as strong dephasers and decouplers of the vector EM components of a vector zero summation in a scalar interference zone. Exothermic interference would have little effect on combustion, but strong endothermic interference can result in engine flameout. Anomalous simultaneous failure of jet and propeller-driven aircraft engines sometimes occurs. For high altitude grids, this can result in engine flameouts, but restart and recovery after unusual loss of altitude has dropped the aircraft out of the scalar interference zone. Several candidate incidents over and near the U.S. exist, particularly where power failures were experienced by multiple aircraft. The more usual exothermic case, however, if substantial could result in anomalous loss of control of the aircraft. For a recent candidate incident, see "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," Associated Press release, Huntsville Times, Feb. 20, 1985, p. 1 and "China Airlines Pilot Denies Cockpit Error," Associated Press release, Huntsville Times, Feb. 21, 1985, p. A-9. Deliberate use of this effect as an anti-strategic-bomber weapon capable of destroying bombers at or shortly after takeoff is possible (the Arrow DC-8 which crashed just after takeoff from Gander Air Force Base, Newfoundland on December 12, 1985 may have been deliberately interfered with by just such a Soviet weapon -- see extensive newspaper and TV news coverage of that event).

In scalar EM interference with aircraft, certain signatures may occur. Different instruments may sharply contradict each other, since their activation and inactivation is no longer a simple function. The pilot's instrument panel, for example, may indicate failure of the control system, when actually it has not failed, and the flight recorder may record that the control system is still operating. In such case, the pilot may take proper action based on his own instrumental readings, yet these actions may be contraindicated for the actual status of the aircraft, resulting in opposing pilot/autopilot actions. Tentative indications of the jetliner that anomalously dropped 32,000 feet and recovered are that apparently such anomalous instrument disagreements occurred. Another signature is that, as the EM charge accumulates in and on the aircraft, its skin charges up electrically, producing a glowing corona similar to St. Elmo's fire. The Arrow DC-8 was observed to be glowing yellow by eyewitnesses, according to the Canadian Broadcast Corporation's 10:00 p.m. network TV news on April 8, 1986.

47. T. E. Bearden, Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization, Tesla Book Co., Millbrae, California, 1984, p. 19-22.

48. For a paper possibly related to the first Bedini scalar wave detector from the "free A-field is a scalar field" view, see Robert R. Lewis, "Aharonov-Bohm effect for trapped ions," Physical Review A, 28(3), Sept. 1983, p. 1228-1235. This paper analyzes the role of the Aharonov-Bohm effect for the bound states of a charged particle in the superposition of two fields: a uniform magnetic field plus the vector potential of a solenoid. Gives exact solutions for a solenoid of zero size and discusses the corrections due to a solenoid of finite size. Shows that a series of new "betatron resonances" appear in the spectrum, dependent on the flux quantum number. The fundamental betatron resonance remains narrow and unshifted for solenoid radii large enough to broaden and shift the cyclotron resonance. See also W. Ehrenburg and R. E. Siday, Proc. Phys. Soc. London, Sect. B62, 1949, p. 8 which, among other things, discusses the effect of the vector potential in electron interferometry, where the insertion of a tiny solenoid between the interfering paths is shown to give an additional phase difference proportional to the flux through the solenoid. The resulting shift in the fringe system is called the Aharonov-Bohm effect.

49. Frank Golden has developed a most important and precise scalar EM detection system to display scalar EM signals on a cathode ray tube. His important measurements have mapped the scalar resonant frequency responses of the entire earth. Two completely different types of scalar instrumentation show precise agreement on the frequency maps. The measurements show that the Soviet Union has for some time entrained the entire earth into giant scalar electromagnetic resonances on two frequencies approximately 12 kilohertz apart. It is the opinion of the author that the beat frequency between these two giant resonances is utilized by the Soviet Union to extract enormous power from the earth, essentially by a mechanism that Golden has uncovered. A complex of four to eight weapon systems might reasonably be powered from the single bi-resonant tapping source. In April 1985 Golden discovered the rather sudden activation of many other scalar frequencies with characteristics resembling command and control communications. In addition, a significant number of additional paired scalar resonances had been entrained upon the earth. It is the

opinion of the author that this showed a major activation of the Soviet strategic scalar electromagnetic defense system, with the power up of from 80 to 100 weapons, along with the complete command and control system. Golden has previously built a prototype of an actual Soviet scalar communications system and demonstrated that it works. Because of the impending mayday and the 1973 statement by Brezhnev indicating that 1985 was a critical planning year, the sudden activation of such massive numbers of weapons and communications was alarming. Accordingly, the author discreetly notified certain contacts to alert the U.S. system. Golden then detected the periodic turn-on and turn-off of the systems, indicating a large command and control exercise rather than preparations for direct tactical use. An additional difference, however, was the change from scalar electric to scalar magnetic field on many of the large transmitters. This is indicative of electrostatic location, targeting, and adjustment of the target on the earth and then switching to the "power input" mode. It is the opinion of the author that this indicated the choice of a geological target and proceeding to attack that geological target in the "slow buildup" mode. Again, proper contacts were so notified.

50. A very good coverage of the Soviet microwave radiation of the U.S. Embassy in Moscow is given in Paul Brodeur, The Zapping of America, W. W. Norton & Co., New York, 1977, passim. None of our U.S. Embassy security personnel knew of scalar interferometry, nor did they possess scalar electromagnetics detectors. Further, orthodox U.S. scientists were also ignorant of scalar electromagnetics. The Soviets were thus in a position to tailor the radiation signals and precisely determine what our on-site personnel would detect. Details of the effects on U.S. Embassy personnel, possibly associated with the microwave radiation, are given by Brodeur. A great number of related releases by the open news media also exist. For example, see Robert C. Toth, "Radiation Devices Bug U.S. Embassy," Huntsville Times, reprinted from the Los Angeles Times, Feb. 8, 1976; "U.S. Screening out Soviet 'Listening Rays'," United Press International release, Washington D.C. Star, Feb. 11, 1976; Norman R. Beebe, "Nixon Exposed to Radiation In Russia?", Associated Press release, Huntsville Times, Apr. 30, 1976; Barton Reppert, "U.S. Scientists Puzzled Over Embassy Radiation Effect," Associated Press release, Huntsville Times, May 28, 1976; Barton Reppert, "Radiation Entering Embassy in Moscow through Phone Line," Associated Press Release, Huntsville Times, May 28,1976. See also Martin Ebon, Psychic Warfare: Threat or Illusion, McGraw Hill Book Co., New York, 1983.

51. For details of the electromagnetic transmission of cellular disease and death, see Vlail P. Kaznacheyev, "Information function of ultraweak light flows in biological systems," Problems in Biophysics, Novosibirsk, 1967, p. 7-18. Kaznacheyev et al, "Conditions necessary for appearance of distant intercellular interactions after UV-radiation," Bulletin Experimentalnoy Biologii i Meditsiny, (5), 1979, p. 468-471; Distant intercellular interactions in a system of two cultures connected by optical contact," Ultraweak Luminescence Biology, Moscow, 1972, p. 224-227; "Distant intercellular interactions caused by UV-radiation," Photobiology of a Living Cell, Leningrad, 1979, p. 221-223; "Studies on the biological role of electromagnetic emission as a factor of adaptive behavior of cells under conditions of latitude changes," Abstracts of the second all-union conference devoted to the problem of human adaptation to different geographical, climatic, and labor conditions, Novosibirsk, 1977, p. 101-104; Vlail P. Kaznacheyev, S. P. Shurin and L. P. Mikjhailova, "Distant intercellular interactions in a system of two tissue cultures," Official Bulletin of the committee on inventions and discoveries affiliated to the council of ministers of the USSR, Discovery no. 122 (19), 1973, p. 3, also printed in Psychoenergetic Systems, 1(3), Mar. 1976, p. 141-142; "Apparent Information Transfer

Between Two Groups of Cells," Psychoenergetic Systems, 1(1), Dec. 1974, p. 37.

52. For evidence of anomalous scalar reflection out of the scalar grid caused by a Navy chaff drop, See "Navy sabotages traffic lights, TV sets," United Press International, PM cycle, Jan. 11, 1985; "Power blackout blamed on anti-radar particles," United Press International, AM cycle, Jan. 11, 1985; "Power Disrupted Briefly After Navy Plane Drops Metallic Strands," Associated Press, AM cycle, Jan. 16, 1985.

53. A good summary of the early incidence of the Soviet Woodpecker signals is given by Bradley Wells, "The Russian woodpecker: a continuing nuisance," CQ, Nov. 1984.

54. A particularly important Soviet Paper is E. B. Smetanin, "Electromagnetic field in a space with curvature -- new solutions," Soviet Physics Journal, 25(2), Feb. 1982, p. 107-111. In this paper, a solution to the stated problem is obtained, which may be regarded as a classical model of a charged particle that has both a magnetic moment and a nonzero magnetic charge density. Interaction with spacetime curvature can disrupt the gauge invariance of the EM field, and it can also reestablish invariance in an initially noninvariant theory. (Hutchinson has had anomalous breaks of metal caused by standing scalar waves, which seems to be due to a mechanism that produces monopoles and deposits them in the metal at standing-wave nodes. This paper should be directly applicable.)

See also Yu. S. Vladimirov and A. A. Kozlenkov, "6-optics and a general theory of gravitation and electromagnetism," Soviet Physics Journal, 27(12), Dec.1984, p. 1039-1041; V. G. Bagrov and M. D. Noskov, "New exact solution of the Dirac equation. XI," Soviet Physics Journal, 27(12), Dec. 1984, p. 1030-1034. See particularly V. I. Petukhov and I. A. Kuzin, "Weak processes in the field of a gravitational wave," Soviet Physics Journal, 27(12}, Dec. 1984, p. 1025-1029.

In the latter paper it is significant that the decay probability increases in a locally curved spacetime, and that solutions are obtained for circularly polarized waves.

(In EM waves, circularly polarized waves have special characteristics in nonlinear isotropic media: exact solutions exist. Also, when two longitudinal sine waves of differing frequencies are transmitted into a nonlinear isotropic medium, the apparent difference frequency is transmitted as a sine wave, undistorted. This characteristic has been used by Westinghouse to produce improved sonar transmission through seawater, for example. The scalar EM (electrogravitational) wave may be represented as a longitudinal EM wave. Immediately, if one wishes to produce a pure scalar EG wave in the nucleus of an atom, two frequencies should be transmitted such that their difference is the frequency wished. Or circularly polarized waves may be utilized.)

Note that, by implication, one should be able to deliberately tailor scalar EM waves to greatly speed the decay of long-lived radioisotopes, such as are contained in radioactive wastes from nuclear reactors and other processes.

55. For some evidence of continuing persistence of the higher frequency scalar components applied to the scalar grid on Jan. 28, 1986 to localize the effects when attacking the Challenger, see "Poison suspected in unexplained deaths of birds," Huntsville Times, AP release, Feb. 6, 1986. From 1-4 Feb.

1986, various kinds of dead birds -- such as blackbirds, cardinals, sparrows and bluebirds -- plummeted to the ground in Birmingham, Alabama. At first some sort of mysterious poison was suspected, but later was ruled out. During this period, the woodpecker grid was particularly active in weather engineering operations. Multiple sightings of giant cloud radials occurred in various places throughout the U.S. The retention of the high frequency components possibly was intensified in the Birmingham area sporadically during the period. Birds flying into this zone would be killed if their brains acted as scalar receivers and were sufficiently jammed by the signals. Normally Huntsville (further north) is a "hinge point" for increased scalar activity, to start the Jetstream bending northeast to move along the Appalachian mountain chain. However, just before 28 January the jetstream had been diverted further south than usual by the Soviet interference, bringing unseasonably cold weather to Florida and the waiting Challenger on its launch pad. Probably hinge points further south than Huntsville were needed for this operation. If these more southern points were retained for a few days, and the high frequency components were retained, then an inadvertent high intensity area of these components may have sporadically occurred in the Birmingham area, killing the birds.

For description of the Apr. 18, 1986 Titan blast and its ramifications to the spy satellite program, see James Gerstenzang and Ralph Vartabedian," Los Angeles Times News Service, "Titan Blast to Hurt Spy Program," Huntsville Times, Apr. 19, 1986.

56. For another very important paper indeed, see Vu. G. Ignat'ev, "Relativistic kinetics of an anisotropic plasmalike medium with damping in a field of gravitational radiation," Soviet Physics Journal, 27(12), Dec. 1984, (English translation: Plenum, June 1985), p. l066-1069. If we substitute the vacuum "virtual plasma" for the observable plasma, this paper may be part of the smoking gun to show that the Soviet Union has developed scalar electromagnetics. Extrapolating from this paper, the amplitude of the "induced longitudinal field" is proportional to the square of the amplitude of the gravitational wave (GW) in an initially isotopic medium, but reduces to a linear proportionality as the medium develops nonlinearities (anisotropies or inhomogeneities). Whereas conventional GW are considered to be weak, the scalar GW may be very strong since we ourselves determine the variation in the energy density of vacuum it involves. Further, sharply spiked scalar EM (EG) pulses may develop appreciably more amplitude that the linearly calculated amplitude from the magnitudes of the zero-summation components. The end result is that (1) we can produce powerful gravity waves using the scalar EM approach, (2) local spacetime can be curved, producing violation of the conservation laws, and (3) the assumption of a local Lorentz frame made by Einstein as a limitation on ordinary general relativity is falsified.

In passing, note that immediately one also has a clear basis for the excess local production of energy when the battery furnishing power to a DC series motor is sharply spiked backwards in a "recharging" fashion. Further, the ions in the battery (which are several hundred thousand times as massive as an electron) possess a resonance on the order of 16 megahertz in a typical lead-acid battery. Under the correct conditions, the ions collect the excess energy (due to the initial square effect) produced by a sharp pulse of recharging potential. The ions are shocked into resonance and, on reversal, "overshoot" and deliver this excess energy to the plates, recharging the battery. Thus such a system can violate linear conservation of energy because spacetime is locally curved by the spiking, and excess energy from the excited local vacuum is collected by the ions and delivered to the plates. This is precisely the effect upon which the first Bedini free-energy motor depended. The timing and adjustment, however, are highly critical and difficult, and the battery takes a rather severe beating.

57. For details and photographs of highly anomalous exhausts seen on Bennett Island, see "Explosive Events Seen on Soviet Island," Aviation Week & Space Technology, Sept. 26, 1983, p. 31. NOAA-6 and NOAA-7 weather satellite photographs of massive smoke plumes, one of them 150 miles long and nearly horizontal, are shown. Also, in U.S. weather satellite photos of Novaya Zemlya, similar plumes have been detected, making this location a tentative candidate for a second Soviet "exhaust site." See also Letter John M. Miller, Geophysical Institute, University of Alaska, Jan. 4, 1985 with attachment, "Bennett Island plume cases recently found." Satellite photos showing actual circular breaking of arctic ice have also been taken near Wrangel Island. See Sam Bishop, "UA photos show Soviets break ice for missiles," Fairbanks Daily News-Miner, Dec. 15, 1984, p. 1,3; Craig Covault, "Soviet Abi1ity to Fire through Ice Creates New SLBM Basing Mode," Aviation Week & Space Technology, Dec. 10, 1984, p. 16-17. Wrangel Island in fact has a bizarre history: it belongs to the U. S., but has been illegally given to the Soviets and occupied and used by them. For details, see "Wrangle over Wrangel and the Weather War," Don Bell Reports, 32(9), Mar. 8, 1985.

58. For details of anomalous blinding (probably by scalar EM means) of U.S. satellites, see Philip J. Klass, "Anti-Satellite Laser Use Suspected," Aviation Week & Space Technology, Dec. 8, 1975, p. 12-13. For details of anomalous failures and recoveries of a British satellite, see Joe Schwartz, Nature, 280, July 12, 1979, p. 95.

59. For other possible evidence of traces of Soviet scalar EM activity over the oceans, see George D. Curtis, "An electromagnetic radiation pattern over the ocean," Undersea Technology, 5(8), Aug. 1964, p. 29-30, 40. In producing a scalar wave, there exists a sort of "signal to noise" ratio of the desired scalar wave to an accompanying, undesired normal EM component. Due to practical equipment imperfections, the generated scalar wave inevitably is accompanied by a small residue of ordinary EM wave. While measures can be taken to further purify the scalar wave, a residue always remains. Curtis may have detected the weak EM residue accompanying early Soviet interference gridding for weather control testing over the U.S.

60. Interactions which occurred between Khrushchev and renowned physicist and Nobelian P. Kapitsa are of interest. Khrushchev desired absolute defense of the Soviet Union so that the Soviets might be able to launch any action desired without risk or serious concern. Kapitsa informed the Soviet leader that, if a means of total neutralization of foreign missiles was to be found, it could only come from a group of new principles in physics which was called " energetics." The term "energetics," of course, was contracted from "psychoenergetics," and is essentially the expanded physics/electromagnetics that this author has dubbed scalar electromagnetics.

Strangely, a portion of a book by Marshal Grechko was deleted from its English translation at the specific request of the Soviet Union to the U.S. State Department. The passage stated: "Of particular importance is basic research aimed at discovering still unknown attributes of matter, phenomena, and the laws of nature, and developing new methods for their study and use to reinforce the state's defense capabilities."

The most definitive Soviet book on military strategy is titled Military Strategy (Voyennaya

Strategiya). See V. D. Sokolovskiy, Soviet Military Strategy, third edition, edited with an analysis and commentary by Harriet Fast Scott, Crane, Russak & Company, Inc., New York, 1975. There are three editions, the latest in 1968. In the third edition, the exotic weapons normally proposed for defense against strategic ballistic missiles -- such as particle beams, high energy lasers, exotic nuclear warheads, plasma weapons, antigravitational weapons, etc. -- are discussed and discounted by implication. The statement is made several times that 100% defense against missiles and aircraft is possible. A statement is also made that it is interesting to note that this capability (a solution to the strategic missile defense problem) has been achieved by the Soviet Union, but not by the West (e.g., p. 298.) In the same edition, a previous definition of antimissile defense obviously consisting of radars, computers, and interceptor missiles is deleted (see author's note 190, p. 454).

The Soviets also recognize the great importance of technical surprise: "The experience of war shows that warring sides quite often have tried to employ new types of weapons to this end to achieve surprise, which are a surprise for the enemy or are little known to him. Each side secretly develops new means of warfare in order to employ them unexpectedly. History knows many examples how the employment of a new weapon initially gave considerable success because the enemy, caught unaware and not knowing the combat capabilities of this weapon, was for some time incapable of effective counteraction… The mass introduction of new weapons into an army usually cannot remain secret for long for the other side. -- New means of warfare, and in mass numbers, are impossible to create often and quickly. In this regard, the sides have begun to take other paths concerning the attainment of surprise along with searches for new weapons and their sophistication." V. Ye. Savkin, of The Basic Principles Operational Art and Tactics, Moscow, 1972, U.S. Air Force translation, U.S. Government Printing Office, Washington, D.C. 20402.

61. The Soviets have not discarded most of their old radars, and have continued to develop, build, and deploy giant new radars (in apparent violation of SALT agreements), which U.S. analysts always analyze in conventional fashion. It is fairly straightforward, simple, and inexpensive to modify an ordinary pencil beam radar to produce scalar wave beams and geometrical scalar Fourier expansion forms. Radars can provide a wide array of totally new and unexpected capabilities when utilized in the scalar electromagnetics mode. Even older Soviet radars, with antiaircraft gun sites and older surface-to-air missile sites, will exhibit new and startlingly advanced capabilities against ballistic and cruise missiles and Strategic Air Command bombers. Since there are many thousands of these older Soviet radars still operationally deployed, in the scalar mode the Soviet strategic defense may well approach 100%, just as Military Strategy implies. Also, the antiaircraft and antimissile defense capability of Soviet land forces may be far greater than anything presently estimated by U.S. analysts.

62. To the physicists, engineers, and researchers who have made the tough sledding to this point, congratulations! Here's a bonus for you.

Turn to page 32, equation [7]. Time differentiate it (this is left as an exercise for the reader). Collect terms and study the result intensely.

Now you have the reason why rise time is so important from a scalar EM viewpoint. You can also see why spiking the innate inductance of any system very sharply is so important. Fast rise time greatly increases the value of the time derivative terms, and your gravito-inertial effects go up with the collective squares of your zero-vector summed components. Any system has at least a little inductance. For sharp spikes, the gravitational potential goes very high. The inductance of the system momentarily opposes any current (translation: opposes any current bleed-off, or EM force field bleed-off). During that time you've got the 4-space G-potential to use. If you need to hold it longer to use it, then increase the inductance.

Now you can see why extremely rapid discharge of properly tuned, giant Tesla coils give gravitational and inertial effects sometimes.

You can see why the sharp "EMP" (electromagnetic pulse) from a nuclear explosion or EMP generating device can penetrate the ocean.

That's why the spikes (such as used by the neurophone) can carry messages directly into the brain and nervous system. You also should see why the Soviet woodpecker carriers (a Soviet over-the-horizon radar system) are loaded with scalar EM spikes (Bedini has precisely detected them from 50 Hz to 10 MHz; Golden has detected them into the gigahertz range.) to work on the humans in the targeted area. They're zapping the hell out of us. And they have done something similar at the U.S. Embassy in Moscow for decades.

Bedini has developed a simple, cheap device to mess up the phaselocking of the scalar EM spikes on the woodpecker carriers. It's based on the fact that the Russkies use a lot of circularly polarized transmitters when they make the spikes. Here's what to do: Take 4 strong bar magnets and an old (but operational) long-playing record turntable. Set it on 33-1/3 RPM speed. Glue the magnets on the top of the turntable, standing vertically, symmetrically placed around the turntable's outer edge. Place them so that on one a north pole is up, on the next a south pole is up, then a north up, then a south up. Now turn the turntable on and let it run. This gadget makes a slowly rotating magnetostatic wave. (Yes, it will also make a rotating "magnetic" wave, which is actually two rotating magnetostatic waves.). These rotating magnetostatic waves will interfere with the scalar spikes, changing them to EM (electromagnetism) and slowly rotating their phase. This breaks the phaselock of the Soviet circularly polarized scalar EM signal on the human brain. This gadget will protect the average large room.

But back to our differentiated equation.

Now you should also see why the first Bedini motor is so "devilishly difficult" to adjust. You've got to get the effect and examine it in well under 20 nanoseconds. You'll need an oscilloscope of 350 MHz range or so. Microwave-quality bench equipment is necessary; the ordinary electrical and electronic shop equipment is not good enough. You can also see why the ordinary lead-acid battery takes such a beating. However, a special battery certainly can be designed to take it. Now you know.

Okay, you tigers. Go get 'em.

63. Now here's a final bonus for the persistent reader.

This time, see J. N. Brittingham, "Focus Wave Modes in Homogeneous Maxwell's Equations:

Transverse Electric Mode," J. Appl. Phys. 54, 1983, p. 1179-1189; T. T. Wu and R. W. P. King, "Comment on 'Focus Wave Modes in Homogeneous Maxwell's Equations: Transverse Electric Mode' [J. Appl. Phys. 54, 1179 (1983)]", J. Appl. Phys. 56, 1984, p. 2587; T. T. Wu and H. Lehmann, "Spreading of Electromagnetic Pulses," J. Appl. Phys. 58, 1985, p. 2064-2065; T. T. Wu, "Electromagnetic Missiles," J. Appl. Phys. 57, 1985, p. 2370-2373. Pay particular attention to the latter paper.

Because of the Strategic Defense Initiative (SDI) and other programs, great interest has recently been kindled in using EM pulses to carry energy from a high-powered source to a target. However, so far almost all the work is still couched in terms of purely ordinary (though sophisticated) electromagnetics. It still involves the EM concept of "EM energy flow through the intervening space from the transmitter to the target." As such, the work to date still suffers from all the self-imposed ills and limitations of using the classical force fields concept of EM.

However, the theoretical work of these referenced papers is of great interest to us here.

In the first paper, Brittingham sparked a great deal of interest with theoretical work which promised a permanently focused 3-dimensional concentration of EM energy that propagates in a straight line with the speed of light. That is, this mode used a newly formulated wave packet (called the "focus wave mode") that acted as a non-dissipating EM soliton.

In the second reference, it was concluded that Brittingham's focused wave packet actually satisfied Maxwell's equations, but not the associated boundary conditions.

[Comment by this reviewer: both of these formulations use "linear time" and a "linear frame," ignoring the local general relativity aspects. Indeed, Brittingham's "focused wave packet" can be made by the zero-vector approach! However, it will not be an "EM" packet, it will be a pulse (or oscillating wave, depending upon how it is made) of 5-gravity potential. As such, it contains a multiplicity of locked-in, 5th-dimensional, non-zero EM gradients in and EM gradients out. Collectively these gradients themselves constitute a deterministic "flux of gradients" or an EM flux in the 5th dimension, with no resultant effective current of EM gradients. In other words, it's possible to build a "flux of gradients" and to have a "current of gradients," just as in ordinary EM it's possible to have a "flux of virtual particles" and a "current of virtual particles." The argument concerning the EM boundary conditions of Maxwell's equations is not relevant in such a case. The only boundary involved is the zero itself.]

Be that as it may, the 4th reference is of direct interest. In this paper, Wu has done us a great service. He shows that, under transient excitation, an antenna of finite size can transmit energy to a faraway receiver so that the energy decreases as slowly as one wishes. An electromagnetic pulse of energy with such slow decay is conveniently referred to as an "electromagnetic missile."

The electromagnetic missile has important applications, according to Wu. One possible application is the transmission of information in a manner that is very difficult to intercept or jam. A second possible application is to destroy or disable a distant target. However, Wu feels that so much energy is required [remember, he is speaking of EM non-zeroed force fields traveling through ordinary space] that a like1y radiating source must be something like a nuclear explosion. However, as Wu points out, a nuclear explosion produces a powerful EM pulse which includes frequencies from microwaves to gamma rays. To weaponize the explosion to produce a coherent beam, the nuclear explosion must be controlled to produce a coherent radiating source and a pulsed laser-like EM beam. Wu points out that such a technology is difficult and has not been developed.

We comment as follows: It is certainly true that enormous energy is required if the waves in the pulse remain electromagnetic -- that is, if they have non-zero resultant E and B fields to the external observer.

On the other hand, if we form a complex of waves in opposition so that we control the energy content of each individual component wave, but blend all the waves so that their collective envelope wave has zero E-field and zero B-field resultants, then -- to the external observer -- the EM energy is infolded and unzipped electromagnetically. It is now electrogravitational energy, and limitations of "EM force field energy flowing through space or through a physical medium" need no longer apply. However, we may still look at each EM component wave as an individual EM wave, moving in the Kaluza-Klein 5th dimension that is wrapped-around each and every point in our ordinary 4-space. In other words, to the ordinary 4-spacetime observer, each component wave has a special channel around every point of his space, not through it. Thus we have established multiple EM channels "just outside" each point of observer space. These we call "scalar channels," and EM energy can move through each one as if it were moving through a wire. The relative potentials between the "ends of each wire" determine in which direction the "EM energy current" flows along the wire.

And in each of these "special channels," we just have a normal kind of space (to the internal observer). Wu's paper assures us that we can get as much of the input energy to the distant target as we desire. The problem then is how to break the zero-summed EM envelope (and hence the "zero channel" for each component EM energy wave) in the distant target itself, spilling the infolded contents of all the special channels back into ordinary spacetime as EM energy (either positive or negative).

That, of course is easily done by interference in the distant zone. As previously stated, there is a way in which we can even perform this interference in a single scalar beam. Since the newly formed EM energy emerges from each and every increment of spacetime occupied by the distant target, it already emerges inside the very vitals of the target. It does not have to translate through the intervening ordinary space between transmitter and target. Instead, it goes through the Kaluza-Klein fifth dimension. It translates along a line through our space, but not through our space. With the scalar EM approach, we transmit EM energy through hyperspace, not through normal space.

Hereafter we will adapt and expand Wu's SDI term "electromagnetic missile" to include these electrogravitational aspects. In other words, we include the "scalar EM missile", which consists of an infolded multiple of Wu's ordinary EM missiles moving together coherently in special sheathed hyperspace channels.

In Appendix I we show an anomalous photograph taken by Bob Gladwin, just before the shuttle launch of Nov. 26,1985. While the photograph may be an artifact, it also is a candidate for the test strike of such a Soviet scalar "electromagnetic missile" tested against an aiming point offset from the shuttle launch which followed a few seconds later. The test was accomplished before the shuttle was launched, and offset from it, so that it positively would not endanger the shuttle itself. The conditions under which this spectacular photograph was taken happened to coincide with the special conditions required to photograph such scalar EM energy. Those conditions are clarified and explained in the Appendix.

My deep appreciation goes to Mr. Gladwin for allowing me to use this photograph. As we will point out in Appendix I, the video camera caught a picture of the Soviet scalar EM marker beacon associated with the Titan explosion of April 18, 1986. One can rest assured that the Soviets probably destroyed that vehicle, using a scalar EM missile.

APPENDIX 1: SOME IMPORTANT INDICATORS AND DESTRUCTIVE TESTING

Some of the indicators which bear on the use of the Soviet scalar electromagnetic grid via the Woodpecker EM carrier beams are given in this appendix. along with other information believed pertinent to overall scalar EM effects.

IN ADDITION, THE THESIS OF ACTUAL LIVE TESTING OF THESE WEAPONS BY THE SOVIETS -- TO INCLUDE THE DESTRUCTION OF A TROOP-CARRYING U.S. AIRCRAFT, A SHUTTLE, AND A TITAN MISSILE -- IS ADVANCED AND INDICATORS PRESENTED.

An anomalous photograph taken by Bob Gladwin is presented which possibly may have captured the "smoking gun" -- the strike of a Soviet scalar EM missile at an offset aim point near the shuttle Atlantis, just before its launch on Nov. 26, 1985. However, this photograph is caveated since it may also represent an anomalous but ordinary artifact. captured in some as yet unknown manner. The manner in which such a photograph may have been taken -- when normally the scalar energy does not show on a photo -- is explained.

Another photograph taken by George Suchary after the launch did capture the Soviet "marker beacon" utilized in the test firing, however. This independent photograph is presented, and adds credence to the thesis that the Gladwin photograph captured the test of a Soviet scalar EM pulse missile.

- INDICATORS AND SUCH -

1. Continual cloud radials have repeatedly been seen and photographed over Los Angeles, California or its vicinity.

Los Angeles is also apparently a "hinge point" for bending and controlling the jetstream. Three typical radials seen to form one after the other by Margaret Wilson (and sketched and airbrushed by her) in the greater Los Angeles area are shown in figures 1, 2, 3, and 4. These radials were formed on Sunday, Jan. 26, 1986. Ms. Wilson had heard my Saturday night "Soviet Weather Engineering" broadcast over KABC Los Angeles, and my request that anyone who saw the typical radial cloud formation should photograph or draw it, noting time and location, and send the photos or drawings to me.

As Margaret described it. "To my astonishment, the next day I not only saw one, but three in the sky at the same time.

The time of the first one was Sunday, Jan. 26, at 1:15 p. m.

. . . I stepped out the back door for some minor errand and noticed a fan of long straight clouds in the sky. The sky had been filled with dark clouds most of the morning but this had been breaking up and the western half of the sky was now clear with these lines of clouds spread out on it. I immediately thought of your request and tried to see the bottom of the fan which was hidden behind bushes and a garage. When I stepped into the street, I saw there was a semi-circle beneath them.

Not having film and a wide-angle lens on my camera. I knew I couldn't capture it in its entirety. I am, however, a retired commercial airbrush artist so I quickly sketched the formation. It is illustration #1.

It was in the western sky, the base of the central semi-circle sitting about 15 degrees above the ocean horizon. It was very clear and sharply defined. The semi-circle was broken at the top and the two broken ends curled inward like commas. As I sketched the wind deformed it somewhat, the right half of the whole formation sliding down away from the left.

Ten minutes later it was getting quite clear of clouds except for a bank over the Santa Monica Mountains to the north and there, straight north, another fan was forming. As I watched it took shape -- illustration #2. The same broken semi-circle (with commas') but this time the bases of the rays were feathered about half-way out with end halves long straight wisps.

As I sketched, another fan had begun to take shape further east. It was hard to tell its size but I had estimated the first one at about 15 miles across and the second a little larger and the third larger yet. At the peak of its formation it looked like illustration #3. ...There in the middle of the central circle, hazy and a little faint was what looked like the top of a mushroom cloud. Or was it only a distant thunderhead that coincidentally found its way to a place behind the formation? The arms of the last, largest formation were rather wavy and feathery like some gigantic sea creature. It did not disperse as fast as the others, but held for a couple of hours, though it was pretty distorted by then. I could not estimate its size as I wasn't sure of its altitude.

Were they aircraft contrails? Certainly not! There was an airliner in the sky at the time -- a mere scratch on the smooth blue sky (see illustration #3, right center.).

The three formations were present in the sky at the same time although the first was getting quite distorted by the upper level winds before the last reached its peak.

Tuesday our weather changed from hot and sunny to rainy as the weather man said the path of the jetstream altered, and then there was the space shuttle explosion.

...The Russians seemed very sincere in their sympathy for the loss of our astronauts but only two days later it was printed in TASS and I personally heard Vladimir Posner say that Russia hoped that the United States would now realize what could happen if the S.D.I. program was further pursued. Was this friendly concern or a threat? It is very evident that if the S.D.I. can destroy a missile in flight, it can also destroy one on the ground. This would immediately put all Russian missiles in [reach of] our arsenal. They are already in place in the most desirable target areas -- their own launch sites. I'd be panicky too if I were a Soviet leader."

[Ed. comment: My sincere thanks to this thoughtful lady who took the time to sketch the cloud radial phenomena occurring in the Los Angeles "jetstream pivot point" at this critical time, just prior to the launch and destruction of the Challenger.

Anomalous cloud grid activity (formation and dissolution) in
the Los Angeles area has been positively correlated with Soviet Woodpecker activity by engineer Ron Cole, with 100% correlation in 30 straight Woodpecker measurement runs. Statistically, of course, this
is quite decisive.].

2. Continual cloud radials have repeatedly been seen and photographed over Huntsville, Alabama, associated with weather engineering and grid activity. Typical radials seen by this author are shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. Huntsville is apparently a pivot point or "hinge point" for bending and controlling the jetstream.

3. Spectacular cloud grid patterns have been seen over Huntsville, Alabama and Los Angeles, California. An excellent videoclip of one of the Los Angeles area grids, personally observed and shot by KABC Open Mind talk show host Bill Jenkins, is shown in Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985. An artist's sketch of an earlier gigantic grid, extending from horizon to horizon in every direction, seen over Huntsville, Alabama by Tom Bearden and Ken Moore, is also shown in the tape. The videotape is available from P.O. Box 1472, Huntsville, AL 35807. Sketches and details of another highly anomalous grid pattern observed in the greater Los Angeles area by engineer Ron Cole are shown in figures 5 and 6.

The reason that clouds of water droplets and/or ice particles detect the scalar interferometry and form signature patterns is simple: Consider each H_2O molecule as having two light little hydrogen atoms hanging on to the much heavier oxygen atom. The covalent sharing with the oxygen atom of the electron from each hydrogen atom means that (1) the electrostatic scalar potential between the H ion and the O ion it is bonded to, is rhythmically varying as the electron is shared back and forth, (2) this varying potential contains "electron spin holes" since it is made by the moving electron, (3) two such varying potentials exist since there are two H atoms sharing covalent electrons with the O atom, (4) the two H ions are at an angle of over 100 degrees with respect to each other, (5) the molecule assembly thus constitutes one part (one half, so to speak) of a scalar interferometer with imbedded electron spin holes for electron hooking, (6) Incident scalar waves from outside the system interact with the "half scalar interferometer." This scalar interferometry interaction is coupled to the covalent bonding electron because of the internal spin hole pattern of the molecular half of the interferometer. The coupled electron moves with the scalar pattern's changes, causing an observable interaction with and in the electrical structure and ionic potential of the molecule, (7) the weak H to H hydrogen bonding between molecules connects resulting ionic potential changes to the structuring of the entire macroscopic cloud assembly. Thus by scalar interferometry with the H_2O "half interferometers", an incident scalar EM pattern is detected and translated into patterns of E and B force fields. The H_2O molecules then "line up" in accord with the electrical patterns detected.

Thus the clouds form "signature patterns" according to the incident scalar EM radiation patterns. Giant cloud radials in the targeted area are probably due to the type and shape of antennas -- including the type and shape of the electrical wires establishing the antennas' ground planes -- 640used by the Soviet Union to project the scalar waves. Note that such "radial" antenna patterns were occasionally used by James Harris Rogers in his underground and undersea scalar EM transmission system.

[For Rogers' patents, see U.S. patent numbers 1,316,188, Radiosignaling system, Sept.16, 1919; 1,322,622, Wireless Signaling System, Nov. 25, 1919 (figure 6 of this patent clearly shows a "radial wire" ground plane antenna); 1,349,103, Radiosignaling system, Aug. 10, 1920 (the fourth version in figure 1 shows inside/outside variation which, for the long waves used, to a distant observer appears as a varying magnetostatic scalar potential using opposing B fields); 1,349,104, Radiosignaling system, Aug. 10, 1920; 1,220,005, Wireless Signaling System, Mar. 20, 1917; 1,303,729, Wireless signaling system, May 13, 1919; 1,303,730, Radiosignaling system, May 13,1919; l,315,862, Radiosignaling system, Sept. 9, 1919; also 958,829, Method and Apparatus for Producing High Frequency Oscillating Currents, May 24,1910.].

4. Anomalous power outages in Ventura County, Southern California on Nov. 8 and 9, 1985. Again, engineer Ron Cole, who lives in the affected area, investigated and reports as follows:

"On the evening of 11-8-85 in a rural area of Ventura County known as Carlisle Canyon, a strange electrical phenomenon began to occur at about 10:15 p.m. Our first observation was that random white dots began to appear on the TV screen, then began to increase in size and brightness. The brighter the dots got, the dimmer certain lights in our home became. Yet later we found that, when some lighting circuits were dimming, other circuits were causing lights to shine much brighter than normal. After a bit of checking, we were able to determine that lighting circuits common to pole 1 of the service transformer were the ones that were seeing a reduction in voltage, while all circuits on pole 2 were seeing an increase in voltage.

There is another interesting condition that was observed on the TV screen. The dots were in sync with the vertical field or frame rate of the signal. This means that the interference had to be of a frequency of 59.7 Hz or 29.85 Hz. Normal power line interference will form a grouped dash pattern across the screen and will strobe upward at the rate difference of 59.7 Hz or 29.85 Hz. In the 8 November case, there was no strobing of the dots whatsoever.

These conditions continued off and on for the rest of the entire evening and were observed by no less than six other persons that I know of. They were observed over the entire TV frequency spectrum, low band VHF, high band VHF, and on two UHF channels. Line interference is almost always most prominent on the low band VHF frequencies only.

These anomalous events would take place on a somewhat irregular time pattern; anywhere from 3 to 10 minutes apart. They would always last for about 2 minutes.

At 12:30 a.m. I went to bed and was not aware of any further activities until 6:20 a.m., at which time I was awakened by several sharp snapping or cracking sounds coming from the walls of my house. I rolled over and went back to sleep, only to be reawakened about 15 minutes later by my wife, who informed me that the power and phones were out. My comment was, "Some dummy must have crashed into a power pole," and I dozed off again.

At 10:15 a.m. the power and phones came back on line. Shortly after, our phone began to virtually ring off the hook. About a dozen neighbors up and down the canyon wanted me to check out their electrical wiring and various appliances. All of them were trying to explain their individual encounters and what took place when the power went out. Such things as wires snapping in walls, singing and humming sounds coming from wiring in walls, strong smell of ozone, light bulbs exploding, buzzing coffee pots, and a handful of other phenomena all took place just before the power went out at 6:20 a.m.

A typical example was that of Rob Steel. [Ed: Name changed for this report; his real name is in my files.] Rob lives four miles down the canyon from us. He had just turned on the lights in his barn and was getting ready to feed the livestock, when he realized he had left several needed items in the garage located about 75 yards from the barn. Upon entering the garage, he saw the garage lights began to flicker, then go to a dim orange glow. Upon stepping from the garage area, Steel saw that lights in the barn were so bright that he thought the entire structure was about to explode. He froze in his tracks for a moment and his whole house and garage began to hum. Then everything went out.

An interesting point to inject is that Mr. Steel's home and several others farther down the canyon from us are being fed from a different 19 kv power feeder source than the 14 kv feeder that supplies us and most homes in the canyon.

At about 2 p.m. on 11-9-85 I was in contact with the power utility company (Southern California Edison Company) to inquire as to the cause of the power outage. My first assumption was wrong. No power poles had been knocked down; only a ceramic fuse in the 19 kv feeder system had blown and a two-pole oil breaker had kicked out on the 14 kv feeder system. Load charge recorders at the power company show that both fuse devices kicked out at exactly the same time, but they are in no way common to each other as to load distribution. They only supply power to adjacent areas.

Also, the phone company has no record of its service being out, yet the phone service was restored when the power was restored. The phone company does not use Edison Company's power to power its phones. Again, the only commonality is that the phones and phone lines were in the same area.

Numerous magnetic breakers, used as manual switch devices for well pumps and other services, were on when power was returned. These switches had previously set in their off state.

A strong electrical-type ozone odor was detected at five locations at the time the power went out.

Later we found that a Research and Development laboratory in a nearby area measured 29 Hz intermittent signals late on the night of 11-8-85. These signals were interfering with some critical tests the lab personnel were trying to accomplish. The laboratory itself did not experience a power outage.

At this point in time, the only realistic surmise we can come up with is that a tremendous ELF or scalar ground wave was adding and subtracting to the grounded neutral service side of the power pole transformers so as to sum in and out of phase. When peak nodes were achieved, all line transformers ended up back-feeding the feeder lines out of phase, causing the final power outage.

Twenty-one persons are aware of the strange conditions of the power outage and prior events. [Ed: Their names are in my files.]"

Editor's comment: Another anomalous power outage also occurred in the same area. This occurrence is possibly associated with effects of the scalar interference grid adjustments. If the Soviets were preparing for a major test of the grid in the Launch Phase ABM mode, substantial tests and exercises may have been performed before the actual "wet run" against the shuttle launch on Nov. 26, 1985 -- about two and a half weeks after the anomalous power outage at Carlisle Canyon.

See the Rogers' patents, referenced above, for ideas of how scalar EM signals couple to the power lines and telephone lines.

As this book goes to press, John Bedini has notified me that his new and very sensitive developmental scalar detector positively verifies scalar frequency "spike" activity on the Woodpecker grid (and possibly the ground reference) from 50 Hz to 10 MHz. A normal instrument will definitely not see these signals. John's lab is at Sylmar, California and about 70 miles from Carlisle Canyon.

5. During the Skylab mission, NASA astronauts observed and photographed a mysterious green glow of the atmosphere. The glow was over 1,000 miles long and over 100 miles wide. From Carlisle Canyon, Ventura County, California the sky has also been observed to possess a mysterious green glow from horizon to horizon, on three occasions. [Editor's comment: It is hypothesized that this green glow is associated with grid activity in the Launch Phase ABM System mode.]

6. Just prior to the hapless launch of the Challenger on Jan. 28, 1986, the Soviet Union sharply bent the jetstream down through the center part of the country, then curved it eastward across the Florida panhandle. This exposed the waiting Challenger on the launch pad to severe cold air, brought down from Canada by the sharply deviated jetstream. The use of the scalar interferometry aspects of the Woodpecker grid for weather engineering over the U.S. has been documented previously. [Bearden, Soviet Weather Engineering Over North America, 1-hr. videotape, 1985;

Bearden, "USSR: New beam energy possible?", Defense & Foreign Affairs Daily, 13(111), June 12, 1984, p. 1-2.]. Wilson's cloud radial sightings on Jan. 26, 1986 and Cole's giant cloud grid sightings on Jan. 24, 1986 are directly connected with adjusting the Woodpecker grid and deviating the jetstream, to carry cold air to the Florida launch site.

7. The previous testing of the grid in the Launch Phase ABM/Anti-bomber Mode, using actual jet aircraft as "wet-run" targets, has been demonstrated by various anomalous incidents.

As one probable example, see "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," AP release, Huntsville Times, Feb. 20, 1985, p. 1 and "China Airlines Pilot Denies Cockpit Error," AP release, Huntsville Times, Feb. 21, 1985, p. A-9. In this incident (see figure 7) a Boeing 747, flying at 42,000 feet and bound for Los Angeles, experienced unusual engine flameouts while still out from San Francisco. In a hair-raising plunge toward the earth, the pilot fought to restart the engines and regain control of the aircraft. After a fall of some 32,000 feet in two minutes, the engines finally restarted, and the pilot was able to fly into San Francisco and land safely.

We have already explained how the Soviets are able to track the aircraft's jet engines. Later in this Appendix we show just how accurately an actual "scalar EM missile" may be focused by the Soviets on a hapless target. If a negatively biased pulse or continuous wave scalar EM "energy extraction" beam hits the engines, the extraction of energy will result in engine flameout. (No heat energy, no combustion of the fuel.). This is apparently what happened to this aircraft. In the presence of "splatter" from the extraction beam targeted at the engines, the aircraft instruments were seriously affected. Different instruments have different responses to such an anomalous scalar EM kindling effect. Thus a signature would be that the instruments disagree with each other, and need not record what actually occurred. This appears to have been the case, constituting a signature that scalar EM was targeted against the aircraft. The pilot's instruments apparently indicated that the control system had failed, while the flight recorder apparently indicated that it had not.

Let us assume this was a Soviet test. When the distant Soviet operator turned off the extractor beam, the engines returned to normal state after a "discharge time constant" period. They normalized before the plane struck the ground, so the pilot was able to restart the engines and proceed.

Notice that this was a limited test to cause engine flameout only. If the scalar EM missile had been used against the aircraft in the "pulse" mode, it would instantly have exploded in mid-air.

The crash of an India jet aircraft outside Great Britain may have been a test of the same weapon in the pulsed "destroy aircraft" mode, since there are indications that the aircraft suffered an anomalous midair explosion. Another candidate for such a possible "shoot-down" of a jet aircraft is an (Italian) Itavia DC-9 that crashed into the Mediterranean in June, 1980. Both Itavia and the Italian Transportation Minister Salvatore Formica believed that a missile struck the jetliner and was the likely cause of the crash. It may have been a scalar EM missile and a full-up test by the Soviet Union. See "Plane was probably hit, judge says," Birmingham Post-Herald, Jan. 17, 1981.

Also the present author has knowledge of at least one instance in the early 1980's when an American-made tactical ballistic missile was anomalously destroyed shortly after being launched by troops in Europe. The destruction was anomalous and a large hole was inexplicably burned through the side of the missile -- exactly as if it had been struck by a scalar EM missile test in the anti-tactical missile (ATM) role.

8. As another example of airborne scalar EM activity in the Woodpecker grid, in January 1985 a U.S. Navy-dispensed chaff cloud, dispensed well-offshore from San Diego, California was caught by an unexpected wind and blown toward the city. (See figure 8.). As the chaff cloud moved in, significant failures of electrical systems and electrical components resulted. Power was disrupted to as many as 60,000 homes by the incident.

The sudden, unpredicted wind was suggestive of a wind blowing toward an artificially created low pressure area. Under certain conditions, a specialized structure such as that of a piece of sophisticated chaff will reflect scalar waves of frequencies within its cut bandwidth. If scalar modulation frequencies within the chaff bandwidth were present on the Woodpecker carrier grid above and surrounding San Diego, a myriad of reflecting bits of chaff in the moving chaff cloud would produce myriads of random, invisible "fireflies" of EM energy kindled at a distance, from random scalar interferometry, in a zone surrounding the cloud. As these invisible "firefly" pulses of

EM energy pulsed inside components of the electrical systems and electrical controls in San Diego and vicinity, extensive and random internal electrical interference occurred. This lead to widespread, anomalous electrical failures throughout the city.

Note that this activity preceded the huge Soviet scalar EM exercise of April-May 1985. The grand exercise would certainly have been scheduled well in advance by Soviet strategic forces controlling the scalar EM howitzers and Woodpecker grid weapons, and adjustments and smaller tests would have been being made for some months before the tests.

9. Other anomalous "electrical strikes" certainly exist, suggesting Soviet testing may occur widely throughout the world, yet be reported as ordinary but anomalous phenomena. An example which comes readily to mind is the anomalous blast (or blasts) that struck Bell Island, off the east coast of Newfoundland, on April 2, 1978. Persons on the shore of Conception Bay reported seeing fireballs coming down out of the sky and hitting the island, according to Royal Canadian Mounted Police. Highly anomalous atmospheric electrical activity preceded the blast. This may have been Soviet testing of scalar-created electromagnetic missiles. Extensive information on this incident can be obtained from P.A.C.E., 100 Bronson Ave., Apt. 1001, Ottawa, Canada T1R 6G8.

As another possible example, villages in an area of West Donegal in Ireland have been terrorized by freak "lightning bolts" for over two years. Roofs have been literally stripped off, windows smashed, telephones and television sets blown up, and pipelines wrecked. On Dec. 25, 1984 a mighty blast caused a three-day blackout: it accompanied a snowstorm, the first "white Christmas" in the area in 20 years. See "Bolts from the blue," Fortean Times, Issue no. 45, Winter 1985, p. 5.

10. Scalar spikes or pulses from the Soviet testing may also cause other anomalies, including deaths of unsuspecting persons using the telephone. For a candidate event of this type, see "Killer Phones," Fortean Times, Issue no. 45, Winter 1985, p. 8-9; also AP release, Huntsville Times, May 30, 1985; AP newswire June 2, 1985; and New York Times, Sep. 11,1985.

On May 21, 1985, five days after passing a complete physical examination, 17-year-old Jason F. Findley of

Piscataway, New Jersey was on the phone at his grandmother's house, talking to his girl friend. An odd click occurred on the line, followed by a gasp from Findley and then only the TV playing in the background. His grandmother found him unconscious, still clutching the phone in his hand. Rushed to Muhlenberg hospital, he was pronounced dead shortly afterwards. An autopsy could not determine the cause of death, and the telephone company found the phone working properly, properly grounded, and no evidence of excess electrical charge. Findley was lying on a wooden bed which would not conduct electricity. He had a hemorrhage of the inner ear. There had been an electrical storm over Scotch Plains, New Jersey the night Findley died.

AP wire services were told by the special forensic investigator with the New Jersey medical examiner's office that about six people had died in the U.S. in similar circumstances. Others had reported being knocked unconscious by "high voltage shock" from their telephones. Several weeks before Findley's death, another man from Whitehouse Station, New Jersey had been found unconscious with a telephone in his hand. According to the Consumer Product Safety Commission in Washington, D. C., in 1984 nearly 12,000 people in the U. S. were taken to emergency rooms with injuries related to telephones. A hundred of them died, although exact details were not revealed.

Note that the giant Soviet scalar exercise of May 1985 had been detected and verified by Golden. Substantial activity in the Soviet weapon complex, and in the Woodpecker grid over the U.S, continued throughout May and beyond. A phone line may by chance detect one or more atmospheric scalar EM pulses, just as occurred in the blackout in Ventura County, California on Nov. 8-9, 1985. These scalar pulses can travel along a phone line and be broadcast into a person's ear and brain from the telephone earpiece. The proper scalar EM pulse can induce both "acoustic shock" in the inner ear and "electrical shock" conditions in the nervous system. The pulse may also interrupt the body's electrical timing pulses to the heart, causing the heart to stop without physical damage. The scalar EM shock to the brain and nervous system can knock the person unconscious or, if sufficiently strong, kill him instantly. The condition is worsened if the person is ungrounded, since then he acts as a large, open-ended capacitance in series with the telephone instrument's inductance. In such case, the scalar resonance of the equivalent series LC circuit can be sharply excited and resonated. By the nonlinearities of the body and nervous system, the scalar resonance can be translated into an ordinary electrical pulse or resonance. Since the scalar power is not translated into ordinary power except inside the nonlinearities of the human capacitor/nervous system, no damage to the telephone or phone lines need happen. [At the ComTec exhibit in Las Vegas in 1984, John Bedini demonstrated powering a 200-watt speaker system through a 100-foot length of #40 magnet wire, using scalar EM means. He put some 200 "watts equivalent" of scalar energy through the tiny wire to the speakers, where a small Bedini translator translated it back into ordinary electrical "force field" energy, powering the speakers. This was much to the astonishment of the technical audience, since the wire was quite cool and exhibited no evidence of heating, even though by conventional thinking it should have been instantly melted.]

The bottom line is: With increased Soviet scalar weapons activity over the U.S., the telephone system is likely to become more hazardous to humans using it, when the instrument is connected to the external wire. A radio phone is much safer against anomalous scalar EM "spiking."

11. The metal-softening detected and verified by Frank Golden on Jan. 1, 1986 is certainly indicative of a new mechanism or effect added to the Woodpecker grid. The highly specialized nature of the particular detector/amplifier which Golden was using provides strong evidence that the signal was actually derived by recording the scalar EM signals produced by a human cerebrum (two cerebral hemispheres connected by the corpus callosum, forming a scalar interferometer) when the human was successfully performing psychokinesis (metal-bending). The fact that Golden was able to nullify the signal locally by special methods worked out by him lends further credence to this hypothesis. In any event, three hours of very objective experimental data strongly support this thesis. It is therefore indicated that (1) the metal-softening signal had been added for a purpose, and (2) something made of metal was going to be softened, leading to its failure. Certainly the Challenger would have been an ideal target; softening the metal on its boosters and fasteners by even a small percentage would significantly reduce the vehicle's ability to withstand the

stress during launch. This of course is circumstantial evidence only, but the finger of suspicion is there nonetheless. It is a fact that the metal-softening signal was and is in the Soviet Woodpecker-repertoire; it is a presumption that it was intended for use against the Challenger.

- FINAL DRESS REHEARSAL BEFORE THE BIG SHOW -

12. The shuttle launch from Cape Canaveral on the evening of Nov. 26, 1986 seemed smooth as silk to the army of official observers and onlookers. In reality the Soviet Union clandestinely conducted one or more Launch Phase ABM System tests against the vehicle, using it as a convenient target.

Figures 9, 10, 11, and 12 are included to give the reader some preliminary background on the highly significant events which occurred in conjunction with this flight, and previous to it. These figures are newspaper articles from the News Tribune, Fort Pierce, Florida. The articles are used by permission.

Anomalous atmospheric "booms" and airquakes had occurred in the Florida area for over three and a half years prior to the launch. Sixteen counties, most of them on the northern Gulf coast or close to it, had reported booms during the two-year period preceding April 1984. State and local emergency preparedness officials and U.S. Air Force Base representatives met but could not resolve the mystery of what was causing all the anomalous booms.

In fact, many booms had occurred in and around the gulf coast of Florida, particularly offshore from St. Petersburg, starting from several years earlier.

These anomalous booms were caused by the Soviets adjusting and testing the Woodpecker scalar EM grid for weather engineering and potential weapons use. The booms and airquakes were certainly not confined to Florida alone, but also occurred at many other locations throughout the U. S. and Europe.

Mysterious, highly localized seismic microquakes also had occurred at widely scattered locations throughout the U. S.

On at least two shuttle flights prior to the Nov. 26, 1985 launch, large anomalous "booms" had occurred over or around the launch site after launch. No one in NASA knew what they were or what caused them.

So prior to Nov. 26, the Soviets had already prepared and tested the Launch Phase ABM system mode of the grid against at least two U.S. shuttle launches. However, so eerie and strange are the scalar EM weapons that they had been able to do so without the U. S. government even becoming aware of what was going on. In the Kremlin, they must have marveled at our continued ignorance.

At any rate, they tested the Launch Phase ABM system again on Nov. 26, and this time private U.S. citizens happened to capture at least some evidence of what was going on. It still would have died there without surfacing, except for a curious chain of events. Let me briefly recount those events.

****

Lucius Farish publishes a fine little monthly newsclipping collection called UFO Newsclipping Service (Route 1, Box 220, Plumerville, AR 72121). In it she includes clippings on all sorts of anomalies, including UFO's, strange light phenomena, etc. Because of my long work in UFOs (for some time now, placed on the back burner due to the press of my work in Soviet scalar electromagnetic weapons), I've subscribed to Farish's service for several years.

And this time my unorthodox interests certainly paid off. Farish published a clipping of the article, "NASA officials stumped by strange light, boom," by Susan Burgess, News Tribune writer (see figure 10). The moment I read the article, I realized that the Soviets had actually tested the Launch Phase ABM System's area mode against the shuttle launch of Nov. 26, 1985. Further, George Suchary had taken a picture of the marker beacon as it was slewed away at the conclusion of the test.

With alacrity I contacted Ms. Burgess and, through her, George Suchary. Both were very helpful, and Mr. Suchary gave me permission to use his important photograph of the light moving away after the test.
Ms. Margery Bril , the kind librarian at the News Tribune, sent me four clippings.

After reading the article by Julie Enders, Showcase Editor of the News Tribune, I contacted her also. Although extremely busy, Ms. Enders kindly gave me additional details, including the fact that Bob Gladwin had taken a series of eight stills of the launch itself, and in one frame an anomalous, curving line of light with a splash on the end appeared.

At first I thought it might be the answer to a dream. Very tentative information suggested that the "curved light with a splash on the end" might have occurred in one of the pictures taken during the launch sequence itself, with the light striking directly under the shuttle shortly after it lifted off. I was positive the Soviets could strike at the shuttle (or a missile) with a "scalar EM pulse" missile, but hardly dared hope someone could have actually been lucky enough to get a photograph of it.

From my acquaintance with two masters of paranormal photography -- the indomitable Trevor James Constable and a close friend, Joe Gambill -- I understood how one could possibly get a picture of such a strike -- assuming it occurred -- at night, when the same picture could not have been taken in the daylight. I also understood why the camera might catch it when the human eye would not see it.

It seemed that, between them, these friendly, cooperative persons just might have gathered together the positive evidence to prove that the Soviets had struck at the shuttle with an offset test of a scalar EM missile, using a marker beacon for adjustment. (As it turned out. it was not to be as definite as that!) However, time was of the essence.

The Soviet tests had definitely included a test in the "area burst" mode, which can catch and destroy several missiles fired from a deployed silo complex at once. The clinching evidence I had needed for so long on a Soviet test of the pulse mode -- the scalar EM pulse missile -- just might be there in one of Bob Gladwin1s photographs. But there was no time to lose; such evidence is fleeting. and unless I moved quickly, it was likely to fade away and disappear forever. With persistence, the rest followed swiftly. Mr. Suchary graciously gave me permission to use his photographs. These of course prove that the marker beacon was there. (In addition, it was seen by numerous persons.). The News Tribune gave me permission to print the articles. Mr. Gladwin furnished his negatives and gave me full permission to analyze and use his photographs. Personal friends, Everett and Shirley Edwards, printed the photographs in their private darkroom so the negatives would not be out of safe hands for even a moment.

At the time, this book was almost finished. If the photo of the candidate Soviet use of the electromagnetic missile was accurate, quick work would be necessary if it was to be included. My friend and publisher, John Ratzlaff. had already agreed to swiftly publish this large output as soon as I could get it to him. At Omni Video Publishing, Ms. Electra Briggs was alerted to the potential importance of what was coming. She and Mr. Glenn Shoemaker swung by my house from New Orleans, and a quick, informal videoclip was taped, announcing the book. (Murphy's law -- if anything can go wrong, it will -- got me there. Inexplicably, in my near-total fatigue from months of 19-hour days and 7-day weeks on this project, I spoke the date of the Titan explosion as April 19th, knowing full well it was the 18th! It happens. Electra suggested she slap a correcting statement over the video. Done. Be ready to get it out. It's too important; don't sweat the small stuff.)

Zerox copies of the photos were received. There in the fifth photo (or so it seemed) was an unequivocal strike of the scalar EM missile, in the launch sequence itself. The assumed strike seemed to have occurred right under the shuttle, a few seconds after liftoff.

The evidence seemed complete. All that was left was to receive the actual negatives and verify the sequence. At last we seemed to have the smoking gun. The Soviets would have been caught redhanded.

Based on the advance zerox copies, the manuscript was altered. Everything was ready. And then the negatives came in.

In this business, you learn to meet and accept a lot of failures. Things often go astray, and don't wind up as they seem at first blush. This time was no exception. Murphy got us again! The photos on the negative were in a different order then they had been on the zerox copies.

There's no mistaking the order in which the pictures were taken. There they are on the negatives, one right after the other, and clearly numbered. And the curving light with a splash at the end did not occur during the launch sequence at all. Instead, it was taken before the launch -- just how long remained to be seen. Everett and Shirley took the photo and developed it at both extremes: very dark and very light, to bring out all the details possible.

The results were most interesting. As shown on the lighter print (figure 13), the light does not curve down out of the sky and strike toward the ground, it turns up toward the sky and strikes above the distant treeline. Further, as shown on the darker print (figure 14), it does consist of a little ball of light, just like a scalar EM missile would appear. No apparent structure to hold the light appears in either picture, though one could not positively rule this out completely without microdensitometer tests. (One can inadvertently get artifacts like this by focusing on a ball-like light bulb, for example, and then moving the camera while the shutter is still open. On the other hand, there is no evidence that this happened at all, and no street lamp post or other foreign object in the photo.)

(The remaining photos of the liftoff sequence are shown in figure 15.).

We called Mr. Gladwin. He verified that the photos had been taken in order, one right after the other. He had not had the camera on a tripod, but had steadied it on a railing. In other words, the "light streak with a burst on it" had been taken just before the liftoff. Also, Mr. Gladwin had not observed 'the marker beacon, and had not attempted to photograph any such "light in the sky." In fact, visually he had not observed any such streak and burst of light in any scene he was photographing.

So the question is, exactly what did Mr. Gladwin photograph in that photo, just prior to the shuttle's liftoff? The answer is. we just don't know for sure.

Let's run through the two possibilities:

(1) He may have snapped an artifact while moving the camera. If so, the camera was definitely quite still at least most of the time the lens was open, for the treeline is quite clear in the lighter photo, which reduces light saturation and shows details more clearly. Further, there is no evident support or structure attached to the light itself; it seems to be a "disembodied" light. So this possibility is ambiguous, though reasonable. Mr. Gladwin does not recall any such movement of the camera; it was steady while he was shooting.

(2) He may still have snapped a scalar EM missile. If so, it means the Soviets fired a "small test" of the exothermic pulse or scalar EM missile in advance of the actual shuttle launch. Let's pursue that for a moment, and suppose it to be true.

The Soviets already had the marker beacon up in the sky -- that's certain. Many persons saw it. Soviet trawlers off shore and Soviet satellites could also have given the actual position of the marker beacon. Thus the Soviet LPABM System would have had (a) the actual parameters the computer had used to transmit the beacon, and (b) the actual location at which the beacon emerged. Comparison of (a) and (b) is much like having a "meteorological data message" in field artillery: One has the corrections to apply for specific conditions of the medium that cause deviation.

These corrections can then be applied to calculated or "table" data to adjust firing for real-world conditions.

So far so good. The Soviets had the exact position of the registration beacon, and could apply corrections for a shot of the scalar EM missile. They certainly would know the exact position of the shuttle on its launch pad.

If a fairly substantial offset for the scalar EM missile shot were used, "mett corrections" could be applied and a small scalar missile could be zinged in, stopping momentarily at the final offset "strike" position so that a track could be obtained by the scanner extraction beam. Then by comparing the detected offset position of the "burst" of the scalar EM missile to the detected offset position of the marker beacon, the difference could be applied to the beacon's true position to get the true strike position of the missile. The offset used could then be subtracted to give the location where the missile would have struck without the offset, had it actually been fired at the rising shuttle after liftoff. This "adjusted final strike position" could be compared to the shuttle's expected early trajectory (which varies only slightly from standard for the first few seconds) to determine the expected combat success of the shot. It's a little roundabout, but it's a very safe way to do it if one wants quite positive assurance that the missile will not strike the shuttle and destroy it. Since the Soviets delayed the "area burst" and its concomitant atmospheric "boom" to T + 12 minutes, it is evident that a large offset was used for the area strike -- probably about 10 extra minutes -- to positively assure that there would be no actual damage or threat to the shuttle.

It seems logical, then, that they might fire the small scalar EM missile well offset from the shuttle, and just before its launch, to prevent any possibility of actual damage to the shuttle.

So there you have it. Mr. Gladwin's photograph is consistent with such a careful, safe scalar EM missile test by the Soviets, prior to the actual liftoff of the shuttle. On the other hand, it's not possible to say with certainty that that is what it is. It may simply be a very peculiar artifact. In fac, if it were not for the other indicators, one would most certainly not consider such a bizarre possibility as an offset Soviet scalar EM missile strike at the shuttle.

If the "streak" photo is examined as a candidate for a sca1ar EM missile test, then one other thing must be explained: Why did the film capture the test when observers did not see it at all? That, of course, we can explain!

So back to our presentation. Here I'm going to assume -- and accent that this is a postulation -- that the Gladwin photo actually represented a scalar EM missile test. We will run through a possible sequence of events, based on that assumption. We will hypothesize the LPABM system operation and the Soviet conduct of the test. The major events here in the U.S. may be quite real; those assumed to occur in the Soviet Union, of course, are "author's license" -- but hopefully realistic.

****

The evening of Nov. 26, 1985 is perfect. At Cape Canaveral, the shuttle Atlantis is poised on its launch pad waiting to move majestically into the heavens on its impending journey. As launch time approaches, everything is as smooth as a cat's whisker. There is, however, just one small anomaly.

A mysterious light is hanging up in the sky, staying still for awhile, then gently rocking back and forth. The light is a Soviet marker beacon for a scalar EM howitzer. The howitzer, employed in a special scalar EM channel contained in the Woodpecker transmitters and interference grid, will shoot through the EM carriers at an aiming point offset from the shuttle.

Since this is still to be a nondestructive test, no chances will be taken. The pulse mode will be used, and a "scalar EM missile" -- a bundle of energy buried in hyperspace around its moving zero-point in this world -- will be fired directly at an offset aiming point prior to launch of the shuttle. A substantial offset in distance will be used to prevent posing any danger of inadvertently hitting the shuttle on its launch pad.

The marker beacon is necessary for precision. Its desired position will be calculated and the beacon will be placed on station. By placing this source of continuous EM energy in the area, it can be detected by the scanning "energy extraction beam" and tracked by the distant Soviet operator. Soviet trawlers and satellites monitoring the test will report the true position of the marker beacon, which will deviate from the calculated location due to conditions of the channel medium (the Woodpecker beams) at the time. Comparison of the beacon's true position to its desired, calculated position "registers" the marker beacon and yields the "meteorological corrections" which must be applied to standard firing table calculations. These corrections can be applied to any further shot into the area to correct its aim, just as standard field artillery applies registration corrections in the target area.

When the shuttle boosters ignite, an additional track will be obtained from it. The offset from the received track of the marker beacon and the received track of the shuttle booster flame thus will give the exact correction or "shift" from the marker beacon settings that should be used to set the howitzer's aim. In other words, a "relative" firing, offset from the marker beacon registration track, will automatically compensate for any vagaries of the medium experienced by the beams in transmission.

The Russians, being excellent artillerymen, have simply applied the standard artillery technique of shifting from a known registration point.

The Soviet Launch Phase ABM System is cocked and ready. This is the last test in the "wet run" series in which an offset will be used. If everything goes as planned, future shots will move into the next phase, where actual U.S. targets will be destroyed. The two great powers are already at war, but one of them does not know it yet. Because of the importance of this last test before the next step in hostilities commences, Gorbachev himself is monitoring the test.

The countdown begins. In Florida a great many persons are observing the event. George Suchary has already spotted a strange light hanging in the sky, where none ought to be. Bob Gladwin, resting his camera on a railing, prepares to snap his first picture so that he will be sure to have a sequence completely across the launch.

Deep within the Soviet Union, a large display screen is illuminated. The marker beacon's track is clearly shown on the display, as are synthetic symbols for the shuttle on its launch pad, several Soviet ships off the coast of Florida, and several Soviet satellites.

The control officer barks, "Missile test! Proceed!" The simulation officer presses a switch, and a simulated shuttle launch track appears.

The operator responds immediately, "Launch detected!" The identification officer, scanning the frequency components and the signature of the energy extracted from the launch, calls out, "Missile! Target hostile!" He presses a button, and a light illuminates on the panel, marking the target as a hostile missile. The launch track begins to blink.

The control officer announces, "Prepare to fire!" "Pulse mode!" "Strength one!" "Test!" "Apply burst offset!" The operator rapidly flips switches, setting the mode of fire and the offset. Instantly lights light to show that the computer has computed the shift and offset, and corrections. A red "Ready to Fire" light illuminates and a low siren begins its wobbulating growl.

The control officer orders, "Fire!" The operator presses the fire switch. A "Fire" light appears. A slight, deep shudder is felt briefly as the howitzer fires its scalar EM pulse missile of deadly energy. On the screen, the path of the scalar EM missile is traced out as it speeds toward its target far over the horizon.

In Florida, George Suchary is getting more and more curious about that pesky light. Bob Gladwin is preparing to snap his first photo as the final seconds of the countdown tick away.

On the distant Soviet screen, a sudden small burst appears, superposed over the synthetic shuttle track. The operator yells, "Burst!" On a smaller screen, a vertical profile shows the strike directly underneath the target. Immediately a blinking "Target Killed" light appears as the computer assesses the shot a kill.

The operator excitedly calls out, "Kill !" "Direct Hit!" Cheering breaks out from the firing team and the onlookers. Gorbachev and his entourage are smiling broadly and clapping. The cheering quickly hushes as the crew and participants continue to watch the track of the rising shuttle, and prepare for the launch and the next test.

Bob Gladwin has snapped his first picture (figures 13 and 14) leaving the shutter open a bit to allow for the nighttime conditions. As luck would have it, the scalar EM missile strike Occurs during the time the shutter is open. The film has captured the strike by a special process, even though human eyes have not seen it. (We will explain that shortly.).

At 7:29 the Atlantis's boosters ignite and the vehicle begins to rise smoothly off the launch pad. She will continue to rise, then pitch over to a 45 degree angle and proceed down range..

Bob Gladwin is snapping pictures regularly, one after the other.

In the local Florida area, hundreds and hundreds of persons are watching the flawless launch. The tip-off of the Indian River Community College basketball game and the St. Lucie County School Board meeting are held up while participants watch the Atlantis begin to rise serenely upward.

In the Kremlin, another track has appeared on the control screen display. Once again the sharp-eyed operator snaps, "Launch detected!" The identification officer, scanning the frequency components, tersely rasps "Missile! Target hostile!" Again he presses a button, and a light lights on the panel, marking the target as a hostile missile.

The control officer orders, "Simulate multiple firings!" The simulation officer acknowledges, "Multiple firings simulated!" and presses a button to cause the computer to simulate multiple launches on the display.

Immediately other synthetic target tracks begin to appear on the display screen, one by one, clustered around the shuttle launch at various separation distances, pre-arranged to represent multiple firings from a missile silo complex for this exercise. A special audio alarm sounds as each track appears.

The operator immediately calls out, "Multiple tracks! Missile launches!" and then "Identified hostile!"

The control officer orders, "Prepare to fire!" "Area pulse mode!" "Test!" "Apply time offset ten!"

The operator flips switches rapidly, acknowledging. Lights illuminate on the console, showing status of preparation for firing. The "Ready to Fire" light illuminates and the low siren begins again.

Back in Florida, Bob Gladwin has stopped snapping photographs. George Suchary has gone into his house to find his Polaroid camera and take a picture of that pesky light still hanging up there in the sky.

Back in Russia, the fire control officer orders, "Fire!" The operator presses the fire switch. The "Fire" light illuminates, immediately blinking, and the siren changes tone. On the display screen, all target tracks are blinking. A large light on the console comes on, showing that an "exercise time offset" of 10 minutes has been applied. A digital clock starts ticking off the minutes.

Minutes pass as the offset time ticks down.

Back in Fort Pierce, George Suchary is coming out to take photographs of that darn light, which is still hanging up there!

Folks are streaming back into the Indian River Community College gymnasium for the basketball game. Attendees at the St. Lucie County School Board meeting are wandering back inside, to get ready for the meeting.

More minutes pass.

Back at the control console in the Kremlin, the offset clock ticks down to zero. The siren ceases its wobbulating tone and holds a steady note. The "Fire" light ceases blinking and illuminates with steady brilliance. A deep shudder is felt as the howitzer fires again. On the screen, the rapid trace of the mighty scalar EM serpent is shown as it races toward its target.

Suddenly it reaches the target area, and a large blossoming light appears, covering all the target tracks. The operator announces, "Burst!"

All the target tracks flare brightly, then dwindle and die. The "Targets killed" light illuminates and the operator yells, "Kill! All targets killed!"

Wild jubilation breaks out. Everyone is grinning and clapping. Gorbachev personally congratulates the fire control officer and pumps his hand. The tension relaxes visibly. In the noise, the operator touches a joystick and slews away the marker beacon, preparing to shut down the system.

The final test has been a crowning success. Now the real first phase of active combat begins. "And the fools don't even know it yet!" a grinning Gorbachev delightedly announces.

Back in Fort Pierce, precisely 12 minutes after liftoff, high in the atmosphere above the launch site and a little downrange, a tremendous rumbling boom suddenly rocks the coast for hundreds of miles in each direction. The blast is heard as far north as Charleston and as far south as Key West.

George Suchary is taking pictures (figure 10) as the hanging light in the sky suddenly begins to move away, obeying the controls of a Russian operator thousands of miles away. Bob Gladwin has already finished snapping his photographs (figures 13, 14, and 15) of the launch and taken down his camera.

The anomalous light disappears rapidly in the distance, moving faster than any jet aircraft.

People gradually resume their normal activities, talking about the beautiful evening and the perfect launch. Many of them have noticed the strange light that hovered and bobbled in the sky, then sped away.

The final Soviet registration of their eerie weapons for World War III has been completed. A totally different kind of Pearl Harbor -- more akin to a Trojan Horse -- has been prepared for the unsuspecting United States.

Countdown toward Armageddon begins.

****

George Suchary's photograph of the marker beacon, just after it had started slewing away, is shown in figure 10, as printed in the News Tribune.

Bob Gladwin's photographs of the shuttle launch are shown in figures 13, 14, and 15.

Figures 13 and 14 show two developments of the same shot, one light and one dark. The light one shows the treeline in focus in the distance, with the light streak and burst.

Figure 15 shows liftoff of the shuttle and its rise and tilt over on the first part of its trajectory.

Figure 16 shows a possible detailed interpretation of the first Gladwin photo. The simulated aim point was offset from the shuttle, and the strike was directly at the offset point. The computer computed the burst strike strike of the scalar EM missile from the adjusted registration data, and determined whether or not the target was killed.

The control precision is shown by the small size of the scalar EM missile.

The strike of the weapon is caught by the camera, but not seen by the naked eye. None of the many hundreds of onlookers saw the strike, though all of them saw the shuttle launch.

There is a very good reason for that, though it is not commonly known.

Constable

The reason is that Bob Gladwin's camera was utilizing a form of photography pioneered -- to the best of my knowledge -- by Trevor James Constable. At least I learned about it from Constable's books and articles. I then told it to my good friend, Joe Gambill, and Joe applied and adapted it to the unusual photography which he performs. Both Constable and Gambill are masters of this type of photography. Once I got deeply into scalar electromagnetics, I finally figured out what was going on in this methodology.

Briefly, here's the way it works. Scalar energy, of course, has to be detected by some sort of interference phenomenon. There exists a very special way to get a camera to perform such interference for you.

In the EM spectrum, the infrared and the ultraviolet have a special relationship to each other. The ultraviolet is exactly twice the frequency of the infrared, if the two zones are properly chosen. In other words, the UV is the first harmonic of the IR.

The scalar EM energy, since it operates in the Kaluza hyperspace surrounding every point in ordinary space, may be considered to be composed of pure spin. (That is, it's moving in a dimension where each particle of that dimension is spinning). Harmonics there are spin harmonics.

In this photon-interaction produced level of reality (the ordinary world), the IR and UV bands are more closely connected to the "shadow world" in which scalar energy moves. Thus there is a sort of shadowy "swirling" of the scalar energy near the IR and the UV. Entry of the scalar energy into this world, through the IR and UV windows by a kind of "harmonic interferometry," is normally prevented by the presence of visible light. That is, visible light "squelches" the "paranormal channel" that lurks beneath the IR and UV zones.

To make use of the IR and UV interference source zones, Constable covered his camera lens with a special filter (18A) which is opaque to the visible light spectrum and transparent to the IR and UV. Note that the film is sensitive to the IR and UV region we are talking about, but the human eye is not.

When pointed at a sufficiently strong source of scalar energy, such a specially filtered camera has unique characteristics. It strips out the visible spectrum, but still admits the IR and UV spectrum. The scalar energy surges through these two source windows, and phaselocks harmonically. That is, the scalar energy entering through the UV window fits twice neatly in the scalar energy entering through the IR window. This interference, on the film, is (recall) swirling in frequency. The net result is that the film records the interfering scalar energy as visible spectrum energy.

Infrared film can be used to increase the response in many circumstances.

Both Constable and Gambill have rigorously proven this technique in literally thousands of "paranormal" photographs. (They do not necessarily use my scalar EM terminology, but the technique is rigorous and works, regardless of terms.)

Under nighttime conditions, sometimes the particular lens and film combination will phaselock between IR and UV for incident scalar EM energy. This is particularly true when a relatively strong source of IR is in the vicinity.

While Bob Gladwin was taking his photographs, there was no suppression of the IR/UV phase-lock effect by visible light because it was night and there was hardly any visible light entering his lens.

Thus his camera lens and film formed an IR/UV phase-locked interferometer that detected and recorded the strike of the scalar EM missile underneath the shuttle, even though human eyes did not detect it.

At this point it would be interesting to know whether NASA may have had IR cameras focused on this shuttle launch (which was at night). If so, it would also be interesting to know whether any of them may have developed the IR/UV phaselocked interferometer effect, and captured the strike of the weapon.

Note that daylight shots would automatically break the IR/UV phase-lock possibility, by jamming out the effect with visible light, unless a special filter is used to filter out the visible light but leave both the IR and UV.

One would hope that both NASA and the U.S. Air Force will take note and discreetly add a specially filtered camera to the instrumentation observing and recording all future daylight launches.

To recapitulate the major events at the shuttle launch on Nov. 26., 1985, see figure 17. A marker beacon was placed in the vicinity, high over the area, for precise registration and location of tracks relative to it. A scalar EM missile strike may have occurred just prior to launch, and offset from the shuttle.

Then the shuttle lifted off. The exercise called for simulating multiple launches from a missile field. The area code was computed, and the fire order given. A substantial time delay -- say, 10 minutes or more -- was used to absolutely assure that the actual shuttle was well away from the engage blast.

Twelve minutes after launch, the delayed firing in an area "multiple missile kill" mode occurred, producing a massive boom heard for hundreds of miles up and down the east coast. After the boom, the marker beacon was slewed away, and photographed by George Suchary just as it moved.

The shuttle moved on downrange and went merrily on its way, blithely unaware of the jaws of the tiger that had snapped shut behind it.

To the Soviets, this final test of the preparation phase was highly successful. They were now ready to increase the risk level and fire the first round of World War III.

- THE SHOW BEGINS -

13. On Dec. 12, 1985 the same Soviet weapon tested against the previous NASA shuttle launches and against various aircraft may have deliberately interfered with the controls of an Arrow DC-8 taking off from Gander Air Force Base, Newfoundland. At an altitude of 100 feet, the aircraft -- carrying over 250 U.S. soldiers and civilian crew members -- lost power and sank into the ground tail-low, killing everyone on board. It was a tragedy of enormous proportions to the U.S., and especially to the families and friends of the brave servicemen and crewmembers who lost their lives.

Three Canadian witnesses to the crash were interviewed over the Canadian Broadcast network television news on April 8, 1986 at 10:00 p.m. No flame or smoke issued from the plane before its descent and crash.

However, witnesses reported seeing the aircraft mysteriously glowing with a yellow halo. That is a signature of the use of a scalar howitzer in the "continuous EM emergence" mode, similar to the manner in which several F-111's were downed in Vietnam.

In short, the DC-8's electrical systems were interfered with by electromagnetic energy and noise created throughout each increment of spacetime occupied by the aircraft. A powerful charge was rapidly created in and on the aircraft structures and skin. The "yellow glow" seen by the witnesses was a corona due to the skin of the aircraft acquiring a high electrical charge.

Something also apparently caused the rapid loss of two engines, one after the other. This indicates that the distant Soviet operator may have struck at the engines, one after the other, with a scalar "energy extractor beam." Thus he apparently tracked the aircraft as it moved down the runway, then used two modes against it: one to create serious EM interference with the electrical systems of the aircraft, and one to disable the engines.

With its controls ineffective and power drastically reduced, the doomed aircraft sank to earth, still in its "tail down" configuration from takeoff, and crashed and burned.

Admittedly other factors may also have contributed to the crash. These included reduced lift due to poor engine maintenance, increased weight of the aircraft due to icing, and heavy loading. However, normally the aircraft could still have taken off and flown to its destination. (See "Arrow Air was absolutely safe, FAA official says," UPI release, Birmingham Post-Herald, Apr. 17, 1986.)

To date, no official seems to nave recognized the sinister significance of the "yellow glow" or what it implies. No one seems to have connected the loss of this aircraft, the mysterious booms over U.S. shuttle launches, the Woodpecker interference grid, the anomalous loss of the U. S.S. Thresher, the high altitude "booms" off the east coast of the U.S. some years back, the puzzling loss of two critical Titan launch vehicles in a row, anomalous engine flameouts of civilian jetliners, etc.

Certainly officialdom has not recognized that a Soviet Launch Phase ABM System is being tested directly over its collective head, to include actual destruction of U.S. aircraft, missiles, and space vehicles with concomitant severe loss of American lives.

Direct and diabolical acts of war have been committed upon us by the Soviet Union, but in a fashion that still eludes our erstwhile pundits.

However, it is not as simple as that. To even energize giant scalar EM weapons is to run a risk -- to all mankind -- so horrendous that it boggles the imagination. Truly it deserves Brezhnev's 1975 phrase, "...more frightful than the mind of man has ever imagined."

Shortly we will reveal exactly what he meant.

- Apr. 18, 1986: Death of a Titan -

14. As this book goes to press, the last two U.S. Air Force Titan 34-D missiles fired from Vandenberg Air Force Base in California have blown up shortly after launch. The first one blew up on Aug. 28, 1985 just after lift-off. That loss has been attributed to failure of a high-powered fuel pump, causing a massive oxidizer leak and a smaller fuel leak. The second Titan loss occurred on April 18, 1986 when the Missile blew up 5 seconds after lift-off. Its loss is still under investigation and no determination of cause has been made. Apparently the shuttle and the Titan presently provide the only viable launch vehicles for launching U.S. "spy" satellites. The loss of these sensitive satellites -- if indeed they constituted the payloads -- cannot help but be damaging to our strategic surveillance capability. According to the Los Angeles Times, the single remaining KH-11 satellite was launched in December 1984 and, with an expected life of two to three years, it could stop functioning later this year.

Whether or not significant Woodpecker grid activity existed in the vicinity of the Titan launch of August 1985 is unknown at this time. However, significant activity in the grid definitely occurred before the April 18 Titan disaster and on the same day.

On Easter Sunday, 30, 1986 engineer Ron Cole observed significant cloud signatures of grid pattern activity, correlated with Soviet Woodpecker measurements. On April 18, the present author observed traces of a cloud radial over Huntsville, Alabama and took photographs of it. Preliminary reports from Thousand Oaks, California indicate extensive grid activity again on April 18, the day the second Titan exploded.

At least the second of these two missile destructions shortly after launch is highly suspicious, since the grid positively was active during that time. Also, a clear trail of Launch Phase ABM system indicators exists back to the massive scalar exercise of April/May 1985. The first Titan explosion in Aug. 1985 thus falls within the Soviets' "now let's test them against U.S. launch vehicles" period.

The second Titan loss follows highly suspicious losses of the Arrow DC-8 on Dec. 12, 1985 and the Challenger on Jan. 28, 1986. The same Soviet weapon system that destroyed those targets probably also destroyed one or both of the critical Titans.

At least, on the second Titan explosion, the Soviets left another signature. In one of the videotapes taken of the explosion, the marker beacon can be seen well , above the exploding debris, moving away from the explosion site.

See for yourself! Refer to "Titan Explosion Cripples U. S. Launch, Surveillance Capability," Aviation Week & Space Technology, 124(17), Apr. 28, 1986, p. 16-19. Specifically, look at the color photo of the explosion on p. 18.

Look well above the explosion (about one inch on the page) and to the left (about one inch on the page) of the center of the top of the fireball portion. There you will see a little light, caught in the very act of moving away.

At my rather frantic urging, engineer Ron Cole sought out a copy of the original videotape. He and a team spent two hours, going through this portion of the tape frame by frame. The light is there. It's real. And it moves on out of there, well above the explosion and independent of it.

It's the Soviet marker beacon, caught in the very act. Here I want to express my sincere appreciation to Ron Cole for his untiring efforts to nail this light down. Without his dedication to supporting this investigation, it would have been impossible to say for sure that the light was there.

So we have caught the bear red-handed, with his fingers in the cookie jar. If the marker beacon is there, you may rest assured that the Soviets destroyed that Titan missile, with very high probability. [See also "Investigators Search for Clues to Cause of Titan Blast," Aviation Week & Space Technology, 124(17), Apr. 28, 1986, p. 19-20; also "Titan Accident Prompts Funding, Technology and Policy Questions," AW&ST, 124(17), Apr. 28, 1986, p. 21.].

Continue

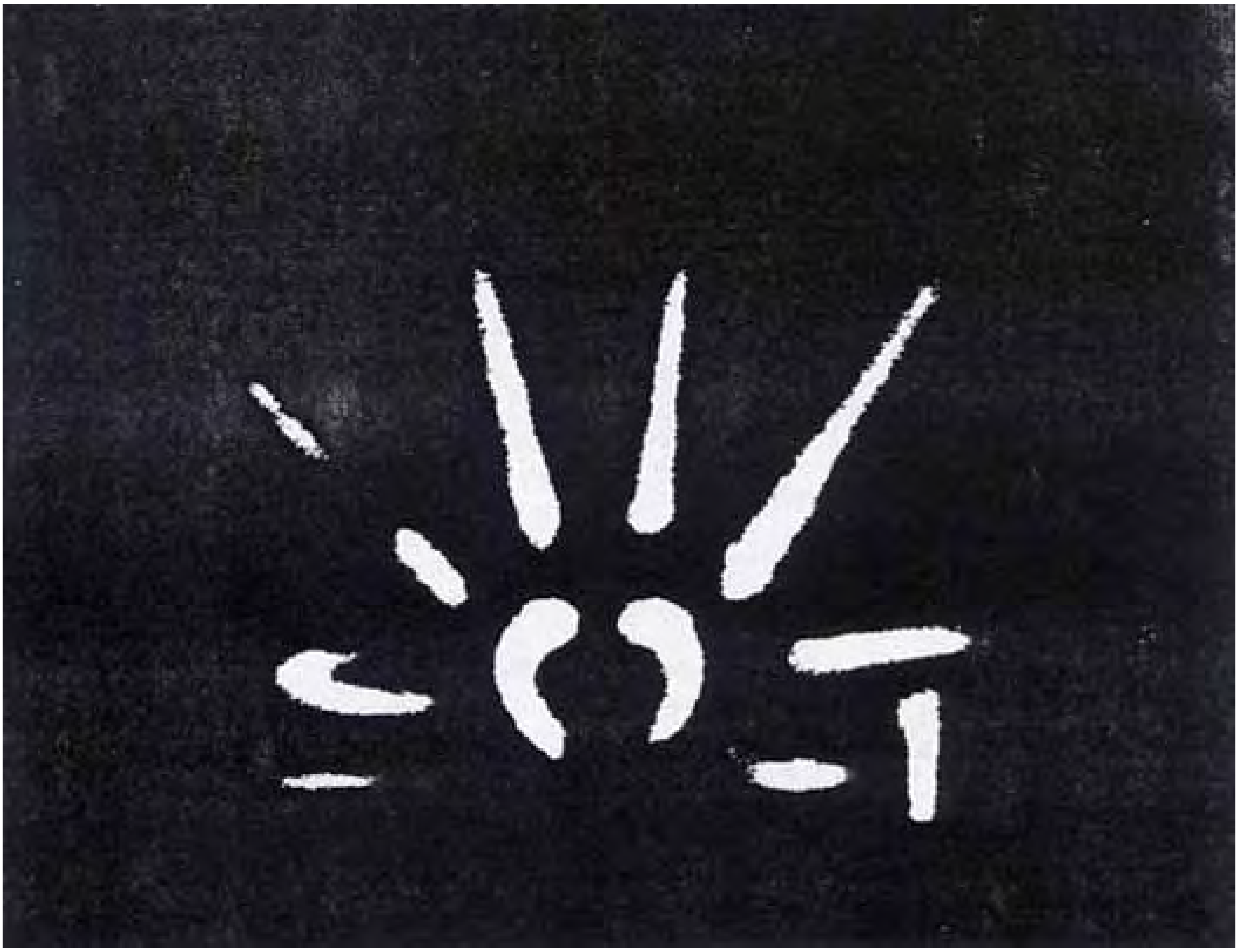

Figure 1. Wilson's first cloud radial

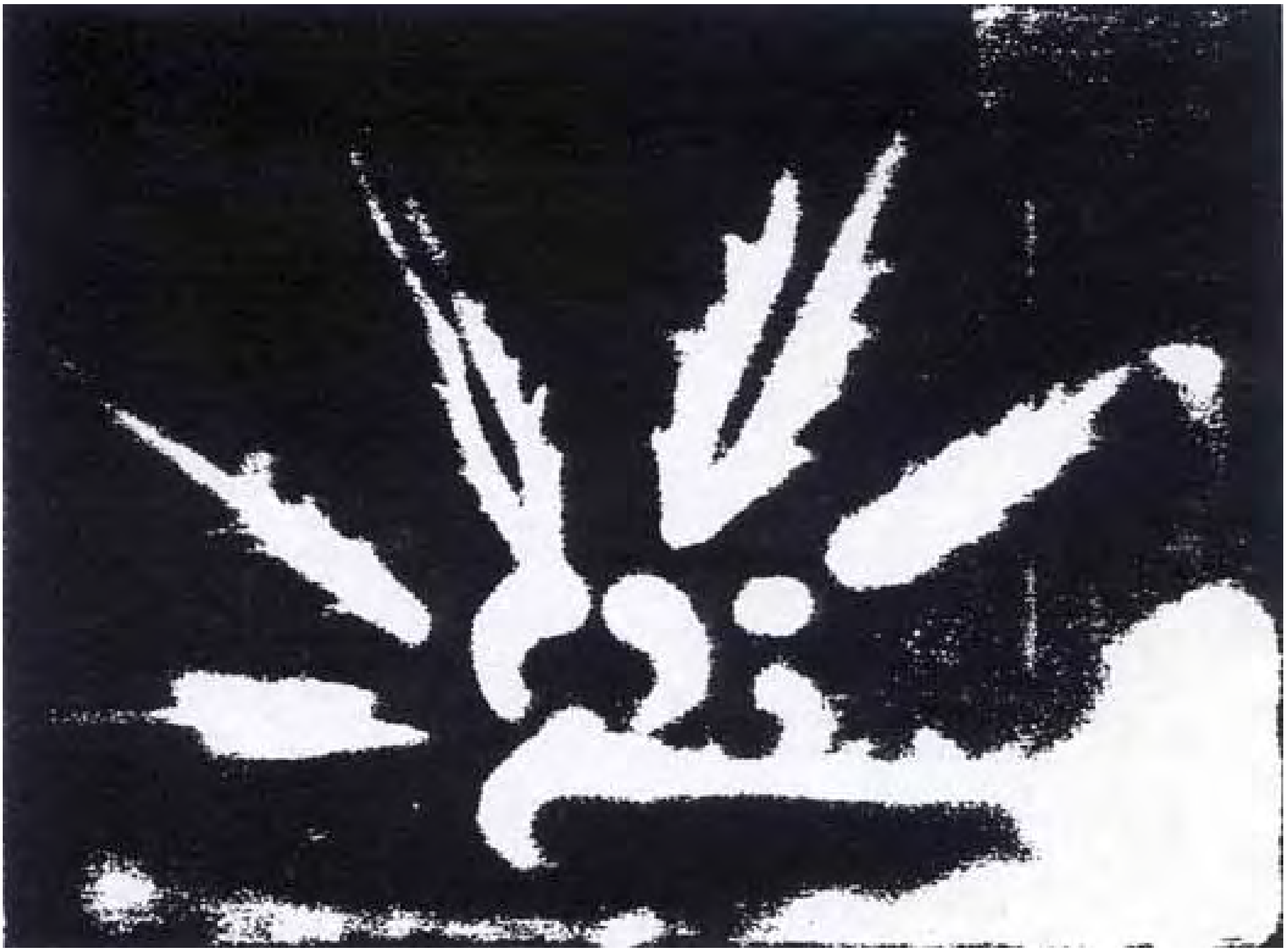

Figure 2. Wilson's second cloud pattern

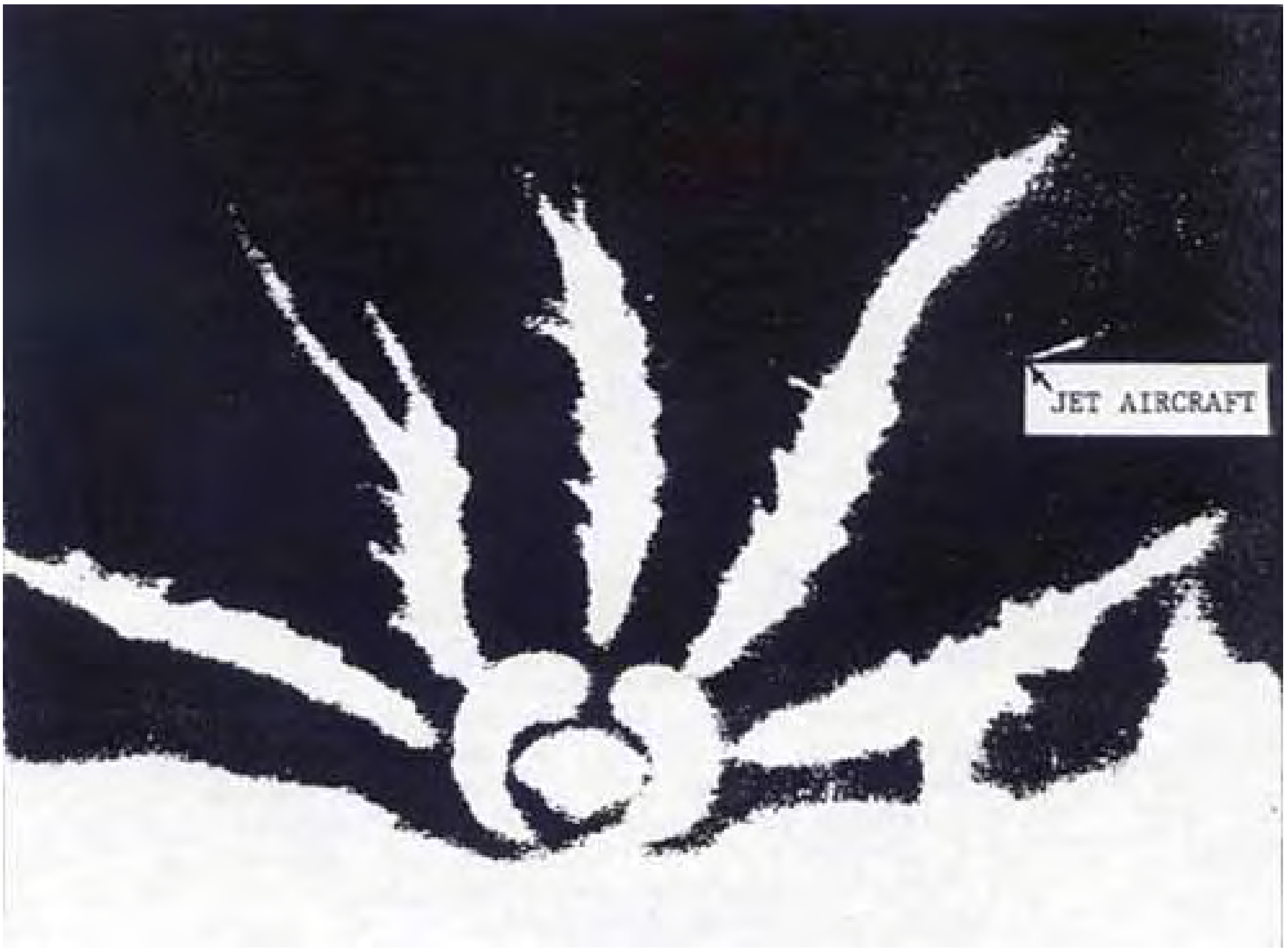
JET AIRCRAFT

Figure 3. Wilson's third cloud radial

Figure 4.

PLACEMENT OF WILSON'S THREE RADIALS

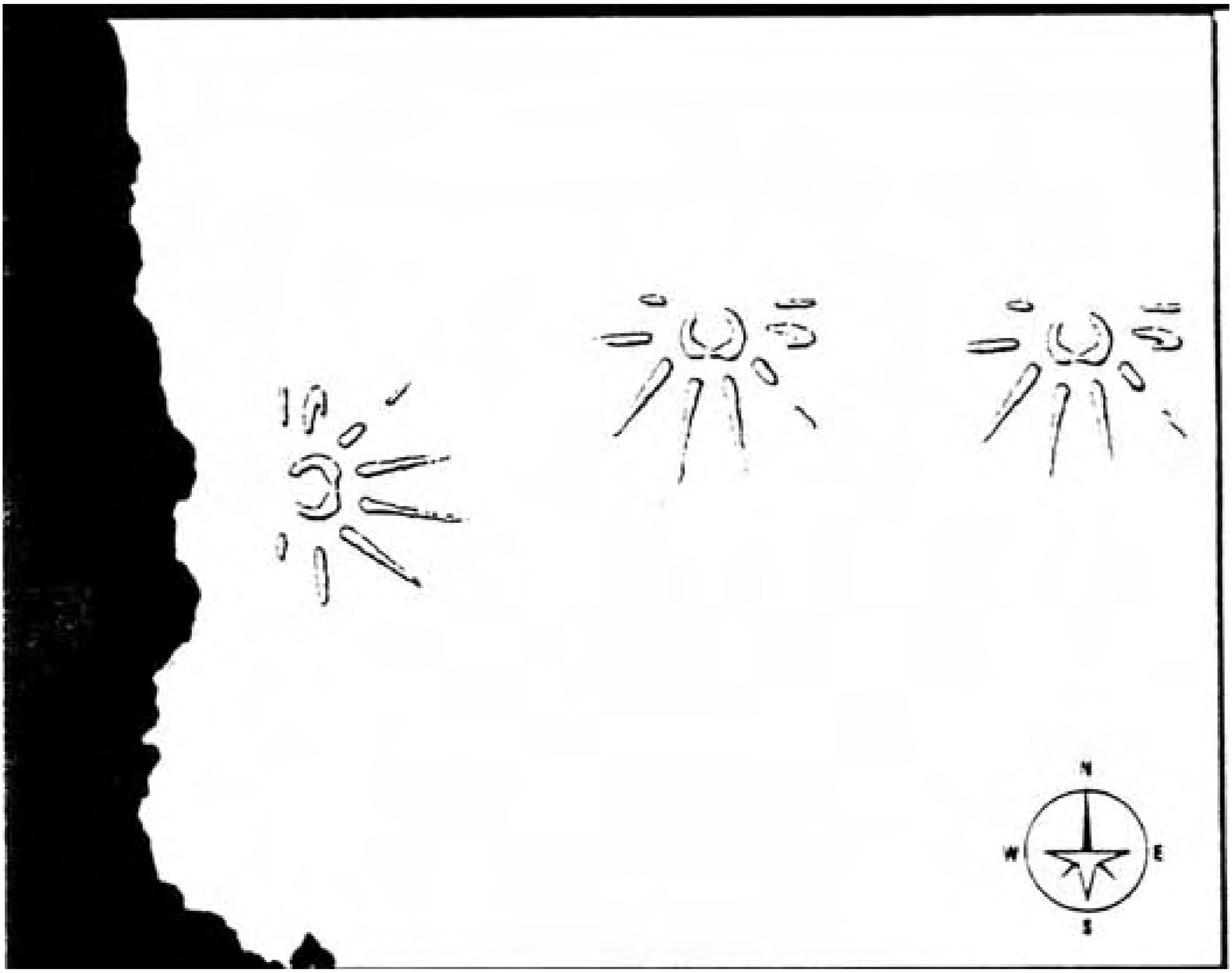

Figure 5.

VIEW FROM WESTLAKE OVER VENTURA COUNTY, CA

1-24-86, 280• WEST/NORTH-WEST

11:45 AM PST

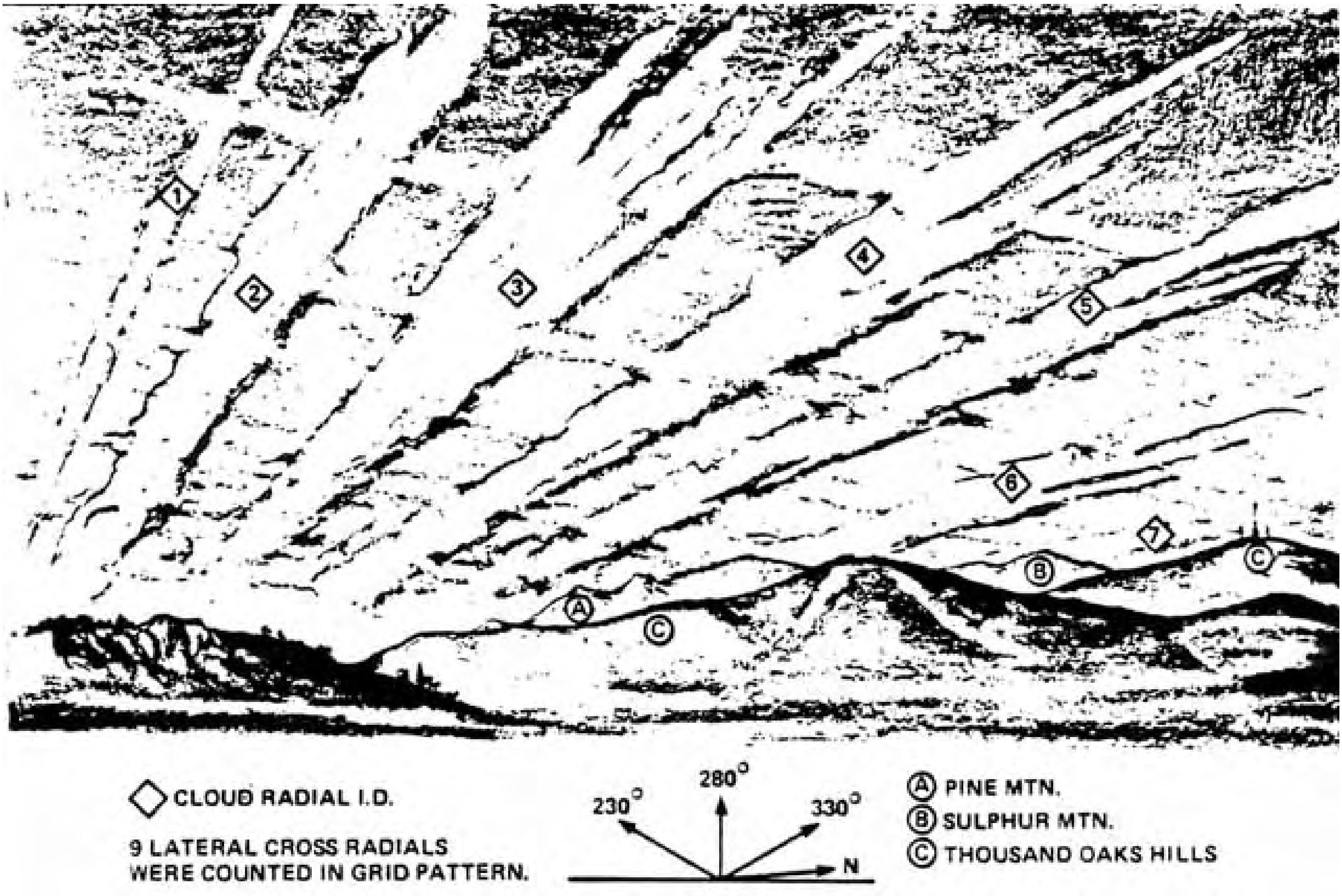
1
2
3
4
5
6
7
A
B
C
C
CLOUD RADIAL I.D.
9 LATERAL CROSS RADIALS
WERE COUNTED IN GRID PATTERN.
230°
280°
330°
N
Ⓐ PINE MTN.
Ⓑ SULPHUR MTN.
Ⓒ THOUSAND OAKS HILLS

Figure 6.

GRID VECTOR APPROXIMATION

1-24-86 VENTURA COUNTY, CA

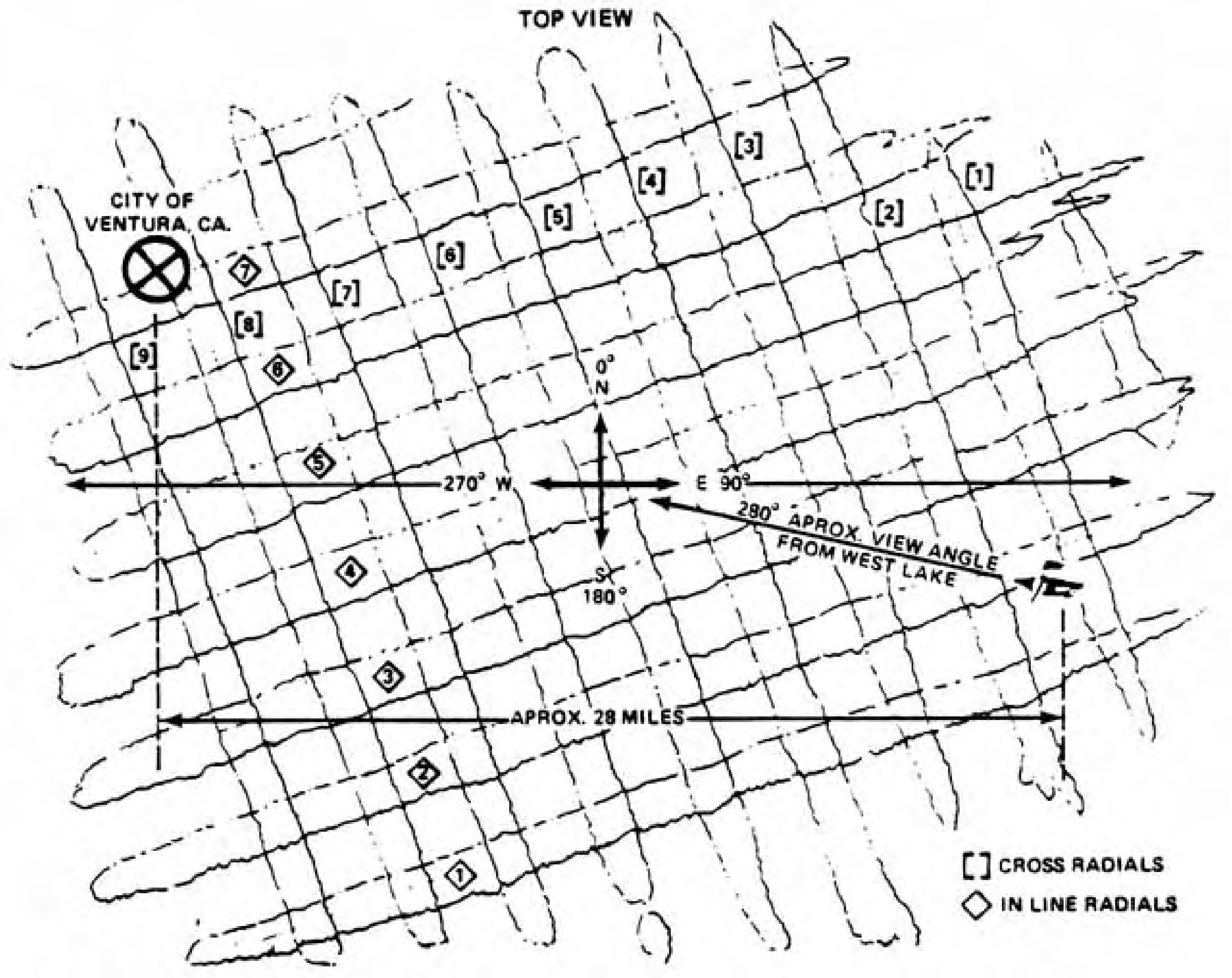
TOP VIEW
CITY OF VENTURA, CA.
[1]
[2]
[3]
[4]
[5]
[6]
[7]
[8]
[9]
7
6
5
4
3
2
1
0° N
270° W.
E 90°
S 180°
280° APROX. VIEW ANGLE FROM WEST LAKE
APROX. 28 MILES
[ ] CROSS RADIALS
◇ IN LINE RADIALS

Figure 7.

SIX-MILE DROP IN TWO MINUTES

CHINA AIRLINES FLIGHT 006
- TAIPEI TO LOS ANGELES -
19 FEB 1985

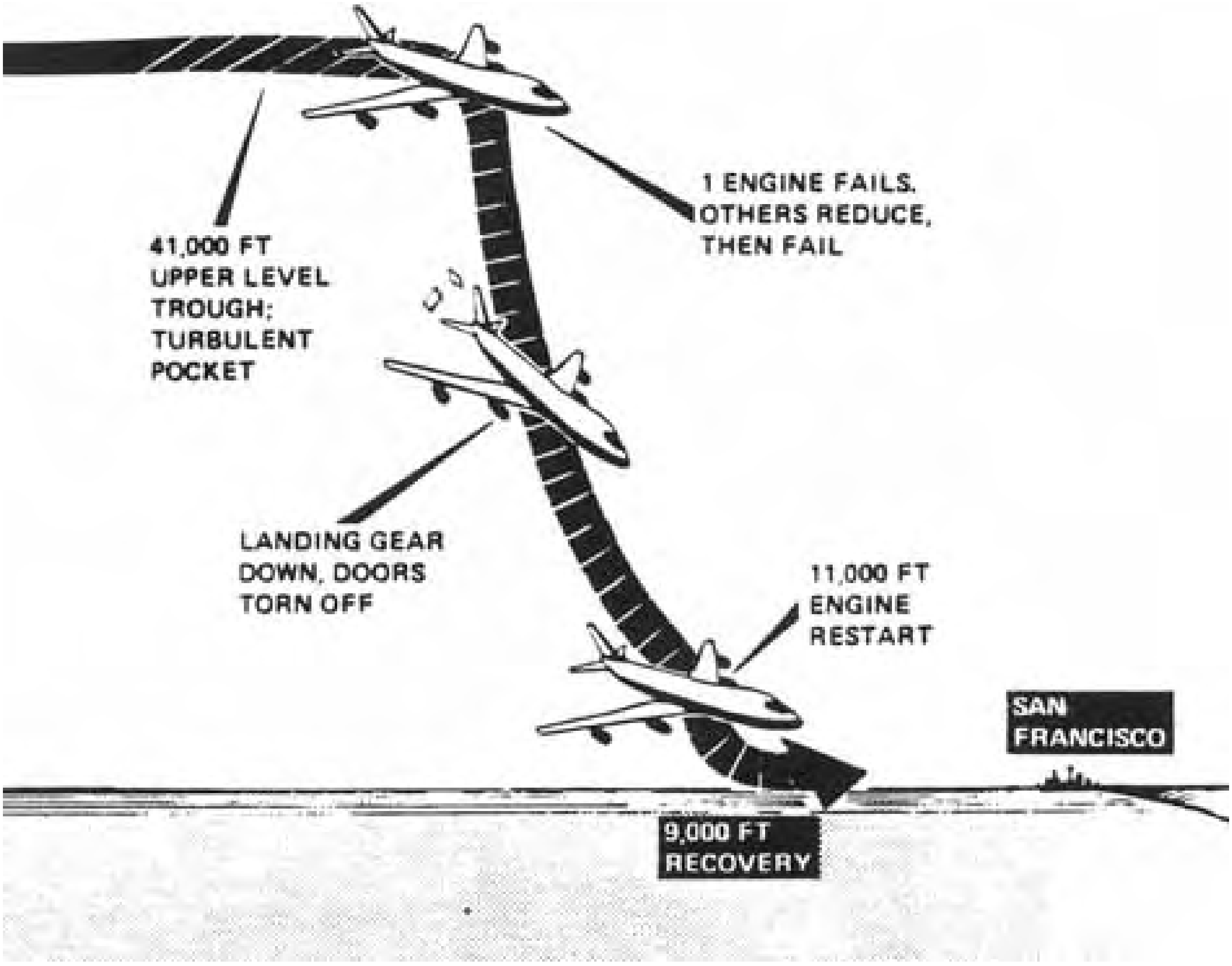
41,000 FT
UPPER LEVEL
TROUGH;
TURBULENT
POCKET
1 ENGINE FAILS,
OTHERS REDUCE,
THEN FAIL
LANDING GEAR
DOWN, DOORS
TORN OFF
11,000 FT
ENGINE
RESTART
SAN
FRANCISCO
9,000 FT
RECOVERY

Return to Fer de Lance Index

[The Tom Bearden Website](#)

[Fer de Lance](#)

Figure 8.

SAN DIEGO CHAFF INCIDENT, JAN. 1985

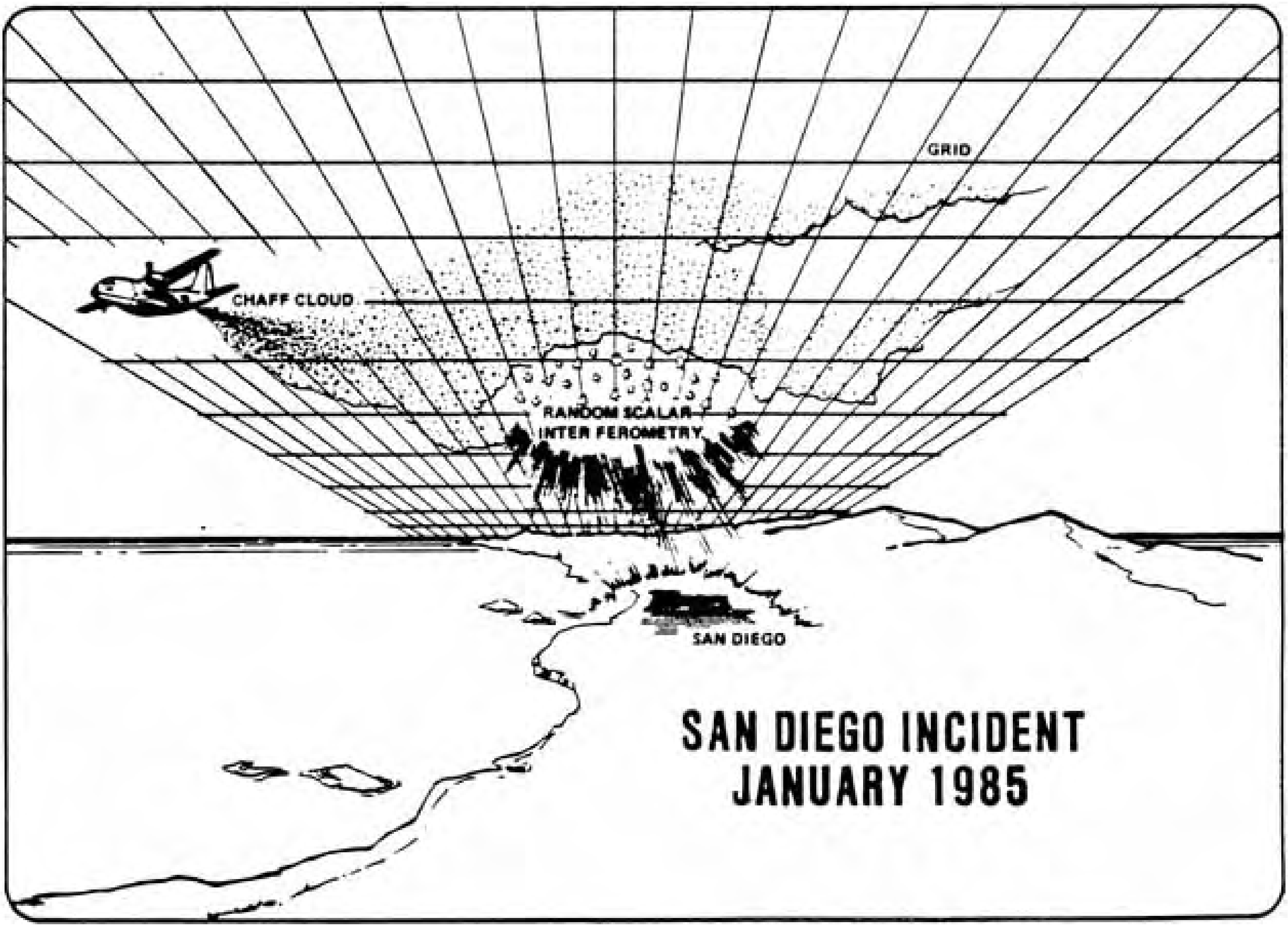
GRID
CHAFF CLOUD
RANDOM SCALAR INTER FEROMETRY
SAN DIEGO
SAN DIEGO INCIDENT
JANUARY 1985

Return to Fer de Lance Index

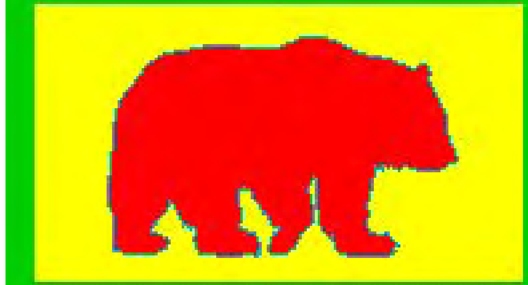

EM Missile Strike, Offset from Night Shuttle Launch

Cape Canaveral, Florida

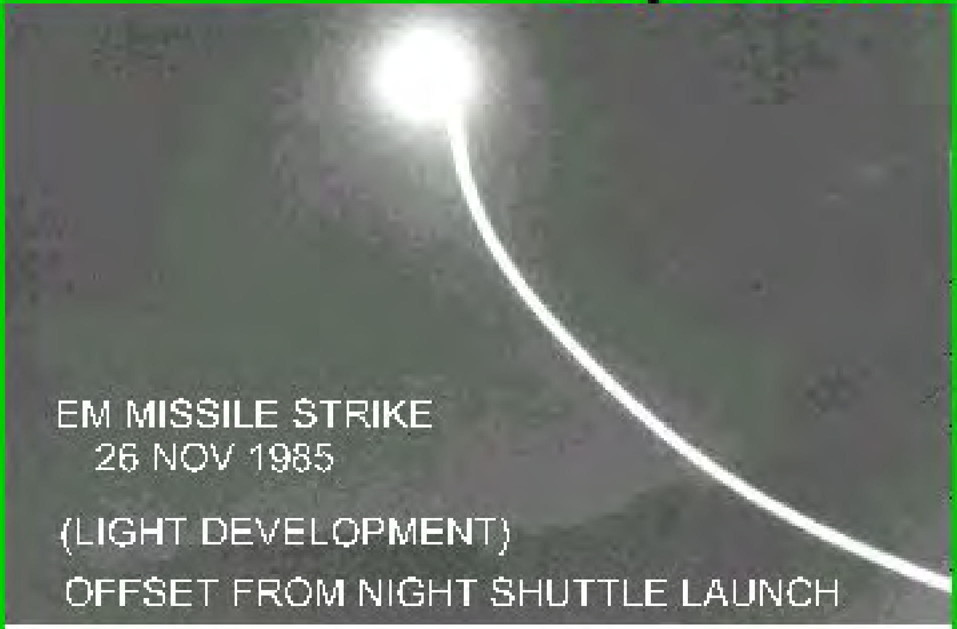

EM MISSILE STRIKE
26 NOV 1985

(LIGHT DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

PHOTO BY BOB GLADWIN

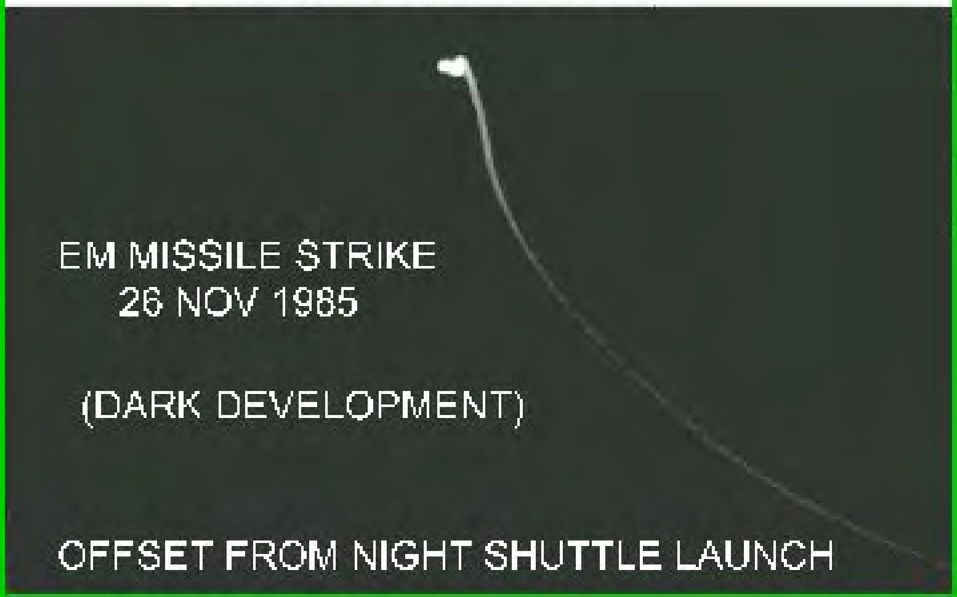

EM MISSILE STRIKE
26 NOV 1985

(DARK DEVELOPMENT)

OFFSET FROM NIGHT SHUTTLE LAUNCH

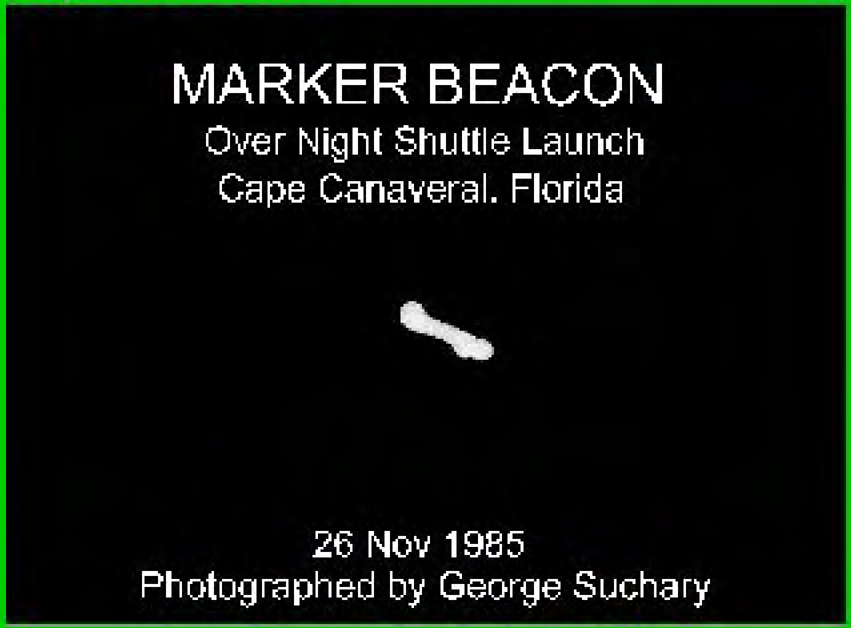

MARKER BEACON
Over Night Shuttle Launch
Cape Canaveral, Florida

26 Nov 1985
Photographed by George Suchary

© T.E. BEARDEN 1995

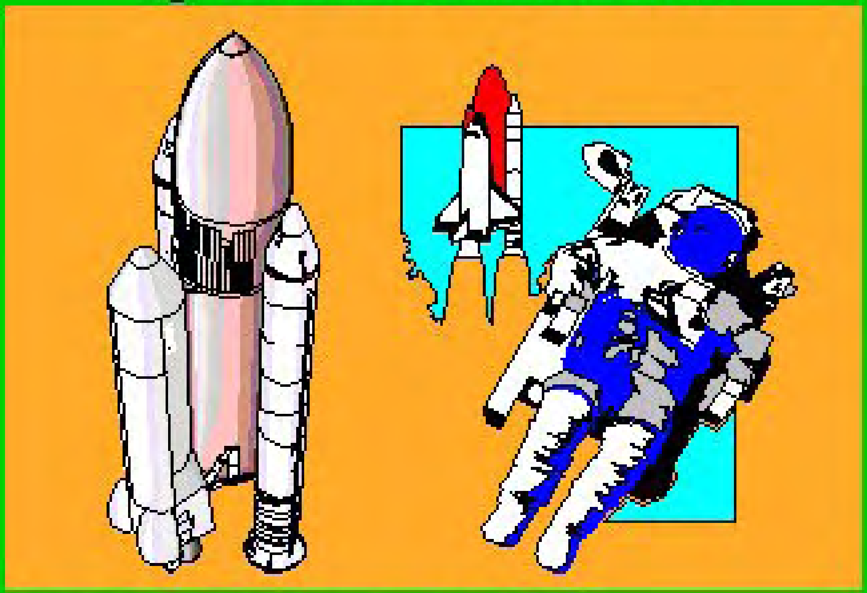

The Smoking Gun on 26 Nov. 1985

Longitudinal EM Wave Interferometry: Exothermic

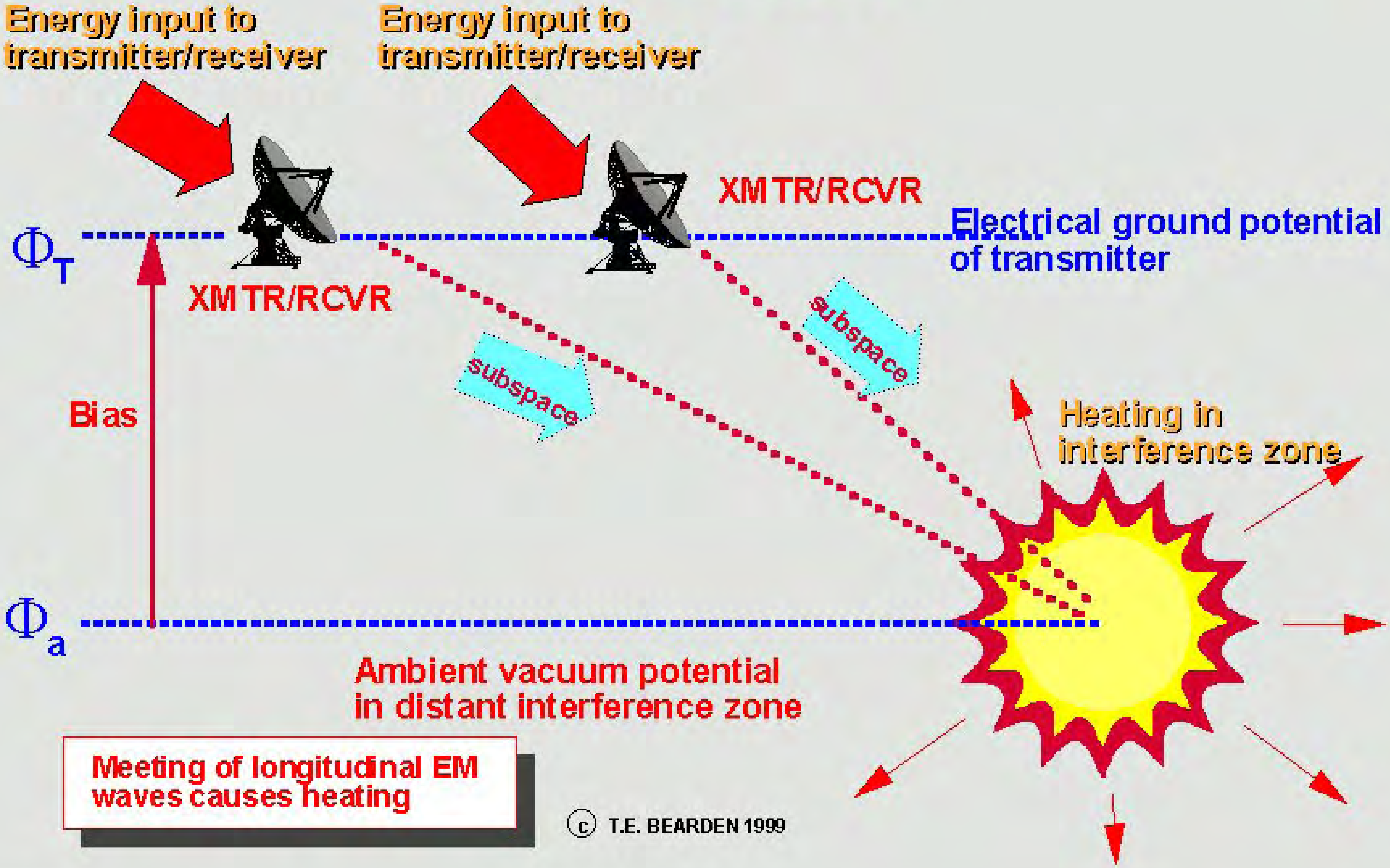

Energy input to transmitter/receiver

Energy input to transmitter/receiver

XMTR/RCVR

XMTR/RCVR

Electrical ground potential of transmitter

Φ_T

Bias

subspace

subspace

Heating in interference zone

Φ_a

Ambient vacuum potential in distant interference zone

Meeting of longitudinal EM waves causes heating

ⓒ T.E. BEARDEN 1999

Longitudinal EM Wave Interferometry: Endothermic

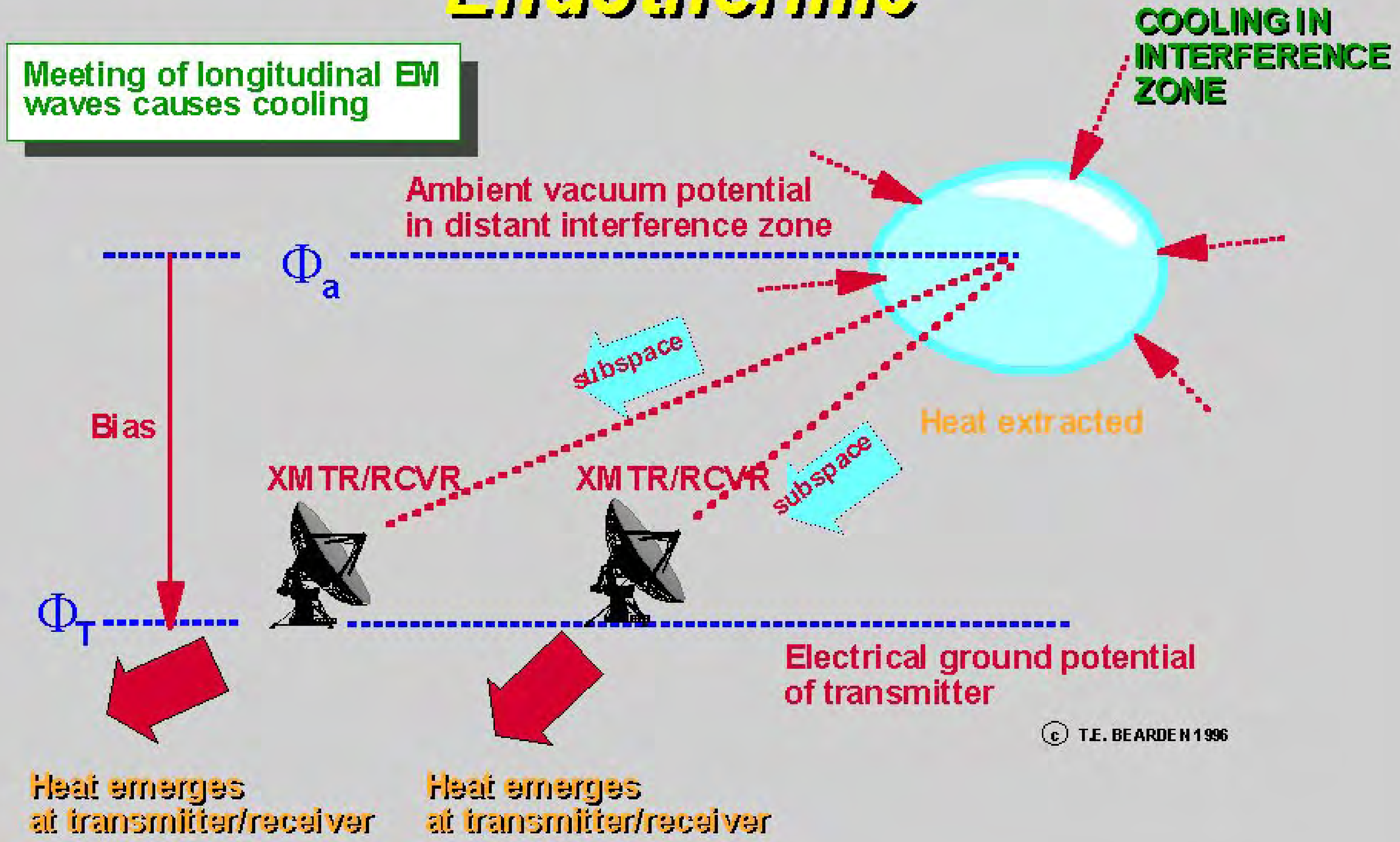

Meeting of longitudinal EM waves causes cooling

COOLING IN INTERFERENCE ZONE

Ambient vacuum potential in distant interference zone

Φ_a

subspace

Bias

Heat extracted

XMTR/RCVR

XMTR/RCVR

subspace

Φ_T

Electrical ground potential of transmitter

© T.E. BEARDEN 1996

Heat emerges at transmitter/receiver

Heat emerges at transmitter/receiver

- More Frightening Than the Mind of Man Has Ever Imagined -

Brezhnev's unheeded 1975 characterization of these weapons as "more frightening than the mind of ran has ever imagined" has very real justification. In a moment we shall explain what he meant. We shall also explain why the Russians may be serious about substantial nuclear disarmament.

There is now a "balance of terror" loose in the world that makes the old MADmen's MAD concept look like a sunny day in Hawaii.

Let me explain just a little of what I mean.

The scalar EM can and does use an ordinary EM beam as a medium. Many, many scalar EM channels can be infolded into an ordinary carrier. Thus the amount of energy stored in the carrier as "electrogravitational" energy has little to do with the feeble amount of electromagnetic energy stored in the EM carrier signal.

Thus in one sense, when the Soviets are employing the big "continent buster" weapons, literally hundreds of millions -- or even billions -- of watts of energy may be infolded into the seemingly weak carrier beams. All this energy is in the scalar EM channels established in the carriers.

Plus which, the scalar beams are employed in a "standing wave" node. Each scalar standing wave is like a gigantic "capacitor": It is highly charged. Under normal conditions, it has a charge and discharge time constant as well. If something were to "short circuit" one of these huge standing waves -- say, one with a thousand megawatts in it -- all that stored electrogravitational charge would discharge through the short circuit into whatever is on the other end of the short, and into the ground itself. In other words, the earth itself in that location would receive a mighty electrogravitational pulse (EGP). The ground in a region surrounding the discharge point would "charge up" just like a capacitor, and would transmit the EGP as a hemispherical current into the earth and spreading around it. This is where the real problem cores in.

The problem, as originally discovered by Wilhelm Reich [orgone energy was and is real, and is simply one form of scalar electromagnetics], is that radioactive material has a very great affinity for this EGP and undergoes a drastic reaction to it. Crudely, it works like this:

Imagine the vacuum as a virtual particle gas. This gas is under extremely high "pressure" or, more correctly, internal Stress. This stress is a gravitational potential.

However, the stress of the ambient vacuum is normally about the same in most localities. So a chunk of radioactive material doesn't experience a difference in vacuum potential across itself or throughout itself. In other words, the scalar EM fluxes in and out of every part of the nucleus are very high, but of equal magnitudes. The radioactive nucleus, however, is "unbalanced"; that is, it has an extra bit of unresolved EG potential which it would very much like to get rid of or discharge, by spitting out one or more particles. That is, it would like to "decay" to a more stable configuration. A nucleus that is trying to decay to a more stable configuration is a radioactive nucleus.

Remember, the radioactive nucleus is involved in a tremendously dynamic virtual particle flux exchange with the vacuum potential. And a part of it isn't quite in balance in that flux. Sooner or later the right "bobble" in the input flux will come along, and the nucleus will decay and emit a particle and/or radiation. By measuring the level of decay activity in a certain amount of the radioactive material, we can determine a "decay constant," or basically the probability that a nucleus will decay in one second. We can then determine how long it will take that chunk of material to decay away to a stable, relatively inert material. Some materials only require a few minutes to decay, while others may require thousands of years.

Now if you suddenly and drastically change the stress of the vacuum in which a radioactive nucleus is imbedded, you change the potential placed on the unbalanced nucleus. You hit it, in other words, with a strong EG potential pulse -- an EGP. That instantly unbalances the nucleus and it decays explosively. That is, it changes the decay constant and the probability of decay.

Basically that's what happens in a nuclear implosion weapon: You hit the radioactive hunk of material in the center with a very high, imploding pressure wave, which is carrying an EGP. The EGP causes all the nuclei to decay nearly instantly, producing a nuclear explosion.

The bottom line is: strong EGPs wreak havoc if they happen to hit anything radioactive. Such as nuclear warheads, in missiles or in storage. Such as nuclear reactors. Etc.

Now do you begin to see the problem? Let me spell it out.

Suppose the Soviets have activated some of their giant scalar EM transmitters, to place a giant weather engineering grid over the U.S., or to knock down a Challenger or a Titan. They've got their scalar EM beams all charged up. One of the giant Woodpecker transmitters has -- let's say -- 500 megawatts in its infolded scalar EM standing waves. Everything is going good. They're socking it to the Americans and the Europeans.

All those multiple scalar EM frequencies they're using to form the patterns, etc. have to all be absolutely phaselocked. Therefore all the transmitters have to be rigorously phaselocked, too.

And suddenly something goes wrong. A major amplifier/transmitter controlling or stabilizing the phaselocking malfunctions. The phaselock is disturbed, and begins to quiver.

Emergency circuits react instantly, trying to stabilize the situation and restore the quivering phaselock. Little "shorts" are being applied to the circuitry and to the "giant capacitor" that is the scalar EM standing waves. Dumps and discharges of EM energy are beginning to flash in the transmitter and associated circuits -- and into the ground. These make the situation more unstable. Feverishly the emergency circuits struggle to contain this thing -- and Wham! they cave in, all at once.

Now you get a mighty and continuing discharge of the scalar EM energy in that standing wave back into the transmitter and into the ground. It burns out the transmitter. And the discharge continues as the "capacitor" drains all its EG energy into the earth in a giant pulse.

Around the transmitter site, the EGP is transmitted radially away, still very much like a standing scalar wave. It reaches the first nuclear facility. Let's suppose it's a nuclear reactor. Specifically, let's suppose it's a nuclear reactor at Chernobyl.

If you'll check Soviet Military Power, DoD, 1986 you'll find that there's a big ABM radar site not far from Chernobyl. And Canadian sources verify that one of the large Woodpecker transmitters seems to be sitting in that general location. Now check "The Soviet Nuclear Disaster," The Wall Street Journal, Apr. 30, 1986, p. 24. Check the small drawing in the lower right corner of the bigger drawing. Here you can see that a Woodpecker transmitter located near Chernobyl would ideally transmit nearly due west to intersect over the U.S., entering across North America from its east. Other Soviet "over-the-horizon (OTH) radar transmitters" go over the north polar regions, to enter the U.S. from the North. The two then form a giant interference grid over the U.S. (You can see that in Soviet Military Power.).

Suppose you were in the field in the U.S., measuring the Woodpecker signals. Suppose the stronger beam was coming from the east. You would be able to see this with your instruments.

Now suppose the giant transmitter near Chernobyl got into the type of failure we're talking about. The emergency circuits panic and cut down the EM carriers, trying to drain away the standing scalar waves in a controlled fashion. Then your instruments would suddenly record the "new" direction as coming from the north, since the east carrier is gone.

That's exactly what engineer Bill Bise detected out in California, about the time it appears the reactor at Chernobyl began to get in trouble.

Let's continue with our scenario. Suppose the emergency circuits at the discharging Chernobyl transmitter contain things for awhile. The standing scalar waves are steadily draining into the earth. The EG potential starts to rise at the Chernobyl nuclear power plant.

In the nuclear reactors, the decay constant of the uranium in the fuel rods changes. Additional radioactivity begins to pour out of the uranium. The reactor alarms go off, and the crew takes measures to reduce the problem. The control rods go further in, to slow it down.

But the EG charge keeps increasing. The radioactivity of the uranium continues to increase. The control rods go further in to counteract this. The electrical controls begin to act up, showing anomalous indications. The radioactivity increases. In panic, the controls are fought until the control rods go all the way in. The reactor should in fact now be shut down completely, in normal circumstances.

Things settle and hold a bit. The scalar EG charge increases. Finally the "shutdown" reactor starts to increase in radioactivity again. The control rods can't handle it. Everything is heating up, including the graphite moderator. Back at the transmitter, the emergency circuits finally are overwhelmed and give in. They fail all at once. The dreaded "uncontrolled short circuit" now exists between the collapsing standing wave and the earth. A gigantic discharge of EGP into the earth occurs at the transmitter site.

In the desperate reactor, suddenly a great EGP appears. Instantly the uranium goes intensely radioactive. A gigantic pressure of heat, steam, graphite and radioactive materials bursts out of the reactor, breaching the building and erupting violently into the atmosphere. The core melts down. The graphite moderator ignites. You've got the realization of the nuclear engineer's worst fears: core meltdown and reactor out of control. It's afire and spewing deadly radioactivity out by the gobs. And the heat! It's like hell itself opened up for business in this spot.

At any rate, something very similar to that seems to have happened at Chernobyl.

This author believes something similar happened in 1957 in the Southern Urals. (See Zhores A. Medvedev, Nuclear Disaster in the Urals, Translated by George Saunders, W. W. Norton & Co., 1979, p. 150.). In the Kyshtym area, an explosion of nuclear wastes stored underground occurred in the winter of 1957-58. The area for miles around was contaminated, and is still contaminated to this day.

Recall, Khrushchev spoke of a fantastic weapon in development in January 1960. Early prototypes of the weapon would have been in development and testing in 1957. Probably at that time there were no emergency control circuits to handle standing wave collapse, or only primitive circuits at best.

If the failure of a transmitter occurred in full-power experiments, the unrestrained collapse of the standing scalar EM wave would most certainly have produced a large EGP in the ground, and instantly activated the nuclear wastes stored nearby. Eyewitness accounts reportedly stated that the wastes exploded, and that is totally consistent with our thesis. It is in fact a scalar EM accident signature.

Any nuclear facility at all is an incredible and monstrous risk whenever scalar EM weapons are active.

Of course, one's nuclear facilities are at risk if one's enemy has scalar EM weapons and fires at them, as we covered earlier in this book.

However, they are at terrible risk even from one's own scalar EM weapons. The consequences ot a relatively simple electronic failure can be catastrophic; not only for the local nation, but for the earth as a whole.

To show what we mean:

Suppose a similar scalar EM accident -- such as seems possibly to have happened at Chernobyl -- occurs at some other giant Soviet "radar" facility. Further, suppose this particular transmitter is not too many tens of kilometers away from a Soviet missile silo complex. Just to really throw in the grit, suppose these missiles are the "big monsters" with--- say -- a 50 megaton warhead in each missile. Suppose there are 10 of them in the complex.

Let's say this accident occurs "catastrophically" -- that is, the emergency control circuits are overwhelmed immediately and a discharge of all the scalar energy begins at once.

A giant EGP strikes the earth, spreads in all directions, and strikes the waiting monster missiles. In their implosion trigger warheads and in their fissionable and fusionable payload, everything goes at once. Actually, it goes better than when normally triggered. So each one goes at about 100 megatons.

You've just had a monstrous 1,000 megatons of nuclear explosion in the earth. You get a giant earthquake. If it's not too far from the ocean, you get a tidal wave also. You get incredible amounts of lethal dirt and debris, radioactive and deadly, hurled violently into the air.

This is just about the equivalent of a "normal all-out nuclear war!"

And what about the other side's detection and early warning systems? Who's going to alert whom, and for what? How many missiles get launched?

I think you see the point.

And I haven't even covered what a giant EGP can do to humans. If it's truly a catastrophic failure, you can get an EGP that can wipe out half a mighty nation in one burst!

And if the discharge happens to tickle the Sun and Moon's feedback loops the wrong way, you'll get convulsions of the earth, mighty burps of the sun raining fire and brimstone on the earth, and a violent increase in the interior heat of the earth's molten core, with a concomitant eruption of that core right up though the mantle.

In other words, it could be "back to the beginning" for the earth, and start all over -- if, that is, some semblance of an earth survived.

The Soviets know full well the implications of their giant scalar EM weapons. (Remember, I estimate they have over a hundred of these monsters, based on Frank Golden's detection of their full-up exercise in April/May 1985.).

Scalar EM weapons have added a totally new dimension to the meaning of the word "terror". If one develops them and his adversary does not, nothing on earth can stand against them. So in Man's present state, the temptation to develop them is nearly irresistible.

On the other hand, if one's adversary already has developed them, one is already totally defeated unless one also develops them. Again, the need to develop them is imperative.

Yet, whenever one activates a large scalar EM weapon, one immediately places the entire earth in deadly peril. The slightest misstep, and it's curtains for everyone. And it's curtains for the earth as well.

So one is damned if one does, and damned if one does not.

We have reached the final turning point in human history. The clock is ticking away the last years -- or days -- of the lives of all humans on earth. There has never been such a threat.

We simply have come to the crossroads. somehow, all Mankind has to change.

Infatuated with the thirst for ultimate power, we are facing two doors. Behind one is a beautiful lady, and the fulfillment of all man's dreams: control of physical reality itself. Behind the other is the deadliest tiger ever dreamed of in our wildest nightmares.

Which is it to be, the lady or the tiger?

SCALAR EM INOCULATION AND ITS IMPLICATIONS

My appreciation goes to Gary North for his openly sharing the above important information with the public. Human nature being what it is, one or more of his more dire scenarios is going to happen sooner or later.

The problem becomes enormously more compounded if the hostile agent or agents possesses sophisticated scalar EM technology.

With scalar technology, there exists the potential that the worm or virus can be implanted from a distance. If scalar scanning devices of sufficient precision can be developed, then major software residing in a chip, a computer, or on a disk can be read from a distance. With scalar EM methods, the ordinary software "guards" are totally ineffective. The scalar "reader" could simply read off the software directly as the disk or device goes about its normal operations.

After the software is read, a germ or a virus is designed to produce a particular problem as desired and upon special stimulus. Then a scalar "imprinter" device is used to produce the necessary digital pulses directly upon the media in which the software resides. The required code has now been .implanted. In effect, the operating system has been injected from a distance, and contaminated with the desired "sleeping agent" -- germ or virus. From there on, it's just like Gary North's scenarios.

So the scalar EM technology can probably be used (at least eventually) to place "mines" in systems in advance, for damage or destruction at a future date. At least it's possible in principle.

However, it's not the injection of computers and equipment that bothers me. It's the injection of human beings. Here's generally how it can be done (and possibly is being done).

Basically, human brains have various rhythmic waves which represent "envelope pulsations" of the entire electrical system. However, it is not the EM force fields represented by these pulsations which are important: it is the summed-zero (scalar) patterning of the entire complex of neuron firings which represents the real activity of the brain -- and represents it thoughts and emotions. A storage system for these patternings exists, but it's a scalar pattern storage, not an E and B field storage. The storage system is in the nuclei of the tissues of the nervous system itself, not in electrical or electrochemical potentials.

At any rate, the human brain rhythms are also "riding along" relatively entrained to the Schumann resonance of the earth's magnetic field. So all the patterning of the brain is grouped together into giant overall orders, and one or more of these is normally synchronized to our natural environment.

If a varying magnetic signal of extremely low frequency -- say 10 hertz -- bombards the brain with greater power than the earth's magnetic field, it is possible for the brain's rhythm to be "forcibly entrained" to synchronize with that signal, rather than with the normal Schumann resonance. If a large number of brains are forcibly entrained to such a signal, and the signal is artificial, then a master coherent phasing of all the brains to the transmitted signal has been accomplished.

And that can be utilized -- and is utilized -- by the Soviet Union. On the Woodpecker signals, Bob Beck has detected up to 16 carriers, each with a 10 Hertz modulation, and with all the 10 Hz modulations in phase.

Let's stress what we're saying here. At times, the Woodpecker signals have shown 16 carrier frequencies which contain the same 10 Hertz modulation, and which are all kept coherently sync-locked.

If those signals and the modulations are strong enough, the brains in a targeted area demodulate the 10 Hertz signal and phase-lock to it -- actually, to them. At that point, the brains are entrained to a 10-Hertz driver, and 16 channels of multimegahertz frequencies are locked-in to the brain.

Now if the Soviets have broken the genetic code of the human brain (which they have) and succeeded in developing the ability to implant thoughts, emotions, pictures, etc. (which at least to a degree they have), then they can simply record the proper signals from human guinea pigs, using scalar technology and instrumentation. They can tailor those signals and add them onto the Woodpecker carriers, using scalar EM techniques. By using most of the sixteen carriers as a means to express Fourier frequency expansions, specific areas and functions of the brain can be targeted.

Notice that the distant transmitters al ready "know" the correct phasing; just use the 10 Hertz modulation signal as a reference. In that case, phasing takes care of itself inside the captured human brains.

The brain has a unique characteristic. If a signal arises inside it, it interprets that it itself originated the signal. Thus for the first time in history , true unwitting "robot" actions and emotions can be engendered in masses of human beings.

But it's still worse. The Kaznacheyev experiments (see references) proved conclusively that electromagnetic signals can carry and transmit cellular disease and death patterns into a targeted biosystem. You can rest assured that the Soviets continued on to actually measure and determine a wide variety of death and disease scalar patterns. All of that can be added on to the Woodpecker carriers so that the population in the intersection grid area is devastated. One can even test it out in advance by producing a modified pattern of a disease in the targeted populace, and keeping the percentage of brains entrained down to a small level. This lowers the number of persons "injected" from a distance to a small percentage of the populace. Then the outbreak of the "new strain" of a familiar disease -- such as influenza, pneumonias, and something like AIDS -- can be monitored to provide targeting and effectiveness data. Notice that a series of diseases was experienced by U.S. personnel assigned to the U.S. Embassy in Moscow, probably as a result of the so-called "weak microwave radiation" of the Embassy by the Soviet Union.

But that's not all. Much more subtle effects can be produced in the targeted populace, without anyone getting wise. Cancers and leukemias should be relatively easy to induce. Also, the immune system can simply be attacked directly. If done slowly, a series of common, ordinary diseases will just seem to get "tougher and tougher to cure." Actually it's not the disease that got tougher, it is the immune systems that got weaker. It is my opinion that precisely that sort of testing -- against the immune system -- has been accomplished by the Soviet Union against the U.S. since about October, 1985. A great percentage of influenza patients during the period from then until now have had complications. So much so that multiple sessions -- as many as 3 to 5 -- of antibiotics are necessary before the patient recovers. At least in the Huntsville, Alabama "hinge pin" area, that seems to have been the case. This author personally experienced this, having influenza and follow-on complications for a period of seven months -- something quite different indeed for him.

But it's even worse.

The human scalar EM nervous system ("inside" the ordinary EM signals and functioning of the system) also has its own resident "software." It has special storage media contained in its brain. its cells, and its genetic material, and its actual nucleons. All of these storage media are continually telecommunicated with via scalar EM means. The "stability" of what each of us is, is largely conditioned by those recordings. As personalities we literally "are" what we have experienced, felt, thought, and did. Our whole "functioning conscious system" has been evolved by our experience, and it continues to be "us" by means of continual reference to the record base.

I hope you now realize that the actual record base itself can conceivably be altered by scalar electromagnetic means, even from a distance. Even by the hidden scalar EM content of 16

Woodpecker carriers, locked into a captured brain and pouring in their deadly content. It is conceivable that even what one is can be taken from him, and what someone decides he will be, he will be.

A totally new -- and ultimate -- form of "slavery" has emerged.

Also, look at all this scalar EM functioning of the human system as a special kind of cybernetics. A scalar EM computing software and hardware system.

What if scalar techniques are used to plant "worms" and "viruses" in this system? With hidden triggers? What about a series of them, so that a range of "eruptive software diseases" can be stimulated at will?

And guess what. Such scalar EM "worms" and "viruses" can even be induced by contact with or consumption of "contaminated" water and food. Or saliva. Or other body fluid.

So you can see the possibilities of "seeding" a population with a staggering variety of time-bombs.

I think you begin to see the point. Now you can change North's "Soviet blackmail" scenario drastically. In a new scenario, Gorbachev simply announces to Reagan that a graduated series of specific sweeping changes are going to happen in our population -- changes which we are absolutely powerless to stop. The last stages are various deaths. The master "trigger" for this series has already been initiated. Incidentally, the Soviet populace has been carefully "seeded" with the proper antidotes. The Russian leader demands certain things (very similar to what North's scenario utilized). Unless his demands are met, a "killer virus" will be activated by a special signal. If so, everyone dies. The first step change in the series will start to occur in a few hours. And it does.

In fact, this scenario could be added to North's scenario, and both used simultaneously. What do you think the President and the Congress would do? Or could do?

There are other even more diabolical possibilities, but these should be quite sufficient to illustrate the point.

Pandora's box has already been spilled, and the end of humanity is ticking away like a time-bomb. It's already nearly midnight, and the watchman hasn't even sounded the alarm yet.

I certainly can anticipate the next question: Have the Soviets already "seeded" the populations of the U.S. and other selected western nations with scalar EM viruses and germs"?

The answer is, I simply don't know. And just like Gary North, I don't know what if anything can be done about it if they have. This researcher certainly doesn't have all the answers. An immediate national crash program to develop scalar EM defenses is essential if we are not simply to be crushed beneath the hammer and sickle. It takes funding, people, facilities, and time.

In some 22 years of unrelenting struggle to convince the system we've got a terrible problem, I've hardly been able to move it off dead-center (no pun intended). This book, produced under extreme workload stress, is all I can do to raise the alarm. I've given it my very best shot.

What happens now remains to be seen. Other hands, hearts and minds are going to have to take up the task and carry it forward. This researcher has reached his limit, and has no more left to give.

1986 APPENDIX II

COMPUTER WORMS AND VIRUSES, AND THE IMPACT OF SCALAR EM

Source: - Gary North, "Electronic Aids (Parts 1 and 2)," Remnant Review, 12(20), Nov. 1, 1985, p. 1-8; 12(20), Nov. 15, 1985, p. 1-8. POB 8204, Fort Worth, TX 76124. Phone (817) 595-2691.

Annotation: - Two deliberately un-copyrighted issues to alert the public to software worms and viruses, and their implications for drastic interruption of society. Describes a hypothetical scenario in which drastic scrambling and destruction of banks' computer tapes, files, and records occur due to software worms and viruses. Written in first person by Gary North. [Slightly edited.]

Article: - "Worms." Earlier this year, I read a very interesting article on a major problem facing computer software (programs) development companies. A program comes on one or more 5.25-inch plastic discs. It takes only a few seconds to copy a program on one disc to a blank disc, which costs $3. Yet these programs normally run at least $250, and usually sell at $495, and sometimes cost thousands. Very few are less than $100. So you have a major temptation: make a $500 asset out of a S3 asset. Insert the $500 program into drive A, write "COPY A:*.*.B:" and hit the "enter key;" sixty seconds later, you have a $500 program in drive B.

There are ways to make this copying more difficult. The companies code the programs, and force you to have a control disc in drive A at all times. These "copy protected" programs are a hassle for users. We cannot put them on "hard (big) disc" easily, and sometimes the control disc dies for some reason. Then what? Your data are locked in your hard disc or on a floppy disc, but you can't get to the data because the control disc is not functioning. You order a replacement. Weeks go by.

Last year, several firms came up with a solution. It is called a worm. A worm is a command that is built deep into the complex code that creates the program itself. These are incredibly complex codes, and it is easy to bury a command in them. They cannot be traced.

What does the worm do? It "eats" things. Say that you are a software thief. You make a copy of a non-copy-protected disc, either to use on a second computer, or to give (or sell) to a friend. The program works just fine. But when the program is copied to a new disc, the worm is "awakened." It bides its time, maybe for many months, maybe for years. The program's user is blissfully unaware that a monster lurks inside his pirated program. He continues to enter data, make correlations, etc. He becomes completely dependent on the program.

Then, without warning, the worm strikes. Whole sections of the data disappear. Maybe the data storage disc is erased. Maybe it is just scrambled. Even his backup data discs have worms in them. Everything he entered on those discs is gone. Forever.

Can you imagine the consternation of the user? He has become dependent on a booby-trapped program. His business could simply disappear. For the savings of $500 (stolen program), he could lose everything he has.

Several firms threatened to insert worms into their programs. But then they backed off. They are afraid that lawsuits initiated against them might go against them in court. They could be hit for damages suffered by the thieving victims. Juries might decide that the punishment, a bankruptcy) was too much for the crime (a $500 theft).

So far, no worms are lurking in any commercial software programs — as far as I know and the industry knows, anyway. But what if a disgruntled programmer was to hide one in a master copy of, say, Lotus 1-2-3, the most popular business program on the market? What if ten thousand copies a month go out for, say, three years? Then, without warning, every company that has started using them loses three years of data? They sue Lotus. Lotus goes bankrupt paying lawyers. No company in the industry is willing to talk about this sabotage threat publicly. Obviously.

LARCENISTS

I just happened to stumble across an article on worms in a computer magazine. It occurred to me that it might be possible to use the worm technique as a form of deliberate sabotage rather than just as a copy protection device. But what did I know? I'm not a computer expert.

I know a computer expert, however. I mean, a real expert — one of those people you occasionally read about. In the world of business, they're called "space cadets." They operate somewhere in between the asteroid belt and Jupiter. But this one is different. He's a businessman, too.

I got him to sit down with me to discuss the problem of worms. It turned out that he has a real fascination for the topic. He tells me that there are advanced design worms, called 'viruses' by 'hackers' — computer freak programming geniuses. "The software virus is the most terrifying thing I've ever come across," he told me. And then he showed me why. My initial scenario is based on only a portion of his estimation of the threat. It gets a lot worse.

He gave me a 90-minute FIRESTORM CHAT interview. He must remain anonymous. He used to be a software developer for programs that were used in the U.S. banking system, but is now employed in a highly sensitive job in a related industry. Therein lies his problem. If he were to tell the story of what he is capable of doing to these banks, his firm might lose a lot of sales. He can't "go public." Let's call him Tom.

Let me summarize briefly some of the details he gave to me. They floored me. They're going to floor you.

(1) Jackpotting. The rush is on in the banking world to get automated teller machines (ATM's) into shopping malls, supermarkets, and in front of every bank. We've all seen them. Just walk up, punch in your card number, ask for cash, and you get it.

In a busy location, one of these machines can hold as much as $250,000 in cash, mostly small bills. These machines are controlled by computer. They are hooked up to the bank's computer system, usually by phone lines. This local line, Tom tells me, is what computer freaks call THE LOOP. The loop is wide open to tampering. He says that what computer thieves are doing is to hook up a cheap Apple II computer, tie into the phone lines, break into the ATM, and get it to empty itself. This is "jackpotting."

He tells me that ATM thieves are hitting banks continually, but nothing is getting to the press. The banks have yet to show a profit with the ATM's so far, which is understandable. They are hoping to get their machines placed in key locations, so "market share" is crucial to their plans. They are suffering horrendous losses in the short run in the hope that long-run profits will pay off, if and when a defense is developed.

The banks are saying nothing because of their fear that if the extent of the losses gets into the press, they will be forced by pressure from depositors — bank runs — to cancel the ATM's. The losses are horrendous, he says. At present, there is no known defense, given the communications technology.

(2) Rounding Off. This is the "preferred" computer bank theft system. Someone on the inside who has access to the software takes advantage of the banks' need to round off numbers. The programs carry numbers out to 13 places. Banks can't use all that space. So when they balance the books (interest rates at, say, 9.873), they just don't count every tenth of a cent. The program is assumed to round off the numbers randomly. What does the bank care? But the thief has set up bank accounts that absorb those random tenths or hundredths of a cent. In millions of dollars worth of transactions (federal funds, etc.), programmers in some cases have stashed away hundreds of thousands of dollars — maybe millions — over a few years. No one knows how much of this goes on.

How could a bank spot this? The books would always balance to the penny. How would the accountants ever know?

I think of a story that Adam Osborne tells in his paperback book, Running Wild. The president of a large firm was looking out his window one day, and he noticed two Rolls Royce cars parked next to each other. He inquired as to the owners. They were two men in the data processing department. He called in investigators, and the cars and the men disappeared. They fled to Brazil and took their cars with them; **Brazil** has no extradition treaty with the U.S. Years later, as

Osborne was writing the story, the firm still hadn't figured out what they had done.

ARSONISTS

These are the fearful ones, far more than the larcenists. These are the practical jokers who get into a major data bank and trash things. It's a kind of multimillion-dollar "Kilroy was here" graffitti.

How easy is it to get in? Incredibly easy. The boy in "War Games" really could have broken into most firms' telephone-connected computers. Computer programs exist that allow the user to hook up his computer to a phone line and randomly dial numbers until they hear the telltale whine of a computer line. It then notes the phone number and goes on its way, searching out more lines.

They can do it by long distance, free of charge. The telephone company has a tough time tracing those who use various sorts of electronic black boxes to call anywhere on earth at no charge. Some people get caught, of course. "The tip of the iceberg," says Tom.

How do they get in? Easy: few systems are protected, once you locate the line. If one is, he says, you create a deliberate error. Most programs then collapse the protective shell, and the hacker finds himself inside the heart of the system. Tom has designed a program that keeps this from happening to his company's programs, but few companies have anything like it.

It's very easy to get in if someone has "logged on," — opened his terminal's connection to the main computer — if the system is connected to phone lines. Or anyone in the company can just tap in, if someone has left his desk and left the computer on. It's common to forget and leave an open terminal.

He showed me. He says anyone can get fired for leaving a computer on. He demonstrated his point. With 40 computers on line, he ran a quick search and found two of them "logged on," despite the fact that it was after hours. All the security in the world can't do anything if a computer line is open. It's like a burglar alarm; it's worthless if you leave the door unlocked or leave the keys lying around. That janitor you hired. Is he a computer illiterate? Or a plant?

Once inside, what can you do? Steal a fortune? Yes, if you really know the system. He told me he could easily steal $3 million from a local bank, even as an outsider. He would then offer to give it back and keep his mouth shut about how easy it was if the bank would pay him 10% the take. He thinks most banks would capitulate for fear of the publicity. In any case, he knows that he probably wouldn't get caught.

How about creating a new identity? The grade-changing scene in "War Games" is true. You could even create a new identity, give yourself high grades in any academic discipline, just by breaking into a university's database. There is very little security here, he says.

But for sheer vindictiveness, for sheer envy, consider the possibilities of a virus-implanter. He gets inside the computer for a major communications link: telephones, large information database, bank wire transfer, or whatever. Then he lays the egg: a tiny, untraceable brief instruction. Inside a huge database are just a few characters. These float inside a system, seeking to devour certain kinds of data, or executing certain routines.

There is a game played by computer freaks called "Core War." They try to implant these killer messages, which seek out each other and battle one another. If you find one morning that yours has been consumed, you lost the battle. That was probably the origin of worms and viruses.

TERRORISM

Say that a revolutionary terrorist group, or some anti-zionist group gets a "ringer" into a system. He might be a computer genius type. Everyone knows they are either oriental, dark-skinned people with accents, or teenagers. The firms don't hire teenagers, but they hire a lot of foreigners. They may even check the guy's credentials. Electronic credentials. (Ha!) Then they turn the guy loose in the system.

The virus is implanted deep inside the system. It can then be transferred to any other bank's computer by means of EFT (electronic funds transfer). Maybe it is triggered when someone with a peculiar name and address opens a bank account. Three days later: Bam! The data disappear. They haul out the back-up tapes. Bam! The virus is on them, too. It is a process of infection, contamination, and incubation. There is no known defense. Not yet. This is the bottom line.

ANTIBODIES

The designer of a virus can also design an "antibody." The antibody is a counter-virus agent that seeks out a virus and destroys it. But like other antibodies, it must be specific. The only way today that an antibody system can be created is to know what kind of a virus is involved beforehand.

Tom says that people are now selling antibodies at very high prices. Who is paying? Big companies that suspect that there is a virus present in their computers. In all probability, the guy selling the antibody created and injected the virus. But now can any businessman prove it? So he pays the blackmail.

NATIONAL DEFENSE

A Soviet agent or American spy working for the Soviets penetrates any of a dozen computers used by the military. He plants a virus. The computers talk to each other, and the virus spreads to all of them. It tells them to execute a certain routine when a certain command is entered at a missile-controlling terminal. That command might interfere with a routine that activates a missile or launches it. Upon reading that command, the virus shuts down the computer, or scrambles the executing program, or scrambles the data. No more "launch on warning." No more launch at all. Dead metal.

Scenario: The President of the United States receives a telephone call on the "red phone" — the direct link to Moscow. He lifts the receiver and says, "Hello."

"Mr. President, this is Michael Gorbachev. You must recognize my voice. I have very little time. I will come directly to the point. You have refused to back down on your threat to implement your Strategic Defense Initiative. You intend to go ahead with space-based weapons. My military staff informs me that they think that the United States has the technology to implement it, and that it could place my nation's military strategy in jeopardy. We cannot allow you to do this.

"If we allow you to deploy the SDI, it will be too late for us to respond effectively. Therefore, we are taking the initiative today. I issued orders this morning to put Soviet military units on immediate alert. We are abiding by your biblical rule to announce the initiation of hostilities before striking. Neither the Japanese nor the Germans gave us this courtesy. If you do not come to terms with us, we will launch a first strike against your nation in three hours. We will delay for one day, if you agree to follow a precise procedure that I will outline shortly.

"At one time we feared nuclear retaliation. We no longer do. Within two hours, you will know why not. I suggest that you instruct your ballistic missile team to prepare your missiles for a strike. Then, to prove to yourself that we no longer are concerned about retaliation, launch one or two of them. As far as I am concerned, launch all of them. But please instruct your senior military commanders to report back to you concerning the effects of their instruction. I suggest that you try launching three or four as a test. We don't care which ones.

"Mr. President, let me tell you what is going to happen. As soon as anyone attempts to launch a missile, that missile's computer guidance system will shut down. It will lock up tight, and you will not be able to unlock it within the time you need to respond to our attack. Two hours and thirty minutes from now, your nation will have no remaining Minuteman III missiles to respond with, should you finally unlock your frozen computers.

"I suggest that you contact your senior officers now. You will have to mobilize them within 60 minutes. The test should take about 30 minutes. I will telephone you again in 90 minutes to present our terms of surrender." Click.

The President calls the Joint Chiefs. If he is lucky, he will be able to locate two of the three in time. They will be paralyzed. Who wouldn't be? But in all likelihood, they will at least test Gorbachev's theory. They will order one or (wo missiles launched. The computer guidance system on both will shut down the system. They will try two or three more, with the same result. They will attempt to launch one from a submarine, with the same result.

The President brings in senior Congressional officials and the remaining Joint Chiefs member to the White House.

Exactly 90 minutes after he had hung up, Gorbachev telephones back. He presents his list of demands. First, the immediate removal of U.S. troops from Europe. Second, the withdrawal of personnel from Diego Garcia Island in the Indian Ocean. Third, the breaking of diplomatic relations with Red China and Taiwan. Fourth, the removal of all U.S. troops from Korea. Fifth, a moratorium on all debts owed to U.S. banks by the Soviet Union and its client states. Sixth, the removal of all Minuteman III missiles from their silos. Seventh, the return of all U.S. submarines to port. If he agrees, and the orders are delivered within two hours, the Soviet Union will delay launching a first strike. The President complies.

They might do it with our communications satellites, Tom says. You might do it with any aspect of U.S. data transmission. The virus could sit dormant in a system for years, and no one would know. Triggered, it would then strike.

THE WEST'S VULNERABILITY

The West has become increasingly dependent on computers. We can no longer function without them. The Third world hasn't. Neither has the U.S.S.R. Their technology is still pre-computer. They are inefficient, but they are far less vulnerable.

Tom says that the world of computers presumes that almost everyone is essentially honest, and that all the brightest programmers must be honest. They aren't. Thus, the members of the entire system — banks, national defense, large and small businesses, public utilities — have opened themselves to attack. The attackers are invisible.

"Nothing I have seen in all my years of computers scares me as much as this does," he says. "The system has been designed in terms of a far older set of standards, especially with respect to security. It is totally vulnerable."

He compares it to plague, or venereal disease. People copy each other's software to save a few bucks. They use public access databases. They use "loops" — the phone lines. Yet these transmission belts of information can become transmission belts of collapse.

This is what I have harped on for twenty years: the potential for a collapse of the division of labor. We become rich by means of a brilliant technology, yet we become dependent on it to an extent that no previous society ever has.

Centralized institutions are most vulnerable, but because we use public transmission lines, from microwave transmissions to cables in the ground, each local unit is vulnerable. Those who would choose to bring down the system need only plant electronic viruses in a handful of major common-use databases or transmission sources, and live years or ten years later, the disease hits.

It could bring down the system if technological defenses are not developed. Nothing on the immediate horizon points to a solution, he says. The silence of those who should know what to do indicates that they don't know what to do, but they don't want panic to spread.

Is this fantasy? Last summer a 23-person team broke in (by phone) to Chase Manhattan Bank's computer.

Break-in: - Continuation of part 1, a treatise on worms and viruses in cybernetic systems.

"Maybe you saw the article buried somewhere in your newspaper. I saw it in the New York Times (Oct. 19):

> "A group of at least 23 teen-age computer users broke into a Chase Manhattan Bank computer installation by telephone in July and August and "significantly damaged" bank records, the Federal Bureau of Investigation said yesterday.
>
> Chase officials insisted that no money was stolen or transferred out of customer accounts as a result of the break-ins, and said no intrusions have been detected recently. But Federal officials, in an affidavit, said that in several cases the youths changed passwords, preventing customers — and in one case a unit of the bank itself — from gaining access to their own computer files."

And where were these teen-agers located? In San Diego, across the continent! It gets even more ludicrous:

Federal officials said that most of the offenders were probably too young to be prosecuted.

Robert D. Rose, the assistant United States Attorney handling the case, said: "We're not yet sure what we are going to do. But these things can get out of hand — it did get out of hand — and we have to treat them seriously."

Treat what seriously! "These things?" What things? If they can't legally treat the electronic trespassers seriously, just what is the man talking about? He is talking about the topic, above all topics, that bank and government officials don't want to face: the vulnerability of their computer records.

I have seen no follow-up on this story in the conventional press. A brief article did appear in the computer-oriented tabloid, Info World (Oct. 28). It turns out that the students had broken into the fifes of Interactive Data Corp. of Waltham, Massachusetts, which maintains the bank's financial records. The break-ins were discovered in late July. They had obtained the toll-free 800 number which was restricted (Ha!) to Interactive Data subscribers. As late as October 9, an illegal entry was observed. In short, it took ten weeks after the break-ins were discovered to put a stop to them.

The response of the bank's bureaucracy was predictable. It will ever be thus: "Bank officials are claiming that the FBI exaggerated the nature of the activities of the suspected individuals. A spokesperson for Chase Manhattan said that Interactive's customers were not prevented from accessing their accounts and that none of Interactive's data was altered or manipulated in any way." In response, FBI supervisory agent John Kelso said that the FBI has sworn affidavits from bank officials that say data has been manipulated or damaged. "That sounds pretty serious to me," he volunteered.

Here is the capper: Interactive Data has 25,000 subscribers who are tied into that toll-free phone line. Try keeping tight security on a system with 25,000 users. Chase Manhattan couldn't. If they can't, who can?

And if Chase Manhattan Bank was vulnerable to 23 teenagers who are too young to prosecute, consider its vulnerability to just one envy driven genius who knows all about electronic viruses. The students that did this were apparently just goofing around. But what if just one malevolent computer freak decided to "get even" with Chase Manhattan? What if he had phoned in just once or twice, implanted a long-dormant data-killing virus, and quit? What if he had tied its detonation to, say, a calendar clock in the Interactive computer? If it took security forces from July until early October 15 to raid the 23 students' homes, they would never have spotted one break-in. They could not have traced it, either. Conclusion: we have a risk-free opportunity for electronic arson. We face a potential electronic epidemic. And when I say "we" I mean the entire financial system of the West.

Sure, all the bank "spokespersons" in the world will tell you, "No problem." But there is a problem. A horrendous problem.

At this point, it really gets interesting. Chase Manhattan Bank has just announced that we will be able to set up our own personal electronic banking facilities with them by buying an expanded version of *Managing Your Money*, Andrew Tobias' home financial management.

Citicorp and Bank of America have opted for *Dollars and Sense* a rival program. You will be able to pay monthly bills electronically, balance your "checkbook," monitor your net worth, buy and sell stocks, etc., etc., just by dialing Citicorp or Chase Manhattan. Fantastic! But despite all the assurances, I get nervous. Yes, I know no one will be able to break in and tamper with the numbers. But 23 teenagers shouldn't have been able to do it, either. And now we're talking about a lot more subscribers than 25,000.

Obviously, the master program used by the banks will prohibit easy entry. Unfortunately, someone has to write the program. Can you imagine the blackmail possibilities? Some hotshot programmer could build-in a bomb, and then threaten to detonate it. In fact, he could merely pretend to have inserted a virus. Who would want to call his bluff? Not Chase Manhattan, I would bet.

CORE WARS REVISITED

In May of 1984, A. K. Dewdney published an article in Scientific American's "Computer Recreations" column. It was a light-hearted-piece on how computer experts can get involved in playing this exciting game of "blow up your opponent's defenses." You know: recreation! In the March 1985 issue, he wrote a follow-up. It begins:

"When the column about Core War appeared last may, it had not occurred to me how serious a topic I was raising. My descriptions of machine-language programs, moving about in memory and trying to destroy each other, struck a resonant chord. According to many readers, whose stories I shall tell, there are abundant examples of worms, viruses and other software creatures living in every conceivable computing environment. *Some of the possibilities are so horrifying that I hesitate to set them down at all.*"(emphasis added.)

It turns out that tne French nave been enjoying a novel on the international implications, *Softwar: La Guerre Douce,* by Breton and Beneich. A translation is scheduled for publication here by Holt, Rinehart & Winston. The study revolves around the sale of a high-power computer to the Soviet Union. The U.S. allows its export because it has a "software bomb" in it. When the U.S. Weather Service announces a certain temperature at St. Thomas in the Yirgin Islands, the program proceeds to subvert every piece of software in the Soviet Union.

A pair of Italian programmers was "inspired" by the translation of Dewdney's original article to dream up a virus (a virus is a computer-to-computer killer, whereas a worm is resident in one man's computer). They figured out that by infecting a disk operating system disk (these start computers and tell them what to do with programs and electronics), and then installing it on disks used by the biggest computer shop in the city, they could create an epidemic. They decided not to do it. In short, the only restraint is self-restraint.

A high school student in Pittsburg wrote a virus that was subtler than a data-destroying virus, which at least tells us that we have a problem. His virus created a plague of very subtle errors in the disk operating system. "All of this seems pretty juvenile," he wrote, but "Oh woe to me! I have never been able to get rid of my electronic plague. It infested all of my disks, and all of my friends' disks." He wrote a program to destroy the virus (an "antidote" to it) but it is not anywhere near as effective as the virus is.

Warning: do not copy disks from your friends' copies. This act of piracy could cost you plenty.

A COMMERCIAL WORM

Just a few days after I wrote "Electronic AIDS, Part 1," 1 read a column in the Washington Times, the conservative (Moonie-owned) daily newspaper. One of the reporters has a computer. He had purchased a newly released program from

Microsoft Co., called "Access." Understand that Microsoft supplies the disk operating system that is used by the IBM PC, the most popular microcomputer. In other words, this is no backyard company. It is one of the two or three software giants in the U.S. (Its owner is under age 30, which tells you something about who is pioneering the microcomputer revolution).

As he was setting up his computer to take advantage of this telecommunications program, a warning flashed on his screen: "The weed of crime bears bitter fruit. Now trashing your program disk." Wham! He lost all his files — probably a couple of year's worth of work. Sure, he was probably smart enough to have made back-up copies, but think of the risk. And what if it had been a worm that kept silent for a few years, infecting all of his back-up disks?

He called Microsoft, and they gave him the runaround. They told him that they were not responsible. Some programmer had put in the worm in order to zap program pirates, but the journalist insisted that he was an original buyer. Tough luck, they told him. Obviously, they didn't know that he was a reporter.

Then he published his article. All of a sudden, the victim was not some average buyer. He was big trouble. Things started moving. InfoWorld (Oct. 28) reports that Microsoft has admitted that a programmer put in the worm, but without permission. The offending text has now been removed, we are assured. But what if it had sat in the master for three years? Here is the premier firm in the software.

This is not idle speculation. It has already happened, verifying my hypothetical scenario within a few days after I published it.

Can you imagine the absolute havoc that a dormant worm or virus could create if it were imbedded in all updates of Microsoft's masters of PC DOS and MS DOS, the operating systems for all IBM microcomputers and IBM-compatible microcomputers? It could cost the U.S. economy billions, and some microcomputer-dependent firms wouldn't survive. Any Microsoft spokesman who says, "It's impossible; it could never happen," has to explain how it already did happen to "Access."

ADAM OSBORNE'S WARNING

You may know the name Adam Osborne. He invented the revolutionary portable computer, the Osborne 1. Before there was an Osborne 2, the company went bankrupt. Compaq, the most successful first-year firm in U.S. history (over $100 million in sales in its 12 months of operations) and others built imitations that were far superior.

That isn't my point, however. Adam Osborne was "present at the creation" of the microcomputer industry. He created Osborne publications, and then sold out to McGraw Hill. He knows what is going on. In his delightful paperback book, Running Wild, which is a history of the microcomputer (desk top) revolution, 1975-82, he offers this warning. He says that three areas should not be allowed to be computerized: 1) bank money transfers; 2) the stock market; 3) elections.

All three are just about fully computerized. Another ten years, or maybe five, and they will be 100% computerized. Several firms allow microcomputer buying and selling of stocks (e.g., Charles Schwab), and New York Stock Exchange floor transactions eventually will be fully computerized, at which time it will be pressured to get rid of the "specialists" who make (and sometimes manipulate) the market, short-term — Richard Ney's hated "Wall Street Gang" — but the price of getting rid of them may turn out to be horrendously high.

"The great fortunes of the 21st century," Osborne predicts, "will be the legacies of the great computer thieves of the 20th."

Three years ago, I used a firm to supply computer services I needed. The head of it was a former businessman, quite young, and a true "space cadet." I've quoted him in the last issue. I call him Tom. He operated in a world far removed mentally from the rest of us. He is a nice fellow, a Christian, and a moral philosopher of sorts.

He ran the operations of the local elections. He did it fairly inexpensively. He told me why: "I want to keep these elections honest. It would be incredibly simple to rig the program to produce whatever outcome I wanted in close races. If I can do it, anyone with enough skill to set up the system could do it."

I asked him if he thought Osborne was correct in his predictions about bank theft. "It would be a piece of cake for me to steal three or four million from any local bank. I could go in the next week, offer to give the money back, keep 10% as a finder's fee, and promise not to tell the press how easy it was to steal. They would probably pay me my 10% just to keep me quiet."

Look, These people are geniuses. Worse, they are geniuses in a very narrow field technically, which is now being used to control darned near everything. This unique intellectual-technical skill is the possession of literally a handful of people, mostly under 35 years of age. They are "fooling around" with Chase Manhattan Bank's computers. What happens when a few of them stop fooling around and get deadly serious?

Computer program designers keep telling us that there is no 100% secure way to defend data banks. Maybe there will be a 98% secure system someday, but not now. To defend itself, the system relies on the integrity of youth. In short, self-government is the major defense.

And where have they learned self-discipline? In the public schools?

"NOW YOU'VE DONE IT!"

About four years ago, I read an article in the *Rolling Stone,* the tabloid aimed at rock music fans. It was the only article I ever read in that periodical. It was a gem.

It described a subculture of students at Stanford University, "hackers." These people are computer freaks. The mainframe computer at Stanford was cheaper to use after midnight, so from midnight to 6 a.m., the hackers gathered at their terminals. They lived on candy bars, junk food, and high-technology dreams.

One of the games they played was breaking into each other's programs. It was considered the mark of a master hacker to be able to crack another hacker's defenses. They would spend hours trying. They were "hacker-crackers."

One bright fellow then designed a classic booby trap. He wrote a program that warned trespassers not to tamper with it. This, of course, alerted every would-be electronic safe cracker to the challenge. It was a complex program, and it took days to crack it. Then, after repeated warnings, the successful trespasser got a surprise. Japanese letters appeared on his screen. Roughly translated, the words proclaimed, "Now you've done it!"

At that point, the victim's computer screen went blank. Then the names of all his own computer files appeared on the screen — files that may have taken years to assemble. One by one, they blipped off the screen. In horror, the victim would stare at the screen, unable to stop the process.

As it turned out, the booby trap was only a practical joke. It really didn't erase all the victim's files. It only listed the names, and then erased them. But for a horrifying few minutes, the victim wouldn't know this.

Hackers play games. Very interesting games.

The kind of people who spend six hours, midnight to 6 a.m., trying to break into each other's programs are different from the rest of us. Among their ranks are some highly individualistic people. Some of them are libertarians. I mean anarchists. They are electronic "don't tread on me" sorts of people. They do not appreciate bureaucracy. They appreciate being pushed around even less.

The folks at Chase Manhattan really do have a problem. Do you attempt to prosecute a legally un-prosecutable kid? A kid who has already cracked your computer system? I don't think you do. You play the role of stern but appreciative banker. "Son, I am impressed by your ability to break in. But understand, we are honest people. There is a code of honor here. You wouldn't want to break that code — of honor, I mean — would you?" Because if this kid gets angry, he can do it again. Quietly. And next time, he deposits a virus.

Of course, Chase may hire a programming team to create an unbreakable system. Sure. "Hire fox A. Give him chain link fence B. Hire him to build fence B around chicken coop C."

TEEN CHALLENGE

Suppose that the public gets wind of the threat to the whole banking system that is posed by viruses. What do the bankers (or anyone else) announce to the public? "We want to assure you that our computer program is impenetrable. No one can break in. It is foolproof."

Here is the challenge — rather like the Stanford program that announced: "Do not trespass." These kids see breaking-in as a challenge, a kind of sport. They do not regard it as theft. They do not regard it as vandalism, even if it costs a company millions of dollars to unscramble. They may be ethical in other respects, but they think of "core wars" as a game.

How would you like to be the 60-year-old banker who doesn't know a byte from usury, but whose public relations department tells him to inform the public that nobody can crack his bank's code? To cite Mr. T in "Rocky III," that bank is dead meat. So are its depositors.

But if he keeps quiet, and the story still gets out about the vulnerability of the system, one or two small "virus-demolished" banks could trigger a collapse of the system, as people do the only smart thing: run for cash. The whole fractional reserve banking system would deflate; only the FED's printing presses could "save the day," in a wave of fiat money.

What I am saying is this: I think that we will see the end of fractional reserve banking in our day. At the very least, I think we will see it subjected to tremendous shocks. People will lose faith in electronic promises made by bureaucrats who do not know anything about the monsters that their efficient computers can be turned into.

Attack on Martinsburg. Now, let's take it a step farther. Some day some state or Federal bureaucrat is going to step on the toes of some genius entrepreneur who has created a software development firm. The bureaucrat will try to wrap this entrepreneur in red tape. Or maybe — just maybe — he will try to sock him with a tax bill that the entrepreneur regards as unfair.

In Martinsburg, West Virginia, there is a large computer. It is owned and operated by the Internal Revenue Service. Into it, over the next five years, the I.R.S. apparently intends to deposit all the records it can assemble on every U.S. taxpayer. This computer database will be the biggest in the world. It is the tool by which the I.R.S. hopes to increase taxpayer compliance. And it may succeed. For a while.

This is one reason for saving all letters to and from the I.R.S. If the I.R.S. becomes dependent on its computer system, which is likely, then any short-circuiting of its database could create havoc for tax collecting. If word gets out that a major failure has hit the I.R.S., the tax revolt could multiply overnight. You would see the deficit become astronomical. If the I.R.S. continues to tie its "voluntary" compliance program to the myth of the "all-seeing computer," then news of the computer's scrambling could backfire.

It is possible that the story of the I.R.S. database is a myth. Maybe they aren't going to build it. But if the public believes that such computer power is at the disposal of the I.R.S., and taxpayers then learn either that the system has been blown, or that it was mythical from the start, the tax revolt could spread like an epidemic. The electronic epidemic-could trigger a tax revolt epidemic.

He who lives on the cutting edge of technology eventually dies on the cutting edge of technology.

"PEOPLE ARE BASICALLY GOOD"

Let's return to my taped interview with "Tom." In a 90-minute interview, we covered a lot of ground. But one topic that stands out in my mind is our discussion of the pre-supposition that goes into the creation of a computer-based society. The computer people have all adopted the assumption that undergirds modern science, namely that, participants are well meaning, that they will not fake their experiments, and that they will play fair. If scientists had to check every aspect of every article, science could not advance very fast.

What about the computer industry? The whole system rests on faith: "Men are not malevolent. They are not envy-driven. They will not deliberately seek to destroy the work of some random victim." Tom says categorically that this assumption is false. There are bad people with tremendous computer skills, and modern society has not restructured its economic institutions to protect itself.

Here is one example of a break-in technique. Someone phones in to a computer, which has been left open temporarily by some user. The lock is unlatched; he needs no key to get in. He then seeks to penetrate the inner core of the program, such as a bank's program. He creates a deliberate error, which all too often triggers a kind of electronic explosion. The protective shell self-destructs, and the invader now finds himself inside the system, where far fewer defense mechanisms exist.

Tom designed his own firm's defense against this tactic. His program automatically records the source of the error, and throws the user out of the program. The program has protection against deliberate errors, but most of them don't, he says. A major error simply collapses the program's outer shell.

In my previous issue, I speculated that a Soviet spy or agent could penetrate U.S. computers. Note: I did not assume that he would simply phone in; I assumed that a disloyal programmer, or a team, could plant the virus as insiders. From there, the virus would spread through the system through normal telecommunications. Several people have written in to tell me that a wrecker cannot destroy the system by penetrating it from the outside. They may be correct. But when informed that I am assuming an inside job by someone with access to a major computer, the critics have admitted that this might be possible.

The weed of crime bears bitter fruit: for honest, computer-dependent people.

FEDERAL FUNDS

The Federal Funds bank transfer lines allow banks to borrow money overnight. Hundreds of billions of dollars go across these lines every working day. The bank's computers communicate with each other by means of this telecommunications hook-up. What if someone were to plant a long-delay virus in the software that operates these transfers? And what banker has even thought about this problem?

What if this scenario was to take place: a virus triggers the disruption of bank records — not a total breakdown initially, but disruptions in the data? It might be weeks or months before auditors recognized the extent of the problem.

As rumors begin to leak out about complex accounting or other data-management problems of major banks all over the U.S. (including off-shore branches), the various banking regulatory agencies would be swamped with crises and outside rumors. Then, all at once, bank computers begin breaking down.

The rumors then explode. The lines appear in front of banks. The only answer at this point is to print up paper money. It would be printed by the hundreds of billions in order to offset the deflationary effects of bank runs (paper money that is pulled out but not re-deposited in another bank).

You could topple the fractional reserve banking system all over the world. The entire payments system could easily become engulfed in chaos. Debits and credits would no longer be meaningful. A pure paper money inflation would replace the manipulated, "fine-tuned" monetary inflation of modern central banking.

All of a sudden, market-created alternative currencies would be revived. It would then be metallic cash that talks loudest. Silver dimes are not electronic. "They can't be infected electronically." They still circulate when banks are "temporarily closed, due to circumstances beyond our control."

The loss of efficiency would be initially horrendous, I would guess. The division of labor would break down. You could then have the crash that lurks in the minds and suspicions of average depositors. Who says it cannot happen? A lot of public relations firms hired by the banks — computer illiterates in high places?

What we have is an internal bank money wire system that is totally vulnerable to some vindictive programmer. There is little doubt in my mind that the bankers are desperately fearful of this sort of vandalism. It could topple people's confidence in the fractional reserve banking system, and confidence is the only thing that keeps it going.

Conclusion. Technologically, there is no solution at this point. I have no heartening message. Maybe later; not now. Keep precious metal coins. Don't assume that it can't happen here. It can. The only thing holding it back is the restraining hand of God, through the temporary self-restraint of a technological priesthood.

[Reviewer's comment: I was unaware of the problem of computer worms and viruses until I read Gary North's two articles. His open release of this important information is appreciated.]

SCALAR EM INNOCULATION AND ITS IMPLICATIONS

My appreciation goes to Gary North for his openly sharing the above important information with the public. Human nature being what it is, one or more of his more dire scenarios is going to happen sooner or later.

The problem becomes enormously more compounded if the hostile agent or agents possesses sophisticated scalar EM technology.

With scalar technology, there exists the potential that the worm or virus can be implanted from a distance. If scalar-scanning devices of sufficient precision can be developed, then major software residing in a chip, a computer, or on a disk can be read from a distance. With scalar EM methods, the ordinary software "guards" are totally ineffective. The scalar "reader" could simply read off the software directly as the disk or device goes about its normal operations.

After the software is read, a germ or a virus is designed to produce a particular problem as desired and upon special stimulus. Then a scalar "imprinter" device is used to produce the necessary digital pulses directly upon the media in which the software resides. The required code has now been implanted. In effect, the operating system has been injected from a distance, and contaminated with the desired "sleeping agent" — germ or virus. From there on, it's just like Gary North's scenarios.

So the scalar EM technology can probably be used (at least eventually) to place "mines" in systems in advance, for damage or destruction at a future date. At least it's possible in principle.

However, it's not the injection of computers and equipment that bothers me. It's the injection of human beings. Here's generally how it can be done (and possibly is being done).

Basically, human brains have various rhythmic waves, which represent "envelope pulsations" of the entire electrical system. However, it is not the Et, force fields represented by these pulsations which are important: it is the summed-zero (scalar) patterning of the entire complex of neuron firings which represents the real activity of the brain — and represents its thoughts and emotions. A storage system for these patternings exists, but it's a scalar pattern storage, not an E and B field storage. The storage system is in the nuclei of the tissues of the nervous system itself, not in electrical or electrochemical potentials.

At any rate, the human brain rhythms are also "riding along" relatively entrained to the Schumann resonance of the earth's magnetic field. So all the patterning of the brain is grouped together into giant overall orders, and one or more of these is normally synchronized to our natural environment.

If a varying magnetic signal of extremely low frequency — say 10 hertz — bombards the brain with greater power than the earth's magnetic field, it is possible for the brain's rhythm to be "forcibly entrained" to synchronize with that signal, rather than with the normal Schumann resonance. If a large number of brains are forcibly entrained to such a signal, and the signal is artificial, then a nester coherent phasing of all the brains to the transmitted signal has been accomplished.

And that can be utilized — and is utilized — by the Soviet Union. On the Woodpecker signals, Bob Beck has detected up to 16 carriers, each with a 10-Hertz modulation, and with all the 10 Hz modulations in phase.

Let s stress what we're saying here. At times, the Woodpecker signals have shown 16 carrier frequencies that contain the same 10-Hertz modulation, and which are all kept coherently sync-locked.

If those signals and the modulations are strong enough, the brains in a targeted area demodulate the 10-Hertz signal and phase-lock to it — actually, to them. At that point, the brains are entrained to a 10-Hertz driver, and 16 channels of multimegahertz frequencies are locked-in to the brain.

Now, if the Soviets have broken the genetic code of the human brain (which they have) and succeeded in developing the ability to implant thoughts, emotions, pictures, etc. (which at least to a degree they have), then they can simply record the proper signals from human guinea pigs, using scalar technology and instrumentation.

They can tailor those signals and add them onto the Woodpecker carriers, using scalar EM techniques. By using most of the sixteen carriers as a means to express Fourier frequency expansions, specific areas and functions of the brain can be targeted.

"Notice that the distant transmitters already "know" the correct phasing, just use the 10 Hertz modulation signal as a reference. In that case, phasing takes care of itself inside the captured human brains.

The brain has a unique characteristic. If a signal arises inside it, it interprets that it itself originated the signal. Thus for the first time in history, true unwitting robot actions and emotions can be engendered in masses of human beings.

But it's still worse. The Kaznacheyev experiments (see references) proved conclusively that electromagnetic signals can carry and transmit cellular disease and death patterns into a targeted biosystem. You can rest assured that the Soviets continued on to actually measure and determine a wide variety of death and disease scalar patterns. All of that can be added on to the Woodpecker carriers so that the population in the intersection grid area is devastated. One can even test it out in advance by producing a modified pattern of a disease in the targeted populace, and keeping the percentage of brains entrained down to a small level. This lowers the number of persons "injected" from a distance to a small percentage of the populace. Then the outbreak of the "new strain" of a familiar disease — such as influenza, pneumonia, and something like AIDS — can be monitored to provide targeting and effectiveness data. Notice that a series of diseases was experienced by U.S. personnel assigned to the U.S. Embassy in Moscow, probably as a result of the so-called "weak microwave radiation" of the Embassy by the Soviet Union.

But that's not all. Much more subtle effects can be produced in the targeted populace, without anyone getting wise. Cancers and leukemia should be relatively easy to induce. Also, the immune system can simply be attacked directly. If done slowly, a series of common, ordinary diseases will just seem to get "tougher and tougher to cure." Actually it's not the disease that got tougher, it is the immune systems that got weaker. It is my opinion that precisely that sort of testing — against the immune system — has been accomplished by the Soviet Union against the U.S. since about October 1985. A great percentage of influenza patients during the period from then until now have had complications. So much so that multiple sessions — as many as 3 to 5 — of antibiotics are necessary before the patient recovers. At least in the Huntsville, Alabama "hinge pin" area, that seems to have been the case. This author personally experienced this, having influenza and follow-on complications for a period of seven months — something quite different indeed for him.

But it's even worse.

The human scalar EM nervous system ("inside" the ordinary EM signals and functioning of the system) also has its own resident "software." It has special storage media contained in its brain, its cells, and its genetic material, and its actual nucleons. All of these storage media are continually telecommunicated with via scalar EM means. Those recordings, largely condition the "stability" of what each of us is. As personalities we literally "are" what we have experienced, felt, thought, and did. Our whole "functioning conscious system" has been evolved by our experience, and it continues to be "us" by means of continual reference to the record base.

I hope you now realize that the actual record base itself can conceivably be altered by scalar electromagnetic means, even from a distance. Even by the hidden scalar EM content of 16 Woodpecker carriers, rocked into a captured brain and pouring in their deadly content. It is conceivable that even what one is can be taken from him, and what someone decides he will be, he will be.

A totally new — **and** ultimate — form of "slavery" has emerged.

Also, look at all this scalar EM functioning of the human system as a special kind of cybernetics. A scalar EM computing software and hardware system.

What if scalar techniques are used to plant "worms" and "viruses" in this system? With hidden triggers? What about a series of them, so that a range of eruptive software diseases can be stimulated at will?

And guess what. Such scalar EM "worms" and "viruses" can even be induced by contact with or consumption of "contaminated" water and food. Or saliva. Or other body fluid.

So you can see the possibilities of "seeding" a population with a staggering variety of time-bombs.

I think you begin to see the point. Now you can change North's "Soviet blackmail" scenario drastically. In a new scenario, Gorbachev simply announces to Reagan that a graduated series of specific sweeping changes are going to happen in our population — changes which we are absolutely powerless to stop. The last stages are various deaths. The master "trigger" for this series has already been initiated. Incidentally, the Soviet populace has been carefully "seeded" with the proper antidotes. The Russian leader demands certain things (very similar to what North's scenario utilized). Unless his demands are met, a "killer virus" will be activated by a special signal. If so, everyone dies. The first step change in the series will start to occur in a few hours. And it does.

In fact, this scenario could be added to North's scenario, and both used simultaneously. What do you think the President and the Congress would do? Or could do?

There are other even more diabolical possibilities, but these should be quite sufficient to illustrate the point.

Pandora's box has already been spilled, and the end of humanity is ticking away like a time bomb. It's already nearly midnight, and the watchman hasn't even sounded the alarm yet.

I certainly can anticipate the next question: Have the Soviets already "seeded" the populations of the U.S. and other selected western nations with scalar EM viruses and germs7

The answer is, I simply don't know. And just like Gary North, I don't know what if anything can be done about it if they have. This researcher certainly doesn't have all the answers. An immediate national crash program to develop scalar EM defenses is essential if we are not simply to be crushed beneath the hammer and sickle. It takes funding, people, facilities, and time.

In some 22 years of unrelenting struggle to convince the system we've got a terrible problem, I've hardly been able to move it off dead-center (no pun intended). This book, produced under extreme workload stress, is all I can do to raise the alarm. I've given it my very best shot.

What happens now remains to be seen. Other hands, hearts and minds are going to have to take up the task and carry it forward. This researcher has reached his limit, and has no more left to give.

APPENDIX III: CONTINUING THE SCENARIO

On May 3, 1986 a stunning loss of a U.S. Delta rocket carrying a critically-needed weather satellite occurred. The rocket suffered a shutdown of its main engine during launch and began to veer off course, causing the range safety officer to destroy it.

The takeoff of the blue and white Delta "workhorse" from launch Pad A at Cape Canaveral at first appeared picture-perfect. About a minute later, six of the nine booster rockets were expended and the remaining three switched on. Shortly after that, the rocket seemed to lose power. The main liquid-fueled engine had suddenly shut down, with the three strap-on boosters still burning. The rocket then had no stability and started to slowly drift off course, creating an increasing angle of attack. Within a few seconds, the nose cone snapped off under the severe stress. The U.S. Air Force range safety officer then sent a destruct command to the stricken rocket, which exploded in a ball of flame.

The Delta, built by McDonnell Douglas, had previously piled up a continuous string of 43 successful launches. Since 1960, 177 launches have occurred, and the success rate has been 94%. The Delta is one of the most reliable rockets ever built.

The unexpected shutdown of the main engine was perplexing. At a news conference following the mishap, William Russell. NASA's Delta project manager. stated: "It was a very sharp shutdown. almost as if it were a command shutdown."

Preliminary data, such as turbine speed and temperatures in and around the engine, "all appeared to be very normal." Russell said.

It may well have been a "command shutdown," unknown to Bill Russell and his NASA engineers.

Preliminary runs of the videotapes of the incident, aired by the national TV news media, show that during liftoff, a mysterious light approached the rocket from below and from the side. rising and striking the vehicle. This light and its path are strongly reminiscent of the first anomalous "light streak and ball" photo taken by Bob Gladwin just prior to the shuttle launch on November 26, 1985.

Let' s suppose that on May 3 the Soviet scalar EM grid was in the launch Phase ABM system mode and was tracking the launch of the Delta rocket.

If the anomalous light that moved up and contacted the rocket was a small ball of continuous electromagnetic energy formed by a distant Soviet scalar EM howitzer, then the rocket's main engine control circuits would have received a sudden and continuous pulse of EM power. This would probably have shutdown the engine by an actual "command" signal, artificially delivered. *

---

*Engineers have now determined that such a shorting of the engine circuitry occurred, causing engine shutdown.

Certainly such mysterious switching of satellite circuits has occurred before, probably due to Soviet scalar EM precision testing. For example, the British Ariel VI satellite has exhibited very strange anomalies in its command and control circuitry -- in fact, with very precise control switching being exhibited as if from an external source.

Thus a very small scalar EM missile may have been utilized by the Soviets to disrupt the control circuits in the rocket. A resulting pulse of power in the engine shutdown command circuit could have "ordered" the shutdown of the main engine -- just as it appeared.

It remains to be seen whether or not the Soviets used a "marker beacon" associated with the disruption of the Delta launch. One news photograph seems to contain a suggestion of such a light, but the videotapes must be studied frame by frame before it can be ascertained exactly what if anything is really there.

If a marker beacon was not used, the distant Soviet operator could simply have "popped out" the continuous EM ball of energy, tracked it and the ascending rocket exhaust, and "joysticked" the scalar EM missile up into the rocket.

Analysis of this incident is continuing, so nothing is certain as we go to press.

However, if this incident was indeed another Soviet kill, then it will have served the Russians well. Much of the world's media attention will now be focused on NASA's launch difficulties, and away from the recent Chernobyl nuclear reactor disaster in the Soviet Union. It's bound to take some of the media pressure off the Soviets.

In addition, it further cripples the U.S. satellite program. Loss of this critical weather satellite payload seriously reduces U.S. weather monitoring capabilities.

Further, the U.S. space program now appears to be all but grounded by the Challenger, Titan, and Delta explosions. For example, it will require about 18 months for the U.S. to redesign the shuttle so that America's shuttle program can get going again. So things are delayed until summer of 1987 at the earliest.

Meanwhile, the Russians are set to grasp the lead in the space race. In 1985 they launched about 100 rockets, five times as many as the U.S. They have a shuttle program of their own, well underway, with a flight expected within months. Earlier this year, they launched Mir, a second generation space station; a U.S. version is still some eight years away.

Charles Vick, of Huntsville, Alabama's Space and Rocket Center, summed it up: "They have their standard military program, a shuttle program and the equivalent of an Apollo program, all within a 10-year period. It's awesome."

An experienced propagandist like Gorbachev certainly understands the advantages of a strong Soviet push into space, such as (1) military advantage, (2) national prestige, (3) scientific achievement, and (4) projecting an image in the world's news media that the Soviets are first in space exploration.

So the failure of the Delta rocket provides many benefits to the Soviet Union.

It also continues to exercise and demonstrate Soviet mastery of the skies over North America itself.

Meanwhile, two U.S. nuclear submarines have "inadvertently" run aground in the last month or so.

Gorbachev again seems disposed toward a summit meeting soon.

If it's held, the agenda may be quite different from the one President Reagan is prepared for.

The clock ticks on toward midnight.

2002 ENDNOTES

1. It is actually a unified field theory since the infolded electrodynamics inside all EM potentials, fields, and waves is utilized—and the internal electrodynamics forms spacetime curvatures and dynamics of those ST curvatures.

2. M. W. Evans, "0(3) Electrodynamics," *Modern Nonlinear Optics,* 2nd Edition, M. W. Evans (Ed.), Wiley, New York, 2001, Vol. 2, p. 79-267.

3. David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II); — "Quantum theory as an indication of a new order in physics." A:, *Found. Phys.,* Vol. 1, 1971, p. 359; B: "Implicate and explicate order in physical law," *Found. Phys.,* 3(2), 1973, p. 139-168.

4. Mendel Sachs, General *Relativity and Matter,* Reidel, 1982; — "Symmetry in Electrodynamics: from Special to General Relativity, Macro to Quantum Domains," *Modern Nonlinear Optics,* 2nd Edition, Wiley, New York, Vol. 2, p. 677-706.

5. M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *J. New Energy,* 4(3), Winter 1999, p. 299-301 gives a basic mathematical expose of scalar interferometry.

6. James Clerk Maxwell, "A dynamical theory of the electromagnetic field," *Phil. Trans. Roy. Soc,* Vol. 155, 1865, p. 71, 459. Orally presented in 1864. This was Maxwell's definitive presentation of his theory, with 20 quaternion and quaternion-like equations in 20 unknowns. Also in *The Scientific Papers ofJames Clerk Maxwell,* edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-604.

7. In T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991 we first exposed the quantum potential weapon and some of its uses and characteristics.

8. A negative energy EMP weapons strikes the target or target area with an extreme pulse of negative energy empty Dirac Sea 4-holes. As a result, the free electrons flowing in electrical currents in the struck target instantly disappear by re-absorption into the Dirac Sea. A living body struck by such a blast is totally and instantly dead in every smallest part, since all its brain and nervous system currents, cellular currents, etc. are instantly and totally extinguished. The Russians tested this weapon in the 1980s in their war in

Afghanistan, against two Afghan villages. I spent two hours in Washington D.C. with the then-representative of the Mujahedin to the U.S. Government, verifying the incidents and going over the exact characteristics of the instantaneous deaths of those villagers. With such a sudden death, the mind-body interconnection loop is instantly severed, so that the mind is instantly snapped loose from its connection to the body—hence my use of the term "MindSnapper." The struck body usually emits longitudinal EM waves for an extended period, so that vultures and other predators will not approach the bodies and eat them. All bacteria, viruses, etc. in the body are also killed instantly, and the body does not decay, even in 30 or 60 days.

9. As can be seen, the negative energy EMP weapon is an all-purpose weapon extremely effective against every kind of target: missiles, aircraft, communications, satellites, ships and particularly aircraft carriers, personnel, generators, power systems, the electronics of nuclear warheads, submarines, tanks and vehicles, and the electrical and electronic controls of any kind of machinery or equipment. Such weapons are the second most powerful weapons on earth. In the first part of 2002, a U.S. 2-carrier task force maneuvered in the South China Sea as a precautionary warning to the Chinese that we would defend Taiwan. Unknown to that fleet, tracking it and trained on it were the newly deployed Chinese negative energy EMP weapons, which could have destroyed the entire task force—everything in it—with a few shots in less than 10 minutes.

10. T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002. Chapters 1 and 2 give a deeper coverage of this and similar foundations principles.

11. As is well known, Russian theoreticians have always been the best nonlinear scientists on Earth, since the beginning. They still are.

12. See T. E. Bearden, "Giant Negentropy from the Common Dipole," *J. New Energy,* 5(1), Summer 2000, p. 11-23 for solution of the source charge problem. On DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org. The problem of the source charge can be solved by treating (with the modern quantum field theory viewpoint) the "isolated" observable charge and its cluster of virtual charges of opposite sign as a composite dipole. Because of the broken symmetry of its opposite charges, any dipole or dipolarity continuously receives virtual energy from the seething vacuum, transduces it into real observable EM energy, and pours out observable photons in all directions at the speed of light. See also E. T. Whittaker, 1903, *ibid.* There is a scalar potential between the ends of each dipole. The scalar potential may be decomposed into a bidirectional set of longitudinal EM phase conjugate energy from the time dimension (from the imaginary plane) and outputs that EM energy in the real 3-space. 3-space conservation of energy flow

is violated, but 4-space conservation of energy flow is upheld. This rigorously follows from the discovery of broken symmetry of opposite charges in 1957—one of the types of asymmetry for which Lee and Yang received the Nobel Prize.

In classical electrodynamics, the charged particle itself is *assumed* to be the source of its own fields and potentials. In other words, the charge is implicitly considered to be *creating*—right out of nowhere—the Poynting energy flow it continuously pours out across the universe in all directions. That makes the charge a perpetual motion machine of the grossest kind, violating the maxim that energy can neither be created nor destroyed.

So the greatest perpetual motion advocates of all times are our own classical electrodynamicists and electrical engineering departments and professors who still ignore the active vacuum interaction that is missing from the classical model. They thereby ignore that the charge is a broken 3-symmetry in its violent exchange of virtual energy with the active vacuum, and merely organizes and emits in observable form a fraction of the energy it continuously receives from the imaginary plane (from the time domain). Yet, that broken symmetry has been known in particle physicists, both experimentally and theoretically, for 45 years.

When Maxwell wrote his theory, neither the electron, the atom, nor the nucleus had been discovered. There was no knowledge of the Drude electron gas, electron spin, etc. There was no knowledge of the nucleus of the atom, and even the atom itself was still contentious. A thin material ether (fluid) was believed to fill all space, so that no point in all the universe was empty of mass. Consequently, E and B were conceived as physical force fields in a material medium—the luminiferous ether. All Maxwell's equations—including the subset selected and modified by Heaviside that today are called "Maxwell's equations"—assume this material ether to this day.

Electricity was thought to be a "thin material electric fluid" running through wires like water through a pipe. The magnitude of the potential was not used; its reaction cross-section was calculated and erroneously used as its magnitude, as it is today (although a good electrodynamicist does know that it is only the *intensity* of the potential, not its magnitude).

Faraday believed that his "lines of force" were taut physical "strings" in this material ether, and that EM perturbations were the vibrations of these taut strings. As he clearly states in his 1873 *Treatise...*, Maxwell sought to capture Faraday's work exactly. He simply *assumed* Faraday's "plucked taut string" waves. Later the measurement in instruments of what was actually the lateral gyroprecession of electrons in receiving wires was thought to prove that the incoming ether vibrations were transverse, like those string waves, because the detected electric **fluid (electron** precession) waves in the receiving wire were transverse. The drift velocity of electrons longitudinally in a wire (in a typical case, several inches per hour) was unknown. A charge *q* was just a piece of electric fluid, much like a cubic centimeter or other volume of fluid; it had no connotation of "charged particle." There was no concept in all of physics of EM energy flowing through space. There was no knowledge that the same fields created in the transmission wire, which perturbs the Drude electrons laterally, also perturbs the positively charged nuclei laterally, in the opposite direction, highly damped but with equal energy.

There was no way to realize that the energetic perturbation of the curvature of spacetime around the sending wire antenna, by the Drude electron perturbation energy, is simultaneously accompanied by an equal and opposite perturbation of the curvature of spacetime in the opposite direction (energy density, not amplitude of the disturbing wave, determines the degree of curvature of the surrounding spacetime). Hence Maxwell omitted precisely one half the EM energy, one half the EM wave, and one-half the electrodynamics. As a consequence, the "Newton's third law recoil" of the nuclei today is thought to spontaneously arise as if some mystical demon observes the Drude electron perturbations, and simultaneously kicks the positively charged nuclei in the equal and opposite direction. Hence Newton's third law, so to speak, is erroneously still omitted from classical electrodynamics, and the better texts say so.

In the 1880s, the Michelson-Morley experiments falsified the material ether. Not a single one of the Maxwell-Heaviside equations were ever changed to this day. Instead, the electrodynamicists eventually announced that, since there was no longer an ether, they were not using one! Then a total non sequitur was incorporated: that in massfree space, the same E and B fields (defined as material entities and thus existing only in mass) still remain but the forces go to zero. Simple algebra reveals this clever, insidious faux pas, using the E-field as an example: First, F == d/dt(mv), so mass is a component of force. (This reveals the great error due to the hoary old notion that a separate force acts upon a separate mass; the mass is a component of the force, and the force is a coupled system of a curvature of spacetime and a charged mass.). Then E == F/q, which rigorously holds when mass is present.

Now for whatever reason, let F = 0. Then F/q = 0 and E = 0 also, else we must discard algebra. What actually remains is easily explained in general relativity terms (available since before 1920). The couplet of a curvature of spacetime and a charged mass is broken up by absenting the mass. What remains is that specific curvature of spacetime. GR assures us that curved ST acts upon mass, and that any force involves a curvature of ST. Yet in more than 80 years, this erroneous old "material ether remnant" in electrodynamics has been fiercely defended. Every textbook today uses and confuses the E and B fields in two dichotomous fashions: (1) as massive force fields existing only in the presence of mass, and as "massless force-free fields" existing only in the absence of mass. This of course is a non sequitur.

13. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.*, Vol. 57, 1903, p. 333-355. The "scalar" potential itself is not even a *scalar entity,* as shown rigorously by Whittaker. Instead, it is a harmonic series of bidirectional EM longitudinal phase conjugate wavepairs, and it decomposes into that series. Hence it is a *multiwave, multivectorial entity.*

14. C. N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance," *Phys. Rev.*, Vol. 96, 1954, p. 191.

15. The term "energetics" simply means *energy as the primary thing in the universe.* Rankine first used it. G. Helm in 1887 revived Rankine's term "energetics" and factored energy into an *intensity* factor and an *extensity* factor (which today we would call potential and kinetic energies). William Ostwald in 1893, 1895, and 1901 raised the concept of "energetics" (energy as the primary thing in the universe) to the forefront. Thus when the Russians developed new extensions and principles in weapons physics, they simply used the historical notion of "energy as the primary thing in the universe" and adopted the earlier name for it: *energetics.*

16. Later when Khrushchev demanded a 100% antimissile defense, famed Russian Nobelist Petr Kapitsa answered, *"If a means of total neutralization of foreign missiles is to be found, it can only come from a group of new principles in physics, called energetics".*

17. See T. E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", *Explore,* 9(2), 1999, p. 59; Part II, *Explore,* 9(3), 1999, p. 61; Part III, *Explore,* 9(4,5), 1999, p. 100-108. It appears that Sachs' unified GR and EM approach is perhaps the best Western physics model to apply to analysis of the KGB energetics weapons. However, because of the third branch *(psychoenergetics:* engineering the mind directly), Sachs' work would appear to need further extension along the lines indicated by the present author in the cited paper.

18. See T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy,* 3(2/3), 1998, p. 12-28; — *Energy from the Vacuum,* 2002, *ibid.* For nuclear transmutation at a distance (e.g., dudding all nuclear weapons by transmuting U235 and weapons grade plutonium), extension is required along the lines indicated by the present author in these references.

19. See M. Sachs, *ibid.,* also M. Sachs, "Relativistic Implications of Electromagnetic Field Theory," in T. W. Barrett and D. M. Grimes, eds., *Advanced Electromagnetism,* World Scientific, 1995, p. 551. The Sachs theory

already expresses the greatest generalization possible of electrodynamics, and fully justifies the preceding view.

20. See Whittaker, 1903, *ibid.;* — "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc. Lond. Math. Soc,* Series 2, Vol. 1, 1904, p. 367-372. The latter paper was published in 1904 and orally delivered in 1903. Per these two Whittaker papers, it is known that (i) ordinary EM fields and waves are the result of interference of two or more scalar potential functions, (ii) scalar potentials are comprised of harmonic sets of bidirectional EM phase conjugate wavepairs, (iii) all ordinary EM fields and waves and potentials have LW EM dynamics internally, and (iv) in theory, this "infolded" electrodynamics can be engineered. Sachs' theory dramatically extends the richness of that internal electrodynamics ignored in the West, and highlights the general relativity (sets of spacetime curvatures, or "engines", involved).

21. E.g., *Marine Observer* was a rich source for Russian energetics weapons testing in remote ocean areas, sighted by ships' captains and crews and entered into the ships' logs.

22. Portions of Khrushchev's statements to the Presidium are given in Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'," *New York Times,* Jan. 15, 1960, p. 1.

23. See Whittaker 1903 and 1904, *ibid.,* for the original basis of scalar interferometry and its creation of any kind of EM field, potential, and wave in the interference zone.

24. Also see M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry, *J. New Energy,* 4(3), Winter 1999, p. 76 for mathematical proof of scalar interferometry. There are several other AIAS papers in the same issue of *JNE,* dealing with interferometry and with the implications of the Whittaker papers.

25. As an example, the U.S.S. Skylark, surface companion of the Thresher, was beset by mysterious "jamming" of multiple types of EM systems in multiple bands. So bad was it that it required more than an hour to get an emergency message to Naval headquarters that the Thresher was in serious difficulty. Later, as the scalar interferometry died away, all the EM systems aboard the Skylark resumed normal functioning. The Skylark was in fact in a "splatter zone" of the underwater interference zone around the Thresher.

26. Circa the end of 1989, the earlier longitudinal EM wave weapons on site in Russia were leased to a rogue Japanese group comprised of Aum Shinrikyo and Yakuza, who have been operating those weapon systems—on site in Russia—since then. Senator Sam Nunn's investigations revealed the presence

of the Aum Shinrikyo in Russia, to include establishing a university there. His investigators missed the Yakuza involvement and also missed the exotic weaponry aspects as one might expect.

27. I concluded from this remark that TEMPEST techniques are probably used at some nearby site not too distant from my home here in Huntsville, to read and record my work as I write it on the computer. That appears to be a good assumption.

28. As a small example of what can be done, see a series of papers by M. W. Evans, P. K. Anastasovski, T. E. Bearden et al, published in *J. New Energy*, 4(3), Special Issue, Winter 1999. The papers are: (a) "On Whittaker's Representation of the Electromagnetic Field in Terms of Two Scalar Potentials, Part I"; (b) "An Experimental Test of the Existence of Whittaker's *g* and/Fluxes in the Vacuum"; (c) "On Whittaker's Representation of the Classical Electromagnetic Field in Vacuo, Part II: Potentials Without Fields"; (d) "On Whittaker's F and G Fluxes, Part III: The Existence of Physical Longitudinal and Timelike Photons"; (e) "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-like Potentials: Canonical Quantization"; (f) "On Whittaker's Analysis of the Electromagnetic Entity, Part IV: Longitudinal Magnetic Flux and Time-Like Potential Without Vector Potential and Without Electric and Magnetic Fields"; (g) "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry"; and (h) "On Extending Whittaker's Theory, Part VI: Photons Without Fields and Vector Potentials".

29. As an example, one part of the strategic deception plan was devoted to convincing the West that Soviet psychoenergetics was nothing but a technical wrinkle upon the rather harmless Western parapsychology. This part of the program was completely successful for decades. Another very clever use of deception is to allow release of some real scientific advances that lead the West, but are not too important to the KGB in the weapons arena. The increasing emphasis on torsion fields is an example of this "deception by giving good information that directs the adversary upon a path one wishes him to take", effectively steering him away from the path one wishes him to avoid. Lately, however, there is an uproar amongst the academicians in the Russian Academy of Sciences over this very point, that torsion fields being presented as the "end all, be all" is simply disinformation.

30. With their deployment of QP weapons in 2002, the Chinese are also on the way to having this capability.

31. The multiply connected spacetime (MCS) or quantum potential weapons went on site circa latter 1989. A laboratory prototype had been tested against the U.S. air attack on Libya in April 1986. For details of the extremely anomalous jamming **of aircraft** and missile guidance and control systems in that

attack on Libya, see Mark Thompson, "Mixed signals may have misguided U.S. weapons," *The Washington Post,* Jan. 22, 1989, p. A-4.

32. One or more QP weapons also apparently induced the Gulf War Disease in MidEast forces in the Gulf War. We have previously explained the peculiar anomalies, how it was done, why the endemic natives did not contract the disease, why the French forces did not contract it but U.S., British, and Canadian forces did, etc. We briefly discussed quantum potential weapons in T. E. Bearden, *Gravitobiology: A New Biophysics,* In T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991.

33. (a) Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," *Jane's Defence Weekly,* 1(3), Apr. 7, 1984, p. 508 gives something of the two Soviet tests of this weapon in Afghanistan. In addition to the fierce blast of empty Dirac Sea 4-holes, these weapons produce a powerful time-polarized EM pulse in the area, producing a very sharp "jerk" in the rate of flow of time, (b) The mind-body coupling mechanism—which cannot withstand such a sharp jerk in the rate of flow of time—has been given by T. E. Bearden, "Mind Control and EM Wave Polarization Transductions," *Explore,* 1999, *ibid.* This sharp "jerk" in time-flow instantly snaps the mind-body coupling mechanism for any living thing in the struck area. A very small version of this weapon was tested twice in Afghanistan, where the bodies in the "mind-snapped" zone died instantly, fell like limp rags, never twitched, and did not decay or get molested by insects or scavengers for the 30 days it required the Mujahedin to fight their way back into the two struck villages. The Mujahedin thought it was some new, extremely fast acting nerve gas, which they dubbed "smerch" or "death wind".

34. Two Ph.D. scientists were working closely with me on this entire effort, as was another close colleague. There was much more involved in the effort that is so briefly mentioned here.

35. Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

36. T. E. Bearden, *Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications,* CTEC Proprietary, Mar. 21, 1998, 200+ page inclosure to CTEC Letter to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. Copies were sent to various other U.S. government agencies, including to NIH, CDC, the USAF, etc.

37. See Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981. An English exposition of much of the Kaznacheyev material is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76. The work of Vlail Kaznacheyev led to the understanding of how to electromagnetically transmit cellular death and disease into targeted cells (or bodies) at a distance, and involved thousands of lab experiments. Highly sanitized work revealing the results but not the mechanisms, were permitted in order to stimulate the West to see if it knew about energetics.

38. See Stanislov Lunev and Ira Winkler, *Through the Eyes of the Enemy: Russia's Highest Ranking Military Defector Reveals Why Russia Is More Dangerous Than Ever,* Regnery, Washington, D.C., 1998. The Red Chinese also were able to decipher the Russian EM disease transmission methodology, early on. As a strange kind of "poetic justice" action the Chinese induced cancer into many Russian spies stationed in the Russian Embassy in China. Quoting Lunev, p. 160: "...I *was not the only person who was stationed in Beijing to come down with cancer...Almost all of the GRU officers stationed in China with me have been diagnosed with cancer. People assigned to the embassy had told me that they were picking up high levels of radiation from the apartment building around the embassy compound. Whether it was a form of high-tech monitoring or a conscious effort to harm our health, it seems to have been successful. So much so that in 1992 the GRU was asked to investigate the high rate of cancer among embassy officials. The findings, however, have never been made public; in part, no doubt, because of the rapprochement between Russia and China."*

39. See my "Porthole Briefing", on www.cheniere.org. Also see "Le Probleme Priore," *Rapport de la Commission de I'Academie des Sciences a Monsieur le Ministre d'Etat charge de la Recherche et de la Technologie,* France, 1982, p. 1-22; J. B. Bateman, "Microwave Magic," Office of Naval Research London Conference report, ONRL C-14-77, 1977; —*A Biologically Active Combination of Modulated Magnetic and Microwave Fields: The Priore Machine,* Office of Naval Research, London, Report R-5-78, Aug. 16, 1978; —"Staging the Perils of Nonionizing Waves," *European Scientific Notes,* ESN 32-3-85-88, 1978; T. E. Bearden, "Energetics Update and Summary," Parts 1, *Explore More,* No. 1, 1996, p. 13-18; Part 2, *Explore,* 8(1), 1997, 53-56; Part 3, *Explore,* 8(2), 1997, p. 42-45; Christopher Bird, "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Explore, 5(5-6), 1994, p. 97-110; R. Cambar, "Rapport general des travaux de la Commission de Controle constituee en vue de verifier l'un des effets biologiques obtenu par l'utilisation de l'appareillage de Priore A. Bordeaux," [General findings of the work of the control commission formed to verify one of the biological effects obtained by use of the apparatus of A. Priore at Bordeaux],

1969; R. Courrier, "Expose de M. le professeur R. Courrier Secretaire perpetuel de l'Academie des Sciences fait au cours d'une reunion a l'Institut sur les effects de la machine de M. A. Priore 26 Avril 1977" [Presentation by Professor R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore.]; Jean-Michel Graille, *Dossier Priore: A New Pasteur Affair,* De Noel, Paris, 1984 [in French]; Antoine Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; A. Priore, *Guerison de la Trypanosomiase Experimentale Aigue et Chronique par L'action Combinee de Champs Magnetiques et D'Ondes Electromagnetiques Modules* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], thesis submitted in candidacy for the doctoral degree, 1973—rejected when the project was suppressed.

40. Eric Perisse, *Effets des Ondes Electromagnetiques et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale* [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984. Although Priore's own thesis was suppressed in 1973, Pautrizel nevertheless finally succeeded in getting this doctoral thesis by Perisse approved on the work, at the University of Bordeaux which had rejected Priore's thesis, after eleven more years had passed. Considering the viciousness of the suppression, this was a major accomplishment.

41. The photo was taken by George Suchary of the registration point (ball of light) high over the Nov. 1985 night shuttle launch from Cape Canaveral. For a photo of the same kind of "ball of light" registration point above the kill of a Titan missile launched from Vandenberg AFB in early 1986, see reference 39 below.

42. The picture by Canadian investigators showing the hole burned through the right fuselage ahead of the engines of the Arrow DC-8 is shown in T. E. Bearden, *AIDS: Biological Warfare,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org , 1988, p. 52.

43. See T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991, p. 82. Two sensors on-board the Arrow DC-8 independently recorded the sharp electrical spike generated by the strike of the EM missile.

44. "Titan Explosion Cripples U.S. Launch Surveillance Capability," *Aviation Week & Space Technology,* 124(17), Apr. 28, 1986, p. 16-19. Specifically see the picture of the explosion on p. 18. It clearly shows the Soviet marker beacon (little ball of light) well above the missile explosion. General background on the Titan explosion and investigation is given in "Investigators Search for Clues to Cause of Titan Blast," *Aviation Week & Space Technology,* 124(17), Apr. 28, 1986, p. 19-20.1 persuaded NBC engineer Ron Cole with a team of engineers to examine the videotape of the explosion, frame by frame. In 17 frames the ball-of-light (registration point) is visible, and moves away completely independent of the explosion. No "bogies on the range" were reported at Vandenberg AFB.

45. Dr. Evans is co-discoverer with Vigier of the B(3) field, now well established but still not in the textbooks. He is the Director of the Alpha Foundation's Institute of Advanced Study (AIAS) and has over 600 papers and books in the literature.

46. M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "On Whittaker's F and G Fluxes, Part III: The Existence of Physical Longitudinal and Timelike Photons," *J. New Energy,* 4(3), 1999, p. 68-71; — "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-Like Potentials: Canonical Quantization," *ibid.*, p. 82-88; — "On Extending Whittaker's Theory, Part VI: Photons Without Fields and Vector Potentials," *ibid.*, p. 79-81.

47. Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): *Essays on the Formal Aspects of Electromagnetic Theory,* World Scientific Publishing, River Edge, NJ, 1993, p. 6-86; — "On the distinction between fields and their metric: the fundamental difference between specifications concerning medium-independent fields and constitutive specifications concerning relations to the medium in which they exist." *Ann. Fond. Louis de Broglie,* 14(1), 1989, p. 37-75; — and D. M Grimes. [Eds.] *Advanced Electromagnetism: Foundations, Theory, & Applications.* World Scientific, (Singapore, New Jersey, London, and Hong Kong), Suite 1B, 1060 Main Street, River Edge, New Jersey, 07661, 1995. In the latter book see T.W. Barrett, "Sagnac Effect: A Consequence of Conservation of Action Due to Gauge Field Global Conformal Invariance in a Multiply-Joined Topology of Coherent Fields," p. 278-313.

48. But see Craig F. Bohren, "How can a particle absorb more than the light incident on it?" *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327. By *resonating* the intercepting charged particle, its reaction cross section increases, and more EM energy is collected and re-radiated by the charge than is input to it by the conventional accounting (which only accounts the input Poynting energy that would be intercepted by a *static* charge). The conventional input energy

calculation therefore ignores the huge nondiverged Heaviside component surrounding the Poynting component and wasted. The resonant particle sweeps out past the usual Poynting component into the unaccounted Heaviside input component, and collects excess energy. The Bohren experiment with resonant charged particles produces COP =18. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another. For independent confirmation of the Bohren experiment, see H. Paul and R. Fischer, {Comment on "How can a particle absorb more than the light incident on it?'}," *Am. J. Phys.,* 51(4), Apr. 1983, p. 327.

49. E. T. Whittaker, 1903, *ibid.*

50. A beneficial use, e.g., would be to use such CSRs to investigate and report the interior of the sun and planets, including the Earth and its core. Since a CSR can also travel "through" the interior of a light beam, in theory such could be dispatched back into some of the nearer stars to also explore and report their interior.

51. This is no more curious than stating that sets of waves and their dynamics can exist in space, without interacting with mass, and that they can then later interact with a mass system to change it in specifically designed manner. Since longitudinal EM waves and their dynamics are "infolded" inside ordinary EM waves, fields, and potentials, they are also infolded "inside" the ambient vacuum potential, which itself is a part of the local vacuum potential. In short, one forms highly complex LW wavesets and functional dynamics, in essentially a sort of "subspace" as popularized by the Star Trek television series. The LW "engines" are identically ST curvature engines. A close colleague of the present author has in fact already invented and rigorously demonstrated such a communication system, but his former funding by several very large communications companies collapsed before he could develop and employ this system for the commercial market.

52. This is easily understood if one shifts one's vision from "space" to "subspace." Subspace is filled with (literally is composed of) longitudinal EM waves and their dynamics. The CSR's are quite unique, enormous systems "singing away" with longitudinal EM waves. So a longitudinal EM wave interferometer can easily locate them by sweeping its beams through subspace, as a radar sweeps to track aircraft in space. Once the CSR target "pings", the system automatically hits it with a profusion of interacting LWs, completing disrupting and destroying its dynamics. Scratch one system. Probably several million such systems were destroyed throughout the West (command centers, power grids, submarines, nuclear bombers, etc.) in rapid fire order, all without any external "spatial" sign other than perhaps a wee bit of very low level "EM noise" briefly as a struck CSR fractured.

53. See M. Sachs, *Quantum Mechanics from General Relativity,* Reidel, 1986. Sachs'unified field theory, e.g., applies equally to the quarks and gluons, as it does to the nucleons, the atoms, molecules, and macroscopic objects and phenomena. See also M. Sachs, "Symmetry in Electrodynamics: From Special to General Relativity, Macro to Quantum Domains," in *Modern Nonlinear Optics,* Second Edition, 3 vols., M. W. Evans (ed.), Wiley, 2001; vol. 1, p. 677-706.

54. G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601.

55. The tragic death of Captain Hess was also very probably such a test. Hess, perfectly normal and happy, laid out his uniform in the early morning and then went for his customary morning jog. He suddenly veered aside, sat down in a path and stabbed himself 26 times, all deep stab wounds, four of which would be totally incapacitating (two right through the heart) to any normal individual experiencing normal shock, etc. To a psychoenergetically controlled individual where shock does not occur nor does emotion or pain, this is understandable. Add this to the incidents involving Captains Button and Svoboda, and one begins to see the picture.

56. W. Heitman and G. Nimtz, "On Causality Proofs of Superluminal Barrier Transversal of Frequency Band Limited Wave Packets," *Phys. Lett. A196,* 1994, p. 154-158. Aichmann and Nimtz and co-workers at the University of Cologne have transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular wave-guide that contains a "barrier" section of reduced dimensions. They have recently transmitted Mozart's 40th Symphony as frequency modulated microwaves through a barrier wave-guide at an FTL group velocity of 4.7 c, receiving audibly recognizable music from the microwave photons that survived their barrier passage. The transit time through the barrier was about 81 picoseconds and was observed to be constant for barriers with widths varying from 4.0 cm to 11.4 cm.

57. Quantum mechanics in fact requires instantaneous transmission of effects. A mishmash of learned contention would have us believe one can "transmit" energy instantly, but not intelligence. This argument has essentially stymied our scientific establishment so that it has not really funded strong investigation in this area. It is a stupid argument: Simply transmit one change, then a second one, then a third one, in serial order, where the "order" contains the intelligence. No different from Morse code. Somehow the pundits so certain that intelligence cannot be superluminally transmitted seem to have forgotten what "intelligence" is. They also seem to have assumed that one superluminal change can be **transmitted** on order, but a second, third, and fourth cannot.

58. Feynman and Wheeler tried very hard to formulate a new absorber theory of electrodynamics to eliminate many of the fierce foundations problems. Wheeler described the effort this way: *"...we may summarize the account of radiation reaction given by Wheeler and Feynman in 1945. In it every charged particle is envisaged as coupled to every other charged particle by a field that is symmetric in time: half advanced, half retarded. Interconnections run forward and backward in time in such numbers as to make an unbelievable maze. That weaving together of past and future seems to contradict every normal idea of causality. However, when the number of particles is great enough to absorb completely the signal starting out from any source, then this myriad of couplings adds up to a simple result: the familiar retarded actions of everyday experience, plus the familiar force of radiative reaction with its familiar sign."* [John Archibald Wheeler, "Law without law," in *Quantum Theory and Measurement.*, preprint, p. 39].

Unfortunately, when the foundations of any science are closely examined in their roots, one runs into the age-old bane of the philosopher: ultimately opposites are identical, at least prior to the observation (separation) action to separate them. It follows that all scientific reasoning is circular at its most primary basis.

59. See Charles W. Misner, Kip S. Thorne and John Archibald Wheeler, *Gravitation,* W. H. Freeman, San Francisco, CA., 1973, p. 71. Wheeler spoke of that inescapable fact in this manner: *"Here and elsewhere in science, as stressed not least by Henry Poincare, that view is out of date which used to say, 'Define your terms before you proceed.' All the laws and theories of physics, including the Lorentz force law, have this deep and subtle character, that they both define the concepts they use (here B and E) and make statement about these concepts. Contrariwise, the absence of some body of theory, law and principle deprives one of the means properly to define or even to use concepts. Any forward step in human knowledge is truly creative in this sense: that theory, concept, law, and method of measurement—forever inseparable—are born into the world in union."*

60. Also see T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* 2001, *ibid,* for a rework and extension of Aristotelian logic to eliminate its present fatal flaw, extending it into a 5-law logic.

61. The very notion of "propagation of energy through space" is radically extended in the extended unified EM and GR being pursued by AIAS and Sachs. Time-like EM energy currents dominate in the vacuum. E.g., see M. W. Evans et al., "Representation of the Vacuum Electromagnetic Field in Terms of Longitudinal and Time-like Potentials: Canonical Quantization," *J. New Energy,* 4(3),Winter 1999,p. 82.

62. Multiwave interactions (between transverse EM waves) produce time-reversed EM waves (phase conjugate replicas of the signal wave input). That method time-reverses EM energy in waveform. To time-reverse EM energy in mass form, simply pump that mass with time-polarized EM waves. As an alternative, it may be pumped with longitudinal EM waves (which are always accompanied by time-polarized EM waves anyway), and the pumped mass is also pumped in the time-domain because of the presence of the time-polarized EM waves in the pumping. We previously filed an invention disclosure with the U.S. Patent Office dealing with how to do that.

63. Mostly studied by Becker. See (a) R. O. Becker, "The control system governing bone growth in response to mechanical stress," *J. Ark. Med. Soc,* Vol. 62, 1966, p. 404; (b) — and David G. Murray, "The electrical control system regulating fracture healing in amphibians," *Clinical Orthopaedics and Related Research,* Vol. unk, No. 73, Nov.-Dec. 1970, p. 169-198. This latter paper is the definitive technical exposition by Becker of his electrical control system ideas and findings. Good references for the educated layperson are (c) R. O. Becker, "A technique for producing regenerative healing in humans," *Frontier Perspectives,* 1(2), Fall/Winter 1990, p. 1-2; (d) Robert O. Becker and Gary Seldon, *The Body Electric: Electromagnetism and the Foundation of Life,* William Morrow and Company, New York, 1985; and (e) Robert O. Becker, *Cross Currents: The Promise of Electromedicine, the Perils of Electropollution,* Tarcher, St. Martin's Press, New York, 1990.

64. A provisional patent application has been filed on the process, and an overview of the method can be seen as the "porthole briefing" on my website www.cheniere.org.

65. (a) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone in vivo," *Nature,* Vol. 204, Nov. 14, 1964, p. 652-654. The complete electrical control system (conventional EM theory) for the healing of bone fractures is given in (b) R. O. Becker, *Clinical Orthopedics and Related Research,* Vol. 73, 1970, p. 169. A paper "putting it all together" is (c) R. O. Becker, J. *Bioelec,* Vol. I, 1982, p. 239.

66. F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, under the heading "5.2 Covariant Quantization" and "5.3 The Photon Propagator" in Chapter 5. The longitudinal and scalar photon polarizations are not directly observable, but only in combination, where they manifest as the instantaneous Coulomb (i.e., electrostatic) potential interaction, *"which emerges as an exchange of longitudinal and scalar photons."* So the common old electrostatic scalar potential is actually comprised of longitudinal EM waves and time-polarized EM waves.

67. With the MidEast increasingly set to explode, we may be facing the unleashing of weapons of mass destruction, a vast shortage and curtailment of

available oil and energy, and a total collapse of the Western economy and therefore most of the world economy, in as little as three or four years, if the impending energy crisis worldwide is not resolved quickly. See T. E. Bearden, ADAS position paper, "The Unnecessary Energy Crisis: How to Solve It Quickly," *ibid.* As this is written, we have already had a warning shot across our bow: the vicious terrorist attacks on September 11, 2001. We are in Afghanistan and elsewhere in a war against terrorism, and anyone can see the escalation in the MidEast toward war between the U.S. and Iraq, and almost certainly a resulting explosion in the MidEast and unleashing of the hostile terrorists assets and mass destruction weapons already inside the United States. Since modern economies are based on cheap energy, and cheap energy is largely based on cheap oil, then with surging energy demands worldwide we shall see "oil wars" erupt and rage worldwide, with terrorist retaliations and attacks—using weapons of mass destruction—occurring in U.S. cities. *Unless energy from the vacuum is quickly developed and deployment begun at high speed, we risk the coming collapse of the world economy, great destruction of populaces, and a new Dark Ages.*

68. E. T. Whittaker, 1903, *ibid.*

69. Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoku Seamount and the Mystery Cloud of 9 April 1984," *Science,* Vol. 227, Feb. 8, 1985, p. 607-611; Daniel L. McKenna and Daniel A. Walker, "Mystery Cloud: Additional Observations," *Science,* Vol. 234, Oct. 24, 1986, p. 412-413.

70. Superpotential theory was initiated by E. T. Whittaker, extending earlier concepts advanced by G. Johnstone Stoney. See E. T. Whittaker, *Proc. Lond. Math. Soc,* Series 2, Vol. 1, 1904, p. 367-372. Nisbet extended the Whittaker and Debye two-potential solutions of Maxwell's equations to points within the source distribution. This is a full generalization of the vector superpotentials (for media of arbitrary properties) together with their relations to such scalar potentials as those of Debye. See A. Nisbet, *Physica,* Vol. 21, 1955, p. 799. Also, A. Ricci in 1901 introduced what might be called the magnetization potential, satisfying a certain equation, as an alternate to the Hertz vector. This was a part of early vector superpotentials. For an excellent overview discussion of superpotentials and related things, see Melba Phillips, "Classical Electrodynamics," in *Principles ofElectrodynamics and Relativity,* Vol. IV of *Encyclopedia ofPhysics,* edited by S. Flugge, Springer-Verlag, 1962. Phillips discusses, for example, that Whittaker 1903 (note: oral delivery; paper was published in 1904) was the first to prove that one can derive a general electromagnetic field from two scalar functions, which are really components of the vector superpotentials, with proper choices of the gauge functions. Whittaker's method is well known in the treatment of transverse electric and transverse magnetic modes of a cylindrical cavity or a waveguide. Nisbet has shown that this result may be extended within the source distribution (assumed

given) by a suitable choice of stream functions. The Debye potentials and the Bromwich potentials are essentially radial components of the vector potentials of which Whittaker potentials are the real parts. So in general the particular integral (the stream potentials) of the inhomogeneous Maxwell equations may be so chosen that the complementary function can be expressed in terms of only two scalars, which are components of the vector superpotentials. The Whittaker and the Debye-Bromwich potentials are special cases of two vector superpotentials.

71. David Jones, *Vancouver Sun Times,* Weekend Magazine, Dec. 17, 1977, p. 17.

72. For more specific information on the production of practical antigravity and an experiment demonstrating it, see T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, Chapter 8. Approach to Antigravity.

73. T. E. Bearden, *Energy from the Vacuum: Concepts and Principles,* Cheniere Press, Santa Barbara, CA, 2002, Chapter 10. Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy. By inadvertently using highly compressed time-energy, cold fusion researchers are actually using—on the tabletop—higher energy density than our present high-energy physics uses in its largest accelerators. The conventional scientific community does not have a single scientific program dedicated to using time as energy, even when one's physics model can legitimately be made of a single fundamental unit. When the joule is chosen as that unit, then time becomes totally a function of energy, just as is mass. The use of mass-energy gave us the nuclear age; the use of time-energy will give us a far greater age, if we can ever get the scientific community off its duff and moving.

74. Further, the second law of thermodynamics has long been known not to apply to a single charge, atom, or molecule or to a small group. For such microscopic phenomena, there will occur reversed reactions a certain small fraction of the time. But also see G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), 29 July 2002, 050601. These researchers have experimentally proven that the second law can be and is violated for much larger scales—up to micro (colloidal) size—and for much longer times (up to two seconds). Hence it is now proven that such "reversed zones" where reactions run backwards for quite a bit, can and do in fact occur on the colloidal scale, sufficient for the formation of large clusters of like charges. Evans and his colleagues have pointed out that these results dramatically affect chemistry.

75. Shoulders has been creating and using such charge clusters for years. See Kenneth R. Shoulders, "Energy Conversion Using High Charge Density,"

U.S. Patent # 5,018,180, May 21, 1991; — "Energy Conversion using High Charge Density," U.S. Patent No. 5,123,039, 1992; — "Method and Apparatus for Production and Manipulation of High Density Charge," U.S. Patent #5,054,046, 1991; — U.S. Patent #5,054,047, 1991; — "Circuits Responsive To and Controlling Charged Particles," U.S. Patent 5,148,461, 1992; — and Steve Shoulders, "Observations on the Role of Charge Clusters in Nuclear Cluster Reactions,"J. *New Energy,* 1(3), Fall 1996, p. 111-121.

76. D. J. Evans and D. J. Searles, *Phys. Rev. E,* Vol. 50, 1994, p. 1645.

77. See Donald Kennedy, "To Publish or Not to Publish," *Science,* Vol. 295, 8 Mar. 2002, p. 1793.

78. R. P. Taleyarkhan, C. D. West, J. S. Cho, R. T. Lahey Jr., R. I. Nigmatulin, and R. C. Block, "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science,* Vol. 295, 8 Mar. 2002, p. 1868-1873. Quoting from the abstract: "In cavitation experiments with deuterated acetone, tritium decay activity above background levels was detected. In addition, evidence for neutron emission near 2.5 million electron volts was also observed, as would be expected for deuterium-deuterium fusion. Control experiments with normal acetone did not result in tritium activity or neutron emissions."

79. Yuri T. Didenko and Kenneth S. Suslick, "The energy efficiency of formation of photons, radicals and ions during single-bubble cavitation," *Nature,* Vol. 418, July 25, 2002, p. 394-397.

80. E.g., see John Archibald Wheeler and Richard Phillips Feynman, *Phys. Rev.* 59, 1941, p. 683; —"Interaction with the Absorber as the Mechanism of Radiation," *Rev. Mod. Phys.,* 17(2, 3), Apr.-Jul. 1945, p. 157-181.

81. Albert Einstein, "Foreword," in Max Jammer, *Concepts of Space: The History of Theories of Space in Physics,* Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

82. Robert A. Fisher, [Ed.], *Optical Phase Conjugation,* Academic Press, NY, 1983, p. xv.

83. Basically, the choice between distant heating in the interference zone, or distant cooling, is as simple as how one biases the ground potential used for the electrical grounds of the interferometer transmitters. Biasing well above the ambient potential in the distant IZ produces divergent EM energy (heat) forming and radiating in the distant zone. Biasing well below the ambient IZ potential produces convergent EM energy (cold) forming and radiating in the distant zone.

84. Mass is mostly empty space filled with EM potentials, fields, and waves. Since these potentials, fields, and waves are simply bundles of longitudinal EM waves and their impressed dynamics, it follows that mass itself

is mostly a "giant superhighway" for the transmission of LWs. This is actually a sort of "subspace" transmission "right through" matter, so to speak.

85. As demonstrated by such dual serial usage by a giant interferometer in the anomalous and explosive "cloud" phenomena near the Kurils in April 1984. For excellent descriptive articles and a hard look at the phenomena, see Walker et al., *Science, ibid.*

86. See "Soviet device uses radio waves as tranquilizer," AP release, *Huntsville [Alabama] Times,* May 19, 1983. See also W. Ross Adey, "Biological effects of low energy electromagnetic fields," *Proc. NATO Adv. Study Inst. on Advances in Biological Effects and Dosimetry of Low Energy Electromagnetic Fields,* Plenum Press, New York, 1983, p. 359-391. Adey apparently was aware only of ordinary U(1) electrodynamics. He describes the **LIDA** device as follows: *"Soviet investigators have also developed a therapeutic device utilizing low frequency square wave modulation of a radiofrequency field. This instrument know as the LIDA, was developed by L. Y. Rabichev and his colleagues in Soviet Armenia, and is designed for the 'treatment of neuropsychic and somatic disorders, such as neuroses, psychoses, insomnia, hypertension, stammering, bronchial asthma, and asthenic and reactic disturbances.' It is covered by U.S. Patent #3,773,049. In addition to the pulsed RF field, the device also delivered pulsed light, pulsed sound, and pulsed heat. Each stimulus train can be independently adjusted in intensity and frequency. The radiofrequency field has a nominal carrier frequency of 40 MHz and a maximum output of approximately 40 watts. The RF field is applied to the patient on the sides of the neck through two disc electrodes approximately 10 cm. in diameter. The electrodes are located at a distance of 2-4 cm from the skin. Optimal repetition frequencies are said to lie in the range from 40 to 80 pulses per minute. Pulse duration is typically 0.2 sec. In a 1-year trial period, the instrument was used on 740 patients, including adults and children. Positive therapeutic effects were claimed in more than 80 percent of the cases treated."* We accent that this is one of the simplest Russian psychoenergetics devices, obsoleted very early on, and thus released very early for public use.

87. Since U.S. researchers are unaware of the EM wave polarization transduction mechanisms, and in general have no instruments for detecting longitudinal EM waves, the specific transduction effects on humans would not be measured by U.S. laboratories. Hence the U.S. scientists could only think of it in conventional electrodynamic terms, and remained totally unknowledgeable about the specific transduction mechanisms in the human body producing time-polarized EM waves and effects from the device.

88. Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'," New *York Times,* Jan. 15, 1960, p. 1.

89. For a simplified introduction to psychoenergetics and the mind-matter coupling mechanism, see T.E. Bearden, "Mind Control and EM Wave Polarization Transductions, Part I", *Explore,* 9(2), 1999, p. 59; Part II, *Explore,* 9(3), 1999, p. 61; Part III, *Explore,* 9(4,5), 1999, p. 100-108.

90. One of the serious goals of the die-hard KGB psychoenergetics scientists appears to be to achieve the capability of directly engineering the entire human collective unconscious mind. The plan envisioned is to convert the human species into a sort of "giant ant species" where each one has perfect, machine-like niche and function (except, of course, for the elite Communist masters). In short, they conceived back there in the 1960s that such a "perfect communist state" for all humanity was in fact possible to achieve technically, with sufficient development of psychoenergetic technology. They are apparently well along the way toward achieving that goal. [The U.S. has not even conceived such a scientific possibility, and modern cognitive science in the West seems bent on the rather astounding proposition that mind itself does not even exist!]. We point out that no type of physical defense is even possible against such a capability engineered into the collective human unconscious, once it is achieved. The only "counter" is to use the same technology (it has to first be developed) to "re-engineer" what will have been done. But our fellows have not even conceived the possibility of the threat, much less the cause or the counter. Hence they continue to totally misunderstand what happened to Captains Button, Svoboda, and Hess in successful mind-control tests achieved perfectly at great distance.

91. To see the nature of the time-density wave (which is basically a longitudinal EM wave along the time axis), due to the perfect correlation between time-energy flow currents and variations with respect to longitudinal EM waves, see T.E. Bearden, "Giant Negentropy from the Common Dipole," 2000, *ibid.* Not stressed in the paper is the fact that the same dipolar negentropy with correlated EM time-energy currents and EM spatial energy currents, exists in the "individual" photon. Further, they can occur in coupled photon-antiphoton pairs, yielding the graviton. For time-polarized EM photons, it appears that two photon-antiphoton pairs are coupled, yielding the "supergraviton." Elsewhere we have briefly and addressed these novel formations as hypotheses.

92. Induction of specific diseases and health changes in the U.S. Embassy in Moscow over four decades, included causing the deaths of three U.S. Ambassadors. It was accomplished by implementing desired disease-pattern "engines" infolded inside field-free potentials. The Johns Hopkins study of the radiation patterns revealed a most significant finding: All the affected persons were in regions of the Embassy where the Russian microwave radiation *fields* were absent. In short, where the field-free—and thus very stable—potentials existed. Since these scientists (and most others in classical EM) consider that EM effects are caused only by fields, they concluded (totally erroneously) that

these health changes could not have been caused by the microwave radiation. To the contrary, they had just clearly proven that the radiation was indeed the cause. They showed a 100% correlation to the *presence of field-free potentials,* and a 100% anticorrelation to where the fields were present. This was for induced phenomena over several years. Had the radiation not been the culprit, then some health changes would have occurred where the fields were present, and some would have occurred where they were absent. That is not what happened at all, and this inexplicable logical non sequitur against common statistical inference was the accepted decision. The truly revealing aspect was that *apparently the entire U.S. Intelligence community had no competence or capability in higher group symmetry electrodynamics.*

93. E.g., see Lewis H. Ryder, *Quantum Field Theory,* Cambridge University Press, 2nd Edition, 1996, p. 147+ which covers the four photon polarizations. See also F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, under the heading "5.2 Covariant Quantization" and "5.3 The Photon Propagator" in Chapter 5.

94. We elaborate on this somewhat, in "Mind Control and EM Wave Polarization Transductions", *ibid.*

95. Again, we point out the unified field approach of Mendel Sachs as probably being the best Western model in which to grapple with these phenomena.

96. David J. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

97. T. E. Bearden, *Gravitobiology: A New Biophysics,* Cheniere Press, Santa Barbara, Ca. www.cheniere.org, 1991.

98. W. A. Rodrigues Jr. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 < v < oo$ in nature," *Found. Phys.,* 27(3), 1997, p. 435-508. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site. It includes corrections to the published version.

99. W. A. Rodrigues Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras,* Vol. 7(S), 1997, p. 457-466.

100. E.g., see A. Enders and G. Nimtz, *Phys. Rev. Lett.,* Vol. 48, 1993, p. 632. Enders, Aichmann, Nimtz and co-workers at the University of Cologne have transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular wave-guide that contains a "barrier" section of reduced dimensions. They transmuted Mozart's 40th Symphony as frequency modulated

microwaves through a barrier wave-guide at an FTL group velocity of 4.7c, receiving audibly recognizable music from the microwave photons that survived their barrier passage.

101. J. D. Jackson, *Classical Electrodynamics,* 2nd Edition, Wiley, New York, 1975, p. 222.

102. See H. A. Lorentz, *Vorlesungen Uber Theoretische Physik an der Universitat Leiden, Vol. V, Die Maxwellsche Theorie (1900-1902),* Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. Figure 25 on p. 185 shows the Lorentz concept of integrating the energy flow vector around a closed cylindrical surface surrounding a volumetric element. This is the procedure which arbitrarily selects only a small component of the total energy flow in space outside a circuit and associated with it—specifically, the small Poynting component diverged by the surface charges into the conductors to power the electrons—and then treats that tiny component as the "entire" vast energy flow (which may be a trillion or more times greater in magnitude). Thereby Lorentz arbitrarily discarded all the huge *Heaviside* energy transport component that does not strike the circuit at all, is not diverged into the conductors, and is just wasted. See Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94. Quoting: *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire... . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire."*

103. When it is associated with positive energy fields and potentials from positive mass, positive energy source charges, we have nominated that long-unaccounted Heaviside energy flow component as the source of the excess gravity holding the arms of the spiral galaxies together (a solution to the dark matter problem). When it is associated with negative energy fields and potentials from negative-energy Dirac Sea holes in 4-space, we have nominated that unaccounted energy flow component as the source of the excess antigravity responsible for accelerating the expansion of the universe (a solution to the dark energy problem).

104. About a year from now, Bedini and I will also release a very novel use of this Heaviside energy flow component in inverted circuits—circuits that

operate almost completely in reverse from normal textbook circuits. This is a totally new kind of circuit that we hope will revolutionize power systems and electronic control systems. Until our intellectual property rights are secured, we are not releasing any details at all.

105. Terence W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Ann. Fond. Louis de Broglie,* 16(1), 1991, p. 23-41. Tesla was able to shuttle energy around in his circuits at will. Barrett went on to improve the process and patent two versions. See T. W. Barrett, "Active Signalling Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996. This is a signaling system in time-frequency space for detecting targets in the presence of clutter and for penetrating media. See also T. W. Barrett, "Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic and RF Phase Conjugation," U.S. Patent No. 5,493,691. Feb. 20, 1996.

106. Bryan G. Wallace, "The Unified Quantum Electrodynamic Ether," *Found. Phys.,* 3(3), Sept. 1973, p. 381-388.

107. William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," Am. J. *Phys.,* 53(7), July 1985, p. 671-679. Most waves caused by a single frequency source have a galloping motion. The only exceptions are the very special cases of absolutely pure moving waves or absolutely pure standing waves. The superposition and resulting interference of two oppositely moving waves generally yield a wave having nonuniform motion; in short, the wave rushes forward rapidly, then slows appreciably, etc. A relativistic example is given in which the source frame observes a very nonuniformly galloping wave whose velocity varies between 0.01 c and 100 c, even though its average velocity is c. Relativistic properties of interfering electromagnetic plane waves are discussed.

108. S. L. Miller, "The Formation of Organic Compounds on the Primitive Earth," *Ann. NYAcad. Sci.,* Vol. 69, 1957, p. 260; — *Bioch. Biophys. Acta.,* Vol. 23, 1957, p. 480-489; — *Science,* Vol. 117, May 15, 1953, p. 528-529; — and H. C. Urey, "Organic Compound Synthesis on the Primitive Earth," *Science,* Vol. 130, 1959, p. 245-251.

109. R. Harada and S. Fox, "Thermal Synthesis of Natural Amino-Acids from a Postulated Primitive Terrestrial Atmosphere," *Nature,* 201(4917), 1964, p. 335-336; S.W. Fox, "How Did Life Begin?", *Science,* 132(3421), July 22, 1960, p. 200-208; — and S. Yuyama, "Abiotic Production of Primitive Protein and Formed Microparticles," *N.Y. Acad. Sci. Ann.,* Vol. 108, June 1963, p. 487-494; — and K. Harada, "Thermal Copolymerization of Amino Acids to a Product Resembling Protein," *Science,* 128(333), 1958, p. 1214; — and M. Middlebrook, "Anhydrocopolymerization of Amino Acids Under the Influence of Hypothetically Primitive Terrestrial Conditions," *Federation Proc.* Vol. 13, Mar. 1954, p. 211.

110. H. C. Urey, "On the Early Chemical History of the Earth and the Origin of Life," *Proc. Nat. Acad. ScL,* 38(4), 1952, p. 351-363.

111. I.e., one may speak of the "potential" of a specific engine of a former life form, and then realize that the potential for that form reduces in magnitude inversely as the radial distance from the Earth. However, it does not reduce to zero until an infinite distance is reached.

112. As an example, a great EM energy flow—produced by the cross product ExH of the earth's magnetic field and the earth-to-electrosphere electric field—flows from West to East around the earth, nearly parallel to the equator. Infolded in that great EM energy flow are all the structures of all the genetics and life forms on earth, or that have ever been on earth and conditioned the natural scalar potential of the earth. The vials in the laboratory experiments of Miller, Fox, and Urey were all continually bathed in that great energy flow, which within it contained the entire life structure and dynamics of the planet. So no "inert" experiments were ever performed, contrary to the interpretations of the scientific community.

113. Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions*, [in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76.] Kaznacheyev, e.g., appears to have headed the Soviet team actually producing the basis for the disease-inducing technology utilized for four decades in the microwave radiation of the U.S. Embassy in Moscow.

114. Specifically, the down payment was $900 million in gold. The lease is probably about $1 billion per year. Thus the Yakuza acquired very powerful strategic weapons, and became a secret power to be reckoned with internationally and strategically. Since 1990, weather engineering over the U.S. and occasional shootdown of aircraft (such as the TWA-800) have been by rogue Japanese fingers on the scalar interferometer weapon triggers, on site in Russia.

115. The Yakuza has also transferred some of the longitudinal EM wave interferometry technology to their Japanese facilities. One such system has apparently been deployed in the power grid, and has been used to kill some government officials who opposed the Yakuza. This information was obtained by some brave Japanese risking their lives to get the information to the present author.

116. Many important KGB weapon tests happened in April of various years; the KGB weapons folks were simply preparing for their May Day reports, and each wished to have a good one.

117. Sen. Sam Nunn's investigation teams verified that the Aum Shinrikyo were actively seeking the "Tesla Weapons", with the Aum Shinrikyo even visiting the Tesla Museum in Beograd for that purpose. We have 100% certain confirmation of the Yakuza and Aum Shinrikyo leasing of the scalar interferometers from the KGB and Russian Mafia (working for the KGB) at the end of 1989 or first few days of 1990. Sen. Nunn's report confirms that the Aum Shinrikyo set up a University in Moscow. Our comment is that this is where the Russians actually taught the Japanese crews the science and technical functions so that they could successfully man and use the scalar (longitudinal EM wave) interferometry weapons.

118. Recall that in the 1950s and early 1960s, Kaznacheyev et al., *ibid.,* had already clearly proven that nearly any form of cellular change, damage, or disease could be induced in cells at a distance by purely electrodynamic means.

119. In the kill of the Arrow DC-8 at Gander, Newfoundland on 12 Dec. 1985, the struck aircraft did charge up with time-density charging, and many pieces of the wreckage did indeed emit longitudinal EM waves thereafter for some time. Consequently, the Canadian recovery crews working the site and handling the materials, were exposed to longitudinal EM wave radiation for varying degrees of exposure. It follows that many of those on-site handlers should have later come down with physical ailments, with symptoms such as flu-like features, weakness, heart palpitations, headaches, dizziness, disturbance of equilibrium, and with various infections because of weakened immune systems. They also should have had weakened cellular regenerative (healing) systems, so their physical problems would endure and largely remain. About five dozen of the recovery personnel did come down with those exact symptoms, and many remained disabled or debilitated for some years.

120. We have also explained (a) why the French forces in the Gulf War did not get GWS, while American, British, and Canadian forces did, (b) why the native population did not exhibit GWS, and (c) the remaining anomalous phenomena associated, such as the semen "infecting" the wife, the children catching similar or related conditions from sustained close proximity to their father's biofields (transmission of engines), etc.

121. Mario Bunge, Foundations of Physics, Springer-Verlag, New York, 1967, p. 176.

122. T. E. Bearden, "Giant Negentropy from the Common Dipole," J. *New Energy,* 5(1), Summer 2000, p. 11-23 gives the solution to the long-vexing source charge problem. The paper is also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org

123. Two examples of inane articles and actions in this respect, revealing no knowledge that every charge in the universe already exhibits COP = oo, are (a)

Graham P. Collins, "There's No Stopping Them: Perpetual Motion is alive and well at the U.S. Patent Office," and (b) Resolution by the Executive Board at the June 2002 meeting of the American Physical Society (APS) as follows: *"The Executive Board of the American Physical Society is concerned that in this period of unprecedented scientific advance, misguided or fraudulent claims of perpetual motion machines and other sources of unlimited free energy are proliferating. Such devices directly violate the most fundamental laws of nature, laws that have guided the scientific progress that is transforming our world."* So the learned APS has just declared the source charge as an unacceptable "free energy" machine and perpetual motion machine, since the charge simply sits there and continuously pours out real observable EM energy in all directions at the speed of light, without any observable EM energy input. Thus the APS has implicitly declared all EM fields and potentials and their energy—which includes every joule of EM energy in the universe, in matter or in space—to be unacceptable byproducts of perpetual motion nonsense. We heartily call upon the APS in the name of science to enforce their resolution in the orthodox academic community, who are the greatest perpetual motion advocates in history. To be consistent with their own resolution, they must root out and destroy—from all texts and universities and from the scientific literature—any mention whatsoever of charge, field, potential, and EM energy. Either they must do that, or stand hoisted on their own petard as a dogmatic group still very much unaware of why the Nobel Prize was awarded to Lee and Yang in 1957 and what the proof—by Wu et al.—of the asymmetry of opposite charges means with respect to the source charge considered in its modern quantum field theory view as a dipolarity of a bare infinite central charge surrounded by clustering virtual charges (also infinite charge) of opposite sign in the active vacuum. Every electrical engineering department and professor and textbook still uses a model and teaches a subject assuming that every charge in the universe freely creates energy from nothing at all. Further, the APS in every annual meeting also tolerates, encourages, and publishes multiple papers and presentations with the same such unacceptable perpetual motion assumptions in them. Such is the lofty understanding of our present organized scientific community, which seems much more politically bent than scientifically bent. The simple Bohren experiment and the entire area of negative resonance absorption of the medium has long been established in nonlinear optics, and such experiments widely yield a nominal COP = 18. In other words, the experiments nominally output 18 times as much energy as the operator inputs, and the excess input energy is furnished absolutely for free from the active vacuum's exchange. In this case the Executive Board of the APS is not even aware of what its own physics discipline has long proven and accepted in particle physics, for nearly a half century. Neither are the self-appointed defenders of the faith, who also are totally unaware of what physics has already proven with a Nobel Prize awarded nearly a half century ago.

124. C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev.,* Vol. 105, 1957, p. 1413. We also quote Nobelist T. D. Lee as follows: *"Since nonobservables imply symmetry, any discovery of asymmetry must imply some observable. The experiment of Wu, Ambler, Hayward, Hoppes and Hudson... established the asymmetry between the positive and negative signs of electricity."* [T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 11.].

125. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, New York, 1998, reprinted with corrections 1999.

126. See "Secret Speech: Did Brezhnev Come Clean?" *National Review,* 29(8), Mar. 4, 1977, p. 248, 250

127. For the energy description of this device and some results, see Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc. 26th Intersoc. Energy Conversion Engineering Conf. (IECEC '91),* Boston, Massachusetts, p. 370-375.

128. Also see my book, *Energy from the Vacuum,* 2002, *ibid,* which includes photographs of the self-powering Sweet device, courtesy of Walter Rosenthal.

129. Roy Brewer was very active in Hollywood Unions against Communists when they were busily penetrating the Hollywood complex. He was a staunch anti-communist, fighting them out of control of the unions, way back when. Ronald Reagan was a personal friend of his of long standing, and Roy had open access to the White House when Reagan was President.

Roy briefed Reagan on the Russian scalar weapons, and Reagan believed him. Three times, Reagan ordered his Secretary of Defense to take action in that area, but neither of the three did so because the U.S. Intelligence Community did not believe it and adamantly opposed any such action. We took quite a beating in those days, for quite a long time (some years), before certain key parts of the Government finally believed us, checked it out experimentally, and found that it worked.

Roy also introduced me to the Duke de Grantsmesnil, one of the last of the great old British intelligence men, now unfortunately deceased. In the hectic 1997 days when we were trying desperately to warn a little friendly nation (and our own) against a coming great strategic strike on the U.S. and other allies, the Duke played a major role in introducing us directly to higher political levels in that little nation - which was the only nation on earth that could counter the pending strike. The first planned strike scheduled in February 1997 was successfully countered, only to be immediately rescheduled. In the second incident, the attack was scheduled for May 1, 1997. We finally realized why the hostile nation could reschedule and go ahead, and needed then to warn the little nation of the reason in some detail. We had tried everything to get certain information through, and all means had failed. A former ambassador of that nation to Great Britain was also assassinated during that struggle. Roy Brewer was one of the key fellows assisting in that attempt, as were other personal colleagues of mine who also were witnesses and direct participants in those desperate events. When all else failed, I finally just faxed the Prime Minister directly, in the clear, some five days before the attack, urging the most urgent and immediate action possible by the Prime Minister. A complete change in command and control of the weapon was required, and that change had to be made and then the hostile nation had to be notified that the change was in place (and probably shown that it was). The Prime Minister accepted the message and the necessary action was carried out immediately. So two days or so before the actual attack was scheduled, that great hostile armada began standing down. Oh, the hostile nation and its leaders could have easily destroyed the entire United States in two hours or so, without fail! But then their own country would have promptly disappeared from the face of the earth. So they stood down. And our own Intel community never even picked it up (they did get my messages alerting the U.S. Government, but thought it was a joke or something). They, of course, also intercepted my open fax to the Prime Minister, and so were all over me for that one for a while thereafter.

Interestingly, at the very time that great armada was just beginning to stand down, then Secretary of Defense William Cohen made the following statement at a conference in Georgia: "Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves...so there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts, and that's why this is so important." Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

Ironically, the Secretary was referring to the earlier scalar interferometer weapons, and now the quantum potential weapons etc. He was completely unaware that a great strategic strike had actually been scheduled and was just being aborted as he was speaking those words.

Roy Brewer is one of those fellows who has been a great American, has served his country admirably during his lifetime, and contributed greatly to preventing a planned take-over by the Communists of the labor unions of America. In addition, he helped set up the contacts with the friendly **liltle foregn nation, so** that on four occasions planned attacks to destroy the United States have been successfully aborted or handled.

As you know, one of the great failures in the Communist plan to have America go communist was their failure with the labor unions. And we owe that failure to the steadfast efforts of some staunch folks like Roy Brewer. He helped save our nation when many of us were still wearing knee pants.

130. Wheeler's principle, as stated in W. Misner, K. S. **Thorne, and J.** A. Wheeler, *Gravitation,* W.H. Freeman and Co., San Francisco, 1973, p. 5.

131. With the new improvements we uncovered, treatment time can be reduced to 5 minutes per patient. Three such treatments one week apart would be what are required. Hence one machine station could treat a patient every five minutes, around the clock. With sufficient portable treatment units, mass treatment for mass casualties would become a practical reality.

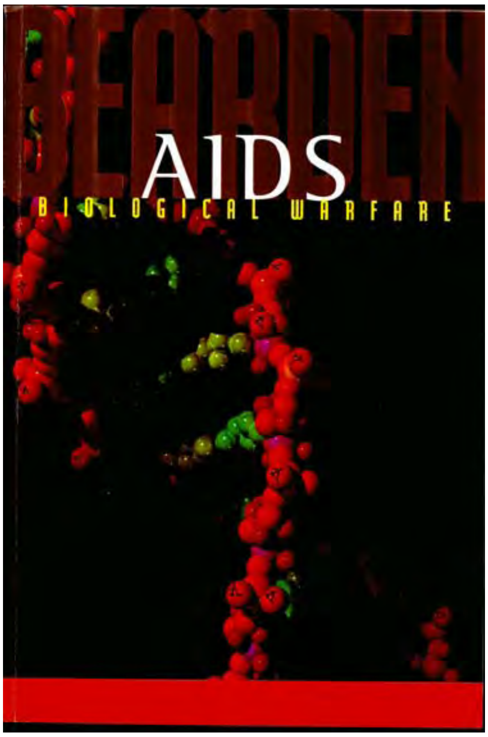

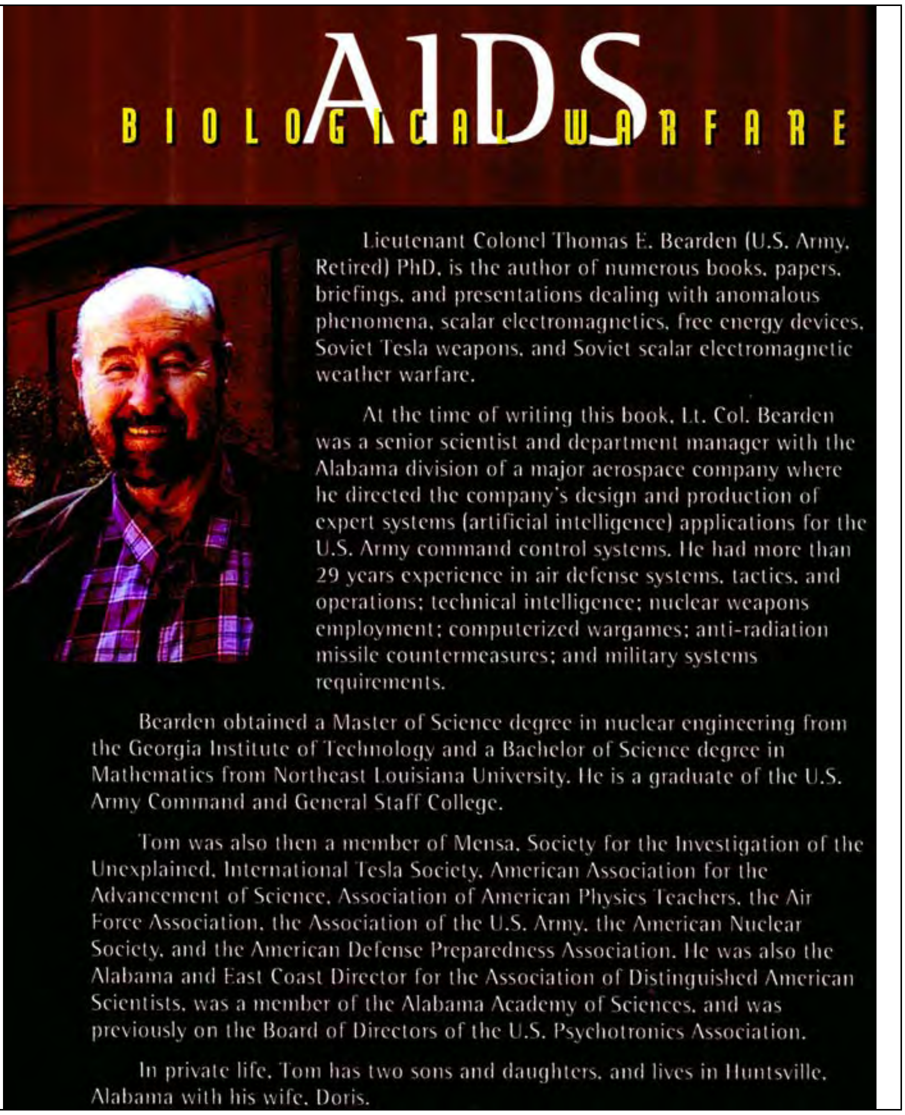
AIDS
BIOLOGICAL WARFARE

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) PhD, is the author of numerous books, papers, briefings, and presentations dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare.

At the time of writing this book, Lt. Col. Bearden was a senior scientist and department manager with the Alabama division of a major aerospace company where he directed the company's design and production of expert systems (artificial intelligence) applications for the U.S. Army command control systems. He had more than 29 years experience in air defense systems, tactics, and operations; technical intelligence; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements.

Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology and a Bachelor of Science degree in Mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College.

Tom was also then a member of Mensa, Society for the Investigation of the Unexplained, International Tesla Society, American Association for the Advancement of Science, Association of American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, and the American Defense Preparedness Association. He was also the Alabama and East Coast Director for the Association of Distinguished American Scientists, was a member of the Alabama Academy of Sciences, and was previously on the Board of Directors of the U.S. Psychotronics Association.

In private life, Tom has two sons and daughters, and lives in Huntsville, Alabama with his wife, Doris.

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. This book should especially be interesting to any person in medical community (being as a doctor or as patient) as it provides factual evidence of successful (but supressed) treatments through electromagnetic means. But before any devices can be understood and developed a new and extended kind of physics must be introduced into our educational system. Only then can this proper foundation be learnt and used by applicative sciences such as medicine. So far such knowledge has been contained in highly classified military circles and hopefully now, after many decades of suppression, it will finally be used for the good of mankind.

Dedication

This book is sincerely dedicated to the millions of persons who have died needlessly of cancer, leukemia, and AIDS in the last two and half decades. They were men, women, and children. Some were weak, some were strong. Some died in agony, some died silently. All of them were our fellow human beings. We remember them. They were important. Their anguish cries out to be heard. God willing, the new electromagnetic medicine so long suppressed will flower, and save the millions otherwise yet to die.

Special Acknowledgements

Mr. Christopher Bird's kind permission to print, his summary of the Priore Affair is gratefully acknowledged, as is the contribution of his own extensive Priore research material to this effort. The author owes a great debt indeed to Chris Bird for his magnificent assistance.

Mr. Robert Whitney graciously contributed rare photographs of Antoine Priore's device and also translations of background papers by French scientific journals. His previous valiant though fruitless efforts to bring the Priore machine into the mainline of Western medical treatment are specifically recognized and commended.

Vital contributions by Drs. Robert B. and Theodore A. Strecker are particularly acknowledged. Their March 1986 alert to the federal government, that AIDS was a deliberate biological attack spawned by Soviet agents working in Western research laboratories, is specially pointed out.

Heroic efforts by Dr. John Seale and other scientists to alert the medical research community to the true source of AIDS is specifically recognized and gratefully acknowledged.

Continuing efforts by Dr. William C. Douglass to "tell it like it is" on the AIDS crisis is specially pointed out. His permission to extract and present an excerpt from his research, pointing out tested healing of viral infections by UV irradiation of the blood, is deeply appreciated.

The unyielding and continuing efforts of Dr. Eva Snead, to alert the national authorities to the fact that there is more to AIDS than meets the eye, is also particularly recognized and appreciated.

Decades of effort by John Crane to bring Royal R. Rife's work out of suppression are specifically recognized and appreciated.

Important scalar EM experiments by John Bedini and Frank Golden are fully recognized and deeply appreciated. Their experimental efforts over a period of years have profoundly influenced my own thinking and understanding.

As always, my great admiration and deep thanks go to my friend Hal Crawford, whose keen artistic vision and skilled hand have contributed key illustrations to my work for years.

Other important illustrations and artwork by Messrs. James Thomas Neumann and Lee Giles are gratefully acknowledged.

Photographs furnished by John Moray showing his father's pioneering free energy work are appreciated. T.H. Moray's genius and achievements will yet be recognized.

Support from the Association of Distinguished American Scientists and Mr. Richard J. Reynolds III is deeply appreciated. Without that support, this effort could not have been accomplished.

Important contributions from a host of other colleagues and persons are deeply appreciated. These persons include Frank Golden, John Bedini, Ken Moore, Tom McLaughlin, Joe Parrot, Randy Davidson, George Fencl, Andrew Michrowski, Stefan Possony, Tony Gideon, Electra Briggs, and many others too numerous to mention.

Finally, nothing that I have attempted could have been accomplished at all without the unflinching and loyal support of my beloved wife, Doris.

To all of you I owe a great deal. If this bizarre struggle to save humanity is ever won, it will largely be due to the significant contributions you have made.

Thomas E. Bearden
Jan. 17, 1988

TABLE OF CONTENTS

DEDICATION iii

SPECIAL ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS vi

LIST OF FIGURES xi

LIST OF TABLES xv

FOREWORD 1
- What This Book Is About 1
- Desperate Diseases, Desperate Remedies 3
- The "Dirty Dozen" 5
- An Analogy 6

THE VACCINE CONNECTION 8
- The "Monkey Bite" Theory 9
- Retroviruses and Recombinants 10
- Cancer Research Produced Ideal Biological Warfare Agents 11
- U. S. "Cancer War" Labs Were Infiltrated 12
- The Deed Was Easily Done 14
- It's Actually Documented 15
- Soviet Biological Warfare Background 17
- We Need An Immediate, Full-Scale Investigation 20

THE SOVIETS KNEW THE U. S. WOULDN'T REACT 24
- Administering the Coup-de-Grace 24
- Murky "Behind-the-Scenes" Intentions 26
- The USSR Has Repeatedly Verified Our Nonreaction 30
- The Asian Tiger Mosquito: A New and Efficient Vector 62

TABLE OF CONTENTS

WORLD WAR III IS RAGING AND WE'VE NEARLY LOST IT 66
- More On Aids Doubling Time 66
- World War III is Already Raging 68
- Communism is Savage Enough to Do Diabolical Things 68
- Western Medicine is Inadequate 69
- The Soviets Will Not Be Decimated 71

EXTRAORDINARY PHYSICS 74
- Maxwell's Lost Unified Field Theory 74
- Heaviside's Mutilation of Maxwell's Theory 83
- Building Upon Whittaker's Fundamental Work 86
- Symmetry and Parity 90
- Charge-Parity-Time and Negative Energy 97
- Free Energy and Antigravity 100
- Phase Conjugation (Time Reversal) 104
- The Phase Conjugate Mirror 114
- Amplifying the Phase Conjugate Effect 117
- The Phase Conjugate Mirror as a Vacuum Triode 122
- Newton's Third Law and the Full Law of Entropy 125
- Scalar Electromagnetics 134
- Scalar EM View of the Vacuum 139
- Charge, Potential, and Curvature of the Vacuum/Spacetime 151
- Twin Flows of Time 152
- Engineering Local General Relativity 154
- Engineering Antigravity 157
- Curved Spacetime and the Disintegrated Vacuum 159
- Separation of Vacuum and Observable States 160
- Partial Potentials, Curvature, and Warping 161
- Force and Mass are Inextricably Intertwined 164
- The Concepts of Zipped and Unzipped Forces 166
- Force, Force Field, and EM Waves in Vacuum 168
- Zero-Summed Vector Systems: Shortcoming in Vector Analysis 172
- Maxwell's Theory was Altered and Curtailed 174
- Scalar Waves and Polarization of the Vacuum 177
- Vector Zero Systems and the Kaluza-Klein Approach 180

TABLE OF CONTENTS

Electrogravitation Amplification Factor 186
Comparison of EM Concepts 187
The Aharonov-Bohm Effect 188
Curl-Free Magnetic Vector Potential (A-Field) 190
Fractional Charges, Magnetic Monopoles, and Magnetic Currents 192
Some Ehrenhaft Experiments 194
What is a Magnetic Monopole? 199
Supercharges and Fractional Charges 202

EXTRAORDINARY BIOLOGY 204
Kervran's Proof of Biological Transmutation 204
Biological Transmutation Has a History 207
Surplus-of-Energy Mechanisms Proposed By the U. S. Army 208
Alchemy and Unusual Critters 210
The Cell Also Lives and Functions in the Virtual State 213
The Kaznacheyev Experiments 214
Structuring and Charging a Biopotential 217
The Cell's Electromagnetic Breathing 218
The Summed Virtual Structures of Kaznacheyev "Death Photons" Physically Kindle the Disease 219
Electromagnetic Infection Results in Physical Disease 221
A Possible Cure For AIDS 221
The Mirror Cytopathogenic Effect and Factors Influencing It 223
Some Biological Warfare Implications 225
What Kaznacheyev Hid: The Role of Phase Conjugation 226
Popp's Master Cellular Communication System 228
Scalar EM Comment on Popp's Communication System 230
A New View of the Nature of Mind and Thought 231
Kindling, Life, Mind, and Negentropy 236
The Priore Machine and Phase Conjugation 247
Phase Conjugates of "Death Photons" are "Healing Photons" 249
The Effect is Universally Applicable 251
In the Mid-70's It Ended 252

TABLE OF CONTENTS

Work of Australians Reid and Barsamian, and the American G. Wilbanks 254
Concepts of Dr. Reid and Associates 259
Other Important Ideas 261
Wider Implications of Structured Field Information 262
Dr. Barsamian's Important Theoretical Insights 264
Royal Raymond Rife 266
Rife's Universal Microscope 267
Rife Proved That Everything Is Alive 272
Rife's Microscope Used Evanescent Waves 275
Rife Revealed a Far More Fundamental, Living Biology 277

DEVELOPING THE ELECTROMAGNETIC CURE FOR AIDS 280
Reversing the Kaznacheyev Effect 280
The Proof: Priore's Work 282
What Is Needed 284
Encouraging Preliminary Work 284
Conclusion 288

SELECTED BIBLIOGRAPHY 290
References Related to Chapters 1 and 2 290
References Related to Chapter 3 297
References Related to Phase Conjugation (Time Reversal) 303
References Related to William J. Hooper's Work 309
References Related to Ehrenhaft's Work 312
References Related to Louis Kervran's Work 314
References Related to Popp's Master Cellular Communication System 316
References Related to Kaznacheyev's Experiments 319
References Related to The New View of The Nature of Mind and Thought 324
References Related to Antoine Priore's Work 333
References Related to the Australian Work 339
References Related to Royal R. Rife's Work 341

TABLE OF CONTENTS

APPENDIX I: The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science (Christopher Bird) (Used by Permission) 346

APPENDIX II: Patents 376

APPENDIX III: A Conditional Criterion for Identity, Leading to a Fourth Law of Logic 428

APPENDIX IV: List of Selected Incidents 444

APPENDDIX V: Dead Man Fuzing: The Real Meaning of the Reykjavik Summit 484

LIST OF FIGURES

No. Title Page

1. Conventional science cannot do the job.....4
2. The "Monkey-Bite" theory.....9
3. The Soviets have delivered an intended knockout.....15
4. The U.S. Embassy in Moscow.....32
5. Decay of a radioactive nucleus.....33
6. EGP fission of a radioactive nucleus.....33
7. Bottled up energy in a standing scalar EM wave is a potential.....34
8. EGP shortout.....34
9. EGP initiation of the Chernobyl event.....35
10. The U-2 high-altitude reconnaissance aircraft.....36
11. The U. S. S. Thresher disintegrates.....37
12. Soviet helicopters have used a "death ray" in Afghanistan.....40
13. Mushroom cloud from the sea off Japan, near a Russian test area.....40
14. The Tesla Shield, a giant hemispherical shell of energy.....41
15. The Tesla shield: A 1969 Virgin Islands incident.....41
16. Tesla shields: Two nested arcs of light in the sky.....42
17. Tesla shield: An expanding hemispherical circle of light.....42
18. Tesla shield: Hemisphere and globes, 1977.....43
19. Large Tesla shield observed by Red Chinese fighter pilots.....43
20. Large glowing "Tesla Globe," Red China.....44
21. 1966 test of giant Tesla globe, expanding deep within the Soviet Union.....45
22. A giant Tesla globe, deep within the Soviet Union in 1979.....46
23. Engines disabled, a passenger jet drops six miles in two minutes.....47
24. Giant "Woodpecker" beams intersecting over the U. S.....48
25. Third Soviet non-destructive testing against shuttle launches Nov. 26, 1985.....50
26. Death of the Arrow DC-8, Dec. 12, 1985.....52

LIST OF FIGURES

No. Title Page

27. Anomalous hole in the right fuselage 52
28. **Part** of the Soviet engineering of U. S. weather, just prior to the death of the Challenger 53
29. Soviet trawler spyship 54
30. Launch of the doomed Challenger, Jan. 28, 1986 54
31. Loss of a Delta rocket, May 3, 1986 56
32. Probable Soviet test of a Tesla Shield over Atlanta, Georgia 60
33. The riddle of time 99
34. Raindrop model of physical change 99
35. Phase conjugate reflection of a time-reversed wave 107
36. Phase conjugation by four-wave mixing 119
37. Soviet launch phase and midcourse strategic defense system 120
38. Power reflection coefficients and oscillation 123
39. Photons have virtual substructures 127
40. Once established, a virtual river is for free 129
41. Local energy conservation can be violated 130
42. T. H. Moray adjusts the controls of his radiant energy device 130
43. Charge affects anything existing in spacetime 136
44. Vacuum/spacetime is pure virtual particle flux 140
45. A simplified charged particle 141
46. A charged particle is a special kind of "spray nozzle." 141
47. An observable charged particle is directly connected to virtual particle vacuum flux 142
48. In a flux gradient (virtual particle river), a charged particle moves itself 143
49. A natural potential is a *disorganized* change in the stress of vacuum 143
50. An artificial potential is an *organized* change in the stress of vacuum 144
51. Gravitational potential is a conglomerate of stresses, organized or disorganized 144

LIST OF FIGURES

No. Title Page

52. EM force fields are releases of gravitational potential via observable charged mass flows 145
53. As beyond (without), so within 147
54. Internested levels of virtual state vacuum interact with neutrino, photon, and mass 147
55. Virtual state patterns are absorbed and emitted by the atomic nucleus 148
56. Nuclei of the universe continually exchange scalar EM waves and virtual particle fluxes 148
57. First order reality 149
58. Higher order (hyperspatial) reality 149
59. A photon is a single oscillation of a carrier wave 153
60. A potential is trapped energy 162
61. A force-field vector is a movement of mass in one direction 171
62. The zero-summed vector system is an engine and an artificial potential 173
63. An analogy for how scalar electromagnetics can engineer Kaluza-Klein unified field theory 182
64. The zero-vector system can engineer Kaluza-Klein theory 182
65. The basic idea of Klein's explanation of where the extra dimension is 183
66. Normal EM bleeds-off G-field 184
67. Scalar EM forces 4-dimensional gravitational field 184
68. The curl-free A-field 190
69. Earth, sun, and moon are locked into a triad system 195
70. Forming magnetic monopoles with a standing scalar EM wave 202
71. Metal fracture by depositing magnetic monopoles 203
72. The Kervran effect 205
73. The Kaznacheyev effect 215
74. A photon is one oscillation of an electromagnetic carrier 215
75. Nested modulations are the key to hyperspatial engineering 216

LIST OF FIGURES

No. Title Page

76. Hyperframes, vacuum, virtual state, minds and thoughts ... 233
77. Internested levels of virtual state vacuum contain mind and thought ... 237
78. The human cerebral cortex is a natural scalar interferometer ... 238
79. Consciousness and life ... 238
80. Life and death ... 239
81. Layers of unconsciousness intersect hyperspatially ... 239
82. Scalar electromagnetics can directly interact with the various levels of human mind and personality ... 240
83. What personality is ... 241
84. Jamming the "older files recall" process produces simple amnesia ... 241
85. To sustain one's personality, continual functioning is required ... 242
86. If the files themselves are changed, the personality itself is changed ... 242
87. The mind and personality can be affected—and changed or controlled—from a distance ... 243
88. The mind and personality are physical things, albeit hyperspatial and virtual ... 246
89. One of Priore's intermediate devices ... 248
90. The large special phase-conjugating plasma tube for Priore's giant device ... 252
91. Capping assembly for the top of Priore's giant plasma tube device ... 253
92. Royal Raymond Rife in his laboratory ... 268
93. Rife's first virus microscope ... 268
94. Side view of Royal R. Rife's prismatic universal microscope ... 269
95. Front view of Royal R. Rife's prismatic universal microscope ... 270

LIST OF TABLES

No. Title Page

1. Interesting Soviet statements 25
2. Microwave radiation of the U. S. Embassy in Moscow 32
3. Khrushchev's 1960 statement 35
4. Death of the U. S. S. Thresher 37
5. Severe EM interference experienced by U. S. S. Skylark 38
6. Massive 1985 strategic May Day exercise 48
7. Shuttle launch anomalies, Nov. 26,1985 51
8. Soviet escalation after 3rd test against shuttle launch on Nov. 26,1985 51
9. Arrow DC-8 disaster Dec. 12, 1985 53
10. Death of the Challenger, Jan. 28,1986 55
11. KGB celebration 55
12. Brezhnev's 1975 proposal 59
13. Symmetry 91
14. A force field is a special local symmetry 92
15. Gravity and local symmetry 92
16. Hindsight is better than foresight 93
17. When symmetry is violated (broken) 93
18. Some proven asymmetries 93
19. Spontaneous symmetry breaking 94
20. Why such drastic difference? 95
21. Aspects of strong local asymmetry 96
22. Flow of time 100
23. Physical change 100
24. Negative energy 101
25. Phase conjugate waves 108
26. Phase conjugate replica 108
27. Distortion correction theorem 108
28. Four-wave mixing 119
29. A zero vector system 135
30. Stress is fundamental 136
31. Vacuum is 140
32. Spin couples vacuum to particle 142
33. Residue universe, residue science 145
34. Reality is a hologram 151

LIST OF TABLES

No. Title Page

35. Physical change ... 153
36. Ordinary General Relativity (OGR) is a special case ... 155
37. Maxwell's equations ... 175
38. Kaluza geometry ... 181
39. 5-D G-field bleed-off (Kaluza 5-D theory) ... 181
40. Physical reality is eleven dimensional ... 183
41. Comparison of EM concepts ... 188
42. The Aharonov-Bohm effect ... 189
43. The living aura: The cell's electromagnetic breath ... 220
44. Lisitsyn's report ... 243

Chapters 4 and 5
contain complex scientific material
and are suggested for those who are
technically proficient.

FOREWORD

What This Book is About

In my book, **"Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons,"** (Tesla Book Co., P.O. Box 1649, Greenville, Texas 75401, 1986), I specifically pointed out that I was of necessity neglecting the biological warfare (BW) aspects. I intended to later write a second book covering the BW aspects.

This book contains the most important item of that planned second book, including the photobiology aspects of BW.

The material is put down in great haste, for it is meant to serve as an urgent warning. There is no time to do the six months additional research and compilation necessary to "flesh it out", sculpture it and meticulously document it, as should be done by a careful scholar.

I apologize to the reader for this lowered standard, but in this case the basic information is far more important than scholarship.

Further, I acknowledge the fundamental contributions of Dr. Robert B. Strecker, M. D., Ph. D., who together with his brother Ted, has for some time urgently called attention to the terrible threat posed by the AIDS epidemic.

I specifically acknowledge his important warning, "This is a Bio-Attack Alert," Mar. 28,1986, which was sent to the U. S. President and Vice-President, Governors of the states, various federal agencies such as Departments of State, Defense, and Agriculture; National Security Agency, Federal Bureau of Investigation, Central Intelligence Agency; and three members of Congress.

Dr. Strecker has also strongly pointed out the covert involvement of Soviet and other communist agents in the epidemic, and the absolute necessity for an *electromagnetic* solution.

Also recognized are the efforts of the other medical scientists, such as Dr. John Seale, M.D., and Dr. Eva Snead, who have tried repeatedly to call attention to the nature of the AIDS initiation, only to have their submitted papers resoundingly rejected by important journals such as New Scientist, **The** Lancet, etc.

Work by Dr. William C. Douglass in alerting the public to the no-nonsense truth about aids is particularly recognized. His monthly newsletter, **The Cutting Edge,** is an oasis of clarity sharply standing out in a sea of misinformation. Permission to reproduce his published research article on UV irradiation of the blood to successfully heal viral infections is deeply appreciated.

The continuing efforts of Dr. Andrija Puharich and Dr. Robert Beck to assist in achieving an AIDS solution are pointed out and recognized. Important contributions by Ed Skilling are also recognized in this regard.

The vital and extensive contributions of Christopher Bird are particularly acknowledged. Chris has unstintingly and unselfishly shared with me the results of his enormous research on the Priore device and affair and the results of his translations from the French. He has most graciously permitted me to publish his important summary of the Priore affair, and I am greatly in his debt. The reader will find that material most illuminating, and well may be shocked at the depth and malevolence of the scientific and political suppression of Priore's rigorously demonstrated electromagnetic cure for cancer, leukemia, and other killer diseases.

Special acknowledgement is made to Robert Whitney, who so nobly labored to bring the Priore technology into the mainstream of medical treatment, only to be ruthlessly suppressed in his efforts. Photographs of the Priore machine were furnished by Bob from his collection.

My special thanks to colleagues John Bedini, Frank Golden, Andrew Michrowski, George Fencl, Ken Moore, Tom McLaughlin, Joe Parrott and Randy Davidson, without whose assistance this project could not have been accomplished.

The material in this present small book is intended to be of any assistance possible to certain brave colleagues who — with little or no funds or resources except what they can tear out with their bare hands — are already laboring around the clock to develop a permanent solution to the AIDS problem.

As a matter of discretion, I shall not mention the names of those colleagues. The solution sought is radical, by orthodox assessment. Yet it is the only possible approach that has any chance whatsoever of succeeding.

What they are doing, I cannot do. If they do not succeed, no one can. But I will say this, and most strongly: In my view, the only hope for humanity lies in the work they are doing, and they have my utmost admiration and support.

Their initial results are exciting and encouraging. Basic phase conjugated electromagnetic "mixed-modulations" signals lethal to bacteria, nematodes (hookworm-like organisms), various amoeba, viruses, etc. at very weak power levels have been synthesized and demonstrated.

In my best estimate/opinion, these researchers must succeed in *two years or less, certainly in no longer than three years,* else the **AIDS** incubation already accomplished by then, and the implementation time delay — even given a cure —together will insure a devastation of humanity unparalleled in all of history.

Let me accent that developing this AIDS cure is far more important than developing our entire arsenal of new military weapons. It is even more important than the development of defenses against the Soviet scalar electromagnetic weapons I detailed in **Fer-de-Lance.** Problems in the Middle East, Korea, and Nicaragua are insignificant compared to the AIDS problem.

Desperate Diseases, Desperate Remedies

Spurred by the looming holocaust of AIDS, and by the strong efforts of my colleagues, I have decided to release specific *information on how to develop electromagnetic healing, essentially against any and all of the "killer" diseases including cancer, leukemia, and AIDS.*

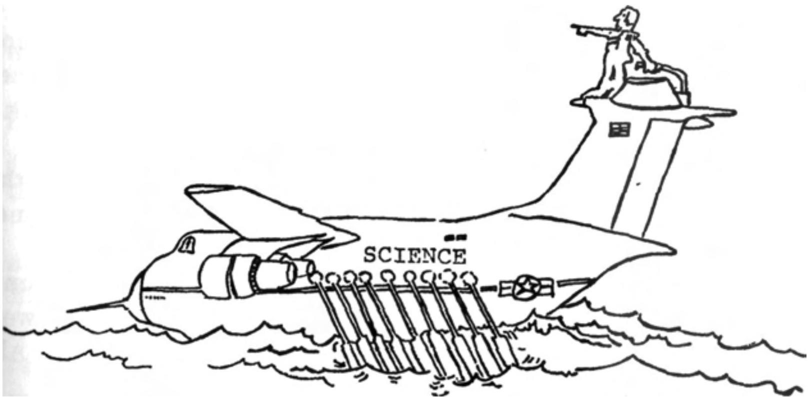

Figure 1. Conventional science cannot do the job.

Unavoidably this also releases a major principle of scalar electromagnetics: what is really going on in phase conjugation. If phase conjugation is truly understood and applied, all the Soviet weapons I have pointed out in **Fer-de-Lance** can be understood, designed and developed.

I'm also releasing specific information on mind and thought that will enable them to eventually be directly detected, measured, and investigated scientifically. Unfortunately, this also will enable them to be technologically manipulated for nefarious purposes.

However, the developing scalar EM technology for the AIDS cure must not be hamstrung. There is no time to apply the standard "Let's classify it!" response. If we do, we'll win the skirmish and lose the war.

As shocking as it sounds, I am convinced that the Soviets have already launched WW III, to the hilt. And it's not at all the kind of war we've been expecting.

There's no time to rail against the Soviets, who have cunningly and in great secret "struck our vitals with the mailed fist." We've been struck in such a way that, *unless we can solve the AIDS problem, in two to three years or less, we are already finished.*

We have been preparing to fight a better World War II. The one we're preparing for is never going to be fought. The real World War III is already raging. Our death knell has been sounded, *and we've not yet even recognized that the battle is underway.*

A great Soviet "first strike to finish us" has already been accomplished. The new "Pearl Harbor" has already happened. The enemy has already overrun our Maginot Line; we are at our Dunkirk. There are no boats waiting, and there are none on the way.

The great Soviet first strike has been delivered with surreptitious biological warfare, not with nuclear weapons or ground forces.

As always in the past, we have almost no troops facing the brunt of the first strike and trying to turn it. This time we have far fewer than we've ever had. Just two or three unorthodox researchers. All the orthodox researchers are manning the wrong ramparts.

Ironically, this time our entire system—the government, the medical establishment, the scientific establishment, the universities — all are far too dogmatic, and have far too sluggish a response time, to offer any assistance at all. They are simply so "out of it" that they don't even know what's really happened.

The bureaucracy simply is so stagnated that, even with a giant paroxysm, it cannot change quickly enough to do the job in time.

The best response the orthodox system could theoretically make —if it immediately placed the nation under martial law, marshalled all the universities and scientists, everything — would require at least seven years to produce results. A more likely response time is 10 to 14 years. That won't do the job.

So the job has to be done from completely *outside* the system. It has to be done with *independent* financing. Financing without a lot of conventional, bureaucratic strings on it. Resources must be committed as needed, immediately, not with a several months' approval process.

Everything in the orthodox establishment simply has to be bypassed.

The government, the medical establishment, the scientific establishment, the universities, everything.

Otherwise, we're as good as finished *right now*.

The "Dirty Dozen"

In the West, there are only perhaps a dozen persons or so that I know of, who are qualified to work on the main task required. They are all unorthodox researchers and scientists.

There are thousands of fine orthodox scientists, of course, who can help with the peripherals — making and testing viral solutions, setting up and operating equipment, doing interminable tests, all the lab stuff, etc.

But only about 12 unorthodox guys are qualified and mentally and emotionally prepared to come to grips with the particular scalar EM technology required for rapid development of an AIDS cure.

Right now, there are just two or three of the 12 working on this, hammering away as best they can, at terrible cost to themselves and their normal endeavors.

Obviously immediate funds are needed for facilities, equipment, and marshalling that dozen unorthodox researchers and a team of scientists — along with support staff and personnel.

An Analogy

I'm reminded of a story from WWII.

On the beach at Anzio, in a battle lull some doughboys were briefly resting near their foxholes, eating lunch from C-ration cans.

A lone German Messerschmitt appeared, closing on them fast and low with machineguns blazing. They were forced to hastily jump into their foxholes as the Messerschmitt strafed their position.

The pilot continued to make strafing run after strafing run, low and "on the deck." Obviously lunch was rather totally disrupted.

Suddenly, in total disgust a soldier leaped up and flung his half-filled C-ration can into the air directly at the approaching aircraft, full in the teeth of the machinegun fire.

Incredibly, the can sailed straight up into the air intake of the hurtling aircraft, where it was sucked into the engine, which sputtered and died.

The surprised German pilot banked his plane upward and bailed out.

He was promptly captured by the very troops he had been strafing only moments before.

Those two or three colleagues of mine are literally "throwing their C-ration cans" in a desperate effort to ward off a monstrous threat.

Those two or three are striving to the limit of their endurance.

This book and the release of the material in it represent my own "Oration can" hurled into the teeth of the actual Soviet threat as a last resort. Hopefully we can "hit the air intake" of the Soviet juggernaut, so to speak, and stop its progress.

The giant Soviet scalar electromagnetic weapons that I detailed in **Fer-de-Lance,** e.g., are easily countered — in six months or less. Primarily, it is necessary to make a pumped phase conjugate mirror adjunct (PPCMA) that operates at the frequency band of interest, and associate it with a receiver. Once a signal is received from the enemy weapon — say, from a Woodpecker transmitter — it is introduced to the PPCMA. A powerful phase conjugate replica energy pulse is returned into the distant hostile transmitter, totally destroying it.

To meet the viral biological warfare weapon, however, we must develop electromagnetic healing.

In an important article, "SDI vs. the plague," Mar. 8,1988, p. 6, the **Washington Times** revealed a startling Pentagon conclusion that the Soviets may counter SDI with genetically engineered weapons.

Valentin Falin, Chief of the Soviets' Novasti Press Agency, is quoted as saying:

> *"We won't copy you anymore: We'll take asymmetrical means with new scientific principles... genetic engineering could be a hypothetical example. Things can be done for which neither side could find defenses or countermeasures, with very dangerous results. If you develop something in space, we could develop something on earth. These are not just words. I know what I'm saying."*

The genetically engineered Soviet counterstrokes exist and are already being taken. This book tells you what the asymmetrical means, new scientific principles, and genetically engineered weapons are.

May God help us all. If we do not move immediately and correctly to augment the effort of my colleagues, the Free World and half of the peoples of the earth will perish before another decade passes.

CHAPTER 1

THE VACCINE CONNECTION

The following quote is from Don Rowe, **Wall Street Digest,** 214 Carnegie Center, Princeton, New Jersey 08540.

"The AIDS Plague will affect society in ways that you cannot now imagine. AIDS is 100% fatal. Scientists and medical research people are not optimistic about an immediate cure to [or] vaccine. They do not expect to develop an effective vaccine within the next two decades. Public health officials estimate that as many as 2.4 billion people (half the world's population) will die from AIDS within the next 15-20 years."

"Economically, the insurance and medical health systems could be devastated in the 1990's. Nothing short of a spectacular medical breakthrough will keep Western civilization from suffering the worst catastrophe in the history of the world."

The first official U. S. AIDS case was diagnosed in San Francisco in 1981. This followed the puzzling outbreak of the disease in Third World areas such as Africa and Haiti in the 1970's.

The U.S. AIDS outbreak at first seemed to be confined to the gay community, such as in San Francisco.

No one seemed to connect the apparently unrelated facts that, before the outbreak in the U.S. gay community, (1) the World Health Organization had accomplished substantial smallpox vaccination in Haiti, and (2) West Coast gays — particularly from San Francisco — had made Haiti a main playground and vacation spot.

Most orthodox medical scientists have not the foggiest notion why such a drastic "explosion" of the virus has occurred.

It has occurred because *AIDS was introduced by contaminated vaccine used by WHO in its smallpox eradication campaign.* A high correlation exists between WHO vaccination and subsequent outbreak of AIDS.

Figure 2. The "monkey-bite" theory.

The "Monkey-Bite" Theory

As the prevailing rationale goes, a naturally-mutated monkey AIDS virus derivative or strain is believed to be the cause of the monstrous plague looming to strike down half of all the humans on earth.

That is, a monkey is supposed to have bitten a man—an African—and transmitted the normally harmless monkey AIDS virus to him. The virus is then supposed to have naturally "recombined" its genetic material with other viral genetic material and the man's human genes, to produce the virulent new strain of AIDS.

Then this new strain, *from its single human host,* is supposed to have rapidly started spreading through Africa and around the world.

That is simply nonsense.

In Africa, for example, monkeys as well as other animals have had their viral diseases for thousands of years. If monkey virus were the culprit, African man would have been devastated by AIDS thousands of years ago. He either would have been dead or—as have the monkeys — he would have developed a certain resistance and tolerance for the disease, so that it is no more harmful than others.

Understand, the "monkey bite plus human cell modification" undoubtedly does occur in Africa with great rarity — once in several thousand years.

However, even if that rare natural recombination of the monkey AIDS virus were to occur in a single human, normal contact between other humans and the infected human is totally insufficient to provide any substantial infection rate *unless the rate of infection of new individuals is greatly assisted and vastly speeded up by some other outside mechanism — for example, by something like AIDS-contaminated vaccine and a mass inoculation program.*

Retroviruses **and Recombinants**

As Dr. John Seale has pointed out, "The AIDS virus (human immunodeficiency virus or HIV) is a lentivirus — a little-studied sub-family of the retroviruses. It is highly pathogenic to man but it differs profoundly from any other virus of humans. It is spreading rapidly in the general population of central Africa, and amongst homosexual men and drug addicts in the West. *It is the first virus to have appeared in mankind for many centuries which is entirely new, highly lethal, and spreading steadily from person-to-person world-wide."* [accent by the present author.]

Dr. Seale has courageously pointed out some other very disturbing facts:

(1) Retroviruses of animal origin, if repeatedly passed from human cell culture to human cell culture, will gain the ability to infect human cells *preferentially.* Indeed, they may then even cease to be infectious to their original animal hosts.

(2) In 1982 Robert Weiss pointed out that, in the laboratory, the genetic code of retroviruses commonly recombine if two of them infect the same cell simultaneously.

(3) In a 1984 article in the **Journal of Virology,** David Baltimore showed an efficient means of overcoming the cell's block against simultaneous infection by two or more retroviruses.

(4) Further, Baltimore reviewed how *this had widely been done during the preceding ten years!*

(5) Outside the laboratory, recombination of multiple simultaneous retroviruses has not been shown to occur naturally in humans.

(6) In animals, there is some evidence that it may occur naturally with extreme rarity — indeed, only once in many thousands of years.

(7) By the early 1970's, the technology to make new recombinant retroviruses virulent to humankind was already developed and had been widely published.

To the layman, the infection and reproduction by retroviruses is peculiar to say the least.

When a retrovirus infects a cell, from its own genetic material (RNA) it first makes a "template" or pattern (DNA) for forming other retroviruses just like it. It then inserts this "DNA template" into the DNA of a cell. The host cell itself is now a "sleeping factory", ready to start producing retroviruses.

When activated, the "template" in the cell causes the factory to go into production of retroviruses. Activation is varied. The "infection production factory" can just remain dormant, be slowed over years, take place almost immediately, or be slowed for years and then "explode" to rapidly increase the infective viruses.

Genetically, the cells of any person infected with AIDS or any other recombinant retrovirus have been permanently transformed. Literally that person has acquired a new set of genes. He or she has been converted into a "factory" for producing—and spreading—the virus.

Cancer Research Produced Ideal Biological Warfare Agents

In 1971 President Nixon signed into law the National Cancer Attack Act. Though certainly not intended by the President and the Congress, *the inadvertent result of this act was that taxpayer's funds were used to research and develop the precise things that constitute the most lethal kind of viral biological warfare.*

The on-going small effort in medical research laboratories to alter animal tumor viruses — including retroviruses —to replicate efficiently in human cells was drastically speeded up, as part of the new "war on cancer." In other words, whether intended or not, much of our cancer research money was converted into the precise kind of research that develops virulent biological retroviruses for potential use in biological warfare.

Not only was the research drastically accelerated, but also the direct participation by Communists and scientists sympathetic to the communist cause was accepted and encouraged.

Astoundingly, this open invitation to the communist scientists was issued at a time when one of the high KGB priorities was to infiltrate our recombinant DNA genetics laboratories! To the Soviets' utter astonishment, we simply opened our labs and invited them in.

The Soviets participated — and penetrated — with alacrity. It is an obvious conclusion that a number ofthe Iron Curtain scientists sent to help staff our laboratories were undercover KGB agents.

In the cancer laboratories, on a wide-spread scale, animal retroviruses were injected and reinjected into human cell cultures, until these retroviruses had become efficient, lethal recombinants favoring human cells as their hosts.

At that point, infection of one or more persons by one of these lethal new retroviruses could unleash a massive AIDS-like plague on all mankind.

A single laboratory accident (or a single act by one Communist agent) could unleash the Great Death Plague.

Weak and ineffective reference to this forthcoming danger was made in 1974 by the Committee on Recombinant DNA Molecules, which had been established by the U.S. National Academy of Sciences. According to **Science,** July 26,1974, p. 303, the Committee made the following rather bland and ineffectual recommendation:

"Scientists throughout the world (should) join with the members of this committee in voluntarily deferring experiments (linking) segments of the DNA from (cancer-causing) animal viruses to... possibly increase the incidence of cancer or other diseases."

So much for the great scientific watchdogs.

U. S. "Cancer War" Labs Were Infiltrated

As we said, ironically we ourselves helped load and cock this potentially disastrous biological warfare cannon. Incomprehensibly, enemy fingers were permitted — even encouraged! — to be placed on the trigger. Many of our medical research facilities involved in this research contained known Communists, dissidents, etc.

For example, see page 106 in Omni, March 1986. Carlton Gajdusek, presently Chief of the National Institute of Health's Laboratory of Central Nervous System Studies and Laboratory of Slow, Latent and Temperate Virus Infections, was asked the question, "Isn't Fort Detrick in Maryland such a biological-warfare research facility?" He responded:

> **"No, emphatically no! There is no defensive or offensive warfare microbiology done at Fort Detrick today. It is the national cancer research facility of NIH. In this facility I have a building where more good and loyal Communist scientists from the USSR and mainland China work — with full passkeys to all the laboratories — than Americans. With night-working U. S. citizens and foreign Communist investigators here, obviously there is no "secret" bacterial warfare activity going on. Even the Army's infectious-disease unit is loaded with foreign workers — not always friendly nationals. It is a valid basic research unit on worldwide problems of infectious -diseases in which no classified or secret activities unfold."**

The point is, what was being produced in these labs were viruses that would grow in human cells. Under tight control, these new lethal agents could be safely studied, and many things could be tried experimentally to see if they would slow, stop, or kill the viruses and heal or benefit the virally infected human cell cultures.

The viruses were being developed to benefit man by aiding in the war on cancer — we accent this most strongly.

However, many of them were precisely the same things that would be developed in a biological warfare lab: lethal new viruses that preferred the human as a host, and against which humans have developed little or no immunity.

The Deed Was Easily Done

Now one can see the subtlety and irony in the Soviet Union accusing the U. S. of having made the AIDS virus at Ft. Detrick!

It's both true and false at the same time.

There is substantial information indicating that the AIDS disease is manmade, and has been deliberately unleashed on mankind by Soviet and other Communist agents, augmented by the international control groups and shockingly inept work by U. S. and European medical researchers.

Human cell cultures deliberately infected with animal viruses were carelessly handled in U. S. and European laboratories — carelessly handled in that far too great (essentially open) access to the new viral strains was permitted.

Our own scientists were apparently unwitting pawns, and Soviet agents performed the "final touchups" to insure release — whether intentional or inadvertent — of the lethal new agents against an unsuspecting world.

Viruses can infect bacteria, since a bacterium is a one-celled organism and the virus is a single molecule.

Cell cultures where vaccine is made from weakened bacteria could easily be contaminated with the desired recombinant viruses available in the cancer laboratories.

To insure worldwide release of the lethal virus, covert Communist agents apparently caused contamination of the smallpox vaccine that was prepared in quantity in Western laboratories.

This contaminated vaccine then was widely used to inoculate Third World peoples and others, in the successful international program to eradicate smallpox.

With only a few agents in key places, the Soviets could easily have taken advantage of the looseness of Western laboratory procedures to deliberately contaminate the animal cell cultures used to make the vaccines.

They had an abundance of agents in the key laboratories, for they were openly invited in and accepted!

The Soviets had their fingers in it, up to the elbow.

And at just exactly the right time, the answer to a Communist's biological warfare dream was available: The World Health Organi-

©Hal Crawford 1979

Figure 3. The Soviets have delivered an intended knock-out.

zation was already involved in its 13-year campaign of massive smallpox inoculation in the Third World, to eradicate smallpox.

All that was needed was to contaminate the WHO vaccine by one or more of the lethal new recombinant retroviruses that had been developed — and were continually being developed — in cancer research laboratories filled with Communists agents.

The opportunity was presented for a great Soviet "first strike" against the West, using biological warfare agents developed in U.S. and European laboratories and administered by an international agency.

The opportunity was ripe for the taking. It was irresistible. You can be sure the Soviets took it.

It's Actually Documented!

Virus-contaminated vaccines were then widely administered to Third World countries—particularly in Africa, South America, and Haiti. Administration of these vaccines by the World Health Organization is precisely correlated to the subsequent outbreak of the AIDS epidemics in those countries.

This strongly indicates that the vaccine itself was the "precursor" agent initiating the start of the AIDS infection.

In a shocking article, "Smallpox vaccine 'triggered Aids virus'," in the London **Times** of May 11,1987, **Times** Science Editor Pearce Wright announced that the AIDS epidemic may have been caused by World Health Organization (WHO) smallpox inoculations in third-world countries.

In an important article in **Easy Reader,** June 4, Jon Rappoport reported the **Times** article and the results of an interview with Robert Matthews, technical correspondent of the **Times.**

According to Rappoport, Matthews informed him WHO itself had somehow suspected that its immunization program against smallpox might be connected to AIDS. An outside consultant was employed to perform an independent study, and the study confirmed that WHO vaccines were indeed — somehow — causing AIDS to evidence in inoculated persons.

Needless to say, WHO promptly buried that report—whose conclusion must have been the last thing WHO wanted to hear.

The consultant then came to the **Times.** The **Times** heard him out and printed the story.

This astounding story has almost startling correlations: *"The smallpox vaccine theory would account for the position of the Central African states as the most afflicted countries; why Brazil became the most afflicted Latin American country; and how Haiti became the route for the spread of AIDS, to the U.S.: Brazil, the only South American country covered in the eradication campaign, has the highest incidence of AIDS in that region."*

Rappoport confirms that the subject of contaminated vaccines—"vaccines which, made in tissue cultures of animal parts, contain 'extra' dangerous viruses" — has been of increasingly serious concern to medical scientists in recent years, and the WHO smallpox vaccines may have been contaminated in this fashion.

A documented U.S. Army incident is a case in point.

New U.S. Army recruits are routinely vaccinated against a range of diseases. One of them is smallpox, in case an enemy should try to use smallpox as a biological warfare agent.

A healthy 19-year old Army recruit developed AIDS after vaccination, and was admitted to Walter Reed Army Medical Center, where he died.

In a paper published in the **New England Journal of Medicine,** Walter Reed medical team scientists reported their discovery of the connection between vaccination and stimulated AIDS disease.

They also warned against the WHO plan to use modified versions of the smallpox vaccine to combat other diseases in developing countries.

WHO's 13-year vaccination and eradication program ended in 1980. The program is credited with saving two million lives a year and preventing 15 million infections.

It may be inadvertently responsible for the ultimate death of billions.

In a meeting of 50 experts near Geneva in May, 1987, it was revealed that millions of new AIDS cases are about to hit Southern Africa. The experts believe that up to 75 million Southern Africans — about a third of the population — may have AIDS within the next five years.

From our previous statements above, it appears almost certain that Communist agents who had penetrated U.S. and European laboratories manipulated the contamination of the altered vaccines.

Further, the Soviets may be stimulating the disease and guiding it electromagnetically, by special implanted signals carried by the giant Soviet woodpecker "over-the-horizon radars" and other Soviet transmitters.

Or, the Soviets may be using their own transmitters and ours, including our own power lines and radio and TV stations, to broadcast signals that "precondition" and greatly amplify our population's vulnerability to certain selected diseases — AIDS and otherwise.

We will address these issues and Soviet subtle EM capabilities later.

Soviet Biological Warfare Background

In World War I, the Russians suffered half a million casualties to gas warfare, including 50,000 dead.

At 10,000 troops per light division, that's the equivalent of 5 divisions killed and 45 divisions incapacitated.

This made a deep and lasting impression on the Soviets. Lenin himself ordered a major effort mounted to develop and manufacture chemical and biological weapons, for he understood the implications of such mass-destruction weapons.

In 1919 the Soviets began work in earnest on BW weapons. Work on them has not stopped since then. It will not stop so long as the Communists remain in power.

To loyal Communists, Lenin's dictate to develop and lead the world in biological weapons is absolute and unquestioned.

By 1937, the Red Army was fully equipped and psychologically prepared to wage chemical and biological warfare.

Continuing today, the Soviet forces have the highest training in chemical warfare of any troops in the world. Red Army soldiers practice regularly, often using actual chemical agents. Decontamination is a standard measure and practiced as a matter of course.

Even though the Soviets, along with many other nations, signed the Geneva Protocol of 1925 which banned the use of deadly chemical and bacteriological weapons, the Soviets have never ceased development of such weapons.

In the 1930's, a secret Soviet bacteriological institute was set up at the town of Suzdal. Prisoner microbiologists worked on a variety of new weapons and vaccines, including bubonic plague and tularemia (rabbit fever). By World War II the Soviets were prepared to retaliate vigorously to any use by the Germans of chemical and biological warfare agents.

After the war, Soviet research and development in chemical and biological weapons continued apace. However, much of the genetic work was hampered by Stalin's stern support of the autocratic control of agronomist Trofim D. Lysenko. Lysenko's crackpot theories on genetics, for example, held that genes did not even exist. After the death of Stalin, Lysenko's work was discredited, and the Soviets set forth immediately to "catch up" with the West in genetic implications such as recombinant DNA techniques.

Thus in the early 1960's the Soviets began sending Russian scientists to the West to catch up with Western genetics research. One of these scientists was David Goldfarb, who later became a leading Soviet expert in the field of molecular genetics. He headed up the Laboratory of Molecular Genetics of Bacteria and Bacteriophages of the Soviet Academy of Sciences in Moscow. Although he did not work on secret Soviet projects, the Soviet authorities stopped him cold when he applied for a visa to leave in 1979. Goldfarb was a specialist in plasmids, antibiotic resistances, and recombinant DNA technology—all very, very useful to the Soviet effort in biowar.

One can readily understand the reluctance of the authorities to release scientists such as Goldfarb. After all, the Soviet "dirty work" with the deliberate AIDS contamination of the World Health Organization's smallpox vaccine had already been accomplished and a great Soviet biological warfare "strike" was already underway, thanks to the massive smallpox vaccination program of WHO.

In 1972, in the Biological and Toxin Weapons Convention, the Soviet Union and every major power agreed not to develop or use biological weapons. The agreement went into effect in 1975. Although it bans the development, production, and stockpiling of biological agents and toxins for hostile purposes, it has not slowed or altered the Soviet program.

In the early 1970's, Soviet leaders made a decision to use genetic engineering to produce new and improved BW weapons. Highly secret institutes for this purpose were set up in Moscow, Leningrad, and Novosibirsk. These laboratories specialize in seeking ways that genetic engineering techniques can be used to develop virulent strains and agents that their opponents cannot resist.

Every Soviet civilian institution dealing with bacteriology and epidemiology is responsible for certain military work as well. The Soviet Academy of Sciences in fact funds appreciable research of such nature at these institutions.

One of the major Soviet centers for recombinant DNA techniques in biological weapons research is in the village of Protvino, near Moscow. Research on chemical and biological weapons is very easy to conceal, and the Soviets are sure that they will not be caught. Satellites cannot detect what is going on in the labs. The payoff is big and the risk is minimal. The Soviets push this area vigorously.

A number of installations capable of producing disease agents and toxins on a large-scale basis and placing them in munitions and delivery systems have also been identified. One such facility is in the city of Sverdlovsk and has a long history of BW research and development and production, with emphasis on anthrax. During early April 1979, an accidental release of anthrax occurred in Sverdlovsk that caused many casualties and deaths in spite of heroic efforts by Soviet doctors and cleanup and decontamination crews.

Note that, according to DoD's 1987 edition of **Soviet Military Power,** page 110, *"Anthrax causes a high mortality rate when the infection results from ingestion (up to 70 percent fatal) or inhalation (almost 100 percent fatal) if treatment is not promptly begun."*

In addition to military BW institutes, military officers and KGB agents are assigned to key posts in the Soviet Union's normal health and science ministries to monitor research and development that has possible military significance.

Also, the KGB has a continuing and highly active program to insert recruited Soviet Scientists in Western genetic research institutes and agencies. This is a high priority effort, and it has been very successful. Also, it obviously says something about the probability of Soviet agents penetrating U.S. and Western European cancer research institutes, particularly those using recombinant DNA techniques. To make it even worse, we simply invited them in with open arms!

Indeed, it was that KGB program plus the "open arms" in Western cancer virology research institutes that enabled the Soviets to set up and bring about a massive "first strike", using WHO and Western laboratories to do the actual dirty work unknowingly.

We *Need An Immediate, Full Scale Investigation*

Annually, President Reagan makes a report to the U.S. Congress on the subject of Soviet noncompliance with Arms Control agreements.

In April, 1984 Defense Secretary Caspar Weinberger put it this way:

"There is an apparent effort on the part of the Soviets to transfer selected aspects of genetic engineering research to their biological warfare centers... Soviet research efforts in the area of genetic engineering may also have a connection with the biological program...Normally harmless, non-disease-producing organisms could be modified to become highly toxic or produce diseases for which an opponent has no known treatment or cure. Other agents, now considered too unstable for storage or biological warfare applications, could be changed sufficiently to be an effective agent."

At the time that Secretary Weinberger made this statement, a devastating biological warfare strike using those exact techniques had already been unleashed on the Western world by the Soviet Union, cleverly disguised under the banner of the World Health Organization's smallpox vaccination program.

It is imperative that the Congress of the United States launch a full investigation of the entire background of the AIDS biological warfare strike, as a matter of the highest priority.

The investigation should include the role of the Soviet Union, fellow-travelers, and Western scientists openly sympathetic to the Soviet cause.

It requires little investigation to see that there is a well-marked trail for several decades, showing that such a biological warfare strike against humanity has long been considered — and even preferred — by those advocating the Communist cause and world domination.

For example, Bertrand Russell, philosophical spokesman for a group of Westerners, explicitly stated this diabolical BW capability for consideration, in his **Impact of Science on Society:**

"At present the population of the world is increasing at about 58,000 per diem. War, so far, has had no very great effect on this increase, which continued throughout each of the world wars... War has hitherto been disappointing in this respect... but perhaps bacteriological war may prove effective. If a Black Death could spread through the world once in every generation, survivors could procreate freely without making the world too full. The state of affairs might be

unpleasant, but what of **it?"**

The blood of every American — and every European, African, and Asian — should curdle at such words. We are now facing the very "Black Death" type of plague referred to by Russell in such glowing terms!

Highly influential followers of the Russell philosophy — and his callous approval of the periodic slaughter of much of the world's population — may have continued to view the great destructive power of new viral warfare agents as the best means to achieve their end: Selective thinning of the world's population.

At least a few such men may have continued to cooperate with the Soviet Union to work powerfully behind the scenes in international organizations, trying to arrange the proper opportunity to achieve their nefarious purpose.

The "East-West cooperative effort" in cancer research created and gave them the precise biological agent custom-made for the job.

WHO's massive, worldwide smallpox inoculation program gave them the precise means to rapidly get the deadly virus into enough human veins to assure a new "Black Death."

All that had to be done was surreptitiously contaminate the vaccine — the easiest thing of all.

This plague has been knowingly, with malice aforethought, unleashed upon mankind. The perpetrators should be hounded down like the wild beasts they are, uncovered in the full light of day, and suitably punished with the swift and sure justice they deserve.

At least one outstanding medical consultant had the courage to report to WHO that their vaccine was generating the AIDS epidemic. When WHO suppressed the evidence, the London Times had the courage to print the story.

I strongly urge every U.S. Senator and Congressman to introduce and/or support immediate legislation that would launch a full-scale, massive investigation, with no stone left unturned. I urge every concerned activist to demand and support such an investigation at the highest level.

By preoccupying ourselves with lesser things — "IranGate," Nicaragua, the Middle East, etc. — we are simply fiddling while Rome burns.

CHAPTER 2

THE SOVIETS KNEW THE U.S. WOULDN'T REACT

Administering the Coup-de-Grace

As we've seen, the present AIDS viruses were deliberately created in Western cancer research laboratories strongly penetrated by Communist scientists and Soviet agents.

In laboratory experiments that were a legitimate and vital part of the war against cancer, animal retroviruses were repeatedly injected into human cell cultures until they recombined to produce **strains** hostile to human life—strains that would thus live in human cell cultures.

Even though their work was accomplished for a totally different purpose, cancer laboratories produced the ultimate, made-to-order biological warfare agents.

The World Health Organization (WHO) is also strongly penetrated — and even controlled at some levels—by communist agents, **as** are many other such international agencies.

And WHO's massive smallpox vaccination program in Third World countries provided the Soviets a made-to-order opportunity to have (ostensibly) "someone else" deliver the actual BW strike.

In other words, the Soviet deception plan was complete.

It was very simple for Soviet agents to contaminate the cultures in which smallpox vaccine was being produced for WHO's inoculation campaign. This insured the production ofvaccine contaminated with AIDS viruses, and the subsequent massive infection of the African peoples and other Third World nations.

It delivered what will be the knockout punch against the West, unless AIDS is countered.

Table 1. Interesting Soviet Statements

"WE HAVE A NEW WEAPON, JUST WITHIN THE PORTFOLIO OF OUR SCIENTISTS, SO TO SPEAK — SO POWERFUL THAT, IF UNRESTRAINEDLY USED, IT COULD WIPE OUT ALL LIFE ON EARTH."
NIKITA KHRUSHCHEV,
1960, SPEAKING TO THE PRESIDIUM

"WE WILL BURY YOU!"
NIKITA KHRUSHCHEV,
TO RICHARD NIXON, "KITCHEN" DEBATE

"IF A MEANS OF TOTAL NEUTRALIZATION OF FOREIGN MISSILES IS TO BE FOUND, IT CAN ONLY COME FROM *A GROUP OF NEW PRINCIPLES IN PHYSICS, CALLED ENERGETICS!"*
NOBELIAN PETER KAPITSA, TO NIKITA KHRUSHCHEV

"EACH SIDE SECRETLY DEVELOPS NEW MEANS OF WARFARE IN ORDER TO EMPLOY THEM UNEXPECTEDLY. HISTORY KNOWS MANY EXAMPLES HOW THE EMPLOYMENT OF A NEW WEAPON INITIALLY GAVE CONSIDERABLE SUCCESS BECAUSE THE ENEMY, CAUGHT UNAWARES AND NOT KNOWING THE COMBAT CAPABILITIES OF THIS WEAPON, WAS FOR SOME TIME INCAPABLE OF EFFECTIVE COUNTERACTION."
V.YO. SAVKIN, THE BASIC PRINCIPLES OF OPERATIONAL ART AND TACTICS. MOSCOW, 1972.

"OF PARTICULAR IMPORTANCE IS BASIC RESEARCH AIMED AT DISCOVERING STILL UNKNOWN ATTRIBUTES OF MATTER, PHENOMENA, AND THE LAWS OF NATURE, AND DEVELOPING NEW METHODS FOR THEIR STUDY AND USE TO REINFORCE THE STATE'S DEFENSE CAPABILITY."
PORTION OF A BOOK BY MARSHALL GRECHKO
DELETED FROM ENGLISH TRANSLATION, BY SPECIFIC REQUEST OF THE U.S.S.R.

"WILL SHORTLY SEE DEVELOPMENT OF NEW WEAPONS, MORE POWERFUL THAN NUCLEAR WEAPONS, AND THEY WILL BE NONVERIFIABLE."
LYSENKO (FIRST NAME UNKNOWN)
USSR REP/SOVIET EMBASSY WASHINGTON, D.C.
WORLD FURTURESCONFERENCE, WASH. D.C, 20 JUL 82.

Murky "Behind-the-Scenes" Intentions

As was pointed out, the WHO inoculation program is well-correlated as the precursor or "initiator" of the AIDS epidemic.

Where WHO smallpox vaccination occurred, the mysterious symptoms of AIDS disease appeared in due time.

Even the eventual spread of the AIDS virus to the U.S. homosexual community is explainable.

For example, the economy of Haiti had earlier been made a shambles. The poor citizens of that impoverished island had been reduced to desperate straits.

In desperation many younger Haitian males had begun to prostitute themselves to visiting homosexual tourists. The Haitian males were not homosexual themselves; rather, they were desperate to gain money to provide for their starving families.

In response, U.S. homosexuals — particularly from San Francisco and New York — had made Haiti their vacation playground. They even ran ads to this effect in their gay newsletters and mngazines.

WHO then vaccinated some 15,000 or so Haitians against smallpox, using the vaccine contaminated with the AIDS virus. This introduced AIDS into Haiti and into the Haitians.

The U.S. homosexuals quickly picked up the AIDS virus by direct contact, and brought it back to the U.S. with them. There the nature of the activities engaged in by homosexuals was responsible for enhanced spread of the virus among members of the homosexual community.

Engineering the AIDS epidemic appears to have been a collusion between Communists and the international control groups.

Apparently the control groups wished to "clean out the African continent and other Third World areas" — possibly to provide room for relieving population pressure in the developed countries. This may have been seen as the way to solve the problem of the ballooning human population — the so-called "biological time bomb."

The Soviets, however, realized that contamination of the smallpox vaccines would result in an epidemic of unparalleled proportions. They recognized the opportunity for safely initiating — and winning — World War III with a great first strike.

The Soviets themselves already possess an electromagnetic cure for diseases such as AIDS, but they have not unveiled it.

All such products of scalar electromagnetics (which the Soviets call energetics) are held tightly by the KGB.

That the Soviets have kept such a cure secret strongly implies that they regard it as a known antidote for biological warfare agents, and thus highly classified.

The importance of this can easily be seen: Presently the West has absolutely zero defense against a biological warfare strike using viruses as BW agents.

In such a strike, fast-acting viral strains can easily be used — say, agents which are highly infectious and kill within two or three days.

Development of a vaccine is impossible. There is simply no time to do that. Further, even if the virus was slow-acting, a vaccine could only be developed if the virus is stable. If it is unstable (constantly mutating), development of a vaccine is impossible.

The greatest vulnerability of the U.S. today is its total vulnerability to unknown BW viral agents. Absolutely nothing we have or are doing in our Department of Defense is of any consequence or protection against new BW viruses for which no vaccine is developed or possible.

The great ease with which such cheap new, highly lethal viral BW agents can be developed means that even a relatively small military power—such as Libya, Cuba, or Iran—can develop and use them against the U.S.

Such BW viral agents can also easily be clandestinely planted across the U.S., to devastate an unsuspecting, unprotected American populace.

In fact, even a single deranged madman can readily develop lethal BW viral agents sufficient to devastate even a great military power. Further, in an open society such as the U.S., that same madman can easily introduce such an agent he develops — and, if he's clever, not even get caught in the act.

Past isolated cases of "contamination" of on-the-shelf drugstore medicines is just a faint prelude of that which is to come.

Indeed, we may unwittingly do it to ourselves by accident, even without a BW strike by the Soviets, small enemy powers, or a madman.

Throughout the West, it is now most fashionable to experiment with the production of new bacteria and viruses. Gene-splicing, rccombinant DNA genetics, etc. are being employed on a grand scale as the latest "scientific marvel."

Even some high school students have experimented with this sort of thing. Many other colleagues, universities, and private persons are proceeding full-tilt with casual experiments.

Most of this "casual experimentation" is uncontrolled. When finished, the residue gets dumped in the trash can or flushed down the drain. In any case, it goes into our biosphere. Some of this residue contains unstable, highly adaptable viruses.

It's only a matter of time until one of these unwitting agents gets loose and survives — one that's lethal, fast, and highly infectious.

Make no mistake: It's a *near certainty* that this will happen. The only question is *when.*

Also, with the recent Supreme Court decision allowing the patenting of a virus or bacteria, a number of laboratories and companies are now feverishly developing new agents to unleash in the biosphere upon persons, animals, and plants.

For the most part, these experiments are tightly controlled. However, it is impossible to test in the laboratory the result of the natural recombination and mutation of these agents once they are loose in the environment and in a free exchange with all the natural bacteria and viruses.

Once loosed in the biosphere, a new viral agent undergoes combinations and recombinations with the greater number of viruses already present. The situation is like an increase in a large factorial number: Only a few additional viruses will result in millions and millions of additional combinations — at least some of which inevitably will be lethal and quick-acting. This is *not* tested in those laboratories making, patenting, and releasing new viruses.

Humans have spent millions of years developing sufficient immunity to the natural BW agents to allow humanity to survive.

Very shortly we are going to have increasingly large numbers of new environmental BW agents emerging all around us. For some of them, it is a foregone conclusion that we have little or no natural immunity developed.

So we are going to enter an increasingly harsh environment, characterized by the rise of great epidemics and pandemics of many kinds.

Inevitably some of these coming scourges are going to rival the "Black Death" in ferocity and lethality.

Effective defense against one or more **such BW viral attacks — either against our forces in the field or against our civilian population — can only be provided by a ready, fully developed technology of electromagnetic healing.**

It does not really matter whether the BW strikes are performed by a hostile foe, a madman, natural mutations within our now drastically stimulated viral environment, or as a result of some particular "residue" discarded by an enthusiastic high school or college student.

The end result — the decimation of humankind — is going to be the same unless we develop and use electromagnetic healing technology.

However, with the present enormously self-hostile state of mankind, one or more enemies or madmen will almost certainly launch a strong, clandestine BW strike against the United States Armed Forces and the U.S. populace.

With EM healing technology, a specific "antidote" to any viral disease — known or unknown **— can be developed in a matter of hours.**

Mass electromagnetic treatment can then be started immediately, even for an entire army in the field or for our population in the heartland. Within hours of recognizing a BW strike, cure rates above 90% can be quickly achieved in exposed personnel.

Further, using electromagnetic healing technology, if necessary a more conventional vaccine can also be rapidly developed — say, in two weeks — for mass distribution and inoculation of the civilian populace.

The bottom line is this: Neither the U.S. Armed Forces nor the U.S. civilian populace can or will survive the coming lethal viral warfare strikes unless we quickly develop electromagnetic healing technology and have the necessary diagnostic and treatment devices already ready and waiting.

The clandestine AIDS strike by the Soviet Union is just the precursor of a great deal more to come.

The USSR Has Repeatedly Verified Our Nonreaction

The Soviet Union already knows that Western science and governments are too stereotyped and dogmatic to react to anything highly unorthodox.

For years, the Soviet have been proving this, over and over, in their development and testing of frightful scalar electromagnetic superweapons of unprecedented power. In the face of overwhelming evidence of this, the West has kept its head buried in the sand like an ostrich.

It wasn't invented in the West. Therefore it cannot be.

Years of substantial Communist indoctrination of students in our universities, by leftist/liberal professors, has placed a massive imprint into our society. An entourage of sympathetic groups has resulted — groups which believe that the Soviet leaders are just like us, that the Stalinist era is long since over, and that the Communist leopard has changed its spots and really isn't trying to take over the world anymore.

As a system, then, much of the West simply refuses to accept anything unorthodox, anything "not invented here", and anything that indicates Soviet aggression.

The Soviets know the U.S. is totally ignorant of advanced scalar EM weaponry. They have proven over and over that the U.S. will not even comprehend that it has been deliberately struck, if the strike is sufficiently unorthodox.

The Soviet scalar EM superweapons have already been documented in **Fer-de-Lance,** Tesla Book Company, 1986, and that documentation will not be repeated here.

Only a short summary of events will be presented, to put the Soviets' confidence into perspective. A longer chronology of selected events is included as Appendix IV.

31 Soviets Knew The U.S. Wouldn't React

The Soviets were quite confident in arranging for a great biological warfare strike, because of the total lack of the U.S. to comprehend incidents such as the following:

1. Beginning in the late 1950's, hidden Soviet transmitters across the street targeted "weak microwave radiation" against the U.S. Embassy in Moscow. Over the years the radiation continued, sometimes continuously, sometimes intermittently. The Soviets electronically induced diseases that resulted in extensive health problems of Embassy personnel, the death of two U.S. Ambassadors, and the sickness of a third. Several U.S. Presidents have protested the radiation. Both U.S. Intelligence and the U.S. scientific community have been totally baffled. They still do not know the true nature of the radiation, nor its purpose.

2. In the winter of 1957-58, an unexplained nuclear event happened near Kyshtym, within the Soviet Union. A huge explosion occurred in the stored nuclear wastes in the disposal section of the Soviet atomic weapons industry located in the southern Urals. Radioactive contamination covered 1,000 square miles. For years the story was opposed — or suppressed — by U.S. intelligence agencies, perhaps to allay public resistance to the U.S. nuclear industry. At the time, Soviet experimentation with large scalar electromagnetic transmitters was well underway. A large transmitter probably failed, discharging its built-up, giant electrogravitational potential (EGP) into the earth. Such a pulse would have propagated through the earth as a wave; the first nuclear material contacted would have been fissioned immediately. The stored nuclear wastes probably exploded due to receipt of an EGP from the inadvertent failure of one of the Soviet Union's new superweapons in development. Survivors reported that indeed the nearby nuclear wastes exploded.

3. In January 1960 Khrushchev announced to the Presidium that a new fantastic weapon was in development. He stated that the weapon was so powerful that, if unrestrainedly used, it could wipe out all life on earth.

Figure 4. The U.S. Embassy in Moscow. The Soviet Union has bombarded this **facility** with weak microwave radiation for decades.

Table 2. MICROWAVE RADIATION **OF THE U.S. EMBASSY** IN MOSCOW

- **SINCE 1959 OR 1960**
- **HIGH LEVEL TARGET (U.S. AMBASSADOR)**
- **GUARANTEES PERSONAL ATTENTION OF**
  - **PRESIDENT**
  - **NSA, CIA, DIA**
  - **TOP CONSULTING SCIENTISTS**
  - **STATE DEPARTMENT**
  - **LEADING SCIENTIFIC INSTITUTIONS**
  - **NATIONAL SECURITY COUNCIL**
  - **ETC.**
- **TWO AMBASSADORS DIED, ANOTHER SICKENED**
- **U.S. REACTION REVEALS ANY KNOWLEDGE OF TESLA ELECTROMAGNETICS**
- **FOUR U.S. PRESIDENTS REQUESTED SOVIETS TO CEASE**
  - **CUT FROM 18 WATTS/CM^2 TO 2**
  - **THEN AGAIN INCREASED**

ORDINARY DECAY

- **EGP MUST BE LARGE ENOUGH**
- **EGP MUST BE LONG ENOUGH**

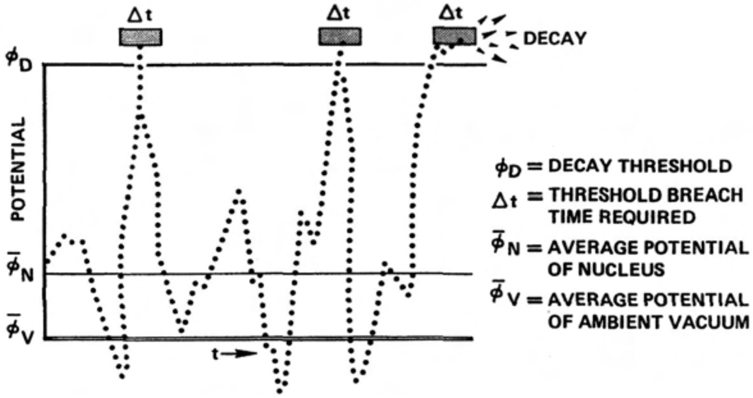
Δt
Δt
Δt
DECAY
φD
POTENTIAL
φ̄N
φ̄V
t→
φD = DECAY THRESHOLD
Δt = THRESHOLD BREACH TIME REQUIRED
φ̄N = AVERAGE POTENTIAL OF NUCLEUS
φ̄V = AVERAGE POTENTIAL OF AMBIENT VACUUM

Figure 5. Decay of a radioactive nucleus. When it absorbs a scalar EM wave (electrogravitational pulse) that is long enough and large enough, the nucleus fissions.

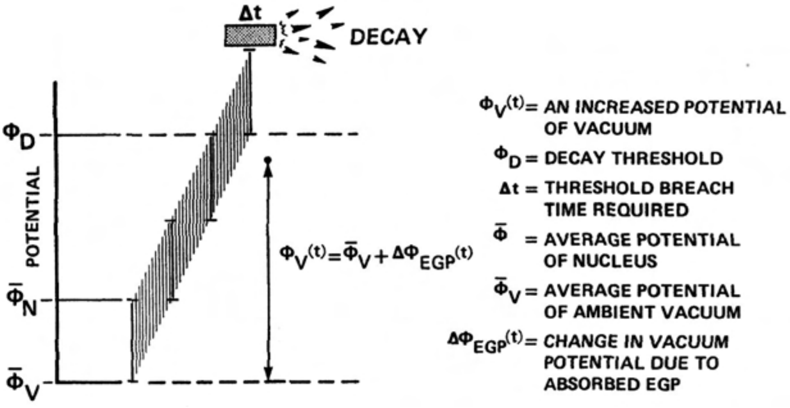
Δt
DECAY
ΦD
POTENTIAL
Φ̄N
Φ̄V
ΦV(t) = Φ̄V + ΔΦEGP(t)
ΦV(t) = AN INCREASED POTENTIAL OF VACUUM
ΦD = DECAY THRESHOLD
Δt = THRESHOLD BREACH TIME REQUIRED
Φ̄ = AVERAGE POTENTIAL OF NUCLEUS
Φ̄V = AVERAGE POTENTIAL OF AMBIENT VACUUM
ΔΦEGP(t) = CHANGE IN VACUUM POTENTIAL DUE TO ABSORBED EGP

Figure 6. EGP fission of a radioactive nucleus. If an initial EGP is large enough and long enough, it will fission any and all radioactive nuclei it strikes.

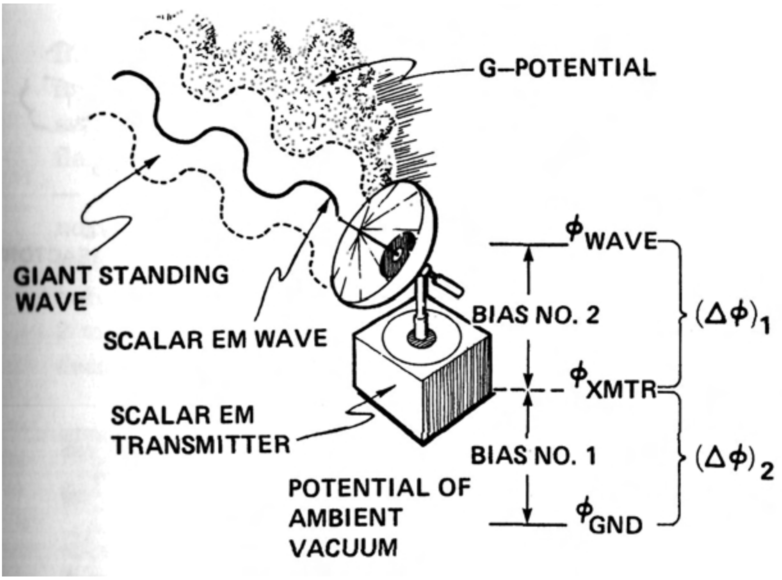
G–POTENTIAL
GIANT STANDING WAVE
SCALAR EM WAVE
SCALAR EM TRANSMITTER
POTENTIAL OF AMBIENT VACUUM
ϕ_{WAVE}
BIAS NO. 2
ϕ_{XMTR}
BIAS NO. 1
ϕ_{GND}
$(\Delta\phi)_1$
$(\Delta\phi)_2$

Figure 7. Bottled up energy in a standing scalar EM wave is a potential.

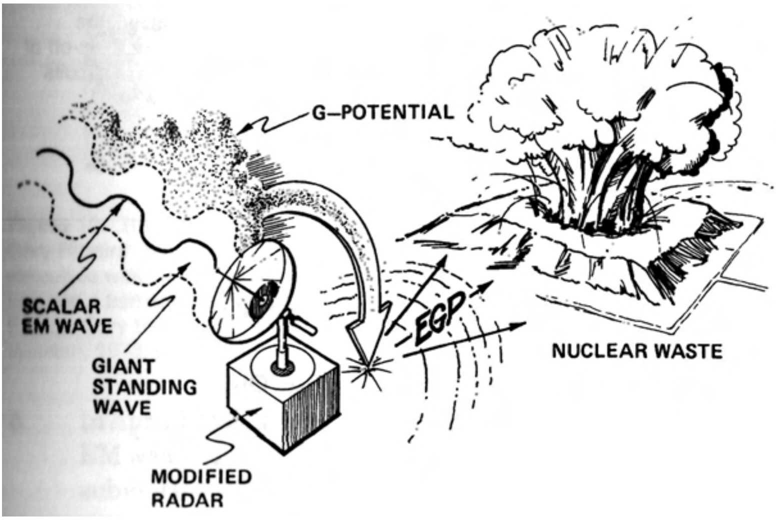
G–POTENTIAL
SCALAR EM WAVE
GIANT STANDING WAVE
MODIFIED RADAR
EGP
NUCLEAR WASTE

Figure 8. EGP short-out Transmitter failure and consequent discharge of a giant standing scalar EM wave potential caused the explosion of stored nuclear wastes at Kyshtym, USSR, in the winter of 1957-58.

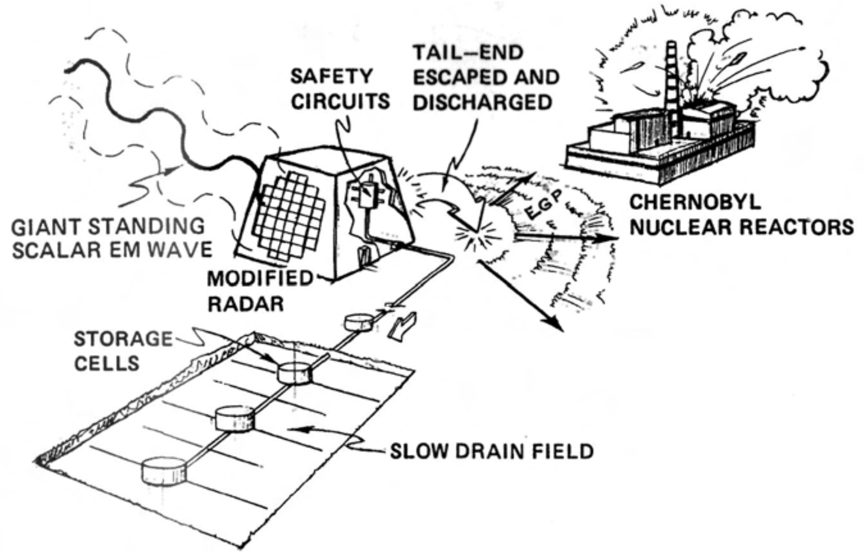
TAIL–END
ESCAPED AND
DISCHARGED
SAFETY
CIRCUITS
CHERNOBYL
NUCLEAR REACTORS
EGP
GIANT STANDING
SCALAR EM WAVE
MODIFIED
RADAR
STORAGE
CELLS
SLOW DRAIN FIELD

Figure 9. EGP initiation of the Chernobyl event. Failure of another giant transmitter near Chernobyl in 1986 eventually caused the eruption of one of the Chernobyl reactors by EGP-pulsing its nuclear fuel rods. When the transmitter failed, the alarmed Soviets shut down the four Chernobyl reactors while waiting for drain-off of the EGP potential by the safety circuits of the disabled transmitter. These circuits finally gave way, producing a mild EGP pulse into the earth. Had they failed immediately, all four reactors at Chernobyl would have violently exploded.

Table 3. KHRUSHCHEV'S 1960 STATEMENT

(Speaking to the Presidium)

**WE HAVE A NEW WEAPON,
JUST WITHIN THE PORTFOLIO OF
OUR SCIENTISTS...
SO POWERFUL THAT,
IF UNRESTRAINEDLY USED,
*IT COULD WIPE OUT
ALL LIFE ON EARTH!***

4. In May 1960 Gary Powers' high-flying U-2 spy plane was mysteriously disabled over the Soviet Union, resulting in its subsequent crash and the Soviet capture of Powers himself. A flash appeared behind Powers' aircraft, as if from an explosion — except that *the flash persisted,* showing that it was not a normal HE explosion at all. A tracking radar, employed in a scalar EM mode, probably was used to cause the electromagnetic "flash ball" and disable Powers' aircraft. Harmless SA-2 missiles fired far below him may have been employed to deceive Western intelligence analysts.

Figure 10. The U-2 high altitude reconnaissance aircraft. On May 2,1960 Francis Gary Powers high flying U-2 spyplane was disabled over the Soviet Union. Some difficulties with the autopilot were first experienced. Then a single dull thump was felt, and a tremendous orange flash lit the cockpit and sky — *and persisted.* (See Francis Gary Powers with Curt Gentry, **Operation Overflight,** Holt, Rhinehart and Winston, 1970, p. 82.)

5. In April 1963 the Soviet Union utilized a newly deployed scalar EM weapon to destroy the submerged U.S.S. Thresher nuclear submarine, off the coast of the U.S. itself. In spite of intense and totally anomalous electronic jamming experienced by the U.S.S. Skylark, surface companion to the Thresher, the U.S. failed to recognize this major indication of a scalar EM attack. *One day later, a Soviet scalar EM weapon placed a giant*

underwater burst in the ocean, 100 miles north of Puerto Rico. With these two tests Khrushchev, still smarting from his Cuban Missile Crisis facedown by John Kennedy, avenged himself and convinced the Presidium to allow him to remain in power for yet awhile.

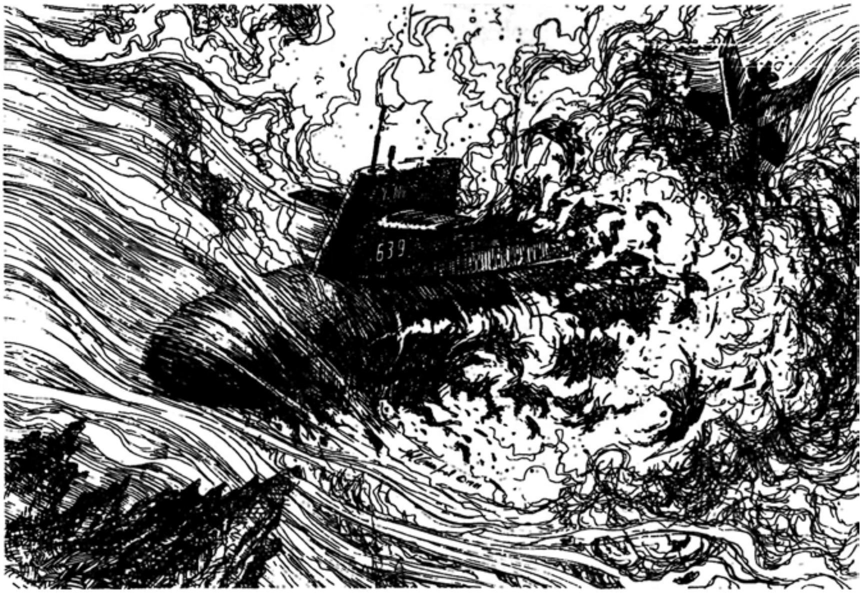

Figure 11. The U.S.S. Thresher disintegrates. © Hal Crawford 1979

Table 4. DEATH OF THE U.S.S. THRESHER

APRIL 10, 1963

- **DURING TRIALS OFF EAST COAST OF U.S. IN ATLANTIC**
- **SURFACE COMPANION: U.S.S. SKYLARK**
- **UNDERWATER PHONE BETWEEN THE TWO**
- **THRESHER DIVED AND WAS LOST**
    - **GARBLED MESSAGE**
    - **DID NOT RESPOND TO CONTROLS**
    - **SANK TO CRUSH DEPTH**
    - **IMPLODED**
- **SKYLARK ENCOUNTERED SEVERE ELECTRONIC INTERFERENCE**

Table 5. SEVERE EM INTERFERENCE EXPERIENCED BY U.S.S. SKYLARK

APRIL 10,1963

- **SKYLARK JAMMED (1.5 HR TO TRANSMIT MESSAGE)**
- **MULTIPLE SYSTEMS**
- **SIGNATURE:**
    - **SCALAR INFEROMETRY**
    - **SKYLARK IN "SPLATTER" ZONE**
    - **THRESHER PROBABLE TARGET**
- **OTHER SYMPTOMS**
    - **ONE SYSTEM KNOCKED OUT**
    - **OTHERS THEN WENT OUT**
    - **LATER RESUMED WORKING**

6. In Afghanistan, in 1981 and at intervals since then, Soviet helicopters have sometimes attacked Afghan guerrillas or villages with a strange new weapon. A body struck by this weapon dies instantly and completely. It falls like a limp rag, **and** does not decay even in 30 days. Every cell, bacterium, virus, etc. in the body has been killed instantly. Nerve gas rockets and chemical rockets are also usually fired to provide **a** "deception cover" for the nature of the real weapon. This eery Soviet weapon has been mislabeled as some strong new gas, **and** is referred to as "Smirch", a combination of the two words "smert" (death) and "semerch" (sandstorm).

7. In April 1984 a monstrously huge scalar EM incident occurred off the coast of Japan, near the Kuriles. Subsequent interviews with the pilots sighting the incident have revealed that several combined tests of the various modes of a scalar EM howitzer occurred. A giant "cold explosion" occurred, raising a giant mushroom cloud to over 60,000 feet and greater than 150 miles diameter in two minutes. A glowing dome appeared, and expanded to monstrous size. According to Dr. Walker and colleagues, **the diameter of this great** glowing, **hem-**

ispherical **shell expanded to several hundred kilometers.** * He and his colleagues were able to rule out any possible known natural events or mechanisms that could have caused the incident.

8. Numerous incidents of hemispherical shells of light and giant balls of light within the Soviet Union have been seen from Iran and from Afghanistan. Such giant artificial phenomena have also repeatedly been seen over the oceans by mariners and airline pilots.

9. Several instances of anomalous disabling of jet engines of commercial airliners have occurred while the aircraft were in flight.

10. Direct and intensive weather engineering over the U.S. itself has been accomplished by the Soviet Union for a decade, using the interference grid resulting from the crossing of two or more giant beams from the woodpecker "over-the-horizon radar" transmitters. Specific cloud signatures (anomalous holes, giant radials, twin giant radials, and razor-sharp giant interference patterns) have continually occurred, but the orthodox system has blithely ignored them.

11. A giant test of over a hundred huge Soviet scalar EM weapons occurred in and around May 1, 1985 for the 40th anniversary of the end of WWII. The entire Western scientific and intelligence establishment remained sublimely unaware of the giant exercise. They had not a single scalar EM signal detector — even though the air, land, and ocean was filled with hundreds of Soviet scalar EM transmitters, chattering away to Soviet armed forces, submarines under the water, etc. Underneath our feet, the earth itself was in giant scalar EM resonance on 54 frequencies — 27 pairs, each of which tapped energy directly from the molten core of the earth and fed it to an ensemble of giant Soviet strategic scalar EM weapons.

---

*See Science, 227 (4584), Feb. 8, 1985, p. 607-611; Science 234 (4775), Oct. 24, 1986, p. 412-413.

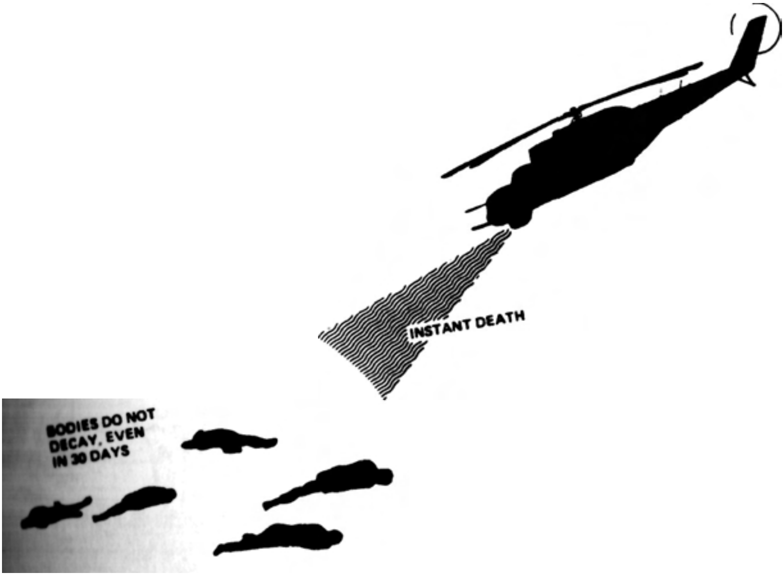
INSTANT DEATH
BODIES DO NOT DECAY, EVEN IN 30 DAYS

Figure 12. Soviet helicopters have used a "death ray" in Afghanistan.

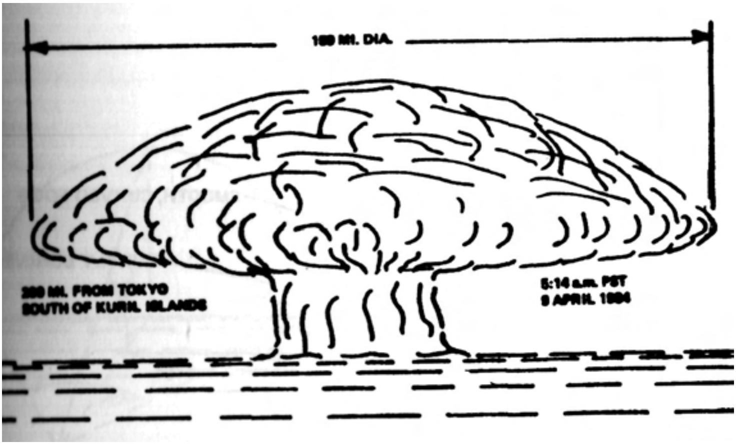
160 MI. DIA.
200 MI. FROM TOKYO
SOUTH OF KURIL ISLANDS
5:14 a.m. PST
9 APRIL 1984

Figure 13. Mushroom cloud from the sea off Japan, near a Russian test area.

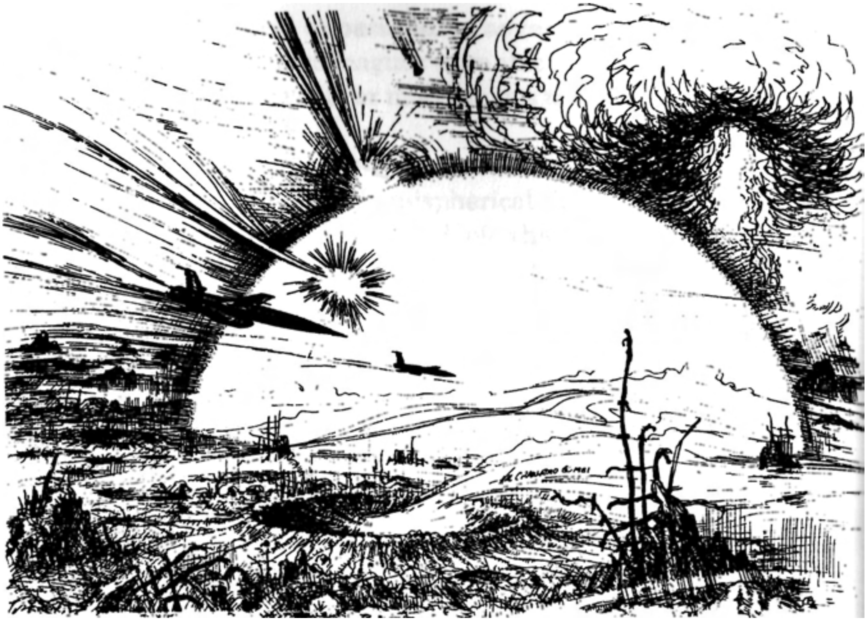

©Hal Crawford 1981

Figure 14. The Tesla shield, a giant hemispherical shell of energy. Seen repeatedly in Soviet tests in out-of-the-way places and remote regions of the ocean.

8:15 PM
THURSDAY
AUG 1969
VIRGIN ISLANDS
SEEN BY MANY RESIDENTS

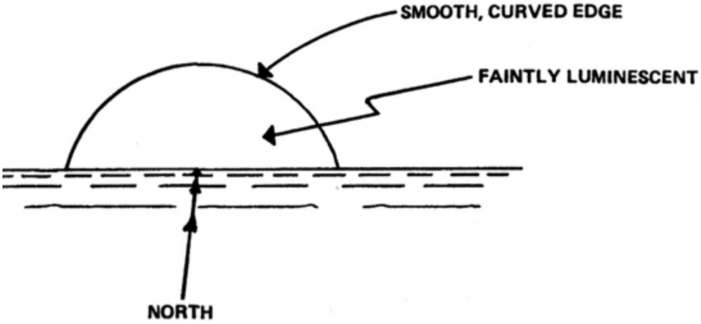
SMOOTH, CURVED EDGE
FAINTLY LUMINESCENT
NORTH

Figure 15. The Tesla shield: A 1969 Virgin Islands incident.

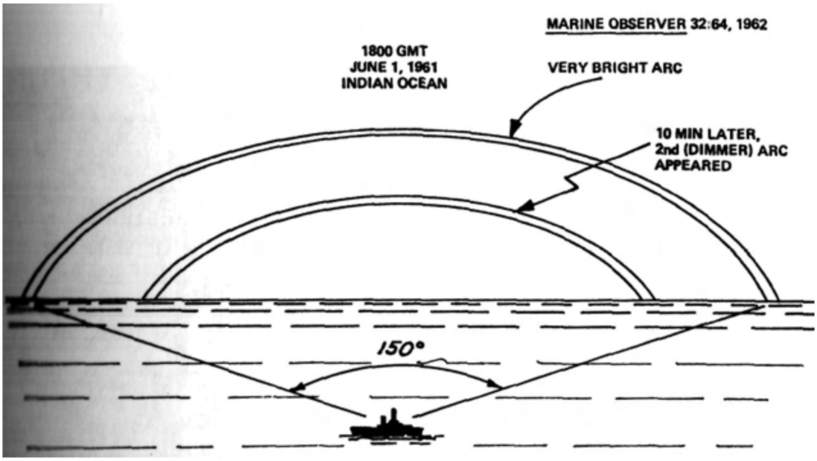
MARINE OBSERVER 32:64, 1962
1800 GMT
JUNE 1, 1961
INDIAN OCEAN
VERY BRIGHT ARC
10 MIN LATER,
2nd (DIMMER) ARC
APPEARED
150°

Figure 16. Tesla shields: Two nested arcs of light in the sky.

MAR 20, 1968
AND WESTERN
NORTH ATLANTIC

MARINE OBSERVER 40:17.1970

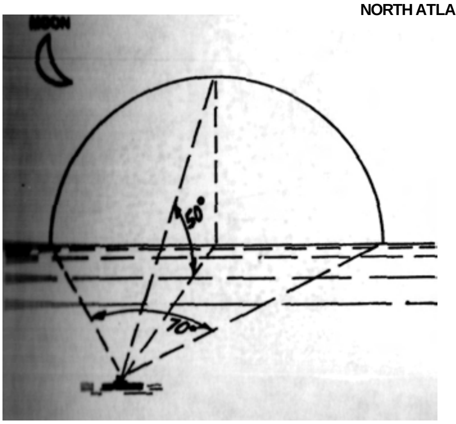

- **SEMICIRCLE**
- **MILKY WHITE LIGHT**
- **QUITE BRIGHT AT FIRST**
- **DIMMED AS ENLARGED**
- **LASTED 10 MINUTES**

Figure 17. Tesla shield: An expanding hemispherical circle of light.

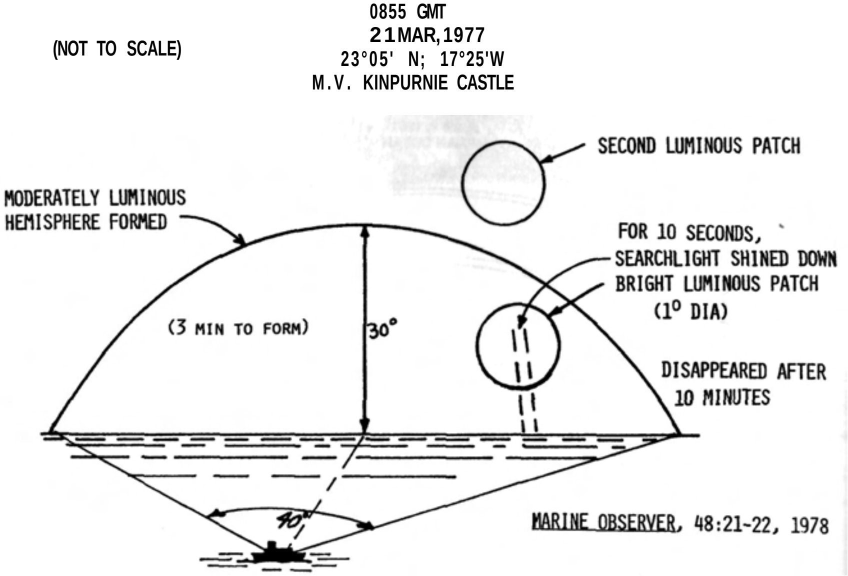
0855 GMT
21 MAR, 1977
23°05' N; 17°25'W
M.V. KINPURNIE CASTLE
(NOT TO SCALE)
SECOND LUMINOUS PATCH
MODERATELY LUMINOUS HEMISPHERE FORMED
FOR 10 SECONDS, SEARCHLIGHT SHINED DOWN
BRIGHT LUMINOUS PATCH (1° DIA)
(3 MIN TO FORM)
30°
DISAPPEARED AFTER 10 MINUTES
90°
MARINE OBSERVER, 48:21-22, 1978

Figure 18. Tesla shield: Hemisphere and globes, 1977.

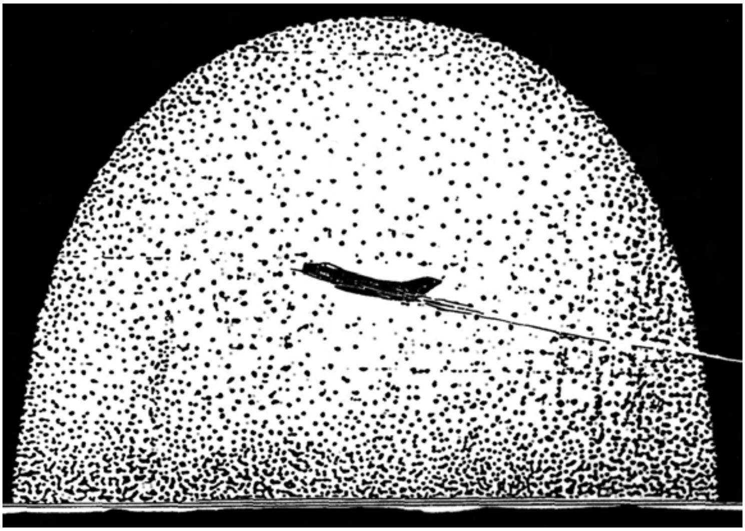

Figure 19. Large Tesia Shield observed by Red Chinese fighter pilots. Electrical difficulties in the aircraft were also experienced.

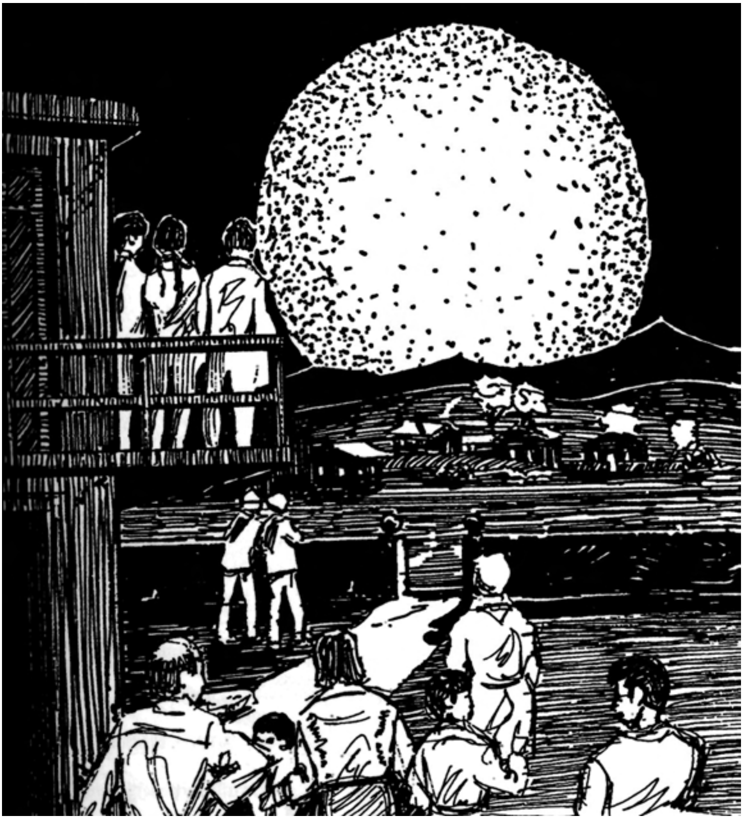

Figure 20. Large glowing Tesla Globe," Red China. Observed by hundreds of **persons.**

45 *Soviets Knew The U.S. Wouldn't React*

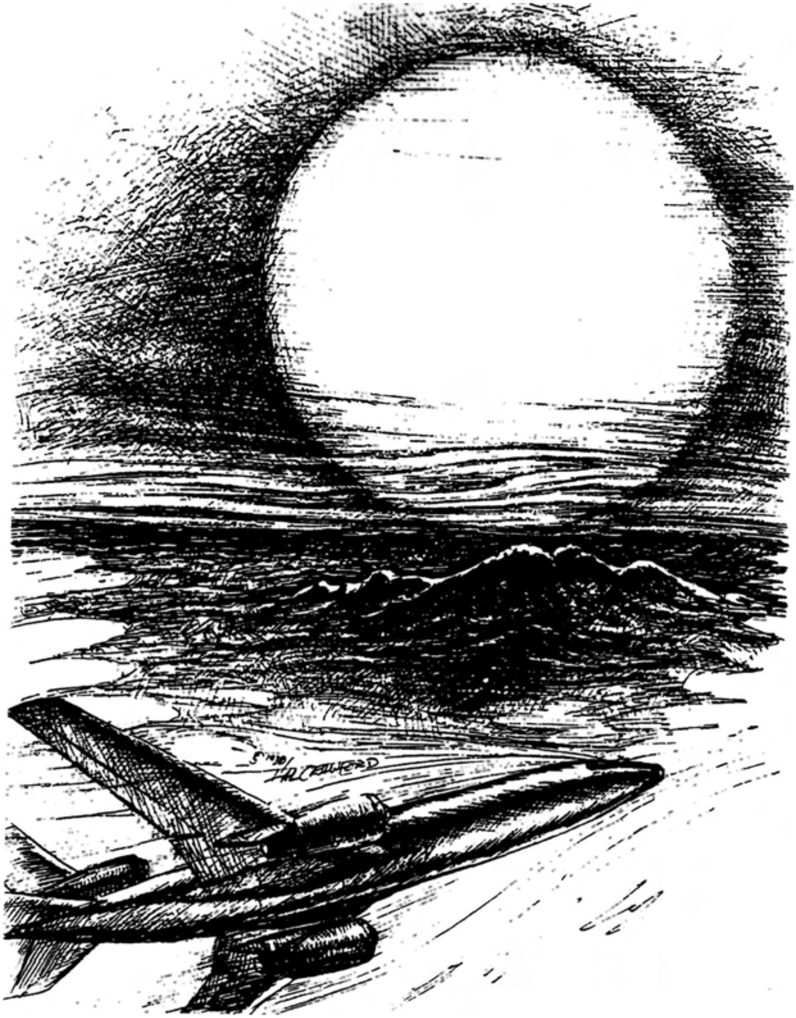

©1980 Hal Crawford

Figure 21. 1966 test of giant Tesla Globe, expanding deep within the Soviet Union. Seen from Teheran, Iran June 17, 1966 by two aircraft approaching Mehrabad Airport. Observed for several minutes.

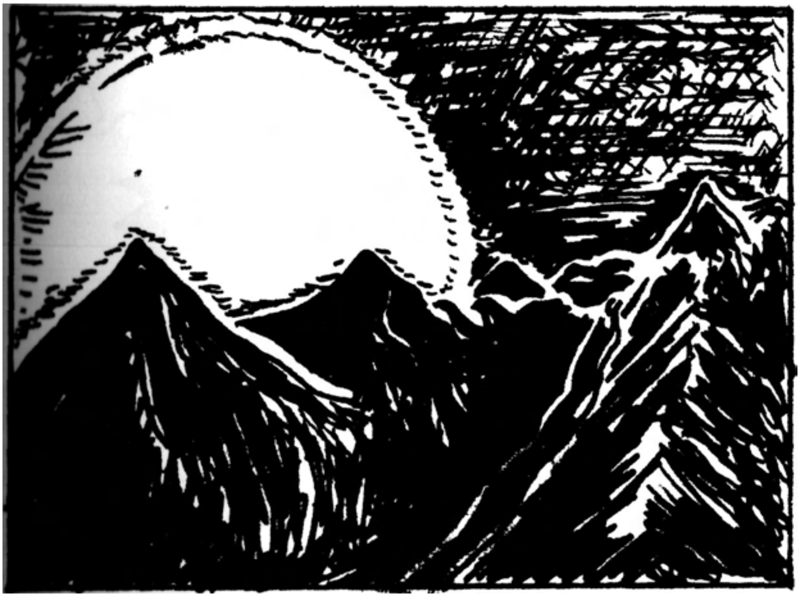

Multiple events seen in Sept. 1979 from Afghanistan

Strange lurid glow that flared silently over the Hindu Kush; as described by
Nick Downie.
The LONDON SUNDAY TIMES, 17 August 1980

Figure 22. A giant globe, deep within the Soviet Union in 1979. Seen several times frorn Afghanistan, looking toward Saryshagan.

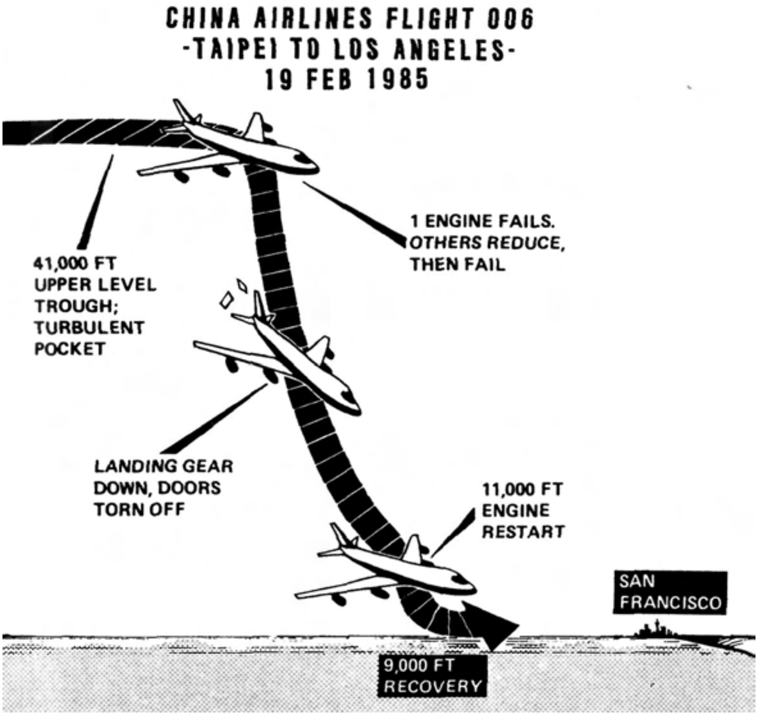
CHINA AIRLINES FLIGHT 006
-TAIPEI TO LOS ANGELES-
19 FEB 1985
41,000 FT
UPPER LEVEL
TROUGH;
TURBULENT
POCKET
1 ENGINE FAILS.
OTHERS REDUCE,
THEN FAIL
LANDING GEAR
DOWN, DOORS
TORN OFF
11,000 FT
ENGINE
RESTART
SAN
FRANCISCO
9,000 FT
RECOVERY

Figure 23. Engines disabled, a passenger jet drops six miles in two minutes.

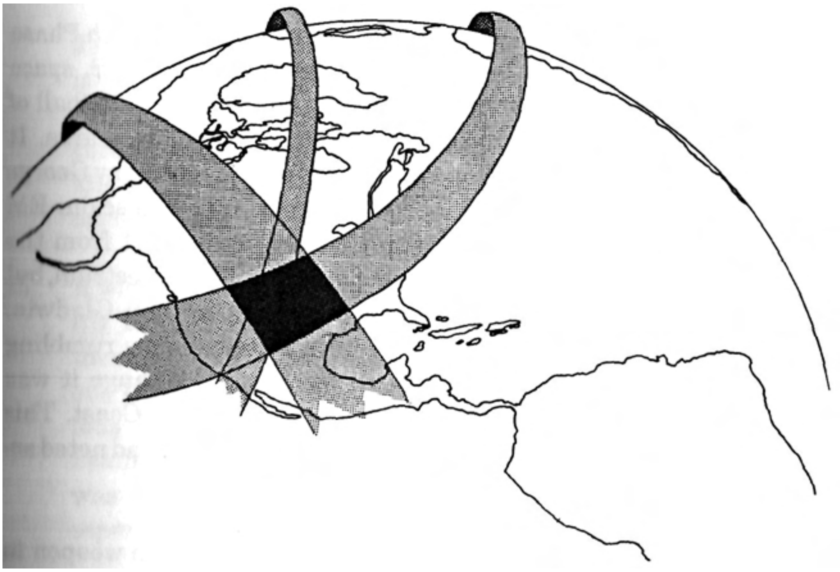

Figure 24. Giant "Woodpecker" beams intersecting over the U.S. Used for weather control, as a launch phase ABM and anti-bomber system, electromagnetic attack of distant populations, and disease conditioning. Using phase conjugate radar mirrors, any amount of energy desired can be delivered to any distant target from which a radar return has been received. Real-time distant holography, to produce powerful energy forms at a distance, can also be accomplished.

Table 6. MASSIVE 1985 STRATEGIC MAY DAY EXERCISE

APRIL/MAY 1985

- **BREZHNEV'S 1972 STATEMENT FULFILLED**
- **40th ANNIVERSARY OF WWII**
- **GORBACHEV NEWLY IN POWER**
- **EARTH IN GIANT SCALAR EM RESONANCE**
  - **54 FREQUENCIES**
  - **27 "POWER TAP" PAIRS**
  - **4 TO 6 WEAPONS EACH PAIR**

- **C^3I SCALAR EM LINKS**
  - **UNDERWATER SUBS**
  - **NAVAL FORCES**
  - **OTHER FORCES**

12. In late November, 1985 the Soviets tested their Launch Phase ABM system over Florida, using the launch of a U.S. space shuttle as a convenient test target. A marker beacon (ball of light) was placed in the sky, hovering near the launch area. It was seen by hundreds of persons, and photographed by George Suchary. During the final countdown to launch, a scalar EM "missile" (ball of intense energy) was fired, offset from the shuttle. This ball of energy was not in the visible spectrum, but may have been accidently photographed by Bob Gladwin. Twelve minutes offset after launch, a tremendous rumbling overhead boom echoed through the area — so huge it was heard for hundreds of miles up and down the East Coast. This was the third U.S. shuttle launch in which NASA had noted associated anomalous phenomena, but did nothing.

13. On December 12, 1985 the Soviets used the same weapon in the anti-bomber mode to destroy an Arrow DC-8 jet just after its takeoff from Gander Air Force Base, Newfoundland. **Over 250 U.S. soldiers and civilian crew members were killed.** Gander has long been a registration point for the Soviet scalar EM weapons. For example, in one 4-day period in 1982, 36 anomalous incidents occurred in and around Gander, involving moving lights, lights in the sky, etc.

14. On January 28, 1986 the Soviets destroyed the U.S. space shuttle Challenger, shortly after its launch. A dozen signatures are associated with this Soviet kill, too many to mention all of them. To give only a few: First, the Soviet ships monitoring the launch site suddenly departed at high speed, four hours before the launch. This was unparalleled. Second, according to General Daniel Graham, *the following evening KGB headquarters gave a party to celebrate the success of the KGB's perfect active measures against the Challenger!* Direct and unmistakable indications of extreme localization of the Soviet transmitters to the Challenger's launch site existed on the day of the launch. In addition, a metal-softening signal was confirmed on the Soviet Woodpecker beams as early as

January 1,1986. Many other indicators also exist. The total is a very strong statement that the Soviets destroyed the Challenger with the same weapon they had tested against three previous shuttle launches, and which they had used to kill the Arrow DC-8 at Gander Air Force Base only six weeks earlier.

15. A whole series of anomalous missile failures occurred, involving Titan, Delta, and Ariane launches. The odds against all these launches failing in that pattern were millions to one. In one Titan failure, video taken of the event showed the anomalous "marker beacon" light in the vicinity, high overhead. Frame-by-frame analysis of the video showed that the light was indeed independent of the explosion, and it moved off separately after the rocket exploded.

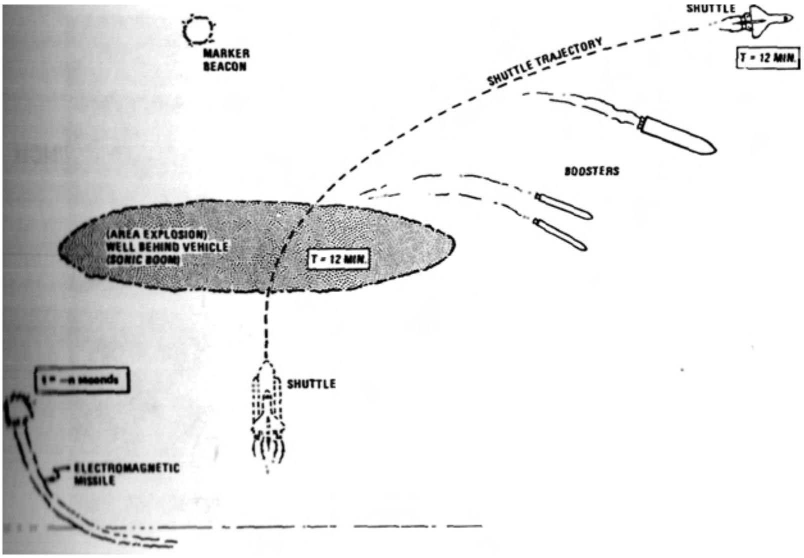
SHUTTLE
MARKER BEACON
SHUTTLE TRAJECTORY
T = 12 MIN.
BOOSTERS
(AREA EXPLOSION) WELL BEHIND VEHICLE (SONIC BOOM)
T = 12 MIN.
T = –n seconds
SHUTTLE
ELECTROMAGNETIC MISSILE

Figure 25. Third Soviet non destructive testing against shuttle launches, Nov. 26, 1985.

Table 7. SHUTTLE **LAUNCH ANOMALIES, NOV. 26, 1985**

- **MARKER BEACON**
    - **PHOTOGRAPHED**
    - **SEEN BY HUNDREDS**

- **OFFSET SCALAR EM MISSILE**
    - **INVISIBLE TO HUMAN EYES**
    - **STRIKE PHOTOGRAPHED**

- **GIANT ATMOSPHERIC BOOM**
    - **T + 12 MINUTES**
    - **HEARD FOR HUNDREDS OF MILES**

- **THIRD LAUNCH WITH SUCH A BOOM**

Table 8. SOVIET ESCALATION AFTER 3RD TEST AGAINST SHUTTLE LAUNCH on Nov. 26,1985.

- **U.S. STILL UNAWARE OF SCALAR EM WEAPONS USE**
- **SOVIETS ESCALATED ACTION**
- **DESTROYED OR DISABLED:**
    - **ARROW DC-8 (DEC. 12,1985)**
    - **CHALLENGER (JAN. 28, 1986)**
    - **TITAN 34-D (APRIL 18, 1986)**
    - **DELTA ROCKET (MAY 3, 1986)**

- **MAY HAVE DISABLED:**
    - **ARIANE II(MAY 31,1986)**
    - **TWO U.S. NUCLEAR SUBS WHICH RAN AGROUND**
    - **STEALTH FIGHTER WHICH CRASHED**

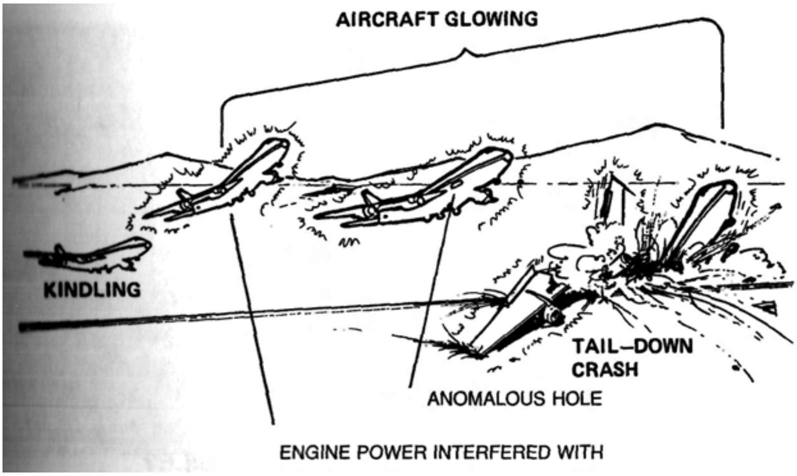
AIRCRAFT GLOWING
KINDLING
TAIL–DOWN CRASH
ANOMALOUS HOLE
ENGINE POWER INTERFERED WITH

Figure 26. Death of the Arrow DC-8, Dec. 12,1985. Three eye-witnesses saw the **doomed** airliner glowing. An anomalous hole was found in the fuselage, forward of **the** engines. Engine combustion was interfered with, by phase conjugate (negative) e**nergy.** Engines were rumbling laborously, not roaring with power, as the aircraft **passed** 100 ft directly over a witness.

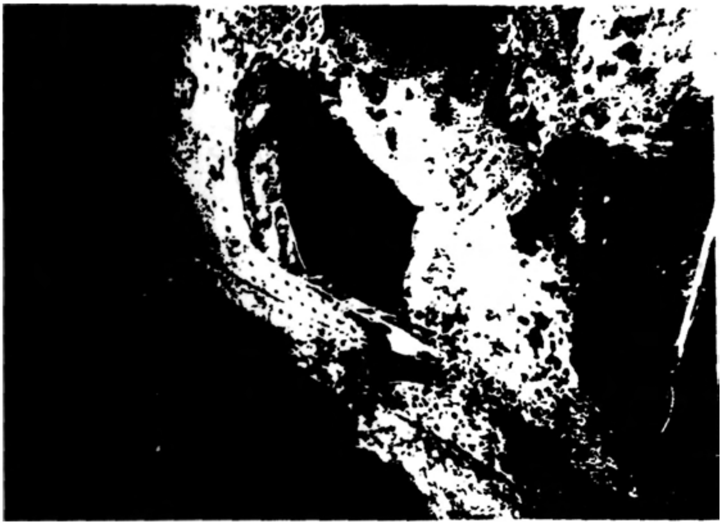

Figure 27. Anomalous hole in the right fuselage of the ARROW DC-8, ahead of the engines. With crossed over-the-horizon radar beams and multiple phase conjugate **radar** mirrors, real-time holography allows the production of a ball of EM energy at a great distance. The ball may be as small as desired, and contain as much energy as desired.

Table 9. ARROW DC-8 DISASTER, DEC. 12, 1985

GANDER AFB, NEWFOUNDLAND

- **AIRCRAFT LIFTED OFF**
- **AIRCRAFT WAS GLOWING YELLOW - ORANGE**
- **SANK TAIL-DOWN AND CRASHED**
- **SIGNATURE OF ENGINE/COMBUSTION INTERFERENCE**
  - **LAUNCH PHASE WEAPON**
  - **ANTI-BOMBER MODE**
  - **ANOMALOUS HOLE IN FUSELAGE, FORWARD OF ENGINES**
  - **STRUCK BY ELECTROMAGNETIC MISSILE?**
  - **EXTENSIVE "BALLS OF LIGHT" WEAPON TESTS IN AND AROUND GANDER**
- **CLOSELY FOLLOWED THIRD TEST AGAINST SHUTTLE NOV. 26,1985, IN ABM MODE**

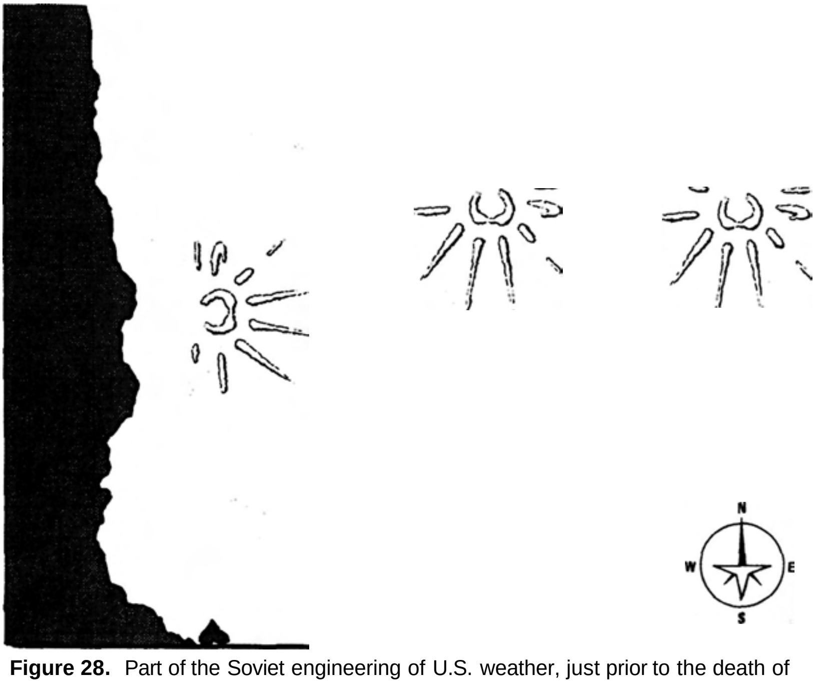

Figure 28. Part of the Soviet engineering of U.S. weather, just prior to the death of the Challenger. Three "giant radial" clouds sighted over the greater Los Angeles area by Margaret Wilson. Appeared one after the other, at 10 min. intervals.

Soviets Knew The U.S. Wouldn't React

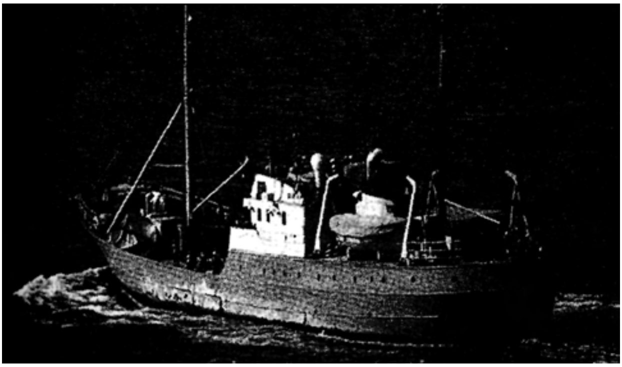

DOD PHOTO

Figure 29. Soviet Trawler spyship. About four hours before the launch of the Challenger, all Soviet ships off coast, that normally monitor each launch, suddenly and mysteriously departed at high speed. Unprecedented event.

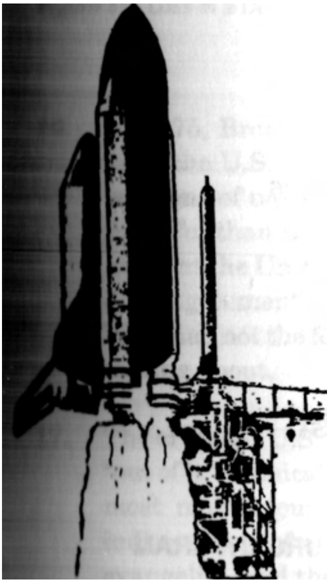

- **SOVIET WEATHER ENGINEERING**
- **SEVERE COLD STRESS**
- **METAL-SOFTENING**
- **BOOSTER SEAL PROBLEM**
- **SOVIET SHIPS ABSENT**
- **LOCALIZATION**
- **AIR TURBULENCE**
- **BIRDS NOT FLYING**

Figure 30. Launch of the doomed Challenger, Jan. 28,1986.

Table 10. DEATH OF THE CHALLENGER, JAN 28, 1986

- **METAL-SOFTENING SIGNAL ON WOODPECKER GRID AS OF JANUARY 1, 1986**
- **SOVIETS ENGINEERED COLD WAVE INTO FLORIDA**
- **SEVERE COLD STRESS ON VEHICLE**
- **GRID LOCALIZATION SIGNATURES**
  - **BIRDS NOT FLYING IN LAUNCH AREA**
  - **BIRD FALLS IN BIRMINGHAM, AL (HINGE)**
- **SOVIET SHIPS/TRAWLERS DID NOT TRACK LAUNCH**
- **BOOSTER SEAL PROBLEM SEVERELY AGGRAVATED**
- **CAUSED SHUTTLE FAILURE**
- **SOVIET KGB HEADQUARTERS CELEBRATED "PERFECT SUCCESS OF ACTIVE MEASURES" AGAINST THE CHALLENGER!**

Table 11. KGB CELEBRATION

EVENING OF JAN. 28,1986

- **AFTER DEATH OF CHALLENGER**
- **KGB HEADQUARTERS**
  - **JUBILANT PARTY**
  - **CELEBRATED SUCCESS**
  - **OF ACTIVE MEASURES AGAINST THE CHALLENGER**

SOURCE: GENERAL DANIEL GRAHAM, URGENTGRAM

- **NOTE: SOVIET SCALAR EM WEAPONS ARE CONTROLLED BY KGB**

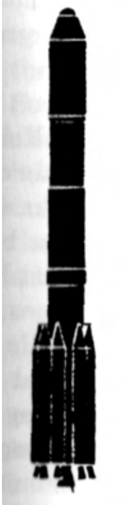

- **SHORTLY AFTER LAUNCH**
- **ANOMALOUS POWER SURGE**
- **"COMMANDED" ENGINE SHUTDOWN**
- **VEHICLE THEN UNSTABLE**
- **RANGE SAFETY OFFICER DESTROYED ROCKET**
- **EXTERNAL RANGE SENSORS DID NOT RECEIVE SURGE SIGNAL**
- **SIGNATURE OF TIME REVERSED WAVE**
- **WHICH LOCALIZED ON ROCKET**

Figure 31. Loss of a Delta rocket, May 3, 1986.

16. In 1975, Brezhnev had surprisingly urged U.S. senators to have the U.S. cooperate in banning the development of new weapons of unparalleled horror. He used the phrase, "more frightful than the mind of man has ever imagined." Shortly after, at the United Nations, Gromyko had even introduced a draft agreement for all nations of the world to sign. Again, the West had not the foggiest notion as to what the Russians were talking about.

17. The orthodox U.S. news media has been, and is, almost totally "out of it." Ironically, at the time of our deadliest peril, and the most momentous period in history, incidents such as the indiscretions of a presidential candidate and of a television evangelist, and the "Iran Gate" actions of the National Security Council in trying to illegally get arms to the Contras, are the "burning issues of the hour" — and the week, and the month, and the year, and HO on ad nauseam.

18. The Congress — great watchdog and the direct representative of the U.S. citizenry — is equally immersed in interminable hearings of little or no substance except news media coverage and publicity.

19. A belabored president, misled and now hemmed in on all sides and on the defensive, has had his power to act greatly constricted. In addition, the orthodox filtering networks surrounding him have effectively isolated him. He has been unaware of the extensive information that shows just how certain it is that the Soviets have fantastic superweapons and have been gingerly using them against us for decades. He and his immediate advisors are fed highly filtered, shaped and interpreted data. U.S. intelligence is very good at doing the things that it does. However, it is also very good at ignoring the things that it ignores! Further, the intelligence community is a past master at "slick-tonguing" the interpretation of anything new, and anything it really doesn't understand, in terms such as "We have no evidence that ..." For decades, the community has consistently underestimated the Soviet threat, and it has also usually projected any unpleasant present Soviet capability in the terms that "If the Soviets keep this sort of thing up, it may prove to be a real threat 30 years from now!" For years U.S. intelligence was unaware of a Toronto inventor's development and public demonstration in the late 1960's of an inertial field generator device, for example, because it simply did not place high value on simple overt intelligence such as reading the newspapers! The entire event was printed in detail in a Canadian newspaper. Only very recently, within the last few years, has the community again realized that the most important intelligence can only come from human (HUMINT) sources. And in relying on the scientific community — which itself knows nothing of scalar electromagnetics — the intelligence community has not received the technical advice necessary for it to properly task its collection agencies to collect information on the Soviet superweapons.

20. Western science has also shown itself to be highly dogmatic and ensconced in the comfortable "status quo." It has regally ignored the important Soviet developments in unified field theory. It has assumed an aristocratic stance, regarding the Soviet scientists as little more than peasants still trying to clean the mud off their boots. In fact, the Soviet scientists have been ahead of the rest of the world in *nonlinear* science — the truly hard stuff— since the beginning. They have always been the best nonlinear mathematicians in the world. Soviet scientific papers regularly appear which deal with local general relativity and the overcoming of the sacrosanct conservation laws as a matter of course. That subject is still a "no-no" in U.S. scientific publications, even though breaking symmetry and consequent violation of conservation laws has long boon proven in particle physics. To my knowledge, with the exception of Hooper's obscurely published work, not a single US. scientific paper appears in the literature showing any detailed investigation of "bucking fields" as such and investigating the behavior of materials exposed to such "sum-zeroed" fields. Yet ironically a whole area of optical physics — phase conjugate optics (PCO) — has shown great activity since we discovered it in the open Soviet literature in the late 60's, and since two Soviet scientists visited Livermore* in 1972 and briefed us upon the subject (as another stimulus to see if we recognized the awesome weapon implications). In PCO, the scalar EM wave formed by two waves 180 degrees out of phase with each other, and locked (modulated) together by a nonlinear medium, is blithely called the *pump wave.* Using it on nonlinear materials produces extraordinary results — such as electromagnetic waves traveling backward in time. Further, the PC effect is known to be universal, and not just in the optics region at all; for example, it has also been done with sound waves. **The obvious experimental and theoretical proof of scalar EM has been in our literature for nearly two decades, and even in PCO our scientists do not yet realize that the** "pump" **wave is a scalar** wave, **a** wave **of**

***Lawrence Livermore National Laboratory**

artificial potential, and a gravitational wave that pumps the nuclei of the atoms in the nonlinear medium. They also still do not understand that it is the nucleus of the atom that produces the phase conjugate replica wave in PCO.

Table 12. BREZHNEV'S 1975 PROPOSAL

AT THE SALT TALKS ON JUNE 13, 1975

- **THE SOVIETS URGED THE U.S. TO AGREE ON A BAN AGAINST RESEARCH AND DEVELOPMENT OF NEW KINDS OF MASS DESTRUCTION WEAPONS MORE TERRIBLE THAN ANYTHING THE WORLD HAS KNOWN!**

- **BREZHNEV PERSONALLY URGED VISITING U.S. SENATORS TO AGREE TO THE BAN.**

- **GROMYKO INTRODUCED A DRAFT AGREEMENT INTO THE UNITED NATIONS MEETING AND URGED THE NATIONS OF THE WORLD TO SIGN IT.**

- **THE WEST HAD NO IDEA WHAT THE SOVIETS WERE TALKING ABOUT!**

- **SOVIET HAWKS THEN GAINED DOMINANCE, AND THE USSR STARTED ITS MASSIVE BUILDUP FOR EVENTUAL WORLD TAKEOVER.**

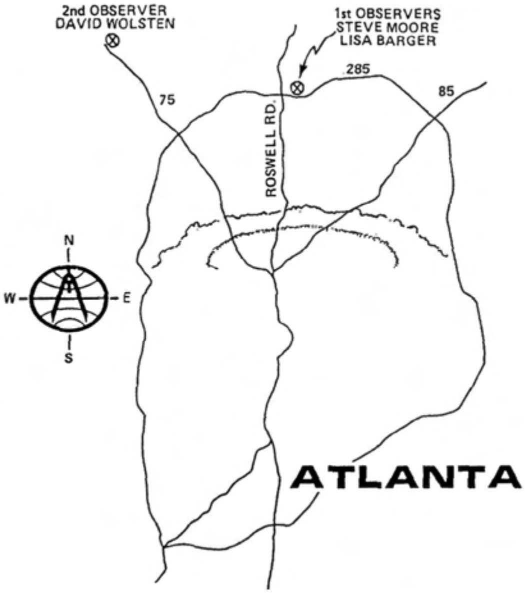
2nd OBSERVER
DAVID WOLSTEN
1st OBSERVERS
STEVE MOORE
LISA BARGER
285
85
75
ROSWELL RD.
N
W
E
S
ATLANTA

A. Location of the incident

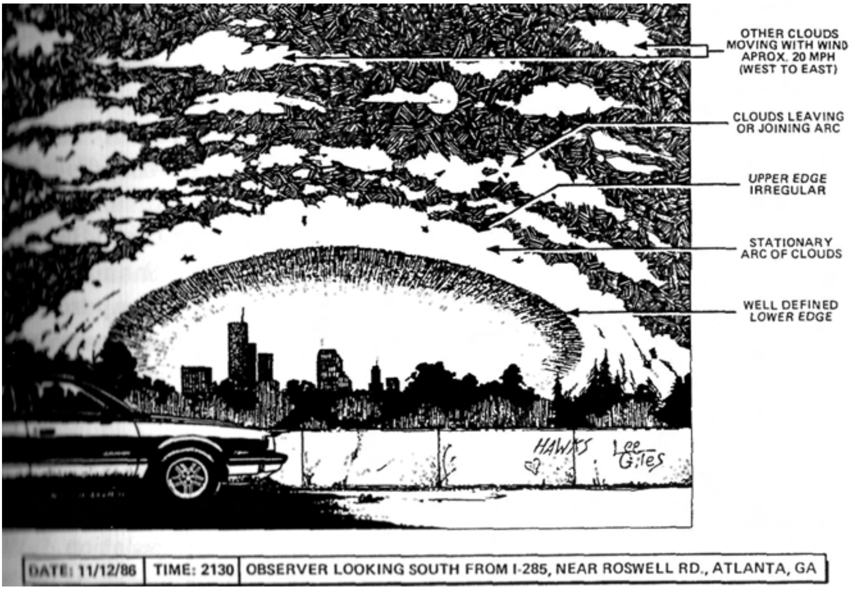
OTHER CLOUDS
MOVING WITH WIND
APROX. 20 MPH
(WEST TO EAST)
CLOUDS LEAVING
OR JOINING ARC
UPPER EDGE
IRREGULAR
STATIONARY
ARC OF CLOUDS
WELL DEFINED
LOWER EDGE
HAWKS
Lee Giles
DATE: 11/12/86
TIME: 2130
OBSERVER LOOKING SOUTH FROM I-285, NEAR ROSWELL RD., ATLANTA, GA

B. Artist's sketch of the incident.

Figure 32. Probable Soviet test of a Tesla Shield over Atlanta, Georgia.

Continually assured that Western science and intelligence remained outrageously ignorant and would obstinately persist in wearing orthodox blinders, the Soviets felt quite safe in striking the U.S. and its Western allies a mortal death blow.

The Soviets knew there would be no retaliation, so long as the blow was sufficiently unorthodox.

After all, the Soviets have been waging a slowly increasing unorthodox war against us for decades, and totally getting away with it.

Consequently the Soviets manipulated the cancer research and the contamination of the World Health Organization's smallpox vaccine to produce the perfect biological warfare "first strike," administered unwittingly by the WHO itself in its mass immunization program to eradicate smallpox. The Soviets continue to manipulate the spreading AIDS viruses electromagnetically so that the West will be devastated, as well as Africa and the Third World.

At least one other spreading AIDS-strain has already appeared in Africa. Testing for exposure to the first strain in most cases did not detect exposure to this second strain.

Officially no one knows whether it will be as infectious as the first strain. However, a suspicious correlation has appeared between hepatitis vaccination in several large U.S. cities and consequent outbreak of AIDS among the persons inoculated.

Further, the latest studies are beginning to show that (1) AIDS may be just as spreadable, though more slowly spread, in the heterosexual community as in the gay community, (2) the AIDS tests themselves are far more unreliable than suspected; one may have the AIDS virus — and serve as an active carrier —for up to three years before the tests begin to detect it, and (3) no one has any real idea as to how many persons have already been infected.

As can be seen, even orthodox scientists are beginning to suspect a coming AIDS infection "asymptote" which may represent an explosion of the disease into a mass pandemic in the world population. In Africa, for example, AIDS is spreading freely among the hetersexual population as well as the homosexual population. * We may be on the verge of seeing a similar phenomenon emerge in the

---

*Although there is much speculation as to why African heterosexuals seem so vulnerable, at present no one knows why it is true.

Western world.

If so, we are looking full in the face of a forthcoming AIDS massacre that makes the 1918 influenza pandemic pale by comparison. Not only will AIDS decimate Africa and the Third World, but it will also decimate the West a little later.

In the Soviet plan, after the AIDS massacre has peaked, the remaining much smaller U.S. and European populations will be decimated and totally powerless. The demoralized and devastated Western nations will easily capitulate and pass under the Soviet hammer and sickle.

In their deception planning, the Soviets even are allowing some contamination of their own citizens by AIDS. Thus they can keep up *a pretense* that they are just as much victimized as we are, and just as much "in the dark" as we are about the whole thing.

Of course, once the AIDS infection does its lethal dirty work in the West, the Soviets can easily eradicate it in their own population at will, using the electromagnetic healing technology they already posess. The Soviet planners regard the loss of a few Soviet citizens along the way as one of the things necessary to insure the success of the Great Strike and the Final Revolution. Forward divisions in the standard Soviet attack are always regarded as such sacrificial pawns to guarantee the success of the follow-on echelons in the attack. In this BW strike, the sacrifice of some Soviet citizens is deemed necessary to the overall success of the attack.

Remember, these are the same type of Soviet communist planners who killed 20 million Kulaks — their own people — to forcibly collectivize Soviet agriculture. For their highest purposes, these planners are still quite capable of ruthlessly sacrificing millions of their own people to achieve their objectives.

The *Asian Tiger Mosquito: A New and Efficient Vector*

As almost everyone now realizes, in the past our government has greatly understated the AIDS problem. Past projections of the limitation oft he spread of the disease have been far more optimistic than warranted.

For example, officially the government is still denying any possibility of insect vectors for the disease.

Don't believe it. In fact, highly efficient insect carriers of the disease are already on the way.

Insects in Africa have positively been verified to be harboring the AIDS retrovirus. Of course, such reports are always accompanied by the bland assurance that "We have no evidence... (that the insect bites will give AIDS to humans).

Let's see now: We're supposed to believe that a monkey bite will do enormously more than that. Even when the monkey didn't have the exact present form of the AIDS retrovirus. Even though monkey bites hadn't done the trick in thousands of years.

But we're supposed to turn right around and believe that an insect carrying the live AIDS retrovirus can bite you, inject into you some of his body fluid containing AIDS viruses, and you won't stand any chance of developing AIDS .

What are we supposed to believe? That the injected AIDS virus from an insect bite won't bother you, but the injected AIDS virus from other causes will?

Nonsense!

Anyone who believes that is a prime candidate for purchasing that bridge down the road...

Of course insect vectors will give you AIDS!

And we've got a real tiger — mosquito, that is, on the way to do just that.

The Asian Tiger mosquito has been introduced into the U.S. (it just "accidentally" appeared, of course!) and is now spreading through the Southeastern states. This vigorous and voracious mosquito prefers man as its target, and has an unusually wide range.

Recently U.S. scientists have reluctantly confirmed the presence of the AIDS virus in the Asian Tiger mosquito. However, as usual we are presented with the same tired refrain, "We have no evidence that... the mosquito's bite is capable of causing a human to develop AIDS."

Of course, you can be sure that eventually we *will* have evidence of precisely that!

The AIDS virus is now known to be able to live outside the body for up to at least 15 days, if not longer.

Also, the AIDS virus is an unstable virus — it's constantly mutating and changing its form. That means it's "highly adaptable" - **that** is, it's continually adapting to its environment, at a rate millions of times faster than normal stable viruses change and

That fact implies that some mutant strain of the virus will almost certainly be adapted to living in the Asian tiger mosquito and being introduced into its human host by the mosquito's bite.*

It's not a matter of "Can it do that?" or "We have (as yet) no evidence of that."; instead, it's a matter of "How long will it be until that occurs and we confirm it?"

The Asian Tiger mosquito apparently is intended to serve as the major vector (carrier), or at least one of them, to eventually rapidly increase the AIDS infection rate and devastate the U.S. population.

AS the vector spreads and the virus adapts to that mode of transmission, the rate of increase of the AIDS infection will itself accelerate. This means that all the projections for "doubling times" are seriously in error.

The latest data from Alabama, for example, shows that the disease is now doubling every six months instead of every year. Also, the doubling interval itself has been shortening, at least up to the p**resent.** Note also that, from the data of the new studies confirming a much longer lag time for the AIDS test to indicate the presence of the infection, even the present data on the number of persons infected and the doubling time are underestimated.

If these and other data are a true indication, it appears that we literally may have an "explosion" of the AIDS infection in no more than two to three years.

In addition, as we said the AIDS virus is an unstable virus. Mutations in the virus itself are occurring so rapidly that producing an effective vaccine essentially will be impossible. Without a totally new means applied to produce an AIDS cure, AIDS simply cannot be **stopped.**

•And one or more strains will eventually evolve that can be infectuously spread like the influenza virus is.

In fact, recenly it has been discovered that the AIDS virus is far more devious than previously susupected.* It attacks in more ways than suspected. It changes in the body, increasing its virulence as the disease progresses.

Macrophages, the immune system's scavenger cells, become infected and carry the virus to other body cells. It appears that macrophages may pass the virus to the T-cells, conditioning the destruction of the immune system. Infected macrophasges are found in rectal material of AIDS patients and in cervical fluid and semen. Thus through infected macrophages the AIDS virus is able to infect partners through exhange of semen, cerical fluid, and rectal material — even though there are no breaks or tears in the skin, and even though the number of free AIDS viruses in the fluids is too small to pass the infection.

Other immune system cells —such as Langerhans cells—are also suspected to be capable of serving as AIDS carriers when infected. The end result is that, as the AIDS patient sickens, the virulence of the AIDS virus increases, so that it attacks a greater variety of cells and can more easily infect them.

Present orthodox methods cannot cure AIDs in time, if ever.

Unless unorthodox technology is applied to solve the problem the West is already defeated.

---

*See Gina Kolata, "Fatal Strategy of AIDS Virus Grows Clearer," Science Times section, **New York Times,** Mar. 22, 1988; see also "Growth in AIDS Virus Virulence Found," **Los Angeles Times,** Apr. 1, 1988.

CHAPTER 3

WORLD WAR III IS RAGING AND WE'VE NEARLY LOST IT

More on AIDS Doubling Time

AS we said, the AIDS problem is much, much worse than officially announced. As stated, in some areas at least, the number of new AIDS cases diagnosed has been doubling every six months instead of every year. New studies have shown that the tests to identify infection by the AIDS virus may themselves have a 3-year lag period in which they do not reliably indicate the presence of the infection.

These facts probably make all previous "officially announced" estimates and projections off by several orders of magnitude (at least three).

If the incubation period (nominal) is, say, 5 years, then there is a factor of up to 2 (exp 10) of AIDS cases already in incubation.

The lag in accuracy of the AIDS tests however, may mean that there are even more AIDS cases in incubation than this.

And we may have a new entry into the equation at any time, once the AIDS virus adapts to its transmission by the Asian tiger mosquito and possibly even other insect vectors.*

Remember, we've got the U.S. data for only a relatively few years since 1981. So it's statistically difficult to project ahead with any accuracy. However, the best interpretation of the data we can make shows a spiraling (speed-up) trend in the rate of acceleration of the disease, indicative of a forthcoming asymptote. If that interpretation of the data holds true, the asymptote represents literally an "explosion" of the disease.

That is, the six-months "doubling interval" is slowly becoming **the s**ix-months "tripling interval," the six-months "quadrupling interval," etc.

The best indication/projection I can tentatively arrive at is this: If the total cure is not accomplished within about two to three years, then - considering the incubation ongoing by then and the speedup

---

***Another HIV virus has now been identified.**

ahead — somewhere between a third and two thirds of all humans on earth will perish.

In other words, nominally half the humans on earth will inevitably die, say within five years (the incubation) after that two to three year point.

Most of the people who die are going to be highly selected. Most of them are going to be Westerners, Africans, Asians, and Brown people.

And all of our tanks, ships, bombers, missiles, nuclear weapons, and armed forces put together can't save a single AIDS victim.

We've got more generals than ever before in our history, including World War II. And not a one of them has a single tactic or a single weapon to use against this mortal attack.

Understand, our generals have not failed us against the enemy they know and recognize. The bellies of our B-52 bombers and the aircraft aboard our distant aircraft carriers are filled with nuclear bombs sufficient to destroy the world many times over. Our powerful nuclear-tipped missiles sit in land-based silos and in sleek submarines cruising silently under the oceans, waiting to respond in an instant to hostile attack, again with the power to destroy the world many times over.

Yet the fine men and women who have built and maintained that awesome force, and who man it in constant vigilance to protect our nation, are absolutely vulnerable and defenseless against the smallest foe of all —just as were the Martians in H.G. Wells' **War of the Worlds.**

With all our awesome forces, we do not have the power to stop a single AIDS virus — or indeed, any other similar new BW virus.

We can be defeated with ridiculous ease by even the tiniest determined hostile nation — or even by a single terrorist, madman, or genetic accident by a high school youth.

And if there's one thing that we can learn from history with iron certainty, it is this: If we can be defeated with ridiculous ease, then most certainly we will be!

Since the strike has already occurred, then our defeat is just a matter of time if we continue our present courses of action.

All other threats — nuclear, chemical, scalar EM, the works — pale in comparison to this seemingly innocuous, doomsday BW viral weapon that has already been unleashed upon an unsuspecting world.

World War III is Already Raging

World War III has long been unleashed and a final, decisive first strike has now occurred. We've already lost the war unless unbelievably heroic measures are taken immediately.

This thing is deliberate. It's fiendishly clever. It's effective. And a smiling Gorbachev — master propagandist — is just part of the icing on the cake to deceive us until it's too late.

Ironically, Soviet announcements have cleverly started claiming that the U.S. has produced the AIDS virus and unleashed it.

That's a very clever propaganda trick to shift suspicion from themselves. Indeed, in the Third World the propaganda has been very effective. Most Third World peoples now give at least credence to the Soviet claim.

Unless we move immediately to effect a total cure, the armed forces and the populaces of all the Western nations will shortly be devastated — perhaps even worse than if they been attacked with nuclear weapons.

With our armed forces and populations devastated, we surrender. We lose.

The Soviets get the U.S. (and European) farmland. That's what they really need.

Of course they also will have rid themselves of about 200 million Americans or so along the way.

That, solves the old "winner's problem" of what to do with, and how to effectively control and care for, a conquered, resentful populace. That problem will have already been taken care of, for most of the U.S. populace will essentially have been eliminated.

There won't be many U.S. military prisoners, for example, since most of the military personnel will also have been eliminated by the virus or viruses clandestinely released.

Communism Is Savage Enough to Do Diabolical Things

The Communist's doctrine is to wage total war. Win any way, with any weapon, at any cost.

ALSO, communism has always taught that it is necessary to destroy capitalism. That equates with destroying the U.S., the major capitalist country. Communists have always taken the long view, willing to struggle for decades to eventually accomplish their aim.

If you think no modern nation could be so incredibly savage as to deliberately destroy untold millions of people, look at the actual record of the Communists.

When the Khmer Rouge took over Cambodia, they murdered a third of their own people.

The Soviets murdered some 20 million or so Kulaks when they collectivized the Soviet farms.

The Red Chinese also slaughtered Chinese peasants by the tens of millions — perhaps as many as twenty or thirty million — when they collectivized the farmers.

Lenin's followers have proved over and over that they are capable of murdering even their own people, on a scale of millions.

Dedicated communists are highly religious people — even fanatics. Deliberate murder to achieve communism's goals is part of the accepted dogma.

The communists are quite capable of killing a billion or more "foreigners" and "enemies," given the chance. Particularly when the goal is to win the "Final Revolution."

The opportunity to strike the decisive final blow without consequences, using as simple and unorthodox a weapon as the AIDS viruses, could simply not be ignored by men who have spent their lives trying to control the world.

The unparalleled strike opportunity was particularly irresistible since someone else (the Western cancer research institutes) would develop the weapon itself, someone else (the World Health Organization) would unwittingly deliver the strike, and the blow could easily be clandestinely arranged (simply by secretly contaminating the WHO smallpox vaccine) and accomplished without repercussions to the Soviet Union.

Western Medicine is Inadequate

But what of Western medicine? Doesn't it always come through to save us?

Present Western medical science is totally inadequate to do the task required.

For example, present medicine is largely based on "slowing or killing the invader of the body." Whether that invader is a germ, bacterium, virus, or whatever.

If there's a mechanical difficulty, then present treatment is to try to mechanically remedy it — remove a bad appendix, set a broken leg, etc.

However, then it's strictly up to the body's own natural healing system - its immune system and its repair system — to actually do the healing and repair/reconstruction.

The problem with AIDS is that it directly affects and suppresses the body's systems for healing and repair.

Once this protection and repair system is in disarray, any disease organism can attack the body and infect it. If the immune system can't handle the problem, eventually the invading disease organism kills its host.

Even if a way is found to destroy the AIDS virus itself, present medicine has not the foggiest notion of how to restore the healing and repair system from the damage and destruction it will have already sustained by massive AIDS infection.

Compare the situation to division combat: Hostile invaders are met at the front of the division by the division's combat forces. A battle ensues, with penetration and damage to the rear area support and sustenance facilities. After the enemy is beaten back or destroyed, the rear support and the combat forces must be renewed and restored. If that cannot be done, the seriously weakened division can still literally be boiled over by almost anything that comes along and hits it.*

---

*For this reason, standard Soviet doctrine is to attack in echelon, say, by three divisions — one behind the other. Thus, even if the first division is annihilated in the attack, the second and third divisions meet a seriously weakened enemy division and blast right over it. Similar tactics in the BW crisis - multiple viral strikes, one after the other — can almost inevitably be expected, now that the "first division" - AIDS - has engaged us.

The Soviet "repetitive attack" method *works,* in case anyone has doubts. For example, see George C. Wilson, "Army training exercises raise troubling questions," The Washington Post, as printed in the Huntsville (AL) Times, Feb. 28, 1988, p. 21A. Repeated U.S. Army tests in the Mojave Desert have shown that, in nonnuclear combat, when attacked by multiple Soviet echelons with 3:1 strength, U.S. forces lose every time! In particular, U.S. troops have poor anti tank defenses.

Modern medical dogma has concentrated almost entirely on the combat forces' problem of destroying and/or repelling the invader.

Western medicine has almost no tools or means of restoring the support and rear maintenance functions.

Even in a paroxysmal effort, dogmatic Western medical science would take decades to totally reorient its medical thinking to where it could even attack the real problem of AIDS and catastrophic sickness due to other viral BW warfare agents.

The same inability to cope with the problem also applies to our government bureaucracy, government agencies, the private companies, the universities, and to Western science in general.

If we have to depend on any or all of our present "system," the war is already over and we are all already doomed. Half of humanity will die within possibly five years after the upcoming two- to three-year "point of no return."

Medical insurance, treatment centers, and even private physicians are quickly going to be bankrupted as the AIDS plague increases. With the collapse of Western medicine, panic and pandemonium will reign in the streets. Martial law will be a necessity, and rioting on a scale unparalleled in American history will occur. We are likely to see our own U.S. dictatorship, in the form of a military junta making a coup and seizing power "to restore order and keep the lid on."

Study carefully the performance of governments and military regimes in such countries as Korea and South Vietnam, when those countries were hard pressed by enemy forces and in great disarray.

You may be looking at your own future.

The Soviets Will Not be Decimated

With the devastation of our armed forces, our governmental and civil agencies, and our population at large, a simple show of strength by the Soviets will collapse any further resistance. The hammer and sickle will wave supreme over the residue.

That's because the Russians *do* have the necessary developed scalar EM medical technology, not only to control and destroy the AIDS viruses, but also to directly restore the immune and repair systems.

The Kaznacheyev experiments alone show they have extensively studied electromagnetic induction of cellular disease and death.* We will have much more to say on that later.

If the Soviets extensively studied cellular death transmission by electromagnetics, you can bet they also studied the reverse: electromagnetic healing, or time-reversing the electromagnetic "death" signals themselves to produce specific "healing" signals for any type of cellular disease.

This is particularly certain since the Soviets also discovered and possessed the necessary "time reversal" electromagnetics: phase conjugation. This had been widely used by the Soviets in developing their enormous scalar EM weaponry, beginning about 1950 or so.

The curative signal technology was kept secret, of course, since the Soviets realized that it provided the only viable defense against viral biological warfare weapons.

The Soviets most certainly have the necessary AIDS cure. They can readily make the correct electromagnetic healing signal, by phase conjugating the AIDS death signal from dying human cells that contain AIDS retroviruses.

They also know we most certainly do *not* have the necessary EM cure. Years of weak microwave radiation of personnel in the U.S. Embassy in Moscow, and inducing illnesses in those personnel and even U.S. ambassadors, have shown that our own intelligence and scientific communities don't know what's really going on there.

Specificially, we've demonstrated (after checking with the DIA, CIA, NSA, and the orthodox scientists) that we do not recognize electromagnetic *biological warfare* —the deliberate introduction of cellular disease and death by electromagnetic means. Not only do we not recognize it, we can't do anything about it when it hits us.

And make no mistake, time-reversed electromagnetic medicine is shockingly cheap and effective, once developed. Specific healing signals for specific diseases — viral or whatever — can be quickly developed (within hours) and immediately used for mass treatment of entire populations.

For example, given the developed healing or "antidote" signal, it can even be recorded and played over ordinary transmitters, using their normal transmissions as carrier waves. A populace continually bombarded by such "healing" television, radio, and powerline signals simply doesn't get the disease. And those already infected recover when continually exposed to such healing signals.

---

* There is now reliable information more closely associating Kaznacheyev with military *institutes near Moscow, engaged in* microwave weapon (directed energy) research

So the Russians can stop the AIDS epidemic cold, anytime they choose to do so.

Electromagnetically they can also gently guide and stimulate the progress of a disease such as AIDS, once the disease is introduced into a targeted population.*

Or, the Soviets can pull out all the stops and induce the disease in a distant population directly, placing the "death signals" on common carriers. Such as on our power lines, radio stations, televisions, and directly by induction from the Woodpecker over-the-horizon radar signals which continually bombard us.

The West cannot do any of these things. Neither can the West counter them, or defend against them effectively.

Unless a totally different tack is taken in the Western effort to counter this viral BW threat, Gorbachev's entrancing smile has already lured us to our doom. We are faced with a direct realization of Lenin's words:

"In the final revolution it does not matter if three-quarters of the earth is destroyed, so long as the remaining quarter emerges communist."

And Khrushchev's ghost will have the last laugh after all. We will have fulfilled the prediction he made to Nixon in their famous "Kitchen debate":

"We will bury you!"

Quite simply, for some decades the Soviets have had a new science — one they call *energetics,* and one which I have called *scalar electromagnetics.* The keystone or "heart" of this new science is time-reversed (phase conjugation) electromagnetics and electrogravitation.

*A candidate for an eery test of just this capability may be the highly anomalous deaths of numerous dolphins off the East coast of the U.S. in 1987. Nearly 500 died in the summer and late fall, apparently as a result of the weakening of their immune system to the common bacterium Vibrio. See Julia Lawlor, "Key: Failure of Mammals' **defenses,"** USA Today, **Jan. 19, 1988, p. 1A-2A.**

Chapters 4 and 5
***deal** with complex scientific material*
and are suggested for those who are
technically proficient.

CHAPTER 4

EXTRAORDINARY PHYSICS

Maxwell's Lost Unified Field Theory

About the time of the U.S. Civil War, James Clerk Maxwell succeded in unifying magnetism and electricity. Actually he did far more than that, in his theory as originally written.

In fact, he had produced a theory which also captured the free interchange between electromagnetic energy and gravitational energy, but no one — including Maxwell himself—realized it at the time.

Maxwell wrote his original theory in quaternion and quaternion-like mathematics. The modern form of vector mathematics had not yet been finalized by Gibbs and Heaviside. It is most instructional to examine some of the fundamental differences between a vector and a quaternion.

In a conventional 3-dimensional vector, one may have three vector components, such as

$$v = ai + bj + ck \qquad (4\text{-}1)$$

where i, j, k are unit vectors in the directions of the x, y, and z axes respectively and a, b, and c are constants.

Obviously if the vector components of vector v are zero, then

$$v = 0 \qquad (4\text{-}2)$$

We shall be interested in the "vector product" of two identical vectors v, where

$$\mathbf{v} \times \mathbf{v} = AA \sin\emptyset = \mathbf{O} \qquad \mathbf{(4\text{-}3)}$$

and A is the length (magnitude) of vector v, Ø is the angle between the two vectors (in this case zero), and O is the zero vector.

Now let us look for a moment at the quaternion situation.

First, in addition to the three vector components, a quaternion also has a separate scalar component, w. So the quaternion q for this situation is

$$q = w + ai + bj + ck \qquad (4\text{-}4)$$

Now when this quaternion is multiplied times itself, the vector part zeros, just as it did for the vector expression. However, the scalar part does not go to zero. Instead, we have

$$q \, X \, q = A^2 = a^2 + b^2 + c^2 \qquad (4\text{-}5)$$

There is a very good physical interpretation of this result. It is a square of the amplitude, hence for the vector part of a wave, it is directly proportional to the energy density of the vacuum, as a function of time, at the particular position. However, we now need to make a short explanation of variation of stress energy density of spacetime.

First, we note that, according to general relativity, the "gravitational potential" is just a conglomerate of potentials of all kinds. Basically, a potential represents a G-potential, and consequently a curvature of spacetime. The potential also represents "trapped energy."

Second, we note that Kaluza combined electromagnetics and gravitation as a unified theory in 1921. Kaluza added a fifth (spatial) dimension to Minkowski's 4-space, and applied Einstein's relativity theory to 5 dimensions.

To Kaluza's delight, a common 5-d potential is responsible for both electromagnetic field and gravitational field. The "bleed-off of this 5-potential in the 5th dimension (which is wrapped **around** each point in our 3-space) is what we know as the **electromagnetic force field.** The bleed-off of this 5-potential *in and through* our 3-space is what we know as the **gravitational force field.**

Since the EM field is very much stronger (by a factor of 10^{42} for electrons) than the gravitational field, it is obvious that most of the bleed-off of the 5-potential is in the 5th dimension, as EM force field. Only a tiny bit is left to bleed-off in 3-space, producing a very weak gravitational field.[1]

We state this fact: as a mass moves in space, it generates increased "activity" with the virtual particle flux of vacuum itself. **The** Increased virtual particle flux activity exchange between vacuum and mass is analogous to a strange kind of "virtual resistance." Since the resistance is virtual, it does not observably slow down an observable object moving in an (unobservable, virtual-particle flux) vacuum.

The increased flux activity represents an increased "virtual energy density" of space time, and an increased "trapped potential" (mass; resistance to an accelerating force) of the moving object. It represents a rotation of the spacetime frame, vis a vis the laboratory observer).

In the virtual vacuum (which contains both positive and negative time), one sees two antiparallel virtual forces: one in positive time, along the velocity vector of the object, and one in negative time (time reversed, or phase conjugated). The reason one sees virtual forces is that, each virtual (subquantal) change in the virtual flux activity represents an individual (unintegrated), separate change, hence a virtual acceleration. The observer (where things are integrated), sees the integral of all these accelerations, hence observable velocity.

The vector sum of these two virtual forces in the vacuum is a zero vector; however, the two taken together represent a stress in the local energy density of vacuum.

Since we may regard an EM wave as a stream of virtual electrons/positrons, each engaging in tremendous virtual particle flux exchange with the vacuum, then the same basic picture applies.

Now for our physical interpretation of (4-5): If we refer to an EM wave moving in the vacuum, the rotation of the frame is maximum (90 degrees). But this same rotation is just the same as additional vacuum stress, so the vacuum stress is maximum.

---

[1] **Electromagnetics is 5-gravity sliding around our 3-space. 3-gravity is 5-gravity oozing** through our 3-space.

This leads to these conclusions: An electrical force field vector represents a local maximum linear stress in spacetime, along the line of the vector. (Note we specifically deny that the electrical force field vector, of an EM wave in vacuum, is transverse. Instead, it is longitudinal. That has been addressed elsewhere by the author and will not be covered further here.)

Another electrical (stress) vector interacting with the first one adds more "urging" stress to the first. However, this action is occurring in the rotated frame of the moving wave, and so is rotated 90 degrees from the electrical velocity vector. Therefore it is lateral (but in a hyperdimension, not in 3-space) at right angles to the electrical velocity vector.

The combined "urging" action of the two vectors thus sweeps out an area with respect to the laboratory observer.

This means that the total "urging" or "stressing" action of the two vectors is analogous to a vector area.

It also means that this "area" function may be taken as the "swirl" of the electrical vector, but in a hyperdimension, not in 3-space. That is, we have described the magnetic force field.

Thus any two electrical vectors that interact will have an "area" or "resistance" component generated. *Any two that interact.* Whether they add vectorially, cross-product multiply, or dot-product multiply.

What is actually happening is that the wave exists in the 5-potential. The E and B fields just represent the oscillations in that 5-potential. They represent oscillations in the bleed-offs of that potential as E-field (longitudinal) and B-field (swirl).

The drag-area represents the accumulation of extra potential — hence the local rotation of spacetime. Since this accumulation is moving (along with the EM wave), as it passes a point it represents a change in the local virtual particle flux density of vacuum at that point, **hence a local curvature of spacetime.**

Hence, the EM wave makes a 5-dimension G-potential wave as it travels. The 3-dimensional gravity wave associated with this is normally very, very much smaller in magnitude — say, by a factor of 10^{-42} or so.

However, if the two vectors interact so as to produce a vector zero resultant, then all the electromagnetic energy of tho two vectors **is** captured. That is, all the "EM vector zero" resultant means is that the EM bleedoff of the 5-space gravitational potential wave has been stopped. The 5-potential is still oscillating, and now all its trapped 5-energy must bleed off as 3-gravity force field.

Mass acts as an accumulator for this "trapped-EM energy turned into local curvature of 5-space." If we continually irradiate a mass with such a wave, the atomic nuclei of the mass slowly charge up with the new energy. Note that this potential delta may be positive or negative, if one adjusts accordingly.

In this fashion one may change the mass of a static object in the laboratory. One may either increase the mass or decrease it, or cause it to float, or even cause it to accelerate upwards.

But to return to our vector interaction and our interpretation of the scalar remainder of the quaternion.

The rule is, when the two EM vectors interact so as to form a zero EM resultant, then the EM energy represented in each of the two vectors has been converted into a special form of 5-space gravitational potential, one that is not bleeding-off in the fifth dimension (electromagnetically), but one which will gradually produce a 3-gravity potential in a mass's atomic nuclei as a function of time, the individual element, permeability and absorption factors of those nuclei, etc.

Therefore in our mathematical theory we ought to have a scalar component remaining when two EM vectors interact to form an EM vector zero resultant. That scalar component represents what is happening in the 5-potential, that will only bleed into 3-gravity.

With exploration of this phenomenology in the laboratory, one can work out the functions, constants, coefficients, and parameters which specify how the "5-G to 3-G and vice versa" component works in conjunction with mass, motion, and other fields.

That's the magic secret of electrogravitation.

It was captured inherently by the quaternion theory of Maxwell published during the American Civil War!

After Maxwell's death, when the scalar portion of the quaternion was discarded (by Oliver Heaviside) to form "modern" EM theory, that also discarded the unified field interaction between electromagnetics and gravitation.

Electromagnetic field and gravitational field were then modeled and regarded as mutually exclusive. EM field, therefore, was thought to produce no specific gravitational effects in the vacuum itself.

Hence when Albert Einstein was formulating general relativity some decades later, he knew only one way to "curve" spacetime: that was gravitationally, by "attraction of mass" forces.

But gravitational force was so weak that only a huge collection of mass would exert enough of it to measurably curve spacetime. That would require a sun or star. Since the observer and his instruments would never be on the surface of the sun or a star, **Einstein** assumed **that the local** spacetime **of the observer would not be curved.**

Hence he severely crippled his general relativity theory. In the West, it remains an assumption to this day. It is not a universal assumption in the Soviet Union, however, since the Soviets have long since written — and developed in the laboratory—unrestricted general relativity with *local* spacetime curvature, and hence *local violation of conservation laws.*

So the scalar part of the quaternion interaction, that remains when the vector part of the resultant is zero, is magic indeed.

That is the magic unified field portion that everyone has been seeking for decades and decades!

It was there at the beginning. Then we inexplicably threw it away!

But to return to our vector/quaternion examples.

Note also that the two vectors

$$V_j = ai + bj + ck,$$
$$v_2 = -ai - bj - ck \qquad (4\text{-}6)$$

sum to zero vectorially when added, such that

$$v_1 + v_2 = 0 \qquad (4\text{-}7)$$

However, quaternions may behave quite differently, even under addition. For example, the two quaternions

$$\begin{aligned} q_1 &= w + a\,i + b\,j + c\,k, \\ q_2 &= w - a\,i - b\,j - c\,k \qquad (4\text{-}8) \end{aligned}$$

sum their vector parts to a vector zero resultant, but do not sum to a scalar zero as well. Instead, they sum to

$$q_1 + q_2 = 2w \qquad (4\text{-}9)$$

AS can be seen, quaternions which have the same vector parts as vectors, do not necessarily yield a complete zero when the vector parts sum to zero. And when two vectors multiply to provide a zero vector resultant, corresponding quaternions may yield a scalar term that is equal to the product of the magnitudes of the two vectors. In this way, the quaternion approach can capture the *stress of the medium,* induced by opposing or multiplying vectors. In the vector approach, the stress of the medium is entirely lost when the two vectors sum or multiply to a zero resultant.

Let us see just how important this "vacuum stress" can be.

First, the "stress in the medium" represents curvature of spacetime when that medium is the vacuum/spacetime.

In other words, the quaternion approach captures the ability to utilize electromagnetics and produce local curvature of spacetime, in an engineering fashion. Heaviside wrote a subset of Maxwell's theory where this capability is excluded.*

---

*Dr. Henry Monteith has independently discovered that Maxwell's original quaternion theory was a unified field theory. See his important "Dynamic Gravity and Electromagnetic Processes," in publication

Note that, by Maxwell's original quaternion theory, however, Einstein's assumption need not be true at all. For example, look at equations (4-5) and (4-9): Here we may utilize electromagnetic force quaternions to produce zeroed EM forces, and an increased stress in local spacetime. In other words, we have curved local spacetime electromagnetically. Since (with electrons) electromagnetic forces are about 10^{42} times as strong as the gravitation force, this local curvature of spacetime is not negligible.

That is, we have produced a scalar effect from zeroing vector operation between electromagnetic forces. I have called this **scalar electromagnetics,** and pointed out that it is truly *electrogravitation.*

We stress again that this violates one of the severely limiting assumptions that Einstein placed upon his theory of general relativity. He assumed that curving spacetime could only be done by the weak gravitational force due to mass. Since gravitational force is so weak, only a stupendous collection of mass — such as the sun or a star — could curve spacetime enough to notice experimentally.

Since obviously the observer and his laboratory instruments would never be located on the surface of the sun or a star, *Einstein assumed that the local spacetime would never be curved!* In other words, the local frame would always be a Lorentz frame. This meant that, locally, the familiar conservation laws of physics would always apply. Curvature of spacetime would only occur at great distances, and at huge collections of mass such as a star or dwarf star.

Einstein did not write a complete, unlimited general relativity. He wrote a sort of "special relativity with distant perturbations."

If Einstein had had electromagnetic theory in quaternions, the scalar "vacuum pressure" parts would have been there for him to ponder. It is highly probable that he would have captured the "electromagnetics-to-gravity conversion remainder" in the quaternion interactions.

If so, he would have written the full theory of general relativity, involving local violation of conservation of energy, a unified field theory, and the direct engineering of

gravitational and antigravity effects on the laboratory bench by electromagnetic means.

In that case, we should long since have navigated all around the solar system, colonized the planets, produced practical free energy devices and power systems, and avoided two great world wars and a host of little ones.

But let us now see if we can make a gravitational wave, electromagnetically.

Again, regard equations (4-5) and (4-9). Suppose these are instantaneous operations of EM force quaternions whose vector parts are varying in magnitude, but in such a manner that the vector parts ways form a zero vector resultant. Now one can see that the scalar part remaining — which represents the stress of local spacetime - is varying as the product of the magnitudes of the vectors in the interaction vary.

This means that one has now produced a scalar wave that represents the local variation of spacetime curvature in an oscillating manner.

Rigorously this is a gravitational wave. It has been produced locally. It has been produced by Maxwell's original unified theory.

Again, I have called this area *scalar electromagnetics.* The Soviets call it *energetics.*

Where local spacetime curvature is varied, conservation laws (energy, conversation, etc.) need not hold. Curved one way, the local spacetime acts as a source (of energy, charge, etc.) Curved the other way, the local spacetime acts as a sink (of energy, charge, etc.)

The Soviets often do not utilize the same restricted kind of general relativity that Western scientists adhere to.

Soviet papers in general relativity regularly point out the complete and unrestricted theory, where local spacetime curvature is allowed. They also point out that all conservation laws may be violated by such local curvature. Thus the Soviets have no unduly dogmatic respect for conservation laws.

Further, by assuming the possibility of local spacetime curvature, Soviet scientists have assumed the possibility of direct experimentation with general relativity on the laboratory bench.

In the West, we have assumed that such cannot possibly be done, because of Einstein's limiting assumption of no local spacetime curvature. Thus Western physicists are strongly conditioned away from electrogravitation.

This is particularly ironic since the basis for just such an experimental theory was produced by none other than Maxwell himself in his original theory of electromagnetism.

Indeed, shortly after the U.S. Civil War, we should have been developing antigravity spaceships. We should have developed electromagnetics a la Maxwell and been on our way to the planets of our solar system. For Maxwell had — admittedly somewhat unwittingly — given us the basis for the necessary engineering theory of unified electrogravitation.

Heaviside's Mutilation of Maxwell's Theory

Well after Maxwell's death, Oliver Heaviside helped to finalize what is today vector analysis.

Then he undertook to "translate" Maxwell's theory from quaternion form to the new vector mathematics form.

Now quaternions were devilishly difficult to calculate in. So much so, that a majority of the electrical scientists (there were not very many of them in those days!) were in despair.

Not to worry! Heaviside took a broadax, figuratively speaking, and simply chopped off the scalar term, leaving only the vector components.

With that artifice, he greatly simplified the calculations to be performed.

Of course, he also threw away the EM stress of spacetime! That is, he threw away the "gravitation" part of Maxwell's theory!

Let me stress this fact most strongly. **After Maxwell's death a single** man**—Oliver Heaviside—directly altered Maxwell's equations, eliminating localized electrogravitation and producing the form of the theory taught throughout the West today as "Maxwell's theory."**

Maxwell's theory has never been taught in Western universities! Only Heaviside's crippled subset of the theory has been taught!

Then, shortly before the turn of the century, a short, sharp "debate" erupted in a few journals — mostly in the journal *Nature*. Only about 30 scientists took part in the "debate."

It wasn't really much of a debate! The vectorists simply steamrolled right over the remaining quaternionists, sweeping all opposition before them.

They simply threw out the remaining vestiges of Maxwell's quaternion theory, and completely adopted Heaviside's interpretation

Thus, a little over a decade later when Einstein wrote his general relativity theory, he did not know that the original work of Maxwell already indicated the unification of gravitation and electromagnetics, and indicated the ease with which local spacetime could be electrogravitationally curved locally and *engineered.*

Accordingly, he placed the scientists of the West on a road which rigorously assumed that a unified field theory was yet to be discovered. It also strongly discouraged any experimentation aimed at curving local spacetime, for it assumed that such could not be done.

After Potsdam and World War II, a frustrated Stalin was to drive his scientists to review the entire scientific literature of the Western world, actively seeking a great new technical breakthrough area such as the Allies had demonstrated with the development and use of the atomic bomb.

Great Soviet institutes — one staffed, for example, with over 2,000 PhD's — were set up to thoroughly review all the Western scientific literature from its very beginning. Anything interesting, anomalous, or unknown was put aside for further examination.

It is a good bet that the meticulous Soviet scientists discovered the difference between Maxwell's original electromagnetic theory and Heaviside's mutilation of it. Great mathematicians that they are, Soviet scientists would have realized the implications of the difference. With their knowledge of unlimited general relativity, they would have made the connection to electrogravitation.

By 1950 they had indeed done so, and were deeply into the development of what they called "energetics", and I have called *scalar electromagnetics.*

They had also reached another milestone about the same time — 1950 or so.

After WWII, both the Soviets and the U.S. were keen on securing the best of the German scientists. The U.S. particularly wanted missile scientists and rocket engineers. The Soviets wanted them too; but they also wanted the German radar specialists and infrared specialists.

The West didn't care about the German radar scientists and engineers, and the IR fellows. The Soviets did, and they got them. That was to prove a most spectacular benefit indeed.

During the war, the Germans had placed extreme emphasis upon radar and radar absorbing materials (RAM). The German scientists had fantastically developed and extended the science of radar cross section — which is the heart of the matter and very, very complex. **They were much further ahead in radar cross section theory at the end of WWII than where the U.S. is today, in the opinion of** some **U.S. radar experts.**

So the Soviets started with a great jump on us in radar knowledge, and they have steadily increased the lead over the years.

In addition, the Germans had developed highly successful radar absorbing materials, and much of the theory to accompany them.

Such materials turn out to be the key to how to build and develop a radar phase conjugation mirror, to produce a time-reversed radar wave.

Thus, because of the German scientists, by 1950 or so the Soviets had already discovered phase conjugation. And they had discovered it in radar first, not in optics!

They would have been primed for the discovery by their great review of Western literature and the foundations of science, since they would probably have noticed that the time-reversed wave is a solution to the wave equation. If so, they would certainly have realized its generality throughout all physics, all frequency bands, and all types of waves.

Superb mathematicians that they are, the Soviets would certainly have made the Kaluza-Klein theory connection, and also realized that phase conjugate waves carry negative energy as well as negative time. They would quickly have seen the gravity and antigravity implications.

So about 1950 or so, the Soviet Union would have started phenomenology experimentation in earnest, with phase conjugate radar mirrors and phase conjugate radars. This is what was referred to as *energetics.* The Soviets began a massive program in energetics about the time of the beginning of the Korean War.

By 1957-8 the Soviets had progressed to the point of a giant scalar EM accident in the Urals which exploded nearby atomic wastes, devastating the area. They had also progressed to development of great new superweapons using their new energetics — weapons to which Khrushchev referred in 1960 when he informed the Soviet Presidium of a new, fantastic weapon in development, a weapon "so powerful that it could wipe out all life on earth if unrestrainedly employed."

About the same time (mid-to-late 50's), the Soviets had also started the eery low-level microwave radiation of the U.S. Embassy in MOSCOW, to see if the U.S. knew of scalar electromagnetics (energetics) and was developing its own electrogravitational weapons and defenses.

Building Upon Whittaker's Fundamental Work

In 1904, a most fundamental paper in the foundations of electromagnetics was delivered by the British mathematician E. T. Whittaker. (E. T. Whittaker, "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.).

In this important paper, Whittaker showed that the electromagnetic force field equations can be replaced with the derivates of two scalar potential functions.

He also derived the most general form of electromagnetic disturbances in the ether.

This means that the coupling of two dynamic scalar functions can replace vector electromagnetics in the vacuum.

Note that Whittaker's work pointing out the overriding importance of scalar fields also accents the erroneously discarded scalar part of Maxwell's quaternion electromagnetic theory even more strongly.

Let me explain now how I got from Whittaker's paper to scalar electromagnetics, Soviet Tesla weapons, free energy, antigravity, and electromagnetic healing.

When I discovered Whittaker's paper, I had already strongly objected that "charges" and electromagnetic vector force fields — as presently included in the Heaviside version of Maxwell's equations — included observable mass. Of course there was no observable mass in the vacuum, hence the prescribed kind of EM force fields could not exist as such in the vacuum.

Obviously the foundations of our ordinary electromagnetics theory were seriously flawed. Although my objections fell on deaf ears, I determined to examine the foundations of EM theory, discover the flaws, and at least point out the necessary corrections to be made.

Though this was an arduous task to undertake and it required many years, slowly the flaws showed themselves, and the necessary corrections slowly became clearer.

Most exciting of all, in working with several unorthodox researchers, I was able to see many of these new ideas tried, adjusted, and demonstrated. In addition, the proprietary discoveries of these colleagues continued to reveal new and unique principles and concepts. The only disadvantage was that I could not reveal the propriety apparatuses and demonstrations of my inventor associates, but only the principles and concepts that developed. In turn, I also developed principles and concepts to explain what they were doing and the results they were obtaining.

So over the years I have slowly been releasing the principles and concepts. Some of them are my own discoveries, many of them are the discoveries of my associates. Some of them are simply a mixture of both.

Early on, it became obvious that the Soviet Union was far ahead on this path, and was already utilizing the new unified field theory to build eery, powerful new superweapons.

Since no one else in the U.S. seemed to be "watching thin particular store" (I was rather universally regarded as some peculiar sort of fool!), I also began to compile information and data on the Soviet weaponization of this unrecognized technology. This information I have released in a series of papers, briefings, and books, the most recent being a 1-hour videotape, "Soviet Weather Engineering Over North America," 1985, and a detailed book, Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons, Tesla Book Co., Greenville, Texas, 1986.

Building upon Whittaker's important work, I formulated a conceptual revision to electromagnetics, which I dubbed *scalar* electromagnetics to accent that the observable EM vector force fields did not exist as such in vacuum, but dynamic scalar fields did. I also wished to call strong attention to the fact that observable force does not exist until an observable particle of mass is coupled to the interference of the two scalar fields (much like in the Aharonov-Bohm effect). The Soviets, of course, call this area *energetics,* Energetics technology has been used in gigantic weapons programs of the Soviet Union for decades, and it appears to be developed under the most highly classified program that the Soviet possess. All development and deployment of energetics weapons is under the KGB and controlled directly by that organization, not by the Soviet Armed Forces.

Peter Kapitsa, the great Soviet physicist, was once pressed by Nikita Khrushchev for a total defense against missiles and air- and space-borne vehicles. Kapitsa replied that it could only come from the new energetics. In 1960, of course, Khrushchev gleefully announced to the Presidium that a new, fantastic Soviet weapon was in the development, "so powerful that, if unrestrainedly used, it could wipe out all life on earth."

Ironically, Khrushchev "jumped the gun" before his new superweapons were deployed. In the fall of 1962 he began inserting long range missiles into Cuba, bracketing the U.S. with nuclear firepower in an attempt to immediately change the balance of power. Kennedy, of course, backed him down "eyeball to eyeball," so to speak, in a blunt confrontation, but promised not to invade Cuba.

Khrushchev, with his days numbered, was desperate to deploy his new superweapons and provide a dramatic demonstration to recover face.

By destroying the U.S.S. Thresher on April 10,1963 and, on the next day, producing a gigantic underwater explosion 100 miles north of Puerto Rico, the Soviets demonstrated that the new superweapons had been deployed. Khrushchev managed to retain his position a while longer.

In the 1960's and early 1970's, I was also deeply involved in the study of paranormal phenomena.

In 1969, I entered the Georgia Institute of Technology to pursue a Master's Degree program in nuclear engineering, graduating in 1971.

In 1973, I published a rather simple paper, "Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena," in which I pointed out the nature of quantum change, gave a new definition of mass and acceleration, and pointed out the fundamental nature of inversion of time. The paper also contained a simplified derivation of Newton's laws of motion, relativistic form. The elements of this paper had been worked out in 1971 while I was finishing my Master's program in nuclear engineering. Finishing the work had been interrupted by a slight sidetrack — a tour in Vietnam from summer of 1971 until summer 1972.

At about the same time, I formulated a fundamental correction to Aristotle's logic, adding a fourth law of logic to Aristotle's three, and a proof of it. The new logic was of great use in discovering and uncovering new concepts in unified field theory.

Incorporating Kaluza-Klein 5-dimensional concepts, scalar EM became a field theory that unifies electromagnetics and gravitation.

Incorporating dynamic sum-zeroed EM vector systems (which are discarded in normal EM theory) allowed the direct engineering of the unified field theory, including structuring the vacuum, curving local spacetime, and producing effects at a distance and in higher dimensions. Actually it allowed the recovery of much of the scalar part of Maxwell's original theory.

I then realized that, inside a vector zero EM force field summation/multiplication, the virtual particle flux of vacuum/spacetime was ordered and controlled locally and macroscopically. This of course violated one of the major assumptions (a postulate) of quantum mechanics; the assumption that the structure of vacuum was randomized, and could not be deliberately ordered, engineered, and curved locally.

Adding phase conjugation (time reversal) aspects and extended quantum mechanical concepts allowed local antigravity and local curvature of spacetime to be included — again, on an engineering basis. It also allowed one to produce a mechanism responsible for Newton's third law, and to engineer the reaction force at will.

Further, it revealed that the law of entropy was simply the *positive time* statement; it showed that there was another half of the law, the *negative time* part or the law of negentropy.

In addition, a startling new concept of mind, thought, life, biofields, disease, and healing emerged from all this — again, on an engineering basis. As we stated in the beginning of this book, it is now an urgent necessity to release my work on the basis for electromagnetic disease and electromagnetic healing. We must produce a very quick, positive treatment and cure of AIDS and other coming lethal viruses before the world is decimated.

Accordingly, this work is being released in this book.

In this chapter we will next present some perhaps surprising material on phase conjugation, from the scalar EM viewpoint, after first briefly explaining symmetry and parity.

In following subsections, we will cover briefly the remaining major concepts in scalar electromagnetics. This will then set the stage for the following chapter, Extraordinary Biology, in which we will deal with the basis for unparalleled electromagnetic healing.

Symmetry and Parity

The basic idea of symmetry is the arrangement of the parts of a body or system about an axis so that two or more parts appear the same with respect to some operation.

The most obvious example is to look in a mirror, where we notice that our image has been reversed, left to right. Yet otherwise there is no difference; and so we may say that the reflection has "mirror symmetry." It's the same except that left and right are reversed.

If you know the details of a system at one point, and at another point you know that the system will possess mirror symmetry, then you can predict exactly what state the system will be in at the latter point. It will be "left to right reversed." You don't have to calculate the laws of reflection and the laws of motion of the system.

All the laws of nature that possess "mirror symmetry" have a special property: If the words "right" and "left" are interchanged in the statement of the law, then the behavior of a system that obeys the law is unchanged.

Thus symmetry became a powerful idea in physics. For a system with a great number of possible interactions, etc., one can usually eliminate a very great number of them due to symmetry considerations. In other words, if one can assume that symmetry of the system will not be violated, then all the possible interactions that would yield "broken symmetry" can be eliminated.

On the other hand, if symmetry is broken in a physical interaction, at least one of the conservation laws is broken in that interaction. Broken symmetry — and hence violation of individual conservation laws — is a well-established fact in particle physics today.

Another concept intimately associated with mirror symmetry is parity. Each particle is assigned a number (+1 or -1), depending upon what kind of particle it is. One adds up these numbers for an assembly of particles, to obtain an overall "parity" number for the system. If parity is conserved in an interaction, then this total number does not change in the interaction.

Table 13. SYMMETRY

- **SOMETHING IS SYMMETRIC IF IT REMAINS UNCHANGED UNDER A CERTAIN OPERATION**

- **A SPHERE IS SYMMETRIC WITH RESPECT TO ROTATION ABOUT ITS CENTER**

- **A CATHEDRAL ARCH IS LEFT AND RIGHT SYMMETRIC ABOUT A VERTICAL LINE THROUGH THE CENTER**

- **LAWS OF ELECTRICITY ARE SYMMETRIC WITH RESPECT TO REVERSAL OF POSITIVE AND NEGATIVE CHARGE**

Table 14. A FORCE FIELD IS A SPECIAL LOCAL SYMMETRY

"..THE FORCE OF GRAVITY IS SIMPLY A MANIFESTATION OF AN ABSTRACT SYMMETRY — A LOCAL GAUGE SYMMETRY — THAT UNDERLIES THE PHYSICS OF THE WORLD."

"ALL FOUR FORCES OF NATURE CAN BE GENERATED IN THIS WAY."

PAUL DAVIES, SUPERFORCE.
1984, p. 115

Table 15. GRAVITY AND LOCAL SYMMETRY

- **THE** LAWS OF PHYSICS CAN BE MADE SYMMETRIC EVEN UNDER LOCAL GAUGE TRANSFORMATIONS

- A GRAVITATIONAL FIELD IS INTRODUCED TO COMPENSATE FOR PLACE-TO-PLACE VARIATIONS

- THE GRAVITATIONAL FIELD IS NATURE'S WAY OF MAINTAINING A LOCAL GAUGE SYMMETRY

- IN THE ABSENCE OF GRAVITY, THERE IS ONLY GLOBAL SYMMETRY

Table 16. HINDSIGHT IS BETTER THAN FORESIGHT

- **FROM THE TWIN REQUIREMENTS OF**
  - **SIMPLEST LOCAL GAUGE SYMMETRY**
  - **LORENTZ-POINCARE SYMMETRY OF SPECIAL RELATIVITY**

- **ONE CAN CONSTRUCT OR INFER**
  - **MAXWELL'S EQUATIONS**
  - **ALL THE LAWS OF ELECTROMAGNETICS**
  - **THE EXISTENCE OF RADIO WAVES**
  - **THE POSSIBILITY OF MOTORS AND GENERATORS**
  - **ETC.**

- **IN REALITY THESE WERE FIRST DISCOVERED EXPERIMENTALLY**

Table 17. WHEN SYMMETRY IS VIOLATED [BROKEN]:

- **A "NON-OBSERVABLE" TURNS OUT TO BE AN OBSERVABLE**
- **AN INVARIANCE IS BROKEN**
- **A CONSERVATION LAW OR SELECTION RULE IS BROKEN**

COMMENTS:

- **A VIRTUAL OBJECT BECOMES AN OBSERVABLE OBJECT**
- **LOCAL SPACETIME BECOMES CURVED, AT LEAST TO SOME POTENTIAL INVOLVING THAT OBJECT**

Table 18. SOME PROVEN ASYMMETRIES

- **POSITIVE AND NEGATIVE SIGNS OF ELECTRIC CHARGE**
- **TIME REVERSAL**
- **RIGHT/LEFT HANDEDNESS**

Table 19. SPONTANEOUS SYMMETRY BREAKING

PRESENT VIEW

1. IF 4-MOMENTUM $K\mu = 0$,

   $\varnothing_{(X)} = \varnothing_{VAC}$

   WHERE $\varnothing_{(X)}$ IS AS COMPLEX
   AS ANY SPIN-0 FIELD
2. **FOR** VOLUME Ω » RELEVANT MICRODIMENSION.
   $K\mu \neq 0$ BUT $K_M \sim 0$
   $\langle \varnothing_{(X)} \rangle \neq \varnothing_{VAC}$ [INSIDE *Q]*
   $\langle \varnothing_{(X)} \rangle \neq \varnothing_{VAC}$ [OUTSIDE Q]
3. INSIDE Ω, SYMMETRY PROPERTIES DIFFER
   FROM THOSE OUTSIDE Ω
4. T. P AND CP MAY BE SLIGHTLY ASYMMETRIC
5. CPT REMAINS INTACT

SCALAR EM VIEW

1. IF ARTIFICIAL POTENTIAL $\varnothing_{ART} >> 0$, VOLUME Ω
   MAY BE SMALL
2. $\varnothing_{ART}$ MAY BE ENGINEERED IN
   - SPATIAL SIZE
   - LOCATION
   - MAGNITUDE
   - SUBSTRUCTURE
3. SYMMETRY PROPERTIES INSIDE Ω MAY
   - DRASTICALLY DIFFER
   - BE DELIBERATELY ENGINEERED
   - BE PATTERNED
4. T, P, AND CP MAY BE APPRECIABLY ASYMMETRIC
5. CPT MAY BE ASYMMETRIC

Table 20. WHY SUCH DRASTIC DIFFERENCE?

IN PRESENT THEORY

$\varnothing_{(X)}$ • IS USED ONLY AS A PHENOMENOLOGICAL DESCRIPTION
- **MICROSCOPIC STRUCTURE OF $\varnothing_{(X)}$ OF NO CONCERN**
- **ONLY LONG-WAVELENGTH LIMIT OF FIELD IS OF INTEREST**
- **VACUUM ENGINEERING IS ONLY A REMOTE POSSIBLITY**
- **ACTION AT A DISTANCE IS NOT POSSIBLE**
- **LOCAL SPACETIME NOT CURVED**

IN SCALAR EM

$\varnothing_{ART}$ • REPRESENTS A LOCAL CURVATURE OF SPACETIME [LOCAL GENERAL RELATIVITY]
- **MICROSCOPIC STRUCTURE OF $\varnothing_{(X)}$ IS DETERMINISTIC AND COHERED INTO MACROSCOPIC STRUCTURE**
- **ALL WAVELENGTHS OF THE FIELD ARE OF INTEREST**

$\varnothing_{ART}$ WAVES ARE REAL, ENGINEERABLE AND ELECTROGRAVITATIONAL
- **VACUUM ENGINEERING IS A REALITY**
- **ACTION AT A DISTANCE IS POSSIBLE**

Table 21. ASPECTS OF STRONG LOCAL ASYMMETRY

- **PROPERTIES OF AN OBJECT MAY DIFFER APPRECIABLY FOR**
    - **DIFFERENT OBSERVERS**
    - **DIFFERENT DETECTING MEANS**
    - **ONE TIME TO ANOTHER**
    - **ONE POSITION TO ANOTHER**

- **CONVERSATION LAWS MAY BE APPRECIABLY VIOLATED**
    - **ENERGY**
    - **CHARGE**
    - **SPIN**
    - **MOMENTUM**
    - **ANGULAR MOMENTUM**

- **LOCAL SPACETIME IS CURVED**
- **LORENTZ INVARIANCE OF VACUUM IS VIOLATED**
- **MAY BE A LOCAL "SINK" OR "SOURCE "**
- **GRAVITATIONAL/INERTIAL EFFECTS FROM EM**
    - **TRANSLATION BETWEEN VIRTUAL AND OBSERVABLE**
- **ELECTROGRAVITATIONAL SOLITONS**
- **ACTION AT A DISTANCE**
    - **TRANSMUTATION EFFECTS MAY EXIST**
- **SCALAR/PSEUDOSCALAR FIELD TRANSLATION**

For years the idea of conservation of parity convinced physicists that it would never be possible to tell right from left in the universe, Then in 1956, Yang and Lee pointed out a class of reactions where parity need not be conserved. This effect — broken parity — was confirmed experimentally, and so certain parts and interactions of the universe do possess asymmetry.

The result of all this was that particle physicists established that

(1) symmetry can be, and is, sometimes broken, (2) individual conservation laws can be, and sometimes are, broken.

Particle physics, however, concerns itself primarily with microscopic interactions. Most of the delightful new things discovered in the microworld have not yet been applied — by *physicists,* that is — in the macroworld.

If you can successfully apply broken symmetry in a major system, for example, you should be able to violate the conservation of energy law and thus produce a "free energy" engine. Several inventors with whom I work have done precisely that. The techniques and technology are still primitive and they need much further development, but the principle has been clearly established and replicated.

Physics is correct in certifying broken symmetry, and you can do it in large systems as well as in microscopic particle interactions. Of course you also *curve local spacetime* as well, something which Einstein and his followers have assumed cannot be done.

Charge-Parity-Time ***and Negative Energy***

The upshot of all this is that the physicists came up with the idea that what is really conserved is a consolidated thing called Charge-Parity-Time. You don't have to conserve any one or two of these three, but only the combined product.

Here they overlooked one key factor that is very important to our thesis: In the photon interaction, if you reverse charge and time, you reverse the sign of the energy.

That is, a photon that is emitted from a positive charge carries negative time negative energy. It is different from a photon that is emitted from negative charge, for that one carries what we have conventionally established as positive energy.

Thus physics has not considered the exclusive use of negative-energy / negative-time photons, even when some of its interactions produced them. Instead, it has just lumped together photons and antiphotons, and considered them both to have positive energy.

Antimatter systems don't have positive energy, for example; they have negative energy. *Here we don't use the positive and negative signs to indicate the direction the energy is traveling, but the fundamental kind, of energy that it is—time forward or time reversed.*

For example, consider the local conservation of energy. As conditoned, one almost always unconsciously thinks of the conservation *of positive* energy, without so stating. Yet one is perfectly free to build a system into which he inputs 100 joules of positive energy, and from which he outputs 300 joules of positive energy and 200 joules of negative energy!

For the real conservation law, that's the equivalent of putting in **100 j**oules of positive energy and outputting 100 joules of positive ***energy.***

Now negative energy will run motors, light lamps, etc. —often much better than positive energy. It can be transmitted, transformed, stepped up or down in "voltage," rectified, received, etc. It can be translated into positive energy or vice versa.

Circuits run "cool" when using negative energy. If you short out a circuit of negative energy so that a violent discharge occurs, it produces cooling instead of heating, greatly lessening the danger of fire and destruction. If you mix negative energy onto a line that is carrying the same amount of positive energy, however, the two negate each other and there is suddenly "no power" on the system.

Think of what you were taught in simple arithmetic. You can take a zero — the absence of any single specific number — and replace it with any set of multiple numbers whose algebraic sum is zero. The presence of this special set of numbers you choose does not violate the condition that the zero means that you could not use just a single number.

Although zero is the absence of just a *single* thing, it can be the presence *of multiple* things.

The same thing works with energy. In the microworld, everything is fantastically energetic, and wildly fluctuating. For example, in a copper wire carrying a current of one ampere, wild momentary fluctuations at extremely high frequencies are occurring.

According to classical electrodynamics, the instantaneous current in a single one of these micro-micro fluctuations may reach thousands of amperes, and the instantaneous voltage may reach 10^{40} volts or more. As can be seen, the instantaneous energy and power are enormous. Yet outside the wire, we see a very placid and peaceful situation.

TIME IS NOT OBSERVABLE

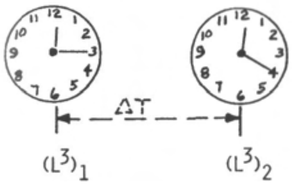

WHAT THEN IS TIME?

WHERE THEN DOES TIME EXIST?

Figure 33. The riddle of time. Unsolved by both physicists and philosophers.

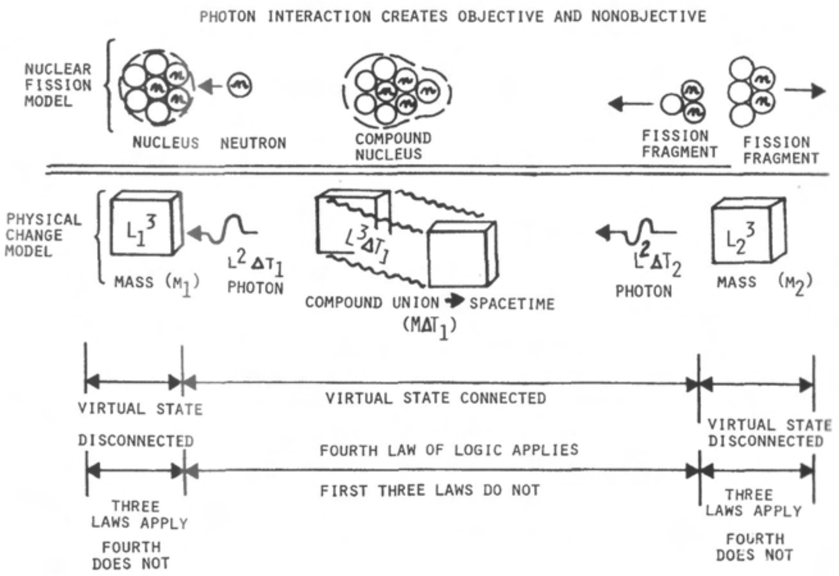
PHOTON INTERACTION CREATES OBJECTIVE AND NONOBJECTIVE
NUCLEAR FISSION MODEL
NUCLEUS
NEUTRON
COMPOUND NUCLEUS
FISSION FRAGMENT
FISSION FRAGMENT
PHYSICAL CHANGE MODEL
L_1^3
MASS (M_1)
$L^2 \Delta T_1$
PHOTON
$L^3 \Delta T_1$
COMPOUND UNION ➔ SPACETIME
$(M \Delta T_1)$
$L^2 \Delta T_2$
PHOTON
L_2^3
MASS (M_2)
VIRTUAL STATE DISCONNECTED
VIRTUAL STATE CONNECTED
VIRTUAL STATE DISCONNECTED
FOURTH LAW OF LOGIC APPLIES
FIRST THREE LAWS DO NOT
THREE LAWS APPLY
FOURTH DOES NOT
THREE LAWS APPLY
FOURTH DOES NOT

Figure 34. Raindrop model of physical change.

Table 22. FLOW OF TIME

- **PHOTON INTERACTION CONSTITUTES TIME CHANGE**
  - $\rightsquigarrow \langle \text{NEGATIVE CHARGE} \rangle \rightarrow + \Delta T$
  - $\rightsquigarrow \langle \text{POSITIVE CHARGE} \rangle \rightarrow - \Delta T$
- **CHARGE-PARITY-TIME [CPT]**

  $\rightsquigarrow \langle e \rangle \rightarrow + \Delta T$ **(note: ~ = photon)**
  - **UBIQUITOUS**
  - **PREDOMINANT**
- **ANTIPHOTON IS A PHASE CONJUGATE REPLICA**

Table 23. PHYSICAL CHANGE

- **COMPOSED OF ACTION**
  - **h = BASIC QUANTUM**
- **TWO CANONICAL VARIABLES**
  - **ENERGY/TIME (dE dt)**
  - **MOMENTUM/LENGTH (p L)**

    ETC.
- **ENERGY/TIME**
  - $h = (+E)(+t)$
  - $h = (-E)(-t)$

Free Energy and Antigravity

Let's say the current flowing in the wire is alternating current at a frequency of one megahertz. Let's say we have one watt of power in the wire (there's one volt on it, and one ampere of flowing current). As a working rule-of thumb for engineering use, the available absolute micropower in an alternating current in a wire can be taken as:

$$P_a = 40 \times f \times P_c \qquad (4\text{-}10)$$

where P_a is the absolute power available, in watts; f is the frequency in Hertz; P_c is the conventional power available, in watts; and 40 is a constant of proportionality.

In our example, the absolute power actually available in the current is 40 million watts! We've got about 20 megawatts of positive power, and about 20 megawatts of negative power, in there — *in the microstructure of the electrical current itself--but disintegrated*

By applying unconventional scalar EM engineering techniques, we ought to be able to take such an "input" (a straight wire with one watt of power at a frequency of a megahertz) and get an "output" of up to 40 megawatts. It is relatively straightforward to get out 1 megawatt; such ratios have actually been achieved on the laboratory bench by my associates.

Of course we will have curved the local spacetime to do such a thing. But it is absolutely permitted to do that, by the laws of nature as best we know them from particle physics.

It is also possible to do that by Maxwell's original theory. And it has been proven in the laboratory.

Table 24. NEGATIVE ENERGY

- **CAN BE**
    - **GENERATED**
    - **TRANSMITTED**
    - **TRANSFORMED**
    - **RECTIFIED**
    - **RADIATED**
    - **TRANSLATED**
- **EASILY POWERS**
    - **AC MOTORS**
    - **DC MOTORS**
    - **LAMPS**
    - **TRANSFORMERS**
    - **RECTIFIERS**
- **CIRCUITS STAY COOL**

In scalar electromagnetics, we deliberately seek to produce such effects and systems by nonclassical application and extension of the modern conservation law CPT. And we include negative energy/ negative time production in those systems and effects.

Here's the extended magic rule: First we write down the expression **CPTEGS.**

In this expression, C stands for charge, P stands for (space) parity, T stands for time, E stands for energy, G stands for gravity, and S stands for entropy.

We are referring here to the algebraic sign of each term, and to a complete photon interaction (absorption and re-radiation) with a charged particle inside an atom of matter. Specifically, we are refferringto the result, after the reaction has occurred.

Normally the photon interacts with one of the electrons in the atom's outer shell. This interaction with the negative electron charge *produces* the "normal" physical world/physical reality we see around us. In the interaction, normal parityis produced, andnormal gravitation (attraction ofmass) is produced. Also, in a series of such interactions, the normal entropy (movement from order to disorder) is txperienced and produced.

So in the expression CPTEGS normally the charge term is negative, and every other term is positive, for the complete photon interaction in an atom.

Now here is the magic engineering rule: In the complete photon interaction with a charged particle in an atom, if one of these terms is reversed in algebraic sign, all the others will be reversed also.

For example, if the photon interacts with the positively charged nucleus,then parity is reversed, the bit of time-advance is reversed, a negative/time-reversed photon is emitted, a tiny bit of antigravity is produced in the nucleus, and negentropy is produced. By negentropy, we mean that the emitted photon exhibits the strange characteristics of time reversal, and hence it will return from disorder to order.

This rule is actually a statement of phase conjugation and time reversed waves, as we shall see later.

For example, using the rule, the production of antigravity is straightforward. Simply produce a great excess of negative energy and negative time in an object or generator by scalar photon interaction with the nucleus, at a sufficient rate.

In negative time, gravitational force is reversed. Thus in negative time gravity is a repulsion, not an attraction. Production of excess negative time in the atomic nuclei of a system produces antigravity in that system. Period.*

---

*At 100 Hz about 500 negative watts/lb is required for flotation. At 1 GHz, the effect is so *weak it cannot be measured, and no amount of power will float* the mirror.

Let me amplify that statement. There are two kinds of energy, negative and positive. Normally, everyone thinks only of positive energy.

Take Einstein's formula for the amount of energy that's "bottled up" and stored in mass:

$$E = mc^2 \qquad (4\text{-}11)$$

That's positive energy and positive mass. It automatically implies that we are looking at things where time is forward-going.

Now suppose we "charge up" that mass with negative energy: Then the mass is negative, like this:

$$-E = (-m)c^2 \qquad (4\text{-}12)$$

And negative mass is repelled by positive mass. (Like mass G-charges attract, unlike mass G-charges repel.)

That's all there is to it! Simply charge up the protons/neutrons of the atomic nuclei of a mass with excess phase conjugate (negative) energy/time, and it exhibits antigravity compared to the earth. Nothing could be simpler.

Again, it's been done in the laboratory by an associate.

So, with respect to a negative energy/negative time generator, you just add more load and draw more negative power from the machine, forcing it to produce more. It produces more negative time, and hence more antigravity and more "repulsion." When the repulsion balances the earth's attraction, the system "floats in air." Draw still more power by adding still more load, and it accelerates upward. Draw less power by decreasing the load, and it sinks downward again. That's all there is to it.

We should have been working upon this, shortly after the U.S. Civil War.

Again, there exist two completely different kinds of electromagnetic energy: positive and negative. These have been totally and hopelessly confused by modern physics, even though physicists sometimes use these terms.

But in dealing with CPTEGS in the photon interaction in an atom, we must clearly keep in mind what kind of energy we are dealing with.

The basic engineering rules are these: (1) If any one of the factors - C, P, T, E, G, or S — is reversed, all the others are reversed. (2) If positive charge is involved in the photon interaction, then time (carried by the emitted photon) is negative and so is the energy carried by it. (3) The simplest scalar EM wave may be considered as two EM waves locked together (modulating each other), where one component wave carries positive energy and time and the other component wave carries negative energy and time. (4) Separation of the two components by interaction with charged particles creates both positive and negative time and energy. (5) If the interacting particle is negatively charged, the positive energy/positive time component will interact with it. (6) If the interacting particle is positively charged, the negative energy/negative time component will interact with it. (7) The interacting particle phase conjugates (time reverses) the component with which it interacts. (8) The primary interaction with negative electron shells of atoms yields positive energy and positive time: constituting ordinary physical reality as we observe it. (9) The secondary interaction with the positive atomic nucleus produces phase conjugated (time-reversed) physical reality. (10) Secondary interaction (nuclear phase conjugation) normally produces Newton's third law of motion. (11) In Newton's third law, the reaction force need not be equal, and need not be antiparallel. Note, however, in such case local spacetime is curved. (12) The present law of entropy is only the positive-time half; the other (negative time) half is the negentropy law.

Phase Conjugation (Time Reversal)

It is a most remarkable fact that Soviet radar engineers and radar scientists are also trained in optics. While this seems little short of astonishing by Western standards, the Soviets have an exceedingly good reason: As a matter of course, most Soviet radar systems are also operable in, or associated with another system that is operable in, a time-reversed (phase conjugate) mode.

In this mode the radar and/or its associated system is also an extremely powerful directed energy weapon, having capabilities undreamed of in the West.

A phase-conjugate (time-reversed) wave actually is a wave that travels backward through time.* That is, it is capable of precisely retracing the path through space, taken by another wave that traveled that path to a nonlinear mirror, stimulating the reflection of the time-reversed wave.

Further, in retracing its invisible path through space, the phase conjugate replica wave does not diverge as do normal waves. Instead, it continually converges upon its invisible trace.

Consequently all its energy arrives back at the distant source that emitted the stimulus wave originally. This was Tesla's original secret of his "wireless transmission of energy at a distance with no losses."**

It can easily be shown, however, that the phase conjugate wave is a legitimate solution to the wave equation. Hence it is a general property of all types of waves: sound, magnetohydrodynamic, etc. It is not confined just to the optical band.

Obviously such a startlingly different wave has phenomenal uses. Literally it ushers in a new physics.

Let us turn now to look at this strange wave that has emerged on our scientific horizon.

In the late 1960's, Soviet researchers cautiously pointed out the possibility of a time-reversed wave. Papers appeared in the open Soviet scientific literature, and shortly thereafter U.S. physicists began to read about it.

In 1972, Soviet scientists visited Lawrence Livermore Laboratory and specifically briefed U.S. scientists on the optical phase conjugation (OPC) phenomenon. Thereafter the effect became of increased interest to American scientists.

(Note, however, that phase conjugation (PC) is a major phenomenon of all physics. It is not confined merely to the optical spectrum, nor just to electromagnetic waves. Phase conjugation is a general phenomenon true of all waves, regardless of type. We will primarily discuss OPC, since most Western experimentation has been in the optics domain, and the literature is rich in that respect. However, we also point out that Western scientists have absolutely no inkling of the mechanism causing OPC. They have simply written a model for a time-reversed wave by taking the equation of a normal wave und writing the conjugation (adding an asterisk exponent) to the appropriate term, and pointed out that this new equation also satisfies Maxwell's wave equation. They also do not clearly recognize that an EM wave carries time and energy, and that a time-reversed EM wave carries negative energy and negative time. Further, they do not at all comprehend the startling implications for generating gravitational fields, including antigravity and inertial effects, directly on the laboratory bench.)

---

*The external positive-time observer sees this time reversal as length (space) reversal.

**The great electrician had actually discovered what today is known as phase conjugation or time reversal of EM waves.

In optical phase conjugation, when an input EM wave (ordinary) enters a nonlinear medium, the medium "reflects" or produces a strange, time-reversed EM wave in response.

Figure (35a) shows the effect. In the figure, E1 is the normal input wave, entering a distorting, nonlinear medium (the "blob") as shown. Wave E_1 moves on through the blob, emerging on the other side and continuing, though now in distorted form.

The medium produces a second wave, E_2, which precisely retraces the steps of E_1 That is, everywhere E_1 is, E_2 now appears. The phases of the E_2 wave are precisely reversed from those of E_1 The only difference in the two waves is that wave E_1 carries positive time and positive energy, while wave E_2 carries negative time and negative energy.

(U.S. physicists are well aware of everything in the above statement except for the negative energy portion.) The historical background of the phase conjugate wave is given in Table 25.

(We strongly stress that the West did not discover the time-reversed wave; the Soviets did. At the time (late 1960's) this appeared in the open Soviet literature, the Soviets had already been using it for about two decades in their giant energetics (scalar electromagnetics) weapons development program.)

However, to return.

The time-reversed wave is called the phase conjugate replica of wave E_1. The overall characteristics of the phase conjugate replica E_2 are summarized in Table 26.

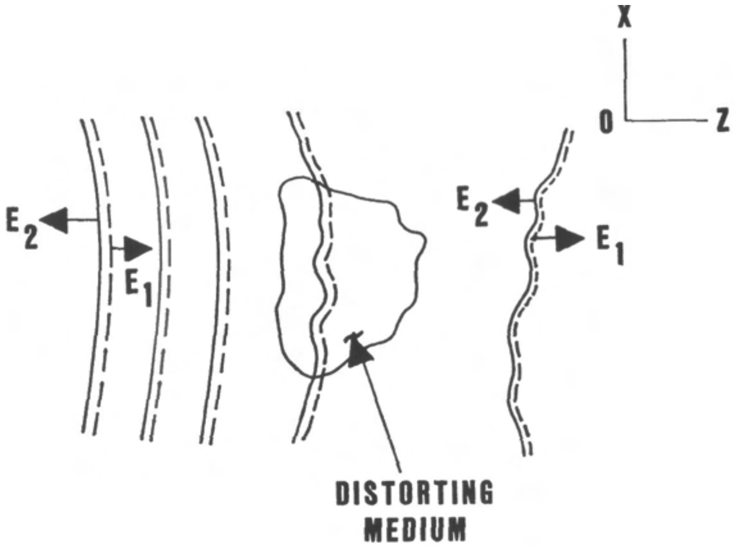
X
0
Z
E2
E1
E2
E1
DISTORTING MEDIUM

a. Phase conjugation by a nonlinear, distorting medium.

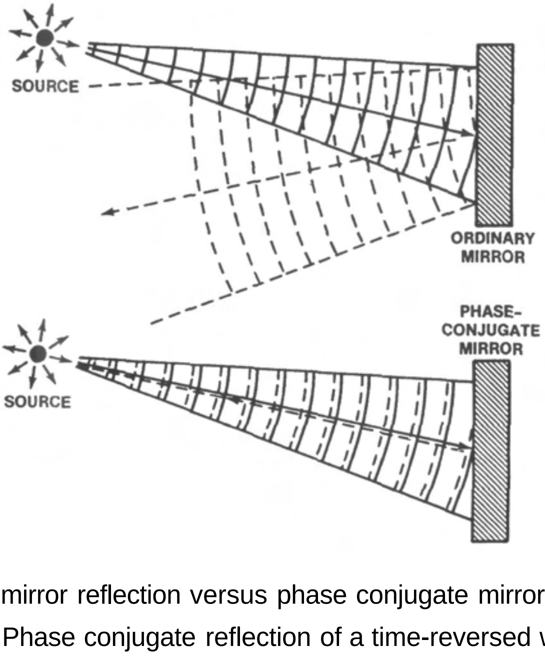
SOURCE
ORDINARY MIRROR
PHASE-CONJUGATE MIRROR
SOURCE

b. Ordinary mirror reflection versus phase conjugate mirror reflection.

Figure 35. Phase conjugate reflection of a time-reversed wave.

Table 28. PHASE CONJUGATE WAVES

- NONLINEAR OPTICS TRANSIENTS
  - DISCOVERED BY SOVIETS
  - LATE 60s-EARLY 70S
- **PRESENTLY** "HOT ITEM"
- INITIALLY THOUGHT RARE
- NOW KNOWN WIDESPREAD
- THIRD POWER OF E-FIELD
- S**TILL** MUCH UNEXPLORED

Table 26. PHASE CONJUGATE REPLICA

- TIME-REVERSED
- CARRIES NEGATIVE TIME
- CARRIES NEGATIVE ENERGY
- "HEALING" PROPERTY
- DISTORTION CORRECTION
- F**ROM** DISORDER TO ORDER
- **MAY** DWARF INPUT WAVE

Table 27. DISTORTION CORRECTION THEOREM

"If a scalar wave E_1 (r) propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near z=0] its phase conjugate replica E_2 (r) , then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1."

We point out here that the basic nonlinearity (nonlinear medium) in spacetime is a single particle of mass. In a photon interaction with that particle of mass, the interacting (input) photon (positive energy, positive time) causes the generation and emission of a phase-conjugated (time-reversed) photon. This time-reversed photon provides the precise mechanism for generating Newton's third law: **For every action there is an opposite and equal reaction.** This will be covered later.

But back to our nonlinear medium, and the two waves E_1 and E_2.

E_1 and E_2 are "locked together" as a scalar EM wave unless something else is specially done to separate them. That is because the nonlinear medium is a *modulator,* and the two waves modulate each other (lock together as a single wave).

(We interrupt again to strongly point out this fact: This directly shows that all EM waves in vacuum exist as scalar EM waves, not as the so-called Maxwell-Hertz transverse waves of conventional theory. Nikola Tesla, the greatest electrical scientist the world ever produced, always adamantly insisted that EM waves were "sound" type waves, not "transverse" waves. For one thing, the vacuum-medium would have to be a solid to produce transverse waves! Tesla even visited Hertz in Europe and tried to convince him of the error of his so-called "proof of transverse EM waves." Tesla found Hertz, adamant on the subject, however, and returned without changing Hertz's mind. Ironically, Tesla was right and all the textbooks in the Western world have been consistently in error as to the nature of the EM wave in vacuum.)

We simply state here that, for a single-frequency EM sine wave in vacuum, a "full photon" of that wave in vacuum is actually a *single sine-wave oscillation.* The positive (positive energy, positive time) half of that sine wave is the same thing as a virtual electron (negative charge). The negative (negative energy, negative time) half of that sine wave is the same thing as a virtual positron — a virtual electron travelling backwards in time, or "time-reversed", which we detect as length-reversed and charge-reversed.

Further the vacuum — including the vacuum inside each relatively empty "mass structure" such as an atom or molecule — is a seething inferno of activity of virtual particle fluxes. It is upon that medium that all physical reality and physical changes are impressed. Thus there is nothing static in the world; what actually exists is *switching of virtual particle fluxes.* What we call a "static thing" is only an equilibrium in this switching activity. And, once switched on, any stream continues until switched off or switched to change it.

For that reason, the full photon has always contained twice as much absolute energy as is provided by the presently used equation

$$E = hf \qquad (4\text{-}13)$$

Where E is energy, h is Planck's constant, and f is frequency.

One half the photon carries negative energy and negative time, while the other half carries positive energy and positive time. In the full photon interaction, one half of the photon interacts and "switches off" the stream. Thus the photon interaction makes a single change of action, energy x time, both switching on the beginning of the quantum change and switching off the ending of it.

The full photon in vacuum is, and has always been, a scalar EM photon. It rigorously is the presently accepted photon plus its directly "hooked on" and accompanying phase conjugated replica photon.

That is, in terms of the present physical view, the scalar electromagnetics view of a full photon in vacuum is that it consists of a photon and its antiphoton, directly combined in serial fashion. Further, this antiphoton part is not identical to the photon, since it is time- and energy-reversed.

The present physics makes no distinction between the two types of photons, hence makes no distinction between positive and negative photon energy. It also still considers "static" things to independently exist. It does not pursue the fact that all things appearing "static" are equilibrium states in nearly unbelievably intense virtual particle streams. Since it does not pursue that fact, it does not build a "switching" model for quantum change, not does it arrive at a method of directly engineering these "switching" operations.

For that reason, present physics has totally missed local general relativity and the engineering of spacetime curvature by electrogravitational methods, as was contained in

Maxwell's original theory.

For the same reason, present Western physics has not worked out the basic mechanism that produces phase conjugation, and it has not realized that all conservation laws can be locally violated with the new methods — methods actually hailing back to Maxwell's original theory!

Now returning again to our nonlinear medium and waves E_1 and E_2:

It is possible to separate (demodulate) wave E_2 from wave E_1 In that case, E_1 is an ordinary EM wave, carrying positive energy and positive time. E_2 is a phase conjugate replica, carrying negative energy and negative time.

Irradiation of an atom in a new sample of material with a normal EM wave such as E_1, just provides the normal photon interaction with the electron shells. That is, the electron part of the full photon adds to the orbital electron's energy state in positive time, raising the orbital electron to an "excited state." (Remember, this has switched on a continuous process of "keep raising the energy level of the electron" in positive time.) The "state-increasing" orbital electron then interacts with the second, positron-part of the full photon, which switches off the increasing, leaving the electron in its "static excited state." Simply put, one half of the full photon turns on an uplifting spray, and the second half of the full photon turns off the uplifting spray, leaving the electron in equilibrium state at a new level.*

Thus in orthodox physics we have taken the view that an electron in a shell absorbs a "photon" from the incident wave, taking extra energy and being raised into an excited energy state.

In the orthodox view, the excited-state electron then decays to release another ordinary photon (since it comes from a negative charge).

What actually happened was that, when the "full photon" of the incident wave is absorbed by the orbital electron, increase of the electron energy is switched on, then switched off, leaving the elec

---

*We state here a magnificient principle without proof. When one furnishes (inputs) electrical power to a system, all the "input work" is expended purely upon switching fields by fighting against them in brute-force fashion. E.g., electrical power furnished to an electric motor does **not** run the motor, but only continually wrestles the internal fields to switch them. Obviously, to build an over-unity device, one must get it to switch its own fields, at least a significant fraction of the time.

tron in equilibrium at a new energy level.

What happened in the middle was phase conjugation. In absorbing the first half of the incident full photon, the electron acted as a phase conjugate mirror (we will explain that term shortly). It emitted a forward-travelling phase conjugate replica (PCR) of that first photon half, and this conjugate replica entered the nucleus, where it is absorbed to cause the Newtonian reaction force. Note that this PCR emission by the electron does not generate recoil in the electron-PCM, as is well known in OPC theory and experiment.

The same type of phase conjugation happened when the state-increasing electron absorbed the second half of the incident full photon. It phase conjugated again, sending this half to the nucleus to acompany the former PCR half. In the nucleus, this second half switches off the increase in excited state of the nucleus that was engendered by absorption of the first half.

The photon interaction is a process for switching on and off a continuous rate of action increase or decrease.

At the conclusion of the full photon absorption, both the orbital electron and the nucleus are in excited states. The nucleus has increased its gravitational mass.

Then the situation precisely reverses. Both nucleus and electron decay, again emitting full photons and again phase conjugating.

The PCR photon emitted by the nucleus reduces the positive mass of the nucleus. The PCR photon emitted by the electron is reversed from the electron's absorption record, reducing the negative charge's energy state, and the electron is said to "decay" from its excited state. The PCR photon emitted by the electron now (to the observer) will appear to carry positive time and positive energy.

(In passing, we point out that, actually, the situation is far more complicated. Both the electron and the positively charged nucleus are enmeshed in continuous streams of virtual particle flux, including intense streams of virtual photons. One must go to n-wave mixing, where n is greater than 4, to fully understand the total reaction. The main points are, the Newtonian reaction forces within the atom are generated by phase conjugation, the photon emitted by the decay of an excited state electron will normally carry what we call positive energy and positive time, and the photon emitted by the decay of an excited state positive charge will normally carry what we call negative energy and negative time. Full coverage of these interactions is beyond the scope of this book, and must wait for another effort in the future.).

Now once again we return to our simplified model of phase conjugation, and waves E_1 and E_2 in our nonlinear, modulating medium!

As we stated, we can engineer the situation so that the phase conjugate replica E_2 can be separated from E_1. A separated E_2, however, is a quite different breed of cat. It is a phase conjugate wave, carrying negative energy and negative time.

E_2 is a magic wave indeed!

Irradiation of the nucleus of an atom in a new sample of material with a phase conjugate replica such as E_2 causes the nucleus of the atom to interact in reverse of its normal interaction.* That is, one of the positive charges in the nucleus interacts with a photon resulting from wave E_2, and goes into an excited "negative energy" state. That state, in decaying, emits an ordinary virtual photon which promptly interacts with a virtual positive charge in the nucleus, which decays to emit the virtual photon, which interacts with another positive charge in the nucleus, etc.

In other words, the negative energy of the absorbed antiphoton is captured by the nucleus, and the nucleus overall does not *immediately* decay to discharge the negative energy.

Instead, the nucleus acts as an accumulator or capacitance, charging up and then discharging far more slowly than the normal photon/electron interaction.

(For the purist, the nucleus does phase conjugate the input negative energy/negative time wave, and phase conjugates it again — upon an electron in the electron shells, through spin coupling This second phase conjugate replica is now just a normal EM wave, and the electron exhibits a normal EM interaction. Thus **our** instruments (Which are electron wiggle detectors) detect a normal

---

*Irradiation of the nucleus with a PCR wave is easily accomplished by using ordinary magnetic resonance, and modulating the PCR wave upon the magnetic carrier, as Antoine Priore did.

EM wave detection and do not discriminate between photon and antiphoton halves of a full photon. For that reason, scientists have not clearly noticed the difference between positive and negative energy, and between photons and antiphotons. They have also not resolved the basic conflict between the particle and wave theories of electromagnetics. *Note that the full photon concept, where the full photon is a normal photon and phase conjugate replica of that photon hooked together into a single sine wave, resolves the long-standing wave-particle controversy.)*

In most of their instruments, present scientists are detecting a secondary normal EM wave for the phase conjugate replica interaction, since the phase conjugate interactions in the instrument are not sorted out, and since the instruments measure only one kind of PCR, that emitted by a negatively charged electron. Hence scientists think the antiphoton has positive energy, and interpret it in positive time. Consequently they tend to accent that the wave is "phase-reversed," and not accent the negative time/negative energy implications since they have not clearly sorted them out.

Let us return to our consideration of the irradiation of the nucleus of an atom (with an incident E_2 phase conjugate replica wave carrying negative energy and negative time, with respect to our electron-wiggle-detecting observer).

AS irradiation continues, the actual "structured charge pattern" existing on the incident E_2 wave slowly charges up the nucleus of the atom with that exact charge structure, in the negative energy/ negative time state.

We shall find that most interesting fact of great utility when we later discuss the mind, thought, cellular control systems, and generating a specific electromagnetic phase conjugation pattern to reverse cellular disease.

We shall also find that it is of great interest in accomplishing antigravity when we use extremely low frequency PCR irradiation.

The Phase Conjugate Mirror

Any system which phase conjugates the input wave and returns a phase conjugate "reflection" is called a phase conjugate mirror, or PCM for short.

But the phase conjugate mirror has some very unique properties, as can be seen in figure 35b.

For example, if you look in a *normal* mirror, you see your own reflection. You see your shoes and legs, for example. Light from your shoe scatters across the mirror, and it hits one spot on the mirror which reflects it at the proper angle to strike your eye.

But that simply doesn't happen with *a phase conjugate* mirror! If you were looking into a phase conjugate mirror (PCM), the light that left your shoe and scattered across the mirror would be reflected right back to the shoe, since it would retrace the path taken by the previous light striking all across the PC mirror from the shoe. So all you would be able to see would be two black dots: the retinas of your eyes. Light reflecting off them and striking the PCM would be reflected back to the source — the retinas — so you could see that light and that light only.

This path reversal is most important to keep in mind. With a phase conjugate signal, you are not functioning with an ordinary signal. Further, it has a "precursor" path taken by the incident wave that stimulated its emission from the PCM. Unless interfered with, the PCR will trace back down the invisible vacuum trace of that incident wave, or back down its wave path, like electricity goes down a wire.

This, by the way, is the secret of Tesla's wireless transmission of energy at a distance with no losses.

When broadcast into space, a normal EM wave (positive time, positive energy) diverges or scatters from its path.

Precisely the opposite happens with a PCR wave. When broadcast (reflected) into space, the PCR wave continually converges back upon the invisible trace taken by its incident stimulus precursor. Thus all the PCR energy is continually converging upon an invisible beam or "wire" through the vacuum, back to the source of the original incident wave.*

If the PCR wave is highly amplified (and means to do this are now well-known), then large amounts of EM energy can be precisely

*One can effectively move the entire imprinted vacuum path, however, by adding an additional input vector wave computed so that the **resultant** input would have come from the desired shifted distant point. In this way the PCR can be steered in space to lead and intercept a distant moving target.

returned to distant points from which a stimulus wave of any kind is received.

If a micro-microwatt is received from a point thousands of miles distant, then a gigawatt can readily be returned precisely to that distant point.

This was Tesla's magic secret. The great electrician had discovered phase conjugation, though he did not use that term. He did, however, point out that he could create an invisible wire through space to a distant point, and could send any amount of energy to that distant point without any scattering losses along the way.

He also accomplished the same thing with sound waves and mechanical waves. He pointed out that he could produce an undiminished physical (mechanical) effect at any distant point, and could eventually split the earth if he added power long enough.

Tesla had given the principles of radar in World War I! He later even spoke of his "big eye that can see at a distance." Here he was probably referring to stimulating the emission of a PCR return from that target. The PCR would automatically correct for the distortion in the intervening medium, and fairly sharp pictures of the distant scene — without showing the size dispersion of distance — could probably be obtained. Note that a non-scattering signal to a distant observer does not make him suffer any loss in field or apparent size and detail of the scanned object, which can be recovered to any resolution, limited only by the state of the art of the physical realization technology.

The implications for far greater resolution microscopy— far heyond the present limitation to about 4,500 — should be immediately obvious. The implications of such a microscope - that can see directly inside atomic nuclei; photograph nuclei and fundamental particles directly; see living viruses, molecules, and atoms directly; and theoretically see even into the virtual state itself— should be immediately obvious.

In discovering phase conjugation, the Soviet Union was attempting to break Tesla's secret of wireless transmission of energy without losses, and the Soviet scientists did so by about 1950. They highly weaponized the effects of phase conjugation, phase conjugate mirrors, and amplified phase conjugate mirrors, both in radar waves, mechanical waves, and sound waves, for nearly two decades. Then the Soviets cautiously stimulated the open literature with a limited paper on optical phase conjugation, to ascertain whether or not we realized the severe weapon implications of time-reversed waves, their production and use, and their amplification.

So phase conjugate mirrors can do very useful things. For example, look again at figure 35a. Suppose we are trying to photograph something on the other side of the distorting medium, say through a distorting gas, such as a turbulent atmosphere. In that case E_1 on the left is, say, light coming from an object on the left and passing through that distorting medium. If we just used ordinary waves in a camera on the right, we would see a very distorted wavefront, as represented by the distorted E_1.

However, if we illuminate the object by sending a wave through the distorting medium, and detect the *phase conjugate* signal that returns, it will have reversed the distortion when it gets to us, and we will get a clear picture. A very nice photo of just this process is contained in David M. Pepper, "Applications of Optical Phase Conjugation," **Scientific American,** 254 (1), Jan. 1986, p. 75.

Phase conjugation has many more uses than we have covered in this brief paper, of course, but this is an important use: the removal of distortion effects from optical systems.

It would, however, be nice to be able to amplify the effect. And so we can.

Amplifying the Phase Conjugate Effect

Let's look now at figure 36, where we will develop how to amplify the phase conjugate replica.

Be careful to notice that we are changing subscripts on you. By convention, optical scientists use A_1 and A_2 as two additional waves, called the *pump waves,* that are used to stress the nonlinear medium, and input signal wave A_4. Wave A_3 then is the stimulated PCM output, or the phase conjugate replica (PCR) wave.

In this figure we show a scheme for amplifying the phase conjugate signal that is generated and returned by the phase conjugate mirror (PCM).

We do this by adding two opposing waves of the same frequency (in this simplified case). These are the two "pump waves", and impressing them upon the PCM is called "pumping" the PCM.

Actually, in producing the pump wave, we are adding two input waves of the same frequency, 180 degrees out of phase with each other. Since the nonlinear medium is a *modulating* medium, the waves are forced to modulate each other and "lock together" as a *single scalar* EM wave of most interesting characteristics. Indeed, its E fields sum to a zero resultant vectorially, and its B-fields sum to B zero resultant vectorially. However, the scalar parts remain and are multiplied together, as in the original Maxwell quaternion theory. Since the magnitudes of the vector components in the scalar summation (quaternion multiplication, since we are addressing modulation) are varying, then the magnitude of the remaining scalar part is varying. We have therefore produced a standing scalar wave in the nonlinear material medium, having zero vector EM gradient components. We have produced a purely scalar EM wave of pure potential — and this is an electrogravitational wave, rigorously, since it represents a time oscillation of the local energy density (local virtual particle flux density) of vacuum. This scalar EM wave passes through the electron shells and enters the nuclei, where it is phase decoupled by the extreme nonlinearity of the violent virtual particle currents of the nucleus. Its energy is then absorbed by the nucleus, raising it to an excited pseudopotential. This potential decays, returning the scalar EM stress wave into the local vacuum and area. We are now pumping the atomic nuclei of the nonlinear PCM material with our two pump waves, and we have created an oscillating local gravitational field around the PCM.

This increased nuclear G-potential, looked at in Kaluza-Klein theory, is a 5-space potential. It is ready to burst out in the 5th dimension (as electromagnetic bleedoff) at the slightest provocation. That provocation is the entrance and absorption of the input signal wave, E_4. Note that E_4 acts as an "initiator" (or a pinprick into a highly inflated balloon, if you will) on the built-up, excited 5-dimensional gravitational pseudopotential of the nuclei. The moment so initiated, the entire 5-space G-pseudopotential collapses, emitting a time-reversed phase conjugate replica (PCR) of E_4. This emitted, highly amplified PCR is labelled E_3. It now "backtracks" down the invisible trace through space taken by the incident stimulus wave E_4, returning to the distant source that originally emitted

Voila! Tesla's wireless transmission of energy to a distant point without losses. *Voila!* The Soviet Launch Phase ABM system and the Soviet Launch Phase Anti-Bomber System, when used in conjunction with an over-the-horizon radar (such as the giant Wood-

Table 28. FOUR-WAVE MIXING

- **NONLINEAR MEDIUM**
- **OPPOSING PUMP WAVES**
  - **CROSS MODULATION**
  - **SCALAR WAVE**
  - **INTO NUCLEI**
- **TRANSMISSION WAVE**
- **REFLECTION WAVE**
- **REFLECTION GAIN**

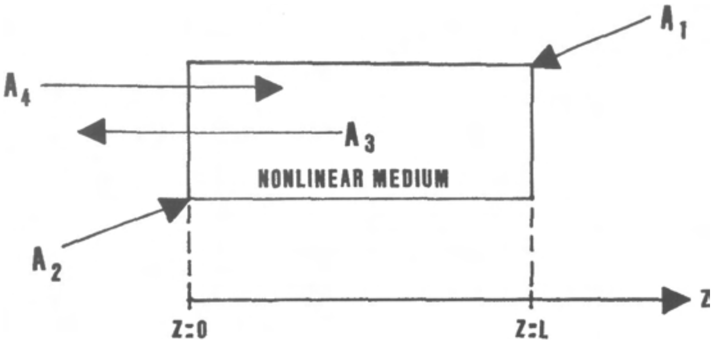

Figure 36. Phase conjugation by four-wave mixing.

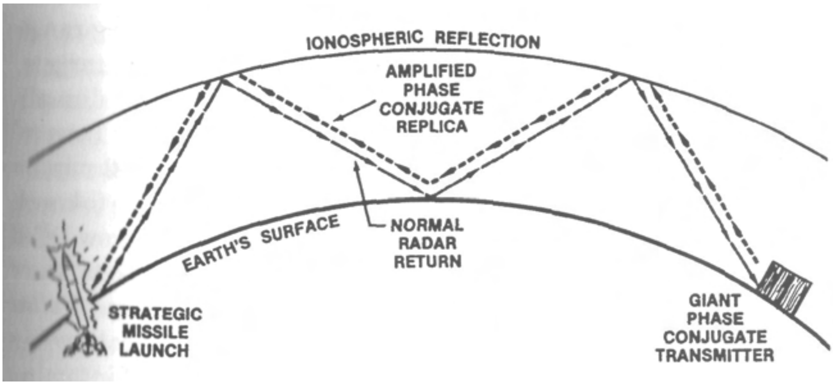
IONOSPHERIC REFLECTION
AMPLIFIED PHASE CONJUGATE REPLICA
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC MISSILE LAUNCH
GIANT PHASE CONJUGATE TRANSMITTER

a. Soviet Launch-phase and midcourse ABM system.

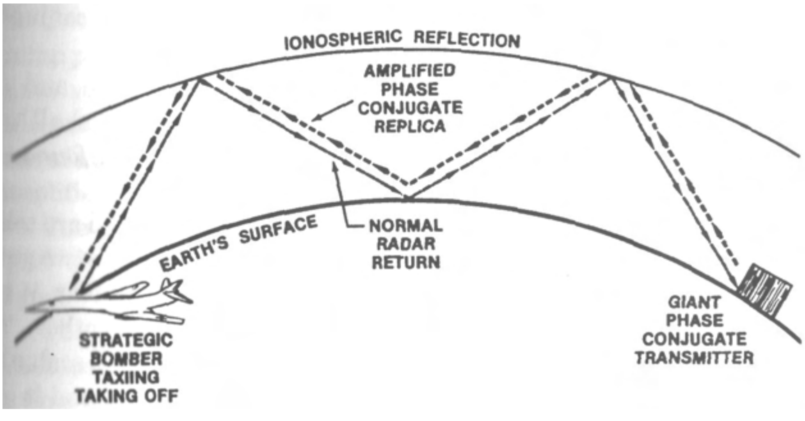
IONOSPHERIC REFLECTION
AMPLIFIED PHASE CONJUGATE REPLICA
EARTH'S SURFACE
NORMAL RADAR RETURN
STRATEGIC BOMBER TAXIING TAKING OFF
GIANT PHASE CONJUGATE TRANSMITTER

b. Soviet launch-phase and midcourse anti-bomber system.

Figure 37. Soviet launch-phase and midcourse strategic defense system.

peckers) to provide a reflection signal from a distant target. *Voila!* The true use of Soviet Stand-Off-Jammers to knock out enemy antiaircraft, air control, and anti-tactical missile radars wholesale. *Voila!* The Intermediate Phase ABM system, where a long range strategic radar is used to provide reflections from incoming targets. "Fire and fire again!" is the rule, and 100% defense is directly achievable. *Voila!* How the Soviets knocked down Gary Powers' high-flying U-2 reconnaissance aircraft. *Voila!* How SA-2 missile system radars were temporarily modified in North Vietnam to knock down several hot new F-111 aircraft. *Voila!* How the Arrow DC-8 aircraft, Titan missiles, Delta rocket, and Ariane rockets were "electromagnetically surged" from inside to destroy them. *Voila!* Why the Soviets almost never throw away old radars, since they can now add PCMs and make them formidable destroyers of incoming targets. *Voila!* How one solves the ABM problems of discrimination and weapon kill; simply pulse-fire everything — nuclear material explodes in a full nuclear explosion when struck with potential pulses of sufficient magnitude and duration.

Summarizing figure 36 and amplification of the phase conjugate replica:

When we consider the addition of the input wave and the resulting phase conjugate wave that will be produced, we shall have a total of four waves. Thus this particular process is called four-wave. *mixing.**

On the diagram in Figure 39, A_1 and A_2 by convention are taken to be the two opposing EM pump waves. Notice that these two pump waves stress the medium. Since the medium is nonlinear, it is a modulator. Hence the two pump waves modulate each other. The two waves thus lock together into a scalar EM wave. This scalar EM wave is an oscillating artificial scalar EM stress wave, and it goes directly into the atomic nuclei of the medium and pumps the nuclei themselves.

The atomic nuclei are now rhythmically stressed with the oscillating scalar wave.

---

*Probably the best single introduction to optical phase conjugation in the English language is David M. Pepper, "Nonlinear optical phase conjugation," **Optical Engineering,** 21(2),**Mar.**/ Apr. 1982, p. 156-183. Especially read Pepper's footnote on p. **166.**

We now input a weak signal (wave A_4) into the "pumped nuclei" medium.

By our CPTEGS rule, the positively- charged nuclei generate a negative energy/negative time wave — in short, a phase conjugate replica wave. Further, it is a highly amplified PCR, and may contain *as* much raw energy as was fed into the pumped mirror by our pump waves. This "seeker wave" or "electromagnetic missile" wave then sets out on its invisible path through space, seeking a distant "A_4 source point" and delivering all its energy intact to that point.

The Phase Conjugate Mirror as a Vacuum Triode

The pumped phase conjugate mirror (four wave mixing) can best be understood as a special kind of vacuum triode.

The pumped artificial potential on the atomic nucleus provides a powerful cathode in positive time, and a powerful plate as well in negative or reversed time. The electron shields, in their spin coupling to the nucleus, act as a grid in positive time. The input signal A_4 itself acts as a relatively weak signal onto the grid of the triode.

The positive charge difference in potential between the powerful pump potential of the nucleus (cathode, in forward time) acts as the plate voltage. The decrease of this potential when discharging (increase of the negative time, negative aspect) acts as the production of a plate signal to the negative time operation (production of the amplified, time-reversed PCR wave).

The excited nucleus immediately decays upon stimulus by the "grid signal" A_4, emitting a "time- and energy-reversed" strong signal — in short, amplified phase conjugate replica wave E_3.

The amplified phase conjugate signal moves from the cathode (nucleus) out to the electron shells and beyond, precisely because it is time-reversed. To a time-reversed signal, the grid (normal input) acts as the final plate (external output) — that is, in negative time, the elements of the triode are reversed. This makes it possible for the material to emit the powerful PCR.

I have called this effect/analogy a vacuum triode, to accent the direct engineering possibilities. My associates have already engineered this effect to obtain enormous amplification of energy ("free energy") and practical antigravity, from solid-state devices, directly

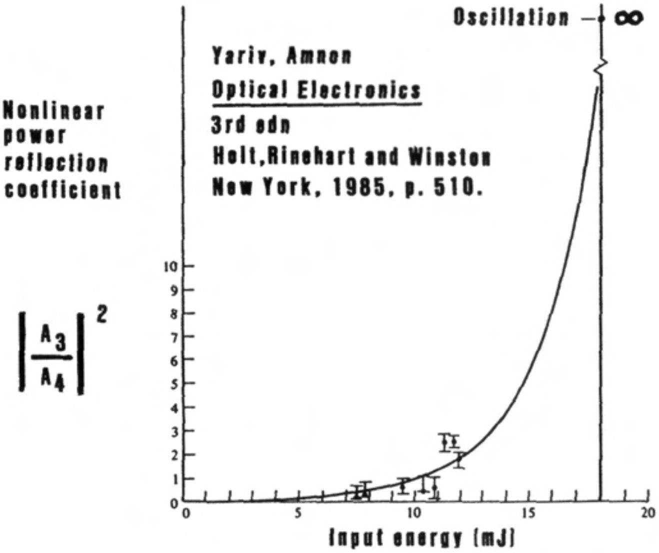

a. Typical nonlinear power reflection coefficients.

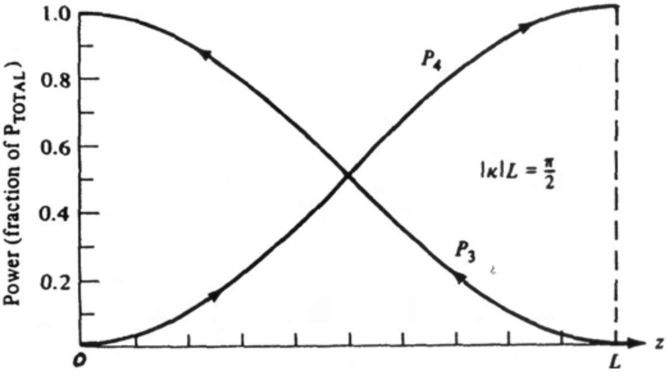

b. Intensities for the oscillation condition.

Figure 38. Power reflection coefficients and oscillation.

on the laboratory bench.

The actual "amplitude gain" of the vacuum triode is the ratio of the amplitude of A_3 to the amplitude of A_4. The square of the amplitude gain is proportional to the power gain; this entity is called the nonlinear power reflection coefficient.

Figure (38a) shows the typical nonlinear power reflection coefficient for a pumped phase conjugate mirror, as a function of the input energy and the depth of the material. As the input signal increases, the gain increases. The ultimate condition would be that the "free oscillation" condition is approached. This occurs when the angle between the input wave A_4 and the pump waves A_1 and A_2 is 90 degrees.

Figure (38b.) shows the oscillation condition and the fraction of the total pump power that is in the input wave A_4 and its output **phase** conjugate replica A_3.

At the oscillation condition, an "infinite" gain exists at the correct depth in the medium. This means that essentially all the energy being fed into the pump waves A_1 and A_2 will now be emitted **in** the output wave A_3. At that point, "saturation cutoff" of the triode output occurs, and it cannot output higher energy, since it has no more.

Notice that, at the oscillation condition, for the slightest output, the full power of the pump waves is fed into the phase conjugate replica. This powerful wave then retraces the path taken by the o**riginal** input signal, unless something else is done to divert it.*

The interesting question then arises: Since we can have infinite **gain,** can we get a special oscillation condition so that the pump wave furnishes its own energy? In other words, can the pump wave "resonate" with the vacuum flux in such a fashion that it is self-pumping and self-perpetuating?

In at least three materials, the answer is yes. Such effects have **been** experimentally accomplished in special forms of barium tita**nate.** See, for example, J. O. White et al, "Coherent oscillation by self-induced gratings in the photorefractive crystal $BaTiO_3$," **Appl. Phys. Lett.,** vol. 40, 1982. p. 450.

*Recall that the PCR can be steered by introducing a minute wave A_5 to vectorially sum with input A_4, In that case, the effective input is the vector resultant of $(A_4 + A_5)$. The apparent distant source point and invisible vacuum path are shifted correspondingly.

Phase conjugation is a universal phenomenon. Theoretically it applies to any sort of wave: electromagnetic, scalar, sound, mechanical, etc. By using it, incredible new effects and an entirely new physics can be accomplished. It is a new and exciting field, little-known outside specialist circles, and still not well-understood. Most discoveries in the field are yet to be made.*

Newton's Third Law and the Full Law of Entropy

For example, the present author noticed that Newton's third law— for every action there is an opposite and equal reaction — is actually a result of phase conjugation. To explain that, we diverge a bit.

First, phase conjugation is also involved in the fact that a charged particle of mass is a little "dynamo" or engine.

Actually, a particle of mass is itself a "nonlinearity" in vacuum spacetime. When a photon (virtual or observable) strikes a particle of mass and is absorbed by it, it's the same as "interacting with a nonlinear medium."

In terms of virtual photons, the charged particle is continually bombarded with them, from the vacuum flux. The charged particle, being a "nonlinear medium", phase conjugates each input virtual photon, emitting a phase conjugate replica.

If it absorbs a normal virtual photon, it emits a "negative energy, negative time" phase conjugate replica.

If it absorbs a normal virtual photon, it emits a "negative energy, positive time" phase conjugate replica — just a normal virtual photon. Note that the particle of mass thus does not exist in time. Or, it exists equally in positive and negative time. Since we only observe particles, it follows that time is not normally an observable, as quantum mechanics has assumed.

The emission of a phase conjugate replica does not affect the momentum of the emitting PCM, as is well-known.

For an observable photon, the particle (nonlinear medium) absorbs the normal photon, which does not affect the energy and momentum of the PCM particle of mass.

If there is no other mass to accept the PCR, it just moves off.

*The U.S. may finally be waking up. SDI is now starting to study EM missiles "in the time domain" as opposed to frequency. See **Av. Wk. & Spurn Tech., Feb. 29, 1988, p. 56.**

If there is another mass to accept the PCR, the PCR is absorbed and phase conjugated by it.

A charged particle in vacuum is continually bombarded by a flux of essentially "randomized" fluctuations. However, on the average these all "average out" to zero. Any particle is automatically in a very high vacuum potential, and is in a state of great agitation at all times, even if it's "at rest" with respect to ordinary observers. It's really "smeared all over the place," so to speak.

When we expose a charged particle to a *normal* potential, the interior virtual particle flux of that potential is disorganized. Hence the particle just "wiggles a bit more frantically," so to speak, randomly in every direction.

However, when we expose a charged particle to an *artificial* potential made by zero summed/multiplied EM force fields or waves, the particle wiggles more frantically in *organized directions.**

This is just as true for virtual particles as it is for observable particles.

Thus the use of such artificial potentials in scalar electromagnetics enables us to overcome the disorder of the internal structure of normal electromagnetics. By this means, we can structure the vacuum, organize electricity and electromagnetic waves, accomplish negentropy, curve local spacetime in complex, deterministic fashions , produce free energy and antigravity, reorganize and transmute nuclei (transmute elements) with miniscule input energy, etc.

Rigorously, this means that the present law of entropy is only half of the full law of entropy. Specifically, it is the "positive time, positive energy" half of the law, and it states that in closed real physical systems, continuing operations of ordered "positive time, positive energy" effects tend to inevitable "disordered" effects produced in the apparatus or system.

In other words, the present law of entropy really states that no ordered system is totally closed. The system will inevitably have energy and actions that escape the system's order, and hence (to the system) this escape represents "loss of order" — or in short, disorder.

The present law of entropy specifically excludes (by implication)

---

* I.e., **the** electronic noise is partially organized. This is a major key toward building scalar EM detectors. It is also a primary cause of many solid-state circuit malfunctions. The exact principle: (Random changes) + (artificial potential) equals (chaos).

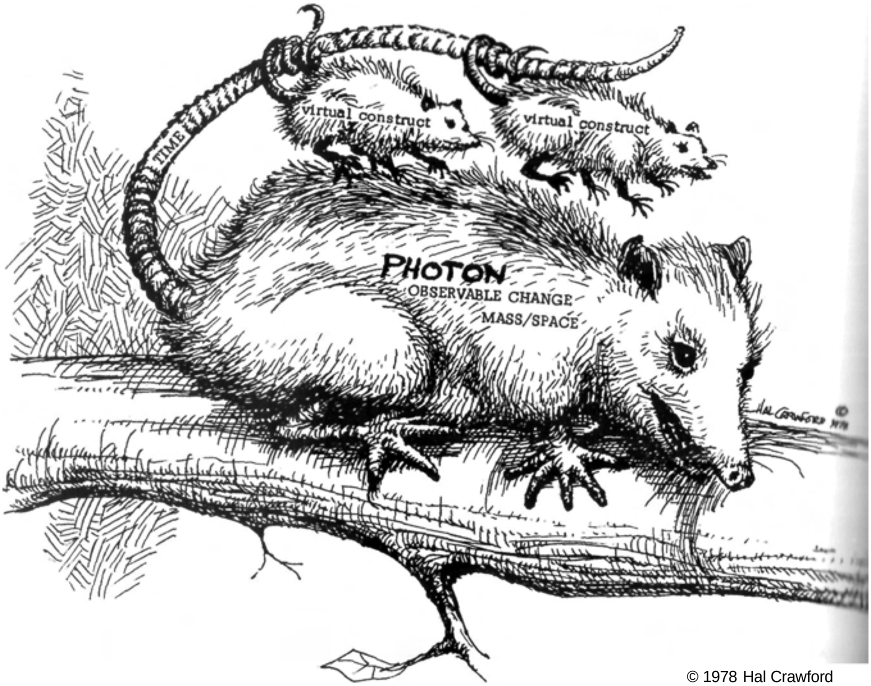

© 1978 Hal Crawford

Figure 39. Photons have virtual substructures. These may be statistical or deterministic.

negative time aspects, where external (disordered) energy time-reverses to again enter the ordered system, in perfect order. By implication, it also excludes negative energy production and operations.

However, **experimentally we can produce both negative energy and negative time, in deliberately time-reversed operations such as phase conjugation and with devices thnt serve as phase conjugate mirrors. Further, we can amplify the phase conjugation and disorder-to-order operations.**

Hence the present law of entropy is incomplete, and states only half the true possibilities. The other half of the law is the **law of negentropy.** In this law, disordered operations outside the **ordered** system undergo time-reversed (and possibly amplified) effects. These reordered energy effects — which may be highly amplified in comparison to the disordered escaped energy—reenter the "ordered system", restoring order. Since this process moves in negative time, it goes from disorder to order. Hence it is negentropic.

Therefore we have solved the old thermodynamic problem of the "eventual decay of the universe into disorder." Not to worry; the other half of the full law of entropy prescribes negentropic operations in negative time/negative energy, and these restore the order of the universe. That takes care of the problem.

It also means that, in highly disordered systems of many degrees of freedom and far from thermodynamic equilibrium (that is, when the disorder law is saturated), then one can expect to see further stress "create" negative time and hence create order emerging and stabilizing from disorder.

And of course that is exactly what is seen. Ilya Prigogine was awarded the Nobel Prize in 1977 for writing a new thermodynamics predicting precisely such effects, and experiments have proved his case.

But to return to Newton's third law:

Newton abstracted his third law from the classical interaction between two colliding balls.

If one ball is at rest and another approaches it at some velocity, the approaching ball carries momentum and kinetic energy. As it "collides" with the resting mass, quantum mechanics tells us what actually happens: Particles of mass in the moving ball are producing virtual photons continually, and these virtual photons strike the particles of mass of the resting ball and are absorbed by it.

The resting mass now acquires extra energy and momentum from the absorbed photons. At the same time, it acts as a nonlinear medium. It produces phase conjugate replica virtual photons, and these time-reversed virtual photons are emitted. These photons produce no change in momentum or energy upon the emitting mass. However, being phase conjugates, they precisely follow back along the paths taken by the first or "stimulus" virtual photons, striking the moving ball. They are absorbed by that ball, producing (to the external observer) negative momentum and negative energy in it.

These subtract from the kinetic momentum and kinetic energy of the moving ball, reducing them.

The virtual photons emitted by the approaching ball perform positive work upon the target ball.

The virtual, phase conjugated photons, emitted in return by the mass of the struck ball as a PCM, reverse to strike the approaching ball to do negative work upon it.

Notice that twice as much absolute work is always done on the system of two balls as we "input" with the approaching ball.

Also notice that, if we directly "engineer" the phase conjugation of the struck system, we can directly tamper with, and drastically change, the production of the negative energy and negative momentum in the moving ball. Hence we can drastically alter Newton's third law. We can now make a Maxwell's Demon.

This alone clearly establishes that it is perfectly possible to build a so-called free energy device. Though subtle mechanisms must be used, it's nothing more spectacular than putting a paddle wheel in a river, and extracting shaft power from the wheel, furnished by the river's current. Phase conjugation yields other exciting possibilities

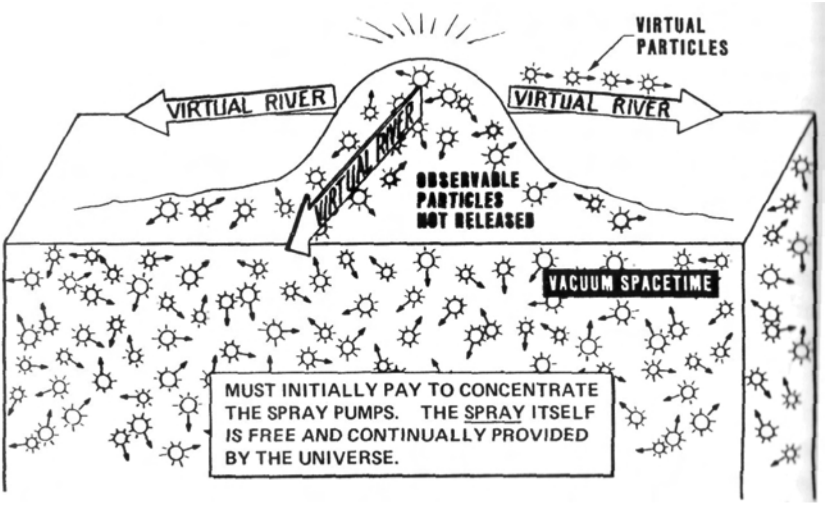
VIRTUAL PARTICLES
VIRTUAL RIVER
VIRTUAL RIVER
VIRTUAL RIVER
OBSERVABLE PARTICLES NOT RELEASED
VACUUM SPACETIME
MUST INITIALLY PAY TO CONCENTRATE THE SPRAY PUMPS. THE SPRAY ITSELF IS FREE AND CONTINUALLY PROVIDED BY THE UNIVERSE.

Figure 40. Once established, a virtual river is for free.

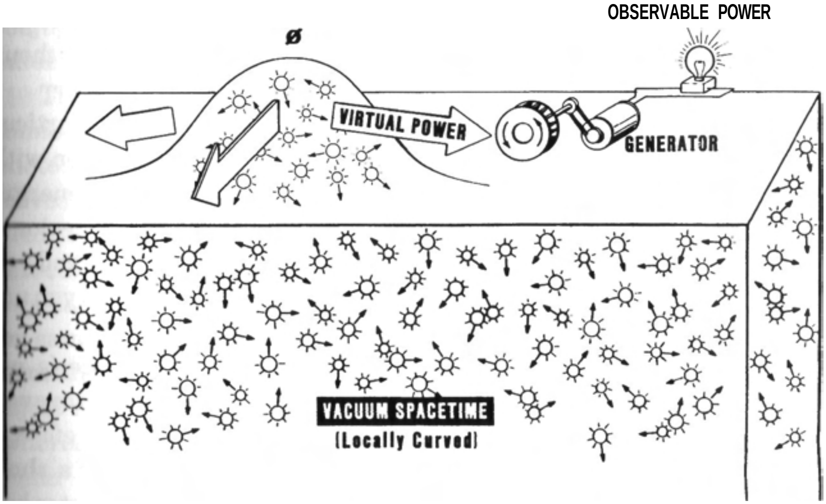
OBSERVABLE POWER
ø
VIRTUAL POWER
GENERATOR
VACUUM SPACETIME
(Locally Curved)

Figure 41. Local energy conservation can be violated.

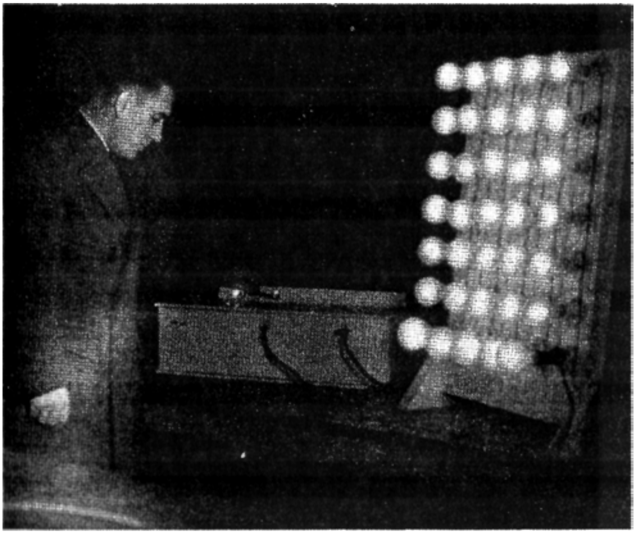

Figure 42. T.H. Moray adjusts the controls of his radiant energy device. Salt Lake City, Utah, in Feb. 1937. Moray obtained 50 kilowatts of power from a 55-lb. device, straight from the ether. Its characteristics (cool circuits, brilliantly lit lamps) reveal that It was negative energy.

too numerous to explore here. For just one example, by phase conjugation, real-time holography can be accomplished without having to separately make and illuminate the holograms.

Now with holography and multiple projectors — and particularly with two intersecting "normal" beams to backtrack upon with ever-converging PCR signals—you can assemble a "form of energy" at an appreciable distance from the projectors in three dimensional space. With phase conjugation, you can do it in real time. And you can do it at a great distance back along the path of any signal you use as an input.

By using PCM holography, degradation with distance is not appreciably experienced. After all, the PCR waves utilized are ever-converging. The energy does not diverge and spread. Therefore, if you can make a 6" diameter ball of intense EM energy at a short distance in the laboratory, you can make that same 6" diameter ball of intense energy at several thousand kilometers distance, using PCR real-time interference holography.

With highly amplified PCM's, you can put far more energy into the distant interference form than you can with normal holography. Conceivably it is possible to assemble so much energy in the holography "object" that it condenses into a real material object! In other words, it materializes or semi-materializes.

Both the "intense light form" usage and semi-materialization of forms (such as light forms, flying geometrical forms, etc.) have been created worldwide by the Soviets for decades, to stimulate UFO reports and provide a deception plan for their development and testing of large, strategic energetics weapons using highly amplified PCMs and PCR holography.*

Other bizarre considerations also arise. For example, in **radionics** one utilizes a "witness"—such as a photograph—of the object one is attempting to send "energy" to. (For example, this might bo a sickly plant one is attempting to spruce up). Now when the plant was photographed, the light from it struck the photoemulsion of tho film, producing a photochemical reaction. It also produced phase conjugation in the atomic nuclei of the film emulsion, since the emulsion acts in some respects as a nonlinear medium.

---

*For an example of highly reliable detection and observation of "Soviet hololography UFOs," see Leif Havik, "Project Hessdalen," MUFON UFO Journal, Jan. 1988, p. 4-7. The present author certainly does not imply that all UFO phenomena are due to Soviet PCR holography, only that some UFO phenomena are.

There exists a continuous, invisible "stimulus photon and its connections" trace in time and space, between the plant when the photo was taken and that photo now, even if one or both of them have moved. That trace is highly persistent (as Australian experimenters Reid and Barsamian have shown). A phase conjugate (time-reversed) signal can be made to retrace that track and travel all the way back to the plant, in the present. The effect of using the photo (or sample of the object) as such a "witness" of the real plant and directing weak PCR energy back along the invisible trace into the object is called radionics. In most cases, the two cerebral hemispheres of the operator's brain are depended on to form the necessary scalar "pump wave" to pump the "witness" and produce phase conjugated signal energy, time-reversed to travel back into the distant object.*

The problem with present radionics is that it is far too weak, and it is "operator dependent", changing with the mood, temperament, and skill of the operator. It does objectively work, however, though Often erratically. In countries such as England, France, and Germany, use of radionics to treat human disease is legal and permitted, though operators (for example, in England) must undergo extensive training and certification. Its results are roughly comparable to the results of normal medical doctors. Radionics does not replace orthodox medicine. It can successfully treat certain things that orthodox doctors have little or no success with (such as lupus), and orthodox medicine can successfully treat many things that radionics has little success with (a broken leg or ruptured appendix, for example). The two treatments are best used in conjunction with each other.

However, we must point out that a new, amplified PCM radionics is likely to emerge in the near future. Such a new approach will offer highly effective, powerful treatment of diseases, even at a distance, by amplified PCR electromagnetics. Again, the radiation in and from a photo or witness of a diseased patient can be used as a "grid signal" into a powerful, pumped phase conjugate mirror. The

---

•Again recall that we observe time-reversal of waves as **spatial path** reversal, not as time **travel.**

PCM will produce a specific, powerful, amplified signal that will travel directly to the patient, be absorbed in his entire system, including the atomic nuclei comprising his cells, and directly reverse the chemistry, toxins, and cellular damage in the patient from that disease. It will also kill or destroy the harmful viruses, bacteria, etc. causing the disease, and it can even reverse the actual genetic changes in the cell caused by disease-inducing viruses.

Of course, any powerful tool is two-edged. It can be used for harm as well as good. I think the reader can appreciate the damage that can be done to systems and persons at a distance by means of such powerful new amplified PCM radionics devices.

You can also see the security problems posed to large computing facilities in banks, federal facilities, state and municipal facilities, the IRS, U.S. military weapon systems, ships, aircraft, etc.

Obviously such devices are going to have to be highly regulated and controlled. Else criminal operators will be killing people and destroying things wholesale, secretly and at a distance, and they will have perfect legal alibis to allow them to escape prosecution.

To summarize: if a photograph or witness pattern of an object or person can be correctly stimulated by phase conjugate electromagnetics so that the emulsion acts again as a PCM, can one not argue that—theoretically — phase conjugate energy from the photograph and the instrument can be sent back along all paths, all the way back to the original object? After all, this has been rigorously demonstrated by University of Sydney researchers. Is there not an actual physical mechanism for radionics? For highly amplified, powerful radionics where the operator himself is removed from the circuit? Just because it changes our present notion of physical reality, must we dogmatically reject the results being shown on the laboratory benches, some of which are in the hard-core scientific literature?

Phase conjugation and time-reversed engineering have a diversity of applications unparalleled in the history of science. They literally present us with a new physics. New rules apply.

Most of the major discoveries in phase conjugation have yet to be made.

Scalar Electromagnetics

The scalar electromagnetic approach modifies classical electromagnetics (EM) to include gravitational waves and effects.

To do this, scalar EM utilizes summed-zero vector forces and force fields to construct polarized vacuum potentials.

Although externally it has zero E and B vector gradients, an "artificial" potential possesses an internal, dynamic E and B vector field structure. This "infolded" (Bohm's term) structure has finite size and is deterministic. The infolded structure of the artificial potential in turn deterministically structures and polarizes the vacuum and curves spacetime locally, contrary to presently assumed limitations of general relativity.

Since in free vacuum a potential normally extends to infinity in a decaying exponential fashion, then this infolded E and B vector structure of an artificial potential extends to infinity in a decaying exponential fashion.

The artificial potential itself may be rhythmically varied either in structure, magnitude, frequency, or all of the above. This produces waves of potential, and waves of the structure of vacuum/spacetime — again, in the free vacuum case, reaching to infinity in a "decaying exponential" fashion. These "scalar EM waves" are gravitational waves.

Each photon (one major wave length of the "carrier", complete with modulations) of the scalar EM wave also has structure (its included modulations).

Modulated waves have compound or "giant" photons — photons containing infolded photons.

Each compound photon of such a scalar EM wave is a vacuum engine. It deliberately structures and patterns — dynamically — the energy density and charge of vacuum.

(1) By canonically varying two or more components of the electromagnetic vector structure of the zero-summed EM force vector system, the local structure of vacuum spacetime is macroscopically varied in its internal composition. This is the compound variation of "curvature within curvature" — and hence hyperdimensional.

Table 29. A ZERO VECTOR SYSTEM:

- **IS AN ACCUMULATOR**
- **DOES "INTERNAL" WORK ON MEDIUM**
- **HAS NO "ENERGY" EXCEPT STRESS OF VACUUM, YET**
- **CAN YIELD THE ENERGY OF ITS COMPONENTS WHEN DISSIPATED**
- **IS A SPECIAL SORT OF "PUMP"**
- **CAN YIELD CONTINUOUS ANENERGY FLOW IF NOT DISSIPATED**

(2) By prohibiting internal canonical variance while coherently varying the amplitudes of all the EM components, the total stress of vacuum/spacetime — and hence its overall curvature — is locally varied, without modifying its structural form. This is simple variation of overall curvature of local spacetime, involving primarily relativistic effects.

Rigorously, each of these two methods produces a localized gravitational wave, where the local stress of vacuum is deliberately patterned as well as oscillated in amplitude.

Both methods may be applied simultaneously (and multiple infolded times) to produce an even more sophisticated gravitational wave, and more sophisticated structuring of the local stress energy density of vacuum.

The resulting unified electromagnetics/gravitation is called **scalar electromagnetics,** since the electrogravitational effects are obtained by deliberately opposing EM vector force fields so that they vectorially sum or multiply to a zero vector EM resultant, while **the** infolded (Bohm's term) EM vector components structure and vary the stress energy density of vacuum.*

*Gravitational potential is just "in-folded and locked in" dynamic electromagnetic forces. The infolded EM energy is locked-in, representing a change in the local energy density of vacuum and hence a curvature in spacetime. Electromagnetic force fields are just the out-folding of the G-potential's inner EM contents into the 5th dimension. Ordinary gravitational force field is just the "trickle leakage" of the G-potential's infolded contents, out into and through 3-space.

Note immediately that scalar EM also deals with internally structured and patterned electricity and electromagnetism.

"Normal" electromagnetics has no deterministic internal structure.

The new electrogravitational (EG) wave is called a scalar EM wave, and is believed to have been originally discovered by Nikola Tesla.

James Clerk Maxwell was aware of the potential for electromagnetics to stress and structure the vacuum ether.

His original electromagnetic theory — written in quaternions, not vectors (which had not yet been completed by Oliver Heaviside) - can allow for these effects to be expressed.

Unfortunately Heaviside's interpretation of Maxwell's work

Table 30. STRESS IS FUNDAMENTAL

- **THE MOST FUNDAMENTAL REALITY IS STRESS**
- **THE COMPONENTS COMPRISING STRESS ARE THE GREAT CAUSATIVE AGENTS**
- **MOST OF THESE ACTIVE AGENTS ARE HIDDEN INSIDE ZERO-VECTOR SYSTEMS**
- **THE COMPONENTS OF STRESS MAY BE STATISTICAL OR DETERMINISTIC**

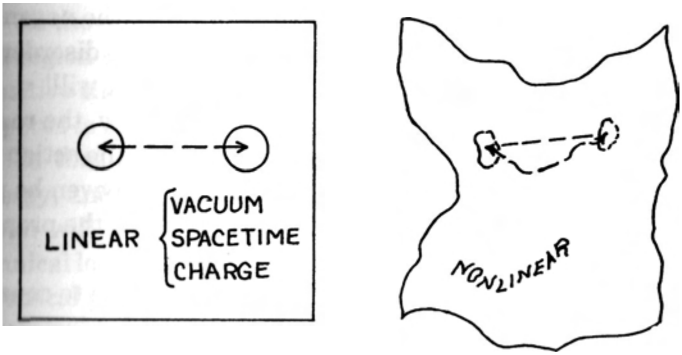

figure 43. Charge affects anything existing in spacetime.

eliminated these effects, and they have remained eliminated from the Western EM theory to the present day.

The use of the scalar EM wave directly engineers the virtual state and the vacuum itself.

Hooper's work represented important work in this respect, as does other related experimental work by Bedini, Watson, Golden, Dea, Faretto, Beck, and other inventor colleagues who do not wish their names mentioned.

Scalar EM theory also bears a strong relationship to the Kaluza-Klein unified theory of gravitation and electromagnetics. At least five dimensions — four spatial and one time — are required as a minimum.

Nested levels of virtual state may be modeled as identical to successive hyperspaces.

Internested levels of zero-summed vector EM force fields allow the direct engineering of those hyperspaces.

Negative energy and negative time flow are especially important when building free energy devices and antigravity. The mechanism producing negative energy and negative time flow is the forced amplified production and absorption of phase conjugated waves.

In addition, phase conjugation is a basic phenomenon of nature: for example, as previously pointed out, it is the direct mechanism that generates Newton's third law of motion. It also adds a negentropy law (for negative energy/negative time).

Accordingly, if we engineer phase conjugation, we can change Newton's third law to our will. And we can invoke the negentropy half of the total law of entropy, so that we can go from disorder back to order. Again, we can make a Maxwell's Demon at will!

In Newton's third law, when manipulated locally, the reaction force need not necessarily be equal in magnitude to the action force It also need not necessarily be parallel to it. It can even be phase conjugated itself to yield an addition force, assisting the propelling of a particular engine or device.

Phase conjugation may also be manipulated so as to cause like magnetic poles to attract, and unlike poles to repel.

Magnetic "force field lines" of a bar magnet normally exist in conjugate co-existing pairs. One of the pair is in observer positive time, and (by convention) runs from the north pole to the south pole. The second line of the pair is a phase conjugate of the first, and exists in negative observer time. Thus, to the "positive time only" observer, it will appear to run from the south pole to the north pole, perfectly retracing the path of its twin.

Howard Johnson uses a "two particle" theory of magnetic field, for example, where one particle (photon) is the phase conjugate of the other. This is identically the same as the scalar electromagnetics concept of phase conjugate pairs of field lines. By very complex, specialized, compound permanent magnets, Johnson is able to partially separate the two lines, and spatially concentrate or diminish the phase conjugation (negative time) component in a given area of operation.

If the negative time component lines are concentrated to outweigh the positive component lines in a local region, then the laws of magnetics are reversed in that region. There like poles attract and unlike poles repel.

In his rotary permanent magnet motor design, he concentrates the negative time component in that part of the rotation where a north rotor pole is approaching a north stator pole. Normally, in this part of the cycle, one has to overcome the repulsion of like poles to "force energy" into the device for later extraction.

However, *now* in that region *like poles attract.* There the north pole of the rotor is attracted toward the north pole of the stator, adding impetus to the rotor.

As the rotating north pole passes the stator north pole, it leaves the region of negative time concentration and returns to a normal "positive time" region again. Now the north poles repel again, and impetus is again added to the rotor.

In that manner Johnson is able to violate local conservation energy, and make a "free energy" permanent magnet motor. The motor will rotate itself and deliver constant shaft power to a mechanical load. By attaching a standard electrical generator, normal electrical power is continually and freely produced by the complete motor generator power system.

Johnson's complex nonlinear magnets severely curve local spacetime, allowing *local* violation of energy conservation.

Scalar EM View of the Vacuum

In the modern quantum mechanical view, the vacuum is not an emptiness, but instead is a plenum.

Today the vacuum is considered to be filled with incredible virtual particle activity. From nowhere, virtual particles continually arise — even with fierce energy — then disappear again into nowhere, so rapidly that they cannot be individually observed.

However, these virtual particles are quite real, for they cause all the forces of nature when they interact in the aggregate with observable particles.

To the observer, the fleeting particles also appear and disappear in both positive time and negative time. That is, the flux consists of both particles and antiparticles.

Thus the vacuum is a seething inferno of virtual particle fluxes.

The concept of an **ether** is again accepted. It refers to this "virtual flux" vacuum.

Note, however, that this is an ether far different from the old material ether that was theorized prior to relativity.

In the new vacuum ether, every imaginable type of particle continually and spontaneously arises (creation) and disappears (annihilation) at every point in the vacuum, according to modern quantum mechanics.

The rate of this seething virtual particle creation and annihilation is essentially unlimited. Hence the "flux density" of vacuum is essentially unlimited.

Further, any virtual particle created has a flux of even finer virtual particles associated, and so on without limit.

The vacuum's virtual particle flux is thus comprised of nested levels of ever finer virtual particle fluxes, in the modern view. We state without proof that **each deeper virtual level may be modeled as a higher dimension** (hyperdimension). **This yields an infinite-dimensional vacuum spacetime (hyperspace) that is identically the infinite nested levels of virtual state.**

At the same time, every nucleus in the universe is continually absorbing and emitting scalar EM (electrogravitational) waves. The emitted scalar EM waves of pure potential are waves in the stress and structuring of the vacuum itself.

VACUUM IS IDENTICAL TO:

- VIRTUAL PARTICLE FLUX
- SPACETIME
- ANENERGY
- ϕ_0
- MASSLESS CHARGE

VACUUM IS:

- MADE OF UNQUANTIZED ACTION
- WITHOUT DEFINITE LENGTH INTERVALS
- WITHOUT DEFINITE TIME INTERVALS
- N-DIMENSIONAL (UNFIXED)

Figure 44. Vacuum/spacetime is pure virtual particle flux.

Table 31. VACUUM IS:

- SPACETIME (L^nT, WHERE $n >= 3$)
- CHARGE (MASSLESS)
- ELECTROSTATIC SCALAR POTENTIAL ($\phi_0 \neq 0$) (ERROR IN PRESENT THEORY)
- BROKEN BITS OF ENERGY (SUBQUANTAL)
- PURE VIRTUAL PARTICLE FLUX
- ϕ -WAVE FLUX
- MULTILEVEL, STRUCTURED, PATTERNED
- A VIRTUAL PLENUM
- AN OBSERVABLE EMPTINESS

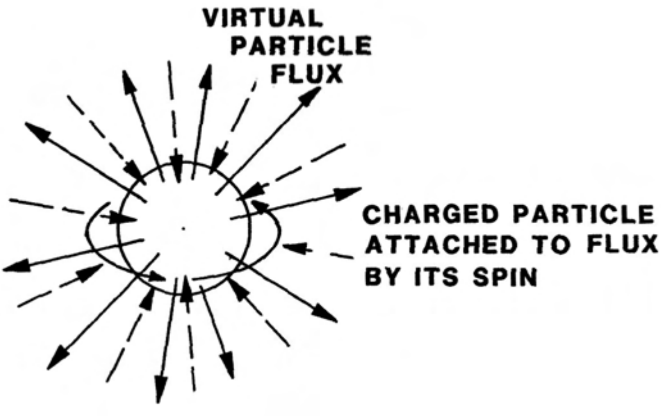
VIRTUAL PARTICLE FLUX
CHARGED PARTICLE ATTACHED TO FLUX BY ITS SPIN

Figure 45. A simplified charged particle.

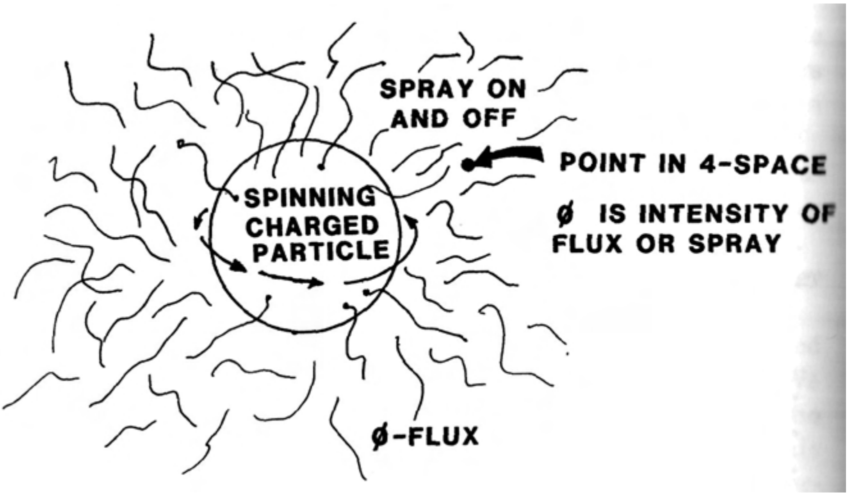
SPRAY ON AND OFF
POINT IN 4-SPACE
ø IS INTENSITY OF FLUX OR SPRAY
SPINNING CHARGED PARTICLE
ø-FLUX

SPRAY IS THROUGH 4 OR MORE DIMENSIONS, NOT JUST 3

Figure 46. A charged particle is a special kind of "spray nozzle."

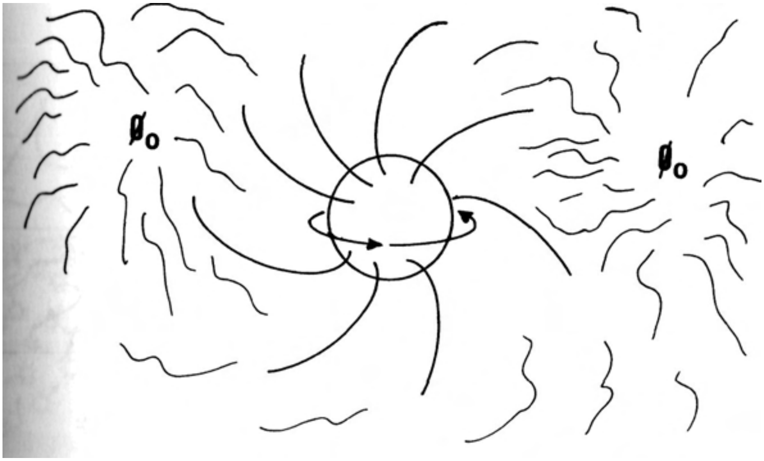

Figure 47. An observable charged particle is directly connected to virtual particle **vacuum** flux.

Table 32. SPIN COUPLES VACUUM TO PARTICLE

- **SPHERICAL ROTATION IS THE KEY**
- **A PARTICLE CAN BE MODELLED AS A SPHERICALLY ROTATING VORTEX OF SPACETIME**
- **ITS MASS IS DUE TO ITS SPIN**
- **THE SPINNING OBJECT IS CONTINUALLY CONNECTED TO ITS ENVIRONMENT**
- **VERY HIGH VALUES OF ELECTROSTATIC POTENTIAL CAN INDUCE RELATIVISTIC CONDITIONS**
    - **CHANGE RATE OF FLOW OF TIME**
    - **EVEN THOUGH VELOCITY IS NONRELATIVISTIC**

E.P. BATTEY-PRATT AND T.J. RACEY,
"GEOMETRIC MODEL FOR FUNDAMENTAL PARTICLES,"
INTL. J. OFPHYS.19. NO. 6, 437-475, 1980

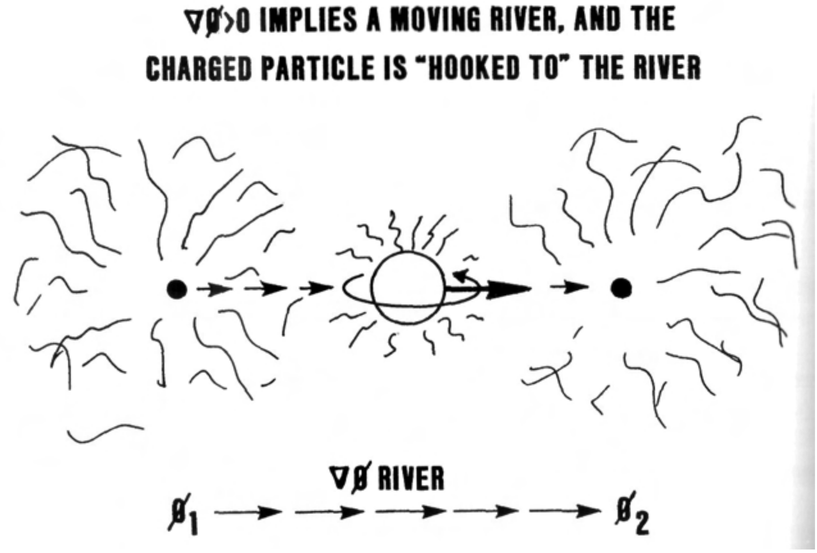
∇ø>0 IMPLIES A MOVING RIVER, AND THE
CHARGED PARTICLE IS "HOOKED TO" THE RIVER
∇ø RIVER
ø1
ø2

Figure 48. In a flux gradient (virtual particle river), a charged particle moves itself.

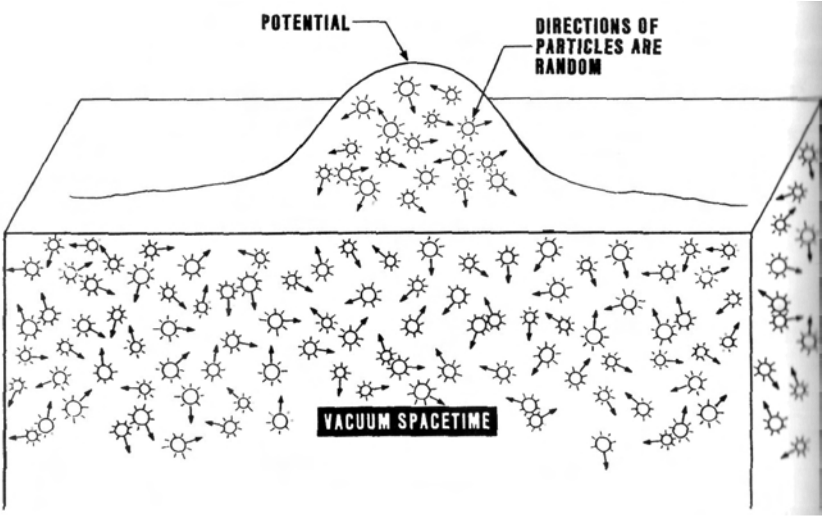
POTENTIAL
DIRECTIONS OF PARTICLES ARE RANDOM
VACUUM SPACETIME

Figure 49. A natural potential is a **disorganized change** in **the stress of** vacuum. It has a random virtual substructure.

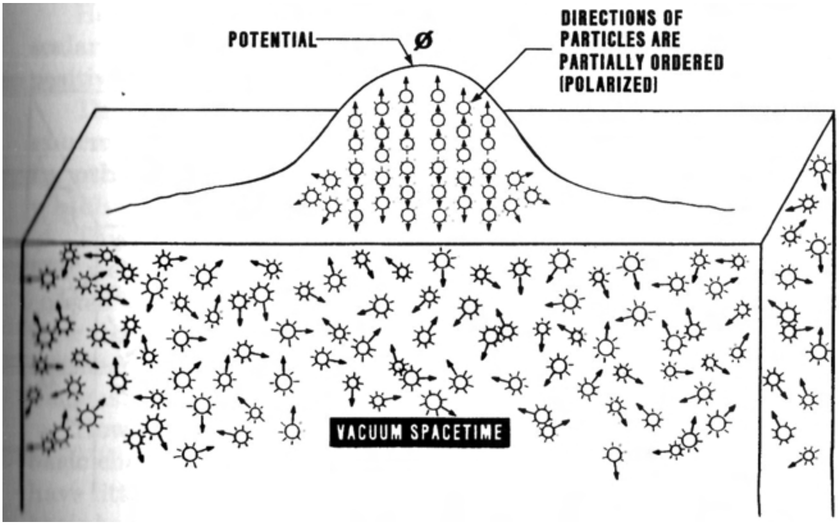
POTENTIAL
Ø
DIRECTIONS OF PARTICLES ARE PARTIALLY ORDERED (POLARIZED)
VACUUM SPACETIME

Figure 50. An artificial potential is an organized change in the stress of vacuum. It has a deterministic virtual substructure.

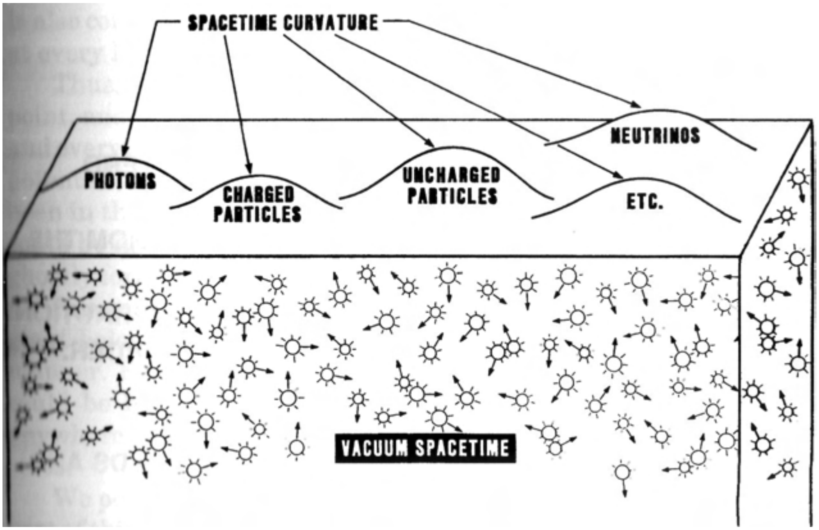
SPACETIME CURVATURE
PHOTONS
CHARGED PARTICLES
UNCHARGED PARTICLES
NEUTRINOS
ETC.
VACUUM SPACETIME

Figure 51. Gravitational potential is a conglomerate of stresses, organized or disorganized.

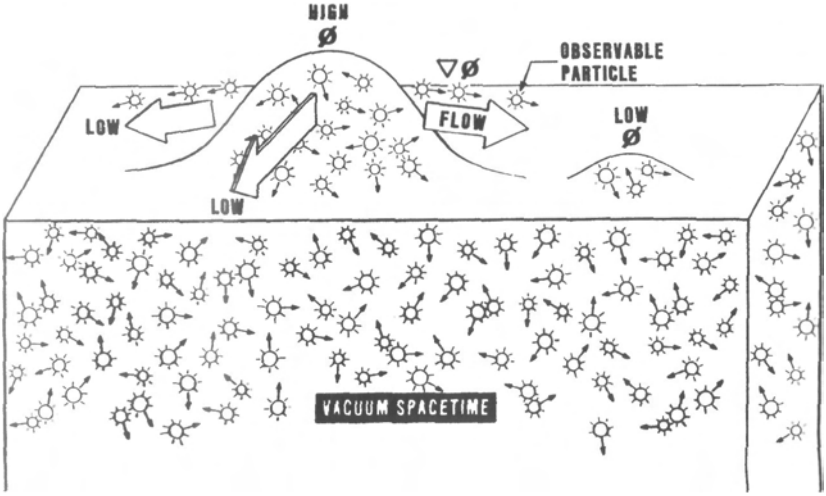

Figure 52. EM force fields are releases of gravitational potential via observable charged mass flows.

Table 33. RESIDUE UNIVERSE, RESIDUE SCIENCE

- **VECTORIALLY, STRESS IS SUM-ZEROED**
- **THE VECTOR FORCE FIELDS REPRESENT NON-ZERO GRADIENTS IN MORE FUNDAMENTAL STRESS**
- **FORCE FIELDS ARE THE RESIDUE LEFT OVER FROM THE SUM-ZEROED STRESSES**
- **PHOTON INTERACTIONS ARE FORCE-FIELD INTERACTIONS**
- **PHOTON-DETECTED REALITY IS ONLY "FIRST ORDER." IT SHOWS ONLY "RESIDUE-LEVEL" REALITY**
- **WE HAVE BUILT A "RESIDUE SCIENCE" AND A "RESIDUE PHILOSOPHY" BASED ON FORCE FIELDS AND PHOTON DETECTION**

Hence at every point in the vacuum, an intense flux of these scalar EM waves, with concomitant interferences, exists — in both positive and negative observer time.

This scalar EM flux causally **drives (constitutes) all the** enormous **quivering of the vacuum** spacetime medium **itself. In other words,** it **drives subquantum (virtual) change, which in turn drives quantum change.**

This is a drastic reorientation of quantum mechanics. In scalar EM, we view that the virtual flux of vacuum is *causally driven or created* by interacting scalar EM waves. This replaces the present **view** that the virtual flux is entirely chance. It opens up the direct engineering and structuring of the vacuum — and hence the atomic nucleus — by scalar EM means.

However, it does not replace the normal *statistical nature* of the **basic** changes of the background vacuum medium. We normally **have** little or no knowledge of the myriad of drivers that cause the basic background waves of that medium. We may, however, deliberately create special vacuum potential waves which are deterministic, and which we have knowledge of.

To the observer, from a purely statistical viewpoint, in this violent vacuum flux *any* and *every* finite pattern of virtual particles is also continually being momentarily created and destroyed — and at every level and in every hyperdimension.

Thus, in the vacuum there continually exists — at any and every point, and in any and every region — the ghostly image of anything and everything, whether in the past, present, or future; and whether **potential,** probabilistic, or actual. Even everything that could have been in the past but wasn't, or might be in the present but isn't, or could be in the future but won't be, is continually present in thin, ghostly form.

Rigorously, the universal vacuum may be taken to be a sort of giant hologram, for the whole is in each part, albeit in ghost-like manner. Everything that is, or was, or shall be; and everything that could be but isn't, wasn't, or shan't be; already exists *at once, anywhere* and everywhere, in this ghostly, holographic, virtual state.

We point out in passing that, conceivably, one can engineer any **part** of this "potential for reality." That is, one can directly engineer physical reality itself, in the scalar EM view.

All that must be done is to amplify (continuously and coherently

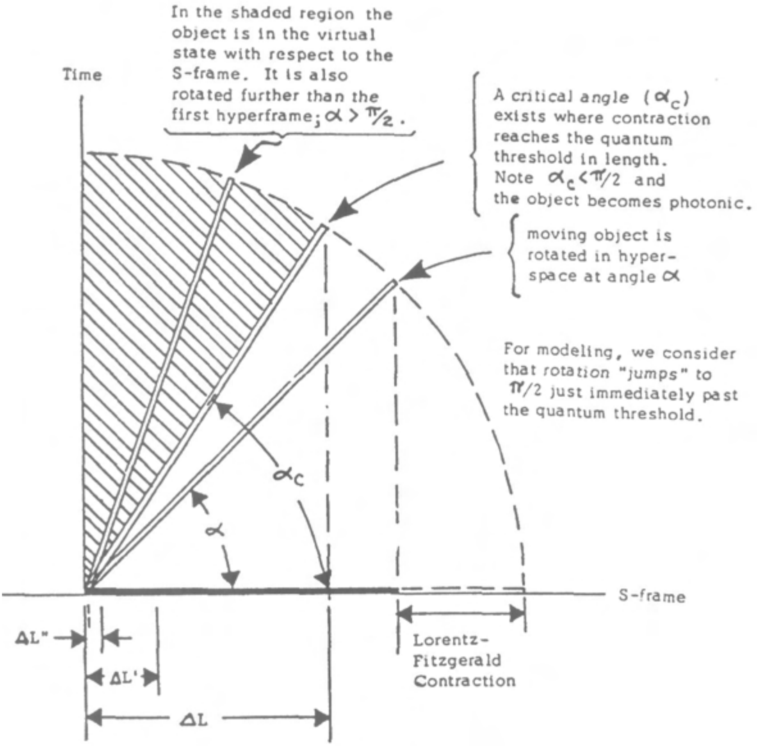
In the shaded region the object is in the virtual state with respect to the S-frame. It is also rotated further than the first hyperframe; $\alpha > \pi/2$.
Time
A critical angle (α_c) exists where contraction reaches the quantum threshold in length. Note $\alpha_c < \pi/2$ and the object becomes photonic.
moving object is rotated in hyper-space at angle α
For modeling, we consider that rotation "jumps" to $\pi/2$ just immediately past the quantum threshold.
α_c
α
S-frame
ΔL"
ΔL'
ΔL
Lorentz-Fitzgerald Contraction

Figure 53. As beyond (without), so within. Hyperspaces are internested virtual states, and rotated orthogonal frames. They are also substructures within and of vacuum state potentials.

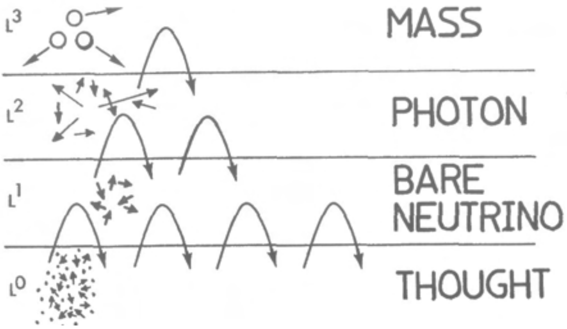
L^3
MASS
L^2
PHOTON
L^1
BARE NEUTRINO
L^0
THOUGHT

Figure 54. Internested levels of virtual state vacuum interact with neutrino, photon, and mass. "Thought" or "mental state" refers to the third and more levels down in virtual state.

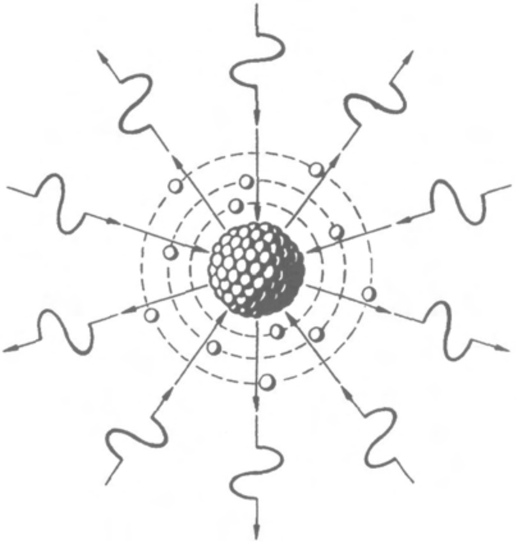

Figure 55. Virtual state patterns are absorbed and emitted by the atomic nucleus. The built-up nuclear "charge" in that particular pattern is a partial potential — for that pattern only.

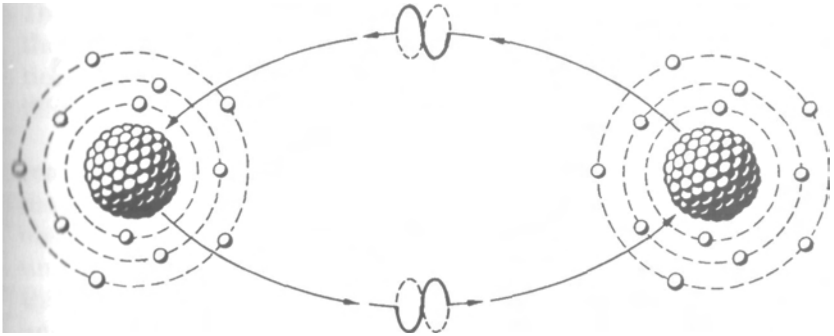

Figure 56. Nuclei of the universe continually exchange scalar EM waves and virtual **panicle** fluxes.

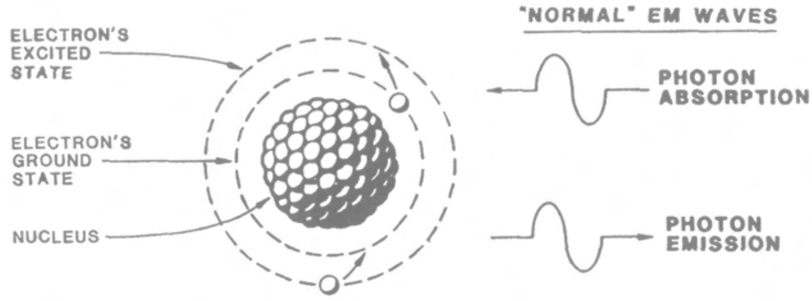
"NORMAL" EM WAVES
ELECTRON'S EXCITED STATE
ELECTRON'S GROUND STATE
NUCLEUS
PHOTON ABSORPTION
PHOTON EMISSION

- **HEAVY NUCLEUS RELATIVELY UNAFFECTED**
- **MACROSCOPICALLY. MASS APPEARS STABLE AND "BRICK-LIKE"**
- **SPECIAL RELATIVITY APPLIES**

Figure 57. First order reality. Rough-hewn by observable photon interaction with electron shells of the atoms of physical mass. This process is "physical observation."

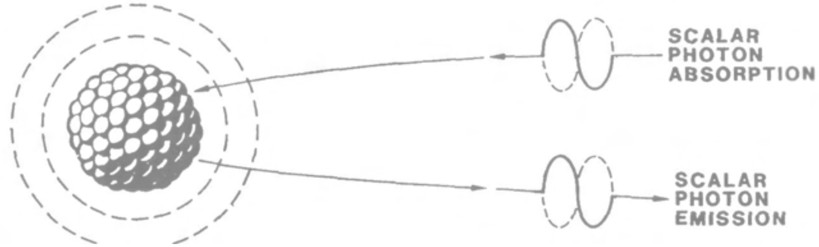
SCALAR PHOTON ABSORPTION
SCALAR PHOTON EMISSION

- **HEAVY NUCLEUS AFFECTED**
- **MACROSCOPICALLY. MASS APPEARS FLUID AND CHANGEABLE**
- **GENERAL RELATIVITY AND KALUZA-KLEIN GEOMETRY APPLY**

Figure 58. Higher order (hyperspatial) reality. It exists as stored in the potentials ot the atomic nuclei of physical mass.

add energy to) one of these "ghost images" virtual state forms. In that case, the form becomes "denser and denser," until eventually it starts to breach the quantum threshold. At that time, the form will be seen emerging as a sort of "coalescing form of light," where the light seems to be forming in little "chunks" or pieces. As the coalescing operation continues, the light-form will eventually become fully formed. If it continues, it will start to coalesce further into *thin, ghostly material* form, appearing much like a faint fog-form. With continued coalescing, it will become a very solid, material form which exerts pressure on tree limbs, breaks them, leaves traces, etc.

Long ago I dubbed this process *kindling.* It is rather widely met in paranormal phenomena of many types.

But back to the basic idea of ghostly images in the virtual states of vacuum.

In this view, Everett's many-worlds interpretion (MWI) of quantum mechanics is literal and real, albeit the additional worlds are to bo taken as virtual and hyperdimensional to the laboratory observer.

(It is rather simple to show that ever deeper, nested levels of the virtual state are exactly the same thing as ever higher dimensions. By "rotating" an object in many-dimensional space, one can show that, after three rotations, a physical object has lost all physical intersections with our normal 3-space, to us now only occupying time. Since this is the precise characteristic of a *thought form,* I used this phenomenon and spatiotemporal model to make a multidimensional space model directly including thought, mind and matter. That is, one is thus able to form a unified field theory of mind, thought, life, and matter — a model that is testable, makes predictions, and actually unites physics and metaphysics. That work and its plications, however, are largely beyond the scope of this book.)

It is stressed that the basic scalar EM image (interpretation) of reality is directly implied by quantum mechanics. It is not idle metaphysics. Other than the scalar EM wave causality for the observer's statistical microscopic dynamics (and the extended model unifying mind, thought, life, and matter), it is not the construct of this author. Instead, it follows from interpretation of the equations and axioms of quantum mechanics (QM) itself, as changed and extended by the "zero vector system" dynamics.

Charge, Potential, and Curvature of the Vacuum/Spacetime

The flux of one type of virtual particle through a point in the vacuum may be examined.

A single type of virtual particle flux (VPF) at a point constitutes one particular type of charge at that point. For example, the flux of virtual photons constitutes the electrical charge.

Further, the magnitude of the selected flux (usually expressed as a flux density) through the point may be taken as the magnitude of the vacuum's charge there in that type of particle. In other words, at a point the magnitude of the vacuum's charge (VPF) is one type of *potential* at that point.

Various kinds of fluxes constitute various kinds of charge, and comprise various kinds of potentials. Immediately we see that one type of charge may always be decomposed into other kinds of charge, as can potentials. This point is beyond the scope of this book and will not be further amplified.

Continuing our example, the magnitude of the electrostatic scalar potential at a point in vacuum represents the magnitude of the flux of virtual photons through the point, and hence the electrical charge at the point.

Table 34. REALITY IS A HOLOGRAM

TO ESTABLISH THIS:

1. **ALL DETECTION IS INTERNAL**
   PHYSICAL REALITY CONSISTS OF THE INTERNAL DETECTIONS OF THE PERCEIVER/OBSERVER.

2. **THE OBSERVER MAY BE AT ANY POINT IN THE UNIVERSE**
   THE UNIVERSE CAN BE DETECTED FROM ANY POINT INSIDE ITSELF.

3. **THE PHYSICAL UNIVERSE IS TOTALLY INSIDE EACH POINT IN ITSELF**
   THIS DEFINES A HOLOGRAM: THE WHOLE IS IN EACH PART.

Twin Flows of Time

In quantum mechanics, each virtual photon is continually turning into an electron/position pair, and vice versa.

By Dirac theory, *a positron is an electron traveling backwards in time.*

Further, pair production produces time-smeared particles — an electron and a positron.

Pair production, then, actually produces two *electrons,* one coupled to (smeared in) a positive piece of time and one coupled to (smeared in) a negative piece of time.

Thus in the vacuum two discretized streams of time, one positive and one negative, are continually created in conjunction with pair production, and destroyed by pair annihilation.

Further, integration of tiny virtual pieces of time (from virtual photons) to form "passage of time" macroscopically is directly associated with the charge (the absorption and emission of virtual particles) of an observable particle.

This is what is meant by an object "existing" (persisting). Its continual virtual photon interactions are integrated by its mass (timeless part) to continually create its energy states and its march through time.

The absorption and emission of an observable photon, however, moves the mass (timeless part) in (comparatively) larger jumps through time. The absorption of the photon connects a positive piece of time to the mass of the particle, converting it to masstime. The subsequent emission of an observable photon "tears off the little time-tail," so-to-speak, and leaves behind a totally spatial mass entity, having no large connection to "the flow of time."

This is what is really meant by "observation."

Because the "observation" disconnects time from the previously "excited mass" (masstime), then it follows that time is not an observable.

It also follows that the twin streams of time are not continuous. Each is broken into incredibly tiny bits.

The bits of time are continually interlaced between negative and positive, and neither of the two streams of time bits is integrated in the vacuum. The vacuum is disintegrated!

A photon — the basic quantum — is composed of discretized energy and time, bound together (integrated) as action (angular momentum). That is, a photon is a little piece of energy, welded to a little piece of time, with no seam in the middle.

Two kinds of photons exist: the normal photon exists as (+dE) (+dt). The phase conjugate photon exists as (-dE) (-dt).

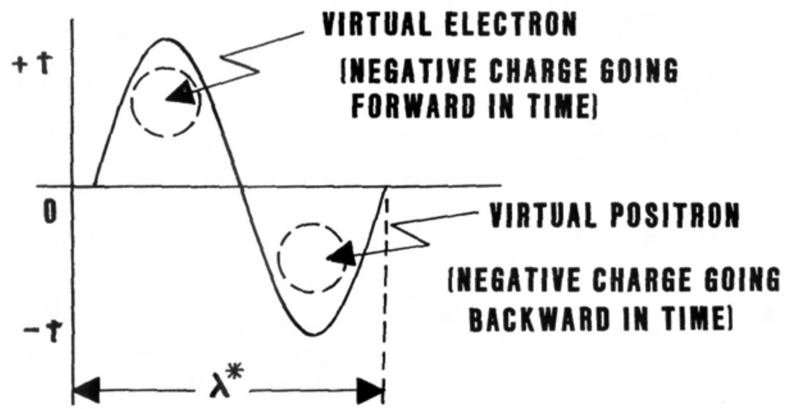
VIRTUAL ELECTRON
(NEGATIVE CHARGE GOING FORWARD IN TIME)
+t
0
-t
VIRTUAL POSITRON
(NEGATIVE CHARGE GOING BACKWARD IN TIME)
λ*

Figure 59. A photon is a single oscillation of a carrier wave. It is also two virtual electrons, one going forward in time, and one going backward in time. With internested modulations (subphotons), the structure of the photon becomes extremely complex.

Table 35. PHYSICAL CHANGE

- **COMPOSED OF ACTION**
- **h = BASIC QUANTUM**
- **TWO CANONICAL VARIABLES**
  - **ENERGY/TIME** $(\Delta E \Delta t)$
  - **MOMENTUM/LENGTH** $(\Delta p \Delta L)$

  ETC.
- **ENERGY/TIME**
  - $h = (+\Delta E)\ (+\Delta t)$
  - $h = (-\Delta E)\ (-\Delta t)$

If the photon is of a certain magnitude, given by Planck's constant, then it is an **observable photon.**

If the photon is less than h in magnitude, then it is a **virtual photon.**

A simple photon in vacuum can also be considered to be one cycle of a sine wave. One half of the sine waves exists in positive time, hence is/contains a virtual electron.

The other half of the sine wave exists in negative time, hence it is/contains a virtual positron — an electron traveling backward in **time.** This second half of the full photon (a single sine wave oscillation) is the phase conjugate (time reversal) of the first half.

As we stated previously, the full photon in vacuum actually exists as a presently visualized "positive time" photon and a phase conjugate "negative time" photon , a "serially linked pair," where **each** photon is only one-half the sine wave. This resolves the conflict inherent in wave/particle duality.

It also has a great deal to say about engineering the vacuum and about detectors and the detection process.

In the spontaneous decay of a virtual photon, it can be seen that one part (the virtual electron) is associated with positive energy and positive time; the other part (the virtual positron) is associated with negative energy and negative time.

The continual "switching" between positive and negative time in an observable particle's absorption and emission of virtual photons **(and** the associated electron/position +t/-t pairs) means that the positive flow of time, to the macroscopic observer, is continually being "started" and "stopped" (created and destroyed).

This continual attachment **and detachment of positive time to and from the "observer particle" creates the "past" and the "future" to which the observer seems somehow to be connected, and yet not connected.**

Engineering Local General Relativity

In a flat spacetime (linear vacuum), the two time flows are balanced, as are pair creation and pair annihilation. In this case, conservation of charge and energy hold.

Table 36. ORDINARY GENERAL RELATIVITY (OGR) IS A SPECIAL CASE

- **OGR HAS ASSUMED AWAY MOST LOCAL GR SYSTEMS, OF THE TYPE WHERE MAXWELL'S EQUATIONS AND THE CONSERVATION LAWS CHANGE LOCALLY.**

- **PHYSICISTS ARE THUS STRONGLY DISCOURAGED FROM CONSIDERING THAT GR CAN EASILY BE ENGINEERED IN A LOCAL EM SYSTEM.**

In a curved spacetime (nonlinear vacuum), one of the two triplets (positive electron/energy/time or negative electron/energy/time) predominates. Hence in a curved spacetime the rate of flow (production) of the two time streams is unequal, as is the production of the two kinds of energy and the two kinds of electrical charge.

A locally curved spacetime thus produces a predominance of either positive or negative electrons, and either positive or negative energy, and a predominance of the addition of either positive or negative time production.

The curved spacetime acts as a local source or sink accordingly, and both conservation of energy and conservation of charge can be locally violated in such a curved spacetime region.

As we stated, the locally curved spacetime produces either positive time or negative time more predominantly, acting as a source or sink for normal (positive) observer time. The local rate of flow of the observer's time can thus be speeded up or slowed down, depending upon the type of local spacetime curvature utilized.

It follows, then, that standing local gravitational waves can result in localized, stable gradients in the vacuum virtual particle flux, between the "high potential" part of the standing wave and the "low potential" part.

This allows the violation of conservation laws such as charge and energy in the local gravitational gradient, utilizing the localized curvature of vacuum to provide a special source.

For proof that conservation laws need not apply to a curved spacetime, see A.A.Vlasov and V.I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity, **Theoretical and Mathematical Physics,** 53(3), June 1983 **(English translation);** Russian, Dec. 1982, p. 406-418. Quoting:

"This result is a particular consequence of the general assertion to the effect that in general relativity there are no energy-momentum conservation laws for a system consisting of matter and the gravitational field."

See also V.I. Denisov and A.A. Logunov, "New theory of spacetime and gravitation," **Theoretical and Mathematical Physics, 50**(1), July 1982, p. 3-76. This paper (p. 3) points out that:

"... **the gravitational field in general relativity is completely different from other physical fields and is not a field in the spirit of Faraday and Maxwell."**

A 1984 Soviet paper by senior Russian physicist C. Yu. Bosgoslovsky, "Generalization of Einstein's relativity theory for the anisotropic spacetime," is also very relevant.*

In composing his theory of general relativity, Einstein assumed that the local spacetime of the observer could never be curved, and instead would always be represented by a Lorentz frame (flat spacetime).

This severe assumption had the effect of "saving" the sacrosanct conservation laws, and maintaining the exclusion of electrogravitation.

Since then, Western physicists have raised Einstein's assumption — and the conservation laws — to a dogmatic belief system — in some cases, to near fanaticism.

To see a clear statement of these limiting assumptions imposed on general relativity to provide the severely restricted version taught in almost all Western universities, see Charles W. Misner, Kip S. Thome and John Archibald Wheeler, **Gravitation,** W.H. Freeman and Co., San Francisco, California, 1973, p. 19-21, 71-72, 367-369.

In other words, Western physicists still use a sort of "special

---

**See* also V.I. Denisov and A.A. Logunov. "The inertial mass defined in general relativity has no physical meaning," preprint p. 0214, Institute of Nuclear Research, USSR Academy of Science, Moscow, 1981.

relativity with distant curved spacetime pertubations," and call it *general* relativity!

The Soviets do not limit themselves to the same restricted general relativity theory as does the West.

Santilli (Ruggero Maria Santilli, **Ethical probe on Einstein's followers in the U.S.A.: An insider's view,** Alpha Publishing, Newtonville, Massachusetts) has documented what essentially amounts to a conspiracy in Western physics to shut out any generalization of Einstein's severely limited relativity.

It is precisely this near-total Western scientific prejudice — along with Heaviside's mutilation of Maxwell's original theory — that has excluded electrogravitation from being developed in the West.

Note also that, by embracing a "locally flat spacetime" assumed to be decreed by the "laws of the universe," in their rigid mindset Western physicists have excluded any possibility of performing simple laboratory experiments in locally curved spacetime.

They do not believe it can be done, so they have never tried.

But let us further address the importance of the negative triplet (electrical charge/energy/time) and its concomitants parity, gravity, and entropy:

Engineering Antigravity

In negative time flow, gravity is a repulsion, not an attraction.

Producing antigravity in an object is simply the production of excess negative time flow in that object.

That is readily achieved by curving local spacetime in the appropriate direction, and forcing the object to produce excess negative energy.

Mass, for example, is just a special form of trapped energy—and trapped energy (potential) is just trapped charge of one kind or another, whether observable or virtual.*

As such, mass represents a "localized potential" in the ambient vacuum, of a certain kind of "trapped energy/time charge."

*Actually, since it is in equilibrium in a dynamic and continual flux, a mass particle is actually a special trapped **rate** of action flow. A defining equation for mass in terms of the rate of action flow was given by the author in **Quiton/Perccptron Physics,** NTIS, 1973. Report AD-763210.

The mass/potential is continually being charged by the vacuum flux, and discharging and emitting charge flux into the vacuum flux. The mass is "in equilibrium" at its fixed potential.

Normal ("positive") mass is simply charged up with positive energy and positive time. And its external electron shells result in it presenting a "negative charge" face to the world.

If we charge up the mass, however, with negative energy and negative time, so that some of its charge potential now is of this second type, the trapped energy will now consist partially of negative energy. In addition, the mass will partially exist in negative time (be charged up partially with negative time).

When excess negative time flow is produced in a mass, the mass begins to be repelled by the earth as well as still being attracted to it in its positive time stream. The net result of the object's fixed attraction and increasing repulsion is that the object begins to lose weight in the earth's gravitational field and gets lighter.

When a sufficient rate of excess negative time flow (sufficient local spacetime curvature) is produced, the mass floats freely.

With additional negative time production rate, it accelerates ("falls") upward.

The rate of excess negative time produced may be controlled very simply by using a "negative time/negative energy" generator and varying the amount of the load applied to the device.

The negative time/negative energy generator can be rigidly affixed to a flying apparatus, serving as a sort of gravitational rocket or gravitational propulsion system.

Of course, the negative energy produced in the propulsion process can be used to light lamps and power electrical loads, as well as host of other things. (Yes, Virginia, it can also be used to generate force fields, power weapon systems, etc.)

Thus one can "eat his cake and have it too"; by using negative energy flow for power, a very practical antigravity spaceship can be developed.

An additional side benefit to the passengers probably exists: they should age at a slower rate when traveling in the vehicle.

As can be seen, since many kinds of VPFs exist in vacuum, the vacuum is filled with (and consists of) many types of "potentials." Overall, this matches very closely the view of general relativity, if we regard the vacuum itself as indefinite "spacetime", where no ordered universal frame with a fixed metric has yet been associated or prescribed.

Curved Spacetime and the Disintegrated Vacuum

From general relativity, additional characteristics of the modern vacuum/spacetime may be inferred or interpreted.

First, we may take a "flat spacetime" to consist of a charged vacuum, such that both the magnitude and structure of the VPF are essentially unchanged from point to point, and the flux of virtual particles and virtual antiparticles are balanced. Equal positive and negative time flows/charges exist.

Note, however, that we can have what casually may appear to be a "flat spacetime" but instead is "warped." This case exists when the magnitude of the overall VPF comprising the vacuum does not change from point to point, but the component structure of the VPF does vary from point to point.

In other words, the overall vacuum potential remains the same from point to point, but two or more individual potentials may canonically vary. Let us further clarify this point by analogy.

Suppose we view the vacuum/ether/spacetime as a special sort of gas, where the "gas" is actually a mixture of many, many gases.

It is a rather peculiar sort of gas; it exists in five or more dimensions and is composed of very strange, fleeting particles which . spontaneously are created and destroyed, each arising out of nowhere and returning to nowhere almost immediately.

There is only minuscule (negligible) time overlap (integration) of the existence of these virtual particles in the main.

That is, mostly each virtual particle — along with its associated energy, movement, momentum, charge, and increment of time — exists almost entirely as an individual.

The energies, movements, momenta, charges, and time increments of the individual virtual particles do not integrate or sum to any appreciable degree in vacuum.

Thus, although casually this ether/gas may seem to contain enormous "energy density", momentum density, charge density and time increment density, these are not *integrated* energy density, momentum density, and time density in the normal sense.

Instead, they are the densities *of disintegrated* energy, *disintegrated* momentum, and *disintegrated* time flow.

In scalar electromagnetics, we have called the disintegrated energy of vacuum **anenergy.** This clearly distinguishes that it exists in a fashion altogether different from normal, *integrated* energy with which we are accustomed.

By integrating anenergy, of course, one obtains ordinary energy—but it first requires an integrating agent. The most **usual integrating agent/operation is the spin of an observable particle of** mass.

Similar considerations apply to momentum and time.

With this more **precise interpretation, the vacuum contains enormous anenergy, but essentially no energy. It contains enormous virtual energy, but no observable energy.**

Unfortunately, however, orthodox physics does not yet clearly differentiate between the two states of energy, virtual and observable (disintegrated and integrated).

Instead, it continues to loosely utilize the term "energy" for both states, whether integrated or not. Thus it exclusively uses the term **"energy** density of vacuum." Rigorously, the term should be "anenergy density of vacuum."

Indeed, most physicists are not clearly aware of the distinct difference between the two different states of energy, insofar as the implication of integration or nonintegration.

Regrettably, then, we will continue to utilize the orthodox term, "energy density of vacuum." but the reader should be clearly aware that the vacuum "energy" and vacuum "stress energy density" as **such** are totally disintegrated.

Separation of Vacuum and Observable States

Because the energy density of vacuum is unintegrated, an integrated physical body — including one's personal body — can **exist** "in" the vacuum medium of incredible "energy" density, yet observably exist in a medium of zero energy density.

The unintegrated vacuum ether has enormous spatial density of *virtual* energy, but essentially has zero spatial density *of observable* energy.

Note, however, that in our hyperdimensional interpretation of the levels of virtual state, the vacuum spatially does not exist in the normal 3-space of observable matter.*

The vacuum occupies the 4th dimension (time) and all spatial dimensions (hyperdimensions) greater than the third — it occupies all levels of virtual state, but not observable state. It occupies all levels of disintegration, but not the levels of observable integration.

Exchanges occur between normal 3-space and the hyperspatial vacuum, through the common connecting dimension, time.

These exchanges (of virtual particles) are integrated by the spinning observable mass particle into the first quantally excited state — a delta (change) representing a single entrapped quantum — of the 5-space potential represented by the particle and its accumulating VPF. At that time a sharp electromagnetic (fifth dimensional) discharge of the quantum occurs, converting the accumulated 5-G quantum to the ordinary EM quantum of the 4th (time) and 5th (infolded space) dimensions.

This accounts for the so-called "collapse of the wave function", the production of unitary quantum change, and the discretizing of observable change.

We note in passing that time is the only "dimension" where everything can be considered to occupy the same "point" or "interval", but it is discretized and chopped into pieces (disintegrated), quite different in nature from the familiar three spatial dimensions of the observable (integrated) universe.

The dimensional separation (spatially) of the seething vacuum "energy" (anenergy) and the 3-space of normal matter allows a physical body to enormously interact with the vacuum at or near equilibrium at the microscopic level, but maintain its macroscopic form and stability as if no interaction were going on.

Partial Potentials, Curvature, and Warping

At any rate, we view the magnitude of the overall "potential" of this vacuum gas mixture, as the magnitude of the entire **"stress**

---

*Here we sharply differ from conventional physics, which hopefully confuses the concepts of vacuum, spacetime (indeterminate), spacetime (determinate), and frame.

ENERGETICS OF A POTENTIAL
Running in Place and Getting Nowhere

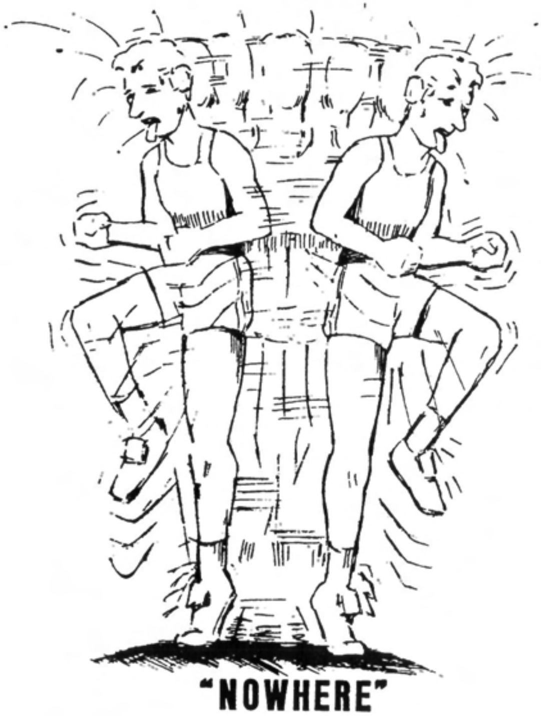

Figure 60. A potential is trapped energy. A potential in vacuum is trapped virtual-state "energy" or flux. A natural potential is trapped random energy. An artificial potential is trapped organized energy.

energy" or "pressure" of the mixture — IF AND ONLY IF this stress energy or pressure were somehow integrated.*

As can be seen, we can now look at each of the component gases in the mix as contributing a partial pressure or partial stress energy (an individual component potential) of the vacuum conglomerate.

We may have the overall pressure remain the same, yet canoni-

---

***Thus it's** really the **potential** stress energy density. The vacuum is nothing but potential!

cally change the partial stress energies (individual potentials).

In that case one has an internally twisted or "warped" spacetime.

A warped spacetime may prove quite useful, for example, for direct translation of one type of "energy" into another. In that fashion, one may conceivably obtain radiation energy at the expense of mass, or mass at the expense of radiation energy.

Or one may convert between potentials where one or more of the converted potentials are as yet unknown to orthodox physics.

Kaluza theory, for example, contains many potentials that are as yet unknown and undiscovered in nature. When undue simplifications are made to arbitrarily exclude the unknown potentials, the adapted theory leads to error in its predictions.

This was originally regarded as a great imperfection, and Kaluza theory fell into disfavor for decades. Then as modern particle physicists began to produce fundamental particles of an astounding variety, a Kaluza-Klein theory of 11 or more dimensions was found to very reasonably describe the emerging situation. Once again, then, Kaluza-Klein theory became a moving force among modern theorists.

One may also convert electromagnetic field energy into gravitational field energy, for example, and obtain enormous amplification of gravitational effects.

This is precisely what scalar electromagnetics seeks to do.

Since the electrical field between two electrons is about 10^{42} times as strong as the gravitational force between them, conversion of the EM charge flux into gravitational charge can conceivably yield up to 10^{42} times as much gravitational force as normal, depending upon the efficiency of the conversion process. Even with a process or device that uses electron currents and has an efficiency of only 10^{-22}, gravitational/inertial gains of 10^{20} are possible.

Note, however, that the levels of vacuum may also be compared to horizontal planes beneath the ocean.

For example, we may take a particular level of vacuum potential, and that level constitutes a sort of included "flat spacetime" of its own. By biasing the potential of a scalar transmitter and receiver to a given level, one may transmit (in either direction) at that level only, Receivers at higher or lower levels will detect nothing.

The immediate application to engineering the nucleus for such things as nonradioactive transmutation of elements, controlled nuclear fusion, electromagnetic healing, and processing nuclear wastes to render them harmless should be apparent.

To recapitulate our analogy, a true "flat spacetime" is a vacuum whose VPF does not essentially vary overall or in each constituent part, between two comparative points.

A "curved spacetime" is a vacuum whose overall VPF varies in overall magnitude, between two comparative points.

A "warped spacetime" is a vacuum whose VPF may or may not **vary** in overall magnitude between two comparative points, but whose partial VPFs vary their ratios, between the same two points. Note that a vacuum/spacetime can be both warped and curved.

It is stressed that we have not limited the number of dimensions **that** can be placed on the vacuum to Minkowski's four. We hold the number of dimensions necessary to represent both vacuum and observable matter to be arbitrary, but greater than four.

To unify electromagnetism and gravitation (which is what scalar electromagnetics is about), we must take five dimensions — four spatial and one time — to be the minimum necessary in modeling the unified theory, similar to Kaluza.

We state now a major characteristic: **Vacuum/spacetime consists entirely of potentials, nothing else, and these are hyperspatial a priori.**

The vacuum does not contain observable force fields nor any **Other** *non-potential* entity.

We also state without proof that in scalar EM theory it seems **very** possible to model the mind and its unification with physical **reality,** in a fashion subject to experimental verification.

Force and Mass Are Inextricably Intertwined

In foundations of physics, it is well-known that, paradoxically, force and mass can only be defined one in terms of the other.

We may use that paradox for a new interpretation of force and **force** fields: a force may be interpreted **to consist of** the time rate of changing the "mv" momentum of a mass system or object, instead of **being the cause of** its dp/dt.

That is, we may take the view that

$$\mathbf{F} \equiv d\mathbf{p}/dt \qquad (4\text{-}14)$$

not just that

$$\mathbf{F} = d\mathbf{p}/dt \qquad (4\text{-}15)$$

We have previously pointed this out and called it the "strong definition of force."

This immediately suggests the question: If force is an effect and not a cause, what, then, causes force?

Quantum mechanics already provides the answer: A differentiating operator must be applied to (coupled to) potential in order to produce a force field. The most usual "differentiating operator" is an observable spinning charged particle.

A charged particle exposed to the variation of electrostatic scalar potential between two points is exposed to a gradient (flow) in the virtual photon VPF, and hence to a directional "stream" of virtual photons. The observable charged particle couples to that stream. **Its** spin integrates the unobservable, disintegrated anenergy of the gradient flux into integrated, observable energy of the particle.

With greater integrated "pressure" on one side of the particle than on the other, the observable particle accelerates and translates This produces or "detects" (and comprises) an electric force between the two points.

Force is the mass particle(s) accelerating, either in time or in mass. Observable force is observable particle(s) accelerating. **Virtual** force is virtual particle(s) accelerating.

In **Quiton/Perceptron Physics,** I gave a new definition of mass that is consistent with this viewpoint:

$$1 \text{ kg. mass} = 17.053 \times 10^{50} \text{ switches/sec} \qquad (4\text{-}16)$$

where a switch is h/4*pi), having the units of action/time, or energy. This directly defines mass in terms of energy accumulation, where the energy accumulated is in equilibrium in a differentiated action flow.

Note that "energy" is the time rate of flow of action. Energy can only be accumulated (as a mass or a potential) by possessing two rates of flow of action simultaneously, one in positive time and one in negative time. This is a very deep statement that affects all of physics, but unfortunately it is beyond the scope of this present book.

To return:

The observable electric force field does not exist as such in vacuum. The vacuum and all its constituents are totally virtual. Only the potential ("potential for the electric force field, given a coupled charged observable particle") exists in vacuum, as pointed out by Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley, Reading, Mass., 1964 vol. 2, p. 2.

The Concepts of Zipped and Unzipped Forces

As noted previously, the potential exists as an unintegrated virtual particle flux (VPF). Further, each little moving virtual particle is smeared in length and time, is coupled to finer VPF, and thus may be said to constitute a minute "force vector" while it exists.

Thus, the electrostatic scalar potential is composed of a myriad of little individual "virtual particle force field" vectors.

Since these little virtual vectors are essentially separate and unintegrated, one may say that the electric field is *unzipped* in vacuum. By "unzipped" we mean that the individual virtual particle E-field vectors are unintegrated.

A potential, for example, is composed of unzipped vector fields since it is comprised of unintegrated, virtual vectors.

To be static, the potential must be comprised of a pair (at least one pair) of unzipped vector fields, one in positive time and one in negative time. These unzipped vector fields must in turn be locked together. In this case, the positive time observer sees nothing (vacuum only), until he "observes" or detects. Normal detectors detect (and in so doing, they zip together) only the positive time unzipped component. The negative time (phase conjugate) vector held retraces the positive time field, in Newton's third law. Or, if generating a wave or pulse, it would be seen as a phase conjugate replica which affected the disturbing source.

Let us be very plain. **A "scalar EM potential" is actually two vector fields zipped together: one is in positive time, and the other is in negative time.** Each half is composed of microvectors. A dynamic gradient in this scalar EM potential (i.e., a scalar wave, for example) actually consists of two waves: one is the normal" positive energy/positive time" EM wave, while the other is the phase conjugate replica. When we detect the normal half (with electron detectors), we actually detect the second half with the atomic nuclei. Since it is a phase conjugate, it produces a negative momentum change on the nucleus {since one has p/(-t)} instead of LeSage's positive momentum change. In other words, it produces a drawing (attracting) force on the absorbing nucleus, instead of LeSage's assumed repelling force on it.

That is the entire secret of gravitation. And of electrogravitation. Knowing that, you can easily work out how to apply the necessary engineering to get antigravity.

But to return.

The potential can also be comprised of many unzipped vector fields at once. These may be deterministic (artificial potential) or randomized (natural potential).

The "force field" is disintegrated (unzipped) in vacuum, and integrated (zipped) on, of, and by an observable charged particle. In vacuum, it is always accompanied by its phase conjugate twin. When detection by an observable particle occurs, the spin of the particle acts as the "zipper," producing an observable force (which is actually of the particle accelerating, not "acting independently upon it.")

Vacuum is unzipped; observable mass is zipped.

Any zipped entity exists observably in 3-space; any unzipped entity exists in hyperspace outside 3-space, and is unobservable and virtual to the 3-space observer.

This interpretation is still consistent with the extended quantum mechanical view, where ultimately all observable forces are considered to be generated by absorption and emission of virtual particles by/from an observable particle (mass), at least insofar as what we may observably measure or detect.

Force, Force Field, and EM Waves in Vacuum

In the new interpretation, consistent with the quantum mechanical view, observable force becomes an effect, not a cause.

One may interpret quantum mechanics as stating that **any observable is an effect, since it is assumed to have a virtual state cause.** In the axioms of quantum mechanics, an operator is required for every observable.

The primary **causative agents are potentials, not force fields, as** is **now** firmly **established in** quantum mechanics.

The Aharonov-Bohm (AB) effect is a striking illustration of the primary reality of the potential rather than of the force field.

In the AB effect, potentials may still exist when the force fields **are** zero, and their interference can still cause real effects. The AB effect has finally become generally accepted by physicists after nearly three decades of controversy.

The immediate result is that, contrary to the outdated view imbedded in classical EM, observable E and B force fields do not exist *as such* in vacuum. This allows a totally new interpretation of the **nature** of the EM wave in vacuum: It is longitudinal. And actually, it exists in longitudinal pairs: the second is the phase conjugate of the first.

But of course we detect transverse waves with our detectors! However, this has a simple explanation.

Examine, for example, a straight-wire antenna. The spinning conduction electrons are longitudinally constrained, as is well known. The electrons are relatively unrestrained transversely in the **wire,** but can only "drift" longitudinally down the wire with a speed usually on the order of centimeters per second. The signal potential, however, moves longitudinally down the wire at almost the speed of **light.**

Obviously it is not the longitudinal movement of conduction electrons that constitutes the "signal", even though classical electromagnetics represents it that way.*

We note that a spinning electron whose movement is longitudinally restrained, but is unrestrained transversely, may be characterized as a gyroscope. Further, the axis of this gyroscope may be taken as lying in a plane perpendicular to the straight wire conduction path, since the electron is relatively unrestrained transversely.

---

*Actually, we do not really know how current/electrons travel down a wire, or even if it does/ they do. See three articles by Chappell Brown in Electronic Engineering Times: "Railgun Research Shoots Holes in Lorentz's Theory," April 6, 1987,p. 49-50; "Anomalies in Electromagnetic Law Spur Debate," Sept. 14, 1987, p. 58; "Electrons and Conduction: Not So Simple After All," Dec. 28, 1987, p. 21-22.

If now disturbed by a vacuum potential gradient, the gyroscopic electron is free to precess at right angles to the length of the wire. In other words, given a disturbing force, the electron will move transversely, or radially in the cross sectional plane of the wire, because of gyroscopic precession.

Thus, assuming a gyroscopic electron and its longitudinal constraint, we will detect transverse oscillations in the electron gas in the wire, IF AND ONLY IF the incoming disturbance in the vacuum is longitudinal, and IF AND ONLY IF the unzipped vacuum disturbance "couples" to (is zipped by) the gyroscopic conduction electrons.

Two things are certain in our electron-detector model: First, by the nature of a gyroscope, the conduction electrons move (precess) at right angles to the "disturbing force." Second, we only detect the movement of the electrons themselves.

Since we do detect transverse oscillations of the precessing electrons in the straight wire antenna, then it follows that normal electromagnetic waves in vacuum—that is, the component half that, interacts with the atom's electron shells and not with its nucleus — are (1) longitudinal, and (2) of such form that they couple to electrons.

Note that Tesla always insisted that electromagnetic waves in vacuum were longitudinal. For example, he called them "electromagnetic sound waves" in the ether. In "Tesla Maps Our Electrical Future," H. Winfield Secor presents Tesla's view of transverse waves in this fashion:

"Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound... He says that a Hertz wave would only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves ***are*** *absolutely impossible."*

In his "The True Wireless," **Electrical Experimenter,** May **1919,** Tesla himself further stated:

"The Hertz wave theory of wireless transmission may be kept up **for** *a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

It should also be pointed out that our analogy permits the option of making a longitudinal wave in the vacuum in such a fashion that It will not couple to electrons.

In that case a normal "electron gas" conduction detector will not detect the new wave. Such a noncoupling EM longitudinal wave is **referred** to as a **scalar electromagnetic wave,** where by use of the **term** "scalar" we call attention to the fact that the vacuum potential **is** varying, but electron coupling to form vector force fields in an **electron** gas does not occur.

Since according to general relativity the variation of a potential in vacuum rigorously is a gravitational wave, then the scalar EM wave is a gravitational wave.*

We often say it is an *electrogravitational* wave to call attention **to** its electromagnetic origin, and to the absence of EM force field "bleeding away" of the 5-space Kaluza-Klein gravitational potential. We will discuss the 5-space G-potential shortly.

The conventional equation

$$\mathbf{E} = -\nabla \varnothing \qquad (4\text{-}17)$$

thus applies in the detecting electron gas, but not in the vacuum. Instead, in the nonmaterial vacuum, rigorously

$$\nabla \varnothing \neq -\mathbf{E} \qquad (4\text{-}18)$$

and this states that a gradient in a vacuum electrostatic scalar potential produces an antiphased electric force on and of a particle, on condition that a particle is present. The intent of equation (4-17) should be stated as

---

*It is also a very powerful gravitational wave, since for electrons it is some 10^{42} times as strong as "mass attraction" gravity waves.

$$\nabla \varnothing = -\mathbf{E} \qquad (4\text{-}19)$$

where the vertical line means "on condition that", and cp means "charged particle" (i.e., "that a charged particle is there.).

Equation (4-19) clearly shows that, in the vacuum, only the organized potential for an E field exists due to the quantity $\nabla \varnothing$ essentially as pointed out less strongly by Feynman et al in **The Feynman Lectures on Physics.**

Certainly, in the quantum mechanical interpretation, it is the potential that is real. The force field can only be made from the potential by a differentiating operation.

Since there is nothing in vacuum to perform this differentiating operation observably, one may interpret quantum mechanics as already prohibiting the existence of observable electromagnetic force fields in vacuum, at least in the form prescribed in classical electromagnetics.

Figure 61. A force-field vector is a movement of mass in one direction.

One may regard quantum mechanics — and physics in general - to be incomplete until electromagnetic theory is changed to incorporate the implications of quantum mechanics.

In scalar electromagnetics a new interpretation is taken, and the "overhaul" of electromagnetic theory is in progress.

We accent that scalar electromagnetics is still embryonic. It is at a stage similar to that where electromagnetism was when experimenters were still rubbing glass rods with cat fur and experimenting with little charged pith balls.

Early experiments are enlightening and encouraging; nonetheless, the phenomenology and the myriad of variables are not at all well-understood, even by the experimenters themselves, and certainly not by this author.

Zero-Summed Vector Systems: Shortcoming in Vector Analysis

It has long been pointed out by this author that a fundamental shortcoming exists in classical vector analysis itself when applied to physical systems. Let us briefly look at the construction of the vector mathematics theory, to clarify this shortcoming.

First, an abstract entity called a "vector" is conceptualized. This vector has both direction and magnizude, as contrasted to a "scalar" quantity, which only has magnitude.

Next, an abstract "vector space" is defined; this is a sort of "container" for mathematical objects called vectors, with certain inherent properties of the system specified. In other words, a vector space is an abstract "system" or "space" which can contain objects called "vectors."

Next certain other properties, operations, and entities are defined, one of which is the is the "zero vector".

Essentially the zero vector is defined by a set of operations; however, one simplified interpretation is "the absence of any vector of finite magnitude." By implication, in the abstract vector theory, all zero-vectors are defined to be identical.*

---

* Because, by their nature assumed in the concept, they have no internal structures. That is, effectively the zero vector has been too restrictively defined as "the absence of any and all finite vectors." It should be redefined as "the absence of any single vector being present alone." In that case the definition would admit "the presence of multiple translation vectors, where their combined translation action is zero."

In the vector addition operation, a group of finite vectors adds to a single vector called the "resultant."

It is standard practice in abstract vector analysis to replace a system of summation vectors by their resultant, since all appropriate properties of the summation system are retained by the vector resultant, in the abstract mathematical model.

This, however, is only possible because the abstract vector space has no such thing as stress, and a zero-vector has no substructure by definition. In the abstract space, only translation and rotation are possible, *by assumption*

It is also possible for two or more finite vectors to add to a zero resultant, in which case the system is replaced with a zero vector. Note that all further action by the summing vector components is eliminated by this assumption.

Again the appropriate properties of the summation system are retained by the zero resultant, in the abstract mathematical model, only because the *medium* (the abstract vector space) is not subject to stress, translation, rotation, and structuring.

Note however, that in physical systems, one cannot simply

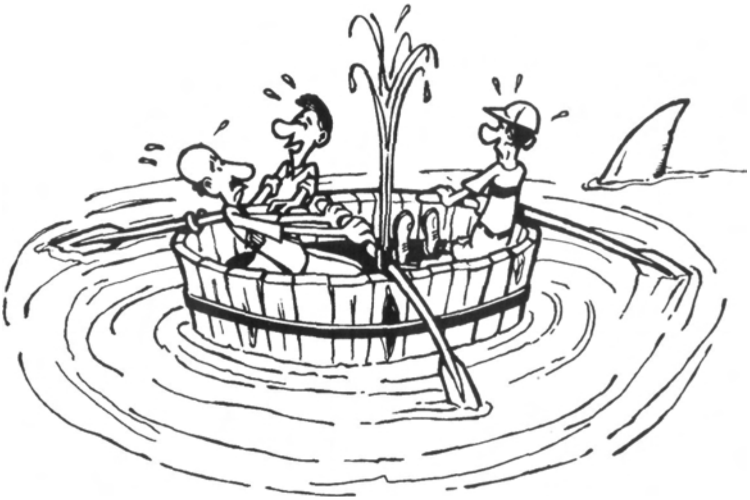

Figure 62. The zero-summed vector system is an engine and an artificial potential.

replace a zero summation system of force vectors with a zero vector and retain all characteristic physical effects caused by the replaced system. The physical medium - including the virtual flux of the vacuum, *is* subject to stress, translation, rotation and structuring.

Physically, the presence of the summing vector component forces creates a stress and a dynamic macroscopic structure in the physical medium in and on which the forces are acting. Therefore the stress of the medium and its structure must be preserved and accounted for, even though a zero vector resultant force exists.

This is particularly true in the vacuum, where internal (infolded) stresses cause charge and energy conversion between one kind of charge/potential and another.

Maxwell's Theory Was Altered and Curtailed

In fact, Maxwell seems to have been well aware of the electromagnetic stress of the ether and its importance. After all, he believed the ether to be material and mechanical, and developed his theory accordingly.

For example, in his Treatise on Electricity and Magnetism, he wrote:

> *"There are physical quantities of another kind which are related to directions in space, but which are not vectors. Stresses and strains in solid bodies are examples of these, and so are some of the properties of bodies considered in the theory of elasticity and in the theory of double refraction. Quantities of this class require for their definition nine numerical specifications. They are expressed in the language of quaternions by linear and vector functions of a vector."*

Maxwell worked out his theory in quaternions since, at the time, the modern form of vector analysis had not been originated by Heaviside and Gibbs.

To recapitulate what we stated earlier: A quaternion consists of a vector part and a scalar part and can readily take into account the stress and strain of the medium. A vector consists only of the vector part, and does not take into account the stress of the medium.

It seems certain that Maxwell knew that the scalar part of the quaternion could be present and vary, even though the vector part was sum-zeroed. Quaternions, however, are devilishly difficult, and even in the time of their founder, Hamilton, few mathematicians and scientists ever mastered them.

Maxwell's theory was transposed (shortly after his death) to its modern vector form by Oliver Heaviside. Many of the characteristics of the scalar part of the quaternion were effectively discarded.

In the remodeling, the "unified" basis was omitted, and electro-

Table 37. MAXWELL'S EQUATIONS

- INVARIANT
  - UNDER CONFORMAL GROUP OF TRANSFORMATIONS
  - IN 4-D MINKOWSKI SPACE
- GROUP INCLUDES
  - TRANSLATIONS
  - ROTATIONS
  - REFLECTIONS
  - INVERSIONS WRT HYPERSPHERES OF M-SPACE
- HENCE MAXWELL'S EQUATIONS INCLUDE TRANSFORMATIONS THAT

CHANGE INERTIAL FRAMES INTO FRAMES THAT ARE NOT INERTIAL.

PROVED IN 1910 BY
H. BATEMAN AND E. CUNNINGHAM

gravitation was accordingly discarded. Heaviside wrote a severely limited subset of Maxwell's theory; he did not by any means capture all of it.*

The new Heaviside vector reinterpretation of EM was far easier to calculate in than Maxwell's quaternions. Many physicists (there were really very few scientists in those days!) immediately loved it.

FitzGerald wrote glowingly of Heaviside's translation (and mutilation!) of Maxwell's theory:

"Since Oliver Heaviside has written, the whole subject of electromagnetism has been remodelled by his work. No future introduction to the subject will be at all final that does not attack the problem from at least a somewhat similar standpoint to the one he puts forward."

Thus it was that Heaviside discarded the basis for scalar electromagnetics and electrogravitation from Maxwell's original work, and **the** so-called "classical Maxwell Equations" appeared for the first time in their modern vector form in Heaviside's work.

Then in the early 1890's a short but vigorous debate on vectors and vectorial methods occurred, involving eight journals, twelve leading publications, and over 30 scientists. Over half the publications were in the leading British scientific weekly journal Nature.

This debate, which Lord Rayleigh characterized by the statement "Behold how these vectorists love one another," together with **the** far greater difficulty of quaternion methods as compared to Gibbs' and Heaviside's vector analysis, spelled the death knell to the position of quaternions.

After some years, the Gibbs-Heaviside vector theory was firmly entrenched. The quaternion theory of Maxwell was forever modified **and** limited to the form we know today as "Maxwell's equations."

What really happened was that electrical physicists, faced with the formidable difficulties of quaternions, simply stampeded down the far easier road blazed by Heaviside, bowling over all protests against Heaviside/Maxwell/Hertz theory.

*Specifically, Heaviside wrote the subset of Maxwell's theory where gravity and electromag**netism do not interact and are mutually exclusive.**

In Western universities today, no one actually teaches Maxwell's true theory! Instead, Heaviside's curtailed reinterpretation — a severely restricted subset of Maxwell's theory — is taught.

And so that is the EM theory that is learned and applied.

One may rightly argue, therefore, that **Maxwell's theory has only been partially applied in its modern translated form, and the modern form of the Maxwell/Hertz electromagnetic theory is very much incomplete.**

To recover the stress factor in Maxwell's original theory, electromagnetics must take into account the *artificial* potential, which has a deliberately ordered, dynamic, deterministic structure. Scalar electromagnetics takes the patterned stress into account and deliberately uses it. Normal electromagnetics only considers the *natural* potential with randomized VPF structure implicitly assumed.

Scalar Waves ***and Polarization of*** *the* ***Vacuum***

Let us use a simple analogy. This will lead us into the profound implications of the deliberate vacuum polarization and curving of local spacetime that is provided by the vector zero summation (artificial EM potential) method of scalar electromagnetics.*

Visualize two sets of opposing and balanced forces pressing on the sides of a plate.

The forces sum to zero, so the resultant force acting on the plate for translation is zero. Hence the plate does not translate (move away or accelerate).

However, the plate is under internal stress (compression), and is in a quite different condition than when it has no external forces at all acting on it.

Now visualize the forces being applied with rigid rods welded to the sides of the plate, so that the forces may alternately "pull" as well as "push." Let the forces rhythmically vary, alternately "pulling" and "pushing", but always remaining balanced so that their summation for a translation resultant is constantly a zero vector.

The plate never accelerates or moves in translation, but it now contains an internal stress wave which rhythmically varies between compression and tensile stress. Rigorously the plate internally

*Vector **multiplications** that yield a zero vector product are also highly significant.

possesses a scalar stress wave, or a wave of "internal stress in the medium."

Now visualize a similar "plate-like" region of vacuum, with its virtual particle flux, instead of the material plate. Visualize two opposing sets of (unzipped) EM force fields, acting in and on the plate-like region, so that they rhythmically increase and decrease, changing direction also, but always with a zero vector summation.

In this case, there is an unzipped zero resultant EM force field (it is zero, by our assumed conditions), but there is a rhythmic oscillation of the intensity of the vacuum flux (intensity of the vacuum potential) and the vacuum structuring in the region. That is, there is a rhythmic and steady oscillation of the stress energy and structuring of vacuum, and hence of the curvature of spacetime, in **the** plate-like region.

Rigorously, this oscillation — which we call a scalar electromagnetic wave—is a gravitational wave, since the local curvature of spacetime is being oscillated.

Further, it differs from a "natural" gravitational wave in several respects:

1. It has a deterministic pattern or substructure,

2. It patterns or "polarizes" the vacuum,

3. It constitutes local curvature of spacetime, and hence localized general relativity - something which ordinary general relativity assumes cannot be accomplished,

4. It deterministically engineers the virtual state and local spacetime,

5. Since it is achieved by converting EM field energy into artificial G-field energy, one can expect tremendous gains of gravitational and inertial effects in and around electrical circuits utilizing such fields precisely,

6. It affects the Schroedinger wave and the probabilities of the

states being propagated forward by the Schroedinger wave. **With the scalar EM wave, one can thus deterministically engineer the emergence of quantum change, and violate one of the fundamental assumptions of quantum mechanics, that of the totally statistical nature of quantum change,**

7. The scalar wave can accomplish direct and localized change of the rate of flow of time (even to its reversal) and variation of mass and inertia, without concomitant translation of matter,

8. By locally curving and patterning vacuum spacetime, a stabilized standing scalar EM wave can provide macroscopic violation of the conservation laws, which rigorously depend upon a locally flat spacetime (Lorentz frame),

9. Since the components in the zero summation may be EM waves, and may be "locked together" and broadcast to a great distance and interfered there, effects at great macroscopic distances may be achieved, in violation of present assumptions of physics, and

10. Since excess negative time flow may be locally produced, antigravity, negative energy, and negentropy may be locally produced.

These are the startling implications inherent in correcting classical electromagnetics' improper treatment of zero-summed EM force vector systems in vacuum.*

By classically replacing the force vectors in a zero EM force field summation with a zero vector and discarding the components, one is totally discarding the fact that such a patterned system of oppositive forces forms a deterministically patterned stress in the vacuum medium.

By this innocuous error, classical EM discards electrogravitation, and avoids unification of gravitation and electromagnetics. It

*And its zero - multiplied force vector systems as well.

also avoids direct engineering of gravitation, vacuum/spacetime, inertia, rate of time flow, free energy, and quantum change.

Generations of Western physicists and electrical engineers have acquired a mindset in only one aspect of the vastly expanded electromagnetics actually available.

Vector Zero Systems and the Kaluza-Klein Approach

Prior to 1921, Theodor Kaluza applied Einstein's new general relativity to five dimensions.

He produced a unified theory of electromagnetism and gravitation.

In Kaluza's theory, the ordinary 4-dimensional gravitational field and the electromagnetic field are but two different aspects of a single more fundamental field: the 5-dimensional gravitational field.

Kaluza's theory was published in 1921, on the personal recommendation of Albert Einstein, who had had Kaluza's paper for two years.

In Kaluza's model, electromagnetics is the 5th dimensional aspect of the 5-d G-field, while the ordinary 4-d G-field is the intersection of the 5-d G-field with our ordinary world.

Five years after Kaluza's epochal paper, Oskar Klein explained where the extra space dimension — the fifth dimension — was. He modified it as "wrapped around" each point in our ordinary 3-space.

In the Kaluza-Klein model, then, an electromagnetic wave— which moves spatially only in the fifth dimension — does not move through our 3-space at all. Instead, on its trajectory it "flows around" each point in our 3-space along its path.

Using Kaluza's model, we may regard the EM field as the normal "bleed-off' or escape of the 5-dimensional gravitational potential in the 5th dimension.

Any gradient in a potential represents a bleed-off of the potential from the high point to the low point. The way to prevent effective bleed-off of a potential at a point is to add an equal amount of bleed-in! The bleed-out and the bleed-in at the point then become balanced, so that an equilibrium state exists in the potential.

By this mean, the magnitude of the potential can be fixed by the two opposing bleed processes in equilibrium.

Table 38. KALUZA GEOMETRY

- **THEORDOR KALUZA, POLISH PHYSICIST**
- **UNIFIED THEORY OF ELECTROMAGNETICS AND GRAVITY [1921]**
  - **5 - DIMENSIONAL SPACETIME**
  - **5 - DIMENSIONAL GRAVITY FIELD**
  - **ELECTROMAGNETISM IS THAT PART THAT OPERATES IN THE FIFTH DIMENSION**

Table 39. 5-D G-FIELD BLEED-OFF [KALUZA 5-D THEORY]

- **FAR EASIER TO BLEED-OFF AS ORDINARY EM FIELD**

- **VERY DIFFICULT TO BLEED-OFF AS ORDINARY 4-D G-FIELD**

- **E.G., BETWEEN TWO ELECTRONS:**

$$\frac{\mathrm{EM}}{\mathrm{G}(4)} \approx \text{[illegible]}$$

- **BETWEEN TWO PROTONS:**

$$\frac{\mathrm{EM}}{\mathrm{G}(4)} \approx \text{[illegible]}$$

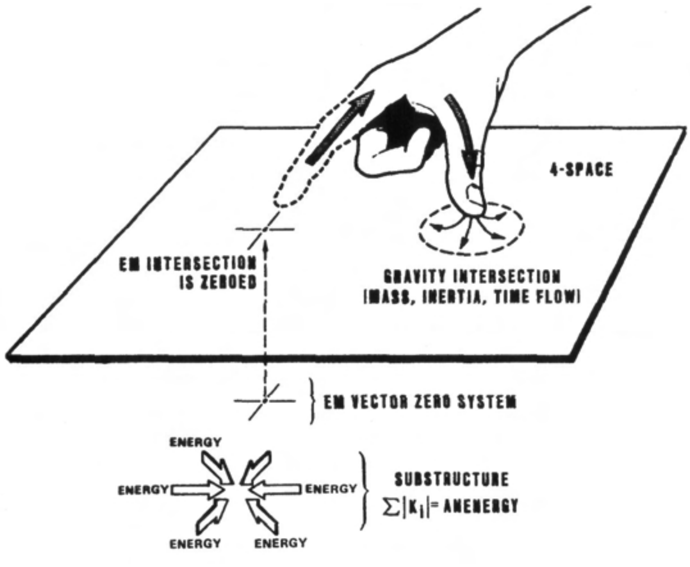
4-SPACE
EM INTERSECTION IS ZEROED
GRAVITY INTERSECTION (MASS, INERTIA, TIME FLOW)
EM VECTOR ZERO SYSTEM
ENERGY
ENERGY
ENERGY
ENERGY
ENERGY
SUBSTRUCTURE
Σ|K_i| = ANENERGY

Figure 63. An analogy for how scalar electromagnetics can engineer Kaluza-Klein unified field theory.

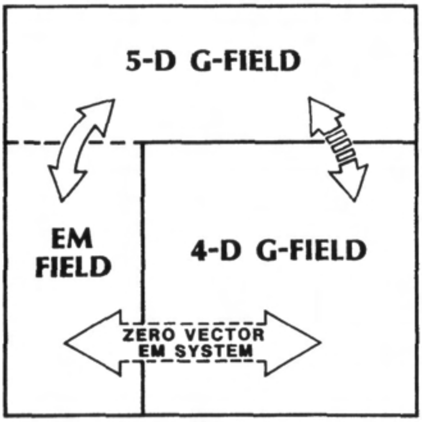
5-D G-FIELD
EM FIELD
4-D G-FIELD
ZERO VECTOR EM SYSTEM

Figure 64. The zero-vector system can engineer Kaluza-Klein theory.

WHERE DID THE EXTRA DIMENSION GO?

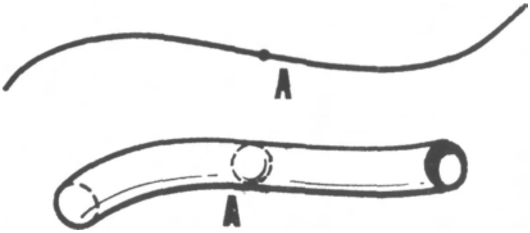

- **FROM A DISTANCE, A HOSE-PIPE LOOKS LIKE A WIGGLY LINE**

- **WHAT WE REGARD AS A POINT IN 3-SPACE IS REALLY A TINY CIRCLE AROUND ANOTHER DIMENSION OF SPACE**

Figure 65. The basic idea of Klein's explanation of where the extra dimension is.

Table 40. PHYSICAL REALITY IS ELEVEN DIMENSIONAL

- **IN THE KALUZA-KLEIN THEORY THE GAUGE FIELD SYMMETRIES BECOME CONCRETE**

- **THEY ARE THE GEOMETRICAL SYMMETRIES ASSOCIATED WITH THE EXTRA SPACE DIMENSIONS**

- **AN 11-SPACE IS REQUIRED BY THE GRAND UNIFIED FORCE THEORY**

PAUL DAVIES, SUPERFORCE.
1984, P. 160-161

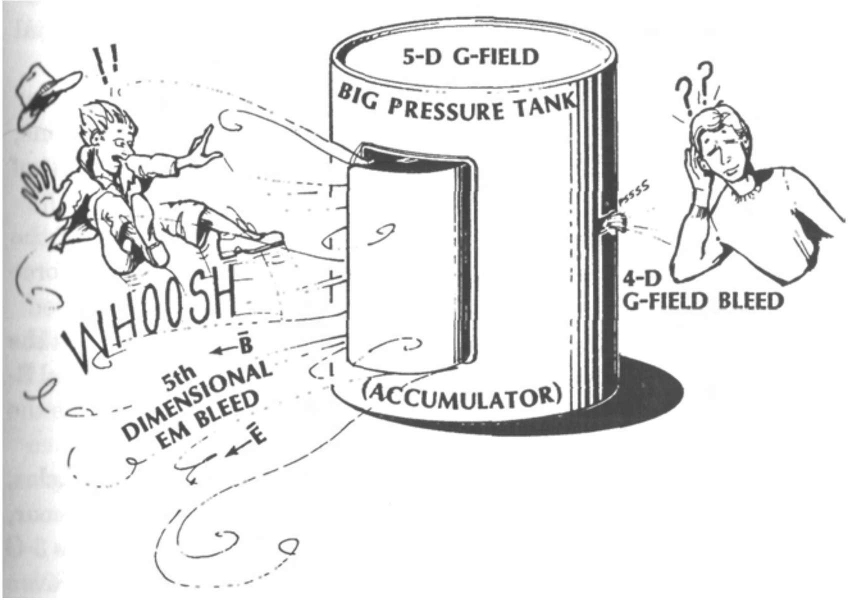
!!
5-D G-FIELD
BIG PRESSURE TANK
(ACCUMULATOR)
??
PSSSS
4-D
G-FIELD BLEED
WHOOSH
$\bar{B}$
5th
DIMENSIONAL
EM BLEED
$\bar{E}$

Figure 66. Normal EM bleeds off G-field.

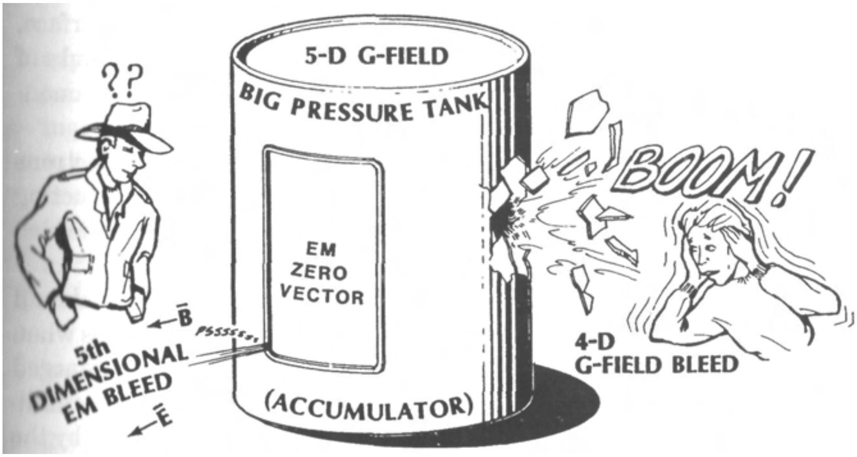
??
5-D G-FIELD
BIG PRESSURE TANK
EM
ZERO
VECTOR
(ACCUMULATOR)
BOOM!
PSSSSSs
$\bar{B}$
5th
DIMENSIONAL
EM BLEED
$\bar{E}$
4-D
G-FIELD BLEED

Figure 67. Scalar EM forces 4-dimensional gravitational field.

Note that this positively establishes that two or more zero-summed EM force field gradients produce and stabilize an artificial potential.

Usually, very little bleed-off of the 5-potential occurs in our normal 3-space. When the 5-space G-field is bled off on electrons, e.g., the resulting 5th dimensional E-field bleed-off is on the order of 10^{42} times as much as the 3-space G-field bleed-off.

This accounts for the enormous magnitude of the ratio of the electrical force between two electrons to the gravitational force between them.

Even when the 5-space G-potential is bled-off on and against the greater inertia of much more massive protons, its 5th dimensional E-field bleed-off is still on the order of 10^{36} or so times as much as the 3-space bleed-off.

Note that if we can stop the acceleration of the charged particles, bleed-off of the 5-space G-potential as EM force field cannot occur, and instead all of it must bleed-off as 3-space G-force field. This 3-G bleed-off normally occurs only minimally in electromagnetics, even for so-called "static" charges. This can be seen as follows:

When one charges a conductor such as a large metal sphere, for example, the electrons are unconstrained on the surface of the sphere. They are in constant agitation and motion on the surface, thus constantly bleeding-off the 5-space potential as myriads of squirming little E-fields. Only a very little is bled off as 3-dimensional G-field.

On the other hand, when one charges a dielectric, the electrons cannot be in as much agitation movement as they are on a conducting surface. Hence slightly more of the 5-dimensional G-potential is bled off as 3-dimensional G-field.

Therefore, charged conductors and charged dielectrics should exhibit slightly different gravitational and inertial properties when exposed to a gravity field potential difference. Highly charged capacitors, for example, should exhibit a small but detectable unidirectional thrust in vacuum. Apparently they do, as exhibited by the Biefield-Brown effect.

In addition, different materials should exhibit slightly different fall rates gravitationally. Again there is some evidence that they do, as exhibited by the Eotvos experiments in a modern reinterpretation by Fischbach et al.*

Electrogravitation Amplification Factor

If the movement of two free electrons could be entirely stopped, all EM bleed-off of the 5-d G-potential between them would be effectively zeroed or blocked. Then the 5-d G-potential would be forced to totally bleed-off into ordinary 3-d G-field.

For two electrons, this could increase the gravitational force between them by a factor of 10^{42}, if the EM bleed were perfectly blocked.

Such perfect blockage of the 5th dimensional EM bleedoff cannot be accomplished in practice. Normally, with the increase in potential caused by zero-summing, agitating electrons in a statically charged object agitate more intensely, moving and squirming in local, erratic constraints.

However, conversion efficiencies on the order of 10^{10} or even 10^{20} appear to be achievable in ordinary circuitry. In that case, gravito-inertial amplification gains of up to 10^{22} or even 10^{32} appear to be achievable. Gains of 10^{10} or more appear to be fairly readily achievable.

Time delay in producing the gravito-inertial effects may be experienced, however, because mass (nuclei) acts as a capacitance to incident scalar waves, exhibiting capacitive charging and discharging "time constants" when charging scalarly for a particular effect or to a particular pattern.

In these achievable amplification ranges, a single milliwatt of electrical power can control enormously powerful local gravitational and inertial effects, once the scalar EM charge is built up in the nuclei of the target material.

Excessive and frustrating "fiddling time" with the circuits is usually required, however, due to the charging "time constant" effect, the general lack of adequate measuring instruments, and the great uncertainty as to the phenomenology and major variables involved.**

---

*Ephraim Fischback et al, "Reanalysis of the Eotvos experiment," Phy. Rev. Lett., 56(1), Jan, 1986, p. 3-6.

••A series of scalar EM instruments, however, is being developed by Bedini, Schnur, Golden, **and** perhaps even Hewlett-Packard.

However, with persistence, in rather ordinary EM circuits and devices, the use of "summation to zero-vector" techniques and "multiplication to yield a zero-vector" techniques can lower the EM bleed-off of the 5-potential sufficiently to allow substantial local gravitational/inertial effects. By beaming and wave interference, effects at a distance can be achieved.

When gravito-inertial amplification of many orders of magnitude are involved, the assumption in ordinary general relativity of a local Lorentz frame (local special relativity, with local conservation laws applying) is readily violated.

Now one may have local general relativity, with concomitant violation of conservation laws, and local broken symmetry on a macroscopic scale.

Soviet papers already strongly point out that such unrestricted general relativity allows the violation of all conservation laws.

In the Soviet Union, scalar electromagnetics is called *energetics.*

Comparison of EM Concepts

Table 41 shows a comparison between three concepts of electromagnetics: classical, quantum mechanical, and scalar electromagnetic views.

As can be seen, scalar EM re-introduces Maxwell's original potential for electrogravitational variation and structuring of the stress energy density of vacuum spacetime. It applies Kaluza-Klein 5-dimensional unified theory in an engineering fashion.

It deliberately utilizes phase conjugation as a tool to achieve startling negative time/negative energy capabilities and results, including "free" energy and antigravity.

With scalar EM, it should be possible to overcome or circumvent most of the severely limiting assumptions of present physics and engineering.

The great conservation laws yield to direct manipulation and violation. The previously inviolate statistical structure of quantum change becomes deterministic and engineerable.

General relativity becomes an experimental scientific discipline in the laboratory, instead of a complex description of effects only observable in distant, massively curved regions such as on the surface of a star or near a black hole.

God does not play dice with the universe after all, just as Einstein suspected.

The Aharonov-Bohm Effect

In 1959, Y. Aharonov and David Bohm published a fundamental paper in **Physical Review** which pointed out the quantum mechanical implications of potentials.

The seminal paper is Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory," **Physical Review,** Second Series, 115 (3), Aug. 1, 1959, p. 458-491.

In classical electromagnetics, the potentials had largely been regarded as mathematical conveniences, having no real effects.

Table 41. Comparison of EM Concepts

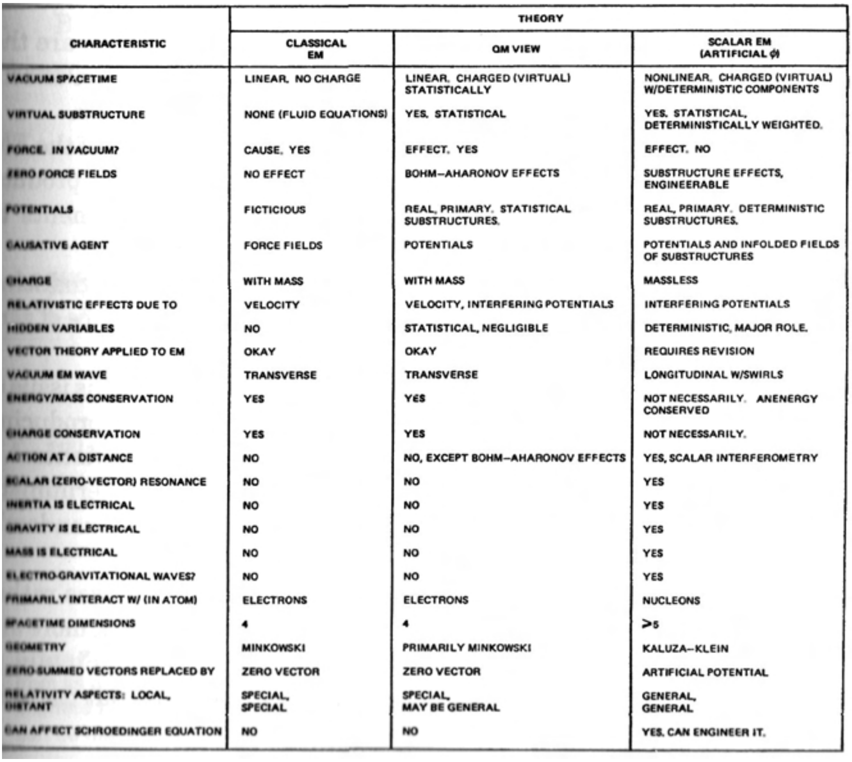

CHARACTERISTIC	THEORY		
	CLASSICAL EM	QM VIEW	SCALAR EM (ARTIFICIAL ϕ)
VACUUM SPACETIME	LINEAR, NO CHARGE	LINEAR, CHARGED (VIRTUAL) STATISTICALLY	NONLINEAR, CHARGED (VIRTUAL) W/DETERMINISTIC COMPONENTS
VIRTUAL SUBSTRUCTURE	NONE (FLUID EQUATIONS)	YES, STATISTICAL	YES, STATISTICAL, DETERMINISTICALLY WEIGHTED.
FORCE, IN VACUUM?	CAUSE, YES	EFFECT, YES	EFFECT, NO
ZERO FORCE FIELDS	NO EFFECT	BOHM–AHARONOV EFFECTS	SUBSTRUCTURE EFFECTS, ENGINEERABLE
POTENTIALS	FICTICIOUS	REAL, PRIMARY, STATISTICAL SUBSTRUCTURES.	REAL, PRIMARY, DETERMINISTIC SUBSTRUCTURES.
CAUSATIVE AGENT	FORCE FIELDS	POTENTIALS	POTENTIALS AND INFOLDED FIELDS OF SUBSTRUCTURES
CHARGE	WITH MASS	WITH MASS	MASSLESS
RELATIVISTIC EFFECTS DUE TO	VELOCITY	VELOCITY, INTERFERING POTENTIALS	INTERFERING POTENTIALS
HIDDEN VARIABLES	NO	STATISTICAL, NEGLIGIBLE	DETERMINISTIC, MAJOR ROLE.
VECTOR THEORY APPLIED TO EM	OKAY	OKAY	REQUIRES REVISION
VACUUM EM WAVE	TRANSVERSE	TRANSVERSE	LONGITUDINAL W/SWIRLS
ENERGY/MASS CONSERVATION	YES	YES	NOT NECESSARILY. ANENERGY CONSERVED
CHARGE CONSERVATION	YES	YES	NOT NECESSARILY.
ACTION AT A DISTANCE	NO	NO, EXCEPT BOHM–AHARONOV EFFECTS	YES, SCALAR INTERFEROMETRY
SCALAR (ZERO-VECTOR) RESONANCE	NO	NO	YES
INERTIA IS ELECTRICAL	NO	NO	YES
GRAVITY IS ELECTRICAL	NO	NO	YES
MASS IS ELECTRICAL	NO	NO	YES
ELECTRO-GRAVITATIONAL WAVES?	NO	NO	YES
PRIMARILY INTERACT W/ (IN ATOM)	ELECTRONS	ELECTRONS	NUCLEONS
SPACETIME DIMENSIONS	4	4	>5
GEOMETRY	MINKOWSKI	PRIMARILY MINKOWSKI	KALUZA–KLEIN
ZERO-SUMMED VECTORS REPLACED BY	ZERO VECTOR	ZERO VECTOR	ARTIFICIAL POTENTIAL
RELATIVITY ASPECTS: LOCAL, DISTANT	SPECIAL, SPECIAL	SPECIAL, MAY BE GENERAL	GENERAL, GENERAL
CAN AFFECT SCHROEDINGER EQUATION	NO	NO	YES, CAN ENGINEER IT.

Table 42. THE AHARONOV - BOHM EFFECT

- **IN FIELD-FREE REGIONS**
  - **E - FIELD IS ZERO**
  - **B - FIELD IS ZERO**
- **POTENTIALS STILL EXIST**
  - **CAUSE REAL EFFECTS**
  - **INTERFERENCE IS KEY**
  - **CONTAIN SUM-ZEROED SUBSTRUCTURE**
- **DOES NOT FOLLOW FROM**
  - **MECHANICS**
  - **CLASSICAL ELECTROMAGNETICS**

However, according to quantum mechanics, the potentials are the real entities, while the electromagnetic force fields are just effects derived from the potentials by imposed operations on them.

Aharonov and Bohm showed that, even in the absence of the EM force fields, the potentials may still exist and interfere to produce real effects in physical systems. They suggested experiments to prove these predictions.

Interference of the potentials is the key mechanism producing real effects in charged particle systems, even in the presence of zero E-field and zero B-field.

Unfortunately, Aharonov and Bohm did not address this issue of zero-summed systems of EM force fields as one way of producing artificial potentials having unique interference characteristics.

That is, by implication they did not consider the deterministically substructured "artificial" potential, but only the randomly structured "natural" potential.

At any rate the "AB effect," as it came to be called, was rather hotly debated and fiercely resisted over the years. Steadily, more and more experiments were added, validating the AB effect. Finally, in 1986 it came to be generally accepted as proven to any but the most die-hard skeptic.

The AB principle represents a violation of both classical mechanics and classical electromagnetics. It is absolutely required, however, by quantum mechanics. And experiments have now established it as proven beyond question.

The potential, then, can be a free and independent field. Further, it can directly act on charged particle systems, even when classical electromagnetics does not recognize the interaction.

Curl-Free Magnetic Vector Potential (A-Field)

As an example, in classical EM the vector magnetic potential (the A-field) had been defined as a mathematical convenience by the equation

$$\nabla \times \mathbf{A} = \mathbf{B} \qquad (4\text{-}20)$$

But if the potentials are real, then conceivably the A field can be loosed from its enchainment to the $\nabla\times$ operator. In that case, it

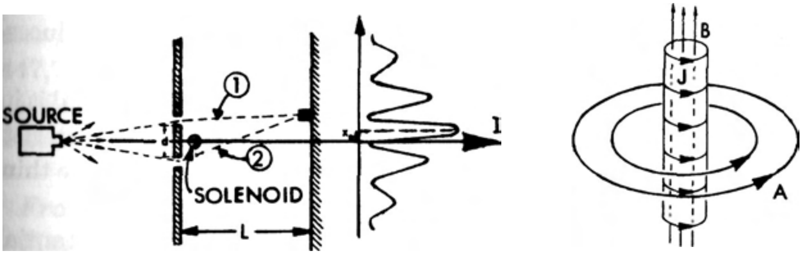

A - FIELD CAN INFLUENCE THE MOTION OF ELECTRONS.

MAGNETIC FIELD AND VECTOR POTENTIAL OF A LONG SOLENOID.

THE FEYNMAN LECTURES ON PHYSICS. VOLUME II

Figure 68. The curl-free A-field. The magnetic vector potential (A-field) may be freed from its attachment to the magnetic field (B-field) and electric field (E-field). It can interfere at a distance to cause real effects. From it, either E-field or B-field, or both, can be made at a distance.

becomes a free, new, and independent field of nature with potentially unique characteristics.

For instance, its defining equation shows that magnetic force field can be made from it, and the rightmost term of the equation

$$\mathbf{E} = -\nabla \varnothing - d\mathbf{A}/dt \qquad (4\text{-}21)$$

shows that its time rate of change makes an electrical field.

Let us explain this in simpler terms, and somewhat more precisely than conventional theory. We will use the Kaluza unified G-EM interpretation and electron flow in our explanation.

In the A-field, we have a certain kind of 5-dimensional G-potential which can bleed-off as EM force fields in two ways: (1) in a swirl fashion, where the vortex producing the swirl moves parallel to electron movement, and the swirling is a torque or spin, and (2) in a linear fashion, where the time rate of change of the A-potential produces a linear E-field on the electrons.

The first bleed-off as given by equation (4-20) constitutes the magnetic B-field, and the second bleed-off as given by the rightmost term of equation (4-21) creates a component of the overall electrical E-field. Bleed off of the electrostatic scalar potential produces the other component of the E-field.

Now in general relativity (GR) theory, "the" G-potential is just a conglomerate of many things, each of which has the characteristic of curving spacetime. "The" gravitational field is not a single thing at all, but is composed of a collection of many things.

Thus if we realize that both the electrostatic scalar potential (Ø-field) and the magnetic vector potential (A-field) are components of the 5-d G-potential, then we see immediately that bleed-off of these two components of the 5-d G-potential creates all normal EM fields.

It follows that, if we produce a zero-vector summation of the two or more EM bleed-offs, we are actually "putting as much back in" to the 5-d G-potential through its A and Ø components as we are taking out electromagnetically.

In that case, the 5-potential is in a state of equilibrium with respect to EM bleed-off in the fifth dimension. It is now forced to bleed-off in the only other way it can: as ordinary 3-dimensional gravitational field. Thus by vector-zeroing EM force fields, we turn KM field energy into G-field energy and vice versa, via the intermediary of the 5-potential.

At any rate, soon after publication of the Aharonov-Bohm paper, experiments showed that, if the magnetic field is trapped inside a long solenoid, a phase shift still is induced in the two-slit electron experiment, even though — classically — no contact of the enclosed magnetic field and the moving electrons occurs. This phase shift is explained by the fact that the freed A-field exists outside the trapping solenoid, even though the B-field does not. Consequently, interaction of this curl-free A-field with the electrons produces a phase shift of the QM interference detection pattern.

This proves that curl-free A-field is real and causes physical effects.

It also proves that a form of "action at a distance" is real, just as required by quantum mechanics.

Years ago, Frank Golden and this author — together with Dr. William Tiller — experimented with "free A-field" devices. Golden went on to develop prototype transmitters and receivers and a prototype underwater communication system.

Since that time, Gelinas has patented several curl-free magnetic vector potential (free A-field) devices. See U. S. patents numbers 4,447,779, May 8,1984; 4,429,288, Jan. 31,1984; 4,429,280, Jan. 31,1984; and 4, 432, 098, Feb. 14,1984. These patents are assigned to Honeywell.

Fractional Charges, Magnetic Monopoles, and Magnetic Currents

Fifteen years ago, a most distinguished gathering of physicists occurred in Trieste, Italy to review the physicist's conception of nature and how it developed. Seven Nobel Laureates were in the prestigious audience. The keynote address, appropriately, was given by P.A.M. Dirac, and — appropriately — was entitled "Development of the Physicist's Conception of Nature."

Dirac called attention to the possibility of fractional charges, and unexpectedly presented some of the data from experiments done lung before by Felix Ehrenhaft.

Years earlier Felix Ehrenhaft, former director of the Physical

Institute at the University of Vienna, had reported a number of experiments where — purportedly — he easily obtained fractional electrical charges.

Whereas Millikan had used liquid oil drops in his experiments to determine the charge of the electron, Ehrenhaft used little solid, red selenium oxide spheres. Not only could Ehrenhaft get a fractional charge, but he could actually vary the size of the fraction by varying the size of the spheres.

(Knowing what has been discovered in the last two decades on the complex bond structuring of water and liquid hydrocarbons, one might very well suspect that a solid — which has one less degree of freedom than a liquid — could possibly yield a very different experimental condition from a liquid, from first principles. However, no one seems to have noted that, or taken it into account.)

(Also, Ehrenhaft's results were replicated by other researchers who used his experimental methods. They published scientific papers, but that made not a whit to the scientific community, which simply rejected the experimental evidence.).

Dirac pointed out that Millikan himself, in his oil drop experiments that established the charge of the electron, reported one anomalous experiment which yielded a fractional charge. He discarded the results of that experiment.

Ending his speech, Dirac stated that it all made him wonder if perhaps Millikan did have a fractional charge in that one experimiment after all. And well he might wonder. Ehrenhaft — almost unknown now to the present generation of physicists — may have found not only fractional electrical charges, but the magic magnetic monopoles as well.

Over and over, Ehrenhaft reported the results of experiments which disagreed with the notions of magnetic theory at the time.

Again, some of these experiments were replicated, and reported by those researchers replicating them.

It didn't make any difference at all.

Now, finding a magnetic monopole — and the accompanying magnetic charge current — would be most satisfying to many theorists. That would make the equations of magnetics quite "balanced" with respect to electricity.

However, no one —at least whose results have been accepted -

has yet found a magnetic monopole or magnetic current.

Stanford researchers have perhaps come closest, and even reported the results of one exciting experiment that found the magnetic monopole. However, the team has consistently failed to be able to replicate those results. And in science, "no replication, no acceptance" is necessarily the dictum.

In a moment we will address a possible cause for this failure to find the magnetic monopole and magnetic charge current.

Some Ehrenhaft Experiments

Let's look at some simple, rather informal experiments that Ehrenhaft showed. We call attention to Alden P. Armagnac's excellent article , "Magic with magnetism," in **Popular** Science **Magazine,** June 1944, p. 130-133, 222, 226. In the first experiment, Ehrenhaft seems to have dissociated water with a permanent magnet.

First, Ehrenhaft seals two rods of pure iron, in holes through opposite sides of the U in a U-shaped glass tube filled with water. Then he uses the two rods — which protrude through the tube into the water — as the pole pieces of a horseshoe magnet — either an electromagnet or a permanent magnet. Dilute sulfuric acid — one percent by volume — is in the water, to give it a very slight acidity.

Bubbles of gas rise up through the twin vertical columns of water, and are collected and analyzed. Most of the gas is hydrogen, as is to be expected from common chemical interaction of the iron rods with the very dilute sulfuric acid. However, some of the gas — ranging from two to twelve percent — is oxygen, and most of the oxygen is produced at the north pole of the magnet.

The pole pieces may be short-circuited with wire, to insure they are at the same electrical potential and prevent any electrical dissociation of the water.

Without the magnet, all the gas liberated is hydrogen, as is to be expected. No oxygen at all is produced.

And another very interesting phenomenon occurs. A strong permanent magnet of the Alnico type gradually loses a marked fraction of its strength — say, 10% in 24 hours. It appears that the energy dissociates water — very similar to electrical energy from a battery draining away when electric current is used to dissociate water.

The real question is: Is this experiment a demonstration of magnetic current?

If it is, then the textbooks are wrong, and there's another constituency in magnetism that has not been taken into account. Further, if it proves magnetic current exists, it has also proven magnetic charge — the magnetic monopole — exists.

If it is not a demonstration of magnetic current, then what is it? Why don't you get any oxygen liberated when the magnetism is removed? What causes the magnet to discharge its magnetic strength?

These serious questions do not appear to have ever been resolved.

In all fairness, however, we must warn the reader that other factors in the water can very seriously affect the outcome of this

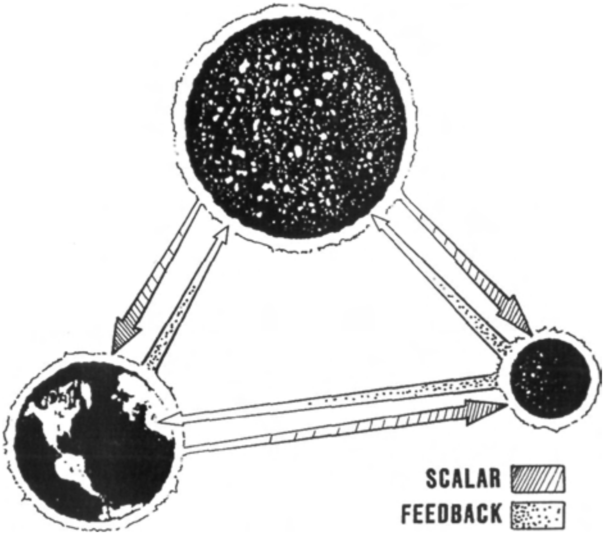
EARTH • SUN - MOON TRIAD

Figure 69. Earth, sun, and moon are locked into a triad system. They exchange both normal energy and phase-conjugate energy. The phase-conjugate energy (negative energy, negative time) streams affect all "subtle-energy" earth experiments

experiment. As we now know, the bonding structure of water is incredibly adaptive. Every bit of water you examine, for example, is absolutely unique in many respects. And in other respects, it's a sort of "single giant molecule."

Be forewarned also that subtle experiments with water are affected by just about everything under the sun. Including the sun!

Ehrenhaft himself was quite embarrassed when this "decomposition of water" failed to be 100% repeatable. It can be done — but you may have to work at it a bit. The best water for this experiment is the most natural, pure ground water you can obtain.

At any rate, it may be that these experiments of Ehrenhaft's are related to the old notion of "watergas" — the changing of water in some fashion so that, as it enters an internal combustion chamber, it decomposes into hydrogen and oxygen, and is ignited by the spark and burns as a fuel. In the typical watergas case, the experimenter has a "green powder" which, when added to water, will allow it to be burned as a fuel when it is used in an ordinary engine. Such a capability, for example, seems to have been exhibited at least once, to the U.S. Navy, by an unorthodox experimenter named Anderson.

Of course, if fractional charges exist and can be controlled (and Ehrenhaft's experiments certainly seem to confirm this), then it is conceivable that a material could be found that would do the "green powder" function. That is, if the bonding electron that bonds hydrogen and oxygen in water were replaced by a fractional charge, the "fractionated water molecule" could just barely be holding together an a liquid. A strong hydroshock — such as going through the jet nozzles in a carburetor — could then cause many of these molecules to break apart, effectively dissociating the water.

A possibly related phenomenon was reported by Le Bon in his Evolution of Matter. The experiment was to hold a piece of ordinary aluminum under the surface of mercury in a container for a period of time. Then the piece of aluminum is taken out and thrown into a container of water, where it violently dissociates the water. At least one colleague was able to replicate this experiment by holding the aluminum beneath the mercury for 36 hours.

These Ehrenhaft experiments may also be related to the extensive use in the Soviet Union (and to some extent, in the U.S.) of magnetic treatment of water pipes so that boilers do not develop so much hard precipitate. Magnetic treatment of water, for example, can rigorously be shown to alter some of a liquid's properties.

Ament, for example, in numerous well-controlled repetitions has clearly shown an increase in auto gas mileage by using a strong magnetic apparatus on the fuel line. Best results are obtained by a magnet which only has one polarity facing the fuel line, and using three stages where the middle polarity is reversed. The effect of this apparatus is definitely scalar EM (gravitational) and expressed in the atomic nuclei in the molecules of the fuel flowing through the line.

Again, the Ehrenhaft experiments may be related to the strange results of Viktor Schauberger, who obtained motive results with water and natural energy that have not been duplicated before or since, so far as anyone can tell.

Schauberger utilized "implosive" energy rather than "explosive" energy — strongly suggesting he was using the negative time aspects of energy, hence negentropy. Schauberger also conceptualized this energy in a spiral form — and that is precisely the type of motion Ehrenhaft observed in many of his bubble experiments. That such negative time/negative energy is "natural" energy will be apparent from our next chapter, where we point out that all living things must use negative time/negative energy processes if they are to defeat the inexorable law of entropy — the increasing disorder of a system as its positive flow of time progresses.

Note, **for** example, **that a** system **that is 20% efficient in positive time would be 500% efficient in negative** time.

Let's look at another of Ehrenhaft's magnetic experiments.

Taking the air gap between two very flat, very parallel pole pieces of sufficient width, one obtains a very accurately-parallel, uniform magnetic field in between them. Now suppose you place some very small particles of material—both magnetic and nonmagnetic — in the air gap.

Common sense assures us that, if closer to the south pole, for example, one of the little magnetic particles will be attracted by **the** preponderance of south polarity. That is, a little magnetic north **pole** will be induced on the side of the particle nearest the south polepiece, and a little south pole will be induced on the particle on the side away from the south polepiece. The little particle will then be attracted to the south polepiece.

Again, common sense assures us that, if closer to the north pole, the situation is reversed and the little particle is attracted to the north polepiece.

Well, that isn't quite what happens. Some move to the south polepiece, and some move to the north polepiece. And it seems to vary by the type of material the little particle is comprised of.

Now try another experiment. Let's make a permanent magnet motor. Understand, this one is not quite "for free," for we'll consume "fuel"—the strength of the magnet and the acidified water solution.

In this one, we use a glass cylindrical cell, with one vertical iron rod at the top and another at the bottom, and the rods serving as the pole pieces of a giant horseshoe magnet. (Or, for ease you can use an electromagnet). The cell is filled with the same acidified solution as before.

Without the magnet, hydrogen bubbles can be seen rising as before.

Now put on the magnet (or turn on the switch). Voila! the situation with the bubbles changes drastically. A spectacular miniature merry-go-round — a "whirligig" — of bubbles forms between the faces of the poles and parallel to them.

Add some copper particles to the solution. The particles will rotate in the same plane, but in opposite directions from the bubbles. The stronger the magnetic field, the faster the particles and bubbles will contrarotate.

If the magnetic field is reversed, the rotation of both the bubbles and the copper particles is reversed.

Only a few months after this was reported in Popular Science Magazine, Brother Gabriel Kane of Manhattan College and Charlea B. Reynolds of the Federal Communication Commission confirmed Dr. Ehrenhaft's discovery of magnetic currents.

They made a drop of copper sulfate solution spin between the polepieces of a permanent magnet. They also made an interposed microscope cover glass spin from the rotation of the copper sulfate.

It was hinted they were heading toward power machinery. If so, nothing else was heard — possibly because they never solved the problem of consuming the magnet's strength and the acidified solution.

Bill Mueller, the present inventor of a free-energy device, has also demonstrated the spinning of a piece of plastic in a magnetic field.

After WWII, Dr. Ehrenhaft returned to his post in Vienna, published later work in French and other scientific journals, and died not long after.

What is a magnetic monopole?

First, let's reiterate what we mean by "magnetic field." It'll take a little bit to get there, so please bear with me.

By electric charge, we mean the virtual photon flux connected with a particle of mass or an object (of mass). One type of charge (by convention, positive) is modeled to constitute a "source" of this flux. The second type of charge (negative) is modeled to constitute a "sink" of this flux.

Everything in electromagnetics is fluid flow, you see. The old guys who formed the theory believed that electricity was a thin material fluid. So they modeled it that way.

They modeled the positive charge as an everlasting "sprayer" and the negative charge as a "sink" or everlasting accumulator.

They reasoned that of course this "stream" of electric fluid could "swirl" or rotate, just as normal fluids can.

This "swirl" in the stream of virtual photons is magnetostatic scalar potential — or magnetic pole for short.*

Obviously there can be two directions of swirl — one right-handed, and one left-handed, with respect to the direction of the virtual photon stream. So there are two kinds of "magnetic pole" — north (right-hand swirl) and south (left-hand swirl).

The strength of one of these "swirls" can vary, of course, and that's a variation in the strength of the corresponding pole.

You can also think of the "pressure" of the swirl, so to speak, **at** least for modeling purposes. In that model, north swirl is "higher

*Rigorously, a permanent magnet does not produce an observable B field in vacuum! Instead, it produces the dynamic, swirling "potential" for that field — i.e., an **unzipped** field. A particle of observable mass placed in this field integrates and "zips" — producing the observed B-field on and of the spinning observable particle.

than ambient" pressure, and south swirl is "lower than ambient" pressure, where "ambient" means "with respect to right-hand swirl." The swirl pressure is then considered to "bleed off/flow" from north pole to south pole. This is modeled by Faraday's "flux lines of force." This bleed off (gradient) from north pole to south pole is called the magnetic force field, or just "magnetic field" for short, even though in vacuum it's really an *unzipped* magnetic field.

Note that the original modelers didn't consider negative time and negative energy at all. And they didn't know anything about phase conjugation — in fact, the electron had not even been discovered when all this model was originally put together.

But back to the model. Considering the swirling fluid as "conserved", if you produced a concentration of one kind of swirl at one place, you had to decrease the concentration elsewhere, which is the same as producing and adding opposite swirl there. Thus, conceptually, in the simplest case (only two swirls allowed), a north pole and a south pole always occur together in a pair — though of course there's no limit to the distance between the pair.

Notice that if you have a conservative field such as we assumed, you are keeping your overall frame of observation Lorentzian — in other words, conservative. You do not have a "curved spacetime" to speak of, by assumption.

Now of course, you can have compound and very complicated magnets and magnetic fields — you don't always have to have everything so neat and simple. In that case, conceptually you could have — for example — one north pole and two south poles.

In fact, it's easily possible to demonstrate that: a bar of magnetic material is simply magnetized so that a north pole is in the middle, and a south pole is at each end. Now bend it around a bit, and you get the point.

However, now let us suppose that we curve spacetime itself, locally — something they avoid in present physics by pure assumption.

Specifically, we curve our local spacetime in such a fashion that there's more right-hand swirl in the local region than there is left-hand swirl there.

In that case, we have a north-pole region there , all by itself. We have a magnetic north monopole. If that "curved spacetime" region of trapped swirl is moving, that constitutes magnetic current, and it may be either positive (north polar) or negative (south polar) flow.

It's as simple as that! Only when you curve local spacetime can you get a magnetic "charge" or monopole. Everyone who's trying to demonstrate a magnetic monopole in uncurved spacetime is doomed to failure a priori. It takes but little reflection to see that, conceptually, a magnetic monopole is a curvature of local spacetime. Indeed, each of the two poles of a bar magnet is precisely that anyway.

Now a monopole can indeed be "made", by bending local spacetime correctly. It can be made in two ways: in the virtual state vacuum itself, or in and on physical materials. It's easiest to make it in the virtual state: just curving the local spacetime will do that. When this "excess swirl flux without corresponding antiswirl flux" is impressed into and onto matter, however, the atomic nuclei are affected, producing excited "magnetic states." In addition, a great deal of physical stress is created in the material, due to monopolar repulsion between the various "nuclear monopoles deposited in it."

If a sufficiently strong monopole intensity (magnetic charge) is produced in the material, the material will violently fragment from internal stress. When a metal, for example, undergoes such **a** "monopolar fragmentation", it is common to see "tiled edges" of the metal rupture, much as twisted little "planks" or tiles of metal unfurled and "split." Such a break may be a positive signature of the involvement of virtual magnetic monopoles (i.e., by the metal having acquired virtual magnetic monopolar charge, causing the disruption).

This precise type of metal fracture has been very well demonstrated in experiments by Canadian experimenter John Hutchinson. Anomalous, powerful breaks in metal have been obtained by Hutchinson where the two ends that break apart both have a north pole or a south pole.

The anomalous **hole in the fuselage of the Arrow DC-8 destroyed at Gander Air Force Base, Newfoundland on Dec. 12,1985,** strongly suggests that an electromagnetic missile struck the aircraft, inducing monopolar effects. That would also match the- **three** eyewitnesses' description of an anomalous glow on part of the aircraft as it faltered and then fell.

Supercharges and Fractional Charges

Now without further ado, let me just state flatly that the only way to get a fractional charge is to have an appreciable curvature of local spacetime.

To get that local curvature, it's going to be very difficult to use liquid. For weak effects, it will almost always change its bond structure and not localize the curvature in the nuclei. So "fractional charges" in liquid are much harder to come by.

However, with solids it's a different thing, particularly if you make the solid material very, very small and nonlinear — precisely like Ehrenhaft's little red selenium oxide spheres. There you can get **the** curvature to localize without so much bleedoff through the bonding lattice.

The greatest electrical wizard of all, Nikola Tesla, reported producing electrons with 50 times the normal electrical charge of an electron. That's readily understandable if you look at what's happening to the charge of a particle: To have 50 times as much charge, the electron must be exchanging virtual photon flux with its surrounding vacuum at 50 times the rate of the normal vacuum.

In other words, *it's the vacuum (spacetime) which must be*

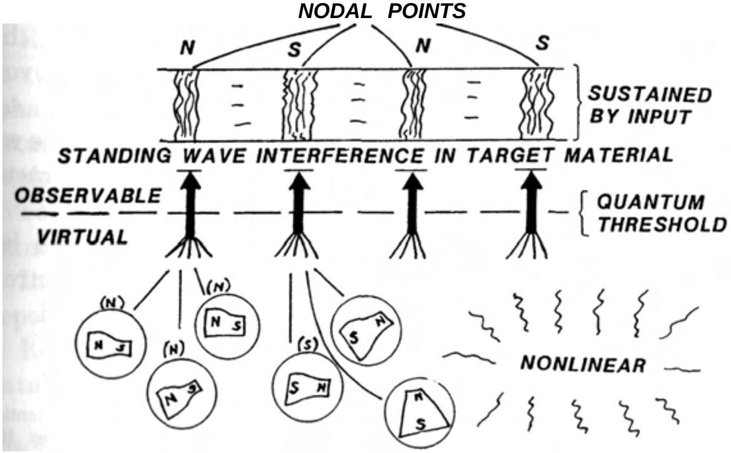
NODAL POINTS
N
S
N
S
SUSTAINED BY INPUT
STANDING WAVE INTERFERENCE IN TARGET MATERIAL
OBSERVABLE
VIRTUAL
QUANTUM THRESHOLD
(N)
N S
(N)
N S
(N)
N S
(S)
S N
S N
N S
NONLINEAR

Figure 70. Forming magnetic monopoles with a standing scalar EM wave. In a nonlinear vacuum (locally curved spacetime), virtual magnetic monopoles can be formed and deposited in materials.

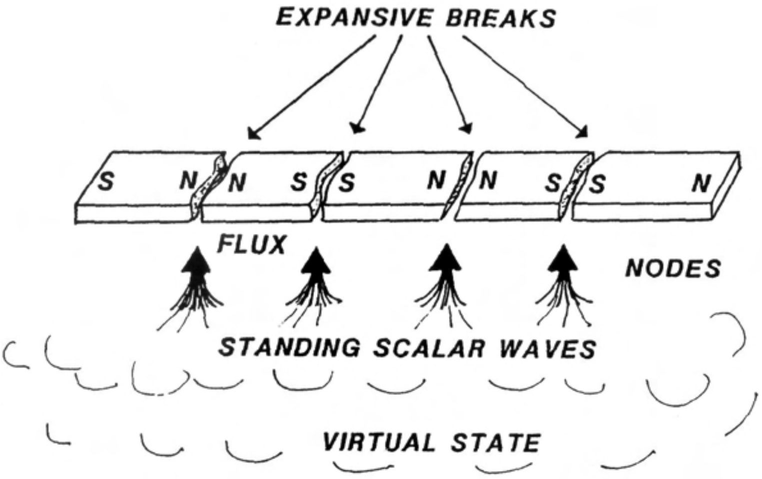

Figure 71. Metal fracture by depositing magnetic monopoles. Demonstrated by Canadian researcher John Hutchinson.

changed, not the electron! Actually, that happens anyway whenever the electron is in an electrical potential field. There are many unrecognized assumptions in the postulate that the electron charge is fixed.

If you curve local spacetime in one direction, you get the charge, of an electron increased. If you curve it in the other direction, you get the charge decreased. In that case it's called a *quark.*

At any rate, our own scientists have been looking in the wrong way and in the wrong place for magnetic monopoles and fractional charges.

Work such as that of Ehrenhaft and others should be re-evaluated. And, most certainly, the basis of the present electromagnetic theory should be both re-examined and reformulated.

---

*Such as the implication that the free electron is always free to move, and that potential surrounding the electron always possesses a translating gradient when it changes. It does not consider the case where the virtual flux exchange between electron and vacuum increases isotropically, centered on the electron. That, of course, represents an *excited charge* of the free electron.

CHAPTER 5

EXTRAORDINARY BIOLOGY

Kervran's Proof of Biological Transmutation

In orthodox chemistry, one of the strongest dogmas is the stubborn insistence that it is impossible to create another element by chemical reaction. Most chemists also insist that all reactions occurring in living systems are chemical in nature. They believe fervently that chemistry can and must explain life itself.

In the early 1960's, a French researcher named Louis Kervran published work which flew directly in the face of the accepted chemistry dogma. Kervran reported the astounding results of his research showing that living plants were able to accomplish limited transmutation of elements. Kervran was then the Conferences Director at the University of Paris, and his first paper was published in **La Revue Generale Des Sciences,** July 1960.

What was so revolutionary was that, according to the prevailing wisdom of science, you can't transmute elements (permanently change the nucleus) except with enormous energy — certainly not with the microvolts and millivolts (and microwatts and milliwatts) that living systems can muster electromagnetically. *

Rutherford, the British physicist who discovered the nucleus of the atom, had shown in 1919 that you can bombard elements with alpha particles and transmute them. The accepted wisdom of today is exactly the same, except that the physicists have used heavier and heavier "bullets" in their artillery approach. No one has tried a controlled approach, for the catchecism is that you have to use the *wham it harder!* approach.

In other words, to most scientists the whole thing had to be preposterous, and Kervran had to be deluded.

Kervran published further details of his work in a book, **Transmutations Biologiques,** Maloine, Paris 1962. But the initial reaction of most scientists was disbelief and skepticism. Few scientists would stoop to repeating Kervran's experiments, which of course they knew could not work anyway.

---

*Note, however.that since gravity is infolded EM, one can have extremely powerful infolded EM, yet only have miniscule electrical (outfolded) residues. Thus the actual "available power" in artificial biopotentials may not be quite so small after all.

Actually the effect is widespread amongst living systems. As Kervran pointed out, the ground in Brittany contained no calcium; however, every day a hen would lay a perfectly normal egg, with a perfectly normal shell containing calcium. The hens do eagerly peck mica from the soil, and mica contains potassium — a single step below calcium in the standard table of elements. It appears that the hens may transmute some of the potassium to calcium.

Further, if one tests this assumption, it is quickly shown to be true. Hens denied calcium but not potassium, stay perfectly healthy and lay perfectly normal eggs. Hens denied both potassium and calcium will be sickly and lay only soft-shelled eggs. If these sick chickens are allowed to peck only mica — which they will frantically do — everything returns to normal again.

Most orthodox scientists nevertheless remained skeptical or downright hostile.

However, a few other scientists began to repeat Kervran's

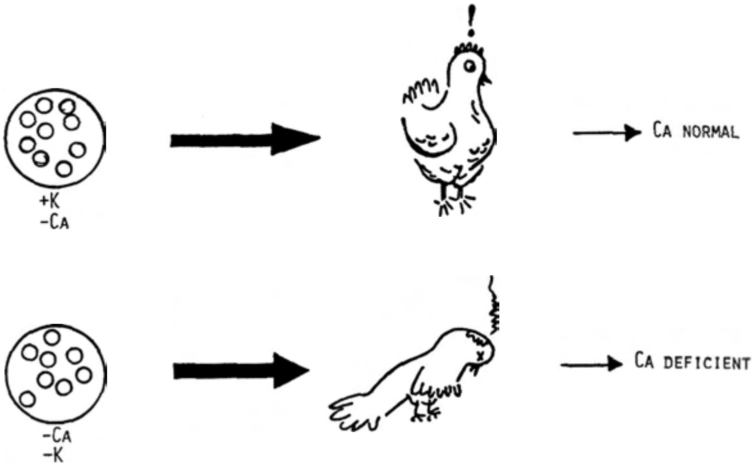

Figure 72. The Kervran effect. A biosystem can accomplish limited transmutation of elements.

experiments and replicate his results. Several of these corroborating scientists were (1) Professor Hisatoki Komaki, Chief of the Laboratory of Applied Microbiology at a leading Japanese university, (2)Professor Pierre Baranger, Head of the Laboratory of Chemical Biology of the Ecole Polytechnique in Paris, and (3) J.E. Zundel, then head of a paper company with a chemical analysis laboratory, and later a chemical engineer of the Polytechicum School of Zurich, Switzerland.

Later work by Zundel was particularly decisive: he utilized the mass spectrometer at the Microanalysis Laboratory of the French National Scientific Research Center, and neutron activation mass analysis at the Swiss Institute for Nuclear Research in Villigen to positively confirm an increase in calcium of 61% to an accuracy of 2%. Such results and instrumentation, of course, removed any doubt that the effect could be due to statistical variation. In the same experiments, the plants increased their phosphorus 29% and their sulphur 36%.

Komaki became head of a research laboratory at Matsushita Electric Company. There he conducted research conclusively proving that microorganisms (including some bacteria and two kinds each of molds and yeast) could transmute sodium into potassium. In fact, he placed a brewer's yeast product on the market that, when applied to composts, increases their potassium content.

Extensive work in the area has been done in the Soviet Union, where results similar to Kervran's have been substantiated.

Thus all doubt (to an open-minded scientist) was removed: living systems are able to change one element into another by some unknown means, using very feeble energy.

A noted French physicist, O Costa De Beauregard, suggested a mechanism for the transmutations, using weak force interactions and advanced waves.

No one — even Kervran himself— thought of negative energy/negative time interactions. The jury is still out on the actual mechanism, but it is absolutely clear that the transmutation does indeed occur.

The Japanese researchers, having replicated Kervran's astounding results to their complete satisfaction, recommended him to the Nobel Committee for a Nobel Prize for such epochal work. Thus Kervran became a Nobel nominee, though he was not granted the prize.

Kervran has since passed away, leaving behind his books and papers that point to a revolution in chemistry and physics - transmutation of elements at very weak energy.

Biological Transmutation Has a History

Actually biological transmutation — and transmutation of elements (alchemy) in general — has a history, of both results and suppression.

Louis Nicolas Vauquelin, a celebrated French chemist, discovered that chickens could produce more calcium in their eggshells than entered their bodies. Hence they had to be able to "create" the calcium, else their own bodies would have been completely depleted.

One of his contemporaries, however — Antoine Laurent Lavoisier — became the "father of chemistry." Lavoisier laid down the dictum that nothing was created. So chemistry fixed upon the notion that the combinations of elements could be shifted, but the element itself could not be transformed.

Not until the discovery of radioactivity did any crack in this solid wall appear. But still, the basic ideas of chemistry said the element couldn't be transformed chemically. It could only be transformed if one blasted the daylights out of it with an atomic or particle bullet. .

Today most chemists still hold that exact same opinion, unshaken.

To resume: Over a century ago, a chemist named Albrecht von Herzeele proved that germinating seeds somehow transmuted elements in the process. In 1873 von Herzeele published a book, **The Origin of Inorganic Substances,** where he showed research proving that plants continuously create material elements.

Even earlier, in 1822 an Englishman named William Prout had studied chicken eggs in incubation. He found that hatched chicks had more lime (calcium) in their bodies than was originally present in the egg!

Another French scientist named Henri Spindler discovered that, a kind of algae called Laminaria could create iodine.

A German researcher named Vogel had planted cress seeds in a

bell jar. They were fed nothing but distilled water; still, when grown they contained more sulphur than had been in the seeds originally.

Lawes and Gilbert, two British researchers, also found that plants could "extract" more elements from the soil than the soil actually contained in the first place.

Baranger performed thousands of meticulous experiments in plant transmutation of elements. He proved that the transmutations do occur. He also discovered that many things affected the germinating seed transmutation process: the time the seeds germinate, the type of light they are exposed to, the phase of the moon, etc.

None of these experimenters understood the transmutation process used by the living organism. But they proved beyond question that the process existed, and universally occurred.

Surplus-of-Energy Mechanisms Proposed by the U.S. Army

There has also been other very positive support for the thesis that if living systems transmute elements, they can produce a net source of energy in the process.

In 1978 an officially-funded effort of the U.S. Army Mobility Equipment Research and Development Command, Fort Belvoir, Virginia positively confirmed that mechanisms for elemental transmutations could occur in biological systems, from an energy consideration.

The work was performed under the direction of Emil J. York, Chief of the Material Technology Laboratory. Solomon Goldfein was the principal investigator for the effort. Robert C. McMillan, Chief of the Radiation Research Group of the laboratory, provided guidance on matters of physics and nuclear physics.

The abstract of the final report (S. Goldfein, Report 2247, **Energy Development from Elemental Transmutations in Biological Systems,** U.S. Army Mobility Equipment Research and Development Command, May 1978. DDC No. AD A056906.) reads as follows:

"The purpose of the study was to determine whether recent disclosures of elemental transmutations occurring in biological entities have revealed new possible sources of energy. The works of ***Kervran,*** *Komaki, and others were surveyed, and it was concluded*

that, granted the existence of such transmutations (Na to Mg, K to Ca, and Mn to Fe), then a net surplus of energy was also produced. A proposed mechanism was described in which Mg adenosine triphosphate, located in the mitochondrion of the cell, played a double role as an energy producer. In addition to the widely accepted biochemical role of MgATP in which it produces energy as it disintegrates part by part, MgATP can also be considered to be a cyclotron on a molecular scale. The MgATP when placed in layers one atop the other has all the attributes of a cyclotron in accordance with the requirements set forth by E.O. Lawrence, inventor of the cyclotron."

"It was concluded that elemental transmutations were indeed occurring in life organisms and were probably accompanied by a net energy gain."

The researchers also concluded that elemental transmutations occurring in life organisms are accompanied by losses in mass representing conversion to thermal energy, and that such energy probably is a net gain when compared to the amount required to effect the transmutation.

All in all, they concluded that the little cell with its feeble energy does quite well! It's in control of cyclotrons, and cyclotron forces, and direct conversion of mass to energy. Pretty good for a little bitty beastie, wouldn't you say?

Actually, one should point out that, according to nuclear physics, an atom gets a little heavier when it absorbs (usually by means of an orbital electron) a normal "positive energy" photon. That is, the addition of positive energy results in the addition of a little bit of "positive mass."

Negative energy, of course, does a similar thing to the nucleus — except that it adds "negative mass." Thus the nucleus of the atom, when it absorbs negative energy, gets lighter. This is seen in the external world as "loss of mass."

With our present nuclear physics, only positive energy is assumed except in extremely rare cases.

Thus the Army study — which was conducted and controlled by some excellent scientists — worked out a "loss of mass" the way they're trained to.

By adding some positive energy, the nucleus would gain some positive mass. **By adding some negative energy, the nucleus would lose some corresponding positive mass. The conventional physics then would equate this "loss of mass" as the direct conversion of mass to energy. And so it is, only it's conversion to negative energy!**

However, by pointing out the cyclotron mechanism in the cell MgATP, the Army researchers have made a most important contribution.

Note also that the whirling motion may be very much related to Viktor Schauberger's work and to Wilhelm Reich's work. Both of them worked with what they viewed as an unusual kind of living, spiraling energy.

All the orbital electrons of an atom also are whirling around in orbit, in the simplest model. Further, these orbits themselves move and rotate or precess.

Similar orbits and shells occur in the nucleus, at least in some models (several rather independent models are used there for specific things.)

It may be that a whirling, spiraling (cyclotron) energy motion is necessary to connect positive energy to orbital electron (negative charge) shells, and to connect negative energy to positive charge shells in the atomic nucleus. *

*Alchemy **and Unusual Critters***

In ancient times, the old alchemists pursued the dream of making gold. Obviously, if one could do that economically, one could become quite wealthy.

Just as obviously, the kings and rulers of the world took rather **a** dim view of such proceedings. After all, much of their own power rested on their ability to get and control gold. And if some "loose cannon" could make all the gold anyone wanted, then the national **treasury** of the king wouldn't be worth a plugged nickel. And that would finish the king, for he would be powerless.

•The spinning/whirling motion may be viewed as integrating the unzipped vacuum flux virtual vectors into zipped observable force vectors —just as great grandma's spinning wheel integrated fibers into continuous threads.

There are some unorthodox researchers today who take the view that the alchemists were stamped out—*not because they failed, but because they succeeded.*

I subscribe to the same view.

T.H. Moray had a process to "recover finely divided gold from quartzite sands." My personal, strong belief is that he possessed a practical transmutation process. His knowledge and techniques, of course, are still possessed by his sons, and reside through them in Cosray Research Institute, Salt Lake City, Utah.

The possession of such a technical secret may be one of the major reasons why the Morays have met with such intrigue, harassment, and suppression over the years.

To speak further on "making gold," we first have to present some details on some special "critters" that live, but that can't be observed through a normal microscope — even an electron microscope.

In that vein, toward the end of this chapter, we will present some of Royal R. Rife's fundamental discoveries. Pay particular attention to his discovery of "finer" living forms — which today we could only refer to as "living energy, virtual-state forms."

Let's call them **critters** for short.

At one time, when the earth was young and the radiation from the sun was different, conditions on earth were much hotter. Great volcanic activity and fiery eruptions were commonplace and nearly continuous. Huge storms, of size and magnitude undreamed of today, swept the primitive atmosphere. The oceans were frenzied.

Under those conditions, many types of "critters" were highly active. Most of the critters, for example, lived in and worked on the atomic nuclei of matter .

After all, the critters are living, virtual-state organisms. There's a continual exchange between the virtual state (the vacuum, or spacetime) and mass (the observable state). An atomic nucleus is like an island in the "virtual state ocean", and the flux interchange is like waves breaking onto the island and then washing back to sea. The critters live in that ocean, and wash upon, so-to-speak, the mass-islands and interact with them.

In those primal days, many of the present great mineral deposits of the earth were created due to the transmutation activities of the critters.

One kind, for example, lived in copper. In an "energetics" sense, this critter "ate" copper and "excreted" gold, so-to-speak. Much of the gold that occurs in great copper deposits today was formed this way in the old days under primal conditions.

When conditions on earth changed, these little "copper critters" ceased their incessant activity and became dormant, just as viruses can do. But the critters are still there in the copper ore, waiting to be activated.

And activate them you can! You can even get the critters into a solution, and then crystallize them out as crystals.

These crystals are what the alchemists of old called the **philosopher's stone,** with the power to transmute base elements into gold. There are several kinds of philosopher's stones; this kind is for copper.

At any rate, you can then place these special crystals on some copper (and add another thing or two), and restore them to a similar primal environment as of old. That is, heat them in an electric furnace. Blast them with terrible electrical bolts. Bathe them in intense ultraviolet light. That's just a nice, refreshing spring day for the critters!*

That stimulates them and revives them. They wake up after a long sleep — and they're immediately "hungry." So they go right to work on the copper. Boom! In a little bit there isn't any more copper, just mostly gold, with a little other miscellaneous residue thrown in, such as black ruby and silver (in the experiments of one of my close colleagues).

The gold is radioactive when first made. Fortunately, all isotopes of gold are very short-lived: just minutes suffice for the radioactivity to die away. So you wait half an hour and everything's okay.

That's all there is to it.

And if you do that and try to capitalize upon it, your life expectancy is about 24 hours.

I don't know whether or not biological systems, in their Kervran-transmutations at weak energy, deliberately manipulate similar "critters." I suspect, however, that they do, at least to some extent.

---

*Note the probable similar effects involved in the Miller-Fox-Urey experiments in biogenesis

The Cell Also Lives and Functions in the Virtual State

Obviously, to transmute elements the living system has to be able to directly affect and influence the atomic nucleus.

It has been shown that this is a cellular capability, for single-celled organisms can do it.

As we shall see, Rife's work showed that the living cell is connected to at least 16 intemested deeper levels of reality than a relative "point" under an ordinary microscope. Further, all levels are structured and organized.

Think of it! Each one of those levels is to the preceding level as microscopy today is to the normal world. Sixteen levels!

I think it's reasonable to state that the life of the cell is patterned and dynamically structured and functioning all the way into the virtual state; indeed, to very deep intemested levels of the virtual state. That is, it also functions *hyperspatially*.

We shall also see that the mind and thought involve these more subtle physical (though virtual) levels.

Thus the living virtual-state levels are a reality, for Rife proved it.

The living organized structures at each level are a reality, for Rife proved it.

The living ordering and control of dynamic functions on all those levels is a reality, for Rife proved it.

Those living virtual-level parts of the living organism — plant or animal — thus affect, function in, and reside in the atomic nuclei of the material that composes its bodily structures.

Beasties like bacteria and viruses also have living, organized energy structures in multiple levels of virtual state. Apparently, for these more primitive life forms, the virtual-state "energy part" can be separated and pass through a filter, then re-engender the physical form and/or itself cause the disease in a host! At least that is what Rife and other scientists showed.

"Bigger fleas have smaller fleas to bite 'em, And so on, ad infinitum."

Of course the living system can "work on" the nucleus and change it a little bit! If it couldn't do so, it couldn't stay alive and function in there in the first place!

The Kaznacheyev Experiments

Dr. Vlail Kaznacheyev is Director of the Institute for Clinical **and** Experimental Medicine in Novosibirsk.

For 20 years he has been directing highly unusual experiments with twin cell cultures. These experiments are vital to understanding disease and healing on a more fundamental basis than is presently utilized by orthodox medical science.

The Kaznacheyev experiments (several thousand) in the Soviet Union proved conclusively that any cellular disease or death pattern can be transmitted electromagnetically, and induced in target cells absorbing the radiation.

In the experiments, two sealed containers were placed side by side, with a thin optical window separating them. The two containers were completely environmentally shielded except for the optical coupling.

A tissue was separated into two identical samples, and one sample placed in each of the two halves of the apparatus.

The cells in one sample (on one side of the glass) were then subjected to a deleterious agent — a selected virus, bacterial infection, chemical poison, nuclear radiation, deadly ultraviolet radiation, etc. This led to disease and death of the exposed/infected cell culture sample.

If the thin optical window was made of ordinary window glass, **the** uninfected cells on the other side of the window were undamaged **and** remained healthy. This of course was as expected in the orthodox medical view.

However, if the thin optical window was made of quartz, a most unexpected thing happened. Some time (usually about 12 hours) after the disease appeared in the infected sample, the same features of disease appeared in the uninfected sample.

This startling "infection by optical coupling" occurred in a **substantial percentage of the tests (70 to 80 percent). From an orthodox medical view, these results were unexpected and unheard of.**

Further, if the originally uninfected cells were in optical contact

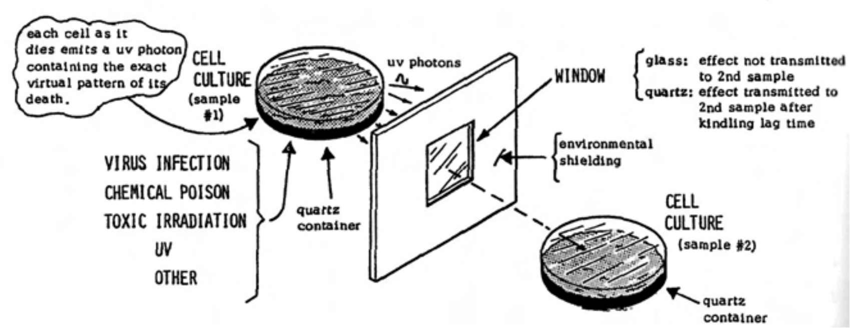
each cell as it dies emits a uv photon containing the exact virtual pattern of its death.
CELL CULTURE (sample #1)
uv photons
WINDOW
glass: effect not transmitted to 2nd sample
quartz: effect transmitted to 2nd sample after kindling lag time
environmental shielding
VIRUS INFECTION
CHEMICAL POISON
TOXIC IRRADIATION
UV
OTHER
quartz container
CELL CULTURE (sample #2)
quartz container

Figure 73. The Kaznacheyev effect. Thousands of experiments proved that (1) cellular disease is electromagnetic, and (2) it can be induced electromagnetically at a distance. Also called the cytopathogenic effect.

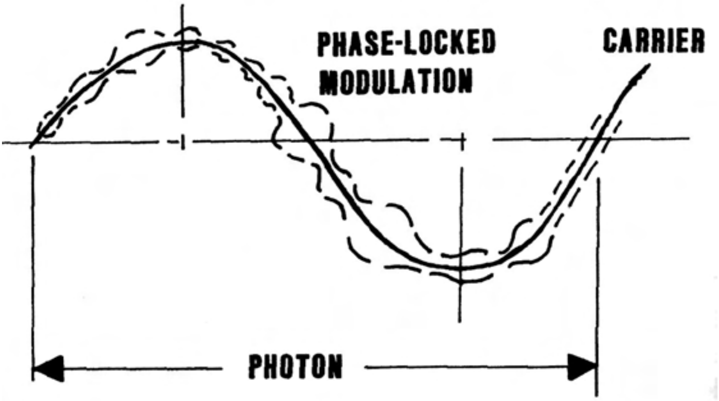
PHASE-LOCKED MODULATION
CARRIER
PHOTON

Figure 74. A photon is one oscillation of an electromagnetic carrier. It may have substructures that are modulations. Such a structure is called a "giant photon," or a "compound photon."

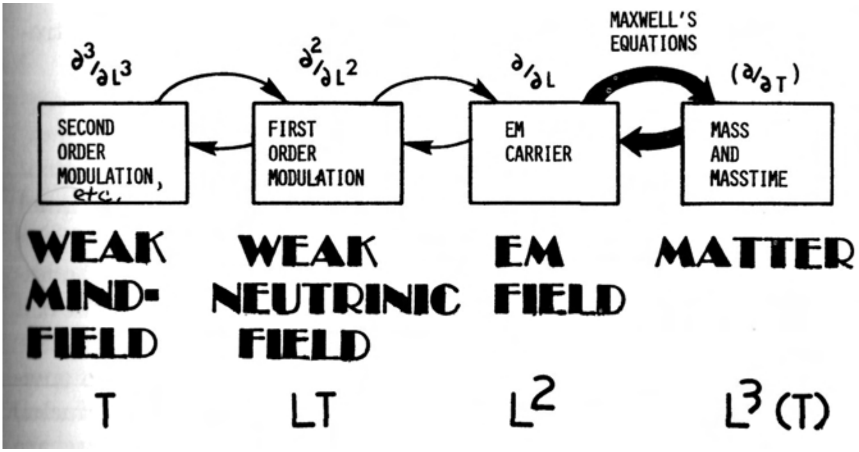
MAXWELL'S EQUATIONS
$\partial^3/\partial L^3$
$\partial^2/\partial L^2$
$\partial/\partial L$
$(\partial/\partial T)$
SECOND ORDER MODULATION, etc.
FIRST ORDER MODULATION
EM CARRIER
MASS AND MASSTIME
WEAK MIND-FIELD
WEAK NEUTRINIC FIELD
EM FIELD
MATTER
T
LT
L^2
L^3 (T)

Figure 75. Nested modulations are the key to hyperspatial engineering.

with the infected cells for 18-20 hours or so, and then were correspondingly exposed (optically coupled) to another uninfected cell sample, symptoms of the infection appeared in this third sample an appreciable portion of the time (20 to 30 percent).

Guided by A.G. Gurvitsch's work that showed that cells give off mitogenetic radiation (photons) that can affect other cells, the Kaznacheyev team sought an answer by looking for photons given off by the infected culture sample as its cells died.

They found that the cells in the infected culture gave off photons in the near ultraviolet when they died. The normal window glass was opaque to these near-UV photons and absorbed them. In that case, the uninfected culture on the other side of the glass was not exposed to radiation by the UV "death" photons from the dying cells, and they remained serenely healthy.

However, the quartz window was transparent to the UV "death photons". When the quartz window was installed, the UV "death photons" passed through it and were absorbed in the uninfected culture on the other side of the window. Most of the time, the uninfected culture which absorbed "death photons" sickened and **died** with the same disease symptoms.

The Kaznacheyev experiments proved conclusively that cellular death and disease patterns can be transmitted and induced electromagnetically.*

Structuring and Charging a Biopotential

Kaznacheyev thus demonstrated that a photon information/regulatory system exists in biological systems due to a continual influx of EM energy from outside the system. That is, the cells of the biosystem are charged with an electromagnetic potential, and additions and changes to the potential are continually received. The cell is thus in minute disequilibrium.

Usually the myriad of continual inputs from the external environment into the cell's potential charge pattern (in its atomic nuclei) may be taken to be potential changes whose substructures are *disordered.* In that case, no specific environmental effect is observed except slight fluctuations without order — a miniscule form of "heating."

However, if a continual *ordered* substructure exists in the input from the external environment into the cell's potential, the cell's potential will gradually "charge up" with that pattern.

An analogy will prove helpful. Imagine an accumulator, a largo pot, that holds a volume of water. Several pipes are connected to tho pot, some are inputs for water coming in, and some are outputs for water flowing in.

Imagine the inputs all containing "blue" water, just in slightly varying shades. The water in the pot is blue, and may slightly rise and fall in level as the input flow rates vary. The water in the pot may also vary slightly in its blueness as the inputs vary. However, it will still be blue.

Now suppose that yellow water starts flowing through one of the input pipes, and at a goodly rate. Slowly the water in the pot will start to turn greenish as a greater percentage of yellow builds up. In other words, the pot slowly charges up with some of the "yellow" charge, in the process acquiring a "green" charge.

---

*We point out that this effect has been investigated in both the infrared and ultraviolet. IR to UV may be taken as a single harmonic interval — an **octave,** musically speaking. The same effect can be reproduced in any other "octave" (single harmonic interval) of the electromagnetic frequency spectrum. The **reversal** of the effect can also be achieved in any harmonic interval. The mechanism for these effects involves the cellular biopotential, Popp's master cellular control system, and the deterministically-tailorcd substructure of photons.

The biopotential in the cell experiments works the same way.

A cell has a biopotential built up, which represents the "nominal equilibrium" of the scalar charge on the cell. This biopotential, being mostly a "sum-zero" of virtual state vectors, is centered in and on the atomic nuclei of the cell, constituting charge patterns in these atomic nuclei. The biopotential extends out of the atomic nuclei, through the electron shells, into and through the molecules, through the internal cell structures and membrane, and outside the cell.

From the atomic nucleus on out through the cell, every layered structure or organization of the cell will layer, structure, and organize the biopotential accordingly.

This organized, structured cellular biopotential is continually receiving "charge patterns" contained in incoming photons absorbed by the cell. The biopotential is also continually exhausting some of its biopotential charge pattern in the photons (heat, light, etc.) that the cell emits.

The Cell's Electromagnetic Breathing

Via structured photon exchange, the biopotential of the living cell thus "breathes in" the virtual state charge structure of its environment, and "breathes out" its own internal virtual state charge structure.

So, in the experiment, the uninfected cells are continually absorbing photons from their surrounding environment, and emitting photons back to it as well. According to our scalar EM view, each photon it absorbs has a substructure that depends upon the part of the environment from whence it came.

These "substructures" are actually patterns of the sum-zeroed virtual vectors comprising the potential of the absorbed photon carriers.

Normally, since a large number of very different substructures are continually being "input" into the cell's potential from the absorbed photons, the substructure of the cell's potential receives an essentially disorganized continual input from the environment. This liquates to the fact that the environment does not normally specifically influence or change the cell's potential with ordered information (organization).*

When a cell dies, it ceases to maintain the bio-dynamics that sustained its artificial potential (that part due to bio-ordering by its organized life processes, above the background level of its "inert matter" potential). The dead cell's built-up artificial potential then "discharges" by emitting a structured photon.**

Since this photon (energy) comes from an *organized potential drop, the virtual substructure of the emitted photon is organized.* The photon, then — among other things — carries the exact organized virtual charge pattern of the dying cell's disease.

We strongly insist on the quantum mechanical view here: All physical changes — chemical, material, mechanical, whatever — at root level are constituted and caused by virtual state interactions, in direct patterns of virtual particle exchanges.

In the full QM view, what's really going on in primary physical reality is just a complex set of patterns and changes in potentials anyway.

The Summed Virtual Structures ofKaznacheyev's "Death Photons" Physically Kindle the Disease

At any rate, the Kaznacheyev experiments showed that the dying cells from the infected culture emitted photons in the near UV that contained artificial (structured) potentials. The virtual-state, patterned-substructures in this photon flux directly represented tho cellular disease pattern caused by the original cell's specific infection.

In other words, as the infected cells died, they emitted "death photons" which contained the **template** pattern of their death condition.

When these "death photons" are absorbed into uninfected cells, their deterministic substructures gradually diffuse into the cell's bio-potentials. Gradually the biopotentials of the new cells are

---

*There may be sufficient ordered input from the environment, however, to have something to do with territoriality in living things, salmon returning to a fixed place to spawn, turtles returning to the same beach to lay eggs, the migration of birds, etc.

Note that this photon is emitted from an atomic nucleus. Hence it is a phase conjugate (time-reversed) photon. It will interact with the biopotential of targeted cells, and thus reach their own atomic nuclei. This is the mechanism for Kaznacheyev's cytopathogenic effect. See particularly C.W. Rietdijk, **Found. Phy , 7(5-6), Jun **77, p.** 351

Table 43. THE LIVING AURA: THE CELL'S ELECTROMAGNETIC BREATH.

- **VIRTUAL EM FIELD**
- **STORES VIRTUAL PHOTONS**
- **ENVIRONMENTAL INPUTS**
- **OUTPUTS BIOPHOTONS**
- **COHERES ORGANISM**
- **TENDS TO STABILIZE**
- **EXPERIMENT ESTABLISHES**

Popp, "Photon Storage in Biological Systems,"
Electromagnetic Bio-Information. 1979

"charged up" with the integrated pattern of the disease.

The master cellular control system of the biosystem is itself a dynamically changing, ordered pattern in the biopotential of the cells, which is centered in the atomic nuclei comprising the cell materials. As the bio-potentials of the cells gradually acquire the "death photon's" substructure pattern, this pattern is also diffused throughout (modulates) the master cellular control system. All the cells in the sample (or in a biosystem) are now slowly charging up with the "death photon" pattern.

As Popp discovered, photons continually "leak out" of the virtual photon master control system of the biosystem. Some of these *leakage photons* are observable photons, **but most are virtual photons.**

Further, they are *structured* photons.

In other words, as leakage photons spill out of the master control system, observable change is now being slowly initiated in the physical structures, biochemistry, etc. of the biosystem's cells — and those changes are in consonance with the integrated "cellular death pattern" of the originally infected cells.

Note particularly that it is already well-known in quantum mechanics/electrodynamics that, when a photon is emitted from **the surface of a dielectric body, the entire dielectric body participates in that emission. If a photon is absorbed on the surface of the dielectric body, the entire**

dielectric body participates in that absorption.

Thus as irradiation by the "death photons" continues, the "death structure" in the irradiated cells increases. It is spread throughout the cell culture by the master communication system, gradually charging the virtual state structure of that system with the death pattern.

Spillage photons from the cellular control system occur throughout the culture. These photons are structured with the death pattern, and gradually affect the cell and its biochemistry physically. The previously uninfected cells thus physically start to acquire and exhibit the symptoms and characteristics of the disease pattern that killed the infected cells.

Electromagnetic Infection Results in Physical Disease

The new cells are now electromagnetically infected and physically diseased.

After all, that is all a cellular disease is in the first place — physical, electrical, and biochemical changes in the normal functioning of the cell.

For a given pattern of changes in the cells, a specific "disease" exists in them.

It absolutely does not matter what causes this exact pattern. If the specific physical pattern is there, the specific disease is there.

Note that **any ghost pattern in the virtual state flux can charge up physical matter — that is, the atomic nuclei of a mass. All that is necessary is that a continual flux of thin virtual pattern continually bathe (irradiate) the mass's atomic nuclei.**

The eventual emergence of this "ghost template pattern" into observable physical reality is called **kindling.** Kindling is charging up one or more atomic nuclei with an integrated virtual charge pattern until the integrated pattern breaches the quantum threshold, resulting in emergence of that pattern into observable physical change.

A Possible Cure for AIDS

One of the things going for the "good guys" and EM defense against AIDS is that cells are a lot tougher than viruses. Thus even non-structured EM signals can be used to effect cures in many disease cases.

In fact, ordinary ultraviolet (UV) irradiation of the blood has already been utilized to cure or control severe infections, including severe viral infections. I am indebted to Dr. William C. Douglass for pointing this out to me, and for permission to reproduce the following information from his important newsletter, **The Cutting Edge, Nov.** 1987, p. 3. The following material is quoted verbatim, with no editing.

"It's amazing what you can find by nosing around in the dusty archives of a good medical library. I came across another remarkable therapy that the AMA and drug industry (or whoever is in charge of supressing non-toxic treatments that work) have shoved down the memory hole."

"Back in 1933, Doctors Hancock and Knott treated a patient dying of septicemia (blood poisoning) with ultraviolet irradiation of the blood.[1] The patient was moribund with a blood stream infection **and** obviously near death. (Remember that this was before antibiotics and there was nothing to lose.) The patient made a complete **and** uneventful recovery."

"Searching further, I found that in 1928 a similar terminal infection was treated by ultraviolet light to the blood. This patient also made a complete recovery.[2] "

"So in 1928, practically in the middle ages, an incurable disease, blood stream infection, was cured with ultraviolet light. With such **a** breakthrough why wasn't it tried again for *5 years?* According to **the** record, *another 6 years* passed before it was tried for a third time.[3]"

"Back in those days infection was the number one cause of death. You can't help but wonder how many lives could have been saved if doctors weren't so resistant to new ideas. Just imagine a cure for **AIDS** being set aside for 11 years. Yet bacterial infections of the blood were uniformly fatal in 1935, just like AIDS is today."

---

[1] Northwest Medicine, 33:200, 1934.

[2] Knott, AM. J. Surg., Aug. 1948, pp. 165-171.

[3] Am. J. Med. Sci., 197:873, 1939.

"Finally, in 1940, 110 cases treated with ultraviolet spectral energy were reported. The results were uniformly good. Between 1940 and 1948 many other conditions were successfully treated, including vein inflammation (phlebitis), polio and asthma. Up to the late 40's *over 40 thousand* treatments were given with ultraviolet blood irradiation."

"And now for the most interesting part. In 1947, Dr. G.P. Miley reported on 79 cases of virus infection.[4] Miley stated that ultraviolet blood irradiation therapy could be relied upon consistently to control an infection of a virus in a safe and efficient manner.[5]"

"AIDS is a virus. AIDS-II is a virus (the HTLV-IV leukemia and lymphoma now sweeping the world). Remember that these killer viruses are within the cell and any chemical agent that enters the cell to kill the virus will often kill the cell as well. But ultraviolet irradiation *kills the virus without harming the cell."*

"A fine piece of crystal can be shattered by exposing it to just the right frequency. You can be standing in the room and the energy from that frequency won't harm you in the least. Viruses have the same characteristics, and so, in my opinion, frequency irradiation of the blood in the ultraviolet range is our greatest hope for curing AIDS."

"But the treatment is simple, safe, inexpensive and unpatentable. That doesn't bode well for its future, at least until a few *senators* get AIDS."

The Mirror Cytopathogenic Effect and Factors Influencing It

The cellular disease induction effect was called the **mirror cytopathogenic effect** (CPE for short) by the Kaznacheyev group. Mirror CPE appeared only when the quartz or mica window was no thicker than 0.8 mm. A. F. Kirkin also duplicated the experiments using a thin plexiglas window.

There are conditions which enhance the effect, and others which inhibit or degrade it. Irradiation of the detector-culture with a low dose of UV prior to its optical contact enhances the effect, increasing it to certainty (99-100%). Increasing the temperature to 38.5 degrees centigrade also enhances the effect (from 37% to 90% for example).*

---

[4] Rev. Gastroenterol. 15 271-277, 1948.

[5] Am. J. Surgery, Aug. 1948, pp. 170.

A necessary condition for the success of the experiment is the rotation of the holder with its two optically-coupled samples at a rate of about 24-25 revolutions per hour. Optical contact between the inductor and detector cells for a minimum of 4-6 hours is necessary, after which the cell cultures can be separated. A longer contact time is necessary for complete development of the irreversible effect.

Both cultures must be maintained in complete darkness throughout the experiment. Use of the detector as a new inductor in a successive state reduces the effect by 20-30%. Three or four such stages is sufficient to eliminate the effect.

There is a seasonal variation in the results. In more than 15,000 experiments, monthly variations and daily variations were noted. (The present author's interpretation of this is that it is due to the monthly variations in the virtual photon substructure input from the moon to the substructure of the cell's bio-potential. The daily variation is due to the daily variations in the virtual photon substructure input from the sun to the substructure of the cell's bio-potential.)

Negative results appear more often in winter. (The present author's interpretation of this is that it is due to the fact that the scalar potentials of the earth and the biopotentials of each living cell on earth are lowered in winter by the weaker flux from the sun.)

Effects are correlated with the polarity of the interplanetary magnetic field. Negative polarity of the field usually precedes the appearance of mirror CPE. (This is because of the positive nuclei — which prefer one direction of the magnetic field over another). Disturbance of the geomagnetic field several days before a culture planting also results in enhancing the mirror CPE effect. (Disturbing the geomagnetic field provides a "dithering magnetic disturbance" in the atomic nuclei which "livens" them. Consequently their readiness to charge-up and emit structured virtual state charges is increased).

Kaznacheyev further discovered that the Sun's activity and the

---

•This is very important. Preconditioning the cells by "dithering" them in the frequency band **of** interest, or in a subharmonic band, "livens" the cells for developing the irradiated pattern. Thin is similar to a dither voltage placed on a missile fin, making it much easier and quicker **for** the fin to move when an actual order is placed on it.

Earth's magnetic field greatly affected the results of his experiments. Large flashes on the sun seem to inhibit the effect. (Such flashes cause substantially increased irradiation of the atomic nuclei by sun-emitted substructures, charging them mostly with this disordered substructure pattern and literally "burying" the disease structure several decibels below it.) In a season of active sunspots, the mirror CPE effect becomes highly unstable. (The sun emissions are sporadically jamming the effect.) Under active sun conditions, the effect varies from 90-100% on some days to complete absence on others.

Some Biological Warfare Implications

The Soviets reported detecting near-ultraviolet photons — bioluminescence — as carriers of the death/disease pattern.

However, scientists at the University of Marburg in West Germany also duplicated the effect in the infrared. This shows that bioluminescent photons in the near UV and in the IR can definitely carry "disease and death" information between cells. Further, integrating a continuing input of such photons coherently integrates the disease or death pattern from the virtual state into the observable state.

Note also that portions of the infrared spectrum are a subharmonic of the near ultraviolet. Harmonics are well-known in nonlinear oscillator theory, and biological systems are filled with nonlinear oscillators. It may be that harmonics and subharmonics are directly involved in the death pattern.

If so, the induction of such "death patterns" upon normal electromagnetic carriers is directly indicated. For example, modulations covering several octaves in the region of 10 gigahertz and above might be constructed that are the analogues of some particular cellular disease. This modulation pattern could then be added to a common microwave carrier — say in the communication band, from 3 to 30 megahertz. Say, that is, to something like the giant Soviet Woodpecker "over-the-horizon radar" signals as carriers.

In that case, a large population could be bombarded, even on the other side of the earth, with "death photons" whose virtual state substructures carry the particular disease pattern. With sufficient time, many of the targeted persons would develop the disease.

Note that, even if the power and/or irradiation time is reduced so that the absorbed "death photons" are insufficient to actually kindle the disease in the targeted population, a heightened change in the substructure of the biopotentials of the cells of the targeted persons is still accomplished.

In that case, a precursor pattern — a predisposition for that disease — exists in the targeted persons.

If the actual disease agent is now loosed on that population, the agent will be far more **infectious and lethal than it otherwise would.**

In this way, even diseases which normally do not kill or seriously debilitate the infected person can suddenly become very lethal agents indeed.

Influenza, the common cold, etc. can become devastating killers if the exposed population has been electromagnetically "pre-conditioned" for enhanced susceptibility.

What Kaznacheyev Hid: The Role of Phase Conjugation

If cellular disease can be electromagnetically induced, can it not be electromagnetically corrected or healed?

If one could time-reverse the exact signal structure (the information) that kindled the effect, and bombard the diseased cells with that reversed pattern, would not the cell deviate back to "normal" and be healed?

The burning question as to whether cellular disease conditions can be corrected by time-reversed disease signals must certainly have occurred to the Soviet experimenters.

It is highly significant that they did not openly publish those results.*

As we have explained in the sections on phase conjugation and scalar electromagnetics, there are really two major kinds of photons:

---

*Recent information indicates the strong connection of Kaznacheyev with the Institute of Physiology and Biophysics and the Frank Institute in Pushkino, just outside Moscow. Since these institutes are deeply involved in microwave and coherent microwave "directed energy" weapons, it is highly probable that the Soviets are applying Kaznacheyev's "death photons" to microwave weapons — such as the Woodpecker transmitters. If so, obviously they would develop phase conjugate countermeasure signals as well.

(1) the "normal" photon carries positive energy and positive time. (2) the "time reversed" or phase conjugate photon carries negative energy and negative time.

Further, the Soviets certainly knew all about phase conjugate signals. After all, they discovered and developed the effect. We discovered it only from the open Soviet scientific literature!

Let us assume that the "death photons" in the mitogenetic radiation emitted by the dying cells are ordinary photons. Their virtual state structures (in positive observer time) are exact "templates" for the disease pattern.

Now suppose we detect the "death photons" with a phase conjugator, which by definition will produce a time-reversed counterpart to the input signal detected. In other words, the death photons are allowed to strike a phase conjugate mirror (PCM). Time-reversed counterpart photons — carrying the exact time-reversed template of the death pattern — will be created and emitted by the PCM.

These newly emitted photons now carry the exact "healing pattern" for that specific "death/disease pattern that was received and detected."

Further, if we "pump" the phase conjugate mirror, we can greatly amplify the output pattern, and hence greatly increase the healing pattern!

If one records the pattern of the "death photons" for a specific disease, one could of course modulate that pattern upon ordinary photons/signals — such as the Woodpecker signals — and accomplish disease induction or precursor conditioning.

By phase conjugating the pattern of the "death photons," one can produce an exact antidote. One can modulate this specific healing pattern upon ordinary photons/signals — such as the Woodpecker signals — and accomplish healing induction for that specific disease.

In other words, one can create the healing pattern — the antidote, if you will, for any biological warfare agent. Cancer, leukemia, AIDS, viral diseases, bacterial diseases, whatever. One can create the antidote within minutes after the first symptoms of the disease or death pattern appear.

One can then simply add the negating (healing) signal to power line signals, television and radio signals, special transmitters, etc. —

and immediately start to "administer the antidote" to the irradiated population one wishes to protect. *Now one can see why the Soviets are so ready to expose the entire world to something like AIDS. It doesn't represent a real problem to them, the instant they decide to negate it.*

So they can devastate the rest of the world, with the assurance that their population is safe.

They can allow some of their own people to develop AIDS — and even some to die of it — as a deception plan to delude the West while Western populations are succumbing en masse.

Then they can snatch their own population right back to health, from "the brink of the grave," so to speak.

Our government must immediately develop the same capability. It is straightforward. As weapons and counterweapons go, it is enormously cheap. It can be immediately and widely implemented. **And** it can protect our population against AIDS or any other biological warfare strike by the Soviet Union.

We can save our people from the AIDS knockout already unleashed upon us by the Soviet Union.

First let us do that. *Then* let us negotiate.

Remember this: **You can negotiate with the Russians only from a position of strength. If you are weak, they will bury you.**

If we do not immediately develop this biological warfare countermeasure, we are already as good as dead.

Popp's Master Cellular Communication System

Dr. Fritz Albert Popp has already discovered and pointed out the "virtual state" master communication system that controls all cells in the body, and all their functions.

Based on a thesis derived to best fit experimental results by Ruth nnd others, Popp postulates that biological systems generally have the capacity to store coherent photons that come from the external world.

In other words, the biosystem is open to environmental communication and exchange.

He has shown that the cell population is in a quasistationary state that is far away from thermodynamic equilibrium, as pointed out by Ilya Prigogine.

Popp also concludes from his analysis that ultraweak photon emission within biological systems can influence chemical reactivity. In fact, his analysis strongly implies that "ultraweak" photon intensity can regulate the whole cell metabolism and related phenomena.

The cell takes up photons from external radiation. This includes both "observable" photons and "virtual" photons. Since it stores virtual photons, it stores charge, or biopotential changes. Since its stored virtual photons may be coherent virtual photons, it effectively "polarizes" or structures its stored photon charge, hence its biopotential.

The cell emits "spillage" photons — both coherent and incoherent — from its stored potential.

Although Popp only uses conventional "unstructured" photons in his analysis, he shows that, at the molecular level, there is a stationary equilibrium, as far as photon storage and emission are concerned, between the molecular photon traps, the cell population, and the external world.

It follows that *coherent* photon/charge inception from the external world can directly and precisely influence the cell's biopotential, hence its functioning and control, by information input.

Incoherent photon inception, on the other hand, can only grossly affect the cell, such as by heating or sporadic effects.

In his "Photon Storage in Biological Systems," Popp points out the master cellular communication and control system as follows:

"The photons which we have measured can be seen as a sort of "waste" from a virtual electromagnetic field with a high coherence. This field has a tendency to become stationary over the whole organism."

After additional analysis, he adds:

"....Consequently, biological systems must exhibit 'holographic' properties to an extremely high degree. The successful trials in

finding 'pictures' of various organs in each other organ, such as the ear, the hands, the eyes (acupuncture, iris diagnosis) support these conclusions. Our assumption that the entire genetic information of the DNA is stationarily delocalized over the body in form of genons may be seen as a further striking example."

"From this we can easily deduce that pattern recognition, as, for example, repair mechanisms and immunity, depends finally on the coherence of the photon field within the body."

Finally, Popp states a most important conclusion:

"...In medicine **new aspects have developed, and not only for cancer problems. Diseases in general can possibly be understood in terms of electromagnetic interactions within the organism."**

Scalar EM Comment on Popp's Communication System

Popp and his colleagues have produced most important work and results indeed. They only need to add the impact of the zero-summed/multiplied electromagnetics (electrogravitation).

As we cover in this book, the biopotential of the cell is rooted in the nuclei of the atoms of the cell's constituent materials. To be sure, every internal physical structure of the cell correspondingly "levels" and structures the biopotential. The overall cellular communication system is actually the exchange of "leakage" photons — both observable and virtual — throughout the overall biopotential of the organism.

Further, going beyond Popp's work, both the biopotential and the leakage photons have extensive, complex internal substructures. Leakage and intercommunication occurs laterally at all levels of the biopotential, and vertically among cells and substructures.

The master cellular control system's primary electrical conductivity path is not through the electron shells of the atoms, but is through the nuclei-to-nuclei scalar EM "biopotential levels" pathway.

With scalar EM methods, organized signals (signals with specific internal nonzero vector EM waves, but which externally sum to zero vector resultant E and H fields) can be constructed for essentially any specific purpose. This includes "killing" a cancer or leukemia cell, destroying a virus, changing the DNA, etc.

This approach can directly reach and manipulate all immune and repair system functions.

The entire biochemistry and functioning of the cell — including its genetics — **is totally engineerable.** The Soviets have long known this, and have long since done it.

Further, a specific "charge pattern" of desired specific immunity (antibodies, etc.) can be designed and used to "charge up" the nuclei of the biosystem. This charge is then maintained by the system to provide permanent immunity. Thus one can develop, for example, an "electromagnetic innoculation" for AIDS, one for cancers and leukemias, etc.

Since the cellular control system is holographic, the "charge pattern" of immunity resides in every cell, including the blood cells.

Injecting a drop of blood from a scalarly immunized animal into another non-immune animal carries the scalar EM immunity pattern into the new animal. That charge diffuses throughout the overall biopotential of the organism, and the charge pattern activates the animal's immune system, including causing it to produce antibodies — according to the EM-transferred antibody template.

Antoine Priore demonstrated this effect numerous times. Thin was one of the great mysteries that confounded the orthodox members of the French Academy of Sciences.

The French Academy did not know of scalar electromagnetics, the cellular biopotential rooted in atomic nuclei of the cellular material, the cytopathogenic effect of mitogenetic radiation from diseased and dying cells, phase conjugation, and phase conjugated electromagnetic healing.

It is little wonder they did not comprehend the operational healing mechanism of the Bordeaux cancer-curing machine of Antoine Priore!

A New View of the Nature of Mind and Thought

The reason that Western science has not discovered what mind and thought are is simple: They have only built instruments and tools to look where the mind and thought *are not to be found.*

Let's look where they are to be found.

For example, consider the ionic discharges in and across the tremendous numbers of synapses of the human nervous system, and the slower discharge and migration of ions across cellular membranes, etc.

Considered as a single biological ensemble, these are much like a vast array of continual and continuing spark discharge vectors, slow charge current vectors, etc.

Overall, within the macroscopic space occupied by the body, these electromagnetic vectors sum almost entirely to a zero vector resultant. Only a tiny nonzero vector residue remains.

However, this vector zero summation has incredibly rich infolded signals, channels, and dynamic relationships (structure) within it.

Modern science/medicine measures the small nonzero electromagnetic residue (the remaining weak E and H fields) and tries to ascertain where and how mind and thought are accomplished, by studying that residue.

However, the residue E and H field remainders are simply the "garbage thrown out" of the glorious engines dynamically operating inside the vector zero summations (and multiplications).

The E and H field residues are the wastes or exhaust byproducts. They are not the functions of the mind/thought process itself; instead, they are the spillage or leakage from those functions.

Take a "frozen time" snapshot of the components of the zero vector summation. Regard the highly complex, infolded structure or pattern these components form.

Now take a second "frozen time" snapshot of the components of the zero vector summation, a very, very short time later. Subtract **the** previous "frozen pattern" from this second "frozen pattern."*

The difference or "delta" between the two patterns represents a myriad of thoughts, hence it represents the contents of the "thinking mind," in the most general sense.

The "mind" is the overall functioning and changing of the entire pattern's substructure, and the ability of an organism to have such **and** do such. The "ability of the organism to do such" requires at least

---

*Similar to the magnificient "double **exposure holography"** work of Dr Robert Powell.

two more nested levels of virtual state—two more hyperdimensions.

In simplest Kaluza-Klein theory, all those electromagnetic component vectors are in the fifth dimension. Hence they are hyperdimensional.

The zero-vector system may well have further infolded, internested levels of zero-vector systems. Systems within systems. These are hyperspatial, going into the 6th, 7th, 8th, etc. dimensions. Rife's microscope, for example, could reveal some of these levels; using evanescent waves, it could resolve some 16 ever-deeper levels of dynamic energy structure.

Everything driving the component vectors themselves is higher dimensional.

"A thought" is the exact change of a localized pattern (localized zero vector summation whose components are changing but remaining zero summed) from one moment to the next.

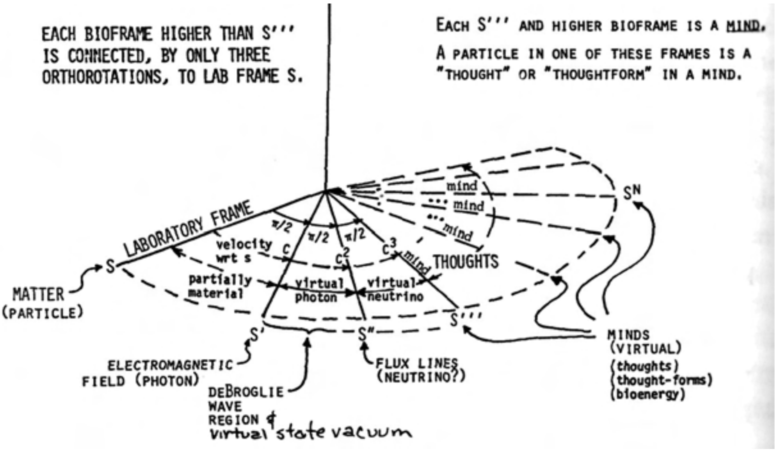
EACH BIOFRAME HIGHER THAN S''' IS CONNECTED, BY ONLY THREE ORTHOROTATIONS, TO LAB FRAME S.
EACH S''' AND HIGHER BIOFRAME IS A MIND.
A PARTICLE IN ONE OF THESE FRAMES IS A "THOUGHT" OR "THOUGHTFORM" IN A MIND.
mind
mind
mind
S^N
LABORATORY FRAME
π/2
π/2
π/2
velocity wrt s
C
C^2
C^3
mind
THOUGHTS
S
MATTER (PARTICLE)
partially material
virtual photon
virtual neutrino
S'
S''
S'''
ELECTROMAGNETIC FIELD (PHOTON)
DEBROGLIE WAVE REGION & virtual state vacuum
FLUX LINES (NEUTRINO?)
MINDS (VIRTUAL)
(thoughts)
(thought-forms)
(bioenergy)

Figure 76. Hyperframes, vacuum, virtual state, minds and thoughts.

Thought is always a pattern change, of the components inside an EM vector zero summation.

Most thoughts are totally "unconscious" (multiple simultaneous or "parallel" thoughts). A few are conscious (serially ordered, singularly considered and processed).

The unconscious mind is totally a parallel processor (many things/thoughts at once).

The conscious mind is totally a serial processor (only one thing/ thought at a time).

Most persons never take the time and trouble to reflect and notice that, consciously, they only can perceive one single thing at a time. Of course, the conscious mind is so rapid that by habit they assume they can perceive many things at once.

Since the conscious mind cannot discriminate the "multiple images in the slide projector at once" of the unconscious mind, that's what makes it unable to "be conscious" of the contents of the "unconscious."

In fact, the unconscious mind is totally conscious — it's just **multiply conscious "simultaneously."** (Which sheds some interesting light on why one can develop multiple personalities, for instance).

When the conscious mind "regards" the unconscious contents, it sees "something" which has "multiple meanings at once."

That's what we call *symbolic.* A symbol is something which can have many meanings at once.

That's why the unconscious content always "symbolizes" something when it's trying to communicate to the conscious mind. It's necessary to interpret the symbols to understand what the message being communicated is.

That's why dreams, for instance, are symbolic. And why it usually takes a trained psychiatrist or psychologist to properly interpret the symbolic representations manifested in mental disorder, and get at the underlying cause being symbolized.

With deliberately constructed scalar EM, any or all of the conscious/unconscious portions of the mind can be available for engineering and control/change. Eventually you will be able to put contents of the mind and memory on a video screen if you wish to.

It will also be possible to interact with the mind's contents electromagnetically. Mental illness will be treated directly, on an engineering basis.

Of course this opens up the frightening possibility of hostile misuse to the detriment of an individual. It will even be possible to change or erase his or her very personality itself. We pray, of course, that such a powerful tool will not be utilized this way, but will be used to heal humans, not hurt or kill them.

Unfortunately, Lisitsyn's work reveals that the Soviet Union has long since applied energetics (scalar electromagnetics) to mind control and mind engineering, including deciphering the genetic code operation. Controlled induction of images and sensations inside the brain — where they are processed as if self-originated — has been reported by the Soviets. Indeed, they have reported being able to control whether or not the induced material should rise to conscious awareness, as well as when it should rise to consciousness.

For years, U.S. intelligence analysts and U.S. scientists simply did not believe that electromagnetic signals could directly influence consciousness.

However, they were forced to reevaluate that position when a Soviet medical machine — the LIDA device — was openly obtained and evaluated.

The LIDA machine — a small unit somewhat larger than a briefcase — has been used in Soviet medical facilities to treat humans for decades.

The device uses a 40 megahertz EM carrier and very complex waveforms (signal modulations with complex mixtures of frequencies, phases, etc.)

When exposed to it, a person is caused to gradually fall into a trance-like, catatonic-like state within a few minutes. That person becomes very still and very quiet.

It was reported that the machine was tested upon a cat and had the same effect upon it.

One American scientist stated that the device had actually been used in North Korean "brainwashing" of U.S. prisoners of war in the early 1950's during the Korean War!

If so, one can see just how long the Soviets have achieved success with electromagnetic biological warfare/medical treatment techniques.

Reportedly, prisoners exposed to the machine were rendered unable to psychologically resist their harsh brainwashing interrogation.

But to return to the brain activity, with its innumerable, zero-summed minute EM vectors, and its extremely rich and complex deterministic substructure.

Note that the ensemble of sum-zeroed vectors forms potentials. These potentials — i.e., the scalar EM — penetrate to the atomic nuclei in the brain and body. Internal changes in the potentials' substructures (i.e., thoughts) also penetrate to the nuclei.

The nuclei continually "charge up" (the particles change their states) to these impressed potentials, including the "potential" or "charge" of each component of the impressed potential. That is, the nuclei charge up with both dynamic overall potential (mind) and thought (individual substructure changes).

Notice that **mind, thought, and memory reside in, and are recorded in, the atomic nuclei. At many virtual state levels. In many hyperdimensions.**

Thus the functioning mind and bio-control systems — including Popp's master cellular communication system, the immune control system, the acquisition control system, and the repair control system - all "reside" in the nuclei as dynamically interacting, patterned potentials or "patterned charge," complete with resonances and frequencies, etc.

We can regard the mind and its interactions as physical, in the extended sense we have briefly developed here.

Everything "mental" can be directly interfaced with and engineered.

One day, for example, "education" will be by direct loading of patterns into the brain/body/cellular biopotentials, much as we now place a diskette in a disk drive and load the computer memory. And then everyone can be educated, in the widest range of skills and to the highest degree ever dreamed of.

Kindling, Life, Mind, and Negentropy

The mass of the atomic nucleus can be charged up with a particular dynamically structured biopotential.

In scalar electromagnetics, I use the term **kindling** to refer to the effect of charging up mass by a specific structured charge pattern to form such a structured potential in the nuclei.

Note that we are actually and physically restructuring the nucleus itself. That is because, if the structure of the potential in the nucleus is changed, local spacetime of the nucleus is polarized in complex electromagnetic structures. The electromagnetic currents, actions, and charges of the affected nucleus adjust automatically to this structure or "grid." *

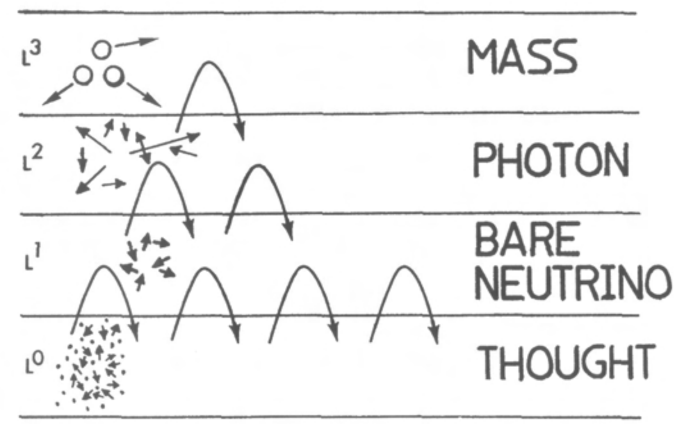

Figure 77. Internested levels of virtual state vacuum contain mind and thought. **The** vacuum also contains their interaction toward and upon mass.

*See Richard E. Prange and Peter Strance, "The semiconducting vacuum," Am. J. **Phy.** 52(1), Jan. 1984, p. 19-21. The vacuum may be regarded as a semiconductor. This semiconductor, particularly in the region close to the nucleus, may be manipulated by subjecting it to external fields, doping, etc. Extra energy density can even be extracted; see H. Paul and R. Fischer, "Comment on How can a particle absorb more than the light incident on it^{7}'" Am. J. **Phy. 51(4),** Apr. 1983, p. 327. For the involvement of negative time/negative energy, see C.W. Rietdijk, **"How** do 'virtual' photons and mesons transmit forces between charged particles and nucleons?" **Found. Phy** 7(5-6), June 1977. p. 351-374.

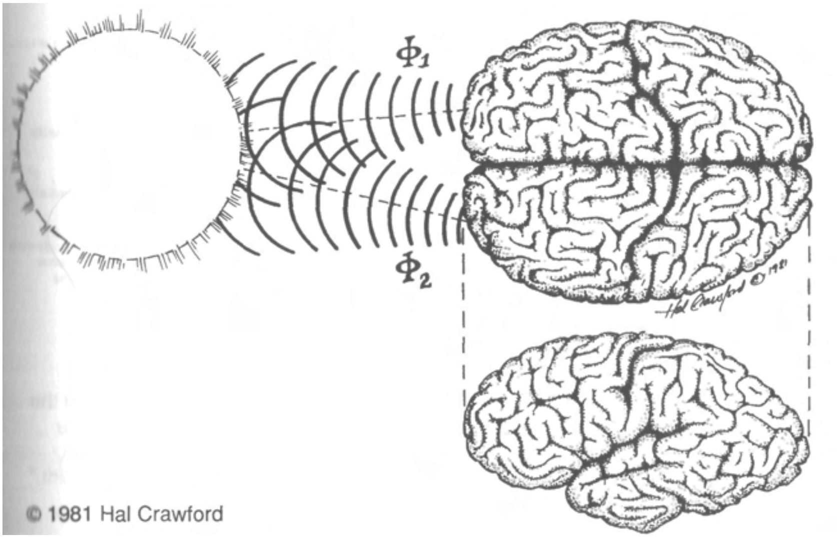
Φ1
Φ2
Hal Crawford © 1981
© 1981 Hal Crawford

Figure 78. The human cerebral cortex is a natural scalar interferometer. It is a virtual state tuner, processor, and transmitter-receiver. It also can produce...and control to some extent...phase conjugate energy and phase conjugate waves.

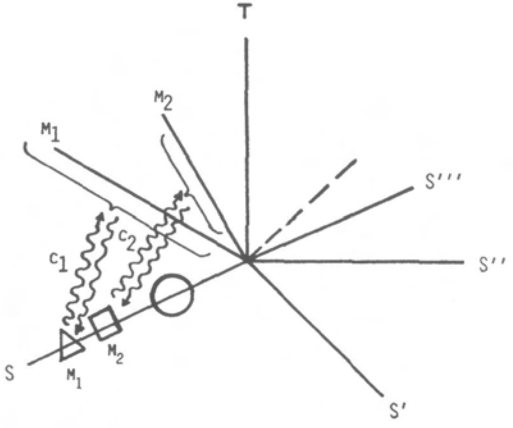
T
M2
M1
C2
C1
S'''
S''
S
M1
M2
S'

Figure 79. Consciousness and life. These phenomena refer to/constitute a deterministic hyperspatial (virtual state) coupling between mass and an ordered hyperworld.

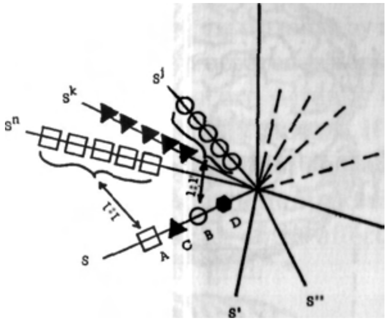
S^j
S^k
S^n
1:1
1:1
S
A
C
B
D
S'
S''

$A \circlearrowright S^n$ is a living system.
A is the physical body, the incepted mass.
S^n is the mind.

$B \circlearrowright S^j$ is also a living system.

To S, D is inert; i.e., D is not preferentially incepted.

To S, S^k is a "disembodied mind." I.e., S^k is "dead," since it is not preferentially incepting its former body, C.

Whenever a proper retuning body, say E, appears in S, S^k will reestablish a coherent crosstalk with E. This constitutes "reincarnation" of S^k.

Figure 80. Life and death. If the tuning or coupling between the mindworid and the tuned, structured mass body is broken, that is physical "death." Discharge of the structured cellular potentials then produces "death photons." Discharge of the overall, structured biopotential of the body constitutes discharge of the "living spirit."

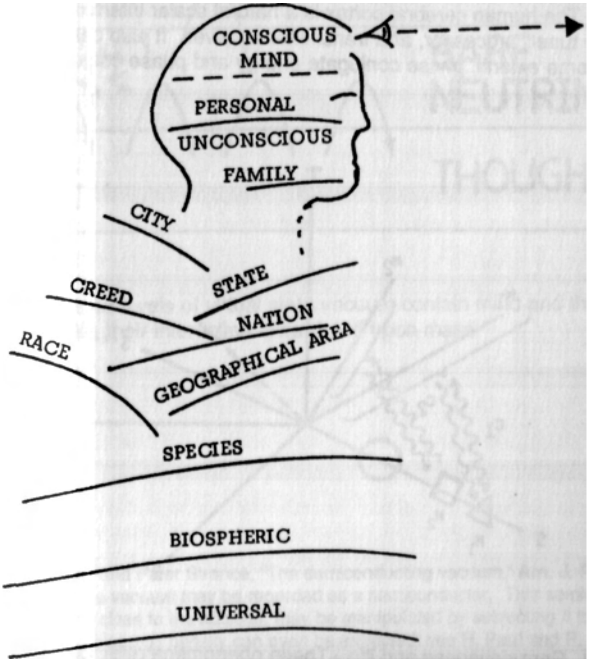
CONSCIOUS MIND
PERSONAL
UNCONSCIOUS
FAMILY
CITY
STATE
CREED
NATION
RACE
GEOGRAPHICAL AREA
SPECIES
BIOSPHERIC
UNIVERSAL

Figure 81. Layers of unconsciousness intersect hyperspatially.

Hal Crawford '79

© Hal Crawford 1979

Figure 82. Scalar electromagnetics can directly interact with the various levels of human mind and personality.

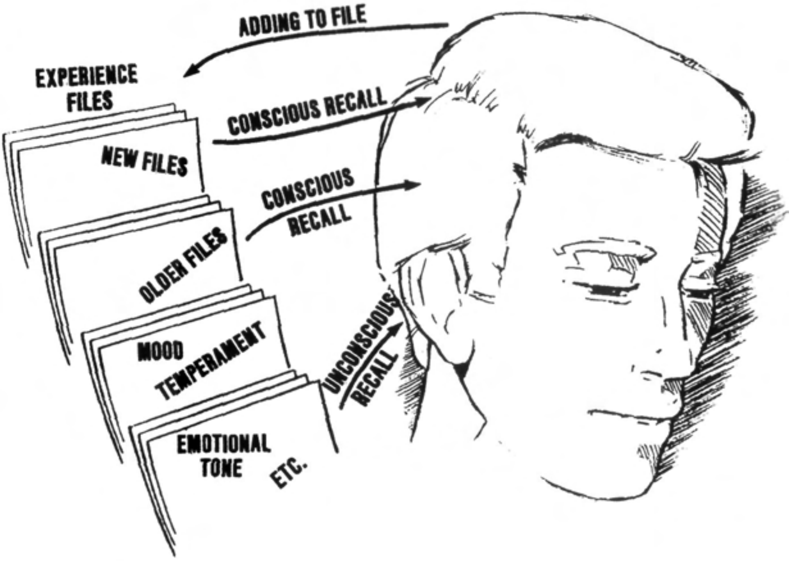
ADDING TO FILE
EXPERIENCE FILES
CONSCIOUS RECALL
NEW FILES
CONSCIOUS RECALL
OLDER FILES
MOOD TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Fgure 83. What personality is.

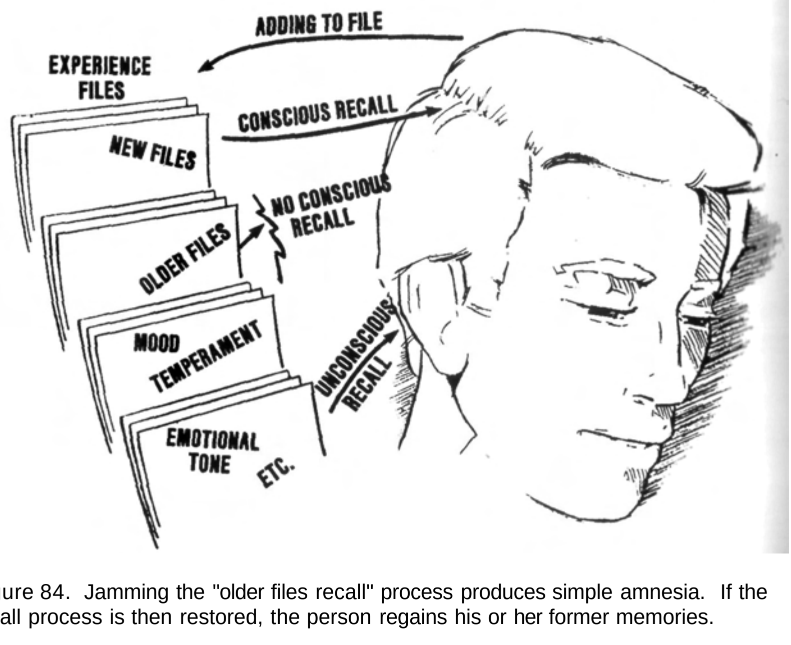
ADDING TO FILE
EXPERIENCE FILES
CONSCIOUS RECALL
NEW FILES
NO CONSCIOUS RECALL
OLDER FILES
MOOD TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Figure 84. Jamming the "older files recall" process produces simple amnesia. If the recall process is then restored, the person regains his or her former memories.

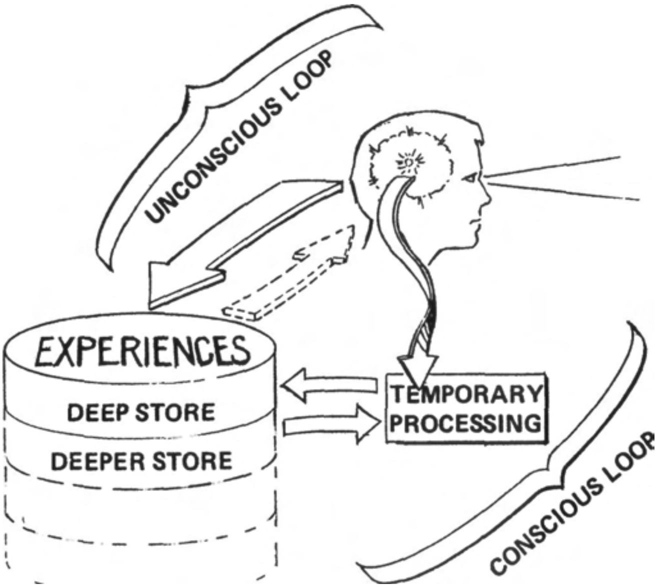
UNCONSCIOUS LOOP
EXPERIENCES
DEEP STORE
DEEPER STORE
TEMPORARY PROCESSING
CONSCIOUS LOOP

Figure 85. To sustain one's personality, continual functioning is required. All channels must be functioning.

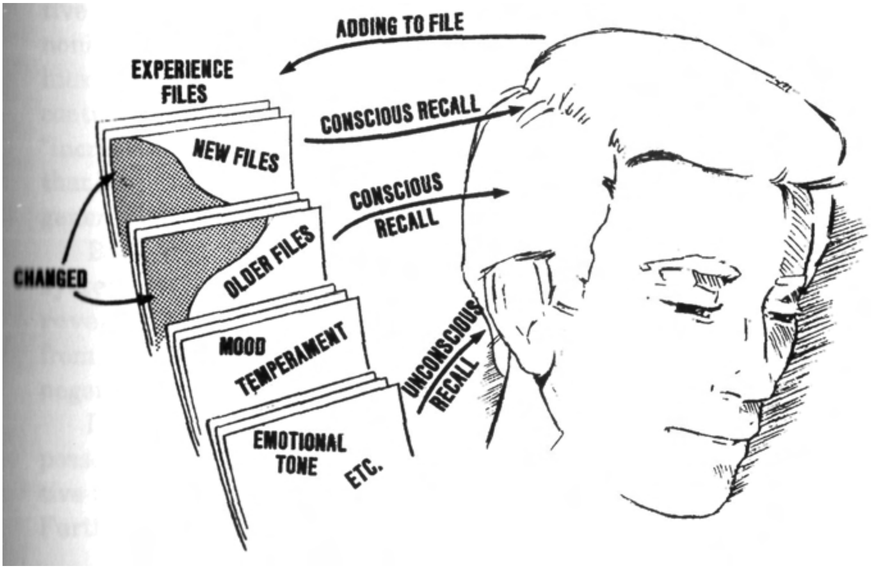
ADDING TO FILE
EXPERIENCE FILES
NEW FILES
CONSCIOUS RECALL
CONSCIOUS RECALL
OLDER FILES
CHANGED
MOOD
TEMPERAMENT
UNCONSCIOUS RECALL
EMOTIONAL TONE
ETC.

Figure 86. If the files themselves are changed, the personality itself is changed. It is possible to alter who or what a living person is.

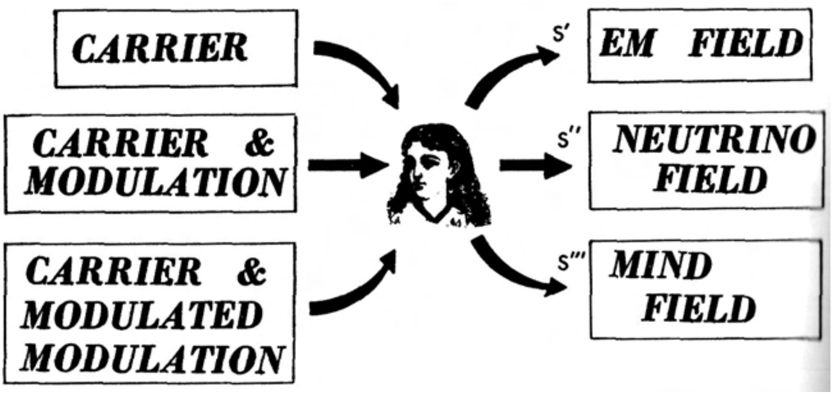
CARRIER
CARRIER & MODULATION
CARRIER & MODULATED MODULATION
S′
S″
S‴
EM FIELD
NEUTRINO FIELD
MIND FIELD

Figure 87. The mind and personality can be affected — and changed or controlled from a distance. The Soviets have developed at least a substantial part of this technology.

Table 44. LISITSYN'S REPORT

- **THEORY DEVELOPED & FITTED**
- **HYSTERESIS MEMORY LOOP**
- **CONTROLLED INDUCTION**
  - **IMAGES**
  - **SENSATIONS**
  - **PREDETERMINED EMERGENCE**
- **23 EEG BANDS**
  - **UP TO 8.1 X 10^{20} HZ**
- **11 INDEPENDENT CHANNELS**
- **BRAIN CODE BROKEN**
- **44 DIGITS OR LESS**

Thus by tailoring the specific scalar waveform and wave substructure with which we irradiate a mass, and utilizing phase conjugate (time reversed) waves, we can directly engineer the nucleus itself when sufficient charge structure has been built up in it to form a structured potential that is powerful enough to alter/ **shape** the nuclear processes.

We can, for example, transmute the nucleus into another form. Here an isomer form is the easiest. An isomer for an atomic nucleus (a specific isotope) has the same number of nucleons (protons and neutrons) overall as that isotope does, but a different number of protons. So it's a different element, but very specially related to its isomeric "brother."

All that it takes to change a neutron into a proton or vice versa, **is** to "flip" a single quark (subparticle of the nucleon).

Living systems can do that and transmute elements, to a small degree.

For example, living systems must use phase conjugation, negative time, and negative energy. To do so, they must function nonlinearly. That is the only way they can defeat the otherwise inexorable second law of thermodynamics: that all (linear) processes continue to more and more disorder as time passes. (That's called "increasing the entropy (disorder)." Unless a biosystem could defeat **that** law, its genetic pattern would inevitably be disordered in a few generations, and the species would disappear.

By using phase conjugation and time reversal, the **living system is** able to reverse the law of entropy in its time-**reversed channel.** That is, in that channel things go in reverse: fro***m*** disorder back to order! That's called "negative entropy" or negentropy" for short.*

It follows a priori that, to do this, the living system of necessity possesses phase conjugate energy systems (negative energy, negative lime systems) that are rooted in the atomic nuclei of its body. Further, it must deterministically manipulate these phase conju-

---

To* show how fully one can integrate physics and metaphysics, we point out that, since the vacuum imprints with every **change, nothing *is* ever lost! Death, where is thy sting?

gate systems.

This allows the precise definition of a living or nonliving system. If the system deliberately utilizes and manipulates phase conjugate energy/time in its atomic nuclei, it is a living system. If it does not deliberately use this, it is an inert or nonliving system.

For example, this sheds a great deal of light on the nature of a virus.

A virus is a single molecule. Ordinarily it clouds the distinction between life and death. Viruses can be crystalized out of solution, formed into "rock-like" crystals, and placed on the shelf, so to speak. There they can remain for many years or centuries, seemingly an inert rock. Then when placed back into their "preferred medium," so to speak, the virus crystals dissolve, separate, and "come back to life" and resume their living function. Yet the virus can be killed, in which case it is just a "hunk of rock," sitting on the shelf. In that case, it won't come back to life when placed back into its preferred medium.

In one case its "deterministic phase conjugate energy system" — its mind, to put it simply — remained in the nuclei of the atoms of the crystaline form. In the other case its "mind" was destroyed or disconnected from the crystalline form.*

A somewhat similar effect can be demonstrated in bacteria.

You can "kill" bacteria, for example, with UV radiation, then hold them in the dark for 24 hours (say, for 12 generations, since a nominal bacterial generation is about 2 hours.) They will remain absolutely "lifeless" and static, with no movement or cell division.

Then if you place them in the sunlight, you will be astounded to see them revive by the hundreds of thousands. In the "special kill" case here, the virtual state substructure of the UV photons just totally "jammed" the biopotential substructure functioning of the

---

*In the living case, the atomic nuclei remained patterned by the structured biopotential In the case of true death, the structured biopotential is lost and the nuclei revert to their "inert" form.

"minds" of the bacteria on one key channel. It didn't physically damage their bodies or "physical tuners." Then when placed in the sunlight, the broad frequency spectrum of the virtual substructure of the photons "dithered" and stimulated the bacteria's body-tuners, much like a "frequency massage" in all the bands. This set the physical tuners vibrating — and therefore responding once again to the biopotential substructure changes of the mind. Hence the bacteria "revived" and swam about — fat, dumb, and happy.

We have stated that **the mind is a physical thing, albeit a virtual or hyperspatial thing. In virtual/hyperspatial reality, it is a functioning, hyperspatial thing.** Interchanges be-

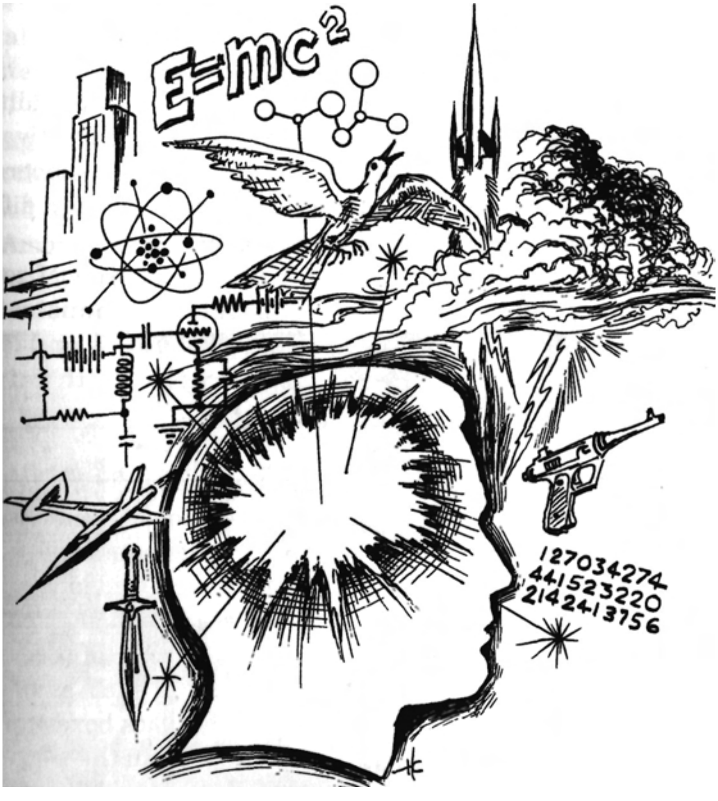

©Hal Crawford 1978

Figure 88. The mind and personality are physical things, albeit hyperspatial and virtual.

tween it and an otherwise inert, functional, 4-dimensional ordinary physical system (a "body") comprise life functions, "thought," cellular control, etc. Those interchanges in a living system occur in the biopotentials residing in the atomic nuclei of the physical body, in the higher levels of biopotential in the body structures such as cells, membranes, and organs, and in the overall biopotential of the entire mass of the body.

We accent that this is not mysticism, but physics.

But now let us return to phase conjugation and the Priore machine.

The Priore Machine and Phase Conjugation

In the 1960's and 1970's, in France Antoine Priore built and tested electromagnetic healing machines of startling effectiveness.

In hundreds and hundreds of rigorous tests with laboratory animals, Priore's machine cured a wide variety of the most difficult kinds of terminal, fatal diseases known today.

Funded by the French government in the amount of several million dollars, **Priore's machines concretely demonstrated a nearly 100% cure of all kinds of terminal cancers and leukemias, in thousands of rigorous laboratory tests with animals. These results were shown to medical scientists as early as 1960.**

Many of the experiments and tests were done by prestigious members of the French Academy of Sciences. Robert Courrier, head of the Biology Section and Secretaire Perpetuel, personally introduced Priore's astounding results to the French Academy.

The operation of the Priore machine was seemingly incomprehensible. Many orthodox French scientists — some of them world renowned — were outraged at the success of such a machine, shrilling that science had nothing to do with "black boxes."

They loudly called upon the inventor to explain the mechanism utilized by his machine, but the inventor either wouldn't or couldn't, explain the curative mechanism.

Priore certainly knew how to build the machine and make it work. It is debatable to this day whether anyone — Priore included — actually understood its principle of operation.

Neither the French Academy nor Antoine Priore knew anything of phase conjugation at the time.

In fact, the entire Western World knew nothing of phase conjugation in the 1960's when Priore was getting his finest results. At that time, only the Soviets knew of time-reversed waves.

Certainly Priore's machine was impressive.

Into a tube containing a plasma of mercury and neon gas, a pulsed 9.4 gigahertz wave modulated upon a carrier frequency of 17 megahertz was introduced. These waves were produced by radio emitters and magnetrons in the presence of a 1,000 gauss magnetic field. Experimental animals were exposed to this magnetic field during irradiation, and the mixture of waves (some 17 or so) coming from the plasma tube and modulating and riding the magnetic field passed through the animals' bodies.

Amongst other things, a plasma can convert a transverse wave to a longitudinal wave. Also, phase conjugate (time-reversed) waves

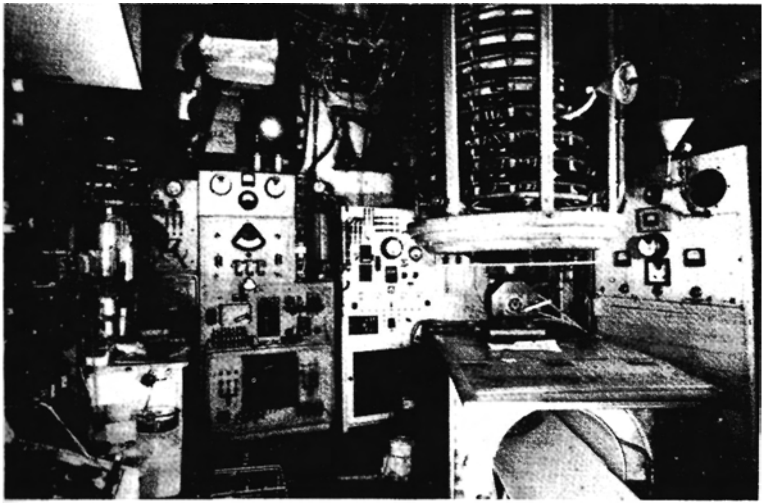

Figure 89. One of Priore's intermediate devices. This device cured terminal cancers and leukemias in thousands of laboratory animals. (Courtesy Bob Whitney)

can be produced by plasmas. Priore's apparatus produced a scalar EM wave/signal with deliberately constructed, infolded components including phase conjugate waves.

One may roughly view a cancer cell as a normal human cell gone "awry" and out of control of the body's master cellular control system. The cancerous cells, viewed as a sort of separate, parasitic group of cells, form a special kind of organism having its own master cellular control system "level," immersed in the host's biopotential.*

There is thus a specific, constant electromagnetic "delta" that differentiates the parasitic cancerous "organism" **from the normal human cellular organism.**

This "delta" can be considered a sort of constant, complex-structured charge existing in the body's atomic nuclei. It's exactly as if the body biopotential had been charged up by Kaznacheyev's "death photons" for that specific cancer condition.

If this cancerous "delta" (which may be rather like a complex intermodulation mix of waves) — or a frequency shifted "transform" of it to a different frequency band — is phase conjugated, a specific healing delta frequency pattern results.

If phase conjugate replicas of a cancer's cell's specific "delta" frequencies are fed into the body having that cancer, the deviation of the cancer cell's master cellular control system will be "time-reversed."

That will return the cancer cell to control of the animal's proper master cellular control system. The cancerous cell will be immediately destroyed, or reverted back to a normal cell of the animal.

A very similar process exists for just about every disease bacterium and infectious agent that attacks the body.

Phase Conjugates of "Death Photons" are "Healing Photons"

The Kaznacheyev experiments in the Soviet Union proved that, any cellular death and disease pattern can be induced by a specific electromagnetic pattern carried on an electromagnetic signal, if the

*One strongly points out the relevance of Dr. Robert Becker's epochal experiments proving that cells can be dedifferentiated and redifferentiated electrically. Those experiments are strong indicators that the cancer cell should be capable of being reverted to its more primitive state — that of the normal cell — by electromagnetic means.

target cells are bombarded with the pattern-carrying carrier signals for a length of time.What was not published of Kaznacheyev's work was the correspondent work showing electromagnetic reversal of cellular death and disease conditions by irradiating the diseased cell cultures by phase conjugate replicas of the pattern-carrying induction signal.

In simplified terms, if an action in forward time induces a condition, then the time-reversal of that action will reverse the condition.

The concept is almost laughably simple. **The time-reversal of an electromagnetic disease process is a specific healing process for that disease.**

In his device, Priore internally structured the carrier photons themselves — making them **vacuum engines.** He phase conjugated his vacuum engines, and then passed these time-reversed vacuum engines down and through a strong magnetic field which thoroughly penetrated all cells of the biological organism being treated.

The scalar components (structured photons) representing the time-reversal of the disease were absorbed and reradiated in all the cells, "charging up" the nuclei of the atoms in the organism to some potential level of the exact "healing and reversing pattern."

In the process, the cancer/leukemia pattern/charge also destroyed the cancer cells, or converted them back to normal cells as appropriate.

As a highly simplified analogy, the complex signal — viewed as a scalar Fourier expansion — represented a "stress" against any abnormal cellular control system encountered and returned it to the normal cell's master cellular control system of the body.

Encountering the normal control system in normal cells, it produced zero stress.

Encountering the abnormal control system in tumerous cells, it produced great stress on it, reverting it to the normal control system of the body's normal cells.

Thus to a normal cell the Priore signal pattern acted somewhat as a comb passing through one's hair.

No interference occurred with the normal cell (one whose scalar control pattern is phase-locked to the body's master cellular control system and in phase with it).

The Priore signal stimulated and "stroked" the normal cell, but did not hurt it at all.

On the other hand, the scalar control pattern in and of an abnormal (cancerous) cell is out of phase with the body's master cellular control system, and is not phase-locked to it.

Hence the Priore signal caused direct interference with the abnormal cell's independent scalar control system. The interference pattern constituted the reconstruction of normal energy directly in the cancerous cell, and also jammed its abnormal scalar life signal.

This destroyed the cancerous cell by two mechanisms: (1) physical energy was kindled directly in the abnormal cell, causing direct mechanical damage, and (2) the cancerous cell, being an "independent living critter," had its scalar life channel (connecting its primitive "mind" to its "body") jammed and stopped.

If the cancerous cell was reverted to a normal cell by the Priore stress before being destroyed, it became just a normal cell and the Priore stress had no further effect upon it.

The Effect is Universally Applicable

Any disease with cellular, biochemical, or genetic basis can be cured in like fashion.

Priore's method, for example, was clearly shown to be able to completely reverse clogging of the arteries with fatty deposits, and to be able to lower the cholesterol level to normal, even in the presence of an abnormally high cholesterol diet.

His method also showed complete mastery and cure of sleeping sickness and trypanosome-induced illnesses.

Often Priore found that every cell of the body — even the hair — must be irradiated and treated ("charged up") with the signal, for the disease pattern was in every cell. The master cellular control system is holographic — the pattern (substructured potential with its dynamic, oscillating components) is in each component (each atomic nucleus, hence in each cell.)

Every structural level of the body larger than the cell also has its own correlated pattern, or modulation, on the overall.

A biological organism can regenerate lost limbs, for example, if it can utilize its natural recovery process in a Priore manner.

Even though Priore's work was presented to the French Academy of Sciences by Robert Courrier, the prestigious head of the Biology Section, the Academy could not understand the device and its functioning.

That was because the Academicians knew nothing of scalar electromagnetics, and phase conjugation, and the Priore machine was a scalar electromagnetic device using phase conjugation.

In the Mid-70's It Ended

In 1974, a change oflocal government lost Priore his government supporters. His support and funding were lost.

Ironically, Priore was just completing a 4-stories-tall apparatus capable of radiating and treating entire human bodies. It would have been capable of curing cancer and leukemia in humans rather than just in laboratory rats. (His previous machines were much

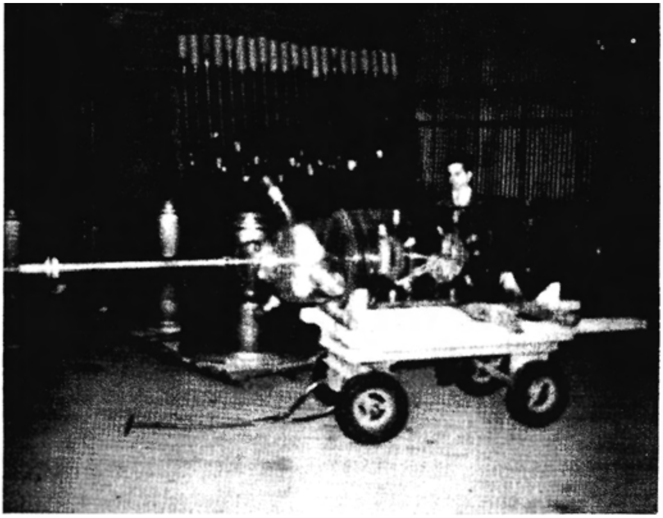

Figure 90. The large, special, phase conjugating plasma tube for Priore's giant device. The final machine would have treated humans "whole-body." (Courtesy Bob Whitney)

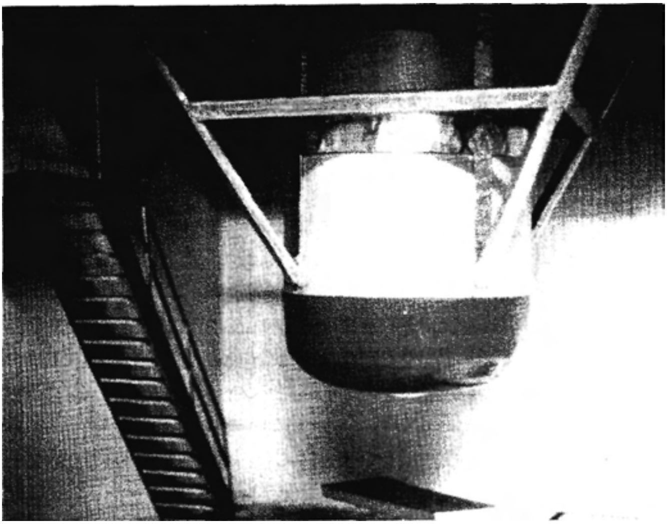

Courtesy Bob Whitney

Figure 91. Capping assembly for the top of Priore's giant plasma tube device. The last device would have treated human patients "whole-body." The cap assembly was on the third floor of the device.

smaller, and only a small animal could be irradiated whole-body.)

In that machine Priore used a "lamp," a special section, in which 17 specific frequencies were mixed and modulated upon the 9.4 gigahertz carrier.

The machine was large enough to irradiate humans over their whole body. **It should have been capable of curing cancer and leukemia in** two **five-minute irradiations, one week apart.**

While Priore was still alive, with my associates I spent nearly two years of my life in an effort to bring the Priore device to market and into the mainstream of medical research and development.

Robert Whitney, Frank Golden, and Tony Gideon played the major role, going to France and negotiating directly with the Priore group and the French government.

Verbal agreements with both the Priore group and with the French government were obtained. The machines were to be built, assembled, and tested in Bordeaux. Then they were to be dismantled and shipped to the purchasers — large medical research and development laboratories where they would be reassembled on site.

A group of wealthy backers was going to provide the necessary millions to fund the effort. Final contracts were being drawn.

Mysteriously, all funding for the project was suddenly withdrawn. Our backers were "leaned on" very hard and strongly threatened. Nothing we could do could revive the project.

Priore never again recovered his funding, and later died. The machine fell into disarray and was disassembled.

However, the proof that the requisite time-reversed signal can be produced, and will ride down a magnetic field, to penetrate every cell and every atomic nucleus in the body, is already in the present hard-core physics literature, if one knows where to look and how to properly interpret the work.

The fact that such signals can reverse nearly every major cellular illness condition of the body also has been proven by Priore and the scientists who worked with him, and it is in the French scientific literature.

The Priore-type cure for AIDS and other diseases only requires the necessary funding and personnel to be redeveloped and re-established.

Work of Australians Reed and Barsamian, and the American G. Wilbanks

At the University of Sydney, Department of Obstetrics and Gynaecology, some magnificent work of great importance to our thesis has been accomplished. Indeed, this Australian research may well be some of the most important work going on in the Western world today.

Dr. B. L. Reid, Dr. S. Barsamian, and their colleagues have produced experimental studies that positively verify pattern and form transmission at a distance, even through Faraday shielding. They have directly shown the reaction to, and results of, an unexplored information field on both living cells and inert crystalline matter.

In addition, they have replicated — and extended — the published results of Kaznacheyev on the cytopathogenic effect. In fact, they have obtained the effect at over 100 meters.

This shows, of course, that the effect is not just mesoscopic, but also macroscopic. Whether we like it or not, structural patterns of cellular disease can positively be transmitted electromagnetically, and kindled in target cell structures at a distance.

Dr. Reid has been interested in the origins of cancer since at least 1958. After much work in the standard models, by the late 1970's he was convinced that the present approach of medical science is wrong. He reasoned that we should pay stricter attention to the nature of growth before examining new growth, i.e., cancer. Growth, of course, is creation — and so the problem became, what is creation?

As a metaphor for biological growth to work with, he and his colleagues chose the growth of crystals. With keen foresight and intuition, they directed their inquiry so as to expose the complicity of some force — if any — external to solute and solvent.

Their work began to show that such a force existed, and it was a type that was unaffected by interposition of a Faraday cage. In other words, *it was not a conventional electrical force.**

Dr. Reid and his colleagues then included biological subjects in the form of living cells. They were familiar with Kaznacheyev's work in Alma Ata, Kazakstan on the subject of long range transmission of pathological (viral) effects on cells. In addition, they were familiar with the seminal studies of Gurwitsch on transmissions by onion roots cells, in the 1920's.

Most of their basic studies on the nature of the new force, however, were made with common salt, drying from aqueous solution, Some of their major results were as follows:

1. The familiar cubic symmetry is altered toward a higher energy form of dendrite crystals when biopolymers such as protein are included in the solvent.
2. A mass of lead 12-20 kg in the crystal space (up to 1 meter distant) causes an interference pattern in the salt crystal display.
3. A prospective dendrite pattern is not realized when a prospective cubic-pattern of the salt is dried in the local space (up to

---

"For one related possible scalar EM (electrogravitation) interaction, see Ya. B. Zel'dovich, "Electromagnetic current and charge due to interaction between a gravitational and a free electromagnetic field," **J. Exper. and Elec. Phy.** (Soviet translation), 16(7), Oct. 5, 1972, p. 302-303, 1 meter). The latter by its presence in space is able to prevent the dendrite expression and enhance the cubic symmetry expression. **A photograph of crystals of cubic symmetry when moved by a vibrating holder (2 cps) has the same effect on the real crystals drying from the solvent.**

4. Dendrite forms of salt crystal result when certain chemical reactions involving proton or ion movement occur up to 30 meters away, separated by brick walls, glass, etc. These reactions were detailed by Dr. Reid and his colleagues. If the mother solution is examined by ac capacitance on a sensitive bridge, variations in capacitance of over 300% accompany the receipt of the signal.

5. If such a transmission from 30 meters distance is regarded as a carrier wave which is modulated by a foreign — and thus detectable — chemical, **the chemical can be detected in the distant atmosphere some 3 months later in the form of a discrete or adiabatic 'cloud' which contaminates sodium chloride drying at the same time.** In this way, copper sulfate dissolved at a site A in June 1985 could be repeatably detected at site B, 30 meters distant and separated by walls, screens etc., by a drying solution of sodium chloride solution in September 1985. Serial dryings of the sodium chloride solution on glass slides showed that each day, the cloud hovered over a space of 1-2 cm at the end of a venturi tube for several hours before 3 p.m., and then moved off at 1 meter per hour in a NW direction. The copper content of the cloud was discernible by crystal structure and chemical means from the adjacent sodium chloride.

6. All of the above crystal forms required ambient oxygen for their manifestation. Atmospheres of nitrogen and argon prevented their display. At the time, the researchers suspected that the effects may have been carried by orbital electrons of oxygen showing unpaired spin which had interchanged with electrons of the copper solution. They then

developed an information-carrying device which could 'infect' the electron cloud of a target molecule sufficient for the chemical display of properties of the infecting electron's parent atom. This concept seemed to explain the experiment of the following paragraph.

7. Two beakers of sodium chloride solution were connected in series with a small voltage source (say a 1.5V dry cell) to provide a transfer of electrons through each beaker. The second beaker in series was sampled for subsequent drying and display of the crystal form. A crystal form of cubic symmetry resulted. When the solvent of the first beaker was altered by the addition of a protein solution (such as albumen or globulin), the crystal pattern of the second beaker was altered from cubic to dendrite pattern as though protein were present in the second solution.

8. During attempts to stabilize salt crystals on glass by means of thin plastic films drying from appropriate solvents (in order to stain the salt with aqueous solutions of dyestuff), it was discovered by chance that the crystalizing process is accompanied by structural patterns on the plastic film so exposed. These patterns were of two major types: (a) vortices, or spirals in the film plane, of 0.1 to 1.0 mm diameter, and (b) smaller curled structures into which the spiral could be seen to merge or grade. From a concurrent mathematical study of vorticeal interaction as designed by Barsamian, the similarity of the curl structures in plastic films (the researchers usually used 0.25% polystyrene in benzol) to computer simulations of vorticeal vectors was shown to be remarkable.

9. The researchers then dispensed with the underlying crystals and exposed the drying polymer film to a variety of spaces where the field, thought to be responsible for many of these tele-effects, was operative. With this technical advance, the team has made rapid progress over the last year or so. They have come to have a more comprehensive grasp of the ether

field than was possible by less direct methods involving formation of crystals.

10. On the biological side, they repeated the Russian studies done at Alma Ata with transmission of the cytopathogenic effect of viruses on animal cells. They used the cytopathogenic effect of colchicine and its relative vinblastine on cells with the following differences from the published Russian work: (a) The effect can be transmitted at least 100 meters between campus buildings, (b) it is not necessary to use a magnetic field for the propagation, (c) it is not necessary to use quartz vessels; glass is quite effective. These experiments came to the attention of a colleague, Dr. G. Wilbanks, Head, Department of Obstetrics and Gynaecology, Medical School, University of Chicago in 1983. Dr. Wilbanks was able to repeat them with similar results. However, in a more faithful copy of the Russian experiments, he used quartz vessels with cells placed in a magnetic field, with the vessels separated by no more than 1 mm. Both the Australian researchers and Dr. Wilbanks shared another thing in common: their uniform inability to interest the editors of several scientific journals such as **Science, Nature, Naturewissenchaften,** and others in publishing papers describing these fundamental experiments and results.

11. Experience with interference of a lead mass in the crystal work prompted the use of several metals in the ambient field of growing cells. Only lead (and to a lesser extent tin) showed the following properties: (a) Cell growth with 0.5-5 Kg lead in the ambient field is depressed sometimes to one third normal, (b) The cells die (dye exclusion test) but then can be revived by placing a small bar magnet in the local space with its poles in an E-W orientation. At least in Sydney, a N-S orientation of the bar magnet does not revive the stricken cells.

12. The effects of energy flowing through living matter have been documented by Kirlian's method. It was therefore of interest

to the Australian researchers to expose living systems to drying films of polystyrene. They found that the staining properties of the ambient-derived image on plastic films parallel those of the real organism.

13. By appropriate use of an imposed potential on living cells, it proved possible to match the field strength and show the actual imposition of the ambient (scalar) field on the biopolymer content of the material during the permeation process. The traces obtained are not unlike noise and would be so regarded by the casual investigator. The design of this apparatus and associated experiments was the work of Dr. Barsamian.

Concepts of Dr. Reid and Associates

Dr. Reid has kindly shared with me some of the concepts he and his colleagues have formulated. These concepts, it is stressed, are consonant with direct physical results they have obtained in the laboratory.

Note that any errors in this explanation are strictly the fault of this author, and not of Dr. Reid and his colleagues.

1. Matter is bathed in an energy sea familiar to quantum theorists. The Australian team prefers to call this sea of energy **ether,** as a relatively bland term, perhaps in a way to rescue it from the oblivion where Einstein banished it. The key to the silence of this sea is its "fractional-charge" charged particle status and low energy values. The former conceals it from physics, while the latter conceals it from chemistry and thermodynamics.

2. The differing energy levels of this sea, combined with its enduring motion, are theoretically and practically best expressed as a **vortex.** Dr. Reid and associates have detailed this theory to explain the creativity of biology. The incipient creativity of the vortex is given material expression by its association (as part of the EM field) with an electron which it

now endows with equivalent novel properties. There is no intuitive reason why the energetic rearrangements of the electron cloud cannot proceed to the nucleus, there to set up a quark or proton rearrangement — or, as the quantum theorist would say, a virtual particle exchange — leading to transmutation. Kervran's work on biological transmutations is relevant; the literature of such effects goes back to 1798 at least! The team suspects that some such mechanism may underlie their experimental result of the transition of sodium to copper.

3. **A further prime property of vorticeal interaction is its proclivity for amplification,** wherein the atomic-size vortex can be built up to the meteorological hurricane by such adherence or affinity. The team has noted the affinity of ambient field vortices to those produced by an electric current in a metallic conductor, with particular reference to the ill-defined boundary between the two. This parasitism was the reason behind the team's use of "carrier waves" from chemical reactions as transporters of fields whose identity of origin were sufficiently specific to trace the field subsequently. It is at least conceivable that the whole process could be amplified and used to parasitize (modulate) a radar beam as carrier. The affinity property of fields may be exhibited by the experience of mariners with what they loosely term "magnetic phenomena" thought to be derived from luminescent plankton. In the South China and Arabian Gulf seas, giant "wheels" of light, 5-15 km in diameter and containing spiraling spokes, revolve about the ship at night, reversing their direction of rotation as they pass the midpoint of the vessel. The key observation in the context of the field affinity concept is that the luminescence is manifoldly enhanced when the ship's radar is switched on. There could thus exist a chain, scalar-EM/plankton-optical, the first leg (scalar EM) of which can be amplified by radar EM.

4. The property of affinity means that **a scalar field's** transform **or copy, of the energy disposition imposed on it by the experience of matter last permeated, may persist as**

> **a** memory **in space,** as many of the Australian observations attest. The team found that recall was not all that difficult technically. **The energy transform for copper** sulfate **existing as a non-dissipated cloud with marked adiabatic properties could be recalled five months after its formation. There is no reason why informationally more elaborate spatial configurations of energy, including human thoughts, will not behave in precisely the same way.** The cell nucleus could conceivably be the candidate for the first site for storage of such energy. The nucleus has a highly ordered structure and an associated high permittivity, and this would make it a real target. In the human subject, there is experimental evidence that the mongoloid polymer constitution will have a higher activity for storage of the ambient field than will the Caucasian.

Other Important Ideas

Dr. Reid also shared other important insights:

Lavoisier in 1748 advanced the concept that matter itself (chemicals) is inert. The activity of all chemicals, according to Lavoisier, resided in their associated ether — and God sent the ether!

Taken to its limit, this suggests that the source of the energy which we call *thought* is a property of the scalar field. Matter (say, the nervous system) interacting with these fields, would be incapable of discerning the source of the field as presented, whether of self or external origin.

In his book **Hidden Variables in Quantum Theory,** Dr. David Bohm felt that in 'elevating' (his word) matter (as particles) from the sea, it might be possible to consider the idea as containing even more elaborate energy transforms such as biological shapes, these transforms themselves undergoing metamorphosis. [Comment by this author: Rife's microscope, using evanescent waves, actually allowed one to directly see and photograph such forms, proving the thesis.]

The ability of the Australian researchers to rescue these transforms on films together with evidence that they could be seeing the results of the exclusively space interaction of virtual forms (say microbe and antibiotic), is also a direct technical step toward validating Bohm's important idea.

Wider Implications of Structured Field Information

As T.D. Lee points out in his important book, **Particle Physics and Introduction to Field Theory,** particle physicists have not attempted to use the structure of the vacuum nor the structure of potentials.

Dr. Barsamian has shared other wider implications of the structured ether.

Over the ages a phenomenon which, as discussed, can have no representation in real (that is, electromagnetic or thermodynamic) terms can hardly have been thought of as possessing specific structure or function. Instead, it signaled its undoubted presence to generations of philosophers and physicists before Einstein, solely through statistical effects.

The writings of Helmholtz, James Clerk Maxwell, and Neils Bohr are examples. The more recent Dirac and Fermi statistics are **a** continuation of the process.

If an alternative definition of science is the continuing refinement of that which we partially knew or took for granted, then the refinement of these statistical approaches ~ with their obligatory uncertainty as defined by Heisenberg — is a worthy goal.

In this sense, a more articulate description of the anatomy of the field is a logical progression from the beginnings of the ether concept in Classical Greece through to the statistics of the nineteenth to the early twentieth century.

A hundred years ago, this orderly progression even reached the concept of a vorticeal atom at the hands of J. J. Thompson. However, **it** was interrupted very effectively by Einstein's publication of the special theory of relativity.

Einstein was fortunate in that he had chosen for his mathematics a pivotal feature of the ether field, its elasticity. The mathematics of elasticity had been evolving — through the developments of Hooke, Young, invariance, and the calculus — to handle its historically increasing complexity (by tensors). This mathematics of elasticity was then available to Einstein. The wisdom of this choice was vindicated by the remarkable special relativity predictions which were born out in subsequent observation and experiment.

There was seemingly no need to suspect, let alone probe, any special properties of the medium underlying the mathematics.*

The success of these predictions effectively blocked any chance of a return to the ether field concept. Indeed, it was seen as an arrogance to despoil the "elegance" of the mathematics. Nor was the success of quantum mechanics — the other leg of the Western World statistical approach to the nature of reality — able to get the physicists to heed Bohm's 1958 ideas on the hidden variables in its fabric or Dirac's 1952 ideas on a lesser uncertainty of the particle-field relationship than Heisenberg had formulated.

Recently Dr. Barsamian has originated the concept of an **alienation of the electron from its field.** The full expression of this could see the field so alienated that it is well-nigh autonomous, and some kind of fundamental entity in its own right.

In this sense, the freed autonomous field — at least conceptually —could proceed with its creativity by hybridization with other fields and so return the product of these asides to the particle. In atoms of the next fragment of matter with which the particle then associates, this product could create novel properties derived from the interchange.

The multiple properties of such a field allow it to be resolved in EM (electrostatics and electrodynamics), in atomic structure (electrons and nucleons), and in pressure (gravity) terms. This universality makes a new concept possible, long beloved of the GUT (Grand Unified Theory) theorists: that there is only one energy.

It is also quite possible that this ubiquity need not be confined to physics. It may penetrate the biological imponderables of growth, new growth and consciousness.

In view of the work of Dr. Reid and his associates, and other work — such as that of Hines and Chimonas showing atmospheric gravity waves launched by auroral currents — the prospect of amplification and transmission of transforms and patterns over long distances,

---

*As, for example, whether or not the concept of an abstract vector space in vector mathematics must be modified to correspond to what we know about the vacuum medium. If the vectors are to correspond to the movements of physical things in the *vacuum medium,* then the vector medium must correspond to the vacuum.

through affinity with long distance carriers, is indicated.

Dr. Reid stated that it would be surprising if other laboratories outside Russia had not progressed as far, if not further, than he and his associates had. Obviously the lack of similar publication indicates the nearly impenetrable editorial curtain confronting Western researchers in this and related areas.

Dr. Barsamian's Important Theoretical Insights

Dr. Barsamian has formulated a theory in relation to the development of cancer that is of great importance.

In his theory, he is developing the approach of classical mechanics to electromagnetic interactions between charged particles and between dipoles. The field associated with the primary moving particle is not neglected in his theory, and this part of it differs from Dirac's "new electromagnetic theory."

The field associated with the primary moving particle, in Dr. Barsamian's approach, is viewed as an important part of the particle's interactions. This field is stressed in the sense that diverse phenomena — from Maxwell's EM field radiating from the remotest regions of the universe to the controlled charge flow in the matter medium — are viewed as manifestations of EM forces.

This of course is unlike the traditional relativity view of EM force action on moving charges.

In Dr. Barsamian's view, the potential vectors (derived by a reatment from fluid mechanics theory) form an EM field and are positioned on a vortex curve. All parameters of such a field can be treated mathematically, and Dr. Barsamian has one or more papers presently in publication.

The duality of such potential vectors (gradients) is to conduct the motion of particles (which are always either charged or dipolar) appearing in nonmatter, which is taken as an incompressible continuum — the "void." Because the final results of these forces are always movement of matter, Dr. Barsamian prefers not to classify them by their origin.

Such a field may persist with great permanence. There are some limitations in summation (triplet conditions) and selectivity (periodicity) of the vortices themselves. Part of these limitations takes the form of particle characteristics. Presently he is rather sure that the formation of patterns of organic, inorganic and liquid crystals (typical to living biomatter) are partially responsive to the effect of the described field.

Dr. Barsamian is attempting to synthesize his own experimental findings on the dielectric behavior of biosystems as they enter the cancer process with the data that Dr. Reid is producing with respect to action at a distance.

With respect to living systems, he has slowly been approaching the conclusion that existing physical laws are just not applicable concerning (1) entropy, or (2) energy breakdown, or (3) random collisions of particles in quantum terms, or (4) electron-ion exchanges in bioelectrochemistry terms.

Under such circumstances, it is difficult to find a referee who sincerely understands the related experiments and their underlying theory. This presents a nearly impenetrable barrier to publication in leading scientific journals.

This is tragic, because these considerations and experimental results at least provide an entry into more fundamental experiments with present lethal diseases such as cancer and AIDS which annually take large numbers of human lives.

The present state of the art is telling everyone loud and clear that something fundamental is wrong with present theory, and that we should come to terms with more novel views.

Comment by the present author: In his prophetic insights, Dr. Barsamian has placed his finger squarely on the fundamental struggle of living systems against the "positive energy/time" laws of present physics. In its struggle, the living system must use negative energy/time, to correct from disorder to order and provide the extra free energy to do so. Otherwise, the steady progress toward increasing disorder in all its actions, would very shortly reduce the living system to such a disordered state that it could no longer reproduce the clear genetic pattern of its own kind. We have already pointed out the hitherto unexpected role that phase conjugation plays in allowing the living system to violate and break the iron dictates **of** conservation, nontransmutation, entropy, and raw statistics.

To live and survive, a living system must have access to other functional levels and other mechanisms beyond those contained in our present physics, electromagnetics, and biochemistry.

The incredibly productive and important work of Drs. Reid, Barsamian, and associates is of the utmost importance. Their work clearly moves everything in biochemistry and the knowledge of cellular functioning a major step forward.

It is a betrayal of the scientific method that such vital work goes so largely unreported in the leading journals. Indeed, the work of these Australian scientists should be galvanized and well-staffed and funded. They can literally usher in a new science, given the chance to further exploit and develop their results.

Royal Raymond Rife

In the 1930's and 1940's, Royal Raymond Rife revolutionized everything that has been done before or since in high resolution optical microscopy.

He also revolutionized everything before or since in cellular biology. He carried cellular structure far beyond anything ever dreamed of at the time or presently. He revealed the direct connection between organized living energy forms and organized biological systems. He revealed that life itself is organized and dynamic, to a far finer level than anything in the textbooks today. **He revealed that our present theory of disease is fundamentally very, very wrong.**

He produced direct, economical, electromagnetic cures of cancer, leukemia, and other such debilitating diseases. His work presages a future mankind could have had, where most debilitating diseases were quickly and economically corrected, and where no poisonous drugs, violent nuclear irradiation, and harsh chemotherapeutic "burning" of the patient would be necessary.

For such epochal work, he was ostracized, essentially imprisoned in a medical treatment facility, broken, condemned, and rejected by his peers. His findings, though printed in reputable publications and journals, were discredited and ridiculed.

He literally was reduced to a non-person by the power of the medical cartel.

Finally escaping from his enforced confinement, he lived out his remaining years and died quietly and unknown.

Rife's Universal Microscope

With his universal microscope, Rife achieved optical resolution of up to 31,000 diameters and magnification up to 60,000 diameters. His microscope could examine living viruses, living bacteria, and other as-yet-undiscovered living organisms and living energy forms that no other microscope before or since could see.

Even today, only the electron microscope furnishes such resolution, and it zaps-to-death the objects that are being examined. Further, it will not at all detect or see the "living energy forms" revealed by Rife's microscope.

To appreciate Rife's accomplishment, let's briefly summarize some of the performance factors of an optical microscope.*

Several factors are important in the functioning of a microscope: Resolution, magnification, and contrast lead the list.

By **resolution** we refer to the ability of the instrument to distinguish a small object. In other words, something like looking at a medical doctor's eye chart, and specifying the smallest print one can clearly see. Resolution is often referred to as resolving power.

By "magnification" or **magnifying power** we mean the ability of the instrument to "blow up" or enlarge the image produced. Roughly, it's the ratio of the apparent size of the object seen through the microscope to the apparent size of the same object seen without the instrument.

Magnifying power is exactly like enlarging a photograph. It won't improve the resolution or make the photograph any clearer, but may make whatever was captured (resolved) easier for the human operator to see. If you just keep enlarging the photograph, it will get more and more grainy, until no further details can be seen with higher magnification.

Contrast refers to the distinctness by which the various parts of the object can be distinguished from one another. To enhance contrast of objects under the microscope, staining of the objects is often used. Very often the act of staining will itself kill the living organisms — such as bacteria — that are being examined.

*The electron microscope reveals only the "dead carnage" of the battlefield after everything is destroyed. With it, one is observing physical residuals, not the living players. For some encouraging modern work, however, see Michael Hercher, "Virometer — an instrument for the measurement of the size of viruses using an optical microscope," **SPIE Seminar Proc.**, 1977, p. 17-22.

Figure 92. Royal Raymond Rife in his laboratory. Courtesy Christopher Bird

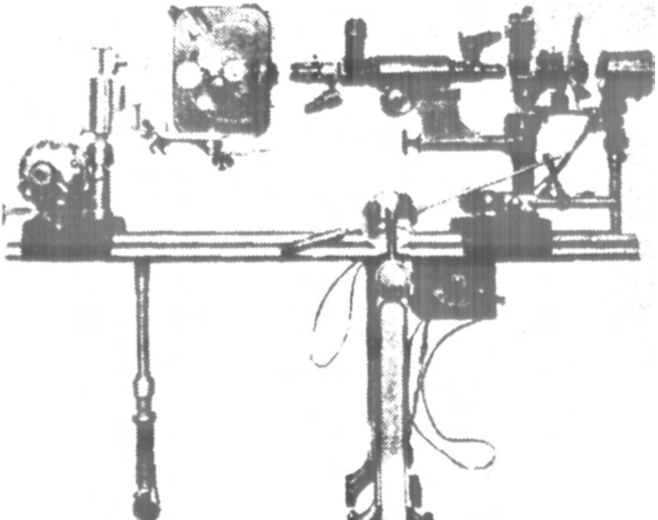

Figure 93. Rife's first virus **microscope.** Courtesy Christopher Bird

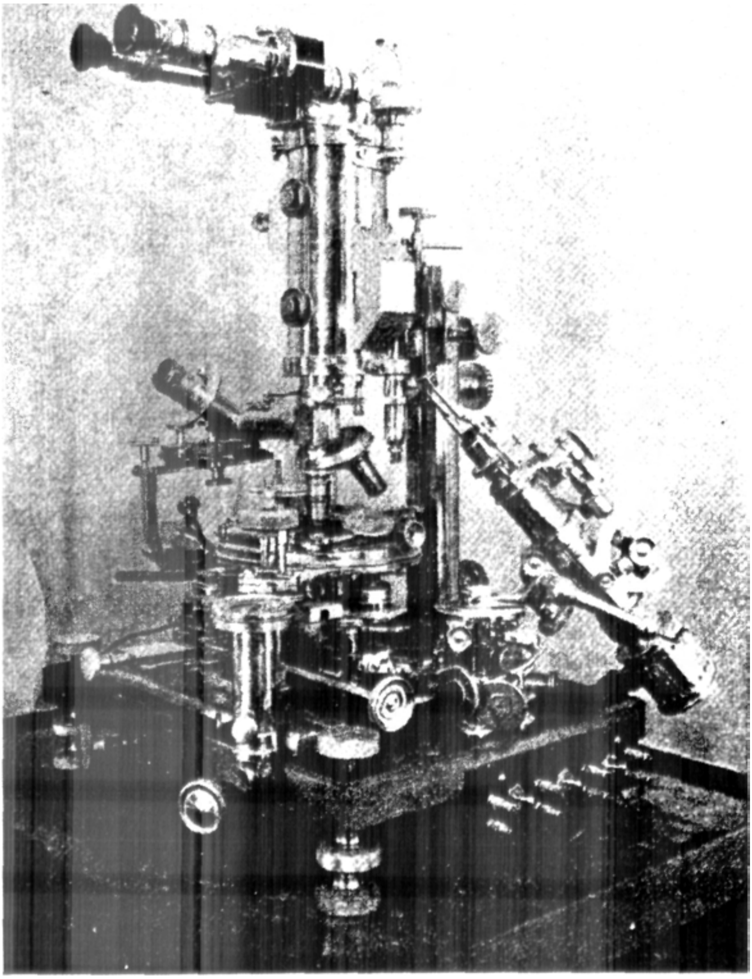

Figure 94. Side view of Royal R. Rife's prismatic universal microscope. Largest and most powerful of five Rife microscopes With it, living filterable viruses could be observed. Courtesy Christopher Bird.

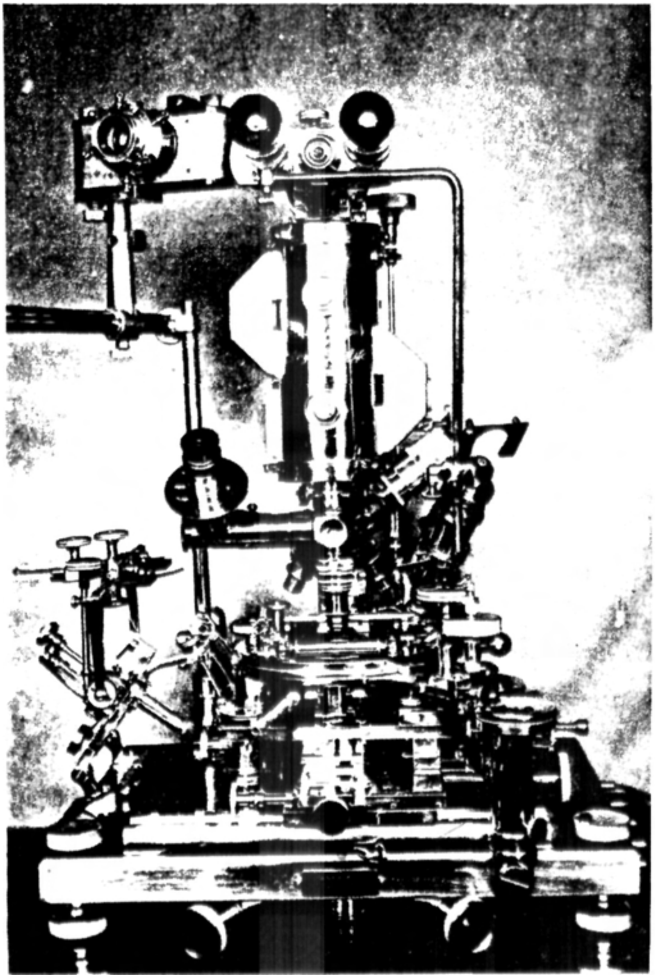

Figure 95. Front view of Royal R. Rife's prismatic universal microscope. Built in 1933 in Rife's San Diego, California laboratory. Courtesy Chris Bird.

We may state the resolution of an optical microscope in terms of the diameter of the smallest object resolved: for example, so many nanometers. Or we may just simply and loosely speak of so many thousands of diameters.

The resolution of an optical microscope depends upon the illumination conditions, its optical system, and the fact that the object viewed diffracts and spreads the light. A wide variety of mechanisms for illuminating, resolving, contrasting, etc. are used in optical microscopes.

According to standard Rayleigh theory, normal optical microscopes are limited in resolution to about half a wavelength. That corresponds to about 250 nanometers for visible light illumination. One can do a little better than that by using ultraviolet light and quartz optics. The UV has a higher frequency than visible light and hence a shorter wavelength. However, beyond 240 nanometers the resolution of an optical microscope quickly disappears.

A few researchers — notably in the 30's and 40's — remarked about other factors, such as "quality of the lenses," affecting resolution. Some of them reported results to about a tenth of a wavelength. Loosely, that translated to somewhere in the order of 5,000 to 6,000 diameters. The dependable upper limit seemed to be — and still is — about 3,000 diameters, however.

Normal optical microscopes simply are almost useless for trying to look at viruses. They will not resolve any but the very largest viruses, and will not resolve the internal features of even those few giants.

Much of today's knowledge of virus structures and shapes comes from the use of electron microscopes. They bombard the object viewed with a fierce rain of energetic electrons. These instruments see nothing of the *functioning* of the virus, for they kill it instantly in trying to view it.

With Rife's universal microscope, the dynamic living functions of the virus could be observed without killing it.

With Rife's microscope, a whole range of complex organisms and structures below the size of bacteria was revealed. Many of these organisms are still not known to present science, even though some of them were written about at the time. Filterable forms of bacteria

— forms which readily passed through filters supposed to easily block their passage — were discovered and reported by Rife and his medical colleagues.

In addition, Rife's microscope — which contained block-quartz prisms and lenses and interference stages — revealed halos around living organisms that other microscopes could not see, even though the object was within their resolving power. Further, it revealed the existence of entire organisms and forms which other microscopes could not see — even though, again, the size of the organisms was within their resolving power.

In other words, Rife's microscope not only revealed smaller physical forms than any other microscope could see, but it also revealed sizeable "living energy" forms which no other microscope could see.

Rife Proved That Everything Is Alive

The degree of smallness to which Rife's microscope would resolve, and the extraordinary energy forms it could detect, showed that direct detection of the virtual state organization of the living organism was accomplished by the instrument.

The science of the day was only just groping its way toward any sort of physics that could explain such an astounding instrumental result. Today, however, in the hard-core literature there is demonstrated proof that the optical limit of resolution can be drastically overcome, using evanescent waves.*

Shortly we will briefly give an "ad hoc" explanation of such evanescent waves.

Electromagnetism has been shown by Kaluza-Klein theory to actually exist in the fifth dimension. In other words, EM itself is hyperdimensional. It flows in the fifth dimension, which is "wrapped around" each point in our ordinary space. It is — to the first approximation — the external environment of every normal point in

*E.g., see T. Sato et al, "Application of evanescent waves to microscopic observation," **Bull. Tokyo Inst. Technol.** (Japan), No. 125, 1974. p. 35-41. See also G.A. Massey, "Microscopy and pattern generation with scanned evanescent waves," **Appl. Opt.** (Poland), l3(3), 1983, p. 247-255.

our 3-space.

Let's say that again.

The electromagnetics "medium" itself is totally external to each point in our space. Each of our points is surrounded in higher space by the electromagnetic medium.

We live inside a totally electromagnetic medium. It's not just an "electromagnetic environment" in our own space; instead, it's an electromagnetic environment in hyperspace.

Everything already is just the internal structure of the electromagnetic medium!

That's what the "vacuum" is.
That's what "spacetime" is.
That's what the "virtual state" is.
That's what physical matter is.

And it's all alive. Totally and completely alive. Everything is alive. There is nothing but life. The electromagnetic medium is alive.

Except within a locally-organized biopotential area, it's just "equally alive in all directions." So it appears inert, where by "inert" *we mean not singly preferential.*

Every biopotential change, at any level in a biopotential, extends "decaying-expotential-wise" to infinity, by standard theory. Part of that potential change exists at every point in the entire spatial universe. And there at that point, so do the biopotential changes for every other biosystem in the universe.

Everything's alive. There is nothing but life. Anywhere. Anytime.

An interchange between this "living, surrounding electromagnetic medium" and each particle of mass in our 3-space continually occurs. This exchange involves the so-called virtual particle flux.

If we model higher dimensions, such as are necessary to include the "particle zoo" discovered by modern particle physics, each of these "higher dimensions" corresponds to a "successively deeper-"

internested level of virtual particle flux—of the vacuum, or of this electromagnetic medium that surrounds us.

The living system orders, structures, and dynamically functions and interchanges electromagnetically throughout many higher "dimensions" — throughout many such internested levels of virtual-state/vacuum/spacetime.

Except for the first layer of virtual state (the 5th dimension in Kaluza-Klein theory), gross (classical) electromagnetic theory totally ignores the infolded, ordered deeper structure of electromagnetism (of the EM medium).

Using such gross electromagnetics, orthodox science thus can only detect and grasp the gross results of the living organism's functioning. Further, it can only detect and grasp those gross results that actually move 3-space matter — observable particles.

So it only detects the grossest interchanges between "mind" or "life" and matter. Specifically, it only detects the final results of that interchange—the gross movement of material particles themselves.

The first hypernumber — "i," the square root of minus 1 — is used to model another dimension at right angles to our normal three. An electromagnetic wave thus is modeled to have two parts: the first is the part that affects or moves a material particle, and that's called the "real" part or "observable" part. The second part of the EM wave is the component that lies in this "imaginary" (unfortunate standard term!) dimension. That imaginary part does not of itself move charged particles, so it is considered to be "something other than real."

Note that what's really done in this sort of orthodox modeling is to define observation as the movement of observable charged particles.

The human "conscious mind" is a functioning part of the overall human mind that has been specifically fitted to function almost totally with our gross bodily detection of the photon interaction.

We thus consciously detect and "are aware of only the first level of reality: the interface between the first layer of virtual-state/vacuum/spacetime/ and the 3-space of observable particles of mass.

Since the physicist must be "conscious" of his observation or instrumental detection, the detection process ultimately has to have a final stage that produces a photoelectric interaction with his body — for that is all the conscious mind detects and processes.

Everything chemical, electrical, etc. — according to quantum mechanics — will involve at its root level only the photoelectric effect.

Rife's Microscope Used Evanescent Waves

However, it is certainly possible to build an instrument which is multi-staged, and one in which higher stages interact primarily with deeper internested levels of vacuum virtual state. That is, an instrument that interacts with higher dimensional phenomena. That is, one that interacts with deeper internested levels of electromagnetics. If these stages interact causally in a vertical manner, then the final interaction with the human body and nervous system can still be "normal electromagnetics" and yet indicate a higher dimensional phenomenon, event, or function.

That is precisely what Rife did: His universal microscope penetrated to a much finer level of reality because its multiple stages used **evanescent waves.**

Here's what we mean by that.

An electromagnetic wave consists of a "real" part and an "imaginary" part, as we have discussed. It's possible, however, to have the real part become zero, and still have the "imaginary" part remain and dynamically vary.

This sort of wave — containing only the "imaginary" part — is said to be one type of **complex wave,** or an **evanescent wave.**

In advanced EM waveguide and optical theory, such evanescent waves can function to guide or determine the real parts of the EM waves. That "real part" is the part that is then going to interact with electrons and move them, and give us a "detection."

In other words, the "real" part of the EM wave will be moved and guided by the "higher dimensional" or evanescent part, or even by separated evenescent parts — pure evanescent waves.

So one can build a device or instrument that utilizes such an effect to reveal what's going on in a whole dimension beyond what we normally see. Waveguides in certain RF radars already do that.

Such instruments are not limited to the resolution determined by the "real" part of the EM wave (such as the "real part" of light).

Such a microscope would not necessarily be limited in resolution by the wavelength of the light it used to illustrate the observed object.

Indeed, if one uses a higher dimensional Kaluza-Klein model (which particle physics of necessity has to consider), one could build a "repetitive stages" instrument which could view many more higher dimensions. That's exactly the same thing as viewing many more deeper internested levels of virtual state.

That is precisely what Royal R. Rife did: **He developed a microscope using multiple stages, special interferences, etc. He built a multistaged evanescent wave instrument that could see into higher dimensions and deeper levels of virtual state.**

Rife's microscopes were thus startlingly different from normal microscopes. That is why he could see phenomena and organisms, ostensibly within the size capability of normal microscopic resolution, which ordinary microscopes could not see.*

Evanescent waves penetrate the virtual state (hyperdimensions) and interact with what to us is "dimensionless" (what does not spatially intersect our 3-space). Interfering multiple evanescent waves reflected from/having interacted with these (to us) nonphysical forms can again reproduce ordinary electromagnetic "light." Thus the evanescent wave microscope, if properly built, allows one to directly "see" (and photograph) what is to us living forms of energy, nonphysical, and without 3-space matter bodies.**

The same scheme can be used to develop instruments capable of directly revealing the living nonphysical world around us — such as the human biopotentials, their internal structures (including

---

*However, Rife's microscope was extraordinarily difficult to focus. Rife often spent 24 hours straight at his universal microscope, just in focusing it.

**If one is going to accept the many-dimensional theories necessary to explain particle physics, one must accept the possibility of higher dimensional living things, which we would see as living energy forms or as thought forms. These would appear nonphysical to us - as is time itself— but would be real and interactive nonetheless.

"thought forms"), etc. What the Soviets call **bioplasma** can be directly observed and photographed in this fashion. We have only to develop the necessary instruments; the science required is difficult, but it is already in the literature.

That is what Rife's microscope did. It was a forerunner of the instruments we need to develop in electromagnetic healing, and Rife was a truly great and unappreciated pioneer.

In one swoop medical science could have jumped a century ahead. The ruthless suppression of Rife and his fantastic scientific breakthrough was one of the most dastardly deeds ever perpetuated by the orthodox scientific establishment.

Rife Revealed a Far More Fundamental, Living Biology

With Rife's powerful universal microscope, it was also possible to view the interiors of the so-called "pinpoint" cells situated between normal tissue cells and just barely visible to ordinary microscopes.

Here is the astonishing living world inside those "pinpoint" cells, as revealed under Rife's powerful instrument: When one of the "pinpoint" cells was magnified, still smaller cells were revealed within its structure. When one of these still-smaller cells, in turn, was magnified, it too was seen to be composed of even smaller cells.

With Rife's microscope, this process could be repeated 16 times. An astonishing internested series of organized levels of a living cell was revealed, far more fundamental than anything that exists in present biological theory.

The present author points out most strongly that these levels correspond electrically to, and are in virtual particle flux pattern exchange with, the electromagnetic potentials of their environment. This includes their own biopotentials that are centered in the atomic nuclei of the atoms comprising the physical material of the cells, and that charge up with specific internal patterns.

The biopotential itself is organized into a corresponding virtual-state series of internested levels and functions. **The structured biopotential of the cell is a living, organized, functioning thing, and its internal functioning literally constitutes the "spirit" or "true nonmaterial deep mind" of the organism.**

All the internested levels are in constant electromagnetic exchange "up and down" with each other, particularly with respect to organized virtual particle flux patterns.

In addition, all the cells are in constant electromagnetic exchange "across" with each other, at all levels. This provides a dynamic, structured, living biopotential for the entire bio-organism (the entire body). Within this potential, dynamic interchange on all levels is continually occurring. This is the basis for the master cellular communication system that Dr. Fritz Albert Popp discovered.

Rife's powerful microscope had revealed nonmaterial functioning life forms (structured, dynamic, living biopotentials) connected to material bodies.

He could follow "filterable" forms of bacteria — actual living biopotential forms of the organisms that could not be separated out by filters, but which would easily pass through any filter.* He could observe interactions of these forms, changes of forms, translations and transmissions of forms, etc. — none of which is detectable by present biological theory or medical science.

Rife had advanced biology and biophysics a century in one jump. As always, orthodox scientists — most of whom in their scientific paradigm are self-admitted materialists — were quite unready to tolerate such heresy.

Obviously the materialistic dogma of the science of his day — and of the science of today — reacted most hostilely to such hogwash. Contrary to the prevailing mystique, most scientists are dogmatically attached to materialism and to the dogma of their present paradigm. Faced with a conflict provided by experiment, most will uphold the dogma and reject the experiment — the exact opposite of the scientific method they espouse.

Rife's revolutionary work was no exception. He incurred the unending, total opposition of powerful individuals controlling the direction of biology and medical science for their own personal gains.

Rife was hounded into court on trumped up charges. Though he was acquitted, he emerged a shaken, broken man and an alcoholic.

---

*More precisely, they would easily flow *around* the 3-dimensional filter, since they were hyperdimensional.

His work was suppressed. His equipment was left to gather dust. He was also forcibly committed to a medical treatment facility.

Finally escaping from his "prison," Rife lived out his remaining years quietly. He died without ever being vindicated for his marvelous, world-shaking discoveries.

But at least he left us a legacy. Persons are still alive who knew Rife and his work, to one degree or another. Some of his microscopes are still in existence, though non-operable due to missing or stolen parts. With proper funding, an enlightened team of scientists and researchers can be assembled to quickly repair Rife's remaining instruments and duplicate his incredible work.

With concentrated effort, Rife's work will yet play a primary role in the development of a direct electromagnetic cure for AIDS. As a byproduct, it will play a role in rapid development of cures for cancer and other debilitating diseases.

Royal R. Rife's contributions may yet help save half of humanity, and prevent a Soviet takeover of the world.

CHAPTER 6

DEVELOPING THE ELECTROMAGNETIC CURE FOR AIDS

By now, how one goes about developing a electromagnetic cure for AIDS — and for cancer and other killer diseases — should be apparent.

One must set up a modern research laboratory and assemble as many of the "dirty dozen" together as is possible. One then tackles the problem head-on, adding the necessary support staff and special consultants.

First the Kaznacheyev effect for AIDS must be isolated and determined. Then it must be reversed, to yield the precise "curative" signal.

Transforms for these Kaznacheyev IR/UV "death photons" and their phase conjugated "healing photons" must be obtained for a lower electromagnetic frequency band — such as, say, 10 to 20 gigahertz.

Away must be found to irradiate the whole human body with the curative signal. The ideal way is to utilize a scalar EM curative transform so that the atomic nuclei of the body — and hence its entire master cellular control system and immune control system—will be "charged up" with the correctly structured "AIDS cancellation message."

The entire procedure and apparatus must be as simple and small as possible. My colleagues have already achieved very promising — even remarkable — progress in this respect.

Reversing the Kaznacheyev Effect

The "Kaznacheyev effect" for AIDS virus condition must be stimulated between cell cultures, so that the disease condition is electromagnetically transferred from one to another. The actual electromagnetic "delta" constituting the contribution of the AIDS infection must then be isolated electronically. The best way to do this is probably to subtract the normal cell radiation pattern from the "cell plus AIDS" radiation pattern.*

With the "AIDS delta" determined, the delta is then fed into the appropriate phase conjugating mirror system, so that its time reversed replica is produced.

The new phase conjugated signal is then the required *AIDS reversal signal* to reverse the effects of the AIDS virus itself, inside the cell where it resides dormant. This " time-reversed" signal will reverse the genetic change in the cell, not just kill the HIV virus.

In a crude way, one is making an **electromagnetic anti-virus.**

For ease of development and treatment, microwave technology is most attractive. Obtaining transforms of the signals in the radar band is ideal, since a wide variety of techniques, instruments, and electronic parts are available for that region. Millimeter waves would be most attractive, for the equipment could then be highly miniaturized.

One can regard it another way also: energy forms (critters) are involved. The action of the AIDS virus in its host cell, upon the DNA of the host, is underlaid by manipulation of energy critters. If one makes the electromagnetic anti-virus form, one is also manipulating the energy critters in their virtual state substrata.

The net result is that essentially the virus pattern — even the virus itself — can be phase conjugated by the energy critters. The result can be to turn the actual virus in the cell into a negative virus, accomplishing recombinant DNA procedures in reverse.

Remember that the signal we seek to use involves negative energy and negative time. We are also engineering the virtual state directly. The ordinary positive energy/positive time/observable state rules and limitations do not necessarily apply. And the "fixed form" first order physical reality as we normally conceive it need not

*Procedures along the lines of the extraordinary double-exposure holographic work of Dr. Robert Powell will probably be necessary. His work on biologically significant spatial frequency spectroscopy has blazed the trail as to how to obtain the specific delta patterns desired. It is hoped that Powell will shortly publish the remarkable results of his 15 years' work.

be so fixed at all.

Physical reality itself can be directly engineered.

The engineering we seek to accomplish is directly upon the probability states propagated by the Schroedinger equation, before observation and collapse of the wave function occurs. **We seek to engineer physical reality before it is born, while it is yet forming.**

Only at such a level can the previous action of the AIDS virus — that in which it combined its genetic material with the genetic material of the host cell — be reversed and undone.

Only at such a level can we convert the infected human body from an "AIDS virus factory" back to an uninfected normal human body without AIDS.

The Proof: Priore's Work

Antoine Priore's pioneering work largely proves that it can be done. Cancers, leukemias, and many other virulent diseases yielded to his phase conjugated signals passed down through a powerful magnetic field to totally penetrate every cell in the treated patient's body.

And Royal R. Rife's work proved that a virus and a bacterium are not at all the "rigidly fixed" physical forms that our normal science has led us to believe they are. Instead, both the organisms and their biochemical and genetic actions can be addressed — and changed — on a much finer level of reality.

Of course, it would be enormously helpful if one had a working Rife microscope.

My colleagues are attempting to rebuild one of the original Rife microscopes, which has parts that were missing from it. They have every hope of having the microscope in action in the future.

Another angle of attack is also possible.

One of my colleagues has discovered a very peculiar, weak electromagnetic signal that will kill viruses, harmful bacteria, toxic protozoa, etc. but not harm living human cells. Only a few volts and a few milliamps are used.

However, there exists a major problem in getting any such very weak signal into every cell in the body — which is required if one is to heal blood diseases such as AIDS and leukemia. After all, that was the reason that Priore utilized a powerful magnetic field of thousands of gauss. The magnetic field penetrated every cell in the body — even those in the bone marrow where all the blood cells are manufactured. By using the all-penetrating field as the carrier, the phase conjugated healing signal pattern could thus be introduced into every cell in the body, bathing it completely, inside and out, with the restorative signal.

Remember, **one must not just get the signal into the cells themselves — instead, one must get the restorative signals directly into, and absorbed in, the atomic nuclei.**

Again, that is why Antoine Priore found it necessary to employ a "rippling" magnetic field. The "ripple" was actually a magnetic wave, and nuclear resonance then provided the magic mechanism to penetrate all the atomic nuclei.

So the initial problem is, how does one provide a mechanism to carry the desired signals into and through each and every cell of the body and into each and every atomic nucleus of the matter of the body?

Obviously one can utilize nuclear magnetic resonance, after the fashion of Priore. If so, the resulting apparatus is going to be extremely large and expensive. It would be highly desirable to do it a different, simpler, much cheaper way.

After many hundreds of back-breaking experiments, one of my colleagues appears to have discovered a completely unique and direct way of introducing the desired EM restorative signals into and completely through every cell of the body, and into every atomic nucleus. Though much additional work to confirm this still remains, the initial results are marvelously encouraging.

Another colleague has succeeded in developing a peculiar sort of detector that should prove adaptable to detecting the actual "biopotential structural patterns" themselves, directly in and out of the atomic nuclei. Though obviously much more work is necessary before the final instrumentation is ready, the preliminary results are again most encouraging.

As I write these words on paper, this work is proceeding, but very, very slowly due to lack of the necessary funds to attack the research problems in force.

What is Needed

Time is running out —most conventional science appears to be driven by special interest groups/Drug Manufacturers who are unable and/or unwilling to counter the AIDS epidemic with anything other than extensive "Addict Style" symptom-reducing drugs that generate dollars, not cures. Though this is not universal, it *is* the conventional norm.

Unconventional disorders or diseases require *unconventional* science for *unconventional* cures.

Immediate funding is required if this awesome threat to humanity is to be stopped. An American public alerted in time to this desperate situation can demand that the government and/or private business immediately address and act on this life-threatening issue in a different, unconventional manner.

Encouraging Preliminary Work

One of my associates, already familiar with scalar EM devices, has exerted every effort to try to reduce the scope of the problem. In literally hundreds of experiments, he has been able to narrow down the search, and obtain at least some very promising results.

He has obtained an initial candidate phase conjugated signal for further test and trial.

He may well have succeeded in discovering a new and unique method to communicate signals directly into the atomic nuclei inside the matter in the human body. This signal presently appears to directly interact with the cellular biopotential and with the body's master cellular communication system.

Remember, however, these are preliminary results. They still must be fully substantiated in a great many more tests. Undoubtedly a great deal more research, analysis, and adaptation is necessary. We certainly cannot presently say we have any sort of "cure" for the AIDS virus or anything else — or that these results are to be considered as proven in any **fashion.**

But what we can say is that my colleague has been able to derive a very complicated phase-conjugated signal, which produces negative energy and negative time of the general sort required. And he may just have made a most marvelous discovery that will point the way to eventual equipment much, much smaller than the 4-stories high machine with which Priore intended to treat terminal human cancer patients.

The initial signal discovered by my colleague, when applied to the body at miniscule voltage, seems to zap all sorts of "bad things" — viruses, harmful bacteria, dangerous protozoa, microworms, you name it — without harming the blood cells, the normal body cells, or the hosted friendly bacteria.

The signal does this at miniscule voltage and amperage.

It can be applied directly to the body through special electrodes. Through a special feature, my colleague has been able to get the signal to reverberate the entire body, all cells, all parts, and even penetrate the atomic nuclei and establish scalar resonance therein.

There turned out to be some extremely strange things that have to be done to the phase conjugated signal before it will accomplish what is being sought. At least some of these "strange things" have been uncovered by my colleague.

One of the peculiarities is that the entire electrical apparatus is part of the input "form" (that conditions the potential wave structure) being phase conjugated and sent to the organism. If a lead-acid battery is included in the apparatus to power it, one will inject the electromagnetic form for the battery acid directly into the organism,* destroyingit. This includes destroying the host's cells. In this case the signal is lethal, not curative. Substitution of a dry cell battery with no liquid electrolyte eliminates the problem. Exactly why a liquid electrolyte has a toxic effect and a "sludge" or solid electrolyte does not, is not understood at this time.

Other such anomalies in the tentative process have been discov ered and compensated for.

However, the way ahead is exciting. It suggests that the body

---

*Via a mechanism similar to that found by Reid and Barsamian (atomic nuclei) can literally be "charged up" (i.e., the living biopotentials can be "charged up" with the signal structure) so that the "disease-proofing" is very lasting, possibly for many years or even for a lifetime.

At least my colleague has pressed this to the point of demonstrating a long-lasting charge being acquired by the body.

For example, at one time his body became so "charged" from his lengthy experiments that a one-inch blue spark often leaped from his fingers when he reached out for something metallic. The discharge was cool, negative energy — living energy, if you will. It should be negentropic, not entropic. It was definitely not the type of energy the orthodox scientific community is accustomed to. And normal electricity will definitely not charge up the body in such a continuing fashion, so far as is known.

Let me clearly state again that we have not yet produced the specific anti-pattern per se. What my colleagues have discovered appears to be a broad-band signal that appears to act hyperspatially, analogous to the manner in which a broad-band drug such as penicillin acts biochemically. Even this much remains to be clearly established.

However, it is a most encouraging and promising first step.

Much more work, and a great deal of experiments to substantiate or adapt these tentative results, still need to be done. Now there is no substitute for rigor and thoroughness — and there is no substitute for clearly and scientifically demonstrating the proof of the concept in the laboratory.

It is not just good intentions that we seek, but solid, concrete, proven results substantiated by proper scientific procedures. Much work remains to be done.

But the preliminary results are very encouraging indeed.

Let me briefly share with you some of the things we foresee, if this present line of successful development continues.

We foresee being able to eventually develop and set up — legally and under proper medical auspices, of course — tested and proven devices that can easily treat up to two or three hundred persons at once. A treatment of about 45 minutes to one hour is all that would be required. Several repetitive treatments a week or so apart, might prove advisory.

In addition,once the entire gamut of the treatment process is validated and proven, and shown to be completely harmless and safe in accordance with legal medical requirements, if need arises we foresee simply adding the signal to ordinary radio and television transmitters — perhaps as simply as modulating the electrical ground. If so, a "maintenance" signal could be established to negate the AIDS virus (or other disease such as cancer and leukemia) in an entire area, and keep out other diseases such as malaria, sleeping sickness, etc.

Another advantage of such machines would be their portability in time of conflict. Also, they would be most useful indeed as defensive measures against biological warfare. For example, the inhalation of only a relative minor amount of anthrax agent is sufficient for 100 percent certain death unless treatment begins promptly. With the portable machines, however, this would be easily negated in any troops exposed to the agent, even before the lethal disease is evidenced. For a totally new agent, a "phase-conjugate delta signal" could be rather quickly ascertained and developed, and the proper settings made on the machines for immediate treatment and immunizing.

And even later, when we proceed to the direct engineering of the living energy form strata themselves, we shall see remarkable cures and remissions of diseases for which medical science offers little hope today. Arthritis, multiple schlerosis, lupus, and other such debilitating diseases come readily to mind. **Even reversal of the aging process should be possible.***

*Two other promising approaches have recently been discovered First, an electrolyte compound has recently been approved by the FDA for clinical testing. This compound has the remarkable property of raising the cellular electrical potential back to that of a strong, healthy cell. In several years of lab animal testing, this alone was indicated to be over 80% effective against cancer. It may also prove effective against diseases such as arthritis, where the body's immune system attacks body tissue with lowered cellular potentials, since it fails to recognize the weakened cells as those of the body. Second, Baylor University reseachers have found that treating blood with certain laser EM radiation kills the AIDS virus, but does not harm the blood cells This means that it will be possible to assure that blood used in transfusions will be AIDS - free, eliminating one sorce of AIDS transmission.

Conclusion

We have now come to the end of our road, literally and figuratively.

All of us have been struck a mortal blow by the Soviet AIDS first strike.

Make no mistake, this is **real.** We and our children are already as good as dead unless we move as we have never moved before.

We have a chance. A slim chance.

Americans have always come through when the chips are down. We can conquer this thing. We can overcome this mighty death blow that has been launched against us. We can defeat the others yet to come.

But we've got to move. *Now.*

Remember, Pandora's box has already been spilled. Even without the Soviet biological warfare strike, hosts of new viruses and different strains of old ones are going to be, and are now being, dumped into the biosphere by our own culture. It is also only a matter of time before terrorists and meglomaniacs turn to the use of this potent weapon against a wide-open society such as ours.

Both our Armed Forces and our civilian populace are totally defenseless against electromagnetic biological warfare. Now. At this moment.

Even without the Soviet BW strike implication, it is only a matter of time until we perish, unless we develop electromagnetic healing and electromagnetic biological warfare countermeasures.

We've come to one of those profound momentary pauses in history that determine the fate of the entire world henceforth.

It's like the parable of the lady or the tiger.

We're facing, so to speak, two doors.

Behind one is the most fearsome and hungry tiger of all time. If we delay, that door will open and we'll get the tiger full upon us. We shall be utterly destroyed. That mighty tiger will consume half the world in his roaring frenzy. Our children, and our children's children — what few of them will be left — will wear the hammer and sickle yoke for eons.

Those few future survivors will be taught strange things. How you and I were the real enemy. How we were absolutely destroyed for the good of all mankind. How glorious and necessary it was to unleash the great plagues upon us. And how heroic were those who performed the "noble deed."

Our own distant children will curse us and revile us, and they will be taught to worship at the throne of a false prophet.

The world will descend into a new "Dark Ages" far more frightful than George Orwell ever envisioned.

On the other hand, behind the second door is the most beautiful lady of all history. Literally all of humanity's dreams of health, beauty, and vitality lie behind that door.

If we open it, we achieve a freedom from disease and a measure of bountiful health for all mankind that has heretofore only been dreamed of. *Even reversal of the aging process itself lies beyond the second door. Health and youth — the dream of the ages—can be ours.*

But we have only a moment to open the second door. It is firmly shut, and we must exert ourselves to the fullest if we are to open it at all.

On the human stage, the first door is already slowly opening, inexorably. In only a few moments it will be open and the tiger will be upon us.

We must move quickly.

Which will it be for mankind, the lady or the tiger?

The next few moments in the human play will most assuredly tell.

SELECTED BIBLIOGRAPHY

Note: An extended bibliography for this book would consist of literally thousands of references. Since there has not been time for that, a few key references on most substantive matters are listed. Where possible, one or more of the listed references will direct the interested reader to the pertinent literature on that subject.

References Related to Chapters 1 and 2

1. Pearce Wright, Science Editor: "Smallpox vaccine 'triggered Aids virus'," **The Times** of London, Mon. May 11, 1987, p. 1, 18.

2. Jon Rappoport, "News blackout on pox vaccine link to AIDS protecting WHO?", **Easy Reader,** June 4, 1987, p. 12.

3. Jonathan Tennenbaum, "Parameters for a research mobilization against AIDS," **Executive Intelligence Review (EIR),** Science & Technology, Mar. 6, 1987. p. 38-45.

4. Steve Connor, "AIDS: mystery of the missing data," **New Scientist,** Feb. 12, 1987, p. 19.

5. Steve Connor, "AIDS: Science stands on trial," **New Scientist,** Feb. 1987, p. 49-54.

6. "The virus reveals the naked truth," **New Scientist,** Feb. 12, 1987, p. 55-58.

7. Michael Gold, "Cancer Cover-up: How honest scientists evaded the issue of renegade cells that contaminated research labs worldwide," **Science Digest,** May 1986, p. 64-67, 80-81.

8. Michael Gold, **A conspiracy of Cells,** State University of New York Press, Albany, New York, 1985.

9. Kenneth J. McLeod, Raphael C. Lee, and H. Paul Ehrlich, "Frequency Dependence on Electric Field Modulation of Fibroblast Protein Synthesis," **Science,** Vol. 236, Jun. 12,1987, p. 1465-1468.

10. "Leukemia Virus Linked to Nerve Disease," **Science,** Vol.236, May 29, 1987, p. 1059-1061.

11. R.C. Desrosiers, "Origin of the human AIDS virus," **Nature,** vol. 319, 1986, p. 728.

12. R.B. Strecker, "AIDS virus infection," J.R. **Soc.** Med., Vol 79, 1986, p.559-560.

13. S.A. Aaronson, "Common Genetic Alterations of RNA Tumor Viruses grown in Human Cells," **Nature,** Vol. 230, 1971, p. 445-447.

14. A. Chandra et al, "Serological relationship between reverse transcriptase from human T-cell lymphotropic viruses defined by monoclonal antibodies," FEBS Lett., Vol. 200,1986, 327-332.

15. E. Dermott, J.K. Clarke and J. Samuels, "The Morphogenesis and Classification of Bovine Syncytial Virus," **J. Gen. Virol.** 1971; 12:105-119.

16. G. Corneo and L.C. Nelli, "Could Bovine Leukemia Virus be a possible Agent of Some Human Lymphatic Leukemias?", **Acta Haemat.** 72; 1984:65-66.

17. C.A. Mims, "Vertical Transmission of Viruses," **Microbiol. Rev.** 45; 1981:267-286.

18. C.W. Molander et al, "Isolation and Characterization of Vi ruses from Fetal Calf Serum," **In Vitro** 1972; 7:168-173.

19. P. Legrain, B. Goud and B. Buttin, "Increase of Retroviral Infection in Vitro by the Binding of Antiretroviral Antibodies," **J.Virol.** 1986; 60:1141-1144.

20. S. Kennedy -Stoskopf and O. Narayan, "Neutralizing Antibodies to Visna Lentivirus: Mechanism of Action and Possible Role in Virus Persistence," **J. Virol.** 1986; 59:37-44.

21. E.K. Knott, "Development of Ultraviolet Blood Irradiation, **Am. J. Surg.** 1948; 76:165-171.

22. Dr. John Seale, "The Origins of AIDS," Mar. 15, 1987.

23. R.B. Strecker and T.A. Strecker, "The Bioattack Alert (on AIDS)," Los Angeles, California, 1986.

24. R.B. Strecker, "Aids Virus Infection," J.R. **Soc. Med.,** vol. 79, Summer 1986, p. 559-560.

25. T.E. Bearden, "AIDS: Urgent Comments on Mankind's Greatest Threat," June 1987. Privately printed and circulated. "Aids: Urgent Comments on Mankind's Greatest Threat; and the Secrets of Electromagnetic Healing," presented in absentia at the U.S.P.A. Annual Symposium, July 30-Aug. 2,1987, Golden, Colorado.

26. **Executive Intelligence Review (EIR)** Special Report, "Global Showdown: The Russian Imperial War Plan for 1988," July 24, 1985.

27. **Executive Intelligence Review (EIR)** Special Report, "The Soviet role in covering up the deadly threat of AIDS," 12(42), Oct. 25,1985. Charges that Soviet infectious-disease experts control the Communicable Diseases Division of the World Health Organization. Nearly hysterical responses were stimulated by this report: for example, less than a week later, **Literaturnaya Gazeta,** Oct. 30, 1985 published a counter, and claimed the AIDS virus may have been produced in U.S. laboratories by the CIA and the Pentagon.

28. Jeanne Toal, "Supervaccines," **Omni,** July 1987, p. 20, 122.

29. The Kaznacheyev work. See appropriate section for references.

30. The Priore work. See appropriate section for references.

31. William Kucewicz, "The gates slam shut on a microbiologist," "Beyond Yellow Rain: The Threat of Soviet Genetic Engineering," Sixth of a Series, **Wall Street Journal,** May 8,1984. p. 34.

32. Department of Defense, **Soviet Military Power,** 1987.

33. William Kucewicz, "A non-stop Russian response to WWI," "Beyond Yellow Rain: The Threat of Soviet Genetic Engineering," Seventh of a Series, **Wall Street Journal,** May 10,1984. p.34.

34. Lewin, Leonard C, **Report from Iron Mountain on the**

Possibility and Desirability of Peace. The Dial Press, New York, 1967. If you want to read something that will curdle your blood, take a look at this book. There exists a surprising body of material where intelligentsia and "think tank" mentalities have grappled with the idea of non-warfare types of "control mechanisms" for humanity — including induction of biological disease on a mass scale, subtle forms of slavery, etc.

35. Bennett,William and Joel, Gurin, "Science that Frightens Scientists," **Atlantic Monthly,** Feb. 1977. Millions of persons in this country were inadvertently injected with a widely-used substance called simian virus 40, or SV40. Tests proved that SV40 caused cancerous cell growth when injected into animals. SV40 was an unrecognized contamination of polio vaccines given before 1962 and of some cold shots given about the same time. The authors conclude by stating: "The ultimate question is not whether bacteria can be contained in special laboratories, but whether scientists can be contained in an ordinary society."

36. Letter, Lynn M. Draft, White House Policy Staff, to Dick Gregory, April 11, 1977. Quoting: "It is true that the SV40 virus was discovered in the early 1960's. It is also true that the virus was found in certain viral vaccines prepared from virus pools grown in monkey kidney cell cultures. However, the virus had not been recognized as a contaminant prior to that time, although millions of people have received the vaccine during the 1950's."
"... contamination of the vaccine was unintentional due to lack of knowledge."

37. Snead, Dr. Eva Lee, M.D. "AIDS - Immunization related syndrome," **Health Freedom News,** July 1987, p. 14-17, 22, 44-45. Strongly and rigorously points out just now little we actually know, even today, about the cause of AIDS and some of the anomalies connected with it. Specifically, SV40 does not appear to have been searched for in AIDS victims. Lack of detection has little meaning, since SV-40 may disappear and not reappear until triggered or promoted by another virus or chemical. Yet SV-40 has been shown to be one of the activators of the HIV or AIDS virus, as reported by Howard E. Gendel man et.al, in an article published in the **Proceedings of the National Academy of the United States of America,** vol. 83, p. 9759-9763, Dec. 1986. Also, SV-40 carries other viruses

"piggyback" and may have that type of role in AIDS disease. SV-40 may predispose one to a secondary infection by nullifying the immune system. SV40 is ideally suited for genetic manipulation, splicing, and the creation of hybrids or mutants.' Dr. Snead rigorously points out that no one has proved as yet that any virus positively causes AIDS. What has been proven is that HTLV-III is present in large percentages of the AIDS victims. Dr. Snead raises many excellent issues which cry out for answers.

38. Pearl Ma and Donald Armstrong, **The Acquired Immune Deficiency Syndrome and Infections of Homosexual Men,** Yorke Medical Books, 1984.

39. Ostrow, David G; Terri A. Sandholzer and Yehudi M. Felman, Eds., **Sexually Transmitted Diseases in Homosexual Men,** Plenum books, 1984.

40. Snead, Dr. Eva Lee, M.D., **aids-update,** Jun. 15,1987. In this incisive summary, Dr. Snead "tells it like it is" about Salk vaccine, how vaccines are made, the lack of sufficient scientific accuracy and safety in making vaccines, the nearly general contamination of vaccines by at least some viruses, some macabre details of the constitution of the vaccine 'soups." etc.

41. Gallo, Robert C, "The AIDS Virus," **Scientific American,** 256 (1), Jan. 1987, p. 46-56.

42. **Institut Pasteur v. United States of America,** United States Court of Claims, Civ. A. No. 730-85. See also Steve Connor, "AIDS: Science Stands On Trial," **New Scientist,** Feb. 12, 1987, p. 49-58. Reveals some direct monetary interests involved in discovery of the AIDS virus, and in royalties for the blood test which detects the antibody to the virus. Extended litigation occurred between French scientists at the Pasteur Institute and Robert Gallo regarding who should receive credit for discovery of the virus causing AIDS. The two sides eventually agreed to share the credit for the discovery, and donate royalties to an AIDS research foundation. Note that Dr. Gallo is head of the Laboratory of Tumor Cell Biology at the National Cancer Institute.

43. Curran, James W. et al, "The Epidemiology of AIDS: Current Status and Future Prospects," **Science,** vol. 229, Sep.**27,**

1985, p. 1352-57. Among other things, this report downplays the role of non-viral agents in causing AIDS.

44. Duesberg, Peter H., "Retroviruses as Carcinogens and Pathogens: Expectations and Reality," **Cancer Research,** vol. 47, Mar. 1,1987, p. 1199-1220. Cites 278 references. Duesburg is a pioneer in the study of retroviruses and a professor in the Department of Molecular Biology and Virus Laboratory, University of California at Berkeley. To show some of the scientific disagreement as to the cause of AIDS, we quote from Dr. Duesberg's article: "AIDS virus is not sufficient to cause AIDS and ... there is no evidence, besides its presence in a latent form, that it is necessary for AIDS." Dr. Duesberg presents an impressive set of technical objections to the present, theory that the accepted AIDS virus actually causes AIDS.

45. **Steadman's Medical Dictionary,** Williams and Wilkins, Baltimore, 1966. p. 874. A German bacteriologist named Robert Koch established three "laws" to determine whether a certain micro-organism is the cause of a specific disease. These laws provide a basis for scientifically investigating infection and its causes. Koch's laws are: (1) Whether it be a bacterium, protozoan, or virus, the micro-organism must be present in all cases of the disease. (2) Inoculating animals with pure cultures of the micro-organism must produce the same disease in them. (3) It must be possible to obtain the micro-organism from these animals and further propagate the micro-organism in pure cultures. (Note that this leaves a real problem with any disease caused by complexes of different organisms, rather than just one.)

46. Altaian, Lawrence K., "Aids Development in Infection; Data Suggest AIDS Rises Yearly After Infection," **New York Times,** Mar. 3, 1987, p. C1. Among other things, points out that the AIDS seropositivity rate rose from 4 percent in 1978 to 68 percent in 1984 among 6,875 members of a San Francisco hepatitis B study group.

47. Moyo, "Fighting AIDS in Africa," **AfricAsia,** May 1987, p. 55

48. Altaian, Lawrence K., "AIDS in Africa," **New York Times.** Nov. 1, 1986, p. Al, A8.

49. Krieger, Nancy, "The Epidemiology of AIDS Riddle," **Science for the People,** Jan./Feb. 1987, p. 18.

50. Nordheimer, Jan, "Florida Pig Farm Poses AIDS Riddle," **New York Times,** May. 2, 1986, p. A7.

51. Saul, Stephanie, "A Focus for the AIDS Mystery," **Newsday,** Jan. 26, 1986, p. 7.

52. Segal, Jakob and Lilli, **AIDS: USA Home-Made Evil: Not Imported from Africa,** date unknown (believed 1986). Pamphlet widely circulated in Africa. The two scientists are a biologist and a biophysicist, affiliated with the Humbolt University of Berlin. In the pamphlet the authors charge that the AIDS virus was manmade, and suggest that it may have been created at Fort Detrick, Maryland. This and other such literature, plus deliberate Soviet propaganda releases charging that "U.S. biological warfare" is responsible for the AIDS pandemic, have significantly influenced widespread belief in Africa that the U.S. government is responsible. Our own thesis is that this merely shows how clever and coordinated the Soviet BW strike really is: the deception plan was complete and ready made. The viruses were made elsewhere (in cancer research labs), someone else unwittingly injected the contaminated vaccines (World Health Organization), and a good cover story was immediately at hand (simply blame the U.S.).

53. Pearce Wright and Thompson Prentice, "WHO seeks evidence over vaccine link to Aids virus," **The Times** of London, May 12, 1987.

54. "Near disaster with the Salk vaccine," **Science Digest,** Dec. 1963.

55. Bernard Fields, Malcolm A. Martin, and Daphne Kamely, Eds., "Genetically Altered Viruses and the Environment," **Banbury Report,** Cold Spring Harbor Laboratory, 1985.

56. The following summary articles are in Science, 239 (4948), Feb. 5, 1988: (1) Peter Piot et al, "AIDS: An International Perspective," p. 573-579; (2) Anthony S. Fauci, "The Human Immunodeficiency Virus: Infectivity and Mechanisms of Pathogenesis, " p. 617-622; (3) James W. Curran et al, "Epidemiology of HIV Infection and AIDS in the United States," p. 610-616; and (4) Richard W. Price et al, "The Brain in AIDS: Central Nervous System HIV-1 Infection and ATDS Dementia Complex," p. 586-592.

References Related to Chapter 3

1. Publications by **T.E.** Bearden:

- **a.** **Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons.** Tesla Book Co., POB 1649, Greenville, Texas 75401, 1986. Extensive coverage of Soviet scalar EM weapons and their testing. Heavily documented and illustrated. See a review of **Fer-de-Lance** in **Defense Science & Electronics,** June 1987, p. 80.
- b. **Excalibur Briefing.** Foreword by John White. Strawberry Hill Press, San Francisco, California 1980. An updated second edition, jointly published by Strawberry Hill Press and Tesla Book Co., 1988.
- c. **Tesla's Secret and The Soviet Tesla Weapons.** 1 hr. 40 min. videotape. 1981. Available from Tesla Book Co. Generally poor quality; lots of information.
- d. "Soviet Psychotronic Weapons: A Condensed Background." **Specula,** Journal of the AAMS, POB 1182, Huntsville Alabama 35807. 1(2), Mar.-June 1978. pp. 18-32. (Journal is now defunct).
- e. **Star Wars Now! The Bohm-Aharonov Effect, Scalar Interferometry, and Soviet Weaponization.** Tesla Book Co., 1984.
- f. **Soviet Weather Engineering Over North America**. 1-hr, videotape, 1985. Available from Tesla Book Co. Excellent quality; a great deal of information.
- g. **Solutions to Tesla's Secrets and the Soviet Tesla Weapons.** John T. Ratzlaff, ed. **Part II: Reference Articles for Solutions to Tesla's Secrets.** Tesla Book Co., 1981.
- h. "Tesla's Electromagnetics and Its Soviet Weaponization," **Proceedings of the 1984 Tesla Centennial Symposium,** International Tesla Society, Colorado Springs, Colorado 1984.
- i. USSR: New Beam Energy Possible?", **Defense & For-**

eign Affairs Daily, 13 (111), June 12, 1984, p. 1-2.

j. T.E. Bearden and Andrew Michrowski, eds., The **Emerging Energy Science,** Planetary Association for Clean Energy, Ottawa, Canada, 1985.

k. "AIDS: Urgent Comments on Man's Greatest Threat; and the Secrets of Electromagnetic Healing," presented in absentia at the U.S.P.A. Annual Symposium, July 30-Aug. 2, 1987, Golden, Colorado.

1. "Soviet Phase Conjugate Weapons: Weapons that use Time-Reversed Electromagnetic Waves," **Bulletin,** Committee to Restore the Constitution, POB 986, Ft. Collins, Colorado 80522, Jan. 1988.

m. "Tesla Electromagnetics: Weapons or Healing?" Presented at the Whole Life Expo, Los Angeles Airport Hilton Hotel, Los Angeles, California, Oct. 3, 1987.

2. William C. Douglass, **The Cutting Edge,** Nov. 1987, p. 3.

3. Stefan T. Possony, "Psy-War: Soviet Device Experiment," **Defense** & **Foreign Affairs Daily,** 12(104), June 7,1983, p. 1-2. Reports on Dr. Ross Adey's investigation of the Soviet LIDA device which is used to bombard human brains with radio waves of complex waveform in the 40 MHz region. Also reports on work by A.S. Davydov of the Ukranian Academy of Sciences, who discovered how the blood-brain barrier can be penetrated by low-frequency radio waves so that brain cells are affected.

4. Stefan T. Possony, "The Tesla Connection," **Defense & Foreign Affairs,** Aug. 1984, p. 12-14, 27. Highly recommended.

5. David Jones, "Israel's Secret Weapon," **Weekend Magazine,** Dec. 17, 1977.

6. Daniel A. Walker, Charles S. McCreery, and Fermin J. Oliveira, "Kaitoky Seamount and the Mystery Cloud of 9 April, 1984," **Science,** 227(4584), Feb. 8,1985, p. 607-611. See also Daniel L. McKenna and Daniel A. Walker, "Mystery Cloud: Additional Observations," **Science,** 234(4775), Oct. 24,1986, p. 412-413 for update to the information in the first article.

7. Robert J. Durant, "An Underwater Explosion — or What?",

Pursuit, 5(2), April 1972, p. 30-31.

8. Gwynne Roberts, "Witness to a Superweapon?", The **Sunday Times,** London, England, Aug. 17, 1980.

9. Central Intelligence Agency (CIA) Information Report No. CO-B-321/15354-66, "Sighting of Unusual Phenomenon on Horizon Near Iranian/USSR Border," Sep. 8,1966, released on 15 Dec. 1978 under the Freedom of Information Act (FOIA).

10. CIA Foreign Intelligence Information Report CO B-324/33601-76, "Aerial Observation of intense Source of Light," Nov. 18, 1976, released on 15 Dec. 1978 under the FOIA.

11. Max Frankel, "Khrushchev Says Soviet Will Cut Forces a Third; Sees 'Fantastic Weapon'", **New York Times,** Jan. 15, 1960, p. 1.

12. Christopher S. Wren, "Brezhnev Calls for Accord against 'Terrifying Arms," **New York Times,** June 14, 1975, p. 1,11.

13. Malcolm W. Browne, "Senatorial Group Received by Brezhnev," **New York Times,** July 3, 1975, p. 2.

14. Christopher S. Wren, "Moscow Now Pressing Disarmament," **New York Times,** Aug. 12, 1975, p. 6.

15. Draft agreement introduced by Andrei A. Gromyko to the United Nations' thirtieth session of the General Assembly on Sept. 23,1975: "Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons." The first article provided that the types of these new weapons would be "specified through negotiations on the subject."

16. "Secret Speech: Did Brezhnev Come Clean?", **National Review,** 29(8), Mar. 4, 1977, p. 248, 250.

17. William Beecher, "Brezhnev termed detente a ruse. 1973 report said," **Boston Globe,** Feb. 11, 1975, p. 1, 12.

18. See a series of articles by C.B. Baker in **Youth Action** NEWS,

POB 312, Alexandria, Virginia 22313: "Soviet Weather Mayhem." Nov., 1983; "Diabolic Soviet Warfare," Nov. 1984; "The Soviet Zapping of America," Dec. 1986; "Fifth Column Treachery," May 1987; and others.

19. "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," Associated Press (AP) Release, **Huntsville (Alabama) Times,** Feb. 20,1985, p. 1. Also "China Airlines Pilot Denies Cockpit Error," AP Release, **Huntsville (Alabama) Times,** Feb. 21, 1985, p. A-9.

20. Paul Brodeur, **The Zapping of America,** W.W. Norton & Co., New York, 1977.

21. "Navy sabotages traffic lights, TV sets," United Press International (UPI), PM cycle, Jan. 11,1985; "Power blackout blamed on anti-radar particles," UPI, AM cycle, Jan 11, 1985; "Power Disrupted Briefly After Navy Plane Drops Metallic Strands," AP, AM cycle, Jan. 16 1985.

22. Bradley Wells, "The Russian woodpecker: a continuing nuisance," CQ **Magazine,** Nov. 1984.

23. "Explosive Events Seen on Soviet Island," **Aviation Week & Space Technology,** Sept. 26,1983, p. 31. Also Letter John M. Miller, Geophysical Institute, University of Alaska, Jan. 41985 with attachment, "Bennett Island plume cases recently found." Incredibly, the U.S. State Department is now attempting to **give away,** to the Soviet Union, the five-island chain of Wrangel, Herald, Bennett, Henrietta and Jeannette Islands and an outer continental shelf portion of Alaska twice the size of California! If this happens, the Bering Sea will become a Soviet pond.

24. Sam Bishop, "UA photos show Soviets break ice for missiles," **Fairbanks Daily News-Miner,** Dec. 15 1984, p. 1,3. See also Craig Covault, "Soviet ability to fire through ice creates new SLBM basing mode," **Aviation Week & Space Technology,** Dec. 10, 1984. See also "Wrangel over Wrangel and the Weather War," Don Bell Reports, 32(9), Mar. 8,1985. See also, "Wrangel Giveaway Imminent," **The Spotlight,** Jan. 18, 1988,p.1,3.

25. Philip J. Klass, "Anti-Satellite Laser Use Suspected," **Avia-**

tion Week & Space Technology, Dec. 8, 1975, p. 12-13.

26. Joe Schwartz, **Nature,** 280, July 12, 1979, p. 95.

27. V.D. Sokolovskiy, **Soviet Military Strategy,** third edition 1968, p. 298.

28. T.T. Wu, "Electromagnetic Missiles," **J. Appl. Phys.** 57,1985, p. 2370-2373. See also "Filter Center," article on SDI, **Aviation Week & Space Technology,** Feb. 29, 1988, p. 55.

29. Michael McClelland, "State officials bend an ear to boom talk in Panhandle," **News Tribune,** Fort Pierce, Florida, Apr. 22, 1984.

30. Susan Burgess, "NASA officials stumped by strange light, boom," **News Tribune,** Fort Pierce, Florida, Nov. 28, 1985.

31. Bernie Woodall, "Residents are baffled by boom," **News Tribune,** Fort Pierce, Florida, Nov. 27, 1985.

32. Julie Enders, "Star of wonder: booming, bright," **News Tribune,** Fort Pierce, Florida, Dec. 6, 1985.

33. April 8, 1986 Canadian Broadcast Network television news interview of three eyewitnesses to the crash of the Arrow Air DC-8 jet liner at Gander Air Force Base, Newfoundland on Dec. 8, 1985. A yellow light was seen on the aircraft.

34. "Arrow Air was absolutely safe, FAA official says," UPI release, **Birmingham (Alabama) Post-Herald,** Apr. 17,1986.

35. "Titan Explosion Cripples U.S. Launch, Surveillance Capability," **Aviation Week & Space Technology,** 124(17), Apr. 28, 1986, p. 16-19.

36. Zhores A. Medvedev, **Nuclear Disaster in the Urals,** Translated by George Saunders, W.W. Norton & Co., 1979, p. 150

37. Gary North, "Electronic Aids (Parts 1 & 2)," **Remnant Review,** 12 (20), Nov. 1,1985, p. 1-8; 12 (21), Nov. 15,1985, p. 1-8. POB 8204, Fort Worth, Texas 76124.

38. Captain John D. LaMothe, **Controlled Offensive Behavior**

— **USSR,** Defense Intelligence Agency (DIA) Report ST-CS-01-169-72, 1972 (Released under FOIA).

39. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and Czechoslovakian Parapsychology Research,** DIA Report DST-1810S-387-75, 1975. (Released under FOIA).

40. John Bentley, **The Thresher Disaster: The Most Tragic Dive in Submarine History,** Doubleday & Co., Garden City, New York, 1975.

41. Canadian Aviation Safety Board (CASB), **Public Inquiry Into the Aircraft Accident Which Occurred Near Gander Airport, Gander NFLD. December 12, 1985,** Vol. 1, Apr. 8, 1986. Hull, Quebec, Canada.

42. Department of Defense, **Soviet Military Power,** 1987.

43. "America's Space Program Sabotaged?" Robert Herzberg, **International Combat Arms,** Mar. 1987. p. 12.

44. Pearce Wright, Science Editor: "Smallpox vaccine 'triggered Aids virus'," **The Times** of London, Mon. May 11, 1987. p. 2.

45. Jon Rappoport, "News blackout on pox vaccine link to AIDS protecting WHO?", **Easy Reader,** June 4, 1987, p. 12.

46. R.B Strecker and T.A. Strecker, "The Bioattack Alert (on AIDS)," Los Angeles, California, 1986.

47. EIR Special Report, "The Soviet role in covering up the deadly threat of AIDS," 12(42), Oct. 25, 1986.

48. Dr. John Seale, "The Origins of AIDS," Mar. 15, 1987.

49. Martin Ebon, **Psychic Warfare: Threat or Illusion,** McGraw Hill Book Co., New York, 1983.

50. For details of mysterious, instant deaths in Afghanistan, see Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," **Jane's Defense Weekly,** 1(13), Apr. 7, 1984, p. 508.

51. " 'Observers' at the August Minuteman Shot?", **The INFO**

Journal, 2(1), Fall 1968, p. 35.

52. Dietrick E. Thomsen, "The Mystery of the Cosmic Gamma-Ray Zaps," **Science News,** 105 (22), June 1, 1974, p. 357-358.

References Related to Phase Conjugation (Time Reversal)

1. Amnon Yariv, **Optical Electronics.** Third Edition. Holt, Rinehart and Winston: New York, 1985. See particularly Chapter 16: "Phase Conjugate Optics — Theory and Applications."

2. Robert A. Fisher, Ed., **Optical Phase Conjugation.** Academic Press, New York, 1983. Collection of papers. An extensive bibliography is given at the end of the book. Note that Fisher's introduction to the phase conjugation effect was as a result of two Russian visitors who briefed scientists at Lawrence Livermore Laboratory on optical phase conjugation. Also note (p.xv, p. 20) that almost any nonlinear optical effect can phase conjugate an incoming beam.

3. B. Ya Zel'dovich and V.V. Shkunov, "Spatial-Polarization Wavefront Reversal in Four-Photon Interaction," **Soviet Journal of Quantum Electronics,** Vol. 9, 1979. p. 379.

4. B. Ya Zel'dovich et al, "Connection between the wave fronts of the reflected and exciting light in stimulated Mandel'shtam-Brillouin scattering," English translation, **Sov. Phys. JETP** 15,109(1972).

5. B. Ya Zel'dovich, N.F. Pilipetsky, and V.V. Shkunov, **Principles of Phase Conjugation,** Vol. 42, Springer Series in Optical Sciences, Theodor Tamir, Ed., Springer-Verlag, New York, 1985.

6. J.P. Woerdman, "Formation of a transient free carrier hologram in Si," **Opt. Commun.,** vol. 2, 1971. p. 212.

7. B.I. Stepanov, E.B. Ivakin, and A.S. Rubanov, "Recording two dimensional and three-dimensional dynamic holograms in bleachable substances," Translation: **Soviet Physics-Doklady-Technical Physics,** vol. 16, 1971. p. 46.

8. A. Yariv, "Three-dimensional pictorial transmission in optical

fibers," **Appl. Phys. Lett.,** vol. 28, 1976. p. 88.

9. A. Yariv, "On Transmission and recovery of three-dimensional image information in optical waveguides," **J. Opt. Soc. Amer.,** vol. 66, 1976. p. 301.

10. R.W. Hellwarth, "Generation of time-reversed wave fronts by nonlinear refraction," **J. Opt. Soc. Amer.,** vol. 67,1977. p. 1.

11. A. Yariv and D.M. Pepper, "Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing," **Opt. Lett.,** vol. 1, 1977. p. 16.

12. R.L. Abrams and R.C. Lind, "Degenerate four-wave mixing in absorbing media," **Opt. Lett.,** vol 2, 1978, p. 94. Erratum is contained in **Opt. Lett.,** vol 3, 1978, p. 205.

13. P.F. Liao, D.M.Bloom, and N.P. Economou, "CW optical wavefront conjugation by saturated absorption in atomic sodium vapor, "**Appl. Phys. Lett.,** vol. 32, 1978, p. 813.

14. A. Yariv, "Compensation for atmospheric degradation of optical beam transmission," **Opt. Commun.,** vol. 21, 1977. p. 49.

15. R.W. Hellwarth, "Third order susceptibilities of liquids and gases," **Progress in Quantum Electronics,** vol. 5, 1977. p. 1.

16. J. Auyeung et al, "A theoretical and experimental investigation of the modes of optical resonators with phase-conjugate mirrors," **IEEE J. Quant. Electr.,** vol. 9, 1979. p. 1198.

17. I.M. Beldyugin and E.M. Zemskov, "Theory of resonators with wave-front reversing mirrors," **Sov. J. Quant. Elect.,** vol. 9, 1979. p. 1198.

18. A. Yariv and T.L. Koch, "One way coherent imaging through a distorting medium using four-wave mixing," **Opt. Lett.,** vol. 7, 1982. p. 113.

19. J.O. White and A. Yariv, "Real time image processing via four-wave-mixing in a photorefractive medium," **Appl. Phys. Lett.,** vol. 37, 1980. p. 5.

20. N.V. Kukhtarev et al, "Holographic storage in electrooptic crystals," **Ferroelectrics,** vol. 22, 1979. p. 949.

21. M. Cronin-Golomb et al, "Theory and applications of four-wave mixing in photorefractive media," **IEEE J. Quant. Elec,** vol. 20, 1984. p. 12.

22. J.P. Huignard, J.P. Herriot and G. Rivet, "Phase conjugation and spatial frequency dependence of wavefront reflectivity in $Bi_{12}SiO_{20}$(BSO) crystals," **Opt. Lett.,** vol. 5, 1980. p. 102.

23. J.O. White et al, "Coherent oscillation by self-induced gratings in the photorefractive crystal $BaTiO_3$," **Appl. Phys. Lett.**, vol. 40, No. 6, 1982. p. 450-452.

24. J.W. Haus, CM. Bowden, and C.C. Sung, "Optical phase conjugation with smooth pump profiles," **Phys. Rev. A, General Physics,** vol. 35, No. 8, Third Series, Apr. 15, 1987. p. 3398-3405.

25. Mary J. Miller et al, "Time response of a cerium-doped $Sr_{0.75} Ba_{0.25} Nb_2 0_6$ self-pumped phase-conjugate mirror," **Opt Lett.,** vol. 12, No. 5, May 1987. p. 340-342.

26. Mark Cronin-Golomb et al, "Passive (self-pumped) phase conjugate mirror: Theoretical and experimental investign tion," **Appl. Phys. Lett.,** vol. 41, No. 8, Oct. 15, 1982. p. 689 691.

27. Philip S. Brody and Richard P. Leavitt, "Dynamic holographic method of imaging phase objects," **Applied Optics,** vol. 26, No. 5, Mar. 1, 1987. p. 913-916.

28. J. Feinberg, "Self-pumped, continuous-wave phase conjugator using internal reflection," **Opt. Lett.,** vol. 7, 1982. p. 486.

29. T.Y. Chang and R.W. Hellwarth, "Optical phase conjugation by backscattering in barium titanate," **Opt. Lett.,** vol. 10, 1985. p. 408.

30. J.P. Huignard et al, "Phase-conjugate wavefront generation via real-time holography in $Bi_{12}Si0_{20}$ Crystals," **Opt. Lett.,** vol 4, 1979. p. 21.

31. David M. Pepper, "Nonlinear optical phase conjugation," guest editorial, **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 155. One of the leaders of the field of nonlinear optical phase conjugation (NOPC) tells one in clear English what all the excitement is about. He also points out that the conceptual extension of these techniques to other portions of the electromagnetic spectrum should follow.

32. David M. Pepper, "Nonlinear optical phase conjugation," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p.156-183. This paper is the best introduction to, and summary of, nonlinear optical phase conjugation (NOPC) in the English language. Also, on p. 156 the author specifically notes that "...these processing techniques can, in principle, be extended to other portions of the EM spectrum (e.g.,, rt, radio, microwave, radars, UV, etc.); and can also involve other fields (e.g., acoustic waves), given the proper nonlinear medium." In other words, phase conjugation is a universal nonlinear phenomenon, unknown until recently. An extensive bibliography of over 200 references is included.

33. Jeffrey O. White and Amnon Yariv, "Spatial information processing and distortion correction via four-wave mixing," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 224-230.

34. R.W. Hellwarth, "Optical beam phase conjugation by four-wave mixing in a waveguide," **Optical Engineering,** vol. 21, no. 2, Mar./Apr. 1982. p. 263-265.

35. I.M. Bel'dyugin, V.N. Seminogov, and E.M. Zemskov, "Possible wavefront reversal of fields using nonlinear-optics methods," **Sov. J. Quantum Electron.,** vol 9, no. 3, Mar. 1979. p. 385-387.

36. I.M. Bel'dyugin, M.G. Galushkin, and E.M. Zemskov, "Properties of resonators with wavefront-reversing mirrors," **Sov. J. Quantum Electron.,** vol 9, no. 1, Jan. 1979. p. 20-23.

37. G.S. Agarwal, Ari T. Friberg, and E. Wolf, "Elimination of distortions by phase conjugation without losses or gains," **Opt. Commun.** 43 (6), Nov. 15,1982. p. 446-450. Using scalar wave theory, the authors show the conditions for complete cancellation of distortion by phase conjugation: (1) scatterer is non-absorbing, (2) infinite phase conjugate mirror, (3) conjugate

wave generated without losses, (4) effects of evanescent waves are negligible outside the scatterer. Comment: By reflecting upon this paper, one can grasp that essentially the same approach can be used to show that phase conjugation generates Newton's third law of motion.

38. G.S. Agarwal and Emil Wolf, "Theory of phase conjugation with weak scatterers," J. **Opt. Soc.** Am., 72(3), Mar. 1982. p. 321-326. The effect of a distorting medium on an incident wave is eliminated by phase conjugation if certain conditions are satisfied: (1) the incident wave contains no evanescent components, (2) the transmitting medium is a weak, nonabsorbing scatterer, (3) backscattering of incident waves and of conjugate waves is negligible, (4) effects of scattered evanescent waves are negligible. Comment: A similar approach can be used to show that phase conjugation generates Newton's third law of motion.

39. T. Sato et al, "Application of evanescent waves to microscopic observation," **Bull. Tokyo Inst. Technol. (Japan),** No. 125, 1974. p. 35-41. Evanescent wave illumination for microscopy can be used for recognizing higher frequency components impossible to recognize by conventional illumination.

40. S. Twareque Ali. "Evanescent waves in quantum electrodynamics with unquantized sources," **Phys. Rev. D,** 7(6), Mar. 15,1973. p. 1668-1675. An identity between evanescent waves and virtual photons is established.

41. J.D. Lawson, "Some attributes of real and virtual photons," **Contemp. Phys. (Great Britain),** 11(6), Nov. 1970. p. 575-580.

42. J.L. Agudin et al, "The behaviour of evanescent waves under Lorentz transformation," **Lett. Nuovo Cimento (Italy),** vol. 23, ser. 2, no. 15, Dec. 9, 1978. p. 547-551. Two linear combinations of transverse magnetic and electrical evanescent waves (the left-and right-handed circularly polarized modes) remain the same in all systems under Lorentz transformation.

43. M. Hercher, "virometer—an instrument for the measurement of the size of viruses using an optical microscope." **Proc. Soc. Photo-Opt. Instrumentation Engineers,** vol. 126,1978. p. 17-22.

44. G.A. Massey, "Microscopy and pattern generation with scanned evanescent waves," **Appl. Opt. (Poland),** 13 (3), 1983. p. 247-255. This article shows how to materially violate the usual "wavelength limitation" of optical systems. Much better resolution can be obtained, and very sharp images can be performed. Comment: This mechanism may be the secret of the Rife microscope, which could see viruses and other small organisms otherwise quite beyond normal optical microscopy's wavelength resolution.

45. H.J. Gerritsen, "Nonlinear effects in image formation," **Appl. Phys. Lett.,** vol. 10, 1967. p. 237.

46. D. Gabor, U.S. patent No. 2,770,166.

47. H.W. Kogelnik, "Controlled transmission of waves through inhomogeneous media," U.S. Patent 3,449, 577.

48. V.I. Bespalov. Ed., **Optical Phase Conjugation in Nonlinear Media,** IPF AN SSR, Gorky, USSR, 1979 (in Russian).

49. B. Ya. Zel'dovich et al, "Phase conjugation by a surface," **Dok. Akad.Nauk SSSR 252,**1980, p. 92. English translation: **Sov. Phys. Dokl.** 25, 1980, p. 377.

50. F.V. Bunkin, D.V. Vlasov and Yu. A. Kravtzov, "On sound phase conjugation with amplification of phase-conjugated wave," **Kvantovaya Elektron.** 8, 1981, p. 1144. English translation: **Sov. J. Quant. Electron.** 11, 1980, p. 687.

51. A.G. Gyulamiryan et al, "Tunable nonlinear four-wave filter," **Opt. Spektrosk.** 52, 1982, p. 387. (In Russian).

52. R.Y. Chiao et al, "Stimulated Brillouin scattering and coherent generation of intense hypersonic waves," **Phys. Rev. Lett.** 12,1964,p.592.

53. S.A. Akhumanov et al, **Problems in Nonlinear Optics (Electromagnetic Waves in Nonlinear Dispersive Media), INI** AN SSSR, Moscow, 1964. (In Russian).

54. N.B. Baranova and B.Ya. Zel'dovich, "Wavefront dislocations and zeros of the amplitude," **Zh. Eksp. Teor. Fiz.** 80,1981, p. 1789. English translation: **Sov. Phys. - JETP** 54, 1981, p. 925. Might one infer from this paper a cautious hint (and

probe) at "do you Westerners know of scalar EM?"

55. V.I. Bespalov, Ed., **Phase Conjugation in Nonlinear Media,** IPF AN SSSR, Gorky, USSR, 1982. (In Russian).

56. D. Gabor, "A new microscopic principle," **Nature** 161,1958, p. 777.

57. Yu F. Kir'yanov et al, "Four-wave mixing in resonantly amplifying media under inversion saturation," **Kvantovaya Elektron.** 8,1981, p. 1734. English translation: **Sov.J. Quant. Electron.** 11, 1981, p. 1047.

58. V.L. Vinetsky et al, "Amplification of coherent beams by dynamic holograms in ferroelectric crystals," **Izv. Akad. Nauk. SSSR** 41, 1977. p. 811.

59. V.P. Kondilenko, G.G. Odulov and M.S. Soskin, "Amplified reflection of phase-conjugate waves in crystals with the linear electrooptical effect in an external electrical field," **Izv. Akad. Nauk. SSSR, Ser. Fiz.** 45, 1981, p. 958.

60. Yu. I. Kucherov et al, "Copropagating four-beam interaction in slowly-responding media," in **Phase Conjugation in Nonlinear Media,** ed. by V.I. Bespalov, IPF AN SSSR, Gorky, USSR, 1982, p. 111.

61. A.P. Brysev et al, "Model realization of a NbLi parametric phase-conjugating sound amplifier," **Pis'ma Zh. Tokhn. Fiz. 1,** 1982, p. 546.

References Related to William J. Hooper's Work

1. William J. Hooper, "Apparatus for generating motional electric field." U.S. patent #3,656,013, Apr. 11, 1972.

2. William J. Hooper, "New horizons in electric, magnetic and gravitational field theory," presented at a meeting of the Particles and Fields Division of American Physical Societies, University of Colorado, Boulder, Colorado, Aug. 21, 1969.

3. William J. Hooper. "New horizons for the BxV theory of gravity," undated, unpublished.

4. William J. Hooper, "Equivalence of the gravitational field and

a motional electric field," Session S6, American Physical Society Meeting, Chicago, Illinois Nov. 28-29, 1958.

5. William J. Hooper, "Equivalence of the gravitational field and a motional electric field," **Proceedings of the Boulder Conference on High Energy Physics,** Colo. Assoc. Univ. Press, 1970. p. 483.

6. William J. Hooper, "The motional electric field generator," **Bull, of the Am. Phys. Soc,** Series II, 15(2), Feb. 1970. p. 209.

7. J.H. Jeans, **The mathematical theory of electricity and magnetism,** 4th edn., Cambridge University Press, 1923. p. 125, 606.

8. Henry W. Kendall and Wolfgang Panofsky, "The structure of the proton and the neutron," **Scientific American,** June 1971.

9. William J. Hooper, "All-electric motional electric field generator." U.S. Patent #3,610,971, Oct. 5, 1971.

10. William J. Hooper, "Similarities of the motional electric and gravitational fields," undated, unpublished.

11. P. Moon and D.E. Spencer, "Some electromagnetic paradoxes," **Journal of the Franklin Institute,** vol. 260, 2955. p. 373.

12. Leigh Page and Norman I. Adams, **Electrodynamics,** D. Van Nostrand Co., 1940.

13. William J. Hooper, **New horizons in field theory,** revised 1965, unpublished. Chapter 1, "Fundamental Fields," is particularly important.

14. G.I. Cohn, **Electrical Engineering** 68,1969. p.441. Pointed out difference between flux linking law and flux cutting law of induction. Comment: Recently the Lorentz force law has been proven false. A modified version of Ampere's law results and saves the day.

15. G.I. Cohn, **Paradoxes of Electromagnetic Induction,** thesis, Illinois Institute of Technology library.

16. Pharis E. Williams, **"The Possible unifying effect** of the **dy-**

namic theory," Los Alamos National Laboratory report LA-9623-MS, Los Alamos, New Mexico, May 1983 .

17. Pharis E. Williams, **On a possible formulation of particle dynamics in terms of thermodynamic conceptualizations and the role of entropy in it,** thesis, U.S. Naval Postgraduate School, 1976.

18. Pharis E. Williams "The Principles of the dynamic theory," Research Report EW-77-4, U.S. Naval Academy, 1977.

19. Pharis E. Williams, "The Dynamic theory: A new view of space, time, and matter," Los Alamos Scientific Laboratory report LA-8370-MS, Dec. 1980.

20. Pharis E. Williams, "The Arrow of time in the dynamic theory," Los Alamos Scientific Laboratory report LA-8690-MS, Feb. 1981.

21. Pharis E. Williams, "The Dynamic theory: Some shockwave and energy implications," Los Alamos Scientific Laboratory report LA-8402-MS, Feb. 1981.

22. Dan C. Ross, "Plane electromagnetic waves in the dynamic theory," Los Alamos National Laboratory report, date unk.

23. Dan C. Ross, "Speed of light in the dynamic theory," Los Alamos National Laboratory, date unk.

24. G. Gilbert and B. McClain, "Fermions and stability in quan turn Kaluza-Klein theories," **Nuc. Phys. B, Part. Phy.s. (Netherlands),** B244(l), Sept. 24,1984. p. 173-185. Oscilla tions of the internal space produce conformal gravity waves in the four-dimensional space.

25. M. J. Perry, "Magnetic monopoles and the Kaluza-Klein theory," **Am. Inst. Phys. Conf. Proc.** No. 116,1984. p. 121-125. Kaluza-Klein theory admits soliton solutions which constitute magnetic monopoles. These monopoles have inertial mass, but no gravitational mass.

26. M.W. Kalinowski, "The nonsymmetric Kaluza-Klein theory." **J. Math. Phys.** 24(7), July 1983. p. 1835-1845.

27. D.J.Toms, "Induced Einstein-Maxwell action in Kaluza-Klein

theory," **Phys. Lett. B (Netherlands),** 129B(1-2), Sept. 15., 1983. p. 31-35.

References Related to Ehrenhaft's Work

1. Alden P. Armagnac, "Magic with magnetism," **Popular Science** Magazine, June 1944. p. 130-133, 222, 226.

2. "Ehrenhaft discovery confirmed by new experiments," **Popular Science** Magazine, date unk., p. 208. [Comment: Later than June 1944, and still during WWII.]

3. "Magnetic current — discovery of the age?", **Radio-Craft** Magazine, Mar., 1944. p. 332.

4. "Discovery of the age?" **Radio-Electronics** Magazine, date unknown, p. 58.

6. Felix Ehrenhaft and Emanuel Wasser, "New evidence of the existence of charges smaller than the electron," **Philosophical Magazine and Journal of Science,** Seventh Series, 5(28), Feb. 1928. p. 225-241.

8. Felix Ehrenhaft, "The microcoulomb experiment. Charges smaller than the electronic charge." **Philosophy of Science,** 8(3), July 1941. p. 403-457.

7. Felix Ehrenhaft and Leo Banet, "Is there 'true magnetism' or not?", **Philosophy of Science,** 8(3), July 1941. p. 458-462.

8. Felix Ehrenhaft and Leo Banet, "Magnetization of matter by light," Nature, Mar. 8, 1941. p. 297.

9. Felix Ehrenhaft, J. Franklin Inst., vol. 230, 1940. p. 381.

10. Felix Ehrenhaft, **Nature,** vol. 146, 1941. p. 25.

11. D. Morichini, **Gilberts Ann. Phys.,** vol. 43, 1813. p. 212.

12. D. Morichini, **Gilberts Ann. Phys.,** vol. 46, 1814. p. 367.

13. M. Sommerville, **Gilberts Ann. Phys.,** vol 52, 1826. p. 493.

14. F. Zantejeschi, **Gilberts Ann. Phys.,** vol. unk. 1829, p. 187.

15. V. Baumgartner, **Gilberts Ann. Phys.,** vol. unk., 1827. p. 508.

16. Felix Ehrenhaft, "Rotating action on matter in a beam of light," **Science,** 101(2635), June 29, 1945. p. 676-677.

17. Gordon Ferrie Hull, "The torque or rotating action in a beam of light," **Science,** vol. 101, 1945. p. 220.

18. F. Ehrenhaft, **Wiener Akad. Anz.,** vol. VII, Mar. 4, 1909.

19. F. Ehrenhaft, **Wiener Akad. Anz.,** vol. X, Apr. 21, 1910.

20. F. Ehrenhaft, **Wiener Berichte,** vol. 119, 1910. p. 815.

21. F. Ehrenhaft, **Physik. Ztschr.,** vol. 11, 1910. p. 619.

22. F. Ehrenhaft, **Physik. Ztschr.,** vol. 39, 1938. p. 673.

23. F. Ehrenhaft, **Ann. der Physik.,** vol. 56, 1918. p. 81.

24. F. Ehrenhaft, **Comp. Rend.,** vol. 190, 1930. p. 263.

25. F. Ehrenhaft, **Ann. der Physik.,** vol. 13, 1940. p. 151.

26. F. Ehrenhaft, **Jour. Frank. Inst.,** vol. 233, 1942. p. 235.

27. F. Ehrenhaft, **Nature,** vol. 154, 1944. p. 426.

28. F. Ehrenhaft, **Phys. Rev.,** vol. 65, 1944. p. 287.

29. F. Ehrenhaft, **Bull. Am. Phys. Soc.** (New York Meeting), vol. 6, 1945.

30. H.S. Renue, **Radio Electronic Engineering (Radio News),** vol. 4, 1945. p. 22.

31. Woldemar Voigt, **Festschrift fuer Heinrich Weber,** 1912.

32. P.A.M. Dirac, "Development of the physicist's conception of nature," **Symposium on the Development of the Physicist's Conception of Nature,** ed. Jagdish Merha, D. Reidel Pub. Co., Boston, 1973. p. 12-14.

33. Olof Alexandersson, **Living Water: Viktor Schauberger**

and the Secrets of Natural Energy, translated by Kit and Charles Zweigbergk, Turnstone Press Limited, Willingborough, Northamptonshire, England, 1982. An excellent bibliography is included in the book.

34. O. Alexandersson, "Implosion contra explosion," **Seklet,** vol. 4, 1965.

References Related to Louis Kervran's Work

1. Publications by Louis Kervran:

- a. **Biological Transmutations.** Crosby Lockwood, London, 1972.
- b. **A la Decouverte des Transmutations Biologiques, une Explication des Phenomenes Biologiques Aberrants.** Le Courrier du Livre, Paris, 1966.
- c. **Preuves Relatives a L'existence de Transmutations Biologiques, Echecs en Biologie a la loi de Lavoisier d'invariance de la Matiere,** Maloine, Paris, 1968.
- d. **Transmutations Biologiques; Metabolismes Aberrants de L'azote, le Potassium et el Magnesium,** Librairie Maloine, Paris, 1962.
- e. "Les Transmutations Biologiques en Agronomie," Maloine, Paris, 1970.
- f. **Biological Transmutations,** Swan House Publishing Co., Binghamton, New York, 1972.
- g. "Alchimie d'hier et d'aujourd'hui," **L'Alchimie, Reve ou Realite,** Revue des Ingenieurs do L'Institut National Superieur de Rouen, 1972-73.
- h. **Transmutations biologiques et physique moderne.** Paris: Maloine S.A., 1982.
- i. **Transmutations a faible energie (naturelles et biologiques).** Paris: Librairie Maloine, 1972.

- j. **Transmutations biologiques.** Rue de l'Ecole-de-Medecine, Paris: Librairie Maloine S.A., 1963.
- k. **Transmutations naturelles.** Librairie Maloine S.A., Rue de l'Ecole-de-Medecine, Paris Vie, 1963.

2. Rudolf Hauschka, **The Nature of Substance,** Vincent Stuart Ltd., London, 1966.

3. V.B. Neiman. ed., **Problems of Transmutations in Nature: Concentration and Dissipation.** (Collection of papers in Russian), Aiastan Pub. House, Erevan, Armenia, USSR, 1971.

4. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973.

5. The Kaznacheyev work and references. See appropriate sections.

6. Dr. Fritz Albert Popp's work and references. See appropriate sections.

7. The Australian work of Reid and Barsamian. See appropriate sections.

8. Fritz Albert Popp et al, eds. **Electromagnetic Bio-Information.** Proceedings of the Symposium, Marburg, September 5, 1977. Urban & Schwarzenberg, Baltimore, 1979.

9. Thomas E. Bearden, **The Excalibur Briefing,** Second edition, Strawberry Hill Press, San Francisco, 1988.

10. N.A. Kozyrev, "Possibility of the Experimental Study of the Properties of Time," Sep. 1967. English translation JPRS 45328, May 1968, p. 1-49.

11. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and** Czechoslovakian **Parapsychology Research,** DST-1810S-387-75, Defense Intelligency Agency, Washington, D.C., 1975. Released under the Freedom of Information Act.

12. Jeff Hecht, "IBM Reports on Dark-Pulse Solitons," **Lasers & Optronics,** 7(3), Mar. 1988, p. 42 -44.

References Related to Popp's Master Cellular Communication System

1. Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979.

2. V.P. Kaznacheyev et al, "On distant intercellular interactions in a system of two tissue cultures connected by optical contact," **Transactions of the Moscow Society of Naturalists,** Vol. XXXIX, 1972. p. 224-227.

3. F.A. Popp, **Biophotonen. Ein neuer Weg zur Losung des Krebsproblems.** Verlag fur Medizin, Heidelberg, 1976.

4. F.A. Popp et al, "Biosignale zur Steuerung des Zellstoffwechsels. Eine Resonanzhypothese der Karzinogenese (vorl. Mitt.)" **Munch, med. Wschr,** vol. 116, 1974. p. 381-384.

5. F.A. Popp, "Molecular aspects of carcinogenesis, in Deutsch et al, eds.: **Molecular Base of Malignancy**, Stuttgart, 1976.

6. A. Gurwitsch, "Die mitogenetische Strahlung der Optischen Bahn bei adaquater Erreguns." **Pfluger Arch. ges. Physiol.**, vol. 231, 1932. p. 254-264.

7. A. Gurwitsch, "Die mitogenetische Strahlung des markhaltigen Nerven," **Pflugers Arch. ges. Physiol.,** vol. 231,1932. p. 234-237.

8. A.A. Gurwitsch, "The problem of mitogenetic radiation as an aspect of molecular biology." (in Russian), Leningrad, 1968.

9. A.A. Gurwitsch, V.F. Eremeiev, and Yu. A. Karabchievsky, "Energy bases of the mitogenetic radiation and its registration on photoelectron multipiers," **Medicina,** Moskva, 1974.

10. J.W. Hastings, "Bioluminescence," **Ann. Rev. Biochem.,** vol. 37, 1968. p. 597-630.

11. L.W. Latmanisowa, "Die mitogenetische Sekundar-strahlung des Nerven," **Pflugers Arch. ges. physiol.,** vol. 231,1932. p. 265-279.

12. H. Ninnemann et al, "Inhibition of respiration in yeast by light," **Biochim. Biophys. Acta,** vol. 205, 1970. p. 499-506.

13. T.I. Quickenden and S.S. Que Hee, "Weak luminescence from the yeast saccharomyces cerevisiae and the existence of mitogenetic radiation," **Biochem. Biophys. Res. Comm.,** vol. 60, 1974. p. 764-770.

14. A. Babloyantz, "Self-organization phenomena resulting from cell-cell contact," **J. Theor. Biol.,** vol. 68, 1977. p. 551-561.

15. R.M. Benolken, "Reversal of photoreceptor polarity recorded during the graded receptor potential response to light in the eye of limulus," **Biophys. J.,** vol. 1961. p. 551-564.

16. H. Breithaupt, "Biological rhythms and communication," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore 1979. p. 1-23.

17. Herbert L. Konig, "Bioinformation - Electrophysical Aspects," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 25-55.

18. Ulrich Warnke, "Information transmission by means of electrical biofields," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds.., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 55-79.

19. Walter Peschka, "On kinetobaric effects and bioinformational transfer by electromagnetic fields," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 81-94.

20. Gunther Becker, "Communication between termites by means of biofields and the influence of magnetic and electric fields on termites," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-In-**

formation: Proceedings of the Symposium, Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 95-106.

21. Bernhard Ruth, Experimental investigations on ultraweak photon emission," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p.107-122.

22. Fritz Albert Popp, "Photon storage in biological systems," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 123-149. A most important paper, presenting what I have been calling the master cellular communication system. Ruth has confirmed the existence of the "degradation radiation," ("death photons"). Experimental results of Ruth and others have led Popp to postulate that biological systems generally have the capacity to store coherent photons which come from the external world, and that ultraweak photon emission from biological systems is governed by photon storage ("charge up") within the cell population.

23. S.S. Sung, "A possible biophotochemical mechanism for cell communication," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 151-174.

24. Helmut A. Fischer, "Photons as transmitter for intra- and inter-cellular biological and biochemical communication — the construction of a hypothesis," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 174-180.

25. Walter Kroy, "The use of optical radiation for stimulation therapy," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-In-**

formation: Proceedings of the Symposium, Marburg, September 5,1977, Urban & Schwarzenberg, Baltimore, 1979. p. 181-193.

26. Ulrich Warnke and Fritz Albert Popp, "Some aspects of magnetic influences on biological systems," in Fritz Albert Popp, Gunther Becker, Herbert L. Konig, and Walter Peschka, Eds., **Electromagnetic Bio-Information: Proceedings of the Symposium,** Marburg, September 5, 1977, Urban & Schwarzenberg, Baltimore, 1979. p. 195-199.

27. A.S. Presman, **Electromagnetic Fields and Life,** Plenum Press, New York, 1964.

28. F.A. Popp and B. Ruth, "Untersuchungen zur ultraschwachen Lumineszenz aus biologischen Systemen unter Berrucksichtig ung der Bedeutung fur die Arzneimittelforchung," **Arzneimittel-Forsch,** vol. 27 (I), 1977. p. 933.

29. B.Ruth and F.A. Popp, "Experimentelle Untersuchungen zur ultraschwachen Photonenemission biologischer Systeme," Z. **Naturforsch,** vol. 31c, 1976. **p.** 741.

References Related to Kaznacheyev's Experiments

1. V.P. Kaznacheyev et al, "Apparent information transfer between two groups of cells," **Psychoenergetic Systems,** 1(1), Dec. 1974.

2. V.P. Kazacheyev et al, "Distant intercellular interactions in a system of two tissue cultures," **Psychoenergetic Systems,** 1(3), Mar. 1976.

3. Captain John D. LaMothe, **Controlled offensive behavior — USSR,** Report ST-CS-01-169-72, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

4. Captain John D. LaMothe and Mr. Louis Maire, **Soviet and Czechoslovakian Parapsychology Research,** Report. DST-1810S-387-75, Defense Intelligence Agency, Washington, D.C. (Released under FOIA).

5. Trevor James Constable, **The Cosmic Pulse of Life,** Merlin

Press, Santa Ana, California, 1976.

6. John White and Stanley Krippner, eds., **Future Science,** Anchor, 1977.

7. B. Ruth, "Experimental Investigations on Ultraweak Photon Emission," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1979). p. 107-122

8. H. Fischer, "Photons as Transmitters for Intra- and Intercellular Biological and Biochemical Communication — the Construction of a Hypothesis," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5,1977 (Munchen-Wien-Baltimore, 1979). p. 175-180.

9. F.A. Popp, "Photon Storage in biological systems," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1979). p. 123-149.

10. A.G. Gurvitsch, **Theory of Biological Field,** Moscow, 1944 (in Russian).

11. A.G. Gurvitsch,"A Concept of 'the Whole' in the Frame of Biological Field Theory," in **Works on Mitogenesis and Theory of Biological Field,** Moscow, 1974, p. 141-147 (in Russian).

12. E.S. Bauer, "General Theory of Living Substance. The Problem of'Living Protein'," **Archives Biol. Sci.,** Vol. 35, Ser. A, Issue 1, 1934, p. 1-36 (in Russian).

13. E.S. Bauer, "Contradictions between the Inner and Outer Work of Living Systems," **Archives Biol. Sci.,** Vol. 35, Ser. A, Issue 1, 1934, p. 37-52 (in Russian).

14. E.S. Bauer, **Theoretical Biology,** Moscow, 1935 (in Russian).

15. S.V. Konev, "On the Question of Nature and Biological Significance of Ultraweak Luminescence of Cells," in **Bioluminescence,** Moscow, 1965, p. 181-183 (in Russian).

16. B.N. Tarusov, A.I. Pllivoda, and A.I. Zhuravlev, "Study of Ultraweak Spontaneous Luminescence of Animal Cells," **Biofzika,** Vol. 6, No. 4, 1961. p. 490-492 (in Russian).

17. M. Rottemeyer, "Modelle zur Interpretation der Ultraschwachen Photone-emission aus Biologischen Systemen," **Diplomarbeit** (Marburg, 1978), (in German).

18. Yu. A. Vladimirov, "Ultraweak Luminescence of Subcellular Structures," **Ultraweak Luminescence in Biology, Proceedings of the Symposium,** Moscow, 1969 (in Russian).

19. I.S. Marchenko, **Biofield of Forest Ecological Systems,** Bryansk, 1973 (in Russian).

20. S.S. Sung, "A Possible Biophotochemical Mechanism for Cell Communication," in **Electromagnetic Bioinformation, Proceedings of the Symposium,**, Marburg, September 5, 1977 (Munchen-Wien-Baltimore, 1971). p. 151-174.

21. P.A.Popp and B. Ruth, "Untersuchungen zur Ultraschwachen Luminiszenz aus Biologischen Systemen uter Berucksichtig ung der Bedeutung fur die Arznei-mittelforschung," **Arzneimit. Forsch.** (Drug Research), Vol. 27, No. 5,1977 (in German).

22. V.P. Kaznacheyev, "Informational Function of Ultraweak Flows of Light in Biological Systems," in **Materials of Scientific Conference: Physical and Mathematical Methods of Research in Biology and Medicine,** Novosibirsk, 1965, pp. 38-41 (in Russian).

23. V.P. Kaznacheyev, N.V. Ignatovich, and S.P. Shurin, "Intercellular Distant Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Ultraweak Luminescence in Biology,** Moscow, 1969 (in Russian).

24. V.P. Kaznacheyev, L.P. Mikhailova, and S.B. Stephanov, "Studies on Cellular Monolayer and Cellular Distant Interaction in High Latitude Conditions," in **Technical-Scientific Progress and Circumpolar Medicine,** Novosibirsk, 1978, p. 133-134 (in Russian).

25. V.P. Kaznacheyev, L.P. Mikhailova, V.N. Sudarev, and S.P

Shurin, "Studies on the Biological Role of Electromagnetic Emission as a Factor of Adaptive Behavior of Cells under Conditions of Latitude Changes," in **Abstracts of the Second Ail-Union Conference Devoted to the Problem of** Human **Adaptation to Different Geographical, Climatic, and Labor conditions,** Novosibirsk, 1977, p. 101-104 (in Russian).

26. V.P. Kaznacheyev, L.P. Mikhailova, V.N. Sudarev, and S.P. Shurin, "Distant Intercellular Interactions Caused by UV-radiation," in **Photobiology of a Living Cell,** Leningrad, 1979, p. 221-223 (in Russian).

27. V.P. Kaznacheyev, L.P. Mikhailova, D.G. Kadayeva, and M.P. Dranova, "Conditions Necessary for Appearance of distant Intercellular Interactions after UV-radiation," **Bulleten Experimentalnoy Biologii i Meditsiny.,** No. 5. 1979, p. 468-471 (in Russian).

28. V.P. Kaznacheyev, L.P. Mikhailova, and S.P. Shurin, "Distant Intercellular Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Regulation of Biosynthesis and Biophysics of Populations,** Krasnoyarsk, 1969, p. 372-374 (in Russian).

29. V.P. Kaznacheyev, S.P. Shurin, and L.P. Mikhailova, Discovery No. 122, "Distant Intercellular Interactions in a System of Two Tissue Cultures," **Official Bulletin of the Committee on Inventions and Discoveries Affiliated to the Council of Ministers of the USSR,** No. 19, 1973, p. 3.

30. V.P. Kaznacheyev, S.P. Shurin, L.P. Mikhailova, and N.V. Ignatovich, "Distant Intercellular Interactions in a System of Two Tissue Cultures Connected by Optical Contact," in **Ultraweak Luminescence in Biology,** Moscow, 1972, p. 224-227 (in Russian).

31. V.M. Inyushin, "The Concept of Biological Plasma and Some Questions of Photoenergetics," in **Questions of Bioenergetics,** Alma-Alta, 1968 (in Russian).

32. E. Makovsky, **Nature and Structure of Living Matter,** Bucharest, 1976 (in Russian).

33. A.L. Chizhevsky and Yu. Shishina, **In the Rhythm of the Sun,** Moscow, 1969.

34. Stefan T. Passony, "Psy-War: Soviet device experiment," **Defense and Foriegn Affairs Daily,** 12 (104), June 7,1983, p. 1-2.

35. V.I. Vernadsky, **Collected Works,** Moscow, 1960, (in Russian).

36. A.L. Chizhesvsky, **The Earth's Echo of the Sun Storms,** Moscow, 1973.

37. A.S. Pressman, **Electromagnetic Fields and Living Nature,** Moscow, 1968, English translation: **Electromagnetic Fields and Life,** Plenum, New York, 1970.

38. H.Z. Konig, "Bioinformation. Electrophysical Aspects," in **Electromagnetic Bioinformation, Proceedings of the Symposium,** Marburg, September 5,1977 (Munchen-Wien-Baltimore, 1979). p. 25-54.

39. V.P. Kaznacheyev, "Information Function of Ultraweak Light Flows in Biological Systems," in **Problems in Biophysics,** Novosibirsk, 1967, p. 7-18 (in Russian).

40. V.P. Kaznacheyev and L.P. Mikhailova, "Ultraweak Radiations in Intercellular Interactions," Novosibirsk, 1981 (in Russian); English Translation: San Francisco: Washington Research Center, 1982.

41. John F. Gilbey, "The delayed death touch," **Secret Fighting Arts of the World,** Charles E. Tuttle Co., Rutland, Vermont, 1963. p. 13-22.

42. Edward W. Russell, **Report on Radionics: Science of the Future,** Neville Spearman, London, 1973.

43. Paul Brodeur, **The Zapping of America,** W.W. Norton & Co., New York, 1977.

44. Waldemar Kaempffert," 'Dead' organisms revived," Science in the News: **Science Digest,** May 1950, p. 77-78.

45 "Rejuvenating light," **Scientific American,** May 1949.

46. Death-rays and life-rays discovered," **Science Digest,** Feb. 1950, p. 56. A "death" frequency seems to be at 2,537 angstroms, while a "life" frequency seems to be at 3,659 angstroms.

47. Lewis R. Roller, **Ultraviolet Radiation,** 2nd edn., John Wiley & Sons, New York, 1965. p. 236-237.

48. T.E. Bearden, Appendix: "Theory of Pulsor(R) operation," and "A partial glossary for scalar electromagnetics and subtle phenomena," in George T.F. Yao, **Pulsor(r): Miracle of Microcrystals,** Gyro Industries, Newport Beach, California, 1986.

49. The Antoine Priore material.

50. The Louis Kervran material.

51. Thomas E. Bearden, **The Excalibur Briefing,** Strawberry Hill Press, San Francisco, 2nd Edition, 1988.

52. Martin Ebon, "Moscow, June 11, 1977," in his **Psychic Warfare: Threat or Illusion,** McGraw-Hill Book Co., New York, 1983, p. 1-11.

References Related to the New View of the Nature of Mind and Thought

Author's Note: So far as I can find, there are no available references other than my own that deal precisely with my particular approach to the nature of mind and thought. There is, however, a serious debate over whether (a) the mind is simply the physical brain, similar to a computer (the prevailing materialist view of orthodox science), or (b) the mind is something else altogether, and metaphysical — at least with respect to the level of scientific knowledge today. My own approach utilizes a non-Aristotlean fourth law of logic (included in an appendix to this book) by means of which the dichotomy between mind and matter is removed. In that fashion, one can model the mind as material, or model matter us mind.

Some references that have influenced my thinking are given **below.**

Selected Bibliography

1. Publications by Thomas E. Bearden:

   a. "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," **Pursuit,** 13(1), Winter 1980. p. 6-10. (Appendix III to this present book.).

   b. "A mind/brain/matter model consistent with quantum physics and UFO phenomena," **Mutual UFO Network Annual UFO Symposium Proceedings,** 1979. Also available through the National Technical Information System, Port Royal Road, Springfield, Virginia 22161. Report AD-A068988.

   c. "An approach to understanding psychotronics." Paper delivered to 1977 Annual Symposium of the U.S. Psychotronics Association. Oglethorpe University, Atlanta, Georgia. 1977. Available through the National Technical Information System (NTIS), Port Royal Road, Springfield, Virginia 22161. Report AD-A027866. (This report is mostly rough notes).

   d. **Excalibur Briefing: Understanding Paranormal Phenomena.** Second edition. Foreword by John White. Strawberry Hill Press, San Francisco, 1988.

   e. "Hyperspace (virtual state) engineering." presented at the 11th MUFON UFO Symposium, Clear Lake City, Texas, June 6-8. 1980. Published in **UFO technology: a detailed examination: 1980 MUFON UFO Symposium Proceedings.** Edited by Walter H. Andrus, Jr. and Dennis W. Stacy. Mutual UFO Network, Seguin, Texas, 1980.

   f. "Photon quenching of the paranormal (time) channel: **a** brief note." Paper presented in absentia at the Third International Psychotronics Congress, Tokyo, 1977. Available through the NTIS, Port Royal Road, Springfield, Virginia 22161. Report AD-A038588.

   g. **Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena." NTIS, 1973. Report AD-763210.**

h. "The Boundary Identity of Exact Opposites: A simple Solution to the Age-old Philosophical Problem of Change." NTIS, Oct. 1975. Report AD-020798.

i. "The fourth law of logic." NTIS, 1979. Report AD-A068987.

j. **Field, Formon, Superspace, and Inceptive Cyborg: A Paraphysical Theory of Noncausal Phenomena.** U.S. Army Medical Intelligence and Information Agency, Report MIIA-1-74, Dec. 1974. Available from the Defense Documentation Center (DDC), Report AD A-005579/8GI.

k. "The One Human Problem, Its Solution, and Its relation to UFO Phenomena." **Pursuit,** January 1976. Available from NTIS, Report AD-A034236.

1. "Writing the Observer Back Into the Equation," Address given to the Princeton Center for Alternative Futures, Princeton, New Jersey, Mar. 5,1976, published in June, 1976. Available from NTIS, Report AD-A027867.

m. "Species Metapsychology, UFO Waves, and Cattle Mutilations, May 23, 1977. Privately published.

n. "The Holography of Being," 1975. Unpublished,

o. **Yoseikan Aikido,** 1973. Privately published.

p. "AIDS: Urgent comments on mankind's greatest threat; and the secrets of electromagnetic healing," presented in absentia at the annual USPA Symposium in Golden, Colorado 30 July-2 Aug. 1987.

2. Mike Costello, "Without a Brain," letter, Fortean Times, 96 Mansfield Road, London NW 8 2HX, England, Spring 1987, p. 78. Information on hydrancephaly, persons with normal intelligence but with large cavities in the area normally occupied by cortical tissue. Most of the data on these anomalous cases is buried in obscure medical journals. (Hydrocephaly — an excess of fluid in the cerebral/spinal system — is much more widely known.). See also Anthony Smith, **The Mind,** Hodder & Stoughton, 1984 for additional information on hydrancephaly.

Quoting from Costello: "Actually cases of this kind have been turning up at autopsies for years, and reported in the appropriate medical journals, but have been disregarded, perhaps because they are so destructive of established beliefs about the relationship between consciousness and the human brain."

Costello goes on to point out that a group at Sheffield University has collected a great deal of data on this subject, from case studies of 253 hydrocephalics and hydrancephalics. There seems to be no correlation between the IQ of these individuals and the amount of residual brain tissue.

One living subject, Roger, has an IQ of 126 and a First Class degree in Mathematics, yet has no more than 5% of the normal amount of brain tissue.

A Swedish documentary on the subject was shown on ITV in 1982.

Fortean Times 38, **p.** 30 also contains an item on surviving adults with no detectable brains.

3. Hugh Everett, **III,** The **Many-Worlds Interpretation of Quantum Mechanics: A Fundamental Exposition,** with papers by J.A. Wheeler, B.S. DeWitt and Neill Graham, Princeton Series in Physics, Princeton University Press, 1973.

4. Louise and Galen Hieronymus, **Tracking the Astronauts in Apollo 11, With Data From Apollo 8 Included,"** Advanced Sciences Research and Development Corporation, Fort Lauderdale, Florida, 1969.

5. Vlail Kaznacheyev's work and publications related to it (see Kaznacheyev references).

6. Antoine Priore's work and publications related to it (see Priore references).

7. Charles Muses and Arthur M. Young, eds, **Consciousness and Reality.** Outerbridge & Lazard, 1972.

8. George Meek, **Healers and the Healing Process,** Theosophical Publishing House, 1977.

9. Harold Puthoff and Russell Targ, **Mindreach,** Delacorte, 1977.

10. Harold Puthoff and Russell Targ, "A Perceptual Channel for Information Transfer Over Kilometer Distances: Historical Perspective and Recent Research," **Proceedings of the IEEE,** 64(3), March 1976.

11. Edward W. Russell, **Report on Radionics,** Neville Spearman, London, 1973.

12. John White and Stanley Krippner, eds., **Future Science,** Anchor, 1977.

13. C. Louis Wiedemann, "Results ofthe N.J. 'Spook Light' Study," **Vestigia Newsletter,** Stanhope, New Jersey, May 1977.

14. David Bohm, **Wholeness and the Implicate Order,** Routledge, Kegan Paul, 1980.

15. Renee Weber, "Reflections on David Bohm's Holomovement: A Physicist's Model of Cosmos and Consciousness," in Valle and von Eckartsberg, editors, **The Metaphors of Consciousness,** Plenum Press, 1981.

16. "The Physicist and the Mystic: Is a Dialogue Between Them Possible?" A Conversation with David Bohm, conducted by Renee Weber, edited by Emily Sellon, **ReVision,** 4(1), Spring 1981, p. 22-35.

17. The Tao of Physics Revisited," A Conversation with Fritjof Capra, conducted by Renee Weber, **Revision,** 4(1), Spring 1981, p. 36-52.

18. "Reflections on the New Age Paradigm," an Interview with Ken Wilber, **ReVision,** 4(1), Spring 1981, p. 36-52.

19. Fritjof Capra, **The Tao** of **Physics,** Shambhala, Boulder, Colorado, 1975.

20. Fritjof Capra, **The Turning Point,** Simon & Schuster, New York, 1982.

Ken Wilbur, **Up From Eden,** Doubleday, 1981.

22. G. Spencer, **Laws of Form,** Julian Press, New York, 1972.

23. Bernard D'Espagnat, **Conceptual Foundations of Quantum Mechanics,** W.A. Benjamin, Menlo Park, California, 1971.

24. Robert Linssen, **Zen: The Art of Life,** Pyramid, New York, 1969.

25. E.P. Wigner, "Remarks on the Mind-Body Question," **The Scientist Speculates, I.J.** Good ed., W. Heinemann Ltd., London,1961.

26. Thomas G. Hieronymus, U.S. Patent No. 2,482,773, Sept. 27, 1949.

27. Langston Day and George De La Warr, **New Worlds Beyond the Atom,** Vincent Stuart Publishers Ltd., London, 1973.

28. Jule Eisenbud, M.D., **The World of Ted Serios,** Pocket Books, Simon and Schuster, New York, 1968.

29. Sheila Ostrander and Lynn Schroeder, **Psychic Discoveries Behind the Iron Curtain,** Prentice-Hall, 1970.

30. Carl G. Jung, **Man and His Symbols,** ed. by Carl G. Jung **and** after his death by M.S. von Franz, Aldus Books Ltd., London. 1964.

31. **The Unfathomed Mind: A Handbook of Unusual Mental Phenomena,** compiled by William R. Corliss, The Sourcebook Project, Glen Arm, Maryland, 1982.

32. Charles Panati, ed., **The Geller Papers: Scientific observations on the paranormal powers of Uri Geller**, Houghton Mifflin Company, Boston, Massachusetts, 1976.

33. Ken Wilbur, **The Spectrum of Consciousness,** The **Theosophical** Publishing House, Wheaton, Illinois, 1977.

34. "Godel's proof," **Scientific American,** June 1965, p. 81-86
As Ken Wilbur puts it, "Thus logically as well as **physically** verification is not an absolute mark of reality. If all **is** to be verified, how do you verify the verifier, since he is part of **the** all?"

35. Erwin Schroedinger, **What is Life? and Mind and Matter,** Cambridge University Press, London, 1969.

36. Teilhard de Chardin, **The Phenomenon of Man,** Harper Torchbooks, New York, 1965.

37. D.T. Suzuki, **The Zen Doctrine of No Mind,** Rider, London, 1970.

38. Frederick S. Perls, **Gestalt Therapy Verbatim,** Real People Press, Lafayette, Indiana, 1969.

39. Frederick S. Perls, Ralf F. Hefferline, and Paul Goodman, **Gestalt therapy,** Dell, New York, 1951.

40. John White, ed., **The Highest State of Consciousness,** Anchor, New York, 1972.

41 J. Blofeld, trans., **The Zen Teaching of Huang Po,** Grove Press, New York, 1958.

42. R.E. Ornstein, **The Psychology of Consciousness,** W.H. Freeman and Company, San Francisco, California, 1972.

43. Joseph C. Pearce, **The Crack in the Cosmic Egg,** Julian, New York,

44. Andrija Puharich, **Beyond Telepathy,** Anchor Press, Doubleday & Company, Garden City, New York, 1973.

45. Jacques Vallee, **Passport to Magonia: From Folklore to Flying Saucers,** Henry Regnery Company, Chicago, Illinois, 1969.

46. Robert Bruce Lindsay and Henry Margenau, **Foundations of Physics,** Dover publications, New York, 1957. See p. 2-3 for the blunt statement that physics has nothing at all to say about a possible real world lying behind its concepts and equations. That is, physics is simply a model, and the assumption of a kind of reality is an interpretation and an hypothesis.

47. Ted Bastin, ed., **Quantum Theory and Beyond,** Cambridge University Press, 1971.

48. Edgar D. Mitchell, **Psychic Exploration: A Challenge for Science,** edited by John White, G.P. Putnam's Sons, New York, 1974.

49. Hector Hawton, **Philosophy for Pleasure,** Fawcett Publications, 1956, Fifth printing, June 1970. (A real gem.)

50. Jiro Kikkawa and Malcom J. Thorne, **The Behavior of Animals,** Plume, The New American Library Inc., New York, 1974.

51. Helena Curtis, **Biology,** Worth Publishers, Inc., New York, 1968 (fourth printing); p. 614-637 contains a good discussion of instinct, learning, and memory.

52. John Cunningham Lilly, **The Mind of the Dolphin: A Nonhuman Intelligence,** Avon, New York, Mar. 1969.

53. Isaac Asimov, **The Human Brain: Its Capacities and Functions,** Signet Science Library, The New American Library, Inc., New York, 1963.

54. R.W. Sperry, "The Great Cerebral Commissure," **Scientific American,** Jan. 1964, p. 42-52.

55. Michael S. Gazzaniga, "The Split Brain in Man," **Scientific American,** Aug. 1967, p. 24-29.

56. Roger Sperry, "Problems Outstanding in the Evolution of Brain Function," James Arthur Lecture, American Museum of Natural History, New York, 1964.

57. Jose M.R. Delgado, **Physical Control of the Mind: Toward a Psycho-Civilized Society,** Harper Colophon Books, Harper & Row, New York, 1971.

58. B.B. Kazhinskiy, **Biological Radio Communications,** Izdatel'stvo Akademii nauk Ukrainskey SSR, Kiev. 1962. Trans, available through DDC, AD 415676.

59. S.K Lisitsyn "New Approach to the Analysis of Electroencephalograms," in **Problems of Bionics (Selected Articles), p.** 16-25, DDC Report **AD** 730045.

60. Evert W. Beth, **The Foundations** of Mathematics, Harper Torchbooks, Harper & Row, New York, 1966. p. 481-518 contains a discussion of the "contradictions" of logic painfully noted by mathematicians shortly after the turn of the century.

61. Stanley Krippner and James Hickman, "West Meets East: A Parapsychological Detente," **Psychic,** May/June 1974, p. 51-55.

62. G.G. Globus, ed., **Consciousness and the Brain,** Plenum, New York, 1976.

63. Karl Pribram, **Emotion: the Search for Control,** McGraw Hill, New York, 1968.

64. Karl Pribram, **Languages of the Brain,** Prentice Hall, Englewood Cliffs, 1971.

65. Charles Hampden-Turner, **Maps of the Mind: Charts and Concepts of the Mind and Its Labyrinths,** Collier Books, Macmillan Publishing Co., New York, 1981.

66. Adam Smith, **Powers of the Mind,** Random House, New York, 1975.

67. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973. (a real gem.)

68. The Louis Kervran material.

69. The Antoine Priore material.

70. Langston Day and George De La Warr, **Matter in the Making,** Vincent Stuart, London, 1966.

71. Sister Justa Smith, "Paranormal Effects on Enzyme Activity," Professional Paper No. 2, **Human Dimensions,** 1(2), 1972.

72. John Pierrakos, **The Energy Field in Man and Nature,** Institute of Bioenergetic Analysis, New York, 1971.

73. Jeffrey Mishlove, **The Roots of Consciousness,** Random House, New York, 1975.

74. Andrija Puharich, "Psychic Research and the Healing Process," in **Psychic Exploration,** ed. John White, Putnam, New York, 1975.

75. Andrija Puharich, **Beyond Telepathy,** Darton, Longman and Todd, London, 1962.

76. Lyall Watson, **Supernature,** Hodder & Stoughton, London, 1973.

77. Jule Eisenbud, "The Mind-Matter Interface," **Journal of American Society for Psychical Research,** 69(2), Apr., 1975.

78. George W. Meek, **From Enigma to Science,** Samuel Weiser, New York, 1973.

References Related to Antoine Priore's Work

1. Antoine Priore. "Procedeet dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes." Republique Francais: Brevet d'Invention P.V. No. 899.414, No.1.342.772,1963. For the U.S. patent, see Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. patent No. 3,280, 816. Oct. 25, 1966.

2. Letter, R. Pautrizel to his colleagues at the University of Bordeaux, Sept. 11, 1979.

3. Letter, R. Pautrizel to his colleagues, Mar. 26, 1969.

4. Jacques Sylvain and Jean-Michel Graille, "Depuis quinze ans le "rayon Priore" divise le monde scientifique." [For fifteen years the "Priore Ray" has divided the world of science.]. **Sud-Ouest Dimanche,** Jan. 27, 1980.

5. "Affaire Priore: L'armee va trancher? [The Priore Affair: Will the Army resolve it?] **Sud-Ouest Dimanche,** May 20,1980.

6. "Affaire Priore: L'armee va tranchera." [The Priore Affair: The Army will resolve it.]. **Sud-Ouest Dimanche,** Apr. 12, 1981.

7. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Action de champs electromagnetiques sur les greffes de la tumeur T8 chez le Rat." [Action of the electromagnetic fields on the graftings of the TB tumor on the rat.]. **Compt. Rend. Acad. Sci.** (Paris), 1964, **259,** 4895-4897.

8. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Effets de champs electromagnetiques sur un lymphosarcome lymphoblastique transplantable du Rat." [Effects of Electromagnetic fields on lympheuplastic lymphosarcoma transplantable from a rat.]. **Compt. Rend. Acad. Sci.** (Paris), 1965, **260,** 2099-2102.

9. M.R. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomenes de regression observes sur les greffes d'un lymphosarcome chez des souris exposees a des champs electromagnetiques." **Compt. Rend. Acad. Sci.** (Paris), 1965, 260, 2639-2643.

10. G. Delmon and J. Biraben,"la croissance du carcinome de Guerin sous Faction de champs electromagnetiques." **Rev. Path,** comp, 1966, 3, no. 2, 85-88.

11. M.R. Riviere and M. Guerin, "Nouvelles recherches effectuees chez des rats porteurs d'un lymphosarcome lymphoblastique soumis a l'action d'ondes electromagnetiques associees a des champs magnetiques." **Compt. Rend. Acad. Sci.** (Paris), 1966, **262,** 2669-2672.

12. R. Pautrizel, M.R. Riviere, A. Priore, and F. Berlureau, "Influence d'ondes electromagnetiques et de champs manetiques associes sur l'immunite de la souris infestee par **Trypanosoma equiperdum."** [Influence of electromagnetic waves and associated magnetic fields on the immunity of the mouse infected with the Trypanosoma equiperdum.]. **Compt. Rend. Acad. Sci.** (Paris), 1966, **263;** 579-582.

13. R. Pautrizel, A. Priore, and F. Berlureau, and A.N. Pautrizel, "Stimulation par des moyens physiques des defenses de la Souris et du Rat contre la trypanosomiase experimentale." [Stimulation, by physical means, of defenses of the mouse and of the rat against the experimental Trypanosoma.]. **Compt. Rend. Acad. Sci.** (Paris), 1969, **268,** 1889-1892.

14. R. Pautrizel, A. Priore, F. Berlureau, and A.N. Pautrizel, "Action de champs magnetiques combines a des ondes electromagnetiques sur la trypanosomiase experimentale du Lapin" [Action of magnetic waves combined with electromagnetic waves on the experimental trypanosoma on the rabbit.]. **Compt. Rend. Acad. Sci.** (Paris), 1970, **271,** 877-880.

15. A.J. Berteaud, A.M. Bottreau, A. Priore, A.N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation entre revolution d'une affection par **Trypanosoma equiperdum** et l'action d'une onde electromagnetique pulsee et modulee." [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave.]. **Compt. Rend.** Acad. **Sci.** (Paris), 1971, **272,** 1003-1006.

16. R. Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, and D. Bernard, "Etat de protection vis-a-vis de **Trypanosoma equiperdum** chex des souris splenectomisees et soumises a une stimulation physique." **ler Multicolloque Europeen** de **Parasitologie,** Rennes, 1 au 4 Septembre 1971.

17. R. Pautrizel, A. Priore, M. Dallochio and R. Crockett, "Action d'ondes electromagetiques et de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholesterole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modiications in the rabbit by the administration of an alimentary hyperchloresterol diet.]. **Compt. Rend. Acad. Sci.** (Paris), 1972, **274,** 488-491.

18. G. Mayer, A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiquea sur 1' orchite trypanosomienne du lapin." [Action of the magnetic fields associated with electromagnetic waves on the trypanosomian orchitis of the rabbit.]. **Compt. Rend.** Acad. **Sci.** (Paris), 1972, **274,** 3011-3014.

19. R. Pautrizel, A. Priore, P. Mattern, and A.N. Pautrizel, "Stimulation des defenses de la souris trypanosomee **par** l'action d'un rayonnement associant de champ magnetique et ondes electromagnetiques." [Stimulation of the defenses of the trypanosomized mouse by the action of a ray associated with a magnetic field and electromagnetic waves.]. **Compt. Rend. Acad. Sci.** (Paris), 1975, **280,** 1915-1918.

20. R. Pautrizel, A. Priore, P. Mattern, A.N. Pautrizel, and A. Capbern, "Guerison de la trypanosomiase chronique du Lapin a **Trypanosoma equiperdum** par l'action combinee de champs magnetiques et d'ondes electromagnetiques modules." **J. Protozool.,** 1975, **22,** no. 3, A 84.

21. R. Pautrizel, "La trypanosomiase experimentale: stimulation des defenses de L'organisme par des moyens physiques." **XVIIth Seminar on Trypanosomiasis Research,** Londres, 22-23 Septembre 1976.

22. R. Pautrizel, P. Mattern, A.N. Pautrizel, and A. Priore, "Effets des champs magnetiques et des ondes electromagnetiques modules sur la trypanosomiase experimentale." [Effect of magnetic fields and modulated electromagnetic waves on the experimental trypansomiasmis.]. **Ann. Soc. Beige Med. trop.,** 1977, **57,** 501-523.

23. R. Pautrizel, P. Mattern, A. Priore, A.N. Pautrizel, A. Capbern and T. Baltz, "Importance des mecanismes immunitaires dan la guerison de la trypanosomiase experimentale par stimulation physique." [Importance of immune mechanisms in the cure of the experimental trypanosomias by physical stimulation.]. **Compt. Rend. Acad. Sci.** (Paris), 1978, 286, 1487-1492.

24. R. Pautrizel, and A. Priore, "Un aspect spectaculaire du bioelectromagnetisme: permettre a l'organisme de se debarrasser d'une infection aig ue ou chronique sans l'aide d'aucune substance medicamentese." [A spectacular aspect of bioelectromagnetism: permitting the organism to rid itself of an old chronic infection without the aid of any medical substance.] **Communication presented to the 104 eme Congres National des Societes Savantes,** Bordeaux, 17-21 Avril 1979 — section Sciences, p. 112.

25. R. Pautrizel, and A. Priore, "Guerison de la trypanosomiase experimentale par association de champs magnetiques et d'ondes electromagnetiques," [Cure of the trypanosomias by associated magnetic fields and electromagnetic waves.]. **Communication presented to the Symposium International des Therapeutiques Ondulatoires,** Versailles 19-20 Mai 1979, p. 9.

26. R. Pautrizel, A. Priore, A.N. Pautrizel and P. Chateaureynaud-Duprat, "Guerison de la trypanosomiase experimentale par l'association de champs magnetiques et d'ondes electromagnetiques: une stimulation des defenses de l'organisme." [Cure of the trypanosomias by associated magnetic fields and electromagnetic waves: a stimulation of the organism's defenses.]. **Communication presented to the Journees Nationales Microondes - Colloque Hertzienne et Dielectriques,** Lille 26-29 juin 1979, p. 31.

27. David M. Rorvik, "Do the French have a Cure for Cancer?" **Esquire** Magazine, July 1975, p. 110-111,142-149. Excellent summary of the Priore Affair, with some details of the working of the machine.

28. J.B. Bateman, "A Biologically Active Combination of Modulated Magnetic and Microwave Fields: The Priore Machine," Office of Naval Research, London Report R-5-78, Aug. 1978. Comes very close when he states that "The possibility that some hitherto unrecognized feature of the radiation from a rotating plasma may be responsible for the Priore effects should not be dismissed out of hand..." Indeed, none of the persons who checked the machine or pondered over it knew anything about phase conjugation — which plasmas can do.

29. J.B. Bateman, "Microwave Magic." Office of Naval Research London Conference Report, ONRL C-14-77, 1977.

30. D.S. Greenburg. "The French Concoction." **Saturday Rev. Sci.,** May, 1978. p. 36-44.

31. Lord Zuckerman, "The great Bordeaux magnetic mystery machine." **Sunday Times Weekly Review,** Jan. 7, 1973.

32. Lord Zuckerman, "Pride and Prejudice in Science. William Randolph Lovelace Commemorative Lecture. **Aerospace Medicine,** 1974, 45, 638-647.

33. L.B. Bateman, "Staging the perils of non-ionizing waves." Office of Naval Research London, **European Scientific Notes,** ESN 32-3: 85-88.

34. R. Courrier, 1977. "Expose de M. le professeur R. Courrier secretaire perpetuel de l'Academie des Sciences fait au cours d'une reunion a l'lnstitut sur les effets de la machine de M.A. Priore le 26 Avril 1977.

35. M. Riviere, A. Priore, F. Berlureau, M. Fournier and M. Guerin, "Phenomena of regression observed on the graftings of a lymphosarcome in mice exposed to electromagnetic fields." Mar. 1, 1965.

36. Jean-Michel Graille, **Le Dossier Priore:** De Noel, Paris, 1984. Published in France.

37. G. Lakhovsky, "Apparatus with circuits oscillating under multiple wave lengths," U.S. Patent No. 1,962,565. June 12, 1934.

38. G. Lakhovsky, "Tube for producing multiple wave lengths." U.S. Patent No. 2,351,055. June 13, 1944.

39. Thomas H. Moray, "Electrotherapeutic apparatus," U.S. Patent No. 2,400,707. Feb. 1, 1949.

40. Robert C. Beck, "Extremely low frequency magnetic fields and EEG entrainment: A psychotronic warfare possibility?" Audiotape; presentation at the 1978 U.S. Psychotronics Association (USPA) Conference, Atlanta, Georgia.

41. T.G. Hieronymus, "History of radionics. Parts 1 and 2." Audiotape; presentation at the 1979 USPA Conference, University of Houston, Texas.

42. Robert C. Beck, "Extreme low frequency magnetic pulses and E.E.G. entrainment: Soviet Psychotronic Weapon?" Audiotape; presentation at the 1979 USPA Conference, University of Houston, Texas.

43. Al Bielek and William Bise, "A magnetic pacer." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

44. Thomas E. Bearden, "Mind and matter interaction." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

45. John Moray, "Sea of energy." Audiotape; presentation at the 1980 USPA Conference, Dayton, Ohio.

46. T. Galen Hieronymus, "Sunspots trigger earthquakes." Audiotape; presentation at the 1981 USPA Conference. Available from USPA, 3459 Montrose Ave., Chicago, IL 60618.

47. Al Bielek and Bill Bise, "Measurements of ELF and its parameters." Audiotape; presentation at the 1981 USPA Conference.

48. Martin Ruderfer, "Misconceptions in established science and their relevance to psychotronics." Audiotape; presentation at the 1981 USPA Conference.

49. Material on the Kaznacheyev experiments. See Kaznacheyev section for references.

50. Material on phase conjugation. See phase conjugation section for references.

51. Material on Ehrenhaft experiments. See Ehrenhaft section reverences.

52. Material on Kervran's work. See Kervran section for references.

53. Extensive proprietary scalar electromagnetics experiments by Frank Golden and John Bedini.

54. Material on Royal Raymond Rife's work. See Rife section for references.

References Related to the Australian Work

1. Private correspondence, Dr. B.L. Reid.

2. Private correspondence, Dr. S. Barsamian.

3. B.L. Reid, "Propagation of properties of chemical reactions over long distance in the atmosphere as seen by crystal growth pattern changes," **Australian Journal of Medical Laboratory Science,** Vol. 7, Feb. 1986. p. 30-35.

4. H.E. Anderson and B.I. Reid, "Vicinal Long Range and Extremely Long Range effects on growth of sodium chloride crystals from aqueous solutions containing protein," **App. Phys.** Comm, vol. 4, 1984. p. 217.

5. W.H. Campbell and J.M. Young, "Auroral zone observations of infrasonic pressure waves related to ionosopheric disturbances and geomagnetic activity," **J. Geophys. Res.,** vol. 68, 1963. p. 5909.

6. G. Chimonas and C. Hines, "Atmospheric gravity waves launched by auroral currents," **Planet Space Sci.,** vol. 18, 1970. p. 565.

7. F.B. Abeles, **Ethylene in Plant Biology,** Academic Press, New York, 1973.

8. G.I. Distler, **Jour. Cryst. Growth,** Vol. 3 & 4, 1968. p. 175.

9. G.I. Distler and V.P. Vlasov, **Thin Solid Films,** vol. 3, 1969. p. 333.

10. G.I. Distler, **Jour. Cryst. Growth,** Vol. 9, 1971, p. 76.

11. B. DeWitt, **Phys. Rev.,** 19C, 1975, p. 297.

12. W. Greiner and J. Hamitton. **Amer. Sci.,** vol. 68,1980, p. 154.

13. P. Davies and T. Birrell, **Quantum Fields in Curved Space,**

    Cambridge University Press, Cambridge, 1982.

14. P. Mitchell, Nature,vol. 191, 1961. p. 144.

15. A Rothen. **Proc. Nat. Acad. Sci.** (U.S.), vol. 72,1975. p. 2462.

16. L. Niemeyer, L. Betroniero, and H.J. Weismann, **Phys. Rev. Lett.,** vol. 52, 1984. p. 1033.

17. T.A. Witten and L.M. Sander, **Phys. Rev. B.,** vol. 27,1983. p. 5686.

18. T.D. Lee, **Particle Physics and Introduction to Field Theory,** Harwood Academic Publishers, New York, 1981. See particularly Chapter 16, "Vacuum as the source of asymmetry."

References Related to Royal R. Rife's Work

1. Royal R. Rife, "Bacteria," Apr. 28, 1939.

2. Royal R. Rife, "The filterable virus of carcinoma," Apr. 26, 1939.

3. Royal R. Rife, "The Rife microscopes," Apr. 28, 1939.

4. Royal R. Rife, "History of the development of a successful treatment for cancer and other viruses, bacteria and fungi." Allied Industries, Dec. 1, 1953.

5. Arthur Isaac Kendall and Royal Raymond Rife, "Observations on Bacillus Typhosus in its filterable state," **California and Western Medicine,** California Medical Association, Dec. 1931.

6. Arthur Isaac Kendall, "Filtration of bacteria; studies in bacterial metabolism," **Science 75,** Mar. 18, 1932. p. 295-301.

   Arthur Isaac Kendall, "The filtration of bacteria," **Science,** Mar. 18, 1932.

8. E.C. Rosenow, "Observations on filter-passing forms...," **Proceedings of the Staff Meeting of the Mayo Clinic,** July 13, 1932.

9. E.C. Rosenow, "Observations with the Rife Microscope of Filter-passing forms of micro-organisms," **Science,** n.s., vol. 76, Aug 26, 1932. p. 192-193.

10. E.C. Rosenow, "Microdiplococci in filtrates of natural and experimental poliomyelitic virus compared under the electron and light microscopes," **Proceeding of the Staff Meeting of the Mayo Clinic,** 17(7), Feb. 18, 1942. p. 99-106.

11. E.C. Rosenow, "Observations of filter-passing forms of Eberthella Typhi (Bacillus Typhosus) and of the Streptococcus from Poliomyelitis," **Proc. Staff. Meet. Mayo Clinic** 7, July 13. 1932. p. 408-413.

12. E.C. Rosenow, "Transmutations within the Streptococcus-Pneumonococcus Group," **Journal of Infective Diseases,** vol. **14, 1914.**

13. "Powerful microscope makes very tiny objects visible," **Science Newsletter,** vol. 33, Jan. 22, 1938. p. 55.

14. "The Rife microscope or "facts and their fate'," Reprint No. 47, The Lee Foundation for Nutritional Research, Milwaukee, Wisconsin.

15. R.E. Seidel, M.D. and M. Elizabeth Winter, "The new microscopes," **Journal of the Franklin Institute,"** Feb., 1944. p. 103-129.

16. R.E. Seidel, M.D. and M. Elizabeth Winter, "Filterable bodies seen with the Rife microscope," Science Supplement, **Science,** Dec. 11, 1931. p. 10-11.

17. "Is a new field about to be opened in the science of bacteriology?" Editorial, **California and Western Medicine,** California Medical Association, Dec. 1931.

18. Christopher Bird, "What has become of the Rife microscope?", **New Age Journal,** Mar. 1976.

19. Peter Tompkins and Christopher Bird, **The Secret Life of Plants,** Harper & Row, New York, 1973.

20. Lida H. Mattman, **Cell Wall Deficient Forms,"** CRC Press, Cleveland, Ohio. 1974.

21. Georges Lakhovsky, **La Cabal: Histoire d'une Decouverte L'oscillation Cellulaire,"** G. Doin, Paris, 1934.

22. Georges Lakhovsky, "Apparatus with circuits oscillating under multiple wave lengths," U.S. Patentno. 1,962,565. June 12,1934. See also G. Lakhovsky," Tube for producing multiple wave lengths," U.S. Patent No. 2,351,055. June 13, 1944.
23. Georges Lakhovsky, **"L'oscillation Cellulaire," Ensemble des Recherches Experimentales,** G. Doin, Paris, 1931.

24. Harold Saxton Burr, **Blueprint for Immortality: The Electric Patterns of Life,** Neville Spearman Ltd., London, 1972.

25. George De La Warr and Douglas Baker, **Biomagnetism,** De La Warr Laboratories, Oxford, 1967.

26. Ruth Beymer Drown, **The Theory and Technique of the Drown H.V.R. and Radiovision Instruments.** (Private printing.). Artists' Press, Los Angeles, 1939.

27. Ruth Beymer Drown, **The Science and Philosophy of the Drown Radio Therapy.** Los Angeles, 1938.

28. Albert Abrams, **New Concepts in Diagnosis and** Treatment. Philopolis Press, San Francisco, 1916.

29. Albert Abrams, **Iconography: Electronic Reactions of Abrams.** San Francisco, 1923.

30. Jagadis Chandra Bose, **Response in the Living and Non-Living,** Longmans, Green & Co.; New York, 1902.

31. Jagadis Chandra Bose, **The Physiology of Photosynthesis,** Longmans, Green & Co.; New York, 1924.

32. D.M. Bose, "J.C. Bose's Plant Physiological Investigation Relating to Modern Biological Knowledge," **Transactions of the Bose Research Institute,** Vol. 37, Bose Research Institute, Calcutta, 1947-48.

33. "Invisible ultra-violet life frequencies made visible," **New Light on Therapeutic Energies,** Source/data unk., p. 51-55.

34. Wilheim Reich, **The Cancer Biopathy.** Orgone Press, New York, 1948.

35. V.M. Inyushin and P.R. Chekorov, "Biostimulation through laser radiation and bioplasma," Kazakh State University, U.S.S.R.

36. Philip Medley et al, "The filterable forms of bacteria," **Journal of Infectious Diseases,** vol. 48, 1931.

37. Gianni A. Dotto, "Method for constructing a thermionic pile," U.S. Patent no. 3,839,771. Oct. 8,1974.

38. Gianni A. Dotto, "Story of the Dotto Ring." notarized in Montgomery County, Ohio on Feb. 8, 1972. Certain parts of Dotto's theory, apparatus, and recommended treatment method suggest potential effects on recombinant genetic structures.

39. Gianni A. Dotto, "Electrostatic Wand," U.S. Patent No. 3,785,383. Jan. 15, 1974.

40. T.H. Moray, **The Sea of Energy.** 5th Edition. History and biography by John E. Moray. Foreword by Thomas E. Bearden. Cosray Research Institute, 2505 South 4th East, Salt Lake City, Utah 84115. 1978.

41. T.H. Moray, "Electrotherapeutic Apparatus," U.S. Patent no. 2,460,707. Feb. 1, 1949.

42. John Moray, "Radiant Energy," in **1st International Symposium on Non-Conventional Energy Technology,** Toronto, Canada. Planetary Association for Clean Energy, 1981. p. 316-319.

43. Antoine Priore's research and devices.

44. Wilheim Reich's orgone energy work and devices.

45. Radionics researchers' work and devices.

46. Jonathan Tennenbaum, "Soviet Work on Electromagnetic Pulse Weapons," EIR Special Report, **"Electromagnetic-Effect Weapons: The Technology and the Strategic Implications,"** Wiesbaden, Federal Republic of Germany, 1988.

47. Ron Iscoff, "Dye Lasers Used to Treat Blood," **Lasers & Optronics,** 7(3), Mar. 1988, p. 26-28.

APPENDIX I

The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science

© 1984, 1988 Christopher Bird (Used by permission)

Forty- four years ago, in 1944, an Italian engineer working as a prisoner and forced laborer for the Germans in the huge submarine base in Bordeaux, approached a French police agent to plead for his **life.** He would be killed when the Germans left Bordeaux, he said, and since they were by that time obviously losing the war, the day of his execution was at hand.

The police officer, who also worked clandestinely for the French underground, told the engineer to get in his car, then simply drove him out of the base and introduced him to the 7th battalion of underground resistance fighters, in the nearby province of Dordogne. There he so distinguished himself in military operations that he was ultimately decorated by the French government.

It was due to his thankfulness to his savior, and his loyalty to his companions-in-arms, that Antoine Priore decided after the war's end to live out the rest of his life in Bordeaux. Thus he became the focus of one of the strangest, and most scandalous, chapters in the scientific history of France or any other nation.

Antoine Priore had earlier graduated from a small provincial school for electricity in Trieste, Italy and become a radar operator in **the** Italian Navy. During this period he observed what to him was an exciting anomaly: some oranges left in a room filled with electrical bric-a-brac had fallen into an assemblage where they seemed to have been preserved in the same fresh state they had enjoyed when **bought** off a fruit stand. Other oranges in the room, bought at the same time, were rotten and putrid.

Stunned by his observation, Priore dreamed throughout the war of one day working out an electrical means of conserving foods in **their** fresh state based on what he surmised was a new, and wholly unexplained, principle. Newton's apple had become Priore's orange.

Occupied during the day as a humble electrical repairman — and **projectionist** in a movie theater — the almost wholly self-taught **Priore** devoted all his free time and his meager resources to research. With the help of his war-time companions, some of whom had attained high rank in the Bordeaux police force, he was able to beg, borrow, steal, scrounge, or otherwise acquire a mini-warehouse of electrical and electronic components and parts. With these he put together a device worthy of Rube Goldberg. Exposing lentil seeds to a magnetic field of 225 gauss and electromagnetic frequencies of 80, 32, 3 and 10 Hertz, Priore's device caused the lentil plants which sprouted from them to grow 12-15 centimeters in length, as against only 5 centimeters for controls not subjected to the same treatment. He got similar results for tulips, asparagus and other plants.

Shifting his focus, he next irradiated fertilized hens' eggs, only to see the chicks hatch in 19 days, instead of the normal 21. Though he could not explain these astonishing results, he realized he had stumbled upon a process basic to the enhancement, or speeding up, of cellular growth.

It was at this point that one of his police friends introduced him to Francis Berlureau, the former Director of Studies at the School for Veterinary Medicine in Toulouse and, at the time of their meeting, director of the Bordeaux abbatoir. Priore asked Berlureau to supply him with various animal tissues for experimentation. For 10 years they worked together, Priore's free time allowing, during which Priore noticed he could get no electrical measurement from a cancerous bull's testicles. Since he realized that, in some way, his newly constructed device (no trace of which remains today, except for a snapshot of it) affected the electrical properties of cells, he put two and two together and his sum of four led him to believe that he might be onto an electromagnetic cure for cancer. Newton's gravity had become Priore's cancer cure.

Berlureau next allowed him to expose a cat with cancer of the mammary glands to radiation of his machine. To make absolutely sure that he was not exposing himself to mockery, the veterinarian had all the histological work done by his friend and colleague, a Professor Drieux at the famous Veterinarian School of Maisons-Allfort, near Paris. Drieux wrote a technical report proving that a tumor taken from a cat had, before treatment, started to become cancerous and, after treatment, had become benign.

By 1953, with the help of a doctor of general medicine, Maurice Fournier, Priore began treating human patients whose cancers had been judged hopeless. The huge file of cases maintained by Fournier, and filed with a notary until after his death, was subsequently mysteriously lost. But a few details were preserved in letters discovered in an old dog-eared file.

Some of these dating to the year 1954 concerned a 12-year old boy, Alain B., whose diagnosis wavered between one of reticulo-histio-sarcoma and a malignant form of Hodgkin's disease. The boy was taken by his parents to Priore, who irradiated him. Though the exact nature of the radiation was not known, 12 years later a Bordeaux physician, after a medical examination, certified that the boy, now become a man of 24, was free of disease.

A second case unearthed from the old file indicated that a patient with cancer of the larynx was able to avoid a laryngectomy and be totally cured after Priore's new ministration.

Fascinated by the principle which he suspected must lie behind the strange Priore Ray, Dr. Berlureau tried to get some Bordeaux University physicists interested in the problem but was laughed out of their offices. He next turned to cancer specialists, beginning with Professor Lachapele, the Director of the Bergonie Foundation, a prestigious center for cancer research, to whom he proposed animal experiments to prove the efficacy of Priore's methodology. His plea met with a stony affirmation on Lachapele's part to the effect that he and his colleagues had no need of the new discovery, inasmuch as "all the patients treated in his hospital were cured and departed in perfect health." As if bound in the chains of his curt reply, years later Lachapele was to become one of the bitter adversaries of Priore's pioneering research.

Only somewhat discouraged, Priore kept up his momentum. He went on to build a new and more complicated version of his treatment device, called the P-l, over the next year. When it was finished he secretly and unofficially began to treat dozens of cancer patients who had been given up by their doctors as incurable. At his funeral in March of last year, among the crowd of mourners was, it is said, a small platoon of older people who had been cured of their terrible afflictions by Priore in the 1950s.

While his findings excited him, he nevertheless felt tremendously frustrated that he could apparently get no one in the world of medicine or science to pay attention to them. Undaunted by his previous rebuffs, his friend Berlureau next introduced the Italian at the end of 1959 or the beginning of 1960 to Professor Tayeau, vice dean of Bordeaux's Medical Faculty. Unlike Lachapele, Tayeau behaved as a true physician and scientist. He sent Priore to two researchers, Biraben, head of the Faculty's Department of Pathological Anatomy, and his assistant, Delmon. The two had been working together on cancerous rats for two years—specifically on animals grafted with T-8 tumors, discovered by the internationally famous team of Guerin and Oberling in Paris, which had proven to be intractable to any form of treatment yet known. To their utter surprise, the tumors in the rats treated with Priore's machine were reduced in volume by 60%, marking the first time in the history of cancerology that the virulent T-8 tumor had in any way been affected by any form of treatment.

Knowing that the mayor of Bordeaux, Jacques Chabans~Delmas-who has kept his post until this day, and was soon to become prime minister of France-- was most interested in the work of Priore (who, he too, had known as a fellow resistance fighter), they also informed Chaban.

Promptly Chaban convoked not one, but two, commissions made up of Bordeaux and Parisian scientists to study the Biraben-Delmon results in detail. Both commissions rejected Priore and his machine out of hand, and without appeal. It is curious that, in the science of our day, a result, undeniable though it may be, seems to have no hearing unless and until all means to effect it can be adequately explained. It was for this ostensible reason that the two commissions decided to so adamantly reject the research: Biraben and Delmon could not explain the nature of the radiation engendered by the Priore device.

One can stress the word *ostensible* here because the principal reason for the rejection lay elsewhere. The decision by the first commission was, in fact, hardly unanimous. But among its members was the same Professor Lachapele who had refused Berlureau's plea for assistance. His opinion was that even the results themselves were of little value because they were obtained, not on spontaneously arising, but on grafted, cancers. The fact that no treatment whatsoever had ever affected a T-8 tumor was totally discounted. As the sole cancerologist on the commission, Lachapele's dictum was preponderant.

When he learned that the rejection of the first commission had actually been a split decision, the Bordeaux mayor asked for the formation of a second commission to re-examine the problem. Fearing a reversal, Lachapele was able to get one of his colleagues, Professor Courtial, director of the Radium Institute in Paris, and one of the so-called top authorities of French cancer research, named to it. It was all but impossible for the other physicians on the new commission now to outvote not one, but two, cancer specialists, so again the antagonists won the day.

At no time did either of the commissions bother to interview Priore himself or to run a supplementary experiment under their own control.

This seemingly incomprehensible attitude on the part of scientific authority was only a foretaste of what was to come, again and again, over the years. Biraben and Delmon went on to do new experiments. They modified either the time after grafting that the radiation was applied, or the length of its duration. This time their efforts were crowned with unequivocal and complete success. The tumors stopped growing and, when still living cells were excised from them and implanted in healthy control animals, none of them became malignant.

Though these results should normally have fascinated any academy of medicine or sciences, the two researchers did not publish them. Why? The reasons horrify or disgust. It seemed that Biraben was simultaneously preparing an examination for the *agregation*, the highest French academic degree leading to a senior university teaching post. In charge of the committee to pass on, and award, this degree was none other than that same Professor Lachapele who told him: "Either you get the degree, necessary to your professional advancement, or you publish your research paper. But not both!" Discouraged, Biraben ceded to this demand but nevertheless continued to work on the research that looked so exciting and promising.

Most mystifying to him was how the machine operated to achieve its startling results. At the 3rd Congress of Biometerology held in 1963 in the Pyrennees mountains, a New York City researcher by the name of Kenneth McLean reported he had been able to obtain regressions on tumors and improve the health of cancer patients by using a magnetic field of a strength of 3000 gauss or more. Acting on this hint, Biraben and Delmon made an electromagnet that put out a field of 4,500 gauss and tried it out on the T-8 tumors but without the slightest success. Obviously, something other than a simple magnetic field was at issue.

In 1966, after others had had the same success with the T-8 tumor by irradiating it with the "Priore Ray," the two scientists finally published a memoir in the Revue of Comparative Pathology in which they stated that neither magnetic fields nor X-rays had any effect on the T-8s and that "only certain devices associating a magnetic field with high frequency waves seem at present to reveal therapeutic properties..."

Their conclusions were too late for, by that time, a campaign to stamp out Priore and his electromagnetic approach to cancer cure was well underway, a campaign that has lasted right up to the present moment.

The all-powerful Lachapele had sealed the fate of the Priore device as far as the local Bordeaux medical community was concorned. Veterinarian Berlureau and Priore next decided to carry their case to Paris. They contacted Professor Guerin at the cancer Institute at Villejuif, the leading French center for cancer research nnd the equivalent of the American National Cancer Institute in Bethesda, Maryland. Guerin, one of the discoverers of the T-8 tumor, which for the first time had been stopped in its tracks by the Priore device, courteously received his guests and heard them relate the whole story of how the device had come into being, starting at the point. when Priore had seen the oranges strangely preserved by some unaccountable electromagnetic effect.

Guerin was sufficiently impressed that he assigned his colleague, Marcel-Rene Riviere, to delve into the whole question. For two years, Riviere, who also had teaching responsibilities at the University of Rennes in Brittany, unremittingly worked to corroborate the Biraben-Delmon findings. On 9 December 1964 a note was sent for publication in the **Proceedings** of the French Academy of Sciences detailing the research and modestly concluding: "... as of now, one may already state that our first observations show that electromagnetic fields used can lead to most interesting data from a point of view of the biological behavior of grafts and their therapeutic action on experimental tumors."

Riviere next decided to see if the Priore Ray could affect another tumoral form that had never been affected by any therapeutic method, the 347 lymphoblastic lymphosarcoma. The results were even more spectacular than for the T-8 tumor. The effects produced were of broader scope and took place more rapidly. A second note was sent to the Academy for publication. The conclusion read: "We can now already affirm that our research offers proof that electromagnetic fields are capable of producing effects on quite different types of neoplasms."

At this point one of the key characters in this extraordinary drama must be introduced. There might have been no drama at all without his appearance on stage. This personage was Robert Courrier, an eminent endocrinologist, who had been named, while still in his 30s, a full professor. Courrier was now perpetual secretary of the Academy of Sciences and later would become President of the Academy of Medicine. Because no scientific paper can be accepted by an academy unless introduced by one of its members, Riviere would have had no chance to see his work so prestigiously published had not Courrier, who knew Riviere well, since he had shepherded him through the winning of his doctorate, taken the responsibility for its introduction.

It was Courrier who, at this point, also took up the cudgel to interest various highly placed French organizations responsible for the administration of scientific projects and their funding. Thus, ho sent a personal letter to the French Minister for Scientific Atomic and Space Research, who immediately offered to try to make funds available for further research on and with the Priore Ray. He also personally asked the Director of the CNRS (National Center for Scientific Research, which coordinates and oversees all such activity in France) to receive Priore and Riviere.

That this meeting was, in its way, somewhat of a disaster, can be explained in part only by a brief resume of the complex character of Priore himself. Priore throughout his life had great difficulty making himself understood in the French language and, as the **years** went by, he even forgot how adequately to speak his own mother tongue, Italian. Added to these twin impediments was his lifelong fear that his discoveries and inventions were prone to being stolen, a fear which led him never to fully explain the exact nature of the complex radiation emitted from his machine, far less the settings which controlled its various parameters. Whether Priore would not, or could not, exactly explain the functioning of his invention (which, as we shall see, went through several increasingly complex generations) is a question to which no precise answer has been given. It would appear that Priore was an excellent engineer gifted more with a God-given intuition than with school-book reasoning and logic. In short, Priore had a combination of talents that could remind one of the same enigmatic personality that was Nikola Tesla, the deductive reasoning behind some of whose discoveries has never fully been unravelled.

Highly placed scientific administrators are neither comfortable with, nor sympathetic to, what they see as self-appointed geniuses who have not run the same academic gauntlet through which they themselves had to pass. Thus the CNRS director took aversion to Priore's somewhat incomprehensible, yet fairly prolix explanations of his technology and only recommended that a physicist be sent to look over his device to properly decipher its working. At the same time, the Minister, together with the head of the general delegation for Scientific Research (DGRST) — still another key body in the administration of the French scientific decision-making process — let it be known to Robert Courrier that they had not understood a tingle thing about Priore's invention despite his best efforts to present it.

Advancing one more step into what was to become for him a 20 year-long expedition into a jungle of scientific intrigue, Courrier next resolutely decided to send to Bordeaux one of his most trusted laboratory workers, Madame Colonge, to repeat Riviere's experiments under her personal supervision. The DGRST director fully Concurred with Courrier's decision, while letting slip his admission that he strongly suspected that Riviere might well have been duped in some manner by Priore. When Courrier asked the minister for travel funds for Madame Colonge, he was refused with the dry remark that such a request was "premature." The now angry Courrier telephoned to reply: "You've been spending millions for programs and hypotheses about the cancer problem. Riviere has been presenting you with facts!"

The physicist who had been sent to try to elucidate the functioning principles of Priore's device, reported that he could make neither "head nor tail" of the machine.

Madame Colonge's experimentation was as prolonged as necessary. She was so meticulous that, in order not to take her eyes off the experimental animals for an instant, she limited her lunches to sandwiches eaten in the laboratory. She returned, profoundly impressed, to Paris.

Nor was Riviere idle during this period. He decided to experiment with the L-52 lymphosarcoma, a tumor similar to, but even more malignant than the 347. This time, he used not rats, but mice, as hosts for the grafts. So successful were his results that, this time, Courrier decided to bar no holds. Instead of simply having a note published in the Academy Proceedings on the quiet, he decided to present it personally, orally, and in all solemnity, before his fellow academicians at an official meeting scheduled for 1 May 1965.

That date was, and is, a turning point in what came to be known in France as the "Priore Affair." From then on the whole French, and even the international, scientific community could be divided into a minority and a majority group, the first that believed in the research, the second that did not know enough details about it, did not or would not believe, or simply didn't give a damn.

Before Courrier could make his presentation, its scheduling and subject were inadvertently and prematurely leaked to the press. Immediately thereafter, a horde ofjournalists arrived in Bordeaux. After one or two of them managed to all but force entry into his lab, Priore closed its doors and, with the help of his friends, wrote a printed press release that stressed his thankfulness to the many people who had helped him over the years rather than providing any comprehensible details about the machine he had brought to birth. Stymied, many of the newsmen travelled across town to seek an explanation from its leading cancerologist, Professor Lachapele, who informed them acidly that the machine was all but useless and unworthy of their attention or their time.

The journalists' reports, founded as they were on rumor of outright lies, roiled the pages of their newspapers and magazines in such a way that they either over-exaggerated the potential or a forthcoming cancer cure, or came close to billing Priore as just one more cancer-cure charlatan. All of which so alarmed, among many others, Dr. Wilhelm Bernhard, world specialist in electron microscopy, that he called his friend Courrier to warn him that his forthcoming presentation to the academy might put his hard-won reputation at risk as well as those of Guerin, Riviere and the Villejuif cancer institute itself.

In the journalistic melee, no one had bothered to read the declaration carefully issued by Riviere from Rennes, where he was occupied with his university courses, which formally stated: "Our experiments are of real interest. Much more experimental research has of course to be done before any therapeutic application on human cancer victims can even be considered. It goes without saying that this will take a certain time and no little effort of many types, both scientific and financial. Our work, as fascinating a perspective as it might hold, in no way, therefore, allows anyone to offer the public hope which could only lead to deception at the present time."

On May Day, Courrier gave his report to an Academy assembly hall crowded with scientists, newspapers and television reporters, photographers and an unusual number of curious bystanders. Accompanied with slides showing histological details, and animals before and after treatment, his lecture was heard out in almost tomb-like silence. When the lights came on again, he announced that he had personally checked the validity of Riviere's findings through the offices of his personal assistant, Madame Colonge. He then went on to say that he had taken the responsibility of presenting three notes to the Academy for two reasons. The first was a ringing declaration and a challenge to critics and skeptics of every stamp. It reads "When it is a question of a problem as serious as that of cancer and when one sees a little light beginning to dawn, one has the obligation to see what this light might represent. One has no right to snuff it out before learning what it may be worth."

The second was a tribute to his colleagues who had done the pioneering work, particularly Guerin, Riviere and Madame Colonge, and a statement of the essence of the problem to be faced down the road. It read: "Attention must naturally be given to the apparatus which Monsieur Priore has conceived and constructed. It has apparently already been examined by several physicists. It is found to be too complicated. While that may be possible, it is nevertheless a fact that Guerin and Riviere have obtained results with it that had to be made public. What is emitted from such an apparatus? I hope Monsieur Priore will allow disinterested physicists to study it at their leisure, for Science cannot tolerate apparatuses enveloped in mystery." Then, as almost a footnote to the history of the moment: "The biological action of magnetic fields is the object of intense research in the United States. In specialized institutes, the influence of these fields on tissue cultures, microbes, plant forms, diastases and certain tumors is under study. Up to now, the results obtained on grafted tumors seem less significant than those which have here been presented."

After Courrier sat down, a leading cancer specialist, Professor Lacassagne, rose to ask snidely why the notes had included no bibliographical references on work done on the bio-effects on tumors from electromagnetic fields, and criticized the experimentation as "impromptu." When Courrier denied this allegation as ludicrous, Lacassagne stalked out of the assembly hall in full view of the audience.

The meeting caused a new eruption of media reports which unfortunately accented one of three aspects of the problem at the expense of the other two. These were 1) the hope that a miraculous cure for cancer was in the offing 2) the contradictory, not to say discordant, reception of the data by various academics and 3) the enigma of Priore's personality.

Remarkably, no serious discussion among the scientists present at the meeting ever took place. This led a foreign scientist, present in Paris at the time, to remark: *"I don't understand. Here is a report given to the most authoritative scientific body in France by one of the most respected and eminent of its members and it is publicly subject to doubt without that leading to any reaction whatsoever."*

The General Delegation for Scientific Research was at this point still open to the idea of providing funds for more research with the Priore Ray. The big stumbling block, however, was one related to niceties involved in relations up and down the scientific hierarchy. To open the way to the allocation of such funds required the approval of the Delegation's own section for cancer research and that section was headed by none other than the same Professor Lacassagne who had so rudely walked out of the Academy's assembly hall.

At the same time the General Delegate diplomatically covered himself by suggesting to Robert Courrier that he had to have more information underscoring the potential importance of the research accomplished. Courrier told him to simply re-read the three notes he had presented to the academy. His matter-of-fact, yet terse, come back then elicited his invitation to a full-dress meeting of scientific experts at the Institute for Scientific Cancer Research.

In this short historical account we obviously cannot go into the ins and outs of what transpired at this meeting or any of the many similar meetings which followed it. A paragraph in a brilliant book, four years in the writing, by the courageous Bordeaux journalist, Jean-Michel Graille and entitled: **Dossier Priore, A New Pasteur Affair** (of which this account is but a tenuous synopsis) must suffice to pointedly characterize the nature of the problem in its most general sense.

Writes Graille: "To read what follows in this chapter might well seem tedious: an enumeration of names and titles, the content of a debate held at an administrative meeting, personal remarks by one ranking personage or another, exchanges of letters following the meeting itself and the official report which came out of it. Tedious but indispensible for not a few reasons. It is important to know who were the participants at this meeting and what of these participants each was trying to represent. It is important to know how such scientific meetings go about their business at the 'top level,' And, finally, it is important to learn about and to understand, in the particular case of the Priore Affair, the behavior and reactions of all concerned. A reading of all this could be difficult (and it won't be the only such passage in this book) yet it is necessary to understand the essence of the dossier in order to be able to create for oneself as clear and well-motivated a personal opinion about it as to subsequently be able to discuss it, or to hear it discussed, with a thorough knowledge of the facts."

In this single paragraph, Jean-Michel Graille has, in my opinion, pointed to both the nub and the difficulty in getting at the essence of the real facts behind a case such as that of Priore's that are so important to its proper understanding, an understanding which can be painted against the backdrop of the history of science and the backdrop of human pettiness and maliciousness or human courage and magnanimity.

One of the participants, Professor Andre Lwoff, soon to become n Nobel Laureate for his work in virology, was violently against the meeting's central issue: namely, whether or not funds should be spent to build a new and better Priore machine. Not only did Lwoff aver that the three notes presented to the Academy never should have been published, but he also opined that since all the work was done, not on spontaneously generated but on grafted cancers, the effects of the machine were hardly impressive. He later added in writing that 1) the patents issued to Priore for his device were nothing but a web of nonsense, 2) the machine itself could never be duplicated based on any description given for it by its inventor and, in a repetition of his oral remarks, 3) the fact that only cancer grafts were experimented with was nothing to shout about. He strangely added that because the animals who had been irradiated subsequently were able to entirely reject new grafts, the whole phenomenon offered no proof that cancer cells could be killed while healthy cells were not. The whole thing came down to a question of immunity, he said, as if that were not of the greatest possible importance.

To which, in due course, Guerin and Riviere replied: "It has been claimed that our experiments are valueless because they were carried out on grafted tumors and that other therapeutic measures were known to get rid of such tumors and their metastases. *We defy those persons who have made such affirmations* to prove, *with the use of such other measures, that animals infected with T-8 tumors can be cured at a percentage rale identical to those obtained by using the device which Monsieur Priore has developed."*

Not a soul has responded to this challenge, then or since.

A second cancer expert at the meeting, a woman of great influence, resorted only to the cavil that the experiments had been of doubtful quality since none of the animals had been *weighed.* The fact that those same animals had survived normally lethal cancers seemed not to have *weighed* with her.

There were many more observations of the same ilk. They seem atrociously paltry, trifling and picayune coming from professionals who, if they no longer believed in the Hippocratic oath to which they once swore, are considered by the public in general, and by cancer patients in particular, at least to be concerned with seeing what a little light on the problem might reveal before extinguishing it, as Professor Courrier expressed it.

At the same time, we must not forget Priore's decidedly difficult personality. He was an inventor determined at all costs that his invention be developed for the benefit of humanity, yet anxious that that same humanity not steal it from him. As author Graille puts it, "His conceptions and attitude directly or indirectly conditioned the overall essence of this affair. Full of enthusiasm, from the very day he discovered that the ray he had developed had a curative effect on a cancerous cat, he developed a single-minded fixation on cancer One could understand and sympathize with him on this score. Here he is, a little Italian immigrant without money or means, and he is going to offer the world a cancer cure. He is so convinced that he wants to move ahead to doing just that. He will never understand or accept the exigencies of Science or Medicine. For him, experiments, controls, verifications and parallel research are a waste of precious time. 'I've made machines which cure cancer. Take them and treat cancer patients. Don't bother with the rest.' Such would be a summation of his point of view."

Through the efforts of persons kindly disposed to the inventor, this point of view was softened and he came finally to understand the-necessity for what has been called scientific rigor, on the other hand. another aspect of his character never changed an iota. This was his determination to preserve the secret of his invention, motivated first of all by his unshakeable desire that it be developed in Bordeaux, the city of his adoption, for the citizens of that city. Deeply rooted was his belief that if he made his secret public, the machine would be taken from the Bordeaux region and further developed by Parisians, those who considered themselves to be in the penthouse of the scientific edifice. Once this was accomplished, he would likely not have one more word to say about the matter. Therefore he continued jealously to conserve his secret and put confidence in nobody.

As Graille generously concedes, he may well have been right. and adds: "All his life he had to go up against men, whether scientists or industrialists, who had but one idea in their heads: to get to the bottom of the inventor's secret in order to build for themselves a machine which they then could exploit for their own account, for their own glory. Many such 'Priore Machines' were to be actually built more or less surreptitiously or clandestinely. Not one of them ever worked."

While one might easily accuse Priore of a limited view, the horizons of the researchers themselves were certainly not as broad us they might have been. Those involved in bio-medicine were content with the results produced by the machine, the workings of which were of no concern to them. A black box, as it were, emitted a ray that definitely affected experimental animals. At the same time, as researchers specifically interested in the cancer problem, they never gave a thought to what the Priore Ray might accomplish in the wider clinical domain of other afflictions.

As for the physicists, they were seemingly not up to the task of comprehending a complex radiation that had miraculously sprung, as from the head of Zeus, out of the intuition of a man they considered to be an undereducated and all but illiterate gadgeteer. Still others, whether physicists, biologists, doctors of medicine or specialists in a dozen other fields, were willing to throw the baby out even before it went into the bath water. In their eyes Priore was just a nobody.

Behind the scenes, many of these scientists resorted to using the press to achieve their own ends. Thus, the chief medical chronicler for Le Monde (the French New York Times), herself a doctor of medicine, was led to write outright lies about the Priore Affair—specifically and falsely stating that cancer patients had been treated with the Priore Ray in the clinic of Professor Lachapele in Bordeaux with not only negative, but disastrous, results.

On the other hand, a journalist for another leading Paris daily, Le Figaro, scrupulously conscious of his responsibility to fairly report what was going on, aptly wrote: "We would like to see at least one thorny point clarified as soon as possible. Several years ago P'rofessor Biraben of the Bordeaux Medical Faculty (who at that time had not become a ranking professor) was involved with the Priore device. According to certain reports from medical circles, his results seemed, even at that time, to have been already quite positive on small animals and he seems to have written a report to that effect, He was advised by highly placed authorities "to keep quiet" and stop talking about this affair. If this turns out to be true, it would be a veritable medical scandal to be judged in the harshest terms."

Could one have put it more succinctly?

The foregoing is to present something of the flavor of what was transpiring in the wide world far removed from the laboratory of

Priore who, at the time unaware of it, was revelling in the fact that his machine had been successfully used by high-ranking French cancerologists and its results reported in three separate notes to the Academy of Sciences.

His courage was also more than buoyed by the arrival on the scene of the commercial director of a large French industrial firm specializing in the intricacies of manufacturing glass components. This man had heard that Priore needed a large tube that was beyond all existing norms and perhaps did not exist anywhere in the world. This tube, it can be stated, contained a rare gas, neon, which when excited into a plasma, seemed somehow to convert the various electromagnetic inputs into a single Priore Ray which surged from the business end of the tube. In the tube were an anode and cathode. Peculiar to the anode was that it had to *rotate* to produce the desired biological effects and this is but one of the anomalies in Priore's equipment which physicists and bio-physicists have to this day been unable to explain.

The manufacturing company, a subsidiary of the internationally known company, Saint Gobain, was looking for a new product. The commercial director thought the new tube might fill the bill, particularly if it could be adapted to a machine that might ultimately cure cancer, a product that indubitably would have an enormous market across the world. There were plenty of problems with regard to the tube, notably those of its large dimension, its resistance and its conductability. When the tube was finally made, it now seemed that. Priore would have to explain his discoveries to the scientists of the company that had made it. One of these was sent to elicit such an explanation but was, so to speak, "shot down in flames" by Priore. So a second attempt was made by Ivan Peyches, a senior executive of the company, and president of the Society of Civil Engineers of France, who made a detailed investigation of the device. His reports were subsequently lost, but there remained an article he published in a leading French journal, Sciences and Technics, a short time before his death in 1978. It bore the intriguing title: "What Are So-Called Paranormal Phenomena?"

In it the engineer wrote: "There was so great an accumulation of components capable of having some kind of action, and being unable to work separately, that the results of measurement were limited to proving that there were no specific rays that issued from the tube (Priore talked about canal rays), no more than there were any X **rays.** On the other hand one could detect a magnetic field which was the end result of a field proper to the tube and of the magnetic field of a solenoid that constituted the experimental chamber, an electromagnetic field with a frequency of 16 megacycles (19 meters) and a high frequency field (metric waves), the whole being pulsed at a very low frequency of an order of one per second. It was impossible, in such an imbroglio, to determine what was necessary and what was sufficient. Priore maintained that the simultaneous action of his various generators was indispensible to achieving his effect."

Peyches then went on to relate how he tried to persuade Priore to offer a more precise definition of his thinking about the workings of his device. He wrote the inventor: "At this point, I would say that all reticence on your part, which in your eyes would be justified by the fear of seeing yourself partially dispossessed of your work, would be of far greater detriment to you than any safeguard of your interests. Moreover, since it has become a question involved in public health, you are no longer entirely your own boss...you absolutely must bring all this to the clear light of day and I don't believe you can do it alone...You must supply all the characteristics so that third parties can reproduce your results." Then, he concluded by citing the words of an academician: "Many phenomena are rejected by the scientific world because they are considered irrational: But it is not B proof of scientific honesty to refuse a priori to try, out of homage to truth, to have a look at them and perhaps to understand them. Will Science one day be able to abandon its taboos?"

It was Peyches' final conclusion that, in the end, Priore was a man of genius who knew absolutely nothing about what occurred in his machine from the scientific point of view. The company which he represented no longer exists since it was bought out by the American firm of Corning Glass.

Industrial interest in the Priore device was not limited to the Saint Gobain subsidiary. Next into the lists was a company in Anguouleme, Leroy-Somer, which specialized in electric motors, generators and later was to branch out into solar power. Its president, Georges Chavanes, took the initiative to write to Priore in 1965 that his company was interested in providing some of the complex electrical equipment needed by the inventor, more particularly high-powered generators, on the condition that Priore move his operation to the company seat at Anguouleme. When the inventor categorically refused, Chavanes tentatively agreed to build a factory to manufacture the Priore device in Bordeaux itself.

The alliance between Leroy-Somer and Priore, shaky at best, lasted two years and blew up on Holy Thursday of 1967. The period was a stormy one for both parties to the agreement. Priore did his best to convince Chavanes to commit himself to building a huge machine with a magnetic gauss strength of 10,000 gauss. In the end he got one that put out only 920 gauss, not much stronger than the machine he had already built which put out 620 gauss. Since the field of action increased with the gauss strength, Priore reasoned that a machine of literally behemoth size would be able to irradiate the whole, or every part, of a human cancer victim lying on a stretcher, whereas the smaller machines had been effective only for small animals or for treating a limited portion of the human body.

Chavanes and his company were aware that it would be a tremendous financial burden to contemplate building the larger machine. So they went ahead with plans for the smaller one while at, the same time putting great pressure on Priore himself to make him feel that he was the least important cog in a new gear, in fact that his status was reduced to being a simple employee of Leroy-Somer. In Graille's estimation, this lack of psychological finesse on Chavanes' part constituted what he called "the blackest pages in the Dossier Priore."

Even the smaller machine was to cost about half a million dollars, a price which today, due to inflation, could be tripled or quadrupled. During a stage in which an intermediate machine was designed by the chief Leroy-Somer engineer, Ribeau, a machine that never did function properly, Chavanes all but forced Priore, who was heavily in debt, to sign a contract which was falsified. The falsification was a matter of one word which was changed in the contract. In a phrase reading that an exclusive license of patents, and subsequent patent modifications, would accrue to the company "for all countries solely for therapy on cancers concerning animals and humans," a word was inserted by hand so that the phrase read. *"concerning particularly* animals and humans" implying that other uses of the machine, whatever they might turn out to be, would also accrue to the company. This one word change was amended on Priore's copy of the contract by calling the word "particularly" a "nullified word," but on Chavanes' copy it was called an "added word."

Leroy-Somer believed it was sufficiently well positioned in the driver's seat to be able to deal on behalf of Priore himself with the French governmental institutions, mainly the General Delegation, concerned with the funding of the new machine. When Priore learned of Chavanes' contact with the General Delegation he wrote a letter informing it that no one had the right to deal in his name. Nor did Chavanes even attempt to cut the Saint Gobain subsidiary, which alone could supply the tube, key to the device's functioning, in on the government funding.

In the meantime, no less a figure than Professor Kastler, soon to win the Nobel Prize in Physics, came down from Paris to inspect. the existing Priore device. He brought with him Delmon, who, we recall, had worked with Berlureau on the first animal experiments and who

NOW, it turned out, was trying to build his own version of a Priore device on the sly without telling Priore. Kastler's bringing Delmon with him to Priore's lab so angered Robert Courrier that he told the physicist he had committed a real gaffe. He also convinced Kastler that Leroy-Somer should build a machine with a power of at least 5000 gauss, but Chavanes refused. There seemed to be no harmony of outlook between the leading industrialist concerned, on the one hand, and the top physical and biological scientists on the other.

While all these, and many more, peripatetics were proceeding, Priore's sister in Italy came down with cancer. Beside himself with **grief,** Priore informed all concerned to commit themselves either to building an intermediate machine correctly, under his supervision or, better, the 5000-gauss machine, and to do this in time to save his sister, or he would wash his hands of the entire matter. Confronted with this ultimatum, the company began to work round the clock to perfect the intermediate machine but engineers involved, believing themselves to be more adroit with respect to its design than Priore, left out a host of what, to them, were unnecessary components. The result was that when the machine was first put to trial, most of the components burned out or otherwise failed, and the machine itself became a useless pile of rubble.

Shortly thereafter, Priore's sister died of cancer. Her grief-s**tricken** brother went into what amounted to total isolation, unwilling to talk to a soul.

The whole Priore affair might have ended at that point, in the early part of 1967, were it not for the entry onto the scene of a key **figure,** Professor Raymond Pautrizel. Born on 3 June 1916 in Basseterre, capital of the French Carribean island of Guadeloupe, at forty years of age he was on the Faculty of Medicine at Bordeaux. He so**on** became known, world-wide, as the "father of parasitological immunity," a title he never accepted, saying that, if others had awarded it, **it** was simply because "he had searched through old scientific publications to find ideas that were as valid for modern research as they were forgotten by modern researchers." And he later was q**uoted** as adding: "It is really too bad that researchers today don't pause from time to tome to dig into studies made by their predecessors, some of which were performed even decades ago!"

Professor Pautrizel was awarded the first academic chair in **France** for immunology, and later a special unit was created for him for parasitological immunology, a subject which is both simple and complex. The simple part involves the fact that various immunological techniques can be applied to diagnosing specific parasites that have invaded an organism in order to develop preventative actions a**gainst** them via vaccines, or curative actions via serums.

When invaded by parasites, organisms react by creating antibodies, specific substances aimed at killing the invaders also known as antigens. These antibodies are liberated, like an attacking army, into the blood. Simple enough so far. The complexity arises because the defending army, the parasites, don't just lie down and die under the attack. They are capable of modifying their "personalities," as it were, and of changing various of their characteristics such that the mechanisms that the host uses to recognize, or detect, the invaders are invalidated. Thus, the substance which an organism would secrete to destroy an invader A becomes incapable of recognizing A, now become A-l, and therefore incapable of destroying it.

The organism at this point seems to realize it has to create a different substance to rid itself of its antagonist but, in the meantime, the metamorphized parasite is getting on with its assigned destructive task. Alternatively, the parasite has another capability: that of itself liberating substances which can annul or annihilate the organism's overall defense system. A sort of "in the blood" version of Star Wars is going on at the microscopic level.

The study and classification of the substances - - call them weapons - - emitted by parasite-attacked organisms allows for the establishment, in turn, of batteries of tests to define the exact nature of the parasites themselves in order to come up with an appropriate therapy or counter-weapon.

This then, is the essence of parisitological immunology, Raymond Pautrizel's area of research. He specialized on a particularly lethal parasite known as trypanosome, the scourge of tropical third-world countries where, in one form, it causes sleeping sickness in animals and humans, in another, equine syphillis, in still othera, other afflications. Over the years, during which he produced a small library of literature on the problem (known mostly to specialists in countries where that problem is acute), Pautrizel and his team discovered, among other things, that the trypanosome can modify itself, again and again, up to 101 times over a period as short as only three weeks.

Even before his work on trypanosomes, Pautrizel, back in 1949, was one of the first researchers to discover what is known as *ambivalence* in drugs, notably histamine. Histamine is a substance which is secreted by an organism as a defense mechanism but if oversecreted by certain cells circulating in the blood, it becomes virulently noxious, mainly by overdialating blood vessels, thus making them permeable to water and leading to edema and even death. This process occurs, for instance, in some human beings who are highly susceptible and over-reactive to bee or wasp stings.

Pautrizel's research on the noxious aspects of histamine led to his finding that the same substance, applied in requisite small doses, is extremely important to the defense system of the organism. Today he stresses the notion of *ambivalence* in many areas of his work and characterizes it as "a key to the biology of the day after tomorrow."

To finish with the background on Pautrizel, before bringing him on stage in the Priore drama, it may be added that only a few years ago, at a formal reception for him attended by the medical elite of France, he was given a Basque *makila,* an iron-bound honorific cane of sculptured wood, in tribute to his work. On it was the incised inscription: "Sometimes to heal, often to alleviate, always to console," an epithet that perfectly characterizes a medical doctor imbued with that kind of rare compassion that marked Pautrizel's character.

When Robert Courrier sent Madame Colonge to Bordeaux, it was Pautrizel whom he asked to provide her with every assistance. In this way, Pautrizel was first introduced to Priore and his device. After witnessing the results obtained with it he was to say: "What stupified me, and led me to ponder the question, was to see the control animals die from their tumors in 3 weeks, while at the same time I could observe that the tumors in the animals under treatment were literally melting away and the same animals were taken back to Courrier's lab at the College de France in Paris in perfect health." ASa result of his thinking about the problem, Pautrizel came to the belief that the machine, however it worked, did not exert any action at all to kill cancer cells but, through as yet unexplained mechanisms, stimulated the afflicted organisms to provide themselves with new immunological weapons that could overpower the cancer cells.

To shed light on this problem, Pautrizel proposed the simple expedient of experimenting, not on cancer-infested animals, but on *in vitro* cultures of cancer cells. He made this proposition to both French and British cancerologists but they were convinced that the Priore device *had to be actually killing cancer cells themselves,.* They could not see the point that, if the machine did not kill cancer cells, then it was doing something else to the body to allow it, and not the machine, to do that job.

Pautrizel's involvement with the British was the result of a team being sent from England to experiment with cancer mice with the Priore Ray. What happened cannot be related in this brief resume except to say that, out of a lack of understanding on the part of certain British cancer experts and malicious conniving on the part of one member of the cancer "aristocracy" in Paris, the experiments were put under a cloud. It was alleged that mice had been substituted somewhere during their long round-trip voyage between

England and Bordeaux to make it look as if a failed experiment had been successful. This did not prevent Sir Alexander Haddow, chief ofthe prestigious Chester Beatty Research Institute for Cancer from stating, at a meeting in Paris, that the Priore machine had been indubitably effective on the English mice and supporting Pautrizel's idea that experiments should forthwith be done to see if the Priore Ray had any effect on cancer cells *in vitro.* Haddow's suggestion backing Pautrizel's recommendation fell on deaf ears.

Because of the emotional turmoil and rancor with respect to cancer that had so long surrounded Priore and the workings of his machine, Pautrizel suggested that it be tried in a completely new area, one he knew so well, namely on afflictions caused by the trypanosomic pathogen. Before these could get underway, however, someone had to persuade the still desolate Priore to return to work. Pautrizel, known to those really concerned with and knowledgeable about the potential of the Italian's invention, at last was able to convince the inventor to cooperate and get back into harness. This ho did with that rare combination of diplomatic tact and warm human sympathy with which only the Pautrizels of this world are gifted.

In the meantime, Riviere had gone on to implant new 347 tumor grafts in rats previously cured of 347 tumors. When none of the tumors developed, that result added one more argument to back Pautrizel's idea that the machine was, in fact, affecting the immunological defense system of the animals. However, when Riviere tried the same procedure with the T-8 tumors, his animals died. This lead to the conclusion that the immunity acquired by the animals to lymphoblastic lymphosarcoma 347 was specific to that tumor. When a note on this research was sent, again through Courrier's good offices, to the Academy, for the first time, it strangely omitted from the listing of the participating researchers the name of Antoine Priore. It seemed that Riviere had been taken to task by fellow cancerologists who believed that Priore was nothing but a naive bumpkin or, worse, a swindler. They had warned him against publishing any papers with which Priore's name would be associated This rank injustice and lack of fair play again sent Priore into a fit of despondency and depression from which he could only be wilth drawn by those subtleties involved in Pautrizel's sympathetic and friendly counsel.

On 25 July 1966, another note was sent to the Academy filed for the first time not under the rubric Cancerology but under the rubric Immunology. It was entitled "Influence of Associated Electromagnetic and Magnetic Fields on the Immunity of Mice Infected with Trypanosoma equiperdum." The conclusion read: "The treatment allows the organism to rid itself of parasites even when these have invaded it in a most intensive way.... There is an enhancement of both the specific and aspecific factors of immunity."

Thus, for the first time, the field of research shifted from the narrower field of cancer to the much vaster domain of immunology. And, for the first time, Pautrizel's name appeared as the senior author on the paper. It also appeared that, for the first time, there should no longer be any problem about experimenting with the I'riore machine. Such was not the case.

Still complicating the whole issue was the fact that Priore himself was using different setting to produce different varieties of radiation depending upon his own intuitive evaluation of the particular biological experiments being run with his machine. He would never reveal the nature of these settings.

At this point there appeared on the scene a new researcher who became Pautrizel's loyal ally, a young woman, Pierette Chateau-Reynaud Duprat. During her work in Paris, she had learned of the Priore controversy, and, against the stern advice of mentors senior by many years to her in the cancer hierarchy, she came to Bordeaux to meet Pautrizel and learn more about the research.

Her work, performed over many years, is too detailed for presentation here but it led to important conclusions. One was that the Priore Ray had no direct effect on the trypanosomes themselves but stimulated and reinforced the defense mechanism of the infested organisms, allowing them to reject the parasitical influence with an effect so durable that they were no longer subject to this influence even after treatment stopped.

Another conclusion was even more important and involved, in part, British research. It pertained to the effects of the machine on both *allografts* or those made between two different individuals of the same species, and *isografts,* or those made between two different individuals of the same genetic line having in common antigens that were characterized by what is called the same histocompatability. The conclusion was that not only was the rejection of *allografts acce*lerated by the Priore Ray but that *isografts* were also rejected. This meant, in sum, that the ray stimulated not only the *defense* mechanisms of the organism but also, and more importantly, its re*cognition* mechanisms. In the case of an isograft, this allowed the recognition of weak antigens that were not recognized in non-irradiated animals. In other terms, where at first the anti-aircraft batteries could not shoot down the aircraft because they could not see them.now they could shoot them down because they could see them. In immunological terms, the ray affected both humoral and cellural, both specific and aspecific, immunity.

Here we must return to the mystery of the settings on the device.

As a result of the new experimentation it seemed that, depending on those very settings, the active ray, complex as it was, could have either similar, totally different, or diametrically opposed effects. Thus it was not a question of a ray having universal effects — a kind of magic bullet capable of killing any target — but of multiple radiations which, due to the complexities in Priore's personal makeup, have unfortunately yet to be sorted out and explained.

Thus, the machine originally designed by Priore, called the P-l when it put out a wave length of from 19-21 meters, had a radical effect on certain animal cancers, on *cellular* defense mechanisms, and finally, but not universally, on organisms infested with **Trypanosoma equiperum,** (hereinafter called T.e.).

A second machine, dubbed the P-2, was at first not able to produce these frequencies. What it did put out was a frequency of 17 meters that *was* universally effective against T.e. and seemed to act. not on the cellular, but the *humoral,* defense mechanisms. The rejection of grafts depends on the cellular defense mechanisms, which partially explains why Pautrizel when using the P-2 machine, selected the T.e. vector, as it is called in microbiology, just because this creature is fought by the organism's humoral defense system.

Consequently, the bio-effects that were successfully attained depend on the varying, not to say quixotic, nature of the radiation At one point Pautrizel actually did experiments on animals infected with *plasmodia* — the vector for malaria which attacks red cells - and found that the settings used were effective while never learning exactly what they were or the exact nature of the radiation. Further more, Priore himself maintained that over the years he had success fully treated cases of human tuberculosis but, again, never revealed which frequencies had been used to achieve this.

Several more notes were sent to the academy on the successful work performed with the Priore Ray on animals affected with T.e. But the central issue remained: how to find out exactly how the- machine worked. It fell, not to civilian scientists, but to those in the- French army service to attempt, at this point, to work out the problem. The army service brought into the picture was the DRME (an acronym which translates as Administration for Research **and** Test Methods), to which Pautrizel had sent a request for funds in 1968.

This request was the subject of a meeting at which were present three of the top names in French science, one representing biology, the second physics and the third, medical physics. The latter two turned in extremely unfavorable reports recommending that no money be wasted on the problem. The biologist, however, turned in a most favorable report and, despite the fact that he was in the minority, his opinion won the day.

As remarkable as was this victory, it was even more stunning and incredible given the fact that this biologist was the same Andre Lwoff who had so adamantly opposed the Priore research a couple of years previously. Lwoff had summoned the courage to completely reverse himself only after he sent one of his most trusted colleagues to do secret experiments with the Priore Ray on mice injected with peroxydase (an antigenic solution) to see if they would produce a higher level of antibodies than non-irradiated animals. This they did so well that Lwoff became convinced that the Priore Ray caused an extremely important increase in immune reactions. These results were never published because, before the experiments could be repeated to be absolutely sure of their results, the machine suffered one of its many interminable breakdowns.

The DRME report was at length, and in length, issued but not publicly since it was protected by a military classification. However, a synthesis of it was finally published in November 1979 by Herbert Gossot, Secretary General for the French Association for Bioelectromagnetism, under the title: "A Scientific Balance Sheet on the Priore Ray." Its contents were as follows:

"The two physicists assigned by the army made a complete analysis of the electromagnetic radiations and magnetic fields activated by the Priore device. They thus determined the spectrum of frequencies which the device emitted. They showed particularly that frequencies in the visible light and infrared range had no biological effect; that there were no X-rays or Y-rays; and that the pulsed ultra-high frequency electromagnetic wave was modulated in amplitude to that of a high-frequency wave. They did a topographic survey of the respective intensities of the various magnetic and electromagnetic fields in the experimental plane of the device. In particular, they determined the spatial repartition in this plane of the density of the strength of the ultra-high frequency wave. They showed that its value was very weak and that it could not produce any kind of overall significant thermal effect imputable to the hyperfrequency ray.

Finally, and most importantly, by using what they had learned about these repartitions, they demonstrated a clear correlation between the biological effects obtained and the intensity of the hyperfrequency ray. What they actually observed was that, on the biological model used, i.e. experimental trypanosomosis of the mouse, there was a diminution of the rate of evolution of the parasitemia that was proportional to the strength of the hyperfrequency wave. To quote them: *'These experiments of correlation are of certain interest: they confirm, if there is still any need of so doing, the* *biological efficacy ofthis device."*

The two physicists, Bottreau and Berteaus, are still interested in rebuilding a Priore device with which additional biological research could go forward. At the same time they suggested to administrative bodies in French science the creation of a special laboratory for bioelectromagnetism to fund more work, a suggestion in which Professor Pautrizel concurred. No action was taken and their report was kept under wraps. In a note they presented to the Academy of Sciences on their investigation, they were not allowed to include the names ofthe laboratories where they worked: in the case of one, the CNRS Magnetic Laboratory at Bellevue near Paris, and ofthe other, the Laboratory of Ultra-Hertzian Optics and Talence near Bordeaux. Why? Because the directors of these laboratories did not want any mud in the Priore affair to be spattered on them.

The next experiment done by Pautrizel was on rabbits whose testicles had been so seriously affected by trypanosomes as to be almost entirely destroyed. After radiation the same testicles took on their normal histological appearance and the rabbits, able to procreate again, in no way abstained from their newly regained ability. **This implied the complete regeneration of an organ that had all but completely degenerated.**

Yetjournalists, who sought out truths about the Priore affair in Paris from high officials they believed would know best about what was going on, continued to be led astray. For example, an American scientific reporter, writing in the **Saturday Review** of **Science** in 1973 saw fit to state: "It is really a question of a mystical problem that has little to do with science." He was quoting Professor Bader, a man who for 15 years held top administrative posts in science that could have allowed him to back the Priore research with all the funding necessary to its accomplishment. At the time Graille's book came out, Bader issued a book of his own about the Priore affair which offers no real idea of what was involved. When I asked several people in France why Bader had written the book, they were unaware of Bader's inmost motivation.

Machinations continued to swirl about the case over the next several years. Behind-the-scenes intrigues, distorted accounts in the press, lethargic attitudes on the part of administrative officials who would not take responsibility to cut an increasingly tight Gordian knot, outright fear of various personalities to become too deeply implicated lest they lose their jobs - - all these, and more, continued their daily round in an atmosphere of "Business AS Usual," and "Don't Risk Your Neck."

To get to the nexus of the situation, we have but to cite the observation of one of the few perspicacious journalists who, in **the** prestigious scientific monthly, Sciences and Life, wrote: "The physicists are convinced that the effective Priore Ray is very complex but to analyze this further some things first have to be made clear. One is to raise the suspicion that has surrounded Monsieur Priore with a fabulous acretion of misunderstandings, insults and accusations of being a swindler over many years. What is needed is a veritable national effort to act effectively and to act rapidly."

Over the next two years the decision-making process of the French government lumbered its way along until it was finally decided to back the construction of a powerful machine. This decision was not favorably accepted in many quarters. As Le Monde would comment: "The decision was made in spite of the disapproval of many scientists. When money is tight, one should pay particular attention to how it is being spent. Such seems not always to be the case. A credit of some $3.5 million francs (or about a million dollars) has just been accorded to finance the construction of a new Priore machine."

The scientists to whom the article referred were in a rage. They understood, at this juncture, that the only way to put an end to the affair was to eliminate Pautrizel who, because of the very success he was having with his research, was seen as a dangerous competitor that might even become one of the top figures in medicine and science on a national, or perhaps, on a world scale. Indeed, it was learned that Professor Courrier had gone to the length of sending a report on Pautrizel's behalf to the Nobel Committee in 1979.

To make a long story short, the large powerful machine, the M-600, was built but a huge tube in it, after functioning for about a week, exploded. Due to the galloping inflation of the 1970's, to replace it would have cost another million dollars. The money was not forthcoming.

In the meantime Pautrizel, ever experimenting with the still functioning smaller machine, was to discover new facts. Mice with their spleens cut out, for example, also could survive injections of T.e. The Priore Ray had important implications for Arterio-sclerosis, since it effected lipid modifications in rabbits given a dietary regime high in cholesterol. This research, published in another note in the Academy Proceedings, instead of being warmly received, only irritated the cardiological fraternity which felt, as some of its members put it, "trapped" by Pautrizel's efforts.

One particularly virulent opponent was Professor Bricault, Dean of the Bordeaux Medical School who, as late as 1980, was telling his own students that the published results were a farce and had never been obtained. The students, who carried out a special investigation of the matter on their own, were able to judge what a farce their own medical dean might represent.

L'Express, the **Time** magazine of France, read by at least half the population of French intellectuals, had the gall to compare the results of the Priore research to those of the infamous Trofim Lysenko of the Russia of Stalin's day. Haughtily **L'Express** added: "Today Priore's defenders explain that his machine has not only cured cancer but, in all probability, altered the immunological characteristics of mice. Were this, in fact, so, all the immunologists, all the geneticists of the world would unite to affirm that a machine capable of changing the genetic patrimony is the discovery of the century, far more important than the atomic bomb or the conquest of the moon. Unfortunately, the history of the whole thing has never been properly elucidated." The article was illustrated with photos distortedly selected to convince viewers that the Priore machine was as serious and effective as the one that purportedly brought Frankenstein to life.

In this poisonous atmosphere the slow work of building the M-600 went forward. To give anyone who was not there a feeling for this endeavor we may now cite verbatim a passage from Graille's book: "The construction and assembly of the prototype —the M-600, that of highest power and variable parameters — were fraught with many uncertainties and delays on the one hand and, on the other, were marked by the stamp of Antoine Priore's sparkling genius.

"To go from an apparatus that developed 1,240 gauss applied over an effective area of some 20 centimeters, to one developing 5000 gauss over an area of 60 centimeters means to take on an extremely risky technical and technological wager. Electrical, mechanical and glass-blowing specialists plunged into the unknown. They had to conceive, make, adapt and put together all the various myriad components almost haphazardly with no precise technical study being previously available. Priore's stubbornness forced them to take on a trial-and-error manufacturing "gimmickry" without precedent. As the thing was put together and preliminary tests made, it became clear that many of the components were unsuitable and that they would have to be modified or replaced. The tube itself, made of pyrex, 60 centimeters in diameter, and 6 meters tall, had to be replaced twice after it imploded. In fact, practically everything had to be reconsidered or readapted. "Everything" meant the parts going to make up a generator of 50 tons in weight. For example, the coil which created the magnetic field: 5.5 tons with 11 miles of copper wire. For example, the numerous cooling circuits which stabilized the thermal equilibrium of the generator and its environment or, additionally, the circuits governing command, control regulation and selection — 6 tons of electrical cables of which 15 miles were of tele-command wiring.

"Priore astonished everyone. Breakdown after breakdown, incident after incident, it was he alone who showed what to do next, indicated the proper steps to take, the right settings to adopt, the right way to assemble the components: He was virtually building his machine by himself, nursing its construction along day after day, all the engineers' studies and efforts actually, and ultimately, serving only as a preliminary attempt, a sketch as it were. When Priore made his presence felt, things began working."

Then.... after the machine was built: "The part of the entire apparatus to generate electricity was set up on a provisional basis. It was so noisy that, while functioning, it woke up the whole neighborhood. The number of experiments had therefore to be curtailed so that the machine would not be used at night. And, all at once, everything came to a halt. The Faraday cage, shielding and isolating Priore's apparatus, was torn and fissured by the shock of the cement pilings that were being sunk into the ground all around to hold up the building under construction. This allowed high-frequency waves to escape which disturbed radio broadcasts emitted by local radio stations, the army, and civilian aircraft for miles around."

Nevertheless during the week or ten days that the machine was in good operation the results of experiments performed with it were more than formidable. First of all, it allowed for as many as forty experiments to be performed on some 280 animals in a remarkably short period of time. Among the discoveries made were: The ray emitted provided the treated animals with an extremely strong immunitary response. Animals whose immune defenses had been attentuated by an immuno-depressant were able to overcome the effects of injected parasites but relapsed a few days later. One could therefore conclude their immune response was much weaker than those normally infested and treated.

Newborn animals, whether treated or not, developed a marked parasitemia leading to their deaths. At the time of death, the parasites had the same antigenic structure as those ofthe innoculum which thus implied that they had met with no defense at all in the infected organisms. This also proved that the Priore Ray did not act directly on the parasites themselves but only by way of an increase in the immune defense system of the organisms. The newborn animals succumbed to their parasitemia because their immune sytem was not yet sufficiently developed to be stimulated by the P-Ray. The phenomenon of a stimulation ofthe immune defenses was demonstrated by the fact that animals which had received soluble antigens developed, after being irradiated, a level of antibodies far superior to the controls.

These and other conclusions were the object of notes presented to the Academy of Sciences by Pautrizel and his team in 1978. Even before, at a colloquium held in Antwerp, Belgium devoted to African human trypanosomiasis, the same team had offered the conclusion that the stimulation of the immune defense system that allowed organisms to throw off the effects of trypanosomiasis had to be very significant in that all attempts to try to effect such stimulation through immuno-stimulants as well known as B.C.G., or Corynebacterium granulosum, had led neither to the cure produced by the Priore Ray, nor to any prolongation of the infected animals' lives, nor even to the slightest modification in the evolution of the Trypanosomiasis.

These three scientific papers did little for the cancerologists who read them except to exacerbate their urge to oppose the Priore research, if not to arouse their outright hatred for the principal experimenter, Raymond Pautrizel. Could this have been because, for over 20 years, the same cancerologists had been working in vain to provoke in cancerous organisms immuno-stimulative reactions by intensively and successively vaccinating them with B.C.G.? Many others had been life-long apostles of chemotherapeutic cocktails of all sorts, or life-destroying ionizing radiations, or, what more recently has become the fashion, of applying the two methods in endless combination.

For this reason, they saw Priore and Pautrizel as nothing more than spoil-sports who had to be destroyed.

One of the opening shots in this campaign was a letter received by Pautrizel to inform him that his request for funds to continue his research through Unit-89, a unit that had been specially set up for him to direct, had been denied. It took many months of investigation for Pautrizel to learn that the real reason for the refusal was because of his work with Priore.

Next Pautrizel was informed that his appointment as director of the same research unit would be extended for only two years, whereas the normal extension for similar units was five years. A third insult came when Pautrizel tried to win a post within his unit for a high-ranking military physician, who had been his student and who had decided to quit the military in order to participate in the fascinating research prosecuted by his mentor. Pautrizel's request for funds to pay this physician, who all his life had been working on tropical medicine closely associated with problems of trypanosomiasis, were refused four times in a row with no cogent reason given. The physician, who in the meantime had volunteered his time without pay, finally became so emotionally overwrought that he gave up his medical career and retired to the countryside where he gave himself over to alcohol. Then Pautrizel tried to get a salaried post for another of his brilliant collaborators (who still works with him). He was told that this man could take up his new functions only if he left Bordeaux. One could go on with many other shocking stories but we will leave it to Graille to conclude: "Everything possible was done to isolate Pautrizel, to separate him from his collaborators. Every single one of these collaborators saw their careers put in jeopardy, compromised, or broken."

As a final insult, when the time came again to renew Pautrizel as director of Unit 89, those responsible, not daring to overstep what even they knew to be decent limits by not extending him, simply abolished the unit. And to add injury to that insult, a doctoral thesis that had now been prepared by Priore, and backed not only by Pautrizel but by Nobel Laureate Andre Lwoff himself, was summarily refused by the President of the University of Bordeaux.

It is perhaps unnecessary to state that the details behind all of this skulduggery could, and did, fill up two chapters of a book and make for the most heart-rending reading imaginable.

So what happened next? In the autumn of 1977, Professor Georges Dubourg, one of the leading lights in Bordeaux's company of surgeons and a friend and admirer of Pautrizel's, came to him to say openly and baldly: "My friend, at the point you've reached, there is only one more way to jolt medical opinion and that is to treat human cancer patients." Pautrizel was hesitant, believing his role to he one of continuing with his animal experiments but where would the funds for that come from now? He therefore asked his old mentor, Robert Courrier's advice. Courrier gave the green light. The treatments were restricted to terminal cancer patients whose immune defense systems had been disastrously weakened by chemotherapy or radiation or both. At least one of them was totally cured. The other lived, without pain, for a period many times longer than predicted by prognosis. Dubourg, Pautrizel and their collaborators wrote up the results and sent them as an official communication to the French Academy of Medicine for publication.

The reply they received from that Academy's perpetual secretary reads: "Experts whom we consulted consider that your work does not fall within the jurisdiction of our members and that it would doubtless find an audience more worthy of its purpose in a more specialized society."

To which Pautrizel formally replied: "Since two of the four signatories of our note are corresponding members of your Academy, could we not benefit from the remarks and comments made by the committee which saw fit to refuse our paper? And even, if this is not too indiscreet a request, to learn the names of the expert members who were consulted which would allow us to get into contact with them directly and to benefit from their singular competence?"

His letter has remained unanswered for four years.

There was nothing more to do except one thing which Raymond Pautrizel, as a man of science, had always been careful to avoid: Get a responsible journalist interested in the case, inform him of all possible details, and let him carry the Priore Affair in all its harrowingly loathsome aspects to the broad reading public. That journalist was Jean-Michel Graille.

For four years, Graille went about his task, publishing three consecutive long articles in his newspaper **Sud-Ouest France** and finally the book to which we have referred and of which this presentation is largely a resume. As early as 1980, Graille would write in his newspaper: "The Priore Affair is simple in essence. It can be reduced to a simple alternative: either the machine developed by Antoine Priore is of no interest and, having shown this, the affair can be considered at an end. Or else the machine is of real and demonstrable medical interest and, if that is officially recognized, he would be allowed to get on with the work. For this dilemma runs the risk, yet again, of being buried under delays and evasions. Beyond all the powers-that-be that have been directly connected to the affair for many years now — the power of finance, the power of medicine, the power of science — perhaps it is now political power with which responsibility lies if it can rise to meet and assume that responsibility through decision."

That was Graille's statement in 1980. His book which came out four years later ends with the sentence: "The Dossier Priore thus depends, from here on out, on a decision that must be taken on the very highest level, and imperatively. This responsibility devolves, in last resort, on the chief of state and on him alone. Will he assume it?"

Would the President of the United States?

Note added by T.E.B.: Antoine Priore is now dead. His machine has been dismantled. The iron dogma and hatred of electromagnetic medicine by bureaucratic science may have doomed hundreds of millions of humans - - whom Priore's device could have saved - - to bitter, agonizing, and unnecessary deaths. Hitler, Stalin, and Mao combined were not responsible for the deaths of so many.

APPENDIX II

PATENTS

English Translation of the French Patent No. 1,342,772 of Antoine Priore, 7 Oct. 1963.

Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of living cells.

Brevet d'Invention. P.V. no. 899.414 no. 1.342.772 Classification Internationale: A61 k-H 05 g

Procedure and assemblage for production of radiation especially serviceable for the treatment of living cells.

Antoine Priore
Requested 1 June 1962, 14.52 hours, Paris
Released by decree (arrete) of 7 Oct 1963

The invention deals in a general manner with radiation capable of penetrating matter. More exactly, it aims at a procedure and an apparatus making it possible to obtain a combination of radiations of different types able to penetrate matter and especially to penetrate intimately living organic tissues in order to produce in them certain effects, particularly in human tissues with a view to a therapeutic effect without destroying essential elements such as the enzymes.

In conformity with the present invention, one emits in a cavity a stream (rayonnement) of electrically charged particles upon which one superposes electromagnetic radiation of the centimeter wavelength range, the wavelength of which is preferably between 3 cm and 80 cm, and one directs the resulting radiation emerging from the cavity on to the object to be irradiated.

This applicant has shown that the penetration and, in particular, the curative effects are very distinctly improved when one gives the electromagnetic radiation a frequency determined as a function of the organ or the tissue to be penetrated or to be treated. For example, a wavelength of 14 cm is suitable for the liver and a wavelength of 19.5 cm for the spleen.

Preferably, the stream of charged particles is accelerated in a particle accelerator in such a way as to increase the force of penetration.

The resulting radiation is advantageously applied and directed upon the target, that is to say, upon the tissue to be penetrated, by the intermediary of a tube which is the site of electric fields and of magnetic fields for acceleration and control, the said radiation being preferably guided and/or reflected by a rotary deflector placed in the tube.

It is often advantageous to modulate or impose rhythm on this stream of particles by means of variable magnetic and/or electric fields so as to augment still more the force of penetration. This rhythm is preferably consistent, especially in medical applications, with the intrinsic period of the tissue to be penetrated or of the neighboring tissues, for example, muscle. These intrinsic periods arc well-known in medicine and are applied, particularly, for diathermy: they are situated in the wavelength range from 1 m to 50 m and more especially from 1 m to 18 m.

Preferably, one arranges to modulate the emission of radiation, the accelerating magnetic and electric fields, and perhaps also the rotatory deflector, to the cardiac rhythm of the subject.

It seems that the result obtained by the invention in the treatment of maladies of living cells (vegetable or animal) are due to certain phenomena which will be described, it being understood that this exposition will not circumscribe the invention.

As a function of its electro-physico-chemical constitution, the cellular pair nucleus-protoplasm is endowed with electric conductivity in direct relationship with ion exchange processes provoked by metabolic phenomena. One finds in tissues the presence of an accumulation of electricity at different potentials according to the different cellular densities of the tissues.

The work of Renshaw, Forbes, Morison, Amassian, de Vito, Ruset, Albe-Fessard, Tauc, Adrian, etc. has shown with the aid of micro-electrodes the existence of slowly oscillating elementary electric activity in the interior of cells; it can be thought that the *rythmeur* (or pace-maker) is achieved by the oscillatory electromagnetic system comprising the cell nucleus. This nucleus, in effect, is made up essentially of tubular filaments of insulating material (related to chitin) containing in its interior an electrically conducting saline liquid, and these filaments, coiled upon themselves, can be considered to constitute real little oscillatory circuits.

The recent work of Warson (sic) in America, as well as that of French scientists, including a communication from Polonsky, Don zon and Sadron presented to the **Academie des Sciences** by Prof. Francis Perrin on 16 May 1960 **(Rec. comptes rend, heb.,** 250 No. 20, 3414-3416) making it clear that experimental samples of **solid** DNA manifest properties analogous to those known in ferroelectric materials, makes plausible the hypothesis that a potential difference may exist between nucleus and periphery of cells. Certain recent theories go even further and liken the cell to an electronic receiver-emitter device normally functioning in harmony with the ambient media. The oscillatory system of demand waves, constituted by the cell nucleus, would behave in accordance with the laws governing semi-conducting materials.

The applicant is led to the conviction that in a normal state of physico-electric equilibrium, the cell nucleus is positively charged, but can acquire a negative surcharge following phenomena analogous to polarization.

The invention, especially, enables organs afflicted by this inversion of their electric potential, particularly in the case of the pathologic negative surcharge of cancerous nuclei, to recover their former equilibrium.

The following description in regard to the attached drawing, given as a non-restricting example, will make it possible to understand how the invention can be realized, the details which emerge both from the text and the drawing being, of course, part of the said invention:

Fig. 1 is a schematic section showing an apparatus for production and emission of an electromagnetic field combined according to the invention.

Fig. 2 is a frontal elevation of the cathode, taken from the right of Fig. 1.

Fig. 3 is a section through III- III in Fig. 1.

Fig. 4 is the overall scheme of the electric supply.

Fig. 5 is a view analogous to Fig. 1 showing another mode of implementation.

Fig. 6 is a section through VI-VI in Fig. 5.

Fig. 7 represents schematically an apparatus for pulsing the electric current.

Fig. 8 is the circuit of an amplifier permitting activation of the apparatus of Fig. 7 at the cardiac frequency.

Fig. 9 is the circuit of an oscillator permitting modulation of the electric current according to a wavelength between 1 m and 18 m.

The assemblage of Fig. 1 contains an apparatus 1 emitting electrically charged particles 2 in a cavity or passage 3, a cyclotron 4 accelerating the particles 2 and sending them into a cavity 5 forming a tube into which merges another cavity 6 acting as waveguide for electromagnetic radiation of frequency in the centimeter range emitted by a magnetron 7. The cavity 8 formed by the joining of tube 5 and waveguide 6 leads into a tube 9 in which the resulting radiation is accelerated and aligned. The interior of the cavity-formed by the assemblage of elements 1, 3, 5, 6, 8, and 9 contains argon at a pressure of 2 mm mercury.

The particle emitter 1 consists of an electron gun comprising a plate 10 and a cathode 11.

The cathode 11 is of molybdenum and has the very special form represented in Figs. 1 and 2. It consists of a rim 11a connected by two aligned spokes 11b to a hub 11c pierced by a hole 11d with its axis along XX'. The rim 11a is in two parts (as one sees in Fig. 1) which may be held together by screws (for example) forming a cavity of revolution 11e traversed by a number of holes 11f parallel to the axis XX' and regularly spaced. The filament for heating, 12, situated in the interior of the cavity 11e and is connected to the power supply conductors 12a.

The best results are obtained with a cathode 11 of molybdenum. The applicant has obtained satisfactory, but slightly inferior, results with tungsten cathodes. It turns out that molybdenum, and to a lesser degree tungsten, are the metals whose valency is closest to the mean valency of the chemical molecules constituting living tissues and more particularly those of human beings. One might well seek to use this fact to explain scientifically the phenomena involved, but it is understood that *the invention is not limited by any scientific explanation.* Concerning, on the other hand, the gas present in the apparatus at low pressure, the best results have been obtained with argon; the applicant has also obtained satisfactory, but slightly inferior, results with other gases of the rare gas series.

Surrounding the tube which constitutes the electromagnetic chamber are arranged: an electromagnet 13 with its winding 13a, placed at the level of the accelerating coils 14 and 15. Other accelerating coils 14,15 and 16 are similarly dispersed around the cavities 3, 5 and 8.

The two semi-circular boxes or "dees" 4a of the cyclotron 4 are placed in the usual manner between the poles of the frame around which are wound the accelerating coils 4b and 4c.

The magnetron 7 is of a familiar type and must be capable of emitting in the cavity 3 a centimeter wave of adjustable wavelength from 3 cm to 80 cm.

The lower portion of the tube 9, for acceleration and alignment, contains a cathode 17 resembling the cathode 11, with a filament 17a. This cathode 17 is supported by a hollow pillar 18 pierced by holes 18a close to its junction with the bottom of tube 9. This pillar 18 communicates with a tube 18b emerging on the axis of a rotary deflector 19 carrying at each end a "crown" of graphite plates 19a inclined at 45° to the vertical. The rotating axis 19b of the rotary deflector is mounted in a support 20 fixed to the interior of tube 19 and carries at its upper extremity magnetic bobs 19c which ensure that it will be set in motion in cooperation with the magnetic bobs 21a mounted on the shaft 21b of a motor 21. The lower extremity of the rotary deflector 19 is composed of a piece of molybdenum or of tungsten 19d in the form of a pyramid whose apex is opposite the open end of the tube 9.

The hollow base 18 and the tube 18b can be of a borosilicate glass of low coefficient of expansion such as that sold under the trademark "Pyrex." They may also be of quartz. Tube 9 itself can be of "Pyrex" as above or of another glass of the quality currently used for the manufacture of electronic tubes, but its bottom 9a, which is traversed by the radiation, is advantageously made of quartz.

The duct 8 joins the tube 9 by way of several tubulures such as 8a and 8b directed in vertical planes towards the plates 19a at a certain angle, which is advantageously about 22.5°. An electromagnet 23 analogous to the electromagnet 13 of the emitter tube 1 is placed around the cathode 7. Similarly, accelerator windings 24 are disposed around the tube 9. This tube also carries, at positions indicated in the drawing, three electrodes 25, 25a, and 25b surrounded by windings 26, 26a, and 26b, respectively. The drawing shows the supply lines, 17b, 17c of the cathode and its filament and that of the plate, 22a.

The basic plan of the electric supply is represented in Fig. 4. The part 27 feeds an initial branch with low voltage alternating current: this consists of a rectifier 28 (e.g., a Kenotron) whose rectified current is modulated at a frequency variable between 30 and 120 pulses per minute by means of a resistance 29, the control apparatus lor which will be described with reference to Figs. 7 and 8. The current so modulated is passed through the electromagnets 13 and *23* in such a manner as to generate, normal to the cathodes 11 and 17, a modulated field of 10,000 to 20,000 gauss.

The part 27 also feeds a variometer (interrupter) 30 which modulates the current of this part at a variable frequency 30 to 120 pulses per minute, the current pulsed in this way serving to feed the remainder of the installation, to wit: The magnetron 7; a converter 31 whose excitation is modulated at a variable frequency *300 to 900 Hz*, yielding a doubly modulated current (first at 30-120 pulses per minute, then at 300-900 Hz) which feeds the coils 15, 16 and 26; another converter 32 producing a low voltage direct current modulated at 30-120 pulses per minute by the variometer 30. This current feeds the motor 21 as well as the motors driving the variometer 30 and the apparatus controlling the resistance 29.

The current produced by the converter 32 also feeds a voltage step-up apparatus 33 consisting of a vibrator followed by a transformer and a rectifier, and producing a direct current varied at 30 to 120 pulses per minute imposed by the variometer 30. The maximum value of the voltage produced by the apparatus 33 is, for instance, 300,000 V, but this value may vary up or down, depending on the power one wishes to operate with.

The current produced by the voltage step-up apparatus 33 feeds the coils 4b of the cyclotron and 24 of tube 9, as well as a rheostat 34 permitting regulation of voltage to the desired value between 5000 V and 70,000 V. This voltage is applied to an oscillatory circuit 35 which produces oscillations at a frequency variable at will of wavelength between 1 m and 18 m. The current available to the output terminals 35a, 35b of this oscillating circuit 35 is thus a high tension current modulated first at 30 to 120 pulses per minute (by the variometer 30) and secondly at a wavelength 1-18 m. This current feeds the coils 4c and 14; the electrodes, 25a and 25b are connected to terminals 35a and 35b, respectively, and electrode 25 is connected to the mid-point 35c.

Cathodes 11 and 17, the "dees" of the cyclotron, and the plates 10 and 22, not shown in the diagram of Fig. 4, are connected to the output of the step-up assembly 33, while the heating current for the filaments 11e and 17a is furnished by the resistance 29.

To operate the apparatus, one adjusts the controls of the resistance 29 and the variometer to the desired rhythm; in medical applications, this is advantageously the subject's cardiac rhythm: This rhythm is thus imposed upon the whole installation. The cathode 13 emits toward the left a stream of positively charged particles 2, which are concentrated by the electromagnet 13 and accelerated by the coils 14, 15 and 16 and by the cyclotron 4. Superimposed on this particle stream in the duct 8 is the electromagnetic radiation from the magnetron 7, which is adjusted to the wavelength found by experience to be most appropriate to the cells which are to be penetrated, e.g., 14 cm for the liver and 19.5 cm for the spleen. The resulting radiation is deflected, directed and accelerated in tube 9 and is directed by way of the base of this tube toward the target to be penetrated.

It must be noticed that the magnetic field of the coils 15,16 and 26 is modulated, by means of the converter assembly 31, at a frequency adjustable *between 300 and 900 Hz.* This modulation has the effect of concentrating the particles, that is to say to detach them from the walls of the tubes, and it also permits an appreciable saving in weight of the iron cores of the coils. One chooses the highest frequencies (i.e., around 900 Hz) when one wishes to produce hard radiation at the exit of tube 9, and the lower frequencies for soft radiation.

The unidirectional magnetic fields of the coils 4c of the magnetron (sic) (cyclotron?) and the accelerator coils 14 as well as the electric field of electrodes 25, 25a, 25b, are modulated by the oscillatory circuit 35 at a chosen wavelength between 1 m and 18 m. In medical applications notably one selects the wavelength that best suits the organ to be treated or its surroundings, such as muscle. As already indicated, experience with diathermy makes it possible to determine the most suitable wavelength.

It must be noted that the resulting radiation already possesses, in tube 8 (Fig. 1) a considerable penetrating force. One could therefore use the assemblage described by omitting tube 9 and terminating the cavity at the end of tube 8 by means of a glass or quartz base, the resulting radiation being accelerated and directed, for example, immediately upstream from the base, by a final coil (not shown) surrounding tube 8. However, tube 9 appreciably improves the results.

Figs. 5 and 6 represent another mode of realizing the assemblage in accordance with the invention, in which the elements playing the same role are indicated by the same signs as in Figs. 1 and 3, modified by primes.

The arrangement of the connecting ducts of Fig. 5 in relation to tubes 1' and 9', to the magnetron 7' and the cyclotron 4', is different from that of Fig. 1 and has been used successfully by the applicant. The waveguide 6' of the magnetron 7' opens into the end of tube 1' and the duct 3', carrying the resulting radiation, divides into two branches: Branch 36, which conducts the radiation directly to tube 9', and branch 37 which conducts it to the cyclotron 4'. This blocks the electromagnetic radiation and accelerates the stream of particles which is passed by way of duct 38 to the tube 9'.

This arrangement can be used with a particle emitter and an accelerator and director tube similar to tubes 1 and 9 of the preceding figures. However, tubes 1' and 9' of Figs. 5 and 6 are constructed in a different manner with regard to their cathodes and anodes.

Tube 1' contains a first electrode 11' exactly like that of the cathode 11 of tube 1, and an identical second electrode 39 furnished with a filament 39a. Tube 9' (Fig. 6) contains in its lower part a first electrode 17' with filament 17'a and an identical second electrode 40 with its filament 40a.

In normal operation, i.e., to produce radiation identical to that described in connection with Figs. 1 to 4, electrode 11' serves as cathode and electrode 39, given a positive potential, plays the role of the plate 10 in Fig. 1, the filament 39a being unheated. Electrode 40 and its filament 40a are disconnected; cathode 17' and plate 22' are supplied as in Fig. 3.

To obtain particularly penetrating radiation, the polarities are reversed: Electrode 11' becomes an anode and its filament 11'e is disconnected, while electrode 39 receives the cathode supply and its filament 39a is heated; electrode 17' (with filament 17'a disconnected) and electrode 22' become anodes, while electrode 40 serves as cathode and its filament 40a is heated. For example, one can establish a potential difference of 250,000 V between 40 and 17', and 50,000 V between 40 and 22'. It is understood that in these conditions the cathode 39 emits to the left a stream of electrons which is concentrated, modulated and accelerated by the various coils and in the cyclotron, the polarities of which must of course be established in the proper direction. This stream of electrons is combined with the centimeter radiation emitted by the magnetron 7', resulting in tube 9' in the emission of very hard x-rays, modulated at the chosen frequencies, combined with the centimeter radiation of the desired frequency.

The assembly shown in Figs. 5 and 6 thus permits one to obtain at will either very hard x-rays or the radiation described in connection with the preceding diagrams.

The following description, referring to Figs. 7 and 9, relates to certain details of the apparatus used for modulating the electric current.

Fig. 7 represents schematically the apparatus for control of the resistance 29 and the variometer 30. The variable resistance 29 consists of a graphite helix 29a immersed in a conducting liquid **29b;** another electrode 29c, also of graphite, partly immersed in the liquid, is set into up-and-down oscillations by a connecting rod 41a linked to a fly-wheel 41. The fly-wheel is set in rotation, through the intermediary of a worm transmission 41b, by an axle 42 which can be driven, thanks to a double clutch 42a, 42b, either by a motor 43, or by the shaft 30a of the variometer 30. The variometer 30 is driven by a motor 44 by way of the worm transmission 44a.

If the fly-wheel 41 is driven by the motor 43 at the proper speed, the resistance 29 causes the supply current of the electromagnets 13 and 23 (Figs. 1 and 4) to vary at the chosen rhythm which as we have seen can be between 30 and 120 pulsations per minute and which can be checked by means of a rotation counter shown schematically in 45. In this case, the motor 44 of the variometer 30 can be stopped and the remainder of the installation is then not pulsed. If, on the contrary, the fly-wheel 41 is engaged in 42b and disconnected from 42a, the motor 44 activates the variometer 30 and the resistance 29 at the chosen rhythm.

The speed of rotation of motors 43 or 44 can be regulated at the required speed, corresponding visibly to the cardiac rhythm of the subject, by acting upon the excitation of these motors by means of a manually adjustable rheostat. If one prefers to regulate the speed of motors 43 or 44 in direct accord with the cardiac rhythm of the subject, one can use an assembly such as that represented schematically in Fig. 8: At 46 there is a contact microphone which, when placed over the subject's heart, produces impulses. These are amplified in the circuit shown and applied to an electromagnet at 47 with a moving core which activates a rheostat; this in turn regulates the current running the motors 43 or 44.

Fig. 9 shows schematically the principle of the oscillating circuit 35. The rectified potential, adjustable between 5000 V and 70,000 V by means of rheostat 34 (Fig. 4) is applied between the terminals 48 and 48a. Terminal 35c (which is also connected to electrode 25, Figs. 2 and 4) is connected to the neutral point on the high tension side of **the** transformer which is a component of the step-up assembly 33 (Fig. 4). The terminals 49 and 49a receive the heating current produced by the resistance 29. The variable condensers 50 and 50a make it possible to regulate to the desired wavelength (which, as seen, is between 1 m and 18 m) the current available at the output terminals 25a and 25b of the oscillator shown.

The modes of implementation described gave been successfully carried out but it is self-evident that these are only examples, and that they might be modified, notably by substitution of equivalent techniques, without going beyond the bounds of the invention. In particular, the electron gun 1 or 1' could be replaced by another charged particle generator.

Resume

The invention includes especially:

> A procedure for obtaining a combination of radiations of different kinds capable of penetrating matter, especially of

intimately penetrating living tissues in order to produce in them certain effects and more particularly in human tissues with a therapeutic effect in mind, consisting of the emission in a cavity of a stream of electrically charged particles, upon which is imposed electromagnetic radiation in the centimeter wavelength range, and the guiding of the resulting radiation emerging from the cavity toward the target to be penetrated.

2. Types of implementation exhibiting the following features taken separately or in the various possible combinations:

    a. The centimeter radiation has a wavelength between 3 cm and 80 cm;

    b. This wavelength is set at the value found by experience to be most suitable for the tissues to be penetrated, e.g., 14 cm for liver and 19.5 cm for spleen;

    c. The particle stream is accelerated by magnetic and electric fields such as those which are used in particle accelerators;

    d. The resulting radiation is accelerated and guided, before its emergence from the cavity, by means of electric and magnetic fields;

    e. The resulting radiation is guided, before its emergence from the cavity, by means of deflecting and/or reflecting surfaces;

    f. The stream of particles and/or the resulting radiation are concentrated and accelerated by means of individual magnetic fields modulated at a frequency between 300 and 900 Hz, the highest frequencies being used to produce hard rays and the lower frequencies to produce soft rays;

    g. The emission of the particle stream, as well as the acceleration and concentration of the radiation resulting at its exit from the cavity, are aided by individual magnetic fields of temporally variable intensity, advantageously pulsed at a rhythm between 30 and 120 pulsations per minute and preferably at the cardiac rhythm of the subject;

h. The assemblages for production of the resultant radiation are pulsed in their entirety at the same rhythm as the magnetic fields according to g;

i. The stream of particles and/or the resulting radiation are accelerated and concentrated by direct magnetic and/or electric fields modulated at a wavelength between 1 m and 50 m and preferably between 1 m and 18 m, this wavelength being advantageously chosen as that which is known in diathermy as suitable for the tissues to be penetrated or for the surrounding tissues.

3. An assemblage making it possible to obtain a combination of a stream of electrically charged particles and a beam of centimeter electromagnetic waves in order to penetrate intimately and to irradiate living tissues and particularly human tissues, the said assemblage comprising at least a particle emitter, means for channeling said particles in a cavity serving as waveguide for an emitter of electromagnetic radiation of which the wavelength is included in the range of centimeter waves and preferably adjustable from 3 cm to 80 cm, means for generating in the cavity magnetic field for acceleration and concentration and means for concentrating and accelerating the resulting radiation at the exit of the cavity.

4. Modes of implementation with the following details taken separately or in the various possible combinations:

a. The particle emitter is an electron gun of which the anode is at the end of the cavity and the cathode is situated further along, this cathode being hollow and placed in the magnetic field of an electromagnetic in order to ensure emission of a stream of particles towards the mouth of the cavity;

b. The cathode consists of a rim connected by two aligned spokes to a hub, the said rim being provided internally with an annular housing containing a heated filament and the said housing communicating with a number of holes arranged annularly and traversing the rim transversely;

c. The cathode is made of a metal of valency close to the mean valency of the chemical molecules comprising the tissue to be penetrated;

d. The cathode is of tungsten or preferably of molybdenum;

e. The cavity contains a rare gas, preferably argon, under a vacuum of the order of 2 mm Hg;

f. The cavity contains a duct carrying at least part of the stream of particles to a cyclotron and a duct bringing back into the cavity the particles accelerated in the cyclotron;

g. The cavity passes through several coils, the supply current for the various coils being capable of undergoing modulation at different frequencies;

h. The downstream end of the cavity is composed of a tube containing, in the part from which the resulting radiation must emerge, a cathode and an electromagnet which may be identical to the cathode and the electromagnetic according to para. a, an anode near the other end, and a rotary deflector consisting of a number of plates arranged *en couronne* on a rotor lacing the incident radiation at such an angle that the radiation deflected and/or reflected is directed toward the cathode, several coils whose supply current can be modulated being distributed over the length of the tube;

i. This tube also contains electrodes supplied by alternating current generating an electric field at the level of the rotary deflector, each of the said electrodes being surrounded by a bobbin of which the supply current can be modulated;

j. Methods are anticipated for modulating, at an adjustable rhythm between 30 and 120 cycles per minute, the supply current of the electromagnetic according to a and h, and preferably to modulate the supply current of the rest of the assemblage at the same rhythm;

k. Methods are anticipated for modulating, at a frequency between 300 and 900 Hz, the supply current of the bobbins surrounding according to i and one or several coils according to g;

1. Methods are anticipated for modulating, at an adjustable wavelength between 1 m and 50 m and preferably between 1 m and 18 m, the supply current of the electrodes according to i, of one or several of the coils generating the magnetic field of the cyclotron;

m. The electrodes of the electron gun consist of two electrodes identical to the cathode according to a, b, c, or d, the cathode of the tube according to h is replaced by a double electrode reproducing the arrangement of the electrodes of the electron gun, and methods are anticipated for reversing at will and simultaneously the polarities of these two pairs of electrodes and the direction of flow of the current supplying the acceleratory coils, a first pattern of polarities assuring the functioning of the apparatus in the conditions which are laid down according to a, and a second pattern of polarities assuring emission in the cavity of a stream of electrons combined with the centimeter radiation and giving rise, at the exit of the said tube, to emission of very hard x-rays.

No. 1.342.772 **M. Priore** **5 planches. - Pl. IV**

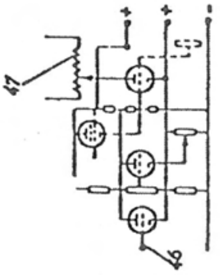

Fig. 18

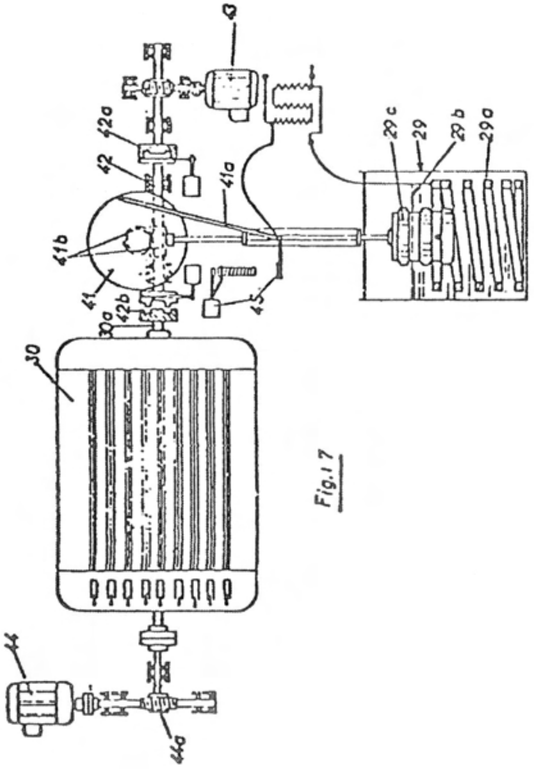
43
42a
42
41a
41b
41
42b
30a
30
45
29c
29
29b
29a
44
44a

Fig. 17

Fig. 9

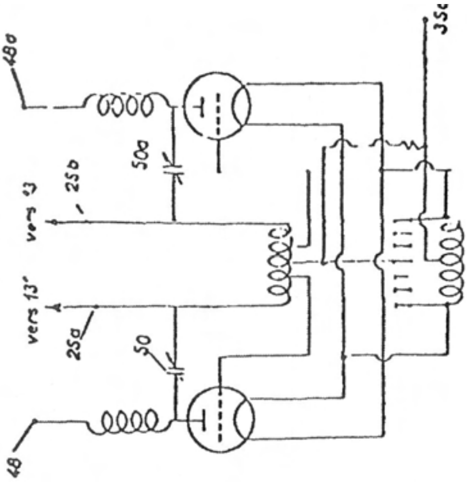
vers 13
vers 13'

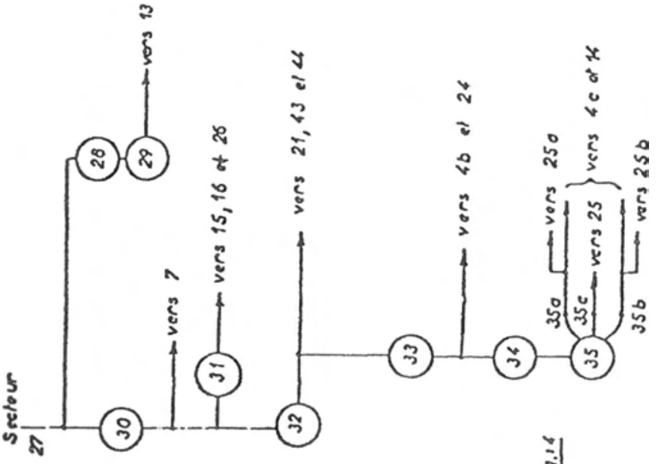
Secteur
27
28
29
vers 13
30
vers 7
31
vers 15, 16 et 26
32
vers 21, 43 et 44
33
vers 4b et 24
34
35
35a
35c
35b
vers 25a
vers 25
vers 25b
vers 4c et 14

Fig. 4

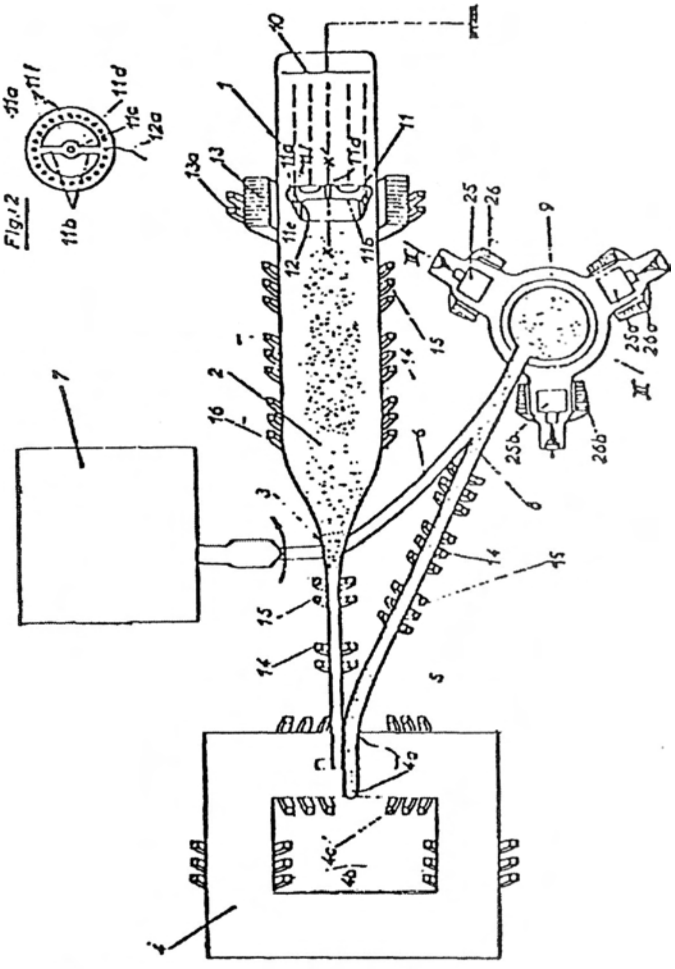
4a
4b
4c
III

Fig. 1

Fig. 2

Fig.: 3

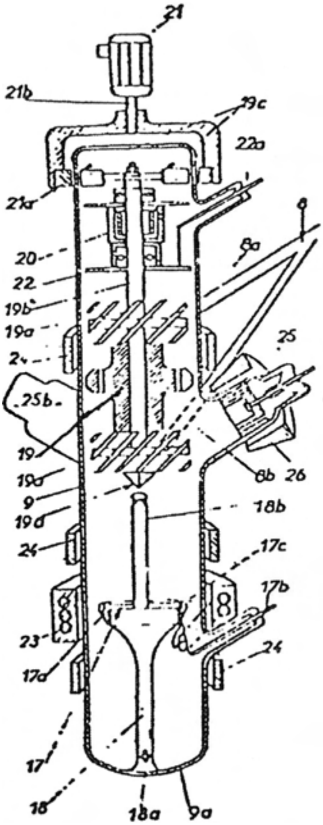
8a
8b
9a

Fig.: 6

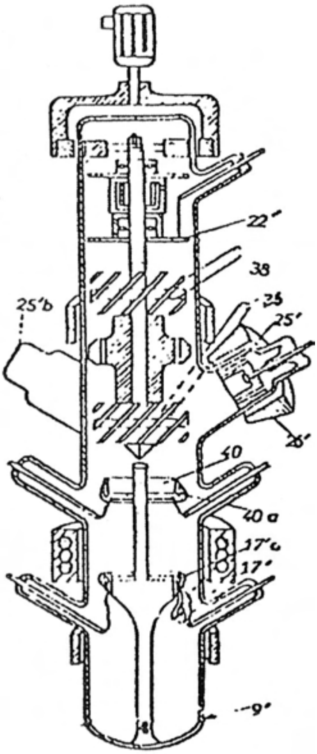
25'b
17'c

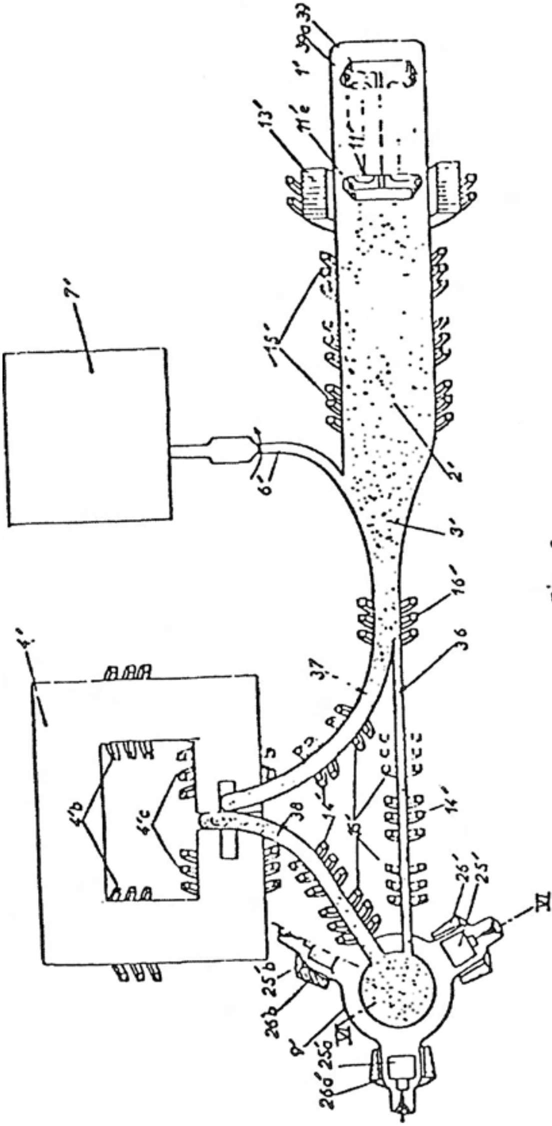
11'e
11'c
P2
4'b
4'c
VI
25'b
26'b
VI
25'a
26'a

Fig. 5

Two Patents of G. Lakhousky

June 12,1934 G. LAKHOVSKY 1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Filed Nov. 13, 1931

Fig. 1.

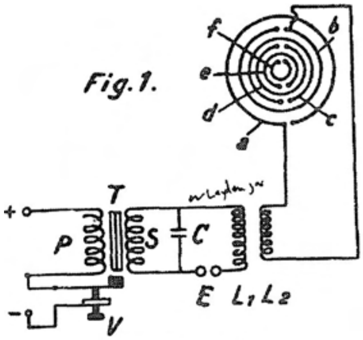
f
b
e
d
c
a
w Leyden jar
T
+
P
S
C
E
L1
L2
-
V

Fig. 2.

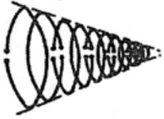

Fig. 3.

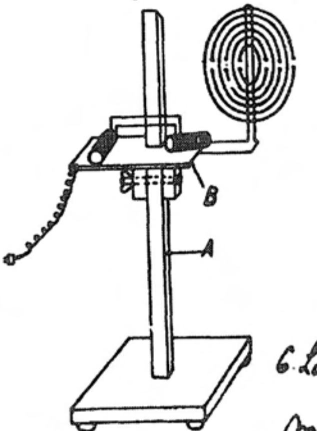

G. Lakhovsky
INVENTOR

By: Marks & Clerk
ATTYS.

Patented June 12, 1934 **1,962,565**

UNITED STATES PATENT OFFICE

1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Georges Lakhovsky, Paris, France

Application November 13, 1931, Serial No. 574,907
In France May 2, 1931

7 Claims. (Cl. 250—33)

The present invention has for its object an apparatus capable of sending out simultaneously different wave lengths so that among these waves or their harmonics there is found almost always one or several capable of producing the best effect sought for any application whatever.

There is used for that purpose the property of open circuits having a fundamental or natural self induction and capacity to oscillate upon a well determined wave length if they are excited by electric impulses emanating from discharges of any source.

The radiating part of the system, producing waves of various lengths, is therefore constituted by a certain number of concentric open rings of different diameters, which are maintained insulated by any appropriate process. These rings may or may not be terminated by small spheres forming capacities.

These rings may be arranged either in the same plane or in different planes so as to form any surfaces, such as portions of cones, spheres, paraboloids, ellipsoids, etc. . . . , these surfaces being of a nature to concentrate the waves in a determined direction in the manner of a reflector, for example.

These rings may be fed by any producing device whatever of high frequency, for example, a combination comprising a trembler coil (or any other transformer) and a high frequency circuit provided with a self-induction coil and a capacity.

One or several points of this high frequency circuit is or are connected to the end of one or several of the rings forming the radiating apparatus. The other rings not connected are excited by induction.

The complete apparatus (coil, high frequency circuits, rings, etc. . . .), may be mounted in a box or upon a small board capable of sliding along a guide or rocking so as to be able to be placed in the most convenient position.

In the accompanying drawing, by way of example, not limitative, there is shown:

Fig. 1 a diagram of the system.

Fig. 2 the position of the rings in the form of a cone.

Fig. 3 the apparatus mounted upon a foot or base.

In Fig. 1 there is seen at T the transformer or coil (P the primary, S the secondary of this coil), V, the screw of the trembler, C the condenser the discharge of which across the spark gap E produces oscillations in the self-induction coil L^1, which in its turn acts by induction upon the self-induction coil L^2; this self-induction coil L^2 is connected on one side to the first exterior ring *a*, on the other side to the following ring *b*. It thus forms an electrostatic field of high frequency which induces oscillations in the other rings *c*, *d*, *e*, *f*.

As a modification, L^2 may be suppressed and the rings *a* and *b* may be connected each to one of the ends of L^1.

One of the ends, or even any point of the self-induction coils L^1 or L^2 may likewise be connected to any one of the rings.

Fig. 2 shows the arrangement of the rings in the form of a cone.

The open circuits *a*, *b*, *c*, . . . *f* can be constituted by solid conductors. They can also be tubular circuits, this being advantageous, since, for the same bulk, the radiating surface is larger. Finally, the area of the cross section of each of these various circuits can vary from one circuit to the other; for instance, this area can decrease from the periphery (ring *a*) to the center (ring *f*) of the radiating system.

Fig. 3 shows the apparatus mounted upon a foot A along which it can slide, in order to be placed at any desired height and upon a hinge B which allows it to be oriented upon the horizon.

I claim:—

1. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters, the split ends being spaced from each other and terminating with small spheres.

2. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and terminated by small spheres forming capacities.

3. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and arranged in one and the same plane the split ends being spaced from each other and terminating with small spheres, in staggered position as regards adjacent rings.

4. An apparatus of the type described comprising a base, a plate slidably mounted on said

base, a support pivoted to the said plate, a series of open and insulated, high frequency radiating circuits mounted on the pivoted support, and a high frequency producing device for the excitation of the said circuits.

5. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits, said circuits being open, insulated, and having dimensions different from each other.

6. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits arranged in the same plane, said circuits being open, insulated, and having dimensions different from each other.

7. An apparatus of the type described comprising a transformer, a high frequency energizing circuit fed by the said transformer and provided with a self-induction coil and a capacity, a series of open and insulated high-frequency radiating circuits having dimensions different from each other, an inductive connection between two of these radiating circuits and respectively each of the poles of the outgoing element of the high frequency energizing circuit.

GEORGES LAKHOVSKY.

Patented June 13, 1944 **2,351,055**

UNITED STATES PATENT OFFICE

2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

George Lakhovsky, New York, N. Y.; Anne-Marie Louise Lakhovsky, administratrix of said George Lakhovsky, deceased, assignor to Henry S. Blum, Chicago, Ill.

Application November 21, 1941, Serial No. 420,006

11 Claims. (Cl. 250—33)

This invention relates generally to electric devices excited by electric impulses and more particularly to multiple wave length conducting and/or producing means. This invention has for its primary object the provision of such means disposed within a vacuum tube or a tube containing rare gas or gases.

My work over a period of more than 20 years has led me to the belief that there is great need for means conducting and/or producing multiple wave lengths which may be employed in many industrial and medical fields. Suitable apparatus for producing waves of various lengths is disclosed in my U. S. Letters Patent No. 1,962,565, dated June 12, 1934. Embodiments of this apparatus have been in use the world over. Because of the need for simplification in such apparatus so that it may be handy and easily used by persons lacking skill in electrical matters, I have concluded that by providing one unit which incorporates oscillation generating means and another in which the wave lengths are produced, the apparatus may be more universally used in the treatment of cellular life in therapy, in the aging or treating of liquids and in other industrial fields.

The present invention deals with those units, preferably combined with one another and constructed as a portable device and as stated above, has for its major object the provision of a unitary multiple wave length producing and/or conducting device in the form of a vacuum tube.

The present invention further contemplates the incorporation in such a tube, of means for varying the effect of wave lengths, or selectively employing the means therefor enclosed in or forming part of said tube.

This invention further seeks to provide a tube of the indicated type incorporating means for generating oscillations.

Another object of the invention is to provide a single unit which contains an oscillation generator, which produces waves of definite frequency, and further inductance means each permitting emanation of wave lengths of various values.

The structural features of my invention also form a material part of this disclosure, the objects and advantages being attained in structures such as shown in the accompanying drawing, which exemplifies the invention. The following specification, based on said drawing, more clearly points out the purposes and advantages of my invention.

In the drawing:

Fig. 1 is a vertical, partial sectional, partial elevational view of a tube incorporating features of my invention.

Fig. 2 is a similar view of an alternate form thereof.

Fig. 3 illustrates in a similar manner a modified form of my invention.

Fig. 4 is a fragmentary sectional view of a modification made in accordance with the invention.

Fig. 5 is a similar view of another modification.

Figs. 6 and 7 are sectional views illustrating types of wires used in either of the forms of my invention as shown in Figs. 1, 2 and 3.

Fig. 8 is a partial sectional, partial elevational view illustrating a tube incorporating a plurality of grids and selective capacity means for varying the effects of said grids.

Fig. 9 is a similar view of a tube, such as shown in Figs. 1 or 2 and incorporating oscillating or vibrating means.

In that form of my invention shown in Fig. 1, there is provided a sealed preferably transparent envelope **10** made of glass or like material affixed in the usual manner to a screw or Edison plug **11**. The latter is conventionally provided with a shell terminal **12** and a central terminal **13**. The support or seal **14** of the tube serves to hold wires such as **15** and **16** which connect to the terminals **12** and **13**.

Connected to the wires **15** and **16** is a primary inductance member in the form of a loose coil **17** which may be disposed axially of the envelope **10**, as shown. The upper end of said member **17** forms a tight wound coil **18** which serves to radiate the strongest induced wave lengths.

Surrounding the primary inductance member **17**, there are arranged a plurality of separate coils of wire each being of different length and capacity. Thus, I provide a coil **19** having greatest capacity and successive coils **20**, **21**, **22**, **23**, **24**, and **25**, each having less capacity than the preceding coil. These coils **19** to **25** form secondary inductance means each permitting emanation of a wave length of different magnitude and value than the other.

I may support the secondary coils in several ways. As shown, I prefer to provide a dielectric member such as a glass sleeve **26** surrounding the primary inductance wire member **17** and to mount the secondary coils **19** to **25** on the outer surface of said sleeve, leaving the coil **18** exposed at the top thereof. In order to effectively support and centralize the glass sleeve **26**, I provide spacing washers **27** and **28** of mica or the like, top and bottom, and provide a supporting crosspiece **29** on the lower portion of the wire **17**.

June 13, 1944. **G. LAKHOVSKY** 2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

Filed Nov. 21, 1941

Fig. 1.

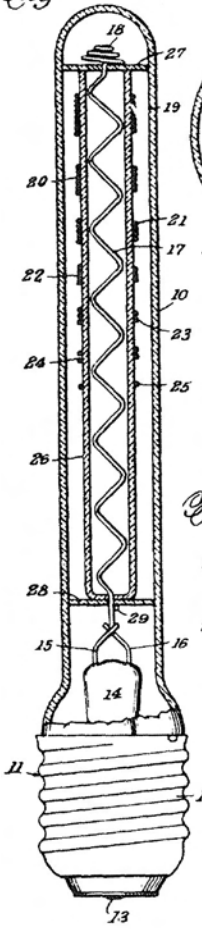
18
27
19
20
21
17
22
10
23
24
25
26
28
29
15
16
14
11
12
73

Fig. 2.

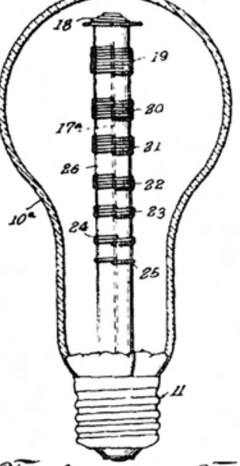
18
19
20
17a
21
26
22
10a
23
24
25
11

Fig. 3.

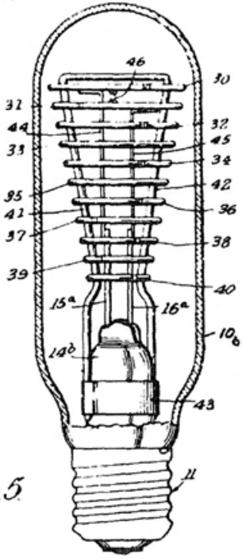
46
30
31
32
44
45
33
34
42
35
36
41
37
38
39
40
15a
16a
10b
14b
43
11

Fig. 4.

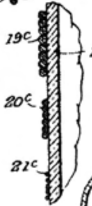
19c
26
20c
21c

Fig. 5.

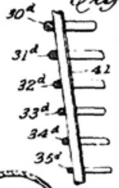
30d
31d
41
32d
33d
34d
35d

Fig. 6.

Fig. 7.

Fig. 8.

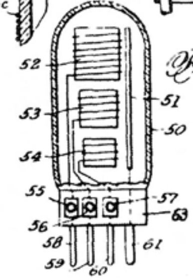
52
51
53
50
54
55
57
63
56
58
61
59
60

Fig. 9.

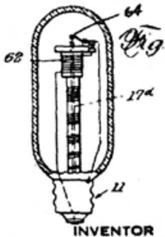

INVENTOR
GEORGE LAKHOVSKY
BY Leon M. Strauss
ATTY.

The tube above described may be employed as indicated for the treatment of cellular structures by connecting it to any one of the rings in the electrostatic field of high frequency, such as shown in said U. S. Letters Patent No. 1,962,565, each of the coils 13 to 25 producing by induction a different wave length, as can be understood.

In the form of my invention shown in Fig. 2, the envelope 10a is somewhat differently shaped, and the wire 17a straight instead of arranged as an open coil. In other respects, the structure follows that described with reference to Fig. 1.

As shown in Fig. 3, there may be arranged a plurality of open rings 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, and 40 in frustro-conical form, for instance, with the largest of said rings at the top and the smallest one adjacent the bottom. I have shown said rings as supported, for instance, by glass posts 41 and 42, said posts being supported on the seal 14b, as by a band 43. I provide extensions for the terminal wires 15a and 16a and connect said extensions 44 and 45 to the upper ring 30 and the next lower ring 31, respectively.

The rings 30 and 31 may become the primary inductance members, whereas the remaining rings constitute the secondary inductance members as before set forth. Electric current conducted to the suitably spaced rings 30, 31 will cause a spark 46 to be drawn between the rings 30 and 31 providing oscillations emanating from electrical discharges therebetween and causing the radiation of waves of various lengths within the envelope 10b and affecting the remaining rings.

In Figs. 1 and 2, I have shown the secondary coils as formed of wires having uniform cross-section. Fig. 4 illustrates how such wires may be made progressively smaller in the succeeding coils 15c, 20c, and 21c, etc. Similarly, I have shown the rings of Fig. 3 made progressively smaller in Fig. 5, said rings being designated by numerals 30d, 31d, 32d, 33d, 34d and 35d, etc. In such various manners, the wave length producing effect of the secondary coils or rings may be arrived at.

The secondary coils or rings may be round or somewhat flattened wire as shown in Fig. 6 or tubular as in Fig. 7, offering a yet greater flexibility of design.

The devices of Figs. 1 and 2 above described may be incorporated in the system shown in said U. S. Letters Patent by connecting one of the rings *a*, *b*, *c*, *d*, *e*, or *f* with plug 11 of Fig. 1, since either terminal 15, 16 is in contact with primary inductance member 17. To this end a conductor (not shown) connects such a ring with said plug or socket.

In Fig. 8, I have shown a modification of the invention which comprises a vacuum tube 50 containing a plate 51 and a plurality of different wave lengths producing grids 52, 53, and 54. I connect each grid with a variable capacity device, such as 55, 55, and 57 carried by said tube 50 and preferably by its base support 63, whereby the grids may be selectively connected into an operating circuit (not shown) by means of the respective prongs 58, 59, and 60, and the plate prong 61. Heating or other exciting means may be incorporated in the tube in a well-known manner. This combination tube 50 is designed to replace a plurality of known tubes each operating with different and variable wave length effect, mere exterior manipulation at said base support brings about change of the capacities 55, 56, and 57 and cutting in one or more of the grids 52, 53, and 54.

The effect of the spark as produced in the form of the invention depicted in Fig. 3 may be further utilized in Fig. 9 in which I also incorporate a vibrator 62 which produces electrical oscillations by a spark intermittently generated between the end of the primary inductance member 17d and vibrator armature 64. In other respects the structure follows that of Fig. 1.

It may be noted that any type of spark producing or oscillation generating means may be substituted for the vibrator shown. The tubes shown in Figs. 3 and 9 may be used in the manner described for Fig. 1 by connecting one terminal of the base as set forth to produce an induction effect. These tubes may be also connected across both terminals with a suitable potential to obtain the spark effect.

From the foregoing it may be seen that I have provided tubes in various forms for the purpose of generating and conducting multiple wave lengths. Other forms of the invention may be produced within the spirit and scope of the invention as claimed.

Having thus described my invention, what I claim as new and desire to be secured by Letters Patent, is:

1. A device of the character described comprising a vacuum tube having an envelope and base, primary inductance means, and a plurality of separate secondary inductance means of different length for radiating waves of various lengths extending above said primary inductance means, all said inductance means being arranged within said envelope.

2. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members spaced apart and independent from each other each member surrounding a respective portion of said primary inductance member and within the field of influence thereof.

3. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members each surrounding the primary inductance member and within the field of influence thereof, each of said secondary inductance members having a different wave length effect than the others.

4. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings.

5. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings, said latter rings being arranged in different planes and being of different length to produce a different wave length effect.

6. In a device of the character described, means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of separate wire coils of different lengths on said dielectric means and in the field of inductance of said wire member.

7. In a device of the character described,

means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of wire coils on said dielectric means and in the field of inductance of said wire member, each coil extending over said wire member and having a different number of convolutions than the others.

8. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings.

9. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others.

10. In a device of the character described means for producing multiple wave length comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others, all of the rings being arranged in frusto-conical disposition.

11. In a vacuum tube having a primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member for affecting the induction of said secondary members.

GEORGE LAKHOVSKY.

T.H. Moray's Patent
Electrotherapeutic Apparatus
Feb. 1, 1949

Feb. 1, 1949. **T. H. MORAY** 2,460,707

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 1

FIG. 1.

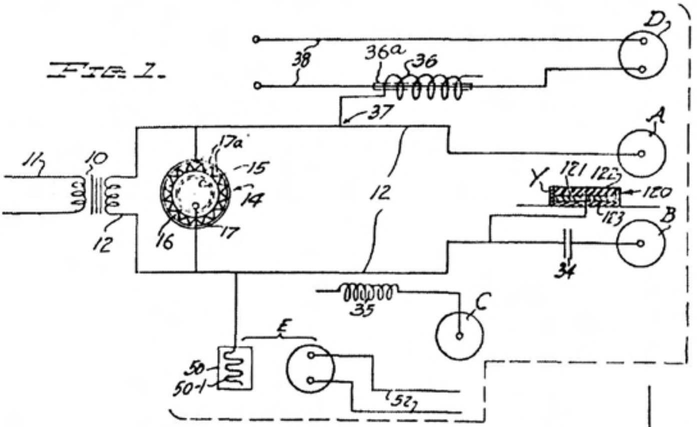

38
36a
36
D
37
A
11
10
17a
15
14
12
Y
121
122
120
123
B
12
16
17
34
35
C
E
50
50-1
52

FIG. 11.

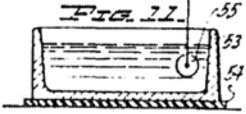

FIG. 3.

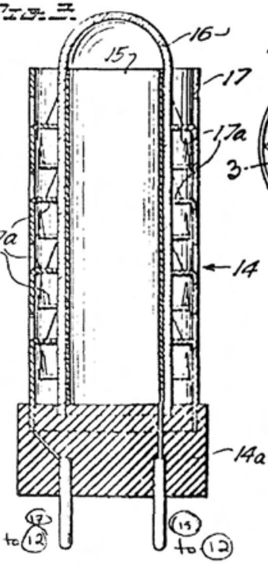

16
15
17
17a
17a
14
14a
17
to 12
15
to 12

FIG. 2.

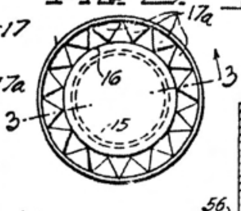

FIG. 12.

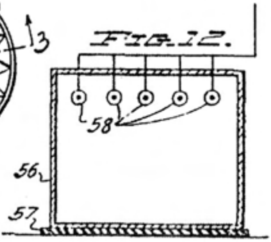

Inventor:
T. H. MORAY,
H. Mallinckrodt and
Philip A. Mallinckrodt,
Attorneys.

Feb. 1, 1949. T. H. MORAY 2,460,707

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 2

FIG. 7.

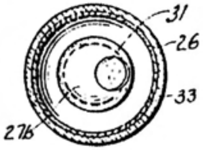

FIG. 5.

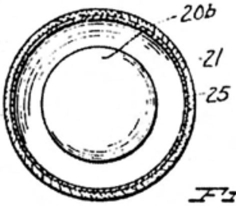

FIG. 9.

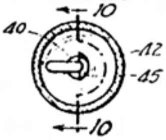
10
40
42
45
10

FIG. 10.

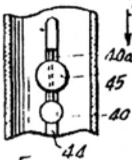

FIG. 8.

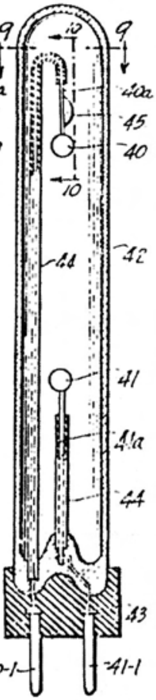
9
10
9
40a
45
40
10
44
42
41
41a
44
43
40-1
41-1

FIG. 6.

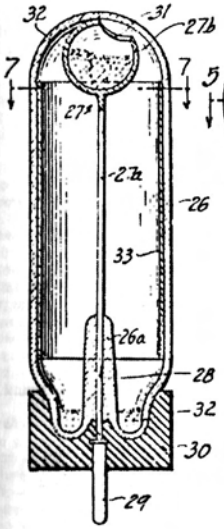
32
31
27b
7
7
27a
27a
26
33
26a
28
32
30
29

FIG. 4.

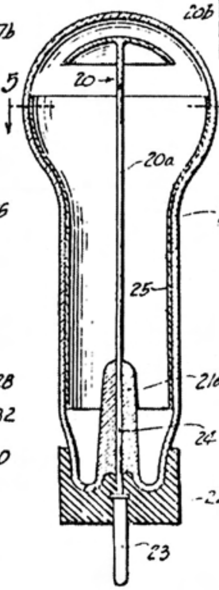
20b
20
5
5
20a
21
25
21a
24
22
23

Inventor:
T. H. MORAY,
C. H. Mallinckrodt and
Philip A. Mallinckrodt.
Attorneys.

Feb. 1, 1949. T. H. MORAY 2,460,707

ELECTROTHERAPEUTIC APPARATUS

Filed April 30, 1943 3 Sheets-Sheet 3

FIG. 13.

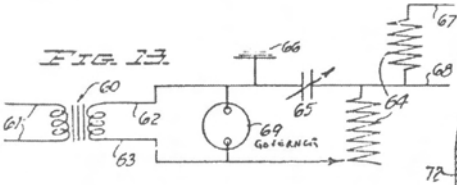
66
67
68
60
61
62
63
65
64
69
GOVERNOR

FIG. 14.

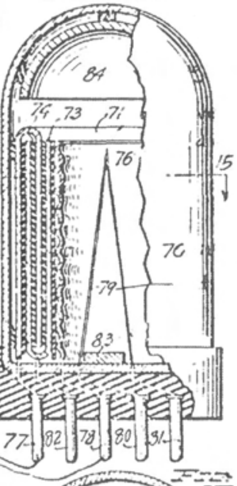
84
74
73
71
72
15
76
15
70
79
72
83
70-1
to 62+63
77
82
78
80
81

FIG. 15.

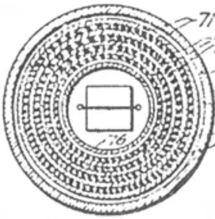
71
74
70
72
73
76

FIG. 16.

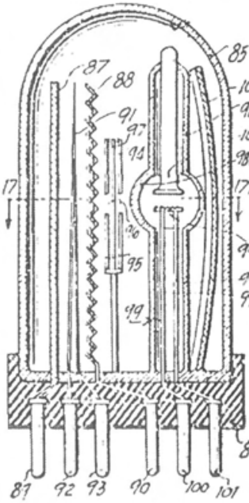
87
88
91
97
94
96
95
99
85
103
98
102
98a
17
17
99c
99a
99b
86
81
92
93
90
100
101

FIG. 18.

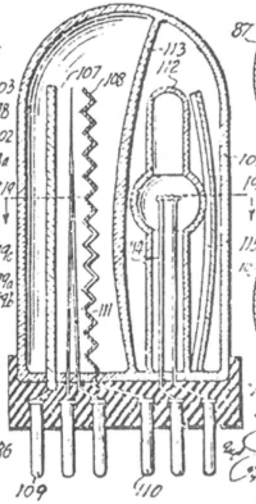
107
108
113
112
105
19
19
111
114
106
109
110

FIG.

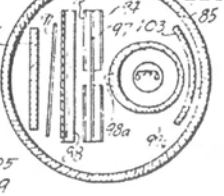
87
95
94
97
103
85
98a
98
99c

FIG.

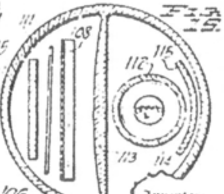
111
108
115
112
113
114

Inventor:
T. H. MORAY.
By
R. Mallinckrodt.
Chas. A. Mallinckrodt.
Attorneys.

Patented Feb. 1, 1949 **2,460,707**

UNITED STATES PATENT OFFICE

2,460,707

ELECTROTHERAPEUTIC APPARATUS

Thomas H. Moray, Salt Lake City, Utah

Application April 30, 1943, Serial No. 485,112

8 Claims. (Cl. 128—421)

This invention relates to electrotherapeutic apparatus, and to methods of applying electrical, radioactive, and other radiant phenomena therapeutically.

The invention is primarily concerned with the use of high potential, high frequency electricity though not necessarily limited thereto, in conjunction with radioactive and other types of electronic and radiation phenomena, for therapeutic purposes.

Among the objects of the invention are the following:

First.—To render highly effective, from a therapeutic standpoint, radioactive and other types of electronic and radiation phenomena, and, likewise, to render highly effective, from a therapeutic standpoint, high potential, high frequency electricity.

Second.—To augment the therapeutic effect of radioactive and other types of electronic and radiation phenomena by the conjoint use of high potential, high frequency electricity, and, conversely, to augment the therapeutic effect of high frequency, high potential electricity by the conjoint use of radioactive and other types of electronic and radiation phenomena.

Third.—To accomplish the above without danger of burning or of otherwise harming the patient.

Fourth.—To provide apparatus for accomplishing the above, which is relatively simple in construction and operation and relatively inexpensive to produce and operate.

Fifth.—To provide novel electronic and radioactive devices especially adapted for use in conjunction with high potential, high frequency electrical therapy.

I have found that, by enveloping a patient in a high potential, high frequency electrical field in such a manner that no closed circuit is completed through his body, radioactive and other electronic and radiation phenomena can be used therapeutically with considerably greater effectiveness than if used alone. The exact reason for this is not known, nor is it known definitely which, the electric field or the radioactive phenomena, acts upon the other to produce the advantageous results. It is thought, however, that the electric field, permeating the body of the patient as it does, attracts the radioactive emanations or radiations and enables them to penetrate considerably deeper into the tissues and vital organs of the patient than would otherwise be the case. In any event, remarkable therapeutic results have been achieved by use of the invention in the treatment

of malignant tumors, arthritis, sinus infections, and various other diseased conditions.

The invention contemplates the use, in therapeutics, of high potential, high frequency electricity to produce diversified forms of radiant energy, such forms being those which have been found best suited, individually, to benefit various human ailments. In accomplishing this purpose, several special discharge tubes have been developed to serve as treatment electrodes, by means of which correspondingly different curative results are obtained. Throughout the practice of the invention, a prime consideration is that only one terminal of any particular circuit shall be in contact with a patient's body at one time, so there will be no flow of current through a closed circuit of which the patient's body is a part. Such a terminal, too, is usually non-heat producing, so there is no danger of burning. In cases where there is a tendency for a tube to produce X-rays or other injurious rays, these are filtered out.

The present application constitutes a continuation in part of a copending application filed by me November 15, 1940, which bears Serial No. 365,798 and is entitled "Method of and device for the therapeutic application of electric currents and rays," and which has now become abandoned.

In the accompanying drawings, which illustrate several embodiments of apparatus preferred for carrying the method of the invention into practice:

Fig. 1 represents a wiring diagram of a preferred embodiment of apparatus for carrying out the method of the invention in general therapeutic work, several independent treatment stations being provided;

Fig. 2, a top plan view of the novel corona regulator of Fig. 1, employed in the circuit to control and adjust the current and as a governor to safeguard the transformer;

Fig. 3, a vertical section taken on the line 3—3, Fig. 2;

Fig. 4, a vertical section taken centrally through one novel type of discharge tube used as a treatment electrode in the apparatus of Fig. 1;

Fig. 5, a horizontal section taken on the line 5—5, Fig. 4;

Fig. 6, a vertical section taken centrally through another novel type of discharge tube used as a treatment electrode in the apparatus of Fig. 1;

Fig. 7, a horizontal section taken on the line 7—7, Fig. 6;

Fig. 8, a vertical section taken centrally through a novel discharge tube used as a treating device in the apparatus of Fig. 1;

Fig. 9, a horizontal section taken on the line 9—9, Fig. 8;

Fig. 10, a fragmentary vertical section taken on the line 10—10, Figs. 8 and 9;

Fig. 11, a fragmentary view in vertical section, and drawn to a reduced scale, of a tub bath capable of use as a treatment station in the apparatus of Fig. 1;

Fig. 12, a view similar to that of Fig. 11, but showing a shower or vapor bath arrangement for the same purpose;

Fig. 13, a wiring diagram similar to that illustrated in Fig. 1, but fragmentary in nature, and of a somewhat different embodiment of apparatus;

Fig. 14, an elevation, partly in central vertical section, of a novel tube used in the apparatus of Fig. 13 in place of the corona regulator of Figs. 2 and 3;

Fig. 15, a top plan view, partly in horizontal section on the line 15—15, Fig. 14, of the tube of Fig. 14;

Fig. 16, a vertical section of another novel tube which may be used in place of the tube of Figs. 14 and 15;

Fig. 17, a vertical section taken on the line 17—17 of Fig. 16;

Fig. 18, a top plan view of still another novel tube which may be used in place of the tubes of Figs. 14 and 15 and of Figs. 16 and 17; and

Fig. 19, a vertical section taken on the line 19—19 of Fig. 18.

In accordance with the invention, provision is made for enveloping the patient in a high potential and, in certain instances, a high frequency electric field, and for applying to the patient, while so enveloped in the electric field, radiations and emanations having therapeutic value.

The apparatus of Fig. 1 is capable of administering various specific kinds of treatment, pursuant to the invention, at the several treatment stations provided. The treatment stations are indicated A, B, C, D, and E, respectively.

For supplying the high potential electric field, a suitable transformer is employed. This may be of any type capable of delivering high potential electricity, say from 10,000 to 30,000 volts. It is preferred, however, to utilize a conventional double magnetic circuit type of transformer, indicated at 10 in Fig. 1, having adjustable, laminated, magnetic shunts (not shown), the transformer being connected across an ordinary power line 11 charged with the customary 115 v. The output lines 12 from this transformer advantageously extend to the treatment stations A and B, respectively. The first secondary of the transformer 10 is preferably direct connected to the second secondary thereof. It is noted that this high potential electricity may be applied, without causing injury, direct to a patient who is not grounded. However, in order to safeguard the transformer 10 from damage by sparking across its output terminals, and to render the high potential electricity more suitable for therapeutic purposes, which is believed to include the automatic changing of the frequency to an extent which depends upon electrical characteristics of the patient's body, a governor or control device 14 is shunted across the leads 12.

This governor or control device 14 is a sparking condenser of high capacity embodying a multitude of spark gaps. A preferred embodiment of

this governor or control device 14 is illustrated in detail in Figs. 2 and 3.

As illustrated, the device comprises a cylindrical, electrically conductive plate 15 surrounded by a cylindrical dielectric 16. An outer cylindrical and electrically conductive element 17 surrounds the dielectric 16 exteriorly. It is provided with a multitude (for example, 250) of inwardly extending prongs 17*a*, which are advantageously formed by stamping out, and inturning, triangular portions of the electrically conductive element 17. The internal plate 15 preferably contacts the interior surface of the dielectric 16, but, in any event, should lie closely adjacent thereto. Likewise, the tips of the prongs 17*a* preferably contact the outer surface of the dielectric. The several elements are advantageously mounted in a plug-in base 14*a*, which is adapted to mate with a suitable receiving socket (not shown) carrying the required electrical connections. The internal plate 15 connects with one of the electric lines 12, while the external element 17 connects with the other electric line 12, as shown diagrammatically in Fig. 1.

It is preferable that the dielectric 16 be in the form of a closed tube or envelope, as shown, and be exhausted to vacuum condition. The multitude of sparking prongs 17*a* product a brush discharge.

Where the dielectric 16 is not a closed tube or envelope, it is preferred that it be of quartz.

The treatment station A is a discharge tube of a novel type, exemplified by the tubes illustrated in detail in Figs. 4 and 5 and Figs. 6 and 7. Either tube is plugged into the circuit of Fig. 1 at a suitably provided, single-terminal outlet. High potential electricity is, therefore, fed directly into the tube, which serves as an electrode. The tube also embodies radioactive material, which supplies radioactive emanations to the patient simultaneously with the electrical discharge.

As illustrated in Figs. 4 and 5, the tube or electrode may comprise an electrically conductive discharge element 20, having a supporting stem 20*a* and a major discharge cap or head 20*b*, which is preferably in the form of a thin, convex-concave plate. The head 20*b* may be spot welded to the end of the stem 20*a*.

The discharge element 20 is enclosed within a tube 21 of dielectric material, preferably glass, the stem 20*a* being fixed in the fused tongue portion 21*a* of the tube. The tube or shell 21 is fitted into an insulating base 22, provided with a single plug-in terminal 23, and an electrical connector 24 extends from the terminal 23 to the stem 20*a*.

The inside surfaces of the side walls of the tube or shell 21 are coated with a radioactive material, as at 25. The coating is conveniently made from uranium salts or powdered carnotite or other radioactive ore. The ends of the tube or shell are left uncoated.

Air is evacuated from the tube 21, and a small quantity of mercury introduced. The mercury is preferably triple-distilled to insure great purity. It is preferred that argon or like inert gas be also introduced.

Since the tube just described is plugged into the circuit of Fig. 1, the discharge element or cathode 20 is charged with high potential electricity, and, in its capacity of a treatment station in the apparatus of Fig. 1, serves as an electrode to similarly charge the patient. The patient is insulated from the ground, and the tube is applied directly to the afflicted part of his body, preferably in close contact with the body.

Because of the construction of the tube, radiation of a radioactive nature is also directed against the patient through the uncoated top end of the tube. This radiation has been found to differ somewhat from the radioactive emanations discharging from the side walls of the tube, and is thought to comprise rays lying close to X-rays on the radiation spectrum. These rays appear to have a definite healing value, and to lack the injurious nature of X-rays. Where a predominantly radioactive emanation treatment is desired, the side walls of the tube are placed against the body of the patient.

Best results are obtained when the discharge element or cathode **20** is made of an alloy metal compounded from copper, lead, sulphur, and, if desired, aluminum. The relative percentages of the several ingredients may vary considerably, but a satisfactory mixture comprises 5.0% copper, 55.0% lead, 30.0% sulphur, and 10.0% aluminum. Should aluminum not be used, the difference may be made up by additional copper.

In preparing the alloy, the copper and aluminum are heated to a molten state, after which the sulphur is added while stirring the mixture. After cooling, the mass is again melted, and the lead, in a molten state, is mixed with it, the molten mass being thoroughly stirred. This new mass is then cooled, being later reheated, and, while hot, rolled to make it ductile, so it can be shaped into the desired forms.

The discharge tube or electrode of Figs. 6 and 7 is similar to that of Figs. 4 and 5, having an enclosing tube or shell **26** which is evacuated. A cathode discharge element **27** is positioned within the shell, being fixed in the tongue portion **26***a*. A conductor **28** connects the stem **27***a* of the element **27** with a plug-in terminal **29**, which extends outwardly of the base **30**. The cap or head **27***b* of the element **27** differs from the cap or head **20***b* of the electrode of Figs. 4 and 5, in that it is spherical in form and hollow. It has an opening **31** formed at its top, contiguous with the top inside surface of the tube **26**. A quantity **32** of radioactive material, which may be the same as used for the coating **25** of the electrode of Figs. 4 and 5, is introduced into the tube or shell **26**, along with a relatively small quantity of mercury, before the tube is sealed tight. Such material **32** is preferably powdered or granulated, and is shaken into the hollow of the head **27***b* through the opening **31** before any given treatment is commenced. The mercury is provided primarily as a getter, and does no harm if shaken into the head **27***b* along with the radioactive substance. The mercury also tends to produce a vapor in the tube, which aids in the operation thereof. As in the case of the electrode tube of Figs. 4 and 5, this tube may have a radioactive coating **33** covering the inner surfaces of its side walls.

The treatment station B of Fig. 1 differs from the treatment station A only in the fact that a condenser **34** is interposed in the electric supply line **12**.

The treatment station C of Fig. 1 differs from the stations A and B only in the fact that the high potential electricity is supplied from the supply line **12** through an inductance **35**.

The treatment station D utilizes a germicidal discharge tube, a preferred form of which is illustrated in detail in Figs. 8, 9, and 10. The high potential electricity is taken by induction from the particular supply line **12** concerned. For this purpose, an induction coil **36** is provided, tapping the line **12** at **37**. A pair of leads **38** from an ordinary 115 v. supply source extend to a plug-in socket connection for the germicidal tube, one of the leads passing through a glass tube **36***a*, Fig. 1, which is disposed within and extends along the length of the induction coil **36**. Thus, high potential electricity is impressed, by induction, upon the ordinary current flowing through the particular lead **38** concerned.

The germicidal discharge tube of Figs. 8, 9, and 10 has a pair of discharge terminals **40** and **41**, respectively, positioned in an evacuated tube or envelope **42**, and electrically connected with plug-in terminals **40—1** and **41—1**, respectively, by means of stems **40***a* and **41***a*, respectively. The tube or envelope **42** and plug-in terminals are mounted in a conventional base **43**. It is preferred that insulating material **44**, such as a ceramic sleeve, cover the major portions of the stems **40***a* and **41***a*. A piece of lithium metal **45**, see particularly Fig. 10, is advantageously secured to the stem **40***a* adjacent the discharge terminal **40** to act as a getter. It may, however, be placed at any other convenient location in the tube. It is preferred that the discharge terminals **40** and **41** be formed of the special alloy previously described. Argon or other suitable inert gas is preferably injected into the tube or envelope **42**, as is, also, a small quantity of mercury. The mercury, by vaporizing, aids electrical arcing between the discharge terminals. As will be noted, the high potential electricity induced in the one lead **38** will manifest at the upper discharge terminal **40**, and will charge the patient simultaneously with the discharge into his body of germicidal rays from the tube.

The treatment station E embodies the tube of Figs. 8, 9, and 10, as above described, but impresses the high potential electricity directly on the patient instead of passing it first through the tube. For this purpose, a discharge device **50**, in the form of a soft, flexible pad in which a coil **50—1** is embedded, taps one of the high potential electric lines **12**. This pad **50** is wrapped around the patient's body adjacent the afflicted portion thereof, thus charging the patient. Any other electrode capable of charging the patient with high potential electricity may be used in place of the pad **50**. The germicidal tube has its terminals **40—1** and **41—1** plugged into a suitable plug-in socket connected to leads **52** which extend to an ordinary 115 v. source of supply. The high potential electricity with which the patient is charged is induced into the germicidal tube, thereby further activitating the discharge therefrom. A certain beneficial discharge from this germicidal tube will be had by induced activation alone, it being unnecessary, in such instances, to plug the tube into the 115 v. line.

Other types of germicidal and discharge tubes may be used in place of the tube of Figs. 8, 9, and 10, as, for instance, the well known infra-red and ultra-violet lamps, to produce results surpassing those ordinarily attained by the use of such infra-red or ultra-violet lamps apart from the apparatus of the invention.

It should be remembered that the patient is insulated from the ground while being treated at any of the treatment stations of the invention.

Figs. 11 and 12 show how a patient is treated, pursuant to the invention, while immersed in an electrically conductive fluid bath. In Fig. 11, a bath tub **53** is insulated from the ground by a layer of insulation 54. A treatment electrode of the type shown in any of the figure groups 4 and 5, 6 and 7, and 8, 9, and 10 is positioned to charge the fluid of the bath with high potential electricity, as well as to discharge healing radiations and emanations into the patient. The particular electrode illustrated is diagrammatic in form and is designated 55. It may be connected into the circuit of Fig. 1 as shown at any of the treatment stations A, B, C, and D. In Fig. 12, a shower or vapor stall 56 is insulated from the ground by a layer of insulation 57. A plurality of treatment electrodes are designated 58, respectively. These correspond to the treatment electrode 55 of Fig. 11. A water spray or vapor, such as steam, may be admitted to the stall 56 in any well known manner (not shown), thus enveloping the patient during treatment.

Another embodiment of apparatus, pursuant to the invention, is illustrated diagrammatically by the wiring diagram of Fig. 13. While no treatment stations are shown, those provided are identical with the several treatment stations designated A, B, C, D, and E in Fig. 1. The distinction in this embodiment of apparatus resides in the fact that a special generator of high frequency electricity is provided in the system.

A transformer 60 has its input terminals connected across an ordinary 115 v. electric power line 61. Electrical conductors 62 and 63 lead from the respective output terminals of the transformer to a high frequency generator of the Oudin coil type, indicated generally at 64, a variable condenser 65 being interposed in the line 62, and the circuit being grounded at 66. Output conductors 67 and 68, leading from the high frequency generator 64, provide connections for the several treatment stations in the same manner as illustrated in Fig. 1.

The transformer 60 may be any ordinary high voltage type. A governor or control device 69 is shunted across the conductors 62 and 63.

In the illustrated instance, the governor or control device 69 preferably takes the form of a vacuum tube, having the construction shown by Figs. 14 and 15, Figs. 16 and 17, or Figs. 18 and 19. These tubes all possess high capacity, and include elements effecting a brush discharge. They serve, as does the device 14 of Figs. 2 and 3.

The tube of Figs. 14 and 15 embodies an outer shell or envelope 70 of insulating material such as glass, a plastic, or fiber coated with shellac. Inside the shell 70 is a bi-cylindrical element 71 formed of electrically conductive material. Separating element 71 from the enclosing shell 70 are spacers 72 made of rubber, Bakelite, or other insulating material. Inter-fitting with the element 71 is a second electrically conductive, bi-cylindrical element 73, the two elements being separated by a dielectric 74. Inwardly of the element 73, and separated therefrom by a dielectric 75, is a corrugated, cylindrical element 76. The shell or envelope 70 is secured in an insulating base 70—1, provided with plug-in terminals. One of the terminals, designated 77, is electrically connected with the element 71, while another, designated 78, is electrically connected with the corrugated element 76. These two terminals connect with the conductors 62 and 63, as illustrated in Fig. 13, and the brush discharge takes place at element 76.

Under certain circumstances, it is desirable that the outer shell 70 be made of quartz glass, and that a filament 79 be provided, the filament being heated by connection, through plug-in terminals 80 and 81, with a source of low voltage heating current (not shown). Plug-in terminal 82, which is electrically connected with element 73, may be used instead of or in connection with the terminal 77, since element 73 acts in a manner similar to element 71. A getter 83 of suitable material, and an insulating and reflecting shield 84 may be provided, as shown. While the tube may have either a high or a low vacuum condition, or may be filled with an inert gas, I have also found it advantageous to fill the tube with a moist vapor. The tube acts as an oscillator for electric currents, and has an enormous capacity, a capacity many times that of a condenser of approximately equal size.

The tube of Figs. 16 and 17 comprises an outer shell or envelope 85, which may be made of metal, glass, or fused quartz. This shell is mounted in an insulating base 86. Inside the shell 85 is a metal plate 87, and, spaced apart therefrom, a corrugated metal plate 88. A plug-in terminal 89, which extends from the base, is electrically connected with the plate 87, and a second plug-in terminal 90 is electrically connected with the corrugated plate 88. These terminals are adapted to connect, through a suitable socket, with the electrical conductors 62 and 63 of Fig. 13.

Under certain conditions of use, it is desirable to have other elements in the tube. These are provided, and may be utilized or not as occasion warrants. A filament 91 is disposed between the plates 87 and 88. It is electrically connected with the two plug-in terminals 92 and 93, which are adapted to be connected to a source of low voltage heating current (not shown). A slit screen, comprising shields 94 and 95, with apertures 96 extending therethrough, is disposed adjacent that side of corrugated plate 88 which is remote from plate 87. The apertures 96 are in alignment with each other, and the shields 94 and 95 are made of lead or other material capable of screening off X-rays. Between shields 94 and 95 is a sheet 97 of material which is readily permeable to X-rays. Within the shell 85 there is also mounted a shell or envelope 98 of glass, quartz glass, or similar material, having a portion 98a which is ground like a lens and directed toward the slit screen. This shell 98 really constitutes a tube within a tube. A filament or cathode 99, comprising electrically conductive legs 99a and 99b and an electron-emitting portion 99c, is disposed within the shell 98, plug-in terminals 100 and 101 being electrically connected to the respective legs 99a and 99b. A bombardment element 102 is disposed within the shell 98 opposite the portion 99c of cathode 99. Within the shell 85, but outside the shell 98, is a reflector 103 directed toward the slit screen.

The tube of Figs. 18 and 19 is essentially the same as the tube of Figs. 16 and 17, being equipped with a shell or envelope 105, a base 106, a plate 107, and a corrugated plate 108, the two plates being connected to plug-in terminals 109 and 110, respectively, which are adapted to connect electrically with the conductors 62 and 63 of Fig. 13. There is a filament 111 and an inner shell or envelope 112, but no slit screen. Instead of a lens portion being provided on the inner shell 112, a partition 113 of lens formation is disposed between the inner shell and the corrugated plate 108. It is fused to the walls of the outer shell 105. Within the inner shell 112 is a filament or cathode 114, which corresponds to the similar element 99 of the tube of Figs. 16 and 17. A reflector 115 is directed toward the lens partition 113.

Reverting now to Fig. 1, there is another advantageous way of treating a patient pursuant to the invention. As shown at Y, a foot pedestal 120 may be provided for making the patient a part of a condenser. The pedestal comprises an electrically conductive plate element or electrode 121, connected electrically with one of the high potential lines 12, and covered by an insulating platform 122 upon which the patient rests his feet while being treated at any of the previously described treatment stations A, B, C, D, or E. The electrode 121 and insulating platform 122 are conveniently mounted in a frame 123, which insulates the plate from the ground. The insulating platform 122 is made of a high quality insulating material, such as first grade hard rubber. In certain instances it is desirable that the device be made in other than foot-pedestal form. For instance, it may be of cylindrical formation for use in a bed against any part of the patient's body.

If desired, the patient may be charged with the high potential electricity by direct contact with a metal or electrically conductive electrode in place of the pad 50 of treatment station E, or of the tube electrodes.

The invention has been described in the foregoing with sole reference to its use for therapeutic purposes. It should be noted, however, that inorganic matter may also be treated to advantage pursuant to the method and with the apparatus of the invention. It has been found that metals, for example, lead, have changed physical properties after treatment in accordance with the above. In instances where the invention is not being used therapeutically, it is not always necessary to insulate the subject from the ground.

Whereas this invention is here illustrated and described with respect to particular specific embodiments thereof, it is to be understood that various changes may be made in such specific embodiments and various other embodiments may be utilized by those skilled in the art without departing from the spirit and generic scope of the invention as set forth herein and in the claims which here follow.

Having fully described my invention, what I claim is:

1. Apparatus for applying radiant energy therapeutically, comprising means for producing high potential, high frequency electricity; a high capacity sparking condenser, and a treatment electrode connected in circuit with the foregoing, said treatment electrode including a discharge element adapted to charge the patient with said high potential, high frequency electricity, and radioactive means adapted to discharge radioactive emanations into said charged patient.

2. Apparatus in accordance with claim 2, wherein the sparking condenser is in the form of a vacuum tube of high capacity having mutually spaced capacity elements adapted to produce a corona discharge.

3. Apparatus for applying radiant energy therapeutically, comprising means for producing high potential, high frequency electricity; a high capacity sparking condenser; and a treatment device connected in circuit with the foregoing, said treatment device including discharge means adapted to charge the patient with said high potential, high frequency electricity, and radiating means adapted to discharge radiations into the charged patient.

4. Electrical treatment apparatus, comprising a high capacity sparking condenser; a treatment outlet electrically connected to said condenser; and means for electrically connecting said condenser to a source of high potential electricity.

5. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a high capacity sparking condenser electrically connected across the high potential output terminals of said transformer; and a treatment outlet electrically connected to said condenser.

6. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a high capacity sparking condenser electrically connected across the high potential output terminals of said transformer; and a plurality of treatment outlets independently electrically connected to said condenser.

7. Electrical treatment apparatus, comprising a transformer for producing high potential electricity; a treatment electrode electrically connected to one of the output terminals of said transformer; an electrical conductor sheathed by insulation electrically connected to the other of the output terminals of said transformer and disposed adjacent said treatment electrode so the subject to be treated may be placed between and in contact with the two; and a high capacity sparking condenser connected across the said outlet terminals of the transformer.

8. In electrical treatment apparatus equipped with means for the supply of high potential electricity and a treatment electrode, a high capacity sparking condenser electrically connected between the said supply means and the said treatment electrode.

THOMAS H. MORAY.

REFERENCES CITED

The following references are of record in the file of this patent:

UNITED STATES PATENTS

Number	Name	Date
628,351	O'Neill	July 4, 1899
647,687	Topham	Apr. 17, 1900
765,470	Friedlander	July 19, 1904
950,842	Davis	Mar. 1, 1910
1,156,317	Santos et al.	Oct. 12, 1915
1,193,018	Howard	Aug. 1, 1916
1,466,777	Winkelmann	Sept. 4, 1923
1,590,930	Falkenberg	June 29, 1926
1,655,783	Gallois	Jan. 10, 1928
2,010,018	Hodnette	Aug. 6, 1935
2,073,428	Schmid	Mar. 9, 1937
2,126,070	Wappler	Aug. 9, 1938
2,192,638	Dixon	Mar. 5, 1940

OTHER REFERENCES

Tousey; Medical Electricity and Roentgen Rays (1910), pp. 493 to 498. Copy in Division 55.

Two Patents of William J. Hooper

PATENTED OCT 5 1971

3,610,971

FIG.1

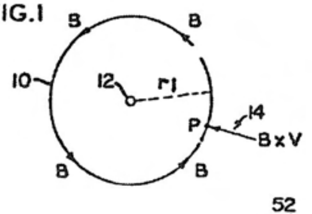
B
B
10
12
r1
P
14
B x V
B
B

FIG. 3

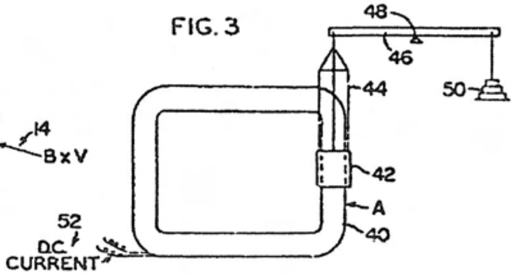
48
46
44
50
42
A
40
52
D.C.
CURRENT

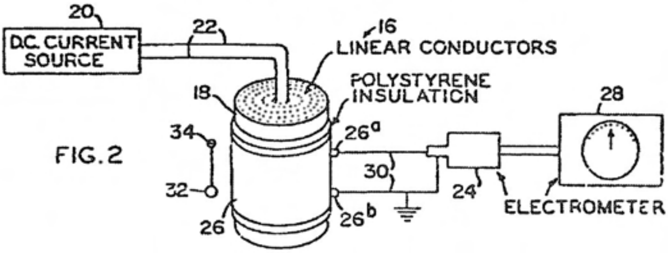
20
D.C. CURRENT SOURCE
22
16
LINEAR CONDUCTORS
18
POLYSTYRENE INSULATION
28
34
26a
30
32
24
ELECTROMETER
26
26b

FIG. 2

FIG. 4

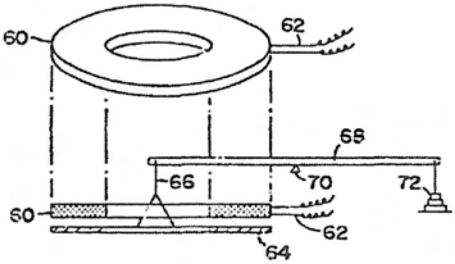
60
62
68
66
70
72
60
62
64

FIG. 5

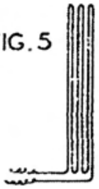

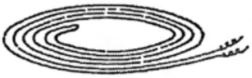

FIG. 6

FIG. 8

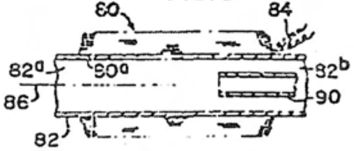
80
84
82a
90a
82b
86
90
82

FIG. 7

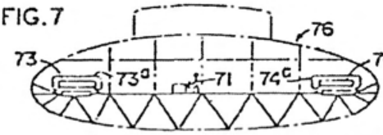
76
73
73a
71
74c
74

INVENTOR
WILLIAM J. HOOPER

BY
Oldham & Oldham
ATTORNEYS.

UNITED STATES PATENT OFFICE

CERTIFICATE OF CORRECTION

Patent No. 3,610,971 Dated October 5, 1971

Inventor(s) William J. Hooper

It is certified that error appears in the above-identified patent and that said Letters Patent are hereby corrected as shown below:

Col. 1, line 14, after "conductor" insert -- with -- and delete "electric charges".

Col. 1, line 45, "10^{17}" should be -- 10^{-7} --.

Col. 2, line 4, "nd" should be -- and --.

Col. 2, line 25, "3.6×10^{12}" should read -- 3.6×10^{-2} --.

Col. 2, line 44, "motion" should be -- motional --.

Col. 4, line 7, "anitgravitational" should read -- antigravitational --.

Col. 4, line 22, "he" should be -- the --.

Col. 5, line 20, "out" should be -- our --.

Col. 6, line 26, after "conductor" delete "o" and insert -- to --.

Signed and sealed this 16th day of May 1972.

(SEAL)
Attest:

EDWARD M. FLETCHER, JR.
Attesting Officer

ROBERT GOTTSCHALK
Commissioner of Patents

United States Patent

[11] 3,610,971

[72] Inventor William J. Hooper
Sarasota, Fla.
[21] Appl. No. 816,356
[22] Filed Apr. 15, 1969
[45] Patented Oct. 5, 1971
[73] Assignee Electrodynamic Gravity, Inc.

[54] ALL-ELECTRIC MOTIONAL ELECTRIC FIELD GENERATOR

5 Claims, 8 Drawing Figs.

[52] U.S. Cl. 310/10
[51] Int. Cl. H02n 11/00
[50] Field of Search 310/10, 11; 324/109; 336/181, 225; 338/284, 297

[56] References Cited

UNITED STATES PATENTS

2,553,875 5/1951 Shaw 338/284 X
3,259,784 7/1966 Vogel 338/297 X

OTHER REFERENCES

Textbook—The Electromagnetic Field in Its Engineering Aspects by G. W. Carter. 2nd Edition, 1967, Published by Longmans—48 Grosvenor St. London W.1, pp. 168, 169, 170 (Copy in Central Library, Washington, D.C.)

The Feynman Lecture on Physics by Feynman, pages 13–6 to 13–11; Addison-Wesley Publishing Co., New York.

Primary Examiner—D. X. Sliney
Attorney—Oldham & Oldham

ABSTRACT: This invention relates to an all-electric generator yielding a motional electric field in the space surrounding the device, but requiring no mechanical movements of its parts in generating this field. The Theory underlying the production of such a field postulates that each moving electron constituting the current in a linear conductor carries with it a loop of magnetic field energy about it.

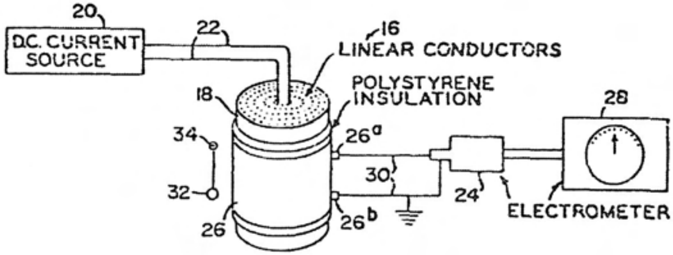

ALL-ELECTRIC MOTIONAL ELECTRIC FIELD GENERATOR

This application is an extension of the teachings presented in my earlier filed pending applications, Ser. Nos. 722,587, filed Apr. 19, 1968 and 803,187, filed Feb. 28, 1969.

PRIOR HISTORY

The detection of a magnetic flux about a current-carrying conductor was first made in 1820 by Hans Christian Oersted. This discovery unified the then separate sciences of electricity and magnetism. My experimental discovery that this magnetic flux actually moves along the conductor electric charges the electric charges constituting the current has great promise of now unifying the three known fields of electricity, magnetism and gravitation.

For a better understanding of the invention, reference should be had to the accompanying drawings wherein:

FIG. 1 is a schematic illustration of a wire containing flowing electrical current indicating the magnetic field generated with respect thereto;

FIG. 2 is a perspective block diagram schematic of a preferred embodiment of the invention;

FIGS. 3 and 4 illustrate modifications of the generator of FIG. 2 adapted to techniques for studying and utilizing gravitational and antigravitational phenomena;

FIG. 5 illustrates the winding of the wire in the device of FIG. 2;

FIG. 6 illustrates the winding of the wire in the device of FIG. 4;

FIG. 7 illustrates the incorporation of a plurality of the devices of the invention into a spacecraft where they might be appropriately utilized; and

FIG. 8 illustrates a coil wound embodiment useful for producing the separation of oppositely charged ions in gases or liquids.

BACKGROUND INFORMATION

The law of Biot and Savart yields the magnetic flux density B (webers/m.^{2}) at a point distant r meters from the conductor carrying a current of I amperes and is given by

$$B=\mu_o I/2\pi r \qquad (1)$$

where $\mu_o=4\pi\times10^{-7}$ webers/amp. meter. With reference to FIG. 1 of the drawings, this flux consists of circular magnetic loops directed counterclockwise, as shown by arrow 10, for current directed in a linear conductor wire 12 at right angles to this paper and outwardly toward the reader.

Electrons producing this flux will be moving perpendicularly into the paper, carrying the flux B with them, at a velocity V meters/sec. Thus, at any point P, distant r_1 from the long linear conductor 12, there will be existent a motional electric field of intensity $E=B\times V$ volts/meter directed radially inward toward the wire in the plane of the paper, as indicated by vector arrow 14, where $B\times V$ is the vector cross product of B and V.

PREFERRED EMBODIMENT

The preferred embodiment of the invention is illustrated in FIG. 2 and consists in part of many insulated linear conductors indicated generally by numeral 16, all parallel, packed and held in close proximity, and connected in series, so as to form a solid package 18 preferably cylindrical in shape. When a constant direct current from a source 20 is caused to flow through wires 22 this device with its axis in the vertical direction, half of the linear conductors 16 will have current flowing upward and half downward. It will be a noninductive device, since surrounding it the magnetic flux due to the upward currents will be superimposed on the flux due to the downward currents, so that no magnetic flux can be measured. On the well recognized "Principle of Superposition of Fields," each one exists however, and acts as though the other was absent.

When the package 18 of linear conductors has a circular cross section there will exist two systems of circularly directed magnetic flux, one directed clockwise and the other counterclockwise as seen from above, and one set moving upward and the other set moving downward. What is of capital interest however is that each of these two systems give rise to equal radially and inwardly directed $B\times V$ motional electric fields distributed laterally in horizontal planes in the space about the cylindrical device. Here we have a field which is not magnetic, is not electrostatic, and which I have shown experimentally to be immune to electrostatic shielding. No such spacially distributed field is known to Physics today except that of gravity. I have shown theorectically that this field should act attractively on electrically neutral matter. I have shown experimentally that the intensity of this field can be measured with the aid of an electrometer system 24, 28 driven by the potential difference induced electromagnetically in the conductors 30 connecting the electrometer to the capacitor plates 26a and 26b of the capacitor 26. The electrometer head 24 actuates a visual dial indicator 28.

At cryogenic temperatures the drift velocity of the electrons in the linear conductors will be greatly enhanced. Theoretical calculations indicate that this velocity in copper at 20° C. is of the order of 3.6×10^{-2} cm./sec. in a current density of 480 amp./cm^2. The random velocity of free electrons in conductors is theoretically estimated at 10^8 cm./sec. At superconducting temperatures when the electrical resistance approaches zero, it is generally believed that the drift velocity of electrons approaches the value of 10^8 cm./sec. This velocity is 10^{10} times greater than the velocity at 20° C. (See "Physics," Halliday and Resnick, pages 681 and 691, J. Wiley, 1962). This indicates that the intensity of the $B\times V$ field about the device, herein described, might possible be enhanced 10^{10} times at or near the absolute zero of temperature. Hence, the invention contemplates that each of the devices described hereinafter will preferably be operated at or below the critical temperature at which the conductor used becomes superconducting. As the state of the art advances new conducting materials will undoubtedly become available with critical temperatures very considerably above the absolute zero.

Not only does the device that has been described in FIG. 2 afford apparatus hitherto unavailable for the generation of a spacial distribution of the motion electric field, but it affords means for investigating the properties of this field in connection with its relation to gravity and antigravitational phenomena. In addition, it affords for the first time, a direct experimental method for determining the electron drift velocities and mobilities in linear conductors made of different materials.

When the cylindrical capacitor 26 is charged by the $B\times V$ field and its potential difference (P.D.) measured, it is determined by the equation

$$\text{P.D. (volts)} = B\times V\cdot l \qquad (2)$$

where l is the distance between the thin capacitor plates. The direct measurement of V meters/sec. can then be made

$$V=\text{P. D. (volts)}/ Bl \text{ m./sec.} \qquad (3)$$

where B is in webers/m.^{2}.

This device will exhibit an attractive force on electrically neutral matter, better defined as a differential force; i.e. the difference between a pull and push, in which the pull is greater. (See reference to general theorem, page 125, The Mathematical Theory of Electricity and Magnetism, by Sir James H. Jeans) The $B\times V$ field is an electrical field acting on all the elementary charges comprising the atoms of matter, electrons, protons and even neutrons, as I believe neutrons also consist of electrons and protons in close bond. The $B\times V$ force being relatively weak does not ionize the atoms of matter, but being immune to shielding, electrically polarizes them. This causes a slight elongation of some of the electrons orbits within the atoms in the direction of the field. The positive charges move slightly into the more intense region of the converging, radially directed field while the negative charges move slightly further way from the device into the weaker portion of the field. Thus the differential force between the pull and the push upon the component parts of an atom results in the polarization of the atoms and a resultant attractive force, equivalent, I believe, to that of gravity.

A small, electrically neutral, simple pendulum 32 suspended close to one side of my device of FIG. 2, about pin 34, should thus undergo a slight measurable deflection under a fairly strong $B\times V$ motional electric field. At cryogenic temperatures this attractive force should be greatly enhanced. When the device is lying horizontally, a gravity meter placed centrally over it should give a measurable reading in milligals.

Modifications of my generator such as are shown in FIGS. 3 and 4 show diverse methods for studying gravitational and antigravitational phenomena which the invention makes possible. FIG. 3 illustrates a rectangular coil 40 made up of a single wire passed many times in the coil configuration so as to provide a large leg A preferably having a circular cross section. However, any cross-sectional configuration would meet the objects of the invention as long as the wires are parallel to each other and closely packed. About one leg of the coil, a cylindrical shell 42, made of electrically neutral, conductive or nonconductive material, is suspended by a carrier frame 44 from one end of a sensitive beam balance 46. The beam pivots about a support bar 48 and carries a load 50 to normally counterbalance the weight of the shell 42. When a fairly strong DC current I is fed through lines 52 and is built up in the coil, the flux density B is generated in the region of the cylindrical weight whose walls are a distance *r* from the central axis of the coil and *n* is the number of linear conductors in the coil.

$$B=\mu_0 nI/2\pi r \text{ (webers/meter}^2\text{)} \qquad (4)$$

If the current I is suddenly cut off the flux B will rapidly collapse. This flux in the act of collapsing should cross the walls of the suspended cylinder with a considerable inwardly directed horizontal velocity of V meters/sec. This should yield a very considerable vertical $B\times V$ in the walls of the cylinder directed either up or down, depending on the direction of the current in the coil.

In this device the motional electric field $B\times V$ is not convergent, but is uniformly vertical, hence the pull and push exerted on the electric components of the atoms in the cylinder will be equal. In other words the action of this field will either enhance or decrease the state of polarization already existing within the cylinder and produced by the gravitational field of the earth. If the polarization is increased, then the gravitational pull of the earth's field should be greater. If the polarization is decreased then the weight of the cylinder should be less. It is conceivable that the polarization could be reduced to zero. In this case, the object should be weightless. If the intensity of the depolarizing $B\times V$ field is still further increased, then polarization in the opposite direction to that produced by the earth's field should be achieved. In this event the earth's gravitational field would act repulsively on the cylinder. This would be antigravity in its truest sense.

The apparatus I have just been describing acts only momentarily (during the time interval in which the flux is collapsing) to change the weight of the cylinder shell 42 (shown in FIG. 3).

It is desirable to have an apparatus with which to experimentally study gravitational and antigravitational phenomena under steady continuous conditions. In FIG. 4 of the drawings is shown another modification of the device shown in FIG. 2. Here I utilize a noninductive coil wound in the form of a solid, low, thick walled, hollow, right circular cylinder, and indicated generally by numeral 60. DC current is supplied through lines 62 to the coil 60. In the region directly beneath this coil disk 60 I place a similarly shaped nonferrous disk 64, suspended by a carrier 66 to the end of a beam 68. The beam 68 is pivotal about a fulcrum 70 and is counterbalanced by a load 72. For positions of disk 64 below coil 60, the $B\times V$ field can only be directed vertically upward when a continuous DC current is flowing in the wires composing the disk 60. It should be noted that this field is practically uniform and parallel close to the disk, hence it is not attractive in its action on neutral matter. Its action on such matter can only change the state of electrical polarization in the matter. Any change thus effected in the state of polarization should cause the earth's gravitational field to act less intensely, on any object placed close to the under surface of the device, causing it to weigh less, become weightless, or to actually be accelerated upward by the anitgravitational action of the earth's field.

FIG. 5 of the drawings illustrates the manner in which the noninductive device in FIG. 2 is wound with wire 22, by turning the linear conductors back on themselves through a 180° turn. FIG. 6 illustrates the way the noninductive windings of the device illustrated in FIG. 4 is achieved. One very long insulated wire is simply turned back on itself at its midpoint and wound double into a solid coil held together in the shape illustrated.

FIG. 8 illustrates a modified embodiment of the invention which comprises a wire coil 80 either simply wound, or noninductively wound, around a hollow tube 82. Electrical current is passed to the coil 80 through wires 84, and at low temperature this coil will produce a radially, outwardly directed, internal $B\times V$ field from the axis 86 of the cylinder to the internal circumference 80*a* of the coil 80. The electrical wire making up coil 80 is preferably a ribbonlike conductor which is coated with any appropriate electrical insulation. A minimum space between adjacent turns with concentricity between adjacent turns is the preferred winding embodiment for coil 80. As many courses of turns as deemed appropriate to generate field strength desired will meet the objects of the invention. The tube 82 is preferably made of material having little or no magnetic permeability. Both the coil 80 and tube 82 are stationary.

An apparatus of the type illustrated in FIG. 8 might be utilized to effect separation and concentration of fluids. For example, in the desalinization of water, a $B\times V$ electric field generated by the passage of the current through the coil 80, acts upon the disassociated ions of the sodium chloride in solution. Hence, by introducing a flow saline water at end 82*a* of tube 82, and providing a concentric separating tube 90 at end 82*b* of tube 82, increments of fluid which are concentrated with chlorine and depleted of sodium may be drawn off through the stationary tube 90, while the increments of fluid which are concentrated with sodium and depleted of chlorine may be drawn off through the space between the exterior walls of tube 90 and the interior walls of tube 82.

FUTURE APPLICATIONS

This country's space projects have become expensive beyond the ability of the average man to comprehend. Achieving lift by means of costly propellants will some day soon be seen as akin to primitive man's use of awkward clubs as a means of exerting force. In a very short interval of time an intense $B\times V$ field should effect the reversal of the gravitational polarization of an object. Once achieved, only a very small expenditure of energy would be required to hold the polarization. Energy from the atoms of the earth would supply the lift and propulsion through the medium of the earth's gravity field. The most formidable problem would appear to be the problem of holding the $B\times V$ generator at cryogenic temperatures. In view of the overall possibilities of this invention, this problem does not appear to be incapable of a solution. A breakthrough in finding superconducting materials or even near superconducting at slightly elevated temperatures from those presently required would greatly aid in the solution of this problem.

FIG. 7 of the drawings illustrates in outline, the future possible application of this device, obtaining regulated lift by an operator 71 adjusting the current into the coils 73 and 74 which incorporate the structural features of the devices of FIG. 4. The antigravity gravity control features of the coils 73 and 74 thus control the lift and movement of vehicle 76. Suitable cryogenic generators 73*a* and 74*a* are associated with each of the coils 73 and 74 to provide the cooling thereof to as near absolute zero as possible to give the enhanced current passage characteristics necessary to the objects of the invention.

While I have demonstrated that AC current can be used in my device, it appears at present that DC current is much to be preferred.

This invention could be uniquely adapted to a communication system from one side of a solid barrier to the other side whether this barrier be solid concrete or battleship armor, by pulsing the $B\times V$ generator coil with signals, Morse Code or even voice modulated signals to be picked up by a capacitor-electrometer circuit on the other side of the barrier. Secret communications could be carried on and with the electrometer circuit as shown in FIG. 2 connected to a relay radio broadcasting unit, so that one might conceivably broadcast by radio from the interior of a solid metallic enclosure.

It is conceivable that the invention might be useful in the highly specialized art of concealed detonation devices. A detonating circuit connected to the capacitor electrometer receiving circuit of my device such as shown in FIG. 2 could be actuated through a solid barrier by our $B\times V$ generating coil placed on the other side of such a barrier.

It is conceivable that for some specialized purposes, it might be highly desireable to transmit electrical power from one side of a fixed solid barrier, concrete or armor plate, etc. by pulsing, DC or AC, in our $B\times V$ coil on one side of such a barrier. A resonant LC circuit could pick up these $B\times V$ pulsations and supply a source of electrical power on the other side of the barrier, in a manner somewhat analogous to my electrical receiving methods outlined in U.S. Pat. Application, Ser. No. 803,187, identified above.

In accordance with the patent statutes, only the best known embodiments of the invention have been illustrated and described in detail, but it is to be understood that the invention is not limited thereto or thereby, but that the scope of the invention is defined in the appended claims.

What is claimed is:

1. Apparatus for demonstrating an electrical phenomenon comprising,

a coil made from an electrically conductive material having a very large plurality of sections compacted in close proximity arranged in substantially complementary relation around a common central axis, where the coil forms a solid cylindrical shape with the sections parallel to the axis thereof, and

means to pass current through the coil so that about half the sections will pass current in a direction opposite to the other sections to cancel magnetic flux.

2. Apparatus according to claim 1 where adjacent sections are electrically insulated from each other, and the coil comprises a single wire bent back on itself to make the sections.

3. Apparatus according to claim 1 which includes means to increase the passage of electrical current through the conductor.

4. Apparatus according to claim 3 where the means to enhance is a cryogenic generator which reduces the temperature of the conductor o or below its critical temperature at which it becomes superconductive.

5. Apparatus according to claim 1 which includes an elongated hollow tube having an axis, the conductor formed around the tube.

Patented April 11, 1972 3,656,013

4 Sheets-Sheet 1

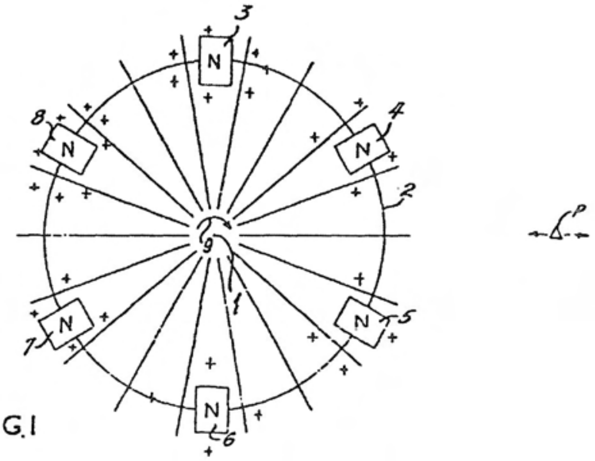
3
8
N
4
2
P
9
1
5
7
6

FIG. 1

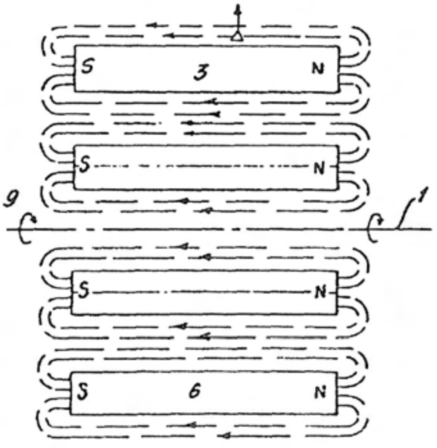
S
3
N
S
N
9
1
S
N
S
6
N

FIG. 2

INVENTOR:
WILLIAM J. HOOPER

BY Oldham & Oldham
ATTORNEY.

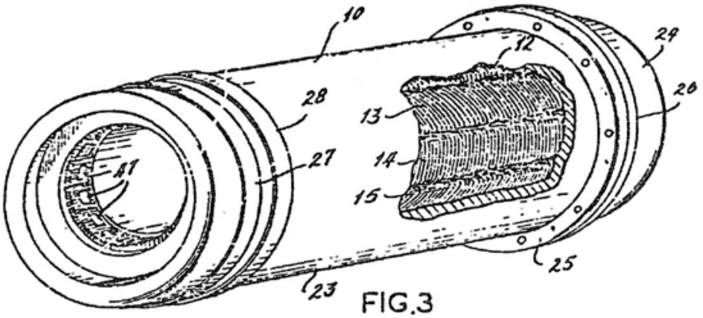
10
12
24
26
28
13
27
47
14
15
25
23

FIG.3

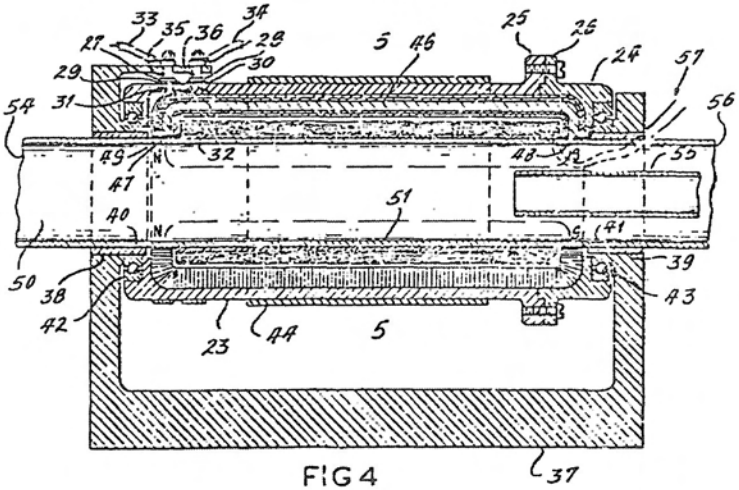
33
35
36
34
27
28
29
30
31
5
46
25
26
24
57
54
56
49
32
48
55
47
N
51
N
S
S
40
41
50
38
39
42
43
23
44
5
37

FIG.4

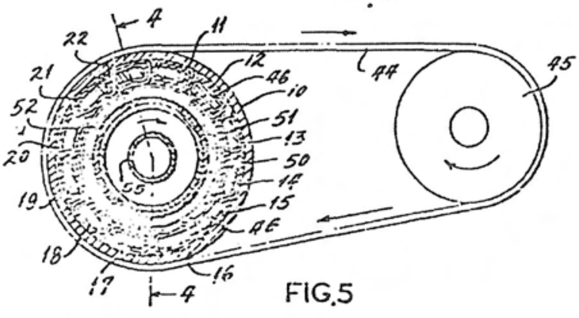
4
11
22
12
46
10
21
52
51
13
20
50
14
55
19
15
46
18
17
16
4
44
45

FIG.5

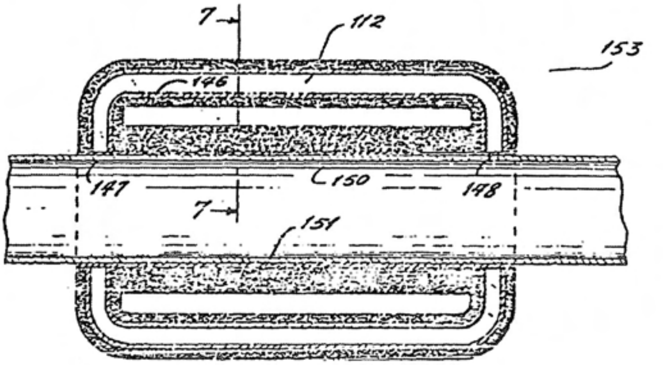
7
112
153
146
147
150
148
7
151

FIG.6

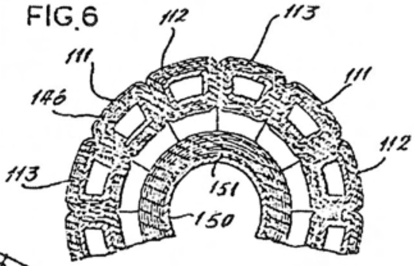
112
113
111
146
111
112
113
151
150

FIG.7

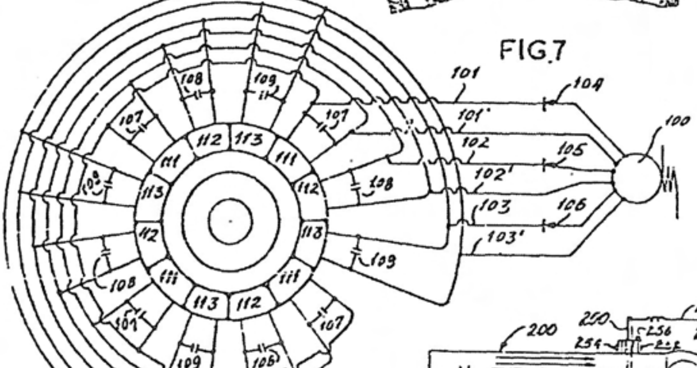
108
109
107
107
101
104
101'
102
105
102'
103
106
103'
100
112
113
111
111
113
112
108
109
112
113
111
111
108
107
107
113
112
109
108

FIG.8

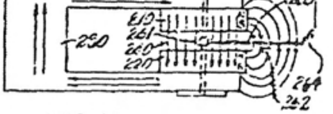
280
250
256
270
200
254
210
261
260
220
290
264
262

FIG 11

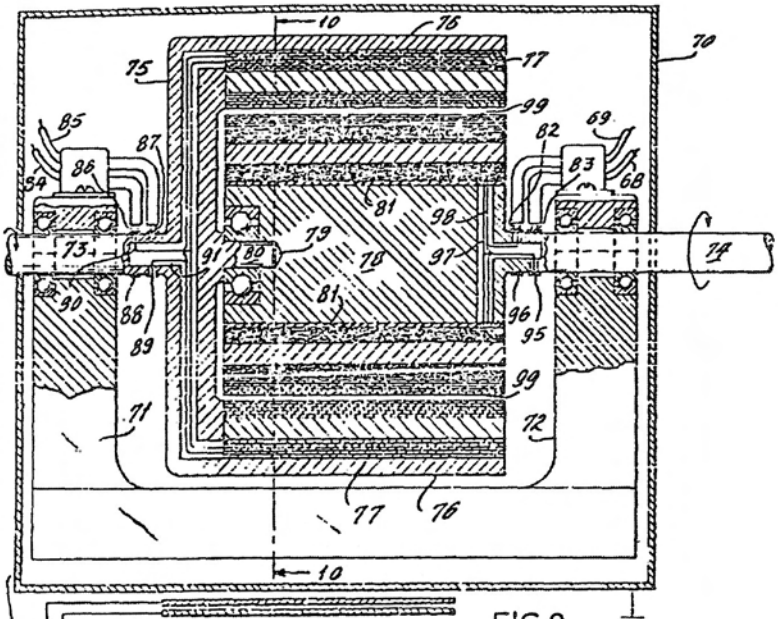

FIG. 9

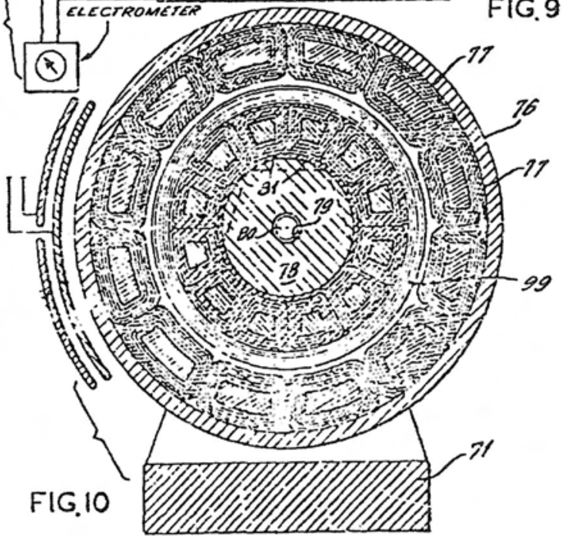
ELECTROMETER

FIG. 10

United States Patent

Hooper

[15] 3,656,013

[45] Apr. 11, 1972

---

[54] APPARATUS FOR GENERATING MOTIONAL ELECTRIC FIELD

[72] Inventor: William J. Hooper, Cuyahoga Falls, Ohio

[73] Assignee: Electrodynamic Gravity, Inc., County of Sarasota, Fla.

[22] Filed: Apr. 19, 1968

[21] Appl. No.: 722,587

[52] U.S. Cl. 310/10

[51] Int. Cl. H02k 1/00

[58] Field of Search ....310/10, 11, 178, 177, 156, 310/42, 268, 211, 169, 216, 269, 259; 103/1; 55/2, 14, 101, 123, 127, 120; 73/194 EM; 328/233–238

[56] References Cited

UNITED STATES PATENTS

3,285,179	11/1966	Resler, Jr.	103/1
3,336,489	8/1967	Volger	310/40
3,431,441	3/1969	Shair	310/11
399,800	3/1889	Thomson	310/115
460,087	9/1891	Hewett	310/46
2,066,343	1/1937	Gillen	310/46
2,558,540	6/1951	Clos	310/46
2,705,762	4/1955	Pile	310/104
3,277,631	10/1966	Sunnen	55/3
3,385,983	5/1968	Bohn et al.	310/11

OTHER PUBLICATIONS

Book–Theory and Calculation of Electrical Apparatus by C. P. Steinmetz; N.Y.– 1917.

Text Book– The Electromagnetic Field in its Engineering Aspects G. W. Carter– 2nd edition– 1967 pages 168 to 171 incl., 328, 329, 334, 335 and 336

The Feynman Lecture on Physics by Feynman, Leighton and Sands Addison– Wesley– Publishing Co., New York; pages 17–1 and 17–2

Primary Examiner—D. X. Sliney
Attorney—Oldham & Oldham

[57] ABSTRACT

Apparatus for producing and demonstrating properties of motional electric fields by means of rotating magnetic flux produced by a plurality of magnets extending parallel with the axis of rotation, said flux of these magnets being put into rotation about a common axis by mechanical or by electro magnetic means.

6 Claims, 11 Drawing Figures

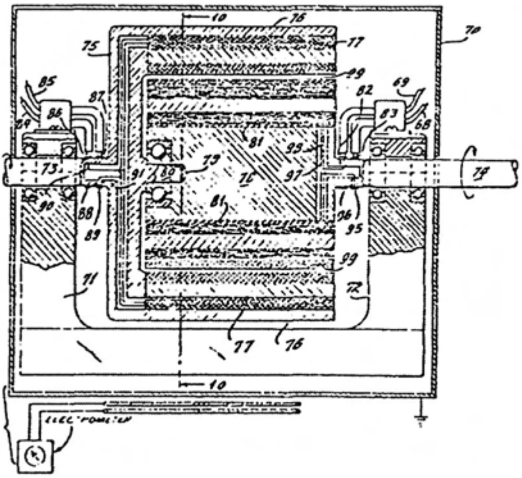

APPARATUS FOR GENERATING MOTIONAL ELECTRIC FIELD

This invention relates to the generation of motionally induced electric fields as distinct from electrostatic fields, and as distinct from the electric field (illustrated by transformer action) which is induced by the time rate of change of magnetic induction. Hence, all reference herein to a motional electric field means the electric field that is generated by the movement of magnetic flux.

The motional electric field generated by rotational motion of magnetic flux is non-uniform in both direction and radial intensity. Particles (large or small) of matter (solid, liquid or gaseous) which are within the ambit of this motional electric field are acted upon with a force which tends to accelerate them. It is well known that a particle carrying a charge Q (coulombs) in an electric field of intensity E (Newton/coulomb) will be acted upon by a force F (Newton) given by the equation $F = EQ$. The electric intensity E of a motional electric field is given by the vector equation $E = B \times V$, where B is the magnetic flux density at a point in space, expressed in webers per square meter, and V is the velocity of the moving magnetic flux at that point in space, expressed in meters per second. What is not so well known, however, is the fact that a non-uniform electric field exerts a force upon electrically neutral matter. In his celebrated text, "The Mathematical Theory of Electricity and Magnetism", Sir James H. Jeans describes, in the case of a non-uniform electrostatic field, how a "slab of dielectric will be sucked in between the plates of the condenser" thus demonstrating the mechanical force produced by such a field. He states (P. 125), "This, as will be seen later, is a particular case of a general theorem that any piece of dielectric is acted on by forces which tend to drag it from the weaker to the stronger parts of an electric field of force." According to Clerk Maxwell in "Electricity and Magnetism" (Vol. II, p. 181), Faraday discovered that the electric field due to electromagnetic induction penetrated within and throughout all materials. The motional electric field is just such a field. Apparatus for demonstrating the "general theorem", stated by Jeans, has, in the case of the non-uniform motional electric field, not heretofore been provided.

Accordingly, it is one of the objects of the present invention to provide a device for generating a non-uniform motional electric field whose force is of magnitude such that it can be sensed.

Since the motional electric field acts within and throughout all matter, such a non-uniform field will exert a physical force on any kind of electrically neutral matter, in a manner quite analogous to that of the electrostatic field with respect to dielectric material.

Another object of the invention is to provide apparatus for obtaining useful work from a non-uniform motional electric field.

The several objectives of the invention may be accomplished by rotating an elongate magnetic field about an axis which is concentric with the field and parallel with the field's elongation. For example, an array of magnets may be mounted on a rotor so that all magnets are parallel with the axis of the rotor, and all have the same polar orientation relative to the rotor. A solenoidal electro magnet, or another array of magnets, with magnetic axis concentric with the rotational axis may also be employed to enchance the flux density of the array of rotating magnets by flux linking with them. When such an array of magnets is rotated about the axis of the rotor, the magnetic field of each is rotated with it. As another example, an elongate magnetic field can be rotated about its axis, without mechanical movement, by the provision of stationary solenoidal electro magnets arrayed like the staves of a barrel and connected to a source of multi-phase alternating current (half wave rectified to prevent reversal of sign) in a manner comparable to the stator winding of a three-phase motor to produce "revolving field." While the composite magnetic field is thus in rotary motion, every charged, or uncharged particle of matter (solid, liquid, or gaseous) within the non-uniform motional electric field thus generated in the neighborhood of the rotor, will be acted upon by a force due to the electric intensity E. Positively charged particles will be forced in the direction of the vector E, negatively charged particles will be forced in the opposite direction, and electrically neutral matter will be pulled from the weaker regions of field intensity to the stronger regions. This latter mechanical force is called a differential force due to the action of the field on the internal electrical polarization of the neutral matter.

The three vectors E, B, and V are always (with a possible exception) mutually at right angles to each other. If the thumb of the right hand is pointed in the direction of the magnetic flux density vector B, and then rotated like a right-handed screw so that it points in the direction of the velocity V of the moving magnetic flux, then the forefinger of the right hand will point in the direction of the electric vector E. The rotation of the composite magnetic flux about its axis produces a motional electric field which will, except at the axial ends of the magnetic field, be quite generally directed radially with respect to the axis of rotation. The possible exception is the case of a magnet having a homogeneous electrically conductive core of perfectly circular cross-section, in which case the force of the electric field may be somewhat off radial.

With a given polar orientation of magnets relative to the axis of rotation, rotation of the magnetic field in a clockwise direction will generate a motionally induced electric field of sign (positive or negative) opposite that which is generated when the rotation is counterclockwise. For any given sense of rotation, reversing the polar orientation of the magnets relative to the axis of rotation likewise reverses the sign (positive or negative) of the motionally induced electric field. Reversing both the direction of rotation and the polar orientation of the magnets (a situation exemplified by viewing the apparatus first from one axial end, and secondly from the opposite axial end) produces no change in the sign (positive or negative) of the motionally induced electric field.

In the accompanying drawings, three embodiments of the invention are exemplified. Also, there are two model schematics to illustrate in its simplest form the principle of the invention, in which a plurality of commonplace permanent magnets are mounted upon a rotor.

The first embodiment is one in which an array of electro magnets is mounted on a rotor having a hollow center, through which particles of matter can be moved and their response to the motional electric field perceptibly demonstrated.

Second, there is an embodiment in which electro magnets are arrayed as in the first embodiment, but are stationary and their composite magnetic field is rotated electromagnetically by phase displacement.

Third, there is an embodiment in which a pair of oppositely rotating rotors, nested one within the other, are each provided, on its circumference, with an array of electro magnets oriented as aforesaid on each rotor, but oppositely oriented on the respective rotors, and wherein the effect of the motionally induced electric fields can be perceptibly demonstrated by the behavior of particles of matter externally of the rotor.

In the accompanying drawings:

FIG. 1 is a diagrammatic view illustrating the effect of rotating an array of spaced permanent magnets about an axis which is parallel with each of the magnets when all magnets have corresponding polar orientation;

FIG. 2 is a diagrammatic view representing a side elevation of the arrangement shown in FIG. 1, and illustrating the external magnetic flux paths of the several permanent magnets;

FIG. 3 is a perspective view of a rotor for generating a radially directed motional electric field in accordance with one embodiment of the invention, wherein electro magnets are arrayed within a hollow rotor with axes parallel to that of the rotor;

FIG. 4 is a sectional view taken along a diametric plane of the rotor shown in FIG. 3, together with its cooperating parts;

FIG. 5 is a sectional view along line 5—5 of FIG. 4, and showing, in addition, driving means for the rotor.

FIG. 6 is a sectional view taken along a diametric plane of a stationary form of apparatus similar to that shown in FIGS. 3, 4, and 5, but wherein the magnetic flux field is rotated electromagnetically by phase displacement;

FIG. 7 is a sectional view taken along line 7—7 of FIG. 6;

FIG. 8 is a wiring diagram for the apparatus shown in FIGS. 6 and 7.

FIG. 9 is a longitudinal sectional view of an apparatus having a plurality of rotors for generating a $B \times V$ electric field;

FIG. 10 is a sectional view taken along line 10—10 of FIG. 9; and

FIG. 11 is a diagrammatic illustration of an apparatus for exhibiting moving field phenomena.

In FIGS. 1 and 2, the principle of the invention is diagrammatically illustrated with the utmost simplicity, wherein a catena of magnetic flux fields, all with the same polar orientation, is produced by arranging an array of rod-type permanent magnets parallel with each other, and parallel with the axis about which they are to be rotated. In FIG. 1, the axis of rotation is designated 1, and may be considered as the center of a shaft of a rotor, preferably composed of material having low, if any, magnetic permeability, and low, if any, electrical conductivity. The periphery of such a rotor is designated by the circle 2. In the form shown, a series of permanent magnets 3, 4, 5, 6, 7, and 8 is appropriately secured to the periphery of the rotor with the respective magnets in equi-spaced relationship, and all with their north poles addressed in the same direction which, as shown in FIG. 1, is in the direction of the reader or, as shown in FIG. 2, to the reader's right. The external flux fields from the respective permanent magnets 3, 4, 5, and 6 are illustrated in FIG. 2, where it will be observed that the external flux path of all magnets is from right to left, a relationship designated by the plus (+) marks in FIG. 1. With such an arrangement, the flux fields of all magnets may be concurrently rotated about axis 1. Considering the rotation to be in the direction of arrows 9 and the polar orientation of the several permanent magnets to be as illustrated, a $B \times V$ motional electric field will be generated both externally and internally of rotor periphery 2 upon rotation of the rotor. The thus generated motional electric field will exert a radially directed force upon every particle of matter, neutral or charged, such as particle P, which is within the ambit of the catena of magnetic flux fields. The magnitude of the so impressed electric field intensity will depend upon the magnetic flux concentration at the particle P and the relative velocity between the flux field and particle P. While, if particle P is in motion, its velocity must be considered in arriving at the relative velocity V (in the formula $E = B \times V$) the disclosure will be simplified if the particle P be considered as stationary, and in such event V will be 2 R times S, where R is the radial displacement of particle P from axis 1, and where S is the revolutions per unit of time made by the rotor. Hence, with the particle P stationary, the magnitude of the force impressed upon it is increased when the speed of rotation of the rotor is increased at a given flux density; and is increased by an increase in the flux density at a given speed of rotation. The direction in which the force of the motionally induced electric field acts upon particle P is always *perpendicular to the magnetic flux field* B and to its velocity V. The sign of the electric force field may be either positive (acting radially outward from the axis of rotation) or negative (acting radially inward toward the axis of rotation), which, for any given polar orientation of the permanent magnets, may be reversed by reversing the direction of rotation of the rotor. With the polar orientation shown in FIGS. 1 and 2, and the direction of rotation indicated by arrow 9, if particle P is a proton or a positively charged molecule, the force of the $B \times V$ electric field will be radially inward as indicated by the solid line arrow attached to particle P; but if particle P is an electron or a negatively charged molecule, the particle will tend to move radially outward as shown by the broken line arrow; and if the particle P is electrically neutral, it will tend to move in the direction of increasing electrical intensity E. Increase of E requires either that both B and V be increasing or that one be increasing at a sufficiently greater rate than the other is decreasing so that the product $(B \times V)$ increases. In the relationship of field and particle shown in FIG. 1, the product $(B \times V)$ increases radially inward.

For practical purposes, however, commonplace permanent magnets produce insufficient flux density in their external flux paths to impress upon a particle of charged matter such as P, an electric force of sufficient magnitude that the force of the electric field can be measured with ease, at least at speeds of rotation within practical limits. Consequently, the practical embodiments of the invention, later to be described, make use of electro magnets of a form capable of producing external flux fields of vastly greater flux density B, as well as higher velocities V.

The term "unipolar induction" has been used to indicate the induction of an electric field in the vicinity of an axially symmetrical magnetic system rotating about its axis of symmetry. There has been disagreement about whether as the magnetic system is rotated the lines of magnetic induction are carried with it or remain stationary while the magnetic system rotates through them. For example, when a bar magnet is rotated about its magnetic axis, it is believed by many that the magnetic field is stationary. I have demonstrated that under certain conditions, at least, the magnetic lines can be made to move, and I have devised certain methods and apparatus for utilizing this phenomena as in the production of localized and high intensity electric fields.

Turning now to FIG. 11 of the drawings, an apparatus for demonstrating certain magnetic phenomena is illustrated in diagrammatic form. A magnetic system, indicated generally by numeral 200, includes a pair of permanent magnets 210 and 220 which are axially magnetized and mounted in axial alignment between the ends of a U-shaped yoke 230 of highly permeable magnetic material. The adjacent surfaces of magnets 210 and 220 being of opposed polarity are spaced apart, providing an air gap 240 therebetween. The lines of magnetic flux are shown in the drawing and indicated generally by numeral 262. A brass rod 250 extends axially through magnets 210 and 220. The magnetic system 200 is rotatable about rod 250. A stiff conductor 260 is electrically connected by a collar 261 at one end in rotatable or fixed relation as selectively desired, with the brass rod 250 at a point in the gap 240 between magnets 210 and 220. The other end of conductor 260 extends substantially perpendicularly from rod 250 to a distance in space away from the magnets and yoke where the magnetic field falls to approximately zero intensity. In order to complete an electrical circuit through galvanometer 270 and conductor 280 to the brass rod 250 and conductor 260, a flexible and extensible electrical wire 264 connects the end of conductor 260 to the galvanometer 270. A gear 252 locked alternatively by set screw 254 to rod 250 or by set screw 256 to yoke 230 can serve to provide driving motion to either the rod 250 or YOKE 230 as selectively desired. Normally, for the purposes of the experiment, the driving motion will be reciprocal because of the connection of wires 280 and 264.

With this apparatus, if the magnetic system 200 is held fixed and a given angular displacement imparted to conductor 260 by rotating it horizontally about or with rod 250 through the air gap at a given angular speed, a voltage is induced which provides a deflection of galvanometer 270. Furthermore, when the conductor 260 is held fixed and the magnetic system 200 rotated through the same angular displacement about rod 250, at the same given angular speed, a similar voltage of exactly the same magnitude but opposite direction is induced in the galvanometer circuit. The same deflection is obtained whether or not the magnets 210 and 220 rotate with the yoke 230 or are held stationary while the yoke alone is rotated. This indicates clearly that the lines of magnetic induction in the quantity rotate about the axis of rotation 250 as if they were rigidly attached to the magnetic system 200.

A most important aspect of the operation of the apparatus of FIG. 11 is noted when the magnets 210 and 220 are held fixed in relation to the conductor 260 and rod 250 and the yoke 230 is rotated about the rod 250. This action causes induction of current in the galvanometer circuit exactly as occurs when the entire magnetic structure or the conductor itself are moved. However, if the magnetic yoke 230 and conductor 260 are held fixed and magnets 210 and 220 rotated, there is no deflection of the galvanometer. Thus, it appears that the yoke structure plays an important part in controlling movement of the magnetic flux. Furthermore, it should be noted that the magnetic flux does not link the galvanometer circuit in the usual sense that this term is used, but that a current is induced in the galvanometer circuit through the action of the conductor 260 in merely cutting across the field extending between adjacent faces of magnets 210 and 220.

It has also been found that the permanent magnets 210 and 220 may be replaced by an electromagnet and the results described above duplicated. Furthermore, it can be demonstrated that the magnetic flux within the air core of a solenoid may be rotated about its magnetic axis. A brass tube formed in the shape of a toroidal C has a winding applied to its entire length with holes drilled in the tubing to receive a brass rod, as 250 in the structure shown in FIG. 11. With the winding energized, the above-described procedures produce similar galvanometer deflections.

This application is concerned with certain methods and apparatus which make use of the phenomena described above.

Referring now to FIGS. 3, 4, and 5 for an illustration of the second embodiment, a tubular rotor 10 is provided on its inner periphery with 12 keystone cross-sectional electro magnets 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, and 22 of the solenoid type. As seen in the cross-section of FIG. 5, the several electro magnets are arranged as sectors of a cylindrical annulus. While in the form shown, the several electro magnets 11–22 are shown in circumferentially wedged relationship, whereby to minimize the likelihood of relative movement radially inward as at standstill, as well as in the interest of achieving substantial uniformity of external flux density throughout the inside circular area of the rotor, it will be understood that when and if desired, the several electro magnets can be circumferentially spaced one from the other, and, if desired, a spacer of relatively low, if any, magnetic permeability interposed between them. Any suitable means may be employed for securing the several electro magnets to the shell of the rotor.

In order to facilitate the assembly of the several keystone-shaped electro magnets on the interior of the rotor as shown, the shell thereof is made of at least two pieces 23 and 24. In the form shown in FIGS. 3–5, the two pieces 23 and 24 are cup-shaped, and are respectively provided with outwardly extending flanges 25 and 26. Once the several solenoids have been positioned within the longer shell piece 23, the shorter shell piece 24 may be applied as a cap so that flanges 25 and 26 abut, and may be appropriately secured together as by riveting, bolting, welding, or cementing.

The shell pieces 23 and 24, as well as their interconnecting means, may be formed of material having little or no magnetic permeability, or in cases where it is desired to shield the exterior from the escape of magnetic flux, they may be formed of material having a high magnetic permeability. Instead of dividing the rotor shell in twain axially, it may be divided in twain radially, in which event the two halves will have axially extending joints when assembled, and may be held in such assembled position by banding or other appropriate means capable of sustaining the centrifugal force to which the rotor will be subject in use.

At the end of shell piece 23, remote from flange 25, there is provided a pair of slip rings 27 and 28 which, in the event the shell piece is formed of electrically conductive material, may be separated from the shell piece by sub-rings 29 and 30, of appropriate insulating material. The slip ring 27 is connected through a conductor 31 to the free end on the inner course of the solenoid winding for each of the electro magnets 11–22. The slip ring 28 is connected trough a conductor 32 to the free end of the outer course of winding in each of the solenoids 11–22. All 12 solenoids are connected so that the direction of electric current flow is the same in all solenoids. The energizing current is supplied from an external source through conductors 33 and 34 and brushes 35 and 36, in a manner well understood in the art.

The rotor is mounted upon an appropriate frame 37, having opposite stationary stub axles in the form of ferrules 38 and 39 which project toward each other. On the projecting portion of each of ferrules 38 and 39, there is provided an inner race 40 and 41 for each of two ball bearing sets whose outer races 42 and 43 are appropriately mounted to the respective ends of the rotor 10, so that the rotor 10 is free-running, with respect to the frame 37, about an axis concentric with the rotor 10.

In the embodiment shown, the exterior cylindrical surface of rotor 10 serves as a pulley for a drive belt 44 which connects the rotor to a source of power 45, but any other suitable means of driving the rotor in rotation may be utilized.

In the form shown in FIGS. 3–5, the apparatus is intended to concentrate the magnetic flux of the several solenoids radially inward, and to facilitate this each of the solenoids is provided with a core 46 of material having high magnetic permeability. As shown in FIG. 4, the several cores 46 are of squat U-shape, and extend, for the most part, parallel to the axis of the rotor, but at each end the cores are curved so as to provide faces 47 and 48 addressed radially inward. To minimize the radially outward escape of magnetic flux from the several solenoids, they are preferably wound of ribbon-like conductor which is coated with any appropriate electrical insulation. The first course of winding begins at 49, and proceeds in helically wound fashion with the minimum of space between adjacent turns about core 46, to the opposite end thereof, whereupon, without interrupting the continuity of the conductor, the second course is wound helically over the first course, with the pitch of the helix reversed from that in the first course. In this way, any gaps between successive turns in the first course are overlapped by turns in the second course. While the drawings show only two courses of conductor about each core 46, it will be understood that, in practice, there may be many more courses.

With the several solenoids connected through the slip rings to a source of direct current as above-described, it will be understood that all the solenoids 11–22 have the same polar orientation with respect to the axis of rotation of the rotor. For example, all core ends 47 are North poles, and all core ends 48 are South poles.

A tube 50 of circular cross-section is mounted in fixed relationship within the respective ferrules 38 and 39. The tube 50 is preferably made of material having little or no magnetic permeability. The tube 50 is stationary, but the rotor 10 rotates about the axis of the tube 50.

A stationary solenoid 51 is fixedly mounted on the exterior of tube 50 in a position such as to fit, with clearance, into the bight of the squat U-shaped solenoids 11–22. Solenoid 51 may be wound directly on tube 50, but in the form shown, it is wound on a cylindrical core 52 of material having little or no magnetic permeability, and the core 52 is telescoped on tube 50. The stationary solenoid 51 is energized through appropriate leads (not shown) from conductors 33 and 34 or other source of direct current, and is so connected that its magnetic polarity is opposite that of solenoids 11–22, that is to say that when, as aforesaid, the core ends 47 are North poles, the adjacent end of solenoid 51 will be its South pole. Thus, magnetic flux of solenoids 11–22 has an external path, through tube 50, which is coincident with, and in the same direction as, the internal path of magnetic flux in solenoid 51.

Thus, the magnetic field generated by each of the solenoids 11 through 22, as well as that generated by solenoid 51, penetrates the interior of tube 50, and when the rotor is driven in rotation, these magnetic fields rotate collectively about the axis of tube 50. During such rotation, particles of matter within the tube 50 are within the motional electric field generated by the rotating magnetic flux. Such an apparatus has a demonstrable effect in the separation and concentration of fluids. For example, in the de-salinization of water, the $B \times V$ electric field generated by the rotating magnetic fields, acts upon the disassociated ions of the sodium chloride in solution. This is explained by the fact that if the rotor 10 be rotated in the direction shown by the arrows in FIG. 5, the $B \times V$ field tends to drive the positively charged sodium cations toward the center, and tends to drive the negatively charged chlorine anions away from the center of tube 50. Hence, by introducing a flow of saline water at end 54 of tube 50, and by providing a concentric separating tube 55 at end 56 of tube 50, increments of fluid which are concentrated with sodium and depleted of chlorine may be drawn off through the stationary tube 55, while the increments of fluid which are concentrated with chlorine and depleted of sodium may be drawn off through the space between the exterior walls of tube 55 and the interior walls of tube 50.

Another utility of the apparatus shown is that of transmuting hydrogen into helium, in which event tube 55 is omitted and electrodes are provided at opposite ends 54 and 56 of tube 50, so that an electric arc may be established between them. Heretofore, difficulty has been encountered in maintaining the plasma of the arc in a relatively straight path between the electrodes as it ends to wiggle and extinguish itself when it makes contact with one of the confining walls. However, by impressing a $B \times V$ electric field upon such an arc, its path may be confined within controllable limits. The $B \times V$ electric field has the further effect of driving hydrogen protons toward the center of the tube, and driving electrons away from the center of the tube. Add one neutron to a hydrogen atom nucleus, and there is produced the isotope known as Dueterium; add one more neutron to the Dueterium nucleus, and it becomes Tritium; add one more proton to the Tritium nucleus, and it becomes Helium.

A third embodiment of the apparatus is shown in FIGS. 6, 7, and 8. With the significant exception that the embodiment of FIGS. 6, 7, and 8 has no moving mechanical parts, its organization is, in general, quite similar to that shown in FIGS. 4 and 5, and hence the last two digits of the reference characters utilized in the previous embodiment. As shown in FIG. 6, a cylindrical tube 150 has a solenoid 151 wound on the exterior thereof, and is energized with direct current from a suitable source through leads 153. A plurality of squat U-shaped solenoids 111, 112, and 113, of keystone-shaped cross-section, of which there are four each or any multiple of four, are wound as described in connection with the previous embodiment, and securely mounted in any suitable way with their pole faces 147 and 148 contiguous with, and addressed toward the axis of, tube 150. In this embodiment, however, the several squat U-shaped solenoids 111, 112, and 113, have cores which are made of material having little or not magnetic permeability, such as tubes 146 of paperboard wound to a keystone shape with hollow interior. As in the previous embodiment, the magnetic polarity of solenoid 151 is opposite that of the solenoids 111, 112, and 113, which is to say that if ends 147 of the solenoids 111–113 are the North poles thereof, the adjacent end of solenoid 151 is its South pole, so that the external magnetic flux path from solenoids 111–113 is coincident with, and in the same direction as, the internal flux path of solenoid 151.

In the embodiment shown in FIGS. 6–8, the several squat U-shaped solenoids 111, 112, and 113 are energized with non-interconnected three-phase alternating current, each phase of which has been half-wave rectified to produce direct current pulsating at alternate half cycles. One phase of the alternating current is connected to all solenoids 111; another phase of the alternating current is connected to all solenoids 112; and the third phase of the alternating current is connect to all solenoids 113, as shown in the wiring diagram of FIG. 8, where a three-phase alternator 100 is provided with six leads arranged in three pairs, to wit: leads 101 and 101′ for the first phase, leads 102 and 102′ for the second phase, and leads 103 and 103′ for the third phase. The respective phases are not electrically interconnected. In each of leads 101, 102, and 103, there is a rectifier 104, 105, and 106, respectively. Leads 101 and 101′ serve all four of solenoids 111; leads 102 and 102′ serve all four of solenoids 112; and leads 103 and 103′ serve all four of solenoids 113. Across leads 101 and 101′, between the rectifier 104 and the respective solenoids 111, there is provided a capacitor 107. Such capacitor is to coordinate the capacity and inductance of input, and to neutralize the effect of induction in solenoids 111 by adjacent solenoids during the off half-cycle of energization. Likewise, a capacitor 108 is connected between leads 102 and 102′, and a capacitor 109 is connected between leads 103 and 103′. It will be apparent to those skilled in the art that since the four solenoids of like phase are arranged in quardrature with each other, and since solenoids 111 will reach their peak of magnetic flux 120° ahead of solenoids 112 reaching their peak flux, and the latter 120° ahead of solenoids 113 reaching their peak flux, a revolving magnetic field is created. Due to the rectification of the alternating current, and the consequent utilization of the half cycles which all flow in the same direction, there will be no reversal of the direction of magnetic flux generated by any of solenoids 111, 112, and 113, and hence, as in the case of the previous embodiment, the external path of all magnetic flux generated in the solenoids 111, 112, and 113 is coincident with, and in the same direction as, the internal flux generated by direct current energization of solenoid 151. When such an arrangement is supplied with alternating current at high frequency, such as a thousand cycles per second, the same results are achievable with this embodiment as with the previous embodiment but without movement of any mechanical parts.

For the purpose of illustrating the effect of the $B \times V$ electric field, and segregating that effect from the effect of stray magnetic or electrostatic forces, reference may be had to FIGS. 9 and 10. The device there shown is intended to be encased in an electrically grounded box 70 of material which has a high magnetic permeability, and which also has the property of electrical conductivity for shielding electrostatic fields.

Within the box 70, there is a compound rotor machine, wherein the respective rotors are driven in opposite directions. In a suitable frame having spaced pedestals 71 and 72, there is mounted a shaft 73 and an independent shaft 74. Both shafts 73 and 74 may be driven from a single source of power, and are preferably driven at the same rotational speed, but in opposite senses. A yoke 75 is fixedly mounted to shaft 73 for rotation with it. The yoke is cup-shaped with cylindrical skirt 76. About the inner periphery of the skirt 76, a plurality os solenoidal electro magnets 77 are fixedly mounted thereto. The several electro magnets extend parallel with the common axis of rotation of shafts 73 and 74, and are of the same polar orientation with respect to those shafts. The several electro magnets 77 have identical solenoids wound, in the manner previously described, about identical cores of material having high magnetic permeability. The solenoids of the several electro magnets 77 are energized from a suitable source of direct current through conductors 84 and 85, brushes 86 and 87, slip rings 88 and 89, and conductors 90 and 91. The respective solenoids 76 are preferably connected in parallel circuit relationship, but, if desired, may be in series so long as uniform polar orientation is maintained.

Shaft 74 is provided with a hub 78 having a coaxial bore 79, into which is rotatably fitted a pilot 80 projecting concentrically from shaft 73. On the outer periphery of hub 78, there is fixedly mounted a plurality of electro magnets 81, each of which is an identical solenoid, and each of which has an identical core of material having high magnetic permeability. As in the previous case, the several solenoids 81 are all energized so as to have the same polar orientation with respect to their axis of rotation. Each of the solenoids is energized from a suitable source of direct current through conductors 68 and 69, brushes 82 and 83, slip rings 95 and 96, and conductors 97 and 98. All of the solenoids 81 are preferably connected in series circuit relationship, but, if desired, may be in parallel.

If desired, and as shown, the several solenoids 81 may be surrounded by a simple solenoid 99 wound helically thereabout and electrically connected in parallel or series relationship with solenoids 81 and to conductors 97–98 so that the current flow through solenoid 99 is in the same direction as through solenoids 81. Thus the winding 99 serves the dual purpose of boosting the magnetic fields generated by solenoids 81 and of mechanically binding the latter to the surface of hub 78.

When, as contemplated, the shafts 73 and 74 are driven in rotation simultaneously, but in opposite directions, the solenoids 77 which rotate with shaft 73 have opposite magnetic polar orientation to these (81) which rotate with shaft 74. The several solenoids 77 are preferably designed to create external magnetic flux fields of flux density approximately equal and opposite to those created by electro magnets 81 and 99, but by the principle of superposition ("Elec. & Mag.", by Ralph W. Winch, 1963 Ed., p. 202), each flux field acts as if the other were not present. Despite the fact that the flux fields created by the *responsive* series of electro magnets 77 and 81 and 89 are oppositely directed, since their rotation is in opposite senses, their resultant effect, insofar as concerns the generation of a $B \times V$ electric field, is cumulative. Accordingly, the external $B \times V$ electric field is intensified at the exterior of the device, and its effect on the outside of shielding box 70 is demonstrable, even though the resultant magnetic flux density there is approximately zero. Such demonstration might be through a capacitor and electrometer as shown in FIG. 9, for example.

In order to achieve the maximum flux density and hence the maximum $B \times V$ electric field intensity, it is desirable to take measures which keep the apparatus cool. This may involve resort to such means as: winding the several solenoids with so-called "super-conductors" such as that known commercially as "Supercon"; or refrigerating the apparatus as by immersing it in a bath of liquid helium which is capable of maintaining a temperature of approximately 4° Absolute.

While several illustrative embodiments of the invention have been disclosed in detail, it is not to be understood that the invention is limited to those embodiments. On the contrary, the principles of the invention are susceptible of application in a vast variety of forms without departing from the spirit of the invention or the scope of the appended claims. The apparatus defined in the following claims is believed by the applicant to demonstrate his unique theory disclosed above.

What is claimed is:

1. Apparatus for generating an electric field, comprising:
a frame;
a cylindrical rotor;
a shaft journalled on the frame and supporting the rotor for rotation about is longitudinal axis;
a plurality of electro magnets, each of which is an identical solenoid having a core formed of a material having a high magnetic permeability, the electro magnets being fixedly mounted on the rotor with each electromagnet extending parallel to the axis of rotation thereof,
means to energize the solenoids of the electro magnets from a direct current source, the solenoids being energized so as to maintain a uniform polar orientation; and
an enclosure surrounding the frame and the rotor, the enclosure being formed of an electrically conductive material the enclosure being grounded.

2. Apparatus according to claim 1 wherein the rotor is a hollow cylindrical rotor, the electromagnets being affixed to the inner cylindrical surface of the rotor.

3. Apparatus according to claim 1 wherein the solenoids are connected in parallel circuit relationship.

4. Apparatus for demonstrating magnetic phenomena comprising:
a frame;
a cylindrical rotor;
a shaft journalled on the frame and supporting the rotor for rotation about its longitudinal axis;
a plurality of electro magnets, each of which is an identical solenoid having a core formed of a material having a high magnetic permeability, the electro magnets being fixedly mounted on the rotor with each electro magnet extending parallel to the axis of rotation thereof,
means to energize the solenoids of the electro magnets from a direct current source, the solenoids being energized so as to maintain a uniform polar orientation;
an enclosure surrounding the frame and the rotor, the enclosure being formed of an electrically conductive material, the enclosure being grounded; and
means surrounding at least a portion of the enclosure to detect and measure the electric field generated during rotation of the energized rotor.

5. Apparatus for demonstrating magnetic phenomena, comprising:
a U-shaped yoke of highly permeable magnetic material;
a pair of permanent magnets mounted in axial alignment between the ends of the yoke, one magnet being mounted to each leg of the yoke;
an electrically conductive non-magnetic rod extending axially through the magnets;
a stiff conductor mounted at one end to the rod for rotation relative to the yoke in a plane perpendicular to the rod and intermediate the magnets, the conductor extending a sufficient distance beyond the magnets so that its other end is located at a point where the magnetic field from the magnets falls to approximately zero intensity;
a galvanometer; and
circuit means connecting the ends of the conductor to the galvanometer to measure current flow through the conductor.

6. Apparatus according to claim 5 wherein the magnets are rotatably mounted to the yoke for rotation about the axis of the rod, and where the magnets are axially magnetized and adjacent faces are of opposite polarity.

* * * * *

APPENDIX III

A CONDITIONAL CRITERION FOR IDENTITY, LEADING TO A FOURTH LAW OF LOGIC

Summary

If logic is regarded as a set of perceptual operations, then logic has a chronotopology (time structure). Identity or nonidentity then results as a decision from an algorithm — a set of perceptual operations and comparisons — in which case the nature of a particular identity is conditional upon the nature of the set of perceptual operations comprising the algorithm.

Ordinary logic does not account for the temporal aspects of perception, merely accounting for the spatial aspects.

In other words, Aristotlean logic is a synthesis of primitive observation, fitted to the partial (spatial) reality emerging from spacetime after the imposition of the monocular (one-at-a-time) photon interaction with matter.

In quantum mechanics, time is a parameter, not an observable. Hence measurement/detection (of observables) deals with primitive observation and Aristotlean logic (topology).

Total reality includes nonprimitive observation — hence, non-Aristotlean logic (chronotopology) — as shown in Young's two-slit experiment.

By applying temporal accounting to each perceptual operation, Aristotle's three laws can be shown to be self-contradictory and incomplete as written. That is, they are topolological, not chronotopological.

A simple derivation of a fourth law is shown and an application rule given which itself may be regarded as a fifth law of logic.

A proof of the fourth law by demonstration is given.

The resulting four law logic is chronotopological. The applica-

tion rule states that either Aristotle's three laws apply explicitly and the fourth law is implicit, or the fourth law applies explicitly and Aristotle's three laws are implicit.

The four-law chronotopological logic is theoretically capable of resolving every present three-law paradox.

Aristotle's Laws and the Paradox of Change

Aristotle's three laws of logic, on which foundation rests all mathematical, physical, and rational thinking, can ordinarily be stated as shown in Table 1.

Table 1. Aristotle's three laws of logical thought

1. $A \equiv A$
2. $A \not\equiv \overline{A}$
3. $A \vee \overline{A}$

A variety of arguments can easily be produced to show that these laws are incomplete; i.e., they do not specify all reality, for parts of reality can be shown to contradict one or more of Aristotle's laws.

Indeed, all "observed" reality can be shown to violate all three-laws.

E.g., the most direct violation is posed by the problem of change, a problem originally propounded by Heraclitus about 500 B.C., and unsolved to this day. Heraclitus pointed out that, for a thing to change, it must turn into something else, and then asked how a thing could be something other than itself?

We may think of a thing — say, α —some feature A of which is said to change. If A changes, it turns into $\overline{A}$, thus violating logic laws one and two. Further, we are considering A as the "changed thing. $\overline{A}$," i.e., something which is somehow both A and $\overline{A}$, so logic law three is violated as well.

Thus, if Aristotle's laws are taken to be all the fundamental laws of logic, then logically there can be no change whatsoever, because change negates all three laws. I.e., either change does not exist or it is totally illogical.

Since all measurements, detections, thoughts, and perceptions are simply changes, then it follows that these operations logically cannot exist. Or, if we assume the "operations" to exist, their outputs cannot exist. If the operations do not exist, then again their outputs do not exist.

So if the products or outputs cannot exist, then by this reasoning no perceived, detected, measured, conceived thing exists. If we then insist that such things do indeed exist, then all is paradoxical and illogical. This is essentially the nature of the paradox posed by Heraclitus.

Heraclitus's change paradox has not been satisfactorily resolved to this day, and rigorously all the rational science of the Western world, being based on paradoxical change (detection, perception, observation) is itself totally illogical by its own logical standards.

Resolving the Paradox of Change

However, the conditions necessary to resolve the problem of change can be stated simply by inspection of the problem as follows: (1) Aristotle's three laws must specify or apply to only that which is not changing, since change violates or negates all three laws; (2) If change is to logically exist, there must exist at least a fourth law of logic, one which applies to change; (3) This fourth law must contain the negations of each of the first three laws, since change negates them; (4) To be consistent, in any particular logical case, either the three laws explicitly apply or the fourth law explicitly applies (i.e., either change explicitly exists in that particular case or it does not); (5) Since all four laws must apply at all times, then when the three laws apply explicitly, the fourth law must be implicit—and when the fourth law applies explicitly, the three laws must be implicit.

With the five stated conditions, a fourth axiom of logic can be written simply by writing down the negations of Aristotle's three laws, and synthesizing these negations into a single fourth law, as shown in Table 2.

Table 2. Negations to Aristotle's laws.

Aristotle's law	Negation	of negation*
$A \equiv A$	$A \not\equiv A$	$A \equiv \overline{A}$
$A \not\equiv \overline{A}$	$A \equiv \overline{A}$	$A \equiv \overline{A}$
$A \vee \overline{A}$	$A \wedge \overline{A}$	$A \equiv \overline{A}$ or $[A, \overline{A}]$

However, even though we can synthesize the negation into a single law — the old "identity of opposites" idea — we still have the problem of understanding such a law. Though at first glance the negations and the synthesized fourth law seem bewildering, we can readily comprehend them if we carefully consider the temporal nature of the process that occurs in logical thinking.

The Importance of Time

Specifically, a finite interval of time is required to perceive, think, detect, or observe an entity — regardless of whether we refer to "physical" or "mental" detection, because both physical and mental processes are temporal. Indeed, we flatly state without further discussion that ultimately the identifying or mapping of physical and mental operations onto each other is what time is a *priori.*

At any rate, we now carefully account for each individual time interval required to think, conceive, detect, perceive, or observe any entity — whether that entity is physical or mental — and we also account for the finite time interval required to perform a logical operation. So we rewrite Aristotle's three laws as shown in Table 3, with subscripted numbers indicating the separate time intervals in each law.

*These negations mean that A and $\overline{A}$ are totally undifferentiated.

Table 3. Temporally accounted laws and negations.

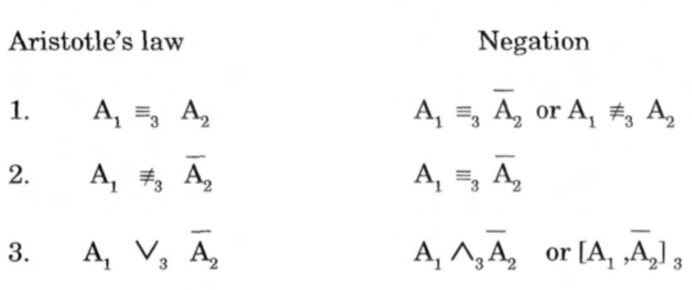

Aristotle's law		Negation
1.	$A_1 \equiv_3 A_2$	$A_1 \equiv_3 \bar{A}_2$ or $A_1 \not\equiv_3 A_2$
2.	$A_1 \not\equiv_3 \bar{A}_2$	$A_1 \equiv_3 \bar{A}_2$
3.	$A_1 \vee_3 \bar{A}_2$	$A_1 \wedge_3 \bar{A}_2$ or $[A_1, \bar{A}_2]_3$

The resolution to the entire mystery so long inherent in these axioms of logic now stands simply revealed: Whether one of Aristotle's laws holds or its negation holds is determined solely by the nature of the logical operation in time interval three.[1]

I.e., the operation in interval three may be regarded as an *algorithm* comprised of subsidiary (assumed) operations in separate time subintervals that, taken together, comprise the overall operation implied by the logic symbol.

Thus in the first law, if temporal tags (time snapshots) are not accounted (i,e., if they do not apply), then Aristotle's laws hold, for the snapshot 1 of A is not differentiated in algorithm 3 from snapshot 2 of A. This then rigorously holds for spatial (L^3) entities, but not for spacetime entities. The snapshots in this case for Aristotle's first law (and the others as well) are *spatial* snapshots. On the other hand, if snapshots 1 and 2 of A are themselves temporally differentiated in algorithm 3, then the negation of Aristotle's law applies, because the *spacetime* snapshots A_1 and A_2 are different. This is immediately apparent, e.g., in a Minkowski geometry representation, where the second snapshot of A will have a time coordinate different from the time coordinate of snapshot 1. This is represented as shown in figure 1, where "A" is taken as a simple magnitude, in this case 5.

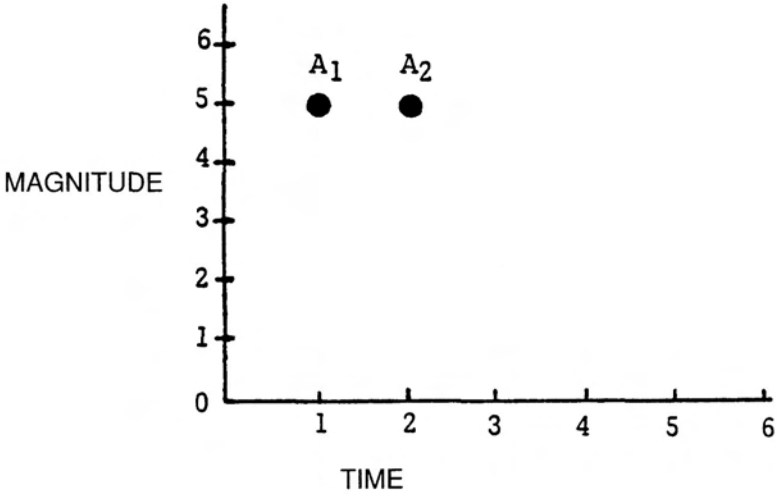

Figure 1. Spatial snapshots of spatiotemporal A differ.

As can be seen from figure 1, 5_1 is not identical to 5_2 *unless we imply the operator* $\partial/\partial t$ *in the time interval three algorithm.*

The negation of the second law may also be simply understood if we use temporal accounting. E.g., suppose we take $A_1 = +1$, $\overline{A}_2 = -1$ and then pose the absolute value operator || for potential use in algorithm 3. If || is not used, then

$$(+1)_1 \not\equiv_3 (-1)_2$$

and Aristotle's second law holds. If || is used, then

$$\left|\, (+1)_1 \,\right|_3 \equiv_3 \left|\, (-1)_2 \,\right|_3$$

and in that case

$$A_1 \equiv_3 \overline{A}_2$$

Note we are taking the view that there is nothing "absolute" or "inherent" about identity or non-identity; instead, each is a conditional result that can only be established by some logical, comparn tive set of operations. If the suboperations comprising the decision algorithm for the identity/non-identity determination are changed, the finding of the algorithm (the decision) may often change. Specifically, one can have the cases

$$\equiv_1 \not\equiv_3 \equiv_2$$

$$\not\equiv_1 \not\equiv_3 \not\equiv_2$$

Thus we advance a *conditional identity criterion* to be incorporated into formal logic: "Identity" or "non-identity" is defined by a decision made as a result of applying an operational algorithm; changing the internal operator components assumed inherent in the algorithm can change the decision. We are stating a fundamental principle that "identity" and "non-identity" are conditional and only conditional; they are never absolute.

Primitive Observation and "Reality"

With these points made, we now turn to the third law. From Table 2, on examination it can be seen that the third law actually is a statement for *monocular* perception, detection, observation, thought, or conception. Indeed, this law says that only a single thing at a time can be perceived, detected, observed, thought, or conceived.

As we pointed out in a previous paper[2] , there is a very good reason for this "law." Primitive man lived almost exclusively in a reality detected by light, by the photon interaction. Even in the absence of visible light, all bodies have temperature, and man is immersed in a "sea" of continual electromagnetic photon with interactions. The photon interaction is monocular - - only one at a time interacts with a particle of mass. Further, photon interaction constitutes the operator $\partial/\partial T$ invoked upon L^3T spacetime. Photon emission carries away time (the photon is made of $(\Delta E\ \Delta T)$, leaving behind an L^3 *spatial* reality, as we have previously pointed out.[3]

So all our primitive concepts, ideas, and notions about reality have come from over four million years of hominid and human experience in the photon-detected partial reality ("physical," "objective," or "spatial" reality) that remains when the time "dimension" (fundamental **variable)** is destroyed from L^3T spacetime, leaving only L^3 space behind. Specifically, our observed macroscopic reality consists of large temporal (mental) aggregates of such spatial results, where we cannot distinguish the tiny temporal separations of the pieces. Thus all our observed/perceived entities are spatial, and further, each perception/observation snapshot results in a frozen, unchanging spatial entity (resulting — in physical detection — from the so-called "collapse of the wave function." The loss of a wave function is simply the loss of time.) We vaguely sense "time" and "change" as the relation between these snapshots — i.e., by causality, or the ordering of the spatial changes — much as we see "movement" in movie frames rapidly projected onto a screen one-at-a-time.

Thus our primitive observations, from which have painfully been formed our relational concepts and ideas, are monocular, unchanging, and spatial. Aristotle's three laws of logic — which indeed may be taken to be only a simple synthesis of our primitive observation and corresponding relational concepts — then exhibit the same characteristics; they are monocular, unchanging, 3-dimensional, spatial, non-temporal relational statements. Any statement that is temporal, changing, or 4-dimensional will thus appear as a logical paradox to this logical shorthand.

But from Young's two-slit experiment, we already know that reality and relationships between its parts are quite different if the photon interaction is not invoked — i.e., if $\partial/\partial T$ is not invoked. Classical reality (as prescribed by Aristotle's laws) is directly violated by an electron in the two-slit experiment, e.g., if and only if photon interaction with the electron is not invoked. Again, this has been simply explained by the present author[4], and Charles Muses as early as 1957 pointed out the absence of any mystery in Young's experiment if the chronotopological aspects were considered.[5]

The Conscious Mind is Fitted to the Photon Interaction

However, what is normally referred to as the "conscious, thinking mind" is simply a functioning temporal (rigorously, chronotopological) mechanism that is painfully built up in the individual's awareness (his mind in the greater sense ofboth thought and awareness, whether monocular or multiocular) by training, conditioning and experience. Its functioning is largely conditioned by one's 90% or so attention to visual stimuli (to the partial reality remaining after photon interaction has been invoked, and to the memory-collated ordering of vast numbers of such photon interactions) and by one's cultural conditioning — which itself has been almost exclusively conditioned and shaped by the monocular photon interaction at base root.

Thus, since the beginning of man, his conscious, rational mind has been trained and constructed to function almost exclusively in basic correspondence with the photon interaction, and his experiential reality consists of the partial reality stripped from fundamental reality by photon interaction.

All "perceived differences," e.g., are created by this deep mindset. As has been previously pointed out,[6] the solitary human problem responsible for all man's inhumanity to his fellow man is directly dependent upon man's almost exclusive detection, observation, perception, and conception of "difference" between humans, these "differences" being due exclusively and totally to the fitting of men's conscious minds to the photon interaction's monocular separation of spatial reality from nonspatial reality, i.e., to

$$\partial/\partial T\,(L^3T) \Rightarrow L^3$$

Such well-nigh total devotion to, and enslavement by, photon interaction also is responsible for the scientist's well-nigh total devotion to, and enslavement by, the present imperfect and incomplete three laws of logic, as presented by Aristotle. The depth of that devotion and enslavement is evidenced by the fact that the resolution of such paradoxes as Heraclitus's problem of change have eluded the best minds of humanity for several thousands of years. Indeed, these paradoxes cannot be resolved by the conscious, rational mind in its present state, for it has been most firmly constructed and fitted to function in accordance with the photon interaction.[7] One cannot hope to resolve any logical paradox by using only those same logical methods that found the situation to be paradoxical in the first place!

Dimensionality

That we need not be constrained by such universal delusion is already shown by binocular vision. Specifically, in viewing a three-dimensional **object, each** eye **never** detects a "third dimension," but detects only **an L^2 2-dimensional** picture. By taking two slightly different 2-dimension snapshots and superposing them, the third dimension is gained. One then essentially sees the resultant superposed pictures as "almost the same but not quite." I.e., the Aristotlean identity algorithm, if satisfied, yields "no difference, hence one object," and if not satisfied, yields "difference, hence multiple (extended) object(s)." So if the two snapshots are almost Aristotlean-identical but not quite, we get an extended-two-dimensional (three-dimensional) object. Otherwise we see two separate, two-dimensional snapshots blurred together (the reader is urged to try this and see).

The point is, "dimensionality" and the identity algorithm are directly related, and geometrically one follows from the other.

Awareness of Time

In very similar manner, we can only gain cognizance of awareness of "time" (as a fourth dimension) by the superposing of two slightly differing (Aristotlean-wise) 3-dimensional snapshots. As is well known, e.g., time is not an "observable" in quantum mechanics; it is a "parameter." Rigorously, the only place such snapshots — each of which is "past" (spatially separated by the annihilation of time in the collapse of the wave function) — can multiply exist is in the "mind" in its most general sense. In a rigorous sense, mind and time can be taken as identical, and the "flow of time" can be taken as the "flow of mind connections or superpositions" of its spatial components. I.e., measured/detected/observed "physical phenomena" are a priori 3-dimensional and spatial, while a mind is four-dimensional and hyperspatial.[8] Spacetime exists mentally but not observably. Time is logically implicit, not explicit.

Temporal Aspects of Logic

Now we return to the temporal aspects of logic. Each perceptual part of each Aristotlean law is fitted to the photon interaction, hence monocular. The logic operation, inherent in the logic symbol in each statement, involves temporal superposition or comparison of spatial perceptual objects. Hence the logic operation is hyperdimensionally a function of mind and injects mind/time into the statement. Yet these laws, being fitted to or synthesizing photon interaction, attempt to prescribe the absence of time, even though writing down the logic operation rigorously invokes time. They are thus totally contradictory, since as written they implicitly violate themselves.

Table 4 summarizes the operations now to be permitted in the time-three algorithm, in developing a new four-law logic:

Table 4. Conditional identity rules.

1. $\left|\overline{S_1, S_2}\right|_3 = (\text{zero})_3 \Rightarrow (\text{identity})_3$

2. $\left|\overline{S_1, S_2}\right|_3 = (\text{nonzero})_3 \Rightarrow (\text{non-identity})_3$

3. $\left|\overline{S}\right|_3 = \left|\,\left|\overline{S}\right|_1, \left|\overline{S}\right|_2\,\right|_3 \Rightarrow$ **monocular separation, differentiation internally**

4. $\overset{S}{\left|\overline{}\right|}_3 = \overset{S}{\left|\overline{}\right|}_3 \Rightarrow$ **nonseparation, no differentiation internally**

Rules one and two simply state that, when snapshots 1 and 2 are superimposed (subtracted) in time interval 3, the resultant snapshot 3 may be zero or nonzero. If zero, snapshots 1 and 2 are said to be identical, and if snapshot 1 is to be labeled A, then snapshot 2 is to be labeled A. If snapshot 3 is nonzero, snapshots 1 and 2 are said to be nonidentical; if snapshot 1 is labeled A, then snapshot 2 is labeled $\overline{A}$.

Rule 3 says that snapshot 3 is a "memory" snapshot, and it may be particulately examined to monocularly separate snapshots 1 and 2.

Rule 4 states that snapshot 3 is not a "memory snapshot" and may not be further separated.

Note that in logic we repeatedly apply these rules in combination, serially or compositely. Note further that Rule 1 must serially apply both rules 3 and 4, as must rule 2 also.

If we take $4|_3$ to mean "rule 4 applied conditional to rule 3 also being applied," and $4|_{\bar{3}}$ to mean "rule 4 applied conditional to rule 3 not also being applied," we may write Table 5:

Table 5. Conditional identity, non-identity, and oneness.

$4|_{\bar{3}} \wedge 1 =$ identity (S_1, S_2)

$4|_{\bar{3}} \wedge 2 =$ non-identity (S_1, S_2)

$4|_{\bar{3}} =$ oneness without separate-ones; oneness, extraordinary and unperceivable; thing-in-itself

$4|_3 =$ "ordinary" one, perceivable separation, "thing-as-separate-from-others"

As can be seen, this type of reasoning also sheds a great deal of light on the long-standing problem of the "thing-in-itself," but that is beyond the scope of this paper.

The Fourth Law of Logic

Now we write the fourth law of logic as follows:

$$4.\quad (A_1, \bar{A}_2)_3 \Rightarrow A_1 \equiv_3 \bar{A}_2$$

where all we have said is that, by rule 4, in snapshot 3 no memory process is allowed, and no separation/differentiation whatsoever of A_1 and $\bar{A}_2$ is permitted. Under these operational conditions for identity, what had previously been called A_1 in snapshot 1 and what had been called $\bar{A}_2$ in snapshot 2 are indistinguishable, hence identical.

Figure 2 shows this concretely, and may be taken as a proof of the fourth law by demonstration.

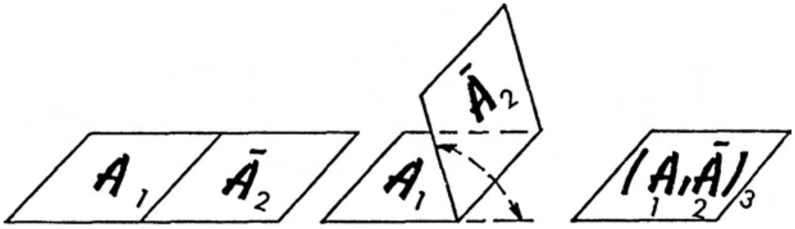

Figure 2. Proof that two opposites can be identified. What was separately perceived as A_1 in time snapshot 1 and what was separately perceived as $\bar{A}_2$ in time snapshot 2, cannot be distinguished in time snapshot 3.

Thus the age-old philosophical dilemma posed by the illogical identity of opposites has a simple resolution if one considers temporal aspects, and introduces temporal conditions for identity or non-identity decisions.

We now write the new four law conditional identity logic as shown in Table 6:

1. $A_1 \equiv_3 A_2$
2. $A \not\equiv_3 \bar{A}_2$
3. $A_1 \vee_3 \bar{A}_2$
4. $A_1 \equiv_3 \bar{A}_2$

Further, we point out that all four laws now apply. Laws 1,2, and 3 are the laws of explicit monocular perception, with implicit binocular perception. Law four is the law of explicit binocular perception, with implicit monocular perception. Both monocular and binocular perceptions must be and are used in each law. So in any situation, either the triad applies explicitly and the fourth law applies implicitly, or the fourth law applies explicitly and the triad applies implicitly.

Indeed, one can **even take the** view that we have prescribed a five-law logic, **the fifth la**w **being taken** as shown in Table 7:

Table 7. A possible fifth law of logic.

$$5.\ \left[1,2,3,\ \wedge(4)\right] \vee \left[4\ \wedge(1,2,3,)\right] \text{ where } () \Rightarrow \text{ implicit}$$

In a previous paper[9], the author has already presented methods to apply this new logic to resolve present paradoxes. At least hypothetically, every present paradox should be simply a statement of the explicit fourth law, and it should be resolvable by explicit application of that law[10].

NOTES AND REFERENCES

1. Specifically, by whether or not exclusivity applies. I.e., we may read Law 1 as "In snapshot (time interval) 3, what was A in snapshot 1 is exclusively identical to (unseparated from) what was A in snapshot 2." Law 2 may be read as "In snapshot 3, what was A in snapshot 1 is exclusively not identical to (is exclusively separated from) what was not-A in snapshot 2." Law 3 reads, "In snapshot 3, what was A in snapshot 1 and what was not-A in snapshot two are exclusively separated." Thus it can be seen that the three laws simply are statements involving whether or not two former perceptions are to be separated in a third perception. These three statements presently prescribe the total separation of the two previous perceptions and prohibit any admixture of the two — the so-called "excluded middle." Thus the three laws prescribe monocular, one-at-a-time perception.

   What we call a "wave" exists in time and is considered to contain an admixture of timeless, static spatial states (such as "wavelengths.") The prohibiting of admixtures thus represents the "collapse of the wave function" and corresponding loss of time. This defines "observation" and explains why all "measurement" and "detection" and "observation" — requiring a collapse of said wave function — are spatial and not spatiotemporal. That is why time is a parameter in quantum mechanics, not an observable.

   Relativity, being constrained by such logic, obviously can find nothing "physical" (observed, spatial, timeless thing frozen by the collapse of the wave function as engendered by or fitted to the photon interaction) that is traveling faster than light — i.e., that violates the conditions implied by the "observing/detecting agent."

That reality need not at all be so constrained is clearly shown by Young's two-slit experiment, the heart of all quantum mechanics. In this experiment, "classical' reality is violated if and only if the photon interaction is not invoked. Classical reality is obeyed if the photon interaction is invoked.
With appropriate change to logic to fit "reality that has not been interacted with by photons" and therefore is spatiotemporal, a new physics becomes possible.

2. Thomas E. Bearden, "Solution of the Fundamental Problem of Quantum Mechanics," January 3,1977, Defense Documentation Center.

3. Bearden, "Photon Quenching of the Paranormal (Time) Channel: A Brief Note," 20 April 1977, Defense Documentation Center.

4. Bearden, "Virtual State Engineering and Its Implications," 1979, Defense Documentation Center.

5. With the possible exception of Kozyrev — whose more technical works on time remain undisclosed to open science — no other person known to this author seems to have grasped the implications of a dynamic structure of time as penetratingly as has Charles Muses. The importance of the time interaction explaining the two-slit experiment (i.e., in explaining wave-particle duality) was noted as early as 1957 by Muses. See, e.g., Muses' introduction to Jerome Rothstein's **Communication, Organization, and Science,** The Falcon's Wing Press, Indian Hills, Colorado, 1958, p. lxii, where Muses pointed out that the celebrated wave-particle paradox remains a paradox only so long as the chronotopological (his word) phases of the phenomena are left unrealized in the analysis. The entire foreword by Muses is a remarkable document which analyzes the structure of time itself. With his hypernumbers Muses can describe the nested structure of time, which is what is actually

being carried by the photon. Further, he can theoretically predict mechanisms by means of which these structures can be orthorotated. It would appear that practical devices should be constructable on the principles elucidated by Muses, and it is little short of astonishing that fundamental work of such importance and application has been thus far little used by theorists, though it is already recognized in the standard literature; e.g., the profound summary paper "Hypernumbers II" in the January 1978 issue of the journal **Applied Mathematics and Computation,** published by Elsevier.

6. Bearden, "The One Human Problem Its Solution, and Its Relation to UFO Phenomena," Defense Documentation Center, January 3, 1977.

7. Which is why a Zen master often gives the student a koan to confound and overwhelm this automatic, robotic mindset and functioning that has been constructed as the student's "conscious mind."

8. Specifically, consciousness/life involves a seven dimensional body/being in an infinite-dimensional universe. See Thomas E. Bearden, "A Mind/Brain/Matter Model Consistent with Quantum Physics and UFO Phenomena," prepared for the 1979 MUFON Annual Symposium, available in the **Proceedings,** MUFON, 103 Oldtown Road, Sequin, Texas 78155.

9. Bearden, "The Fourth Law of Logic," **Specula,** Journal of the American Association of Meta-Science, P.O. Box 1182, Huntsville, Alabama 35807, Vol. 2, No.l, January-March 1979, pp. 30-40; also in publication in Defense Documentation Center. (The journal **Specula** is no longer pulished).

10. No consideration of the foundations of logic and mathematics is complete until one has read Morris Kline, **Mathematics: The Loss of Certainty,** Oxford University Press, New York, 1980.

APPENDIX IV

LIST OF SELECTED INCIDENTS

Author's note: In 1985-1987 a series of anomalous missile and space failures swept the U.S. and the West. These incidents were not unique, as many think, but were actually only a portion of thousands of related incidents showing the Soviet testing of advanced weapons throughout the world. This listing will show the types of incidents referred to.

The listing is not intended to be complete at all, but only representative. Hundreds of incidents of weather engineering, giant cloud radials and grids, suspicious light incidents, etc. could easily be added, were there time and room. No attempt has been made to exhaustively map the "aerial booms" that have blanketed the U.S., Europe, and other parts of the world. The light and "maybe it's UFO's" reports alone would add thousands of incidents to the list.

Our message is succinct: The Soviet Union has developed a new science and weaponized it, since at least the latter 1940's. It has utilized that science — energetics — to develop superweapons and supervehicles. Because of the danger from accidental initiation of nuclear weapons and materials, it is exceedingly difficult and dangerous for the Soviets to make extensive use of these weapons. Accordingly, a biological warfare strike — AIDS — has been utilized to deliver what is to be the knockout blow to the West.

DATE, INCIDENT, LOCATION, REMARKS

Early spring 1966 - Malstrom AFB Montana.
UFO sighting coincident with simultaneous problems in 10 strategic missiles in launch site associated with Malstrom AFB, Montana. *Similar event in Mar 1967 also. (Source:* ***Clear Intent.).***

20 **Mar** 67 - Malstrom AFB Montana.
Flight of 10 strategic missiles in launch site associated with Malstrom AFB, Montana experienced problems with guidance and control system. UFO in area and confirmed on radar. Jet fighters scrambled, results unknown. *Similar event one year earlier.* (Source: **Clear Intent.).**

1972 - Prague.
At a secret meeting of European communist leaders in Prague, Brezhnev lays out firm statement of Soviet intent to be able to dominate the world by 1985. Stated would control the oceans, 90% of the land, and the air and space above. *Notice he said control, not invade and conquer.*

Aug 73 - Over ocean.
USAF Minuteman ICBM is launched from Vandenberg AFB, aimed for Kwajalein missile range. Unidentified object appears and is tracked next to the ICBM's nose cone. Crossed the RVs trajectory. Object about 10 ft long. Seen by 2 separate radars. Three other identical objects were seen in the vicinity. *Probable Soviet advanced flying vehicles. Also a deliberate stimulus to see if U.S. knows of the technology, and how the U.S. reacts. (Source:* ***Clear Intent.).***

18 **Oct 73 -** Near Mansfield, Ohio.
U.S. Army helicopter is intercepted by glowing red object that converged on it with terrific speed. Object placed some sort of "tractor beam" on helicopter. Object was gray metallic structure, 50-60 ft long. Helicopter was stopped in mid air. Radio blacked out. Official report filed by pilot. *Probable advanced Soviet flying vehicle. Also a deliberate stimulus to see if U.S. knows of the technology, and how the U.S. reacts.*

21 Oct 73 - Near St. Joe, Indiana.
20 or 30 lights sighted circling in the sky near St. Joe, Indiana near a woods fire, by firemen. *Probable Soviet holographic tests to stimulate UFO reports.*

1973 to date - Bennett Island.
Bennett Is. exhaust plumes detected by U.S. weather satellites. *Well over 100. spotted so far. Weirdly, 5 islands and part of the continental Alaskan area are in process of being secretly given to the USSR; Bennett Island is one of them.*

May 1974 - Lake Michigan.
Blue-green glowing ball of light seen to fall into Lake Michigan. *Soviet scalar EM testing.*

8 July 74 - Lake Okeechobee, Florida and Atlantic Ocean.
Orange red ball of light lights up sky over south Florida. Crashes in area of Lake Okeechobee. Was accompanied by a large boom. Several other booms 5-7 min. after "object crashed." Pilots also saw object plummeting in from over Atlantic. *Probable Soviet scalar EM weapons testing and calibration.*

Nov. 74 - Saryshagan. USSR.
Construction of Tora facility begins at Soviets' Saryshagan test facility. Believed to be directed energy facility. *Probably connected with Soviet scalar EM weapons systems development.*

Oct? 74 - Antarctic.
Ozone hole over Antarctic first detected. Did not appear in measurements from 1957-1973. Note did not appear until exhaust plumes noted from and around Bennett Island, with substantial activity. *Looms over the continent every Oct. Diminishes through Mar. More severe each year since 1974.*

1975 - USSR.
Soviet article in **International Life** speaks of weather war, changing the nature of lightning, increasing the power of lightning, and directing electric charges of tremendous power against specified targets. "Atmospheric electricity" can be used to suppress mental activity of large groups of people. *Tip of the iceberg. Unclassified reference to Soviet's scalar EM (energetics) weapon program. Note involvement of "lightning" in destruction of U.S. missiles a decade later.*

16 **Feb 75 -** Caribbean.
R.M.S. Carmania (UK) observes bright white circular light appear and rise to 20 degrees, leaving comet-like trail, circle and disappear. Repeated four more times at exact hourly intervals. *Soviet scalar EM weapons tests. May have been a shipborne device.*

13 June 75 - Kremlin.
Brezhnev calls for ban on frightful weapons of mass destruction. *Major speech. Repeated proposals to U.S. Senators.*

2 **Jul 75** - Kremlin.
Brezhnev repeats his proposed ban on development of frightful new weapons, to a group of U.S. Senators. *To visiting U.S. Senators. No one knew what he was talking about.*

Aug. 1975 - Kremlin.
Ponomarev calls for ban on frightful new weapons of mass destruction. *To visiting Congressmen.*

23 Sep 1975 - United Nations.
Gromyko presents draft agreement to UN General assembly, for banning development of frightful new weapons. *30th session of UN Gen. Assy. No one knew what the Soviets meant.*

Oct - **Dec 75** - Satellite, over the Indian Ocean.
U.S. satellites illuminated or "blinded" over the Indian ocean. *Source 10-1,000 times as strong as natural sources. 5 incidents. One 4 hrs. long. Soviets wasted no time after Sep. fiasco.*

28 **Oct** to **11** Nov **75** - Mid-U.S. and Canada.
Series of "UFO's" seen at Malstrom AFB, Montana; Loring AFB, Maine; Minot AFB, North Dakota; Wurtsmith AFB, Michigan and Canadian Forces Station, Falconbridge, Ontario, Canada by reliable military personnel. Reported to NORAD. Some objects seemed like helicopters. Radar tracked objects. One object at 72,000 ft. was 100-ft sphere, with seeming "craters" around it.

Malstrom F-106's scrambled on one object but could not close on it in darkness and low altitude. Efforts by Air Guard helicopters, SAC helicopters and NORAD F-106's were fruitless. Probable Soviet activity in and around the SAC bases and Canadian radar site. (Source: **Clear Intent.).**

30 Oct 75 - Wurtsmith AFB, Michigan.
UFO's and "unidentified helicopters" sighted (multiple incidents) at or near Wurtsmith Air Force Base, Michigan. Helicopter hovered over the weapons storage area. Radar reported low-flying objects. Returning KC-135 tanker a/c diverted to intercept one UFO. Observed two objects, unable to close. Visual and radar contacts. Low on fuel, KC-135 returned to base to land, again seeing lights over the weapons storage area. (Source: **Clear Intent.)** *Soviet Activity.*

7 **Nov** 75 - Near Lewiston, Montana.
Giant glowing disc, football field sized, hovered over Minuteman missile launch site K-7, just south of Lewiston, Montana. When it rose to 1,000 ft., NORAD radar picked it up. Two F-106's were scrambled. UFO rose to 200,000 ft. and disappeared from NORAD radars. *Probable advanced Soviet flying vehicle, using antigravity propulsion system.* (Source: **Clear Intent.).**

Nov 75 - Semipalatinsk, USSR.
Large amounts of gaseous hydrogen with traces of tritium begin to be detected by USAF/TRW early warning satellites.
Nuclear debris apparently related to nuclear facility at Semipalatinsk labelled PNUTS (possible nuclear underground test site.) Probable large scalar EM development and test site.

Jan 76 - Cannon AFB, New Mexico.
Two UFO's reported near flight line at Cannon AFB, New Mexico. 25 yards in dia., gold or silver in color with blue light on top, hole in the middle and red light on bottom. Observed and reported by security police. (Source: **Clear Intent.).** *Probable Soviet activity.*

31 **Jan 76** - Eglin AFB, Florida.
MG Lane, CG, Armament and Development Test Center, Eglin AFB, Florida reported a UFO sighting near an Air Force Radar site from 0430 to 0600 Eastern Standard Time. Photographs were taken. Later press release called it "building lights." (Source: **Clear Intent.)** *Probable Soviet Scalar EM activity.*

1976 - U.S. Embassy, Moscow.
State Dept. declares U.S. Embassy in Moscow an unhealthful post. Installs Aluminum window screens against Soviet microwave radiation, which has been underway since 1950's.
Provided 20% pay extra differential. Embassy radiated with weak microwave radiation since 1950's. Two U.S. Ambassadors died, Stoessel ill with strange leukemia-like malady. Soviet stimulus to see if we knew about scalar electromagnetics, phase conjugate waves, and EM disease induction. (Stoessel died recently of cancer).

1976-U.S.
Soviet nuclear scientist named Rudakov visits U.S., proposes cooperative effort in some a**reas of** fusion research. Showed work he **had**

been doing and results obtained. U.S. officials reacted drastically, classified his talk as top secret, even ripped off blackboard and carried it away.*Very strange incident. Soviet open, public scientific work considered Top Secret by the U.S. One wonders why.*

30 **Mar** 76 - Netherlands.
Huge atmospheric boom occurs over northern part of Netherlands. *Through 1983, in 10 years 7 events took place.*

22 Jun 1976 - North Atlantic.
Expanding white sphere of light developed for about 10 min. then faded. Other associated light phenomena observed. *Seen and logged by passing ship.*

4 **Jul** 76 - Soviet Union.
Giant Soviet Woodpecker transmitters activated. *3-30 MHz band. Worldwide interference with communications.*

30 **Jul** 76 - Fort Ritchie, Maryland.
Several UFO's sighted in and around Ft. Ritchie, Maryland. One hovered over the ammunition storage area. Reported. *Temperature inversion at the time provided a convenient pseudoexplanation. Could not have caused the incidents.* (Source: **Clear Intent.)**

10 **Sep** 76 - Lithuania.
European Airways flight 831, Moscow to London, at 31,000 ft over Lithuania sees blinding single ball source of light at constant altitude about 5-6,000 ft. below them and above a lower cloud. *Pilot contacted Soviet authorities. Received negative response and suggestions he should not ask questions. Probable deliberate stimulus of flight so, from reports to British authorities, could ascertain whether Britain knew anything about scalar EM weaponry.*

19 Sep 76 - Near Teheran, Iran.
In early morning hours, UFOs were spotted north of Teheran, Iran. Two F-4's attempted to intercept. The first lost all communications and instrumentation at 25 miles distance. When the F-4 turned away, equipment was restored to functioning. Second F-4 acquired radar lock at 27 nm. and attempted to close. Object moved to stay at 25 nm distance. Second object came out of it, headed straight for the F-4. F-4 attempted to fire AIM-9 missile and at that instant weapons

control panel and communications went off. F-4 took evasive action. A third object was seen to emerge from first. The F-4 returned to base. Each time it passed through 150 degrees magnetic bearing, it lost communications and INS fluttered 30 to 50 degrees. A civilian airliner also experienced loss of communications in the vicinity. *Signatures of scalar electromagnetic weapons and antigravity.*

1976 - Soviet Union, Sweden.
Swedish national authorities detect anomalous radionuclides — similar to those from nuclear explosions — over Sweden. No accompanying seismic activity. Correlated with known Soviet activity at Semipalatinsk. *In other words, nuclear explosions were occurring, but they did not shake the earth. Total anomaly. Indicates transformation of the explosive energy. Positive signature.*

1976 - Philadelphia.
Legionnaires Disease precipitously strikes Legionnaires attending convention in downtown Philadelphia. Kills 34, makes 187 others seriously ill. Similar malady struck Odd Fellows convention at same hotel in 1974. *Odd person with apparent transmitter and antenna was expelled. Informed members: "It's too late! You are all dead!" Probable scalar EM conditioning of convention with an immune suppression pattern for the disease.*

21 **Jan 77** - Columbia.
Columbia A/L Avianca night flight. Instruments went haywire. Saw enormous, zigzagging glowing light, traveled at fantastic speed back and forth. Blinked landing lights, UFO responded. Sped away at incredible speed. *Probable Soviet advanced performance antigravity flying vehicle. Such have been widely sighted over South America.*

3 **Mar 77** - Columbia.
Columbia A/L Avianca night flight. Pilot abruptly changed course to circle a huge, glowing UFO. Object had cigar-shaped fuselage and was twice as big as the aircraft. Blinked lights, UFO responded. Sped away. Air traffic control radar tracked object at ten times the speed of normal aircraft.
Probable Soviet advanced performance antigravity flying vehicle. Such have been widely sighted in South America.

24 **Mar 77 -** Off coast of Spanish Sahara, Africa.
M.V. Kinpurnie Castle, off coast of Spanish Sahara, Africa, observes complex, dynamic light phenomena. Third phenomenon was the formation of a semi-circular area of moderate luminosity, a hemisphere of light. Took 3 min to form about a previous luminous patch. Then another luminous patch formed over the hemisphere. After about 7 min., phenomena dispersed. *Soviet test of Tesla shield and globes.*

11 Jun 77 - Moscow, USSR.
Soviet biophysicist hands Robert Toth, Los Angeles Times journalist, a report containing proof that certain particles existed in cells, carried information, and were radiated. Both seized by KGB and charged with possessing Soviet state secrets. Interrogated, released. *Neither knew what they actually had. Soviet Academician I.M. Mikhailov pronounced secret such phrases as "micro-organism self-radiation... by means of vacuum particles in space." Use of such particles was discussed, and was secret. Paper contained basis of Soviet energetics use in biology. Positive connection of Soviet classification to mitogenetic radiation and scalar EM waves ("vacuum particles" indicate virtual flux structuring).*

13 July 77 - New York.
Massive electrical blackout strikes New York City. Affects 8 million persons. Staggering series of "natural" electrical shocks, mechanical failures, and human errors occurred. Anomalous "lightning strokes" supposedly initiated the collapse of the power grid. *Probable Soviet test.*

1977 - Ocean off Yugoslavia.
M.V. Dolphin, between Yugoslavia and Israel, observed extensive light balls and repeated phenomena, appearing and disappearing. Men were disoriented. Compass was disrupted. Okay after the phenomena dispersed. *Probable Soviet test. Stimulus to see if Britain recognized the weaponry and knew anything about it.*

13-14 Dec 1977 - Leitchville, Australia.
Large aerial light plus 3 smaller ones observed for four hours by multiple persons. Lights simulated activity of a "mother ship" and smaller craft assigned. Changed colors. Then a shower of orange

lights occurred. *Probable Soviet test. Part of deception plan, to create "UFO" phenomena to hide weapons testing. TR wave holography.*

Dec 77 to **Jan 78 -** Off East U.S. Coast.
Mysterious atmospheric explosions occur off U.S. Atlantic Coast. *Some accompanied by light phenomena. Soviet tests and calibration (registering the long-range artillery, so to speak.)*

13-14 Dec 77 - Leitchville, Australia.
Large arrays of light phenomena sighted by multiple witnesses. *Probable Soviet creation. Reported as "UFO" phenomena. Holographic creation of "UFO" phenomena as part of deception plan to disguise tests of the weapons worldwide. TR wave holography.*

1977-78 - North Pacific Ocean.
Two huge hot spots in North Pacific ocean, 2,000 mi by 1,200 mi. *May have caused El Nino and one of most severe winters on record. May have been caused by Soviets as a weather experiment.*

27 **Jan 78 -** U.S. East Coast.
Boeing 727 off east U.S. coast experiences engines failing, 1 at a time, in all three engines. A/c plunges 8,000 feet; engines restart one at a time. *First time ever that all three engines on a 727 had failed. Probable Soviet test.*

Feb 78 - Rocky Mtns. NW of Las Vegas.
Four aircraft crash in a 6-day period in the rugged Rockies, NW of Las Vegas. *Connected with a crashed military a/c from Nellis AF Base. Should be further investigated.*

Feb 78 - Texas coast. Nova Scotia. Charleston, SC.
Atmospheric booms occur. *Multiple booms in Charleston. Soviet tests. Scalar EM weapons.*

2 **Apr 78 -** Bell Island, Nova Scotia.
Mysterious fireball, explosion accompanied by anomalous electrical effects. *Small bldgs blown apart, trees scorched, fireballs seen. Probable Soviet test. Apparently even registered on U.S. satellite.*

10 **May** 78 -Off Florida.
Women pilots in race experience anomalous "time loss" on their watches from Fort Myers to Freeport (Bahamas). *39pilots. Watches synchronized. Up to 3 min. lost. Definitely not normal EM effect. Probably due to sporadic natural scalar interferometry and time-reversed waves generated in area from time to time. "Bermuda Triangle" mechanism?*

May 78 - Florida.
Unexplained lighted objects sighted in sky over remote sector of Florida. Tracked on radar at a Navy EW center. Naval personnel also observed them through binoculars. *Probable Soviet advanced flying vehicles, or tests of intense holograms with energy sufficient to ionize air and reflect radar.*

18 **May** 78 - England.
Several atmospheric explosions shake houses over a wide area in Hull, Scunthorpe, Holderness coast, and Brough and Grimsby. *Probable Soviet weapon testing and calibration.*

Jun 78 - Indiana.
Large ball of light streaks across sky, stops, hovers, then explodes with a large boom. *Possible test. Standard Soviet incident to generate deception.*

7 **Sep** 78 - Netherlands.
Large acoustical event (boom) occurs over the Netherlands. *Soviet scalar EM weapons testing and calibration.*

Sep 78 - Tabas, Iran.
Anomalous Richter 7.4 earthquake strikes Tabas, Iran, killing 25,000. No aftershocks. Anomalous depth. Soviets needed to relievo pressure in rocks at nearby Askhabad on the Soviet/Iran border, else lose the city to a large earthquake. *36 hrs. earlier, Soviets exploded monstrous 10 Megaton underground nuclear explosion, to disguise what caused the earthquake. Almost definitely a Soviet- induced earthquake.*

Oct. 78 - USSR.
U.S. DOD weather satellite purportedly photographs a semi-rectangular object climbing out of atmosphere at 4,000 to 4,500 **mph.**

Object was over the Soviet Union. *Probable Soviet advanced flying vehicle, using antigravity propulsion system, if photograph is real.*

31 **Oct** 78 - Fairbanks, Alaska & Lake Ontario.
Red and yellow flash seen descending through sky S. of Fairbanks, Alaska by multiple observers. One minute later, U. of Alaska seismograph registered a tremor 1.5-2.0 on Richter scale. Ten days prior, a similar phenomenon was observed over Lake Ontario and a photo taken of the light. *Probable Soviet scalar EM wpns testing calibration.*

Nov 78 - Kuwait.
Large flying vehicle lands at Umm. Alaish oil pumping sta., Kuwait. Disrupts elect, comm, and pumping activities. Leaves after 7 min. *Probable antigravity propulsion system. Soviet vehicle. One of a series of "UFO" sightings in oil fields of Kuwait. Positive scalar electromagnetics and time-reversed wave signatures.*

Nov - Dec 1978 - Kuwait.
Series of "UFO" incidents over Kuwaiti oil fields occur. Kuwait Gov't. investigates, makes official report. *U.S. Pentagon noted incidents. Probable Soviet-induced incidents.*

1979 - British Columbia & Caspian Sea.
Ariel 6 satellite experiences anomalous turnoff of two power supplies. *Over British Columbia and Caspian Sea, when sun is shining. Probable Soviet stimulus.*

Jun 79 - Vienna.
Pres. Carter signs SALT II agreement in Vienna. After that, his acts and speech are erratic and he seems to be in a depressed state. *Suggestive of exposure to EM behavior modification. Soviets had possessed the LIDA technology for 29 years.*

Sep 1979 - USSR.
Strange giant semicircular light phenomena observed in USSR, toward Saryshagan. Observed several times. Source: London **Times.** *Observed from Afghanistan by British cameraman Nick Downie. Definitely Soviet tests of a Tesla globe and/or shield.*

22 Sep 79 - South Atlantic.
U.S. Vela satellites detect anomalous flash over South Atlantic, as if from a nuclear event. *U.S. Govt Agencies still disagree as to whether this was a nuclear event or not. May have been a Soviet test, or a test of a scalar EM weapon by another nonhostile country.*

25 Dec 79 - Afghanistan.
Soviets invade Afghanistan. *This is the Soviet "Spanish Civil War," where wpns, training, and the state of their ready reserves are tested. It is not primarily the regulars who are fighting there. Some scalar weapons — e.g. Smirch death ray — are tested also.*

1980 - Afghanistan.
Soviets kill 225 women and children in a village in the province of Maiden, near Kabul, Afghanistan. A month and a half later, bodies still undecayed. Bodies were also shot, to try to deceive Mujaheddin as to cause of death. *Bodies perfectly preserved. No smell, no bugs, no scavengers had eaten them. Probable Soviet test of a "death ray." Definitely scalar electromagnetic weapon signature.*

Feb 1980 - Duncanville, Texas.
About 100 startlings suddenly start falling from flight, instantly dead. *One of several such anomalous bird falls. Connected with very high frequency components added to Woodpecker signals for sharp localization. Birds entering such an area are killed.*

May 1980 - Seminole, Florida.
Start of mysterious plague of assorted illnesses striking healthy persons in a 20-block section of Seminole, Florida. Animals affected also. Lack friendly bacteria in their intestines. *Signatures of EM induction of disease and kill of intestinal bacteria. The lack of friendly bacteria is important. May indicate one direction of biological warfare planned by Soviets.*

June 1980 - USSR.
Huge, luminous hemisphere of light seen within Soviet Union, from Kuwait. Probably over Caucasus region of USSR. Source: personal communication from observer. *Tesla shield. A definite test, unequivocal.*

July 1980 - U.S. and Canada.
Richter 5.1 quake centered in northern Kentucky rumbles through 14 states and into Canada. *No known fault. Was predicted two wks in advanced by an associate.*

7 **Nov 80** - Applegate Valley, Oregon.
Mysterious explosion, earth rumbling, brilliant light turns night into day in Southern Oregon. *Associated with subsequent earthquake next day?*

8 **Nov 80** - California.
Richter 7.5 earthquake off coast of Southern California. Shakes CA and Southern Oregon. *Note anomalous phenomena in S. Oregon the preceding day.*

Dec 1980 - South Atlantic.
U.S. Vela satellites detect another anomalous nuclear-like flash over the South Atlantic. *In infrared only. Obviously not a normal nuclear wpn.*

28 **Dec 80** - South Atlantic.
Great bluish-white flash over the South Atlantic. Clear sky except 2 or 3 small cumulus clouds, one about 600 ft. above ship. *Possibly connected with the second Vela flash detection, or generated by the same system. One of many such incidents.*

1 **Jan 81** - Somerset, Pennsylvania.
Pilots of two a/c report the descent of a bright, shiny object with fiery tail over Pennsylvania. Large boom rumbles across most of Western PA. *Possible Soviet scalar EM testing. Part of deception plan.*

Jan 81 - Minnesoto, Wisconsin, South Dakota.
Great skyflashes and rumbling noises occur around Morris, Minnesoto and Wisconsin and South Dakota. *Possible Soviet scalar EM testing. Part of deception plan.*

Feb 81 - San Jose California.
Two pilots menaced by 10 ft. dia. light which flew directly at aircraft over San Jose, CA airport. Plane climbed, light followed, then sped away. *Totally consistent with exercise of the Soviet Launch Phase*

Anti-bomber system's real-time holography aspects of the Woodpecker OTH radars.

6 **May** 81 - Fredericksburg, Maryland.
EC-135 a/c crashed near Fredericksburg, MD. Severe nose-down from 29,000 ft. Apparent loss of all power. At 5,000 ft., either fire or minor explosion occurred. At 1,500 large explosion destroyed the aircraft. *Possible Soviet test of OTH radar anti-bomber weapon system?*

Mid-81 - Afghanistan.
Soviet helicopters use mysterious, highly lethal agent in Afghanistan. Causes death so quickly victims do not have time to make the slightest move. Rpt by Jane's Defense weekly. Referred to as Smirch gas or Smert gas. *Probable test of a Soviet scalar EM "death ray."*

20-23 Jan., 1982 - Gander, Newfoundland.
"Object with red, green and yellow lights sighted in the Gander, Newfoundland area more than 36 times. Alternately hovered and darted in a complex "flight" pattern. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception plan. Gander is a major "registration point."*

3 **Feb** 82 - Newfoundland.
Red and green flashing lights travelling across the horizon at 2,000 mph reported by 3 Armed Forces officers at Gander, Newfoundland. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception plan. Possible major "registration point."*

Feb 82 - Ankasaray, Turkey.
Cluster of high, multiple lights pass over Turkish town of Ankasaray, 200 km. from Ankara. Dogs were disturbed also. *Possible real time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception of plan. Gander is a major "registration point."*

Sep 82 - Ankara, Turkey.
Cluster of 10-14 high, multiple lights pass over Ankara, Turkey in clear night sky. Dogs were also disturbed. *Possible real-time holographic phenomena produced by Soviet scalar EM weapons as part of the worldwide deception plan.*

Early 82 - St. Petersburg, Florida.
Extended series of booms occur off coast of Florida, near St. Petersburg. *Continued lengthy period of time. Years of Soviet scalar EM interference grid adjustments have been registered in Florida and off St. Petersburg. Booms are so numerous that meetings of civil defense, federal, and state officials have been held.*

June 82 - Washington/Canada border and off-shore.
Extended series of booms occur over Pacific Ocean, centered about Port Angeles and Sequim near Washington state/Canadian border. *Persisted for several weeks. As many as 10 blasts per day. Fireballs and lights also seen. Soviet extended scalar EM tests.*

18 June 82 - Northern Pacific Ocean.
Two JAL flight crews observe large, yellowish-white globe of light over the North Pacific Ocean 700 mi. east of Hokkaido, Japan. Covered one fourth of the sky. Light gradually dimmed and disappeared. *Tesla globe. Definitely a Soviet weapons test.*

June 1982 - Red China.
Several Red Chinese giant light hemisphere and globe incidents. In one instance, a/c electrical power systems began to fail in several Red Chinese fighter a/c when encountering a huge hemisphere of glowing light. *Tesla globes. Tesla shields. Balls of light. Definitely Soviet weapons tests.*

24 **Sep** 82 - Maryland, New Jersey.
Anomalous airquakes off the coast of Cape May and Northern Maryland, and Atlantic City, New Jersey. *Adjustment of Woodpecker grid. Soviet tests. 3 weeks later also.*

14 **Oct** 82 - Maryland, New Jersey.
Anomalous airquakes off the coast of Cape May and Northern Maryland, and Atlantic City, New Jersey. *Adjustment of Woodpecker grid. Soviet tests. 3 weeks earlier also.*

16 **Nov** 82 - Delaware, New Jersey.
Five mysterious booms rock Sussex county, Delaware and southern New Jersey. *Adjustment of Woodpecker grid. Soviet tests.*

1983 - Loma Linda, California.
Photos and information about the Soviet LIDA machine released. 40 MHz carrier, complex waveforms. Induces catatonic state in humans and animals. *Shows consciousness can be affected by EM means. Was used on U.S. prisoners of war in Korea. Suggests similar capabilities are possible for Woodpecker. U.S. scientific community had said impossible. President Carter affected in Jun. 79?*

Early 1983 - Bennett Island.
Large exhaust plume detected coming from Bennett Island and vicinity, one on-shore and two offshore. Plume was 6 mi. wide, 155 mi. long and reached 23,000 ft in height. *U.S. weather satellites detected it. Nearly horiz. jet, 155 mi. long. Definitely not natural. At angle ofperhaps 1.5 degree above horizontal.*

27 **Sep** 83 - England.
British RAF/Panavia Tornado aircraft loses all electrical power and crashes. *Possible Test? Tornado system is triplexed.*

22 **Oct 83** - Off Okinawa.
Two a/c operating off U.S.S. Midway apparently lost w/o trace, about 127 mi. E. of Okinawa, under anomalous conditions. *Causes unknown. Additional data needed.*

8 **Nov 83** - Western Europe.
Anomalous earthquake occurs in Western Europe. All foreshocks and aftershocks were missing. Called the "Leige" earthquake. Definite symptoms of manmade (Soviet) origin. *Large boom occurred over Netherlands the next day over 35,000 sq. km. area. Soviet scalar EM weapon testing.*

9 **Nov 83** - Netherlands.
Large acoustic event occurs over Netherlands. Detected over 35,000 sq. km. First mistaken as a Dutch quake. *One day after the anomalous Leige quake. Soviet wpn testing.*

Dec 83 - Alma Ata.
Stored munitions exploded at Dolon (Soviet) airfield. 150 km south of Alma Ata. One of six that mysteriously exploded in 7 months. *May be evidence that another nation also has Scalar EM wpns, and*

is checking the Soviets. If so, probably was warning that Soviets also would suffer if they moved.

1984 - Novaya Zemlya.
First mysterious plume detected from Novaya Zemlya by U.S. weather satellites. Sometimes plumes are 200 mi. long. *Novaya Zemlya contains Soviet atomic facilities including weapon facilities.*

1984
Senior Soviet physicist G. Yu. Bogoslovsky publishes paper, "Generalization of Einstein's Relativity theory for the anisotropic space-time." Such Russian research published regularly. Would be impossible in U.S., where similar attempts meet discredit, humiliation, and rejection by leading U.S. physics journals and departments. *The West has decreed that general relativity must not be local, and above all must not be lab-bench experimental. Open Soviet publication indicative of experimental development of local space-time curvature technology.*

1982 - 84 - Florida (& elsewhere).
Airquakes, booms (hundreds). *16 counties, near coast.*

28 Jan 84 - Pennsylvania, New Jersey.
Bright orange light streaks across sky from Pa. to N.J.'s Cape May. Similar "fireball" explodes in Clementon. Two other sky explosions observed. *Possible Soviet activity. TR wave holography and pulses. Part of Soviet deception plan.*

Feb 84 - in space.
All three payloads of the Challenger flight fail. *2 commication satillites, 1 balloon.*

10 **Feb** 84 - Netherlands.
Large atmospheric boom occurs over Netherlands. *7 others thru 1983.*

29 Mar 84 - Bennett Island.
Bennett Island exhaust plume detected by U.S. weather satellite. *Over 100 such anomalous exhaust plumes detected there.*

9 **Apr.** 84 - S. of Kurils toward USSR.
Giant rapid cloud tower rose, then faded. Small semicir. halo of light replaced it. Halo expanded to full circle of light. Expanded to giant globe then faded. *Cloud 60,000 ft in 2 min. Light globe expanded to about 380 mi. dia. and center over 200 mi. high. Source:* ***Science*** *Journal. Definitely a Soviet test of multiple weapon modes.*

26 **Apr.** 84 - Bennett Island.
Bennett Island exhaust plume detected by U.S. weather satellite. *Over 100 such anomalous exhaust plumes detected there.*

May 84 - S.E. U.S.
Series of devastating storms breaks out over S.E. U.S. in early May. Tornadoes, large hail, downbursts. Some loss of life. Persisted through night into early morning hours. Several instances of rapid thunderstorm top warming were observed in satellite photography. Occurred near downbursts. Very heavy plume activity at Bennett Is. noted in Apr. and May, 1984. *Warming of tops of thunderstorms can produce downbursts. In literature. Note that Soviet scalar EM wpns can perform such top-warming of thunderstorm anvils, and induce downbursts on command.*

13 **May** 84 - Severomorsk, USSR.
Soviet missile storage facility at Severomorsk Naval Base on the Kola Peninsula 1450 km north of Moscow blows up in a series of vast explosions. Over 200 wounded. Soviet North Fleet lost about 1/3 of its missiles. Speculation of an "electromagnetic cause, possibly by reflection of radio waves from ionosphere." Detected by satellites; first thought to be nuclear explosion. *One of six such explosions of Soviet ammunition storage sites in a period of seven months. Definitely not accidental. Probably use of scalar EM weapons by another nation, not hostile to U.S., to warn the Soviets that they also will suffer if they move.*

15 **May** 84 - Bobrwjk Airfield.
Major blast occurs at Bobrwjk Airfield, 138 km SE of Minsk and the base of a Badger-equipped air-to-surface missile regiment of the Smolensk Air Army. 10 of the 11 ammo storage bldgs blow up and admin, and access control bldgs. are destroyed. Note that fires and explosions were still raging at Severomorsk, from 13 May event. *Same note as above.*

25 **Jun** 84 - Schuerin, East Germany.
Explosion occurs at an ammunition depot at Schuerin, south of Wismar in East Germany. Windows blown out within 10 mi. radius of depot. *Same note as above.*

27 **Jul** 84 - Kuril Islands.
Slowly expanding ball of light seen by Boeing 747 crew near the Kuril Islands. Expanded for 10 min. Shell was sharply defined. Light semi-transparent and nearly perfect half-circle. *Tesla globe. Bennett Is. exhaust plumes detected by U.S. weather satellite. Definitely a Soviet test. Related to 9 Apr. 84 incidents S. of Kurils.*

29 **Aug 84** - Edwards AFB, California.
B-l bomber prototype crashes in flight test about 10 mi. NE of Edwards AF Base.

1983-84 - Ireland.
For two years, freak lightning terrorizes villages of West Donegal, Ireland. Roofs stripped off, windows smashed, pipelines wrecked, telephones and TVs blown up. *No known cause for the anomalous lightning. Shows natural potential in area is disrupted.*

25 **Dec 84** - Ireland.
Freak lightning continues in West Donegal, Ireland. Causes 3-day power blackout. *Extensive plume activity at Bennett Island.*

1985 - Worldwide.
After years of a superb safety record, airline industry suffers a series of disastrous crashes. 20 major air crashes occur worldwide, with more than 2,150 lives lost.

Early 85 - Worldwide.
KGB begins a worldwide campaign to blame the U.S. for starting the AIDS epidemic. *Effort continues. Disinformation; is believed in 3rd world.*

Jan 85 - San Diego, California.
U.S. Navy chaff caught by wind, blown in over San Diego. Power knocked out to 60,000 homes. *Symptoms of presence of Soviet scalar EM grid in San Diego urea.*

7 **Jan 85 -** Cantrell, Alaska.
Cantrell, Alaska. Blinding white light passes in sky overhead, lights up sky and mountain range for 30-40 seconds for 16 mi. Est. 3-mi. altitude. *Probable Soviet test. Note that plume activity at Bennett Is. is heavy in Jan. 85.*

Feb 85 - Zimbabwe.
Zimbabwe suffering worst electrical storm season in history. Death toll up to 116. 21 persons killed by a single bolt in one case. *Anomalous lightning. Shows natural potential in area disrupted.*

17 & **18 Feb 1985 -** Anchorage, Eklutna Flats, Northway, Alaska.
Several anomalous lights and sky booms incidents in Alaska. Skyflashes, balls of light, booms, etc. Seen both from ground and by pilots of aircraft. *Probable Soviet realtime holography tests as part of worldwide deception plan.*

19 Feb 85 - Out from San Francisco, California.
China Airlines Flight 006 Boeing 747 enroute to Los Angeles. Engines fail. A/c falls 32,000 ft. in 2 min. Restarts engines, lands at San Francisco. *Note previous incidents over Alaska on 17 & 18 Feb. Probable Soviet test of launch phase anti-bomber system capability of woodpecker transmitters.*

10 **Mar 85 -** Tblisi & Tallin, USSR.
Soviet Academy of Sciences announces an Aeroflot a/c had been followed by a UFO for 800 mi. *Soviets occasionally produce light phenomena in their own territory and near Soviet a/c as part of their deception plan. They do not wish to appear different.*

1 **May 85 -** Near Nashville (monitoring station).
In and around May Day 1985, the Soviet Union held a giant exercise of all its strategic scalar EM weapons. 27 "power taps" into the earth — each two frequencies 12 KHz apart — were activated. The earth was thus in forced gravitational resonance beneath our feet on 54 frequencies. Each tap estimated to power up to six of the giant weapons; thus over 100 major strategic scalar EM weapon transmitters were activated. Frequency bands were filled with scalar EM transmissions, command and control, including to subs on station underwater. *Frank Golden monitored this startling giant exercise, which was probably held for Gorbachev, and this author observed it by courtesy of Golden's special detection equipment. Thus the Soviets*

met the 1985 timetable established by Brezhnev in 1972 at a secret meeting in Prague of European Communist leaders. Western nations do not appear to have the equipment to monitor scalar EM emissions, and so Western intelligence agencies are believed to have totally missed this giant exercise and its significance. At conclusion, two power taps (four frequencies) were left in the earth to power "ready firing batteries," so to speak.

11 **Jun** 85 - Lanzhou, China.
China B747 crew encounters a huge, expanding yellowish glow of light, with very intense spot at center. *Probable Soviet test.*

11 **Jun** 85 - Mongolia.
Chinese jumbo jet, Peking to Paris flight, encounters giant "UFO" light form over Mongolia. 10 km. wide. Lit sky for 50 km. Extremely bright spot at center. Paced aircraft for 2 minutes. *Probable Soviet test. Soviet operator tracked a/c exhaust with long range antibomber system. Added giant light form for deception, and to create UFO reports, which are largely ignored.*

23 **Jun** 85 - Iowa, Nebraska, Kansas, Illinois, Kentucky.
Thunderstorms over Iowa, Nebraska and Kansas organized into single circular pattern rainstorm 200,000 sq. km. in area. Declined until dawn, then rejuvenated to unleash severe weather over Illinois and Kentucky. *Such organization never before had been seen. Now called "mesoscale convective complex." Shows unique new weather pattern involving giant organization.*

2 **Aug** 85 - Dallas/Ft. Worth, Texas.
Delta Airlines flight 191, a Lockheed L-1011, encounters microburst while approaching runway at Dallas/Ft. Worth International Airport. Crashes and kills 134 of the 163 persons on board. *While such microbursts occur in nature, this was during a period of intense Soviet scalar EM activity. Note that warming the tops of thunderstorm anvil causes a downburst. May have been just accident. May have been test.*

23 **Aug** 85 - Italy, Adriatic Coast.
More than 28 "UFO" sightings reported along Italy's Adriatic coast soon after **midnight.** Light phenomena. *Probable Soviet test. Part of deception plan. Scalar KM holography (real-time).*

28 **AUG 85 -** Vandenberg AFB, California.
Titan B fails after launch at Vandenberg AFB, California. Believed to have carried sensitive U.S. satellite. *Suspicious.*

Sep 85 - Kourou.
Ariane rocket explodes at Kourou. 3rd stage engine failed to ignite properly, forcing loss of power. Destruct then ordered. *French officials suspect Soviet sabotage. Probable Soviet test.*

Sep/Oct 85 - Cape Canaveral, Florida.
Anomalies observed on two U.S. shuttle launches by NASA. *No action taken; NASA does not recognize scalar EM weapons capabilities nor testing over its Cape Canaveral launch site. Nor does it connect the mysterious "booms" that have rocked the Florida coast for several years with the Soviets or with the anomalies noted on its two shuttle launches, such as giant booms.*

2 **Sep 85 -** Mexico City, Mexico.
Worst hailstorm in 50 yrs. strikes Mexico City, leaving some streets more than a foot deep in ice. 1-hr pelting. 25 bldgs collapsed. 4,000 roofs heavily damaged. 1 killed, 185 injured. Power blacked out. *May have been due to extensive Soviet long range weather engineering. 17 days prior to suspicious giant Mexican earthquake.*

5 **Sep 85 -** Australia.
Premier Wran, Australia, along with Minister of Agriculture Hallam, observe a brigh spherical "UFO" from their government jet. It was photographed. Radar contact showed object at 100,000 ft. *Probably Soviet advanced performance antigravity flying vehicle.*

18 Sep 85 7:15 pan. - Catalina Is. Channel Is.
Cookman observes giant cloud radial, awesome to behold, between Catalina and Channel Islands, off California coast.
18 "funnel" strobes, dynamically boiling. Activity indicates great energy poured into earth and atmosphere. Mexican quake the next morning. Indicator of Soviet scalar EM connected to quake.

19 **Sep 85 7:18** a.m. - Mexico City, Mexico.
Large earthquake strikes Mexico City, causing great damage. Largest of several. *Preceded wks earlier by giant anomalous hailstorm. Evening before, giant radials showed extensive Soviet power*

transmission activity. Positive indications of Soviet test.

8-9 **Nov** 85 - Ventura County, California.
Highly anomalous power blackout in Carlisle Canyon, Ventura County, Calif. Accompanied by severe electrical anomalies totally beyond normal electromagnetics. *Investigated and reported by engineer Ron Cole, who lives in the area. TV, light, phone, electrical anomalies plus intermittent 29 Hz signals.*

Nov- Dec 85 - California.
Mysterious hum continues to baffle people at Pacific Heights, Calif, and Marina district on San Francisco Bay. Previously had had same hum. Note such hum in England persisting over long period. Note 8-9 Nov 29 Hz. anomalous elec. signal at Carlisle Canyon, Ventura County, CA. *Possible indications of Soviet scalar EM activity. Area appears to be one of Soviet register / test points.*

26 **Nov** 85 - Cape Canaveral, Florida.
Launch of space shuttle Atlantis from Cape Canaveral, Florida is accompanied with high anomalies. (1) Seconds before launch, Bob Gladwin inadvertently photographs nearby strike of what appears to be an electromagnetic missile, probably of Soviet origin. (2) Anomalous light is hanging in the sky, near the launch site: a marker beacon, placed there by the Soviet scalar EM weapon system being tested. Seen by hundreds of persons, and photographed by George Suchary. Printed in Fort Pierce **Tribune.** (3) 12 minutes after shuttle is launched, when it is well down range, a tremendous rumbling atmospheric boom occurs, heard 600 mi. up and down the East Coast. Marker beacon then slewed away by distant Soviet operator at high speed. *With offset EM missile strike, Soviets tested the weapon to be used on the Arrow DC-8 two weeks later at Gander AFB, Nfld. Marker beacon used in precise registration. Note it was evidenced again above the explosion of the Titan on 18 Apr. 86 at Vandenberg AFB. Boom is from energy induced over a large area; which can catch multiple missiles being launched, a large Naval task force at sea, or a whole flight of airborne bombers, fighters, or helicopters. Also can catch an entire organization on the ground, or an installation - - wiping out all electronics, exploding explosives and ordnance, stunning or killing personnel, etc.*

*Again hundreds of NASA engineers and scientists did not recognize what was **happening** at the **site** and over their own heads.*

12 Dec 1985 - Gander, Newfoundland.
Arrow DC-8 crashes immediately after takeoff at Gander AFB, Newfoundland. Lost power just after takeoff; engines "rumbling" abnormally rather than roaring, according to eyewitness whom the a/c flew directly over, 100 ft up. Aircraft seen by three eyewitnesses to have an orange glow on it. Symptoms of negative energy introduced into engines, cancelling some of heat energy of the engines and seriously interfering with combustion process, causing loss of power. Glow may have been steady, small electromagnetic ball of energy against the aircraft, causing it to electrically charge and exhibit corona. Anomalous hole in right fuselage, just ahead of engines, as if from internal explosion; however, no explosive residue (it was tested). One other anomalous hole found. Wild swings of some instruments recording just at/after liftoff — ignored by investigation. *Note previous 36 light incidents in and around Gander, Jan. 1982. Gander Airport is major stopover for Soviets, and a major "registration point" for scalar EM weapons. Note anomalies in shuttle launch 2 wks earlier, and apparent actual test of the small EM missile then, offset so as not to destroy that shuttle.*

Aircraft crew one of Arrow's best. Aircraft should have flown okay, if not interfered with. Serious indications that the DC-8, with over 250 U.S. soldiers and air crew, was deliberately destroyed by test of a Soviet launch-phase antibomber system.

Note metal-softening Soviet signal detected on 1 Jan 86 by Frank Golden may actually have been on the Woodpecker transmissions on Dec 12,1985. This metal-softening signal probably played a part in the destruction of the Challenger on 28 Jan 86.

1 **Jan** 86 - Near Nashville, Tennessee.
Frank Golden discovers a metal-softening signal on the Soviet Woodpecker interference grid. Performs experiments to verify the metal-softening ability of the signal. *Signal is of such a nature that only Golden's instrumentation could have detected it. He then scalarly cancelled the signal to totally prove it.*

24 **Jan** 86 - Ventura, California.
Engineer Ron Cole observes and sketches a giant grid cloud pattern over Ventura, Calif, at 11:45 p.m. PST. Giant cloud fingers, separated from each other, formed the crossed-grid pattern. *Shows Woodpecker scalar EM power very high and active.*

26 **Jan** 86 - Northern California.
Strong quake hits Northern Calif. 5.5 on Richter scale. Hollister area, about 120 mi. south of San Francisco. Several power outages in southern Bay area, where quake knocked out transformers, downed 21,000 V. power line, and set off burglar alarms. *May have been induced by Soviet scalar EM transmitters.*

26 **Jan** 86 - Los Angeles, California.
In greater Los Angeles area, Margaret Wilson observes 3 classic giant radial clouds form at 10 min. intervals: 1:15, 1:25, and 1:35 p.m. Appeared one after the other, till all three were there. Last one lasted about two hours, gradually distorting. *Shows a great amount of structuring to capture and turn the Jetstream, to direct cold weather to the Florida Panhandle and catch the Challenger before its impending launch, exposing it to drastic cold conditions. Definitely Soviet engineered.*

27 **Jan** 86 - Santa Monica, California.
On evening prior to the Challenger disaster, Al Matthews photographs a strange bar-grid cloud formation, absolutely rectangular bars, from the Los Angeles freeway. Pattern approx. over Santa Monica Bay, due north of Catalina Island. Edges and ends very straight, as if drawn in sky by a giant draftsman. *Shows severe high frequency content, and fine-point localization of Soviet scalar EM transmitters, in preparation for forthcoming launch of the ill-fated Challenger.*

27 **Jan** 86 - Florida Panhandle.
Avalanche of freezing air temperatures to record lows across the Southeast. Low temperature records set all across the South. Jetstream bent far south, across Florida Panhandle. *Definitely caused by Soviet weather engineering. Preceeded by giant radial cloud signatures in "hinge" areas of Los Angeles and Huntsville, Alabama, showing steering of Jetstream.*

28 **Jan** 86 - Cape Canaveral, Florida.
Low temperature records shattered all across the South. Into 30's in Florida and into teens elsewhere. *Definitely result of Soviet weather engineering, and definitely directed against the Challenger rocket on the launch pad.*

28 Jan 86 - Off Florida Coast.
About 4 hr. before the launch of the Challenger, all Soviet ships off-coast, which normally monitor shuttle launches, suddenly depart at high speed. *Soviets do not want any suggestion that the ships' radars and transmitters may have interfered with the Challenger launch. Total anomaly. Unprecented event.*

28 Jan 86 - Cape Canaveral, Florida.
National radio and TV newsmen comment on the highly unusual fact that no birds are flying in the launch area. *Reason: For severe localization, high frequency content of the scalar EM signals had been drastically increased. A bird's brain will detect and resonate to this energy, if the wavelength approaches the diameter of its brain. This resonance will cause great pain and/ or death of the bird. Thus the birds flee the area when this begins to occur — or they die. When birds accidentally fly into such a high frequency scalar EM signal, their nervous systems are instantly jammed. They die and fall from the sky.*

28 Jan 86 - Cape Canaveral, Florida.
Space shuttle Challenger, launched from Cape Canaveral after exposure to cold weather conditions, disastrously explodes. Booster seal problem. Encounters air turbulence during launch, at most critical altitude. *Soviet metal-softening signal has been added to grid. Extreme localization in area because of high frequency content added; shown by birds having vacated the area. Anomalous departure of Soviet ships offshore, 4 hrs. earlier. Three previous tests of the weaponry in that exact launch area, against the three previous shuttle launches. Kill of the Arrow DC-8 about 6-112 weeks earlier. Actual celebration by the KGB of perfect success of their active measures against the Challenger! U.S. space program crippled and set back severely.*

29 Jan 86 - Moscow.
At KGB headquarters in Moscow, the KGB holds a jubilant party to celebrate the perfect success of their active measures against the Challenger. *Note that Soviet scalar EM weapons — development, deployment, and use — are under the control of the KGB rather than the regular Soviet armed forces.*

31 **Jan 86 -** Mid-U.S.
Richter 5.0, Earthquake rattles 9 states and Great Lakes. *Possible test of weapon that killed the Challenger, in the earth transmission / earthquake induction mode.*

2-4 **Feb** 86 - Birmingham, Alabama.
In Birmingham, Alabama dead birds plummet from the air to the ground for three days. Blackbirds, cardinals, sparrows, bluebirds, etc. No trace of poison in carcasses. *Only a few days after kill of the Challenger. To divert the Jetstream further south and hit Florida, the Soviets had moved the normal "hinge pin"—where scalar EM energy, either positive or negative, is focussed —from Huntsville to Birmingham. Hinge pin area was still activated. The high frequencies used for extreme localization against the Challenger were lethal to birds. Therefore birds that inadvertently flew into this zone were instantly killed and dropped from the sky.*

10 **Feb.** 86 - Europe.
Super Frelon (French) helicopter crashes, with 11 killed. Aeronautique Navale aircraft. Part of a spate of crashes of French military aircraft in four months (through June 86). *Defense Minister Andre Giraud launched a special investigation.*

23 Feb 1986 - Irish Sea.
Over the Irish Sea in an airplane, England's Prince Charles sights a brilliantly glowing, red object hovering near them. His pilot reports the object to air traffic control. Four other aircraft traveling through that area radioed similar reports. *Probably either a Soviet craft or a Soviet holographic scalar EM demonstration as part of the overall deception plan, and/or a stimulus to a member of the Royal Family to observe what action is taken by British Ministry of Defense. By actions, could ascertain whether or not Britain knew what was going on.*

30 **Mar** 86 - Ventura, California.
Engineer Ron Cole in Ventura CA observes significant cloud patterns, positively correlated with his measurements of the Soviet woodpecker transmissions. *Cloud pattern changes were directly correlated to signal changes on the Woodpecker. Shows cloud signatures can be indications of specific Woodpecker activity.*

31 **Mar** 86 - San Francisco, California.
Strong earthquake, Richter 5.3, rocks San Francisco Bay area. Third in region in 3 days. Three aftershocks within minutes. *Note correlation of Woodpecker activity previous day.*

31 **Mar** 86 - San Francisco, California.
Crew error on Delta flight accidentally shuts down engines after normal takeoff from San Francisco. Engines restarted. *Definitely a crew error. Possible human effects from earthquake energy in area. Same type incident on Jun. 30, 1987.*

3 **Apr.** 86 - Florida.
Loud boom shakes houses and rattles windows over Brevard County, Florida. One of a whole series over the years. Not far from Kennedy Space Center. *Soviet adjustment of scalar EM grid.*

Apr 86 - France.
French Armee de L'Air has lost two Jaguars and two Alphajets in April 86. Part of a spate of losses in 4 months, leading Defense Minister Giraud to launch a special investigation.

7 **Apr** 86 - Sacramento, California.
Charlotte King of Sacramento calls UPI and predicts a major earthquake will strike in Canada, Alaska, Aleutians, or Japan. She had predicted eruption of Mt. St. Helens in 1980. On Apr. 27th, she predicted quake that shook Mexico three days later.
Note that human brain is a scalar EM interferometer, capable of transmitting and receiving scalar EM waves. Unfortunately, not conscious — output of system is in deep unconscious. Some persons are more sensitive than others, and can detect the scalar activity from the increasing stress in rocks that is leading toward an earthquake. Charlotte King is such a person, having over an 80% accuracy of prediction of quakes. She will be able to detect the increasing stress in rocks, whether naturally or artificially induced.

9 **Apr.** 86 - Aleutian Islands.
Three quakes rock the Aleutian Islands. Predicted by Charlotte King on Apr. 7. *Documented prediction.*

18 **Apr.** 86 - Huntsville, Alabama; Thousand Oaks, California.
T.E. Bearden sees and photographs remnants of a giant radial cloud

in Huntsville, Alabama. Extensive Soviet woodpecker grid activity noted over Thousand Oaks, California by Engineer Ron Cole. *Both show increased scalar EM activity in the interference grid.*

18 **Apr** 86 - Vandenberg AFB, California.
USAF Titan 34-D rocket explodes nine seconds after launch from Vandenberg AFB, California. Video captured the Soviet marker beacon **(Aviation Week & Space Technology).** Engineer Ron Cole and team study video of the explosion frame by frame and verify the anomalous beacon light is present, is separate from the explosion, and moves away separately after the explosion. *Sensitive satellite believed on board the rocket. U.S. Space surveillance program crippled. Probable Soviet kill. TR wave adjunct with Woodpecker.*

25 **Apr** 86 - New Mexico.
NASA research rocket misfires over the New Mexico desert. No public announcement made at the time. Was first failure in 25 consecutive missions of the Nike-Orion rocket. Said NASA spokesman Kukowski: "It looks like we're snakebit."
Possible Soviet test kill.

25 **Apr** 86 - California; Chernobyl, USSR.
Bill Bise detects the sudden loss of the East-West Soviet woodpecker transmitter he is monitoring in the field, leaving the North-South transmitter. The failed transmitter was about 20-30 km. from Chernobyl. The *Soviet transmitter sites safety circuits go-in full, trying desperately to hold back the giant scalar EM standing-wave potential that has been built up, and slowly and safely drain it away into the earth. In alarm, Soviets shut down nearby reactors at Chernobyl. Any "flash discharge" escape as a flash-over pulse of the scalar EM potential will produce a giant EGP pulse into the earth. When this EGP pulse reaches the four reactors, they will violently explode, full-up nuclear, like giant hydrogen bombs — perhaps 1200 megatons equivalent. Later the Soviets will claim that the operators shut the plant down to do some "unauthorized experiments."*

26 **Apr** 86 - Chernobyl, USSR.
*With the giant scalar EM standing wave almost all drained away, the safety circuits finally **fail,** dumping a small EGP into the earth. The nearest reactor at **Chernobyl** gets **it,** and the uranium fuel roils*

suddenly go half-decayed, instantly heating and blowing out the reactor, destroying it. Nuclear radiation fallout spreads over much of Europe.

3 **May** 86 - Florida.
Delta 178-rocket destroyed by range safety officer at T + 91 after two anomalous power surges occurred in the rocket, causing premature shutdown of 1st stage Rocketdyne RS-27 engine 71 sec. into the flight, and rocket to go aerodynamically unstable. Previous 43 launches had been successful. Since 1960, 177 launches 94% successful. One of most reliable rockets ever built. *Soviet weapons test. One of a series. TR wave adjunct with Woodpecker.*

7 **May** 86 - Aleutian Islands.
Great earthquake measuring 7.7 Richter strikes Aleutian Islands. Centered about 100 miles SE of the island of Adak. Tsunami alert issued for the North Pacific, but no major waves ever materialize. Thousands evacuate coastal areas. A *U.S. researcher known to this author used a powerful scalar interferometer to block the tsunami. The researcher must remained unnamed.*

9 **May** 86 - Redstone Arsenal, Alabama.
All power on Redstone Arsenal, Alabama suddenly and mysteriously fails all at once. *Similar related incident on May 12. No explanation.*

12 **May** 86 - Huntsville, Alabama.
In Huntsville, Alabama — adjacent to Redstone Arsenal — on local radio program Feedback, public radio discussed the mysterious failure of all power on Redstone Arsenal on Fri., May 9, 1986 and requested that, if anyone knew why all circuits would go off at once, to please get in touch. *Immediately the station's radio transmitter went off the air! (Almost as if "Big Brother's listening to you!")*

18 **May** 86 - Dijbouti.
French Breuguet Atlantic aircraft crashes in Dijbouti, killing 19. *Part of spate of crashes of French military aircraft in 4 months that leads French Defense Minister to open special investigation.*

19 **May** 86 - Brazil.
Brazilian fighter jets are sent to chase multi-colored "ping- pong ball sized" UFO's in skies over Brazil. 4 jets, two Mirages and 2 F-5's

chased them at speeds of 840 mph. Tracked on radars. Chase lasted 3 hrs. Planes could not catch them; ran low on fuel and returned. Discussed on TV by Brazil's Air Force Minister Brig. Gen. Otavio Moreira Lima. *Soviet scalar EM holographic production; distant Soviet operators having fun and games as part of the deception plan, generating UFO incidents and reports?*

22 **May** 86 - California.
High-speed orange "X" object with lights seen above Sonoma CA before dawn. At one time the object slowed and hovered.

29 **May** 86 - Huntsville, Alabama.
Huntsville and Madison County, Alabama electrical power dipped momentarily and mysteriously, causing a seconds-long outage in businesses and homes, playing havoc with traffic signals and computer centers throughout the area. Some traffic signals continued to malfunction for 30-45 min. after power was restored, then mysteriously resumed normal functioning. Some phones also went out inexplicably, and many battery-powered radios went off, and were awry for seconds or minutes before recovering. *Obviously this was not caused by any kind of ordinary electromagnetics. Shows anomalous and high activity (pulsed) in the Soviet scalar EM grid (Huntsville Ah. is a pivot point).*

30 **May 86** - French Guinea.
Ariane 2 rocket fails because of a faulty ignition system for the third stage engine. Rocket destroyed by ground command 4-1/2 min. after launch from Kourou, French Guinea. The third stage has failed 3 times since start of Ariane's operations in 1979.
Possible Soviet test.

Tad Szulc, Washington journalist, implies that French Defense Minister Andre Giraud may have discussed the possibility of Soviet sabotage of the rocket with Washington.

May? 1986 - Straits of Gibralter.
USS Atlanta nuclear sub runs aground in Straits of Gibralter and suffers "serious damage," according to reports in Madrid. A hole was made in one ballast tank and the sonar wrecked.
*Another **U.S. nuclear sub** ran aground within a month or so (earlier?) of this incident. C**ould something** have interfered with their navigation systems underwater? Continued Soviet Scalar EM tests?*

11-12 July - California.
Secret USAF crashes approx. 14 mi. NE of Bakersfield, CA along the Kern River. Site secured. Possible Stealth fighter. Crippled aircraft in trouble seen and photographed by Andy Hoyt, who turned photos over to USAF (properly so). Aircraft not on fire when came down. Did not explode in the air. *No information as to whether any Soviet scalar EM hanky-panky involved or not. Suspicious because follows a series of anomalous missile failures. Woodpeckers definitely have good anti-stealth capability.*

11 July 86 - Tennessee and Georgia.
Moderate quake with sonic boom shakes SE Tennessee and Northern Georgia. 3.7 on Richter scale. Centered just east of Chattanooga Tennessee on the TN/GA line.

12 July 86 - Michigan, Ohio, and Kentucky.
Mild quake in Michigan, Ohio, and Kentucky. 4.0 on Richter scale.

18 July 86 - Los Angeles, California.
Sharp aerial boom rattles Los Angeles Basin at about 0930 hrs. FAA and USAF unable to identify a potential source from an aircraft.

Aug 86 - Soviet Union.
Two huge new Soviet "radars" discovered on Western Soviet border by U.S. intelligence. One near Skrunda, on Lithuanian border. Other near Mukachevo, on the Czech border. *With phase conjugate adjuncts, represent significant strategic weapon system deployments.*

2 **Sep** 86 - Baltic Sea.
Finnish ferry Scandinavia in Baltic Sea observes a strong light rise from the sea near the Finnish Coast. After 3 to 4 seconds, a giant mushroom reached 500 feet into the sky and covered the moon. *Probable Soviet test.*

12 Sep 86 - Dallas, Texas.
Mysterious, powerful "explosion" rocks South Dallas. No source found or explained. No supersonic jets in area.

Selected ***Incidents***

15 **Sep** 86 - Australia.
Katherine, Australia. Object with flaming orange **tail** moved **from** sky toward ground at 45 degree angle. Katherine's electrical **power** suddenly cut off and phone lines in the town disrupted. ***Probable*** *Soviet scalar EM generated.*

Oct-Nov 86 - Los Angeles, California.
Giant radial clouds observed in Los Angeles area.

Oct 86 - Reykjavik, Iceland.
Gorbachev and Reagan meet at Reykjavik, Iceland. Gorbachev proposes "zero option." *Seeks to get nuclear missiles removed, since they constitute "dead man fuzing" because of nuclear warheads, and severely limit pulse use of Soviet scalar EM weapons. Adamantly insists on bottling up SDI — because if space laser is converted to scalar EM laser, its power (effective) goes up about* 10^{20} *times, and single shot could devastate Russia.*

10 **Oct** 86 - Vandenberg AFB, California.
U.S. forced to scrub another missile launch when a USAF Scout satellite malfunctions during final countdown at Vandenberg AFB, Calif. *The booster's gyro system failed a checkout sequence at about T-5 hr. Scalar EM can inertially interfere also.*

13 Oct 86 - ABC News.
On Peter Jennings' ABC News show, Gennadi Gerasimov, Soviet spokesman, indicated that SDI as presently designed was of no concern to Soviets. Real concern is that U.S. will make a "great technical breakthrough" and deploy new, modified devices, presenting an unacceptable threat to the Soviet Union. *Here he is thinking of changing the laser to a scalar EM laser.*

6 **Nov** 86 - Albuquerque and Phoenix.
Power failures at Albuquerque air traffic control tower and a radar station near Phoenix, Arizona occur. Radar controlling high altitude air traffic from Texas Panhandle to southern California was knocked out for 40 min. First such incident in Albuquerque center. The Arizona radar was out for 59 hrs. *Note similar incident at Montreal on Nov. 11, 1986.*

10 **Nov** 86 - Huntsville, Alabama.
Three giant radial cloud formations in Huntsville, Alabama photographed by T.E. Bearden. One is a twin giant radial. *Shows heavy activity in scalar EM interference grid.*

10 **Nov** 86 - SE U.S.
Red glowing light with tail flares across sky in Southeast about sunset. Seen over parts of TN, AL, GA, MS. Ground searches conducted in several areas for possible crash. *Probable Soviet scalar EM activity. Note heightened activity in grid same day.*

10 **Nov** 86 - Near Nashville, Alabama: Hemet, California.
Frank Golden measures scalar EG field of earth — highly agitated and dynamic, about half as much as when Soviets induce earthquakes. Anomalous winter storm in progress. Cold records broken in Mid-U.S. Heat records broken in Florida. Storm penetrates very deeply southward. *Soviet weather engineering.*

11 **Nov** 86-Canada.
Air traffic control radar in Montreal, Canada region broken down. Backup system employed. *Note similar occurrences to Albuquerque and Phoenix on Nov. 6, 1986.*

12 **Nov** 86 - Atlanta, Georgia.
Stationary arc (dome) of clouds sighted over Atlanta. Well defined lower edge. Upper edge irregular. Multiple witnesses. Clouds moved into top of dome and flowed around it. Very strange sight. *Probable Soviet "gentle test" of placing a suppressive covering over a distant U.S. city. Performed at night, at minimal level (no glow) so no great notice would be taken. Dome was stable against 20 mph winds and other moving clouds that collided with it.*

13 **Nov** 86 - Huntsville, Alabama.
Twin giant radial cloud observed moving over Huntsville, Alabama by T.E. Bearden. West to East. At nearly noon. At dusk, single giant radial cloud was also spotted over Huntsville.
Significant scalar EM activity in grid.

17 **Nov** 86 - off Alaska.
JAL Flight 1628 observes giant UFO while enroute from Reykjavik, Iceland on mid-leg of Europe to Tokyo. Crossed Beaufort Sea off

Alaska's north coast. Sighted flashing lights which moved alongside the aircraft. Pilot descended, so did light. Glimpsed craft in profile, twice as big as aircraft carrier. On board radar located it. USAF and civilian controllers reported brief radar echoes, but tapes did not show it. Approaching Fairbanks, Alaska. Pilot made 360 degree circle, object stayed with him. Finally disappeared as he was approaching Anchorage. *Probable Soviet test as part of deception plan. TR wave holography.*

27 **Nov** 86 - Fort Pierce, Florida.
Flash of light in sky and sonic boom occurred vicinity of Fort Pierce, Florida. *Probable Soviet EM test. Note that a classified payload was scheduled for launch in near future at Cape Canaveral nearby.*

4 **Dec** 86 - Huntsville, Alabama; Hemet, California.
Twin giant radials seen and photographed over Huntsville AL. Large radial cloud seen in skies over Hemet, Calif, and photo printed in **The Press Enterprise,** Hemet, CA 5 Dec 86.

4 **Dec** 86 - Cape Canaveral, Florida
Atlas/Centaur rocket successfully launched at Cape Canaveral, orbits a communications satellite for DOD. *The Soviets had to allow one or two of the missiles to succeed, else even the lethargic U.S. bureaucracy would have caught on. Grid activity indicates they probably just tracked it without the other goodies.*

5 **Dec** 86 - Vandenberg AFB, California.
MX ICBM launched from Vandenberg AFB. Two of RVs did not deploy although all impacted within the Kwajalein Missile Test Range. *Note: Soviets let one go from Vandenberg.*

21 **Dec** 86 - Huntsville, Alabama.
Twin giant radial photographed over Huntsville Alabama.

23 **Dec** 86 - Huntsville, Alabama.
Small flash in lower sky over Huntsville, Alabama observed by T.E. Bearden. Followed by faint boom. Momentary loss of power occurred in large area of Huntsville. Positive correlation between the aerial flash/boom and the loss of the power. *Soviet test.*

7 - 21 Jan 1987 - Florida.
Several giant radial clouds seen in vicinity of Fort Lauderdale, Florida. *Indicates substantial action in scalar EM grid.*

20 **Jan** 87 - U.S.
Highly unusual bend in Jetstream. Completely down West Coast, curving east under Texas, NE along SE USA. *Shows great action in scalar EM grid. Significant weather engineering.*

20 **Jan 87** - Vandenberg, AFB, California.
Minuteman missile destroyed in launch. *Note substantial action in scalar EM grid.*

23 **Jan** 87 - Huntsville, Alabama.
Bearden photographs twin giant radial over Huntsville AL. Cloud fades with great and unusual rapidity. A private-type jet, similar to a Learjet, continually circled under the strange twin radial cloud, as if observing and/or photographing it. *Perhaps someone in the U.S. system wised up.*

28 **Jan** 87 - Vandenberg AFB, California.
Minuteman missile successfully launched from Vandenberg AFB. Note: *Anniversary of the Challenger disaster.*

7 **Feb** 87 - Alabama - Tennessee border.
Loud aerial boom or explosion shakes area from Tennessee, New Market, Monte Sano through Huntsville Alabama area.

10 **Feb** 87 - Huntsville, Alabama.
Bearden photographs twin radial cloud in Huntsville, AL. Cloud stayed fixed, although winds present and other clouds moved. Note big storm in NE - Cape Cod, etc. *Associated with weather engineering.*

11 **Feb** 87 - Huntsville, Alabama.
Bearden photographs another twin giant radial in Huntsville, Alabama. *Shows activity in Woodpecker grid.*

Feb 87 - Vandenberg AFB, California.
USAF Titan launched from Vandenberg AFB, Calif.

12 Feb 87 - California.
Blue Angel F-18 a/c crashed near Brawley, CA. Fourth military air crash in the Imperial Valley since October. *Note Woodpecker grid activity on previous day.*

26 Feb 87 - Cape Canaveral, Florida.
GOES-H weather satellite launched from Cape Canaveral on a DELTA rocket.

26 Mar 87 - Cape Canaveral, Florida.
Stunning failure of an Atlas/Centaur rocket launched from Cape Canaveral. May have been hit by lightning. Failed suddenly 51 sec. after liftoff. Veered out of control and destroyed. *Note: Possible Soviet scalar EM involvement. Much to do later about multiple lightning strikes; very suspicious.*

June 1987 - Vandenberg AFB, California.
Minuteman 3 missile fired from Vandenberg AFB, Calif, is aborted and destroyed in flight. *Note similar incident in July.*

12 **Jun** 87 - Cape Canaveral, Florida.
Trident III launched successfully from flat pad.

21 **Jun** 87 - Alaska.
Three strong earthquakes shake parts of remote Alaskan Peninsula.

30 **Jun** 87 - Los Angeles, California.
Delta 767 flight 810 crew inadvertently shuts down engines shortly after takeoff from Los Angeles. Starts to fall, restarts engines and recovers. *Definitely crew error. Similar incident on Mar. 31, 1986 after takeoff from San Francisco.*

12 July 87 - Over Pacific Ocean.
Minuteman missile is destroyed in flight over the Pacific Ocean because of a problem in flight. Launched from Vandenberg. *Note similar incident the previous month.*

16 **Jul** 87 - Huntsville, Alabama.
Anomalous largo "boom" occurs at Huntsville, Alabama.

17 **Jul** 87 - Huntsville, Alabama.
Another anomalous, large "boom" occurs at Huntsville, Alabama. *Note that Huntsville is a pivot point on the Soviet grid. Adjustments in the grid often "spill out" energy, causing such a boom. It causes a sharp boom if pinpoint spillage occurs. It causes a more general quaking, rumbling sound of longer duration if spillage occurs over an area.*

20 July 87 - Cape Canaveral, Florida.
Tident III successfully launched from flat pad.

14 **Oct.** 87 - Nellis Air Force Gunnery range.
Top secret plane crashes; possible stealth fighter. *Did the Soviets kill another one?*

9 **Nov.** 87 - Vandenberg, AFB, California.
Minuteman II missile is destroyed seconds after launch. *Unnamed anomalies cited. Did the Soviets interfere?*

4 **Jan.** 88 - Madison and Huntsville, Alabama.
Giant radial sighted for several hours. Three mysterious booms rattle Huntsville. *Soviet scalar EM activity continues.*

APPENDIX V

*DEAD MAN FUZING** *THE REAL MEANING OF THE REYKJAVIK SUMMIT*

Soviet Foreign Minister spokesman, Gennadi Gerasimov, following failure of the Iceland Summit, said in a radio interview that SDI (U.S. Strategic Defense Initiative), as presently designed, is of no concern to the Soviet Union. The real Soviet concern, he indicated, is that the U.S. would make a great technical **breakthrough** and deploy new devices in space as a modification to SDI **presenting** an unacceptable threat to the Soviet Union.

Peter Jennings' ABC News Show, Los Angeles, CA
13 October 1986.

Gerasimov let the cat out of the bag!

With their deployed scalar electromagnetic weapons, the Soviets would immediately hold the winning hand-if these weapons could be unrestrainedly brought to bear.

As pointed out in **Fer-de-Lance: A Briefing On Soviet Scalar Electromagnetic Weapons,** Tesla Book, Co., 1986, the use of massive Soviet scalar EM weapons can only be gingerly applied. Activation and use of the Soviet scalar EM weapons is fraught with grave danger to the entire earth and to the whole human species because of the presence of nuclear facilities in the desired target area, and because of the "backlash" potential against the Soviet Union's own nuclear weapons and facilities.

The explosion at Chernobyl was apparently caused by the accidental failure of a nearby scalar EM transmitter, and the eventual impulsive loss of part of the electrogravitational potential built up by the transmitter before it failed. The resulting electrogravitational pulse (EGP) in the earth "pulsed" the nearest nuclear material... feul rods in the reactor at Chernobyl.

Had the EGP not been depleted prior to its escape, all four reactors at **Chernobyl** would have exploded with the same intensity as if they had been "triggered" by a fission explosion.

---

*The term "Dead Man Fuzing", means that, even though your opponent kills you, a weapon or effect of your own will inevitably result which will then kill your opponent.

In short, if unrestrainedly used, the EGPs from the Soviet scalar EM weapons would initiate most of the nuclear weapons where they are in the targeted area-in storage, in silos, in aircraft, etc. Target area nuclear reactors and nuclear wastes would also be violently initiated into nuclear explosion. The resulting vast megatonnage of nuclear explosions, much of them "in the dirt, " would provide an unthinkable holocaust. The effects of these giant explosions, deadly fallout and dense clouds of smoke and dirt, would blanket the earth with nuclear effects greater than is presently expected from an all out nuclear war, in which many of the weapons are expected to be destroyed or lost without being exploded nuclearly. Deadly nuclear radiation would blanket the globe, contaminating it for thousands of years. From the covering of the earth by dense smoke and dirt clouds, a new and deeply bitter global ice age-not just a "nuclear winter"-would almost certainly be initiated. In 1960 Khrushchev appropriately characterized these fantastic weapons by stating that they could wipe out all life on earth if unrestrainedly used.

A country's nuclear weapons and facilities, then, become its inadvertent "dead man fuzing" to prevent any substantial employment of large scalar EM weapons against it. So long as nuclear "dead man fuzing" is in place in a country, its adversary can only utilize scalar EM superweapons against it with great difficulty, and then only sparingly. At present, both the U.S. and its NATO allies are "dead man fuzed" by the presence of Western nuclear weapons-and to some extent, by the presence of nuclear reactors and stored nuclear wastes.

Obviously the Soviet Union's task of world domination—and the mind-rending risks it must take in employing large scalar EM weapons—would be greatly expedited if the obstacle of "deployed U.S. and NATO nuclear weapons and facilities" could be eliminated or minimized.

Also, Gorbachev must be acutely aware that the U.S. is finally beginning to look at the possibility of scalar EM weaponry. He would assume that either the U.S. is already on the way to acquiring scalar EM weaponry of its own, or else the time is imminent when it will do so. If this happened, and the U. S. then were to deploy SDI **scalar** EM lasers in space, the effects of each laser would be enormously increased by the electrogravitational amplification factor. Reasonable gain increases of the yield of each laser shot might approach, for example, on the order of 10^{20}. In that case, each laser could suddenly devastate a whole state or region with one or two shots, instead of just being able to destroy one little missile. The Soviets know that Reagan's Star Wars weapons-if modified to the "improved" kind and deployed in space—could deliver a surprise first strike in seconds, and devastate Russia before all the Star Wars space weapons could be knocked out by Soviet ground-based scalar EM beams.

The Soviets also know that it is not too difficult to modify lasers to become such powerful scalar EM lasers.

This means that Gorbachev needs to insure that the Star Wars weaponry is not actually tested in space. Testing in space requires developing at least prototype "space deployment" SDI weapons, to get them up in space to test them from there in the first place. And even such "prototype" SDI weapons would be sufficient to devastate the Soviet Union, given that the prototypes employed the technological breakthrough represented by scalar EM weapons. This is what Gerasimov was really referring to!

Thus Gorbachev simply cannot allow such space testing of SDI weapons, if he hopes to be able to safely seize the opportunity to employ his own already-deployed, massive scalar EM weapons during the "window of time" that would be provided by dismantling of U.S., NATO, and Soviet nuclear weapons.

And-given that he gets agreement and implementation of the "zero option"-Gorbachev knows he will only have a narrow "window" of time in which he can hope to achieve world domination. It can be only a short time until the U.S. produces results in scalar EM weapons, and promptly develops such weapons of its own. Faced with that eventuality, Gorbachev would then have no choice but to turn away from the Communist dream of world domination, and reach a mutually agreed accommodation with the West. Both the West and the Soviets then would have to take immediate and drastic measures to prevent the rapid development and spread of the relatively cheap scalar EM weaponry throughout other nations, particularly those motivated from fanatical beliefs and quite capable of "blowing up the earth" deliberately.

Thus, before the U.S. becomes aware of scalar EM weapons and builds its own, Gorbachev desperately needs to get rid of the "dead man fuzing" problem posed by the nuclear weapons and facilities possessed by the U.S. and its Nato allies, so that he can utilize his superweapons to dominate the world without destroying it.

And he needs to block any chance that the U.S. would be able to have space-tested high energy laser prototypes which could be quickly modified to scalar EM weapons, fired into space, and used to counter Soviet aggression or even destroy the Soviet Union in retaliation.

In short, today the paramount and almost desperate need of the Soviet Union is to reduce American dead-man fuzing and keep the SDI space laser genie bottled up in the lab on the ground.

Accordingly, in Iceland, in what appeared to be a stunning concession on nuclear arms, Gorbachev suddenly offered Reagan a complete dismantling of all nuclear weapons-Reagan's "zero option." Actually this move had been planned by Gorbachev from the beginning; it was not a spur-of-the-moment offer. The reason was simple: If the U.S. could only be persuaded to quickly get rid of its nuclear weaponry, the Soviets would hold an absolutely winning hand. Their massive, deployed, operational scalar EM weapons could then be unleashed with minimal restraint. The extremely delicate and incredibly dangerous problem of bringing the large scalar EM "continent-buster" weapons to bear, without inadvertently causing world destruction, would be vastly eased if only the ordinary U.S. nuclear facilities such as nuclear power plants and radioactive wastes had to be reckoned with.

Fortunately, Ronald Reagan fervently believes in defense .

He made the right decision, whether or not it was for the right reasons.

Indeed, he probably knew in advance that SDI was to be targeted in Iceland by the Soviets, and that sweeping concessions might be offered by them for that purpose. According to one report (" 'Blinded' Satellites," Rowland Evans and Robert Novak, **Washington Post,** Oct. 29,1986), just before Reykjavik the CIA-sent a report to Reagan with the warning that "the sole reason for Gorbachev's wanting the Iceland summit was to offer unprecendented inducements for Reagan to drop SDI."

Fortunately, Reagan rejected Gorbachev's proposal, because he does not wish to give up what he truly regards as a potential defense against incoming Soviet missiles.

Apparently neither Reagan nor his advisors were aware of the real reason for Gorbachev's unexpected and breathtaking proposal for what would essentially be near-total nuclear disarmament.

The Soviet ground-based scalar EM lasers could sweep the heavens of the presently conceived SDI weapons, as **Gerasimov** obliquely implied. However, this would not change the barrier represented by U.S. nuclear dead man fuzing, unless the nuclear weapons had been removed.

Meanwhile, Gorbachev's chess game to remove the dead man fuzing is certainly not over. The Soviets have mounted an intense propaganda campaign urging nuclear disarmament and cancellation or restraint of SDI. With a Democratic Senate now in, and with the approaching end of Reagan's term in office, the Soviet effort to clear the U.S. nukes-and the dead man fuzing-out of the way for Soviet deployed superweapons, while keeping the SDI genie in its laboratory bottle, can only be intensified. Already the Soviets are beating every drum to get substantial antinuclear segments of the U.S. scientific, congressional, political, and pacifist communities aroused to oppose SDI and urge a quick zero nuclear option.

The treaty for dismantling certain missiles in Europe has already been signed by the President, and its ratification by the U.S. Senate appears imminent. More is yet to come, and quickly.

In its aching desire for peace, if an unsuspecting America falls into Gorbachev's trap, shackles SDI and removes its dead-man fuzing, the Soviet Union will be free to enforce its will wherever and whenever it wishes.* If that happens, our beloved stars and stripes will be replaced with the atheistic hammer and sickle. And the noblest experiment of all—freedom for the common person—will have come to an ignoble end.

---

*Just as **Brezhnev forecast** for 1985, **at a** secret meeting in 1972 of the communist party leaders of E**urope.**

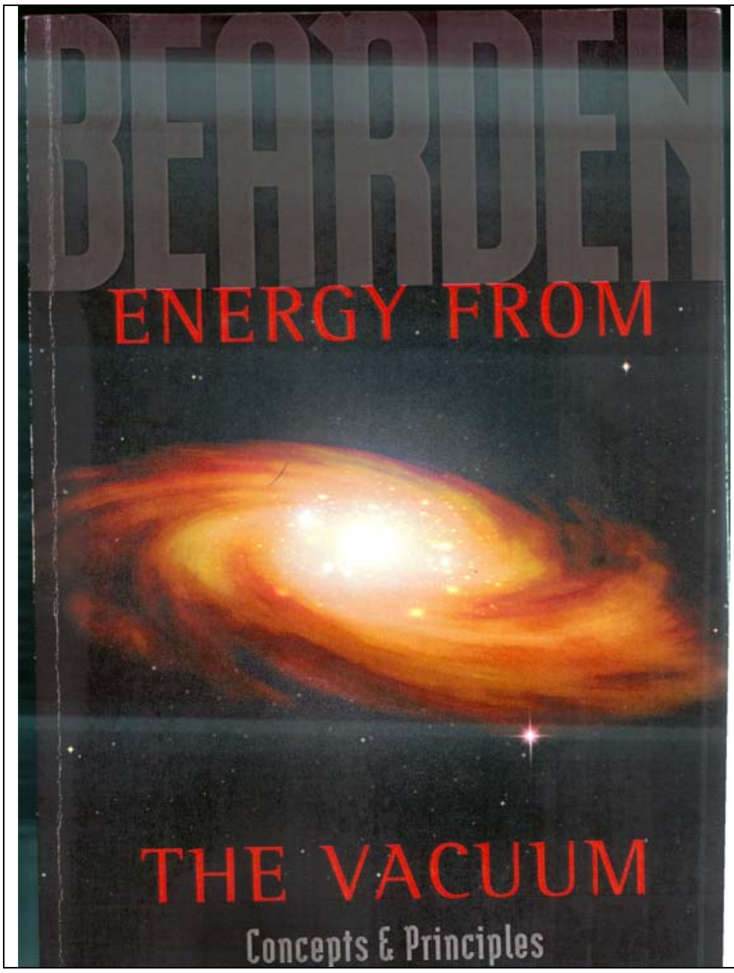

BEARDEN

ENERGY FROM

THE VACUUM

Concepts & Principles

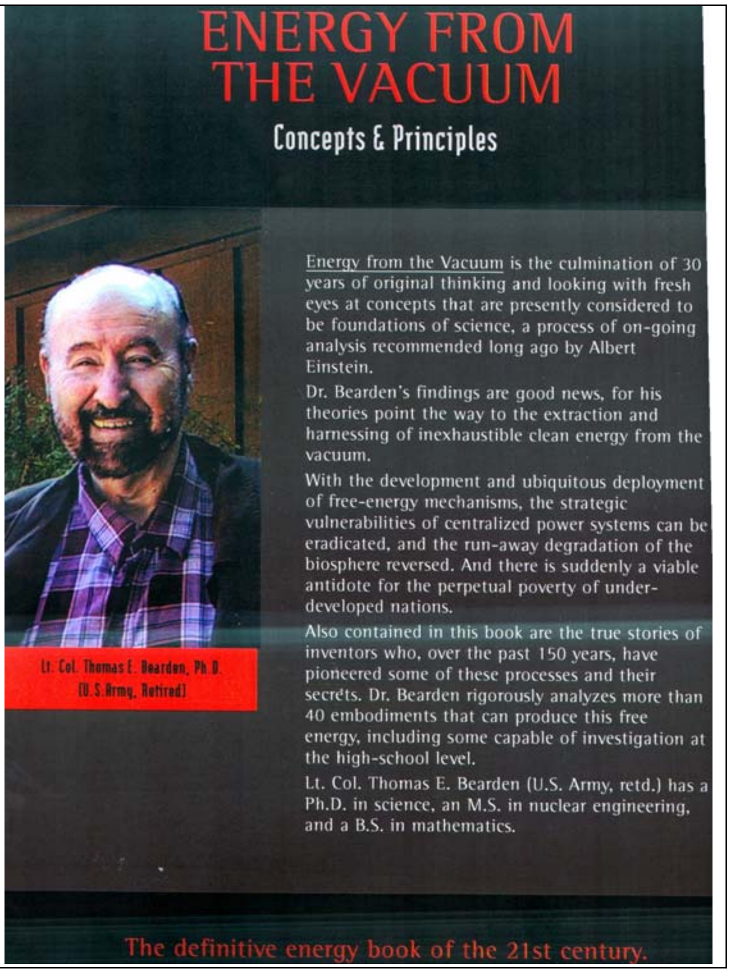

ENERGY FROM
THE VACUUM

Concepts & Principles

Energy from the Vacuum is the culmination of 30 years of original thinking and looking with fresh eyes at concepts that are presently considered to be foundations of science, a process of on-going analysis recommended long ago by Albert Einstein.

Dr. Bearden's findings are good news, for his theories point the way to the extraction and harnessing of inexhaustible clean energy from the vacuum.

With the development and ubiquitous deployment of free-energy mechanisms, the strategic vulnerabilities of centralized power systems can be eradicated, and the run-away degradation of the biosphere reversed. And there is suddenly a viable antidote for the perpetual poverty of under-developed nations.

Also contained in this book are the true stories of inventors who, over the past 150 years, have pioneered some of these processes and their secrets. Dr. Bearden rigorously analyzes more than 40 embodiments that can produce this free energy, including some capable of investigation at the high-school level.

Lt. Col. Thomas E. Bearden (U.S. Army, retd.) has a Ph.D. in science, an M.S. in nuclear engineering, and a B.S. in mathematics.

Lt. Col. Thomas E. Bearden, Ph.D.
(U.S. Army, Retired)

The definitive energy book of the 21st century.

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it. If you are an amateur electronic guy/girl try to make some good use of this knowledge and post your results. Try to use physics vocabulary as is presented in this book because as the author concludes, there are many good inventors around but they each use their own vocabulary and explanations of their inventions which are only understandable to them.
And please don't be naive and expect you'll be buying those devices in your local supermarket any time soon. Although some progress is being made it's still to slow, especially with the suppression factors watching over it. The knowledge has been handed to you and it's expected that at least a few of us make some use of it. Gather together and try to share your results with the rest of us.

TABLE OF CONTENTS:

Acknowledgements xi

Foreword xiv

Chapter 1 1

Foundations and Scientific Mindset 1

1.1 Introduction 1

1.2 Organized Science Often Resists Innovative Change 24

1.3 Scientific Disagreement or Street Fight? 32

1.4 A Few Significant and Unresolved Scientific Problems 36

1.5 Electrodynamics is Still Developing 51

1.6 A Summarizing Letter 55

Chapter 2 75

Background and Concepts 75

2.1 Comments on the Foundations of Classical Electrodynamics .. 75

2.2 Lorentz Regauging of the Maxwell-Heaviside Equations 108

2.3 The Suicidal Closed Unitary Current Loop Circuit 119

2.4 In Summary 122

Chapter 3 125

Giant Negentropy, Dark Energy, Spiral Galaxies and Acceleration of **the** Expanding Universe 125

3.1 Broken 5-Symmetry of the Dipole 127

3.2 Reinterpreting Whittaker's Decomposition of the Scalar Potential 137

3.3 **How** the Dipole Charges Transduce Time Energy 141

3.4 Reordering of the Local Vacuum: Giant Negentropy 143

3.5 Some Implications of Giant Negentropy 143

3.6 How Circuits Are Powered 153

3.7 The Heaviside Component vs. the Poynting Component 155

3.8 Dark Positive Energy: The Unaccounted Heaviside Component 157

3.9 A Short History of the Discarding of the Heaviside Dark Energy 158

3.10 The Problem of the Accelerating Expansion of the Universe **164**

3.11 In Conclusion **166**

Chapter 4 171

Setting the Stage for Understanding Overunity Power Systems 171

4.1 Introduction 173

4.2 Demonstration Gedanken Experiment **180**

4.3 Iterative Retrorefiection and Collection in a Circuit **194**

4.4 Tesla's Iterative Retroreflection Work **197**

4.5 Perspective on Tesla vs. Marconi 203

4.6 Conclusion **209**

Chapter 5 219

Selected Approaches to Overunity Power Systems 219

5.1 Introduction 219

5.2 Tesla's Shuttling of Potential Energy and Barrett's Extension 220

5.3 Moray's Radiant Energy Device 220

5.4 Anti-Stokes Emission and Similar Processes 225

5.5 Gain in Intensely Scattering Optically Active Media 227

5.6 The Fiber Fuse 230

5.7 Multiple Feedforward and Feedback Semiconductor Loops. ..235

5.8 Negative Resistors 236
5.9 Suppression of Free Energy Systems Has Been Common 288
5.10 In Conclusion 289
Chapter 6 293
Additional Approaches to Overunity Power Systems 293
6.1 Introduction 294
6.2 Energy Conversion and Energy Amplification 295
6.3 Other Systems of Electrical Interest 304
6.4 Some Magnetic Processes of Interest 341
6.5 Conclusion 381
Chapter 7 385
Aharonov-Bohm Effect, Geometric Phase, and the Motionless **Electromagnetic Generator** (MEG) **385**
7.1 Introduction 387
7.2 Aharonov-Bohm (Geometric Phase) Effect 390
7.3 Abbreviated History of the Geometric Phase 391
7.4 Possible New Aspect of the Curl-Free Magnetic Vector **Potential** 393
7.5 Use of Aharonov-Bohm Effect in the MEG 397
7.6 Collection of Field Energy in the MEG 403
7.7 Experimental Path to Development of the MEG 408
7.8 **Additional** MEG Functions 411
7.9 In Conclusion 419
Chapter 8 421
Approach to Antigravity 421
8.1 Introduction 422
8.2 Selected Theories of Gravitation 423
8.3 Selected Antigravity and Inertial Engineering Research Efforts 431

8.4 The Author's Engineering Approach to Local Antigravity.....454

8.5 In Conclusion..471

Chapter 9 473

The Supersystem and Remarks on Gravity, Antigravity, and Testing 473

9.1 The Supersystem and Its Considerations..............................476

9.2 Differences between Effects in COP»1.0 and COP<1.0 EM Systems..........479

9.3 Primarily Concerned With Dirac Sea Holes and Not Lattice Holes.........487

9.4 Peculiarities of Spacetime Curvature and Dirac Hole Interactions.......488

9.5 Charging and Discharging a Capacitor................................490

9.6 The Open-Loop COP>1.0 EM Power System...............................495

9.7 Remarks on Dirac Sea Holes, Currents, and Negative Energy..............501

9.8 Remarks on the Observed Acceleration of the Expanding Universe.........505

9.9 A Solution to the Hole Current Problem and Close-Looping..509

9.10 Negative Energy: Localized and Nonlocalized.........................510

9.11 Tests Reinforcing the Antigravity Mechanism.........................514

9.12 Susceptible Unitary System without Conversion of Dirac Sea Hole Current.........519

9.13 Unitary System with Conversion of Dirac Sea Hole Current.............521

9.14 Non-unitary Outrigger Array System..................................521

9.15 Some Instrument Considerations.......................................525

9.16 Still Anomalous Aspects..528

Chapter 10 537

Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy 537

10.1 Explaining Time as Energy ... 539
10.2 Mechanism Generating Flow of a Mass through Time ... 540
10.3 Time-Polarized EM Waves and Longitudinal EM Waves ... 542
10.4 Time Reversal Zone and New Nuclear Reaction Mechanisms ... 554
10.5 Revision of the Conservation of Energy Law ... 560
10.6 Some Resulting *Low Spatial Energy* Nuclear Reactions ... 566
10.7 Time-Energy Can Generate Instrumentation Anomalies ... 570
10.8 Additional Implications for the Future ... 577

Chapter 11 583

Final Perspective: Permissible COP>1.0 Maxwellian Systems 583

11.1 Introduction ... 583
11.2 Source Charges, Curved Spacetime, Efficiency, and COP....585
11.3 Truncations of Maxwell's Theory and Discard of System Classes ... 590
11.4 Examples of COP>1.0 Electromagnetic Systems ... 594
11.5 Motionless Electromagnetic Generator (MEG) ... 596
11.6 Physical Considerations of Gauge Freedom ... 599
11.7 Special Modeling of the Vacuum ... 600
11.8 Every Charge and Dipole in an EM System Receives and Transduces Vacuum Energy ... 601
11.9 Energy from the Vacuum Powers Every Electrical Power Line **and** EM Circuit ... 601
11.10 Significant Developments in Negative Resistance ... 602
11.11 In Conclusion ... 605

Annotated Glossary of Selected Terms 609

Appendix A 739
Appendix B 767
Index(notincludedbecauseyoucansearchinPDF)774
ENDNOTES 849

Acknowledgements

This book is dedicated to those lone researchers and stalwart inventors who have succeeded in building working model COP>1.0 electrical systems in the face of overwhelming odds and strong scientific opposition. We particularly mention Gabriel Kron , T. Henry Moray, John Bedini, Jim Watson, Ed Gray, Frank Golden, Floyd Sweet, Howard Johnson, Ken Shoulders, Bill Fogal, the Correas, and Teruo Kawai. There are quite a few others, of course, and no slight is intended by not listing all their names.

Also, I particularly acknowledge the contributions of my long-suffering close colleagues Dr. Lee Kenny, Dr. Jim Hayes, Ken Moore, and Steve Patrick, whose efforts on the bench have produced the working motionless electromagnetic generator (MEG) experiments as well as several variants. Our first MEG patent was issued in April 2002 and a second patent application has been filed. The MEG represents the first practical macroscopic application of the geometric phase (extended Aharonov-Bohm effect) for EM power systems. I particularly thank Lt. Col. (Retired) Ken Moore for watching my back and keeping me alive. I'm also grateful to both Dr. Fred Wood Sr. and Dr. Fred Wood Jr. for their personal support, action, and encouragement. I particularly acknowledge more than two decades of close cooperation and work with John Bedini, who is one of the most creative inventors on this planet; nearly a decade of work with Floyd Sweet and his vacuum triode amplifier; more than a decade of work with Frank Golden in COP>1.0 motors and uncurled A-potential communication systems; and a special thanks to Don Jensen for his important contributions and work in special combustion systems. Working with these inventors, I was able to physically examine working COP>1.0 power system prototypes and also experience and study the novel phenomenology that results in such systems.

With deep appreciation I acknowledge the kind assistance and helpful critiques of Dr. Myron W. Evans and other members of the Alpha Foundation's Institute for Advanced Study (AIAS) such as Alex Labounsky, Alain Beaulieu, and Bob Flower. Dr. Evans encouraged me to write down the basic concepts of the theory of permissible COP>1.0 Maxwellian systems I have been working on for three decades. This book is the result of that effort.

A special note of thanks goes to Dr. Dave Hamilton of the Department of Energy, who oversees a DoE restricted website provided for the important

ongoing scientific work **and drafts of the AI**AS. On that site, now there are more **than** 100 AIAS papers, many **of them dealing** with EM energy from the vacuum in rigorous fashion. Many of these AIAS draft papers have been published in leading journals such as ***Foundations*** *of Physics Letters, Physica Scripta, Optik,* etc. **This** is an invaluable scientific effort that clearly establishes the importance and potential for extracting copious EM energy directly from the active vacuum. It also extends the present U(1) electrodynamics to the much more extensive O(3) group symmetry electrodynamics, in which such things as I discuss in this book can indeed be accomplished on the bench and mathematically modeled.

The outstanding work and encouragement of Dr. Hal Fox and Dr. Eugene Mallove are particularly recognized, and I am grateful to them for their continued publication of important information and developments in cold fusion and the entire field of anomalous energy systems. I am also grateful to Dr. Robert Flower for his cogent explanations of some tough issues and points in quantum field theory that were beyond my ken before his explanation. Dr. Hal Puthoff introduced me to his cosmological feedback mechanism, and I am grateful for his important work with his colleagues in the areas of inertia, permissible thermodynamic extraction of energy from the vacuum, and possible mechanisms of antigravity.

A special thanks goes to Tim Ventura and to Jeffrey Cameron and the engineers of Transdimensional Technologies, Inc. for their courtesy and demonstrations of Transdimensional's successful lifter and rotary devices, and for an introduction to the field of inertial propulsion and antigravity now ongoing under the NASA Breakthrough Propulsion Physics (BPP) research program.

To Marcia Stockton I owe very special thanks indeed for her tireless efforts in reading and proofing the manuscript and translating the graphics. Also, I am much indebted to Michael Anderson for the art direction, and to Tony Craddock for making it possible. I am deeply grateful to Tony for his unfailing encouragement and support, and for operating and maintaining my website www.cheniere.org. I also deeply appreciate the continued contributions of Mike Rieker, of Network Innovations, who furnishes the website pro bono. A special thanks also goes to Dr. Steve Greer, who gave my wife and I great assistance and personal uplift at a time of personal tragedy. Dr. Greer has also now taken up the effort to develop and deploy practical electrical power systems, freely taking their EM energy from the vacuum.

Stalwart modern researchers in the overunity movement such as Jerry Decker and Walter Rosenthal are particularly to be commended for both their work, assistance to others, and their communication of work in the field. I am much indebted to Walt for furnishing the pictures of the closed-loop Sweet vacuum triode amplifier used in this book.

To Jean-Louis Naudin, I also am very grateful for his extraordinary experimental work and extensive publication of valuable information, his personal replication of a variety of systems, and his kind permission to use several illustrations in the book. His important website, http://jnaudin.free.fr/, is a cornucopia of information, experiments, and results in this entire area, with important links to many other appropriate web sites.

Also, to more than a dozen special persons — including those now passed on such as Glenn Foster and R. J. Reynolds III — who encouraged and supported these efforts through the years, I am very grateful. Throughout the years their kind words and friendship have made a very great difference to me personally and professionally, and they still do.

Finally and most importantly, I acknowledge the continuing and unwavering support of my dear wife, Doris, without whose encouragement and tolerance this work could never have been accomplished. She made all these long years of seven-day weeks and 18-hour days possible and humanly bearable. Without her, this book simply could not have been written.

Foreword

The original purely scientific purpose of this hook was interrupted en route by two events: The first was a sudden heart allack and onslaught of severe hypoxia resulting from the resurgence in Spring 2001 of a 33-year old chronic mycoplasma infection, of the modified biological warfare kind produced by a combined U.S./Canadian team, which got loose in Canada and Florida in the 1950s. I contracted this "mysterious, unknown ailment" in 1968 while in the U.S. Army and stationed in Quebec, Canada as the U.S. Army Liaison Officer to the Canadian Armament Research & Development Establishment (CARDE). The second event was the sudden terrorist attack on the U.S. on September 11, 2001, which propelled this nation and others into a war against international terrorism.

For the mycoplasma, Kenn Thomas recognized and informed me exactly what the difficulty was, and I was finally tested and diagnosed correctly in December 2001 due to the efforts of a courageous family physician. Accordingly, in December 2001 we started a yearlong (or longer) regimen of antibiotics and supplementary oxygen, with additional medication to control the runaway heart fibrillation associated with this virulent form of the disorder. If we make it through this yearlong program, we will probably survive with at least some significant recovery. Now on a much-reduced regimen, this made me realize the absolute necessity of passing the "energy from the vacuum" information and baton, so to speak, to those sharp young graduate students and post doctoral scientists who are interested in this area. They simply must not have to spend 30 years of their lives getting to this point. Instead, they should start where I am now, correct any errors I have inadvertently made, and go much farther.

Accordingly, we sharply changed the intent and content of the book to comply with that new purpose. We also deeply considered the potential U.S. and World energy implications of that second shocking event of 9/11 when international terrorists struck the World Trade Center in New York Center and the Pentagon, having also targeted the White House.

In the new asymmetric form of warfare, the internal aspects of a nation - its civilian population, its vulnerable installations, its energy infrastructure, its food crops, and its transportation modes — are the targets. Terrorist teams and groups in significant numbers are already inside every targeted developed nation, and these teams have weapons of mass destruction including biological weapons and even nuclear weapons. For confirmation of the hostile nuclear weapons already secreted in large U.S. cities, read Lunev's book[1] to see how the former Soviet Union brought them in, along with the Spetznaz teams to detonate them on call. The nuclear weapons and the Spetznaz teams are still here. Other hostile nations have inserted terrorist teams with weapons such as anthrax, smallpox (camel pox is nearly identical to smallpox, and widely available), bubonic plague, etc. At the end of the 1991 Persian Gulf War, most of Iraq's calutrons and its nuclear program escaped the weapon inspectors and were hidden. In the few calutrons we did get, weapons-grade U235 residues were found. In a formal threat study for a foreign nation, begun before the Gulf War and delivered after its end, my analysis team and I estimated that Iraq had produced sufficient U235 in 1991 for from 7 to 12 nuclear weapons, particularly of the gun assembly type. My personal estimate now would be that Iraq has made sufficient weapons grade U235 to build three dozen to six dozen nuclear weapons. If so, then some of those weapons are undoubtedly long since spirited into the United States, and planted in our large cities.

In modern asymmetric strategic war, the first phase of the war is conducted in "peacetime". The first phase is to insert the weapons of destruction into the targeted nation and hide them at the intended targets, along with the teams to detonate and unleash the weapons when desired. In that sense, the first phase of WW III has already been accomplished.

According to Vice-President Cheney,[2] *"The war on terrorism will not be over in our lifetime. It is different than the Gulf War was in the sense that it may never end. At least not in our lifetime. The way I think of it is, it's a new normalcy."*

With much of the cheap oil of the world in its borders, the MidEast is a powder keg ready to explode. If a clash between the U.S. and Iraq occurs, Saddam Hussein has already clearly shown that he will attack and destroy his foe's energy infrastructure and he will also use weapons of mass destruction. Simply witness the hundreds of oil wells set afire in Kuwait by Iraqi forces during the Persian Gulf War, and the use of chemical weapons on his own people. As this is being written, Pakistan and India are skirting the brink of war, and both have nuclear weapons. Either side that decides on war **will have to make a n**uclear strike on the other. They have no antimissile defenses, hence **making the** "first knock-out" or pre-emptive strike is each **nation's** only **chance of survival.** Hopefully cooler heads will prevail, but **If that** war erupts, the Mid East, oil, and present energy infrastructures will never again be the same, for the entire world.

---

[1] Stanislov Lunev and Ira Winkler, *Through the Eyes of the Enemy,* Regnery, Washington, D.C., 1998, p. 26.

[2] Vice President Richard B. Cheney, October 21, 2001.

There are other very serious U.S. vulnerabilities of gas and oil pipelines, port facilities, Gulf of Mexico oil rigs, nuclear power plants, power substations, hydroelectric dams, long power transmission lines on towers, etc. that are too numerous to enumerate. With but a moment's reflection, one realizes that even normal and cheap guerrilla-type or commando-type forces with C-4 explosives and timers can destroy a very substantial portion of the energy infrastructure with minimal risk and loss of assets. For those who wonder about a nuclear power plant's vulnerability, visualize a seized Boeing 747 loaded with fuel and flown directly into the main reactor container . If meltdown occurs, the nuclear tragedy at Chernobyl in 1986 will appear tame by comparison. There are also several other electromagnetic ways (such as a portable EMP "shooter" hauled in the back of a personal SUV) to instantly destroy the electronic controls of the nuclear power plant, seriously risking runaway and core meltdown.

Consequently, increasing attention is now focused on the extreme vulnerability of the centralized energy infrastructure in the U.S. and every other developed nation. So what started for me as a purely scientific endeavor to set out the concepts and principles of permissible COP>1.0 EM systems suddenly assumed a role of potentially critical strategic significance to the U.S. and to the entire developed world. Quite simply, some 20,000 terrorists in teams already in the United States can - at this very moment —destroy large portions of the nation's energy infrastructure at will, easily and cheaply. The economy of any modern developed nation depends on the availability of cheap energy. So this kind of hostile strategic capability means that the power to readily destroy a nation's economy — and thus defeat a mighty nation itself in the ensuing economic collapse — has passed into the hands of hostile forces already in country and awaiting the order to commit their assets.

In this new asymmetrical form of warfare, it would appear that the rapid development and deployment of significant distributed power sources is now a critical element of national survival for many nations of the world. Decentralization **does** little **good, however, if fuel** for the decentralized generators must remain **centralized and distributed** by very large and complex systems.The **fuel and Its distribution** are still extremely **vulnerable resenting** easy **targets allowing paralysis** of the decentralized generators. Self-refueling *{self-powering)* of the electrical power systems is therefore a mandatory additional requirement.

So sober assessment dramatically reveals the need for self-powering electrical power systems, freely extracting their EM energy from the local vacuum. Thus an engineerable theory of permissible COP>1.0 EM systems has become essential to the national survival of the U.S. and many other nations, so that an effective decentralized energy infrastructure — with the seething vacuum as the "localized and inexhaustible free fuel supply" — can be rapidly designed, developed, and deployed.

Sadly, the leaders of the scientific community are still in a "business as usual" mode with respect to energy, and intend to remain so. They do not even recognize what powers and *always has* powered an EM circuit and system: EM field and potential energy extracted directly from the vacuum by the broken symmetry of the source charges and dipoles. Since no scientific research on EM energy from the vacuum is going to be launched by the scientific leaders at the top, then the results of the broken symmetry discovered in 1957 are going to have to be applied in the scientific community from the bottom up. Consequently, a rather desperate need to communicate everything possible to the young graduate students and post-docs has become imperative, if the nation is to survive.

From the beginning of my effort in COP>1.0 systems, and then my effort on this book and prior to the events leading to the attack on September 11, 2001, the appalling *humanitarian* need for COP>1.0 electrical power systems taking their energy from the vacuum has been obvious and overwhelming. One third of the human population usually goes to bed hungry at night. One third is infested with worms and parasites of one kind or another, in addition to other diseases. Simple things such as clean water, common sanitation, medical treatment, and jobs to support one's family are difficult things in many parts of the world.

In the final analysis, the only way to get a suffering, impoverished, and undeveloped nation up and going is to provide cheap clean energy, and to do it with systems which get their "fuel" for free, such as from the vacuum itself. Once cheap clean energy and self-powering systems are widely available, then the economy and infrastructure of that nation can be slowly and steadily built up to get the society moving. Roads, bridges, hospitals, schools, teachers, industry, and jobs then follow as the economy rises. Until there is a cheap energy infrastructure available and growing, essentially fuel-free, none of that is going to happen in most impoverished nations because presently the world's fuel is monopolized, controlled, and

fought over, and it will become even more expensive as the supply vs. **demand** ratio falls, prices escalate, and additional wars - almost certainly involving weapons of mass **destruction** - are engendered. With present aid concepts, at best the politicians and warlords usually get the bulk of whatever assistance is sent in, and the warlords on the take from foreign masters will continually plunge the country into turmoil, war, and human tragedy. *Assistance is necessary,* but the eventual aim has to be to enable and empower the common citizens to be self-supporting, educated and trained, relatively free from disease, and living in a nation having a growing and developing infrastructure, jobs, and a decent economy. Availability of cheap and affordable energy — and *clean* energy if we wish to save the precious environment — is the primary key to "freeing and empowering the developing nations" as well.

Along with this omnipresent, desperate humanitarian need, the increased strategic need for decentralized energy systems freely fueled from the vacuum itself has led me to make every effort possible — particularly after the heart attack in 2001 and in continued hypoxia from severe chronic mycoplasma infection — to finish the book and at least outline the major concepts and principles that enable the construction of a sorely-needed COP>1.0 electrical power systems theory, followed by the development and rapid deployment of actual self-powering systems. The Alpha Foundation's Institute for Advanced Study (AIAS), spearheaded by Dr. Myron Evans, has also been vigorously preparing and publishing papers in leading scientific journals and other publications[3] showing that EM energy from the vacuum is permissible, possible, and feasible.

In my three decades of work with multiple inventors of working prototype COP>1.0 electrical systems, I have been privileged to experience and work with several actual COP>1.0 circuits and prototype systems, study them, and slowly comprehend some of the advanced and unusual principles involved. In short, thanks to the courtesy and kindness of others, I have been able to gradually acquire at least an elementary understanding of the mechanisms which enable COP>1.0 EM power systems. With this background, one can understand why conventional EM systems and circuits do not develop COP>1.0 or self-powering, even though all the energy collected and used in every electrical power system comes from the vacuum and not from the shaft energy input to the generator, chemical energy dissipated in a battery, wind energy collected by a windmill (that just "feeds" the shaft of the generator), or from the water power dissipated in a hydro turbine (the hydro turbine also only turns the shaft of the generator, which does not actually furnish the power to the external power line connected to it).

---

[3] E.g, *Modern Nonlinear Optics,* Second Edition, 3 Vols., edited by M. W. Evans, Wiley, 2001; M. W. Evans et al, "Derivation of 0(3) Electrodynamics from the Irreducible Representations of the Einstein Group," *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-187; "Anti-Gravity Effects in the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 601-605; "Classical Electrodynamics Without **the Lorentz Condition: Extracting Energy** from the Vacuum," *Physica* ***Scripta,*** **61(5),** May 2000. p. 513-517; "Runaway Solutions of the Lehnert Equations: The **Possibility of Extracting Energy from the Vacuum,"** *Optik,* 111(9), 2000, p. 407-409.

It seems appropriate that other researchers should be able to start where 1 am now, understand what I have learned or think I have learned, find and correct any errors, and simply go forward with alacrity — much farther than I have been able to go, and much faster. This is particularly true of the young researchers who have excellent theoretical and experimental skills and a vigorous lifetime ahead of them. When an old and tiring runner falters with the baton, it is time to pass it on to the fresh young runners who can carry it to the successful end of the race.

The purpose of this book, then, is to lay out in some little detail what I have learned in all these years of sustained effort, but in such a way that it ties directly into good physics. In some cases, the tie-in to physics is unavoidably to a somewhat esoteric area such as the geometric phase and to time as a special form of energy. COP>1.0 EM systems are not simple, and they involve special phenomenology rarely encountered in conventional EM circuits and systems. COP>1.0 EM systems simultaneously involve a combination of particle physics, modern active vacuum theory, advanced materials science, higher group symmetry electrodynamics, quantum mechanics, and general relativity. They also involve areas where foundations of physics are presently in error. We found it necessary to think in terms of the *supersystem,* defined as having three components, which are (i) the physical power system itself, (ii) the active local vacuum, and (iii) the active local curvatures of spacetime. All three components of the supersystem interact and exchange energy with each other continuously. Any legitimate COP>1.0 power system must freely receive, collect, and use some excess energy from its active environment — in this case, the latter two components of its supersystem. By definition, this COP>1.0 function requires disequilibrium between the system and its active environment, so that classical equilibrium thermodynamics does not apply. Instead, the thermodynamics of systems far from **equilibrium** in **their** energetic exchange with an active environment applies.

However, since the "environmental energy" we are extracting is virtual photon energy from the vacuum and energy from local curved spacetime, the fundamental definitions of classical thermodynamics and classical electrodynamics have to be re-examined and altered. A closed system, e.g., can no longer be defined just as one closed to the transfer of mass across its boundary, but permitting energy to exchange across it. Since the advent of general relativity in 1915, we have known that energy and mass are the same thing. Whenever any system changes its energy, it also changes its mass by the well-known $E = mc^2$ rewritten as $dm = dE/(c^2)$. Further, in the fierce virtual particle exchange of the vacuum with every part of every system, both virtual mass and virtual energy continually cross the system boundary. Neither classical thermodynamics nor classical electrodynamics deals with the transduction of time-energy into spatial energy in a system, which is a totally different type of energy exchange accomplished by charge.

Further, certain phenomenological areas — such as strong gradients — are already well known to violate present thermodynamics. These are presently being researched under the aegis of *extended nonreversible thermodynamics.* Finally, as the present laws of classical thermodynamics are stated in the textbooks, *every charge and dipole in the universe already violates those laws, and has done so since the beginning of the universe.* Let us put it bluntly: Every charge in the universe already freely and continuously pours out EM energy in 3-space in all directions, without any observable EM energy input. That is the well-concealed *source charge problem,* known but ignored by the leaders of the scientific community for. a century. *All* EM fields and potentials and their energy come from those source charges, according to electrodynamics itself. Either we must give up the conservation of energy law entirely, or else we must accept the fact that unobservable *virtual* EM mass and energy are continuously absorbed

The vacuum by the source charge, transduced into real observable EM energy, and then re-radiated in 3-space in all directions as observable EM energy, creating the associated fields and potentials reaching out across the universe. That this problem continues to be ignored by the scientific community — some 45 years after the basis for its solution has been proven and is well known in particle physics — is totally inexplicable. But it continues to be resoundingly ignored, which means that *the scientific community itself is and remains totally responsible for the energy crisis and the associated vast despoilment of the biosphere, because it adamantly* *insists on using a seriously flawed electrical energy science whose hoary basis is more than a century old.*

Accordingly, we have simply extended the first law of thermodynamics appropriately, pointed out that the second law does not apply absolutely but only as an approximation when conditions are appropriate, and revised the third law and the zeroth law. In this book we simply show the results of that revision, particularly in Appendix A.

We were also able to trace the history of electrodynamics, and to find exactly how the permissible Maxwellian systems far from equilibrium with an external active environment — which are included in Maxwell's 1865 seminal theory and even in Heaviside's vector truncation later — were arbitrarily discarded by Lorentz with his symmetrical regauging of the Maxwell-Heaviside equations. The purpose of such arbitrary regauging was and is to provide simpler equations that can usually be solved analytically without resorting to numerical methods. In the name of mathematical simplicity and ease, that entire great class of Maxwellian systems far from thermodynamic equilibrium in their supersystem was just arbitrarily discarded. Those Maxwellian systems are still arbitrarily discarded in every electrical power-engineering textbook today.

So the first requirement for a COP>1.0 EM system becomes *building-in a mechanism to violate the Lorentz symmetry condition, at least for a significant fraction of the time.* Further, the standard closed current loop circuit — including the entire external circuit in the same closed current loop as the source dipole inside the generator is what already *self-enforces* the Lorentz symmetry condition. Consequently, in any legitimate COP>1.0 electrical system, the operation of that closed current loop circuit must be violated for some portion of the system's operation. This of course places the system in disequilibrium with its active supersystem environment, allowing intake of extra energy from the environment and therefore permitting COP>1.0. So one of the immediate requirements in COP>1.0 electrical systems research is to seek out and find known or suspected mechanisms that accomplish just such violation of the Lorentz symmetry condition. In the book, I show a sampling of the various mechanisms I found other inventors using or that I uncovered myself, or that already exist in physics.

Finally, we discovered (and found the basis for it was shown in the hard literature in Heaviside's original work) that every generator and battery as an energy transducer already extracts from the vacuum and outputs from its terminals a far greater *nondiverged* energy flow than is accounted in the

Poynting diverged energy flow theory (which captures only that small **fraction that** is **diverged** into the circuit and captured by it). The total output energy flow from the terminals of a generator is also far greater than the mechanical energy input to the shaft of the generator. So in its *energy transduction* function, every dipolar electrical circuit ever built is already a "free energy circuit" and a COP >>1.0 system, extracting its utilized electrical energy from the vacuum and extracting additional energy also. But presently the electrical power system is unwittingly designed and constructed by engineers to deliberately destroy the source of that enormous energy flow from the vacuum faster than the system can use some of it to power its load. Wryly put, that does not seem to be the proper way to build energy systems or "run an energy railroad," even though it's always been done that way.

This led to recognition of a great scientific faux pas, perhaps the greatest in history. One must keep one's sense of humor! So with tongue in cheek, let us state it this way: *There is not now, and there never has been, a single electrical engineering department, electrical engineering professor, or electrical engineering textbook that teaches what powers an electrical circuit. Yet the basis for it has clearly been in particle physics since 1957, as certified by the award of the Nobel Prize to Lee and Yang. What powers every electrical circuit and system is the broken symmetry of the primary power system's source charges and source dipole (i. e., the broken symmetry of the opposite charges on the ends of the dipole) in its fierce energy exchange with the seething vacuum, once that dipole has been made. The modern "charge " is also a dipolarity, once its associated clustering virtual charges of opposite sign are considered. Further, any dipole will freely extract EM energy from the vacuum, convert it into real, usable, observable EM energy, and pour out real, usable EM energy continuously and indefinitely, so long as the dipole remains intact.*

The external circuit attached to a generator is not powered by the mechanical energy that is input to the generator shaft. The external circuit attached to a battery is not powered by the chemical energy expended by the battery. Astonishingly, every electrical power system and electrical circuit ever built, and every one built today, was and is powered by energy extracted and transduced directly from the active vacuum by the broken symmetry of the opposite charges on the ends of the source dipole (or a charge considered as a modern dipolarity), once that dipole is formed. A Nobel Prize was awarded to Lee and Yang in 1957 for the prediction of broken symmetry (strongly in 1956). Wu and her colleagues proved it **experimentally** in early 1957. So revolutionary a change in physics was that startling discovery of broken symmetry, that — in a nearly unparalleled action — the Nobel Committee awarded the Nobel Prize the same year, December 1957, to Lee and Yang. *One of the broken symmetries proven was that of opposite charges such as are on the opposite ends of every dipole, or such as constitute a classical "isolated charge" when the charge and its clustering virtual charges of opposite sign are considered as in modern physics theory. Rigorously, every source dipole is continuously extracting EM energy from the vacuum and pouring it out of the "terminals" of the dipole in all directions in 3-space. And so is every charge as well.* That the leaders of the scientific community continue to ignore this solidly proven fact is astonishing. But they do.

We also came to realize that the great Gabriel Kron, who was never allowed to reveal the secret of his *open path,* had in fact discovered the precursor of that very broken symmetry of opposite charges (of dipolarity). His open path is simply the recognition that any two points in the universe, that are at differing potentials and therefore can be considered as a dipolarity of opposite charges, acts as a source dipole that already extracts virtual photon energy from the vacuum, transduces it into real, observable photon energy, and pours out that energy in all directions. In reaching that understanding, he had applied general relativity to rotating electrical machines, simulated various forms of electromagnetics equations, and dealt deeply with quantum mechanics and the Schrodinger equation.

Sadly, the electrodynamics model used in electrical power engineering still does not even model the active vacuum, *much less* a broken symmetry in the virtual photon exchange between that active vacuum and all the charges in every electrical system. The electrical engineering model does not model the *supersystem,* hence cannot be used to design and produce COP>1.0 systems. *A priori,* that seriously truncated and crippled model arbitrarily discards all permissible Maxwellian systems far from thermodynamic equilibrium (i.e., with broken symmetry) in their vacuum flux exchange. And those are precisely the systems that are permitted to perform the five magic functions of great interest to the COP>1.0 energy researcher. Such a disequilibrium system can permissibly (i) self-order, (ii) self-rotate or self-oscillate, (iii) output more energy than the operator inputs (the excess energy is freely received from the active environment), (iv) power itself and its load simultaneously (all the energy is freely received from the active environment), and (v) exhibit negentropy. *Every charge and dipole in the universe already performs all five magic functions.*

It is **the purpose** of this book **to** clearly show that all EM systems are already powered by vacuum energy, and always have been. We cite the references and experimental proof that this is true, and the interested reader can check the extensive references himself or herself. The intent is to show that only a part of what has been discovered and is already known in particle physics is presently included in the woefully inadequate and ancient electrical power-engineering model. And frankly, unless there is a groundswell from the bottom of the scientific community, the leadership of the scientific community has no intention whatsoever of funding or allowing significant research in extracting copious EM energy from the vacuum (very simple) and learning to use it effectively (more difficult). A single 200 million dollars of good research in EM energy from the vacuum would solve the energy crisis forever, and dramatically reduce the pollution of the biosphere. The scientific community will spend many times that amount on their favored "big science" projects, but not one cent on EM energy systems powered by energy extracted from the vacuum.

Also, we show a sampling of the present work ongoing in inertial propulsion and antigravity research. Working with Sweet, in the 1980s I designed and Sweet performed a highly successful antigravity experiment.[4] I share with the reader the concepts and theoretical basis for that experiment, why it worked and how it worked, and the excellent results achieved. One of the tragic losses to science was the loss of the Sweet COP»1.0 vacuum triode amplifier (VTA) system with the death of Floyd Sweet. Nonetheless, we apply the mechanism we conceived so long ago to propose a mechanism generating the antigravity recently confirmed to be accelerating the expansion of the universe. We offer the long-neglected, Lorentz-discarded giant nondiverged Heaviside energy flow component accompanying every field-charge interaction as the probable generatrix of the excess gravity holding the arms of the spiral galaxies together. And we offer that same component — when connected with a flow of *negative* EM energy and with *negative energy* EM fields and potentials — as the specific generatrix of the excess antigravity that is accelerating the expansion of the universe. Both theses are testable on the laboratory bench.

Our target audience is several fold. Even though presented conceptually, the material is not oversimplified, and it is targeted at (i) the sharp young grad students and post-docs interested in extended (higher group internal symmetry) electrodynamics and in researching and developing COP>1.0 electrical power systems, (ii) the forward-looking electrical engineering professors who will at least consider that extended electrodynamic models and the laws of nature do permit COP>1.0 electrical power systems, and (iii) the still open-minded leaders of the scientific community, such as the National Academy of Sciences, National Science Foundation, the U.S. Department of Energy, the great national laboratories, etc. In *one area* of physics much better models of electrodynamics enabling COP>1.0 power systems, antigravity, and novel propulsion systems already exist (and a few such systems *experimentally* exist in the hands of independent inventors as well), while in a *second area* of physics that part of electrodynamics dealing with permissible COP>1.0 EM systems and unified field theory has been arbitrarily excluded at great cost to society, to the environment, and a tragedy to science and technology. Additional secondary targets are (iv) governmental leaders responsible for advanced research and development planning and funding, to argue that the "energy crisis" is a direct result of an inexplicable and sustained shortfall in the scientific mindset, and not as a result of the laws of nature and the need for cheap oil, coal, and gas or nuclear power plants, (v) science writers, journalists, and entrepreneurs to sensitize them to (hopefully) a coming energy revolution of unparalleled proportion, (vi) strategic and tactical planners in our Department of Defense to make them aware that the great logistics burden of modern highly maneuverable armed forces can and will be dramatically lowered by rapidly developing and using self-powering power systems "fueled" directly by the local active vacuum, and by adapting novel new propulsion methods, (vii) theoretical and experimental physicists to make them aware that the easiest thing in the world is to extract copious EM energy freely from the vacuum; just pay *once* to make a little dipole or assemble a charge and do not let it be destroyed, (viii) the Department of Energy upper level managers to bring into their awareness the coming advent of (and terrible need for) self-powering decentralized EM power systems worldwide, and the pressing need for a major role to be played by the DoE, (ix) the struggling but serious COP>1.0 researchers and inventors, to give them a technical overview — rather than just anecdotal material — of much of what has been done, tried, and conceived, and what seems to be successful and what is still to be determined, (x) the environmentalists devoted to cleaning up the biosphere and saving this precious planet, to make them aware that the answer to the world's energy problems and saving the biosphere cannot come to them from their conventional scientific consultants, who are trained in applying the very scientific disciplines responsible for the problem in the first place, and (xi) **the long suffering undeveloped** and impoverished nations and peoples of the earth. Our fervent message is one of hope after all, *once a* ***change*** *occurs in the scientific mindset.* There is a cheaper, cleaner, and better way to get their economies and nalional infrastructures up and running, and it is imminently doable.

---

[4] Floyd Sweet and T. **E. Bearden,** "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc 26th Intersoc.* ***Energy Conversion*** *Eng. Conf. (IECEC '91),* Boston, Massachussets, 1991, p. 370-375.

The main message of this book is that there is not now and there never has been an *electrical energy availability* problem. Every large centralized power generator actually outputs enough EM energy flow, extracted directly from the seething vacuum, to power all electrical loads on earth, if all the energy pouring from its terminals were caught and utilized. However, there is certainly an *EM energy flow interception* problem, to get more of the available energy flow from a source dipole intercepted and diverged into the circuit in a manner not complying with the Lorentz symmetrical regauging condition. Also, once the excess energy has entered the circuit and been collected, there certainly is a problem in *using the intercepted and collected energy in the circuit without destroying that part of the generator or battery — the source dipole — that is actually extracting all the energy from the vacuum.* Or at least to use the circuit's collected energy in such fashion as to destroy the source dipole in the generator *slower* than the energy is used to power the external load.

That conglomerate of the two sub problems — the "energy flow *interception, collection, and usage* problem — is the single energy problem. There is no other energy problem, and there has never been any other. One finds it very ironic that, in the conventional agencies, universities, the DoE, the National Science Foundation, the National Academy of Sciences, the national laboratories, etc., not one of them is working on the real, solitary electrical energy problem, even though they are importantly spending billions of dollars purporting to do so.

We fervently hope that the content of the book will reach a sufficiently interested audience to begin to change that century-old erroneous scientific mindset. We hope that the book, along with the important contributions of other COP>1.0 inventors and researchers having success, will spark a new creative drive in the scientific spirit that solves the problem of cheap, clean energy in three years or less. This is a doable. I firmly believe in the scientific method and what it can do once it is unleashed from the century-old dogma that COP>1.0 systems are against the laws of nature, physics, and thermodynamics. Just turn the young scientific and engineering tigers loose on the problem, give them some funding, and watch them do it.

As we stated in a formal rebuttal to a charge of perpetual motion nonsense, it only takes one white crow to prove that not all crows are black. And with respect to COP>1.0 EM systems far from thermodynamic equilibrium with their active environment, every charge and dipole in the universe is already a "white crow" pouring out EM energy freely and continuously with the speed of light, and exhibiting COP = ∞. This is in fact reluctantly recognized in physics, but it has essentially been buried from the attention and knowledge of generations of electrical engineers and most physicists.

A COP>1.0 experiment well known to nonlinear optics is the Bohren-type experiment, demonstrating the known negative resonance absorption of the medium. A typical experiment of that type outputs some 18 times as much EM energy as the experimenter inputs by standard calculations. We note that a single replicable experiment can and does negate any theoretical model that contradicts it. Hundreds of such materials experiments proving the feasibility of COP>1.0 EM systems are performed every year in any major nonlinear optics laboratory without a second thought. Together with the known but avoided source charge problem and proven broken symmetry of any dipolarity, it is almost unthinkable that our electrical engineering departments still do not incorporate the active vacuum exchange — and every EM circuit's broken symmetry in it — into the classical electrodynamic model used. In short, to really understand what powers every EM circuit, just now the electrical engineer would have to go across campus to the particle physics department and investigate broken symmetry of opposite charges — as are on the ends of the source dipole, once created in the generator or the battery, or as are involved with the single "isolated charge" considered with its associated virtual charges of opposite sign.

It is high time that our electrical power engineers and scientists heed the message of all these demonstrated white crows of nature, and begin developing and deploying self-powering COP>1.0 electrical power systems with alacrity.

We have a biosphere, a planet, a strong world economy, and an uplifting for the long downtrodden poor people of the world to recover. We strongly urge the scientific community to accept the challenge and meet it head-on.

Tom Bearden

22 July 2002.

Chapter 1

Foundations and Scientific Mindset

> *"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."* [Albert Einstein]{1}

1.1 Introduction

Einstein's quote is a beautiful statement that scientists should never take the present understanding — and the present models — as absolute. That is the approach the present author has taken for some 30 years, in a struggle to comprehend that class of electromagnetic (EM) systems that are open systems in disequilibrium in their *virtual photon energy exchange* with the active vacuum, and exhibiting a broken symmetry in that exchange. The statement beautifully expresses that the major problems encountered in grappling with such EM systems have been the existing errors and non-sequiturs in classical electrodynamics and other parts of physics. Many difficulties have resulted from the continued propagation of a 137-year old classical electrodynamics model formed before electrons, atoms, nuclei, positrons, the active vacuum, special and general relativity, quantum mechanics, etc. were discovered or known.

1.1.1 EM Foundations Are Incomplete and Contain Errors

In any model, there are many assumptions. Even when a model is well-fitted and well-developed, it still applies only when the foundations assumptions on which it is based are true or are not too much in error. Whenever one or more of the fundamental assumptions is violated by phenomena uncovered, then there is a new class of phenomena where the model does not hold, or does not hold well and is only an approximation at best. In that case, either the existing unsatisfactory model must be improved and advanced, or a new model must be constructed.

Great scientific minds continue to point out Einstein's beautiful principle in different words. E.g., expressing the thought of Stephen Hawking, one of the great physicists of our day:

> *"All we ever know is our models, but never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself.*" [As stated by George Zebrowski] {2}.

Excellent scientists — Feynman, Wheeler, Bunge, Evans, Barrett, and many others — have indeed pointed out that classical EM theory is seriously flawed. In the words of Bunge {3}:

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state."*

The author also found it imperative to return to many of the original seminal papers of physics, particularly in electrodynamics. The major concepts in those papers led to the present classical EM model. This was particularly true of the work of Poynting {4a, 4b} and Heaviside {5a-5c}, who independently and simultaneously arrived at the notion of the flow of EM energy through space[5]. Their work occurred in the 1880s, after Maxwell was already deceased. It also necessitated reviewing Lorentz symmetrical regauging of the Maxwell-Heaviside equations, where Lorentz arbitrarily discarded all permissible COP>1.0 Maxwellian systems.

The science of this "EM energy flow through space" is controversial to this day. Which is the real "EM energy flow vector" as such has never been adequately resolved, and there continue to be polite debates about it {6}. One should also be aware that physicists really do not know what many things — including energy — *are*. The definitions of these fundamentals are still uncertain, as stated rather poignantly by Feynman {7} in this quote:

---

[5] As we shall point out later, the concept of energy flowing through 3-space is a non sequitur and requires substantial revision today, to be consistent with the nature of observation and the fact that no observable continuously persists. Any observation is an instantaneous frozen 3-space "snapshot" at a single instant, gone the next instant when almost immediately replaced by another such frozen snapshot. In between observations, not mass but *masstime* exists. The same is true for 3-space, which only exists as the output of an observation process. Prior to observation, *spacetime* exists.

> *"It is important to realize that in physics today, we have no knowledge of what energy is."*

As another fundamental example, Feynman {8} also pointed out that we really do not know what *force* is[6] either! Quoting:

> "*One of the most important characteristics of force is that it has a material origin, and this is not just a definition. ... If you insist upon a precise definition of force, you will never get it!"*

1.1.2 Physics Is Not the Mathematics, But What It Manipulates

Following Feynman's ansatz, one should realize that the physics is not really in the mathematics itself, but in the physical meaning of the concepts that the mathematics manipulates. Here again, the inimitable Feynman {9} cautioned against over-attachment to the mathematics itself. He said it very clearly:

> *"Mathematicians or people who have very mathematical minds, are often led astray when "studying" physics because they lose sight of the physics. They say: "Look, these differential equations – the Maxwell equations – are all there is to electrodynamics it is admitted by the physicists that there is nothing which is not contained in the equations. The equations are complicated, but after all they are only mathematical equations and if I understand them mathematically inside out, I will*

---

[6] The problem is the ubiquitous and erroneous use of a "separate" force acting on a "separate" mass. Actually mass is a *component* of force, as can be seen from F ∑d/dt(mv). Hence there is no such thing as a separate force in empty space, acting upon a mass. Instead, the massless 4-field in space acts upon mass to produce force *in that interaction*. Force is not a cause, but an effect of an ongoing interaction — as is any observable. An observable is a continuing series of frozen 3-space LLL snapshot entities given by the result of continual application of ∈/∈t (LLLT) ♥ LLL by photon emission. The formation of the LLLT (spacetime and masstime) in between observed m as LLL, is produced by photon absorption of the previous m = ∈/∈t (LLLT).

> *understand the physics inside out." Only it doesn't work that way. Mathematicians who study physics with that point of view – and there have been many of them – usually make little contribution to physics and, in fact, little to mathematics. They fail because the actual physical situations in the real world are so complicated that it is necessary to have a much broader understanding of the equations."*

Such matters and similar thoughts had bothered me even back in the early 1950s, when I could not find a single professor or dictionary of physics that *logically* defined a field or a potential. Considered rigorously, the definitions all fell apart and violated elementary logic. Improper, insufficient, or just plain *wrong* definitions in physics have continued and are still widespread to this day. To give a single example from an excellent book by Kraus {10}: On p. 60, Kraus gives the formula for a potential referred to as the "absolute potential" of a charge source. Quoting:

> *"This potential... is, by definition, the work per coulomb required to bring a positive test charge from infinity to the point* r_1*."*

Kraus erroneously "defines" the scalar potential identically as work, which is equating the cause with the effect. The work that dissipating a potential does or can do, is not the potential itself! Else "human" means nothing but how well one can chew one's food or drive an automobile. Actually, Kraus gives one theoretical way to measure or calculate the effect of the potential's *local intensity at a point.*[7] Note that what is measured is the energy *diverted from* the potential at that point, around that test charge. This does not specify the *entity* (the potential itself) at all, but only what has been diverted from it. It is rather like confusing the whirlpool (water diverged from the normal river flow) in a river as being the river itself.

---

[7] To show the non sequitur, a million more positive test point charges could be brought from infinity to that same point, and the potential — without any change in it — would cause the same amount of work to be done upon each of those charges. As can be seen, even noted professors can speak rather casually when they "define" the potential's point intensity as "the" potential (which extends over all space, not just at that one point). There may be different intensities at different points in the potential, but the potential itself remains one-and-the-same thing regardless of which one or all of the point intensities are discussed. And neither the point intensity of the potential nor the potential is the work that is done upon a moving charge by that potential or its intensity.

Integrating what has been diverted from it does not yield the potential itself! At best, the measurement gives an indication of the *intensity* of the potential at a point, insofar as its reaction with charge is concerned. More rigorously, what is being utilized is the potential's reaction cross section presented to a *unit point static charge* at that point. The same unit point charge, if placed in particle resonance, will sweep out more geometrical area and exhibit a greater reaction cross section. That will increase the energy collected divergence of the energy flow that is being moved around the particle itself. By normal calculations, the resonant charge may collect some 18 times as much energy as is possible at the same point by the same charge in static mode — e.g., as shown by Bohren {24, 25}.

A definition must present an identity. Examining the proposed definition as "the potential is identically … work" one sees the problem immediately. *The potential exists whether or not there is a positive test charge, or whether one moves such a charge in from infinity or not, and whether there is any work done or not.* None of that is what a potential *identically is*, but only one aspect of what it *does* or *can do* or *can cause.* One leaves as an exercise for the reader the task of further examining dictionaries of physics and textbooks, to try to find a satisfactory definition of that common scalar potential.[8] We do not believe the reader will find it.

Yet any good textbook will also contain some real gems of great insight, simply said. As an example from Serway {11}, we eventually took a most marvelous cue, of how to get around classical thermodynamics' prohibition against heat energy "running uphill" from hot to cold. In an insightful statement, Serway said {11}:

> *"The second law [of thermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or of creating order out of*

---

[8] The scalar potential identically is actually a harmonic set of phase conjugate longitudinal EM wavepairs, as shown by E. T. Whittaker in 1903 {85}. Even Whittaker, however, misinterpreted his wavepairs only after interaction with that ubiquitous unit point charge assumed at every point in space. He gave two *effects* of that interaction, not the *cause* (which exists prior to interaction) *and the effect* (which exists only after interaction). Reinterpreting to get at the causal wave, each wavepair is a matched set of two waves; an incoming EM longitudinal wave in the time domain prior to interaction with a charge, and — after the time-energy wave is absorbed by the charge, an emitted outgoing EM longitudinal EM wave in 3-space. The so-called "static" potential is not static at all, but is a dynamic, ongoing 4-space process. This follows the re-interpretation by the present author {12} of Whittaker's decomposition — a reinterpretation then found to be consistent with quantum field theory {19} and with broken symmetry of opposite charges in particle physics {73}.

> *disorder. It merely states that such a reversal of the natural flow requires an influx of energy…"*

This essentially states the *law of entropy*, where to reverse entropy (disorder), one must apply ordering (energy).[9] We also note that the original concept of entropy was as *dissipation of potential*. There are of course different ways to apply the energy, and it need not be by the operator himself. Chapters 4, 5, 6, 7, and 8 of the present book give some unusual ways. A special paper {12} published by the author in 2000 gives another. Indeed, when the flow of time is predominantly reversed, so is the "entropy" of a situation, since the "videotape is running backward", so to speak. In that case, the entire classical thermodynamics must be extended to essentially include its own opposite. The making of a small time-reversal zone where such things happen is as simple as involving a predominance of antiphoton interactions with the charged particles in that zone, rather than a predominance of photon interactions. What we are saying is that the notion of irreversibility in thermodynamics is not necessarily absolute. It usually assumes a "time forward" situation, and may not hold in a "time-reversed" situation. In our chapter on cold fusion, we will present some specific and quite startling nuclear interactions that occur as a result of the time reversal of the coulomb barrier (repulsion of like charges in a forward time situation) into a coulomb attraction between like charges in a time-reversal zone and time-reversed situation.

---

[9] Now notice what Serway's statement means with respect to the classical thermodynamics "definition" of closed system. Thermodynamics defines a *closed system* as one in which mass is not exchanged across its boundary, but energy can be. That definition permits a closed system to receive excess energy from its active environment, and thereby reduce its entropy. A specific example is simply potentializing a circuit, prior to movement of the current. Hence the entropy of a closed system does not necessarily increase, but may decrease or increase if the system is in disequilibrium (difference in energy received from its active environment and energy escaping back to the active environment from the system. For the same system, if energy exchanges across its boundary equally in both directions, the system is said to be in "equilibrium" with respect to external energy exchange. One must be very careful in interpretation of the second law of thermodynamics! One has a very different "closed thermodynamic system" when it is in *energy exchange disequilibrium*, than when it is in *energy exchange equilibrium*. Equilibrium is the condition of maximum entropy. For substantial disequilibrium condition, entropy cannot be computed, but is less than the entropy of the same system in equilibrium. The entire ansatz of thermodynamics may be violated once time-energy is transduced into internal EM energy inside the system. That is a fundamental disequilibrium, performed by every charge in the universe. Hence of necessity we have advanced and utilized new definitions of "open system" and "closed system" in our approach in this book, as discussed in the Appendix.

If Serway's statement were reversed, it would then be a statement of the *law of negentropy*. It would state that self-ordering (i.e., freely receiving energy from the active environment) in a system could indeed "push heat uphill from a cold object to a hot one", and the system could simultaneously emit energy in the process. Every charge does it!

We did find it necessary to correct the classical thermodynamic definitions of "open system" and "closed system". To define a closed system as closed only to mass transfer, but open to energy transfer, is a gross non sequitur. Since general relativity was published in 1915, energy and mass are known to be the same thing (*mass is just a special form of energy*), hence the term "mass-energy" (mass as energy) in physics. Whenever energy crosses the boundary of a system, the system's mass changes and mass (or certainly mass change) has also crossed that boundary. Indeed, as we shall point out, in 1917 Hilbert specifically pointed out that in general relativity there can be no energy conservation equations of the kind usually employed elsewhere. The fact that general relativity falsifies much of the present foundations of classical thermodynamics seems to have been either ignored or missed by most of the scientific community, although we will quote leading Russian scientists who have noticed it and are aware of it.

The reader is thus warned that, henceforth, when we use "open system" we mean one where either energy or mass or both exchanges across the system boundary. When we use "closed system" we mean one in which *neither* energy nor mass exchanges across the boundary. In short, the notion of a "closed system" has been redefined into what classical thermodynamics calls an "isolated system". We already know from particle physics and the active vacuum (and from general relativity and the change of spacetime curvature with every change of spatial energy or mass-energy), and from the giant negentropy involving time energy transduced into 3-space energy and vice versa, that there is no such thing in all the universe as a truly closed system.

In Appendix A, we have discussed how extension and change to classical thermodynamics must be made. We do this by extending the first law, refuting any absoluteness of the second law and third law, dealing with the zeroth law in a new way, etc. We also urge the better theoreticians to re-examine classical thermodynamics along such lines, to modernize and upgrade it. We believe that the present scientific work to extend thermodynamics so it fits those situations now known to violate it should include additional considerations such as we present in Appendix A.

In short, with the new definitions Serway clearly states the difference between classical equilibrium thermodynamics, where no excess energy from the environment is received, and the thermodynamics of systems far from equilibrium with their active environment (using the new definition of open system for clarity), in which case excess net energy from the environment can be received and used in electromagnetic systems, providing COP>1.0 systems or even COP = ←systems. The windmill, sailboat, and waterwheel are age-old examples of disequilibrium systems where of course the energy of mass in motion is what is transduced. It is our objective in this book to point out the use of electromagnetic systems in energetic disequilibrium with their active environment (the active vacuum and curved spacetime) to provide just such negentropy.

1.1.3 Time Is Energy and Must Be Considered As Such

In a physics model, one's choice of fundamental units is arbitrary.[10] As an example, in one type of physics only a single fundamental unit — length — is employed. All other entities then become functions of length.

We are also free to choose the *joule* as the single fundamental unit in our physics model. The result that mass is a function of energy is now familiar and quite accepted, by the famous formula $E = mc^2$. Indeed, as can be seen, in that equation mass and energy are one and the same thing, since c^2 is a dimensionless constant. However, in our new model time becomes a function of energy similarly, and thus time is identically energy.

Let us perform a thought experiment. Suppose we take some spatial EM energy in 3-space, and compress it by the factor c^2. What can we do with it? If we leave it there in 3-space, it is known as "mass". If we place it on the fourth Minkowski axis ict, it is known as time because t is the only variable on that axis, and the only "place we can set it".

So to first order, *time has the same energy density as mass*. Multiplying an amount of time t (in seconds) by c^2 gives the decompressed spatial energy E that the time t will transduce into. In short, $E = tc^2$ also.

Now we notice what special relativity has to say about the relationship between time-energy and mass-energy. When the mass-energy increases

---

[10] E.g., see Jackson, *Classical Electrodynamics*, 2nd Edition, Wiley, 1975, p. 811-812; *ibid.*, 3rd Edition, 1998, p. 775. Jackson wryly remarks on how much excess heat and passion have been unnecessarily expended on that subject.

(e.g., as a function of velocity), time "dilates" or “decreases”. In other words (hint), there is a sort of special new extension to the conservation of energy law: If mass (3-space) gains some 3-spatial energy, then time loses some time-energy. Since time is spatial EM energy compacted by c^2, then *the relativistic energy changes in the time domain are enormously greater than the corresponding relativistic changes in spatial energy in the photon*. To the external 3-space observer, this is the injection of negentropy with respect to the present 3-spatial conservation of energy law. In the reverse case (as when the object reduces its velocity), if mass (3-space) loses some energy, then time gains some time because the time dilation reduces — i.e., time contracts or densifies (less time has more energy). To the observer, this is the production of entropy with regard to 3-space energy conservation. Yet the observer misses the fact that, accompanying entropy in 3-space is negentropy in time energy on the 4th Minkowski axis, and vice versa.

We believe this to be a rather dramatic extension to the previous concept of entropy in thermodynamics. We get 3-space entropy in physical processes only by gaining 4th dimensional negentropy accordingly. One can thus appreciate the impact upon the entire subject of thermodynamics, once time-energy is accounted and the new interplay of entropy and negentropy are accounted. We believe that this mechanism may involve the fundamental mechanism for both entropy and negentropy. Neither is produced alone; both are always produced in tandem. This of course is not what we “observe” since all observation is 3-spatial. Hence to observe the system and see 3-space entropy or negentropy is to hide the simultaneous 4th dimensional negentropy or entropy that unobservably accompanied the observable that we did see.

Again we call attention to the original meaning of entropy: the dissipation of potential (i.e., potential energy).

Once we understand that *time* is a special form of energy (we discuss this shortly), we may input the energy required to "move heat energy backwards" — i.e., we may directly engineer negentropy — by transducing some time-energy into 3-spatial energy. We can do it easily by time reversal, and every charge in the universe does it already. On the other hand, we can do it by breaking the symmetry of time-energy flow — which is as simple as forming a little dipole. The broken symmetry of unlike charges — and therefore the dipole — was discovered by Lee and Yang {13a-13c}, who strongly predicted it in 1956. Wu *et al.* {14} confirmed it experimentally in early 1957. This was such a revolutionary change to physics that Lee and Yang were awarded the Nobel Prize in the same year, 1957 — an almost unprecedented action.

Its broken symmetry tells us that the dipole's unlike charges continuously absorb virtual photon energy from the active vacuum, transduce it into real observable energy, and pour out real, observable EM energy in all directions in 3-space. That puts an entirely different perspective on *what really powers every dipolar EM circuit; i.e., what produces the flow of energy pouring from the terminals of every generator and battery, filling all space around the external circuit and its conductors*. The EM energy pouring from the generator or the battery is not due to the generator outputting some of its own available internal energy (from the shaft energy input to the generator, transduced into magnetic field energy inside the generator, and then dissipated to separate the charges and form the source dipole between the generator terminals) or the battery transducing and outputting some of its available chemical energy (which is only dissipated inside the battery to separate the internal charges and form the source dipole between the battery terminals). We will explain that later, and explain how any dipole or charge simply pours out real, observable energy continuously in all directions, without any *observable* input of energy. The input energy is there, but it is in *unobservable* (virtual) form.

In 1971 while a graduate student at the Georgia Institute of Technology, I realized the mechanism that generates the "passage of time" insofar as the observer is concerned {15}. With a little more work, this gave the clue in the 1990s for the mechanism generating little momentary *time-reversal zones* (TRZs) {16} in the electrolyte utilized in cold fusion experiments. Hence we proposed the use of little time-reversal zones forming momentarily in the electrolyte after loading of the palladium lattice of the electrodes with hydrogen or deuterium. Such time-reversal zones can form in the region of excessive positive charge accumulation, since positive charge can be said to receive negative EM energy from the time domain and output negative EM energy in 3-space.

This led to uncovering an entire class of new nuclear reactions — fusion reactions at low spatial energy but high temporal energy — in these little temporary time-reversal zones (TRZs), in which like charges attract and unlike charges repel {17}, followed by rapid decay of the TRZs where the normal law of attraction and repulsion is restored again.

In theory, a fermion cannot be time-reversed because the Pauli exclusion principle prohibits it. However, a boson can be time-reversed, so fermions can be time-reversed in pairs, where each pair acts as a *quasi*-boson. The

TRZ completely overcomes and *reverses* the "coulomb barrier" between a pair of like fermion charges.[11]

As an example, two deuterium ions in a little TRZ would momentarily form a quasi-nucleus of helium, since the two positive deuterons are attracted so closely that each enters the strong force region of the other. At the same time, the strong force is weakened by the partial time reversal of the gluon forces, so that the quarks are nearly freed. As the other ions in the surrounding solution then move to destroy the momentary time-reversal zone, their movement forces the TRZ to decay back to a time-forward zone — with the decay action starting from every point in spacetime inside the nucleons themselves[12]. The TRZ decay-induced action thus strikes the *nearly freed* quarks first, and decay can occur by easy quark flipping while the gluon forces are still much reduced and not back to normal strength. Hence as the gluon forces return toward normal, the quasi-nucleus decays by the strong force increasing and overpowering and drawing the quasi-nucleus into a full-blown helium ion — i.e., an alpha particle —flipping the partially-freed quarks as necessary to do so (that is not necessary in equation [1] below.

During decay of the TRZ, the weakened strong force grows much more rapidly than the Coulomb force zeroes and then increases. Consequently, the quasi-nucleus of two D+ ions merely draws together due to the rapidly increasing strong force, forming an alpha particle without quark flipping. Four H+ ions — four protons — in a quasi-nucleus in a TRZ will undergo quark flipping twice when the TRZ decays, thereby resulting in an alpha particle.). So that explained the anomalous formation of the alpha particles in the experiments. The interaction for two ions of deuterium is given by:

$$^{2}H_{1}^{+} + {}^{2}H_{1}^{+} \Downarrow {}^{4}He_{2}^{++} \qquad [1]$$

---

[11] In a TRZ, the law of attraction and repulsion of charged particles is reversed.

[12] Any moment in time exists everywhere simultaneously throughout the universe. In short, time is a multiple connection in 3-space. Hence in any time-reversal zone (a TRZ) where time is reversed, then starts fading away and back to a time-forward zone (TRZ), the changes induced by the "fading back to TFZ" *simultaneously* involve every point in the 3-space of that TRZ that is changing (decaying). An easy change is quark flipping, since the quarks are almost freed in the TRZ to begin with. The reason the reaction proceeds in that direction is that the recovery of the strong force is much faster than the restoring of normal electrical repulsion, hence the quasi-nucleus is drawn further together into a full nucleus, constituting a legitimate nuclear transmutation at low spatial energy but high time-energy.

This interaction between two deuterons in solution does not occur in a normal time-forward zone because of the coulomb barrier preventing the two deuterons from entering each other's strong force region and acting as a quasi-nucleus of helium, so particle physicists have missed it. It does exist in a time-reversal zone; such zones form and then decay back into normal time-forward zones. This is possible because the coulomb barrier is momentarily reversed. The use of such TRZs in particle physics opens up thousands of new nuclear reactions, all at low *3-spatial* energy, but involving very high *time*-energy. Because of the extreme energy density of time, *these time-energy-induced reactions are actually much higher-energy reactions than high-energy physicists presently consider and utilize!* Indeed, it opens up a new kind of far more energetic "high energy physics".

Cold fusion experimenters have unwittingly opened a window upon a vast new particle physics, previously overlooked by our scientists because they have disregarded the use of time-energy, time reversal zones, and transmutation of time-energy into 3-spatial energy in their nuclear reactions. However, a few physical theorists attempting to better explain particle physics have recognized the importance of the time domain, and probing work in that respect is occasionally done {18}. In our view, it is not accidental that more than 600 successful cold fusion experiments have been obtained, by a variety of researchers in many labs in several nations. We strongly suggest that much of the conventional physics community has firmly placed its collective head in the sand, and is refusing to grapple with the startling new time-energy physics that is being initiated by cold fusion research.

In 1999 (published in 2000) {12} we finally discovered a great new symmetry in EM energy flow, whereby time-energy flow symmetry and 3-space energy flow symmetry are each individually broken, while an overlooked and more fundamental 4-symmetry energy flow — between the time-domain and 3-space — is sustained. *The result is that all EM energy in 3-space comes from the time domain locally and returns to the time domain locally, in a giant negentropic circulation.*[13] Together by the discovery of relevant quantum field theory work by Mandl and Shaw {19}, this now lends strong support to the use of time-energy in physics as a practical matter for strenuous investigation.

---

[13] This is also understandable from the implications of the observation process, which yields a frozen 3-space snapshot existing only at a single instant. Thus any 3-space energy existing there in that frozen instant, had to just come from 4-space (from the time-domain via the giant negentropy process, if we take the view that "the past exists only in time itself".). For the frozen snapshot to "change", time must be added to it, which converts it to a 4-space process again. So the "3-spatial" observed

As an example, Mandl and Shaw {19} treat the four polarizations of the photon. Neither the longitudinal nor the scalar photon is directly observable, but in the presence of charge the two are observable in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. This argument, translated from particle terminology to wave terminology, directly fits our re-interpretation {12} of Whittaker's 1903 decomposition of the scalar potential {85}. For the combining mechanism of the fields of the photons, we must account for *the field as a ubiquitously assumed interaction with the detecting/observing unit point charge*. Thus we must account for the absorption of the incoming time-polarized wave or photon, the transduction of that excitation energy of the charge into longitudinal EM wave/photon energy, and subsequent emission of that excitation energy in 3-space. That is what happens for a negative charge. For the positive charge, the process is time-reversed, hence occurs in opposite fashion. Or as an alternative, the positive charge can be said to continuously receive negative time-energy from the time domain and emit negative spatial EM energy.[14]

There is an *energy polarization transduction function* of charge, whereby it transforms received time-polarized photon energy into emitted longitudinal photon energy in 3-space (for the negative charge, and vice versa, for a time-reversed positive charge). This transduction appears to have been overlooked in physics prior to our recognition of it. It can in fact be used to generate an acceptable definition of charge itself. Charge is the

---

energy must come from an immediately previous 4-space process, and must return to an immediately following 4-space process.

[14] As we will find in our chapter on antigravity, there is good reason to treat the positive charge as a source of negative energy and negative energy fields. However, this is pertinent only prior to observation of the charge, while it is still a 4-spatial unobserved negative energy electron entity. For the observable charge, one has already conjugated — after all, the positron is observed as if it were an electron going backwards in time, which we observe as an electron with its charge reversed and with parity reversal (of its spatial direction). For the observable charge, we have already reversed the negative energy fields into positive energy fields by simply reversing their direction and the time associated with the photon (quantum of the EM field).

continuously active entity which performs that ongoing process or those ongoing processes of energy transduction between the time domain and 3-space (between the causal unobserved 4-space process and the observed 3space snapshot). Or in other words, it is an active process connecting 4-space cause and 3-space effect, and connecting the unobserved (such as virtual) to the observed (such as mass).

So for a dipole, the "causal" time-polarized EM wave or photon as a 4-space entity comes to the dipole[15] (3-spatial as observed) and is absorbed by the detecting negative charge or dipole, then is re-emitted as the longitudinally polarized EM wave or photon in 3-space. *That absorption and remission is what charge does, since it is an entity for performing that process*. The emitted energy in turn is absorbed by the nearby positive charge, retransduced into time-energy, and re-emitted back to the time domain. This ongoing very special 4-space energy circulation (even with a virtual charge in the vacuum) is what a scalar potential *identically is and is doing*, at every spatial point of itself, inducing vacuum polarization and “point dipoles” in the virtual state in the process. Recognition of these missing functions allowed at last a solution to the long-vexing problem of the source charge and its associated fields and their energy, often called the most difficult problem in both quantum and classical electrodynamics {68}. We discuss that solution later.

1.1.4 The Search for COP>1.0 Circuits and Systems

A very long search and much intense study and reflection eventually revealed the concepts and principles of those long-neglected disequilibrium Maxwellian systems that permissibly output more energy than the operator inputs. The active environment — not the operator — simply inputs the rest of the energy. Such disequilibrium systems are indeed permitted in Maxwell's theory {20}, and are also still prescribed by Heaviside's severe curtailment of it {21} into what is two vector equations with variables unseparated, rather than Maxwell's 20 equations in 20 variables.

---

[15] Prior to interaction of the incoming time-energy with the observable charge or dipole, it interacts with the virtual particles of the vacuum, generating vacuum polarization. Neither the virtual particles nor the time-energy are observable; only the effects of their conglomerate interactions with observable charges are observed. Hence one can take the particle view that virtual particle energy is continually absorbed, or one can also take a quantum field theory view that time-energy is continually absorbed. The two are always present in combination.

When Lorentz[16] symmetrically regauged the Maxwell-Heaviside equations, he arbitrarily discarded the entire class of Maxwellian systems that are far from equilibrium in their exchange with their active (vacuum) environment. Lorentz revised (symmetrically regauged) the Maxwell-Heaviside equations to make them amenable to separation of variables and closed analytical solutions, thus reducing the onerous chore of numerical methods. This Lorentz symmetrical regauging is given in most EM textbooks {22}, and we show it in Chapter 2. The symmetrically regauged Lorentz equations are not Maxwell's equations, nor are they the truncation of Maxwell's theory by Heaviside *et al*. Considering an active environment, under our altered thermodynamics definitions Lorentz implicitly selected and retained only the equilibrium class of Maxwellian systems, while *arbitrarily discarding* the entire disequilibrium class. He thus discarded all those Maxwellian systems permitted to produce COP lying in the range 1.0 < COP Ω←.

Maxwell's electrodynamics is a *material fluid flow theory* and it assumes a material ether. Anything that fluid systems can do, electrodynamics systems can do, at least in theory, because their mathematical models are the same form. So when one cites known examples of fluid-driven physical systems where the energy to run the system is freely furnished by the active environment, analogous electrodynamic systems in active environments — and in disequilibrium exchange with that environment — must also exist in nature. Indeed, particle physics requires it and proves it. These are the very systems arbitrarily discarded by Lorentz symmetrical regauging in every university.

So there exists a direct analogy between fluid systems and classical electromagnetic systems. The common windmill, waterwheel, and sailboat demonstrate by analogy that open EM systems far from equilibrium — powered by free EM "winds" and "energy flows" in the active vacuum environment — also exist in consonance with natural law. They are no more mysterious than a solar cell power system, which is after all a recognized "free energy" or "overunity" system, as is the humble charge (thought to be the source of all EM energy, fields, and potentials). In physics, the powering of systems by receipt and use of energy from their active environment leads directly to the thermodynamics of systems far from thermodynamic equilibrium in their energetic exchange with that active environment.

---

[16] Actually first accomplished by Ludwig Valentin Lorenz in 1867, then by H. A. Lorentz much later. Lorentz was given the credit erroneously. Lorenz actually derived electromagnetic theory independently in his paper. See J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys*, Vol. 73, July 2001, p. 663-680.

It follows that the seeming absence of such *electrodynamic* systems arises not because they are impossible but because present-day circuits and systems are ubiquitously designed to self-enforce an inherent energy equilibrium with their active vacuum environment. The closed-current loop circuit turned out to be the Lorentz self-regauging demon involved in destroying the COP>1.0 capabilities of every EM circuit. So little by little, we unraveled the long tedious trail of Maxwell's electrodynamics and what had happened to those missing Maxwellian-Heaviside systems far from equilibrium with the active vacuum.

We learned how, where, and by whom those permissible overunity Maxwellian systems were discarded. That is, we found what happened to all those Maxwellian *open disequilibrium systems* — originally included in Maxwell's and Heaviside's theories — where such a system receiving and using excess energy from its active environment[17] is permitted by the laws of physics, electrodynamics, and thermodynamics to:

---

[17] For the discerning reader, of necessity we have revised the foundations of the ancient classical thermodynamics, as further discussed in Appendix A. We refer the reader to Bimalendu N. Roy, *Fundamentals of Classical and Statistical Thermodynamics*, Wiley, New York, 2002, and to any good book on the history of thermodynamics. Also particularly good is Kondepudi and Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures*, Wiley, Chichester, 1998, reprinted with corrections in 1999. The foundations of classical thermodynamics (and its fundamental definitions) were basically frozen prior to the advent of Maxwell's 1865 seminal theory, and well before the 1880s discovery (modeling) of the flow of EM force field energy through space. Some of the fundamental definitions of thermodynamics now will not withstand critical review in terms of "meshing" with the Heaviside/Poynting material fluid energy flow theory. Neither will they withstand the new concepts of energy such as mass-energy, time-energy and transduction between time-energy and spatial energy by every charge. E.g., thermodynamics defines an *open system* as one that exchanges energy and mass with its surroundings. Yet it defines a *closed system* as one closed only to mass exchange, not to energy exchange. *If the energy exchange is analogous to material energy flow and changes the mass of the system, then that definition of closed system is a non sequitur.* From general relativity, mass is simply energy anyway, and so "mass" exchanging across the boundary of the system is actually energy exchanging, and vice versa. Since Maxwell's theory is a *material fluid* theory, the Poynting and Heaviside energy flow models are *material fluid flow energy* models by analogy. The specialized thermodynamics definition of *closed system* rigorously will not logically allow the exchange of "material fluid energy flow" (or energy as matter)

since it prohibits matter flow, but the specialized definition of *open system* would and does. In short, with respect to material fluid energy flow, the concept of the "closed system" in thermodynamics has forced itself to become the *isolated system* instead, which is unacceptable since we must model EM energy flow exchange between the environment and the system. So we must change the thermodynamic definitions of *open system* and *closed system*. Else there cannot be any EM energy-mass or mass-energy flow between environment and system, which totally violates what is well known to be happening in all EM systems. In modern physics, every charge and every dipole already have such energy exchange with the active vacuum environment, and it is never zero; instead, it is of enormously high magnitude. Without that exchange, as we advance in this book, there cannot even be a "source charge" or associated EM fields and potentials and their energy, reaching across space. In other words, without it we can have no electrodynamics at all. So we have uncovered a fundamental and major contradiction between classical thermodynamics and general relativity, as well as between thermodynamics and the "material EM fluid energy flow" model used in electrodynamics. What classical thermodynamics calls a "closed system" permitting energy flow exchange, we must now logically regard as an *open* system because it is open to energy exchange across the boundary! The previous notion of the *isolated* system — with no exchange of either energy or mass — is what we must treat as a truly *closed system*. And there is none such in the universe, as we know in particle physics (e.g., because of the discovery of broken symmetry in 1957 and because of the well known active vacuum exchange with every EM charge and dipole). So we have corrected these direct contradictions between "EM as a material fluid energy flow theory" and the old classical thermodynamics.

As the reader will later see, this strongly affects our new definitions of *efficiency* and *coefficient of performance*. The new definitions we advance are rigorous, and they also apply to COP>1.0 EM systems, and even to self-powering (COP = ↔) EM systems. They also hold for very novel new energy processes such as quantum potential energy in a multiply connected space, multiple retroreflections and re-use of the same energy, conversion between time energy and spatial energy, and 4-space giant negentropy energy flow circulation.

But one can no longer be allowed to equate *efficiency* with *coefficient of performance*. Now they are never the same thing, just as a six-foot tall man and a six-foot tall doorway are never the same thing, even though they have the same height magnitude. We warn the reader that a great deal of thought and study must be put into appreciating these suddenly encountered changes to the quite old classical thermodynamics definitions. The changes are absolutely necessary. Bluntly put, in light of much more modern knowledge, a rigorous foundations analysis reveals the classical thermodynamics as well as disequilibrium thermodynamics to contain logical contradictions, such as its direct contradiction with general relativity and the EM material fluid energy flow theory. Either we give up or dramatically change the present EM energy flow theory, or we must make the necessary foundations changes to thermodynamics. We have chosen the latter option in this book, and the reader is forewarned of that dramatic change. The specific changes and rationale are discussed in Appendix A.

(1) self-order,

(2) self-oscillate or self-rotate,

(3) output more energy than the operator inputs (the excess being freely received from the active environment),

(4) power itself and its load simultaneously (all the energy being freely received from the active environment), and

(5) exhibit negentropy.

We vigorously pursued those long-lost Maxwellian systems, and we eventually found them. We also found real experiments {23, 24, 25, 26, 27, 28} and real devices {29a, 30, 31, 32, 33, 34, 35a, 36a, 36b} that performed one or more of those fabulous five functions, though there was often no realization by the experimenters, inventors, and scientists of the actual mechanism involved. Eventually my colleagues and I were also able to produce a successful experimental device {37, 38a-38c, 39}, the motionless electromagnetic generator (MEG), which outputs more energy than we input to it. We cover the MEG in Chapter 7. Presently a cooperative research program is ongoing with the National Materials Science Laboratory of the National Academy of Sciences of a friendly foreign nation, to develop and market commercial power systems based on successful laboratory experiments with the MEG. At this writing, we are also strongly seeking the extensive funding required to set up a physics lab and complete the final research allowing production of power systems.

1.1.5 Additional Very Important Implications

We also formally proposed {40} that the vast nondiverged EM energy flow component — Heaviside's "dark" nondiverged energy flow component, accompanying every reaction of a charge with a field or a potential, but arbitrarily discarded by Lorentz and modern classical electrodynamicists — is the generatrix for the extra gravity holding the arms of the spiral galaxies together, after all the dark matter is accounted.[18]

---

[18] Heaviside himself recognized the gravitational implications of his extra component of energy flow, which is in closed circular loops. Beneath the floorboards of his little garret apartment, years after his death, handwritten papers were found where Heaviside used this component for a unified EM approach to gravitation. See E. R. Laithwaite, "Oliver Heaviside – establishment shaker," *Electrical Review*, 211(16), Nov. 12, 1982, p. 44-45. Laithwaite felt that Heaviside's postulation that a flux of gravitational energy combines with the (**EΔH**) electromagnetic energy flux, could shake the foundations of physics. Quoting from Laithwaite: *"Heaviside had*

In addition, my close colleague Bedini and I have filed a patent application upon a very special process to "freeze-frame and lock-in" a disequilibrium Maxwellian system in its otherwise far-from-equilibrium state, so the apparent disequilibrium operation of the system can be maintained stably as a new equilibrium condition. This appears to be a method to produce and utilize what Kondepudi and Prigogine refer to as a *nonequilibrium stationary state*. This stabilization is necessary in order for such a system to maintain its COP>1.0 excitation and steadily output more energy than the operator inputs, or to steadily power itself and its load {41}.

We also found that COP>>1.0 EM systems (and some COP>1.0 systems) produce (as a function of the COP) a current of Dirac sea holes (positrons) *in the local vacuum environment itself*, from the output section back to the input section. At COP<1.0 and COP not too greatly above 1.0, a Dirac sea hole in the vacuum almost immediately interacts with an orbital electron in the material lattice of the system. This converts the negative energy, negative mass "vacuum hole or state" into a lattice hole, which is attached to the large positive mass of the ion left by the disappearance of the electron. This "lattice positron" type of problem has been known in semiconductor design for some decades. Normal EM circuits do not usually meet the phenomenon overtly because the semiconductor designers controlled it in the semiconductors themselves by use of appropriate donors and acceptors.

There is a great difference between the actions of Dirac sea holes in the vacuum prior to observation, and lattice holes in materials (after observation). So there is a great difference between the action of a “positron” on spacetime before its interaction with mass and observation, and its action on spacetime after it interacts with mass and is observed.

For COP>>1.0, significant phenomenology and novel effects occur because a substantial fraction of the Dirac sea holes (unobserved positrons) sweeping from output to input do not convert to lattice holes (observed and bound positrons) along the way. For substantial values of COP above 1.0, Bedini's invention (patent application filed by Bedini and the present

---

originally written the energy flow as $\mathbf{S} = (\mathbf{E}\Delta\mathbf{H}) + \mathbf{G}$, *where* **G** *is a circuital flux. Poynting had only written* $\mathbf{S} = (\mathbf{E}\Delta\mathbf{H})$. *Taking p to be the density of matter and* **e** *the intensity of a gravitational force, Heaviside found that the circuital flux G can be expressed as* $p\mathbf{u}$ 4 $c\mathbf{e}$, *where* **u** *represents the velocity of p and c is a constant.”*

author)[19] covers the master process for intercepting and transducing this appreciable flow of negative energy from the output section back to the input. Otherwise, that flow — if not intercepted and not converted — will "eat" extra input electrons from the power supply, thus acting as a novel "extra load" appearing in the input section. That extra load then draws additional current and power from the external power supply by electron-hole annihilation[20].

With the Bedini invention, the negative energy (unobserved positron) flow appearing at the input section is transduced into a flow of positive energy (a flow of electron current) from the input section back into the system. The process deliberately uses the "interaction and observation" process to phase conjugate the charge and reverse its direction of flow! In that case, the otherwise detrimental *negative* energy output back through the COP>1.0 system (which is nature's decay process for COP>1.0 interactions) is changed to a beneficial *positive* energy input within the system itself, freely received from the vacuum environment. This process is then used to close-loop the system for self-powering in a "locked" and stabilized disequilibrium condition — a nonequilibrium stationary state.

If we consider mass to be a special kind of positive energy state, then positive energy states represent curvatures of spacetime that are positive gravity. Negative energy states generate antigravity (the time-reversal of gravity).

---

[19] Bedini personally discovered and implemented the solution before the exact nature of either the problem or the solution was recognized! My contribution was to recognize the nature of the problem and the mechanism used in the solution. We also stress that, contrary to conventional treatment, all EM circuits do involve not only lattice holes in the conductors and components, but also Dirac sea holes in the local vacuum. That the Dirac sea is involved whenever there are EM fields is clearly shown by Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," *Adv. Theor. Math. Physics*, Vol. 2, 1998, p. 963-985.

[20] With a proper change in the curvature of local spacetime, pair annihilation can occur with no accompanying photon radiation. The condition is that the part of the curvature of spacetime representing the energy change of the otherwise emitted radiation, does not "relax" even though it is an "excited state". Rigorously, the emission of the radiation from pair annihilation occurs in two steps: (i) first the local spacetime is curved for and by the energy excitation, as a *static change* of the curvature, and (ii) then that curvature relaxes back to its former value, propagating that specific curvature in space, which is recognized as the radiation propagating in space. If the spacetime curvature excitation does not relax, there is no photon emission and propagation.

Before their interaction and observation, the Dirac sea hole (positron) currents — produced in natural COP>1.0 processes in the universe — are still negative energy electrons in 4-space. They are not 3-positrons until interaction with matter has occurred. These "negative energy electrons" generate negative energy EM fields, including both the Poynting energy flow component and the Heaviside nondiverged energy flow component. These negative energy EM fields appear to be *generating the antigravity* whose effects are seen in the distant parts of the universe. They produce far more powerful effects than the accounted electrogravitation effects in astrophysics, which only uses the gravitational effects of the very much smaller Poynting energy flow component. Also, astrophysicists do not use the "positron before observation," when it is still a negative energy electron with negative mass. The accumulation of the negative energy (Dirac holes) in space (in the surrounding vacuum) is connected with massive objects and violent processes, where very strong gradients are produced.[21] As a result, very large negative energy fields and potentials are produced. In turn, this results in cumulating antigravity. This cumulating and interacting antigravity appears to be the mechanism for the forces *accelerating* the expansion of the universe — rather than it decelerating as would be the case if the net gravity there were positive. In Chapter 8 we propose this explanation for the observed acceleration of the expanding universe — and the basis for the explanation can be and has been successfully demonstrated in a legitimate overunity EM circuit or system {42}. We offer this in honor of Heaviside, who first discovered the gravitational aspect of his huge nondiverged energy flow, but did not live to publish it. He also did not consider the Dirac sea prior to hole interaction with matter, as it was not yet formulated, so he had not yet recognized the way to produce and utilize the practical antigravity potential of his discovery.

1.1.6 A Scientific Dilemma

There would seem to be a sufficient abundance of techniques, devices, processes, and theoretical works to impel a crash project in the scientific community to develop successful overunity electrical power systems {43a-43d}. This would be especially appropriate at this time, since the escalating energy crisis now threatens to slow the world economy, and then eventually collapse it. At this writing, the MidEast has heated considerably, and appears headed for another MidEast war, with resulting severe interruptions of cheap oil supplies from the MidEast. Nonetheless, in spite of Russia having opened its oil field spigots to try to fill the need, a crisis in oil appears to be looming.

---

[21] Kondepudi and Prigogine, *ibid.*, p. 459 already point out that strong gradients produce situations that violate present thermodynamics. Research in these and other situations violating present thermodynamics is going forward under the caption of "extended thermodynamics" research. A discussion of the area is given by D. Jou, *Extended Irreversible Thermodynamics*, Springer-Verlag, New York, 1996.

So *why* does the U.S. scientific community still so adamantly oppose the very notion of Maxwellian systems freely extracting EM energy from the vacuum? Why is there no outcry pointing out what the hoary old Lorentz regauging really means in terms of equilibrium or disequilibrium of the regauged system with the active vacuum? The unequivocal participation of the vacuum in a continuous energy exchange with the charges and dipoles of every EM system has long been affirmed by particle physics. All the fields and potentials — and their energy — manipulated in any EM circuit come from the vacuum, as proven in particle physics for 45 years. Why have our electrical scientists not understood — from the broken symmetry of the vacuum exchange with the opposite charges of the source dipole between the charged terminals of every generator and battery — that *EM energy from the vacuum powers, and has always powered, every electrical power system and circuit ever built?* Why have the later rigorous broken symmetry findings of particle physics not been incorporated to update the ancient classical EM theory used to design and build electrical power systems, nearly a half-century after those broken symmetries were discovered and proven? Why do our classical electrodynamicists continue to assume that every charge and dipole in the universe is a perpetual motion machine, freely creating energy out of nothing and pouring it out across all space at the speed of light?

Therein lies one of the real problems of present science — its historical and continuing resistance to "out-of-the-box" thinking[22] and to research that overcomes conventional strictures. More than 40 years after the basis for the vacuum-energy powering of every dipolar system (and of every observable charge once its clustering virtual charges of opposite sign are accounted) was discovered and proven in particle physics, all our university electrical engineering departments continue to erroneously teach that the shaft energy input to the generator powers its external circuit. Internal EM energy in the generator, transduced from the input shaft energy, does not directly add a single watt to the external circuit. Instead, it only continually forces the generator's internal charges apart, to continually remake the source dipole, which then extracts energy from the vacuum and pours it out of the terminals, filling the space surrounding the conductors of the external system for the system to intercept and utilize some of the available external energy flow.

---

[22] "Out-of-the-box" thinking is a widely used concept among planners and program formulators, when conventional thinking will not suffice to solve a major problem with which they are struggling. Conventional thinking is considered "in the box" thinking, so by demanding out-of-the-box thinking, a problem demands an unconventional solution outside those normally proposed. In short, some new thoughts and concepts are required. Much lip service is given to the concept as a favored buzzword, but few proposed programs with truly "out of the box" approaches will be funded. In the energy field, none at all are funded that are truly "out of the box", whether or not that or similar phraseology is used.

Once the dipole is established, it will extract and transduce EM energy from the vacuum and pour it out in all directions at the speed of light, without ceasing. Else, the Nobel Committee should admit its grave error and revoke the Nobel Prize awarded to Lee and Yang. Real observable EM energy extracted and transduced from the vacuum's virtual energy is precisely what the "broken symmetry of the opposite charges" on the ends of the dipole *means*. Indeed, all the forces of nature are already considered as generated by the interaction of virtual particles with observed particles. Since force produces energy changes in the system affected, then it follows directly that energy changes are produced by the interaction of virtual particles with observable particles.

However, our power system engineers ubiquitously use the closed current loop circuit. This inane circuit self-enforces the Lorentz symmetrical regauging condition. It uses half the energy captured by the external circuit (from intercepting some of that "energy flowing around the circuit from the seething vacuum" that is copiously pouring out of the generator terminals) to ram the spent electrons back through the back emf of the source dipole itself. That scatters the dipole charges and kills the dipole and its broken symmetry — and also kills the free flow of transduced EM energy from the vacuum.

That insane circuit kills the source dipole in the generator faster than it powers its external loads! So one has to keep rotating the shaft of the generator, to keep producing a magnetic field inside the generator, so that this magnetic field energy can continue to force the charges back apart and continually reform the source dipole.

In short, our engineers build the equivalent of an electrical windmill, then — so to speak — force it into a closed barn so the environmental wind cannot reach it any longer. It is little wonder that we ourselves then have to do work on that "electrical windmill" to crank it around!

We pay the power company to engage in a giant wrestling match inside its own generator and *lose*. We also wildly pollute the planet with hydrocarbon and nuclear wastes, poison species (including killing of X numbers of humans every year via the disruption of their body functions and health), and enhance global warming. We are slowly strangling our biosphere and ourselves. None of that is necessary.

1.2 Organized Science Often Resists Innovative Change

The history of science is littered with examples where the scientific community has ignored the principle behind Einstein's statement quoted at the beginning of this chapter. As pointed out by Smith {44}, science has become reoriented toward profit. Quoting:

> *"...science is not the danger; scientists encouraged to do bad science to survive are."* ... *"...changing the way modern science is funded is an enormous undertaking, but it is a necessary one if we want to protect our future. Call it managed risk."*

Science's resistance to change is so well known to historians of science that it is rather universally accepted — although usually not made explicit to undergraduates. Further, scientists are under great pressure to conform:

1. After all, science is *patronized*; someone must fund the laboratories, the research supplies, the salaries and personnel benefits, etc.
2. To procure and protect its patronage, science has become quite organized, particularly with respect to how the funds — received and channeled down from the top — are cut into individual packages (research grants and research programs) and made available for competition among the "performing" chain of universities and research laboratories.
3. In the last few decades, there has risen an increasingly fierce demand by universities that the scientific researchers (i) be successful in attracting outside funds and (ii) file patents assigned to the university. So fierce has this demand become, that the research professor's continued livelihood may literally depend upon his or her success in bringing in extra funding. Further, much of his time is now spent in writing proposals to

compete for these "packaged funds". When he wins them, the kind of research and the areas of research are already stringently defined, and he dare not deviate — else there will shortly be no more funding packages won, no funding for his graduate students, and soon thereafter there will be no job of any importance for the professor!

4. In the United States, all the government National Laboratories and our universities are intensely seeking and filing patents! So small, independent inventors cannot deal with these organizations, without risking and almost guaranteeing the loss of their patents and intellectual property. Funding dangled in front of the inventor, much like a carrot dangled in front of a horse, is often tied to "march in" rights {45, 46} calling for surrender of the inventor's patent rights altogether, *whenever the government — i.e., a single bureaucrat — wants to take it*. All that has to be done is to declare that the inventor is not getting it developed and to market fast enough. Science has thus become more avaricious and — some inventors would even say — it increasingly involves overt and covert piracy of intellectual property rights. From personal experience, reluctantly I would not argue with that statement. Simply ask Larry Fullerton of Time Domain Corporation about his struggle with a National Laboratory over patent rights to his ultrawideband communications invention and technology.[23] It eventually resulted in a "draw" of sorts. He did not lose his rights, but the government gained them also, in competition with him. The government *circumvented* his patent, even though they did not succeed in taking it. We were delighted to recently see that Larry (the company is Time Domain, Inc.) received its 74th patent in this technology, as well as a ruling that will allow the technology to at last go to market.

5. The result is a dramatic increase in the pressure on working scientists and independent inventors to conform, and to "play the game by the rules". Then everybody up the scientific food chain is fed, and is happy and secure. The journals happily publish the research papers and results, the professor gets

---

[23] Stephen Fenichell, "Radio Flyer," *Discover*, 22(5), May 2001. Fullerton's technology has been given a limited go-ahead by the FCC, which has drafted standards and regulations in the area as of February 2002.

funding for his graduate students, the university gets that wonderful overhead cut of the research funding — such as half or it or more — and the entire apparatus is like a very large and tidy Titanic adhering to its ponderous course toward the iceberg. Meanwhile, truly new and innovative science discoveries — vulnerable and desperately needing nurturing funding during their initial embryonic state — get shoved aside, crushed, and starved in the funding rush to adhere to performance of the *prescribed* funding packages.

6. In this environment, the day of the "defenders of the scientific faith" has arrived! A small percentage of conventional scientists who are dogmatic and vociferous, are now very prominently attacking any novel experiments and ideas with a vehemence seldom seen in organized science. It is again reminiscent of some of the noted scientific attacks in history, e.g., as pointed out by Hellman {47}. Yet, because of the financial pressure upon the scientific community, and the increased pressures to conform, there is little restraint of the dogmatists and they are almost never called to task. Cold fusion is a current example. The American Physical Society has recently issued a statement condemning perpetual motion machines — yet the society's members continue to condone and use a classical EM model that assumes every charge in the universe to continuously be creating energy from nothing. Even the American Physical Society has not recognized what broken symmetry of opposite charges means with respect to the common dipole and dipolarity in every circuit. Nuclear reactions at low spatial energy (which means at extraordinarily high total energy when the c^2*-compressed* time energy is considered) do indeed sometimes occur in carefully controlled experiments, whether or not we yet sufficiently understand the reactions theoretically, and regardless of whether we can get the anomalous results to happen every time. Yet this area of nuclear interactions at low *spatial* energy {48} — and unknown to the scientific community, at very high *time* energy {49, 50, 51} — has been savaged by these self-appointed spokespersons for the "official" community, none of whom even account the compressed energy in the time increment portion of the photon. More than six hundred successful experiments in multiple laboratories, by respected scientists in multiple nations worldwide, are now

> rather resoundingly ignored. Yet the replication rate for good cold fusion experiments is certainly higher than many of the replication rates for novel and little-understood phenomena in large particle accelerators, and the cold fusion experiments are also *far cheaper*.[24] While particle accelerators are "popular" in their ability to garner huge funding, their cost/benefit ratio compared with, say, cold fusion experiments, may be abysmally low. Simply examine the decades of effort and many billions of dollars expended on the search for *warm* fusion (using spatial energy only). What has it produced, in terms of watts of power on the power lines? When will it produce any electrical power of any significance? Prototype cold fusion power systems have in fact been produced and patented {52}. With seed money from the scientific community and using a higher symmetry electrodynamics, cold fusion power could proceed at a rapid pace.

Strangely, the ever-present pressure to conform to *that which is already known and accepted* has often made science its own worst enemy throughout its history. Establishment scientists and the "system itself" now probably block — and have blocked over the decades — more innovative scientific research than does any other factor {53}.

1.2.1 Many Scientists and Historians Have Pointed It Out

The scientific community is well known to have always been highly resistant to novel ideas and innovations. Here are some selected pertinent comments regarding this phenomenon, where organized science is itself the obstacle to the advance of science, and where such has been recognized for many decades:

---

[24] As an example, see R. P. Taleyarkhan *et al.*, "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science*, Vol. 295, 8 Mar. 2002, p. 1868-1873; Charles Seife, "'Bubble Fusion' Paper Generates a Tempest in a Beaker," *ibid.*, p. 1808-1809. See also Donald Kennedy, "To Publish or Not to Publish," *ibid.,* p. 1793. *Science* had the courage to publish the peer-reviewed results of a tabletop sonoluminescence experiment that apparently produces nuclear reactions. Editor Kennedy essentially advises all protagonists on both sides to cut the rhetoric and allow the scientific community to do its replication work, to either validate or refute the successful experiments of Taleyarkhan *et al*. This action by *Science* is a shining beacon to remind the scientific community that science is based on experimental method, and that prevailing theories cannot refute new experiments that contradict them. Instead, laboratory bench experiments must decide such an issue.

"Every great scientific truth goes through three stages. First, people say it conflicts with the Bible. Next they say it had been discovered before. Lastly they say they have always believed it." [Louis Agassiz, 1807-1873.]

"There are three steps in the history of a great discovery. First, its opponents say that the discoverer is crazy; later that he is sane but that his discovery is of no real importance; and last, that the discovery is important but everybody has known it right along." [Sigmund Freud].

"Anybody who has studied the history of science or worked as a scientist knows that whenever something novel is discovered or proposed, there is a polarization of scientists, with hostility and bitterness that may last for generations. What wins arguments is scientific fact, and that may change as the years go by. A good example of this is the geological theory of continental drift, as proposed by Wegener in 1912. When I studied geology around 1950, continental drift was acknowledged in my undergraduate textbook as a crank theory. The first serious confirmation was in 1956, and it was finally established as the dominant theory in the early 1970s. Until that time, anybody who admitted that he or she believed in continental drift was the subject of derision and scorn. Sorry, folks, science is not and has never been the 'idealized portrait painted in textbooks'." [Allan Blair] {54}

"... the four stages of response to any new and revolutionary development [are]: 1. It's crazy! 2. It may be possible — so what? 3. I said it was a good idea all along. 4. I thought of it first." [Arthur C. Clarke]. {55}

"...I suggest that most revolutions in science have taken place outside the lofty arena of the refereed journals, and with good reason. The philosophy by which these journals govern themselves virtually precludes publication of ideas that challenge an existing consensus." [William K. George] {56}

"At every crossway on the road that leads to the future, tradition has placed against each of us, 10 thousand men to guard the past." [Maeterlinck].

> *"An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning."* [Max Planck] {57}.

> *"Peer review is widely seen as a modern touchstone of truth. Scientists are roundly drubbed if they bypass it and 'go public' with their research… The first limitation of peer review is that nobody can say quite what it is… A more pernicious danger is that peer review may reject the important work. As Charles W. McCutchen, a physicist at the National Institutes of Health, has put it, peers on the panel reviewing a grant applicant 'profit by his success in drawing money into their collective field, and by his failure to do revolutionary research that would lower their own ranking in the profession. It is in their interest to approve expensive, pedestrian proposals.' "* [Jonathan Schlefer] {58}.

The sheer massive size and inertia of the modern scientific establishment also exert mind-numbing difficulty in "hearing" and recognizing an innovative scientist's message, even a message of utmost importance, and even if it gets through the censors. For example:

> *"We used to be able to say things once; if the message was reasonable, it had a good chance of becoming a permanent part of the structure of the field. Today, a single publication is lost; if we say it only once, it will be presumed that we have changed our mind, and we therefore must publish repeatedly. This further fuels the large publication volume that requires us to repeat."* [Rolf Landauer] {59}

1.2.2 Some Specific Examples

There are hundreds of examples of new discoveries in science that have at first — and often for an extended period of years — been severely obstructed and ridiculed. Here are just a selected few:

1.2.2.1 *Conservation of Energy*

Von Mayer {60}, the discoverer of the modern statement of the conservation of energy and the mechanical equivalent of heat, was severely chastised for his "insane" work. He was hounded and severely ridiculed. This extremely harsh treatment, together with domestic problems, drove him to a suicide attempt and a nervous breakdown, and into psychiatric treatment for some years. Toward the end of his life, his principle of energy conservation had so greatly increased the ease of calculations and the understanding of systems that the same scientific community — due to the commendable efforts of Helmholtz, Clausius, and Tyndall — began to recognize his great contributions and lionize him. In 1867, he was made a member of the nobility, dying in 1878 with his "insane" work by then well respected. He was fortunate to have the "cur dog attack" reversed in his lifetime. Most scientists with novel discoveries are not so fortunate.

1.2.2.2 *Continental Drift*

Alfred Wegener {61} proposed the theory of continental drift in 1912. The reception was extraordinarily hostile. So ferociously was he ridiculed for the notion that huge continents of rock could "float" and "drift" that his very name, "Wegener", was often used as a synonym for "utter idiot". To refer to someone as "a Wegener" was to cast the greatest slur possible upon that person's mental powers and to label him a gibbering lunatic. Only in the 1960s when sea-floor spreading from ocean ridges was discovered, proving that ocean basins are not permanent features, did Wegener's concept of continental drift concept finally gain acceptance.

1.2.2.3 *Kinetic Theory of Gases*

As pointed out by Paul Nahin {62}:

> *"J. J. Waterston's paper on the kinetic theory of gases, in 1845, was rejected by the Royal Society of London. One of the referees declared it to be 'nothing but nonsense, unfit even for reading before the Society.' ... "Waterston's dusty manuscript was finally exhumed from its archival tomb forty years later, because of the efforts of Lord Rayleigh..."*

Lord Rayleigh was the Secretary of the Royal Society when he had Waterston's paper reprinted nearly a half-century after submitted. Lord Rayleigh also gave an introduction to the paper, regretting it lying so long unpublished when its content was quite important. His introduction is a way of explaining the delay.[25]

---

[25] Waterston's paper was finally published as John James Waterston, "Free and Elastic Molecules," *Phil. Trans. Roy. Soc. Lond.*, Vol. 183, 1892, p.1-79. Lord

1.2.2.4 *Surgical Pain Deemed Necessary*

It has *always* been this way in science. As another example, the famed surgeon Alfred Velpeau wrote in 1839:

> *"The abolishment of pain in surgery is the chimera. It is absurd to go on seeking it today. 'Knife' and 'pain' are two words in surgery that must forever be associated in the consciousness of the patient. To this compulsory combination we shall have to adjust ourselves."* [Martin Gumpert] {63}.

Wryly we observe that today a similar attitude of "we must glory in the pain" — where the "pain" is due to the yoke of COP<1.0 EM systems and of the second law of classical equilibrium thermodynamics — consumes most modern electrical power system scientists and engineers.

1.2.2.5 *The Photoelectric Effect*

Almost every household now knows Albert Einstein's epochal achievements. Yet his formative three papers — on Brownian motion, the photoelectric effect, and special relativity — were published in 1905 while he was working in the Swiss Patent Office. The most renowned physicist of the time was Max Planck. Planck was embarrassed that a scientist who was not even employed in physics was doing such important work in physics. So Planck and other scientists arranged for Einstein to be awarded a chair in physics at a proper university. In their letter to the university, they pointed out Einstein's brilliance in his papers. They also then excused him for straying down the road of the photoelectric effect, because — as they put it — everybody knew that was foolishness, but persons of such brilliance could be forgiven a few such little bobbles along the way. Years later, in 1921, Einstein was awarded the Nobel Prize for Physics, for his work in theoretical physics and *especially* for his explanation of the photoelectric effect.

The Einstein incident is a typical illustration of Arthur C. Clarke's cogent observation:

> *"When a distinguished but elderly scientist states that something is possible he is almost certainly right. When he states that something is impossible he is very probably wrong*." [Clarke's First Law]

Rayleigh's introduction and Waterston's paper also are given in Jefferson Hane Weaver, *The World of Physics*, Simon and Schuster, New York, 1987, p. 632-651.

1.2.2.6 *Amorphous Semiconductors*

Stanford R. Ovshinsky's development of amorphous semiconductors {64, 65} is another modern example. "Everybody knew" that a crystalline structure was necessary in order to have a semiconductor at all; in short, a semiconductor formed out of non-crystalline material was deemed to be totally impossible, never mind this “phase” change that Ovshinsky advanced. However, some young graduate students, e.g., began to look at Ovshinsky's amorphous materials and his phase change approach.

Finally funded by the Japanese, Ovshinsky's company, Energy Conversion Devices, Inc. (ECD) simply placed its amorphous semiconductor devices into working equipment. Copy machines appeared with his amorphous semiconductors installed in them — with the machines and their amorphous semiconductors working very well. More graduate students and post-docs enthusiastically entered the area, did research, and wrote dissertations.

As a result, amorphous semiconductors finally became accepted, and they are now part of the established technology and scientific knowledge base. The interested reader can simply look up Ovshinsky's company and statistics on http://www.ovonics.com . The Japanese have reaped a continuing bonanza from the sales of amorphous semiconductors, because of the shortsightedness and bias of the U.S. scientific community.

1.3 Scientific Disagreement or Street Fight?

It is one thing to scientifically disagree — even strongly — in a technical and courteous fashion, and quite another to engage in ad hominem attacks. The first is science; the second is a cur dog fight. Unfortunately the history of science far too often reveals "cur dog fights" instead of respectful scientific disagreements {66}.

A recent modern example of legitimate research and qualified researchers still being savaged in a "cur dog fight" manner is cold fusion (low energy nuclear reactions). Quoting one learned orthodox scientist, whose name is withheld, speaking to a learned scientist in cold fusion:

> *“How stupid do you think we are? My assessment of you and your colleagues is that you are complete frauds or totally mad. There is no known physical principle that would support the kind of results that you claim your technology can accomplish, nor is there any credible argument why there should be such a principle."*

We answer that "cur dog attack" unemotionally and scientifically. The key is in the scientist's own phrase: "no *known* physical principle." If the experiments work and are replicated, but are not understood by the conventional theory, then they refute the prevailing theory, whether the physical principle is understood or not. Else scientific method has been totally abandoned in favor of "truth by authority and dogma". The experiments clearly show that a previously *unknown* physical principle is at work. The real task then is to rediscover this new principle — that being one of the primary ways that science advances. We propose new principles in this book as a possible explanation of those experiments.

The dogmatic scientist's statement was made in spite of some two hundred (200 at that time; now more than 600) scientific experiments worldwide, in many nations, many at prestigious institutes, where dozens of scientists have reported positive and anomalous results in cold fusion experiments.

So yes, there is indeed a very "credible argument" — *the results and replicability of the experiments*, which are supposed to be the *decisive* statement of science. If that view in science has changed, then we are no longer practicing the scientific method. Instead of accepting successful and replicable experiments and seeking to change the model, the scientist is insisting that we must first understand the principle and thus have a model. This of course is a total violation and reversal of the scientific method. We used aspirin effectively for decades without the slightest notion as to the mechanism enabling its beneficial actions. The mere fact that "there is no known physical principle" for the results achieved has nothing to do with the validity of the replicated experimental results. Instead, it merely substantiates that *there should be a vigorous scientific program to uncover the new principle or principles obviously involved*, since the old model has either failed or been revealed as too limited.

The derogatory statement by the dogmatic scientist is also made from the viewpoint of the conventional nuclear physics model. The conventional physics has not taken into account that all 3-spatial electromagnetic energy associated with charges and dipoles and their fields and potentials in fact comes from the time domain {85, 86, 12, 19}. It has also not taken into account that time itself (as in the time component transported by the time-polarized photon or a time-polarized EM wave) may be comprised of extraordinarily dense energy. Indeed, time appears to be spatial EM energy compressed by the factor c^2, so it has the same energy density as mass {67}, which we pointed out earlier. The smaller the *spatial* energy of the photon, the greater its time component and hence its *time*-energy in seconds expressed in decompressed spatial energy joules[26]. The highest energy particle physics is not *spatial*-energy physics as presently practiced, but *time*-energy physics, where some of the time-energy of photons is transduced into spatial energy. One second of time transduced (decompressed) into spatial energy yields approximately $9\Delta 10^{16}$ joules. Further, every negative charge in the universe continuously accomplishes that decompression, and every positive charge continuously accomplishes the recompression.

In other words, not all physical principles were discovered in the particle physics of, say, five years ago, as the literature since then clearly shows. To assume that everything is already known today is ludicrous; science is never completed. The new principles reported in the present book now are at least "candidates" for the physical principles that do support cold fusion results. There is also another powerful argument for these hypothesized principles: they have produced the final resolution of the source charge problem — something for which the arch skeptic quoted has no solution whatsoever, and which he himself cannot explain. *If the arch critic cannot produce an alternate solution to the source charge problem, and is unaware of that new principle that may solve it, let him go and learn the new principle. Until then, he unwittingly assumes that every charge in the universe is a perpetual motion machine, continuously creating and pouring out EM energy in all directions at the speed of light.* So we return his own argument and approach to him: he should attack the conventional scientific community for accepting the source charge, while having absolutely no principle to explain how it continuously pours out observable EM energy without any observable EM energy input. In short, he should practice what he preaches; else he brands himself a total hypocrite. The alert reader will note that none of the arch skeptics doing all the ad hominem attacks on cold fusion and COP>1.0 EM systems are practicing what they preach.

---

[26] The spatial energy of the photon decreases linearly as the frequency is lowered, while the time component in seconds increases linearly. But the highly compressed time energy, comprising that time component, increases nonlinearly (by the factor $c^2 \div t$). Hence the highest energy photons are actually the low frequency photons — something completely alien to particle physicists — and not the high frequency photons chased by "high energy physics". Indeed, high energy physicists are practicing a high spatial energy physics, which — overall — is a much lower energy physics than radar, microwave, VHF, or ELF, *if the total energy of the photon, to include its time energy, is considered.*

What we are saying is this:

(a) If one's physics knowledge was current five years ago, it may be stale today. A few years ago, no one believed the expanding universe was accelerating. Today we know that it is, rather unequivocally. And by "no known physical mechanism". We will later present a strong candidate for that missing mechanism. The results of experiments will substantiate or refute it.

(b) The proposed principles in this book, e.g., are newer, and may shed light on the mechanism for the cold fusion results as well as other phenomena. The experimental results themselves are irrevocable; any proposed explanation requires validation.

(c) The new principles do explain cold fusion and are consistent with the phenomena encountered in multiple experiments by multiple researchers in many laboratories. We readily admit that the stale and incomplete principles presently utilized in nuclear physics *do not* explain cold fusion. Neither do they explain how a charge continuously pours out EM energy. So do the skeptics attack all those physicists who believe in source charges and their provision of the fields and potentials and all EM energy? Of course not. They themselves believe in that greatest of all “perpetual motion” *faux pas*.

(d) The proposed new principles also solve the source charge problem, which is still ignored by most conventional physicists and electrodynamicists, even though often referred to as the "most difficult problem in electrodynamics" {68}.

(e) Scientists should not be close-minded, but should consider new proposals and let the experimental results decide their validity or falsity — precisely the position taken by the journal *Science* in publishing the results of some new cold fusion experiments. That is the scientific method in action.

(f) Those scientists who remain close-minded, and viciously attack experimentally demonstrated new processes and mechanisms, are guilty of practicing dogma and not science. They are in fact guilty of being what they so frequently charge: *pseudo-scientists*.

When science does not allow proposed new mechanisms and principles to be considered in science following demonstration of new phenomena inexplicable by present models, then science is no longer practicing scientific method. When any scientist rejects these demonstrated new experimental phenomena out-of-hand, he is practicing dogma rather than science. The “friendly skeptic” attitude is welcomed and appropriate. The cur dog attack has no place in science, but only among cur dogs — and those who behave like them.

1.4 A Few Significant and Unresolved Scientific Problems

1.4.1 Time as Structured Energy

Time also has internal structure and dynamics {69, 70}, as we explain later, as well as the same energy density as mass. So it should not be surprising that a tiny bit of time-energy, transduced into ordinary spatial EM energy, might be capable of energetically inducing a wide variety of nuclear reactions. One has transduced not only "gross" energy, but also energy structuring and dynamics — which *act internally* on any object with which interaction occurs. Further, the overall spatial energy component and the time-energy component in a photon are canonical. Being quanta, all observable photons have the same fixed "total magnitude" with respect to angular momentum. The photon is also "made" of $(\div E)(\div t)$ which may be taken as $(\div E)(c^2 \div t)$ in terms of *purely spatial* energy equivalent content, assuming that the time energy is decompressed into spatial energy. As can be seen, the so-called "low (spatial) energy" photons are precisely the photons that transport the greatest time-component. When that time-energy component $(\div t)$ is converted to spatial energy $(\div E_C)$, the formula is $(\div E_C) = c^2(\div t)$. As can be seen, the converted time energy will produce far more spatial energy in the *converted* interaction, than was carried by the photon in its spatial energy $(\div E)$ component prior to interaction and time-energy transduction. Hence the highest energy photons are actually low frequency photons — under the circumstances where time-energy transduction into spatial energy is involved. With transduction, the highest energy particle physics can be conducted at low frequencies and low spatial energy if conversion of time-energy occurs in the interaction.

Since cold fusion involves transduction of a little bit of the time-energy, the total lack in physics of knowledge of time-energy transduction into spatial energy explains why conventional nuclear physicists find the cold fusion results so confusing and astonishing.

1.4.2 The Obsolete Foundations of Classical Electrodynamics

Another modern example of science's resistance to change is the failure to update classical electrodynamics to include the active vacuum interaction and the interaction with the local curved spacetime[27], as we shall discuss in Chapter 2. It has been scientifically established (in particle physics) that there can be no equilibrium of any physical system without accounting for the vacuum interaction.[28] To simply observe a physical electrical power system sitting there stably and running, is to prove its vacuum interaction in order to even *have* such equilibrium.

An interesting point then arises since every charge and dipole in the system is a *broken equilibrium* in the exchange between the material system and the vacuum. Not only does an electrical circuit receive some energy from the vacuum, but also the energy from the vacuum is *massively hemorrhaging* from the vacuum into the system (and back out)! The isolated charge, e.g., is infinite if one removes the screening clustered virtual charges of opposite sign, as is well known in modern physics. Further, the screening virtual charges also have infinite charge, again as is well known. The difference between these two infinite values of charge, e.g., gives the standard finite observed charge of the charged particle. The lay reader may wish to positively verify that statement; e.g., as given by Nobelist Steven Weinberg.[29] Quoting:

> *"[The total energy of the atom] depends on the bare mass and bare charge of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite*

---

[27] We again stress the concept of the *supersystem*, introduced in footnote 11.

[28] E.g., see T. D. Lee, *Particle Physics and Introduction to Field Theory*, Harwood, New York, 1981. On p. 380-381, Lee shows how there is no symmetry of matter alone, but only of matter and vacuum.

[29] Steven Weinberg, *Dreams of a Final Theory*, Vintage Books, Random House, 1993, p. 109-110.].

> *because it depends on the infinite bare mass and charge, and the energy shift ... that is infinite because it receives contributions from virtual photons of unlimited energy."* [Steven Weinberg].

So a simple charged particle actually involves a polarization of the vacuum involving two charge energy functions — the bare charge and the charge change functions — each of which has a known and recognized COP = ←!

A priori, the conventional system's overall actions must incorporate functions (whether intentional or unintentional) that continuously adjust to provide *net equilibrium in that exchange by adjusting the hemorrhaging-out to equal the hemorrhaging-in*. In this book, we will spell out just what those unintentional functions are, such as the ubiquitous closed-current-loop circuit, and how to beat them so that excess energy from the vacuum can be utilized by the system to power the loads.

1.4.3 What Powers an EM Circuit or Electrical Power System?

Another modern example of science's resistance to change is the continued engineering of electrical power systems with the erroneous notion that mechanically powering the shaft of a generator directly powers the power line. The notion is that the generator transduces some of the mechanical shaft energy into output EM energy added to the power line {71}. To the contrary, *all EM systems are powered by energy extracted from the vacuum*. They are not powered by the mechanical energy we input to the shaft of a generator, or by the chemical energy in a battery. In this book, we discuss at some length what powers the EM system, and we have previously published the basis for our "shocking" statement {12, 72}. The basis for how an EM system is powered by energy from the vacuum has been known in particle physics for nearly a half-century {73}, since the experimental proof of broken symmetry {74}, including the broken symmetry of opposite charges (and hence of any dipole) in its fierce energy exchange with the active vacuum {75}.

A generator or battery expends all of its available energy to separate its internal charges and form its source dipole between the terminals. Once made, the dipole's broken symmetry — in its violent energy exchange with the active vacuum — converts virtual photon energy absorbed from the vacuum into observable EM energy, and emits it — *pours it out* — through the terminals and along the power line, filling all space around the conductors.

If one wishes to appreciate the enormity of the vacuum changes engendered by formation of that source dipole, simply visualize those two infinite charge energy functions (from Weinberg's quotation above) for every charged particle on the ends of that dipole. The dipole then is a great set of offsetting infinities in EM energy (photon energy) boiling and being ordered in the vacuum. Making a simple little dipole or producing a simple charge initiates into motion an enormous set of infinite energy changes in the vacuum! None of these startling, enormous vacuum energy changes and functions is modeled in classical electrodynamics and electrical power engineering. So when we speak of the Heaviside nondiverged energy flow component pouring from the terminals of a generator, as being a trillion or ten trillion times as great in magnitude as the intercepted Poynting component, the reader should not be appalled. The reader should recall that we are dealing with the difference between sets of infinities that combine to provide finite differences. These finite differences can be very small, or they can be extraordinarily large, depending on the exact situation. In the case of a source dipole formed in an ordinary generator or battery, the difference of the infinities is very large.

Figure 1, adapted from Kraus {76}, illustrates that small part of the external energy flow around a typical transmission line that is intercepted by the surface charges and their fields, and that is drawn into the wires to power the circuit as these charges are driven axially back and forth across the conductors. The surface charges are constrained to the "drift velocity" (usually a few inches per hour) movement down the wire by repulsion of the electrons ahead of them.

The spinning, *longitudinally restrained* electrons precess laterally, thus withdrawing across the transverse axis of the wire. So the laterally precessing electron withdraws a little ways, withdrawing that first small portion of its fields and their energy increased by intercept of additional energy from the outside energy flow. The further part of the fields is not withdrawn into the conductor, and is not used to power the electrons.

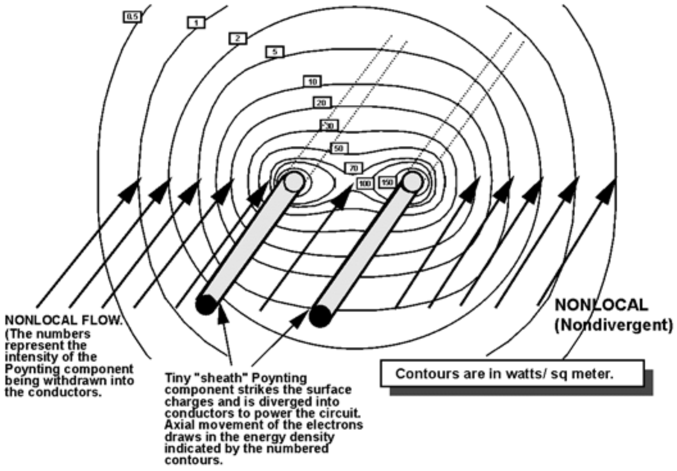
0.5
1
2
5
10
20
30
50
70
100
150
NONLOCAL FLOW. (The numbers represent the intensity of the Poynting component being withdrawn into the conductors.
Tiny "sheath" Poynting component strikes the surface charges and is diverged into conductors to power the circuit. Axial movement of the electrons draws in the energy density indicated by the numbered contours.
NONLOCAL (Nondivergent)
Contours are in watts/ sq meter.

Figure 1-1 Poynting (caught energy) flow contours surrounding a transmission line.

Most of the available energy flow misses the circuit and is not drawn into the conductors. However, that large component is not shown by Kraus, and Poynting did not consider it. Heaviside {5c} pointed out that the remaining flow component is hardly reduced (hardly changes direction) from the entire flow component before the extraction of the small Poynting component. Given sufficient intercepting charges outside the wires in separate receiving circuits, the total remaining energy flow that could potentially be intercepted is enormous — far more than the feeble amount of energy that is input to the generator shaft or that is in the chemical energy of a battery.

This is easily established by actual experiments placing intercepting charges in separate "receiving" circuits in the otherwise nondivergent energy flow outside the conductors, or one may show it with Bohren's experiment {24}. Collecting additional energy completely outside conductors is one part of the COP>1.0 operation of the motionless electromagnetic generator (MEG) {37}, to be discussed later in this book. The Sweet device {29a} established sustaining self-oscillation of the barium nuclei in a barium ferrite magnet, by using the surrounding vacuum as a semiconducting medium.[30]

1.4.4 The Incompleteness of Aristotelian Logic

Another false notion usually perpetuated in much of science and mathematics is that Aristotelian logic is complete and consistent {77, 78}. To the contrary, it is both incomplete and inconsistent, as is easily shown. Let us use the symbols "A" for a particular thing, "Ā" for that which is not-A, "∑" for "is identical to," "≢" for "is not identical to", "˅" for "or", and "˄" for "and". Using these symbols, the three laws of Aristotelian logic may be expressed as follows:

$$A \sum A \qquad [2]$$

$$A \not\equiv \bar{A} \qquad [3]$$

$$A \vee \bar{A} \qquad [4]$$

The first law states that a thing is identical to itself. The second states that a thing is not identical to that which is not itself. The third states that a thing is either itself, or it is something else. Those are the three laws of Aristotelian logic. As written, the process of perception, observation, etc. *has been excluded*. More on that in a moment.

Specifically excluded are the laws

$$A \not\equiv A;\ A \sum \bar{A};\ A \wedge \bar{A} \qquad [5]$$

Which says that a thing is not identical to itself but is identical to something else that is not itself.

Heraclitus posed a profound challenge to Aristotelian logic, a challenge that has not been adequately resolved by Aristotelians to date. He simply observed that, for a thing to change, it had to change into something else. “But then,” he asked, “how can a thing be itself but also something else as well?” In philosophy that is still known as the unresolved "problem of change". It has no resolution in Aristotelian logic, but can be resolved in an extended logic we shall advance. The solution to Heraclitus' problem of change is the nature of observation, as advanced in this book.

---

[30] E.g., see Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys.,* 52(1), Jan. 1984, p. 19-21. The vacuum may be regarded as a semiconductor. In particular, the vacuum in the region close to the nucleus of a superheavy element is analogous to the inversion layer in a field effect transistor. The authors introduce the idea of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that so can be the vacuum.

We now critique Aristotelian logic, by observing a simple Venn diagram of the type used in "proving" logic theorems in axiomatic logic. See Figure 1-2.

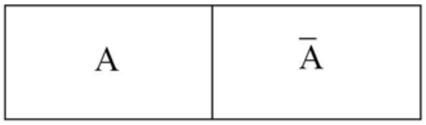

1-2a. Usual Venn diagram of A and A.

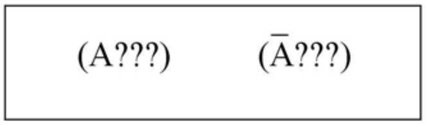

1-2b. Applying second and third laws to divider.

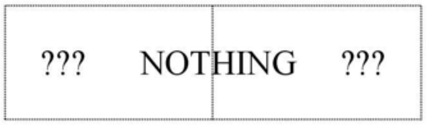

1-2c. Applying second and third laws to divider and outer boundary.

Figure 1-2 Applying Aristotle's laws to Venn diagram used to prove logic theorems destroys the entire diagram.

In Figure 1-2 a, we have introduced a dividing boundary line between A and Ā. *Unfortunately that dividing line belongs entirely to both A and Ā.* So it violates all three Aristotelian laws, and must be removed.[31] Therefore, we remove it in Figure 1-2 b. Now we have neither a discernible A or a discernible Ā, but we have removed naught but the boundary separating them, so we may argue that they are both still there although not discernible. However, if they are not discernible, we cannot distinguish what A is or what Ā is, and so we cannot discern whether they are identical or not. Here again this violates all three laws of Aristotelian logic, so we have to remove the diagram. Indeed, the outer rectangle line is also such a boundary, between the "inside" and the "outside" (the not-inside), and so it must be removed. In compliance with Aristotle's laws, we are left with the trivial diagram shown in Figure 1-2 c, which is nothing at all.

---

[31] A clever fellow once proposed regarding the boundary as a total discontinuity, being neither A nor Ā. However, in that case A and Ā could never meet, so there could not be a cause interacting with not-cause to produce an effect. So that suggestion as to how to "fix" Aristotelian logic does not hold.

If this Venn diagram method is objected to, then we must insist that all those papers and texts using that method of proof of logic theorems be either abolished or corrected!

Sometimes the objection is raised that, as far as the center boundary line is concerned in Figure 1-2a, the line "belongs to A on the left and belongs to Ā on the right". Fine! Then we have a unique situation where a line (call it L) is an entity made of two entirely different things, which we may refer to as L_L and L_R. Yet $L \sum L$, without any qualifications as to "sides" L_L and L_R. It can be shown that every point in L is a point in L_L, and is simultaneously a point in L_R. The point in L is obviously identical to itself, by Aristotle's first law. Further, the same point in L_L and in L_R is identical to itself, by the same first law. But L_L is the set of all such points, and so is L_R, and so is L. Therefore $L \sum L_L \sum L_R$, since each one consists of each point that the others consist of. There is absolutely no distinction between the three things themselves, unless we are to violate Aristotle's laws.

Note that in passing from left to right across the boundary, a "change" occurs. This simply re-resurrects Heraclitus' original objection that, according to Aristotelian logic, there can be no change. That is, a line cannot change from "a different entity as seen from the left" compared to that entity as seen from the right. None of Aristotle's three laws contains a "left and a right" for A or Ā.

So the question of logic is a little more complex than writing three simple laws and drawing simplified Venn diagrams. In short, one cannot have Aristotelian logic without having something else *outside* it, which follows directly from Gödel's theorem and proof {78}. Note that we are introducing the required notion of "perception" into the formal notions implied as axioms in Aristotle's laws, as that "something outside the prescription of the laws themselves" and in addition to them. We are perfectly free to *use* Gödel's theorem and observation.

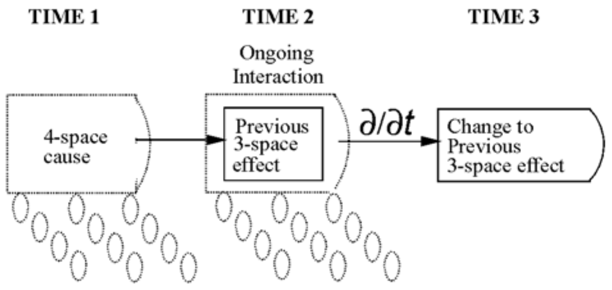

Figure 1-3 Mechanism for observation of change to a previous observation.

See Figure 1-3. We now add the notion that a thing is a perceived, thought, or observed thing. So let us use the symbol ⇒to mean *explicit*, i.e., *output* by a given perception, thought, or observation process as the result of a ∈/∈*t* operator having been applied to whatever exists prior to perception, thought, or observation. The symbol ⇔is used to mean *implicit*, e.g., when one observes an observation "Y" at time two, and later is deciding in time three whether or not that observation "Y" in time two is identical to a previous observation "X" made during time one, then *memories* of the observations of X and Y are involved in time three rather than the observations X and Y themselves, and so the observations of X at time one and Y at time two — by whatever manner they were observed and according to whatever decision algorithm is used — is said to be "implicit".

We note that we can know nothing about the so-called "thing in itself" *without* thought, perception, observation, or other process involving ∈/∈*t*. Further, at the moment the ∈/∈*t* operator is applied, time momentarily ceases. The resulting perception, thought, or observation exists therefore as a "frozen output" at that single moment in time. To have it "persist" or exist a moment later, we have to apply the ∈/∈*t* operator *again*, and stop time again, so that we again perceive, think, or observe.

But at any moment later than when we made a particular perception, thought, or observation of "A", *that* specific "perceived A" no longer exists, except in our memory as a recording of "observed A" that we can *continually recall*. Our conscious mind is a very fast serial processor, with only one "perception" or "thought" at each fleeting moment — only one slide at a time in the slide projector, so to speak. But it is very rapid. However, our so-called "unconscious" mind is totally conscious, just multiply so. It is a massively parallel processor, and has "a great many slides in the slide projector" at any given time.[32] Hence in our genetic multiprocessor mind, we can indeed record, recall, compare, etc. — and perform all the modern massively parallel computer operations.

Let us now re-examine the three Aristotelian laws. Let us label the "slide snapshots" with the time instants when each "perception, thought, or observation" is or was consciously made, by use of a subscript, such as A_1, which means *what was perceived, thought, or observed as "A" in time-snapshot 1*. Let us also note that each of the symbols $\sum$, $\not\equiv$, $^\vee$, and $^\wedge$ actually involves the output of a *comparison and decision algorithm* in the massively parallel processor, after the two observations X_1 and X_2 were made. Let us use the symbol ♥ to mean "results in the decision that" or "implies that". As a check for identity, e.g., simple comparison algorithms for determining identity or nonidentity might be

$$\{[A_1 - A_{(2)}] = [0]_3 \heartsuit_4 (A_1 \sum A_2) \qquad [6]$$

$$\{[A_1 - A_{(2)}] \prod [0]_3 \heartsuit_4 (A_1 \sum A_2) \qquad [7]$$

Thus we more precisely rewrite Aristotle's laws as

$$A_1 \sum_3 A_{(2)} \qquad [8]]$$

$$A_1 \not\equiv_3 \bar{A}_{(2)} \qquad [9]$$

---

[32] E.g., this can be seen by a moment's reflection. At any one time, the "unconscious" mind is controlling and directing a great multitude of ongoing physical processes, is also processing short-term and long-term memory processing, filing conflicts for resolution or later presentation to the conscious mind in symbolic fashion, etc. One can physically measure the electromagnetics associated with this activity, after the mind transduces its time-polarized EM interactions into 3-space EM actions. The process is two-way, and certain recorded EM stimuli will be "reverse-processed" back to affect the time-polarized EM operations of mind. Mind operations are electromagnetic; but they use time-polarized (scalar) photons and time-polarized (scalar) EM waves which are unobservable a priori. Mind operations are also electrodynamically engineerable, but that is beyond the scope of this treatise.

$$A_1 \curlyvee_3 \bar{A}_{(2)} \qquad [10]$$

Where — surprise! We have now accounted for all those different times and periods of observations, processing, comparison, decision, etc. We are dealing not with some mystical "thing-in-itself", but with interactions (perceptions, thoughts, observations, associations) in the mind and psyche of the observer. And hidden in time 3 is the application of a decision algorithm such as given in [6] and [7] above.

Now the first law [8] states that in time one an observation was made and named "A", notated (A_1), by comparing it to a previously recorded observational memory of what we call "A". The existence of that *memory* of what we call A is implicitly assumed in the first law, as well as the others. We do not show the memory itself in [8], [9], and [10]. In time two an observation X_2 was made (not notated), but it is not then known at the moment whether that X_2 is A or $\bar{A}$. So in time interval three the decision algorithm to determine identity or non-identity occurred in the mind, and the zero output of that algorithm [6, 7] established that $A_{(2)}$ was actually identical to A_1, *according to the decision algorithm actually used. The algorithm matters, and it too is a variable.*

A similar process occurs in law two [9], but this time the algorithm had a different output. Note that the little line over $X_{(2)}$ to make it $\bar{A}_{(2)}$ was not assigned until time interval four (not shown) after the decision algorithm had given its output in time interval three.

The second law merely states that the operation of the decision algorithm in time interval 3 found the two snapshots (in time 1 and in time 2) not to be identical, *by the decision algorithm and comparison process utilized. Again, the algorithm matters, and it too is a variable that must be taken into account.*

All three laws [8, 9, 10] written one after the other *assume* that the identity algorithm does not change between snapshot 1 and snapshot 2, in all three of them. This is the key point.

With two different versions of the identity decision algorithm, the results of two different comparisons may differ. If the identity decision algorithm does not change during the time between snapshot one and snapshot two, then we have one case. This is like a person with good color vision, looking at a red marble beside a black marble. That observer clearly distinguishes the marbles, and to him they do *not* appear identical. Every time he looks again, they still differ because he did not change his decision algorithm between looks. On the other hand, a color-blind observer cannot distinguish the marbles, and to him they are seen as identical. Every time he looks, they are seen as identical, because his identity decision algorithm did not change.

Or, suppose a proponent argues that the "rightness" or "leftness" of the dividing line L, between A and Ā in the Venn diagram, can be taken into account. He is actually invoking a different algorithm (it belongs to the left and therefore to A) in one time snapshot than the algorithm (it belongs to the right and therefore to Ā) used in the other time snapshot. So his right side and left side of a line implicitly invokes the very point we are making: *the algorithm utilized is a variable and must be accounted.*

The point is this: *identity — whether in perception, thought, or observation — is not absolute, but depends upon the precise nature of the operation of the perception, thought, or observation process utilized and specifically on the "decision algorithm" used for "determining" identity or nonidentity*. For example, two antennas certainly "see" quite differently, if one is a VLF (very low frequency) antenna and the other is an IR (infrared) antenna! As another example, two observers in different frames may see a particular object as quite different observed things! In n-space, an observer in a frame at right angles to the lab frame, will see any mass in the lab frame as a wavefront going at light speed; in short, as a photon or photonic object, whereas the observer in the lab frame will continue to see it as just a common old mass object. So the "same object" *physically* differs to the two different observers, according to their frame of reference. An observer whose frame is rotated by three orthogonal rotations from the lab frame, will see that object as an "object existing in time only," i.e., as simply a sort of "thought" image, so to speak. To be more scientific, he may assign it to something called the "virtual state".

In short, we can violate any and all of the laws of Aristotelian logic, because identity per se is *perceived, thought, or observed* identity — the output of a variable decision algorithm — and is not absolute.

For clarity, we add a fourth law that violates all Aristotle's three:

$$A_1 \sum_3 \bar{A}_{(2)} \qquad [11]$$

All this really says is that, in time 3, the decision algorithm being used was changed from what it had been in times 1 and 2, and now could not distinguish between what was seen in snapshot 1 and what was seen in snapshot 2, even though to some "ultimate" observer the two were distinctly different.

In addition, we say that it "implicitly includes" the negations of all three Aristotelian laws. We then add an "application rule" (or a fifth law, as one wishes) as follows:

$$\{ [(A_1 \Sigma_3 A_2) \Rightarrow \curlywedge (A_1 \not\equiv_3 A_2)] \Leftrightarrow \}_5 \curlyvee \{ [(A_1 \Sigma_3 A_2) \Leftrightarrow \curlywedge (A_1 \not\equiv_3 A_2)] \Rightarrow \}_5 \qquad [12]$$

This winds up producing an extension of Aristotle's logic, so that the following explicit laws emerge:

$$A_1 \Sigma_3 X_{(2)} \heartsuit A_1 \Sigma_3 A_{(2)} \qquad [13]]$$

$$A_1 \not\equiv_3 X_{(2)} \heartsuit A_1 \not\equiv_3 \bar{A}_{(2)} \qquad [14]$$

$$A_1 \curlyvee_3 \bar{A}_{(2)} \qquad [15]$$

We also have the exact opposites of those three laws implicitly. The opposites can all be congealed into a single fourth law:

$$A_1 \Sigma_3 \bar{A}_{(2)} \qquad [16]$$

To make sense of these, we also have the master application rule or 5th law:

$$\{ [(A_1 \Sigma_3 A_2) \Rightarrow \curlywedge (A_1 \not\equiv_3 A_2)] \Leftrightarrow \}_5 \curlyvee \{ [(A_1 \Sigma_3 A_2) \Leftrightarrow \curlywedge (A_1 \not\equiv_3 A_2)] \Rightarrow \}_5 \qquad [17]$$

We prefer to refer to this extension as a "four-law logic", where equations [13], [14], and [15] are the normal "explicit" laws, with the implicit law [16] understood but not explicitly written, and where equation [17] is the master application rule that puts it all together. But if one wishes to be rigorous, we have specified a 5-law extended logic that contains but expands Aristotle's 3-law logic. The application rule is the fifth law.

The point is this: In every case, we have a part of the perception, thought, or observation that obeys the explicit laws, and we also have a part that obeys the implicit laws. If opposites are not *explicitly* identical, then they are simultaneously *implicitly* identical. If opposites are *explicitly* identical, then simultaneously they are also *implicitly* not identical.

The new approach then covers what we observe more completely. As an example, any "wave" (observed) is implicitly a non-wave (particle), but not observably so. If its particle nature is observed and therefore explicit, its wave nature is unobserved and therefore implicit, and vice versa.

One also resolves such dilemmas as stating something "does not exist" but having to first call it into existence to even state it does not exist. In the new approach, it simply does not *explicitly* exist, but *implicitly* exists. *Not-being* is something like that: explicitly it doesn't exist, but implicitly it exists. The great problems of the philosophers were never solved, mostly because the logic brought to bear was Aristotelian and incomplete. The ultimate answers they sought all involved the fourth and fifth logic laws. So the ultimate answers all appeared to them to involve such things as "the accursed necessity for the identity of opposites." Precisely!

In physics, there has been the same problem over whether a fundamental particle is a particle or a wave. Physicists argued fiercely until they realized the futility of further argument, and settled for the duality principle. That simply states, well, in one case it can be a particle as observed in your experiments, and in another case it can be a wave as observed. So simply treat it as whichever one is useful! The philosophers, on the other hand, split into different schools, where each school had its "interpretation" usually involving a very "smooth" position statement of the school's position. In physics also, there are eight or more "interpretations" of quantum mechanics, for example. So the same centuries-old problem is still with us today in modern physics.

Opposites are no longer the formidable opponents they once seemed. Instead, they become more like the two sides of a coin, lying with only one side up. To see heads or tails is to decide which is explicit. The other is always there, implicitly. But notice also that the "two sides" require a 3-dimensional object, and not just the 2-dimensional "head" or the 2-dimensional "tail". This juxtaposition of "identical" opposites is strongly met with in modern physics. For example, quoting Lee[33] :

---

[33] T. D. Lee, "Space Inversion, Time Reversal and Particle-Antiparticle Conjugation," *Physics Today*, 19(3), Mar. 1966, p. 23. Positive charge is really *observation* imposed on negative charge moving backwards in time, and the positron is really observation imposed upon an electron traveling backwards in time, etc.

> *"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into minus, running time backward and turning things inside out."*

Any addition to the Aristotelian laws of logic is actually a higher dimensional form of it. Three-law Aristotelian logic is in fact fitted to primitive observation — the result or output of single-photon interaction and observation processes.

This "four-law" logic (or "five-law" logic if one includes the master application rule as another "law") has proven very useful over the years, particularly in bioenergetics applications. It has also been most useful in attempting to decipher that branch of Russian energetics weapon science called "psychoenergetics", and in attempting to decipher the mind-body coupling mechanism, but that is beyond the scope of this chapter {79}.[34]

1.4.5 Substitution of Effect for Cause

Refer to Figure 1-3 again, to see the relationship of the unobserved cause, the interaction of cause and a previous effect, and the production of the resulting "new effect" as a change to the old effect or the output of a second replica of it. The greatest non sequitur in electrodynamics — and in parts of physics such as mechanics — is the widespread substitution of the *effect* for the *cause,* as a result of considering an observable to *persist in time* without any physical interaction. We discuss that more fully in Chapter 2, particularly with respect the notion of the same EM field existing both as a component of a material environment and as a component of a nonmaterial environment, but with the "force reducing to zero" in the latter. Jackson {80} points out the prevailing view of electrodynamicists as follows:

> *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable*

---

[34] The interested reader is referred to my website, www.cheniere.org, for papers dealing with the nature of the mind-body coupling mechanism and the body-mind coupling mechanism, as well as direct engineering of the mind, and limited information on the weapons referred to.

> *quantities such as force somehow involve the product of charge and field."*

On the other hand, Bunge {81} very clearly stated that the standard electrodynamics and physics itself are in terrible shape. Here is a direct quotation:

> *"... the best modern physicist is the one who acknowledges that neither classical nor quantum physics are cut and dried, both being full of holes and in need of a vigorous overhauling not only to better cover their own domains but also to join smoothly so as to produce a coherent picture of the various levels of physical reality."*

We shall try to further clarify these unresolved difficulties with the field concept in Chapter 2.

1.5 Electrodynamics is Still Developing

1.5.1 Statement by Penrose

Roger Penrose, one of the leading physicists of our time, pointed out the still-developing nature of electrodynamics with respect to the developments from conventional gauge-theoretic interpretation of Maxwell's equations. Quoting Penrose {82}:

> *"These facts should not, however, deter theoretical or experimental physicists from seeking alternative descriptions, unconventional formulations, surprising electromagnetic effects, or radical generalizations. The various articles in this book provide the reader with a great variety of different kinds of approach to developments of this nature. We have historically motivated accounts, suggestions for new experiments, unconventional viewpoints and attempts at generalizations. We also see novel and ingenious formulations of electromagnetic theory of various different kinds... I am sure that this book will make it clear that electromagnetism is a subject that is in no way closed to stimulating new developments. It is very much alive as a source of fruitful new ideas."*

1.5.2 An Example of a Ubiquitous Error

One of the rather horrid "bad examples" of ubiquitous errors in electrodynamics is the conventional illustration of a so-called planar EM wavefront moving through space, as shown in Figure 1-4.

A succinct evaluation of that diagram and notion is given by Evans {83}, one of the great electrodynamic theoreticians of our times:

> *"I broadly agree ... that the transverse plane wave view of vacuum electrodynamics is the biggest blunder of twentieth century science."*

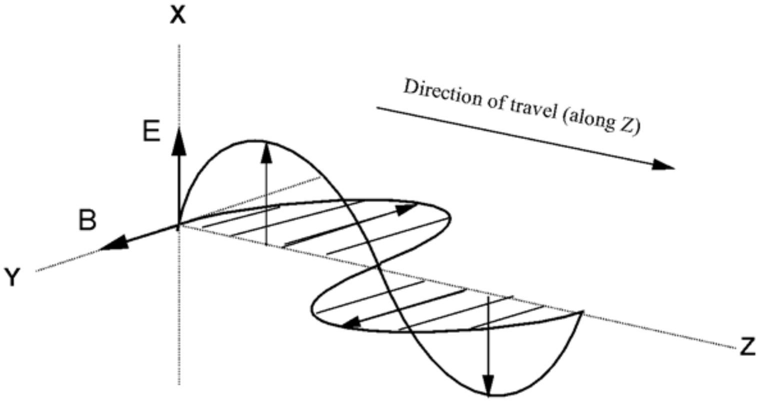

Figure 1-4 The erroneous but ubiquitous supposed illustration of the transverse EM wave in 3-space.

Dr. Robert H. Romer, former Editor of the *American Journal of Physics*, also chastised the horrid diagram shown in Figure 1-4, purporting to illustrate the transverse plane wave traveling through 3-space. In endnote 24 of his noteworthy editorial, Dr. Romer {84} takes that diagram to task as follows:

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we*

> *should] find out how that diagram came to contaminate our literature in the first place."*

Indeed, each point along the z-axis, is a momentary "3-space frozen slice" of the ∈/∈*t* operator being applied to the ongoing 4-space interaction of the causal field in 4-space interacting with the previous "observed effect" (in this case, the previous 3-space slice) to produce a change to it. So the entire diagram is a set of "frozen 3-space observations", in which nothing at all is moving, and where the fields shown are in the mass (detecting/observing) medium. There is no such thing as a "traveling EM wave in 3-space", because "moving" or "changing" or "traveling" *a priori* must involve time as well as 3-space. What we are looking at in Figure 1-4 is the iterative and very rapid results of continual "3-space observations" (frozen 3-snapshots) of an on-going 4-space interaction.

1.5.3 An Extremely Important EM Omission

There is an enormous amount of development yet to be done in electrodynamics! As a simple example, for nearly a century there has been waiting quietly a fundamental "internal" or "infolded" electrodynamics by Whittaker {85, 91a, 86}, based on more primitive initial work by Stoney {87}, existing inside — and *comprising* — all conventional EM potentials, fields, and waves. This internal electrodynamics {88} has been essentially ignored in the West, but has already been weaponized in the weapons laboratories of several nations under the aegis of *energetics*. The U.S. Secretary of Defense, Secretary Cohen {89}, alluded directly to some of these weapons in a speech in 1997 as follows:

> *"Others [terrorists] are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our [counterterrorism] efforts."*

The weapons are already being utilized against various nations in the form of terrorist acts in an undeclared war {89}, and have been so used since the 1950s.

This internal longitudinal-wave EM and its dynamics is a much more fundamental electrodynamics (actually a subset of a unified field theory) consisting of correlated longitudinal EM wavepairs in both the 3-space and time domains. Each Whittaker decomposition wavepair is comprised of an incoming longitudinal wave — a "time-polarized" EM wave {90a-90c} — in the time domain, perfectly correlated with an emitted longitudinal EM wave in 3-space. See Figure 1-5. This multiwave structuring and dynamics in both the time domain and in 3-space simultaneously *comprises* the scalar potential.

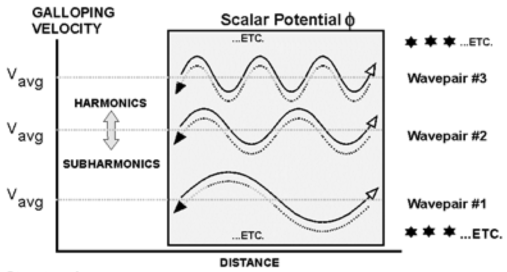

The Structure Is:

- A harmonic set of longitudinal phase conjugate wavepairs.
- In each wavepair the two waves superpose spatially after detection, but travel in opposite directions.
- Prior to detection, the convergent wave set is in the imaginary plane, and hence is not observable. It is EM energy incoming to the potential (dipolarity) from the time domain.
- The charges receive the complex convergent EM energy, transduce it into real EM energy, and emit enormous energy at the speed of light in all directions -- which includes bidirectional pairs in 3-space (after the reaction, being after "observation").
- This produces the fields and potentials from the "source charge or dipole."
- Mandl and Shaw argue that the scalar (time-polarized) photon and longitudinal photon are observable only in similar pairs, which then makes the instantaneous scalar potential. Thus quantum field theory work strongly supports the "negative resistor" interpretation of the scalar potential and our solution to the source charge and source dipole problem.

Figure 1-5 The scalar potential is a harmonic set of phase conjugate longitudinal EM wavepairs.

We specifically stress that the dynamics occur in the *time* domain (in its specific structuring, propagation pace, etc.) as well as in 3-space.

Deliberately changing the internal longitudinal EM wave structuring produces precise energetics (both energy and dynamics) in spacetime itself. This is a special form of "spacetime curvature structure and dynamics" which we call a *spacetime curvature engine* or *vacuum engine*. Together with superpotential theory in the original form initiated by Whittaker {91a} and extended by others {91b, 91c, 91d}, this internal electrodynamics — together with scalar potential interferometry {92} — creates all ordinary EM fields and waves {79}. Superluminal communication is possible using the internal longitudinal EM wave structure of the EM fields, potentials, and waves {93a, 93b}.

1.6 A Summarizing Letter

Our thrust in this book is to attempt to (1) reveal some of the major flaws in electrodynamics that have prevented development and use of COP>1.0 electrical power systems, and (2) point the way toward the development and use of such systems. To give a general summary of the ground this book will cover, we conclude with an informal E-mail letter I sent to the recent editor of *American Journal of Physics*, pointing out the greater implications of his cogent observation of that "dreadful diagram" and where it leads. Indeed, it (i) leads to a dramatic change in electrodynamics, (ii) changes the way in which we regard "propagation of EM energy through 3-space", (iii) solves the long-vexing problem of the association of the source charge or source dipole with its fields and potentials and their energy, (iv) changes dramatically the way we view what powers an electromagnetic circuit or power line, (v) allows extraction and use of copious EM energy from the vacuum, and (vi) solves the energy crisis permanently.

Here is the letter sent informally to Dr. Robert H. Romer — *slightly edited* to improve the grammar, correct one error of misstatement, and with reference citations added and listed at the end of this chapter:

To: Dr. Robert H. Romer Thursday, Mar. 8, 2001
American Journal of Physics
Amherst College, Box 2262
Amherst, MA 01002

Personal communication

Subject: Implications of your cogent comments on that atrocious illustration

Dear Dr. Romer:

This is not a submission of a manuscript, but a personal communication to you on a matter of great importance in physics, directly related to endnote #24 of your seminal editorial {94}.

It will take a little exposition, so please bear with me and read this when you have the spare time to do so. We are going to show you how your keen insight can be extended to solve some extraordinarily formidable foundations problems in present electrodynamics. We will also show how to apply the implications of your insight to totally solve the present electrical energy crisis permanently.

In your endnote #24, {94} you took to task (quoting):

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

Dr. Romer, you have lifted the corner of a dark cover concealing one of the most important flaws in electrodynamics and in fact in all of physics: the unwitting and pervasive substitution of the effect for the cause. A marvelous extension to the present physics is enabled if one removes this terrible non sequitur in physics, and particularly in electrodynamics. For openers, one solves what has been called the most difficult problem in electrodynamics (the problem of the source charge and the association of its fields and potentials and their energy) {100}. I will solve that problem for you in this informal write-up. I will also explain how to extract enormous EM energy from the vacuum, anywhere and anytime, easily. Extracting it is easy; catching it and using it to power loads without killing the extraction process is another matter.

One also gets a unified field theory, engineerable by novel electrodynamic means, as is steadily being shown by a series of rigorous Alpha Foundation's Institute for Advanced Study (AIAS) papers published in various leading journals (and more than 90 of them carried on a Department of Energy website restricted primarily to DoE scientists). Dr. Myron Evans, Director of the AIAS, has over 600 papers published in the literature, including such journals as *Physical Review*, *Foundations of Physics*, *Physica Scripta*, etc. Many of the other AIAS co-authors are excellent theoreticians and scientists.

The early pioneers of electrodynamics (Maxwell etc.) all assumed a material ether filling all space {95}. To these early scientists, there was not a single point in the entire universe that was devoid of matter, because the ether was present there. Hence their outlook as to the nature of EM fields etc. was quite material. Faraday conceived his "lines of force" as physical, taut strings, so that perturbations were "plucking these taut strings".

Maxwell himself points out in his famous "Treatise" that he specifically captured the thinking of Faraday in his theory. In fact, Maxwell wrote a *material fluid flow dynamics* theory.

In the light of more modern knowledge, let us see the impact of these and similar early *but still retained* erroneous electrodynamics assumptions.

First, observation/detection is totally spatial, as is well known in quantum mechanics. In fact, observation is a $\in/\in t$ operator imposed upon 4-space (LLLT spacetime), yielding a frozen instantaneous snapshot LLL of an ongoing 4-space dynamic process. At the next instant, that particular *previous* observation no longer persists. Why?

Well, no observable *persists*, since it is only an instant frozen 3-space snapshot, at a single point in time, *a priori.* Here again we have another horrendous non sequitur in all of physics: the assumption that observables "continue to exist" and therefore persist in time in a passive manner. In fact, there is an interactive process that generates their (seeming) persistence, involves time, and continually changes mass into masstime and back to mass, etc. With the reader’s permission, we will pass discussing that mechanism until another time (pun intended!).

What we conceive as "an observable such as mass, traveling through space and persisting in time while doing so", is actually an iterative, continual series of these frozen 3-space snapshots or observations, much like the frames of a motion picture film. We ourselves mentally add the "continuity" to provide "the sensed motion", but rigorously what is actually "observed" is not continuous, but is a vast continual series of those frozen 3-space snapshots.

We're getting directly at that atrocious diagram!

Each snapshot is an *effect*, not a cause, because it was the output of the observation process whereby a 4-space *causal* entity (non observed *a priori*) interacts with a previously observed frozen entity (say, a unit point charge at some point in space) to produce the observation (change or effect generated in that interacting observed charge) as the "next instantaneous observation".

The usual "representation" of a "3-space EM wave" propagating in 3-space is indeed atrocious, just as you stated! It is actually just an iterative succession of such instantaneously frozen snapshots in 3-space, one after the other. There is no such thing as that set of snapshots independently existing in spacetime, *prior to interaction with charge* in that series of interactions and observations, unless we wish to discard quantum mechanics and the laws of logic.

There is, however, a continual iterative stream of those observations — those frozen 3-space snapshots — that we interpret (erroneously) by recall from memory as the "EM wavefront propagating in 3-space". As you eloquently pointed out, that is not so, and it is atrocious. An observation, being an absolutely frozen entity, cannot "move through time" anyway, since it cannot persist, nor can it move. A change to an observable can only be another observable snapshot of the ongoing 4-space entity and action, which is then compared to the first snapshot and a difference noted {96} [See Figure 1-3].

In short, a great stream of "frozen effects" (frozen instant observations) does not constitute a "picture" of the ongoing 4-space action, but only a series of frozen 3-space intersections involving the interaction of that fixed observed (3-space) charge with the ongoing causal 4-space entity. If we add and integrate a series of 3-space pieces, we *do not* get a 4-space entity! Instead, we get a longer or bigger 3-space slice/piece, but one for which each piece of it only existed at a single point in time as a 3-space "slice" at that moment. That is precisely what is wrong with that horrible illustration.

But it is also "wrong" with electrodynamics itself! Electrodynamicists mistakenly conclude that the same *effect* "series of static 3-slices" — which they might call, e.g., the "field" in a case where the field is the subject — is the same as the 4-space continuous *causal* field prior to observation interaction with charge at all. Well, LLLT is definitely not LLL, nor is it n(LLL) where n is some large but finite number of 3-slices LLL.

So the field concept is dichotomously used in two contradictory manners in electrodynamics:

(1) it is considered to be in 4-space prior to the observing/detecting interaction with the observable (such as a unit point 3-space charge), and it is also considered a 3-space entity *after* that interaction. The dimensions of the two entities are not the same, and neither are the dynamics. The causal EM wave is dynamic and 4-spatial, the effect "3-space EM wave" (ugh!) is static and 3-spatial at each instant it is "observed". Assuming that the two are the same thing is a non sequitur.

In fact, it substitutes the *effect* for the *cause*, a rather gross violation of the causality principle itself.

(2) then an attempt at glossing over the illogic is used by the smooth statement that "the field (meaning that atrocious series of frozen 3-space snapshots) remains in the absence of charge, but the force goes to zero." Well, if the field is defined as a force field, it cannot be a "non-force field in the same observation!" Else, opposites are always identical.

So of what importance is all that?

It is of great importance. Let me show you one very startling thing that comes out of correcting this "biggest foundations non sequitur in physics, that of substituting the effect for the cause".

Consider a very special paper by E. T. Whittaker {97} in 1903. [I can send you a pdf file of the paper if you do not have it and are interested]. In this much-neglected paper, Whittaker decomposed the so-called "static" potential into a harmonic set of bidirectional longitudinal EM wavepairs, where each pair consists of a longitudinal 3-space wave (an *as-observed* wave) and its longitudinal phase conjugate wave (*considered unwittingly as having also interacted with charge, and therefore as being shifted into 3-space as an "as-observed" wave with inverse parity*).

Whittaker — as has everyone since him — unwittingly assumed the "iterative continual observation" interaction in there for the phase conjugate wave also. In so doing, he came up with two *effect* "waves" that are the outputs of the assumed observation process. Neither of these effect waves would be a wave in spacetime {98} at all, but the two in ensemble are an example of the same thing you objected to in your cogent commentary on that abominable "illustration".

In short, Whittaker invoked *observation* as a process with two effects and no cause, rather than with a cause and an effect, with the interaction with the observing/interacting unit point charge being assumed for both waves. Whittaker and everyone since seem to have made the same error in interpreting that seminal Whittaker decomposition. This misinterpretation of the decomposition has until now hidden one of the greatest secrets of all times in Nature's electrodynamics!

Let us correct the interpretation, and uproot that great secret to the light of day.

First, for observation to occur, one must have a cause acting upon the affected (observable) entity, and an effect (observable change) must be produced in, on, or of that affected (interacting) entity. One *must not* have two effects (two observables) and the affected entity (another observable and therefore another effect)! Again, assuming that one has three effects (observables) constituting the observation process is a logical non sequitur of first rank.

Let us now correct that logical non sequitur made by Whittaker and other physicists and electrodynamicists, and see where it leads us.

First, we go to particle physics, where broken symmetry {99} was discovered in the 1950s. Lee, e.g., received a Nobel Prize for his work in that area {75}. Lee also showed that any dipole is a broken 3-symmetry in its violent energy exchange with the active vacuum.

Well, a scalar potential is a dipolarity; always a potential is actually a difference between two potentials, so to speak. So the potential itself represents a broken 3-symmetry in an energy flow exchange with the active vacuum.

Let us further examine that interesting broken symmetry aspect. It means that the "static" potential is a process whereby energy is received from the vacuum in *one* form, not observable, and hence unusable, but is output in *observable* (usable) form. In short, the dipolarity or dipole receives and absorbs (QM view) virtual photons, integrates them into observable magnitude, and emits real, observable EM energy ("continual observation" snapshots of the latter being assumed).

In physics, all observation is 3-spatial, as is well known. And 3-space is the realm of the observed. The EM energy from the vacuum is not received in 3-spatial (observable) form, else there would be no broken 3-symmetry of the dipolarity.

We can also *experimentally* verify that there is no 3-space input of EM energy to the potential — e.g., to the potential between the ends of any source dipole, because we cannot measure any 3-space observable energy feeding the charges of the dipole. Instead, observable 3-space energy is continuously pouring out of the dipole.

Let us continue now with the notion of a real dipole of separated source charges, with our "scalar potential" between its ends, so we have something concrete in mind.

First, our instruments prove there is a continual emission of EM energy in all directions in 3-space (of the kind in that atrocious diagram; a "series of iterative film-snapshots"). That is the way it is conventionally represented, as if observed at each and every point successively in that 3-space, and along every radial.

But secondly, our instruments also prove there is no such observable *input* of EM energy in that same 3-space *to* the dipole. Bummer! That is precisely what has stopped electrodynamicists from solving that vexing problem of where all the energy pouring out of the source dipole is coming from, and how! They unconsciously assume that the only possible source of the EM energy input is in 3-space. That is totally wrong. Further, the conservation of energy law does not require that energy be conserved in 3-space! Instead, rigorously — if we are using a 4-space (i.e., spacetime) model — it requires that energy be conserved in 4-space. The assumption of the extra condition of 3-space energy conservation is an added and arbitrary extra condition.

Since electrodynamics has not resolved this "source dipole and its associated fields and potentials" problem, it has stopped work on the problem and implied that every dipole in the universe is a perpetual motion machine, freely creating all that EM energy it continuously pours out across space, in all directions, creating its associated fields and potentials. That implicit assumption, if true, of course destroys the conservation of energy law.

It isn't true, and the energy conservation law is alive and well!

We can experimentally prove that the source dipole *does* continuously pour out energy in all directions in 3-space, without ceasing, as follows:

In a gedanken experiment, we set instruments every 300 million meters or so, along a radial line from an origin in the lab. With the instruments and clocks synchronized, we suddenly form a dipole at the origin. One second later, the first instrument reads. A second later, the second instrument reads. And so on. But it is not a "passing pulse". Whatever reading the instrument makes as the forward edge of the energy flow reaches it, is then continuously maintained thereafter. This proves that the energy is poured out continuously and at the speed of light, and in any (and all) directions in 3-space, *and it continues to pour out at exactly the same rate so long as that dipole remains intact.*

Dipoles in the original matter of the universe have been pouring out EM energy in that fashion for some 14 billion years, and they have not "exhausted" their unobserved energy input source yet.

A newly formed simple dipole, e.g., in one year will have poured out energy into a spherical volume of space that is a light-year in radius. Wait another year, and that volume of space whose energy density has been changed will be two light years in radius. The dipoles in the original matter have filled the universe itself with that energy outpouring from them.

In other words, an incredible amount of EM energy has been and is being poured out into space from every dipole in the universe.[35] And electrodynamicists have had not the foggiest notion of where that mind-boggling amount of outpoured EM energy has come from.

Unless we wish to totally discard the conservation of energy law, we must have an equal input of energy *from outside 3-space*, going into every dipole continuously. In 4-space, that only leaves the fourth axis, along which and from which the input energy must move into the dipole.

And so it does. If we re-interpret that phase conjugate half set of the Whittaker decomposition, *before* observation has occurred, it then is a harmonic set of longitudinal EM waves moving in the time-dimension, *into* the source dipole (parity is not inversed because no interaction with charge has occurred to shift it into 3-space.

So voila! We have strangely (but quite rigorously) solved what Sen {100} referred to in this manner: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."*

We express our reasoning as follows: (i) we know from particle physics that the source dipole[36] is a broken 3-symmetry in its flux exchange with the vacuum. This means that we know the energy is received from the vacuum in an *unobservable* form, absorbed by the dipole, and emitted as *observable* EM energy. It remains to translate that into classical electrodynamics rather than quantum physics {101}. (ii) The *time* domain is the only domain outside 3-space, in the standard 4-space model. (iii) Anything in the time domain exclusively, is indeed nonobservable, since the $\in/\in t$ observation operator destroys time and all its internal structure and dynamics whenever observation occurs. (iv) Looking for a "3-space" input is looking for an "observable" EM energy input, which would in fact disagree with the known broken 3-symmetry of the source dipole.

---

[35] Later we will see that, in a time-forward situation, the negative charge pours out positive EM energy while the positive charge may be said to pour out negative EM energy. Charge conservation then implies that an energy balance is maintained overall.

[36] Because of the broken symmetry of the opposite charges on its opposing ends.

Let me digress now to speak of fundamental units. As is well known, the fundamental units one chooses to make his physics model represent an arbitrary choice. One can, if one wishes, make a perfectly valid (but mentally nightmarish!) physics using only a single fundamental unit. In that case, all other "fundamental units" in the present model become functions of that single fundamental unit.

Suppose, e.g., that we make the joule our only fundamental unit. Then "mass" becomes totally a function of energy — and we have no heartburn with that one since the Einstein revolution and the nuclear age. But then "time" also becomes totally a function of energy — and that surprises us, because we have unconsciously been taught (erroneously) that "time is a flowing river down which a mass drifts like a drifting boat". That is not the nature of time at all; the totality of the photon interactions with a mass create that mass's "motion through time". I can later explain that to you also, if you wish, but let us pass it for now.

It turns out that time is spatial energy compacted by the factor c^2, so it has the same energy density as mass, but in the time axis instead of 3-space. Intuitively, if we are interested in cause and effect, the energy of the cause (time) should be equal to the energy of the effect (in this case, mass). After one reflects a moment, one also sees that "time energy" is required to "drive" a mass through time, just as "spatial" energy is required to "drive" a mass through space. The notion that mass in an inertial frame moves through "empty space with no reaction" is false; it continually moves through the energetics of spacetime, and interacts continuously with it.

Anyway, from still another viewpoint there is no magic in EM energy currents moving in the time domain! Quantum field theory already recognizes multiple polarizations of the photon, including transverse, longitudinal, and time-polarized. Thus, it implies a time-polarized EM wave also. A longitudinal EM wave *moving in the time domain* oscillates along its line of travel. That is indeed oscillating its energy density in the time domain, so that it is "time-polarized". Hence it is a proper time-polarized EM wave, and appears to "gallop" {102} or vary its speed periodically.

The solution to the "source dipole" problem is that the phase conjugate wave half-set of the Whittaker decomposition, when reinterpreted, is the *incoming* EM energy in the time domain, continuously input to the charges of the dipole. The charges interact in the imaginary plane (the time domain), and absorb the time-energy, then transduce it into 3-space, and emit it as Whittaker's set of real observable longitudinal EM waves in all directions in 3-space (as observed). The time-energy waves that are input to the dipole cannot be observed, since observation destroys time and its constituents.

All 3-spatial EM energy comes from the time domain! Broken 3-symmetry of the source dipole immediately releases our arbitrary additional restriction on nature's energy conservation law — i.e., our insistence that the energy input for conservation must be input in 3-space. By removing these arbitrary "shackles" from nature's feet, we allow nature to joyously resume her much-preferred special 4-symmetry in EM energy flow: the circulation of EM energy from the time domain into 3-space, and outpouring of that energy in all directions in 3-space, at the speed of light {103}.

Now let us solve the source-charge problem as well. That one is now easy to resolve.

We know from quantum mechanics that any "isolated" observable charge is actually clustered around by virtual charges of opposite sign. So we simply take one of these clustering opposite charges while it exists, and a differential piece of the observable charge, and we have a composite dipole. Hence the "isolated charge" may be represented as a grouping of such momentary dipoles, each (while it exists) with a scalar potential between its poles, and hence each subject to the same decomposition and reinterpretation we have done.

That is why the source charge can "sit there" and pour out EM energy in 3-space (as observed) continuously, and indefinitely. It is also continuously absorbing EM energy from the time domain. As a set of composite dipoles, it is a set of broken 3-symmetries in EM energy flow. Hence it exhibits nature's preferred 4-symmetry in energy flow, between the time-domain and 3-space.

Note also that, to the 3-space observer, this 4-symmetry is purely negentropic. It is a continuous and ongoing (and expanding) reordering of the vacuum energy, in the form of the reinterpreted Whittaker decomposition.

We can easily engineer negentropy! Just make a little dipole, and nature happily starts pouring out energy and reordering part of the vacuum, with that reordering spreading at the speed of light, steadily increasing all the while.

Reinterpreting Whittaker's marvelous decomposition tells us many things: (i) the source charge or dipole does indeed continuously receive and absorb its input energy from the time domain, (ii) it continuously transduces its absorbed energy to real 3-space energy (as observed!), and (iii) it continually outputs its absorbed and transduced time-energy as 3-space energy emitted in all directions in 3-space (as continually observed).

So the terrible foundations problem of the source charge and source dipole is resolved, and the energy conservation law is maintained.

We have also uncovered a most important thing: Given a little broken 3-symmetry in EM energy flow, t-symmetry in EM energy flow is also broken. A more primary 4-symmetry {104} between time-energy flow and 3-space energy flow emerges *automatically*. In short, every potential, every dipolarity, and every charge is such a broken 3-symmetry and an example of the new, preferred 4-symmetry of EM flow, with input flow in the time-domain (nonobservable!) and output flow (as continually observed in a series of 3-snapshots) in all directions in 3-space.

This also resolves the logical cause and effect problem for observation.

This 4-symmetry is a purely negentropic process, once the dipole is established! Well, we should have known that from the gauge freedom axiom in gauge field theory anyway! That axiom says that we can change the potentials anytime, freely and at will. In electrodynamics, that means we can change the potentials at will, anytime we want to. But that means *we can freely change the potential energy of a Maxwellian system at will, whenever we wish*. It costs nothing — at least in theory; in the real world we have to pay a little for switching — to suddenly potentialize an EM system, prior to the time the Drude electrons relax and current flow begins.

I have applied this great new 4-symmetry in EM energy flow, that is freely evoked and persists indefinitely after one pays a little to make the dipole and thus produce a little broken 3-symmetry, to produce EM energy freely from the vacuum. Without going into it, I refer you to our forthcoming papers {105}. Indeed, we can use this negentropic process to extract as much EM energy from the vacuum as we wish, anywhere, anytime, for peanuts. Let us now move to discuss that area.

First, we point out another astounding result that comes from resolving this "source charge and source dipole" foundations problem.

Generators do not use any of the shaft energy input to them — even in transduced form — to power their external circuits! A generator itself does not add a single watt to the power line, and neither does a battery's chemical energy dissipation add a single watt to its attached circuit.

Neither does burning all those hydrocarbons, consuming all those nuclear fuel rods, building all those dams, etc. to provide our conventional power systems and grid. None of that enormous destruction of the biosphere and pollution of it has ever directly produced one watt on the electrical power line.

Just follow the energy flow, in light of what we now know. Here's how it works.

Typically, we burn some fuel to boil water and make steam, and use the steam to power a steam turbine, which forcibly rotates the shaft of the generator, thereby inputting mechanical shaft energy into the generator. So far, so good. That took care of all the hydrocarbon burning and fuel rod consuming, extracting and transporting the oil, etc.

This input of the shaft energy forcibly rotates the rotor against internal resistance, forming an internal magnetic field. Assuming a 100% efficient generator with no internal losses whatsoever, this means that the *mechanical shaft energy input* has now been totally transduced into *internal magnetic field energy*.

So what does that magnetic field energy *do*? It is totally dissipated upon the internal charges of the generator, performing work on them and forcing the positive charges in one direction and the negative charges in the other direction. That dissipation of the energy in the internal magnetic field *forms a source dipole inside the generator, connected to the terminals*.

And that is all the generator does. Period. *None* of the energy transduced from that shaft input, went roaring out of the terminals and down through space outside the conductors of the power line. Not a single watt. So let us turn to particle physics to find out what happens next, because *it does not yet appear in the electrodynamics model, even though proven.*

The internal source dipole, *once formed*, is a great broken symmetry in the vacuum flux, as we discussed and as is well known in particle physics. But the proven and well-known vacuum interaction with the generator and the dipole charges is not even modeled in the classical EM theory used by the scientists and engineers to build electrical power systems — much less a *broken symmetry* in that active exchange! That's atrocious, since it's been proven in particle physics for nearly a half century, Nobel prizes awarded, etc. But the hoary old 137-year old Maxwell-Heaviside model, further curtailed by Lorentz symmetrical regauging, does not incorporate what has already been proven in physics. *It does not accurately model the situation as it is and as it is known to be.*

Let us continue.

Once that dipole is formed, it extracts *enormous* EM energy from the vacuum, and pours it out of the terminals of the generator, filling space surrounding those conductors of the attached external power line. It is a huge EM energy flow, trillions of times larger than what we account after Lorentz taught us to disregard almost all of it.[37]

Only the tiny little bit of that EM energy flow in space along and surrounding the conductors — the little "boundary layer" that skims down the surface of the conductors — will strike the surface charges in the conductors and get diverged into them to potentialize the Drude electrons and "power the power line and its circuits and loads". *All the rest of the giant EM energy flow in space surrounding the conductors, and generally parallel to them, misses that power line entirely and is just wasted.*

Check the original papers by Poynting {4a, 4b} and by Heaviside {5a, 5b, 5c}, who independently discovered the flow of EM energy through space (as if continually observed!) in the 1880s, after Maxwell was already dead. I can furnish the citations required. From the beginning, Poynting only considered that component of the energy flow that actually enters the circuit. He considered only the "boundary layer" right on the conductor surfaces, so to speak.

Heaviside considered that component that enters the circuit, and also uncovered and recognized the *gigantic* component in the surrounding space that *does not* enter the circuit but misses it entirely and is wasted. [Added] Here is Heaviside's {106} own statement:

> *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire… . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from*

---

[37] However, contrary to what is in the textbooks, **E** Δ **H** energy flow is not in observable EM field form **E** and **H** until it interacts with charge. Instead, the flow is in the form of virtual energy currents in the vacuum. We point out that the fields **E** and **H** utilized in the Poynting energy flow theory are effects and not causes. They are rigorously defined as **E** and **H** only after the causal 4-fields have interacted with charge. E.g., **E** is "defined" as force per unit point massive charge — i.e., as the force created by interaction of the "vacuum field" (curvature of spacetime) with a unit point charge. No unit point charge interaction, no force or force field intensity per interacting unit point charge.

> *the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire."* [End of Added].

But Heaviside had absolutely no explanation for the enormous and startling magnitude of this energy flow that "misses the surface charges of the conductors and is wasted". You can see an elementary illustration of the "point intensity" of this Poynting diverged energy flow component in Kraus {107}. Kraus's figure 12-59, p. 576 shows a good drawing of the *Poynting* component *being withdrawn* from the total EM energy flow filling all space around the conductors [see our Figure 1-1 in the present chapter].[38] Most of that available energy flow is not intercepted and thus not diverged into the circuit to power it, but just "wasted." The *remaining* huge component discovered by Heaviside is not shown on Kraus's diagram. Kraus's numbers on the contours represent the amount of power (watts per sq meter) being withdrawn from each contour, by the very limited axial movement of the electrons and the very ends of their associated fields into the wire.[39]

---

[38] The energy flow in space near the surface of the wires strikes the surface electrons and their fields, potentializing them and creating a force field with respect to inner electrons. This drives the surface electrons axially into the wire, since they can only move down the wire with the drift velocity, nominally on the order of a few inches per hour. The short inwards drawing of those small portions of the fields and of the increased potentials attached to the charges that move into the wires, interacts with the internal charges in the interior of the wire, powering the Drude electrons throughout the conductor and the circuit. For a discussion of the importance of the surface charges, see J. D. Jackson, "Surface charges on circuit wires and resistors play three roles," *Am. J. Phys.*, 64(7), July 1996, p. 855-870.

[39] Obviously the increased fields on the withdrawing surface charges reach toward infinity radially outward from the wire. Thus most of this increased field on a withdrawing electron remains outside the wire, radially reaching toward infinity, since the electron and its field cannot be withdrawn further than the diameter of the wire. As the surface charges and their fields change in Jackson's discussion, obviously those distant parts of the field remain outside the conductors. Hence there remains an extensive field and field energy outside the conductors and the circuits

So Heaviside {106} spoke cautiously of the *angles* and *relative directions* of the flow components; he did not wish to be scientifically destroyed for pointing out such an inexplicably large EM energy flow, far larger than the known energy input to the generator. There was then no such thing yet discovered as the electron, the atom, the nucleus, special relativity, general relativity, quantum mechanics, the active vacuum, etc. Poynting {4a, 4b} never even thought of the "nondiverged component that misses the circuit", nor did he consider it.

Lorentz, however, understood the extra Heaviside component and its vast magnitude, but he also could not explain it in any fashion. Even the great Lorentz could not risk publishing or advocating such an enormous energy flow, lest he be called a "perpetual motion nut" and destroyed. So he reasoned that, since that stupendous nondiverged energy flow component misses the circuit and powers nothing, it has "no physical significance" (his words).

Jackson in his famous 1975 *Classical Electrodynamics*, 2nd edition, uses essentially the same phrase {108a} as did Lorentz. And so do most other electrodynamicists.

So Lorentz {109} originated a little trick of integrating the energy flow vector itself around a closed surface surrounding any volume element of interest. Obviously, that zeros all *nondiverging* EM energy flows — including precisely that inexplicable and enormous Heaviside "dark energy" component. The integration trick does retain that small, *diverged* Poynting component that enters the circuit. In addition, our instruments measure energy dissipation from the circuit, and the energy has to *enter* the circuit to *be dissipated* from it. So our instruments and their measurements will indeed agree with the Poynting energy flow component. Lorentz thus arbitrarily discarded accountability of trillions of times as much EM energy flow as was retained and accounted.

In a later book by Lorentz, one can see that little trick {109} that is still used by electrodynamicists {110}.

---

that is never drawn into the wire. Additional separate receiver circuits with resistive loads, adroitly placed in this "externally remaining" field energy flow, can be made to intercept additional EM energy and will perform additional work in those external loads in the secondary circuits, separate from powering the loads in the primary circuit.

The Lorentz trick does not cancel the *actual* flow of the Heaviside dark energy component around every circuit! It just drops it from any accountability.

Indeed, we do precisely a similar thing for the "field" and the "potential". There is not a single text in the U.S. that calculates the magnitude of the *field itself*, prior to point interaction. Instead, we are taught to calculate the reaction of that field at a point in it, with a unit point static charge at that point. In short, we calculate *what is actually locally diverged from the field or potential by that little point static unit charge*, and call it "the magnitude of the field". At best, it is indicative of the field intensity at a point, because we have prescribed the magnitude of the static point charge's *reaction cross section* with the field, *not* the overall magnitude of the entire field itself. Another major non sequitur! By identifying the "field" as "that which is diverged from it", we gravely err. We do the same for the potential, again using its reaction cross section for a unit point charge at a point in the potential.

There is hardly a living electrodynamicist, it seems, who has calculated the magnitude of the field itself, or of the potential itself! All calculate the respective reaction cross section (and the *static* reaction cross section at that!) and erroneously call that the "magnitude of the field" or the "magnitude of the potential". It is no such thing. No thing filling all space is identical to a little something diverted from one point in itself, else we discard all logic.

Now we can return to our source dipole and its extraction of an *enormous* energy flow from the vacuum, once we account for the long-neglected (for more than a century) Heaviside dark (nondiverged, unaccounted) energy flow component. That is the essence of one of my papers {111}.

The easiest thing in the world to do is to extract enormous usable EM energy flow from the vacuum, *from the time domain*. Every circuit and every electrical power system already does it, and every electrical power system is powered by vacuum energy, *not* by burning all that coal, oil, etc. None of that does anything to actually power the circuit. It only makes dipoles.

If you will check the characteristics of the ubiquitous closed current loop circuit, you will discover a diabolical thing: that closed current loop circuit forces all the spent (depotentialized) electrons from the ground return line back through the source dipole (back through the back emf). It is easily shown that precisely half the energy *collected* in the circuit from that feeble Poynting energy flow component is then used to perform work on those dipole charges and scatter them, thereby destroying the source dipole and abruptly shutting off all energy extraction from the vacuum. The other half of the collected Poynting energy is dissipated in the external loads and losses. That means that half the collected Poynting energy is used to kill the source dipole, and less than half is used to power the loads {112}.

Well, we must input at least as much energy (assuming a 100% efficient generator with no losses) to the shaft of the generator to *restore* the dipole, as was used to *destroy* it. This means we shall always have to input *more* energy to the generator shaft, than we get dissipated in the loads.

That is precisely what is responsible for our ubiquitous COP<1.0 systems. *COP* is Coefficient of Performance, and is the average load power output in the external circuit divided by the average shaft power we pay to input to the generator.

COP < 1.0 comes from (i) the ubiquitous usage of the closed-current-loop circuit, and (ii) the prevailing notion that electrical power systems far from equilibrium in their energetic exchange with the active vacuum are impossible to build. In fact, every electrical power system is *already* just such a COP > 1.0 system, as far as the energy flow out of the generator or battery — compared to the shaft input energy to the generator or battery — is concerned.

As is well-known in the thermodynamics of open systems far from equilibrium with their active environment (in this case, the active vacuum), such a system is permitted to: (1) self-organize, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself must input to the system (the excess energy being freely received from the active environment, in this case the active vacuum), (4) power itself and its loads simultaneously (all the energy being freely received from the active environment, in this case the active vacuum), and (5) exhibit negentropy.

But by designing all our systems so that they use more of their collected energy from the vacuum to kill their source dipoles than they use to power their loads, we have foolishly wasted the planet's resources, vastly overcharged the consumer, artificially created a great energy crisis, fomented wars for precious oil and other energy resources, polluted the planet, enhanced global warming, and strangled species.

In short, we pay the power company to have a giant wrestling match inside its generators and lose! And we pay our electrical engineers to keep designing and building such asinine systems!

This sad state of affairs is what is now upon us as a great and increasing energy crisis, polluting and destroying the biosphere, etc. and threatening to eventually collapse the world economy.

It is astounding that, since the basis for the above has been in physics for nearly a century (Whittaker decomposition) and for nearly a half century (broken 3-symmetry of the source dipole, as well as the active vacuum), the hoary old classical EM model has not been updated to incorporate what has already been proven in particle physics. Such is inexplicable and unconscionable.

It is also astounding that no electrical engineer realizes that energy extracted from the vacuum powers every electrical power system, and few if any professors are aware of it either.

This is where your cogent realization of the terrible non sequitur in that atrocious "wave in 3-space" standard diagram leads.

Our AIAS (Alpha Foundation's Institute for Advanced Study) advanced theorists now have a series of papers published in the hard literature pointing out the feasibility of extracting EM energy from the vacuum {113a, 113b}. About 20 other AIAS papers along such lines or related have been submitted to the various journals and are in the referee process.

I just wanted to contact you informally and, for your personal information, show you how perceptive and correct your objection to that atrocious diagram is. You have lifted the corner of the veil on electrodynamics' confusion between effect and cause, with effect being widely used as the cause. The entire notion of a "separate force" acting on a "separate mass" in mechanics is also a non sequitur. If we define force as $F \sum \in/\in t(mv)$, then we see that mass is a *component* of force! So here is another major and uncorrected non sequitur. This one was also largely responsible for the dichotomy of the field concept, where the "same field" is used in both a force (mass-containing) manner and a force-free (without mass) manner, as if the two were the same!

I very much enjoyed your editorial, and would urge you to publish additional material along that same vein. You are striking at the very heart of the problem, and every bit of insight and change in those terrible non sequiturs will result in enormous progress in electrodynamics and physics.

Sincerely,

Tom Bearden, Ph.D.

Lt. Col., U.S. Army (Retired)
Director, Association of Distinguished American Scientists
Fellow Emeritus, Alpha Foundation's Institute for Advanced Study

----

Chapter 2

Background and Concepts

"[Hypotheses made without realizing that they are being made] ...are what Poincare has called "unconscious" or "natural" hypotheses — a type which one hardly ever challenges, for it seems too unlikely that we could make progress without them. Nevertheless it should be the endeavor of the physicist always to drag them out into the light of day, so that it may be perfectly clear what we are actually doing." [Lindsay and Margenau] {114}

"Equilibrium states are the only ones that are capable of explicit analysis in thermodynamics... " [Lindsay and Margenau] {115}

"Non-equilibrium conditions cannot be specified by variables of state, and their entropy cannot be computed. ...the condition of equilibrium is the condition of maximum entropy. " [Lindsay and Margenau] {116}

2.1 Comments on the Foundations of Classical Electrodynamics

2.1.1 Faraday, Maxwell, Quaternions, and Heaviside

For our purposes, we start with the important experiments of Faraday {117}, which produced experimental integration of electricity and magnetics {118}. Faraday's own work contained very little mathematics, and so he himself did not capture his own seminal experimental work in adequate theoretical fashion. That task was undertaken by James Clerk Maxwell {119a, 119b}, who deliberately set about to capture Faraday's results in a proper mathematical theory, using quaternion-like and vector expressions. Of the quaternion algebra, Maxwell had this to say {120}:

> *"...the virtue of the 4nions lies not so much as yet in solving hard questions as in enabling us to see the meaning of the question and its solutions. "*

Quaternion algebra {121a, 121b} is an advanced algebra created by William Rowan Hamilton, brilliant Irish mathematician. Hamilton was widely regarded as a "new Newton" by his peers. Hamilton coined the word "vector", wrote two books on quaternions, and hoped the books would find extensive use in physics. Vectors and tensors later came to suffice for most problems of that day, and so quaternions faded to the background.

Hamilton also grasped the importance of a science of time, and realized that such a science was possible. E.g., in 1837 Hamilton stated:

> *"The notion of time may he unfolded into an independent pure science... a science of pure time is possible."*

In that statement, Hamilton anticipated that a physics model can be comprised using only a single fundamental unit, and that unit could be the second.

In the early 1890s, quaternions in electrodynamics were discarded in a short "debate", mostly in the journal *Nature,* and vector electromagnetics was adopted. So Maxwell's 20 quaternion equations {20} in some 20 unknowns contain a wealth of electrodynamics {122} that is not included in Heaviside's {123} later vector curtailment after Maxwell's death, or in the modern tensor formulation. As an example, Nikola Tesla's patented circuits cannot be properly understood in either a vector EM or tensor EM analysis. However, very novel functions performed in them can be seen in a quaternion EM analysis, as has been clearly shown by Barrett {124}. Any electrodynamicist, who has looked at Tesla's work only in terms of tensor or vector classical electrodynamics, has no concept of what Tesla was actually doing. Since that is precisely how most academics have examined Tesla's work, they have had little idea of the functions he was actually able to accomplish in circuits — functions that U.S. electrical engineering departments still cannot perform.

Along with other electrodynamicists at the time, Faraday and Maxwell believed in a material ether — a thin material fluid filling all space. Hence there was thought to be no point in the universe that was devoid of mass. For that reason, the EM field conceived in mass and the EM field conceived in "space" — actually, in the "luminiferous ether" thought to fill all space — were identically force fields and material entities, at least in the minds of the electrodynamicists at the time. The equations of Maxwell, the later truncation by Heaviside and others, and the final truncation by Lorentz are all still material fluid flow models. They still assume the material luminiferous ether, more than a century after its falsification.

See again Figure 1-3 in Chapter 1. Contrary to the conventional view, force is not a fundamental causative entity, but is a system of two coupled entities. It involves a nonobservable, dynamic, causative 4-space entity (which may be modeled as a curvature of spacetime) interacting with a previous "frozen" observable (a 3-space mass). The output of this interaction is another observable (i.e., a change to the previous observable or an iterative output of its replica). Observables do not continuously exist, but continually recur as continual outputs of the iterative observation process (photon interaction).

Mass is a *component* of force, as is apparent if we define force F as $F \equiv d/dt(mv)$, which gives $F = m(dv/dt) + v(dm/dt)$. Both terms on the right of the expansion contain mass; hence force F contains mass as a *component.* Both terms on the right also contain nonmass terms; hence force F also contains nonmass coupled to and interacting with mass. So there can be a force generated by a nonmassive non-force entity in spacetime interacting with a mass, and the force only exists when that interaction is ongoing. The ether was erroneously considered to be a material system, containing mass, and hence capable of supporting a force. Maxwell modeled his fields (in what we know today to be mass-free space) as material force fields in the material ether.

In the 1880s, the Michelson-Morley experiments {125a-125d} destroyed the material ether. However, the fundamental Maxwell-Heaviside equations *assuming* those material force fields in the vacuum — which depended on the notion of a material ether and are a function of it — have never been altered, even to this day. Instead, electrodynamicists just assume the force "somehow" goes to zero in the field in space, but the same field remains. The equations still assume the same old material ether.

This is a non sequitur; a *force* field cannot be the same as a force-free *non-force* field. As an example, the EM force field in observable matter is an output of observation. It is therefore a *3-space effect containing mass.* The force-free field in mass-free spacetime (4-space) is a causal 4-space component of massless spacetime. *A priori,* it exists prior to interacting with the charged mass. It contains no observable mass and hence no force since mass is a necessary component of force. It independently exists without the presence of the charged mass at all.

In short, the field in massless spacetime is a specialized curvature (changed region) of that spacetime. *Spacetime consists of spacetime,* to slate a tautology. Once one fixes the concept of the medium, then any changes to the medium can only consist of "medium-stuff. So any "changes" in spacetime can only be changes of 4-space itself, in the Einstein general relativity sense. This area is handled quite well by Sachs' unified field theory {126a, 126b}, extending and completing the work of Einstein.

Sachs' epochal theory provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. Further, Evans' extension of electrodynamics into a higher 0(3) group symmetry turns out to be demonstrably a special subset of Sachs' electrodynamics.[40] Hence, for the first time, an *electromagnetically* engineerable unified field theory is available. Much of general relativity, quantum mechanics, etc. can be engineered directly by Evans' electrodynamics subset of Sachs' overall unified field theory. I am much indebted to Evans {127} for the following statement with respect to O(3) electrodynamics:

> *"In 1992 it was shown (Phvsica B. 192, 227, 237 (1992)] that there exists a longitudinal component of free space electromagnetism, a component which is phaseless and propagates with the transverse components. Later this was developed into a Yang-Mills theory of electromagnetism with O(3) Lagrangian symmetry. This theory is homomorphic with Barrett's SU(2) electrodynamics and has far reaching implications in field theory in general. Recently it has been recognized to be a sub theory of the Sachs theory of electromagnetism, based on the irreducible representations of the Einstein group of general relativity.[41] The Sachs theory produces a non-Abelian structure for the electromagnetic field tensor. The O(3) electromagnetism also has implications for the* ***potential*** *ability of extracting energy from the vacuum,* ***and its*** *topological implications are currently being* ***investigated*** *by Ranada. The O(3) electromagnetism has* ***been tested*** *extensively against empirical data, and succeeds* ***in*** *describing interferometric effects and*

---

[40] E.g., see M. W. Evans, "The Link Between the Sachs and 0(3) Theories of Electrodynamics," in M. W. Evans (Ed.), *Modern Nonlinear Optics,* Second Edition, Wiley, 2001, part 2, p. 469-494.

[41] Evans et al, "Derivation of 0(3) Electrodynamics from the Irreducible Representations of the Einstein Group," *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-

> *physical optical effects where the conventional Maxwell-Heaviside theory fails. Implicit in both the O(3) and Sachs theories of electromagnetism is the ability to extract electromagnetic energy from curved space-time. Working devices based on this ability would help solve the current energy crisis."*

2.1.2 Mass Is a Component of Force

The dichotomy of using the field concept in two mutually contradictory fashions as something massless and also as something massive — still unresolved in electrodynamics — is clearly summarized by Jackson's important statement {128} of the prevailing view:

> *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."*

In trying to rationalize the "massive field versus massless field" dichotomy, Jackson {129} also states:

> *"...the thing that eventually gets measured is a force..."*
> *"At the moment, the electric field can be defined as the force per unit charge acting at a given point."*

Actually, field is not force, so cannot be force per unit charge. Field *produces* force when it interacts upon charge. *Causing* something as an effect, but then *being* that same effect identically, is a confusion of cause and effect and a non sequitur.

We point out that spacetime is charged, but with *observably massless* (virtual) charge. So the field in spacetime is observably massless and therefore observably force-free. That massless field then interacts with an observable massive charge in the "observable" world, to create an observable force and an observable forcefield. In space, from a quantum mechanical view we may consider the field to have *virtual* force x time, since the field may be regarded as a curvature of spacetime acting upon the *virtual* charged masses in the ever-bubbling *virtual particle fluctuations* of the vacuum permitted by the Heisenberg uncertainty principle. Nonetheless, that is only the "potentiality" for an observable force, as pointed out by Feynman and Wheeler, should the field — including its interactions with the virtual charges of the vacuum to produce forces and accelerations on those virtual particles — interact upon an observable charged mass.

Obviously, the force-free field as a product of spacetime curvature and massless (virtual) charge, and the force field as a product of *that* observably force-free field but virtual force field (in the vacuum) with observable charged mass, cannot be the same thing. Assuming they are the same assumes an oxymoron, and confuses cause and effect. This illustrates the continuing dreadful foundations difficulties propagated in classical electrodynamics.

To resolve these EM foundations difficulties so long maintained and cherished as dogma, the scientific community at the highest priority should (1) fund extensive and rapid work by our very best theoreticians and foundations physicists to heal these century-old festering sores, (2) extend electrodynamics back to a higher group symmetry theory, as in O(3) EM theory for example, and (3) provide a rigorous dictionary of science foundations concepts and terms which does contain valid definitions and full discussions — clearing misdefinitions and disinformation now notable in the electrodynamics literature.

Nobelist Feynman, co-founder of quantum electrodynamics, was careful to state the notion of the field in mass-free space very differently. Quoting Feynman {130}:

> *"...the existence of the positive charge, in some sense, distorts, or creates a "condition" in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a force is called an electric field."*

Feynman's imagery clearly invokes a "distortion condition in spacetime" {131}, eventually coupled to a charged mass and interacting with it — and *then* one has a force and a force field. He erred in still somewhat making the force separate from the charged mass, in the phrase "it feels a force" — which ignores the fact that the mass of the interacting charge is a *component* of the force, not separate from it and not "feeling it as a separate external thing". He pointed out that the field in spacetime represents only the *potential for* a force field, *if a* charge should be placed there. And he points out that the field is a "condition in space" — in the full sense, a condition in spacetime.

We comment that, in a sense, spacetime and vacuum may be considered one and the same. Vacuum contains many fluctuations and virtual state dynamics, all of which are curvatures of spacetime or produce curvatures of spacetime. Vacuum has energy density, and is therefore a potential. This energy density has dynamics, so the ambient vacuum potential has dynamics, and so therefore does spacetime {132}.

However, a dynamic of the potential is precisely what a field *is.* It follows that dynamics of spacetime are what fields are. So fields (as force-free entities in spacetime) may be taken more accurately as the dynamics of the vacuum-spacetime, or as dynamics of the ambient vacuum-spacetime potential, or as dynamics of the ambient curvature of spacetime.

Again quoting Feynman {133}:

> *"We may think of E(x, y, z, t) and B(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the charge there did not disturb the positions or motion of all the other charges responsible for the fields."*

Here Feynman very clearly shows that E and B involve functions of spacetime *(x, y, z, t)* only, prior to interaction, but do not physically exist until interaction with charge has occurred. But he has not quite yet completely eliminated the "force field" as a separate force acting on the charged matter independently. This centuries-old fundamental mistake in mechanics — that force is independent of and external to mass — has simply been propagated into electrodynamics, without correction. Physicists such as Feynman and Wheeler have been trying very hard to correct it.

The observed/measured force field is an output of observation, and observation is a *d/dt* operator imposed upon spacetime LLLT. All observables are 3-spatial, and any observation is an instantly frozen "3-space snapshot" at a single point in time. A single observation has no temporal existence *a priori,* since it is only an instantaneous 3-space snapshot existing at one moment only. Hence an observable — being 3-spatial and frozen — does not persist *as such* in time. Indeed, the next instant after an observation, that particular snapshot does not exist any more. [Refer again to Figure 1-3 in Chapter 1].

2.1.3 Mechanism for the Flow of Time and Related Matters

It is well recognized that the nature of time is itself an unresolved question. For our purposes, we are interested in the energetics[42] approach. So we apply that approach to arrive at a mechanism generating the "flow of an object through time". First we will choose our fundamental units.

2.1.3.1 *Getting Creative with Fundamental Units*

In physics, the choice of fundamental units one chooses for one's model is arbitrary. Usually mass, length, time, and charge are used, but a perfectly valid model can be generated using only a single fundamental unit.

This is well known to modern physicists and leading electrodynamicists, but sometimes comes as a shock to electrical engineers! Jackson {134} expresses it very succinctly: Quoting:

> *"For example, theoretical physicists active in relativistic quantum field theory and the theory of elementary particles find it convenient to* ***choose*** *the universal constants such as Planck's quantum of action and the velocity of light in vacuum to be dimensionless and of* ***unit magnitude.*** *The resulting system of units (called 'natural' units) has only* ***one*** *basic unit, customarily chosen to be length. All quantities, whether length or time or force or energy, etc., are expressed in terms of this one unit and have dimensions which are powers of its dimension. There is nothing contrived or less fundamental about such a system than one involving the meter, the kilogram, and the second as basic units. It is merely a matter of convenience."*

So let us consider what happens if we use the joule as the single fundamental unit for a model of physics. Then each of the entities "mass", "length", "time," and "charge" will become totally a function of energy. Since the dawn of relativity and the nuclear age, the notion that "mass is energy" is no problem, and everyone is familiar with Einstein's $E = mc^2$. Solving for m in that expression, one realizes that spatial energy E has been compressed by the factor c^2, to yield what is called "mass m". More rigorously, compressing the 3-space energy by the factor c , *and then leaving it in 3-space,* produces what we call "mass". However, if we remove it from 3-space and place it over on the fourth Minkowski axis *ict*, it becomes what we call "time" because *t* is the only variable in *ict*. Hence time has or can have similar energy density to mass. Specifically, 1 sec => $9x10^{16}$ joules of spatial EM energy, when transduced (decompressed) into spatial energy by a transducing charge.

---

[42] **Energetics** is an **old term from the early birthing** days of electrodynamics. Russian scientists resurrected **the term to cover their use of an** expanded electrodynamics in a unified field-engineering **manner, particularly in new** superweapons developed and tested after WW II and deployed in 1963 and subsequently.

In our new model using only the joule for our fundamental unit, we can also legitimately state that "time is energy", and be rigorously accurate. In that case, we must see what happens to 3-space EM energy when it is shifted to the time domain.

First, we take "EM energy in 3-space" as EM wave energy, in the usual transverse EM wave model (see Figure 1-4 of Chapter 1) of a field entity. We choose, say, the vector E oscillating in the x-direction, and another field entity (say, the vector B) oscillating in the y-direction. If we then rotate each of these vectors into the time axis, then that is an orthogonal rotation for each, which is what the velocity *c* actually is. So we have to do two orthogonal rotations, and the two resulting *c's* multiply to give c^2. With these two orthogonal rotations into the time axis, we have completely rotated the EM field wave energy vectors into the time axis. We have also compressed that EM field energy by c^2.

Hence "time as energy" is just EM spatial energy compressed by the factor c^2, but rotated into the time-axis as "time" rather than remaining in 3-space as "mass". Again, by this second analogy time has the same energy density as mass, but the highly compressed spatial energy has been rotated into the time domain to "produce time" rather than remaining in 3-space to "produce mass" {135}.

This sheds new light upon the scalar (time-polarized) photon of quantum field theory. Note that, prior to compression of the wave energy into time, the magnitude of the electric field *E* in the EM wave was oscillating lengthwise along its x-direction, which means that, with respect to x, it was a longitudinal EM wave. A similar consideration exists for the oscillation of the magnitude of *B* in the y-axis. Along the time-axis, we also have a "time-energy" vector having magnitude, and the magnitude of that time-energy can only vary along the time axis. Oscillating the time energy produces a longitudinal EM wave in the time dimension (on the fourth axis). What is actually varying is the time-density itself. This is a *time-polarized* or *scalar* EM wave, where by use of the term "scalar" we mean that it has no vector component in 3-space. In fact, this variability of time density produces a "galloping motion" in the speed of light itself, which has been recognized {136}.

We have uncovered a direct connection between the time-polarized EM wave (which can only oscillate its time-energy magnitude longitudinally in the time-domain) and the longitudinal EM wave in space. Such waves can only be observed when coupled or paired with an interaction existing between them, very similar to what Mandl and Shaw {167, 168} argue for the time-polarized and longitudinal photons of quantum field theory.

Again see Figure 1-3 in Chapter 1. Any observable is an instantaneous, frozen 3-spatial snapshot (or 3-intersection) of a 4-dimensional ongoing event. "Observation" itself may be taken as a process where a *d/dt* operator is invoked upon 4-space (spacetime), leaving a purely 3-spatial output. However, the resulting observable does not "exist or persist in time", since rigorously it is what was "caught" in only a single, frozen "3-slice" at one single instant, forever fixed.

For that reason, time is not an observable. It cannot be directly observed, even in principle, since *a priori* it is discarded by the observation process itself!

For further work along such lines, we recommend the Sachs unification approach to a combined general relativity, quantum mechanics, and electrodynamics, generalized from a topological standpoint {126a, 126b}. To work electromagnetically in that theoretical approach, we recommend the 0(3) electrodynamics as extended by Evans {137a, 137b}.

2.1.3.2. *The* **Four Polarizations of** *Photons* **and** *EM Waves*

In quantum field theory, four polarizations of a photon are recognized {138}. These are the *x-, y-, z-,* and /-polarizations, where *x-, y-,* z- and *t-* are assumed as the four dimensions in a 4-space. By agreement, z- is used as the direction of motion of a free photon or an electromagnetic wave. So at least in theory, there must also be four polarizations of electromagnetic waves, even though not all these waves are yet experimentally known.

The *x- and y-* polarizations are the familiar transverse photon and the transverse wave. The z- polarization along the line of propagation gives the longitudinal photon in 3-space and the longitudinal EM wave in 3-space.[43]

---

[43] Here we are **applying the conventional** model of the "EM wave in space", which is known to be erroneous **but is** in **ubiquitous** usage. See Romer, {84} for a scathing characterization **of that** erroneous model **But here we are stuck with it, so will use it** as "just a model".

2.1.3.3 *Imperfect Longitudinal EM Waves* ***in 3-Space***

When attempting to produce longitudinal EM waves (LWs) in 3-space, from transverse EM waves (TWs) input to some sort of polarization transduction process, only imperfect LWs are produced and a residue of TW content remains. The resulting imperfect LW is referred to as an *undistortedprogressive wave* (UPW). Some work has been done on UPWs {139}. UPWs are expected to have remarkable characteristics including wave velocities either slower or faster than *c {*140}.

The *t*- polarization in the time dimension is quite unique: The *spatial* energy overall is in equilibrium and not vibrating at all; instead, the energy of the photon or wave is vibrating in the time domain and therefore exists in the time domain. That is called a "scalar photon" (time-polarized photon). Its wave version does not yet seem to be known in the literature, although in 2000 we uncovered its secret hiding place {12}. It was unexpectedly hiding in Whittaker's {85} decomposition of the scalar potential in 1903, but had been slightly misinterpreted.

2.1.3.4 *Photons andEM Waves Carry* ***Energy and Time***

On the other hand, the concept of "EM waves flowing in 3-space" may be in need of a thorough overhaul {141}. A photon *y* is a "piece of angular momentum" in the form of $y = (dE)(dt)$. Hence the photon carries an increment of spatial energy d*E* and also an increment of time-energy d*t*. The time-energy component *(dt)* may be regarded as ordinary spatial energy that has been compressed by the factor c^2 {142}.

As can be seen, since *c* represents an orthogonal rotation in n-space, the multiplication or division by *c* and by powers of c changes the dimensionality of an entity, as seen by the observer in the laboratory frame.

So the photon transports two types of energy: (i) a "weak spring" (spatial, decompressed) energy d*E*, so to speak, and (ii) a "very stout spring" (time) energy d*t*, so to speak.

When a mass *m* absorbs a photon *(dE)(dt)*, the *(dE)* component is compressed spatially by c^2, turning it into an extra amount of mass d*m*, so that the mass becomes $(m + dm)*d$t the same time, the *(dt)* component is joined, so that what results is $(m + dm)*dt$. In short, mass *m* is changed to masstime *mt* by photon absorption. So in the absorption of a photon y by a mass *m*, we have

$$y+m\text{->} (dE)(dt) + m \text{->} (m + dm)dt$$

In short, the mass *m* turns into masstime *mt,* by absorbing a photon, and masstime *mt* is as different from mass *m* as impulse *Ft* is different from force *F.* We point out that "mass" *m* alone does not even exist in time, but masstime *mt* does exist in time. This is proposed as a simple but fundamental correction to much of present physics. Further, the state "mt" is an excited, time-charged state, excited by very dense time-energy.

For the simplest case, in the next instant the excited state *mt* decays and a photon is re-emitted, and so we have

$$(m + \Delta m)\Delta t \rightarrow (\Delta E)(\Delta t) + m \rightarrow \gamma + m \qquad [12]$$

So emission of a photon changes the excited masstime state (time-charged state) back to mass (uncharged state with respect to time-charging), in the simplest case.

See again Figure 1-3 in Chapter 1. "Observation" and "observable change" are in fact generated by the photon interaction. The total photon interaction with a mass creates the "flow of that mass" through spacetime, macroscopically as seen by the observer and microscopically in myriad streams of virtual time changes. Mass is continually charged to the masstime state, and the masstime state is continually decayed to the mass state. Time flow itself thus has a myriad internal EM energy streams and a vast dynamic EM energy substructure.

The photon itself is not observable, as it exists prior to interaction. It is thus a "causal" entity in 4-space. It *is not* a simple 3-space observed "frozen snapshot" particle! When the causal photon is absorbed, its spatial energy component goes through the c compression function, thus adding a small amount of new mass to the previous mass (previous effect), but simultaneously connecting its d*t* time-tail, so that the slightly increased mass in fact now exists as masstime and not mass. Observation has not been completed at that point.

In the next instant, a photon is re-emitted (from masstime, never from mass!), in the simplest case removing that extra little mass increase and orthogonally rotating it back into 3-spatial EM energy, coupling that spatial energy increment d*E* to that "time-tail" d*t* and tearing it away as the photon is re-emitted. That leaves behind a frozen 3-space snapshot of the mass (the interacting particle) and completes the "observation". In the simple case, this is a replica of the previous particle (frozen 3-space snapshot) with which the causal photon interacted. So this process accounts for the so-called "persistence in time" of an object or mass, as seen by the observer's continual recall process accompanying his stream of iterative observations. Mass does not *continuously exist* in time, but *continually recurs* in time.

As is well known, what we call "observable" change must involve the expression in equation [13] in the fashion discussed and with the observation mechanism given.

We first pointed out the preceding mechanism for the "flow of a mass through time" in 1971, as a graduate student at Georgia Institute of Technology, and later published it in a crude paper {143} in 1973. It still needs a far more thorough theoretical treatment, which perhaps some graduate student will take up as the subject of his or her doctoral thesis.

2.1.3.5 *Photon Interaction: Mechanism Generating Flow Through Time*

See Figure 2-1. Mass does not really "travel through time" continuously *per se*, but proceeds with an overall serial change mechanism as

$$m \rightarrow (mt) \rightarrow m \rightarrow (mt) \rightarrow m \rightarrow \ldots \qquad [14a]$$

where *(mt)* symbolizes a nonobservable ongoing "coupled" interaction state prior to observation completion. Equation [14a] represents the results of the continual photon interaction process, observation process, and "flow of a mass through time" process, at the highest single-photon interaction or quantum change level, and as "seen" or measured by the external observer.

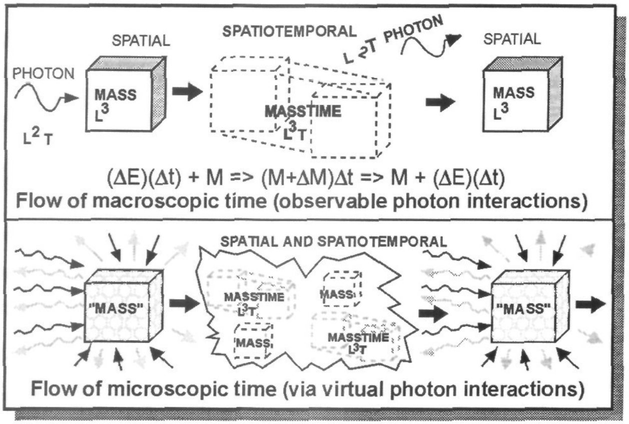
SPATIAL
SPATIOTEMPORAL
L^2T PHOTON
SPATIAL
PHOTON
L^2T
MASS L^3
MASSTIME L^3T
MASS L^3
$(\Delta E)(\Delta t) + M \Rightarrow (M+\Delta M)\Delta t \Rightarrow M + (\Delta E)(\Delta t)$
Flow of macroscopic time (observable photon interactions)
SPATIAL AND SPATIOTEMPORAL
"MASS"
MASSTIME L^3T
MASS
MASS
MASSTIME L^3T
"MASS"
Flow of microscopic time (via virtual photon interactions)

Figure 2-1 Mechanism for the flow of a mass through time.

Also, in that overall quantum level "flow" there are anyt number of ongoing streams and "sublevels" in the underlying subquantum level (in the virtual vacuum and in its virtual photon interaction with observable mass). We might write one of these "internal time-stream interaction flow substructures" as:

$$(\delta m) \rightarrow (\delta mt) \rightarrow (\delta m) \rightarrow (\delta mt) \rightarrow (\delta m) \rightarrow \qquad [14b]$$

In equation [14b] we have used parentheses, because all terms individually are nonobservable. We hypothesize that each stream continues until one of its terms participates in a summation which eliminates it by enfolding it into another interaction.

When a mass is observed as in [14a], a photon has been emitted (we observe the mass as the "effect" remaining). Time has been stripped away by the resulting *d/dt* operation, leaving a frozen 3-spatial snapshot, which we will see as (having been) a particle (simplest case). That occurs just after major ("observable") photon emission from the masstime state {144}. Immediately another observable photon is absorbed, and so state *mt* occurs. The particle of mass actually oscillates at a very high rate between the *m* and *mt* states — so high a rate that by arranging the interaction conditions one may interact with it either as a wave (react predominantly in the *mt* state) or as a corpuscle (react predominately in the *m* state).

Hence we propose that the process in [14a] may account for the duality of particle and wave.

2.1.3.6 *The Overall Flow of Time Has an Internal Dynamic Structure*

During the transition in any mass to masstime state by reaction of the mass with an "observable" photon, a myriad of fleeting virtual photon interactions involving very tiny *(dE) (dt)* components occurs with the mass *m.* These tinier increments of time, and their increments of energy, constitute internal structures in the time flow process. Therefore they may be considered as "energy currents" or "time-like energy currents" and dynamic structures or streams inside the flow of time. This is particularly straightforward if we use a model having only a single fundamental unit, the joule. In that case, time is energy, and we are speaking of energy flow and its constituent internal structures of energy flow.

So the d*t* component of masstime at the observable-photon action level has a myriad of energy-time structured dynamics infolded within it. Hence the *mt* state is very dynamic in time, particularly for fundamental particles. The *mt* state is in fact a "collection of time-energy dynamics" and therefore "wavelike".

See again Figure 2-1. A major point is that ***mass*** does not emit a photon; *masstime* does. Mass "travels through time" by an extremely high oscillation between corpuscle-like state *m* and wavelike state *mt*, and with a vast internal "dynamic streams" of other such high oscillations between corpuscle-like *m* states and wavelike *mt* states.

The concept can be very much extended, of course, but this suffices for our concept of energy currents in time, the interaction of such energy currents with mass in a mass system, and the internal dynamic structuring of the "flow of time".

This internal structuring is important in the event of *"time-energy" charging.* There the internal structure of time has experimental consequences. We will discuss those consequences in later chapters.

2.1.4 The Ubiquitous Substitution **of** Effect **for** Cause

See Figure 1-3. To repeat, no observable exists or persists *as such* in time. An observable *as such* only exists at all when time has been momentarily stopped. It then quickly changes into (observable x time) form, for a time interval after a photon absorption. Then it changes again to observable form by photon emission and the corresponding *d/dt* differentiation imposed by that photon emission.

What happens when we think we see an "observable moving through space" is that we make or suppose a continual series of very fast *d/dt* 3-spatial snapshot observations, one after the other in serial fashion (much like the individual frames of a motion picture film). For more than a million years, human brain and mind processing have "always interpreted" this serial set of iterative operations occurring in the eyes and in the physical senses — and continually recalled and processed by the mind — as an "observable persisting in time", which it is not.

This age-old "natural" and inbred, instinctual *mental practice* and assumption by humans — and therefore by scientists — has provided a major problem in physics and especially in electrodynamics. It has resulted in the substitution of the "effect" for the cause in a great number of models. Even a rapid and continual series of 3-space effects after reaction of an observable (frozen snapshot) with a 4-space causal entity, does not and cannot constitute that 4-space entity which interacts with the previous "frozen snapshot" observable and causes the resulting "comparative change or sameness" of the new observable with the former. Hence, e.g., Romer's scathing condemnation {84} of the conventional drawing of an "EM wave in space".

In general relativity, it is straightforward. The curvature of spacetime — the cause — is not the mass (the 3-space observable effect) that is further changed or created again in the ongoing interaction, as seen in the "next

3-space snapshot" (next output of observation) when comparing that snapshot to the previous one.[44]

The same non sequitur — unwitting substitution of effect for cause — has existed for hundreds of years in mechanics, for example. There it is strongly passed on in the erroneous old notion of a separate force acting upon a separate mass. *Prior to* the interaction, no "force" exists. *During* the ongoing interaction, force exists because it is the product (interaction) of a "non-force" causal entity (e.g., a field in 4-space, as a certain curvature of spacetime there) and a previous effect (a 3-spatial frozen snapshot called "mass"). The force exists during the interaction and only during the interaction, when both cause and previous effect are coupled (and will produce the new effect at the completion of the next serial *d/dt* operation as a photon is emitted). The force can only exist when the mass (previous effect) is coupled to the cause (the curvature of spacetime), since *a priori* mass is a *component* of force {145}, and curved spacetime is another *component* of it. After the interaction "ends" in a new *d/dt* observation, the mind compares the new effect with the former, to determine whether there is a "change" or a replica (sameness). This comparison of the new observation (observable) with respect to the former, *occurs in the brain and mind processing*, and gives the sense of "persistence of an object in time" as well as of an object "changing in time".

So a great faux pas in physics — and especially in electrodynamics — is this widespread substitution of the effect for the cause. Understanding this non sequitur clearly, and correcting it, is a prerequisite to understanding the principles and concepts of legitimate COP>1.0 Maxwellian systems. Also, if we would ever hope to adequately correct electrodynamics, this tremendous faux pas must be recognized and rooted out of the model. Indeed, the same is true for mechanics and other branches of physics.

2.1.5 Some Additional Foundations Non Sequiturs

Another gross non sequitur is made in calculating the amount of EM energy flow continually being diverged around an intercepting unit static charge at a point in space, from a potential in which that point is located, and then calling this "local intensity" of the collecting interaction the "magnitude of the potential" itself.

---

[44] Rigorously, if one wishes to connect spacetime and mass, spacetime can be interpreted in *masstime* units, but not ***mass*** units And vice versa. Again, masstime *mt* is as different from mass *m* as impulse Ft is **different** from force *F*.

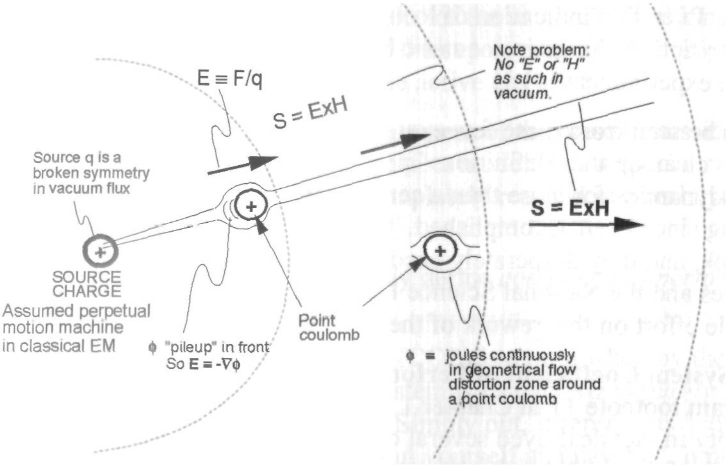
E ≡ F/q
S = ExH
Note problem:
No "E" or "H"
as such in
vacuum.
Source q is a
broken symmetry
in vacuum flux
S = ExH
SOURCE
CHARGE
Assumed perpetual
motion machine
in classical EM
ϕ "pileup" in front
So E ≡ -∇ϕ
Point
coulomb
ϕ ≡ joules continuously
in geometrical flow
distortion zone around
a point coulomb

Figure 2-2 Field and potential are defined in terms of the diversions from them around a unit point static charge assumed at any spatial point

(See Figure 2-2). At best, it is a measure of the *local intensity of the diversion of energy flow from* the "bidirectional" EM longitudinal waves comprising a potential, as they are diverged around the unit point static charge. It is not the magnitude of the entire potential, filling all space, nor does it prescribe the spatial energy magnitude of that entire potential {146}. It is a small energy flow, *diverted* by an assumed standard fixed charge at a single point in the potential, from the overall energy flows comprising the potential. Calling this little "diversion of a bit of energy" the magnitude of the potential is rather like calling a little whirlpool in a river the "magnitude of the river". Or more exactly, it is like calling the amount of little "swirl diversion of a river's flow" around a standard small fixed rock the "magnitude of the river".

A similar non sequitur exists in the usual calculation of the "field" magnitude at a point, which is not the "entire field entity itself at all. It is merely what is *diverged from* the "field as it exists in spacetime" at a single point in it, by an assumed fixed standard point charge at that point that is interacting with the field. The field takes into account the patterning of the diverged flow, e.g., the difference in potential (difference in pressure) across the unit point charge. That is not the field itself at all, and the textbooks and dictionaries of physics are completely wrong in calling it such. Again, at best, it is an indication of *the field intensity* at that point, and the reaction cross section of the field to an assumed unit point static charge. Even that indication of local intensity of the field will change if the intercepting charge is not static but is in resonance, as shown by the Bohren experiment {24}.[45]

As can be seen from these few examples, a great deal of work needs to be done to clean up the "definitions" promulgated in physics and electrodynamics for more than a century. It is inexplicable[46] that this has not long since been accomplished. These errors permeate almost every textbook, and they desperately need correcting. The National Academy of Sciences and the National Science Foundation should fund the best possible effort on this rework of the foundations of physics.

2.1.6 System Coefficient of Performance (COP) and System Efficiency

See again footnote 11 in Chapter 1 and Appendix A. In 1915, general relativity in fact destroyed several of the fundamental definitions (axioms) of classical thermodynamics. It necessitated our correction of some thermodynamical definitions (open system and closed system) and our more rigorous definitions of COP and efficiency advanced below.

Once mass is recognized as energy and the two are just different sides of the same coin, there can be no thermodynamically closed system which passes energy across the system boundary without passing mass across it. Any system changing its rate of flow through time also changes its mass and its energy, and vice versa. Energy has mass characteristics, and any mass system with added or subtracted energy — either kinetic or potential

---

[45] Many Bohren-type experiments are continually done in nonlinear optical labs in universities and elsewhere. The effect (excess radiation from the medium) is euphemistically called "negative absorption of the medium," "negative resonance absorption of the medium," etc. Such terminology avoids clearly recognizing that more energy is output than the scientists (erroneously) calculate was input. Bohren faced the issue head-on, and clearly stated that more energy was output than was input (or than was thought to be input).

[46] A true pessimist might point out that it may be quite understandable, considering that it solidly blocks COP>1.0 EM electrical power systems and electro-gravitational anti-gravity systems from being developed by our universities, which would solve the energy and transportation crises forever. Thereby it also keeps about $1.5 to 2 trillion a year pouring into the coffers of controlling financial groups behind many great interlocking corporations.

— will also change its mass relativistically as well as its other characteristics such as inertia and gravitational attraction. As pointed out by Logunov and Loskutov {635}, in 1917 the new characteristics of general relativity led Hilbert {632} to observe that in general relativity there can be no energy conservation rules of the usual classical kind. This also follows from our consideration of the supersystem with multiple active environments rather than a single active environment.

We have formally destroyed any absoluteness of the present classical thermodynamics, which is just an imperfect model. It is a special case which can be approached but never completely reached in a real dynamic system.

With that in mind, we now more exactly define *coefficient of performance* and *efficiency* of a system.

The term "coefficient of performance" (COP) is a ratio whereby the useful energetic output performance of the system is characterized *purely with respect to the operator's energy input.* Simply put, it represents "how much you get for what you have to input yourself and pay for". It does not characterize the performance of the system with respect to the *total energy input from all sources.* For our purpose, in general there are two major ways of expressing this COP:

(a) COP == (work accomplished in the load during a representative time of operation) divided by (energy input by the operator during that representative time of operation). Or for short, COP = (useful work out in the load) + *{operator's* energy input). We stress that there may be other free inputs of energy to the system, in addition to the operator's input, but only the operator's input is accounted in the COP calculation.

(b) COP == (average power out in the load) divided by (average power input by the operator).[47] Again, there may be additional average power inputs that are freely input from the environment, without cost to the operator, but only the *operator's* input is accounted. We stress that "power input" is another of electrical engineering's misnomers, but we use it here because it is universally used in power system engineering.

---

[47] This is in standard electrical engineering terminology, which is mangled. For a more precise physics statement of (b), COP == (the average rate at which the system dissipates energy in its load to do useful work) / (the average rate at which the operator must furnish energy to be dissipated as work in the input section of the system to make it operate and do its output work).

For a system such as a transducer, which merely changes the form of the energy in some fashion and does not perform work in an external load (the transducer is its own load), we may express the COP as

a. COP == (effective energy output) / (energy input by the operator). Here we accent that there may be additional energy inputs to the system from the active environment, and these inputs are "free" and are not input by the operator, so they are not accounted in the COP calculation.

b. COP == (average power output) / (average power input by the operator). Again, additional average power inputs may be freely received from the environment, but they are not accounted in the COP calculation.

The "efficiency" E of a system is a ratio less than or equal to 1.0 (or a percentage less than 100%), where E indicates the percentage of the total input energy (from all sources) that is dissipated in the load as useful work. It follows that (7 - E) indicates the percentage of the total input energy that is dissipated in the internal losses in the system, not directly resulting in useful work by the system in its load. So the efficiency E of a system may be expressed as:

a. E == (total work output in the load during a representative operational period) / (total energy input to the system, from all sources, during that period).

b. E == (average power output in the load) / (average energy rate of input to the system, from all sources, during that representative operational period.).

We accent that no system can have an efficiency greater than 100%, for then it would be "creating energy from nothing" {147}. No system can dissipate or convert energy that it does not first receive. The conservation of energy law states that energy can neither be created nor destroyed. This means that *there actually are no energy sources per se, in the sense that the source creates the energy,* even though we use the terms "energy source" and "source of energy" informally. E.g., Semiz {148} states it this way:

> *"The very expression 'energy source' is actually a misnomer. As is known since the early days of thermodynamics, and formulated as the first law, energy is conserved in any physical process. Since energy cannot be created or destroyed, nothing can be an energy source, or sink. Devices we call energy sources do not create energy, they convert it from a form not suitable for our needs to a form that is suitable, a form we can do work with."*

A medium inefficient system can easily have a COP>1.0, if it receives additional energy from its active environment, and if that extra free energy is sufficient to overcome its internal inefficiency losses. An example is a common windmill, where a very good one may have an efficiency of less than 50% due to friction and drag losses in the gears and wind spillage losses in the blades, etc. Yet the operator himself does not have to input any energy at all, for the windmill to operate continually. In this case, the COP»1.0 and approaches infinity, but the efficiency of the windmill is still less than 50%.

Another example is the home heat pump, which may also have an efficiency of less than 50%. When acting as a refrigerator, its theoretical maximum COP = 8.22 under nominal conditions {149}, and a well-designed 50%-efficient home heat pump will produce COP = 4.0 when conditions are suitable.

On the other hand, if the only net energy input to the system is that energy that the operator inputs, then the COP < 1.0. This is a system in equilibrium with respect to any energy exchange with its external environment — *except* with respect to the energy input by the operator (i.e., with respect to its "fueling" by the operator, so to speak) and with respect to the energy subsequently dissipated in the loads and losses. If the system is 100% efficient (has no internal losses or conversion losses at all), its COP = 1.0. Almost all real systems do have internal losses, so their efficiency is E < 100%. In that case, the system in equilibrium with its environment, and having E < 100%, will also exhibit COP<1.0

irrevocably. Indeed, the same number will give both the efficiency E and the COP, because numerically they are then the same.[48]

For electromagnetic systems, the state of confusion between efficiency E of the system and the COP of the system is due to one fact. *Numerically* the two are always equal in *equilibrium* systems — and the Lorentz regauging condition enforced by the closed current loop circuit self-enforces the equilibrium condition of the system with respect to its active environment. Because electrical engineers usually have zero experience with COP>1.0 electrical systems, they tend to loosely and erroneously use the two terms "efficiency" and "COP" as if they meant the same thing. They do not.

Now suppose that we have an open system, far from equilibrium in its energy exchange with its active environment. Suppose that the system's efficiency E is very poor, so that E = 20%. Now suppose that the environment inputs twice as much energy as does the operator. Let the operator's energy input be E_1. Then the total energy input to the system, from both the operator and the environment, is $3E_1$. The efficiency is only 20%, so the system outputs W, as work in the load, of $W = 0.2(3E_1) = 0.6 E_1$. The COP of the system is the work out divided by the operator's input, which is $COP = W/E_1$, which is $COP = 0.6\ E_1 / E_1$ which gives COP = 0.6. As can be seen, even though this system receives twice as much additional free energy as what the operator inputs, it is so inefficient that **its COP<1.0.**

Suppose we have a similar system with the same energy inputs from the operator and from the environment, but now the system's efficiency E is E = 90%. Then the total energy input to the system, from both the operator and the environment, is $3E_1$. The efficiency is 90%, so the system outputs W, as work in the load, of $W = 0.9(3E_1) = 2.7 E_1$. The COP of the system is the work out divided by the operator's input, which is $COP = W/E_1$, which is $COP = 2.7\ E_1 / E_1$, which gives COP = 2.7. As can be seen, a more efficient system in the same energy input situation, now outputs more work than the energy input by the operator. Energy is conserved at all times; the excess energy for the additional work was in fact freely input to the system — which is an "open" system far from equilibrium in its exchange with its active environment.

---

[48] We emphasize that COP and efficiency are two quite different concepts, however, even when their numerical values are the same. A 6-foot tall man and a 6-foot tall door have the same number for their **height,** but only a fool would consider them the "same thing".

Thus a system far from equilibrium in its energetic exchange with its active environment, is *permitted* to exhibit COP>1.0 even though its efficiency is always less than 100%. On the other hand, if the system is very inefficient and the energy input from the environment is not too great, the system will still exhibit COP<1.0. But the operator will pay less for his energy costs to operate that inefficient system, than he would pay to operate it if he himself had to furnish all the energy input.

Two cautions are emphasized:

(a) First, electrical engineers use the term "power" to mean "energy flow rate without dissipation or change of form", as well as "energy flow dissipation and change of form rate." This is inexact and unfortunate, but it appears ubiquitously throughout the electrical engineering literature. So we are essentially "stuck with it" for the moment. From a rigorous physics point of view, energy flow without divergence or change has absolutely no power, because power is defined as the rate of change of the form of that energy flow — e.g., the rate of dissipation (scattering) of that energy flow, which in physics is a rate at which work is being performed.

(b) Second, it seems that more than half the engineers and scientists — and many textbook authors — do not clearly understand the difference between COP and efficiency, and often interchange these two terms as if they were the same. They are not at all the same, and a great deal of confusion exists in the casual engineering literature because of using them interchangeably. That is why we have clarified them and given some simple explanatory examples, sufficient for our purpose in this book.

2.1.7 COP>1.0 and the Question of Perpetual Motion

Legitimate scientific researchers attempting to work in open EM systems far from thermodynamic equilibrium are often subjected to rather vicious ad hominem attacks, charging that they are perpetual motion machine addicts and therefore raving lunatics. This of course is not true. We are simply trying to do with Maxwell's "electric fluid" systems what nature already shows us can be done with material fluid systems. We are in fact seeking the equivalent of "electromagnetic heat pump" processes and systems. It is the strident critics who are thoroughly confused and who reveal an immature knowledge of physics and thermodynamics.

As we explained, prior to their Lorentz symmetrical regauging, the Maxwell-Heaviside equations already clearly prescribe and permit COP>1.0 EM systems, including EM power systems. In addition to the present EM systems in equilibrium in their exchange with their active vacuum environment, the non-regauged equations also include Maxwellian systems far from equilibrium in their energetic exchange with an active environment, such as the active vacuum. Classical *equilibrium* thermodynamics — stridently waved as a sacred banner by the critics accusing COP>1.0 researchers of being "perpetual motion nuts" and such EM systems as being "perpetual motion nonsense" — does not even apply to such *disequilibrium* systems! One should simply read the literature of the thermodynamics of such disequilibrium systems {160a-160f}. Or one should read why Prigogine was awarded the Nobel Prize in Chemistry in 1977, for his contributions to such open systems in disequilibrium.

We accent again that, under our new rules where we corrected the thermodynamic definitions of open system and closed system, we are in compliance with general relativity. Further, every EM system must be taken as an open system, since in the supersystem energy (with mass-like changes resulting across the system boundary) is always exchanged between the three supersystem components.

Fortunately there are also rational, reasonable critics such as Angrist {150, 151} where the criticism is properly scientific and not ad hominem, even though still misguided. Usually the rational perpetual motion critic applies the three laws of classical *equilibrium* thermodynamics, to classify "perpetual motion machines" of class 1, 2, and 3 after equilibrium thermodynamics laws 1, 2, and 3. This of itself has no intellectual content for open systems in disequilibrium with their active environment. The classical equilibrium thermodynamics — including laws 1, 2, and 3 interpreted only in terms of the operator's energy input — does not apply to *open systems far from equilibrium* — which are the only systems permitted to exhibit COP>1.0 in the first place. And we insist that all EM systems are open systems far from thermodynamic equilibrium, since every charge is, and all fields and potentials come from those source charges. Indeed, such open disequilibrium systems are *permitted* to violate the laws of classical thermodynamics! As an example, the entropy of such an open disequilibrium system cannot even be computed {152}, and it is less than the entropy of the same system in equilibrium, since the equilibrium state is the state of maximum entropy.

The rational critic then falls upon the horns of a dilemma, by admitting that, yes, such open dissipative systems do exist and do permissibly exhibit

COP>1.0. He then labels these *as, false or fictitious perpetual motion machines* {150, 151}. In short, when such COP>1.0 machines are real and admittedly exist, they are not *"true"* perpetual motion machines but *fictitious* perpetual motion machines. My comment is that they are not perpetual motion machines at all, but permissible disequilibrium machines freely receiving excess energy from their energetic exchange with their active environment. They are no more mysterious than a windmill, a sailboat, a watermill, or a solar cell bank.

One must be precise. A bicycle is a free-energy COP>1.0 machine while going downhill. It is not such at all when going uphill. The difference is that gravitational input of energy aids the bicycle's motion in the downhill case, and opposes it in the uphill case. The active environment gives the system additional energy in one case, and takes energy from the system in the other case. Disequilibrium exchange can cut either way!

It is precisely the EM system far from equilibrium with its active environment — where the environment puts in additional energy — that the serious overunity researcher is seeking. It is a perfectly valid pursuit, even by the standards of the perpetual motion critics, who really should study *both* kinds of thermodynamics, not just *one.* They also should study how Maxwell's equations were curtailed and truncated; one result is that all Maxwellian disequilibrium systems were — and are — discarded *arbitrarily* by the Lorentz symmetrical regauging.

The only criticism one needs to raise vis a vis COP>1.0 and perpetual motion machines is this: If an inert (no hidden internal source) machine receives no net energy from its external active environment, but is in equilibrium in its energetic exchange with that environment,[49] then classical thermodynamics rigorously applies to it and the machine cannot and will not exhibit COP>1.0. However, all EM systems contain charges, and these are indeed "hidden sources" of EM fields and potentials and their energy. It *rigorously follows that, since they contain hidden energy sources, all EM systems are potentially COP>1.0 systems a priori.* Further, *it follows that some ubiquitous design characteristic of the system itself must prevent the system from achieving COP>1.0, since nature does not prevent it.* That characteristic feature is the ubiquitous closed current loop circuit, directing all spent electrons in the external circuit right back through the source dipole in the generator or other primary power source.

---

[49] Here we quibble again and point out that Lorentz's symmetrical regauging actually does require that the system receive and take on excess EM potential energy from its environment (the active vacuum and the local curved spacetime). However, this excess energy can only be received in the form of two injections of field energy, where the fields are equal and opposite. Hence the energy is "bottled up" as a stress potential, with net zero translation force available to it. Since this stress potential energy cannot translate electrons, it cannot do work in an external load. It merely changes the stress of the system itself. The net result is actually a rotation of the frame in which the system exists, away from the "laboratory frame" of the observer. In short, the Lorentz condition establishes a new "equilibrium condition" for the system, by rotating its frame. It also implies that EM energy is continuously received by the system from both the active vacuum and local curved spacetime, in order to maintain the rotated frame and **the** new equilibrium condition. Further this

It is indeed mandatory that — if at all possible - the inventor or independent researcher *identify* a legitimate source of environmental energy that his purported COP>1.0 device receives and collects, and how it receives it and then uses it to power the loads. That means he should identify where and how his system violates the Lorentz self-regauging condition imposed by the standard closed current loop circuit. If the inventor himself cannot explain it — and this does happen — then someone else must explain it for him. If the machine works under proper independent test and replication, then criticism that no explanation exists for it in normal physics is of no relevance. It is a spur to the theorist to enlarge his horizons and to find the proper physics that *does* explain it.

We stress that we cited the works of Angrist because his voice is reasoned and he does not stoop to name-calling and ad hominem attacks. He also does point out a historical rash of machines that clearly must comply with classical equilibrium thermodynamics, because they have no discernible or stated extra input of energy from their active environment and have no self-asymmetrical regauging. When such machines have no external free input of energy from their active environment, they are systems in equilibrium with their active environment, and cannot exhibit COP> 1.0 — exactly as pointed out by Angrist. With such a reasoned critic, a scientific dialog is possible, and his main thrust — that no *equilibrium* system can exhibit COP>1.0 — is quite valid. With the vehement critic, it is already a cur dog fight from the outset and there can be no scientific discussion at all.

The only contention one has with a rational critic, whose criticism is based on classical thermodynamics, is when he then generalizes that *all* EM

continuous input of two energy flows, equal and opposite, does continuous internal work on the Maxwellian system to increase and maintain its stress.

systems fall within this class. That conclusion is a non sequitur, and ignores the entire science of disequilibrium thermodynamics and a vast body of scientific literature and scientific experiments in electrodynamics. When stated for electrodynamic systems, it also ignores the actual documented history of electrodynamics, and ignores Lorentz's *arbitrary discard, from all accountability,* of the Heaviside giant nondiverged energy flow component accompanying every circuit and system. It also ignores Lorenz's and then Lorentz's symmetrical regauging of the Maxwell equations to give new and simple equations easier to solve. This arbitrary regauging did in fact unwitting discard all open disequilibrium Maxwellian systems — precisely those permitted to exhibit COP>1.0.

See again Figure 1-1. The energy entering the circuit is entering it from the surrounding space outside the circuit. Heaviside showed that the energy flow component remaining alter the Poynting component is withdrawn is orders of magnitude greater than the Poynting component entering the circuit and powering it. *It can easily be demonstrated that every generator outputs far more EM energy in space surrounding the power line attached to it, than the operator inputs as mechanical energy to the shaft.*[50]

---

[50] E.g., just use a gedanken experiment followed by the Bohren {24} experiment as validated by Paul and Fischer {25}. Consider a 100% efficient DC system, powered by a DC generator, using a closed Current loop circuit. All the Poynting energy received by the system appears at first blush to be dissipated in the external load. That is not true, but let us not yet belabor the point. Include the complete circuit loop back through the source dipole in the generator. The back emf voltage across the source dipole between the terminals is precisely equal to the forward emf voltage between the terminals but around the external half-loop. The same current flows through both the forward voltage drop and the reverse voltage repotentialization. So precisely as much energy is dissipated to force the current through the back emf, as is recovered from the load.

Accept the conventional view that the shaft energy input to the generator is transduced into magnetic energy, which is then dissipated to force those charges back though the back emf. Now examine the energy flow in space surrounding the external circuit, using both the Poynting theory and Heaviside's theory. We have accounted for the Poynting diverged component, but we have not accounted for the huge remaining nondiverged Heaviside component. From whence comes this excess remaining energy flow that did not enter the circuit, was not included in the Poynting calculation, and was not expended in the circuit but was just wasted? To prove the excess energy flow really exists, perform the Bohren experiment {24} where the intercepting charges in the circuit are in resonance, and thus "sweep out" a larger geometrical cross section of interception of the impinging energy flow. These charges do sweep beyond the static unit charge cross section conventionally assumed in the definitions of field intensity (e.g., of the fields E and H in the Poynting vector S in S = (E x H). If the defining unit point charges are in resonance and sweep out a

Heaviside knew this in the 1880s. All the energy flow that is intercepted by the circuit and enters (i.e., the Poynting component) at essentially right angles, hardly even changes the angle of flow of the remaining huge flow of energy surrounding the circuit and missing it. Quoting Heaviside {153} directly:

> *"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire.... Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire. "*

So every electrical power system is rigorously a COP>1.0 "energy gating" system already, if the neglected, enormous Heaviside energy flow through space around the circuit (and missing it and wasted) is re-accounted. In that case, much greater energy pours out of the terminals of every generator, than the feeble amount of mechanical energy we input to the shaft.

We do not have an *energy availability* problem, and we never have had one! Instead, we have an *energy flow interception and collection* problem

greater cross section, then the assumed Poynting vector S, for the static case, changes in magnitude by some ratio k for the resonant case to the vector S_R, so now $S_R = k_1E \text{ x } k_2H = k_3(E \text{ x } H)$. Since k_3 is just the ratio of the actual geometrical cross section swept out by the charge to the standard geometrical cross section swept out by the static charge, then for a static charge $k_3 = 1.0$, and for a resonant charge $k_3 > 1.0$. Hence the Bohren experiment, with $k_3 = 18$ or so, produces 18 times as much collected (Poynting) energy "out" as we erroneously think we input by normal calculations ignoring the input Heaviside nondiverged component. The Bohren experiment {24} proves the argument; validation by Paul and Fischer {25} appears in the same journal issue.

— and we always have had one, due to the ubiquitous use of the closed current loop circuit.

Texts — e.g., such as by Kraus {154} — do show the Poynting (intercepted and collected) component of the EM energy flow surrounding the power line attached to the generator terminals, but none of them today go into the remaining nondiverged Heaviside component. Each of Kraus' contours {76} of energy flow in space, around those power line conductors, shows *only that part of the energy flow in space that is being drawn into the circuit.* It does not show the remaining huge energy flow that (i) is not intercepted, (ii) does not enter the circuit, and (iii) is wasted. Presently no texts illustrate this Heaviside nondiverged energy flow component, and no text accounts for it except to point out that the Poynting component must be considered to be indefinite, since it can be accompanied by an additional energy flow component. The texts then erroneously conclude that the excess flow can have no physical significance.[51]

One can easily prove the existence of additional energy flow to be collected, in addition to the Poynting component. E.g., simply examine the Bohren experiment {155} and its replication by Paul and Fischer {156} — or one can repeat the experiment oneself. The Bohren experiment outputs some 18 times as much energy as is input by normal "Poynting component" calculations alone, because it also intercepts additional energy from the unaccounted and ignored *free* Heaviside input component remaining and not diverged by a *static* unit point charge. Bohren simply places the charge in resonance, so it sweeps out a greater geometrical reaction cross section in the energy flow, thereby penetrating into the free — but arbitrarily discarded and unaccounted — Heaviside component, and intercepting and collecting additional energy. Bohren is perfectly aware that the excess energy caught by the resonant charges comes from the surrounding vacuum. However, he seems quite unaware of Heaviside's discovery and cautious statement of the excess energy involved in *every* field-particle interaction, and particularly in electrical circuits.

---

[51] E.g., J. D. Jackson, Classical Electrodynamics, Second Edition, Wiley, 1975, p. 237: *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences. "* Jackson errs in stating that the added energy flow vector cannot have any physical consequences. It cannot, of course, unless one catches some of it and uses it to power a load, without using half of the caught energy to destroy the source dipole faster than the load is powered. One of the major applications of it is to use it with Dirac sea negative energy 4-holes in the vacuum, where the holes produce negative energy fields and potentials in surrounding spacetime, acting back upon the system which is the source of the 4-holes to produce practical antigravity.

There are several such "white crows" which are legitimate COP>1.0 EM processes in physics. It only takes one small white crow to prove that not all crows are black. As another example, the papers of Letokhov {157a-157d} are especially recommended. So long as excess energy is received freely from the environment, the system is permitted to be in disequilibrium where COP>1.0.

So far as this author could discover, none of the perpetual motion machine critics have noticed another very simple fact: The conventional electrodynamics they utilize to refute COP>1.0 Maxwellian systems implicitly assumes that the source charge — which is any and every charge in the universe — continuously *creates* and pours out energy in 3-space without any energy input at all. Their own conventional classical electrodynamics contains a "perpetual motion machine assumption" of the most fundamental and grossest kind. The ardent skeptics of overunity EM systems — in using that implicit conventional assumption — are themselves the greatest perpetual motion advocates of all, though perhaps unwittingly. In 2000 we resolved that long-vexing problem[52] of the source charge {12}.

The more strident critics simply set up the "equilibrium thermodynamics" straw man, then knock it down and attack independent COP>1.0 experimenters with it, erroneously implying that all overunity researchers seek such in an equilibrium system. They fail to notice orthodox science's implied assumption that every charge and every dipole in the universe is already assumed to be a COP => inf. perpetual motion machine, freely *creating* all that continuous outpouring of energy from nothing. None of the critics have noticed and critiqued this greatest of all perpetual motion assumptions right in their own ranks, without a single exception, and the most ubiquitous. Undoubtedly that is because the scientific community honestly and freely admits that it has not heretofore solved that problem {68}, and also points out that the outpouring of energy is an experimental fact and therefore unquestionable.

The better critics — at least the rational ones such as Angrist — already admit that there exists a class of legitimate "fictitious perpetual motion machines" which are powered continuously without the operator inputting the energy. The phrase "fictitious perpetual motion machine" is an attempt to retain the classical thermodynamics straw man, when dealing with the thermodynamics of a system far from equilibrium in its energy exchange with its active environment. That phrase has been a non sequitur since the award of the Nobel Prize to Ilya Prigogine in 1977. We stress again our new use of all EM systems as open systems, in accord with general relativity.

---

[52] After publication of our paper, we also discovered very powerful support for our solution in quantum field theory. E.g., see F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984/1993, Chapter 5.

So when rigorously analyzed from the breadth of physics and thermodynamics, and not just from a selected very narrow portion, the entire "perpetual motion" critical literature is inappropriate or — more exactly — is inappropriately focused upon a fictitious or false target. It is simply a classic example of reasoning from a false premise.

Not one such critic seems to have noticed that every generator and battery already pours out more energy than what is input to the shaft, and that it is easily proven experimentally via the thousands of experiments showing negative absorption from the medium. None seems to have bothered to read the original Heaviside, Poynting, and Lorentz papers to see how this startling fact was buried more than a century ago.

Finally, a self-appointed "expert" who does not even know what furnishes the energy to power every electromagnetic circuit, cannot be taken seriously in his "learned criticism" , regardless of what journal publishes it. If he continues to accept the unsolved source charge problem, implying that all charges freely create EM energy from nothing, he is not expert in any facet of powering electromag letic systems and circuits.

2.1.8 Prescribed Systems Versus Changes to Maxwell's Equations

As we stated, Maxwell's theory {20} was some 20 quaternion-like and vector equations in some 20 unknowns. By discarding the scalar component of the quaternion to retain only the vector subcomponent, and by discarding many of Maxwell's potentials, Heaviside {158} truncated the Maxwell equations to essentially the present familiar set of four vector equations - or in potential form, to two equations with variables unseparated.

While this truncation discarded a great deal of Maxwellian electrodynamics phenomenology, from a thermodynamics view the resulting *Maxwell-Heaviside* equations still contain and prescribe two types of Maxwellian systems:

(1) systems in equilibrium with their active environment, such as the active vacuum, and

(2) systems far from equilibrium with their active environment, such as the active vacuum.

Systems in the first class of Maxwell-Heaviside systems — i.e., in equilibrium in their energetic exchange with their active environment — rigorously obey classical equilibrium thermodynamics {159}. No system in that Maxwellian class can output more energy than the operator himself inputs to the system. The greatest coefficient of performance (COP) that these systems can accomplish, is COP = 1.0 for a theoretically perfect system without internal losses whatsoever, and thus with 100% efficiency of conversion of their input energy into work in the load. Since any real system has internal losses and thus is less than 100% efficient, the *practical* systems of this first class exhibit COP < 1.0.

These Maxwellian systems of the first class are the electrical systems always "understood" by ardent skeptics who proclaim that no other kind of Maxwellian system exists, that therefore no Maxwellian system can exhibit COP > 1.0, and that a claim to COP>1.0 is perpetual motion lunacy and not science at all. These fellows often launch scathing and unjustified personal attacks upon any serious researcher trying to work with that second class of permitted Maxwellian systems! It would be scientifically beneficial if these self-appointed defenders of the faith would simply study some more physics and thermodynamics, review the history of Maxwell's theory and how it was altered, and understand the well-established thermodynamics of systems far from equilibrium.

This false charge — that there can be no COP > 1.0 EM system because that would violate the laws of physics and the laws of thermodynamics — has been insidiously indoctrinated into generation after generation of physics and electrical engineering students. Today most of our electrical engineers and electrical scientists have absorbed it and believe it without question. Our power-engineering scientists believe it, and so advise the government. Hence our government believes it on the advice of our best scientists, such as the National Academy of Sciences and the National Science Foundation. Our great national laboratories also believe it and are just as mentally conditioned. The universities — especially the electrical engineering departments — also believe it, and provide a mainstream suppression of the sharp young graduate students and post-docs that would otherwise have long ago revised this mess and solved the energy crisis permanently.

Again we are reminded of that first quotation by Einstein, which we included at the beginning of Chapter 1, and which admonished us to ever examine the foundations of our science to root out and correct the errors. We are also reminded of another quote by Tolstoy:

> *"I know that most men, including those at ease with problems of the greatest complexity, can seldom accept even the simplest and most obvious truth if it be such as would oblige them to admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives."* [Count Leo Nikolaevich Tolstoy].

The second class of Maxwell-Heaviside systems — rigorously prescribed and permitted by Heaviside's truncated Maxwellian theory — is comprised of systems in disequilibrium with their active environment. A system in this class does not obey classical equilibrium thermodynamics. To the contrary, this system obeys the well-known *thermodynamics of systems far from equilibrium* {160a-160f} with their active environment — such as the active vacuum.

Such a system *is permitted to:*

(i) self-order,

(ii) self-oscillate or self-rotate,

(iii) output more energy than the operator himself inputs and thus exhibit COP>1.0 (the excess energy is freely received by the system from its external active environment),

(iv) power itself and its load simultaneously (all the energy is freely received by the system from its external active environment,[53] without any operator energy input being required), and

---

[53] We accent the concept of the supersystem. consisting of three components: (a) the system and its dynamics, considered as in normal electrical engineering (with a flat spacetime and an inactive local vacuum), (b) the active local vacuum and its dynamics, and (c) the active local curvatures of spacetime and their dynamics. All three components of the supersystem interact with each other. In conventional electrical engineering, the assumption in U(1) electrodynamics of a flat local spacetime and equilibrium with the local vacuum exchange eliminates any

(v) exhibit negentropy, which every charge and every dipole in the universe already does {12, 68}.

Let us now see why Maxwellian EM systems of the second class *are not being designed and built as such,* even though they are permitted by the laws of physics and reinterpreted thermodynamics, and even though they are prescribed in the Maxwell-Heaviside equations before further truncation by Lorentz {161a, 161b}. We may use the Maxwell-Heaviside equations in vacuum, because the energy flow intercepted by the external circuit pours out of the generator terminals and fills all space surrounding the power lines. It is therefore a transmission-reception problem, prior to the "potentialized current-transport" problem in the circuit.

2.2 Lorentz Regauging of the Maxwell-Heaviside Equations

2.2.1 Introductory remarks.

The domain of Lorentz's symmetrically regauged equations is only a small subset of the domain of the Maxwell-Heaviside equations they replace. Indeed, the later Lorentz {161a, 161b} symmetrical regauging discards an entire class of Maxwellian systems permitted by nature and by the Maxwell-Heaviside equations before they are symmetrically regauged.

Lorentz's symmetrically regauged EM equations discard all Maxwell-Heaviside systems that are far from equilibrium in their energetic exchange with the active vacuum. They retain only that subset of the Maxwell-Heaviside theory wherein the system being described is in forced equilibrium {162} in its exchange with its active vacuum environment. Hence the present Lorentz-Maxwell-Heaviside theory, by which EM circuits and electrical power systems are designed, produces only systems in self-forced equilibrium with the active vacuum, *specifically during the symmetrically regauging discharge* of the circuit's excitation energy. The closed current loop circuit in fact discharges half its collected energy to destroy the source dipole in the generator, thereby destroying that dipole's extraction from the vacuum and furnishing of the energy flow pouring out of the terminals. These circuits kill themselves faster than they can power the load, and they use more energy to kill the energy flow from the vacuum than they use to power their load. Hence all present EM systems

disequilibrium of the system with its active environment, hence arbitrarily discards all Maxwellian disequilibrium systems - precisely those that can exhibit COP>1.0.

rigorously conform to classical equilibrium thermodynamics, and exhibit a coefficient of performance (COP) of COP<1.0 since any real system also has losses {163}.

2.2.2 Technical regauging of the Maxwell-Heaviside equations {164}. For asymmetrical-regauging (A-regauging) considerations, we are speaking of A-regauging the potential energy in and around a circuit. We include not only the Poynting energy flow component that is diverged into the circuit conductors, but also the remaining Heaviside *nondiverged* energy flow in space surrounding the conductors. This means that the energy is in field energy (E-field and B-field) form, both overtly as ordinary EM fields and covertly or "infolded" inside the corresponding scalar potentials {165, 166, 167, 168}, or both. Consequently, we must analyze Maxwell's equations as we would for radiating energy, rather than employ only the $\mathbf{j}\phi$ circuit analysis conventionally utilized, where the collected energy is sluggishly transported by the Slepian vector $\mathbf{j}\phi$. We show in this book that asymmetrical self-regauging (ASR) allows permissible overunity operation of electromagnetic engines and devices {169}, if other requirements are fulfilled also.

In Gaussian units, Jackson {170} shows that Maxwell's four equations (vacuum form) can first be reduced to a set of two coupled equations in the $(\mathbf{A}, \Phi)$ representation as follows:

$$\nabla^2\Phi + \frac{1}{c}\frac{\partial}{\partial t}(\nabla \bullet \mathbf{A}) = -4\pi\rho \qquad [15]$$

$$\nabla^2 \mathrm{A} - \frac{1}{\mathrm{c}^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla(\nabla \bullet \mathbf{A} + \frac{1}{\mathrm{c}}\frac{\partial \Phi}{\partial t}) = -\frac{4\mathrm{p}}{\mathrm{c}}\mathbf{J} \qquad [16]$$

The result is two coupled Maxwell equations rather than four. Jackson shows that potentials A and Φ in these two equations are arbitrary in a specific sense, since the A vector can be replaced with $\mathbf{A'} = \mathbf{A} + \nabla\Lambda$, where A is a scalar function and VA is its gradient. The B field is given by $\mathbf{B} = \nabla\times\mathbf{A}$., so that the new B' field becomes

$$\mathbf{B'} = \nabla\times(\mathbf{A} + \nabla\Lambda) = \nabla\times\mathbf{A} + \mathbf{0} = \nabla\times\mathbf{A} = \mathbf{B} \qquad [17]$$

In other words, the B field has remained entirely unchanged, even though the magnetic vector potential has been asymmetrically changed. However, if no other change were made, then the electric field E would have still been changed because of the gradient $\nabla\Lambda$. In that case the net change would be asymmetrical, because one obtained a "free" E-field which could then do work on the system — either beneficially or detrimentally, depending upon the specific conditions, geometry, and timing. To prevent this excess "free" E-field from appearing, the electrodynamicists simultaneously and asymmetrically regauge (transform) the scalar potential Φ so as to offset the E-field change due to the regauging of equation [15]. In short, they also change Φ to Φ'„ where

$$\Phi' = \Phi - \frac{1}{c}\frac{\partial \Lambda}{\partial t} \qquad [18]$$

With that additional change, now the net E and B fields remain unchanged {171, 172}, even though both potentials have changed and the fundamental stored energy of the system has changed, as has the stress of the system. Unchanged force fields just mean that only a set of zero-summation forces (a zero-summed stress system) has been utilized to effect the change in potential energy. It also means that the net summation of the two *asymmetrical* regaugings has been entirely *symmetrical* {173}.

Jackson points out that, conventionally, a set of potentials $(\mathbf{A}, \Phi)$ is habitually and *arbitrarily* chosen by the electrodynamicists such that

$$\nabla \bullet \mathbf{A} + \frac{1}{c}\frac{\partial \Phi}{\partial t} = 0 \qquad [19]$$

This *net symmetrical* regauging operation creates a new and simpler Maxwellian system, with different system stress and different system potential energy. It successfully separates the variables, so that two inhomogeneous wave equations result. This procedure yields a new and simplified system, and the new Maxwell's equations for it are as follows:

$$\nabla^2 \Phi - \frac{1}{c^2}\frac{\partial^2 \Phi}{\partial t^2} = -4\pi\rho \qquad [20]$$

$$\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\frac{4\pi}{c}\mathbf{J} \qquad [21]$$

The two previously coupled Maxwell equations [15] and [16] (potential form) have been replaced by equations [20] and [21], to leave two much simpler inhomogeneous wave equations, one for Φ and one for A. These are *new equations for a new system!*

Of course, this arbitrary *net symmetrical* regauging (let us use the term ***S-regauging)*** is quite useful for purposes of simplifying the theory and for easing calculations. But its unquestioning and rather universal usage has arbitrarily eliminated the freedom of the system designer to *asymmetrically* regauge the system's potentials, and *use* the resulting excess free force fields to change the stored energy in the system without the operator performing extra work upon the system. So we advance the condition for violating this S-regauging, violating the exclusion of net A-regauging, and violating the "frozen gauge" process as

$$\nabla \bullet \mathbf{A} + \frac{1}{c}\frac{\partial \Phi}{\partial t} \neq 0 \qquad [22]$$

Any regauging of the potentials that complies with equation [22] will *a priori* produce one or more net excess forces in the system, as well as a change in potential energy of the system. By controlling the regauging, the system designer is then able to control where, how, and when these excess forces appear, and whether they enhance the system's operation or hinder it. These net forces can then be used to perform work and dissipate the excess potential energy taken on in the asymmetrical regauging. That is what we do when we ourselves add potential (and potential energy) to an EM system to enable it to do work! *If we always have to asymmetrically regauge the system to get it to do subsequent work, why do we not just let the system asymmetrically regauge itself so we get the input energy freely and also get the resulting work freely?* The gauge freedom axiom in quantum field theory assures us that nature will indeed freely change the potential energy of any system for us, if we but arrange it and permit it.

In short, we have had a gauge freedom principle for some time, which guarantees us that COP>1.0 EM systems are permitted and possible. Yet we have failed to realize it, or take any advantage of it. So we continue to pay to asymmetrically regauge (potentialized) all electrical power systems, and to insure that we have to continue to do it, we specifically design the systems so they will then re-enforce Lorentz's symmetrical regauging condition.

This is another of those "inexplicable aberrations of the scientific mind" referred to by Nikola Tesla!

2.2.3 A Humorous Comment but an Exact Analogy

Again see Figure 1-1 in Chapter 1, and see Figure 2-3. In Figure 2-3, we show how Lorentz's integration trick {174a, 174b} discarded the huge nondiverged Heaviside component of energy flow outside the conductor, while retaining the small Poynting component that strikes the surface charges and gets diverged into the conductor to power the electrons. In justifying his integration trick, Lorentz stated that all the rest of that wasted Heaviside energy flow

See Panofsky & Phillips
Classical Electnqtv and Magnetism. 2nd edn
Addison Wesley 1962.
p 178-181

The Lorentz procedure arbitrarily discards the enormous Heaviside component that misses the circuit entirely and is wasted This results in a non sequitur of first magnitude in EM energy flow theory

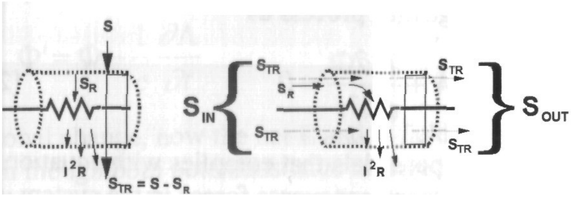

2-3a. Lorentz surface integration

2-3b. Actual S in and S out.

Figure 2-3 Lorentz's integration trick to discard the enormous Heaviside non-diverged energy flow component

The Heaviside component is often 10 trillion times the Poynting component but is simply wasted in ordinary single-pass energy flow circuits

Heaviside component (nondiverged)

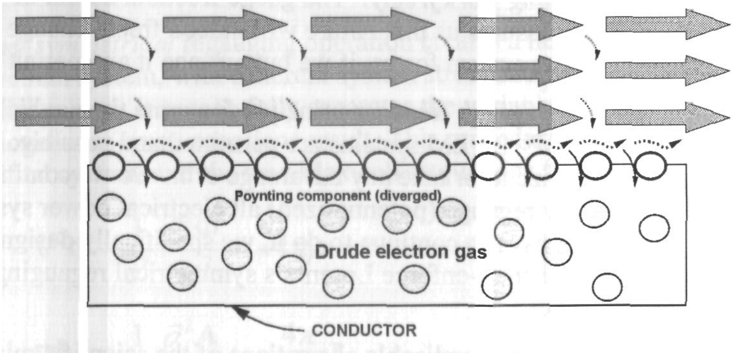

Figure 2-4 Heaviside nondiverged and Poynting diverged energy flow components.

(Figure 2-4) was "physically insignificant" (his term) because it did not strike the circuit, was not intercepted, and did not power anything. Well, Maxwell's theory is a material fluid flow theory, so let us evaluate Lorentz's justification in a fluid flow analogy.

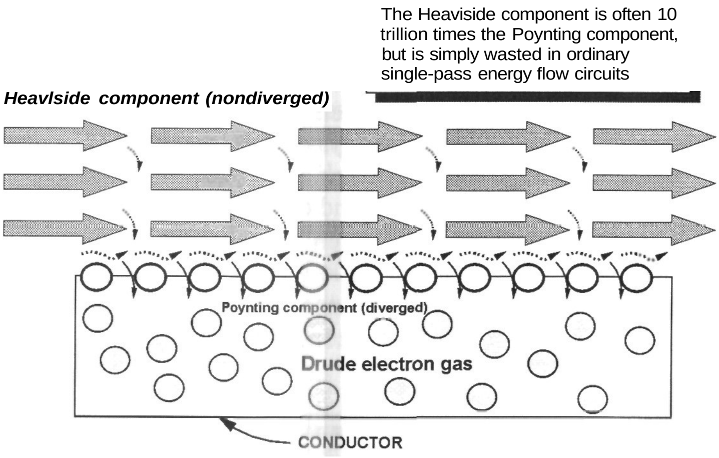
The Heaviside component is often 10 trillion times the Poynting component, but is simply wasted in ordinary single-pass energy flow circuits
Heavlside component (nondiverged)
Poynting component (diverged)
Drude electron gas
CONDUCTOR

Figure 2-4 Heaviside nondiverged and Poynting diverged energy flow components.

Contrast Lorentz's statement if applied to the ocean wind on a sailing ship, where the wind represents the total energy flow and the ship represents a single Maxwellian system under consideration.

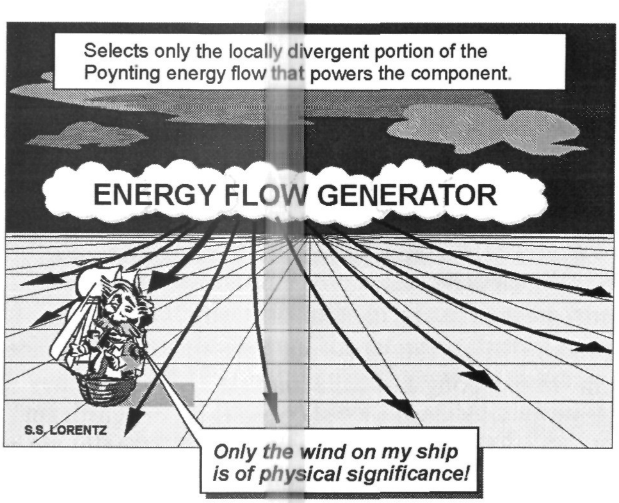
Selects only the locally divergent portion of the Poynting energy flow that powers the component.
ENERGY FLOW GENERATOR
S.S. LORENTZ
Only the wind on my ship is of physical significance!

Figure 2-5 Lorentz in his sailing ship in an ocean wind, with no other ships around.

As shown in Figure 2-5, Lorentz's argument is analogous to stating that the huge component of the wind *that does not* strike the ship's sails is "physically insignificant" — in other words, it has no physical use. Now that is true for the single ship under consideration, as shown in Figure 2-5. None of that wind component that misses the ship's sails does anything at all to power *that ship* (that individual system). Hence it is indeed physically insignificant *to that ship itself.* Of course, if we flare the sails wider, the ship will catch more wind and develop more power. That is an analogy to the Bohren experiment, which resonates the charges (the "sails" intercepting the incoming EM energy flow "wind") so that they sweep out a greater geometrical cross section. Hence 18 times as much energy is caught, as when the charges are not resonated (the sails are not flared to a greater area).

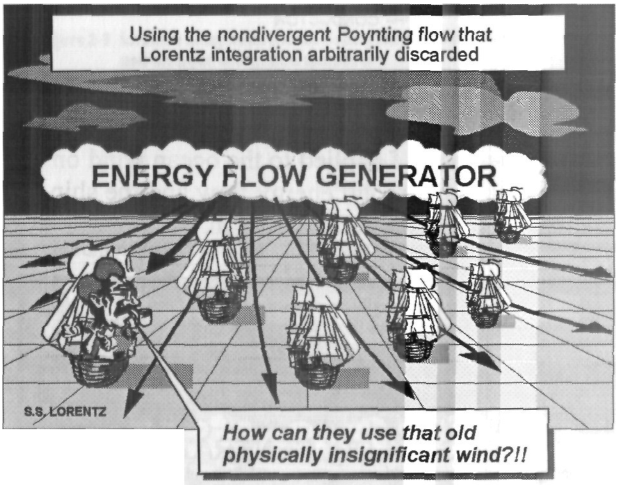
Using the nondivergent Poynting flow that
Lorentz integration arbitrarily discarded
ENERGY FLOW GENERATOR
S.S. LORENTZ
How can they use that old
physically insignificant wind?!!

Figure 2 6 Lorentz in his sailing ship in an ocean wind, with many other ships alongside.

However, now see Figure 2-6. If we put some more ships (some more systems) with "intercepting and collecting sails" in there, they will be powered quite well with some of that "physically insignificant" wind. So Lorentz (following earlier work by Ludwig Lorenz) really did advance and impose upon electrodynamics a non sequitur of first rank.

And the wind-fluid analogy helps to "lighten things up" a bit here and inject a little good-natured humor!

2.2.4 Discussion

Equation [22] has now become the first *necessary but not sufficient* principle of asymmetrical self-regauging (ASR) for overunity electromagnetic engines and devices. That is, we are now free to A-regauge in a *nonconservative* field manner, where excess net force fields automatically appear and remain for our use in the system. All that remains is for the designer to:

(i) evoke "nonconservative" (asymmetrical) regauging in accord with equation [22];

(ii) have the system assembly in disequilibrium with its energetic external environment, so that the system itself performs this nonconservative A-regauging in some sector or part {175} of its normal operation,

(iii) insure that any "free force" that is produced and allowed to do work upon the system will do some positive work upon the system, so as to increase its collected energy, and

(iv) assure that the "free" forces that are otherwise produced by the circuit design (this includes, e.g., the back emf) do not seriously degrade the system (i.e., do not substantially reduce its energy available to dissipate as work in the load). Usually this means that, contrary to the standard closed current loop circuit, the circuit must not dissipate half its collected Poynting energy to dissipate the charges in the source dipole and destroy it. *These four considerations we now advance as the primary or master principles of asymmetrical self-regauging of electromagnetic engines and devices for permissible overunity efficiency.* We will meet additional principles in a later chapter when we discuss what the form of the energy output of an overunity EM system is, and the considerations for close-looping a COP>1.0 EM system into stable disequilibrium and continuous self-powering.

It is understood that the asymmetrical regauging portion of the circuit or system must be a broken symmetry in the energetic exchange between the active vacuum environment and that portion of the system. We exclude, g., solar cell-powered systems, windmill-driven generators, etc.

For simplicity, the system designer may wish to A-regauge only one potential, producing a desired "free" force F_1 to work on the system and increase its energy, while allowing any other extraneous force F_2 to appear, but just "bottling up" that additional force F_2 so that it cannot do any work at all on the system and therefore cannot degrade the system's operation. A most useful ASR corollary that immediately suggests itself is to simply A-regauge onto that force operation that normally produces the back-drag in a system or device, to either *eliminate* the back-drag altogether or reverse its algebraic sign so that it becomes an *assisting* force instead of a *hampering* force {176}.

In overunity systems, this becomes a startling and unexpected exercise in novel new phenomenology effected by tempic potential differences, tempic force, etc. It involves not only the system, but also the entire *supersystem* consisting of the physical system, its associated altered local vacuum, and its associated curved local spacetime. See Chapter 9 for our expose of these phenomena, and how one must transduce[54] the effects of a very novel kind of tempic force back emf, if the COP>1.0 system is to remain in disequilibrium with the active vacuum and thus maintain its stable COP>1.0 operation.

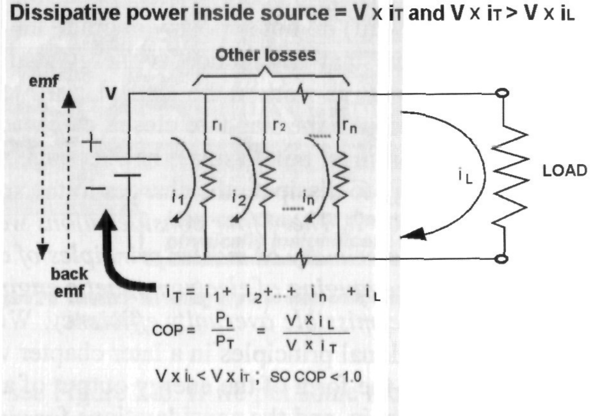
Dissipative power inside source = V x iT and V x iT > V x iL
Other losses
emf
V
+
−
r1
r2
rn
i1
i2
in
iL
LOAD
back emf
iT = i1+ i2+...+ in + iL
COP = PL/PT = V x iL / V x iT
V x iL < V x iT ; SO COP < 1.0

Figure 2-7 Why a conventional closed current loop circuit is limited to COP<1.0.

Again see Figures 2-3 and 2-4, and equations [20] and [21]. Electrodynamicists have previously *assumed away* the capability of work-free "refueling" of an electromagnetic system directly from the local vacuum {177}. They have deliberately designed their systems *to prevent self-evoking and positively utilizing asymmetrical self-regauging of their own potential energy during system excitation discharge.* (See Figure 2-7).

---

[54] A patent application on the process has been filed by John Bedini and the present author.

In fact, by forcibly pumping the ground electrons from the external circuit back up through the primary source dipole in the generator or battery, our closed current loop circuits self-enforce the Lorentz condition and require that precisely half the regauging energy freely intercepted and collected in the external circuit is used to destroy the source dipole providing the regauging energy. This destruction of the energy source is accomplished faster than the external load is powered. Hence all such systems with losses are COP<1.0 systems. For more than 100 years, this "self-destroying, suicidal system" is the only kind of electrical power system our energy scientists and engineers have designed and built. To inject a little humor, we can only describe this in Tesla's words[55] as *"one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history. "*

Thereby electrodynamicists have narrowed the inherent flexibility of the Maxwellian EM gauge theory model to an artificially symmetrized subset of Maxwell-Heaviside-Lorentz equations that utilize a single fixed gauge. This prohibits the free use of the system's vacuum energy freely received by the source dipole's broken symmetry in its violent exchange with the active vacuum. If the system increases its dipolarity (regauges freely), it simultaneously increases the speed at which it destroys the dipolarity! *Maxwell's equations themselves do not require this, and neither does Heaviside's truncation of them into the resulting Maxwell-Heaviside equations. Lorentz's symmetrical regauging of the Maxwell-Heaviside equations does require it.*

The Maxwell-Heaviside theory without Lorentz symmetrical regauging allows regauging asymmetrically to place the system in disequilibrium so that it freely receives and utilizes excess energy from the surrounding vacuum. Then it can again asymmetrically regauge by dissipating that freely received energy in an external load. *The theory specifically does not require destroying the primary source dipole as is done in the arbitrary closed current loop circuit between load and primary source dipole in the generator or battery.*

" Nikola Tesla, "The True Wireless, *Electrical Experimenter,* May 1919.

By arbitrary habit, electrodynamicists have discarded the Maxwell-Heaviside major EM overunity mechanism: *Asymmetrically self-regauging accomplished by the system itself, without requiring external work being done on the system by the operator,* in order to freely gather in and collect excess potential energy in the system from its broken symmetry with its active vacuum environment. In this self-A-regauging, associated *excess net force* does appear, whereupon that excess collected/stored regauging energy can then be used to "drive" the forces against load resistance and system losses to freely power the system and its loads {178}. They have also discarded the major overunity corollary whereby the designer is free to utilize *system self-regauging asymmetrically so that the normal back-drag or back-emfwork — usually done upon the system in its closed current loop configuration to reduce its kinetic energy — vanishes or reverses sign and becomes beneficial work done in the opposite sense, to freely increase the energy ofthe system.*

The failure to do this is bluntly met with, when one does succeed in producing a $COP \gg 1.0$ system momentarily by suddenly and very powerfully breaking the system's equilibrium in its vacuum energy exchange. Nature then exhibits a very novel kind of Dirac sea hole reaction (powered by tempic force generated in the locally curved spacetime and resulting effects generated in the altered semiconducting vacuum). This reaction sharply jerks the system (or the system plus its external power supply) back into equilibrium and $COP<1.0$ operation. In Chapter 9, apparently for the first time in the literature[56] we present this novel mechanism exhibited by nature as a new kind of "tempic back emf phenomenon in systems in disequilibrium. We also explain how to handle the problem and in general how to use it for "freeze-framing" and stabilizing the system's disequilibrium operation in sustained $COP > 1.0$ configuration while powering itself and its load.

All that is required for "self-powering" $COP \gg 1.0$ EM engines directly powered by the vacuum, is that (1) initially the engine must asymmetrically self-regauge to produce $COP \gg 1.0$, and (2) positive *transduced* feedback (from the output) of a very special and previously unknown kind — where negative energy is transformed into normal electron current and energy — must be utilized to its input, in a clamped and controlled fashion {179}. The clamped positive feedback can be designed and applied, once the fundamental engine is designed for self-regauging and overunity efficiency, and once the novel feedback and transformation requirements are understood and incorporated. The remaining excess output can then be utilized to power loads, while the engine seemingly "powers itself."

---

[56] The basis for it, however, is in Dirac's electron theory published in 1930, nearly three quarters of a century ago. See P. A. M. Dirac, "A theory of electrons and protons," *Proceedings of the Royal Society of London, Series A,* 126(801), Jan. 1, **1930, p. 360-365.**

We stress, of course, that this system is just analogous to the windmill, and it continuously receives a "free energy wind" from its active environment. The conservation of energy law is obeyed at all times, as are physics and thermodynamics.

In a later chapter we present the concepts and processes for these novel phenomena involved in close-looping a COP>1.0 system for self-powering operation.

2.3 The Suicidal Closed Unitary Current Loop Circuit

The present Lorentz-Maxwell-Heaviside theory, by which EM circuits and electrical power systems are designed, produces only systems in net usable equilibrium with the active vacuum, specifically during the symmetrically regauging discharge of the circuit's excitation energy (Figure 2-7). Hence all present EM systems designed by that model rigorously conform to classical equilibrium thermodynamics, and exhibit a coefficient of performance (COP) of COP<1.0.

Since the Maxwell-Heaviside equations do not restrict the designer to systems in equilibrium with their active environment, some ubiquitous characteristic of the standard circuit approach must be self-enforcing the Lorentz symmetrical condition as far as using any excess free potential energy of the system to freely power loads. That is the closed-loop circuit, as we have discussed. Let us now illustrate what goes on in such a circuit.

Taking into account Figure 2-4, Figure 2-8 shows a modified diagram of what is actually happening between the vacuum and the dipole to power the circuit. Figure 2-9 shows how a pair of conductors constitutes an extended series of parallel dipoles. Once the dipole is formed by dissipating some chemical energy in the battery, the dipole extracts energy from the surrounding vacuum and pours it out of the battery terminals, producing an energy flow filling space around the conductors of the external circuit. Only a very small portion (Figure 2-4) of the very large energy flow is intercepted by the external charges in the conductors, and diverged into the conductors to power up the Drude electrons.

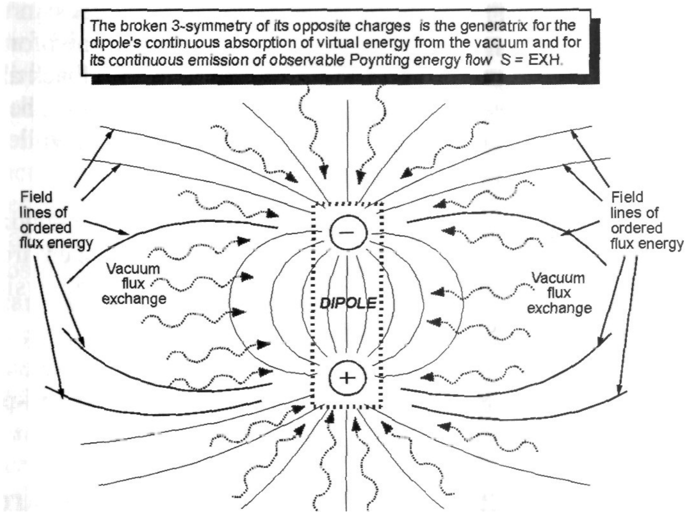

The broken 3-symmetry of its opposite charges is the generatrix for the dipole's continuous absorption of virtual energy from the vacuum and for its continuous emission of observable Poynting energy flow S = EXH.

Field lines of ordered flux energy
Vacuum flux exchange
DIPOLE
Field lines of ordered flux energy
Vacuum flux exchange

Figure 2-8 Virtual photon flux exchange intensity **with the charges of a** dipole.

As can be seen from Figure 2-7, half of the energy collected upon the internal electrons to drive them as $j\phi$ is expended as work required to pump the spent electrons from the ground return wire back through the source dipole in the battery. This scatters the charges of the dipole (in this case, by changing the chemistry of the battery plates) and destroys the dipole. The other half of the collected energy is dissipated in the external circuit's losses and in the load.

Then some more chemical energy has to be dissipated by the battery to perform work upon the internal charges and restore the dipole (in this case, to change the plate chemistry of the battery plates). So even if this process is 100% efficient, the battery has to dissipate as much chemical energy as was electrically dissipated upon those former dipole charges to scatter them and destroy the dipole.

In Figure 2-7, we show the small internal resistance of the conductors, as well as the load resistor. Half the energy dissipated by $j\phi$ is expended just to forcibly ram the spent electrons back through the back emf of the source dipole. The other half is expended in the circuit's internal losses and in the load resistance. Therefore, less energy is expended as usable work in the resistive load than is expended to destroy the dipole and cut off the flow of free EM energy from the vacuum.

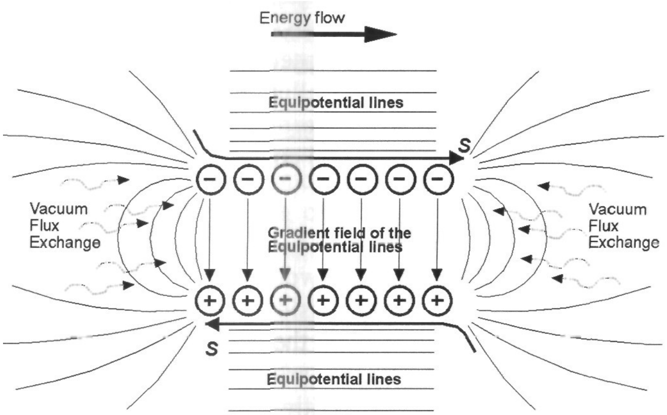

Figure 2-9 Paired conductors as an extended set of dipoles.

The battery has to input as much energy to restore the dipole as was used to destroy it. The battery has to expend at least as much chemical energy as equals half the entire energy originally collected in the external circuit and expended as $\int j(t)\phi(t)dt$. It must expend chemical energy $CE \geq \frac{1}{2} \int j(t)\phi(t)dt$. The energy RE dissipated as work in the load resistor R is $RE < \frac{1}{2} \int j(t)\phi(t)dt$. Thus CE>RE, and this system is forced to exhibit

$$COP \; ^{\circ} \; \frac{RE}{CE} = \frac{< \frac{1}{2} \grave{o}\, j(t) f(t) dt}{^{3} \; \frac{1}{2} \grave{o}\, j(t) f(t) dt} \; \text{Þ} \; COP < 1.0 \qquad [23]$$

As can be seen, this system symmetrically regauges itself during its excitation discharge, so that it self-enforces the Lorentz condition during that period. The circuit self-enforces COP<1.0, even though all the energy — both the energy powering the external circuit and also destroying the internal source dipole in the battery to enforce the Lorentz condition during discharge — is freely received from the active vacuum via the broken symmetry of the source dipole,[57] once established, in its energetic exchange with the active vacuum.

---

[57] The broken symmetry of opposite charges is one of the broken symmetries predicted by Lee, {13a} and by Lee, Oehme, and Yang, {13b}. Broken symmetry was proven experimentally in early 1957 by Wu *et ai,* {14}. So profound were the implications for physics that the Nobel Prize was awarded to Lee and Yang in latter 1957, the same year! E.g., see T. D. Lee, "Weak Interactions and Nonconservation of Purity," Nobel Lecture, Dec. 11, 1957 in *T. D. Lee, Selected Papers,* Gerald **Peinbcrg,** Ed., Birkhauscr, Boston, 1986, Vol. 1, p. 32-44.

The problem is that *the source dipole is destroyed faster than the load is powered,* as shown in Figure 2-7. Because of the inexplicable scientific pre-occupation with the conventional closed current loop circuit, the planet and biosphere have been increasingly polluted for more than a century, by burning more hydrocarbons, building more dams, using more nuclear fuel rods, etc. to provide ever more electrical energy.

2.4 In Summary

COP>1.0 operation in EM systems via work-free *asymmetrical* regauging has been inherent in the Maxwell-Heaviside equations for more than a century. Since Lorentz first did so circa the latter 1880s, electrodynamicists have just conveniently and arbitrarily *assumed away* COP>1.0 systems by (i) artificially limiting the theory to only that fraction of Maxwell-Heaviside systems in usable net force field equilibrium in their vacuum exchange during their excitation discharge, and (ii) using the ubiquitous closed current loop circuit to self-enforce the Lorentz symmetry condition during that excitation discharge. In other words, our engineers have only been building systems which self-enforce the overly restricted Lorentz-regauged theory.

Electrodynamicists have never seriously tried to build overunity circuits, nor until recently have they explored the discarded Maxwellian systems by removing the Lorentz regauging condition and examining EM energy from the vacuum.[58]

---

[58] Under the directorship of Dr. M.W. Evans, the Alpha Foundation's Institute for Advanced Study (AIAS) has been gradually developing the physics of EM systems far from equilibrium with their active environment (the local active vacuum and local curved spacetime). Some of the papers of interest are: M. W. Evans et al., "Derivation of the Lehnert field equations from gauge theory in vacuum: Space charge and current," *Found. Phys. Lett,* 13(2), Apr. 2000, p. 179-184; Evans et al., {38b}, {38c}, {113b}, {233a}, {233b}, {233e}; {233g}, {233j}; Evans et al., "Schrodinger Equation with a Higgs Mechanism: Inherent Vacuum Energy," *Found. Phys.* (in review); — "Vacuum Energy Flow and Poynting Theorem from Topology and Gauge Theory," *Physica Scripta* (in review); — "Energy from the Vacuum," *Physica Scripta* (in review); — "Some Notes on Asymmetric Regauging", *J. New Energy* 4(3), Special Issue, Winter 1999, p. 325-326; — "Inconsistencies of the Maxwell-Heaviside Theory of Electrodynamics: The Aharonov-Bohm Effect," *J. New Energy,* 4(3), Special Issue, Winter 1999, p. 236-240; — "Spontaneous Symmetry Breaking as the Source of the Electromagnetic Field," *Found. Phys. Lett.* (in press), A monumental work. *Modern Nonlinear Optics,* Second Edition, 3 vols., ed. M. W. Evans, Wiley, 2001 gives a broad coverage of higher group symmetry electrodynamics and energy from the environment (the local active vacuum and the

Every electrical power system ever built is and has been powered by electrical energy freely extracted from the seething vacuum exchange with the source dipole in the system, due to that dipole's broken symmetry. But the engineers and scientists have designed and built only those electrical systems which self-enforce Lorentz's symmetrical regauging.

The time for energy scientists and engineers to correct this procedure and cease this inane practice is long overdue, following the direction pointed out by the AIAS.

---

local curved spacetime). Some 60 AIAS papers are published as a Special Issue of **the** *J, New Energy,* 4(3), Winter 1999. More than 100 AIAS papers are also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/.

Chapter 3

Giant Negentropy, Dark Energy, Spiral Galaxies and Acceleration of the Expanding Universe

"... the discoveries made in 1957 established not only right-left asymmetry, but also the asymmetry of the positive and negative signs of electric charge. In the standard nomenclature, right-left asymmetry is referred to as P violation, or parity nonconservation. The asymmetry between opposite signs of electric charge is called C violation, or charge conjugation violation, or sometimes particle-antiparticle asymmetry." [T. D. Lee]. {180}

"The ends of a dipole, being oppositely charged, exhibit charge conjugate symmetry violation. An 'isolated' charge, considered with its clustering virtual charges of opposite sign, also exhibits symmetry violation." [T. E. Bearden, this book, below].

"A generally acceptable, rigorous definition of radiation has not as yet been formulated." ... "The recurring question has been: Why is it that an electric charge radiates but does not absorb light waves despite the fact that the Maxwell equations are invariant under time reversal?" [B. P. Kosyakov]. {181}

"It [the energy transfer flow] takes place, in the vicinity of the wire, very nearly parallel to it, with a slight slope towards the wire... . Prof Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference of a quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature of the electric field in the vicinity of a wire supporting electric current. The lines of electric force are nearly perpendicular to the wire. Their departure from perpendicularity is usually so small that I have sometimes spoken of them as being perpendicular to it, as

they practically are, before I recognized the great physical importance of the slight departure. It causes the convergence of energy into the wire. " [Oliver Heaviside]. {182}

"The funny thing about the dark matter is it doesn't show up in light, X-rays, gamma rays, infrared, ultraviolet, radio waves, or submillimeter. We're running out of places to look for it." [William Keel]. {183}

"... baryonic, ordinary matter — the stuff of stars and of people — makes up just over 4% of the energy and matter in the universe. ...about 30% of the stuff in the universe is dark [unobserved] matter. The remaining two-thirds, theorists believe, is a mysterious 'dark energy' or 'quintessence' — a large-scale antigravity-like effect that is making the universe expand ever faster.... " [Charles Seife]. {184}

"And yet, curiously enough, we do not know exactly what charge is, only what it does. Or, equally significantly, what it does not do."[M. P. Silverman]. {185}

*"**Charge** is the ongoing circulation of EM energy flow between the time and 3-space domains, as seen by the observer. Negative charge is the ongoing absorption of EM energy input from the time domain (from ict), transduction of the absorbed energy into 3-space EM energy, and re-emission of the EM energy in all directions in 3-space. **Positive charge** is the ongoing absorption of EM energy input from 3-space, transduction of the absorbed energy into the time domain, and re-emission of the EM energy in the time domain. If we wish, we may also consider the positive charge as radiating negative energy, while the negative charge radiates positive energy. Once we consider vacuum polarization, we recognize that this circular flow process is involved at each "dipolar point" of the polarized vacuum. "* [T. E. Bearden, this book, below].

"The matter density of the Universe is extremely low. On average there might be one atom per cubic metre of space. The major constituent of the Universe is believed to be some kind 'dark energy', which is pushing the Universe apart. " [Matthew Colless]. {186}

3.1 Broken 3-Symmetry of the Dipole

To recap from Chapters 1 and 2: In 1956, Lee and Yang {187} strongly predicted broken symmetry in physics, suggesting experiments to detect it in the weak interaction. In early 1957, Wu *et al.* {188} experimentally demonstrated broken symmetry, thus furnishing the experimental substantiation of broken parity in the weak interaction. The emergence of broken symmetry was such a dramatic change to the view of physics, that in December of the same year, 1957, Lee and Yang were awarded the Nobel Prize.

One of the asymmetries shown by Lee *et al.* is the asymmetry of opposite charges. That means that a dipole, which consists of slightly separated opposite charges, exhibits broken symmetry in the seething vacuum virtual particle flux. In short, a dipole — any dipole or dipolarity — must absorb or receive virtual photon energy from the active vacuum, integrate at least some of it to observable state, and re-radiate that integrated EM energy in observable form in 3-space.

We stress: *The ends of a dipole, being oppositely charged, exhibit broken symmetry. An 'isolated' charge, considered with its clustering virtual charges of opposite sign, is a set of composite dipoles and also exhibits broken symmetry.*

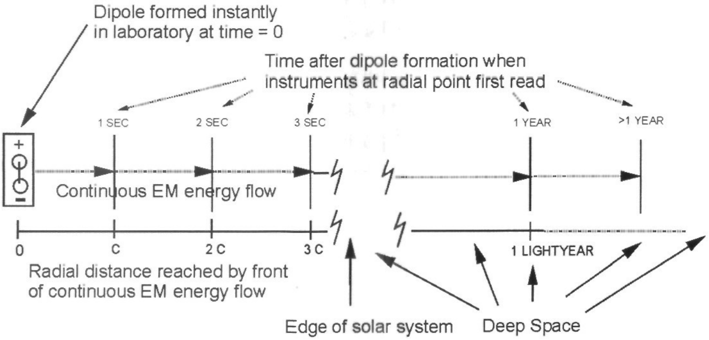
Dipole formed instantly in laboratory at time = 0
Time after dipole formation when instruments at radial point first read
1 SEC
2 SEC
3 SEC
1 YEAR
>1 YEAR
+
−
Continuous EM energy flow
0
C
2 C
3 C
1 LIGHTYEAR
Radial distance reached by front of continuous EM energy flow
Edge of solar system
Deep Space

Figure 3-1 The source dipole experiment, demonstrating continuous energy flow.

Let us now consider a gedanken experiment, as shown diagrammatically in Figure 3-1, to illustrate Kosyakov's quotation cited above. At a point in the laboratory corresponding to the origin of a 3-dimensional frame, a dipole is produced instantly by separating charges, such as in a flash-charged capacitor. Along two radial lines in 3-space from that origin, we have previously placed perfect EM field and potential detectors, at distances *ct* apart, where *t*= 1, 2, 3, 4, *n*,... and so on in seconds, with the radials extending across the universe, and where c is the speed of light. We neglect the transient pulse that momentarily occurs at each instrument package, since we are examining steady state conditions.

At the end of one second, the instruments at *t* = 1 instantly read, and those readings remain thereafter. One second later, the instruments at *t* = 2 instantly read, and those readings remain thereafter. And so on with each of the instruments progressively farther from the origin reading at the proper light-time travel, and with all field readings then *remaining continuously* thereafter.

This gedankenexperiment demonstrates that what is emitted from the dipole, from its moment of creation at the origin, is a continuous flow of EM energy in all directions in 3-space.[59] This energy continuously flows from the dipole, indefinitely or as long as the dipole physically exists. Other than paying once to initially separate the charges and make the dipole, we have neither added nor input any additional energy. Yet energy continues to flow out from the dipole in all directions, unceasingly, as shown by the continued of the reading on each field energy density instrument, once the leading edge of the energy outpour reaches that instrument and passes it by. In short, the experiment shows that it is not a "pulse" that passes each instrument, but a continuous and sustained flow of EM energy, remaining steadily thereafter so long as the dipole remains.

One year after making the dipole, the instruments at one light year distance suddenly read, and their readings remain. *The energy density of a volume of space one light year in radius has now been changed, and that change in spatial energy density continues to spread outwards at light speed.*

[59] More exactly, what is observed is a spreading change in the energy density and point-like polarization of the vacuum itself, and of the local curvature of spacetime. At each point, once the change of energy density and ST curvature occur, the energy density and ST curvature remain fixed in a new equilibrium condition, steadily fed by the energy radiating from the dipole. See our point made in {181}. The discovery of giant negentropy {12} requires a change in how we view the very process of energy propagating in space. Now it has become a quite different *propagation of energy circulation between time and space domains at each affected point in space.*

Yet there has been no input of *observable 3-space* EM energy flow to the dipole whatsoever, as our instruments also verify. All observable energy flow in 3-space is *outward* from the origin where the dipole is located.[60]

We have shown that the *observable* EM energy flow from a source dipole is not conserved in 3-space, which is consistent with a broken 3-symmetry since symmetry implies conservation. Either we must discard the conservation of energy law altogether, or it must be conserved in 4-space since we have proven the dipole does not conserve energy in 3-space. If we assume that the experiment must conform to the conservation of energy law, then 4-symmetry EM energy flow must exist since 3-symmetry energy flow is broken. The energy steadily pouring out of the dipole in all directions in 3-space must therefore be received by the dipole *charges from the time-domain,*[61] assuming we are working in Minkowski 4-space. In particle physics, this is essentially the same as receiving the energy from the virtual domain, and therefore is consistent with the proven broken symmetry of every dipole and dipolarity.

The energy flow is indeed received from the time domain, as we have argued {189} in a wave manner consistent with (i) quantum field theory as shown by Mandl and Shaw in the photon model {190}, (ii) broken charge conjugation symmetry of particle physics using the virtual photon and observable photon concepts, and (iii) a re-interpretation of Whittaker's decomposition of the scalar potential {191} when applied to the scalar potential between the ends of the dipole {12}.

We state the obvious: A dipole is a Maxwellian system. This gedanken experiment also proves by demonstration that a Maxwellian system — e.g., a dipole — is permitted by the laws of nature, physics, and thermodynamics to output more energy than the operator inputs, since we input nothing to the source dipole[62] once we have made it. After paying to produce the dipole initially, the operator has made no other energy input to the dipole at all, yet the dipole will continue to pour out observable energy in 3-space for the next 14 billion years if the dipole remains and is not destroyed. In addition, every positive charge and every negative charge in the universe form a dipole, regardless of the separation distance between them. That dipole is also performing the gedanken experiment process. Every dipolar EM circuit contains innumerable such dipoles — e.g., across "open paths" between the various points in the circuit, as shown by Kron {192}. Kron {193} stated:

---

[60] Obviously, we are utilizing the prevailing notion of the flow of energy through 3-space, which is actually a non sequitur. No 3-space observable entity can "move" in 3-space alone; it requires time to move, and so any movement is in 4-space a priori. We previously discussed this and the *dldt* operator being applied by the observation process. But for simplicity, we are using the conventional view here, so the negentropy of the dipole can be easily grasped.

[61] Again, we have stated the positive time case only. In the spreading special 4-circulation, energy appears in 3-space at each point dipole from the time domain (courtesy of the action of the negative charge of the dipole), and then returns from 3-space back to the time domain (courtesy of the action of the positive charge, which is a time-reversal of the action of the negative charge).

> *"... the missing concept of "open-paths " (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of two nodes. (Previously - following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

By considering all dipoles in the universe, one has in fact finally defined Kron's open path. One has also defined the master mechanism responsible for the incredible EM energy density of the vacuum — the so-called "zero-point EM energy". Overall, every electrical circuit thus has innumerable energy flow broken 3-symmetries in it, with respect to exchange of EM energy flow between the time and 3-space domains. To sum all these broken 3-symmetries to an overall net 3-symmetry during the discharge of the free excitation energy of the circuit, obviously requires some specific macroscopic cohering function in the circuit. Moreover, one exists. Specifically, the closed current loop between the external circuit attached to the generator and the internal source dipole created between the terminals of the generator, is what enforces the circuit's net 3-symmetry and Lorentz symmetrical regauging. That is an arbitrary man-made condition, and not a law of nature, physics, or thermodynamics.

---

[62] Or to a charge with its clustering virtual charges of opposite sign, if we redo the experiment with a single "isolated" observable charge and consider it more exactly as a set of composite dipoles, each of which exhibits broken symmetry in the vacuum virtual particle flux.

By our gedanken experiment's demonstration and the existence of innumerable disequilibria between the active environment[63] and the system in an ordinary EM circuit, it follows that a Maxwellian circuit must be *capable* of exhibiting a coefficient of performance (COP) of COP>1.0. It is a giant collection of open dissipative systems, far from equilibrium in the vacuum exchange.

That concept of course is contrary to the entire present mindset of electrical engineering and Maxwell-Heaviside-Lorentz classical electrodynamics. Those disciplines merely continue to accept the source charge as if it freely created from nothing all that energy in its associated fields and potentials reaching across the universe. Ironically, so long as the giant negentropy of the common dipole and its disequilibrium with its active environment (the local active vacuum and the local active curvatures of spacetime) are unaccounted, engineers and scientists will continue to produce the inane Lorenz/Lorentz-regauged power systems they produce today and have produced for more than a century. Also, they will continue — wittingly or unwittingly — to implicitly prescribe that every charge and dipole in the universe violates the conservation of energy law and is thus a *perpetuum mobile* of the grossest kind: a system which massively and continuously creates EM energy from nothing.

We now point out something deeper and very important. In modern physics, the *observed* positive charge is regarded as the time reversal of the *observed* negative charge. See Figure 3-2. If the negative charge is responsible for EM energy flow from the time-domain into 3-space, then *as observed* the positive charge must represent a time-reversed situation. We observe time reversal in the charge domain as a parity reversal and a charge reversal, but with the energy remaining positive. In other words, one imagines that one records a process on videotape, and then runs the video backwards to observe the "time-reversed" situation. When observed, both the sign of the positive charge and its direction of motion appear to us to be reversed from the negative case.

---

[63] In paragraph 3.10 and in Chapter 9 we will introduce — more deeply — the notion of the supersystem, consisting of three parts: (1) the system and its dynamics, (2) the active vacuum and its dynamics, and (3) the active spacetime curvatures and their dynamics. All three components of the supersystem interact with each other. In our gedanken experiment, the system (component 1) must receive the excess energy from the system's external environment, which consists of components 2 and 3 of the supersystem.

It follows that the positive charge as observed involves a flow of (positive) energy from 3-space back to the time domain. If we wish, before observation we can accept that "observed positive energy inflow" as an unobserved outward radiation of negative energy. Thus, a dipolarity involves a continuous flow of energy from the time domain into 3-space, and back from 3-space into the time domain (Figure 3-2). The energy flow circuit (circulation) is completed by a remaining energy flow across the dipolarity in the time-domain only, from the positive charge where the energy enters the time domain, to the negative charge where it exits again.[64]

- 4-spatial energy flow circulation between time domain and 3-space
- Broken energy flow symmetry in time axis and in 3-space
- Energy flow symmetry in 4-space
- A sort of "Lorentz symmetrical regauging" of 4-space
- Thus the EM energy in 3-space does not build-up and explode the universe

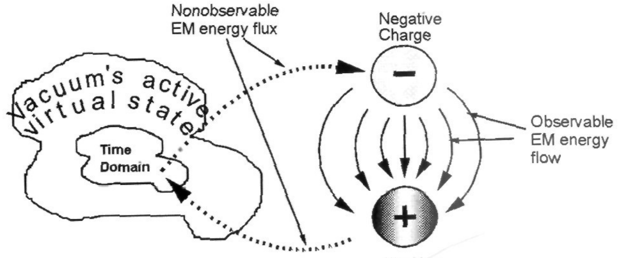
Nonobservable EM energy flux
Negative Charge
−
Vacuum's active virtual state
Time Domain
Observable EM energy flow
+
Positive Charge

- What spreads from the source charge or dipole is a circulation, and the polarization of the vacuum

Figure 3-2 Circulation of energy flow in a dipole, between time domain and 3-space.

---

[64] Rigorously, we should speak of +*qt* and -*qt*. Charge *q* is an observable, and therefore an instant frozen snapshot. No observable exists continuously in time, but instead it recurs continually as the photon interactions — iteratively producing the *d/dt* (LLLT) = LLL observations — continue. The moment a charge *q* absorbs a photon, it changes to *qt* in form. The moment the charge-time *qt* emits a photon, it changes back to *q* in form. This incessant change of *q* -> *qt* -> *q* -> *qt* —» .... is continual, at every level of the macroscopic charge, with its entire ensemble of photon interactions, including both virtual and observable photon interactions. As can be seen, the "charge as it persists" and "moves through time" has a rich dynamic substructure essentially ignored in present electrodynamics. It is, however, recognized by leading theoreticians and Nobelists.

However, even each of the clustering virtual charges and each of the differential charges of the observable charge, may be further broken into a central charge and clustering charges of opposite sign, in the modem view of the vacuum and its polarization. Hence, at any level — electron or positron — we actually have a dual and slightly separated bidirectional flow, at least in the observable equilibrium condition. The "positive energy" and positive flow of time is usually associated with "inert" matter with the electrons on the outside, so that the observer observes normal matter through the "forward time" situation mostly. Hence much of the accompanying backwards or time-reversed flow — from the positive charges in the nuclei, etc. — is "buried in the vacuum virtual state" in the Dirac sea and in Dirac sea hole flows.

Indeed, the vast hierarchy of nested structures and nested dynamics in charge as it exists in the vacuum, is recognized in physics, as is the fact that it involves infinite energy. E.g., quoting Nobelist Weinberg {194}:

> *"[The total energy of the atom]... depends on the* *bare mass* *and* *bare charge* *of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite because it depends on the infinite bare mass and charge, and the energy shift... that is infinite because it receives contributions from virtual photons of unlimited energy."*

Here again we have met Kron's "open path" in a new and unrecognized form.

Hopefully we now are in a position to rigorously define charge — or try to — for the first time in physics.[65] **Charge** is the ongoing circulation of EM energy flow between the time and 3-space domains, as seen by the observer. **Negative charge** is the ongoing absorption of positive EM energy input from the time domain (from ict), transduction of the absorbed energy into positive 3-space EM energy, and re-emission of the EM energy in all directions in 3-space. **Positive charge** is the ongoing absorption of EM energy input from 3-space, transduction of the absorbed energy into the time domain, and re-emission of the EM energy in the time domain. Or contrary-wise, one may regard the positive charge as the ongoing absorption *of negative* EM energy input from the time domain, transduction of the absorbed negative energy into 3-space, and re-emission of the negative EM energy in all directions in 3-space. The negative charge then absorbs the negative EM energy impinging on it from 3-space, and transduces it into negative energy flowing back into the time domain.

Any *overall* curvature of spacetime, as seen by the observer, results in one of the two flows *predominating* as seen by the observer. Usually, at COP < 1.0, this Dirac sea hole current does not bother our circuit's operation, but remains confined to the local vacuum. It is often not bothersome or is minor for other than very high COP's. The reason is that, for such moderate and small COP's, most of the Dirac sea hole currents react either with (i) the seething local vacuum virtual particle cauldron, or (ii) the atoms in the material lattice to form lattice holes.

As the COP increases, the fraction of Dirac holes interacting with the lattice to form lattice holes increases. As the COP further increases greatly, a large fraction of the Dirac sea holes do not react with the local vacuum cauldron (they become stable forms in that cauldron) or the lattice, but flow into the input section of the system to react with and "eat" incoming electrons from the external power source.

If the inflow of electrons is insufficient to fill the arriving Dirac sea hole current at the input section, the hole current passes back into the feeder line and into the distant power source, eating electrons all the while. This includes eating electrons in the distant power source, which simply "sees" that it must furnish additional current and voltage (power) toward the distant circuit or system.

---

[65] This is still a first order, macroscopic definition. For a more comprehensive statement, one must spell out the hierarchy and dynamics of the virtual state, the participating mass of the charged particle and its generatrix, the operations and dynamics of the *d/dt* observation operator, etc. We have not yet succeeded in getting all that into a single reasonable but sufficiently short statement!

In that operational COP region where strong interaction with the lattice occurs, the attachment of the lattice positron (lattice hole) to the mass of the atom or other ion, results in a very sluggish response of the overall ion with EM fields, compared to the response of an electron or a negative energy state electron hole (Dirac sea hole). Hence much of conventional electrical power circuit theory ignores the lattice holes and hole currents which no longer have negative energy because they have reacted and been "observed". Semiconductor designers, however, account for it and handle it in the semiconductor (in a given range of operation) primarily by the use of donor and acceptor materials.

At COP»1.0, however, the Dirac sea hole current must be strongly taken into account because it produces substantial and novel new effects in the power system itself. This Dirac sea hole current "flows in reverse" with respect to the circuit — from output back to the input. In short, it flows *through the vacuum* from the output section of a unitary COP»1.0 system, back through the middle section of the system, and back into the input section where it "eats" incoming electrons from the external power supply. The power supply thus must furnish electrons to "kill (power the killing of) the Dirac sea hole current arriving in the input section", while still furnishing the necessary amount of additional electrons to power the circuit's losses and loads.

The hole current arriving back in the input section of the unit poses an additional "power system load". These holes must be filled with incoming electrons (which disappear into the Dirac sea hole in the vacuum without radiation, due to the concomitant relaxation or change of the local curvature of spacetime) before any incoming electrons are left to power the circuit in normal fashion. So the hole current appears as a highly undesired extra "load" formed in the input and "eating input power from the external power supply".

We further discuss this situation and how to deal with it, in Chapter 9.

Finally, we propose a formal modification of the concept of "EM energy flow through space" (perhaps the first since its formulation independently by Heaviside and Poynting). The EM energy flow connected with the charge is a circulation of EM energy flow between the time domain and 3-space and back, *as a result of dipolarity*. This 4-circulation flow occurs at every point in 3-space (at every point dipole in the 3-space vacuum).

The formation of an observable charge results in an immediate, continuing, spreading re-organization (polarization) of the polarization of the vacuum virtual particle flux itself. It is this *spreading polarization of the vacuum* — and its accompanying giant negentropy EM energy flow circulation in 4-space connected with the spreading dipolarity and the observable "source" charge — that moves outward in all directions at the speed of light in 3-space from the newly formed charged particle or group of them.

We also define charge itself as that *continuously spreading dipolarity in 3-space*[66] accompanied by the giant negentropy EM energy flow 4-circulation from the time domain to 3-space and back.

The giant negentropy circulation in 4-space — represented by the spreading vacuum point-polarization and hence broken 3-symmetry — is what (observably) moves out in all directions from the newly formed observable charge. Since we "observe" only the 3-space aspect, we "see" it as the spreading appearance of 3-space EM energy (the polarization change of the vacuum itself) moving in 3-space, because we "see" it as the iteratively observed frames of a motion picture. *As observed,* that is one way of expressing it. As it actually *exists and happens prior to observation?*[67] the so-called "flow of EM energy in space" is not that at all. It is the radial spread of the appearance of a giant 4-space negentropy EM energy flow circulation between the dipoles formed in the polarization of space. In that process, the special 4-symmetry between time and space is maintained but with broken 3-energy flow symmetry and broken time-energy flow symmetry with respect to the 3-space observer.

---

[66] I.e., the spreading partial restructuring and re-organization of the vacuum energy or energy flux.

[67] The sharp-eyed reader will note that we clearly distinguish between the operation of a "tree falling in the forest" without an observer to see it, and a "tree falling in a forest" as an observer sees it. Without observation, there is no observed tree, no observer, and no "observed falling". Yet there exists the "tree x time", the "observer x time", and the "falling x time" whether or not their observation occurs with its resulting *d/dt* differentiation operation. One set (the causal set) is 4-spatial, and the other set (the event set after observation) is 3-spatial. The cause is 4-spatial and unobserved, and the effect is 3-spatial and observed. They are not the same, though related by the *d/dt* observation operator and by the interaction of the 4-space cause with a previous 3-space frozen observation to change it to a new 3-space frozen observation. Note that one can and does have a cause without an effect in any legitimate theory of causality, but not an effect without a cause. The cause only assumes the potential for interaction and observation, whereas the effect assumes that the interaction has occurred, yielding the observation as the effect.

The 4-circulation (giant negentropy) of the EM energy flow "from" or "connected with" a charge, is intimately connected with the spin of the charged particle, since the charged particle spins in both the time domain and in 3-space, as discussed in paragraph 3.3 below. We believe the spin itself is the basic giant negentropy generator.[68] If that hypothesis is true, then giant negentropy, the circulation of EM energy between time-domain and 3-space and back, broken 3-symmetry and broken time-energy symmetry (as seen by the observer), the flow of EM energy through space (as observed), the nature of charge itself, and polarization of the quantum mechanical vacuum are all just different sides of the same coin.

Without amplification, we state that his giant negentropy outflow process has additional important ramifications in electrodynamics, general relativity, particle physics, and unified field theory. However, we leave further discussion of the subject for a future paper or book and for the advanced theorists.

3.2 Reinterpreting Whittaker's Decomposition of the Scalar Potential

As previously stated, we applied and re-interpreted Whittaker 1903 decomposition of the potential {85} between the ends of the dipole. The scalar potential is actually a harmonic set of bidirectional longitudinal EM wavepairs, where each wavepair consists of an outgoing EM longitudinal EM wave in real 3-space and an incoming EM longitudinal EM wave in the complex plane. Hence we have a new and novel EM energy flow symmetry in 4-space, where (as seen by the observer) the broken 3-symmetry of the dipole has removed the usual arbitrary imposition of an additional condition of 3-space energy flow symmetry as well. Simply making a dipole permissibly breaks the 3-space conservation of energy flow symmetry, and moves to 4-dimensional conservation of energy flow. Both EM energy conservation in 3-space and in the time domain are individually broken, but conservation of energy flow is now upheld between the time-domain and the 3-space domain. In blunt terms, we "consume a little time" to get any EM energy that exists in 3-space.

---

[68] We leave to a future sharp young graduate student the business of expressing the angular momentum (spin) of the charged particle as a result of giant negentropy in EM energy flow between the time domain and 3-space. It can probably be done by deeply considering the giant negentropy process for an "isolated" observable charge considered as also containing its clustering set of virtual charges of opposite sign.

The broken symmetry of a source dipole — or of a source charge considered with its clustering virtual charges of opposite sign as a set of dipoles — produces a continuing giant 3-negentropy, as seen by the observer. He only sees the steady 3-space outpouring of energy, and sees that it will continue as long as the dipole or charge remains intact.

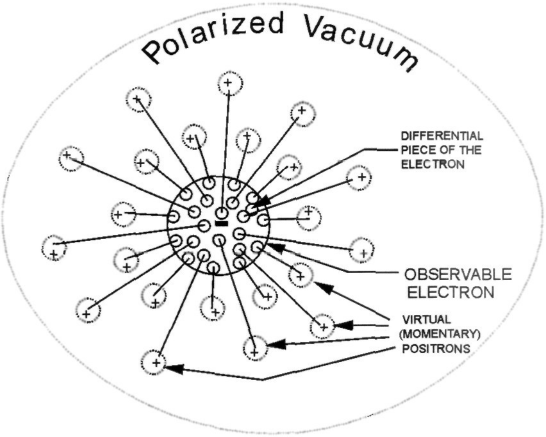
Polarized Vacuum
DIFFERENTIAL PIECE OF THE ELECTRON
OBSERVABLE ELECTRON
VIRTUAL (MOMENTARY) POSITRONS

Figure 3-3 Observable electron charge as a set of composite dipoles.

To clarify the "isolated charge" as a set of composite dipoles, see Figure 3-3. There is no such thing as an *isolated* observable charge in 3-space; instead, the observable charge is surrounded by attracted virtual charges of opposite sign in the active vacuum. We take a differential piece of the observable charge and a momentary virtual charge of opposite sign, and call the two a "composite dipole". An observable charge is simply a set of composite dipoles connecting virtual and observable energies and energy flows.

Hence any source charge (classical sense) may be considered such a set of composite dipoles. So any source charge also multiply initiates the broken 3-symmetry and broken t-symmetry in EM energy flow (as interpreted by the observer), but conserves energy flow 4-symmetry between the time domain and the 3-space domain. In 4-space, there is an energy flow circulation from the time domain to 3-space and back to the time domain again.

This special circulation of energy is strikingly similar to, and may be an extension of, Heaviside's unpublished theory of combined electrodynamics and gravity, where he used closed circulation of EM energy as exhibiting gravitational effects (in modern terms, producing spacetime curvature) {195, 172b.}. This approach or extensions of it may yet have a profound impact on physics, as commented upon by Laithwaite {196}. Certainly Heaviside considered his own energy flow theory. So it appears that he considered the excess energy flow — that "misses" the interaction with the circuit or assumed unit point charge and is not utilized — as producing gravity. That of course would be in perfect agreement with modern general relativity, since any change in the local energy density of vacuum is also a change in the local curvature of spacetime. We have called that Heaviside nondiverged and unaccounted EM energy flow by the label *dark positive energy*. See our further discussion of unaccounted (dark) positive energy in paragraph 3.8 below.

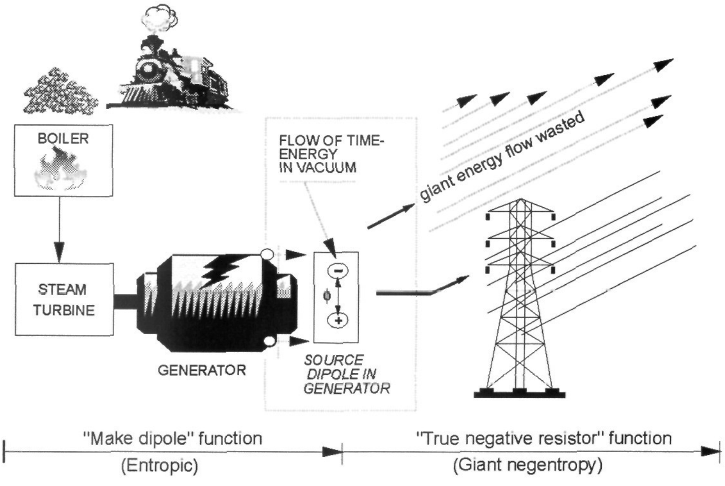
BOILER
STEAM TURBINE
GENERATOR
FLOW OF TIME-ENERGY IN VACUUM
ϕ
SOURCE DIPOLE IN GENERATOR
giant energy flow wasted
"Make dipole" function
(Entropic)
"True negative resistor" function
(Giant negentropy)

Figure 3-4 Dipole in generator powering circuits, taking energy from the time domain.

See Figure 3-4. This 4-symmetry is a far more fundamental energy flow symmetry than either 3-symmetry or t-symmetry. Both the "isolated charge" and the dipole are open systems far from equilibrium with their external active environment (the active vacuum and the active curvatures of spacetime). They are permitted to perform the five "magical" functions of disequilibrium thermodynamical systems. Such a system can: (i) self-order, (ii) self-oscillate or self-rotate, (iii) output more energy than the operator inputs, (iv) power itself and its loads (the "load" of a source dipole or source charge may be said to be comprised of the associated fields and potentials reaching across all space, and the continuous flow of energy necessary to maintain those fields and potentials), and (v) exhibit negentropy.

The source charge and the source dipole exhibit all five functions. We often refer to these characteristics of a charge or dipole as its *giant negentropy* functions {197}.

It only takes a single white crow to prove that not all crows are black. Every charge and dipolarity in the universe is already a Maxwellian system far from thermodynamic equilibrium with its local active environment — the very class of Maxwellian systems that Lorenz/Lorentz symmetrical regauging of the Maxwell-Heaviside equations arbitrarily discards. That local environment consists of the local curvatures of spacetime and the local active vacuum. The system together with the two components of its active environment is called the *supersystem*. All three components of the supersystem continuously interact with each other and exchange energy. All is dynamic, continuously. Any appearance of a "static state" in the system is actually an equilibrium condition stabilized in the ongoing supersystem dynamics.

So one does not have to "prove" that Maxwellian systems exhibiting COP>1.0 (even COP = ∞) are possible. Every charge and dipolarity in the universe is already just such a Maxwellian system, even though all such systems are erroneously discarded by Lorenz/Lorentz symmetrical regauging. Either one must accept asymmetry and COP>1.0 EM systems in electrodynamics, or one must purge from the theory all the charges and dipoles and all effects from them. That destroys all the charges, all the fields, and all the potentials.

That also destroys electrodynamics itself, as well as chemistry, particle physics, and most of modern science. The arch critic of overunity EM system research would do well to acquaint himself with some additional physics. Either that, or practice what he advocates, and give up all charges and dipoles and all effects from them — including in chemistry, particle physics, materials science, etc. When he holds on to the charges and dipoles while insisting on the impossibility of COP>1.0 Maxwellian systems, his cherished conventional EM theory eats itself by its own tail and becomes an oxymoron.

3.3 How the Dipole Charges Transduce Time Energy

We present a simplified analogy[69] that allows one to visualize how the dipole transduces the incoming time-like longitudinal EM energy flow into 3-space real EM energy flow for negative charge, or vice versa in the case of positive charge. With a little oversimplification, a charge may be said to spin 720° degrees in one complete rotation. For our purposes, it spins 360° in the imaginary plane (over in *ict),* and then spins 360° in the real plane (in 3-space). Thus the negative charge can absorb the incoming EM energy flow in the complex plane, transduce or "flip" the absorbed EM energy into 3-space to begin its 360 degrees in 3-space, and the excitation can decay during that 3-space spin part of the cycle. The only "variable" in *ict* is the t, so any energy flow has to be "by means of the variation of the t-variable". We already argued in a previous chapter that time can be treated as highly compressed spatial energy, having the same energy density as mass. So absorbing a very tiny amount of time-energy and transducing (decompressing) it into spatial energy produces enormous spatial energy *(t* multiplied by the factor c^2).

The negative charge absorbs a *little bit* of positive time energy, transduces it into a much larger amount of 3-space excitation energy, and re-emits it in 3-space as an observable, real flow of EM energy radially outward in all directions. The positive charge receives the large amount of 3-spatial EM energy and absorbs it, transduces it back into time-energy (highly compressed energy) and re-emits it in the time domain as a little bit of highly compressed time energy.

Conversely, we may consider that the positive charge absorbs a little bit of negative time energy, transduces it into a much larger amount of negative 3-space excitation energy, and re-emits this negative 3-space energy in all directions in 3-space. That consideration is important in the cold fusion phenomenology, and in fact accounts for the formation of little fleeting "time-reversal zones" in the electrolyte solution where the law of attraction and repulsion of charged particles is momentarily reversed. The omission of the giant negentropy mechanism and the negative energy mechanism from particle physics reaction theory is why there do not exist such nuclear transmutation actions at low spatial energy (but high time energy) in conventional particle physics. All such permissible reactions have been arbitrarily omitted from the discipline.

---

[69] **For** simplification, the analogy treats "spin" of the charged particle as if it really were something like that of a spinning top, at least in its 3-spatial component. However, in physics the entity **spin** is not confined to a flat plane or even to 3-space. C**onsequently,** we caution that the analogy is primarily a mnemonic aid.

For the positron (or any positive charge), we assume that the process is reversed: 3-space energy is continuously absorbed (from the active vacuum exchange) by the positive charge during its 360° spin in 3-space. This 3-space excitation energy is then flipped into the time domain as time-charge or time-energy, as the charge starts its second 360° spin in the time domain. During that latter part of its 720° spin cycle, the spinning charge re-emits the time-charge energy (time-excitation) as a flow of a little bit of very dense EM energy in the time domain.

The effects of these mechanisms in systems of charges in motion are met with as time dilation effects, spacetime curvature effects, frame rotation effects, etc. in relativity theory.

A charge that has received excess EM energy from the time domain is said to be *time-charged* or *time-excited.* Since time-charge or time-energy is c^2 denser than spatial energy, a tiny amount of time-energy excitation or time-charge may be re-emitted as substantial spatial energy over a period of time. In short, the decay of time charging or time-excitation of charged particles can occur slowly and over time. The decay is marked by the presence of longitudinal EM wave radiation, by mysterious ionization phenomena appearing in some Geiger counters depending on their individual time-histories (individual initial time-charging condition), and by excess energy appearing in electrolytes and emitted as heat where such decay of time-charge is occurring.

Usually in ordinary circuits and EM systems the time-charge and time-charge-decay effects are minimal, being offset by their opposites in close adherence to equilibrium. However, when this is not the case and time charging does appreciably occur, then novel phenomena result in those circuits and systems. Such effects have occurred for some time in instruments associated with rigorous electrolyte experiments at U.S. Naval research facilities at China Lake {198}, but the researchers have not recognized the cause. We have pointed out the time-charge and time-decay solution to those novel phenomena, and explained why apparently "identical" instruments need not respond to the stimuli in the same manner {199}. Along the way, we also explained several of the low spatial energy (thus high time-energy) transmutations, including giving the typical reactions producing the excess deuterium, tritium, and alpha particles. We also explain those new nuclear reactions at low spatial energy (but high time-energy) in Chapter 10.

As we stated in Chapter 1, paragraph 1.1.3: In such electrolyte experiments where the palladium electrodes load with concentrations of positive charges on hydrogen ions of one form or another, then in the adjacent solution there can arise fleeting **time-reversal zones.** Such a zone is simply a very tiny volume of fluid in which antiphotons temporarily comprise the majority of the ongoing photon interactions with the local ions in solution in that zone. The result is that the normal (time-forward) law of attraction of unlike charges and repulsion of like charges is reversed; *in the TRZ, like charges attract and unlike charges repel.* This allows two positive ions such as H+ to be drawn together so tightly that each enters the (reduced) strong force region of the other, forming a quasi-nucleus. When the other ions move to eliminate the TRZ, during the decay of the TRZ the strong force increases more rapidly than the EM force. Hence, the preferred method of decay of the now-excited quasi-nucleus is by quark flipping to allow the quasi-nucleus to become a new nucleus. This nuclear transformation interaction allows the production of the excess deuterium, tritium, and alpha particles.

3.4 Reordering of the Local Vacuum: Giant Negentropy

A giant negentropic reordering of a fraction of the surrounding vacuum's energy is initiated when a dipole or charge is formed. This reordering is deterministic, and is the 4-flow energetic structure shown by Whittaker {85}, slightly reinterpreted, with perfect coordination of the converging EM energy inflow in the complex plane with the outflowing converted real EM energy in 3-space. Hence this negentropic reordering introduces "hidden order" into the vacuum's otherwise disordered energy. This reordering of a fraction of the vacuum's disordered energy also changes the Gibbs statistics assumed in quantum mechanics, replacing it with an "already chaotic" statistics containing hidden order. We propose this as a possible contribution toward resolving the presently unsolved problem of the missing chaos in quantum mechanics {200}.

3.5 Some Implications of Giant Negentropy

Giant negentropy has many envisioned impacts upon physics, such as:

(a) Allows development and use of negentropic EM circuits (COP>1.0 systems) rather than entropic circuits. Once a broken 3-symmetry with stable 4-symmetry is established (as by forming a dipole), it continues indefinitely unless deliberately destroyed.

(b) Allows direct negentropic reordering of a practical and usable fraction of the vacuum's energy at will, at any point in the universe, easily and cheaply.

(c) Provides for electrodynamic engineering of the vacuum energy and of spacetime curvatures clustered in specific templates (the concept of vacuum engines or spacetime curvature engines). This has significant medical applications {387c}.

(d) Utilizes time as energy (per our previous discussion) and thus allows use of the Whittaker phase conjugate half of the biwave set as an EM energy flow within the flow of time. Observably this translates to the continuous receipt of free time-energy by the formed broken 3-symmetry (e.g., the charge or dipole). Accordingly, the charge or dipole continuously pours out real observable EM energy in 3-space, without any observable EM energy input (time is not observable, and neither is the time-polarized photon alone).

(e) Allows "burning a little time for fuel", so to speak, in order to transduce it into a very large spatial energy flow, due to the enormous compression of 3-spatial energy represented by time-energy by a factor of c^2.

So from the giant 3-negentropy discovery, we have arrived at some intriguing findings:

(a) As is well known in particle physics, the opposite charges of a dipole constitute a broken 3-space symmetry in the violent flux exchange between the active vacuum and the dipole.

(b) This dipole's broken 3-space symmetry in EM energy flow, provides a *relaxation* to a more fundamental EM energy flow symmetry between the time domain and 3-space, and therefore in 4-space.

(c) There is no law of nature or physics that requires 3-symmetry of EM energy flow as an additional condition applied to 4-symmetry of EM energy flow. Instead, 4-symmetry energy flow is more basic, by the hierarchy rule.[70]

---

[70] Quoting Steven Weinberg: *" ..there is a hierarchy of symmetries; whatever symmetry unites the gravitational and strong nuclear forces with the electroweak*

(d) The dipole is a practical and very simple means of "breaking" the additional 3-flow symmetry condition in EM energy flow our systems have previously exhibited, and relaxing to the fundamental 4-flow symmetry *without* 3-flow symmetry.

(e) So long as the dipole statically exists (e.g., imagine an electret suddenly formed, or a charged capacitor with no leakage), real usable EM energy circulation between time and 3-space, at each point in 3-space, will continuously move out from the dipole at light speed in all directions. In the view of the observer, at the same time, reactive EM power (actually, time-energy) will continuously flow into the dipole from the time-domain (the complex plane), and be transduced into real EM power output radially outward in 3-space from the dipole.

(f) Observably, a dipole or dipolarity and its scalar potential thus comprise a *true negative resistor* system of the most fundamental kind. The dipole continually receives EM energy in unusable form (reactive power, which cannot perform real work), converts it to usable form (real power, which can perform real work), and outputs it as usable, real EM energy flow (real power) in 3-space. Further, it spreads this "transformation of time-energy into 3-space energy at every dipole point" outward in 3-space in all directions.

(g) So at its formation the dipole initiates a continuing *giant negentropy* — a progressive reordering of a substantial and usable portion of the vacuum energy {201} that begins with dipole formation and continues. Further, this reordering of vacuum energy flow at every point in space continuously spreads in all directions from the initiation point, at the speed of light. Dipoles in original atoms formed shortly after the beginning of the universe, have been pouring out real EM energy for some 14 to 15 billion years or so (i.e., in the prevailing view; the age of the universe and whether it is expanding are still controversial {754}), and have reordered a fraction of the vacuum's energy, where the magnitude of the re-ordering varies inversely as a function the

---

forces is broken roughly a hundred million million times more strongly than the symmetry that unifies the weak and electromagnetic interactions. The puzzle of explaining this enormous difference in fundamental energies is therefore known in elementary particle physics today as the hierarchy problem." [Weinberg, *Dreams of a Final Theory,* Vintage Books, Random House, 1993, p. 205], radial distance from the dipole.[71] This is in fact the solution to the long-vexing problem of the source charge and its associated fields and potentials and their energy, reaching across the universe. The charge together with its associated clustering virtual charges is a set of composite dipoles, hence a multiple *broken 3-symmetry* energy flow system.

(h) If the dipole is destroyed, the ordering of the vacuum energy ceases, leaving a "separated chunk" of reordered vacuum energy dynamics that continues to expand at the speed of light in all directions, steadily reducing in local intensity as it expands.

(i) At any very small volume in space, from the dipole dynamics of the universe it follows that a great conglomerate of reordered vacuum flows and fluxes — some continuous, some chopped — is continually passing through that volume. Further, the situation is totally nonlinear, so that direct wave-to-wave interactions occur continuously amongst these energy flows and waves. We hypothesize that this is the actual physical mechanism constituting Puthoffs cosmological feedback mechanism {202}.

(j) Further, in 1904 Whittaker {203} showed that any EM field or wave pattern can be decomposed into two scalar potential functions. This initiated what is called *superpotential theory.* Each of the two potentials for those functions, of course, decomposes into the same kind of harmonic longitudinal EM wavepairs as shown in Whittaker 1903, plus superposed dynamics. In other words, the interference of scalar {204} potentials — each of which is actually a set of longitudinal EM waves, and not a scalar entity {205} at all, but a *multivectorial* entity — produces EM fields and waves and their dynamics. Hence we hypothesize that the Whittaker interference of the propagating reordered EM energy entities, continuously occurring at any point in space, generates the zero-point EM field energy fluctuations of the vacuum itself. Indeed, an AIAS group paper by Evans *et al.* {206} has already shown that just such "scalar interferometry" produces transverse EM fields and waves in the vacuum at a distance.

---

[71] The field energy density in the fields produced by the source charge varies inversely as the square of the radial distance. The potential energy density in the potentials produced by it will vary inversely as the distance.

Since energy can neither be created nor destroyed, there really are no true *energy sources* in nature in the sense that they *create* energy and pour it out {207}. Instead, what we call "energy sources" merely convert the energy already there in one form, into another form. All energy sources are no different from a solar cell in the sun, or a windmill in the wind, at least in principle. They are actually energy transducers rather than energy sources. This corresponds well with the conservation of energy law wherein not a single joule of energy can be created or destroyed. Instead, the form of that joule of energy can be continually changed. Every joule of energy present in the universe after its formation is still present, and is still doing joule after joule of work in iterative interactions with transducers that change its form.

A priori, we can measure no real 3-space input of EM energy to the unchanging charge but we can measure real 3-space EM energy pouring from it. Energy must be input to it from the active vacuum in a *nonobservable* form, and converted by it into an *observable* form that is re-emitted, usable, and produces what we call the "fields and potentials" and their energy, associated with that "source charge". As is common usage, we will continue to use the term "source charge" or "source dipole", but with the understanding that we refer to a special kind of *energy transducer.*

When we "make entropy", we must do work. Even so, doing work is not really entropic unless we also lose control of the energy[72] — whose form was merely changed by doing the work — and thence lose any *further* ability to use that energy (change its form, or transduce it) to do useful work. And when we observe entropy, we also do work in the observing. Some of the erroneous notions about entropy desperately need correcting. E.g., quoting Weinberg {208} for a lay definition of entropy:

> ***"..a certain quantity called entropy:*** *To define entropy, imagine that some of the system's temperature is very slowly raised from absolute zero. The increase in entropy of the system as it receives each small new amount of heat energy is equal to that*

---

[72] An interesting point is that energy is energy and remains energy. "Disordered" ultimately means "unusable" or "inconvenient" and nothing else. Disordered energy is still energy, and if we do not lose it from the system we can in fact have the **system** automatically reorder it, e.g. by use of simple retroreflection. The entire concept of entropy must be rigorously interpreted in terms of "energy whose reuse is lost", at least by the processes being performed by the system considered. In our view, "entropy" itself is just additional usually "lost" energy to be recovered.

> *energy divided by the absolute temperature at which the heat is supplied."*
>
> *"....entropy, which always increases with time in any closed system..."*

However, we point out that apparently there exists no completely closed system in the entire universe. In the notion of entropy and the escape of the energy from their "closed" system, thermodynamicists themselves ubiquitously negate the latter premise that entropy in a completely closed system always increases with time. That simply does not follow. The real situation is as follows: (i) Entropy in a half-open system with no allowed external input of energy, but with allowed escape of energy, does decrease with time if the energy of the system decreases with time by escaping.[73] Hence the cooling of "hot" water with time, if and only if the excess energy in the system can have a net escape to the (cooler) environment as time passes, (ii) Energy in a completely closed system (closed to both energy and mass), with no input and no output of energy at all, simply remains constant because it is not dissipated; simply apply the conservation of energy law. It may rearrange within, and the subcomponent energies may disorganize with respect to range, but the total energy remains. Entropy actually only deals with the loss of ability to utilize the energy. To wit, the assumed (and apparently true, so far as we can tell) conservation of energy in the present universe is an example. A closed system is in this sense a special case of a system in equilibrium with its active environment.[74] I.e., any "closed system" is such a special case.

---

[73] In case the reader has not thought of it, until the energy "escapes" it is not dissipating or dissipated from the dissipating "object". A "hot" object is energetically excited but not hot; the *energy dissipated* from it (including from every component within it) is scattering, and hence disruptive and "hot" in its interaction with a test instrument (or one's finger). The thermometer (or any other instrument) is external to the rest of the system (even if imbedded in it). Hence merely to "measure the temperature" is actually to pass energy out of the system into the thermometer (or other instrument), so that the thermometer measures the heat of the dissipating energy that has just escaped from the system. For these and similar reasons, Romer — former editor of *A. J. Phys.*— pointed out that "heat" should not even be used as a noun {84}.

[74] The concept of energy always requires that one can write "energy of. ..(X)" and fill in the (X). We usually write the phrase "of...(X)" as "X energy". Thus we write the phrase "energy of heat" as "heat energy," the "energy of the vacuum" as "vacuum energy", and the "energy of the EM entities" as "EM energy". At least in its usage, energy always requires a medium or entity "having" or "exhibiting" the energy. As Feynman pointed out, we really do not know what "energy" per se *is.*

(iii) In a system closed with respect to output (escape) of energy, but with the system opened only to allow external input of energy, the system energy will increase, simply by the conservation of energy law. That system's entropy will decrease and its negentropy will increase. (Its mass will also increase, and so it violates the thermodynamic misdefinitions of closed system; see Appendix A). It is well known that the entropy of an open system far from equilibrium cannot even be computed! (iv) Energy in an open system where input of energy and escape of energy both occur, may either increase, decrease, or remain constant, depending upon whether the input rate exceeds the output rate, or vice versa, or the energy input and output rate are equal. Therefore such a system can exhibit overall entropy, overall negentropy, or overall equilibrium, (v) Net energy scattering rate (rate of energy escaping the system in disordering fashion) from a hot system determines the observed temperature of that system at the time of observation. The system itself is cold, prior to escape of the energy. The "heat" is due to the interaction of the escaping energy with an external object, in its disordering fashion. Inside the system prior to escape, the energy is ordered hence cold.

A system in equilibrium in its exchange with its active external environment is in a state of maximum entropy *for that specific rate of energy exchange with its environment.* As we will point out in Chapter 9, by deliberately balancing the *system-and-vacuum* exchange with the *system-and-curved spacetime* exchange, so that the two exchanges are equal and opposite at a desired rate of exchange, one may establish (at least in theory) system equilibrium at whatever specific rate of energy exchange is desired.[75, 76] In this fashion, e.g., an "impossible" molecule normally intensely unstable and decaying in a nanosecond or less, can in fact be stabilized once an appropriate balance between the two opposing energy exchanges with the system is achieved at the required specific rate of energy exchange. This is actually a new Lorentz symmetrical regauging

---

[75] This is necessary, e.g., to close-loop a COP>1.0 EM unitary system taking its input energy from the vacuum, as we shall see in Chapter 6.

[76] The Fogal superluminal transmission system in fact uses a conventional signal input to change this "symmetrical regauging" condition in consonance with the amplitude of the input signal. Thus suddenly Lorentz symmetrical regauging becomes a way to transform a conventional signal into a signal consisting of changes in the stress energy potential of the vacuum. That constitutes a longitudinal EM wave, and so it is not limited by the speed of light. To use this process, a Fogal semiconductor must be used as the modulator in the "transmitter", and another must be used as the demodulator in the "receiver".

condition; change of symmetrical regauging is very useful in this fashion. In this way, a great new chemistry with reactions presently deemed impossible is in the offing, and it is already being developed by at least one U.S. company. We do not name the company, under our tacit agreement with the scientists doing the work.

To operate conventional entropic systems,[77] first we must input energy to the inert system to force it out of equilibrium and back into excitation (back into an asymmetrically rcgauged system, violating the Lorentz condition) so that it has some excess energy (and negentropy). Then we allow the decay of the excited system's excess energy to pass through a load and be dissipated from it, doing work for us by powering the load. Conventional systems usually do not recapture the "escaping dissipated energy" but just waste it. *"Dissipating" energy is merely allowing it to escape from the system and from further system control.* Then we brutally force more excitation energy into the system, opening it and breaking its equilibrium to do so. Again we dissipate the system's excitation energy in the load to produce some more work, usually wasting an appreciable fraction of the excitation energy.

In EM systems, it is a little more complex. We continually "switch on" a negentropic flow of EM energy from the vacuum by making a source dipole. In the circuit, we collect a small part of the resulting huge energy flow in space surrounding the circuit conductors. Then we use half the collected energy to destroy the dipole source of the energy flow, while less than half of the collected energy is dissipated in the load (the rest of that second half is dissipated in the external circuit's losses). Hence we must continually restore the source dipole and pay more to do it than we get out in the load as useful work. We make a perfectly good negentropic system, and then forcibly destroy its negentropic ability faster than we power the load and get any use out of it. That's a terrible way to make and use power systems, but it is what has been done for more than a century.

So with our present entropic EM systems (actually our *self-killing negentropic* EM systems) we continually wrestle nature fiercely to the mat, so to speak, by brute force, repeatedly, to bind her into 3-space EM energy conservation. All the while, nature protests our continual entropic brutality by providing the Newtonian third law reaction force[78] back upon our causative wrestler performing the "forcing". To do entropic engineering, we have to continually input 3-space energy to the wrestling mechanism or engine, losing a bit of the input energy in the inefficiencies, and fighting the "back emf", "back mmf", or Newtonian third law reaction that is nature's cry of protest all the while. Those are nature's penalties for imposing a monstrous 3-space EM energy flow symmetry[79] upon her as an *additional, highly undesired, and highly unnatural condition* added upon her beloved special 4-symmetry energy circulation between the time domain and 3-space.

---

[77] By "entropic system" we mean a system whose excitation energy decays and is dissipated from the system or at least from its further control, so that the entropy (lost or unusable energy) of the system continually increases until equilibrium (total decay of the excitation energy) is achieved. As can be seen, this also can be used to define entropy and increase in entropy. What is usually unappreciated is that *equilibrium is actually a counterbalance between ongoing entropy and negentropy.*

In short, because we so love entropic engineering and 3-space EM energy flow symmetry, we have to provide the continual input energy to our entropic processes by burning fuel, damming rivers, erecting windmills, building waterwheels, erecting solar cell arrays, building and charging chemical batteries, etc. In the process, we destroy and pollute the biosphere on a giant scale as we rip down forests, strip-mine and drill the earth, pile up thousands of tons of radioactive nuclear wastes —that will be hot for thousands of years — from our power plants and related facilities, and spew streams of pollutants into the atmosphere, the rivers, the oceans, etc. We do all that biosphere destruction because we inexplicably insist upon placing cruel chains on nature by brutally enforcing 3-space energy flow symmetry upon her. In the process, we adamantly require adherence to classical equilibrium thermodynamics (as given a somewhat altered explanation above, and also in Appendix A).

We do not have to do it that way. In avoiding solving the source charge problem for so long (more than a century), we have been extremely backward in our thinking and in our consequent development of entropic engineering technology.

---

[78] Note that Newton's third law reaction is implicit in the special negentropic energy flow circulation of the dipole, or of the charge considered as a set of composite dipoles. The same energy that moves into 3-space from the time domain, then returns from 3-space to the time domain, at every point in the 4-circulation. If we consider the positive charge as a "source" of negative energy, then there is a concomitant circulation of negative energy from the time domain to a point in 3-space, and back to the time domain.

[79] The 3-space symmetry of energy flow equilibrium is actually between the energy we input (from its external environment) to the system, and the energy that escapes from the system back into its external environment, either in its losses or in its loads. As we stated, equilibrium condition is a balance between ongoing entropic and negentropic operations.

For our dullness, we have had to pay and pay continuously for insisting on doing such atrocious entropic work and forcing nature to do it that way. In so doing, we "tie nature's feet" with that added arbitrary 3-symmetry in energy flow. We ourselves prohibit nature from performing the giant negentropy she so dearly loves and much prefers. We also arbitrarily and *meanly* discard the bountiful electromagnetic energy flow that nature loves to furnish us so freely by her vast preference for giant negentropy. We meanly discard nature's bountiful giant negentropy banquet free for the taking, by our vast thirst for giant entropy and doing violence to nature. Because we act like brutes instead of scientists, we destroy nature in spite of nature's continuing attempt to give us all the EM energy we wish, freely except for a little switching cost.

A far better way is to cooperate with nature and "let nature make her beloved copious negentropy" as she yearns to. To do that, we now can see the startlingly simple mechanism. We simply make a little dipole, *once*, entropically. So we have to pay for *making* the dipole, *once*, and we have to do just a little gentle violence to nature, *once*. Then we need do no more violence, if we just leave the dipole intact and do not destroy it. In short, we do have to tap nature gently on the shoulder. We do not have to brutally knock nature down to the mat, repeatedly.

When we make the dipole, we make a little bit of "broken 3-symmetry" in the universe's energy flow. Voila! Nature sings for joy at finally having her feet freed from the shackles of the horrid 3-symmetry energy flow. In great glee, she instantly relaxes into her dearly preferred giant negentropy and 4-symmetry. She instantly sets to re-ordering a substantial and usable

---

[80] E.g., so far as I am aware, there have been no funded research programs to discover how to deviate and collect some of the freely flowing energy from a permanent magnet or an electric dipole. Instead, electrodynamics has smugly continued to call it "statics" as if there were no dynamics, and ignored the giant negentropy circulation that is "seen" as the steadily outflowing field energy in 3-space, and erroneously called the "static field". There is no such thing as "statics" in the entire observed universe *a priori;* the very process of observation itself is dynamic. What we see as "static" is continual recurrence of a given condition or state or magnitude, as beautifully pointed out by Van Flandern {628}. One might even say that, not only is nature energetic, nature *is* energy, constantly changing and in motion in innumerable ways at once! One might even define "energy" as the process by which nature both acts and observes its actions, in every form and structure and possibility.

portion of the vacuum energy of the universe, in all directions at the speed of light, spreading her preferred giant negentropy EM 4-energy circulation flow from the dipole toward the ends of the universe at the speed of light. As long as we do not destroy the dipole (the broken 3-symmetry) that gently breaks the 3-symmetry shackles, nature's feet remain freed from brutal 3-space symmetry in EM energy flow. In that case, she delightedly continues to reorganize a portion of the vacuum energy, with the reordering spreading radially outward at the speed of light. Simply making a common dipole or charge[81] sets nature to going about reordering an ever-increasing fraction of the entire vacuum energy of the universe, and continuing indefinitely.

Simultaneously with freeing her from enforced 3-space EM energy flow symmetry, in great gratitude nature pours out an immense real EM energy 3-flow from that little dipole or charge. She will continue to pour it out forever for us, if we do not destroy the dipole or charge.

Instead of the present universal entropic engineering, the smart thing to do is make just a little bit of entropy wisely, using the work we get to break 3-space energy flow symmetry (basically, to make a dipole). *Then leave that mother of all negative resistors and free energy flow generators alone and forever unchanged!* Concentrate on intercepting, extracting, and using the free 3-space energy copiously flowing forth from the giant negentropy, without destroying the dipole that is freely providing it.

3.6 How Circuits Are Powered

Let us now look at the great magnitude of the energy flow that nature gives us from that dipole. We have to get into the subject a bit, because EM energy flow theory has been rather thoroughly confused for more than a century.

First, batteries and generators *do not* use their available internal energy — the shaft energy we input to the generator, or chemical energy available in the battery — to power the external circuit. Instead, each uses its available internal energy {209} to perform work on its own internal charges, forcibly separating the charges to form the *source dipole.*

See again Figure 3-4. *All the hydrocarbons ever burned, all the nuclear fuel rods ever consumed, and all the dams ever built, added not one single watt directly to the power line.* All the energy from those activities was input to the generator shaft after normal losses en route, to provide internal magnetic energy available to the generator. In turn, the generator used that available internal magnetic energy only to do internal work on its own internal charges to force them apart, forming the source dipole connected to the terminals. Generators are energy transducers only; they do not directly power their own external circuits.

---

[81] As we previously explained, a charge is just a set of composite dipoles.

Batteries and generators expend their internal energy available to them, to make the source dipole, and for no other purpose! None of their internal energy is used to power their external circuit. It never has been, and it never will be.

Once the source dipole is formed, it does all the hypothesized 4-functions we pointed out previously. It induces the spreading giant negentropic reordering of the vacuum energy, extracts (transduces) EM energy from the continuously reordering vacuum, and pours out from the terminals of the generator (or battery) a vast 3-flow (as observed) of EM field energy along the external circuit. As indicated by Kraus's illustration of the Poynting component {210}, this giant EM energy flow fills all space surrounding the circuit, out to an infinite lateral radius.[82] The energy flow is generally parallel to the conductors of the circuit. Only a tiny component of this flow — due to the surface charges of the conductors and the little boundary layer of energy flow that slides along the surface of the conductors — strikes the surface charges and gets diverged into the conductors (by the lateral withdrawal of the surface charges — with the "stub" or "base" of their field energies — laterally into the conductors. The electrons move mostly laterally, withdrawing from one side of the conductor surface to the opposite. Only the tiny component of its field vector integrated over that small distance is withdrawn into the conductors to power the electrons. This small amount of "withdrawn" energy is the diverged Poynting component, "collected" by the circuit. It also is the small component then used (dissipated) to power the Drude electrons and the circuit.[83] All the rest of that vast EM energy flow in the surrounding space, that pours forth from the terminals, just misses the circuit entirely. It roars on off into space and is wasted.[84]

---

[82] We accent that Kraus, along with other authors, only shows the 3-space *Poynting component* of that flow; i.e., he shows the very small amount of that external 3-space energy flow that is diverged into the circuit to potentialize the Drude electrons and power the circuit. Kraus *does not* show the remaining Heaviside component that is not diverged.

[83] As shown by Kraus, some of the energy flow at the various radial distances from the wire is withdrawn, as the surface electrons and the "stubs" of their near fields precess laterally into the depths of the wire when potentialized. Jackson, *Classical Electrodynamics,* Second **Edition,** Wiley, 1975, p. 223 also points out the nonlocal

The diverged, utilized, and accounted energy flow component — the *Poynting component* — is only a tiny, tiny fraction of the entire giant EM energy flow produced by the source dipole for every circuit.

3.7 The Heaviside Component vs. the Poynting Component

Unless Heaviside and Lorentz did rigorous calculations in a work we have not yet uncovered, we could not find any development of the functions and equations required for computing the numerical ratio of the unaccounted Heaviside "dark energy flow" *{nondiverged* energy flow component that misses the circuit), to the tiny Poynting energy component thai strikes the surface charges of the conductors {211} and is diverged into the wires to power the circuit when the surface charges precess laterally into the wire.

Consequently, we performed a very crude "special case" estimation {212} — a back-of-the-envelope type, with highly simplifying assumptions — for a very simple circuit in which one resistor is powered by a DC source. Our crude estimate showed that about 10^{13} times as much EM energy flow *misses* the circuit, is *not diverged,* and is *wasted* — as *strikes* the circuit, gets collected, and then is dissipated in the circuit to power the load and losses. Until electrical physicists re-examine the energy flow theory and

---

involvement of the transverse current. Quoting: *"...the transverse current... extends over all space, even if J is localized. "* One notes that present electrical power systems concentrate exclusively on the J current, making no attempt to collect and utilize the recognized but rather ignored transverse current.

[84] Note, however, that this Heaviside energy flow component differs from ExH. The Heaviside component is totally nondiverged (being actually in curl or swirl form), and since E and H are both diverged energy components assumed around an **Interacting** unit point static charge, there is neither E nor H in the electrical engineering sense prior to interaction with charge. So the energy flow in both the Poynting and Heaviside components is in a force-field-free form until intercepted by and interacted with charge. Poynting's component, e.g., assumes only that energy co**mponent** that does interact with charge and get withdrawn into the circuit. **Heaviside's** component assumes only that energy component that does not interact with charge and does not get withdrawn into the circuit.

again recover the Heaviside unaccounted component in it, that brute force estimate will have to suffice as at least an illustrative example.[85]

What does that 10^{13} ratio mean? For a little 1 watt generator of that DC type with that specific size conductors and that specific resistor, the Heaviside unaccounted energy flow component was about *10 trillion watts,* occupying all space surrounding the wire,[86] if all of it could be intercepted, collected, and used to power loads {213}. But the little circuit was only intercepting and collecting — and using to power loads — about 10^{-13} of the available energy flow surrounding the circuit, *if all* of it could have been intercepted, collected, and used to power loads.[87]

We have never had, and we do not now have, an *EM energy* shortage or problem. We never will have one, due to nature's bounty in providing us with the giant negentropy of the common dipole and of the charge considered as a composite dipole. Instead, we have the two problems that

(i) only a tiny, tiny component of the available 3-energy flow extracted from the vacuum (decompressed from time-energy) by the source dipole and poured out of the terminals of the power source, is caught and used by the circuit, and

---

[85] We would very much welcome a much more accurate functional expose by a capable higher group symmetry electrodynamicist!

[86] We also point out that none of this nondiverged energy flow component is accounted in conventional electrodynamics and electrical engineering (the closest it comes is Jackson's mention of the transverse current over all space), and neither has it been accounted or even investigated by biophysicists seeking to determine the effects of EM radiation and fields upon biological systems. We strongly suspect (and have done a little work on the problem) that these "force-free fields" and neglected excess "strange energy" flows do have interactions throughout the highly nonlinear cells of the body, and can indeed have profound long term effects on long-exposed biological systems. We particularly stress the potential for time-charging and subsequent *combined scalar and longitudinal photon pair* decay of time-charge excitation in the body. Several major areas of electrodynamics — such as the transverse current effects, Heaviside component effects, and internal Whittaker structuring (engines) effects of EM radiation — have not been investigated by the scientific community.

[87] We accent that time-energy may be regarded as spatial EM energy compressed by the factor c^2. Hence one second = 9×10^{16} joules of EM spatial energy, if decompressed by time-energy interaction with charge. The crude estimate for one circuit of 10^{13} times as much energy being wasted in the Heaviside energy flow component as is caught and used in the Poynting component, is thus quite reasonable when one recognizes that *all* the EM energy flow appearing in the circuit or outside it actually comes from the time domain, hence constitutes *decompressed time-energy.*

(ii) in the closed current loop circuit, half of that small spatial EM energy component that is intercepted and caught, is used by the circuit only to destroy the source dipole and cut off the free flow of EM energy from the vacuum faster than it powers the load.

We have previously discussed this further in several papers {214}.

One must occasionally keep one's sense of humor. The real problem that the Department of Energy should be working on, with massive resources, is simply the alteration of the closed current loop circuit so that it does not destroy the source dipole faster than it powers the loads. Secondly, they should then be working on how to close-loop the COP>1.0 EM systems that readily emerge from that research. Yet it appears that this fundamental energy problem — and the *only* energy problem — has no conventional work being openly done on it anywhere in the open Western world.[88]

3.8 Dark Positive Energy: The Unaccounted Heaviside Component

Now we restate a deceptively hidden thing: The EM field and potential are defined as the observed effect of the deviation from them by an assumed unit point charge. Then that same field and potential are assumed to be in mass-free space, before the interaction with charged matter. Hence the field and potential have each been defined erroneously as both a cause and an effect. This is a gross non sequitur. There is also a total lack of accounting that all EM energy in 3-space comes from the time domain. There is no consideration that every dipolarity and every charge represents a giant circulation of EM energy between time and 3-space. In ensemble, this non sequitur and these omissions are responsible for most of the present formidable foundations problems in electrodynamics.

---

[88] Using the U.S. as an example, the U.S. classified community is highly compartmented, and the real "organization" of a given compartment often consists of laterally connected "same clearance" compartments in multiple outside agencies rather than vertical chain of command through one agency. We accent that some highly compartmented "lateral" classified organizations in the West and elsewhere already have working COP>1.0 systems, and have had them for several decades. Some of these laterals in the U.S. are very dedicated to the duly elected government of the United States, with good national security reasons for the fact that they "officially do not exist". Some are not so dedicated at all, and have their own agendas having little if anything to do with official U.S. government policy. These latter "rogue" laterals often have convenient connections with outside shadowy groups of the High Cabal (Churchill's phrase). A similar situation exists in several other major foreign nations. Substantial electrodynamic technology has been developed (and weaponized) for decades, in secret, and this technology does not appear in the open news, the university texts, or the scientific journals. Such is the case for COP>1.0 EM power systems, which have actually existed in secret for decades.

We submit the following new hypothesis: *The present "definition" of field and potential as a function of interception and divergence by a unit point static charge are only indications of their point intensity, obtained as an estimate after their observed interaction with a unit point static charge.*[89] *Thus the actual field and potential contain not only their Poynting components that interact with the unit point charge, but also their Heaviside nondiverged components that do not interact with it. The true "magnitudes" of the massless field and potential prior to interaction with charged mass are many orders of magnitude greater than the magnitudes of their standardized point intensities* {215}.

In short, we propose that *every field and potential, in its interaction with charged mass, produces not only the accounted energy aspects presently calculated by electrodynamicists and astrophysicists, but also contains Heaviside's vast unaccounted "dark positive energy" component presently discarded.* We further propose that this ubiquitous and unaccounted "dark positive EM energy" component in the interactions of the fields and potentials of the spiral galaxies is what is producing the excess gravity that is holding the spiral arms together. We specifically do this in honor of Heaviside, whose unpublished papers clearly establish that he recognized the overwhelming gravitational importance of his extra and unaccounted energy flow component {172a, 172b, 196}.

3.9 A Short History of the Discarding of the Heaviside Dark Energy

3.9.1 Independent Work by Heaviside and Poynting

Let us now see how the enormity of the EM energy flow from the dipolar source was treated in the early electrodynamic theory, and how it is treated in the received view today.

---

[89] This is recognized by leading electrodynamicists, but not by most electrical engineers building power systems!

In the 1880s after Maxwell was already deceased, Poynting {216} and Heaviside {217} independently (and rather simultaneously) discovered EM energy flow through space. Before that, the concept did not appear in physics. Poynting {216} published prestigiously, while at first Heaviside published more obscurely {217}, then finally more prestigiously {218, 219}.

With respect to circuits, from the beginning Poynting assumed only that small amount of EM energy flow that enters the circuit from the surrounding space. Here are Poynting's {220} own words:

> *"This paper describes a hypothesis as to the connexion between current in conductors and the transfer of electric and magnetic inductions in the surrounding field. The hypothesis is suggested by the mode of transfer of energy in the electromagnetic field, resulting from Maxwell's equations investigated in a former paper ("Phil. Trans.," vol. 175, pp. 343-361, 1884). It was there shown that according to Maxwell's electromagnetic theory the energy which is dissipated in the circuit is transferred through the medium, always moving perpendicularly to the plane containing the lines of electric and magnetic intensity, and that it comes into the conductor from the surrounding insulator, not flowing along the wire. "*

As can be seen, Poynting considered only the energy flow actually entering the wire, and subsequently being dissipated in the circuit. Hence Poynting never considered the huge EM energy flow component around the circuit that is *not* diverged, *misses* the circuit entirely, *does not* contribute to the energy dissipated by the circuit, and is *wasted.* In short, there is a vast "dark energy flow" associated with every dipolar interaction — a huge energy flow component which Poynting never considered.

Heaviside's theory was an extension of what Poynting had considered, and Heaviside also corrected Poynting as to the direction of flow. Heaviside was fully aware of the enormity of the "dark energy" flow missed by Poynting, but had absolutely no explanation as to where such a startlingly large EM energy flow — pouring from the terminals of every dipole, generator, or battery — could possibly be coming from. Consequently, Heaviside was very cautious in referring to it, usually doing so only obliquely in terms of the angles and components. In Heaviside's {221} own words:

> *"It [the energy transferflow] takes place, in the vicinity ofthe wire, very nearly parallel to it, with a slight slope towards the wire... . Prof. Poynting, on the other hand, holds a different view, representing the transfer as nearly perpendicular to a wire, i.e., with a slight departure from the vertical. This difference ofa quadrant can, I think, only arise from what seems to be a misconception on his part as to the nature ofthe electric field in the vicinity of a wire supporting electric current. The lines ofelectric force are nearly perpendicular to the wire. The departure from perpendicularity is usually so small that I have sometimes spoken ofthem as being perpendicular to it, as they practically are, before I recognized the great physical importance ofthe slight departure. It causes the convergence ofenergy into the wire. "*

As can be seen, Heaviside was fully aware that the energy flow diverged into the wire was only a minuscule fraction of the total. He was fully aware that the remaining component was so huge that the energy flow vector remaining — after the divergence of the Poynting component into the circuit — was still almost parallel to the conductors. However, he had no explanation at all of where such an enormous and baffling energy flow could possibly originate.

Had Heaviside strongly stated the enormity of the nondiverged component of the energy flow, he would have been viciously attacked and scientifically discredited as a perpetual motion advocate. His words were measured and cautious, but there is no doubt that he recognized the enormity of the nondiverged EM energy flow component.

We have chosen to call that huge unaccounted component the "Heaviside dark energy component" in his honor, since he actually discovered it. By the word "dark" we mean "unaccounted", which hides it from scientific view. We have also nominated it as the previously unsuspected source of the extra gravity holding the arms of the spiral galaxies together.

3.9.2 Lorentz Disposed of the Problem Rather than Solving It

Lorentz then entered the EM energy flow scene to face the terrible problem so quietly raised by Heaviside. Lorentz understood the presence of the Poynting component, and also of the extra Heaviside component, but could find no explanation for the startling, enormous magnitude of the EM energy pouring out of the terminals of the power source (pouring from the source dipole) {222} if the Heaviside component was accounted. Had he retained and advanced this enormous dark energy flow component, even the great Lorentz would have been castigated as a perpetual motion advocate who did not accept the law of energy conservation. Even today, when one points out that far more energy is pouring out of the terminals of every generator than the small mechanical energy input to the shaft, one is immediately labeled a raving lunatic — even though the Bohren experiment {24} clearly proves the presence of an extra energy flow component not usually intercepted. It is not difficult to understand how such a bombshell would have been received in the 1880s! At least today, we have particle physics and its broken symmetry of opposite charges, together with its active vacuum, to explain where the excess energy comes from and how.

Unable to *solve* the dark energy flow problem by any rational means, Lorentz found a clever way to *avoid* it. He reasoned that the nondiverged Heaviside component was "physically insignificant" (his term) because it did not even enter the circuit. Since it did nothing of any physical consequences, or so he reasoned, then it could just be discarded.

So Lorentz simply integrated the entire energy flow vector (the vector representing the sum of both the Heaviside nondiverged component and the Poynting diverged component) around an assumed closed surface enclosing any volume of interest {223}. *A priori,* this mathematical procedure discards the dark *Heaviside* energy flow component because of its nondivergence. It retains only the intercepted *Poynting* diverged component that enters the circuit.

A century later, electrodynamicists are still happily avoiding the dark energy flow problem by continuing to use the Lorentz integration procedure {224} to dispose of all but the Poynting component that enters the circuit and is then dissipated by the circuit. As a result, the "Poynting energy flow" has come to be loosely regarded as "the" entire EM energy flow, though electrodynamicists find it necessary to give stringent warnings about it. E.g., Panofsky and Phillips {225} state it this way:

> *"...only the entire surface integral of N [their notation for the Poynting vector] contributes to the energy balance. Paradoxical results may be obtained if one tries to identify the Poynting vector with the energy flow per unit area at any point."*

Most electrodynamicists note the freedom to add a vector — few call it an *energy flow* vector, though that is the type of vector being discussed, and one must add apples to apples — which has zero divergence. Jones {226} states:

> *"It* is *possible to introduce the Poynting vector S, defined by S = ExH, and regard it as the intensity ofenergy flow at a point. This procedure is open to criticism since we could add to S any vector whose divergence is zero without affecting [the basic integration procedure's result]."*

Jackson {227} says it even more plainly, and also uses Lorentz's "no physical significance" argument for disposing of any energy flow vector with a zero divergence. Quoting:

> *"...the Poynting vector is arbitrary to the extent that the curl ofany vector field can be added to it. Such an added term can, however, have no physical consequences."*

Any energy flow vector, which is the curl of a vector field, will have zero divergence, by elementary vector algebra. In short, to be pertinent at all, it must be an *energy flow* vector (since energy flow is what S = E x H is all about. Since the curl of any vector has no divergence *a priori,* then any energy flow vector that is a curl of a vector field will be part of the Heaviside dark energy flow component, rather than part of the Poynting energy flow component. It will also be discarded by Lorentz's closed surface integration.

We refer the reader again to Figures 2-5 and 2-6 in Chapter 2, p. 113-114. Electrodynamicists err in assuming that such a divergent free vector (energy flow) can have no physical consequences. If one inserts *additional intercepting charges* into that formerly nondiverged energy flow component, the charges will immediately diverge some of the formerly nondiverged energy flow around them and hence "collect additional energy". In other words, "divergence" is the result of a physical process, accomplished by intercepting charges. We can easily change "nondiverging" energy to diverging energy, by simply interacting it with charges. We can also "bunch" the energy flow lines closer together (as in some waveguides), thereby increasing the local flow intensity and the concomitant energy collected by each intercepting charge in the densified energy flow stream.

So whether the Heaviside energy flow component is of physical consequence or not, depends on whether or not we "add more sail to the sailboat", or "add more sailboats" before the wind. That is most certainly a useful physical consequence, and with a little ingenuity we ought to be able to use it to power loads "for free" or nearly so. There are other consequences also, as we used in deriving the negentropy of the dipole or charge. There the input of a non-Poynting energy flow component certainly has universal and physical significance — e.g., to the curving of local spacetime, conditioning the active vacuum, furnishing all the energy to form the EM fields and potentials from their source charges, etc.

Schwarz {228} expresses it this way:

> *"There will be many opportunities in which the interpretation ofE x H as a rate offlow ofenergy per unit area will be profitable. In most cases ofpractical interest, such an interpretation is valid, although it must always he kept in mind that only the integral of S over a closed surface can be physically measured... Just how it is that the connections to the energy source, say a battery, are at the ends ofthe wire, yet energyflows in through the sides, should be pondered by the reader. "*

Note carefully that Schwartz recognizes the difference between the observed (deviated or diverged) part of the EM energy flow and the unobserved nondeviated or nondiverged part of it. However, he did not follow it up and apply it back to the basic "definition" of the field and potential as what are diverged from them.

For recommended changes to the Poynting vector, Jones {229} presents many conditions the changed vector must fulfill. Then he falls back on the Lorentz closed surface method again, but without realizing that he therefore (i) first includes both the diverged and nondiverged component, and (ii) invokes a procedure that arbitrarily discards the nondiverged component. In thus disposing of the problem, Jones says:

> "It *does not seem likely that an expression satisfying all these conditions will be simple........fortunately, we are rarely concerned with the energyflow at a point. In most applications we need the rate at which energy is crossing a closed surface."*

Finally, we note that even today, a debate on what the Poynting vector is or should be, is still politely ongoing. As an example, it has been ongoing for more than 40 years in the *American Journal ofPhysics* alone as well as in other journals {230a-230i}. One thing is clear: The Poynting vector does not represent the total energy flow at right angles through a square meter in space surrounding the conductors of the electrical circuit. It is, however, the *component or fraction* of that actual energy flow through the square meter in force-field-free form *that will subsequently be diverged and collected in the circuit in force-field form and dissipated in the circuit loads and losses.*

3.10 The Problem of the Accelerating Expansion of the Universe

As briefly stated earlier, we introduce the notion of the supersystem, consisting of three parts: (i) the system and its dynamics, (ii) the active vacuum and its dynamics, and (iii) the active spacetime curvatures and their dynamics. All three components of the supersystem interact with each other.

In Chapter 9, we explain in some detail how a COP>1.0 EM circuit — and especially a COP»1.0 EM circuit — has an induced flow of negative energy (Dirac sea hole current) from the output section of the system, back through the system, and into the input section. Reaching the input section, if the hole flow is not completely filled by incoming electrons, the remaining Dirac hole current will flow on back into the feeder line from the external power supply, and into the distant power supply itself, eating electrons and electron current all the while.

We also explain our reinterpretation of the *unobserved* (non-reacted) Dirac hole with negative energy or negative mass, without time reversal and parity reversal. This dramatically differs from the "as observed" (interacted) Dirac positron with positive mass, and with charge-reversal and parity reversal from the electron. The Dirac hole is "as unobserved and non-reacted" and it is still an *electron* having negative mass and negative energy. The Dirac positron with positive mass, opposite charge, and opposite direction is "as observed and reacted with charged matter". The unobserved Dirac hole is the "cause" of an observational interaction; the Dirac positron is the "effect" of an observational interaction. The two are very different because the cause differs from the effect. Here again, the confusion of cause and effect has had very detrimental consequences in the use of Dirac's electron theory.

Very powerful astronomical phenomena, such as giant astronomical explosions of great energy, involve COP>1.0 and COP»1.0 processes, particularly when the unaccounted Heaviside energy flow is accounted for every field and potential interaction with charged mass. Hence all such phenomena also provide very powerful "explosions" of negative energy

(Dirac sea holes moving outward in the vacuum). The negative energy is the equivalent of negative mass, so long as it is unobserved.

So there is a "dark negative energy explosion" in the Dirac Sea vacuum, accompanying every large astronomical explosion seen through our telescopes and other sensors. This accompanying dark negative energy explosion has not been accounted in astrophysics until now.

We thus live in an astronomical universe of neglected "giant dark negative energy explosions" as well as one of accounted "giant positive energy explosions".

The dark (unaccounted) negative energy radiating outward in the vacuum from gamma bursts, x-ray bursts, etc. produces antigravity. The spreading negative energy, moving at light speed, is a change to the vacuum itself, which also interacts upon spacetime to negatively curve it, producing negative gravity. For those energetic processes of COP sufficiently greater than one, the processes produce more antigravity than gravity. As of this writing, we are in fact filing a patent application on this method of performing and producing antigravity, and it will be filed before this book is published.

Over the 14 billion or so years of the observed universe, the dark negative energy content of spacetime has been slowly changing by the results of giant astronomical explosions and other such processes producing negative energy radiated outward into the Dirac sea vacuum. Slowly the vacuum itself has undergone transformation from the perfect Dirac sea with all holes filled by particles, to a vacuum where there is a slowly growing fraction of negative energy (negative mass) unobserved Dirac sea holes.

Thus, slowly the antigravity generated by the gradual altering of "space" (the vacuum) itself has been growing for some 14 billion years. At some point in the past, the excess antigravity effect accumulating in space reached the equal of the "outside" gravity acting upon energetic explosive astronomical phenomena. From that point in time, the antigravity effect has been increasing slowly. The result is that the present violently explosive astronomical phenomena, by adding a sudden and sharp increase in their external space antigravity, produce acceleration away from all positive mass and positive energy phenomena. Hence astronomers now observe the acceleration of the expanding (positive energy/positive mass) universe, instead of the previously expected slowing (due to positive gravity) of the expanding universe (conventional view).

The antigravity-producing phenomena have not been openly investigated in Western science because Western science has arbitrarily discarded COP>1.0 EM systems, primarily with the Lorentz regauging of the Maxwell-Heaviside equations. Hence, Western science has not experimentally studied the type of system that produces the effect in the first place.

On the other hand, it is possible to directly investigate the antigravity-producing phenomenology on the laboratory bench, by utilizing overunity EM systems with very high COP (e.g., COP = 10^6 or more). The Sweet experiments {231} did in fact produce such phenomenology on the laboratory bench, but these experiments were privately performed under proprietary conditions {232}. We discuss the Sweet experiments and device in a later chapter.

In science, such experiments are quite properly deemed interesting but anecdotal until replicated independently. Similar experiments should certainly be repeated by the scientific community, using various interactions (some given in this book) that develop COP>1.0. The problem is nowhere near as difficult as hot fusion, but it has not been tackled and overcome because of (1) the prevailing and erroneous scientific mindset against COP> 1.0 electromagnetic processes, and (2) the present "kill" of the supersystem and its effects in the standard U(1) electrodynamics model.

In Chapter 8 we further discuss the antigravity effect in practical or potentially practical systems where COP » 1.0.

3.11 In Conclusion

In this Chapter, we have stressed what we believe to be a great new principle of giant negentropy, leading to direct and easy extraction of EM energy from the vacuum in copious quantities. That principle is that nature retains *her preferred 4-symmetry* of EM energy flow, while breaking 3-symmetry of EM energy flow and the symmetry of EM time-energy flow on the fourth axis. It is implemented by making a common dipole or charge.

The electrical energy problem is *not* due to any difficulty in inducing giant and indefinitely continuing EM energy flows from the vacuum! The easiest thing in the world is to extract EM energy — any amount, anywhere, anytime — from the seething vacuum, by making a simple dipolarity. The problem is only in then intercepting and collecting some of the energy to dissipate in loads, and doing it *without* using half the collected energy to destroy the dipole producing the energy, faster than the loads are powered. In other parts of this book we have given some of the many mechanisms that can be applied to attack this "power the load without killing the source dipole" problem, and have also included such in papers in the literature {233a-233u}.

We have pointed out the implications of the new principle of negentropy in EM energy flow, and how all our circuits and electrical power systems actually use it but then arc designed so as to kill the negentropic process's potential for self-powering electrical power systems.

Presently we are told by the conventional scientific community that the dream of freely extracting EM energy from the vacuum, and using it to efficiently and easily power our electrical needs cleanly, is either a fool's concept of perpetual motion, or the science of the next century. It is neither; it is good physics, and it is already *present and misused* in all our electrical power systems and circuits — and always has been. Unlimited electrical energy from the vacuum is here and now, if we but overcome the prevailing mindset and grasp it.

We have argued that all electrical loads and circuits are now and always have been powered by just such EM energy extracted directly from the vacuum by the giant negentropy principle, evoked by the source dipole. We have removed the artificial and erroneous notion that batteries and generators provide some of their available 3-space internal energy to the external circuit. Instead, we have emphasized that batteries and generators dissipate their available energy to form and then continually reform their source dipole, and nothing else. Once their source dipolarity is formed, a battery or a generator is nothing but an energy gate, gating and transducing enormous EM energy from the vacuum. We only have to keep inputting energy to keep reforming the dipole that our external circuit is designed to destroy faster than the load is powered.

In dealing with the function of the dipole in extracting the EM energy from the vacuum and pouring it out to power the circuit, we have shown errors in present energy flow theory and how they came to be made. We also indicated the enormous magnitude of the EM energy flow actually **extracted** and present with every conventional circuit, but wasted and **arbitrarily** removed from accountability. Some processes for intercepting, co**llecting,** and using additional amounts of this available Heaviside "dark EM energy" flow component, surrounding every circuit, have been mentioned, and references have been given to other papers more fully discussing such processes. In a previous paper {234} we have proposed that the Heaviside dark energy, accompanying all EM field-charge interactions, is responsible for the excess gravity observed to be holding the spiral arms of those distant spiral galaxies intact — in short, we have proposed an unaccounted Heaviside *dark positive energy* solution to the well-known *dark matter* problem in astrophysics.

The ordinary closed current loop circuit inherently yields a system COP<1.0, because it destroys its source dipoles faster than it powers the load. Indeed, the circuit self-enforces the Lorentz symmetrical regauging condition during discharge of its free excitation (potential) energy.

In addition, we have presented an hypothesis to explain that negative energy (negative mass) created in the Dirac sea vacuum by giant astronomical explosion phenomena produces the excess antigravity in the space surrounding those entities that is accelerating the expansion of the universe. We have also indicated how this hypothesis can be investigated and tested on the bench. We have provided one experiment whose results did support the hypothesis.

The present world energy crisis is real and increasing, while at the same time the availability of oil is beginning to decrease and oil becomes more costly.[90] Unless a substantial fraction of the "electricity from oil" curve is shifted to "electricity from the vacuum" rather quickly, the economic collapse of the Western world, followed by concomitant collapse of other economies, may ensue within a surprisingly few years, and even within the present decade.

---

[90] As this was first being written in Dec. 2001, there was a temporary relief in the oil crisis, due to the sudden decrease in travel, air traffic, etc. after the terrorist attack on Washington and New York on September 11, 2001. As the present war on terrorism has continued, the MidEast is now heating up toward another war, and oil prices are again trying to climb. Iraq, Iran, and Libya called for an oil embargo on the United States, for example, which would almost immediately plunge the nation into an oil crisis. Meanwhile, should effective terrorist attacks occur upon the highly vulnerable energy distribution infrastructure of the United States and other allied nations, the oil crisis will again emerge catastrophically and almost immediately thereafter. It is held off for now, only because Russia under Putin is seeking U.S. financial assistance to further develop its rich oilfields. When OPEC votes to curtail production and then curtails it, Russia opens the spigot and keeps the prices down. Putin is seeking to become a trusted U.S. oil partner, in the role long played by Saudi Arabia but now declining due to perceived Saudi payments to terrorist organizations and reluctance to close down terrorist activities in their own country. So in a strange way, a new alliance between the U.S. and Russia has prevented the oil and energy prices from escalating right through the roof.

We believe that a transfer of much of the "electricity from oil" curve to an "electricity from the vacuum" curve can be quickly accomplished, if the government and the scientific community can be induced to move with full priority and vigor in the directions indicated in this book. However, to get that done, it apparently will be done over the writhing bodies of our leading scientific organizations, literally dragging the organizations to face the proven fact of giant negentropy of every charge and dipole in the universe over their strenuous objections.

As a major objective, we fervently hope that undergraduate students, graduate students, post-doctoral scientists, other scientists, engineers, and environmentalists will interest themselves in these new principles and viewpoints. We have called attention to the new electrodynamics pioneered by AIAS, Evans, Sachs, Barrett, Lehnert, and others, and indicated the capability of this emerging electrodynamics to model this new *energy from the vacuum* functioning.

It has been said that no idea is accepted until its time has come. We strongly believe that the time for cheap, clean extraction of electrical energy from the vacuum is an idea whose time has arrived. Hopefully our efforts in this book will at least conceptually show the interested reader the outline and course that energy from the vacuum can and will take in the scientific community and in the world at large.

We accent again that modeling the extraction of EM energy from the vacuum demands the use of higher symmetry electrodynamics, particularly a theoretical model capable of expressing the interactions of the three components of the supersystem. Barrett and Grimes {235} said it beautifully:

> *"In the case of electromagnetism, the theory was first simplified before being frozen. Maxwell expressed electromagnetism in the algebra of quaternions and made the electromagnetic potential the centerpiece of his theory. In 1881 Heaviside replaced the electromagnetic potential field by force fields as the centerpiece of electromagnetic theory. According to him, the electromagnetic potential field was arbitrary and needed to be "assassinated" (sic). A few years later there was a great debate between Heaviside and Tait about the relative merits of vector analysis and quaternions. The result was the realization that there was no need for the greater physical insights provided by quaternions if the theory was purely local, and vector analysis became commonplace.*

> *The vast applications of electromagnetic theory since then were made using vector analysis. Although generations of very effective students were trained using vector analysis, more might be learned physically by returning, if not to quaternions, to other mathematical formulations in certain well-defined circumstances. As examples, since the time when the theoretical design of electromagnetism was frozen, gauge theory has been invented and brought to maturity and topology and geometry have been introduced to field theory. Although most persons view their subject matter through the filter of the mathematical tools in which they are trained, the best mathematical techniques for a specific analysis depend upon the best match between the algebraic logic and the underpinning physical dynamics of a theoretical system. "*

The problems we face today are ever more demanding of the necessary EM change pointed out by Barrett and Grimes, along with many others. Let us hope their words are heeded. Given rapid development of higher group symmetry electrodynamics models, we shall very rapidly see the solutions to many previously intractable human problems — including cheap clean energy from the vacuum, understanding and directly using antigravity as well as gravity, a revolutionary new medical therapy, a new chemistry and biochemistry, and a new appreciation for the electromagnetic interactions of mind and matter.

Let us hope this new scientific revolution comes swiftly.

Chapter 4

Setting the Stage for Understanding Overunity Power Systems

[Dissipative structures] *"This leads precisely to a distinction between 'equilibrium structures' [which may be understood in terms of classical equilibrium thermodynamics] and 'dissipative structures'. The latter are formed and maintained through the exchange of energy and matter in non-equilibrium conditions."* [I. Prigogine] {236} [Internal brackets added for clarity.]

[The energy in space itself] *"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing energy and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea of ghost particles of all varieties — virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Paul Davies] {237}

[Modernizing the fluid analogy of EM]. *The intensity of the vacuum virtual particle flux may be treated as the intensity of a potential, thus by analogy corresponding to pressure in a fluid. Polarity (charge) is a process for establishing a change in the intensity of the vacuum potential — and thus a change in the ambient pressure of the vacuum fluid. Difference between vacuum pressures at two separated points creates a force upon any charge placed between those two points. Just as a differential in pressure between points in a fluid induces currents, a differential in vacuum pressure induces energy flow currents in the vacuum fluid, which are called 'fields'. All*

EM circuits, charges, potentials, and forces in electrodynamics involve altering the vacuum to include altering its pressure and thereby producing the forces that the resulting pressure gradients (fields) induce upon charges. All observable EM energy in space is EM energy intercepted and collected on charges, and it comes from the vacuum via these fundamental vacuum-engineering dynamics." [T. E. Bearden, private communication to a colleague, 2001].

[Impact of EM energy from the vacuum.] *"If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete — and so would many of our worries about environmental pollution." "Don't sell your oil shares yet — but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development: 1. It's crazy! 2. It may be possible — so what? 3. I said it was a good idea all along. 4. I thought of it first."* [Arthur C. Clarke] {238}.

[Broken symmetry of opposite charges.] *"Since nonobservables imply symmetry, any discovery of asymmetry must imply some observable. The experiment of Wu, Ambler, Hayward, Hoppes and Hudson... established the asymmetry between the positive and negative signs of electricity."* [T. D. Lee] {239}.

[On implications of asymmetry of opposite charges of a dipole.] *"In particle physics, the proven asymmetry of opposite charges establishes that the source dipole — with opposite charges on each end—* is *a broken symmetry in the violent virtual photon flux of vacuum. The dipole charges constantly receive and absorb virtual photon energy from the seething vacuum, and re-emit that excitation energy. By the very definition of broken symmetry, some of this absorbed* *virtual* *energy must be integrated into* *observable* *3-space EM energy, and emitted by the dipole in observable form, in all directions in 3-space. Hence there is a rigorous basis for the source dipole extracting virtual EM energy from the vacuum, integrating it, and continuously pouring it out in* *observable and usable EM energy form. It has been in particle physics since 1957, being certified by the award of the Nobel Prize to Lee and Yang in that year. "* [T. E. Bearden, private communication to a colleague, 1995.]

[On powering an electromagnetic circuit] *"In the United States there is not now, and there never has been, a single electrical engineering department, professor, or textbook that knows and teaches what directly powers an electromagnetic circuit. Neither is it in any publication of the National Academy of Sciences, National Science Foundation, National Academy of Engineering, Department of Energy, or the great national laboratories. Yet the proof that every circuit is powered by EM energy extracted directly from the vacuum has been in particle physics for 45 years, since the award of the Nobel Prize to Lee and Yang in 1957, following the experimental proof by Wu et al. earlier that same year. "* [T. E. Bearden, private correspondence, 2002].

4.1 Introduction

In this chapter, we present a selection of potential overunity notions, some inventions, and some experiments. Its purpose is to show some of the more relevant work that has been done, and hopefully to stimulate new thinking by the interested reader. Much of the work addressed is by other inventors and scientists, and we give an appropriate commentary from our viewpoint. Not all the concepts presented can be clearly explained, though some can and are. This is not intended to be a "kit of parts" or an "assembly instructions" type of presentation. For that, the interested researcher will have to rely on his or her own ingenuity and lots of bench experiments, because no such "kit of parts with instructions" for COP>1.0 electrical power systems exists at present {240, 241}.

Because of the broken symmetry of its opposite charges, we do know that any dipolarity — such as a permanent magnet, a charged flat-plate capacitor, an electret, or a simple EM scalar potential — continuously pours out EM energy circulation extracted from the vacuum. E.g., see Figure 4-1, which shows a very simple "free energy generator" that will generate a flow of EM energy indefinitely, even by ordinary U(1) electrodynamics theory (by standard Poynting theory). E.g., quoting Buchwald {242}:

> *"[Poynting's result] implies that a charged capacitor in a constant magnetic field which is not parallel to the electric field is the seat of energy flows even though all macroscopic phenomena are static."*

This is easily seen by applying the standard Poynting formula, S = E x H. With the H-field of the magnet at right angles to the E-field of the capacitor or electret, the Poynting flow of energy S is maximized and orthogonal to E and to H.

The problem is not in extracting energy - even enormous energy — from the vacuum The problem is collecting and using the energy to power loads, without destroying the source dipole

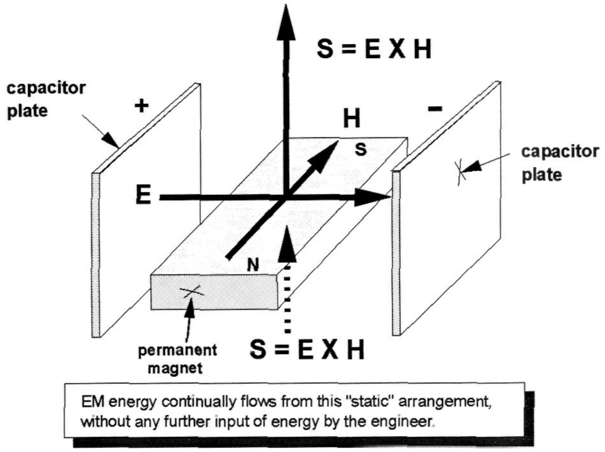

EM energy continually flows from this "static" arrangement, without any further input of energy by the engineer.

Figure 4-1 Poynting generator providing endless EM energy flow from the vacuum.

Buchwald errs, e.g., in assuming the macroscopic phenomena are static. They are not; they are highly dynamic in the same sense that a perfect macroscopic whirlpool in a river, seemingly static, nonetheless involves a continuous flow of water through it and a continuous dynamics. So far as can be observed, it appears there is no such thing as a truly "static" object. As an example, just to "exist" an object must "travel through time", which is itself a special kind of dynamics in the time domain.

A beautiful explanation of what is meant by "static" is given by Van Flandern[91]:

> *"To retain causality, we must distinguish two distinct meanings of the term 'static'. One meaning is unchanging in the sense of no moving parts. The other meaning is sameness from moment to moment by continual replacement of all moving parts. We can visualize this difference by thinking of a waterfall. A frozen waterfall is static in the first sense, and a flowing waterfall is static in the second sense. Both are essentially the same at every moment, yet the latter has moving parts capable of transferring momentum, and is made of entities that propagate."*

Now suppose that the capacitor in Figure 4-1 is rotated until E and H are either parallel or anti-parallel. In that case, the Poynting theory predicts zero EM energy flow. That is an incomplete assessment; it means that there is no overt field energy flow. There is, however, still quite a lot of covert or "infolded" field energy flow ongoing anyway, inside the static fields E and H themselves.[92] To see this, we first turn to the static potentials.

Between the charged plates of the capacitor or electret, there exists an electrostatic scalar potential ϕ. Between the polar ends of the permanent magnet, there exists a magnetostatic scalar potential Φ. We may independently decompose both ϕ and Φ via Whittaker decomposition {243}, applying a necessary correction {244} to Whittaker's interpretation. Then in all space surrounding the two dipoles, we have a hidden harmonic set of bidirectional longitudinal EM wavepair flows of energy comprising each of the static potentials ϕ and Φ filling all space[93] around the assembly in Figure 4-1. Each of those potentials ϕ and Φ is formed of a combination of scalar (time-polarized) and longitudinal photon pairs, as strongly argued by Mandl and Shaw {245} and as previously discussed in an earlier chapter. Further, each of the "static" fields can be decomposed into two scalar potential functions, as shown in 1904 by Whittaker {246}. Then each of the base scalar potentials further decomposes into the set of time-polarized EM wave and longitudinal EM wave pairs. The result is that every "static" field or potential in the universe is actually a set of longitudinal EM wave dynamics.

---

[91] Tom Van Flandern, "The speed of gravity — what the experiments say," *Phys.* ***Lett.*** *A.*, Vol. 250, Dec. 21, 1998, p. 8-9.

[92] Indeed, by the asymmetry of its opposite charges, each of the dipoles is co**ntinuously** pouring out EM energy extracted from the seething vacuum.

[93] In the conventional view. We corrected that view in {12}, to agree with quantum field theory and to use a *cause and effect* instead of the *two effects* used by Whiltaker.

Further, in each coupled Whittaker wavepair as reinterpreted, there is perfect correlation between the time-domain EM energy wave (scalar or time-polarized EM wave) and the 3-space longitudinal EM wave. So even a "static" EM field, such as from the permanent magnet and from the charged capacitor or electret, does involve an enormous number of hidden internal EM energy flows in longitudinal and scalar EM waveform.[94]

---

[94] One of the great omissions of modern Western electrodynamics is the failure to follow up on direct engineering of this "vast, internal, infolded, far more fundamental" scalar and longitudinal electrodynamics hidden inside all present overt EM potentials and transverse EM fields and waves. Russian secret weapon scientists under the former KGB called this internal electrodynamics the "information content of the field" and have strongly weaponized it. A deception plan has been used to successfully convince the West that the Russians are speaking only of conventional spectral analysis. Use of this internal EM structuring, in the form of longitudinal or scalar wave interferometry, formed the first giant superweapons deployed by the former Soviet Union in 1963. To see a proof that ordinary EM fields and waves and potentials can be formed at a distance, *arising out of the local spacetime in the interference zone,* see M. W. Evans et al., {79}. Longitudinal EM waves easily pass through the entire earth and ocean relatively undiminished, since mass is mostly empty space filled with fields, potentials, and waves — i.e., filled with incredible numbers and bundles and dynamics of longitudinal and scalar EM waves. Mass is thus a great superhighway for the passage of longitudinal EM waves, and longitudinal interferometry can be performed right through the earth and ocean, on the other side of the earth from the transmitters. Such interferometers can also easily focus inside the earth or beneath the surface of the ocean. Hence in April 1963 the first giant Soviet LW interferometer deployed was used to kill the U.S.S. Thresher beneath the ocean, off the East Coast of the U.S. Intense production of EM signals — emerging everywhere in the space in that underwater interference zone placed around the submarine — filled the sub and its electronic controls. This jammed the electronic controls of the sub, rendering it helpless. The hapless sub then sank to crush depth and imploded. The signatures are clearly present in the jamming and cessation of multiple electronic systems functioning in the sub's surface companion, the U.S.S. Skylark (which was in a sort of surrounding "splatter" zone and survived because it was not individually targeted). Today at least 10 nations have scalar or longitudinal wave interferometer weapons, to one extent or another, other nations are working on them, and the Yakuza also has them. These were apparently the weapons

Ultimately there is no frozen waterfall, but only the dynamic waterfall always reappearing the same but filled with dynamic movement and energy flow.

These latter EM energy flows *infolded inside* the scalar potentials and the **fields** exist for a single charged capacitor or electret alone, or a single magnet alone. They also exist for any two points in a circuit that have differing electrical or magnetic potentials with respect to the ambient vacuum potential (energy density). This latter fact, of course, was part of the secret of Kron's *open path.* Indeed, every dipolarity in the universe — between any two potentials or any two differing charges anywhere — is an open path and an energy flow source. E.g., quoting Kron {247}:

> *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of two nodes. (Previously— following Maxwell— engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search. "*

As can be seen, Kron had discovered how to create currents — *currents could be made to flow* — in individual branches between any two nodes in **the** circuit, physically connected or not. We point out that deliberate use of Kron's "open path" in a network circuit constitutes a strong violation of **the** standard closed current loop circuit's self-enforcement of the Lorenz/Lorentz symmetrical regauging condition. Hence it is a magnificent method for achieving COP>1.0 circuits. It is the belief of this

confirmed in 1997 by Secretary of Defense Cohen in these words: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So* there *are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts."* [Secretary of Defense William Cohen at an April 1997 DoD counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.]. As the reader can see, the Cold War **was** not so "cold" after all. This mere "tip of the iceberg" should also give the reader a little more insight as to why a war on terrorism is so necessary to all civilized nations.

author that such "open path" functioning in his patented circuits was part of Tesla's secret, though he did not think of it in those terms. However, Tesla's ability to shuttle potential energy independently in his circuits — invisible to a tensor analysis — was rigorously shown by Barrett {124}. Barrett extended the method and obtained two patents in the field of communications {286a, 286b}.

Thus the COP>1.0 problem *is not* in the part or function (the dipole or the charge) that extracts real EM energy flow from the vacuum. That is extraordinarily simple and straightforward. Engineers and scientists have been using it for two centuries without understanding what they were doing. Every charge, dipole, charged capacitor, electret, and permanent magnet already continuously extracts EM energy from the vacuum. The two-fold COP> 1.0 problem is in (i) *separately intercepting and catching (in an external circuit) some of the copious free flowing energy from the "extractor"*, and then (ii) *separately dissipating that collected energy in an external load without killing the "extractor" itself* — or at least only killing it more slowly than the load is powered. The most elegant way to violate the closed current loop circuit is with Kron's open path, typified by Tesla/Barrett potential energy shuttling. Since vector and tensor analysis will not even show it or model it, then a higher group symmetry electrodynamics is imperative in modeling COP>1.0 EM systems and their functions. Much of that work remains to be done.

Our main purpose in this chapter is to point out some selected efforts and approaches by scientists, engineers, and inventors. We will give a brief description and a few details, and leave it to the researcher for further experimentation and reflection. Hopefully, knowing what some others have tried and the results they achieved will prove helpful to the serious researcher.

We also strongly encourage the reader to again read this section after studying Chapter 9. In any COP>1.0 system, the excess energy must come from the "active environment". The active environment of an EM system consists of (i) the active vacuum and its interactions and dynamics, and (ii) the active local spacetime curvatures and their interactions and dynamics. Contrary to naive belief and the received view, overunity electrical power systems — particularly very high COP systems — are neither simple nor easy, but are extraordinarily difficult and highly nonlinear even when their appearance is deceptively simple. Chapter 9 will give a brief expose of some of the difficulties for higher C OP systems. There we will refer again to the intricate circuit analysis work of Gabriel Kron {248}, to Dirac sea electron theory {249} both prior to observation and after observation, and to the need for higher group symmetry O(3) electrodynamics by Evans {250} which is a subset {251} of the electromagnetics used in the unified field theory of Sachs {252}.

Chapter 9 indicates how mathematically complex these systems really are, even when they appear deceptively simple. The phenomenology is highly complex, and it has not yet been scientifically explored because of the erroneous received view that COP>1.0 EM systems are not permitted by Maxwell's theory or the laws of physics. That is a non sequitur. More exactly, *they are not permitted in that equilibrium class of Maxwellian systems described and selected by Lorentz's arbitrary symmetrizing of the Maxwell-Heaviside equations, following their previous symmetrizing by Ludwig Valentin Lorenz* {253}. Prior to applying the Lorenz-Lorentz symmetrical regauging, the Heaviside equations — which in field form are already a truncation to four vector equations of Maxwell's 20 quaternion-like equations — do prescribe both disequilibrium COP>1.0 and equilibrium COP<1.0 Maxwellian systems. The variables are not separated (potential formation) and the equations are difficult to solve, usually requiring numerical methods. With Heaviside's truncation and Lorenz-Lorentz regauging applied, the resulting "special case" equations — now for only that class of COP<1.0 Maxwellian systems that receive and use no net input from their active environment — continue to be used because closed analytical solutions can usually be had and numerical methods can be largely avoided.[95]

So with tongue in cheek, we humorously remark that COP>1.0 electrical power systems have not been developed or known because the early scientists were a bit lazy and wished to avoid ubiquitous use of numerical methods. These days, with computers, proper math programs, and sharp graduate students to easily crunch the numbers, there is no longer any reason for continued arbitrary discard of Maxwellian COP>1.0 systems.

---

[95] For some reason, electrodynamicists ignore the fact that the Lorentz regauging does assume two *additional* equal and opposite energy exchanges of the system with its external active environment, but carefully selected so that all the excess energy received in the system is purely in the form of a stress potential. Hence this energy continually does internal work on the system to produce additional stress, but cannot translate electrons through an external load to power it. Also, this regauging condition represents an effective rotation of the frame of the described system out of the laboratory frame. To then argue that this severely altered system (increased stress, constant additional internal work being done on the system, frame rotation, altered potential energy) is the same system prior to Lorentz regauging, and that such a system is not receiving excess energy from its active vacuum environment, is naive in the extreme and a non scquitur.

For such reasons, no legitimate theory of COP> 1.0 EM power systems has been developed, except for light shed by Prigogine {254a-254f} and others {255a-255c} as to the necessary active environment, disequilibrium nature, and disequilibrium thermodynamics of COP>1.0 systems. Further, the energy flow theory (Poynting's theory) has not been further extended to include the Whittaker decompositions {243, 246} as should long ago have been done. Neither has it been extended to account for the arbitrarily discarded giant Heaviside nondiverged energy flow component.

The researcher should be aware that the energy input to a system is always in the voltage[96] (the *potential;* one input volt of potential will cause one joule of energy to be collected upon each interacting unit point electric charge in the system; one volt of drop in potential results when one joule of collected energy is dissipated per unit point collecting charge in the system.). The fields and their energy are functions of the potentials and their energy, not the other way round. After potentialization of the circuit, the subsequent movement of the potentialized Drude electrons, as current in the system, is connected with (and part of) the *dissipation* from the circuit of previously collected/intercepted energy from the vacuum.

4.2 Demonstration Gedanken Experiment

4.2.1 Work and Energy

Doing work is defined as *changing the form of energy.* No energy is lost per se when work is done; the energy is only changed from one form to another. All the energy since the beginning of the universe is still present,[97] and it is still repeatedly changing its form to do work, over and over. The point is, a single initial joule of energy can be "used" (changed in form) to do more than one joule of work, if multiple changes of form occur serially. *One joule of energy changing its form gives one joule of work, only if a single and total change of form of that joule of energy occurs and then no further change of form occurs.* If the joule changes form once, a joule of work is done and a joule of energy still remains in a second form. If this second form of the joule of energy is not lost, but is retained and then changed in form, a second joule of work is performed and there is still a joule of energy left in different form. And so on. Tricking a system to self-induce multiple iterative form changes of each joule of input energy (or at least of some fraction of each input joule), without losing the energy after each change of form, is one way of providing COP>1.0 in a system.

---

[96] The reason is that all 3-space EM energy — in the circuit or elsewhere — must come from the time domain to each 3-space point and return from that point to the time domain. The scalar potential (voltage) actually is an ongoing process whereby, for any interacting charge, energy in the time-domain and energy in 3-space are functionally combined in that 4-circulation at every point dipole in space. In quantum field theory, e.g. as shown by Mandl and Shaw {19}, the *combination* of the time-polarized photon and the longitudinal photon in 3-space is observed as the instantaneous scalar potential, but neither photon is *individually* observable. We comment that an "observable" 3-space potential is the result of an iterative set of *d/dt* differentiations of the ongoing "combination" of scalar (time-polarized) photons having their energy oscillations in the time domain, with longitudinal photons having their energy oscillations in 3-space along their line of motion. Hence each *d/dt* differentiation of the ongoing 4-space combination yields the instantaneous 3-space potential. Cumulating, one has the ordinary 3-space potential or "voltage" of macro electrodynamics.

Remember that COP (coefficient of performance) is best defined as *the usable work output obtained, divided by the EM energy that the operator inputs.* Since "usable work output" is just usable "change of form of the energy", it can be seen that multiple changes of form of every joule of energy input by the operator can lead to COP>1.0.

This illustrates that one must be careful in applying the law of conservation of energy. For a thorough analysis, several features must be identified and accounted: (i) The flow of the energy impinging upon the collection, (ii) the collector itself and the type of collection process utilized, (iii) the linear or nonlinear operation of the collection/collecting process, (iv) the dissipation process used to dissipate the collecting mechanism, (v) what happens to the energy flow that was collected and then its collection "dissipated", (vi) what type of energy flow was collected, and then its collecting "dissipated", (vii) what happens to the collector after the dissipation interaction, (viii) what happens to the energy in different form after each previous change of form, (ix) whether single-pass or multiple-pass of the energy flow is used, (x) whether single collection or multiple collecting is used, and (xi) the effects of any nsymmetrical regauging of a local component or region (e.g., as in simply increasing one of the potentials of the region) that in turn acts as a free energy flow.

---

[97] However, "present" means in spacetime, not just in 3-space. The EM energy we think of as being in 3-space is macroscopically in continuous circulation from the tine domain to 3-space and back to the time domain. Microscopically it is in *continual* circulation.

See again Figure 2-4 in Chapter 2. In general, the available energy flow surrounding a circuit will be far greater than the intercepted and collected energy in that circuit, particularly in a "single pass" collection — i.e., where the impinging energy only passes once. In analyzing a system or circuit, one must be very careful to ascertain whether the system utilizes any mechanisms to enhance and increase the energy collection fraction. If it does, then a necessary (but not sufficient) criterion for COP>1.0 has been identified in the system. Letokhov's negative absorption of the medium (i.e., excess energy emission) {256} and Bohren's experiment {257} are examples. Bohren's experiment is easily replicated by any nonlinear optics lab, and outputs 18 times as much energy as the conventionally calculated energy input.[98]

If the impinging Heaviside nondiverged component of the energy flow beyond the collecting/collector is rerouted and passed back to impinge again in the system — even on another collector or other collectors — this must also be taken into account.

Conventionally, the "magnitude" of the fields[99] E and H in the Poynting (collected) energy flow S = E x H are "defined" by the collection from the fields by an assumed unit point static charge at each point in space. In short, the conventional fields have already been "defined" as a standard linear collection upon standard charged *mass, from the field entities as they*

---

[98] Energy conservation is not violated! Instead, the conventional calculation assumes a static intercepting unit point charge, which has the minimum reaction cross section in the stream of bidirectional longitudinal EM energy flows comprising the potential, etc. By placing the intercepting charged particles into particle resonance, the particles' reaction cross section is increased because each resonating charge sweeps out a greater geometrical cross sectional area. Hence the charges collect more energy from the energy flow comprising any "field" (actually field intensity) or "potential" (actually potential intensity) than is assumed in the standard input energy calculation. In short, Bohren's experiment takes advantage of that arbitrarily discarded Heaviside energy flow component that remains and is present but not considered in the conventional Poynting calculation and field intensity calculation. The experiment in fact proves the existence of the unaccounted but ubiquitous giant Heaviside nondiverged EM energy flow.

[99] At best, the 3-space point intensity of each 4-space field is so defined, for a forthcoming reaction with unit point static charge, and where the forthcoming reaction does not involve lingering time - charging but only simple time-excitation followed by full decay of the time-excitation via normal photon emission.

exist in 4-space prior to interception and collection, and where the collector itself is absolutely fixed and "static". So nonlinearity in energy flow collection — by additional disruption of the impinging energy flow induced by the dynamics of the collector itself— may easily increase the "reaction cross section" of the collector. The static reaction cross section of a unit point static coulomb and the dynamic reaction of a unit point resonant coulomb may differ appreciably.

Without stating it this clearly, Bohren has shown that the interaction cross section for a single-pass collection by one static particle may be increased by a factor of 18 in a given example, *by inducing resonance and thus increasing the "collection" dynamics of the collecting particle* in the impinging S-flow {257}. The work has been independently substantiated by Paul and Fischer {258}. It also is a direct proof of the existence of Heaviside's neglected nondivergent energy flow component surrounding every field and charge interaction.

So indeed there are COP>1.0 EM processes in the literature. The Heaviside component also pouring from the terminals of every generator and battery, already proves that *every generator and battery ever made already was and is a COP> 1.0 energy transducer, because far more energy pours from the terminals of the generator and battery than the shaft energy input to the generator or the chemical energy dissipated in the battery.*

4.2.2 Extending the Work-Energy Theorem

The standard work-energy theorem in mechanics may be conventionally stated as: *The change in the kinetic energy of a particle is equal to the net work done on the particle by the net force acting upon it.* As in any "model", here there are assumptions that are seldom taken into account. First, the collected/collecting kinetic energy of a particle is due to its increased interaction rate with the vacuum flux, because of its acceleration through that flux.[100] In mechanics the increase in flux interaction (and hence "Newtonian 3rd law resistance" is just treated as a sort of resistance to acceleration, and known as *inertia.*[101] The actual energy of the vacuum flux interaction rate is a function of the square of the particle velocity through that flux — hence the formula $K = 1/2\ mv^2$ for the kinetic energy K of a mass m moving through the flux of space at velocity v. Relativistically, the mass of the particle is moving through the ambient potential flux, or we may view it from the particle's frame of reference where the ambient potential flux is moving through the particle but interacting with it. Only a single pass of the flux impinges upon the "particle collector", so there is only a "single-pass" type of energy collection ongoing.

---

[100] As we cover elsewhere in this book, any charge (which, together with its clustering virtual charges of opposite sign is a dipolarity) is a broken symmetry in the fierce flux of the vacuum. This means that every charge in the universe absorbs virtual energy from the vacuum, coherently integrates some of it into observable energy, and pours out that observable energy in all directions. The negative charge does that for positive energy; the positive charge does it for negative energy. The latter can be better interpreted as the return of positive observable energy from space to the positive charge, thence disintegration of that positive energy back to virtual energy, and subsequent re-radiation of that virtual energy back to the vacuum. One **can** in fact model the **Bohren** experiment in terms of the extra vacuum flux

All energy in the vacuum exists in "vacuum flux" or "change in vacuum flux" form. There is no force in the vacuum, although there is polarization that will produce a force upon a charge when one is introduced. The force occurs only after the "change in vacuum flux" that we call a field is interacting with a mass. The force is *already* a change to the mass, a priori, since mass is a part of force by $F = d/dt\ (mv)$. If energy is then added by a AK energy flux (usually mistakenly called a force) performing work upon mass m, this constitutes an additional energy flux or flow impinging on the

---

encountered and transduced by the resonant charge, as compared to flux encountered and transduced by the static charge. In the same vacuum flux density, a resonant charged particle will intercept and absorb more virtual photon energy than the same particle in static condition. Hence its broken symmetry will convert more absorbed virtual photon energy into real observable emitted photon energy.

[101] If one wishes to understand that inertial reaction, here is how it works. Consider the virtual charges of the vacuum as an analog to the Drude electron gas in a conductor. At the instant an observable charge starts to move, it is momentarily affecting all the surrounding virtual charges — an enormous number, which are initially repelling and attracting it in all directions intensely. To move, the observable particle must create a change in that entire participating virtual particle "gas". Just as the Drude electron gas, the vacuum "gas" of virtual charges has a certain relaxation time constant. So a finite time delay occurs before the vacuum gas relaxes and movement of the observable charge occurs. As the relaxation continues apace, the particle being acted on by a force accelerates. Even mass with a net charge of zero is filled with charges, and even the neutrons in the atoms are continuously changing into protons and back, etc. So all mass continually exists in the "charge and charge-changing" state. Inertia is in fact due to the relaxation time curve of the vacuum virtual charges "gas". We dealt crudely with this in Bearden, "Quiton/Perceptron Physics," NTIS Report AD-763-210, 1973, p. 11; hopefully a far better treatment will eventually be undertaken by qualified higher group symmetry electrodynamicists.

particle, imposing an additional "energy collecting" and gradient across the mass. Rigorously, the dynamic system comprised of the dK gradient coupled to the particle, and interacting with it, identically is a force.[102] Depending upon the situation, it may accelerate or decelerate the particle, or change its direction, or a combination. This is summarized in the usual work-energy formula

$$W = K_f - K_i = \Delta k \qquad [4\text{-}1]$$

where the work W that was done is given by the change dK in kinetic energy of the particle from its initial kinetic energy K_i to its final kinetic energy K_f.

In applying the conservation of energy law, usually physicists utilize either the work-energy theorem or an approach closely paralleling it. This application is valid in any linear single-pass collection situation. It may not be valid — and usually isn't —for a multi-pass, multicollection system where the same energy, changed in form to accomplish work in a single reaction, is then changed in form iteratively, again and again, in the same system or material. In such case, one joule of input energy can and will accomplish multiple joules of work in and on the system. So for our purposes we must modify the work-energy theorem to give

---

[102] We stress again that the hoary old mechanics, some four centuries old, is seriously in error in assuming a separate force acting on a separate mass. That remains an inexplicable non sequitur in modern physics, in its very foundations. We call attention to how classical electrodynamicists' wrestle with the problem. E.g., quoting Jackson, *Classical Electrodynamics*, 2nd edition, Wiley, 1975, p. 28: *"...the thing that eventually gets measured is a force... " "At the moment, the electric field can be defined as the force per unit charge acting at a given point. It is a vector function of position, denoted by E."* Our comment is that the "definition" is in error, including the use of the word "field" for "field intensity", when field and its intensity are two entirely different things. A proper definition should be stated as *"...the observable electric field intensity can be defined as the force per unit static charge and mass, existing at a given point in an interacting gradient of the virtual particle flux of the vacuum, as inferred from the changes induced upon the unit static charge and mass. "* The field intensity is a vector function of position, charged mass, and vacuum virtual particle flux gradient. The standard definition substitutes an effect (of collection) for a cause (prior to collection), and this error is widespread through physics, from mechanics to electrodynamics and on into particle physics. On p. 249, Jackson shows how the electrodynamicists largely capitulate and continue to promulgate the substitution of effect for cause. Quoting: *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."* We comment that actually mass is a component of force, as shown by $F = d/dt(mv)$.

$$W = \lambda(K_f - K_i) = \Delta k \qquad [4\text{-}2]$$

where λ is the averaged energy collection fraction, adjusted for multipass, multicollection and also for asymmetrical self-regauging (self-increase of the potential in, and hence the energy flow from, one or more regions of the active system's source). In formula 4-2, K_f and K_i and Δk still refer to single-pass collection only.

In the new work-energy theorem, $\lambda = 1.0$ for that vast number of conventional cases that have successfully utilized the conventional form of the theorem. These are almost always single-pass, single collection systems. The new COP>1.0 systems have deliberately enhanced energy collection — and hence multiple *energy collection and work amplification*[103] — such as multi-pass collection. For these COP>1.0 systems, the expanded work-energy theorem form applies and in general $\lambda > 1.0$. In an ideal COP>1.0 system, $\lambda >> 1.0$. In some versions of the Patterson Power Cell®, for example, $\lambda = 1200$. In the Sweet device (discussed later), a very high value of lambda resulted, so that $\lambda = 1.5 \times 10^6$. We sometimes refer to λ as the *energy amplification factor, energy collection fraction, Poynting amplification, collection amplification,* etc. Bohren and others have referred to it indirectly as the *interaction cross section,* the *resonance absorption cross section,* and the *negative absorption cross section.*

Again, we strongly iterate that *there is no conservation of work law in physics!* When energy is input to a system, acts on it, and is "dissipated" by a change of its form (doing work), the "dissipated" energy is still available after the work (change of the energy's form) is accomplished. However, usually by "dissipated" energy we mean energy that has permanently left the system (or a component in the system) and cannot be recovered and used by the system. The original input energy is not work, because it did not undergo any change of form prior to interaction with the system to change its form. So one does not do "work" as input, which is a non sequitur. Any work accomplished in the input section of a system is due to inefficiencies of the system,[104, 105] usually allowing extra change in form of some of the energy in non-useful processes called "input losses". As we stated, it is quite possible for a single joule of original input energy to perform many joules of work. During the subsequent billions of years following the formation of the observable universe and the energy in it, every joule of the primordial energy is still here. And every joule of it has been steadily performing joule after joule of work since then. Further, it has continually been recycled between 3-space and the time domain and back, and is continually being so recycled today.

---

[103] There is no "conservation of work" law in physics or thermodynamics. It is energy that must be conserved. A single change in form of K joules of energy gives an equal number of K joules of work done upon the interacting mass system. However, there remains an equal number of joules available, just in a different form. If that second K joules of energy is then changed in form again and not allowed to escape from the system, an additional K joules of work is performed in the system, without the operator having added any extra input energy. And so on, until the inefficiency of the system and some inevitable losses allow all the "recycling" energy to escape from the system. We strongly believe that ignition, quenching, and re-ignition phenomena such as in gamma ray bursters and x-ray bursters are due to exactly such processes where iterative change of form of each initial joule of energy occurs.

Suppose input energy to a system has been dissipated (changed its form) once, doing work (single change of form of energy) in the system. If the dissipated energy in a different form is then recovered or partially recovered — without escaping the system — to change its form yet again in the system and produce additional useful work, that system is

---

[104] Almost always due to a negative feedback mechanism such as Newton's third law, self-regauging symmetry, the closed current loop circuit, inertia as due to a self-regauging 4-symmetry, etc.

[105] E.g., a transformer would be an overunity system, because the current in its primary is separate from the current in its secondary, if these two currents were completely decoupled and independent. In that case, the transformer would simply propagate and "couple" or pass energy emitted from primary to the receiving secondary, without back-coupling. However, the closed current loop in the secondary, routing all spent electrons in the secondary back through the back mmf of the secondary coil, produces a back-field coupling interaction with the primary, and this "back mmf/back emf" coupling requires energy dissipation in the primary section to overcome. In short, the transformer is designed so that the secondary lights the primary to a draw and the designer loses. With some losses and inefficiencies in the device, that makes the transformer a COP<1.0 device (unless free external energy is received in the input). The interested reader might extract several hints and suggestions in this book, passim, of things that overunity researchers have tried in their search to break up the back-coupling of a transformer. The simplest way, of course, would be to place a true negative resistor shunt across the secondary, so that part of the secondary current does not return back through the primary.

considered to be an open system far from single-dissipation equilibrium with its external environment.[106] In that case, instead of the first dissipation simply "returning" the energy to the external environment, the energy is recovered by another change in form (which may simply be a change in direction in the internal environment). The thermodynamics of an open system far from equilibrium with its active environment applies, whether that "environment" is internal or external to the system. It is "external" to the subcomponent of the system forcing the "local change of form" of the energy.

4.2.3 Gedanken Experiment: Multicollection by Regenerative Feedback

After a COP>1.0 collecting process with subsequent proper discharge of the energy, more work output as "dissipation of the collected-energy" exists than the energy being input by the operator or experimenter. By regenerative, clamped, carefully controlled positive feedback of a portion of the output *positive* energy[107] to the system input, the system can be converted to a self-powered (self-asymmetrically-regauging) system that powers both itself and its loads. However, see Chapter 9 for special considerations of the Dirac sea hole current and excess negative energy that arises for COP»1.0 EM systems. Positive clamped feedback will fail dramatically for any COP»1.0 EM system, and for many COP>1.0 systems it will also fail unless conversion of negative energy — arriving at the input section — to positive energy is accomplished.

To ease visualization of the energy multicollection process, consider a gedanken experiment iteratively retroreflecting a steady small EM energy input. Particularly see Figure 4-2.

---

[106] The vacuum environment is also "internally" penetrating throughout any macroscopic system, to the finest level! So the notion of separated "external" and "internal" environments requires strong qualification in any system engaged in iterative, work-amplifying interaction with its local vacuum.

[107] The output may contain appreciable negative energy as well as positive energy, however. In that case, in the feedback loop the negative energy must be transduced into positive energy by regauging, else it will appear in the input section of the system as an extra system load there in the input itself. It will "eat" incoming electrons from the external power supply, so that the external power supply has to power this "extra load" in the input section as well as powering the normal system load presented. This reduces the system COP to COP<1.0.

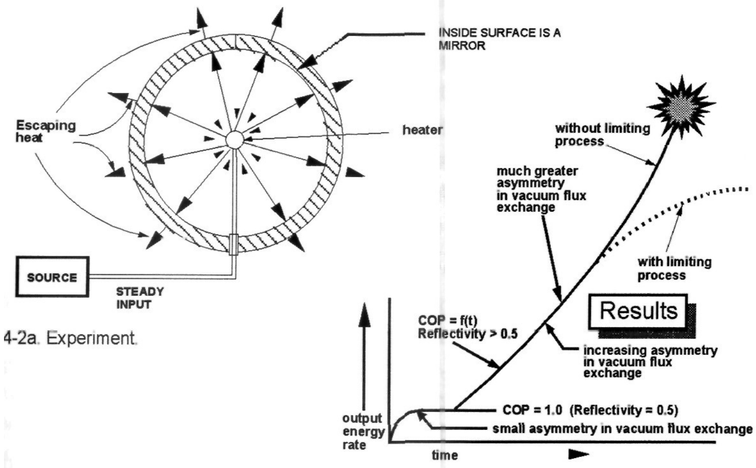
INSIDE SURFACE IS A MIRROR
Escaping heat
heater
SOURCE
STEADY INPUT
without limiting process
much greater asymmetry in vacuum flux exchange
with limiting process
Results
COP = f(t)
Reflectivity > 0.5
increasing asymmetry in vacuum flux exchange
COP = 1.0 (Reflectivity = 0.5)
small asymmetry in vacuum flux exchange
output energy rate
time

4-2a. Experiment.

4-2b Result

Figure 4-2 Heater wire in a hollow sphere with certain retroreflectivity properties.

As shown, a heater-emitter is placed inside a closed hollow sphere, through a tiny hole for the two wires connecting the heater wire into the external power supply (the external power supply is not shown). The heater emitter is 100% efficient; any joule of energy absorbed by it is re-radiated by it. One joule of energy (one watt of steady "input power" — conventional electrical engineering terminology where rate of work and rate of energy flow are confused!) is furnished from the power supply to the heater element inside the sphere. The inside of the sphere is coated with a substance that acts as a phase conjugate mirror (PCM) reflector with, say, 0.5 reflection coefficient. In other words, when scattered photons are emitted from the wire in all directions, they strike the inner surface of the sphere. Half are retroreflected precisely back to the spot on the resistor from whence they came, where they are absorbed and "re-collected" to further heat the resistor, causing emission of additional photons from it. The other half of the absorbed photons on the inner surface of the sphere diffuse through the cylinder walls as heat, and escape as heat radiation from the sphere to the outside world.

If that sphere's operation could be maintained intact and stable, without change of the physical characteristics and mechanism, then the buildup of energy in the heater wire inside the cavity — and the buildup of the energy icing emitted from the outside of the sphere — would be ever increasing. It would increase without bounds as time passed. Wait a short time, and the sphere would be outputting 10 watts of power while one would still be inputting one watt. Later, the sphere would be outputting 1,000 watts of power, while one would still be inputting one watt. And so on. Indeed, the rise in energy density of the output would be asymptotic, and would increase toward infinity. The only limitation as to the energy output for a steady 1 joule per second input, would be the limit at which the mechanism goes unstable and changes occur to dampen and curtail the process.

Before the reader objects on conservation of energy grounds, one should realize that this rise in the energy density inside the sphere is an example of iterative regauging. The principle of *gauge freedom* — one of the axioms of quantum field theory and well established — assures us that *the potential energy of a system can be changed at will, and freely.*[108] What the gauge freedom principle does *not* state is a *mechanism for providing the additional potential energy and a source for it* — nor do the quantum field theorists and electrodynamicists. Without such a source and mechanism specified, the conventional gauge freedom principle assumes total violation of the conservation of energy law.

---

[108] Conventional scientists are almost always very careful to regauge symmetrically, so that the two excess force fields that result are equal and opposite, and sum to a net zero resultant field. That "zero-summed system" comprises a stress potential *a priori.* So such symmetrical regauging alters the stress in the system and alters its potential energy, but only in the form of additional stress energy. There is no *net* force and "force field energy" that can be used to dissipate that additional free stress energy in an external load, thereby doing free work. This regauging is a purely arbitrary practice by the electrodynamicists. We are far more interested in "asymmetrical regauging", where only one potential is freely changed or the two are changed unequally, resulting in a *net nonzero force field* that can then be dissipated in an external load to perform free work. Only by asymmetrically regauging an otherwise inert system can usable field energy be added and any work then be done by it anyway. In conventional systems, we do the "asymmetrical regauging" by adding the voltage (potential difference). As current flows with dissipation of energy from the circuit in the loads and losses, the system is able to develop "power" and thus do work in the external load. Unfortunately, the electrodynamicists are still obsessed with symmetry, so they ubiquitously employ in all electrical power systems the closed current loop circuit. This guarantees that the circuit will kill its source dipole (the source of the potential and potential energy generated from the virtual flux of the vacuum by the broken symmetry of the opposite charges of the dipole) faster than it powers its load. That way, lovely symmetry is maintained, beautiful free energy and negentropy are avoided and tossed away, and ugly pollution and destruction of the biosphere continues worldwide to fulfill the ever-increasing and insatiable thirst for electrical power.

In our supersystem view, the gauge freedom mechanism providing the energy is straightforward: the excess energy is freely supplied from disequilibrium in the active vacuum interaction with the system and in the disequilibrium represented by the local curved spacetime. Since a change in system potential energy is also a free change in spacetime curvature, then it follows that simply changing the potential energy of the local vacuum (the vacuum is also an electromagnetic system!) and of the local curvatures of spacetime is also "for free", or it can be. One does not have to perform work oneself upon spacetime to curve it! The mere *presence* of a spatial energy change — including a free potential energy change allowed by gauge freedom — is sufficient to freely curve spacetime also. That is a real energy change, because any curvature of spacetime acts back upon mass. Higher group symmetry electrodynamics does indeed include vacuum energy and energy current {259a-259c}, and it is possible to extract useful EM energy from the vacuum.

But back to our sphere, where the internal potential energy of the sphere is rising asymptotically without bound.

In the real world, of course, the reflection coefficients and the materials characteristics will change as the energy density changes, the sphere will heat, etc., and these changes will start damping the perfection of the retroreflection process to limit it to some finite plateau value which may be a COP>1.0 or COP»1.0 condition. Or, the materials will melt or soften so that the sphere ruptures and explodes, sharply quenching the process entirely and emitting a violent burst of energy to signal the disruption of the localization of the process.

Nonetheless, a successful real bench experiment similar to this should be possible with some tinkering, and it should readily yield COP>1.0, after one waits for build-up and stabilization at some level. We would hope that a sharp young graduate student may eventually prepare a doctoral thesis on this experiment or a variant. The principle is demonstrable.

Similar buildups by regenerative feedback and multiple collections by the collectors do appear to occur in nature, up to and including such phenomena as gamma ray bursts, x-ray bursts, etc. These occur in some systems such as exploding gases (and in some cases, even in the upper atmosphere of the Earth). In such a system, the "physical particles in suspension in space" do move and continually disrupt the geometry for the increasing buildup, resulting in decay of the process after a time delay.

However, there is a finite time during which the movement of the particles is still insufficient to appreciably break the geometry and cause quenching.[109] During this "nearly linear stage", the build-up applies and the energy density very rapidly increases. Build-up occurs and continues while the relaxation time of the initial countering symmetry-restoring mechanism for the old or former state or condition (the initial reaction) is occurring. A well-known conventional example is the Lenz law reaction. Then as the geometry changes significantly, a rapid damping of the energy density increase occurs, quenching the asymptotic rise and "discharging" or "decaying" back to a less-than maximum condition. However, the decay position or state (in the quenched condition) will be different from the initial condition before asymptotic rise (regauging) set in. The system usually will have itself absorbed and collected additional potential energy, and so it will be "hotter" or more energetic. This is often called the "afterglow". In the simple case, parts of the system will be in greater motion (more energetic).

An intermediate plateau can appear as quenching occurs and it can stabilize. However, if no stable plateau is reached during the damping, then further rapid quenching and a rapid reduction of the increased energy density occurs as the geometry changes become decisive and break the asymmetric self-regauging. Thus the "damping" of the process may yield a plateau of steady COP»1.0 operation, or it may simply go immediately into full catastrophic quenching and decline back to some lesser level of COP>1.0 operation, but still more energetic than the beginning condition. The latter case produces a sort of "afterglow" - - observed in the gamma ray bursters, e.g. — of increased energy density from the beginning of the quenched phenomenon. Indeed, in that new system of more energetic gases or particles, once the new state is stable, continuing regenerative feedback versus the new geometry can then reinitiate or "re-ignite" another "burster" followed by subsequent quenching. And so on.

We submit that the gamma ray bursters fit this schema and are consistent with it. So do the phenomena observed in intensely scattering photoactive media on the nonlinear optical laboratory bench. We hypothesize that the gamma ray burster and similar violent burst emission phenomena are generated by this mechanism or a similar version of it.

---

[109] Regardless of what we call it or what it has been conventionally named, any impulsive, explosive process involves at least momentary broken symmetry, resulting in instant and countering broken symmetry. Then a relaxation time passes, and the countering broken symmetry is dissipated by change of the system to a new potential energy state, the new dynamics state, etc.

Figure 4-3 diagrammatically shows a proposed range of such "excess energy" emissions due to this regenerative positive feedback and multiple collection mechanism. As can be seen, this may place our view of very powerful astrophysical emission phenomena in a completely different light, arranged energetically by the length of the initial "nearly linear" phase of exponential increase. There is a scale of stability levels up to just before the Big Bang itself {260}. A so-called Big Bang, of course, would result when the "containment" ability of 4-space itself is breached, resulting in a rupture of 4-space and a consequent violent blow-out into n-space (where n>4), with a very rapid "false vacuum" created *outside* the blow-out region and pumping itself up by asymmetrical self-regauging in a "new" 4-space. When a new stability level is reached outside the blowout hole from the spawning 4-space, it represents an outside "new" 4-space and a new, infant 4-spatial expanding 4-universe, freshly born.

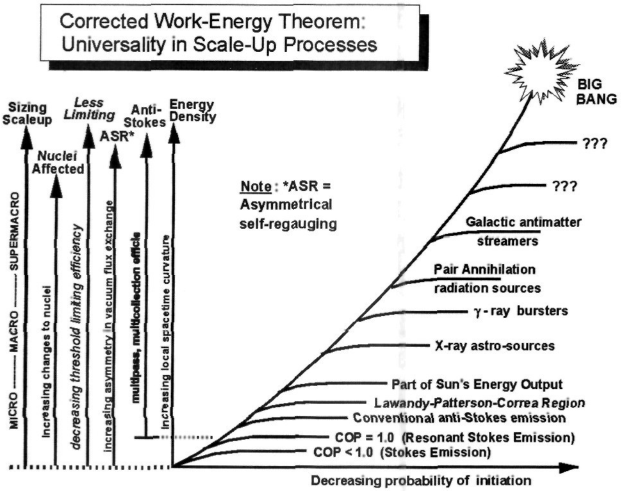
Corrected Work-Energy Theorem:
Universality in Scale-Up Processes
Sizing Scaleup
Less Limiting
Anti-Stokes
Energy Density
ASR*
Nuclei Affected
MICRO —— MACRO —— SUPERMACRO
Increasing changes to nuclei
decreasing threshold limiting efficiency
increasing asymmetry in vacuum flux exchange
multipass, multicollection efficie
increasing local spacetime curvature
Note: *ASR = Asymmetrical self-regauging
BIG BANG
???
???
Galactic antimatter streamers
Pair Annihilation radiation sources
γ - ray bursters
X-ray astro-sources
Part of Sun's Energy Output
Lawandy-Patterson-Correa Region
Conventional anti-Stokes emission
COP = 1.0 (Resonant Stokes Emission)
COP < 1.0 (Stokes Emission)
Decreasing probability of initiation

Figure 4-3 COP>1.0 **stability levels for the asymmetrical self regauging process invoked by multipass and** multicollection **of same energy.**

So from inside the spawning universe, the burster phenomenon would be so great as to "burst" 4-space, producing a sort of "super black hole" as seen by the spawning universe. We hypothesize that the asymmetrical self-regauging mechanism produces a great burst of EM energy in its universe if a stability plateau is reached that is below the threshold of 4-space rupturing. If such a stability plateau is not reached, the asymptotic rise in local spatial energy density increases until the primary 4-space ruptures to produce a "blowout" seen from the old universe as a sort of black hole. There is a sort of "white hole" in the new universe being spawned on the other side of the blowout. This white hole and its associated phenomena produce the birth of another 4-universe outside the original spawning 4-universe. At least we propose that as a hypothesis.

We propose that this could be a legitimate process for the birth of multiple 4-space universes in the infinite-dimensional cosmos.[110] It has an added advantage of accommodating the vexing question of "What existed before the beginning of time (in this universe)??? The answer then would be, *"The existence of time in another universe that birthed this one. "* We emphasize our interpretation of time as just a special form of EM energy, and not at all a "mysterious river down which we float helplessly as if in a drifting boat in the river's current."

4.3 Iterative Retroreflection and Collection in a Circuit

The normal circuit attached to a generator or other external power supply is the "external" portion of the overall closed current loop circuit passing back through, and including, the source dipole in the power supply. In a DC external circuit, we may consider that the energy flow pours forth steadily from the terminals of the generator (from the dipolarity between them) into space surrounding the external circuit. Since the external conductors act as a sort of guide, this *transmitted* external energy flow is generally parallel to the external conductors.

The external circuit further acts as a *receiver.* The surface charges and their fields intercept and interact with some of the transmitted energy flowing parallel to the conductors and outside them. This interaction diverts a component of that passing energy flow into the conductors, powering the Drude electrons. This powers the external circuit. The nondiverted component of the passing energy flow — usually much larger — escapes on out into space beyond the external circuit. According to Heaviside's energy flow theory {261}, this escaping nondiverged energy flow may be interpreted to exist in closed circulations.[111] Laithwaite

---

[110] After all, a "dimension" ultimately is just a fundamental variable in a physics model or, in other words, a "degree of freedom" in mathematical modeling. We may construct any number of dimensions we desire, merely by assigning a sufficient number of variables, including both overt variables and hidden variables, etc.

[111] We point out the interesting possibility of re-interpreting this "closed circulation" Heaviside component as the giant negentropic circulation of EM energy from the

{172b} has specifically commented upon the potential importance of this aspect of Heaviside's theory.

As we shall later address, Heaviside worked out but never published a draft theory of gravitational effects from this extra escaping energy flow {262}. Interestingly, since Lorentz's integration of the energy flow vector around a closed surface surrounding any volume element of interest, the discarding of this "excess nondiverged energy flow" has been ignored by electrodynamicists, and it has been ignored by gravitational theorists.

Consider a theoretically "perfect" generator, loss free. Consider the presence of the Heaviside nondiverged component emerging from the generator terminals — and filling all space around the external circuit. That is the part of the energy outside the circuit that is not caught and processed by the circuit, but is just wasted. By definition, in this "perfect" situation the magnitude of the Poynting or "caught" component of energy flow will precisely equal the magnitude of the mechanical energy input flow (rate) to the generator.[112]

But the presence of the additional huge Heaviside component rigorously means that far more energy pours from the terminals of every battery and generator than the amount of mechanical energy the operator inputs as shaft horsepower. This easily demonstrated fact — e.g., proven by the Bohren experiment {24} — continues to meet with a resounding silence from electrodynamicists and physicists, and particularly from those dogmatic skeptics so keen on defending the classical thermodynamics of closed or equilibrium systems as if they also applied to open systems far from equilibrium in their exchange with their active environment. We refer particularly to the dogmatists so supremely confident that COP>1.0 EM systems are prohibited by the laws of thermodynamics. These stalwart fellows would thus exclude every charge (and its associated fields and potentials) in the universe, were they logical and speaking scientifically. So they strangle at the gnat of COP = 4, and swallow COP = oo while proclaiming that COP>1.0 EM systems cannot exist.

The dogmatists have achieved a major nefarious goal, however. They have silenced most inquiry into such matters by electrical engineering

---

time-domain to the negative charge, to 3-space, to the positive charge, and back to **the** time-domain. To our knowledge, that has not been examined.

[112] That is merely a statement of the Lorentz symmetry condition self-enforced by the closed current loop circuit itself.

professors, young graduate students, and young postdoctoral scientists. They have glorified classical equilibrium thermodynamics into a religion rather than the partial and imperfect model it really is. They are directly responsible for the continuing rape and pillaging of the planet's resources, pollution of the biosphere, global warming, and the increasing death of species. They are indirectly responsible for the deaths of more human beings than Hitler, Mao, and Stalin combined. Most of the blame for environmental pollution problems concerned with power and electricity is squarely at their feet. Sadly, the environmentalists seeking to save this planet and biosphere have not yet recognized the true nature of the scientific Taliban that has generated the environmental problem. In tolerating its own scientific "terrorism", the scientific community seriously aborts much of the scientific method.

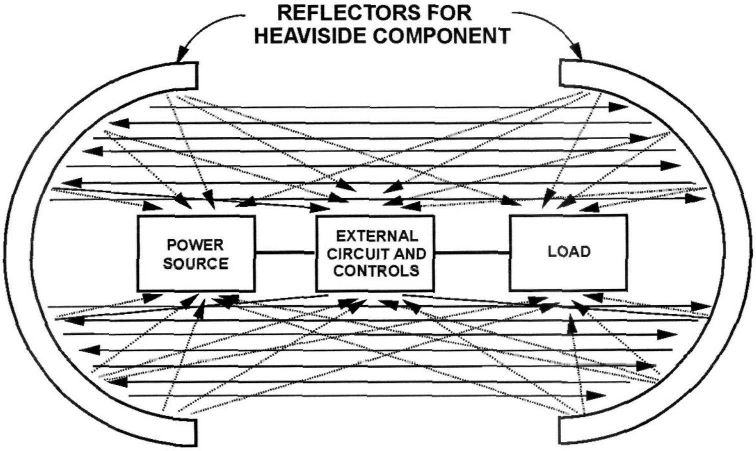

Figure 4-4 Iterative retroreflection of the "passed" **Heaviside component back and forth across a** circuit -- **thematic.**

See Figure 4-4. The neglected large nondiverged energy flow component need not be wasted. As an example, part of it can be captured and used, by applying near-field antenna techniques and such methods as Bohren's resonance increase of the intercepting charge's reaction cross section.[113] It is possible to intercept part of that excess energy that is there and available. E.g., even after it passes on by the immediate circuit, the Heaviside energy flow component can be intercepted beyond the circuit and retroreflected back across the same circuit again. In this way — at least in theory — extra Poynting energy can be collected and recollected by the same external circuit for a single given input energy.

---

[113] Or other methods we cannot yet divulge because of patent situation.

In essence, such an "iterative, multi-collecting circuit" is now fed with transmitted energy from multiple (iterative) paths simultaneously. This is *physically doable,*[114] but it is not a simple procedure, and antenna theory and especially near-field antenna theory must be utilized as well as focused retroreflection techniques. To directly model it, higher group symmetry electrodynamics (such as quaternion EM, O(3) EM, etc.) must be used. Precisely such phenomena are met in modern experiments in intensely scattering optical materials {263a-263b, 264a-264b} and in Bohren-type experiments where more energy is collected than the operator has input via standard calculations {257, 258}.

4.4 Tesla's Iterative Retroreflection Work

4.4.1 Tesla's Single Wire Circuit

See Figure 4-5, which diagrammatically shows one of Tesla's developments, which used a retroreflective single wire circuit between two large separated capacitances. This was the basic approach used in his magnifying transmitter and in several other circuits. Here the dipolarity exists between two accumulators, one at each end of a long conductive path with a load in the middle. AC operation is used, by setting up oscillation end-to-end along the "single wire" circuit, from accumulator to accumulator. Even so, no overunity operation will be experienced unless passive retroreflection of the normally wasted Heaviside component of each energy flow "passage" occurs back from each receiving end, so to speak {265}. That is possible if specialized dielectric material and construction is used in the end capacitors.

---

[114] E.g., some of Tesla's actual patented circuits do it, but that part of their functioning is not described by any vector or tensor analysis. However, a higher **group** symmetry electrodynamics analysis of the circuits will indeed show the effect clearly as a "shuttling" of potential (regauging) around the circuit at will. E.g., see T. W. Barrett, {124, 286a, 286b}.

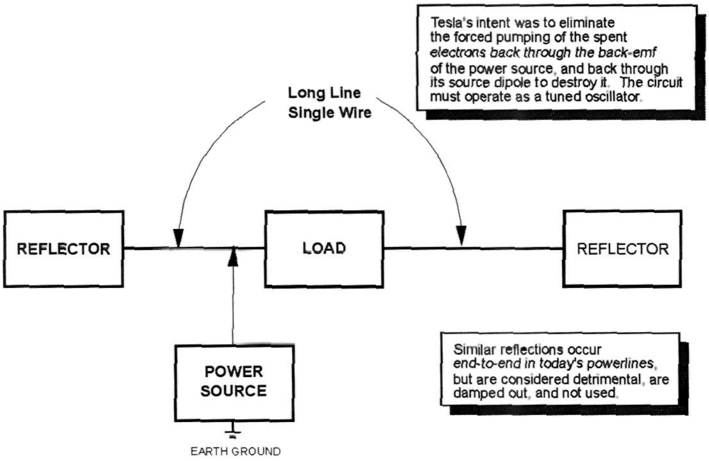
Tesla's intent was to eliminate the forced pumping of the spent electrons back through the back-emf of the power source, and back through its source dipole to destroy it. The circuit must operate as a tuned oscillator.
Long Line Single Wire
REFLECTOR
LOAD
REFLECTOR
POWER SOURCE
EARTH GROUND
Similar reflections occur end-to-end in today's powerlines, but are considered detrimental, are damped out, and not used.

Figure 4 5 Tesla's iteratively retroreflective single wire circuit.

Further, it may usually be necessary to place the individual charges comprising the "current surging to and fro" into particle resonance at either the frequency being transmitted or some subharmonic of it. This schema is doable, but it is difficult and most everything affects it, including even the impurities in the conductors.

4.4.2 Strain to Stress Potential Energy Transduction

See Figure 4-6. This is a diagrammatic representation of a fundamental "strain-to-stress" transduction mechanism we believe was involved in Tesla's "single wire circuit" (i.e., his iterative retroreflecting circuit).

At each end of the circuit, the electrons are "trapped" in the open outside end. Hence when EM energy rushes to the capacitor on one end of the circuit, the forward emf tries to strain the dielectric from the inner plate toward the outer plate. However, the trapped electrons in the outer open wire cannot move, so they produce an equal back-emf and back force trying to strain the dielectric from the outer plate to the inner plate. The result is that the translational field "strain" energy (dipolar charging) is transduced into stress potential (monopolar charging) energy. Electrons try to strain the dielectric from the inner plate toward the outside with a "forward emf so as to charge the capacitor normally, while the trapped external electrons simply push back the other way, exerting a back emf force that is equal and opposite to the straining force from the forward emf and simultaneously tries to "charge" the capacitor in the opposite direction.

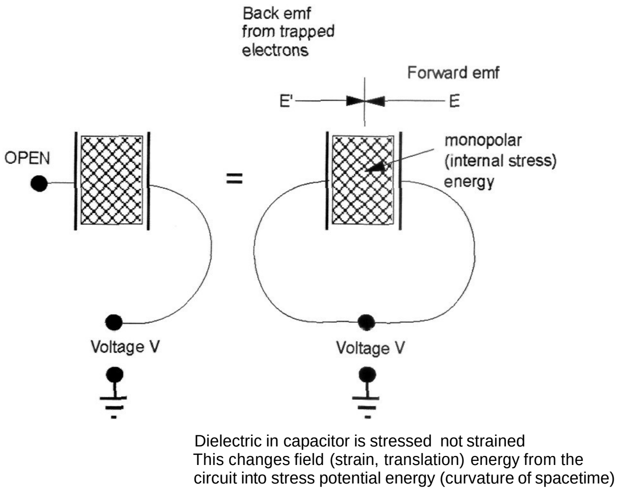
Back emf
from trapped
electrons
Forward emf
E'
E
OPEN
=
monopolar
(internal stress)
energy
Voltage V
Voltage V
Dielectric in capacitor is stressed not strained
This changes field (strain, translation) energy from the
circuit into stress potential energy (curvature of spacetime)

Figure 4-6 Transaction of translational field energy into stress potential energy.

Here is the importance of the "vector zero resultant" stress (monopolar) charging that occurs: The input EM strain field energy familiar in dipolar charging of a capacitor is converted to a dielectric stress (monopolar) potential for monopolar charging. This changes *translational* field energy (external energy) into *stress potential* energy (internal energy), adding an equal amount of back-emf translational field energy in so doing. In short, the charging captures an equal amount of charging energy from its external environment — in this case, from the back emf of the trapped external electrons in the outside open wire. This is one way to capture and use the back emf energy normally reducing the energy available.

We point out that this is charging the capacitor by *Lorentz's symmetrical regauging,* and it represents a rotation of the frame of the capacitor itself out of the laboratory frame. As is well known, the trapped energy of a capacitor in a rotated frame is not the same as the energy of the same capacitor in the nonrotated lab frame.

The single-wire system destroys the symmetry-enforcing function of a closed-current-loop circuit conventionally used. In short, it does not return the "spent electrons" from the ground return line back through the source dipole of the outside power source, so it does not use half the collected energy in the circuit to destroy the source dipole and its extraction of EM energy from the vacuum (to destroy its broken symmetry). Because it retains the broken symmetry, Tesla's single wire system is an open system far from thermodynamic equilibrium with its active vacuum environment. It is thus permitted to exhibit those five magic functions of such open dissipative systems: (1) self-ordering, (2) self oscillation or self-rotation, (3) outputting more energy than the operator inputs, (4) self-powering of itself and its load, and (5) exhibiting negentropy.

4.4.3 Potential Energy Shuttling in a Circuit

The net result is to "shuttle" the (stress) potential energy back and forth in the circuit, between the ends. See Figure 4-7. When this stress potential energy shuttling is done at the natural (or tuned) resonant frequency of that single wire circuit as an oscillator, the monopolar charge transfer is optimized and amplified by the resonance effect, with an increase in the energy being shuttled. It is this type of Tesla single wire circuit that can in theory exhibit COP>1.0.

Tesla's single wire system also forces the change of energy between internal energy and external energy. A deeper analysis will also involve Dirac sea hole current effects as well as lifting of electrons from the Dirac Sea and combination of unobservable Dirac sea holes and electrons without radiation. However, we leave that for the advanced theorists. A taste of it is included in Chapter 9.

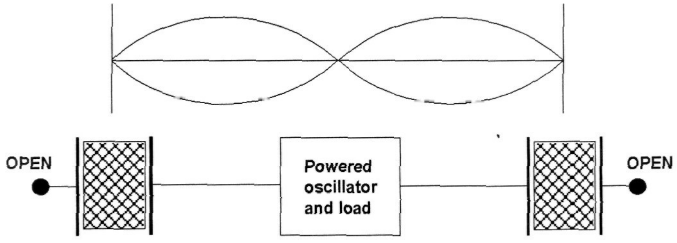

open capacitor to change field energy into stress potential energy

Note: System is designed with a certain resonant frequency At that frequency, the current can become mostly "squeeze" and "relax" as a standing wave emerges and stabilizes

open capacitor to change field energy into stress potential energy

Whether COP>1 0 or not depends on successful additive collection and use of the Heaviside nondiverged energy flow component due to the multipass and reflection operation

Figure 4-7 Tesla's resonant single wire circuit **operation.**

But is there any evidence for this "shuttling" of the monopolar potential in Tesla's circuits? Yes, there is, and indeed some of Tesla's patented circuits do exhibit such energy shuttling in them, as rigorously shown by Barrett {124, 286a, 286b}.

4.4.4 Other Considerations

For COP»1.0 systems and some COP>1.0 systems, the concepts of Chapter 9 must also be applied, since the system outputs not only positive energy but also negative energy. The negative energy — in the form of Dirac sea hole current — then sweeps back through the system from its output section to its input section. At various places, these "vacuum holes" eat electrons (as on atoms, causing ionization).[115] In short, the Dirac sea holes are converted to more normal lattice holes, and thus can be dealt with in the same fashion as semiconductor designers handle lattice hole problems inside their semiconductor designs. Positron (hole) effects are well known in materials science, and are increasingly being used in actual systems such as the antimatter (positron beam) microscope {266a-266b}.

We accent that a simple LC oscillating circuit does not produce COP>1.0 of itself. If, however, the charges in the currents surging to and fro are in particle resonance at the "feed frequency", then they will "sweep out" a greater geometrical cross sectional area to intercept additional Heaviside energy flow existing outside the normal "static charge" Poynting energy flow. This is similar to the "intercepting resonant particles " process utilized by Bohren {257}, and it allows output of more EM energy than the operator inputs *in the calculated Poynting component* of his own input energy flow. It *does not* output more EM energy than the operator inputs in the sum of his Poynting input component and his Heaviside input component.

The real energy conservation law is that the input energy from all sources and of all types must equal the output energy of all types, where by

---

[115] The union of a Dirac hole (4-space negative energy electron before observation) and **a** positive energy electron before observation, can occur without radiation. The two curvatures of spacetime involved are equal and opposite, so they just get **replaced** by a flat spacetime, which involves a filled Dirac sea (all the holes filled with electrons). In short, the energy can go into changing the stationary local double re**curvature** of local spacetime, by a sort of "superposition of two spacetime curvatures", to give a net flat spacetime, rather than dynamically radiating EM energy away at the speed of light. However, if either of the two interacting particles has been observed, this is not the case and standard pair annihilation will be observed. The point is that particles in their unobserved *mt* state rather than their observed *m* state, can interact without observation, similar to the way that wave-to-wave interaction occurs in nonlinear optics.

"output" we mean output as useful energy and output as loss or non-useful energy.

So conservation of energy flow is not violated, because one also inputs an additional *Heaviside nondiverged* component, unknown to Bohren. Circa the 1880s or early 1890s, Lorentz arbitrarily discarded it from accountability. One does not have to "pay" for that *nondiverged* and *unaccounted* input energy flow component. Except in continually applying Lorentz's integration trick to discard it, the nondiverged Heaviside component has been almost totally neglected in the textbooks since Lorentz arbitrarily discarded it back there circa the late 1880s. However, it is roughly dealt with in near field antenna theory, sometimes under the topic of "Poynting energy bunching". In the near field, e.g., there also may be a substantial *longitudinal* E-field component. One region may be oscillating back and forth along the flow direction, while the region in quadrature may be flowing in only one direction.

By causing the intercepting charges themselves to resonate and sweep out a greater reaction cross section in an energy flow, the normal definitions of "field intensity" and "potential intensity" are no longer true, because their reaction intensity to a unit point *static* charge has been used as their stated "magnitude". The conventional definition is only for an indication of the *intensity* of the energy flows comprising the field and potential. With the unit point *resonant* intercepting charge rather than the assumed *static* charge, the reaction cross section for the same charge increases, and so the "apparent magnitude of the (intercepted) field intensity" also increases beyond what is in the textbook. So far as I am aware, there is no previous precise term associated with this phenomenon, except to lump it in the euphemism of "negative absorption of the medium" or "negative resonance absorption of the medium" in certain materials {267}. It is more nearly a Poynting energy amplification process, since the "intercepted" or Poynting component of the energy flow is increased. Thus it can be a mechanism for COP>1.0, if the closed current loop circuit's use of half the collected energy in the circuit to kill the source dipole is also defeated.

Nonetheless, energy bunching or Poynting component amplification by altering the reaction cross section function of the intercepting particle is a legitimate way to provide excess "free" energy input for COP>1.0 EM power systems.

4.5 Perspective on Tesla vs. Marconi

Tesla was neither deluded nor mistaken, though some electrodynamicists have mistakenly considered him in that light. On the other hand, Jackson — one of the leading electrodynamicists of our time — pays Tesla a significant compliment as follows {268}.

> *"In U.S. patent no. 787,412 (April 18, 1905), reprinted in Nikola Tesla, Lectures and Patents and Articles. Nikola Tesla Museum, Beograd, Yugoslavia (1956), this remarkable genius clearly outlines the idea of the earth as a resonating circuit (he did not know of the ionosphere), estimates the lowest resonant frequency as 6 Hz (close to the 6.6 Hz for a perfectly conducting sphere), and describes generation and detection of these low-frequency waves. I thank V. L. Fitch for this fascinating piece of history. "*

We also point out that many present science texts err by continuing to credit Marconi with the *discovery* of radio, although he certainly did much to pioneer it and received a Nobel Prize in 1909. Hertz must be given credit for the first deliberate signal transmission per se, during the period 1886-1888. As his transmitting antenna, Hertz used an end-loaded dipole with a gap, and as his receiver he used a resonant square loop antenna with a gap {269}. Sparks produced at the gap in his transmission antenna resulted in sparks appearing at the gap in his receiving antenna.

Marconi became obsessed with Hertz's results nearly a decade later, and in mid-December 1901 Marconi announced the reception of signals at St. John's, Newfound that were transmitted across the Atlantic from Cornwall, England. He later recorded radio signals transmitted around the Earth, and began regular trans-Atlantic message service in 1903.

Marconi's patent {270}, however, was eventually overturned by U.S. Supreme Court decision {271}, because Tesla's patents (and demonstrations) substantially preceded it. Tesla also demonstrated radio transmission before Marconi, and Marconi actually used a modified Tesla circuit in his famed transmission in 1901. Nonetheless, this is not to detract from Marconi's accomplishments; he definitely put "wireless communications" on the map. It is simply to give *discoverer of radio* credit where credit is due.

Tesla was also closer to the truth of the form of an EM wave in space than are modern electrical engineering and most electrodynamics texts. In the present book we have stated that what exists in 3-space as the so-called "transverse EM wave" are actually longitudinal EM waves and functions of them, with the energy entering each 3-space point from the accompanying time-polarized (conjugate) EM wave at that point (i.e., from the time domain) and emerging in 3-space in the form of a longitudinal EM wave, in 1:1 correspondence.[116] This is followed by a return of the energy from 3-space back to the time domain. This "circulation" from the time domain to 3-space and back to the time domain occurs at every point dipole in 3-space. Only the presence and combination of both the time-polarized photon and the longitudinal photon at each point in an EM wave makes the wave "observable", when we translate into wave phenomena the photon phenomena pointed out by Mandl and Shaw {19}. What spreads as EM energy flow in space (from a source charge) is the local polarization of the vacuum, associated with this giant negentropy circulation from time to the negative charge, to the positive charge, and back to time.

The notion of the transverse EM wave flowing through space is an artifact of the ubiquitous *substitution of the effect for the cause* in Maxwellian electrodynamics. E.g., a completely erroneous "definition" of the causal field (as it exists in space prior to interaction with charged mass) is given as the *effect of its interaction with a ubiquitously assumed unit point charge at every point in space where the field is present.* This in fact substitutes the effect (existing *after* the interaction) for the cause existing *prior to* the interaction. That of course is a gross non sequitur.

We give a concrete example from the first edition of a standard dictionary of physics and mathematics {272}:

> *"**electric field...** 1. one of the fundamental fields in nature, causing a charged body to be attracted to or repelled by other charged bodies; associated with an electromagnetic wave or a changing magnetic field. 2. Specifically, the electric force per unit test charge."*

---

[116] Later we shall deal a little further with how EM energy iteratively recurs statically in 3-space and propagates dynamically and continuously in 4-space, in more detail. The solution to the long-vexing particle-wave duality is the process of observation. Before observation, an entity is dynamic and continuous in spacetime, not 3-space. After observation (by emission of a photon), it is a momentary frozen 3-snapshot (particle) at a single instant of time. At the next immediate instant, a photon is absorbed, reconverting the "static snapshot" into a dynamic 4-entity again. Since observation (photon emission) and de-observation (photon absorption) iterates at extremely high rate, the entity and each tiniest part of it oscillate back and forth between being a frozen 3-entity (particle) and a dynamic 4-entity (wave).

In the first definition, the field is given as the "cause" of the interaction that exists between like and unlike charges. In the second definition, the field is given as the effect — of the interaction with a test charge — of *something undefined* as the cause.

An entirely different "definition" is given by Gray and Isaacs {273}:

> *"**electric field** ... The space surrounding an electric charge within which it is capable of exerting a perceptible force on another electric charge."*

Here it is defined as simply the space around an electric charge, and not really the cause or effect of anything, unless that local spacetime has been curved (which it has).

Not to belabor the point, this sort of thing led Feynman to "define" the EM field in this fashion {274}:

> *"...the existence of the positive charge, in some sense, distorts, or creates **a "condition"** in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a force is called an electric field. "*

So Feynman made the electric field a *condition in space* surrounding a charge, where the condition can produce (has a potentiality for producing) a force on another charge. He also was aware that force is an effect of the reaction of the causal field with charged mass, in the following statement {275}:

> *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present... One of the most important characteristics of force is that it has a material origin..."*

In Maxwell's theory, the electron, atom, and nucleus had not yet been discovered, and electricity was considered a thin material fluid, flowing in a wire (hence the name "currents") much like water through a pipe. The surrounding space was considered filled with a thin material fluid called the luminiferous ether. A very rough notion existed that the "shaking" of the ether disturbance coming in to an antenna would disturb or perturb the electric fluid in the wire directly. Hence, if one "measured" a transverse oscillation of the electric fluid in the wire due to this "detection", it conclusively "proved" that the incoming ether-perturbations were also transverse. On that notion — together with Faraday's notion that "field lines" were taut physical strings, so that "field perturbations" were naught but plucked oscillations of these taut strings — the entire notion of the transverse EM force field wave in space was created.

It really is not that way, as is easily shown with modern re-examination of what happens in the receiving wire, considering the Drude electron gas {276} and electron drift velocity. Today we know that the electrons move *longitudinally* down the wire only with a small drift velocity (usually a few inches per hour), while the signal (field) races longitudinally down the wire at essentially light speed or nearly so.

The electrons are longitudinally constrained in the wire because of the vast number of electrons ahead of them that tend to repel their longitudinal movement. The electrons are far less constrained laterally in the wire. However, each electron is also spinning, and its 3-space performance may be compared to a spinning gyro, whose spin axes are now longitudinal restrained. Accordingly, the electrons readily *precess laterally* in the wire[117] when perturbed *longitudinally* by the incoming longitudinal 3-space components of the "wave in vacuum" — by gyroscope precession theory. The electrons at the surface layer of the wire are somewhat less constrained longitudinally, and they "slip" a little down the wire in each oscillation, accounting for the very slow electron longitudinal drift velocity.

Accordingly, one of the most important dynamics of the circuit is the dynamics of the surface charges, as pointed out by electrodynamicists {277}.

If the incoming waves were true lateral waves, there would be no longitudinal force component on the laterally precessing electrons, and coherent current would not flow longitudinally in the wire at all. Instead, one would just get some longitudinal to and fro noise in the coherent lateral motions of the electrons, due to electron collisions.

---

[117] It is this lateral precession movement of the electrons that is "detected" in most of our instruments. To the early electrodynamicists prior to the discovery of the electron, this detected lateral motion was considered the "shaking of the material electric fluid" in the detecting wire. It was thought that the incoming "electric fluid" from space also had to be material (in the material ether) and transverse, because otherwise the oscillations from space could not have mechanically transferred their "lateral oscillation" to the electric fluid in the wire. Faraday's "material lines of force" concept and the notion that EM disturbances (waves) were simply the mechanical pluckings of these taut strings which also involved similar "material ether" concepts.

Nonetheless, it seems we are stuck with the ubiquitous transverse EM wave model. For more than 100 years, scientists have universally substituted effect for cause in the case of much of electrodynamics — as in the continuing substitution of what is *diverged from* the field and potential at each point in them, by an assumed static point charge, as being the magnitude of the field or potential itself at that point. At best, what is diverged from the field or potential is an indication of the point intensity of the energy flows comprising the field or potential at each point, and in the case of the field, of the prevailing direction.

So electrodynamics is still riddled with that nearly all-pervasive error, and it seems it will likely never be corrected. The leaders of the scientific community do not seem to wish it done and will not allow it to be funded. In short, the community is so comfortable with that grave error that it will almost defend it to the death. The so-called transverse EM wave in 3-space has come to be accepted as if it were a law of nature.

Occasionally a journal will point out the terrible non sequitur in the prevailing notion of the EM wave in space as a "plane wave of oscillating orthogonal E and H fields along the x- and y- axes moving at light speed in the z- axis direction." Romer, former editor of *American Journal of Physics,* provides a prime example {278}. Romer takes to task:

> *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate." "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

At any rate, once we understand how the real EM wave exists in space, then we can sympathize with Tesla's view of it from the following {279}:

> *"Tesla upholds the startling theory formulated by him long ago, that the radio transmitters as now used, do not emit Hertz waves, as commonly believed, but waves of sound... He says that a Hertz wave can only be possible in a solid ether, but he has demonstrated already in 1897 that the ether is a gas, which can only transmit waves of sound; that is such as are propagated by alternate compressions and rarefactions of the medium in which transverse waves are absolutely impossible."*

In Tesla's own words {280}:

> *"...I showed that the universal medium is a gaseous body in which only longitudinal pulses can be propagated, involving alternating compressions and expansions similar to those produced by sound waves in the air. Thus, a wireless transmitter does not emit Hertz waves which are a myth, but sound waves in the ether, behaving in every respect like those in the air, except that, owing to the great elastic force and extremely small density of the medium, their speed is that of light."*

From quantum field theory, the instantaneous scalar potential in space is actually a coupling of a scalar photon and a longitudinal photon {19}. The only EM component in 3-space is that longitudinal photon.[118] Transforming to wave language, Tesla appears to have been correct in his view of the longitudinal form of the EM wave in 3-space, whereas all the texts are still wrong even today.[119] However, Tesla was also noted for grandiose pronouncements, a trait which certainly did not endear him to the struggling electrodynamicists of the time, even when he was correct and they were wrong. An example is the following {281}:

> *"The Hertz wave theory of wireless transmission may be kept up for a while, but I do not hesitate to say that in a short time it will be recognized as one of the most remarkable and inexplicable aberrations of the scientific mind which has ever been recorded in history."*

Unfortunately, in electrodynamics we are still stuck with that "most remarkable and inexplicable aberration" today, and we will remain stuck with it until the leaders of the scientific community (such as the National Academy of Sciences and National Science Foundation) recognize it and fund a complete overhaul of the foundations of classical electrodynamics, which is so sorely needed. On the other hand, *one should not hold one's breath while waiting for the mountain to come to Muhammad! One might as well ask the stars to fall from the sky as to ask the leaders of the scientific community to correct the serious foundations errors in physics and especially in electrodynamics.*

---

[118] Hence it is not surprising that all EM in 3-space decomposes into bundles of longitudinal EM waves and their dynamics, as shown by Whittaker {85, 91b}.

[119] However, Tesla did not appear to realize the involvement of time-energy transduction to longitudinal KM 3-space wave energy and vice versa.

4.6 Conclusion

Hopefully we have set the stage for our next chapter, where we will be advancing mechanisms and approaches that have been tried by a variety of researchers. Several of these researchers have indeed produced COP>1.0 devices and processes. However, at this writing none of the devices produced by these inventors and researchers are in production and on the market. As we mentioned in Chapters 1 and 2, there are very powerful forces that have been applied to prevent such development.

Briefly, such suppression has resulted due to one or more of the following:

(a) Sheer greed by the group that gathers around the inventor/researcher, including multiple competing investors and/or "backers". Often a struggle for control develops, and the project is decimated in the process. In many cases a legal struggle also develops, with the result that the legal status of the invention may be fouled beyond repair.

(b) Naivete on the part of the inventor and his surrounding group. The inventor needs the frequent advice of a patent attorney, a regular corporate attorney, a person familiar with research and development projects and procedures, a person expert in testing (particularly with sophisticated multichannel digital oscilloscopes which gather multiple channels of data simultaneously and integrate under the curve, an accountant, and a person skilled in negotiation with potential investors and backers. Most inventors do not have access to all that, or only have access to it after much of the die has been cast and serious mistakes have been made.

(c) Animosity and fierce attack from the scientific community and self-appointed "defenders of the faith". Few inventors are prepared for the onslaught that awaits them if they succeed in producing a genuine COP>1.0 electrical system. Slander and libel are just part of it. An inventor or researcher caught up in always trying to defend himself against vicious ad hominem attack is an inventor successfully deviated from his work. The family also often suffers.

(d) Just now, after the Enron fiasco much of America is becoming aware that powerful organizations often pull out all the stops in a blind pursuit for money and power. Of course they have done so for a very long time; this is nothing new. The inexperienced inventor or researcher is ill prepared for the fact that portions of government organizations,[120] national labs, universities, etc. may be precisely that way also! U.S. government laboratories also file patents; check it out! Check out Fullerton's defense of his patent rights for his ultrawideband communication system work.

The people in any large organization form a bell-shaped distribution curve. About 9% of that curve represents folks who have — to put it mildly — little ethics when it comes to position, power, and money. Some of these 9% fellows will go to quite some length to twist the patent or part of it or control of it away from an inventor. An inventor who has never heard of "march-in" rights (a neat little "theft" phrase favored by DARPA, for example) is often an easy mark. Government agencies will be on an interesting invention and its hapless naive inventor like vultures on a fat calf. They dangle some money and a contract to "assist the inventor to get his invention developed". There is a little clause in the fine print of the contract where the government reserves "march-in" rights. This is a hidden bombshell that really says the government — *anytime it wishes* — can just unilaterally declare that the funded inventor isn't getting the invention to market fast enough, so a single bureaucrat just *seizes it* — in the name of the government, of course. Anyone who has been in serious aerospace contracting very long is also aware of the deals between a few government fellows who control award of the funds, and their "favored contractor". If enough business is steered to the contractor by the government person, it is implicitly understood that when the government fellow retires, he goes to work for that company with a Vice President position, lucrative salary, and a fat stock option. It's easy to become an instant millionaire or multimillionaire that way. A small percentage of government contracts are actually contaminated in such a hidden fashion. The universities usually are much more open and blatant, and just put in a clause that says, "All patents revert to the university!" Also, the inventor must be aware of the difference between "nondisclosure" and "non-circumvention". Just try getting the latter clause from a university, or either clause from a government agency (yes there are rules for "proprietary information", and they are widely violated). Non-circumvention means the outfit agrees not to change a coil or something, and go and patent your invention. Nondisclosure means they just agree not to talk or write about it openly.

---

[120] Interestingly, the government is not allowed to copyright its own publications by federal servants. Neither should the government be allowed to patent its own inventions by federal servants and federal agencies. This has set up direct conflict in the federal funding channels for the taxpayers' funds intended for independent research, providing a wonderful opportunity for hanky panky by that lower 9% of the people comprising any large group. By the bell-shaped distribution curve, those comprising that 9% will rise to the opportunity. They are manipulative, greedy, and some are immoral and self-serving. Partially coupled with a similar low "bad apple percentage" in the contracting game, this enables many travesties in the name of federal research and development with respect to patents and misdirection of funds.

(e) And yes, in addition to these many "petty" little conspiracies, there really is a *great* conspiracy — in fact, there are many of them. Churchill just referred to the entire set of conspiring cartels as the "High Cabal". This is a loose-knit set of extraordinarily wealthy cartels that between them control much of the financial power of the world. You can only find the principals behind a couple of hundred interlocking corporations, if you penetrate deeply enough. Much of their great financial income comes from the control of energy at its various levels, through a net of interlocking corporations. Something like $2 trillion dollars per year is their "take" from energy worldwide. Oil plays a great role in that. These "high cabal" fellows make or break nations financially, as is almost being done to Argentina as this is being written. They also have been stopping COP>1.0 electrical power systems since shortly after the turn of the century, using whatever means are necessary — fair or foul. That is why we still are burning oil, coal, and natural gas, and why our automobiles and trucks run off gasoline and diesel, and not off free electrical energy from the vacuum. The high cabal intends to keep it that way. Some of the major primary means of "stopping" an invention have been as follows:

a. Manipulation and improper use of the law (Gray's production line was shut down by law officers before his engines could roll off there and onto the market).

b. Involving the unsuspecting inventor in drug charges by planting drugs on his property surreptitiously, then secretly notifying the drug enforcement officials who suddenly raid the premises and discover the narcotics, resulting in the inventor going to jail since he has been caught "red-handed".

c. Threatening the inventor's life and the lives of his family, to frighten him away from what he is doing.

d. Arranging a fatal accident for the inventor, such as a fatal auto accident, a fall from a high place, or drowning.

e. Assassination (hit men are fairly simple to hire; a good one is a bit more difficult to hire but is doable. A good hit man might assume a disguise and fake a robbery, shooting and killing the target in the "robbery". In an extreme case, he just fires at a distance with a professional sniper rifle that is silenced. A variation is the use of a "throwaway" assassin, a fellow who has been deeply conditioned by very harsh methods, to deeply involve his autonomic nervous system. He can be spotted by the "glaring eyes" effect of autonomic nervous system involvement. This fellow is only marginally functional in real life, and requires a handler. He also must "fixate" for some minutes (usually from 20 to 30) on the target, by staring at him for a period. Then — say, the next day — while the target is at lunch in a public restaurant, the throwaway walks in, walks straight to the target, and empties a pistol in him at point blank range. He remains there until the police arrive and arrest him. He is immune even to torture; nothing can be gotten out of him, even by electrical shocks on his genitals. There are two levels of throwaways; that is the older and less functional version produced by very harsh means. The more modern version is without the glaring eyes with white showing above the iris; this fellow is more functional, fixates very quickly (couple minutes to five minutes), and can operate without the close immediate presence of the handler.

f. Making the inventor an offer he "cannot refuse", in the old Mafia sense. In the latter, nicely dressed persons come to the inventor's house, meet with him, and offer him a large sum of money to quit what he is doing for the rest of his life. The going sum varies between $5 million and $10 million, usually. The offer is real and it cannot be refused; either the inventor accepts it, or he and all his family are dead when these gentlemen leave the premises. If the inventor refuses anyway, he and his family are killed.

Usually the bodies also just "disappear" and are never found or heard from again.

g. These days, variations are used such as a careful crash of another auto into the back of the inventor's car so he is really shaken up badly but not seriously hurt. He goes to the hospital for checkup, in case of concussion etc. Everything is seemingly okay, but he is kept overnight for observation. That night he is given a hypodermic of air in his veins. Next morning his family receives a call that the inventor has suffered seizures from an apparent concussion after all, and is fighting for his life. If he doesn't die, he becomes a human vegetable. Piece of cake.

h. Then there's the ice dart dipped in curare. The curare paralyzes the muscles of the heart and lungs; the inventor goes down in convulsive floundering, gasping and unable to breathe or to pump blood. He expires very quickly. The curare breaks down in the body, the ice melts, and you have a fellow the medics find died of convulsive heart seizure, etc. Works like a charm. This is an adapted "dirty spooks" trick, adapted from those branches of the intelligence communities of the world that specialize in assassinations.

i. Use of "gaming", which was originally developed to influence or change the behavior of foreign leaders. Imagine writing a scenario, as for a movie. Only you have the deep psychological profile of the targeted individual. You also have the deep psychological profiles of quite a few other individuals who have "knee-jerk" reactions in a certain direction useful to you in your games (in your scenarios). These are called "agents of influence" since, to get them into interacting with the target, only a phone call or other contact need be made and a stimulus given, and they are off and running, self-powered and going after that target with their knee-jerk response. So very sophisticated computers are used, the scenario (intended plot of events to happen) is prepared (the artificial intelligence programs will even give the probabilities of success of all the variants), and a recommended scenario variant is derived and accepted. Then the controller has one or more agents working **for** him that contact the players (the agents of influence). They get their phone calls or stimuli, react in predictable fashion, and come pell-mell at the inventor, interacting all over the board with him. The range of games is enormous, from stock schemes to lawsuits to infiltrating agents of influence into the internal organization of the inventor to money laundering, etc. The controller follows the progress, and only once in awhile makes slight adjustments (may need to trigger other agents of influence, send in one actual agent, etc.). Over the last 20 years, the gaming has become very expert. So many games can be launched against a single inventor that he usually is finished or nullified by one of them, eventually. Having been gamed over 200 times in the last two decades, I plan to write a book one day on the subject of the games and what to watch for.

j. Finally, here is a sad and disconcerting rule, usually true. If you gather six persons together that you have known for years or even all your life, and they are the nucleus of the organization around you, watch what happens when success occurs and you really do pull EM energy from the vacuum and power loads, lights, etc. One-sixth of the population, seeing that, will shortly contract a very strange malady known as "gold fever". It is not just an *obsession;* instead it becomes an all-consuming *possession* almost in the biblical sense. An individual possessed by gold fever would kill his own grandmother to try to gain control of that system. If one or more of these "gold fever" transformations occur, that sets the group into profound conflict and vicious infighting, on every front. You get sabotage of equipment, sabotage of demonstrations, stealing of equipment, legal shenanigans, just name it. This has destroyed many groups. It will destroy many more.

k. A variation is to send in a special agent who penetrates the group and becomes a member. Indeed, the high cabal has certain persons who specialize in just such penetration and are specialists at setting up disruption. They have had plenty of practice, and are very good at their trade. The identities of some of these "special agents" would surprise the casual observer. The so-called "overunity community" is rather strongly penetrated and misdirected by a number of such special agents, in addition to a greater number of agents of influence.

1. Finally, the technical knowledge of most inventors and researchers in this field has been largely limited to ordinary electrical engineering and electronics. This means that the inventor often forms his own lexicon and way of thinking about how he does the "magic" COP>1.0. It may be difficult or impossible to connect the way he thinks back to legitimate physics. Further, most of the COP>1.0 community, the majority of electrical engineers, and nearly half the physicists do not understand the technical difference between the *efficiency* of a power system and its *coefficient of performance.* We have insured that the reader of this book will know the precise difference — it's even in our glossary! Further, the entire group around the inventor, and the inventor himself, may not have a clear understanding of the source of the extra energy, the mechanism by which it is obtained, or even the difference between a system in equilibrium with its active vacuum environment and one in disequilibrium with it. Most engineers and a great number of scientists also are still unaware of the dramatic difference between the thermodynamics of equilibrium systems and the thermodynamics of disequilibrium systems. Many have no notion of what the active vacuum is or does. So the inventor may actually be able to build the device himself, but may later err and actually build it away from COP>1.0 with further development.

m. Another problem can be measurement. Sometimes very simple errors — such as "measuring" highly nonlinear, spiky signals with an RMS meter, or using an inferior oscilloscope, or using inferior probes with good oscilloscopes, tricks a relatively inexperienced inventor himself into believing his unit achieves COP>1.0 when it doesn't. Also, some standard engineers will insist on calorimetry, which is inappropriate for COP>1.0 EM systems since they output a mix of both positive and negative energy. The positive energy will heat the calorimeter fluid and the negative energy will cool it simultaneously, and the "expert measurer" will always measure the difference and thus COP<1.0 with his calorimeter (unless the inventor is a rare one who has also transduced the negative energy fraction of his output into positive energy). Yet the conventional engineer or scientist — who has never seen negative EM energy or a COP>1.0 EM system in his life — will be so confident from his "positive energy only" thinking that he will be completely convinced that a real COP> 1.0 system, measured with a calorimeter to show net positive energy COP<1.0, is *actually* a COP<1.0 system instead of the error in measurement that it is. One of the purposes of this book is to try to explain such things and get a better procedure and protocol made available for the testing.

n. Finally, there are a few actual charlatans who deliberately deceive both the group around them (if it's sufficiently naive) and naive investors and backers. E.g., if one builds a huge starter motor (very high current, very high torque, cannot run long without burning up!) and puts it in an automobile, then fills the back of the automobile full of batteries, that beast will draw 400 amps and leap across the pavement or warehouse floor with an agility one is shocked to see. Then if the inventor is always "having difficulty with his controller unit", and "everything will be fine once the controller problem is licked" or some such, many small and naive investors will leap to invest their life savings eagerly. There are many other scams also, and unfortunately every experimental field, with great financial promise if successful, does attract them. Here the old adage is best: If it is just too good to be true, it probably isn't. One should insist on a technical explanation of the source of the excess energy, the mechanism by which it is taken from the environment, how asymmetrical self-regauging of the circuit occurs, etc. If these questions cannot be answered, the inventor himself does not know what he is doing. At best, he cannot explain it; or he may just be honestly mistaken. In the worst case, the inventor knows it is not genuine, but is deliberately deceiving the investors. So my advice would be to first assess the scientific level of the understanding of the inventor group. Then get an independent test of the device or see one done properly. Then decide whether what you have seen and heard is real or not.

o. Meanwhile, there are from time to time legitimate COP>1.0 systems initially developed. The place where all of them have failed is where the inventor group seeks to obtain the appreciable financing (something like 30 million dollars is required these days) to complete the phenomenology experiments, modeling, scale-up factor determination, and prototyping before a production model can be built. To date, every one of the legitimate COP>1.0 systems has foundered and failed right at that sheer vertical cliff. Our own MEG system is at that cliff at the moment, so that cliff is very near and dear to the hearts of the five of us who have struggled so long and invented the MEG. Time will tell whether we succeed where all the others have failed, or whether we also just get defeated by the cliff. As of this writing, we have not yet obtained the necessary capitalization. But we are giving it our very best try, and we believe we will win it eventually.

Chapter 5

Selected Approaches to Overunity Power Systems

> [On point-contact transistors]. *"First, the theory underlying their function is imperfectly understood even after almost a century..., and second, they involve active metal-semiconductor contacts of a highly specialized nature. ...The manufacturing process is deceptively simple, but since much of it involves the empirical know-how of the fabricator, the true variables are almost impossible to isolate or study. ... although the very nature of these units limits them to small power capabilities, the concept of small-signal behavior, in the sense of the term when applied to junction devices, is meaningless, since there is no region of operation wherein equilibrium or theoretical performance is observed. Point-contact devices may therefore be described as sharply nonlinear under all operating conditions. "* [William B. Burford III and H. Grey Verner] {282}.

5.1 Introduction

In this chapter we present a selection of potential overunity notions, inventions, and experiments by various inventors and researchers. Its purpose is to show some of the more relevant work that has been done, and hopefully to stimulate new thinking by the interested reader. It is not intended to be all-inclusive, but merely representative. Most of the work addressed is by other inventors and scientists, and we give an appropriate commentary from our viewpoint as best we understand the invention and the process used. Not all the concepts presented can be clearly explained, though some can and are.

This is not intended to be a "kit of parts" or an assembly instructions type of presentation. For that, the interested researcher will have to rely on his or her own ingenuity and lots of bench experiments, because no such "kit of parts with instructions" for COP>1.0 electrical power systems exists at present {283,284}.

5.2 Tesla's Shuttling of Potential Energy and Barrett's Extension

See Figures 4-4 and 4-5 in Chapter 4. In his patented circuits, Tesla accomplished the novel function of "shuttling" the potential itself (and therefore potential energy) around in the circuit at will. This startling function accomplished by Tesla is impossible to see in a conventional vector or tensor analysis of his patented circuits, but it is clearly evident in a higher symmetry EM analysis of them. So it does not appear in the conventional texts, which use vectors or tensors for their circuit analyses. Barrett, one of the leading higher group symmetry electrodynamicists and also one of the pioneers of ultrawideband radar, utilized quaternion electrodynamic analysis to clearly show this novel effect in Tesla's actual circuits {285}. Thus, Tesla was able to produce and move around asymmetrical self-regauging in his circuits, wherever he wished. That has been totally missed by scientists prior to Barrett's revealing analysis.

Barrett then went on to extend Tesla's basic mechanism and employ it for communication and radar purposes, obtaining two patents {286a-b}.

It would appear that use of such shuttling techniques is probably applicable to reducing the back emf in conventional nonshuttling circuits, leading to new COP>1.0 mechanisms. The definitive answer remains to be determined by some future researcher.

5.3 Moray's Radiant Energy Device

One of the remarkable examples of a successful COP>1.0 system was the radiant energy system of Dr. T. Henry Moray of Salt Lake City, prior to World War II. Figure 5-1- shows Dr. Moray and his final 50 kW unit, powering an assembly of light bulbs {287}. The device weighed 55 pounds, used an antenna connected as a "signal input" from the ether, but needed no input of electrical energy by the operator in order to achieve its steady power output.

As shown in Figure 5-2 and Figure 5-3, the various Moray tubes were complex in structure. In addition to various grids, they usually contained (in quadrature) three or more amorphous semiconductor pellets pressed in large presses and sintered with locked-in stresses. One points out the presence of optical effects including phase conjugation, possible use of the built-in stress as a sort of self-pumping and mild self-amplification, possible iterative feedforward and feedback loops due to iterative phase conjugation, collecting of additional energy from the usually nondiverged

Heaviside component via the iterative reverberation of energy, and non-interfering fields in quadrature.

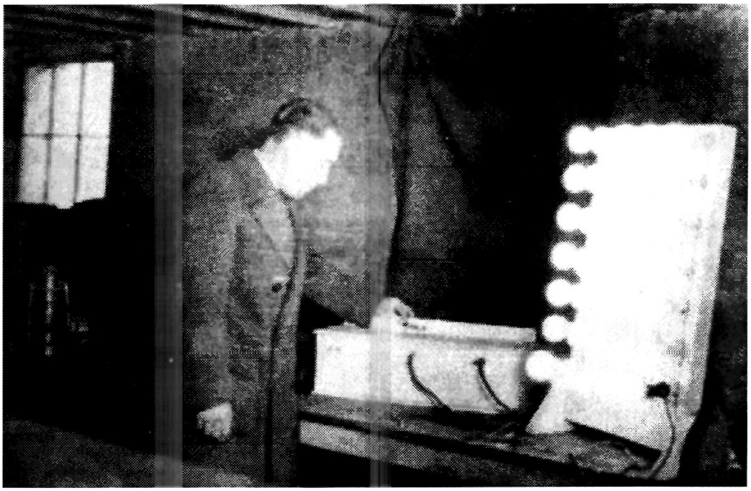

Figure 5-1 T.Henry Moray's 50-kW unit.

The blown quartz tube envelope, in the dark of an enclosure, is open to the passage of infrared (heat) and ultraviolet, which spans a harmonic interval. This allows the passage *of difference* frequencies as sine waves without breakup, even in a nonlinear isotropic environment. Moray stated unequivocally that the tubes would not work unless the envelopes were made of blown quartz. The pellets (shielded from outside light) would nonetheless receive IR input from the ambient temperature, and probably triggered at least weak UV due to the radioactive decay continuously ongoing in the semiconductors. We point out that the UV band is roughly twice the frequency of the IR band, so the "difference frequency" is roughly the IR band itself. In short, it may be that "difference frequency" IR is thus processed in Moray tubes in much more sophisticated fashion than has previously been suspected, including in a self-amplifying (regenerative) manner.

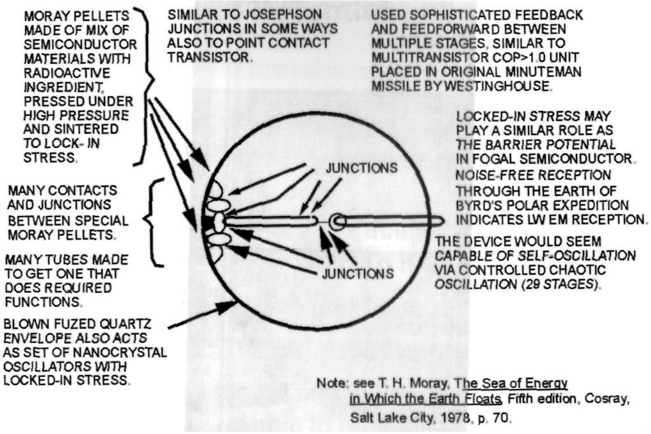
MORAY PELLETS MADE OF MIX OF SEMICONDUCTOR MATERIALS WITH RADIOACTIVE INGREDIENT, PRESSED UNDER HIGH PRESSURE AND SINTERED TO LOCK- IN STRESS.
SIMILAR TO JOSEPHSON JUNCTIONS IN SOME WAYS ALSO TO POINT CONTACT TRANSISTOR.
USED SOPHISTICATED FEEDBACK AND FEEDFORWARD BETWEEN MULTIPLE STAGES, SIMILAR TO MULTITRANSISTOR COP>1.0 UNIT PLACED IN ORIGINAL MINUTEMAN MISSILE BY WESTINGHOUSE.
LOCKED-IN STRESS MAY PLAY A SIMILAR ROLE AS THE BARRIER POTENTIAL IN FOGAL SEMICONDUCTOR.
JUNCTIONS
NOISE-FREE RECEPTION THROUGH THE EARTH OF BYRD'S POLAR EXPEDITION INDICATES LW EM RECEPTION.
MANY CONTACTS AND JUNCTIONS BETWEEN SPECIAL MORAY PELLETS.
THE DEVICE WOULD SEEM CAPABLE OF SELF-OSCILLATION VIA CONTROLLED CHAOTIC OSCILLATION (29 STAGES).
MANY TUBES MADE TO GET ONE THAT DOES REQUIRED FUNCTIONS.
JUNCTIONS
BLOWN FUZED QUARTZ ENVELOPE ALSO ACTS AS SET OF NANOCRYSTAL OSCILLATORS WITH LOCKED-IN STRESS.

Note: see T. H. Moray, The Sea of Energy in Which the Earth Floats, Fifth edition, Cosray, Salt Lake City, 1978, p. 70.

Figure 5-2 Diagram and some novel features of T. Henry Moray's special tube.

The pressed semiconductor pellets were placed in the tube in quadrature, in a sort of "point or junction contact" with each other.[121] Each semiconductor contained a mix of triboluminescent zinc (a semiconductor material), a radioactive material (probably uranium dioxide), and another material of unknown composition, rumored — but not confirmed — to possibly have been diatomaceous earth which also has some unusual properties one may look up. There were some 29 amplifying tube stages in the device. A U.S. patent {288} was eventually granted on an electronic therapeutic unit containing Moray tubes, but the patent on the power system was never granted, although the application was kept current for some years by the Moray family. Eventually the power device patent application files disappeared from the U.S. Patent and Trademark Office.

Although its full technical details were never released, the Moray device is of interest because Moray demonstrated it many times to engineers and scientists, under rigorous measurement conditions, for years. These expert witnesses then signed affidavits to the results achieved by Moray in the tests. Moray also discovered that he could produce an additional 50 kW of power simply by inserting an additional tap just prior to the formal output.

---

[121] Indeed, the junctions of such sintered pellets are most interesting. Such a grainy junction can be compared to a host of very close point contact junctions, particularly with finely-ground powdered materials used in the pressing and sintering. In essence, Moray's junctions contained a great many "point contact transistors", and it is known that such transistors can behave (particularly under pressure!) as true negative resistors. This also appears to be one of Moray's secrets.

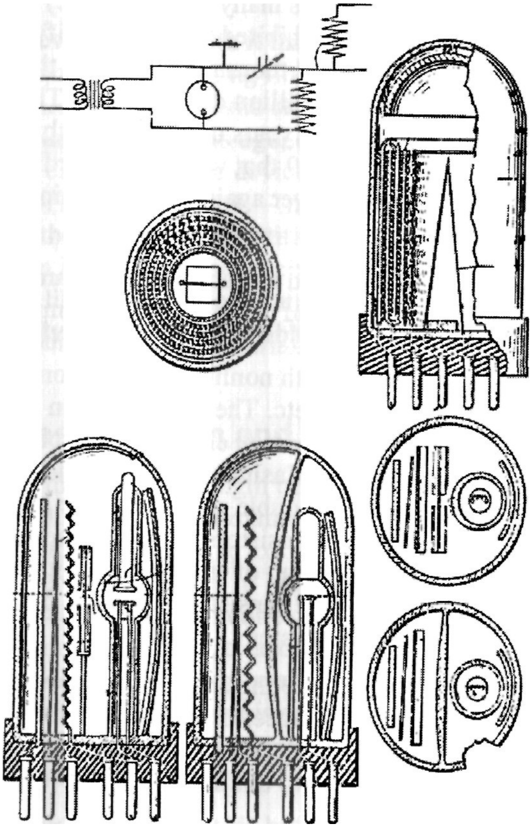

Figure 5-3 A variety of Moray tubes showing their complexity.

Repeated assassination attempts were made against Dr. Moray. While driving his car on the streets of the city, on several occasions he was fired at from an alley or from the sidewalk, but he had bulletproofed the car because of continual threats and ambushes. He was also shot in his own lab in a physical struggle with one would-be assassin, but survived although seriously wounded. A Russian company tried to steal the device, and a Russian agent was planted in Moray's lab, gradually sending essentially all the technical details of the Moray device to Russia. That agent later destroyed the unit right in Moray's lab by suddenly smashing it with a hammer.

Moray's tubes were quite expensive and were painstakingly made by hand. He fabricated a number of tubes (as many as 30 or more) before he obtained one that "worked" and exhibited the self-amplifying effect. Over the years, he spent several hundred thousand dollars on the project, which is the equivalent of some twenty million dollars today. The 29 tubes in his final 50 kW unit were the result of years of labor and fabrication of several hundred rejected tubes to get the 29 that worked. Accordingly, once those 29 were destroyed, Moray was never again able to obtain sufficient financing to rebuild his unit.

However, Moray had demonstrated a special type of transistor years ahead of the first transistor patent. He had demonstrated the importance of multiple junctions in transistor materials. Optical and nuclear functioning of the tubes was also woven in with nonlinear oscillations, multiple feedforward and feedback loops, etc. The sophistication of the tubes and feeds alone almost certainly guaranteed chaotic self-oscillation. We suspect but cannot prove that the Casimir effect may also be involved in some of his conducting metallic double surfaces.[122] He had demonstrated a remarkable "tube-with-internal-transistors" that produced energy output self-gain, without any operator input of energy. In addition, he had successfully used his special tube/transistor hybrids to produce a very successful COP>1.0 EM power system.

After Moray's death, his two sons John and Richard continued to try to rebuild the unit, but were never successful due to lack of sufficient funding and lack of a proper scientific team. They do know how to build the tubes, however, and they do know how to build the unit.

It appears that no electrodynamicist or physicist skilled in higher group symmetry electrodynamics ever examined the Moray device or worked with Moray. No particle physicist or electrodynamicist skilled in non-Abelian electrodynamics ever worked with the device. Since broken symmetry was not discovered and proven until 1957, no technical person who worked closely with the unit was ever aware of the various broken symmetries by which some of the virtual energy of the vacuum can be and is transformed into observable EM energy.

---

[122] The "grain" of the stressed semiconducting pellets may be important with respect to the Casimir effect versus the frequency. In short, Moray's contact surfaces between his semiconductors were "grainy" and these grains would have constituted Casimir surfaces. For the importance of the shape of the double surfaces, see M. Chaichian et al., "Quantum theories on noncommutative spaces with nontrivial topology: Aharonov-Bohm and Casimir effects," *Nucl Phys B,* Vol. 611, 2001, p. 383-402.

Certainly no one ever worked with it that understood the peculiar action of the difference frequency in traveling linearly through an isotropic nonlinear medium. John Moray, who stalwartly carried on the Moray lab in Salt Lake City until a few years ago, does know how to build the tubes, and he knows the circuitry for the device. That he does not understand the "deep theory" is of no consequence — no one else does either. In the opinion of this author, a funded scientific team — carefully chosen and working with John Moray — could probably rebuild the Moray device. We would very much like to see a Moray unit in production and contributing to permanently solving the energy problems of the world. Sadly, that does not seem likely.

5.4 Anti-Stokes Emission and Similar Processes

Stokes' law states that the wavelength of the fluorescent light emitted by a phosphor or other luminescent substance is always greater than the wavelength of the radiation used to excite the fluorescence {289}. This condition was observed by Stokes in 1852 {290}. From an energy viewpoint, the collecting medium collects and dissipates some of the input energy in the medium itself (e.g., to accelerate moving molecules) and then emits the remainder of it. The "system" consisting of the medium, the experimenter's input, and the output is such that it usefully outputs as luminescence less energy than the experimenter inputs. Hence it exhibits a COP<1.0, by absorbing excess energy in its "environmental surroundings".

Usually such a "law" generally holds but is sometimes violated in at least a lew cases. Stokes' law is no exception. It applies for most cases of fluorescence and luminescence, but it does not hold for a few of them. In a few cases, the emitted wavelength is precisely the same as the input wavelength. This means that the collector/absorber re-emits all the collected input energy as fluorescence or luminescence. Here the COP = 1.0. This condition is known as the *resonance condition* and the emission is called *resonance emission.*

In other well-known cases, Stokes' law is violated and the energy emitted is greater than the energy absorbed. In other words, the system exhibits a **COP>** 1.0 *with respect to energy emitted versus energy input by the experimenter.* This is known as anti-Stokes emission. It has been well known for six decades that anti-Stokes emmision processes do indeed output more energy than one has to input oneself {291}. This excess energy emission is usually euphemistically attributed to "additions from the internal energy of the molecule" and referred to as "negative resonance absorption by the medium" or just as "negative absorption by the medium".

Unless the excess energy taken from the physical medium of the system is replaced from outside the physical system, the anti-Stokes emission process produces cooling in the physical system and is self-quenching once the medium loses all the energy it can afford. For the overunity researcher, a huge hint is to search the literature for anti-Stokes type emission phenomena that do not result in media cooling. The Letokhov phenomena {292} are in fact such phenomena where the excess emitted energy is received from outside the physical system medium, and therefore from the surrounding active vacuum and/or the surrounding local curvature of spacetime.

Other than for a very few physicists — e.g., Bohren {24}, Letokhov {157}, H. Paul and R. Fischer {25} — there appears to have been little *effective* discussion of where the excess energy "extracted from the molecule" really came from initially, and how it is replenished to the molecule (if at all), and particularly *whether or not it has to be replenished to the molecule by the operator.* There has been no discussion of the local potentials as dipole asymmetries in the virtual energy exchange with the vacuum, although there has been inadequate discussion of the involvement of the Poynting energy flow (e.g., by Bohren, Paul, and Fischer). There has been no discussion of the possible involvement of the Heaviside nondiverged component of the energy flow, which is additional to the Poynting component but unaccounted by today's electrodynamicists. To my knowledge, the broken symmetry of every dipolarity has not been discussed with respect to such experiments.

The reader can see the point. If a mechanism exists or can be evoked to freely resupply the lost energy to the molecule from the other two components of the supersystem, then the resulting *self-replenishing* anti-Stokes emission phenomenon can be utilized as a legitimate and continuing COP>1.0 process. It remains to be seen whether this can be worked out. If we wish, the giant negentropy mechanism of the source dipole and source charge {308} may be considered universally proven cases of self-replenishing anti-Stokes emmission, already exhibited by every charge and dipole in the universe.

At least one audacious scientist — Letokhov — has suggested that similar processes in certain systems can perhaps comprise true Maxwell's demons, yielding excess energy output and perhaps eventually a COP> 1.0 operation {293}. Russell has suggested that a similar COP>1.0 process is involved in the fiber fuse phenomenon {299}.

Over the last few decades there has been slow but important work in an optical area of possible anti-Stokes emission phenomena {294a-c}. A variety of strongly scattering, optically active media have been shown to produce substantially more energy output than the experimenter inputs to stimulate the output. However, most are being done in laser-like pumping situations and so the pump energy has to be furnished by the operator. This means that the "replenishment" energy to the medium is furnished by the operator, resulting in overall COP<1.0 performance. However, in self-oscillation conditions, the medium can be self-pumped (self-replenished, taking its replenishment energy from the active vacuum exchange when viewed in the supersystem), and in such conditions COP>1.0 performance is possible, at least in theory. Presently we know of no one who has achieved it — except in the destructive fiber fuse effect — but the work seems to be progressing steadily toward that eventual outcome.

5.5 Gain in Intensely Scattering Optically Active Media

This modern work — which may be regarded in one sense as extending the anti-Stokes emission effect — is largely being pursued in intensely scattering optically active media. A very nice example is given by Lawandy's experiment {295a-d}, shown in Figure 5-4. Figure 5-4a shows a comparison experiment in water without an intensely scattering optically active medium but with a fluorescent dye to show the emission interaction. Figure 5-4b shows the same experiment with titanium dioxide particles added to the fluid. The TiO_2 particles are sized so that their optical resonance is within the laser's frequency domain.[123]

First, as shown in Figure 5-4a, a small, weak laser beam is aimed into the solution, and a small "warm little glow" of fluorescence results just around **that** region where the laser beam interacts with the solution. Scattering in **the** medium is rather normal, and provides nothing of any great interest to the overunity researcher.

---

[123] In passing, we point out the involvement of the Bohren effect due to particle resonance. This collects additional energy from the usually nondiverged large **Heaviside** energy flow component.

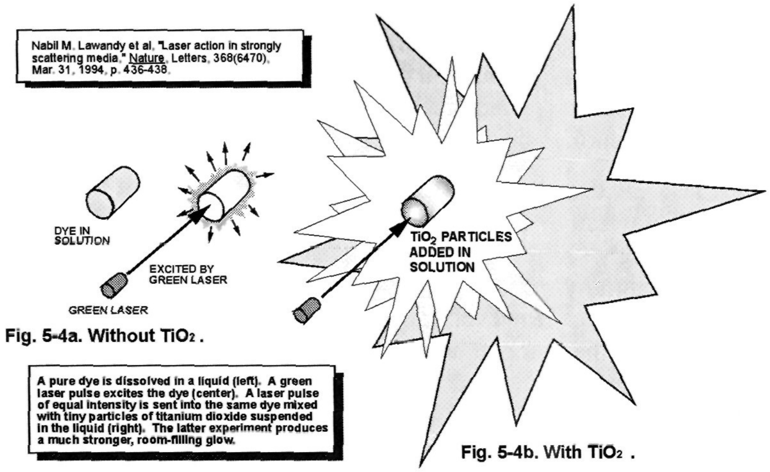

Nabil M. Lawandy et al. "Laser action in strongly scattering media," Nature, Letters, 368(6470), Mar. 31, 1994, p. 436-438.

Fig. 5-4a. Without TiO_2.

A pure dye is dissolved in a liquid (left). A green laser pulse excites the dye (center). A laser pulse of equal intensity is sent into the same dye mixed with tiny particles of titanium dioxide suspended in the liquid (right). The latter experiment produces a much stronger, room-filling glow.

Fig. 5-4b. With TiO_2.

Figure 5-4 The Lawandy experiment

Then the titanium dioxide particles are added, and the same weak illuminating laser beam is directed into the colloidal solution. Figure 5-4b shows the new and spectacular results.[124] Immediately a very bright, room-filling emission of scattered coherent light fills the entire room. The intense optical scattering includes extensive retroreflections for some "ping pong" between individual particles, so that the optical gain is enormous. A highly enhanced energy emission is now evoked, for the same energy input by the operator as was used in the first experiment without the TiO_2 particles. Lawandy's experiment is inexpensive and can be repeated in any university nonlinear optics laboratory and by many individual experimenters. It gives results every time, without fail.

Oddly, no one seems to clearly state that, for a single pulse of input laser energy to the suspension of TiO_2 particles, precisely how much energy was input and how much energy was then emitted by the medium. Instead, "gain" is mentioned, but never the COP = (energy output by medium emission) / (energy input by operator). One suspects that journal referees would probably not allow such a clear and unequivocal statement of overunity, but would apply quibbling and spin control to prevent stating that more energy is output by the medium than is input to it by the operator.

---

[124] This is a very easy and very convincing argument for the presence of the long-ignored Heaviside component of energy flow arbitrarily discarded by Lorentz more than a century ago, and still discarded by modern electrodynamicists.

It follows that, if sufficient "ping pong" iterative retroreflection and multiple collection occurs in the medium, the COP>1.0 because more work is done on exciting the medium than the energy input to it by usual (Poynting) calculations. No one accounts the energy unwittingly input to it by the ignored Heaviside component accompanying the Poynting component. Usually the output/input is just referred to as "optical gain". We stress, however, that we are considering only the actual laser beam energy being furnished into the colloidal solution as input, and we are not calculating the efficiency of the laser, its pumping, etc. Anyway, for this stimulated emission process, it is possible that the COP>1.0, and one can see the parallel to anti-Stokes emission. In short, we are considering the TiO_2 solution itself as exhibiting a true negative resistor action.

The field of high gain stimulated optical emission is advancing rapidly {296}, although many experiments still use external pumping furnished by the operator. If the operator has to pay for the replenishment (pumping) energy as well as the input energy, then the COP<1.0 overall. But the experiments are tending toward sustained self-oscillation and self-pumping conditions. If such conditions are obtained in sufficient magnitude, the process will then become a legitimate COP>1.0 operation overall. In that case, it will be usable as a basis from which to develop self-powering COP>1.0 electrical power systems, particularly for infrared "heating" systems etc.

The more recent experiments have shown positive feedback loops both in the time-forward and time-reversed paths; trapping of light flow energy (both time-forward and time-reversed) in large random walks of more than 1,000 individual interactions; weak Anderson-type localization; and constructive interference of forward time and reversed time light paths. since such experiments can be performed in the infrared, they point toward a potential *"vacuum-energy-powered overunity heater"* as a **feasible** achievement in the future. In our opinion, there should be a determined and major Department of Energy program oriented to develop exactly that kind of system.

Such a heater can become self-powering by the presence of greater governed positive feedback during self-oscillation conditions, which will allow sufficient excess collection due to *multipass multicollection* from the usually wasted giant Heaviside energy flow component. This process —

with the self-excitation occurring spontaneously as a "kick-in" positive feedback process in an exploding gas — probably accounts for the phenomena observed in the gamma ray burster and other such violent "burster" cosmic phenomena. Re-ignition, afterglow, and similar effects are observed in gamma ray bursters. They are also observed with *remarkable* similarity in the latest experiments in intensely scattering optically active materials the laboratory. Similar phenomena occur in x-ray bursters as well, and perhaps even in the recently observed and confirmed gamma ray emissions from intense storm clouds.

Finally, we point out that many magnetic materials are also photorefractive, and they readily produce nonlinear optical effects at various frequencies. Some barium compounds are typical examples. As an example, multivalued phase conjugate reflection can occur {297}. Such effects did occur in the Sweet vacuum triode amplifier, to be discussed shortly. It may be that Sweet's conditioning of his magnets conditioned their barium nuclei into self-oscillation, self-pumping, and thereby into *self-replenishing* stimulated optical-type emission at ELF frequencies.

5.6 The Fiber Fuse

See Figure 5-5. Although presently a one-shot destructive process, another multicollection phenomenon exhibiting a novel COP>1.0 process is the *fiber fuse* {298a-e}. This is a well-known but unusual effect in optical fibers and in fiber optic cables.

See Figure 5-5a. Once initiated by simply heating the cable (as with a butane cigarette lighter), the fiber fuse ignites and is self-sustaining. In those susceptible systems utilizing power on the order of one watt or so, the ignited fiber fuse propagates down the cable slowly, at about a meter per second nominally. It is evidenced by a bright white light propagating down the fiber, resembling the burning of a fuse for explosives — which led to the name "fiber fuse".

In fiber optic cables, it occurs only in those that contain germanium in their cores, and not in those containing just silicon in their cores. It moves down the fiber or cable at approximately one meter per second, and it is unlimited in the length of cable that it will continue along and march to the end. In the core, the marching fuse is actually "hopping", so to speak, and it melts a little "hole" about every centimeter or so, blowing out the material from the hole as it occurs. This "steadily and slowly marching" series of core melts results in the catastrophic destruction of the optical fiber core of the cable and thus destroys the functioning of the cable.

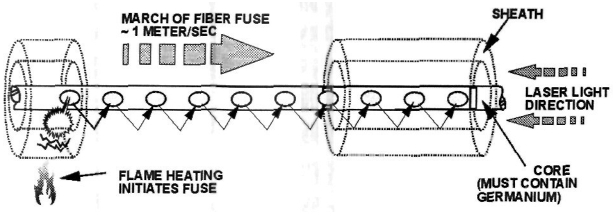

Figure 5-5 A. Fiber fuse destroys core of fiber optics cable of indefinite length, pitting core with holes.

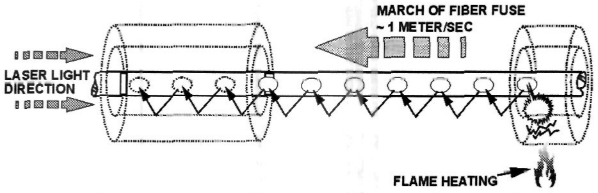

Figure 5-5 B. Reversed second fiber fuse often *restores* pitted core in cable, filling holes back up.

Figure 5-5 The fiber fuse.

The mechanism has been theorized to be self-propelled and self-focusing. Other experiments have found evidence for thermal shock waves and chemical reactions.

But see Figure 5-5b. Astonishingly, after a — say — 1 kilometer fiber optics cable is destroyed by such a fiber fuse, it is often possible to "heal" the damage by reversing the direction of the normal laser light propagating down the cable, while re-initiating the fiber fuse from the other end by heating the cable on that end. In that case, sometimes the *reverse-ignited* liber fuse will then march right back down the cable, precisely refilling those previously-melted little holes every centimeter or so, *restoring the functioning of the cable!* If ever a phenomenon cried out for further painstaking investigation to determine its actual mechanism, the fiber fuse is it.

Truncated coupled-wave theory applies to this phenomenon. A rigorous analysis by Russell {299} of power conservation in dielectric gratings has shown that truncated coupled-wave theories do not rigorously conserve power. Predictably such conclusions have been disputed {300a-300d}, however, and are adulterated by the same old use of Lorentz-surface integration of the Poynting vector around the grating and discarding of the nondiverged Heaviside energy flow component. Those disputing the results and claiming the impossibility of COP>1.0 systems continue to ignore the unaccounted Heaviside energy flow component accompanying every EM energy flow situation. In fact, they continue to unwittingly discard that component, and so have never performed *a fully* adequate analysis of any circuit or process. They also are unaware that all EM energy in 3-space comes directly from the active vacuum — specifically, it enters 3-space from the time domain and then returns to the time domain.

As long as the Lorentz surface integration scheme continues to be ubiquitously used without understanding what it does with respect to energy flow accounting, the dispute will not be dissolved. Presently both sides still use the Lorentz integration and discard the available Heaviside component. So they are really disputing whether the *Poynting diverged component* of the input energy can be equal to the energy dissipated. The answer to that is simply "No!"

Of course in a lab test the fiber fuse can be initiated with small Poynting energy input by precision point heating. In that case, the total energy required to melt the holes and move the material out of them, every centimeter or so down that (unlimited) length of cable, can be made as large as one reasonably wishes, simply by making the cable as long as one wishes. In short, COP»1.0. As can be seen, it can be experimentally proven that the input *Poynting* energy component can be far less than the total energy dissipated in damaging the cable. On the other hand, the energy dissipated in damaging the cable *is not* greater than the sum total of the usually calculated Poynting component together with the neglected extraordinarily large Heaviside nondiverged component. Conservation of energy is upheld, but only if the long-neglected Heaviside component is also accounted.

It stands to reason that the reverse fiber fuse can also be initiated by less input Poynting energy component that the total energy dissipated in restoring the cable. At that point, conventional theory has a problem. One can start with a good cable, put in a little EM energy twice — once to ignite the fiber fuse and once to ignite the reverse fiber fuse — and one winds up with a good fiber optics cable and having done lots more work inside the cable than the amount of energy one input by conventional considerations. We leave the solution to that problem as an exercise for the reader. Again, unless one includes the long-neglected Heaviside component, there is no way to make the energy flow balance, and one will have an experiment that seemingly contradicts the conservation of energy law. It doesn't, and that is the exercise the reader should understand.

So unless one uncovers and specifies the external source (Heaviside's energy flow component) of excess energy feeding the fiber fuse, it is a demonstrable experiment whose efficiency ξ seemingly winds up being $\xi > 1.0$. It works every time and seemingly violates the conservation of energy law *as that law is conventionally and imprecisely stated for an equilibrium system only.* The COP>1.0 fiber fuse phenomenon does not violate overall conservation of energy because the system is not in equilibrium with the other two components of its supersystem. It only violates (permissibly!) the conservation of energy statement for a system in equilibrium. The "system" consisting of the fiber fuse and the cable is *not in equilibrium in its energy exchange with its active environment (with the other two components of the supersystem, consisting of the active vacuum energy exchange and the active local curvature of spacetime).* We point out, of course, that the Heaviside component is flowing in space outside the physical system, hence in the vacuum and therefore in the external environment.

Specifically, every dipole in the system is extracting and pouring out energy from the vacuum, and only the small Poynting diverged component of this available energy flow is ordinarily caught. The extreme nonlinearity (and some resonances!) of the particles in and exposed to the extraordinarily locally heated fiber fuse do sweep out a greater reaction cross section than do cold, static charges, in a fashion similar to the separate demonstration of such extended geometrical reaction cross section in the Bohren experiment {301}. *Hence the fiber fuse subsystem and the associated particles of melted matter from the holes do intercept, diverge, and collect some of the accompanying but usually nondiverged Heaviside component.*

See again paragraph 2.1.6 in Chapter 2. Iterative retroreflective systems — such as in intensely scattering nonlinear optical materials — are not necessarily limited to COP < 1.0. Such a system is not necessarily limited to only the energy input by the operator, and it is not limited to only the operator's input Poynting component. Instead, the system can receive an input of excess energy from (i) its active local vacuum, and/or (ii) the local curvatures of spacetime. Its efficiency ξ, however, is always $\xi < 1.0$, even when its COP is clearly greater than 1.0.

Since scientists disputing the energy aspects of the fiber fuse only have assumed an equilibrium system and have not considered the neglected Heaviside energy flow component, the present arguments against COP>1,0 performance of the fiber fuse are quite irrelevant.

We also point out that an optical phenomenon is involved, and that the core materials in the melt are optically active. Hence, appreciable iterative phase conjugate retroreflections occur in the fiber fuse itself and particularly in the melt materials (in the "inferno" of the "spot" itself). This asymmetrical self-regauging process gets more than one joule of work from one joule of original input Poynting energy, because of the iterative (and reciprocating) change of form of the energy. That operation produces iterative joules of work for each joule of original input energy, plus the unaccounted original input of the Heaviside component.

Again, there is no law of *conservation of work* in nature. The present conventional energy-work theorem erroneously considers only a "single interception and change of form" of the energy collected in the system, before loss of the energy from the system. Therefore, it only considers each joule of input energy performing one joule of work on the system to "excite" it to a higher potential energy state. It also considers that, once a joule of energy is then dissipated from the system as work in an internal load or process, it has departed the system forever and is not returned to do additional work by having its form changed yet again! In intensely retroreflective or intensely scattering systems, those conventional assumptions do not necessarily hold. For analysis of the fiber fuse phenomenon, the extended work-energy theorem must be utilized.

Finally, we point out the connection of this fiber fuse phenomenon to the similar "runaway" phenomena we have previously postulated as involved in producing the x-ray burster, gamma ray burster, etc. where iterative retroreflection and thus asymptotic rise in energy density is involved.

Another possibly related phenomenon is the unexpected explosion of porous silicon, as discovered at the Technical University of Munich {302}. The porous pellet involved in the unexpected experiment is covered with a layer of hydrogen atoms. In the presence of oxygen penetrating the oxygen barrier, an extraordinary and superfast chain reaction is initiated. The power of the explosion is some 10 times that of TNT. In the inferno portion of the fiber fuse, the "melt" as it puddles and then particulates, should be in the presence of oxygen and hydrogen from air in the cable. This dramatically differs from the explosive pellet, but may be sufficient to ignite at least in a local area rather than explode violently. If so, between the ablating particles of silicon there could exist iterative phase conjugation, thus producing an asymptotic rise in temperature and energy, perhaps sufficient to sustain the fuse. At least this possibility should be investigated.

5.7 Multiple Feedforward and Feedback Semiconductor Loops.

DeSantis *et al.* {303} showed that feedback systems with a multipower open loop chain can produce COP > 1.0 performance. In semiconductors, the present author hypothesizes that complex feedforward and feedback loops involving semiconductors can induce time reversal effects, hence some movements of current against the voltage. In short, there can arise a true negative resistance effect, as demonstrated in the Fogal transistor.

Indeed, a Westinghouse frequency converter using 64 transistor stages and similar sophisticated feedforward and feedback mechanisms was placed in the original Minuteman missile, then was later deliberately modified by Westinghouse engineers to prevent its demonstrated COP > 1.0 performance.[125] The specific germanium transistor used in the converter was also withdrawn from the market.[126] In those days, power meant weight and weight was at a premium in missile design. The power supply was very tightly designed, so that the weight could be minimized.

The converter had been deliberately designed to be 90% efficient, which was quite the state of the art. However, placed in the missile, the average converter would produce from 105% to 115% times as much output power as input power. (We are stating a COP > 1.0, *not* ξ > 100%!) This was more output power than the tightly designed succeeding stages were designed to accept; consequently those overloaded follow-on circuits often burned out and deadlined their Minuteman missiles — something of great strategic concern at the time. The increased rate of Minuteman missiles out of action due to this problem placed great pressure on Westinghouse executives. In turn, they simply ordered the engineers to "fix it!"

Westinghouse simply "fixed the problem" very directly. The engineers reduced the output power of the converter by adding limiting diodes etc. Very quietly, Westinghouse then obtained several patents {304a-c} surrounding the technology, but no further mention of it appears in the literature. The Westinghouse work continued for at least a decade, judging by the spread of the patent dates.

---

[125] Our information came directly from private conversations of a close scientific colleague with the Westinghouse engineers who thoroughly tested the converter causing the Minuteman problems, and who then altered it to prevent the problems.

[126] Germanium has some unique properties, particularly with respect to transducing surface longitudinal waves to transverse waves and vice versa.

Such multi-loop, multi-feedback transistor circuits can apparently be developed to exhibit true negative 4-resistor effects. With specialized clamped positive feedback of the Bedini type, such open dissipative systems can then exhibit stable COP>1.0, since the energy is taken from the vacuum via the broken symmetry of dipoles, and then discharged asymmetrically in the loads so as not to destroy the source dipole negative 4-resistors. However, the reader should see Chapter 9 for special considerations and phenomenology.

5.8 Negative Resistors

We define a *true negative resistor* as a circuit component or system that receives energy from the environment and outputs useful energy, either in altered form or to power loads usefully. This contrasts sharply with a *differential negative resistor,* which exhibits negative resistance in one part of its regimen, but where all the excess energy to move the current against the voltage in that regime was previously drawn from the circuit itself and therefore was not freely received from the active external environment. Instead, the operator paid for it already. We are not interested in differential negative resistors, which are well known in the literature {305} and do not enable COP>1.0 system performance, so we will not discuss them further. Hereafter when we use the term "negative resistor", it is understood that we mean "true negative resistor" unless the term is otherwise qualified.

The true negative resistor is an open dissipative system *a priori,* and equilibrium thermodynamics therefore does not apply. Instead, the thermodynamics of open systems far from equilibrium applies. The negative resistor freely receives energy from outside the system (from the . environment), and "dissipates" it in interception and collection actions inside the system, to freely increase the available potential energy in the system.

In circuits, the main characteristic of a negative resistor is that the environment freely furnishes some excess energy to (i) power the load, and/or (ii) move the current back against the voltage, particularly when shunted across the back emf region of the source dipole. The *operator* does not have to furnish this excess energy dissipated to propel the current backwards or dissipated to power the load!

One must incorporate the entire supersystem into the analysis of a negative resistance element or system. Since the operator does not furnish the input energy dissipated by the negative resistor, then the active vacuum or curved local spacetime — or both — must furnish the input energy to it. One must consider local curvature of spacetime interactions with the negative resistance and the active vacuum interactions with it, as well as the interactions between active vacuum and local spacetime curvature. In other words, one must again include the active environment of the Maxwellian system that Lorenz/Lorentz symmetrical regauging arbitrarily discarded.

We specifically do not include the well known "differential" negative resistor, such as a tunnel diode, which in one regime moves current against the voltage, all the while drawing all its enabling energy to move current against the voltage from an overall previous energy input by the operator. Unless the input enabling negative resistance energy is freely received from the external environment, true negative resistance is not involved.

Several true negative resistors other than the point-contact transistor have been produced. We review a few of them in the following subparagraphs.

5.8.1 Kron's Negative Resistor

One of the greatest American electrical scientists was Gabriel Kron. Working for General Electric (GE), Kron built a true negative resistor in the 1930s, which could power itself and the network analyzer[127] at Stanford University, under a GE support contract with the U.S. Navy. Here **is** a direct quote from Kron {306} to show what we refer to:

> *"When only positive and negative real numbers exist, it is customary to replace a positive resistance by an inductance and a negative resistance by a capacitor (since none or only a few negative resistances exist on practical network analyzers.) "*

In that sentence we believe that Kron was forced by the censors to insert **the** words "none or" in the expression in parentheses. Particularly note the suggestion that in the imaginary realm it may not be the same as the accepted practice of using inductance as a positive resistor and capacitance **as** a negative resistor. Of course, the fourth Minkowski axis is - *ict,* and so this is very probably an oblique reference to the fact that energy can also "flow in" from the time domain. Kron was quite knowledgeable in general relativity as well as advanced electrodynamics, and he applied full general relativity to rotating EM machines etc. In another quote, Kron {307} also revealed that he had the negative resistor but was not allowed to use it to openly power the Network Analyzer. Quoting and reading through the "spin control":

---

[127] The network analyzer was a large analogue simulation capable of modeling sophisticated equations — such as Maxwell's equations, the Schrodinger equation. **etc.** — for use in designing, simulating, and analyzing naval communications systems or proposed systems. It was one of the most advanced simulations of its day. Kron was the chief engineer and scientist on the project, and one of the ablest electrical physicists the United States ever produced.

> *"Although negative resistances are available for use with a network analyzer, in practice it is more convenient to use a second type of circuit, in which the positive and negative resistances are replaced by inductors and capacitors and the dc currents and voltages are replaced by ac currents and voltages of fixed frequency. The use of the second type of interpretation is equivalent to multiplying the wave equation by i =* $\sqrt{-1}$*."*

After the word "although..." Kron includes the blunt statement that negative resistors were indeed available for use with the network analyzer. But if the "convenient" replacement of the positive and negative resistors by inductors and capacitors was made, then there was no COP>1.0 involved, and there was no true negative resistance involved. Kron is trying to clearly tell us that fact in the above quotation. He also strongly hints that energy from the time domain on the axis *ict* is involved, but that multiplication by $\sqrt{-1}$ discards this "energy from the time domain".

Kron's negative resistor secret has never been released by General Electric, Stanford University, or the U.S. Navy and it is not going to be released. It has never been deciphered outside those groups, with the possible exception of knowledge by Floyd Sweet of Kron's negative resistor construction. Sweet worked in General Electric but not on the Network Analyzer project. Kron, however, was Sweet's mentor and patron, and Sweet often spoke of Kron in glowing terms. In our opinion, Sweet's vacuum triode amplifier was a further development of Kron's early negative resistor, as Sweet understood it.

We further believe that Kron's "open path" is simply the other two components of the supersystem, being (i) local curved spacetime, and (ii) the active local vacuum. Obviously the *external environment* connects any two points in the physical circuit, in addition to the circuit's physical connection (the "closed path"). Between any two points at differing potentials in an electrical circuit, there exists a potential and a dipolarity. Therefore, there exists a broken symmetry because of the dipolarity. The potential difference of that dipolarity is also a curvature of spacetime; i.e., it is a "field" between the two points, and the field is identically a curvature of spacetime. Hence from any two such dipolar points in the circuit, there exists the giant negentropy flow of EM energy that we pointed out in 2000 {308}. This "open path" EM energy flow between any two unlike charges (any dipolarity) in the universe, not intercepted by the circuit and thus nondiverged, is discarded by Lorentz's integration of the energy flow vector around a closed surface enclosing any volume element of interest. It appears that Kron not only recognized (in different terms) what we are calling the "supersystem", but also recognized the equivalent of Heaviside's nondiverged EM energy flow component from any dipolarity — referring to that energy flow as the "open path" flow of energy. It is almost certain that he fully recognized what Lorentz symmetrical regauging did to the Maxwell-Heaviside equations. Thus it appears that Kron was the first scientist to discard Lorenz/Lorentz symmetrical regauging, producing his negative resistor by doing so.

We do not have firm knowledge, however, but only circumstantial evidence that Kron recognized the flow of EM energy from the time domain into 3-space, and vice versa, though he was certainly capable of seeing that. Nevertheless, this would also be included in his notion and discovery of the "open path", since it is included in every dipolarity, generating the scalar potential between the poles of the dipolarity. It may also have occurred to him from his excellent grasp of general relativity and spacetime curvature {309}.

Kron's concepts are certainly capable of being modeled and used to produce COP>1.0 EM circuits and systems, including his negative resistor. Kron did tremendous work in advancing EM networks and power systems. He left behind a great legacy of network and system analysis capable of dealing with the COP>1.0 electrical power system problem {310a-e}. Unfortunately, that legacy has been little used, if at all, in design of electrical power systems and power distribution networks today. To our knowledge, there are no electrical power systems deliberately using Kron's open path.[128]

5.8.2 Chung's Negative Resistor

A highly simplified diagrammatic illustration representing the Chung negative resistor is given in Figure 5-6. In a July 9, 1998 keynote address at the Fifth International Conference on Composites Engineering in Las

---

[128] A possible exception is provided by Barrett's extensions {286a, 286b} of Tesla's "potential shuttling" technique {124}, which can only be seen in higher group symmetry electrodynamics such as quaternions. However, as far as we know, Barrett has **not** applied energy shuttling techniques to electrical power systems.

Vegas, Dr. Deborah D. L. Chung,[129] professor of mechanical and aerospace engineering at State University of New York (SUNY) at Buffalo and Director of the Composite Materials Research Laboratory, reported that she had observed apparent negative resistance in interfaces between layers of carbon fibers in a composite material prepared under pressure.[130] The negative resistance was observed in a direction perpendicular to the fiber layers. By varying the pressure at which the composite material was prepared, Chung *et al.* were able to control whether the material would exhibit negative, minimal, or positive resistance. Apparent negative resistance was also observed in carbon fiber cement-matrix composites and in bare carbon fibers held together by pressure.

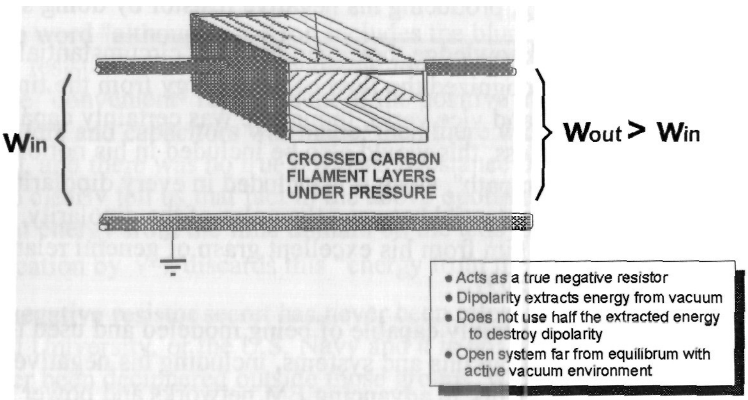
Win
Wout > Win
CROSSED CARBON
FILAMENT LAYERS
UNDER PRESSURE
• Acts as a true negative resistor
• Dipolarity extracts energy from vacuum
• Does not use half the extracted energy to destroy dipolarity
• Open system far from equilibrum with active vacuum environment

Figure 5-6 Chung's negative resistor (diagrammatic).

A formal technical paper was later published by Wang and Chung {311}. A few years earlier a patent {312} had been issued to Chung and assigned to The Research Foundation of State University of New York, Albany, NY for a superconductive material using a laminate combination of carbon fibers and metal.[131] The patent mentioned mesoporous carbon filaments, strikingly similar to our previous discussion of the surprising explosion of porous silicon at the Technical University of Munich, as well as possibly a special kind of "porosity" consideration in the fiber fuse phenomenon (in the material of the fuse itself).

---

[129] Professor Chung holds the Niagara Mohawk Chair in Materials Research at SUNY and is internationally recognized for her work in smart materials and carbon composites. The reader may peruse a curriculum vitae, research work, and papers at http://www.mae.buffalo.edu/people/faculty'chung/researchl.doc.

[130] The preparation of "locked-in" stress in Chung's carbon fiber material is reminiscent of Moray's similar preparation of his amorphous semiconductors with locked-in stress also.

An initial offer on the State University of New York at Buffalo website to provide a technical package with respect to the negative resistor — for which a patent had been filed — to those submitting a nondisclosure agreement was abruptly withdrawn. The project may have been classified, or — far less likely — it may have been retained as deeply proprietary by the University. We suspect that both the technical package and the patent application have been classified by one or more U.S. government agencies.[132]

In our opinion, Chung's fruitful discovery deserves rigorous and continuing investigation, particularly in higher group symmetry electrodynamics and in terms of supersystem interactions to include Dirac sea hole current interactions. We also believe Chung may have been required by referees (or by declassifying authorities, if such were involved) to use the term "apparent" in the paper. The reported measurements in the technical paper do seem to unequivocally include true negative resistance with current running backwards against the voltage. It is not clear, however, whether this is Dirac hole current or electron current or a mix of both.

We also call attention to the fact that Dirac sea holes in the local vacuum — as will be discussed in Chapter 9 — will and do run backwards against the voltage. Further, such hole current can be transformed from negative energy to positive energy by the Bedini process shown in Chapter 9. The pressure used by Chung *et al.* in fabricating the material produces a stress potential in the material and thus decomposes into our reinterpreted Whittaker 1903 decomposition, preciously discussed. So giant negentropy is involved in the composite material, which implies that true negative resistance can be involved, at least in some operational regimes as pressure etc. are varied. Such results were found by Chung *et al.* and are reported in the paper.

---

[131] We stress that the issuance of this unclassified patent does not rule out a possible negative resistor (hence COP>1.0) patent that may have been classified. Some sensitive defense work in special materials is indeed performed at SUNY at Buffalo.

[132] At least the gist of the work is not lost, however, due to Naudin's duplication of the Chung effect and placing the results and building instructions on his website in France; see http://jnaudin.free.fr/cnr/.

Naudin reported independent success with a simpler variant of the Chung negative resistor {313}. This assures that at least the effect will not just be buried.

5.8.3 The Fogal Charge-Barrier Device

Figure 5-7 shows an illustration from one of William Fogal's patents {314} on a new kind of transistor (more accurately, a charge barrier device). Fogal's device incorporates a very special electrolytic capacitor[133], which introduces optical-type functions as well as the more familiar capacitor functions.

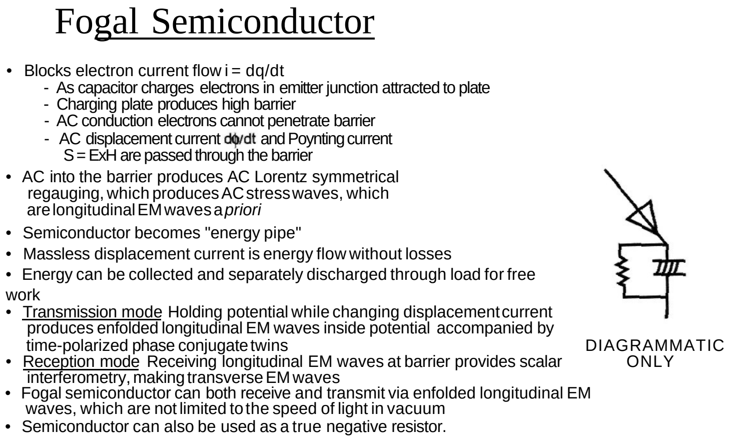

Fogal Semiconductor

- Blocks electron current flow $i = dq/dt$
  - As capacitor charges electrons in emitter junction attracted to plate
  - Charging plate produces high barrier
  - AC conduction electrons cannot penetrate barrier
  - AC displacement current $d\phi/dt$ and Poynting current $S = E \times H$ are passed through the barrier
- AC into the barrier produces AC Lorentz symmetrical regauging, which produces AC stress waves, which are longitudinal EM waves a *priori*
- Semiconductor becomes "energy pipe"
- Massless displacement current is energy flow without losses
- Energy can be collected and separately discharged through load for free work
- Transmission mode Holding potential while changing displacement current produces enfolded longitudinal EM waves inside potential accompanied by time-polarized phase conjugate twins
- Reception mode Receiving longitudinal EM waves at barrier provides scalar interferometry, making transverse EM waves
- Fogal semiconductor can both receive and transmit via enfolded longitudinal EM waves, which are not limited to the speed of light in vacuum
- Semiconductor can also be used as a true negative resistor.

Figure 5-7 Novel characteristics of the Fogal semiconductor.

Most semiconductor materials are optically active materials. Further, the individual electrons in a signal with a frequency well below light frequencies, are "jittering" with higher frequencies including the entire optical band. A "frequency" of a signal thus is imbedded in a jittering higher frequency complex set, and that set is quite able to interact in optical fashion *if the complex itself is used and controlled.* Fogal's process does appear to both use and control that jittery high frequency signal complex.

---

[133] The exact method of manufacturing this electrolytic capacitor, and its internal structuring, is proprietary to Fogal and a major capacitor manufacturer, who is under a nondisclosure agreement with Fogal.

Hence Fogal's process can "tunnel under" the normal *detected* force-field frequency responses of the individual electron,[134] and directly affect, change, and alter its higher-frequency environmental changes in the force-free field or field-free potential condition. In this way, Fogal has carried the electron (and hole) functions of a transistor to functions well beyond those in the present texts {315}. He may have also extended quantum mechanical tunneling effects to signals between widely separated regions of spacetime, since superluminal communication is possible with the Fogal semiconductor properly used. Longitudinal EM waves — which Fogal's semiconductor is able to utilize and control — can and do move at superluminal speeds {93a) because it uses the long-neglected "infolded" longitudinal electrodynamics revealed by Whittaker in 1903 and 1904. Further, superluminal solutions to Maxwell's equations do exist {93b}.

The charge barrier device is formed in a bipolar design that can be incorporated in Metal Oxide Semiconductor (MOS) designs as well as multiple gate devices. The device also produces a Hall effect EM field internal to the device. The Hall effect magnetic force will force electron flow and angular spin of the electrons in the same direction to the top of the conduction bands in the crystal lattice on semiconductor devices, unlike Silicon On Insulators (SOI) devices that force electron flow to the surface of the semiconductor lattice. This forcing function, together with the "extended tunneling" effects on the embedded EM environment of the electrons and holes, produces a unique effect. The device is able to stop and "pin" electrons while continuing to process the signal energy flow (the entire jitter signal complex) in which the electrons were coupled and moving. Thus the Fogal transistor is able to operate upon and with mass-free spacetime itself, and curvatures thereof. It is the first practical semiconductor capable of performing *engineering of the active vacuum.*

In the last few years we have stressed that EM waves in spacetime are longitudinal EM waves *a priori* — but they are accompanied by time-polarized waves (phase conjugates). Hence the two in combination become observable, as pointed out in quantum field theory by Mandl and Shaw {316}. *We detect effects a priori and not their causes.* We detect transverse *electron precession waves* in our detector circuits, and not the causal waves interacting with them to produce their precession. *The Fogal transistor is able to process the causal longitudinal EM waves directly, prior to their interaction with matter.* It in fact opens up an entirely new region of "detected physical reality" to our vision and detection, a region far more extensive than the development of infrared and ultraviolet sensors opened up to our sight.

---

[134] The electron itself is enveloped in the jittering of whatever background it is exposed to and interacts with. Hence a coupling interaction exists to arrange and **control** the jitter of the electron as well.

The pioneers of electromagnetics, who erroneously assumed a material ether and EM force fields in space, confused the transverse "material detected waves" (electron precession waves) in the detector with the "material undetected waves" they assumed to be incoming from space. Note that neither the electron nor the atom had been discovered when all these assumptions were used to determine how the EM wave in space was to be viewed and modeled. The transverse waves observed in the interacting/detecting of the Drude electron gas — i.e., the detected electron precession waves — were considered to be direct interceptions of the incoming causal waves from space. This already substituted the effect for the cause, as we have previously discussed.

Further, Faraday envisioned the "material field in space" as consisting of a series of "taut physical strings". A "perturbation" of this field in space was, in his view, simply the transverse wave plucking of those strings. Maxwell, who closely followed Faraday's experimental work and concepts, simply assumed the transverse "plucked string waves" advanced by Faraday. Therefore, the erroneous notion of the transverse EM wave in space arose and stayed firmly fixed in the theory.

One can check Whittaker's decomposition of the scalar potential into longitudinal EM wave pairs, then realize that the energetic vacuum is naught but a giant scalar potential — which can be decomposed by Whittaker's method. In short, as a scalar potential, space itself is a writhing mass of longitudinal EM waves and their highly nonlinear dynamics and interferences, or it can be modeled in that fashion. The energy continually enters 3-space from the time domain to the negative charge of every point in space considered as a point dipole, thence to the positive charge, thence back to the time domain. Notice the giant 4-space circulation involved. One also notes the polarization of the vacuum continually happening and arising spontaneously, as is well known in quantum mechanics. We also suspect that this or a similar mechanism is at the root of spontaneous broken symmetry in physics.

An interesting thing emerges from Whittaker's 1904 paper {317}, where he showed that all EM fields and waves can be decomposed into differential functions of two scalar potentials. Each of these two base scalar potentials can be decomposed by Whittaker's earlier 1903 paper {85} into a set of longitudinal EM waves. All EM fields, potentials, and waves are comprised of longitudinal EM waves and their internal dynamics, which also comprise waves of spacetime curvatures and their dynamics. Further, each longitudinal EM wave in 3-space is accompanied by its phase conjugate twin in the time-domain.

This far more primary "internal electrodynamics comprised of pure spacetime dynamics" — comprising and creating the recognized EM potentials, waves, and fields, their internal dynamic spacetime structuring and also the structuring of time — has not been developed and applied in physics theory. This internal or "infolded" electrodynamics is the primary "medium" in which the Fogal semiconductor appears to work and work upon. The Fogal semiconductor seems able to modulate and demodulate this internal structuring or parts of it, as has been demonstrated.

Imperfect longitudinal EM waves — called *undistorted progressive waves* - are already known to be able to travel at superluminal velocity, e.g., as shown by Rodrigues and Lu {318}. Indeed, Maxwell's equations have both subluminal and superluminal solutions, as shown by Rodrigues and Vaz {319}. Further, these waves can carry information. However, the *technology* of employing these waves has been lacking {320}. The use of scalar EM wave interferometry to produce normal EM fields and EM field energy at a distance has been shown by Evans *et al.* {321}.

As pointed out by Rodrigues and Lu {318}, perfect longitudinal EM waves travel at infinite velocity. This eliminates the difference between potential, field, and wave; a perfect longitudinal EM wave can also be taken as an instantaneous scalar potential with net force-free iterative regauging (varying at each point instantly and simultaneously). Both ways of looking at it produce the same end result: instantaneous remote energy change of spacetime energy density itself, in general relativity terms, or of the energy density of the vacuum, in particle physics terms.

Elsewhere, (e.g., Chapter 2), we also pointed out the "total photon **interaction"** mechanism that generates the flow of an entity through time. We also have advanced the concept of a "time-reversal zone" (TRZ) as a region where antiphoton interactions with charges predominate over photon interactions. We used these new concepts to propose exact nuclear **reactions** for the production of excess deuterium, tritium, and alpha particles in cold fusion experiments. In such an "excited state" TRZ region, the law of attraction and repulsion of charges is reversed. In the TRZ, like charges attract and unlike charges repel. Further, the gluon forces are reduced, so the quarks in the nucleons are much more lightly bound. In such TRZ zones, dense clustering of like charges occurs due to their mutual attraction.

A time-reversal zone is created in one part of the Fogal transistor, greatly assisting in the "electron pinning" action. That is, the usual strong mutual repulsion (the Coulomb barrier) of the electrons at the pinning site vanishes in the TRZ inside the Fogal semiconductor, and the resulting mutual attraction of the electrons greatly facilitates the pinning process far beyond what is usually obtained against "like charge" repulsion in a time-forward zone. Indeed, the TRZ in the Fogal semiconductor also materially enhances the electron pinning process beyond what can usually be obtained with cryogenics. At least some indication of trapping and perhaps pinning has been reported in amorphous semiconductors at room temperature {322a-b}.

The Fogal device is able to remove EM noise (electron noise) from existing EM signals, displays, etc. by 96% or more, by pinning the electrons and simultaneously time-reversing the spatial EM signal back to its original noise-free state prior to hosts of interaction electron collisions. This "noise removal" capability has been rigorously tested by independent parties, including at a leading university in the Netherlands,[135] and by several large electronics companies and several other universities. By pinning the electrons (and electron currents) while still processing and propagating the EM energy effects, the Fogal device can move noiseless EM energy without loss at room temperatures, at superluminal velocities, and reconnect the lossless propagated energy to electrons past the pinning region at a receiving region, releasing the electrons in the receiving region. The Fogal semiconductor therefore already accomplishes *room temperature superconductivity* as well as tunneling.

The Fogal device's basic mechanism (consisting of all the submechanisms) is to restore and move EM energy in force-free form, using direct regauging of the vacuum instantaneous scalar potential itself (or the vacuum potentials themselves). It operates directly upon the longitudinal EM waves comprising the vacuum potential. Hence it directly engineers the active vacuum and curved spacetime. I.e., it is the first *supersystem engineering* device of (hopefully) many more to come.

---

[135] The university researchers verified room temperature superconductivity being produced by the Fogal semiconductor in their independent verification measurements. They also measured that some 96% of the noise already present in a circuit was removed by Fogal's semiconductors in Fogal's circuits.

In the "jitter" environment of the space in which the individual electron is embedded, there exist many (and essentially all) field gauges, simultaneously. One of these, of course, is the Coulomb gauge where there exists an instantaneous scalar potential. Usually the instantaneous scalar potential is considered only a *near-field* effect, with no contribution at any significant distance. However, as Jackson points out {323}, transverse radiation fields are given by the vector potential alone. Also, in the Coulomb gauge the scalar potential "propagates" instantaneously everywhere in space {324}.

Along with other electrodynamicists, Jackson avoids the obvious superluminal communications implications of the instantaneous scalar potential by stating that *"..it is the fields, not the potentials, that concern us."* As he points out, a detailed *(though conventional)* discussion of causality is given by Brill and Goodman {325}. Again as pointed out by Jackson, in the Coulomb gauge the quantization of the photon involves only the vector potential. *The use of the scalar potential in the Coulomb gauge, together with deliberate engineering of the internal longitudinal Whittaker wave structure of that potential, directly allows distant engineering via multiply connected spacetime, even inside the nuclei of distant objects. This is in fact a unique and marvelous engineering application and extension of Bohm's quantum potential concept.*

My "pinning" the electrons, apparently Fogal is able to essentially "hold" the vector potential and all the force fields while the instantaneous scalar potential *persists and now extends over all space, not just to the near field since all formerly moving field dynamics that would limit it have been frozen.* If this effect is substantiated, I have nominated it as the "Fogal effect". Further, this unrestricted instantaneous scalar potential now i lunges in field-free form (pure regauging form) over all space according to the signal energy being perpetuated and propagated[136] by the Fogal semiconductor independently of its pinned electrons. In this fashion, the Fogal semiconductor is indeed capable of "instantaneous" communication over essentially unlimited distances. At least one independent test has verified the Fogal instantaneous communication effect, but the test is still held proprietary by the testing agency.

---

[136] Here we again point out that the usual notion of a signal as energy propagating in 3-space is wrong. It propagates in 4-space, but not in 3-space. Also, what propagates is the spreading point-like polarization of the vacuum itself.

Since Fogal is able to alter and control the internal longitudinal EM wave Whittaker structuring of the scalar potential used in this Coulomb gauge manner, he is able to accomplish superluminal (and even instantaneous) communication, so long as a Fogal semiconductor is available in both the transmitter and the distant receiver. In the transmitter, the Fogal component is necessary to *infold* the information into the internal LW structure of the scalar potential to be used in the Coulomb gauge, thus transforming the input signal information into output longitudinal EM wave information directly in spacetime itself. The Fogal component in the receiver is necessary in order to detect, transduce, and *outfold* the infolded longitudinal EM wave information inside the instantaneous scalar potential. We have stressed that these internal longitudinal EM waves and their dynamics are spacetime curvatures and their dynamics. Hence their transmission and detection is a viable process for superluminal communication, including instantaneous communication.[137]

Further, once the electrons are pinned, all "signal fields in charged matter" (i.e., all EM force fields) are frozen and localized. This, we believe, involves a new physical effect introduced by Fogal that is analogous to the Aharonov-Bohm effect.[138] So I have called it the "Fogal effect". That is, the force-fields are frozen and localized by the pinning, but the field-free potentials then remain, including the instantaneous scalar potential ϕ and probably the curl-free magnetic vector potential A as well. The result is that the instantaneous scalar potential can be directly *regauged* everywhere at once by the changes of magnitude of the "frozen fields", rather as if they were "static voltage", possibly simultaneously regauging the field-free A-potential — all as part of a new kind of Lorentz symmetrical regauging. After all, a single Lorentz symmetrical regauging {326} changes both the potentials A and ϕ. Changes in symmetrical regauging produce stress waves, which are longitudinal EM waves. Continual changes of symmetrical regauging also continually changes both the potentials A and ϕ, *including over all space when operating in the Coulomb gauge.* This is one explanation of the beautiful effect that Fogal has discovered and is utilizing.

---

[137] We believe that physics books will have to be substantially rewritten, once the internal longitudinal structuring of the scalar potential (and of EM fields and waves) is employed directly in open technology. We further believe that the Fogal semiconductors are the first pre-developments in that coming vast new technology, at least in the open Western scientific world.

[138] There have long been, and still are, efforts to "fit" the Aharonov-Bohm effect into slightly modified classical electrodynamics after all. E.g., see C. J. Carpenter, "Electromagnetic energy changes due to charges moving through constant, or zero, magnetic field," *IEE Proc. A*, 138(1), Jan. 1991, p. 55-70; — "Electromagnetic energy and power in terms of charges and potentials instead of fields," *IEE Proc. A*, 136(3), 1989, p. 101-113; G. M. Graham and D. G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo," *Nature*, Vol. 285, 1980, p. 154-155; T. H. Boyer, "Semiclassical explanation of the Matteucci-Pozzi and Aharonov-Bohm phase shifts," *Found. Phys.*, 32(1), Jan. 2002, p. 41-49; — "The Aharonov-Bohm effect as a classical electromagnetic-lag effect: An electrostatic analogue and possible experimental test," *Nuovo Cimento*, Vol. 100B, 1987, p. 685-701; — Classical electromagnetic deflections and lag effects associated with quantum interference pattern shifts: considerations related to the Aharonov-Bohm effect," *Phys. Rev. D*, Vol. 8, 1973, p. 1679-1693; — "Does the Aharonov-Bohm effect

We accent that, with pinned electrons frozen in place, the normal ϕ and A in matter (defined in terms of interaction with charge and the movement of charge) cannot occur. With time reversal zone utilized, electron pinning is greatly eased and the instantaneous scalar potential ϕ now assumes the primary active propagating role in surrounding space. Neither ϕ nor A is free to develop the normal $\partial\phi/\partial t$ and $\nabla\times A$ *spatially propagating dynamics of the force fields (effect fields) observed in matter.* In the Fogal process, the instantaneous scalar potential "signal" (the causal signal in 4-space) is generated in the transistor when it pins the electrons and freezes the force fields into "equal and opposite" form. Since no net force field can develop, the instantaneous scalar potential is a stress potential in spacetime itself and simply changes its magnitude instantly, at each and every point in space that it occupies, when the Fogal chip's input EM fields change in its signal input.

In the output, the net force fields remain zero and frozen, even though the individual force fields change constrained to "equal and opposite" manner. If one wishes, one may take this part of the Fogal output — the field free changes in the stress energy density of the vacuum — as one expression of *modululating the instantaneous scalar potential's energy density in perfect correspondence to the field changes in the normal EM input to the Fogal transistor.*

exist?". *Found. Phys.,* Vol. 30, 2000, p. 893-905; — "Classical electromagnetism and **the** Aharonov-Bohm phase shift," *ibid.,* p. 907-932.

The instantaneous Fogal changes in magnitude (e.g., instantaneously induced upon a distant collecting charge) are no longer limited to near field effects, or to the speed of light — and communication is not bound to the retarded wave. Now the induced changes of the instantaneous scalar potential exhibit far field effects and these changes — being naught but these special Lorentz condition magnitude regaugings — are instantaneous. This of course is a form of *action at a distance.* We stress that the best astronomical evidence also points to a startlingly superluminal speed of gravitation (curvature of spacetime, which after all is just a longitudinal EM wave or set of them) itself, as pointed out by Van Flandern.[139] If so, since any change of energy in spacetime is a curvature of spacetime, and one form of spacetime curvature can move at superluminal velocity, so can other forms of spacetime curvatures. In our favorite expression, *it only takes one white crow to prove that not all crows are black.*

As is well known, the de Broglie waves of matter always move faster than light, as do appropriate quantum tunneling processes. Further, the quantum potential used by Bohm in his beautiful hidden variable theory of quantum mechanics {347} also appears instantly throughout space, wherever it will be, and with the value it will have. The question of whether a quantum potential can be used to communicate has been posed but not definitively answered in the previous literature. We propose a resounding "Yes!" if one uses the internal structure of the quantum potential for one's communication, applying some component such as the Fogal semiconductor to do the infolding and outfolding.

We also propose that a process for forming and using an instantaneous scalar potential in the Coulomb gauge can be interpreted as the formation and use of a quantum potential or its direct equivalent. If that assumption is true, then the Fogal semiconductor seems to be the first component ever openly built and patented which produces and uses Bohm's quantum potential, *including* its infolded Whittaker structuring and dynamics. It also definitively answers in the affirmative the question of whether the quantum potential and the scalar potential in the Coulomb gauge can be used in superluminal communications.

---

[139] Tom Van Flandern, "The speed of gravity - What the experiments say," *Phys. Lett. A,* vol. 250, Dec. 21, 1998, p. 1-11. Van Flandern gives an extremely important, clear, and succinct summary showing that experiments refute the standard far field assumption of light speed in gravitational effects. Laboratory, solar system, and astrophysical experiments for the "speed of gravity" yield a lower limit of $2 \times 10^{10}c$.

Thus Fogal has apparently found a way to directly modulate intelligent change of magnitudes into and upon the internal longitudinal EM wave structure of the instantaneous scalar potential (and upon the quantum potential), as well as iterative "pure *net-field-free"* regauging of the instantaneous scalar potential {327}. It seems to have previously escaped notice that varying the Lorentz symmetrical regauging of a circuit or system easily produces pure longitudinal EM waves. This strongly implies that Fogal has surmounted and removed the barrier to practical superluminal communication and even further has reduced the barrier to practical instant communication.

Several tests — in one of which the present author participated — have indeed shown Fogal transmission of signals, including video signals, between two widely separated points with less delay than light-speed signals exhibit, and also with video bandwidth infolded inside a DC signal. In at least one series of tests in a large company's own laboratory and by its own engineers, communications from a ground-based Fogal chip transmitter apparatus occurred between the ground station, through the uplink into the satellite, back through the downlink to earth, into a ground-based transmission system with repeater stations, and into the receiving Fogal transistor unit in the lab. The transmissions occurred with *no discernible time delay,* something previously deemed impossible in all the textbooks. This was particularly significant since even the conventional time-delays in the repeater station amplifiers and the satellite receiver-transmitters were apparently eliminated. This further accents that the infolded "longitudinal EM waves superhighway" inside all EM fields, potentials, and waves — including inside matter such as the ocean and the earth itself— is a multiply connected space phenomenon. It is the most ideal communication medium, enabling deep space communication without delay, etc.

The implications for nondelayed future remote control of robot spacecraft and robots landing on the surface of planets and asteroids are obvious, as are the implications for direct exploration and measurement of conditions *inside the interiors* of the planets, the sun, the moon, and the earth via **internal** scalar interferometry. We accent that Fogal has done private experiments exploring this vast new physical reality revealed by the "internal longitudinal EM wave dynamics" and its interferometry, and it is mind-expanding (some would say *mind-boggling!).* It is also a great new step in physics and in our very notion of physical reality itself.

T**he** vacuum itself, being highly energetic, is indeed a scalar potential. **Further,** it has any and all gauges. Since potentials superpose, we postulate that Fogal's use of the instantaneous scalar potential is directly superposed on the corresponding instantaneous Coulomb gauge scalar potentials of the entire vacuum. His discovery of how to modulate the instantaneous ϕ magnitude without allowing modulated EM field propagation, constitutes a very special case of tunneling right through "space" itself, using one aspect or "component" of space.

In short, to be a little dramatic one might even apply the term *subspace communication* to this long distance tunneling effect. That interpretation would also seem consistent with the deterministic variation of the internal stress energy density structuring of the vacuum that is produced in the Fogal process.

From another viewpoint, Fogal has succeeded in not only finding the missing chaos of quantum mechanics, but in deterministically engineering and using it.

This is also consistent with Whittaker's decomposition of the scalar potential into bidirectional longitudinal EM waves {328} and his connection between time-polarized EM waves and longitudinally polarized EM waves as reinterpreted (12) to be consistent with quantum field theory {19}. We point out that a single point in time is simultaneously "connected" to every point in the universe. By using a transmission system that back-pedals from 3-space into the time domain, an instantaneous connection to any other spatial point in the universe is therefore available, if one's technology can take advantage of it. Fogal's technology does.

It is also consistent with quantum tunneling experiments that have demonstrated the transmission of intelligent signals (such as Mozart's 40th symphony) at superluminal speed between two points in a waveguide {329a-b}. Regardless of how one chooses to model intelligent superluminal communication, it has been experimentally demonstrated — multiple times and by multiple experimenters in multiple laboratories — that such can occur.

The difference in those superluminal experiments and the performance of the Fogal chip is that the chip uses iterative regauging of the magnitude of the instantaneous scalar potential, as well as deterministic alteration of the scalar potential's internal longitudinal EM wave substructuring, and probably also modulates the internal structuring of a quantum potential. The presently proposed quantum tunneling mechanisms apparently do not do this, at least as they are presently modeled. Further, the quantum tunneling mechanism is difficult to apply; the Fogal mechanism is fairly simple and straightforward to apply, given the Fogal semiconductor at both the transmitting and receiving ends of the communication system.

Pure longitudinal EM waves readily penetrate matter to great distances, since matter is mostly empty space between widely separated mass particles. On the microscopic scale, matter is comparable to something like the solar system, except that intense fields and potentials (bundles of longitudinal EM waves and their dynamics) occupy the space between the particles. All this vast empty space inside matter, being a potential and comprised of longitudinal EM waves, is a giant "superhighway" for the passage of longitudinal EM waves and their dynamics, as indeed is matter itself which is mostly such empty space. Further, even the masses themselves are such, since mass is simply highly compressed spatial energy density, and therefore is highly compressed longitudinal EM wave energy and dynamics. It is just a "denser" superhighway.

Hence a longitudinal EM wave system can readily be developed to "see through matter" using Fogal semiconductors in transmitter and receiver. Fogal in fact has made at least one proprietary prototype demonstration unit that demonstrates this startling capability. The vision is very clear, due to the remarkable absence of noise in the Fogal process. It is also shockingly unique, and it changes our very notion of physical reality. We are not at liberty to further discuss this new "sub-submicroscopic" reality revealed in such experiments. We simply interpret it as jumping the presently assumed restrictions of observation, and allowing the direct or correlated observation of the virtual state entities.

As alluded to, the previously modulated "Fogal type signals" emitted by all matter and other physical entities can also be displayed on a screen using the Fogal process, revealing a hitherto unsuspected physical reality and EM signals environment. In this prototype system, Fogal showed a remarkable and startling new extension of our very perception of physical reality itself. This is another reason why we expect that physics books will have to be rewritten, once the Fogal semiconductors are on the market and available in practical equipment. As every physicist knows, any dramatic increase in our ability to sense finer levels of physical reality inevitably leads to new and startling advances in physics itself and changes our very knowledge of the nature of physical reality.

The ability to bypass and "park" force fields by pinning their source electrons in a time-reversal zone, while continuing the flow of noise-free EM energy at will and at infinite velocity, obviously enables COP>1.0 systems. A pure longitudinal EM wave already has infinite energy, for such a wave appears across the entire universe simultaneously, instantly changing the stress energy density of the entire vacuum in which the longitudinal EM wave is present.[140] The back emf and back mmf fields in an EM system or electrical power system can be frozen and held from restoring the Lorentz condition in the circuit, using the Fogal semiconductor {330}. We presently have an agreement with Fogal for the licensed use of his semiconductors in all such energy unit applications.

A quantum potential is modeled in a multiply connected spacetime, so that a given joule of input energy to one multiply connected point simultaneously appears in every other multiply connected point.[141] Since the Fogal semiconductor apparently can operate in quantum potential or Coulomb gauge fashion, it can act as a *direct energy amplifier,* with respect to the lab observer in a singly connected space. In short, the Fogal process dramatically extends the concept of "conservation of energy" to include conservation in multiply connected space and therefore incorporates observed energy amplification in singly connected space. Again, a broken symmetry (in singly connected space) reveals a higher symmetry (in multiply connected space), and the hierarchy principle is alive and well. In short, with the Fogal semiconductor used in power systems, the law of conservation of energy must be extended to include energy symmetry in multiply connected space but asymmetry in singly connected space. Obviously, such functioning totally transcends the limitations of the present thermodynamics models.

We carefully point out that this is not "energy creation" but simply the energy following the known kind of statistics in physics that allows a thing to be in multiple different places simultaneously. What is actually "instantly transmitted" is the simultaneous regauging of the energy density of the vacuum. At any point in space, the vacuum has an energy density and thus is a scalar potential with a given magnitude. Changing the energy density is simply regauging, which is freely permitted under the gauge freedom axiom of quantum field theory. So instant amplification of energy" over a region of points is simply the instant regauging of the potential already present at each point in the region. It is *a permissible multiply connected space violation* of the present more limited "special case" energy conservation concept, which implicitly assumes a singly connected space. We have been working with Fogal for some time in this area, and we expect to rapidly develop and introduce commercial EM power systems based on the Fogal semiconductor's abilities in multiply connected space, as soon as it is in production and available. Until then, the details of our work in that area must remain highly proprietary.

---

[140] This should not be too surprising, since the bare charge and the bare mass of every electron each has infinite energy, as is well known in particle physics. For confirmation in very straightforward language, see Nobelist Steven Weinberg, *Dreams of a Final Theory,* Vintage Books, Random House, 1993, p. 109-110.

[141] If we wish, we can even take time itself as a special kind of quantum potential, since time also displays such a "multiplicity of connection" in 3-space. In one sense, time is the ultimate quantum potential, since a single point in time is "connected" to every single 3-space point in the universe, simultaneously.

5.8.4 Point-Contact Transistor.

Figure 5-8 diagrammatically shows a point-contact transistor with n-type base in a typical circuit for power gain. The point of the contact is under pressure on the material with which it is in contact.

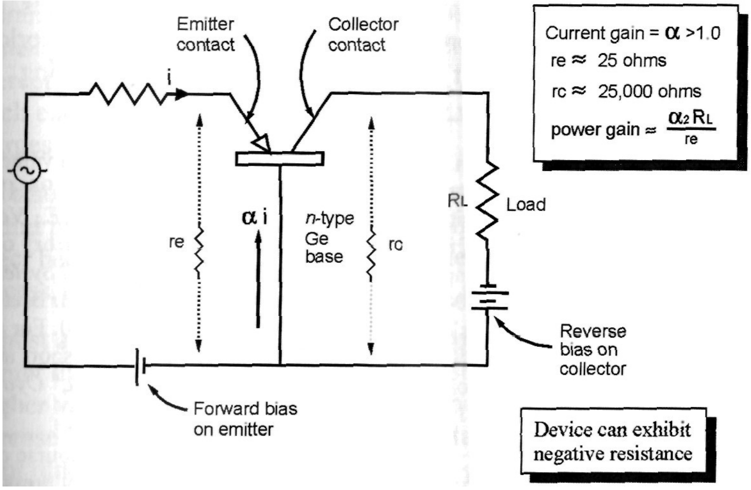
Emitter contact
Collector contact
i
Current gain = α >1.0
re ≈ 25 ohms
rc ≈ 25,000 ohms
power gain ≈ α2 RL / re
RL
Load
α i
n-type Ge base
re
rc
Reverse bias on collector
Forward bias on emitter
Device can exhibit negative resistance

Figure 5-8 Point contact transistor with n-type base.

The point-contact transistor often behaved in true negative resistor fashion, outputting more energy than was input to it. Its production was always far **more** of an art than a science.

As can be seen from our quotation at the beginning of this chapter, the original point-contact transistor was never thoroughly understood technically.[142] Specifically, there does not appear to have ever been any deliberate attempt to capitalize on the ability of the transistor to become a true negative resistor under certain manufacturing techniques and under certain conditions.[143] Instead, the variations and difficulties in manufacture resulted in use of the point-contact transistor being essentially bypassed rather quickly, by advancing to other transistor types more easily manufactured and with less manufacturing variances. In reducing the manufacture of other types of transistors to a science rather than an art, the point-contact transistor and its further development and optimization into reliable true negative resistors were abandoned.

We present this transistor as a straightforward and relatively inexpensive area in which young researchers in COP>1.0 systems and phenomenology may wish to begin, assuming they have access to semiconductor facilities in a university or similar or are willing to work meticulously under a jeweler's loupe. A rich combination of effects in the transistor awaits optimization and control. We predict that we shall yet see, on the world market of the future, such point contact transistors reliably exhibiting negative resistance. It will be necessary, however, to also take into account the symmetrical self-regauging characteristics of the common closed-current-loop circuit. By adroit switching of a true negative resistor in and out of a closed-current-loop branch, or placing it in parallel with the back emf of that branch to reduce it, asymmetry can be introduced into the overall closed current loop circuit, defeating the Lorentz symmetrical regauging. Defeating the Lorentz condition is essential to COP>1.0, of course.

---

[142] E.g., different theories are given in: (a) W. Shockley, "Theories of high values of alpha for collector contacts on germanium," *Phys. Rev.,* Vol. 78, 1950, p. 294; (b) W. R. Sittner, "Current multiplication in the Type A transistor," *Proc. I.R.E.,* Vol. 40, Apr. 1952, p. 448-454. Also of interest is (c) W. van Roosbroeck, "Theory of the flow of electrons and holes in germanium and other semiconductors," *Bell System Tech. J.,* Vol. 29, Oct. 1950, p. 560-607. See also (d) J. Bardeen and W. H. Brattain, "The transistor, a semiconductor triode," *Phys. Rev.,* Vol. 74, 1948, p. 230. For a more modern re-examination, see Shuji Hasegawa *et at,* "Electronic transport at semiconductor surfaces — from point-contact transistor to multi-tip STM," *Oyo Buturi,* 70(10), 2001, p. 1165-1171 (in Japanese).

[143] E.g., the point contact is usually under pressure, and this pressure of course can be varied. The full phenomenology of points (which increase voltage) and pressure, complicated by surface effects as well, has not been worked out in physics. Note the similarity of the pressure in the point contact transistor to the pressure in the Chung negative resistor. Note the almost certain involvement of the "overpotential" of chemistry and electrode chemistry, as well as the "double surface" effect of the small gap between the point and the substrate on which it rests in contact. Part of the gap probably even involves the Casimir effect as well. Any electrical signal variation in that very complex point junction will vary the overpotential, the stress potential, the "point increase in voltage" effect, etc. As can be seen, the phenomenology of the point contact is remarkably complex and rich in several areas of physics and in a great variety of physics variables. Obviously, the reasonable control of all these highly nonlinear variables — and their mutual interactions — is a difficult matter. Hence the readiness with which point contact transistors were dropped with substantial sighs of relief.

We envision such a development as an ideal "negative resistance shunt" to add across the secondary of a transformer, and another possibly across its primary and across the external power supply as well, in producing more amenable and easily fabricated COP>1.0 electrical systems. We also point out that, once the back emf or back mmf forcing of equal power dissipation in the primary of a transformer or in the emf section of a circuit is dramatically reduced or eliminated, one does not require "large intensity" potentials and voltages to collect a great deal of power in the intercepting and receiving external circuit. Again, by simple W = Vq, as much energy W can be collected from any nonzero potential V as there are charges q to intercept.

We also envision such a true negative resistor being close-looped by the Bedini process, thus producing a small "self-powering" transistor, which in effect becomes a small self-powering "battery". This is absolutely a doable process, and it will be done once the Fogal semiconductor is in production.

Almost all semiconductor materials are also optically active materials, and a point discharge into such materials represents a very sharp regauging (higher voltage) discharge at a point or into a very small area, due to the increase in potential at the tip where it contacts the base material. The point junction is under pressure, so a stress potential exists there. In addition, the well-known "point" effect also increases the potential in the junction pointer itself, from its base to the point. The point contact phenomenology of different materials — one conductive and one semiconductive — is of much interest, and with novel phenomenology.

The fact that the point-contact transistor in its most usual formulation primarily uses holes more than electrons, is also of much interest in COP>1.0 situations. Holes in a circuit move against the voltage. The trick is to let the Dirac holes *before observation* move in an open path from the ground return line against the back emf to the potentialization line, and **transduce** the moving hole current (via the Bedini process) into electron current after the holes have already freely reached the vicinity of the high side of the circuit.

When Dirac sea holes (causal positrons before observation and thus prior to their alteration to lattice holes and parity reversal) and Dirac sea hole current are also involved with point-contact transistors, the resulting phenomenology has been but little investigated and none of it is in the present textbooks. However, from recent work with positron probes and positron microscopes to examine semiconductors and semiconductor materials, it is known that the holes (positrons) tend to be repelled from nuclei in the material into defects (voids) in it {331}. These effects have been investigated, for example by Triftshauser *et al.* {332}.

Both the mechanical stress potential (which is fundamentally electromagnetic) and the heightened junction potential decompose via our reinterpretation of Whittaker's 1903 decomposition of the scalar potential. This leads to optical-type pumping in both the time domain as well as the 3-space domain. Hence novel optical-type effects and time-reversal of material states can be involved, leading to a very complex set of phase conjugate phenomena, time-reversal phenomena, etc.

Certain Hall effects employed in conjunction with a point contact transistor could be a fruitful area of investigation. As an example, narrow Hall bars with junctions between current and voltage leads of various geometries could be investigated. Widening the junction from the normal square-cornered shape can sometimes produce a negative Hall resistance.

The junction of the point contact involves asymmetrical self-regauging, iterative time-reversal retroreflection, increased Poynting and Heaviside energy flow components, optical scattering processes inside the junction materials, etc. The transistor can indeed be manufactured so that these highly nonlinear effects sum to a negative-resistor-like movement of the output current against the voltage, although with so many other phenomena involved it will require some hard work and research in order to develop and stabilize it — and understand it.

5.8.5 Bedini's Negative Resistor Process in a Battery {333}.

5.8.5.1 Some Pertinent Lead-Acid Battery Processes

A lead-acid storage battery is a highly nonlinear device. The plates may be plante, pasted, or tubular etc. They are usually highly porous and have a thin oxide layer. They may have other coatings and the electrolyte solution may include additives to enhance operation. The plates consist of a mix of crystalline materials in most cases. The plate materials include acicular (needle-like) crystals, particularly the negative plate. The plates usually are porous, knit, and have gridded separators.

The chemistry of the battery electrolyte is complex and — even today — not all of it is well understood, particularly when mass transport is accounted. In our discussions, we will be including some of the previously neglected phenomena during mass transport (ion sluggishness in moving) that are responsible for some of that lack of understanding. In the battery chemistry, there are many double surfaces between the ions and also there is a variety of point voltages. There is a separate theory for double surfaces.

One of the most important phenomena is that the ion currents in the electrolyte are confined to the battery, and do not pass out into the external circuit. Much of the electron currents in the external circuit are blocked at the plates from moving within the electrolyte. Consequently, we will limit our discussion to two major currents of interest[144] in a battery-powered system, and these currents are effectively isolated from each other (at least to first order). Those currents are (i) the ion currents in the electrolyte, confined to between the plates and therefore internal to the battery electrolyte, and (ii) the electron currents between the outside of each plate through the inside of the plate and on out through the external circuit and **then** around to the inside and thence to the outside of the other plate. Chemical changes occur on the plates themselves as a result of these two currents.

It is usually assumed in power systems that these two currents are in phase or almost entirely so. That need not be true at all, and Bedini takes specific advantage of deliberately arranging these two currents to be antiphased part of the time. Part of the Bedini effect can be understood from this fact alone; the other part must consider one other most unusual electrical current that is encountered only in battery-powered systems or COP»1.0 systems.

So indeed there are *several* (many) currents in a lead acid battery [334], not just one, and we will consider three of them — the two previously mentioned, and an additional Dirac sea hole current in the local vacuum **itself.** The reader interested in greater technical depth should also turn to the literature of electrode chemistry and electrochemistry {335a-f}, which is quite complex.

---

[144] Of course, there are many other currents in a battery as well, but — although an oversimplification — these two major currents together with the Dirac sea hole current will suffice for our introductory modeling and explanation purposes.

5.8.5.2 Dirac Sea Hole Current

In a battery-powered COP»1.0 system (and in some COP>1.0 systems), there is a previously unrecognized significant flow of Dirac hole current in the local vacuum, running from the output section of the system to the system input section and thence through the conductors into the battery terminals and onto the battery plates, and on into the battery electrolyte chemistry. This hole current may be greater — *evenfar* greater — than the normal electron current running from input to output.

So in the battery-powered overunity system with positive feedback, not only does an electron current run from the battery through the connecting conductors to the input section of the system, but also a Dirac hole current pours back out of the system input section, and — if not intercepted and transduced — through the conductors and back into the battery. All the while, this Dirac hole current is "eating" electrons being furnished by the battery to the system. The net result is that the battery "sees" an extra load to be powered, in addition to the ordinary load presented to it by the normal system input section. It "sees" the situation as an "extra load" because it first has to furnish sufficient electron current to fill the hole current and "kill" it, before it can furnish the additional electrons to power the system in normal fashion.[145]

---

[145] An odd effect occurs in the battery, however, when the battery completely discharges in the normal sense. Actually, it is still fully charged with positive energy charge, but also has been simultaneously charged with negative energy charge. With the excess negative energy holes continuing to charge the battery with negative energy charge, the battery becomes increasingly *charged with net negative energy.* There is almost no limit to the negative energy charge the battery (i.e., the local Dirac vacuum portion of its supersystem) can take, except the eventual development of antigravity effects due to the inverse and increasing curvature of local spacetime in which the battery is embedded. After a battery has been used to power a COP»1.0 system for some time, one can take the battery off, place it on a normal battery charger, and it will then "eat" positive power for an extended period of time — e.g., a week or longer. During this period, one is steadily negating the inverse curvature of the local vacuum via the increased positive energy collecting in that spacetime. The battery again finally passes through zero voltage and net charge again, and suddenly starts charging up again in a "normal" sense with positive energy charge. This process actually increases the lifetime and function of a battery in a startling manner. Bedini has used batteries almost devoid of charge and energy in the normal sense, and powered systems very well with them, even for longer than a year continuously.

A COP»1.0 EM system is an excited system far from equilibrium in its exchange with its active environment. For all such excited systems, there exist significant decay mechanisms to bring the system back into equilibrium with its environment, since equilibrium yields the lowest energy state and maximum entropy. The Dirac sea hole current (negative energy current) is nature's unexpected natural decay mechanism to sharply decay a COP»1.0 system back to an overall COP<1.0 system when the external power supply is considered also.[146]

So when Dirac hole current back through the local vacuum and through the system is involved, the battery first has to furnish electron current to fill these holes at the rate they are appearing in the input section, thereby furnishing a primary "hole-filling" electron current component which "disappears" into the Dirac vacuum and is lost.[147] Only after the holes in the hole current are filled — and the hole current is being steadily nullified

---

[146] It decays the *true negative resistor* power system back into being only a *differential negative resistor section* of a slightly extended metasystem including the power system and its external power supply. This novel decay mechanism also prohibits close-looping the now differential negative resistance power system for self-powering, unless the hole current is first converted into electron current, as discussed in Chapter 9 and in paragraph 4.9.6 below.

[147] The filling of these Dirac sea holes is not pair annihilation in the usual sense, and no radiation is emitted. Instead, the energy of the radiation that would otherwise be emitted is added to the negative energy curving the spacetime. Hence the net curvature of spacetime relaxes because of superposition. In effect, the vacuum/spacetime itself reabsorbs the energy as relaxation spatial energy added to the curvature energy of the curved spacetime, reducing that curvature. Particularly see Mendel Sachs, *Quantum Mechanics from General Relativity: An Approximation for a Theory of Inertia,* Reidel (now Kluwer), 1986. A generalization of quantum mechanics is demonstrated in the context of general relativity, following from a generally covariant field theory of inertia. Nonrelativistically, the formalism corresponds with linear quantum mechanics. In the limit of special relativity, nonlinearity remains and several new features are derived: (i) Particle-antiparticle pairs do not annihilate; an exact bound state solution is derived corresponding with all experimental facts about annihilation/creation — which, in approximation, gives the blackbody radiation spectrum for a sea of such pairs, (ii) A result is proven, without approximation, that is physically equivalent to the Pauli exclusion principle- which in linear approximation gives the totally antisymmetrised main-body wave function and Fermi-Dirac statistics, (iii) The hydrogen spectrum is derived, including the Lamb shifts, in agreement with experiment; new results are found for high-energy electron-proton scattering, (iv) Finally, several applications to the elementary particle domain are demonstrated, in agreement with results from experimental high-energy physics.

— can the battery then send additional electrons which do not disappear and which are accepted in the system input section to power the system. We discuss this Dirac hole current in some detail in Chapter 9, including how to convert it into useful input electron current to the system, reducing the current load on the external generator and enabling self-powering. Bedini invented the method of transducing available negative energy hole current in the vacuum into electron positive energy flow to charge the battery and power the system, preventing decay of the COP»1.0 system back to COP<1.0.

We shall not discuss the complexity of the additional chemical reactions ongoing in the battery electrolyte and on the plates. The two main mass-containing currents and the negative energy Dirac hole current will suffice to explain the Bedini negative resistor formation and usage, at least to first order and sufficient to stabilize systems.

5.8.5.3 Changing Symmetrical to Asymmetrical Regauging

To first order, we reiterate that the electrons flow from one plate of the battery through the external circuit to the other plate, but the lead ions do not. There is thus a sharp "two separate currents" interface at the plates and thus a sharp separation between the electron current and the ion current. In short, there are actually two half-circuits in the battery-powered system if one chooses to use them semi-independently.

The current carriers in the two half-circuits are quite different. Therein lies the opportunity to separate and antiphase the two currents, using one to recharge the battery while the other is used to power the loads. It also allows a good opportunity to asymmetrically regauge both half circuits, greatly and freely increasing the Poynting energy component intercepted, collected, and used by each half-circuit's carriers. This is the basis for a profound "opening" of the system and influx of excess vacuum energy.

The two half circuits meeting at a common interface allow a stress potential upon that interface to *individually* use the two halves of what is normally Lorentz's symmetrical regauging. That usually *symmetrical* regauging (if the complete circuit were unitary) is now comprised of two separated *asymmetrical* regaugings, one for each of the two circuit halves, because the net stress potential (net Lorentz symmetrical regauging) — consisting of two opposing fields and forces — is split into two now-individually-separated and separately utilized force fields.

We accent this process by an analogy. Suppose we suddenly place a scalar potential ϕ upon a point in the middle of a transmission line. Immediately the potential ϕ races off (spreads) in both directions simultaneously, at nearly light speed. Further, in each direction there is a moving gradient of ϕ, hence an E-field by $\mathbf{E} = -\nabla\phi$. These E-fields or "emfs" are oriented in opposite directions, so there is a movement of charge from the middle of the line toward each of the ends. The same "dephasing" current effect can be applied to any conductive medium or to the interface between two conductive circuits.

In Bedini's negative resistor case, the placement of the potential ϕ — for its movement in both directions — is on the interface (the face of the lead plates) between the two half circuits and therefore between the two different current carrier types. An initial "Lenz's law" reacting with the suddenly placed E-field onto the plate surface, creates an E-field in opposition, forming a stress potential ϕ which is much higher than the normal potential of the battery (in a nominal case, 100 volts instead of the 12 volts of the battery).

Immediately this potential "takes off in both directions", with a powerful E-field and emf back into the outside circuit and its load, while a powerful E-field and emf also proceed into the battery electrolyte in the opposite direction. So the ions in the battery have a sudden, much higher energy interception and collection forced upon them, with the excess emf now being in *battery-charging mode.* The electrons in the outside circuit simultaneously have an excess emf and potentialization *in the load-powering mode.* Let us see how the two very different current carriers react to these "dual emfs" in the overall circuit.

5.8.5.4 Mass-to-Charge Ratios of the Two Current-Carrier Types

We stress the dramatic difference in the *mass-to-charge (m/q) ratio* of the two currents. The lead ions' m/q ratio is some few hundred thousand times the m/q ratio of electrons, as we recall. For our purposes, all we need to know is that the m/q ratio for the lead ions is *very much larger* than the m/q ratio for the electrons. For a given force, the ions in the internal half circuit in the electrolyte will therefore respond very much slower than will the electrons in the external circuit half.

5.8.5.5 Hysteresis Between the Responses of the Two Half-Circuits

Because of this great disparity in the m/q ratios of the two mass currents, there is obviously a relatively significant hysteresis (time delay) between the *response* of the more massive ion current and the *response* of the far **less** massive electron currents that interact at the plates to try to change the ion current and its momentum. This time-delay between ion response in the electrolyte (the effect) and electron pileup urging (the cause) can be adroitly manipulated. Specifically, it can be used to alter the local vacuum potential and cause it to add excess energy to the ions in charging mode **and** simultaneously to add excess EMF to the electrons in circuit-powering mode. In short, the response delay can be manupulated to "asymmetrically regauge" the mass currents in the system, freely changing their potential energy, and dramatically increasing how much potential energy is available for battery recharging and how much is available for powering the external circuit (loads and losses). Further, during the antiphase condition of the two current halves, the battery can be charged simultaneously while the external circuit half and load is being powered. Excess energy is fed into both actions by the *overpotential* (the stress potential) created at the surface of the plates.

5.8.5.6 Evoking the Initial Bedini Negative Resistor Effect

Figure 5-9 shows the first phase of Bedini's negative resistor process. Consider the battery in normal load-powering operation. From the external circuit, a very sharp leading edge rise of a pulse of electrons and potential is sent to the battery plates in "back-popping" or "battery charging" mode. The back-popping electrons drive in nearly instantly, piling up on the negative plate[148] and trying to force the heavy ions to start moving in the battery recharging direction. The pile-up that results in the stopped electrons on the battery plate surface represents an increase in local current density, hence an increased potential. This produces the 100 volts potential, during the time that the ions have not yet responded to any appreciable degree.

Due to much larger m/q ratio of the sluggish for a moment they lag due to their greater inertia, and this allows the much more agile electrons to "pile-up", producing a substantial overpotential as the local current density increases. The lagging ions are being steadily overpotentialized during this "lag time", up to about 100 volts in a normal 12-volt battery. At the same time, electrons are being forced back out of that higher 100-volt area and into the external circuit and its load, which had been at 12 volts. Recall now that these electrons can move longitudinally only at the drift velocity. Hence all electrons in the external circuit are now highly overpotentialized, and in *load-powering mode. The* overpotentialized electrons are thus producing excess power in the external circuit, much more than normal.

---

[148] We are not using conventional "positive current and positive plate" as the high side of the potential. We are using "electron current and the negative plate" as the driving potential plate.

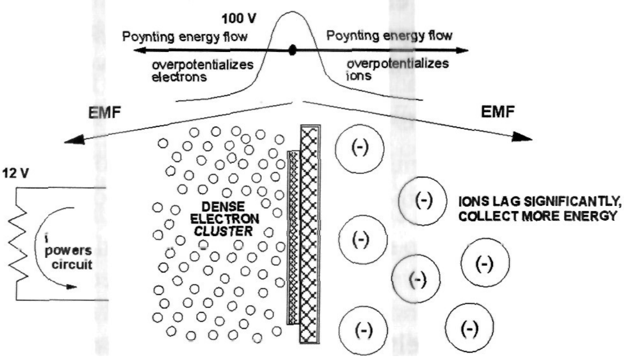

1 Sharp leading edge voltage pulse produces intense electron clustering and large potential (negative resistor) *at* interface *Sluggish ions only beginning to move*

2 Bidirectional Poynting energy flow from potential overpotentializes both the electron cluster and the ions, which are only beginning to move in charge mode

4 Large back emf Into circuit returns overpotentialized electron current, powering circuit with excess energy pouring from negative resistor potential Continues during ion delay

5 Simultaneously overpotentialized ion current is overcharging battery

Figure 5-9 Phase I of the Bedim negative resistor process in a storage battery.

Then the overpotentialized ions very slowly (compared to the electrons!) slow, stop, and begin to move in the opposite direction. They reluctantly respond and move in *battery recharging mode.* During that ion-response lag time, and the slower initial portion of the response, the electrons in the pulse continue to furiously surge in and pile-up on the negative plate, overpotentializing both the ions and the external circuit's electrons, while also some of them are being impelled back out into the external circuit to power it with extra overpotentialized energy dissipation. The charge density at that plate sharply increases due to the pile-up where the charges are "squeezing" together (clustering). There is a much higher potential suddenly rising in the squeezed charge cluster, because of the increased charge density there. As we stated, this potential nominally may be about 100 volts during this initial phase.

We call attention to the simple equation W = VQ, where W is the potential energy added to charges Q exposed to voltage V.

Simplified, the excess energy W_{IONS} freely impressed upon the ions is

$$W_{IONS} = f(100 - 12)(Q_{IONS}) \qquad [4\text{-}3]$$

where Q is the total coulombs of charge of overpotentialized ions and f(100 -12) is the magnitude of the overpotentialization of the ions.

At the same time, an excess energy is impressed upon the electrons in the external circuit by the same potential extending along the conductors into the external circuit. Again, this excess energy WE is given by

$$W_E = f(100 - 12)(Q_E) \qquad [4\text{-}4]$$

where Q_E is the total coulombs of charge of the overpotentialized electrons and f(100 -12) is the magnitude of the overpotentialization of the electrons.

But the emf on the electrons from the pile-up is directed in load-powering direction, as can be seen. Consequently the ion current and the electron current have been deliberately dephased by 180°, and so the overpotential energy of the ions is delivered in battery-recharging mode, while the overpotential energy of the electrons is delivered in circuit load powering mode.

Recapitulating: Pulsed pile-up of excess electrons on the negative plate interface between the two currents, while the ions are beginning to respond or only sluggishly responding, produces a much higher potential (an *overpotential)* on the sluggish ions — nominally some 100 volts in a 12-volt battery. On the negative plate, momentarily there is now a much higher voltage (with respect to the positive plate) than normally exists in the 12-volt battery. This voltage *overpotentializes* both the reluctant charging ions in the battery solution between the plates, and the powering electrons back into the circuit in powering mode due to the reversal of the emf. Since there is, say, 100 volts across the battery momentarily, there is also 100 volts now across the external circuit momentarily. Accordingly, overpotential excess powering of the external circuit load is suddenly evoked, while at the same time overpotential recharging of the battery is also evoked.[149]

In short, the ion current in the battery and the electron current into the external circuit have been placed 180° out of phase, achieving one major requirement for a COP>1.0 electrical system: *violating the integrity of the closed current loop circuit.* The battery is recharging at the same time that the external circuit is being powered, *from the same free overpotential,* and both recharging and circuit powering are driven by increased emf.

---

[149] Another of Bedini's innovations is to shunt the excess voltage (say, above 14 volts) into an external capacitor on the circuit side. In that way with his overpotential he can be (i) overcharging the battery, (ii) powering the load, and (iii) storing excess energy in that capacitor, from the altered vacuum — all simultaneously.

We strongly accent that the overpotential at the plates represents a change in the local vacuum potential, and it *identically is* part of that now-altered local vacuum potential. The vacuum, since it contains enormous EM energy in virtual state, is a very powerful EM potential. Any EM potential in our circuits is automatically a change to the ambient vacuum potential, or a change to another potential that is such a change to the vacuum potential. In the most exact sense, this is a method of overpotentializing the plate interface with excess energy from the vacuum, and then letting that energy flow onto the ions to recharge the battery and onto the electrons in the external circuit to power it and its load.

The Bedini overpotential at the battery plates decomposes via Whittaker 1903 {85} as reinterpreted and previously explained, so that excess energy is entering 3-space there, *from the time domain.* Further, the piled-up electrons on the plates and the ions (as charges) in the solution receive such potential energy from the increased potential on them via the same decomposition process. So the creation of the Bedini overpotential on the battery plates, together with dephasing the two currents, is the creation of a true negative resistor at the plates, freely receiving energy from the external vacuum (from the time domain and virtual state) and transducing it into real potential energy and emf on the internal ions and on the external circuit electrons.

Hence Bedini has invented a process for creating a true negative resistor inside a storage battery, and for suddenly thrusting the system out of equilibrium with both the active local vacuum and the active local curvatures of spacetime. As such, the thermodynamics of open systems far from thermodynamic equilibrium applies, and that system is permitted to exhibit COP>1.0, while complying with energy conservation and the laws of physics and thermodynamics. With adroit use and collection of the excess energy, the externally collected energy can be used to close-loop the system and power all its functions. So the system is permitted to power itself and its loads, with all the energy being received from the vacuum via the broken symmetry created. Any overpotential is a dipolarity *a priori,* since any potential is. Hence creating an overpotential is precisely producing an extra broken symmetry of that dipolarity right there at the interface between the two half-circuits and the two dephased, localized currents.

During the "back-popping" pulse signals, one should not think of the energy pulses that Bedini inputs to the battery as the "powering" energy. Instead, one must think of each pulse as "triggering" and "timing" energy which initiates certain other key *negentropic* interactions to freely occur, once the electron pile-up occurs. The resulting negentropic interactions then add substantial additional energy (from the local active vacuum) to the ions in the ion current and to the electrons in the electron current. The freely added energy can be appreciably more than the switching or triggering energy that is dissipated as the "input by the operator".

In short, Bedini deliberately "switches"' and "triggers" certain kinds of vacuum exchange interactions, effectively creating a *true negative resistor* in the battery itself.[150] Due to the broken symmetry of the increased dipolarity *(overpotential)* that Bedini makes in "electron pile-ups" urging reluctant and delayed ion response, the vacuum furnishes extra virtual particle flux to this pile-up of electrons on the plate, which produces an enhanced Poynting energy flow that interacts with the ions in the battery electrolyte. Being charges, these ions thus transduce some of the excess absorbed virtual photon energy into real observable energy, thus increasing their potentialization and energy.

Bedini's method does the following: (i) It forms a true negative resistor in an unexpected way, upon the plates between a pile-up of electrons and the ions in solution in a common lead acid battery, (ii) it uses that negative resistor to extract excess energy from the vacuum and furnish it both to the ions forced into charging mode and simultaneously to the electrons in load powering mode, and (iii) it adds several other stimuli (such as Lenz law effects) which further amplify the negative resistor and enhance the effect, increasing the excess energy extracted from the vacuum and collected in the battery-charging process and also in the circuit-powering process simultaneously.

Specifically, the delay in ion response is adroitly allowed for and manipulated by Bedini to place the battery in ion current recharging mode while the signal pulse electrons between the plates and the external circuit are simultaneously placed in external circuit powering mode. By manipulating the hysteresis and adroitly timing the electron pulses and pulse widths, *Bedini breaks the usual forced Lorentz symmetry of the excitation discharge in a usually closed current loop containing both the*

---

[150] Again we stress that any dipolarity or potential is a negative resistor, producing giant negentropy {12}.

battery's source dipole and the external load. This is possible since his method deliberately opens the system so that vacuum energy enters freely, increasing the potentialization (energy collection) of the ions in the battery solution and upon the electrons between the plates and the external circuit as well.

We stress that Bedini has chosen to *avoid the* usual dissipation of half the energy collected in the external circuit to do nothing but kill the source dipolarity between the battery plates. He works on "that half of the circuit" that is usually just called the "back emf region" and ignored, and he separates and dephases that half of the circuit from the other half. By interrupting that normal "back emf battery-discharging section dynamics and converting it to "forward emf battery recharging section dynamics, while simultaneously powering the external load, Bedini temporarily produces and utilizes a negative resistor right there on the surface of the battery plates themselves.

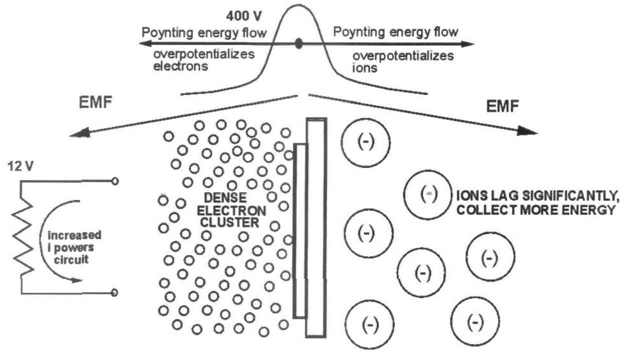

1 Sharp trailing edge of voltage pulse produces intense Lenz law effect Sharply increases electron clustering and potential (negative resistor) at interface Sluggish ions delay speedup further increasing the Lenz law effect

2 Increased bidirectional Poyntjng energy flow from potential overpotentializes even further both the electron duster and the ions, which are only beginning to accelerate in charge mode

4 Large new increased back emf into circuit returns overpotentialized electron current powering circuit with new excess energy pouring from negative resistor potential Continues during ion delay

5 Simultaneously, re-overpotentialized ion current is overcharging battery even faster

Figure 5-10 Phase II of the Bedini negative resistor process in a storage battery

5.8.5.7 Further Increase of the Negative Resistor Effect by Lenz's Law

See Figure 5-10. To further increase the Phase I effect just before it would end, and requiring precise timing of his switching of the pulse leading edge and trailing edge, Bedini then invokes a second phase by carefully controlling the timing for the sharp *cutoff of* the "stimulus pulse" creating the negative resistor. This is usually invoked just as Phase I is preparing to end, but experimentation and adjustment for optimization to the individual circuit conditions may be required.

Invoking the Lenz law reaction by minimizing the stimulus pulse cutoff time, Bedini sharply increases the already-increased negative resistor overpotential by a Lenz's law induced voltage surge, and sharply raises it to as much as 400 volts nominally. The process also sustains the negative resistor overpotential for a longer period, while increasing it again during this second phase. Thus even more free energy from the altered local vacuum potential is delivered to the ions in charging mode inside the battery, while simultaneously even more energy is delivered to the external circuit electrons in powering mode. The overpotential period is also extended. In this second phase, the extra energy W_{IONS} added to the ions in recharging mode is given by

$$W_{IONS} = f(400 - 12)(Q_{IONS}) \qquad [4\text{-}5]$$

and the extra energy W_E added to the electrons in system powering mode is given by

$$W_E = f(400-12)(Q_E) \qquad [4\text{-}6]$$

The total energy added to the system in Phase I by the negative resistor is

$$W_T = W_{IONS} + W_E = f(100-12)(Q_E + Q_{IONS}) \qquad [4\text{-}7]$$

The total energy added to the system in Phase II by the negative resistor is

$$W_T = W_{IONS} + W_E = f(400-12)(Q_E + Q_{IONS}) \qquad [4\text{-}8]$$

And so the total energy added to the system by the Bedini process is just the summation of equations [4-7] and [4-8]. With adroit switching, Bedini need only "pay" a small fraction of that freely received excess energy, in his own operator's input pulse energy and switching costs.

So by invoking a novel negative resistance effect directly upon the driving plate of the battery, Bedini creates (in the first phase) and then further enhances (in the second phase) a "nearly free" overpotential and overpotential period inside the battery. This excess potential directly upon the electron-pileup plate acts in both directions — out into the electrolyte between the plates to overpotentialize the ions in charging mode, and back out into the external circuit in powering mode to overpotentialize the electrons now powering the load. During a fraction of the operating cycle, Bedini recharges the battery while powering the circuit simultaneously, and thus has invented a novel method for extracting energy from the vacuum and curved spacetime to enable a COP>1.0 power system.

This is just a description of one fundamental period where Bedini applies his negative resistance process.

Several other places in the operation of the circuit lend themselves to additional phases of negative resistor formation and usage, and Bedini does use them. We do not discuss them here, since our purpose is only to advance the *fundamental principle involved.*

5.8.6 Independent Replication of the Bedini Negative Resistor Effect

Independent replication of the Bedini effect was accomplished early on by Nelson and by Cole, and also by Watson. Nelson has chosen not to further publicize his own work for some years, perhaps due to his employment as a microwave switching engineer in a large (and conservative) aerospace company. Cole is deceased, and Watson abruptly withdrew —forcibly, in our opinion — from public research and all further contact with his close colleagues after successfully demonstrating (in 1984) an 8-kilowatt power system utilizing a modification of the Bedini process. We are almost certain that Watson received the "offer he could not refuse", and that Watson-type COP>1.0 power systems are in fact already powering exotic underground facilities that "officially do not exist" but which we are aware of. We briefly discuss Watson's 8 kW generator at the end of this Chapter.

For non-battery powered systems, as we explain in Chapter 9, any unitary $COP \gg 1.0$ EM system may experience a Dirac sea hole current decay mechanism which will produce an effective extra "load" in the system's input section, including on out into the external power system. As a general rule (sometimes violated), we estimate that the effect begins to be of importance at $COP > 10^3$ or so. Certainly it is of great importance in the range of $COP > 10^5$. To operate stably at $COP \gg 1.0$, this decay mechanism must be dealt with and prevented. For close-looping into self-powering mode, this negative energy feedback must be converted to positive energy feedback, thus allowing self-powering mode where all the input energy is freely received from the vacuum.

For battery powered systems, the hole current and "extra load in the input section" becomes of significance near COP = 2.0 or even less, because of interaction with the complex battery chemistry and the tendency to charge **the** battery with negative energy as well. So battery powered systems present a special problem for close-looping to achieve sustained self-operation. Bedini has long solved and demonstrated solutions to that problem in a variety of successful experimental systems.

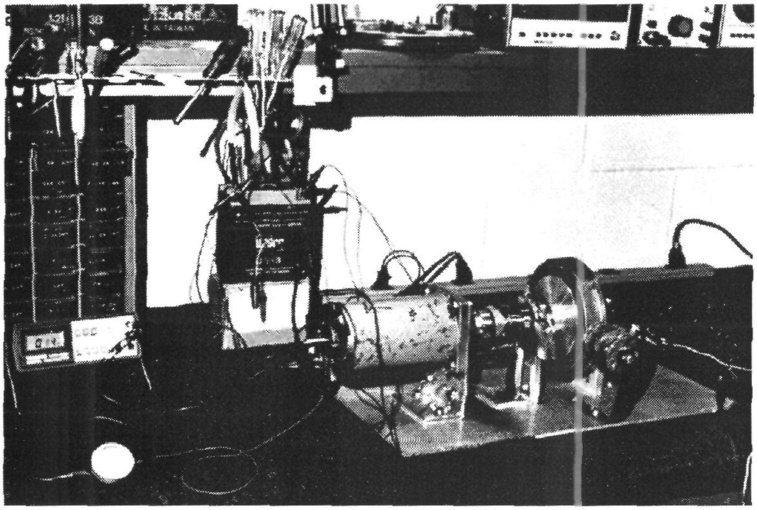

Figure 5-11 **Bedini motor** similar to the **type which ran continuously off a battery for three years, simultaneously recharging the battery.**

To utilize Bedini's battery-powered system at COP > 2.0, or to close-loop it for sustained operation, one must add another Bedini process that transforms negative energy current into positive energy current. That is Bedini's *transducing* negative resistor process {336}. We describe that process briefly in the next paragraph below.

Bedini has built many different motor-generator versions of systems self-powered by use of his negative resistor processes. Figure 5-11 shows a Bedini motor similar to one that once ran off a single battery for more than three years without stopping, using the forerunner of the Bedini back-popping process for creating a negative resistance in the battery. Bedini is still building and testing such power systems; one of his prototype motor-generators today, capable of operating and self-powering off a single battery, is shown in Figure 5-12.

5.8.7 Bedini's Transducing Negative Resistor.

Figure 5-13 gives a simplified block diagram of Bedini's transducing negative resistor process, which asymmetrically regauges negative EM energy into positive EM energy. With this process Bedini is able to intercept the Dirac sea hole current running from the output section of a COP>1.0 EM system to its input section, and divert the negative energy Dirac hole current to "reverse charge" a capacitor. The "reverse charge" consists of providing a hole current to the anti-negative plate of the

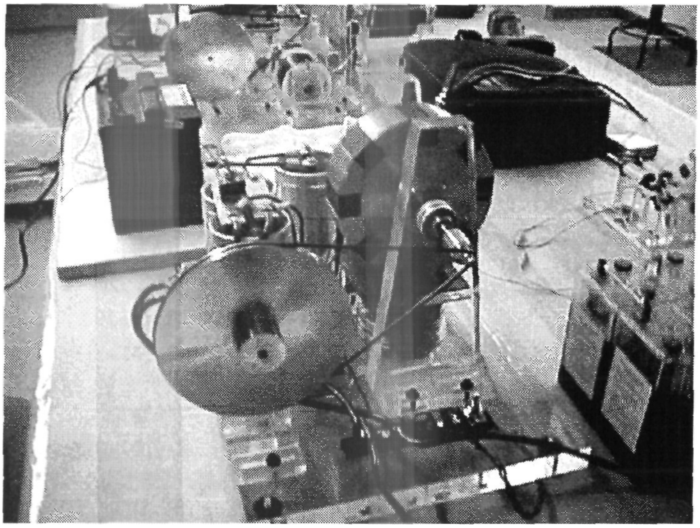

Figure 5-12 Modern Bedini motor prototype which operates off a single battery, powering the load and recharging the battery also.

capacitor, so that holes acting as positrons pile up in the local vacuum of **that** plate. This attracts the trapped electron charges in the dielectric, straining the dielectric in the direction of the hole collection. At the same time, the other plate draws electrons from the system that would otherwise be in currents that fight the back emf of the primary source dipole for the system. That reduces the back emf of the system during the charging of the capacitor; in effect, the back emf has participated in charging the capacitor with excess electrons on the negative plate.

Then Bedini reconnects the charged capacitor to the conductors leading to **the** power supply, discharging the capacitor by electrons furnished to the negative driving plate of the battery powering the system, and simultaneously to the input section of the system itself in normal powering mode. Note that the holes simultaneously discharge from the capacitor's anti-negative plate to the ground side of the battery, further charging the **battery** (the power source!) and the load as well.

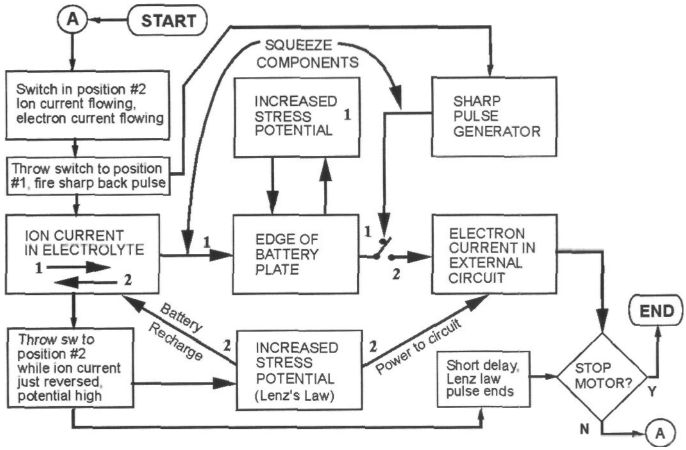
START
A
SQUEEZE COMPONENTS
Switch in position #2 Ion current flowing, electron current flowing
INCREASED STRESS POTENTIAL 1
SHARP PULSE GENERATOR
Throw switch to position #1, fire sharp back pulse
ION CURRENT IN ELECTROLYTE
1
2
1
EDGE OF BATTERY PLATE
1
2
ELECTRON CURRENT IN EXTERNAL CIRCUIT
END
Battery Recharge
Throw sw to position #2 while ion current just reversed, potential high
2
INCREASED STRESS POTENTIAL (Lenz's Law)
2
Power to circuit
Short delay, Lenz law pulse ends
STOP MOTOR?
Y
N
A

Figure 5-13 Bedini transducing negative resistor process block diagram. Adjusting **timing of rear edge of pulse adds additional Lenz law** increase.

Though it does not appear in the conventional textbooks,[151] one can power a load directly by Dirac hole current, but one does so by using the hole current in exactly reverse fashion from how one would use electron current to power the load. Semiconductor components are also critical and usually should be avoided. In short, one simply reverses the terminal that is considered the "driving terminal", and feeds the hole current to that "anti-negative" terminal as now the temporary driving terminal.

The Bedini process thus allows the Dirac hole current (negative energy) produced by a COP>1.0 EM system to be fed back to the input section and transduced into positive energy input to power the system with electron current and also to recharge the battery with positive energy.

---

[151] Lattice hole currents of course do appear and are considered. However, even though the circuit designers consider a lattice hole as a positron, it is not such at all, because the m/q ratio of the ion with one + charge is quite different from the m/q ratio of a true positron. The Dirac hole in the vacuum is effectively the "4-positron" — still in negative energy Dirac 4-electron state — before its observation or interaction with mass. It is a true negative energy entity, and thus a negative mass entity. Prior to observation, it actually should be said to be a "negative energy x time" entity and a negative "masstime" entity. It is also a negative energy electron entity. This subtlety is essential if one wishes to produce successful antigravity experiments on the bench, as we point to in this book and as were exhibited by the Sweet vacuum triode amplifier at COP > $1.5x10^6$.

This converts what would otherwise be negative energy feedback to the input section of the system (and to the battery) to positive energy feedback **to** its input section and to the battery. Instead of seeing the negative energy current in the input as an additional "load" to be powered, the external power supply sees it as excess positive energy input by the environment to the system, reducing the "load" represented by the system. It thus reduces the power required by the system from the external power supply.

When the amount of "transduced" positive feedback electron current to the input section equals the "normal electron current draw" required by the system, the battery (or generator) can be disconnected and the unit will continue to operate in the self-powering mode, fed completely by the positive energy from the Bedini transduction process. The other *supersystem components* (comprising the external environment of the system) will continue to power the system and its loads and losses, with all the input energy to the system being taken from the combined active local vacuum and local curvature of spacetime.

5.8.8 Gray's "Splitting the Positive Pole" Engine.

One of the real pioneers in early application of causal positron energy was inventor Edwin V. Gray of Van Nuys California. In WW II Gray learned radar during his tour in the U.S. Navy, having previously attended advanced engineering school in the Army until discovered to be under age 15 and discharged. After the war, by 1958 he had learned to perform what he called "splitting the positive pole", using "cold electrical energy" (his name for negative EM energy) as well as the normal positive electrical energy. His first motor was operating by 1961. Preparing to put units into production, in 1974 the Los Angeles District Attorney confiscated all Gray's records and prototypes. To escape bogus charges, Gray later pleaded guilty to two minor infractions and was released. His prototypes were never returned. After several mishaps, moves, and more prototypes, Gray died in Nevada under mysterious circumstances in April 1989. We quote directly from Lindemann {341}:

> *"In the early 1980's, Gray offered the U.S. Government his technology to augment Reagan's SDI program. He actually wrote letters to every member of Congress, both Senators and Representatives, as well as the President, Vice-President, and every member of the Cabinet. Remarkably, ... Gray did not receive a single reply or even an acknowledgement!"*

In fact, during his early experiments, Gray discovered he could power light bulbs with cold energy, etc. Later he was to demonstrate this fact in a sensational manner: With a light bulb lit from his generator, he simply immersed the lamp — electrical leads and all — in water, where it remained lit. With his hands in the water, he would have been electrocuted or severely shocked if the electrical energy involved had *been positive* (diverging) energy. Instead, it was *negative* (converging) energy.[152] Another demonstration was to hit the coil of an electromagnet with some 3,000 volts, which hurled the electromagnetic into the air — but did not heat it. The electromagnet's coil would have been instantly melted if the energy had been positive, and the magnet would have been heated.

The reader should recall the timing of Dirac's theory of the electron {498a-b, 249}, which included the negative energy electron as early as 1930. Dirac also pointed out that we would observe this negative energy electron as a positive energy electron of opposite (positive) charge and positive mass, going in the opposite direction. The *observed* 3-positron and the *unobserved* negative energy 4-electron are quite different entities, producing quite different EM fields. The positron produces positive energy EM fields and positive gravity while the negative energy 4-electron produces negative energy EM fields and negative gravity.

In 1930, to his previous theory of the electron Dirac introduced the vacuum energy: the electron sea filled with electrons occupying negative energy states. The holes are treated as the positive electrons (as if observed). This vacuum energy medium is now referred to as the "Dirac Sea." First, Dirac attempted to identify these holes with protons, but that was falsified. Later the positron was experimentally discovered.

---

[152] By convention, "positive" energy has been associated with diverging EM energy, which is heating. Negative EM energy is associated with converging EM energy, which is cooling. In scalar interferometry {79}, one can control whether the distant EM field energy emerging in space in the distant interference zone shall be diverging (positive) EM field energy or converging (negative, cooling) EM field energy. This is determined simply by the biasing of the electrical grounds of the interferometer transmitters. The weather control referred to in 1997 by Secretary of Defense Cohen {89} is accomplished by using multiple scalar interferometers, some to gently heat the atmosphere in certain selected local areas (forming low pressure areas by expansion of the air) while cooling it in other selected local areas (forming high pressure areas by contraction of the air). By then moving these highs and lows around adroitly, one can capture and steer the giant jet streams, thereby steering the weather as one wishes, and directing significant weather changes into an area. By placing fairly sharp turns on the jet streams, one can induce great spin energy, thereby spawning tornadoes. Interestingly, by sharply pulsing the interferometer and controlling its electrical ground bias, distant cold explosions or hot explosions can be produced at will. Such weapons have been tested worldwide, and many test incidents are given in our older books F*er-de-Lance* and *Gravitobiology.* Information on such incidents is available on website http://www.cheniere.org .

One of the problems with present conventional application of Dirac's theory is that the positron is not used in its *unobservable causal negative energy 4-electron state* by modern theorists and experimentalists, but *after observation* where its direction has been reversed, its charge has been reversed, time has been stripped away, and the mass has become positive. So it is actually used as a positive energy entity having positive mass, and producing positive energy EM fields and positive gravity. This defeats the entire magic of the causal (unobserved) 4-positron as the electron carrier of negative mass and negative energy {*convergent* electromagnetic energy) in currents in the vacuum (spacetime) itself, rather than the electron carrier of positive mass and positive energy in ordinary 3-space. By replacing the unobserved (causal) vacuum 4-positron (negative energy electron) with the observed (effect) 3-positron with positive mass and positive energy, the entire vast area of negative energy EM fields — and thus direct antigravity via electromagnetic means — is discarded.[153] But the physicists avoided having to grapple with practical negative mass and practical negative energy (cool mass and cool energy). It also set back the pace of physics in discovering and engineering antigravity transportation for nearly three quarters of a century.

It appears that Gray's thinking about cold energy was more prosaic than the somewhat esoteric considerations of the Dirac Sea theory. Conventional electrical engineering circuit theory considers positive charges moving around a circuit (although some electronics technical circuits required thinking in terms of electron current). Hence the "high" side of the standard circuit was the positive polarity, or the positive pole of the terminals of a generator or battery. From this "positive pole", positive current was sent around the external circuit to the ground return line by the forward emf (forward voltage). However, forcibly returning to the positive pole from the ground return line (i.e., from the negative pole) were those same positive charges, being forced back up by the back emf and requiring expenditure of half the external circuit's collected Poynting energy.

---

[153] When the huge nondiverged Heaviside energy flow component is appreciably of "convergent" or of negative EM energy rather than positive EM energy, the antigravitational aspects are substantive and observable (and practical). The unaccounted Heaviside negative energy flow component is often a trillion or more times as great as the accounted Poynting negative energy flow component. Hence antigravity technology is possible and practical, as we present in a later chapter.

So to Gray, the positive pole had two currents, one useful and one usually detrimental. The positive pole had (i) outgoing positive emf current powering the external circuit's loads and losses, doing the beneficial work (dissipating energy) in the external circuit with its load, and (ii) incoming positive current being forced back up against the back emf, doing the harmful work against the back emf and using up half the energy previously collected out there in the external circuit. Gray was aware that the work against the back emf was equal to the work done by the forward emf. So he set out to split these two currents apart, and catch the incoming current against the back emf separately, then use that normally wasted or oppositional energy to further power loads and/or recharge the batteries powering the system. Something like this was the way he was thinking.

That is what Gray called "splitting the positive pole". He also reasoned (and found) that the "reverse current" must be cooling instead of heating, since it moves against the voltage and acts backwards from regular current dissipated in resistors, etc. This precisely matches the cold (negative) energy phenomena met in the Sweet vacuum triode amplifier and in the motionless electromagnetic generator, as well as other COP>1.0 electrical systems.

Gray's later demonstrations showed that he had recognized this process or something very similar to it, in his "splitting the positive pole".

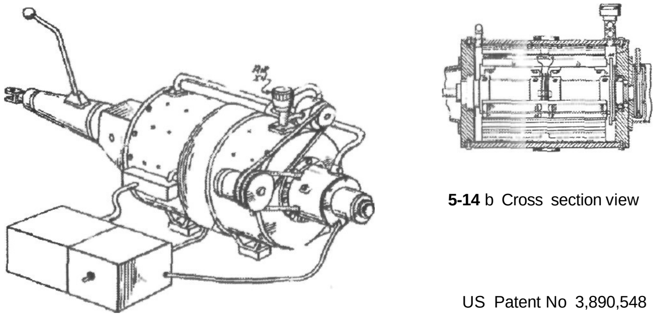

5-14 b Cross section view

US Patent No 3,890,548
issued 17 June 1975

5-14 a The Gray engine

Figure 5-14 Gray's "splitting the positive" COP>1 0 engine.

To take advantage of his concept and principles, Gray developed a special engine shown in Figure 5-14. In 1975 Gray received a patent on this engine {337} along with his special means of battery powering and special switching circuits.

Gray's engine uses pulsed capacitive discharge to power a magnetic motor, with recharging of the powering capacitors taking place *between* successive discharge positions of rotor and stator. The motor runs by initiating repulsion rather than attraction, which — apparently without Gray recognizing it in such terms — invokes the notion of the "superpole" pioneered by Bedini. The superpole is an important concept in COP>1.0 research, though most researchers in the field have never heard of it. John Bedini conceived it several decades ago and demonstrated the surprising difference between a superpole and a common pole of double strength.

When two equal and like magnetic poles are facing each other in repulsion with a small gap between them, a net stress magnetic potential (pole) is formed in that gap which has an energy density some four times as strong as the normal field energy density in the field from one of the two opposing like poles. Hence the local energy density of the vacuum in the gap has been substantially increased above the energy density in the field region between a north and a south pole facing each other. Further, this energy density is in a "field-free" Lorentz-regauged form, i.e., a stress potential form. From that superpotential *outward,* there exists a normal magnetic field. But in it, there is no net magnetic field, but there is a "stress" field nonetheless.

Compared to an ordinary magnetic motor using attraction as well as some repulsion, the exclusive use by Gray of the Bedini "superpole" principle caused an extra *symmetrical* regauging[154] of the local vacuum energy input to the system (to the gap between the poles) as the magnetic superpole

---

[154] Every conventional electrodynamics text already assumes and uses the Lorenz-**Lorentz** symmetrical regauging as being absolutely free. Symmetrical regauging assumes that the energy of the system has been freely changed twice, but in such a way that the two free fields formed are equal and opposite, forming a stress potential. Gray found how to split apart and separately use these two "free-flowing exc**hanges** of energy between system and active vacuum" that comprise the overall stress potential of the system, and thereby violated the Lorenz-Lorentz symmetry condition, freeing the two equal and opposite EM fields and their free energy to be used individually and beneficially to (i) provide extra power to the load and (ii) recharge the powering batteries so they did not run down. Using a different method. **Bedini** has done likewise.

(stress magnetic scalar potential). However, this stress potential is comprised of bidirectional longitudinal wavepairs, as shown by Whittaker in 1903. This superpole or stress potential thus can and does "split" into two different fields in antiparallel directions from each other. Since the north pole is deemed "positive" in magnetics, then using the two anti-fields comprising the free positive superpole is indeed "splitting the positive pole". Gray coined a very apt phrase for what he was doing.

Each of those antiparallel fields also involves an associated and unaccounted phase conjugate field {338a-c}, hence it is a time-reversed field (after observation). So opposing time-polarized EM waves also accompany Whittaker's decomposition waves. In short, the scalar potential accomplishes *optical-type pumping* in the time domain, and can be used to demonstrate time-reversal (as in Becker's bone-healing method).

Before observation (interaction with charged mass) and while still in the causal state, *the EM field from the positive pole is actually an EM field from a negative energy 4-electron, and the field can be taken to be a negative energy field!* In our view, positive energy enters every point dipole in the polarized vacuum at the negative charge, thence to the positive charge, thence back to the time domain. From the positive charge, the positive energy is leaving 3-space back to the time domain. Hence positive EM energy diverges from the negative charge and converges on the positive charge, then "disappears", so to speak, back to the time domain from whence it came. The "disappearance" of positive (divergent) energy is the disappearance of heat energy; hence it is a cooling process rather than a heating process. Or said another way, *negative* EM energy enters 3-space at each point dipole, entering from the time domain to the positive charge. Thence it moves to the negative charge, and back to the time domain. We see or detect the positive energy flow in 3-space directly, while the negative energy flow in the Dirac vacuum is evidenced by the "backwards" interaction creating Newton's third law reaction force.

By splitting the positive superpole, one has effectively destroyed the Lorentz symmetry condition. Hence the excess energy in the superpole will be capable of performing real work in the system in two directions. One direction is from the system to the external environment so that it produces the dissipation of energy (therefore work) in the loads and losses. The other direction produces a "negative resistor" or negentropic action that adds that much excess usable energy to the system back into its input section. By adroitly intercepting and switching that energy into capacitors and/or batteries, Gray was able to charge capacitors, recharge batteries, and even power loads with cool energy (as demonstrated later by immersing a lighted electrical line and a lit light bulb in water).

A positive-energy resistor becomes a negative resistor when fed with negative energy. Then it converges positive EM energy from the external environment and outputs it into the circuit in which the resistor is connected. Note that this is the same statement as the resistor receiving (extracting) negative energy from the circuit and emitting it back to the time-domain. *That's like extracting back emffrom the circuit and tossing it offto the time domain.* A conductor to positive energy flow is an insulator or dielectric to negative energy flow. Hence the normally "conductive" water was a neat insulator for Gray's "dunking" demonstration of negative power, as were his own normally conductive hand and body. To negative energy, a coil seems to produce capacitance rather than inductance, while a capacitor seems to produce inductance rather than capacitance.

Thus when the "positive pole is **split"** in the system, the system has broken **its** Lorentz regauging and is an open system in disequilibrium with the external vacuum environment, freely receiving excess energy from it.[155] With adroit switching and energy capture, such a system is permitted to exhibit COP>1.0 because it has violated the free Lorentz regauging once the excess energy in the superpole region is discharged and used.

---

[155] When one analyzes the supersystem of any Lorentz-regauged EM system, one realizes that the extra stress potential energy — entering the system and "locked up" as a stress potential due to the Lorentz condition — is actually a special relativistic **rotation** of the frame of the power unit away from the laboratory frame. That is, the system when rotated away from the lab frame, takes on free symmetrical regauging energy in that rotated frame as an excess stress potential. Gray's "splitting the positive" process simply separated the two appositive fields comprising the stress potential, then produced excess free energy from that *broken* free regauging Lorentz condition. Literally Gray splits Lorentz's stress potential into halves, the halves being its two opposing EM fields and their energy. This is equivalent to the system being rotated back into the lab frame, releasing its excess energy it possessed in the rotated frame back into the lab frame. Richard "Scott" McKie has also filed a patent on an invention deliberately rotating a capacitor's frame and fully charging it — while it is in its rotated frame. The rotated capacitor can be charged with less energy furnished from the lab frame, than is obtained back out of the capacitor in the lab frame when the charged capacitor is suddenly rotated back and discharged. At least in **theory,** McKie's overunity process should be workable if the switching costs are negligible. Our contribution to McKie was to advance the theoretical mechanism for **the** experimental mechanism he had discovered and was using.

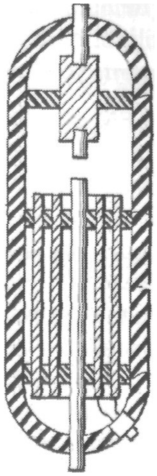

From US Patent No **4** 661747
issued 28 April 1987

Also in U S Patent No 4595,975
issued June 17,1986

Figure 5-1S Gray's conversion tube used with power systems,

Gray also utilized a special conversion discharge tube, shown in Figure 5-15. This tube uses the fact that a very sharp pulsed discharge in it results in a temporary COP»1.0 condition. This COP»1.0 condition instantly involves causal Dirac sea 4-positrons, moving as negative mass current and negative energy current in the local vacuum, which sweep back from the output side of the discharge to the input side.

We also point out the severe overpotential that must accompany any such violent pulse discharge process. The overpotential is a departure from equilibrium conditions, and is the difference between the actual potential and the equilibrium potential. The overpotential is well known and utilized extensively in electrochemistry {339a-c}, but is little used in conventional circuit theory. Nonetheless, it is involved in cold fusion processes due to the double layer phenomena {340a-b}, and it is involved in all electrode discharges (as from a cathode to an anode). It is also involved widely in electrochemistry and ion processes in solution.

Figure 5-16 shows Gray's circuit containing his conversion tube. Figure 5-17 shows Bedini's successful replication of the Gray conversion tube. Figures 5-18, 5-19, 5-20, and 5-21 show four pertinent pages from Bedini's laboratory notebook on his replication of the Gray conversion tube and the Gray motor system.

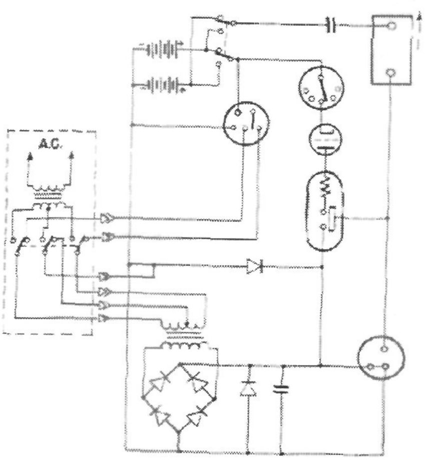

From U S Patent No 4 661 747
issued 28 April 1987

Also in US Patent No 4 595 975
issued June 17, 1986

Figure 5-16 Gray's circuit containing his conversion tube.

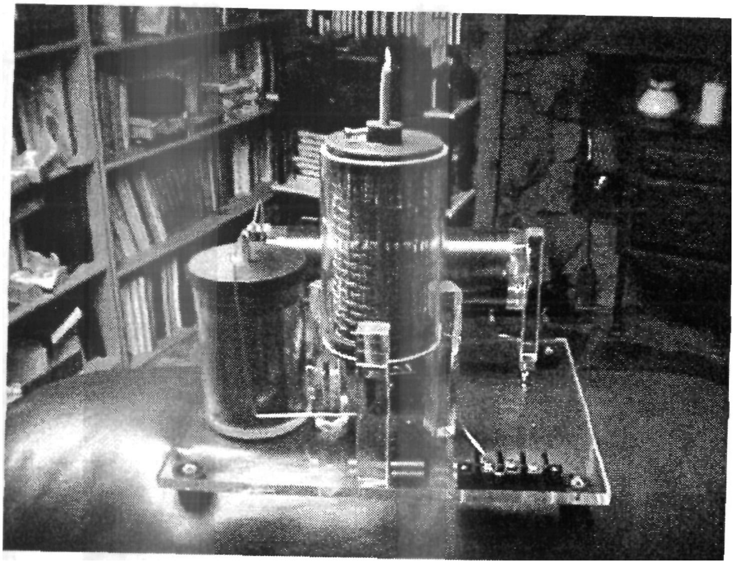

Figure 5-17 Bedini's successful replication of the Gray conversion tube.

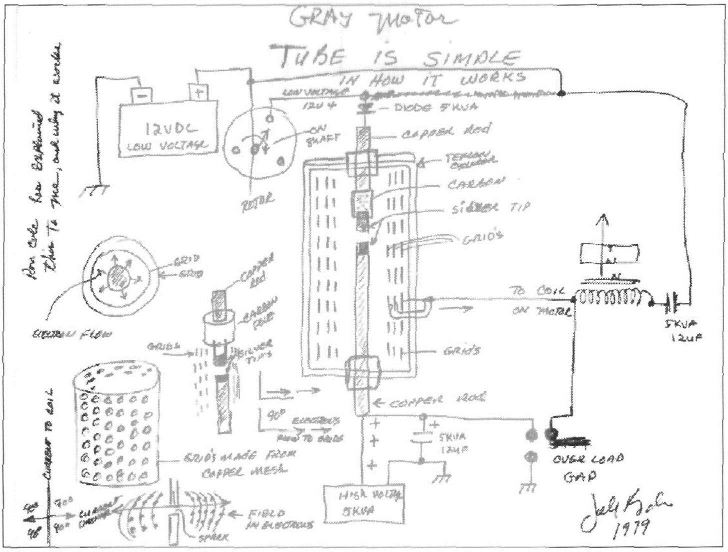
GRAY MOTOR
TUBE IS SIMPLE
IN HOW IT WORKS
12VDC
LOW VOLTAGE
LOW VOLTAGE
12V+
DIODE 5KVA
ON SHAFT
ROTOR
COPPER Rod
TEFLON CYLINDER
CARBON
SILVER TIP
GRIDS
Ron Cole has Explained this to me, and why it works
GRID
GRID
ELECTRON FLOW
COPPER ROD
CARBON TUBE
GRIDS
SILVER TIPS
TO COIL
ON MOTOR
5KVA
12uF
GRIDS
COPPER ROD
90°
ELECTRONS FLOW TO GRIDS
5KVA
12uF
CURRENT TO COIL
GRIDS MADE FROM COPPER MESH
OVER LOAD GAP
HIGH VOLTAGE
5KVA
90°
90°
CURRENT CHANGE
FIELD IN ELECTRONS
SPARK
1979

Figure 5-18 Page 1 from Bellini's lab notebook on the Gray conversion tube.

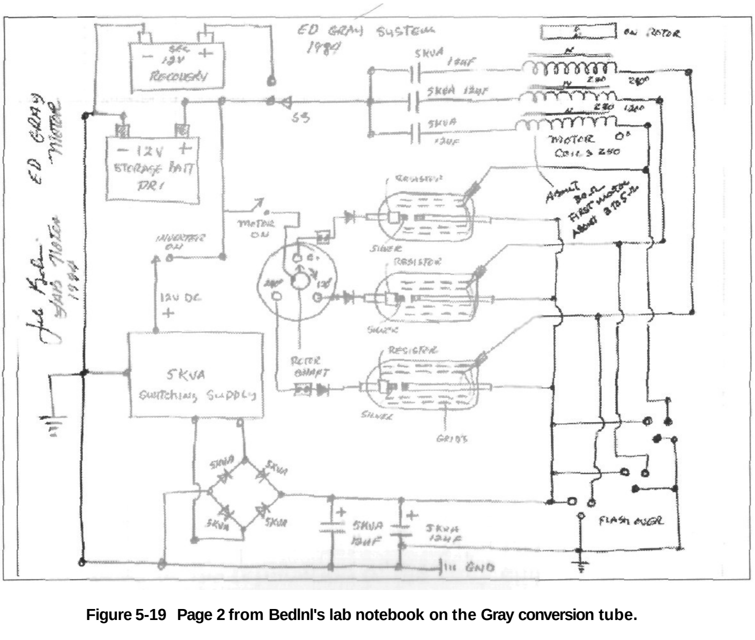
ED GRAY SYSTEM
1984
ON ROTOR
SEC
12V
RECOVERY
5KVA
12uF
5KVA 12uF
5KVA
12uF
280
240°
280
120°
MOTOR 0°
COILS 280
ED GRAY MOTOR
12V
STORAGE BATT
PRI
SS
MOTOR ON
INVERTER ON
RESISTOR
SILVER
ABOUT 30Ω
FIRST MOTOR
ABOUT 3 TO 5Ω
LAB MOTOR
1984
12V DC
+
0°
240°
120°
RESISTOR
SILVER
5KVA
SWITCHING SUPPLY
ROTOR SHAFT
RESISTOR
SILVER
GRIDS
5KVA
5KVA
5KVA
5KVA
5KVA
12uF
5KVA
12uF
FLASH OVER
GND

Figure 5-19 Page 2 from Bedlnl's lab notebook on the Gray conversion tube.

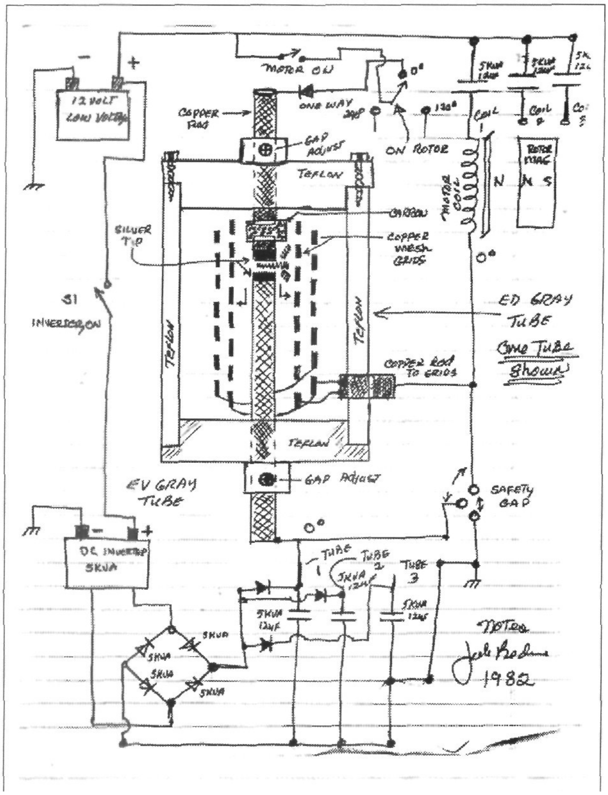

−
+
12 VOLT
LOW VOLTAGE
MOTOR ON
0°
ONE WAY
240°
120°
COPPER ROD
GAP ADJUST
ON ROTOR
5KVA 12uF
COIL 1
COIL 2
COIL 3
ROTOR MAG
N
N S
MOTOR COIL
TEFLON
CARBON
SILVER TIP
COPPER MESH GRIDS
0°
S1
INVERSION
ED GRAY TUBE
One Tube Shown
COPPER ROD TO GRIDS
TEFLON
GAP ADJUST
EV GRAY TUBE
SAFETY GAP
DC INVERTER 5KVA
0°
TUBE 1
TUBE 2
TUBE 3
5KVA 12uF
5KVA
Notes
Jack Bedini
1982

Figure 5-20 Page 3 from Bedini's lab notebook on the Gray conversion tube.

In science, when a novel experiment is replicated independently, that experiment then is established as a scientific fact. It then becomes the task of the theorists to either change the theory to accommodate the experimental results, or come up with a new theory to explain them. With Bedini's successful replication of Gray's work and extension of it, that work has now been scientifically established. With his "splitting the positive pole", Gray was referring to an actual, useful mechanism he evoked and used in the processes ongoing in his motor, in his conversion tube, and in his power system.

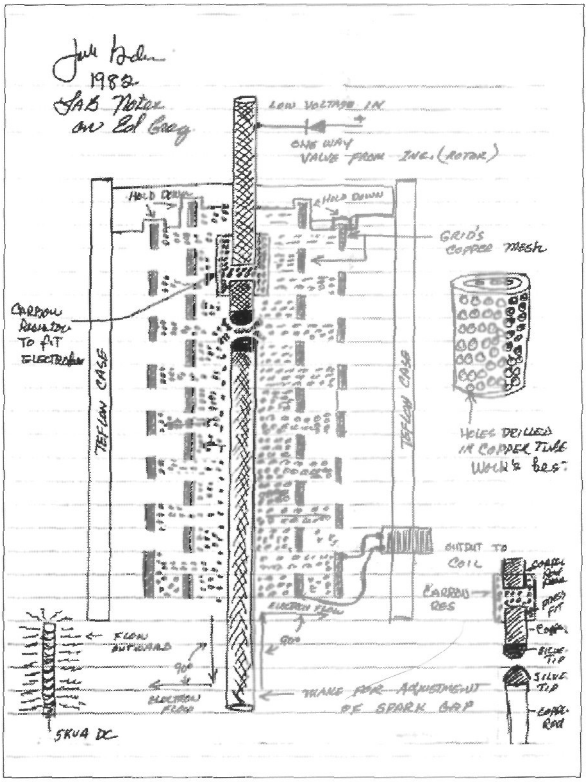
John Bedini
1982
LAB Notes
on Ed Gray
LOW VOLTAGE IN
+
ONE WAY
VALVE FROM INC. (ROTOR)
HOLD DOWN
HOLD DOWN
GRID'S
COPPER MESH
CARBON
RESISTOR
TO FIT
ELECTRODE
TEFLON CASE
TEFLON CASE
HOLES DRILLED
IN COPPER TUBE
WORK'S best
OUTPUT TO
COIL
CARBON
RES
ELECTRON FLOW
90°
FLOW
OUTWARD
90°
ELECTRON
FLOW
THREAD FOR ADJUSTMENT
OF SPARK GAP
5KVA DC
COPPER
ROD
PRESS
FIT
COPPER
SILVER
TIP
SILVER
TIP
COPPER
ROD

Figure 5-21 Page 4 from Bedini's lab notebook on the Gray conversion tube.

For further discussions about the Gray motor, conversion tube, and the history of it all, one is referred to the book by Lindemann {341}.

5.8.9 Watson's Self-Powering Generator

See Figure 5-22. In 1984, Jim Watson demonstrated a self-powering 8-kilowatt generator at the International Tesla Conference in Colorado Springs. This generator was a modified extension of smaller motors and generators built by Bedini over a period of years. At the conference, engineers from the audience were invited to the platform to perform their own independent measurements of the unit, while it was running and powering its load. The engineers directly confirmed that the system was delivering power to the load (as could be seen visually also), and that it was recharging its batteries as well.

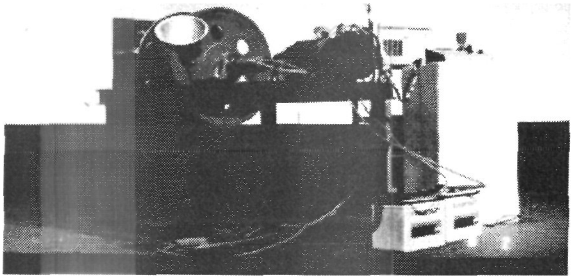

Figure 5-22 Watson's self-powered 8 kW generator demonstrated in 1984.

Interestingly, one of Watson's batteries was stolen that night. The reason was obvious. First, at the conference there was present one of the major persons charged with suppression of COP>1.0 electrical power systems world wide, along with several of his henchmen. Since they have been involved in suppressing many dozens of similar COP>1.0 systems, they are thoroughly familiar with both the theory of COP>1.0 electrical power systems and the various types of devices and mechanisms that have been invented in that area over the past century.

One feature of Watson's unit was that it charged the batteries with negative energy (i.e., with Dirac sea hole current, or "cold energy" as sometimes referred to), and could use negative energy for powering the unit. Hence, his batteries became charged with negative energy as the machine ran and powered its loads. After much demonstration and running, such a negatively charged battery has its local curved vacuum charged with an appreciable number of unfilled Dirac holes. Such a negatively charged "battery supersystem" will then absorb positive energy charging (electron current charging) for a protracted period before it will start to change its voltage and accept the positive charge. Instead, during that protracted initial period the battery will just "eat electron currents and positive energy" from the battery charger, with no apparent effect and no change in Its discharged state or its voltage.

A very simple way to absolutely ascertain whether or not a unit such as Watson's (without a Bedini negative energy converter) is genuine, is to **simply** take one of his batteries from a unit that has been running quite a bit, and recharge it on a conventional battery charger. If it exhibits that two-phased recharge phenomenon, it proves that the unit is real and is indeed a true COP>1.0 device, or certainly has the potential to be one. Coupled with the demonstration and independent measurements, this removed all doubts about the validity of Watson's COP>1.0 system.

So it appears that certain parties who were concerned about that system decided to do the "simple test" to see if Watson's device was real. It was real, and so they discovered. At that point, the High Cabal — Winston Churchill's term — recognized they had another successful unit and inventor to suppress.

Bedini and I were in close contact with Watson over a period of years, as often as three or more times a week. Later, some time after that conference, abruptly Jim ceased all communication with his former close colleagues. For two days a rather crazy recording **was** on his phone, stating the Watsons had moved. Then that disappeared. Watson never again contacted either Bedini or me, or his own financial backer (the late R. J. Reynolds III). Even Reynolds could not find Watson's whereabouts at that time.

We are quite certain that Watson received the "offer he could not refuse". Simply put, this is the old Mafia trick of visiting the target, offering him a deal, and he has no choice but to accept or die. When this is used to suppress COP>1.0 systems, and some medical systems, the offer is usually: *"Accept the money and you and your family will live very comfortably, but you will never openly work on this again or communicate it. You either accept our generous offer, or you and your family will be dead when we leave here."* Assuming that happened Watson had no choice but to accept the several million dollars offered him.

He is still alive, because Jeane Manning physically ran into him at a conference — and recognized him and spoke to him — whereupon he withdrew abruptly and left the conference very hastily.

5.9 Suppression of Free Energy Systems Has Been Common

This is only one of many cases we know of where such suppression of overunity systems and inventors has been accomplished. Others are not so benign. Marinov was killed with a longitudinal EM wave shooter in Europe, and his body was thrown off a building to make it appear a suicide. The police allowed the body to lie on the pavement for a protracted time (it was emitting longitudinal EM waves from the time-charging action of the strike). When finally moved, the pavement glowed in that area that had been underneath the body. Only one weapon on earth will kill a person in that manner, and that is a longitudinal EM wave "shooter".

Another prominent researcher was killed with a Venus technique (look it up, it's a distortion of the wave front) modified beam that plunges the heart into violent fibrillations. He rushed from a restaurant exclaiming, "They're killing me!" and died of a resulting heart attack. The present author was hit with just such a weapon in a restaurant here in Huntsville. A friend and I actually saw the weapon and the person wielding it. Knowing what it was and recognizing the symptoms, my seated companion and I were able to bolt out of the nearby back door during that minute or so we could still physically function. I personally saw the device and the person doing the beaming, and I have a witness to the event.

The partner of a colleague in Australia was killed on an upper floor by such a shooter from the street below, right through the walls. The assassin was observed, and was seen putting the bazooka-sized shooter back in his car, then speeding away. There are many other true incidents we could also relate, including a person killed by an ice-dart dipped in curare (a very professional hit method, used by some of the sinister arms of various intelligence systems). So there is more than just scientific work involved in pursuing COP>1.0 EM power systems. There can also be a direct threat to one's life, essentially at any time.

Almost any serious overunity researcher, who succeeds, will have had confrontations with assassination attempts, suppression attempts, threats, deliberate legal entanglements, etc. The present author is no exception. It is sad that one of the prerequisites for serious scientific work in this area is o**ften** a legal gun permit (for both the inventor and his wife) and either legally carrying a hidden weapon or never being very far from a weapon when working at home. As an old soldier, of course, one just shrugs and accepts it, and gets on with the job to the best of one's ability, albeit remaining very watchful and always on guard.

5.10 In Conclusion

In this chapter we have introduced many of the approaches to COP>1.0 EM systems or processes, or the possibility for such in some novel mechanisms known in science. Recalling our previously presented solution to the source charge problem, we emphasize to the reader that all assumed 3 - space EM processes are actually 4-space processes prior to observation, continually observed in iteratively frozen 3-snapshots by iteratively invoking the *d/dt* operator that produces observation. Hence from the source charge (and source dipole) solution we realize that all EM energy in 3-space comes from the time-domain and returns to the time domain. It iteratively appears along the "line of propagation" in 3-space at each successive position, but does not continuously propagate in 3-space. It does continuously propagate in 4-space. Since all the EM energy in any EM circuit comes from the time domain via the broken symmetry of the source charges and source dipoles, then we need to apply this greater vision to the COP>1.0 mechanisms and approaches so that we see the 4-space to 3-space to 4-space interplays.

The notion of the negative resistor is particularly significant and direct in that respect, and we have presented several versions of true negative resistors. For those wishing to experiment, we would suggest the easiest thing to start with is the point contact transistor. Or, simply go to Naudin's website {313} and download his instructions for making a Chung-type negative resistor, build it, and do experiments with it by varying the pressure on the fibers etc.

Again, we have not presented "parts kits" with instructions, but we have presented things that have worked or can reasonably be expected to work with some effort. We have also discussed the principles involved, to the best of our ability. In a few cases, we have directed the reader to information on websites where building instructions for selected negative resistors and other devices are posted.

As in any other scientific endeavor, we also urge the interested reader to read the literature and check the references cited, or as many as possible. Increasing one's understanding of the area is necessary to increase one's probability of success in the area.

We also wish the reader good results and success in his or her research and experimenting. The energy crisis is not going to go away, and the demand for oil will double in the lifetime of many of our middle-aged or younger readers. The cheap oil peaks shortly and declines in availability forever thereafter {342}; this alone will drive up the price of oil and seriously impact the economies of the nations of the world. To keep from destroying the world economy and the biosphere, it is absolutely essential that COP>1.0 electrical power systems — including self-powering systems taking their input energy from the seething local vacuum — be developed to produce the necessary electrical energy needs of the world, cheaply, cleanly, and easily.

If one or more of the readers should succeed in developing and producing such systems, then the entire purpose of this book will have been fulfilled.

Chapter 6

Additional Approaches to Overunity Power Systems

> *"Can the quantum potential carry a signal? If it can, we will be led to a violation of the principles of Einstein's theory of relativity, because the instantaneous interaction implied by the quantum potential will lead to the possibility of a signal that is faster than light... a signal has in general to be a complex structure, consisting of many events that are ordered in definite ways. In terms of our language, each signal is a supersystem of events, while each event is in turn a system of subevents.... At present, the answer is, of course, not known. "* [Bohm and Hiley] {343}. A comment:[156]

---

[156] We comment that the "event" involves observation: i.e.. it is conceived as an *observed* event. In that exact sense, no signal in space carries "events" because there is no interaction with mass and no *d/dt* operation induced on entities (curvatures) of spacetime. The effect (observed) 3-space field cannot propagate faster than light, because it cannot even propagate in the first place! The totality of light's EM interactions with matter is what makes the observed field or observed signal to appear to be propagated at light speed. It is not actually "propagating", but being iteratively recreated from point to point in space *if observed at every point.* Signal in spacetime is causal and hence nonobservable; the interaction of causal signal with charged mass is observable and makes an "event". When we clearly differentiate the nonobserved (causal) 4-field or dynamic from the iteratively observed 3-field or dynamic, then the *nonobserved 4-field* or dynamic is not limited to light speed. Propagation along the time axis, e.g., can be at "infinite velocity" because a single point in time is connected to every point in the universe simultaneously. Since all EM energy in space comes from the time domain locally at each point dipole in 3 - space and returns to the time domain from that same point dipole, it is the notion of "propagation continuously through 3-space" at *any* velocity that is a non sequitur. Absolutely nothing propagates through 3-space! If general relativity is re-interpreted **to account** the difference between observed and unobserved (effect and cause), then the unobserved cause can propagate superluminally without violation of the "observed event (effect) propagation at light speed" of general relativity. It is the notion that anything propagates in 3-space that is the non sequitur, and it should be removed from special and general relativity as well as electrodynamics.

> *"...contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish."* [Aharonov and Bohm] {344}.

> *"...for any force varying as the inverse square of the distance, the potential of such force satisfies both Laplace's equation and also the wave equation, and can be analyzed into simple plane waves propagating with constant velocity. The sum of the waves, however, does not vary with time, i.e., they are standing waves. Therefore the force potential can be defined in terms of both standing waves, i.e., by a global, or non-local solution, and by propagating waves, i.e., by a local solution changing in time... Thus, Whittaker's mathematical statement related the inverse square law of force to the force potential defined in terms of both standing wave (i.e., global) and propagating wave (i.e., local) solutions. The analysis also showed that the electromagnetic force fields could be defined in terms of the derivatives of two scalar potentials."* [Barrett] {345}.

> *"What might appear to be empty space is, therefore,* ***a*** *seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing energy and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it it actually swimming in a sea of ghost particles of all varieties — virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Davies] {346}.

6.1 Introduction

In this chapter, we continue to present a selection of potential overunity systems, inventions, approaches, and selected mechanisms. We also make the reader aware **that quantum** mechanics, an entire branch of physics, requires nonlocality and "instant propagation" in space. We utilize the quantum potential from Bohm's hidden variable theory of quantum mechanics {347} to impress the nonlocality. Further, the staid old scalar potential — which we meet as voltage in a circuit — also exhibits nonlocality and instant propagation in the Coulomb gauge {348}. Our purpose is to show that electrodynamics is not limited to the cut-and-dried subject as usually taught to electrical engineers and applied by them, but also covers a wide range of phenomenology that electrical engineers usually do not consider or account.

6.2 Energy Conversion and Energy Amplification

6.2.1 The Mead-Nachamkin Zero Point Energy Converter

It is well known that the vacuum interacts with electrical charges and dipoles, and also with the electrons in an atom. For example, Lamb received the 1955 Nobel Prize in physics jointly with Polykarp Kush for experiments measuring the small displacement later called the "Lamb shift" of 0+ne of the energy levels in atomic hydrogen {349}.

Casimir {350a} proved that closely separated conducting plates also alter the energy density of the vacuum, thereby becoming an "energy extracting" method, although the energy is miniscule. It is quite real, however, as shown in meticulous experimental work by Lamoreaux {351}. **The** principle that vacuum energy interacts with matter, plates, and electrical double layers and circuits is well proven experimentally.

In 1996 Mead and Nachamkin were granted a patent on an overunity EM power system process for extracting zero-point energy of the vacuum {352}. If one closely examines the patent wording, the device is properly patented as an energy converter and does not overtly state that it is a COP>1.0 system. It is a COP>1.0 system, however, since the input energy is freely received from the vacuum and not input by the operator himself. **One** embodiment of the invention inserts a tiny coil between the plates to ex**tract** the energy. See Figure 6-1. The energy is based on the Casimir e**ffect** {26}, which was accurately measured in 1997 by Lamoreaux {27} a**nd** since then by others.[157] Further investigation of the force's behavior **with** the shape of the conducting plates has been performed by Chen *et*

---

[157] See also U. Mohideen and A. Roy, "Precision measurement of the Casimir force from 0.1 to 0.9 microns," *Phys. Rev. Lett.*, Vol. 81, January 1998, p. 4549.

al.[158] Numerous other papers of interest on the Casimir effect have also been published {353a-d}.

Because energy cannot be created or destroyed, any COP>1.0 system is an *energy converter,* a priori. It must receive the excess energy from the environment (in this case, from the curvature of spacetime and from the local active vacuum) and convert that energy to a form usable to power or assist in powering the system loads and losses.

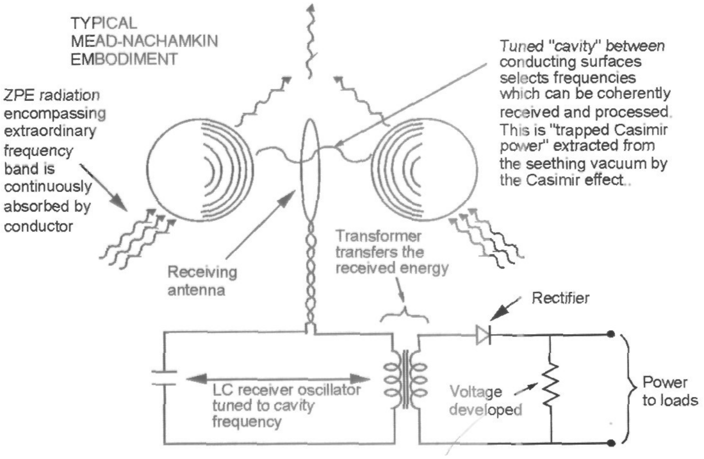
TYPICAL
MEAD-NACHAMKIN
EMBODIMENT
ZPE radiation encompassing extraordinary frequency band is continuously absorbed by conductor
Tuned "cavity" between conducting surfaces selects frequencies which can be coherently received and processed. This is "trapped Casimir power" extracted from the seething vacuum by the Casimir effect..
Receiving antenna
Transformer transfers the received energy
Rectifier
LC receiver oscillator tuned to cavity frequency
Voltage developed
Power to loads

Figure 6-1 Mead-Nachamkin conducting surfaces, antenna, and receiver for extracting EM energy from the vacuum,

The Mead-Nachamkin patent is interesting because the physics of the device is proven and accepted. Getting sufficient power output to be useful for electrical power purposes, however, remains to be demonstrated, since the Casimir effect is very tiny.

6.2.2 The Quantum Potential As an Energy Amplifier

In Bohm's hidden variable theory of quantum mechanics {354}, an extra potential is added to the Schrodinger equation, and is known as a *quantum potential.* Figure 6-2 lists the novel characteristics of this potential, and

---

[158] F. Chen, U. Mohideen, (G. L. Klimchitskaya, and V. M. Mostepanenko, "Demonstration of the Lateral Casimir Force," *Phys. Rev. Lett,* (in press as of the time of this writing).

Quantum Potential Characteristics

- No point source; not radiated
- Just appears instantly
- Quantum potential between two particles
  - Interaction does not vanish as spatial separation becomes very large
  - Instantaneous connection
- Depends on quantum state of system as a whole
- System parts can be greatly separated

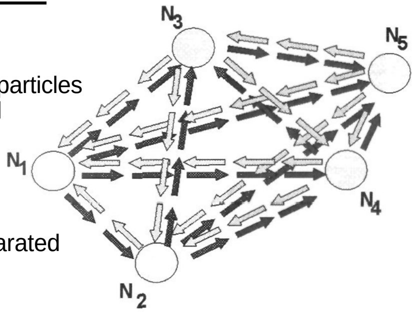

Figure 6-2 Quantum potential characteristics (table).

Figure 6-3 lists technical features of Bohm's hidden variable theory from whence the quantum potential is taken {355}. It can "move instantaneously", at "infinite velocity", which means it just suddenly appears everywhere it will be, and does not "propagate at a finite speed" through singly connected space at all {356}.

$$V_{quantum} = \frac{h}{2m}\frac{\nabla^2\sqrt{\rho}}{\sqrt{\rho}}$$

A quantum particle moves as if it were subject, in addition to its external potentials, to a potential which is a function of its own probability distribution.

Bohm's hidden variable theory assumes:

- **Particle and wave function real and separate**
- **Wave function obeys Shrodinger's equation**
- **Particle obeys classical mechanics**
- **Particle couples to wave function through a quantum potential**

Figure 6-3 Technical features of Bonn's hidden variable theory.

The potential is after all *potential energy.* With the quantum potential (QP), a joule of energy placed upon one collector, where that collector is connected via a QP with one or more distant collectors, will also result in a joule of energy (or some fraction thereof) instantly emerging upon and in each of those distant participating (QP-connected) stations simultaneously. It appears that, if only a fractional participation — say, 0.50 — exists with one or more of the distant stations, each of those stations will experience the appearance of that fraction (in this case, 0.5 joule) of the energy input to the initiating station. Any station can input energy, and it will appropriately appear at every other participating station, regardless of where these stations are located in ordinary 3-space as seen by the external observer.

One way of expressing this "instantaneous QP connection" between separated objects is to model the different objects (or stations) as linked together in a multiply connected space, with a certain QP connection or collection fraction. Presence of energy at one station in the multiply connected group, results in the simultaneous appearance of the energy — or an appropriate fraction of it — at every other station in the group.

To understand multiply connected space, consider a single point on a plane. Now consider 999 more points superposed on/in that same point. But suppose that those thousand points are really located at different locations in 3-space, though superposed in this "crazy space" containing a plane where all of them are superposed at one point.

As the ordinary external observer, suppose you input a joule of energy into one of these — to you — separated points, which happens to be the original point and is located in your laboratory. Supposed the fractional connection is 1.0 (the connection is perfect in multiply connected space). Voila! Instantaneously one joule of energy appears at each and every one of those (to you) widely separated points. If you vary your input energy (e.g., as in communication), it varies precisely the same in each of those distant QP-connected points.

Those who continue to advocate that superluminal communication of information is impossible are ignoring present experimental proof that it does occur. As a single example (there are others), it has been shown by Enders and Nimtz {357} that Mozart's 40th symphony can be transmitted as frequency modulated microwaves through a barrier wave guide between two points in a waveguide, via quantum tunneling, at 4.7 times the speed of light c. We point out that this is achieved using *transverse* EM waves.

If *longitudinal* EM waves are used together with longitudinal modulation (which engineers might think of as a sort of "pulse compression" technique), the only limitation to the speed of communication is given by the impurity of the longitudinal EM wave one is able to make. The Fogal semiconductor, e.g., can be used to transduce ordinary transverse EM wave signals into longitudinal EM wave modulations inside a scalar potential. Hence that scalar potential can move in the Coulomb gauge at infinite velocity. It requires another Fogal semiconductor circuit on the receiving end to receive and transduce the longitudinal EM waves and their longitudinal modulations back into transverse EM waves for processing and amplification by normal means. This capability of the Fogal semiconductor circuits has been demonstrated through satellite uplinks and downlinks coupled with long transmission line links, with essentially no delay from transmission to reception. One may think of the Fogal superluminal process as *tunneling through subspace.*

The best way to vary the energy is apparently by use of a variable stress potential (which translates to a longitudinal variation or modulation). In short, one must modulate the "internal energy" of the vacuum, so to speak, and not the "external energy". A reasonable discussion of internal energy versus external energy is given by Chen {358}. If one longitudinally modulates the internal energy of the vacuum potential — which, being a scalar potential, decomposes by the Whittaker decomposition {359} — then superluminal communication is not only possible but also practical. We expect Fogal to eventually place just such a system on the market, shortly after his semiconductors get into precision production.

One possible method of producing modulated internal stress potentials is by use of a Davis noninductive resistor {360}. Another is by use of a biwound coil with two windings, so that the two currents through two equal "superposed" coils are equal and opposite. We point out that, while there has been published much stuff and nonsense about such biwound coils, the truth is quite mundane. Simply examine the Lorentz symmetrical regauging applied to the Maxwell-Heaviside equations. It is the deliberate production of two equal and opposite force fields, which is in fact the creation of a stress potential and extra stress energy. By varying the magnitude of the Lorentz forces comprising the stress potential, the stress potential is varied in amplitude. Hence the concomitant longitudinal EM waves are varied in amplitude. A biwound coil is a variable Lorentz-regauging coil, *a priori.*

By oscillating the magnitude of any Lorentz regauging (which in the simplest case is simultaneously changing both opposing fields equally and oppositely), one can make longitudinal EM waves. By appropriately modulating both Lorentz regauging components, one can make longitudinal EM modulations upon those concomitant longitudinal EM wave carriers.

So with some careful tuning and adjustment, and a little nonlinear material in the core,[159] one should be able to have the superposed appositive magnetic fields lock together *(modulate,* rather than linearly *add)* and produce a magnetostatic scalar potential stress signal in and from a nonlinear core material. We caution, however, that the biological effects of the radiation of such stress potentials internal to spacetime have not been investigated. The experimenter is cautioned that he experiments at his own risk, and he is cautioned to experiment only at low power levels until the phenomenology in a particular case is understood.

---

[159] The opposing waves must modulate (multiply), not mix (add), and modulation is a nonlinear process. Linear mixing of opposite waves or fields will not cause the

If we compare the stress potential oscillation (wave) to a sound wave, the similarities suggest that the stress potential oscillation can be treated as a longitudinal EM wave, if the two superposed opposing EM waves "lock together". Simple addition is insufficient; modulation is necessary. Hence the nonlinearity of the conductors and core material is of significance.

For instantaneously propagating potentials and pure longitudinal EM waves, any conceived difference between the two becomes somewhat academic. The characteristic of interest is the instant appearance of the input energy at multiple points via a multiply connected space.

To the ordinary observer, use of a quantum potential and multiply connected space measurably yields a fantastic *energy amplifier.* One furnishes one joule, and gets out — say — 1,000 joules in those distant locations altogether. As shown in Figure 6-4, the QP system is a communication system of enormous implications, particularly as humans travel into space across the solar system and even beyond. Unfortunately, by using appreciable power in the initiating station, the QP system also leads to weapon systems of incredible destructive capabilities. Some five nations {361} now possess such quantum potential weapons, which are the dominant weapons on Earth. We do not further discuss the quantum potential weapons aspects in this book.

Back to our little energy exercise, inputting one joule and getting out some 1,000 joules. Suppose these 1,000 participating points are not distant at all, but are gathered closely together into a local material but still QP-connected with a connection fraction of 1.0. Now we input a joule of energy to that device, and it develops and outputs one kilojoule. It is hypothesized that this becomes the ultimate negative resistor. It will also exhibit separated Dirac hole current phenomena as the COP increases and, if sufficient COP is developed, it will exhibit antigravity effects such as

---

fields to "lock together" into a common stress potential. Nonlinear mixing (modulation) will. Hence the nonlinear core is required for conversion of appositive fields to stress potentials and **Lorentz** regauging.

those shown by the Sweet vacuum triode amplifier discussed in this chapter.

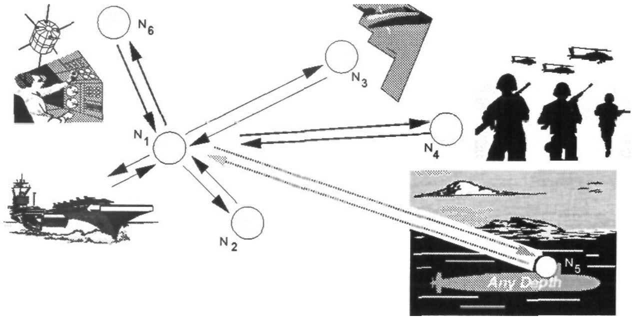

- Quantum potential decomposes per Whittaker 1903 into bidirectional longitudinal EM waves
- Can insert additional longitudinal waves, thus information using a Fogal semiconductor
- Can receive "signal" and discriminate the information using a Fogal semiconductor

Figure 6-4 Quantum potential communication system.

Strangely, overall conservation of energy is not violated by the quantum potential COP>1.0 energy system. It's just that the energy now is furnished from space (from the vacuum) in a *multiply connected* manner. However, all the excess energy is still furnished by the active external environment - i.e., from the time domain, according to the giant negentropy process {12} — though that environment now is quite novel and multiply connected. So we still have a system analogous to a windmill, though *a very novel windmill and a very novel wind indeed.*

Without belaboring the point, action-at-a-distance is absolutely required in quantum mechanics. Further, photon correlation experiments have long demonstrated experimentally that such action-at-a-distance effects do exist, just as predicted. This was a refutation to Einstein's objection to quantum mechanics' "spooky action at a distance".

We point out without further comment that, if a self-powered process is Invoked in the initiating station, then all other participating stations also contain that self-powered process. However, note that one does not really wish to "invoke" a quantum potential connection and use it except occasionally and very quickly! Otherwise, if a strong QP is turned on and **left** on among many distant stations, anything and everything happening to each one of them can instantly appear in all of them, so long as the "constant of connectivity" remains a constant or nearly so.[160] For that reason, strong quantum potential connections can be self-destroying and strong effects do not seem to normally appear in the physical world.[161] The proper energy use of a quantum potential is thus to use it only for fleeting "inductions" of effects in the desired stations, in *momentarily* multiply connected spacetime, at least until the self-limiting behavior is ascertained.

Apparently, a quantum potential can sometimes be formed in the limit of an iterative, mutually phase conjugating operation of EM signals passed back and forth between two phase conjugate mirrors, or among many of them. Such iterative mutual phase conjugate reflection of signals — particularly pumped signals — is technically known as iterative *self-targeting.* Sometimes we use a more picturesque term, *ping-pong,* for iterative self-targeting.

Conceivably, strongly coupled, strongly pumped optical self-targeting mirrors can — in the extreme limit — initiate or partially initiate a quantum potential. The superposing beams of energy may be thought of as first narrowing to laser-like beams, then gradually subsiding into pure longitudinal waves with transverse components "canceled" to form simply "stress energy". In short, if this reaction occurs, the energy is transformed into a stress energy potential, which acts either as a scalar potential in the Coulomb gauge or as Bohm's quantum potential. In our view, the pure longitudinal wave has infinite velocity, as pointed out by Rodrigues and Lu {93a}, so Lorentz-Fitzgerald contraction results in "zero-separation distance", which in turn results in a multiply-connected spacetime.

One points out that the "input energy" to the initiating station may be deliberately structured internally, including altering the internal Whittaker longitudinal EM wave functions comprising the input electrical field energy. In this manner, in theory deliberate spacetime curvature sets and dynamics can be introduced into the distant but multiply connected targets. A form of highly advanced electro-biological warfare induction of disease has been developed by at least one nation (Russia) in this regard, as an extension of Kaznacheyev's experiments showing induction of cellular disease at a distance. It appears that the cells of a living body are interconnected by a very weak quantum potential in such manner. Again, a great deal more research on that possibility is needed.

---

[160] Fortunately, it appears that this connectivity constant is not really constant, but reduces as the feedback of each station's fluctuations to the other stations increases. The result seems to be a "leveling off process reaching a plateau and a fixed amount of gain in the overall system. In short, the system usually seems to be "self-adjusting" and "self-limiting" as the overall gain increases. Nature is riddled with quantum potentials, but this self-limiting process keeps most so tiny that they are not even noticeable.

[161] However, the validity or invalidity of this statement for some of the highly energetic phenomena of astronomy remains to be determined. It also remains to be determined as far as the quantum potential's role in generating the zero point energy fluctuations of the active vacuum.

Given a quantum potential connection between separated objects including our initiating station, one can also add vacuum *engines* (energy structuring of curvatures of spacetime itself) as our input to the initiating node. Those vacuum engines will then simultaneously appear unabated (or only partially abated) in the multiply connected separated objects. In this way, in theory one can use the QP and QP energy to perform highly anomalous *momentary* engineering of nuclides, materials, fields, etc. in widely separated, distant objects and areas. By repeating the momentary engineering "pulses", distant objects and areas can be profoundly altered.

The use of vacuum engines in medical applications is of particular future interest.[162] Very radical extension of those techniques can be achieved once quantum potential technology is developed and available in the open scientific community.

But let us ignore the weapons, free our scientific imagination, and let it roam. As an example, in the future a vacuum engine (spacetime curvature engine) might be developed and utilized in a world-wide quantum potential to *momentarily* strike and eradicate all AIDS viruses on Earth, and do only that. Or, the engine could be slanted to simply *momentarily* slightly change those viruses into an innocuous kind of new virus. The cells of the individual human body are already connected, it appears, by a very weak quantum potential. By connecting an external initiating station to that body's quantum potential, the energy of the body could be directly scanned, all abnormalities noted, and tailored engines created and *momentarily* used to treat the body and correct its abnormalities and malconditions. This, we believe, may well be a medical therapy of the future, a few generations from now. [163]

---

[162] See Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981; Vlail Kaznacheyev *et al.,* "Distant intercellular interactions in a system of two tissue cultures," P*sychoenergetic Systems,* 1(3), Mar. 1976, p. 141-142; Vlail Kaznacheyev, "Apparent information transfer between two groups of cells," *Psychoenergetic Systems,* 1(1), Dec. 1974, p. 37. Unfortunately this QP induction of cellular disease has already been used by the former Soviet Union to develop a QP extension of the **force**-free structured potentials method first used to induce cellular disease and **health** changes in U.S. personnel in the former U.S. Embassy in Moscow for several decades. Now QP weapons using such techniques can target the entire population of a targeted nation. Using the internal Whittaker structuring of quantum potentials and EM fields and waves, *electromagnetic biological warfare* has already been born and highly developed. The actual dimensions of possible biological warfare strikes are far greater than has been mentioned in the open literature.

6.3 Other Systems of Electrical Interest

6.3.1 Russian Parametric Oscillator Power Systems

In leading physics institutes and laboratories in the 1930s, Russian scientists built COP>1.0 parametric oscillators, including some of appreciable size. The work is documented in the literature {362a-k}, both in Russia and in France. The devices were developed and tested in several Russian laboratories. With linear loads, the oscillators would progressively build to self-destruction. With nonlinear loads, the devices would stabilize and power themselves and their loads, at a level depending on conditions — reminiscent of the behavior shown in Figure 4-3 in Chapter 4 for stability COP>1.0 levels for the asymmetrical self-regauging process.

Standard references for parametric oscillation in EM circuits are available. However, with standard practices none of these conventional oscillators produces COP>1.0, primarily because the conventional closed current loop circuit is applied without special modifications. For overunity work, the

---

[163] Apparently, tests of such QP capabilities in weaponry have been gingerly conducted for direct induction of disease patterns in the population of a targeted nation or area. One symptom of such a test is that a very small number of widely isolated cases of the same disease breaks out simultaneously, in isolated cases scattered widely and puzzlingly throughout the population, *without the possibility of any direct carrier vector connection.* Apparently, the test inductions have deliberately been held mostly to the "shadow state" (or upper virtual state, just below observable state) in most of the targeted populace, to prevent giving away the tests. However, the human immune system and cellular regenerative system do respond to such "near-observable-state" signals and engines, when sustained or repeated in time, by a sort of coherent addition (integration) over time. The bell-shaped distribution curve predicts that a few individuals, randomly scattered throughout the populace, will have quite lower induction thresholds and will evidence the actual disease after a certain time, whereas in almost all the other members of that population the "disease" will never evidence in the observable state during that time. The test is deliberately terminated when only a few observable cases occur, which are sufficient to prove the efficacy of the test and the method.

Russian and French literature (as we referenced) should be meticulously studied and analyzed in a higher group symmetry electrodynamics, to see how the Lorentz condition is broken during discharge of circuit excitation energy in the load, and how the COP>1.0 Dirac hole current decay mechanism is avoided or contained.

The Russian COP>1.0 parametric oscillation work appears to have been suppressed by the Communist regime just prior to WW II. After the war, all such technology passed under the rigid control of the KGB (by whatever name it took and takes from time to time) and into the special weapons research and development area — still highly classified in Russia. We know of no similar work presently ongoing in the West, although at least an English translation of one of the major summary documents of the Russian parametric oscillation COP>1.0 power system developments was performed under NASA auspices {363}. However, in conventional U.S. parametric oscillation research, I know of no case where COP>1.0 operation is being sought, or where the Dirac sea hole current decay mechanism is even mentioned or recognized {364}.

6.3.2 Sweet Vacuum Triode Amplifier

During the 1980s and 1990s, the present author worked —sometimes closely for extended intervals — with inventor Floyd Sweet, the inventor of the Sweet vacuum triode amplifier (VTA). Indeed, we gave the unit its VTA name, at Sweet's request.

Fig. 6-5 (pg. 384) shows a diagrammatic illustration of the fundamental VTA construction. Two coils in quadrature are between two barium ferrite "brick" magnets, of the kind formerly used in many power audio systems. The vertical coil is the "input" signal coil, and the horizontal coil is the "output" power coil. The input consisted of a 33 microwatt 10 volt 60-Hertz signal, and the output was a 120-volt 60-Hertz 500-watt signal. The barium nuclei in the magnets were specially preconditioned by Sweet so that they were in powerful self-oscillation with the surrounding energetic vacuum. A double-edged razor blade or piece of shim stock placed on one of these magnets would oscillate back and forth, incessantly, without cessation — showing that the magnetic field itself was "waving" back and forth (Figure 6-6). It was also continually doing work against air **resistance,** by moving the air.

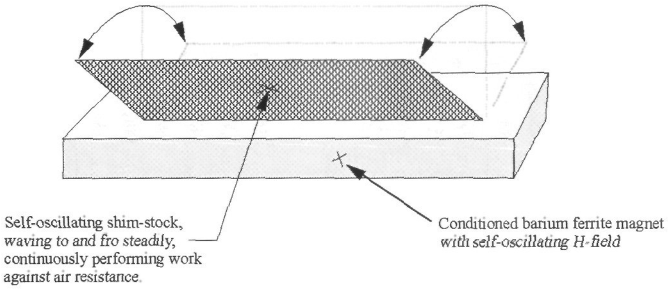
Self-oscillating shim-stock,
waving to and fro steadily,
continuously performing work
against air resistance.
Conditioned barium ferrite magnet
with self-oscillating H-field

Figure 6-6 Blade waving continuously on Sweet's conditioned barium ferrite magnet.

Sweet had a Master's degree in electrical engineering from a leading university and handled the mathematical theory very well. He had worked at General Electric for many years, and Gabriel Kron was his patron and mentor. Sweet often spoke glowingly of Kron, and I came to believe that Sweet's VTA was probably an outgrowth of, or very similar to, Kron's negative resistor.

- Two activated magnets facing, with fields in self-oscillation.
- Barium ferrite magnets, barium nuclei in self-oscillation
- Barium nuclei self-pumped
- Two coils in quadrature, load is lamps, six watts
- Next unit produced 500 watts output with 330 microwatts input; later made 5 kW VTA

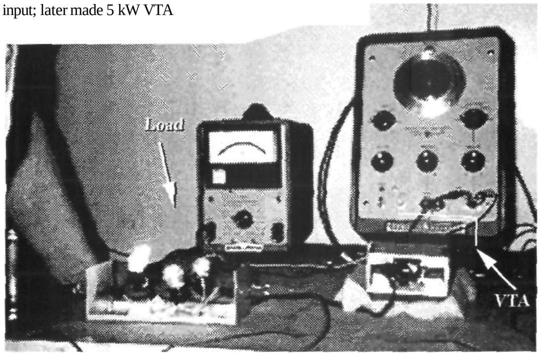

Figure 6-7 First Sweet VTA producing 6 watts output.

When I first met Sweet, his little VTA system was producing a 6-watt output, enough to light four 1.5 watts of auto lamps (Figure 6-7. But there was no question as to the genuineness of the device. Nothing was hidden, and Sweet allowed me to measure the device at will, disassemble it, play with it, and examine it in any fashion. It was genuine, and not a hoax or trick. I also locked up one of his specially conditioned magnets for 24 hours, with a piece of shim stock sitting on the flat of the magnet and waving to and fro continuously, steadily performing work by moving air. When I opened the lock the next day, the shim stock was still there on the magnet and oscillating, having continuously done work against the air resistance for 24 hours with absolutely no energy input by the operator. And it was still working. Indeed, that single "kinetic" permanent magnet destroys all objections to COP>1.0 EM systems, including those that are self-powering and thus have COP = ∞.

It seems little known that the vacuum around nuclei, in some cases, can be treated as a semiconductor, e.g. as discussed by Prange and Strance {365}. It is also known that nuclei do exhibit resonances at ELF frequencies. In particular, the vacuum in the region close to the nucleus of a superheavy element is known to act in a fashion analogous to the inversion layer in a field effect transistor. Prange and Strance introduce the idea of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that the vacuum can be similarly manipulated by appropriate means. We personally suspect that the semiconducting vacuum can be and is resonantly involved in any ELF resonances of the nucleus, which can occur in lighter nuclei such as barium.[164]

The virtual particle flux of vacuum, regarded as noise, may provide noise amplification of the coherent self-oscillation frequency between the semiconducting vacuum and the barium nucleus. Certainly, analogous noise amplification of signals is known in electrical physics {366}.

---

[164] Here our concept of the supersystem may be of utility. The "lock-in" or "freeze-framing" of an equilibrium condition for a system — such as in a state of nuclear ELF self-oscillation — is stabilized and made a "new equilibrium condition" when the force field reaction from the curved local spacetime into the system is equal and o**pposite** to the force field reaction from the local active vacuum into the system. We believe Sweet's undisclosed activation process was a method for synchronizing those two force field reactions and making them equal and opposite in one short discharge. We hope that this speculation is of use to future experimenters trying to duplicate Sweet's activation of his magnets into sustained and powerful supersystem self-oscillation.

I hypothesized that Sweet's activation process treated the vacuum surrounding the barium nucleus in such fashion, so that he was able to establish self-oscillation between the local activated vacuum and the concomitantly activated barium nucleus. Since barium ferrite is optically active, it may be that Sweet discovered how to get sufficient "self-pumping gain" for the self-oscillation to endure and not die away in a decaying oscillation manner. Since the "gain" of the second VTA as a self-pumping device was some 1,500,000 (Figure 6-8), obviously Sweet's activation method introduced a powerful state of self-oscillation.[165]

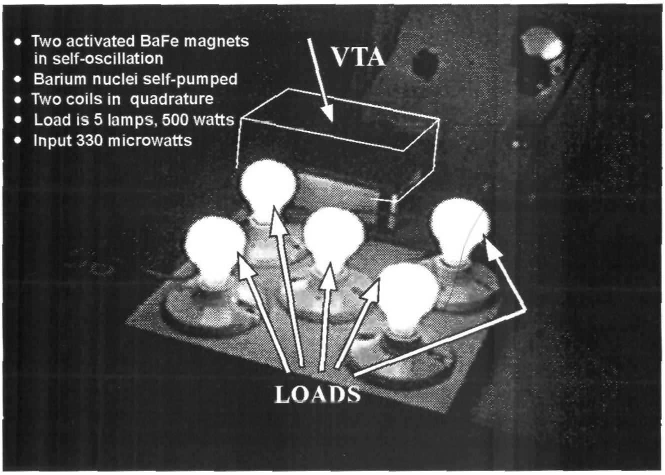
• Two activated BaFe magnets in self-oscillation
• Barium nuclei self-pumped
• Two coils in quadrature
• Load is 5 lamps, 500 watts
• Input 330 microwatts
VTA
LOADS

Figure 6-8 Second Sweet VTA producing 500 **watts and C0P=1,500,000.**

---

[165] Here we point out that "small" things with very close "double surfaces" of opposite charge or potential usually have very large fields. The tiny Lamb shift, e.g., has a local energy density greater than the surface energy of the sun. Jackson, *Classical Electrodynamics,* Second Edition, 1975, p. 10-11 points out that *"field strengths of the order of 10^9-10^{15} volts/cm exist at the orbits of electrons in atoms, while the electric field at the edge of a heavy nucleus is of the order of 10^{19} volts/cm."* It should be obvious to the reader that, if our speculation is true and nuclear oscillation of the barium nuclei is involved, the production of optical gains of 1.5 x 10^6 in Sweet's self-pumped, optically acting barium ferrite magnets is not surprising because of the extreme magnitude of the oscillating nuclear fields. In that case, optical gain converts directly to true power gain because the pumping energy is furnished freely by the environment and not by the experimenter.

6.3.2.1 Mapping the Magnetic Field of the Magnets

Figure 6-9 shows a small magnetic field mapping device built by test engineer Rosenthal for Sweet. Sweet "scanned" and mapped the consistency of the magnetic field from his candidate magnets (barium ferrite magnets bought from surplus stores at the time). If the consistency varied over 12 to 15%, the magnet was useless because it would not "hold" the activation and retain it. Magnets whose magnetic field variation did not exceed 10% were ideal. So Sweet only found about 1 in 10 or even 1 in 30 magnets that would retain the self-oscillation state when initiated.

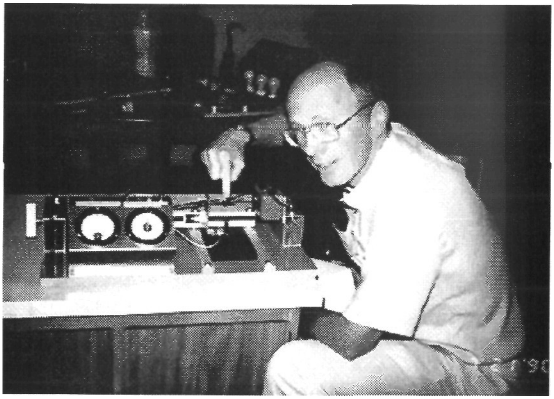

Figure *6-9* Test engineer Rosenthal points to a magnetic field strength mapping device he built for Sweet.

Sweet may have actually used two appreciably higher frequencies in his activation process, with the difference frequency between them being a precise ELF frequency (60-hertz, 100-hertz, 400-hertz, etc. as he desired). That would make sense, when one realizes that the magnet is a highly nonlinear medium and then treats it (to first order) as an isotropic nonlinear medium. In that case, the two wave frequencies actually used would be subject individually to the normal overshoot, breakup, reconstitution, etc. inside the material of the magnet — but the difference frequency would behave as if it were a sine-wave frequency {367}.

This makes sense, since the output wave of Sweet's activated magnet was essentially a pure sine wave. Also, the presence of noise-enhanced heterodyning can sometimes be used to amplify a difference signal even further {368}. As we stated, since the activation is a self-oscillation between the local semiconducting vacuum and the barium nuclei, one wonders whether much of the rest of the vacuum virtual particle flux can be treated as noise, and whether an effect is obtained for noise amplification of the coherent self-oscillation frequency.

L'vov and Prozorova {369} point out some interesting characteristics of the formation of self-oscillating spin waves which occur above parametric excitation. These oscillations result when internal stability does not occur, and they evidence themselves as oscillations of magnetization. The frequencies of the oscillations usually lie in the range from tens of kilohertz to tens of megahertz. At small above-threshold ratios, the shape of the oscillations is nearly sinusoidal. At larger ratios, the shape differs appreciably from sinusoidal. At still larger ratios, the oscillations become chaotic.

If Sweet did use the "difference frequency" conditioning, one suspects he may thus have been able to utilize larger above-threshold ratios, where the two frequencies actually transmitted would have resulted in chaotic oscillations. It is interesting to speculate that, in that case, for a suitably chosen barium ferrite magnet, the difference frequency still can behave sinusoidally, but now much stronger. If so, that might account for the tremendous COP Sweet attained in his second unit, which exhibited a COP = 1.5 x 10^6, and could be pushed even higher.

6.3.2.2 Sweet VTA for Closed-Loop Self-Powering Operation

Courtesy of Walter Rosenthal, we now present a visual record of the Sweet closed-loop self-powering version of the VTA and its momentous successful acceptance test on May 1, 1990, as closely observed by Rosenthal.

Figure 6-10 shows Sweet and a different version of the VTA, with the coils in a slightly different configuration and the windings altered. This is the two-magnet, 2-coil VTA that he built and tested for self-powering operation. Figure 6-11 shows the unit mounted on a special test fixture built by engineer Rosenthal. This fixture and its components allowed the sensing of the sine wave output power at the zero-value crossover points. To switch in the special clamped feedback system for self-powering operation of the input, and simultaneously switching away from external power, the switching had to be done very accurately at that zero-point crossover. Feedback would not stabilize and hold if the switching were done at any other point in the output cycle.

Unit independently examined and testing observed by test engineer Walter Rosenthal

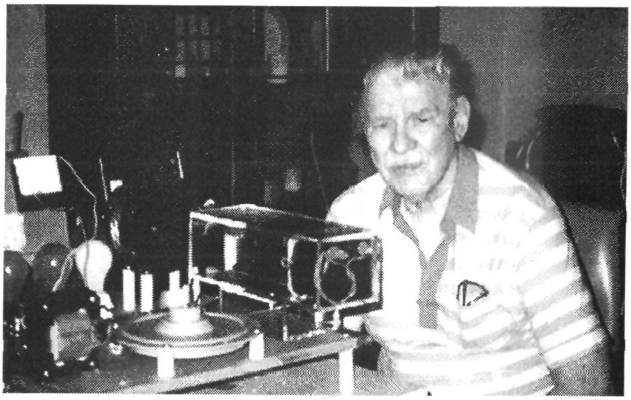

Figure 6-10 Sweet with 2-magnet 2 coil VTA, loads, and instruments for historic self-powering operational tests on May 1,1990.

Figure 6-12 shows the set-up for the tests, with the unit and loads as well as some of the instruments used. Figure 6-13 shows the closed-loop VTA powering itself and its loads, after being initially set into operation with outside input power by the operator, and then adroitly switched into self-powering by switching the input from external to internal at an output

Fixture allowed unit to be started into operation on external power then switched to self-powering internal power at the zero crossing of the *output* voltage The device in its fixture has a row of test points (disconnects) on every conductor going to the device

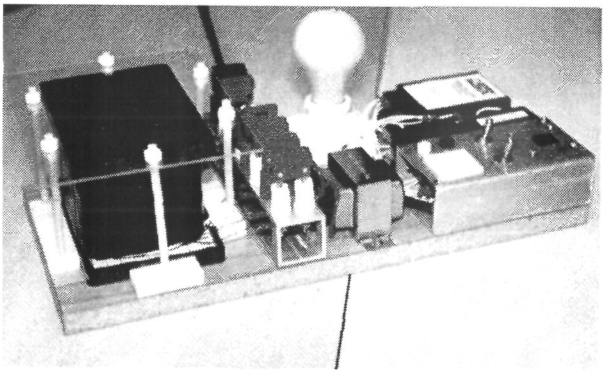

Figure 6-11 Two-magnet two-coil VTA mounted on a special test fixture built by Walter Rosenthal.

power zero crossover point. This picture shows the unit in self-powering and stable operation, taking all its input energy from the vacuum, including the power necessary to power its own internal losses and the power necessary to power its controls and loads.

Independently examined and testing observed by test engineer Walter Rosenthal

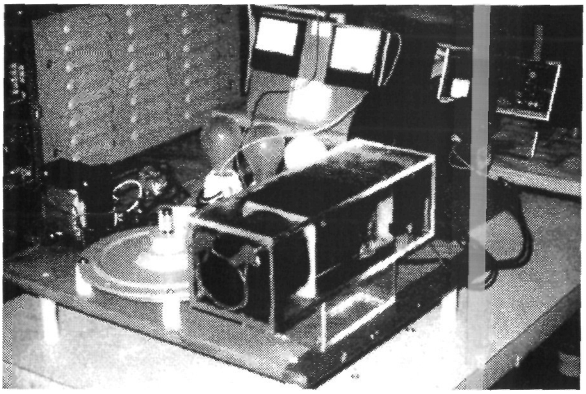

Figure 6-12 The 2-magnet, 2 coil VTA, loads, and instruments for its historic self-powering operational tests on May 1,1990.

Formal self-powering test on 1 May 1990 The loads and the VTA are powered by electrical energy extracted directly from the activ e vacuum Independently observed by professional test engineer Walter Rosenthal

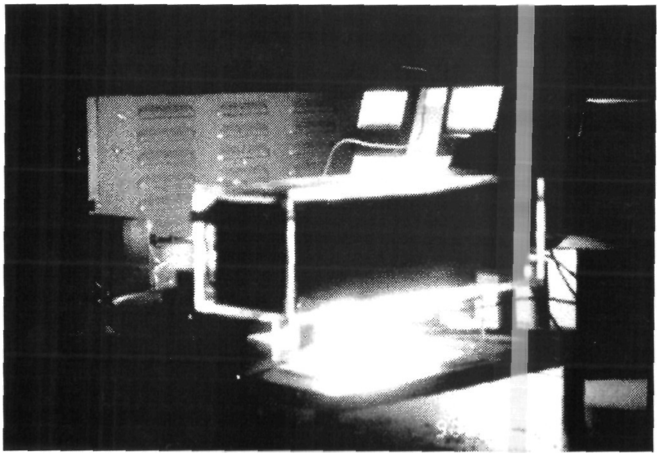

Figure 6-13 Historic self-powering operational test of the VTA May 1,1990 with no energy Input by the operator.

We strongly stress that this operation is permitted by the laws of physics and thermodynamics for an open EM system far from equilibrium in its exchange with its active environment. The active environment in this case is the combination of the active local vacuum and the active local curvatures of spacetime. The closed-loop condition is a new and stable equilibrium condition in the dynamics of the mutual exchanges between all three supersystem components: the system, the active vacuum, and the curvatures of spacetime. The *system* is in stable disequilibrium; the *supersystem* is in a stable new equilibrium condition.

6.3.2.3 Discussion

We do know that the output current of the second Sweet device was comprised of a significant Dirac hole current[166] as well as electron current — showing the expected heavy involvement of the altered local vacuum (the Dirac sea) and the curved local spacetime. The negative energy majority of the output revealed itself readily. If the output leads were shorted, ice formed instantly on the shorted leads from the moisture in the air. That is, an electrical short resulted in the sudden production of cold (negative) EM energy rather than positive EM energy.[167] That can only occur if the majority of the output is negative energy, and any sudden surge in output (as from the shorted leads) increases the ratio of negative energy to positive energy so that very strong net cooling results.

For many predominantly mechanical components (simple electrical motors, light bulbs, etc.), Dirac hole current applied to the positive lead will power the circuit equally as well as electron current applied to the high voltage (negative) side, if the grounding is simply by a conductor and not a "vast reservoir" of charge. In such case, the wire chosen to conduct or direct the Dirac hole current should be made very small. It can and will carry intense hole current and will cool rather than heat.[168] The wire to the electron side of the device should be made normal sized, appropriate to the electron current to be carried.

---

[166] We do mean hole current in the Dirac sea, consisting of Dirac holes in the vacuum itself, prior to their interaction with charged matter to form lattice holes in matter. We do not mean the standard lattice hole current at all.

[167] In local curved spacetime, Dirac holes constitute "negative energy EM fields and potentials". In such case, one can have negative energy entities and reactions analogous to positive energy entities and reactions in 3-space. On the time axis, one also has negative energy entities and currents, potentials, waves, fields, etc. Because **classical** Maxwell-Heaviside-Lorentz theory eliminates any modeling of the local vacuum or the local curvature of spacetime, these negative spacetime curvature **(negative** energy) potentials, fields, and currents have not been added into the electromagnetic theory. That of course is a serious shortcoming of the normal theory.

Several other inventors over the years have been able to activate a barium ferrite magnet into sustained self-oscillation of its material particles or domains, in similar but weaker fashion for varying lengths of time — up to a month or so in the longest time that I am aware of. Sweet's magnet would continue in self-oscillation indefinitely unless (i) the magnet was sharply shocked mechanically (as with a hammer blow), (ii) the magnet was heated to too high a temperature, (iii) the leads of the output circuit were electrically shorted, or (iv) external artificial field interference (electrical pulse shock, particularly from longitudinal stress waves) was encountered.[169,170]

Unfortunately, Sweet signed multiple conflicting agreements with several financial backers, and his work was legally entangled so deeply that it was never straightened. He eventually died, taking with him his activation secret. He was often threatened, and was once fired at by a professional assassin using a silenced rifle. He survived the incident because, being old and a bit feeble, he stumbled and fell over the front steps just as the assassin fired. The bullet snapped by his ear, right where his head had just been. The incident was reported to the FBI, but with nothing else to go on, the assassin was never caught.

---

[168] This is very probably the secret of Tesla's "single wire system". At an exposition in Nevada, Bedini demonstrated just such a system, by powering 600 watts of audio speakers at the end of a long, very thin wire across the stage from his power amplifier. If he had been using positive EM energy, the wire would have instantly melted.

[169] During that period, the Russians were playing with earthquake induction via scalar interferometers, at various places around the world, including inside the U.S. heartland. Sweet's machine would sit above a natural quake and not be affected. But when a quake within many hundreds — even several thousands — of miles was artificially induced, the Sweet device would react to it. A fairly close artificially induced quake would destroy the self-oscillation of the magnets, and sometimes a very close artificially induced quake would actually explode the intensely activated magnets. So for several years one had a way of ascertaining which quakes were natural and which were artificially induced, by simply having Sweet confirm how his magnets interacted.

[170] For those who do not believe such an activity, here is a direct statement by then-Secretary of Defense William Cohen: *"Others are engaging even in an eco-type of terrorism whereby they can alter the climate, set off earthquakes, volcanoes remotely through the use of electromagnetic waves... So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts. "* Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, **University of Georgia, Athens, Apr. 28, 1997.**

By whatever means, we do know that Sweet's magnets had to be activated by a special process that Sweet used and never fully revealed. An activated or "kinetic" magnet, it seems, was actually something a few of the "older magnetics guys" had known about, and Sweet had learned of it either from them or from Gabriel Kron. Sweet was also a transformer designer and expert, and he remarked that he had also observed specialized self-oscillation in certain transformers. Although most power engineers approached were either nonplussed or completely skeptical of a "self-oscillating field in a permanent magnet", self-oscillation of the fields, spin waves, etc. in permanent magnet materials *are* known to materials scientists today, especially to those working with magnetic effects in thin film materials {370, 371a-b}.

From a general relativity viewpoint, particle spin has a unique spin field of its own {372}, and there are such spin fields involved and operating in magnetic systems. The simple classical electrodynamic approach to magnetism and magnetic assemblies — or the view of a magnet as simply a bar with a north pole and a south pole — is quite insufficient when one wishes to research magnetic assemblies that will act as open systems far from equilibrium with their active environment, and thus will be capable of COP> 1.0 operations.

Recent work in coupled spins has revealed long-range coherence effects, previously unsuspected, as well as dual spin states participating in such phenomena {373a-b}. For example, Johnson has long used a 2-particle approach to spins and spin flipping in his own highly nonlinear magnetic assemblies {374}. His recent laboratory work has also shown similar and unsuspected *long-range coherence* effects in large multiple-magnet macroscopic systems {375} such as large assemblies of neodymium-iron magnets. This is strongly suggestive of the expected general relativistic spin field previously mentioned.

This work by Johnson is very important, since it may reveal a hitherto unsuspected mechanism for decay — of an otherwise nonconservative magnetic field around a closed loop — back into a conservative field. If one seeks an all-permanent-magnet rotating motor, a nonconservative magnetic field around a rotation loop must be established and sustained if one is to have self-rotation and self-powering. A nonconservative magnetic field around such a rotation loop means that the system acts as an open system far from equilibrium with its active vacuum and active curved spacetime environment, by means of some asymmetry mechanism such as multi-valued magnetic potential {376a-c, 377a-g}, elucidation of exchange forces {378a-c} at selected positions and directions, etc. It also requires both a stable broken 3-symmetry and a stable 4-symmetry.

On the other hand, in our opinion, Sweet employed spin waves triggered into self-oscillation in the nuclei of the barium atoms in his barium ferrite magnets. Indeed, such nuclear spin waves in magnets have been known for some time {379}. We postulate that, as in any other COP>1.0 power system, a self-rotating all-permanent-magnet motor must also involve Dirac sea hole current interacting in the various magnetic components. It is reasonable to suspect that these also interact with the electrons and their spin fields, perhaps producing the active mechanism providing that observed long-range coherence that apparently can decay the nonconservative magnetic field around the rotation loop, causing it to revert back to a conservative field and thereby stop the self-rotation and self-powering. If so, that long-range coherence must be violated and the violation stabilized. However, we could find no papers in the literature that deal with such possibilities. Hence it is a subject for further investigation by COP>1.0 researchers.

6.3.2.4 The *Highly Successful Anti-Gravity Test*

On one trip to California to work with Sweet, we noticed that, at full power output of 500 watts, his unit weighed a little less on the bench than when not powered. Struck by this anomaly, I considered the negative energy aspects of the output — as evidenced by the Sweet's demonstration of icing when the output leads were shorted — and the very high gain, which added up to a predominance of negative energy in the output.

Pondering this from a general relativity viewpoint, I reasoned that: (i) increase of positive EM energy density in space increases the curvature of spacetime to produce additional positive gravity G_P; (ii) increase of negative EM energy density in space must increase the curvature of spacetime in the other direction, and thus it must produce negative gravity G_N.; and (iii) the total gravity G_T produced in the immediate spacetime of an object producing both positive gravity G_P and negative gravity G_N, was $G_T = f(G_P + G_N)$. Thus if we increased the output of the VTA unit, we must inevitably increase the weight reduction (antigravity) effect since the predominance of the energy output was negative energy, and this output was in the neglected negative energy fields aspect as well as in the local vacuum (spacetime) itself. After several back-of-the-envelope estimates and educated guesses over some weeks, I concluded that if the efficiency of the antigravity process[171] were 10% and the unit was pushed to 1500 watts output, it would lose all its weight and hover weightless.

However, because of the nonlinear local spacetime, magnetic monopoles[172, 173] were deposited in Sweet's magnets when the unit was producing power, and the density of these monopoles varied as the output power. Increasing the power output thus increased the tensile stress of the magnets, and at some point they would explode into fragments, like a hand grenade.[174] This explained why conditioned magnets sometimes exploded when experiencing fluctuations from artificial earthquake induction testing.

---

[171] We again accent the difference between efficiency and COP. Here we are definitely speaking of efficiency, not COP.

[72] Dr. Robert Flower has confirmed that the 'symmetrized" (or "extended") Maxwell equations (Eqs. 7.1 in Springer's *Encyclopedia of Physics,* Vol. 16, p. 431) include non-zero magnetic charges and magnetic currents. Antenna engineers (Henning Harmuth, and others) know that these terms must be included for proper description of boundary conditions, even though, at the final step of computation, their integrals are set equal to zero (apparently to snuff out magnetic monopoles). For Dirac's original theory of the monopole, see P. A. M. Dirac, "Theory of magnetic monopoles," *Phys. Rev.,* Vol. 74, 1948, p. 817-830. See also F. A. Bais and B. J. Schroers, "Quantisation of Monopoles with Non-abelian Magnetic Charge," *Nuclear Physics B,* Vol. 512, 1998, p. 250-294; G. Lochak, "The Symmetry Between Electricity and Magnetism and the Problem of the Existence of a Magnetic Monopole," in T. W. Barrett and D. M Grimes. [Eds.] *Advanced Electromagnetism: Fundations, Theory, & Applications,* World Scientific, 1995, p. 105-147. Particularly see T. W. Barrett, "On the distinction between fields and their metric: the fundamental difference between specifications concerning medium-independent fields and constitutive specifications concerning relations to the medium in which they exist." *Annales de la Fondation Louis de Broglie,* 14(1), 1989, p. 37-75; — "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in Lakhtakia, A. (ed.): *Essays on the Formal Aspects ofElectromagnetic Theory,* World Scientific, River Edge, NJ, 1993, p. 6-86; — "The Ehrenhaft-Mikhailov effect described as the behavior of a low energy density magnetic monopole-instanton." *Annales de la Fondation Louis de Broglie,* Vol. 19, 1994, p. 291-301. The work of Ehrenhaft, which is now substantiated and thoroughly replicated by Mikhailov, has produced experimental evidence of the magnetic monopole in many replicable experiments. In these experiments, magnetic particles move in a magnetic field along the lines of force. Reversal of the magnetic field, H, causes a reversal of the magnetic particle's motion, which is not the case with magnetic dipoles, since that can only occur for a monopole. See V. F. Mikhailov, "Observation of magnetic monopoles in the field of a line conductor," *J. Phys. A: Math. Gen.,* Vol. 18, 1985, p. L903-L906; — "Observation of the magnetic charge effect in experiments with ferromagnetic aerosols." *Annales de la Fondation Louis de Broglie,* Vol. 12, 1987, p. 491-523. See also M J. Perry, "Magnetic monopoles and the Kaluza-Klein theory," *Am. Inst. Phys. Conf. Proc,* No. 116, 1984, p. 121-125; E. V. Smetanin, "Electromagnetic field in a space with curvature — new solutions," *Soviet Physics Journal,* 25(2), Feb. 1982. p. 107-111. [English translation]. See also T. T. Wu and C. N. Yang, "Dirac monopole without strings: monopole harmonics," *Nuclear Physics,* Vol. B107, 1976,

Having no explosive control facilities, I warned Sweet not to push the unit past 1,000 watts, for safety reasons. At my urging, Sweet made a new output load of ten 100-watt lamps in sockets, so the load could be varied from 100 to 1,000 watts easily.

---

p. 365-380; — "Dirac's monopole without strings: classical Lagrangian theory," *Phys. Rev. D.,* 14(2), July 15, 1976, p. 437-445.

[173] We emphasize that a magnetic monopole can be taken simply as a highly localized magnetic scalar stress potential. See, e.g., P. A. M. Dirac, "Theory of magnetic monopoles," *Phys. Rev.,* Vol. 74, 1948, p. 817-830. Dirac pointed out that the existence of a single magnetic monopole in nature would allow an explanation for charge quantization. We believe that the reverse of Dirac's argument may be true also: The existence of charge quantization should allow an explanation for the magnetic monopole. However, we also surmise that a magnetic monopole exists only in (and as) a sharply localized curvature of spacetime. Jackson, *Classical Electrodynamics,* second edition, 1975, p. had this to say about magnetic dipoles and Dirac's argument: *"Dirac's argument... is that the mere existence of one magnetic monopole in the universe would offer an explanation of the discrete nature of electric charge. Since the quantization of charge is one of the most profound mysteries of the physical world, Dirac's idea has great appeal. "*

[174] Sweet in fact did explode several magnets by pushing the power output of the device. Pushing the output was easy (and is easy for any unit outputting mostly negative energy), because all one had to do was connect additional load.

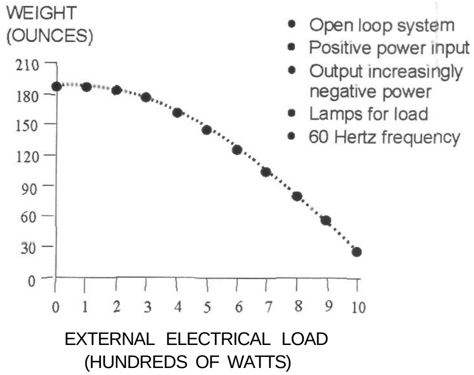

Figure 6-14 **Results** of the"pushed"VTA **antigravity test:** Weight **of VTA on the bench versus power output.**

Sweet completed the new load section and performed the experiment in California after 1 had returned to Huntsville, reading off the results to me over the phone. I took down the data, and then plotted the curve. Figure 6-14) shows the beautiful results. At 1,000 watts the unit had lost 90% of its weight. This was electrifying, because it was an unprecedented and highly successful antigravity experiment. Even my original "rough" assumption of levitation at 1500 watts was not too far from the indicated 1250 watts or so from the actual curve.

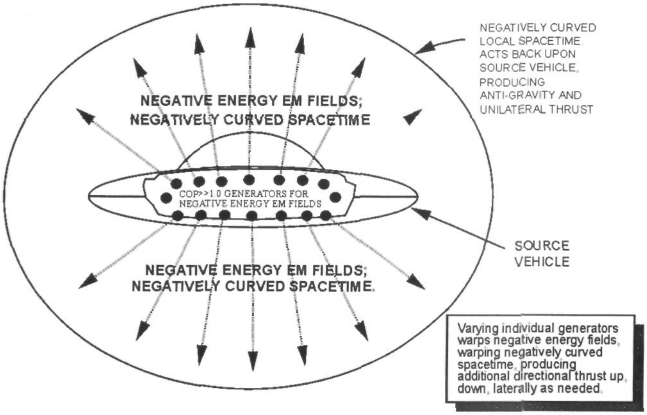
NEGATIVELY CURVED LOCAL SPACETIME ACTS BACK UPON SOURCE VEHICLE, PRODUCING ANTI-GRAVITY AND UNILATERAL THRUST
NEGATIVE ENERGY EM FIELDS; NEGATIVELY CURVED SPACETIME
COP>>1 0 GENERATORS FOR NEGATIVE ENERGY EM FIELDS
SOURCE VEHICLE
NEGATIVE ENERGY EM FIELDS; NEGATIVELY CURVED SPACETIME.
Varying individual generators warps negative energy fields, warping negatively curved spacetime, producing additional directional thrust up, down, laterally as needed.

Figure 6-15 Proposed mechanism for practical antigravity.

Figure 6-15 shows my proposed mechanism for antigravity. We simply use negative EM field energy generators to produce a negative EM energy field in space surrounding the source vehicle. The associated positive energy from the same energetic power source can power the vehicle simultaneously. The local reversed curvature of spacetime now interacts back upon the mass of the vehicle, to produce an antigravity component. By adjusting the output of the power source, the amount of negative energy — and hence the amount of antigravity component — can readily be varied and controlled. By controlling the directions of the field components, one can have unidirectional antigravity force, or antigravity and thrust force vectors simultaneously. Hence such a vehicle could be maneuvered and controlled, and be self-powering. This at least was and is my concept of a practical antigravity technology.

6.3.2.5 Later Developments and the End of the VTA

Later I prepared a paper on the VTA, including the antigravity experiment's results. We succeeded in getting the paper published {380} but I deliberately did not include the real details of the negative energy aspects as I wished to withhold that information for patenting. I prepared the entire paper, but placed Sweet's name first, as was appropriate since he was the inventor of the VTA.

Unfortunately — but understandably! — Sweet became a bit paranoid after the sniper assassination attempt (he was shot at by a distant sniper using a silenced rifle), and then would never allow a formal demonstration of his antigravity test. I strongly wished to present it to a major national U.S. scientific committee for independent testing and confirmation — which could have been done without Sweet having to reveal his activation process for the magnets. However, Sweet would not hear of such. Indeed, he fearfully kept the antigravity capability a secret from most of his backers. He strongly believed that he would be killed almost immediately if he revealed our antigravity experimental success. He may have been correct.

In fairness, the inventor and his backers of necessity have different priorities from those of scientists doing open research in a university. But at least a record of the test results is in the literature, though certainly not as rigorously or as extensively as I wished. Sometime later, I withdrew from any further association with the VTA unit, after the backers changed, a lawsuit broke out, and the entire project unraveled. Sweet later died without ever revealing his full secret for such powerful activation of self-oscillation in permanent magnets at low ELF frequencies.

The loss of the Sweet vacuum triode was a severe blow to science and a real setback to the scientific progress that could already have been made. Nonetheless, I contributed what I could, documented what I could, and made a videotape of the operation of the open-loop 500-watt device in the power mode. I also made a video of as much as Sweet would reveal of his activation process. I was not allowed to videotape a second antigravity experiment, even though I frequently and strongly requested it. By that time, Sweet mortally feared for his life. Perhaps what we have written in this book will be sufficient to provide some needed clues to some future graduate student or post-doctoral scientist who will succeed in recovering the necessary activation of magnets and both the COP»1.0 power capability and the antigravity experiments.

Meanwhile, I have filed a provisional patent application on the mechanism for antigravity, the gist of which I worked out solely on my own back in 1971 at Georgia Institute of Technology, and that I include in Chapter 8. As far as I am aware, no one else had thought of it previously, or has thought of it in that exact form *{using the causal positron prior to observation, and while it is still a negative energy, negative mass electron, as the source charge for negative EM field energy, and also using the causal positron's corresponding giant Heaviside nondiverged negative energy flow for substantial negative curvature of spacetime.).*

6.3.3 Intense Discharges in Ionized Gas

Intense sudden discharges in ionized gases are especially of interest due to the presence of optical frequency components and the involvement of iterative optical retroreflection etc. These processes seem to be involved in several investigations and inventions. Further, it appears that such intense discharges do produce bursts of negative EM energy as part of their overall discharge energy.

In the case of plasmas, some are known to transduce an input transverse EM wave into an output longitudinal EM wave, and vice versa. We know now that such a longitudinal EM wave is also accompanied by a time-polarized (scalar) EM wave in the time domain. The connection with the Whittaker 1903 decomposition of the potential, and with the new 4-symmetry negentropy law, is apparent. Several COP>1.0 researchers — notably Shoulders {381a-f} and Paulo and Alexandra Correa {382a-e} — have performed extensive research into anomalous discharge phenomena, and also have developed experimental devices capitalizing on the novel effects available.[175, 176] The Correas' U.S. Patent No. 5,449,989 includes many tests of a Correa glow discharge apparatus, one of which in 20 minutes of operation produced 0.988 kilowatt-hour of energy output for an input of 0.258 kilowatt hours. Using the known negative resistor characteristics, the Correas achieved sustained self-oscillations in the discharge.

As an aside, we point out that the well-known anti-Stokes emission — which produces more energy output than the experimenter inputs, and hence exhibits COP>1.0 — has long been validated in certain gases such as aniline vapor {383} and H_2 {384}. Stimulated anti-Stokes radiation occurs in many solid materials such as oxides and doped oxides {385}. Assuming the excess energy comes from the internal energy of the molecules, this is a prime candidate for free replenishment of that furnished internal energy of the molecule, via the broken symmetry of the molecular dipoles in their fierce virtual energy exchange with the local active vacuum and local curved spacetime. Of course, conventionally the energy furnished by the molecules, etc. is not replenished from the vacuum, hence anti-Stokes emission can be used for coherent light cooling effects {386}. Apparently the Correas' process accomplishes the energy replenishing, hence the COP> 1.0 performance over a 20 minute period.

Formation of longitudinal EM waves in the plasma forms in certain gas tubes, such as in Priore's gas-filled tubes {387}, may also occur in the Correa experiments and probably do. Longitudinal EM waves, of course, always involve their concomitant, accompanying time-polarized EM waves, since the combination comprises the scalar potential and hence the local curvature of spacetime provided by that scalar potential. The replenishment mechanism involved in longitudinal EM waves and scalar potentials — time-energy transduced into 3-space energy — apparently is

---

[175] Both Shoulders' electron cluster discharges and the discharge phenomena of Paulo and Alexandra Correa have been independently verified. Fox and other scientists have verified COP>1.0 in Shoulders' work, and Mallove has formally observed a self-powering Correa system. Both these highly reputable scientists are well known, have the highest integrity, and I would trust either of them with my life. Further, we point out that strong gradients — as exist temporarily in strong discharge phenomena — are already known to violate thermodynamics. That is one of the areas being researched today to advance thermodynamics, under the aegis of "extended thermodynamics". See Appendix A.

[176] In addition, the Correas have demonstrated a form of energy that demonstrates latent heat. See discussion in Paragraph 6.3.9.

invoked and utilized in the Correas' sustained self-oscillations in the plasma discharge.

6.3.4 Shoulders' Electron Cluster Processes

Valone {388} has summarized Ken Shoulders' charge clusters and processes as follows:

> *"An unusual energy source is the clustering of electrons by a discharge needle into a high density bundle equaling Avogadro's density of a solid.... Ken Shoulders has patented a process (Pat. #5,153,901) that produces electron clusters with such high energy density, they equal processes exceeding 25,000 degrees Celsius upon impact. Yet, he only uses 20 microjoules to produce the effects. The clusters travel at a maximum of one tenth the speed of light and penetrate any substance with accuracy and sharp precision. It is similar to xenon clustering techniques currently used at megavolt energy levels. Low energy nuclear transmutation of the target has also been achieved with this process. Using a deuterium-loaded palladium foil, only the bombardment areas show transmutation into silicon, calcium, and magnesium with electron clusters upon analysis with X-rays. Jin and Fox {389} have postulated that the high velocity electron clusters achieve results similar to accelerators, including penetration of the nucleus, with substantially less power. The new physics of like-charges clustering in bundles under low power conditions opens a wide range of applications including spacecraft maneuvering microthrusters. Overunity efficiency* [sic: should be COP] *is 9 to 1, or higher."*

Ken Shoulders is an accomplished experimental scientist who has discovered and patented many aspects of electron charge clusters, including their COP>1.0 characteristics {390}.

According to his experiments, one of the most remarkable characteristics of these charge clusters seems to be their dramatic reduction of the permittivity of space {391}. Shoulders has noted the remarkable removal Of **positive** charge during the self-formation process of the clusters {392}. In our opinion, the involvement of the supersystem of these charge clusters is shown by that phenomenon. We hypothesize that an accompanying Dirac hole current is shown by the experimental occasions where — in addition to the Shoulders' negative charge clusters — additional charges unite but show effects of having opposite charge {393}. Such changes would change the permittivity of space, which Shoulders has noted. They could also create a time-reversal zone (TRZ), which would account for the clustering of like charges since in a TRZ the law of attraction and repulsion of charges is reversed. In such a zone, like charges attract and unlike charges repel. This could explain why Shoulders has observed both negative charge clustering and positive charge clustering.

This would also explain the remarkable separation of the positive charges from the negative charges — they are simply *repelled* by the negative clusters in a TRZ.[177] This of course immediately involves curvature of spacetime effects, including local changes in the speed of light etc. As stressed by Shoulders, the entire area cries out for extensive and sophisticated experimental investigation on a continuing basis, as well as some powerful new theoretical work. Shoulders has also noted both "black" and "white" (dark and bright) clusters, including their combination and separation, strongly suggesting Dirac holes as well as the electrons in the charge clusters. He has proposed a form of "inertial propulsion" based on some of the mechanisms and asymmetrical forces observed in his experiments, and such asymmetry is possible in curved local spacetime conditions.

The relation of Shoulders' effects to cold fusion phenomena was shown by an experiment in which a deuteron-loaded palladium cathode was impacted by a charge cluster {394}. An explosive-like reaction occurred, and an X-ray analysis of the impact crater showed a number of elements not seen in the nearby palladium surface away from the impact site. Elements such as oxygen, calcium, silicon, and magnesium were detected in the impact site, but not in the surrounding palladium lattice. We would suggest the presence of TRZ-based nuclear transmutation reactions (via quasi-nucleus formation followed by quark-flipping in nucleons as the

---

[177] Since the same TRZ phenomenon generates specific nuclear transmutation reactions at low spatial energy, explaining the anomalous transmutation of elements in cold fusion electrolyte experiments, a strong indication exists that it is a viable mechanism. In our mechanism proposed for the flow of time, a TRZ is produced for charges whenever antiphoton interactions predominate over photon interactions. This indeed implies a change in the permittivity of the local space, local curvature of spacetime, and local changes in the speed of light — all phenomena Shoulders notes in his experiments.

TRZ decays to a time-forward zone). Additional information on this charge cluster phenomenon has been given by Fox {395}.

Other Shoulders patents are pending and his research is continuing. It is our understanding that at least one application is presently being developed for the market.

6.3.5 Mills's Hydrino Energy Cells and Developments

Dr. Randell Mills {396a-396c}, {397} has proposed a cosmic sea of hydrinos (the hydrino is a proposed reduced energy state of the hydrogen atom) to explain his tested energy cells. He envisions — and produces a theory for — hydrogen that is reduced to a lower energy state (the "hydrino"). He has gathered considerable evidence for his theory, and has accounted for some sixty or so infrared spectral lines in the sun's radiation that have previously not been explained. Independent tests appear to have validated his energy production claims.

Several major companies are presently working with Mills and his parent company, BlackLight Power, Inc. Progress can be followed from the company website, www.blacklightpower.com. Developments are ongoing in light, power, plasma, and new compositions of matter. A technical book by Mills {398} is freely downloadable from the website, giving his overall theory, as are other publications by Mills *et al.* {399a-h}. Critical objections to various parts of Mills' theory have been given, e.g., by Zimmerman {400}, with Mills responding strongly and — in our opinion — successfully.

Certain concepts utilized by Mills are, in our opinion, valid and insightful - whether or not the classical quantum mechanics developed by him is found to hold, and whether or not it agrees with ordinary quantum mechanics (which has problems of its own). The point is that *the experiments work.* It is thus the job of the theorists to find what changes must be made to the theory for it to remain consistent with experiment. If they do not like Mills' theory, the proper procedure is to advance one that **is** consistent with all the experimental results, and also to show where Mills' theory is inconsistent with the results.

In our own view, an *observed* positive charge is the output of an ongoing time-reversal process and therefore represents a "time-reversed entity" as observed. Before observation (imposition of *d/dt* operator), the positive **charge** is thus a negative 4-charge comprised of *qt*, not *q*. From that viewpoint, the H+ ion, which is just a hydrogen nucleus (a proton) without an electron shell, may be regarded as a time-reversed 4-entity prior to observation. Further, any energy change in and of the hydrogen atom results in a change in the local energy density of spacetime (ST) *a priori,* hence a change in the local ST curvature.

If we regard the entire hydrogen atom as a set of such ST curvatures and their dynamics, then Mills' consideration of a lower positive energy state *hydrino* simply implies the addition of an additional negative ST curvature to the ST curvature set associated with the normal hydrogen atom. We note that binding energy of the nucleons is already negative energy, and thus just such a negative spacetime curvature. But the binding energy is itself a spacetime curvature *capable of being increased,* as can be any spacetime curvature. In the supersystem, addition of such ST curvature is possible, and it is *doable* electrodynamically, particularly when Heaviside's normally unaccounted huge energy component is accounted and utilized.

In our opinion, Mills' theory should probably be recast into a higher group symmetry electrodynamics — such as O(3) — which has been shown to be intimately connected with Sachs's unified field theory. This would provide an adjusted model that is not only testable but also *engineerable.* Nonetheless, the concepts of Mills' theory and the theory should be closely examined as they stand, and at least some of the key concepts are engineerable as witnessed by his successful experimental results. That is better than what the critics have offered! A higher group symmetry recasting would enhance and expand his theory even more, and correct any errors that might be present. It would also allow a full theoretic treatment (i.e., a supersystem treatment) of Mills' main concepts, which we believe have much merit.

Further, given Mills' hydrino as we approached it, two hydrinos will each possess such a set of spacetime curvatures (each will possess such a *vacuum engine).* Later in this book — and elsewhere {401} — we have pointed out that, in a time-reversal zone (TRZ), the law of electromagnetic charge attraction and repulsion is reversed. Accordingly, in a TRZ two or more hydrinos will attract each other so closely that each enters the now-reduced strong force region of the other, forming a quasi-nucleus. As the TRZ decays away back to a normal time-forward zone, because of their power law form the strong forces restore back to normal much faster than the electromagnetic forces do. Hence during its "decay", the quasi-nucleus is energetically compacted into a permanent nucleus, either flipping quarks or giving off one or more particles as necessary to balance its new energy state in the process. We have previously argued that a similar process where a TRZ induces formation of quasi-nuclei also yields the major transmutations {402} demonstrated in a large number of cold fusion experiments worldwide.

The present author has previously proposed an engineerable mechanism (the totality of all photon interactions with a mass) that generates the flow of that mass through time {403}. By manipulating the arrow of time between forward vector and reversed vector, the strong forces in a nucleus (or for a particle) can be directly manipulated in magnitude (and to some extent, even in sign although the latter involves gluon and quark manipulations in the nucleons). By adroit use of these manipulated strong forces and TRZ formation and decay, the energy states of nuclei can indeed be manipulated, including reduced in energy level —just as Mills assumes and presents experimental evidence for, and as cold fusion experiments present experimental evidence for.

We note that two such interacting hydrinos must each possess a specific vacuum engine (specific set of ST curvatures and dynamics). Novel new interactions are available in TRZ's because of manipulation and even reversal of electromagnetic forces, strong forces, etc. and because of nuclear fusion of like charged particles into quasi-nuclei followed by decay into stable nuclei, etc. These novel new particle interactions caused by manipulation of strong spacetime curvature engines and their magnitudes are presently missing from particle physics. Mills, however, does come at them from a somewhat different direction in his own concepts.

The reason for the absence of dominant general relativity effects in most particle physics is that the "total photon interaction" mechanism generating the flow (persistence) of a mass or field through time has not been recognized and used. Consequently, in general only those far weaker ST curvatures due to the weak G-force have been considered. To Mills' great credit, he is aware of the gravitational significance of manipulating much stronger forces and spatial energy densities such as the electromagnetic fields and their energy densities.

The U.S. Patent and Trademark Office then refused a new patent to Mills on one of his processes, and revoked another Mills patent even after already agreeing to award it within two weeks. BlackLight Power, Inc. initiated and lost a lawsuit against the Commissioner of Patents, seeking to reverse this unusual action by the U.S. Patent Office. However, the company's patent portfolio has continued to grow, with extraordinary international applications {404a, 404b}, and several European patents have been filed appropriately. BlackLight Power Inc. seems well funded (some millions of dollars) and well staffed, and we expect to see BlackLight Power, Inc. products entering the market place. Those products will probably be related to fuel cells, plasmas, power converters, generators, separation methods, and novel compounds materials.

and

6.3.6 Switching Circuits with Extended Electron Gas Relaxation Time

See Figure 6-16. Suppose an instantly applied voltage to the high side of the circuit instantly potentializes a pair of circuit conductors, with the other acting as ground. The charges in the conductor are thus instantly potentialized and excited, so that through the external circuit there exists an E-field E, where $\mathbf{E} = -\nabla\phi$ between the high side of the circuit and the ground return side. An acceleration force on the energized electrons thus exists. It requires a finite time for the energized electrons in this force field E to break their static friction or inertia, move, and accelerate to the requisite slow drift velocity down the wire. During that delay or "electron relaxation time", no current is initially in progress, then a little current is flowing, and finally all of the current is flowing. This delay time is called the electron gas relaxation time.

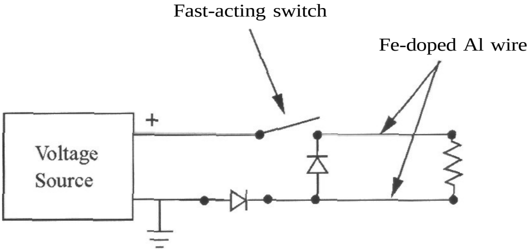

Note: May use many strands of Fe-doped Al wire instead of one, for each conductor

Figure 6-16 Extended-relaxation-time circuit with diodes.

In the usual copper conductor, the delay time arguably may be on the order of 10^{-16} sec. This is so slight a delay that it can be ignored, for all practical purposes. Current — with resulting dissipation of the collected potential energy — usually occurs "as soon as" we apply the voltage, for all practical circuit work. However, suppose we could increase that electron gas relaxation time to, say, one millisecond? Then for an appreciable number of microseconds, there would be essentially no current flowing at all in the now-potentialized circuit.

We previously proposed a mechanism whereby a degenerate semiconductor alloy (say, of a tiny bit — perhaps 2% or so — of iron alloyed in aluminum wire) is utilized for the conductors of the external circuit. This should give a vastly extended electron relaxation time, somewhere on the order of a millisecond. By obtaining a millisecond, one can excite the circuit with current-free potential alone, which becomes pure regauging during the electron relaxation period.

We strongly point out that this is an asymmetrical regauging; the E-field previously mentioned is present, but the electrons are still frozen in the circuit since their relaxation has not begun. The gauge freedom principle, recognized in quantum field theory, assures us that we do not have to perform work to simply inflow potential energy to change the potential energy of the system. In real circuits, we may have to pay a tiny bit for switching to connect and disconnect the external potentialization source, but we can be at pains to make that very efficient and minimal {405}. The point is, other than minor switching costs, the dramatic change in the potential energy of the system is cost-free and no current is drawn from the external potential source. Hence — in the language of the electrical engineer — we draw no power from the source of potential, during this potentialization time.

Suppose we switch away the excitation source prior to its decay while the electrons are still frozen. Simultaneously, we complete the "freed" potentialized circuit with a strategically placed diode, as shown in Figure 6-16. Again, we pay a little switching cost, but it can be minimized. The electrons in the *now asymmetrically potentialized circuit* are still frozen, so no power has yet been drawn by the system from the external power supply furnishing the voltage only. In this way, pure asymmetrical regauging is used to excite the circuit with additional potential energy, without requiring work (except minimally for switching).

A little more time passes, and suddenly the electrons in the potentialized circuit wake up and move. We stress again that the full emf is acting on the electrons, and now they are finally free to move. There is a net nonzero regauging E-field (force field) also. The excited closed current loop circuit then discharges in normal Lorentz symmetrical fashion, killing the source dipolarity as current is driven back up through the diode, but all the work in the load is "free". The circuit specialist will also recognize the usefulness of additional components associated with the shunt, such as a capacitor or even an LC oscillator for AC work.

If L*E* is load work obtained and *SE* is switch energy utilized, this approach yields COP = *LE / SE* and COP > 1.0 is possible. The reason is that the *primary* source dipole providing the potential for potentialization of the electrons and regauging of the circuit is removed from the closed current loop circuit and the effect of current through the back emf. Thus the primary source dipole is not destroyed faster than the load is powered. It is only depleted by the small amount of switching energy we utilized. This appears to be a major way to avoid the COP<1.0 limitation of the usual closed current loop circuit.

This is a circuit with legitimate COP>1.0 capability, without violation of the laws of thermodynamics or physics. We break the Lorentz condition between the initial "external circuit" and the primary source dipole for the regauging energy. After regauging, we "cut loose" the circuit with its own primary dipolarity and newly transferred "separate source dipole". *By such adroit use of electron gas relaxation time, we can use one source dipole to freely make another of equal intensity, without any depletion of the first.*

The difficulty is in getting the Fe-Al alloy wire, which must be made in an inert atmosphere since Al corrodes easily and the two elements have very different melting temperatures.

Alternatives can be visualized, such as to utilize a capacitor whose plates are made of the Fe-Al alloy. This might be useful in the triode-capacitor, to be discussed next. A series of pulses might then be used to charge the capacitor, without depletion of the primary source dipole in the external source of potential, if it is repeatedly switched in and away. An interesting question arises if such a capacitor with tailored "hysteresis" in its charging is utilized in various circuits and schemes containing a coil of useful time delay also, particularly since the electron gas relaxation phenomenon involves decaying oscillations. Interesting phase angle results can be constructed in such circuits, but we leave the resolution of this question to future experimenters.

6.3.7 The Triode-Capacitor and Other Curious Components

See Figure 6-17. In this rough concept by the present author, the notion is to implant a grid of very small copper wire mesh inside the dielectric and near one or both of the capacitor plates, and then to utilize grids as if they were grids of a triode tube, for gating and controlling the displacement current in the capacitor and the dielectric strain. As an example, when charging the capacitor, the grids can be used to increase the displacement current (which also increases the voltage on the capacitor and dielectric strain). This increases the charge and energy in the charged capacitor, without putting in as much "enhancement and gating" energy as the extra energy obtained in the charged capacitor. If desired, we can also incorporate the 2% Fe-doped Al wires in the grid, or connected to it, to further move the grid toward a "pure current-free potential" operation.

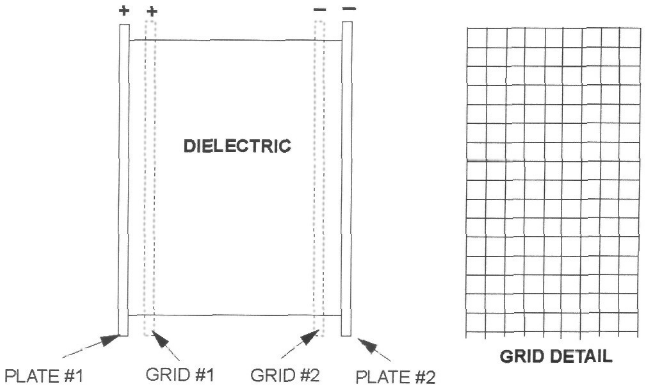

Figure 6-17 The triode-capacitor concept.

The scheme is that a voltage on this grid will intercept only a small current during charging of the capacitor, but can have an appreciable voltage upon it. The charging potential is placed across the capacitor, and simultaneously the grid has a similar potential placed upon it. The external source of charging potential is then disconnected, and the capacitor charges with more energy rearranged in its circuit than the energy one dissipates in the grid. The similarity to an old vacuum tube triode is apparent. For the purist, additional grids with tetrode and pentode functions can also be added.

One or more inventors has unwittingly tried to approach this in one way or another, but have not directly incorporated this schema and have not used the grid. They do not seem to have obtained positive results with those different schemas without the grid, nor have they specified the difficulties and phenomenology encountered, etc. We were curious about possible results of combining this (the triode grid in the capacitor) with the known elimination of capacitor losses by appropriate pulse charging {406a-b}. It seems that experimentation and research with the triode capacitor and its variations is definitely indicated.

We were also curious about the Davis non-inductive resistor {407a}, which is a special kind of "capacitor" that becomes an inductance-free resistor. One may argue that the resistor may be multiply inductive after all, though it is *net* noninductive. It may well be "equally and oppositely inductive," which is quite a different thing altogether, since it produces an artificial stress potential (and in the presence of AC, it produces a longitudinal EM stress field wave) in spacetime even though it produces a *net* zero reactance. In short, it produces a Lorentz-regauging, directly, and this can be a varying Lorentz regauging producing longitudinal EM stress waves. Whether this odd characteristic can be utilized in a COP>1.0 process remains to be seen. Preliminary indications are that it can be so utilized.

The various versions of the biwound coil, wound with two conductors at once and then the currents passed through the conductors in opposite directions, also presents interesting capabilities. Here again, it is multiply inductive, but may be tuned and adjusted to be *net* noninductive. We strongly suggest comparison with the well-known hairpin dual dipole antenna.

In short, this biwound coil arrangement — with various variations — can produce (and modulate) a magnetostatic scalar stress potential in the local vacuum and, in the case of AC, a longitudinal EM wave. This would appear to be a direct change (and oscillation) of the stress potential of the vacuum itself, which can be decomposed via the Whittaker process into a harmonic set of phase conjugate longitudinal EM wavepairs. The potential longitudinal EM wave communication possibilities can be seen, although one might require a Fogal transistor to receive the longitudinal EM wave modulations of such signals.

Further work in this area of unusual capacitors and components is left to the interested reader for his or her experimentation and research.

6.3.8 A Surprising Thing About Thermodynamics and Reservoirs

Thermodynamically, it is fashionable to state that one cannot take energy from a reservoir at constant temperature. That is true for *equilibrium* conditions, but not for *nonequilibrium* conditions {408}. That is, we cannot take energy from a reservoir *in equilibrium* at constant temperature. We can indeed take energy from a reservoir at constant temperature *but not in equilibrium.* More exactly, we can take energy from a nonhomogeneous reservoir at constant temperature.

We quote Hsu-Chieh Yeh {409} for a vivid statement of this little-recognized fact:

> *"From Planck's statement of the second law of thermodynamics it is generally inferred that it is impossible to construct an engine which produces work at the expense only of heat taken*

> *from the air or the ocean. ...[It is demonstrated that]... when the air and the ocean are combined as a nonhomogeneous reservoir of uniform temperature, it is possible to construct an engine which produces work by extracting heat from the said reservoir. This does not constitute a violation of the second law of thermodynamics, rather that the "reservoir" in the Planck's statement must be clearly stated as being in equilibrium."*

Yeh's article falsifies the general, mistaken inference that Planck's statement of the second law of thermodynamics requires it to be impossible to construct an engine that produces work at the expense only of heat taken from the air or ocean. The experiment described in Yeh's paper demonstrates that when the air and the ocean are combined as a nonhomogeneous reservoir of uniform temperature, it is possible to construct an engine that produces work by extracting heat from the said reservoir. This does not constitute a violation of the second law of thermodynamics, but rather shows that the "reservoir" in Planck's statement may be in equilibrium or disequilibrium, and the statement holds only for the equilibrium case. Also, an entire reservoir can be in "overall" equilibrium, but in raging disequilibrium between various parts of itself. Classical thermodynamics only applies with the reservoir or part of it that is used by the system in its supersystem, when that part is in equilibrium.

We bring this out to show the analogy to our concept of the two additional components of the supersystem as the "reservoir" or external environment of the system itself.

We also point out that the concept of temperature itself is not strictly defined under disequilibrium conditions, as pointed out by Gabriel Laufer {410}.

The proof and a schematic diagram of a machine to do just such an energy extraction process at constant temperature in an environment in disequilibrium are shown by Yeh {408}.

6.J.9 Latent Heat Work of the Correas

In addition to their glow discharge work previously discussed, Paulo and Alexandra Correa have demonstrated a method to optimize some anomalous form of energy from the sun, which transduces into heat, producing an anomalous T_0 - T difference registered in Orgone Accumulators[178] invented by W. Reich in 1939-1940. They have also employed the optimized difference in this "latent" heat — their term — to drive a low dT Stirling engine of the MM6 type, to perform free work {41 la-b}. The Correas employ their own new theory of the aether {412} to explain the effect.

---

[178] **This** is Reich's terminology, not used in conventional physics. We suspect that **proper** application of quantum field theory with its four photon polarizations could

Against the assertion that this is just normal solar energy warming, they have demonstrated the successful driving of the Stirling engine for several hours during nighttime {413}. With their latest improved version, they report speeds of the Stirling engine of 150 rpm during daytime (with dT values reaching 22.5° C), and speeds of 30 to 80 rpm during nighttime (with dT values of 3.1 to 9° C). There appears to be no obvious standard thermodynamic explanation or standard theory to explain how the anomalous heating occurs — although once it does occur, the explanation from then on is thermodynamically normal.

The Stirling energy is driven on a difference in temperature created between the top plate and bottom plate of a box by this anomalous energy. It is the production of the heat that is anomalous, not the driving of the engine. For comparison, imagine a bank of solar cells that would somehow operate at night, when no normal radiation from the sun was striking them. The Correa effect demonstrates such an anomaly.[179]

Dr. Eugene Mallove, an excellent scientist with impeccable credentials, has personally observed the Correa accumulator system driving the Stirling engine, and has examined it inside and out {414}. Though small, the heat and the motive power are real and the engine is really driven, in the absence of any kind of conventional batteries or other known power source.

be applied to transform the notion of "orgone" energy into specialized EM energy. We further suspect that orgone energy is really the transduction of the time-polarized photon energy into normal photon energy. We are assured by quantum field theory and the giant negentropy solution to the source charge problem that the instantaneous scalar potential involves this process.

[179] We point out that local regions of the sun may well produce Aharonov-Bohm effects, and probably do. Such reactions in the sun would localize some of the B-fields and emit curl-free magnetic vector potential radiation into space, striking the earth. Close examination and analysis of the possible reaction of uncurled A-potential radiation with the accumulator box and plates in the Correa process might yield a novel mechanism for the observed heating (scattering) effects, including the difference between daytime and nighttime levels, periodic variations, etc.

While the Correas utilize some of Reich's orgone and aether terminology, they are quite aware that the energy being converted is not normal transverse EM wave energy. Quoting {415}:

> *"Aetherometry demonstrates that what traverses space is not transverse electromagnetic radiation (and certainly not sensible heat), but longitudinal electric ambipolar radiation emitted from the sun. All that electromagnetic radiation consists of is a local production of photons (for those who care to listen): aether energy is not composed of photons, actual or virtual. "*

We can generate sympathy for that statement, although we would still not rule out time-polarized (scalar) photons and longitudinal photons from quantum field theory, each of which is individually nonobservable, but when paired the two comprise an instantaneous scalar potential. As we previously discussed, the classical notion of the EM force fields E and B implicitly assumes interaction of the unseen causal "field-as-it-exists-in-massfree space" with charged mass having been accomplished, with the E and B being the 3-effect of this interaction after observation. Thus these force fields E and B are 3-effects (after interaction and after observation) rather than 4-causes prior to interaction and observation. We suspect that the Correas are trying to get at the causal 4-photon, rather than the usually assumed 3-photon. Further, a virtual photon spends much of its time existing as a dipole comprised of a virtual electron and a virtual positron. Therefore it spends much of its time in the presence of the giant negentropy 4-circulation, involving transduction of time energy into spatial energy and vice versa. It is this negentropy circulation represented by the virtual photon (or accompanying it, whichever one chooses) that the Correas are confronting as the "aether energy".

As previously discussed, an analysis of the Drude electrons in the electron gas of a detector builds a convincing case for EM waves in space existing in longitudinal EM waves, *if the spinning and longitudinally restrained electrons precess and act as gyros.* The detected transverse EM waves are in fact transverse precession waves of the gyro-electrons when perturbed by incoming longitudinal EM waves (longitudinal photons appearing in 3-space from the time-domain's scalar photon couplet). As Feynman put it {416},

> *"We may think of E(x, y, z, t) and B(x, y, z, t) as giving the forces that would be experienced at the time t by a charge located at (x, y, z), with the condition that placing the*

> *charge there did not disturb the positions or motion of all the other charges responsible for the fields."*

The E and B force fields — and transverse photons — are not the form in which EM energy *causally* exists in 4-space prior to its interaction with charged mass. In that causal form, EM energy seems to exist in field-free form, as a direct curvature of spacetime — at least in Sachs's unified field approach. Further, this form may be represented as the causal precursor of longitudinal EM waves accompanied by time-polarized EM waves.

We will not repeat our past discussion of the great necessity to change the very notion of "propagation through 3-space", but that is the issue. In our opinion, the Correas have had to struggle with the use of this causal EM energy as it exists in 4-space prior to interaction with charge and prior to observation. Hence their view is solidly based that the photon does not exist until after the interaction — which in fact has much support once we remove the still-present assumption in classical electrodynamics that a material ether exists.

Our own recommendation is that this anomalous energy should be deeply investigated with the view that all EM energy in 3-space comes from the time-domain and the 3-space aspect is in terms of longitudinal photons, as we have proposed in our solution to the source charge problem previously cited. That is the only way that the nature of the EM wave in vacuum can be reconciled with quantum field theory's findings {19} and with Whittaker's decomposition of the scalar potential {85}, slightly corrected {12}.

Mallove also points out {417} that, if this aether energy view of the Correas has validity, it *"...suggests that physics is not just wrong, it is profoundly wrong — catastrophically wrong about even the most fundamental questions of existence, such as: 1) Is there an energetic aether? and 2) What is the nature of light?"*

We profoundly agree with Mallove's observation, and we believe we have explained the reason why physics is indeed profoundly wrong (widely substitutes effect for cause). Further, classical physics is decidedly wrong about the nature of light — from quantum field theory, experimental proof of superluminal communication, and action-at-a-distance effects.

In our view, the Correas presently have a replicable experiment that, once independently and thoroughly investigated, may force a deeper examination of the foundations of physics. Hopefully, the Correas' experiment and others like it will finally force the separation of effect from cause, presently so mangled in physics and particularly in electrodynamics.

If the National Academy of Sciences and the National Science Foundation can be convinced of the horrible manglement of cause and effect in physics and especially in electrodynamics, perhaps they may yet wake from their long slumber and get the foundations corrected to eliminate such non sequiturs. If so, we shall have a magnificent new physics as a result.

If not, then the NAS and NSF will continue to profoundly reduce the security of this nation, since several nations have already made and weaponized such corrections to physics to produce what is called "energetics". Further, extraordinarily powerful energetics weapons have **already** been developed and deployed (and even employed), putting the U.S. at a strategic disadvantage because of the continuing reluctance of the NAS and NSF to correct the glaring and extraordinary errors in both physics and electrodynamics — and especially in electrical engineering.

Meanwhile, we believe that the work and experiments of Paulo and Alexandra Correa are at the forefront of capturing and transducing energy from the vacuum, in whatever form one chooses to consider the vacuum energy itself. They are also pointing the way to the corrections and changes necessary in the "old physics" if we are to better understand the world we live in.

6.3.10 Swiss Overunity Converter

We must also include the "Testatika" devices invented by Paul Baumann in **a** religious commune, Methernitha, adjacent to the village of Linden in Switzerland[180]. Figure 6-18 shows one of the machines, which over a period of 20 years were developed in various output sizes up to 30 kilowatts. Once started, the device is self-powering while also powering a very substantial external load (up to 25 kW). Many engineers and scientists have observed the machines in observation, examined them, and measured one or more in operation, so there is little doubt of their **authenticity.**

---

[180] The website for Methernitha is www.methernitha.com. One may read the statements of the religious group for oneself.

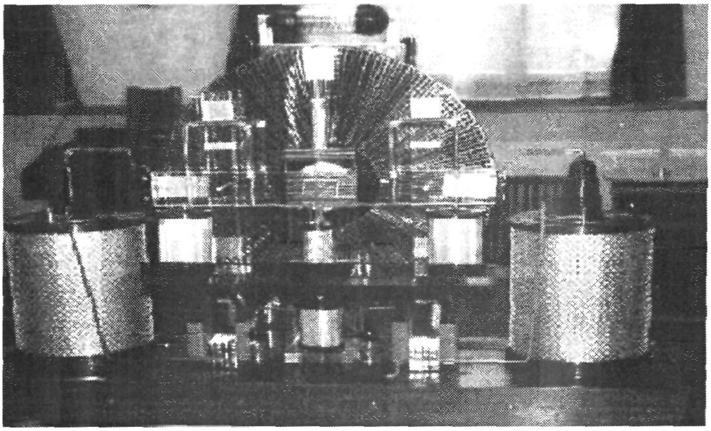

Figure 6-18 Swiss Methemitha machine.

However, there the science stops and mysticism and speculation begin. The religious commune releases no real details, except deliberately tantalizing clues. When questioned, the reply is that humankind in its present state cannot be trusted with such free energy. Almost all those examining the device have thought in ordinary electrical engineering terms, so very little comes from their examination because standard electrical engineering does not contain any theoretical explanation of COP>1.0 electrical power systems whatsoever. Baumann himself sometimes cryptically replied that his study of lightning discharges in nature gave him the operational principle. As we shall see, that may indeed be a very substantial clue as to the possible technical COP>1.0 mechanism, and one that does not exist in conventional classical electrodynamics.

The Methernitha (Baumann) machine resembles a Wimhurst machine with contra rotating static wheels, made of plastic or other insulating material with copper ribs and structures on one side and aluminum on the other. Use is made of the discharge into and from large capacitors and large coils. Also, at least one horseshoe-shaped permanent magnet assembly at the bottom is utilized. The major operating feature seems to be the contra rotating disks of a "Wimhurst" static generator machine, a magnetic field at right angles, and the sharp discharge of high voltage into and from large capacitors and large coils. The voluminous material on the Internet and in articles about the machines is mostly of little real use in contributing understanding or in proposing any legitimate mechanism.

We have not "solved" the machine's mechanism, since that would require close study of the device for a protracted time, and there has been no opportunity or time for that — although Baumann invited this author to come and see the machine some years before he died. Regrettably, I was unable to make the trip, and Baumann himself has now been dead for some years. We will therefore offer some technical suggestions pertaining to mechanisms that may be operating in the machine (or others like it).

First, we call attention to a paper by Aguirregabiria {418}. Quoting:

> *"An ohmic ring that rotates with constant angular velocity in an external uniform magnetic field is considered as a simple model for a current generator. Under the assumption that all quantities vary slowly in time, the lowest-order approximation to the surface charge density is found. The flux of the Poynting vector through the loop surface is also computed. Unlike the examples that are given in textbooks, this flux is not always incoming: It has the outgoing direction around the loop parts where the electrons are moving against electrostatic forces. "*

In other words, contrary to textbooks, there are positions where such a ring produces currents in one direction, and positions where it produces current in the other direction — *in fact moving against the voltage and E-field, and thus serving as a true negative resistor in those latter regions.* This may be a profound clue to the negative resistor operation of Baumann's machine.

This unexpected negative resistor effect is in theory usable to provide a true negative resistance. If there are transformers in the machine (and there are in some of them at least), then if a true negative resistor were used as an external shunt across the external circuit connected to the secondary of a transformer, the transformer could be operated as a true COP>1.0 device. The back-field connection from secondary to primary would be reduced by the fraction of the return current to the secondary that is shunted in parallel to that current "pumped" back up through the secondary.

This should be considered as an important possibility in the Baumann machine's operation.

Second, a charging and discharging capacitor involves not only energy flow in a direction perpendicular to the plates, but also an energy flow in a direction parallel to the plates {419}. Each of the rotating "plates" is actually a rotating capacitor of sorts, with a dielectric between two plates of special construction and dissimilar metals. The dielectric and the plates thus involve a complex intermingling of the negative resistor effect, a positive resistor effect, magnetic energy flows (both in entropic and negentropic manner in various regions), etc.

Baumann once handed two dissimilar plates, with a dielectric between them, to an engineer who measured it and found a startling several hundred volts across the plates! To date, no one is known to have been able to reproduce that effect.[181] We suggest that, if the dielectric material is layered and photorefractive in nature, it may be related to the negative resistors built by Chung {420}. If so, the physical pressure of the plates against the layered dielectric may also be quite important. By varying this pressure, Chung and her team found that their device could be made to exhibit negative resistance, positive resistance, or zero resistance. This leads us to speculate that Baumann may have used a "stressed dielectric" capacitor as a device having a built-in negative resistance. In that case, build-up of such a surprisingly high voltage would indicate high stress locked into the negative resistance dielectric.

We also call attention to a paper by LaPointe {421}, which will be discussed later under inertial propulsion and antigravity. Essentially, true Dirac hole current in the adjacent vacuum/space can be generated by powerful electrical discharges, particularly from capacitors or coils. When the unaccounted Heaviside energy component is also considered as well as the Poynting component, LaPointe's despair at the difficulty of obtaining sufficiently large voltage gradient across a very tiny distance is dispelled.

---

[181] We are reminded, however, of the peculiar phenomenon that can occur in large power capacitors stored on a warehouse floor. If their leads are not shorted, the capacitors will build up a self-charge gradually, and this can become quite lethal. For that reason, such large capacitors when stored will have their leads shorted, to prevent this "free cumulative charge and energy" from appearing and posing a serious hazard. To our knowledge, there has been no complete scientific explanation advanced for this "self-charging" effect in large storage capacitors. Our own hypothesis is that it is probably an effect of unaccounted Heaviside nondiverged energy flow components associated with various field/charge interactions in the area and with ordinary Poynting EM energy flow in the area, including within the earth itself. Further, with respect to the surface of the earth there is an increase in potential with altitude. Hence from any point above the earth's surface and a point on the surface, there exists a dipolarity and a "Kron open path". The broken symmetry of that dipolarity means that within it a continuous 4-circulation of EM energy flow occurs as well as point-dipole polarization.

If this speculation holds, then in examining lightning strikes and their more bizarre phenomena, Baumann may well have hit upon the fact that such discharges also generate negative energy as well as positive energy, although it appears that the technical concept of negative energy was foreign to him, at least in such terms. In short, he could have been unknowingly charging his large capacitors by Dirac sea hole currents, somewhat similar to Bedini's process.[182] In that case, most of the "negative resistance" operation could be explained by Baumann's transduction of large bursts of negative energy into large bursts of positive charging energy and current in his capacitors.

Or, said another way in terms of Aguirregabiria's effect, Baumann may simply have been charging capacitors in that "negative resistance" Aguirregabiria region where the current does flow backwards against the voltage.

At least we have proposed some known though rare phenomena, which Bedini and my four colleagues and I have discovered in independent efforts. We strongly suggest that the Baumann device may unwittingly incorporate several of these mechanisms, but that neither Baumann nor the other members of Methernitha seem to have any technical notion of them.

With that, we leave it to the interested researcher to perform his or her independent Baumann-type system experiments and build-ups. Eventually, as is always true in science, the experiment must determine the truth of what is really happening. Given the successful experiment, then a model must be contrived or fitted that (i) explains the new effects produced, and (ii) complies with normal EM operation of normal COP<1.0 systems.

6.4 Some Magnetic Processes of Interest

There are some 200 or more magnetic effects in the literature, and only about half of them are well understood. Of the remaining half, some are partially understood and some are not understood at all. The COP>1.0 researcher interested in magnetic systems would do well to study magnetic materials science {422a-422c}, known magnetic effects, and particularly those magnetic effects that are still anomalous such as in leading edge superconductivity research. The new area of *spintronics* {423}, e.g., is rapidly advancing and will have future impact on COP>1.0 magnetic systems. We do not further discuss spintronics in this book, but merely call this rapidly emerging area to the attention of the interested COP>1.0 researcher.[183]

---

[182] Bedini and the present author have filed a patent application upon this exact mechanism and process, and typical embodiments for accomplishing it. Bedini discovered the effect and used it for two or more decades, and after some intense struggle with the problem I was able to find the exact technical mechanism, including literature justification. This becomes a system to close-loop a "unitary" type of COP>1.0 system, which is covered in the patent application also. A different system can be used to close-loop a non-unitary COP>1.0 system, and that method is contained in the second MEG-type patent application of my four colleagues and I who invented the motionless electromagnetic generator (MEG).

Let us examine a very few magnetic effects of interest, as a sampling of what we are referring to.

6.4.1 Self-Powered Magnetohydrodynamic Motors

Interesting self-powered magnetohydrodynamic motor work has been shown by Bednarek {424}. He shows a multi-branch rotor in a galvanic cell composed of sulphuric acid electrolyte with copper and zinc electrodes. The unit develops an emf of about 1 volt, with the rotor placed in a vertical magnetic field such as from a large rectangular permanent magnet, where the field of the magnet is a few teslas in strength. The electrolyte and the rotor revolve in opposite directions. To change the direction of revolution of the rotor, the magnetic field direction is reversed. A linear self-powered magnetohydrodynamic motor is also shown, and has a similar principle of operation.

The interesting point is that these motors work, but less efficiently, when a water solution of table salt is used as the electrolyte. With the worldwide availability of seawater, this becomes an interesting possibility of "renewable energy" application. We leave it to the reader to further investigate this proven experimental technology as desired and to determine its practicality.

6.4.2 Multivalued Magnetic Potential

Conventional system designers work with conservative fields, as shown in Figure 6-19. Conservative fields arise from a single-valued potential — a potential that has only one value at each point in 3-space that it occupies. Consider a rolling ball of mass *m* on the "oval track" shown in Figure 6-19, starting from point A and rolling on the right side path. With the ball at rest at point A, we give it a slight push to velocity V_i adding kinetic energy $K_i = 1/2\ m(V_i)^2$. At point A, the ball has its maximum potential energy P_A due to gravity.

---

[183] For technical information, see (a) Michael Ziese and Martin J. Thornton, Eds., *Spin Electronics,* Springer-Verlag, 2001; (b) D. D. Awschalon, N. Samarth, and D. Loss, Eds., *Semiconductor Spintronics and Quantum Computation,* Springer-Verlag, 2001.

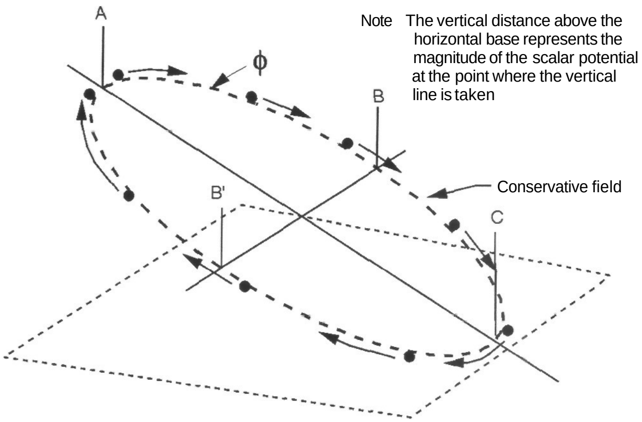

Note The vertical distance above the horizontal base represents the magnitude of the scalar potential at the point where the vertical line is taken

A circular closed path in potential ϕ is given by A-B-C-B'-A
Line integral from A around any closed path back to A is conservative
Along path A-B-C, work W(1) may be extracted from rolling ball
Along path C-B'-A, must do work W(2) on ball, where W(2) = - W(1)

Figure 6-19 Conservative field use (single-valued potential).

As it moves to the right because of our push, the ball will accelerate due to the force of gravity and increase its kinetic energy until it reaches point C at the bottom, reaching its maximum linear velocity Vc and kinetic energy Kc at point C. Its change (PA - Pc) in potential energy at from point A to point C has been converted to kinetic energy $[\frac{1}{2}\ m(V_1)^2 + \frac{1}{2}\ mVg^2]$ at point C.

Continuing on around on the left half of the path, as the ball rises toward A again, the portion $(P_A - P_C) = \frac{1}{2}\ mVg^2$ —of the ball's kinetic energy at C that was added by the force of gravity from A to C — is returned to potential energy P_A. The ball reaches its lowest kinetic energy $\frac{1}{2}\ m(V_1)^2$ at A and also its lowest velocity V_1. For a perfect lossless system, once started in motion the ball would rotate around the track indefinitely, freely changing potential energy into kinetic energy and then back to potential energy. Nevertheless, it would not do any outside work, for that would represent losses or dissipations of energy from this conservative system. Consequently, the ball would quickly run down if work were being done, even if the system were otherwise "perfect".

If one integrates the change in potential energy around the track, the net change in potential energy is zero. The work done on the ball by increasing gravity to increase the ball's energy in one half-cycle, is taken back from the ball back when it climbs back out into decreasing gravity in the other half cycle. If one integrates the changes in kinetic energy around the track, once the initial velocity and kinetic energy are produced by outside forces in a perfect system, the net change in kinetic energy is zero. Of course, it is easily seen that the ball gains kinetic energy on its downward half of the track, and then returns the kinetic energy on its upward half of the track.

This situation is said to involve a *conservative fields* and the system will not produce any net energy to use as free work. Indeed, a real system will almost always have some friction and other losses around the path {425}, so the ball will gradually lose net initial energy given it, slow down, and eventually come to rest at point C, the lowest potential energy of the system.

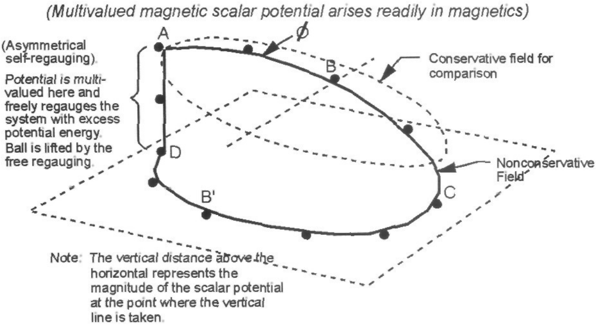
(Multivalued magnetic scalar potential arises readily in magnetics)
(Asymmetrical self-regauging).
Potential is multi-valued here and freely regauges the system with excess potential energy. Ball is lifted by the free regauging.
A
ϕ
B
Conservative field for comparison
D
Nonconservative Field
C
B'
Note: The vertical distance above the horizontal represents the magnitude of the scalar potential at the point where the vertical line is taken.

A circular closed path in potential ϕ is given by A-B-C-B'-D-A.
Potential ϕ is multi-valued from D to A.
Line integral from A around this closed path back to A is nonconservative.
Along entire path A-B-C-B'-D, work may be extracted.
No work is required to go from D to A. When ball arrives at D, it also arrives at A simultaneously. I.e., the potential "jumps" in magnitude.

Figure 6-20 Field may be non-conservative for a multivalued potential.

A charge circulated through a closed current loop circuit also moves through a conservative field region in similar manner. Hence, there is no net excess energy input available for such a conventional circuit. It follows that one must arrange for inputting all the input energy (intercepted and collected by the circuit) that is then dissipated from the circuit's loads and losses. Unfortunately, half the collected circuit energy also goes to destroying the dipole, with less than half powering the load. Therefore, that circuit always exhibits COP<1.0.

Now see Figure 6-20, where we provide an analogy using a rolling ball around a circular track, in a gravitational potential. In this case, the potential has a multiple magnitude A and D at point A-D, which is the same point (we plot a point in the potential's magnitude, not a point in 3-space, although there can be close approximation).

Approaching magnitude D from point B', the value of the potential is steadily reducing and is much lower prior to reaching A-D than it was when leaving magnitude A. During this phase, the ball is gaining kinetic energy from the gravitational potential, and can be used to perform useful work (up to all the energy gained, in a perfect system). Upon passing through point A-D, instantly the ball is again at point A on the right. We accent that the "instantaneous jump" on the diagram represents a free insertion of excess energy (asymmetric regauging energy) from outside the system. Rigorously that is a broken symmetry, which also breaks the internal energy conservation of the system. In short, the "ball" has been instantly "lifted" (in this simplified analogy) back to a higher potential by a free insertion of excess energy from the external environment. The main point is that a surge of potential energy into the "ball system" occurred, freely input by nature and the environment. The system can now go through the "doing work" routine again, traversing from A back to D and through it.

Note that this is a nonconservative system, because it continually receives a free and sudden input of excess energy from its environment. So this system — because of its broken symmetry — can continually do work and keep on going.

Indeed, so long as that free energy input from the environment occurs without fail, this is a "self-powering" system, completely complying with **the** laws of physics and thermodynamics. It violates the equilibrium thermodynamics because the system is periodically not in equilibrium with its environment. Consequently, the system can exhibit those five magic functions we spoke about previously. It can exhibit (i) self-organizing (in this case, freely getting that little ball from low potential energy D back to high potential energy A), (ii) self-oscillation or self-rotation (the ball will continue to go around and around the loop, even doing a little work in the process), (iii) output more energy than the operator inputs (in this case, the operator is not inputting any energy at all, so the energy output is indeed more than the operator furnishes), (iv) power itself and its load simultaneously (all the energy is being input from the external environment at the insertion of excess energy at point D to move the ball **back** to position A), and (v) exhibit negentropy.

This is an analogy to a nonconservative field and a multi-valued potential. In the case where a potential has discontinuous values at a single point, with the value depending upon whether the detecting charge (the "ball", so to speak) is to the left or to the right of that point, one has a *multivalued potential*. This multivalued potential actually represents a change of potential energy in the system, freely occurring without operator input. In short, it is an asymmetrical self-regauging, violating Lorentz's symmetrical regauging condition.

Actually, multivalued *magnetic* potentials arise naturally in magnetics theory, but —foolishly, in our view — theoreticians do all in their power to minimize or eliminate their consideration {426a-k}. They consider such a nonlinear change as being embarrassing and troublesome, and to be gotten rid of at all costs! However, if deliberately used and optimized, rather than eliminated, incorporating a multivalued magnetic potential can provide a *nonconservative magnetic field* (analogous to the illustration), where $\int \mathbf{F} \bullet \mathbf{ds} \neq 0$ around a rotary permanent magnet loop. The multivalued potential represents a broken symmetry that further produces a nonconservative field.[184] In theory, such use of the multivalued potential and the resulting nonconservative field can enable a "self-powering" permanent magnet rotary engine, operating as a negative resistor freely extracting and using magnetic energy from the broken symmetry in the system's energetic exchange with the active vacuum.

However, note that the multivalued potential represents a point of sharp "self-regauging asymmetrically" by the circuit, with the regauging potential energy coming from the external environment. It requires that some external process in the exchange between environment and system must be automatically invoked at the multiple-value point, so that there is a sharp and sudden entry of excess potential energy received by the circulating ball (or by the circulating charges in an electrical circuit, or by the circulating flux in a magnetic circuit).

One such means of evoking such a sudden surge of excess energy at a point, momentarily, is given by Lenz's law, as discussed in the magnetic Wankel engine. Another means is by Johnson's sudden evocation of the exchange force, which momentarily can produce a pulse of energy density even several hundred times as strong as the usual field energy at that point. For example, imagine that the ball in Figure 6-20 represents a magnetic rotor in a magnetic potential represented by the height of the ball above the dotted line base. The point (D-A) then represents the results of the system having suddenly injected an instantaneous exchange force at point D, to freely increase the potential and potential energy of the rotor system back to point A. The rotor is immediately lifted back at its starting potential energy situation at point A, freely, by this sudden and free evocation of the exchange force. We will briefly discuss the exchange force in our discussion of the Johnson engine.

---

[184] In physics, the appearance of a force and its subsequent action to perform work is nature's way of restoring symmetry to a situation where symmetry has been violated. As can be seen, the net force can be used to translate something (such as current) and do work, thereby dissipating the excess potential energy received from the broken symmetry condition.

6.4.3 Dromgoole Effect as an Example of a Novel Magnetic Effect

An example of a novel magnetic effect is the *Dromgoole effect* {427}. This is an interesting phenomenon where a voltage placed on a solenoid wrapped around a longitudinally magnetized iron wire may be increased up to 300 times in magnitude by twisting the wire through 90 degrees. If a scheme can be worked out whereby this dramatic increase in voltage potentializes and moves very substantial current, then perhaps the extra output can — at least momentarily — produce more output energy than the work required to twist the wire.[185] At least that could be a working hypothesis from which to launch experiments to see if it is possible.

The reason this may be possible is that any amount of energy W one wishes can in theory be collected from any finite potential intensity ϕ, according to the simple equation $W = \phi q$, where q is the collecting charge exposed to and interacting with the potential intensity ϕ. In the case of magnetics, q is analogous to the pole (magnetic charge), where the north pole is positive magnetic charge. The ϕ then becomes the magnetostatic scalar potential.

So if we can produce ϕ with only a little expenditure of energy, and then have that ϕ potentialize a very large amount of charge q, we can collect upon the charge q much more energy than what we ourselves expended. The ϕ produced is a change to the local vacuum potential, and hence the collection of energy W on charges q is actually a collection of EM energy from the local altered vacuum potential itself. Any potential we make becomes a change to the local vacuum potential, and hence a change to the local active vacuum. That does not seem to appear in electrical engineering, which model does not even incorporate modeling the active vacuum or its potential, much less a change to it.

---

[185] The scheme must prevent *at least an appreciable fraction* of the spent electrons-from the external circuit and load — from being forcibly rammed back up through the coil against its back EMF. Otherwise the Lorentz symmetry condition applies and the arrangement will not produce COP>1.0.

While we discuss a few interesting magnetic effects in this paper, there are many more. The interested researcher is referred to such easier sources as Burke {428} and Cullity {429}. For more complex scientific sources and explanations, other publications are available {430a-430r}. In addition, it is helpful if the researcher is aware of some of the foundations problems in physics and electromagnetics {431a-431e}. Our point is that there are more than 200 known effects in magnetics, and only about half of them are well understood. For the other half, the understanding ranges from "partially understood" to "not understood at all." The latter half of the magnetic effects provides a rich ground for investigation by researchers seeking an asymmetrical self-regauging mechanism.

The COP>1.0 researcher must be prepared for an extended self-education period, and appreciable study and work. Some research discipline is highly recommended, such as starting one's own database and rigorously maintaining it up to date with one's latest interests. For the experimenter, a good lab notebook, meticulously kept and regularly posted, is an absolute necessity. Reading and searching the scientific literature is also highly recommended. It is not as simple as applying the principles one learned in university or technical school. Those techniques and principles are involved, but at some point in the circuit they must also be violated. Else COP>1.0 systems would long ago have been developed and marketed by sharp young students, graduate students, and post-doctoral scientists.

6.4.4 Exchange Force and Johnson's Approach

Refer again to our discussion of conservative versus nonconservative fields, above.

If one wishes to build a self-powering rotary permanent magnet motor, obviously one must evoke a condition whereby $\int \mathbf{F} \bullet d\mathbf{l} \neq 0$ around a single full rotation of the rotor (and where F is the instantaneous magnetic force between stator and rotor). For permanent magnets without any external switching, etc., that condition can only arise if an excess force is freely triggered by changes occurring freely in the magnetic materials themselves.

Fortunately, magnetic materials are almost always highly nonlinear, and **so** such self-triggering opportunities abound. They can occur because there are a multiplicity of mechanisms (dozens and dozens) going on simultaneously in the materials of the magnets of both the rotor and the stator, and not just the standard polarity and field repulsion and attraction that electrical engineers use. These materials processes also involve direct and ongoing exchange of energy between the active vacuum and the local materials in the magnet.

The trick is to take one of these automatically ongoing processes that is already there and freely available but just not adroitly used, and deliberately evoke and enhance it to induce excess energy at the appropriate time, with a force in the appropriate direction. In simple terms, if you have eight more horses available to put into the harness to pull the load, and you only have two normal horses but wish to pull more load, then by all means you should hitch up and use at least one more additional horse, and take him from the six that are usually just standing by and not working at all. They may run back and forth a bit, while watching the others work, but that of itself contributes nothing.

One such additional force and free regauging mechanism always available is the exchange force, which can be checked in Feynman's three volumes of physics {433} and also in the reference by Cullity {429} previously **cited.** Good science dictionaries {432} also include helpful information. Reading the literature will add enormously to one' data base of effects and methods.

Howard Johnson's approach is to arrange highly nonlinear assemblies of geometrically asymmetrical magnets into complex configurations. At certain points in a magnetic field, these Johnson assemblies will trigger the sudden brief self-generation (actually a sudden release) *of exchange force* between stator and rotor. The exchange force is a known phenomenon {432, 433, 434}, though still not completely understood theoretically. It **can** be evoked asymmetrically by evoking electron-spin flipping, for example, which has led to the development of a new field of technology, **spintronics.** *Microscopically,* the exchange energy is much greater than the magnetic field energy, sometimes by a factor of 10^3 to 10^4 {435, 436}. *Macroscopically,* the resulting exchange energy can momentarily still be many times the magnitude of the ordinary magnetic field energy; e.g., by a factor of 200, which Johnson often achieves.

If **the** stimulated momentary exchange forces are oriented to always add to the thrust upon the rotor in the "powering rotation" direction, then the **multivalued** potential condition and $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ are met. In that case, self-powering is permitted, if enough of these exchange force triggerings occur so that the extra potential energy added to the rotor is sufficient to overcome the load and the losses being powered.

The self-triggering of each exchange force appearance produces an excess burst of force and energy[186] input into the system from its active supersystem environment. Thus, repeatedly the system is momentarily converted (in each exchange force burst) into an open system in disequilibrium in its energy exchange with its active environment, freely receiving excess energy from it. Because the exchange force "input energy burst" is short, multiple such "input bursts" must be used in a single rotation cycle so that the total energy input by all of them is significant.

In that case, COP>1.0 performance is permitted by the laws of physics, thermodynamics, and nature. Conservation of energy is not violated. Classical equilibrium thermodynamics with its infamous second law does not apply to the Johnson system, since the system is periodically an open system far from equilibrium and receiving excess energy from its active environmental exchange. *A priori* the Johnson system has increased its negentropy overall, and that negentropy (increased order or increased potential energy) can then be dissipated (disordered) to produce free shaft horsepower if the bursts of exchange force are properly coherent in direction and timing.

We diverge for a moment: As is well known, the equilibrium condition in a system is the condition of maximum entropy in the system; any disequilibrium condition reduces the entropy *a priori* because it is an excited state of the system containing additional potential energy. It is worth rigorously clarifying the infamous second law of thermodynamics. Quoting Lindsay and Margenau {437}:

> *"[The]...statement of the second law: (a) the entropy... is a variable of state, (b) Its value, for a closed system, can never decrease."*
>
> *... "Non-equilibrium conditions cannot be specified by variables of state, and their entropy cannot be computed. ...the condition of equilibrium is the condition of maximum entropy."*

---

[186] Technically speaking, the exchange force is not a magnetic field force, but a force that arises independently due to quantum mechanical considerations. Nonetheless, it is a real force arising in magnetic materials and affecting magnetic materials, as in permanent magnets.

In the first statement, the reader should note the assumption of a closed system in the first subparagraph, and the absolute requirement that the calculated entropy be a variable of state.

Also, the quotation can be falsified. Oddly, the "closed system" in classical thermodynamics is defined (illogically) as a system that does not exchange mass across its boundary, but may and often does exchange energy across it. We negate any absoluteness of that statement by Lindsay and Margenau by citing a counter example of a "closed" system with continuously increasing energy, fed by transfer of energy from the environment, and with no escape of the energy or very little escape of it. A specific example is a photon absorption by a mass particle. This system achieves increasing energy (order) as the process continues, hence achieves negentropy. This may be considered a fluctuation, of course, but it still places severe limitations on this law of thermodynamics and in fact negates any absoluteness of it. For such reasons, in this book we have redefined "closed system" as one that exchanges neither energy nor mass with its environment, and we recognize that there are no such systems in the universe. We have defined an "open system" as a system that exchanges either energy or mass or both across its boundary, so that we do not encounter the problem of the counter example cited. Further, general relativity requires an increase in the mass of any system that increases its potential energy, and a decrease in the mass of any system that decreases its potential energy. Hence energy exchange at all with the system, involves mass exchange since mass and energy are the same thing. The classical thermodynamic definition of a "closed system" has thus been falsified since 1915, with the definition becoming only an approximation rather than a generally valid definition.

In the second subparagraph of that first statement by Lindsay and Margenau, the reader should note that the closed-system assumption must be violated *a priori* if the entropy does decrease, and vice versa. If the system is broken into a set of subsystems, then the only way the entropy of the overall closed system to decrease is for one or more of the subsystems to be open (new definition!) and energy (order) to pass out of the system. Then an interesting thing emerges: For order (energy) to remain in the system as such, the subsystems taken as a whole must produce as much negentropy as they do entropy. Energy from an ordered subsystem can be emitted in disordered form, but then it has opened that subsystem and has entered the space between parts (subsystems) of the overall system. In other words, in a closed system, *any increase in entropy requires the subsystems to become open subsystems.* Again, the statement of this law of thermodynamics eats itself. To stay in the overall system, this scattered energy *outside the subsystems* must then interact totally with another part of the system, and so on. This introduces disorder to the succeeding parts that interact. Therefore, the second law of thermodynamics itself *internally* violates its own "closed system" assumption because, to operate at all, the law requires continuing interaction between the active local vacuum environments and the subsystem components. In short, it requires a very special kind of overall or average equilibrium in an unavoidable energy exchange between the local vacuum and all the parts of the system. The source charge problem already demonstrates the universal violation of the second law and the thermodynamic definition of "open system", but both classical electrodynamics and classical thermodynamics have ignored this source charge problem for more than a century. Our solution to it was published in 2000.

Quite simply, there is no such thing as a truly closed system in the first place. Kondepudi and Prigogine come close to this statement in the following quotation[187]:

> *'Anyway, equilibrum thermodynamics covers only a small fraction of our everyday experience. We now understand that we cannot describe Nature around us without an appeal to nonequilibrium situations. The biosphere is maintained in nonequilibrium through the flow of energy coming from the sun, and this flow is itself the result of the nonequilibrium situation of our present state in the universe."*

In short, all systems on the planet — and we ourselves — are immersed in a nonequilibrium state a priori. Rigorously there is no such thing as an absolute equilibrium state on the planet, except as an approximation.

Now consider a perfectly insulated system, so that no heat can pass from the system outside it. An interesting constraint then exists on those "open subsystems" producing disorder (entropy). Unless equal reordering occurs in the subsystem-to-subsystem reactions, then disordering (heat) grows a priori. But this is not observed to happen in well-insulated systems approximating our theoretically perfect example! Otherwise, the temperature of a well-insulated system would increase until system rupture and failure. And experimentally that does not happen.

---

[187] Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, 1998, p. xii.

It follows that, to maintain the internal equilibrium between subsystems and a constant internal temperature, a negentropic process is clandestinely involved. We submit that this process is revealed in our discovery of giant negentropy of the negative charge, and what may be said to be the giant entropy of the positive charge — i.e., in the discovery of the common " 4-circulation" of energy surrounding a dipole from the time domain to the negative charge of a point dipole in 3-space (thereby entering 3-space once emitted by the negative charge), thence to the positive charge of the point dipole, and thence back to the time domain. For a single charge, the well-known vacuum polarization provides virtual charges of opposite sign, to convert the "isolated charge" into a set of composite dipoles, as previously explained.

The second law of classical thermodynamics, considered in a more modern light, appears to conceal hidden giant negentropy and hidden giant entropy, in the ongoing 4-circulation of EM energy in the supersystem. *It is not possible to eliminate the supersystem or the interchange between its parts; particle physics told us in 1957 that there is no equilibrium of any system without this ongoing exchange.* Any thermodynamics attempting to discard the supersystem exchange (which involved both mass and energy) is at best an approximation for special "reasonably well-behaved" situations.

If the entire system is not in net equilibrium with the external environment (i.e., if there exists disequilibrium between the separated parts of the supersystem), then classical thermodynamics does not *absolutely* apply to that system. The system is no longer *absolutely* describable by "variables of state".

Those objecting to COP>1.0 in an EM system on the grounds that it would violate the second law of thermodynamics (which already violates itself), would be well-advised to restudy the very notion of the second law and the thermodynamics definition of *open system.* Compare relativity's equating mass as energy. Then ponder the thermodynamics of open systems far from equilibrium with their active environment. Every system in the universe is open, and it has an ongoing exchange with its proven active environment (local active vacuum and curved spacetime). This exchange Includes and exchange with every particle in the system. As pointed out by Lee:[188]

---

[188] T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 46-47.

> *"...symmetry implies conservation. Since our entire edifice of interactions is built on symmetry assumptions, there should be as a result a large number of conservation laws. The only trouble is that almost all of these conservation laws have been violated experimentally."... "...this difficulty could be resolved by introducing a new element, the vacuum. Instead of saying that the symmetry of all matter is being violated, we suggest that all conservation laws must take both matter and vacuum into account. If we include matter together with vacuum, then an overall symmetry could be restored."*

The *system itself* is always in disequilibrium; only the supersystem can exhibit equilibrium. The second law of thermodynamics specifically does not and cannot apply to a system far from equilibrium, because of its implicit assumption of overall equilibrium without the active vacuum exchange. Also, a deeper balance is required between the hidden asymmetries existing between the subsystems and their local vacuum (and local spacetime curvature.

Indeed, *one cannot even calculate* the entropy for a system that — overall — is far from net equilibrium with its active environment. We quote Lindsay and Margenau even more strongly {438}:

> *"Equilibrium states are the only ones that are capable of explicit analysis in thermodynamics..."*

And again {439}:

> "... *variables of state have meaning only if they define an equilibrium state. Hence the quantity we are seeking will be meaningless unless it refers to equilibrium states."*

While we are at it, let us also address a serious flaw in the first law of thermodynamics. We again use Lindsay and Margenau for a succinct statement of the First Law {440}:

> *"First law of thermodynamics. A complete statement of the first law comprises two assertions: (a) heat is a form of energy, (b) Energy is conserved."*

All that really says is that energy is conserved. It does not state that it is conserved in an object. It states that, whether the system is in equilibrium or not, energy is conserved. If heat is taken as disordered energy, then it merely states that overall the energy is conserved, whether ordered or disordered with respect to some ordering criterion. It does not state that the disordering is conserved, and it does not state that disordering must increase or decrease. But *it does implicitly assume that all energy at some most basic level is ordered, else it cannot be energy (order).* So it assumes that, at higher levels, energy can be disordered (incoherent). However, at the underlying basic level, it is and remains perfectly ordered — else it could not remain "energy and could not be conserved. As an EM example, in so-called "heat", *every scattered photon retains its perfect order; it is only the photon ensemble that is "disordered".*

In short, each "basic piece" of energy is perfectly ordered, but the *ensemble* of the pieces may be disordered. Therefore, entropy applies only at a level higher than the basic energy quantum. Contrary to the assumptions of classical (macroscopic) thermodynamics, processes which directly engineer the basic energy quanta[189] — more exactly, the *action* quanta, consisting of *energy x time,* since energy cannot be "engineered" or changed in 3-space without also being engineered "in time" as well — are time-reversible. Hence they can be negentropic — simply because every observable system is "open" to, and in continuous energy exchange with, its active time environment (and also its active vacuum environment). Also, no system changes its spatial energy in any fashion, including ordering or disordering, without interacting with spacetime and spacetime curvature dynamics. It also changes its time-energy.

So in our view the notion of "disordering" and "disordering of energy" must be carefully reconsidered, as to exactly what is and is not being disordered, when the assumed "disordering" occurs, at what level it occurs, where and how the compensating reordering occurs, etc. We also point out that the simple discovery of giant negentropy {12} as the solution to the long-vexing source charge problem already removes the "absoluteness" of classical thermodynamics. Giant negentropy already violates the assumptions of classical thermodynamics at the elemental level in every physical system. Indeed, every charge in the universe already falsifies any "absoluteness" of the assumptions of classical thermodynamics.

This problem in the old classical thermodynamics has long been indirectly solved in particle physics, with the discovery of broken symmetry. As Lee states so clearly {441},

---

[189] Actually, energy is discretized, not quantized. Energy x time (i.e., *action)* is **quantized.**

> *"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into minus, running time backward and turning things inside out."*

We point out that a symmetry involves a conservation law, such as are stated in classical thermodynamics, and a broken symmetry involves a broken conservation law. So the discovery of broken symmetry in physics was a profound change affecting all physics, including the staid old classical thermodynamics. Lee further points out the new complexity of concepts {442} such as symmetry (which is behind every conservation law, including the first law of thermodynamics):

> *"At present, it appears that physical laws are not symmetrical with respect to C, P, T, CP, PT and C. Nevertheless, all indications are that the joint action of CPT (i.e., particle <-> antiparticle, right <-> left and past <-> future) remains a good symmetry."*

So unless the first law is stated in terms of modern CPT symmetry, it does not absolutely apply! Further, every charge is changing time-energy into spatial energy or vice versa. Yet there is nothing about time-energy and its transduction into spatial energy, or vice versa, in the present textbook statements of the thermodynamics. The term "heat" does not refer to the presence of energy at all, but to the scattering (disordering) and escape of energy.[190]

Considering heat as "energy of the system", or "heat energy" of the system, is a grand non sequitur. Rigorously, "heat" refers to the reduction of higher levels of ordering of energy, and since the gist of energy is ordering, reduction of ordering is the very antithesis of energy! "Heat energy" thus is an oxymoron. Before the "escape", there is no "heat energy" (ugh!) in the system at all. The energy is present in the system *not as disordering, but as ordering, a priori.* If it were in the system, it would not have escaped nor would it be escaping from the system. More energetic molecular motion, e.g., is actually more energetic ordering, simply at an excited state (of greater energy!).

---

[190] Think closely: We never take the temperature of a "system"! We take the temperature of the disordered energy (heat) leaving that system or its subsystems. We do measure the effect of the emitted disordered energy. But that has already left the system and is in the local vacuum (a second component of the supersystem). Thermodynamics might be usefully redone more exactly in terms of the supersystem. We leave that task to some budding young future thermodynamicists for a recommended doctoral thesis.

We stress again (and strongly advise the researcher to read) Romer's strong objection to the use of heat as a noun {443}, and we suggest that the entire subject of classical thermodynamics needs a thorough revision to tighten up its terminology, correct its definition of closed system, eliminate its conflict between the first and second laws, and remove its inappropriate consideration of heat as "energy". Otherwise, the presentation and general interpretation of thermodynamics itself will continue to be one of the great confusion factors one encounters in trying to think clearly about extracting EM energy from the active vacuum environment to produce and utilize **COP>** 1.0 systems.

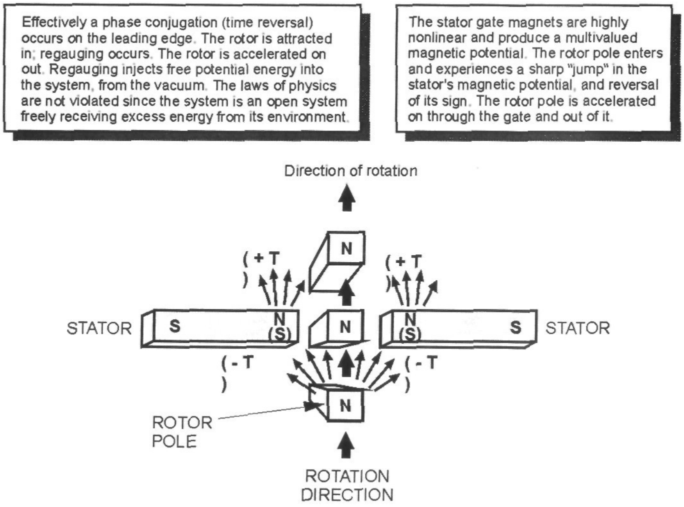
Effectively a phase conjugation (time reversal) occurs on the leading edge. The rotor is attracted in; regauging occurs. The rotor is accelerated on out. Regauging injects free potential energy into the system, from the vacuum. The laws of physics are not violated since the system is an open system freely receiving excess energy from its environment.

The stator gate magnets are highly nonlinear and produce a multivalued magnetic potential. The rotor pole enters and experiences a sharp "jump" in the stator's magnetic potential, and reversal of its sign. The rotor pole is accelerated on through the gate and out of it.

Figure 6-21 Concept of operation of Johnson's patented magnetic gate.

Johnson's approach produces a peculiar kind of multivalued potential. One **part** is conservative and of magnetic energy in nature, and the other part is nonmagnetic, being direct exchange force and energy effects on the **participating** atomic nuclei arising from quantum mechanical considerations. Nonetheless, the net *hybrid* potential can be multivalued at various points around the stator, and that is all that is required for broken symmetry. If $\int \boldsymbol{F} \cdot d\boldsymbol{l} \neq 0$ around a closed path, regardless of which set of forces F represents, then that overall composite force field acting upon the rotor or object moving around the closed path is nonconservative. In that case, self-powering is permitted.

Figure 6-21 diagrammatically illustrates the operation of a nonconservative force-producing magnetic gate in Johnson's approach to a permanent magnet motor. As Johnson has shown, by using a multivalued hybrid potential (MVP) in his gates so that the resulting excess exchange force is properly oriented in direction and timing, a rotor magnet is attracted into a highly nonlinear stator gate region where the MVP is located. When it enters the MVP region where the exchange force is initiated, the rotor encounters a dramatic jump in the net potential with a change of polarity as the exchange force suddenly fires automatically. In turn, this produces a sudden accelerating tangential force in the region, usually against the prevailing magnetic force in the normal back mmf region!

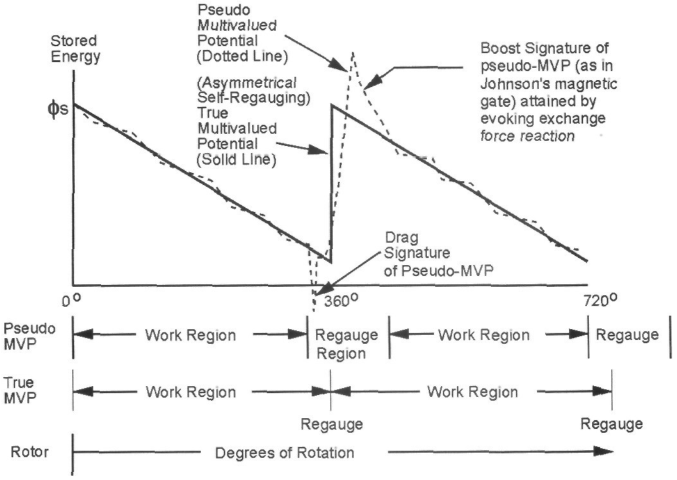

Note Single regaugmg per notation shown Johnson uses multiple regaugings per rotation

Figure 6-22 Johnson's "refueling" a permanent magnet engine by asymmetrical **self regauging. A multivalued potential is momentarily created by precisely evoking an exchange force reaction.**

Johnson often confounds professors of electromagnetics by showing them that his special gate can produce attraction between like poles, with a north rotor pole being drawn into a north stator pole in that exchange force region. This can be understood by tracing the spin field flow patterns, but cannot be understood by simple "like poles repel and unlike poles attract". In short, Johnson knows he is adding a force completely different from the magnetic force the professors have in mind as the sole acting force. It is rather like pushing two like poles of two little magnets together. You can do it because an additional force — your mechanical force pushing them together — is acting and overcomes the normal repulsive force between the two poles.

The sudden exchange force burst produces a sudden accelerating tangential force in the normal deceleration (back mmf) region of the stator-rotor arrangement. In short, Johnson utilizes a deliberately evoked burst of exchange force to freely overcome most or much of the back mmf (back-drag) region. This in turn produces a net driving force around the rotor's complete rotation, because the overall field combination is nonconservative.

Rigorous force meter measurements taken at 0.01-second intervals prove that this occurs as the rotor enters and passes through Johnson's gate. The results of one such experiment are given in Figure 6-22.

Johnson thus uses highly nonlinear magnet assemblies of novel design to create a special kind of multivalued potential in his magnetic gate by adroitly evoking exchange forces — *extra quantum mechanical forces* in addition to the usual magnetic field forces. The MVP produces a sudden extra rotational force, backwards to the back-drag (back mmf) region of the stator-rotor magnet arrangement where the rotor magnet assembly is entering and passing through the conventionally repelling like-polarity situation. The result is that a net north rotor pole is seemingly attracted in[191] to the otherwise repelling stator north pole region, then accelerated out the other side by the normal magnetic repulsion.[192] Net drive force results, instead of a net zero force. Engineer Ken Moore, a close colleague of the present author, independently tested a Johnson gate and verified the net driving force effect {444} within the limits of our little laboratory.

---

[191] In reality it is "forced in" against the repulsive magnetic force, by a superior force of a different nature, suddenly evoked.

[192] I.e., as the exchange force dies, the forces now acting are the normal repulsive magnetic forces, but the rotor has moved to a position where now this repelling force accelerates the rotor in the "load powering" rotational direction. In other words, the observer sees a like pole suddenly "attracted" into a like polarity region, move partially through it, and then suddenly be accelerated out the other side. But the "attracting in" force and the "accelerating out" force accelerate the rotor in the load-powering rotational direction.

Technically, Johnson's apparatus asymmetrically and suddenly regauges its exchange forces in such a manner as to momentarily overpower the back mmf region of the engine's rotation, thus producing a net multivalued potential and a net accelerating force around the overall rotation loop. However, correlated distant effects in magnetic materials are known. If distant compensation does occur, then the closed-loop capability might conceivably be defeated, if the total compensation is sufficient to provide an $\int \mathbf{F} \bullet \mathbf{dl} = 0$ overall condition. Hence the research probably demands attention to maintaining both *local* asymmetry and *distant* asymmetry. If the distant correlation effects occur, only in that manner can an overall nonconservative field represented by $\int \mathbf{F} \bullet \mathbf{dl} \neq \mathbf{0}$ be maintained around a completely closed magnetic loop.

This may be the problem with which Johnson appears to have been struggling for so long. He did solve the problem once, and produced a self-rotating permanent magnet motor that the present author personally tested over a period of about 2 hours. During a two-hour test period, the motor apparatus was allowed to self-rotate for up to 15 minutes at a time. Shortly after it became known that Johnson had successfully produced such a prototype, his laboratory was mysteriously broken into. The only thing taken was the magnet assemblies on that specific little successful demonstrator — in a laboratory with more than a hundred thousand dollars worth of magnets and magnetic assemblies of many kinds present.

As a personal aside, I have been associated with Johnson, off and on, for many years. He is a man of the highest integrity and an indefatigable researcher. It is my fervent hope that Johnson will succeed in his permanent magnet rotary motor project. All I can say as this goes to press, is that he is getting very close again, now that he has help and good machining accuracy, etc. for higher precision buildups. In the near future, I hope to be among the first to drink a good glass of champagne to celebrate his success.

6.4.5 Radus Effect: Magnets with Preferred Flux Path Memory

In the early NASA space program, magnetic boots for holding the astronaut to the skin of the space vehicle when outside it were required. A Westinghouse engineer named Raymond J. Radus had found that a permanent magnet applied to a dual flux path (Figure 6-23) would "remember" the flux path taken initially, until switched to the other flux path in an initial condition, after which it would remember that path as preferable {445a-445c}. A permanent magnet with such "memory" will split its flux in two parallel flux paths by some ratio, e.g., 20% in the non-preferred path and 80% in the preferred path. This preference can be switched at will by proper manipulation. So Radus developed switchable magnetic astronaut boots for the space program, using his magnetic memory and flux path switching technique {446}.

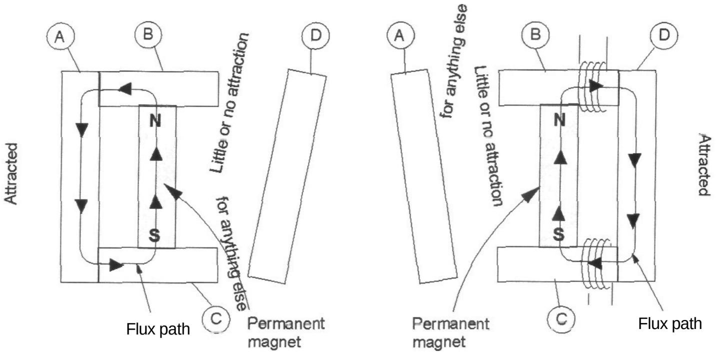

Figure 6-23A.
Once the magnetic circuit is made, there is little attraction for anything else

Figure 6-23B.
A single pulse through a coil can reverse ttie direction of the flux path

Figure 6-23 One example of the Radus effect.

The Radus circuit differs from ordinary permanent magnet circuits in three ways: (i) It exerts strong magnetic force at one end while exerting hardly any at the other, (ii) the strong and weak magnetic poles can be switched end-for-end easily and at will, thus, in effect, turning the magnetism on and off at one given side of the circuit, and (iii) once switched, it remembers its direction of greatest magnetic pull indefinitely.

The original Radus boots were excellent. For the acceptance tests, an engineer clad as an astronaut walked across the bottom of a steel beam in a high bay research area, upside down against the pull of Earth's gravity (Figure 6-24). He *stepped as* he walked, putting his foot "down" and then picking it "up" {447}.

There is no problem in finding modern magnets strong enough to hold the astronaut firmly in such an upside position. The problem with simple magnetic boots using such strong magnets is that, once the foot is planted, the astronaut cannot pick up the foot again. Consequently, the best the astronaut can do with such magnetic boots is walk rather laboriously by "scooting" the foot forward, with the boot remaining in contact and the astronaut unable to pick it up.

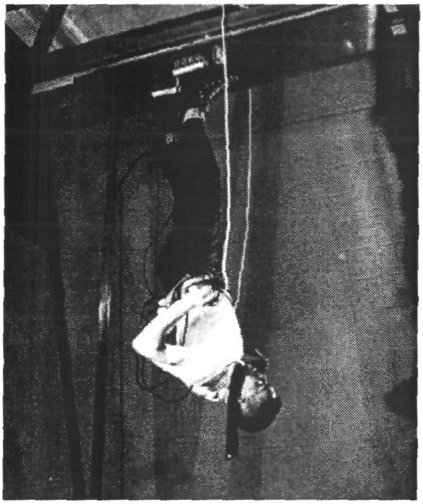

Engineer *walked* upside down on the underside of a steel beam in a high-bay area, wearing the Radus boots He used a stepping movement, not shuffling

Figure 6-24 Acceptance test of the Radus astronaut boots.

The Radus boots completely solved that problem. The permanent magnet fields are switched away from the "boot sole contact with the beam" for that foot that the astronaut wishes to lift, so he can lift it easily and take another step. Then the fields are switched back in again so that the strong field is on the boot sole as he places his foot down. This switching of the fields allows him to walk in a manner resembling normal walking, though a little slower. To do that switching by normal "battery and electromagnet coils" would be prohibitively bulky and heavy — and awkward. With the Radus boots, the astronaut could walk in a manner very similar to slow but normal walking.

However, it is easily seen that an adaptation of the Radus process could conceivably be used to produce a self-switching, self-powered permanent magnet motor. Being a permanent dipole, a magnet is already a particular kind of "free energy generator", since it continuously extracts magnetic energy directly from the active vacuum due to its dipole asymmetry in the virtual photon flux of the vacuum.

The Radus boots were abandoned quickly. NASA then developed the present inferior "shuffler" kind of magnetic boots where the astronaut cannot lift his boot from the surface, but must "scoot" his feet along in a sliding and painfully awkward fashion {448}.

Even today, it is little known that in many virgin magnets fresh from the factory, their very first use conditions them with a Radus-type memory! *That fact can be used, e.g., to create magnets whose fields appear normal, but which deviate from the normal behavior of ordinary magnets, and which produce anomalies in their magnetic fields.*

The Kawai engine {31} uses a novel means of switching the flux path itself instead of overpowering the magnetic poles in an electromagnet. As this is written, Bedini has been notified that his patent with a "path-switching" effect control method has been accepted and will be issued. Although control of the Kawai engine was seized by the Yakuza, we expect to see that some of the first COP> 1.0 motors on the world market will be magnetic motors using the Radus effect or other flux path switching method.

6.4.6 Hole Currents and Electron Currents

In a conducting solid, there are both hole currents and electron currents moving in response to a forcing electric field. In the simple case, electrons move in one direction along the conductor and lattice holes move (migrate) in the other. However, when additional fields are added, the situation can become very complex. The holes and electrons can move oppositely or together in the same direction, depending on the arrangement of the fields and voltages. One can even use fields at right angles to the conductor to affect the currents, and even stop them. This is a rich subject for the inventor, and one that does not appear to have been adequately explored for potential energy and power purposes. This is particularly true when the conductor material is bent or shaped, etc. This area is so broad that it will not be treated here; instead, we refer the reader to Burke {449} for a basic idea of some of the effects and current combinations that can be obtained. From there, an interested experimenter should search the materials science literature on electron and hole current responses of materials.

6.4.7 The Magnetic Wankel Engine

For baseline comparison, Figure 6-25 shows a conventional linear magnetic motor. As can be seen, due to the difference between magnetic poles, a magnetic field exists along the line of the linear track, from the end having the magnets separated the least to the end having the magnets separated the most.

Figure 6-26 diagrammatically shows the scheme of operation of the magnetic Wankel engine. It is basically a linear magnetic motor, wrapped into almost but not quite a complete circle. A set of permanent magnets, each at an angle to the various radial lines of the device, comprises a slightly widening spiral stator that forms most of a circle. A circular rotor with a sector magnet is mounted inside this spiral stator. An end gap exists in the stator as shown, so that the stator is not a completely closed ring. The direction of rotation of the rotor is clockwise as shown. For demonstration of the principle, the beginning air gap between rotor and stator is 0.1 mm and the ending air gap is 5 mm.

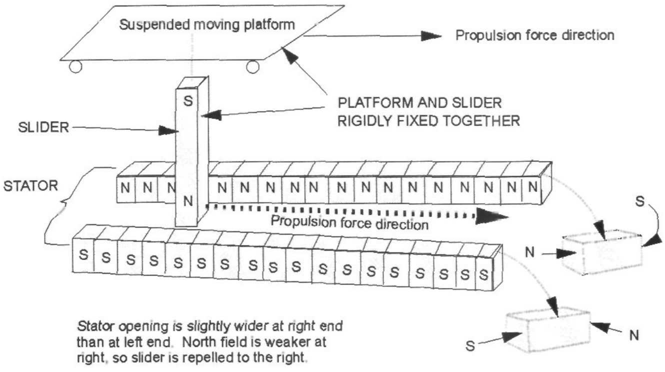
Suspended moving platform
Propulsion force direction
PLATFORM AND SLIDER RIGIDLY FIXED TOGETHER
SLIDER
STATOR
S
N
N N N N N N N N N N N N N N N N N
Propulsion force direction
S S S S S S S S S S S S S S S S S S
S
N
S
N

Stator opening is slightly wider at right end than at left end. North field is weaker at right, so slider is repelled to the right.

Figure 6-25 Type of conventional linear magnetic movement device

An electromagnet is mounted along the perimeter of an angular sector of the rotor, completing the stator gap. The electromagnet is weakly magnetized by a weak trickle current in the associated coil, furnished by a coil with a make-and-break magneto point gap. The electromagnet is magnetized, say, with the north pole facing radially outwards, and the south pole facing radially inside. In the stator, the permanent magnet north poles are facing radially inward toward the rotor, but at an angle, and the south poles are facing radially outward but at an angle.

Tangentially the north pole of the rotor is in a nonlinear magnetic field, and it will experience a clockwise force and acceleration from position 1 where the air gap between rotor and stator is the 0.1 mm minimum, to position 2 where the air gap reaches the 0.5 mm maximum.

If this were all that was involved, the engine would not produce COP> 1.0 because the tangential field is conservative unless additional *free* energy is introduced to overcome the back mmf region in the stator gap. When the rotor crossed the end gap in the stator between point 2 and point 1, very sharp and dynamic magnetic braking due to the back mmf would be done back upon the rotor magnet by the field of the stator magnets at position 1.

Since any real machine will have at least some friction and drag, the actual COP would be less than 1.0.

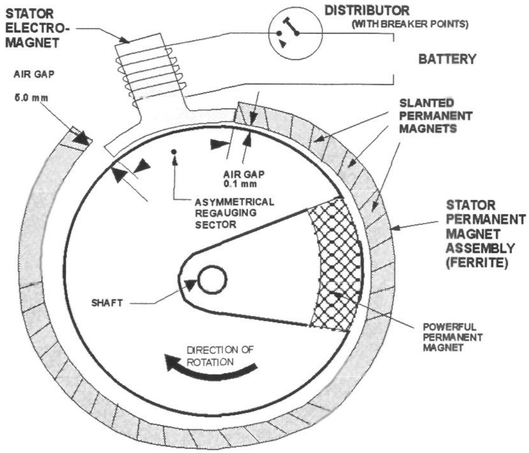
STATOR ELECTRO-MAGNET
AIR GAP
6.0 mm
DISTRIBUTOR
(WITH BREAKER POINTS)
BATTERY
SLANTED PERMANENT MAGNETS
AIR GAP
0.1 mm
ASYMMETRICAL REGAUGING SECTOR
STATOR PERMANENT MAGNET ASSEMBLY (FERRITE)
SHAFT
POWERFUL PERMANENT MAGNET
DIRECTION OF ROTATION

figure 6-26 Magnetic Wankel engine with asymmetrical regauging section.

Let us use the notion of the magnetostatic scalar potential (roughly, magnetic pole strength) to examine a new situation in the end gap.

Technically, let us regard a single unit north pole in the rotor, going from position 1 to position 2 (the acceleration cycle, where the engine will deliver shaft horsepower against a load), and going from position 2 to position 1 (where the magnetostatic scalar potential must be suddenly regauged asymmetrically to equal or exceed the potential at position 1, in order for the rotor to continue unabated or with even further acceleration. That is, when the rotor enters the "back mmf' end gap between position 2 and position 1, a sharp and sudden increase in the "stator magnetostatic scalar potential" must be accomplished, so that the potential in that region is equal to or greater than the potential at position 1. This effect, nearly freely obtained, is what is required for a self-powering magnetic Wankel engine.

In normal machines, conventionally this asymmetrical regauging part of the cycle is where the design engineer forcibly inputs energy from outside the system to do brute physical work on the rotor to forcibly "reset" the machine, and to forcibly wrestle its potential energy storage back to initial conditions. In short, the operator himself arranges to furnish all the excess energy from outside the system that is required to brute force "regauge" the potential at that point, thus effectively creating a multivalued potential instead of a single valued potential. A multivalued potential achieved only by the operator himself furnishing the extra potential energy will not produce COP> 1.0. It will in fact produce $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ of the motor section, but only at the expense of extra energy that the operator himself had to input and pay for. In that case and in a real system with some system losses, the net work out because of $\int \mathbf{F} \bullet \mathbf{dl} \neq 0$ will still be less than the total energy input by the operator.

The forcible "reset" work is conventionally done by simply overpowering the field and reversing it (building it up equal and opposite in the other direction), but with energy totally input by the operator and not "free energy input from the environment" at all. The operator first pays to "kill" the existing field, and then pays to establish a field in the opposite direction.

To obtain COP>1.0 and self-powering, we must trick something else or some other process into furnishing that asymmetrical regauging energy — or most of it — for regauging of the magnetic Wankel engine in that stator gap zone. In other words, instead of engaging in the conventional wrestling match against the back mmf, we must let something else provide most of the energy for the wrestling.

During rotation of the stator from position 1 to position 2, we have been maintaining (and paying a little for) a tiny trickle current and small voltage from the battery into the coil around the electromagnet. As the rotor enters the stator gap, suddenly a sensor sharply breaks the distributor points, momentarily inducing a sudden powerful voltage in the coil. With a very short delay, a very sharp surge of current appears in the coil, producing a sharp and suddenly increased magnetostatic scalar potential (pole) in the gap region. That is the "multivalued potential" effect, where we pay a little to achieve it suddenly at that point, by invoking Lenz's law.

The effect is that suddenly the rotor is raised to the same or greater magnetostatic potential as exists at position 1, and "almost freely though not quite". If equal, the rotor suddenly is in a region with no back mmf, hence it experiences no deceleration braking. If the sudden potential is greater than the potential at position one, the stator in this normally back mmf region now actually experiences a further acceleration (a forward mmf) in that region.

Note that no *radial* mechanical work can be done on either the electromagnetic pole piece (part of the stator) or on the rotor, since neither the stator nor the rotor can move *radially*. However, there momentarily exists a clockwise circumferential magnetic field on the rotor in the stator gap, due to the gradient between the sharply regauged pole piece magnetostatic potential and the potential at position 1.

So we pay a little energy continuously (tiny trickle current, sharply broken breaker points) to get much more energy density momentarily in that small back mmf gap region only. The former back mmf in the stator gap is sharply eliminated by the Lenz force and converted to forward mmf. The rotor experiences a continuous acceleration throughout a complete rotation, due to the judicious use of an artificially induced multivalued magnetostatic scalar potential.

If the average shaft power output during the complete rotation cycle is made greater than the average power input to the asymmetrical regauging circuit during that same rotation cycle, the engine will produce COP>1.0. This type of engine is also easily close-looped, since the excess output is **not** electromagnetic energy but mechanical shaft rotation energy.[193] Hence the problem of the Dirac sea hole current (discussed in Chapter 9) is eliminated.

Such engines have been built and placed in an automobile to power it, in Japan {450}, though there is no information on the exact overall COP. The design was lighter and smaller than a gas engine of the same power, and it was a pygmy when compared to other electric engines of similar power. **The** prototype 45-hp unit weighed 155 pounds compared to 440 pounds for a comparable electric motor. The rotary engine was compact enough to fit inside a two-foot cube. The engine was in development by Kure Tekko, a sizable firm that supplies auto parts to Toyo Kogyo, the Mazda maker. To my knowledge, no hard data on the input electrical power utilized for the trickle current and current-breaker has been made available by the Japanese. The principles, however, are quite clear and easily analyzed.

This may be one of the Japanese COP>1.0 engines suppressed by the Yakuza.

---

[193] For example, a geared or belted arrangement can be used to drive a very small but efficient DC generator that replaces the battery. Many other efficient arrangements are possible.

The researcher might like to consider using a strategically-placed Wiegand wire sensor (discussed in the next section) as the free "generator" providing a pulse of electrical energy to the pulse-magnetizing coil in the stator gap region at the precisely appropriate time. If that or some other similar "self-furnished" pulse of sufficient power can be delivered to the pulse-magnetizing coil, then the system would self-initiate a multivalued magnetic potential in that gap region. In that case, the driving magnetic field around the loop need not be conservative, and self-rotation would be possible without violating any laws of nature or electrodynamics. At least a small toy demonstration model might be possible, simply to illustrate the principle. Presently we know of no one who has tried it.

6.4.8 The Wiegand Effect

If a Permaloy® or other suitable magnetic wire is properly tensioned and worked by repeated torsion while under tension, the skin of the wire is work-hardened and caused to have different magnetic characteristics from that of the internal core. This type of wire is called a *Wiegand wire*, or *pulse wire*, or "self-nucleating magnetic wire" (SNMW™).

When the pulse-wire experiences a certain level of ambient magnetic field strength, it will automatically switch its magnetic state, very sharply, producing a sharp magnetic field pulse. The effect is generally called the *Wiegand effect.*

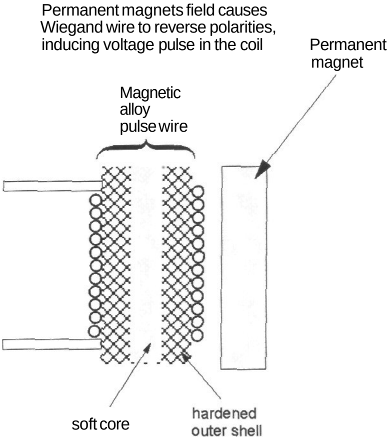
Permanent magnets field causes
Wiegand wire to reverse polarities,
inducing voltage pulse in the coil
Permanent magnet
Magnetic alloy pulse wire
soft core
hardened outer shell

Figure 6-27 The Wiegand effect

If a coil of many turns of fine electrical conductor is wound around the Wiegand magnetic pulse wire, then when the wire suddenly alters its magnetic state, the sudden magnetic field pulse "cutting" the conductor bundle will produce a sharp electrical pulse. Thus the apparatus produces a sharp pulse of electrical energy, when the ambient magnetic field intensity reaches the pulse-initiation value. No input of outside energy is necessary. The apparatus simply gates the energy from the vacuum when it sharply changes its magnetic polarity.

Wiegand effect sensors may achieve voltage pulses up to a nominal 12 volts, in a typical application.

Numerous patents were issued to Wiegand prior to his death {451a-451h}.

Figure 6-27 diagrammatically shows the primary parts of a typical Wiegand sensor application, where the Wiegand wire is also known as a *pulse wire.* The operation of the pulse wire itself is shown in Figure 6-28. Figure 6-29 shows a typical rotary Wiegand effect pulse transmitter.

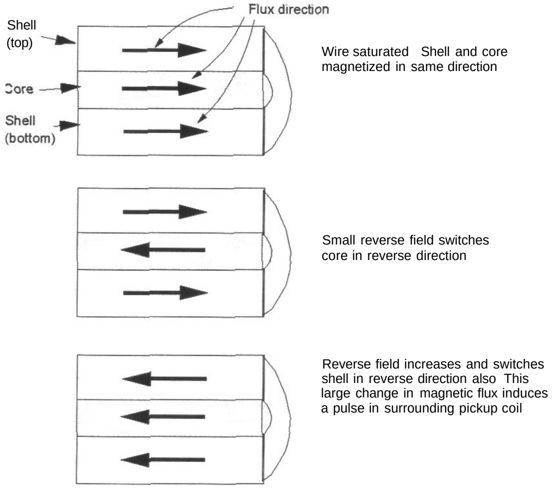
Flux direction
Shell (top)
Core
Shell (bottom)
Wire saturated Shell and core magnetized in same direction
Small reverse field switches core in reverse direction
Reverse field increases and switches shell in reverse direction also This large change in magnetic flux induces a pulse in surrounding pickup coil

Figure 6-28 Operation of the Wiegand effect in a pulse wire.

Pulse wires find many modern "switching" uses, since one does not have to connect them to a power supply in order to obtain the electrical signal pulse generated in a magnetic field as it changes. The magnetic experience of **the** wires can be arranged (using a "resetting magnet") to automatically **reset** the pulse wire to its original condition after the initiation magnet has **fired** it. Many identification cards for personal access to restricted facilities use these wires. Some magnetic memories use the effect. At least one European automobile firm has even used Wiegand wires to trigger the timing of automotive ignition systems.

The pulse characteristics created internally by the Wiegand wire represent a sharp, asymmetrical, self-regauging of the magnetics. The combined magnetic field consisting of the ambient magnetic field and the Wiegand wire magnetic pulse is momentarily a nonconservative magnetic field. It follows that a closed-loop integration $\int \mathbf{F} \cdot d\mathbf{l}$ around some paths in the combined magnetic field of the system do not sum to zero.

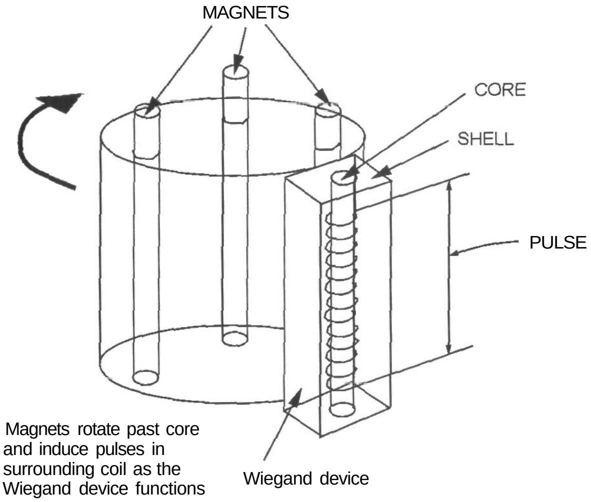
MAGNETS
CORE
SHELL
PULSE
Magnets rotate past core and induce pulses in surrounding coil as the Wiegand device functions
Wiegand device

Figure 6-29 Operation of the rotary Wiegand pulse generator

In theory, at least, it should be possible to utilize something like the magnetic Wankel engine together with exactly self-triggering and self-resetting Wiegand units to form a self-powering permanent magnet motor. In short, referring to our discussion of the magnetic Wankel engine, it might be possible to use a special Wiegand sensor and pulse generator to furnish the properly timed pulse to the coil that enables the continuous operation of the Wankel engine. Obviously this type combination would have occurred to the Japanese scientists working on the magnetic Wankel engine.

We accent that the resulting system would be an open dissipative system, far from equilibrium in its energetic exchange with the vacuum during the Wiegand pulsing. The disequilibrium is achieved by the momentary self-regauging of the magnetic gap in the magnetic Wankel engine.

My colleagues and I have experimented with Wiegand wires and Wiegand sensors to a limited extent. The effect is quite real, fully documented in the scientific literature, and the wires and sensors are not too expensive. A typical coil of 1,000 turns on a 3-cm Wiegand wire will produce pulses of about 2 volts in a 1,000-ohm resistor. The pulse width (half maximum height) is about 20 microseconds. Essentially the Wiegand wires are immune to stray magnetic fields. Viewed on an oscilloscope, the pulses are very clean without spurious oscillations or hash. The field required to switch a typical Wiegand wire is typically about 150 Oersteds. The resetting field is quite a bit smaller, being about -20 Oersteds. With further work, it is probably possible to come up with Wiegand wires whose coils produce up to 20 Volts in a 1,000-ohm resistor. If 25 of these 20-volt Wiegand pulse generators wired in series could be induced to fire simultaneously across a 1,000-ohm resistor, something like a momentary half-watt device could be built. That might be enough to operate a very small version of the magnetic Wankel engine, etc. purely for demonstration purposes.

For practical power, of course, one would need to find a way to increase the power output in each Wiegand generator pulse. So far, no one has been able to successfully find a way to do it, at least to our knowledge. If the reader finds a way to do that, then the reader should certainly patent the process and use it in a self-powering magnetic Wankel engine!

So to those experimenters wishing to experiment with something relatively inexpensive, we would suggest investigating the possibility of incorporating Wiegand units into a magnetic Wankel-type small magnetic motor, and attempt to get a self-powering little unit, or try to build a self-powering version of the rotary Wiegand pulse generator. It is at least possible in theory, and if achieved in practice it will prove that self-powering EM engines are perfectly possible.

6.4.9 Kawai Path-Switching Motor

Japanese inventor Teruo Kawai has invented a process for adroit self-switching of the magnetic path in magnetic motors, rather than directly overpowering the magnetic field and then forcibly reversing it. The process reduces most of the back mmf in an otherwise rather conventionally switched magnetic motor, and results in producing a COP that is approximately double the stated efficiency rating of the unmodified motor. Understand, the Kawai motor always has an efficiency less than 100%, even when its COP>1.0! Hence if one starts with a *high efficiency* magnetic motor — e.g., efficiency of 0.7 or 0.8 — one can obtain a modified motor having a COP of 1.4 to 1.6. Two Kawai-modified Hitachi high-efficiency magnetic motors were tested by Hitachi engineers, and the tests did indeed show COPs of 1.4 and 1.5, respectively, under rigorous and totally independent testing conditions.

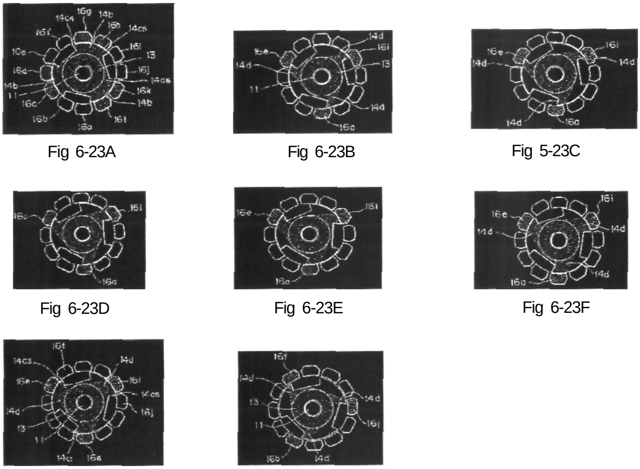
14c4
16g
14b
16h
14c5
16f
16e
16l
13
16d
16j
14b
14c5
11
16k
16c
14b
16b
16l
16a
14d
16e
16i
14d
13
11
14d
16a
16e
16i
14d
14d
14d
16a
16e
16i
16a
16e
16i
16a
16i
16e
14d
14d
14d
16a
16f
14c3
14d
16e
16i
14c5
14d
16j
13
11
14c
16a
16f
14d
13
14d
11
16i
16b
14d

Fig 6-23A

Fig 6-23B

Fig 5-23C

Fig 6-23D

Fig 6-23E

Fig 6-23F

Fig 6-23G

Fig 5-23H

Figure 6-30 The Kawai path-switching process.

Figure 6-30 shows eight snapshots A, B, C, D, E, F, G, and H of the rotor advance of a typical Kawai engine, taken from Kawai's 1995 patent {452}. This is one end rotor/stator side of a two-rotor device, where a similar rotor/stator device is on the other end of the central shaft *11.* In snapshot A, pole piece *14* had three outward teeth *14b* dispersed equally around the circumference, alternated with three notches. An end magnet *13* provides the source of flux passing through the pole piece. With the electromagnets DC-energized, their core materials *16c, 16d, 16g, 16h,* and *16k, 16l* are shown shaded by flux from central magnet *13* out through teeth *14b.*

In snapshot B, electromagnets *16a, 16e,* and *16i* are energized. The shaded area shows the sharp convergence *14d* of the flux from magnet *13* through pole piece *14* and the edge of teeth *14d.* Since the electromagnets are magnetized in attracting mode, the rotor will experience a torque tending to widen the flux path from magnet *13* to the activated electromagnets. Thus a clockwise torque exists on the rotor, and it will start to rotate clockwise.[193] Note that each electromagnet is operating independently of the other two.

As shown in snapshots C, D, E, and F, the rotation of the rotor continues clockwise, widening the connecting flux path to the three activated electromagnets. During this time the torque on the rotor is clockwise.

In snapshot G, the flux path to the activated electromagnets is fully widened. In addition, the leading edges of the three teeth are just beginning to enter the domains of the next electromagnets *16j, 16b* (see snapshot A) and *16f.* This is getting symmetrical to the original position shown in snapshot B.

Now the electromagnets *16i, 16a,* and *16e* of snapshot G are deactivated, as shown in snapshot H, and electromagnets *16j, 16b,* and *16/*are activated. This action asymmetrically self-regauges and resets the engine back to the original situation shown in snapshot B, but one more electromagnet beyond. The action cycle begins anew. As can be seen, in each complete rotation of the shaft, each of the three teeth of the rotor will be asymmetrically self-regauged multiple times. So multiple total asymmetrical self-regaugings, resettings, or "refuelings" are utilized per shaft rotation.

In each motor coil, at energization a tooth is just entering that coil, Energized in attractive mode with respect to the ring magnet around the shaft, the flux in the pole piece jumps from fully widened flux (and small **or** vanishingly radial torque on the rotor) to angled and narrowed flux (with full radial clockwise torque on the rotor). This is an asymmetrical self-regauging action. As previously explained, the narrowed flux and its angle exert a clockwise accelerating tangential component of force upon the rotor. By the asymmetrical self-regaugings of the engine, each coil is de-energized prior to beginning to exert radial back emf— which it would do if it remained energized as the trailing edge crossed it and again narrowed the flux path).

So the Kawai engine uses normal magnetic attraction to accelerate the **rotor** for a small distance, and then asymmetrically self-regauges to zero attraction to eliminate the back-mmf drag portion of the attractive field. It regauges to zero as the reset condition.

---

[193] The Kawai patent contains a small error of misstatement. It states at first that the torque and rotation are counterclockwise, and then shows it moving correctly clockwise from figure to figure in the series.

For appreciable power and smoothness, the typical Kawai engine uses an extensive number of asymmetrical self-regaugings per axle rotation, for example 36 times on each end, or a total of 72 times for the two ends. The force field of each coil, accompanying its increased magnetostatic scalar potential, is oriented radially inward, so that radial work cannot be done by the coil on the rotor because the rotor does not translate radially. Advantage is taken of the initial clockwise acceleration force initially produced, and self-regauging eliminates the counterclockwise decelerating drag force that would be produced without the regauging.

The major benefits are that (a) multiple self-regaugings occur for a single rotation of the rotor assembly, enabling high weight-to-power ratio, (b) each electromagnet is energized only when positively contributing to the clockwise torque that drives the rotor, and (c) each coil is de-energized to asymmetrically self-regauge the system during those periods when the coil would otherwise create back-drag (counterclockwise torque) if it remained energized.

A conservative cycle is one in which the work done in the back-drag (mmf) region is equal to the work done in the forward boost (forward mmf) region. Self-eliminating the back-drag portion of the cycle is a form of asymmetrical self-regauging, and makes the net field highly nonconservative. The Kawai approach is thus permitted to attain COP>1.0. In his U.S. patent, Kawai quotes performance measurements showing COP = 3.18. As we stated, Hitachi engineers measured two Kawai-modified Hitachi magnetic motors which exhibited COP =1.4 and 1.6 respectively.

Since the output is shaft horsepower, in principle the Kawai engine can be close-looped for self-powering, without consideration of the Dirac sea hole current. With its external close-looping system (not shown), it falls within the second close-looping class, that of an "outrigger" system. Only the unitary system closed-looping class requires consideration of hole currents in the Dirac Sea.

The Kawai process and several other Japanese overunity systems have been blocked from further development and marketing by the Japanese Yakuza. The present author and the Board of Directors of CTEC were physically present and in final negotiations (having reached agreement!) with Kawai to market his self-powering motors worldwide, when control of the Kawai company and process were suddenly seized on the spot in our presence.[195]

6.4.10 Bedini's Skin Depth Magnetic Switching

Bedini found a most interesting variation of flux path switching: he switches the direction of just a little fraction of the ongoing magnetic flux in a material flux path, on the skin of one side of it and with only a slight "skin depth".[196]

Suppose a flux path material is used as the stator pole of a permanent magnet motor. A large permanent magnet is used to magnetize that short length of flux path, as shown one the left in Figure 6-31, which illustrates a concept for a magnetic motor. For convenience, imagine that the small magnet on the right is not yet placed on the material flux path extension. The flux path extension is energized by the south north pole of the large permanent magnet on the left, and used as the stator in a rotary magnetic motor.

---

[195] To anyone who understands Japanese culture, Kawai paid me one of the greatest compliments possible. After his company representative called from Japan, Kawai shipped his *original proof of principle engine* — which first successfully demonstrated his flux-path switching process without being optimized for COP>1.0 — to me, here in the U.S. In the Japanese manner, he had entrusted me with his single most valued possession. From years of aikido experience, I fully understood the meaning of this unprecedented gesture, and expressed my deep appreciation accordingly. From our first meetings with Kawai and his group, I have and always shall have the utmost respect for Teruo Kawai. In my view, it is a sad loss to humankind to have his engine buried and withheld so brutally. Kawai had already achieved a closed-loop self-powering version, and it is this version upon which we reached agreement. In less than 24 hours after that agreement, the Kawai engine and company — and Kawai's own fate — were seized and no longer in Kawai's hands. For those unfamiliar with the deep spread of the Japanese Yakuza, the following are recommended: (a) Brian Bremner, "How the Mob burned the Banks: The Yakuza is at the center of the $350 billion bad-loan scandal," *Business Week,* Jan. 29, 1996, p. 42-43, 46-47; Michael Hirsh and Hideko Takayama, "Big Bang or Bust?" *Newsweek,* Sept. 1, 1997, p. 44-45; Adam Johnston, Yakuza: Past and Present," downloaded from the Internet, Committee for a Safe Society, Organized Crime Page: Japan: History and present status of Japanese organized crime.

[196] Close colleague Ken Moore also independently discovered this effect in his own experiments, but then we found that Bedini had discovered it earlier. That was fine with Ken; he greatly admires Bedini and the marvelous discoveries Bedini has made. I greatly admire both of them, and am happy to call them best friends. After all, science has a human side as well as a materials side. Without one's close colleagues and best friends, science alone would be a miserable enterprise!

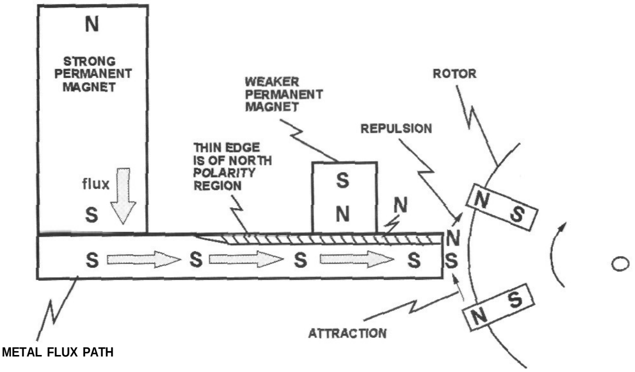

Figure 6-31 Bedlni skin depth magnetic edge switching.

As can be seen, the north pole of the rotor, shown in the figure, is attracted to the resulting south pole of the stator. Power can be taken from the shaft of the rotor while it is being attracted in to the extended north pole of the stator flux extension. As it begins to pass the extension end, obviously if nothing else is done it will develop back mmf "drag" by being attracted back toward the north pole of the flux extension as it passes it.

However, suppose that —just when the rotor south pole is passing dead center of the stator north pole extension, one suddenly pulses a small flat coil placed on the face of the extension, which serves the function of the small permanent magnet shown on the right of the material flux path extension. Further, suppose the magnetic field produced by the coil is oriented as the field of the small permanent magnet, but is much weaker than the magnetic field of the large permanent magnet on the left. A very curious effect then occurs. This energized coil magnetizes the edge of the flux path extension, to a small "skin depth" as shown, with a *north* polarity. The magnetic flux is now bidirectional in that same material flux path extension.

Further, this edge-switching action is performed so that it is this "partial north polarity" of the now-altered extension that faces the rotor north pole as it begins to pass the extension and enters what previously would have been a back mmf section. Instead, now that former back mmf section is a weakened forward mmf region, *further accelerating the rotor.* This action breaks the Lorentz symmetry around the rotation path of the rotor, resulting in a net nonzero (nonconservative) magnetic field. Hence the entire rotor exhibits a net drive force around its loop. The rotor will accelerate until the mechanical work performed upon the shaft, friction, etc. is equal to the net magnetic energy furnished to the rotor.

The beautiful part of this motor is that one does not have to "overcome" by brute force — and then reverse — the north polarity of the entire stator extension. In theory, the overall energy added to the rotor is greater than the energy that must be added for the "edge switching" of the magnetic flux path extension, because the entire back mmf of the motor has been eliminated and replaced by a weaker forward mmf. If the process of furnishing the "skin depth" weaker magnetic north field to the edge of the flux path extension is made sufficiently efficient, this engine also can be made self-powering, at least in theory.

The former "motor" design has been converted to a "motor/generator" design, by the addition of the edge-switching action. The novel factor is the use of bidirectional flows of magnetic flux in a single material flux path.

An additional improvement is to rig a Wiegand sensor and pulse generator or sensors in conjunction with the rotor action, to freely furnish and kill the edge-switching pulse to the face-magnetizing coil on the flux path extension, at the appropriate times. This is an area where even high school students will be able to experiment.

Bedini's patent on the edge switching process and its various embodiments and improvements has been issued {453}.

6.4.11 Bedini's Flux Path Switching with Energy Capture

One additional embodiment of the Bedini flux path switching motor is shown in Figure 6-32. Here the entire flux path is switched, by moving a small coil on the bottom side of the material flux path, as shown in the figure. As Radus showed {454}, it turns out that a much weaker magnetic field in that left bottom position will switch the flux path of a powerful permanent magnet field at the left top of the path, as shown.

Now, however, Bedini adds a transformer coil around the right end of the core (the right end of the extension is still the stator pole in the magnetic motor diagrammatically shown in Figure 6-32). When the flux path is switched back and forth, this transformer coil produces an AC signal output, which can power a load as shown, and can even be used to power the switching power supply. The trick is that it requires less energy to make the vertical flux path preferable to the horizontal flux path, than the energy delivered into the flux path by the permanent magnet. The potential **for** self-powering of a sufficiently efficient magnetic motor is obvious.

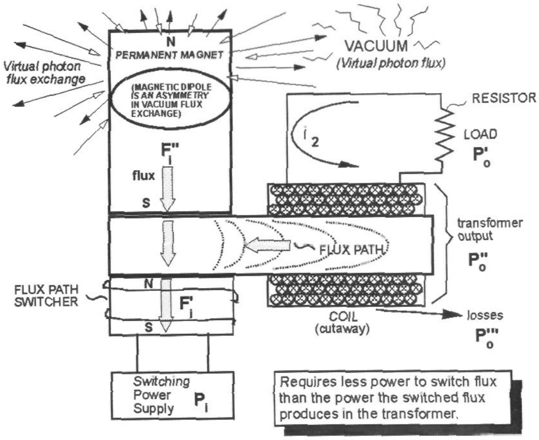
N
PERMANENT MAGNET
Virtual photon flux exchange
VACUUM
(Virtual photon flux)
(MAGNETIC DIPOLE IS AN ASYMMETRY IN VACUUM FLUX EXCHANGE)
RESISTOR
i_2
LOAD
P'_o
F''_i
flux
S
FLUX PATH
transformer output
P''_o
FLUX PATH SWITCHER
N
F'_i
S
COIL (cutaway)
losses
P'''_o
Switching Power Supply P_i

Requires less power to switch flux than the power the switched flux produces in the transformer.

Figure 6-32 Bedini's total path switching and energy recovery process in a transformer.

So the motor can be powered while simultaneously acting as a partial generator. For efficient switching, this motor/generator can in theory develop COP>1.0 and can even be made close-looped and self-powering. We strongly accent that the primary excess energy input to the system is from the vacuum itself, via the broken symmetry of the large permanent magnet dipole at the left top of the material flux path. As we have repeatedly stated, the only energy problem is how to capture and utilize the EM energy freely received, transduced, and poured out from the vacuum by the source dipole's broken symmetry, without using half of it to destroy the source dipole.

6.4.12 Transformer Secondary Shunted by a Negative Resistor.

See Figure 6-33. This embodiment assumes the availability of a true negative resistor — possibly an adaptation of a point contact transistor or other device. As shown in the figure, a transformer exhibiting COP> 1.0 can be developed by adding a shunting *true* (not differential!) negative resistor across the terminals of the secondary, adjacent to the secondary. In this manner, the Lorentz symmetry of the closed unitary current loop containing the external loads and losses, and normally the secondary of the transformer in series, is broken.

Part of the return spent current is now "pumped" back up from ground side to power side of the transformer secondary without passing through the secondary. The pumping work is providing by dissipation (in the negative resistor) of energy freely obtained from the active vacuum environment. In short, one adds a negentropic process in parallel with the secondary coil, in the secondary external circuit, breaking the otherwise enforced Lorentz symmetrical regauging of the discharge of the collected EM excitation energy in the secondary circuit. This provides the capability for a permissible COP>1.0 transformer.

ON EACH HALF CYCLE, ONE OF THE NEGATIVE RESISTORS SHUNTS PART OF THE RETURN CURRENT AROUND THE SECONDARY, FROM LOWSIDE TO HIGH SIDE, THUS DECREASING THE BACK-FIELD COUPLING FROM SECONDARY TO PRIMARY ACROSS THF TRANSFORMER. THIS LESSENS THE ENERGY DISSIPATION REQUIRED IN THE PRIMARY FOR A GIVEN ENERGY FLOW TRANSFER FROM PRIMARY TO SECONDARY

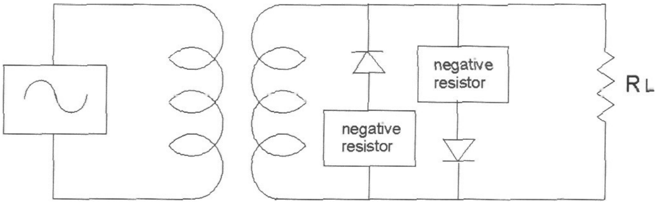

Figure 6-33 Transformer shunted by a negative resistor.

In theory, if the negative resistor can bypass all the current in the secondary, or almost all of it, the transformer can become a COP»1.0 device.

This overunity transformer can use step-up of the voltage from primary to secondary, and exhibits an asymmetry in its forward and backward field coupling between primary and secondary. Thus the dissipation of energy in the primary circuit need not be as great as the dissipation of energy in the secondary circuit. The *energy flows* are of course in perfect conservation, but there is no law of nature requiring the *energy dissipations* to be conservative between the primary and the secondary. It is strictly the backfield coupling from secondary to primary that enforces equal energy dissipation in the primary as in the secondary, in normal transformers. And that backfield coupling's strength is a function of the return current through **the** transformer secondary. By reducing that current, the back coupling to the primary is reduced, which in turn reduces the current draw of the primary from the external power supply.

No laws of physics, electrodynamics, or thermodynamics are violated by the asymmetrically coupled transformer power system with broken symmetry in its secondary to primary circuit couplings. In short, there is less backfield coupling from primary to secondary than there is from secondary back to primary. The additional energy dissipated in the secondary circuit is freely extracted from the vacuum by the negative 4-resistor shunt.

Again, this application assumes the availability of a true negative resistor, to use as a shunt of the secondary.

The beauty of this application is that, once achieved, such a COP>1.0 transformer can easily be close-looped for self-powering by standard "governed" positive feedback. No concern as to hole current effects usually need be accounted for unless the COP becomes very large.

6.4.13 Eddy Current Multiplier.

Figure 6-34 shows an example of an eddy current multiplier, adapted from Burke {455}. This is given as a somewhat unusual system which may exhibit novel effects and possibly can be adapted into a COP>1.0 system (e.g., by using permanent magnets instead of the coils and triggering things into unbalanced condition using a rotary pulse wire generator. We simply list it here as a candidate for the experimenter.

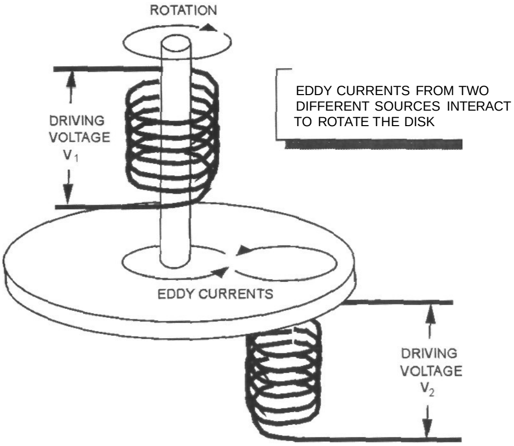

Figure 6-34 Eddy current multiplier.

In Figure 6-34, the shaft of the conducting disk is connected to a rotor, which could have arrangements for rotary type of Wiegand pulse generator of Figure 6-29 (with self-initiation and self-recovery) around it (not shown in Figure 6-29). One would start the rotating disk by hand. These Wiegand sensors would then automatically and repeatedly deliver a series of electric pulses to the coils, momentarily generating the two eddy currents shown, so that dynamic force and rotary acceleration is obtained during the endurance of the pulses. In short, one would "electrically kick" the coils so that they continually deliver a propulsive "magnetic kick" to the conducting disk to repeatedly generate eddy currents to rotate it. The rotor is powered as shown by the interactions of the two unbalanced induced eddy currents. Such an arrangement could possibly exhibit COP>1.0.

To experiment with such an apparatus, the major parts that must be developed are (i) the rotor and coils arrangement (the motor), (ii) the integrated rotary Wiegand effect pulse generator with self-recovery, and (iii) the proper timing arrangement. Each should be developed and optimized individually in the researcher's experiments, then combined and optimized together. This is an experimental area that can be cheaply investigated, e.g., by a college student or high school student for a science project. Whether COP>1.0 is achievable or not depends upon the efficiencies that can be obtained with the various processes involved. It is possible, at least in theory, for such an arrangement to be self-rotating once rotary motion is underway.

6.5 CONCLUSION

In this chapter we have advanced many approaches to COP>1.0 systems, that have been tried or considered by various inventors and researchers in the field. Many of these experimenters have attained concrete, replicable results, including results published in the hard literature.

We hope this sampling shows the reader that indeed we can have "self-powering" electrical power systems and COP>1.0 electrical power systems whenever the scientific community will come out of its preoccupation with the present seriously restricted electrodynamics and electrical power engineering practice and fund the necessary research. Simply funding a few graduate students for research and doctoral theses in these areas would be very useful and immensely more productive for electrical energy purposes than all the billions and years spent on chasing hot fusion.

Hopefully the iron dogma of the scientific community against COP>1.0 electrical power systems is slowly changing. If so, and if the community will turn loose those sharp young graduate students and post-doctoral scientists on the problem, very shortly there will no longer be an electrical energy problem on Earth, ever again. We strongly refer the reader to Deffeyes' book {456} to rigorously validate the coming increase of the price of oil, in the middle of the present decade. See particularly the quotation from Deffeyes in our Endnote {342}. With the present extreme volatility of the MidEast and the continuing war on terrorism, the probability is high of eventual severe disruption of oil supplies and the energy infrastructure of the U.S.

An independent review of Deffeyes' book is given by Young {457} and we recommend that the reader digest that commentary also.

If the scientific community will fund the necessary research on permissible COP>1.0 Maxwellian systems, the quick results will solve the energy crisis, clean up the biosphere — including much cleaner air in our cities for the populace to breathe — provide cleaner streams, and result in far less polluted oceans. It will stop the pollution-caused destruction of species, and dramatically reduce the hydrocarbon combustion contributions to global warming. There will no longer be a necessity to consider "burying" in the bottom of the sea the excess CO2 from massive hydrocarbon combustion to provide the much-needed electrical power, because it will no longer be necessary to burn the oil, coal, and gas to get the power. And there will no longer be any necessity for expensive and highly vulnerable nuclear power plants with their resulting nuclear wastes that will remain radioactive for thousands of years.

If cheap, clean EM energy from the vacuum is rapidly developed, we may *finally* get the scientific community's head wrenched back out of the big nuclear power business, as a colossal waste of research manpower and taxpayer dollars — as well as providing ticking time-bombs for the terrorist teams now already in the United States to attack and destroy. It is so easy for a person with appropriate technical knowledge to build a portable electromagnetic pulse (EMP) "shooter" that one shudders at the consequences of in-country terrorist teams that can do so. Try imagining what we would face when the electronic controls of one or more operating nuclear power plants are suddenly disabled and destroyed with the reactor on line and providing power. Chernobyl might look like a mild spring breeze in comparison, should we have one or more actual meltdowns.

It is not a pleasant consequence to contemplate, and that alone may eventually dictate the overwhelming urgency of achieving decentralized, vacuum-energy-powered electrical power systems as soon as humanly possible. The present centralized, cumbersome energy infrastructure is vulnerable beyond all sanity, considering the thousands of terrorists already in country, with weapons of mass destruction as well as more conventional sabotage means. As a single example, the former Soviet Union clandestinely brought in nuclear weapons into the U.S., and

Spetznaz teams to detonate them on order. Such weapons (and teams) are already in large U.S. cities and population centers. The reader is referred to Lunev's book,[197] for details about how the nuclear weapons were introduced.

After all, we do have to live and survive in the real world, and not just in an idealistic scientific environment. We are already at war, with alerts of potential danger to our nuclear power plants already having been given on several occasions. In the present very disorderly and very dangerous world we live in, many of the former options we have had for decades are rapidly ending. That includes the option for the leaders of the scientific community to continue "business as usual", if they wish this nation and their own families to survive. There is an old adage in war: *Whatever your determined enemy can do to hurt you desperately, he will do as his highest priority.* Sadly, in my opinion the leaders of the scientific community — unwittingly — are presently doing almost everything they can do to insure that the energy infrastructure goes, the national economy goes, a nuclear meltdown possibly occurs, and this nation and much of civilization perishes in a new dark ages.

---

[197] Stanislav Lunev (with Ira Winkler), *Through the Eyes of the Enemy,* Regnery, Washington, 1998. Lunev is the highest ranking GRU defector. He is a former Colonel in the GRU, the military counterpart of the KGB. On p. 22-33, he summarizes the Spetznaz capabilities. On p. 22-27, he summarizes Spetznaz use of nuclear weapons already on American soil. On p. 26, he gives some of the ways in which the Soviets easily brought nuclear weapons into the U.S. On p. 30 he confirms **Russian** seismic weapons. Use of EMP weapons is on p. 30-31. Use of very, very low frequency weapons to destroy the human brain, put people into a zombie-like state, and aid in brainwashing is confirmed on p. 31. The present war on terrorism has many more facets than are presented by the conventional news media.

Just as the dinosaurs were confronted with a new and dangerous threat, and failed to adapt, the electrical power infrastructure is now confronted with a potent threat against its very vitals and — so far — is completely failing to adapt. It is going to have to adapt or perish, which presently implies that we all perish along with it. Those who make it their nefarious business to see to it, will see to it.

So let us "see to it" first, scientifically, as rapidly as human research and development can do the job. The very existence of civilization may well depend upon it.

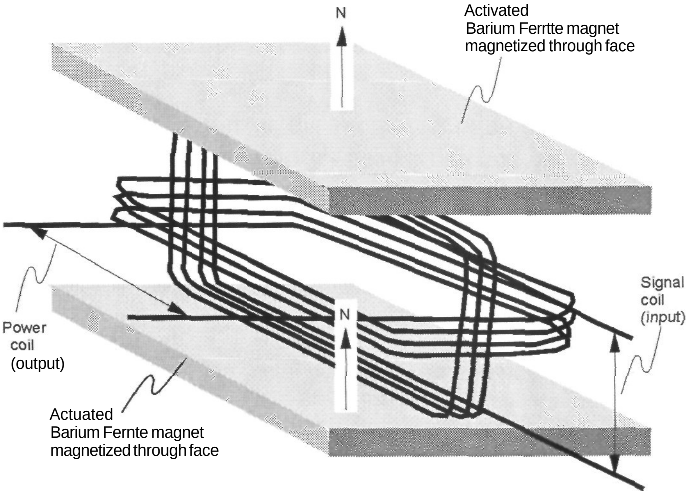
N
Activated
Barium Ferrtte magnet
magnetized through face
Signal
coil
(input)
N
Power
coil
(output)
Actuated
Barium Fernte magnet
magnetized through face

Figure 6-5 Sweet vacuum triode amplifier (VTA) construction.

Chapter 7

Aharonov-Bohm Effect, Geometric Phase, and the Motionless Electromagnetic Generator (MEG)

[Decision that hydrogen fuel is the future]. *"We are at the peak of the oil age but the beginning of the hydrogen age. Anything else is an interim solution. The transition will be very messy, and will take many technological paths ...but the future will be hydrogen fuel."* [Herman Kuipers, Royal Dutch Shell.].

[Vacuum energy is real] *"...the acceptance of a structured vacuum described by an O(3) gauge group leads directly to the existence of novel charges and currents in the vacuum. These are conserved, or Noether, currents and charges and are clearly topological in origin. They spring from the fact that the vacuum is a topological space. Four such entities emerge: [1] A topological vacuum electric charge, also proposed empirically by Lehnert et al. [2] A topological vacuum electric current, also proposed empirically by Lehnert et al. [3] A topological vacuum magnetic charge, proposed also by Barrett and Harmuth. [4] A vacuum topological magnetic current, proposed also by Barrett and Harmuth.*

Each of these four objects can provide energy, which can be loosely termed 'vacuum energy': energy coming from the topology of the vacuum. " [Evans] {458}.

[Aharonov-Bohm effect] *"...the Aharonov-Bohm effect is a local gauge transformation of the true vacuum... [which] produces a vector potential from the true vacuum. [This gauge transformation] produces topological charge..., the electromagnetic field, which carries energy, and the vacuum charge current density first proposed by Lehnert ... and developed by Lehnert and Roy... ".* [Evans and Jeffers] {459}.

[Geometric phase] *"The concept of the geometric phase is closely related to the effect of a state function acquiring a non-zero phase factor after a quantum system undergoes a cyclic evolution. This means, that if a system after some time returns to its original state, the result of such an evolution will be recorded in the phase of the wavefunction. Moreover, this phase factor can be measured by interfering initial and final states. ... Although there are no widely recognized practical applications of the geometric phase, its experimental observations have been reported in many fields of science. The largest group of experiments have been carried out on polarized light ...and polarized neutrons ... The geometric phase has also been observed in magnetic resonance experiments ..., mesoscopic structures ...and molecular systems ...Analogues of GP — the Hannay angles — have been shown to exist in classical mechanical systems ..., the most famous example of which is the Foucault pendulum."* [Aleksiejunas][198]

[Practical use of geometric phase]. *"The motionless electromagnetic generator is the first practical macroscopic use of the geometric phase. "* [T. E. Bearden, 2001, to a correspondent].

[Impact of extracting energy from the vacuum] *"If they [quantum fluctuations of vacuum] can be [tapped], the impact upon our civilization will be incalculable. Oil, coal, nuclear, hydropower, would become obsolete — and so would many of our worries about environmental pollution." "Don't sell your oil shares yet — but don't be surprised if the world again witnesses the four stages of response to any new and revolutionary development: 1. It's crazy! 2. It may be possible — so what? 3. I said it was*

---

[198] I am much indebted to Dr. Rimvydas Aleksiejunas for this clear statement of the geometric phase. See his very elegant summary, " Introduction to The Geometric Phase Decomposition," on his Internet website at http://signalogram.free-hosting.lt/Research/gp dec.htm. His own publications in scientific journals and proceedings are also listed on **the** website, and many are available for downloading. His doctoral thesis dealt **with geometric phase** and he is a specialist in that field.

> *a good idea all along. 4. I thought of it first."* [Arthur C. Clarke] {460}.

7.1 Introduction

In this chapter we give a history, some information, and comments about the Aharonov-Bohm effect {473}, which was generalized by the Berry phase {477} for adiabatic conditions, which itself was further generalized by Aharonov and Anandan {480} to remove the adiabatic limitation. The modern all-inclusive term is *geometric phase,* which has been found to be involved in a very wide variety of physics areas. In its embodiments, the motionless electromagnetic generator (MEG) does utilize the Aharonov-Bohm effect (geometric phase) in a unique manner that we explain. We also remark on what we hypothesize may be some related new aspects of the curl-free magnetic vector potential itself.

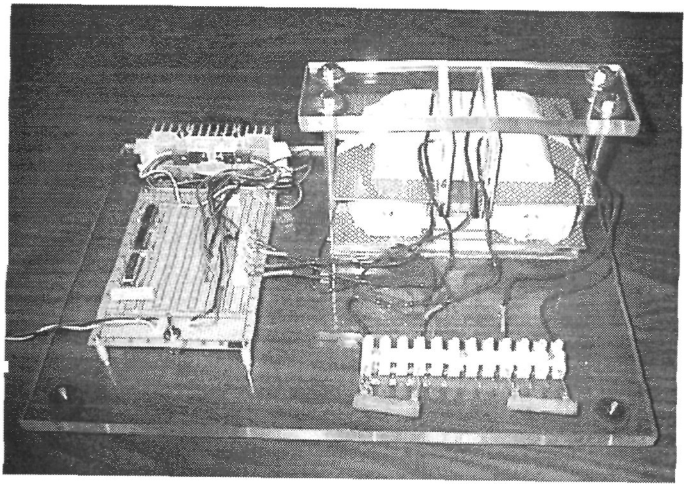

Figure 7-1 Motionless Electronic Generator (MEG)

The motionless electromagnetic generator (MEG) (Figure 7-1) was invented by Patrick, Bearden, Hayes, Moore, and Kenny {461} after years of research and experimentation. It has been technically described elsewhere {462}, and scientific explanations of the fundamental mechanism by which it successfully extracts useful EM energy from the active vacuum have been published in the recognized literature {463, 38a, 464}. Other closely related papers on extracting EM energy from the vacuum have been published or are submitted and in review {465a-465h}.

A version of the MEG has also been independently replicated by Naudin {466}, and in fact by others who did not wish to publish their results. By permission, Naudin's Mark II MEG replica is shown in Figure 7-2. Consequently, the MEG is a validated and replicated COP>1.0 system, as required by proper scientific methodology. Further, sufficient information has been released for any competent laboratory team to replicate a version of the MEG.

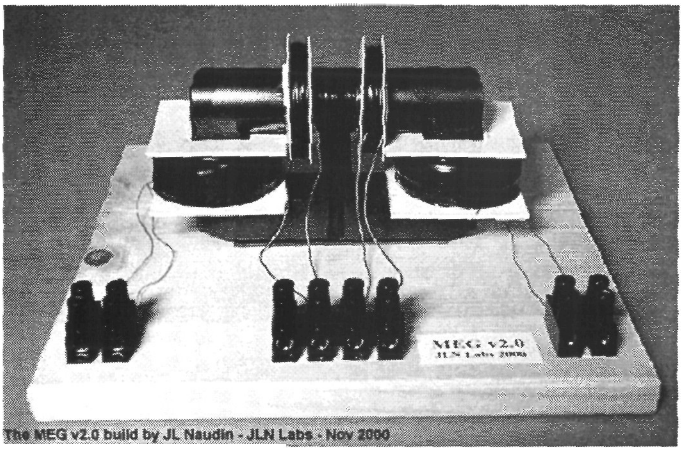

Figure 7 2 Nuadin's Mark II Replication of the MEG.

We give an overview of the MEG, the principles of its operation, the status of the project, and some indication of future plans.

7.1.1 Status

As this is written, the first MEG patent has just issued {37}. A second Provisional Patent Application has been filed, and at this writing is in process of being replaced by a formal regular patent application. Two other patent applications are in preparation. All rights to the MEG are assigned to Magnetic Energy Ltd., Huntsville, Alabama, U.S.A., whose CEO is Dr. James L. Kenny.

In addition, Bedini and the present author have filed a provisional patent application on the process and embodiment for close-looping a COP>1.0 system for self-powering operation, and are in process of replacing it with a regular patent application also. Magnetic Energy Ltd. has permission to utilize the Bedini-Bearden close-looping process. In addition, Magnetic Energy Ltd. has its own proprietary close-looping process, for which a patent application is being prepared. The Company is also involved in serious discussions with several large companies and financial groups for potential financing. When that funding is obtained, we will open a substantial research and development laboratory here in the United States, serving a separate production company to be set up with our financial partners to produce and market the MEG. Meanwhile, laboratory work to complete the research and development of the MEG is also ongoing under a cooperative arrangement with the National Materials Science Laboratory of the National Academy of Science of a friendly foreign nation formerly under the Soviet bloc.

7.1.2 Future Plans

Our plans are straightforward. We will continue to pursue and update our patents and file additional ones as additional discoveries are made. We will continue our negotiations with major financial parties until satisfactory and mutually beneficial capitalization is obtained (we expect that to occur before the end of 2002). At that time we will establish a modern Magnetic Energy Ltd. U.S. laboratory devoted to advanced research and development, as well as a separate production company in partnership with our funding partners. The final research necessary on the MEG (four areas of physics as well as the electrodynamics) will be completed in somewhat less than one year after setup of the lab and lab team, and the production company is expected to have the first engineering development production prototype at the end of that first year. We expect it to be a self-powering unit. We will simply put the first MEG units into production and on the world market immediately thereafter.

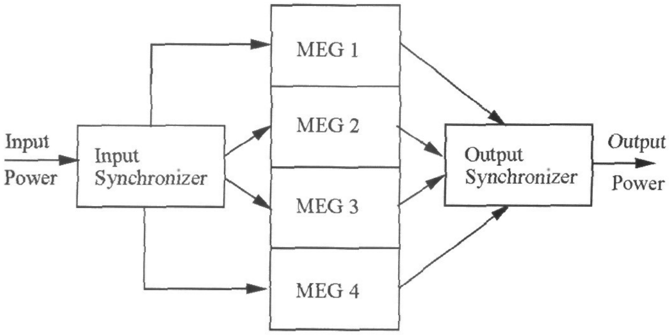

Figure 7-3 Block diagram of envisioned **modular** MEG design.

See Figure 7-3. The first production MEG is envisioned as a modular 2.5 KW unit together with a synchronizer unit simultaneously developed, so that up to six 2.5 KW MEG units can be synchronized in a power system array. This will yield power units from 2.5 to 15 KW. We are struggling to have the closed-loop process completed and ready in that same first year, and if possible the very first MEG units will be close-looped and self-powered. At the end of the second year or before, we plan to have the close-looped basic MEG unit in 10 KW size, again with synchronizer so that up to six MEG units can be synchronized in an array. That will cover the 10 - 60 KW power system size, which is ideal for homes and small businesses. Once the units go on the world market, we believe that the market will be extraordinarily dynamic and productive. We intend to license multiple companies for production and competition for various markets, with no single company being given total monopoly worldwide.

7.2 Aharonov-Bohm (Geometric Phase) Effect

The Aharonov-Bohm (AB) {473} effect, or geometric phase (GP) {477, 480} effect, is now well known {467} in quantum mechanics and quantum electrodynamics, but does not exist in classical electrodynamics. The effect proves the reality of the potentials and their primacy over the fields. Further, it shows that electron effects can be generated by the magnetic vector potential A even in the absence of the magnetic field B.

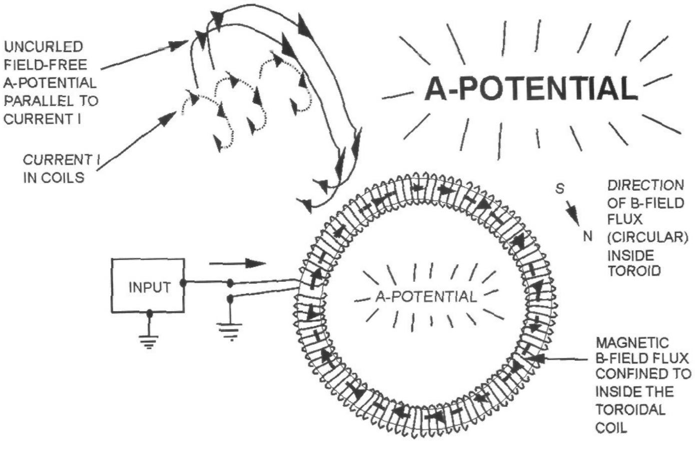
UNCURLED
FIELD-FREE
A-POTENTIAL
PARALLEL TO
CURRENT I
CURRENT I
IN COILS
A-POTENTIAL
S
N
DIRECTION
OF B-FIELD
FLUX
(CIRCULAR)
INSIDE
TOROID
INPUT
A-POTENTIAL
MAGNETIC
B-FIELD FLUX
CONFINED TO
INSIDE THE
TOROIDAL
COIL

Figure 7-4 Aharonov Bohm effect for a tightly wound toroidal coil.

According to Feynman {468}, it required 25 years for quantum physicists to clearly face the issue eventually pointed out in 1959 by Aharonov and Bohm of the primacy and separate action of the force-field-free potential, in regions where no force fields were present (Figure 7-4) {469}. After the effect was experimentally demonstrated by Chambers {474} in 1960, it then required another 25 years before physicists would accept it. Even today, another 16 years after full acceptance in 1986, to our knowledge the Aharonov-Bohm effect has not previously been utilized in any practical electrical power system {470}, much less one that has a COP>1.0. As Aleksiejunas[198] states (see his quotation at the beginning of this chapter), there have not previously been any widely recognized practical applications of the geometric phase. Now there is a practical application of major worldwide importance, since the operation of the MEG depends upon the geometric phase (Aharonov-Bohm) effect to produce real electrical power from the active vacuum itself.

We also believe that the curl-free potential may not have been adequately investigated, particularly along the lines of its utilization in the MEG.

7.3 Abbreviated History of the Geometric Phase

The earliest work with a state function acquiring a non-zero phase factor after a quantum system undergoes a cyclic evolution was published in 1956 by Indian scientist Pancharatnam {471}. Pancharatnam noticed phase shifts of cyclic evolutions in optical phenomena. This means that, if the system leaves its original state and after some time returns to it, the result of the system's evolution will be recorded in the phase of its wavefunction. In addition, this phase factor can be measured by interfering the system's initial and final states.[198] Sadly, Pancharatnam's anticipation of the geometric phase factor did not receive the attention it so justly warranted, and his paper languished almost unnoticed for nearly four decades, when it was finally rediscovered. This was an excruciating scientific tragedy to Pancharatnam, comparable to that which befell Waterston {472}.

The effect of the field-free potential and the geometric phase issue were pointed out strongly by Aharonov and Bohm {473} in 1959 in a leading U.S. physics journal, causing much consternation in physics. Before that paper, most physicists still believed that all electromagnetic forces and effects were induced only by the force fields,[199] and not by the potentials.

---

[199] Indeed, the force fields do not even exist in mass-free space, but only in matter. Force is defined as $F \equiv \partial/\partial t\ (mv)$ so that mass is a *component* of force. Feynman states: *"...in dealing with force the tacit assumption is always made that the force is equal to zero unless some physical body is present... One of the most important characteristics of force is that it has a material origin... " "...the existence of the positive charge, in some sense, distorts, or creates a "condition" in space, so that when we put the negative charge in, it feels a force. This potentiality for producing a*

In the prevailing belief, when the fields were zero there were no EM effects. Indeed, the potentials were largely regarded as mathematical conveniences.

Another point of consternation was that the A-potential is not uniquely determined, but is determined only up to a particular gauge transformation. So there can be two (or even more) magnetic vector potentials A and A' corresponding to a specific magnetic field. Aharonov and Bohm raised a disturbing question: How can two different "causal" A potentials produce the same effect? We will rigorously answer this question shortly.

While the experiments suggested were difficult, they were performed shortly after the publication of the Aharonov-Bohm paper. The very next year, in 1960, Chambers {474} experimentally demonstrated the Aharonov-Bohm effect, though not with complete decisiveness. A 1983 paper by Tonomura *et al.* {475} was significant, and a 1986 publication of further work of Tonomura *et al.* was decisive {476}.

In 1984, Berry generalized the Aharonov-Bohm effect into the geometric phase {477}. In 1985 the Tonomura effort plus experimental work by Webb *et al.* {478} settled the matter, as strongly stated by Schwarzschild the next year {479}. In 1987, Aharonov and Anandan {480} further generalized Berry's geometrical phase, freeing it from its adiabatic

force is called an electric field. " So Feynman was fully aware that not the force field but only the potential (i.e., the potentiality) for producing the force field exists in mass-free space. The electrodynamicists continue to be confused on that point. E.g., on p. 28 of his *Classical Electrodynamics,* 2nd edition, Wylie, 1975 Jackson states: *"...the thing that eventually gets measured is a force... ",* confirming that force is an observable and thus a *d/dt* operation imposed upon spacetime LLLT to give a purely spatial entity LLL. On p. 249, Jackson also states that *"Most classical electrodynamicists continue to adhere to the notion that the EM force field exists as such in the vacuum, but do admit that physically measurable quantities such as force somehow involve the product of charge and field."* The two parts of the conventional assumption, as stated by Jackson, contradict each other. We point out that Jackson's book has been a standard of excellence for years, and no EM library is complete without it. It merely shows the extraordinary difficulty of rooting out the exasperating and mind-numbing confusion in physics between the observed effect and the unobserved cause. Force is an effect, after interaction (observation or measurement), hence cannot be the cause (which exists prior to interaction or measurement). Else, there is no difference between cause and effect, which destroys causality itself. This confusion is still there in even the hoary old classical mechanics, now some centuries old. By that standard, electrodynamicists have a long time remaining before they will have propagated this non sequitur" as long as the mechanical physicists have.

restriction and generalizing the geometry. Much discussion of the AB effect and Berry's geometric phase occurred {481a-481c}, but the scientific consensus was in: *The Aharonov-Bohm effect is real and the geometric phase is a widespread fact of nature.*

In 1985, Olariu and Popescu {482} published a very good overview and discussion of the AB effect up to that time, with hundreds of references.

Thousands of papers on the AB effect, the Berry phase, and the geometric phase have since appeared in the literature, and continue to appear at a rather steady and prolific rate. A few are listed for a sampling {483a-483k, 484a-484i}. The original Aharonov-Bohm effect has now expanded and spread widely through physics and into a myriad other effects and areas, such as the scalar AB effect, the Aharonov-Casher effect, magnetic vector potential vortices, quantum tunneling, chemical reactions, chaotic oscillation, and others. These related extensions are far beyond our interest here. We are, however, interested in what we hypothesize may be a new aspect of the perturbed curl-free magnetic vector potential A. We shall address that aspect briefly.

7.4 A Possible New Aspect of the Curl-Free Magnetic Vector Potential

We propose what may possibly be a unique new aspect of the curl-free A-potential — and thus of the Aharonov-Bohm effect — that has been uncovered in our MEG experiments. We caution, however, that this is a tentative interpretation and remains to be further validated.

First we state that the equation, $\mathbf{B} = \nabla\times\mathbf{A}$, usually accepted as the "definition" of the magnetic vector potential A, is not even a definition, because no equation is a definition.[200] If we rewrite the equation as the identity $\mathbf{B} \equiv \nabla\times\mathbf{A}$, then it is a proper definition — though not of the A-potential, but of the B-field itself, *given* that the A-potential exists and is primary. It also states that the B-field is simply a magnetic vector potential having swirl; i.e., the notion of the B-field captures that swirl aspect of a swirling A-potential. Hence the statement that B = zero implies only that **the** A-potential remaining is swirl-free and hence a linear EM energy current flow. Note that the B = 0 condition does not specify the magnitude of the remaining A-potential, but only specifies that it has no curl. This is a rigorous answer to the original question posed by Aharonov and Bohm of how two different A-potentials can have the same magnetic field B (the same swirl). It is no more mysterious than two different rivers having a whirlpool of exactly the same swirl magnitude.

---

[200] No equation is a definition, *a priori.* It merely states the relationship of the total magnitude of the entities on the left to the total magnitude of the entities on the right. It does not define any entity on either the left or the right. One of the gross non sqeiturs in physics texts is to continue to present equations as "definitions".

At least to first order, it appears we have arrived at a "semi-definition" of the A-potential along these lines:

> *The magnetic vector potential A is an EM energy current in 4-space, which can be swirling or linear, or can have two components where one is swirling and the other is linear. The swirling component of the A-potential is also known as the B-field, and the linear component is known simply as the "curl-free A-potential". The A-potential exhibits a remarkable property: If its swirl component is localized and confined, the remaining space where the swirling A without localization would have been is filled by a remaining unswirling linear A-potential energy current. In short, by swirl-localization the swirling A-potential can be freely 'regauged' into two separated entities, with the sum of their energy being greater than the apparent energy of the previous swirling A-potential. Hence swirl-localization is a method of asymmetrical self-regauging of the A-potential itself with the excess energy coming directly from the active vacuum.*

If we take the view that any curled A is identically a B-field, then the "natural" A is not curled, because it requires a curl operation applied to it to produce a "curled A" or B-field. Therefore the "natural" magnetic vector potential A is simply a magnetic energy current in space, since it is a vector and has direction {458}. That is a *longitudinal EM energy flow process?*[201] We shall be very interested in the uncurled A-potential; i.e., in that linear energy current in space that is uncovered and freely produced when the B-field is localized. Indeed, we shall perturb it to make a large E-field, which we intercept and use to furnish emf to the secondary circuit and power loads.

---

[201] Interestingly, around a toroidal localization zone, the uncurled A-potential does curve, but is only linearly following the curving current and the moving E-field from the moving charges in the current. In the MEG, the uncurled A-potential is even more complex, since we use each half of the toroid-like localization zone differently, with the B-flux going one direction (say, counterclockwise) in the left branch and the B-flux going in the other direction (say, clockwise) in the right branch, and with a crossover point in the middle.

Given a localized B-field so that we have an uncurled A-potential outside the localization region: If we now increase the magnitude of the A, that produces an increased linear energy flow rate outside the localization zone, in addition to increasing the B-field in the localization zone. If we decrease the magnitude of the A, that produces a decreased nonlocalized energy flow rate and a decreased localized B-field. In our view it follows that, when we oscillate the magnitude of the A-potential by any means whatsoever, we oscillate the nonlocalized linear EM energy current flow rate as well as the localized B-field. In fact, that creates a normal oscillation of the localized B-field, and also creates a nonlocalized *longitudinal wave* of electric field E by the equation $\mathbf{E} = -\partial\mathbf{A}/\partial t$, if the conditions localizing the curl operator — and hence localizing and holding separate the concomitant B-field produced by a changing E-field — continue in effect.

We have produced a mechanism for generating a nonlocalized longitudinal E-field wave without a nonlocalized B-field component being allowed to act outside a confined locality. We stress that a B-field component is produced, but remains localized in the special core material (that is one of the conditions required in the MEG).

We have also produced a mechanism for generating a nonlocalized longitudinal E-field wave with another very novel feature: the magnitude of the E-field and thus the magnitude of its energy density (its intensity) does not depend on the *magnitude of the input energy* used to perturb the A-potential, but only depends on how sharply we perturb it. Thus we have produced a mechanism for energy intensity amplification and magnification, freely. This is possible because the system is an open system far from equilibrium with the active vacuum, and specifically with the curl-free A-potential as an energy current at large in the local vacuum where it can be intercepted, collected, and used to power loads. The energy intercepted from the uncurled A-potential by the output section(s), is *additional* energy in excess of what is intercepted from the E-fields generated in core material B-field magnetic flux changes.

However, see our discussions of the giant negentropy process, earlier in this book. Any longitudinal EM "wave in 3-space" is accompanied by a longitudinal EM wave on the fourth Minkowski axis, in the time domain. EM energy does not observably propagate continuously in 3-space, since observed propagation is a series of frozen iterative observation processes a priori). The "energy flow" actually consists of a propagation in 4-space of a *continual local circulation* of EM energy from the time domain (time-polarized EM energy) to each interacting point dipole in 3-space via the negative charge of the point dipole, thence to the positive charge of that point dipole, thence back to the time domain. In a time-reversed zone, the circulation is in the opposite manner. So what spreads as the "longitudinal EM energy current in 3-space" is actually the spreading point polarization of the vacuum and this giant negentropy 4-circulation involved in every point dipole of that polarization.

We believe this previously unsuspected process is involved in the Aharonov-Bohm effect (which is a sort of polarization in the time domain), other aspects of the geometric phase, fundamental polarization of the vacuum itself, and the very notion of "propagation of EM energy through space". At any rate, the hypothesis does fit the phenomena observed to happen in the MEG, so it warrants additional careful experimental investigation as well as attention from leading theorists.

Further, the MEG is practical and produces practical electrical power, but with a *purely electrical reaction* in the output secondary coils of the transformer section.[202] Hence there is finally a very practical use for the geometric phase: It extracts additional *nonlocalized* EM energy from the vacuum (and local curved spacetime) in addition to extracting and retaining *localized* energy. Both the extra "free" nonlocalized EM energy and the localized EM energy from the vacuum can be intercepted, collected, and utilized to power practical fuel-free MEG-type power systems.

We point out that the rate of changes of all E-fields produced in the secondary and primary coils also produce B-fields proportional to the time rate of change of those E-fields in each case. These secondary B-fields are retained in the core rather than reaching out into surrounding space. Their rate of change then further produces E-fields, which pass out of the core and interact with the coils. Hence there are very novel "multiple feedforward and feedback" regenerative loops in the MEG's highly nonlinear operation, as well as chaotic oscillations.

---

[202] This follows from the localization of all B-fields in the core material — including the B-field of the permanent magnet, those B-fields produced by dE/dt operations, and those B-fields produced inside the secondary coils, etc. All E-fields produced, however, including in the changing flux of the core and in the $d\mathbf{A}/dt$ operations of the external uncurled A-potential — are free from that localization restriction and can pass out of the core and interact with the output coils of the transformer section. In short, we have produced the world's first purely electric "induction-free" magnetic transformer device. The voltage and current, measured directly across the secondary coil's ends, is in phase within 2 degrees and even that 2 degrees can be eliminated. We expect the theorists to have a field day in modeling these new phenomena.

7.5 Use of the Aharonov-Bohm Effect in the MEG

From the standpoint of electrical power systems, our own interest in the AB effect focused upon one important characteristic: the clear separation (and localization) of the B-field (the swirl component of the A-potential), leaving the remaining "natural" uncurled field-free A-potential as a free linear flow of energy current in the vacuum. This operation is iterative, and is produced by the core for every B-field produced in the transformer section. So a localized dynamic B-field makes nonlocalized dynamic E-fields, which themselves make further localized dynamic B-fields, which then further make nonlocalized dynamic E-fields, and so on. The result is a very extensive chain of feedforward and feedback field energy loops.

7.5.1 Separation of the Swirled and Linear A-Potentials

The AB effect has long been known in toroidal coils and in long solenoids. In a theoretically perfect toroid, for example, the AB effect separates and localizes the entire B-field inside the toroid, while the *curl-free* magnetic vector potential A appears in surrounding space outside the toroid. By our new interpretation and hypothesis, based on AIAS theoretical work published in the literature, this curl-free A-potential represents a nonlocalized longitudinal EM energy current in space.[203,204] Further, the curl-free A-potential is an extra flow of EM energy that can be intercepted, collected, and used while also still using the entire "normal" B-field energy localized in the core. The neat thing is that the E-field made from the

---

[203] This is addressed in M. W. Evans, T. E. Bearden, and A. Labounsky, "The Most General Form of the Vector Potential in Electrodynamics," *Found. Phys. Lett.,* 15(3), June 2002, p. 245-261. See also M. W. Evans et al., "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," *Found. Phys. Lett,* (in review).

[204] See also B. Lehnert and J. Scheffel, "On the Minimum Elementary Charge of an Extended Electromagnetic Theory," *Physica Scripta,* vol. 65, 2002, p. 200. In extended EM theory, the elementary charge may no longer be considered as a fundamental constant of nature, but is expressed in terms of the Planck constant and the velocity of light.

A-potential, and thus providing the actual intercepted and collected EM energy, can be made as large as desired, merely by manipulating the rise time and decay time of the pulses used to perturb the A-potential external to the core. Further, the same perturbation also perturbs the B-field flux in the core, giving rise to nonlocalized E-fields whose magnitude again depends upon the rise time and decay time of the perturbing pulses. Hence the dA/dt and the dB/dt become energy amplification causes, where all the excess F-field energy is furnished freely from the vacuum via a special kind of asymmetrical regauging and gauge freedom. In that sense, the MEG produces and applies a hitherto unrecognized form of direct field energy amplification by asymmetrical regauging.

We noted that this mechanism represents the formation of a locally curved spacetime[205] and engineering of the magnitude of the curvature itself. By freely increasing the local spacetime curvature, the appearance of additional EM energy density in the region follows *a priori*. That is, one still has all the B-field energy produced by the permanent magnet, but it is just confined to the space inside the core. All the energy one "pays for" or inputs with current and voltage to the input coil, in the normal sense, is the energy to *perturb* the localized B-field and the curl-free A-potential. However, even before perturbation to evoke energy amplification, there is already additional energy current that has appeared outside the core, where the magnetic field B would have normally extended. All the A-potential energy[206] is "extra" energy in addition to the normal B-field energy that would otherwise have appeared out there but is now confined inside the core.

---

[205] More accurately there is one specific curvature of spacetime in the core of the MEG, representing the confined magnetic field B (the localized swirling A-potential), and there is another different curvature of spacetime outside the localization zone, filled with the nonlocalized curl-free (non-swirling) A-potential as a nonlocalized linear energy flow in space. Further, in the MEG we freely change and increase the magnitude of those spacetime curvatures, by the *rate of change* of the input energy rather than by its *magnitude*, and hence we easily increase the magnitude of the energy density of local space both inside and outside the core. That is the fundamental principle of the MEG"s energy amplification (free regauging).

[206] Some scientists even deny that the curl-free magnetic vector potential has any energy! We answer that it produces energetic effects upon charged particles, hence must either contain energy itself or direct and gate the interaction of some local vacuum energy with those charges. In the MEG, the uncurled A-potential is oscillated, which does not appear to yield a *transverse* EM E-field wave at all. Instead, it appears to yield a *longitudinal* E-field wave with an associated scalar KM wave in the time domain. We also explained what appears to happen to the concomitant B-field wave produced by any changing E-field in the secondary

Further, the uncurled A-potential outside the core reduces its magnitude only inversely as the radial distance, whereas the normal B-field that would otherwise be there in the absence of localization would have reduced its magnitude inversely as the square of the radial distance. This far less rapid fall-off of the magnitude of the nonlocalized curl-free A-potential than the former nonlocalized magnetic field B can result in a remarkable increase in communications range — when a normal EM field antenna with oscillating EM field transmissions is compared to a curl-free A-potential antenna radiating the same power but as curl-free A-potential oscillating transmissions. Golden[207] has applied that effect for long-range communications.

With the application *of the free* Aharonov-Bohm effect produced by the core material and permanent magnet combination, rigorously there has appeared much more available EM energy from the local vacuum than the energy one inputs to the device as one's input energy "payment". Again, this is because the system is an open system far from equilibrium with the active vacuum processes such as the appearance of the uncurled A-potential and the iterative appearance of powerful extra B-fields and E-fields. The rise time and decay time of the input perturbation pulses produces an energy amplification effect both in the core and in space outside the core. When one utilizes this extra uncurled A-potential energy, as well as the now-confined B-field energy, as well as the extra B-fields and E-fields appearing from the iterative processes, then one can extract and use more energy than one himself "paid for" or input to perform the perturbation. Indeed, one can easily extract and use more energy from the curl-free A-potential alone, than one has input. By simply controlling the rise time and decay time of the input pulses, the magnitudes of the resulting E-fields formed in the uncurled A-potential (and in the perturbed trapped B-field perturbations inside the core) can be controlled at will.

(output) coils of the transformer section, etc. The core material withdraws and holds localized, all such magnetic fields formed in the internal MEG core processes.

[207] In the late 1970s Frank Golden demonstrated such an uncurled A-potential a**ntenna** which attained a range of 20 miles for a tiny CB radio, compared to a range of only 1 mile for the same CB radio with a normal EM field antenna. The energy density of the potential drops off inversely as the distance, while the energy density of the field drops off inversely as the *square* of the distance.

We also point out that all these amplified E-fields produced by multiple actions have both a conventional Poynting energy flow component and an unaccounted, much larger Heaviside energy flow component. Thus the total energy amplification factor can be very large (several orders of magnitude). With efficient interception and collection, in the future we expect to be able to produce at least 100 kilowatts of power out of the same size core and buildup now producing 80 watts or less in a highly nonoptimized set of laboratory experiment MEGs.

We again call the reader's attention to our previous discussions in this book, showing that all EM energy in EM circuits comes from the vacuum anyway (particle physics view) or from the time-axis (relativistic view). This has previously been hidden by the failure of classical electrodynamics to solve the source charge problem, which we also solved in this book after our paper {12} in 2000. Circuit theorists avoided it by indirectly assuming the source charges freely created from nothing all the EM energy in their fields and potentials. The use of the assumed fields and potentials and their Poynting energy was calculated, but never was the source of those fields, potentials, and their energy presented or accounted.

7.5.2 Using the Uncurled A-potential as a Linear Energy Current

A very simple equation allows utilizing the extra nonlocalized A-potential energy, even though it has no curl and therefore no magnetic B-field. That is $\partial \boldsymbol{A}/\partial t = -\boldsymbol{E}$. This equation means that, if one perturbs that large field-free A-potential outside the toroid, it produces an E-field, which we have argued must now be a longitudinal EM wave due to the perturbations *ifthe B-confining operation is still imposed.* By oscillating the A-potential perturbations, one produces an oscillating E-field that is also an oscillating longitudinal electric E-field wave without an accompanying B-field wave.

The oscillating B-fields in the core also produce E-fields proportional to $\partial \boldsymbol{B}/\partial t = -\boldsymbol{E}$. These extra E-fields are in phase with the E-fields produced by the perturbed external A-potentials, considering the wavelength of the frequencies at which the MEG is operating (30 to 80 kHz nominally).

Interestingly, the time-oscillating rate of change of the E-field in this longitudinal E-field wave does produce a magnetic B-field, but it is confined to only the core material and not in the nonlocalized external spatial region *because ofthe localization capability ofthe core material.* The core simply changes its permeability and holds the additional B-fields and their flux in there, thus forming additional uncurled A-potential outside the core. The E-fields produced by variations of the core B-field flux is not confined, but freely passes out of the core material

If that B-field (curled A) is sharply localized by an ongoing localization capability, the longitudinal E-wave has its produced magnetic B-wave component stripped from it and localized. In that case, the process produces a purely longitudinal E-field wave outside the localization region — with the possible exception that a longitudinal B(3) field wave component pioneered by Evans {485} *may be produced as well. That is left as an open question presently unresolved in our MEG experiments. We tentatively believe that the B(3) field wave component is in fact produced, but also forcibly localized in the core along with the magnetic field B. If so, its perturbations will also contribute to producing extra E-fields.*

See Figure 7-5. In the MEG, the localization is accomplished as a unique function performed freely by the special nanocrystalline layered core material. So contrary to the case of the solenoid or toroid, one does not to have to "pay" any EM energy to the localizing component to obtain the localization function itself.

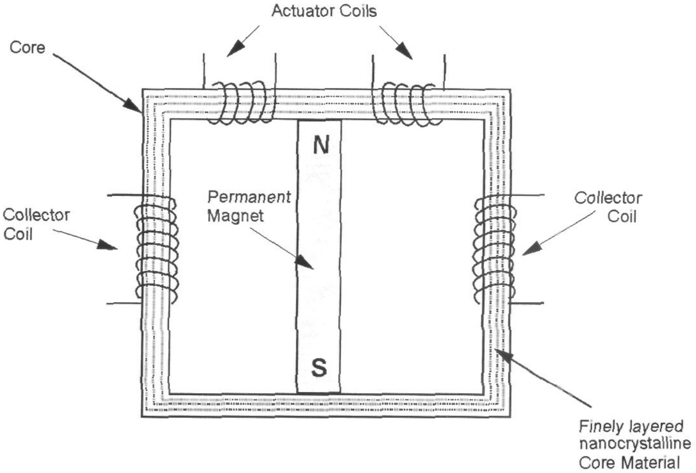

Figure 7-5 Diagrammatic drawing of the MEG.

Further, if one deliberately uses a nearly rectangular perturbing pulse that has very sharp rise time and very sharp decay time, the nonlocalized E-fields, resulting localized secondary B-fields, and resulting nonlocalized secondary E-fields produced can be made very large, with rapid rates of change also. In other words, the magnitudes of the nonlocalized E-fields and localized B-fields produced can be adjusted at will, including in their higher orders. The magnitudes of these evoked fields do not depend on the *magnitude* of the input perturbation energy, but on its *rate of change*. This is very important, because it means the MEG permissibly achieves energy magnitude increase in its nonlocalized output E-field waves by simple instantaneous power deliberately prepared in its low-energy input perturbation waves.[208] In short, the motionless electromagnetic generator uses perturbation of the Aharonov Bohm effect as an energy amplifying mechanism.

This also affects all the iterative processes, hence all the other E-field waves produced outside the core material.

With the localization process remaining and ongoing (simply because the core material remains), the B-field continues to be stripped away, leaving the perturbed A-potential to produce a longitudinal E-field wave (ignoring the possible presence of a B(3) field also). In the MEG transformer section, the output (secondary) coils are wound around the nanocrystalline core that performs a B-field and B-flux localization within the core itself. Hence a purely net E-field interaction occurs with the electrons in the secondary coils, and the output of these coils is an output having the output voltage in phase with the output current.[209] The AC current in the secondary coils does in fact still produce magnetic field B-flux and B-waves in them, but the core material draws the extra B-flux and B-waves inside and localizes them within the core. The fact that localization is *not quite* 100% at the imperfect junctions of the ends of the permanent magnet and the core material, gives a tiny bit of leakage of

---

[208] It also means that the MEG operation is highly nonlinear, with very extensive feedforward and feedback field transduction loops. For the MEG, its scale-up, mathematical modeling, and engineering simulation are not simple things at all, but highly complex. In addition, control and stability are more complex, since one must utilize nonlinear oscillation theory (including chaotic oscillation theory) as well as nonlinear oscillation control theory. In addition, a higher group symmetry electrodynamics is required for the modeling and simulation. Consequently, because of these considerations and the geometric phase aspects, as well as the additional close-looping considerations for self-powering operation, an expensive physics lab must be established as well as the usual electrodynamics laboratory aspects. The result is an expensive final research and development program required to complete the MEG to final production prototype units, ready to start rolling off the assembly lines. A minimum capitalization of $20 million is required.

[209] In the actual lab experiment MEG, there is a slightly imperfect junction between the ends of the permanent magnet and the core material. This results in a very tiny bit of leakage of nonlocal B-field —just enough to produce a 2° phase shift between the voltage and current in the secondary coil's output. With a better junction, this phase shift can be further reduced or eliminated.

B-field, so a tiny B-field component does escape, interact, and generate the *2%* or so variation from the complete in-phase condition of the secondary output voltage and current. Again, that can be further reduced or eliminated by improving the junction.

So essentially there is a purely electric field interaction with the electrons in the output coil current of the MEG, hence the result that the output current is in phase with the output voltage — something previously unheard of in a simple secondary coil of a transformer. This is a major practical macroscopic application of the Aharonov-Bohm effect and therefore of the geometric phase.

7.6 Collection of Field Energy in the MEG

To collect and use EM energy in a power generating system, first the EM energy must be evoked from local spacetime (the local vacuum) by the broken symmetry of the source charges and dipoles in the circuit {12}. In the open system far from equilibrium in the resulting EM energy flow, the operator must usually pay for (i) the energy required to *initially* produce the requisite source dipolarity and (ii) the energy required to perform necessary switching and control functions, unless close-looping is incorporated. Using the permanent magnet as a source dipole in the MEG, and using the special nanocrystalline core material to generate the Aharonov-Bohm effect to furnish excess energy from the vacuum, the MEG operator does not have to *furnish* the evoked and available field energy at all. So the first "payment" is totally avoided. Further, the magnetic dipole is not destroyed by passing magnetic flux back through its back mmf. No additional input energy is required to continually restore the source dipole, since it is never destroyed.

In an open-loop MEG system, the operator only has to pay for the second part: the switching and control function energy. In later closed-loop MEG systems, a method of controlled and positive feedback has been developed so that, once the system is up and operating and powering its load, the necessary switching and control function energy can be siphoned off the output and fed into the input in governed fashion, first being transduced into positive energy input form. Once the process is properly synchronized and governing properly, the operator can throw a switch for feedback and transduction, and the unit will be "self-powering", or "fueled entirely by **the** local vacuum energy", so to speak.

Let us consider the open-loop MEG system we have presently attained in a series of embodiments.

The total emf force F collected from a given nonzero electric field E, e.g., depends only upon the amount of available intercepting charge, by the simple equation $F = Eq$. Similarly, from any potential ϕ the amount of collected EM potential energy W depends on the amount of intercepting charge q, by $W = \phi q$. Hence in the MEG a legitimate COP>1.0 system is possible if (i) more energy is collected and used from the evoked EM energy flow than is input to gate that energy flow's appearance from the active environment, and (ii) none — or appreciably less than half— of the energy collected in the circuit is used to destroy the source dipolarity.

By invoking the AB effect in a local region, the vacuum contributes additional energy density in the space outside that region, and that energy density can be sharply perturbed so that substantial field energy amplification appears by asymmetrical regauging. The excess energy to provide COP>1.0 performance is freely furnished from the active vacuum (and from the local curved spacetime) {38b, 38c}. The excess energy can be intercepted in the perturbed curl-free A-potential region outside the transformer section, collected, and used to power additional loads, in addition to being intercepted in the transformer secondary region to power the secondary circuit and main load.

Similar contributions of extra energy and extra energy collection outside the core localization region occur for the iterative processes also.

At the same time, the B-field inside the core material is perturbed, but the core material self-varies its permeability to adjust to the additional magnitude of the perturbed B-field. The production of any B-field oscillation inside the core does produce an electric field E, which is not localized in the core. Consequently, energy from the perturbed core-localized B-fields is transduced into nonlocal E-field energy, passing out of the core and interacting with the output secondary coils of the transformer section in purely electric field fashion.

Because of the purely electrical interaction, the voltage and current output from the secondary coils of the transformer section are in phase, and not 90° out of phase as in a transformer. The actual 2° experimental phase shift of current with respect to voltage is a measure of the slight inefficiency of the junctions of the permanent magnet ends with the B-localizing core material.

In the MEG, we are using a novel new kind of component: an *external-induction-free cored inductor*, which might alternately be called *a purely electrical inductor — at least externally.* In short, the inductive action remains localized in the core, while the electrical action escapes the core and interacts with the coil to drive the current electrically. So here is another excellent, novel, and practical use of the Aharonov-Bohm effect and the geometric phase: Eliminate the inductive reaction of a coil's output, and transduce it into a purely electrical action. At this writing, we are preparing to file a separate patent application upon this latest application and process, and it will already be filed by the time this book is published.

We accent that any field or potential, once made, already involves an ongoing giant negentropy and overunity process {486}, as we previously discussed in this book. Simply calculate how much energy it requires to separate and form the source charge — which can be as simple as charging a tiny little sphere — and recognize that the field formed by that charge is an altered energy density of surrounding space and a polarization of the vacuum. Once the charge is formed, the corresponding alteration and sustaining of spatial energy density change and vacuum polarization freely moves radially outward at the speed of light from the charge, from the very moment it is produced. So a steady flow of observable EM energy is speeding out in all directions from the source charge. An enormous change in the energy density over that external spatial region surrounding the source charge occurs and spreads steadily, moving radially outward in all directions at speed c.

Hence a magnificent and rapidly extending curvature of spacetime — which energetically acts upon any available charged matter therein, by conventional general relativity — is formed and spreads outward in all directions from any source charge, once formed. How much effect on charged matter is exhibited by this freely created and spreading spatial EM energy density change, depends only on the amount of interacting/receiving charged matter in that surrounding space that is available to interact with the energy of the new fields and new potentials.

If **the** source charge or source dipolarity (in this case, primarily the magnetic dipole of the permanent magnet) is left alone and not destroyed, an unlimited amount of energy can be intercepted and discharged from a local load, again and again, from the charge's surrounding field or potential because the energy in the field and potential is steadily replenished from the active vacuum. The only problem is discharge the collected energy in the load(s) so as to leave the original source charges (the magnetic poles) inviolate, and not to destroy them. The MEG accomplishes this function.

This also rigorously follows from the gauge freedom axiom of quantum field theory. Consider this: the ability to freely change the potential — assumed by all electrodynamicists — can be applied directly to the load to potentialized it without any current moving, if there exists a somewhat lengthy electron gas relaxation time in that part of the load that is potentialized. After potentialization, one can switch the source of free potential away, and complete the load circuit separately, so that the excess potential energy is dissipated from the load, powering it. In that case there is no dissipation of the primary source dipole. *Gauge freedom already assumes one can produce EM energy from the local vacuum freely, anywhere and at any time.* In real life, one has to pay a little for the switching, but not for the energy input itself.

Why the scientific community for decades has fiercely opposed the funded experimental investigation of the implications of gauge freedom is quite inexplicable.

As can be seen from the conservation of energy law alone, in theory energy can be continually intercepted and collected from these free fields and potentials, once the source dipolarity (or source charge) creation has been paid for. Further, by applying gauge freedom to potentialized the load without energy dissipation, with a little adroit switching the collected energy can then be dissipated entirely in a separated external load circuit, so that the free potential energy completely exits that circuit without re-interacting with the source dipole and its back emf or mmf.

We predict that the above application of gauge freedom will eventually power most of the world's electrical systems, when the scientific community gives up its present "ostrich position". Once the engineers can build it, the scientists do not have to agree.

That such is not done in our conventional circuits is purely due to the limited vision of the scientific community and its refusal to fund appropriate research. To date, it has not even recognized what actually powers every electrical circuit, or that a common practice (the closed current loop circuit) has eliminated COP>1.0 and self-powering Maxwellian power systems for more than a century. The entire electrical energy crisis hurtling at the nations of the world is due entirely to the shortsightedness of the scientific community, particularly since 1957.

Once the energy is pouring (transmitted) from the dipole and out of the terminals of the generator to fill space around the external conductors, then a straightforward energy transmission and reception problem exists. The source dipolarity produces the steady emission and flow of energy filling space around the circuit. In the "receiving external circuit" the surface charges and their fields must intercept some of that available EM energy flow and diverge it into the conductors to power the circuit electrons. The intercepted and diverged component entering the circuit is the Poynting component, while the nonintercepted and nondiverged component missing the circuit and usually wasted, is the extra Heaviside component.

Fortunately, every source dipole pours out enormously more EM energy than our feeble conventional circuits intercept and catch to provide the Poynting calculations. So to produce COP>1.0 in the MEG, all we have to do is insure that we have (i) a mechanism easing the interception and collection process, and (ii) sufficient "antennas" containing sufficient conduction electrons (and their fields) to intercept and collect more energy than we pay to make the source dipole and perturb it. Since we pay nothing at all to make the source dipole (which is the permanent magnet), and since the magnetic flux back through the permanent magnet does not destroy the dipole, then all we have to pay for is the perturbation and for switching and control.

The resulting field energy generated outside the core localization region is primarily a function of the rise and decay times for the input perturbation pulses and not the energy of those pulses. From a little input energy with fast little changes, we can easily generate very powerful oscillating E-fields and provide dramatically increased "antenna collection" area with **the** necessary magnitude of intercepting charge. It rigorously follows that COP>1.0 is easily obtainable.[210, 211]

[210] Again, for the MEG, changing and controlling the rise time and decay time of the input pulses is a method for directly engineering and controlling the general relativistic curvature of the local surrounding spacetime, because it controls the magnitude of the E-field that is created by $\mathbf{E} = -\partial\mathbf{A}/\partial t$ from the curl-free A-potential outside the core. This also directly engineers and controls the magnitude of the polarization of the local active vacuum. That simple fact allows the use of a unified field theory operation, with a net energy interchange between the system and **the** local active vacuum. To understand the MEG, one must think in terms of the *supersystem.*

[211] All EM energy at any point in space must have been "gated" from the time-domain and it must return to the time domain (as, e.g., by the source charge for it). It **requires** quite a shift of one's mind to realize that all 3-space energy (such as "energy at each point in a potential") involves a gating process and a giant negentropic 4-circulation of EM energy ongoing at each point dipole. Much of the extraction of **energy** from the vacuum has been hidden by thinking in terms of "inputting energy" instead of "inputting or arranging energy flow gating" by the source charges and sourcedipoles.

The MEG is an open system far from equilibrium with its active environment (its active local vacuum and its active local curvatures of spacetime). By attending to the interception and collection part of the overunity problem, the MEG can permissibly collect and use more EM energy to power its loads than one oneself has to input to it. In short, one only has to input a little switching and perturbation energy with properly shaped waveforms, to generate very powerful fields and appreciable field energy available for collection and use. With sufficient collection mechanisms in place, the unit easily produces COP>1.0. Then with proper closed-loop principles implemented, the MEG can be self-powering.

7.7 Experimental Path to Development of the MEG

In our experiments with the Radus effect previously discussed in Chapter 6, paragraph 6.4.5, we had already done extensive work with switching magnetic fluxes for less switching energy than can be intercepted from the switched flux. We had also experimented with various cores for transformer and coil variants, including eventually the new layered nanocrystalline cores.

In our experiments, we noticed a remarkable thing. With certain modern nanocrystalline core materials, and using a permanent magnet to provide B-flux, we found that *the core material would extract and hold essentially all the permanent magnet's B-flux inside the core.* Field measurements right against the outside of the poles of the powerful little neodymium-iron-boron magnet itself showed very little B-flux spilling out into space! In short, we had discovered how to make a magnetic system that *nonlocally* propagated only the uncurled magnetic vector potential A, without any operator energy input. We realized we had discovered a marvelous and free way to freely obtain an excellent Aharonov-Bohm effect, similar to the way a nearly perfect toroid would localize the B-field inside the core, with the uncurled A-potential appearing in space surrounding the toroid. The advantage of our approach over the toroid was that we did not have to furnish any current and voltage to obtain the localized B-flux and the non-localized curl-free A-potential energy.

Further, if we perturbed the B-field flux in the core, we consequently perturbed the uncurled A-potential in space surrounding the core. From both sources — the perturbed B-flux in the core and the perturbed

A-potential outside the core — we could produce very large E-fields and concomitant very large voltages capable of driving more "power" through the secondary coils of the transformer section than the power required to perform the switching and perturbation.

We could easily prove this by perturbing the localized B-flux (and the concomitant surrounding uncurled A-potential automatically) with weak energy. We proved that increasing the rate of the rise time and decay time of even a weak input EM energy perturbation could produce quite large E-fields, including 1,000 to 5,000 volts (or even 20,000 volts!) appearing in the secondary coil if desired. We also proved another thing: *the special nanocrystalline core material in some buildups self-adjusted its permeability, continuing to localize the perturbed and alternating B-fields and their fluxes while freeing the resulting E-fields and passing them out of the core.* In short, the permeability of the specially selected core material automatically adjusted with the perturbation of the localized B-flux and nonlocalized A-potential, so that all the oscillating B-flux remained localized in the core. Further, the highly layered nanocrystalline material produced essentially no eddy currents at all, hence no heating, even in the presence of dramatic B-flux oscillation and dramatic E-field intensities in the surrounding uncurled A-potential space and in the output (secondary) coils of the transformer section of the MEG.

The absolute proof that the core changed its permeability was in the lack of a phase lag between voltage and current in the output from the output coils. Here a remarkable thing was observed: the output current was in phase with the output voltage, within about 2°. In short, under certain circumstances the output coils wound around the nanocrystalline core would exhibit essentially a purely E-field interaction, with only minuscule B-field leakage — something deemed impossible in the typical electrodynamics textbook. This meant that the excess E-fields produced by the perturbed A-potential in space outside the core also reacted with the output coil's conduction electrons in a purely electrical manner! This was because the B-fields locally produced in the output coils by the interacting changing E-fields, *were in fact instantly drawn into the core material as they were produced, and were held there so they did not react in magnetic fashion with the electrons in the output current from the secondary coils.*

So this is a dramatic localization of Faraday's law of induction. Consequently, one entire group of embodiments of the MEG uses a purely electrical interaction with the secondary (output) coils of the transformer section.[212] Or more exactly, the core material draws in all the magnetic fields from the reactions of the output coils, leaving only the electrical reaction outside the core — including in the coil conductors, which obviously are wound outside the core material.

On the other hand, the perturbed localized magnetic B-flux inside the core material did produce the normal E-field to be expected, as a function of the time-rate of change of the B-field. Such an E-field is not localized in the core material, but extends outside the core and interacts with the electrons in the output secondary coils of the MEG's transformer section. Again, the magnitude of this E-field is a function of the time-rate of perturbation of the localized B-field flux in the core.

The nanocrystalline core also has excellent high frequency response, with negligible eddy currents. Thus we can "fine tune" the engineering of the rise time and decay time of the input perturbation pulses.

Hence we accomplish several kinds of "gated energy amplification" in the MEG: (i) the amplified magnitude of the external E-field interacting with the external collectors (such as the output secondary coils) due to perturbation of the nonlocalized external uncurled A-potential, (ii) the amplified magnitude of an additional E-field interacting with the external collectors (such as the output secondary coils) due to perturbation of the localized internal B-field flux inside the core material, (iii) a large number of additional derivatives of higher order B-fields and E-fields are formed and utilized, and (iv) adding additional collectors (either coils wound on the core or external "receiving antenna circuits" produces additional power for heavier loads, for a given perturbation input.

We had exhibited a cored output coil with a remarkable adaptation of the AB effect. Essentially the output of the secondary coil was driven by the large electric field component created from the uncurled A-potential perturbations *and* by the large electric field component created from the B-field flux perturbations inside the core. This peculiar transformer-like section operated from the E-field fluxes emerging or produced *outside* the transformer core material, rather than from the magnetic field flux being produced and oscillated *inside* the core material itself.

---

[212] Again we accent that the magnetic field reactions do occur, and the B-fields are in fact produced. But the resulting B-fields are also extracted into the core material and localized therein. Even the magnetic B-fields produced in the input and output coils themselves, is so localized and not allowed to interact with the electron currents through the coils to produce the conventional 90° phase lag between voltage and current.

Obviously we had found a remarkable but legitimate COP>1.0 process and embodiment — in fact, a whole set of them. We had also invented a new kind of "transformer" or energy generator-transducer.

From that we progressed to discovering ways to a dual use of both the perturbed B-flux in the core and the E-fields from the perturbed A-potential outside the core. To first order, that turned out to be fairly simple: any changing B-flux inside the core material and thus inside the output (secondary) coils of the transformer section, produces not only B-field which is held and localized, but also the changes of this local "pinned" B-field further produces E-field oscillations which are not pinned in the core. Hence the electric field half of a normal transformer action could and did occur in the secondary coils, but the magnetic field half did not occur *in the coils themselves*. The MEG used and uses a nearly pure E-field interaction in the output coils of its transformer section secondary, where that net interacting E-field comes from a variety of E-field generating mechanisms.

We filed our first patent, and then filed a second patent application soon after. Presently we are preparing to file our third and fourth patent applications on additional phenomenology and mechanisms uncovered in our experiments. As we are writing this section, our first patent has issued {37}.

In the view of the present author, we have uncovered a remarkable new application of the Aharonov-Bohm effect and the geometric phase, one that is practical and eminently useful in novel new electrical power systems.

7.8 Additional MEG Functions

Figure 7-6 diagrammatically shows the core material inside the input coil for either open-loop operation using an external power supply or closed-loop (self-powering) operation using clamped and governed positive feedback. As can be seen, the B-field and B-field flux are confined to the core material (which is in the form of a closed transformer coil path, as shown diagrammatically in Figure 7-5). The external space surrounding **the** core material is filled with the uncurled A-potential, due to the Aharonov-Bohm effect being imposed by the core material. The two leads from the input coil are shown at the bottom of Figure 7-6.

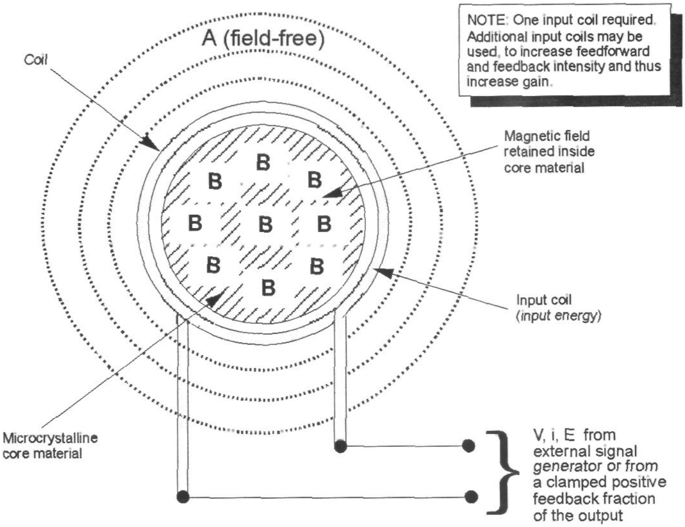
A (field-free)
Coil
NOTE: One input coil required. Additional input coils may be used, to increase feedforward and feedback intensity and thus increase gain.
Magnetic field retained inside core material
B
Input coil (input energy)
Microcrystalline core material
V, i, E from external signal generator or from a clamped positive feedback fraction of the output

Figure 7-6 Input coil for open- or closed-loop **operation.**

We pulse the input coils with nearly rectangular pulses each having a sharp leading edge and a sharp trailing edge, so that the resulting rapid time-rates of change of the two edges create very large perturbations of the B-flux in the core and of the A-potential outside the core. The large E_2-field resulting from the fast rate of change of the perturbed B-field in the core is not localized, but appears in space outside the core. In addition, an E_3 field created directly in the coil itself also appears in space outside the core. In addition, the sharply perturbed A-potential outside the core creates a third large E_1-field in space outside the core. The net E_t-field in the surrounding outside space is the vector summation of the E-fields produced by these processes, or to first order (neglecting the further differentiations)

$$\mathbf{E}_t = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 \qquad [7\text{-}1]$$

We also may adjust the pulse width to generate a further Lenz law effect, increasing the three component fields E_1, E_2, and E_3 if desired. To do that, the pulse must be narrow enough to suddenly decay while each of the fields E_1, E_2, and E_3 is still rising and not yet reached maximum. As can be appreciated, considerable pulse shaping and timing can be adjusted for optimization of the output voltage, current, and power when interception and collection from E_t is accomplished in the secondary coils. In addition, it can be appreciated that slightly differing materials delay times do occur in the production **of** $\mathbf{E}_1$**,** E_2, and E_3. Hence there is appreciable optimization that can be and is accomplished on a given MEG variant and build-up type.

More complex waveforms of various shapes can be used to further optimize the output. It is also possible to accomplish further optimization and control by adjusting the shapes and timing of the feedback signals in the case of the closed-looped MEG system. For the externally powered open-loop version, additional optimization and control can be achieved by adding additional simultaneous pulses fed into the input coils in various fashions.

In addition, the response of the output coils produces feedback into the core material, into external space, and into the input coils. These in fact provide three more E-field *return* or feedback inputs, so that the total feedback input is

$$\mathbf{E}_f = \mathbf{E}_4 + \mathbf{E}_5 + \mathbf{E}_6 \qquad [7\text{-}2]$$

Without doing the more complex iterations for this feedback and feedforward set of loops, to first order the total input E_I to the input coil may be considered to be

$$\mathbf{E}_I = \mathbf{E}_t + \mathbf{E}_f \qquad [7\text{-}3]$$

Equations [7-1], [7-2], and [7-3] are for simple analogy only; in the real MEG, full nonlinear feedforward and feedback analysis techniques must be utilized because of the iterative operations, and much more sophisticated optimization techniques are appropriate. Also, chaotic oscillations can and do appear, and techniques for control of such oscillations are indicated, particularly along the lines indicated by Fradkov and Pogromsky {487} without use of probability etc. Quoting:

> *"1. There is ... great benefit of using the modern nonlinear and adaptive control theory. ...2. There is no need to distinguish periodic and chaotic behavior. Accurate control is possible without accurate prediction. ...3. There is no need to define chaos in order to control it. ...4. There is no need to use probability in order to control systems with seemingly random behavior. "*

For any given MEG build-up design, one experiments with these optimization and control factors and mechanisms to provide the best performance, and to provide rock-steady performance.

Figure 7-7 shows the output coil for either open-loop operation from an external power supply or closed-loop (self-powering) operation using clamped and governed positive feedback. Here the situation is partially reversed. There are three major inputs to the output coils, and these are (i) the external input from the total outside E-field, (ii) the internal input from the perturbation changes in the confined magnetic flux inside the core through the output coils, and (iii) the ensuing feedforward and feedback from all E-fields.

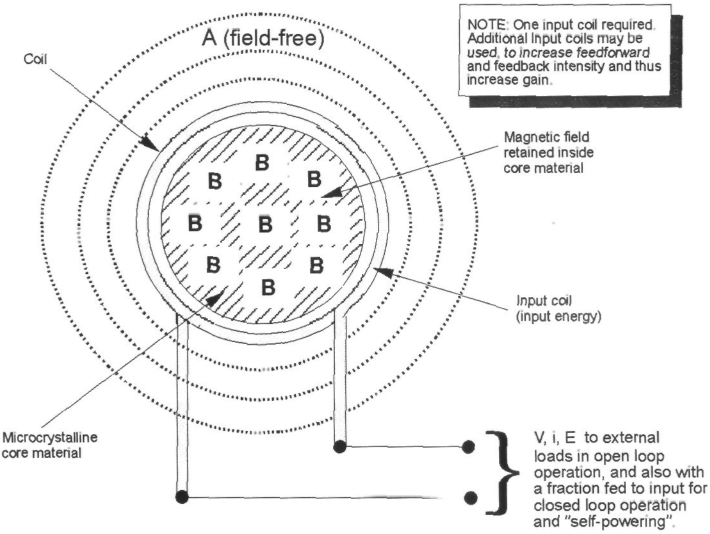
A (field-free)
Coil
Magnetic field retained inside core material
B
B
B
B
B
B
B
B
B
Input coil (input energy)
Microcrystalline core material

NOTE: One input coil required. Additional input coils may be used, to increase feedforward and feedback intensity and thus increase gain.

V, i, E to external loads in open loop operation, and also with a fraction fed to input for closed loop operation and "self-powering".

Figure 7-7 Output coil for open- or closed-loop operation.

Again, this is simply an easy way to visualize the situation to first order. In the real case, sophisticated feedforward and feedback analysis is required and utilized. Optimization usually consists of two steps: (1) the simpler, more direct optimization techniques such as adjusting pulse width, pulse rise time and decay time, actual pulse form, etc., and (2) the deeper and more complex optimization arrived at from detailed feedback and feedforward loop analysis as well as nonlinear oscillation and nonlinear control theory analysis.

In the actual optimization analysis, the feedback and feedforward loops are mathematically endless and thus should be optimized for quick decay and stability. On the other hand, the MEG also has the ability to self-oscillate over the entire unit, in which case the feedback and feedforward loops are optimized for stable self-oscillation of the entire unit. The self-oscillating

MEG is permitted since it is an open system far from equilibrium with its active environment.

So if desired, an entirely different kind of optimization can be utilized for self-sustained nonlinear oscillation and closed loop operation. Because of the gain achieved in the large E-fields produced by the perturbed uncurled A-potential in space surrounding the core, there is sufficient excess energy collection available in the output coils to use the available feedback and feedforward loops in the system itself to achieve system self-oscillation. The system can be operated as a self-oscillating open system freely receiving excess EM energy from its active environment — the local active vacuum, local curvatures of spacetime, and local freely furnished uncurled A-potential and resulting nearly free E-fields from the A-potential's perturbations). As we are writing this, our experiments are now in this direction, and we are preparing the material for filing a formal patent application on this additional process. The patents will be filed, of course, by the time this book is published.

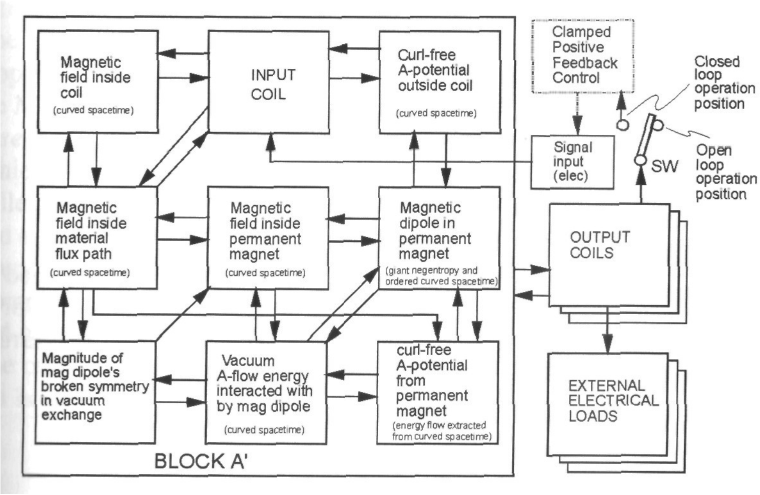
Magnetic field inside coil
(curved spacetime)
INPUT COIL
Curl-free A-potential outside coil
(curved spacetime)
Clamped Positive Feedback Control
Closed loop operation position
Signal input (elec)
SW
Open loop operation position
Magnetic field inside material flux path
(curved spacetime)
Magnetic field inside permanent magnet
(curved spacetime)
Magnetic dipole in permanent magnet
(giant negentropy and ordered curved spacetime)
OUTPUT COILS
Magnitude of mag dipole's broken symmetry in vacuum exchange
Vacuum A-flow energy interacted with by mag dipole
(curved spacetime)
curl-free A-potential from permanent magnet
(energy flow extracted from curved spacetime)
EXTERNAL ELECTRICAL LOADS
BLOCK A'

Figure 7-8 Energy gain process using feedforward and feeback subprocesses

A highly simplified block diagram of energy gain process using feedforward and feedback subprocesses is shown in Figure 7-8. As can be appreciated from this high order diagram, the actual feedforward and feedback loops are extremely complex, and a great deal of complex analysis is required for a self-oscillating, stable, controlled system. We also stress that control measures to insure and hold stable self-oscillation must be included in the overall features implemented in this self-powered, self-oscillation variant of the MEG. Not only must we use chaotic oscillation techniques to achieve the self-oscillation, but also we must use advanced techniques for stable control of such nonlinear oscillations, as indicated by Fradkov and Pogromsky {487}.

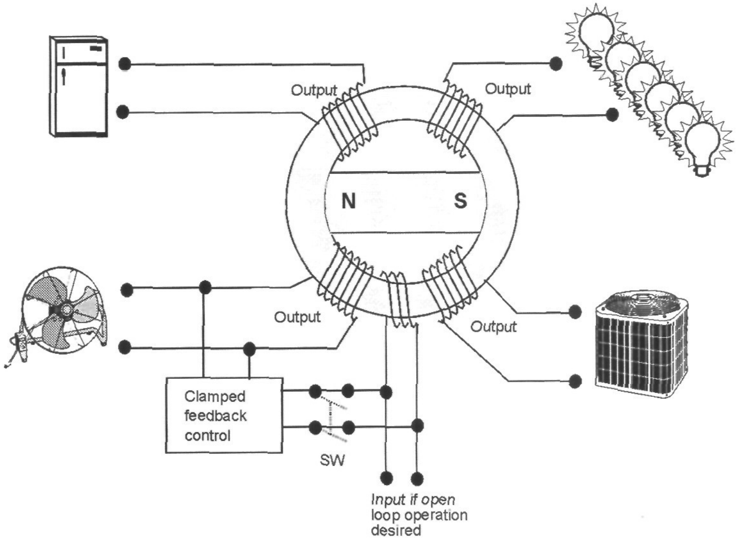
Output
Output
N
S
Output
Output
Clamped feedback control
SW
Input if open loop operation desired

Figure 7-9 Typical MEG embodiment application on site.

In Figure 7-9, we show a typical diagrammatic embodiment envisioned for the MEG in a household, using a large MEG to later be developed. For interim installations, the single large MEG will be replaced with an array of MEGs together with a synchronizer, as shown in Figure 7-10.

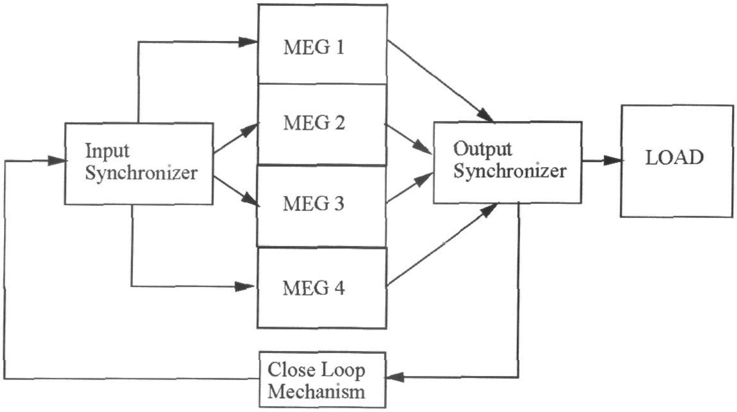

Figure 7-10 MEG self-powering array concept for increased powering applications.

Figure 7-11 shows a MEG unit where additional special collection antennas feeding separate circuits are employed in an "outrigger" fashion. One major advantage of this configuration is that the complexity of close-looping and its control are materially reduced, although the complexity of the MEG build-up and stabilization is increased. Another advantage is the *ease* with which close-looping for self-powering operation can be achieved, as shown in Figure 7-12. In this case, the external antennas are collected to independent output circuits (not shown) in the output section, and each of these independent outrigger outputs is a COP<1.0 circuit if we count the perturbed A-potential energy input as an "operator's" input. Completely normal clamped positive feedback can be used with sensing and governing, to allow self-powering and self-adjusting to varying loads. The outrigger embodiment can be made to almost entirely avoid the Dirac sea hole current problem.

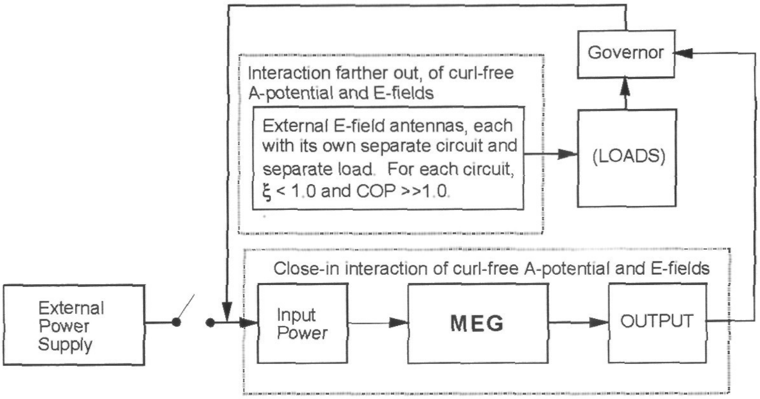
Governor
Interaction farther out, of curl-free A-potential and E-fields
External E-field antennas, each with its own separate circuit and separate load. For each circuit, ξ < 1.0 and COP >>1.0.
(LOADS)
Close-in interaction of curl-free A-potential and E-fields
External Power Supply
Input Power
MEG
OUTPUT

Figure 7-11 MEG unit with outrigger antennas for extra energy collection.

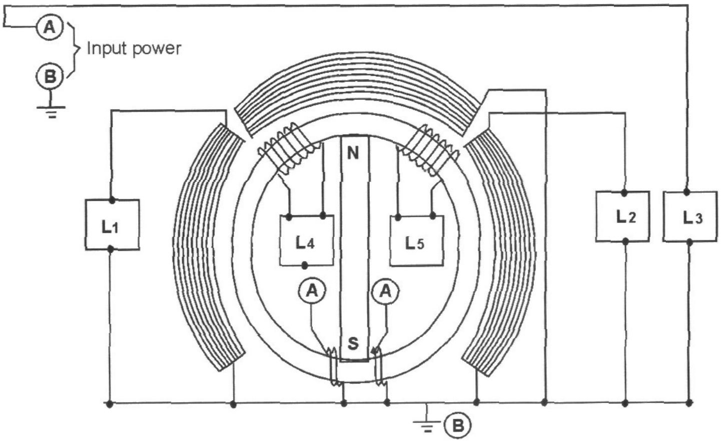
A
Input power
B
N
S
L1
L2
L3
L4
L5
A
A
B

Figure 7-12 Use of outrigger antennas for damped positive feedback method of close-looping.

As can be seen, the MEG has many possible configurations and embodiments, each of which has special advantages and uses. One of the embodiments we discussed is the deliberate use of self-oscillation forced by the feedforward and feedback loops. That is a separate embodiment, and capable of a very high power per pound ratio, but it also requires more sophisticated sensing and control techniques.

7.9 In Conclusion

By the time this book is published, Magnetic Energy Ltd. is hopeful it will have achieved its major funding requirement, and be well along the way to completing the research for full production and marketing of MEG units on the world market. We have arrived at that part of development — the need for substantial funding (several tens of millions of dollars) — where every previous COP>1.0 EM power system has died. We intend to do everything within our power to see that we obtain the funding, finish the complex research task, and get the MEG onto the world market.

For additional information on the MEG and on its preliminary replication by Naudin, the reader is referred to other technical articles {488} and to Jean-Louis Naudin's website.[213] Actual measurements for one of Magnetic Energy Ltd.'s MEG test units is given in reference {488a}. The MEG patent {37} shows the variation of COP for a test MEG with various conditions. Curves for Naudin's measurement of the performance of his replicated MEG unit are given on his website.

The MEG project is a most interesting project, because of the involvement of so much advanced physics and several different physics areas, as well as the prerequisite for higher group symmetry electrodynamics. Once funded, we are confident we succeed in crashing through the barriers and getting working MEG power units — including self-powering MEG units — on the market. We also believe the MEG will represent a substantial contribution to alleviating the coming worldwide energy crisis. Hopefully it will materially hasten the growth and evolution of the present highly centralized — and terribly vulnerable — national energy infrastructure into a highly decentralized — and much less vulnerable — national energy infrastructure. We also believe it will provide a great contribution to helping alleviate global warming, reducing pollution of the biosphere, and reducing the production of hazardous wastes. If so, it will be a continuing contribution, not only to humanity today, but also to generations yet unborn.

---

[213] Naudin's website MEG material and information is at http://jnaudin.free.fr/html/meg.htm. Particularly see his Mark III replication, shown powering a light on that webpage.

Chapter 8

Approach to Antigravity

[Large-scale antigravity exists] *"...baryonic, ordinary matter—the stuff of stars and of people—makes up just over 4% of the energy and matter in the universe. ...about 30% of the stuff in the universe is dark [unobserved] matter. The remaining two-thirds, theorists believe, is a mysterious 'dark energy' or 'quintessence'—a large-scale antigravity-like effect that is making the universe expand ever faster.... "* [Charles Seife] {489}.

[On the nature of the vacuum] *"...the (electromagnetic) vacuum is the ground state of the quantized electromagnetic field. ...The Dirac electron vacuum is that state with all negative energy electron states fully occupied."* [Aitchison] {490}

[On forces and how charge exists] *"Vacuum polarization, in general, alters the effective force law. Forces, in quantum field theory, are understood as being due to the exchange of virtual quanta ...In the case of QCD and QFD, ... a crucial new feature is that the force-field quanta themselves carry the 'charge' of the force-field, i.e. it is as if the photon of electromagnetism carried electromagnetic charge."* [Aitchison] {491}

[Possible gravitational circuits] *"There is an implication ... that gravitational excursions might be controllable by gravitational 'circuits', analogous to electronic circuits, which could lead to a controllable form of negative gravity."* [Robert Neill Boyd] {492}.

[Gravitational circuits] *"The supersystem concept utilizes a unified fields approach so that specialized electrical circuits in COP» 1.0 condition generate strong Dirac sea hole currents of negative energy in the immediate local vacuum. Thus a strong negative energy field is formed surrounding the circuit and changing the curvature of local spacetime as an action by the source*

> *circuit. This produces a direct antigravity field around the circuit or physical system, acting upon the mass of the system. The net gravitational interaction by that circuit or physical system is therefore the summation of the local positive gravitational field and the new local negative gravitational field produced by that negative energy field."* [T. E. Bearden, correspondence to a colleague, February 2002.]

> [On the multiplicity of spacetime curvature components] *"The full mathematical expression for the curvature of four-dimensional space-time ...is given by something called the Riemann curvature tensor. This is a somewhat complicated object, needing twenty real numbers at each point in order to specify it. These twenty numbers are referred to as its* components. *The different components refer to the different curvatures in different directions in the space-time."* [Penrose] {493}.

> [Physics studies and prepares imperfect models, not perfect truth] *"All we ever know is our models, but never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself"* [Hawking] {494}.

8.1 Introduction

In this chapter, we first present a highly abbreviated background of the theory of gravitation, sufficient for our experimental orientation. Of special interest are unified field theories, which cover general relativity, electrodynamics, and quantum mechanics in one overall master theory. Even more of interest is an engineerable unified field theory, which is presently closest approached by the theory of Mendel Sachs {510}.

We then present a short overview of work that has been done in the fields of lifters and inertial propulsion, listing some of the inventions of interest, the approaches, results achieved, and our comments on each.

Finally, I present my own approach to antigravitation, and give some results of the Sweet device's antigravity test (the test was designed by the present author) that successfully reduced the weight of a device by 90%.

In our view, this experiment fully demonstrated the veracity of our theoretical approach.

8.2 Selected Theories of Gravitation

8.2.1 Newton's Theory and a Negative Mass Extension

The first mechanism for gravity of interest to our purposes was advanced by Newton in his work on mechanics {495}. Newton's law of gravitation may be stated as {496}:

> *"A particle of mass* m_1 *attracts a particle of mass* m_2 *a distance d away with a force* $F = G\frac{m_1 m_2}{d^2}$ *in the direction of the line joining the particles. "*

Of course, it should have stated "in a time forward zone". The situation is precisely reversed in a time-reversal zone. It is also reversed for negative mass-energy.

In the Newtonian view, antigravity would mean repulsion of mass rather than attraction of mass. Given that negative mass is possible, then a negative mass would exhibit antigravity — i.e., it would be repelled by all other positive masses.[214] This can be seen in the quoted equation, by

---

[214] The discussion of what attracts what and what repels what, in a mix of negative and positive masses and negative and positive energies, is still controversial in physics. Different theorists have taken differing positions, and there is no formal experiment that has decided between them. In our opinion, the successful antigravity test of the Sweet vacuum triode amplifier answered that question experimentally, although independent replication was not accomplished. The scientific discussions heavily involve discussions (and assumptions) of the nature of "gravitational" mass, inertial mass, and electrodynamic mass etc. Again, there is no decisive selection among several possibilities, and there is no clear and universally accepted definition of **any** of those three types of mass.

In addition, all the discussions have included a fundamental error: They **assume** a separate 3-force acting upon a separate 3-mass, which is unequivocally false. Another fundamental error is the assumption of an observed (effect) 3-positron interacting with normal matter rather than a non-observed (causal) 4-positron (negative energy 4-electron) in massfree spacetime. In effect, this applies the CPT (charge-parity-time) symmetry, and converts the direction of the 3-positron to the antidirection of the 4-positron while also changing the negative mass of the 4-positron to positive mass of the 3-positron. Further, in physics there is still no firm **definition** of either force or mass, but only circular definitions between the two. We are using the simple definition of mass as $F = d/dt\ (mv)$, where mass is a component of force F and hence there is no separate force F acting upon a separate mass m. We interpret that equation as "the force *F* observed in a new observation is proportional to the time-rate of change of the ongoing 4-space interaction between the just-observed 3-mass *m* and the average interchange of time-energy and space-energy represented by v in the ongoing dynamics between spacetime and mass *m.* Whatever adjective is prefaced upon the *F* is also prefaced upon the *m* in the *mv* term.

We politely but firmly disagree with one prevailing notion that positive mass *attracts* both positive and negative mass, while negative mass *repels* both types. Further, we believe that the Sweet vacuum triode amplifier antigravity experiment falsified that notion. We hold to the dictum that $m == E/c^2$, where E is real spatial energy and may be positive or negative, c^2 is always positive, and therefore negative mass is highly compressed negative spatial energy while positive mass is highly compressed positive spatial energy. (Imaginary mass is a separate consideration.) We also prefer to reason in "single fundamental units" physics where the joule is the only fundamental unit. Gravity then becomes a function of the joule, whether electromagnetic, mechanical, or whatever. *We consider the energy in spacetime to he the source of the curvature of spacetime that is known as " gravitational field".* If the joule is of positive energy, the accompanying gravity is positive. If the joule is of negative energy, the accompanying gravity is negative. And we consider gravity as the 4-space causal spacetime curvature. So for gravitational interaction theory, we insist on examining the interaction of the curvature of local spacetime with the mass-energy. Further, we postulate that the gravitational effect upon one mass from a second mass is an improperly posed concept. It actually is an effect of the local curvature of spacetime around the first mass, where that curvature of spacetime was induced by the second mass's energy. Further, we consider that positive energy fields surrounding a positive mass will curve spacetime in the "positive gravity" direction. The positive mass is attracted by all other positive masses. Negative energy fields surrounding that positive mass system curve spacetime in the "negative gravity" direction so that the positive mass of the system is at least partially repelled by all other positive masses in the universe.

We further prefer to think in terms of the *supersystem* and the unobserved causal 4-field, not the frozen iterative 3-field snapshots defined only after iterative . observations. As we stated previously, the "observed" 3-positron is, in our view, a very different beast from the as-yet-unobserved 4-positron. The 3-positron is a frozen momentary "effect" entity with no dynamics at all and no continuous existence in time (and hence in spacetime) at all; instead, it continually *recurs* as the iterative effect of iterative observation by iterative application of the *d/dt* observation operator. The 4-positron is still a negative energy electron with negative charge, and still going in a non-reversed direction since broken parity has not been applied (because observation has not been applied), a causal entity with ongoing dynamics, continuously existing in spacetime, not just in 3-space. Until interacted and observed, it also has negative energy and negative mass-equivalency.

If one wishes to accurately deal with spacetime effects from the "positron," in our view one must deal with the 4-positron before observation, and hence with the supersystem. The 3-positron is not a *spacetime* entity, but only a frozen *3-spatial* snapshot. Scientists, including Dirac himself, were horrified at the negative mass implications of the negative energy states. Therefore, they moved the interpretation simply replacing positive mass m_1 by negative mass $-m_1$. That two negative masses would attract can be seen by replacing positive mass m_1 by negative mass $(-m_1)$, and also replacing positive mass m_2 by negative mass $(-m_2)$.

Now consider mass m_1 to be the rest of the net positive mass in the universe, and mass m_2 to be the "effective" mass in a local supersystem consisting of the local physical system, its local active vacuum, and its local active curvatures of spacetime — in short, in a local supersystem consisting of the physical system and its local active environment, and the interchanges between them. Let $m_2 = m_p + m_n$, where m_p is the amount of local positive mass (or mass-equivalent) in the system and m_n is the amount of negative mass (or mass equivalent) in the system. This supersystem now simultaneously exhibits both positive gravity (attraction) to all net positive masses in the universe and negative gravity (repulsion) to them as well. The net gravitational response of the supersystem depends upon the fraction of the effective mass that is positive and the fraction that is negative — in short, upon the net vector sum of the two gravitational forces.

to the 3-positron, and thereby twisted the interpretation to substitute the effect for the cause. Hence circuit theory treats the *lattice hole* as a positron, yet that "hole" acts very differently from a true Dirac positron (a Dirac sea negative energy electron) in free space.

Consequently, matter-antimatter interactions and the interaction between positive energy and negative energy are quite different, or they can be if so manipulated prior to observation. In the matter-antimatter interaction, the initial curvature of spacetime formed by the union decays by radiation of EM positive energy and the "effect particles" disappear back into the uncurved vacuum from whence they came. In the positive energy and negative interaction, there is a change of the local curvature of spacetime without its decay and subsequent re-radiation of EM energy. Equal amounts of interacting positive and negative EM field energy thus are simply transduced into stress energy potential of the vacuum and consequent sustained change in the curvature of the local spacetime. Indeed, the radiation of pair annihilation energy is obtained only at the expense of negating (relaxing) the local polarization of the vacuum (and the two consequent curvatures of local spacetime) represented by the independent positron and electron. Indeed, normal pair annihilation is a sort of Lorentz symmetrical regauging of the locally asymmetrical vacuum.

The reader is certainly free to disagree with our conclusions; we are merely explaining them. Experiment will eventually sort out the truth or falsity of the approach. So far, based on the Sweet experiment designed by the present author, a highly successful experiment was consistent with this approach.

In a "mixed" mass system containing both positive and negative component masses, the fraction of the total mass that is positive and the fraction that is negative is thus significant. When the mixture's negative mass is less than its positive mass, the mixture would exhibit a "reduced" gravitational attraction to other masses of the universe. Then the ratio of negative mass to positive mass might be taken as the "gravitational shielding" coefficient (fraction).

When the amount of positive mass and the amount of negative mass were equal, the overall mass system would exhibit no net gravity or antigravity, and would "float" or "hover" in the Earth's gravitational field. It would be equally attracted and repelled by the Earth's positive mass (and by all other positive mass in the universe, hence could — at least under certain conditions — *appear inertia-free.).* When the amount of negative mass exceeds the amount of positive mass, the overall mass system would exhibit a net antigravity effect. It would be repelled more strongly by the Earth's positive mass, as well as more weakly by the distant positive masses of the universe, and hence would accelerate upwards and away from the Earth. Notice that the key to antigravity in the Newtonian approach is the availability of negative mass, but in the new sense in which we use it. Note also that we did not discuss the separate issue of imaginary mass.

If we translate that into modern terms, where we consider not the "mass" but the condensed energy comprising the mass, then m_p and m_n become positive mass-energy and negative mass-energy terms respectively. Given the ability to produce electromagnetically produce substantial negative field energy in the immediate vacuum environment of the system, where the source of the field is the physical system, then the creation of a direct antigravity component in the system's environment — the other two components of its supersystem — becomes not only possible but also feasible.

The key to antigravity becomes the *availability of negative field energy in the local space surrounding the physical object, with the physical object containing the source of that surrounding spatial negative energy field.* Voila! This impels us directly to electrodynamics and a modified Dirac sea {497, 498a, 498b), one with excess empty holes as source charges producing negative energy EM fields. We deliberately create these unique source charges by actions taken in a COP»1.0 system. This gives the key to the approach to antigravity taken by the present author.[215] However, we use as negative energy not just the accounted *Poynting* energy flow component in space where the field is present, but also the unaccounted *Heaviside* energy flow component, which is much greater in magnitude (in a nominal case, nearly 10^{13} times greater). In this way, we dramatically increase the amount of antigravity effect that can be obtained by a given EM power dissipation in the physical electrodynamic system itself. *This antigravity effect of a Heaviside component of EM negative energy flow has not been accounted in previous analyses.*

8.2.2 Riemann Surfaces and Riemann Geometry

The next development of interest to us is the doctoral thesis of Riemann {499}, which founded the theory of Riemann surfaces. Three years later, **he** founded Riemann geometry in a famous lecture. Riemann was working on a unified explanation of gravity and electromagnetism when he died of tuberculosis a dozen years later. It was Riemann's multiply-connected topologies that eventually resulted in the view of electric charge producing lines of force trapped in the topology of space {500}. Had Riemann lived, we believe that engineerable unification of electrodynamics and gravity would have occurred early on. In that case, we might have already had practical antigravity vehicles and explored the solar system and beyond.

8.2.3 Heaviside's Excess Energy Circulation Flow and Gravity Theory

The next theory of gravity of interest to us is that of Oliver Heaviside. Heaviside noted that his extra EM energy flow term was a trapped EM energy circulation {501}. He realized that, if that trapped EM energy circulation were taken as gravitational in nature, his energy flow theory could be converted into a unified theory incorporating both electrodynamics and gravitation. After his death, his handwritten notes on **the** theory were found beneath the floorboards in the little garret apartment he had long occupied. No one knows what other papers he may have had; **the** place was ransacked and looted prior to the discovery of these remaining papers hidden beneath the floorboards — papers the looters had missed.

---

[215] We also point out that a Dirac sea with empty holes (excess negative energy states) exists in (or constitutes) a negatively curved spacetime *a priori*. So we get back to the other two components of our "supersystem", the active vacuum and its dynamics and the local curvature of spacetime and its dynamics.

Other scientists such as Laithwaite have felt that Heaviside's approach might still shake the foundations of physics itself. Quoting Laithwaite {502}:

> *"Heaviside had originally written the energyflow as S = (ExH) + G, where G is a circuital flux. Poynting had only written S = (ExH). Taking p to be the density of matter and e the intensity of a gravitational force, Heaviside found that the circuital flux G can he expressed as pu — ce, where u represents the velocity of p and c is a constant."*

8.2.4 Mach's Principle

After Heaviside, the next item of interest to us is Mach's principle {503}, which essentially states {504}:

> *"The properties of space have no independent existence but depend on the mass content and distribution within it. Thus, acceleration dealt with in Newtonian mechanics can only have meaning if it is an acceleration with respect to the stars or with respect to something equally well-defined."*

Mach's principle does not adequately deal with spacetime, and certainly does not even acknowledge or encompass how energy exists in 3-space at each point dipole, as an entering flow from the time domain into the negative (charge x time). It does not include transduction by the negative charge into 3-space energy, passage of the 3-energy to the associated positive charge, transduction back into time energy by the positive charge, and re-emission of the transduced time-energy back into the time domain, by the positive charge. In short, Mach's principle will never get one a practical antigravity system, regardless of how one applies the principle.

8.2.5 Einstein's General Relativity

Next comes Einstein's epochal general relativity theory of gravitation {505}. In that theory, Mach's principle is interpreted as pointed out by Misner, Thorne and Wheeler {506} in the following quote:

> *"...Einstein's theory (1) identifies gravitation as the mechanism by which matter there influences inertia here; [and] (2) says that this coupling takes place on a spacelike hypersurface..."*

The great contribution of Einstein's general relativity was to remove the notion of the "field occupying space", where field and space are separate things. Quoting {507}:

> *"If we imagine the gravitational field, i.e., the functions* g_{ik} *to be removed, there does not remain a space of the type (I) [Minkowski Space], but absolutely* *nothing, and also no 'topological space'. For the functions* g_{ik} *describe not only the field, but at the same time also the topological and metrical structural properties of the manifold. ... There is no such thing as an empty space, i.e., a space without field. Space-time does not claim existence on its own, but only as a structural quality of the field. "*

In passing, we note that in the conventional sense we can and do have many fields "in the same space", which is merely the statement that a net field has internal structure, and so does a net spacetime. Extending that notion, we can conclude that spacetime is in fact naught but a set of fields, both static and dynamic. The total energy density of all those fields at any point in space constitutes *a potential* with that energy density at that point, and with that internal field dynamics and structure. So because of its incredible field energy density, space is also *identically* a giant potential with internal structure and dynamics. With that observation, we rather casually join the seething vacuum of particle physics, the spacetime of general relativity, and the fields of electromagnetics as *different aspects of the same thing.* We point out that our concept of the supersystem therefore only identifies the three separate things into which this "one-thing" has been arbitrarily divided in much of physics.

Although he tried more than once, Einstein himself failed to achieve a successful unified theory of gravitation and electrodynamics. One thing he added in his earlier theory was a cosmological constant, which he felt was necessary to explain a stable closed universe. He later regretted that assumption, but today with the discovery of the acceleration of the expanding universe, which can only be caused by antigravity, his cosmological constant has assumed a new and vital role in the search for the source of this observed antigravity effect in the cosmos.

8.2.6 Kaluza-Klein 5-Dimensional Unified Theory

Using five dimensions, a unified theory of general relativity and electrodynamics was advanced by Kaluza {508} as extended by Klein (509}. This is a notable extension of Einstein's theory, and a very substantial literature exists in the Kaluza-Klein area.

8.2.7 The Sachs Unified Field Extension of Einstein's Theory

Another notable extension of Einstein's theory, which does include electrodynamics and the results of quantum mechanics as well, is the theory by Sachs {510}. Starting with Einstein's general relativity view and the group symmetry of its underlying mathematics, Sachs extended Einstein's theory into a fundamental theory of matter in all domains. Einstein's group formalism and the spinor-quaternion calculus in a curved spacetime provide the genesis of matter fields. The theory encompasses globally covariant field equations for inertial features of matter including its force manifestations. Unification is provided by the generalized spinor formalism. Quoting a popularized summation {511}:

> *"Novel features of general relativity are shown to emerge, in elementary particle physics as well as astrophysics. Especially noteworthy ...is a factorization of Einstein's field equations to a quaternion field formalism which, by way of a single covariant field, includes both gravitation and electromagnetism coupled to matter field equations — whose linear (low energy) limit recovers quantum mechanics. The ...original approach also furnishes derivations, not previously available, of several observed properties of inertial mass and gravitational force. "*

For a more technical description of how the unification of electrodynamics with general relativity is derived, we quote directly from Sachs {512}:

> *"It was demonstrated earlier that the factorization of Einstein's second-rank, symmetric tensor field equations, by removing the space and time-reflection elements from the underlying covariance group, yields a 16-component quaternion metrical field equation. The 16 equations were then re-expressed in a second-rank tensor form, breaking up into 10 $\oplus$ 6 equations. The 10 equations are in one-to-one correspondence with Einstein's symmetric tensor equations for gravitation; they are even under reflections. The remaining 6 equations are in antisymmetric tensor form; they are odd under reflections. Taking the covariant divergence of the latter equations then puts them into one-to-one correspondence with the structure of Maxwell's equations for electromagnetism. What was shown earlier was that the four of Maxwell's equations with sources followed. What is shown in this paper is that the other four of Maxwell's equations that are source-free also*

> *follow, thus indicating the absence of magnetic monopoles from this theory, as in the usual Maxwell formalism. This shows that the factorization of Einstein's field equations to the quaternion form fully unifies the gravitational and electromagnetic manifestations of interacting charged matter."*

To freely download several of Professor Sachs's important technical papers, one is referred to his website at http://www.compukol.com/mendel/articles/articles.html.

The O(3) electrodynamics of Evans and Vigier has been shown to be a special subset of Sachs's electrodynamics {513a}. Hence O(3) can be used to engineer much of the Sachs unified field theory phenomenology, including — in our view — both gravity and antigravity. The first theoretical indication of this has been published by the AIAS (Alpha Foundation's Institute for Advanced Study) {513b}. Shortly we discuss our own views on how to directly perform engineering of antigravity.

A great many excellent publications on general relativity and the background of scientific knowledge in gravity exist; e.g., the text by Wheeler *et al.* {514} is particularly good. Einstein himself wrote a book on his special and general theory for the nonspecialist {515}. A collection of English translations of some of the original papers on relativity is provided in a book by Einstein, Lorentz, Minkowski, and Weyl {516}. Many other gravitational theories or changes to major theories have also been proposed by other scientists, but are not discussed here. For our purposes, an abbreviated discussion will suffice to understand a fundamental approach to feasible antigravity.

Since we are interested in actual working systems, a short review of some of the efforts in asymmetric lifting and propulsion experiments is useful.

8.3 Selected Antigravity and Inertial Propulsion Research Efforts

We give only a small sampling of the many efforts performed or ongoing in this area over the last few decades, and some still ongoing today.

8.3.1 T. Townsend Brown

Our first example is the work of T. Townsend Brown. Brown was working **on** a "gravitor" before he finished high school, and continued work in electrogravitation throughout his life. Space is a dielectric, and Brown reasoned that the displacement current in a capacitor dielectric could influence the space dielectric itself.[216] Hence he built, tested, and patented several variations of lifting devices based on this concept {517a-517e}. Essentially, he found that certain high capacitance dielectrics could be subjected to high voltage (50 to 300 kilovolts), and would exhibit thrust in the direction from the negative to the positive pole.

Here we point out our contention that all EM energy in space comes from the time domain via the negative charge, and re-enters the time domain via the positive charge. This is consistent with Brown's finding. Although he demonstrated many such lifting and thrusting devices, Brown was never able to obtain sufficient financial and scientific backing to bring a practical embodiment of his inventions into practice. His work, however, has been replicated by multiple experimenters, including NASA researchers (see below) and several researchers contributing results to Jean-Louis Naudin's website http://jnaudin.free.fr/html/lifters.htm.

8.3.2 Jonathan W. Campbell

Recently a U.S. patent {518} on a thrusting capacitor, by Campbell, has been issued and assigned to the U.S. Government (NASA), and made available for commercial licensing. The patent write-up states that thrust from capacitors is well known, but has been too small to be practical. The patent purports to remedy that impracticality. A version of the NASA device has been replicated by the Naudin researchers also. So an adaptation of Brown's work has been patented by the U.S. government with intentions of licensing. This would appear to validate much of Brown's approach and efforts.

8.3.3 The Dean Drive

Our second example is the inertial drive patented by Dean {519a-519e}. Unfortunately, the effects of the Dean drive later appeared to be a problem in accurate measurement. Nonetheless, work still continues along the direction shown by Dean. One may argue that mechanical materials also contain charges and have a certain kind of capacitance. Hence, mechanical motion of the parts of the machine does move these charges and dielectrics, opening a possibility of interacting with the dielectric of space. For example, Sweet's VTA involved a sustained self-oscillation between the local vacuum and the barium nuclei in his barium ferrite magnets.

---

[216] We have not yet explored the gravitational effects — if any — of our concept of the triode capacitor, discussed in paragraph 6.3.7 and diagrammed in Figure 6-17. We will examine it as soon as we have the proper equipment.

Dean's approach still has merit, even if a particular embodiment gave a negative result when very accurate measurement was performed. It is particularly of interest that mechanical force is essentially derived from electromagnetic forces (and fields) at basis. These base EM fields involve (and identically are) curvatures of spacetime, and hence the motions of the inertial drive mechanical apparatuses do involve at least some curvatures of spacetime and dynamic changes to them.

Our comment is that, when the additional but neglected Heaviside nondiverged energy flow component is accounted, these field interactions internal to matter can produce many orders of magnitude greater curvature of spacetime than is accounted by the diverted Poynting energy flow component only.

The "inertial antigravity" approach may be visualized as the use of excess gravitational thrust in a specific direction. If that direction is against the local pull of gravity, and is more powerful, then "lift against the gravitational pull" is obtained. Rigorously, this is not true antigravity but manipulated excess positive gravity.

In our approach to antigravity, one way to approach the problem is to have the mechanical apparatus also the source of an intense *negative energy* EM **field,** producing an intense flux of Dirac sea holes into and in the local surrounding spacetime. The excess charge removed from the Dirac holes can in fact be used in the electrical powering of the physical system, as was demonstrated in the Sweet VTA antigravity test. Then movements of the mechanical parts could involve movement of strong negative energy fields, hence strong curves of local spacetime that are local *strong negative gravity fields*. Or, better yet, movement of the charges themselves will also produce field-induced movement of the Dirac sea hole negative energy. This appears to be a practical method to manipulate the metric itself, along the lines proposed by Puthoff *et al.*[217]

8.3.4 Henry Wallace

Other work has been by Henry Wallace, an engineer working for General Electric in the 1970s. His work dealt with the interactions of relatively moving masses of various kinds, and it is described in his patents {520a, 520b}. The discussion under the Dean drive probably applies to Wallace's work as well.

---

[217] H. E. Puthoff, S. R. Little, and M. Ibison, "Engineering the Zero-Point Field and Polarizable Vacuum for Interstellar Flight," *J. British Interplanetary Soc,* Vol. 55, 2002, p. 137-144.

8.3.5 J. F. Woodward

Next is the work of J. F. Woodward, resulting in one or more patents {521} and multiple articles in the legitimate physics literature {522a-522k}. Woodward uses an adapted theory based on Mach's principle and on the thesis that the mass of a capacitor can be made to vary from its normal mass by rapidly charging and discharging it — say, at a frequency of 20 kHz. This predicted effect is testable, although some sophistication in the testing is required.

Woodward's publications show the results of actual experiments, measuring a very small change in the mass, and are worthy of much further investigation. Here we note that the actual mechanism generating the inertia of an object is still an open issue, with several different contending theories.[218] Whether the gravitational mass and the inertial mass are identical or not, also remains an issue. If the exact local mechanism generating the inertia of a mass could be decisively found, then direct methods of changing the inertial mass and the gravitational mass might be a real possibility, and research to try to find such methods would be strongly indicated.

Our further comment is that the mechanism that generates mass itself has never been specified in formal physics,[219] particularly when one views an observable (such as mass) as continually recurring in 3-space rather than continuously existing in 3-space, due to the *d/dt* operator imposed on spacetime by the observation process. Before the reader assumes that mass is well understood, a reading of Jammer's scholarly work {523} is highly recommended.

From our viewpoint we stress that observed mass cannot continuously exist, but must be an effect and the instantaneous frozen output of the *d/dt* observation process invoked on an ongoing 4-space process. Hence, from this view there is indeed a "process" that (continually) creates mass — it is at least the emission of EM angular momentum (a photon) from the *masstime* entity. Similarly, there is a "process" that continually destroys or converts any mass — it is the absorption of a photon and conversion of the mass m into masstime $(m + dm)dt$. Mass "flows through time" via this time charging and discharging, in the form $m \Rightarrow mt \Rightarrow m \Rightarrow mt \Rightarrow m$... etc.

---

[218] E.g., (a) A. Rueda and B. Haisch, "Inertia as reaction of the vacuum to accelerated motion," *Phys. Lett. A*, Vol. 240, 1998, p. 115-126; (b) B. Haisch, A. Rueda and Y. Dobyns, "Inertial Mass and the Quantum Vacuum Fields," *Ann. Phys.*, Vol. 10, 2001, p. 393-414.

[219] The present author made an early **and** crude attempt at doing it in his "Quiton-Perceptron Physics." 1973, NTIS Report AD-763-210.

Again we stress that masstime mt is not mass m, but is as different from it as impulse Ft is from force F. We crudely proposed such an initial mass-generating mechanism in 1973 {524a}, with additional information published in 1998 {525a}, which may have bearing on the mechanism involved in Woodward's experiments.

8.3.6 E. R. Laithwaite

Eric Laithwaite was for some years a professor with the Imperial College in London. He was one of the pioneers of the linear electric motor and of portions of the magnetic levitation (MagLev) train concept. An illustrious career headed for greater things (perhaps knighthood, Fellow of the Royal Society, etc.) was abruptly derailed in 1973 after he lectured to the prestigious Royal Institution by invitation and demonstrated that a heavy gyroscope, difficult to lift with one hand while not turning, could be lifted easily with that same hand, once turning at speed. Laithwaite's suggestion that Newton's laws of motion might be in difficulty proved much too controversial. The Royal Institution (apparently for the first time in its 200 year history) did not publish proceedings of that invited discourse.[220]

One of the remarkable things Laithwaite achieved was a form of "indefinite motion" device that he demonstrated circa 1970 {526}. The Centennial Centre of Science & Technology in Ontario was looking for working models to decorate the entrance hall. Laithwaite showed that if a linear-motor primary was rolled into a cylinder, so that it became simply the stator of a conventional induction motor,[221] a steel washer (about 3 cm. in diameter) could be made to roll continuously in a vertical plane around the inside of the stator. A combination of centrifugal force and magnetic attraction kept the washer in contact with the stator at all times. At the same time, a pivoted copper cylinder could be held in the center of the stator and would be seen to rotate on its own axis in the opposite direction.[222] Laithwaite completed and delivered a working model to the Centennial Center of Science & Technology. The reader should note the similarity of Laithwaite's proven experimental device to the magnetic Wankel engine.

---

[220] With some humor we point out how shocking is our proposal that such a thing as "continuous" propagation of an object or energy through 3-space does not exist. Instead, small differential "pieces" of it continually "recur" at successive positions, much like the projection of the frames of a motion picture frame, where continual observation is the "projection process". Nonetheless, in our proposed solution to the source charge problem, either that is true or every charge in the universe is a perpetual motion machine, freely creating energy out of nothing. In our view, that is the impact of the long-neglected source charge problem in quantal and classical electrodynamics. It has produced the wrong and naive notion of what "propagation through space" actually entails.

[221] Something similar to diagrammatic Figure 6-26 in Chapter 6.

In addition to his attention to Heaviside's unpublished draft notes and manuscript on a combined theory of electromagnetics and gravity (previously discussed above) {502}, Laithwaite spent the latter years of his life working on sophisticated gyroscopic systems {527a-527h}. Ultimately he felt he had successfully produced a mass transfer effect {528} and the basis for a propulsion system. He and William Dawson applied for a patent in 1993, which was granted in 1995 {529}, with a U.S. patent granted in 1999 {530}. Regretfully, Professor Laithwaite died in 1997.

8.3.7 Floyd Sweet's Vacuum Triode Amplifier (VTA)

See again Chapter 6 and Figures 6-5 through 6-14 on the work of Floyd Sweet, previously discussed in paragraph 6.3.2. My concept of antigravitation used in the experiment is shown in Figure 6-15. By now, the sharp-eyed reader will recognize that my method actually constitutes a conceptual extension and definitization of the Dirac electron theory, reinterpreted to recover negative mass, negative energy, negative energy fields, and negative energy potentials. The defining feature is the deliberate use of a $COP \gg 1.0$ main power system (perhaps constituting the power system of a flight vehicle) as a source of locally strong negative energy EM fields generated by the concentrated Dirac sea holes (causal 4-positrons being negative energy 4-electrons prior to observation) locally produced in the power system. We collaborated with Sweet on the vacuum triode amplifier for more than a decade, and in fact suggested the name of the device.

The antigravity approach designed by the present author was based on realizing that negative energy sources (e.g., 4-positrons as negative mass 4-electrons prior to their reaction and observation) produce *negative*

---

[222] We have had the distinct notion that the fundamental *contrarotating* mechanism found by Laithwaite is probably an analogous mechanism — or very similar — to the mechanism utilized by Bauman in his Swiss converter (Methernitha device). As pointed out, it also has remarkable similarities to the magnetic Wankel engine, except that the magnetic Wankel does not utilize two contrarotations.

energy EM fields. The field can reach out over all space, so a negative EM energy field represents a reversed curvature of a large volume of spacetime. Hence it represents a considerable antigravity field — particularly when one considers the unaccounted huge Heaviside component of the negative energy field and charge interaction. Since the field strength falls off as the inverse square of the distance from its source, the strong antigravity effects of negative energy EM fields are much stronger locally than at a distance. This is particularly true when the conventionally ignored Heaviside nondiverged EM energy flow accompanying the accounted diverged Poynting energy flow of every field-charge reaction is accounted. This usually ignored factor dramatically increases the antigravity intensity produced by the local negative EM energy field by many orders of magnitude, often by nearly 10^{13}.

My antigravity concept (then and now) is shown in Figure 6-15 of Chapter 6. At the time, the only available device I knew of that produced a substantial percentage of output negative EM energy was Sweet's vacuum triode amplifier. So I persuaded Sweet to fabricate a new output section (an output load similar to Figure 6.8 but containing ten sockets for ten 100 watt lamps). Thus the output of the unit — which adjusted automatically to the available load — could be increased in 100-watt increments by merely screwing in the lamps one at a time and making laboratory readings after each new 100-watt load increase was added. This was done and the concept was successfully tested by Sweet {531}, who was making the measurements in California and reading them off over the phone to me (I was in Huntsville, Alabama). The experiment worked beautifully.

The setup and progress of the experiment were intended to materially increase the strength of the *negative energy* EM field and Dirac sea hole flux already produced to a lesser extent by the Sweet vacuum triode amplifier (VTA) in its normal operation. Thus the hypothesized negative energy antigravity field, produced weakly by the negative energy from the device in its ordinary power output range of 500 watts, was to be substantially increased at 1,000 watts. The response to the net gravitational field surrounding the object would thus be the resultant of two factors: (i) the normal response of the device's positive mass to the positive gravitational field of the Earth etc., and (ii) the additional response of the device's positive mass to the local antigravitational field produced around it. The measurable result was expected to be a decrease in the measured weight of the object on the laboratory bench.

As discussed, this stratagem succeeded dramatically. The weight of the vacuum triode amplifier was decreased, smoothly and controllably, by

90% on the laboratory bench (see again Figure 6-9).[223] The results were later published in a joint paper {531} prepared by the present author. I placed Sweet's name first on the paper because he was the inventor of the VTA, which demonstrated the effect once properly modified as I requested. Note that the extra antigravitational field effect on an object is indeed a local effect, which the experiment also showed. This effect was obtained at a COP>1,500,000 (actually, at nearly a COP = 3,000,000) and at ELF frequency.

The ELF frequency is important. As we explained previously, the highest energy photons of all are the lowest frequency photons, when the highly compressed time-energy component of the photon is accounted as well as the decompressed spatial energy component. Hence, for very appreciable local curvature of spacetime to produce considerable antigravity, the antigravity source must produce negative energy fields consisting of negative energy photons at low frequencies (such as 60 Hz). Else, the curvature of spacetime is so small that little effect will evidence. My best "back of the envelope" estimate of the efficiency of the antigravity process in the VTA was 10%.

8.3.8 Ning Li and (separately) E. Podkletnov

Very interesting work was performed by Ning Li {532a-532d}, a scientist originally with the University of Alabama in Huntsville (UAH), who returned to China for a while but is now back in Huntsville. Quoting David Brown from the UAH website from 1997:

> *"In 1989, Dr. Ning Li of UAH predicted that if a time varying magnetic field were applied to a superconductor, charged and deformed lattice ions within the conductor could absorb enormous amounts of energy via the*

---

[223] Many times I have been asked how such small output power increase (from 500 to 1,000 watts) could possibly achieve such dramatic antigravity results. The answer - due to the neglected Heaviside nondiverged energy flow component — lies in multiplying the power output by the "Heaviside energy amplification factor *k* " where *k* may be as large as 10^{13}. A 500-watt accounted increase, multiplied by 10^{13}. is actually an incredible increase in the total EM negative energy field density being created in the local vacuum. In short, a very little "accounted" power increase of the COP»1.0 VTA output is actually a very vast negative power output increase when the long neglected k-factor is accounted. It is just this missing k-factor and the failure to account for the negative energy KM fields generated in COP»1.0 discharges in violent astronomical phenomena that seem to be responsible for generating the excess antigravity that is accelerating the expansion of the universe.

> *magnetic moment effect. This acquired energy would cause the lattice ions to spin rapidly about their equilibrium positions and create a miniscule gravitational field. Dr. Li's calculations showed that if these charged, rotating, lattice ions were aligned with each other by a strong magnetic field, the resulting change in local gravity would be measurable."*
>
> *"Podkletnov and Nieminen (1992)* {533} *made the accidental discovery that a single-phase, dense, bulk, high* T_c*, superconducting, ceramic disk spinning at 5,000 rpm can produce a 2 percent reduction in the weight of non-conducting, non-magnetic objects placed over the spinning disk. UAH and MSFC [Marshall Space Flight Center, a part of NASA] [cooperated] on a joint research project to independently confirm the results of the Podkletnov experiment and to validate Dr. Li's theory of gravity modification via superconductor. On March 26th, 1997, as a part of this project, the joint UAH-MSFC research team produced the largest high temperature superconducting disk ever manufactured in the USA. This disk measures 12 inches in diameter and is 0.5 inches thick"*

We now explain the experiments, from our point of view of the mechanism for antigravity. We remark that, at least in theory, the charges in such a rotating superconducting disk have an inward "discharge" acceleration force that is a function of the distance of the charge from the rotation axis center. Hence, in the rotating frame, some of these charges are very strongly "accelerating" toward the center, as are their spins whose magnetic vector is orthogonal to the E-field vector. A flow of EM energy exists perpendicular to the rotating flat disk and is accelerating at right angles to the disk surface. This is analogous to a *discharge* (sudden acceleration) phenomenon, and it will also produce accelerating negative energy Dirac sea 4-holes in the vacuum component of the supersystem of the spinning disk. So from the spinning disk there is a sort of "wide beam" of Dirac sea 4-hole current in the vacuum, radiating at right angles to the disk surface, with heavier concentrations toward the outside of the disk. The result is that objects in the Earth's gravitational field, and also placed in the path of this negative energy Dirac sea 4-hole current, is acted upon by two superposed curvatures of local spacetime: (1) the positive curvature due to the Earth's gravitational field, and (2) the negative curvature due to the negative energy field's curving local spacetime. The total net gravitational effect will be due to the algebraic sum of the superposed two spacetime curvatures — the normal positive curvature from the Earth's gravitational field and the second a negative curvature created by the Dirac sea hole current's negative energy fields (Heaviside energy flow component considered). The expectation is therefore a *reduction* in the weight of the object, as is shown in the experiments.

The UAH-MSFC group was unable to confirm Podkletnov's findings.[224] Ning Li and the NASA group later split amicably. Ning Li continued her theoretical work, eventually ceased publishing papers for a period to protect the techniques and experimental results achieved in her work, and focused upon potential applications. She apparently returned to China for some time, but is now back in Huntsville, Alabama and still experimenting. We suspect there may also be a Chinese research effort into antigravity, using the approach taken by Ning Li. Perhaps information from that program will be released in the future.

But what of the work of Podkletnov since his original experiments? NASA continued to try to replicate his disk and his work, with Podkletnov acting as a consultant. Presently, NASA has paid an independent research firm to replicate the disk to full specifications, which has just been accomplished as this is written. At this writing, the experimental results of this new attempt are not yet available as the work is still in progress. It should be available by the time this book is published.

Meanwhile, a relatively new paper by Podkletnov and Modanese has been published on the Los Alamos National Laboratory website.[225] This method uses high amperage sharp pulse discharges from a superconducting ceramic electrode, to produce an anomalous radiation in a focused beam. Significantly, the beam does not noticeably attenuate through different materials. The beam exerts a short repulsive force on small movable objects on the propagation axis, with the force proportional to the mass of the objects.

---

[224] E. Podkletnov and R. Nieminen, "A possibility of gravitational force shielding by bulk $YBa_2Cu_30_{7-x}$ superconductor," *Physica C,* Vol. 203, 1992, p. 441-444.

[225] Evgeny Podkletnov and Giovanni Modanese, "Impulse Gravity Generator Based on Charged $YBa_2Cu_10_{7-y}$ Superconductor with Composite Crystal Structure," carried on website http://xxx.lanl.gov/abs/physics/0108005.

We interpret the results as completely consistent with our own proposed excess gravity and antigravity mechanism. We know that such sharp discharges produce significant Dirac sea 4-hole currents in the local vacuum component of the supersystem. Such currents in a beam interact very little with intervening materials, and the major spacetime curvature effects are concentrated upon the propagation axis. On that axis, force effects can be expected upon small independent objects, due to the *additional* negative curvature added to the local spacetime in their supersystem. The impulse produced on the object should indeed be proportional to the mass of the object and independent of its composition, since it is a gravitational force effect induced by local curved spacetime. The force should be repulsive in the beam, since that is the direction of the antigravity force that will be induced in objects in the axis of the beam, by the beam of negative energy Dirac 4-holes. A positive gravity force would mean attraction between the object in the propagation path and the source of the beam. A negative gravity force would therefore mean repulsion.

So Podkletnov and other scientists are still working on a successful effect. It will be interesting to see what results are reported in the future, and whether they will deviate from their present attempt to explain the effect by a combined quantum gravity and anomalous vacuum fluctuations approach.

8.3.9 NASA's **Breakthrough Propulsion** Physics **(BPP) Project**

NASA's Breakthrough Propulsion Physics Project was established in 1996. It seeks a breakthrough in space transportation involving "(i) propulsion that requires no propellant mass, (ii) propulsion that attains the maximum transit speeds physically possible, and (iii) breakthrough methods of energy production to power such devices." Investigations involve coupling of gravity and electromagnetism, the quantum vacuum, hyperfast travel, and superluminal quantum effects. The project is managed by the Glenn Research Center, sponsored by the Advanced Space Transportation Program, with overall management by NASA Marshall Space Flight Center, Huntsville, Alabama.

NASA continues its investigations and experiments in methods to change **the** force of gravitational attraction. The scientist in charge of the project is David Noever of the Marshall Space Flight Center. The work is still considered a work in progress, with the need for both experimental investigation and meticulous design of experiments, instrumentation, and measurement techniques.

The overall NASA program also spreads to other supporting laboratories and projects. For example, at the Jet Propulsion Laboratory (JPL) Kulikov {534} has measured the difference between inertial and gravitational mass of a boson in a heat bath. At least one invention seems to have resulted or to be in progress, assigned to JPL.

As mentioned, the U.S. Patent by Campbell {518} on a capacitor device producing a net thrust force (a more modern version of T. Townsend Brown's work) has been issued and assigned to NASA.

8.3.10 Scientific Conferences

There are regular conferences (several per year) being held in the U.S. and elsewhere, dealing with antigravity and electrogravitic propulsion. We give a little-noticed but pertinent example: In a nice presentation in 2001, LaPointe {535} proposed a new method for producing anti-matter. Currently antimatter production is accomplished in very large accelerators, and is therefore quite expensive. For increased experiments, a cheaper and much smaller (and much more available) process is highly desirable.

In the Klein paradox, which involves very large potential drop across a very short distance — less than the Compton wavelength of the antimatter particle — matter and anti-matter pairs are produced. LaPointe hypothesized that it may be possible to use the Casimir effect to push one side of the step into the negative energy range. He is at least *thinking toward* negative EM energy sources that produce negative EM fields and negative gravity in the surrounding space, without explicitly enunciating it in that fashion.

We already briefly discussed our own approach to antigravity, which is a much simpler and cheaper anti-particle producing mechanism, as demonstrated by the Sweet device. In short, overunity EM systems with COP»1.0 do produce free flows of causal (before observed) 4-positrons (still in negative energy 4-electron state prior to interaction) in local spacetime immediately surrounding the system. Consequently, LaPointe's efforts are in what we believe to be the correct direction, but he is still using a very difficult mechanism to engender the antimatter production. Also, the positrons or other antimatter sources of the negative energy fields must remain in the causal (unobserved and non-interacted) 4-state, in order to produce negative energy EM fields and potentials and antigravity.

8.3.11 James Corum *et al.*: Rectified Slepian AC Force Resonator

Another example is some promising work by Corum *et al.* {536}. We quote from the abstract and summary of his paper:

> *"Corum presented an experimental paper on the use of the Heaviside force in conjunction with a Slepian antenna as a form of space drive using nothing more than the classical Maxwell stress tensor. Slepian proposed the same thing in 1949, but came to the conclusion that it would not be useful, since the time average of the resultant AC force would be zero. Corum's contribution has thus far been two-fold: (1) In conjunction with Dr. Alan Barnes of WVU he has experimentally shown that the AC version of the Slepian Resonating Antenna does produce a force, and has measured it to within 3.6%, and (2) has designed a way for Hartley's variable capacitor rectification to be used with the Slepian resonator such that the rectification results in a DC force component. The first experiment has already been achieved, the second experiment is the logical next step. If successful, the result would be quite revolutionary: a true space drive."*

Our comment is that this unilateral thrust force work is important. In our view, the Heaviside energy flow component does not appear in Slepian's vector $\mathbf{j}\phi$, which contains only the diverged EM energy flow surrounding a circuit or component, but it is present immediately outside the circuit where the Slepian vector exists. Hence there is an enormous, extra energy flow closely associated with every EM circuit, that is unaccounted since Lorentz's closed surface integration of the energy flow vector discarded it quite arbitrarily.

We already know from the Bohren experiment (and thousands of others similar to it) previously cited that resonant charges do collect excess EM energy from a given EM energy flow input, compared to the collection from that flow by the same charges in nonresonant state. In using the Slepian resonator and then rectifying its output, it would appear that both the diverged (Poynting) energy flow component and the Heaviside nondiverged energy flow component are first resonated and then rectified. In addition, excess energy is collected — much more than is in the normal "static charge" formula assumptions for Poynting energy flow and in assuming only a "single pass, single interception and collection" of the energy.

An important factor then is obvious: Once the energy flow is surrounding the circuit, there exists a certain local curvature of the immediate external spacetime. The rectification represents not only a change in the Slepian vector current in the circuit, but also a change in this local ST curvature interaction upon the circuit matter and charges. In short, it should produce a DC component in the direct excess interaction of the rectified local curvature of spacetime. This should produce *excess direct propulsive force,* just as Corum theorizes and the experiments support to date.

Further, the oscillation of the large, *unaccounted* Heaviside nondiverged energy flow component involves the oscillation of a "dark energy" component that may be orders of magnitude greater than the *accounted* Poynting component's magnitude. By involving this huge energy flow component in oscillation and synchronized to the resonant Slepian circuit, Corum has unwittingly involved the supersystem as a COP»1.0 EM "energy converter" system. Hence, there is appreciable Dirac hole current (in causal 4-form) in the local vacuum, not reacted with the materials in the system and therefore not converted to material lattice holes attached to positive mass. The time rate of changes of the negative energy fields and potentials are thus significant for antigravity effects, in the approach taken by the present author. In short, we believe Corum *et al.* are in fact producing significant evidence for a workable approach to unilateral thrust force.

This work by Corum *et al.* thus is highly promising, is solidly based experimentally, and certainly bears watching closely. We would strongly suggest that Corum *et al,* analyze their proposed mechanisms and results obtained in terms of supersystem analysis rather than system analysis, using higher group symmetry electrodynamics, and recover and include the long-neglected Heaviside energy flow component.

8.3.12 Gravitec, Inc. and H. Serrano

A patent {537} was issued to H. Serrano on Oct. 5, 2000 and assigned to Gravitec, Inc. whose principal investigator and corporate officer is Tim Ventura. Based on this invention, Gravitec, Inc.. has developed and demonstrated a rotary device that develops unilateral thrust. Purdue University began a small but steady effort to test the unit, first in the atmosphere and later under high vacuum conditions. The first tests have been successful, and a paper {538} from the Purdue study shows it is not the ion wind effect, which is orders of magnitude too small to account for the demonstrated thrust. Further, Jean-Louis Naudin has replicated and tested the Serrano thruster, and it works beautifully {539}. Ventura also hopes to develop a toy kit, with safe power supply, which would allow persons to perform their own successful lifter experiments and perhaps rotary experiments at nominal cost.

8.3.13 Transdimensional Technologies, Inc.

There is also ongoing important propulsion work at Transdimensional Technologies, Inc. of Huntsville, Alabama {540}. Founded in 2000 by its chief scientist, Jeffrey Cameron, the company has developed several novel propulsion and lifting mechanisms. Independent tests of some of these systems have been accomplished by Jean-Louis Naudin, and the results can be seen on his website (previously cited).

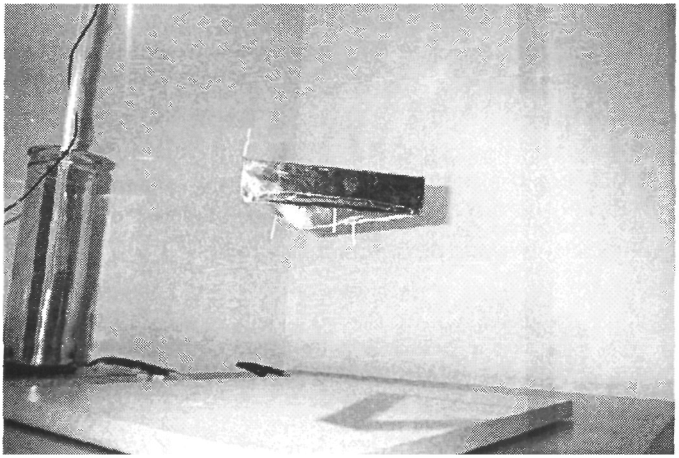

Figure 8-1 Naudin's **replica of the T**ransdimensional Technologies **lifter.**

As we go to press, Transdimensional's most recent propulsion effect is called Power3 (power "cubed"). It uses an asymmetric capacitor to produce the thrust in space, and the principle has been successfully tested in a vacuum. A larger, triangular lifter made mostly of pipe was demonstrated in vacuo to NASA officials in Huntsville, Alabama. Cameron has patents pending on magnetocaloric hybrid rocket engines and asymmetric gravitational waves {541}. A replica of the Transdimensional lifter successfully tested by Naudin is shown in Figure 8-1, by courtesy of Jean-Louis Naudin. Naudin's replica of the Transdimensional Technologies rotational device is shown in Figure 8-2.

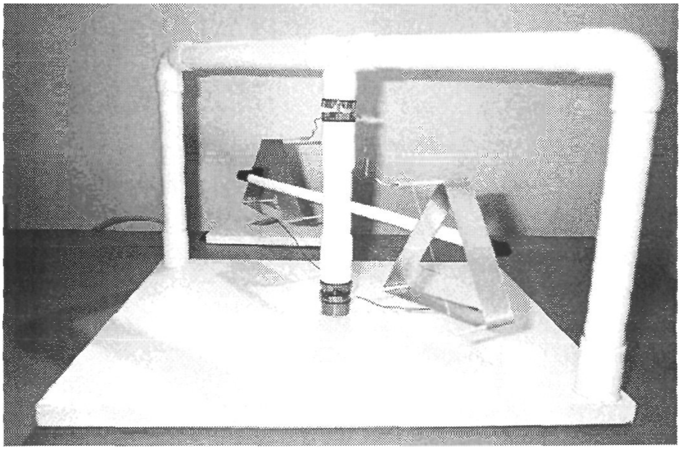

Figure 8-2 Naudin's replica of the Transdimensional Technologies rotational device.

We understand that new Transdimensional Technologies systems, now in patent process and still closely held, represent dramatic advances in the state of the art. The corporation foresees eventual commercial licensing of some of its developments, and the development of practical lift vehicles and space propulsion systems within 3 to 5 years.[226]

---

[226] Ken Moore and I personally visited Transdimensional Technologies here in Huntsville, and we found a very straightforward, enthusiastic, and hard-working group. We were treated with great courtesy, so much so that Jeff Cameron graciously cut short his vacation day and came in to show us the lab equipment and give a demonstration. We very much appreciate the work that Transdimensional Technologies is doing, and we certainly wish this fine group outstanding success in their efforts. We also are very appreciative for their courtesy and consideration in showing us the demonstrations and for their cogent discussions with us, which **were** most helpful.

8.3.14 **American Antigravity** (Tim Ventura's Company)

Tim Ventura is a very energetic lifter and antigravity researcher, who has done work independently and for others such as Transdimensional Technologies. The results of his many experiments and replications, tests, photos, etc. are on his website at http://www.americanantigravity.com. He also has several videos showing his testing, as well as detailed instructions for a researcher to build his own demonstration unit and verify the tests. Tim kindly furnished some needed information in this area, and put me in touch with other researchers and websites, which I appreciate.

8.3.15 Jean-Louis Naudin

Jean-Louis Naudin is a tireless researcher in the fields of COP>1.0 systems and processes, and in the field of lifters, inertial thrusters, and possible antigravity systems. His website is a cornucopia of information, patent references, inventions, processes, and replications or replication attempts by Jean-Louis and others who regularly work on such things and communicate with him. The website is http://jnaudin.free.fr/. From there, one may click on the various parts of the site and on the links to other sites to gain a wealth of information on these subjects.

His lifter work, including successful replications and instructions on how to build lifters, can be accessed from http://jnaudin.free.fr/html/lifters.htm, and will keep the interested researcher busy for many productive hours. We especially commend Jean-Louis for his extraordinarily important efforts and for making all this work available to researchers worldwide. Once on the site, one is free to choose from a very wide variety of projects, subjects, etc. The information value of this site is very, very high. We consider Naudin's site to be the single best information site on the web for COP>1.0 system information and for the new space propulsion methodology and effort. Necessary links to other related sites of interest are also provided.

Good hunting!

8.3.16 Other Research Work and Inventions

While we have given a sort of quick overview of work done in this field, there have been many other patents filed by numerous inventors and additional work done by them. This is a dynamic field and it is moving; the biggest problem is that it is still sorely hampered by highly limited funding and an attitude by much of the scientific community that "if it has not been invented here, it doesn't exist".

In the table in Figure 8-3 below, we list some of the pertinent patent numbers for reference, with the emphasis on U.S. Patents. The Naudin website previously referenced has a great deal of information on these thrusters and lifters, including instructions for building and testing several of the simpler proof-of-principle versions. There are links to other sites with even more information. Naudin also includes the results of his successful replication and testing of a variety of these devices and inventions, as well as the successful replication by other experimenters. Naudin also gives construction instructions and information on some of these, so that independent researchers can build test units and test them if desired.

For those researchers interested in pursuing this area further, we would heartily recommend a determined search on the Internet, using Google (www.google.com) as their search engine. A good place to start is Jean-Louis Naudin's web site, and then run the cross-links. The combination will turn up quite an amount of specific material, from which the individual researcher can choose which approach and experimental method he wishes to pursue.

Figure 8-3.

Selected Inertial Propulsion System Patents

T. Townsend Brown, "A Method of and an Apparatus or Machine for producing Force or Motion, British Patent 300,311, Nov. 15, 1928.

Norman L. Dean, "System for Converting Rotary Motion into Unidirectional Motion," USP #2,886,976, May 19, 1959.

T. Townsend Brown, "Electrokinetic Apparatus," USP #2,949,550, Aug. 16, 1960.

T. Townsend Brown, "Electrokinetic Transducer," USP #3,018,394, Jan. 23, 1962.

T. Townsend Brown, "Electrokinetic Generator," USP #3,022,430, Feb. 20, 1962.

T. Townsend Brown, "Electrokinetic Apparatus," USP #3,187,206, Jun. 1,1965.

Henry Wallace, "Method and Apparatus for Generating a Secondary Gravitational Force Field," USP #3,626,605, Dec. 14, 1971.

Henry Wallace, "Method and Apparatus for Generating a Dynamic Force," USP #3,626,606, Dec. 14, 1971.

Fernand Estrade, "Device for Transforming Kinetic Energy," USP #3,807,244, Apr. 30, 1974.

Leo J. Novak, "Centrifugal Mechanical Device," USP #3,810,394, May 14, 1974.

Everett H. Benson, "Inertia Engine," USP #3,863,510, Feb. 4, 1975.

Oscar Mast, "Propulsion System," USP #3,889,543, Jun. 17, 1975.

Arthur N. Lehberger, "Centrifugal Propulsion Drive and Steering Mechanism," USP #3,897,692, Aug. 5, 1975.

Howard A. Gaberson, "Vibratory Locomotion Means," USP #3,916,704, Nov. 4, 1975.

Calvin I. Cuff, "Device for Converting Rotary Motion Into a Unidirectional Linear Motion," USP #3,968,700, Jul. 13, 1976.

Nicholas Joseph Schnur, "Method and Apparatus for Propelling an Object by an Unbalanced Centrifugal Force with Continuous Motion," USP #3,979,961, Sep. 14, 1976.

Calvin I. Cuff, "Device for Converting Rotary Motion Into a Unidirectional Linear Motion," USP #3,998,107, Dec. 21, 1976.

George Knap, "Orbital Propulsion Apparatus," USP #4,087,064, May 2, 1978.

Calvin I. Cuff, "Device for Converting Rotary Motion into Unidirectional Motion," USP #4,095,460, Jun. 20, 1978.

Robert L. Cook, "Device for Conversion of Centrifugal Force to Linear Force and Motion," USP #4,238,968, Dec. 16, 1980.

Ladislaw G. Srogi, "Mechanical Propulsion System," USP #4,242,918, Jan. 6, 1981.

Harry S. Melnick, "Unidirectional Force Generator," USP #4,261,212, Apr. 14, 1981.

Frederick L. Dehen, "Apparatus for Converting Rotary Motion to a Rectilinear Force," USP #4,347,752, Sep. 7, 1982.

Harry S. Melnick, "Mechanical Power Transmitting System," USP #4,398,431, Aug. 16, 1983.

Peter Kleber, "Apparatus for Acceleration-Free Mounting of a Body in a Spacecraft," USP #4,408,740, Oct. 11, 1983.

Fred de Weaver, III, "Propulsion System," USP #4,409,856, Oct. 18, 1983.

Ronald K. Linde, "Centrifugal Reaction Chamber," USP #4,450,141, May 22, 1984.

Fred de Weaver, **III,** "Propulsion System," USP #4,479,396, Oct. 30, 1984.

Joseph Colla, "Mechanical Propulsion System," USP #4,577,520, Mar. 25, 1986.

Elmer M. Dobos, "Propulsion Apparatus," USP #4,579,011, Apr. 1, 1986.

Brandson R. Thomson, "Apparatus for Developing a Propulsion Force," USP #4,631,971, Dec. 30, 1986.

James E. Cox, "Dipolar Force Field Propulsion System," USP #4,663,932, May 12, 1987.

Alvin C. Peppiatt and Alfred J. Peppiatt, "Impulse Drive," USP #4,674,583, Jun. 23, 1987.

Henry North, "Apparatus for Producing a Force," USP #4,712,439, Dec. 15, 1987.

Harry S. Melnick, "Continuous Force and Impact Generator," USP #4,726,241, Feb. 23, 1988.

Oscar F. A. Peterson, "Apparatus for Producing a Directional Unit Force," USP #4,744,259, May 17, 1988.

James D. Mundo, "Universal Propulsion Powerplant and Impulse Drive Unit for Self-Propelled Vehicles," USP #4,770,063, Sep. 13, 1988.

Lancelot I. Kethley, "Gyroscopic Propulsion Device," USP #4,784,006, Nov. 15, **1988.**

Charles Fulop, "Flywheel," USP #4,788,882, Dec. **6, 1988.**

Charles E. Rogers and Fred Van Arsdell, "System for Propulsion and Positioning of a Transitory Object," USP #4,801,111, Jan. 31, 1989.

Paul J. Montalbano, "Conversion of Rotational Output to Linear Force," USP #4,856,358, Aug. 15, 1989.

George J. Zachystal, "Device for Obtaining a Directional Centrifugal Force," USP #4,884,465, Dec. 5, 1989.

James E. Cox, "Dipole Accelerating Means and Method," USP #4,891,600, Jan. 2, 1990.

Hector Serrano, "Propulsion Device and Method Employing Electric Fields for Producing Thrust," WO 00/58623, Oct. 5, 2000.

Brian C. Motts, "Airship," USP #4,967,983, Nov. 6, 1990.

Lyle M. Mason, "Centripedal Device for Concentrating Centrifugal Force," USP #4,991,453, Feb. 12, 1991.

Alexander D. Kidd, "Gyroscopic Apparatus," USP #5,024,112, Jun. 18, 1991.

Paul J. Montalbano, "Conversion of Rotational Output to Linear Force — a Transmission," USP #4,042,313, Aug. 27, 1991.

Andrew T. Rodgers, "Controllable Gyroscopic Propulsion Apparatus," USP #5,054,331, Oct. 8, 1991.

Mortimer S. Delroy, "Gyrostat Propulsion System," USP #5,090,260, Feb. 25, 1992.

Kemal Butka, "Propulsion System," USP #5,111,087, May 5, 1992.

Rex L. Schlicher, Steven M. Rinaldi, David J. Hall, Peter M. Ranon, and Charles E. Davis, "Nonlinear Electromagnetic Propulsion System and Method," USP #5,142,861, Sep. 1, 1992.

Thomas L. Navarro and James D. Isaacson, "Apparatus with Rotatably-Driven Asymmetrically-Accelerated Eccentric Mass for Generating Translational Force," USP #5,150,626, Sep. 29, 1992.

Theordore R. Bristow, Jr., "Method and Apparatus for Converting Rotary Motion to Lineal Motion," USP #5,156,058, Oct. 20, 1992.

John C. McMahon, "Energy Transfer Device," USP #5,167,163, Dec. 1, 1992.

James W. Black, "Non-Linear Propulsion and Energy Conversion System," USP #5,182,958, Feb. 2, 1993.

James R. Taylor, "Electromagnetic Energy Propulsion System," USP #5,197,279, Mar. 30, 1993.

Michael J. Sohnly, "Magnetohydrodynamic Propulsion System," USP #5,211,006, May 18, 1993.

Ross C. Wood, "Stabilization System for a Freely Rotatable Platform," USP #5,256,942, Oct. 26, 1993.

James F. Woodward, "Method for Transiently Altering the Mass of Objects to Facilitate Their Transport or Change Their Stationary Apparent Weights," USP #5,280,864, Jan. 25, 1994.

Robert C. Willis, "Spaceship Propulsion by Momentum Transfer," USP #5,305,974, Apr. 26, 1994.

Harold L. Hull and Donald E. Joslin, "Reusable Mass Propulsion System," USP #5,313,851, May 24, 1994.

Kemal Butka, "Propulsion System," USP #5,334,060, Aug. 2, 1994.

James Harvey, "Impulse Converter," USP #5,335,561, Aug. 9, 1994.

Maurice Mitchell, "Net Kinetic Energy Differential Guidance and Propulsion System for Satellites and Space Vehicles," USP #5,377,936, Jan. 3, 1995.

Howard M. Woltering, "Rotating Eccentric Weights Vibrator System," USP #5,388,469, Feb. 14, 1995.

Richard O. Marsh, Jr., "Centrifugal Force Drive Machine," USP #5,388,470, Feb. 14, 1995.

Kemal Butka, "Propulsion System," USP #5,410,198, Apr. 25, 1995.

Thomas L. Navarro, "System for Generating Controllable Reference Environment and Steerable Translational Force

from Interaction Therewith," USP #5,473,957, Dec. 12, 1995.

Richard L. Lieurance, "Centrifugal Inertia Drive," USP #5,488,877, Feb. 6, 1996.

John C. Claxton, "Centripetally Impelled Vehicle," USP #5,557,988, Sep. 24, 1996.

Ezra Shimshi, "Apparatus for Energy Transformation and Conservation," USP #5,673,872, Oct. 7, 1997.

Richard E. Foster, Sr., "Inertial Propulsion Plus/Device and Engine," USP #5,685,196, Nov. 11, 1997.

James D. Booden, "Electromagnetically Actuated Thrust Generator," USP #5,782,134, Jul. 21, 1998.

George J. Howard, "Propulsion System," USP #5,791,188, Aug. 11, 1998.

Timothy J. Stopplecamp, "Boot Strap Propulsion System," USP #5,831,354, Nov. 3, 1998.

Eric Laithwaite and William Dawson, "Propulsion System," USP #5,860,317, Jan. 19, 1999.

Jonathan W. Campbell, "Apparatus and Method for Generating Thrust Using a Two Dimensional, Asymmetrical Capacitor Module," U. S. Patent # 6,317,310 Bl, Nov. 13, 2001. Assigned to U.S. Government.

8.3.17 Mechanism for the Antigravity Accelerating the Universe's Expansion

Finally, in addition to sorely needed better means of space propulsion, one of the real incentives for antigravity research is provided by the search for a source of the now-established mysterious antigravity that is accelerating the expansion of the universe. Elsewhere in this book we have advanced our own proposal for the mechanism generating that *antigravity* effect in the cosmos.

Simply put, a very violent astronomical process may be considered a giant and very violent discharge phenomena, since it first involves a giant excitation followed by a giant release (discharge) of that excitation. All such phenomena involve strong broken symmetry during the discharge itself, with the concomitant strong available of causal 4-sources of negative energy and negative energy fields. These processes — in their causal unobserved state - do produce locally strong negative energy EM fields from the source causal 4-positrons (Dirac sea holes prior to observation as positive mass 3-space positrons having positive energy and reverse-oriented). When the unaccounted and very large Heaviside nondiverged EM energy flow component accompanying every field and particle interaction is accounted, the causal negative "mass-energy" states produce strong antigravity fields, particularly considering the *lower frequencies* of many of the phenomena.

With the long-neglected Heaviside nondiverged energy component considered and not just the accounted Poynting diverged component, the total amount of negative EM field energy available in these violent process is dramatic. Accordingly, just as was the Sweet VTA in its increased power and its ELF pulsations, these giant astronomical sources are also giant negative "mass-energy" sources and strong antigravity sources as well. In short, such astronomical sources produce strong antigravity components repelling each other across the universe. Hence the observed acceleration of the expanding universe; the violent astronomical phenomena are continually "pulsing" the universe with antigravity and hence continually accelerating its expansion velocity.

In 2000, we also published a proposed mechanism for the source of the excess *positive* gravity holding the arms of the spiral galaxies intact {542}.

8.4 The Author's Engineering Approach to Local Antigravity

See again our discussion under Newton, subparagraph 8.2.1 above, and our discussion of a negative field energy extension to it. For a major laboratory, we believe that antigravity is straightforward,[227] if one first produces a COP»1.0 situation in a higher group symmetry electrodynamics process. This provides a very strong "displacement current" of causal 4-positrons (before parity inversion, time-reversal, etc.) as true negative energy "particles" or "states" or "Dirac holes" in the surrounding 4-vacuum. The actual "negative mass-energy field" is very large, due to the unaccounted Heaviside nondiverged negative energy component. Consequently, a substantial antigravity field is established.

---

[227] However, this is based on the assumption that the lab team first gives up some erroneous but widely held misconceptions, and really does a little "thinking outside the box" as they are prone to suggest.

8.4.1 Sweet Antigravity Test.

See again our discussion in paragraph 8.3.7. A practical realization of antigravity was clearly shown by the Sweet antigravity test designed by the present author. See previous discussion under paragraphs 6.3.2 in Chapter 6, and see again Figures 6-5 through 6-15.

In Chapter 6, the Sweet vacuum triode amplifier device shown in Figure 6-8 had a very large COP = 1,500,000 at 500 watts output with a 33 microamp, 10 volt operator's input. It had an even greater COP when pushed to 1,000 watts output because of only a slightly greater input.

The Sweet VTA used an induced powerful self-oscillation between the local vacuum and the barium nuclei in the barium ferrite magnets. This resulted in what is called a "kinetic" magnet (and one magnetized through its flat face). In a kinetic magnet, its nuclear magnetic field (in this case, in the barium nucleus) is self-oscillating (waving) in an angle about its local direction, as shown in Figure 6-6 in Chapter 6.

The conditioning was performed with the barium ferrite magnet itself in forced resonance at its resonant condition (some 11 to 12 Hz, in the case of Sweet's magnets). Then the 60 Hz signal was impressed orthogonally to the primary direction of the resonant magnet. After conditioning each of the two magnets, they were placed together on a separation box, carefully keeping the corresponding ends together similar to their original orientation during activation, but in attracting mode (through their face) across the separator. The assembly itself was then further conditioned, again orthogonally to the coupled oscillating fields. This "synchronized" the two magnetic self-oscillations and phase-locked them together in a common self-oscillation with the local vacuum.

Similar treatment of the vacuum as a semiconductor, together with self-oscillation between vacuum and nuclei, has been advanced independently by Prange and Strance {543}, who showed that the vacuum may be regarded as a semiconductor. In particular, they showed that the vacuum in the region close to the nucleus of a superheavy element is analogous to the inversion layer in a field effect transistor. It is similarly analogous for the barium atom, though Prange and Strance do not discuss that element. They introduced the concept of the inverted vacuum. Just as a semiconductor may be manipulated by subjecting it to external fields, doping etc., it appears that so can be the vacuum — along the lines shown by Prange and Strance and along the lines demonstrated by Sweet's conditioning of his kinetic magnets.

8.4.2 Bearden's Proposed Antigravity Mechanism

Any EM process that produces a COP»1.0 condition, such as can be done in intensely abrupt ionic discharges and other processes, will also produce intense causal (4-space) Dirac sea holes in the immediate surrounding vacuum. This is the negative energy mechanism, conceived by the author in the 1970s and finalized in the 1980s, for producing a negative energy causal (unobserved) EM field[228] in the immediate space surrounding a system. The overunity system must be the "source" of that negative energy EM field. So the negative energy EM field in surrounding local space is "attached" to the mass and energy dynamics of the "unit's supersystem" as being its "source or generatrix". Its attachment to the system mass produces the antigravitational force upon that mass.

Negative energy state charges (Dirac holes) are also *source charges.* Prior to observation converting them (as observed) to 3-space positive mass-energy positrons with positive energy and going the opposite direction, they must be accepted as negative energy source 4-electrons "pouring out negative energy" in all directions in 3-space, continuously, via the giant negentropy process for source charges. Since they pour out "negative" energy, we may consider that negative energy field action at any point in 3-space to be a continuous absorption of positive energy from 3-space and a return of that extracted energy to the time domain as increased time — in short, time dilation.

Either way one considers it, the 4-space Dirac holes do produce negative energy EM fields, potentials, and spacetime curvatures and these must be accounted during their existence *prior to observation and interaction.* Further, their accompanying giant Heaviside negative energy flow component must also be accounted. In other words, do not reverse the spatial direction and thus the location and direction of these fields, do not turn them into positive energy fields by such reversal of direction, do not turn the negative mass into positive mass, and do not reverse the sign of the source charge (the Dirac sea hole). All that happens after interaction of the Dirac sea holes with matter, forming the usual lattice holes or *observed* 3-positrons.

---

[228] We strongly stress that the continuous, dynamic, ongoing causal 4-field before it interacts with charge and is observed, is a totally different entity from the iteratively and continually observed frozen "effect" 3-field in matter after the 4-field has interacted with charged matter to produce an observation or "event". The great barrier to using electromagnetic fields in practical antigravity application has been due to the substituting, in electrodynamics, of the iterative "effect" 3-field of the observed positron for the continuous "causal" 4-field of the unobserved negative energy 4-electron (the Dirac sea hole). The "effect" 3-field is simply the iterative, continual energy gradient of the diverted EM field energy flow around the assumed intercepting unit charged particle.

Therein lies the magic secret of antigravity. Instead of thinking of the *fields* interacting on the craft or system that is producing them,[229] think of these negative energy fields interacting upon the *local spacetime surrounding the craft or system, and curving that spacetime negatively. Then* think of this negative curvature of spacetime interacting back upon the craft or system in general relativity fashion. *That* is what will and does produce practical antigravity upon the interacted craft or system, once the Heaviside energy flow component is accounted! But to understand its magnitude and the practicality, one must also account for the long neglected Heaviside nondiverged component.

The mechanism is also consistent with general relativity and Wheeler's succinct statement that mass acts on spacetime to curve it, and curved spacetime acts back on that mass to move it or produce a translation force upon it. We are simply applying that fact of general relativity to produce a local antigravity effect of a locally curved spacetime upon the source system (in this case, the power system itself, which is the source of the negative field energy in the vacuum.

The entire notion of "curvature in spacetime" — and indeed of spacetime itself— is a notion applying *before observation and interaction to produce the observed effect* — i.e., prior to the *d/dt* operation occurring upon an ongoing 4-space process to produce a momentary frozen 3-space snapshot (an observed effect). Think carefully. Wheeler's rule does not mean that 3-space acts back on the mass (actually on the masstime), but that 4-space (spacetime) acts back on it. Spacetime does not exist in the observed object after the application of observation's *d/dt* operation has occurred and until a new causal interaction begins. The *d/dt* operation strips away the time, and hence most of the energy (i.e., all the time-energy) of the spacetime that is to do the interacting.

---

[229] The ongoing interaction of the vacuum/spacetime with the source of the unobserved 4-fields is their cause, and the observed 3-fields are the effect. The effect *a priori* does not interact upon the cause! So thinking that "one should make a 3-field in a craft or system that will interact with it and produce antigravity force upon it" is a non sequitur. Instead, one more correctly thinks that "one should make a 4-field *from* a craft or system which will interact with the local spacetime to curve it negatively. Then that negative curvature of spacetime is itself a negative energy 4-field that will interact back upon the observed craft or observed 3-system to produce observed antigravity force upon it.

As source charges, the Dirac sea holes — before observation and in their vacuum state condition — produce a locally curved *spacetime* around the source unit. That induced *(negative energy x time)* curvature interacts back upon the physical system that is the source of the field producing the negative spacetime curvature. It specifically acts upon the mass of the 4-system when it exists as *(mass x time) form.* That interaction produces *antigravity* force and momentum upon the system, rather than *positive gravity* force and momentum[230]. Adding negative EM field energy (in the causal, unobserved state) is precisely the same as adding extra negative masstime, i.e., extra *(-m)t,* since the 4-field is being used in unobserved state.

Using the principle of superposition of fields[231], the external observer sees the "weight" of the unit (the force developed on it by the *net* gravitational field — see again our discussion of the Newtonian approach under paragraph 8.2.1 above) — as being reduced. The intensity of the *negative energy x time (negative angular momentum or negative action)* EM field (including with the Heaviside component accounted) produced by the unit determines the intensity of the antigravity produced, and hence determines the fraction of the unit's weight[232] that is observed to be "lost". See again Figures 6-14 and 6-15 in Chapter 6.

---

[230] Note that, if the Dirac hole interacts with the mass lattice of the system, it converts to a "hole" that is attached to the positive mass of an ion. Hence the "observed" positron is said to have "positive mass", which is a non sequitur for the state in which the "positron" exists prior to observation or interaction. The lattice hole 3-positron and the true spatiotemporal "4-positron as still a negative energy 4-electron'" are two very different entities, and their fields and concomitant field effects are very different.

[231] Note that identification of spacetime and field in general relativity extends the notion of superposition of multiple fields to the notion of superposition of multiple curvatures of spacetime. Thus, it is the *net* superposed curvature of spacetime that interacts back upon the physical mass system, to produce gravity (either positive or negative, depending on the sign of the net curvature).

[232] We remind the reader that the "weight" of an object on Earth is a measure of the force developed on the object by the (Earth's) gravitational field. "Loss of weight" is simply "reduction of that net force" due to the combination of the Earth's positive (attractive) gravitational field and the Dirac hole curvature of the object's local spacetime producing a negative (repulsion) gravitational field.

At sufficient intensity of this local negative energy 4-field in spacetime with the unit as its source, the back interaction upon the unit will balance the action of the earth's gravitational field. At that point, the unit will hover and "float" against the force of terrestrial gravity.[233] By further increasing the intensity of the negative energy EM field, the unit will accelerate upward, now being repelled by the local earth more than it is attracted by it. By "angling" the negative energy EM field direction — or by producing a second such negative energy EM field from the source unit as an "angled add on" superposed on the first one — one also provides unilateral propulsion in a given direction.

As can be seen, all the requirements for a practical antigravity vehicle and negative energy propulsion system are met. The only real problem in this approach is the requirement to first attain a COP»1.0 EM power system at ELF frequencies.[234] There is, however, an additional problem of correcting one's thinking to be along the lines indicated.

Antigravity becomes straightforward if one first does overunity discharges or steady output in a system with COP»1.0 and understands the difference between the observed "effect" 3-field and the unobserved "causal" 4-field. Sweet's VTA unit had a beginning COP = 1,500,000 at 500 watts output. The unit was pushed to nearly double its COP, by connecting additional load so that the output was now 1,000 watts. The additional input draw was minimal.

Designing the experiment with that in mind, I convinced Sweet to build a new output load unit, and to perform the experiment. In advance, I roughly predicted levitation of the 6-pound device at about 1500 watts output power, by some back-of-the-envelope estimates assuming a 10% efficient process. The projected curve (again see Figure 6-14 in Chapter 6) shows that levitation would have occurred at about 1250 watts or so.[235]

---

[233] Sweet later did push the VTA to hovering and upward acceleration, with the unit on a tether. After he was shot at by a distant sniper, thereafter he was very fearful **and** refused to show many experiments, including the antigravity experiment and particularly its extension to the hover point and upwards acceleration extension. In my opinion he had good reason for his paranoia and his fear for his life.

[234] The reason for ELF frequencies is simple: for the most powerful curvatures of spacetime, one wishes to use EM fields whose photons have the greatest *total* energy (including time-energy). That of course is the lower frequency photons, not the higher frequency photons. So ELF is the way to go, not gamma rays etc.

Here is why we advocate a Dirac sea hole's negative energy field as a practical antigravity inducing mechanism.[236] At COP»1.0, there is a very appreciable vacuum energy density difference between the operator's energy input and the energy output of the system, *particularly when the Heaviside nondiverged energyflow component is taken into account as well as the Poynting energyflow component?*[237] There is a production of negative EM mass-energy x time (causal Dirac sea holes) in the immediately surrounding vacuum. That represents the production of a causal, unobserved antigravitational field (curvature of spacetime) with the system-associated anticharge as its source.

Meanwhile, the Earth's causal, unobserved gravitational field is still normal and unchanged. Now the two causal fields superpose, and interact with the system mass. As seen by the external observer, the net effect is a reduction in the weight of the object — the effect of the interaction of the system mass with the two superposed causal gravitational fields simultaneously.

Reasoning that increasing the COP of the Sweet unit should increase the ratio of output negative energy to output positive energy even further, 1 calculated that doubling the output of the device would result in a sufficiently intense negative mass-energy field in the immediate space around the unit, to approach or reach levitation of the unit. However, as the increasing curvature of spacetime would further increase the deposition of magnetic monopoles in the magnets,[238] at some point the magnets would simply explode like hand grenades if the unit were pushed too far. With no explosive control facilities available, it appeared that doubling the output was the maximum that could safely be done under the circumstances. My own estimate showed that 1500 watts output would be required for levitation, but one dared not go beyond 1000 watts or one would risk a fragmentation explosion and injury or death.

---

[235] I strongly warned Floyd not to exceed a 1 KW load, no matter what. In the resulting local curved spacetime surrounding the barium nuclei in the magnets, it appears that magnetic monopoles are deposited or appear. Since these barium ferrite magnets were made by pressed powder metallurgy, they had little tensile or torsion strength, and under sufficient stress they would explode violently like a fragmentation hand grenade. In his little lab (a converted bedroom), Floyd had no explosive facility. He did explode several magnets, but while cautiously remaining some personal distance from the unit as its output was increased. We warn the reader that experimentation with such procedures is at his own risk, and all proper laboratory safety precautions must be utilized.

[236] A provisional patent application has now been filed on the process in 2002.

[237] In a nominal circuit, the unaccounted Heaviside energy flow may be up to 10^{13} times as great as the Poynting energy flow component. This is a completely unaccounted, vast EM energy flow accompanying every field and particle interaction. Hence one can understand why such an energy density for a macroscopic system with COP»1.0, is sufficient to achieve on a macroscopic scale what LaPointe is attempting on a microscopic scale. The total macroscopic scale energy density is so great that the required LaPointe microscopic energy density {535} is present or approached.

I convinced Sweet to build a 1,000-watt output load box, which simply contained ten sockets for ten 100-watt light bulbs. This was a modification to the 500-watt VTA unit shown in Figure 6-8 in Chapter 6. By successively installing the bulbs one by one, the load could be adjusted in 100-watt increments up to 1 kilowatt.

The foregoing is the reasoning behind how I designed the experiment, and the purpose was to show very substantial and smooth weight loss of the VTA unit itself. The experiment was performed in that 100-watt incremental manner, with highly successful results as shown in Figure 6-14 in Chapter 6.

With each 100-watt increase in power output past 500 watts, the fraction of negative energy produced was greater. Hence the unobserved 4-positron fields (Dirac sea hole negative energy fields) were stronger, because there were more unobserved 4-positrons. This meant that the local negative mass-energy curvature of spacetime was increased also.

In turn, this negative curvature of spacetime interacted back upon the mass of the VTA to add negative mass equivalent, or more simply, antigravity force. The observed weight of the unit was the sum of the normal gravitational weight (due to earth's gravity curvature of the local spacetime), and the abnormal antigravity weight (negative weight) due to the antigravity curvature of the local spacetime superposed. So as the fraction of negative curvature of spacetime increased, the net weight of the VTA was observed to decrease.

---

[238] Contrary to popular opinion, one does meet with effective magnetic monopoles in a magnetic field — including in the magnetic field of a permanent magnet — in a strongly curved local spacetime. The stronger the field, the greater the magnitude of the effective monopole charge.

As stated, Sweet was performing the experiment on the bench in California, and reading the instrumental results over the phone (I was in Huntsville, Alabama) where I recorded them. The spectacular results are shown in Figure 6-14 in Chapter 6.

The high COP (corresponds to gain) of energy density present in the Sweet VTA system's output section — with a greater percentage of it being negative energy — versus input energy one inputs to the input section, is what is important. In short, increasing the load effectively increased the gain, which further increased the ratio of the negative mass-energy to the positive mass-energy. This in turn steadily decreased the observed weight of the object.

Anything outside the system mass of that very high COP system will "see" the weight of the system mass as having decreased. Weight is just so much mass making so strong a *net* G field (spacetime curvature) in the surrounding spacetime. Reduce the *net* G-field that the system mass is normally making (by having it *simultaneously* make the additional antigravity field which then vectorially adds to the mass's normal G-field), and the external observer, external scales, etc. will see the object as having very much reduced its weight.

Positive spatial EM energy density of a system makes positive gravity (as seen by the external observer in the lab frame) because of the way it curves the immediately surrounding spacetime and the way that curved spacetime interacts back on the system's mass-energy. Negative spatial EM energy density makes negative gravity because it curves the immediate spacetime the opposite way. A system under those "high negative energy field" conditions will produce a great deal of antigravity, because the immediate spacetime surrounding is dramatically curved by the negative energy. That was my concept when designing the experiment, and it worked beautifully. So that is how the Sweet device reduced its weight on the bench by 90%, for a 1 kilowatt total output and minuscule input power {544}.

If one were dealing with the source in a flying vehicle system, there is another effect due to momentum. Note that force F is defined as $\mathbf{F} \equiv \partial p/\partial t = +\partial/\partial t\ (\mathrm{mv})$. That is in a positive energy density local spacetime. When an extra negative energy density is added to the local spacetime, then there are two forces, a new one being $\mathrm{F} \equiv -\ \partial/\partial t\ (\mathrm{mv})$ and the normal one being $\mathrm{F} \equiv +\partial/\partial t\ (\mathrm{mv})$. These forces are created by any change in momentum in space containing an extra increment of negative EM field energy.

If the two forces balance, then there is essentially no net centrifugal force produced on the object (or any occupants in it) in a turn. In theory, maneuvers that are far more powerful can be made, without undue G-stress to any occupants inside. This of course remains to be tested in future work and experiments, and so must be taken as a strong hypothesis at present. It could be tested, given availability of the proper COP»1.0 energy system or process on board a test vehicle. Laithwaite's demonstration of the dramatic reduction in force necessary to lift a rotating gyroscope compared to lifting the same heavy gyroscope when not rotating, may have been a partial test of this "inertial resistance decrease" effect.

The Sweet device was the only available COP»1.0 device which could have been used for the experiment. That is why I designed the experiment and convinced Sweet to perform it. Had it not worked, much of what I had worked out for more than a decade would have been falsified — which of course is one of the purposes of experiment: to validate or falsify a thesis. In this case, I breathed a great sigh of relief because — in my view — the spectacular results completely validated my approach to antigravity, and also much of my approach to COP>1.0 systems.

Sweet was never in a position to be able to go with this to the scientific community. The VTA was completely fouled legally, by various entangling agreements Sweet had made with different backers, without ever changing or resolving any previous agreement. Also, he realized that his life really was in danger should he try to take the system openly to the scientific community. Consequently, he never tried and in fact adamantly turned all further work away from the antigravity effects.

Aging and rather defenseless, Sweet also began playing all sorts of games to prevent having to have the project validated independently, even for its COP>1.0 capability. He was mortally fearful of being killed if he had it resoundingly validated. He was motivated by the necessity to survive and not be killed, and he firmly believed that any such attempt for full scientific validation and open publicity would absolutely guarantee his quick death. His fears appear to have been well founded.

His most carefully guarded secret was his activation process for the barium ferrite magnets, which he guarded with his life. He also began telling different persons all sorts of different stories and "mechanisms" for the activation, which were clever ruses and disinformation designed to throw them off the true trail and prevent his activation secret from being uncovered. I made a personal videotape of the activation process, or what he released of it, before he began such disinformation. So we do know much of the activation process, though not all of it.

Eventually, with changes in backers yet once again, Sweet died and never revealed the full secret of strong self-oscillation activation in permanent magnets to anyone. I knew and know part of it, even most of it, but still not all of it. What I understand about it has been briefly included in this book. From there, the interested researcher must find his own way.

The VTA was real, my gravitational mechanism is real, and — in my opinion — validated. However, it has not been *independently* validated in formal order, which is the proper scientific requirement. I did try very hard to convince Sweet that we should take this to the leaders of the scientific community and have it independently verified (e.g., by UCLA), but he would not hear of it after the assassination attempt. By that time, he was so afraid from the constant threats and constant stalkers that he had totally decided against any such releases. He also would not permit it to be taken to a major scientific conference and demonstrated, even though several scientists and engineers knowledgeable and experienced in measurement engineering did measure and certify the device's operation. But they were never allowed to see it in the antigravity mode. He would not discuss it with them, and he would not allow it.

Sadly, with the death of Sweet there passed away the VTA and its major secrets. I regard the VTA as one of the great scientific accomplishments of all time, and I regard Sweet as one of the most capable lab experimenters I ever met and one of the greatest inventors I ever met. The loss of the VTA and the absence of its full, independent scientific testing and validation were in my opinion two of the greatest losses to science of all time.

8.4.3 Discussion

Particle physicists speak of antimass and antigravitational field, but do not speak of anti-electromagnetic field and do not like negative mass. Antimass is actually "anti-mass-energy". When did one hear any scientist speak of a *negative energy* electromagnetic field or *negative energy* force field (except possibly as a hypothetical "tractor beam")? Or a *negative energy* electrical or magnetic potential? Yet the EM fields from an accumulation of Dirac sea holes (causal state) as a source charge are precisely negative energy EM fields. The EM potentials from an accumulation of Dirac sea holes are precisely negative energy potentials. Further, these fields and potentials produce opposite curvatures of spacetime as compared to positive energy causal EM fields and potentials from conventional source positrons and electrons.

Conventionally, scientists speak and think of *the positive energy* electromagnetic field. In our view, they do not apply the implications of the negative energy state part of the Dirac 1930 electron theory {545}, because of a built-in horror of negative mass (and antigravity!). Recall again our discussion of Laithwaite; a promising career was cut down in mid stride because he "mentioned the unmentionable": antigravity and possible failure of Newton's laws (they *do* fail in a curved spacetime!). His exhibiting to the Royal Society a very heavy gyroscope, very difficult to lift when placid, but easy to lift with one arm when spinning, was the final straw. The British scientific community simply cut him down and ostracized him from the rewards and positions he would otherwise have been given.

To conceal the relationship of the negative energy fields, hide negative mass, and pretend it does not exist, much scientific effort is exerted to consider the positive energy, positive mass positron *only after it has been observed.* So scientists can assign this "observed" 3-positron a positive mass and positive energy but with spatial direction reversal. That immediately moves to an *effect* field and an *effect* positron, not a causal field or a causal Dirac sea hole. It effectively eliminates the antigravity field of the negative mass of the unobserved negative energy Dirac electron, which is produced by the *causal* (unobserved) Dirac sea hole's electromagnetic field, not the *effect* (observed) positron field which then is "seen" to have positive mass but with field direction reversed.[239]

The positron is quite a different beast before it is observed, just as is a mass (and just as is any other observable, and just as is a field). For one thing, the positron before observation and the positron after observation have different dimensionalities; the unobserved causal entity is 4-spatial and continuous,[240] while the observed effect entity is continually and iteratively 3-spatial. Mass before it is observed exists as masstime, not mass.

---

[239] Note also that the causal field, being unobserved, is a 4-field, while the observed (effect) field is a 3-field. By reversing the parity, one has made the antigravity field **into** a gravity field, by substituting the effect for the cause and erroneously using the old non sequitur of a separate 3-force acting upon a separate 3-mass. Also, when one reverses its spatial direction, one converts a *positive* gravity field into an *antigravity* **field.**

[240] Along with Bohm, we point out that — strictly speaking — in 4-space one does not have "things", but only "processes". To speak of a "4-space field" is to redefine the field as an ongoing process, rather than as an observed frozen 3-snapshot or "3-thing". The solution to the long-vexing duality problem is there, but we leave its solution to the reader. It is necessary to extend Aristotelian logic itself if one is to resolve the duality problem.

Simply consider the delayed choice two-slit experiment and Wheeler's famous comment that no phenomenon is a phenomenon until it is an observed phenomenon. Once the observation occurs, one can even change the "if it had been observed but was not" previous past. One gets the magic from the delayed choice two slit experiment without the observation, and then by making the observation such that the "magic" is not revoked *after the fact.* The whole experiment proves that, if one insists on applying the observation, one kills the magic, even if one thinks the magic has already occurred and must have already occurred, but it has not been *observed* to have occurred.

Causal 4-spatial "things" are not things at all, but are ongoing processes and their cessations have not "occurred" to yet produce a "thing"! The *observed* event is 3-spatial and has occurred. It is something in the past, not the present. That is different from something ongoing (happening or unhappening)[241] in 4-space, before having been observed and therefore a frozen snapshot taken in 3-space. As is well known in quantum mechanics, observation is a physical interaction and it indelibly changes and stops the "ongoing 4-process", converting it to a momentarily frozen "3-thing" as an effect of the observation process at that particular instant. So any "thing" is continual and iterative, *not continuous,* in its travel through time, and in its "observed existence in 3-space".

Because of such, physicists also have difficulty realizing that mass — being an observable — does not and cannot exist continuously in time. It cannot even *continuously exist!* No observable can. Masstime exists in time, but mass does not, as we previously discussed. Observation is a *dldt* operator imposed upon an ongoing 4-space process, destroying the time portion and producing an instantaneous, frozen snapshot at that instant. It's just *d/dt* (LLLT) => LLL. That is why all observation is 3-spatial, as is well known in physics. We long ago pointed out how the common photon interaction — the total photon interactions (both virtual and observable) with a mass — is what drives the mass through time, so to speak {15}. But it does it iteratively, frozen snap after frozen snap —just like progressively showing the frames of a movie film.

---

[241] When a frozen instant 3-space snapshot (observable) interacts with a second causal process to generate another frozen snapshot (observable), the first observation "unhappens" because it is thereby moved into the "past", disappearing from 3-space forever. With every change of an observable (with every observation), even if the new effect appears identical to the previous one, it is still a *new* observable (a new effect) where a former effect "seems" to have recurred (been created, happened). Even if it is spatially the same as the previous observation output, it is at a different point in time and hence is the result of the *d/dt* observation operator being iteratively applied. The serial outputs — even of what we loosely call an "entity existing in time with no 3-space change" — are *successive and different* serial outputs from a serial time-differentiating process.

What really has to be changed is the old notion by Heaviside and Poynting that 3-energy propagates continuously through 3-space.[242] No observable can propagate continuously in 3-space, but only in 4-space and then only during its nonobservable state! So a "mass propagating through time" (i.e., just sitting still and "existing", so to speak) is doing so in an iterative fashion by recurring iteratively in observation output at the same spatial position. The mass turns to masstime, then to mass, then to masstime, etc. Or $m_1 \Rightarrow m_1t \Rightarrow m_2 \Rightarrow m_2t \Rightarrow m_3\ldots$ etc. In short, *observation continually recurs at an enormous rate,* so *an observable (effect or output of that observation process) continually recurs at an enormous rate,* destroying the previous observation as the next one occurs.

Because of the history of thinking in 3-space terms rather than 4-space terms, one of the ubiquitous errors widespread in physics is the substitution of the effect (the frozen output of the observation process) for the cause (that 4-space initiator process which continuously exists prior to observation, hence unobserved). So causes (4-space initiators which are unobservable a priori) interact with a previous "frozen 3-space snapshot" or observable, to produce a change (either a temporal change or spatial change or both) in or to or of that observable (that 3-snapshot) to a new observable (3-snapshot). If the object is seen to iteratively change in time but not position or form, it is said to "exist" or "persist" at rest continuously in time. It doesn't. It *continually recurs*[243] in that form, but does not *continuously exist* in that form!

---

[242] And also the hoary old mistaken notion that the "same observable thing" can continuously exist in time or persist.

[243] More rigorously, the continual recurrence is at the quantum level. So a larger mass or other observable actually is a great horde of very small "continual recurrences" of tiny incremental pieces of itself. At any time interval of observation via observable photons, a vast number of nonobservable (virtual) "continual recurrences" occur in the "observed object". In short, the "observed object" is the **end** summation and result of a vast number of infinitesimal changes. Hence the entry of statistics and the uncertainty principle. Each one of the infinitesimal

The antigravity mechanism has been implicitly present in the Dirac theory of the electron since his 1930 paper {497}, where he brought forth the Dirac sea and the anti-electron, if the *causal* Dirac sea negative energy states are applied to general relativity before they interact with charge and are "observed". General relativity deals with spacetime and spacetime curvature. Therefore, it is concerned with causal 4-fields, 4-potentials, and 4-waves in 4-space prior to observation.

As we stated previously, there is really no such thing as a "3-space EM wave", e.g., as rather strongly pointed out by Romer {546}.

One can produce and have an EM field of *negative* EM energy in spacetime, coming from the 4-positron charges of the anti-charged mass of a system, as well as one can produce and have an EM field *of positive* EM energy in space, coming from the charged mass of a system. But negative energy EM fields have been swept away from our minds by interpreting the negative energy 4-electron as the 3-positron, after the negative energy 4-electron has interacted with charged matter (been observed as the *effect),* rather than how it exists in spacetime prior to observation (as the *cause).*

So physicists reverse the direction and the charge, and make the mass positive, which is how the Dirac sea hole is "seen" *after observation.* That makes it a material lattice hole and a special kind of positive ion. Before observation, it is not seen at all, but it exists as negative action (energy x time) and therefore in "negative masstime" state form. Before it has been observed, it's still there in spacetime, as a curvature of spacetime. It's just a *negative energy state curvature* in the vacuum/spacetime.

Make lots of these negative mass-energy state Dirac holes from a source system, such as one producing a very powerful negative energy EM field, and voila! That's it. The system is producing — and surrounding itself by — its own negative energy EM field (negative curvature of 4-space), which is also its own local antigravity field. That antigravity field superposes with the prior local curvature of spacetime, altering it. The *net* curvature of local spacetime is itself a field which interacts back upon the mass system to produce "net gravity" whose sign may be positive or negative, depending on the sign of the net spacetime curvature. If the two back-acting curvatures of spacetime are opposite and equal, that represents the source object existing in a zero gravitational field, and thus "hovering" in the Earth's gravitational field, for example.

changes may be modeled causally, but since there is no individual information available on each one of them, they can only be calculated statistically.

The use of negative energy EM fields and Dirac sea holes to alter the curvature of the local spacetime in a negative direction provides the present author's engineering approach to practical antigravity. Note that our approach gathers in the gist of both of the major two approaches to antigravity: (1) the shielding theory as typified by Modanese {547}, and (2) the opposing force theory as typified by Ling Ni and Torr {548}. The opposing force, considered as a field that is identically a curvature of spacetime, is also a "gravity shielding" force via superposition of fields. We state unequivocally that a theoretical model of the new approach can be constructed from Sachs's theory {510}, and implemented in the 0(3) electrodynamics of Evans and Vigier {549}. Further, it is a directly engineerable theory, without the necessity for superconductivity, spinning wheels, etc. as demonstrated by the highly successful Sweet experiment previously discussed.

We quote Evans {550}, on the appearance of the B(3) field in general relativity, to indicate the nature of such a future theoretical model:

> *"I have chosen a metric that gives circular polarization (as observed), but in the framework of general relativity using the irreps of the Einstein group according to the Sachs theory. This is a closed field theory which is analytical, using a Lie group, the Einstein group. As such there are no particles, only spacetime curvature, so everything in physics reduces to spacetime curvature, i. e. to a metric vector and its components. This is then incorporated in the Sachs theory to produce a metric four-vector which is a generally covariant component of a quaternion-valued metric with sixteen components, the most general form of the metric allowed by general relativity and topology. The Einstein-Sachs equations for electromagnetism are six metrical equations in six unknowns, the components of an antisymmetric field tensor of electromagnetism. By choosing a metric that corresponds to circular polarization, $B^{(3)}$ appears automatically.*
>
> *This method gives the $B^{(3)}$ field from Einstein's theory of general relativity applied to electromagnetism by a consideration of the irreps of the Einstein group. It follows that whenever the conjugate product $-gA^{(1)} \times A^{(2)}$ is observed, the $B^{(3)}$ field must also be observed, otherwise general relativity fails. Examples of*

> *the $B^{(3)}$ field at work are the inverse Faraday effect and RFR [Radiatively induced fermion resonance]. "*

Evans and the AIAS have already taken the first theoretical steps showing that antigravity does indeed arise from Sachs's theory {551a-551c}.

We are convinced that practical antigravity only occurs at extremely low frequency (ELF) for reasons[244] explained previously. Certainly that seems to be true for our approach to it. The highest "total energy" photon is the low frequency photon with low spatial energy and a larger time increment. The larger time component has the same energy density as mass. So one multiplies the time component increase (in seconds) by roughly $9x10^{16}$ to get the increase in spatial energy equivalency. Halve the frequency of a photon, and one increases its *total energy* by $4.5x10^{16}$ {552}.

High energy physicists have not realized that the time component of the photon has enormously more energy than the spatial component, and that low frequency photons contain enormously more trapped energy than do high frequency photons. So they largely ignore the time-energy component, focus on the spatial energy component, and erroneously consider *high energy* physics as *high frequency* photon physics. Instead, present high frequency photon physics is only a high *spatial* energy physics, not a high *total* energy physics at all. Today's high energy physics is really still a very low total energy physics, compared to what can be done with the time-energy component of the photon rather than its spatial-energy component. By transducing and using the highly compressed time-energy, experiments can be done on the lab bench with fairly simple and cheap equipment, that actually utilize *higher total energy nuclear reactions* (though at low spatial energy) than are presently obtained in the largest accelerators.

That is also why there is presently no understanding in the conventional community of the "low spatial energy transmutations" of cold fusion, where the experiments unwittingly use the high time-energy component. The scientific community has not even recognized the time-energy aspects, and has not recognized that time-energy physics is an ultrahigh total energy physics, where the present "high energy physics" is tame by comparison.

---

[244] Dramatic increase in the magnitude of the time-component, hence an enormous increase in the time-energy component with only a small decrease in the spatial energy component total energy of the photon. The *total energy* thus dramatically increases, when the spatial energy equivalent of the time-energy is considered. The present conservation of energy law in physics is inadequate because it does not account time energy at all; thus completely discards the most fundamental EM energy mechanism of all: the giant negentropy of source charges and source dipoles.

The previously unknown time-energy induced transmutation interactions are the precursors to, and the opening salvo of, a new *ultrahigh total energy physics* struggling to be born in spite of the adamant opposition of the conventional scientific community. Allowing for the time-energy component, for the mechanism generating the flow of time, and for formation and decay of time-reversal zones, we previously wrote the exact new nuclear reactions for the production of the excess deuterium, tritium, and alpha particles {553} in numerous cold fusion experiments. We also explained the highly anomalous nuclear instrument effects — due to (i) time-charging, (ii) the slow decay of time-charge with radiation of Whittaker longitudinal waves, each accompanied by its phase conjugate twin time-polarized wave, and (iii) the previous time-charge history of the instrument. These puzzling anomalies have been observed at U.S. Navy research facilities at China Lake now for some years in rigorous electrolyte experiments {198}. We summarize that work in Chapter 10 of this book.

With a little effort and development, practical antigravity appears to be straightforward for any good modern laboratory. But the lab physicists should first (i) rethink the entire subject, from its very fundamentals, to include considering the positron before observation and not after observation, (ii) remove the Lorentz symmetrical regauging from electrodynamics and recover the presently missing COP >1.0 and COP»1.0 Maxwellian systems again, and (iii) rapidly develop COP»1.0 systems and processes to use in antigravity experiments and development. At least in this approach, the COP»1.0 EM systems — once developed — are an immediate gateway to practical antigravity devices, unilateral thrust propulsion systems, and antigravity vehicles. And of course they are the immediate gateway to practical transport vehicle power systems, fuel-free, clean, and self-powering.

Hopefully, such a scientific reassessment will be performed, and it may provide the gateway for practical transportation and travel to the ends of **the** solar system and beyond, to be achieved in the first two decades of this century.

8.5 In Conclusion

We have briefly reviewed the presently developing field of inertial space propulsion and antigravity, and presented our own approach to practical antigravity as well as an informal experiment that strongly supported that approach. We have strongly stressed that the causal field from the system as its source constitutes a curvature of the local spacetime superposed upon the pre-existing local spacetime curvature. Instead of thinking of that causal field from the source system as acting back upon the source system, we have stressed thinking of that *net superposed curvature of local spacetime* acting back upon the source system and producing a net gravitational force upon it. That net gravitational force may be either positive or negative, or zero, depending upon the sign of the net curvature of spacetime that is interacting back upon the system.

We presented and explained the results of the Sweet VTA antigravity experiment as probable verification of our approach to antigravity. Nonetheless, we also emphasized that the formal independent replication required by the scientific method has not been accomplished, and it remains to be accomplished before the approach is to be considered as scientifically validated.

We believe that this approach to antigravity can be modeled and engineered by use of the Evans-Vigier O(3) electrodynamics as a subset of the Sachs electrodynamics that is a part of the Sachs unified field theory.

If so, then tentatively we have laid out the basis for eventual practical antigravity systems and a practical antigravity technology. We fervently hope that this approach will in fact lead to practical technology and working transportation means during the next decade.

Finally, we hope that these notes and concepts will stimulate many sharp young graduate students and post-doctoral scientists to examine, model, correct, and greatly enlarge the ideas and principles advanced. If so, then the purpose of this chapter will have been realized.

Chapter 9

The Supersystem and Remarks on Gravity, Antigravity, and Testing

"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert and featureless, but alive with throbbing energy and vitality. A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea of ghost particles of all varieties - virtual leptons, quarks, and messengers, entangled in a complex melee. The presence of the electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is not at rest: it is being continually assaulted by all manner of other particles from the vacuum." [Paul Davies] {554}.

"Space acts on matter, telling it how to move.. In turn, matter reacts back on space, telling it how to curve." [Wheeler's principle] {555}.

"...the Aharonov-Bohm effect is a local gauge transformation of the true vacuum... [which] produces a vector potential from the true vacuum. [This gauge transformation] produces topological charge..., the electromagnetic field, which carries energy, and the vacuum charge current density first proposed by Lehnert... and developed by Lehnert and Roy... ". [M. W. Evans and S. Jeffers] {556}.

"The energy density (En_d) in curved spacetime

is given in the Sachs theory by the quaternion-valued expression

$$En_d = A_m j_m^* \qquad [1]$$

where A_μ *is the quaternion-valued vector potential and* j_m^* *is the quaternion-valued 4-current as given by Sachs ...Equation [1] is an elegant and deeply meaningful expression of the fact that electromagnetic energy density is available from curved spacetime under all conditions; the distinction between field and matter is lost, and the concepts of "point charge" and "point mass" are not present in the theory, as these two concepts represent infinities of the closed-field theory developed by Sachs from the irreducible representations of the Einstein group. The accuracy of expression [1] has been tested to the precision of the Lamb shifts in the hydrogen atom without using renormalization of infinities. The Lamb shifts can therefore be viewed as the results of electromagnetic energy from curved spacetime."* [Myron W. Evans] {557}.

"Old habits of thought die hard, and it is not always easy to remember that a particle with negative mass if pushed to the right will move to the left; or that in a given reference frame a particle with imaginary mass can have infinite speed, in which case its energy is zero." .... "... extremely powerful energy sources may occur if particles of negative mass really exist in nature." [Banesh Hoffman] {558}.

"The observed (3-spatial) positron is a quite different entity from the unobserved (4-spatial) Dirac hole. Its direction has been reversed, and its mass and energy have also been reversed. Before the 4-spatial hole in the vacuum has interacted and been observed, it has negative energy and therefore negative mass equivalency, as well as negative charge. As a source charge, it also is the source of negative energy fields and negative energy potentials — all in the

nonobservable 4-state. " [T. E. Bearden, private communication to a correspondent, 2001].

"The...visible structure of orthogonal' 1-networks possessing coupled open-paths and closed-paths represents only an introductory step in the development of a correct topological theory for conventional electric networks...An electric network is like an iceberg...Its visible portion is only a small fraction of the submerged invisible structure. " [Gabriel Kron, who developed special tensor methods and diakoptics to deal with the large number of visible and invisible multidimensional p-networks surrounding the branches of every electric circuit.] {559}.

"...symmetry implies conservation. Since our entire edifice of interactions is built on symmetry assumptions, there should be as a result a large number of conservation laws. The only trouble is that almost all of these conservation laws have been violated experimentally. "... "...this difficulty could be resolved by introducing a new element, the vacuum. Instead of saying that the symmetry of all matter is being violated, we suggest that all conservation laws must take both matter and vacuum into account. If we include matter together with vacuum, then an overall symmetry could be restored" [T. D. Lee] {560}.

"One perhaps oversimplified way to deal more conventionally with the new unified field approach to electrodynamic systems is to utilize the notion of the 'supersystem', comprised of three components: (I) the ordinary EM system and its dynamics—conventionally assuming flat spacetime and inert vacuum, (2) the local nonlinear vacuum and its dynamic interactions with the system and with spacetime, and (3) the local curvatures of spacetime and their dynamic interactions with the system and with the local

> *active vacuum. In the rich feedforward and feedback interactions that result, there emerges a dramatic new physics, new electrodynamics, and new chemistry including a new biochemistry. This is in fact a much simplified conceptual approach (good to 'first order', so to speak) to the multinetwork and multipath work of Gabriel Kron. It also allows the retention of symmetry, considered so necessary to particle physicists. "*
> [T. E. Bearden, private communication to a correspondent, 2001].

9.1 The Supersystem and Its Considerations

In classical electrodynamics, the Maxwell-Heaviside-Lorentz equations implicitly assume a local flat spacetime and no net action by the vacuum {561}. Both these implied assumptions are false. All real EM systems involve interactions with both curvatures of spacetime and the active vacuum. Yet to try to include both additional interactions requires a unified field theory approach, such as that of Sachs {562}. Unfortunately, any unified field theory is replete with quite difficult mathematics, little of which can be illustrated pictorially or simply.

What is required — at least for most electrical engineers and applied electrical scientists — is a conceptual way of visually thinking about the EM system in a way that is consistent with unified field theory to something like first order, but which can also be conceptualized easily.

9.1.1 The Supersystem Concept

Accordingly, we introduce a new and very simple way of thinking of electrical circuits in terms of a unified field theory. We utilize the notion of the supersystem, consisting of three components: (1) the EM system and its dynamics, considered in conventional fashion (as if the local vacuum were inactive and the local spacetime were flat but perturbed), (2) the local interactive vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. All three components of the supersystem interact. This produces both feedback and feedforward functioning; e.g., energetic changes in the system produce corresponding changes in spacetime curvatures. In turn, the alterations in spacetime curvature produce nonlinear changes in the active vacuum, which in turn acts back upon the system. So every action in a system produces a corresponding set of feedforward and feedback reactions from and with its complex environment, and vice versa.[245]

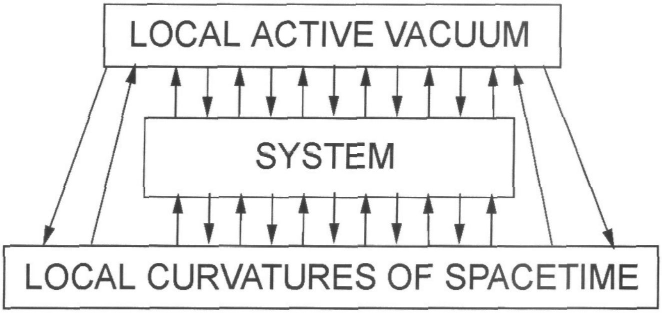

Figure 9-1 The Supersystem concept (block diagram).

Figure 9-1 shows a diagrammatic illustration of the supersystem concept. Figure 9-2 illustrates the effect of the classical EM assumptions of a flat spacetime and a locally inert vacuum. As can be seen, the classical assumptions *arbitrarily discard the* interaction effects upon and in the system due to the active vacuum dynamics and due to the local spacetime curvatures and their dynamics.[246]

As stated, all three components of the supersystem interact with each other. This is a dramatic change to classical U(1) electrodynamics. Hence the total interactions — with the system — of both the active vacuum and the locally curved spacetime, and their interactions with each other, are more general than either classical EM, general relativity, or the quantum mechanical vacuum of particle physics prescribes individually.

---

[245] We hypothesize that the feedback contains the mechanism generating Newton's third law as a restoration of symmetry, but leave that to the master theorists to explore and decide.

[246] As one example, the nondiverged Heaviside component of energy flow — previously discussed in earlier chapters — is part of the supersystem activity. This **flow** alters both the local active vacuum and the local spacetime curvature associated with an electrical power system and its connected external circuits. Since the resulting local "environmental" changes in the supersystem are extensive, then one can conceive ways in which these changes can be interacted with additional intercepting/receiving systems to produce extra "EM energy from the vacuum".

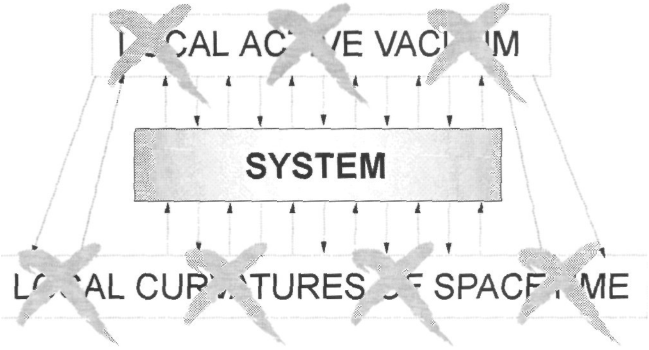

Figure 9-2 Classical EM destroys two-thirds of the supersystem.

Immediately it can be seen that the concept of energy from the vacuum is a "supersystem active and not negligible" concept involving a unified field theory approach *a priori.* It can also be seen that a "system" in equilibrium must be in equilibrium with its active environmental exchange — its exchange with the active local vacuum and the active local curvatures of spacetime. Since energy "exchange" with the system means both incoming and outgoing exchange, it follows that the equilibrium system has equal incoming and outgoing energy exchange. It follows that the symmetrically regauged system has equal incoming and outgoing energy, and so the symmetrically regauged system must be in additional stress from two net equal and opposite direct interaction forces, one general relativistic and the other from the active vacuum. Further, this "equal and opposite" equilibrium stress state must be accompanied by equal and opposite interaction effects between the two environmental media of the system.[247,248]

[247] We point out but do not pursue the fact that this is a very important new concept associated with *coherently* changing the symmetrical regauging, which allows a new physics, a new electrodynamics, and a new chemistry including biochemistry. For one thing, it permits such things as a self-maintained excitation state that simultaneously continuously discharges its excitation energy. This is simply a redefinition and more general broadening of the "source charge" concept. From the normal system view (assuming away the vacuum and curved spacetime interactions), an "impossible" system state can nonetheless be possible and stabilized, by changing the *intensity* of the Lorentz regauging condition. As an example, new molecules or ions can be made and stabilized in an excited state, that otherwise would be so unstable that the excitation would decay in a nanosecond with the state disappearing also. We believe this alone will lead to a great new biochemistry and medical

9.1.2 Time and Space Effects and Positive and Negative Energy

From curved spacetime considerations, we must include the effects on the system, and on the vacuum, of dynamics in the *time* domain as well as in the *spatial* domain. Indeed, in Sachs's theory space and time do not separately exist, but only spacetime. We consider time as highly compressed EM energy — as heretofore pointed out, where time is spatial EM energy compressed by the factor c^2 and having energy density similar to mass. We must include not only EM *energyflow* along the time-domain (fourth Minkowski axis *ict),* but also EM *force, potential, energy oscillations*, etc. in the time domain itself. We shall have to deal with time-force, time-energy, time-potential, time-energy current, longitudinal EM waves in the time domain (time-polarized EM waves), etc. {563}. The complexity of the full supersystem analysis is thus easily seen and appreciated. In most cases, a higher group symmetry model for a particular situation will prove intractable for closed solutions, and numerical methods must be used.

Further, when curved spacetime interacts with the active vacuum, we shall also have to consider *negative energy* currents in the nonlinear vacuum itself, as well as in the time-domain. We may also speak of the *negative energy density* in the vacuum as a special kind of negative energy *potential* in the vacuum that is created by a negative energy potential in the time domain. We may speak of a gradient in this vacuum negative energy potential as a negative energy force or "force in the negative energy domain". Applied to the time domain, these concepts constitute a causality condition placed on the Dirac sea, in the same vein as more technical causality considerations given by Finster {564}.

9.2 Differences between Effects in COP»1.0 and COP<1.0 EM Systems

In this chapter, we utilize the Dirac hole in a different manner from the received interpretation. The received interpretation assumes an electron hole in the vacuum, with the hole filled with a negative energy electron

---

therapeutic science; one company is already working in this area, with positive results.

[248] We leave to the advanced theorists the connection between this "stable equilibrium state under a change in symmetrically regauged stress", occurring as a *net-force-free* intensity change in the conventional Lorentz symmetrical regauging of the Maxwell-Heaviside equations.

having negative mass, negative energy, and negative charge. In that interpretation, when the electron is lifted out of the hole, the hole is said to be *a positive energy* electron with positive mass and positive charge, (i.e., when interacted and observed). In short, with reaction one has a complete transformation of negative energy into positive energy, negative mass into positive mass, etc. In our view, this totally destroys the notion of "how the hole exists in the vacuum prior to its interaction and observation". That kind of positron only occurs after a physical interaction and observation has occurred — in short, *the positron exists only after an interaction with the Dirac hole has been made, to change the very nature of the negative energy hole.* The positron is the 3-effect of an invoked extra 4-interaction, and is not at all the "negative energy electron" or hole as it actually exists prior to being interacted and transformed. The positron is a different beast from the beast we wish to harness and use. The conventional interpretation leaves the question of the unobserved and unreacted "hole" left behind by an electron lifted out of the hole, before another interaction is performed on the hole to change it. So it is an incomplete accounting (as all observations are).

In our interpretation, the Dirac "hole" in the vacuum is itself the true negative energy electron before its interaction and observation. It is the "causal" 4-space negative energy electron, while the received interpretation is only for the "observed" or "effect" positron. In our view, the received interpretation thus leaves out a goodly part (and the most interesting part) of the Dirac sea hole. When the negative energy hole is filled, we consider the *electron* in the hole to be a normal, positive energy, positive mass electron. Hence the summation of the energy of the filled hole is zero, as is its summation mass and its summation charge. This accords with the properties of the "normal" vacuum, said to contain a sea of such "filled Dirac holes".

Our purpose is to use *emptied* Dirac sea holes before their interaction or observation, and thus to use them as *causal 4-electrons having negative mass and negative energy, but where each of those quantities is multiplied by time. So the hole has negative masstime and negative action.* Unless otherwise stated, that is our interpretation of a Dirac sea hole, and that will be how we use it. This way we avoid having to replace a true negative energy, negative mass, positive charge causal (unobserved) 4-electron by a positive energy, positive mass, positive charge observed (effect) 3-electron. We do not just artificially "dispose of" the question of the "nature of the hole that is left behind" when the positron is observed. We consider the (hole x time) as the fundamental negative energy 4-electron. We will sometimes refer to that as the "4-positron".

Normal "forward flow of time" tempic force

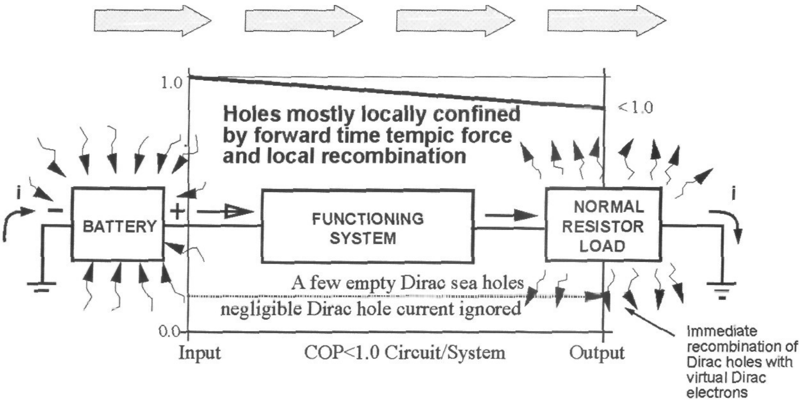
1.0
<1.0
Holes mostly locally confined by forward time tempic force and local recombination
i
BATTERY
FUNCTIONING SYSTEM
NORMAL RESISTOR LOAD
i
A few empty Dirac sea holes
negligible Dirac hole current ignored
0.0
Input
COP<1.0 Circuit/System
Output
Immediate recombination of Dirac holes with virtual Dirac electrons

Figure 9-3 Time potential gradient across a C0P<1.0 EM system.

See Figure 9-3. In a COP<1.0 EM system, the tempic potential gradient (tempic force) is oriented from input to output and on out beyond the system. A very slight Dirac hole current moves outward from the output, but is immediately "eaten" by the virtual electron flux fluctuations of the vacuum in the space immediately adjacent to the output. There is a slight Dirac hole current also exiting the system in its losses, but again these are immediately "eaten" by the virtual electron flux fluctuations of the vacuum in the immediately surrounding space.

No excess tempic force

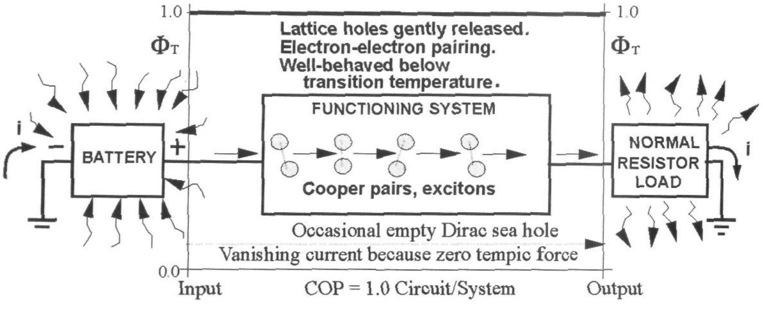
1.0
Φ_T
Lattice holes gently released.
Electron-electron pairing.
Well-behaved below transition temperature.
1.0
Φ_T
i
BATTERY
FUNCTIONING SYSTEM
Cooper pairs, excitons
NORMAL RESISTOR LOAD
i
Occasional empty Dirac sea hole
Vanishing current because zero tempic force
0.0
Input
COP = 1.0 Circuit/System
Output

Figure 9-4 Time potential gradient across a COP = 1.0 EM system.

See Figure 9.4. In a superconductive system or section, COP =1.0 and on the average there is no net gradient of the time potential across the system, hence no time-force across it in either direction. Accordingly, one sees ready Cooper pairing of charges and formation of excitons. We shall not concern ourselves further with the superconductive EM system; such a system has not yet produced COP>1.0.

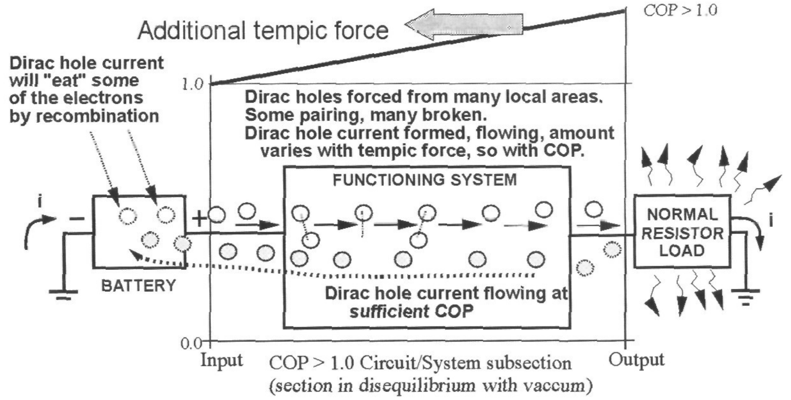
COP > 1.0
Additional tempic force
Dirac hole current will "eat" some of the electrons by recombination
1.0
Dirac holes forced from many local areas. Some pairing, many broken. Dirac hole current formed, flowing, amount varies with tempic force, so with COP.
FUNCTIONING SYSTEM
i
-
+
BATTERY
NORMAL RESISTOR LOAD
i
Dirac hole current flowing at sufficient COP
0.0
Input
COP > 1.0 Circuit/System subsection (section in disequilibrium with vaccum)
Output

Figure 9-5 Time potential gradient across a COP»1.0 EM system.

See Figure 9.5. In a COP>1.0 EM system, there is a net tempic force across the system from the output section to the input section. However, unless the COP»1.0, the hole current in the local vacuum is mostly used up in interacting with the system material and converting to lattice holes (lattice positrons). Hence at "nominal" or moderate COP>1.0, except in fast pulsed discharges one *usually* has little to worry about, concerning Dirac sea holes in the vacuum, since those readily convert to lattice holes as already accounted. One effect of concern is that, if the semiconductors in the system are exposed to sufficient extra lattice holes or interacting 4-holes, their donor-acceptor arrangement is disrupted, spoiling the action of the semiconductors and perhaps destroying them.

When COP»1.0 (and even for some COP>1.0 systems), then there results a substantial Dirac sea hole current in the local vacuum occupied by the system, flowing through the supersystem from the output section back to the input section. This is an additional hole current component from the

Dirac hole current component converted into lattice hole current.[249] This unconverted Dirac sea vacuum hole current runs through the system from the output section back to the input section of the system and beyond. If not "filled" there in the input section by the incoming input electrons from the external power supply, the 4-hole current can and will pass on back into the feed line and to the distant external power supply, *eating electrons and electron current all along the way.* Usually semiconductors in the path of a substantial 4-hole current will either malfunction or be quickly destroyed.

In the input section and on back into the external power source, the effect of the unrestrained Dirac sea hole current's reaction is to interact and convert into lattice holes all along the route, which lattice holes then "eat" incoming electrons from the distant power source. Thus, the external power source "sees" the hole current as a dynamic additional load in the system's input section itself (and in the external conductors leading to it) *that also must be powered with additional positive energy electron current.*

---

[249] Here we wish to recognize the perceptivity of Peter Lindemann, who recognized the importance of "cold electricity" in COP>1.0 systems without going deeply into the more technical aspects of negative EM energy. As an example, we cite his book, *The Free Energy Secrets of Cold Electricity,* available from Clear Tech, Inc., Box 37 Metaline Falls, WA 99153, U.S.A. Particularly note his efforts to explain the Gray overunity device via the negative energy that can be associated with very strong discharges. Gray referred to this appearance of an associated negative energy component associated with the normally positive EM energy as "splitting the positive pole" (a "pole", after all, involves a stress potential). The Gray engine drawing is shown in Figure 5-14, his conversion tube is shown in Figure 5-15, and his circuit for the conversion tube is shown in Figure 5-16. As Lindemann explains, Gray also apparently succeeded in converting the negative energy into positive energy in his conversion element switching tube. Bedini's successful replication of the tube along with pertinent lab notes are shown in Figures 5-17, 5-18, 5-19, 5-20, and 5-21 by permission. Since Bedini is one of only two or three persons truly experienced in converting negative energy into positive energy, his successful replication is significant and decisive. Gray's process thus transformed negative energy into positive energy, allowing positive energy COP>1.0 self-powering systems. Bedini and the present author have filed a joint patent application filed on the process explained in this Chapter. We recognize Ed Gray as being a real pioneer in the area of COP>1.0 electrical power systems and in energy conversion. His pertinent patents are E. V. Gray, "Pulsed Cap. Electrical Discharge Engine," U.S. Patent #3,890,548, 1976; "Efficient Power Supply Suitable for Inductive Loads," U.S. Patent #4,595,975, **June 17, 1986.**

Aside from destruction of semiconductors etc.,[250] the overall effect is to increase the draw of power and current from the external source, so that the COP» 1.0 power system is thereby converted back to a COP<1.0 system!

The increased power draw due to the Dirac hole current can be remarkable. For an external battery power supply, if the hole current is sufficient to pass on into the battery itself, the battery may *charge* with negative energy while also powering the circuit! A remarkable phenomenon then takes place. A battery having been appreciably charged with negative energy may appear to be discharged (and so it is, with respect to positive EM energy). Yet, it may also continue to power a circuit. Sometimes it will, sometimes it won't, depending on the circuit. However, on connecting a normal positive energy battery charger to recharge it, the battery *precharged* significantly with negative energy will "eat" positive energy for an extended period, filling its localized Dirac Sea holes in its negative charge all the while, with the external voltage of the battery not changing at all. When all the negative energy holes are filled in the internal vacuum environment of the battery, and the negative energy charge is thus removed, suddenly the battery will start recharging with the input positive energy in a normal manner.

The batteries in Watson's generator (again see Figure 5-12) exhibited that phenomenon. Sometimes in recharging one of his batteries that had been extensively used, the battery would "eat" positive energy from the charger for a week or more before finally filling up its negative energy charge and then recharging normally from that point on.

The Dirac 4-hole current is nature's decay mechanism for the COP»1.0 system, sharply bringing it back to COP<1.0. Any successful COP»1.0 EM system must contain a mechanism for defeating (or transducing and using) this Dirac sea vacuum 4-hole current into its input section and beyond. We shall cover that subject shortly. The 4-hole current decay mechanism can *sometimes* become significant for semiconductors in the lattice 3-hole stage at COP = 5 or more, though the exact figure varies as the type and size of system and the semiconductor type. At COP > 10 or so, some systems will frequently begin to ravage themselves back into equilibrium with the vacuum, and destroy their overunity performance, unless the 4-hole current phenomenon is dealt with.

---

[250] As an example, novel effects may also happen in resistors, capacitors, and coils, and the effects depend on both the materials and construction of these components. A resistor can act as a true negative resistor and cool rather than heat, a capacitor can have its current lagging its voltage, and a coil can have its current leading its voltage. Mixtures of all three of these effects can happen in each of the three components, since each usually has some aspects of resistance, capacitance, and inductance. The emfs and back emfs in the circuit can also be altered and disrupted in a manner varying with the individual circuit and the individual operating conditions. Nonlinear oscillations and discharges of negative energy are particularly difficult to foresee and understand. Effects in dielectrics may vary widely, as can effects in conductors because of their internal impurities, etc.

Let us recap. We have assumed — to first order — a sweeping generalization that COP<1.0 systems exhibit an overall average spacetime curvature across the system (from input section to output section and load) such that excess positive energy electrons exist in the Dirac sea, without any appreciable accompanying excess Dirac sea holes interacting in the circuitry itself.[251] The slight excess of Dirac sea holes that are generated is "swept" forward out of the output section, and "eaten by the vacuum interaction" (particularly by the vacuum fluctuations) immediately upon exit from the output section. In that case, we do not have to concern ourselves particularly with Dirac Sea 4-holes and Dirac sea 4-hole currents, or at least they are usually negligible in most cases.[252]

For COP»1.0 systems, however, the overall average spacetime curvature across the system — from output section and load to input section — is such that Dirac sea 4-holes and their currents do independently continue to exist and independently flow in the local vacuum occupied by the physical system. Specifically, we shall have to consider the *tempic* effects upon such holes and hole currents, which are oriented from the output section back through the system to the input section.[253] In the COP»1.0, the

[251] In short, the requisite number of Dirac holes have interacted with and do interact with the lattices, converting into the normal "lattice hole" type of positron connected with positive mass, positive energy, but slowly migrating around the circuit in the opposite direction from the electrons. Here the true Dirac hole current exists only in short "hops" from atom to atom as the lattice holes are steadily filled by the "hopping" electrons.

[252] As stated, a kind of "hole" phenomena still exists in the circuit, but now as lattice holes and currents. Since these are attached to ions each having a mass much greater than that of an electron, the lattice hole currents may usually be neglected except in semiconductors, where they are taken into account in the very design of the semiconductor component itself.

[253] When we state a concept such as "negative mass" or "negative energy", we have conceptualized it *as if observed.* Hence, unwittingly we have made it an instantaneous, but iterated, concept. Rigorously, no such thing as mass or energy

effect of the increased tempic force is to sweep a Dirac sea 4-hole current back through the system from output to input, and even on out into the conductors from the external power supply, and on further out into the external power supply itself. The 4-hole current continues its "reverse feedback into the powering end of things" until quenched by filling it with incoming electrons. This conversion of 4-hole current to filled Dirac sea 4-holes by induction of lattice holes also squelches the concomitant negative EM energy fields, potentials, and waves associated with Dirac 4-holes as sources.[254] The reason is that a filled Dirac hole is a dual system of two opposite charges, hence produces counteracting positive and negative energy fields.

The "4-hole-eating and filling process" is therefore the exact decay mechanism that squelches the natural electrogravitation associated with Maxwellian circuits which use Dirac 4-holes and 4-hole currents, negative energy fields, negative energy potentials, etc. When the circuit converts its free Dirac 4-holes into lattice holes, all electrogravity capability is lost, insofar as practicality is concerned.

For electrical power system purposes, the external environment of the COP»1.0 system furnishes an extra "electrical load" demand upon the external power supply, so that the extended system consisting of the (external power supply + power system) decays back into overall COP<1.0 condition and operation. For this reason, any appreciable involvement of Dirac 4-hole currents in a COP>1.0 EM power system must be *transduced* into positive energy electron currents, which requires just a special form of regauging. One need not confine regauging and gauge freedom to positive energy potentials and positive energy fields alone.

From the vacuum consideration and particularly from Dirac's theory of the electron, in COP>1.0 EM systems and circuits we shall consider that the Dirac sea 4-holes involve negative energy states (negative mass), and that positive energy states involve positive mass. In our approach, we shall consider a negative energy state (Dirac 4-hole) in the vacuum as having

---

independently exists except at one instant in time. Instead, mass exists in a continual iteration of *m* —> *mt* —> *m* —> *mt*... and energy exists in a continual iteration of *E* —> Et—»E—» *Et*.... and so on. So during the masstime or energy-time (action) state, time and either mass or energy are intimately interactive. See our previous discussion in this book of the mechanism for the flow of a mass through time.

[254] Obviously that kills the antigravity aspects, because it kills the negative energy fields that would give negative curvature of local spacetime.

negative mass, which means that antigravity can be involved as well.[255] In this treatise, we consider only electron states, thus only negative energy and positive energy electron states.

9.3 Primarily Concerned With Dirac Sea Holes and Not Lattice Holes

We also stress the dramatic difference between a Dirac sea "electron 4-hole" in the vacuum (prior to observation) as compared to a lattice 3-hole in a semiconductor or other material component (after interaction and therefore after observation). In the case of a lattice 3-hole, usually the "positron" (hole) is attached to an ion, whose mass (i) is very much greater than the mass of an electron, and (ii) is also positive. In other words, a lattice hole is just a part of an ordinary positive ion. So the theory of "lattice holes" and "lattice positrons" does not primarily apply to the situation we are discussing. On the other hands, semiconductors exposed to any significant currents of pure Dirac sea 4-holes in the vacuum usually react and are nearly instantly destroyed because of the immediate disruption of their donor-acceptor arrangement and functioning. Interaction (filling the incoming Dirac 4-holes with contributed electrons) causes sudden mass disarrangements in their atoms and nuclei and lattice bonds. Thus, in the presence of Dirac sea 4-holes and 4-hole currents, usually any interacting solid-state electronics of the circuit or system will be dramatically affected. For switching in the presence of holes and hole currents, in most cases one may have to turn away from semiconductors and utilize mechanical or electro-mechanical switching, although this requirement varies with the type of semiconductor, the COP, the system geometrical layout, and the subsystem operations and layouts. Sometimes semiconductors may still be used if they are "brutes", with current handling capabilities well beyond what is "normally expected" in that circuit.

---

[255] We assume that the important thing about mass is its localized energy. Hence we associate positive mass with positive localized energy (with respect to the ambient energy density of the massless vacuum) and we associate negative mass with negative localized energy (with respect to the ambient energy density of the massless vacuum). In this view, the empty Dirac sea hole has both negative energy and negative mass. Note that this differs from the conventional assumption, which unwittingly addresses the positron as having been observed, producing a directional reversal. The rigorous test of our thesis is that concentrations of Dirac sea holes *must* generate antigravity *potential,* while currents of Dirac sea holes must generate antigravity *force.* The conventional view, assigning positive mass to the positron after observation, would produce gravity, not antigravity, but the force would be reversed by the reversed parity after observation. We argue that antigravity has been hidden and lost by its parity reversal in the assuming positive mass for the positron *as if that were the mass of the positron before it is observed.* In an overunity system with powerful steady-state Dirac 4-hole currents, the 4-hole currents can be differentiated by the cooling (negentropy) effect of the COP»1.0 system—when shorted to produce a "current surge" — rather than heating (entropy) effect. The magnitude of the hole current (which determines the average presence of the hole density) can be differentiated by a measurable weight loss in the COP»1.0 system. The successful antigravity production of the Sweet COP»1.0 device — discussed later in this chapter — experimentally verified these hypotheses.

9.4 Peculiarities of Spacetime Curvature and Dirac Hole Interactions

When considering the interaction between *curved* spacetime and the Dirac vacuum, immediately we violate the premise that all electron states in the Dirac Sea are always filled (which condition we surmise — along with Dirac — only exists on the average in a flat spacetime region). In a curved spacetime, we can have sustained excess 4-holes in the vacuum, or we can have holes that accept an electron falling into them *without emitting radiation.* The radiation-free "pair annihilation" reaction is possible when the spacetime simultaneously changes its curvature appropriately and thereby "absorbs" and holds the excess EM energy that would otherwise be emitted if the spacetime remained flat or decayed back to the flat state.[256] So long as that "energy-storing ST curvature" remains, there is no photon energy emission. If the "energy storing ST curvature set" has a net curvature instead of being flat, and then decays to a flat spacetime, then the previously retained energy is emitted as photon energy.

In that case the reaction provides the conventional emitted pair annihilation energy. In *that* case, the emitted energy may also transform into emitted particles in reactions well known in particle physics.

---

[256] Eerily, this interaction represents a situation where opposite reaction forces (the vacuum upon spacetime, and spacetime upon the vacuum) can create a special kind of stress in the system and changes in that stress. This leads to a peculiar kind of stress waves (scalar waves — longitudinal EM waves, so to speak) that are not covered in circuit theory anywhere. A symptom is that these stress waves react with components, amplifiers, preamps, etc. to create electrostatic waves. The effect of these waves is seen as the "crawling" of electrostatics all over and in the electrical equipment that is in one's laboratory. The effect can readily damage oscilloscopes, etc.

The normal emission of photons by pair annihilation need not occur, if the excess decay energy is absorbed *and held* by the local spacetime changing and holding its energy density and therefore (i) changing its curvature and (ii) stabilizing in that new excited state instead of decaying back to a flat spacetime.[257] Under such "non-emitting" pair recombination, a Dirac Sea 4- hole may simply "eat" a free 4-electron, and instead of emitting radiation the process simply affects the curvature of spacetime (and the state of the local Dirac sea) and changes it accordingly. A new "equilibrium state" is therefore reached in the supersystem between the opposite forces of the vacuum acting on spacetime and spacetime acting on the vacuum.[258] In other words, we consider change in the curvature of spacetime to exhibit EM energy *source* (emission) or *sink* (absorption) capability, with respect to the nominal ambient flat spacetime and with respect to conventional EM interactions.[259]

Or simply put, the energy emitted by the combination of positive and negative energy in the vacuum may be "locally radiated", "locally absorbed", or "locally transferred" to the other local component of the supersystem: the local curvature of spacetime. *It is all spacetime curvature dynamics.* By definition, a flat local spacetime consists of a vacuum with all its Dirac sea holes filled. Therefore, if "filling a 4-hole" results merely in relaxation of a set of local ST curvatures back to a flat spacetime, no EM radiation need be emitted, contrary to the usual pair annihilation phenomenon where additional energy (curvature of spacetime) is involved.

---

[257] Our view is that in normal pair annihilation this curving of spacetime to absorb the emission energy occurs first, and then the new ST curvature decays by what is called "emission of photon energy". We regard the photon as a specialized ST curvature traveling at light speed. So what is emitted as a photon is in fact the formation of a locally structured spacetime curvature (an "engine") traveling at light speed. It is a propagating spacetime curvature "engine" or "set of engines". In the process we are describing, the emission of the photon is the decay of the locally formed ST curvature set. If that set does not decay, the locally formed ST curvature set remains and no photon is emitted.

[258] This new kind of equilibrium is extraordinarily important, particularly in chemistry, where it will allow the "locking" of impossible combinations of particles into strangely stable "impossible molecules" with ongoing "impossible chemical reactions". The further discussion of the startling new chemistry that emerges is well beyond the scope of this paper, but such chemical states are included in nonequilibrium thermodynamics as *stationary states* — e.g., see Kondepudi and Prigogine, *Modern Thermodynamics,* Wiley, 1998, Chapter 17. At least one company has already succeeded in initiating this startling new kind of chemistry, with several patents granted and several more in process. Similar new "impossible stabilized excited states" and "impossible nuclear reactions" also occur in nuclear physics.

[259] The classical EM assumption of the flat local spacetime is in fact an assumption that the decay of the intermediate "ST curvature absorber/sink" occurs instantly.

Similarly, in a ST curvature, there can conceivably be electrons without "holes". If the curvature is negative, there *will* be unfilled Dirac sea holes present. Of course this is just a normal positive energy 4-electron, where the flat spacetime assumption in classical electrodynamics is in error. Instead, the local spacetime is positively curved, producing positive gravity. Whenever the energy density of spacetime-vacuum interaction changes from the "ambient vacuum and flat spacetime" condition, either positively or negatively, that constitutes an appropriate curvature of spacetime as well as a change in the vacuum dynamics. Therefore, in classical electrodynamics the assumption of a normal electron (or any other charge or mass) existing in an uncurved spacetime and an unaltered vacuum is a non sequitur. The local spacetime has in fact "curved" to an extent of holding the energy of one electron mass.[260] Further, the Dirac Sea has also changed to contain one unfilled and unobserved Dirac 4-hole.

With these things in mind, let us now introduce a conceptual model for dealing with these novel 4-hole phenomena that are experienced in COP>1.0 systems — and that have been the bane of many a would-be energy researcher attempting to develop a system extracting its EM energy from the vacuum and using it to also power the system itself.

9.5 Charging and Discharging a Capacitor

In electronic circuits, we usually think of charging a capacitor with electrons, and then discharging the capacitor to emit electrons again. However, we can charge the same capacitor with Dirac sea holes, or with a combination of Dirac sea holes and electrons. Then we can switch the capacitor adroitly and force it to give us back electrons. This "regauging" effect can thus be used to transduce negative energy current into positive energy current. So let us examine the charging and discharging process more closely. This is part of the Bedini process for handling the 4-hole current problem in overunity EM systems, to allow close-looping a system into self-powering operation.

---

[260] Mostly physicists think only of the *spatial* energy change. However, the time domain of spacetime also acts as a sink or source. Essentially one kilogram of mass only has the energy of 1 second of time. Hence the minuscule *mass* of the electron results in such a minuscule "sink" in the time domain that it has been inadvertently ignored. However, during that fraction of time where the electron exists as "mass time" and "charge time", it has considerable time energy, and *that* curvature of local spacetime cannot be neglected with respect to supersystem interactions.

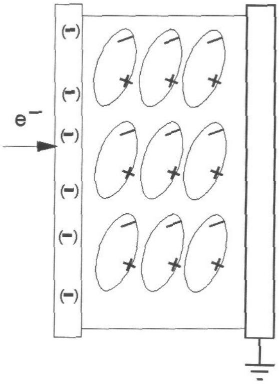

Figure 9-6 Normal strain in a dielectric capacitor that charges with electron current

Figure 9-6 shows the normal strain induced in a capacitor containing a dielectric between its plates. We charge the left plate with electrons, which repel the negative ends of the molecules in the dielectric, straining them to the right.

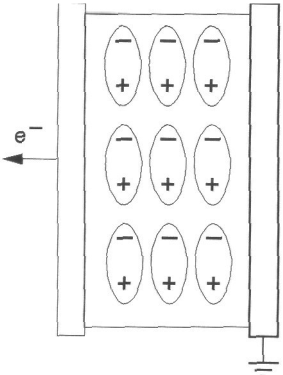

Figure 9-7 Strain relief in a charged dielectric capacitor that discharges.

Figure 9-7 shows the relief of this dielectric strain, by allowing the piled electrons on the left plate to be pushed back into the circuit from the left plate. Note that the electrons emitted in the relaxation of the dielectric strain pass back into the circuit from the same plate that was charged. This is a special kind of "reflection" symmetry. For a perfect capacitor, the total

Vq joules stored in the charged capacitor and drawn from the circuit, is returned to the circuit.

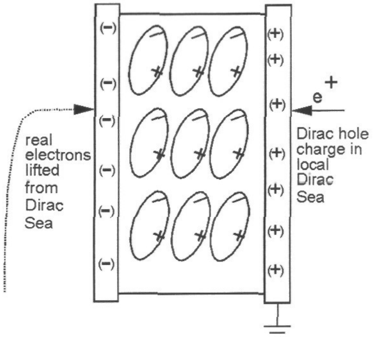
(−)
(−)
(−)
(−)
(−)
(−)
(+)
(+)
(+)
(+)
(+)
(+)
(+)
e+
Dirac hole charge in local Dirac Sea
real electrons lifted from Dirac Sea

Figure 9-8 Strain in a dielectric capacitor supersystem that charges by Dirac hole current and lifted Dirac electrons

Figure 9-8 shows a remarkable thing when we charge the capacitor with Dirac sea holes. Suppose we charge the rightmost plate of the same capacitor used before, while the leftmost plate is temporarily disconnected from the circuit (or connected to another capacitor). Now Dirac Sea holes pile up in the local vacuum connected with the right plate. Whether or not they interact with the right plate, these positively charged holes attract the negative ends of the dielectric molecules, again straining the top of the dielectric molecules to the right. The strain of the dielectric is a force interacting upon the local vacuum, and it lifts real electrons from the Dirac Sea (or pulls them from the circuit or other attached capacitor) and onto the leftmost capacitor plate. The charged capacitor *in its supersystem* thus has additional functions that occur. The total *Vq* joules stored in the charged capacitor, were this time drawn (converted) from the applied Dirac hole current and the resulting interaction upon the vacuum, to lift real electrons onto the leftmost plate.[261]

---

[261] The reader should recall that a force is something automatically produced by nature to restore symmetry to a broken symmetry condition. A "4-hole negative energy force in the Dirac Sea" produced on one plate of the capacitor will evoke a "positive energy force outside the Dirac Sea" on the other plate as an attempt to restore symmetry. To produce this "positive energy force", positive energy electrons are lifted from the local Dirac Sea of that plate.

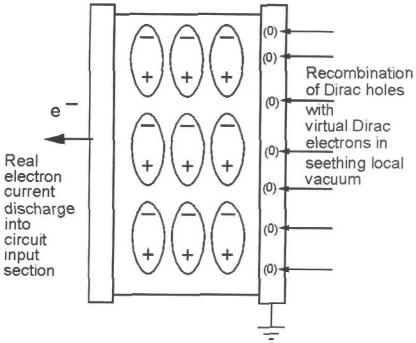
e⁻
Real electron current discharge into circuit input section
(0)
(0)
(0)
(0)
(0)
(0)
(0)
Recombination of Dirac holes with virtual Dirac electrons in seething local vacuum

Figure 9-9 Strain relief in a dielectric supersystem that discharges after charging by Dirac hole current and lifted Dirac electrons.

Figure 9-9 shows the strain relief of the charged capacitor in Figure 9-8. First the charged capacitor is re-connected to the circuit or element to which one desires to provide with electron current. The strained dielectric relaxes, and electrons in the leftmost plate are returned to the circuit. The Dirac holes in the rightmost plate discharge back into the local vacuum (a supersystem interaction) and are "eaten" immediately by the readjustment of local curved spacetime.[262]

The result is that the Dirac sea 4-hole current arriving in the input section has been intercepted by a waiting capacitor, and has been effectively transposed into normal electron current that is furnished back to the input section from the capacitor. This additional electron current helps power the circuit and reduces the power drawn from the external power source. The 4-hole current feeding this capacitor, together with transduction of the 4-hole current into electron current, effectively comprises a true negative resistance process. In other words, energy in unusable (in this case, detrimental) form is received, transformed into energy in usable (beneficial) form, and that usable energy is discharged back into the circuit input to assist in powering the system.

Transducing negative energy into positive energy — and vice versa — in the system does not violate the conservation of energy law in the supersystem, when most generally stated. With such regauging (which involves supersystem reactions), it is the *absolute value* of the energy that is conserved in the system, not its algebraic sign. Regauging (by gauge freedom) is after all an effect or condition locally imposed, but with appropriate energy exchange with the rest of the external universe via spacetime curvature dynamics exchange and vacuum dynamics exchange. The algebraic sign *is* conserved in the universal supersystem by the additional compensating reactions occurring in the other two components of the universe's supersystem.

---

[262] From experiments, it seems that Dirac holes forming at the "output" section of a **COP<1.0** device or system are almost instantly filled with electrons in the violent fluctuations of the zero-point energy. When forming at the "input" section of a **COP>1.0** device, this is not the case, and usually the holes can only be filled by incoming electrons in the power input, unless special provisions are taken to convert **the** arriving hole current (in the input section) into electron current.

The reader should strongly realize that energy is only required to be conserved in the universe's supersystem, not the system component alone. Again, this is an open system far from equilibrium with its active environment. In that case, energy is conserved in the ensemble of system and environment, but not necessarily in the system itself. This is no stranger than a common windmill or waterwheel.

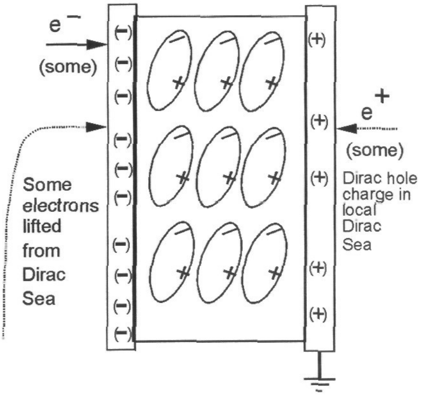
e^-
(some)
Some electrons lifted from Dirac Sea
(−)
(+)
e^+
(some)
Dirac hole charge in local Dirac Sea

Figure 9-10 Strain in dielectric in a capacitor supersystem that charges with a mix of electron current, Dirac hole current, and lifted Dirac electrons.

Figure 9- 10 shows a "mix-charging" variant of the 4-hole current charging of a capacitor, in which some real electrons are fed from the circuit to the left plate. Also, the Dirac sea hole current arriving in the input section is intercepted to charge the right plate, and some electrons are also freely lifted from the local Dirac Sea onto the left plate. Figure 9-11 shows the discharge of this type of charging of the capacitor, and the relief of the strain in the dielectric.

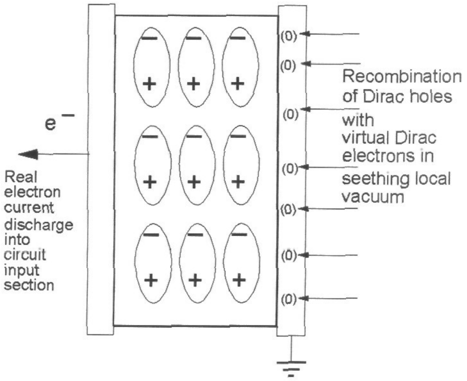
e⁻
Real electron current discharge into circuit input section
(0)
(0)
(0)
(0)
(0)
(0)
(0)
Recombination of Dirac holes with virtual Dirac electrons in seething local vacuum

Figure 9-11 Strain relief in a dielectric that discharges after charging of capacitor supersystem with mix of electron current, Dirac hole current, and lifted Dirac electrons.

In Figure 9-11, for example, we accent that the holes on the right plate are maintained only so long as there is a net strain on the dielectric and a net force across the capacitor. When the capacitor is discharged, the holes are discharged into the local vacuum in such a way that they now instantly recombine with virtual Dirac electrons in the vacuum fluctuations.

There are other ways and "tricks" of transducing the Dirac sea hole 4-current into electron current, but these methods are still proprietary to Bedini and will not be included in this presentation. We accent that all such "tricks" are actually methods of re-engineering the other two components of the supersystem — the active local vacuum and the local curvature of spacetime. This is using and applying what we have referred to as "engines".

9.6 The Open-Loop COP>1.0 EM Power System

Figure 9-12 shows the use of the transduced Dirac hole current as real input energy to "catch and lock" a COP>1.0 system in its overunity condition and stabilize its operating point for stable self-powering of the system and its load. The COP>1.0 operation reverses the normal tempic potential across the system, so that Dirac hole 4-current arrives in the input section, preparing to "decay" the system back to COP<1.0 condition.

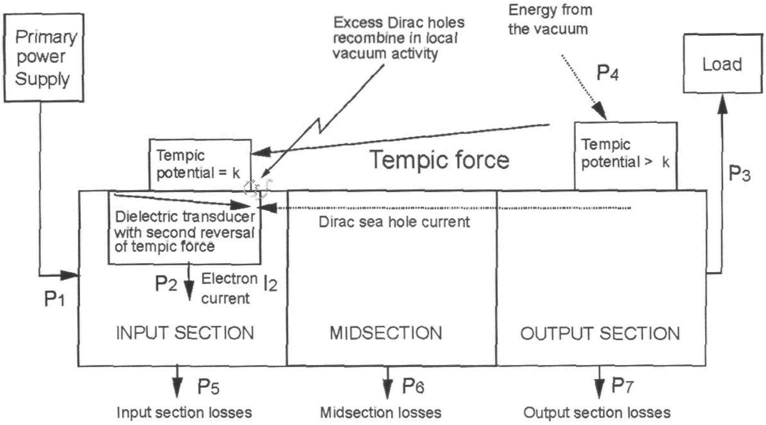
Primary power Supply
Excess Dirac holes recombine in local vacuum activity
Energy from the vacuum
P4
Load
Tempic potential = k
Tempic force
Tempic potential > k
P3
Dielectric transducer with second reversal of tempic force
Dirac sea hole current
P2
Electron current
I2
P1
INPUT SECTION
MIDSECTION
OUTPUT SECTION
P5
Input section losses
P6
Midsection losses
P7
Output section losses

Figure 9-12 Open-loop use of transduced Dirac hole current energy as real input energy to reduce power input required from the primary power supply.

However, the arriving hole current is intercepted by an adroitly switched dielectric transducer, which again reverses the tempic force. As explained above, the capacitor is charged with 4-hole currents (or a mix) and then discharges electron currents into the system — in this case, into the input section itself— by adroitly engineering the local vacuum state and the local curvature of spacetime.

The input section does not care where it receives its electron current from; it merely needs sufficient for its powering needs. Hence, the input section happily accepts this transduced electron current to partially power the system. This in turn reduces the "system load" that the external primary power supply "sees". The primary power supply thus reduces the current and power that it is furnishing to the system. In short, now the external power supply sees the combination of the normal system load and a "negative resistor (negative load)" represented by the transduced current and power. Its required input power is reduced accordingly.

9.6.1 Close-looped System: A True Negative Resistor

Figure 9-13 shows the use of the same process given in subparagraph 9.6 above and in Figure 9-12, except that now there is additional 4-hole current received and available, for transduction into normal electron current into the input section of the system. Sensors and governing are added to this operational mode, so that the system may *adjust and regulate its self-furnished electron current* to meet the total needs of its losses and its load. The system thus adapts the transduction of Dirac sea hole current: to fit its needs and is capable of regulated self-powering operation.

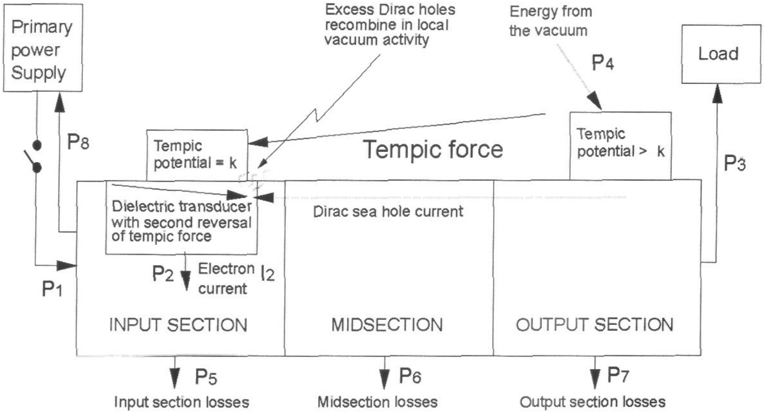
Primary power Supply
Excess Dirac holes recombine in local vacuum activity
Energy from the vacuum
P4
Load
P8
Tempic potential = k
Tempic force
Tempic potential > k
P3
Dielectric transducer with second reversal of tempic force
Dirac sea hole current
P2
Electron current
I2
P1
INPUT SECTION
MIDSECTION
OUTPUT SECTION
P5
Input section losses
P6
Midsection losses
P7
Output section losses

Note: Power outputs P3 and P8, and current I2 have sensors reporting to the Governor in the output section. The governor adjusts the potential energy in the output section to maintain steady output or adjustment of output to demand.

Figure 9-13 Close-looping the C0P>1,0 system with transduced hole current.

The system is first started in open-loop mode. The governor then gradually increases the transduction of 4-hole current into electron current, and the supply of that transduced electron current to its own input. This steadily reduces the "current draw" from the external primary power supply. When the "external current draw" reduces to zero (the Kron condition), the system "locks" into that position, with automatic governing to adjust the transduced electron current as the loads and losses vary. At that point, the external power supply is disconnected automatically.

This "locked and governed" system is now a *self-powered system,* freely taking its energy from the active vacuum and local curvatures of spacetime, to power both itself and its load. So far as we are aware, this is the first time that the technical means of locking and stabilizing COP>1.0 operation in an EM system in an open loop and closed-loop operation has been given.

If sufficient Dirac sea 4-hole current is available and transduced, excess electron current can be available at the system input section, even when the system is fully self-powering. In this case, the additional current can be fed back to the power line, so that the self-powering system not only powers its own load but also helps power the power network. Or additional load can be added and powered.

We point out that Gabriel Kron apparently had a solution to these problems in the late 1930s, and that solution has been suppressed to the present day. We give the following quotations to show Kron's knowledge of these things.

Quoting Lynn and Russell {565}:

> *"Kron has never published details of his method of making the polyhedron self-organizing, although his published results show that in this state it has some remarkable properties, associated with harmonic integrals on multiply connected spaces."*

In reference to Lynn and Russell's remarks, we point out that the Dirac sea 4-hole current from the output to the input is a special case of Kron's more general discovery of the multiply-connected open path existing through space between any two points in a circuit that are at differing potentials.

Quoting Kron {566}:

> *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of two nodes. (Previously —following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which created 'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

Again quoting Kron {567}:

> *"Although negative resistances are available for use with a network analyzer, in practice it is more convenient to use a second type of circuit, in which the positive and negative resistances are replaced by inductors and capacitors and the dc currents and voltages are replaced by ac currents and voltages of fixed frequency. The use of the second type of interpretation is equivalent to multiplying the wave equation by* $i = \sqrt{-1}$."

The reader should particularly note that in the first part of the sentence Kron stated unequivocally that negative resistances were available for **use** on the network analyzer, a large analogue simulation used by the Navy **and** administered at Stanford University with the contractor being General Electric. The censors missed that one.

Again quoting Kron {568}:

> *"When only positive and negative real numbers exist, it is customary to replace a positive resistance by an inductance and a negative resistance by a capacitor (since none or only a few negative resistances exist on practical network analyzers).*

We believe the censor caught this one, and forced Kron to add the words, "none or". Kron had at least three negative resistors, and possibly more.

Again quoting Kron {569}:

> *"Now a value E of the negative resistances, at which the generator current becomes zero, represents a state at which the circuit is self-supporting and has a continuous existence of its own without the presence of the generator, as the negative resistances just supply the energy consumed by the positive resistances. ... When the generator current is positive the circuit draws energy from the source, and when the current is negative the circuit pumps hack energy into the source. At zero generator current the circuit neither gives nor takes energy, and theoretically the generator may be removed. "*

Here we believe the censor forced Kron to add the word "theoretically". In this quotation Kron reveals that he was well aware of the process for close-looping a system — starting it in open-loop mode and bringing the negative resistor currents up to match the system needs nnd reduce the generator current to zero, then disconnect the generator. He also was quite aware that excess transduced energy from the negative resistor could be available at the input, in which case energy could be furnished back to the power network while the system powered itself and its load simultaneously.

We also point out that Dirac's definitive electron theory — covering the Dirac Sea and Dirac sea holes — was published in 1930, the year the present author was born. In the late '30s and in the 40's Kron certainly was well aware of that work. In addition, Kron applied full general relativity to electrical circuits and machines, and he was far ahead of his time. His work is also still far ahead of the present time, at least in the open scientific community. He also was aware of developments in particle physics, and so he did have the gist of the supersystem concept in his mind {570}, however he may have thought of it. His concept of the "open path" effectively includes the effects upon the system from the other two components of the supersystem, so it appears that Kron's "system plus its open path" roughly corresponds to our concept of the "supersystem."

Further, Kron was a mentor to Floyd Sweet, who also worked at General Electric but in a different project. Sweet often spoke glowingly of Gabriel Kron.[263] Sweet almost certainly knew the secret of Kron's negative resistor. It is the present author's opinion that Sweet's vacuum triode amplifier was derived from Kron's negative resistance unit. With both deceased, we will never know for certain.

9.6.2 Suppression of COP>1.0 Systems Should be Investigated

During our decades of research, we ran across several cases where U.S. scientists actually produced working COP>1.0 systems.[264] None of these overunity systems was ever allowed to go forward with independent development. This is an area that cries out for a full and vigorous Congressional investigation, and appropriate legal action should be taken by the U.S. government against the suppressing parties, whomever they might be {571}. Simply put, the U.S. government, the U.S. taxpayer, and the scientific community have been ripped off and deluded in the field of electrical energy for more than half a century. That continues to the present, as witness the disappearance of the Watson generator (accompanied by Watson and family breaking off all contact with their close colleagues) and the suppression of the Kawai device and process. The latter suppression was accomplished in 1996 in the personal presence of the present author and the members of the Board of Directors of our company, CTEC, Inc. An accidental overunity frequency converter was even initially placed in the original Minuteman missile, but was altered back to COP<1.0 operation when the electronics past the converter kept failing because of overload power furnished by the converter.

---

[263] The present author worked with Sweet off and on for a decade or more, and designed the antigravity experiment that Sweet performed.

[264] These incidents were previously exposed in a draft book, *Matters Arising,* carried on the Internet for some months and then withdrawn. Some things must wait for their appointed time, and so must certain kinds of information.

9.7 Remarks on Dirac Sea Holes, Currents, and Negative Energy

The subject of the Dirac Sea, Dirac holes, Dirac hole currents, and negative energy still has elements of strong disagreement within the scientific community and among leading theorists. Often the "conclusion" one reaches depends on the way the starting assumptions are made and then on how the mathematics is applied. There is no definitive experiment establishing that negative energy cannot exist; to the contrary, many things do occupy negative energy states. The binding energy of the nucleus, e.g., involves a loss of mass (loss of positive energy) and hence involves negative energy states, as is well known. Here the "negative mass" effect is immediately and experimentally verifiable, and is recognized. The total mass of the bound nucleons is less than the total mass of those same nucleons when unbound, and this fact is widely known and accepted.

A single white crow is sufficient to prove that not all crows are black. Hence a single experimental proof of mass-energy "loss" and negative energy states establishes that negative energy and negative mass are real. Indeed, every photon absorption by mass and every photon emission by masstime prove the same point universally: mass is a variable. The same assembly of protons and neutrons in a nucleus, having lost mass, also has "lost" some of the gravity that the positive masses of those unbound protons and neutrons would otherwise produce. Therefore, there is an appropriate amount of "negative gravity" already acting in conjunction with positive gravity, in every atomic nucleus. It is associated with the negative energy states of the bound nucleons in every nucleus.

Consistent with the initial quotations and comments at the beginning of this chapter, we present an interpretation by which we can understand some of the major novel phenomena that arise in laboratory experiments with COP>1.0 electromagnetic systems, including COP»1.0 systems.

We propose that COP»1.0 EM interactions are widespread in highly energetic astronomical interactions {572} and even in highly energetic phenomena on Earth, when the supersystem changes are also accounted.[265] We argue that novel negative energy phenomenology — such as encountered in the COP>1.0 experiments by Magnetic Energy Ltd. in various versions of the COP>1.0 motionless electromagnetic generator (MEG) {573, 574}, by Bedini {575} in his induction of negative resistor actions inside the batteries of battery-powered COP>1.0 EM systems, and by Sweet and the present author in experiments with the Sweet COP» 1.0 vacuum triode amplifier system {576} — is also generated in these COP»1.0 highly energetic astronomical interactions, yielding the same kind of negative energy and negative energy currents.

---

[265] As an example, in looking at sprites and other such sharp electrical discharge phenomena, we may be observing the prompt *decay* of COP»1.0 interactions which were self-initiated in the atmosphere.

This Dirac hole negative energy generation of COP»1.0 EM processes is proposed as the mechanism which directly produces the negative energy fields and potentials that generate the mysterious antigravity accelerating the expanding universe {577}. This is a reasonable assumption, even though it is against the received view, because it is based on some laboratory experiments, not just theory. It also is the unexpected effect of the cosmological constant introduced by Einstein {578}, and it is consistent with Zeldovich's demonstration that the energy density associated with Einstein's repulsive cosmological constant is — or is associated with — vacuum energy.[266] The received view *is not* based on direct experiments with COP>1.0 EM systems, but only on COP<1.0 EM systems. Until the received view is broader based, it does not adequately represent nature. Hence it does not represent a valid refutation of experimental negative energy systems producing antigravity processes.

In Bedini's experiments and in the MEG experiments, antigravity is essentially absent or so small as to be nearly nondetectable. However, these experiments generally have COP of 1.5 to 4.0, mostly, and sometimes 10 or 12 (we have experimented up to 20 or more stably, and unstably at nearly 100). Hence the Dirac hole current is relatively small in them, since the magnitude of the Dirac current is a function of (COP - 1).

Sweet's COP > 1,500,000 experiments produced a very large flow of negative energy back across the system, substantially curving the local spacetime and reducing the weight of the system on the laboratory bench. Figure 6-7 shows the first Sweet device, which produced six watts output in open-loop operation. Figure 6-8 shows a later Sweet device which produced COP = 1,500,000 and was pushed past that COP to produce the antigravity in the laboratory, smoothly reducing its weight by 90%. Figure 6-9 is a plot of the actual readings taken during the antigravity experiment. Sweet performed the experiment in California, reading off the measurements to me over the phone as he made each one.

---

[266] Believed to have been demonstrated in 1967. But check Ya. B. Zeldovich and A. A. Starobinsky, "Particle creation and vacuum polarization in an anisotropic gravitational field," *Zh. Eksp. Teor. Fiz.*, Vol. 61, 1971, p. 2161 *(Sov. Phys. - **JETP,*** Vol. 34, 1972, p. 1159); Ya. B. Zeldovich, "Vacuum theory — A possible solution to the singularity problem of cosmology," *Uspekhi Fizicheskikh Nauk,* vol. 133, Mar. 1981, p. 479-503. Translation in *Soviet Physics — Uspekhi,* vol. 24, Mar. 1981, p. 216-230.

Another indication of the negative energy effect present in the output section of the Sweet device was tested by suddenly shorting the output leads. This resulted in instant freezing of moisture from the air, so that the leads were instantly frozen and covered with ice. Contrast this to the normal heating of shorted wires, where the energy is positive.

Sweet never revealed his activation process, whereby he initiated very strong self-oscillation — of the nucleus of the barium atom in barium ferrite magnets — with the local vacuum treated as a semiconductor. One part was known to the present author and another part was known to Rosenthal, a professional test engineer. In addition, how Sweet converted his excessive hole current is not completely clear. It appears that he somehow used his coils in quadrature together with special circuitry to interact the excess hole current into his magnets, directly with the self-oscillating barium nuclei and thence with the other nuclei in his magnets. If so, it created the "rough equivalent" of increasing the associated binding energy offset and locked with concomitant additional local spacetime curvature.

If this speculation is correct, he "locked" the nuclei of his magnets to the local curved spacetime, maintaining the overall local binding energy in the nuclei themselves while producing a great excess of "binding energy" (negative energy) in the immediate vacuum. This would explain the antigravity, and would be consistent with the experimental results as well as the increase in antigravity as the output load of the device was deliberately increased.

We must use the supersystem concept, as in the quoted comment at the beginning of the chapter, to enable a better understanding of these novel phenomena. Together with the Dirac negative energy production and high COP system, the hypothesis of "additional negative energy acting on local spacetime immediate to each nucleus" allows the generation of the observed antigravity to be understood. However, we accent that this mechanism is still a hypothesis, and it must be further substantiated by experiment.

9.7.1 Lattice Holes

In a material conductor or other component of an electromagnetic circuit, the conventional model envisions currents in the Drude electrons as the electrons moving by "hopping" from molecules or atoms in the material lattices into the Drude gas, with each hop leaving behind a positively charged "vacated electron position" called a *lattice hole,* or just a "hole". Subsequent Drude electrons are attracted to fall into the holes, etc. This net movement of the hopping electrons down the conductor, etc. comprises the conventional electron current through the conductor. We point out that the actual movement of the electrons in their "hopping" is vastly more energetic than the very feeble net "drift" of the overall ensemble of Drude electrons longitudinally down the wire as current. The energy of the total electron activity ongoing in a simple 1-foot length of copper wire would power a city, if it could all be harnessed and utilized.

At the same time the electrons hop with a slow net progress down the wire, the "holes" in the material lattices are also collectively seen by the external observer as if "slowly changing their net positions" longitudinally in the opposite direction. One speaks of a corresponding "hole current" or "lattice hole current" in the opposite direction to the electron current. However, there are subtle differences in the flow of the two currents, because an electron in the Drude gas is far more kinetic than the positive ion left behind by an electron that has "hopped up into the electron gas". Nonetheless, the positive lattice ions do "rock" physically from the induced forces upon them, as the holes are opened and filled. In conventional electrical power systems, this equal and opposite energy reaction in the ions of the lattice materials, induced by the applied fields, is usually ignored. In most classical EM power system theory the holes and their "composite equal and opposite energy and work in the system" are also ignored, except in semiconductors and semiconductor circuits.

9.7.2 Newton's Third Law Is Accounted in the Supersystem

Because half the EM energy and work in the circuit is usually ignored (e.g., in the Maxwell-Heaviside-Lorentz standard equations), electrodynamics is said to be free from Newton's third law reaction. This may be a non sequitur, and it follows from the assumption that — for an EM wave, field, or potential in space — no interaction exists with the active vacuum or with the curvature of spacetime. In fact, those reactions do exist in the supersystem. The field, wave, and potential do indeed interact with both the quantum mechanical vacuum and the general relativistic spacetime.

Indeed, the field, wave, and potential may be said to *identically be* changes induced in general relativistic spacetime. The absence of the "transmitting" physical system leaves us with the remaining two components of the supersystem, each of which extends to infinity. Again, Figure 9-1 shows the actual interactions (the supersystem) and Figure 9-2 shows the erroneous and arbitrary "kill of the supersystem" by standard classical electrodynamics.

When higher group symmetry electrodynamics such as O(3) is applied, then the interactions between the three parts of the supersystem emerge and must be accounted. This is particularly true of COP>1.0 EM systems. However, as Finster points out {564}, the Dirac sea effects exist *wherever* there are EM fields, potentials, and waves. We add that curved spacetime effects also exist *wherever* there are EM fields, potentials, and waves. And wherever Dirac sea effects exist, or wherever curved spacetime exists, then there the local vacuum and local spacetime interact with each other.

However, as we shall point out, usually most of the Dirac sea hole effects can be neglected in COP<1.0 systems and in most lower gain COP>1.0 systems because of the time-energy aspects. For COP»1.0 EM systems, the time-energy aspects are partially reversed, and instead of carrying the hole effects away from the system, the time-force from system output to input forces the Dirac sea holes to flow back in the local vacuum across the system from its output to its input. This flow constitutes an additional "load" appearing in the input section and requiring additional electron current from the external power supply to fill the holes. If the power supply does not furnish sufficient electrons to fill the hole current at the input section of the system, then the unfilled hole current passes back into the feeder line toward the external power supply, and even into the distant power supply itself. In its progress, it eats incoming current all along the way, requiring that more current be furnished by the distant generators.

9.8 Remarks on the Observed Acceleration of the Expanding Universe

As a candidate for *dark negative energy*, the present author has proposed vacuum negative energy and currents in curved spacetime regions partially containing excess *empty* Dirac sea holes as well as normal *filled* Dirac sea holes. The process by which these excess Dirac holes are produced has been advanced, and is supported by at least one dramatic experiment. One result of such "negative energy areas" of negative spacetime curvature would be "jumps" increasing the velocity of light and high-energy particles traversing these altered regions of spacetime. In fact, there is some evidence for just such jumps, according to a hypothesis by Richard Lieu of the University of Alabama in Huntsville {580a, 580b}. Otherwise, high-energy particles produced in distant processes in the cosmos would not be detected on earth, but would be annihilated in highly energetic collisions.

The departure of large-scale spacetime from homogeneity would also result in some loss of correlation in the presently fitted red shift versus distance relationship. Again, there is tentative indication that such may prove to be the case. The most distant galaxy recently detected has a red shift of 6.56, which — assuming homogeneity of space — implies look-back of 14 billion years or so.[267] (conventional view) Best estimates of the age of the universe based on the Big Bang assumption is some 14.7 billion years. It would seem that 0.7 billion years is a very short time for that observed galaxy to have formed. If even more distant galaxies are observed with larger telescopes to become available, then by the redshift correlation to homogeneous space one would be observing galaxies "before the beginning of the universe" — an obvious non sequitur. It appears that the redshift correlation to velocity and distance, as well as the homogeneity of space and in fact the Big Bang itself— are already in some difficulties and may be in serious trouble in the near future.

The empty Dirac sea 4-holes in the vacuum are special "Dirac negative action quanta" and the holes and their dynamics form negative action potentials, currents, and fields. This type of vacuum condition containing some empty Dirac sea holes is referred to as a *Dirac-polarized vacuum.*

However, we do not envision the Dirac-polarized vacuum as just simple *static* polarization, but as possessing rich dynamics such as negative energy currents, potentials, fields, waves, etc. including in the time-energy medium in the time domain. We also include the interaction of the dynamic Dirac-polarized vacuum with spacetime.

---

[267] I am indebted to Prof. Stanley Jeffers for this information, in his E-mail discussions with the Alpha Foundation's Institute for Advanced Study fellows. Dr. David Roscoe contributed the fact that the Big Bang theory has so many parameters that nearly anything can be "fitted in". But he also pointed out that W. M. Napier, *Astronomy & Astrophysics,* Vol. 310, 1996, p. 353 gives a very thorough analysis of the quantized cz phenomenon, which does present a difficult obstacle to Big Bang theory. Napier is **the** originator of the cometary catastrophe theory of biological extinction. One way or another, the theory of the Big Bang is in trouble, even though many theorists are still complacent about it.

Due to ongoing COP»1.0 processes involved in highly energetic astronomical phenomena, Dirac sea 4-hole regions and dynamics are believed to occur and exist in the vacuum across the observable universe. These occur especially in appreciably curved spacetime regions with large tempic potential gradients and thus large *time-forces* across the region.[268] The 4-holes and 4-hole currents in the vacuum are generated in great quantity in large cosmic explosions. Also, the energetic fluctuations of the local vacuum energy can separate a Dirac hole from its occupying electron by lifting the electron (pair production), and so can a curved spacetime. In a properly curved spacetime, fluctuation can lift an electron from a hole, and enter a stationary state by accommodating reaction of the local spacetime curvature. In that case, hole and electron go their separate ways. The hole and that or another electron may later recombine, or in sufficient local spacetime curvature, they can remain apart. We point out that every positive energy electron is accompanied by an appropriate local curvature of spacetime that prevents the accoutrement of a free Dirac sea hole,[269] and thus prevents the electron from disappearing.

The more energetic the local region, the more severe and violent are local spacetime curvatures in that region, and — in our view — the more frequent the occurrence of COP»1.0 interactions producing significant Dirac 4-hole currents, negative energy EM fields and potentials, etc. Indeed, the measurements that yielded the accelerating expansion of the universe were measurements of the relative velocities existing with stellar explosions known as type la supernovae. In our view, in explosions of such violent nature — particularly including in optically active gases such as where gamma ray bursters form — an appreciable fraction of the hole currents are transduced into positive energy electron currents. This produces — at least for a finite time — a net positive feedback, self-powering COP»1.0 phenomenon, resulting in very rapid asymptotic output energy increase. This is what we have hypothesized as producing the gamma ray burster itself. *Quenching* of this asymptotic rise in COP occurs only after the physical geometry of the "closed-looping" gas particles is broken a short time later. The maximum output energy is determined by a function of the *rate* at which the geometry is broken and the corresponding *rate* at which damping of the asymptotic rise and quenching occurs. The afterglow remaining in the gas is due to the extra positive energy produced in it by the transduction of negative energy into positive energy during the asymptotic rise of the self-powering burster, with excess positive energy being absorbed in the gas while the burster magnitude is rising. After damping and quenching, the gas has been heated (that is how the geometry was disturbed in the first place, to produce damping and then quenching). Re-ignition can occur whenever a new close-looping COP»1.0 situation spontaneously occurs. As in the gamma burster, this can often happen as the disrupted geometry of the gases stabilizes.

---

[268] We also conjecture that these regions may account for a tentatively observed slight change in the fine structure constant, but leave that to the advanced theorists to validate or falsify.

[269] This needs a bit of modification. Rigorously, the so-called "isolated" observable electron is clustered around by *virtual* positrons (Dirac sea 4-holes) in the vacuum. So a positron influence is present, though from virtual positrons instead of observable positrons. The magnitude of both the positive charge and the internal or "bare" negative charge is infinite. The difference, however, is finite and is the "observed" magnitude of the observable central charge. There is thus a net observed shielding of the bare charge of the electron, and this is taken into account to yield the observed charge of the electron.

All 3-space EM energy comes from the time domain, as previously developed. In the more violent positive energy COP»1.0 regions, there is an increased conversion of time-polarized EM energy into 3-spatial energy and back. On the average there is a higher "time-energy conversion" or "tempic" potential. In those regions, a tempic conversion potential is higher at the higher center magnitude of the spatial energy density. Hence there are strong tempic forces from "inside to outside", providing a "tempic broom" effect to sweep outward the excess Dirac sea holes separated from their electrons in the curved spacetime. The result is the presence of increasingly unoccupied Dirac sea 4-holes toward the outside of violent regions of spacetime,[270] until the effect is overcome by the decreasing spacetime curvature. The 4-hole current radiating out from an asymptotically increasing (exploding) burster provides an asymptotically increasing (exploding) negative energy and negative mass reaction with spacetime. That reaction explosively creates curved spacetime for negative gravity (repulsive gravity), acting upon the local system and upon distant masses as well. The result appears to be the cause of the acceleration of the expanding universe.

It follows that re-ignition of the gamma burster can occur if, as the quenching continues, the geometry stabilizes and reaches another "close-looping" positive feedback condition. In fact, just such re-ignition of the gamma burster is very often observed.

---

[270] I.e., the violent astronomical phenomena — while generating and increasing are more violent at the center, and less violent at the outside.

Therefore, we hypothesize that the more distant violent astronomical regions have greater amounts of repulsive gravity being produced by their local Dirac-polarized vacua particularly during their explosive asymptotic energy density rise. This produces violent streams or bursts of negative energy outflow, which is equivalent to negative mass outflow and which provides outgoing negative curvature of spacetime (antigravity). At sufficient distance from us, the increasingly earlier, more violent events and resulting greater curved spacetime regions have created sufficient cumulated Dirac-polarized vacua to override the normal attractive gravity between galaxies, etc. and produce net repulsive gravity. This results in observed distant acceleration of the expanding universe {581a-581f}.

9.9 A Solution to the Hole Current Problem and Close-Looping

Bedini and the present researcher have filed a patent application {582} on the above special capacitive process to *convert* or *transduce* the detrimental Dirac sea hole current reaching the system's input section into electron current, and put it to use in powering the system. This transduction process receives the hole current as a special potential, then regauges the "negative energy" potential into a "positive energy" potential. This regauging is accomplished by charging the capacitor with Dirac hole current and then discharging electron current from it. The transformed positive energy potential is then applied to the Drude electrons, producing an ordinary electron current in the input that can be used to power the system itself. If sufficient excess positive electron current is provided, then the system can "feed itself in self-powering mode and also feed power back into the main power supply line.

When conversion of the arriving hole current in the input section produces an output "positive energy" electron current equal to the system's normal power needs (Kron's condition), the external power supply can be disconnected and the system will remain powered by the steadily converted negative hole current arriving in the input section.

Our patent-pending process allows this otherwise devastating special form of "negative energy feedback" to be converted into "positive energy feedback", allowing the system to be close-looped successfully for self-powering of the system and its load. Regulators of the transduction process may be added to cover the situation when the load is erratic and changing.

At Magnetic Energy Ltd., we have also filed a patent application for an outrigger method of close-looping a COP>1.0 system, which greatly eases the problem. This method is covered in Paragraph 9-14 below.

9.10 Negative Energy: Localized and Nonlocalized

It sometimes appears there are as many views on negative energy as there are scientists who mention it. Simply put, the Dirac theory of the electron requires negative energy states. The negative energy electron state was envisioned by Dirac as a sort of "hole" in the ambient vacuum, where the hole could contain a positive energy electron. The combination of equal positive energy of the electron and negative energy of the hole itself, would give a "normal" vacuum zero state of the "filled hole". In other words, the normal or ambient vacuum consists of holes filled with electrons. Considered in that fashion, such a vacuum is referred to as the **Dirac Sea** or the **ambient Dirac Sea,** or the **ambient Dirac vacuum.**

If positive energy is added to a filled Dirac hole, the electron can be lifted out, leaving the hole. Note that adding positive EM energy, e.g., is actually the addition of spacetime curvature and dynamics to the Dirac sea vacuum. In short, it is a supersystem interaction, not a system interaction.

In that case, the curvature of the local spacetime interacts upon the electron mass, producing a force that lifts it from the hole. The result is the appearance of a positive energy, positive mass electron and a Dirac sea hole (a 4-positron) in a curved spacetime. That is known as *pair production,* and it is a supersystem interaction. When a positron and an electron meet, the electron falls into the hole, changing the Dirac sea which also changes the local ST curvature. That change in local ST curvature is a change in the energy density of space, and is normally radiated away as a photon or photons as the local curvature decays back to a flat spacetime. The emitted radiation may further decay into particles, etc. by several reactions well known in particle physics. That "conventional" set of supersystem interactions is known as *pair annihilation.*

If we deliberately change the ST curvature appropriately, however, the pair annihilation radiation need not be emitted since the excited local spacetime curvature does not decay back to a flat spacetime to emit it.

Dirac's original paper tried to identify a negative energy state vacuum hole as the proton, but later that was falsified and it was taken as the positron. There is a very strong aversion in physics to considering negative mass; hence, the mass of the positron was proclaimed positive (for many reasons not really connected with our thesis). Nevertheless, some physicists (in the minority) still speculate or theorize that negative mass "can exist." Not only "can" it exist, but also it *does* exist widely and one can demonstrate it in actual electrical circuits. It certainly exists in the binding energy of the nucleons in the nucleus of an atom. Again, a single white crow is sufficient.

We take a very straightforward view: we consider a positron — an electron observed with positive mass and "observed" to be traveling backward in observer time — to be a different (transformed) entity than a Dirac 4-hole traveling forward in observer time.[271] The observation operator accomplishes both velocity reversal and time reversal, and — since time is not an observable, the time reversal is "seen" as a reversal in spatial direction. *Prior to observation, neither reversal has been imposed.* The major distinction, we hypothesize, is that the observed positron has positive mass in 3-space, which is concentrated positive energy, and therefore produces positive energy electrical fields and positive gravitational field. Mass is an observable, hence, it is the iterative frozen 3-space snapshots of an ongoing 4-space process. Observation (and any observable) is the effect of applying the operator $\partial/\partial t$ upon spacetime LLLT in which a 4-space process is occurring. As we stated earlier in this book, our view is that no observable *continuously persists* (in time), but continually and iteratively *recurs* by the continual ripping of time from the ongoing 4-space reaction, to see the 3-space intersection at a frozen moment. We also advanced the *mechanism* that generates the flow of a mass through time, consisting of a mechanism for $\partial/\partial t\,(\mathrm{LLLt})$ - photon emission, converting *mt* to *m* — and the mechanism for $\int_t (\mathrm{LLL})dt$ — photon absorption, converting *m* to *mt.*

---

[271] This point is stressed. The *unobserved* positron is not identically a positive mass electron traveling backwards in time, but a negative mass, negative energy electron traveling forward in time. When Dirac theory of the electron was written, broken symmetry and the CPT theorem had not been discovered. Further, there was no distinction made in the difference between the observed entity (the *effect* of the observation interaction) and the non-observed entity (the *cause* of the observation interaction).

[272] We stress that we are once again confronting the ubiquitous substitution of the effect for the cause in classical electrodynamics. The observed positron and the unobserved positron are two quite different entities —just as the "cause" of an observation is quite different from the "effect" of the observation.

We may also consider mass to be condensed 3-spatial energy (say, EM energy) compacted by c^2. Modern physicists tend to consider the mass-energy itself as the source of the gravitational "attraction of mass" rather than the mass itself {583}. Let us accept that assumption, but differentiate between the observed 3-positron with positive mass and opposite direction in positive observer time, and the unobserved Dirac sea 4-hole with negative mass and without change of direction in the observer's positive time. The 3-positron is an *observed effect,* while the unobserved Dirac sea 4-hole is an *as-yet-unobserved cause.*

We also surmised that all EM energy in 3-space comes from time, by our solution {584a} to the long-vexing source charge problem {68}. In our view, we may then consider that the positive energy compressed as 3-space mass has been extracted from the time domain of surrounding spacetime, via entry of that time-energy into the surrounding 3-space at each 3-space point, producing a surrounding spacetime curvature which is known as the *gravitational field.* This gravitational field represents "withdrawn" 3-space energy from surrounding space, and withdrawn time-energy from the time associated with each surrounding 3-space point.

Hence the gravitational field may be said to consist of "negative energy" since it represents *lowered* energy of spacetime due to the curvature. This viewpoint yields results consistent with the received view that the energy of the gravitational field is negative. In our view, it is negative precisely because it represents "withdrawn energy" from the normal energy density of that otherwise flat space. The withdrawn nonlocal energy has been localized (as what we call "mass"). Simultaneously that withdrawn 3-spatial energy also represents "withdrawn energy" from the time-domain, since all the withdrawn spatial energy came from the time domain in the first place. The result is the curvature change in spacetime, that is taken as the gravitational field.

The 3-space energy comprising the mass is positive because it is *added to* the normal energy of an otherwise flat spacetime, at that concentrated locality.

The "energy in the gravitational field" is negative because it represents lowered energy density (i.e., lowered "pressure" in a fluid analogy) of that surrounding nonlocal spacetime compared to when that spacetime was flat

Thus when we add positive energy (a positive energy state) to 3-space, we create normal "positive" gravity because "mass" is a process of withdrawing some positive energy from the energy density of surrounding 3-space.

Here is the shocker: When we add localized negative energy (a localized negative energy state or "4-hole" — or a localized negative mass!) to a local 3-space region, we create "negative" gravity (antigravity) in surrounding space because the withdrawn energy from the local 3-space "hole" must have been added to the energy density of the surrounding spacetime, once we apply conservation of energy. Or in other words, the surrounding spacetime changed (curved) in interaction to the production of the hole.

If we wish, for simplicity we can just consider localized positive energy to generate positive gravitational charge in the surrounding nonlocal spacetime. We can consider localized negative energy to generate negative gravitational charge in the surrounding nonlocal spacetime.

This is where our concepts differ from the rather arbitrary way that a Dirac sea hole is interpreted in the received view. In spacetime, we try to account for the entire energy process and we apply a conservation law by including the supersystem interaction in every case. We insist that positive energy concentrated in positive local mass must have been extracted from surrounding spacetime.[273] That "negative energy" in the surrounding spacetime is what is known as the *gravitational field of the source mass* (where the energy was withdrawn to). Similarly, we insist that positive energy extracted from a localized region of the ambient vacuum to leave a "Dirac sea hole" must have been spread out into the surrounding spacetime, where it would properly be known as the *antigravitational field of the source Dirac hole.*

As an analogy in very simple lay terms, the holes associated with a local mass system or region involved in COP>1.0 operations create higher "pressure" outside the system or region with respect to other objects. Higher pressure between any two objects pushes the objects apart. The masses associated with a mass system create lower "pressure" outside the system or object with respect to other objects. Lower pressure between any two objects results in the objects being pushed together by the higher pressures beyond them.

---

[273] Alternatively, if we add local spacetime curvature, the surrounding nonlocal spacetime interacts as mediated by the supersystem changes in all three components. What we are saying is that there is no such thing as just a system, vacuum, or spacetime change. Any change to one is accompanied by a change to the other two, *a priori.*

We submit that this new view provides an exact (and testable) antigravity mechanism. It appears to have been validated by the Sweet antigravity experiment {576}, which was designed in advance by the present author with increasing the local negative energy reaction in mind.

9.11 Tests Reinforcing the Antigravity Mechanism

9.11.1 Dirac Sea Hole Current Effects in a COP»1.0 EM System.

For COP»1.0 electromagnetic systems, we recall that the temporal energy will flow negatively across the system from output section to input section. Hence the Dirac sea hole current in the local vacuum flows back across the system, from its output to its input section. Regarding the output section: Specifically, *all COP> 1.0 systems produce both positive and negative energy outputs. The greater the COP of the system, the greater the fraction of negative energy produced with respect to the fraction of positive energy produced.*

The positive energy will flow on out of the system output section (as into the external load) as in a normal circuit, but the negative energy output will flow from the output section through the local vacuum back across the system to its input section.

At the input section, the arriving Dirac holes appear to the input electrical current as a special kind of "load" added to the system input section. Hence, as previously stated, the external source must input electrons (as electron current) to "fill the holes" prior to having any input electrons available to power the circuit in normal fashion.

That is nature's unexpected *overunity decay mechanism* for COP>1.0 EM systems. The COP>1.0 system is in a condition of disequilibrium *a priori,* since the system exhibits negentropy and only systems in energy flow disequilibrium with their active environment can do that. Any COP»1.0 EM system is therefore in a highly excited state, and all excited states in nature have decay mechanisms. A highly excited state usually has a prompt decay mechanism, and the COP»1.0 system is no exception.

This emergence of the Dirac hole current from the system output section back to the system input section, is nature's decay mechanism to trigger COP»1.0 EM systems back into COP <= 1.0. With this current, a "negative energy load" is added to the input section, causing the system to draw additional electrical power from the external power supply to "power" the killing of the Dirac holes[274] as well as to "power" the normal circuit (the system itself).

If the external power source connected to the COP»1.0 system is not furnishing sufficient electron current to kill the hole current in the input and power the system as well, the excess hole current will pass back through the conductors and on into the external power source, "eating current" and power as it goes. The product of the hole current and the voltage gives a negative power imposition, wherever the hole current interacts with the power system and its external power supply, even back into the distant huge power generator in the distant power plant. For a battery-powered circuit, e.g., excess hole current will pass into the battery and discharge it (change its chemistry in discharge mode) more quickly.

Indeed, a lead-acid battery (and some other kinds as well) can be supercharged with Dirac holes. What this actually means is that the local vacuum — in which the battery and its components and chemistry are embedded and in a continual exchange with — is altered to increase the number of empty Dirac sea holes devoid of electrons. At the same time, the local spacetime is also negatively curved. The end effect is that the battery is charged with negative energy, rather than positive energy. We propose that a battery or capacitor possessing an extreme negative energy charge should be lighter in weight than the same capacitor when uncharged.

As previously stated, some years ago, Jim Watson developed a very large multi-kilowatt COP>1.0 battery-powered generator, publicly demonstrating it at a technical conference (see again Figure 5-12). Engineering measurements at the conference verified that the batteries were constantly charging while the load was constantly powered at 8 kilowatts.

The longer Watson's system would run, the more charged his batteries became with *negative energy.* After one of his batteries had been used for a couple of weeks in extensive experiments, one could obtain a most novel effect when it was placed on a normal "positive energy" battery charger. The battery's excess "Dirac hole charge" or "negative energy charge" accumulated in its local vacuum would "eat positive power and electrons" from the battery charger for an extended period, sometimes as long as a week, with no symptoms of "charging" the battery at all (no increase in battery voltage at all). During this period, the battery charger was indeed recharging the *battery supersystem.* It was filling the excess Dirac sea hole charge in the vacuum with the steadily supplied electrons. In the process, it was gradually relaxing the previously formed negative spacetime curvature.[275] When all the Dirac holes were filled, the relaxed spacetime then was relatively flat (normal) as was the local vacuum, and suddenly the battery would charge-up normally again with positive energy, with its voltage increasing in quite normal fashion.[276]

---

[274] By converting separated incoming electrons and incoming holes to resulting filled Dirac sea electron holes, in the input section itself.

With tongue in cheek, we point out to the COP>1.0 researcher that these effects are not described in the available storage battery manuals, no matter how technically they are written. To the electrical engineer, we point out that they are not in his textbook.

Whenever Dirac sea holes have accumulated in the input area of the COP>1.0 system and are being filled by the external power supply, normal ammeters and voltmeters will only "see" that the external power system is furnishing more current. That is true, because the holes are an additional load being powered. The holes "eat incoming electron current and power", and require the external power supply to "power the killing of the holes" as well as "power the system".

The unwary overunity experimenter can successfully have the first phase of an overunity system working and not realize it, because his initial COP is too high and he does not know about the associated hole current decay of overunity systems back to underunity systems. His system may actually be at COP>1.0, but the negative energy component being produced in his output section is back-flowing and continuously degrading his input section by adding a "new load" right in the input section.

---

[275] This proves that a system's supersystem can indeed be negative-energy-charged, so that the local spacetime around that system is negatively curved, even though the system itself seems normal and functions normally. If the negative energy charge is . sufficient, conceivably the antigravity produced will be sufficiently great that the system itself "floats" or even levitates. We hypothesize that an intense negative energy beam can indeed be technically created which will accomplish this "negative-energy-charging of the supersystem" effect when continuously absorbed by a targeted system's local spacetime supersystem component, as the system is continuously irradiated by the beam. In that manner, levitation beams, tractor beams (both attracting and repelling) are at least hypothetically possible. If this proves to be a valid hypothesis, we foresee the day when such beams will revolutionize transportation and handling of otherwise heavy loads and very heavy systems. Even the heaviest vehicle could then negatively charge its own supersystem's local curved spacetime, thereby floating in the earth's gravity, in the atmosphere, in space, and under the ocean.

[276] We deliberately omit discussion of the extraordinary all-purpose weapons implications of negative energy electromagnetic pulse (EMP) directed energy weapons, already developed and deployed by several major nations.

So his overall power draw from the external power source can be such that his positive power output in the load is less than his total power draw at the input of his system. Unless he is aware of the process and how to handle it, he may never realize that his system is actually at COP>1.0 condition, when the negative energy component is credited as an output "plus".[277] *COP>1.0 research starts by evaluating the absolute value of the output energy, then working from there!* The experimenter must be ever aware that he can and will have Dirac sea 4-holes and currents in any unitary COP>1.0 electrical power system, and particularly in those with very high COP. He must understand what the 4-hole currents and the negative energy actually do and how they function in the overunity system unless corrected.

Any COP»1.0 EM system demonstrates the Dirac sea 4-hole current phenomenon as nature's special decay mechanism to reduce overall performance of the COP»1.0 power system and its external power supply to COP < 1.0 — unless one *corrects* the problem and prevents the decay process (developed later in this chapter).

One symptom of the negative 4-hole currents is their cooling capability rather than heating. Another is that the excess current (and power) drawn from the external power supply also does not heat the input section of the system. Another is that semiconductors operating within the Dirac 4-hole current flow are likely to fail immediately and permanently.

We strongly suggest that, for high COP systems experimentation, the experimenter use a system of synchronized clocks where the lab clock is synchronized with a remote clock prior to the experiments, then the two clock times are examined after significant experimentation at COP»1.0. If deviation beyond normal error in synchronizing occurs, then the involvement of spacetime curvature effects is indicated. Two decades ago, anomalous experiments by Golden — with whom I was working at the time — charged the area with negative energy, producing a significantly curved local spacetime which disrupted local clocks and watches for four days as the charge gradually decayed away. Another indicator is that Geiger counters and other nuclear radiation detectors often "read" when bathed in vacuum 4-hole current, because of the ionization changes induced in them and also because of their individual previous "time-charge" histories. After all, ionization instruments "read" anything which causes ionization of their internal gases. Dirac sea hole current, if interacting with the gas, is in fact a very special new form of "ionizing" radiation.

---

[277] In other words, in the presence of both positive and negative energy output, true COP is not "positive energy out" divided by "positive energy input by the operator", but is more appropriately "total absolute value of the energy output" divided by the "positive energy input by the operator".

9.11.2 Antigravity Hole Current Effects in a COP»1.0 EM System.

Any COP>1.0 EM system can exhibit minor, usually barely detectable antigravity effects due to the presence of the negative local energy 4-holes in the hole current if the hole current is not transduced or corrected. It can also produce minor antigravity "thrust" due to the flow of the local energy holes. Usually these effects are not even noticeable in an overunity system unless it has very high gain (very high COP, say in the range of 10^5 to 10^6 or more.).

However, if the EM system's COP » 1.0, and reaches 10^5 to 10^6 or more, the antigravity and thrust effects may become significant and easily detectable. At 10^9 they become highly significant and are very detectable: The unit will lose appreciable weight on the laboratory bench, and may even levitate easily.

An example is provided by the Sweet vacuum triode amplifier discussed above, with the results that were given in Figure 6-9. Such effects are also what generate the anomalous levitation phenomena of free-spirited John Hutchison.[278]

9.11.3 Negative Energy Effect When Leads Shorted

When the leads of a COP>1.0 EM system are shorted, a noticeable cooling effect can sometimes be detected. If COP»1.0, shorting the leads will result in instant icing from the water vapor in the surrounding air.

If one accidentally shorts oneself across the leads of a COP»1.0 system, the effect is to freeze the nerves along the pathway of conduction. Experimenters are cautioned to use every safety precaution. Their experimentation is at their own assumed risk.

---

[278] John Hutchison, "The Hutchison Effect apparatus." *Proc. Intl. Symp. New Energy,* Denver, Colorado, May 12-15, 1994, p. 191-198.

Anecdotally, some older researchers reported that their constant handling of overunity circuits resulting in charging their bodies, apparently with negative energy. The body after all may be considered as a special kind of complicated capacitor or set of capacitors, so this is reasonable, based on Bedini's and Watson's demonstrated negative energy charging of a battery. Some of the older overunity researchers reported that, after charged in that manner, they could produce a 3-inch spark from their fingers, or a "slight shock" into the body of a person they touched.

There is even a special form of acupuncture in the East in which the practitioner uses special exercises to charge the body with negative energy in this fashion. The "electro acupuncture" is then administered by the operator twirling the inserted acupuncture needles between his fingers, and deliberately administering a jolt of negative energy from his body into selected points in the body. I have personally tested a person who could deliver strong "electric shocks" into one's hand or body, and who could consistently produce a 1-inch spark discharge between the hands at will. That person could also deliver a very astonishing "jolt" to one's body, particularly in the vicinity of the liver.

There appear to be similar rare negative energy phenomena in semiconductors {585a-585c}. Anecdotally, it appears that the body does have structures or methods of storing and using Dirac sea holes (negative energy) and mechanisms for accessing the negative energy charged in its supersystem. This is suggested as a fruitful line of future research for interested biophysicists. Again, we call attention to the fact that, whenever EM fields, potentials, and waves are present, Dirac sea effects are also present and involved. Little or no biophysical experimentation has been done utilizing higher symmetry electrodynamics and analysis of the supersystem effects. Contrary to the received view, the understanding of EM effects on biological systems is not mature at all, but is still in its infancy.

9.12 Susceptible Unitary System without Conversion of Dirac Sea Hole Current

A *unitary system* is simply a self-contained single power system with an overall closed current loop between the source dipole and system loads and losses. A unitary system is contrasted to an array system of synchronized separate power subsystems without a common closed current loop, with the array subsystems connected to the source dipole by energy transmission-reception only. Hence the array system utilizes the Kron open path concept.

Beginning at about COP = 1.2 to 1.5, a few highly susceptible unitary COP>1.0 electrical power systems will experience significant Dirac hole current from output to input. This will produce a "negative energy load" in its input section and feed line from the external power system, resulting in decay back to an overall "system plus negative energy load" of COP<1.0. At the COP = 2.0 level, in the susceptible system the effect will forcibly decay the COP back down to overall COP<1.0. It does this by forcing excess power to be furnished by the external power system to "power" (fill) the Dirac hole current arriving in the input section of the power system. In short, it converts the system to a composite system containing an additional Dirac negative energy load in its input section.

Or in other words, it forces the operator to "pay" for both the *positive* energy output and losses of his system and the *negative* energy output and losses of his system.

In addition to usual powering of its load and losses, such a system is also "powering" the return of its own locally altered vacuum back to a more normal ambient vacuum, instead of stabilizing its supersystem in a stationary disequilibrium state. It is a supersystem interaction and effect. A COP>1.0 system that significantly exhibits this phenomenon is said to be a "susceptible" system.

One of the first tasks of the COP>1.0 experimenter is to ascertain whether he is working on such a highly susceptible COP> 1.0 system. Ironically, he does this by observing his COP>1.0 system kick itself back into a COP<1.0 system. If he carefully observes the initial COP>1.0 condition followed by decay into COP<1.0 condition, he must then realize that not all is lost! In short, he must realize he has developed a susceptible system. If he obtains COP>1.0 only to watch it subsequently decay back to COP<1.0, he has just discovered he is working on a susceptible COP>1.0 system, *and that is experimental progress.* He isn't there yet, but he is en route. In that case, he must concentrate on transducing the excess hole current into usable electron current, to "stabilize" and "lock" his system in its COP>1.0 operating condition.

9.13 Unitary System with Conversion of Dirac Sea Hole Current

Using the Bedini conversion process, a unitary system can be close-looped to maintain COP> 1.0 and stable overunity operation. The 4-hole current reaching the input section is intercepted by the Bedini mechanism, and transformed to ordinary electron current furnished to the input section. Thus the transduced negative energy feedback becomes a positive energy feedback, by conversion. This converted positive energy in the input section will reduce the demand for incoming electron current from the external power supply, increasing the COP>1.0 even more.

A feedback and governing system is adjusted until the system's converted positive energy feedback current reaches the Kron condition. At that point, the transformed electron current is sufficient to power the system with the external power supply disconnected. The governing and control system thus disconnects the external power supply, and the system is smoothly transitioned to self-powering operation, powering both itself and its load.

We stress that all the energy to power the system and its load is now being furnished from the active local vacuum interaction and from the local curved spacetime. No laws of nature, physics, higher symmetry electrodynamics, or thermodynamics are violated. Conservation of energy rigorously applies, but to the supersystem. The system is an open, *stabilized* disequilibrium system — of the kind referred to by Kondepudi and Prigogine as nonequilibrium stationary states — as a freely receiving all its energy from its active external environment. No standard U(l) electrodynamic analysis will show this condition or this mechanism, and neither will electrical engineering. Higher group symmetry electrodynamics such as SU(2) or O(3) will show it easily. For a modern treatise on O(3), we refer the reader to the recent *tour deforce* by Evans {586}.

9.14 Non-unitary Outrigger Array System

See Figure 9-14, which is a diagrammatic representation of a COP>1.0 array system that is non-unitary. In the non-unitary (outrigger array) system, no part of the outrigger array works at $\xi > 1.0$. All external parts operate at $\xi < 1.0$, but independently and with COP>1.0. Each of the external outrigger systems is an open system freely receiving excess energy from its active environment, because it is an "energy-receiving" system and the energy is freely transmitted from the central system. The central part of the array is the unit whose purpose is to stimulate or evoke an excess energy flow from the vacuum, so that the outrigger receiver/power units can individually intercept and utilize radiating energy flow surrounding the central unit.[279] The usable (output) energy intercepted and captured (collected) from a given free flow of energy depends totally on the intercepting reaction cross section — usually the total array amount of intercepting charge, together with the degree of resonance (sweeping a larger geometrical cross section) of the intercepting charge.

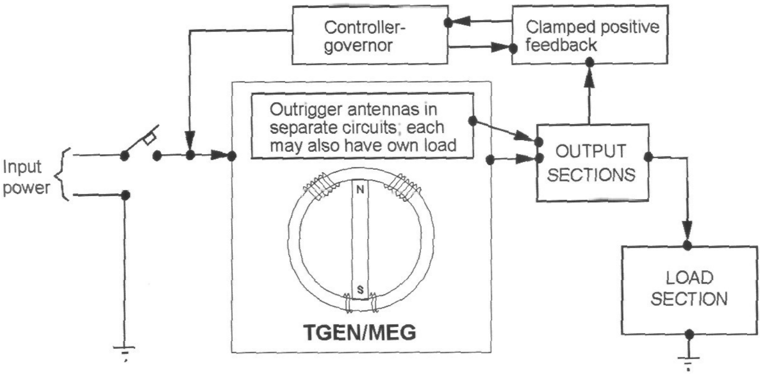
Controller-governor
Clamped positive feedback
Outrigger antennas in separate circuits; each may also have own load
OUTPUT SECTIONS
Input power
N
S
TGEN/MEG
LOAD SECTION

Figure 9-14 The nonunitary (outrigger array) system and its functioning.

Each outrigger unit (subsystem) has its own complete closed current loop *separate* system and load, plus its "antenna" or "interception section" or receiving component for intercepting and collecting the free flow of energy coming from the central source unit (the energy transmitter). Thus, the loads in the outrigger units are not closed-current-connected with the central source unit at all. Each outrigger unit has its own independent electrical ground and its own closed loop system with its load.

---

[279] We emphasize that, when the usually unaccounted Heaviside nondiverged energy flow component surrounding the conductors of the external circuit is accounted, every EM power system is already a $COP \gg 1.0$ EM energy transduction system. Consequently, in the received view the array system can be operated at $COP>1.0$, or even in self-powering mode, even though the central system is "said" to be at $COP<1.0$. It is at $COP<1.0$ only because of the self-applied Lorentz symmetrical regauging and ignoring the excess unused energy flow available to and surrounding the circuit, but not conventionally used. If we account the COP for what we have to furnish to the central unit, then the COP of each array unit is $COP = \infty$. The point is that we ourselves do not have to pay any *extra* input energy for the output energy of each array circuit.

This schema in fact separates the two functions of Poynting energy flow theory: (i) the extraction and radiation of the EM energy flow from the seething vacuum, and (ii) the interception, collection, and dissipation of some of that available energy flow, to power the loads and losses of the intercepting system.

In this fashion, multicollection from a single energy flow (generated from a central source "transmitter") occurs by direct, independent interception and collection in each of the COP<1.0 outrigger units. However, the input power to each outrigger units is *free* with respect to that outrigger circuit itself. We of course must pay to generate the flow of energy from the central unit, but that can be done in normal COP<1.0 manner (such as COP = 0.95) while powering its load.[280] Let the output power of each numbered unit (including the central source unit) be given by P_i, n be the number of power units, and P_T be the total power output. Then the total output P_T of the array system is given by

$$P_T = \sum_i P_i = P_1 + P_2 + \cdots + P_n \qquad [2]$$

The total input power P_I is simply the input power $P_{I(1)}$ to the central source unit. The COP is given by COP = $P_T/P_{I(1)}$.

In the motionless electromagnetic generator (MEG) array, for example, the central unit is the transformer-like unit, deliberately using a special core to contain the magnetic flux of a permanent magnet in contact with the core material, thereby stimulating the Aharonov-Bohm effect and freely filling the surrounding space with the uncurled magnetic vector potential A. We will explain the non-unitary (outrigger array) system using the MEG as an example.

---

[280] We again stress that, from any nonzero potential ϕ, or from any EM field in space, any amount of EM energy can be collected if there is sufficient intercepting and collecting charge. For example, the simple equation $W = \phi q$ applies for the potential, where W is the energy collected from a potential with point intensity ϕ by collecting charges q. A similar equation applies with respect to collecting as much force (emf in the various outrigger circuits) as desired. Here a single given E-field E can be assumed, and $F = Eq$. What is happening in the outriggers is that they are intercepting and collecting some of the excess Heaviside energy flow component that is wasted by the central circuit.

The MEG freely induces the Aharonov-Bohm effect,[281] thereby extracting excess energy from the vacuum in the form of that novel field-free A-potential. Being field-free, A represents an energy flow without swirl (without curl). When the MEG input section is perturbed, the surrounding A-flow is perturbed also, giving a pulsed or AC flow of energy that can be intercepted and utilized by each outrigger circuit as an E-field energy interception.

The central unit's $COP<1.0$ function is to output almost as much, or as much, energy as is required to power the MEG, while also evoking that excess A-flow and serving as a $COP \gg 1.0$ energy flow transducer. The function of each outrigger is then to intercept some of the pulsed A-flow (as E-field energy via the equation $dA/dt = -E$), in its antenna or interceptor section. Each outrigger has its own independent and conventional closed loop circuit. It simply is a receiver and user of extra Heaviside EM energy produced from the central MEG unit in the new form of the curl-free A-potential. Perturbing the A-potential produces a large E-field impinging on all the external outrigger subunits. Hence these units freely collect and use some of that available excess energy. The COP of the total array is thus $COP>1.0$. For best results, tuned antenna theory and near field theory should be utilized.

Suppose we operate the MEG central unit at $COP = 1.0$. In that case, there is no 4-hole current of significance associated with the central unit or the outriggers, and transduction of hole current into electron current for close-looping is not necessary. Instead, sufficient outriggers are employed so that the total energy W_O caught and dissipated in their loads by all the outriggers is given by $W_O \geq W_I$, where W_I is the input energy in the main central section. Since $COP = 1.0$ for the central unit in isolation, the total input E_{IN} is $E_{IN} = W_T$ (the additional huge Heaviside nondiverged EM energy input component is there also, but just neglected in conventional engineering). The total output W_A of the entire array is $W_A = W_O + W_T$. The overall COP_A of the entire array is thus

$$COP_A = W_A/E_{IN} = (W_O + W_T)/W_T = (W_O/W_T + 1.0) > 1.0 \quad [3]$$

---

[281] The reader should recall that the Aharonov-Bohm effect is not included in classical U(1) electrodynamics. See Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," in A. Lakhtakia (ed.): *Essays on the Formal Aspects of Electromagnetic Theory,* World Scientific Publishing, River Edge, NJ, 1993, p. 6-86.

Hence the array has a COP>1.0, even though every single component of the array has $\xi < 1.0$. With the outriggers fully functioning, many schemes of operation can be used. E.g., the energy from the outriggers can be used to steadily charge a capacitor bank by "shuttling". We define shuttling as use of two capacitor banks: (1) a capacitor bank which is continually connected and charged by the outriggers as their effective output, and (2) an input capacitor bank for the central unit that is continually connected and recharged from the outrigger capacitor bank. By adroit switching, the charge of the input capacitor bank is maintained by replenishing from the output capacitor bank, and the input capacitor bank powers the entire system, including the output load in the central section.

The non-unitary outrigger array construction is one means of close-looping a non-unitary overunity power system without bothering with the transduction of hole current into electron current. It is the easiest of all COP>1.0 systems to close-loop for self-powering.

An additional adaptation is for each outrigger also to continuously power a separate load, and only contribute *intermittently* along with the other outriggers to charging the output capacitor bank.

Another adaptation is to operate the central unit in the array in solitary self-powering mode.

These and similar modes of nonunitary array system operation allow close-looping the system for self-powering of itself and its load, without concern **for** 4-hole current transduction. A real advantage of such nonunitary systems without worrisome 4-hole current is that transistors can readily be used for switching etc. in them, without the destruction of the transistors by a damaging Dirac 4-hole current in the supersystem.

The nonunitary overunity system can also be close-looped by *conventional* (simple) "clamped positive feedback" means if good isolation (e.g., optical isolation of the circuitry switching) is used in the feedback loops.

9.15 Some Instrument Considerations

9.15.1 Time Effects

As previously indicated, there are time-domain effects (potentials, forces, and currents) associated with a COP>1.0 EM system, particularly as the COP increases toward COP»1.0. A useful method of indicating time-effects is to utilize a transistor in a circuit capable of measuring recombination time in the transistor.[282] NASA has such circuitry available, but I do not have the reference at hand. We leave it as an exercise to the reader to explore that possibility.

For "very large COP" devices, ordinary watches, clocks etc. may be affected. Particularly significant is the slowing of timepieces in the presence of substantial time-domain effects due to substantial negative energy. In addition, for high gain devices (and for some at lower COP), the negative energy charge can build up in the local area, and may then require some days to dissipate. As we stated, in experiments years ago with Frank Golden, we experienced just such alteration of multiple kinds of clocks and watches in an area. It required four days for the negative energy charge in the area to gradually drain off (for the excess Dirac sea holes to gradually be filled with electrons, as the tempic potential gradually returned to normal).

Effects that were even more dramatic were exhibited in experiments conducted by Sid Hurwich in Canada some years ago {587}. In a series of experiments lasting a half-hour, Hurwich inertially jammed a police revolver so that a human finger could not pull its trigger. *The watches of the witnesses did not change their setting during that entire half-hour.* This cannot be understood by conventional electrical system analysis, but potentially it can be understood by supersystem analysis.

9.15.2 Calorimetry Is Taboo for Overunity System COP Measurement

It is standard dogma that calorimetry is a "true" measurement of power, and is always accurate in the hands of a skilled practitioner. In general, that is true for a COP<1.0 system, where one is concerned only with *positive energy* in the input and output sections. However, with respect to COP>1.0 systems, nothing could be farther from the truth. Indeed, for a significant COP»1.0 system, a calorimeter is one of the most inaccurate instruments that can possibly be used.

As previously stated, all COP>1.0 EM systems produce a mixture of negative and positive energy in their output section. The higher the COP, the greater the percentage of negative energy and the greater the flux of Dirac 4-holes back to the input section. The output energy, immersed in a calorimeter, will both cool the water with its negative energy output fraction and heat the water with its positive energy output fraction. Hence, the calorimeter will show the *difference* between the water's simultaneous heating and cooling. If the calorimeter measurement is misinterpreted as the *total energy output*, this "definitive test" will always erroneously show COP<1.0. After all, it itself already applies the decay mechanism for converting any overunity system to an underunity system.

---

[282] This suggestion is courtesy of John Schnurer some years ago, and is appreciated.

The only good use of the calorimeter in an overunity EM system is to verify that negative energy is present, when comparing its measurements to more sophisticated electrical measurements of the system. It can be used along with other measurements to ascertain (1) the total energy output, whether negative or positive, (2) the difference in the two forms of energy output and the algebraic sign of the resultant (positive if calorimeter heated, negative if calorimeter cools), and (3) thereby help determine the output's energy fractions (positive energy divided by total energy, negative energy divided by total energy).

Any test group insisting on testing a purported COP>1.0 electrical system with a calorimeter as the *definitive statement of energy output* is totally devoid of knowledge of COP>1.0 systems and their phenomenology. Usually such "test groups" tend to regard themselves as "measurement experts" (which they well may be, in COP<1.0 positive energy systems!). That does not make the team even minimally knowledgeable or qualified in COP>1.0 system measurement. To the contrary, the team members have zero experience or knowledge of COP>1.0 systems. The proposed test group may consist of physicists, thermodynamicists, electrical engineers, technicians, or all four, but they still will have no knowledge or expertise in COP>1.0 EM power systems unless they have actually worked with such systems before, and at length.[283] This is particularly true of COP»1.0 systems. My advice to all COP>1.0 researchers is simply to keep the "calorimeter addicts" out of one's laboratory in the first place.

9.15.3 Other Instrumental aspects.

Here we just point out some simple and obvious things, familiar to the experienced researcher but sometimes not appreciated by the novice. One is usually dealing with nonsinusoidal waves, pulses, spikes and the like. As is well known, RMS meters are useless for measuring such nonsinusoidal electrical entities, since they are designed and calibrated to measure sine waves only, or to measure DC. One must use a good differential sampling scope with multiple synchronized channels, with professional software to perform accurate integration under the curve, and with special differential probes for the scope. Simultaneous triggering of the multiple measurement channels on the same time line is required. The necessary high quality probe set may cost as much or more than the oscilloscope!

---

[283] To date, we have found none who have actually worked with COP>1.0 electrical systems. We have, however, found quite a few who, though totally inexperienced with COP>1.0 systems and not knowledgeable, have regarded themselves as "real experts" in such systems. They claim to already know how COP>1.0 systems work and behave, and assume just ordinary electrical system phenomena.

Dirac sea hole currents are generally not separately measurable, so one measures where the hole currents are not significantly present. Results may be compared to measurements where the hole currents are known to be present. When directly measuring hole currents, the conventional meters often read "backwards" and interpret a negative energy output at the system output section as subtracted from the positive energy, and interpret a negative energy output at the input section of the system as additional input electron currents from the external power supply. If the electron currents from the external power supply are not present, then measuring hole currents will draw electron currents from the power system inside the meter itself, resulting in a "backward measurement". Again, the meter actually will measure a real electron current, whether from the external power supply or from a conversion within the meter itself.[284]

Recombination time in semiconductors can sometimes be used in instrumented circuits to differentiate between negative hole current and electron current, by observing the change in recombination time due to the negative energy currents. It will differ from the change in recombination time due to positive energy currents.

Again, what is needed in the field is a set of solid, reliable instrumentation specifically developed for these peculiar measurement phenomena involving both positive energy and negative energy. To my knowledge, no such thoroughly designed and tested instrument package presently exists. Further, there are not even any standards for such, since there are apparently no standards for negative energy measurement at all.

9.16 Still Anomalous Aspects

Several aspects of COP>1.0 systems are still not properly understood. The behavior of curl-free A-potential, sometimes produced in such systems, is a case in point. This type of energy can exhibit characteristics as if it were electrostatic energy, which my associates sometimes refer to as the "crawly stuff. It exhibits the characteristic of "crawling" all over the surfaces (outside and inside) of the objects in the area, including the scopes, producing internal electrostatics and possibly magnetostatics everywhere. This can easily destroy the amplifiers etc. in a good scope or an expensive meter. It is best to transform this field-free A-potential into an ordinary "EM field energy AC wave", and then utilize it in that fashion.

---

[284] What is very much needed is a meter of sophisticated design where this "drawing of excess electron current from within the meter itself has been taken into account in the design, so that the meter functions correctly for either positive energy measurement or negative energy measurement.

Scalar stress waves are also sometimes produced, leading to anomalous effects. This is best regarded as a "time-polarized" EM wave, or more accurately, as a combination of scalar (time-polarized) EM waves and longitudinal EM waves in 3-space. For macroscopic waves in COP>1.0 systems, the Mandl and Shaw 1:1 correlation between scalar photons and longitudinal photons can apparently be corrupted, so then one has "abnormal" (improperly correlated components which normally correlate to make "ordinary" EM waves). These abnormal EM scalar stress waves also exhibit the "crawly" phenomenon, and can be hazardous to one's instruments.

In addition, some instruments such as ionization counters, Geiger tubes, etc. read in the presence of such waves as if nuclear radiation were present (which it is not). Also, the individual time-charge history of a given instrument determines or substantially influences how the instrument reacts in the presence of the "corrupted" scalar stress waves. Essentially, at a given time any instrument has a small "time-energy charge" depending upon its past history. The corrupted scalar stress wave accomplishes further "time-charging" of the instrument, which combines with the initial time-charge to produce more time-charge, ordinary EM waves by scalar EM interferometry {588}, or abnormally correlated longitudinal EM wave emission.

In an ionization-measuring instrument, *ionization is ionization, by whatever means achieved.* And in the presence of the "abnormal" radiation and charging, ionization-type "nuclear radiation" instruments will vary from one to another, *depending upon the individual time-history ofeach instrument.* So from a group of identical instruments made the same day on the same assembly line, one may read anomalously and several others read normally. Or two may read anomalously, while all the rest read normally. A very good example of this anomalous measurement effect is seen in the rigorous electrolyte experiments in U.S. Navy research facilities at China Lake {589}. The present author has previously given an explanation of this time-charging and decay phenomenon {590}.

Negative energy reacts in different ways with different materials. The more nonlinear the material, the greater the effects and unpredictability of the interaction, as a rule of thumb. In a copper wire, Dirac hole current introduced into the wire will "eat" Drude gas electrons, reducing the "pressure" at that point or area in the wire. This "combining of electrons and causal Dirac holes is not normal pair annihilation, and radiation does not occur. Instead, the two curvatures of spacetime (one by the positive energy of the electron and the other by the negative energy of the Dirac hole before observation) interact and adjust.

The Drude gas in the copper further back toward the external power system still has normal (higher) gas pressure, and so the differential in pressure induced in the Drude gas will drive electrons from the "normal" pressure areas into the "reduced" pressure area where some of the Drude electrons disappeared after being "eaten" by the Dirac holes. To look at that in lattice hole terms, the "eating" of Drude electrons in a selected volume of the wire leaves positive ions, which are "lattice holes" vacated by the eaten electrons. So these excess positive charges attract electrons from further up the wire nearer the distant power supply, filling the lattice holes to bring the Drude gas back up to "normal". However, that also draws and "eats" additional power from the distant power supply. Meanwhile, the Dirac 4-holes fed from the input of the COP>1.0 system back into the power feed conductors continue to interact with the lattice and eat electrons to create more ions (lattice holes).

One may visualize the Dirac sea holes producing excess lattice holes, which then causes the Drude electron gas to change so that the "average" demand current increases. In short, we must also consider the effect of Dirac hole charges or currents upon the lattice material itself, and the structure of holes and electrons as charge carriers in the material lattice.

Semiconductors are very vulnerable to negative energy due to its disruption of their donor-acceptor and band gap design functions. With mild negative energy exposure, a semiconductor may just cease operation entirely and later recover when the negative energy application ceases, or it may fail and never recover. For pulsing, strong "opposing force pulses" (stress pulses) are created deep inside the semiconductor materials. In case of sharp pulsing, sometimes an exposed semiconductor will just suddenly explode in a nice little fireball about one foot in diameter. This along with some minute stinging fragments from the explosion assures that the COP>1.0 bench researchers stay alert and on their toes. Needless to say, when experimenting with COP>1.0 systems and circuits, one should exercise all safety precautions, including safety glasses, until the **"negative** energy beasts have been tamed and controlled". One should not experiment with such things unless one is an experienced and safety conscious experimenter. One experiments at one's own risk.

Negative energy effects in coils and toroidal coils are largely still unknown, though we have outlined some of the major effects. For a normal helical coil, one can derive a concept of the effects by considering the lattice effects in the wire. In this case, introduction of hole current at one end of a helical coil will induce electron current from the other end and into the coil. Thus, hole current can be "switched" back and forth between the ends of a helical coil to induce an ordinary AC electron current in the coil, accompanied by normal magnetic fields etc. A point can be reached where the coil is actually in resonance with the frequency of the switching, and with the "seeming capacitance" of the hole current introduction in each switching. With the proper material in the coil and a keen balance between the negative energy capacitance effect and the normal positive energy induction effect, *in theory* a single coil can be turned into a self-oscillating, self-powering system.

With core material inside the core of a coil, the situation dramatically changes. Often the type reaction experienced will appreciably depend upon the core material, and changing the core material can dramatically change the effects upon the coil and the effects that the coil prc luces. One may even use dynamic changes in the core material to produce corresponding dynamic changes in the coil. A toroid may act differently from an ordinary helical coil, as does a flat (spiral-wound) coil. Here the experimenter must simply perform his own experiments to seek out and understand the phenomenology he uncovers.

For exploratory experimenting with core-filled coils, we advise strong caution and testing at very low power only. Sometime effects can apparently be experienced by the researcher (headaches, nausea, dizziness, etc.) It is suspected from some evidence (but not proved) that the core material's chemical characteristics can sometimes be "imprinted" upon the human body chemistry, from the novel radiation of a core-filled coil or some core-filled coils. This is particularly true in the presence of strong stress potentials, such as from bi-wound coils with fields in opposition, and particularly if the biwound coil contains a nonlinear material core. The nonlinear core induces modulation (multiplication) of the two waves, rather than linear mixing. Thus, the two waves are "locked together" to form a true stress potential and local spacetime curvature sets, when the biwound coil has a nonlinear material core. In short, one is now engineering pure general relativity, locally.

Again, we iterate that the experimenter should be knowledgeable and take all safety precautions. He experiments at his own assumption of risk.

The appearance of an extra stress potential (field-free) in a biwound coil is in fact a new Lorentz symmetrical regauging, and the equivalent of rotating the frame of the coil. As signals are introduced, this is the equivalent of continually changing the degree of the frame rotation, which is a general relativistic effect, with direct curvatures of spacetime being produced. In short, one is producing spacetime curvature and rotation engines, previously discussed.

The magnitude of the stress potential (field-free) falls off inversely as the distance. Thus one's body can be exposed to it, including the body chemistry, every cell and every part of every cell, the nervous system, etc. Placing signals on the biwound coil can conceivably result in slight variation of the scalar stress potential, including inside the body, and inside each cell and each part of each cell. Variation in the stress potential ϕ in a little local region inside the body produces $-\nabla\phi$, which is an electric field E. The rate of change of this E-field then produces a magnetic field H.

Consequently, very minute EM fields are produced throughout the body and all its parts, or can be. It is as if the entire body is "electromagnetically washed" in internal EM fields throughout its matter, structures, and dynamics. The hydrogen bonding actions ever ongoing are particularly sensitive and can restructure appreciably in the presence of this stimulus. They can also "imprint" and take on a specific induced structure and dynamics, correlated to the persistent changes induced from the biwound coil. As can be seen, such EM biological effects have not been appreciably investigated and documented in the Western scientific literature. Hence our urging and insistence that the researcher use all caution, and our warning that he experiments at his own will and assumption of risk. Neither the publisher nor the present author is responsible for any accidents or blunders that may occur.

Capacitor effects also vary between types of capacitors, when applying negative energy or a mix of positive and negative energy. Some capacitor dielectric materials are also photo-refractive and are semiconducting materials, hence photoactive. The rich effects capable of being generated in capacitors — by EM fields, waves, potentials, and pulses of negative energy or negative and positive energy mixed — are largely unexplored and not catalogued.

Almost all modern permanent magnetic materials are also photo-refractive or photoactive materials as well, and many are semiconducting materials. Hence a rich variety of effects can be produced in permanent magnets by the introduction of negative energy fields, potentials, waves, and pulses.

As can be seen, a very rich combination of effects exists for "normal-appearing" coils, capacitors, semiconductors, and magnets when applying appreciable negative energy or controlled mixes of positive and negative EM energy. This is a still largely unknown materials science region, and one that has not been systematically and rigorously explored and catalogued, if at all.

In addition, there are effects from *pulsing* negative energy (Dirac hole currents) that are not completely understood. These appear to occur almost — or partially — at random, and also seem to be affected by the natural surroundings, by the various materials, etc. The "imprinting" phenomenon, when it occurs, seems to be augmented by pulsing, and sometimes by the pulse shape and pulse repetition frequency as well as the surroundings.

We also legally state that we are not responsible for any risks the researcher chooses to take, but are only alerting him to be very careful and use very good safety practices. Any experimenter is himself responsible for stringent adherence to good safety practice and common sense, and he legally assumes his own personal risks for all experiments he chooses to undertake.

Another effect to be alert to is the effect of "charging" the local area's supersystem with an altered active vacuum and curvatures of spacetime. This phenomenon is usually experienced as a "slow growth" of the COP of the system, over a long time (months or years). Tiller in private discussions has reported such phenomena in his detector built at Stanford University. Golden and I experienced such phenomena with a small COP>1.0 system Golden was developing. Over a period of five years, the system slowly increased its COP to about 2.0. Moved several hundred miles, the system then exhibited COP of about 0.9, but with occasional fluctuations of overunity temporarily.[285] Golden had gradually charged his local area with the "engine" or spacetime curvature operational and functional form of his unit. Thus in that repeatedly charging area, gradually the operation of the unit improved. We referred to this *engine-*

---

[285] In fact, Golden and I were astounded at these results, until Tiller explained it to us and explained his own experiences with "growing" his detector. Thanks to Professor Tiller, thereafter we had this new phenomenon firmly fixed in our lexicon.

charging of the environment with a particular effect or dynamic form as *kindling.*

The kindling or "charging to favor the dynamic form of the engine or device" problem, should it occur, is in fact usable because it is possible to produce it everywhere on earth, and fairly permanently. To use the kindling effect once discovered, the mechanism is to build several identical units. Keep the original unit working in the original area, and "outrigger" the others at radial distances as distant as possible, but maintaining COP>1.0 for the outriggers. Then simply run both the outriggers and the central unit, steadily. The "structural charge" will grow and reinforce at the outrigger locations, then spread on beyond the outriggers, and a much larger area will now be "charged with that dynamic form". With sufficient additional outriggering (three to five, or seven in an especially difficult case) where the successively internal outriggers arc retained and operating, the charging will eventually spread around the entire Earth, and *it will become a permanent fixture of the Earth environment itself.*[286] Thereafter, anytime one of the units is turned on anywhere on earth, it will perform with its normal overunity condition. In other words, not only can one grow the dynamic form (engine) in a supersystem in a given locality, but one can also kindle it into the entire Earth's supersystem permanently {591}. One is strongly reminded of Feynman's observation that physicists with their big accelerators may in fact be partially creating their own reality. He pointed out that often the physicists look and look for a new, predicted particle. At last and after years of difficulty, finally a "glimpse" of the particle occurs in an accelerator and is reported. The physicists quickly crank up the other accelerators and look intensively, and — after some moderate difficulties — they find that particle. Thereafter, *every time an accelerator turns on and looks, that particle is readily seen, easily.*

The supersystem's spacetime is part of the local spacetime of the Earth, and all supersystems also interact with each other. As Kron put it {592}:

> *"An electric network differs from all other types of non-electric networks in that it is always surrounded in all directions to infinity by an invisible dynamic electromagnetic field of its own creation. ... Each Inductance L produces its own magnetic field, each*

---

[286] We have proposed an application of this phenomenon to kindle and maintain a specific disease organism's "anti-engine" as a startlingly new kind of medical therapy for permanently eradicating a particular disease vector from the earth.

> *capacitance produces its own electric field and each resistance R creates a thermodynamic field of its own. Moreover, there exists a continuous interaction between the magnetic, electric, thermodynamic, and radiation fields. Thus any correct and basic theory of the electric network must include all the parameters of its surrounding electromagnetic field and must satisfy all the field equations of Maxwell. The theory must also dovetail with the Hamiltonian and Lagrangean equations of dynamics containing both time and space derivatives, as well of the equations of irreversible thermodynamics. "*

So decades ago, Kron was struggling with the "supersystem" problem, in different terminology, and he eventually originated *diakoptics* {593}, an entirely new way of breaking the problem down piecemeal so it could be solved. But Kron strongly stressed that the interactions of the fields and potentials of a circuit and all its parts reach to infinity and have an n-dimensional set of degrees of freedom and an infinity of spaces. The fields and potentials of every charge reach to infinity (for those charges in matter from the beginning), and so every charge in the universe interacts with every other charge and its fields, potentials, and dynamics. Spacetime itself is filled with these interacting EM fields, potentials, waves, and dynamics. Little wonder that the horrendous set of ongoing EM interactions at every point in the vacuum has such incredible energy density and fluctuations. We may even regard the fluctuations and interactions as completely causal, but — with no information on each interaction — they must be treated as totally statistical. Ultimately nature confounds our Aristotelian logic and insists that, in the limit, one is dealing with 5-law logic as developed earlier in this book.

Since the earth near an ongoing COP>1.0 system experiment has innumerable electrical aspects, materials, etc., it follows that the surrounding earth is indeed in direct interaction with the experiment. It then follows that "slow conditioning" of the surrounding area should be possible, say, after some years as Golden experienced. It also follows that the effect diminishes in magnitude with distance from the experiment.

The "growth of the activation (kindling) to include the entire Earth" is a slow process, requiring years and much expense to aid the "spreading" — as the physicists may be experiencing, according to Feynman. But it is a doable.

We highly recommend that the researcher work on COP>1.0 systems that are not charge and location dependent, so that the kindling problem does not arise. We have assumed non-kindling COP>1.0 systems throughout most of this book.

Chapter 10

Cold Fusion: Low Spatial-Energy Nuclear Reactions at High Time-Energy

"This book is about time. I would have liked to have named it Time, the Forgotten Dimension..." **[Ilya** Prigogine]. {594}

"The notion of time may be unfolded into an independent pure science... a science of pure time is possible." [Sir William R. Hamilton, 1837]. {595}

"...contrary to the conclusions of classical mechanics, there exist effects of potentials on charged particles, even in the region where all the fields (and therefore the forces on the particles) vanish. " [Y. Aharonov and D. Bohm] {596}.

"...global behavior greatly modifies the very meaning of space and time. Much of geometry and physics is based on a simple concept of space and time, generally associated with Euclid and Galileo. In this view, time is homogeneous. Time translations may have no effect on physical events. Similarly, space is homogeneous and isotropic; again translations and rotations cannot alter the description of the physical world. It is quite remarkable that this simple conception of space and time may be broken by the occurrence of dissipative structures. Once a dissipative structure is formed, the homogeneity of time, as well as space, may be destroyed. " [Ilya Prigogine] {597}.

"As we expand our observation, we extend our concepts. Thus the simple symmetries that once seemed self-evident are no longer taken for granted. Out of studies of different kinds of interactions we are learning that symmetry in nature is some complex mixture of changing plus into

minus, running time backward and turning things inside out. " [T. D. Lee] {598}.

"The desirable features of a system of units in any field are convenience and clarity. For example, theoretical physicists active in relativistic quantum field theory and the theory of elementary particles find it convenient to ***choose*** *the universal constants such as Planck's quantum of action and the velocity of light in vacuum to he dimensionless and of* ***unit magnitude.*** *The resulting system of units (called 'natural' units) has only* ***one*** *basic unit, customarily chosen to be length. All quantities, whether length or time or force or energy, etc., are expressed in terms of this one unit and have dimensions which are powers of its dimension. There is nothing contrived or less fundamental about such a system than one involving the meter, the kilogram, and the second as basic units. It is merely a matter of convenience."* [J. D. Jackson] {599}.

[Violation of Second Law of Thermodynamics] *"We experimentally demonstrate the fluctuation theorem, which predicts appreciable and measurable violations of the second law of thermodynamics for small systems over short time scales... Our results show entropy consumption can occur over colloidal length and time scales."* [Evans et al.][287]

[Implications] *"This result has profound consequences for any chemical or physical process that occurs over short times and in small regions. "* [Evans et al.][288]

[Implications] *"The smaller a machine is, the greater the chance that it will run backwards. "* [Evans et al., *ibid.]*

[Implications] *"Physicists knew that at atomic scales over very short periods of time, statistical mechanics is pushed*

---

[287] G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.*, 89(5), 29 July 2002, 050601.

[288] BBC News, **18 July 2002, http://news.bbc.co.uk/l/hiscitech/2135779.stm.**

> *beyond its limit, and the second law does not apply. Put another way, situations that break the second law become much more probable. ...the new experiment probed the uncertain middle ground between extremely small-scale systems and macroscopic systems and showed that the second law can consistently be broken at micron scale, over time periods of up to two seconds."* [Chalmers] [289]

10.1 Explaining Time as Energy

See Chapter 1, paragraph 1.1.3 and 1.4.1; see also Chapter 2, paragraph 2.1.3. As highlighted in the quotation from Jackson {599}, it is well known that the fundamental units chosen to model physical phenomena are arbitrary. In a physics model whose only fundamental unit is the joule, time is obviously modeled solely as a function of energy, just as mass is modeled today as a function of energy — hence the term "mass-energy". We may also equally well use the term "time-energy", since in a model using the joule as its only fundamental unit the second is purely a function of energy. Specifically, for our purposes 1 sec = 9 x 10^{16} joules of spatial energy, compressed by the factor c^2 and placed on the 4th Minkowski axis to constitute one second of time. If the compressed spatial energy had been left in 3-space, it would be recognized as *mass*. For our purposes, *time* has the same energy density as mass.

The flow of time becomes a special form of EM energy flow, analogous on the fourth axis to mass flow (translation) in 3-space. Further, as previously pointed out the overall flow of time has extensive internal structuring of component time-energy flows. However, time-flow is in the imaginary plane or connected with it, since in Minkowski space the 4th dimension is modeled as -i*ct*. Note that the only variable in *ict* is the *t*, hence anything varying on the fourth axis is a variation of time and time-energy *a priori*.

Further, we can apply conservation of EM energy to the time domain as well as in 3-space. And we can have "broken symmetry" in time-energy flow as well as in 3-spatial energy flow, so long as the net energy flow is conserved. The most fundamental conservation of energy law is giant negentropy: the automatic circulation of energy from the time domain into 3-space via decompression by the negative charge, thence the slight flow of 3-space energy to the positive charge where the 3-space energy is recompressed, thence from the positive charge back into the time domain.[290] That was our solution to the long-vexing problem of the source charge and its associated 3-space fields and potentials and their 3-space energy. From special relativity, one already knows that time can dilate and then *un*-dilate, but each action is associated with an appropriate change in spatial energy. While that change in spatial energy can be kinetic, it does not have to be. It can be a change in potential energy instead. In short, a static object rotates its frame with respect to the observer (lab) frame, whenever it absorbs or emits potential energy.

---

[289] Matthew Chalmers, "Second law of thermodynamics 'broken'." *New Scientist*, 19 July 2002, http://www.newscientist.com/news/print.jsp?id=ns99992572 .

We also pointed out the connection of time-energy to electromagnetics and to quantum field theory, via the four polarizations of the photon considered to travel in the z direction in 4-space. These four polarizations are with the energy vibrating along *x*-, *y*-, *z*-, and t- respectively. Vibration of the spatial energy of the photon along the *x*- or *y*- axis gives an ordinary transverse photon. Vibration of the spatial energy of the photon along the z- axis gives the longitudinal photon. Vibration of the time-energy component along the t- axis gives the scalar or time-polarized photon. We assume a similar four polarizations for EM waves in 4-space.[291]

10.2 Mechanism Generating Flow of a Mass through Time

As we covered in the referenced paragraphs in Chapters 1 and 2, the fundamental generatrix of the flow of a mass through time — and the structuring of that flow — is the total photon interaction with the mass and all its parts. This includes both observable and virtual photon interactions.

As a recapitulation: Mass *m* does not exist in time, since it is an observable and observation is the imposition of d/*dt* upon spacetime LLLT, removing the T to give a "frozen LLL snapshot." In short, mass m is an *effect* (of the iterative very high speed observation process), not a *cause* (which is an *input* to the observation process). This is why all observation is 3-spatial rather than spatiotemporal: it yields the 3-space *effect* (at a given time, of a given interaction), not the 4-space cause of the interaction and not the ongoing 4-space interaction.

---

[290] Note the relationship with such things as time dilation, which represents a broken time-flow symmetry, where net energy flow from the time domain to the 3-space domain occurs. Hence time dilation is associated with the increase of spatial energy, such as kinetic energy due to the relativistic velocity of the moving object.

[291] In this book, we have limited our discussion to four dimensions. If we model in five or more dimensions, then additional symmetries and broken symmetries **are** possible, as are new kinds of multi-dimensional EM waves.

By absorbing and emitting photons, observed mass *m* may be said to move through time in a succession of *m* => *mt* => *m* => *mt* => *m* ... jumps, each involving change of physical state as a photon is absorbed by *m,* changing it to *mt* which is nonobservable. Then a photon is re-emitted by *mt,* changing it back to *m* and completing another observation with this new *m* as the output (effect). We note that masstime *mt* is as different in nature from mass *m,* as impulse *Ft* is from force *F.* In mass *m,* every tiniest differential *dm* is also interacting with photons continuously (even though they may be *virtual* photons), hence *dm* also exhibits its own "digital flow through time" as a part of the structure of the overall flow of the parent mass through time.

Interaction primarily with the *mt* state yields wavelike interactions. Interaction primarily with the *m* state yields particle-like interactions. This is proposed as a possible resolution of the wave-particle duality problem, which has never been resolved but just accepted without further argument. Previous consideration of wave-particle duality has assumed that "mass has two states, one particulate and the other wavelike". We correct that assumption by pointing out that two different "states in which mass exists" are involved, and the two are quite different: one is mass *m,* and the other is masstime *mt.* We stress the function of the myriads of virtual photon interactions during an *mt* state, where an observable photon has not yet been re-emitted. These virtual photon reactions create streams of virtual *dm* => *(dm)dt* => *dm* => *(dm)dt* =>. . . . .on all the differential parts of mass *m,* which creates the rich structuring inside the "time-flow" of the overall observable mass *m* through time. The rich dynamics of the *dm* -> *(dm)dt* streams also accounts for the seeming existence of mass *m* as a set of 4-dynamics which interpret as a set of waves. We note that the emitted photon carries with it the internal structuring of the spatial energy domain and the time-energy domain. Herein we have the basis for the Kaznacheyev's demonstrated electromagnetic induction of cellular damage and disease at a distance.[292]

Finally, we tentatively assume that, prior to its interaction, the antiphoton carries negative spatial energy and positive time increments, while the photon carries positive spatial energy and positive time increments. In a very small region where an excess of antiphoton interactions occurs temporarily, time-reversal occurs and the physics "runs backward". It has recently been experimentally proven that such "reversing" of the physics processes occurs up to micron size and for times up to two seconds.[293]

---

[292] Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian|, Novosibirsk, 1981.

10.3 Time-Polarized EM Waves and Longitudinal EM Waves

10.3.1 Waves, Photons, Spacetime Curvature, and Observation

Spacetime (ST) curvature involves not only the change of *spatial* energy density of local spacetime but also a change of its *time-energy* density. That component of the total ST curvature due to time-energy density change — i.e., time-polarized EM photons and waves — can be far more powerful, because of the enormous energy density of the time-energy compared to the far lies dense 3-spatial energy. As previously stated, time may be regarded as having the same energy density as mass, except that the highly compressed 3-spatial energy is moved to the 4th Minkowski axis to constitute "time" rather than being left in 3-space to constitute "mass".

Hence ST curvature can be made very powerful when one uses longitudinal phase conjugate EM *wavepairs,* as shown by Whittaker {600} in 1903 with slightly corrected interpretation as argued by the present author {615}. Each longitudinal EM wave in 3-space is accompanied by a precisely coupled time-polarized (scalar) EM wave in the time domain, which latter wave is a longitudinal EM wave on the fourth Minkowski axis involving changes in the t variable in the axis *-ict.* We previously discussed the giant negentropy that this represents with respect to the source dipole. At various times in the past, we have emphasized the time-domain aspects of the coupled phase conjugate longitudinal EM wave-pair by calling such a coupled wave by terms such as *time-density EM waves, time-polarized EM waves,* etc. These terms imply additionally altering and structuring the *time density* and *time-energy density* of spacetime rather than merely altering and structuring its *3-spatial energy density.* Such time-density induced ST curvatures may be nearly 10^{17} times stronger than ordinary spatial energy density ST curvatures {601}.

Interestingly, at least one researcher — Matsumoto {602} — has realized the basic implication in cold fusion processes of using the strong EM force as the agent of spacetime curvature. He has shown consistency of the postulated EM-induced ST curvature approach with some important cold fusion results. However, Matsumoto has utilized the conventional interpretation of spacetime curvature due to alteration of *3-spatial energy density* of spacetime. This misses the primary time-density functions, but does include the resulting spatial-energy functions after transduction from time-energy to spatial-energy. His work, however, does support the involvement of a more intense form of energy in the cold fusion effects.

---

[293] See footnotes 287, 288, and 289.

There are intriguing but still mysterious relationships to be noted between cold fusion and deep astrophysical mechanisms. Continuing reference is made to these similarities, including black holes and white holes. Black holes and white holes involve spacetime curvatures — extreme, but still spacetime curvatures. So one would suspect that the entire scale of spacetime curvature — including black holes and white holes in the limit — should be of interest to cold fusion researchers, since in their use of time energy they are inadvertently involving more intense forms of energy than ordinary spatial EM energy.

Most nuclear scientists and even some relativists seem unaware of the much more powerful spacetime curvature effects obtained by alteration of the *time density* of spacetime rather than just its *spatial energy* density. Physicists have largely omitted specializing on time density ST curvatures because they have largely omitted the absorption and emission interactions of the time components transported by photons and EM waves.

Longitudinal EM phase conjugate wavepairs (each wavepair a coupling of a time-polarized EM wave and a longitudinal 3-space EM wave) of the type shown by Whittaker {600} in 1903 involve net *time density oscillations* as well as 3-spatial energy density oscillations, according to our reinterpretation. The infolded "inner EM content" of all conventional EM potentials, fields, and waves thus contains extraordinary dense time-energy, normally lost in the *d/dt* observation interaction.

In short, the infolded "internal Whittaker electrodynamics is a pure "general relativistic electrodynamics" and a unified field dynamics. Further, it can be engineered, which leads to startling new physical processes previously thought to be impossible and forbidden by the laws of nature.

Prior to observation, the internal highly compressed time-energy structures associated with the interior EM structure of all conventional EM potentials, fields, and waves are extremely powerful spacetime curvatures.

In short, all "conventional" electrodynamics has an internal structure of strong spacetime curvatures and dynamics.

Organization and patterning impressed on, or inserted in, this internal structuring can provide extraordinary actions in materials, living biological systems, cells, etc. Adding deterministic structuring inside these normal EM entities is referred to by Russian energetics weapon scientists as engineering the *information content ofthe field.* This is the approach behind deliberate engineering of *spacetime curvature engines,* or vacuum engines — or just "engines" for short.

In theory, any physical change to a mass can be engineered by inserting such "engines" into the mass. Any mass-energy is a mass-energy EM potential with an internal structuring (engine) for all the mass's internal dynamics. This engine is referred to as the *resident engine* of that mass (and all its dynamics). Introducing to a given mass a potential with a desired "delta engine" structure, results in the superposition of the two potentials, resulting in exchange sharing of the engines. In that manner, a desired "delta engine" diffuses into the mass's potential structure and its resident engine, changing it to a new resident engine which now acts on the mass at every level to change it. Potentials superpose, and their engines diffuse and superpose as well. Direct engineering of matter — even inside the nuclei and inside the nucleons themselves — can be performed by introducing the desired engine or engines for the changes desired.

This area has been highly weaponized by one or more nations. E.g., the Russians — in several decades of so-called "microwave radiation of the U.S. Embassy in Moscow — induced health changes and diseases in U.S. personnel in the Embassy by inducing deliberate "disease engines". Such EM-induced diseases and health changes were generated in the zones where the potentials were field-free, and therefore stable for stable engine diffusion.

Note that the common scalar potential decomposes by Whittaker's 1903 decomposition {600} into the longitudinal EM wavepairs we are speaking of. Also note that, per a second paper by Whittaker in 1904 {91b} that initiated superpotential theory, all EM fields and waves can be decomposed into two scalar potential functions. Each of those base potentials involved in the potential functions further decomposes by Whittaker 1903 into a harmonic set of phase conjugate longitudinal EM wavepairs as previously discussed. Therefore, the infolded EM inside all the "conventional" EM fields, waves and potentials is indeed a highly organized set of such bidirectional longitudinal EM wavepairs and thus a highly organized set of extreme localized ST curvatures. In other words, it is comprised of these special "engines" of deterministic spacetime curvatures and their dynamics.

Constructing the potential by assembling one's own set of phase conjugate longitudinal EM wavepairs to form it, one can insert any kind of ST curvature set — any kind of vacuum engine desired — inside the base potential(s), hence inside the two potential functions comprising a given set of EM fields and waves utilized as "carriers". The specific internal structure of a "carrier" radiation field or potential designed to carry such engines is called the "information content of the field".[294] The Russians have long had a disinformation campaign to suggest that the term refers to conventional spectral analysis. It does not.

Johns Hopkins scientists did detailed measurements inside the Embassy and confirmed that all the anomalous health changes and diseases had occurred only in areas where the EM fields B and E were absent. That is, the health changes occurred only where the potentials were field-free and therefore stable and unchanging, so that an exposed body was immersed in a steady sea of "vacuum engines" being diffused throughout the targeted body's EM potentials in and on its cells, its cellular membranes, etc.

In general relativistic terms, every physical set of changes and altered dynamics (syndrome) of a specific disease has a specific *delta* set of spacetime curvatures and dynamics associated with it, which interacts on the body, its chemistry, its electrodynamics, its cells, etc. A desired set of

---

[294] Indeed, in theory one can *construct, functioning robotic systems* of these internal engines. In clandestine Russian weapon research, it appears that at least elementary "causal robots" of such internal engine type have been developed and tested. In a sense, these are the ultimate "nanobots" made of infolded longitudinal EM waves and their dynamics. In theory, command and control functions (via longitudinal EM wave communication) can be included, as can certain kinds of weapons effects based on scalar interferometry. The eventual development of such startling weapons seems intended to produce an army of eerie "internal EM" robots which can be "sent in'" *through the interior of EM fields, waves, and potentials,* to targets anywhere on earth, beneath its surface, in space, or beneath the ocean. This may be intended to become the ultimate robot or "high tech" warfare, and it is also the ultimate "information" warfare. There are no shields against this type weaponry. Defense, however, is accomplished by first using longitudinal wave "scanning" to pick up and track the systems, much like radar picking up aircraft. Once detected, scalar interferometry is utilized to disrupt and destroy them in place. Progress on these robotic weapon systems appears to be well along, as does progress on their countering longitudinal EM wave weapons.

ST curvature deltas for a given disease was inserted inside the base potentials in a combined Whittaker 1903 and 1904 decomposition of the stable field-free potentials in the Embassy. The Russians thus induced the desired engines for induction of the specifically desired disease(s) in the assigned personnel in those field-free potential zones in the Embassy. The purpose was to use the tests as (i) a guaranteed means to provoke high level U.S. governmental attention, and (ii) a probe to see if the U.S. knew about such technology and had developed defenses. Other tests of such disease induction in entire populations have also been conducted, but these are beyond the scope of this book.

We point out one very important thing: the cells of the body have their normal "resident engines" and resident engine dynamics inside their internal fields and potentials and waves. For any specific disease or disorder condition, there exists a specific associated change (delta) in the normal set of vacuum engines and their dynamics, in the cells and the body. In short, there exists a specific "delta engine" for any cellular disease or disorder, added to the normal engine.

It is not the immune system that heals. Instead, it is the cellular regenerative system. The natural healing method used by the cellular regenerative system is not chemical. The regenerative system uses the phase conjugate wavepair composition of increased scalar potentials (on cellular membranes and every part of the cell) to "pump" the cells in the time-domain. This creates an "anti-engine" for the specific disease delta or disorder, and it slightly amplifies this anti-engine within the limited capabilities of the body. This action places this amplified anti-engine also into the stricken cells' engine complex. Engines superpose, just as do potentials. This action by the regenerative system has the effect of slowly producing complex anti-forces in the body and inside the cells themselves, eliminating the detrimental engine delta, restoring the normal resident engine and thus time-reversing the stricken cells back to a previous healthy condition. It does not kill pathogens at all; that is the function of the immune system. In the 1960s and early 1970s, Antoine Priore in France demonstrated remarkable cures of terminal tumors, infectious diseases, clogged arteries, and other disorders in laboratory animals,[295] unwittingly amplifying the body's own master regenerative mechanism. His work was suppressed in the mid-70s.

---

[295] See (a) A. Priore, *Guerison de la Trypanosomiase Experimental Aigue et Chronique par L 'action Combinee de Champs Magnetiques et D 'Ondes Electromagnetiques Modules* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], Priore's thesis submitted in candidacy for the doctoral degree, 1973. The thesis was rejected when the project was suppressed. See also (b)

Check our website, www.cheniere.org, for a 119-slide "porthole" briefing on the present author's extension whereby specialized mixes of ordinary EM waves can be used to initiate and perform this "delta engine elimination" healing process, and it can be highly amplified well beyond the body's normal healing capabilities.

In 1998 we seriously proposed to the DoD, the USAF, the NIH, and other U.S. government institutions that a crash development program be initiated to develop portable units capable of rapidly (5 minutes per patient) treating mass casualties from coming terrorist attacks with weapons of mass destruction such as anthrax, smallpox, bubonic plague, Ebola virus, etc. Sadly, no one in the U.S. government had the foggiest notion of what we

---

"Le Probleme Priore," *Rapport de la Commission de 1'Academie des Sciences a Monsieur le Ministre d'Etat charge de la Recherche et de la Technologie,* 1982, p. 1-22; (c) Antoine Priore, "Precede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; (d) Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (e) Antoine Priore, "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968; (f) A. J. Berteaud and A. M. Bottreau, "Analyse des rayonnements electromagnetiques emis par l'appareil Priore," [Analysis of the electromagnetic radiations emitted by the Priore apparatus], D.R.M.E., 1971, p. 3-12.; (g) R. Courrier, "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reunion a LTnstitut sur les effets de la Machine de M. A. Priore le 26 Avril 1977," [Presentation by Professeur R. Courrier, Perpetual Secretary of the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore]; (h) R. Pautrizel, A. Priore, M. Dallochio and R. Crockett. (1972) "Action d'ondes electromagnetiques el de champs magnetiques sur les modifications lipidiques chez le Lapin par l'administration d'un regime alimentaire hypercholestrole." [Action of electromagnetic waves and magnetic fields on provoked lipidic modifications in the rabbit by the administration of an alimentary hypercholesterol diet.] *Compt. Rend. Acad. Sci. (Paris),* Vol. 274, 1972, p. 488-491. The same university that rejected Priore"s doctoral thesis in 1973, accepted a doctoral thesis on the same subject some 11 years later due to the persistence of Pautrizel: See (i) Eric Perisse, *Effets des Ondes Electromagnetiques et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale [Effects of Electromagnetic Waves and Magnetic fields on Cancer and Experimental Trypanosomias],* Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

were discussing. At NIH, we never got out of their "policy" (i.e., "spin control") department. Not a single capable scientist reviewed the extensive package. The attack on the World Trade Center and the Pentagon on Sept. 11, 2001 is just a forerunner of that which is to come. We could have been prepared to save hundreds of thousands of Americans, even millions. We shall not be ready, and the first really professional anthrax or smallpox (or camel pox, which is easily available and almost exactly the same as smallpox) attack on a major U.S. population center may leave from one to three million Americans dying in the streets and in their homes. Presently nothing will save them. It could have been different.

10.3.2 Photons and Waves in Minkowski Spacetime

Photons are made of angular momentum (action); i.e., of *energy x time.* Waves, being made of photons, carry both spatial (decompressed) energy and time (highly compressed) energy — although physics has largely omitted the *time-energy-carrying* aspect, and only accounted for the *spatial energy-carrying* aspects. An EM wave in space is actually a wave of the particular spacetime curvatures that produce angular momenta in interacting charged masses. The standard notion of E and B fields at right angles in a plane in 3-space, traveling along the *z*- direction and with the plane orthogonal to *z*-, and with E and B field vectors oscillating back and forth, is nonsense, as acidly commented upon by Romer {603}, the former editor of *American Journal of Physics.*

Nonetheless, using the standard way of modeling the "transverse" EM wave: In conventional electrodynamics, a *transverse EM wave* is one where the 3-spatial energy density is considered to be varying rhythmically at right angles to the direction of wave propagation, and the time-density is assumed to be non-varying, or at least altered only by special relativistic considerations in certain cases.[296] A longitudinal EM wave has its spatial energy density fixed in *x*- and *y*- directions, and its energy density is oscillating along *z*-, the direction of wave propagation, by a "compression" and "rarefaction" process much like squeezing an accordion. This, however, leads to a "galloping" motion where the wave is alternately very superluminal and very subluminal at given instants, but moving at an *average* velocity of *c* {604}. On the other hand, a *time-density (scalar) EM wave* by definition is not varying in its 3-spatial energy density, and therefore is not varying in *x*-, *y*-, or *z*-. Instead, the energy is in the form of time-energy, and its variation is in time density. So its energy density is oscillating along the fourth Minkowski axis, *ict*, as an oscillation of the energy density of the variable *t.* It is therefore polarized in the t- direction.

---

[296] We point out that a wave of varying spatial energy density is moving in a varying curved ST a priori, by general relativity. If the "wave" were in a flat spacetime as normally assumed, the EM flat spacetime would eliminate the EM wave itself. **This** has been particularly pointed out by Sachs and by Evans.

10.3.3 Time-Charging, Internal Structuring, Vacuum Engines, and Diffusion

Photon interaction as a *spatial* energy collection or energy exchange process is well known in physics. However, the time-charging or time-excitation interaction between the mass and the time component carried by the photon is usually ignored.

As is well known, an atomic electron that absorbs a transverse photon increases its spatial energy and changes into an excited spatial energy state. When the electron subsequently "decays" from that excited energy state to a lower state, the spatial energy differential between the two states is emitted as the spatial energy component of an emitted photon. This is the *spatial energy density* reaction of a transverse photon with an absorbing and emitting mass particle.

Unaccounted for by physics, at the time of the energetic excitation of the mass (i.e., when the photon is absorbed), a time excitation — time charging — also occurred because the time-component of the absorbed photon is also attached to the mass, converting it to *masstime,* a new entity. Re-emission of the photon is accomplished by the decay of both the *spatial energy* excitation and the *time-energy* excitation. In the previously excited state after photon absorption, however, one did not have the entity "mass", but the entity *masstime* having greater topology. If mass is an observable and taken to be 3-spatial (as is conventional), then photon absorption and emission oscillates between 3-space and 4-space effects, generating both spatial energy and time-energy (or *angular momentum)* excitation and decay.

Further, only *masstime* emits a photon. This immediately casts a very different light on such things as radioactive elements, which are in an initially excited state. Discussion of this aspect, however, is beyond the scope of this treatise.

On the other hand, a *scalar (time-density)* photon {605, 606} has a rather different interaction, regardless of how physics presently treats it. Any atomic electron at any energy level is continually receiving and emitting virtual photons - transverse, longitudinal, and scalar- in its virtual photon exchange with the energetic vacuum. Hence it is continually "connecting to" or "absorbing" time components, thereby changing its nature from mass to masstime, back to mass, back to masstime, etc. Not only the entire electron but also each differential piece of it is performing that dance between mass and masstime. This is a continual changing from 3-space to 4-space and back. There thus is a violent "chatter" of the electron (and each differential piece of it) in the myriad levels of its *time* exchanges, giving it

(i) An average rate of time component exchange,

(ii) An average "time-charge", and

(iii) A dynamically changing, specific structuring of its time components.

Whenever a mass absorbs an observable photon, the mass converts to masstime, with an "extension in time" as well as extension in 3-space. During the time-interval created in the masstime, myriads of virtual photons are absorbed and emitted by that mass component of the masstime entity. Hence every dt portion of the masstime state after observable photon absorption, is "internally structured" by the entire virtual photon interaction set's At's occurring within that time interval.

By tailoring at least the larger elements of this internal structuring of the masstime state's virtual photon bombardment, by a deliberately induced virtual photon bombardment, the emitted photon from the masstime decay will carry away with it a desired internal "vacuum engine" structure of organized curvatures of spacetime. In short, two potentials superpose; that is one of their primary features. When they superpose, they also diffuse their internal structures (internal vacuum engines) into each other.

We speak of the *diffusion ofa specific vacuum engine* placed in one interacting EM potential, field, and or wave as *dimensioning* the potential, field, or wave. Adding a dimensioned potential, field, or wave to other potential(s), field(s), or wave(s) diffuses the dimensioning (the engines) between all of the participating entities. With steady radiation of such dimensioned signals into a targeted object, field, wave, or potential, the dimensioning is gradually created (kindled) in the target, simply diffusing throughout it and gradually accumulating. If the irradiating signal carries a specific disease — e.g., necrotizing fasciitis — then gradually in the irradiated body there will emerge all the symptoms of necrotizing fasciitis. If the irradiating signal carries a specific disease antiengine — e.g., for necrotizing fasciitis — then gradually the necrotizing fasciitis condition will be eliminated, with the disease organisms also being eliminated.

This is a more fundamental extension of the disease vector concept. Not only can a disease vector be induced by a physical organism such as a virus or bacterium, but it can also be introduced by simply capturing the fundamental vacuum engine pattern of the disease and using it to condition or dimension signals used to steadily irradiate the targets.

This is the explanation of the Kaznacheyev experiments {607}: thousands of experiments performed by Vlail Kaznacheyev and his colleagues demonstrated the induction of almost any kind of remote specific cellular death and disease patterns in targeted cells, by absorption of radiation emitted from appropriate dying or diseased "emitter" cells with that condition, if the experiments were performed in the dark.

Instead of looking for biogenesis causes in amino acid precursors to cellular life in meteorites etc., one might well look into the *vacuum engine structures* inside the light from the sun, the planets, the stars, and the universe in general. E.g., dead sperm can often fertilize female cells, if not too much physical damage and deterioration to the dead sperm cells has yet occurred. This is understandable, since for a time a dead but undamaged sperm cell will have the same resident engine as a living sperm cell.

10.3.4 Some Cosmological and Gravitational Effects

Each of the observable-sized dt's in all the serial masstimc excited states of the mass moving through time, is internally and precisely structured. **The** precise time history of any mass — together with all its interactions, experiences, etc. — is present inside the time stream of ongoing interactions of that mass with photons. These virtual photon changes are in fact deterministic and come from every part of the universe, as shown by Puthoff {608} in his self-regenerative cosmological feedback cycle. We have merely extended his cosmological feedback to include the time **domain** as well as the energy domain. We hypothesize without further **discussion** that this internal structuring of the time-domain for every mass or object will eventually be connected with the quantum concept of the ge**ometric** phase {609}.

When a transverse photon interacts with a mass in the presence of other **photon** interactions, the mass can act as a phase conjugate mirror and form the phase conjugate replica photon (an antiphoton). For a longitudinal **photon,** the antiphoton *prior to observation* is a scalar photon. We propose that in the unobserved causal state the photon and antiphoton can couple (strong interpretation of the distortion correction theorem) and form a spin-**2 graviton.** According to quantum field theory, one may already interpret the common scalar potential as being comprised of such spin-2 gravitons {610, 611} if we speak of the situation prior to observation.

As opposed to the accounted *spatial energy charging* of the familiar transverse photon interaction, the longitudinal photon interacting with mass will cause immediate phase conjugation by the mass because of the accompanying scalar photon (accompanying antiphoton prior to observation). When one pumps with longitudinal EM waves, one also inadvertently pumps with time-polarizcd EM waves simultaneously.

From its background "time-energy pumping" in its vacuum exchange, the mass forms the phase conjugate replica, which couples to the incident longitudinal EM photon. This forms (wave view) bidirectional longitudinal EM wavepairs (conjugate pairs), comprised of (particle view) internal coupled longitudinal photon-antiphoton pairs. Each of those internal photon-antiphoton pairs is a spin-2 graviton.

Each such conjugate longitudinal wavepair actually contains a time-density (time-polarized) or scalar EM wave. So the interaction of LWs and their accompanying scalar waves immediately initiates the formation of time-density waves, which then primarily perform a *time-charging* interaction. In addition, even for an equilibrium time-charging state, the "masstime" state of the interacting electron contains a specific structure of the time elements in the electron-time's time component. Each of these infolded time elements is also time-charging the electron.

10.3.5 Charging and Excitation Considerations For Cold Fusion Purposes

So there are several new types of charge or excitation that are involved in cold fusion and other phenomena. One may speak of charging and excitation such as gravitational charging, time charging, time-energy excitation, time-energy charging, etc.

Both the longitudinal photon and the scalar (time-polarized) photon are in fact known in physics, but usually neglected, at least at the end of calculations. E.g., Gray {612} puts it this way about the four kinds of photon polarizations and the habitual elimination of the higher ones:

> *All that quantum theory can say is that there are two transverse photons, a longitudinal photon and a scalar photon. It must be decided by other means which are "desirable" states and then it is customary to "eliminate" the longitudinal and scalar photons by invoking a*

> *subsidiary condition, now somewhat weakened from the Dirac form.*

For our purpose in examining the cold fusion experiments, there exists a completely unsuspected "time-charge" set of excited states for an atomic electron, a proton (as in an H+ ion), etc.[297] There also exists a "time-structuring" of the time-charge excited time-state, which we will see is important in certain highly anomalous instrument phenomena. Uncovering this time-charging and time-structuring interaction has been one of the major results of this author's long effort to decipher the functioning of the healing (cellular regeneration) system of the body as well as his efforts to decipher the fundamental nuclear transmutation mechanism in cold fusion transmutation interactions at feeble *spatial* energy.

In summary: For longitudinal photon interaction, the atomic electron "absorbs" the longitudinal photon and its accompanying scalar photon, being simultaneously spatial-energy excited (charged) by the longitudinal photon and time-energy excited (time-charged) by the scalar photon. This time-excited state (of masstime) then subsequently decays, emitting a longitudinal photon accompanied by a scalar photon in the process.

---

[297] See again footnotes 287, 288, and 289. From our mechanism propelling a mass through time, time-charging and discharging play the causal role. The acquisition of dt by mass *m* produces masstime *mt,* then the subsequent photon emission decay of *mt* back to *m* produces "propagation of mass *m* forward through time" by one "jump". Time charging with -dt and subsequent photon emission is what produces "propagation of mass m backwards through time" by one "jump". Evans et al. proved that even up to the micron scale physical particles and their dynamics can run backwards in time for up to two seconds or more. The fluctuation theorem proved by Evans and his colleagues in 1993 shows us that, at the scale of a proton and neutron, there can become a high probability of significant time reversal of the physical dynamics. This means that the law of attraction and repulsion of charged particles -such as two H+ ions in solution — "runs backwards" or is reversed, so that momentarily like charges attract and unlike charges repel. We have called the region in which this reversing of the physics occurs a *time reversal zone* (TRZ). With the reccentwork of Evans et al., the existence and occurrence of TRZs even well above the proton (the H+ ion) level is now clearly established, including the reversing of the physical dynamics. These concepts have not yet been applied to cold fusion interaction results by the leading cold fusion researchers, but they now stand as explaining the most probable mechanism for the observed phenomena. In this Chapter we have written some of the typical new nuclear reactions that can occur in such a TRZ, and the exact products of these reactions are widely reported in several hundred successful cold fusion experiments.

We shall return to time-charging and time-charge decay when we examine and explain the odd instrumental anomalies experienced for some time in experiments at China Lake.

10.4 Time Reversal Zone and New Nuclear Reaction Mechanisms

It follows that the flow of a mass through positive time must involve a preponderance of reactions with photons rather than antiphotons, since *net* absorption and re-emission of photons (rather than antiphotons) is involved. Otherwise, the absorbing and emitting mass would not accomplish little "jumps" forward in positive time.

Suppose we deliberately arrange a situation where the target mass is interacting with a preponderance of antiphotons. In this case, the mass will be time-reversed, as will be the EM energetics.[298] This leads us to the novel concept of a *time reversal zone* just described in footnote.

A time-reversal zone (TRZ) is a region of space, or electrolyte, etc. in which given masses or charges of interest are interacting on the average with more antiphotons than photons. Consequently, in the TRZ the usual EM energetics are reversed and normal charge reactions appear to "run backwards" insofar as the spatial 3-space observer is concerned.

For example, in electrolytes in a region of highly loaded positive charges in a loading palladium lattice with properly prepared surface, a very great number of double surfaces exist. Hence a great number of scalar interferometries continually occur from that large number of double surfaces acting as scalar interferometers {613}. Because of the highly excess positive charge loading, these interferometers are predominately fed by negative EM energy from the positive charges. Hence, in the distant interference areas in the electrolyte outside the palladium lattice, some negative energy fields and negative energy potentials are produced dominantly. An excess of antiphotons is produced in those interference zones where an excess of antiphotons appear due to the scalar interferometry. In *those* little interference zones, the simple positive ions are momentarily bathed in an excess of antiphoton interactions. This is therefore a time-reversal zone of momentary time-reversed EM energy flow.

---

[298] This *is not* time-travel in the classical science fiction sense. For time-travel, the traveling object must remain moving in its own forward time, while the entire remainder of the universe must be reversed in its time and must move backward in time to a past coordinate. That is not what is happening here, and no one is suggesting we can time-reverse the rest of the universe! Instead, energy can be reversed in time, as proven in nonlinear phase conjugate optics. So can mass-energy and charge-energy, as shown by the Dirac theory of the electron. An *observed* positron is an electron observed while traveling backwards in time, so to speak, with respect to the observer. But it is observed in observer forward time as traveling in the opposite spatial direction, having positive mass and positive energy, and with the sign of its charge reversed from negative to positive.

From a palladium lattice loaded with H+ or D+ ions, and also having a proper surface with many work grooves and thus many small interferometers, random fluctuations in the scalar interferometry occur in the adjacent electrolyte. Some of these random scalar interferometries continually form fleeting time reversal zones (TRZs) in their interference zones in the surrounding electrolyte. The very large energy required for the formation of each TRZ and its reversal of the normal laws of attraction and repulsion of charges, is available from conversion of time energy from the time domain as a result of the giant negentropy mechanism in 4-space ongoing in the loaded positive charges (614, 615} in the palladium lattice.

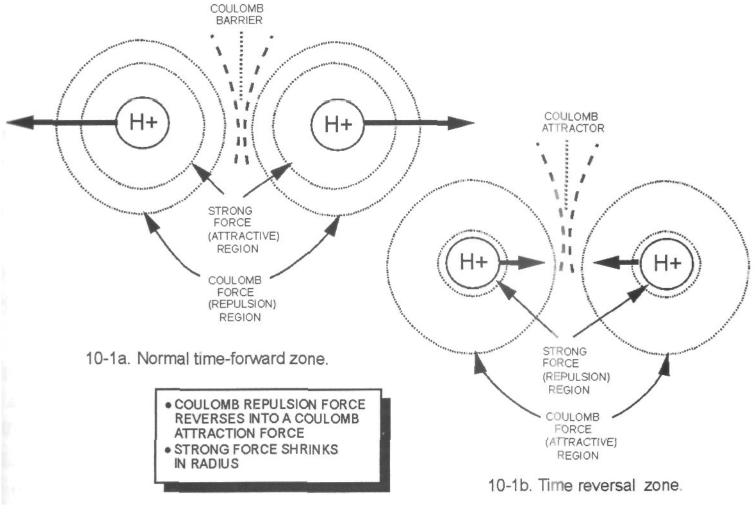
COULOMB BARRIER
H+
H+
STRONG FORCE (ATTRACTIVE) REGION
COULOMB FORCE (REPULSION) REGION
COULOMB ATTRACTOR
H+
H+
STRONG FORCE (REPULSION) REGION
COULOMB FORCE (ATTRACTIVE) REGION
• COULOMB REPULSION FORCE REVERSES INTO A COULOMB ATTRACTION FORCE
• STRONG FORCE SHRINKS IN RADIUS

10-1a. Normal time-forward zone.

10-1b. Time reversal zone.

Figure 10-1 Forces on nuclei of simple ions in time-forward and time-reversal zones.

See Figure 10-1. In Figure 10-la, two hydrogen ions (two free H+ protons) in a normal electrolyte and in a normal time-forward zone are shown. The

Coulomb barrier dramatically increases the forces of repulsion between the two H+ ions as their kinetic energy may be driving them momentarily toward each other. This Coulomb barrier becomes so strong that it stops the approach of the two ions and forces them back apart, or to deviate aside from their paths, before each can ever enter the very short-ranged strong force region of the other. Consequently no nuclear reaction occurs, but only a common chemical reaction. The maintenance of that Coulomb barrier is all that prevents energetic ions from being driven together closely enough to engage the strong force and cause the ions to form a new nucleus (cause a transmutation).

Now see Figure 10-lb, for the same two ions that suddenly find themselves in a momentary time-reversal zone (TRZ). In a temporary TRZ, suddenly like charges attract and unlike charges repel, exactly in reverse of the *normal* behavior of charge attraction and repulsion. The usually increasing Coulomb barrier (repulsive force between the two approaching H+ ions) has disappeared and been replaced by an increasing Coulomb attractor (attraction force). Further, the strong force has been partially reversed and much weakened since the gluon forces are dramatically reduced and fluctuating.[299] The attractive strong force is now a partially *repulsive* force and so it is much weaker. Consequently, the quarks in a proton or neutron are not nearly so strongly bound as they are in a normal time-forward zone (TFZ).

A TRZ represents a highly time-charged excited local state in that local region of the electrolyte. The moment a TRZ is formed, the surrounding ions in the electrolyte outside the TRZ immediately move or deviate their movements to reverse this TRZ action[300] and convert it back to a time-forward zone (TFZ). Hence once it is born in the electrolyte, the TRZ decays rapidly due surrounding ion movement changes, and even more rapidly due to the rapid variation in the fluctuating scalar interferometers in the loaded palladium lattice. In most cases, the random fluctuation in the scalar interferometry is so rapid that the TRZ decays back to a normal TFZ before the two approaching H+ ions can reach each other (reach each other's reduced strong force repulsion region). However, in a certain percentage of approaches, the two approaching like-charged ions will "reach each other" — i.e., each will enter the weakened strong force region of the other before decay of the TRZ occurs.

---

[299] Again we stress the tremendous energy density of the time-energy involved. These cold fusion interactions are actually higher *total* energy reactions than nuclear physics presently uses, even though the *spatial* energy density of the reactions is very small. Together with the now-proven fluctuation theorem of Evans et al., the TRZs and reversal of the normal Coulomb barrier into a Coulomb attractor open up a vast new nuclear chemistry of direct nuclear transmutation reactions at low spatial energy but at very high tempic energy. The ability to directly engineer the quarks and the gluon forces, e.g., opens up a new chemistry of direct subparticle engineering heretofore considered quite impossible by chemical means.

[300] Their approach increases the fraction of photon reactions relative to antiphoton reactions, since the other ions are moving in a TFZ on the average.

See Figure 10-2. In Figure 10-2a, we show the case in which the TRZ lasts long enough for the two H+ ions to reach each other's strong force repulsion zones and form a quasi-nucleus. Here they vibrate back and forth in a dynamic dance around the zero net force axis between their repulsive strong forces and their attractive Coulomb forces. In this case, a new kind of nuclear reaction is set up to occur. Decay from this excited quasi-nucleus state in a TRZ can occur into a real nucleus in a normal TFZ. Energetically, because the quarks are so loosely bound now[301] the preferred decay mechanism when TRZ -> TFZ is by appropriate *quark flipping.*

As a little bit more time passes, the interferometry changes and the TRZ decays, returning back to a normal TFZ. In a change from TRZ to TFZ, the energy change is enormous because time-energy is involved, with energy density equal to mass. The only difference in a proton and a neutron is the orientation of one quark (three quarks make up each of the particles). As TRZ decay starts and progresses, the strong force changes back to a powerful attractive force. It increases its strength far more powerfully and quickly than the Coulomb attractive force reduces its reversal and changes back to a repulsive force. The addition of the extra energy from the now-increasing strong attraction force simply causes one nearly-freed quark in one of the two H+ ions to flip, converting that H+ ion into a neutron n. Instantly the preferred decay product is the deuterium ion D+. So the H+ and the n simply draw a bit closer together, each now fully inside the rapidly increasing attractive strong force region of the other. This converts what had been a quasi-nucleus of two H+ ions into a real D+ nucleus. This result is shown in Figure 10-2b.

---

[301] To appreciate the sheer raw power of the time-energy control of interactions, simply examine the enormous *spatial* energy density required in high energy physics to try to approach the "free quark" condition. Enormous and very expensive *spatial energy* accelerators are used, at the absolute state of the art. By going to the use of the extraordinarily dense time-energy, quarks can readily be freed or nearly freed in electrolytes, and they are nearly freed in more than 600 relatively simple and inexpensive cold fusion experiments {753) on the bench-top to allow new nuclear reactions by quark flipping, where a proton turns into a neutron or vice versa.

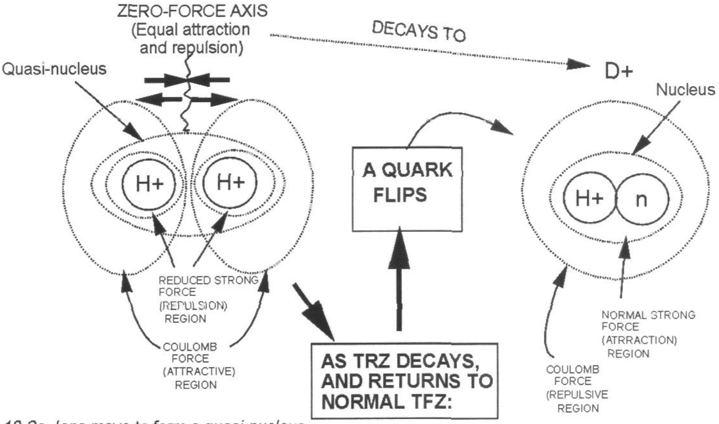
ZERO-FORCE AXIS
(Equal attraction
and repulsion)
DECAYS TO
Quasi-nucleus
H+
H+
D+
Nucleus
H+
n
A QUARK FLIPS
REDUCED STRONG FORCE (REPULSION) REGION
COULOMB FORCE (ATTRACTIVE) REGION
AS TRZ DECAYS, AND RETURNS TO NORMAL TFZ:
NORMAL STRONG FORCE (ATRRACTION) REGION
COULOMB FORCE (REPULSIVE REGION

10-2a. Ions move to form a quasi-nucleus.

10-2b. Quasi-nucleus decays by quark flipping into a new nucleus.

Figure 10-2. Production of quasi-nucleus and its decay by quark-flipping into a new nucleus as the time-reversal zone decays back to a time-forward zone.

Recapitulating: As the TRZ subsequently decays back to a normal time-forward zone (TFZ), the new quasi-nucleus becomes an excited state, and decays. However, the quasi-nucleus decays by very novel means. Because of the time reversal, the energy changes induced in the decay *start* at each spacetime point inside the quasi-nucleus, deep inside the quarks, and proceed outward. The first interaction of the decay mechanism is with the quarks comprising the nucleons (in this case, the protons). With the gluon forces still very much weakened, quark flipping becomes the preferred decay mechanism. Hence one quark in one proton flips its orientation (that is the mechanism of decay!) and the nucleus — now a nucleus comprised of a proton and a neutron — becomes a nucleus of deuterium.

As can be appreciated, the clustering of different types of like-charged but relatively simple positive ions in TRZs in solution, with subsequent decay of the TRZ into a TFZ, initiates a revolutionary new family of nuclear reactions at low spatial energy (but very high temporal energy), completely contrary to, and not included in, the present forward-time high-spatial energy (but very low time-energy) reactions model known in particle physics.

In the future, as these new time-energy-based reactions are extended and mastered, scientists will simply assemble desired nuclides in solutions in the laboratory, at low spatial energy (but high time-energy) and at will. We point out but do not pursue further the fact that these new reactions also open up cheap, practical, electrochemical means of altering nuclear and chemical wastes. We therefore foresee a dramatic development in this area, in order to clean up the presently polluted biosphere of much of the chemical and nuclear waste contaminants.

Indeed, we propose that a very strong program in mastering these new nuclear reactions be launched and supported by the U.S. Department of Energy, so that the nuclear wastes storage problem can be permanently solved, much more cheaply, and without having to store hazardous nuclear wastes for centuries. Instead of merely *talking* about "out-of-the-box thinking and research", the DoE should actually do some of it, over the violent objections of the entrenched conventionalists of the national laboratories. Their own great national laboratories are the major part of the problem, not the solution, for both innovative nuclear reactions and innovative energy systems! The problem is that "in-the-box" entrenched conventionalists do not perform or allow "out-of-the-box" research that strongly threatens their vested interests.

The TRZ decay conversion of two protons into a deuteron — by easily flipping one quark in one proton — is an example of the so-called "nuclear reaction at low energy" that has been so controversial *to* the orthodox scientific community, even in the face of some 600 successful cold fusion experiments. Contrary to the assumption of the conventional physics community and the skeptics, these are not "low *total* energy" physics reactions at all. Instead, they utilize energy density on a level commensurate with that achieved in the largest accelerators available and even much higher. *This is actually a much higher total energy physics than the present conventional high energy physics heretofore known or used.*

Particularly in a deuterium-enriched (deuterated) electrolyte, a variety of simple ion arrangements of D+ ions and H+ ions can occur. Many of these combinations and arrangements can and will occur in TRZs if sufficient loading of the palladium lattice is achieved. Below we will present and explain a few of these new nuclear reactions that have occurred in successful cold fusion experiments. First, to be tidy we must correct the present statement of the conservation of energy law by extending it to include time-energy.

The excess heat usually experienced in the electrolyte is explained by the added negentropy (added energy) from the time-domain, which then dissipates randomly in the solution, producing excess heating. Later we will cite strong experimental evidence for the involvement of such time charging (time-energy charging), with subsequent time-charge (time-excitation) decay as ordinary photons in cold fusion experiments.

10.5 Revision of the Conservation of Energy Law

10.5.1 Unaccounted Transduction Extends the Conservation of Energy Law

Along with the re-emission of all the previously absorbed spatial energy, all the previously absorbed time-charge of the excited electron may be re-emitted with a longitudinal photon/scalar photon pair (a graviton) when the time-excited state decays. In that case, no energy-charging permanent change results to the electron's original energy state. Hence no "electron wiggle" will be created and our normal "electron wiggle" detectors performing such non-transducing longitudinal EM wave interactions will not exhibit an indication. In short, the "pure" time-charging EM LW interaction is unobservable to present instruments. All they "see" is the "same electron" sitting there and "flowing through time".

However, if two or more such gravitons are simultaneously emitted from various reactions in time-decays, the emitted scalar photons and waves may interfere at some little distance. In that case, transverse EM waves and ordinary EM energy will be produced in the interference zone {613}. To the external observer, these EM fields and forces will simply seem to arise spontaneously and emerge from every point in spacetime inside that interference zone, moving outward from there in conventional "propagation through space" manner. In that case, the normal "photon" (actually, graviton) decay processes plus scalar interferometry have resulted in transducing some time-energy into ordinary spatial EM field energy. Such reaction provides the excess heat demonstrated in successful cold fusion experiments.

This is also a true action at a distance reaction, presently unaccounted in particle physics. Action at a distance is required in quantum mechanics, and it has been experimentally demonstrated multiple times in distant photon correlation experiments. Once any action at a distance is admitted because it has been experimentally proven, there is no logical restriction that only a single mechanism exists for it. Consequently, many of the present foundations assumptions of physics and some of the things previously abandoned because of their conflict with those assumptions will have to be re-examined {616}.

Scalar interferometry is indeed a mechanism for action at a distance. While still mostly absent from open physics, it has already been weaponized by some 10 nations of the world. Indeed, U.S. Secretary of Defense William Cohen confirmed in 1997 that (novel) electromagnetic processes are already being used to initiate earthquakes, stimulate volcanoes into eruption, and engineer and control the weather {617}. Whether or not it is present in the particle physics texts, scalar interferometry is being widely used clandestinely in several armies today. It has also been used to shoot down missiles and aircraft as tests. The so-called "Cold War" was not quite as cold as the news media and U.S. scientific community were led to beleieve.

Because of the ubiquitous involvement of unaccounted time-energy in all photon interactions, an extension to the present form of the conservation of energy law is required. The present (special case) law is

$$E(E)_1 + E(m)_1 = E(E)_2 + E(m)_2 \qquad [1]$$

where E(E) = energy in conventional non-massive energy form, E(m) = energy in mass form, subscript 1 refers to the measurement at time 1, and subscript 2 refers to the measurement at time 2. This is just the well-known Einstein formulation that the sum of the ordinary energy and the mass-energy is conserved.

However, that law is for a special case. It assumes that there is no transduction of time-energy into either mass-energy or ordinary energy. I.e., it assumes 3-symmetry and t-symmetry in EM energy flow. Since every charge in the universe already violates both that 3-symmetry and t-symmetry of energy flow, the fundamental assumption in the Einstein formulation is falsified by the general case. The conventional conservation of energy law is a special case where only the two mentioned energy flow symmetries are upheld. When they are violated, then transduction of EM energy between 3-space and time occurs. For that case, the conservation of energy law must be extended to

$$E(E)_1 + E(m)_1 + E(t)_1 = E(E)_2 + E(m)_2 + E(t)_2 \qquad [2]$$

The **new** law simply states that the sum of the ordinary spatial energy, the **mass-energy,** and the time-energy is conserved. This extension of the conservation of energy law becomes important in successful cold fusion experiments, where transduction of time-energy into ordinary heat energy occurs and where one encounters very anomalous reactions differing from those reactions known and accepted in conventional nuclear physics.

10.5.2 Transduction Between Time-Charging and Spatial Energy-Charging

The oscillating energy (spatial and temporal) in a moving EM wave in the vacuum represents a moving oscillation of spacetime curvature. Simply put, the so-called electromagnetic wave in space is actually a gravitational wave in space, since it is an oscillation of the local energy density of spacetime and therefore an oscillation of the curvature of local spacetime. It "transports" gravitons (scalar photon/longitudinal photon pairs) rather than merely photons. So let us now consider it as a gravitational wave.

In the absorber under special conditions, a small fraction of the graviton wave time-density absorbed will sometimes change (transduce) its time-energy into what we call transverse photon energy effects {618}, with something like a 10^{17} gain in spatial energy in joules compared to the time in seconds that actually transduces. In other words, 1 second of time transducing into EM transverse wave energy (using the standard EM modeling approach) will yield approximately 10^{17} joules of spatial EM energy. Since one can change energy from one form to another, *one can change time-energy into 3-spatial energy* and 3-*spatial energy into time-energy,* the two being merely opposite sides of the same coin, and both always involved in any situation exhibiting electrodynamic dipolarity.

In one's physics model, the choice of fundamental units is purely arbitrary. Everything can be expressed in a single fundamental unit, as pointed out by Jackson (see quotation from Jackson at the beginning of this chapter).

In a situation where 3-symmetry and t-symmetry in EM energy flow are broken, part of the *time charging* of the atomic electrons (or ions) sometimes changes (transduces) into *spatial energy charging* and vice-versa. This cross-dimensional decay of time-excitation states results in the emission of some extra transverse photons that may be either photons or antiphotons, depending on the exact interactions. The point is, *tremendous excess EM spatial energy can appear and either be emitted as heat or interacted to perform new nuclear interactions as extraordinarily high local energy density.* This is the master mechanism enabling anomalous cold fusion transmutation reactions. The amount of time transduced will be so small that it will usually escape notice, unless stringent efforts are used to measure changes in the local flow of time.

The excess photon energy (excess 3-space EM energy) appearing in these rather random but powerful transduction interactions often accounts for the production of anomalous excess heat - for example, as occurs in many cold fusion experiments. Where scalar interferometry is maximized, a more exact action will be detailed below for the specific nuclear transformations that occur by novel new nuclear reactions. These novel nuclear reactions occur at low *initial* spatial energy but with some time-energy transducing into additional very high spatial energy to initiate the interactions.

10.5.3 Resolving the Energy Conservation Problem in Cold Fusion

When transduction is involved, standard "transverse EM waves only" computations will seemingly yield violation of energy conservation — e.g., from the emission of excess heat. That is an erroneous interpretation; energy overall is conserved, but some of the unaccounted *time-energy flow* is transduced into 3-spatial transverse photon energy dissipation. Neither 3-spatial energy nor mass-energy nor time-energy alone, or in any pairing, need be individually conserved; only the combined *total energy of all three energy components* is conserved. This resolves the present major "sticking point" between the cold fusion researchers and the conventional "transverse wave" nuclear and electrodynamic communities regarding energy conservation difficulties and perpetual motion accusations. Those making the accusations and those defending against them are both using only a special-case conservation of energy law, and a more general law is required.

Let t = time in seconds, and W= energy in joules. Rounding the speed of light to 3 x 10^8 meters/sec, the energy W_{TW} emerging as transverse wave EM energy in 3-space by the transduction of time into 3-space energy is approximately

$$\text{One sec} = c^2 \text{ joules} = 9\times10^{16} \text{ joules} \quad [3]$$

and

$$t \Rightarrow W_{\text{TW}} = 9\times10^{16} \times t \quad \text{(in joules)} \quad [4]$$

Let us present the new energy conservation law in slightly different form. Let W_{TW} = energy measured in spatial energy form, conveniently taken as transverse wave (TW) EM energy. Let W_m = energy in mass form (energy compressed by c^2). Let W_t = energy in time form (energy compressed by c^2).**The** present conservation of energy law is

$$W_{\text{TW}} + W_{\text{m}} = W_{\text{(total)}} \quad [5]$$

The new extended conservation of energy law is

$$W_{TW} + W_m + W_t = W_{(total)} \qquad [6]$$

The conventional particle physics accounting will see only

$$W_{TW} + W_m = W_{(total)} \qquad [7]$$

The conventional accountant will be unable to see

$$t \Rightarrow W_{TW} \qquad [8]$$

so he will observe that the experimenter is reporting that

$$[W_{TW} + W_m]_1 \neq [W_{TW} + W_m]_2 \qquad [9]$$

which in his lexicon is a violation of the conservation of energy law itself. In short, he will identify it as a typical example of a claimed perpetual motion machine or interaction. Then the defender of the faith will wax eloquent, applying labels such as "perpetual motion nonsense", "voodoo science," etc. He himself is practicing a hoary old "voodoo" science and ignoring newer science. None of these defenders of the faith even know what powers an electromagnetic circuit or how it is that every little charge in the universe is able to continuously pour out observable EM energy in all directions, with no observable EM energy input.

On the other hand, a reasoned response will be given by real scientists still following scientific method, when faced with the numerous replicated successful experiments. The scientific method insists on the absolute priority of the replicable experiment over the predictions *of any* theory. Consequently, such reasonable scientists will advise that the protagonists cease their vitriolic rhetoric and let real science take its course; i.e., let further experiments confirm or refute reported successful results of cold fusion experiments.[302]

---

[302] Precisely such a reasoned response by chief editor Donald Kennedy of the journal *Science* has recently been given, providing a breath of fresh scientific air into an attempt by "defenders of the conventional faith" to destroy publication of important experimental results in sonoluminescence. The article which courageous editor Kennedy approved for publication — over intense objections of the professional critics — is R. P. Taleyarkhan et al., "Evidence for Nuclear Emissions During Acoustic Cavitation," *Science,* Vol. 295, Mar. 8, 2002, p. 1868-1873. We heartily commend Editor Kennedy for this classic example of how science should be conducted — so that experiments, not prevailing dogma, determine what is valid and what is invalid.

What the professional skeptic will not see is that his own expression of the conservation of energy law has been incomplete, and that it is his own inadequate "measurement yardstick" assumption that is now in error. The *extended* conservation of energy law of course still holds, but the *conventional* special case law is violated because its fundamental assumption — that time is not transduced to spatial energy in the experiments —has been violated in the experiment.

Indeed, with the publication of the work of Evans et al., experimentally proving the fluctuation theorem and the high probability of time reversal zones appearing for short but significant periods at up to micron levels, the second law of thermodynamics has been rigorously proven to be violated. Hence the conventional scientific community's rather dogmatic objection to the extra heat of cold fusion interactions is no longer scientifically justifiable.

Understanding of the extended conservation of energy law and the permissible violation of the second law of thermodynamics resolves the present thorny dissension within the scientific community over the excess spatial energy produced by cold fusion. It also resolves the issue of what could possibly be a source of such intense energy density as is required for nuclear transmutation reactions.

Any ordinary system is indeed an open system, freely and continuously exchanging energy with its *temporal* environment as well as the seething active vacuum.[303] If the system is in temporal energy equilibrium in that exchange, then the present restricted conservation of "spatial and mass" energy law applies and one does not have to consider the interactions between supersystem components. If the system is placed in disequilibrium with its temporal energy environment, then supersystem components interact asymmetrically and the extended conservation of energy law applies. In that case, excess spatial energy can be and will appear in the system whenever time transduction into spatial energy is occurring.

---

[303] But see our Appendix A. Present thermodynamics rather inanely defines a "closed" system as one where no mass exchange crosses its boundary with its environment, but energy exchange can and does. In that view, a "closed" system can be far from equilibrium in its energy exchange with its active environment, and thus demonstrate the five magic attributes of such disequilibrium systems. In this book we have deliberately chosen to treat a "closed" system as one in which neither energy nor mass exchange crosses the system boundary — which classical thermodynamics refers to as an "isolated" system. We explain all this (and why) in Appendix A.

The implications of cold fusion energy transduction are that

(1) Time is being "burned" or transduced as a "fuel" and source of energy,

(2) Transduction of only one microsecond per second will yield $9\text{x}10^{10}$ watts steadily. That's 90,000 megawatts of energy per second — the equivalent of 90 large 1,000-megawatt power plants — if all the transduced energy were collected and used to power loads with 100% efficiency.

(3) Thousands of new nuclear reactions are made possible and feasible by local transduction of time-energy into 3-spatial EM energy.

(4) Appropriate capture and use of a portion of the excess energy can be accomplished to power loads.

Even if we capture and utilize only 1% of the transduced time-energy to power the loads, where the *time-energy to spatial energy transduction* is only the transduction of 1 microsec/sec, that system will produce the normal EM energy to power 900 megawatts in the load. That's almost 90% of the capability of a modern 1,000-megawatt power plant. That is quite a respectable electrical power system. It only requires the transduction and "burning" of one microsecond of time per second, without any nuclear wastes or hydrocarbon byproducts.

As can be seen, transducing time into energy is by far the most efficient way to go to develop new power plants for the future. This is particularly true since every charge in the universe already transduces time-energy flow into spatial EM energy flow. Indeed, we may consider that the negative charge transduces time-energy flow into the spatial outflow of EM positive energy, while the positive charge transduces negative time-energy flow into an outflow of EM negative energy. The process is useful not only for power applications, but also for transport applications as we briefly pointed out in Chapter 8.

10.6 Some Resulting *Low Spatial Energy* Nuclear Reactions

Refer again to paragraph 10.4 above. Summarizing: The formation of time-reversal zones (TRZs) is what enables the appearance of the new nuclear reactions in the electrolyte in cold fusion experiments. It does this by changing the Coulomb barrier between like charged ions into a Coulomb attractor. Further, the rate at which TRZs form and decay in the solution is a function of the frequency and intensity of the scalar interferometry from the loaded palladium lattice. This means that the appearance of the new nuclear interactions in the electrolyte is also a function of the degree of loading of the palladium lattice, as well as the "double surface" interferometry preparation of the palladium, its cracks, etc. These are the major variables. Once the major variables provide the solution for significant formation of TRZs, then the following reactions can and do appear.

The sample reactions we present here are just a few of the thousands of new reactions possible. These example reaction products have been found in a great many successful cold fusion experiments in various laboratories around the world. Some 600+ successful experiments have now been done, and a sizable literature has built up and continues to be built up on the experiments and their results.

For our sample reactions, the primary mechanism of interest consists of several parts:

(i) Formation of a TRZ, in which two positive ions (we shall look at H+ and D+), now attract each other. Also, the gluon forces in the proton and neutron, in the ions of interest, are significantly lessened, so that the quarks in each proton and neutron are very much more loosely bound. Consequently, decay from an excited state by quark flipping to turn a proton into a neutron or vice versa becomes an attractive option. The strong force is therefore reduced in radius.

(ii) Two positive ions now attract each other so closely that each enters the edge of the weakened strong force region of the other. This forms a *quasi-nucleus* because of the limited involvement of the weakened strong force.

(iii) The other ions in solution surrounding the TRZ immediately move (as soon as the TRZ is formed) to negate the TRZ and decay it back to a time-forward zone (TFZ). This action initiates the decay of the TRZ.

(iv) As the TRZ lessens and then turns back into a TFZ, the strong force expands its size faster than the novel Coulomb "like attracting like" force reduces and reverses.

(v) The rapidly expanding strong force of each of the two ions in the quasi-nucleus fully envelops the other ion, increasing the disequilibrium of the two-ion quasi-nucleus. A condition is reached where the most probable mode of decay is the flipping of one quark in one of the positive nuclei.

(vi) The quark flips, turning that H+ ion into a neutron n. This is a quasi-nucleus of deuterium (D+), still in an excited state but with less excess potential energy in the excited state. At this point, the TRZ is vanishing and the TFZ state is returning.

(vii) As the TRZ vanishes and TFZ state increases back to normal, the most probable mode of decay increasingly is to a full deuterium nucleus. Hence the H+ and n simply draw completely into normal deuterium binding position, bound by the normal strong force. This constitutes a normal D+ ion, now existing in a normal TFZ.

(viii) So by formation and decay of a significant TRZ, two H+ ions have been drawn together into a quasi-nucleus consisting of two H+ ions partially bound by a weakened strong force. As the TRZ decays and the Coulomb repulsion resumes, the strong force increases back to full strength faster than the Coulomb repulsion force returns and repels the two H+ ions. Since the quarks are much more loosely bound than normal, a quark in one of the H+ ions has flipped, turning the ion into a neutron n. The H+ ion and the neutron, partially bound already in a quasi-nucleus, simply "tighten" into a normal D+ nucleus as the TFZ condition is fully resumed.

The nuclear reaction equation for the above interaction may be written as

$${}^1H_1^+ + {}^1H_1^+ \Rightarrow (({}^1H_1^+ + {}^1H_1^+)) \Rightarrow ({}^1n_0 + {}^1H_1^+) \Rightarrow {}^2H_1+ = {}^2D_1^+ \qquad \textbf{[10]}$$

In equation [10] we use the left superscript as the number of nucleons in the ion nucleus, the right subscript as the protons in the ion nucleus, and the right + superscript to show the overall charge of the ion and its sign. We use the expression in parentheses to show the involvement of a time-reversal zone (TRZ). Double parentheses show a stronger TRZ than single parentheses. Thus from double parentheses to single parentheses to no parentheses shows the initial formation and subsequent decay of the TRZ.

Another interaction involves a D+ ion and an H+ ion, to produce tritium. This is:

$$^{2}D_{1}^{+} + {}^{1}H_{1}^{+} \Rightarrow (({}^{2}D_{1}^{+} + {}^{1}H_{1}^{+})) \Rightarrow ({}^{2}D_{1}^{+} + {}^{1}n_{0}) \Rightarrow {}^{3}H_{1}^{+} = 3T1^{+} \quad [11]$$

That reaction transmutes a deuterium ion and hydrogen ion into a tritium ion. Without showing it, in a TRZ three $^{2}H_{1}^{+}$ ions may also attract into a quasi-nucleus, and as the decay of the TRZ occurs, first one proton turns into a neutron and then a second one does also, by quark flipping. This nuclear reaction also produces a fusion into tritium.

Another reaction between two deuterium ions is"

$$^{2}D_{1}^{+} + {}^{2}D_{1}^{+} \Rightarrow (({}^{2}D_{1}^{+} + {}^{2}D_{1}^{+})) \Rightarrow [2({}^{1}H_{1}^{+} + {}^{1}n_{0})] \Rightarrow {}^{4}He_{2}^{+} = \alpha \quad [12]$$

That interaction — particularly in pre-deuterated electrolyte solution — gives the excess a particles produced in a great number of the cold fusion experiments.

A rarer but still occurring interaction is:

$$4[{}^{1}H_{1}^{+}] \Rightarrow (({}^{2}D_{1}^{+} + {}^{2}D_{1}^{+})) \Rightarrow {}^{4}He_{2}^{+} = \alpha \quad [13]$$

Indeed, reaction [13] may occur to completion and formation of a particles in some of the transmutations in an electrolyte, while it proceeds only to the intermediate phase in others. In that case, both deuterium ions and a particles may be produced out of the same overall "chain of reactions" where some reactions proceed all the way and some do not.

These interactions are directly using and transducing time-energy by use of the TRZ to time-reverse the Coulomb repulsion law for like charges. From the reactions, excess energy given off as heat may and will occur.

The nuclear physicist and nuclear chemist can immediately see a great many new nuclear reactions now made possible by the adroit production and decay of TRZs. A great number of these new nuclear reactions are now possible by this means, which are impossible in ordinary nuclear chemistry without the deliberate formation of TRZs.

These few examples serve to illustrate the process, and these reactions represent results already achieved in numerous cold fusion experiments worldwide, by many researchers and multiple laboratories. Now we turn to other evidence strongly supporting the novel nature of these reactions rather than several other proposed reaction equations yet to be proven. The other proposals do not explain the next type of new experimental reaction.

10.7 Time-Energy Can Generate Instrumentation Anomalies

Strong evidence for time-charging and time-charge decay is experimentally demonstrated in electrolyte experiments where time-reversal zones (TRZs) are operating, and thus where some localized processes are running backwards in violation of the second law of thermodynamics. Weak time-charging of photon-absorbing and emitting matter, with special patterns of spacetime curvatures (engines) induced by time-density oscillations and structuring, is persistent in matter, having "charge-up" and "decay" times that may last for minutes, months, or even years. A very wide range of previously anomalous phenomena, both in inert and living matter, is explainable by these new interactions. This is very probably also related to the "memory phenomena" noted by Kondepudi and Prigogine[304] as outside the present thermodynamics, but being worked on at the forefront in what is called "extended thermodynamics". Quoting Kondepudi and Prigogine, p. 460:

> *"...interesting memory effects ...appear for long times (as compared to characteristic relaxation times). ...nonequilibrium processes may have 'long-time tails'. In other words, the approach to equilibrium is not exponential, as was generally believed, but polynomial ...which is much slower. ... As a result, Nature has a much longer memory of irreversible processes than it was thought before. Again this shows that local equilibrium is an approximation, albeit a very good one. "*

In this book, we focus primarily upon some typical anomalous measurements in China Lake experiments that are directly involved with time-charging and decay effects, lending strong support to our thesis of the novel TRZ-induced nuclear reaction chains. Note that these instrumental anomalies are also strong evidence for the association of time-charging and time-charge decay with the proven fluctuation theorem of Evans et al.

---

[304] Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998, p. 459. For a discussion of extended thermodynamics, see D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

at the Australian National University. We believe this to be a contribution to that very important work of Evans and his colleagues.

10.7.1 Ionization Detectors, Transduction, and EM LW Interferometry

Ionization processes obviously are processes where the partial transduction of LW time interaction into 3-spatial energy interaction would yield ordinary excited energy states of the affected electrons or ions. In turn, this would lead to ionization (e.g., of the atoms and molecules of a gas, liquid, etc.) Hence, ordinary ionization detectors such as Geiger Miiller tubes have the innate *capability* of detecting such transduction that generates ionization, and thereby have some capability of serving as "detectors" of longitudinal EM wave interactions (and time-charging) ongoing in the detector. In the usual situation, no excess time-charging results and no time reversal zones result. In that case, symmetry of energy flow in the time-domain exists, as does symmetry of energy flow in 3-space. Absolutely normal chemistry and nuclear chemistry results, and the nuclear detectors function normally, as familiar to every laboratory.

However, when significant TRZs emerge and persist for short periods, the symmetry of both time-energy flow and 3-spatial energy flow is broken. In this case, transduction of time-energy into 3-spatial energy (and vice versa) can result in anomalous ionization in nuclear ionization detectors, even in the absence of normal transverse EM nuclear radiation. The trick is to use and compare different detectors and different types of detectors to clearly demonstrate anomalous ionization detection effects.

As an example, a Geiger Muller tube will detect any ionization of its internal gas, regardless of what caused that ionization. If the ionization is caused by normal ionizing radiation from nuclear reactions, the instrument will detect and read a resulting internal ionization of the gas inside the tube.

On the other hand, if time-charging and decay are involved, then the asymmetry resulting in 3-spatial energy excitation of the gas inside the Geiger Muller tube (due to transduction of some time-energy into excess 3-spatial energy) may be sufficient to ionize the gas. In that case, the instrument will read "as if nuclear radiation were present. Note that there is indeed "ionizing radiation" present in the gas. However, now this ionizing radiation is not transverse EM wave in nature, but is due to the time charging and decay, and transduction of time into 3-spatial energy in the detecting gas.

This capability of an ionization detector to ionize in the presence of non-transverse wave ionizing radiation and energy can be used to provide an indication of time-density waves being formed and interacting in a process (such as a cold fusion process) where some transduction from time-energy to 3-spatial energy occurs.

Further, the previous "time history" of the individual instrument plays a part in whether the instrument "reads" or not for a given transduction situation. There will be a statistical variation of the actual "already present" low-level time-charge (and in its internal structure and engines) between instruments, even those made in the same factory on the same day from the same batch of materials and parts. That is because the time-charging and discharging history of each instrument has been different.

The production of transduced TWs, however, must be a function of scalar interferometry in the gases or other detecting media of the instrument, since scalar interferometry *creates* all transverse EM waves, as shown by Whittaker {619} and confirmed by Evans *et al.* {620}. In such interferometry, the entire inner structures of the transverse EM waves formed in the interference zone also interact wave-to-wave. Thus, this kind of interferometry is extremely sensitive to the exact internal longitudinal EM bidirectional wave structure (internal engine substructure) of the involved potentials.

As a consequence, the ability of an ionization detector to transduce a specific set of LWs and give ionization detection will vary appreciably, including for different variations in the instrument's original manufacturing process, and even for the specific past photon interaction history and experience of the individual detector itself. One detector's set of cumulated internal time-charges (and hence spacetime engines) may vary considerably from those of a second detector of exactly the same type. In general, multiple ionization detectors are unlikely to all detect a given time-density EM wave emission and interaction set {621}.

Even for two detectors of the same brand, from the same manufacturing plant, and from the same batch of manufacture, it is likely that significant differences in LW detection of a specific TDW set will occur because of the "past history" time-charge differences of the individual detectors. Indeed, a likely phenomenon is that, when one ionization detector detects the LW emissions (transduces them), several others will not detect them at all. This is one of the peculiarities of the new unified field area that must be overcome by further research in order to develop reliable, calibrated TDW and transduction detectors. It is one of the primary problems - if not *the* primary problem - of cold fusion research to resolve this *"time-charge experience history"* problem, so that transduction detectors having calibrated, uniform responses to given TDW sets and interactions can be provided.

At present, no one has the foggiest notion as to how to "calibrate" a nuclear ionization detector that one desires to use as a time-energy transduction detector. Speculating, we would envision such calibration (in the transduction detecting functional response) to become possible when small, standard, calibrated sources of transduction are developed and available. In that case it will be possible to formulate procedures whereby varying the output of the transduction source will vary the transduction irradiation of the GM tube. The level of transduction intensity required to create ionization in the instrument will be a direct indicator. When small, calibrated sources of TDWs are also available, it will be possible to irradiate the GM tube to a specifically desired rate and level of time-charge and decay. In that case, an array of GM tubes can be "standardized" or "calibrated" for transduction detection. So far as this author is aware, those developments have yet to be accomplished or even undertaken. Indeed, the need for such a development program has not even been realized.

10.7.2 Experimental "Detector Transduction Phenomena" in Electrolysis

Let us look at one set of ongoing scientific experiments where just such anomalous detector results have been obtained.

Researchers at the Naval Air Warfare Center at China Lake, and at the Department of Chemistry, University of Texas at Austin, have detected precisely the kind of "anomalous" radiation and transduction effects we expect to be in the ionization processes of multiple kinds of detectors {622}.

According to a report by Miles and Bush {623}, anomalous radiation at China Lake was first detected by the exposure of dental X-ray films in two experiments producing excess power (excess TW energy emission). Film used in a control study during these experiments showed no exposure. There was also no exposure of similar films in more than 20 experiments where no excess power was present. In other words, the presence of excess heat (excess TW energy emission) strongly indicates the presence of transduction. In turn, the presence of transduction implies the presence of scalar interferometry wave interaction, where TDW waves and LW waves **are** absorbed and **a** mix of LW and transduced TW waves are emitted consisting of gravitons (paired scalar and longitudinal photons). This only occurs when significant time-charge excitation has built up. Taken in reverse, the presence of LW wave absorption and emission — with transduction into detected excess ionization energy — directly indicates the emission of "anomalous" TDW or graviton radiation, as detected by the resulting TW exposure of the x-ray film. So the "anomalous" radiation detections are consistent both forward and backward.

The film in manufacture is quite strongly quality-controlled. Also, it is used (interacts and indicates) only once. Its past time-charge history has minimal variations between two samples of the film, because there is no repetitious past *detection interaction* history. Hence one would expect a high degree of uniform film detection interaction responses from film to film, to the same graviton emission set. That is precisely what occurs.

We conclude that the behavior of the film in the combined China Lake experiments clearly shows the time charging, graviton emission and interaction, and transduction nature of what is happening when it is known that no normal nuclear radiation is present.

Geiger-Muller (GM) detectors and sodium iodide (NaI) detectors were also utilized when electrolysis experiments using heavy water were ongoing. We again accent that a Geiger-Muller tube does not detect nuclear radiation per se; instead, it detects *anything* that will cause its internal gas to ionize sufficiently. Sufficient transduction in graviton absorption-emission interactions in a Geiger-Muller tube will cause the counter to indicate, because it ionizes the gas and produces an ionization discharge. However, both the specific transduction and scalar interferometry aspects of the ongoing experiment are involved, as well as the previous background time-charge history of the Geiger-Muller tube counter.

Several Geiger-Muller detectors gave anomalously high readings, reaching some 73 sigmas above normal background counts. Most experiments (i.e., most GM detectors), however, gave normal radiation counts, and no anomalous count rates were ever observed when the experiments were turned off. So these anomalous results are differentiated by using multiple, carefully calibrated Geiger-Muller tube detectors. Our interpretation here is that

(i) There was a variation in the presence of transduction and scalar interferometry from experiment to experiment,

(ii) The majority of the experiments did not produce significant time-charge and sufficient transduction or scalar interferometry to cause detection (ionization discharge) on most (typical) Geiger-Muller tubes, and

(iii) The transduction effects in the ongoing experimental process were mostly of the rapid variety, and not due to long-term "charge-up" effects conditioning the time-charge aspects and structuring of the experimental apparatuses.

(iv) By using multiple detectors, the probability of one or more of the instruments having increased time-charge from specific time-charge histories was increased.

(v) With sufficient experiments and sufficient detectors, there resulted a high probability of having at least one or more detectors capable of detecting the levels of graviton radiation and transduction expected from the experiments.

(vi) A direct correlation would be predicted — and was observed — in the appearance of the anomalous radiation effects and the expected time periods required to load the palladium with deuterium. As reported by Miles and Bush, *ibid.*:

> "... *the anomalous radiation would appear within a few hours in the co-deposition experiments where the palladium is loaded with deuterium as it deposits from solution. In contrast the appearance of anomalous radiation required days of electrolysis for the palladium rods that load much slower.*"

We previously discussed the major variables indicated from many cold fusion experiments to include the strong correlation with the degree of loading of the palladium lattice. Hence the effect pointed out by Miles and Bush would indeed be predicted by the present approach and proposed mechanism.

Our interpretation is that the experiments show the time-charge rate effect to be expected in such graviton radiation and transduction interaction phenomena using collection of ions (in this case deuterium) that are much heavier than electrons. The faster the deuterium loaded, the greater the buildup of the interaction of the deuterium in phase conjugating and self-targeting iterative interactions — and therefore the greater the increase in scalar interferometry interactions — inside the palladium lattice. The rate of graviton radiation production and transduction production increases as some function (not necessarily linear!) of the rate of loading of the **deuterium.** As the rate of graviton and transduction production increases, so does the expectation of anomalous ionization effects in the nuclear radiation detectors. And so does the rate of production of TRZs and the appearance of the new nuclear cold fusion reactions at low spatial energy but very high time-energy.

10.7.3 Experiment Confirms It

A clinching observation is given by Miles and Bush, *ibid.*, as follows:

> *"One GM-detector would measure anomalous radiation while another GM-detector would be "blind" to any anomalous effect. A few experiments, nevertheless, gave simultaneous anomalous effects from two different radiation detectors."*

Here our comment is that this precisely fits our expectations for demonstrating involvement of

(i) Time-charging and decay interactions,

(ii) Consistency with transduction expectations,

(iii) Consistency with the fact that two different radiation detectors may or may not exhibit or detect the anomalous ionization effects, depending upon the individual past time-charge histories of the two detectors, and

(iv) Predictable variations in the individual experiments (based on palladium lattice loading time, specific geometry, ionization state of the chemical solution, internal time-charge state and decay rate, and structuring of the apparatuses and their parts, etc.) when transduction-associated phenomena are considered.

The anomalous non-nuclear, non-transverse wave — but ionizing — EM radiation detected at China Lake, and the resulting peculiarities in the measurement apparatus functioning, do strongly support the present proposed approach, the proposed graviton radiation phenomena, and the time-charging and decay and transduction mechanisms presented. It is decisive that sometimes two detectors both indicated, and sometimes one indicated and the other would not. Sufficient replications have ruled out any defect in a single given instrument as the cause of the anomalous detections. That immediately eliminates from consideration all purely transverse EM wave phenomena, conditional only upon the premise that the instruments were properly functioning and calibrated, and that sufficient replications were made for decisive statistics. Those conditions were all met in the numerous experiments producing these results over a long time and in many repetitions.

As a final thought on the subject, we pose the possibility that the longitudinal EM waves involved may be the B(3) longitudinal field waves advanced convincingly by Evans {624a-624c}. We would recommend the attention of leading theorists and experimentalists to this possibility, as a matter that should be investigated both theoretically and experimentally.

10.8 Additional Implications for the Future

10.8.1 Longitudinal EM Waves and Undistorted Progressive Waves

There is a revolution in electrodynamics presently underway, due to the unique characteristics of *longitudinal* EM waves[305] and *pseudo-longitudinal* EM waves. A summary by Rodrigues and Lu {625} is particularly revealing.

A pure longitudinal EM wave has infinite energy and infinite velocity — i.e., it appears everywhere at once. "Propagation through space" does not apply; instead, one has stumbled into that weird region where spacetime is multiply connected. In the real world, one meets imperfect longitudinal waves — or *pseudo-longitudinal* EM waves — that still retain a transverse energy density variation residue. These waves are called *undistorted progressive waves* (UPWs).

Such UPWs can move in space faster than the speed of light. This should not be too surprising, since superluminal velocity of intelligent signals has been positively demonstrated by several researchers in quantum-tunneling experiments. As an example, Nimtz {626} and his colleagues have beamed Mozart's 40th Symphony through a barrier in a waveguide at a speed of 4.7c. So the old notion that information cannot be transported at superluminal velocities is passe. The Fogal semiconductor has been used in experimental transmitter-receiver pairs to demonstrate superluminal transmission, such as TV pictures and programs. Direct signal transmission from earth to satellite back to earth and through many en route amplifiers and processors has been demonstrated by Fogal to occur far beyond the speed of light. Again, theoretical summary papers by Rodrigues *et al.* {627a-627f} are revealing.

---

[305] We have pointed out a direct way to generate longitudinal EM waves, by simply oscillating the magnitude of the **Lorentz self-regauging** of a system.

10.8.2 Death Knell of the Speed of Light Limitation

Another giant shock has emerged in astrophysics during the last few years, and it is now threatening a dramatic revision to physics. Heretofore, it has been fashionable to assume that gravity itself propagates at light speed. A great deal of very comfortable physics theory has been built around that assumption.

Well, the assumption is not true. Experiments in astrophysics continue to refute it. It now is almost certain that the speed of gravity is at least 2×10^{10} c. A beautiful though somewhat controversial paper by Tom Van Flandern {628} summarized the entire situation. He articulates the comfortable general relativity view held to date {629} as follows:

> *"...GR proposes that such changes appear to act instantaneously in the 'near field', but eventually show their true, light-speed-delayed character in the 'far field', which is conveniently beyond our present ability to observe. The necessity of this dual behavior is to prevent the logical need for changes to continue to appear to act instantaneously at ever increasing distances, even to infinity."*

In fact, Van Flandern {630} points out that general relativity already admits superluminal effects in the near field region. He also pointed out the startling nature of the experimental results to relativists, and points out that an attachment to a paradigm has been involved:

> *"To many, this result* [astronomy's verification that gravity's speed $v_g >= 2 \times 10^{10} c$] *is so contrary to 'common sense' in the light of relativity theory as to be absurd. But Thomas Kuhn has cautioned all scientists to avoid the trap of becoming so steeped in a prevailing paradigm that it starts to seem like common sense and makes other ideas sound and feel wrong. Eventually, even one's professional status can become linked to a prevailing paradigm. "*

As Van Flandern further emphasizes, it is not difficult to show by computer calculations of astrophysical situations that any strong limit on the speed of light destroys conservation laws:

> *"...anyone with a computer and orbit computation or numerical integration software can verify the consequences of introducing a delay into gravitational interactions. The effect on computed orbits is usually*

> *disastrous because conservation of angular momentum is destroyed."*

So here we have a new paradigm in the making. This is a startling change to standard relativity. Either we have to give up the conservation of angular momentum, or we have to accept superluminal speed of gravity in the far field. Years of struggle to detect quadrupole gravitational wave radiation have failed; no direct detection has ever been made. Indeed, the rather arbitrary nature of the entire quadrupole radiation assumption is not required by general relativity, as long pointed out by leading Russian scientists. For example, Vlasov and Denisov {631} *et al.* bluntly state it in these words:

> *"...Einstein's well-known quadrupole formula, which is usually employed in general relativity for calculations, is not a consequence of general relativity and is not contained in it. "*

Therefore, it may not really be so surprising that physicists have not been able to detect such quadrupole radiation in a great many laboratory experiments. It is not in general relativity at all.

Of course, our thesis developed in this book is that what we have long called "EM radiation" is in fact gravitational radiation, always involving paired scalar and longitudinal photons and hence time-density waves and longitudinal EM waves. We have pointed out that prior to observation there is no such thing as a separate photon, but there is such a thing as two correlated photons, one time-polarized and the other longitudinally polarized. We believe that there is sufficient meat in our proposed explanation of the anomalous cold fusion results and the anomalous instrument results at China Lake to warrant serious and strenuous investigation of these hypotheses and proposed mechanisms.

Another oddity - usually never stated by Western physicists - has long been known but ignored. When Einstein characterized the gravitational field as a pseudotensorial field, and not a field in the sense of the Faraday-Maxwell field, the consequences were that the 3-space energy laws of the familiar kind really do not exist in general relativity. Shortly after Einstein published his theory of general relativity, this absence of familiar energy conservation was pointed out by Hilbert {632} in these words:

> *"I assert... that for the general theory of relativity, i.e., in the case of general invariance of the Hamiltonian function, energy equations... corresponding to the energy*

> *equations in orthogonally invariant theories do not exist at all. I could even take this circumstance as the characteristic feature of the general theory of relativity. "*

With our publication of the giant negentropy of 4-space once arbitrary 3-space symmetry and time-symmetry in EM energy flow is removed, one of the truths involved in Hilbert's remark was finally recognized. In addition, what Sen called "the most difficult problem in classical and quantum electrodynamics" {633} has been explained by giant negentropy {12}. Also, it is consistent with an appropriate reinterpretation of Whittaker's 1903 decomposition of the scalar potential {600, 615}. Finally, it is consistent with quantum field theory {634}.

Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that, in general relativity, conservation laws for 3-space energy, momentum, and angular momentum — in the sense we are accustomed to in the rest of physics —in principle may be impossible. That is because one can readily remove 3-space energy symmetry when curved spacetime is permitted. Again, Russian scientists such as Logunov and Loskutov {635} have long pointed out the "unthinkable" and "astounding" fact that the predictions of general relativity are not unique.

In the West until recently relativists were conservative, and this resulted in the sidelining of innovative Western physicists who challenged the prevailing interpretation or extended it. With the experimental falsification of one of the fundamental tenets — that gravity moves at speed c in the far field — of the prevailing interpretation of general relativity, it appears that we may be approaching an asymptotic burst of great new theoretical work and a dramatic new reinterpretation. If it ever gets into production, we believe that the Fogal semiconductor will usher in that great change of communication speed paradigms that now appears imminent.

Nonetheless, general relativity — whatever the modifications and extensions now called for — will remain a very useful tool indeed. There is simply nothing else on the horizon that can replace it, string theory notwithstanding. The reinterpretation is likely to shed additional light on many present problems such as the nature of spacetime itself, the nature of dark matter, and new insight and progress on unifying physics.

In the new openness that should result, we hope that a great extension and reinterpretation of electrodynamics will also be undertaken, and that it will reveal the underlying powerful structured general relativity infolded inside conventional EM potentials, fields, and waves, as originally revealed by

Whittaker {636}. It may even be that one of the novel concepts {637} we have proposed in this book will yet see the light of day in practical systems.

If so, then hopefully much of what we are addressing in this book will be incorporated in the emerging new physics of the twenty-first century.

Chapter 11

Final Perspective: Permissible COP>1.0 Maxwellian Systems

"Are you certain of that?" [Augustine, in his famous confrontation with the Skeptics, who held that nothing can be known with certainty.]

"[It is usually]...assumed that the magnitude of potential energy is irrelevant, being arbitrary to the extent of an additive constant. " [Block and Crater]. {638}

"The potentials V and A, defined ... in terms of the fields E and B, are considered to be nonobservable in Maxwellian electromagnetism, since the energy density associated with them is zero... " [Lakes]. {639}

"[The total energy of the atom] depends on the bare mass and bare charge of the electron, the mass and charge that appear in the equations of the theory before we start worrying about photon emissions and reabsorptions. But free electrons as well as electrons in atoms are always emitting and reabsorbing photons that affect the electron's mass and electric charge, and so the bare mass and charge are not the same as the measured electron mass and charge that are listed in tables of elementary particles. In fact, in order to account for the observed values (which of course are finite) of the mass and charge of the electron, the bare mass and charge must themselves be infinite. The total energy of the atom is thus the sum of two terms, both infinite: the bare energy that is infinite because it depends on the infinite bare mass and charge, and the energy shift... that is infinite because it receives contributions from virtual photons of unlimited energy. " fWeinberg]. {640}

11.1 Introduction

The central message in this book is that all EM devices are already powered — though in an inappropriate fashion — by EM energy freely extracted from the seething vacuum. It is also intended to be a clarion call that we must rapidly develop the *appropriate* vacuum-energy powering of our EM circuits and electrical power systems, or else the survival of the nation and of the biosphere is at stake rather immediately.

For this summary chapter, we wish to place that specific area — EM energy from the vacuum — in sharp focus, with the major facets together in one place.

Consequently, for this closing chapter we have chosen and slightly adapted a previous technical rebuttal to the objections of "perpetual motion", "violation of the second law of thermodynamics," "impossible", etc. that we wrote in 2001 in the referee process, to the referees of one of the AIAS papers submitted to a major physics journal. The rebuttal was successful, and it was accepted and upheld by the referees. The paper was then published by the journal, after very vigorous refereeing.

That is as it should be.

Scientific journals provide a proper forum for the free expression and debate of scientific ideas and work. They do not and should not ensure that every paper conforms to commonly held concepts or to the most widely accepted model. As we stated, Chief Editor Kennedy of the journal *Science* has only recently had to restate and accent the principle that science is free from "truth by prevailing authority," and bound by "truth by replicable experiment". Insuring that every paper meets the accepted models and notions would be enforcement of religious belief, not experimental science. Ideas which are wrong will be refuted experimentally; experiment is still the basis of scientific method. In our rebuttal, the argument of the conventional skeptics is thoroughly refuted by experiments already well known in the literature, as we detail.

So we thought it appropriate to close the book with this strong statement of its primary pbjective: economical, clean, copious electrical energy from the vacuum.

11.2 Source Charges, Curved Spacetime, Efficiency, and COP

11.2.1 The Source Charge Problem

The received interpretation of the Maxwell-Heaviside theory treats the source charge and source dipole as perpetual motion machines, freely and continuously pouring out EM energy in 3-space in all directions with no energy input. Experimentally one verifies that there is no *3-space observable energy input* to the charge or dipole. In accepting that model, a scientist objecting to COP>1.0 EM systems has unwittingly hoisted himself upon his own perpetual motion machine petard, by already accepting every source charge and dipole as precisely that to which he is objecting. To explain the source charge, either one holds to the U(1) Maxwell-Heaviside theory and surrenders the conservation of energy law, or one accepts a change to U(1) electrodynamics to add the concomitant input of energy flow from the active vacuum (from the *time* domain, which in Minkowski geometry is all that is left outside 3-space) to that charge or dipole. That corresponds to a violation of the second law of thermodynamics, as — as we discussed in Chapter 10 — Evans and his colleagues at the Australian National University have already experimentally proven that the second law can be violated up to micron scale and for times of up to two seconds.[287] At the much smaller fundamental charged particle level and far shorter "switching" time of one spin completion, the probability is 100% and a certainty that the second law is violated. Hence the broken symmetry of the charged particle considered in the modem sense as a dipolarity (considering its associated **clustering** virtual charges of opposite sign).

Either way, the classical Maxwell-Heaviside theory must be changed. Every charge and every dipole — which themselves are Maxwell-Heaviside EM systems *a priori* — refutes the erroneous argument that no Maxwell-Heaviside system can output more energy than the operator himself inputs from an external power supply. To adamantly continue to challenge something already *experimentally demonstrated by* every charge and dipole in the universe is the epitomy of scientific blindness.

As have others, Sen {641} called this source charge paradox the "most **difficult"** problem in electrodynamics. However, the basis for its solution - absorption and transduction of virtual KM energy from the vacuum via the broken symmetry of a charge or dipole in its vacuum energy exchange — has now been known in particle physics for nearly a half-century {642a-642d}. It is not included in the Maxwell-Heaviside-Lorentz theory. Neither the proven vacuum interaction with Maxwellian system nor the broken symmetry in that interaction has been added to the model.

The present author has proposed an EM wave solution {643} to the source charge problem, by reinterpreting Whittaker's {644} decomposition of the scalar potential (as between the ends of a dipole) and treating the "isolated charge" and its clustered virtual charges {645} of opposite signs as a set of composite dipoles. Each of the dipoles then becomes a broken symmetry in the energetic exchange with the vacuum. Hence the charge pours out observable EM energy with no *observable* energy input because the input energy is in virtual photon form.

In the solution, the EM energy is input to the charge or dipole from the time-domain, so conservation of EM energy flow is upheld in 4-space, while time-like symmetry and space-like symmetry are broken individually. Powerful support for the proposed solution was given in a quantum field theory argument by Mandl and Shaw {646} nearly three decades ago. We challenge any scientist skeptical of COP>1.0 Maxwellian systems to produce a solution of the source charge problem in Maxwell-Heaviside-Lorentz theory alone. In conventional theory, every charge in the universe experimentally exhibits COP = oo. Any theory conflicting with a replicable experiment is falsified a priori.

11.2.2 EM Waves Imply Curved Spacetime

The oscillating energy of an electromagnetic wave is continuously changing its local spatial energy density. This is an *oscillating change of spacetime curvature* if one accepts general relativity (GR). Hence the EM wave in space is *always* moving in a locally curved spacetime; else, no EM wave exists because no change in spatial energy density exists. To reject that, is to reject GR in its entirety as well as all the experiments consistent with GR. If one accepts GR, then the classical U(1) EM *assumption* of the EM wave moving in a fiat spacetime is a non sequitur. It is unscientific to be suspicious of a higher group symmetry O(3) electrodynamics which *does* correct this known non sequitur in U(1) electrodynamics, while limiting oneself to U(1) electrodynamics *with* that known error. One might as well be suspicious of tensor algebra because it is more comprehensive than vector algebra.

There are many EM experiments such as the Aharonov-Bohm effect {647a, 647b} which violate Maxwell-Heaviside EM, as is well known

{648, 649a-649i}. The foundations of the Maxwell-Heaviside theory are well over a century old, and Maxwell's theory was substantially truncated in the 1880s (more on that below). The considerable physics developed since then has required dramatic extensions to the Maxwell-Heaviside theory such as developing non-Abelian gauge field EM theory, quantum electrodynamics, and modeling of the interaction of vacuum and matter in any material system.

11.2.3 On EM Systems with COP>1.0

Conservation of energy requires that all energy output by an inert system must be input to it {650}. It does not require that the *operator* input all or even any of the energy utilized. The active environment can permissibly input part or all of it, else there could be no such thing as a windmill — or a charge, or a dipole, or a solar-cell electrical power system.

The *coefficient of performance* (COP) of a system compares the system's useful energy or work output to the *operator's* energy input. The *efficiency* ε of a system compares its useful output to its *entire* energy input. No inert system can output more useful energy than its total input, hence $\varepsilon <= 1.0$, being $\varepsilon = 1.0$ for a theoretically lossless "perfect" system and $\varepsilon < 1.0$ for a real system with losses.

Whether the operator must input all the energy that is output by the inert system depends upon whether the system is (i) an open system in disequilibrium with its active environment, and (ii) *designed* to accept and utilize excess energy from that environment, such as from the active vacuum. If appreciable environmental energy input is received and utilized, then a system with losses may permissibly exhibit COP>1.0, even though its efficiency ε is $\varepsilon < 1.0$. The common home heat pump is an example. While the efficiency of a home heat pump may be $\varepsilon = 50\%$, its **COP** may be COP = 4.0.

As does the skeptical scientist, the electrical engineering community erroneously assumes that an inert electrical power system that outputs more energy than *one oneself* inputs to it — e.g., a generator outputting more EM energy flow than the magnitude of the mechanical energy input to the generator shaft — is against the laws of physics. To the contrary, it is permitted by the laws of physics and thermodynamics, *once the modern* ***Vacuum*** *interaction is included and disequilibrium exists in the interaction* ***between*** *vacuum and system.* Not only is it permissible, it is a universal fact once the arbitrarily discarded Heaviside energy flow component is re-ac**counted,** and the source charge problem is reluctantly placed back on the table. If we account all that EM energy pouring freely from the source charges in any EM circuit, then the actual output of EM energy from that circuit is enormous. It is many orders of magnitude greater than the magnitude of the input energy that the experimenter or operator provides.

We do not have to reprove the active vacuum; it has long been proven in particle physics. We also do not have to reprove the disequilibrium between electromagnetic systems and the active vacuum environment; that also has been proven since 1957 by the prediction and discovery of broken symmetry. One of the proven broken symmetries is that of opposite charges, such as are on the ends of a dipole. Hence every dipole or dipolarity in electromagnetics already freely absorbs virtual photon energy from the vacuum, transduces (coherently integrates) it into observable photon energy, and re-emits observable EM photons in all directions at the speed of light. If there are no COP>1.0 Maxwellian systems, then there can be no Maxwellian charges, dipoles, or dipolarity, and therefore no resulting fields and potentials from these sources — an absurdity falsified by every standard two-wire electrical circuit and ever charge and dipole in the universe. Any potential is a dipolarity, hence exhibits broken symmetry. That is why from any finite potential, as much EM energy as desired can be intercepted and collected, by the simple equation $W = \phi q$, where W *is* the collected energy in joules, ϕ is the scalar potential intensity in joules per unit point static coulomb, and q is the intercepting charge in coulombs.

Systems far from equilibrium in their energy exchange with their active environment {651a, 651b} can permissibly perform five important functions impossible to equilibrium systems. Such a disequilibrium system can:

(1) Self-order,

(2) Self-oscillate or self-rotate,

(3) Output more energy than the operator inputs (the excess energy is freely received from the active environment),

(4) Power itself and its load (all the energy is freely received from the active environment), and

(5) Exhibit negentropy.

Any charge or dipole already exhibits all five of these functions — forbidden by classical thermodynamics but permitted by the well known thermodynamics of systems far from equilibrium in their exchange with their active environment {651a, 651b}.

It *only takes one small white crow to prove that not all crows are black.* The known performance of the charge and the dipole totally refutes the notion that no EM system can output energy without operator-arranged input from other than the active vacuum. The entropy-consuming reactions from the fluctuation theorem, experimentally proven by Evans et al. in 2002, also is a sufficient proof to destroy any argument against permissible COP>1.0 EM systems. Since all the EM field energy and potential energy in a power system or circuit must come from the source charges and dipoles, it follows that all our EM energy systems already take *all* their E*M* energy from the active vacuum, *not* from the operator's input. A solar cell array also refutes the notion that the operator has to input the energy, but the solar cells' environmental energy input may not be ubiquitous or dependable. The active vacuum is both.

Most power system electrodynamicists avoid the particle physics solution {652} to the source-charge problem {641} involving disequilibrium exchange with the active vacuum. Instead, they adhere strictly to the Maxwell-Heaviside-Lorentz theory with its *assumed* equilibrium between system and vacuum. Therefore they avoid modeling the vacuum interaction and solving the problem of open dissipative EM systems which freely and dependably receive energy from their vacuum environment in unusable form, translate it into usable form, and furnish it for further capture and use. They therefore cannot resolve the source charge problem in their model, for the model has already artificially excluded the solution.

By definition, an EM system in equilibrium cannot output more useful energy than the operator inputs. However, considered as an *energy **transducer,*** every EM system is in disequilibrium with the vacuum because of the charges and dipoles in the system, and it continuously receives energy from the vacuum. Because of the broken symmetry of the charge and the dipole, some of this energy from the vacuum is discharged by every charge and every dipolarity in the circuit as observable energy *a priori.* This creates a truly enormous outpouring of rather disorganized (in this case, *uncollected)* EM energy extracted from the vacuum and poured out without being intercepted and diverged into the circuit, where it is organized for dissipation and use. The circuit "uses" only that amount of this enormous available outpouring EM energy that it first organizes (collects) and directs. That is its "organized excitation" energy, commonly known as *input energy* macroscopically — the Poynting energy flow component.

Upon completely discharging its excitation (input energy), the inert system outputs all its received (organized) energy, from whatever source. So every EM system already outputs far more energy than the experimenter inputs — when the system's *nondiverged, non-organized, and nondiverged* output energy flow is considered as well as its *diverged, organized, and collected* energy flow component. The fact that the *diverged Poynting* energy received in the external circuit enters the circuit from the surrounding space is shown by Krauss {653} (See Figure 1-1 in Chapter 1). We show it in Figure 2-4 of Chapter 2 as well. The remaining and nondiverged Heaviside energy flow component in space, which misses the circuit and is wasted, is also shown in Figure 2-4 of Chapter 2 and contrasted to the diverged (caught) Poynting component. To understand this experimentally demonstrable fact {653}, we need a bit of EM history.

11.3 Truncations of Maxwell's Theory and Discard of System Classes

11.3.1 Concept of EM Energy Flow Through 3-Space

The concept of the connection of EM energy and space did not occur until 1876 in a faulty work by Croll {654}. In the 1880s Heaviside {655} showed the huge nondiverged component filling all space around a circuit (see Figure 2-4 in Chapter 2). Poynting {656} never even considered the external nondiverged flow missing the circuit, but assumed only the *diverged* energy flow component that enters the circuit. Figure 2-4 in Chapter 2 shows both components of EM energy flow laterally, so they can be seen clearly.

It is stressed that the Heaviside nondiverged component of EM energy flow around every circuit is far greater (by many orders of magnitude) than the circuit's intercepted, diverged, and collected Poynting energy flow component. This follows from Heaviside's own original papers. So when the entire output energy flow is considered, every generator and every battery is already a $COP \gg 1.0$ *energy transducer,* outputting far more energy flow than the rate at which mechanical energy is input to the generator shaft or the rate at which chemical energy is dissipated in the battery.

In the 1880s the electron, atom, nucleus, special and general relativity, quantum mechanics, quantum electrodynamics, and active vacuum had not been discovered. There was no knowledge of any possible active vacuum interaction source for such an enormous energy flow pouring from the terminals of the generator or battery and surrounding the circuit, filling all space around it. Heaviside's extra energy flow component was thus shocking and inexplicable, smacking of "perpetual motion" and total violation of the conservation of energy law. Heaviside — a brilliant, self-taught hermit who never attended university - spoke cautiously and obliquely of its magnitude in terms of angles with respect to a reference direction {657}.

11.3.2 Elimination of the Non-Intercepted Energy Flow Component

A perplexed and concerned Lorentz {658} cut the Gordian knot and arbitrarily *discarded* the Heaviside nondiverged component, eliminating the quandary. Lorentz stated that the huge nondiverged component had "no physical significance" (his phrase) since it powered nothing. Of course it **has** no physical significance, and powers nothing, unless one intercepts it and uses it!

By analogy, the component of a great wind on the ocean that is *outside* the intercepting sails of a single sailing ship, does not contribute to further powering *of that ship.* Figure 2-5 in Chapter 2 shows that situation. However, that nondiverged wind component can power a whole flotilla of ***additional,*** *separate* sailing ships with separate sails, as we showed in Figure 2-6 in Chapter 2. Electrodynamicists {659} have continued to utilize Lorentz' clever integration trick (Figure 2-3 in Chapter 2) to dispose of the huge nondiverged Heaviside energy flow component surrounding **every** circuit and missing it, and therefore pouring from the terminals of every generator. Most electrical engineers today do not realize that the original purpose of Lorentz's integration of the energy flow vector around a closed surface assumed surrounding every volume element of interest was merely to get rid of that perplexing and bothersome Heaviside nondiverged energy flow component while retaining the Poynting diverged energy flow component.

11.3.3 Lorentz's Previous Discard of all COP>1.0 Maxwellian Systems

Previously (circa 1890s) Lorentz had symmetrically regauged Heaviside's **equations** that had already seriously curtailed Maxwell's seminal equations of 1865. This Lorentz symmetrical regauging retained only that class of EM systems that are in equilibrium in their seething energy exchange with their active vacuum environment. It discarded that entire permissible class of the Maxwell-Heaviside systems far from equilibrium with the active vacuum. In short, it *arbitrarily* discarded all COP>1.0 Maxwellian **systems,** and all those even *permitted* to exhibit COP>1.0.

The ubiquitous use of closed current loop circuits containing both the ex**ternal** circuit and the source dipole inside the generator forcibly applied (and continues to forcibly apply) the Lorentz symmetrical regauging. Thus all electrical engineers began using circuits which self-enforced equilibrium in the vacuum exchange, so that no *net usable* energy[306] that could translate electrons and do work in an external load —could be received from that source.

However, so long as the enormous Heaviside nondiverged energy flow component outside the circuit remained, then separate outrigger antenna circuits and systems could always be inserted in this external energy flow to divert some of it into separate additional circuits. This would have opened up the COP>1.0 *ensemble* of systems and separate outrigger receiver circuits (with separate loads) fed by a single central generator, such as one version of the motionless electromagnetic generator of Chapter 7.

Consequently Lorentz integrated the entire energy flow vector around a closed surface assumed around any volume element of interest {658}, as shown in Figure 2-3. This integration trick — still universally applied {659a-659c} in all the texts today — arbitrarily discarded that troublesome and massive Heaviside energy flow component that could not be explained, and for which no source was known.

---

[306] Lorentz symmetrical regauging does indeed require that the system receive excess energy from the environment (the active vacuum). It requires that the potential energy of the system is changed twice and continuously, which means the excess energy in each of those two changes must continuously enter from the active vacuum and do internal work on the system. Two new free force fields are created, each with its field energy, which is additional energy being continuously dissipated in the circuit to perform continuous work. However, the two free changes of energy in the system have been precisely selected so that these two force fields are equal and opposite. The two free changes of energy thus produce a *stress potential* without a net translation field. This means that the local spacetime has been curved, in a manner representing a rotation of the system frame away from the lab observer's frame. Further, the stress potential represents the continuous entry of extra stress energy from the vacuum, whose two force fields continually perform opposing local *internal work* within the system to stress it, while unable to translate electrons and power external loads. Contrary to the received view, the Lorentz-regauged Maxwell-Heaviside system is a dramatically altered and different system than the original system before the arbitrary application of the Lorentz condition. The system's inertial frame has been rotated, it has taken on extra energy from the vacuum that can only be used to do internal work upon the system, and that internal work on the system is continuously being done to increase the system's stress and hold it at the new level.

11.3.4 The Heart-Breaking Result of Lorentz's Actions

The combination of these two actions by Lorentz — symmetrical regauging of the model equations and discarding the excess Heaviside nondiverged energy flow component — unwittingly resulted in the design and building of only those electrical power systems which symmetrically and automatically regauge their discharge of their excitation energy. The symmetrical regauging is self-applied by the closed current loop circuit. That circuit uses half of its collected Poynting energy to destroy the source dipole in the generator, and the other half to power the loads and losses in the external circuit {660}. Conventional EM systems are unintentionally designed to destroy their own source dipoles — and their free extraction of organized EM energy from the vacuum — faster than they power their loads.

11.3.5 Agreement with Classical Equilibrium Thermodynamics

These curtailments of Maxwell's theory resulted in the exclusive building of equilibrium EM systems. These systems obviously obey the classical equilibrium thermodynamics with its infamous second law. Lorentz had unwittingly discarded all that entire class of Maxwellian systems in disequilibrium in their exchange with their active vacuum environment; in those days, there was nothing known about an "active vacuum environment" or broken symmetry in its virtual energy flux. Consequently, Lorentz had discarded all those Maxwellian systems that *do not* obey classical equilibrium thermodynamics, but do obey the newer thermodynamics of systems far from thermodynamic equilibrium with their active environment. He discarded all permissible COP>1.0 Maxwellian systems.

However, there was a total agreement between all the EM systems actually built using Lorentz's symmetrized equations and all measurements in their ubiquitous closed current loop circuits — all self-enforcing the Lorentz condition. This led to the iron dogma that all EM systems must inevitably comply with classical equilibrium thermodynamics — the circuits that were built *were* repeatedly and universally observed to do so!

11.3.6 Dogma that COP>1.0 EM Systems are Perpetual Motion Machines

Finally, this total agreement (now for more than a century) between the theoretical model used, the circuits actually built, the measurements performed, and classical equilibrium thermodynamics led to the present fallacious but iron dogma that EM circuits and power systems cannot output more energy than we ourselves input. Today the vast majority of scientists consider it absolutely proven that the laws of nature prohibit

COP>1.0 electrical power systems, and particularly prohibit self-powering systems that freely power themselves and their loads (i.e., EM systems analogous to "windmills in a free wind"). As we stated, every charge and dipole falsifies the "COP>1.0 EM systems are impossible" dogma, as does the fluctuation theorem and its experimental proof by Evans et al.

Sadly, no one seems to have realized that we do not presently have COP>1.0 electrical power systems only because engineers have specifically designed and built their EM systems to destroy their own source dipole faster than they power their load. Ironically, every circuit and every power system used and measured by our engineers already outputs far more EM energy than is input by the operator or by turning the shaft of the generator, if the Heaviside nondiverged energy flow component is re-accounted. The Bohren experiment clearly proves it experimentally {662}, along with thousands of other similar experiments in "negative resonance absorption of the medium" {157a-157d).

11.4 Examples of COP>1.0 Electromagnetic Systems

The motionless electromagnetic generator (MEG) {661} (Figure 11-1) does catch and use some of this available free energy, as explained below. So does the familiar Bohren experiment {662a, 662b} by resonating each intercepting charge so it sweeps out an increased geometrical reaction cross section in space, reaching past the usual Poynting interception cross section (for a static charge) and thereby absorbing additional energy from the otherwise non-intercepted Heaviside energy flow component.

Every charge and dipole in the universe also already acts as an electrical "windmill in a free wind," having a COP = oo.

The present author has also nominated the long-neglected but vast nondiverged Heaviside EM energy flow, accompanying every field and charge interaction, as the long-missing source of some 90% of the gravity holding the arms of the spiral galaxies intact {663}. Our *supersystem* examination of several COP>1.0 EM systems developed by various inventors has also uncovered a proposed solution for the source of the distant antigravity accelerating the expansion of the universe {664}.

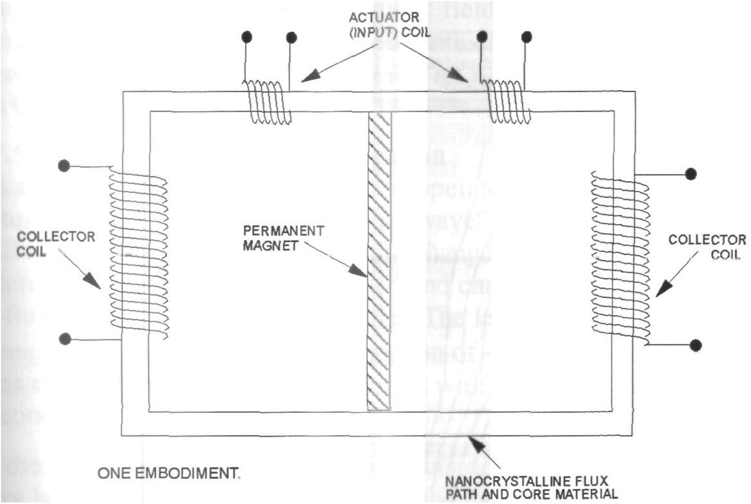

Figurei 11-1. Transformer section of the MEG.

Experimental proof that EM energy COP>1.0 is both possible and demonstrated is already in the literature — e.g., as in the Bohren experiment {662} which outputs 18 times as much energy as one inputs (see Figure 11-2). As another experimental example, the Aharonov-Bohm (AB) effect {647, 665, 666} freely adds additional nonlocalized field-free A-potential energy while all the magnetic B-field energy is sharply localized but still present and available. The AB effect is routinely implicated in physics, but apparently was never applied in an electrical power system prior to its use in the motionless electromagnetic generator

Bohren's experiment {662} was quickly and independently replicated by Paul and Fischer {667}. It can be replicated by any university laboratory at will, totally negating the conventional COP < 0 assumption. The AB effect is also easily replicated, using a toroidal coil, a long solenoid, or a permanent magnet and the special cores used in the MEG. When the toroid system's *total available output energy flow* is measured and accounted — including both the confined magnetic field energy flowing inside the toroidal windings and the field-free A-potential flowing outside — the toroidal coil system outputs more energy flow than the operator furnishes from his external power supply by conventional accounting {668}.

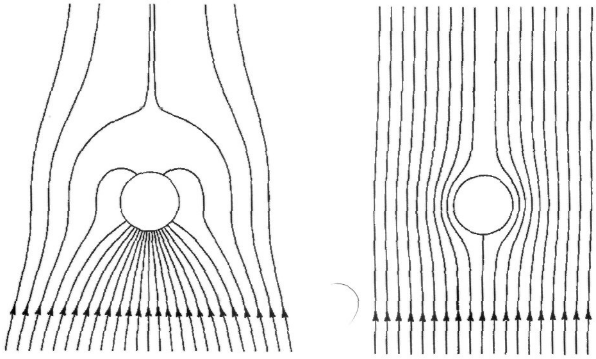

At or near resonance frequency, in the case shown the energy collection fraction (reaction cross section) increases dramatically.

Figure 11-2 The Bohren experiment outputs 18 times as much energy as one inputs

In conventional terms, for a theoretically perfect toroid the operator only "pays for" the magnetic field energy "output component" confined inside the toroidal windings. The presence of an extra energy output component outside the toroid produces COP>1.0 performance as *an energy transducer.* It *does not* produce system COP>1.0 performance in a closed-loop circuit containing the primary power supply's source dipole.

The AIAS has previously nominated more than a dozen mechanisms as promising COP>1.0 systems for investigation, research, and development {669}.

11.5 Motionless Electromagnetic Generator (MEG)

11.5.1 Overall Description

In the MEG {661} (Figure 11-1), one uses a nanocrystalline material core path with the unique characteristic of drawing in almost all the magnetic field energy of an emplaced permanent magnet. The B-field flux from the magnet is withdrawn from its normal position in surrounding space and sharply confined to that material core path. Around the core there freely appears an *additional* curl-free magnetic vector potential A energy flow, since the surrounding spacetime is still curved due to the interaction exchange of vacuum and magnet dipole, and therefore still contains extra energy. The operator inputs no energy whatsoever after the magnet is once emplaced, yet has appreciable magnetic field energy and field-free magnetic vector potential energy available to utilize. As an *energy transducer*, the MEG has a COP»1.0. As a power system powering loads, its COP>1.0.

11.5.2 An Example of MEG Operation

As a replicable example, suppose one operates the MEG core at less than saturation, and places a small "square wave" pulsed signal upon a primary coil wound around the core. The core changes its permeability as the internal magnetic B-flux is perturbed and changes, so all the perturbed B-flux is retained in the core material. The large E-fields produced by the changing B-flux in the core as a function of $-\partial \boldsymbol{B}/\partial t$ are *not* localized, but pass readily out of the core and interact with the coils, particularly with the secondary.

In the A-potential region in the space outside the core, one also produces very large E-fields by $\mathbf{E} = -\partial \mathbf{A}/\partial t$. At the frequencies being used (40 to 80 kHz nominally), all these E-fields produced by the core flux perturbations and by the external A-potential perturbations are essentially in phase. The large E-fields thus coherently add and interact directly with the secondary coil acting as a receiver.

In the secondary, the very large and sharply changing E-fields interacting with the coils also produces large B-field flux changes in the core. However, again the core changes its permeability and holds-in all the extra B-flux. The changes in that secondary B-flux also make additional non-localized E-fields, and so on.

The surprising and shocking result is that the transformer secondary undergoes a coherent, purely E-field reaction with the electrons in its coils. This results in the output voltage and current being in phase (within 2°, since there are inevitably some small remaining inductance effects due to a tiny bit of external B-field leakage from the junctures of the ends of the permanent magnet with the core material).

The magnitude of the powerful E-fields produced by these processes — and interacting with the secondary coils — depends on the rise and decay times of the input signal's pulse edges. So the *average* power input to the signal coil may be small, but — using sharp rise and decay times of the input pulses — very large E-fields will be produced in the adjacent surrounding space outside the transformer section and interacting with the secondary coils, as well as in the sharply changing B-flux localized in the core.

One then uses transmission-reception theory and near-field antenna theory for separate receiving circuits containing loads, added in the large E-field region in space outside the core. Those E-fields arise from both the changing A-potential outside the core and the changing localized flux inside the core. We stress again that the E-fields produced in the core are not localized, but pass outward into the external space to interact with whatever interceptor/receivers are placed out there.

A "outrigger array" version of the MEG uses separate external receiving circuits, containing loads, to intercept, collect, and use more of the EM energy available in the surrounding space outside the core.

Within physical space limits, the energy intercepted and utilized by these *separate, independent receiving circuits containing loads* can be increased by adding more receiving circuits with loads or adjusting the rise and decay time of the input pulses to provide greater collected voltage. These *separate* loads will be powered by the incident E-fields' interception in the external receiving circuits. This is not at all a *transformer* action, but a *near-field transmission and separated multiple antenna circuits' reception* action.

There is no back-current coupling between the external circuits and loads with the primary circuit of the transformer section. In the external A-potential region alone, the secondary coil in the transformer section will interact with the large incident E-fields from that adjacent space and from the changing B-flux inside the core, producing at least as much energy as one inputs to the primary. Actually, that process alone will produce more output energy in the secondary than the operator inputs in the primary.

If one operates the core below saturation, the magnetic flux in the core will also be switched, giving additional intercepted E-field energy in the secondary coil attached to its load. The total work out in all loads is permissibly greater than the total energy dissipated in the input by the operator's energy input {668}. The MEG is an open system far from equilibrium with the active vacuum {670}.

11.5.3 The MEG Is Replicable

Any university laboratory can replicate this effect, and demonstrate more work in resistive loads than is input to the MEG primary from the external power supply. For replication, we suggest initially designing to COP = 1.5, since beginning at COP = 2.0 there *can* be other effects not in the U(1) EM literature, but which are dealt with in the actual MEG. Some of the methodology {671} used for COP >2.0 is still proprietary information not yet releasable due to patent law requirements, while some of it was included in Chapter 7.

Without elaboration we advance the *supersystem* concept, and strongly insist that, to properly analyze the performance of any similar EM circuit or system, one must analyze the dynamics of the entire supersystem consisting of (i) the physical system and its dynamics, (ii) the curved local spacetime and its dynamics, and (iii) the local vacuum and its dynamics {672}. All three components of the supersystem are mutually interacting.

For a COP>1.0 system, the mutual interactions are utilized from disequilibrium processes evoked between all three supersystem components. U(l) electrodynamic analysis system in an inert vacuum and a flat spacetime is insufficient. Previous AIAS analysis {661} has addressed the supersystem by use of higher O(3) symmetry electrodynamics to include vacuum dynamics (energy currents, both space-like and time-like) and curved spacetime dynamics.

11.6 Physical Considerations of Gauge Freedom

Additional physical considerations must be accounted when *gauge freedom* is invoked or utilized with respect to physical systems. Applied to electrodynamics systems, gauge freedom assumes the ability to freely change any potential of the system being studied. From the physics of the situation, a system potential change concomitantiy assumes changing the *potential energy ofthe system itselffor* which gauge freedom changes are being theoretically applied to system-descriptive equations.

If one *can freely increase* the potential energy of an EM system, the extra energy has to come from a physical asymmetry of the system with its active environment — the active vacuum. Else, one must surrender the conservation of energy law. At worst, gauge freedom would then imply the creation and destruction of energy at will. At best, it would imply a return to the view that potentials are mere mathematical conveniences with no physical reality — a notion falsified for decades by the AB effect {647}. ***freely** decreasing* the potential energy of the system by simply depotentializing it, has a similar vacuum asymmetry requirement for de-excitation of the system, where energy must flow from the system to the vacuum.

11.7 Special Modeling of the Vacuum

Having EM energy density, the vacuum *is and should be modeled as* a scalar potential of great magnititude. For deeper (supersystem) analysis, one will also have to examine the internal EM longitudinal phase conjugate wavepair structure and dynamics of the scalar potential vacuum {672}, along the lines shown by Whittaker {644} in 1903 and re-interpreted by the present author {643} in 2000. From quantum field theory arguments, Mandl and Shaw {646} give strong support to the proposed reinterpretation. They argue that neither the scalar (time-polarized) photon , nor the longitudinal photon is independently observable, but the combination of the two is observable as the instantaneous scalar potential.

Transforming this argument to EM waves yields a solution consistent with the reinterpretation {643}. The result is that all 3-space EM energy comes from the time-domain, in a manner not previously utilized in EM circuit analysis. It is experimentally established by proving (measuring) that the source charge and dipole do not *receive ob*servable EM energy in 3-space, but continuously *output* observable EM energy in 3-space. Simply making and destroying dipoles in an electrical system turns out to be a method for direct engineering of negentropy and entropy respectively — or, from another viewpoint, it directly engineers regauging.

The greatest unrecognized problem in conventional electrical power system engineering is its total failure to include the making and destroying of dipoles as direct engineering of negentropy and entropy respectively.

Every EM potential in an EM system is a change to the energy density of the ambient vacuum potential — and thus is a spacetime curvature — or it is a change to an intermediate potential that *is* such a vacuum potential change. Vacuum exchange with every EM system and asymmetry in that exchange, is ubiquitous and cannot be prevented. Whenever one makes or changes an EM potential, field or wave, and whenever one regauges, one invokes and utilizes asymmetry in the interaction of the active vacuum and the system, either wittingly or unwittingly. Else one must discard the conservation of energy law altogether. There is no symmetry of a mass system alone {673}, without including the vacuum's asymmetrical interaction with it. The reason is simple: all observables involve asymmetry {674} *a priori,* and mass is an observable.

11.8 Every Charge and Dipole in an EM System Receives and Transduces Vacuum Energy

In particle physics, every dipole — and every charge considered with its associated clustering virtual charges of opposite sign — is recognized as an asymmetry in the vacuum flux {675}. The very definition of broken symmetry means that a charge or dipole receives and absorbs virtual photon energy from the vacuum, transduces it into observable form, and re-radiates it as real, observable EM energy. This results in an observable outpouring of 3-energy from the charge or dipole, with no observable 3-energy input.

To save the conservation of energy law for physical EM systems, one must therefore recognize and model the *physical vacuum asymmetry mechanism* **that** allows this demonstrated performance of charge and dipole. In electrical power systems, that required asymmetry mechanism is the known broken symmetry of the vacuum's interaction with the source dipole created in the generator.

11.9 Energy from the Vacuum Powers Every Electrical Power Line and EM Circuit

11.9.1 What the Generator Powers and What Powers the Circuit

Shaft power input to a conventional generator allows the generator to pr**oduce** its source dipole, nothing else. The generator does not transduce any of its input shaft mechanical energy to add EM energy to the external circuit connected to its terminals. Instead, it first transduces its mechanical shaft energy input into internal magnetic field energy.

Then it totally dissipates that magnetic energy to separate its own internal c**harges,** forcing the positive charges in one direction and the negative charges in the other, making the source dipole connected to its terminals. That is all that the mechanical shaft power input to a generator does — and **that** is all that burning hydrocarbons, consuming nuclear fuel rods, **damming** rivers for power stations, putting up windmill-powered g**enerators,** using solar energy input to solar cells, or consuming chemical energy in a battery accomplishes.

O*nce made,* the source dipole — via its broken symmetry in its vacuum exchange - freely receives and absorbs enormous virtual photon energy **from** the active vacuum, transduces it into observable (3-space) form, and pours out real, observable EM energy in 3-space from the generator terminals to which it is connected. As shown in Figure 1-1 of Chapter 1, this outpouring of EM energy from the generator terminals fills all space around the external circuit {653} with a vast energy flow almost parallel to the conductors. Figure 2-4 is a lateral view so that the two flow components can be clearly seen and compared.

11.9.2 Gedanken Experiment

We advance a gedanken experiment to illustrate. Consider a perfect generator with no losses, and with two perfect conductors and a purely resistive load R_1 connected across its terminals. Call the mechanical energy input to the shaft of the generator W_{in}. Call the Poynting diverged energy flow component W_r that is intercepted by the attached circuit and powers R_1. Call the Heaviside nondiverged energy flow component W_h that misses the circuit and is wasted. As an *energy transducer,* the total energy output W_{out} from the terminals of the generator is

$$W_{out} = W_r + W_h \qquad [1]$$

However, $W_{in} = W_r$, for this theoretically perfect system. So

$$W_{out} = W_{in} + W_h \Rightarrow W_{out} > W_{in} \qquad [2]$$

Further, when accounted, $W_h >> W_{in}$, so

$$W_{out} >> W_{in} \qquad [3]$$

We conclude that all EM systems — *as energy transducers freely extracting and transducing EM energy from the active vacuum*— already exhibit COP » 1.0. However, for *useful work* in a closed current loop circuit, the COP = 1.0, unless additional antenna-receiver circuits or another scheme such as Bohren charge resonance is used to increase the energy collected and dissipated in the load while *not* increasing the energy used to dissipate the source dipole.

11.10 Significant Developments in Negative Resistance

True negative resistance (e.g., current moving against the voltage) also exists and is experimentally replicable. In true negative resistance, the current in a circuit moves against the voltage instead of with it, suggesting time-reversal effects in the mechanism by which EM force and charge interact.

11.10.1 Chung's Negative Resistor and Naudin's Replication

Chung {676} has reported replicable negative resistance in the interfaces between layers of carbon fibers in a composite material. By varying the pressure used when the fibers are made, the resistance can be controlled to later be negative or positive as desired. A simpler version of Chung's negative resistor using off-the-shelf materials has been successfully replicated and tested by Naudin {677}. (See Figure 11-3). Naudin's website {677} provides measurements and full details necessary for repetition. He also includes sources of the carbon fiber materials needed for independent replication.

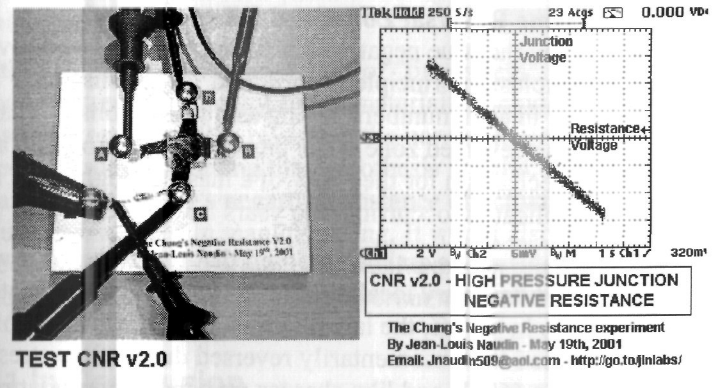
Tek Hold 250 S/s
23 Acqs
0.000 VDC
Junction Voltage
Resistance Voltage
Ch1 2 V Bw Ch2 5mV Bw M 1 s Ch1 ∕ 320m
A
B
C
D
The Chung's Negative Resistance V2.0
Jean-Louis Naudin - May 19th, 2001
CNR v2.0 - HIGH PRESSURE JUNCTION NEGATIVE RESISTANCE
The Chung's Negative Resistance experiment
By Jean-Louis Naudin - May 19th, 2001
Email: Jnaudin509@aol.com - http://go.to/jnlabs/
TEST CNR v2.0

Figure 11-3 Naudin's replication of Chung's negative resistor.

11.10.2 The Point Contact Transistor

There is also evidence that the original point contact transistor — which used spring pressure on the point contact — often exhibited true negative resistance. Certainly this transistor was never adequately understood {678}, and it is not understood today, since its supersystem has never been analyzed.

11.10.3 Kron's Negative Resistor

Gabriel Kron, early pioneer in electrical systems theory, reported developing a negative resistor in the 1930s, on the U.S. Navy's Network Anayzer project {679} at Stanford University. He was not permitted to reveal the details of his negative resistor, but did provide strong hints in certain statements in his papers in the literature {680a, 680b}. His method for a self-organizing network depended on his "open path" {681} which he also was not permitted to reveal.

The present author worked for some years with Floyd Sweet, a Kron protege", who did produce a successful type of powerful negative resistance system {682} known as the vacuum triode amplifier, apparently based on Kron's original negative resistor. See Figures 6-5 to 6-13 in Chapter 6. As shown in Figure 6-6, Sweet induced self-oscillation in barium ferrite magnets {683a-683c} (which are made by powder metallurgy methods using pressure applied during their manufacture) and extracted energy by induction from the resulting self-sustained magnetic field oscillations. His device was independently tested by Walter Rosenthal {684}, a senior test engineer of long experience with a major aerospace company. See Figure 6-9 and Figure 6-13.

11.10.4 Proposed Time Reversal Zones and Effects

One may take the view that true negative resistance involves time-reversal, which by the Pauli exclusion principle is permitted for bosons {685} and possibly for fermions in even numbers, acting as quasi-bosons (perhaps only possible in a time-reversed zone). The present author proposed a photon interaction mechanism for the "flow of a mass through time" in 1971, with initial publication occurring two years later {686}.

Extending this mechanism, the author proposed that *time reversal zones* (TRZs) can occur in momentary local regions in semiconducting materials or electrolytes {687}. In a TRZ, the law of attraction and repulsion of charges is hypothesized to be momentarily reversed due to time-stress or time-domain pumping {688}, and like charges may momentarily attract. In Chapter 10, we used these time-reversal zones to explain the anomalous results of the 600 or so successful cold fusion experiments worldwide. We also gave the exact new nuclear reactions for the production of anomalous deuterium, tritium, and alpha particles in many of these experiments.

When under strong stress, anomalous attraction of clusters of like charges is experimentally known, e.g. in Shoulders' {689} experiments and patented processes {690}. When the output leads of Sweet's vacuum triode amplifier device were suddenly shorted, extreme cooling occurred rather than heating. Ice condensed instantly upon the shorting contact and leads from the sudden extreme cooling of the metal in the moisture in the air, dramatically showing a converging EM energy process rather than a diverging one.

11.10.5 Effects in Optically Active Semiconducting Materials

In addition, four-wave-mixing, phase conjugation, and time-reversal effects are well known in semiconducting materials {691}, almost all of which are optically active. Optical activity and microlasing is known in semiconductor powder {692}. Further, the vacuum itself can exhibit semiconducting properties {693}.

11.10.6 Special Functions Produced by Stress

Apparently pressure (stress) in many of these devices and experiments can play an important role in the negative resistance process. Based on harmonic sets of paired couplings of a longitudinal time-like wave and a longitudinal wave in any scalar potential {643}, the present author hypothesizes that such stress — treated as a stress scalar potential — accomplishes time-domain pumping in the optical sense, due to the presence of the coupled time-polarized longitudinal waves in the time domain.

This time pumping is a special form of "time stress". Thus in optically pumped or stressed optically active materials (including electrolytes), the time-stress creates a TRZ, reversing the normal force laws upon and between charges. Assuming this hypothesis is valid, this could account for the variation of Chung's negative resistance effect with the variation of the pressure (stress) used in manufacturing. It would also apply to the laser-**created** stress in semiconducting powder experiments, etc. The hypothesis should be further investigated for experimental validation or refutation.

11.11 In Conclusion

As Penrose {694} states,

> *"Electromagnetism is a subject that is in no way closed to stimulating new developments."*

Indeed, electrodynamics is still very much an advancing subject, and in classical electrodynamics there are foundations errors and problems, a few of which we have addressed. As Bunge {695} pointed out more than three **deca**des ago,

> *"...it is not usually acknowledged that electrodynamics, both classical and quantal, are in a sad state. "*

From a foundations viewpoint, Einstein {696} beautifully summed it up — and advanced a requirement for continually examining the foundations of electrodynamics and every other field of physics — in these words:

> *"..the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools **of thought,** as something obviously, immutably*

> *given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."*

Though they have long been neglected, legitimate COP>1.0 electrical systems are good science, or they can be. They were present in the original Heaviside truncation of Maxwell's theory prior to Lorentz's symmetrical regauging. That regauging arbitrarily selected only those Maxwellian systems in equilibrium in their exchange with the active vacuum. The energy flow theory was further curtailed by Lorentz's integration trick, which discarded the huge vacuum-furnished non-intercepted Heaviside energy flow component accompanying every EM system. Every EM system already outputs for more EM energy flow than the energy that the operator inputs; thus every EM system is already a COP»1.0 system, with respect to *energy transduction.* The only way this giant negentropy of every EM system can be hidden is to ubiquitously continue that integration trick to discard the giant Heaviside energy flow component.

The present Lorentz-regauged Maxwell-Heaviside equations are not consistent with *any* bipolar EM system or circuit. The ends of any dipolarity in the universe — including any electron on Jupiter and any positron in the star Sirius — constitutes a broken symmetry in the seething vacuum flux. Actually, that is what Gabriel Kron discovered and called his "open path". Hence, that "dipole" continuously emits observable EM energy extracted from the vacuum.

Engineering use of the ubiquitous closed current loop to contain both the primary source dipole and the load — which destroys the source dipole and extraction of vacuum energy faster than the load is powered — is responsible for the successful predictions of the equilibrium EM model (which of course completely fails to explain the source charge or source dipole). That circuit is unwittingly designed to self-enforce the Lorentz symmetry condition and therefore self-enforce those predictions.

The question as to whether COP>1.0 electrical power systems are possible *and practical* must be resolved by direct experiment, and not by established models and cherished concepts that already arbitrarily exclude such systems! Such theoretical refutation by models that in fact contradict the experimentally observed COP = oo action of every charge in the universe, cannot be trusted at all with respect to COP> 1.0 judgment. Well-known supporting and confirming experiments already in the literature — such as the Bohren experiment yielding COP =18 — have been cited, and these are replicable by any university laboratory. The Bohren-type experiment is well known and routinely performed in every nonlinear o**ptics** laboratory, under the guise of "negative resonance absorption of the medium".

It is indeed a critical time to re-examine the foundations of our electrical **power** science and engineering, along the lines so eloquently stated by Albert Einstein. With the impending oil crisis, we face a world economic crisis of epic proportions unless we rapidly develop and widely deploy sell-powering disequilibrium electrical power systems.

Also it is a critical time when the leaders of the scientific community — specifically the National Academy of Sciences, the National Science **Foundation,** the great national laboratories, and the universities — must at last face and correct the excruciating *faux pas* that has now been perpetuated in electrodynamic theory for more than a century. In short, the le**aders** of our scientific community must now show whether they are scientists practicing and upholding experimental basis, or are dogmatists **defending** an ancient faith. The fact that the source charge problem c**ontinues** to be propagated in electrodynamics 45 years after the award of the Nobel Prize to Lee and Yang, is no longer excusable.

A thorough review and correction of the foundations of EM theory is re**quired** for the survival of this nation and much of civilization. It is also re**quired** for the cleanup of the increasingly polluted biosphere, survival of the threatened species of the Earth, reduction of present global warming **trends,** and the survival of this fragile and beautiful planet itself.

Annotated Glossary of Selected Terms

action at a distance: An action (effect) occurring in a local system, for which no separate local 3-space cause exists, but for which a causal correlation exists between a distant action and the local effect.

advanced electromagnetic field: An electromagnetic field that is a solution of the classical Maxwell equations, which is positioned on the future light-cone of spacetime. It may be possible to involve such a field in the tempic force that may exist between the tempic potential at the output of an overunity system and the tempic potential that exists at the input of the system, where the interaction of the tempic force with the local nonlinear vacuum drives Dirac sea hole current from output section to input section.

advanced electromagnetic potential: An electromagnetic potential that is a solution of the classical Maxwell equations, and is positioned on the future light-cone of spacetime. Presently this potential has not yet been given a physical interpretation. It may be possible to involve its effects in a specific system by our notion and use of the instantaneous tempic potential difference (the tempic force) that may exist between the tempic potential at the output of a system and the tempic potential at the input of that system if the two differ.

advanced electromagnetic wave: An electromagnetic wave that is a solution of the classical Maxwell equations, and which is positioned on the future light-cone of spacetime.

Aharonov, Yakir: Renowned Israeli quantum physicist and former student of David Bohm. Co-discoverer of the Aharonov-Bohm effect. Together with Bohm, Aharonov authored a fundamental paper in 1959 that pointed out the *primary importance of the potentials* rather than the force fields. The force fields are only made in and of the charged particle system itself, and do not exist as such in the vacuum. The force fields are thus not primary electromagnetic causes at all, but are *effects* due to the interference of potentials in a particular charged particle system.

Aharonov-Bohm effect: Quantum mechanical phenomenon theoretically pointed out in 1959 by Yakir Aharonov and David Bohm whereby interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields. A system undergoing a cyclic evolution from an initial state back to it will acquire a phase difference, which can be detected by interfering the initial and final states. Theory was extended later to Berry phase, then again to geometric phase.

In classical EM terms, a more important *energy-amplifying* principle - advanced by the present author — emerges from the AB effect. My colleagues and I have used this principle in our motionless electromagnetic generator (MEG) to provide COP>1.0. Consider a source of magnetic field B, such as a permanent magnet. All the magnetic field energy B from the permanent magnet may be drawn into and confined in a local material path, as in a special nanocrystalline transformer core. Then in the space surrounding the localized magnetic field energy, the active vacuum will freely replenish — via the broken symmetry of the permanent magnet dipole in its exchange with the vacuum — the magnetic energy in that space that would otherwise have been filled with magnetic field B (curled magnetic vector potential A). However, nature replenishes the energy in space outside the B-localization region with the *curl-free* magnetic vector potential A. In this fashion a system designer may cause the active vacuum to provide more available magnetic energy from the permanent magnet that normal, with each of the two energy forms being in separate spatial regions and paths, and occurring in different forms of magnetic vector potential energy. By inputting only a little energy in the input coils of the transformer section, and controlling the rise time and decay time of the pulses, one can produce very large E-fields from the A-potential via $E = -dA/dt$. From the simultaneously perturbed confined B-field flux in the core, one also produces E-field simultaneously and — at the frequencies used — in phase with the E-fields produced outside the core. The net result is interaction of large E-fields with the output coil, producing an output power *proportional to the rate of change* of the weak input signal energy rather than its magnitude. In this fashion, additional E-field energy from the vacuum enters the output coils to power the secondary circuit and its loads. Also, the output of the entire transformer section has the voltage in phase with the current, in contradiction to the actions of all previous transformers.

AIAS: Alpha Foundation's Institute for Advanced Study, directed by Dr. Myron W. Evans, a noted scientist with more than 600 papers in the literature. The AIAS is an unusual "think tank" comprised of Fellows and Fellow Emeriti in various places in the world, who cooperate in advanced electromagnetics studies via E-mail. The Director, Dr. Myron W. Evans, is a theorist of note, as are several other Fellows. The AIAS has been forging ahead with the production of O(3) symmetry electrodynamics, particularly as an important subset of Mendel Sachs' unified field theory. Dozens of AIAS group-authored papers have been published in various leading journals such as *Foundations of Physics, Physica Scripta,* etc. More than 100 AIAS papers are carried on a restricted Department of Energy website for reference by DoE and AIAS scientists. The work is particularly significant in preliminary theorizing how EM energy can possibly be extracted from the active vacuum.

algebras: The various types of mathematics in which the theoretical models for physics and other sciences are embedded. Some important algebras utilized in physics are vector algebra, tensor algebra, quaternion algebra, Grassmann algebra, Pauli algebra, Clifford algebra, etc. There are many others. The complexity of the symmetry and of the topology varies significantly between these algebras. Thus, an electrodynamics model in a lower topology algebra such as vectors or tensors will not reveal or permit nearly so great a set of electrodynamic functions and operations as will electrodynamics embedded in a higher topology algebra such as quaternion algebra or Clifford algebra. See topology.

ambient vacuum: The average or typical "standard" active vacuum (i.e., active spacetime), removed spatially from any large collections of observable charge and observable mass.

ambient vacuum potential: The *average* or typical standard vacuum considered as a scalar potential because it possesses energy density as well as structure and dynamics on many levels. As a scalar potential, the vacuum potential (considered without zero-point fluctuations) decomposes into a harmonic set of "bidirectional" longitudinal EM phase conjugate wavepairs, by Whittaker's 1903 decomposition as reinterpreted by the present author in 2000. Dynamic structuring of the vacuum thus consists of dynamics impressed upon these wavepairs. Since the vacuum is observably mass-free, Whittaker's phase-conjugate half of the decomposition must be considered prior to its interaction with mass, hence in the imaginary plane **and** arising from the time domain. This consideration led to the discovery by the present author of the more fundamental 4-symmetry in EM energy flow between the time domain (imaginary plane) and real 3-space, when broken 3-symmetry is present (as with any dipolarity or with any **potential).**

Powerful support for this solution is also given by F. Mandl and G. Shaw, Quant*um Field Theory,* Wiley, 1984, Chapter 5. Mandl and Shaw argue tthat the longitudinal and scalar (time) polarizations of the photon are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, directly fits our re-interpretation of Whittaker's 1903 decomposition of the scalar potential.

angular momentum: The momentum or "leverage" of the linear momentum of a moving body, with respect to an axis or reference point. Angular momentum has the same units as action; that is, it is energy multiplied by time, or momentum multiplied by length. In terms of particles, it is also known as *spin.*

anti-circuits: Somewhat normal-appearing electromagnetic circuits which primarily process and use antiparticles and anti-electrons rather than particles and electrons, or else process a combination of the two simultaneously. Operations of the anti-circuit must be analyzed and understood in terms of its *supersystem* interactions. The circuits process and utilize both negative energy and positive energy EM fields and potentials, hence are closely connected to unified field engineering and antigravity. One aspect enabling such circuits is the use of the supersystem's curvatures of spacetime in both negative curvature and positive curvature fashion so that an appropriate change of spacetime curvature precludes the radiation normally accompanying pair annihilation, etc. The key to the circuits is T-reversal and C-reversal operation, along with the matching use of spacetime curvatures, and also the use of the Dirac "positron" (negative energy electron) — and its fields and potentials — prior to its observation. The unobserved Dirac negative energy electron in the vacuum is an entirely different entity than is the observed Dirac positron after its interaction with charged matter, after which it has become a "lattice hole" or "positive charge on a positive mass ion". The response of normal circuit components to vacuum Dirac sea hole operations is dramatically different than for Dirac positrons, and for certain applications special versions of circuit components are required. At this writing, together with John Bedini the present author has submitted what may be the first patent application on the fundamental process for anti-circuits and their operations with antimatter and negative energy currents, fields, and potentials as well as matter and positive energy currents, fields, and potentials. Another patent application by this author **is** in preparation, covering the anti-gravity aspects. The details of these mechanisms are still proprietary at this writing in order to protect our intellectual property rights, but limited information is in this book.

anti-electron: A negative-energy electron, observed as *a positron.* Under time reversal, electric charge reverses in sign. Dirac did not predict the positron per *se,* but a negative energy electron that we would *observe* in reversed fashion. Present theorists assume the interaction of the 4-space negative energy electron with matter — i.e., observation — to have occurred. This transforms the negative energy, negative mass electron to a positive energy, positive mass positron traveling in the opposite direction. In so doing, antigravity from the negative energy 4-electron source charge, due to its negative energy EM fields and negative energy potentials, is arbitrarily transformed by the spatial reversal into positive gravity fields rather than the antigravity fields they were. The secret to practical antigravity is to utilize (in anti-circuits) the negative energy 4-electrons prior to observation, *without* transformation of (i) negative energy fields to positive energy fields, (ii) negative mass to positive mass, and (ii) reversal of the field direction. In short, simply consider the unobserved negative energy 4-electron as what it is, and utilize it in its unobserved vacuum state. The conventional transformation process also unwittingly discards the concept of anti-circuits, anti-fields, anti-potentials, and anti-currents, which discards half the rich electromagnetic circuitry that can be built and utilized. In the present book, we only touch lightly on this latter subject, which hopefully will be the subject of a future book after certain patents are filed.

antigravity: Negative gravity or reversed gravity. Use of negative curvature of spacetime to produce a force on a gravitational source (physical system) that is opposite to the normal force of gravity of the earth etc. on that system. Antigravity is achieved by the use of negative-energy, negative-mass source charges in the mass component of the supersystem prior to observation, while the source charges are still 4-spatial. This usage produces negative energy fields and negative energy potentials to negatively curve the local spacetime (ST), so that these negative *local* ST curvatures interact back upon the positive mass source system associated with the vacuum negative energy sources in its supersystem. Again, to understand the operation of an antigravity circuit or system, one must analyze it in terms of its *supersystem* interactions.

antimatter: Matter consisting of atoms that are composed of anti-electrons (positrons), antiprotons, and antineutrons and such. Also loosely refers to the antiparticle corresponding to a particle; an antiparticle may be regarded as a particle traveling backward in time, or "phase conjugated", or "time-reversed. Since time is not an observable, the observer sees the antiparticle {*after its observation by interaction with mass)* in forward time as being reversed in direction and charge.

antiparticle: Conventionally, a counterpart to an ordinary fundamental particle, *observably* having identical mass, lifetime and spin, but with charge and magnetic moment reversed in algebraic sign. We again point out, however, that the antiparticle differs when in its unobserved 4-spatial state. Only by analyzing the particle's supersystem can the true unobserved antiparticle be understood, as well as its fields and potentials in that condition.

antiphoton: A time-reversed photon. Presently the photon is considered its own antiparticle. *Observably* it is, but in its pre-observation condition it is not. An alternate but as yet unaccepted view is that, since the unobserved photon is comprised of (dE)(dt) — i.e., a piece of energy welded to a piece of time with no seam in the middle, the antiphoton may be assumed to be comprised of (-dE)(-dt) with respect to the external observer.

anti-Stokes radiation: The radiation coming from an anti-Stokes emission process from an intensely scattering medium, wherein more energy flux is emitted than is input by the experimenter-operator. The excess emitted energy is furnished by other processes in the medium. For true overunity operation, the excess energy furnished by the energetic medium must be freely replenished to the medium by the active vacuum's interaction with it.

antiwave: The time-reversal (phase conjugation) of a reference wave.

A-potential: The magnetic vector potential A, conventionally thought to be "defined" by the equation $\mathbf{B} = \nabla\times\mathbf{A}$. However, no equation is a definition (an identity is required). If we replace the "=" sign by "==", we have $\mathbf{B} \equiv\nabla\times\mathbf{A}$, which is actually seen to define the field B as just a curled A-potential, or as the curled component of the A-potential, *after* interaction with magnetic charge. The A-potential or an additional component of it can and does also exist in uncurled fashion, and will emerge in the uncurled state in the space outside any local path (e.g., the path in the interior of a toroidal coil) that retains the B-field therein. That process of separating the curl component of the A-potential from its uncurled component is known as the Aharonov-Bohm effect (which also affects the wave function). The motionless electromagnetic generator (MEG) uses a special nanocrystalline core material in a transformer configuration, which draws in and retains the B-field from a permanent magnet outside the core but firmly in contact with it on both ends. As a result, the vacuum freely replenishes the withdrawn magnetic energy that would otherwise have existed in the space outside the core, but the energy outside the B-field localization zone remains and appears as a separated and uncurled A-potential. In this way, the MEG is converted to an open system not in equilibrium with the vacuum, and can therefore permissibly output more EM energy in its output coils than the operator inputs into the input coils, the excess being freely furnished from the vacuum via the broken 3-symmetry of the permanent-magnet dipole and the resulting 4-symmetry EM energy flow from the vacuum's time domain into the dipole. The MEG is actually a practical macroscopic EM power system application of (1) the well-known Aharonov-Bohm effect and (2) the effect of the separation of the curled and uncurled potentials by a long tight solenoid or by a tight toroid, but in this case by the nanocrystalline magnetic core performing that separation function. In effect, the output coils interact with double (or greater) energy than would be available in just the B-field of the core (and of the magnet). Dual interaction with both the internal core B-field flux and also the external A-potential is obtained by pulsing, so that the pulse edges produce strong E-fields in space, with the E-fields striking the output coils in a transmission-reception near-field fashion. As is well-known., $d\mathbf{A}/dt = -\mathbf{E}$, and also $f(d\mathbf{B}/dt) = \mathbf{E}$. These two fields produced by the perturbed internal B-flux in the core and the pertubed external uncurled A-potential outside the core are in phase. This is how the extra pulsing A-potential energy outside the core is made to interact as -E-field energy with the Drude electrons in the output coils, while the changing magnetic flux in the core through those coils also simultaneously produces a second -E which interacts with the coil's conductors in normal E-field fashion. The total E-field interacting with the output coils is proportional to the *time rate of change* of the input perturbation signal, not the magnitude of its energy. By adjusting the leading and trailing edges of the perturbation pulses, the magnitude of the E-field interactions with the output coil is determined as desired. Further, the usual output B-field produced in the output coils by dE/dt is directly extracted into the core and held, so that *a purely electrical interaction and output occurs in the output coils.* This is easily shown since the output current and voltage of the output coil are in phase, within a degree or two. The MEG may be said to the world's first purely electrical induction transformer, freely separating the magnetic interactions and bottling them up, while allowing the electrical interactions.

asymmetrical regauging: A change of either the scalar potential ϕ or the vector potential A, or both, so that a *net resulting excess force together with a net change of total system potential energy* occurs in the system. During this change, an asymmetry exists in the active vacuum's energetic exchange with the system. The result is that the potential energy of the system is freely changed by energy from the vacuum, and a free net force is also produced in the system. This new force is then free to dissipate the excess potential energy by translating electrons through a load and powering it. By contrast, the Lorenz/Lorentz *symmetrical* regauging produces two free force fields that are equal and opposite, so there is no *net* resultant translation force available to translate electrons through the external load and power it. Thus Lorentz regauging arbitrarily applied to the Maxwell-Heaviside equations selects only that resulting subset of Maxwellian systems that are in equilibrium in their interaction with the external active vacuum. The potential energy of the system is changed, but only as a "stress potential" doing internal work upon the system to form and maintain its increased stress. The extra energy also rotates the system's frame out of the laboratory frame. Lorentz thus unwittingly discarded an entire class of permissible Maxwellian systems that are not in thermodynamic equilibrium with the vacuum. The Lorentz condition is initially violated each time anyone potentializes an EM circuit, but the closed-current-loop circuit design with the same charged-particle energy carriers around the loop — including back through the back emf of the source dipole in the generator — enforces a form of Lorentz symmetrical regauging during discharge of the system's excitation energy. Further, the response of the electrons in the usual electrical circuit with copper conductors is so rapid that the asymmetrical regauging (potentialization) is almost instantly converted to symmetrical discharging by movement of the electrons as current. When one section of the otherwise closed current loop uses a dramatically heavier charge as the energy carrier (as a battery uses ion current between its plates, but uses electron current between the outside of the plates and the external circuit), it is then possible to dephase the two independent carrier currents so that the battery will be charging while simultaneously the circuit is being powered, by a negative resistor (sharply increased scalar potential or voltage) produced at the plate surface interface of the two carrier current types. This proven Bedini process is an example of a true overunity process, capable of producing a self-powering system by means of adroit switching and timing.

asymmetry: Lack of symmetry, or "broken symmetry" in a general sense. Since a specific symmetry is accompanied by a specific conservation law, the breaking of that symmetry represents the violation of that conservation law. Particularly note that a broken symmetry at one level often (usually) results in a new symmetry at a higher level, necessary to maintain the asymmetry at the lower level. The resulting "levels" of symmetry are known as *hierarchies* of symmetry.

asymmetry of the dipole: Because of the proven asymmetry of opposite charges, any dipole or dipolarity is a broken 3-symmetry in its fierce exchange of EM energy with the active vacuum, because of the known broken symmetry of the opposite charges on its ends. Any two points in the universe (in or out of the circuit), that are at differing ambient potentials, constitutes such a broken symmetry continuously pouring out EM energy extracted from the vacuum. A dipole or dipolarity continuously receives and absorbs copious virtual energy from the seething virtual photon vacuum flux, and not all this absorbed energy is reradiated back to the vacuum in virtual form. Instead, some of it is coherently integrated into observable form and re-radiated in all directions as usable, real 3-space observable photons and EM energy. We found that the input energy to the dipole is from the time domain of 4-space. Hence the dipole's broken 3-symmetry results in the appearance of a new and far more fundamental 4-symmetry in energy flow between the time-domain and 3-space. This was the solution to the long-vexing source charge (or source dipole) problem in classical and quantum electrodynamics. Reinterpreting the Whittaker 1903 decomposition of the scalar potential reveals the mechanism for this 4-symmetry between time-domain EM longitudinal wave energy flow into a dipole and 3-space EM longitudinal wave energy flow out of the dipole, due to the presence of the proven broken 3-symmetry of the dipole. *EM energy flow is conserved in 4-space, but not in 3-space. Further, the 4-*symmetry energy flow from the source dipole (or source charge) continues as long as the source dipole (or source charge) remains intact so that the broken 3-symmetry remains intact. It is strongly indicated that this asymmetry of the dipole or dipolarity is what Gabriel Kron finally discovercd as his "open path".

back emf: The counter-electromotive force, usually between the end charges of the source dipole in the external power source for an electrical circuit (in its overall closed current loop). When the electrons in the circuit are forced against the emf— as when the "spent" or depotentialized charges from the ground reference of the circuit are forced back up through the emf of the source dipole to re-potentialize the charges — then work is done upon the source dipole to scatter its charges and destroy the dipole. Removing the dipole stops its receipt of vacuum energy from the time - domain (which was due to its broken 3-symmetry, now removed). It also stops the former dipole's transduction of that energy, and its former output of 3-space energy flow along the circuit and in space around the circuit. When this flow is occurring, a small component of it strikes the surface charges in the conductors and is diverged into the circuit (by axial precession of the surface charges withdrawing the "immediate stubs" of their fields and potentials into the conductor) to power it as the Poynting energy flow component. The huge remainder of the energy flow in space surrounding the circuit simply misses the circuit and is utterly wasted. Destruction of the source dipole removes the transduction of EM energy from the vacuum. It thus stops the powering of the circuit. So additional energy must then be input to the power source to perform work on its scattered internal charges, force them apart again, and reform and restore that source dipole, in order to resume powering the circuit by energy extracted from the vacuum by the dipole. It is simple to show that such a closed loop circuit uses more of its collected energy from the vacuum to destroy its source dipole, than it uses to power its load. Hence the circuit self-enforces COP<1.0.

back emf (tempic): There are actually two emfs in a circuit, not one. There is a 3-space emf and also an entirely unrecognized *tempic* emf. In either case, under proper circumstances there can be a back-emf of either one or both. See discussions under **electromotive force, back** and under **tempic back emf.** The tempic back emf is very important to COP>1.0 circuits, where the system's output energy density in spacetime is greater than the energy density at the input section of the system. This difference in tempic stress potential provides a tempic force sweeping from system output section back through the system to its input section. The current is a current of Dirac sea holes (negative energy 4-electrons) moving in that local vacuum, from the output section back through the system to the input section. At the input section, this Dirac sea hole current reacts with incoming electrons from the external power system. It therefore "eats" incoming electrons, so that the power supply must now furnish additional electron current to (a) power the negation and "killing" of the Dirac sea hole current by filling those holes with electrons in the input section of the circuit, and (b) simultaneously furnish the additional usual current to power the system's input. The result is that the tempic back emf is nature's mechanism for decaying the excited COP>1.0 energy state represented by a system in overunity functioning condition. Unless this hole current problem is overcome, this decay mechanism will destroy the net overunity, and it will prevent any attempt at close-looping the circuit for self-powering by use of clamped and governed positive energy feedback from output to input. The Bedini method can be used to transform the detrimental Dirac hole current into ordinary electron current, providing the self-powering feedback loop needed. Or, Bearden's outrigger method of using a multi-system set of external COP<1.0 receiving circuits powered by radiation (Poynting EM energy flow) from a single central power source (energy flow radiator) can be used. Then already-positive energy feedback from one or more of the separate outrigger COP<1.0 receiver circuits can be fed back in normal fashion to power the central power unit. Note that the outrigger concept uses the bridging concept, so that an external COP<1.0 system can be powered from a central radiating system, and the external COP<1.0 system can collect and output more energy than is output in the *accounted Poynting component* of the central radiating system. The excess energy radiated comes from the radiated Heaviside component.

bare electron: The true electron without its partial shielding of virtual positrons that are attracted toward it from the active vacuum, and that cluster around the electron, partially shielding its charge as seen by an external observer. The bare electron has infinite energy and infinite charge, as does the shielding cluster of virtual positrons. The difference between these two infinite values of opposite charge is the finite "externally observed" value of the electron charge listed in texts and handbooks. This strange dipolarity of the source charge thus exhibits an asymmetry in its exchange with the active vacuum virtual photon flux. Hence this mechanism also solves the source charge problem — how the charge continuously pours out EM energy in all directions at the speed of light, establishing its fields and their potentials reaching out into or even across the universe, without any input *of observable* EM energy to the source

Barret, Terence W.: Noted modern electrodynamicist and consultant who works in SU(2)xSU(2) gauge symmetry extended electrodynamics, and other models, and was also one of the pioneers of ultrawideband radar, Barret is one of the vigorous modern electrodynamicists continuing to develop and extend the science of electrodynamics and the electrodynamics model itself.

Bedini, John: Noted audio engineer, leading overunity energy researcher, inventor of the Bedini process for creating a true negative resistor inside a battery used as primary power to a circuit, and inventor of the Bedini overunity process for killing the back mmf (magnetomotive force) in a magnetic motor. Bedini is also the inventor of the renowned Bedini amplifiers well known to audiophiles, the BASE process for holographic sound, and a remarkable process for cleaning the harshness from digital audio disks. In addition, he discovered a mechanism for transducing into usable positive energy the detrimental Dirac Sea hole current (negative energy electron flow) backflow in an overunity system, flowing from the output section back through the system to the input section and even on out into the external power supply. The Bedini process transduces this detrimental current into normal electron current, thereby providing most or all of the required input power to the input section. Bearden then explained the technical process for Bedini's discovered transduction procedure, and Bedini and Bearden filed a patent application on this process.

Berry phase: Generalization by Berry of the Aharonov-Bohm effect, for an adiabatic situation. Later Aharonov and Anandan further removed the adiabatic condition and generalized the Berry phase to the geometric phase. See Aharonov-Bohm effect, geometric phase.

blocking: In a Fogal charge-barrier semiconductor, stopping or partially stopping the flow of current in a circuit, or between two points, by *pinning the electrons without stopping the EM energy flow*. In ordinary usage, "stopping the progress of whatever charge flow or current is being referred to.

Bohm, **David J.:** World-renowned physicist and originator of the hidden variable theory and interpretation of quantum mechanics. See David Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II," *Physical Review*, 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II). Together with his student Aharonov, Bohm co-authored a fundamental paper on the Aharonov-Bohm effect wherein interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields. From Bohm's hidden variable theory, the quantum potential has been highly weaponized by five nations, and these eerie weapons operating instantaneously at a distance through multiply connected spacetime are presently the dominant strategic weapons on earth.

Bohm's hidden-variable theory (HVT) of quantum mechanics: A major interpretation of quantum mechanics, formulated by David Bohm and published in *Physical Review* in 1952. Bohm's theory makes all the correct predictions and also eliminates many problems in quantum mechanics, such as the "measurement problem" and the — now recognized — problem of the missing chaos (i.e., the missing "hidden chaotic order."). Bohm's theory implies that physical reality — or much of it — can in fact be deterministically engineered. A key concept is the use of a *quantum* *potential*, which is multiply connected. Clandestine superweapons using **Bohm's** quantum potential (QP) have been developed and deployed by several nations. Present mutual assured destruction "standoff between these nations is based on surviving counters to QP weapons rather than surviving nuclear weapons.

Bohren, Craig F.: Distinguished Professor of Meteorology, Pennsylvania State University, noted for work in scattering theory. In an experiment **detailed** in C. F. Bohren, "How can a particle absorb more than the light incident on it?" *American Journal ofPhysics,* 51(4), Apr. 1983, p. 323-327 Bohren shows that in nonlinear conditions a resonant particle may absorb and output 18 times as much energy as is impinging upon it from the usual **field** definitions based on a static intercepting charge. The additional **energy** is "absorbed from the vacuum"; actually, it is absorbed from the nondiverged Heaviside energy flow component accompanying the accounted Poynting input component, but arbitrarily discarded by Lorentz. The resonant particle sweeps out a geometrical interception area (reaction cross section) greater than that of the same charge in a fixed static condition. Hence it intercepts part of the always-available Heaviside "dark energy" flow surrounding the Poynting energy flow component of every field/charge interaction.

Bohren experiment: Revealing experiment by Craig F. Bohren, replicated by Paul and Fischer, which shows that a given charge when in particle resonance absorbs more energy from a field or potential then the same charge does when in "static" position. This experiment is a direct experimental demonstration of the existence of Heaviside's nondiverged energy flow component around every field/charge interaction. See discussion under Bohren, Craig F. The Bohren experiment can be repllicated by any university nonlinear optics laboratory and outputs 18 times as much energy as is input by normal calculations (neglecting the Heaviside nondiverged EM energy flow component of the input).

boson: A particle having integral spin, which obeys Bose-Einstein statistics but not the Pauli exclusion principle. Bosons include the photon, alpha-particle, and any nucleus of even mass number.

bridge: A component or process which passes EM energy between two **isolated** circuits, but does not pass dq/dt between them, and which breaks the normal rigidly field-locked power dissipations in the two circuits. Space itself would appear to be one such bridge for EM energy flow.

bridging **function:** The function of passing EM energy flow between two isolated circuits, but not passing dq/dt between them, while simultaneously breaking any normal rigidly field-locked power dissipations in the two circuits. Use of the uncurled A-potential where $\nabla\times\mathbf{A} = 0$ is one available method of bridging otherwise isolated circuits, particularly in the outrigger embodiment of the motionless electromagnetic generator and other COP>1.0 systems.

broken symmetry: Fundamentally, a condition in which two parts of some configuration or shape, on opposite sides of some divisor or condition or boundary, are not similar but differ. In vacuum energy physics, we are very interested in the fact that — because of the broken symmetry of opposite charges — any charge (with its associated clustering virtual charges of opposite sign) and any dipole or dipolarity represents an asymmetry in its fierce energy exchange with the vacuum. This implies that something virtual becomes observable; i.e., part of the virtual EM energy continuously absorbed from the vacuum by the charge or by the dipole is changed into observable form and re-emitted as real, observable EM energy. The remainder is re-radiated as virtual EM energy. The charge's transduction of absorbed virtual photon energy and its re-emission as observable photon energy was proposed by the present author in 2000 as the solution to the long-vexing problem of the source charge and its associated fields and potentials (and their energy) reaching across the universe.

bypass resistor: In electrical theory, a resistor to pass or conduct current around (in parallel to) some other component. If another effect such as use of the overpotential, pumped phase-conjugate reflection, quantum well, quantum tunneling, negative energy, or charge blocking is added, the bypass resistor's function becomes extremely complicated. The bypass resistor may be complex, e.g., or it may be a true negative resistor in some cases.

carrier wave: A fundamental wave that is modulated by another wave or other waves, and hence "carries" the other modulating waveform(s). By stripping off the carrier in a demodulator, the carried waveform(s) emerges.

Casimir **effect:** The attraction of two conducting parallel plates in space, placed very close together, due to their influence on the active vacuum and on the vacuum's interaction with the plates. A part of the vacuum spectrum between the plates is "cut off and blocked or shielded, so that the vacuum potential between the plates is less than the vacuum potential outside them. The resulting potential gradient (force) on each plate from outside pushes the plates together. The Casimir effect unequivocally proves that it is possible to directly engineer the vacuum so as to produce forces and energy and physical effects in material systems. While it is a "small" white crow, it only takes one white crow to prove that not all crows are black. It is sufficient to conclusively prove that EM energy can be extracted from the vacuum. From there to practical EM power systems, fed by vacuum energy, is simply a product of modifying the prevailing electrodynamics and electrical engineering to incorporate the supersystem components and interactions, and then to gradually discover and develop the necessary technology. In this book we attempt to point the way by using the supersystem and providing some of the major necessary principles and concepts.

Casimir, Hendrik Brugt Berhard (1909-2000): Noted scientist who in 1948 correctly predicted that two parallel conducting plates, placed very near each other in a vacuum, would experience an attractive force due to their influence on the electromagnetic vacuum.

causality: The relationship between "cause" and "effect", or the interaction of 4-space cause upon a previous 3-space observation (effect), to produce a new 3-space observation (change in the previous observable or production of an exact 3-space replica). The basic notion in a causal systemis that the system's response to an input signal is dependent only upon past input values, not upon future input values. In a multiply connected space, the causality principle is dramatically changed and requires very careful accounting.

chaos: In dissipative dynamical systems, the dynamical evolution or ordering of the system that is aperiodic and highly dependent upon initial conditions of the system. The trajectories of the system move on a *strange attractor,* which is a fractal subspace of the phase space. The mathematical equations describing chaotic behavior are very nonlinear and so complex that at the present time they cannot be computed or predicted. Control of chaotic oscillations thus is ongoing in its scientific development, with the most advanced control work having been done by Russian scientists, but now openly available.

charge: *Charge* is the ongoing circulation of EM energy flow between the time and 3-space domains, as "seen" by the observer in 3-space. *Negative* charge is (i) the ongoing absorption, by a mass particle, of EM energy input from the time domain (from *ict),* (ii) transduction of the absorbed energy into 3-space EM energy, and (iii) re-emission of the EM energy in all directions in 3-space. *Positive charge* is (i) the ongoing absorption, by a mass particle, of EM energy input from 3-space, (ii) transduction of the absorbed energy into the time domain, and (iii) re-emission of the EM energy in the time domain. Alternately, the positive charge can be considered as (i) the ongoing absorption, by a mass particle, of EM negative energy input from the time domain (from *ict),* (ii) transduction of the absorbed negative energy into 3-space, and (iii) re-emission of the EM negative energy in 3-space. A similar inversion can be utilized to represent negative charge. Time reversal of EM energy can be seen to result from the combined broken symmetries of EM energy flow (in 4-space) between time and 3-space, of two opposite charges and thus of dipolarity itself. See paragraph 3.1 of Chapter 3.

charge, electrical (q).: In the sense of q being charged mass, where the mass m_q of charge q is associated with the phenomenon defined as charge (see above definition). In our view, an electrical charge q associated with a charged particle or mass may be defined as $q \equiv m_q\phi_q$, to first order. This fits the common use of "the charge" as an expression for a charged mass, and explains why a gathering of charges produces a potential ϕ reaching out from the "source gathering of charge q". The actual "charge" action associated with the mass is due totally to ϕ_q, and to the broken symmetry of the ϕ_q dipolarity in the vacuum flux exchange with m_q. The charge q can be further broken down into a set of composite dipoles if the gathering of a screen of virtual charges of opposite sign in the vacuum, surrounding the mass of the "bare charge" inside the gathering, is considered. At first order, ϕ_q can be expressed (from *a particle view)* as a change in the local vacuum virtual photon flux (VPF), due to the VPF exchange between vacuum and m_q. *In a classical approach,* the vacuum may be considered as a giant scalar potential comprised of bidirectional EM wave flows according to Whittaker's 1903 decomposition, except we must re-interpret the phase conjugate half-set of Whittaker's decomposition as incoming from the time domain or complex plane. The charge then may be considered to diverge so much energy flow from the time domain into 3-space "around" the mass, resulting in ϕ_q as the steadily diverged energy surrounding the mass m_q. The ϕ_q component is actually the massless electrical charge and massless of itself. This approach would then yield a definition of massless charge itself, which at present is undefined in physics, as pointed out by Feynman and others.

In the expression $q \equiv m_q\phi_q$, the ϕ_q represents a change to the local vacuum potential ϕ_{vac}. Both ϕ_q and ϕ_{vac} decompose into Whittaker phase conjugate longitudinal EM wavepairs. As can be seen, by altering or engineering these constituent wavepairs, a particular ϕ_q or particular local vacuum potential ϕ_{vac} can be given a deterministic internal structure and dynamics.

Thus the interaction of a given electron or group of electrons may be made **quite** different from the interaction of another electron or group of "apparently identical" electrons. Such structuring dynamics are met with in the internal structuring of water, e.g., where the fierce activity of H-bond making and breaking can be treated as a special ϕ_H and its "internal structure and dynamics" can be engineered. A leading scientist pursuing that type of effect is Jacques Benveniste, in France. Predictably, the o**rthodox** establishment has severely castigated these experiments, defending conventional models as absolute while completely ignoring the experimental results that refute those conventional models as being incomplete.

charge carriers: Anything — usually fundamental particles, ions, etc. — which consists of or contains charges and can be potentialized and move.

charge, magnetic: In the sense of charged mass, a magnetic charge is what is commonly called a "magnetic pole". The mass m of the magnetic charge is associated with the exchange of EM energy between the time domain and the 3-space domain. Essentially, the energy exchange action **known** as "charge" in the massless sense is the magnetic scalar potential, or monopole, when decomposed as per Whittaker and as reinterpreted by the **present** author, and as pointed out in quantum field theory by Mandl and Shaw. Note that our personal view discriminates between *massless charge and charged mass.* Massless charge is an ongoing action and process, involving energy transduction from one form (time energy or 3-space energy) to another (3-space energy or time energy). A charged mass is a **mass** associated with such action. Unfortunately, in the prevailing usage the word "charge" is just loosely used for both senses without discrimination. Hence one must carefully differentiate between the use of "charge" for "charged mass", and its use for an "ongoing process".

charge-parity-time (CPT) theorem: C is *charge conjugation,* which changes the quantum number of every particle into its antiparticle. P is the reflection operation known as *parity,* which turns an object into its mirror image and rotates it 180 degrees about an axis perpendicular to the mirror. T is t*ime reversal.* Any violation of time-reversal symmetry is always accompanied by a violation of CP symmetry (by violating either C or P symmetry) and vice versa.

Electrodynamics and the strong interaction preserve CP symmetry to great accuracy. Weak interactions do not conserve CP symmetry. T symmetry violation has been shown at CERN, the European particle physics laboratory in Geneva, Switzerland.

charge trapping (also "pinning"): Holding charges "pinned" or "trapped" at a site or in a location, by a barrier or force, so that they do not flow as $i = dq/dt$. A clustering of the pinned charges at the pinning site often occurs.

charge-barrier: A process or component or function — particularly in a semiconductor such as the Fogal semiconductor — that blocks the movement of charges q as current dq/dt.

classical electromagnetic theory: The classical theory of electrodynamics, in its modern form begun by Maxwell with his seminal-paper orally presented in 1864 and published in 1865. Today many variations have been made, particularly with respect to the basic group symmetry of the theoretical equations.

Clifford algebra: A special higher topology algebra founded by Clifford, which includes as subsets many other algebras of lower topology.

closed system: In the present approach, a system that in theory does not communicate with its environment, and does not exchange energy or matter between system and environment. Notice that this directly conflicts with the notion in thermodynamics (both classical and disequilibrium) that a closed system is closed only to mass exchange across its boundary, but open to energy exchange across it. General relativity has falsified that notion since 1915. Any system undergoing a change of energy also undergoes a change of mass, since mass is a special form of energy.

An ideal model in which dynamically and energetically the system is considered to be "isolated" as if nothing else existed. In short, the system is considered as if it were in some "magic box" and nothing outside the box can ever enter the box or affect the system inside the box, and nothing in the system can ever leave the box or affect anything outside the box. No electromagnetic system is closed. In short, we have corrected the thermodynamics "closed system" with open energy transfer, to a true closed system with neither energy nor mass transferring. That corresponds to the present thermodynamics notion of the "isolated system".

One performs a non sequitur when proclaiming treatment of any EM system as a "closed system", and of course modern physics substantiates this. The broken symmetry of the opposite charges on the ends of a dipole — and of an observable charge together with its clustered virtual charges of opposite sign — has been proven since 1957. The absence of any closed system is particularly true in electrical power systems. Here one is indeed allowed to input energy into the system to excite or potentialize it. That **of** course assumes that the system was "opened" long enough to take on excess energy (which also changed its mass)! Then as the system operates, losses and dissipations occur, in which case energy is considered to "escape" from the system permanently. In other words, the system was also continually "opened" so that energy could escape. In any EM system, every charge and dipole is already an open system, with free exchange of energy between 3-space and the time domain via that charge. In particle physics, every charge and dipole is a broken symmetry in its virtual particle flux exchange with the active vacuum.

The ancient thermodynamics definition of a "closed system" as one that exchanged heat but not mass, was formulated before thermodynamicists even knew that heat was a form of EM energy, or even knew what energy was. Hence we have changed that definition, since it is a non sequitur in view of modern physics.

There is really no such thing as a truly closed system in the universe, since *every* system is embedded in the active vacuum and is an open system in an energy exchange with the vacuum. Further, any change of energy in the system curves the local spacetime, which in turn interacts back upon the system. If that energetic exchange of the system with its "active vacuum and active curved spacetime environment" is symmetrical, then the system is in equilibrium with respect to its environmental exchange. In that case, for a limited number of purposes the system may be treated as if it were a closed system. In classical electrodynamics, Lorentz regauging of the Maxwell-Heaviside equations further reduced them to describe only Maxwellian systems in equilibrium with the active environment including the active vacuum. Consequently, any *net* interaction between the external vacuum (or the active local curved spacetime) and the system was excluded from the CED model. Obviously a *broken equilibrium* (i.e., broken symmetry) in that actual physical interaction is also not included in the CED model. Nor is it included in the classical (equilibrium) thermodynamics model with its archaic second law.

coefficient of performance (COP): Ratio of the usable energy output of a system divided by that portion of the total energy input that is input by the the operator's or experimenter. The COP is thus a measure of "return for using the operator's or experimenter's input energy" to produce — or direct the production of— useful work by the system". There are other ways to calculate the COP ratio; e.g., the average output power divided by the average operator input power. Contrast COP to the *efficiency* of a system, which is the ratio of the total usable output energy divided by the *total input energy from all sources*, including by the external active environment as well as by the operator. The efficiency of a system means how efficiently it produces useful output from its *total* input. The COP is the performance the system renders for the *operator's* input. Note that even a highly inefficient system — say, with efficiency $\xi = 0.25$ — may nonetheless have a COP>1.0, if the operator himself inputs less energy than the system outputs and the environment inputs all the rest of the output energy as well as all the energy in the system losses. As an example, a common home heat pump may have an efficiency of 40%, but a COP = 4.0. A windmill may have an efficiency of 20%, but a COP = ∞.

Also, for many systems one may evaluate their coefficient of performance for (1) the performance of the system as an energy transducer only, i.e., the ratio of energy the system outputs (whether used or not) divided by the energy the operator inputs; or (2) the performance of the system as a work producer, which is the useful work accomplished by the system divided by the operator's energy input. As an *energy transducer,* every generator and every battery has a COP»1.0, when both the Poynting diverged energy flow output and the Heaviside nondiverged energy flow output are accounted. Yet for useful work, the same system may have COP<1.0, because all the Heaviside nondiverged energy flow component is wasted, and some of the Poynting diverged energy flow component is wasted also.

Analyzed as an energy transducer only, every dipolar power unit — such as a generator or battery — outputs enormously more energy flow rate from its terminals than the mechanical energy rate input to the shaft of the generator, and than the chemical energy dissipated in the battery. So as an energy transducer, every dipolar power source is already a system exhibiting COP»1.0, prior to Lorentz' integration trick to discard the enormous Heaviside nondiverged energy flow output component.

cold fusion: An ad hoc term applied to the transmutations at low spatial energy (but high time-energy) achieved in electrolyte experiments, particularly when using specially prepared palladium electrodes after their loading with positive H+ or D+ ions, and particularly when deuterium is present in the electrolyte and the loading is by D+ ions. The *total* energy of the photon is greater the lower the photon's frequency and therefore the lower the spatial energy of the photon. Hence the cold fusion interactions actually use much higher *total* energy nuclear reactions than does present high (spatial) energy physics. Since "high energy" physics erroneously eliminates consideration of time-energy, the present high energy physics community is unjustifiably hostile to the very notion of cold fusion. The fact that little experiments on the lab bench use far higher total energy than the community's proud large accelerators is a bitter pill many high energy physicists cannot *tolerate*, let alone view with scientific objectivity. So again there is a classic confrontation between science from scientific results and "science as known in the prevailing dogma".

composite dipole: Dipole formed as one part of an isolated, observable "charged particle", consisting of a differential element of that observable charge on one end and a momentary clustering virtual charge of opposite sign. This is one way of dealing with the modern "infinitely strong dual dipoles" view of the charge, where the bare charge in the middle is infinite as is the clustering virtual particle charge surrounding that bare charge and partially shielding it. The difference between these two infinite charges is finite, and is routinely calculated to show that it is the standard value of the charge as seen by an external observer.

Cooper pair: In superconductivity theory, a dynamic, weakly bound pairing of electrons in a superconducting material below its critical temperature. In this pair, if the energy state with wave number sigma and spin 1/2 is occupied by an electron, then so is the state with wave number -sigma and spin -1/2. The Bardeen, Cooper, Schrieffer (BCS) theory uses this concept to provide a detailed microscopic theory of superconductivity.

COP (Coefficient of Performance): See coefficient of performance.

cosmological feedback principle: A self-regenerating cosmological feedback cycle ascribed by Puthoff as the source of the vacuum EM zero-point energy (ZPE). One may assume the existence of EM zero-point energy by fiat as part of the boundary conditions of the universe, or concieive of its generation by the quantum-fluctuation motion of charged particles that constitute matter. Puthoff calculated the latter possibility, assuming that the ZPE spectrum (field distribution) drives particle motion, and that the particle motion in turn generates the ZPE spectrum. This provides a self-regenerating cosmological feedback cycle, which in fact is consistent with the general relativity assumption that curvature of spacetime affects mass energy, and mass energy changes affect the curvature of spacetime. It would be interesting to see if Puthoffs concept can be made compatible with the broken symmetry of a dipolarity and Kron's open path concept.

curl concept of fields: The concept of a field as the curl of a vector potential (the vector potential may not be known). Any field so conceived has a divergence of zero *a priori*. One must be careful, however, because electrodynamics calculates only the field's reaction cross section with an assumed fixed unit point static charge at a point in space, leaving the field itself undefined. In short, the calculation of "the field" is actually a representation of the local point intensity of the field's interaction with unit point static charge. It is not the magnitude of the entire field (or potential) itself, *a priori,* but only of its reaction at a point with such a charge. In short, the field's reaction cross section with static charge is often misinterpreted as the magnitude of the field; at best it is an indication of the field intensity to static charge at a point.

curl-free magnetic vector potential: A field-free magnetic vector potential A without curl (swirl), hence lacing the B-field with which an A-potential is usually associated. For simplicity, one may visualize the curl-free A-potential as massless magnetic charge (and energy) in linear motion. Toroids and long solenoids have the characteristic of confining the B-field (curled potential) and holding it inside. However, drawing energy from any potential whose source charge or source dipole is not destroyed simply results in the full value (sum of the A-potential and B-field components) of the potential being replenished from the active vacuum, via the 4-symmetry energy flow mechanism in the presence of the broken 3-symmetry of the dipole. Hence one may easily increase the available potential energy density at will. This is a special form of gauge freedom allowing one to freely alter the potential energy of any electromagnetic system at will. Contrary to the dictates of EM textbooks, gauge freedom rigorously allows the direct and *free* amplification of potential energy of the EM system *at will;* the excess energy being freely furnished from the active vacuum processes. This principle incorporated in gauge field theory, has been known for decades and not previously used in the design, production, and use of COP>1.0 electrical power systems, prior to the invention of the motionless electromagnetic generator (MEG).

curved spacetime: A four-dimensional geometry (spacetime or ST) is used in general relativity, where the ST curvature is determined by the distribution of spatial energy such as mass-energy. In the general relativity view there is no such thing as an "empty" space "filled" with some energy substance. Instead, there is a combined space and time as a single entity, *spacetime,* which is itself energetic and is the very geometry we use in our mathematics. Any change in the local energy density of that spacetime constitutes a *curvature* of spacetime. In the new, more extended approach, any change in the spatial energy density or time-energy density of spacetime, or a combination of changes in both, constitutes a curvature of spacetime. In 4-space modeling, time is treated as a special, highly condensed form of EM energy. Specifically time may be regarded as ordinary 3-spatial EM energy compressed by a factor of c^2 and existing in the time domain rather than in 3-space. Time-energy therefore has essentially the same energy density as mass.

dark energy (negative): (Dirac holes, Dirac hole potentials, Dirac hole currents): A strange form of vacuum energy that creates repulsive gravity, causing the expansion of the universe to speed up (as observed and not otherwise explained by conventional theory). As a strong candidate for the dark energy, the present author has proposed a vacuum negative energy and negative energy currents in a negatively curved spacetime region partially containing excess Dirac sea holes devoid of electrons, along with normal filled Dirac sea holes. These special "Dirac negative action quanta" then form *negative action* (negative angular momentum, or — for simplicity — negative energy) potentials, currents, and fields. This type of vacuum condition containing some empty Dirac sea holes is referred to as a Dirac *-polarized vacuum.* However, it is not just simple static polarization, but possesses rich dynamics. Such dynamic Dirac sea hole regions occur in the vacuum in curved spacetime regions with tempic potential gradients and strong 3-spatial energy gradients, and thus time force existing across the region. The holes occur since the energetic fluctuations of the local vacuum energy can separate a Dirac hole from its occupying electron. These may then recombine, or in sufficient local spacetime curvature can remain apart. The more violent the region, the more severe are local spacetime curvatures in that region, and the greater the fraction of empty Dirac holes that is maintained on the average, in their continual forming and recombining. In those regions, a tempic potential is higher at the higher center magnitude of the spatial energy density. Hence there tempic forces from "inside to outside", providing a "tempic broom" effect to sweep outward the excess Dirac sea holes separated from their electrons in the curved spacetime. The end result is the presence of increasingly unoccupied Dirac sea holes toward the outside of violent regions of spacetime, until the effect is overcome by the decreasing spacetime curvature. These empty Dirac holes in regions and currents are a special form of vacuum negative energy, negative energy fields, and negative energy potentials, creating negative gravity (repulsive gravity) upon ordinary mass exposed to them. So the more distant violent astronomical regions have greater amounts of repulsive gravity being produced by their local Dirac-polarized vacua and negative energy fields and potentials. The result of this effect is that, at sufficient distance from us, the increasingly earlier, more violent events and resulting greater curved spacetime regions have created sufficient Dirac-polarized vacua to overcome the normal gravitational attraction between galaxies, etc. Even farther out, the increase in Dirac-polarized vacua overrides the normal attractive gravity between galaxies, etc. and produces net repulsive gravity. In turn this results in acceleration of the observed distant expansion of the universe. Dirac-polarized vacuum and its special dynamics can be accommodated in Sachs' unified field theory, which unites physics from quarks to the cosmos. See also **Dirac** sea **holes.**

On the laboratory bench, Dirac hole currents and effects are met with in a circuit or system producing COP>1.0, or very sharp 3-spatial gradients such as in violent electrical discharges. The effects become non-negligible at about COP = 2.0 and sometimes less. The tempic force back across the, system from output to input is sweeping the Dirac holes through the system. This can play havoc with attempts to close-loop the circuit for self-operation. In semiconductor components, strange effects are met due to their use of donor and acceptor materials, and temporary or permanent failure of semiconductors can result. The simple way to visualize this is that, in the system's supersystem, the spacetime component is curved and the vacuum component is highly nonlinear. All three components — overunity system, local nonlinear and polarized vacuum, and local curved spacetime — are interacting vigorously with each other. An operating overunity system represents an excited state of that local region of vacuum and curved spacetime, hidden in the euphemism "polarized" which is inadequate to describe the dynamics. With the dynamics, one may visualize that the Dirac sea hole currents are swept back through the system, seeking to recombine with electrons to eliminate the "polarization" and its dynamics, thus decaying the excited state back to equilibrium and destroying the overunity condition. In a robust system whose internal parts do not appreciably interact with these Dirac hold dynamics, the dynamics will reach the input section and extend back into or toward the primary power supply. In the input section, the feeder line, and in the power source itself, this results in "eating electrons". The Dirac sea effects thus act as a separate load imposed within and around the system input section or even back in the feeder line to the external power supply, in addition to the system load itself in the system output section. Simply put, the power supply must now furnish extra electrons to "power" the Dirac sea hole recombinations first, and additional electrons to power the system in its "normal" fashion. The result is to "draw additional power" from the source, where "drawing power" now includes the additional "eating of electrons" within the power supply itself, the feeder line, and the input section of the system.

Bedini and Bearden have filed a patent application on processes to convert the detrimental Dirac sea hole current reaching the system's input section.

transforming it into ordinary electron current. Magnetic Energy Ltd. has developed two other processes that handle the Dirac sea hole current problem, allowing close-looping for self-power ng operation. One of these processes converts negative energy into positive energy and the other avoids the negative energy problem. The first one allows this otherwise devastating special form of "negative feedback" or "degenerative feedback" to be converted into "positive feedback" or "regenerative feedback", allowing the system to be close-looped successfully for self-powering. Bearden has proposed yet a different approach using an "anti-circuit" method, but at this writing the method is still proprietary until intellectual property rights are secured.

dark energy (positive): Our term for the vast Heaviside nondiverged energy flow component surrounding every field/charge and potential/charge reaction but unaccounted in present electromagnetic theory after being arbitrarily discarded by Lorentz circa the 1890s. Present theory only accounts for that small amount of energy flow (the Poynting component) diverted locally around the interacting charge. All the rest of the energy flow comprising the field or potential, and missing the intercepting point charge, was discarded arbitrarily by Lorentz and is still discarded today, using Lorentz's integration trick. In honor of Heaviside, who **did** recognize the gravitational implications of his dark energy flow component, the present author has nominated this vast, unaccounted "dark energy" component as being responsible for the extra gravity holding the spiral arms of the spiral galaxies together. See **spiral** galaxy.

So one must recognize that the term "dark energy" is indefinite, since there are two kinds: (a) dark positive energy, and (b) dark negative energy. Dark positive energy consists of "normal" energy as often connected with particles, collected upon them and affecting their momenta, etc. It has dynamics and effects consistent with that kind of energy. Dark negative energy is a condition of the vacuum itself. It represents a sort of "empty container and empty container dynamics" (negative energy dynamics) in the vacuum itself, which "eats positive energy", so to speak. In philosophical terms, dark positive energy represents unaccounted actions and effects from unaccounted "extra presence of real substance and things and actions". Dark negative energy represents unaccounted actions and effects from "extra absence of real substance and things and actions beyond zero". Dark negative energy exhibits precisely appositive effects to the effects produced by dark positive energy. E g., dark energy creates the excess gra*vity* that is holding the arms of the spiral galaxies together. Negative energy resulting from spacetime curvature conditions in highly energetic astronomical phenomena produces the *antigravity* that is responsible for the acceleration of the expansion of the universe. We have nominated the negative energy Dirac sea hole phenomena as the cause of the antigravity that accelerates the expanding universe. Interestingly, these negative energy phenomena present in sharp discharge (strong gradient) phenomena appear to be responsible for the fact that such phenomena are known to violate present thermodynamics, and are being studied under the aegis of "extended thermodynamics". See discussion of areas violating thermodynamics in the last chapter of Kondepudi and Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, New York, 1998, reprinted with corrections in 1999. A reference is D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

degenerate semiconductor: A semiconductor whose conductivity approaches that of a metal.

demodulation: From a carrier containing a modulated signal, recovering the signal that was used to modulate it.

deterministic pattern or "template": The exact arrangement, by plan and deterministic action, of a group of entities. The deterministic actions and their "controlling forces" are simply called "dynamics", "structuring", etc. An example might be deterministic field structuring and actions of photons in a "herd" or "group". Or of a set of spacetime curvatures, into a dynamic structure called a *spacetime curvature engine* — or *engine* for short, so long as the meaning is clear. Or by hidden order in the virtual photons in the vacuum flux, or a deterministic set of changes to that vacuum flux (the vacuum flux that is called the "vacuum potential." We have called this *internal EM dynamics* function — of deterministically patterning or templating the virtual flux (particle view) of the potential, or of deterministically patterning or templating the longitudinal and time-polarized EM wave decomposition (wave view) of the potential - *dimensioning* the potential. See **dimensioning.** We refer to the deterministic structure and its structural dynamics as the "engine template".

dielectric: Literally, "against electric (current)", originally coined to mean "opposes the flow or conduction of electric fluid". In today's language, a dielectric is a non-conducting material used to oppose or prevent the flow of electric charges. Even empty vacuum exhibits some dielectric resistance. "Dielectric" is also used to refer to a nonconducting material (often between the metal plates of a capacitor or elsewhere) that can (i) sustain a change in potential across a distance (i.e., an E-field), and (ii) serve as an insulator. Space itself is a dielectric, or at least is said to exhibit certain dielectric qualities.

dimension: A certain primary geometrical physical attribute, such as length, used to describe the separational relationships of physical phenomena. Actually, "dimension" just means "fundamental quantity of separation" and "geometry" in the most general sense. Dimensions thus become a sort of set of fundamental variables describing the *extensivity* of system and the dynamics of that extensivity. The number of dimensions taken in advanced physics usually represents a somewhat arbitrary choice for good modeling fit. In modern Kaluza-Klein theory, for example, it becomes necessary to accept some 10 or 11 dimensions in the usual particle physics case. Sometimes more than 20, or even an infinite number, of dimensions may be used. General relativity (GR) has incorporated the interaction of the abstract geometry (spacetime) with mass, so that *geometry is a real, active, dynamic entity*. Until Sachs' extension of Einstein's work into a unified field theory and Evans' incorporation of O(3) electrodynamics as an important subset of Sachs' unified theory, GR has largely remained a nonexperimental discipline, at least in the laboratory. With the new Sachs-Evans approach and with recovery of (a) the Heaviside dark energy flow component (b) overunity EM systems arbitratily discarded by Lorentz symmetrical regauging, and (c) modeling of the complete supersystem including system, vacuum, and spacetime in mutual interaction, GR suddenly becomes substantially engineerable by novel electromagnetic means, both in the laboratory and in operating devices. We are thus entering a great new age of electrogravitation and a new physics. Research in antigravity lifters, COP>1.0 electrical power systems, medical therapy utilizing engines and anti-engines, and cold fusion are the elements of the new science that are struggling to be born and funded.

dimensi**oning:** Coined term meaning "forming an exact pattern or template of structures", particularly of nested curvatures of spacetime, where the pattern/template may also be dynamic, to include patterning and dynamics in the time-energy domain as well as the 3-space energy domain. It involves the internal structuring of the "dimensioned" entity. The term "dimensioning" makes sense in unified field theory, but is awkward in normal EM theory, which erroneously assumes that EM energy propagates in an uncurved spacetime. The mere presence of a change in local spatial energy, created by the wave energy itself when present in that local region, curves the local spacetime. Note that the background potential may be changed in a steady state level as well, which means that the frame of the system and its field dynamics has been rotated away from the laboratory frame. A time-polarized EM wave, e.g., changes the time-energy density and also curves local spacetime very powerfully. Hence the EM wave rigorously travels in curved spacetime, and in fact identically is a propagating, oscillating curvature of spacetime. What is missing from present physics is all accounting of the extraordinary Heaviside nondiverged energy flow component accompanying the accounted Poynting diverged energy flow component. Thus present physics *does not properly account for* the gravitational aspects of "electrodynamic field and potential energy in space".

Dimensioning also implies a deliberate form or structure (with dynamics) created in the virtual photon flux of the vacuum, in the virtual photon flux comprising a scalar potential, or in the infolded longitudinal-wave EM inside all EM potentials, fields, and waves. It implies an associated set of spacetime curvatures and their dynamics. The latter are referred to as "engines", "vacuum engines", "spacetime curvature engines", etc.

Dimensioning the potentials or signals or carrier waves of conventional electrodynamics involves adding hidden *vacuum engines* to them. The term *dimensioning* is descriptive and was chosen because adding such templates is equivalent to increasing the EM topology by adding mathematical dimensions. Since an EM wave or field or vector potential can be decomposed into two scalar potentials, then "dimensioning" one or both of the scalar potentials can "infold" the desired hidden structures or patterns (vacuum engines) inside an EM carrier wave, EM field, or vector potential created by the interference of the two dimensioned base potentials. This is referred to as *dimensioning* those EM field and wave entities. Internally structuring the two scalar potentials via arrays of longitudinal EM wave emitters and applying the necessary dynamics (modulations), is the primary manner to directly engineer dimensioning. When dimensioning is present, two EM waves that appear identical as seen on the oscilloscope may have dramatically differing internal structures (dimensioning). Hence they may produce drastically differing effects when they interact with the same or similar objects. *The only way in which the action of a vacuum engine can be resisted is to create and simultaneously introduce the exact antiengine for it — a special sort of forcible symmetrical regauging, so to speak.* Once the action has been completed in the receiving object and it has been physically changed, the only way it usually can be undone is to form a precise antiengine for that action, amplify the antiengine, and let the antiengine act upon the object for a sufficient time to "time-reverse it back to its previous physical state and condition". The concept has great utility in electromagnetic healing, not discussed in this book. We point out that the decades-long induction of diseases, health changes, etc. in U.S. personnel in the U.S. Embassy in Moscow occurred in forcefield-free regions. In short, the changes occurred in regions where the potentials were stable and unchanging, and so their engines were stable and constantly applied to the exposed personnel. The Soviets were simply radiating a high level target, the U.S. Ambassador and his staff, to guarantee high-level U.S. attention. By U.S. reaction to the deliberate electromagnetic "engine" induction of disease and health changes (three Ambassadors sickened and later died), the Soviets could positively ascertain whether we knew of the "infolding" or "dimensioning" technology, and whether we had developed defenses.

dipole: In the simplest case, two separated charges of usually equal magnitude and of opposite sign. More generally, a localized positive charge or charge distribution, and a localized negative charge or charge distribution, such that the net charge summation is zero, while the positive and negative charge distributions do not precisely superpose point to point but only on the average. We point out, however, that in regular physics electrical *charge* has no proper definition. As is well known in particle physics.due to the opposite charges on its end a dipole is a broken symmctry in the virtual photon flux of the vacuum. By definition of broken symmetry, this means that *some* of the virtual energy flux continuously absorbed from the seething vacuum by the charges of the dipole is not re-radiated as *virtual* energy. Instead, it is integrated *coherently* and re-radiated in 3-space as real emitted EM energy, establishing the fields and potentials associated with that source charge or source dipole. The dipole is therefore the most fundamental true "negative resistor" since it freely and continuously receives EM energy in unusable form, transduces it into usable form, and re-emits it in usable EM form.

It follows that any positive charge in the universe and any negative charge in the universe form a dipole and a broken symmetry in the vacuum flux. Hence every such dipolarity in the universe already extracts energy from the vacuum and puts it out in 3-space, or absorbs energy from 3-space and returns it to the vacuum. Kron called this an "open path" electrically, well before broken symmetry was formally discovered and experimentally prove in l957. Kron was also able to use the open path in and around special electromagnetic circuits to cause the circuits to become true negative resistors, freely powering the network analyzer at Stanford Universoty even in the 1930s.

In classical EM theory, the problem of how every charge and dipole can continuously emit EM energy to form all the associated fields and potentials and their energy has been an unsolved problem — sometimes called the most difficult problem in electrodynamics, both classical and quantal. For the solution and a discussion, see the present author's paper "Giant Negentropy from the Common Dipole," *Journal of New Energy,* 5(1), Summer 2000, p. 11-23. Symmetry represents a conservation law, and broken symmetry represents violation of that conservation law. Because of the broken 3-symmetry of the dipole, the EM energy flow is not conserved in three dimensions — in this case, violation of conservation of 3-space EM energy flow is not conserved. Neither is the time EM energy flow conserved in the 4th dimension. Instead, EM energy flow is conserved in 4-dimensions, between the inflow in the time-domain and the outflow in 3-space in all directions, or vice versa. The hidden inflow of EM energy to the dipole (and to a charge) is actually from the complex plane (time domain). Thus all EM energy in 3-space is actually received by the source charge or source dipole from the time-domain, and returned to the time domain. What we call "propagation of EM energy through 3-space" is actually the spreading of this 4-circulation to successive 3-space points. The importance of the internal structuring of time-flow and of longitudinal EM waves in the time domain (dimensioning in the time-domain) is apparent, since it produces very powerful engines that can act upon matter in any fashion desired. Indeed, the Russians have developed clandestine robot engine systems comprised of sets of spacetime curvatures and their dynamics (which are just sets of longitudinal EM waves and their dynamics). Any function of a weapon system — sensing, weapons (using scalar interferometry), communications, propagation, etc. — can be incorporated. The capabilities of such systems, such as travel through continuously emitted EM signal beams from satellites, communications stations, power lines, etc., is eerie. Such a system can reside in any convenient EM potential or field as a "dimensioning" of the potential or field. Such robot engine systems have been clandestinely tested, but discussion of them is beyond the scope of this book. Suffice it to say that these systems are ushering in a dramatic revolution in future warfare, and the present research in physical nanobots is already decades behind.

Very powerful support for the author's solution of the source-charge problem is given by F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised edition 1993, under the heading "Covariant Quantization of the Photon Propagator" in Chapter 5. In their deeper coverage of the photon polarizations, Mandl and Shaw point out that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential.

Translated from particle terminology to wave terminology, Mandl's and Shaw's argument directly fits my re-interpretation of Whittaker's 1903 decomposition of the scalar potential.

Dirac, Paul A. M.: Noted English physicist, 1902-1984, who shared the Nobel Prize for developing wave mechanics and predicting the existence of the positron. He also originated the notion of "holes" or negative energy states in the vacuum, where the holes are ordinarily filled with electrons. The "vacuum replete with these Dirac structures" (all filled with electrons) is referred to as the *Dirac Sea.* If some of the holes are empty, then that is the Dirac sea of a curved spacetime. We strongly stress that a Dirac sea **hole** (negative energy electron with negative charge, negative energy EM fields and potentials, and with negative mass-energy) prior to observation (prior to its reaction with mass) is a quite different entity from the positron (after reaction with mass, so that the positron has positive mass-energy, **positive** charge, positive energy EM fields and potentials, and has been reversed in its 3-space propagation direction). Specifically, the Dirac sea **hole** produces negative energy fields and potentials, whose powerful Heaviside nondiverged components enable to perform practical antigravity. The (observed) positron produces gravity, not antigravity, and also positive energy fields and potentials.

Dirac sea: A concept and description (modeling) of the vacuum as containing — amongst other things — a "sea" of an infinite set of negative energy slates ("holes") for electrons, almost all of which are usually filled with electrons in the absence of any spacetime curvature of significance. In Dirac electron theory, the relativistic wave equation for the electron has four components, which correspond to two spin orientations and two energy conditions: positive and negative. Ordinary electrons are in the positive energy condition. So then the vacuum must also possess a "sea" of negative electron energy conditions. These negative conditions are called "holes" and are ordinarily filled or mostly filled with electrons which have fallen into these negative energy states. The Dirac sea is this vacuum sea of negative energy electrons (and a perhaps a few empty holes). In forward time, if energy is added to a negative energy electron, electrons can be **lifted** from this sea, since the added energy creates a local curvature of spacetime. If spacetime is locally curved, additional holes may be created so that, when the holes meet normal positive energy electrons, the electrons fall in and disappear, creating the appearance of an "electrical energy sink." This can happen so that the local ST curvature merely relaxes back to an uncurved spacetime, therefore avoiding radiation from "pair annihilation". If the curvature is in the opposite direction (i.e., if positive energy is added to the vacuum), electrons may be lifted from the Dirac sea by the resulting spacetime curvature energy, thus producing a flow of electrons "right out of the nonlinear vacuum" and generated by the spacetime curvature, and creating the appearance of an "electrical energy source" from the vacuum itself. In that case the energetic holes left behind also remain, and are affected by all EM fields etc. These holes thus form negative energy currents, fields, potentials, waves, and other electromagnetic entities such as negative energy charging and discharging (excitation and de-excitation).

In an overunity EM system, there is a higher energy density in the output section than in the input section. This includes time-energy density. So a higher tempic potential exists at the output than at the input. A gradient in the tempic potential thus exists, which constitutes a tempic force from the output section back through the system to its input section, and even on back through the conductors to the external power source, and even back into the distant external power source itself. Thus the power source has to furnish additional electron current. It must "power the holes", so to speak, by furnishing the electrons that are "eaten" by the Dirac hole current to eliminate it, which represents a novel kind of load in the input section of the system. So the power supply must furnish more power, so as to furnish the normal power to power the system, and the extra power to "power the elimination of the Dirac sea hole current as a back-reflected load". This is nature's way of decaying the excited state represented by COP>1.0, and restoring equilibrium and COP<1.0. By the Bedini process, the otherwise detrimental Dirac sea hole current reaching the input section is transformed into ordinary electron current, thus serving as input power to the system. By adjusting the system and optimizing it, this strange kind of "positive power input from converted Dirac hole current negative power feedback" can be adjusted to equal the input power required by the system. At that point the external power system can be disconnected, and the unit continues in close-looped and self-powering operation, including powering its load as well. Gabriel Kron succinctly described this condition and adjusting for it. Such a system allows energy to be continuously extracted from the vacuum, and a load to be powered continuously, with all the energy to run the extracting system being freely received from the active vacuum.

Dirac sea hole: A negative energy state for an electron, in the Dirac theory (in the Dirac sea). Literally the hole (negative energy state) when not occupied by the electron. In flat spacetime, almost all Dirac sea holes are filled with electrons. A small fraction on the average are not filled due to the local fluctuations of the vacuum energy continually lifting electrons from their holes and then having them recombine, usually so quickly that no radiation occurs in the observable state. The situation in the virtual state can be described by Feynman diagrams. This received view can be falsified in a curved spacetime. There the holes can exist and flow as currents, as validly as positive energy electrons can flow as currents, without the incessant "filling" activity of the vacuum itself. In addition, in such a hole-stabilizing curved spacetime region, the hole can "eat" or recombine with a real electron, so that the electron vanishes but without radiation. If we wish to join the two views, then we take Feynman's and Wheeler's view that the field does not exist in vacuum as such, but only the potential for it exists — in short, only the engine exists. The engine can then be said to be an ordering (chaotic self-ordering) condition present on the statistical random variation of the fluctuations of the vacuum. With longitudinal EM wave technology, this "order in the midst of disorder" — this templating and engine technology — can be directly engineered.

Dirac sea hole current (positron current): Just as a current of positive energy electrons can be generated, a current of negative energy Dirac sea holes can also be generated if the local spacetime is properly curved. Mathematically these are usually considered to be "positron currents", except they have no positive mass, because a hole will "eat" an electron. In an overunity electrical power system, the tempic potential is greater at the output of the system than is the tempic potential at the input section. There thus exists a "tempic force" or "tempic broom" sweeping Dirac sea hole current (positron current) out of the input section of the system into the external power supply. Simple instruments will record this current as if it were electron current being furnished from the external power supply to the system. Also, the Dirac sea positron current will "charge" the battery with Dirac sea holes (mathematical positrons), which in the conventional sense discharges the battery. A battery may become highly charged with Dirac sea holes. If so, then when placed on a normal battery charger, the electrons in the charging current will be "eaten" for an extended period by falling into the Dirac sea holes, and the battery will not acquire any charge until they are filled — whereupon the battery will sudden begin charging up normally again.

We also point out that the vacuum, containing the Dirac sea, actually is not "massless" at all, but has zero resultant mass on the average, as a result of the simultaneously presence of positive and negative mass in average equal abundance. Any curvature of spacetime must violate this "net sum zero" mass, hence produces inertia in all masses. Note that a negative mass can also exhibit a negative inertia; i.e., it accelerates and enhances the action of any force acting upon it, instead of opposing it. In short, negative mass exhibits negative inertia. We suggest that this effect is probably involved in the more violent explosive phenomena of astrophysics.

disequilibrium: Broken equilibrium, which in turn involves breaking appropriate conservation laws unless compensated.

displacement current: The apparent continuation of a conduction current through a region of space without a conductor, e.g., between the plates of a capacitor. Rather than being the motion of charges in a conductor, the displacement current is related to the rate of change of an E-field, and thus to the rate of change of the strain of the dielectric.

dissipative structure: In open systems far from thermodynamic equilibrium, a region of self-ordered behavior of a special kind in matter, characterized by symmetry-breaking, multiple choices, and correlations of macroscopic range. By correlations we mean statistically reproducible relations between distant parts of the system. Formation of a dissipative structure represents the birth of complexity in that region by self-organization. Usually a dissipative structure emerges only after the driving constraints reach a critical value. As the embryonic "engine" technology advances, however, the direct engineering and control of dissipative structures will become commonplace and routine.

distortion correction theorem: "If a scalar wave $E_1(r)$ propagates from left to right through an arbitrary but lossless dielectric medium, and if we generate in some region of space [say near $z = 0$] its phase conjugate replica $E_2(r)$, then E_2 will propagate backward from right to left through the dielectric medium, remaining everywhere the phase conjugate of E_1." Quoted from Amnon Yariv, *Optical Electronics*, 3rd Edn., Holt, Rinehart and Winston, New York, 1985, p. 500-501.

What is missing from the distortion correction theorem is hidden in the phrase *"if we generate in some region of space ... its phase conjugate replica. "*. Since electrodynamicists and optical physicists have been unaware that all EM energy at any point in space comes from the time domain to that point, then —*prior to detection by interaction with charge* — the actual phase conjugate wave exists in the imaginary plane (in the time domain) and *not* in 3-space. The continuous interaction of that imaginary plane wave with charges, to actually produce the force-field wave in space, is *assumed* in the standard nonlinear optics. In short, one is looking at the "effect" wave, not the "cause" wave. However, since time is not an observable even in theory, it is this "effect" wave in 3-space that will indeed be measured by our instruments, because they must interact the imaginary or phase conjugate field with charge in order to detect or observe it (measure it).

The proof of this is experimentally already known but not recognized. If one reflects an EM wave off a normal mirror, the mirror must absorb the incoming wave, which transfers the wave momentum to the mirror to repel it, then the mirror must re-emit it, which adds that much more momentum to the recoil. So the mirror recoils with twice the momentum of the reflected wave. That produces a wave going in the other direction. If the incoming wave is continuous, one then has "paired" waves going in **opposite** directions). However, if an incoming wave strikes a phase *conjugate* mirror (PCM) material and is phase conjugated, a wave going in the opposite direction also emerges, but it synchronizes point by point with every spatial point occupied by the incoming wave. Also, *the phase conjugate mirror does not recoil at all.* This is well known in phase conjugate optics. It clearly shows that the origin of the "detected" phase conjugate wave is quite different from that of a normal reflected wave. Indeed, by our reinterpretation of Whittaker's 1903 biwave decomposition of the scalar potential, at every point along the detected phase conjugate wave, the energy enters that point from the time domain, and not from some sort of reflection from the PCM. The energy also returns to the time domain from that point. Its presence at the point produces a separation of virtual point charges, or the well-known polarization of the vacuum. So the flow of energy is from the time domain to the negative charge of the point dipole, thence to the positive charge of the dipole or others immediately surrounding it, thence back to the time domain. That is the operation in a "time-forward" sense to the external observer. In a "time-reversed" situation, the 4-circulation appears in the opposite direction. We propose that this fundamental 4-circulation is responsible for and constitutes spin itself. The direction of flow determines the 'polarity of the charge" we obseve. In this view, the point dipolarity in the virtual state with its ongoing 4-circulation can be taken as the very definition of energy itself.

Druge, Paul: Important German physicist, 1863-1906, who formulated the theory of the free electron gas inside conductors, using it to formulate a theory of metallic resistance.

Drude electron gas: The electrons that have broken loose from their orbital bonds in atoms in a conductor, and are thus free to "wander around" and "migrate" — or flow as current when exposed to an emf. Also called just the "electron gas". Note that, in the metal lattice of the conductors, each position vacated by a Drude electron is a "hole" or positively charged little region. Not only do electron currents exist as net negative charge migration in conductors and materials, but also hole currents exist as net positive charge migration from atom to atom, or a sort of "ionization migration flow". Electrons are continually recombining with these "ion holes", however, to reduce the hole current. These considerations are particularly significant in semiconductor materials.

duality principle: In quantum mechanics, the principle that photons and other particles propagate like waves and interact like particles, so that their description as only waves or only particles is inadequate.

Edison, Thomas Alva: Noted American inventor, 1847-1900, whose staggering portfolio of inventions included the storage battery, electric light bulb, phonograph, motion pictures, carbon microphone, and many others. Edison actually ran multiple "patenting research" laboratories as discovery and patent generating facilities. Not only was he an inventor himself, but he also employed a very good staff of researchers for invention and discovery.

E-field (electric field): The electric field intensity E; that field on an electric charge (charged mass) which produces an electrical force resulting in a linear motion of the charge. In reality, this electric field *as usually defined* does not and cannot exist in empty space devoid of charge, but only in and on charged matter, since mass is a component of force by $\mathbf{F} \equiv \partial/\partial t(m\mathbf{v})$. More specifically, the E-field is the local intensity of the field at a point, in its interaction with a unit point static charge fixed at the point. Even more specifically, it is the output of an observation/detection operation, which is a *d/dt* operator imposed on the 4-space causal field interaction with the 3-space charge, resulting in a frozen 3-space "slice" or snapshot of the deviation of energy *from* the field, forced to flow around that intercepting charge, at that single instant. In short, the E-field as a force field is an *effect* and not a *cause.*

The greatest error in all electrodynamics is to then assign this same *frozen effect force field intensity snapshot* — or an implicitly assumed iterative sequence of such snapshots — as the dynamic and nonfrozen *causative field itselfm* 4-space, prior to the field's interaction with charged matter. This is the substitution of effect for cause, a grave non sequitur.

When EM theory was founded, the founders believed there was no single **point** in the entire universe where mass was absent; instead, a thin material "luminiferous ether" was considered to fill all space. Time was considered immutable, and there was no concept of a dynamic *spacetime.* Hence at each point in such a material ether, the reaction with charged matter (etheric matter) was thought to universally exist. Hence the field was indeed a very material entity, in their concept!

With the destruction of the material ether notion in the 1880s, not a single Maxwell-Heaviside equation was changed to eliminate its implicit assumption of the material ether. Hence this enormous error of confusing the *effect* field with the *causal* field (prior to observation-detection-interaction) is ubiquitous in electrodynamics and therefore in physics itself. The two fields (causal and effect) do not even have the same dimensionality; one is dynamic in 4-space while one is frozen in a 3-space snapshot. This has led to the false notion that an EM force field wave "movesthrough" 3-space, which does not happen at all. What really happens is that, if we assume a continual iterative *d/dt* observation operator, we have a continual series of such frozen 3-space snapshots, like the frames of a motion picture film. But none of those snapshots have any inherent motion at all. For a statement of the falsity of the conventional illustration of the "EM wave in space as an orthogonal pair of oscillating E and H fields in a transverse plane", see Robert H. Romer, "Heat is not a noun." *American Journal of Physics,* 69(2), Feb. 2001, p. 109. In endnote 24. **p** 109, AJP editor Romer takes to task *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate. " "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place.* "Also see charge.

E-field, **motional:** In classical EM theory, the electric field produced by moving a conductor in a magnetic field. Often called *motional inductance* of the field. In the conventional theory, motional inductance is deemed to be identical to the inductance caused by a time rate of change of the magnetic field, experienced by the moving conductor. Or vice versa, experienced by a stationary conductor in a moving magnetic field.

E-field, static: In conventional theory, the electric field in the space surrounding a static charged particle, induced by the presence and action of the charged particle. There is no notion as to the *causative mechanism for this static* field of and from a charge, which implicitly is assumed to just continuously create and pour out EM energy in all directions, thus grossly violating energy conservation. Consideration of this anomaly leads to what has been called the most pressing problem in both quantal and classical electrodynamics: the question of the formation of the fields and potentials associated with the "source charge". Either the charge totally destroys the conservation of energy law itself, or there must be a *virtual* (nonobservable) energy input to the charge, since instruments establish that no observable EM energy is input to it.

For a solution to this problem, see the present author's "Giant Negentropy from the Common Dipole," *Journal ofNew Energy,* 5(1), Summer 2000, p. 11-23. We later found our solution to be consistent with quantum field theory and with a slight reinterpretation of Whittaker' s 1903 decomposition of the scalar potential.

Ehrenhaft, Felix: Scientist who performed numerous experiments believed to demonstrate fractional charges. We point out that classical electrodynamics assumes a local flat spacetime, a notion long falsified by general relativity, since 1915. In a local curvature of spacetime, the dynamics of the charge (modern view as the finite difference between two local infinite charges forming a dipolarity) can change the observed value of the net finite charge. So finite fractional charges, or even supercharges, are quite possible.

Einstein, Albert: German (Jewish) physicist, 1879-1955, who emigrated to the U.S. and formed special and general relativity theory, discovered and formed the theory of Brownian motion, recommended the development of the atomic bomb to President Roosevelt, and profoundly affected the course of science and the modern world,

Einstein, Podolsky, and Rosen (EPR) paradox: Thought experiment in 1935 for measuring two correlated photons separated and at distant locations, protesting that quantum mechanics requires that the particles communicate instantly, which is a "spooky action-at-a distance" (Einstein's phrase). The experiment eventually was performed and it works, just as required by quantum mechanics. The paradox inspired the development of hidden-variable theories.

Einstein's general theory of relativity: Mathematical theory of gravitation in which the gravitational force is mathematically described by a curvature in spacetime. In other words, the geometry and its dynamics have become physical agents, doing away with the old notion of the agent being something separate that is occupying inert spacetime. It is important to note that, in general relativity, mass-energy acts on spacetime to curve it, and the curvature of spacetime acts back on mass-energy to translate or change it. What seems missing from GR — at least in its practice — is the reaction of time-energy upon the geometry to curve it, and that corresponding reaction of the spacetime curvature back upon time and time-energy to change it.

Further, we consider that an exact pattern (template) of curvatures of spacetime and their dynamics exists and is back-acting upon every mass system and its dynamics. This is the *engine* concept. The most general way to engineer matter is not by brute force translation forces, but by forming spacetime engines. Once formed, these engines act upon the interior of any mass placed in the engine region, to eliminate the deltas existing in that muss's resident engine and the impressed engine. *The forces arise from the local spacetime itself, as curvatures in that local spacetime, so the action is from "inside" to "outside" an object, starting and continuing from every point in the object.* The curvatures of spacetime itself furnish the energy for the continuing actions upon the interior of the mass, once we have paid to form the engine. The mass can be acted upon at any level, from its gluons and quarks to its nuclei, to its atoms, its lattice forces and binding energy, etc. Significantly, the Sachs unified field theory applies to physical reality from beneath the quarks and gluons to the entire universe itself, and folds in quantum mechanics, electrodynamics, and general relativity into a single unified field theory that is (in theory) engineerable by higher symmetry electrodynamics means such as Evans' O(3) electrodynamics.

Einstein's postulates: (1) All the laws of physics are equally valid in all inertial frames of reference, (2) the speed of light is the same to every inertial observer, and (3) the observable local effects of a gravitational field are indistinguishable from those arising from acceleration of the frame of reference.

The first is called the special relativity principle, the second is called the law of light propagation, and the third is called the equivalence principle. It is now known that postulates (1) and (2) are not independent of each other. Note that the third postulate implies that any local force is due to a "gravitational field" (acceleration of the frame of reference). It also implies that gravitational force has a mechanism, since quantum mechanics assigns a differential operator, operating upon a potential, to create every force. By extended inference, it can be seen that the third postulate also implies that there should exist a method to turn any type of force (field) into gravitational force (field).

Note that postulate 1 may be violated if two frames remain inertial but each local spacetime contains different hidden engines. So the laws of interaction physics may indeed vary for the same object placed in first one of the frames and then in the second. Engines are ongoing general relativity actions, even in inertial frames. The action may not be translation of the bulk object, but extensive changes to the interior of the object and its interior dynamics at any and all levels.

The Russian physicist Sakharov has strongly postulated that gravitation is not even a fundamental field of physics, of the kind used by Maxwell in his electrodynamics. Instead, in this view G-field is always comprised of, and made from, other fields and interactions upon them. This would seem to follow straightaway since change of any kind of energy in spacetime curves that spacetime, therefore producing a gravitational change. So such has always been implicit in general relativity; it just took Sakharov to formally propose it. See A. D. Sakharov, "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," *Soviet Physics Doklady,* 12(11), 1968, p. 1040-1041 [English translation.] Sakharov's article (in Russian) in *Dokl. Akad. Nauk SSSR,* Vol. 177, 1967, p. 70-71.

Einstein's postulates (second postulate extended): Let us do a little "tinkering" with Einstein's postulates. For the first one, we point out that in the new approach using dimensioning and subspaces, the notion of an inertial frame now has acquired additional complexity! The "gross translation" effects — which tends to be what physics ultimately focuses upon today — can remain the same, and yet vacuum engines can be buried up inside the inertial frame to act upon a mass in non-translating ways. Everything translation-wise can remain the same, but now matter can be transmuted and transformed, etc., still right there in an inertial frame. So when the infolded electrodynamics inside all fields, potentials, and waves are utilized, the notion of "inertial frame" has acquired a higher topology, so that one inertial frame may differ very substantially from another inertial frame, even if the two are not in translation motion with respect to each other. This of course moves one from special relativity into full general relativity, and even there into a slight extension. Special relativity can remain special relativity on the surface, and yet general relativity effects occur "within" the objects postulated to exist in this inertial frame. The "inertial frame" concept in the postulates is now just a sort of special "equilibrium case" of a higher topology "inertial frame" notion. Further, general relativity is extended from the notion of "just translation" to the notion of "internal engines and internal dynamics also". These engines can be created in mass, in fields and potentials and waves, and in the vacuum itself.

The second postulate now stands as a "special case" of a far more general second postulate. The conventional second postulate actually defines (and limits) "light" to bulk changes in the entire vacuum potential magnitude. It only applies to light signals which are "bulk upheaval" disturbances of the vacuum potential (spacetime), so to speak. The limitation to light speed is an envelope wave speed restriction. Specifically it need not apply to the "inner EM/GR realm" of longitudinal EM waves infolded inside the vacuum potential, other potentials, signal carriers, etc. Superluminal communication is now included in this "hidden variable" or longitudinal EM wave communication. The Fogal semiconductor is capable of infolding and outfolding EM signals, including producing the internal longitudinal EM waves, and during their infolded state these signals are permitted to move at superluminal velocity. This has been experimentally demonstrated by Fogal.

So the second postulate needs to be restated to include the present statement as a special case. Notice that the concepts "light" and "speed of light" have now taken on much richer, extended meanings, since there is "outfolded" light and "infolded" light, so to speak. Light now is not just the "external perturbation wave" in the ambient vacuum potential medium (in flat spacetime). Instead, it can also be an "internal perturbation longitudinal wave" inside a "net flat spacetime" which does not show on the "surface" at all. Indeed, it may move in the time dimension or in 3-space. It now includes the transmission of an extraordinary "vacuum engine" as well as the usual transmission of just an ordinary signal. The ultimate expression of this extension has unfortunately been the development of such "engines" as infolded robot systems for use in warfare.

These infolded vacuum engines can be visualized as very special sets of longitudinal EM waves and their dynamics, consistent with the Whittaker decomposition. Whittaker's U(1) symmetry electrodynamics decomposition— particularly when enriched in the higher symmetry O(3) electrodynamics — applies to the ambient vacuum as a potential (it is a potential since it has energy density). So the "subspace" inside spacetime, so to speak, is just a set of LWs and their dynamics, and that subspace itself is engineerable, just as is the "bulk curvature of spacetime" itself. Further, we may now consider that space itself is just a vast superhighway for such superluminal LWs and their dynamics. Mass is mostly empty space with only an occasional particle here and there and with fields, potentials, and waves in between. Hence huge masses such as the ocean and earth are actually superhighways for the direct passage of longitudinal EM waves, "engines", special robots, etc. The degree of "transparency" is a function of the degree of cleanness of the LWs — where by "clean" we mean "absence of transverse EM wave residues".

The third postulate has also been dramatically extended and requires revision. As presently stated, antigravity effects do not appear to be included in present physics heretofore. The "gravitational effects and external accelerated effects" notion in conventional physics are still founded purely on *ordinary bulk translation.* Now they need not be. For example, the system can appear to be an ordinary system, as judged by normal "external light" observation. Yet its internal vacuum engines can cause local violation of most so-called "laws of nature." The reason a "law of nature" can be violated is that one is now able to violate the premises upon which it is postulated, and which are required to hold if the law is to hold. Or another way to look at it: any conservation law is the result of a symmetry. Breaking that symmetry allows violation of that specific conservation law. All external symmetries can be broken internally, even if the overall symmetry still applies externally!

The infolded EM inside the "conventional" EM fields, waves, and potentials — and inside the vacuum potential and spacetime itself-dramatically changes the present theories of electromagnetism, quantum mechanics, and general relativity. Further, for the first time the Sachs unified field theory allows an *engineering* approach that simultaneously crosses, includes, and extends all three disciplines in the required manner. The approach can be placed into good engineering models based on the Sachs work, particularly using the O(3) electrodynamics subset spearheaded by Evans, and the models can be tested and verified or falsified. See **frame, reference;** and **superluminal communication.**

electric field: (See also discussion under **E-field).** The *conventional* definition is this: "1. a region in space in which a stationary electric charge experiences a force due to its charge. 2. the electric force per unit test charge." Quoted from *Dictionary of Science and Technology,* p. 720. That definition is flawed. Consider the "spatial aspects" of part 1. If an electric field is defined as a force per unit charge (per Part 2), it is not a region in massless space but is an effect of something existing in massless spacetime and interacting with the charged 3-spatial mass. The electric field *prior to interaction and observation of the interaction* is indeed related to a region in "spacetime", because it is identically a curved spacetime region. In the Sachs approach, general relativity and electrodynamics have become one and the same. So prior to interaction, the *field as it independently exists prior to interaction with charge* has become a curved spacetime region, and the field-in-space is *purely* a spacetime curvature. Again, there is no such thing as a massless force. It is the hoary old mechanics notion of a separate force acting upon a separate mass that is a total non sequitur. The interaction of the massless field in space with a mass to produce a force, *of which that mass is actually a component,* is a separate consideration from the field as it exists in space prior to interaction. The result of this entity's interaction with charged mass has nothing to do with *what an electric field* is but with *what it does to charged mass.* Further, the force evidenced on a charged particle is not due to its charge; the particle has that charge always, whether or not there is a net force evidenced on it!

Part 2 is a little better. The only problem is, it totally disagrees with the notion of the field as an entity existing in spacetime prior to interaction with charged mass. In other words, here it's a "force on a charge", which is an effect (force consisting partially of the charge) occurring after the interaction. Again, the conventional definition states only what the field *does or accomplishes,* not what it *is.* Hopefully the reader can now begin to see what terrible difficulties await anyone who seriously attempts *foundations definitions.* Feynman sums it up admirably in one of his famous statements: "*Everything we know is only some kind of approximation, therefore, things must he learned only to be unlearned again or, more likely, corrected.* "An additional difficulty is that the very notion of a separate massless force — independent of the mass on which it operates - is fictitious, yet continues to be widely used. Recalling that any true definition is an identity and not an equation, let us start with the definition of force: $\mathbf{F} \equiv d/dt(m\mathbf{v})$. As can be seen, force is a system comprised of (i) a mass component and (ii) a nonmass component. The force is not something separate from the mass, at all! Further, force is not even unitary. It's a system comprised of two coupled components; the mass and a coupled (interacting) gradient in a potential flux. *Note that v involves motion through, and interaction with, the virtual flux comprising the vacuum potential. Note that dv/dt involves a gradient in that flux interaction with the mass.* Unfortunately mechanics is still using the medieval concept of a "separate force acting upon a separate mass." That is wrong, unless we change the fundamental definition of force itself. See, e.g, Robert Bruce Lindsay and Henry Margenau, *Foundations of Physics, Dover P*ublications, New York, 1963, p. 283, where the authors emphasize that a "field of force" at any point is actually defined only for the case when a unit mass is present at that point. The reason can be seen: mass is a necessary component of force, and without mass present there cannot be a force present. For that reason, the force field in matter is a quite different beast from that "same" force field in empty spacetime. The first is a 3-spatial snapshot of an ongoing 4-spatial interaction, while the second is just a curved region of that "empty" spacetime without any ongoing reaction necessarily.

electric dipole: See **dipole.**

electrodynamicist: A specialist in electrodynamics — the physics of the relationships between electric, magnetic, and mechanical phenomena — and especially a specialist in electrodynamics *theory.* Essentially all electrodynamicists are versed in U(1) electrodynamics, and a much smaller number are versed in more advanced (higher group symmetry) electrodynamics such as O(3) symmetry electrodynamics, unified field theory electrodynamics, non-Abelian electrodynamics, etc. The electrodynamics field is actually a very vast field indeed, touching everything in physics, chemistry, etc. When one needs a particular electrodynamics problem solved, it often calls for a special category of electrodynamicist and not just an electrodynamicist in general. One must therefore sometimes examine the background of the electrodynamicist making a statement or a pronouncement; electrodynamicists after all do have differing views and differing disciplines within the field. And something may be perfectly true and acceptable in O(3) electrodynamics, yet not even exist in U(1) electrodynamics. As an example, the Aharonov-Bohm effect exists in O(3), but not in U(1). U(1) assumes a flat local spacetime (falsified by general relativity almost a century ago) and an inert vacuum (strongly falsified by modern particle physics and quantum mechanics), while O(3) is perfectly content to include and model curved spacetime and active vacuum also.

electromotive force (EMF): A difference of potential created by sources of electrical energy — i.e., by separated charges (dipoles) — which can be utilized to drive electrical currents through external circuits. As a difference of potential, that constitutes a gradient and hence a force. Also, open circuit voltage, or the limit of the potential difference between the source's terminals as the current in the external circuit approaches zero.

This term is poorly named and also quite confused in the standard theory and usage. First, em/is supposedly not a *force* at all. It is a difference in potential; hence it is measured in volts. It is very difficult to imagine a "difference in potential" at a single point unless one is dealing with a multiply connected space. So if the difference in potential is between two adjacent points, then that constitutes a force between the two points. Also, emf is not considered an E-field, since that is a field of force. Yet it is in fact a $\nabla\phi$, restricted to the direction between the two measurement points, and $-\nabla\phi$ is considered to be an E-field in the EM theory! As can be seen, one will not find a truly logical definition of potential or emf in the present texts and science literature.

For this reason the standard theorists say that the term emf is sometimes used as equivalent to a difference in potential, but that strictly speaking it should be applied only to a source of electrical energy. The problem is that most conventional theorists are unaware that a potential *identically is* a bidirectional flow of energy in 3-space, but a 3-space flow where *the* energy *enters from the time-domain (complex plane) and is caught in 3-space by interaction with charge, in a frozen snapshot (observation). The energy actually enters each 3-space point in the 4-space potential, from* the *time domain.* It moves across a dipole in space (present due to the quantum mechanical polarization of the vacuum), to the opposite charge, thus returning from the immediate region of that second 3-space point (or points) back to the time domain. Since conventional theorists are unaware that a potential is a hidden negentropic energy flow circulation process involving broken 3-symmetry, broken t-symmetry, and a special 4-symmetry between t and 3-space, the theorists have been unable to consider the potential in general as a direct source of energy flow. Yet they are forced to consider a potential difference across the terminals of a source as "somehow a flow of energy, that generates forces upon the electrons to propel them through the circuit as current flow."

In the new viewpoint, a resolution is achieved of these incongruities by more accurately defining the E-field with respect to the potential as $\mathbf{E} = -[\nabla\phi \bullet q]/|q|$. It is noted that $-[dA/dt] \bullet q/|q|$ also produces an E-field, so that the full definition of the E-field $\mathbf{E}$ is $-[(\nabla\phi + d\mathbf{A}/dt) \bullet q]/|q|$. Note that dividing q by |q| does not remove the q unit from the corresponding dimensional equation, since |q| is dimensionless. However, one volt represents one joule (of excess energy) *diverged from the potential,* per coulomb of interacting charge. *The energy density flow in an electric circuit is in the voltage, in the emf, and it is not in the current per se.* The electron current (drift current) carries the excess energy that is being dissipated as work in the circuit, but in a nominal circuit the drift current only moves a few inches per hour. Half the energy collected in the potentialized circuit is dissipated in the external circuit and external losses, and half is dissipated in forcing spent electrons in the ground return line back up through the back-emf of the source, performing work inside the source to dissipate its separation of charges. Left alone, any dipole will furnish emf indefinitely, freely receiving the energy from the time domain (complex plane) and outputting it in 3-space via the reinterpreted Whittaker 1903 decomposition, in a circulation where at each point in it the energy enters from the time domain and thence returns back to the time domain. The free flow of EM energy will continue unabated from the undestroyed dipole, regardless of how many external loads and losses are powered (how much energy is collected and dissipated from the ongoing flows representing the potential). So *electrical circuits should be developed which allow few or no electrons to be forced back up through the source dipole to dissipate its separation of charges and destroy the dipole.*

electromotive force, back (back emf): Back emf is a loose term for the *counter-electromotive force.*

When the source is powering a circuit or circuit element, the term "back emf is used to refer to the forcing of current (electrons) back up through the source against its emf, thereby performing work inside the source to *dissipate* its separation of charges (its dipolarity). When a potential source of electrical energy — such as a capacitor, an inductor, or a rotating machine — is receiving energy from a source and converting it to *dissipative* power and work, that "powered" device is said to develop a *counter-emf.* A source has a built-in emf *a priori,* and this emf is holding its charges separated so that it is dipolar. When electron current is forced backwards through the source against its emf, work is performed against the internal resistance of the source to dissipate its separation of charges, thereby *dissipating* the source dipole. If electron current is not allowed to flow back up through the source against its back emf and through its built-in counter-emf, no dissipation of the source dipole occurs, even though energy density (not *electron current* dq/dt but *energy flow* dq/dt) is flowing from the source and can be collected in *collectors* in the external circuit to store the energy as a created local energy source. This collected energy can then be shuttled to a separate load circuit and released, to provide a "heat pump" cycling operation and overunity coefficient of performance.

Nikola Tesla in fact knew how to shuttle energy at will in a circuit in such fashion, as Barrett's quaternion analysis of Tesla's actual patented circuits has revealed. [See T.W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Annales de la Fondation Louis de Broglie,* 16(1), 1991, p. 23-41.] Note that a rigorous analysis of those same circuits, using tensors or vectors (lower topology algebras), will not show the effect. One

needs a higher-topology algebra such as quaternions or Clifford algebra, or **the** significant properties of such circuits will be entirely missed.

When the original source dipole is not dissipated because electron current is **not** passed back up through its back-emf to scatter the charges, the o**riginal** source dipole will continue to furnish emf indefinitely because of **the** proven asymmetry of its opposite charges in the seething vacuum energy flux.

The terms emf and counter-emf are normally applied only to recognized sources of electricity. Emf is computed as the algebraic sum of the **potential** differences acting in a circuit attached to a source. Back-emf (counter-emf) is computed as the potential difference internally through the source, between its terminals. The back-emf is thought to be necessary in producing *a powered device or component in the circuit.* Actually, *its* ***main*** *application (the ubiquitous closed current loop where all the external circuit current is passed back through the back-emf of the source dipole) is to prevent COP>1.0 electrical power systems from being developed and utilized, andfor every electrical circuit to self-enforce Lorentz symmetrical regauging during its de-excitation phase.*

elect**ron:** A stable elementary particle in all atoms, and having — in "f**orward** time" — a negative charge of 1.602x 10^{-19} coulombs, a spin 1/2, *and a* mass of 9.11x10^{-31} kilograms. If time is reversed, the charge (but not the masss) of the electron is reversed and it becomes a positron. Note this is an equival**ent** *observed* 3-positron, in a material lattice after interaction with charged matter, and not the mathematical equivalent represented by Dirac sea holes in the 4-vacuum prior to interaction with mass.

The "electron" may also exist as a negative energy, negative charge, negative mass-energy 4-state (Dirac hole) in the vacuum itself. In that state, the negative energy 4-electron is the source of negative energy fields and negative energy potentials. When the associated nondiverged and unaccounted Heaviside energy flow component is also accounted, the Dirac sea hole negative energy 4-electron can be manipulated in "anti-circuits" to directly engineer local antigravity, quite strongly and practically.

electron gas: A system of electrons with only weak interactions, so that the electrons may be regarded as moving independently. In many ways, such a collection of electrons can be treated as a gas. See Drude Electron gas.

electrostatic scalar potential: A *static* (stationary, in the sense that a perfect waterfall would appear stationary) ordering in the virtual photon flux of vacuum. The notion of "static" in the term "electrostatic" is misleading and should be replaced by "stationary" since the underlying active vacuum medium is always in violent motion, and there is nothing truly "static" in the universe. But an *equilibrium* or *stationary* condition can exist as an external envelope condition, even with continuous violent motion internally, so that a stationary entity or form persists. An example is a perfect whirlpool in a river.

The electrostatic scalar potential decomposes into a harmonic set of bidirectional phase conjugate longitudinal EM wavepairs, as shown by E. T. Whittaker in 1903. Hence the electrostatic scalar potential is an overall curvature of spacetime with internal longitudinal EM order and dynamics.

ELF: Extremely low frequency. Used to describe EM fields whose frequencies lie in the range from just above 0 Hz (direct current) to 300 Hz. This includes power line frequencies (60 Hz in U.S. and 50 Hz in Europe) and frequencies used by certain U.S. military ELF communication systems such as for submerged submarines.

EMF: Acronym for electromotive force. See discussion under electromotive force.

EM field: Electromagnetic field. In its loose usage, it refers in a very broad way to any field, force, or energy associated with electromagnetic interactions, charges and currents. Thus "EM field" includes electrostatic fields, magnetostatic fields, electromagnetic fields (including radiation and induction), vector-potential and scalar-potential fields, Hertz potentials, Fitzgerald potentials, Whittaker potentials, Debye potentials, etc.

In its precise usage, "EM field" conventionally refers to an EM "force" field, such as the E-field, D-field, B-field, and H-field. In classical electromagnetics (CEM), as a holdover from the old assumption of a material ether, the force fields are considered (erroneously) to exist in mass-free space, and they are also considered to be the primary causes of electromagnetic phenomena. This is a non sequitur, and it involves substituting the effect for the cause.

In the new viewpoint, *the force* fields do not exist in vacuum, but only in and on and of the charged particles themselves, where mass is present to form one component of the force. The force fields are therefore effects and not primary causes. The primary causes of all electromagnetic phenomena is taken to be the potentials; however, the hidden

Stoney/Whittaker/Ziolkowski internal structures of gradient-free scalar potentials may interfere to cause potential gradients in and on charged particle systems. [In O(3) and other higher group symmetry electrodynamics, this internal structuring and dynamics is far richer.] This therefore includes distance-free scalar potential interferometry and action at a distance, as well as vacuum engines.

The new view uses Sachs' unified field theory, and considers that the EM field in the vacuum, or an EM potential in the vacuum, involves and identically is curvature of spacetime. It follows that the internal EM structures of "ordinary" EM are also spacetime curvatures, where the specific pattern of these ST curvatures constitutes a very special kind of spacetime engine that acts upon any local mass in a specific manner determined by the individual pattern of curvatures. Thus the new view is a unified field theory where GR has become EM and EM has become GR, and where higher symmetry EM — such as O(3) group symmetry EM — can involve or produce gravitational, temporal, and inertial effects as well.

EM force field in the vacuum (a false concept : The erroneous notion that there exist forces in the vacuum, so that the set of forces at a set of spatial points forms a force field. Presently classical electrodynamicists accept the same E-field in vacuum as exists upon a charged mass when the vacuum E-field interacts with that charge, but it is asserted that the force is zero. This is easily falsified by defining force F as $\mathbf{F} \equiv \partial \boldsymbol{p}/\partial t = \partial/\partial t(\mathrm{mv})$. By expanding the term on the right, one sees that both terms in the expansion include mass. Mass is thus a *component* of force. If mass is absent, there is no force present, thus no force field present. The field in space and the "same" field in mass differ precisely by a mass term in their fundamental units. The field in massless space may be taken as a curvature of spacetime. Zero field in space means a flat spacetime; a nonzero field in space represents a curvature of spacetime. A nonzero field in mass means a curvature of spacetime (a 4-space entity) interacting with mass (a previous 3-space observation) to produce a force consisting of that field and mass interaction. There is no separate force acting on a separate mass, and mechanics has been wrong in assuming on for hundreds of years. Instead, there is a force produced by the interaction of a massless 4-field (spacetime curvature) entity with 3-spatial mass. It is that *interacting and ongoing entity or process* that is a force field. Speaking of such a force field in space, is speaking of the force field that would exist in that space if every point had a unit point static charge and charged mass. As Feynman and Wheeler pointed out, no force field as such exists in space, but only the potential for a force field exists there, in case some charged mass is brought in to interact and have the forces developed upon its charges.

EM generatrix: Whatever generates electromagnetics, particularly the fields. Since all EM energy comes from the time domain and returns to it, time is the ultimate EM generatrix. The circulation of EM energy from the time domain to 3-space and back, or vice versa, is a function performed by the broken symmetry of every charge, considered as two infinite dipolarities with a finite difference that is the observed charge listed in texts and handbooks.

energetics: Energy as the primary thing in the universe. Rankine's term later raised to the forefront by Ostwald. Modern Russian term for advanced higher group symmetry electromagnetics, where the electrodynamics is a subset of an engineerable unified field theory. In the U.S., a model approaching the Russian model is given by Evans's O(3) electrodynamics as a special subset of the electrodynamics used by Sachs in his epochal unified field theory. Thus there is a growing, engineerable unified field theory (energetics theory) in the U.S., being provided by Evans, Sachs, and others. A solid basis for energetics can be realized from the giant 4-space negentropy EM energy circulation involved with every charge, so that all "EM energy in space" and time itself comes from that circulation.

energy: Usually erroneously "defined" as the "capacity to do work". This is a total misdefinition since it would define energy as the ability to change its own form. The closest to a legitimate definition of energy appears to be: *energy is the invariance of that which is varying* (Lindsay). However, note that this definition violates Aristotelian logic (but does not violate five-law logic developed in this book). Another try at it might be: *"Energy is any deterministic or coherent structuring, either dynamic or stationary, existing in the virtual particle flux of vacuum."* While that sounds good, it still involves the nonchanging aspect of the changing flux — much like considering a whirlpool in a river as being a "static" thing. So it is still "the invariance of that which is varying", which is an identity of opposites. We conclude that 4- or 5-law logic is necessary in order to define energy, since the defining logic must encompass the identity of opposites. Feynman simply stated: *"It is important to realize that in physics today, we have no knowledge of what energy is_."* [Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics*, Addison-Wesley, Reading, MA, Vol. 1, 1964, p. 4-2.]. In Aristotelian logic, there is no acceptable definition of energy. In 5-law logic, the second definition holds since opposites can indeed be identical. Now to more practical aspects:

For our purposes, there are three major types of energy in 4-space: (i) 3-spatial energy, which includes energy in one, two, or three of the spatial dimensions, (ii) mass-energy, which is 3-spatial energy compressed by the factor c^2 but still in 3-space, and (iii) time-energy, which is 3-spatial energy compressed by the factor c^2 but now placed in the time domain. **Note t**hat time-energy has the same density as mass-energy, and hence the same spacetime-curvature effect and gravitational capability.

The major implication of the new view is that the conservation of energy law is extended: Now it is the total of the 3-spatial energy, mass-energy, and time-energy that must be conserved. Transduction between time-energy and either of the other two forms, or both, will violate the conventional conservation of the sum of the 3-spatial energy and the mass-energy; every charge and every dipole is an example. Use of time-polarized and longitudinally polarized photons and EM waves particularly may involve the extended conservation law, e.g., as utilized in cold fusion and its lime reversal zones, and as involved in the outpouring of observable EM energy from every charge in the universe, with no observable energy input.

energy collection/collecting: In electromagnetics, placement of an interceptor/collector — e.g., a charged particle — into an impinging EM energy flow, so that some of the flow is diverted around the intercepting particle. We may say that the flow of energy has been "reordered" into a stream diverged around the intercepting charge. The diverged part of the total energy flow is what is calculated as the Poynting energy flow component, and is what is referred to as the "collected" energy or the "potential" of the charge.

However, neither charge nor potential are acceptably defined in conventional electrodynamics, but they can be defined in the new approach. We are now able to see something more fundamental. By the giant negentropy mechanism, the charge is involved in both the time-ene**rgy** stream and the vacuum flux set of virtual particle spatial-energy stream**s.** "Charge" is not just related to the mass of the particle, but is the entire ongoing set of the dynamics as well as the mass (the "nondynamics", so to speak). When a charge is placed in a field, that field is a change to the local dynamics, which is a change to the charge itself. Both the dynamics and the mass (the "nondynamics" constituting the charge are altered. So one does not have the "field acting on a separate charge", but instead we have the charge with its previous set of dynamics transformed into the altered charge with a new set of dynamics and a new mass. Similary when a charge is placed in a potential. Charge is also not a 3-

spatial situation and process, but a 4-spatial situation and process. If the internal structuring is also important, it becomes an n-spatial situation where n>4 and even where n » 4. To deal with sophisticated engines, n » 4 is necessary in the modeling utilized.

energy dissipation: Refers to the scattering or disordering of energy; at fundamental basis, it refers to the changing of an altered charge (see energy collection/collecting) back to an unaltered charge. In other words, it is a special decay process whereby an "excited charge" (so to speak) decays back to its "normal state" by releasing dynamics (energy) back to the vacuum. The energy itself is not destroyed, but its "collecting pattern" or "flow pattern" as coupled to the charge is disrupted and altered or destroyed There is also the connotation that *the dissipated energy is lost as far as any re-use of it by the system.* We point out that only collected/collecting energy can be scattered; it is the collecting *process* (the entire alteration of the normal charge dynamics) that is "destroyed" - or more exactly, *transformed* — not the energy itself.

energy, electromagnetic: From a vacuum and quantum mechanical viewpoint: a deterministic or coherent structuring, either dynamic or static, existing in the virtual-photon or charged-particle flux of vacuum. Since a photon may also be considered a positron-electron pair (if we multiply each by time so that they are qt rather than q in fundamental unit form), the flux of photons in the polarized vacuum necessarily also includes the flux of charged virtual electrons and positrons. From a spacetime viewpoint: a curvature of spacetime, or set of such curvatures. If a deterministic set of spacetime curvatures (a template) is involved, then an engine is involved and the energy is said to be "dimensioned" (internally structured).

energy flow: Generally considered as the passage of energy, in any form, in one direction or generally in one direction in a given 3-space. To first order, our definitions of energy treat energy flow in a given space as (1) a directional change (propagation) of the vacuum energy and its form, or (2) as a directional propagation of a curvature of spacetime or a set of curvatures of spacetime. *The terms "energy flow" and "flow of energy through space" are not precisely defined in physics! They are not going* ***to*** *be defined, at least not in ordinary Aristotelian logic terms, since energy itself is not precisely defined, and one of the best struggles with the concept of energy concludes that it involves invariance in the midst of change, which is a violation of the third law of Aristotelian logic and therefore of the other two as well.*

Indeed, the "EM wave flowing through 3-space" does not exist (see previous comments quoted from Romer). What does exist is an ongoing 4-space process, assumed to be continually and iteratively interacting with a unit point charge to "detect" or "observe" it as an iterative series of frozen 3-space snapshots, in memory-recalled serial order much like the frames of a motion picture film. Nonetheless, we must keep our sense of humor and proceed as if we knew exactly what we were speaking of when we use the concepts "energy", "energy flow", and "energy propagation through 3-space". If one dwells on the problem without humor, one is led to Hawking's grim but perceptive statement: *"All we ever know is our models,* but *never the reality that may or may not exist behind the models and casts its shadow upon us who are embedded inside it. We imagine and intuit, then point the finger and wait to see which suspect for truth turns and runs. Our models may get closer and closer, but we will never reach direct perception of reality's thing-in-itself"* Ultimately we do not know physical reality itself, but our mental modeling of it, as best we can "fit" our modeling through comparison with our sensory and perceptive sensations.

energy flux: The passage **of** energy **flow, in any form, through a unit area, and** in any given direction (usually all directions). This includes both passage in a single direction, or passage in multiple directions at once. See flux. Sometimes taken as that component of the net flux through a unit area at right angles to the area.

energy, free: In our usage, free energy is what is obtained from any environment or entity which freely emits, furnishes, or makes available energy which either can be collected and used as furnished, or can be transduced into usable form and collected and used to power loads and do useful work. The word "free" does not apply to *costs,* but rather to the source *of the energy.* Since energy can neither be created nor destroyed, energy is always available free for the taking (collecting and using) if we bit learn how. *We never pay for energy per se, but for the manipulation and processing of the energy.* The notion of extracting and using free EM energy from the vacuum involves five key concepts: (1) the local vacuum/spacetime itself is filled with EM energy of generally unusable form (2) some sort of organizing principle is utilized in one part of the circuit or system to cohere the vacuum energy and transduce it into usable EM form (every charge already individually does that; we just have to learn to do it in a larger ensemble), (3) a gating mechanism is utilized to divert the vacuum energy now in usable EM form from the active vacuum to the external circuits or load parts of a device, (4) the coherent gated extra energy available in the external circuits or load parts is intercepted by those circuits or load parts, couples to them, and potentializes them with extra EM energy, so that the energy is collected by the system and regauges the system asymmetrically, and (5) the external circuits or load parts then scatter or dissipate the coupled extra energy to produce useful work (physical changes in the load), without destroying the organizing and gating mechanism that is extracting and furnishing usable EM energy from the vacuum.

energy source: A function or entity which furnishes energy by either gating it from one energy reservoir to our system or locality, or by transducing available unusable energy into usable form and then emitting the usable energy. Technically speaking, since energy cannot be created or destroyed, there is really no such thing as a *creative* "energy source", though the term "source" seems to imply such. So-called "energy sources" actually gate and/or transduce energy, changing it from one form to another, from one direction to another, etc. All EM energy in the universe comes from the time domain into 3-space vacuum, onto charge, and back to the time-domain, in a giant ongoing negentropy 4-circulation as freely provided by the "source charges".

engine: Deterministic pattern of structures, with their concomitant dynamics, in the flux of the vacuum. Also, deterministic patterning of curved spacetimc curvatures and their dynamics that comprise an overall larger spacetime curvature, or internally comprising an *overall* "uncurved" spacetime.

engineering, entropic: Engineering or designing systems and processes in equilibrium with their external active environment, and whose operation "spills" or wastes all the energy input to them, and which do not freely recover and use the dissipated or wasted energy. Entropic systems — which are the products of entropic engineering — thus have COP<1.0 and function in forward-time, and classical equilibrium thermodynamics with its infamous second law rigorously applies because the system is designed to self-enforce the equilibrium condition, often called by other names such as Lorentz symmetrical regauging.

engineering, negentropic: Engineering or designing open systems and processes far from equilibrium in their exchange with their active environment, whose operation freely "receives" excess energy from the environment, so that they permissibly output more energy than the operator inputs, the excess being received from the external environment. As an example, simply making a common dipole or assembling a charge initiates a giant negentropy operation that persists indefinitely (as long as the dipole or charge persists).

entropy: Thermodynamics state term representing a measure of an equilibrium system's capacity to undergo spontaneous change. In short, a quantity that measures the availability of a system's energy for conversion to work. The entropy of an open system not in equilibrium with its active environment cannot even be computed, because the state cannot be computed.

Entropy is given by dS = dQ/T, where dS is an infinitesimal change in entropy for a system absorbing an infinitesimal quantity of heat dQ at absolute temperature T. In statistical mechanics, entropy is a measure of the disorder in a system and is given by S = k lnP + C, where S is the value of entropy (disorder) for a system in a given state, P is the probability of that state's occurrence, k is a fixed constant, and C is an arbitrary constant. Intuitively, the concept of entropy corresponds to the concept of disorder. We specifically call attention to the fact that the "S" used in representing entropy is not at all the S used for Poynting energy flow in electrodynamics.

entropy, negative (negentropy): In a sense, the reversal of disorder, or reversal of entropy. Another way of defining negentropy even more rigorously — for electrodynamics — is to define it in spacetime as a bro*ken 3-symmetry in EM energy flow with concomitant relaxation to a more primary 4-symmetry between time-energy flow and 3-spatial EM energy flow, where time-energy flow is transduced into 3-spatial EM energyflow and then back to time-energy by the entity — such as a common dipole — causing the broken 3-symmetry.*

equilibrium: *In physics,* the state of a system in which opposing forces balance each other — i.e., sum to a vector zero. However, the *vector* zero summation does not represent "total absence" of everything, and the forces and their energy remain even though comprising a net vector zero system. Even a pure number zero may also be taken as a set of nonzero numbers; e.g., (+5 -3 -2). So even zero need not be total absence, but can be a "balanced set of presences". Or zero can be regarded as an overall equilibrium condition comprised of a set of disequilibria conditions.

Indeed, a vector zero summation system of nonzero forces comprises a stress potential, and it contains stress-energy doing work upon the stressed system. In *mechanics,* a system of particles — where the net force acting upon each constituent particle is zero — is said to be in equilibrium. *In chemistry, t*he state of a solution where the forward and reverse reactions are equal and balance each other, so that the chemical reactants and products maintain the same amounts. Again, the notion of equilibrium involves the "stationary" notion (similar to a perfect whirlpool in a perfect river) of the overall invariance of some set of ongoing dynamic changes.

ether: An extremely fine material substance originally theorized to exist in space in lieu of sheer emptiness. The problem is considering space as emptiness in the first place, which it is not. Instead, space (more rigorously, spacetime) is a plenum and highly active (as in the general relativity view a half-century after Maxwell's seminal paper). So there is no need to "fill empty space" with something, because in the modern view there is no such thing as empty space.

Evans, Myron W.: Noted chemical physicist, primary developer of O(3) symmetry electrodynamics, and Director of the Alpha Foundation's Institute for Advanced Study (AIAS). Papers produced by the AIAS are critical to understanding extraction of EM energy from the active vacuum, engineering of vacuum engines, and development of a revolutionary new healing method by time-reversing diseased or damaged cells back to a previous healthy state.

excess collected energy: Collection by an interceptor/collector of a greater fraction of EM energy flow impinging upon it and also surrounding it, than is normal. See **Bohren experiment.**

exciton: A bound electron-hole pair in a semiconductor. Excitons are temporary formations and they decay by recombining, but they have a specific half-life before recombining during which they migrate through the semiconductor crystal. When the hole and electron recombine, then — if the curvature of spacetime relaxes to a flat spacetime, which is not overtly stated in classical EM theory — the recombination releases energy in the form of a *photon,* or sometimes in the form of several *phonons.* **Note** that the "hole" is an "absence of an electron" from a position in a lattice usually containing an electron, so that it exhibits a net positive "change in the charge field" at that point. Specifically, this "lattice electron charge hole" is not a Dirac Sea hole in the vacuum, which is an entirely different matter. On the other hand, if the hole and electron recombine and the curvature of spacetime does not relax to a flat spacetime *because of other curvature-sustaining involvements,* then no "pair annihilation" radiation is emitted. It is the relaxation of the curved spacetime (when hole and electron are separated) back to a flat spacetime that causes photon or phonon emission.

exothermic: Literally, "giving off heat", or emitting scattered EM energy. Producing time-forward EM energy from a "source" by reducing the ordering (internal energy) of the source.

Faraday, Michael: Noted English physicist and chemist, 1791-1867, who discovered field theory and electromagnetic induction. Faraday also invented the dynamo and formulated his laws of electrolysis. His work directly inspired Maxwell, who vowed to study no other electromagnetics theory until he had captured Faraday's work in a mathematical theory.

Ferminion: A fundamental particle which has spin angular momentum of (n+1/2)h, and which obeys Fermi-Dirac statistics. Examples are baryons and leptons, which are subject to the Pauli exclusion principle.

Feynman, Richard P.: Noted physicist, Nobelist, and co-developer of quantum electrodynamics. Feynman is particularly noted for his refreshing three volumes of physics: see Richard P. Feynman, Robert B. Leighton and Matthew Sands, *The Feynman Lectures on Physics,* Addison-Wesley, New York, 1963. The book is particularly noteworthy in that it gives the student some insight into foundations problems in physics. E.g., in Vol. I on p. 2-4, Feynman defines the electric field not in terms of force per unit charge per se, but in terms of its *potentiality* for producing a force only when a charge is present for the force to be developed upon.

fiber fuse: Effect in optical fibers or fiber optic cable, whose core contains germanium, whereby heating the cable at one spot with a small flame will launch a focused traveling heat energy effect which melts a small hole in the core about every centimeter, and marches down the cable at about a meter per second to the end of the cable, destroying the optical transmission capability and ruining the cable. Oddly, by igniting the fiber fuse from the other end of the damaged cable, a *reverse* fiber fuse effect will often march back down the cable in similar fashion except that it fills in the melted holes in the core, restoring the transmission capability of the cable. If the energy required to melt all the holes (or fill them again) is taken into account, the amount of energy required to launch the fiber fuse is much less. This phenomenon does not appear to have an adequate technical explanation at present.

field: In mathematics, a set with two binary operations. In **electromagnetics** in more common mathematical terms, it is a region consisting of magnitudes and/or vector directions assigned to some active entity (e.g., spacetime itself) at each point in the region.

In physics, "a region of space marked by a physical property, as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region." [Thanks to *Webster's II New Riverside University Dictionary.]* Note the very special use of calculation of magnitude: one does not calculate the magnitude of the field itself, which fills an entire region. Instead, one calculates the magnitude of the *reaction* of that field with some interceptor/reactor at a given point — i.e., the *effect* of that field upon some mass entity. That is at best a measure of the *local intensity* of the field, not the magnitude of the entire field per se. For that reason, often we speak of calculating the field intensity, which is far more accurate than saying that we are "calculating the magnitude of the field". So we do not really calculate the "magnitude of the field," and in fact such a calculation does not occur in any textbook. Instead, we calculate the *reaction cross section* — in the case of electrodynamics, for an assumed unit point static charge — of the field at a point in the region or space occupied by the field. *We actually calculate what is deviated or diverged from the field, around an assumed interceptor, not the field itself.* This means that we calculate the result or *effect* of a detecting/observing action, and a frozen 3-space snapshot of that ongoing interaction, since all observation is 3-spatial. To then state that this same field exists in space prior to the interaction, and is what causes the interaction, is a gross non sequitur and confuses the effect with its own cause. This ubiquitous non sequitur widely permeates electrodynamics and mechanics, and it is one of the greatest scientific logical errors in history.

field, scalar: In the standard view, a field, to any point of which is assigned a magnitude only. However, "magnitude" of the EM field in the standard usage refers only to the amount of energy deviated (diverged) around a fixed intercepting unit point charge, per the conventional "detection by reaction" model. Again, note the serious logical errors in the conventional usage. One cannot define the field in terms of what is diverged from it! At best one has defined the "intensity" of the field's interaction with a given standard interactant (charge). So in electrodynamics the term "scalar field" would be better replaced with "scalar field intensity", at least for many applications.

field, vector: In the conventional view, a field, to any point of which is assigned both a magnitude and a direction. Again, the standard view also implements that assumed reaction with a detecting unit point static charge, and the "magnitude" used is just the difference between the amount of energy piled on the "front" of the interacting unit point static charge and the amount of energy pile on the "rear" of the interacting unit point charge.

The "direction" is from the center of the "high energy" pileup to the center of the "low energy " pileup around this fixed intercepting unit point charge. Again, note the serious logical errors. One cannot define the field in terms of what is diverged from it and the direction of the difference in "energy pressure." At best one has defined the "intensity" of the field's interaction with a given interactant (charge), and the "direction" established by the difference between the frontal and rear energy pileups on the interacting charge. The term would be better replaced with "vector field intensity".

flat (uncurved) spacetime: In our view, a vacuum/spacetime whose virtual particle flux intensity — and therefore its local energy density — does not change as a function of spatial position or averaged time. That is, in our view a flat spacetime is also identically a quantum mechanical vacuum whose virtual particle flux and energy density are constant in magnitude and internal constitution with respect to spatial position and in averaged time. Note that in this extended view, a flat spacetime (vacuum potential) can still contain vacuum engines, and thus still be *dimensioned. In that ca*se, contrary to normal physics, the flat spacetime itself will still interact with an embedded object and *change* it internally, even though it does not *translate* it externally. To perform transmutation of elements, e.g., one is interested in just such engines and not in spatially translating the element being treated. This dimensioning or activation of the local flat spacetime can result in an otherwise unexpected, seeming violation of one or more laws of nature in the dimensioned region, and easily results in violation of conventional physics which does not take into account engines and dimensioning of a flat spacetime. Note also that this allows a differentiation between the actions of two inertial frames, if the frames are differently dimensioned. As we stated, this is a deliberate extension of orthodox physics. See also dimensioning, and Einstein's second postulate —extension of.

fluctiation theorem: Theorem advanced by Denis J. Evans et al. at the Australian National University in 1993 that appreciable and measurable violations of the second law of thermodynamics occurs for small systems over short time scales. See D. J. Evans, E. g. D. Cohen, and G. P. Morriss, Phys. *Rev. Lett.* Vol. 71, 2401 (1993). In 2002 Evans and his colleagues experimentally demonstrated the proof of this theorem in rigorous fashion. See G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.* 89(5), 29 July 2002, 050601. The proof also directly proves the occurrence in electrolytes and fluids of time-reversal zones (TRZs) where the dynamics runs backwards for a short time. This validates Bearden's proposal of the TRZ and its consequent reversal of the Coulomb barrier to a Coulomb attractor as the mechanism producing attraction of like charges into quasi-nuclei, which then can decay by quark flipping (if necessary) as the TRZ decays and reverses back to normal. This provides a mechanism for transmutation of elements in these TRZs as a result of the fluctuation theorem and its proof by Evans et al.

flux: The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time. The term flux density is now more commonly used. Loosely, "flux" refers to the perpendicular flow through an area, while "flux density" refers to the magnitude or intensity (rate) of the flux. In some cases such as a flux of particles of different velocities, the number of particles may be multiplied by the average velocity to give an average flux density that is weighted.

flux density: The amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time.

Fogal semiconductor: A special semiconductor invented by Bill (William) Fogal which, in its operational regime, utilizes an integrated semiconductor, very special tantalum capacitor, and feedback resistor to accomplish charge blocking (blocking of electron current flow dq/dt) while passing displacement current $d\phi/dt$.

The Fogal semiconductor can also accomplish amplified phase conjugation of signals as well as infolding (translation of input transverse EM waves to output longitudinal EM waves) and outfolding (translation of received input longitudinal EM waves to output transverse EM waves). Used in communication systems, it opens the use of the unlimited "infolded" electromagnetics bandwidth. Since it may communicate using longitudinal EM waves, it is also usable for superluminal or a specialized "tunneling" communication through the "interior" of normal EM waves, potentials, and fields. For example, it can easily transmit color TV signals through the "interior" of a common DC potential. See William J. Fogal, "High Gain, Low distortion, Faster Switching Transistor," U.S. Patent No. 5,196,809, Mar. 23, 1993; - - "High Gain, Low Distortion, Faster Switching Transistor," U.S. Patent No. 5,430,413, July 4, 1995, a continuation of his earlier patent.

Fogal, William: Researcher and inventor of the Fogal semiconductor (two patents so far), as well as a superluminal infolding-outfolding communication system and other remarkable longitudinal EM wave processes. The Fogal semiconductor allows the use of the internal longitudinal wave electrodynamics comprising all ordinary EM fields, waves, and potentials, and thus will usher in a revolutionary new communications science and extended electrodynamics technology. See also discussion under Einstein's second postulate extended.

force: In mechanics, identically the time derivative of momentum, or $\mathbf{F} \equiv \partial/\partial t(mv)$. One must realize that it is not the mass in motion, but the single undivided quantity "mass-motion" or "mass-velocity" (momentum) that is changed. Present EM theory erroneously takes force fields to exist in empty, massless vacuum but with zero force! However, the mechanical definition is an identity, not an equality. The observable mass in vacuum is zero, so substituting m = 0 gives $\mathbf{F} \equiv \partial/\partial t(0 \cdot v) = 0$. This appears to be the conventional reasoning. However, $\mathbf{E} = \mathbf{F}/q = \mathbf{0}/q = \mathbf{0}$ also. So if mass is zero, the field E used in "force field definition" disappears. All that says is that the field in mass-free space and the field in mass are totally different entities,

Since force can be induced into and of the mass-motion target at a distance, in the new approach there is a more fundamental mechanism that produces force itself, and that is a curvature of spacetime, in agreement with general relativity. Distant curvatures of spacetime to produce forces on exposed charged mass systems may also be engineered by longitudinal EM wave interferometry, which does constitute "action at a distance." The fields arise in the local spacetime in the distant interference zone, thus arising from within and around any physical mass object located in that zone. Hence it produces an "inside-to-outside" type of action that is presently unknown in orthodox physics. The produced fields can start inside the nucleons of the atoms and move outward, and work directly upon the quarks and gluons first. This opens a great number of entirely new nuclear reactions presently not developed in particle physics.

force **field:** An 3-space *effect* that exists as a local vacuum virtual particle flux interaction upon and with a hypothetical unit point mass or unit point charge assumed at every point in 3-space. The effect (force field) results from a ***causal*** force-free 4-field (a curvature of spacetime) interacting with an observed hypothetical unit point mass or unit point charge assumed at every point in space. Hence the invocation of observation invokes a *d/dt operator* (the interacting charged mass) upon the causal field, providing a frozen 3-snapshot of the ongoing 4-interaction. An iterative series of such 3-observations is said to constitute a "3-space force field" as if it existed in time, which it does not. It is seen (iteratively observed) as a series of such frozen snapshots, which mentally we conceive as "existing in time" instead of the "recurring at points in time" that it really is. We unconsciously invoke that substitution of iterative effect as being continuity in space and time, when we use terms such as "propagation through space".

force-free propagation: Propagation of force-free disturbances in the vacuum flux; i.e., propagation of curvatures of spacetime or sets of such curvatures, together with any associated dynamics.

foundations of mathematics: The concepts, operations, assumptions, postulates, logic, definitions, and axioms of algebra systems.

foundations of physics: The postulates, assumptions, concepts, and definitions upon which a scientific physics model rests. Included are also the postulates, concepts, assumptions, definitions, axioms, logic, and operations of the mathematics in which the scientific model is expressed, as well as the postulates, concepts, assumptions, definitions, and operations of the physics entities and functions assigned to the mathematical symbols. Thus "scientific truth" -— even mathematical truth; see Morris Kline's *Mathematics: The Loss of Certainty* — is actually a *qualified* truth. What is true in one model (one set of postulates, assumptions, concepts, and definitions, including those in the algebra) may not be always true when one or more of the postulates, assumptions, concepts, and definitions is/are changed, and/or when the topology of the mathematics in which the model is expressed is changed, etc.

four-space (4-space): The "normal" spacetime used in physics, consisting of three spatial dimensions and one time dimension. Minkowski spacetime.

fourth law of logic: The law that a thing and its opposite become identical on the boundary, i.e., in multiple simultaneous observations. *Identity* is merely the loss of observational or perceptual distinction; when formerly "two" objects are simultaneously observed by a single observer, and two slides have been shown at once in the slide projector. It does not matter at that point what each was individually observed to be before when it was singly observed. Now there is no observable distinction or difference. We previously showed that the fourth law is implicitly assumed in the three laws of Aristotelian logic anyway, and published a simple proof of the fourth law. Without the fourth law, Aristotelian logic "eats itself." As a simple example, the Aristotelian third law eliminates the center vertical line division of a simple Venn diagram between A and not-A, because the division line itself is happy to belong to both A and not-A simultaneously. So Aristotelian logic destroys the very Venn diagrams used to "prove" or "demonstrate" its logical propositions. What results in the four law logic is that either (1) the three laws apply explicitly and the fourth law applies implicitly, or (2) the fourth law applies explicitly and the three laws apply implicitly. This *application rule* — i.e., *either (1) or (2)* — itself can be taken as a fifth law of logic, thus extending and completing Aristotelian logic which is presently incomplete. The fourth law is also the law of the **paradox:** i.e., when something is true but violates one or more of the Aristotelian laws, it is a case where the fourth law applies explicitly and the three laws apply implicitly. Wave-particle duality is such a fourth law case.

four-wave mixing: A nonlinear multiwave mixing effect when four waves mix (interact) in a highly nonlinear situation so that wave-to-wave interaction occurs.

fractional charge: A hypothetical particle having a charge less than the electron's charge. Often called a quark, where its charge is 1/3 or 2/3 the charge of the electron. A few years ago, Stanford University researchers found what was believed to be evidence of free quarks, but that later was doubted. There is also now the fractional Hall effect. In the 1920's Ehrenhaft reported the discovery and production of fractional charges of all sizes, and his experiments have been replicated by Mikhailov, with papers by Mikhailov and Barrett.

In the new approach, when time-energy is utilized, quarks can be freed or nearly freed rather easily in matter inside "time-reversal zones", since the energy emerges from the local curved spacetime and propagates from within **the** nucleus outward and the gluon forces are substantially reversed. Note also that time is multiply connected, and energy can enter an object anywhere within it or even everywhere within it, since each and every differential piece of the object exists in every single increment of time. It would appear that, as the new approach develops, free quarks will become a normal event rather than one intensely sought but terribly difficult to come by experimentally. We flatly predict that, in a sufficiently good time-reversal zone, free quarks can be produced because of the enormous energy density (in the time component) being utilized.

frame, Lorentz: A frame of reference that is not accelerated with respect to the laboratory observer. In this case, the frame may be rotated (moving at a constant velocity) but it is not rotating (the frame is not accelerated with respect to the observer, and spacetime is not curved). In this case special relativity applies, as do the conservation laws. We do add a caution that full general relativity still applies to the interior structure and dynamics (the engine) comprising the flat spacetime. Two *overall* flat spacetimes may produce very different interactions inside matter, if their internal engines are materially different and dimensioned.

frame, reference: A spatial, organized, measured lattice placed in "emptiness" (space, spacetime). Normally refers to a 3-dimensional, spatial frame (simplest case) or to a 4-dimensional spacetime. All objects and points in the "universe" or spatial frame are considered to simultaneously coexist at separate, measured points in the frame. Differs from the vacuum in that, rigorously, vacuum has no existing definite lengths and no existing definite time intervals, as these appear only after measurement or detection, and are relative to the observer and to the interactions ongoing as well as in the detection process itself. The "laboratory frame" is the static reference frame of the observer or measurement. A separate reference frame may be assumed to exist for any fixed or moving object, or centered on any point in another frame. When a type of frame is assumed, the entire class of *translation type* physical interactions that can occur has been restricted to an assumed set or type. In other words, given the frame, the *conventional* physics of translation has been assumed. One of the greatest restrictions of an assumed "frame" is to rule out the consideration (existence) of other higher dimensions, hence of dimensioned engines.

In the new unified field theory approach, the other higher dimensions are always available and cannot be ruled out in general, but only in some special case. Every curvature of spacetime, and any internal additional curvatures comprising that primary curvature, *adds one or more new dimensions.* In our view, a spacetime may be "flat" in overall envelope curvature, but consist of internally structured deterministic curvatures or "engines". In this view, normal inertial frames, e.g., may still contain engines, which will not affect the normal *bulk translation* rules, but may affect any or all of the nontranslation mechanisms, including the very laws of nature in many cases. Engines have not been developed and used in conventional open physics, but were weaponized clandestinely by the former Soviet Union. See **Einstein's second postulate extended.**

free energy: Excess energy freely received from its external active environment by an open system that is far from thermodynamic equilibrium in its exchange with that environment. There is a separate thermodynamics for such open disequilibrium systems. (See **thermodynamics of open systems.** See also Appendix **A** for a more detailed discussion of the thermodynamics aspects). Such a system is permitted to exhibit five "magic" functions: it is permitted to (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator hims**elf** inputs (the excess energy being freely received from the active environment), (4) power itself and its load simultaneously (all the energy bein**g** freely received from the active environment), and (5) exhibit negentropy. Every charge and dipole in the universe already performs all five of those functions.

free **energy** system: Any open system not in thermodynamic equilibrium, which freely receives excess energy from an external environmental source, **and** collects and utilizes this free excess energy to freely power external loads, so that a permissible system COP>1.0 is achieved, as allowe**d** by the nonlinear thermodynamics of open systems in disequilibrium with their environment. (See **thermodynamics** of open systems.). A free energy system may also be "self-powering" if "close-looped" with governed positive energy feedback from output section to input section, meeting the Kron self-powering condition. Operation of free energy systems can only be properly modeled and understood in terms of their super*system* interactions.

Implicit in the notion of free electromagnetic energy systems are six key concepts: (1) the local vacuum/spacetime is known to be filled with hidden (i.e., virtual**)** energy in violent motion, i.e., the local vacuum/spacetime is a violent virtual energy flux and therefore a scalar potential, (2) a broken symmetry - i.e., some sort of organizing and gating mechanism, such as the common dipole — in this hidden (virtual) energy flux of the vacuum is utilized as a source, to cohere the virtual energy and divert it as energy $\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$, flow from the broken symmetry initiation point onto the external circuits or load parts of a device, (3) the coherent gated extra virtual energy flow available in the external circuits or load parts couples to (inter**acts** with) the mobile charges in those circuits or load parts, so that gradients developed in the interacting virtual flux form force fields on and of the particle masses, driving the mobile charges along the circuit and through current-impeding circuit components such as loads, (4) in the interaction of the driven charges with the driving virtual energy flux, their spin and erratic motions integrate a tiny "coupling fraction" of the driving virtual energy into observable (i.e., macroscopic field) energy, thus creating the **E**-fields and B-fields, (5) the current-impeding external components or load parts scatter or dissipate the coherence of the driven charges and thereby dissipate the collected energy or change the form of the co**llected fi**eld energy, thereby producing useful work, and (6) little or none of the excess collected energy in the circuit is split off and dissipated in the gate (asymmetry component serving as the S-flow source) to destroy the gate and thereby destroy its asymmetry in the vacuum flux, which is what is extracting and producing the free energy flow.

gate: A component or function that extracts some part of a flux or flow, compacting it into a stream, and "sending it out" in organized, useful energy flow form. The fundamental gate for all EM energy in 3-space is the dipole or dipolarity — the "isolated charge" is actually a dipolarity, when the virtual charges of opposite sign clustering around it in the active polarized vacuum are considered. The source of the 3-space EM energy continuously emitted by any charge or dipole is an equal inflow of EM energy from the complex plane (from the time domain). Thus EM energy flow is locally conserved in four dimensions as a 4-circulation, but not in 3-space. EM energy flow in 3-spacc is conserved over all space, since there are essentially equal amounts of positive and negative charge in the universe and in the virtual charge of the vacuum.

gauge: In electromagnetism, usually considered a possible choice for (change of) electric scalar potential and magnetic vector potentials, in such a manner that there is no any *net* change of force fields so that the change satisfies Maxwell's equations. There are indeed two changes of force fields, but the two are selected so that the two new force fields are equal and opposite, constituting a stress potential and continuously performing internal work on the system to create and maintain the excess system stress. The equal and opposite force fields, however, cannot translate electrons through loads to freely power them. We prefer to consider any change of potential — even a single potential changed, or both changed so that a net force field does result — as a "regauging", for it is an implied primary assumption *within the normal use ofthe term "regauging."* If only one potential is changed, a net force field results in the system and we call that "asymmetrical regauging". If both potentials are changed, but in such a way that a net force field still results, that is also an asymmetrical regauging. When both potentials are changed precisely in such highly selected manner that the two force fields created are equal and opposite, the force fields sum to a *translation* vector zero, and that is symmetrical regauging or *Lorentz regauging.* However, the energy of the two fields remains, having been transposed to a stress potential. Each of the two "balanced" equal and opposite force fields is continuously performing work in the system, but all that is created is additional system stress. Hence symmetrical regauging is obtained only by altering the stress and stress energy of the system, with two disequilibria in energy flow between the external environment and the system. This in fact constitutes both a curvature of the local spacetime and an alteration of the local active vacuum. It is a rotation of the frame of the regauged system out of the laboratory frame. In effect, the net EM field energy of the two new fields created by symmetrical regauging is transformed into a change in the local spacetime curvature, a change in the local vacuum, increased stress in the system, and rotation of the system's frame away from the lab frame. We focus upon the energy aspects, since in a system any change of potential - whether symmetrical or asymmetrical — is a change of the internal energy of the system, and therefore of its excitation. In a system, conventional "symmetrical force fields" regauging results in a *nonusable continous* energy change in the system (actually two of these changes), and a change in the internal stress of the system. In short, it is a special case of adding a *stress potential* to the system, which we consider a special form of *asymmetrical regauging of the supersystem* consisting of (i) the system and its dynamics, (ii) the local active vacuum and its dynamics, and (iii) local spacetime curvature and its dynamics.

gauge freedom: The axiom in gauge theory that one can freely change the gauge (and therefore the potentials and the potential energy) of an EM system, so long as the net new forces resulting from the change will sum to a zero vector resultant so that the system remains "symmetrical" in its force-field functioning with respect to the system prior to the change of potentials. Electrodynamicists do not seem to have recognized that the freedom to change the potential of a system at will, is *a priori* the freedom to make a purely negentropic change to increase the system's potential energy, or a purely entropic change to decrease the system's potential energy. Gauge and gauge freedom are actually formal restrictions of an even more important asymmetrical regauging principle: *a potential in a system can be freely changed at any time, thus freely changing the potential energy energy of the system freely and at will.* The additional symmetry assumption is an arbitrary added requirement not imposed by nature.

Further, it appears that physicists have not recognized that a longitudinal EM wave is an oscillating gauge varying in its magnitude. Said in reverse, oscillating *the Lorentz regauging condition for a circuit makes that circuit a generator and emitter of longitudinal EM waves.* This also provides a clue for detection of longitudinal EM waves, similar to what occurs in a Fogal semiconductor used for that purpose. A longitudinal EM wave detector must detect a variation in the symmetrical regauging condition, which is a stress change detection.

gauge symmetry: Abstract mathematical symmetry of a field, that relates to the freedom to regauge (change the value of) potentials, without affecting the *net resultant* values of the field quantities. This is a special case of the higher principle that any potential (and thus the potential energy of a system) can be changed freely and at will. However, changing the gauge of an EM system in that sense changes the potential energy of the system but *only in stress potential fashion.* In other words, it does nontranslating work on the system, which is called "internal work." The lack of a net resulting force in the system, created by the regauging symmetry, prevents use of this excess free energy to translate system charges and power an external load. All the excess free energy that could otherwise be used to power loads freely, is instead "locked up" to only perform internal work on the system to increase and maintain system stress.

gauge theory: A field theory using a field that has one or more gauge symmetries. The restriction to symmetry is a carry-over of the old erroneous notion that potentials are just convenient mathematical figments, and that fields are the primary causes of all EM phenomena. For the overunity researcher, broken gauge symmetry at a given level (usually in 3-space EM energy flow) is the feature that must avidly be sought in experimental systems if one would use asymmetrical regauging to allow free powering of loads. Hence one is primarily interested in a wide variety *of broken symmetry effects* and phenomenology. Particularly, unless the automatic self-enforcing of symmetrical gauge transformation during excitation discharge of the closed current loop system is violated, no EM system can produce COP>1.0. Further, local spacetime curvature between local input and local output is always involved in an overunity EM power system. The tempic (time) potential of the output end of an overunity system differs from the tempic potential of the input end, so that the gradient (the tempic force) is directed from output to input — exactly the reversal of the case for a COP<1.0 power system.

Thus in both COP<1.0 systems and COP>1.0 systems, a local spacetime curvature is involved when the system is in operation. In COP< 1.0 systems, the tempic force sweeps along from input to output and on into space beyond, where the violent interactions of vacuum absorb the associated Dirac sea hole current immediately. In COP>1.0 systems, the tempic force sweeps along through the local vacuum from the output section back through the system and into the input section, and thence on back through the external conductors to the external power source. This reversed Dirac sea hole current may "eat electrons" in the normal currents

internal to the system, and in the input section and back towards the external power supply it will "eat input electrons in the input current." Thus the Dirac sea hole current in the COP>1.0 system "eats input power" and serves as a curious new kind of electrical "load" or "internal loss" in the **input** section of the system. The reversed Dirac sea hole current is nature's way of decaying the local curvature of spacetime back to a flat spacetime, and decaying the COP>1.0 excited system disequilibrium state back **to an** equilibrium state. In any overunity electrical power system, this Dirac sea hole current must be taken into account and either transduced into useful power, minimized, or eliminated. An extension to gauge field theory is needed in which the tempic force, the tempic potentials and their gradients, and the supersystem are all modeled and taken into account. It appears that the Sachs unified field theory is capable of handling that requi**rement,** which also can be accommodated by O(3) electrodynamics to provide engineering of the resulting unified field theory.

gauge **transformation (in electromagnetics):** Conventionally taken to be the a**ddition** of the gradient of some function of space and time to the **magnetic** vector potential—and simultaneously the addition of the negative of the partial derivative of the same function with respect to time, divided by the speed of light, to the electric scalar potential. The resulting symmetri**cal** change assumes that the stress energy of the system is changed **freely,** but by design it deliberately prevents any stress difference in the system to produce a net force that could expend the free change of energy by doing work in an external load. Changing the stress energy alone is a curvature of spacetime and a purely gravitational change. It is also a **rotation** of the frame of the system away from the laboratory frame. See discussion under gauge theory.

general theory of relativity: Einstein's theory of gravity in which the gravitational force is represented by a curvature in spacetime, and in which spacetime is an active entity. We may look at all forces as being due to curvatures of spacetime interacting with mass.

geometric phase: A phase difference acquired by the state function of a quantum system that undergoes a cyclic evolution from an initial state and back. This phase difference can be detected by interfering the final and initial s**tates.** The modern generalization of the Aharonov-Bohm effect and the Berry phase.

giant nege**ntropy (of the dipole):** The continuous receipt from the time domain **of spacetime** (the vacuum), by one charge of any dipolarity (including **a** single charge's fundamental dipolarity), of enormous longitudinal EM wave energy flow in condensed time energy form, transduction of that absorbed energy from the complex plane into real 3-space, and continuous emission of that enormous energy flow in all directions in 3-space, thence receipt of that 3-space energy by the other charge of the dipole or other adjacent dipoles, retransduction of the energy back into time-energy, and re-emission of that energy back into the time domain. This 4-circulation giant negentropy process is evoked and continues automatically and indefinitely after a little energy has been expended to separate the charges and make the source dipole. The EM waves in the input from the time domain and the output in 3-space are perfectly ordered and correlated, which is a giant ordering of the vacuum/spacetime and is thus a giant negentropy. In conventional circuits, the source dipole in the generator or battery, once made, performs this function. Only a minuscule fraction (some 10^{-13} in a nominal case) of the energy pouring out of the generator's or battery's terminals strikes the surface charges of the attached external circuit and is diverged into the conductors as the Poynting component that powers the Drude electrons and the circuit. All the rest of the energy flow (the Heaviside nondiverged component) outside the circuit is just wasted by conventional circuits.

gluon: A particle that carries the forces between quarks. In quantum chromodynamics, the gluon plays a role analogous to the photon in quantum electrodynamics.

graviton: In the quantum theory of gravitation, the graviton is the quantum of the gravitational field. It is massless, with spin 2. In the new theory we may take the graviton as a coupled scalar and longitudinal photon pair.

gravity: The phenomenon characterized by the physical attraction between any two material bodies of positive mass, specifically due to the trapped positive energy in the masses. If one particle is of negative mass and energy, the normal gravitational attraction between the two particles is reversed so that they repel, which produces antigravity. The fundamental mechanism of gravity can be found once one realizes that Maxwell omitted the incoming EM wave interaction with the nucleus, and not with just the electron shells. When he wrote his theory, the electron, atom, neutron, proton, and neutron had not been discovered yet. The lateral precession movement of the spinning Drude electrons — which are restrained *longitudinally* to a very small drift velocity in detecting wires (antennas) — later came to be considered proof of Maxwell's transverse EM wave in the vacuum. The detected electron precession wave is oscillating laterally, hence proves the longitudinal EM wave interaction. Elsewhere we have pointed out that the EM wave in space is longitudinal EM wave accompanied by a corresponding time-polarized (scalar) EM wave, and it also follows from Whittaker's 1903 paper as reinterpreted by the present author in our Giant Negentropy paper.

gross particle translation: Refers to translating the entire particle — and its internal structures and dynamics — in space, rather than internally structuing the particle's potential (massless charge). Note that, if the internal structuring (engines) of the particle is changed during translation (e.g., by passing through a region of the flat spacetime having an altered internal engine structure), the particle may have an anomalous interaction with another particle or process having a different internal structure or "dimensioning." The mass-energy comprising the particle's nucleus, electron shells, etc. can be dimensioned. Mass so dimensioned or conditioned then may act differently in chemical reactions, effects on living systems, etc. than normal "inert" mass that has no specific deterministic dimensioning.

group: A mathematical structure that is a complete set of transformations on a mathematical entity that leaves the entity unchanged.

group, higher symmetry: In a given theoretical area, a particular theory can be expressed in various group transformations. When the group transformation (group symmetry) utilized is more complex than the "standard" group symmetry usually employed, it is referred to as a *higher* group symmetry. The mathematics of the model is then said to be a higher group symmetry mathematics. Since the groups have names such as U(1), SU(2), O(3), etc., the transformations of the mathematics are often referred to by **those** names. Hence characterizing phrases are used such as U(1) symmetry electrodynamics, SU(2) symmetry electrodynamics, O(3) symmetry electrodynamics, etc.

group, Lie: A group whose transformations act continuously. An example of a Lie group is given by rotations in ordinary space.

group theory: The general mathematics of symmetry transformations; i.e., the the**ory** of all transformations that leave the transformed entity or entities unchanged.

gyroelectrons: Spinning electrons considered as gyroscopes during the portion of their spin in 3-space, so that when longitudinally perturbed they can and will laterally precess. Considering (in a gross first order model) the conduction electrons in the Drude gas in a conductor as little gyros, because they are (i) spinning on an axis, (ii) severely restrained in their ability to move forward longitudinally down the conductor as current, and (iii) much more free to precess laterally in the conductor. Thus when the gyroelectrons in the receiving wire antenna, perturbed by an incoming longitudinal EM wave in the vacuum, are longitudinally perturbed, they precess laterally so that our instruments (which detect electron wiggles, not the incoming vacuum EM wave) do detect "transverse waves." However, they are detecting the transverse gyroelectron precession waves, not the waves in the vacuum. They are detecting the *effect* of the interaction of the incoming waves with the gyroelectrons, not the *cause.*

harmonic series: A set of frequencies or oscillations consisting of integer multiples of a fundamental frequency.

Heaviside, Oliver: Noted English self-taught physicist and brilliant electrodynamicist, 1850-1925. who played a role in discarding Maxwell's quaternions, and also played a role in forming vector mathematics and formulating the vector reduction of Maxwell's theory from 20 quaternion equations in some 20 unknowns to the present 4 vector equations. Heaviside's curtailment of Maxwell's theory still contained both (i) systems in equilibrium with their active environment and (ii) systems in disequilibrium with it. To separate variables and enable closed solutions, *Lorentz further* curtailed the Maxwell-Heaviside equations by symmetrically regauging them, thereby *arbitrarily* discarding all disequilibrium Maxweilian systems and retaining only those in equilibrium with their active environment. This regauging unwittingly and erroneously discarded all that class of Maxweilian systems permitted to exhibit COP>1.0. The first such curtailment of the Maxwell equations was by Ludwig Valentin Lorenz in 1867, but H. A. Lorentz (who did a similar symmetrical regauging nearly decades later) was later given the credit.

Heaviside also discovered the enormous *nondiverged* energy flow surrounding every circuit, far greater in magnitude than the feeble Poynting *diverged* component that actually enters the circuit. But he was at a loss to explain what the source of such a startling flow of energy could be. Lorentz understood the Heaviside component, but also could not understand its source. Hence he reasoned that it "had no physical significance" since it did not power the circuit. Lorentz then introduced the trick of integrating the entire energy flow vector around a closed surface surrounding any volume element of interest. This neatly discards the vexing giant Heaviside *nondiverged* energy flow component, and retains only the *diverged* Poynting component that actually enters the circuit. The latter is of course the energy that will later be *emitted* from the circuit, which is what we will measure when we "measure the circuit."

Electrodynamicists are still using this method of dropping the Heaviside energy flow from any further accounting consideration. Hence they do not develop systems which intentionally increase their energy collection from this readily available energy flow around every circuit. The Bohren experiment conclusively proves the physical existence of the Heaviside nondiverged component, even though it is no longer accounted.

In his last years Heaviside realized the gravitational importance of his extra energy flow component, and worked out an electrogravity theory based on it. He did not live to publish it, but his papers on it were found some time after his death, beneath the floorboards of his little garret apartment. A learned society then published the papers.

Heaviside energy flow component: The enormous *nonintercepted* and nondive*rged* energy flow component pouring out of the terminals of every battery or generator, or charge or dipole, and missing the external circuit (or external intercepting charges) entirely so that it is does not contribute to the intercepted (Poynting) flow component that powers the circuit. Also, the enormous energy in the EM field, potential, or wave that misses the interaction of the field, potential, or wave with the assumed unit point static electric charge, and thus is not intercepted and collected (and accounted!) in the unit point static charge assumption. In honor of Heaviside, the present author has nominated this "dark Heaviside energy" component as the missing source of excess gravity known to be holding the arms of the spiral galaxies in place. In the interactions of Dirac hole current with a unit point charge (to determine the negative energy fields and potentials in a COP>1.0 EM system), there is also a missing component of the negative energy flow that is comparable to the Heaviside component for positive energy flow. These flows of surplus negative energy **from** COP>1.0 astronomical interactions—with the associated negative energy fields and negative energy potentials produced by the **negative-**energy Dirac holes in these negative energy flows — is hypothesized by the present author as producing the antigravity that generates the observed acceleration of the expanding universe.

Hertz, **Heinrich:** Noted German physicist, 1857-1894, who together with Heaviside and others transformed Maxwell's theory from quaternion algebra to vector algebra, and who also experimentally produced Hertzian (EM) waves, thus confirming Maxwell's theory. These experiments by Hertz first detected EM energy propagating as described by Maxwell's equations, thus providing the experimental impetus that allowed Maxwell's theory to predominate. The oscillation or "cycle" is named after Hertz, by specifying *X* number of oscillations or cycles per second as a frequency of *X* Hertz. One type of potential is also named after him.

hidden bidirectional EM waves: Those bidirectional EM wavepairs that comprise the scalar potential are indeed "hidden" with respect to our normal "electron translation" (i.e., electron wiggle) detectors and instruments. Nonetheless, properly functioning devices such as Fogal's semiconductor allow the utilization, transmission, and reception of the vast longitudinal EM wave dynamics that are infolded inside all conventional EM fields, waves, and potentials, including the giant potential comprising the ambient vacuum (spacetime). Use of this infolded electrodynamics is akin to using "subspace" or higher dimensions, depending upon how one chooses to model it mathematically. It is also a direct method of using a unified field theory.

hidden variable theory: A class of quantum mechanical theories which considers that the quantum state of a physical system is not a complete specification after all. The "hidden variables" are those additional components necessary to provide the "complete state" of the system. By utilizing the infolded longitudinal EM wave electrodynamics, one is in fact using a macroscopic hidden variable theory. This becomes a subset of Evans' O(3) electrodynamics, which in turn is an important subset of Sachs' unified field theory. Sachs' theory encompasses and subsumes electrodynamics, general relativity, quantum mechanics, etc., so the use of this macroscopic hidden variable theory through special functions provided by components such as Fogal semiconductors is a powerful and directly engineerable extension of conventional hidden variable theory. The physics of this new, engineerable, unified field theory area is also called *energetics.*

hole, lattice: Vacancy, in an otherwise filled energy band position. Sometimes called *negative ion vacancy.* In a semiconductor, an electron may be lifted to the conduction band by thermal energy, which produces u hole-electron pair (an exciton). Also, one of the sites in the crystal lattice can be occupied by an acceptor impurity atom, so that the electron is accepted there, leaving a hole showing excess positive charge. This is a lattice hole, and the propagation of such holes being formed is called electron hole current, or hole current for short. We point out that the *laittice hole* exists only after interaction of the Dirac sea vacuum hole with matter. i.e., only after observation. In that case, the 3-positron (hole) has positive mass and positive energy, produces positive energy fields and potentials, produces gravity rather than antigravity, and is "seen" to be going in the opposite direction to the Dirac sea hole. Prior to observation, the hole is a

Dirac 4-hole in vacuum, having negative energy and therefore negative mass equivalence, negative charge, and without a direction reversal. This 4-hole is a source of negative energy fields and negative energy potentials in **the** surrounding space, and thus of antigravity back-interaction (upon the physical system connected with the active vacuum component of its supersystem) due to the negative curvature of the local spacetime.

hole, Dirac sea: In the presence of fields (i.e., in the presence of curved space**time),** some of the negative energy states in the Dirac sea are not filled with electrons. In that case there are some empty 4-holes in the Dirac sea, in local spacetime itself. The Dirac hole is a negative energy state in the **vacuum** itself. It too *seemingly* has a positive charge, as seen by the external observer after interaction with matter. Holes are also mathematically equivalent to positrons, in case observation has been evoked, so long as the lack of positive mass is not a consideration Note that our view differs from the received view, but is based on experimental evidence which can only be gained from COP» 1.0 EM system experiments. We know of no such COP» 1.0 EM system experiments involved in the received view.

hole current: (1) In a semiconductor: Usually, the part of the current in a seminductor that is composed of migration of lattice holes. ***Mathematically,*** within certain constraints this may be considered a positro*n* current. The hole migration is usually slower than the electron migration producing electron current. (2) In the vacuum itself. In the *Dirac sea, in* **the** presence of fields the Dirac 4-holes flow in the opposite direction to the local electron current, providing a Dirac sea hole current or posit**ron** current. [See discussion under hole.] An electrical power system in disequilibrium with its vacuum interchange and having COP>1.0, will have a **greater** *tempic potential* at its output section than at its input section. Thus it will have a tempic force or tempic broom, sweeping holes from local instabilities to form a hole current (may be both a Dirac sea hole current and a lattice hole current) flowing from the output section of the system back through the system into the input section, and on out of the input section into the primary power source. *Hole current transduction may be* used to convert this detrimental hole current energy transport into usable **electron** current energy transport, or anti-circuit techniques and mechanisms can be used.

hole current transduction: Hole current in a COP>1.0 open system, where **the** system is in disequilibrium with the vacuum *a priori,* is usually detrimental to (he system and may destroy the overunity condition. A hole current from the input section of the system, fed back into the primary power supply, requires that extra electron current be furnished since some of the electrons in that furnished current will recombine with the oppositely moving Dirac vacuum holes and disappear to fill those holes. Hole current into a battery in "normal" fashion discharges the battery rather than charges it (although reversing the battery terminal connections allows charging in normal fashion using the hole current). Thus a battery can be discharged "far below zero", so to speak, by charging it with excess holes in the "normal connection." A capacitor may be charged with hole current, in which case the dielectric will undergo compressive strain into the positron-charged plate. However, when the capacitor is subsequently discharged, the dielectric will expand and recover, with its electron charges pushing electrons out of the negative side plate and into the external circuit. Hence a capacitor can be *charged* with positron (hole) current, and then can be adroitly switched so as to *discharge* electron current. In theory so can a battery because of its capacitance. This is a transduction of entropy to negentropy, a process for which a patent application has been filed by Bedini and Bearden. Gauge freedom already permits that operation, since gauging in electromagnetic systems is freely changing the local potential energy of the system, and is either negentropic (increasing the system's energy) or entropic (decreasing the system's potential energy).

hole injection: In an n-type semiconductor, holes can be emitted by applying a metallic point to the semiconductor's surface. We suspect (but have not proved) that sharp gradient pulses may also induce some of these holes as Dirac sea 4-holes and Dirac sea hole current rather than positron and lattice holes current. In that case, a mix of positive energy and negative energy (and gravity and antigravity) would be produced.

hole trap: An impurity in a semiconductor that can trap a "hole" by releasing electrons to the conduction or valence band.

hyperspace: Refers to a space of more than four dimensions, and specifically to those spatial dimensions outside the normal three. The term is used for both a space of n dimensions, where n>4, or for one dimension in such a space, beyond the normal four.

identity of opposites: On the boundary between A and not-A, A and not-A are identical (i.e., no difference can be discriminated). If all of a thing is collected, one reaches the boundary of that thing and it turns into its own opposite. The philosopher's problem of the "accursed necessity for the identity of opposites" cannot be solved in Aristotelian 3-law logic, since that logic is incomplete. It is readily solved, and part of, a more extended logic such as 5-law logic. The result is that "identity" is not an absolute entity, but is a decision by the observer reached by application of a comparison algorithm to separate observations made at differing times. An example is provided by a colorblind observer who sees as identical two marbles which, to an observer with color vision, are seen as red and black and therefore not identical. Aristotelian logic omits the action by the observer, the times of the multiple observations (only a single thing at a time is observed or perceived, *a priori),* and the decision algorithm used *by the* observer to decide whether two observations are identical or not identical.

inertial frame: A frame of reference in which force-free bodies move along straight lines, and the postulates of special relativity are valid. In the new a**pproach,** this notion is extended. Even in an inertial frame, force-free **bodies** may move along straight lines, the postulates of special relativity may remain valid, and yet the EM hidden variables may induce other chan**ges** in the usual laws of nature due to the action of hidden vacuum engines infolded in the *dimensioned* local flat spacetime (vacuum **potential).** See Einstein's second postulate extended.

infolded real vector components: Simply take a set of real force vectors that altogether sum to a vector zero resultant. This "vector zero system" is a zero vector with respect to forced *translation* of a particle or mass. It is not am absence of the vectors at all, but is the presence of specifically hidden and "infolded" real vectors that do not initiate bulk translation of the affe**cted** object or system, but perform *internal* work upon the system. Such vector systems are stressing systems of a special kind, where the internal components of the stress are not random but deterministic. Hence they are electrodynamic *engines,* in the new approach. To a somewhat arrogant mathematical scientist who once objected that zero vector resultant systems had no effect, I humorously challenged him to put it to experimental test. I suggested he stand between two elephants pushing against him in opposite directions, while I would stand between two fleas doing the same to me, and we would compare notes. He apparently was not amused at the suggestion.

infolding: With reference to signals, infolding means "placing the signal **or modulation** or change upon one or more of the internal bidirectional wave**pairs** comprising the scalar potential. Since all EM potentials ultimately represent changes to the ambient vacuum potential, one has conditioned or altered the vacuum itself when infolding is utilized, infolding of E*M* signals, vectors, and directions produces *engines*. By Whittaker's decomposition of the potential and Bearden's reinterpretation, these vacuum engines can be designed to affect matter — including atomic nuclei, nucleons, etc. — in any manner desired.

information content of the field: Russian euphemism for the deliberate structuring of the hidden longitudinal EM waves and their dynamics, that comprise any EM potential, field, or wave, so that the potential, field, or wave is *dimensioned* and contains active *engines.* This area was extensively weaponized under KGB (now changed in name) control rather than the regular Russian armed forces, and a disinformation campaign was waged to convince Western analysts that the term refers to simple spectral analysis. It doesn't. Instead, it refers to infolding and outfolding, and the use of longitudinal EM waves, dimensioning, and engines. The overall science is called "energetics" by those weapon scientists.

interferometer: An instrument or system in which a wave is split into two waves or beams, which after traveling over different paths are subsequently reunited and interfere. The use of longitudinal EM waves (LWs) allows ready penetration of mass by the LWs traveling through the interior Whittaker composition of a mass's internal EM fields, waves, and potentials. Longitudinal EM wave interferometers should prove useful in probing the interior of planets, the sun, etc. They pass easily through the earth, so that interferometry can be obtained at targets on the opposite side of the earth, or inside the earth. These are the weapons that then Secretary of Defense Cohen referred to in 1997 as being used to engineer the weather, initiate volcanoes into eruption, or initiate earthquakes.

Josephson effect: Flow of Cooper pairs (superconducting electron pairs) across a thin dielectric separating two superconducting electrons, in the absence of a voltage drop. E.g., the pair of electrons can "tunnel" through u thin insulating barrier.

joule: A measure of energy; one watt for one second in electrical terms. In physics the unit of work or energy in the MKS system, being the amount of work done by one newton acting through a distance of one meter, equal to 10,000,000 ergs.

Kirchhoffs current law: At any time, the sum of the instantaneous currents flowing into a node or point in a circuit equals the sum of the instantaneous currents flowing out of that node or point. This does not have to be true if (1) relativistic effects are appreciable, or (2) electron production and annihilation processes — such as Dirac sea hole current, lattice hole current and recombinations are occurring and non negligible, (3) the Kron open path is being used, or (4) localized time reversal zones (TRZs) are utilized. Another way of stating where the law need not hold is when there are nonlinear interactions between the three components of the superset (the physical system and its dynamics, the active vacuum and its dynamics, and spacetime and its curvature dynamics.).

Kirchhoff's voltage law: At any time, the sum of all voltage rises in a closed loop circuit equals the sum of all drops in that closed loop circuit. The law assumes a flat spacetime. This does not have to be true if relativistic effects are appreciable, if true negative resistor effects are present, if Kron open paths are utilized, and if time-reversal effects arc present. It also need not be true in anti-circuit techniques, or when they are involved in a particular portion of an otherwise normal circuit.

Lamb, Willis Eugene Jr.: American physicist and Nobelist who discovered the hyperfine structure of the hydrogen spectrum, and first measured the interaction of the vacuum energy that generates a change in the energy levels of the hydrogen atom, according to Dirac's theory and quantum electrodynamics.

Lamb shift: A very small correction to the first excited state of the hydrogen atom due to interaction of the vacuum. Willis Lamb experimentally confirmed this prediction of quantum electrodynamics to high accuracy. The local energy density of the little Lamb shift is greater than the surface energy density of the sun. However, in matter there are such a vast number of similar small interactions generating very powerful fields randomly, that the sum total of all these changes sums to a vector zero for most summation distances of interest.

lamellar current: Thin layers of current; thin sheets of current.

laser: A device that generates light in which all the photons are exactly in step and produce a coherent beam. Originally the word "laser" was an acronym that stood for "light amplification by stimulated emission of radiation."

law of physics: A restriction placed upon the physical mechanisms that can occur under certain broad assumptions as to the background situation and environment. Because of the assumptions made, an expressed law of physics is not absolute, but relative to the assumed conditions. Any "law of physics" can be violated if the conditions contained in its assumptions are violated. It is important that a scientist be aware that physics is comprised of models and the "laws of physics" are very general and wide-ranging models based on very specific assumptions. Further, by Godel's theorem and the theory of models, no model is perfect anyway, and none will ever be. When one assumes that physical laws are not based on any assumptions but are universal and absolute truths, one has immediately moved to dogma and not science.

Lawandy patents: Patents covering and related to a form of anti-Stokes emission (excess emission of energy) in strongly scattering, optically active media such as TiO_2 particles in suspension. This is a new form of lasing, without the requirement for population inversion.

Letokhov, V. S.: Russian pioneer in excess emission phenomena from stimulated media.

Letokhov effect: The excess emission (excess emission of energy) phenomena in strongly scattering, optically active media, usually referred to as "negative absorption of the medium." See particularly V. S. Letokhov, "Laser Maxwell's Demon," *Contemporary Physics,* 36(4), 1995, p. 235-243.

localized: Confined to a small region of a large system rather than being extended through the system or at a distance.

localized virtual photon flux of vacuum: Refers first to a local region of the vacuum, and then to the flux density of virtual photons in that local region. If that flux is deterministically structured, then the local region also contains an *engine* that will act internally upon masses and mass systems in that region. The local region also contains hidden variables.

longitudinal component: In EM fields, the longitudinal component of the fields along the direction of travel, as opposed to the transverse components at right angles to the direction of travel.

longitudinal E-field: Any electrical field in 3-space may be described using two components, a transverse component and a longitudinal component. Of course one or the other may be a vector zero. In 4-space, there is a time component also, as an additional "transverse" component.

longitudinal gradient: Compression wave gradients (differences) in density or pressure, considered in a direction from their origin along the direction of primary movement in the medium. Contrast to transverse waves, which act perpendicular to their direction of propagation. In electrodynamics, longitudinal EM waves in 3-space are said to be "longitudinally polarized." Longitudinal EM waves in the time-domain are said to be "time polarized." The latter are also referred to as "scalar waves" since they have no polarization vector component in 3-space and thus arc "scalar" with respect to 3-space. This is a misnomer, however, since these "scalar waves" are perfectly good time-polarized EM longitudinal vector waves. For the first time, the giant negentropy process allows one to make and use time-polarized EM waves, since they automatically accompany all longitudinally polarized EM waves and it is straightforward to make longitudinal EM waves in 3-space (just oscillate the Lorentz regauging condition).

longitudinal interior wave: An infolded longitudinal gradient that oscillates in magnitude along the line of travel "inside the vacuum potential or a normal potential." In conventional theory, the electric portion of the EM wave is a transverse field oscillation, and the longitudinal component is essentially zero. In the new "infolded" wave approach, there is no overall transverse gradient in the vacuum potential, but there are such longitudinal waves. This is rather like a "pressure wave" transmitted under water, without causing a surface wave involving physical upheaval and fall-off of the water. The speed of the longitudinal EM wa**ve** is not limited to c, the speed of transverse EM wave light in the ambient vacuum..

longitudinal wave: A "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave. Hence, a wave composed of alternating densifications and rarefactions, where we focus upon the longitudinal component of the changes. One can think of it as a sort of "accordion" wave, rhythmically shrinking and stretching along the line of motion as it travels. Note that rhythmically and symmetrically varying the two potentials involved in regauging creates longitudinal EM waves. So one pseudonym for longitudinal EM waves is "gauge oscillation" waves. Matter is mostly empty space filled with EM potentials, fields, and waves—all of which are bundles of longitudinal EM waves and their functional dynamics. Hence matter (such as the ocean and the earth) is a vast superhighway for the passage of longitudinal EM waves. Note that the longitudinal wave represents an oscillating energy density in space and/or time (scalar waves), and hence is an oscillating curvature of spacetime. All forces of nature are generated by curvatures of spacetime. Hence all forces of nature can be generated by the correct generation of longitudinal EM waves and their impressed dynamics, Further, this means that all the physical *causal* functions of the exhibited **effects generated** in a material system can be captured in longitudinal EM wavesets. A longitudinal wave functional system deliberately having such causal functions of a physical functioning system is known as a causal **system robot** (CASR). Such robots are far beyond the presently developing nanobots, and have been developed rather extensively by at least one nation. Such robots can easily pass through the "interior" of any normal

EM signal or wave or potential or field, which again is nothing but a giant superhighway for longitudinal EM waves. In theory one can communicate with a CASR via longitudinal EM waves if communications functions are built in to the CASR. In the future, one would expect that planetary exploration, underground and undersea exploration, exploration of the interiors of the sun, moon, and planets, could be enabled by the further development of specialized CASRs.

Lorentz, Hendrik Antoon: Renowned Dutch physicist and Nobelist. 1853-1928, who developed the Lorentz transformations and Fitzgerald-Lorentz contraction. Lorentz shared the Nobel Prize for discovering the Zeeman effect. He is also responsible for symmetrically regauging Maxwell's equations (some time after Lorenz did it in 1867), and for introducing an integration trick that neatly disposes of the vexing huge Heaviside nondiverged energy flow component accompanying every field/charge and potential/charge interaction.

Lorentz invariance principle: An inertial frame is independent of the velocity of the frame relative to any other inertial frame. This is not necessarily true with respect to engines inside inertial frame objects, since the engines involve acceleration dynamics.

Lorentz regauging: Changing the two potentials in the Maxwell-Heaviside equations in potential form so that the resulting equations have the variables separated, and there is no net force introduced by the two new forces, which are deliberately chosen so as to be equal and opposite and thus sum to zero. An unfortunate effect is that this discards all Maxwellian systems far from equilibrium in their vacuum exchange, hence discards all permissible Maxwellian COP>1.0 systems. Electrodynamicists have not recognized that Lorentz regauging represents the addition of stress potential energy and continuous internal work and stress to a regauged system. It also represents a rotation of the frame of that system out of the laboratory frame by some angle. Further, oscillating the Lorentz regauging condition produces longitudinal EM wave generation and emission.

Lorentz symmetry: The symmetrical regauging by Lorentz of Maxwell's equations.

Lorenz, Ludwig Valentin: Scientist who first changed Maxwell's equations — two years after their publication in 1865 — by (effectively) symmetrically regauging them, thereby unwittingly discarding all Maxwellian systems far from thermodynamic equilibrium with the active vacuum. See L. V. Lorenz, "On the identity of the vibrations of light with

electrical currents," *Philosophical Magazine*, Vol. 34, 1867, p. 287-301. In this paper [L. V.] Lorenz gave essentially what today is called the [H. A.] **Lorentz** symmetrical regauging. When H.A. Lorentz also symmetrically **regauged** the Maxwell-Heaviside equations, the physics community **adopted** the procedure and resulting further truncation of the Maxwellian theory. That truncation continues to this day.

low energy **nuclear reaction** (LENR): Actually a bit of a misnomer, since it refers to a low *spatial* energy, high *time* energy reaction. Cold fusion, or the anomalous transmutations of elements that occur in certain electrolytes with specially prepared palladium electrodes. The transmutations occur even though in conventional theory there is insufficient *spatial* energy to accomplish them, because they use time energy. Time energy is spatial EM energy compacted by the factor c^2, hence time has the same energy density as mass. The formation of temporary, fleeting time reversal zones (TRZs), with action from from inside-to-outside, is the genesis of a host of new nuclear reactions missing from conventional particle physics since the physicists do not account time as energy, and do not use such TRZs. For more on this hypothesis and the specific reactions generated, see T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *Journal of New Energy*, 3(2/3), 1998, p. 12-28 as well as Chapter 10 of the present book.

magnetic **dipole:** Paired north and south magnetic poles of equal strength ever microscopic or macroscopic distance.

magnetic induction: The act of a change in magnetic field producing an electric field at right angles, which in turn produces an electron flow.

magnetic spin: The intrinsic angular momentum, known as spin, of electrons, protons, and neutrons, etc. The electrons combined in an atom or ion have a resultant angular momentum that is the combined intrinsic spin of the electrons and the angular momentum due to their motion around the nucleus. There is also a magnetic moment associated with this angular momentum (spin) whenever it is nonzero. Therefore atoms or ions with nonzero spin are *magnetic* atoms or ions. This is often loosely reffered to as "magnetic spin." It is actually the magnetic moment of nonzero spin.

magnetostatic scalar potential: Continuous mathematical function whose value in space, at any point it occupies, represents the potential energy collected upon an assumed unit magnetic north-seeking charge (pole). There is a magnetostatic scalar potential between the poles of any magnet. This scalar potential is also subject to **Whittaker's** 1903 decomposition into a harmonic series of magnetic longitudinal phase conjugate wavepairs. Reinterpretation of Whittaker's wavepairs should be made according to Bearden, "Giant Negentropy of the Common Dipole," *J. New Energy,* 5(10, Summer 2000, p. 11-23. This reinterpretation agrees with quantum field theory; see Mandl and Shaw, *Quantum Field Theory,* Revised Edition, Wiley, 1993, Chapter 5.

matter: Highly condensed 3-spatial energy, compressed by factor c^2 and remaining in 3-space, and with the observation process *d/dt* having been applied to the fundamental masstime entity to provide mass as a frozen 3-space snapshot by stopping the flow of time momentarily. Matter, of course, has internal structure, order and organization, and dynamics. All matter is in a continuous energetic exchange with its active vacuum (4-spatial) environment. Matter stripped of its EM fields, waves, and potentials is called mass. Mass is an observable and therefore a frozen series of continual 3-space snapshots, formed by incessant photon absorption to change the mass *m* to masstime *mt,* with following incessant photon emission to change masstime *mt* back to mass *m.*

Maxwell, James Clerk: Brilliant Scottish scientist, 1831-1873, who first formulated the unified theory of electricity and magnetism. See *The Scientific Papers of James Clerk Maxwell,* 2 vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597 for his definitive paper, "A Dynamical Theory of the Electromagnetic Field," first published in *Royal Society Transactions,* Vol. CLV, 1865. In that paper, (see Niven's book) his general equations of the electromagnetic field are given in Vol. 1, Part III, "General Equations of the Electromagnetic Field," p. 554-561. On p. 561, he lists his 20 variables. On p. 562, he summarizes the different subjects of the 20 equations, being three equations each for magnetic force, electric currents, electromotive force, electric elasticity, electric resistance, total currents; and one equation each for free electricity and continuity.

Maxwell's famous book, *A Treatise on Electricity and Magnetism,* Oxford University Press, Oxford, 1873, Second Edition 1881 (Maxwell was already deceased), Third Edition, Volumes 1 and 2, 1891. Foreword to the second edition was by Niven, who finished the work. Maxwell had dramatically rewritten the first nine chapters to simplify his mathematics because of the resistance to it, with much new matter added and the forma contents rearranged and simplified. Maxwell died before finishing the rcsl of the second edition. The rest of the second edition is therefore largely a reprint from the first edition. The third edition edited by J. J. Thomson was published in 1892, by Oxford University Press, and later was published unabridged by Dover Publications, New York, 1954. J. J. Thomson finished the publication of the third edition, and wrote a "Supplementary Volume" with his notes. A summary of Maxwell's simplified equations are given in Vol. II, Chapter IX of the third edition. However, Maxwell had gone (in his second edition) to some pains to reduce the quaternion expressions himself, and not require the students to know the calculus of quaternions (so stated on p. 257). We note that Maxwell did not finish the second edition, but died before that. He actually had no hand at all in the third edition as to any changes. The Second edition was later finished by Niven by simply adding the remaining material from the previous first edition approved by Maxwell. The printing of the first nine chapters of the thirg edition was already underway when J. J. Thomson was assigned to finish the editing of the manuscript.

Maxwell's work profoundly affected the course of science and the course of human history. Maxwell, Heaviside, Hertz, Gibbs, Lorcntz and others participated in dramatically curtailing Maxwell's original 1865 quaternion theory to the much more limited subset that it is today. Since the early part of the 20th century, work has slowly been ongoing to move EM theory back to a higher group symmetry theory, which ironically it already was when Maxwell first created it.

Max**well's** electrodynamics: Simply put, Maxwell's theory of electrodynamics consists of Maxwell's equations. His fundamental theory consisted of some 20 quaternion equations in 20 unknowns, appearing in his 1865 paper. After Maxwell's death and some curtailment by Maxwell himself, Heaviside modified and sharply curtailed these equations, into the familiar four equations of today, as did Gibbs and Hertz. Lorentz further curtailed the Maxwell-Heaviside equations, by symmetrically regauging them. This simplified their mathematical solution, but it also inadvertently and arbitrarily discarded all open Maxwell systems far from thermo**dynamic** equilibrium with their active environment (such as the modern active vacuum). See James Clerk Maxwell.

Maxwell's Theory, fluid dynamics analogy: *Considering and adapting the fluid analogy approach of James Clerk Maxwell:* The intensity of the vacuum virtual particle flux may be treated as the intensity of a potential, thus by analogy corresponding to pressure in a fluid. Polarity (charge) is a **process for** establishing a change in the intensity of the vacuum potential - and **thus** a change in the ambient pressure of the vacuum fluid. Dif**fence** between vacuum pressures at two separated points creates a force u**pon** any charge placed between those two points. Just as a differential in pressure between points in a fluid induces currents, a differential in vacuum pressure induces energy flow currents in the vacuum fluid, which are called 'fields'. All EM circuits, charges, potentials, and forces in electrodynamics involve altering the vacuum to include altering its pressure and thereby producing the forces the resulting pressure gradients (fields) induce upon charges. All observable EM energy in space is EM energy intercepted and collected on charges, or assumed to be, and the spatial EM energy comes from the vacuum via these fundamental vacuum-engineering dynamics which all involve the giant negentropy 4-circulation between the time domain and 3-space.

Michelson, Albert Abraham: German-born American physicist, 1852-1931, who invented an interferometer (Michelson interferometer) named after him. Michelson used his interferometer in a renowned experiment in the 1880s designed to measure the velocity of the earth through the ether, which would produce a shift in his interferometer pattern. No such shift was observed, which indicated the absence of an ether wind and was largely responsible for the downfall of the concept of the material ether that had been previously assumed, including by Maxwell.

motionless electromagnetic generator (MEG): Transformer-like COP>1.0 electrical power generator invented by Patrick, Bearden, Kenny, Hayes, and Moore, ""Motionless Electromagnetic Generator," U.S. Patent # 6,362,718, Mar. 26, 2002. A second patent application is processing and several more are in preparation. In particle physics, any dipole (including a permanent magnet) is a known asymmetry in the fierce vacuum flux, continuously pouring out real EM energy extracted from the vacuum. The MEG continuously powers its transformer section core with a permanent magnet dipole, more than doubling the EM energy made available from **the** active vacuum via the permanent magnet dipole's broken symmetry in its exchange with the vacuum. [See M. W. Evans, P. K. Anastasovski, T. E. Bearden et al., "Explanation of the Motionless Electromagnetic Generator with O(3) Electrodynamics," *Found. Phys. Lett.,* 14(1), Feb. 2001, p. 87-94; "Explanation of the Motionless Electromagnetic Generator with the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(4), Aug. 2001, p. 387-393]. The MEG nanocrystalline core material extracts the magnetic field B of the permanent magnet from the surrounding space and localizes it in the core material. The active vacuum asymmetry freely replenishes the withdrawn magnetic B-field energy in space surrounding the core, but with uncurled magnetic vector potential A rather than magnetic field B. This is an application *of geometric phase,* specifically the well known *Aharonov-Bohm effect,* with several thousand papers in the literature. The MEG then transduces and outputs magnetic B-field energy inside the **core** material and some of the field-free magnetic vector potential A energy in space surrounding the core, effectively having available more than double the energy being made available from the permanent magnet, or normally available to a transformer. The output power of the MEG is a function of (i) the time rate of change of a small amount of input EM perturbation energy to the input coil, plus (ii) the intercepting and collecting secondary charges available in the secondary coil, (iii) additional energy from chaotic oscillations, (iv), sophisticated feedforward and feedback regenerative self-oscillations, and (v) the geometric phase (Aharonov-Bohm effect). It thus **is** a rigorous practical application of the geometric phase effect to a macroscopic power system, and in fact is the first such application in practical electrical power systems. The noted feature of the MEG is a purely electrical interaction in the output coils, in which the voltage and energy are in phase within 2°, differing from all other transformer-like systems. Thus the MEG directly delivers real power rather than apparent power. See Aharonov-Bohm effect.

multicollection (of energy): Multiple collection and re-collection in a circuit of the same energy flow from a source dipole, by rerouting either (1) the dissipated Poynting energy back around to the collectors for another pass by them, with additional collection, or (2) retroreflecting the nondiverged Heaviside EM energy flow component back and forth across the surface charges in the circuit conductors, iteratively diverging additional energy into the conductors to potentialize and power the Drude electrons, or (3) a combination of the two. The interesting point is that this effect **can** "use" the same energy to do repeated work, multiple times. It takes advantage of the fact that work is merely the change of form of energy, and that the new form can again be changed (as for example, back into the **first** form again) to do additional work, again and again so long as the energy in each transformation is held in the system without escaping. This process permissibly violates classical equilibrium thermodynamics and its **infamous** second law, since it is an open system freely receiving excess energy from its active environment and thus far from equilibrium with that active environment. As is well known, such systems are permitted to exhibit COP>1.0 and even self-powering of the system and its load simultaneously.

negative absorption of the medium: Excess emission from an active medium, as in the Letokhov work and in Bohren's experiment. The excess energy is first absorbed from the vacuum interaction (actually from Heaviside's unaccounted energy flow component that usually does not interact with the system), and then it is re-emitted. This is actually a true "Maxwell's demon," as pointed out by Letokhov.

negative resistor: Any component or function or process in an EM system that freely receives energy from outside the system in unusable or disordered form within the system, transduces that energy into usable form, and outputs it so that it can do useful work. In a *true* negative resistor, that is the *net* function performed. We specifically do not include "differential" negative resistors such as the tunnel diode, thyristor, and magnetron, which dissipate and disorder more energy from the system itself, overall, than the fraction that they order and furnish back to the system in their negative resistance regime. In anti-circuits, true negative resistors are quite natural components. Every charge and dipole in the universe is already just such a negative resistor, freely extracting unusable EM energy from the seething vacuum, transducing it, and continuously pouring out real observable EM energy in all directions in 3-space. A negative resistor can produce the five magic functions: (i) self-ordering, (ii) self oscillation or self-rotation, (iii) output more energy than the operator inputs, (iv) power itself and its load freely, and (v) exhibit negentropy. The condition for self-powering of a system containing a true negative resistor was given by Gabriel Kron.

negentropic: Functioning to produce or involve negentropy; reordering previously scattered energy without requiring the operator to input energy to perform the reordering. A simple example is passive retroreflection of radially scattered EM energy. The conventionally unaccounted enormous and nondiverged Heaviside energy flow component surrounding every circuit and every field/charge interaction is an example of a dynamically reordered and organized energetic vacuum, and thus an example of a **giant** negentropy process. The continuous observable EM energy freely pouring at the speed of light from every dipole and charge in the universe is also **an** example of the giant negentropy process.

negentropic engineering: Expending a little bit of energy to form a dipolarity, and then intercepting, collecting, and using some of the enormous EM energy freely and continuously extracted by the dipolarity from the vacuum, indefinitely, without using any of the extracted EM energy to scatter the charges in the dipole and destroy it. Retaining the dipole intact avoids shutting off the furnishing of copious free energy **flow** from the vacuum to the attached external circuit. As an analogy, it allows an electrical windmill operating in (driven by) a free electrical wind from the external environment of the system.

Newtonian recoil: Newton's third law reaction from the affected entity back to and upon the effecting entity. The "action" is the action from **the** effecting entity upon the affected entity. The "reaction" is the action from the affected entity back upon the effecting entity. Newton's third law actually implies that the cause interacts upon the effect and the effect reacts equally back upon the cause, thus producing a strange kind of equilibrium. In a time-reversal zone, the "recoil" is reversed in direction by parity reversal. Tentatively, one may interpret from pumped phase conjugation that the Newtonian third law reaction of one single acting force can be amplified by properly directing the actions of other acting forces. An interesting aspect is that a pumped phase conjugate mirror does not recoil, no matter how powerfully it is pumped or how powerfully amplified is the emitted phase conjugate replica wave.

Newton's third law: For every action there is an opposite and equal reaction. This law should be extended and restated something like this: "for every action there is an opposite and equal reaction if the causative mechanism of the reaction is allowed to occur without interference and is not first redirected away from its target." An example of a violation of Newton's third law as usually stated is provided by the emission of a phase conjugate replica (PCR) wave from a phase conjugate mirror material. No matter how powerfully pumped, the PCM does not recoil by the emission of the PCR wave. The reason is that the antiwave cause of Newtonian third law recoil was intercepted by nonlinear multiwave interaction (wave-to-wave **interaction)** before reaching its target nuclei and interacting with them, and was diverted away from the atom and on out of the phase conjugate mirror material. Since the causative mechanism for Newtonian recoil did not occur on the mass of the mirror, the mirror did not recoil. We also point out that, if Newton's third law is indeed engineerable and controllable, then COP>1.0 systems and self-powering EM systems are permissible a priori. In electrical system terms, this simply means that, if one can rid the system of its equal back emf or back mmf, the system can be legiti**mately** self-powering because it no longer self-enforces the Lorenz-Lorentz symmetrical regauging condition.

non-Abelian electrodynamics: An electrodynamics based on a group theory in which the group operation is not commutative and hence not Abellian in nature. Contrast to an Abelian group in which the group operation is commutative.

nonequilibrium thermodynamics: Colloquial term referring to the newer thermo**dynamics** of self-organizing systems and structures, far from euilibrium in their energetic exchange with their active environment, Since the **1960s,** there has emerged in physics (and in thermodynamics!) this surprising new area: the physics of far-from-equilibrium systems and the self-organization and large-scale ordering of such systems. The equilibrium conventionally referred to is 3-space equilibrium, with no consideration of temporal equilibrium, although some pioneers of the field such as Prigogine are quite aware of the temporal disequilibrium also involved. Note that such "3-disequilibrium" thermodynamical systems are also in temporal energy flow disequilibrium as well as 3-disequilibrium, but overall are still in 4-equilibrium energetically. See **giant negentropy (of the dipole).** For a necessary thermodynamics discussion, see Appendix A in this book.

nonlinear material (optical): A substance whose optical properties vary with the intensity of the light.

nonlinear optical functioning: Functioning in a manner prescribed by nonlinear optics and quantum optics, particularly with respect to time-reversal effects, use of phase conjugate mirrors, phase conjugate reflection, wave-to-wave interaction, and pumping for amplified phase conjugate reflection. To quote Knight, "A phase conjugate mixing device has remarkable optical properties: perfect retroreflection, the complete cancellation of optical aberrations and inhomogeneities in wavefronts, and the ability to reverse an optical beam and return it to its source." A most interesting modern science, eventually adaptable to COP>1.0 systems, is developing in the study of intensely scattering optical materials.

nonlinear phase conjugate optics: See **phase conjugate optics.**

nonlocalization: Producing or rendering an effect or effect-producing agent at a distance, rather than at the immediate location of the effector. Extended throughout a large region or system, or to a distant part of a large region or system, rather than being confined to a small region of the large region or system. Nonlocalization effects are absolutely prescribed in quantum mechanics, and are experimentally confirmed. They have been largely ignored in classical electrodynamics and electrical engineering. Nonlocalization can result from the existence and use of a multiply connected space, as in Bohm's quantum potential and hidden variable theory. Engineering of action at a distance by engineering the quantum potential through multiply connected space gives one method of energy amplification. This method has unfortunately been secretly weaponized In five nations of the world, and none of the technology is available for use in electrical power systems for energy amplification and solving the energy crisis.

non-Riemannian geometry: Riemannian geometry is the geometry **of** Riemannian manifolds — i.e., manifolds having an infinitesimal squared distance *ds* that is everywhere positive. A non-Riemannian geometry violates this property. [A manifold is a topological space with certain prescribed characteristics.]

nucleons: A collective name for the neutrons or the protons in a nucleus. In the nucleus, the neutrons and protons are continually changing one into the other by exchange of the charge. What is called a neutron and what is called a proton is thus somewhat arbitrary, so long as one makes a rule and stands by it. The interesting thing is that, in a nucleus, there is already continous and highly rapid transmutation activity between nucleons, all the time, without giant accelerations and operator input of very intense spatial energy. In short, nature shows in all matter that matter can be and is continually transmuted from instant to instant, and is "constant" only on the "**average".** Since most present models used for this are statistical rather than causative, it strongly suggests that we should look for a causal mechanism. The use of the time-reversal zone and time-energy has been nominated in this book as providing a causal mechanism for transduction, which can be directly engineered. It appears that cold fusion research already accomplishes transmutations using the suggested mechanisms, Several specific nuclear transmutation reactions applying the new mechanisms, and consistent with hundreds of cold fusion experiments, are given in the present book.

O(3) electrodynamics: An electrodynamics extended from a two-dimensional inner group symmetry to a 3-dimensional inner group symmetry. O(3) electrodynamics also becomes an important subset of Sachs' unified field theory. The inner group symmetry of O(3) can be taken as normal space, which immediately opens up engineering the vacuum and using spaetime curvature engines to directly engineer matter for us.

open system: A system that communicates with its environment, and exchanges energy, matter, or both between system and environment. With the possible exception of a few theoretical or hypothetical systems, all systems in the universe are in fact open systems. An open system may be in equilibrium with its active environment, so that it cannot accept, store, and utilize any excess energy from the environment. Another open system may be in disequilibrium with its active environment, in which case it can accept, store, and utilize excess energy from the environment. When the operator inputs EM energy into an open equilibrium system, he automatically breaks its equilibrium condition with the external active environment, decreasing the system's entropy. This is an excited state, and the system will subsequently decay back to equilibrium (maximum entropy) condition by emitting or dissipating its excess energy. The reader interested in the thermodynamics should read Appendix A in this book, since the thermodynamics definition (or misdefinitions!) of "closed system" actually permits the system to be open with respect to energy exchange with its active environment — either in equilibrium or disequilibrium — and only prohibits exchange of mass across the system boundaries. By that definition, a "closed system" can be far from energy flow equilibrium with its external environment, and hence permitted to perform COP>1.0 or even COP = ∞. Hopefully our discussion in Appendix A will clarify the position we have taken on this matter.

optical switching: Switching of optical signals, or switching using optical frequency switching techniques, components, and circuits.

orthodox transistors: Ordinary transistors in use today, i.e., as contrasted with the Fogal semiconductor and its special functions.

orthogonal rotation: Simply speaking, a rotation from one direction to a direction at right angles to it. The Lorentz transform of special relativity, e.g., is a formula for the rotation of an object and all its parts away from its line of motion, as a function of the object's velocity, in a direction toward the time axis. The speed c just represents a full orthogonal turn, so that the projection intersection of the object with the observer 3-space, along the original line of motion, is zero.

outfolding: Translation by a system of received input longitudinal EM waves entering inside its potentials and fields, into conventional EM waves present in the system and then output from it. The opposite *of infolding.*

overpotential: Essentially the overpotential (as in an electrode in a solution) is a shift in the Fermi level necessary to allow the electron in the electrode metal to have energies overlapping with vacant acceptor levels in molecules adjacent to the electrode in the solution. The overpotential effect appears at small "double surfaces" of things in very close contact, particularly in electrochemistry and in solid state physics. The overpotential enables electron transfer, e.g., by tunneling. The overpotential increases with the log of the reaction rate that occurs at the overpotential location. For a small reaction rate, the overpotential is small; but for a small increase in the overpotential there occurs a dramatic increase of reaction rate. Fogal has utilized the overpotential effect in his semiconductor, to great advantage in accomplishing unusual charge blocking and other special effects. See **tunneling.**

overunity coefficient of performance (COP>1.0): More energy out or work out of a system than the operator or experimenter must furnish **and**

input. The excess energy is input from the active environment. See coefficient of performance.

parity: Deals with studying a system of a sequence of events as if the events were reflected in a mirror. A system has parity if it undergoes no fundamental operational change if replaced with its mirror-reflected twin.

partially time-reversed region: A region of spacetime in which an object temporarily experiences not only the ordinary forward flow of time, but also an extra induced backwards flow of time, usually a fraction of the time- forward time stream. It appears that the combination of the two is what constitutes time dilation in relativity. In a mixed time-forward (photon interactions) and time-reversed (antiphoton reactions) zone, if time-forward predominates, the law of attraction and repulsion of charges is partially reversed, leading to overall reduction of the forces — which remain in their usual directions. If the time reversal predominates, the law of attraction and repulsion is reversed, and in that region like charges attract and unlike charges repel. If time reversal and optical pumping are combined, the reversed law forces of attraction of like charges and repulsion of unlike charges are amplified also.

particle physics: The branch of physics using accelerators to study high-energy particle collisions, to determine properties of atomic nuclei and of the elementary particles themselves. It appears that conventional particle physicists do not recognize or consider time-energy interaction and time-reversal zone nuclear interactions at all. Hence the present "high (spatial) energy" physics is a far lower "total" energy physics than is the physics where time-energy is deliberately utilized. When time-energy is utilized, then the total energy of the photon varies inversely as a function of its frequency rather than directly as a function of its frequency. Lower frequency photons have very much greater total energy than higher frequency photons. For practical electrogravitation, therefore, one is interested in lower frequency bandwidths such as the ELF region.

Patterson effect: The intense multiscattering, multipass, multicollection of energy density flow in an assemblage of palladium-clad spherical bead capacitors, as the palladium adsorbs hydrogen ions from the fluid in which the beads are immersed, and thereby charges up as a capacitor, becoming an energy-flow generator, and nonlinearly increasing the energy-flow as the charge of the capacitors increases.

Patterson Power Cell®: Patented overunity power unit by James Patterson, which has been independently validated to yield COP>1.0. However, continuing difficulty has been experienced with the palladium metal used in the effect. See James Patterson, "System for Electrolysis of Liquid Electrolyte," U.S. Patent No. 5,372,688, issued Dec. 13, 1994. Patterson has several other patents in the area.

perpetual motion machine: The rather archaic and dogmatic assertion that no machine can indefinitely operate, since no closed system can input its output power back to itself while powering a load, because powering itself and its load violates the *equilibrium* thermodynamics (i.e., the thermodynamics of a closed system). Every system is already an open system in a violent energetic exchange with the active vacuum, so the pronouncement is a non sequitur assuming a false straw man — that every machine is a truly closed system or is in energy flow equilibrium with its active environment and can be treated as if it were a closed system. (The thermodynamic definition of "closed system" is misleading; it refers to a system where no mass exchange crosses the system boundary, but energy flow exchange can and does freely cross it. See Appendix A to this book for a proper discussion). Further, the equilibrium thermodynamics cited does not necessarily apply to either an open thermodynamic system far from energy equilibrium, or to a closed thermodynamic system in energy flow disequilibrium, as is now well known. After all, a common windmill will operate so long as there is a wind or until something breaks. But since it involves a mass flow across the system boundary, it is said to be a true open thermodynamic system. In electrodynamics, a closed system (with no matter exchange across its boundary) can yet have copious energy flow exchange across its boundary, and thus be far from EM energetic equilibrium with its external active environment. The charges and the dipoles in matter operate in a free "time-energy flow wind" and virtual particle flux wind in the vacuum, which are constant and unfailing, and thus charges and dipoles have been freely and continuously outputting real EM energy flow into 3-space for some 14 billion years. All assertions that overunity EM systems are perpetual motion nonsense are completely refuted by every charge and dipole in the universe. To repeat such gibberish only reveals the ignorance of the person making the improper assertion.

Ironically, those dogmatists fond of hurling perpetual motion labels at overunity researchers already recognize the overunity legitimacy of open systems far from thermodynamic equilibrium with their active environment; they refer to such systems *as false perpetual motion machines* or *fictitious perpetual motion machines.* All of them also accept the classical electrodynamics assertion of the source charge, from which all EM fields and potentials and their energy come. So the very dogmatists so confidently pontificating that perpetual motion machines are ludicrous, themselves already implicitly accept that every charge and dipole in the universe is precisely such a perpetual motion machine, freely and continuously creating energy right out of nothing. No legitimate free energy researcher in his wildest nightmares has proposed such perpetual motion machines on such a universal scale. The greatest perpetual motion nuts on earth are precisely those pundits so confidently castigating the notion that an electrical system can output more energy than the operator himself must input. We will have a modicum of tolerance and respect from them, when they can demonstrate how they can purge all those "perpetual motion charges and dipoles" -—and all their EM fields and potentials— from their own accepted electromagnetic model. Obviously if they do that, they simply destroy all EM energy in space, in systems, and in the entire universe. This is the absurdity of the position these pundits advocate so ignorantly.

phase conjugate: Of an entity: its time reversal. For waves: often referred to as *wavefront reversal.*

phase conjugate mirror: A nonlinear material which, when provided with an input wave, emits a phase conjugate replica (time reversed replica) of that input wave. The input wave is by practice called the *signal wave.* The time-reversed wave output is by practice called *the phase conjugate replica.*

phase conjugate optics: That portion of nonlinear optics dealing with multiwave mixing, formation of time-reversed waves, phase conjugate reflection (PCR), phase conjugate mirrors (PCMs), pumping PCMs to amplify their emitted PCRs, nonlinear interferometry, distortion correction, self-targeting, laser tracking, adaptive optics, and so on.

phase conjugate replica (PCR): Of a wave: the time-reversed duplicate of a signal (input) wave, after its detection by interaction with charge. We point out *there may still be some foundations difficulty, however, with the form in which the phase conjugate replica wave exists in spacetime prior to its interaction with charge and prior to its observation (observing the effect of its interaction with charged matter).* Indeed, there is very serious foundations difficulty with the very notion of an EM transverse wavefront in space, traveling at the speed of light, for no such thing exists. See Robert H. Romer, "Heat is not a noun," *Am. J. Phys.,* 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of AJP of the concept of heat in thermodynamics, where heat is not a substance, not a thermodynamic function of state, and should not be used as a noun. In endnote 24, p. 109, he also takes to task *"...that dreadful diagram purporting to show the electric and magnetic fields of a plane wave, as a function of position (and/or time?) that besmirch the pages of almost every introductory book. ...it is a horrible diagram. 'Misleading' would be too kind a word; 'wrong' is more accurate. " "...perhaps then, for historical interest, [we should] find out how that diagram came to contaminate our literature in the first place."*

phase conjugating action: In nonlinear optics, the process or action of producing a time reversed replica of an input wave._For a particle, producing its antiparticle. For a living cell, dedifferentiating it back to some previous form, genetics and all. In the presence of a predominately time-reversal zone, however, the cell is in a predominately phase conjugating mode *a priori*. Phase conjugating it a second time converts it from dedifferentiating to redifferentiating mode. We hypothesize that these combinations are the actual mechanisms generating cellular dedifferentiation and redifferentiation in the cell, tissues, and body. A future book will address that issue; meanwhile, see the 119-page "porthole briefing" on the author's website, www.cheniere.org.

phase conjugation: In nonlinear optics, the novel nonlinear mixing of waves which generate an output wave — called the phase conjugate replica or time-reversed replica — that precisely retraces the path previously taken by the input wave that stimulated the action.

phonon: A quantum of the lattice vibration energy in a crystal. This mechanical vibration energy is considered a field, and the phonon is the quantum of that mechanical energy field. Lattice vibrations in a crystal thus are characterized by waves in the lattice, where the waves are comprised of ordered phonons. Interesting effects occur between electron currents and phonons, due to electron-phonon interaction.

photon: In physics, the basic action quantum. Conventionally considered to be the basic quantum of the EM field. There is a foundations problem with the latter concept, because of the dichotomous use of "field" for both the *effect* force field in 3-space after observation (detection) and the *causal* force-free field in 4-space prior to observation (detection). Strictly speaking, the photon is the quantum of the causal 4-space massless **field,** since the photon possesses both spatial EM energy and time increments. In curved spacetime, the photon may possess mass because the time flow is a mix of time-forward and time reversal, and thus cause and effect are slightly intermixed as well. We point out that, in a negatively curved spacetime, the photon mass is negative, the photon energy is negative, **and**

the photon fields are negative energy fields. Presently the negative energy photons do not appear to be utilized in physics except haphazardly if at all. This may represent a large class of photon interactions with matter that have been inadvertently omitted.

ping-pong: Colloquial term for iterative retroreflection.

pinned electrons: Electrons held stationary by a magnetic field, or other set of forces that react to "push them back into their positions and hold them fast" whenever they try to spatially move away.

pinning: The act of restricting (by restoring forces and fields) spatial movement (particularly of electrons) due to a disturbing magnetic field or other external force.

polaron: An electron in a substance, where the electron is trapped in a potential well due to polarization charges on surrounding molecules. The exction in a substance is analogous to an exciton in a semiconductor.

positron: An antielectron (time-reversed electron), possessing a positive charge and positive mass. *Mathematically* an empty *negative energy electron* state in the Dirac sea whose usual occupying electron has been removed, leaving a four dimensional "Dirac sea hole," may also be regarded as a positron for many purposes, at least *after observation resulting from interaction with matter,* so long as the remarkable difference of the Dirac 4-hole prior to observation from the Dirac 3-positron a**fter** observation is understood. Mathematically, the lattice hole previously occupied by an electron can be considered a positron-equivalent (i.e., as if observed), again so long as the real differences are kept in mind, between a Dirac hole *before* observation and what results after its observation by interaction with mass. This is important, because otherwise one is again substituting the effect for the cause. In addition, if the two are confused, **then** one also arbitrarily discards the negative energy fields and negative energy potentials of the source 4-hole in the vacuum prior to observation. That effectively discards anti-circuits and the ability to engineer practical electro antigravity.

power: In physics, the time rate at which work is done (at which *the form of energy* is changed). In electrical engineering, the term "power" is erroneously but ubiquitously also used to loosely mean "energy." Hence such odd and improper expressions as "draw power from the source," which is a non sequitur. Nonetheless, such non sequiturs are so universally utilized in electrical engineering that they have become a part of the literarure itself.

Prigogine, Ilya: Noted Belgian chemist and Nobelist, and one of the pioneers of the thermodynamics of open systems far from thermodynamic equilibrium. Prigogine was awarded the 1977 Nobel Prize for his work on the theory of dissipative structures in nonequilibrium thermodynamics.

Poynting, John Henry: An English physicist who lived from 1852-1914, whose contributions to electromagnetics included the energy flow theory now known as Poynting flow. He contributed the Poynting theorem, which states that the rate of EM energy loss in a specific region of space is equal to the sum of the dissipation rate (heat loss) flowing across the boundary of the region. Concurrently with Oliver Heaviside, Poynting conceived the notion that energy flowed through space, which previously had not appeared in physics. Heaviside actually published first, but obscurely; Poynting published in a prestigious journal, so the theory came to be known as "Poynting theory."

Poynting energy: A loose term (not rigorous!) used to loosely describe the flow through space of Poynting energy density S = ExH, which is that component of the total energy flow in space that will be *intercepted and collected* by the intercepting charges, circuit, etc. It actually describes not the energy flow density *per se*, but the component of it that would be intercepted by a unit point charge, or that actually enters an intercepting circuit. Poynting started with the assumption of the component of the energy flow that enters the circuit, *a priori.* Only the work of Heaviside considered the remaining component of the energy flow that is not intercepted by the charge or by the circuit. Because of the startlingly great magnitude of this remaining nondiverged Heaviside component, Lorentt integrated the entire (Heaviside) energy flow vector around a closed surface surrounding any volume element of interest. This discarded the entire nondiverged Heaviside component, and removed any necessity of considering or explaining how every generator and battery already pours from its terminals up to 10^{13} times as much EM energy as the presently accounted energy input to the shaft of the generator or the chemical energy available in the battery. We believe that the discarded and unaccounted Heaviside nondiverged energy flow component accompanying every field-particle interaction is responsible for the puzzling excess gravity holding the arms of the spiral galaxies together. In short, we believe it is the solution to the "dark matter" problem, now often believed to be a "dark energy" problem.

Poynting energy density flow: A somewhat more rigorous term used to describe the flow through space of the Poynting energy density flow

$\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$ component that would be intercepted by a unit point charge or an intercepting circuit.

Poynting energy flow: A loose term (not completely rigorous) used to describe the flow of energy density $\mathbf{S} = f(\mathbf{E}\times\mathbf{H})$. Rigorously, both E and H are "defined" only after their reaction with a unit point static charge. Hence S is the resulting component of the overall energy flow that interacts with the charges (with the circuit) and is diverged into the circuit to power it. It is not the entire energy flow, which also includes a vast Heaviside nondiverged component that does not strike the charge and interact with it, and hence does not interact with the circuit but is just wasted. See Heaviside energy flow component and Poynting energy.

Poynting generator: Colloquial term for any dipole or potential difference that represents a broken symmetry in the vacuum's virtual particle flux, and hence extracts virtual energy from that asymmetry, and gates it out as a gushing energy flow in 3-space. When both electric field and magnetic field are involved, the component of the energy flow that would be intercepted and diverged by a unit point charge is the Poynting energy flow S=f(ExH).

Poynting vector: The vector S, given as $\mathbf{S} = f(\mathbf{E} \times \mathbf{H})$, used to represent the Poynting (intercepted and collected by a unit point coulomb) energy density flow component. In Heavisidc's version, $\mathbf{S} = \mathbf{E} \times \mathbf{H} + \mathbf{G}$, where G is an energy circulation flow term of trapped EM energy flow. Heaviside later realized that such a closed energy circulation flow exhibits substantial gravitational effects; in modern terms we would say that it represents an appreciable local curvature of spacetime. Unpublished notes on the gravitational aspects were found beneath a floorboard in his little garret apartment, years after his death. We also point out one limitation of the present use of the Poynting vector: It has rather universally been associated only with the flow of *positive* EM energy, and not with the flow of negative energy. Indeed, the concept of negative EM energy is still consufed, with different views and positions, and still in debate without complete scientific consensus.

PPCM: Pumped phase conjugate mirror. A phase conjugate mirror material that has introduced to it appositive or "pump" waves (in other words, *stress* waves). In effect, pumping is "stressing" or "squeezing" ryhtmically. It is also varying a stress potential, which emits longitudinal EM waves, each accompanied by a correlated scalar (time-polarized) wave. In the phase conjugating process, up to all the energy in the pump waves can be output as the energy in the emitted phase conjugate replica wave. Thus the PPCM may be viewed as an *amplifier,* very similar to a triode. When the time-components of the photons in the signal wave input and also in the pumping waves are considered, then there is a functioning of the pumping also ongoing in the time-domain. There is a rhythmic variation of stress in the time-domain also, which creates the correlated time-polarized EM wave accompanying the longitudinal wave from the 3-space pumping.

So the time-reversal aspect (formation of the phase conjugate replica wave) may be thought of as a rhythmic time-stress wave (pumping or stressing in the time domain) which, by alternately decreasing and increasing the time potential of the mirror, produces two temporal force effects. During the decreases in temporal potential of the mirror, the time-force from the input wave and from the pump waves is forward into the mirror material, allowing the input and interaction of these waves. During the increase in temporal potential of the mirror, a reversed temporal force from the mirror back to the past (temporal potential earlier in 4-space) of the input "signal" wave. The pumping energy (excess temporal potential energy) adds the gain to the phase conjugate replica wave,

probability: The probability of a simple event is the ratio of the number of times it occurs to the total number of trials (for a large, essentially infinite number of trials). Note that "trials" constitute things that have occurred (are past, have been observed). Probability has resisted rigorous logical definition by Aristotelian logic; the reason is that it is an expression of the fourth law, the identity of opposites. Probability is an expression of the future (that which has not occurred) in terms of the past (that which has occurred). Consider *throwing a die to land with one face up.* Thinking of the event (as if it has just occurred, in terms of the most recent past), one can conceive six ways of looking at it. In other words, one can conceive of six "most recent pasts", however, these have not occurred. We were merely thinking of them as if they had occurred. By the fourth law, the *total* "most recent pasts that have not occurred" is identical to the "most immediate future" that has not occurred, on the common boundary between past and future. If one therefore *collects* all six "most recent **pasts** not occurred" they turn into the most immediate future, by the fourth law of logic. One may argue that only a single one of the events will actually occur. Here one counters with the observation that, when it occurs, it is the past, not the future, and then it has not been collected so as to move **to t**he boundary. In theory, the hidden variable approach makes it possible to directly engineer those probabilities while the Schrodinger wavefunction is still propagating and has not yet collapsed to constitute an "observation."

I.e., it is theoretically possible to engineer Wheeler's "pre-reality," and determine in advance just which of the outcomes will occur.

Puthoff, Harold: Noted physicist active in a variety of fields and investigations including in psychoenergetics, extended electrodynamics, energy from the vacuum, inertia, antigravity, parapsychology, remote viewing, psychokinesis, zero-point energy of the vacuum, and cosmology. We are particularly impressed with Puthoffs cosmological feedback principle. Together with Cole, Puthoff also proved that thermodynamics does not prohibit extracting usable EM energy from the active vacuum. See Daniel C. Cole and Harold E. Puthoff, "Extracting Energy and Heat from the Vacuum," *Phys. Rev. E,* 48(2), Aug. 1993, p. 1562-1565.

QM vacuum: The quantum mechanical vacuum, which models the vacuum as an intense flux of virtual particles, each appearing and disappearing in such extremely short time intervals as to be unobservable individually—although as physics develops, from time-to-time what is observable changes because of increased observation techniques. Violent fluctuations of extremely short duration constantly occur in the QM vacuum flux. Various calculations estimate the energy density (in mass units) as up to 10^{80}-10^{90} grams per cubic centimeter (in mass density units), or even greater. The vacuum is sometimes described as two separate components: (1) the ambient or average vacuum in deep space, which can be taken as a scalar potential having an enormous energy density, and (2) the fluctuations superposed upon the ambient vacuum as local violent potential energy changes. Finally, we note that any EM potential in a system is *a priori* a change to the local vacuum scalar potential.

quantized: Adjective which means (i) reduction of some observable quantity to multiples of some small, indivisible unit, or, (ii) expressed in terms of quantum theory.

quantum: A single "particle" of the quantity *action* (angular momentum). Often confused as a single quantity of energy alone.

quantum electrodynamics: The theory of photons and electrically charged particles and their interactions. The use of the term "quantum" implies that the EM radiation's discrete photon nature is significant, so that quantum theory must be employed.

quantum **field theory:** A quantum mechanical theory in which "... a phsyica1 held is considered as a collection of particles and forces, and observable properties of an interacting system are expressed as finite quantities rather than state vectors." *[Dictionary of Science and*

Technology.] More simply (after Davies, *The New Physics,* 1989): "The theory that describes the quantum effects of a classical system of fields defined on space-time and satisfying various partial differential equations." Particles are taken as quanta of fields similar to photons being quanta of the electromagnetic field. Each type of particle is represented by an operator that obeys certain commutation laws. The theory relates gluon fields and intermediate vector boson fields to the strong and weak interactions, and allows an overall theory of fundamental particles and their interactions.

quantum potential: A special potential added to the Schrodinger equation by Bohm in his hidden variable theory of quantum mechanics. The quantum potential is a multiply connected entity, so that it "occupies" widely separated but multiply connected points, events, or objects. It is also an extraordinary energy amplifier, since any energy input to one of the multiply connected points simultaneously and instantly appears in every other connected point, regardless of distance or location in the universe. In real life, the quantum potential also has a "coefficient of multiple connectivity", so that only a fraction of the energy input to one multiply connected point will appear in the other points of the multiple connection group. The quantum potential has been strongly weaponized by five nations, and quantum potential weapons are the dominant weapons on earth, being more powerful than nuclear weapons. In theory the quantum potential together with engines and antiengines could be used to treat and cure a given disease in all persons on earth simultaneously, or in one nation or area simultaneously. Sadly, engines have been developed to generate diseases in a targeted populace, rather than engines to heal the populace. Russia and Brazil have had the quantum potential weapon for some time, as have two nations friendly to the United States. In 2001, China also deployed the quantum potential weapon.

quark: Fundamental particle theorized to comprise hadrons (particles such as protons, neutrons, and mesons) by combining in twos or threes. There are six varieties of quarks. The difference between a proton and a neutron, e.g., is merely the orientation of a single quark. In a time reversal zone (TRZ) in the electrolyte in cold fusion experiments, the gluon forces can be significantly reduced, so that the quarks are nearly freed. Two protons (H+ ions) can attract together in a TRZ so closely that each enters the reduced strong force region of the other, forming a quasi-nucleus. As the ions in solution outside the TRZ move to cancel the TRZ, it decays and the quasi-nucleus becomes an increasingly excited state, seeking to decay. Preferred method of decay is by quark flipping. So one quark flips, so that the quasi-nucleus has become a quasi D+ nucleus. As the TRZ vanishes and time-forward resumes, the strong force recovers faster than the Coulomb barrier returns. Consequently, the quasi-nucleus of D+ merely tightens into a real nucleus of D+. This is one of the standard transmutation reactions in cold fusion experiments, forming excess deuterium.

quaternion: Expression comprised of the sum of four terms, one of which is real and three of which contain imaginary units, and where the terms can be writtten as the sum of a scalar and a three-dimensional vector.

quaternion algebra: The algebra of quaternions and their mathematical operations. Quaternion algebra is of higher topology than either vector or tensor **algebra.** Maxwell's original 1865 equations are some 20 equations in 20 unknowns, in quaternion and quaternion-like algebra. Heaviside and others reduced the algebra to vector algebra and some four equations; at his death, Maxwell himself was rewriting the equations in his *Treatise* to purge the quaternions, because of the controversy over the difficult quaternion form. The present vector equations taught in university as "Maxwell's equations" are in fact Heaviside's truncation of the Maxwell theory. Going to tensor algebra electrodynamics does not recover the full range of Maxwell's original 1865 quaternion EM theory.

quintessence: Literally, the "fifth essence". In short, *dark energy*. We have nominated the arbitrarily discarded, vast Heaviside nondiverged energy flow component surrounding every interaction of charge with field as the dark energy that is the source for the excess positive gravity holding the arms of the spiral galaxies in place. See dark energy, positive.

reactive power: Electrical power, measured in volts-amps-reactive (vars) that cannot do work as is. For sinusoidal waveforms, the formula for reactive power is $IV \sin\theta$, where θ is the phase angle between the voltage **and the current.**

real hidden dynamics: The dynamics of vacuum engines, or of "internal energy" or "internal work" ongoing in a system or volume of space. These internally organized forces, fields, and energies involve dynamic curvatures of spacetime, and it is this dynamics of spacetime itself that provides **the** continuous work and dynamics of internal energy. See Einstein's second postulate extended; infolded; dimensioning.

real hidden vectors: The "hidden" or "infolded" real vector components, e.g., of a zero vector summation system. The individual hidden vector cannot spatially translate the entire bulk system, but it can interact inside the system to perform *internal work* upon it or upon some particular portion of it. Such a system of a finite number of infolded deterministic vectors is said to constitute a "dimensioned" stress potential. Specifically dimensioned potentials, fields and waves thus can be constructed, so that they carry specific vacuum engines (spacetime curvature engines). The interactions of these dimensioned EM entities may differ from normal EM interactions in chemistry, physics, biological systems in general, cells in particular, etc. The long neglected combination of Whittaker's 1903 decomposition of the scalar potential, together with his 1904 demonstration that any EM field or wave pattern decomposes into two scalar potential functions, has neglected the development and use of the "internal EM" theory and technology for a century. One or more nations has, however, highly weaponized "dimensioned EM fields, waves, and potentials" as evidenced by the peculiar and anomalous results of the decades-long microwave radiation of the U.S. Embassy in Moscow. See dimensioning.

real power (active power): Power, measured in watts, that can do work, including generating heat. For sinusoidal waves, real power is given by the formula $I \cdot V \cdot \cos \theta$, where θ is the phase angle between the voltage and the current. Note that when higher group symmetry EM effects are involved (such as the Aharonov-Bohm effect), then the "normal" behavior of circuit components can be altered dramatically. As an example, in the MEG the secondary (output) coils of the transformer section are caused to exhibit a purely electrical interaction with the electrons and electron currents in the coils themselves. Consequently *the output current and voltage are in phase, in complete violation of conventional transformer theory and conventional assumptions about the interaction of inductors.* So the output of the MEG is real power, and phase angle does not enter into it.

regauging: In electrodynamics, changing the scalar potential or the vector potential, or both. If one or both of the potentials is/are changed so that a net new force is created in the system, that is an *asymmetrical* regauging. If only one potential is changed, a net force results and that is an asymmetrical regauging. If both potentials are changed but carefully selected so that the free force resulting from one change is equal and opposite to the free force resulting from the other, that is a net *symmetrical* regauging, often called "Lorentz regauging" or the "Lorentz condition". We note that longitudinal EM waves are produced very simply by rhythmically oscillating the magnitude of a stress potential, which is naught but coherently oscillating the individual vectors of a set of EM fields that sum to a zero vector resultant. The locked-in stress in T. Henry

Moray's semiconductors used in his radiant energy device tubes and the locked-in stress in the Chung negative resistor thus have to do with the production and use of longitudinal EM waves, time-polarized EM waves, and stress potentials.

reflection: The rebound of light or other energy from a surface.

refraction: The bending of light as it passes from one medium to another or through a density change in a given medium.

retarted wave: An electromagnetic wave that is derived from retarded fields, which in turn are derived from retarded potentials. The retarded wave at a point or region of space has arrived there or exists there as the result of the dynamics of charges and currents that existed at earlier times elsewhere. Compare to advanced electromagnetic wave.

retroreflecting: Precisely reversing the path of a reflected wave or object.

Sachs, Mendel: Noted scientist who has completed the work of Einstein and developed a unified field theory covering the universe from the tiniest part to the entire conglomerate. Since O(3) electrodynamics has been shown by Evans to be an important subset of Sachs' theory, then for the first time we have a higher group symmetry electrodynamics and a unified field theory that is largely engineerable by novel electrodynamic means, using the O(3) model.

Sachs' unified field theory: An extension and completion of Einstein's work that includes the universe and its dynamics from the tiniest part to the entire conglomerate, and includes quantum mechanics, electromagnetics, gravitation, and the strong and weak forces. See particularly Mendel Sachs, *General Relativity and Matter*, Reidel, 1982; Quant*um Mechanics from General Relativity*, Reidel, 1986. Sachs' unified theory provides a great generalization of general relativity, quantum mechanics, and electrodynamics reaching from the quarks and gluons to the entire universe. Much of the model is also engineerable by higher group symmetry electrodynamics such as the O(3) electrodynamics pioneered by Evans and Vigier.

scalar: Characterized by magnitude only. However, with respect to polarization, a scalar photon is a term in use for a *time-polarized* photon, where the EM energy oscillates along the time-axis. Hence a scalar wave, following the same terminology, is an EM wave composed of scalar photons, hence whose EM energy is oscillating along the time-axis. The term "scalar" is used in the sense that no vector in 3-space exists even though a vector (and a variation of its magnitude) exists along the time axis. The effects are observed as an oscillation in the rate of flow of time, hence a "time-density" oscillation. The term "scalar" with respect to polarization implies only that there is no vector component in 3-space.

scalar entity: In ordinary vector analysis, an entity thought to be completely characterized by magnitude alone. Actually this is incomplete; a scalar entity must contain a time-like vector component, if it is to steadily persist in time. What we usually do unthinkingly is assume a continual series of successive observations or *d/dt* operations upon the *kt* nature of a scalar *k.* We conceive these individually frozen 3-space snapshots to "continuously exist" since we can readily recall the "stream" *of d/dt* differentiations in our mind and memory. Since all observation is 3-spatial, then no observable *persists continuously* in time; instead, it *continually recurs* in time, instant to instant. This "iterative successive stream of frozen 3-space snapshots" exists in the continual and rapid recall from our memories. Compare this to the similarity of observing a sufficiently rapid series of still pictures on a screen, where the individual pictures differ by a time sequence taken of some phenomenon. One "sees" the recreation in one's mind of the flow through time of those events and entities so photographed. Yet at any one instant, there is only a single frozen snapshot on the screen, and no motion at all was "physically" observed.

scalar electromagnetics: Colloquial term for the electrodynamics that arises from considering transverse EM waves, longitudinal EM waves, time-polarized EM waves, electrogravitation, superluminal EM signals, interferometry, nonlinear optical functions, time-as-energy, and the infolded electrodynamics inside all usual EM fields, waves, and potentials, In secret superweapon projects, the Russian term for scalar electrodynamics is *energetics.* It is divided into three major fields, depending upon the intended target. Targeted against inert matter, systems, and fields, the same term is used: *energetics.* Targeted against the physical and chemical processes and biological fields and potentials of living organisms, the term used is *bioenergetics.* Targeted against the mind's time-polarized EM functioning and dynamics, the term *psychoenergetics is* used. In the latter, higher group symmetry electrodynamics and extensions to quantum field theory play significant roles. Significant progress and strategic testing of all three branches of energetics has been evidenced worldwide. Progress in the energetics branch was confirmed by a statement by U.S. Secretary of Defense Cohen in 1997. Quoting: *"**Others** are engaging even in an eco-type of terrorism whereby they can alter **the** climate, set off earthquakes, volcanoes remotely through the use of* *electromagnetic waves...So there are plenty of ingenious minds out there that are at work finding ways in which they can wreak terror upon other nations...It's real, and that's the reason why we have to intensify our efforts "* Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

scalar potential: Usually considered synonymous with the electric potential. The concept of potential was developed by Poisson, Green, and others from about 1813 to 1827. Three notions involved in the development of the concept of the potential are: (i) there is the collection of storage of energy in the system, via the collection or storage of some entity therein, and this "potential energy" can be released and dynamically used, (ii) the potential is a scalar function whose spatial rate of change is a vector force, and (iii) the dynamics of the stored energy is suspended by the "storage" or "collecting" entity; in other words it is a sort of "suspended action" waiting to be unleashed. *It requires some action—such as interception and divergence—to release the leashed energy.*

The second notion is false, since no force exists in the absence of mass because mass m is a *component* of force F via $F \equiv dp/dt = (m)dv/dt + (v)dm/dt$. It is corrected by stating that its spatial rate of change *when interacted with charged matter* is a vector force. For the third notion, we note that the concept of a "virtual" particle or particle flux means that it cannot become observable unless some interaction is imposed. So we may handle the "third notion" requirement by assuming the potential in modern terms to be composed of hidden virtual particle flux, or hidden wave flux, in wave-to-wave interaction. Physics tries to define the electrical potential as energy per coulomb, or $\phi = W/Q$. That equation, however, is actually not a definition at all (no equation is a definition), but rather is an algorithm for calculating the magnitude of the excess energy collection/collecting upon a coulomb of charge, when that coulomb is immersed in the potential It in fact is the *reaction cross section* of the potential, not the potential entity itself. This now begins to give us a clue, Potential is *energy* in some fashion, and the term "potential energy" is a redundancy. So the entity must have essentially the same definition as energy! **Potential** energy, collected energy, stored energy, whatever—it is energy first and foremost. How it is collected and what it is collected upon or in, determines the "type" of potential. It also cannot be logically defined as the work expended by it upon a unit charge. The causal entity itself is not the action taken; just as a man is not definable as "how far he has run" in some race.

scattering processes: Collision processes in which an incident particle such as a photon, electron, or proton, of known energy or direction or both, collides with another entity, and emerges from the collision with a different energy or direction or both.

Schrodinger, Erwin: Noted Austrian physicist, 1887-1961, and Nobelist who discovered the wave equation named after him as the Schrodinger equation, and performed other vital work in wave mechanics.

Schrodinger equation: In quantum mechanics, an equation that describes the propagation of waves associated with subatomic particles, or more generally, that describes the evolution over time of the quantum state of a system. The fundamental quantum mechanical equation governing the behavior of wave functions.

self-energy: In a quantum mechanical system, the energy associated with the emission and absorption of virtual particles. In a classical system, the energy associated with the interaction among parts of the system itself.

self-organization: (After Davies, *The New Physics,* 1989): "Spontaneous emergence of order, arising when certain parameters built in a system reach critical values." Simply put, self-organization is the spontaneous formation of patterns of localized order in a system that begins in a more homogeneous state. In short, the system moves from a simpler to a more complex state. Self-organization is one of the permissible features of an open system far from thermodynamic equilibrium with its active environment. Note that the localized orders represent free potential energy formations (free self-regauging) that in theory can then be dissipated in loads to perform free work. Hence the capability of a self-organizing system to perform some free work, or in other words to exhibit COP>1.0.

self-potential (e.g., of the electron): The potential created in and of the charge (e.g., the electron) by its virtual particle flux exchange with the vacuum. Also, an asymmetry in the virtual photon flux of the vacuum, caused by the interaction of the charge (e.g., the electron) with the vacuum flux. From here, we note that the self-potential must consist of real observable 3-space EM energy flow or flows from the electron, because of the asymmetry. However, the charge is not the primary source *per se* **of** the energy flows that establish the self-potential, but instead is the asymmetric gate in the vacuum flux that "gates out" the potential and its energy **flux** The asymmetry in the vacuum flux is the actual *source* or gate. Indeed, in classical terms the energy is received by the charge in the form of converging longitudinal EM energy waves from the complex plane (from the time domain). The spin of the charges of the dipole transduces the absorbed complex EM energy into real observable EM longitudinal wave energy, and pours it out in all directions in 3-space. See discussions under Whittaker and phase conjugate.

self-targeting: Between two entities, where each acts partially or wholly as a phase conjugate mirror, self-targeting is the iterative phase conjugation or "ping-ponging" and rapid convergence of beams between the two PCMs.

S-flow: The Poynting energy density flow S, where S = ExH. Heaviside's equation is given by S = ExH + G. In the latter equation, E is the electric field, H is the magnetic field strength, and G is a trapped closed loop flow of energy density. Note that, from the very definitions of E and H, the Poynting, energy flow represents only that portion of the total energy flow in a spatial region that is diverted by interception and interaction with charges. Heaviside's additional component G, which is not part of the field interactions but a result of the huge noninteracting part of the energy flow outside the interacted ExH component, does appear to be responsible for an enormous unaccounted EM energy flow component that generates the excess gravity required to hold together the arms of the spiral galaxies.

signal wave: In nonlinear optics, the signal (wave) input to a phase conjugate mirror (PCM), that stimulates the PCM to emit a phase conjugate replica (PCR) wave in return. See discussion under distortion correction theorem.

Slepian current: The vector $j\phi$, which in a current loop represents the energy unit area per coulomb of flowing charge, per unit of time, that is collected/collecting from the Poynting S-flow and dissipated in that current loop in its loads and losses.

Slepian flow: "Flow *of collected-energy* density's dissipation" in a current loop, and therefore related to the vector $j\phi$ in the current loop, which represents the energy per unit area per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

Slepian vector: Refers to the vector $j\phi$ in a current loop, which represents the energy density per coulomb of flowing charge, per unit of time, that is being dissipated in that current loop in its loads and losses.

solid-state parametric oscillator: A parametric oscillator circuit employing solid-state devices. A parametric oscillator is a device that oscillates by rhythmically changing a parameter (such as capacitance or inductance, etc.). Simple switching between differing values of the parameter is often used to induce the periodic change of the parameter. One way of close-looping a COP>1.0 EM system is by using a bit of the positive output energy to do parametric switching, in effect converting the system to a self-oscillating energy flow system. The AC flow of EM energy is then much more easily intercepted, collected, and dissipated to power loads without killing the source dipole extracting and furnishing the free flow of EM energy from the active vacuum.

soliton: A "solitary wave" consisting of a wave pulse which propagates without changing its shape and without dispersion.

source dipole: Any dipole, considered from the viewpoint that real observable EM energy continuously pours out of it without any observable EM energy input. There is no true "source" as such in nature, since the energy is not created but only transduced and gated. In the case of the dipole and electrical charge as sources of EM energy, fields, and potentials, the broken 3-symmetry of the dipole initiates an automatic and more primary 4-symmetry, where EM longitudinal wave energy flows into the negative charge of the dipole from the time domain (complex plane), is transduced into normal EM longitudinal wave energy in 3-space, wherein the 3-space energy is absorbed by the positive charge of the dipole, transduced back into time-energy and reradiated back into the time-domain (fourth Minkowski axis). We roughly state that the dipole "emits observable EM energy" into all directions in 3-space as observable EM energy. However, what actually happens is that this *energy 4-circulation process* is what is emitted into and along 3-space, along every radial direction, with the process occurring at every spatial point along that direction. Any charge (taken together with its clustering virtual charges of opposite sign as a dipole) then inserted at any point in space along a radial line, will dipolarly intercept the ongoing process. We will observe the interception by its 3-effect on the intercepting and interacting 3-space charge, but we will not observe the 4-process itself. This deeper examination of the EM field-charge interaction requires a new interpretation of what is meant by "propagation of EM energy through 3-space", which is a non sequitur. Recall again Romer's acid objection **to the** totally illogical picture in every elementary textbook of the so-called "EM wave through 3-space" as a flat x-y plane orthogonal to the line of **motion** z, and where the E-field vector is oscillating along one axis while the magnetic field B is oscillating along the other. That representation is totally false, just as pointed out so strongly by Romer.

spacetime or space-time: The entity represented by the "fused product" of space and time, to give "space x time". Spacetime is also dynamic geometry, in the general relativistic interpretation. We may also consider spacetime to be a giant potential, and also active geometry, and also active vacuum, and also virtual particle flux of the vacuum, and also hidden EM wave flux. If we wish, we can also consider spacetime to be comprised of longitudinal EM waves and impressed dynamics.

spacetime curvature engine: See deterministic pattern and dimensioning.

special relativity: Einstein's theory of inertial frames in which, if two systems are moving uniformly in relation to each other, one cannot determine anything about their motion except that it is relative. Each of the two frames is said to be "rotated" with respect to the other, but not accelerating. The velocity of light in space (the vacuum) is constant, and is independent of the velocity of its source and the velocity of an observer. All the laws of physics are the same in all inertial frames of reference.

spin: The intrinsic angular momentum of a particle such as an electron, proton, neutron, photon, graviton etc., even when at rest, as if it were a top spinning about an axis, but had to spin 720 degrees before it turned "full circle." Spin is quantized, and is always described as a half or whole spin, e.g., - 1, - 1/2, 0, 1/2, 1, etc. A spinning charged particle such as an electron thus demonstrates a magnetic moment, due to the circulation of charge in the spinning. In the nucleus of an atom, the spin of the nucleus is the resultant of the spins of the nucleons (particles comprising the nucleus).

spiral galaxy: A member of the class of galaxies having arms that extend in the form of spirals. About 70% of all galaxies are spiral galaxies: only about 10% of the gravity necessary to hold the spiral arms together can be accounted for by known gravitational sources. This has led to the theorizing of dark (hidden) matter of exotic and unknown form, which would be responsible for the excess gravity. Later it was theorized that some sort of "dark energy" might be responsible for the extra gravity. The present author has proposed that the *unaccounted* enormous Heaviside component of the EM energy flow surrounding every field/charge interaction, and not entering into that interaction, is the missing source of the excess gravity. The dark Heaviside energy flow was arbitrarily discarded by Lorentz in the 1890s, and it is still discarded today by all electrodynamicists and astrophysicists.

Stoney, George Johnstone: Well-known scientist (1826-1911) who first estimated the value of the electronic charge (1874) and was the first to advocate the particle nature of electricity. In 1891, he named the unit of electronic charge the *electron.* In 1896-7 he laid the foundation for Whittaker's 1903 bidirectional wave decomposition of the scalar potential.

string wave or "plucked taut string" wave: The transverse wave on a taut string produced when the string is plucked. The string wave itself stays on the string and does not travel through the slapped external medium.

subspace: In vector mathematics: a subset of a vector space which is closed under vector addition and scalar multiplication operations. In the new approach, we may consider the active ambient vacuum as a giant scalar potential. This ambient potential then decomposes into longitudinal EM waves via Whittaker's 1903 decomposition. One can then refer to longitudinal EM waves deliberately transmitted "inside" that Whittaker "subspace" as "subspace communication". The Fogal semiconductor is able to infold normal EM signals into such subspace signals (longitudinal EM wave signals) in the transmitter, and to outfold the longitudinal EM waves back into normal EM waves in the receiver. Hence the Fogal semiconductor can be utilized to perform subspace and superluminal communication. These capabilities have been demonstrated by Fogal in the laboratory. As this is being written, Fogal is in serious negotiations with several large companies for the funding to get his chips into volume production and onto the marketplace in very sophisticated and advanced equipment representing a great leap forward in technology. There are, however, powerful interests who do not wish the Fogal semiconductor to ever be produced and marketed. We have worked with Fogal for some years, and — once the semiconductors are available in quantity — we will be developing and marketing COP>1.0 electrical power systems utilizing his semiconductors and technology, under license from Fogal.

superconductivity: Property of some materials or metals whereby at very low temperature their resistance reduces to zero. These materials also exhibit many additional properties that are anomalous compared to normal materials. Much research has gone into the attempt to find materials that exhibit superconductivity at higher temperatures, with the goal of reaching room-temperature superconductivity. Room-temperature superconductivity is also just COP = 1.0. EM systems far from equilibrium with the active vacuum, and which exhibit COP>1.0 at room temperature, may involve most of the odd entities also observed in superconducting materials (polarons, lattice holes and hole currents, Dirac sea holes and Dirac sea hole currents, "charging" by Dirac sea hole excitation, vortices, etc.) The energy density in the output section is greater than the energy density input by the operator to the input section. Considering the giant negentropy mechanism and the dipolarity of the input and output sections, the tempic potential of the output section of a COP>1.0 EM system is greater than the tempic potential of the input section. Hence there is a tempic gradient across the entire system, and a tempic force from output to input. This may "sweep" out a great deal of the produced "superconductivity" entities, and particularly may result in a Dirac sea hole current in the vacuum, out of the input section back into the external power source, as a novel output of the COP>1.0 system. Feedback from output to input may also be in the time domain, corresponding to what Gabriel Kron referred to as the "open path". By recombinations and other means, the feedback in the tempic domain may be transduced in the input section of the device to ordinary current in forward time. When the voltage and current in this transduced tempic feedback are properly adjusted to that being supplied from the external power source, the external source can be disconnected and the COP>1.0 system becomes self-powering, while simultaneously powering its load also.

superluminal: Exceeding c, the speed of light in vacuum. Longitudinal EM waves can travel either slower or faster than the speed of transverse wave light in vacuum. See W. A. Rodrigues and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds 0 < v<oo in nature," *Found. Phys.*, 27(3), 1997, p. 435-508; W. A. Rodrigues and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras,* Vol. 7(S), 1997, p. 457-466.

superluminal communication: The movement of intelligent signals faster than the speed of light in vacuum, and the transmission and reception of such signals and extraction of their transported information. Longitudinal EM waves are not limited to speed c, whereas transverse EM waves are. The infolded electrodynamics inside all EM fields, potentials, and waves consists of longitudinal EM phase conjugate wavepairs, with their impressed dynamics. Hence this "subspace" is a superhighway for such waves, including those moving at superluminal velocity. In essence this is a very special kind of "tunneling" which might be referred to as *subspace tunnelling.* The limitation to light speed rigorously applies to the modeled transverse EM waves which are a sort of "bulk upheaval" waves similar to waves on the surface of the ocean. As in the ocean wave analogy, "pressure" waves underneath the ocean may travel very much faster than the bulk upheaval waves on the surface.

When making longitudinal EM waves, the process is imperfect and the resulting longitudinal EM waves will have a transverse residue as well. The resulting "transverse-noisy" LW is called an *undistorted progressive wave*. Undistorted progressive waves may move slower than the speed of light or faster than the speed of light. The degree of transparency of matter to LWs is an inverse function of the "transverse-noisiness" of the LWs.

superposition: The simple linear addition and subtraction of two or more values, states, etc.; one of the key principles in field theories and in the concept of potentials. When the situation is sufficiently nonlinear, however, interaction of waves and potentials occurs instead of simple superposition.

supersystem: A higher level or "metasystem" of interactions between the classical system, the active vacuum, and the active local curvatures of spacetime. Hence the supersystem consists of three interacting components: (1) the EM system and its dynamics, considered in conventional fashion as if the local vacuum were inactive and the local spacetime were flat; (2) the local interactive vacuum and its dynamics, and (3) the local curvatures of spacetime and their dynamics. In component (2), one must consider the Dirac sea, Dirac holes and hole currents, with concomitant potentials and forces impressed on them. In component (3), the time-domain interactions (time-energy, time-force, time-potential, and currents of energy along the fourth Minkowski axis) must be accounted to include negative energy.

All three components of the supersystem interact with each other. In general, the vacuum interactions may be taken from particle physics, and the spacetime curvature reactions may be taken from general relativity. However, this will yield only a rough or "first order" analysis, since it does not produce the tri-reactions of all three components at once, but mostly reactions between pairs of the components. The concept is easy to conceptualize, however, and yields valuable clues and insight in many areas, such as chemistry (including biochemistry), electrical physics, electrodynamics, thermodynamics, and electrogravitation. It is particularly useful in understanding the phenomena encountered in COP>1.0 EM power systems.

symmetrical regauging: In electromagnetics, changing the scalar potential to a different value, and simultaneously changing the vector potential to a different value, with the two changes precisely selected so that the two extra forces (each created by one of the changes) are equal and opposite — thus canceling each other's effect with respect to bulk translation and doing no net *external* work. The two forces remain and can perform internal work upon the system. Note that the stressing effect of the two new forces is not canceled, and that the stress of the system is altered a priori. See regauging.

tantalum capacitor: Electrolytic capacitor with a tantalum or sintered-slug anode, such as a solid tantalum, tantalum-foil electrolytic, or tantalum-slug capacitor.

targeted pinpoints: Those separated points which engage in self-targeting between themselves and some separated signal energy source (as in Fogal's semiconductor). The effect is to sharply narrow the retroreflected source signal energy from the source onto the separated points by means of the mutual ping-ponging. In this fashion much more energy density from the source can be concentrated on the targeted pinpoints. In effect, *phase conjugate shooting* — the return of the phase conjugate reflected scattered energy directly to the original target emitting the signal wave in pinpoint fashion - is evoked. See distortion correction theorem.

tensor: A special mathematical entity which is a generalized vector element of a vector space, with more than three components, each of which is a function of the coordinates of an arbitrary point in space of an arbitrary number of dimensions.

Tesla, Nikola: Famous American emigre electrical engineer, inventor and eccentric (1856-1943), whose numerous inventions fueled much of the electrical 20th century. In essence he produced the AC system, including polyphase systems. Also, Tesla invented radio transmission, not Marconi, according to a U.S. Supreme Court decision that upheld the prior patents of Tesla over Marconi. In his famous experiment, Marconi in fact utilized a variation of one of Tesla's circuits. This does not denigrate Marconi's many accomplishments, but just sets the record straight on who invented radio transmission. The impetus to the development of commercial radio transmission technology and the industry itself came from Marconi.

thermodynamic equilibrium: Stable state eventually reached by an isolated system or one in a well-behaved active environment. One may consider it a "balanced" or "preferred" state. It is important to note that, in particle physics, there can be no equilibrium of a mass system unless the vacuum interchange with that system is considered. Electrodynamics mistakenly continues to omit the active vacuum interaction with a Maxwellian system, hence is seriously flawed in that respect.

thermodynamics of open systems: Systems far from thermodynamic equilibrium with an active environment are permitted to exhibit five "magic" functions. Such a system can (1) self-order, (2) self-oscillate or self-rotate, (3) output more energy than the operator himself inputs (the excess energy being freely received from the active environment, (4) power itself and its load simultaneously (all the energy being freely received from the active environment), and (5) exhibit negentropy.

time: In our view, time is spatial EM energy compacted by the factor c . Time thus has the same energy density as mass, but in the time domain rather than in the 3-space domain. Also, time is not observable *a priori;* mass is an observable. Time persists since it is persistence itself. No observable persists *a priori,* since observation is a *d/dt* operator imposed upon LLLT spacetime, yielding a frozen LLL snapshot which does not even exist the very next instant. When we say "mass exists in time" (or any observable continuously persists), we are unwittingly invoking a very rapid iterative *d/dt* operator repeatedly, and recalling these iterative sequential "frozen snapshots" or observations from memory. One may produce and utilize EM energy flow in the time domain, as in 3-space. Time, being energy and energetic, has tempic potential (energy density), tempic potential gradients, and therefore has "tempic force." The photon is composed of both spatial energy and time-energy as its two canonical constituents. The lower the frequency, the higher the energy of the photon when the time-energy is considered, since the magnitude of the time-component of the photon increases as the spatial energy component decreases, and the energy of the photon in that case increases as c^2 times the increase in time magnitude, while decreasing linearly proportional to the decrease of the spatial energy magnitude. When transduction of time into spatial energy occurs, or vice versa, the energy conservation law is modified to specify the sum of the spatial energy, time-energy, and mass energy.

time-charge, time charging: Since time is a special highly compressed form of energy, then time-excitation or *time-charging* also occurs. Also, since tempic longitudinal EM waves accompany all longitudinal EM waves (per reinterpretation of the phase conjugate half set of the Whittaker decomposition), then irradiating a mass with substantial longitudinal EM wave radiation will time-charge the mass, exciting it with time-energy. There are excited states and internal structuring (engines) in time **charging.**

much like the excited states and internal structuring (engines) in energy excitation of normal charge. When the longitudinal EM wave radiation is removed, the time-charged entity then slowly emits longitudinal EM wave radiation as the time-excitation slowly decays via the giant negentropy process of the charges.

time-energy: Time may be considered as spatial EM energy compressed by the factor c^2. See also discussions under time and time charging.

time excitation: See time charging.

time-flow generatrix (mechanism): Since any observable is a frozen nonpersistent snapshot in 3-space, then the common expression "flow of a mass through time" requires an active generating mechanism to add time to mass, converting it to masstime, which is persistent. Masstime *mt* is as different from mass *m* as impulse *(Ft)* is from force *F*. The flow of a mass through time is generated by the ubiquitous interaction of photons with the mass. This includes all photons, both virtual and observable. Thus for the ambient vacuum, there is an ambient flow of time, since there is an ambient rate of total photon interaction. The "macroscopic" movement of the mass through time is the result of (i) absorptions and emissions of observable-sized photons, and (ii) the continuous integration of virtual photon absorptions into observable-sized photons. The macroscopic flow of time thus has myriads of "sublevels" of infolded time flows and engines. Hence the flow of time has a marvelous, internal, dynamic structure.

time-forward: Refers to the flow of an entity through time in the forward direction, as seen by an external observer. Time-forward flow occurs when (to the observer) there is a predominance of time-forward photons interacting with the body that is moving in forward time. Contrast to time-reversal of an object, as seen by the external observer, which occurs when (to the observer) there is a predominance of time-reversed photons (antiphotons) interacting with the body that is observed to be time-reversed. Pumping a given living mass (in the nonlinear optical sense) with time-reversed EM waves pumps the object in the time domain. That drammatically extends nonlinear optics, since the pumped living object's cells will be observed to slowly reverse back to a previous physical state or condition. Since time-polarized EM waves accompany all longitudinal EM waves via reinterpretation of Whittaker 1903, pumping a living mass (a body) with longitudinal EM waves also pumps it with time-polarized EM waves, time-charging the body which continues the pumping in the time domain after the longitudinal EM wave radiation is removed.

time-forward EM wave: A normal EM wave moving in forward time as seen by the observer. Otherwise known as the retarded wave. The external observer comfortably sees its forward progress as expected.

time-polarized electromagnetic wave: An electromagnetic wave whose vibrations occur in the time domain only. May be referred to in a quantum field theory sense as a *scalar* EM wave. It is a longitudinal EM wave in the time domain. It may also be said to be a longitudinal EM wave in the complex plane, since in Minkowski space the fourth axis may be taken as -ict. The time-polarized EM wave is indicated in quantum field theory, or at least its time-polarized photon is theorized and called a "scalar" photon.

time-polarized photon (scalar photon): A photon whose vibrations occur in the time domain only, also known as a "scalar" photon. This polarization of the photon is utilized in quantum field theory. Neither the scalar photon nor the longitudinal photon is individually observable. However, the combination of the two in the presence of charge is observable as the instantaneous scalar potential.

time reversal: For EM waves, the process of forming the phase conjugate wave. For a particle or a mass, the process of pumping the particle or m;iss with time-polarized EM waves, so that the resident spacetime curvature engine in the mass is amplified and phase conjugated, forming a precise, amplified antiengine which then slowly time-reverses the mass back to a previous physical condition and state. We accent that *time reversing a single object or single group ofobjects* is not the same thing as "travel into the past" so popularized by science fiction. For time travel to the past, the entire universe and everything in it — except the traveler — would have to be time-reversed. That would not seem possible by any stretch of the imagination today! On the other hand, time-reversal of a single thing such as a particle or a wave — or even a group of things such as a group of particles or waves — is not only feasible but is readily achievable. A Dirac sea hole is a negative energy electron, for example, having negative energy and therefore negative mass prior to observation. After observation, it is seen as a positive energy, positive mass electron of opposite charge to the conventional electron and reacting with fields in opposite direction from the conventional negatively charged electron. In short, after observation (interaction with mass) it becomes the positron.

time reversal zone: A region of space in which the majority component of the overall time flow mechanism runs backwards, and hence the fundamental mechanism generating the flow of a mass through time is reversed. The "time running backwards" is due to the predominance of interactions with antiphotons. It follows that, in the time-reversal zone (TRZ), there is a predominance of antiphoton interaction over photon interaction. In a TRZ, the law of attraction and repulsion of charges is reversed; like charges attract and unlike charges repel. Also, induced action by the time reversal forces start *inside and throughout* an affected mass in that zone, originating at every point in that local spacetime zone and moving outward. It thus involves a novel action "from inside to outside", beginning even inside the nucleons, and first affecting the quarks and gluons. Deliberate use of time-reversal zones and their subsequent decay allows a great new class of dramatic nuclear reactions, including simple transmutation and loosely called "cold fusion" reactions, at low spatial energy but very high time-energy. High energy physics has neglected the use of time-energy, opting to go for high spatial energy. Hence the high energy physicists have actually eliminated a much higher energy physics (at low spatial energy so readily examined on the bench far *more ch*eaply) than the science they have long developed.

time-reversed wave: A phase conjugate reflected wave, which is a wave moving in reversed time. Since the external observer "sees" in forward time only - and even then he sees in an iterative series of frozen 3-space snapsholt —he *observes* the time-reversed wave as a 3-spatial wave (series of iterative frozen snapshots) traveling in the opposite direction as its generatrix wavefront (the *signal wave).*

topological: Invariant under transformation by continuous mappings, or related to the properties of geometric shapes and figures that remain invariant under transformation by continuous mappings.

topological properties: Those properties of a geometrical space that are unchanged by continuous distortion of the space.

topology: Mathematics of the properties of geometric configurations invariant under transformation by continuous mappings. In short, the mathematics of the large-scale structure of curved spaces.

transverse (with respect to waves): Wave oscillatory movement perpendicular to the wave's primary direction of principal propagation, as in the movement of a plucked taut string in a wavelike motion traveling down the string but with the oscillations being lateral. Based on Faraday's notion that the field lines of the electromagnetic field were physical material lines (in the material ether) similar to taut strings, Maxwell assumed the transverse wave model in his theory. The entry of longitudinal waves, into the theory can easily be demonstrated by rhythmically and coherently varying the magnitudes of the Lorentz symmetrical regauging condition, producing "stress potential waves" which are in fact what longitudinal EM waves really are. The production of a longitudinal EM wave in 3-space will be accompanied by the simultaneous production of a time-polarized companion wave.

tunneling: In quantum mechanics, a phenomenon in which a particle penetrates and crosses a small region where the potential opposing such passage is greater than the particle's available energy. The phenomenon is thought to be impossible according to classical physics, but occurs nonetheless. Further, the speed of transit through the "tunneling" separation can be superluminal. Tunneling of music between two separated points in a waveguide has been measured at more than four times the speed of light.

A more extended type of tunneling is the travel of longitudinal EM waves infolded inside the normal EM potentials, fields, and waves. Indeed, this is a more fundamental electrodynamics, and it can be shown that interferometry of this "internal" electrodynamics creates the external electrodynamics by a form of longitudinal EM wave interferometry. E.g., see M.W. Evans, P.K. Anastasovski, T.E. Bearden *et al.*, "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *Journal ofNew Energy,* 4(3), Special Issue, Winter 1999, p. 76-78. This is part of the technical basis for the "scalar interferometry weapons" developed by the Russians after World War II. Today, some 10 nations have such weapons, more are developing them, and even the Japanese Yakuza has such weapons.

unified field theory (engineerable): A unified theory of the four forces o| physics — the electromagnetic, gravitational, strong, and weak forces that is not just an intellectual model but which is also engineerable on **the** laboratory bench and in actual physical systems using higher symmetry O(3) electrodynamics as a special subset of Sachs' unified field theory.

unitary Q_E (charge of the electron): Fixed or unit charge q of the electron. This notion is held because q is not decomposed into its component elements, but treated as if it were an indivisible unit, charge and mass and all. The mass of a fundamental particle at rest is quantized, and the massless charge (self-potential) of the particle is discretized as a function of the background potential (virtual particle flux density) in which it is embedded or to which it is exposed. In our view the definition **of a** charge q is $q \equiv m_q\phi_q$, at least to first order. (The ϕ_q may be internally structured with deterministic "engines", so that one electron need not bo identical to another, in its actions in some cases). The *electrical charge* of the "charged mass m_q" should be taken to be the massless ϕ_q component. For an electron, $q_e \equiv m_e\phi_e$. The *electrical charge* of the electron should be taken to be the massles ϕ_e component. We note that the charge is also structured, since ϕ_e or ϕ_q decomposes via Whittaker 1903 into a harmonic setof EM longitudinal phase conjugate wavepairs. Our reinterpretation of this decomposition then yields the solution to the problem of the "source charge" and its associated fields and potentials and their energy. The massless charge ϕ_e is a dipolarity and hence a broken 3-symmetry in EM energy flow, invoking a more primary 4-symmetry in EM energy flow so that EM longitudinal wave energy is continuously incoming from the time domain, being absorbed by the charge q of which ϕ_e is a component, being transduced into real EM energy in 3-space, and emitted in all directions in 3-space at the speed of light, creating the fields and potentials and their energy that are associated with charge q. If a charge (say, an electron) is placed in an additional potential ϕ_1, then the total massless charge of the electron(i.e., its virtual vacuum flux exchange with mass me) is $(\phi_e+\phi_1)$.

As can be seen, the massless charge of the electron (its virtual flux activity) has now changed. The electron is then said to *be collecting and to have collected* excess energy (which may be either positive or negative, depending upon the sign of ϕ_1). Further, either ϕ_e or ϕ_1 or both can be internally and deterministically *dimensioned* (structured in Stoney-Whillaker-Ziolkowski fashion). When its massless charge (its own ϕ_e) potentialissodimensioned,thenthedimensionedelectronissaidto contain or transport an *engine* or a *vacuum engine* or a *spacetime curvature engine.*

vacuum: Space devoid of observable matter. In modern theory, the "empty" space is in fact bristling with very rapid fluctuations of EM energy, remaining in the virtual state. It is also filled with a violent, fluctuating flux of virtual particles, appearing and disappearing so quickly that an individual particle does not persist long enough to be individually detected. Thus the vacuum is extraordinarily energetic, but the energy is in very special form (fleeting violent fluctuations and virtual particle flux). Nonetheless, because it contains enormous energy, the *average* vacuum may be considered a potential. All potentials then are changes to this vacuum potential, or to an intermediate potential which is a change to it. In another sense, the vacuum, potential, and spacetime are all one-and-the-same thing. This means that general relativity's spacetime curvature is ubiquitousand extraordinarily dynamic. The vacuum very near a mass also can act in some cases as a semiconductor, capable of supporting self-oscillation between the mass and the local vacuum. Considerations of the active vacuum and its nature lead to the notion of the *supersystem,* which is comprised of three parts: (1) the physical system and its dynamics, (2) the active vacuum and its dynamics, and (3) general relativity and its dynamics (since physicists persist in separating the notions of spacetime and vacuum). No analysis of any mass system is complete unless an analysis of the entire supersystem and the interactions of the three components, one with the other, is performed.

Every charge is "embedded" in the active vacuum by its potentials and fields, which themselves are spacetime curvatures and alterations of the local vacuum. This is easily seen by $q \equiv m_q \phi_q$, where ϕ_q is a change to the local vacuum potential *a priori.* In the quantum view, the fields and potentials may be taken as organized polarization patterns and dynamics (currents) in the vacuum virtual particle flux.

vacuum engine: Also referred to as a *spacetime curvature engine.* From the particle view: a deterministic internal pattern or *template* deliberately created as a deterministic structuring and dynamics in the virtual flux comprising a scalar potential. It can also be created in, and transported by, the two scalar potentials that comprise a static or dynamic force field, or a moving EM wave in the vacuum. It can also be created in the local vacuum potential, in which case it has a "charge-up" creation time and a "discharge time," exhibiting a sort of "time constant" very much like a capacitor charging and discharging.

In the wave view: the vacuum engine is a deterministic internal pattern or template deliberately created in the Stoney-Whittaker-Ziolkowski internal biwaves and wave products comprising the scalar potential. Again, it can be implanted in the vacuum or in an EM signal as a carrier. The self-energy (scalar potential) portion of any electric charge or magnetic charge can also be a "carrier" that is conditioned *{dimensioned)* with a vacuum potential. The process or act of conditioning any of the carriers with vacuum engines is called *dimensioning* that carrier. Russian energetics used in secret weapons programs refers to the vacuum engines of a dimensioned carrier as the *information content of the field.* Such vacuum engines, e.g., were used in the potentials of the microwave radiation of the U.S. Embassy in Moscow for several decades, inducing a variety of health changes and sicknesses in the personnel of the embassy. Interestingly, all these health changes occurred in the field-free regions of the Embassy. There the field-free potentials existed and were stable and unchanging. Hence a body exposed in that region was steadily exposed to the specific engine being induced, and time-charged with that engine accordingly. The deterministic set of spacetime curvatures comprising this charge then acted upon the cells and body to slowly change it to the damaged or diseased condition. See dimensioning.

vacuum flux exchange: The virtual particle interaction between the active vacuum and a mass, particularly a charged particle.

vacuum's **virtual particle flux:** In the particle viewpoint, the quantum mechanical vacuum identically is a virtual particle flux. Present physics, however, is prone to refer to the vacuum as an implied emptiness *containing* a virtual particle flux, even though that is known to be in error. In our view, energy, vacuum, spacetime, virtual particle flux, and dynamic geometry are all synonyms. One must note that the ageless contradiction of the wave vs. particle view has still not been resolved in physics, even by the duality principle. The duality principle is merely an agreement to quit fighting, and to feel free to consider a thing a particle in one case or circumstance and a field in another case or circumstance, *whichever works.*

vector: In mathematics, an entity completely characterized by a magnitude and a direction.

vector potential: A potential having — at any point in it — both direction and magnitude. The direction, e.g., may be linear, curvilinear, or swirling. The term "vector potential" is often used as the magnetic vector potential A. When it is swirling, the magnetic vector potential A may be said to be the magnetic field B. However, the swirling magnetic vector potential (the magnetic field B) may be "bottled up" and contained locally inside a long tight solenoid, or inside a toroidal coil, or inside a special nanocrystalline material used as a closed transformer core as in the motionless electromagnetic generator. In that case nature freely replenishes the energy in the surrounding space which would have been occupied by the nonrestricted, nonlocalized B-field (curled A-potential) but replenishes that space with an uncurled (linearly flowing) magnetic vector potential A. This in fact at least doubles the available energy flow available, without any additional EM energy input by the operator. The well-known Aharnov-Bohm effect clearly demonstrates this effect, and it is deliberately employed in the motionless electromagnetic generator (MEG) of which the author is one of the five co-inventors. Present dictionaries of physics and many textbooks are confused as to the definition of magnetic vector potential A and its relation to B, many stating that the A is defined by the equation $\nabla\times\mathbf{A} = \mathbf{B}$. However, no equation is a definition; it is an expression of equality of two magnitudes, and not a definition of either side. If we substitute the more stringent $\nabla\times\mathbf{A} \equiv \mathbf{B}$, we see that the $\nabla\times$

operator acting upon the A potential is precisely what a B field *identically is*. Hence the A potential exists in two forms: (1) in curled form as identically the B-field, and (2) in uncurled field-free form as the linearly moving A-potential energy current that has no imposed $\nabla\times$ operator and thus has no curl. This makes total sense with $\nabla\times\mathbf{A}\equiv\mathbf{B}$.

vector field: Conventionally, the assigning of a magnitude and a direction (that is, a vector) to each point in a region of space occupied by a field. In our view the field is in fact a change to the active energetics of that spatial region, such that the region will interact with any appropriate entity (e.g , **a** charged mass) placed at any point in the region and produce a force upon that entity.

vector zero: See discussion under zero vector.

virtual antiphoton: An antiphoton that spontaneously appears and disappears, so swiftly that it cannot be individually observed. It is real while it exists. Coherent integration of a stream of virtual photons, however, adds to produce an observable quantum of energy, which is observable in its effects on a charged particle.

virtual particle: A fleeting quantum particle that spontaneously appears and disappears, so swiftly that it cannot be individually observed, so that it exists only temporarily. It is real while it exists. The virtual particle does not satisfy the usual relation between energy, momentum, and mass because it is under the aegis of the Heisenberg uncertainty principle. The virtual particle can have any amount of energy momentarily, so long as the product of its energy and the time interval of its existence is less than the uncertainty principle's minimum magnitude.

The coherent interactions of large numbers of virtual particles with a **mass** or charge can combine to generate real observable effects (forces). In quantum field theory, all forces of nature are caused by the interaction **of** the mass entity with virtual particles. Interaction of a mass with virtual photons, e.g., is projected to cause all mechanical and electromagnetic forces.

virtual particle flux exchange: Interaction with or exchange of a flux of virtual particles.

virtual photon: A photon that spontaneously appears and disappears in the vacuum so quickly that it cannot be individually observed.

virtual photon flux: See definitions under **flux** and under **virtual particle.** For electromagnetics, the vacuum can often be modeled as a flux of virtual photons. One gets the virtual charges also, since the photon is also considered to consist of a positron-electron pair. That is, the photon is continually separating into a positron-electron pair, whose particles then annihilate each other to form additional photon(s), etc.

virtual photon flux exchange: The mutual exchange of virtual photons, as between the vacuum and a charged mass, or between two colliding masses.

virtual photon flux of vacuum: In our view, the vacuum identically is a flux of virtual particles. For many electromagnetic cases, it can just be modeled as identically a flux of virtual photons, where the word "photons" in general usage refers to both photons and antiphotons.

virtual positron: A positive electron (positron) which spontaneously appears disappears in the vacuum so quickly that it cannot be individually observed. More accurately, in the 4-vacuum, a negative energy electron state that spontaneously appears and disappears so quickly that it cannot be individually observed. A virtual Dirac sea hole.

Whittaker, Edmund Taylor: Well-known English mathematician and physicist, 1873-1956, who made major contribut ons to mathematics and physics. Whittaker dramatically extended Stoney's earlier work of 1897-8 an authored two papers in 1903 and 1904 giving the basis for the theory of scalar electromagnetics (Russian energetics, not including the quantum potential portion). His decomposition of the scalar potential into a harmonic set of longitudinal EM wavepairs, where each pair is a phase conjugate pair, may be reinterpreted to solve the long-vexing problem of the source charge or source dipole and its continuous emission of EM energ*y* in 3-space, providing its associated potentials and fields and their energy. The problem was solved by the present author's reinterpretation in that manner, published in 2000. This was later found to be consistent with quantum field theory as well, by interpreting the observability of the combined scalar and longitudinal photon pair as opposed to the nonobservability of either the scalar or longitudinal photon individually.

Work: The changing of the form of energy. One joule of energy whose form is totally changed in a reaction, does one joule of work in that reaction. However, it is strongly stressed that after that joule of work is perform**ed,** one still has a joule of energy left to use (with which to do additional work). This remaining joule of energy is just in different form than was the initial joule of energy. So one can do joule after joule of work from a single *initial* joule of energy, so long as the energy is not dissipated and lost after each change of its form. In any one single change of form, only o**ne** joule of work can be done from one joule of energy. So energy is conserved through many changes of form, but work is not. To work in COP>1.0 systems, one absolutely must root out from one's consciousness the mistaken notion that only one joule of work can be done from a joule of input energy. One of the places where extra joules of work are performed by each joule of energy is that of strong gradients. Strong gradients are known and recognized to violate thermodynamics. Work is further divided into **internal work** and **external work.**

work, external: The changing of the form of the collected energy of a body or system by dissipating it upon an external load or entity is the external work *done by* a body. The changing of the collected energy of a body or system by performing work upon it is the external work *done upon* a body or system.

work, internal: The changing of the form of the internal energy of a body or system. It may also involve the changing of internal parts of the system.

work-energy theorem: In mechanics, a theorem that the change in kinetic energy of a particle in motion is equal to the work done by all the forces acting upon the particle.

zero-point energy: The energy remaining at a temperature of absolute zero, due to the Heisenberg uncertainty principle and the quantum mechanical theory of an oscillator. Also, the minimum energy of a system due to its quantum fluctuations, resulting from its incessant virtual particle activity. Quantum mechanically, no system of interest (including even spacetime itself) can have zero energy. The zero-point energy of the vacuum is the lowest energy vacuum state, with fluctuations taken into account. Even at low energies, quantum fluctuations continually arise, and result in an incessant, extremely rapid, and violent "jittering" of the energy momentarily present. The minimum energy due to these quantum fluctuations is called the *zero-point energy.*

zero vector: Candidates are: (i) A vector having no length (no magnitude) or specific direction; (ii) absence of any and all finite vectors; (iii) resultant of a system of multiple, finite vectors whose vector sum has **no** specific direction but may have nonzero scalar magnitude.

The second definition differs drastically from the first. That is, we **cannot** say that a vector resultant zero has no infolded magnitude, if we **consider** its nonzero components, all of which have magnitude. (See the discussion under **zero).** It only has no net magnitude for *translating* force. It is, in fact, *a. stress potential,* and it is one with a deterministic internal structure or *engine.* By rhythmically and coherently varying each component of a zero vector system, one easily produces longitudinal EM waves,, each of which is accompanied by a correlated time-polarized EM wave. In vector analysis, the zero vector axiom states that any vector v plus a zero vector is eq**ual** to a zero vector plus v. Note that, being an axiom, this is not proven, but just assumed for any zero vector. We very much disagree with that blanket axiom; it reduces vector analysis to a highly special single case, where all zero vector systems are to be considered equal, and totally "inactive." That is true only for bulk translation. In the real world that need not be true at all; specifically, it is not true for two different zero vectors having different internal structure and dynamics (different internal engines).

There are an infinite number of different zero vector resultant systems, and so zero vectors may differ, if their infolded component "substructure" is considered. In the real world, things that are vectors very often have energy. The energy is infolded and hidden, so that it represents not only (i) a local curvature of spacetime, but also (ii) a structured, deterministic set (a template) of curvatures of local spacetime. Therefore this zero vector system is actually a *vacuum engine.* Any mass system placed in such a potential, will be acted upon in its internal parts by that local vacuum engine; in short, it will have internal work performed on its interior. Further, the zero vector resultant system is a "stress potential," since it does contain "collected internal energy."

So between zero vector resultant systems, the "infolded, hidden energies" differ dramatically in (i) total infolded and trapped (collected) energy, (ii) specific internal action patterns (templates), (iii) general relativistic effects, and (iv) type of engine. Accordingly, the actions the zero-vector system induces in an exposed mass may be designed in advance. This is vacuum engineering. This spacetime engineering. This is the formation and use of engines. It is also one of the bases of Russian energetics.

The'problem is the *vector algebra itself.* In the abstract algebra, a "vector space" Is regarded as an inert thing, where the only actants are the overall net nonzero vectors, and there is no interaction between an overt vector and a covert set of vectors. The vector space is not allowed to have an internal realm, stress, hidden engines, or to interact with the vectors representing the physical system. So vector algebra lacks the capacity to model such higher topology phenomenology. Consequently, for over a century electrodynamicists — thinking primarily of spatial translation — have routinely discarded such "zero vector systems" as if they represented the total absence of any finite vector. *This they do, in the unsuitable* *assumptions of the vector algebra.* This they do not do, in the real world, whether we are aware of it or not.

The common practice in electrical physics of replacing a zero-summed system of nonzero vectors with a vector zero indicating a "complete absence of vectors" is very much incomplete. This present practice should be changed; it has reduced the topology of electromagnetics far too much, and gutted vast, important parts of the theory insofar as an extended electromagnetics is concerned. The vector zero should be replaced with a special organized vector zero system and a dimensioned scalar potential.

The interactions of the "hidden" vectors — the *engines* — should be incorporated. By failing to do this, physics excludes the ability to engineer electrogravitation, spacetime, the virtual state, local general relativity, free energy, effects at a distance, and the probabilities of the states propagated by the Schrodinger equation. It also excludes unification of forces. As another more human example, for every cellular disease condition there is a specific engine resident in the cells. An amplified antiengine for that disease pattern can fairly readily be generated. That antiengine, embedded in ordinary EM fields and potentials, can be irradiated into a diseased body with that disease, and it will time-reverse those diseased cells slowly back to normal. This is actually the basis of the body's own healing system the cellular regenerative system studied mostly by Becker. Unwittingly using this "antiengine" approach, Priore and a team of French scientists provided startling cures of terminal cancers, infectious diseases, clogged arteries, etc. in laboratory animals in France in the 1860s and early 1970s. The work is strongly documented and replicated in the French scientific literature, but the work was suppressed in the early 1970s in the change of the French government. See our "porthole briefing" on our website www.cheniere.org for our extension of the Priore method, for greater ease portability, and rapid treatment of mass casualties.

In the new view, a zero vector may be a system of nonzero vectors that vectorially sum or multiply to zero for spatial translation purposes, but nonzero components may still exist "inside" and dynamically function inside the zero-resultant envelope. This substructure can act on nonlinear or resonant systems. Any or all of the nonzero components may be time varying, or none of them may be. The zero vector system may thus be a dynamic engine and a dimensioned scalar stress potential. In addition, such a vector zero system is considered to be also a scalar stress potential, whose magnitude is equal to the sum of the absolute values of the components. In addition, a separate type of stress potential may be included for the magnitude of the absolute values of the swirl components (torques). Other associated potentials may be included for the system, particularly if the system is in multiple dimensions, if it consists of more than one type of virtual particle flux, if several fluxes intercommute, etc.

zero vector summation systems: Zero summations of real, nonzero vectors, so that the real vectors are still present but do not cause *spatial translation* of the exposed mass or charges. Instead, the zero vector summation has an internal structure and dynamics, hence an internal engine. Also see discussion under zero vector.

Appendix A
Relation to Equilibrium and Disequilibrium Thermodynamics

"The second law [ofthermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or ofcreating order out of disorder. It merely states that such a reversal ofthe naturalflow requires an influx ofenergy... " [Raymond Serway] {697}

"It *is in no way possible, either by mechanical, thermal, chemical, or other devices, to obtain perpetual motion, i.e., it is impossible to construct an engine which will work in a cycle and produce continuous work, or kinetic energy, from nothing. "* [Max Planck] {698}.

"What might appear to be empty space is, therefore, a seething ferment of virtual particles. A vacuum is not inert ***and*** *featureless, but alive with throbbing energy and* ***vitality.*** *A 'real' particle such as an electron must always be viewed against this background of frenetic activity. When an electron moves through space, it is actually swimming in a sea ofghost particles ofall varieties - virtual leptons, quarks, and messengers, entangled in a complex melee. The presence ofthe electron will distort this irreducible vacuum activity, and the distortion in turn reacts back on the electron. Even at rest, an electron is* ***not*** *at rest: it is being continually assaulted by all manner of other particles from the vacuum."* [Paul Davies] {699}

"I assert... thatfor the general theory ofrelativity, i.e., in ***the*** *case ofthe general invariance ofthe Hamiltonian function, energy equations... corresponding to the energy equations in orthogonally invariant theories do not exist* ***at all.*** *I could even take this circumstance as the characteristic feature of the theory of general relativity. "* **[D. Hilbert]** {700}

*"**In formulating the** equivalence principle, Einstein actually abandoned the idea of the gravitational field as a*

> *Faraday-Maxwell field, and this is reflected in the pseudotensorial characterization of the gravitational field that he introduced. Hilbert was the first to draw attention to the consequences of this. ... Unfortunately, ... Hilbert was evidently not understood by his contemporaries, since neither Einstein himself nor other physicists recognized the fact that in general relativity conservation laws for energy, momentum, and angular momentum are in principle impossible. "* [Logunov and Loskutov] {701}

> *"No amount of theory can refute a replicable experiment that contradicts the theory. A single replicable experiment can refute any theory that is in conflict. That is the very heart of the scientific method. "* [Myron Evans, private correspondence].

> *"It only takes a single white crow to prove that not all crows are black. "* [Anon.]

> *"One of the most important principles of physics, that disorder, or entropy, always increases, has been shown to be untrue. "* [BBC News] {702}

> *"One of the most fundamental rules of physics, the second law of thermodynamics, has for the first time been shown not to hold for microscopic systems. ... [The experimenters]... found that the change in entropy was negative over time intevals of a few tenths of a second, revealing nature running in reverse. ...But over time intervals of more than two seconds, on overall positive entropy change was measured and normality restored."* [Chalmers] {703}.

> *"We experimentally demonstrate the fluctuation theorem, which predicts appreciable and measurable violations of the second law of thermodynamics for small systems over short time scales... Our results show entropy consumption can occur over colloidal length and time scales. "* [Wang et al.] {704}.

A.1 Introduction

For excellent background books on thermodynamics matters, at various levels of difficulty, the reader is referred to texts by Gyftopoulos and

Beretta {705}, Kondepudi and Prigogine {706}, Chalidze {707}, and Dugdale {708}.

It has long been known in physics that at sufficiently small levels (e.g., atomic levels) the second law of thermodynamics does not always hold true. Recently, as quoted above, Wang et al. {704} have extended this area of permissible violation of the second law to much more appreciable times (a few tenths of a second) and lengths (the micron scale). Without further discussion, we note that this places regions and times of negentropy well-within circuitry switching times and interaction with fundamental chargedparticles. Hence these experiments are beginning to point the way to practical negentropic devices. When coupled with the demonstrated continous giant negentropy of a charge (considered with its clustering virtual charges of opposite signs), these experiments now strongly support the feasibility of practical overunity EM circuits and systems. The interested reader is encouraged to continue to read the literature along those lines, and stay abreast of developments in that field.

In this book we have presented considerable discussion of time-energy (time as a special, highly compacted form of energy). We have also discussed the three components of the supersystem, including the system, the local active vacuum, and the local curvatures of spacetime. All three componets of the supersystem interact with each other continuously, and such Interactions are not modeled in classical thermodynamics (CTD), classical electrodynamics (CED), or electrical engineering (EE). Hence one should not be surprised that the supersystem model and supersystem analysis involve contradictions (actually, *extensions)* of these three models, which are special cases and approximations only.

All COP>1.0 EM systems result from supersystem component interactions. So COP>1.0 EM systems are in conflict with CTD, CED, and EE since these three models exclude those interactions.

To falsify any absoluteness of the three models considered, one or more solid, demonstrable, replicable experiments are presented to establish validity of the new higher symmetry approach. This is sufficient to negate the conventionally accepted absoluteness of the lower symmetry CTD, CED, and EE models. Scientifically, these experiments suffice to demonstrate that those three models describe "*special-case* domains of nature" rather than full generalizations of nature. It is clearly understood that the three models remain very useful and very often quite sufficient, just as Newtonian mechanics is useful and often sufficient, without use of relativistic mechanics.

The experiments also falsify any interpretation of the three models as "proof that COP>1.0 systems are impossible. Accordingly, many of the accepted statements and assumptions of the three models must be corrected for COP>1.0 research, as must be the conventional interpretation that the laws of nature (as understood based on those three models) prohibit COP>1.0 EM systems. This follows directly from the quotation by Evans at the beginning of this appendix. Experimental evidence of COP>1.0 is abundant and decisive, such as the many hundreds of routine "negative resonance absorption of the medium" experiments as typified by Bohren's paper {709} and experiment and Letokhov's work {710}. Also, every charge in the universe already falsifies all three models, as amply discussed in this book. That fact for at least a century has generated the source charge problem — the problem of the source charge and its associated fields and potentials and their energy, reaching across the universe. The solution to this problem was published by the present author in 2000, based on the experimentally proven and well-established broken symmetry of opposite charges, such as are on the ends of a dipole.

In the extended approach, conservation of energy in 4-space does not require simultaneous conservation of energy in 3-space. The concept of energy conservation in 4-space does not exist in the three models mentioned, which contain only 3-space energy conservation. There exists a "giant negentropy" 4-space energy flow (circulation) mechanism, as we developed in this book, and it is experimentally demonstrated by every charge in the universe. In a classical 3-spatial energy conservation analysis and view, the giant negentropy action may continuously perform useful work in a system, since it already furnishes all EM field energy and potential energy in every system. All 3-space energy conservation is an extra condition imposed upon 4-space energy conservation, and that condition may be violated at will without violation of the general conservation of energy law.

To the three conventional models mentioned, 4-symmetry in energy flow with broken 3-symmetry, will appear to be perpetual motion and a total violation of the more limited 3-space conservation of energy law. It *does* permissibly violate 3-space energy conservation, which is just a special case a priori. But it is not a violation of energy conservation overall, since we are using a more fundamental energy conservation law (4-symmetry) than the 3-symmetry used in the mentioned CTD, CED, and EE models,

In short, the three models must be extended to provide 4-space energy conservation with simultaneous violation of 3-space energy conservation, accompanied by simultaneous time-symmetry energy flow violation. The two broken symmetries — of 3-space energy flow and of time domain energy **fl**ow — thus act as a special extension of the notion of symmetrical regauging, accomplishing it by "equal and opposite-effecting" fields, one in **the** lime domain and one in 3-space.

Ironically it then follows that what have heretofore been conceptually called "perpetual motion machines" with respect to 3-space energy conservation are simply attempts — successful or unsuccessful — to build analogues of a windmill turning in a free wind and powering a load. In short **t**hey are attempts to duplicate in systems what every charge in the un**iverse** already demonstrates unequivocally. *Nature's* CTD, CED, and EE do not prohibit such systems — again, as evidenced by every charge in the universe demonstrating exactly such "perpetual motion" (we prefer the exact term "giant negentropy") and violation of present *human* CTD, CED, and EE. None of the original charges from the beginning of the universe has slowed and stopped its motion, or stopped pouring out EM energy freely, in some 14 billion years or so. And charge may be taken as the most basic Maxwellian system. Further, all EM fields and potentials, and their energ***y,*** comes from their source charges. So those so adamant against perpetua**l** motion machines should first purge their own three models of their *assumed* perpetual motion machines. If this is done, they will purge the three models of all charges, fields, and potentials — and all EM energy in the **process.** In short, the three conventional models used by those same ske**ptlcs** so loudly condemning "perpetual motion machines" will eat **themselves"** by swallowing their own tail, so to speak.

in short,all three models are easily falsified by every charge in the universe**,** as well as by every EM field and potential, and by the existence of EM energy itself.

Either their models must surrender any pretense at absoluteness and accept being useful approximations only, or else the conservation of energy law **itself must** be surrendered.

Conventional CEM, CTD, and EE have just hidden this appalling fact and convealed the serious flaws in the modeling by ignoring what Sen {711} reffered to as " t*he most difficult problem in classical and quantum electrodynamics.* " We believe it will very difficult to find a single electrical engineering department or professor or text that deliberately acquaints electrical engineering students with the source charge problem and its implications.

As an example, in Chapter 9 we presented an experimentally verified and patent-pending method for a special energy conversion method. This method is proven on the laboratory bench. With this method, a system accepts time-energy and negative energy input, converts it to ordinary positive 3-spatial energy, and uses the converted input energy to close-loop an EM system far from equilibrium. With that conversion process invoked in a governed feedback loop, the COP>1.0 system becomes "self-powering" (an informal term referring to the system freely receiving all its input energy from its active environment). Self-powering, of course, involves the system continuously operating and performing useful work, analogous to a windmill in a wind and powering a water pump. Rigorously, the self-powering COP = oo EM system becomes directly powered by energy from the broken symmetry induced in its supersystem component interactions.

In Chapter 10 we presented the novel concept of the time-reversal zone (TRZ), which uses the extraordinary EM energy density of time to reverse the Coulomb barrier in nuclear reactions in an electrolyte {712}. This enables nuclear transmutation reactions at low *spatial* energy (but at very high *total* energy when time energy is also considered). There are some 600 successful cold fusion experiments worldwide, in multiple laboratories, by multiple researchers, and most of these experiments demonstrate examples of the specific nuclear reactions made possible by the actions of the TRZ. The time-energy anomalies are directly demonstrated in rigorous electrolyte experiments at U.S. Navy research facilities at China Lake, as we have also discussed in this book. So the experimental results are replicated and well proven. Hence there is adequate, replicated experimental demonstration that additional nuclear reactions exist outside the conventionally known nuclear reaction model. Therefore the set of conventional nuclear reaction models also is reduced to a special case domain, with any absoluteness of those models already experimentally falsified. Either that, or we must abandon scientific model where successfully replicated experiments falsify any theory in contradiction to the experiments.

In our proposed solution to the source charge problem {71 lb}, we found a broken 3-symmetry in energy flow while 4-symmetry is maintained. This is a more primary principle of energy conservation. We called this principle "giant negentropy", which is laughably simple to demonstrate experimentally, since every charge and dipole in the universe already demonstrates it. We showed that such giant negentropy is consistent with quantum field theory, particle physics, and a slight corrective reinterpretation of Whittaker's 1903 decomposition of the scalar **potential** The experimentally proven and theoretically justified giant negentropy falsifies any absoluteness of the CTD, CED, and EE models in conflict with it.

Again, this does not mean that one abandons the CTD, CED, and EE models! They are still very useful models in most cases of interest. It means that these present models are still incomplete, and extended models must be developed and used for COP>1.0 EM system research. It also means that none of these three models is valid as a prohibition of the permissibility of COP>1 EM systems.

All EM energy in 3-space comes from the time domain and returns to it. Since all three models exclude time-energy and its transduction into spatial energy, all three are in conflict with the proven giant negentropy model and with the unlimited experimental demonstrations. Accordingly, the experimentally demonstrable giant negentropy of the charge and dipole strongly points out the inadequacy of the three models and the nuclear reaction models, which do not account for time reversal zones, or time-energy and its transduction into a very large magnitude of 3-space energy.

In paragraph A.2 below, we briefly discuss some of the fundamental but permissibie supersystem violations of these older models of CTD, CED, and EE, and provide some comments. It is obvious that all three classical models need very serious updating and extension. We point out that exceptions to present thermodynamics models and "laws" already are recognized{713}, and work on such exceptions is being performed under the heading of extended thermodynamics {714}.

In effect, all three models *exclude* every charge, dipole, or dipolarity in the universe - which therefore excludes all matter, fields, and energy - something completely untenable in modern electrodynamics and modern physics. Since the fundamental nature of the "heat" notion of CTD is electromagnetic field energy, and all EM fields and potentials (and their energy) are excluded if their source charges are excluded, then heat itself is excluded by the three models. That is, CTD is also totally falsified by CED'S exclusion of the source charge's demonstrated giant negentropy, which in effect excludes heat itself. This is a non sequitur of such magnitude that it absolutely requires the complete rework of CTD, CED, and EE.

A.1.1 Featurcs of the Development of Thermodynamics

In the development of thermodynamics, a few key events of interest are listed in Figure A-1 below.

Selected Events in the History of Thermodynamics {715}

1673	Huygens builds a motor driven by gunpowder explosion.
1690	Papin uses steam to move a piston.
1712	Newcomen develops a steam engine.
1723	Stahl popularizes phlogiston and Becher's concepts.
1724	Boerhaave proposes that heat is a fluid.
1733	Bernoulli publishes a text that is the first truly statistical treatment of kinetic theory, and publishes an update in 1738. It is almost forgotten until 1859.
1765	Watts invents a steam engine that is more than six times more effective than Newcomen's engine.
1783	Lavoisier points out the weakness of phlogiston theory with respect to combustion.
1786	Lavoisier and Laplace provide a Memoir on Heat.
1789	Lavoisier publishes his *Elementary Treatise on Chemistry,* which presents the law of conservation of mass.
1791	Prevost publishes a theory of heat and radiation exchange. Richter founds stoichiometry.
1798	Thompson demonstrates the conversion of work into heat, in his *Enquiry Concerning the Source of Heat which is Excited by Friction.*
1799	Davy's experiments seriously challenge phlogiston theory.
1800	Herschel explores the heating ability of different frequencies of light.
1802	Lussac establishes the thermal expansion of gases.
1803	Dalton forms his atomic theory of matter, setting the stage for atomic chemistry.
1804	Leslie shows that light and radiated heat have similar properties.
1806	Young formulates a precursor to the modern kinetic energy.
1811	Poisson develops a mathematical theory of heat, based on Fourier's work. Avogadro presents his law. Berzelius states that electrical and chemical

forces are the same and that atoms are electrically charged.

1819 Dulong and Petit show an empirical law on the specific heat of solids.

1824 Carnot studies the thermal efficiency of machines, introduces reversible transformations, and generates the foundations of thermodynamics in his "Reflections on the Motive Power of Fire".

1829 Coriolis defines "kinetic energy".

1834 Clapeyron presents the first version of the second law of thermodynamics, based on steam engine studies.

1837 von Suerman experimentally verifies Clapeyron's version of Carnot's equations.

1842 Mayer equates the conservation of energy law and shows the equivalence between heat and mechanical energy.

1843 Waterston anonymously publishes an accurate kinetic theory of gases. It is essentially ignored.

1845 Waterston submits a beautiful paper on the kinetic theory of gases to the Royal Society. It is rejected and ignored for more than 40 years, until finally resurrected by the President of the Royal Society.

1847 Helmholtz generalizes the conservation of energy law for mechanical, caloric, electric and magnetic energy.

1848 Lord Kelvin introduces the absolute temperature scale. Joule finishes establishing the exact relation between heat and mechanical work; his work is almost ignored.

1850 Clausius states the second law of thermodynamics.

1851 Thomson derives the second law of thermodynamics based on Carnot's work.

1855 Rankine publishes *Outlines of the Science of Energetics.*

1857 Clausius establishes heat as energy distributed among particles.

1858 Clausius introduces the idea of the mean free path of a colliding particle.

1859 Kirchhoff establishes the thermodynamic foundation of radiation theory, and completes the basic radiation theory in 1862.

1860 Maxwell provides studies in kinetic theory of gases and gives the distribution of molecular velocities. The paper, "Illustrations of the Dynamical Theory of Gases," was read in 1859.

1860 Maxwell shows a discrepancy between experiment and the kinetic theory of diatomic gases. It was not resolved until the early stages of quantum theory.

1861 Kirchhoff advances the notion of the blackbody.

1865 Clausius, Rankine and Kelvin provide the concept and definition of entropy.

1860-1877 Boltzmann and Gibbs provide the foundations of statistical classical mechanics.

1867 Maxwell publishes *On the Dynamical Theory of Gases*, his major work on the kinetic theory.

1874 Thomson points out the reversibility paradox.

1877 Boltzmann defines entropy in terms of probabilities.

1873-1878 Gibbs publishes several important papers on thermodynamics, widely influencing scientists.

1879 Boltzmann establishes the relation between radiation intensity per unit area of a blackbody, and the absolute temperature. Stephan formulates the "Stephan-Boltzmann law" $R_T = \sigma T^4$.

1879-1899 Discoveries of various laws and effects, particularly at the micro level.

1900 Planck proposes the quantum of action and derives the correct radiation spectrum for blackbodies. (He published the definitive paper in 1911).

1905 von Smoluchowski and Einstein independently investigate Brownian motion. Einstein publishes a paper on the photoelectric effect.

1906 Nernst advances his heat theorem, where at absolute zero temperature both heat capacity and entropy change go to zero. This is later recognized as the "third law of thermodynamics".

1909 Einstein corrects Planck's blackbody derivation.

1910	Perrin experimentally validates Einstein's calculations on Brownian motion, providing a strong argument for the existence of atoms.
1911	Rutherford proposes the nuclear model of the atom. Sackur points out the need for an absolute definition of entropy. Nernst's experiments provide strong support for the new quantum theories.
1925	Born, Heisenberg and Jordan formulate quantum mechanics from matrix algebra. Einstein points out that the analogy between quantum gases and molecular gases is complete; photons and molecules have both wave and particle characteristics.

Figure A-l. Key Events in the History of Thermodynamics up to the Development of Quantum Mechanics.

During the early period of science, a continuing struggle between various notions of esoteric fluids and fluid flow occurred. French scientist Lavoisier (1743-1794), regarded as the father of chemistry, is identified with the concept of heat as an invisible fluid, which he named "caloric". The notion was that one drove this invisible fluid — this caloric — out of a material by some action such as mechanical friction or combustion. This produced the "manifestations" of what was loosely called "heat". In short, caloric was considered to be "heat" which was itself an esoteric fluid. In his famous Treatise {716}, published posthumously, atoms were still in dispute. However, Lavoisier listed the known elements, and included light and heat as elements. Asimov pointed out {717} that Lavoisier had "... era*dicated one imponderable fluid, phlogiston, but it was only partly through his influence that caloric, just as false, remained in existence in the minds of chemists for a half a century.* "

Caloric dogmatically existed in the minds of scientists deeply into the nineteenth century, until Maxwell and Boltzmann produced theories that strongly supported the concept of atoms as the building blocks of matter. In the late 1790's, Count Rumford (Benjamin Thomson) showed that the supplay of heat that could be "driven from matter" was unlimited, which totally contradicted the caloric theory {718}. Yet the notion of caloric persisted until Maxwell's theory eventually killed it, once application of that theory showed that heat was comprised of EM energy.

The fundamental ansatz of thermodynamics was formed and solidified before the advent of modern physics, and even before Maxwell's theory of electrodynamics, and before heat was known to be electromagnetic energy. It has unjustifiably retained an aura of "absoluteness" and dogma, even though its foundations concepts are based on outdated notions already scientifically falsified as far as any absoluteness is concerned. As Mallove states {719}:

> *"It is astonishing to realize that many modern conceptions (or laws') in the science of heat— thermodynamics — arose during the nineteenth century, a period of utter confusion about the fundamental nature of heat. How could it have been otherwise, given that the very existence of atoms was still in question! One sees the shakiness of the claim that the laws of thermodynamics had reached a state of 'near perfection' in the twentieth century ...when they in fact rested on this very flawed foundation. "*

During that formative period, the notion of the conservation of energy **was** also being searched for, as if in a thicket of brambles. For some time, any notion of energy conservation was seen as directly contradicted by caloric, and — in formulating the heretical notion that energy is conserved — pioneers such as Mayer {720} were severely castigated.

In 1850 Clausius resolved the conflict between caloric and conservation of energy in a seminal paper {721}.

Mallove {722} explains how the theory developed from there:

> *"In one fell swoop Clausius 'scooped' Kelvin and cast into precise form both the First and Second Laws of Thermodynamics — energy conservation, and the limit of the Carnot efficiency."*

Clausius' statement of the Second Law is: *"It is impossible for a self-acting machine, unaided by an external agency, to convey heat from one body to another at a higher temperature. "*

In 1851, Thomson (Lord Kelvin) stated the Second Law as:

> *"It is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects."*

In 1865 — the same year that Maxwell published his seminal paper in his electrodynamics {723} — Clausius derived entropy and showed the two laws of thermodynamics expressed in the same way as the old caloric theory. Clausius used the concept of "entropy" to represent the *"dissipated potential".* In a public speech with the title "The Entropy of the Universe Tends to a Maximum," Clausius included how thermodynamics requires an eventual heat death for the universe. His view was that

> *"The energy of the universe remains constant (first law), while entropy increases. "* {724}

We stress that the notion of the flow of EM energy through space did not originate until the 1880s, when Heaviside and Poynting formulated it independently. Neither relativity — with its notion of a combined spacetime - nor the Minkowski space had been born. Neither had quantum mechanics, quantum electrodynamics, quantum field theory, etc.

For a succinct and very useful listing of the major developments in the history of thermodynamics and statistical mechanics, one is referred to the work of Biggus {725}.

A.1.2 Inadequate Fundamental Definitions in Thermodynamics

In the view we take in this book, energy is the all-important concept, in all its forms, no matter how esoteric. The interactions of energy with matter must therefore be extended to include all known energy interactions, not just those prescribed by CTD or CED. Consequently, two fundamental definitions of thermodynamics are faulted by the supersystem concept.

First, thermodynamics defines an *open system* as one that is open to the exchange of energy and mass across its boundary. No concept or discussion of virtual energy, virtual mass, time-energy, or multiply connected space effects is included. Further, mass is actually a special form of energy, and hence this definition only prescribes the exchange of energy across the system boundary.

Second, thermodynamics defines a *closed system* as a system with no mass exchange across its boundaries, but heat (energy) exchange is permitted. This definition has been a non sequitur since general relativity was established in 1915, and also since caloric (heat) was identified as energy rather than a thin material fluid that was forced or beaten out of matter by doing work on the matter. Since then, energy and mass are also known to be the same thing, as established by general relativity. Whenever the energy of a system changes, its mass changes a priori. So a system thought of as "closed to any mass transfer across its boundaries" still undergoes mass change due to the exchange of energy across its boundary. Further, the definitions of "closed system" and "open system" are not mutually exclusive. As an example, a system exchanging energy but not mass across its boundary can be classified as a "closed system", and also as an "open system" where normal exchange of mass is blocked but the mass of the system still changes as the systems absorbs and emits energy.

Rigorously, even the energy of a system cannot individually change in 3-space, without first invoking 4-space change! Instead, the "energy x time" — the action — of the system is what moves and changes in 4-space, Observation applies a *d/dt* operator, resulting in a 3-space energy observed as having already happened or been fixed.

The modern view of gravitation is that the energy in mass generates the gravitational field, not the mass per se {726}. Also, mass is an observable, hence it is a continuously recurring 3-spatial snapshot that does not even continuously exist as mass, but alternately as mass and masstime. The assumption of the continuous existence in mass is falsified by the quantum mechanical nature of observation itself; hence all three models — CTD, CEM, and EE — are falsified.

Nonetheless, let us continue to use the conventional notion of "flow of energy" and "flow of mass" for convenience.

The thermodynamic definition of "open system" ignores the state in which mass exists as "masstime". No "change of mass", flow of mass, etc. occurs. Instead, masstime can flow or change, but mass is an observable and **thus** frozen 3-space snapshot at one instant. Time is "outside" any 3-space material system, hence when mass becomes masstime, it already "exchanges across the system boundary". *For a mass merely to continue to exist requires the exchange ofmass across the 3-space system boundary.* Indeed, only "masstime" can be changing, so in failing to consider masstime, the definition is falsified as it stands, in its very assumption of "continuous change of mass" when there is no such thing in all of nature. As stated, a similar consideration applies to energy, which does not **and** cannot "continuously exist in 3-space" at **all,** and cannot *continuously* change. Also, *time is highly concentrated energy* as we develop in this book, and time certainly "exchanges across the boundary" of any mass system. Time may be taken as a sort of "multiple connection" in space, since in theory each point in space is said to "exist simultaneously" in the same time-point.

With quantum potential connection, external energetic processes may be "superposed instantly in multiply-connected space" with one or more points of the "closed system". Hence, if there exist processes in the system to transduce time-energy into 3-spatial energy — as in the well-proven cold fusion experiments — matter changes including nuclear transmutations can and do occur. Then the "system closed against mass motional exchange through 3-space" (which does not even concretely exist) is not closed against mass changes at all.

To eliminate the mass-change effects of energy transfer across the system boundary, the present thermodynamic definition of a *closed system* must be completely recast into what thermodynamics presently calls an *isolated system* - one in which neither energy nor mass is exchanged across the system boundary. No such system exists in the universe, since merely to exist it has to be open to continual energy exchange between time and 3-space domains, mass has to continually interchange between mass and masstime, and energy has to continually interchange between energy and (energy x time). The blunt truth is that *only open systems exist in nature.*

All systems are open systems in their supersystem interactions, and they can and do exchange time-energy and mass changes due to spacetime curvatures and active vacuum changes. Any local spacetime curvature made by an energy change in a system or nearby, also reacts back upon the system from its local curved spacetime environment. It may react in a totally energetic fashion, or it may react to produce mass in the system (as in pair production). In the book we pointed out that the supersystem components external to the system can also interact in the system to eliminate mass (as when a Dirac hole combines nonradiatively with a Dirac **election).**

Finally, the notion of an "open system" as one in which energy and mass both can be exchanged across the system boundary must be extended to include a system which can also exchange time-energy, spacetime curvature energy, virtual particle flux activity and energy, and other supersystem effects across the boundary. Further, the notion that a "closed system"can be open to energy exchange but not mass is totally falsified in modern physics' virtual particle exchange between environment and system. Virtual mass — which is physically real — is continually exchanged across the system boundary. Though each virtual mass particle is individually unobservable, coherent summations are observable.

To properly evaluate the conventional thermodynamic definitions of open system, closed system, and isolated system, simply select and use a physics model containing only one fundamental unit, the joule. Now all basic entities are made of energy and functions of energy. So all basic entities are forms of energy; e.g., both mass and time are special forms of energy. As presented previously in this book, to even "exist in time" a system is continually exchanging and transducing 3-space energy and time energy, across its boundary at an enormous rate. Every system is an open system, a priori, and the thermodynamic "closed system" is a non sequitur. The thermodynamic "isolated system" is revealed as the early attempts to deal with observation; i.e., the observed system as a series of frozen 3-space snapshots (effects), each stripped from a 4-space ongoing process at a single moment by the observation process *d/dt* (LLLT) = LLL. Each snapshot is an isolated system (which is what observation produces). Now we understand that this "isolated system" existed only at a single frozen instant, continually recurring, and never existed continuously in time at all,

So in view of today's physics, the old thermodynamic "isolated system" concept is an early attempt to deal with observation of the system without input or input (without dynamics, hence frozen). "Closed system" is a non sequitur since mass is energy and energy is continuously exchanged across the system 3-boundary in several forms. Both virtual mass and virtual energy are continuously exchanged across the system boundary. It follows that the only type of system existing continuously in nature is the *open system.* The other two "types of thermodynamic systems" are at best only approximations for special cases.

Under modern physics and the new definitions, we conclude that there is no such thing as a truly closed or isolated system continuously existing in the universe. Hence that very definition disappears, *as far as any absoluteness*. It is retained only as a *useful approximation in special cases,* when the more esoteric mass effects and energy effects are intentionally omitted or negligible, and only classical effects are of importance. Any scientist dogmatically objecting to COP>1.0 EM systems on thermodynamics grounds alone, simply reveals his lack of knowledge (or lack of acceptance) of modern physics.

This is similar to the situation existing between Newtonian mechanics and relativistic mechanics. We can still use the Newtonian model for many practical matters, since the relativistic effects are often small and nearly inconsequential. Similarly, the present CTD concept of "closed system" now assumes that kind of relationship to *one class* of open systems in which the more esoteric reactions are negligible or inconsequential. But rigorously, all systems are now open systems, removing a gross non sequitur in CTD which is especially violated by COP>1.0 EM systems including every charge in the universe.

This is a correction to both equilibrium (classical) thermodynamics and disequilibrium thermodynamics. Unfortunately the two disciplines have continued to use the same (old) definitions for closed system and open system. Hence both equilibrium and disequilibrium thermodynamics at present are useful approximations, and their "laws" must not be misinterpreted as "absolute". Thermodynamics remains a very useful model, but further development is also indicated. Indeed, presently such scientific development and extension of the thermodynamic model is being done under the heading of "extended thermodynamics".

A.2 Comments on Problems with Thermodynamics

Quoting, from Kondepudi and Prigogine {727}:

> *"Every system is associated with an energy and an entropy. When matter undergoes transformation from one state to another, the total energy remains unchanged or is conserved; the total entropy, however, can only increase or, in idealized cases, remain unchanged."*

The absoluteness of that statement is falsified by the solution {12} of the long-vexing problem in CED of a source charge and its associated fields and potentials. The observable charge is a charged mass and thus a Maxwellian system. It is also a thermodynamic system, meeting the present erroneous definition of a "closed system" utilized in both equilibrium and disequilibrium thermodynamics. No observable mass crossed the system boundary, although virtual mass continually crosses it.

The charge is "fed" by a more esoteric energy input in disordered form (the virtual particle flux energy of vacuum), yet it outputs ordered observable EM energy spreading radially outward at the speed of light across the universe. The charge increases and sustains the organized energy density of a volume of space depending on the radius *ct*, where *t* is the time since the formation of the charge and *c* is the speed of light in empty space. This continous outflow of observable energy is a continuous and giant negentropy, exhibited by every source charge and dipole. When positive charge, negative charge, polarization of space, and both positive and negative energy are considered, what is spreading is precisely a reordering of the active vacuum — in short, a giant and increasing negentropy. Since the universe is filled with charges, this is a natural process continuously ongoing everywhere in the universe.

Re-quoting Serway (697}:

> *\'The second law [of thermodynamics] does not rule out the possibility of pushing heat uphill, as it were, from a cold object to a hot one, or of creating order out of disorder. It merely states that such a reversal of the natural flow requires an influx of energy... "*

Our comment is that any mechanism by which the system receives extra usable energy in forward time, or is time-reversed, enables the system to permissibly violate any narrow interpretation of the second law of thermodynamics. The second law was formulated in accord with forward time assumed, and with the old definition of "closed system", which erroneously permitted energy exchange but ignored the mass effects of energy exchange. Therefore, the second law is a special case for equilibrium in its energy exchange with its active environment, but effectively allowing its own negation anyway, and it always has. In the new approach with altered definitions, all EM systems are open systems, and they may also be brought into disequilibrium (in excited states or potentialized condition). Thus in theory all systems are permitted to violate the second law, *particularly when broken symmetries in their supersystems are involved.* This is especially true of EM systems, where all the EM field energy and potential energy in the system— even in the matter itself— comes from the active vacuum environment (particle physics view) by means of the broken 3-symmetry of the source charges and dipoles. The condition of 4-symmetry in energy flow exchange together with broken 3-symmetry and t-flow symmetry, is a disequilibrium condition and an excited state. Every charge and dipole already exhibits this state and the giant negentropy condition.

To first order, then, one may say that the present second law assumes no broken 3-symmetry or broken t-flow symmetry in energy flow, while violation of the second law assumes broken 3-symmetry and broken t-flow symmetry.

Quoting Serway {728}:

> *"Thermal equilibrium is a situation in which two objects in thermal contact with each other cease to have any net energy exchange due to a difference in their temperatures."*

Our comment is that net energy exchange — to include time-energy exchange and negative energy emission, absorption, and transduction are not modeled in thermodynamics and are arbitrarily excluded in **the** quoted statement. Since both are exhibited experimentally, the statement is falsified as far as being absolute. Instead, it is a special case and a useful approximation when some of the esoteric energy considerations are negligible.

Quoting Serway again {729} on the zeroth law of thermodynamics:

> *"If objects A and B are separately in thermal equilibrium with a third object, then A and B are in thermal equilibrium with each other. "*

A corollary is stated by Serway {730} as:

> *"...two objects in thermodynamic equilibrium with each other are at the same temperature. "*

Our comment is that these statements are true if the time-energy exchange is identical among all three. Otherwise, they are not necessarily true. It is not generally true, e.g., for a system in an accelerated frame. Also, the zeroth law is not necessarily true if negative energy (and/or its transduction is involved with one or more of the objects.

An example: A and B are at the same temperature as C, as measured by a thermometer. But A is also emitting negative energy, which is being absorbed by B, to maintain A and B at the temperature of C. A and B are not in real thermal equilibrium in their supersystem interactions, even though they are at the same macroscopic temperature. A second example: A is emitting negative energy to maintain its temperature, while B and C are not. The supersystem of A is not in real thermal equilibrium with the supersystem of either B or C, even though the observed temperatures of A, B and C are the same.

Quoting {731} on what is meant by the term "heat":

> *"...heat is treated as another form of energy that can be transformed into mechanical energy. " ... "The word heat should be used only when describing energy transferred from one place to another. That is, heat flow is an energy transfer that takes place as a consequence of temperature difference only."*

Quoting Lindsay and Margenau {732}:

> *"A complete statement of the first law comprises two* ***assertions:*** *(a) heat is a form of energy, (b) Energy is* ***conserved."***

The statement in (a) that heat is simply a form of energy is misleading, and the statement in (b) that energy is conserved must be expanded to include total energy, not just heat and the presently accounted forms of energy. Specifically virtual particle energy, time-energy, and energy from the curvature of spacetime must be included.

More correctly, heating is a form *of energy disordering* or energy scattering. As previously pointed out, entropy itself was and is conceived as *the dissipation (irretrievable loss) of potential energy.* Heat thus would be the process of energy disordering, which means that heat is a form of the process of entropy (loss of control of the energy). Note the strong non sequitur with the notion that heat is energy. *Heat cannot logically be both energy and the dissipation of energy.* The concept of energy (its mere presence) is not identical to the concept of scattering of energy or transfer of energy, but the concept of heat is identical to it. The thing is not identical to something done to the thing or to transfer of the thing to another place. A form of scattering of order is not a form of ordering. Transfer of heat is a transfer of disordering of energy, not the transfer of energy per se. Hence, considering heat as simply "heat energy" is a fundamental non sequitur of classical thermodynamics, and remains so to the present day and in the present textbooks. For an example of strong support of this position, see Romer {733}. Quoting {734}:

> *"The work done on (or by) a system is a measure of energy transfer between the system and its surroundings, whereas the mechanical energy (kinetic or potential) is a consequence of the motion and coordinates of the system. Thus, when you do work on a system, energy is transferred from you to the system. ... one can refer only to the work done on or by a system when some process has occurred in which the system has changed in some way."*

Our comment: Consider a system which "ping-pongs" the energy it receives from the environment, between two parts of itself. In each part, the form of the energy is changed by that interacting part and that does work upon the part. But the new form of energy remaining after the work was done in that component is not lost but is ping-ponged to the other component, where it does work upon that component, and so on. In this case, each joule of energy input from the environment does more than a single joule of work, with the work performed being limited only by the ability of the system to sustain and contain the "ping-ponging". There is work but no entropy if the energy doing the work (changing its form) is not lost from system control!

There is no law of "work conservation" in nature, and a joule of input energy can do more than one joule of work (as in this example) if repetitive changes of its form are considered. Thermodynamics arbitrarily excludes such cases, and hence is incomplete. Again, the conventional perception of the absoluteness of thermodynamics is falsified. It models a special case where the energy can be "used" to do work only once, and then all control of the new form of the energy (it is all still there, just in a different form) is lost.

For a particular substance, the heat capacity C is usually defined as *the amount of heat energy needed to raise the temperature of the sample being tested by one degree Celsius.* This definition assumes only one "use" of the *energy to* perform work in the substance, and completely ignores "multipass,multi-use" of the energy to perform more than one joule of work from one joule of input energy. As an example, if an equal amount of positive and negative energy is steadily input, then no amount of the positive heat energy input component would raise the temperature at all. The modern active vacuum, with a local energy density in the region of 10^{80} or more grams per cubic centimeter (expressed in mass units), nevertheless is harmless to living systems immersed in it. It has equal positive and negative energy densities, and so normally inputs the same enormous amount of positive and negative energy in its overall interactions with a living system.

The first law of thermodynamics is usually written as

$$\Delta U = U_f - U_i = Q - W. \qquad [1]$$

This is a highly special case, which does not include time energy transduction (spatial energy transduced into time energy and vice versa), or energy transduced from the active vacuum via the broken symmetry of an electrical system in continuous exchange. The definitions of work and heat have to be changed and drastically extended.

As an example of a legitimate exception, when time-energy T_E is transduced into available spatial energy inside the system, we have

$$\Delta U = U_f - U_i = Q + T_E - (W_D + W_L) \qquad [2]$$

where W_D is the energy that was dissipated in the system losses without doing useful work and W_L is the energy leaving the system but doing useful work in a load L. $(Q + T_E)$ represents the total net energy entering the system, while $(W_D + W_L)$ is the total energy leaving the system, either as losses or to do useful work in the load. Here we are also incorporating virtual energy under the title of time-energy, since both time and virtual energy are unobservable, and virtual energy "exists in time".

Now suppose Q entering the system, and collected for use in it, is freely furnished from the active vacuum via the broken symmetry of the electrical system's dipolarity. Suppose T_E is spatial energy freely transduced from the time domain by ongoing processes in the system, and thus T_E also enters the system freely. Then suppose that the system processes are deliberately adjusted by the operator so that

$$(Q + T_E) = (W_D + W_L) \quad [3]$$

In that case,

$$\Delta U = U_f - U_i = 0 \quad [4]$$

yet continuous loss energy WD is escaping from the working system, useful work W_L is continuously being done in the load, and useful energy $(Q + T_E)$ is continuously entering the system from the active environment (the active vacuum and the local curvatures of spacetime). This system is totally permissible in nature (and is already operating in every charge and dipole in the universe). It is an example of an open system far from equilibrium in *3-space* energy flow, but in equilibrium in *4-space* energy flow. Yet it is artificially excluded by classical thermodynamics.

Classical thermodynamics thus excludes every charge and dipole in the universe, and therefore is an extraordinarily curtailed model. It is a special case approximation having no absolute authority as a proper limiting model of EM systems.

The first law of thermodynamics for infinitesimal changes may be stated as

$$dU = dQ - dW \quad [5]$$

Following the previous reasoning, for involvement of time-energy transduction (with virtual energy included) we have

$$dU = dQ + dT_E - (dW_D + dW_L) \quad [6]$$

If dQ and dT_E are freely received in the system from the active vacuum and local curvatures of spacetime respectively, and if

$$dQ + dT_E = dW_D + dW_L \quad [7]$$

then

$$dU = 0 \quad [8]$$

So with no change in its internal energy at all, the microsystem or microprocess is receiving excess energy continuously, and also is continuously emitting energy that is doing useful work with some losses. The system is in fact in *energy 4-flow* equilibrium, and doing useful work, without the operator inputting any of the energy.

So the *corrected* standard statement of the first law allows an equilibrium system to steadily produce output work W, if its "energy input Q" is freely received. The first law of thermodynamics, properly extended and reinterpreted, does indeed admit "electrical windmills" operating in free electrical winds.

Now turning to the second law. In effect, the second law prescribes those processed in nature — more exactly, in the prescription of the first law — that may or may not occur. It is a law of limitation. It states that, of all the energy conversion processes permitted by the first law, only certain types of energy conversions can take place. A priori, the validity of its prescription depends totally on the nature of the functions involved in the system considered! However, since we dramatically extended the first law, using neww processes experimentally determined to take place, then we have automatically extended and changed the second law because we used experimentally *proven* functions. The old second law is now revealed as a special case, with no absolute limitations. The perception that it prescribes absolute limitations has been falsified. It is still useful, however, but to be careful one must completely re-examine (and possibly rework) the second law whenever the energy functions of a system are changed. We strongly stress that *the second law is a conditional law, and its application is totally dependent upon the system energy functions behaving in the limited manner assumed by the second law. Whenever the energy functions of the system are changed outside those assumed by the second law, the second law is permissibly invalidated for that system.* In short, in that case the second Iaw is simply a totally inadequate model for that system, and cannot be used as any sort of criterion for stating the "permissibility" or "impossibility" of that system. As an analogy, a model specifying scalars only cannot be applied to a vectorial system as a decisive logical criterion orrestriction.

Quoting Seray {735}:

> *"From an engineering viewpoint, perhaps the most important application [of the second law] is the limited efficiency of heat engines. Simply stated, the second law states that a machine capable of continuously converting*

> *thermal energy completely into other forms of energy cannot be constructed."*

Our comment is that, if we believe that energy is conserved in the entire universe, then the universe itself already violates Serway's statement. If we consider the seething virtual particle flux energy of the active vacuum as a special form of thermal energy, then every charge and dipole in the universe already is a heat engine that falsifies the second law of thermodynamics. The charge and dipole continuously absorb this special form of thermal (virtual flux) photon energy, and also continuously convert it into real, observable photon emission in 3-space. This is negentropy on a giant scale, already demonstrated by every charge and dipole and thus experimentally falsifying the present second law.

Classical thermodynamics also assumes irreversibility. While many of the physics models and equations prescribe reversible reactions, thermodynamics takes the view that events occur with a much greater probability in one direction in time than in the other. This of course is usually borne out in our perception of the gross observable universe around us. However, charge conservation alone requires equal amounts o| charge and time-reversed charge. With the nearly ubiquitous time-reversed charge and charge conservation, it follows that some factor not accounted for in thermodynamics is operating in the universe. And so it is, but it is a general relativistic factor. We shall use a biological example of the cellular healing mechanism we discovered in the living body {736}, extending Becker's work and Priore's work.

Normally most processes do not involve deliberate net transduction of time-energy into spatial energy, nor do we model and account for the specific set of ongoing interactions between every part of the system **and** its concomitant set of spacetime curvatures acting back upon it. Call that set of spacetime curvatures the *resident spacetime curvature engine* ***and its*** *dynamics,* or just *its engine* for short. The engine is comprised of structured energy and energetics, and it is already experimentally demonstrated in nonlinear phase conjugate optics that energy can be **time** reversed. The engine's spacetime curvatures act directly upon the **mass-**energy dynamics of the system. Time-reversing the system's operation is a matter of producing a sufficiently amplified, phase conjugated engine to act in reverse upon the physical system and overpower its resident **engine.** In that case, the thermodynamic system and its functioning can be **made to** run backward.

In fact this is the fundamental mechanism used by the body's own cellular regenerative system to heal its damaged cells. To a limited degree, the body is able to "pump" its cells in the time-domain (with time-energy), thus producing an amplified "phase conjugate energy and engine dynamics".

Pumping a living cell with time-energy is interesting. The resident engine serves as the "signal input wave" in conventional nonlinear phase conjugate optics, and the output time-reversed engine (the amplified *anti-engine*) serves as the "phase conjugate replica wave". A sufficiently simplified anti-engine acts upon the cell and all its tiniest parts, time-reversing all energetic operations in exact phase, proportion, and timing. This is the mechanism for thermodynamic reversibility. It is also the "secret" to eventually reversing aging.

This action is upon all the energetics of the cell and its parts, including the mass-energy and mass-energetics as well.

The result of the action of the anti-engine is to slowly time-reverse the damaged cell and all its parts back to a previous state and dynamics. Electromagnetic amplification of the process was unwittingly demonstrated by Priore {737} and by Becker {738}. We refer the reader to our website {739} and to our "porthole briefing" for pertinent technical information, and for a most important extension and great simplification of the Priore/Becker approach and methodology.

Present classical thermodynamics does not consider or model the reaction of local spacetime curvatures and their dynamics upon the system. Hence all time-domain pumping that occurs is occurring rather randomly in each and every microscopic part of the system. In short, there is no organized anti-engine, and an organized anti-engine is what is required for time-reversing the system's operations back to a previous state. However, by mimicking, the body and pumping the physical process in the time-domain, with sufficient pumping a precise, amplified anti-engine will be formed and act upon the system — *any* system. The applicability is limited only by the state of development of the appropriate time-pumping technology.

In theory (and as experimentally demonstrated unwittingly by Priore and by Becker), this anti-engine can and does overpower the resident engine. It can in fact time-reverse the pumped physical processes in the system itself.

So the classical thermodynamics assumption of irreversibility for heat engine energy processes is not absolute but only provisional. Specifically, it assumes no overall time-domain pumping process that provides a precise anti-engine and time reversal of the system energetic process of concern and its dynamics.

If the second law is extended to encompass and model the new concepts such as broken symmetry and time-energy, then the extended second law will encompass reversible heat engine processes, Maxwell's demons, COP>1.0 electrical systems, COP» 1.0 electrical systems (electrical windmills), time-reversal of the dynamics, reversibility of that which has been irreversible, etc.

As given by Kondepudi and Prigogine, the third law of thermodynamics essentially states {740} that

> "... *at the absolute zero of temperature the entropy of every chemically homogeneous solid or liquid body has a zero value."*

This is also known as Nernst's heat theorem, after W. Nernst who proposed it in 1906. The third law states that

$$S \to 0 \text{ as } T \to 0 \qquad [9]$$

Certain violations of the third law are already known. E.g., C=O: carbon monoxide molecules can align in various ways in a crystal, so there is entropy associated with C=O crystals at 0°K. In this case, A and B are equally probable and the following equation can be used to calculate the entropy (called residual energy):

$$S = nR\,[1/2\ln(1/2) + 1/2\ln(1/2)] = -nR\ln 2 \qquad [10]$$

From the new viewpoint, the definition of T = 0° is open to change. Rigorously, the present third law assumes no local broken 3-symmetry of the vacuum's flux exchange with the system, and it also assumes no broken time-energy flow symmetry. In the new approach, both these symmetries can and are broken by every charge while a higher 4-symmetry is maintained. This leads to a giant negentropy, so that the "calculation of entropy" via Nernst's theorem fails. A fundamental charge at T = 0° thus continues to exhibit giant negentropy, in total violation of the third law of classical thermodynamics. Hence the third law is also devoid of any absoluteness, and it is reduced to a special case statement. It remains useful where the more esoteric energy interactions are negligible.

We have now demonstrated (by citing actual substantiating experiments) that all the laws of CTD are falsifiable insofar as being absolute. Hence they have *no* absoluteness. Instead, they are special case models and "approximation laws" only. They remain very useful and applicable, However, so long as the more esoteric energy interactions of the supersystem do not play a major role in the system operation.

We conclude that the present form of classical thermodynamics and disequilibrium thermodynamics are special case models — as indeed is every model, by Godel's theorem and proof alone — with no absoluteness. They do not prohibit COP>1.0 EM systems, nor can they necessarily be used to rule out COP>1.0 EM systems.

This also falsifies all current arguments that COP>1.0 EM systems violate either the first, second, or third law of classical thermodynamics. Calling COP>1.0 systems "perpetual motion machines of the first, second, and third kind", according to which law they supposedly violate, is a total non sequitur. Indeed, every original charge in the universe has already demonstrated just such "perpetual motion" for some 14 or so billion years, and happily violates all three laws of the present CTD. Every charge is in violation of the "laws" of the present models of CTD, CED, and EE, and it is still in motion and still freely pouring out EM energy. We have no abillity to examine the notion of "perpetual" past the entire past existence of the universe.

All three models examined — CTD, CED, and EE — are seriously out of date and in need of revision and extension, because any absoluteness of their predictions and assumptions is already falsified by actual experiments and examples. Some work is already ongoing to extend thermodynamics, but is not nearly extensive enough. A thorough overhaul of the very foundations is what is needed for all three models. The three models remain useful as special case approximations, but their received use to arbitrarily rule out COP>1.0 EM systems is dogma, not science because such use is already experimentally refuted by every charge in the universe.

Appendix B

Maxwell's Demon: Alive and Thriving

Introduction

ITHERE is a long history of discussion over "Maxwell's demon", a fictitious being that Maxwell proposed to "sort" molecules striking a barrier between two volumes of gas initially at the same temperature. By intentionally opening a trap door between two for hotter, more energetic molecules to pass through from the left side to the right side, but shutting the door for lower lemperature molecules so that they are retained on the left side, eventually a temperature gradient would "freely" exist between the two reservoirs. In theory, then, if one had the demon working "for free", one would be able to utilize the temperature difference between the reservoirs to perform free work. Then one could repeat the demon's sorting procedure, to obtain more free work again. And so on.

The prevailing discussions have ignored the main feature: the fact that the demon is part of the "external environment" (i.e., not under control of the system, and operating independently of any energy input by the operator). Maxwell's demon represents an analogy to the system's environment furnishing free energy (free ordering and sorting). A lot of breath and mental gyrations have gone into the discussions over the years, with very little concrete experimental examination. The prevailing concept in science, which still adheres to the "absoluteness" of the second law of thermodynamics, is that Maxwell's demon is impossible.

Recently, the former editor of *Nature,* Dr. John Maddox, wrote a very clear emeritus editorial on Maxwell's famous demon, repeating the standard conclusion that such is not possible. Since there are specific examples in physics of what can be called "experimentally proven Maxwell's demons," the present author prepared a respectful commentary to *Nature* to refute the standard erroneous thesis in the interest of science. since experiment is supposed to be decisive in physics, we hoped that citing overwhelming experimental proof would be sufficient. It wasn't.

Almost symbolically (i.e., "free the demon!"), I submitted the commentary to the editor by I -mail on the evening of July 4, 2002. The very next day I received rejection (not too unexpected) by E-mail with no further discussion accepted. So I wrote a polite note in response, and then posted my commentary note on my website.

In this appendix, I include my E-mails of both 4 July and 5 July 2002, and then include the commentary that was submitted. The reader can read the cited emeritus editorial in *Nature,* and then decide for himself or herself whether experiment still rules the day at *Nature* where COP>1.0 and COP = co EM systems are concerned.

E-mail to *Nature* 4 July 2002, submitting the Commentary.

Dear Editor:

Attached is a significant commentary on John Maddox's excellent emeritus editorial on Maxwell's demon, titled "Slamming the Door," *Nature,* 417, 27 Jun. 02, p. 903.

Forms of Maxwell's demon are very much alive and thriving, and experimentally confirmed, as we develop very clearly in this commentary. They are unwittingly used widely in electrodynamics, though seldom recognized.

Most sincerely,
Tom Bearden, Ph.D., Magnetic Energy Ltd.

E-mail to *Nature* 5 July 2002, submitting a comment on the rejection.

Dear Ms XXX:

Thanks for the prompt reply.

Since *Nature* will not be publishing the note, I have now posted it on my website (some 80,000 hits per day from all over the world). It can be **seen** at http://www.cheniere.org/articles/maxwells%20demon.htm

This is a respectful article, and I just thought the experimental information needed getting out there to the scientific community. I also understand you cannot discuss it further. I only wish to note that it is based on cited replicable experiments, which are supposed to be the decisive thing in science and scientific method. It seems a bit odd that you would not publish the results of several replicable and replicated experiments **that** falsify a prevailing theoretical thesis and conclusion. Nonetheless, I respect the wishes of *Nature,* and I do enjoy the Journal each week.

Best wishes,
Tom Bearden, Ph.D.

Within days of this rejection, the paper by Wang et al. {741} was published, extending the known physics "Maxwell's demon" regime (i.e., where the second law of thermodynamics is known to be violated) from the atomic scale and very short periods of time, to the micron scale in size and time up to a few tenths of a second.
Here is the rejected commentary I submitted to *Nature.*

Maxwell-type demons are alive and thriving

The recent emeritus editorial by Maddox on Maxwell's demon [742] is a short overview of one historical aspect of would-be negentropy, but much greater thought on the subject needs to be given. Indeed, Maxwell's demon is alive and thriving, but just in a different form not requiring a nimble-fingered microscopic being. We now discuss such experimentally proven demons.

Consider the prevailing ansatz that Maxwellian systems outputting more EM energy than the energy input by the operator or experimenter are impossible. Such proposed systems are said to constitute examples of forbidden perpetual motion machines because they violate the second law of classical (equilibrium) thermodynamics. That prevailing ansatz has been falsified in classical electrodynamics since the 1880s, and in particle physics for 45 years, but eerily it is still the prevailing scientific opinion nonetheless.

The ansatz can be experimentally disproved in classical electrodynamics as follows: Lay a charged capacitor or electret on a permanent magnet so the E-field of the capacitor or electret is at right angles to the H-field of the magnet. That optimizes $\mathbf{S} = f(\mathbf{E} \times \mathbf{H})$, and hence optimizes the continuous flow of Poynting EM energy from that simple device, even though the fields, seem to be "static". As Buchwald states [743], *"[Poynting's result] implies that a charged capacitor in a constant magnetic field which is not parallel to the electric field is the seat of energy flows even though all macroscopic phenomena are static. "* [744] This simple device will freely pour EM energy indefinitely, so long as it remains intact.

We consider a special but universal kind of related Maxwell's demon problem: the vexing problem of the source charge, sometimes called "the most difficult problem in classical and quantum electrodynamics." [745] Any charge pours out observable EM energy continuously in all directions, at the speed of light, with no *observable* EM energy input. This continuous flow of EM energy establishes the charge's related fields and potentials and their energy, reaching even across the universe for very old charges. In the classical electrodynamics (CED) model and therefore in electrical engineering, there is no solution to this problem because CED does not model *nonobservahle* EM energy inputs such as the virtual particle flux exchange of the charge with its vacuum environment.

With no solution available in CED, then the present CED and electrical engineering assume that every charge in the universe *freely and continuously creates energy out ofnothing,* and pours it out in all directions, forming its associated fields and potentials and their energy. Either we must solve the problem in the classical EM model, or totally surrender the conservation of energy law in the model in its present limited form. Or — as is presently the case — we may continue to ignore it, as has been done during the 45 years since the proof of broken symmetry in particle physics [746].

On the other hand, particle physics models and utilizes virtual energy **in** the seething vacuum, including the interactions of the vacuum with charge, With the 1957 experimental proof of broken symmetry by Wu et al., [746] certified by the award of the Nobel Prize to Lee and Yang later that **same** year, one of the proven broken symmetries in the vacuum virtual energy exchange is the asymmetry of opposite charges, such as are on the ends of a dipole. For a dipole or dipolarity, then by the very definition of broken symmetry *something virtual has become observable.* We diverge for a moment, and then will return to this proven asymmetry of the opposite charges of the dipole.

If we take the modern view of the bare charge clustered by virtual **charges** of opposite sign, then the magnitude of the bare charge in the middle is infinite (including its energy) and so is the magnitude of the charge of the surrounding clustering virtual charges. [747] So for an "isolated **observable** charge" we actually have a sort of "infinitely powerful dipole". Yet **the** difference between these two infinite values of charge is finite; it is routinely calculated as the observed charge of an elementary charged particle, the value that is listed in standard texts and handbooks.
The asymmetry of this "opposite composite charges" model of the "isolated observable charge" now explains the long-vexing source **charge** problem. The observable "composite" charge continuously absorbs **virtual** photon energy from the seething vacuum exchange, transduces it **into** observable photon energy excitation, and this excited state **continuously** decays to emit real observable photons in all directions at the speed of light. When *virtual* energy input as well as *observable* energy output is accounted, the charge's proven asymmetry in the vacuum exchange **makes** the source charge an open system far from equilibrium in its active

environment. As an open disequilibrium system the charge is thermodynamically permitted to (1) self-order (hence coherently integrate absorbed virtual photon energy into observable photon energy), (2) self-oscillate or self-rotate (spin), (3) output more energy than the "operator" **inputs** (the operator inputs none at all), and (4) exhibit negentropy. Every charge in the universe is already doing those four functions. It is also a system having a COP (coefficient of performance, or useful energy output divided by the operator's energy input) of COP = ∞, since the operator inputs nothing.

In short, every charge in the universe already is a special kind of known "Maxwell's demon", whose operational mechanism is long since proven in particle physics but does not appear in classical electrodynamics or electrical engineering. It is a dependable Maxwell's demon, since all charges in the original matter in the universe have been pouring out real EM energy freely, using this asymmetry mechanism, for some 14 billion years, **and** they continue to do so. Further, any charge one assembles in the laboratory instantly starts up its Maxwellian demon performance automatically, and it will continuously pour out EM energy in all directions indefinitely, so long as the charge remains intact. So this demon is easily created and demonstrated experimentally. The Maxwell's demon for freely extracting copious EM energy from the vacuum is one of the easiest entities to invoke in all of physics

This is **not** a violation of thermodynamics, since classical *equilibrium thermodynamics* with its infamous second law does not apply because the charge is far from equilibrium in its exchange with its active vacuum environment. Although the COP = ∞, the *efficiency* of the charge system -defined as the useful energy output divided by the total energy input from all sources — is never more than 100%. Hence this is not a perpetual motion machine, and the charge rigorously does not create energy; it only transduces energy input to it in a novel form. In this case, the charge can permissibly act as if it were a free electrical windmill turning in a novel free electrical wind due to its asymmetry in the vacuum flux.

So every charge in the universe exhibits COP = ∞, clearly proving that Maxwell's demon in an improved form exists and is thriving throughout the universe. There would be no EM fields and potentials and their energy, and hence no electrical circuits or power systems, were it not for this asymmetry of the charge in its virtual energy exchange with the vacuum, making the charge a true Maxwell's demon.

In the usual CED model and electrical engineering, the 1867 L. V. Lorenz symmetrical regauging (later credited to H. A. Lorentz) [748] of the equations carefully selects only those Maxwellian systems that have two simultaneous equal and opposite disequilibria with their active environment. This assumes that the potential energy of the system is freely changed twice (by gauge freedom), but only in such highly selected manner as to form two equal and opposite new "free" force fields in the system. These two force fields perform *internal work* inside the system continuously, increasing its stress (and its stress potential, thereby curving local spacetime). Since there is no *net resultant* force field, the two force fields are unable to translate electrons in the circuit to do free *external* work in the external load. So the symmetrically regauged Maxwellian system has been altered: its energy has been changed twice, forming a stress potential inside it and increasing and maintaining that stress on the system, and the frame of the system has been somewhat rotated out of **the** laboratory frame. The symmetrically regauged Maxwellian system most decidedly is not identical to the system prior to regauging.

We note merely that the common Lorentz symmetrical regauging unwittingly assumes two Maxwell's demons of yet different kind, each freely fueled from the local vacuum environment, and each continuously performing internal work upon the system to produce and maintain system stress.

Gauge freedom is thus revealed as a special kind of Maxwell's demon, since (1) it assumes that the potential energy of a system can be freely changed at will at any time, and (2) unless we abandon the conservation of energy law, that energy change must have involved energy exchange **with** the external environment. So the gauge freedom axiom of quantum field theory assumes two specialized Maxwellian demons — i.e., mechanisms for transfer of energy between environment and system, so as to continuously perform internal work upon the system. In this case the two demons are equal and opposite, and fight each other to a draw insofar as performing any useful external work.

Oddly, in present electrical circuits the ubiquitous but *arbitrary* closed current loop circuit — passing all spent current from the external circuit back through the source dipole in the generator against the dipole's **back** emf— self-imposes Lorentz regauging of excitation discharge in the circuit and prevents COP>1.0 functioning. This is not required by thermodynamics in general! Indeed, several areas are already known to violate present thermodynamics. Sharp discharges (strong gradients), for example, are known to violate it. [749] Other known areas where thermodynamics is violated include rarefied media, and anomalous memory effects in materials. Modern research is being conducted in such areas under the heading of "extended thermodynamics". [750].

Since experiment and not theory is primary in science, let us consider some additional actual Maxwell demon experiments. The well-known "negative resonance absorption of the medium" produces more output energy than the operator inputs to the experiment, with hundreds of these experiments done every year routinely by many nonlinear optical laboratories. As an example, Bohren's version of such an experiment [751] commonly outputs 18 times as much EM energy as the operator inputs, exhibiting COP = 18. Independent replication of Bohren's work by Paul and Fischer [752] is reported in the same journal issue.

Added to the charged capacitor lying on a permanent magnet, there thus exist plenty of proven, replicated experiments which can be easily performed to demonstrate a Maxwellian demon operating in a Maxwellian system and freely producing COP>1.0, or even COP = oo. Because of the demon (the operational mechanism), these experiments all involve open systems far from equilibrium with their active environment. Hence they are not limited by the second law of classical equilibrium thermodynamics.

In short. Maxwell's demon is very much alive in many forms and is thgiving after all. It has just been hiding in different form than what is usually suspected.

T.E. Bearden, Ph. D.
Magnetic Energy, Ltd.

ENDNOTES

1. (a) Albert Einstein, "Foreword," in (b) Max Jammer, *Concepts of Space: The History of Theories of Space in Physics,* Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

2. George Zebrowski, "The holdouts," *Nature,* Vol. 408, 14 Dec 2000, p. 775; expressing the thinking of Stephen Hawking, one of the leading physicists of our day.

3. Mario Bunge, *Foundations of Physics,* Springer-Verlag, New York, 1967, p. 176.

4. (a) J. H. Poynting, "On the transfer of energy in the electromagnetic field," *Phil. Trans. Roy. Soc. Lond,* Vol. 175, Part II, 1884, p. 343-361; (b) — "On the Connection Between Electric Current and the Electric and Magnetic Inductions in the Surrounding Field," *Phil. Trans. Roy. Soc. Lond,* Vol. 176, 1885, p. 277-306.

5. (a) Oliver Heaviside, "Electromagnetic Induction and Its Propagation," *The Electrician,* 1885, 1886, 1887, and later — a series of 47 sections, published section by section in numerous issues of *The Electrician* during 1885, 1886, and 1887; (b) —*Electromagnetic Theory, 3 vols., Be*nn, London, 1893-1912. Second reprint 1925. An unabridged edition of all three volumes, with a critical and historical introduction by **E. Weber,** was published by Dover, New York, 1950. See particularly (c) - "On the Forces, Stresses, and Fluxes of Energy in the Electromagnetic Field." P*hil Trans. Roy. Soc. London,* 183A, 1893, p. 423-480. In this prestigious publication, Heaviside discusses the Faraday-Maxwell ether medium, outlines his vector algebra for analysis of vectors without quaternions, discusses magnetism, gives the EM equations in a moving medium, and gives the EM flux of energy in a stationary medium. On p. 443, he credits Poynting with being first to discover the formula for energy flow, with Heaviside himself independently discovering and interpreting this flow a little later by himself in an extended form.

6. E.g., a polite debate has continued at intervals, for about 30 years, in the Am. J. P*hys.* alone.

7. Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics,* Addison-Wesley, Reading, MA, Vol. 1,

8. *Ibid.,* Vol. 1, 1964, p. 12-2.

9. *Ibid.,* Vol. 2, 1964, p. 2-1.

10. John D. Kraus, *Antennas,* Second Edition, McGraw Hill, 1988.

11. Raymond A. Serway, *Physics for Scientists and Engineers with Modern Physics,* Third edition, updated version, Saunders College Publishing, Philadelphia *et al,* 1990, p. 620.

12. T. E. Bearden, "Giant Negentropy from the Common Dipole," J. *New Energy,* 5(1), Summer 2000, p. 11-23. Also on website www.cheniere.org.

13. (a) T. D. Lee, "Question of Parity Conservation in Weak Interactions," *Phys. Rev.,* 104(1), Oct. 1, 1956, p. 254-259; (b) — with Reinhard Oehme and C. N. Yang, "Remarks on Possible Noninvariance under Time Reversal and Charge Conjugation," *Phys. Rev.,* 106(2), 1957, p. 340-345; (c) — and C. N. Yang, "Parity Nonconservation and a Two Component Theory of the Neutrino," *Phys. Rev.,* 105(5), Mar. 1, 1957, p. 1671-1675.

14. C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes and R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev., Vol.* 105, 1957, p. 1413.

15. We published it crudely two years later. See T. E. Bearden, *Quiton/Perceptron Physics: A Theory of Existence, Perception, and Physical Phenomena,* NTIS Report AD-763-210, 1973.

16. A time-reversal zone is a momentary tiny region of the fluid wherein antiphoton reactions with the main ions predominate over **photon** reactions.

17. See T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy,* 3(2/3), 1998, p. 12-28. A slightly expanded paper by the same title is published in *Explore,* 8(6), 1998, p. 7-16.

18. E.g., see Xiaodong Chen, "A New Interpretation of Quantum Theory — Time as Hidden Variable," Los Alamos National Laboratory web site, arXiv:quant-ph/9902037 v3, 11 Feb. 1999. Using two more **time** variables as the quantum hidden variables, Chen derives Dirac's field equation under the principle of quantum physics, and then extends **the**

method into quantum fields with arbitrary spin number. Chen's model offers very interesting unifications of quantum mechanical behavior.

19. F. Mandl and G. Shaw, *Quantum Field Theory,* Wiley, 1984, Revised Edition 1993, Chapter 5. This is a deeper coverage of the photon polarizationd. Mandl and Shaw argue that the longitudinal and scalar (time-polarized) polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. This argument, translated from particle terminology to wave terminology, directly fits and strongly supports my reinterpretation {12} of Whittaker's 1903 wave decomposition of the scalar potential {85}. The reinterpretation also represents the discovery of the giant negentropy process favored by nature, easily evoked by breaking 3-symmetry (as by forming a simple dipole).

20. James Clerk Maxwell, "A Dynamical Theory of the Electromagnetic Field," *Roy. Soc. Trans.,* Vol. CLV, 1865, p 459. Read Dec. 8. 1864. Also in *The Scientific Papers of James Clerk Maxwell, 2* vols, bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597. In this seminal paper, Maxwell's equations of the electromagnetic field are given in "Part III, General Equations of the Electromagnetic Field," p. 554-564. On p. 561, Maxwell lists his 20 variables. On p. 562, he summarizes the different subjects of the 20 equations, being three equations each for magnetic force, electric currents, electromotive force, electric elasticity, electric resistance, total currents; and one equation each for free electricity and continuity.

21. B. J. Hunt, *The Maxwellians,* Cornell University Press, 1991 gives the history of the alteration of Maxwell's field theory by Fitzgerald, Heaviside, Lodge, and Hertz. In an appendix, Hunt makes a comparison between the standard expressions of the Maxwell field equations and their expression in quaternionic form.

22. E.g., J. D. Jackson, *Classical Electrodynamics,* 2nd Edition, Wylie, 1975, p. 219-221.

23. Willis E. Lamb Jr. and Robert C. Retherford, "Fine structure of the hydrogen atom by a microwave method," *Phys. Rev.,* 72(3), Aug. 1, 1947, p. 241-243. Lamb received the 1955 Nobel Prize in physics jointly with Polykurp Kush for experiments measuring the small displacement later called the "Lamb shift" of O+ne of the energy levels in atomic

hydrogen. This is a small "vacuum energy" effect, but the *energy density* involved is greater than the surface energy density of the sun.

24. Craig F. Bohren, "How can a particle absorb more than the light incident on it?" *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327. Under nonlinear conditions, a particle can absorb more energy than is in the light incident on it. Metallic particles at ultraviolet frequencies are one class of such particles; insulating particles at infrared frequencies are another.

25. See also H. Paul and R. Fischer, "'(Comment on "How can a particle absorb more than the light incident on it?')," *Am. J. Phys.,* 51(4), Apr. 1983, p. 327 for a replication and validation of Bohren's experiment,

26. H. B. G. Casimir, "On the attraction between two perfectly conducting plates," presented at a meeting of the Royal Netherlands Academy of Arts and Sciences on 29 May, 1948. Published in the same year in *Proc. Koninklijke Nederlandse Akad. van Wetenschappen,* Amsterdam, vol. 51(7), 1948, p. 793-796. In English. It is an extrapolation of the more detailed paper of Casimir and Polder, 1948.

27. S. K. Lamoreaux, "Demonstration of the Casimir Force in the 0.6 to 6 micro m range," *Phys. Rev. Lett.,* 78(1), Jan. 6, 1997, p. 5-8. Lamoreaux gives a rigorous experimental verification of the Casimir effect.

28. Franklin B. Mead and Jack Nachamkin, "System for Converting Electromagnetic Radiation Energy to Electrical Energy," U.S. Patent No 5,590,031. An invention for extracting energy from the vacuum via the proven Casimir effect.

29. (a) Floyd Sweet and T. E. Bearden, "Utilizing Scalar Electromagnetics to Tap Vacuum Energy," *Proc. 26th Intersociety Energy Conversion Eng. Conf. (IECEC '91),* Boston, Massachusetts, 1991, p. 370-375. Sweet's device produced 500 watts output from a 33 microwatt **input** by the operator. A highly successful anti-gravity experiment designed by the present author was also performed, and is reported in the paper. Unfortunately, Sweet later died and never fully revealed the activation secret by which barium ferrite magnetic materials could be in self-oscillation at 60 Hertz. Weak self-oscillation of such permanent magnetic materials at higher frequency is now known, of course. E.g., see (b) V. E. Zakharov and V. S. L'vov, "Parametric excitation of spin waves in ferromagnets with magnetic inhomogeneities," *Sov. Phys. Solid State,* 14(10) 1973, p. 2513. Particularly see (c) V. S. L'vov, *Wave Turbulence Under Parametric Excitation: Applications to Magnets,* Springer Series in

Nonlinear Dynamics, Springer-Verlag, New York, 1994. Self-oscillation can in fact be induced in permanent magnet materials. We comment that, while a permanent magnet is in self-oscillation, a simple set of wires or a coil around the magnet will receive EM energy by induction, and can power a very tiny load in a closely adjacent circuit.

30. T. E. Bearden, "Bedini's Method for Forming Negative Resistors in Batteries," *J. New Energy,* 5(1), Summer 2000, p. 24-38. Also carried on restricted DoE website http://www.ott.doe.gov/electromagnetic/ and on the present author's open website http://www.cheniere.org.

31. Teruo Kawai, "Motive Power Generating Device," U.S. Patent No. 5,536,518.Jul. 25, 1995. My board of directors and I were physically present when the Kawai system, company, etc. were seized by what can only have been the Yakuza (Japanese mafia), even though Kawai was being backed by one of the wealthiest investors in Japan. I had just reached an agreement with Kawai to produce and market his overunity motor worldwide. Several legitimate Japanese overunity systems, ready for production, have been stopped in this fashion. Kawai's process, applied with efficient switching such as photon-coupled switching, essentially doubles the COP of the motor to which it is applied. If one starts with a high efficiency (e.g., Hitachi) permanent magnet electric motor of 0.7 or 0.8, one can expect to produce COP = 1.4 and 1.6 respectively, the actual results obtained by Hitachi tests of two Kawai-modified Hitachi motors. We accent that the efficiencies of the motors remain less than 100% at all times.

32. Gabriel Kron, "Electric circuit models of the Schrodinger equation." P*hys. Rev.* 67(1-2), Jan. 1 and 15, 1945. Quoting, p. 41: *""Now a value E of the negative resistances, at which the generator current becomes zero, represents a state at which the circuit is self-supporting and has a continuous existence of its own without the presence of the generator, as the negative resistances just supply the energy consumed by the positive resistances. ... When the generator current is positive the circuit draws energy from the source, and when the current is negative the circuit pumps back energy into the source. At zero generator current the circuit neither gives nor takes energy, and theoretically the generator may be removed."* In this passage, Kron was required to insert the word "theoretically". Kron was the chief scientist for General Electric on the Network Analyzer project at Stanford University before WW II. In the 1930, he did successfully build negative resistors (at least three), and he

did successfully power the network analyzer with them. He was never permitted to release the technical details, although he placed judicious hints in several of his papers, and slipped certain key phrases past the censors. Interestingly, Sweet was his protegee, and it is conceivable that the basis for Sweet's device may have been Kron's original negative resistor. Independent tests and verification of the Sweet device were performed by Walter Rosenthal, a retired professional test engineer of considerable experience.

33. Shoukai Wang and D. D. L. Chung, "Apparent negative electrical resistance in carbon fiber composites," *Composites, Part B,* Vol. 30, 1999, p. 579-590. Apparently the referees required the replacement of the term "negative electrical resistance" by the term "apparent negative resistance". In the paper, simply examine the actual results shown in the curves of the negative-going current, against the voltage, with the line passing directly through the origin. That is a true negative resistance, not a differential one or an "apparent" one. A patent was filed on this invention.

34. (a) T. H. Moray, *The Sea of Energy,* 5th Edn., Salt Lake City, 1978. The Moray device weighed 55 pounds and produced 50 kilowatts of power in the late 1930s, and was demonstrated to scientists and engineers numerous times, Moray made tubes which actually contained very special semiconductors, three of them in contact, each pressed at high pressure during a sintering process. A very eerie chain of circumstances eventually destroyed the device. Moray rode in a bulletproof car, and he was sometimes fired at from the sidewalk or an alley. A double agent, working for both the U.S. and the Soviet Union, infiltrated his lab and eventually destroyed the device. The Morays were never given a proper scientific hearing and were never funded for full development. T. Henry Moray was shot and wounded in his own laboratory, wrestling with a would-be assassin. Yet so many expert engineers and scientists witnessed and tested the machine in operation that there is no question of its legitimacy. Unfortunately, the reference cited is not at all a satisfactory scientific examination of the Moray device. A related U.S. patent was issued which at least shows some details of the Moray tube: See (b) Thomas H. Moray, "Electrotherapeutic Apparatus," U.S. Patent No. 2,460,707, Feb. 1, 1949. I personally received a treatment on a Moray device for severe pain, and the relief was rapid and remarkable.

35. See (a) William B. Burford III and H. Grey Verner, *Semiconductor Junctions and Devices: Theory to Practice,* McGraw-Hill,

New York, 1965, Chapter 18: Point-Contact Devices. The very first transistor — the point contact transistor — also often exhibited negative resistance. Quoting, p. 281:

> *"First, the theory underlying their function is imperfectly understood even after almost a century..., and second, they involve active metal-semiconductor contacts of a highly specialized nature. ... The manufacturing process is deceptively simple, but since much of it involves the empirical know-how of the fabricator, the true variables are almost impossible to isolate or study. ... although the very nature of these units limits them to small power capabilities, the concept of small-signal behavior, in the sense of the term when applied to junction devices, is meaningless, since there is no region of operation wherein equilibrium or theoretical performance is observed. Point-contact devices may therefore be described as sharply nonlinear under all operating conditions."*

Certainly degenerate four-wave mixing and therefore time reversal can occur in in semiconductors. E.g., see (b) R. K. Jain, "Degenerate four-wave mixing in semiconductors: application to phase conjugation and to picosecond-resolved studies of transient carrier dynamics." *Opt. Eng.,* 21(2'), March/April 1982, p. 199-218. We also point out that hole current can be emitted from an n-type semiconductor having a metallic point placedd on its surface, and hole current will move against the voltage (electron current view). A process (patent pending) by Bedini and Bearden have shown that a capacitor may be charged by hole current, and then discharge electron current, effectively transducing negative energy into positive energy. It is therefore possible (Bedini's method) to switch a capacitor from a hole-current charging circuit by a hole-emitting process, to a bypass capacitor across a source such as a battery after the capacitor has reached higher charge voltage than the source. The capacitor will then charge the battery and power the circuit as well, during that portion of its discharge cycle while its voltage is above that of the battery. This is a legitimate COP>1.0 process, invented by Bedini.

36. (a) William Jay Fogal, "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,196,809, Mar. 23, 1993; (b) — "High Gain, Low Distortion, Faster Switching Transistor." U.S. Patent No. 5,430,413, July 4, 1995. The Fogal semiconductor can be rigged to perform the permitted degenerate four-wave mixing and therefore time reversal. It can thus be rigged as a true negative resistor, where the current moves against the voltage.

37. Stephen L. Patrick, Thomas E. Bearden, James C. Hayes, Kenneth D. Moore, and James L. Kenny, "Motionless Electromagnetic Generator," U. S. patent # 6,362,718, March 26, 2002.

38. (a) T. E. Bearden, "Energy from the Active Vacuum: The Motionless Electromagnetic Generator," in M. W. Evans (Ed.), *Modern Nonlinear Optics,* Second Edition, Wiley, 2001, Part 2, p. 699-776; (b) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al.,* "Explanation of the Motionless Electromagnetic Generator with 0(3) Electrodynamics," *Found. Phys. Lett,* 14(1), Feb. 2001, p. 87-94; (c) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et ah,* "Explanation of the Motionless Electromagnetic Generator by Sachs's Theory of Electrodynamics," *Found Phys. Lett.,* 14(8), Aug. 2001, p. 387-393.

39. A version of the MEG was also independently replicated by J. L. Naudin. See his website at http://jnaudin.free.fr/html/meg.htm .

40. T. E. Bearden, "Dark Matter or Dark Energy?", *J. New Energy,* 4(4), Spring 2000, p. 4-11.

41. See Chapter 4 for a high-level expose of this methodology.

42. Sweet and Bearden, {29a). Sweet performed an antigravity experiment designed by the present author, and successfully reduced **the** weight of an object on the laboratory bench, by 90%. Sweet's solid state magnetic vacuum triode amplifier produced a COP = 1,500,000, so a **very** appreciable Dirac sea hole current was available to work with. Conversion of the Dirac sea hole current into electrical current via the present Bedini-Bearden process would have eliminated the antigravity and produced a very large increase in output electrical power.

43. See (a) T. E. Bearden, "The Unnecessary Energy Crisis: How to Solve It Quickly," ADAS Position Paper, June 2000; (b) Bearden **{12};** (c) Bearden {30}; (d) T. E. Bearden, "EM Energy From The Vacuum: Ten Questions With Extended Answers," September 2000. The present author has been calling for just such a program. These papers are on restricted DOE Website http://www.ott.doe.gov/electromagnetic/ and also are on the author's website at http://www.cheniere.org.

44. Nathan Smith, "Marketable Results vs. Good Science?," *Sci. Am.*, 284(5), May 2001, p. 8-9. Quote is from p. 9.

45. This is the actual phrase often used in such Government contracts. What it means is that the Government (which means the government

employee controlling that program and its funds) can suddenly declare that the Inventor is not getting his invention developed rapidly enough, and the Government then seizes the patent to "get it developed more quickly". The inventor is suddenly left hanging, no longer in control of his own intellectual property. Or the Government has the power to block development altogether.

46. Often the universities just bluntly state that *all patents revert to the university*.

47. Hal Hellman, *Great Feuds in Science*, Wiley, New York, 1998. We particularly call attention to Hellman's dissertation on Wegener's persecution. Today, the legitimate overunity EM researcher and the cold fusion researcher are regarded much like Wegener was regarded — as the very epitome of a fool. This label is appended by "experts" who are not even aware of what really powers an EM circuit or an electrical power distribution system.

48. A photon is a piece of angular momentum (called "action" by the physicist), consisting of a little piece of spatial energy (ΔE) and a little piece of time (Δt), welded together as $(\Delta E)(\Delta t)$ with no seam in the middle, so to speak. But now look at that little $(\Delta E)(\Delta t)$. Since the photon is quantized, $(\Delta E)(\Delta t) = k = \hbar$. Any change in (Δt) requires that (ΔE) change inversely; i.e., $(\Delta E) = \hbar/(\Delta t)$. So increasing (Δt) decreases (ΔE) accordingly. Let us call the spatial energy s compacted into (dt) the equivalent *energy* of the photon. Then s is given by $s = c^2(\Delta t) = 9\times10^{16}(\Delta t)$. So as the frequency of a photon lowers, the (Δt) time component increases proportionally, and the total equivalent energy in the photon increases dramatically by many orders of magnitude! As an example, suppose we halve the spatial energy (ΔE). That doubles the original time interval (Δt), which gives the new $s = c^2(2\Delta t)$ or $s = 2c^2(\Delta t)$ for the new equivalent energy. So while we halved the *normal* spatial energy formally transported by the photon, we multiplied the *equivalent* time-energy by the enormous factor of $2c^2$ which is approximately 18×10^{16} in the MKS system. It follows that, if we use the equivalent time) energy of the photon and transduce a little of it into normal spatial energy, the energy density of those nuclear reactions now seen only in distant energetic objects such as pulsars, exploding stars, etc. can be achieved right on the laboratory bench. *The highest energy physics has not yet been born in the physics community, and what is presently being done in the high-energy*

accelerators is child's play compared to what can be done with time-energyengineering.

49. To see by another method that time is energy, one notes that the choice of fundamental units chosen for one's physics model is arbitrary. So we choose a single fundamental unit, the joule, to make a new physics model. Thereupon all other entities are energy and energy functions only, Hence, time is totally a function of energy. Actually, time is spatial energy compressed by the factor c^2, so it has the same energy density as mass, except the highly compressed spatial energy is placed in the time domain to appear at "time" rather than in 3-space to appear as "mass".

50. E.g., in cold fusion experiments where the time-energy is partially transduced in a time-reversed manner, the gluon forces binding the quarks in the affected neutrons and protons are dramatically reduced and the quarks are therefore nearly freed. This is being done in an apparatus of perhaps two cubic feet in volume, rather than a giant accelerator 20 miles in diameter and costing billions of dollars. So far, the largest accelerator using 3-space (decompressed) energy is just a bit short of the energy required to free the quarks. Use of time (compressed) energy allows such reaction as freeing the quarks to be easily accomplished, while the accompanying decompressed *spatial* energy is very low.

51. Time energy has the same energy density as mass, but that highly compressed energy has just been placed in the 4th axis *ict* of Minkowski space rather than in 3-space. In *ict* the only variable is t, so the variation of the compressed energy can only appear in the *t* variable portion. See Chapter 2.

52. (a) James Patterson, U.S. Patents 5,318,675; 5,372,688; 5,607.563 (b) — and Dennis Cravens, U.S. Patent No. 5,607,563.

53. Ironically, as suppression of innovative research in science increases, science may eventually face turning into the kind of dogmatic system that it struggled against for some 300 years before escaping. When science defends a theory that experiments refute, as it does today in the case of cold fusion, then it is very close to that debacle that can spell the end of science itself.

54. E. Allan Blair, Letter to the Editor, *Newsweek,* Mar. 3, 1997, p. 14.

55. Arthur C. Clarke, "Space Drive: A Fantasy That Could Become Reality," *NSS... AD ASTRA,* Nov/Dec 1994, p. 38.

56. William K. George, in *Editing the Refereed Scientific Journal, edited* by Robert A. Weeks and Donald L. Kinser, IEEE Press, New York, 1994. p. 227-228.

57. Max Planck, in G. Holton, *Thematic Origins ofScientific Thought, Harva*rd University Press, Cambridge, MA, 1973.

58. Jonathan Schlefer, editor of *Technology Review,* published at MIT. in an October 1990 editorial.

59. Rolf Landauer, "Fashions in Science and Technology," *Physics Today*, 50(12), Dec. 1997, p. 62.

60. Julius Robert Von Mayer, "Bemerkungen iiber die Krafte der unbelebten Natur," in J. von Liebig's *Annalen der Chemie* (vol. 42), 1842; Die *organische Bewegung in ihrem Zusammenhange mit dem Stoffwechsel, (Organic* Motion in Its Relation to Metabolism), 1845. His papers were published as *Die Mechanik der Warme,* 1867, with third edition in 1893.

61. Alfred Wegener, *Die Entstehung der Kontinente und Ozeane, 1915;* fourth revised edition, 1929. A Dover reprint of the fourth revised edition, translated into English, *Origin of Continents and Oceans,* was published in 1996.

62. Paul Nahin, *Oliver Heaviside: Sage in Solitude,* IEEE Press. New York, 1988, p. 225.

63. Martin Gumpert, *Trail-Blazers of Science.* New York, Funk and Wagnalls Company, 1936, p. 232.

64. This area is now referred to as *ovonics,* where the "ov" is directly taken from "Ovshinsky". The materials have found use in copy machines, fax machines, liquid crystal displays, and optical memory disks. Ovshinsky has obtained over 200 patents and continues to work in the energy conversion field. His development of a new storage battery, e.g., broke all the Department of Energy tests and doubled the energy density permitted. Once in awhile the good guys beat the scientific suppression system.

65. Stan and Iris Ovshinsky were honored with the American Chemical Society's Heroes of Chemistry Award for 2000. They were honored for 40 years of sustained effort and contributions in the

sustainable and renewable energy field, including amorphous silicon cells for roofs of houses, highly specialized storage batteries, etc.

66. On the other hand, many scientific disagreements are conducted in quite proper and gentlemanly fashion. An example is the continuing decades of argument over energy flow theory, in *Am. J. Phys.* The problem has still not been officially resolved; we would hope that the material in this book will shed new light upon it, by pointing out the exact difference between the long-neglected nondiverged component as well *as* the diverged component, and also pointing out what Lorentz's trick of closed surface integration of the energy flow vector really does.

67. It follows quite naturally that, once we compress spatial energy by c , we may place the compressed energy in 3-space, in which case it is known as "mass", or we may place it in the time domain, in which case it is known as "time". If that cannot be done, then physics errs in using time as a variable on the fourth axis, and fundamental units are not arbitrary after all.

68. E.g., see D. K. Sen, *Fields and/or Particles,* Academic Press, London and New York, 1968, p. viii. Quoting: *"The connection between the field and its source has always been and still is the most difficult problem in classical and quantum electrodynamics."*

69. See T. D. Lee, "Can Time Be a Discrete Dynamical Variable?". *Phys. Lett,* 122B(3, 4), Mar. 10, 1983, p. 217-220. Lee showed that time is a discrete dynamical variable across the entire range of mechanics: from classical to nonrelativistic quantum mechanics, and then to relativistic quantum field theories. In all stages of mechanics, time can be treated as a discrete parameter, and it can be treated as a bona fide dynamic variable, Hence time has dynamics and structure, and therefore energy.

70. E.g., one may assume that the entire virtual state is an internal structuring and dynamics existing inside the macroscopic flow of time. The virtual state dynamics obviously exists in spacetime (which is assumed to be continuous), but we also consider any "virtual particle" (e.g., a virtual electron) as if it had been suddenly "observed" momentarily That assumes d/dt (LLLt) => LLL, a frozen 3-space snapshot. However, since that cannot really be observed, then we are actually considering something analogous to "if we could observe something in spacetime that cannot be observed in space alone". The only place left for it to

!observably exist anyway, if we could in fact observe it" is time, if space is ruled out.

71. It is necessary to trace the energy transductions and dissipations inside the generator. Ignoring those internal dynamics, the casual observer sees that one inputs shaft energy to the generator, and out comes some electromagnetic energy in the external electrical circuit, pouring from the gnerator terminals. We catch a lesser amount out there in the external circuit or power line, so we think we input some energy, had some internal losses and what gets collected in the conductors of the power line is the remainder of what we input, after transductions and losses in the generator. That is not what happens at all! Embarrassingly, an *enormous* flow of energy comes out of those generator terminals, if all of it is accounted. This includes (i) the tiny component that strikes the surface charges of the external conductors and gets diverged into the conductors, thus being captured by the circuit, and (ii) the enormous remainder of the energy flow in the surrounding space, which misses the circuit altogether and is just wasted. Something like a trillion times as much EM energy may be pouring out of the terminals of the generator and *being wasted,* as we mechanically input to the generator shaft, or as we "catch" and use in the external circuit. As later chapters discuss, Poyntin ; never considered the component that misses the circuit entirely. Heaviside discovered it but could not explain its startling magnitude or its source. Lorentz knew of this enormous flow component. He could not explain it, and stated it "had no physical significance" since it powered nothing. Lorentz originated a little integration trick that discarded it neatly from a11 accountability.

72. If the energy input to an inert power system generator were what powered the attached external circuit, then only the Poynting component would be emitted from the terminals of the generator. In that case, Heaviside's discovery of the additional nondiverged component would be wrong, and Lorentz was also wrong in recognizing it by stating it had no *physical* significance. It also means that a dipole formed between the terminals can have no broken symmetry in its active vacuum flux exchange, which is falsified by particle physics {14, 73,}. Also, no charge could pour out energy and create fields and potentials in surrounding space, and neither could a dipole. That is easily falsified by actual experimental measurement. Our reinterpretation of Whittaker's wave decomposition of the scalar potential would also be wrong, as would Mandl's and Shaw's argument {19} that a combined pairing of a scalar

photon and a longitudinal photon are observable, producing the instantaneous scalar potential.

73. T. D. Lee, *Particle Physics and Introduction to Field Theory,* Harwood, New York, 1981,*passim.* In 1957, Lee and Yang received the Nobel Prize for their work in predicting broken symmetry. This included the broken symmetry of opposite charges — and hence of a dipole.

The very definition of broken symmetry, applied to the source dipole once it is formed in the generator, reveals that the dipole itself receives virtual photon energy from the seething vacuum, transduces it, and outputs it as real observable EM energy pouring out of the generator terminals and out through 3-space surrounding the entire power line. This is the source of the tremendous energy flow from the terminals, including a large component missing the circuit (as discovered by Heaviside in the 1880s). This large nondiverged component of the energy flow was discarded arbitrarily by Lorentz because there was absolutely no explanation of its source, and because it was thought to have "no physical significance" (Lorentz's term) since it powered nothing.

74. Wu et al. {14} experimentally showed that the weak interaction violates parity (spatial reflection) and thus exhibits broken symmetry.

75. T. D. Lee is noted for his broken symmetry work with charges, and for establishing the broken symmetry of separated charges of unlike sign. In 1957, Lee and Yang were awarded the Nobel Prize in Physics for their investigations of weak interactions between elementary particles.

76. John D. Kraus, *Electromagnetics,* Fourth Edition, McGraw-Hill, New York, 1992, p. 578.

77. We strongly feel that anyone interested in logic must read Morris Kline, *Mathematics: The Loss of Certainty,* Oxford University Press, New York, 1980. Contrary to popular opinion and the impression projected by mathematicians, mathematics is not a body of unshakable truths aboul the physical world, and mathematical reasoning is not exact and fallible.

78. One also needs to be aware of Godel's proof. See Kurt Godel, "Uber formal unentscheidbare Satze der *Principa Mathematica* une verwandter Systeme" ("On Formally Indeterminable Propositions of the *Principia Mathematica* and Related Systems," in *Monatshefte fur Mathematik und Physik,* Vol. 38, 1931. This is the publication in which Godel's Proof first appeared, which states that within any logical

mathematical system there are propositions or questions that cannot be proved or disproved on the basis of the axioms within that system. Therefore, it is uncertain whether or not the axioms or arithmetic will give rise to contradictions, since they can.

79. A critical reference for the longitudinal EM interferometry *weapons* is M. W. Evans, P. K. Anastovski, T. E. Bearden *et al.*, "On Whittaker's Representation of the Electromagnetic Entity in Vacuo, Part V: The Production of Transverse Fields and Energy by Scalar Interferometry," *J. New Energy,* 4(3), Winter, 1999, p. 76-78. This paper is a short but rigorous proof of a scalar interferometer's ability to produce EM fields and energy in the distant interference zone. Either diverging (heating) energy or converging (cooling) energy — in the normal sense, extraction of energy — can be created in a distant interference zone. By pulsing and controlling the circuit ground bias of the transmitters, one produced distant hot explosions or cold explosions. In a military target in that distant zone, the effects arise from the local spacetime, thus from everywhere within the targeted vehicle. At least eight nations now have such weapons, which were apparently the weapons that Defense Secretary Cohen {89} referred to in 1997.

80. Jackson {22}, p. 249.

81. Bunge, {3},p. 182.

82. Roger Penrose, "Foreword," in T. W. Barrett and D. M Grimes [Eds.], *Advanced Electromagnetism: Foundations, Theory, & Applications,* World Scientific, (Singapore, New Jersey, London, and Hong Kong), 1995.

83. Myron Evans, President of Alpha Institute's Institute for Advanced Study, private communication, 1999. Dr. Evans is the author of over 600 scientific papers, the prestigious editor of several scientific book series, and a noted foundations scientist. He is the co-discoverer and proponent of the Vigier-Evans B(3) magnetic field and a proponent of the proposed fundamental extension of U(1) electrodynamics to O(3) electrodynamics. O(3) electrodynamics has now been revealed as an important subset of Sachs' unified field theory.

84. Robert H. Romer, "Heat is not a noun," *Am. ./. Phys.,* 69(2), Feb. 2001, p. 107-109. The quotation is from footnote 24, p. 109.

85. E. T. Whittaker, "On the Partial Differential Equations of Mathematical Physics," *Math. Ann.,* Vol. 57, 1903, p. 333-355.

86. But see Chapter 2 for a reinterpretation and correction of Whittaker's decomposition of the scalar potential. His harmonic set of bidirectional phase conjugate longitudinal EM wavepairs in 3-space is replaced with a harmonic half-set of longitudinal EM waves converging in the time domain upon the dipolarity; absorption and transduction by the charges; and re-emission as the other harmonic half-set of longitudinal EM waves diverging out from the source dipolarity in all directions in 3-space. The correction allows a very novel solution to the long-vexing problem of the source charge and its associated fields and potentials and their energy {12}.

87. (a) G. Johnstone Stoney, "Microscopic Vision," *Phil. Mag.,* Vol 42, Oct. 1896, p. 332; (b) — "On the Generality of a New Theorem," ***Ibid.,*** Vol.43, 1897, p. 139-142; (c) — Discussion of a New Theorem in Wave Propagation," *ibid,* Vol. 43, 1897, p. 273-280; (d) — "On a Supposed Proof of a Theorem in Wave-motion," *ibid.,* Vol. 43, 1897, p. 368-373.

88. Some Russian scientists refer to the hidden internal electrodynamics as the "information content of the field".

89. Secretary of Defense William Cohen at an April 1997 counterterrorism conference sponsored by former Senator Sam Nunn. Quoted from DoD News Briefing, Secretary of Defense William S. Cohen, Q&A at the Conference on Terrorism, Weapons of Mass Destruction, and U.S. Strategy, University of Georgia, Athens, Apr. 28, 1997.

90. For a basic discussion of these photon polarizations, see (a) Richard Feynman, *Quantum Electrodynamics,* 1961, 1963. Quantum field theory recognizes four polarizations of the photon, to include x- and y-transverse polarizations in 3-space, longitudinal or z- polarization in 3-space, and the time-polarized photon (which may be taken to be a longitudinally-polarized photon in the time axis). Existence of these four modes of polarization or energy vibration in photons implies the existence of four correspondingly polarized EM waves. Among other things, Feynman discusses the "sum over four polarizations" of photons and such The term "Gupta-Bleuler" theory is the formalism that has grown up around this problem. For a deeper discussion, see (b) Lewis H. Ryder, *Quantum Field Theory,* Second Edition, Cambridge University Press, 1996, p. 147 *et seq.* For an even more advanced discussion, see (c) F.

Mandl and G. Shaw {19}. For discovery of the coupling of time-polarized and 3-space longitudinally polarized EM waves as the higher EM energy flow symmetry resolving the source charge problem, and thereby reinterpreting Whittaker's decomposition of the scalar potential {85}, see (d) Bearden {12}. We point out that the latter reinterpretation can also be "cranked through" the superpotential theory of Whittaker, to involve all EM potentials, fields, and waves.

91. (a) E. T. Whittaker, "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," *Proc Lond. Math. Soc,* Series 2, Vol. 1, 1904,p. 367-372. The paper initialed what today is called *superpotential* theory. It was published in 1904 and orally delivered in 1903. For a sampling of other superpotential work, see (b) W. H. McCrea, *Proc. Roy. Soc. Lond. A,* Vol. 240, 1957, p. 447; (c) A. Nisbet, *Physica,* Vol. 21, 1955, p. 99; (d) P. Debye, *Ann. Phys.,Leipzig,* Vol. 30, 1909, p. 57. For an excellent overview discussion of superpotentials and related things, see (e) Melba Phillips, "Classical Electrodynamics," in *Principles of Electrodynamics and Relativity,* Vol. IV *of Encyclopedia of Physics,* edited by S. Flugge, Springer-Verlag, l962.

92. (a) See Evans et al. {79} for proof that scalar interferometry can produce all normal transverse EM fields, waves, and patterns. The paper appears in (b) P. K. Anastasovski; T. E. Bearden, C. Ciubotariu, W. T. Coffey, L. B. Crowell, G. J. Evans, Myron W. Evans, R. Flower, S. Jeffers, A. Labounsky, B. Lehnert, M. Meszaros, P. R. Molnar, J.-P. Vigier, and S. Roy, "The New Maxwell Electrodynamic Equations: New Tools for New Technologies," *J. New Energy, 4(3),* Special Issue of A1AS papers. Winter 1999. The volume consists of some 60 papers by the Alpha Foundation's Institute for Advanced Study (AIAS), advancing electrodynamics to a non-Abelian, gauge theoretic higher topology theory in O(3) internal symmetry. Some ten of the papers directly address Whittaker's work and extend it.

93. (a) W. A. Rodrigues, Jr. and J.-Y. Lu, "On the existence of undistorted progressive waves (UPWs) of arbitrary speeds $0 < v < \infty$ in nature," *Found. Phys.,* 27(3), 1997, p. 435-508 show that Maxwell's equations do possess superluminal solutions. A slightly corrected version is downloadable as hep-th/9606171 on the Los Alamos National Laboratory web site. Also see (b) W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Solutions in Vacuum of the Maxwell

Equations and the Massless Dirac Equation," *Adv. Appl. Clifford Algebras,* Vol. 7(S), 1997, p. 457-466.

In classified Russian weapons labs, this internal longitudinal wave (both time-domain and 3-space domain) structuring of all normal (transverse) EM potentials, fields, and waves is euphemistically called the *information content of the field.* For decades there has been a successful disinformation campaign to lead scientists of other countries into believing that the phrase means *standard spectral analysis.* It doesn't. It is also possible to communicate at superluminal speed, using longitudinal EM waves inside this infolded "inner" electrodynamics. Since mass is mostly empty space filled with a particle here and there — together with EM potentials, fields, and waves — then mass is a vast "superhighway" for direct superluminal communication using longitudinal EM waves.

94. Romer, {84}.

95. See E. T. Whittaker, *A History of the Theories of the Aether and Electricity,* two vols., Nelson, London, 1951, 1953. Reprinted by Dover.

96. Note that we have just resolved the age-old problem of change, which can be summarized in the query stated so long ago by Heraclitus: "For a thing to change, it must turn into something else. But how can a thing be itself and something else also?" We have explained the enigma by pointing out that the "thing" has no persistence in time, and is continually changing into yet another thing, either a replica (in which case we sense "no change" and hence "persistence" by memory comparison), or a partial replica with some differences (in which case we sense that "the object has changed itself in some way"), or a thing with little or no replication (in which case we sense a "different thing"). It is the unseen nature of the *d/dt* observation process that has deluded us into such irresolvable logical conflicts — and into an incomplete Aristotelian logic based on observation alone, which does not contain the resolutions to such problems because a higher-order logic is required.

97. Whittaker, {85}

98. See Chapter 2 for a discussion of the substitution of effect for cause, and of the nature of observation.

99. *Wu etal,* {14}.

100. Sen, {68}.

101. Also see Mandl and Shaw {19}.

102. William G. Harter, John Evans, Roberto Vega, and Sanford Wilson, "Galloping waves and their relativistic properties," *Am. J. Phys., 53*(7), July 1985, p. 671-679. A relativistic example is given in which the source frame observes a very nonuniformly-galloping wave whose velocity varies between 0.01 c and 100 c, even though its average velocity is c. Relativistic properties of interfering electromagnetic plane waves are discussed.

103. We point out, but do not further pursue, that this situation is reversed when time is reversed. This has very interesting applications in our little "time reversal zones" that we found were so important in the electrolyte experiments of the cold fusion researchers, and which accounted for a new class of nuclear interactions previously unknown. This does reverse the conventional law of attraction and repulsion of charged particles. It also directly affects the quarks and gluons, partially or even nearly dissolving the gluon forces holding the quarks in place, so that the quarks are nearly freed and are easily flipped.

104. Simultaneously accompanied by *separately broken* time-energy flow symmetry and *separately broken* 3-spatial energy flow symmetry.

105. Evans etal., {38b, 38c}.

106. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

107. Kraus, {76}.

107. E.g., see (a) Jackson {22}, p. 237. Jackson points out that *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences. "* The first sentence is true, but the second sentence is a non sequitur. Poynting only considered the small component of the overall energy flow that enters the circuit — in short, that strikes the surface charges in the conductors and is diverged into the wires to power the Drude electrons. That is the diverged part of the energy flow. All the rest of the flow that misses the circuit is obviously not diverged or collected in the circuit, and so it has zero divergence. It is elementary vector algebra that the curl of any vector field has zero divergence. So the extra nondiverged Heaviside energy flow component — filling all space around the conductors in a power line, e.g., but missing the power system entirely and just wasted — can indeed be expressed as the curl of a vector field That energy flow is real, however, and if intercepted and collected, it then does become divergent. In that case, it does have highly significant

consequences, since the total energy flow in that component may be a trillion times the energy flow in the Poynting component alone.

For a rigorous experiment demonstrating the reality and significance ol that "physically insignificant energy flow" that "can have no physical consequences", see (b) Bohren {24}. Under nonlinear conditions, a particle can absorb more energy than is in the light incident on it by normal calculations. Electrodynamicists do *not* calculate the magnitude of an entire field or potential, e.g., but only the intensity of a *divergence from* it at a single point, as given by an intercepting unit point static charge. If that same charged particle is placed in particle resonance, it sweeps out a greater geometrical reaction cross section. Hence it intercepts and **collects** (diverges) more impinging energy — from that long-neglected otherwise nondiverged component of the flow or field that is not diverged by the *static* unit point charge assumed conventionally. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another. See also (c) Paul and Fischer, {25}, who replicated Bohren's experiment.

109. H. A. Lorentz, *Vorlesungen iiber Theoretische Physik an der Universitdt Leiden,* Vol. V, *Die Maxwelhche Theorie (1900-1902),* Akademische Verlagsgesellschaft M.B.H., Leipzig, 1931, "Die Energie im elektromagnetischen Feld," p. 179-186. The section is called "Der Poyntingsche Energiestrom," and begins on p. 183. See his Figure 25 **on** p. 185.

110. E.g., see W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism,* Addison-Wesley, Reading, MA, 1962, 2nd edition, p. 181**,**

111. Bearden, {12}.

112. A simple way to see this is to note that the back-potential across the half-loop internal to the dipole is precisely equal to the forward potential externally between the ends of the dipole but across the **external** circuit half-loop. Since the same current q runs through both these potentials, then by Vq the same amount of energy is dissipated upon **the** charges of the dipole — to scatter them and destroy the dipole — as is dissipated in the external circuit in its losses and the load. Thus **more** energy is used to destroy the dipole than is used to power the load. We have to input at least as much shaft energy to restore the dipole as was used to destroy it. Hence the COP< 1.0.

113. See to (a) M. W. Evans, P. K. Anastasovski, T. E. Bearden *el al*, "Classical Electrodynamics without the Lorentz Condition: Extracting Energy from the Vacuum," *Physica Scripta,* 61 (5) . May 2000, p. 513-5 17 in addition to that paper previously quoted. See also (b) "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik,* 111(9), 2000, p. 407-409.

114. Robert Bruce Lindsay and Henry Margenau, *Foundations of Physics,* Dover, New York, 1963, p. 25.

115. *Ibid.,* p. 213.

116. *Ibid.,* p. 217.

117. M. Faraday, *Experimental Researches in Electricity and Magnetism,* Vol. 1, Taylor and Francis, London, 1839; Vol. 2, Richard & John E. Taylor, London, 1844; Vol. 3, Taylor and Francis, London, 1855. A 1965 reprint of all three volumes is available from Dover, New York.

118. We also acknowledge the independent contributions of Joseph Henry. In 1830, Henry independently discovered magnetic induction, anticipating Faraday's own discovery of it. Faraday, however, published first. Two years later, Henry discovered self-induction. He died in 1878, and the unit of inductance, the *henry,* was named for him in 1893.

119. (a) Maxwell, {20}. Also see (b) James Clerk Maxwell, *A Treatise on Electricity and Magnetism,* Oxford University Press, Oxford, 1873, Third Edition, Volumes 1 and 2, unabridged, Dover Publications, New York, 1954. This standard citation widely used in the literature is in fact confusing; the First Edition was published in 1873, a second rather substantially altered edition was later published after Maxwell's 1879 death of stomach cancer, and a third edition, slightly altered from the second edition, was published even later. So controversial were quaternions at the publication of Maxwell's first edition in 1873, that part of what was to become the second edition — published after Maxwell's 1879 death — was truncated by Maxwell himself prior to his death and prior to the posthumous publication of the second edition. Thus there is considerable difference between the first and second editions, and first and third editions, of Maxwell's *Treatise,* compared to his original 1865 paper {20}. "Maxwell's theory" is actually his 1865 paper; everything else is a later truncation of Maxwell's theory.

120. Maxwell's letter to P. G. Tait in 1871, *Archives,* Cavendish Laboratory, Cambridge.

121. For a discussion, see (a) A. M. Bork, "Vectors versus quaternions - the letters in *Nature,"* in *Am. J. Phys.,* Vol. 34, Mar. 1966, p. 202-211. For a cogent presentation of what might have been discovered much earlier in physics if quaternions had not been cast aside in favor of the much more limited vectors, see (b) James D. Edmonds Jr., "Quaternion quantum theory: New physics or number mysticism?", *Am. J. Phys.,* 42(3), Mar. 1974, p. 220-223.

122. Edmonds, {121b}. Edmonds presents a cogent argument showing the significant impedance of the rate of discovery in physics, that **resulted** from the ill-advised casting aside of the far more comprehensive quaternion theory in favor of the much more limited vector theory.

123. Heaviside, {5a, 5b}.

124. T. W. Barrett, "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," *Annales de la Fondation Louis de Broglie,* 16(1), 1991, p. 23-41, Barrett later extended some of the energy-shuttling techniques shown by Tesla and obtained U.S. Patents 5,486,833 and 5,493,691.

125. (a) A. A. Michelson and E. W. Morlcy, "Influence of motion of the medium on the velocity of light," *Am. J. Sci.,* Vol. 31, Series 3, 1886, p. 377-386; (b) — "The relative motion of the earth and the luminiferous aether," *Am. J. Sci.,* 34(3), 1887, p. 333; (c) — "On the relative motion of the earth and the luminiferous aether," *Phil. Mag.* 24(4), 1887, p. 449. See also (d) A. A. Michelson, "The relative motion of the earth and the ether," *Am. J. Sci.,* (4), 3, 1897, p. 475.

126. (a) Mendel Sachs, *General Relativity and Matter: A Spinor Field Theory from Fermis to Light-Years (Fundamental Theories of Physics),* Reidel, 1982. Sachs provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also (b) Mendel Sachs, "Symmetry in Electrodynamics: From Special to General Relativity, Macro to Quantum Domains," in M. W. Evans, Editor, *Modern Nonlinear Optics,* Second Edition, Wiley, **New** York, 2001, Part 1, p. 677-706.

127. Myron W. Evans, Director of the Alpha Foundation's Institute for Advanced Study, private communication, 2001.

128. Jackson, {22}, p. 249.

129. Jackson, {22}, p. 28.

130. Feynman, {7}, Vol. 1, p. 2-4.

131. Physicists often speak of spacetime as "space", since in Minkowski geometry the time portion is modeled along a separate spatial axis used in addition to the ordinary three spatial axes. So Feynman's "distortion" is actually Riemannian spacetime curvature.

132. As an aside, note that Feynman also utilized a dipolarity. There is always a scalar potential between the poles of any dipole. Further, this "dipole potential" decomposes according to Whittaker's 1903 decomposition{85}, as re-interpreted by Bearden's giant negentropy approach {12} in 2000, and as relates to Mandl and Shaw's 1984/1993 argument{19} that only the associated time-polarized photon and longitudinal photon together are observable, and then as the instantaneous scalar potential. If we accept Mandl and Shaw's powerful argument, then the Whittaker 1903 decomposition must be reinterpreted to be in accordance. That is what was done by the present author {12} to resolve the source charge problem as well as the source dipole problem. In addition, the resolution must be in accord with the theoretical and experimental findings of particle physics, to include the broken symmetry of the opposite charges on the ends of a dipole. The giant negentropy process uncovered by Bearden {12} fulfills all the necessary requirements. It also places the theme of this book — *Energy from the Vacuum: Concepts and Principles* — on a solid theoretical and experimental basis.

133. Feynman, {7}, vol. II, p. 1-3.

134. Jackson, {22}, p. 811-812.

135. We have previously postulated this feature where one second is equal to spatial EM energy compressed by the factor c^2, so that time is just extremely compressed EM spatial energy. See Bearden, {17}.

136. (a) Harter et al., {102}. It may well be that galloping due to time-density variation may provide an alternative explanation to quantum tunneling, which can readily occur at superluminal speed. E.g., see (b) A Enders and G. Nimtz, *Phys. Rev. Lett.,* Vol. 48, 1993, p. 632. Their superluminal tunneling experiments transmitted 8.7 GHz microwaves (free space wavelength 3.4 cm) traveling in a rectangular waveguide that contains a "barrier" section of reduced dimensions. Mozart's 40th Symphony was transmitted through the barrier waveguide as frequency

modulated microwaves, at an FTL group velocity of 4.7c, receiving audibly recognizable music from the microwave photons that survived their barrier passage.

137. (a) M. W. Evans, "The Link Between the Sachs and 0(3) Theories of Electrodynamics," in M. W. Evans, editor, *Modern Nonlinear Optics,* Second Edition, Wiley, New York, 2001, Part 2, p. 469-494; (b) — "0(3) Electrodynamics," in M. W. Evans, editor, *Modern Nonlinear Optics,* Second Edition, Wiley, New York, 2001, Part 2, p. 79-267.

138. E.g., Ryder, {90b}, p. 147 *et seq.*

139. Rodrigues and Lu, {93a}.

140. Rodrigues and Vaz, {93b}.

141. The reader should recall that to this day there is no satisfactory resolution of the wave-particle duality principle. Physicists finally just shook hands and agreed to quit fighting, since the hullabaloo was getting them nowhere. The problem does have a probable resolution, but not in Aristotelian logic.

142. Bearden, {12} discusses the relation between time-energy and 3-spatial energy, and reveals a more fundamental 4-symmetry between time energy flow and spatial energy flow. In this 4-symmetry EM flow, time symmetry in energy flow is broken and simultaneously 3-space symmetry in energy flow is broken, but 4-symmetry in energy flow is maintained. The basic transduction of time-energy into 3-space energy and vice versa is accomplished by the interacting charge.

143. Bearden, {15}, p. 11. In the paper we also gave a force-free definition of mass, in terms of the transduction of angular momentum across the 3-spatial boundaries of a mass particle.

144. Rigorously, *mass* does not and cannot emit a photon; *masstime* does. But an observable mass m can be absorbing and emitting **enormous** numbers of virtual photons while not yet having absorbed another observable photon and decayed to emit an observable photons. Differential (virtual) pieces of a "frozen" observable mass m are **indeed** changing continually into masstime and back into mass, by absorbing and emitting virtual photons per equation [14b], while the observable mass m has not yet completed its absorption and emission of an observable **photon** This "knitting" together of observable and virtual state interactions is one of the great centralizing features of physics.

145. The problem is due to the hoary old erroneous notion in mechanics that a separate force acts upon a separate mass to move it.

146. There appears to be no paper or textbook which calculates the actual "energy magnitude" of an entire potential or an entire field, including over all space occupied by it. All texts advance the "local intensity" indicator calculation, and then call that the "magnitude of the potential". That is a non sequitur of first rank. A skilled electrodynamicist will call it the intensity of the potential.

147. It appears that the only region where this statement is in dispute is in various models attempting to explain the creation of the universe. Since that is *a priori* a creation event, then physicists struggle with explaining how something (the existing universe) was created out of nothing (the absence of any slightest bit of the existing universe). We will leave that debate to the astrophysicists and cosmologists. The "accursed identity of opposites" has been the bane of philosophers for more than 3,000 years.

148. Ibrahim Semiz, "Black hole as the ultimate energy source," *Am. J. Phys.,* 63(2), Feb. 1995, p. 151.

149. David Halliday and Robert Resnick, *Fundamentals of Physics, Third* Edition Extended, Wiley, New York, 1988, Vol. 1, p. 518.

150. Stanley W. Angrist, "Perpetual Motion Machines," Sci. *Am.,* Vol. 2I8, Jan. 1968, p. 114-122.

151. Stanley W. Angrist, "Perpetual Motion," *Encyclopaedia Britannica,* Bicentennial Edition, Macropaedia Vol. 14, 1976, p. 102-105.

152. Lindsay and Margenau, {114}, p. 217.

153. Heaviside, {106}, p. 94.

154. Kraus, {76}, Figure 12-59, p. 578.

155. Bohren, {24}.

156 Paul and Fischer, {25}.

157. (a) V. S. Letokhov, "Laser Maxwell's Demon," *Contemp. Phys., 3*6(4), 1995, p. 235-243; (b) — "Generation of light by a scattering medium with negative resonance absorption," *Zh. Eksp. Teor. Fiz.,* Vol, 53. 1967, p. 1442 (English translation is in *Sov. Phys. JETP,* 26(4), Apr. 1968. p. 835-839); (c) "Stimulated emission of an ensemble of scattering particles with negative absorption," *ZhETF Plasma,* 5(8), Apr.

15, 1967, p. 262-265; (d) — "Double y - and optical resonance," *Phys. Lett. A,* Vol. 43, 1973, p. 179-180.

158. It is Heaviside's equations that are usually taught in University as "Maxwell's Equations," and labeled as Maxwell's equations in most undergraduate texts. In addition to Heaviside' work, similar changes were made by Gibbs in the United States and by Hertz in Europe. Every sophomore physics student and electrical engineering student should be given a short treatise on the background development of electrodynamics, and how, when, and by whom Maxwell's 20 quaternion equations were modified into the present four vector equations. The treatise should also point out what was omitted from Maxwell's theory in its truncations. It should be made very clear to every student that his elementary EM subject errs dramatically from physical reality when it assumes a flat local spacetime and an inert local vacuum. *By discarding the physically active environment of all Maxwellian systems,* only a highly limited subset of nature's electrodynamics is captured in the standard electrical engineering model. Specifically, systems far from equilibrium in their exchange with that environment are arbitrarily discarded. Those are precisely the Maxwellian systems permitted to exhibit coefficient of performance greater than unity. In short, the truncation of Maxwell's theory arbitrarily discarded all electrical windmills turning — or partially turning — in **a** free electrical wind.

159. As we pointed out, there are *tin electromagnetic systems at all in* that first class — and never have been — if we account for the arbitrarily-ignored Heaviside energy flow pouring from the source dipole and the energy flow pouring from every charge, and consider the situation prior to self-enforced implementation of Lorenz/Lorentz symmetrical regauging (by the closed current loop circuit). The remarks — pertaining to EM systems supposedly in equilibrium with the active vacuum — only apply if we neglect that Heaviside component and also neglect the source **charge** problem, and even then only if the circuit self-enforces that equilibrium in its exchange with the active vacuum. But since it has been so long **and** ubiquitously neglected, we will continue to speak of the "two classes" of EM systems from that "conventional" view. It is rather like our concession to electrical engineering, where the word "power" is universally **misused.** However, the reader should be aware of what we are doing and why.

160. See (a) Ilya Prigogine, *Nonequilibrium Statistical Mechanics,* Wiley-Interscience, New York, 1962; (b) — *From Being to* ***Becoming:***

Time and Complexity in the Physical Sciences, W. H. Freeman and Company, San Francisco, 1980; (c) D. Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998; (d) G. Nicolis and I. Prigogine, *Self-Organization in Non-Equilibrium Systems: From Dissipative Structures to Order through Fluctuations,* Wiley, New York, 1977. For a good educated layperson's overview, see (e) Gregoire Nicolis, "Physics of far-from-equilibrium systems and self-organization," Chapter 11 in {f} Paul Davies, Ed., *The New Physics,* Cambridge University Press, Cambridge, 1989, p. 316-347.

161. Actually, Ludvig Valentin Lorenz first performed the equivalent of symmetrical regauging of Maxwell's equations, only two years after Maxwell's seminal paper was published in 1865. See (a) Ludvig Valentin Lorenz, "On the identity of the vibrations of light with electrical currents," Phil. *Mag.,* Vol. 34, 1867, p. 287-301. Not much attention was paid to Lorenz's work, however, and the later symmetrical regauging of the Maxwell-Heaviside equations by H. A. Lorentz is what "stuck" and came to be used. For a proper historical discussion, see (b) J. D. Jackson and L. B. Okun, "Historical roots of gauge invariance," *Rev. Mod. Phys.,* Vol. 73, July 2001, p. 663-680.

162. This point is important. See Gabriel Kron, "The Frustrating Search for a Geometrical Model of Electrodynamic Networks," circa 1962. General Electric publication, p. 11-128. Quoting p. 114: *"...the missing concept of "open-paths" (the dual of "closed-paths") was discovered, in which currents could be made to flow in branches that lie between any set of* ***two*** *nodes. (Previously -following Maxwell - engineers tied all of their open-paths to a single datum-point, the 'ground'). That discovery of open-paths established a second rectangular transformation matrix... which cre****ated*** *'lamellar' currents... " "A network with the simultaneous presence of both closed and open paths was the answer to the author's years-long search."*

When we design and build a circuit or system in compliance with the Lorenz/Lorentz-regauged equations, it follows that *the circuit itself must contain some feature and function that self-enforces the Lorenz/Lorentz condition — at least during the excitation discharge of the energy that has been freely caught by the circuit. The Lorenz/Lorentz condition is not a law of nature, nor is it automatically applied without a specific circuit cause and function physically applying it.* That function is automatically applied by the standard closed current loop circuit.

As Kron stated, his long search for a true negative resistance finally succeeded when he discovered the "open path". Actually Kron had discovered that any two points in the universe at different potentials (charged to different intensities) produced a flow of EM energy. In other words, he had discovered in more classical EM terms the broken symmetry of opposite charges of differing charges, such as on the ends of any dipole or across any dipolarity.

163. Superconducting *sections* in a system may be loss-free, but the entire superconducting *system* has losses and exhibits only COP<1.0 overall, if the refrigeration burden is accounted. In the superconducting section alone, without accounting for the cooling overhead, the COP = 1,0 and the efficiency is 100% also.

164. After Jackson, {22}, p. 219-221; 811-812.

165. See Whittaker, {85}. Whittaker mathematically decomposes the scalar potential into a hidden "bidirectional" series of EM wavepairs in a harmonic sequence. Each wavepair consists of the wave and its phase conjugate. Thus any Aq) (as across the terminals of a dipolar electrical power source) *a priori* involves an ordered, hidden, bidirectional EM wave flow. I.e., d(fi)) *identically is* such a hidden bidirectional EM energy flow.

166. But see Bearden, {12}. In the paper we conceptually re-interpreted Whittaker's 3-space (observation assumed, hence effect waves) bidirectional wavepairs. Our reinterpretation yields a correlated harmonic half-set of converging *unobserved* causal EM longitudinal waves (time-polarized EM waves) in the time domain prior to interaction with charge so as to prepare for being observed, and a harmonic half-set of diverging EM longitudinal waves (after observation due to the convergent time-polarized EM waves interacting with the observable charges) in 3-space in all directions. We strongly stress that observation does not occur without the absorption of a scalar photon and the follow-on emission of a longitudinal photon. In a time-reversed situation (phase conjugate charge condition), to the observer it appears to have happened in the other direction.

167. Particularly see Mandl and Shaw, {19}.

168. However, Mandl and Shaw fail to totally identify the "combining mechanism". They do not account the detailed interaction of the detecting/observing unit point charge, and thus fail to clearly account for the absorption of the incoming time-polarized wave or photon, the

transduction of that excitation energy of the charge into longitudinal EM wave/photon energy, and the subsequent emission of that excitation energy in 3-space in all directions as longitudinal EM waves/photons. So Mandl and Shaw do not clearly account for *photon (or wave) polarization transduction by the ubiquitously assumed charge.* The "causal" time-polarized EM wave or photon comes in and is absorbed by the detecting charge or dipole, then the absorbed energy is transduced and re-emitted in 3-space as the longitudinally polarized EM wave or photon in 3-space. (Actually, since all space is charged, what is really emitted into 3-space is the giant circulation of EM energy between time and 3-space, which is a more rigorous statement of what "propagation of EM energy through 3-space actually is). Recognition of these missing interactions and transductions allowed a solution at last to the long-vexing problem of the source charge, often called the greatest problem in both quantum and classical electrodynamics, as given initially in Bearden, {12}. Now, of course, a greater problem has reared its head: the notion that EM energy continuously propagates in 3-space is a total non sequitur, along the lines objected to by Romer {84}. We therefore have some very fundamental rethinking to do for all of classical physics. This is really where the problem of wave-particle duality has pointed since the beginning.

169. Asymmetrical self-regauging collects just a tiny bit more of the enormous energy that is flowing outside the conductors and usually wasted. E.g., if the self-regauging EM system collects as **jϕ** some 10^{-12} of the energy flow instead of the more usual 10^{-13}, the system can have a COP =10, *if it* does not then expend half the collected energy to destroy the sourcc dipole faster and require its reconstitution more energetically. And so on.

170. Jackson, {22}, p. 220-223.

171. So-called "canceling" opposing EM fields are actually produced, which sum to a *vector zero system* with respect to translation, which the electrodynamicists erroneously discard by assumption. We point out, but do not further pursue, that the locally produced *field energies* of the opposing fields in a zero-summed vector system remain and add, even though the fields offset each other *translationally.* The local energy of the field is proportional to the square of the local magnitude, and that is always positive regardless of field orientation. Thus "trapped" and structured EM stress energy has been localized in spacetime in the symmetrical regauging of CEM. This rigorously is a gravitational change

and therefore a local curvature of spacetime *a priori.* Indeed, the field energy remains in a zero-vector summation system, and has been changed into stress potential energy. It represents continuous internal work being performed on the system to stress it, by energy continuously fed in from the active vacuum environment but only in appositive fields of equal magnitude. The discarding of the zero vector summation system thus is a discarding of an available stress potential and its energy, a rotation of the system frame out of the laboratory frame, and a continuous exchange of energy from the vacuum that performs internal work upon the system but no external work in the load. The energy is not translational, but it is there nonetheless, and it causes a spacetime curvature change. The electrodynamicists have largely ignored the fact that gauge symmetry transformations are accompanied by local gravitational (spacetime curvature) changes that violate the prevailing flat local spacetime assumption of classical electrodynamics.

172. See (a) H. J. Josephs, "The Heaviside papers found at Paignton in 1957," *IEE Monograph No. 319,* Jan. 1959, p. 70-76. See also (b) E. R. Laithwaite, "Oliver Heaviside — Establishment Shaker," *Elec. Rev.,* 211(16), Nov. 12, 1982, p. 44-45. Heaviside became aware of this gravitational relationship, and worked out a draft theory of electrogravitation based on his energy-flow theory (which added nondiverged closed-loop circulations of energy which vectorially summed to net zero). Those "trapped EM energy flow loops" were *gravitational,* in Heaviside's concept. His hand-written notes on his eleelrogravitation theory were found long after his death, beneath the floorboards of his little garret apartment. In honor of Heaviside, we have nominated his huge nondiverged energy flow component — erroneously discarded by Lorentz — as the source of the extra gravity in the arms of the spiral galaxies, holding them together. In short, we have proposed that as the solution to the so-called "dark matter" problem.

173. Physicists love symmetry, turn to it at every opportunity, and will do almost any mathematical manipulation to obtain it, because they regard it as "beautiful." And so it is. Performing work, however, requires breaking symmetry to produce excess energy, which interacts with mass to produce a force, and then requires using that force to do the work. If we wish to build a system that produces more work output than the work that we ourselves have to do upon it to run it, then *a priori* that system must be *asymmetrical* and — to many physicists — *ugly.* Let us all drink a toast to more efficient *ugly* asymmetrical systems that also do not require **energy.**

input by the operator, but only by the active environment! Give us "ugly" electrical windmills that turn freely in electrical winds, rather than the inane "beautiful" present systems which destroy their energy winds faster than hey power their loads.

174. While Lorentz appears to have done it earlier, one can see his use of the surface integration in (a) Lorentz, {109}. In 1896 he also included it, using a more general theorem by Volterra, so that Poynting's theorem arises as a special case; see (b) H. A. Lorentz, *Versl. K. Akad. W. Amsterdam,* Vol. 4, 1896, p. 176.

175. This is usually performed by the source dipole.

176. Thus the long and tedious effort of my colleagues and I to create circuits that asymmetrically self-regauge their potential difference (source voltage) without allowing all the spent load current to pass back through the primary power source dipole and do work upon it to scatter the charges and destroy its dipolarity.

177. Again we strongly stress that asymmetrically regauging a potential in the system also regauges the local vacuum potential and the intensity of the vacuum potential's bidirectional EM energy exchange with the system. It alters the local curvature of spacetime. It directly alters the stored energy of the system and of the supersystem. See again Whittaker, {85}, to see this clearly. See particularly our use of the *supersystem* concept in Chapter 8, and as mentioned in Chapter 1 and the present Chapter 2.

178. Circuits and electrical power systems universally and routinely break Lorentz symmetry for *excitation* (potentialization) of the system at the initial onset. Just to potentialize a system in theory costs nothing, although it freely changes the potential energy of the system. The gauge freedom axiom of quantum field theory also guarantees that such change of the potential energy of the system is essentially "for free". It conclusively proves that (i) asymmetrical self-regauging exists and is done in electrical circuits, since the source dipole in the power supply accomplishes that initial self-regauging of the external circuit completely for free once the dipole is made, and (ii) the Maxwell-Heaviside equations do permit asymmetrical self-regauging, in violation of Lorentz's arbitrary condition prohibiting it.

179. This is the patent-pending Bedini process, which is processing through the U.S. Patent Office as we go to press.

180. Lee, {73}, p. 184.

181. B. P. Kosyakov, "Radiation in electrodynamics and in Yang-Mills theory," *Soviet. Phys. Usp.*, 35(2), Feb. 1992, p. 135-142. Quotes are from p. 135 and p. 141. Our proposed solution to the problem posed by Kosyakov (which is actually the broken symmetry of a charge) is that negative charges pour out positive energy and positive charges absorb positive energy. We "see" the positive charge as pouring out "reversed direction positive field energy" after we interact the energy with a laboratory charge to "observe". Considering the well-known polarization of the vacuum by any charge, one realizes that a 4-symmetry circulation of the EM energy is what is "emitted" and spreads throughout all space, since the vacuum polarization is also spreading. The actual energy flow is between the time domain and 3-space and return, at any elemental dipole in the polarization of the vacuum that is spreading.

182. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

183. Prof. William Keel, quoted in James McWilliams, "Through a Galaxy, Darkly," *Huntsville (Alabama) Times,* May 14, 2000, p. A-18.

184. Charles Seife, "Peering Backward to the Cosmos's Fiery Birth," *Science,* vol. 292, June 22, 2001, p. 2238.

185. M. P. Silverman, *And Yet It Moves: Strange Systems and Subtle Questions in Physics,* Cambridge University Press, Cambridge, 1993, p. 127.

186. Matthew Colless, Australian National University, one of the survey team leaders in the project leading the world in mapping galaxies, Mar. 2001.

187. (a) Lee, {13a}; (b) Lee, Oehme and Yang, {13b};

188. Wu et al., {14}.

189. Bearden, {12}.

190. Mandl and Shaw, {19, 168}.

191. Whittaker, {85}.

192. Gabriel Kron, "Invisible dual (n-1) networks induced by electric 1-networks," *IEEE Trans, on Circuit Theory,* CT-12(4), Dec. 1965, p. 464-470.

193. Kron, {162}. The quote is from p. 114.

194. Steven Weinberg, *Dreams ofa Final Theory,* Vintage Books, Random House, 1993, p. 109-110.

195, See Josephs, {172a}. Heaviside's hand-written notes contained his theory of electro-gravitation, based on his theory of energy flow. The papers were found beneath the floorboards in his little garret apartment.

196. Laithwaite, {172b}. Laithwaite feels that Heaviside's postulation that a flux of gravitational energy combines with the (ExH) electromagnetic energy flux, could shake the foundations of physics. Extracting from Laithwaite; *"Heaviside had originally written the energy flow as S = (ExH) + G, where G is a circuital flux. Poynting had only written S = (ExH). Taking p to be the density of matter and e the intensity of a gravitational force, Heaviside found that the circuital flux G can be expressed as pu -ce, where u represents the velocity of p and c is a constant. "*

197. Hence the title of cited reference {12} above.

198. Melvin H. Miles and Benjamin F. Bush, "Radiation measurements at China Lake: Real or Artifacts?", *Proc ICCF-7 (International Conference on Cold Fusion* — 7, Vancouver, BC, Canada, Apr. 1998, p. 101.

199. Bearden, {17}.

200. See P. V. Elyutin, "The Quantum Chaos Problem," *Sov. Phys. Usp.*.Vol. 31, No. 7, 1988, p. 597-622. Elyutin discusses the crisis in quantum mechanics because of the missing chaos (hidden order). Quantum mechanics is known to be wrong unless this order can be found, because otherwise it does not predict the ordered macroscopic universe. Thus any proposed solution to the quantum chaos problem is worthy of Investigation, even if it turns out to be in error.

201. Unfortunately, entropy is one of those concepts in physics for which there are several differing major views. We consider entropy to be comparable to a positive resistor: organized energy goes in, disorganized energy comes out. Entropy is usually taken to be a measure of the disordering of energy. It does not *eliminate* the energy; it just makes it more difficult to *use* it. For our work in energy from the vacuum, we take the very simple view that a negentropic process is like a negative resistor: it receives energy in a form unusable to us, transforms it, and outputs it in

a form that is usable. In short, a negative resistor receives disordered energy and reorders it, so to speak. So it exhibits negentropy. We completely avoid the various notions of "information" and attempts to equate information and energy. We do point out, however, that a time-reversal process in one form or another is usually involved. In that sense, e.g., Newton's third law would be a negentropic process and involve time reversal in the mechanism generating the reaction-causing agent.

202. H. E. Puthoff, "Source of Vacuum Electromagnetic Zero-Point Energy," *Phys. Rev. A,* 40(9), Nov. 1, 1989, p. 4857-4862.

203. Whittaker, {91}. The paper was published in 1904 and orally delivered in 1903. This paper initiated superpotential theory.

204. As Whittaker showed in 1903, {85}, the scalar potential is actually a harmonic set of bidirectional EM longitudinal EM wavepairs, where each pair is composed of a longitudinal EM wave and its phase conjugate replica. Only because classical electrodynamicists have erroneously defined the field and potential as their own reaction cross sections with a unit point static charge, has the "static" potential been misidentified as a *scalar* entity, which it is not. The energy diverged from a uniform potential, around a fixed static point unit charge, is actually the set of divergences around the intercepting charge of the energy flows of all those EM waves comprising the potential. The sum total of all these individual wave divergences indeed has a scalar magnitude, but the magnitude of the total energy divergence from the potential is neither the potential itself nor its magnitude.

205. We point out the obvious: A "scalar" mass in 3-space actually him a time-vector since it moves through time continually, just to continue to exist. Further, it involves a special form of energy (energy compressed by c^2) which time is taken to be. Since we may choose any form of energy we wish by simple transduction, we may take it as compressed EM energy. So the mere continued existence of any mass proves conclusively that EM energy can and does ubiquitously flow to, from, and through the time dimension. The combined continued existence of numerous masses **proves** conclusively that the flow of time can have a myriad internal electromagnetic energy flows. An equilibrium between (i) an inflow of EM energy to a transducer from the time dimension, and (ii) an **outflow of** EM energy in 3-space from the transducer, will be seen as a discrete excitation (potential energy) associated with the transducer. Hence **the** notion of the charge. Every charge in the universe may be said to pour out

energy (negative charge pours out positive energy; positive charge absorbs positive energy — the equivalent of pouring out negative energy — and any charge results in a polarization of the vacuum. Hence what we call "spacetime" or "vacuum" is in fact a giant circulation of energy from the time domain to 3-space, and back to the time-domain. In short, a giant negentropy process continually sustains the active universe and all its components and processes.

206. Evans et al., {79}.

207. Semiz, {148}.

208. Weinberg, {194}, p. 286.

209. The available internal energy of a generator is the transduced shaft energy we input to it, say by turning the generator shaft with a steam turbine or hydro turbine. The shaft mechanical energy input is transduced into internal magnetic field energy once the rotor rotates. This internal magnetic energy is then dissipated on the internal charges inside the generator, forcing the positive charges in one direction and the negative charges in the other, thereby making the source dipole between the terminals of the generator. The available internal energy of a battery is the chemical energy possessed by it at any given time, and available for performing work on the internal charges to force them apart to form a dipole between the plates (and the battery terminals).

210. Kraus, {76}. Figure 12-60, a and b, p. 578 shows a good drawing of the Poynting (intercepted) energy flow being withdrawn into the conductors from the surrounding space. It does not show the Heaviside nondiverged component remaining, which is even larger, not diverged into the conductors, and wasted. The reason for the remaining Heaviside non-diverged component is that the electrons precess laterally across the conductor, thus withdrawing into the conductor only a small nearby section of the fields of the surface electrons — the ones (plus their fields extending into space) which intercept the impinging energy flow. Because of the short distance radially across the conductor's cross section, a surface electron has very limited "field and field energy withdrawal" length.

211. (a) J. D. Jackson, "Surface charges on circuit wires and resistors play three roles," *Am. J. Phys.,* 64(7), July 1996, p. 855-870. See also (b) Mark A. Heald, "Energy flow in circuits with Faraday emf," *Am. J. Phys., Vol.* 56, 1988, p. 540-547; (c) — "Electric fields and charges in elementary circuits," *Am J Phys.,* 52(6), June 1984, p. 522-526. The

surface charges in the conductors of a circuit are of enormous importance to the powering of the circuit.

212. See T. E. Bearden, "Energy Flow, Collection, and Dissipation in Overunity EM Devices," *Proc. 4th Internat. Energy Conf.*, Academy for New Energy, Denver, CO, May 23-27, 1997, p. 5-51. In Figure 5, p. 16 the fraction of the Poynting energy flow that is intercepted and collected by the circuit is roughly shown to be on the order of 10^{-13} of the entire Poynting energy flow available. That number is based on a very crude estimating procedure, but even if off by several orders of magnitude, it shows that the Heaviside nondiverged energy flow component is orders of magnitude greater than the diverged Poynting component, for single-pass of the energy and collection only once.

213. Rigorously, there is no *power* in an EM energy flow, regardless of how great in magnitude, if it is not altered in form or diverged. That is because "power" is rigorously the time rate of doing work (time rate of *changing the form* of energy), not the time rate of energy flow per se. Exactly, the Heaviside dark energy flow component was some 10 trillion joules per second in magnitude, but it had zero watts of power. Unfortunately, electrical engineers just loosely refer to nondiverged **energy** flow as "power", regardless of the non sequitur. Hence inane terms **widely** used in electrical engineering (and texts) such as "draw power from **the** source", etc. Rigorously, *the power* (rate at which energy is changed in form) is developed locally in each component having losses or performing work — i.e., changing the form of the energy.

214. (a) T. E. Bearden, "On the Principles of Permissible Over **Unity** EM Power Systems," *J. New Energy,* 4(2), Fall 1999, p. 16-39; (b) Bearden, {17}; (c) — "Use of Asymmetrical Regauging and Multivalued Potentials to Achieve Overunity Electromagnetic Engines," *J. New Energy,* 1(2), Summer 1996, p. 60-78; (d) — "Regauging and **Multivalued** Magnetic Scalar Potential: Master Overunity Mechanisms," *Explore,* 7(I), 1996, p. 51-58; (e) — "The Master Principle of EM Overunity and the Japanese Overunity Engines," *Infinite Energy,* 1(5&6), Nov. **1995-Feb,** 1996, p. 38-55; (f) — "Use of Regauging and multivalued Potentials to Achieve Overunity EM Engines: Concepts and Specific Engine Examples," *Proc. Internat. Sci. Conf, "New Ideas in Natural Sciences," St. Petersburg, Russia, June 17-22, 1996, Part I: Problems of Modern Physics,* 1996, p. 277-297.

215. In an oblique kind of sense, this is even "recognized" to some extent by leading electrodynamicists. E.g., quoting Jackson, {22}, p. 237: *"...the Poynting vector is arbitrary to the extent that the curl of any vector field can be added to it. Such an added term can, however, have no physical consequences. "* So Jackson follows Lorentz's lead that any additional non-intercepted and uncollected energy flow permissibly accompanying the Poynting intercepted and collected component, has no physical significance. Both Lorentz and Jackson err as to the *finality* of the statement. The accompanying Heaviside component can have no physical significance *unless it is intercepted and collected by means other than the single circuit or mechanism considered, and the single pass of the entire energy flow onto and past the receiving entity (circuit, charge, etc.).*

216. E.g., see Poynting, {4a. 4b}.

217. Heaviside, {5a}.

218. Heaviside, {5b}.

219. Heaviside, {5c}. Heaviside discusses the Faraday-Maxwell ether medium, outlines his vector algebra for analysis of vectors without quaternions, discusses magnetism, gives the EM equations in a moving medium, and gives the EM flux of energy in a stationary medium. On p. 443, he credits Poynting with being first to discover the formula for energy flow, with Heaviside himself independently discovering and interpreting this flow a little later by himself in an extended form.

220. J. H. Poynting, "On the connexion between electric current and the electric and magnetic inductions in the surrounding field," *Proc. Roy. Soc. Lond.* Vol. 38, 1984-85, p. 168.

221. Oliver Heaviside, *Electrical Papers,* Vol. 2, 1887, p. 94.

222. (a) T. E. Bearden, "On Extracting Electromagnetic Energy from the Vacuum," *Proc. IC-2000,* ibid., 2000; (b) Bearden, {38a}.

223. Lorentz is believed to have done this earlier (perhaps circa 1890s). In 1902 he published the method in a book, which strongly implies it was first done in an earlier paper. See Lorentz, {109}. Figure 25 on p. 185 shows the Lorentz concept of integrating the Poynting vector around a closed cylindrical surface surrounding a volumetric element. Many thanks to Marcus Reid for furnishing a copy of the actual Lorentz reference from a library in Leipzig, Germany.

224. E.g., see (a) Panofsky and Phillips, {110}; (b) W. Gough and J. P. G. Richards, *European J. Phys.*, Vol. 7, 1986, p. 195.

225. Panofsky and Phillips, {110}, p. 180.

226. D. S. Jones, *The Theory of Electromagnetism,* Pergamon Press, Oxford, 1964, p. 52.

227. Jackson, {22}, p. 237.

228. W. M. Schwarz, *Intermediate Electromagnetic Theory,* John Wiley & Sons, New York, 1964, p. 280-281.

229. Jones, {226}, p. 53.

230. For typical references, see (a) J. Slepian, *Am. J. Phys.,* 19, 87 (1951); (b) Mario lona, *ibid.,* 31, 398 (1963); (c) Udo Backhaus and Klaus Schafer, *ibid.,* 54, 279 (1986); (d) C. J. Carpenter, *IEE Proc. A* (UK), 136A(2), Mar. 1989, p. 55-65; (e) J. A. Ferreira, *IEEE Trans. Edit.,* 31 (4). 1988, p. 257-264; (f) Mark A. Heald,Am. *J. Phys.,* 56(6), 1988, p. 540-547. The debate has also appeared in many other leading journals, e.g., (G) T. H. Boyer, *Phys. Rev. D,* 25, 3246 (1982). Interestingly, (h) M. Abraham and R. Becker, *The Classical Theory of Electricity and Magnetism,* Blackie. London, 1932, p. 146 and p. 194 give two examples of the controversy over the Poynting vector. Finally, see (i) D. F. Nelson *Phys. Rev. Lett.,* 76(25), June 17, 1996, p. 4713-4716 for advanced **work** requiring a greater generalization of the Poynting vector.

231. Sweet and Bearden, {29a}. Sweet's device produced 500 watts lit a 33 microwatt input, which is approximately a COP = $1.5x10^6$. By steadily increasing the COP to $3x10^6$, the weight of the Sweet unit was reduced smoothly, reaching 90% reduction. The results of this highly successful anti-gravity experiment are reported in the paper. Unfortunately, Sweet later died and never fully revealed the activation secret by which barium ferrite magnetic materials could be triggered **into** strong self-oscillation at 60 Hertz or other ELF frequency.

232. Weaker self-oscillation of such permanent magnetic materials at higher frequency is known. E.g., see (a) V. S. L'vov, *Non-Linear Spin Waves,* Moscow, 1987; (b) L'vov, {29c}. See also (c) A. G. **Gurevich,** and G. A. Melkov, *Magnetization Oscillations and Waves,* CRC **Press,** Boca Raton, Florida, 1996. The latter authors give hundreds of **additional** references.

233. E.g., (a) M. W. Evans, P. K. Anastasovski, T. E. Bearden *et al,* "Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 589-593; (b) — "Anti-Gravity Effects in the Sachs Theory ol Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 601-605; (c) Evans et al., {38b}; (d) Evans et al., {38c}; (e) -- "Operator Derivation of the Gauge Invariant Proca and Eehnert Equation: Elimination of the Lorentz Condition," *Found. Phys.,* 39(7), 2000, p. 1123-1130; (f) - "O(3) Electrodynamics from the Irreducible Representations of the EInstein Group," 2001, *Found. Phys. Lett.,* 15(2), Apr. 2002, p. 179-187.; (g) - "Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys.* Lett., 13(3), June 2000, p. 289-296; (h) Evans et al., {113b}; (i) Evans et aI., {113a}; (j) - "On the Representation of the Maxwell-Heaviside Equations in Terms of the Barut Field Four-Vector," *Optik* 111(6), 2000, p. 246-248; —"Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* (in press); (k) — "The New Maxwell Electrodynamic Equations: New Tools for New Technologies. A Collection of 60 papers from the Alpha Foundation's Institute for Advanced Study. Published as a Special Issue of the *J. New Energy,* 4(3), Winter 1999. See also (1) Bearden, (222b}; (m) Bearden, {38a}; (n) Bearden, {43d}; (o) Bearden, {43a}; (p) Bearden, {30}; (q) Bearden, {40}; (r) Bearden {214}; (s) Bearden, {1214e}; (t) - "The Final Secret of Free Energy," *Magnets,* 7(5), May 1993, p. 4-26; (u) Bearden, {214f}.

234. Bearden, {40).

235. Terence W. Barrett and Dale M. Grimes, (a) Preface, p. vii-viii, in (b) *Advanced Electromagnetism: Foundations, Theory, and Applications,* Terence W. Barrett and Dale M. Grimes (eds.), World Scientific, Singapore, 1995.

236. I. Prigogine, "Irreversibility as a symmetry-breaking process." N*ature,* Vol. 246, Nov. 9, 1973, p. 67-71.

237. Paul Davies, *Superforce: The Search for a Grand Unified Theory of Nature,* Simon and Schuster, New York, 1984, p. 105.

238. Clarke, {55}, p. 38.

239. T. D. Lee, *Symmetries, Asymmetries, and the World of Particles,* U. Wash. Press, Seattle, 1988, p. 11.

240. That is, it does not exist *openly* in the open literature. Several inventors do have working laboratory COP>1.0 experiments, and at least three of them have models which could rapidly be developed into production systems — say, in one year or less. However, the exact and detailed construction details are understandably maintained as deeply proprietary by the inventors, until their intellectual property rights can be secured, including foreign patent rights. On the other hand, many other inventors mistakenly believe they have COP>1.0 systems, but do not because of instrumental error, considering nonsinusoidal waves as measurable with RMS meters, or effects covered in Chapter 6 which they do not comprehend. In addition, there have been those more interested in selling stock than in producing a working system.

241. But see Kawai, {31}. The only available *open* details approaching the "kit of parts and instructions" point is the Kawai process, and it will work just as in the patent if carefully built. If very efficient photon-coupled switching of the flux path is used for build-ups, and if very high efficiency (0.7 to 0.8) magnetic motors are modified with the Kawai process, one can expect the COP of the modified motors to be double the conventional efficiency of the original base motor. So one can obtain COP = 1.4 to 1.6, e.g., by using the necessary and available Hitachi high efficiency motors (one at 0.7 and the other at 0.8) to start with. Hitachi engineers physically tested two high efficiency Hitachi motors modified hy Kawai, at COP = 1.4 and COP = 1.6 respectively. Unfortunately, the Kawai process, engines, and company have been seized by what can **only** be described as the Yakuza (Japanese Mafia) and are being held off **the** market. The seizure was accomplished here in Huntsville, Alabama, witnessed by the CTEC board of directors and I, when Kawai and **CTEC** were in final negotiations to rapidly develop and market the Kawai **system** Several other COP>1.0 Japanese engines have also been seized and withheld by the Yakuza.

242. Jed Z. Buchwald, *From Maxwell to Microphysics,* University of Chicago Press, Chicago and London, 1985, p. 44.

243. Whittaker, {85}.

244. Bearden, {12}.

245. Mandl and Shaw, {19}.

246. Whittaker, {91a}.

247. To see how arduous a search Kron himself made prior to athieving a proper model for an electrical circuit, see Kron, {162}. The quote is from p. 114.

248. For example, Kron, {192}.

249. P. A. M. Dirac, "A theory of electrons and protons," *Proc. Roy. Soc. Lond. A,* 126(801), Jan. 1, 1930, p. 360-365. Here Dirac introduced the vacuum energy — the electron sea filled with electrons occupying negative energy states. The holes are treated today as the positive electrons. This concept is now referred to as the "Dirac Sea," and the concept has been expanded to include negative energy "holes" for other kinds of particles. Dirac first attempted to identify these holes with protons, until later the positron was discovered.

250. Evans, {137b}.

251. (a) Evans, {137a}; (b) — "The Link Between the Topological Theory of Ranada and Trueba, the Sachs Theory, and 0(3) Electrodynamics," *ibid.,* vol. 2, p. 495-499.

252. (a) Sachs, {126a). The Sachs theory provides a great generalization of general relativity and electrodynamics reaching from the quarks and gluons to the entire universe. See also (b) Sachs, {126b}.

253. Indeed, the symmetrizing of Maxwell's equations was first accomplished by Lorenz. See (a) Lorenz, {161a}. In this paper Lorenz gave essentially what today is called the Lorentz symmetrical regauging. AS pointed out by Barrett, the Lorenz paper gave f(t-r/c) functions. Fitzgerald said that Lorenz's functions were essentially the same as his, and Fitzgerald became a leading proponent of "retarded potentials". But it is believed that Fitzgerald was unaware of Lorenz's work until the 1880's, so he is given credit for parallel development. Some people speak of Fitzgerald-Lorenz functions. This is a regauging, but the term "gauge" (inspired by railroad gauges) was first introducec by Hermann Weyl in the 1900s. He used it for a change in length and was quickly countered by Einstein - Weyl's theory was not relativistic. The idea resurfaced in the 1920s when quantum theory was being formulated, but this time it meant "change in phase" and not "change in length". At any rate, when the renowned H. A. Lorentz performed the symmetrizing of the Maxwell-Heaviside equations in the 1880s. it was adopted immediately and has been rather universally applied since then. See also (b) Jackson and Okun {161b} for a history of gauge invariance.

254. (a) Prigogine, {160a}; (b) G. Nicolis and I. Prigogine, {160d}; (c) T. Petrosky and I. Prigogine, "Quantum Chaos, Complex Spectral Representation and Time-Symmetry Breaking," *Chaos, Solitons, and Fractals,* Vol. 4, 1994, p. 311-359; (d) D. Kondepudi and I. Prigogine, "Thermodynamics, Nonequilibrium," *Encyclopedia ofApplied Physics,* Vol. 21, 1997, p. 311-337; (e) D. Kondepudi and I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, Chichester, 1998; (f) T. Petrosky and I. Prigogine, "Laws of Nature, Probability and Time Symmetry Breaking," *Physica A,* Vol. 263, 1999, p. 528-539.

255. (a) G. Nicolis, and I. Prigogine, *Exploring Complexity,* Piper, Munich. 1987; (b) — *Exploring Complexity: An Introduction,* Freeman, New York, 1989; (c) G. Nicolis, {160e}.

256. Letokhov, {157a-157d}.

257. Bohren, {24}.

258. Paul and Fischer, {25}.

259. E.g., see (a) Patrick et al., {37}. For rigorous explanations of **how** the energy is extracted from the vacuum, see (b) Evans et al., {38b}; (c) Evans et al., {38c}. The vacuum can and will furnish an electromagnetic "wind" from which useful EM energy can be extracted. Indeed, every source charge and dipole already extracts and continuously pours out EM energy from the vacuum, thereby establishing its fields and potentials **(and** their energy) across the universe.

260. A relationship between gamma bursters and a big bang is suggested from some recent observations. On Jan. 23, 1999 the Hubble Space Telescope caught an optical flash simultaneous with an initial gamma burst. When calculated, this burst was second in energy only to the big bang itself, and generated about 3.4 x 10^{54} ergs, which is the equivalent of converting two solar masses instantaneously into energy. There was also no polarization of the afterglow, which has been taken as evidence of beaming. However, such beaming could also be due to the beaming involved in the intense iterative phase conjugate reflections of the asymmetrical self-regauging process in intensely scattering, optically active, energy-amplifying media.

261. Heaviside, {5a-5c}.

262. Josephs, {172a}; Laithwaite, {172b}.

263. (a) Diederik Wiersma and Ad Lagendijk, "Laser Action in Very White Paint," *Physics World,* Jan. 1997, p. 33-37; (b) K. Totsuka, G. van Soest, T. Ito, A. Lagendijk, and M. Tomita, "Amplification and diffusion of spontaneous emission in strongly scattering medium,"*J. Appl. Phys.,* Vol. 87, 2000, p. 7623-7628.

264. (a) A. Lagendijk, J. Gomez Rivas, A. Imhof, and R. Sprik, "Propagation of light in disordered semiconductor materials," in *Photonic Crystals and Light Localization in the 21st Century,* edited by C. M. Soukoulis, Kluwer, Dordrecht, 2001, p. 447-473; (b) H. Cao *et at,* "Spatial Confinement of Laser Light in Active Random Media," *Phys. Rev. Lett., 8*4(24), 12 Jun. 2000, p. 5584-5587.

265. We also point out the possibility of the accumulator at each end being an intensely scattering, photoreactive, energy-amplifying material, a subject discussed a bit later in this chapter.

266. (a) Werner Triftshauser *et al, Phys. Rev. Lett.,* Vol. 87, 2001. An overview is given by (b) Philip Ball, "The Positron Probe," *Nature,* Vol. 412. Aug. 23, 2001, p. 764.

267. For example, see Letokhov, {157a-157d}.

268. Jackson, {22}, p. 363, third footnote.

269. Kraus, {10}, p. 4.

270. Marconi's patent is Gugliehno Marconi, "Transmitting Electrical Signals," U.S. Patent 586,193, July 2, 1897. Filed Dec. 7, 1896, with 56 claims. This patent was overturned in 1943 by the U.S. Supreme Court, and prior work by Nikola Tesla for transmission of EM signals was upheld. Most textbooks and journals, etc. to this day continue to falsely credit Marconi with discovery of the transmission of EM signals and thus the discovery of radio. This is not to discredit the important work that Marconi did; without him, signal transmission would have been much longer in coming.

271. United States Reports: Cases Adjudged in the Supreme Court of the United States, Vol. 320, Marconi Wireless Telegraph Co. of America vs. United States, June 21, 1943, p. 1-80. The U.S. Supreme Court ruled that Tesla's fundamental radio patents had anticipated all other contenders, including Marconi.

272. *Dictionary ofScientific and Technical Terms,* Daniel N. Lapedes (editor in chief), McGraw-Hill, New York, 1974, p. 300. On page 517 of the second and expanded edition (1978), the same definition of electric field appears and is unchanged.

273. H. J. Gray and Alan Isaacs, *A New Dictionary ofPhysics,* Longman, London, second edition, second impression, 1976.

274. Feynman, {7}, p. 2-4.

275. Feynman, {7}, p. 12-2.

276. P. K. L. Drude, *Ann. Physik,* Vol. 1, p. 566; Vol. 3, 1900, p. 370, 869. Here Drude first develops a theory of conductivity in a metal based on the assumption that there is a gas of free electrons in the metal. This gas of conduction electrons is often referred to as the "Drude gas." Note that this was shortly after J. J. Thomson's discovery of the electron in 1897.

277. E.g., see Jackson, {211a}.

278. Romer, {84}, endnote 24, p. 109.

279. H. Winfield Secor, "Tesla Maps Our Electrical Future."

280. Nikola Tesla, "Pioneer Radio Engineer Gives Views on Power," *New York Herald Tribune,* Sep. 11, 1932.

281. Nikola Tesla, "The True Wireless, *Electrical Experimenter,* May 1919.

282. Burford and Verner, {35a}. Quote is from p. 281.

283. That is, it does not exist *openly* in the open literature. Several inventors do have working laboratory COP>1.0 experiments, and at **least** three of them have models which could rapidly be developed into production systems — say, in one year or less.

284. But see Kawai, {31}.

285. Barrett, {124}.

286. (a) Terence W. Barrett, "Active Signalling [sic] Systems," U.S. Patent No. 5,486,833, Jan. 23, 1996; (b) — "Oscillator-Shuttle-Circuit (OSC) Networks for Conditioning Energy in Higher-Order Symmetry Algebraic Topological Forms and RF Phase Conjugation," U.S. Patent No. 5,493,691, Feb. 20, 1996.

287. No proper scientific treatise on the Moray device has ever been published, to our knowledge, nor has a proper theory of its operation been developed. Anecdotal material and some technical material is contained in (a) T. H. Moray, *The Sea of Energy,* 5th Edn., Salt Lake City, 1978 (foreword by T. E. Bearden). The fourth edition of the book, titled (b) *The Sea of Energy in Which the Earth Floats,* contains additional material but is long out of print. Additional information is given in (c) Keith Tutt, *The Search for Free Energy,* Simon & Schuster, 2001, p. 33-68. (d) Paul Brown's study of the Moray work is "The Moray Energy Device: Operational Parameters, Design Criteria, and Considerations," date unknown. Researcher and inventor Bruce Perreault has based much of his own work with his radiant energy valve on Moray's work. Perreault proposes a different interpretation. In the present book we have tried to point out the various aspects of the Moray device that resemble other presently known effects — most of which were still unknown in the 1920s and 1930s when Moray's fundamental work was accomplished.

288. Moray, {34b} "Electrotherapeutic Apparatus," U.S. Patent No. 2,460,707, Feb. 1, 1949. This patent clearly shows the complexity of the various special tubes containing his pellets that Moray developed.

289. H. *C.* Dake and Jack DeMent, *Fluorescent Light and Its Applications,* Chemical Publishing Company, Inc., Brooklyn, New York, 1941, p. 51-52 gives a succinct discussion of Stokes emission and anti-Stokes emission.

290. Sir G. G. Stokes, "On the Change of Refrangibility of Light," a memoir, 1852.

291 For example, in 1935 Prileshajewa showed that there is an energy difference involving as much as 1.1 V between the exciting light and the fluorescence of aniline vapor.

292. V. S. Letokhov, {157a-157d}.

293. Letokhov, *ibid.*

294. E.g., (a) R. N. Bhargava and D. Gallagher, "Optical Properties of Manganese-Doped Nanocrystals of ZnS," *The Am- Phys. Soc,* 72(3), Jan. 17, 1994; (b) I. A. Izmailov et al., "Generation of Stimulated Noncoherent Radiation in Light-Scattering Media Exhibiting Chemical Reactions," *Soy. J. Quantum Electronics,* 12(5), May. 1982, p. 588-594; (c) R. Pappalardo and A. Lempicki, "Brillouin and Rayleigh Scattering in Aprotic Laser

Solutions Containing neodymium," *J. Appl. Phys.,* Apr. 1992, p. 1699-1708.

295. (a) Nabil M. Lawandy *et al.,* " Laser action in strongly scattering media," *Nature,* Vol. 368, Mar. 31, 1994, p. 436-438. For a good lay article with color pictures, see (b) Ivars Peterson, "Boosted light: Laser action in white paint," *Science News,* 145(15), Apr. 9, 1994, p. 228-229. See also (c) P. Mandel, "Lasing without inversion: A useful concept?," *Contemp. Phys.,* Vol. 34, 1993, p. 335; (d) O Kocharovskaya, "Amplification and lasing without inversion," *Physics Reports,* Vol. 219, 1992, p. 175.

296. Some useful papers are: (a) Diederik S. Wiersma and Ad Lagendijk, "Light diffusion with gain and random lasers," *Phys. Rev. E,* 54(4), Oct. 1996, p. 4256-4265. The authors discuss experimental considerations for a medium that both multiply scatters and amplifies light, and distinguish three regimes; (b) H. Cao *et al.,* "Random Laser Action in Semiconductor Powder," *Phys. Rev. Lett.,* 82(11), 15 Mar. 1999, p. 2278 2281. The authors observed random laser emission in all directions, with coherent feedback, in semiconductor powder, (c) Diederik S. Wiersma, Meint P. van Albada and Ad Lagendijk, "Coherent Backscattering of Light from Amplifying Random Media," *Phys. Rev. Lett.,* 75(9), 28 Aug., **1995,** p. 1739-1742. (d) D. S. Wiersma, P. Bartolini, A. Lagendijk, and R. Righini, *Nature,* Vol. 390, 1997, p. 671. The authors report Anderson localization of photons in GaAs powder, (e) Ad Lagendijk *et al,* "Microscopic Approach to the Lorentz Cavity in Dielectrics," *Phys. Rev Lett.,* 79(4), 28 July 1997, p. 657-660. (f) Pedro de Vries and Ad Lagendijk, "Resonant Scattering and Spontaneous Emission in **Dielectrics** Microscopic Derivation of Local-Field Effects," *Phys. Rev. Lett.,* 81(7), 17 Aug. 1998, p. 1381-1384. (g) Gijs van Soest, Makoto Tomita, and Ad Lagendijk, "Amplifying volume in scattering media," *Opt. Lett.,* 24(5), Mar. 1, 1999, p. 306-308.

297. S. Itoh *et al.,* "Simulational and experimental studies on anomalous reflectivity of phase conjugate wave," *Ferroelectrics,* vol. 170, 1995, p. 209-217.

298. See (a) D. P. Hand and P. St. J. Russell, "Solitary thermal shock waves and optical damage in optical fibers," *IEE Colloquium on 'Non Linear Optical Waveguides', LEE Digest* No. 88, London, England, 1988, p. 101-103; (b) — "Soliton-like thermal shock-waves in optical fibres: origin of periodic damage tracks," *Fourteenth European Conference on*

Optical Communication (ECOC 1988), Sep. 11-15, 1988, Publication No. 292, 1(2), 1988, p. 111-114; (c) — "Solitary thermal shock waves and oplical damage in optical fibers: the fiber fuse," *Opt. Lett.,* 13(9), Sep. **1988,** p. 767-769; (d) D. P. Hand and T. A. Birks, "Single-mode tapers as 'fibre fuse' damage circuit-breakers," *Electronics Lett.,* 25(1), Jan. 5, 1989, p. 33-34. For an excellent lay description with photographs, see (e) Ivars Peterson, "Fibers with Flare," *Sci. News,* 140(13), Sep. 8, 1991, p. 200-201

299. P. St. J. Russell, "Power Conservation and Field Structures in Uniform Dielectric Gratings," *J. Opt. Soc. America A.,* 1(3), Mar. 3, 1984, p. 293-299. Russell is referring to apparent violation of local conservation of energy. It is true in the conventional flawed energy flow analysis considering only the Poynting component, because the fuse produces more work output (melting the holes) than the original operator's input of Poynting energy to initiate the fuse effect. It does not produce more work output than the total of the energy input by the operator in both the acounted Poynting component and the unaccounted Heaviside component. It merely shows that the extra fiber fuse ignition actions do absorb extra energy from the Heaviside component, perhaps by particle resonance (Bohren effect) of the melt and vapor particles. Also, intense heating produces intense ionization and thus intense dipoles, thus producing broken symmetry in the fierce exchange with the active vacuum. Multiple retroreflections of the EM energy flow also occur in the nmelt and vapor particles, leading to multiple collections in the ping-pong of a joule of energy between two particles as phase conjugate mirrors.

300. E.g., see (a) H. Kogelnik, "Coupled wave theory for thick hologram gratings," *Bell Systems TechnicalJournal,* Vol. 48, 1969, p. 2909-2947; (b) Z. Zylbergerg and E. Marom, "Rigorous coupled-wave analysis of pure reflection gratings," *J. Opt. Soc. Am.,* Vol. 71, 1981, p. 811-818; (c) — - "Rigorous coupled-wave analysis of grating diffraction — E- mode polarization and losses," *ibid.,* Vol. 73, 1983, p. 451-455; (d) - "Comments on analyses of reflection gratings," *ibid.,* Vol. 73, 1983, p. 399-401.

301. Bohren, {24J.

302. Justin Mullins, "Superpowerful Explosive Arrives with a Bang." New *Scientist,* Aug. 4, 2001.

303. Romano M. DeSantis *et al.*, "On the Analysis of Feedback Systems With a Multipower Open Loop Chain," Oct. 1973, AD 773188, available through the Defense Technical Information Center or the National Technical Information Center. The paper clearly shows that such systems can produce overunity coefficient of performance.

304. See (a) J. H. Andreatta, "High Power Switching Amplifier Wherein Energy is Transferred to a Tuned Circuit During Both Half Cycles," U.S. Patent No. 3,239,771, Mar. 8, 1966; (b) Tom L. Dennis Jr., "Highly Efficient Semiconductor Switching Amplifier," U.S. Patent No. 3,239,772, Mar. 8, 1966; (c) Heber J. Morrison, "Square Wave Driven Power Amplifier," U.S. Patent No. 3,815,030, June 4, 1974.

305. For a good but conventional coverage of many differential negative resistors and negative conductors, see Adolf Fritz Schwarz, "Negative Resistors and Negative Conductors," Doctoral Thesis, Technische Hogeschool, Delft (Netherlands), Feb. 1969, available from the U.S. National Technical Information Service (NTIS), Springfield, VA. True negative resistors are dismissed as impossible in Schwarz's opening remarks. And they *are* impossible in equilibrium systems not receiving and using any net excess energy asymmetrically from their environment, which are the only systems conventionally considered by electrical engineers and electrodynamicists.

306. Gabriel Kron, "Numerical solution of ordinary and partial differential equations by means of equivalent circuits." *J. Appl. Phys.*, Vol. 16, Mar. 1945a, p. 173.

307. Kron, {32}, p. 39.

308. Bearden, {12}.

309. E.g., see Gabriel Kron, "Equivalent Circuits for Oscillating Systems and the Riemann-Christoffel Curvature Tensor," *General Electric Review*, Jan. 1943, p. 25-31.

310. See (a) S. Austen Stigant, "Gabriel Kron on Tensor Analysis, A bibliographical record," *BEAMA Journal*, Aug. 1948, for a full bibliography of Kron's publications up to 1948. Also useful are (b) Gabriel Kron, "A Method to Solve Very Large Physical Systems in Easy Methods," *IRE Transactions on Circuit Theory*, 2(1), Dec. 1953, p. 71-90; (c) P. L. Alger, editor, *The Life and Times of Gabriel Kron, or Walking Around the World, and Tensors*, Mohawk Development Services, Inc.,

Schenectady, NY, 1969; (d) Banesh Hoffman, " Cron's Non-Riemannian Electrodynamics," *Rev. Mod. Phys.*, 21(3), 1949, *x* 535-540; (e) K. Kondo and Y. Ishizuka, "Recapitulation of the Geometrical Aspects of Gabriel Kron's Non-Riemannian Electrodynamics," Research Association of Applied Geometry, *Memoirs of the Unifying Study of Basic Problems in Engineering and Physical Sciences by Means of Geometry*, Gakujutsu Bunken Fukyu-kai, Tokyo, Vol. I, 1955, p. 185-239 [particularly see footnote 1, p. 222];

311. Wang and Chung, {33}.

312. Deborah D. L. Chung, "Superconductor-metal laminates and method of making," U.S. patent no. 5,059,582, issued Oct. 22, 1991. A superconducting laminate having at least one layer of metal and at least one layer of superconducting material. The metal layer and the superconducting layer are bonded. The layer may also include carbon fibers from various precursors. The superconductor may be a composite material. The invention also includes a method of making the laminates. Assigned to The Research Foundation of State University of NY, Albany, NY.

313. Jean-Louis Naudin, website http://jnaudin.free.fr/cnr/.

314. Fogal, {36a, 36b}.

315. As one example, Fogal's device has to be able to manipulate Dirac holes in the vacuum itself, at least to some extent. Once the erroneous flat spacetime assumption of classical electrodynamics is removed and curved local spacetime considered, we can and will have Dirac holes in the local vacuum itself, especially in COP»1.0 systems. In a COP»1.0 system, not all the Dirac holes have interacted with orbital electrons in atoms to "eat" an electron and produce a positive ion — with a "lattice hole" or "lattice positron" where the electron had been. There is a vast difference between a Dirac hole positron (which has a negative mass equal in magnitude to one electron-mass) and a lattice hole (which is attached to the relatively huge positive mass of the ionized atom. Hence the mobility and phenomenology of Dirac positrons and lattice positrons are very different. There presently are no books on the dynamics of the Dirac negative energy electron holes in COP»1.0 EM circuits and systems.

316. Mandl and Shaw, {19}.

317. Whittaker, {91a}. The paper was published in 1904 and orally delivered in 1903. Note: On p. 368 midpage, Whittaker's equation for the y-component of vector-potential "a" is missing the exponent "2" on the **del** operator.

318. Rodrigues and Lu, {93a}.

319. Rodrigues and Vaz, {93b}.

320. It is certainly not lacking in the secret military weapons game. Some 10 nations of the world have weaponized longitudinal EM wave weapons over the past few decades, and almost every major weapons laboratory in the world has now "discovered" longitudinal EM waves. Even the Yakuza (Japanese Mafia) has acquired its own portable longitudinal EM interferometry weapons.

321. Evans et al., {79}.

322. (a) V. I. Arkhipov, *et al.,* "Negative transient currents in amorphous semiconductors." *Internat. J. Electronics* (UK), 51(6), 1981, p, 735-742; (b) — and A. I. Rudenko, "Negative currents caused by injection-controlled polarization," *Solid. St. Commun.,* 28, 1978, p. 675-676.

323. Jackson, {22}, p. 222.

324. Jackson, {22}, p. 223.

325. O. L. Brill and B. Goodman, *Am. J. Phys.,* vol. 35, 1967, p. 832.

326. For example, see Jackson, {22}, p. 219-221.

327. We stress that electrodynamicists, conventional theory, and quantum field theory all assume the principle *of gauge freedom,* the **total** freedom to change the gauge (and therefore any potential) in a Maxwellian system at any time, without cost or work. In the real world one may have to pay a little switching costs, but by simply regauging one can freely change the EM energy physically present at a point in 3-space **occupied by** the Fogal "freed" instantaneous scalar potential, once the electrons are pinned. In short, the Fogal process is able to directly "gate" the instantaneous transduction of EM energy existing at multiple points **in** space, simultaneously. There is nothing in the gauge freedom **principle** that restricts (i) regauging the instantaneous scalar potential at will, **or** (ii) using these changes at a distance to produce force fields upon electrons, thereby moving them — so long as one employs a Fogal semiconductor in

the "distant receiver" in order to be able to produce the back-transition from the field-free state to the force-field state by removing the anomalous pinning restriction. To the advanced theorists, we leave the debates over causality considerations for this demonstrated process!

328. Whittaker, {85}.

329. (a) W. Heitman and G. Nimtz, "On Causality Proofs of Superluminal Barrier Transversal of Frequency Band Limited Wave Packets," *Phys. Lett. A196,* 1994, p. 154-158; (b) Enders and Nimtz, {136b}.

330. The reader should recall that the Lorenz/Lorentz condition, which limits the circuit to COP< 1.0, is self-applied by the circuit when it rams all "spent" electrons in the ground return line back through the emf of the circuit. One need not do that, and need not produce that force field that destroys the primary source dipole and its extraction of EM energy from the vacuum via its broken symmetry. As the reader has undoubtedly now observed, the "back-emf' and "back mmf are those aspects in an electrical circuit and magnetic circuit respectively, that kill the dipole and physically self-apply and self-enforce Lorenz/Lorentz symmetrical regauging. Hence the intuitive concern of COP>1.0 experimenters to try to discover how to avoid or lessen the "back emf or the "back mmf in their circuits and systems.

331. Philip Ball, "The positron probe," *Nature,* Vol. 412,23 Aug. 2001, p. 764.

332. Werner Triftshauser et al., *Phys. Rev. Lett.,* 87, 067402 (2001).

333. Bearden, {30}.

334. For the chemistry and currents of the battery, see any good battery reference, such as Colin A. Vincent and Bruno Scrosati, *Modern Batteries: An Introduction to Electrochemical Power Sources,* Second Edition. Wiley, New York, 1997.

335. (a) J. O'M. Bockris, A. K. N. Reddy, and M. Gamboa-Aldeco, *Modern Electrochemistry: Fundamentals of Electrodics,* 2nd edition. Kluwer/Plenum, NY 2000; (b) J. O'M. Bockris and A. K. N. Reddy, *Modern Electrochemistry: Electrodics in Chemistry, Engineering, Biology, and Environmental Science,* 2nd edition, Kluwer/Plenum, NY 2000; *(c)* J. Koryta, J. Dvorak, and L. Kavan, *Principles of Electrochemistry,* 2nd ed.. Wiley, NY, 1993; (d) Wolfgang .1. Lorenz and Waldfried Plieth,

Electrochemical Nanotechnology: In Situ Local Probe Techniques at Electrochemical Interfaces, Wiley, NY, 1998; (e) B. E. Conway, *Electrochemical Supercapacitors: Scientific Fundamentals and Technological Applications,* Plenum, NY, 1999; (f) Allen J. Bard and Larry R. Faulkner, *Electrochemical Methods: Fundamentals and Applications,* Second Edition, Wiley, NY, 2001.

336. At this writing, a provisional patent application on the process has been filed by Bedini and the present author. By the time of publication, hopefully this will have been turned into a formal patent application, with foreign patents also applied for.

337. Edwin V. Gray, "Pulsed Capacitor Discharge Electric Engine," U.S. Patent #3,890,548, June 17, 1975.

338. (a) Per Whittaker, {85}. However, see (b) Bearden, {12} for a necessary correction of Whittaker's interpretation to accord with a quantum field theory view of the instantaneous scalar potential. After observation, the phase conjugate EM wave half of the Whittaker wavepair is indeed going in the opposite direction to the normal longitudinal EM wave half But prior to observation, it exists in the time domain, and in the time domain it is precisely accompanying the 3-space longitudinal EM wave, with the absorbing and then emitting charge interacting in the middle between the two. See (c) Mandl and Shaw, {19}, to see that the combined scalar (time-polarized) photon and longitudinal photon are observable as the instantaneous scalar potential. This means that the corrected Whittaker decomposition produces an associated time-polarized EM wave in the time domain and longitudinal EM wave in 3-space. That is the corrected interpretation of Whittaker's phase conjugate longitudinal EM wavepair.

339. (a) J. O'M. Bockris, "Overpotential: a lacuna in scientific knowledge,"J. *Chem. Edu.,* 48(6), June 1971, p. 352-358. This is a simplified introduction to overpotential theory. The energy density that exists at minute "double layers" in contact, which act as tiny capacitors with extremely small separation of plates, is enormous and far greater **than** what the power engineer meets in his normal engineering practice. **Sec** also (b) Bockris and Reddy, {335b}, particularly Chapter I which deals with the overpotential. See also (c) Bockris et al, {335a}.

340. E.g. see (a) L. Bertalot *et al.,* "Study of deuterium charging in palladium by the electrolysis of heavy water: heat excess production", *Nuovo Cimento,* Vol. 15 D, 1993, p. 1435; (b) J. O'M Bockris *et al.,* "Cold

fusion as a consequence of high fugacity among hydrogen isotopes," Intl. *J. Hydrogen Energy,* Vol. 17, 1992, p. 445.

341. Peter Lindemann, *The Free Energy Secrets of Cold Electricity,* **Clear** Tech Inc., Metaline Falls, WA, 2001.

342. Particularly see Kenneth S. Deffeyes, *Hubbert's Peak: The Impending World Oil Shortage,* Princeton University Press, Princeton, NJ, **2001.**

Quoting: *"Deffeyes finds that world oil production will peak within five wars and that there isn't anything we can do to stop it. New exploration and production technologies can't save us, and plans to open the Arctic National Wildlife Refuge and other areas to drilling offer no more than a small and soon-to-be-forgotten blip on the production curve. While long-term solutions exist in the form of conservation and alternative energy sources, they probably cannot — and almost certainly will not — be enacted in time to evade a short-term catastrophe of shortages, soaring prices, and global economic, agricultural, and possibly political disturbance. "* In short, Deffeyes rigorously demonstrates that a world energy crisis of staggering proportions is just around the corner.

343. D. J. Bohm and B. J. Hiley, "On the intuitive understanding of nonlocality as implied by quantum theory," *Found. Phys.,* 5(1), Mar. 1975, p. 93-109. Quote is from p. 107.

344. (a) Y. Aharonov and D. Bohm, "Significance of Electromagnetic **Potentials** in the Quantum Theory," *Phys. Rev.,* Second Series, 115(3), Aug. 1, 1959, p. 485; (b) — "Further Considerations on Electromagnetics In **the** Quantum Theory," Phys. Rev., 123(4), Aug. 15, 1961, p. 1511-1524.

345. Terence W. Barrett, "Electromagnetic Phenomena Not Explained by Maxwell's Equations," A. Lakhtakia, ed., *Essays on the Formal Aspects of Electromagnetics Theory,* World Scientific Publishing, River Edge, NJ, **1993,** p. 11 and generally p. 6-86.]

346. Davies, {237}, p. 105.

347. David J. Bohm, "A Suggested Interpretation of the Quantum **Theory** in Terms of 'Hidden' Variables, I and II." *Phys. Rev.,* 85(2), Jan. 15, 1952, p. 166-179 (Part I); 180-193 (Part II).

348. Jackson, {22}, p. 223.

349, Lamb and **Retherford,** {23}.

350. (a) Casimir, {26}. This paper is an extrapolation to two attracting plates of the more detailed paper of (b) H. B. G. Casimir and D. Polder, "The influence of retardation on the London-van der Waals Forces," *Phys Rev.* 73(4), Feb. 15, 1948, p. 360-372. Casimir's paper was an extension of earlier work by Verwey and Overbeek on the stability of colloidal systems, including the interaction of electric double layers; see (c) E. J. W Verwey, J. T. G. Overbeek, and K. van Nes, *Theory of the Stability of Lyophobic Colloids*, Elsevier, Amsterdam, 1948.

351. Lamoreaux, {27}.

352. Mead and Nachamkin, {28}.

353. E.g., see (a) P. W. Milonni. R. J. Cook and M. E. Goggin, "Radiation pressure from the vacuum: Physical interpretation of the Casimir force," *Phys. Rev. A,* Vol. 38, 1988, p. 1621; (b) C. E. Carlson, T. Goldman and J. Peres-Mercader, "Gamma-ray bursts, neutron star quakes, and the Casimir effect," *Europhys. Lett.*, Vol. 36, 1996, p. 637; (c) M. Chaichian *et ah,* "Quantum theories on noncommutative spaces with nontrivial topology: Aharonov-Bohm and Casimir effects," *Nucl Phys B.* Vol. 611, 2001, p. 383-402; (d) V. Dodonov, "Nonstationary Casimir Effect and Analytical Solutions for Quantum Fields in Cavities with Moving Boundaries," in *Modern Nonlinear Optics,* 2nd Edn., 3 vols., Wiley, 2001, Vol. 1, p. 309-384.

354. Bohm, {347}.

355. See T. E. Bearden, *Gravitobiology: A New Biophysics,* Tesla Book Co., Chula Vista, CA, 1991.

356. One should not be nonplussed at the notion of infinite velocity. A pure longitudinal EM wave is — at least in theory — capable of moving at infinite velocity, and is a proper solution to Maxwell's equations. See (a) Rodrigues and Lu, {93a}. Also, the ordinary potential in the Coulomh gauge does indeed move at infinite velocity; see Jackson, {22}, p. 223. In addition, Bohm's interpretation of quantum mechanics {347} uses the quantum potential, which instantaneously appears everywhere it will have magnitude.

357. Enders and Nimtz, {329b}.

358. R. Chen, "Cancellation of Internal Forces," *Am. J. Phys.,* 49(4), Apr. 1981, p. 372.

359. Whittaker, {85}.

360. R. L. Davis, "Non-Inductive Electrical Resistor," U.S. Patent No. 3,267,406, Aug. 16, 1966. Davis uses a Mobius-loop capacitor as an inductance-free resistor. Our comment is that, more rigorously, the resistor is *not* noninductive. It is "equally and oppositely inductive," which is quite a different thing altogether when one considers the difference between internal energy and external energy. We hypothesize that the "resistor" — or some version of it — produces an artificial scalar potential stress in spacetime — apparently varying the stress energy tensor itself— even though it produces a *net* zero reactance.

361. These are Brazil, Russia, China, and two other nations friendly to the U.S. The dominant weapons on earth today are quantum potential weapons, not nuclear weapons. The second most dominant weapons are probably the negative energy EMP weapons deployed by several nations, including China.

362. E.g., see (a) L. I. Mandelstam and N. D. Papaleksi, "On the parametric excitation of electric oscillations," *Zhurnal Teknicheskoy Fiziki,* 4(1), 1934, p. 5-29; (b) — "On resonance phenomena with frequency distribution." *Z.f. Phys.,* No. 72, 1931, p. 223: (c) — "Concerning asynchronous excitation of oscillations," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934; (d) — "Concerning asynchronous excitation of oscillations," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934; (e) —- "Concerning nonstationary processes occurring in the case of resonance phenomena of the second class," *Zhurnal Tekhnicheskoi Fiziki,* 4(1), 1934. See also (f) A. Andronov, "The limiting cycles of Poincare and the theory of self-maintained oscillations," *Comptes-Rendus,* Vol. 189, 1929, p. 559; (g) — and A. Witt, "On the mathematical theory of self-excitations," *Comptes-Rendus,* Vol. 190, 1930, p. 256; (h) — "On the mathematical theory of self-excitation systems with two degrees of freedom," *Zhurnal Tekhnicheskioi Fiziki,* 4(1), 193; (i) — "Discontinuous periodic movements and theory of multivibrators of Abraham and Bloch," *Bull. De l'Acad. Ed Sc. De l'URSS,* vol. 189, 1930. See (j) S. Chaikin, "Continuous and 'discontinuous' oscillations," *Zhurnal Prikladnoi Fiziki,* Vol. 7, 1930. p. 6; (k) — and A. Witt, "Drift in a case of small amplitudes," *Zhurnal Teknicheskoi Fiziki,* 1(5), 1931, p. 428; (1) — and N. Kaidanowski, "Mechanical relaxation oscillations," *Zhurnal Teknicheskoi Fiziki,* Vol. 3, 1933, p. 1.

363. L. Mandelstam [L. T. Mendel'shtam], N. Papalexi, A. Andronov, S. Chaikin and A. Witt, "Report on Recent Research on Nonlinear Oscillations," Translation of "Expose Des Recherches Recentes Sur Les Oscillations Non Lineaires," *Technical Physics of the USSR,* Leningrad, Vol. 2, 1935, p. 81-134 [NASA Translation Doc. TTF-12,678, Nov. 1969].

364. In Alexander L. Fradkov and Alexander Yu. Pogromsky, *Introduction to Control of Oscillations and Chaos,* World Scientific Series on Nonlinear Science E, Series Editor Leon O. Chua, World Scientific, New Jersey, 1998 — as a typical example — I found no mention of the Dirac hole current decay mechanism or of the active external environment (the active vacuum and its dynamics, and the active local curvatures of spacetime and their dynamics).

365. Richard E. Prange and Peter Strance, "The Semiconducting Vacuum," *Am. J. Phys.,* 52(1), Jan. 1984, p. 19-21.

366. John F. Lindner *et ah,* "Noise Enhanced Propagation," *Phys. Rev. Lett.,* 81(23), 7 Dec. 1998, p. 5048-5051. Lindner suggests that the mechanism appears to be general, in which case it could supply the mechanism that coherently "locks together" the barium nuclei self-oscillations in Sweet's activated barium ferrite magnets.

367. Owen Flynn, "Parametric arrays: A new concept for sonar," *Electronic Warfare Magazine,* June 1977, p. 107-112. Any two sine-wave frequencies as simultaneous drivers combine to produce a sine-wave difference frequency propagating in an isotropic nonlinear medium, essentially without sidebands or reverberations. Its pattern has a main lobe approximately equal to that of the high frequency drive, but devoid of sidelobes. The level of the propagating difference frequency is proportional to both the product of the two fundamental drive levels and to the square of the desired value of difference frequency.

368. M. I. Dykman *et al.,* "Noise-enhanced heterodyning in bistable systems." *Phys. Rev. E,* 49(3), Mar. 1994, p. 1935-1942

369. (a) V. S. L'vov and L. A. Prozorova, "Spin Waves Above the Threshold of Parametric Excitation," in (b) A. S. Borovik-Romanov and S. K. Sinha, Eds., *Spin Waves and Magnetic Excitations,* North-Holland, Amsterdam, 1988.

370. Gurevich and Melkov, {232c}. The authors treat high **frequency** processes, i.e., oscillations and waves, in magnetically ordered substances, including ferromagnetic and antiferromagnetic resonances, spin waves, nonlinear processes, and high frequency manifestations of **interactions** between the magnetic system and other systems of magnetically **ordered** substances (crystal lattice, charge carriers). The treatment is classical except that a quantum mechanical approach is used in the study of relaxation processes, which determine the energy losses of magnetic oscillations and waves. Experimental results but not techniques are included. Self-oscillation and self-oscillation criteria arc on p. 279.

371. (a) G. S. Kandaurova and A. E. Sviderskii, "Observation of autowave state and stable dynamic structures in multidomain magnetic films," *Sov. Tech. Phys. Lett.,* 14(5), May 1988, p. 346-347; (b) V. S. **L'vov,** {29c}.

372. Richard T. Hammond, "New fields in general relativity," *Contemp. Phys.,* 36(2), 1995, p. 103-114.

373. (a) J. Baugh et al., *Science,* Vol. 294, 16 Nov. 2001, p. 1505. For a simpler synopsis of the paper, see (b) Warren S. Warren, "Multidimensional Symmetry in a Three-Dimensional World," *Science,* Vol. 294, 16 Nov. 2001, p. 1475-1476.

374. E.g., see Howard R. Johnson, "Permanent Magnet Motor." U.S. Patent No. 4.151.431. Apr. 24. 1979.

375. Howard Johnson, private communication, January 2002.

376. In magnetics, multi-valued potentials (and hence the possibility **for** nonconservative fields) arise naturally in the theory, but are considered a nuisance to be eliminated or "averaged away" by magnetics engineers. E.g., see (a) Z. Ren, "A 3D vector potential formulation using edge element for electrostatic field computation," *IEEE Trans, on Magnetics,* **31(3),** May 1995, p. 1520-1523; (b) — "New technique for solving three-dimensional multiply connected eddy-current problems." *lEEProc, 1*37A(3), May 1990, p. 135-140; (c) J. C. Verite, "Calculation of multivalued potentials in exterior regions," *IEEE Trans, on Magnetics,* **Vol.** MAG-23(3), May 1987, p. 1881-1887. Amazingly, these and other similar methods are directed into converting the multivalued potential into single-valued regions by a process of multiple, highly selected cuts. In other words, the real attempt is to eliminate the multivalued aspect, in which case the theorists are in fact simply eliminating the asymmetrical

self-regauging provided by the multivalued potential and hence are discarding the ability to use multivalued potentials in true overunity processes.

377. (a) Z. Badics, "Transient eddy current field of current forced three-dimensional conductors," *IEEE Trans, on Magnetics,* 28(2), Mar. 1992, p. 1232-1234;(b) R. Albanese and G. Rubinacci, "Magnetostatic field computations in terms of two component vector potentials," *Int. J. for Numerical Methods in Engineering,* Vol. 29, 1990, p. 515-532; (c) Bearden, {214f}: (d) P. Hammond and T. D. Ssiboukis, "Dual finite-element calculations for static electric and magnetic fields," *IEE Proc. A,* 130(3), 1983, p. 105-111; (e) P. R. Kotiuga, "On making cuts for magnetic scalar potentials in multiply connected regions," *J. Appl. Phys.,* 61(8), 1987, p. 3916-3918; (f) J. Penman and J. R. Fraser, "Complementary and dual energy finite element principles in magnetostatics," *IEEE Trans, on Magnetics,* 18(2), 1982, p. 319-323; (g) E. A. Tolkachev and Ya. M. Shnir, "Non-trivial topology effects in the interaction of a monopole with a line current," *Phys. Lett. A,* 155(8, 9), May 27, 1991, p. 464-466.

378. (a) Feynman, {7}, Vol. II, Chapter 37; (b) T. Ambrose, R. L. Sommer, and C. *L.* Chien, "Angular Dependence of Exchange Coupling in Ferromagnetic/Antimagnetic Bilayers," *Phys. Rev. B,* Vol. 56, 1997, p. 83; (c) H. Sang, Y. W. Du, and C. L. Chien, "Exchange Coupling in $Fe_{50}Mn_{50}/Ni_{81}Fe_{19}$ Bilayers: Dependence on Antiferromagnetic Layer Thickness," *J. Appl. Phys.,* Vol. 85, 1999, p. 4931.

379. V. A. Tulin, "Nuclear Spin Waves in Magnets," *Fis. Niskikh Temp.,* Vol. 5, 1979, p. 965-993 [in Russian].

380. Sweet and Bearden, {29a}.

381. (a) Kenneth R. Shoulders, U.S. Patent # 5,153,901; (b) — "Energy Conversion Using High Charge Density," U.S. Patent # 5,018,180, May 21, 1991; (c) — "Energy Conversion using High Charge Density," U.S. Patent No. 5,123,039, 1992; (d) — U.S. Patents 5,054,046 and 5,054,047, 1991; (e) — "Circuits Responsive To and Controlling Charged Particles," U.S. Patent 5,148,461, 1992; (f) — and Steve Shoulders, "Observations on the Role of Charge Clusters in Nuclear Cluster Reactions," *J. New Energy,* 1(3), Fall 1996, p. 111-121.

382. (a) Paulo N. Correa and Alexandra N. Correa, "Pulsed Abnormal Glow Discharge Technology," *Infinite Energy,* 2(7), Mar./Apr. 1996, p.

18-21; (b) — "Other Applications of the PAGD Technology Besides Energy Conversion," *ibid.*, p. 22-27; (c)— "Electrochemical Transduction of Plasma Pulses," U.S. Patent No. 5,416,391, May 16, 1995; (d) "Energy Conversion Systems," U.S. Patent No. 5,449,989, Sept. 12, 1995; (e) — "Direct Current Energized Pulse Generator Utilizing Autogenous Cyclical Pulsed Abnormal Glow Discharges," U.S. Patent 5,502,354, Mar. 26, 1996.

383. In 1935, Prileshajewa showed that there is an energy difference as much as 1.1 v between the exciting light and the fluorescence of aniline vapor. This added energy is attributed to additions from the internal energy of the molecule. Thus, anti-Stokes emission has been validated as a COP>1.0 mechanism since at least 1935. For our purposes, replenishment of the internal energy of the molecule must be obtained from the local active vacuum; else one simply has the standard optical cooling process which cools the gas.

384. E.g., see J. K. Brasseur et al., "Coherent Anti-Stokes Emission in a Continuous Wave Raman Laser in H_2," carried on website location http://www.physics.montana.edu/optics/jlc/Anti-Stokes.pdf.

385. For a typical study, see J. Silver et al., "Studies on anti-Stokes emission from metal oxides (doped with Er3+ and Yb3+) stimulated by 632.8 nm laser excitation," *6th Int. Conf. Phosphors and Display Tech.*, San Diego, California, 2000.

386. For a typical paper of hundreds on the subject, see C. Zander and K. H. Drexhage, "Cooling of a Dye Solution by Anti-Stokes Fluorescence," in *Advances in Photochemistry*, edited by D. C. Neckers, D. H. Volman, and G. von Bulau; Wiley, New York, 1995.

387. (a) Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (b) — "Apparatus for producing radiations penetrating living cells," U.S. Patent No. 3,368,155, Feb. 6, 1968. For an explanation, see (c) T. E. Bearden, "The Porthole Briefing," http://www.cheniere.org.

388. Thomas Valone, "Future Energy Technologies," *J. New Energy*, 5(2), Fall 2000, p. 41-46; summary is quoted from p. 44-45.

389. Shang-Xiang Jin and Hal Fox, "Characteristics of High-Density Charge Clusters: A Theoretical Model," *J. New Energy*, 1(4), Winter 1996, p. 5-20.

390. Shoulders, {381a-381e}.

391. Ken Shoulders, "Permittivity transitions," *J. New Energy,* 5(2), Fall 2000, p. 121-137.

392. *Ibid,* p. 122.

393. *Ibid,* p. 122.

394. K. Shoulders and S. Shoulders, {381 f).

395. Harold L. Fox, "Energy for the Future: High-Density Charge Clusters," *Modern Nonlinear Optics,* Second Edition, edited by M. W. Evans, Wiley, New York, 2001, Part III, p. 623-653.

396. See (a) Art Rosenblum, "Randell Mills — New Energy and the Cosmic Hydrino Sea," *Infinite Energy,* 3(17), Dec. 1997-1998, p. 21-34. Also see (b) Randell Lee Mills, "Energy/Matter Conversion Methods and Structures," Australian Patent No. 668678, Nov. 20, 1991. Particularly see (c) Eugene Mallove, "Dr. Randell Mills and the Power of BlackLight," *Infinite Energy,* 2(12), Jan.-Feb. 1997, p. 21, 35, 41.

397. Randell Mills *et al.,* "Lower-Energy Hydrogen Methods and Structure," U.S. Patent No. 6,024,935, Feb. 15, 2000. Some 60 pages, and an astounding 499 claims. For the patent, an independent verification was performed by Johannes Conrads, Institute for Low Temperature Plasma Physics at the Ernst Moritz Arndt University in Greifswald, Germany.

398. Randell L. Mills, "The Grand Unified Theory of Classical Quantum Mechanics," www.blacklightpower.com/book/shtml.

399. (a) Randell L. Mills et al., "Identification of Compounds Containing Novel Hydride Ions by Nuclear magnetic Resonance Spectroscopy"; (b) — "Spectroscopic Identification of Novel Catalytic-Reaction of Atomic Hydrogen and the Hydride Ion Products"; (c) — "Black Light Power Technology: A New Clean Energy Source with the Potential for Direct Conversion to Electricity," presented to the Intl. Conf. On Global Warming and Energy Policy, Nov. 26-28, 2000 and to the National Hydrogen Association Meeting, March 6-8, 2001; (d) — Plasma Power Source Based on a Catalytic Reaction of Atomic Hydrogen"; (e) R, M. Mayo and R. L. Mills, "Direct Plasmadynamic Conversion of **Plasma** Thermal Power to Electricity for Microdistributed Power Applications," *40th Ann. Power Sources Conf. Proc*, June 10-13, 2002; (f) R. Mills et al., "Novel Catalytic Reaction as a Potential New Energy Source," 223rd

ACS National Meeting, Orlando, Florida, Apr. 6-11, 2002; (g) — "Novel alkali and Alkaline Earth Hydrides for High Voltage and High Energy Density Batteries," *Proc. 17th Ann. Battery Conf. On Appl. And Advances, Calif. State Univ., Long Beach, CA, Jan. 15-18* (in press); (h) - "Synthesis of Diamond Films from Solid Carbon"; (i) Mills and Ray, "Spectral Emission of Fractional Quantum Energy Levels of Atomic Hydrogen from a Helium-Hydrogen Plasma and the Implications for Dark Matter," *Int. J. Hydrogen Energy,* 27(3), 2002, p. 301-322.

400. Peter D. Zimmerman, "An Analysis of Theoretical Flaws in So-Called Classical Quantum Mechanics and of Experimental Evidence against CQM," 2001.

401. Bearden, {222b}, p. 674.

402. E.g., it yields the excess deuterium, tritium, and alpha particles; see Bearden, {17}.

403. (a) Bearden, {222b}; Bearden, {15}.

404. (a) "Ion Cyclotron and Converter and Radio Power Microwave Generator," International Pub. No. WO 01/21300 A2, 29 March 2001; (b) "Ion Cyclotron Power Converter and Radio and Microwave Generator," International Pub. No. WO 01/22472, 29 March 2001. Both filed with World Intellectual Property Organization.

405. But see Samuel D. Harper. "The energy dissipated in a switch," *Am. J. Phys.*, 56(10), Oct. 1988, p. 866-889. Unless special attention and effort is paid, switches can indeed consume a great deal of energy! As Harper points out, it can be a substantial fraction of the energy transfer in the circuit. In a capacitive loaded circuit this is especially true. One should consider very efficient switching such as optically coupled switching arrangements.

406. (a) F. Heinrich, "Entropy change when charging a capacitor: A demonstration experiment," *Am. J. Phys.*, 54(8), 1986, p. 742-44. For an expanded discussion, see (b) I. Fundaun, C. Reese, and H. H. Soonpaa, "Charging a capacitor," *Am. J. Phys.*, 60(11), Nov. 1992, p. 1047-1048. Also see (c) Tyler A. Abbott and David J. Griffiths, "Acceleration without radiation." *Am. J. Phys.*, 53(12), Dec. 1985, p. 1203-1211. An extended charge can accelerate without radiating, under specified conditions.

407. (a) Davis, {360}. It would appear that a sine wave across this resistor may generate a **rhythmic** variation in what is called "internal

energy"; i.e., stress energy inside the system, hence stress energy in the associated vacuum. Since this is analogous to a sound wave consisting of rarefaction and expansion, it is a method of generating a longitudinal EM wave. This may also be involved also in the Fogal semiconductor. For a discussion of internal forces, which occur in equal and opposite pairs and are quite different from the resultant of forces, see (b) Chen, {358}, p. 372 and passim.

408. See Hsu-Chieh Yeh, "Remark on the second law of thermodynamics," *Am. J. Phys.,* 52(8), Aug. 1984, p. 720-723. As an example, from the common charge or dipole at constant temperature, we can indeed receive and take electrical energy because the charge (as a set of composite dipoles) and the dipole are not in 3-space energy flow equilibrium. Instead, they constantly receive virtual EM energy from the active vacuum, transduce it into 3-space EM energy, and continuously emit **that** energy in all directions — all at constant temperature. When we consider the supersystem of the charge-system and its environment, it is in disequilibrium and hence exhibits the effect pointed out by Yeh.

409. Yeh, {408}. Particularly see the illustration on p. 721.

410. Gabriel Laufer, "Work and heat in the light of (thermal and **laser**) light," Am. *J. Phys.,* Vol. 51, 1983, p. 42.

411. (a) Paulo N. Correa and Alexandra N. Correa, "A Modified Orgone Accumulator (HYBORAC) as a Drive for a Low Delta-T **Stirling** Engine, Part 1," *Infinite Energy,* 7(41), 2002, p. 23-29; (b) — "Orgone Accumulator Drive (HYBORAC) for Low Delta Stirling Motors, video presentation, ABRI Media Production.

412. Paulo N. Correa and Alexandra N. Correa, *Introduction to Experimental Aetherometry,* Vols. 1 and 2, 2001. Summaries of these **two** volumes may be freely downloaded from http://www.aetherometry.com

413. Paulo N. Correa and Alexandra N. Correa, "A Modified Orgone Accumulator (HYBORAC) as a Drive for a Low Delta-T Stirling **Engine** Part II." *Infinite Energy,* 7(42), 2002 (in press)

414. Eugene F. Mallove, "Demonstrating Aether Energy," *Infinite Energy,* 7(41), 2002, p. 6-8.

415. As quoted by Mallove, *ibid.,* p. 7.

416. Feynman, {7}, Vol. II, p. 1-3. Emphasis in original.

417. Mallove, {414}, p. 7.

418. J. M. Aguirregabiria, A. Hernandez, and M. Rivas, "Surface charges and energy flow in a ring rotating in a magnetic field." *Am. .J. Phys.,* 64(7), July 1996, p. 892-895.

419. F. Herrmann, "The unexpected path of the energy in a moving capacitor," *Am. J. Phys.,* 61(2), Feb. 1993, p. 119-121.

420. Wang and Chung, {33}.

421. Mike LaPointe, "Antimatter Production at a Potential Boundary," presented at the AIAA Joint Propulsion Conference, Breakthrough Propulsion Physics Session #1, July 9, 2001.

422. E.g., (a) Charles Kittel, *Introduction to Solid State Physics,* Seventh Edition, Wiley, New York, 1996 in various chapters contains excellent introductions to a more in-depth view of magnetic phenomena. A very useful book, written with clarity is (b) B. D. Cullity, *Introduction to Magnetic Materials,* Addison-Wesley, Reading, MA, 1972. For nuclear effects, (c) A. Abraham, *Nuclear Magnetism,* Oxford. 1961 is still excellent and quite comprehensive.

423. (a) Particularly see the lay overview by Philip Ball, "Meet the spin doctors...", *Nature,* Vol. 404, 2000, p. 918-920. This article deals with spin-injection from magnetic semiconductors. In addition to up and down, electron spin also has an additional degree of freedom called "color"; see (b) "Colourful electrons solve puzzle of ferromagnetism," *Physics World,* 14(7), July 2001; (c) also see the accompanying article by Fu Chun Zhang.

424. Stanislaw Bednarek, "Self-powered magnetohydrodynamic motors," *Am. J. Phys.,* 64(1), Jan. 1996, p. 90-92.

425. In a true superconducting system, it is possible to keep the current moving around in a resistance-free conductor indefinitely. However, no net work can be produced by the system, and in fact when one counts the energy one inputs to maintain the extremely low temperature of the environment, then to maintain the supersystem in such **"indefinite current** circulation" one has to continuously input energy to cool the **system. On** the other hand, an electron in an atom, orbiting the nucleus, is in **net** equilibrium on its path, so it will maintain its orbit indefinitely until perturbed sufficiently to break loose.

426. (a) Badics, {377a};(b) P. Hammond and J. Penman, "Calculation of inductance and capacitance by means of dual energy principles," *IEE Proc,* 123(6), 1976, p. 554-559; (c) Hammond and Ssiboukis, {377d}; (d) Yongzhong Huo, "A classification of thermodynamical potentials for two-variable transition systems," *Meccanica* [Netherlands], 30(5), Oct. 1995, p. 475.494; (e). Kotiuga, {377e}; (f) J. N. Murrell and S. Carter, "Approximate single-valued representations of multivalued potential energy surfaces," J. *Phys. Chem.,* 88(21), Oct. 11, 1984, p. 4887-4891; (g) Penman and Fraser, {377f}; (h) Ren, {376a}; (i) Z. Ren, "Solving 3D static field problems by dual formulations using potential variables," ***2nd** Int. Workshop on Elec. and Magn. Fields,* Leuven, Belgium, May 1994; (j) Ren, {376b}; (k) Verite, {376c}; (1) V. Vittal, "A generalized procedure to obtain first integrals for non-conservative dynamical systems: Application to power systems," *IEEE ISCAS '89,* 1989, p. 1984-1987.

427. G. W. O. Howe, "Effect of Torsion on a Longitudinally-Magnetized Iron Wire," *Wireless Engineer,* 29(344), May 1952, p. **115**-117.

428. Harry E. Burke, *Handbook of Magnetic Phenomena,* Van Nostrand, New York, 1985. This is a one-of-a-kind handbook on a great number of magnetic effects and phenomena.

429. Cullity, {422b}.

430. As examples: (a) Barrett, {345}; (b) Gurevich and Melkov, {232c}; (c) William Weltner, Jr., *Magnetic Atoms and Molecules,* Dover, Mineola, NY, 1989; (d) Giorgio Bertotti, *Hysteresis in Magnetism,* Academic Press, New York, 1998; (e) L'vov, {29c}; (f) Sidney Platt, *Magnetic Amplifiers: Theory and Application,* Prentice Hall, Englewood Cliffs, N.J., 1958; (g) Maciej J. Ogorzalek, *Chaos and Complexity in Nonlinear Electronic Circuits,* World Scientific, New Jersey, 1997; (h) Jammer, {1b}; (i) Borovik-Romanov and Sinha, Eds, {369b}; (j) Paul Davies, ed., {160f}; (k) G. I. Shipov, *A Theory of Physical Vacuum: A New Paradigm,* Moscow, 1994; (1) Nicolis and Prigogine, {255b}; (m) Richard Feynman, *Quantum Electrodynamics,* 1961, 1963; (n) V. 1. Fushchich and A. G. Nikitin, *Symmetries of Maxwell's Equations,* D, Reidel, 1987; (o) Mandl and Shaw, {19}; (p) M. W. Evans, Ed., ***Modern** Nonlinear Optics,* Second Edition, 3 vols., Wiley, 2001; (q) Robert C. O'Handley, *Modern Magnetic Materials: Principles and Applications,* Wiley, New York, 2000; (r) Sachs, {126a}.

431. As examples: (a) Lindsay and Margenau, {114}; (b) Richard P. Feynman, *The Character of Physical Law,* MIT Press, Cambridge, 1965:; (c) Feynman, {7}; (d) Ilya Prigogine, *The End of Certainty: Time, Chaos, and the New Laws of Nature,* Free Press, New York, 1996, 1997; (e) Mendel Sachs, *Einstein Versus Bohr: The Continuing Controversies in Physics,* Open Court Publishing Company, 1988.

432. For those unfamiliar with magnetics and the exchange force(s), see H. J. Gray and Alan Isaacs, *A New Dictionary of Physics,* second edition, 1976 printing, *ibid.,* p. 194 for a precise definition. The force from the sudden exchange of charge is known as the *Heisenberg force.* That due to the exchange of spin direction is known as the *Bartlett force.* The force associated with exchange of both spin and charge is known as the *Majorama force.* Quoting: "These exchange forces may be interpreted as being due to the continual emission and absorption of virtual particles that carry the shared property from one particle to another. The exchange of virtual hadrons between interacting particles of the atomic nucleus is intimately involved in nuclear exchange forces.

433. Feynman, {7}, Vol. II, Chapter 37 covers magnetic materials including exchange forces, spins, and spin effects.

434. *The Encyclopedia of Physics,* Robert M. Besancon (editor), Second Edition, Van Nostrand Reinhold, New York, 1974, p. 518.

435. Cullity. {422b}. p. 132.

436. Kittel, {422a}, p. 443 (footnote 1).

437. Lindsay and Margenau, {114}, p. 216-217.

438. *Ibid.,* p. 213.

439. *Ibid.,* p. 215.

440. *Ibid.,* p. 214.

441. T. D. Lee, "Space Inversion, Time Reversal and Particle-Antiparticle Conjugation," *Physics Today,* 19(3), Mar. 1966, p. 23.

442. Lee, {73}, p. 187-188.

443. Romer, {84}.

444. Kenneth D. Moore, "Testing of a Johnson Magnetic Gate," *Explore,* 9(6), 2000, p. 70-75. Also available on www.cheniere.org.

445. (a) Raymond J. Radus and William G. Evans, "Apparatus Responsive to Direct Quantities," U.S. Patent #2,892,155, Jun. 23, 1959; (b) R. J. Radus, "Permanent-Magnet Circuit using a 'Flux-Transfer' principle," *Engineers' Digest,* 24(6), June 1963, p. 86; (c) (anon.), "On-Off Permanent Magnet with a Memory," *Electronic Equipment Engineering,* 11(4), Apr. 1963, p. 22, 26.

446. "Human fly has magnetic sole," *Electrical Engineering,* Apr. 1963, p. 294.

447. Malcolm McCaid, *Attraction and Repulsion,* Oliver and Boyd Ltd., 1967 has an entire chapter on switching magnets, and has two photos of an engineer hanging upside down and walking on the underside of an elevated steel beam, wearing the Radus magnetic boots.

448. For a description of "shuffler" boots, see A. W. Hangar and A. A. Rosener, "The use of permanent magnets in zero-gravity mobility and restraint 'footware' concept," *IEEE Trans, on Magnetics,* Vol. MAG-6, No. 3, Sep. 1970, p. 464-467. These crude shuffler boots are quite a contrast to the elegant Westinghouse/Radus stepping boots earlier developed and used.

449. Burke, {428}.

450. David Scott, "Magnetic "Wankel" for electric cars," *Popular Science,* June 1979, p. 90-91.

451. E.g., (a) John R. Wiegand, "Bistable Magnetic Device," U.S. Patent No. 3,820,090, June 25, 1974; (b) — "Asymmetric Bistable Magnetic Device," U.S. Patent No. 3,866,193, Feb. 11, 1975; (c) - "Pulse Generation by Changing Magnetic Field," U.S. Patent No. 4,309,628, Jan. 5, 1982; (d) — "Pulse Generator Using Read Head with Wiegand Wire," U.S. Patent No. 4,263,523, April 21, 1981; (e) - "Switchable Magnetic Device;" U.S. Patent No. 4,247,601, Jan. 27, 1981; (f) — "Method of Manufacturing Bistable Magnetic Device," U.S. Patent No. 3,892,118, July 1, 1975; (g) — "Multiple Pulse Generator;" U.S. Patent No. 3,793,532, Feb. 19, 1974; (h) — and John E. Opie, "Pulse Generator with Shaped Magnetic Field," U.S. Patent No. 4,484,090. Nov 20, 1984.

452. Kawai, {31}.

453. John C. Bedini, "Device and Method of a Back EMF Permanent Electromagnetic Motor Generator," U.S. Patent No. 6,392,370, May 21.

2002. Uses a regauging process for capturing available back **EMF** electromagnetic energy in the system. The available energy may be captured and used to re-energize the battery, and/or sent in another direction to be used as useful work, or dissipated in the system.

454. See {445c}, p. 22, 26. Quoting: *"The circuit differs from ordinary permanent magnets in three ways: (I) It exerts strong magnetic force at one end while exerting hardly any at the other, (2) the strong and weak magnetic poles can be switched end-for-end easily and at will, thus in effect turning the magnetism on and off, and (3) it remembers its direction of greatest magnetic pull indefinitely. "*

455. Burke, {428}, p. 250.

456. Deffeyes, {342}.

457. Stuart Young, *Nature*, Vol. 414, 29 Nov. 2001, p. 487.

458. Evans, {137b}, p. 84.

459. M. W. Evans and S. Jeffers, "The Present Status of the Quantum Theory of Light," in *Modern Nonlinear Optics,* Second Edition, 3 Vols., edited by M. W. Evans, Wiley, 2001, Vol. 3, p. 150.

460. Clarke, {55}, p. 38.

461. Patrick, Bearden, Hayes, Moore, and Kenny, {37}. A second patent is pending, and two more are in preparation.

462. Bearden, {222b}.

463. Evans etal, {38b, 38c}.

464. T. E. Bearden, James C. Hayes, James L. Kenny, Kenneth D. Moore, and Stephen L. Patrick, "The Motionless Electromagnetic Generator: Extracting Energy from a Permanent Magnet with Energy-Replenishing from the Active Vacuum," Sept. 2000, carried on restricted DoE website http://www.ott.doe.gov/electromagnetic/. A search on the internet will turn up other sources where the paper can be downloaded.

465. (a) Evans et al., "The Aharonov-Bohm Effect as the Basis of Electromagnetic Energy Inherent in the Vacuum," *Found. Phys. Lett.,* 2002, (in review); (b) — "Derivation of the B(3) Field and Concomitant Vacuum Energy Density from the Sachs Theory of Electrodynamics," ***Found.*** *Phys. Lett.,* 14(6), Dec. 2001, p. 589-593; (c) — "Operator Derivation of the **Gauge** Invariant **Proca** and Lehnert Equations:

Elimination of the Lorenz Condition," *Found. Phys.,* 30(7), 2000, p. 1123-1130; (d) — "Runaway Solutions of the Lehnert Equations: The Possibility of Extracting Energy from the Vacuum," *Optik,* 111(9), 2000, p. 407-409; (e) — "The Effect of Vacuum Energy on the Atomic Spectra," *Found. Phys. Lett,* 13(3), June 2000, p. 289-296; (f) Evans et al., {113}; (g) — "Derivation of the Lehnert Field Equations for Gauge Theory in Vacuum: Space Charge and Current," *Found. Phys. Lett.,* 13(2), Apr. 2000, p. 179-184;(h)Bearden, {222b}.

466. Details are on Jean-Louis Naudin's website, in the section dealing with the MEG, which is http://jnaudin.free.fr/html/megv21.htm#megstat.

467. For an introduction to the field and its literature, see J. Anandan, J. Christian, and K. Wanelik, "Resource letter GPP-1: Geometric phases in physics," *Am. J. Phys.,* Vol. 65, 1977, p. 180-185.

468. A discussion by Feynman of the Aharonov-Bohm effect is given in Feynman, {7}, "15-5: The vector potential and quantum mechanics," p. 15-8 to 15-14.

469. A similar phase change also is induced outside the localized presence of an electrostatic scalar potential fi. See Feynman, {7}, p. 15-10. Quoting: *"...the effect of an electrostatic field is to produce a phase change given by the negative of the time integral of the scalar potential fi"*

470. Ya. M. Blanter, C. Bruder, R. Fazio, and H. Schoeller, "Aharonov Bohm-type Oscillations of Thermopower in a Quantum-dot Ring," *Phys Rev. B,* Vol. 55, 1997, p. 4069 typify the kind of "power" considerations that have been analyzed in the past. Only a minuscule thermopower treatment is considered.

471. S. Pancharatnam, "Generalized theory of interference, and its applications," *Proc. Indian Ac ad. Sci.,* Vol. A44, 1956, p. 247-262.

472. See Nahin, {62}, p. 225. Quoting: *"J. J. Waterston's paper on the kinetic theory of gases, in 1845, was rejected by the Royal Society of London. One of the referees declared it to be 'nothing but nonsense,* ***unfit*** *even for reading before the Society.'... "Waterston's dusty manuscript was finally exhumed from its archival tomb forty years later, because of the efforts of Lord Rayleigh... "*

473. Aharonov and Bohm, {344a, 344b}.

474. R. G. Chambers, "Shift of an electron interference **pattern** by enclosed magnetic flux," *Phys. Rev. Lett.,* Vol. 5, July I960, p. 3-5.

475. A. Tonomura *et al,* "Observation of Aharonov-Bohm effect by electron holography," *Phys. Rev. Lett.,* Vol. 48, 1982, p. 1443-1446.

476. A. Tonomura *et al.,* "Evidence for Aharonov-Bohm effect with magnetic field completely shielded from electron wave," *Phys. Rev. Lett.,* Vol. 56, 1986, p. 792-795. In the latter paper, the magnetic fields were completely and painstakingly shielded from the electron wave in a very careful experiment. The experiment demonstrated the expected relative phase shift as well as direct evidence for flux quantization. This was a definitive experiment clearly proving the Aharonov-Bohm effect. In 1999, Tonomura was awarded the Franklin Institute Award for this definitive proof of the Aharonov-Bohm effect.

477. M. V. Berry, *Proc. R. Soc. Lond. A,* Vol. 392, 1984, p. 45.

478. R. A. Webb, S. Washburn, C. P. Umbach, and R. B. Laibowitz, "Observation of h/e Aharonov-Bohm oscillations in normal-metal rings," *Phys. Rev. Lett,* Vol. 54, 1985, p. 2696-2695.

479. B. Schwarzschild, "Currents in Normal-Metal Rings Exhibit Aharonov-Bohm Effect," *Phys. Today,* Vol. 39, Jan. 1986, p. 17-20.

480. Y. Aharonov and J. Anandan, "Phase change during a cyclic quantum evolution," *Phys. Rev. Lett.,* Vol. 58, 1987, p. 1593-1596.

481. (a) V. Chandrasekhar *et al., Phys. Rev. Lett.,* Vol. 55, 1985, p. 1610; (b) S. Datta *et al., Phys. Rev. Lett,* Vol. 55, 1985, p. 2344; (c) S. N. M. Ruijsenaars, "The Aharonov-Bohm Effect and Scattering Theory," *Ann. Phys.,* Vol. 146, 1983, p. 1-34.

482. S. Olariu and I. Iovitzu Popescu, "The Quantum Effects of Electromagnetic Fluxes," *Rev. Mod. Phys.,* 57(2), Apr. 1985, p. 339-436.

483. (a) R. Bhandari, "Polarization of light and topological phases," *Phys. Reps.,* Vol. 281, 1997, p. 1-64; (b) — and J. Samuel, "Observation of lopological phase by use of a laser interferometer," *Phys. Rev. Lett.,* Vol. 60, 1988, p. 1211-1213; (c) R. Y. Chiao et *al.,* "Observation of a lopological phase by means of a Mach-Zehnder interferometer," *Phys. Rev. Lett,* Vol. 60, 1988, p. 1214-1217; (d) T. Bitter and D. Dubbers, "Manifestation of Berry topological phase in neutron spin rotation," *Phys. Rev. Lett,* Vol. 59, 1987, p. 251-254; (e) Y. Hascgawa *et al.,* "Geometric

phase in coupled neutron interference loops," *Phys. Rev. A.,* Vol. 53, 1996, p. 2486-2492; (f) A. G. Wagh *et al.,* "Experimental separation of geometric and dynamical phases using neutron interferometry," *Phys. Rev. Lett.,* Vol. 78, 1997, p. 755-759; (g) D. Suter, K. T. Mueller, and A. Pines, "Study of the Aharonov-Anandan quantum phase by NMR interferometry," *Phys. Rev. Lett.,* Vol. 60, 1988, p. 1218-1220; (h) D. Loss and P. M. Goldbart, "Persistent currents from Berry's phase in Mesoscopic systems," *Phys. Rev. B,* Vol. 45, 1992, p. 13544-13561; (i) C. A. Mead, "The geometric phase in molecular systems," *Rev. Mod. Phys.,* Vol. 64, 1992, p. 51-85; (j) J. H. Hannay, "Angle variable holonomy in adiabatic excursion of an integrable Hamiltonian," *J. Phys. A.,* Vol. 18, 1985, p. 221-230; (k) Y. Aharonov et al., "Aharonov-Bohm Type Forces Between Magnetic Fluxons," Hewlett Packard Report HPL-BR1MS-97-24, Oct. 1997.

484. (a) F. V. Gubarev and V. I. Zakharov, "The Berry Phase and Monopoles in Non-Abelian Gauge Theories," *Internat. J. Mod. Phys. A,* Vol. 17, 2002, p. 157-174; (b) J. M. Robbins, "Topological Phase Effects," Hewlett-Packard Labs Technical Report HPL-BRIMS-97-03, Jan. 31, 1997; (c) E. Sjoqvist, "Anandan-Aharonov-Casher oscillations in a simply connected region, *Phys. Lett. A.,* Vol. 270, 2000, p. 10; (d) E. Sjoqvist et al., "Geometric phases for mixed states in interferometry," *Phys. Rev. Lett.,* Vol. 85, 2000, p. 2845; (e) D. J. Thouless, *Topological Quantum Numbers in Nonrelativistic Physics,* World Scientific, Singapore, 1998; (f) J. Mickelsson, "Vacuum Polarization and the Geometric Phase: Gauge Invariance," *J. Math. Phys.,* Vol. 39, 1998, p. 831-837; (g) W. T. Lee *at al.,* "Scalar Aharonov-Bohm Effect with Longitudinally Polarized Neutrons," *Physica B,* Vol. 241-243, 1998, p. 216; (h) A. Cimmino et al., "High Precision Measurement of the Topological Aharonov-Casher Effect with Neutrons," *Nuclear Instr. and Methods A,* Vol. 440, 2000, p. 579; (i) Ali Mostafazadeh, *Dynamical Invariants, Adiabatic Approximation and the Geometric Phase,* Nova Science, 2001.

485. E.g., see (a) M. W. Evans and L. B. Crowell, *Classical and Quantum Electrodynamics and the B(3) Field,* World Scientific, Singapore, 2001; (b) M. W. Evans, *0(3) Electrodynamics,* Vol. V *of The Enigmatic Photon,* Kluwer, Dordrecht, 1999.

486. Bearden, {12}.

487. Fradkov and Pogromsky, {364}. The quote is from p. 359-360.

488. (a) Bearden, {38a}, (b) Bearden *et al., {*464}, (c) Evans et al., {38b}, (d) Evans et al., {38c},

489. Seife, {184}.

490. I. J. R. Aitchison, "Nothing's Plenty: The Vacuum in Modern Quantum Field Theory," *Contemp. Phys.,* 26(4), 1985, p. 339, 360.

491. *Ibid.,* p. 372.

492. Robert Neill Boyd, "Black Holes and Negative Energy", carried on Internet website http://www.rialian.com/rnboyd'blackholcs-negative.htm.

493. Roger Penrose, *The Emperor's New Mind.* Penguin Books, 1991, p. 210.

494. Stephen Hawking, per Zebrowski, {2}.

495. Isaac Newton, *Philosophiae Naturalis Principia Mathematica,* 1687.

496. Quoted from D. W. G. Ballentine and D. R. Lovett, *A Dictionary ofNamed Effects and Laws in Chemistry, Physics, and Mathematics,* Fourth Edition, Chapman and Hall, New York, 1980. p. 227.

497. Dirac, {249}.

498. See also (a) P.A.M. Dirac. "The quantun theory of the electron," P*roc. Roy. Soc. Lond. A,* 117(777), Jan. 2, 1928. p. 610-624; (b) — "The quantum theory of the electron, Part II," *ibid.,* 118 (779), Mar. 1, 1928, p. 351-361.

499. Georg Friedrich Berhnard Riemann, "Foundat ions of a general theory of functions of a single complex variabable." Doctoral Thesis, 1851. In H. Weber, ed., *B. Riemann: Gesammelte Mathematische Werke,* 2nd Ed., paperback reprint, Dover, New York, 1953.

500. As pointed out by Charles W. Misner, Kip S Thorne and John Archibald Wheeler, *Gravitation,* W. H. Freeman and Company, San Francisco, 1973, p. 220.

501. Josephs, {172}.

502. Laithwaite, {172b}.

503. Ernst Mach, *The Science ofMechanics.* 1912, Chapter 2, section 6.

504. This statement of Mach's principle is quoted from Ballentine and Lovett, {496}, p. 201.

505. Albert Einstein, "Die Grundlage der Allgemeinen Relativitatstheorie," ["The Foundation of the General Theory of Relativity,"] *Annalen der Physik,* Vol. 49, 1916, p. 769-822. (English translation in *The Principle ofRelativity,* Methuen, NY, 1923.).

506. Misner, Thome and Wheeler, {500}, p. 543.

507. A. Einstein, Appendix V to his *Relativity, The Special and the General Theory,* translated by Robert W. Lawson, Crown Publishers, New York, 1961, p. 155.

508. Theodor Kaluza, *Sitz. Peuss. Akad. Wiss.,* Vol. K1, 1921, p. 966. English translation, "On the Unification Problem of Physics," is in Lochlainn O"Raifeartaigh, *The Dawning of Gauge Theory,* Princeton University Press, Princeton, 1997, p. 53-58. Einstein had held Kaluza's paper since 1919 before recommending it for publication by the journal.

509. {a} Oskar Klein, *Zeit. f. Physik.* Vol. 37, 1926, p. 895. An English translation, "Quantum Theory and Five-Dimensional Relativity," is in Lochlainn O"Raifeartaigh, *The Dawning of Gauge Theory,* Princeton University Press, Princeton, 1997, p. 59-69. See also {b} Oskar Klein, *Nature, Vol.* 118, 1926, p. 319.

510. Sachs, {126a}.

511. Sachs, *ibid.,* quoted from the publisher's summation on the book jacket itself.

512. Mendel Sachs, "On Unification of Gravity and Electromagnetism and the Absence of Magnetic Monopoles," *IL Nuovo Cimento* B, Vol. **114,** 1999, p. 123.

513. (a) Evans et al., {233f}; (b) Evans et al., {233b}.

514. Misner, Thorne, and Wheeler, {500}.

515. Albert Einstein, *Relativity: the Special and General Theory,* translated by Robert W. Lawson, Crown Publishers, New York, 1961, fifteenth edition.

516. A. Einstein, H. A. Lorentz, H. Minkowski and H. Weyl, with **notes** by A. Sommerfeld, translations by W. Perrett and G. B. Jeffery, *The*

Principle ofRelativity, Methuen and Company. Ltd., 1923. Also available in paperback from Dover.

517. (a) T. Townsend Brown, "A Method of and an Apparatus or Machine for producing Force or Motion, British Patent 300,311, Nov. 15, 1928; (b) — "Electrokinetic Apparatus," U.S. Patent # 2,949,550, Aug. 16, 1960; (c) — "Electrokinetic Transducer," U.S. Patent # 3,018,394, Jan. 23, 1962; (d) — "Electrokinetic Generator," U.S. Patent # 3,022,430, Feb. 20, 1962; (e) — "Electrokinetic Apparatus," U.S. Patent # 3,187,206, Jun. 1, 1965.

518. Jonathan W. Campbell, "Apparatus and Method for Generating Thrust Using a Two Dimensional, Asymmetrical Capacitor Module," U. S. Patent #6,317,310, Nov. 13,2001.

519. (a) Norman L. Dean, "System for Converting Rotary Motion into Unidirectional Motion," U.S. Patent # 2,886,976, May 19, 1959. See also (b) John W. Campbell, Jr., "Report on the Dean Drive," *Analog,* Sept. 1960; (c) — "Instrumentation for the Dean Drive," *Analog,* Nov. 1960; (d) — "Final Report on the Dean Drive," Analog, Dec. 1960. Also (e) G. Harry Stine, "Detesters, Phasers, and Dean Drives," *Analog,* Jun. 1976.

520. (a) Henry Wallace, "Method and Apparatus for Generating a Secondary Gravitational Force Field," U.S. Patent # 3,626,605, Dec. 14, 1971; (b) — "Method and Apparatus for Generating a Dynamic Force," U.S. Patent # 3.626.606. Dec. 14, 1971.

521. J. F. Woodward, "Method for Transiently Altering the Mass of Objects to Facilitate their Transport or Change their Stationary Apparent Weights", U.S. Patent #5,280,864, Jan. 25, 1994.

522. (a) J. F. Woodward, "A New Experimental Approach to Mach's Principle and Relativistic Gravitation," *Found Phys. Lett.,* 3(5), 1990, p. 497-506; (b) — "Corrigendum: A New Experimental Approach to Mach's Principle and Relativistic Gravitation," *Found. Phys. Lett.,* 4(3), 1991, p. 299;(c) — "Measurements of a Machian Transient Mass Fluctuation,[1]' *Found. Phys. Lett.,* 4(5), 1991, p. 407-423; (d) -- "A Stationary Apparent Weight Shift from a Transient Machian Mass Fluctuation," *Found. Phys. Lett.,* 5(5), 1992, p. 425-442; (e) — Maximal Acceleration, Mach's Principle, and the Mass of the Electron," *Found. Phys. Lett.,* 6(3), 1993, p. 233-244; (f) — "Making the Universe Safe for Historians: Time Travel and the Laws of Physics," *Found. Phys Lett.,* 8(1). 1995, p. 1-39; (g) - "Erratum," *Found. Phys Lett.,* 8(3), 1995, p. 303; (h) - "Killing Time."

Found. Phys. Lett., 9(1), 1996, p. 1-23; (i) — "Laboratory Test of Mach's Principle and Strong-Field Relativistic Gravity," *Found. Phys. Lett.,* 9(3), 1996, p. 247-293; (j) — "Twists of Fate: Can We Make Traversable Wormholes in Spacetime?", *Found. Phys. Lett.,* 10(2), 1997, p. 153-181; (k) — Mach's Principle and Impulse Engines: toward a Viable Physics of Star Trek?", presented at the NASA Breakthrough Propulsion Physics Workshop, Cleveland, Ohio, Aug. 12-14, 1997.

523. (a) Max Jammer, *Concepts ofMass, in Classical and Modern Physics,* Harvard University Press, Cambridge, MA, 1961. One also suggests the reading of Jammer, {1b}.

524. (a) Bearden, {15}, p. 11. One kilogram of mass represents 17.053 x 10^{50} switches per second of action quanta h/(4*PI) across the boundary of the mass, in a flux exchange between the mass region and its external environment. The acceleration force can be derived from the difference in "inswitching" and "outswitching" of action quanta (as the masses m_i of the system iteratively change between mass and masstime states). The paper, albeit crudely, derives Newton's three laws of motion in relativistic form and the square law of gravitation. It also indicates the mechanism (reaction) upon a mass that generates the mass's rate of progression through time. For a somewhat expanded treatment, see (b) T. E. Bearden, "Vacuum Engines and Priore's Methodology: The True Science of Energy Medicine," Parts I and II, *Explore,* 6(1), 1995, p. 66-76; 6(2), 1995, p. 50-62.

525. Today we would also point out that to first order time and mass have essentially the same energy density — i.e., they are both spatial energy compressed by c^2. So the "switches" of action quanta we used in 1973 may be taken to actually represent the time-charging and discharging, of a specific region of space. Rotation of time-energy from the time-domain into 3-space connection with a mass is a mass increase as well as a change of mass to masstime. Rotation of mass-energy out of 3-space back into the time domain (during photon emission) is a mass decrease, (a) Our paper, "Giant Negentropy..." {12}, proposing a solution to the source charge problem, argues that all 3-space EM energy comes from the time domain anyway and returns to it. The source charges normally absorb time-energy, and transduce it to 3-space EM energy or vice versa, but the 3-space energy radiated away is in expanded spatial energy form and not compressed into mass-energy. It is the very release of the c^2 compression of the time-energy that is the "transduction" process to provide the **radiated**

spatial EM energy. But if mere rotation in 4-space is used without **that** expansion/transduction, then the time energy appears as mass. This is consistent with the mechanism we advanced for the flow of a mass *m* through time. Note that when a photon is "absorbed", the spatial energy component of the photon is in fact compressed into mass energy, slightly changing the mass. But when it is re-radiated, it is radiated in expanded or "spatial" EM energy form. Without the expansion and transduction, pure rotation of the compressed EM energy between the time domain and the 3-space domain provides the notion of "time-charging" and "time discharging". Indeed, the very anomalous phenomena occurring in radiation detectors in electrolyte experiments at China Lake, already provide direct evidence for just such time-charging, as pointed out in (b) Bcardcn, {17}.

526. E. R. Laithwaite, "Rack and pinion motors," *Electronics and Power,* Vol. 16, 1970, p. 251-252.

527. Of interest: (a) E. R. Laithwaite, "Unexplained phenomenon at Savoy Place," *Electronics and Power,* Oct. 1972, p. 360-362 recounts the discovery of a free energy device violating the second law of equilibrium thermodynamics; (b) Laithwaite, {526}, gives some particulars of the device; (c) —*All Things Are Possible: An Engineer Looks at Research and Development,* I*PC Electrical-Electronic Press Ltd., Dorset House, London, 1976; (d) Laithwaite, {172b}; (e) — "Miles Walker — a pioneer at Met-Vick and UMIST," *Electrical Review,* 211(12), Oct. 15, 1982, p. 33-34; (f) — "Roll, Isaac, roll — Part I," *Electrical Review,* 204(7), Feb. 16, 1979, p. 38-41; (g) — "Roll, Isaac, roll — Part II," *Electrical Review,* 204(11), Mar. 16, 1979, p. 31-33; (h) — "The laws of science and of England," *Electrical Review,* 202(23), June 16, 1978, p. 27-28.

528. E. R. Laithwaite and W. R. C. Dawson, "Mass Transfer: An Introduction based on Experimental Work at the University of Sussex," Sep. 1994. A subtitle is: A system for the transfer of mass derived from the principle of conservation of momentum. On website http://jnaudin.free.fr/html/IPEmain.htm.

529. See WO95/30832, Nov. 16, 1995.

530. Eric Laithwaite and William Dawson, "Propulsion System," U.S. Patent Number 5,860,317, Jan. 19, 1999.

531. Sweet and **Bearden, {**29a**}.**

532. (a) Ning Li and D. G. Torr, *Phys. Rev.*, Vol. 43D, 1991, p. 457; (b) — *Phys. Rev.*, vol. 46B, 1992, p. 5489; (c) — *Bull. Am. Phys. Soc*, Vol. 37, 1992, p. 948. Also see (d) D. G. Torr and Ning Li, *Found. Phys. Lett.*, Vol.37, 1993, p. 948.

533. E. E. Podkletnov and R. Nieminen, *Physica C*, Vol. 203, 1992, p. 441-444.

534. Igor K. Kulikov, "Inertial and gravitational masses of bosons at finite temperatures," in NASA technical package sent in response to queries of "Study of Inertial and Gravitational Masses of a Boson," *NASA Tech Briefs*, 26(2), Feb. 2002, p. 56. The report by Dr. Kulikov of the Jet Propulsion Laboratory (JPL) is carried on the website as http://nasatech.com/TSP/PDFTSP/NPO30325.pdl'. This is a theoretical study of the relationship between the inertial mass and gravitational mass of a self-interacting neutral scalar boson in a heat bath. The difference in the two masses was determined as a function of the temperature of the heal bath. The theory for distinguishing the difference is given. Since it is presented as an inventor's report, we assume that a patent is either pending by or to JPL or already granted to it.

535. Mike LaPointe, {421}.

536. Jim Corum, T. Keech, D. Gray, P. Pesavento, and M. Duncan, "The Electro-magnetic Stress-Tensor as a possible Space Drive Propulsion Concept, presented at the AIAA Joint Propulsion Conference, July 9, 2001.

537. Hector Serrano, "Propulsion Device and Method Employing Electric Fields for Producing Thrust," WO 00/58623, Oct. 5, 2000.

538. William B. Stein, Electrokinetic Propulsion: The Ionic Wind Argument," Energy Conversion Laboratory, Purdue University, West Lafayette, Indiana 47906, Sep. 5, 2000. This is a Gravitec report. Results can be downloaded from http://foldedspace.com/data.html, and an even more rigorous future test is planned.

539. A motion picture of the Naudin's replica thruster in dynamic thrust performance is shown on http://jnaudin.free.fr/html/liftbldr.htm, along with information on many other successful "lifters". This website is a cornucopia of information on inertial propulsion and electrogravitation, including successful independent replication and tests.

540. Ryan Mahoney, "Huntsville Firm Developing New Propulsion Technology," Birmingham Business Journal, Birmingham, Alabama July 19,2001.

541. E.g., see J. Cameron, "An Asymmetric Gravitational Wave Propulsion System," presented in the Breakthrough Propulsion Physics session of the AIAA Joint Propulsion Conference, July 9, 2001.

542. Bearden, {40}.

543. Prange and Strance, {365

544. Privately, Sweet strongly affirmed to me that later he took the risk and pushed it a little further in output, actually levitating the system several times, including accelerating it upward, hovering it, and then letting it sink slowly down to come to rest on the bench.

545. Dirac, {249}.

546. Romer, {84}; see endnote 24, p. 109.

547. E.g., as by Giovanni Modanese, "On the theoretical interpretation of E. Podkletnov's experiment," Extract from report UTF-391/96, OANL gr-qc/9612022, presented to the *I.A.F. Congress 1997,* Number IAA-97-4.107.

548. Ling Ni and D. G. Torr, {532a-532d}.

549. Evans, {137b}; Evans, {485b}.

550. M. W. Evans, personal correspondence.

551. (a) Evans et al., {233b}; (b) — - "Development of the Sachs Theory of Electrodynamics," *Found. Phys. Lett.,* 14(6), Dec. 2001, p. 595-600; (c) Evans et al., {233a}.

552. As we point out in our discussion of cold fusion, use of this enormous time-energy density component in a time-reversal zone allows reducing the force of the gluons in a nucleon and nearly freeing the quarks. The time-reversal zone also reverses the law of attraction and repulsion of charges; now like charges attract. So the coulombic attraction of two H+ ions can attract them together so closely that each enters the strong force region of the other, forming a quasi-nucleus. As the other ions in solution move to negate the time-reversal zone, the zone decays, restoring the normal repulsion of the two protons in the quasi-nucleus. However, being exponential the strong force increases much more rapidly than the normal

coulombic repulsion grows. The result is to force the two protons even closer together, into a full nucleus with a full strong force between them. This forms a nucleus of deuterium, as one quark in one of the protons is flipped while still relatively free, to form a neutron. Different but similar reactions exist to produce the excess tritium shown in many cold fusion experiments, and a similar interaction — especially between two deuteron ions — forms an alpha particle.

553. Bearden, {17}.

554. Davies, {237}, p. 105.

555. Misner, Thorne, and Wheeler, {500}, p. 5.

556. Evans and Jeffers, {459}, p. 150.

557. Evans, {137a}, p. 477.

558. Banesh Hoffman, in Foreword to Yakov P. Terletski, *Paradoxes in the Theory ofRelativity,* translated from the Russian, Plenum Press, New York, 1968, p. vi, viii.

559. Kron, {162}, p. 466.

560. T. D. Lee, {239}, p. 46-47.]

561. This is well known to modern theorists, but is especially accented by Sachs and by Evans. E.g.. to even have an EM wave in space, the energy density of space has to be changing as the wave energy oscillates. This is an oscillating curvature of spacetime, a priori. So all EM waves move in a dynamically curving spacetime.

562. Sachs, {126b}.

563. These concepts are generally consistent with 0(3) electrodynamics. For rigorous exposition of 0(3) EM, see Evans, {137b [.

564. Felix Finster, "Definition of the Dirac Sea in the Presence of External Fields," *Adv. Theor. Math. Phys.,* Vol. 2, 1998, p. 963-985.

565. J. W. Lynn and R. A. Russell, "Kron's Wave Automaton," Journal unk, date unk, p. 131. We would appreciate any researcher furnishing us a completed citation for this reference. We have a copy of the paper, but without full citation details.

566. Kron, {162}. The quote is from p. 114.

567. Kron, {32}, p. 39.

568. Kron, {306}.

569. Kron, {32}, p. 41.

570. See Stigant, {310a} for a bibliography of Kron's **publications.**

571. Several overunity researchers have been assassinated by forces and parties unknown. Many very clever means are used to do the dirty work.

As an example, there is a standard EMF "shooter" which uses Venus modification of the wavefront, to produce a wave that instantly disrupts the control of the heart's beating. There are two versions at least; one about the size of a pocketbook and for close range (30 feet) only, and one about the size of a bazooka and with longer range (up to 100 yards). About 1 minute exposure to the beam of the small weapon will result in death of the targeted individual. I have personally been hit by the small version of that device (and I have a witness who was with me at the time, and also was exposed to it). We survived by instant recognition of what it was, and by bolting immediately out the nearby back door of a restaurant here in Huntsville. If the kill is successful, the autopsy will show that the person died of a heart attack.

Stan Meyer, an overunity researcher with numerous patents, rushed from a restaurant exclaiming "They're killing me!" and expired with just such a heart attack, possibly hit inside the restaurant by a close-range shooter.

Harry Mason's associate was killed in Australia in his apartment by a longer-range, bazooka-sized shooter. The assassin was spotted putting the shooter back inside the car below, and speeding away.

Sometimes a silenced rifle is used for sniping at long range. Sweet was fired at by one such would-be assassin — luckily just as he stumbled and fell on the steps of the foyer to his apartment complex. The bullet cracked right by his ear as he fell forward, passing where his head had just been the moment before.

Another "shooter" uses longitudinal EM wave peak power pulses. This type kill is instant and very permanent. This unit can be set to stun or render the targeted person unconscious, or set to kill. It can also disrupt and stop running automobile engines (there is another version for that also, using negative energy EMP weaponry). However, it time-charges the struck human body so that the body emits longitudinal EM waves for some time after the kill. Marinov was killed with such a shooter (apparently by

KGB-related foreign intelligence operatives) and his body then thrown off a rooftop to make it appear to be suicide. Where his body lay, the pavement glowed when the body was removed. Only one weapon on earth will kill a body so that it will induce luminescence in concrete or asphalt that it lies on for a period.

Another favorite weapon is an ice dart dipped in curare. I was threatened once and told I would be killed by such an ice dart weapon if I traveled in a 2-month period. Just four days past that period, and traveling and changing planes in Dallas-Fort Worth airport, a person was killed three feet from me in broad daylight, right inside the airport, with the exact symptoms of a curare ice dart kill. That was apparently just to teach me "they" were serious.

"They" is that set of cartels that Churchill called "the High Cabal". Bedini and I had severely offended the High Cabal by Bedini's successful transmutation of copper (and other things) into gold. John often makes his own power transistors for his amplifiers, and the slag in the electric oven turns into black ruby laced with gold and silver, confirmed by assay. We have had numerous other assassination attempts, too numerous to iterate. Suffice it to say that developing a good COP>1.0 EM power system can be hazardous to one's health.

Any significant researcher should be wary of "meeting with a sudden suicide" on the way to the supermarket. Another thing to beware of, is a calibrated auto accident where your car is rammed from the rear, and you are shaken up considerably. An ambulance just happens to be passing by moments later, and it will take you to the hospital. If still conscious, the researcher must not get in the ambulance unless accompanied by a watchful friend who understands the situation and the danger. Otherwise, he can easily get a syringe of air into his veins, which will effectively turn him into a human vegetable. If he goes to the hospital safely, he must be guarded by friends day and night, for the same reason, else he runs a high risk of the "air syringe" assassination during the night.

Simply trying to do scientific work, I find it necessary to often carry (legally) a hidden weapon. Both my wife and I have gun permits, and we frequently and legally carry concealed weapons.

As early as the 1930s, T. Henry Moray — who built a successful COP>1.0 power system outputting 50 kW from a 55 lb power unit - had to ride in a bulletproof car in Salt Lake City, Utah. He was repeatedly fired at by snipers from the buildings or sidewalk, with the bullets sometimes sticking in the glass. He was also shot by a would-be assassin in his own laboratory, but overpowered his assassin and recovered.

Over the years, probably as many as 50 or more overunity researchers and inventors have been assassinated for their efforts, and particularly For their success. Some have simply disappeared abruptly and never been heard from again.

These days the use of longitudinal wave shooters (such as killed Marinov) and such is growing, since no signature is left for a normal autopsy to discover. The Yakuza, e.g., uses such "shooters" in Japan to assassinate persons or politicians having incurred their displeasure. Others are subjected to the more traditional Yakuza assassination: Three or four assassins suddenly jump the unsuspecting target, stabbing him repeated and with extreme rapidity. His body may have 50 to 70 deep stab wounds — through the heart, the abdomen, liver, lungs, etc. — before the body strikes the pavement.

Other very sophisticated methods of legal entrapment, gaming, planting narcotics on one's premises and tipping off the drug enforcement agencies for a drug raid in the middle of the night, etc. are utilized.

Widely used is "gaming"', which uses deep psychological profiling to select a "vulnerability". The supercomputer with the game prepares a "scenario" much like writing a movie script of predictable interactions among the players to be chosen. The game involves "agents of influence" (AOIs — persons who are dogmatic, perhaps even radical, and have knee-jerk responses easily triggered, and whose vector direction when they "jump" is along or near the desired response direction), and one or two agents to do the phone conversations with the AOIs and trigger them into action), as well as a game controller or director. The computer programs also use prediction and artificial intelligence, and will forecast the probability of success of the selected game. A very extensive database of deep psychological profiles on useful AOIs and all major COP>1.0 researchers is kept in the computer. When the controller selects the game, the computer spits out the most recommended AOI players. The controller accepts or changes and then accepts, and the game scenario is then prepared. The agents trigger the AOIs, and the unwitting target is suddenly besieged from every direction, along a direction he already has an affinity or "vulnerability" towards. The controller follows the course of the game and the outcomes of all interactions between the target and the AOIs. The agent or agents add "nudges" or additional actions to overcome delays, accomplish redirection when the direction goes awry. etc. Over the decades, these games have gotten very effective indeed, and are quite difficult to content with. I laving survived about 200 of these games over the last two decades, I may yet write a book on some of the major ones that have been used against overunity researchers. The games are not confined just to the overunity field, however, and sometimes are also used to directly intervene in politics and other key situations, {754}. Machiavelli's methods are not dead; they thrive and have been highly computerized.

572. More than a dozen potential mechanisms for overunity EM interactions and systems are given in Evans et al., {233i}. Many of these interactions are likely in highly energetic astronomical phenomena such as the gamma ray burster and the X-ray burster.

573. Evans et al., {38b, 38c}.

574. Bearden, {38a}.

575. Bearden, {30}.

576. Sweet and Bearden, {29a}.

577. (a) A. G. Reiss *et al., Astron. J.,* Vol. 116, 1998, p. 1009; (b) S. Perlmutter *et al., Astrophys. J.,* Vol. 517, 1999, p. 565.

578. A. Einstein, *Sitzungsber. Preuss. Akad. Wiss.,* Vol. 1917, 1917, p. 142. Einstein came to regret his introduction of the cosmological constani as his "greatest blunder". Ironically, this may yet turn out to be one of his greatest contributions.

579. Prange and Strance, {365}.

580. (a) R. Lieu, "The effect of Planck scale space time fluctuations on Lorentz invariance at extreme speeds," *Astrophys. J. Lett.,* 2002 (in press). A preprint is on http://xxx.lanl.gov/abs/astro-ph/0202443. See also (b) Philip Ball, "Time gives rays a break," http://www.nature.com/nsu/020304/0209304-5.html .

581. In 1998, experimental evidence of the accelerated expansion of **the** universe was provided by two groups of astronomers; see (a) Riess *et al.,* (577a}; Perlmutter *et al.,* {577b}. For an attempt to provide a new model for the negative antigravity, see (c) L. Parker and A. Raval, "A new look ill the accelerating universe," *Phys. Rev. Lett.,* Vol. 86, 2001, p. 749-752; (d) Philip Ball, "New model of expanding universe," *Nature News Service,* Science Update, Nov. 12, 2001. A good coverage for the educated layperson is (e) Mario Livio and Alan Sandage, *The Accelerating Universe: Infinite Expansion, the Cosmological Constant, and the Beauty of the Cosmos,*^{1}' Wiley, New York, Dec. 2000. See also (f) P. D.

Mannheim, "Implications of cosmic repulsion for gravitational **theory.'"** *Phys. Rev. D,* Nov. 15, 1998, p. 3511.

582. J. C. Bedini and T. E. Bearden, " Process For Transducing **Forms** of Electromagnetic Energy," U.S. Provisional Patent Application, 200 I. A formal U.S. Patent Application is in preparation as of this writing.

583. For a deep consideration of what is "meant" by mass, and how subtle the concept really is, the reader should refer to the scholarly work of Jammer {523a}. The more casual reader should be prepared for somewhat of a shock. Mass is not at all a "cut and dried" concept, as usually assumed. Indeed, the simple materialistic view of mass seems to be very much opposed by modern physics findings.

584. (a) Bearden, {12}. If we insert the absorption of a scalar (time-polarized) photon by the ubiquitously-assumed unit point charge and the subsequent re-emission of a longitudinal photon as the process for "combining" a scalar and longitudinal photon, then our solution is also consistent with (b) Mandl and Shaw, {19}. Mandl and Shaw argue that the longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (i.e., electrostatic) potential. Our comment is that this argument, translated from particle terminology to wave terminology, then fits our re-interpretation of Whittaker's 1903 decomposition of the scalar potential in (c) Whittaker, {85}. It also strongly suggests the need for a complete reinterpretation of what is meant by "propagation of EM energy through 3-space".

585. E.g., see (a) V. I. Arkhipov *et al.,* "Negative transient currents in amorphous semiconductors." *Internal. J. Electronics* (UK), 51(6), 1981, p. 735-742; (b) — "An analysis of the dispersive charge transport in vitreous 0-55 As_2S_3 : 0*45Sb_2S_3," *Phys. Stat. Sol.* (A), Vol. 54, 1979, p. 67-77: (c) — and Rudenko, A. I., "Negative currents caused by injection-controlled polarization." *Solid St. Commun.,* 28, 1978, p. 675-676.

586. Evans, {137a, 137b, 251}.

587. David Jones, *Vancouver Sun Times,* Weekend Magazine, Dec. 17, 1977, p. 17. Article on Sid Hurwich's Toronto gun-jamming demonstration, stopping of watches, and jamming of a gun.

588. For proof of production of ordinary EM fields and energy at a distance by scalar **interferometry,** see **Evans** et **al., {**79}.

589. Miles and Bush, {198}, p. 101.

590. Bearden, {17}.

591. We point out, but do not further discuss, the obviously startling impact on the validity of the random statistical analyses normally used in scientific work if such "hidden variables" are present in the Earth's supersystem and interact with the experiment.

592. Kron, {192}, p. 466.

593. (a) Gabriel Kron, *Diakoptics: The Piecewise Solution to Large-Scale Systems,* MacDonald & Co., London, 1962. See also (b) Kron, {310b}.

594. Ilya Prigogine, Preface, (160b), p. xi.

595. Sir William Rowan Hamilton, 1837. Hamilton originated quaternion algebra, the Hamiltonian, etc. Hamilton's quaternion algebra has a higher group symmetry than either vectors or tensors. Since Maxwell's theory was first formulated in Hamilton's quaternion algebra and later reduced to vectors and then to tensors, ironically the first successful EM theory has been sharply curtailed rather than being further developed and extended.

596. (a) Aharonov and Bohm, {344a, 344b}.

597. Prigogine, {160b}, p. 104.

598. Lee, {441}, p. 23.

599. Jackson, {22}, p. 811-812.

600. Whittaker, {85}.

601. This is because time may be regarded as compressed spatial length, where the L compression factor is c. Since the energy of a wave is proportional to L^2 (where L is the wave amplitude), then the release of ST time-density as ST energy density provides nearly $9x10^{16}$ times as much effect per second as simple energy density curvature alone does per joule. For convenience, we refer to this as the "10^{17} gain factor" one obtains by releasing the compressed length energy from its time-density ST curvatture form.

602. T. Matsumoto, "Mechanisms **of Electro-Nuclear Collapse,"** *Proc, ICCF-7 (International Conference on Cold Fusion - 7),* Vancouver, BC, Canada, April 1998, p. 98.

603. Romer, {84}, endnote 24, p. 109.

604. See Harter et al., {102}. For a simple analogy, we may think of the longitudinal EM wave as a sort of velocity-modulated wave; i.e., it is rhythmically changing its velocity about some average value, where that value is usually taken to be c. Another way to visualize it is as an "accordion" wave, expanding and contracting along the line of travel.

605. There are several types (polarizations) of photons utilized in physics. E.g., see (a) Ryder, {90b}. On p. 147+, Ryder discusses four polarization states of the photon. Two of these are "transverse" —i.e., the well-known x- and y-polarizations of light; one is called "longitudinal," i.e., z-directed and related to Evans' B(3) field, and one is "scalar" or timelike.

The time-like (scalar) photon we are using in this paper seems to be yet another brand, something like a cross between the Ryder's third and fourth classifications. That is, our "scalar" photon is named such because when it is imperfect it propagates (i.e., in the 7 direction) at finite speed. When perfect, it just appears instantly everywhere it will be, without propagation through space in the usual sense. Hence it can be at an infinite number of points simultaneously - very close to the notion of the pure Coulomb gauge or of Bohm's quantum potential - or it can be impure and propagate at subluminal or superluminal speeds. The main notion we intend in our present use of the "scalar" photon for this treatise is that its oscillation is in the time domain and along the time axis. We use a model which produces the flow of time itself, and a time-component structuring of that flow — something which is still absent from physics. So our time-like (scalar) photon in one part very much resembles Ryder's discussion of the "scalar" photon, but is not just that alone. For instance, both our longitudinal photon and our scalar photon seem to have internal "photon" structures that we do not discuss in this treatise.

A still deeper discussion of photon polarization is in (b) Mandl and Shaw, {19}. The longitudinal and scalar polarizations are not directly observable, but only in combination, where they manifest as the "instantaneous" Coulomb (electrostatic) potential. So by this route one arrives at a well-established mathematical theory that allows a hidden substructure within **the photon itself.** Mandl and Shaw include the photon's four polarization states as related to the four independent degrees of freedom available in the vector potential A. Suppression of the longitudinal and scalar polarizations is accomplished by "fixing the gauge". This reduces the independent degrees of freedom from 4 to 2, giving the conventional electrodynamics. However, when one raises serious foundations flaws in that resulting conventional electrodynamics, it is obvious that the entire structure is still on somewhat shaky ground. An even better mode might be to consider the fundamental photon entity as existing in all degrees of freedom, and the various "kinds" of photons existing due to restrictions or partial restrictions placed on one or more of these degrees of freedom. This of course opens the stage to interactions that change one (restricted) type of photon into a different (restricted) type. Thai is essentially the basis for my use of the notion of "photon transduction". Tt also leads to the transposition of time into energy and vice versa, which I use in my own approach, but which is still missing from nuclear physics. Since experimentally the cold fusion phenomena seem to require that characteristic, we have retained it.

We are indebted for much of the preceding discussion to private correspondence from Bob Flower, a fine scientist and knowledgeable in this area. Flower has pointed out that the conventional photon theory outlined by the cited authors is missing the possibility that nonlinear phase-conjugate optics could make the photon's longitudinal and scalar polarizations manifest in an experiment. Or put another way, if the instantaneous Coulomb potential could be decomposed into a time-like component and a longitudinal component, it would allow the two hidden polarization states to be controlled directly. This would result in anomalous modifications of the EM vector potential, leading to measurable electromagnetic or quantum effects. Flower's observation seems to be the closest to our own view, stated more simply in the present paper.

606. The result of our yet-unfinished struggle in these *very* swampy photon waters is not entirely consistent with the conventional view. Nonetheless, the new approach does seem to fit and explain most of the anomalous cold fusion phenomena, as well as a wide variety of other phenomena previously deemed beyond the pale of physics itself. E.g., by noting that *mind is time-like in its operations,* and therefore directly uses time-polarized (scalar) photons, waves, and electrodynamics, then mind operations become totally electromagnetic, but of a dramatically extended electrodynamic nature presently being approached in quantum field theory That approach leads directly to a testable and engineerable theory of mind

on all its levels (conscious, subconscious, unconscious, species unconscious, biospheric or all-species unconscious, and so on). But since the mind can interact with the body and produce at least the initial spatial forces carrying out intent in the physical world, it follows that some process must exist whereby the "time-like" photon interacts with and changes one or more of the other photons in the conventional view. Our use of *transduction* in fact encompasses that, based on quantum field theory as described by Ryder {90b} and by Mandl and Shaw {19}. Further, the present author's transduction reinterpretation of Whittaker's 1903 and 1904 papers {85, 91a} means that Whittaker's work as reinterpreted strongly supports this approach mathematically. So the approach used directly incorporates *mind and matter interaction,* again on multiple levels. While this area of *engineerable* mind and matter interaction is well beyond the scope of this treatise, it has become possible to begin the physics of the living being — even of the entire living universe — and to foresee direct engineering of the various realms of mind itself. Indeed, Russian psychoenergetics - particularly the highly weaponized secret part of it - already is involved in just that capability in deployed operational equipment of great portent. Since real results and real explanations of diverse metabiological phenomena have arisen from the approach as well as a very deep "more ordinary" physics including the necessary corrections to electrodynamics, we have continued to follow it to see where it eventually leads. A future book is planned on the more positive and beneficial uses of such startling technology. A preliminary view can be had on the author's website, www.cheniere.org, in the "porthole concept" briefing. A patent application has in fact been filed on this startling new method of medical healing.

607. (a) Vlail Kaznacheyev and L. P. Mikhailova, *Ultraweak Radiation in Intercellular Interactions,* [in Russian], Novosibirsk, 1981 [an English exposition of much of the Kaznacheyev work is given in (b) Vlail Kaznacheyev, "Electromagnetic Bioinformation in Intercellular Interactions," *Psi Research,* 1(1), Mar. 1982, p. 47-76.] See also (c) Kaznacheyev *et al.,* "Distant intercellular interactions in a system of two tissue cultures," *Psychoenergetic Systems,* 1(3), Mar. 1976, p. 141-142. In the same issue, see (d) "A Comment" by William A. McGarey, p. 143; (e) "A Comment" by Arthur C. Hastings, p. 143-144. See also (i) Yu. A. Vladimirov, "Ultraweak luminescence of subcellular structures," *Ultraweak Luminescence in Biology: Proceedings of the Symposium,* Moscow, 1969 [In Russian], Robert Becker also showed in his healing of

otherwise intractable bone fractures that a simple scalar potential (now known to involve the direct transduction between scalar and longitudinal photons and vice versa), that cellular dedifferentiation and redifferentiation of red blood cells occurred in the scalar potential region across the damaged bone cells, and directly in such manner as to "eliminate the delta engine" in those damaged bone cells in the fracture. The red cells shucked their hemoglobin, turning back (time-reversing) to a more primitive cell (earlier state). Then they redifferentiated (time-forwarded) into the type cells that make cartilage. There is still a delta engine between cartilage and bone! So these cartilage-like cells redifferentiated again into the kind of cells that make bone. These latter cells were deposited in the fracture, healing it. This process is in fact used today in various medical hospitals, without any real understanding of the physics mechanism involved. In short, it has long been experimentally proven without understanding the actual mechanism by which it works. See (g) Robert O. Becker, "The direct current field: A primitive control and communication system related to growth processes," *Proc. XVIInternal. Congr. Zoology,* Washington, D.C., Vol. 3, 1963, p. 179-183; (h) — and Charles H. Bachman and Howard Friedman, The direct current system: A link between the environment and the organism," *New York State J. Medicine,* Vol. 62, April 15, 1962, p. 1169-1176; (i) — and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," *Bull. N.Y. Acad. Med,* Second Series, 48(4), May 1972, p. 627-64. See also (j) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone *in vivo," Nature,* Vol. 204, Nov. 14, 1964, p. 652-654.

608. Puthoff, {202}.

609. See (a) A. Shapere and F. Wilczek, Eds., *Geometric Phases in Physics,* World Scientific, 1989. See also (a) M V. Berry and J. M. Robbins, "Indistinguishability for quantum particles: spin, statistics and the geometric phase," *Proc. Roy. Soc. LondA,* Vol. 453, 1997, p. 1771-1790.

610. See Mandl and Shaw, {19}, Chapter 5 and simply reinterpret the coupled scalar and longitudinal photon pair as a spin 2 graviton.

611. In fact, Sweet's vacuum triode amplifier probably used the domains and nuclei in a barium ferrite permanent magnet in such a graviton reaction manner. In an epochal experiment designed and conceived by the present author, Sweet reduced the weight of the device

on the lab bench, smoothly and efficiently, by 90 degrees. For **the** description, see Sweet and Bearden, {29a}.

612. Reginald I. Gray, *Unified Physics,* Naval Surface Warfare Center, Dahlgren, VA, 1988, p. 2-19.

613. For the basic explanation of a scalar interferometer, see Evans et al., {79}.

614. See Chapter 3, paragraph 3.1 through 3.5.

615. Bearden, {12}.

616. As an example, see Bunge, {3}, p. 179-180. Quoting: " *The most complete direct interparticle action theory of electromagnetism is WHEELER and FEYNMAN'S (1949). Its central postulate is Fokker's action principle for a collection ofpoint charges.... from which the equations of motion for charged particles follow. As can be seen, only particle variables occur in this theory, which is an electrodynamics in AMPERE'S sense and therefore the opposite of a unitary field theory. (The theory is mathematically equivalent to an "adjunct" field theory, but this adjunct field, far from being free, is attached to the particles.) The theory has been abandoned by its creators because it is inconsistent with the quantum theory, in particular with the photon hypothesis. Yet it continues to appeal to many not only because it smacks of mechanism and operationalism, but also because no self-forces appear in it and consequently no infinite self-forces (and the corresponding self accelerations) But this is too small an advantage compared with its shortcomings even at the classical level. Indeed, the theory makes use of special relativity - as is apparent from its central axiom...yet it is inconsistent with it for SR borrows from CEM the assumption that e.m. signals propagate in a vacuum with a constant speed c and moreover lead an existence, which, while it lasts, is independent of both emitters and absorbers. In other words, by denying the existence of free radiation fields propagating in space, the action at a distance formidation of CEM contradicts its own basis: it is logically inconsistent.* "

Our comment is that the "photon" interaction itself must be re-examined in light of the demonstrated giant negentropy of every change and dipole in the universe. Allowing for the quantum field theoretic combination of the scalar and longitudinal photons as an observable "photon" that is actually a spin-2 graviton in 4-space rather than 3-space. the entire nature of "propagation of EM energy through space" changes. No photon has ever interacted in just 3-space; instead, it interacts in

4-space. In the strong interpretation of giant negentropy, there is really no such thing as a "single observable photon" in the universe. Instead, there is a combined scalar/longitudinal photon pair, which when interacting in matter causes transverse EM wave movement of the interacting/detecting charges which are restrained by (i) the inertia due to their own mass, and (ii) in circuits, the back repulsion of all other like charges longitudinally ahead of the excited charge that absorbs the "photon" (actually, the graviton). Considering the longitudinal wave "bunching and relaxing" actions resulting in galloping, then the conventional assumption of the constant speed of light c in vacuum is destroyed. What we perceive as "light propagating in space at speed c" already moves in a galloping and jerky motion *about* speed c — now moving superluminally, now moving subluminally — with only an *average* speed of c overall. In turn, anything which affects the performance and "galloping mechanism" exhibited by the longitudinal EM wave can in theory be used to change the speed c. With the coming advent of the Fogal semiconductor and direct superluminal communication, all such objections based on previous rather classical assumptions will simply fall by the wayside. We also stress that the speed of gravitational field is shown in astronomical events and effects to be extraordinarily superluminal by many orders of magnitude. Again, this greatly upsets many of the modeling assumptions in present physics. Physics is based on a set of models and a set of experiments and observed effects. If the experimental effects negate the theory, one has to change the theory or be guilty of abandoning scientific method altogether, thereby turning science into dogma.

617. Cohen, {89}.

618. See Rod Donnelly and Richard Ziolkowski, "Electromagnetic field generated by a moving point charge: A fields-only approach," *Am. J. Phys.*, 62(10), Oct. 1994, p. 916-922. In conventional language, both the transverse and longitudinal field components always try to be born initially, whenever an EM wave is being generated. Generation of the transverse component produces a function which zeroes the longitudinal component. The function appears in the spacetime representation of the transverse EM fields, obtained by Donnelly and Ziolkowski using special transforms. The phase conjugating and retroreflecting aspects of cold fusion experiments produce spacetime curvatures, directly affecting the "canceling function." The altered function becomes an "LW-reduction function" which is postulated to determine the mix of transverse and longitudinal field components that emerge in the interactions.

619. Whittaker, {85} shows that the scalar potential is **comprised of** bidirectional longitudinal EM wavepairs. Whillaker, {91a} shows **that all** fields and waves are comprised of two interfering scalar **potential** functions. Hence all fields and waves are comprised of **functions** of longitudinal and scalar wave interferometry.

620. Evans et al., {79}.

621. Usually the detecting GM tubes will detect the interferometry of the incident longitudinal EM waves with the weak LWs being emitted in the tube itself by time-decay of its time-charges. The resulting transverse wave EM energy ionizes the tube's gas, and the GM tube "reads." If little or no time-decay is occurring, then there will be little or no transduction into TWs, hence little or no ionization. In that case the instrument will not read.

622. Miles and Bush, {198}, p. 101.

623. *Ibid.*

624. (a) Evans, {137b}; (b) Evans et al., {233a, 233f}.

625. Rodrigues and Lu, {93a}.

626. Enders and Nimtz, {329b}.

627. (a) W. A. Rodrigues, Jr. and J. Vaz Jr., "Subluminal and Superluminal Electromagnetic Waves and the Lepton Mass Spectrum," *Kluwer Ac. Pub. Proceedings,* <hep-th/9607231> on Los Alamos web site; (b) P. Letelier and W. A. Rodrigues Jr. (Eds.), *Gravitation: The Spacetime Structure,* World Scientific Publishing Co., Singapore, 1994; (c) J. Vaz Jr. and W. A. Rodrigues Jr., "On the equivalence of Maxwell and Dirac Equations, and Quantum Mechanics," *Internal. J. Theor. Phys.,* Vol. 32, 1993, p. 945-958; (d) J. Vaz Jr. and W. A. Rodrigues Jr., "Maxwell and Dirac Theories as an Already Unified Theory," *Adv. Appl. Clifford Algebras,* Vol. 7 (S), 1997, p. 369-386; (e) W. A. Rodrigues Jr. and J. E. Maiorino, "A unified theory for construction of arbitrary speeds (0 < v < oo) solutions of the relativistic wave equations." *Random Operators and Stochastic Equations,* Vol. 4, 1996, p. 355-400; (f) Rodrigues and Vaz, {93b}.

628. Torn Van Flandern, "The speed of gravity - What the experiments say," *Phys. Lett. A,* Vol. 250, 21 Dec. 1998, p. 1-1 1.

629. *Ibid.,* p. 9.

630. *Ibid.,* p. 10.

631. A. A. Vlasov and V. I. Denisov, "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity," *Theor. Math. Phys.,* 53(3), June 1983 (English translation), p. 1208-1216. Translated from *Teoreticheskaya i Matematicheskaya Fizika,* 53(3), Dec. 1982, p. 406-418 [in Russian].

632. D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21.

633. Sen, {68}, p. viii.

634. Mandl and Shaw, {19}.

635. Logunov, A. A. and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," *Sov. J. Part. Nucl.,* 18(3), May-June 1987, p. 179-187.

636. Whittaker, {85}

637. I.e., that spacetime == scalar potential == virtual particle flux == Whittaker biwave set. If that holds, then when we engineer the internal structure of the scalar potential, we also simultaneously engineer the internal structuring of the vacuum and spacetime itself, with deliberately formed spacetime curvature engines (vacuum engines). This is vacuum engineering — by different means — of the sort proposed by Nobelist T. D. Lee, {73}, p. 380-381. On p. 383, Lee points out that the microstructure of the scalar vacuum field (i.e., of vacuum charge, or — in our view — of the vacuum potential) has not been utilized. Particularly see Lee's own attempt to indicate the possibility of using vacuum engineering, in his "Chapter 25: Outlook: Possibility of Vacuum Engineering," p. 824-828.

638. Ingram Block and Horace Crater, "Lorentz-Invariant Potentials and the Nonrelativistic Limit," *Am. J. Phys.,* 49(1), Jan. 1981, p. 67.

639. Roderic Lakes, "Experimental Limits on the Photon Mass and Cosmic Magnetic Vector Potential," *Phys. Rev. Lett.,* 80(9), Mar. 2, 1998, p. 1826. Our critique is that the given equations define nothing at all, since no equation is a definition. If one replaces the "=" with "≡" in the "definition" of the A-potential in terms of B, one then has B==curl A, which is actually the definition of B in terms of A. If A has no energy, then neither can B, which is just the swirl of the A-energy. If A has no curl, B vanishes. If A has no energy, B vanishes. Yet $dA/dt = -E$, whether

A has curl or not, and E has energy. Ergo, A must have energy if **any** operation upon it is to produce **E-field** energy, else one has **violated the** conservation of energy law and advocates creation of energy from **nothing.** Regardless of how one looks at B = curl A and B = curl A, neither the equation nor the identity defines A in terms of B, because B is always, just one aspect of A and yet A can exist without any curl at all — without B as demonstrated in the Aharonov-Bohm effect. But in no fashion can B exist without A. Therefore A is primary and B is derived, and if A has no energy then B can have no energy. V has energy and energy density, since in fundamental units any point of the potential V is identically the joules of energy collected upon a unit point positive electrical charge assumed at that point. If it had no energy, then no joules could be collected from it.

640. Weinberg, {194}, p. 109-110.

641. Sen, {68}, p. viii.

642. (a) Lee, {13a}; (b) Wu et al., {14}; (c) Lee, {239}, p. 57-59; (d) Lee, {73}, p. 184.

643. Bearden, {12}

644. Whittaker, {85}.

645. The clustering of virtual charges around observable individual charges — due to the polarization of the vacuum — is well known in quantum electrodynamics. Macroscopically one sees the charge as partially shielded by the cluster, so the clustering charges are known to have macroscopic significance. At close enough distances, one sees the "bare" charge with reduced shielding.

646. Mandl and Shaw, {19}.

647. (a) Aharonov and Bohm, {344}; (b) R. G. Chambers, {474}.

648. E.g., see (a) Barrett, {345}; Barrett and Grimes [Eds.], {235b}.

649. (a) Evans *et al,* {113a}; (b) — "Self-Inconsistencies of the U(l) Theory of Electrodynamics: Michelson Interferometry," *Found. Phys. Lett.,* 12(6), Dec. 1999, p. 579-584; (c) — *J. New Energy,* Vol. 4, Winter 1999, p. 231-235; (d) — "Derivation of the Lehnert Field Equations from Gauge Theory in Vacuum: Space Charge and Current," *Found Phys Lett.,* 13(2), April 2000, p. 179-184; (e) Evans et al., {233g}; (f) Evans **et** al.. {233e}; (g) — **"Equations of the** Yang-Mills Theory of Classical Electrodynamics," ***Optik, 111***(2), 2000, p. 53-56; (h) **Evans** el **al.,** {113***b}.***

Also (i) more than 100 papers on restricted U.S. Department of Energy Internet website http://www.ott.doe.gov/electromagnetic/.

650. Obviously we exclude active systems such as matter that decays in nuclear fashion. Those are not inert systems, but are "pre-excited" systems whose matter has had excess energy since its formation during the birth of the universe.

651. (a) Kondepudi and Prigogine, {160c}; (b) — "Thermodynamics, Nonequilibrium," *Ency. Appl. Phys.* Vol. 21, 1997. p. 311-337.

652. I.e., the known broken symmetry of the charge (regarded as a set of composite dipoles) in its vacuum energy exchange.

653. Kraus, {76}, p. 578, Figure 12-60, a and b.

654. James Croll, *Phil. Mag.,* Series 5, Vol. 2, 1876, p. 242-254.

655. Heaviside, {5a-5c}.

656. Poynting, {4a, 4b}.

657. Heaviside, {106}.

658. E.g., see Lorentz, {109}.

659. E.g., (a) Panofsky and Phillips, {110}; (b) Gough and Richards, {224b}. (c) Jackson {22}, p. 259, points out the arbitrary nature of the Poynting vector and disposes of any added nondivergent energy flow vector with the words *"Such an added term can, however, have no physical consequences."*

660. Simply examine *Vqt* around the forward emf section (the external circuit between the generator terminals and the ends of the source dipole inside the generator), and - *Vqt* across the generator terminals (and the source dipole).

661. Evans etal, {38b, 38c}.

662. Bohren, {24}. Independent replication is reported in the same journal issue; see (b) Paul and Fischer, {25}.

663. Bearden, {40}.

664. See Paragraph 8.3.16 in Chapter 8 for the explanation.

665. Olariu and Popescu, {482}.

666. Bertram Schwarzschild, "Currents in normal-metal rings **exhibit** Aharonov-Bohm effect," *Phys. Today,* 39(1). Jan 1986, p. 17-20.

667. Paul and Fischer, {25}.

668. One actually inputs far more electrical energy flow from **the** external electrical power supply than is conventionally accounted, if one accounts for the huge nondiverged Heaviside energy flow component that Lorentz arbitrarily discarded. That energy flow is actually present, but Is not accounted once Lorentz' integration trick is applied to integrate the entire energy flow vector around a closed surface assumed around any volume element of interest.

669. Evans et al., {113a}.

670. As is any EM system when the nondiverged very large Heaviside energy flow component is accounted in addition to the diverged Poynting energy flow component.

671. Bearden and Bedini, {582}.

672. As an example, two scalar potentials fi_1 and fi_2 may have precisely the same intensity magnitude, but different Whittaker internal bidirectional longitudinal EM wave structures. The effects of the two externally identical potentials — as in various chemical reactions — can be quite different. See Bearden, {17}.

673. Lee, {73}, p. 380-381.

674. Lee, {73}, p. 184.

675. For succinct summary of the major symmetries and broken symmetries in physics, see Lee, {239}, p. 57-59.

676. Wang and Chung, {33}.

677. Jean-Louis Naudin, "Chung's Negative Resistance Experiment", on website http://jnaudin.free.fr/enr/enrexpl.htm.

678. Burford and Verner, {35a}.

679. Early but sophisticated analogue simulator. E.g., see Kron, {306}, p. 173.

680. E.g., see (a) Kron, {32}, p. 39. Quoting: *"Although negative resistances are available* ***for*** *use with a network analyzer...* "; — "Numerical solution of ordinary and partial differential equations by

means of equivalent circuits." See also (b) Kron, *J. Appl. Phys.,* Vol. 16, Mar. 1945a, p. 173. Quoting *"...since none or only a few negative resistances exist on practical network analyzers) "* (accent added). In the second quotation, arguably Kron was required to insert the underlined words. In the first quotation, the censors missed the clarity of the statement.

681. Kron, {192}.

682. Sweet and Bearden, {29a}.

683. Self-oscillation in such magnetic materials (which are also optically active) is well known. E.g., (a) L'vov, {29c}; (b) Borovik-Romanov and Sinha [Eds.], {369b}. See particularly (c) Gurevich and Melkov, {232c}.

684. Rosenthal publicly verified the results of his tests of the Sweet device to the 26th session of the Intersociety Energy Conversion Engineering Conference (IECEC '91), Boston, Massachusetts, 1991.

685. David M. Pepper, "Applications of Optical Phase Conjugation," 254(1), Jan. 1986, p. 83.

686. Bearden, {15}, p. 2-10.

687. Bearden, {17}.

688. Time-polarized photons are recognized in quantum field theory, implying the existence of time-polarized EM longitudinal waves. In coupling with spatial longitudinally polarized EM waves in wavepairs, the coupling is observable as scalar potential energy (stress energy), based on an argument by Mandl and Shaw {19}. Opposing time-polarized longitudinal EM waves therefore constitute "time-stress", "time-stress potential", and optical pumping in the time domain, which is hypothesized to reverse the normal "forward time" law of attraction and repulsion of charged particles.

689. Kenneth Shoulders and Steve Shoulders, {38If}.

690. U.S. patent numbers 5,018,180; 5,123,039; 5,054,046; 5,054,047.

691. Jain, {35b}.

692. Cao et al, {264b}.

693. Prange and Strance, {365}.

694. Roger Penrose, {82}, p. v.

695. Bunge, {3}, p. 176.

696. Albert Einstein, {la}, p. xi-xii. Appendix I.

697. Raymond A. Serway, *Physics for Scientists and Engineers* ***with*** *Modern Physics,* Third Edition, updated version, Saunders College Publishing, Philadelphia, 1990, p. 620.

698. Max Planck, *Treatise on Thermodynamics*, 3rd ed., Dover, New York, 1945.

699. Paul Davies, *Superforce: The Search for a Grand Unified Theory of Nature,* Simon and Schuster. New York, 1984, p. 105.

700. D. Hilbert, *Gottingen Nachrichten,* Vol. 4, 1917, p. 21.

701. A. A. Logunov and Yu. M. Loskutov, "Nonuniqueness of the predictions of the general theory of relativity," *Sov. J. Part. Nucl.,* 18(3), May-June 1987, p. 179.

702. David Whitehouse, "Beads of doubt," BBC News, July 18, 2002, http://news.bbc.co.Uk/l/hi/sci/tech/2135779.stm.

703. Matthew Chalmers, "Second law of thermodynamics 'broken'," *New Scientist,* July 19, 2002, carried on New Scientist website at http://www.newscientist.com/news/print.jsp?id=ns99992572.

704. G. M. Wang, E. M. Sevick, Emil Mittag, Debra J. Searles, and Denis J. Evans, "Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales," *Phys. Rev. Lett.,* 89(5), July 29, 2002, 050601.

705. Elias P. Gyftopoulos and Gian Paolo Beretta, *Thermodynamics: Foundations and Applications,* MacMillan, 1991. This is an elaborate text developed over the years at MIT.

706. Dilip Kondepudi and llya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures,* Wiley, NY, 1999.

707. Valery Chalidze, *Entropy Demystified: Potential Order, Life and Money,* Upublish.com, 2000. Contains an excellent nonmathematical exposition of the second law.

708. J. S. Dugdale, *Entropy and Its Physical Meaning,* Taylor & Francis, 1996. This book is of a more mathematical bent, but a very thorough treatment.

709. Craig F. Bohren, *Am. J. Phys.,* 51(4), Apr. 1983, p. 323-327.

710. V. S. Letokhov, {157}, ibid.

711. (a) D. K. Sen, *Fields and/or Particles,* Academic Press, London and New York, 1968, p. viii. Thirty-two years later, we published the solution to that problem in (b) T. E. Bearden, "Giant Negentropy from the Common Dipole," *J. New Energy,* 5(1), Summer 2000, p. 11-23. Also carried on DoE restricted website http://www.ott.doe.gov/electromagnetic/ and on www.cheniere.org.

712. T. E. Bearden, "EM Corrections Enabling a Practical Unified Field Theory with Emphasis on Time-Charging Interactions of Longitudinal EM Waves," *J. New Energy,* 3(2/3), 1998, p. 12-28.

713. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics,* ibid., p. 459.

714. E.g., see D. Jou, *Extended Irreversible Thermodynamics,* Springer-Verlag, New York, 1996.

715. Sources for this table are: (a) Roy Porter, Consultant Editor, *The Biographical Dictionary of Scientists,* Second Edition, Oxford University Press, NY, 1994; (b) Kondepudi and Prigogine, {703}, ibid.; (c) Serway {697}, *ibid.;* (d) K. Martinas, L. Ropolyi, and P. Szegedi (eds.), *Thermodynamics: History and Philosophy,* World Scientific, London, 1991; (e) Stephen F. Mason, *A History of the Sciences,* New Revised Edition, Collier Books, NY, 1962; (f) E. T. Bell, *Men of Mathematics,* Simon & Schuster, NY, 1986; (g) Charles Singer, *A Short History of Scientific Ideas to 1900,* Oxford University Press, London, 1959; (h) William C. Reynolds, *Thermodynamics,* Second Edition, McGraw-Hill, NY, 1968; (i) Paul Edwards (Ed. in Chief), *The Encyclopedia of Philosophy,* Vols. 1-8, Macmillan, NY, 1967; (j) I. Asimov, *Asimov's Biographical Encyclopedia of Science and Technology,* Second Revised Edition, Doubleday, Garden City, NY, 1982; (k) Jeff Biggus, *Sketching the History of Statistical Mechanics and Thermodynamics,* 2002, http://history.hyperjeff.net/statmech.html, (1) J. B. Jones and G. A. Hawkins, *Engineering Thermodynamics,* Wiley, New York, 1986; (m) Kenneth Wark, *Thermodynamics,* Third Edition, McGraw-Hill, 1977.

716. Antoine Laurent Lavoisier, *Elementary Treatise* ***on Chemistry,*** 1798.

717. **I.** Asimov, *Asimov's Biographical Encyclopedia of* ***Science*** *and Technology,* Second Revised Edition, Doubleday. Garden City, NY, 1982.

718. Benjamin Thompson (Count Rumford), *Enquiry Concerning the Source of Heat which is Excited by Friction,* 1798.

719. Eugene F. Mallove, "The Mysteries and Myths of Heat: A Brief History of Hot and Cold," *Infinite Energy*, May-June 2001, issue #37.

720. Julius Robert Mayer, 1814-1878. In 1842, Mayer's classic paper on conservation of energy was published in *Annalen der Chemie.* He then published several papers himself. With no recognition and bitterly attacked, Mayer suffered a nervous breakdown and attempted suicide, being committed to institutions for several years. From 1858 he began to receive well-justified recognition. He was lauded by the scientific community and received the Royal Society's Copley Medal in 1871.

721. Rudolf Clausius, "On the Moving force of Heat and the Laws of Heat Which May be Deduced Therefrom," 1850.

722. Mallove, {719}.

723. James Clerk Maxwell, "A Dynamical Theory of the Electromagnetic Field," *Roy. Soc. Trans.,* Vol. CLV, 1865, p 459. Also in *The Scientific Papers of James Clerk Maxwell, 2* vols. bound as one, edited by W. D. Niven, Dover, New York, 1952, Vol. 1, p. 526-597.

724. Actually, it is not *proven* that the energy of the universe is constant. With the recent experimental findings that the expansion of the universe is accelerating, the kinetic energy of the universe could be increasing. Whether this is offset by an appropriate change in the physical universe's supersystem remains to be seen.

725. Jeff Biggus, "Sketching the History of Statistical Mechanics and Thermodynamics," http://history.hyperjeff.net/statmech.html.

726. For a discussion of the modern death of matter and the death of materialism, see Paul Davies and John Gribbin, *The Matter Myth,* Simon & Schuster, NY, 1992. Quantum mechanics shook the foundations of materialism and the clockwork world, and chaos has **finished the** devastation. For a scholarly dissertation on mass, see Max **Jammer,**

Concepts of Mass, in Classical and Modern Physics, Harvard University Press, Cambridge, MA, 1961.

727. Dilip Kondepudi and Ilya Prigogine, *Modern Thermodynamics: From heat Engines to Dissipative Structures,* Wiley, New York, 1998, p. 3.

728. Serway, {697}, p. 507.

729. Serway, *ibid.,* p. 507.

730. Serway, *ibid.,* p. 508.

731. Serway, *ibid.,* p. 527.

732. Lindsay and Margenau, *Foundations of Physics,* Dover, New York, 1963, p. 214.

733. Robert H. Romer, "Heat is not a noun," *Am. J. Phys.,* 69(2), Feb. 2001, p. 107-109. This is an editorial discussion by the Editor of *Am. J. Phys.* of the concept of heat in thermodynamics, where heat is not a substance, not a thermodynamic function of state, and should not even be used as a noun.

734. Serway, *ibid.,* p. 527-528.

735. Serway, *ibid.,* p. 588.

736. T. E. Bearden, *Energetics: Extensions to Physics and Advanced Technology for Medical and Military Applications,* CTEC Proprietary, Mar. 21, 1998, 200+ page inclosure to CTEC Letter to Gen. (Ret.) Walter Busby, Deputy Secretary of Defense for Counterproliferation and Chemical and Biological Defense, March 21, 1998. This communication proposed a crash program to develop a portable suitcase-size treatment device based on this effect, for use in treating mass casualties after forthcoming terrorist attacks by weapons of mass destruction. Treatment of a patient would have required only five minutes. Copies were also sent to the U.S. Air Force, National Institutes of Health, and other key government agencies. Three and a half years later, the attack on the World Trade Center and the Pentagon — and a foiled attack on the White House — occurred on September 11, 2001. Today the nation is at war against international terrorism, and extensive changes in defense of the heartland are also under way. However, there still appears to be no real solution to the problem posed by mass casualties, particularly which professional

strikes can produce casualties in the hundreds of thousandsor a few millions.

737. (a) A. J. Berteaud and A. M. Bottreau, "Analysc des rayonnements electromagnetiques emis par I'appareil Priore," [Analysis of the electromagnetic radiations emitted by the Priore apparatus], D.R.M.E., 1971, p. 3-12; (b) A. J. Berteaud, A. M. Bottreau, A. Priore, A. N. Pautrizel, F. Berlureau, and R. Pautrizel, "Essai de correlation entre l'evolution d'une affection par Trypanosoma equiperdum et 1'action d'une onde electromagnetique pulsee et modulee," [Trial of the correlation between the evolution of a disease by Trypanosoma equiperdum and the action of a pulsating and modulated electromagnetic wave], *Compt. Rend. Acad Sri.* (Paris), Vol. 272, 1971. p. 1003-1006: (c) R. Cambar, "Rapport general des travaux de la Commission de Controle constitute en vue de verifier l'un des effets biologiques obtenu par l'utilisation de l'appareillage de Priore A. Bordeaux," [General findings of the work of the control commission formed to verify one of the biological effects obtained by use of the apparatus of A. Priore at Bordeaux], 1969, I vol., (d) R. Courrier, "Expose par M. le Professeur R. Courrier, Secretaire Perpetuel de L'Academie des Sciences fait au cours d'une reur on a L'Institut sur les effets de la Machine de M. A. Priore le 26 Avril 1977," [Presentation by Professeur R. Courrier, Perpetual Secretary ot the Academy of Sciences, made at the meeting of the Academy on the effects of the machine of M. A. Priore.] (e) G. Delmon and J. Biraben, "La croissance du carcinome de Guerin sour Faction de champs electromagnetiques."[The growth of carcinoma treated by the action of electromagnetic fields], *Rev. Path. Comp.*, 3(2), 1966, p. 85-88. (f) G. Mayer, A. Priore, G. Mayer and R. Pautrizel, "Action de champs magnetiques associes a des ondes electromagnetiques sur l'orchite trypanosominne due lapin," [Action of magnetic fields associated with electromagnetic waves on the typanosomian orchitis of the rabbit], *Compt. Rend. Acad. Sci.* (Paris), Vol. 274, 1972, p. 3011-3014. Many more reports of the novel cures are in the French medical literature, particularly by Pautrizel (g) Priore's doctoral thesis (which was rejected when the project was supressed) is: *Guerison de la Trypanosomiase Experimental Aigue et Chronique par L 'action Combinee de Champs Magnetiques et D 'Ondes Electromagnetiques Modules.* [Healing of intense and chronic experimental trypanosomiasis by the combined action of magnetic fields and modulated electromagnetic waves], thesis submitted in candidacy for the doctoral degree, 1973. (h) Antoine Priore, "Apparatus for producing radiations penetrating living

cells," U.S. Patent No. 3,368,155, Feb. 6, 1968;(i) — "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816, Oct. 25, 1966; (j) — "Procede et dispositif de production de rayonnements utilisables notamment pour le traitement de cellules vivantes," [Procedure and Assemblage for Production of Radiation Especially Serviceable for the Treatment of Living Cells], Republique Francais Brevet d'Invention P.V. No. 899.414, No. 1,342,772, Oct. 7, 1963; (k) Eric Perisse, *Effets des Ondes Electromagnetiqv.es et des Champs Magnetiques sur le Cancer et la Trypanosomiase Experimentale* [Effects of Electromagnetic Waves and Magnetic Fields on Cancer and Experimental Trypanosomias], Doctoral thesis, University of Bordeaux No. 83, March 16, 1984.

738. (a) R. O. Becker, "The neural semiconduction control system and its interaction with applied electrical current and magnetic fields," *Proc. XI Internatl. Cong. Radiology,* Vol. 105, 1966, p. 1753-1759, Excerpta Medica Foundation; Amsterdam; (b) — "The direct current field: A primitive control and communication system related to growth processes," *Proc. XVI Internatl. Congr. Zoology,* Washington, D.C., Vol. 3, 1963, p. 179-183; (c)———"Electromagnetic fields and life," *Psychoenergetic Systems,* Vol. 3, No. 1-4, 1979, p. 119-128; (d) - - "A technique for producing regenerative healing in humans," *Frontier Perspectives,* 1(2), Fall/Winter 1990, p. 1-2; (e) - - "The Biological Effects of Magnetic Fields — A Survey," *Med. Electron. Biol. Engr.,* Vol. 1, 1963, p. 293-303; (f) — - "The bioelectric field pattern in the salamander and its stimulation by an electronic analog," *IRE Trans. Med. Electronics,* Vol. ME-7, July 1960, p. 202-207; (g) - - and Charles H. Bachman, and Howard Friedman, "The direct current system: A link between the environment and the organism," *New York State J. Med.,* Vol. 62, April 15, 1962, p. **1169**-1176; (h)———and Joseph A. Spadaro, "Electrical stimulation of partial limb regeneration in mammals," *Bull. N.Y. Acad. Med.,* Second Series, 48(4), May 1972, p. 627-64; (i) - - and Carlton F. Hazlewood, Abraham R. Liboff, and Jan Walleczek, *Electromagnetic Applications In Medicine,* NIH-OAM Electromagnetics Panel Report, Jan. 15, 1993; (j) B. Baker; J. A. Spadero, and R. O. Becker, "Electrical stimulation of articular cartilage," *Ann. N.Y. Acad. Sci,* Vol. 238, 1974, p. 491-499; (k) C. A. L. Bassett; R. O. Becker, and R. J. Pawluk, "Effects of electric currents on bone *in vivo," Nature,* Vol. 204, Nov. 14, 1964, p. 652-654. There are many other publications by Becker, who is a true pioneer.

739. http://www.cheniere.org.

740. Kondepudi and Prigogine, {727}, p. 90-91. See also W. Nernst, *A New Heat Theorem,* Dover, New York, I969.

741. Wang et al., {704}.

742. Maddox, John, *Nature,* 417 (2002), p. 903.

743. Buchwald, Jed Z., *From Maxwell to Microphysics* (University of Chicago Press, 1985), p. 44.

744. For a better meaning of "static" as dynamic but steady state, see Tom Van Flandern, *Phys. Lett. A,* 250 (1998), p. 8-9.

745. E.g., Sen, D. K., *Fields and/or Particles* (Academic Press, 1968), p. vii.

746. Wu, C. S. et al., *Phys. Rev.,* 105, 1957, p. 1413.

747. Weinberg, Steven, *Dreams ofa Final Theory* (Vintage Books, Random House, 1993), p. 109-110.

748. Jackson, J. D. and L. B. Okun, *Rev. Mod. Phys.*, 73, 2001, p. 663-680.

749. Kondepudi, Dilip and Ilya Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures* (Wiley, New York, 1998), p. 459.

750. Jou, D., *Extended Irreversible Thermodynamics* (Springer-Verlag, 1996).

751. Bohren, Craig *F.,Am. J. Phys.*, 51, Apr. 1983, p. 323-327. Under nonlinear conditions, a particle can absorb and re-emit more energy than is in the light incident on it. Metallic particles at ultraviolet frequencies are one class of such particles and insulating particles at infrared frequencies are another.

752. Paul, H. and R. Fischer, *Am. J. Phys.*, 51, Apr. 1983, p. 327.

753. See Hal Fox, and Mitchell Swartz (JET Technology, MA), "Progress in Cold Nuclear Fusion - Metanalysis Using an Augmented Database," presented at ICCF-5, Monaco, April 9-13, 1995. The authors extracted information from over 3,000 papers on cold fusion collected by Fox's company. Over 600 of these experiments reported excess heat."

754. E.g., see Petar K. Anastasovski, H. Fox, and K. Shoulders, "A New Approach to the Cosmic Red-Shift and to the Cosmic Microwave Sources," J. New Energy, 1 (2), Summer 1996, p. 79-87."

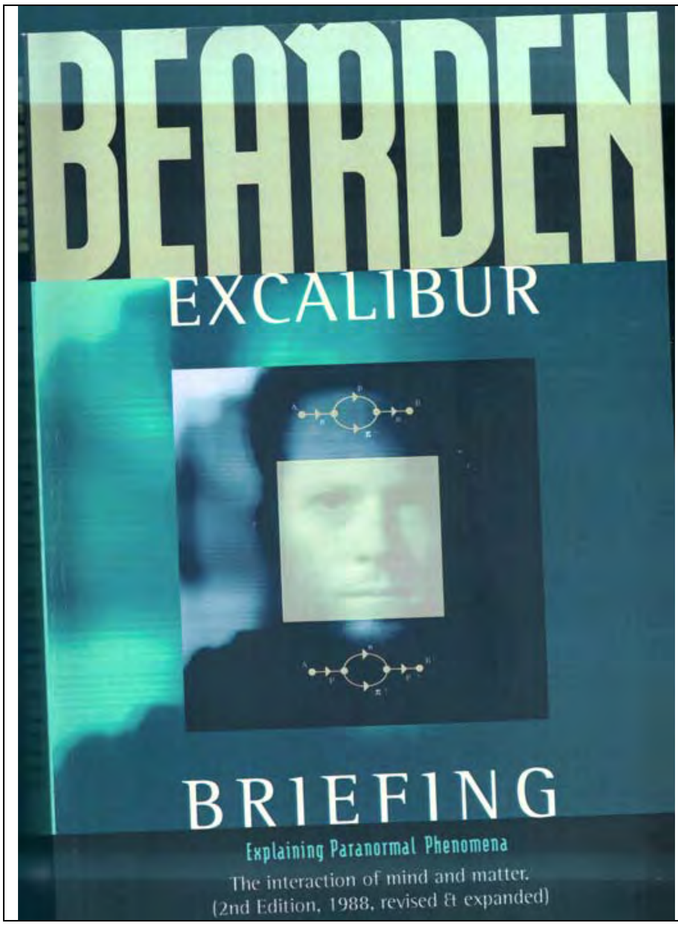

BEARDEN

EXCALIBUR

BRIEFING

Explaining Paranormal Phenomena

The interaction of mind and matter.

(2nd Edition, 1988, revised & expanded)

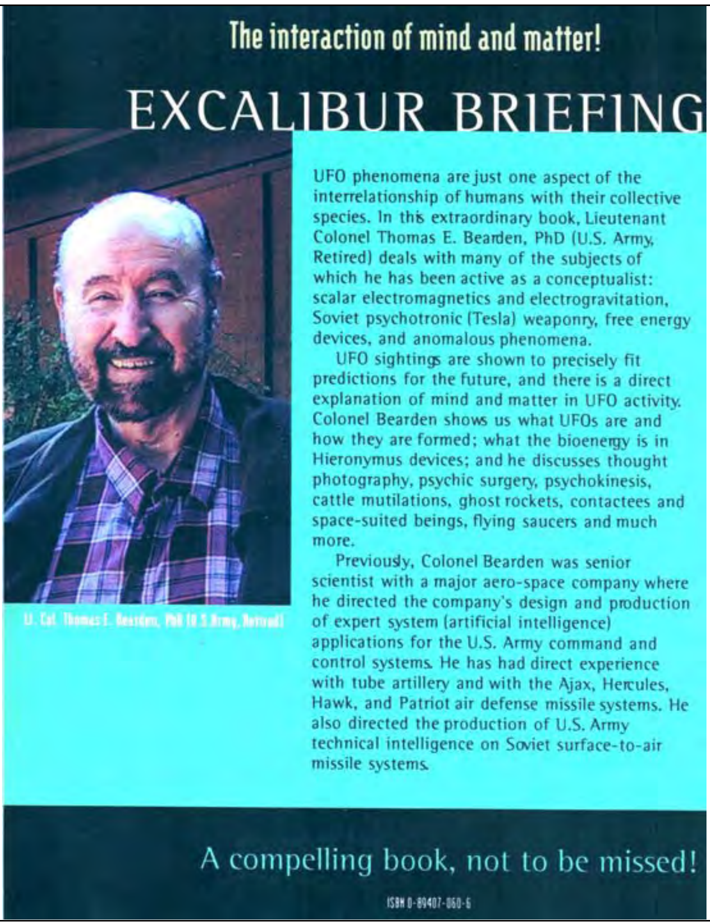

The interaction of mind and matter!

EXCALIBUR BRIEFING

UFO phenomena are just one aspect of the interrelationship of humans with their collective species. In this extraordinary book, Lieutenant Colonel Thomas E. Bearden, PhD (U.S. Army, Retired) deals with many of the subjects of which he has been active as a conceptualist: scalar electromagnetics and electrogravitation, Soviet psychotronic (Tesla) weaponry, free energy devices, and anomalous phenomena.

UFO sightings are shown to precisely fit predictions for the future, and there is a direct explanation of mind and matter in UFO activity. Colonel Bearden shows us what UFOs are and how they are formed; what the bioenergy is in Hieronymus devices; and he discusses thought photography, psychic surgery, psychokinesis, cattle mutilations, ghost rockets, contactees and space-suited beings, flying saucers and much more.

Previously, Colonel Bearden was senior scientist with a major aero-space company where he directed the company's design and production of expert system (artificial intelligence) applications for the U.S. Army command and control systems. He has had direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot air defense missile systems. He also directed the production of U.S. Army technical intelligence on Soviet surface-to-air missile systems.

Lt. Col. Thomas E. Bearden, PhD (U.S. Army, Retired)

A compelling book, not to be missed!

ISBN 0-89407-060-6

Sweating done by red_gonzo.

Note: sometimes formatting of the text may be a little fuzzy. It's all because the OCR process didn't go as smoothly as I expected. Also there may be some errors inside text so be vigilant. In any case share this book, send emails to people about knowledge contained in it.

Lieutenant Colonel Thomas E. Bearden (U.S. Army, Retired) Ph.D., was formerly senior scientist and department manager of a major aerospace company where he directed the design and production of expert systems (artificial intelligence) applications for the U.S. Army command and control systems.

He has over 29 years of experience in air defence systems, tactics, and operations; technical intelligence; command and control expert systems applications; nuclear weapons employment; computerized wargames; anti-radiation missile countermeasures; and military systems requirements. He has direct experience with tube artillery and with the Ajax, Hercules, Hawk, and Patriot air defense missile systems. He also directed the production of U.S. Army technical intelligence on Soviet surface-to-air missile systems.

Colonel Bearden obtained a Master of Science degree in nuclear engineering from the Georgia Institute of Technology, and a Bachelor of Science degree in mathematics from Northeast Louisiana University. He is a graduate of the U.S. Army Command and General Staff College, as well as of several U.S. Army artillery and missile schools.

He holds a Sandan (third degree black belt) in Yoseikan aikido and was the first president of the Yoseikan Aikido Association of North America. He is a former singer, guitarist, songwriter, and recording artist. He is member of Mensa, the Society for the Investigation of the Unexplained, the International Tesla Society, the American Association for the Advancement of Science, the Association of the American Physics Teachers, the Air Force Association, the Association of the U.S. Army, the American Nuclear Society, the Association of Old Crows, the American Defense Preparedness Association, and the Alabama Academy of Sciences. He is also East Coast Director for the Association of Distinguished American Scientists.

Tom Bearden is the author of several books and numerous papers, briefings and presentations, dealing with anomalous phenomena, scalar electromagnetics, free energy devices, Soviet Tesla weapons, and Soviet scalar electromagnetic weather warfare; he is also the producer of several videotapes on these subjects.

Table of Contents

Foreword, by John White . . . xv
Introduction to Second Edition . . . xvii
Introduction to First Edition . . . xxix

Part One: A Sampling of Specific Paranormal Phenomena

The Vestigia Mystery light Experiment . . . 35
Unexplained Generation of Electricity . . . 42
Hyperfield Generation of Photon Patterns About A Bar Magnet Collector . . . 43
Hyperfield Circulation Patterns Above a Magnetized Railroad Track . . . 46
The Cyborg Effect and Orthorotational Sensitivity . . . 49
Additional Factors Consistent with the Vestigia Experiment . . . 51
Remote Viewing . . . 52
The Moray Radiant Energy Device . . . 54
An Atlantean Power Crystal? . . . 60
The Crystal Skull . . . 61
Metal Bending . . . 65
Thought Photography: Stella Lansing . . . 65
The Displacement Effect . . . 68
Pavlita's Psychotronic Generators . . . 71
Cell-like Invisible Tupoidal Forms . . . 71
UFOs . . . 74
    1. Black Ring UFO . . . 74
    2. Anchor Patterns . . . 74
    3. Bifringement . . . 76
    4. UFOs in Water . . . 76
    5. UFOs in Clouds . . . 76
    6. UFO Seen by the Author . . . 76
    7. The 1966 Dennis Billings UFO Sighting . . . 82
    8. The Second Billings UFO Sighting . . . 87
    9. The 1973 Pascagoula UFO Sighting . . . 88
    10. Other UFOs . . . 89
Falkville Spaceman . . . 89
Men-in-Black . . . 94
Sasquatch . . . 95
UFOs Over the Soviet Union . . . 99
The Sighting Over Petrozavodsk . . . 103
French Experiments with Psychotronics . . . 104
Dowsing . . . 104
The Strobe Effect . . . 105
Kirlian Photography . . . 1ll
Interaction of Biofield, Thought and Spark Discharge . . . 1ll

Kervran Effect . . . 114
Psychic Surgery . . . 114
A Deliberate Experiment with Kindling . . . 118
Fay Clark's "Fireflies" . . . 118
"Foo" Fighters . . . 119
Water Monsters . . . 119
Cattle Mutilations . . . 123
Big Mama . . . 123

Part Two: A Theoretical Background for Understanding PT, UFOs, and Psi Phenomena

Some Unexplained Mysteries of Physics . . . 129
Uri Geller . . . 130
My Approach to Psychotronics . . . 131
    1. Reality as a Paranormal Bridge with Two Ends . . . 133
    2. A Fundamental Correction to Classical Logic . . . 136
    3. A Physical Example . . . 139
    4. New Definition of Zero . . . 140
    5. Einstein's Postulate . . . 141
    6. Synchronicity of Concepts . . . 142
Primitive Perception . . . 142
Two-Slit Experiment . . . 146
Photon Quenching of the Paranormal (Time) Channel . . . 151
Raindrop Model of Quantized Change . . . 155
Kindling Effect . . . 156
Bioenergy Collectors . . . 157
Psychometric Effect . . . 159
All Mass Is Charged: Everything Is Electrical . . . 160
A Funny Thing About Electromagnetic Field . . . 163
Extinguishing Electrical Currents . . . 165
Death Transmission Via the Paranormal Channel . . . 166
Disease and Death Induction . . . 169
Dynamic Thought Constructs: Biofield Mechanisms . . . 170
Radionics . . . 171
Psychotronic Devices: A Simple Space (Quiton) Amplifier . . . 172
Typical Hieronymus Device . . . 172
Another Kind of Hieronymus Device and the Inceptive Cyborg Effect . . . 175
Probability: Throw of a Die . . . 176
Negative Time . . . 177
Time Oscillation, Quantum Change, and Mass . . . 177
Mass Is a Time-Differentiation . . . 178
Reality Is Like an Opossum . . . 179
Biofield Concepts . . . 180
Mass, Force, and 3-D Frame Rotation . . . 180

Mass is pi Rotation of Hyperspatial Flux 184
Jitter or Dither Modulates the Primary Flux 185
The Untold Story of Quarks 186
Einstein's Spherical Model of the Cosmos 188
A Charged Particle Is a Dual Closure of the Universe 188
Adding and Subtracting Dimensions 189
Mind Is Objective 191
The Skin Transducer Effect: Turning a Thing into Its Opposite 193
Everett's Many-Worlds Interpretation of Quantum Mechanics 193
Spacetime Curvature Is Tricky 193
Multiple Realities (Many Worlds) 194
Clustered Orthogonal Worlds 196
Crosstalk Between Multiple Realities 199
Biofields: And Maverick Worlds 200
Hyperfield Flux in a Magnetic Field 201
Life and Death 202
The Holographic Hyperchannel Effect 209
Mind Linkage 212
Angels, Imps, and UFO Tulpas 214
Dreams and Unresolved Conflicts 215
Ghost Rockets and Flying Saucers: Materialized Tulpoids 218

Part Three: New Military Applications of Psi Research

Background to Psychotronic Research in the U.S. and the U.S.S.R 223
Psychotronic Background Since WWII 224
Electron Dissolution 225
The 1973 Contactee Wave and Cattle Mutilations 226
Radiation of the U.S. Embassy 229
Disease and Death Transmission 229
OtherWeapons 230
Hyperspace Howitzer Operation 230
A Possible Operational Test of Psychotronic Antiaircraft Weapons 236
Population Conditioner 238
A Midcourse ABM System 241
The Bell Island Incident 242
A Terminal ABM System 245
Sweep Mode, Sweeping Away a Nuclear Detonation 246
Zarg, Armageddon, and Linkage 247
Virtual States and Hyperspaces 254
Feynman Diagrams 256
Nature of an Electromagnetic Wave 261
Neutrinic Longitudinal Polarization Waves 262
A New Look at Modulation 265
Hyperspatial Aspects of Modulation 269

The Eureka Concept . . . 272
Biophotons and **Virtual EM Field of a** Bio-organism . . . 273
The Priore Machine . . . 273
The Neurophone . . . 278
Soviet "Woodpecker" Signals . . . 279

Part Four: Soviet Phase Conjugate Directed Energy Weapons

Weapons That Use Time-Reversed Electromagnetic Waves . . . 285

Glossary . . . 297

Bibliography . . . 329

List of Photographs

1. The Vestigia mystery light experiment (© Vestigia, Stanhope, NJ. Used by permission.) . . . 37-38,40
2. Hyperfield flux in a magnetic field (Courtesy George Meek) . . . 44-45
3. Remote viewing (Courtesy Dr. H.E.Puthoff) . . . 53
4. The Moray radiant energy device (©Dr. T.Henry Moray. Used by permission.) . . . 54-55
5. An Atlantean power crystal? (Courtesy Al Bielek) . . . 59
6. The crystal skull (©El Cariso Publications. Used by permission.) . . . 62
7. Metal Bending (© *New Realities*. Used by permission.) . . . 66
8. Clockwork patterns (© Mrs. Stella Lansing. Used by permission.) . . . 67
9. Displacement effect (Courtesy George Meek and Paul Trovillo) . . . 68
10. Pavlita's psychotronic generators (Courtesy James L. Hickman) . . . 69-70
11. Cell-like invisible tulpoidal forms (©Trevor James Constable. Used by permission.) . . . 72
12. Form for potential materialization (©Trevor James Constable. Used by permission.) . . . 73
13. New Mexico UFO (Courtesy Wide World Photos) . . . 75
14. Oregon UFOs (Courtesy Project Bluebook) . . . 77
15. Black Ring UFO (Courtesy Project Bluebook) . . . 78
16. Anchor patterns (Courtesy Ira Maxey, Project Bluebook, Brad Steiger, Fay Clark) . . . 79-80
17. Bifringement (Courtesy Project Bluebook) . . . 81
18. UFO in water (Courtesy Project Bluebook) . . . 81
19. UFO in clouds (Courtesy Project Bluebook) . . . 83
20. UFO seen by author (Courtesy of the late Ray Palmer) . . . 83
21. Second Billings UFO (© 1978 The American Association of Meta-Science) . . . 85
22. Spanish UFO (Courtesy Spanish National Air Ministry) . . . 91
23. Ohio UFO (Courtesy *Columbus Evening Dispatch)* . . . 91
24. Heflin UFO (Courtesy Rex Heflin) . . . 92-93
25. Falkville spaceman (Courtesy Wide World Photos) . . . 96
26. Men-in-black coin (Courtesy Warren York) . . . 98
27. Sasquatch (Courtesy Wide World Photos) . . . 101
28. Bigfoot cast (Courtesy Wide World Photos) . . . 102
29. The strobe effect . . . 106-107
30. Kirlian photography (Courtesy Danniel H. Kientz) . . . 108-110
31. Interaction of biofield, thought, and spark discharge (Courtesy James L. Hickman) . . . 112-113
32. Psychic surgery (Courtesy George Meek) . . . 115
33. Kindling of thought bioenergy in snow (Courtesy Mrs. Ann Baune) . . . 116
34. A deliberate experiment with kindling (© 1978 The American Association of Meta-Science) . . . 117
35. Fay Claris "Fireflies" (Courtesy Fay dark) . . . 120
36. 'Too" fighters . . . 121
37. Pleiosaur-like creature hauled from the sea by Japanese fishermen Courtesy World Wide Photos) . . . 122
38. Loch Ness monster (© Academy of Applied Science, Boston) . . . 124
39. Cattle mutilation (Courtesy Tom Adams and Walt Andrus) . . . 125
40. Big Mama (Courtesy Bill Jackson) . . . 126
41. Hieronymus device (Courtesy James B. Beal) . . . 219
42. Ghost Rockets (Courtesy of the late Ray Palmer) . . . 244

List of Figures

1. Hyperfield circulation patterns above a magnetized railroad track . . . 47
2. The cyborg effect and orthorotational sensitivity . . . 49
3. Reality: a paranormal bridge with two ends . . . 133
4. A fundamental correction to classical logic . . . 137
5. A physical example . . . 139
6. Primitive perception . . . 143
7. Two-slit experiment . . . 147
8. Photon quenching of the paranormal . . . 150
9. Raindrop model of quantized change . . . 154
10. All mass is charged: everything is electrical . . . 161
11. Death transmission via the paranormal channel . . . 167
12. Typical Hieronymus detector . . . 173
13. Biofield concepts . . . 181
14. Rotation of 3-D frames . . . 182
15. Mass is re-rotation of hyperspatial flux . . . 183
16. Einstein's spherical model of the cosmos . . . 188
17. Adding and subtracting dimensions . . . 190
18. Turning a thing into its opposite . . . 192
19. Multiple realities (many worlds) . . . 195
20. Clustered orthogonal worlds . . . 197
21. Crosstalk between multiple realities . . . 200
22. Biofields: time-clustered orthoframes . . . 201
23. Hyperfield flux in a magnetic field . . . 203
24. Life and death . . . 204
25. Consciousness, life, and intent . . . 207
26. The holographic hyperchannel effect . . . 210
27. Output of stage i . . . 211
28. Layers of consciousness . . . 213
29. Dreams are unresolved conflicts . . . 216
30. Collective conflict initiates prophetic tulpoids . . . 219
31. Examples of UFO waves . . . 227
32. 1973 Contactee wave . . . 228
33. Cattle mutilations . . . 231
34. Induction of Legionnaire's disease . . . 232
35. Hyperspace howitzer operation (2 figures) . . . 235
36. Nuclear Hyperspatial howitzer . . . 237
37. Amidcourse ABM system . . . 240
38. A terminal ABM system . . . 248
39. Sweep Mode: sweeping away a nuclear detonation . . . 249
40. ZARG . . . 252

FOREWORD

by John White

This book deals in a comprehensive fashion with such a vast array of topics — from physics, psychology and evolution to advanced weaponry, military strategy and international politics — that many of you may become anxious, even fearful on reading it. In fact, this assembly of previously uncorrelated information and scientific breakthrough may be the most disturbing you have ever received.

This cannot be helped. This book represents an important step in our efforts to avoid global war involving not only nuclear armaments but also what appears to be a revolutionary development in Soviet military technology — psychotronic weapons. They are the modern equivalent of the legendary invincible sword, Excalibur.

The author is former U.S. Army officer Lt. Col. Thomas E. Bearden, now retired. During his twenty-year military career specializing in artillery, nuclear weapons, and military intelligence, Tom has pondered and researched many subjects that he covers here. His access to the full range of both military and NASA documents will be evident in his book. What may become highly classified data in the future is being revealed to you, our nation's ultimate policymakers, for an urgent reason that Tom has called Operation Fer-de-Lance, the presumed Soviet plan for domination of the planet through nuclear-psychotronic aggression and warfare. For the moment I would simply like to note two statements that indicate the basis for our sense of urgency:

1. On 13 June 1975, Soviet Premier Leonid Brezhnev urged a ban on research and development of new kinds of weapons "more terrible than anything the world has known."
2. In January 1977 former Air Force Chief of Intelligence, Maj. Gen. George J. Keegan (Retired) told *Newsweek* magazine, "The Soviets are working on dramatically exotic new weapons, twenty years ahead of anything ever conceived in the U.S. — so awesome as to lead the Soviets to believe that in the coming decade they would be capable of total neutralization of our ballistic and submarine missiles."

The U.S. military intelligence community interprets these signs in conventional terms as perhaps indicating the development of laser rays. Tom disagrees. His book will deal with the possibility that these weapons are psychotronic or paraphysical in nature.

In fact, we can say that in a strict sense, these are not weapons but devices — breakthrough technology with an enormous potential for the destruction of humanity. They are a second-or third-generation outgrowth of radionic devices presently used in unorthodox medicine for diagnosis and healing. Extremely compact, they could be carried in a suitcase and planted near targets — ICBM sites, communications centers, etc. Even if one of these devices were found and examined, it would only look like an electronic switching circuit or a small transmitter/receiver set, rather than a bomb or explosive.

The energy and the mechanism involved in psychotronic damping of nuclear, chemical or electromagnetic technology is, from all indications, unrecognized by official Western science but nevertheless inherent in it. That also has been deeply pondered by Tom, and this book will give a full explanation of the physics, mathematics and technology involved. It is not an exaggeration to say, as Tom does, that the explanation offers a "physics of metaphysics" — a fully developed theory of paranormal phenomena that elegantly unites physics and psychology. The ramifications of what Tom calls perception theory lead into every field of human knowledge and inquiry with revolutionary results — from science, theology and philosophy through the full spectrum of paranormal phenomena such as UFO's and psychokinesis into the elementary school classroom, the supermarket shopping basket and the voting booth. They most certainly lead into the safety of your homes and the integrity of your minds.

Before you turn to *The Excalibur Briefing,* I want to leave you with these thoughts. The material you are about to read will certainly be denied by our own government and military and probably by a major segment of the scientific community. It is not unlikely that Tom will be dismissed as a paranoid crackpot. Such tactics have been employed before against brilliant, courageous people who challenge the fundamental dogma of the establishment as Tom does here. Nevertheless, I want to reassure you that Tom is sane, sober, and serious.

Of course, such a person could still perhaps be mistaken. In that regard, I believe you will find Tom's presentation to be persuasive. He himself is careful to distinguish speculation from fact, inference from descriptive statement. Even if Tom were paranoid, however, the question arises: *What if there is good reason for paranoia?* To phrase it positively: We should welcome truth and enlightening insight whenever offered, regardless of the source. Do not let *ad hominem* arguments affect your critical judgment about the soundness of this presentation. Assess the work, not the man, for the work is always greater. And having judged the work affirmatively, render true acknowledgement — not lip service — by sharing the ideas in this book with others. Only free, informed, vigorous effort in the court of world opinion can defuse the possibility of Operation Fer-de-Lance and psychotronic mind control leading to total enslavement of the human race.

In closing, I want to point out that Tom's fundamental concern is the unification and evolution of the human species to a higher form of life through a process that he terms "mind linkage." This stage of growth to higher consciousness is explored in the latter part of the book. It is a breathtaking portrait of how the spirit of science can lead to a science of the spirit, to a physics of metaphysics.

INTRODUCTION TO THE SECOND EDITION

Since the first edition of *Excalibur Briefing* was published, a great deal of progress has been made on the actual physics of unusual and paranormal phenomena. Indeed, so much progress has been made in the ten years since the 1978 or so cutoff of most of the Excalibur material, that I really should rewrite the entire book. However, time will not permit that luxury at present I am simply too hard-pressed in the strenuous efforts to move a free energy machine to market and to convince our government that unorthodox Soviet superweapons present us with a problem of the gravest magnitude.

Accordingly, the inclusion of one of my latest papers, "Soviet Phase Conjugate Directed Energy Weapons," together with an updated glossary for the entire emerging field, will have to suffice.

The important thing is that **a separate new field of physics is beginning to emerge.** It turns out that what we've gotten on to is a unified field—one can now turn electromagnetic energy into gravitational energy and vice versa. Indeed, gravitational potential is now revealed as simply infolded electromagnetics, and electromagnetics is simply the outfolding of the internal contents of gravitational potential.

Colleagues have already proven it on the laboratory bench, by building both "free energy" devices and practical antigravity devices.

This new, extended electromagnetics—electrogravitation—I have dubbed **scalar electromagnetics.** The key to it is to let the EM force fields fight themselves to a "cancellation," forming **a** vector zero. If we do this in a nonlinear medium (modulator), the summing/multiplying EM vector forces are locked together and remain as an infolded EM system inside the EM vector zero. This now is a gravitational system. By merely varying (in phase) the summing EM components, one varies the local energy density of vacuum. Rigorously that is a gravitational wave. Further, it is a very powerful G-wave, for we are utilizing EM forces some 1042 times as strong as the normal weak G-force.

Classical EM has taught us to simply discard the components in the zero vector summation, replacing the system of opposing EM forces with a vector EM zero. This of course makes all vector EM zeroes equal. However, it totally ignores the space-time/vacuum stress of the summed forces, and the resulting gravitational implications. Therefore we have been taught to ignore any gravitational aspects of normal EM circuits, and indeed we proceed to release the G-stress effects in the circuits when we build them as we have been taught

Ordinary EM represents the release of the internal stress of gravity. By teaching us to make nonzero EM, scientists have taught us to always release electrogravitational potentials as electromagnetic forces. Thus in classical EM—and in most of the circuits we build—the EM force fields are the primary causative agents. Classically, when these force fields reduce to zero, there is no more electromagnetics going on. That is true, but there can still be a very large amount of gravity and inertial effects going on!

However, quantum mechanics teaches us that the EM force fields are not at all the primary causes. Instead, the potentials are the primary things; EM forces are simply made from these potentials by differentiating operators. And when the force fields are zero, the potentials still exist. Further, when these "forceless potentials" interfere—

even at a distance—real EM effects cun still be obtained, in total violation of ordinary EM and ordinary mechanics!

This is called the Aharonov-Bohm Effect, and in 1986 it was finally proven to most physicist's satisfaction. Only the most diehard skeptics still object to it, according to *Physics Today,* January 1986.

At any rate, when we have zeroed the E and B fields, we can still have the potentials, according to modern physics. These potentials can interfere to produce electromagnetic force effects—even at a distance.

Further, the zero summation can be dynamic; that is, the vector components can be varying, so long as they always sum to a zero vector. Actually such a zeroed system now has an infolded, deterministic pattern of dynamic forces, according to the pattern we deliberately construct. Further, the energy of each force component infolded is varying, and thus the overall energy density of local vacuum is varying.

As we said, rigorously this defines a gravitational wave. So the zero-vector approach allows EM wave energy to be turned directly into gravitational wave energy. Now it is the curvature of spacetime that is being varied locally. That's a really astounding thing.

We can now deliberately engineer the vacuum/spacetime itself, with all that that statement implies. Locally we can bend spacetime itself, which Einstein assumed could never be done except by a huge collection of mass—such as a sun, star, neutron star, etc. But here we can do it with simple things such as opposing coils, opposing E-fields, etc. We've just violated one of the sacrosanct, severely limiting assumptions of ordinary (Einstein's) general relativity!

That is, up to now most Western physicists have assumed that the local spacetime/vacuum cannot be curved. They assume that, where we the observer and the laboratory instruments are, only special relativity applies. Locally the observer always has a "Lorentz frame," in that view. This assumption (the local Lorentz frame) means that all the conservation laws apply, and physical reality stays well-behaved where the laboratory is and where we are. Ordinary relativity becomes just a sort of "special relativity with distant perturbations."

The only curvature of spacetime of any consequence is only to be found at or near distant stars, suns, black holes, and other huge collections of mass, in this highly restrictive view. Further, this assumption implies that one can never have an experimental general relativity, since where one is experimenting (locally) there is always essentially zero spacetime curvature!

However, one need not have such a formidable restriction at all. Two bucking transformers can make a B-field vector-zero a little bit More complex circuits can easily be designed to sum (or multiply) force fields to zero very well. When we do this zero summation, we also get an amazing gain as follows: Consider two electrons in free space. As is well known, the electric field force (repulsion) between them is about 10^{42} times as strong as the gravitational field force (attraction) between them. So the electric force is so very much stronger that the comparatively weak gravitational force can be entirely neglected, in most applications. However, if we "strangle" that electric field force by zero summation, all the opposing electric forces are converted directly into gravitational potential!

Interference of two such potentials can directly generate gravitational forces. So, if we could zero-sum their E-fields perfectly, we could instantly amplify the gravitational force between the two electrons by a factor of 10^{42}. With such a gravitational force amplification factor, our altered "electrical" circuits would actually become powerful electrogravitational circuits. Further, appreciable gravity effects would now be experienced by any mass, object, or device exposed to the action of the circuits. This is the fundamental secret that Nikola Tesla discovered, one which he never openly revealed.

Further, by locking opposing EM waves together, we may easily construct purely scalar EM (electrogravitational) waves and beams. In such case, amazing new phenomena are encountered: (1) one can increase or decrease the mass and inertia of an object; (2) one can produce a unilateral force in and on each and every nucleon in a body, thus accomplishing the long-sought "antigravity"; (3) a unilateral force drive—a "space drive"—can be constructed; (4) transmutation of elements becomes fairly simply done, with miniscule energy; (5) effects at a distance—such as cold explosions or hot explosions—can be accomplished; (6) negative energy and negative time effects can be created and utilized; (7) Newton's third law can be manipulated and violated at will; (8) negative entropy effects can be created and utilized; and (9) direct energy extraction from the vacuum can be accomplished, etc. So many new effects are encountered that it will take a generation of graduate students and university laboratories to get a good handle on most of them.

The reader is urged to remember these rules: (1) The destructive interference of scalar EM waves (electrogravitational waves) produces EM energy in, or extracts it from, the interference zone. (2) The destructive interference of ordinary EM waves produces gravitational potential energy. (3) The release of the ordinary EM force fields represents the bleed-off of a 5-dimensional gravitational potential as ordinary electromagnetics. (4) The bottling-up of the ordinary EM force fields represents the forced bleed-off of a 5-dimensional gravitational potential as ordinary 4-dimensional gravitational force. (5) The EM field and the ordinary (4-space) G-field are just separate bleed-off s of the same 5-dimensional gravitational potential field.

Thus, I make no bones about it; we have a new physics in the offing—one that is eminently practical and easily engineered by any modern electromagnetics laboratory.

Since the first edition of this book was issued, it has also become obvious that at least two other phenomena are involved in UFOs, and intermingled with the "materializations from the collective unconscious" true UFO phenomena. These two additional things are:

(1) A great number of "lights at night" and "mysterious fireball" type phenomena—and giant balls of light and glowing hemispheres of light of gigantic size—are actually due to deliberate Soviet long range, high energy, real-time holography using phase conjugation. These phenomena are to accomplish orientation and alignment of giant Soviet weapons, and at the same time deliberately evoke UFO reports and scientific skepticism, providing perfect deception.

(2) Several nations of the world—at least two—have been building and flying antigravity spaceships all over the solar system, since the early 1950's. A very small number of UFO ship reports and encounters are due to these ships and their crews. However, control of these ships is not under any duly elected government, but passes out of the government concerned, into and under the control of some mysterious group

which I loosely refer to as "the control group." This is the same mysterious group which has been suppressing free energy and antigravity inventors and inventions for a great number of years.

Thus in *Excalibur Briefing* we are still addressing the primary UFO phenomena: Materializations from the collective human unconscious. Coverage of the other two parts of the overall reported UFO phenomena will be made elsewhere.

So it seems that, under my hard-pressed circumstances, the best thing to do for the second edition of *Excalibur Briefing* is to include something that shows some of the drastic progress that has been made. By adding a glossary and a current paper, the book will also be made much more usable. I also have prepared an index for the first edition, but it cannot be included because of space limitations. Admittedly that is regrettable and still not the optimum, but it will have to suffice. Perhaps in the future I will be able to do a complete rewrite of the book, for *Excalibur III.*

Until then, I ask the reader's indulgence and understanding.

* * *

In the first edition we accented the hyperspatial aspects of the problem, but also pointed out that higher dimensional spaces and deeper nested levels of virtual state are in fact one and the same thing. The reader should keep in mind that when we "inter-nest" vector zero systems, we are performing direct engineering in hyperspace.

Let me be more specific. When you zero sum a group of EM force vectors, you place those component vectors "inside the zero," so to speak. In modern language you have placed ordered patterns directly inside the violent flux of virtual particles comprising the vacuum. The modern concept of the vacuum is that it is like a seething plasma or gas. However, it's a very strange gas, for each particle arises spontaneously out of nowhere (according to quantum mechanics), and almost immediately turns into something else and/or disappears. Thus the vacuum plasma or gas is rather peculiar; the particles comprising it appear and disappear so rapidly that they cannot individually be detected. However, while a virtual particle exists, it is in motion, and so a violent flux of these "phantom particles" comprises the vacuum/spacetime.

However, just as many fleeting particles in the vacuum are moving backward in observer time as are moving forward in it! So this wild "vacuum gas" of ghost particles exists in both positive and negative time at once!

Anyway, when we make a zero vector EM system, each component force in the summation represents a gradient in this virtual particle flux. That is, we have "patterned" and organized the flux—polarized it, so to speak. And that now overturns one of the fundamental assumptions of quantum mechanics.

Quantum mechanics assumes that the violent flux of vacuum ghost particles is totally randomized. That is, no deterministic pattern exists in the flux. These virtual changes collect or cohere statistically, not deterministically, in the standard assumption. When they happen to collect/cohere sufficiently, the threshold of an observable quantum change is breached, and a quantum change occurs.

It can readily be seen that, in the ordinary view, the occurrence of these quantum changes is totally statistical. And indeed it is, so long as the virtual particle flux is randomized. However, with zero vector summation (or multiplication) we have produced deterministic patterns in the flux—and hence in the cohering and collecting of this flux on observable charged particles. Therefore, these deterministic "virtual state" ordered patterns will result in the appearance of deterministic patterns in the occurrence of observable quantum changes.

Now we have begun to engineer physical reality directly; that is, one can now begin to pre-determine which statistical change shall be favored, or in the extreme case, which one shall occur, and whether or not it shall even occur at all.

The renowned physicist, David Bohm, has written a theory of "hidden variables" which could fit and describe all of physics (nothing out there contradicts it). It has largely been ignored, and will continue to be ignored unless it can show appreciable application that normal theory cannot cover. With the scalar EM approach, we can now directly provide hidden EM force field vectors, ordering the virtual state vacuum flux, and we may vary these ordered virtual patterns at will to produce startling new results in charged particle systems. So now we can directly engineer Bohm's "hidden variables," and the vacuum itself—something ordinary theory cannot cover. Drastic changes in the physicist's view of fundamental physical reality, and of what he can do about it, now result

But to return: We wish to engineer additional hyperspaces. (The first zero-vector summation places the component vectors in the fifth dimension.) If we take one of the component EM vector forces in the first zero vector summation, and we then take a point on it as yet another "reference zero", we may encapsulate a second series of zero-system summation vectors into and within that zero-point. In this case we have placed these second vectors into yet a deeper level of virtual state vacuum. That is, we have reached into yet another hyperspace with our engineering and control. Indeed, we have placed the second system EM vectors into the sixth dimension. So we have become true hyperspace engineers.

The process is repeatable as desired. We may reach any hyperspace we wish, as our technology and control develop.

Now we see the impact In modern Kaluza-Klein geometry, one needs about 11 or so dimensions if one is to account for quarks, charm, etc. Now we have a direct tool to reach into any one or all of these hyperdimensions and affect the exact particle or component desired. We will be able to engineer matter—including the nucleus, and including inside the proton and neutron—in almost any way we wish. It all follows from scalar electromagnetics, which allows the direct engineering of hyperspace in all its dimensions.

Remember, engineering in hyperdimensions is just the engineering of nested levels of vacuum potentials. Even more startling, we can also place "thoughts" and effects directly in the mind itself, including the unconscious mind.

As we stated in the first edition of this book, we may model the mind as simply an objective 3-space, three turns away from the 3-space of the ordinary physical world. If we nest two or more additional zero-vector systems in our first one, then we have placed affects ("vectors") directly inside the mind itself. If we know the code of the human brain-tuner, we can now manipulate the mind directly. Electromagnetically, by scalar EM. And it appears that the Soviet Union has begun to reach this frightening capability.

Disease patterns may be induced upon ordinary EM carriers in the same manner, by nesting the zero-vector systems. That is, of course, if one has measured the scalar patterns evident in the induced disease, say, in the Kaznacheyev-type experiments with the desired disease patterns. The reverse is also true: Nesting the patterns in the proper context on EM field carriers can cure diseases such as cancer and leukemia, as evidenced by the Priore device.

One can also make a practical "free energy" device by setting up a standing scalar wave resonance in a localized region or device. From the "high stress" point to the "low stress" region of vacuum formed by such a standing scalar wave, a literal "river of energy" is flowing in the vacuum itself. Once the river is established, only the "leakage" need be replenished. Hence, if one builds a special sort of "paddle wheel" to react with the river and turn in it, one obtains shaft power freely from the localized curvature of spacetime. In that case, an ordinary generator can be driven by the shaft torque to provide very cheap, or"free," electrical energy.

Several inventors—three of whom I work with in one respect or another—now have demonstration models of such machines, which take diverse forms.

Because of the extreme importance of these emerging potentialities, in recent years I've concentrated intensely upon the properties of the vacuum as spacetime, and uncovered shortcomings (and downright errors in many cases) in ordinary physics and electromagnetics. Slowly the scalar EM picture has emerged, and a series of papers has been written and published to develop the concepts of the new approach.

The scalar approach also includes the ability to engineer negative energy and negative time flow. The reader is referred to Chapter 4, "Extraordinary Physics," of my forthcoming book, *AIDS: Soviet First Strike*, Tesla Book Co., P.O.Box 1649, Greenville, Texas 75401 for a summary of the vast importance of the time-reversed EM wave (phase conjugation).

Let me here just accent a rule of great importance: It is actually the key to negative energy, negentropy, antigravity, free energy, transmutation, Tesla's wireless transmission of energy at a distance without loss, Tesla's original magnifying transmitter, and many other things.

First we need the expression CPTEGS, where C is charge, P is parity, T is time flow, E is energy, G is gravity, and S is entropy. We shall be concerned with the algebraic sign of each term in the expressioa The rule deals with photon interaction with a charged particle inside an atom. Normally, incident photons interact with an electron in the electron shells, so C is normally negative. This produces normal parity, and causes the normal or positive flow of time, s T is normally positive. The normal kinds of energy, gravity, and entropy are experienced, so energy is positive, gravity is positive (attraction), and entropy is positive (progressive interactions in a system tend to more and more disorder).

The astonishingly powerful rule is this: In a photon interaction with a charged particle in an atom, if any one of the terms of the expression CPTEGS is reversed in algebraic sign, all the other terms are reversed in algebraic sign.

Thus if one makes scalar EM, the wave (and its photons) passes through the electron shells and interacts with (is absorbed in, and raises to an excited state, which then decays to emit a wave and photons) the highly nonlinear, positively charged nucleus. This represents a reversal of C, hence a reversal of P, T, E, G, and S also.

Consequently, a time-reversed wave—a phase conjugate replica of the incident wave—is emitted by the decaying excited nucleus. In nonlinear phase conjugate optics, this is the mechanism that produces the phase conjugate wave. This is also the mechanism that produces a very large amplification of the incident wave when a "pump wave" (two waves 180 degrees out of phase) are used in the nonlinear (modulating) medium of the phase conjugate mirror.

(The reader is referred to any good book or summary paper on phase conjugate optics; for example, see David M. Pepper, "Nonlinear optical phase conjugation," *Optical Engineering* 21(2), Mar./Apr. 1982, p. 156-183.).

When one pumps a phase conjugate mirror material sufficiently, then "tickles" the release of the total nuclear excitation energy as phase conjugation replica waves (negative energy), appreciable negative time is produced in the decaying nuclei of the mirror. In negative time, gravity is a repulsion, not an attractioa The mirror thus produces antigravity, and gets lighter. If it is forced to sufficiently high amplification, it can be made to float in air. At higher amplification still, it will accelerate upward.

Some other references of interest to this new edition follow:

1. **Robert A. Fisher, Ed., *Optical Phase Conjugation,* Academic Press, New York, 1983. Collection of papers. An extensive bibliography is given at the end of the book.**
2. **Ya. Zel'dovich, N. F. Pilipetsky, and V. V. Shkunov, *Principles of Phase Conjugation,* Vol. 42, Springer Series in Optical Sciences, Theodor Tamir, Ed., Springer-Verlag, New York, 1985.**
3. **Amnon Yariv, *Optical Electronics,* Third Edition. Holt, Rinehart and Winston: New York, 1985. See particularly Chapter 16: "Phase Conjugate Optics—Theory and Applications."**
4. **Y. Aharonov and D. Bohm, "Significance of Electromagnetic Potentials in the Quantum Theory." *Physical Review,* Second Series. 115(3), Aug. 1,1959. p. 485-491.**
5. **R. G. Chambers, "Shift of an electron interference pattern by enclosed magnetic flux." *Physical Review Letters.* 5(1), My 1,1960. p. 3.**
6. **Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics.* Volume n, Section 15-5, p. 15-8 to 15-14.**
7. **E. T. Whittaker, "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions." *Proc. Lond. Math. Soc.,* Vol. 1,1903. p. 367-372.**
8. **T. E. Bearden, *Solutions to Tesla's secrets and the Soviet Tesla weapons.* Tesla Book Co., Greenville, TX, 1981.**
9. **T. E. Bearden, *Tesla's secret and the Soviet Tesla weapons.* Presented in absentia at the First International Unorthodox Energy Symposium, Toronto, Canada, 1981. Available from Tesla Book Co.**
10. **T. E. Bearden, "Soviet weather engineering over North America." 1-hour videotape, 1985. Available from P.O. Box 1472, Huntsville, AL 35807.**
11. **Gwynne Roberts, "Witness to a super weapon?" *The Sunday Times,* London, England, Aug. 17,1980.**
12. **Greg Rippee, "Mushroom cloud sighted off Japan." *Los Angeles Daily News,* Apr. 11,1984. p. 1,8.**
13. **Martin Ebon, "Moscow, June 11, 1977." in his *Psychic Warfare: Threat or Illusion.* McGraw-Hill Book Co., New York, NY, 1983. p. 1-11.**
14. **A. K. Lapkovskii, "Relativistic kinematic equations and the theory of continuous media." *Soviet Physics Journal* 21(6), June 1978-Dec. 1978.**
15. **B. N. Belyaev, "On random fluctuations of the velocity of light in vacuum." *Azvestiya Vysshikh Uchebnykh Zavederdi, Fizika.* 11, Nov. 1980. p. 37-42.**
16. **Ingram Bloch and Horace Crater, "Lorentz-invariant potentials and the non-relativistic limit" *American Journal of Physics.* 49(1), Jan. 1981. p. 67-75.**

17. T. E. Bearden, *Comments on the new electromagnetics: Part I: discrepancies inpresent EM theory.* Tesla Book Co., Greenville, TX. 1982.
18. T. E. Beaiden, *Comments on the new electromagnetics: Part 2: the secret of electrical free energy. Tesla* Book Co., Greenville, TX. 1982.
19. T. E. Bearden, *Toward a new electromagnetics: Part III: clarifying the vector concept.* Tesla Book Co., Greenville, TX. 1983.
20. T. E. Bearden, *Toward a new electromagnetics: Part IV: vectors and mechanisms clarified.* Tesla Book Co., Greenville, TX. 1983.
21. T. E. Bearden, "USSR: New beam energy possible?" *Defense & Foreign Affairs Daily.* 13(111), June 12,1984. p. 1-2.
22. T. E. Bearden, *Star Wars Now! The Bohm-Aharonov effect, scalar interferometry, and Soviet weaponization.* Tesla Book Co., Greenville, TX. 1984.
23. T. E. Bearden, "A conditional criterion for identity, leading to a fourth law of logic." *Pursuit.* 13(1), Winter 1980. p. 6-10.
24. Max Frankel, "Khrushchev says Soviet will cut forces a third; *sees* 'fantastic' weapon." *New York Times.* Jan. 15,1960. p. 1.
25. David Bohm and B. J. Hiley. "On the intuitive understanding of nonlocality as implied by quantum theory." *Foundations of Physics.* 5(1), Mar. 1975. p. 93-109.
26. Derek Wood, "USA probes mystery cloud reported off *Japan." Jane's Defence Weekly.* 1(18). May 12,1984. p. 716.
27. Yossef Bodansky. "Soviets 'testing chemical agents in Afghanistan'." *Jane's Defence Weekly.* 1(13), Apr. 7,1984. p. 508.
28. "Explosive events seen on Soviet island." *Aviation Week & Space Technology.* Sep. 26,1983. p. 31.
29. Archibald E. Roberts, "The Tesla Factor." *Bulletin. Committee to Restore the Constitution,* POB 966, Fort Collins, CO 80522. Nov. 1984. p. 1-6.
30. See a series of articles by C. B. Baker in *Youth Action News,* POB 312, Alexandria, Virginia 22313. "Diabolic Soviet Warfare," Nov. 1984; "Soviet Weather Mayhem," Nov. 1983; "The Soviet Zapping of America," Dec. 1986; "Fifth Column Treachery," May 1987, and others.
31. D. W. McCann, "The Enhanced-V: a satellite observable severe storm signature." *Monthly Weather Review.* Ill, 1978. p. 887-894.
32. I. J. R. Aitchison, "Nothing's plenty: the vacuum in modem quantum field theory." *Contemporary Physics.* 26(4), 1985. p. 333-391.
33. G. E. Stedman, "Ring interferometric tests of classical and quantum gravity." *Contemporary Physics.* 26(4), 1985. p. 311-332.
34. Peter Hobbs, "Holes in clouds?" *Weatherwise.* 38(5), Oct 1985. p. 254-258.
35. T. D. Lee, *Particle physics and introduction to field theory.* Harwood Academic Publishers, New York, NY 1981.Second printing with corrections, Nov. 1982.
36. Ruggero Maria Santilli. *Ethical probe on Einstein's followers in the USA.: An insider's view.* Alpha Publishing, POB 82, Newtonville, MA 02160.1984.
37. Bryan G. Wallace, "The great speed of light in space coverup." *Scientific Ethics,* 1(1), Feb. 1985. p. 2-3.
38. Bryan G. Wallace, "Letters to the Editor." *Physics Today.* 36(1), 1983. p. 11,
39. Bryan G. Wallace, "Letters to the Editor." *Physics Today.* 37(6), 1984. p. 15.
40. M. E. Ash, I.I. Shapiro and W. B. Smith. *Astron. J.* 72,1967. p. 72.
41. Bryan G. Wallace. *Spectrosc. Lett.* 4,1971, p. 79.
42. Bryan G. Wallace. "The unified quantum electrodynamic ether." *Foundations of Physics.* 3, 1973. p. 381.
43. D. E. Thomsen, "Was a fifth force felt?" *ScienceNews.* 129(3), Jan. 18,1986. p. 38.
44. S. Olariu and I. Iovitzu Popescu. "The quantum effects of electromagnetic fluxes." *Reviews*

of Modern Physics. 57(2), Apr. 1985. p. 339-438. Thorough review of the entire subject of the Aharonov-Bohm effect, from the standpoint of conventional physics. A very comprehensive list of references is included.

45. V. I. Denisov and A. A. Logunov. "The inertial mass defined in the general theory of relativity has no physical meaning." *Teoretichedskaya i Matematicheskaya Fizika,* 51(2), May, 1982. p. 163-170. (in Russian). English translation in *Theoretical and Mathematical Physics*
46. A. A. Vlasov and V. I. Denisov. "Einstein's formula for gravitational radiation is not a consequence of the general theory of relativity." *Teoreticheskaya i Matematicheskaya Fizika.* 53(3), Dec. 1982. p. 406-418. (in Russian). English translation in *Theoretical and Mathematical Physics.*
47. Richard A. Blade, "Modification of Maxwell's equations in free space to account for nonzero photon mass and scalar electromagnetic waves." *Proceedings of the Tesla Centennial Symposium,* an IEEE Centennial Activity, Colorado College, Colorado Springs, Colorado, Aug. 9-11,1984, p. 91-92.
48. Jack Y. Dea. "Scalar fields: their prediction from classical electromagnetism and interpretation from quantum mechanics." *Proceedings of the Tesla Centennial Symposium,* 1984. p. 94-98.
49. Robert W. Bass. "Self-sustained non-Hertzian longitudinal wave oscillations as rigorous solutions of Maxwell's equations for electromagnetic radiation." *Proceedings of the Tesla Centennial Symposium,* 1984. p. 89-90.
50. T. E. Bearden. "Tesla's electromagnetics and its Soviet weaponizauon." *Proceedings of the Tesla Centennial Symposium,* 1984. p. 119-138.
51. Theodor Kaluza. *Sitz.* Berlin Preuss. Acad. Wiss. 966,1921.
52. O. Klein. *Z. Phys.* 37.895,1926.
53. Charles W. Misner, Kip S. Thome and John Archibald Wheeler. *Gravitation.* W. H. Freeman and Co., San Francisco, CA, 1973. p. 19-21, 71-72, 163-165, 367-369. Presents the limiting assumptions imposed on general relativity, to provide the severely restricted "ordinary general relativity" taught in almost all Western universities. The Soviets are not nearly so attached to such artificial assumptions.
54. V. I. Denisov and A. A. Logunov. "New theory of spacetime and gravitation." *Teoreticheskaya i Matematicheskaya Fizika,* 50(1), Jan. 1982. p. 3-76. (In Russian). English translation in *Theoretical and Mathematical Physics,* 50(1), July 1982. p. 1-48.
55. Raymond C. Gelinas. "Apparatus and method for determination of a receiving device relative to a transmitting device utilizing a curl-free magnetic vector potential field." U.S. Patent No. 4,447,779. May 8,1984.
56. Raymond C. Gelinas. "Apparatus and method for modulation of a curl-free magnetic vector potential field." U.S. Patent No. 4,429,288. Jan. 31,1984.
57. Raymond C. Gelinas. "Apparatus and method for demodulation of a curl-free magnetic vector potential." U.S. Patent No. 4,429280. Jan. 31,1984.
58. Raymond C. Gelinas. "Apparatus and method for transfer of information by means of a curt-free magnetic vector potential field." U.S. Patent No. 4,432,098. Feb. 14,1984.
59. Thomas E. Phipps, Jr. "Critique of accepted theories of special relativity and electromagnetism." *Scientific Ethics.* 1(2), Apr. 1985. p. 17-23.
60. Bertram Schwarzschild. "Currents in normal-metal rings exhibit Aharonov-Bohm effect" *Physics Today.* 39(1), Jan. 1986. p. 17-20.
61. William J. Hooper, "All-electric motional electric field generator," U.S. Patent No. 3,610,971. Oct 5,1971.
62. William J. Hooper, "Apparatus for generating motional electric field." U.S. Patent No. 3,656,013. Apr. 11,1972.

63. William J. Hooper, *New horizons in electric, magnetic and gravitational field theory.* Akron, Ohio. Electrodynamic Gravity, Inc. 1974.
64. Oleg D. Jefimenko. "Gravitation and antigravitation." *Proceedings of the West Virginia Academy ofScience.* 52(2), 1980. p. 106-118.
65. R. C. Jennison. "Relativistic phase-locked cavities as particle models." *Journal ofPhysics A: Mathematical and General.* 11(8), Aug. 1978. p. 1525-1533.
66. T. D. Keech and J. F. Corum. "A new derivation for the field of a time-varying charge in Einstein's theory." *International Journal of Theoretical Physics.* 20(1), 1981. p. 63-68.
67. Raymond Kromrey, "Electric generator." U.S. Patent no. 3,3374,376. Mar. 19,1968.
68. M. J. Perry, "Magnetic monopoles and the Kaluza-Klein theory," *Am. Inst. Phys. Corf. Proc.* No. 116,1984. p. 121-125. Kaluza-Klein theory admits soliton solutions which constitute magnetic monopoles. These monopoles have inertial mass, but no gravitational mass.
69. M. W. Kalinowski, "The nonsymmetric Kaluza-Klein meory," /. *Math. Phys.* 24(7), July 1983. p. 1835-1845.
70. D. J. Toms, "Induced Einstein-Maxwell action in Kaluza-Klein theory," *Phys. Lett.* B (Netherlands), 129B(l-2), Sept 15,1983. p. 31-35.
71. Fritz Albert Popp et al, eds., *Electromagnetic Bio-Information. Proceedings of the Symposium,* Marburg, September 5,1977. Urban & Schwarzenberg, Baltimore, 1979. See particularly Fritz Albeit Popp, "Photon storage in biological systems," p. 123-149.
72. A. G. Gurvitsch, *Theory of Biological Field,* Moscow, 1944 (in Russian).
73. V. P. Kaznacheyev et al, "Conditions Necessary for Appearance of Distant Intercellular Interactions after UF-radiation," *Bulleten Experimentalnoy Biologii i Meditsiny,* No. 5, 1979, pp. 468-471 (in Russian).
74. Antoine Priore, "Method of producing radiations for penetrating living cells," U.S. Patent No. 3,280,816. Oct. 25,1966.
75. David M. Rorvik, "Do the French Have a Cure for Cancer?", *Esquire,* July 1975, p. 110-111, 142-149. Excellent summary of the Priore Affair, with some details of the working of the machine.
76. J. B. Bateman, "A Biologically Active Combination of Modulated Magnetic and Microwave Fields: The Priore Machine," Office of Naval Research, London Report R-5-78, Aug. 1978. Comes very close when he states that "The possibility that some hitherto unrecognized feature of the radiation from a rotating plasma may be responsible for the Priore effects should not be dismissed out of hand..." Indeed, none of the persons who checked the machine or pondered over it knew anything about phase conjugation— which plasmas can do, and which the Priore machines did.
77. J. B. Bateman, "Microwave magic." *Office ofNaval Research London Conference Report,* ONRL C-14-77,1977.
78. R. Courrier, 1977. "Expose de M. le professeur R. Courrier secretaire perpetuel de 1'Academie des Sciences fait au corns d'une reunion a l'lnstitut sur les effets de la machine de M. A. Priore le 26 Avril 1977."
79. Jean-Michel Graille, *Dossier Priore: A New Pasteur Affair.* De Noel, Paris, France. 1984.
80. Thomas H. Moray, "Electrotherapeutic apparatus," U.S. Patent No. 2,400,707. Feb. 1,1949.
81. T. D. Lee, *Particle Physics and Introduction to Field Theory,* Harwood Academic Publishers, New York, 1981. See particularly Chapter 16, "Vacuum as the source of asymmetry."
82. Royal R. Rife, "The Rife Microscopes," Apr. 28,1939.
83. Arthur Isaac Kendall, "Filtration of bacteria; studies in bacterial metabolism," *Science 75,* Mar. 18,1932. p. 295-301.
84. R. E. Seidel, M.D. and M. Elizabeth Winter, "The new microscopes," *Journal of the Franklin Institute,* Feb., 1944. p. 103-129.
85. Christopher Bird, "What has become of the Rife *microscope?", New Age Journal,* Mar. 1976.

86. Lida H. Mattman, *Cell Wall Deficient Forms,* CRC Press, Cleveland, Ohio. 1974.
87. T. E. Bearden, *AIDS: Soviet First Strike,* Tesla Book Co., 1987. (In press).
88. T. E. Bearden, "AIDS: Urgent Comments on Mankind's Greatest Threat, and the Secrets of Electromagnetic Healing," presented in absentia at the U.S.P.A. Annual Symposium, July 30-Aug. 2,1987, Golden, Colorado.
89. 'Tesla Electromagnetics: Weapons or Healing?", presented at the Whole Life Expo, Los Angeles Airport Hilton Hotel, Los Angeles, California, Oct 3,1987.
90. Stefan T. Possony, "The Tesla Connection," *Defense & Foreign Affairs,* Aug. 1984, p. 12-14, 27. Highly recommended.
91. Daniel A. Walker et aL "Kaitoky Seamount and the Mystery cloud of 9 April, 1984," *Science,* 227(4584), Feb. 1, 1985, p. 607-611. See also Daniel L. McKenna and Daniel A. Walker, "Mystery Cloud: Additional Observations," *Science,* 234(4775), Oct 24,1986, p. 412-413 for update to the information in the first article.
92. T. E. Bearden, *Fer-de-Lance: A Briefing on Soviet Scalar Electromagnetic Weapons,* Tesla Book Co., Greenville, TX. 1986.
93. T. E. Bearden, "Soviet Phase Conjugate Directed Energy Weapons: Weapons That Use Time-Reversed Electromagnetic Waves," 1987 (in this publication).
94. Yossef Bodansky, "Soviets testing chemical agents in Afghanistan," *Jane's Defence Weekly,* 1(13), Apr. 7,1984. p. 508.
95. Draft agreement introduced by Andrei A. Gromyko to the United Nations' thirtieth session of the General Assembly on Sept. 23,1975; "Prohibition of the Development and Manufacture of New Types of Weapons of Mass Annihilation and of New Systems of Such Weapons." The first article provided that the types of these new weapons would be "specified through negotiations on the subject."
96. "Secret Speech: Did Brezhnev come Clean?", *National Review,* 29(8), Mar. 4,1977, p. 248, 250.
97. "Jetliner drops 32,000 feet; 400 aboard; 50 are injured," Associated Press Release, *Huntsville (Alabama) Times,Feb. 20,* 1985, p. 1. Also "China Airlines Pilot Denies Cockpit Error," AP Release, *Huntsville (Alabama) Times,* Feb. 21,1985, p. A-9.
98. *Asahi Evening News,* Tokyo, Japan, June 22,1982. Significant airline sighting by Japan Air Lines flights 403 and 421 of a giant, expanding globe of light on June 18,1982, in the North Pacific, 700 km. east of Kushiro.
99. T. E. Bearden, "Search for a Unified Theory of Electromagnetics and Gravitation: The Russians Got There First and Weaponized It," 1987 (in publication).
100. T. E. Bearden and Andrew Michrowski, Eds., *The Emerging Energy Science,* P.A.C.E., Ottawa, Canada, 1985.

T. E. Bearden
December, 1987

INTRODUCTION TO THE FIRST EDITION

Today there are a great many books available on UFO's, paranormal phenomena, and wild science, and one more such book ought to say something quite different if it is to be justified. Accordingly, in this book I have tried to say something quite different, perhaps in some cases too different. And so, to assist the reader's comprehension, I would like to explain something of how the book came about.

For the last fifteen years or so I have been interested in paranormal phenomena to one extent or another. Fifteen years ago I started studying Shotokan karate, and later switched to Yoseikan aikido, a most profound and absorbing martial art. I had already started serious reflection on the nature of perception, and after some years of aikido practice I occasionally began to experience what might be called paranormal perception. For example, occasionally, when violently attacked by several opponents, I would suddenly discover myself standing calmly on the sidelines watching the robot-like figures go through the motions of a play, motions that had already occurred in one kind of reality and that were now simply being repeated. On the rare occasions when I experienced this pure zen state of aikido, the techniques became flawless. In fact, the state was not separable into attacker, defender, and technique — rather a single union of the universe occurred. Unfortunately, although I eventually reached Sandan (third-degree black belt), I never permanently attained the complete aikido oneness state; this does not normally occur until the sixth dan level.

However, the occasional experiencing of this pure unity made a deep impression on me, for such aikido is pure meditation in motion, and only in such a state can the individual grasp the nature of perception. Only when conscious activity has stopped ("mind like a moon on still water") is such awareness possible. For several years I then laboriously developed a theory and interpretation of aikido for Westerners and eventually published it privately.

In 1971, while finishing a Master of Science degree in nuclear engineering at Georgia Tech, I put together a first crude theory of perception, which was published in 1973, many rewrites later. By that time I was already deeply engrossed in the study of paranormal phenomena, and especially unidentified flying objects. I had seen several UFO's (the earliest in 1950), and no longer doubted that the UFO phenomena represented some sort of new reality experienced by a great many people.

The UFO organizations at that time seemed to be in disagreement with each other. For twenty-five years or so almost everyone studying the phenomena had been trying to prove the extraterrestrial hypothesis, i.e., they started with the assumption that UFO phenomena were caused by, say, a group of young stalwarts over on planet X of star Y, getting beered up on a Sunday afternoon, hopping into an old space jalopy, and motoring over to planet Earth to jolly the natives. So they vigorously cataloged the UFO reports that fit the picture, and embarrassedly ignored those that did not. Yet most of the phenomena consisted of strange lights at night, not steel-hulled spaceships. And most of the contactee cases — particularly those involving apparent communication and abduction — were regarded as just too far-out for engineers and scientists to swallow. Almost everyone in the UFO organizations seemed to be convinced that Earth was being visited by spacemen from another planetary culture more advanced than our own, and so the UFO enthusiasts looked for hardware and spaceships, no more and no less.

The tabloids were a mess — filled with sensational nonsense about the purported UFO invasion of earth. Some invasion! If the would-be invaders possessed such an advanced technology, why had they spent twenty-five years flitting around the backside of the country and frightening the rural people, instead of simply landing an expeditionary force and having done with it? It seemed that almost no one was really looking at the phenomena in a totally unbiased manner. When one did so, without especially selecting the data, the strangeness of the data went up by orders of magnitude.

So I decided — rightly or wrongly — not to compete with the catalogers, but to try to come to grips with the possible *science* evidenced by the phenomena. I wanted to examine what kind of science was necessary to describe the phenomena and the mechanisms producing them. It rapidly became obvious that our scientific framework was not adequate to explain these events. In fact, our logical framework was not adequate. So I decided to look for a new scientific framework, knowing that the effort would be long and arduous and might not succeed at all.

All the unexplained phenomena seemed to me to be interconnected; for example, a study of Sasquatch will sooner or later lead to the study of UFO's and thence to the study of the other phenomena. So it appeared that they were all different aspects of the same overall reality. Since that reality most definitely involved the mind as well as matter, it appeared necessary to develop a new science incorporating both mind and matter. And so the search and the gradual accumulation of insight continued.

The trail led directly to the great paradoxes of science, the great unsolved problems of philosophy, and the three laws of logic on which all logical disciplines and sciences are founded. Using my perception theory, it was possible to see through the age-old necessity for the identity of opposites, how opposites could be identified, and when they were identified. Thus a fourth law of logic could be written and the system closed.

Slowly the nature (or, more precisely, a model of the nature) of mind became clearer. In addition, mechanisms for materialization and dematerialization appeared, as well as mechanisms for metal bending, thoughtform materializations (tulpas), and telepathy.

But the international strategic balance also could not be ignored. From my military background, the Soviet military buildup toward an attack capability was apparent as early as 1970-71. Moreover, from my work on the Moray radiant energy device I had discovered what I believed to be the basic theoretical mechanism used by T. Henry Moray to tap the zero point energy of free space. (This mechanism powerfully amplifies virtual state into observable state.) And positive evidence suggested that the Soviets had obtained the secret of how to make the Moray detector tubes. Putting this together with psychotronics produces an astounding weapon capability — and the Soviets must have had such capabilities already on site and others in development.

I looked for evidence of these weapons, and the evidence was there, cleverly — even diabolically — concealed.

When I then realized the *entity* nature of the collective human unconscious, everything fell into place. Suddenly the different major waves of UFO's made very good sense, as did the cattle mutilations, the Soviet buildup, Khrushchev's 1960 statement about fantastic weapons, Brezhnev's 1975 reference to the development of weapons more frightful than man had ever conceived, the long years of microwave radiation of the U.S. Embassy in Moscow, the Legionnaires' disease, etc.

The intelligence community had completely missed the fact that the Soviets had been working on a psychotronic Manhattan Project for thirty years. Not a soul in officialdom would believe that such capability could be achieved by surprise. This is not because of a deep conspiracy of the government (although I believe one of our high intelligence officials may be a Soviet agent, a sort of Penkovskiiy in reverse, who has been personally responsible for preventing any serious U.S. governmental research into psychotronics and UFO's); rather, it is because of all the personal arrogance, bias, power-seeking, manipulation, and maneuvering that now make up so much of our top managerial base in government.

So I decided to put out the information for all to see. There is still time — but just barely, I think — to build psychotronic defenses and prevent the launching of the Soviet attack. But we must move swiftly if Fer-de-Lance, the Soviet attack plan, is to be frustrated, for the eleventh hour is here.

Moreover, there is a single human problem responsible for all the ills that man causes man, and that problem has a technical resolution that is well within our grasp. I firmly believe that by psychotronics, the minds of all men can be linked into a great supermind having absolute psychokinetic mastery over time and space — literally a godlike being, in whom the collective human unconscious has a collective conscious instead of the four billion fragmented individual personal conscious minds it has at present. The initial psychotronics linkage mechanism is just a tool; after linkage is accomplished and stabilized, such props will not be necessary.

This is my real goal. Even if the U.S.S.R. launches Fer-de-Lance, annihilates NATO and China, and obliterates all Western retaliation with a stunning blow, and even if the collective unconscious then psychokinetically lashes back with the violent destruction of the Eurasian continent, linkage is still the goal and the final resolution. Should the apocalypse come to pass, there will still be time, in the midst of the horrendous collapse of civilization, to gain the final solution.

And it is the final and total solution. All the parables in the Book of Revelation will be realized. The grave literally will give up its dead — for with the overcoming of time and dimensionality, all that ever has been or ever will be, *is,* in one great nowness, allness, Omega, "I am that great I AM."

That is the goal, and it is toward that end that this book is written.

Please note that a listing of the research and reference works mentioned in the text can be found in the bibliography.

PART ONE

A SAMPLING OF SPECIFIC PARANORMAL PHENOMENA

In this section I show a sampling from the world of the paranormal, including UFO's of various kinds, psychotronics, Loch Ness monsters, Sasquatch, bigfoot, superfoot, psychokinesis, remote viewing, psychic photography, and a host of other phenomena— all explicable by the theories presented in Part Two.

I fervently believe that a theory can be constructed to explain our lives, including all the strange phenomena as well as the more familiar happenings. In this book a serious attempt has been made to point the way by developing what I believe to be the basic concepts, and by showing that a host of phenomena do seem to be consistent with the schema.

Let us hope that science and scientists are not after all dead, and that they will shake the cobwebs from their eyes and set about particularizing a new paradigm of mind and matter. When the top theoreticians and experimentalists wake up and join in the effort, my task will be finished.

THE VESTIGIA MYSTERY LIGHT EXPERIMENT

Let me start by introducing you to Vestigia, a very fine, relatively new organization that has been formed to study mysterious and unusual phenomena in a scientific manner. This organization, of which I am happy to be a member, is headed by Robert E. Jones and the investigation of the mystery light is spearheaded by James McGrath and Bill Wagner. (The informative *Vestigia Newsletter is pub*lished by the organization. Persons interested in the objective scientific study of unusual phenomena are cordially invited to contact Vestigia at P.O. Box 1183, Perth Amboy, NJ 08861.)

Vestigia's ongoing investigation of a mystery light in Morris County, New Jersey is a classic example of the manner in which mysterious phenomena may be studied in a scientific manner, and it should serve as a model for other investigative groups. Vestigia's experimental work has exemplified careful planning, close teamwork, proper scientific equipment and preparation, and the patience and persistence that are essential when performing fieldwork experiments in the field of unusual phenomena.

1 wish to stress in advance that Vestigia does not endorse my own theory of the possible nature of mystery lights; in fact, the organization very properly remains openminded. Also quite properly, they practice scientific conservatism in studying the possible nature of the phenomenon; at the present time they strongly favor the theory of the piezoelectric effect. Substantial evidence for this thesis includes the presence of a known geological fault in the local area, the presence of piezoelectric minerals in the local granite (e.g. thirty percent quartz), close association of the incidence of the phenomenon with subsequent local earth tremors, conditions favoring the accumulation of a surface charge (e.g., metal railroad tracks oriented parallel to the geological fault), and enhancement of the phenomenon under study. And this is exactly how Vestigia is in fact proceeding.

In science, the prevailing world view is not to be dismissed lightly since a great deal of experimental data has been gathered to substantiate it. And a single experiment cunnot establish the need to correct that solidly based view, even if the experiment at first appears puzzling and not in accordance with the accepted view. Only after many exhaustive studies have found the world view inadequate, may a paradigm properly be changed. From a rigorous scientific viewpoint, then, I am saying that this classic experiment is consistent with my own orthorotation theory, but not that the one experiment proves my theory. The experiment at this point is also perfectly consistent with the more conservative hypothesis that the mystery light is due to the piezoelectric effect. Additional different kinds of experiments are required before either theory will hold. For example, in my own theory, the mechanism of orthorotation must be verified elsewhere, and in the piezoelectric theory, the mechanism that spatially separates the effect from the crystalline material where such an effect normally occurs, must be explained and verified by other experiments.

With these precautionary remarks to the reader, let me now describe the beautiful experiment initiated and still being performed by Vestigia. I will then present my own theoretical explanation, again stressing to the reader that it must be treated as a hypothesis at this point.

Mystery lights occur in many locations throughout the United States and Canada. Vestigia has determined over ninety of them to date. They are recurring phenomena and are often seen near some natural or man-made feature such as a stream, hollow, road, railroad track, field, or mountain.

The mystery light under investigation by Vestigia is located along a one-mile stretch of railroad track in Washington Township, Morris County, New Jersey. The light has been seen by hundreds of people at one time or another over the last twenty to forty years. Dr. C. Louis Wiedemann describes it in the *Vestigia Newsletter:* "Appearing at night, it is said to bob or sway from side to side in the manner of a swinging lantern, and when approached it vanishes, sometimes reappearing at a more distant point."

Quoting from the *Vestigia Newsletter:*

> The preplanned battery of experiments and observations was conducted on the night of November 20-21, 1976. Three observation posts were established along one mile of railroad track, and these were manned in shifts by small teams of observers who maintained contact with team leaders via portable radio communicators. Communications were monitored and instructions were issued from a command post which was maintained in two vans joined in a T-configuration adjacent to the central observation post (Post 2).
>
> The vans contained electronic equipment valued at $30,000, which was wired to sensing devices along the railroad tracks. Four thousand feet of silver plated fine copper wire was laid between the rails and attached directly to an amplifier and oscilloscope in one of the vans, to detect variations in electrical frequency and amplitude. Other cables were attached directly to the rails, providing a capacitance test. Other devices included a Geiger counter, thermometers, a methane gas detector, a barometer, a parabolic listening device, a radar unit, an infrared sensor, and a sensitive compass. Personnel at observation posts were equipped with binoculars and cameras. One tripod-mounted camera equipped with infrared sensitive film was located near the rails at Post 2.
>
> James McGrath, a member of Vestigia's executive board, acted as advisor to the team. Team leader was electrical engineer William S. Wagner who was assisted by Jack Mazzuchelli. In all, sixteen persons were actual members of the team, and a total of twenty-one persons participated that night. A number of these were members of the Unexplained Phenomena Club of the County College of Morris, headed by the club's president, Steve Hopler, and vice-president Mary Ann Barron.

Shown here are six of the seven infrared photos of the mystery light which were taken by the Vestigia team. The frames shown are frames two, three, four, six,

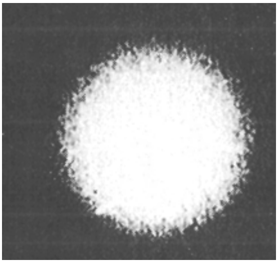

Frame Two

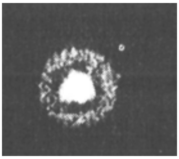

Frame Three

Photo 1. The Vestigia mystery light experiment.

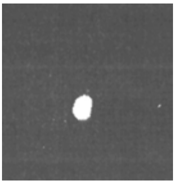

Frame Four

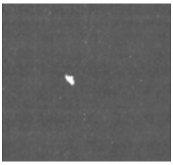

Frame Six

seven, and eight. The camera with infrared film was located at Post 2. Frame 1 was a test frame. Simultaneously with the recording of frame 2 at Post 2, observers looking from Post 1 reported visual observation of a light. The light was not seen by observers at Post 2. A second camera with ordinary film recorded only a pinpoint of visible light.

In the photograph, frame two recorded an image of an apparently plexiform spheroid of dynamic granules of light invisible to the eye of the cameraman. At the same time, observers from Post 1 looking back reported visual observation of a light.

In frame three, within about fifteen seconds the image has decreased in size and the plexiform spheroid is now a thinner corona around a dense, nearly hexagonal center.

In frame four, a dense but small irregular or oval image can now be seen. Frame five, not shown, is quite similar.

In frame six, a substantial decrease in size is apparent.

In frame seven, again the image size has increased, and again it possesses a dome central core surrounded by a thinner spheroidal corona of dynamic granules of light. Also, several separate satellite images are now visible at some distance from the main spheroid.

Frame eight is very similar, except that the satellites are no longer visible and the light in both the core and the corona is not quite as dense.

To describe the experiment precisely, the following passage is taken from Dr. C. Louis Wiedemann's article in *Vestigia Newsletter* No. 2, Spring 1977:

> Following several hours of equipment testing and calibration, at 10:21 P.M. two observers at Post 1 radioed the message that they had a "light" under visual observation between their position and Post 2, at an apparent distance of 300 to 400 feet from Post 1. Post 2 observers immediately scanned the area with binoculars and radar, but no object or light was detectible from that position. The yellowish-white light remained visible from Post 1 for approximately one minute and thirty-five seconds, initially approaching the tracks from a lower position off to one side, but soon positioning itself over the rails, and then seeming to sway from side to side in a pendulum-like manner. Its size was estimated between three and six inches in diameter, and for awhile the observers believed it was approaching them. When it vanished, the disappearance was sudden, as if the light had been "switched off."
>
> During the period in which the light was under observation, personnel at Post 2 took pictures of the apparent darkness, and continued to search with radar and binoculars. Two observers at Post 3 briefly saw what appeared to be an orange light between their position and Post 2. While the main light was under observation, several of the instruments in the command post gave unexplained readings. Most notable among these was the oscilloscope's display of an electrical frequency in the range of 40,000 Hertz. Additionally, the capacitance test showed the apparent presence of a conductive body between or over the rails. Nine minutes after the light had vanished from view, the Geiger counter (whose sensor was between Posts 2 and 3) indicated a full scale reading of greater than 5 mr per hour, lasting ten seconds. Normal background radiation was then recorded for five to six minutes (0.1-0.2 mr/hr), whereupon another full scale reading was displayed for approximately seven seconds.
>
> The pictures were enlarged from small images which appeared on the infrared film which was exposed through a 135 mm telephoto lens from Post 2 by Richard Larkin. Film employed was Kodak high speed infrared No. 2481, which is sensitive to infrared radiation to 900 nanometers, with maximum sensitivity from 750 to 840 nanometers. The lent opening was f/3.5 with the apera-

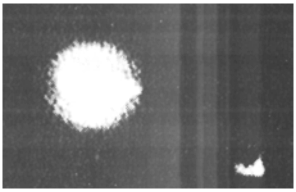

Frame Seven

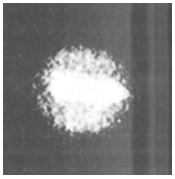

Frame Eight

> ture set at infinity. The second frame (frame 2) was exposed between 3 and 5 seconds, and the others were varied in exposure from 3 to 10 seconds. Intervals between exposures were between 10 and 15 seconds.
>
> High speed (non infrared) film in a camera which was hand-held by James McGrath at Post 2 also produced several images of an apparent point source of light above the rails (not shown here). Jack Mazzuchelli observed the area through binoculars during the entire time period, and he was unable to discern any light whatever, although he could perceive the rails and the railbed in the faint illumination provided by starlight shining through a partially overcast sky.

As I stated previously, Vestigia has not yet reached a firm conclusion as to the causative agent, but strongly favors the piezoelectric theory, since if offers the simplest explanation at this point. Again, to quote the *Vestigia Newsletter:*

> No absolute conclusions can yet be drawn, other than the apparent fact that a phenomenon or set of phenomena did occur. Team leader Wagner theorizes that the phenomenon might be similar to so-called earthquake lights which have been reported to precede seismic tremors in Japan. Interestingly, there were explosionlike shocks or "loud booms" reported by residents of two housing developments several miles distant from the site on December 5 and 6,1976, and these shocks were recorded by seismographs at Columbia University and elsewhere. Intensity was (-)0.25 to (+)0.38 on the Richter scale. Wagner posulates a possible piezoelectric effect from tension on subterranean rock deposits.
>
> The possible relationship of a piezoelectric effect will be investigated further. Data gathered in Japan tends to indicate the development of subterranean stresses during the passage of a cold front. In the light of this possibility, it is interesting to note that there was a drop in barometric pressure from 29.9 to 29.76 culminating less than one hour before Vestigia recorded the start of the phenomena. A light snowfall preceded the appearance of the light.
>
> Wagner believes that the initial discoveries tend to indicate that the following conditions must be present for the phenomenon to be manifested:
> (1) A geologic fault in the local area.
> (2) The presence of certain minerals, especially quartz (quartz comprises 30% of the granite in the area under investigation).
> (3) Conditions should permit the accumulation of a surface charge (e.g.: the presence of railroad tracks, because they are metallic).
> (4) The tracks must run parallel to the fault.
> (5) The phenomenon's manifestation tenc
> sure changes as well as extreme changes in temperature.
>
> Wagner states, "Our indications are, at this point, that the lights precede earthquakes or tremors."

Regardless of what hypothesis we begin with, it must be consistent with all the major features of the Vestigia mystery light experiment.

(1) The explanation must show why a light with structured patterns formed and changed. The structured patterns changed geometrically in shape and intensity, as well as in movement and location. This dynamic formation and change was recorded on infrared-sensitive film. The geometrical shapes of the light structures must be accounted for.

(2) Ordinary film recorded only a pinpoint source of light.

(3) The light which was photographed in infrared was not seen by the photographer himself, nor by eight other persons who were observing at the time from Post 2.

(4) Either that light or another was seen visually from a different station.

(5) At the same time as the visual appearance of the light, instruments recorded the apparent presence of a conductive body between the rails or over them.

Specifically, the electrical capacitance between the rails was observed to change — or at least indications to that effect were received on the instruments.

(6) Nine minutes after the main light vanished, a Geiger counter began to read full scale for about ten seconds, then lapsed back to background reading. Five or six minutes later, full scale reading was repeated for about fifteen seconds. About a minute after that, full scale reading was again recorded for about seven seconds.

(7) At the same time as the main light visual sighting, an electrical frequency of approximately 40,000 Hertz was recorded on the oscilloscope.

There were also environmental conditions which may well have been involved in the experimental effects.

(8) Barometric pressure dropped from 29.9 to 29.76, culminating less than an hour before the phenomena were recorded.

(9) There was a temperature drop and a light snowfall preceded the appearance of the light.

(10) The geological conditions strongly favored seismic stress and piezoelectric effects. For example, the local granite contains about 30 percent quartz, and there is a known geological fault not far from the experimental site. About fifteen days after the experiment, seismic shocks were indeed recorded at Columbia University and elsewhere, and they were audibly detected by residents of two local housing developments. The measured intensity of the seismic shocks was from(—)0.25 to (+)0.38 on the Richter scale. It thus appears highly probable that seismic stresses existed on the subterranean rocks at the time of the experiment and that piezoelectric activity was present in the local granite quartz due to these stresses.

(11) Conditions also favored the accumulation of surface charge, and the railroad rails (good conductors) were oriented parallel to the geological fault.

(12) Several thousand feet of wiring existed in the immediate area to connect the different stations and the instruments to the sensors. About four thousand feet of fine copper wire, silverplated, was laid between the rails and fed to an amplifier and an oscilloscope to detect variations in electrical frequency. The capacitance test also used cables attached to the railroad rails themselves. The Geiger counter's sensor was also positioned at the experiment's hot spot and wired to the equipment van. Additional equipment and wiring was also used.

Finally, preliminary surveys of the area were carried out by persons of known paranormal sensitivity several days before the main experiment. Before conducting the major experiment, Vestigia interviewed persons who had seen the light and also conducted preliminary site studies using psychics. One nationally known psychic was brought to the site by Vestigia's president Robert Jones, and asked to walk along the tracks. The psychic reported impressions of some personal tragedy, perhaps a gruesome murder. Later several other persons with heightened psychic sensitivity were brought to the site, and they reported psychic impressions.

In all, three psychics were brought to the track. The only directly correctable result was that all three had a reading in the same area on the stretch of track, within sixty feet of the sighting of the night of November 20-21. The psychics' own interpretation of their readings differed considerably.

UNEXPLAINED GENERATION OF ELECTRICITY

Some additional data exist to indicate the generation of electricity by rocks and the generation of unexplained electromagnetic energy radiated from the earth, ocean, or atmosphere.

For several years T. Townsend Brown has worked in electrogravitics and holds many patents in the field. Some of his laboratory experiments have positively demonstrated a direct connection between electric currents, electric charge, and gravity, although the exact nature of this connection has not yet been fully explained. Nevertheless, it can be shown that highly charged objects do not behave in a gravitational field exactly as predicted by classical theory.

Furthermore, in 1974 Santilli succeeded in proving that one of the classic assumptions of physics is wrong, namely that electric field and gravitational field are mutually exclusive. Santilli proved that gravitational field and electric field are either partially or totally the same thing, in contradiction to present assumptions.

Over a decade ago, Curtis reported the measurement of very low frequency electromagnetic signals generated over the ocean that are not accounted for by any presently known mechanism or theory. Although Curtis published several articles on this effect, no further investigation has appeared in the literature.

Finally, there is a certain location where the generation of an anomalous electromagnetic signal occurs continuously. This is McGregor Missile Range (U.S. Army), where Army units fire air defense missiles such as Nike Hercules and Hawk. As a young captain, in 1961 I headed up the Hawk Evaluation Team at McGregor Range and directed the Hawk missile firing. About sixty-five kilometers out in front of the Hawk missile sites and downrange, there was a most peculiar source of electromagnetic noise, which could considerably affect a Hawk illuminator (continuous wave) radar when the radar antenna was oriented in that direction at a low elevation. In that case the radar would continually noise-lock (make and break lock-on) on the noise signal, dithering slightly as it continued to try to resolve a coherent radar return signal. This noise source was always present, in all seasons and temperatures, and so we simply made each Hawk unit commander and his personnel aware of the phenomenon. On one occasion 1 even vectored a helicopter with an investigation team to descend on the site and inspect it. All they found was the usual surface of the desert floor, with sparse growths of greasewood and cactus. There was nothing which could serve as a generator or reflector.

At one of our other missile sites in the U.S., there was a strange reflector mechanism located at some distance from the site and at about 15,000 feet altitude. **This** was rigidly fixed in location and unvarying, and the unit radar personnel often used it to calibrate the range of the radar. They simply locked on the reflecting point and fine-adjusted the radar range dials to indicate exactly the known range to the reflecting point. Here again, personnel on their own initiative vectored an aircraft directly through the point and searched the air in that area, but nothing at all could be found.

HYPERFIELD GENERATION OF PHOTON PATTERNS ABOUT A BAR MAGNET COLLECTOR

The magnetic field has more fundamental components which integrate in the virtual state to form it when two successive orthorotational thresholds (quantum thresholds) are breached. Furthermore, in the picture of a bar magnet taped to the

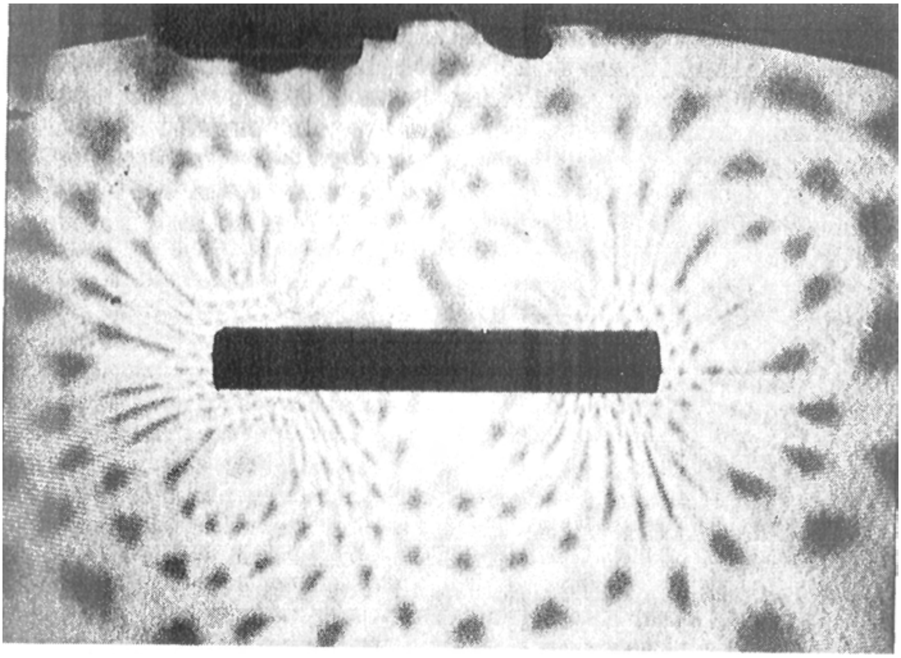

Magnet 3/4" dia. x 6" long taped to faco of a color TV tube approximately 1200 Gauss.

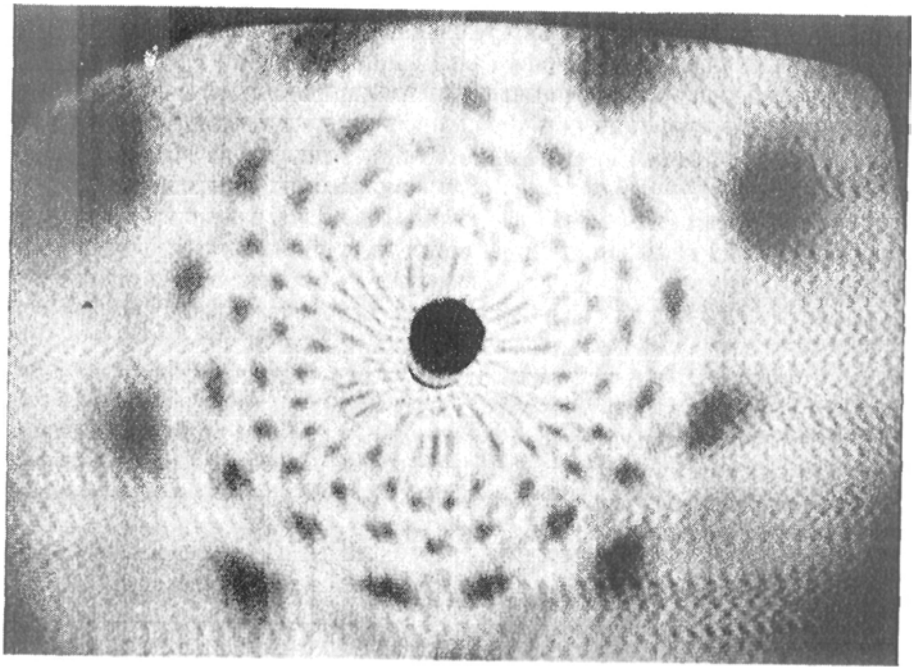

South pole end of same magnet tapod to face of color TV tube.

Photo 2. Hyperfield **flux in a magnetic field.**

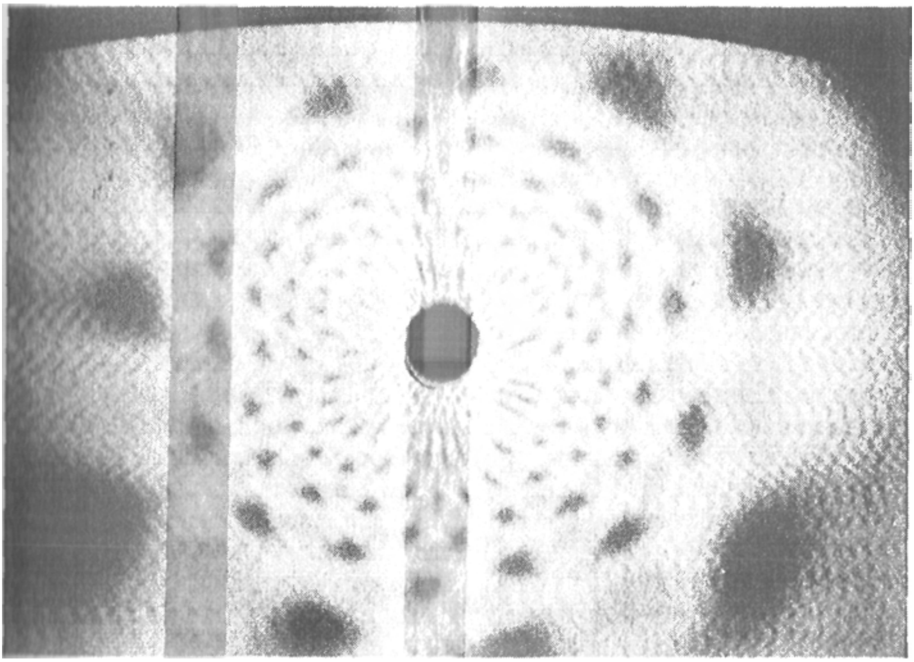

North pole end of same magnet taped to face of color TV tube.

face of a color television tube, one can see that the hyperfield orthogonal circulations in the magnetic field arc not symmetric about the two poles; i.e., the pattern is quite different around the north magnetic pole from that around the south magnetic pole.

Around the north pole, the hyperfield circulation is emitted in the form of a hexagonal solid. Thus there is an intense hexagonal orthorotation about the north pole. Radiating conically out longitudinally is a toroidal null zone, at an angle to the ongitudinal axis projected away from the north pole into space. At an even higher angle, another concentric cone is radiated out in the same direction; this higher cone is an emission cone where photons are being generated. Thus viewed from the north end, and at a distance from the magnet, the pattern shown in the right figure of the slide emerges, The pattern is a ring of light or light bubbles, surrounding a concentric ring neutral zone, which in turn surrounds a hexagonal center of intense emission.

If frame three of the Vestigia mystery light photos is examined meticulously, this exact pattern is seen, allowing for the dynamism of the phenomena and for slight nonlinearity of the spatial medium in which the photons are emerging by orthorotation.

On the other hand, looking at the left hand figure on the slide, we see that the outer emission ring is *so* thin that it has effectively disappeared; it is radiated at right angles to the magnetic longitudinal axle. The middle null area is also much wider. The densely emitting core area now has a relatively circular cross section at a distance. So the pattern of photon emission that is seen from this end is simply a dense round core. Frames four and five of the Vestigia photos demonstrate variations of this effect, viewed at a slight off-axis angle.

When the magnetic field is observed from the side, only the hexagonal dense emission core of the north pole is seen. In the south pole area, we are looking into a large null area, so no emission at all is seen. Frame eight is approaching this result, with a slight deviation from an orthogonal viewing axis. Also, frames seven and eight exhibit composite structuring from several (three or more) side-by-side magnets of square-end configuration, where the magnet cross section is wide, looking at the end. In that case, the top right picture would be elongated, giving a stretched hexagonal dense core surrounded by a concentric null zone of oval shape, further surrounded by a concentric thin emission zone of oval shape. In frames seven and eight, the outer thin emission zone is roughly spherical, not oval. However, a combination (further composite) of the lower right picture case and the upper right picture case could give the pattern exhibited by frames seven and eight.

And as I show clearly elsewhere, orthorotated hyperfield energy usually emerges with a "burst-in" effect which, in free space, assumes a spherical shape. Thus one would expect a dynamically bubbling sphere of light at the initial "burst-in" of the phenomena, and frame two exhibits this effect.

So the expected hyperfield generation of photon patterns about a localized region of magnetic field (such as a bar magnet or composite grouping of bar magnets), when that localized region of field is in the collection or burst-in and decay gradually mode, would account precisely for the particular structured patterns photographed by Vestigia.

Again, this does not prove the hyperfield hypothesis, but shows that the structural patterns obtained are consistent with it.

HYPERFIELD CIRCULATION PATTERNS ABOVE A MAGNETIZED RAILROAD TRACK

Figure 1 shows the hyperfield circulation patterns above a magnetized railroad track. The general setup of the Vestigia experiment is shown. The rails are oriented parallel to the fault line, which means that the rock stresses are oriented perpendicular to the rails. Thus the telluric currents piezoelectrically induced by the quartz and other minerals in the local granite are also generally oriented perpendicular to the rails and below them.

These piezoelectric currents thus constitute effective coils about the rails, magnetizing them weakly along their length, but this is in addition to their normal magnetism, if any, picked up due to the specific local and atmospheric electricity changes.

In the hyperfield flow pattern for two magnets which are lengthwise parellel and viewed from the end, a primary hyperfield circulation pattern exists both between the two rails and in the air above them. The location is the same, whichever end of the parallel magnets one looks into; however, the circulation direction changes with the change of ends. As the strength of the magnetization of the two magnets changes, the vertical distance above the rails to the plane of the center of the hyperfield circulation pattern (primary) changes. The stronger the field, the higher the center of the pattern, and vice versa.

In addition a secondary (weaker) hyperfield circulation area exists over both rails as shown. As the piezoelectric currents fluctuate in direction and intensity, corresponding fluctuations in the magnetization of the two rails occur.

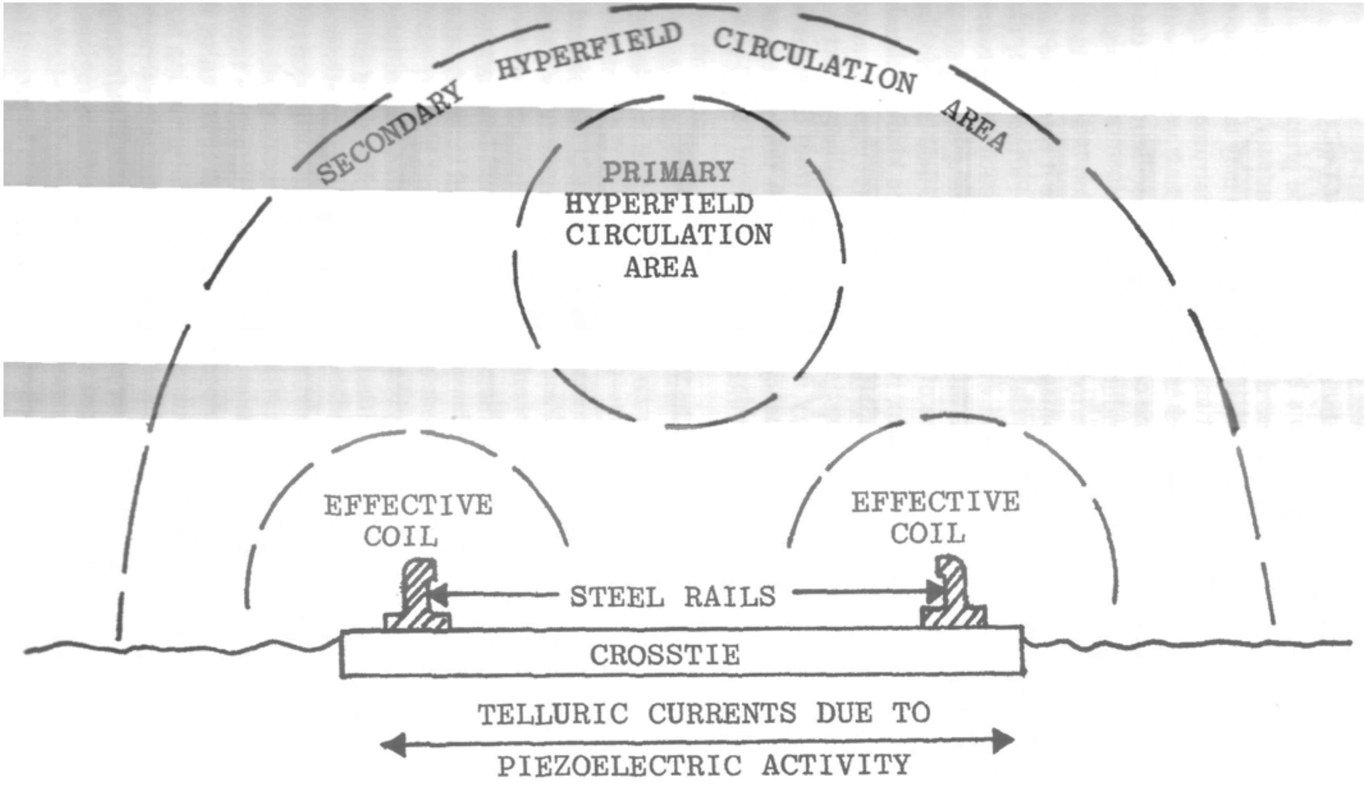
SECONDARY HYPERFIELD CIRCULATION AREA
PRIMARY HYPERFIELD CIRCULATION AREA
EFFECTIVE COIL
EFFECTIVE COIL
STEEL RAILS
CROSSTIE
TELLURIC CURRENTS DUE TO
PIEZOELECTRIC ACTIVITY

Fig. 1. Hyperfield circulation patterns above a magnetized railroad track

Thus, over the rails there exists a sort of undulating tubular tunnel of primary hyperfield circulation, surrounded by a larger undulating tunnel of half-elliptical cross section. In addition, the terrain and foliage features nearby will have their own hyperfield circulation areas, with flowing currents of hyperfield circulations between them and undulating tunnels over the rails. The final result is a complex flow pattern, which is further modified by atmospheric, ionic, and electrical currents. So a mystery light can kindle into a burst in one of the primary circulation areas and move, usually in a faltering or drift-dart-pause manner, along the undulating tunnels above the railroad track. Structural changes may then be noted as the geometrical aspect between the light and the observer-instrument changes, as additional collection-kindling occurs, and as discharge dampening occurs with the passage of time, due to the discharge time constants of the oscillations of the time dimension itself that are induced by kindling and burst-in.

Note that in the kindling process, the hyperfield circulation in a localized area suddenly bursts into magnetic or electromagnetic field. Since it enters the laboratory frame at right angles, orthogonal relationships result. Furthermore, the appearance of a kindled magnetic field form will change the capacitance between the rails, and a capacitance test will indicate this, as if a conducting body were there. The localized magnet or geometrical piece of electromagnetic field is locked in and maintained as long as the hyperfield flux remains above the kindling threshold. However, since the burst-in of the hyperspatial flux into electromagnetic energy involves the shocking into oscillation of the time channel, an oscillator effect is produced, and electromagnetic signal energy may be radiated from the light form, usually through one of the magic frequency windows which thinly interconnect orthogonal three-spaces. One of these windows is located at about 38,000 to 40,000 Hertz.

Such a signal was indeed recorded in the Vestigia experiment.

Also, at night time the time channel quenching effect of visible light is materially reduced since most visible light is absent. Further, the light-shielded quartz in the underground rocks has two magic windows: one in the infrared and one in the ultraviolet. So the area of the Vestigia experiment constitutes a gigantic kindling mechanism par excellence. With sufficient piezoelectric activity, the collected hyperspatial flux exceeds the kindling threshold required for burst-in, and localized triggering of a discharge or discharges can be initiated in the proper area. The thousands of feet of wire further enhance the collecting and kindling mechanisms. From a physics viewpoint, the living observer is the initiator who initiates the collapse of the wave function, in turn initiating observation. The same ability or property also is required to initiate the kindling of hyperspatial flux into the normal three-space, in my opinion. This is consistent with the rather wide collection of data presented here to show that a living observer can generate or extinguish electrostatic, magnetic, and electromagneic fields at a distance, even through a superconductive farraday shield. The thought photography examples presented also show that electromagnetic field (light) can sometimes be generated by a living human being, and impress the camera film with the visible light spectrum, even though the light is not consciously seen by the observer's eye. It also appears that frequencies of something other than light, in the infrared and ultraviolet frequency regions, can enter through IR/UV windows and then rotate into the ordinary visible light spectrum and ordinary visible light energy. Such effects also appear to be present in the

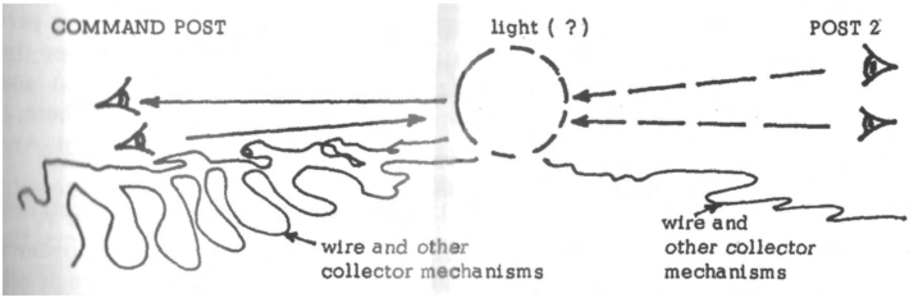

Fig. 2. The cyborg effect and orthorotational sensitivity

Vestigia experiment, since the cameraman himself saw no light, but other posts did. Thus an additional directional effect of the rotation into visible light seemed to be involved.

Since so-called thought energy is simply hyperspatial flux which has a minute one-to-one coherence in its crosstalk with the laboratory frame, the section explaining thought photography deals with hyperspatial mechanisms that are also applicable to mystery lights.

THE CYBORG EFFECT AND ORTHOROTATIONAL SENSITIVITY

In Figure 2, I show how it is possible that one team sees the light and a second team does not.

In the inceptive cyborg effect, each person, each instrument, the wiring, other collectors including the sensors, the other paraphernalia of the environment such as railroad tracks, rocks and their minerals, etc. are all cyborged together into a single instrument in the hyperfield domain. This is precisely because the hyperfield domain is the time channel, and all of these exist at the same time. Further, in the experimental locality, the difference in time between observers, equipment, devices, and aspects is small, because the distances between pieces are relatively small. Hence to the hyperfield channel an appreciable degree of coherence exists between the various components of the hypersensory cyborg. Yet the overall pattern of coherence has high points and lower points in it, and pattern groupings that vary with the grouping of components. Thus the subgrouping at the command post, for example, may be expected to differ in its hyperfield characteristics from the subgrouping at Post 2. Furthermore, the bulk of the collection wiring, and the rails, were outside the command post, so the greatest collection might be expected at this station.

In addition, the human components of the subgroupings vary appreciably in their sensitivity to hyperfields. Some persons are more positive or attuned than others, particularly in their unconscious minds. In fact, some persons are quite negative and are squelchers of the hyperfield effects — the negative psi effect is well known in experimental parapsychology. However, it is reasonable to assume that there were no psi-negatives on the Vestigia experiment; but there was probably an appreciable variation in the degree of psi-positive exhibition in the various personal unconsciousnesses. So it is understandable that one post may not see the light, while another would. And with the appreciably greater collection and kindling outside the command post, we would expect the sighting to occur there, if anywhere, as long as a real psi-negative was not present in the command post. And the experiment seems to have exhibited this effect as expected.

Also, I must point out that the visual mechanism for detecting orthorotating energy is different from that for detecting normal visible light, unless the orthorotating hyperenergy is extremely strong and overrides the squelch effect of normal light. (At night the degree required to go into visible light is less.) At any rate, the orthorotational hyperfield energy comes in through the infrared and ultraviolet windows for the most part, and then orthorotates into the visible light region, changing its frequency as it is absorbed by or interacts with a receiver.

Further, the orthorotational energy does not travel through three-dimensional space in the accepted sense that a signal normally does; rather, it is entering this three-dimensional space all at once from a region orthogonal to the entire three-space. It is therefore entering all components of the tuning-receiving instantaneously and simultaneously, or holographically, if one extends the term. Each subgrouping component of the cyborg exerts its own tuning-damping constant(s) upon this holographic signal entry, giving differing degrees of receptive sensitivity, or signal strength, or signal to noise ratio.

So the signal enters each part of the cyborg holographically and the reception characteristics of that portion dampen or enhance the degree of orthorotation into the normal electromagnetic field. Further, the actual orthorotation into the electromagnetic field is occurring primarily at the actual sensor.

For a human being, the primary sensors are the retinal surface of the eye and the skin. The retina is actually a complex structured piece of the brain, so to speak, and extensive nervous system complexes are located there which do extensive on-site signal processing. The unconscious mind of the individual directly interacts at the retina with signals received by the retina, either by hyperspatial orthorotation or by ordinary light. So the individual nervous system is highly variable in its interactions with the hyperfield energy that is orthorotating into ordinary electromagnetic energy in the interaction itself. Briefly, the unconscious state of the observer partially shapes and forms the observed.

In a camera, the sensor is the film. It is thus entirely possible for a camera to record light energy that is holographically rotated in from hyperfield energy without the cameraman himself seeing anything at all. This is especially true when the camera film is sensitive to infrared and ultraviolet to an appreciable degree. A camera containing infrared-sensitive film will interact readily with hyperfield energy because it is open in the magic window between orthogonal frames. A camera containing ordinary film, which is less sensitive to infrared, will interact much less in the magic window because it barely extends into it.

Thus, with other variables held constant, one would expect the infrared camera to record hyperfield energy to a much greater degree than the camera containing ordinary film, and indeed the Vestigia experiment has demonstrated this effect.

Furthermore, the experiment exhibited heightened kindling and collection outside the command post due to the distribution of wiring, components, and rails.

ADDITIONAL FACTORS CONSISTENT WITH THE VESTIGIA EXPERIMENT

First, a cooling effect is often associated with the appearance of paranormal phenomena; this effect is well documented in the literature.

A corollary effect also exists when hyperfield energy is being collected and kindled. Just as lowering the electronic noise in an amplifier increases the fidelity of the signal, lowering the local environmental temperature reduces the "noise" that is blocking the collection and kindling of hyperfield energy. Specifically, reducing the temperature reduces the photon emission and absorption activity of the environment. This lowers the background electromagnetic noise which inhibits orthorotation of the hyperfield energy into electromagnetic field energy. Since photon activity is reduced, the consequent photon squelching of the time channel (precisely detailed elsewhere in this book) is also reduced. This in turn lowers the threshold required for kindling to occur.

Also, a Geiger tube often acts as a collector and amplifier of bioenergy and hyperfield energy. (The association of Geiger counter readings with UFO phenomena has been frequently noted in the literature.)

The Geiger tube may become a detector of hyperfield energy, because the tube's kindling threshold may be significantly lower than that of the environment or of other components such as the observer's eye. Kindling across the Geiger tube's lower threshold simply creates an electrostatic field inside the gas in the tube. In fact, it is the potential ionic current that is the collector, since hyperfields accumulate along probability paths as well as along observationally determined paths, communicating with both future and past. At any rate, the kindling of an electrostatic field in the Geiger tube may result in ionization of the gas if the field is of sufficient intensity. An avalanche discharge may occur and the Geiger counter may indicate a reading (usually appreciable or full scale). This may occur even though no visual phenomena are seen and even though other sensors receive no indication. This reading of the Geiger counter is of course not due to radioactivity. A Geiger *counter does not indicate radioactivity;* rather it indicates appropriate ionization of its internal gas. In this case, the kindling of hyperfield energy into electrostatic field occurred, generating the ionization in the Geiger tube gas.

Note that this occurred in consonance with the visual sighting and the infrared film photos, and when no visual sighting resulted and photos were not taken. Furthermore, as I point out elsewhere, the kindling and burst-in of energy has a discharge time constant, due to the discharge time constant of the oscillation of time that occurs at kindling and burst-in. Depending on the initial amplitude of oscillation at burst-in, the time required to lower the burst amplitude to kindling threshold and below varies. When the amplitude drops below the kindling threshold, kindling and burst-in cease and the hyperfield-induced electrostatic field in the Geiger tube vanishes, stopping the hyperfield-induced ionization of the gas in the tube. The Geiger counter reading then drops back to normal background reading.

The variable amplitude at burst-in and the consequent variation in the time required for the Geiger counter to cease reading the hyperfield-induced ionization were demonstrated in the Vestigia experiment. Furthermore, the Geiger counter showed a reading twice when visual phenomena were not seen and cameras did not record. Of course, no photos were taken during the Geiger counter readings.

Finally, the drop in barometric pressure also lowers the molecular collision activity of the local atmosphere, and lowers the rate of production of collision ions, electric currents, and photons induced by the collision activity. This reduces the electromagnetic noise, which is squelching the collection mechanism and raising the kindling threshold. It reduces the kindling threshold, making orthorotation of hyperfield energy into electromagnetic-electrostatic-magnetic field energy easier.

Thus, according to hyperfield theory, the major features of the Vestigia mystery light experiment are seen to be within normal expectations. Again, this does not prove the hyperfield-orthorotational hypothesis, but it is consistent with it and therefore constitutes an important piece of evidence for the theory.

Also, I should stress that the piezoelectric theory is in fact included in the hyperfield theory, and is a vital part of it. In other words, the hyperfield theory is an extension of the piezoelectric theory, and in my opinion the extension is necessary if the piezoelectric effect is to fit the experimental results. There are no contradictions between the two theories since one is merely an extension of the other.

REMOTE VIEWING

In a classic series of carefully controlled experiments at SRI International, physicists Hal Puthoff and Russell Targ have shown that the human mind can produce evidence of its ability to see at a distance, and also to anticipate the future.

In their protocol one experimenter goes to a distant target, which is selected randomly for him en route, and a subject in the laboratory tries to describe the target scene, both orally and by sketching on paper. The experimenter in the field uses a random number generator to select one sealed envelope out of a group held by a controller accompanying him. The selected envelope is then opened, and that target becomes the target for the experiment. The targets are preselected with rigorous protocol by a group not involved in the experiment.

In the experiment shown here, the target selected was a rather unusually shaped playground merry-go-round, with handrails made of bent pipes, as shown in the two photographs. The laboratory subject, in addition to describing the scene verbally, made the sketches shown at the right. As can be seen, the channel is noisy, but there is definitely signal in it.

Puthoff and Targ also obtained some of the best experimental results when the laboratory subject was asked to describe the scene which the field team would see in the future, the target selection via random number generator not yet having been made. A wide variety of subjects — including persons hostile to parapsychology — were used by Puthoff and Targ in their experiments, with good results.

Each particle of mass in our bodies represents one closure of the entire universe — yielding a holographic reality — and deeper communication with ourselves is identical to communication with the universe, including any part of it, at any distance. Furthermore, in hyperspace the future and the past are all present. Since a particle does indeed exhibit a four-dimensional component for 1/137 of the time, each particle does connect to the future and to the past. With selective tuning

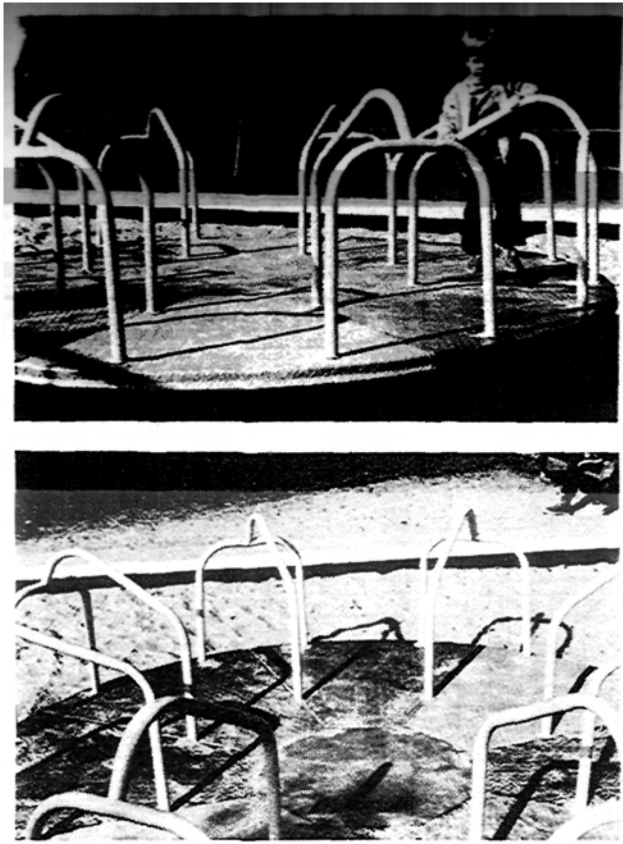

Target: Playground Merry-Go-Round.

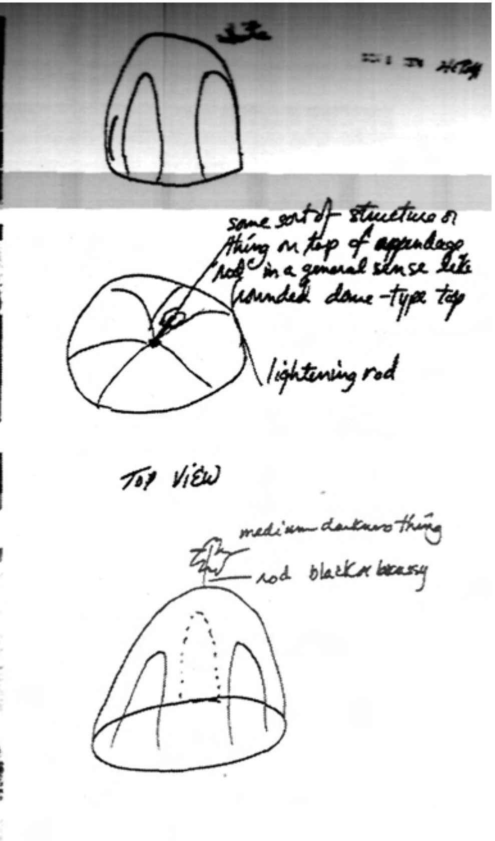
some sort of structure or
thing on top of appendage
rod in a general sense like
rounded dome-type top
lightening rod
TOP VIEW
medium darkness thing
rod black or brassy

Responses.

Photo 3. Remote viewing.

and kindling any part of this holographic reality is accessible. However, because of the smallness of a single selective signal in the midst of the totality, the channel is quite noisy. For this reason skilled psychics ~ persons who have been found to have a greater fidelity of selective tuning - can be expected to produce better results than the normal person.

Puthoff and Targ's results have also been successfully repeated by other experimenters.

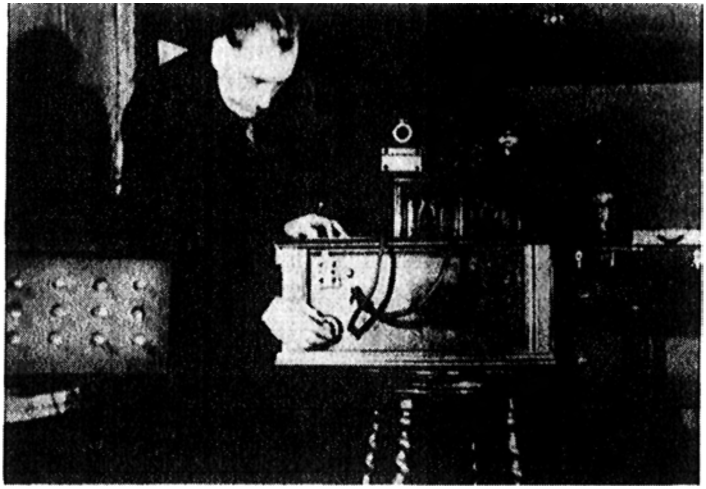

T. Henry Moray makes some final adjustmonts In tuning his radiant energy device to tap zero-point vacuum energy.

THE MORAY RADIANT ENERGY DEVICE

In the early 1900's, Dr. T. Henry Moray of Salt Lake City produced his first device to tap energy from the metafrequency oscillations of empty space itself. Eventually Moray was able to produce a free energy device weighing sixty pounds and producing 50,000 watts of electricity for several hours. Ironically, although he demonstrated his device repeatedly to scientists and engineers, Moray was unable to obtain funding to develop the device further into a useable power station that would furnish electrical power on a mass scale.

As a boy, Moray had been deeply inspired by the greatest electrical genius of all time, Nikola Tesla. His imagination was especially fired by Tesla's claims to have knowledge of an energy source greater than ordinary electricity, and by Tesla's emphasis on frequencies as the stuff of the universe. When Moray finished high school in Salt Lake City, he went abroad to study, and took resident examinations for his doctorate in electrical engineering from the University of Uppsala, Sweden, during the period 1912—14. Returning home, his diploma and credentials were interrupted by World War I, and the University mailed him these items in I918 after the war.

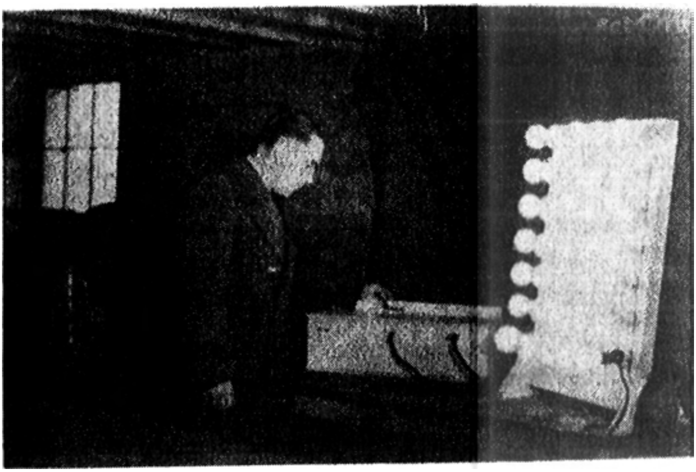

Here **the** Moray radiant energy device is providing free power to 35 100-watt lamps and a 1200-watt Iron.

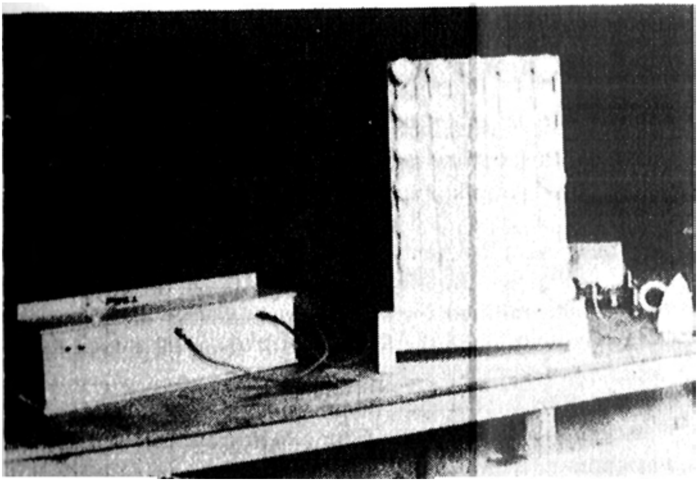

Once tuned in, the power continues and is free for the taking.

Photo 4. The Moray radiant energy device.

Shortly thereafter, Moray produced his first elementary device that delivered measureable electrical power, and he continued to work diligently on energy devices when he had time. In the 1920's and 1930's he steadily improved his devices, particularly his detector tube, the only real secret of the device according *to* Moray himself. In his book, *The Sea of Energy in Which the Earth Floats,* Moray presents documented evidence that he invented the first transistor-type valve in 1925, far ahead of the officially recognized discovery of the transistor. In his free energy detector tube Moray apparently used, inside the tube itself, a variation of this transistor idea — a small rounded pellet of a mixture of triboluminescent zinc, a semiconductor material, and a radioactive or fissile material. His patent application (for which a patent has never been granted) was filed on July 13, 1931, long before the advent of the Bell Laboratories' transistor.

In test after test Moray demonstrated his radiant energy device to electrical engineering professors, congressmen, dignitaries, and a host of other visitors to his laboratory. Once he even took the device several miles out in the country, away from all power lines, to prove that he was not simply tuning in to energy being clandestinely radiated from some other part of his laboratory. Several times he allowed independent investigators to completely disassemble his device and reassemble it, then reactivate it themselves. In all tests, he was successful in demonstrating that the device could produce energy output without any appreciable energy input. According to exhaustive documentation, no one was ever able to prove that the device was fraudulent or that Moray had not accomplished exactly what he claimed. On the other hand, the records are full of signed statements from physicists, electrical engineers, and scientists who came to the Moray laboratory as doubting Thomases and left with the complete conviction that Moray had indeed succeeded in tapping a universal source of energy that could produce free electrical power.

But in the face of all of this, the U.S. Patent Office refused to grant Moray a patent, first, because his device used a cold cathode in the tubes (the patent examiner asserted it was common knowledge that a heated cathode was necessary to obtain electrons) and, second, because he failed to identify the source of the energy. All sorts of irrelevant patents and devices were also presented as being infringed upon or duplicated by Moray's work. Each of these objections was patiently answered and nullified by Moray; nonetheless, the patent has still not been issued to this day, although the Morays still keep the patent application current.

One of Moray's efforts to develop the machine involved an association with the Rural Electrification Agency for a short time before World War II. At that time, the R.E.A. was apparently infiltrated by Communist sympathizers and high level officials of a decidedly pinkish tinge. These officials continually urged Moray to turn over all details of his device to the Soviet Union, and even arranged the visit of high-level Soviet scientists to the Moray laboratory to see the device in operation. On the initiative of the Communist-infiltrated R.E.A., one person — let us call him "Z" — spent about two months in Moray's laboratory, and succeeded in worming his way into Moray's full confidence. Moray eventually disclosed most of the details of the construction of his special valve to "Z" — the only person in whom Moray ever completely confided.

There is strong reason to believe that "Z" was a Soviet agent, and that this is where the Soviets first obtained the impetus to develop the hyperspace amplifiers later used in their psychotronic weapons.

However, Moray became quite alarmed at the continued attempts of his R.E.A. contacts to get the device into the hands of the Russians. He eventually concluded that he had become involved with a governmental group filled with radicals and reactionaries. Moray became concerned that fifth column activity was actually directed against him in an attempt to steal his device. Quotes from the *Salt Lake Tribune* of December 2, 1941, attributed to Representative Thomas D. Winter, imply that Moray's suspicions of the R.E.A. were well founded, since Representative Winter called for a full-scale investigation of the R.E.A. on the grounds of Communist infiltration. Indeed, Moray was wounded by gunshot in his own laboratory on March 2, 1940, which seems to suggest that his fears and suspicions were based on fact. Moray severed his association with the R.E.A. in February 1941.

However, his basic unit had been destroyed by a hammerslinging witness in 1939; it is not clear whether or not this was the work of "Z" or of someone else. According to his son, John Moray, the man who broke his unit, as well as other interested groups, wanted complete disclosure of the materials and construction - nothing more or less. John Moray, who operates the Research Institute in Salt Lake City, has been trying to continue his father's work since the basic unit was destroyed. Dr. Moray himself died in May 1974.

According to John Moray, highly energetic cosmic rays from space are tapped by the machine, which is in subharmonic resonance with this high-frequency energy level, and it converts this energy level into a usable form of electricity. However, John Moray's use of the term "cosmic ray" is not necessarily the same as that of modern physics, but in fact is the same as the present concept of "zero-point" energy of vacuum. T. Henry Moray envisioned all space filled with tremendously high frequency vibrations carrying vast and incalculable amounts of pure raw energy. He envisioned the dynamic universe as a turbulent source of great energy, just as the ocean waves carry energy throughout the earth. This was also the vision of Tesla, and after him of Clifford and Einstein, who turned the attention of general relativity to the problem of the nature of the vacuum itself. Clifford and Einstein founded a branch of physics that has come to be known as quantum geometrodynamics, the study of energetic change that occurs in little pieces, including the concept that very small lengths of space, or pure vacuum, themselves oscillate at great frequency and with great energy. In Moray's day relativity was still a strange and unproven branch of physics, suspected and rejected by most of the physicists of the day, and quantum physics was still in the process of being worked out. There was no theory whatsoever predicting that empty space itself not only contained prodigious quantities of energy, but in fact *was* prodigious quantities of energy. But slowly, over the decades, the picture has changed, and the modern followers of quantum geometrodynamics assert the truth of Tesla's original vision. Today we know that one cubic centimeter of pure vacuum contains enough energy to condense into 10^{80}—10^{120} grams of matter! Thus the major part of Moray's thesis — that vacuum itself contains unlimited energy — is vindicated today. In this sense empty space is like a gigantic, restless ocean, and Moray's free energy "tapping" device is no more mysterious than the water wheel. In other words, his thesis that the energy is there to be tapped is correct; it only awaits a practical method to tap it in order to solve the energy problems of mankind forever.

Moray thus is vindicated as a man ahead of his time, who simply built a device before any theory existed to explain its operation. Moray met with constant opposition, and his life and that of his family was anything but pleasant. Certain pressure groups constantly tried to force him into selling for almost nothing, or to disclose the secrets of his valve's construction. He was often attacked and he was sometimes shot at on the city streets. In fact, his life was threatened so often that he was forced to install bulletproof glass in his automobile. His laboratory was broken into, some of his components and papers were stolen, and his dogs were continually being killed. Moray was shot in his own laboratory, and he himself always carried a gun. Because he was harassed ceaselessly, over the years he became understandably suspicious and close-mouthed about his work. He sometimes greeted visitors at his desk with a loaded pistol lying on the desk within easy reach, and occasionally confirmed that he would not hesitate to shoot if he were attacked or threatened. One of his greatest fears was that big interests would take up his invention and simply shelve it to keep it from benefiting the public. When large companies made him offers, he always demanded written guarantees that the device would be put into production and sold to everyone, once it was developed.

Moray also realized the potential of his device as a weapon. He was fully aware that the only difference between a controlled energy device and an explosion is the rate of release of the energy. According to John Moray, his father received an offer to go to Japan in 1938 because he had built a deathray which operated off the device. According to John, a representative of the Japanese government came to the laboratory and offered a position to Moray, but he refused. Allegedly Moray had managed to rig the device so that it projected a ray of energy in a beam, and could kill mice instantly at over fifty feet. When radiated by the beam, the mice were carbonized immediately. They appeared frozen, but nothing was left except their shells, and they simply fell apart when touched.

In one experiment Moray ran his device for 157 hours without any connection to external power sources, and produced over fifty kilowatts of power during the test. He also found that an additional fifty kilowatts could be added by simply providing another tap further back in the circuit. When he shut the device off from this test, he had proven once and for all that the device was generating electrical energy from free and natural sources without batteries or external power. During this test nothing in the machine heated up; instead, all parts of the circuit ran absolutely cold. This alone is totally unexplainable by ordinary electromagnetic theory, and it strongly implies the truth of Moray's assertion that the device simply collects the energy in each of its stages in a subharmonically resonant manner, in synchronization with the extremely high frequency cosmic oscillations. In other words, since the parts of the machine ran cold, it is clear that the energy was being simultaneously collected at each stage rather than being processed through the individual stages in serial order, since serial processing in the conventional electrical sense would lead to resistance heating of the circuit elements.

Moray's device used twenty-nine stages of his special detector valves, which were difficult to produce, costing about five hundred dollars each. Only about one in four proved suitable for operation. He also had difficulty in obtaining sufficiently pure materials to make his special mixture for the pellet that enabled tube operation in a one-way gating fashion. Moray explained that his device was based

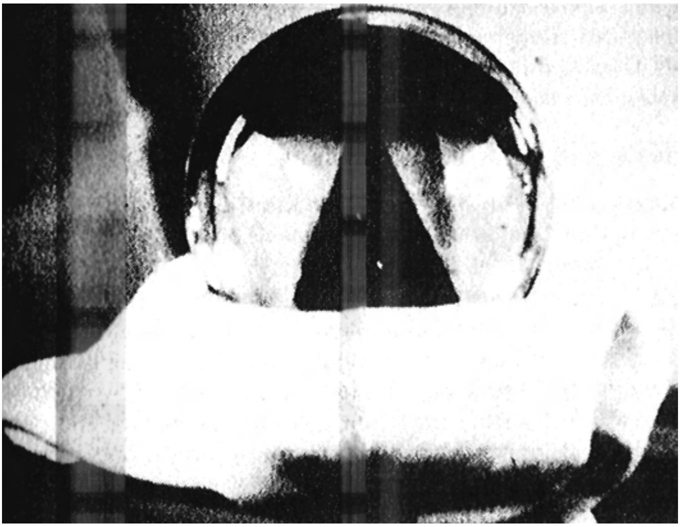

Photo 5. An Atlantean power crystal?

on the discovery of a mixture that would act as a one-way gate for the high frequency oscillations of space, so that the energy could go through the material more readily in one direction than another. Thus it was like a one-way gate valve to an ocean wave; the energy "water" could flow in in each valve, but was prevented from flowing back out. The assemblage of multiple stages thus provided a series of collectors which contained enough energy to be useful. Theoretically there was no limit to the number of collectors that could be added, and so there appeared to be no limit to the energy that such a device could produce.

AN ATLANTEAN POWER CRYSTAL?

The next two photographs are of a most amazing quartz crystal sphere removed from a chamber in a pyramid underneath the surface of the ocean in the Bermuda Triangle. Dr. Ray Brown purportedly removed the object in a dive while treasure-hunting with several friends after a storm had stirred the sandy ocean floor in 1968. Brown describes how he swam into a chamber in the pyramid after swimming around it three times, and how he saw seven stone seats surrounding two handlikc stone holders for the crystal in the center of a table or dais, with a two inch metallic rod running down vertically from the roof of the stone chamber and pointing directly at the crystal. Brown plucked the crystal from the holder and ascended to the surface; he found that his companions had brought some nonoperative objects to the surface which appeared to be some sort of strange electrical apparatus.

Brown himself had a strange foreboding and a strong impulse against returning to the site, and so he refused to do so. Three of his companions reportedly returned at a later date, and their boats were found floating abandoned in the Triangle. The three men were never seen again.

The crystal does indeed have some strange properties. In the first picture one can clearly see a very sharply outlined, pyramid shaped inclusion inside the quartz. The pyramidal inclusion is dark, but the crystal is absolutely clear. However, the second photo shows that the quartz is milky when seen from a different angle. Here three pyramids can be seen.

Other objects have been seen and photographed in the crystal from time to time. In *The Pyramid Guide,* Vol. V, No. 4, there is a photograph of the crystal showing a very distinct, humanlike iris and eyelid appearing in it, making it resemble nothing so much as a giant human eye. In that photo the three pyramids are not visible.

While I have not personally investigated the crystal and examined it in detail, some general remarks can be made. (1) Brown reported finding the pyramid after a huge storm had stirred the ocean, including the ocean's sandy bottom. This had revealed outlines on the bottom resembling buildings, which had attracted the attention of the treasure hunters. (2) Such a storm, over the ocean's great water mass, is an excellent kindler of bioenergy. In other words, Brown and his companions were definitely in a place which should have had an enormous kindling of bioenergy from the biosphere and the collective human unconscious. (3) Sufficient publicity has been given the Bermuda Triangle's mysterious disappearances to have its alleged ability to open a window between worlds implanted deeply into the unconsciousnesses of a large number of persons. (4) Brown was with a group of persons who were sharing the same sensations with the storm and surrounding area, so there was sufficient coherence of their personal unconsciousnesses to accomplish at least some degree of tuning. (5) The legends of Atlantis were in a great revival period in the late 1960's from the popularizing of Edgar Cayce's work and other material, as indeed was the Egypt/pyramid/Atlantis mythology in general.

So conditions were very ripe for a tulpoidal contact (see Part Two) of extensive magnitude — perhaps even to the point of tuning in the ruins and pyramid for a period. Brown also reported sensing an omniscientlike presence, which is symptomatic of many such tulpoidal contacts or materializations.

In such a case, Brown could have actually swum into such a pyramid that had materialized from the collective unconscious — which after all has the psychokinetic power of Uri Geller times ten to the four billionth power or so, and which largely maintains the format of our physical reality in the first place. The collective unconscious has no difficulty in creating even many alternate worlds or materializing and dematerializing entire areas, let alone in producing a single pyramid and a ruined city.

Brown's action in plucking the crystal out of its holder and swimming back to the surface would also have constituted a tuning/kindling which was "returning with the crystal into the normal world" — in which case it could have been tuned into complete stability so that it would endure, rather than fade away as do most tulpoids.

Until we have developed psychotronic instruments to scan the crystal and reveal its true story, we cannot say for certain that any of the above actually happened, and that some sort of delusion or fakery was not involved. The point is, it could have happened just as Brown reported it, in the theory I have constructed in the second part of this book.

Until we free our minds from the present bias that physical reality is always permanent and unchanging, we shall never be able to adequately investigate such phenomena as Brown's strange crystal.

THE CRYSTAL SKULL

It appears that at least some of the ancient peoples understood certain aspects of psychotronics and even possessed some psychotronic devices.

An example of such a psychotronic device is shown in the next two pictures, reprinted here with the kind permission of Bill Cox, from *The Pyramid Guide,* No. 17, May-June, Vol. Ill, No. 5.

To quote directly from *The Pyramid Guide:*

> **These are two views of the famed crystal skull... completely hand-formed, possibly centuries in the making from a single block of hard quartz crystal. It was recovered from beneath the altar ruins, within the jungle-covered, 7 1/2 acre citadel of the ancient Mayan city of Lubaantu'n by explorer F.A. Mitchell-Hedges in 1927. A several-stories-high pyramid was also found within the citadel complex. The skull, possibly of Atlantean and/or Babylonian origin may have been brought to the Americas by the Phoenicians around 1,000 B.C. Frank Dorland says historical records suggest the skull may have been held in the London inner-sanctum of the Knights Templar prior to the Holy Wars. The high priests probably executed ceremonial magic with the skull; a ribbon-like prism in the cranium acts as a magnifier. Hollow light pipes projecting upward from the base permit light reflections from the eye-sockets. The moveable jaw easily coordinated motion with speech. Lesser quality crystal skulls exhibited in the British and Paris museums, and miniature versions, lack the Mitchell-Hedges skull's brilliant craftsmanship**

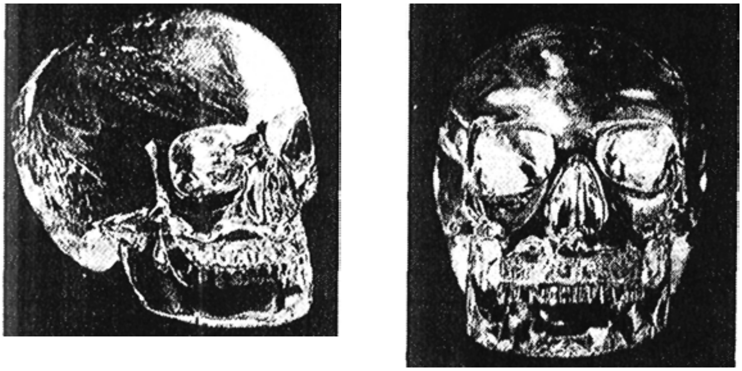

Photo 6. The crystal skull.

Two views of the famed crystal skull. Completely hand-formed, possibly centuries in its making from a single block of hard quartz crystal. It was recovered from beneath the altar ruins within the jungle-covered seven-and-a-half acre citadel of the ancient Mayan city of Lubaantu'n by explorer FA. Mitchell-Hedges in 1927. A several-stories-high pyramid was also found within the citadel complex. The skull, possibly of Atlantean and/or Babylonian origin, may have been brought to the Americas by the Phoenicians around 1,000 B.C. The high priests probably executed ceremonial magic with the skull; a ribbon-like prism in the cranium acts as a magnifier. Hollow light pipes projecting upward from the base permit light reflections from the eye sockets. The moveable jaw easily coordinates motion with speech. Lesser quality crystal skulls exhibited In the British and Paris museums, and miniature versions, lack the Mitchell-Hedges skull's brilliant craftsmanship.

One of the major African tribes is also reported to have once possessed such a crystal skull as a holy object and the center of their religion. Their skull was reportedly stolen by a white thief in the early 1800's,

So let us now discuss sonic of the peculiar properties of such a device.

First, quartz is a very hard substance and not easily worked at all. Making such a perfectly carved piece, with removable hinged jaw and eye lenses, must have consumed a tremendous amount of manual labor and considerable time, even centuries. Such an effort is not lightly undertaken; usually there is an overpowering reason for such persistence. In this case, there may well have been such a reason, and ancient priests may well have known it.

Quartz has magic qualities from a psychotronic point of view. It is open to both the ultraviolet and the infrared, and so it possesses two magic windows into the hyperframes, bioframes, orthogonal frames, and other realities. Since the third and higher bioframes (orthoworlds) are all mind worlds, quartz has open magic windows into the mind and mind energies. Thus Frank Dorland surmises correctly that the device can amplify and focus mental energies.

Furthermore, quartz is a piezoelectric crystal. Being crystalline, it possesses a lattice, a regular array of molecular stages. Thus it possesses a large number of identical stages, which is important for hyperchannel amplification.

Properly shielded against visible light, or in dim light, the crystal skull thus can function as a mental amplifier in the hands of a skilled user. Also, the quartz it piezoelectric; multistaged electrical piezoelectric currents are natural hyperenergy-hyperfield collectors, and under the proper conditions the quartz can produce heat and light, since collected hyperfield energy can be orthorotated into electromagnetic field by kindling. It can even produce these effects at a distance due to the hyperchannel circulation patterns surrounding a magnetic field or a localized area of magnetic field. The easiest place to produce effects with the skull would be inside another human head (in the brain).

So, in the hands of a skilled practitioner, the skull is (1) a collector of bioenergy, capable of turning this into electromagnetic energy and physical effects, (2) a form resonance inductor, i.e., a sort of super witness which matches the general shape of the human skull and is capable of resonating with a human skull (and thus with its contained brain) at a distance through hyperspace, and (3) a psychotronic device capable of producing psychotronic effects such as the augmentation of healing energy, psychic attack, telepathy, clairvoyance, precognition, and psychokinesis, where these initiative abilities are possessed by the user or practitioner.

Thus, in the hands of an adept, the crystal skull is indeed a formidable device, and one can well understand how the motivation existed to make it, even though its production may have taken centuries. Since in the ancient days knowledge was zealously guarded and usually incorporated directly into religion, such psychotronic artifacts as the crystal skull would have been regarded as the most powerful, hence the most sacred, of objects. One can therefore speculate that the elite of the elite of the priesthood were the only ones who knew of and were taught how to use the skull, and that this information was guarded extremely closely. The legend of such a skull existing as a central part of the religious magic of an African tribe is not beyond belief. Ordinary crystal balls, which are certainly made more easily, may have originated from the secret rites connected with the skull. An ordinary good quality, clear quartz crystal may well have been the substitute for priests who did not belong to the supreme inner circle, or the device used when a skull was not available.

I have seen at least one picture of an optical image which spontaneously formed in the crystal skull; the image closely resembled that of the Mayan astronomical observatory at Chichen Itza.

To give a simple example of how the skull may have been used by a skilled priest, let us suppose the priest had an enemy he wished to disable or kill. With the proper psychic exercises he could have induced an increase in his own psychic energy. Then, using the crystal skull, he could simply have visualized the image of his enemy's face, and visualized the energy or light pouring into his enemy's head. The stability of the imagery is more important than effort of will.

At any rate, steady application of this imagery starts the bioenergy projection into the enemy's head and brain, and it collects there. Collection to the kindling threshold starts turning the bioenergy into electromagnetic field. At this point, a host of effects could be produced, depending upon the specific wish and induction by the priest. The effects include sharp pain (such as migraine headache); a leaden mass or frozen mind in which the brain activity itself is slowly frozen and stopped; the generation of a strong, overpowering emotion such as rage, hatred, panic, or terrible sadness; a sharp attack on the mind itself, as if a saw or knife were slashing a physical object to bits; numbness of the body or selected portions of it; interruption of the heart's functioning; kindling of the brain and nervous system into seizures, convulsions, or violent muscular spasms (capable of breaking bones); and an increase in blood pressure until the vessels swell and burst. In the hands of a priest who was strong in psychokinetic ability, this amplified ability would have been projected through hyperspace to the unfortunate victim's vicinity; objects such as stones could suddenly have levitated and hurled themselves at him, bushes and small trees could have enveloped him, etc. All in all, the skull can be used as a terrible weapon.

A few of these effects, and others I have not described, are presently in use by certain lunatic fringe groups in the United States; they use basic devices other than a crystal skull. I have been subjected to several strong attacks, and can vouch for their effectiveness if the target person is unskilled in shielding himself or evading such attack. In two cases I was nearly killed, and in a third case in substantial jeopardy.

Use of the crystal skull by a group of priests would give even more power, for from such a group the kindling level at which bioenergy orthorotates into electromagnetic field energy is much more easily approached. The inceptive power is of course usable for either good or evil purposes. Good purposes would include amplifying psychic healing; curing fevers and other infections; stopping internal bleeding; dematerializing growths and tumors; placing a person into deep anesthetic sleep so that a surgical operation can be performed; treating soil, seeds, or plants to assist sprouting and growth; treating emotional disorders with cancellation thought forms (contrary forms), and projecting them directly onto the unconscious to neutralize the disturbing form; and communicating telepathically at a distance.

A 300-year fabrication period is more credible if we recognize that the quartz crystal would have been usable from the beginning, just as is a crystal ball. So the high priests could have used the rare clear crystal (the skull is made from a single crystal), while priests recently initiated into the level that knew of crystal devices worked on it when it was not needed for use. Such a practice would explain why other skulls less finely worked have been found; these had simply not been as highly finished as the Mitchell-Hedges skull.

Furthermore, cutting off an enemy's head and exhibiting the skull was a common practice among many primitive tribes as a way of showing power. Crystal devices could have been used as a means of combining psychotronic effects with prevailing religious practice and beliefs. The fact that this eventually yielded a psychotronic amplifier of great power in the hands of a skilled priest would have led to the zealous guarding of the knowledge and the device from all but a very few of the elite.

METAL BENDING

The next remarkable sequence of photos from a Super 8 movie film taken by James Bolen, editor and publisher *of Psychic* (now *New Realities)* magazine, shows Israeli Uri Geller breaking a dinner fork in two through mind over matter. The fork, which Bolen personally verified as being intact before the demonstration, gradually becomes pliable at its midsection as Geller rolls his thumb and index finger over it. The fork finally breaks apart, the prong part clinging slightly to the handle just before it drops away, suggesting that the stainless steel momentarily became plasticlike.

Here the kindling is taking place in the nuclei of the metal atoms and also along the lines of field force in the bonding lattice. The protons in the nuclei may become quarklike due to partial extinguishing of the pions which create their electrical charges. The softening of the electrical field directly softens the ability to hold electrons and rearranges electron orbits. In addition, kindling is occurring in the orbital electrons, partially extinguishing their charge also. The overall effect is for the metal to become plasticlike as lattice bonds are severely weakened. For a metal break due to these effects, the metal in and near the fracture should show the results of severe electromagnetic fields at the localized level, e.g., as if subjected to intense heat from an internal laser action. This has indeed been found to be the case in other experiments, notably by Byrd and Franklin.

THOUGHT PHOTOGRAPHY: STELLA LANSING

Stella V. Lansing is a remarkable person who has captured a large number of authentic thought photographs on film, and in whose presence others often get thought photographs when they operate their own cameras. For several years her work has been investigated by the psychiatrist and noted researcher, Dr. Berthold Schwarz, who has published numerous papers in scientific journals on his studies of Mrs. Lansing. The late J. Dunninger also observed Mrs. Lansing's work and concluded that it was genuine, and that no magician could accomplish by sleight of hand or other trickery the remarkable results she obtained.

In Stella Lansing's photographs all the evidences and signatures of tulpoidal or thought energy kindling into the first biofield, electromagnetic energy (light), are seen (see Part Two). Typical, metallic-looking, clamshell, UFO-like images are often obtained, most often with several objects arranged in a clocklike circle symbolically: the unconscious may well be aware that the time channel is the paranormal channel, and symbolically express this in a clockwork pattern suggestive of time.

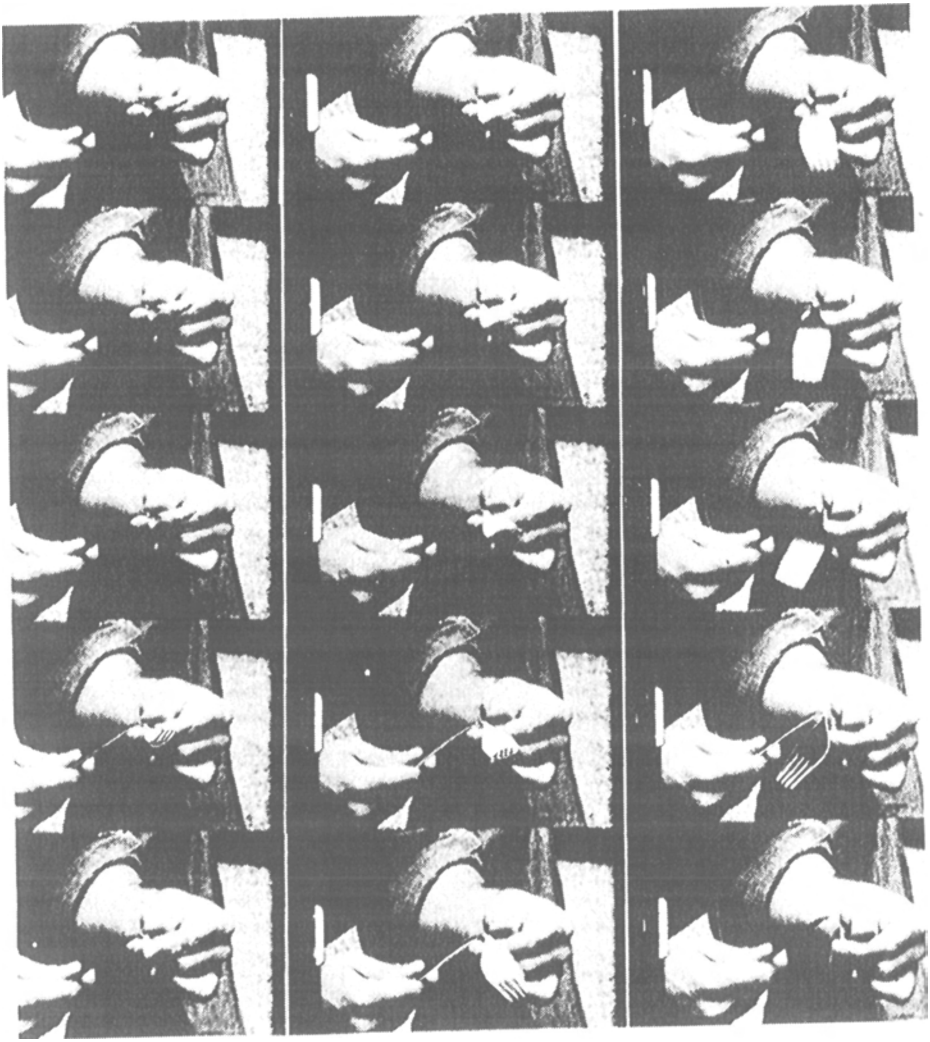

Photo 7. Metal bending.

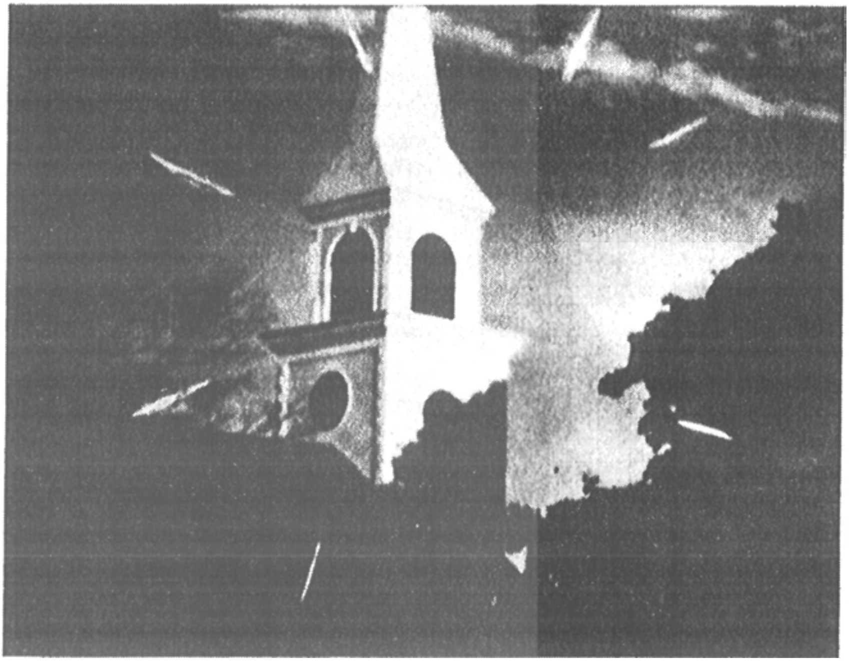

Late evening clockwork pattern of "clamshell-like objects" obtained when photographing a church steeple.

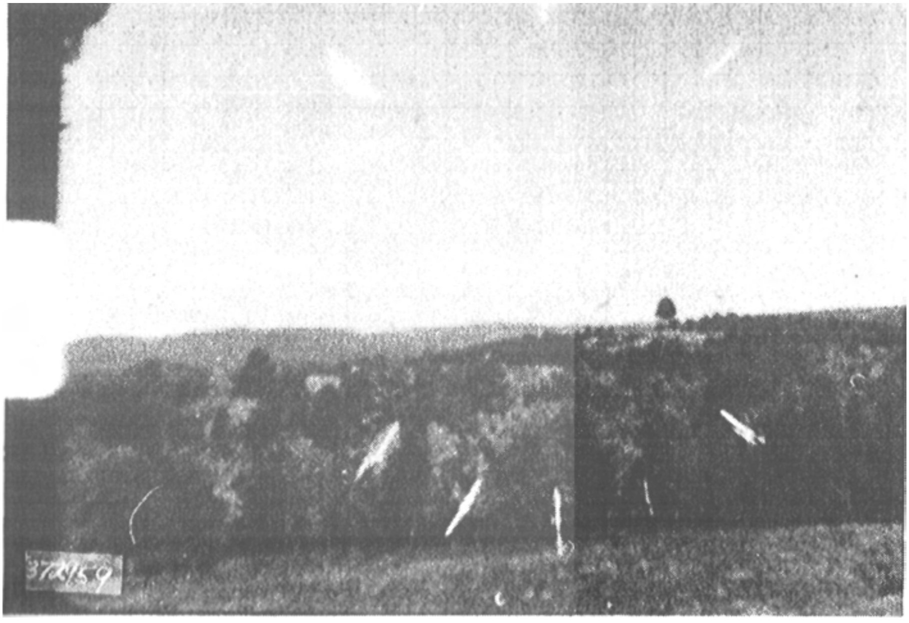

First daytime clockwork pattern, Route 47, Amherst, Mass., September 1972.

Photo 8. Clockwork patterns.

Just as Geller and Kulagina have been noted to transfer psychokinetic ability to other talented persons who were in their presence for a period of time, Mrs. Lansing's clockwork patterns and psychokinetic abilities appear to have stimulated another person's abilities in similar fashion. Paul Koslouski is a researcher who worked with Mrs. Lansing for a short time, and in his own thought photographs the clockwork pattern so characteristic of Mrs. Lansing's pioneering efforts dutifully appeared.

Mrs. Lansing is a brave and gracious lady, and her significant work with thought photographs is important to the understanding and advancement of parapsychological science.

THE DISPLACEMENT EFFECT

The next picture was taken about 1938 in Chicago with a camera on a tripod and a timed release. It was taken by Paul Trovillo, who is the gentleman sitting on the bench with somewhat of a problem of transparency and partial disappearance.

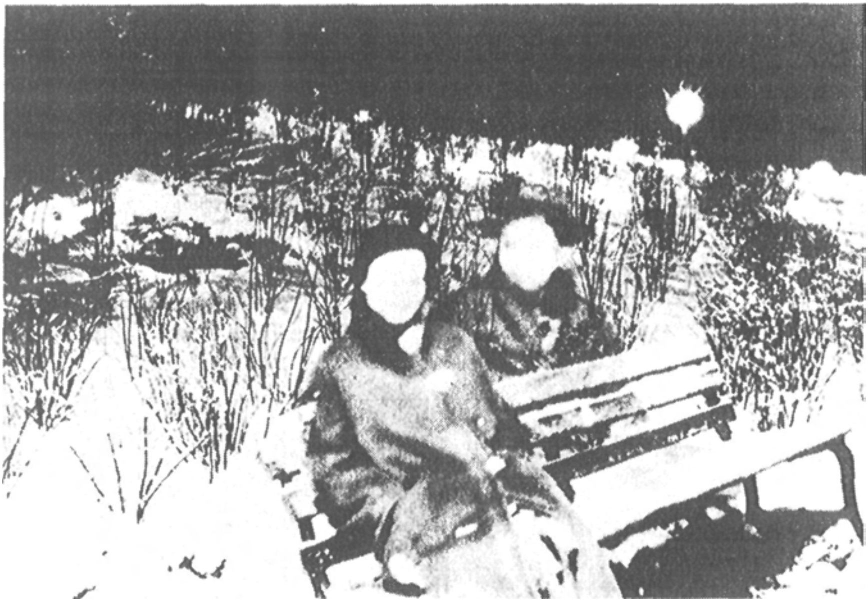

Photo 9. Displacement effect.

The light is by an overhead arc light, which is rich in ultraviolet (one of the magic window frequency bands in the hyperchanncl crosswalk between orthogonal worlds). Pan film was used in the camera, and the picture is of Paul and FerneTrovillo. Note the bushes through Paul's left shoulder and through the slats of the bench at approximately Paul's waist. Also note that the background is in excellent focus.

1. Robert Pavlita, Czechoslovakian engineer, and his daughter.

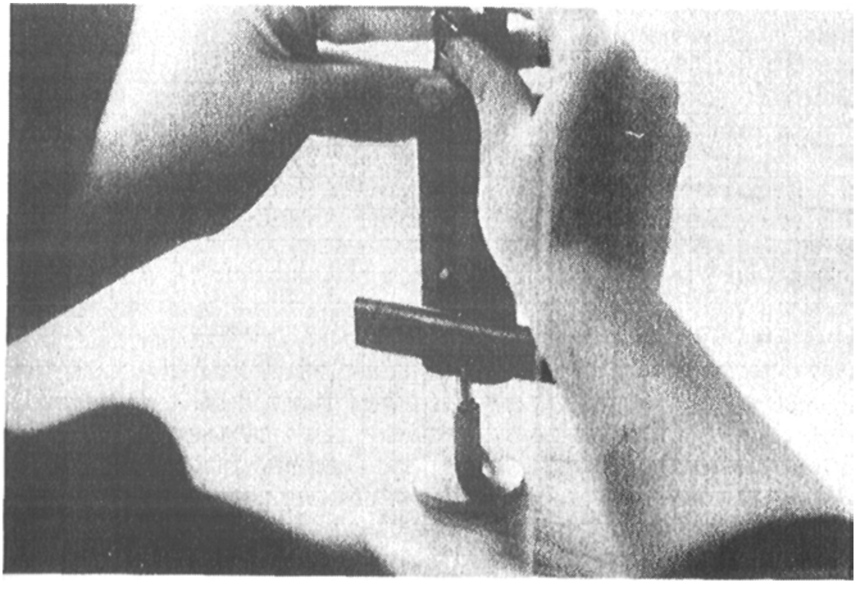

2. Device becomes a psychoenergetic "magnet" when charged by the human being—and will pick up materials, including nonmagnetic objects.

Photo 10. Pavlita's psychotronic generators.

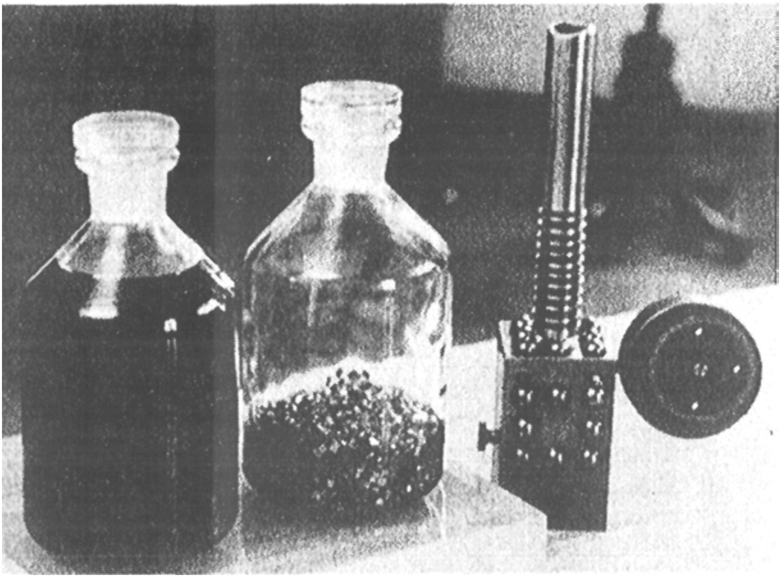

3. A generator that can precipitate solutes from solution. The jar to the left contains a dense solution; when the generator was placed over identical solution in the jar to the right, the precipitate shown in the jar resulted.

Occasionally the kindling of bioenergy (which is greatly eased in ultraviolet light) leads to eddies of hyperchannel connections between separated portions of the ordinary frame. The normal three-dimensional frame is constructed action quantum by action quantum, and each quantum is fissioned and fused individually to give individual increments or jumps of time and length. Appreciable kindling intermixed in the frame construction can provide other geometrical paths for light, so that the light rays during one or more segments of their straight-line path arc actually flipped into and back out of hyperspace. In cases such as this, one can often see right through a physical object, or even inside it (see Part Two).

I have worked with a psychic group in the American Association of Meta-Science, and have seen one experiment where similar effects were obtained, with the camera on a tripod, in a picture taken by my wife.

Excellent pictures of the Loch Ness monster were recently taken by Doc Shiels. These pictures have been subjected to computer analysis by Ground Saucer Watch (GSW) - and GSW was alarmed at the discovery that the monster in the portrait exhibited transparency. Indeed, for a materialization less than complete, which the monster must be if it is to keep disappearing, one would expect or even demand such transparency, even though a large body of water is a good collector for kindling of bioenergy and tulpoidal forms (UFO's are also notoriously connected with rivers, lakes, and the ocean).

Note also that the Dennis Billings UFO picture demonstrates the same sort of partial transparency. There is a readily understood physics that applies to these effects, but the effects are not contained in the present monocular interpretation of physics and reality.

PAVLITA'S PSYCHOTRONIC GENERATORS

In Czechoslovakia an engineer named Robert Pavlita has been building and demonstrating psychotronic generators for about thirty years. These are devices of various shapes and sizes and composed of materials such as steel, iron, brass, and copper, and they often contain mysterious symmetries of pattern, holes, and shape. The devices for the most part charge up with thought energy when subjected to human contact or eye contact, and then perform various functions such as precipitating a solution, turning, picking up or magnetizing nonmagnetic materials, and killing insects. The devices thus accomplish orthorotation and kindling of thought energy. Pavlita has admitted that his devices could be scaled up to serve as weapons — unfortunately, every tool can be used for either good or evil. He has demonstrated a great many times the killing of insects by pointing a charged rod at them. He also admitted that in an accident his daughter became semiparalyzed and her life was in grave danger, requiring him to work feverishly without stopping for about two days to build a second device to reverse the effects of the first. Pavlita has also demonstrated that his psychotronic devices can either stimulate or retard plant growth, depending on which polarity the plant is exposed to.

The loose information net of which I am a part has recently learned that Pavlita is now in charge of a secret project that has succeeded in building two weapons, one of about two hundred miles range and the other of unlimited range. No further details on the weapons are available.

Pavlita is a pioneering genius, and it is to be hoped that he will eventually release the secret of his psychotronic devices to the world for use in the healing arts and other beneficial sciences. Bioenergy/hyperenergy is real, and Pavlita has scored a major breakthrough in building simple devices to harness it.

CELL-LIKE INVISIBLE TULPOIDAL FORMS

Trevor James Constable has succeeded in evoking and capturing many of these forms on film. The first photo was taken by Constable on 25 August, 1957 in the Mojave Desert, and shows an amoebalike or cell-like bioplasmic form containing shapes that resemble nucleoli, vacuoles, and other cellular structures. The invisible object was at first directly overhead, but moved to positions that enabled inclusion of local terrain in the pictures.

In material presented in Part Two of this book, I show that all that a physical form or physical body is is a total tulpoid, *i.e.,* a form that can be shown to exist in each and every mind. All that a living biosystem is is a one-to-one coherence established between such a total tulpoid (physical form) and a monocular tulpoidal universe (mind world or hyperframe). Furthermore, all possibilities exist concretely in the various hyperworlds. Thus Constable's plasma creatures do indeed exist concretely—but in invisible, orthorotated hyperworlds (mind worlds). The revolutionary discovery by Constable has been a direct, straightforward way to evoke and photograph such bioplasmic tulpoidal forms. His experimental work is of absolutely fundamental importance, and I would not hesitate to compare its ramifications to those of the Michelson-Morley experiments.

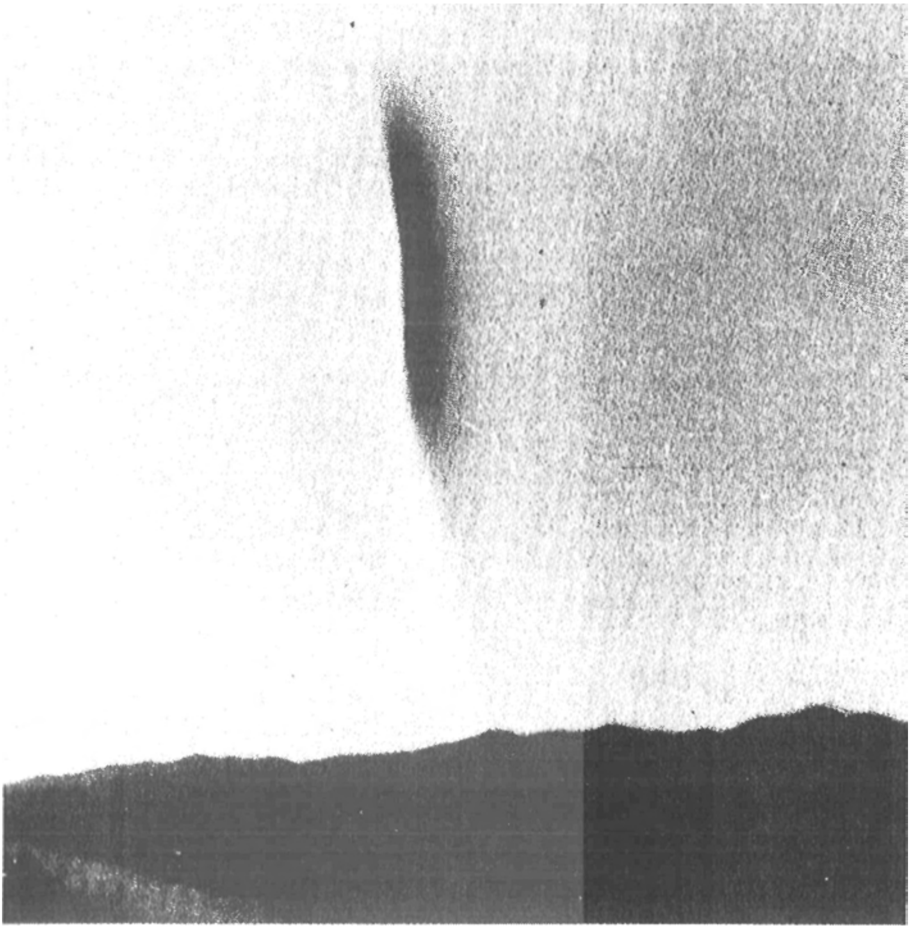

Photo 12. Form for potential materialization.

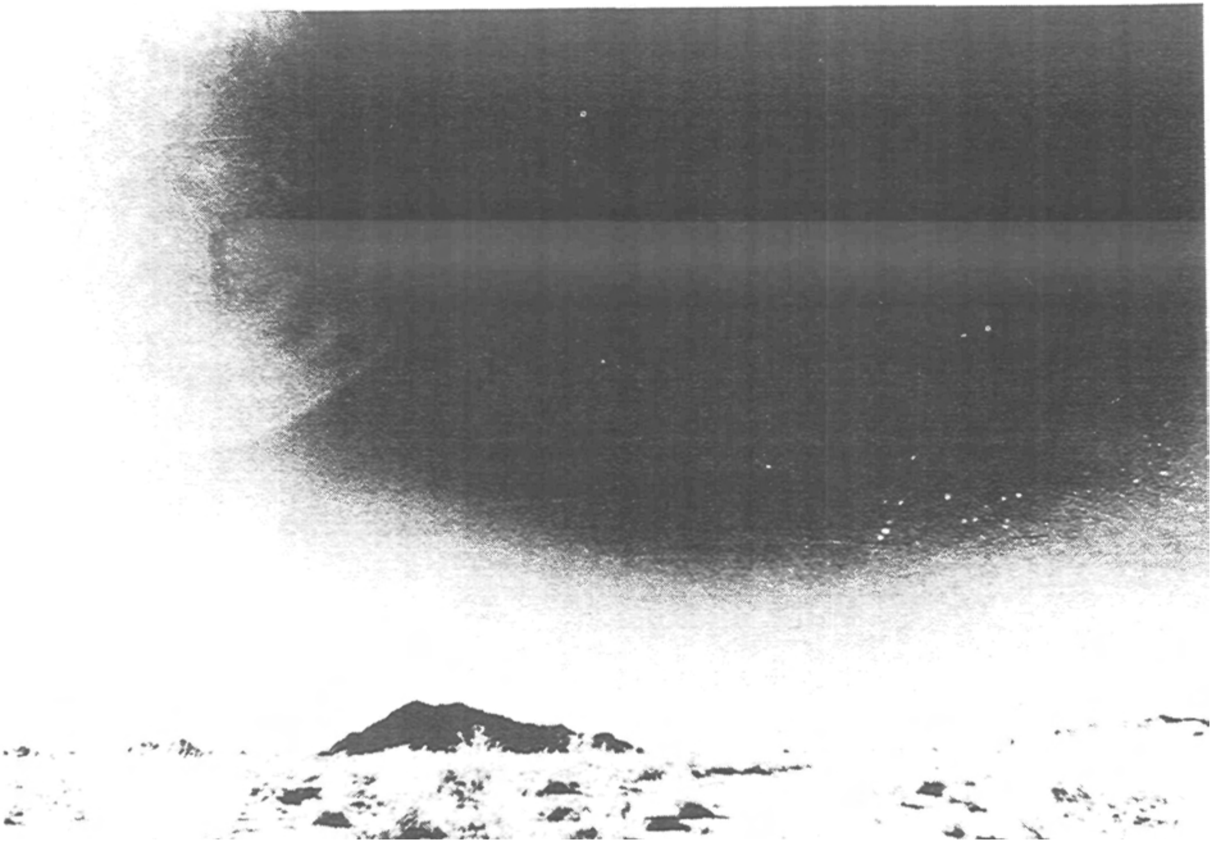

Photo 11. Cell-like invisible tuipoidai forms.

These experiments are also directly related to the work of Miller and Fox, who have shown that, when the probable atmospheric conditions *of* the earth of millions of years ago are exposed to electrical discharges and ultraviolet radiations, amino acids are formed; and then these proteinoids will form microspheres having cellular characteristics, even the ability to grow and reproduce. Every possible life form is already present in virtual state reality, and under proper selective conditions can be kindled into the first bioframe (electromagnetic field) and photographed, or even into the zeroth bioframe (full materialization).

In the UFO photographed in strong potential condition for materialization by Trevor James Constable, note the form of a disc on edge, only a portion of its form being tangible even to infrared. Note also the whitish lobe of the tulpoidal force field extending earthward from the discoidal region. Highspeed infrared film was used, no filter, shot just before sunrise near Giant Rock, California. Constable's pioneering work is remarkable laboratory evidence for orthorotation and materialization of tulpoidal energy forms.

UFO'S

The following pages show some of the most extraordinary UFO sightings and other paranormal phenomena of recent years. Photograph 13 shows a noiseless UFO seen by a university student near Las Cruces, New Mexico, in 1967. (The student reported that, as he changed plates in his camera, the object disappeared. He had been taking photos of rock formations for a class in geology.) In the Spring of 1950, two photographs of a highly materialized UFO were taken by Paul Trent on his farm outside McMinnville, Oregon. The first photo (photo 14) showed the UFO at an angle, while the second (not shown) caught the UFO in direct profile.

1. Black Ring UFO

From Project Bluebook, here is an example of a tulpoidal UFO moving along through the sky and exhibiting the "black ring" effect. In this case only the edge frame is visible, and that edge is generating the black ring or "light quenching" effect. Note that the inside portion of the tulpoid is quite transparent, for one can see the clouds behind it clearly.

2. Anchor Patterns

One of the interesting aspects of combined orthorotation into our normal three-dimensional space and out of it from hyperspace is an anchorlike pattern, shown in the next four pictures. Photo 1 was taken by Ira Maxey, and it shows several tulpoidal forms, which are intersecting our frame in a white or photon-emitting anchorlike pattern. The curved part of the anchor pattern is due to the spinning or orthogonal rotation, and the shaft portion intersects it at right angles. Note also the clouds, whose water content is highly conducive to kindling of the tulpoid — into the first bioframe, in this case.

The second photo from Project Bluebook shows another anchorlike tulpoid — this time the orthorotation is reversed, giving photon absorption — moving through the sky over a dock area by the water.

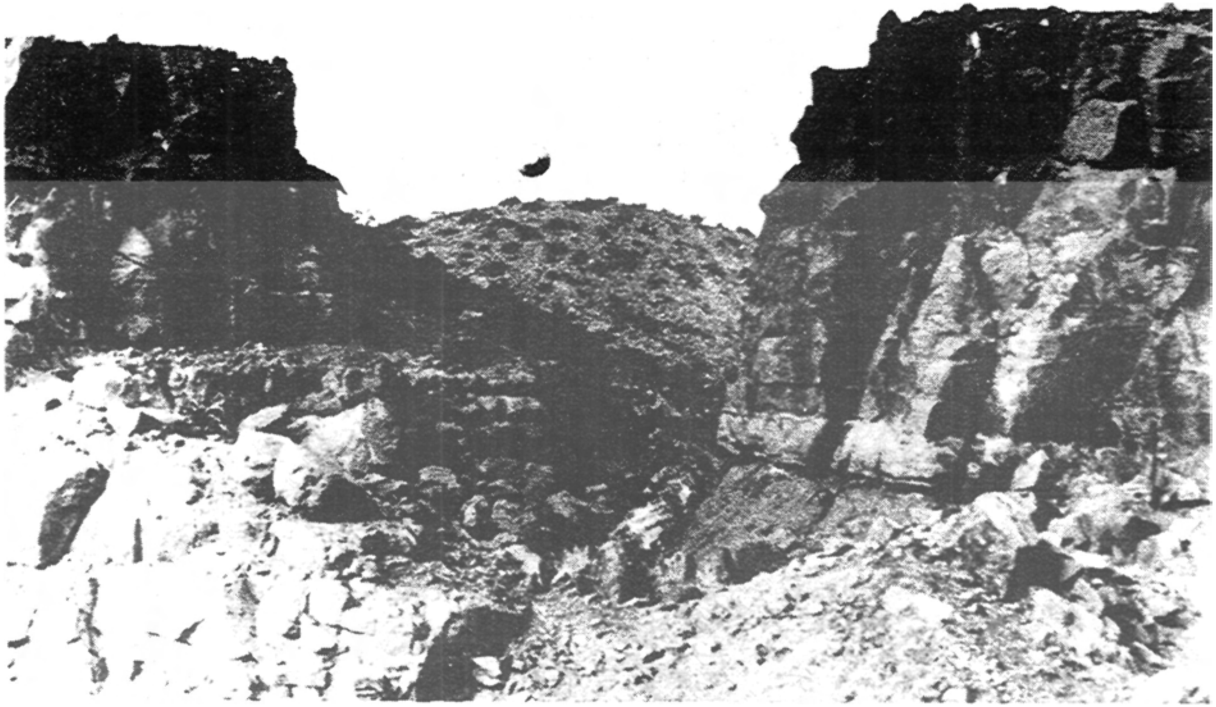

Photo 13. New Mexico UFO.

The third photo, taken in Wales, is one of several which clearly captured another anchorlike tulpoid in a photo-emitting mode.

The fourth photo, taken by Fay Clark in a darkened room, shows very similar photonic patterns being emitted by Ingo Swann's hands after Swann was asked to try to produce energy to be photographed. In the rightmost arc the shaftlike perpendicular bisector can be seen, associated with Ingo's arm. Bioenergy and mind-energy are the same, whether being emitted from a single human system/mind or from a collective system/mind.

3. Bifringement

Another common effect demonstrated by UFO's is bifringement, clearly shown in this photo from Project Bluebook. Here a UFO in highly materialized form is shown just beginning to turn back into hyperdimensional space. Interestingly enough, in hyperspace a tulpoidal form can turn in two or more directions at once, and this one appears to be doing so. A photon-absorbing turn is indicated by the outer bifringement ring, which is white. These can also be compared to photon interference patterns commonly demonstrated by light waves from two or more sources. Just as there are interference patterns in our three-dimensional space, there are interference patterns in higher dimensional spaces. The three-dimensional intersection in our frame of one of these higher dimensional interference patterns may be black, white, or consist of both black and white components.

4. UFO's in Water

Water is a substance that changes its state and form in a fantastic complex and is therefore one of the very best collectors for bioenergy or tulpoidal energy. That is, in fact, why water is so universally associated with living biosystems, for biosystems need a good collective medium in which to establish coherent tuning. So it is not accidental that water is necessary to life as we know it, or that our bodies are largely made up of water. For the same reason (ease of kindling), UFO's and many other tulpoidal forms have a high degree of association with water. Nessie and other water monsters tune in from time to time in large bodies of water such as lakes, rivers, and oceans.

This photo from Project Bluebook shows a UFO tulpoid in the first bioframe (photonic) moving along in the water.

5. UFO's in Clouds

Again from Project Bluebook here is a picture of a UFO tulpoid kindling in water, this time in the moisture in clouds. Clouds are also usually electrically charged, which enhances the tuning effect, particularly when the clouds are darkening and so are highly charged and also densely loaded with moisture. The darker clouds shield out more of the visible light, reducing the hyperchannel (time) squelching effect of visible light.

6. UFO Seen by the Author

The next UFO is one of several seen by me, and it provided a very curious, anomalous phenomenon in my memory. Furthermore, this UFO was sighted by hundreds of persons over a four-state area, so there is no question that something was there in the upper atmosphere.

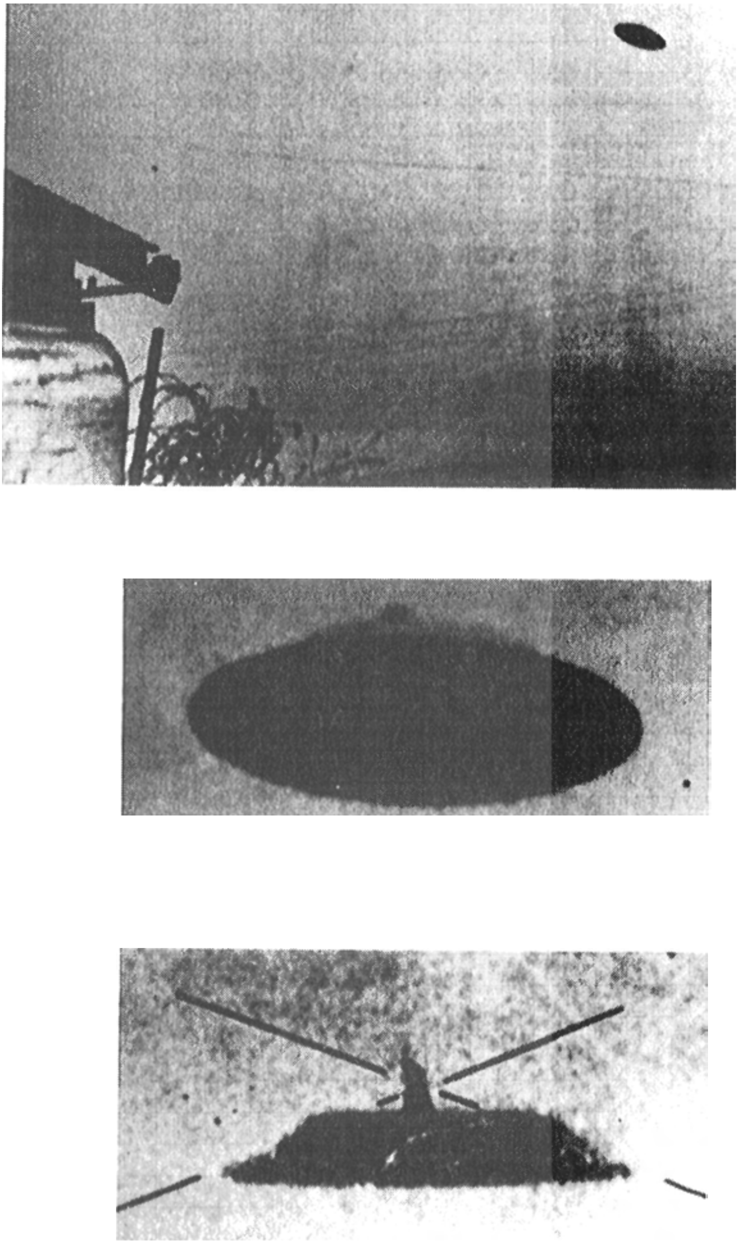

Photo 14. Oregon photos.

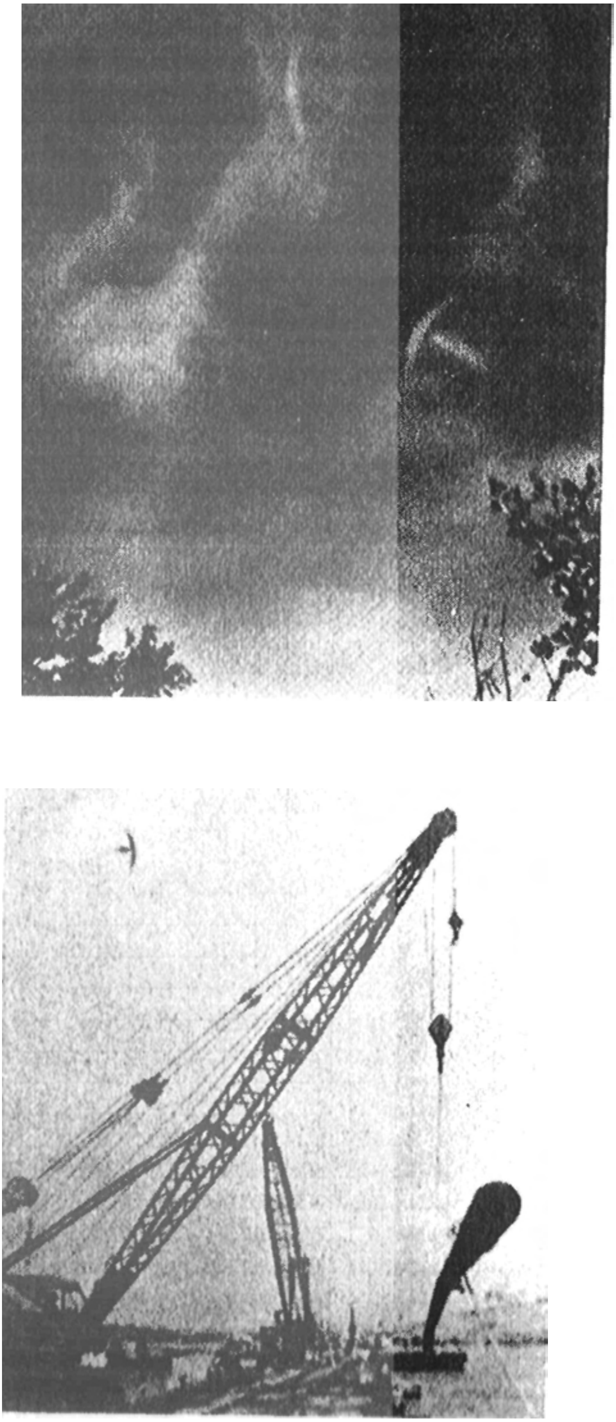

Photo 16. Anchor patterns.

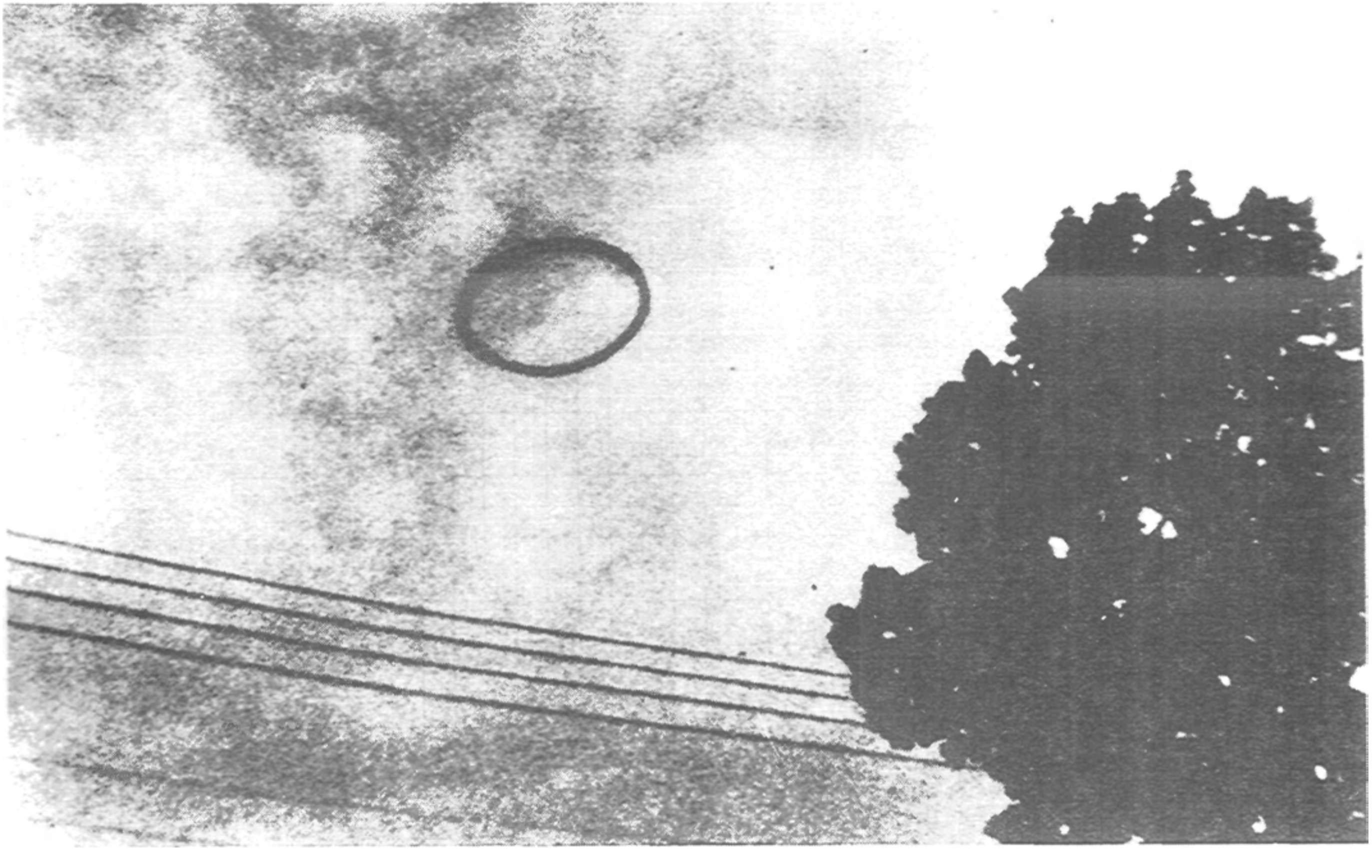

Photo 15. Black ring UFO.

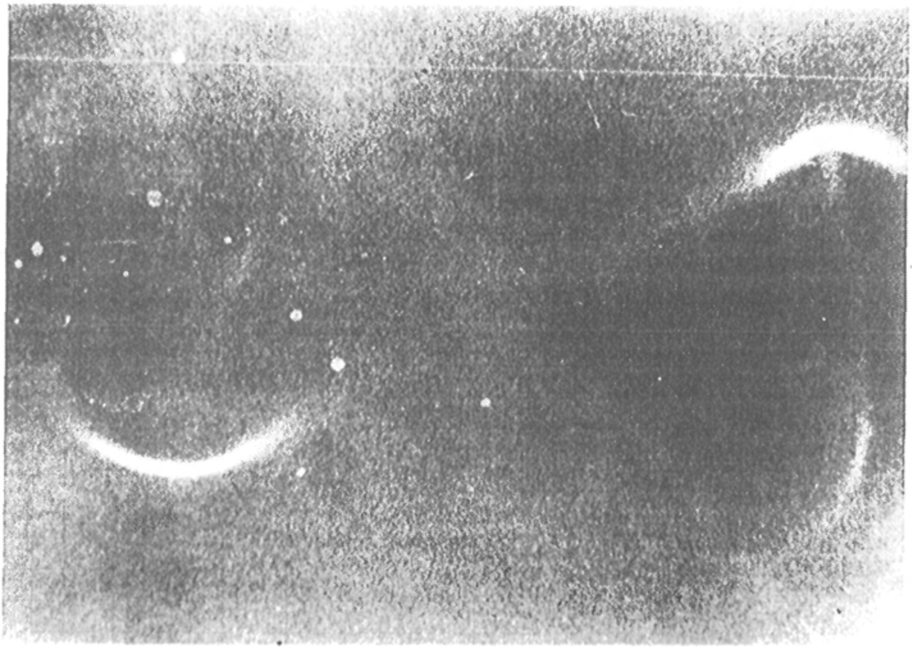

Photo 17. Bifringement.

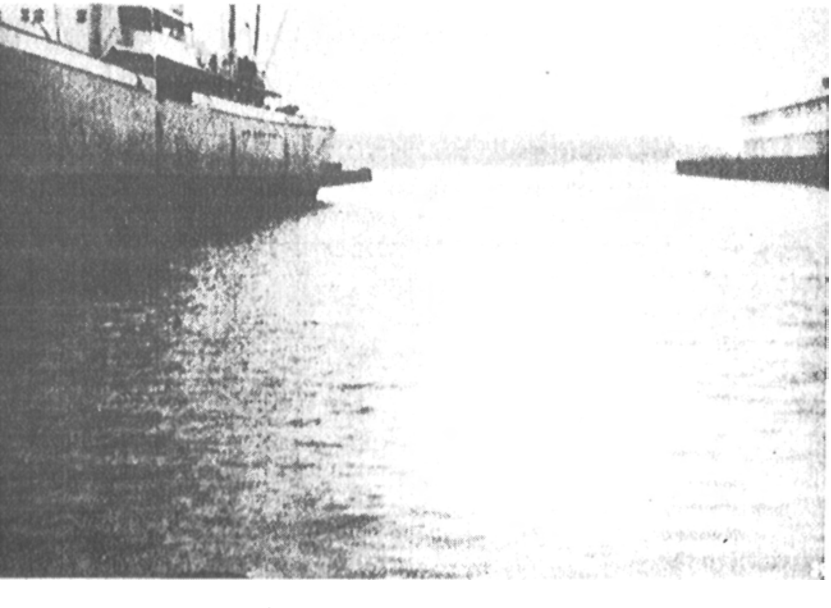

Photo 18. UFO In **water.**

The sighting occurred in June 1950. At the time I was home with my father in Woodville, Mississippi for the summer vacation from college. All my family were in the timber business — pulpwood and logs — and I was working for the summer with them.

It was a beautiful, warm, clear summer day with almost no clouds at all. Coming out of the woods on a pulpwood truck with the last load of the day (about 4:30 or 5:30 PM), the truck was slowly making its way across a large open field to the county road. Suddenly a light was seen in the extreme upper atmosphere, glowing with a fierce white radiance, very similar to the light emitted by a magnesium flare or by burning magnesium. Some object was apparently entering the atmosphere from outside it, but the velocity of reentry was far slower than reentry vehicles that would be fired many years later on U.S. and Soviet rockets. As the fiercely glowing light smoothly descended, a change occurred. At that point gray smoke also appeared in the white smoky trail being left by the object, and the trail became turbulent from that point on. Then an orange color appeared at the front of the descending, tumbling object, in the light being emitted. The smoke trail from this orange glowing color was darker gray. The object tumbled slowly downward, burning fiercely, until utterly consumed, still at an extremely high altitude (I would estimate still above 100,000 feet). In the sky the smoke trail that had been left persisted for a while, until it was gradually dispersed by high atmosphere winds. In the picture a great deal of dispersion has already occurred.

So far as is known to me, and I have had considerable missile experience, neither the U.S. nor the Soviets were firing any sort of vehicle in 1950 that could account for the phenomenon. Furthermore, even if we postulate a reentry vehicle from a large rocket, its velocity was far below normal reentry velocity, and this cannot be accounted for. The object also could not have been a meteor because its velocity was far too low.

For the doubter let it be pointed out that even if a high drag, low ballistic coefficient reentry vehicle comes into the earth's atmosphere on a downwards plunge and rapidly slows down, its initial velocity is quite high, and its rapid slowdown is quite apparent. This phenomenon did not exhibit such a characteristic.

7. The 1966 Dennis Billings UFO Sighting

At about 7:00 P.M. on the night of September 23, 1966, Dennis Billings and his wife and daughter sighted a UFO at close range. The close UFO was one of a total of four sighted by the family in about one hour. The sighting occurred in Decatur, Alabama. The UFO phenomena occurred after anomalous physical effects were noted. From the alley by his back yard Dennis, who had taken out the garbage, began to hear sounds from inside a neighboring house at some distance as if he were physically inside the room in which the sounds were occurring. These sounds consisted of conversation and dish-rattling in the kitchen where two women were apparently cleaning up the table after the evening meal and washing dishes in the sink. Then the sounds changed to those from another room where children apparently were playing games. Peculiarly, the sounds seemed to originate from wherever Dennis focused his attention, such as the garbage can, the house, etc.

Shortly afterward his daughter followed him into the yard as he left the house to go to a football game, carrying his binoculars. She spotted a cluster of lights moving along slowly in the sky above the treeline, and called his attention to them.

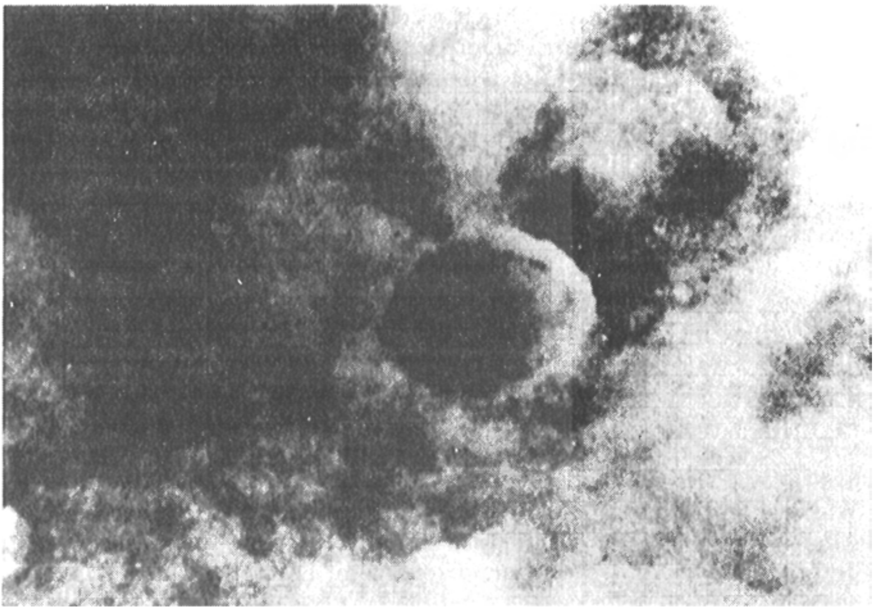

Photo 19. **UFO** in clouds.

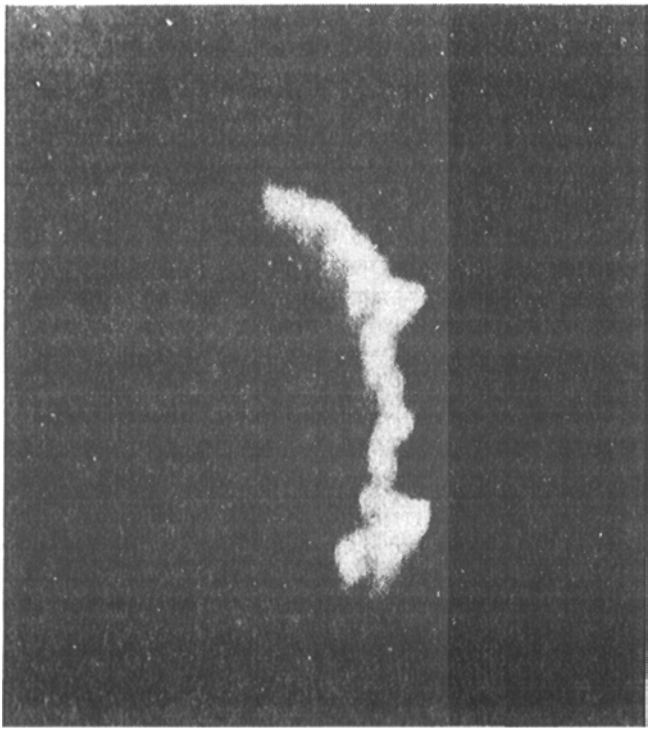

Photo 20. UFO seen by the author.

Through binoculars the cluster of brightly colored, flashing lights could be seen to be one object about two blocks from Dennis's house, moving very slowly and making no sound, about fifty feet above the trees.

Just then Dennis's wife turned on the exterior house lights which illuminated the entire back yard. As she came out into the yard, an object, a fiery red, glowing, round ball of light about twenty-five feet in diameter, appeared just above the neighboring house in back, the same house from which the anomalous sounds had been heard a few minutes previously. Strangely, the glow of the object did not appear to illuminate the rooftop of the dwelling above which it hovered.

The glowing light seemed to be composed of a tremendous amount of energy, and there was violent fluctuation, turbulence, and movement of the light energy within the sphere. Another way to describe its dynamic characteristics is to visualize it as being composed of individual particles of light or incandescent particles of highly charged matter moving rapidly in a circular plasma. The light itself appeared to be comprised of hundreds or even thousands of individual light sources confined to the circular shape.

Within a few seconds the individual lights began to coalesce and form into a solid light with no agitation, dimming in incandescent intensity and changing from red to orange to yellow. As the circular shape of light grew dimmer, a rocket-shaped object formed in the center. After the light had finished changing to yellow and was much dimmer, the rocket-shaped object began to rise, with particles of light from the sphere clinging to it and falling off toward the ground, where they evaporated and disappeared. The rocket-shaped object had no wings or tail fins, and it picked up speed and sped away, streaming an exhaust tail of light or flame behind it, similar to a welding torch's flame or a flaming sparkler. The object made several spirals as it rose and then headed in a straight line toward the stars. As it rose, it also made a slight arc toward the moon, which was full or nearly full, and passed across its face. As it crossed the face of the moon, its rocket-shaped outline could be clearly seen.

Astonished, Dennis and his wife began discussing what they had seen, still standing in the back yard. At this time the anomalous reception of unusual sounds began again. At first Dennis seemed to hear a train on a distant track although there are no railroad tracks in the vicinity. The sound seemed to come closer until he seemed to hear a train passing by immediately in front of him with its wheels clicking on the joints of the steel rails. Then they heard other sounds similar to automobiles and trucks on a freeway, again very close, although there is no freeway in the vicinity either. The whining motors and honking horns of these unseen vehicles could also be heard. Turning first in one direction and then in another, Dennis could hear the anomalous sounds seemingly originating from wherever his attention was directed. Other sounds such as airplanes flying in the sky also came in. Then he heard another sound like a high-pitched tuning fork, and it began to increase in frequency until Dennis's eardrums were hurting.

With this high-pitched frequency sounding and his ears smarting, Dennis looked up and saw a bright object positioned in the sky, larger than a star, and approaching him and his wife. As it approached on a straight line toward their back yard, it got bigger and brighter, until it stopped directly overhead, hovering 100—150 feet directly above the yard. The object appeared to be circular or oval in shape, and it descended with the flat side down. On its rear end the object was

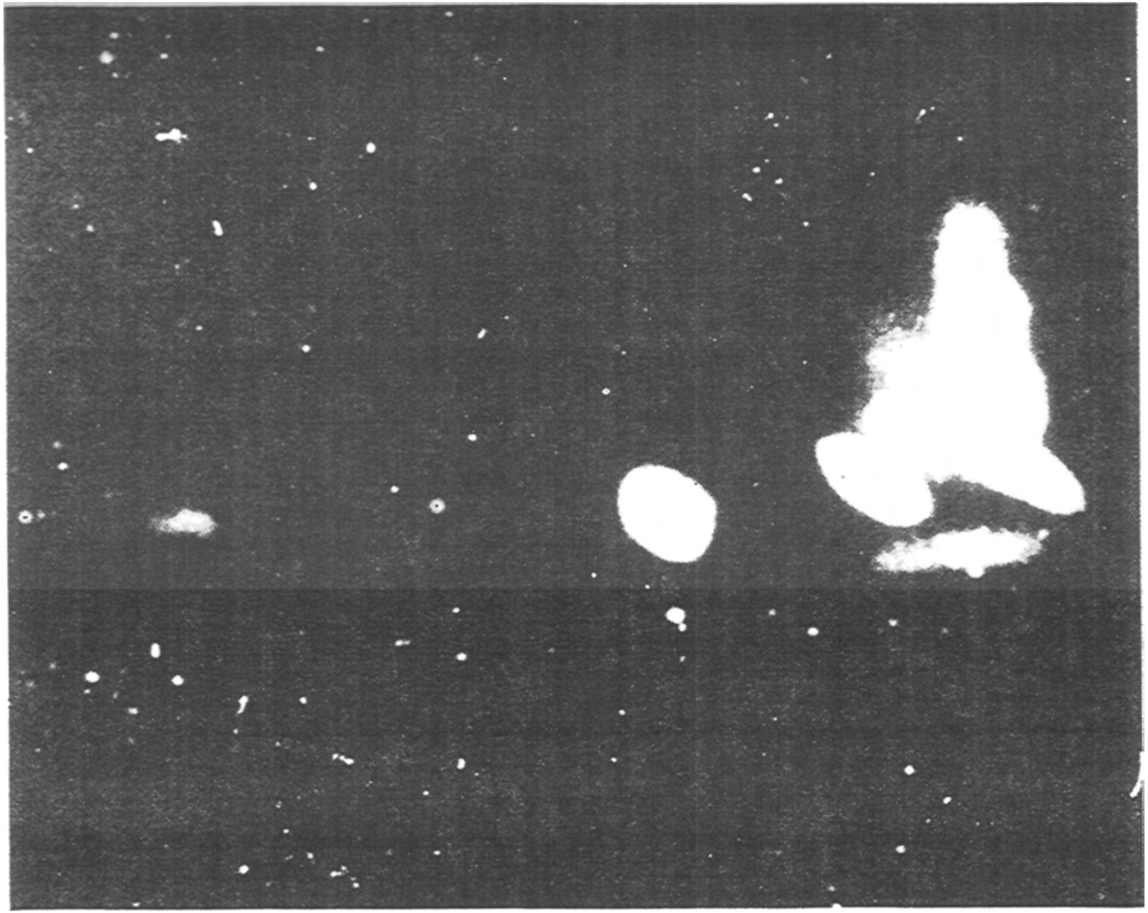

Photo 21. Second Billings UFO.

rimmed like a horseshoe with twelve washtub-sized blinking lights, each of a different color. These lights appeared to be controlled so that as one light blinked on, the other lights went off. In this way one light at a time blinked on and off in successive positions, so that a single light seemed to rotate around a complete half circle and cross at the open end, changing color as it progressed. On one side of the horseshoe the light positions were bright red, on the next yellow, on the next pale green, on the next a darker blue-green, and the final light on the opposite side of the horseshoe pattern was a very bright blue. Underneath the center of the object was what appeared to be a hemispherically-shaped plastic capsule or shield. This capsule was completely blacked out; yet through what appeared to be smoked glass, movements or shadowy activity could be seen. Dennis assumed that this was the exhaust section of the craft and that what he was seeing was probably smoke coming out of a black hole in the craft.

Around the outer perimeter of the craft, and about two feet out from it, was an illuminating light which circled the object like a halo or corona. Dennis had the eerie sensation that the object was intelligently controlled and that he himself was under close observation.

At this time Mrs. Billings became extremely frightened, ran into the house and got the baby, and crawled under the bed until the craft had gone.

As she left in fright, the craft lifted up in the sky a little and began to move slowly away, so slowly that Dennis could walk along beneath it as it moved across the back yard and into the front yard. Walking under it and looking straight up at it, he was able to observe closely the blinking lights and their operation. When the craft had been motionless, the lights had blinked on and off slowly, so that they moved in a slow, pulsating motion from position to position around the horseshoe pattern, changing color as they went. As the craft began to move and pick up speed, the lights began to increase in pulsation frequency and thus to appear to move faster around the horseshoe pattern. As the craft moved even faster, the light pulsations became so short and fast that they seemed to be spinning around the horseshoe in one continuous blur of motion.

As the object left Dennis's yard, it moved into the sky and away, but not too far, until it had positioned itself over the airport (Pryor Field) as a large bright object. At this time Pryor Field was serviced by Southern Airlines. As one of the commercial aircraft took off, the UFO followed it back directly over Dennis's house, then left the aircraft and repositioned itself in the sky in its original position over the airport, hovering motionlessly.

After hovering motionlessly for about fifteen minutes, the UFO suddenly shot down in an arc toward the ground, passing below the tree line and out of Dennis's view.

Dennis waited for about ten more minutes, when suddenly the UFO shot back up into the same position in the sky, seeming to follow in reverse the same arc on which it had descended. It remained in this position, hovering in the sky above Pryor Field for another ten to fifteen minutes. Then a similar object approached and joined it, and the two seemed to merge or dock together. They remained motionless for about five minutes more, then moved across the sky, still joined, until they disappeared below the horizon.

According to the newspapers the next day, this object or something similar had been seen all over the eastern half of the U.S., and had been described as "multicolored clouds."

8. The Second Billings UFO Sighting

If Dennis and his wife had experienced only one UFO, the incident would have quietly faded away in the unwritten annals of ufology. However, such was not to be the case.

Dennis vowed to himself that, if such an incident ever occurred again, he would get a photograph of the object. And just before Labor Day in 1970, Dennis got such a chance, and indeed made good his intention.

On that night, with almost totally clear skies and a bright moon, Dennis's wife came rushing into the house in Decatur, Alabama and excitedly reported to Dennis that she had seen another UFO. Rushing into the yard, all Dennis could see were some distant white lights — too far away to say positively they were UFO's. However, recalling the multiple UFO's in the previous incident, Dennis and his wife took up separate watches on each side of the house. Dennis also thought of his Polaroid camera left in his car parked in the driveway, so he went to get it.

Standing between the trees, Dennis looked up into the sky and saw a UFO suddenly triggered into position, as if his thought and intent had pulled a switch and the object had suddenly appeared, positioned perfectly for his observation, pouring light from underneath for a distance of about twenty to thirty yards. This light seemed to emit some immense force and was colored like the fire, whiteness, and glare of a welder's torch. The tail diminished as the craft lowered. Dennis snapped a picture when the craft's tail had shrunk to only four or five feet long. He pulled the tab on the polaroid film and ran into the street under an outdoor light to see if he had recorded a UFO. When he saw he had done so, he then noticed the object had descended and made a left turn, moving eastward down Nelson Street slowly at about five miles per hour. Its porthole lights were still clearly visible and no sound was being made. Dennis hurried after the UFO and caught up with it at the end of the street near a T.V.A. easement, where large electrical power lines were located, running generally north to south. He got slightly ahead of the object as it traveled parallel to the power lines and as the craft passed directly over him. At the very instant he pressed the shutter, all lights in the craft instantly extinguished as if his mind had been read and the photograph deliberately prevented. The now darkened, cold, lifeless-appearing craft moved smoothly away, passing over Mallard Park where a Marine Corps Band was playing. The second photograph showed no lights whatsoever.

One of the local newspapers then printed the story and the picture, and Dennis received a call from one of the national UFO agencies. After giving the agency his photograph, his camera was mysteriously stolen from his automobile, and the agency then claimed to have lost his photograph. However, they underestimated the determination of Dennis Billings.

Dennis simply got on a plane and went to Washington, and proceeded to the office of the head of the UFO agency. The president of the organization had said that he would be out of town for several days and would be unable to see Dennis. Nevertheless, Dennis went to his office, where his secretary repeated that the gentleman was out and would not return for the day. Dennis barged into the office over the protestations of the secretary — including the threat to call the police and have him removed — and proceeded to search for his photograph, eventually finding it in the desk drawer of the president of the organization. He retrieved his photo and returned to his home. Without his personal determination to retrieve his property, the photograph — like so many others over the years — would have been "lost." The immediate question that occurs is: Isn't it curious that one of the major UFO organizations in the U.S. — supposedly devoted to solving the UFO mysteries — exhibited the desire to misappropriate a really good picture of a UFO and see that it disappeared from the scene?

At any rate, Dennis got his photo — and a beautiful photo it was. Furthermore, the incidents experienced by Dennis Billings are classic examples of my theory. Many anomalies in the incidents will not fit orthodox theory or orthodox ideas about reality. The "pieces of light" seem to be almost intelligent. There was a close coincidence of thought/intent/expectation and kindling burst into observable reality, plus a great deal of sky-watching and expectation by Dennis between the first UFO incident and the triggering of the second (i.e., a long kindling time, just as Patterson experienced before he got his excellent movie shot of a sasquatch). And there was close experience of very distant sounds (to be expected in hyperspace, where ordinary distance can be shortened or even eliminated). Furthermore, there were multiple witnesses to the UFO incidents, and Dennis got his photograph. Psychologically, it may even be significant that for several years Dennis vowed to take *a* picture of any other UFO he might later see, and when he got *a* picture, he was prevented from getting another. And on the photograph obtained, one can see that the "craft" form of the tulpoid is only partially materialized, since the stars can be seen right through the vehicle's midsection. This is similar to the strange transparency of the Doc Shiels photograph of Nessie (the Loch Ness monster), which Ground Saucer Watch found when they performed a computer analysis of his photo. (The second Shiels photo showed that the water was being forcibly moved by Nessie, so the tulpoid was at least partially materialized.) Yet such is to be expected if one is dealing with partial materialization of tulpoids.

Finally, the entire area is located in the Tennessee Valley Authority's project. Great electrical transmission lines go through the area, and there are many long lakes and the Tennessee River. So an excellent kindling medium is provided, and many UFO's are seen over the Tennessee Valley.

The Dennis Billings photo remains one of the best UFO photographs ever taken, and coupled with Dennis's astute personal observations, it provides a keystone insight into the actual nature of the UFO phenomena.

9. The 1973 Pascagoula UFO Sighting

In October 1973 one of the greatest UFO flaps of all time occurred in the United States, and created headlines because of the bizarre increase of contactee incidents. One of the most sensational of these was the purported abduction of two men in Pascagoula, Mississippi, who were allegedly taken aboard a saucerlike craft, examined, and released.

Shortly afterwards another incident happened in Pascagoula, which drew little attention. In this case, underwater phenomena were involved, because two U.S. Coast Guard petty officers investigated the object and made physical contact with it. An official Navy investigation was made and a report written. Unusual characteristics, highly suggestive of a device using psychotronic effects, were displayed by the object. I suggest that this device may well have been a Soviet psychotronics device, taking advantage of the UFO flap as a screen to carry out clandestine activities.

In Pascagoula an illuminated underwater object followed the boat of Raymond and Rayme Ryan, who tried to beat it away with an oar. The nine witnesses, including two Coast Guard investigators, described the object as metallic, rod-shaped, two to three feet in length, and four inches in diameter. It was attached to an oval-shaped beam ahead for four to six feet. It looked like a mini-submarine and moved at four to six knots in four to six feet of water. Although it sounded like metal when struck, it could not be struck consistently, and its light, when illuminated by a flashlight, turned off immediately and then on again after about one minute. The portion of the Ryans' oar that was under water could not be seen from the surface when the object's light beam fell on it.

At this point I would like to make some general remarks about the nature of UFO's. (1) The UFO-generating mechanism and the psychotronic mechanism are the same. (2) The Soviets may even be able to psychotronically evoke particular types of UFO phenomena at will. (3) A small number of UFO phenomena could even be generated by the Soviet Union as part of its psychotronics deception plan. However, most UFO phenomena are definitely not of Soviet initiation, (4) An explanation of the UFO phenomena is briefly sketched out in my section on psychotronics. (5) Some of the NASA photos taken of moving lights in orbit around the earth may have been of Soviet-initiated effects. Like the radiation of U.S. personnel in the U.S. Embassy in Moscow, this effect could have been intended to evoke U.S. high level reaction, to positively assess whether the U.S. had any knowledge of psychotronics, and if so, what that level might be.

10. Other UFO's

A brilliant, "crystalline-like" UFO was seen over Madrid, Spain in 1968 and was chased by Spanish Air Force planes. Whenever people are significantly structured in their thinking, symbolically "sharp" and "crystalline" tulpoids often are seen, representing their ideas. Many such UFO tulpoids are seen over Iron Curtain countries, and over countries such as Spain that have one dominant religion and culture.

In October 1973 the governor of Ohio, John Gilligan, saw this UFO phenomenon along with over 100 other people. The police were notified by the governor, and this photo was taken by the *Columbus Evening Dispatch.* Note that the tulpoids remain in the first bioframe (see Part Two), and even though their intersections change shape, the second one remains in a "burst mode."

In 1965 Rex Heflin took four pictures of a highly materialized UFO near Santa Ana, California. In the fourth photo the dematerialization "photon absorption" effect is shown: if orthorotation into higher dimensional space is made in one direction, photon (light) is emitted. If the orthorotation is made in the opposite direction, photons are absorbed. In one case "glowing light(s)" are associated with the UFO's hyperdimensionality, while in the other case "black rings" or "smoke" are associated.

FALKVILLE SPACEMAN

These two pictures are black-and-white versions of two of the photos taken by **Falkville, Alabama Police Chief, Jeff Greenhaw, in an encounter during October**

1973. Greenhaw had received a call from one of his constituents to say that a UFO had landed in her pasture, and he was proceeding to investigate the incident. In his police cruiser he carried a polaroid camera which he used in his police work.

Driving down a narrow gravel road, Jeff noticed a metallic-clad, suited figure walking down the road toward him, and he stopped his car to get out and have a look. For some strange reason his instinct was not to pull his pistol on the creature, and instead he started taking pictures as the creature/tulpoid walked toward him. The creature was holding something exposed in one of his hands, as can be seen in one of the photographs. As the tulpoid approached with no apparent intention to stop, Jeff then switched on the blue light above his cruiser, and the rotating, flashing light caused the creature to turn about and run away down the road. Jeff jumped into his cruiser and gave chase, eventually reaching about forty-five miles per hour without catching the tulpoid. His car spun off a curve and Jeff still had the photos — polaroids that he developed.

Bizarre things then happened to Jeff Greenhaw. His wife left him and divorced him (I suspect this had nothing at all to do with the UFO incident), he was fired from his job as police chief, vicious rumors were circulated that it all had been a hoax, and that a person had been dressed in an aluminum suit and photographed to give the appearance of a spaceman. Various fire suits were positively identified as having been the suit used. Jeff was discredited. Finally his trailer was burned down.

Before the trailer was burned, a friend of mine here in Huntsville, Warren York, visited Jeff and made photographic slides from the original polaroids with Jeffs permission. Shortly after that I had a chance to study these first generation slides for several hours, and some interesting things turned up.

The tulpoid's left hand is positively reversed in a mirror image fashion; it opens outward instead of inward as does a normal hand. (This cannot be seen on the photos shown here, but on the first generation slides it is readily seen.) Being an aikidoist, I am totally familiar with the degree of bending and flexing of most human joints, particularly the arms, wrists, and hands. And the mirror imaging of the left hand of the tulpoid is beyond question.

Furthermore, one can establish a clear shoulder line in some of the other photos, and a line just under the hat or "nazi-helmet" portion of the figure. From the other photos a remarkable thing emerges: there are only two places between the edge of the helmet and the chest that are sufficiently wide and sufficiently thick to contain a human head. The person would either have to have a neck about twelve inches long, or he would have to have a neck shorter than a Neanderthal's, with the head lolling down completely on the chest. Again, this can be clearly established in the original first-generation slides.

Finally, I have a letter in my files that establishes as a falsehood one of the rumors about Jeff. One person had admitted dressing in the suit and faking the incident, when in fact what actually happened was that Jeff went to his house immediately after the incident to ask this person to accompany him back to the area and search for the spaceman. When confronted in late 1977 by his family and friends, this person broke down and confessed that he had lied about the incident.

Finally, some other investigators dressed a person in an aluminum foil suit and photographed him, and there is no resemblance to the tulpoid photo. And Warren York investigated all the firesuits supposedly used in the hoax, including new suits used by the various fire departments and also by NASA at Redstone Arsenal. None of them resembled the tulpoid photo.

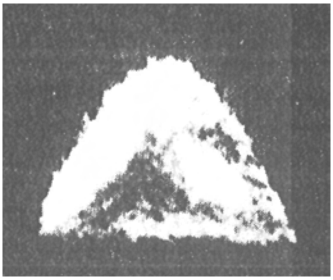

Photo 22. Spanish UFO.

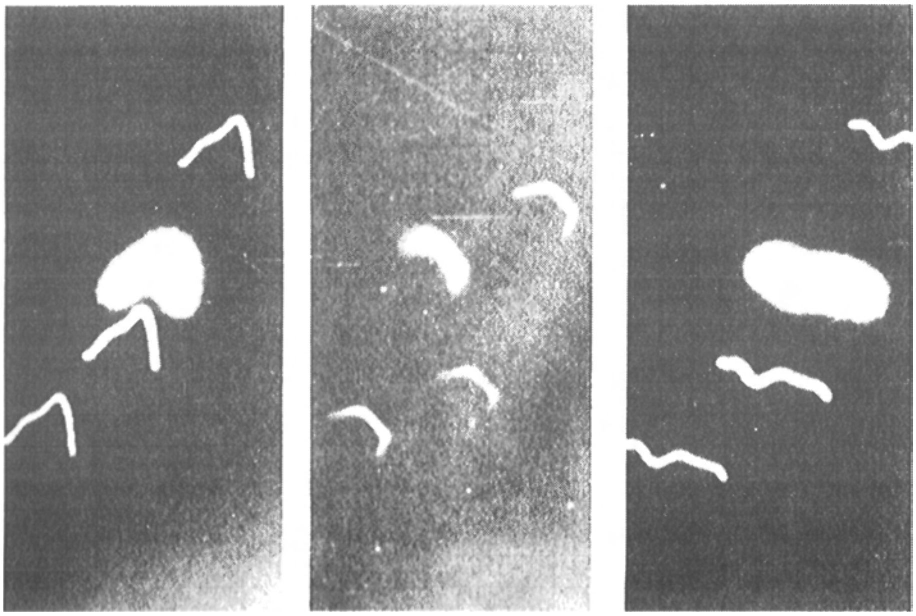

Photo 23. Ohio UFO.

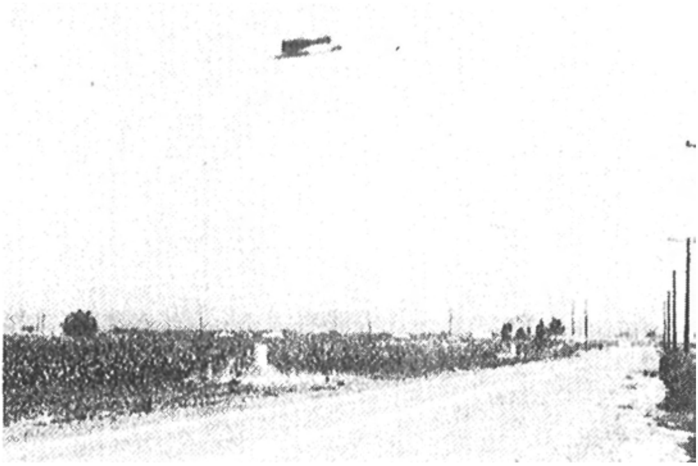

Photo 1

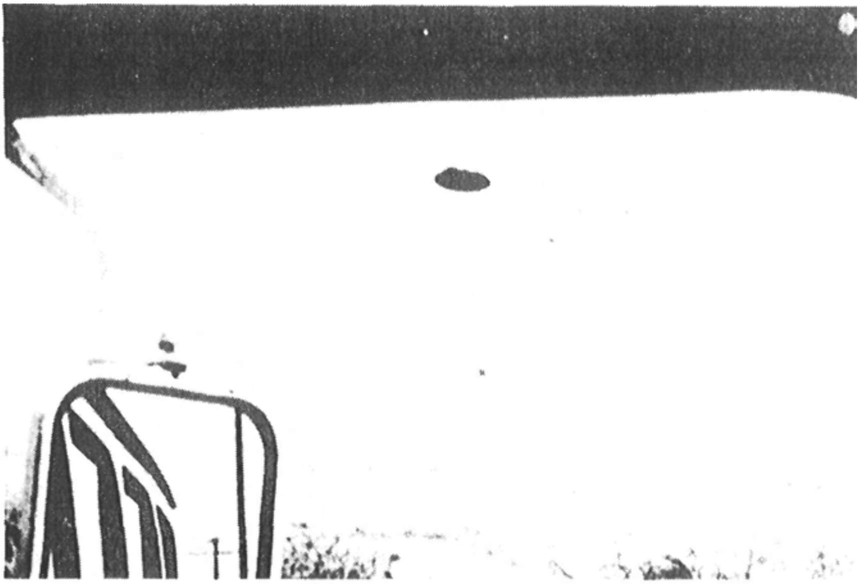

Photo 2.

Photo 24. Heflln UFO.

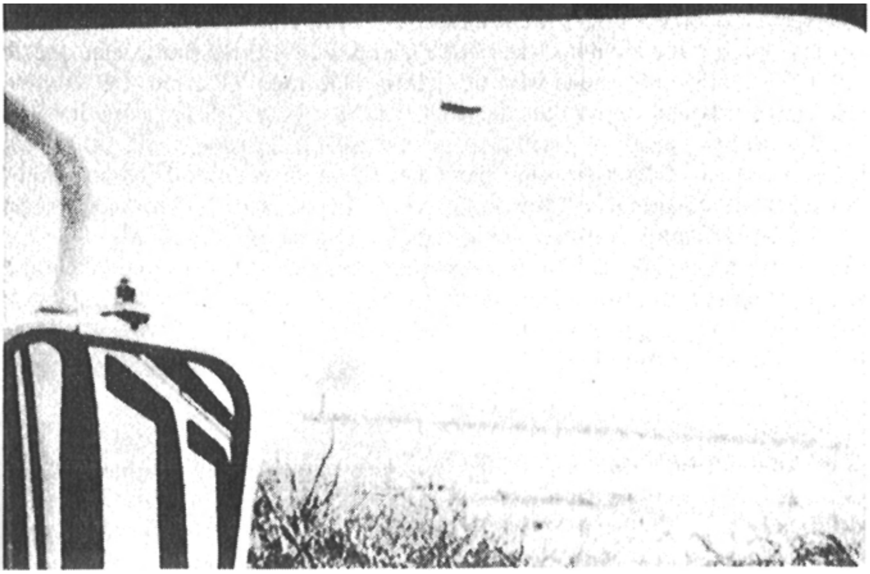

Photo 3

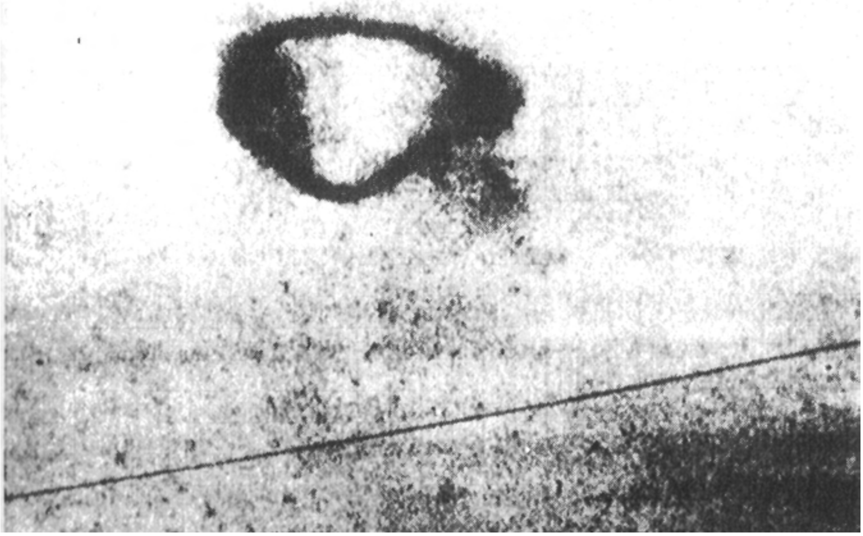

Photo 4 (enlarged)

Several photo experts have concluded that the photos are false, and several have concluded that they are not. A computer analysis concluded that they were false. However, what exactly could a computer show that would differentiate between a normal physical object and a well-materialized tulpoid?

In my opinion the incident was a fake *if and only if* three things can be established; (1) an individual exists with a horribly deformed left hand that is mirror-image backwards and opens outside away from the body; (2) the same individual either has no neck at all or else has a neck almost a foot long; and (3) the same individual can run forty-five miles per hour, faster than any other human in the world. Such an individual — if he existed — would be readily identified, and there would be no difficulty in ascertaining that he was the one involved.

There are other things about the case that I am not at liberty to divulge at this time out of respect for Jeff Greenhaw's confidence and privacy. But even so, the incident has a very high probability of being tulpoidal, and Jeff Greenhaw stands fully exonerated, in my opinion.

MEN-IN-BLACK

The kindling of virtual state tulpoidal forms into the zeroth bioframe of a physical reality (materialization) can result in the materialization of living, breathing, functioning forms. There is essentially no limit on the format that may be obtained. Sasquatch, Nessie, fairies, demons, angels, mothmen, skunkmen, yeti, kangaroos, etc. are just a small sampling — as indeed are UFOnauts such as little green men, space-suited astronauts, androids, robots, etc. These (usually) fleeting phenomena are physically real when in the materialized state (zeroth bioframe); unfortunately the tuning which is accomplishing the collection and kindling is usually unstable, and the stability to maintain the materialization lasts only a short time.

One of the consistent phenomena experienced by many UFO investigators is that of men-in-black. Typically, oriental-featured men in dark suits appear and demand the investigator's material, instructing him to cease his efforts. Usually they have almond-shaped eyes that never blink, and speak telepathically without moving their lips. Variations, of course, occur.

Here is an incident that happened to a friend of mine in the city where I live (Huntsville, Alabama). At the time my friend was an ardent UFO investigator and even wore a radiotelephone so that he could be contacted quickly at any hour to go and investigate UFO incidents in the field. He is the friend who investigated the Jeff Greenhaw case and took slides of the original Greenhaw photos before Greenhaw's trailer was burned.

A friend of his, a girl with psychic abilities, was approached one evening in her apartment by two medium-sized, almond-eyed men of oriental cast, in typically dark suits. These tulpoidal gentlemen's eyes never blinked and they spoke to the girl telepathically without moving their lips. They handed her an envelope containing two ancient coins and instructed her to give the coins to my friend, and inform him that he must keep them on his person at all times. After other messages that completely frightened the girl, the two tulpoids left, never to be seen again.

In tears from fright, the girl contacted my friend and between fearful sobs urged him to take the envelope and get it away from her, as she never wanted to see it again or the two men who had brought it.

My friend picked up the envelope and later examined the two coins. I also personally examined them: to anyone who is sensitive, the coins radiated an invisible unpleasant tulpoidal energy.

The investigator then photographed the coins, and daringly placed one on a string around his neck against my own advice to seal them up against light. He promptly broke out in a rash from head to foot and was shortly forced to remove the coin and seal it away. Interestingly, photos of the coins seemed to have about half the effect of the actual coins.

For a period of two years I kept my copies of the photos sealed against light and stored in a file cabinet. To my surprise, several times when I went to retrieve the photos, they would alternately be there and not there. I simply state what occurred, without any change or justification. Eventually, however, their effect seemed to decay, and at that time there was no longer any problem in recovering them from the filing cabinet.

The men-in-black syndrome is based on our own unconscious tuning; since each of us has some unpleasantness in the unconscious, sometimes the tuned-in men-in-black can be very nasty indeed.

I am personally aware of several good researchers into the paranormal who have been forced to retire from the field to save their own lives and the lives of their families. As late as April 1978 a very good friend of mine has been forced to do just that because of overwhelming danger to his family from materialized tulpoidal phenomena.

In my own case I have found it expedient to work closely with a group of psychic friends in the American Association of Meta-Science. This group is capable of clairvoyantly peering into the unconscious to ascertain the most unpleasant blockages present, and forming energized thought forms to go in and cleanse the unconscious of most of the unpleasantness. In this way the unconscious can be kept relatively free of unpleasant tuning that might otherwise tune in something like the men-in-black. My own life and experience has not been free of tulpoidal tuning; however, it has been free (at least so far!) of tuning in anything really unpleasant, or nasty, or threatening to my family.

But I continue to be very grateful to my psychic companions for their understanding and assistance, and I continue to avail myself prudently of their services when needed.

SASQUATCH

In many regions of the earth there exist legends of giant, apelike creatures of various sizes, hues, and dispositions. In the high Himalayas this takes the form of the yeti, a humanoid creature sometimes seen in the snows of high and remote fastnesses. Expeditions have in fact uncovered and photographed the yeti's tracks, and sightings by reliable observers have been made. Several parties have also reported hearing the yeti's strange vocal sounds.

In the remote Siberian wastes the Soviets have also reported such a manlike humanoid creature, even purporting to have captured one a century or so ago.

In the U.S. the tulpoidal form similar to this has been the sasquatch of the Pacific Northwest, known since Indian times, and observed hundreds of times by a variety of persons over the years. Nevertheless, in spite of intensive searches Bigfoot continues to elude all pursuers while continuing to be sighted and leave footprints.

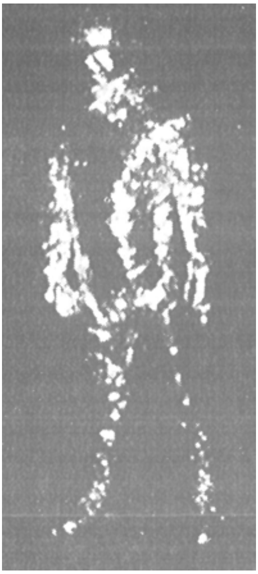

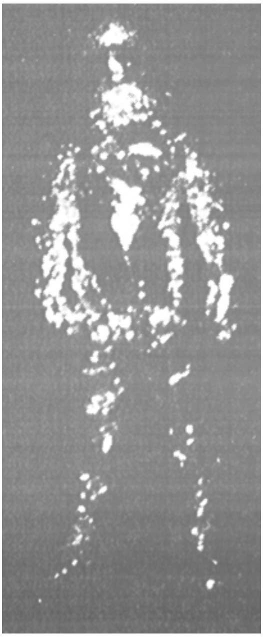

Photo 25. Falkville spaceman.

But Bigfoot cannot be caught by packs of hounds. (This has repeatedly been tried. His lair is never uncovered, nor can he be tracked to it. He has never been found dead of old age or injury. The bones of sasquatch have never been uncovered in caves or elsewhere.)

Furthermore, not just one type of footprint or one type of creature is seen; instead, a variety of shapes/sizes/types of creatures appear. This is one particular clue to the fact that we are not dealing with a species of shy and elusive wildman, but instead with a tulpoid.

Sasquatch cannot normally be found — as indeed Nessie and the yeti cannot — because shortly after each appearance he is no longer present in this bioframe.

However, as interest in sasquatch/Bigfoot grows and the effort to discover or uncover him intensifies, the kindling of the tulpoidal form intensifies. Today bigfoot-type creatures of all varieties and hues are being seen throughout the U.S., from Maine to Florida, from Alabama and Tennessee to Oregon, Washington, and California. With this intensity of kindling, it is getting easier to evoke a bigfoot-type tulpoid, and as it continues, sasquatch is probably going to be tuned in and stabilized. If so, the past itself will just be changed, and we will be able to dig up ancient bones of sasquatch, and find his lairs, and track him down with dogs and capture him. In that case, sasquatch will become a respectable animal and will stabilize in our reality. Tulpoidal activity from the collective human unconscious can be stabilized and the entire past, present, and future changed by the awesome 4×10^9 dimensional psychokinetic mastery of spacetime exerted by the species unconscious.

We also should not be at all surprised if the same thing happens to the placid Loch Ness monster. There is probably going to be a family of pleisosaurs living in the Loch eventually (and in many other bodies of water across the globe) — and spacetime will be changed (past, present, and future) in our zeroth bioframe to accomodate this small shift in the collective human unconscious. In reference to changing the past, it is interesting to note that Wheeler has recently demonstrated that, in the two-slit experiment (which contains the basis of quantum mechanics), it can be decided *after the fact* whether or not a photon has interacted with two slits or with one. Thus in at least one sense the past can be changed!

After several years' intensive search for and pursuit of the elusive Bigfoot in the Pacific Northwest, Roger Patterson finally obtained a short movie strip of a very large, seven-foot tall female with pendulous breasts and crested head, walking through a forest clearing with giant strides. This is perhaps the best frame from that movie strip. Sasquatch tulpoids have been known since ancient time by the Indians who lived in the area. It would be interesting to see what results could be obtained in materializing or partially materializing the Bigfoot tulpoid with a Reichian cloudbuster.

Bigfoot-like creatures are materializations from the collective human unconscious shaped and formed by the layers (race, culture, nation, territory, state, region, family, etc.) between the collective unconscious and the individual unconscious minds of the observers. The exact format depends on all the higher level "kneading" of the "dough" that is being forced out into materialization.

In the Pacific Northwest the Indian knew of the sasquatch long before the coming of the white man; in that region, a great deal of tuning has been accomplished, and the sasquatch is a rather well-tuned animal form, having well-

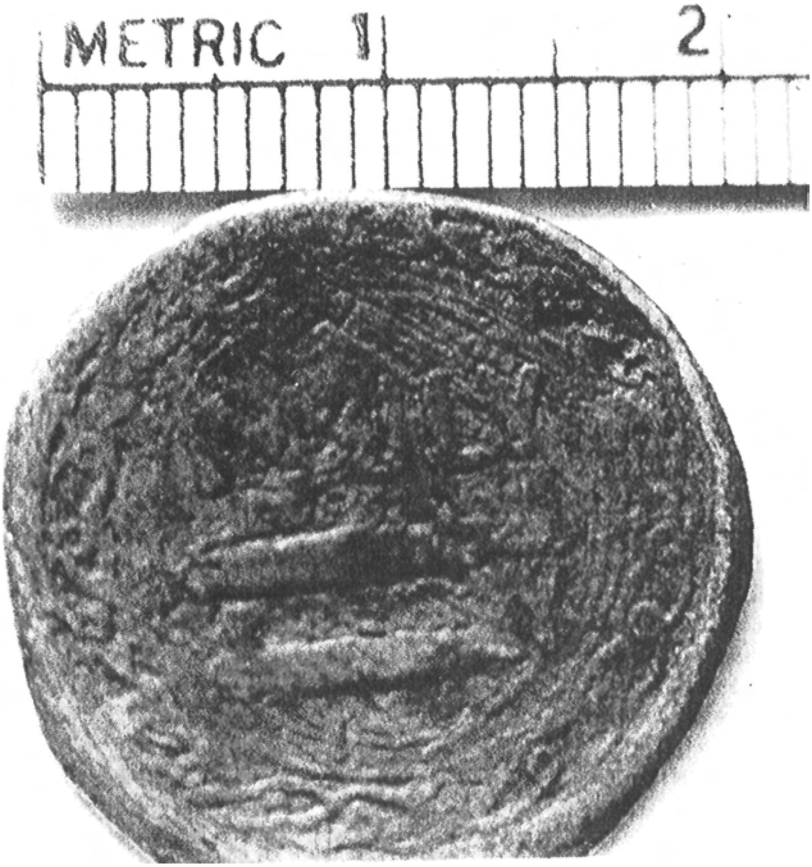

Photo 26. An ancient coin handed to a psychic friend of a UFO investigator by tulpoidai men-in-black.

defined footprints as shown in Photograph 28. In the East, the yeti has also been tuned for centuries, and a cast made from his footprint is quite well formed — though also quite different from that of a typical Pacific Northwest sasquatch. In New Jersey, on the other hand, for the last two or three years there have been a large number of Bigfoot sightings, but here there has not been the same degree of tuning for several centuries, and consequently the footprints look rather messy, like blobs left by the creature before it vanished back into dematerialization. Finally, in one case a person well-known to me stumbled into dematerialization phenomena of a very strong nature. One of the incidents involved the materialization of a really gigantic "superfoot" creature, which walked around all over the base of a mountain, leaving giant footprints two feet or so long. Though monstrous in size, superfoot is really ill-tuned and thus ill-defined, with footprints that are decidedly blobby in form.

UFO's OVER THE SOVIET UNION

In 1961 locomotive engineers on a freight train from Moscow to Golytsino, when about forty kilometers from Moscow, sighted a disc-shaped object topped by a cupola approaching the train. The object slowed to pace the train and escort it for about one kilometer. Suddenly the locomotive engines went dead, and could be restarted only after the mysterious object disappeared.

On December 29, 1967 an Iliouchine 14 aircraft enroute from Zaporoje to Volgograd sighted an approaching UFO. As the UFO neared the aircraft, suddenly the aircraft's engines stopped in midair and the plane began slipping sideways toward the earth, carrying its panic-stricken crew to certain destruction. When the crippled aircraft was about 700 meters from the ground, the UFO abruptly disappeared. At the same instant the aircraft engines sputtered and roared to life again. The pilot righted the aircraft as lift returned, narrowly averting a crash.

On September 24, 1962 first one oval object then several others escorted a TU-104 airliner enroute from Vorkouta to Omsk and maneuvered around it with astonishing speed and agility. The eerie aerial escort continued for most of the flight, and the objects were extensively observed by both the passengers and the crew. The objects positively could not have been other planes, balloons, satellites, celestial objects, or other known natural phenomena.

In Latvia in 1965, R. Vitolniek, a well known astronomer, together with several scientific colleagues, sighted a triangular-shaped formation of four spherical objects, one at each of the points of the triangle and one a little larger in the middle of the triangle. Two of the scientists, Ian Mederis and Robert Vitolniek, made immediate calculations and insured that the objects were not satellites or normal orbital objects. The objects steadily shrank in size, leading Vitolniek and his colleagues *to* conclude that the objects were traveling away from the earth into outer space.

At least two Soviet astronauts, Colonel Bykovsky and V. Terechkova, have observed UFO sightings and reported them to the Soviet authorities. Many other UFO sightings have occurred in the Soviet Union and the Iron Curtain countries; **these** are just a sampling.

In fact, on several occasions Soviet defense forces have been committed into massive action against UFO's.

In 1959 numerous UFO's flew over Soviet rocket launchers at Sverdlovsk. In addition to being visually observed by a large number of personnel, they were detected and tracked on Soviet radars. Soviet military units were ordered into action immediately. Interceptors and antiaircraft artillery batteries went into action and the air defense missile units opened fire. In spite of the fierce engagement of the elusive UFO's, none were hit or damaged and no results were obtained by the Soviet combat forces. Men, machines, guns, rockets, missiles and radars could not cope with the UFO's. The impact in the Soviet military of several such military actions must have been enormous.

In fact, such incidents as this probably were responsible for the establishment of an extremely high level and highly classified Soviet investigation of UFO phenomena. Given the decision to launch such an investigation — and it would be inevitable for the Soviets to take such a step when their air space was under open invasion by possibly hostile forces — the Soviets would have assembled a formidable scientific team to tackle the problem. Furthermore, given the highly classified military reports of such encounters, reports of UFO incidents at other Soviet installations and by other Soviet military personnel, and the steady occurrence of the normal range of UFO sightings and incidents over cities and villages, the Soviet system would have tolerated no scientific nonsense in opposition to the mission to discover the nature of the phenomena. In other words, the Soviet team from the beginning would have concentrated on finding the explanation and nature of the phenomena, not on whether or not such phenomena were real or were occurring.

In my opinion, a connection between UFO's and paranormal phenomena — and the new psychoenergetics which also was formulated by 1959-60 — was realized about 1960. The high classification of the Soviet psychoenergetics program then led directly to an official Soviet policy to debunk UFO's and paranormal phenomena if the work of any outside research group became too serious. Some Soviet scientists who had no connection with the weapons programs were allowed to work a little on paranormal phenomena and on UFO's. This was simply to provide a deception plan for Western scientists; in their contacts with their Soviet scientific colleagues, these Western scientists would see only a rather puzzling collection of simple laboratory experiments on telepathy, ESP, simple psychotronics, psychokinesis, etc. A limited number of Soviet papers dealing with these subjects, but of no particular theoretical content or very exact statement of experimental apparatus and conditions would be permitted in the open journals. However, close scrutiny of such activities would be maintained by the Soviet state, and no real breakthroughs would be permitted to occur or to be transmitted to the West. Ziegel, Dubrov etc. owe the limited work they were permitted to do to this Soviet deception policy.

They also owe the surveillance and harassment they have undergone to the Soviet policy of insuring that they would be permitted to do nothing significant, or to communicate nothing significant to their Western colleagues. Meanwhile, a rigorously controlled and highly classified Soviet psychotronic weapons development program continued apace. And the radiation of the U.S. Embassy (and the U.S. ambassador to the Soviet Union) in Moscow with weak radiofrequency microwaves containing psychotronic modulations began about 1959-60 and con-

m

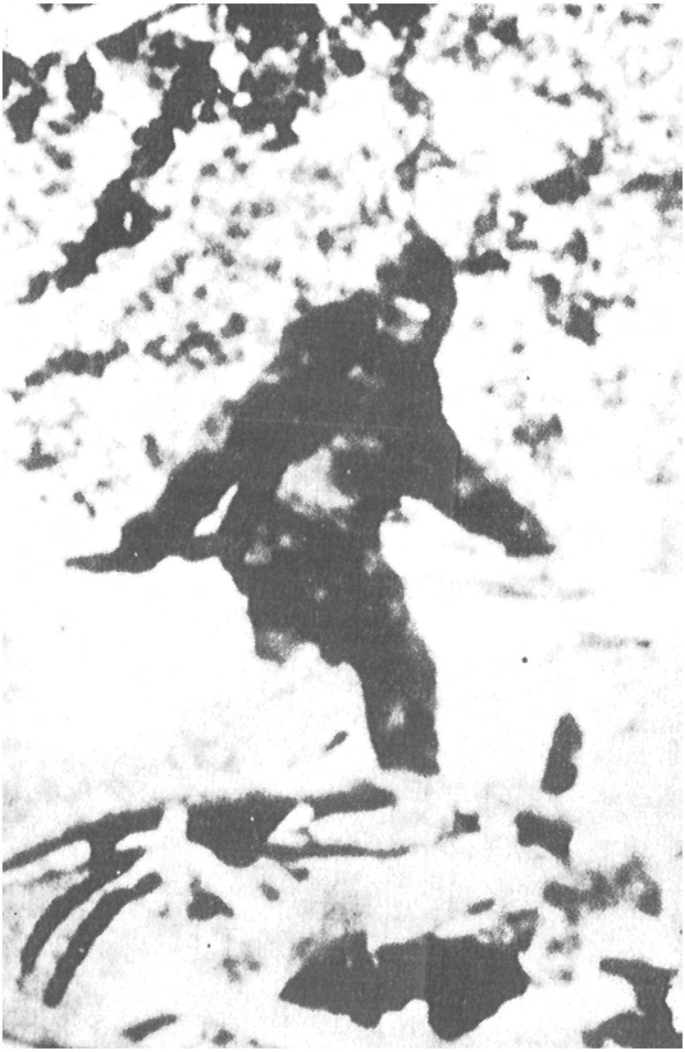

Photo 27. Sasquatch.

Photo 28. Bigfoot cast.

tinues to the present. As explained elsewhere, this radiation was in fact another very clever Soviet intelligence probe: By observing how the U.S. reacted, they could determine whether the U.S. had any knowledge of psychotronics and thus whether or not the U.S. had developed secret psychotronic weapons and counters of its own.

The Sighting over Petrozavodsk

Even more spectacular phenomena have occurred in Soviet skies and are continuing to occur.

For example, a mysterious phenomenon of great magnitude occurred in the skies over the Soviet city of Petrozavodsk at about 4:00 A.M. on September 20, 1977. Petrozavodsk is in Soviet Karelia about 130 miles from the Soviet-Finnish border.

First, a gigantic mass of light flared up in the skies over the city, glowing brilliantly, and hovered there. Fantastic streamers of light poured from the glowing hemispherical dome of light and radiated down towards the city. Tass, the official Soviet news agency, reported that it looked like nothing so much as a shining "jellyfish sending down a multitude of fine beams." It gave, Tass said, "the impression of a torrential rain of light." Where these fine beams touched windows, many windows were broken or perforated with small holes.

Later the glow died down and the jellyfish became a single giant bright semicircle. It then gradually moved off in the direction of Lake Onega. But all was not yet over. Even after the bright hemisphere had moved away, great gray clouds appeared over the city with a semicircular aperture bright red in the center. These phenomena lasted for ten minutes or longer.

Being over the city and lighting up the sky so fantastically, the phenomena were seen by a large number of Soviet citizens and reported by them to the Soviet authorities. In fact, Tass admitted in its initial release that informative reports from witnesses were still pouring in and being analyzed. This interesting phenomenon is worthy of some remarks by way of explanation.

Not only is pressure from the Soviet threat transmitted through the collective human unconscious (ZARG) to the skies of Europe and North America; in addition, a backpressure is generated, which provides the impetus for initiation of appropriate phenomena over the Soviet Union and the Iron Curtain countries. Psychologically, it may not be accidental at all that the original shape resembled a giant jellyfish; the implication is that the psychotronic sting will also be transmitted to the Soviet Union by backlash whenever the Soviet psychotronic weapons are unleashed on the West. The shapes are also psychologically significant. First the Portuguese man-o-war's stings will be applied (this accounts for the tentacles toward the city); then the all-seeing eye (symbolized by the bright hemisphere) will check the results. And the results will be gloom and despair (gray clouds), surrounding death, blood, and destruction (the red center) left in the aftermath.

Thus in this materialized symbolic dream event of ZARG's, there is a clear warning and an omen to the Soviet Union about the end result that awaits it on its present course to attempt to conquer the world and unleash a great psychotronic barrage in the process.

Since so many Soviet citizens saw and reported this fantastic phenomenon in alarm, the Soviet regime could not ignore it or simply debunk it. The Tass release thus tried the oblique dodge: the phenomenon was still being evaluated, reports were still being received and checked out, and an official of the Petrozavodsk observatory, Yuri Gromov, was quoted as saying the phenomenon was still a mystery but was confirmed by a large number of eyewitnesses. The Pulkovo observatory in Leningrad was quoted by Tass as speculating that the phenomenon was due to a meteor. And of course there one can expect the matter to be left in limbo.

An interesting development related to this UFO sighting occurred in the U.S. The October 3, 1977 issue of *Aviation Week and Space Technology* magazine carried a brief disposal of this phenomenon. The brief insert, Spook Show, referred to a "Soviet press ballyhoo" of the incident and stated that the incident was actually the launching of the Soviet Cosmos 955 classified military satellite from secret launching facilities at Plesetsk. This AWST insert is credited to Washington Staff.

My comment is that the launching of the Cosmos 955 may well have been the initiator of the event — a rocket firing does produce a tremendous ionized trail of burning exhaust gases rising up into and out of the atmosphere, and this is a bioenergy collector par excellence. However, a simple review of the time sequence of the phenomenon, its location, and its shape is sufficient to destroy any notion that one could mistake it for a simple rocket launch. And rocket launches had occurred there many times before, with no such phenomena or mass reportings by the population, and with no such news release by Tass.

FRENCH EXPERIMENTS WITH PSYCHOTRONICS

Aime Limoge in France is building some truly unusual gadgets that are being tested and that work. Limoge travels widely in the Soviet Union, and attempts to duplicate some of the unusual Soviet devices he has been permitted to see. Some of Limoge's devices have a partially psychotronic nature, although none of his reports mention this.

One of his gadgets, a portable and battery-powered electroanesthesia device, was considered for purchase by the U.S. Army.

Another of his gadgets is a magnetic inductor of hypnosis, which has been used in medical treatment by L'Institut de Recherche d'Electro-anesthesie. Limoge is reported to have seen such a device in the Soviet Union, possibly because one part of the Soviet psychotronics deception plan calls for the deliberate release of information of this kind, in a manner to suggest that this is the latest state of development in the USSR. In fact, their research efforts probably go far beyond this point. However, we should note that this device can be used as a most efficient brainwashing instrument.

DOWSING

To understand how dowsing works, we must first understand something about higher dimensional spaces. It should be evident that, if there is a type of energy being sensed by the dowser, that energy must be coming in or arriving from outside this three-dimensional world, since our ordinary science deals with the energies existing in this world, and none of these ordinary energies can explain dowsing. However, because of the work that has been done showing that dowsers are extremely sensitive to magnetic fields, this energy must be somehow related to the ordinary magnetic or electromagnetic fields known to science.

In the human being the mindworld must itself have a coupled kindling system to cohere its changes and kindle them into electromagnetic energy and energize the physical body. The body system which does a great deal of this is the acupuncture system of points, plexuses, and meridians. Also, all powerful emoters, such as the endocrine gland system, are deeply involved in the kindling of the biological organism's mind energies into electromagnetic energies to energize the physical body. The body and its biorhythms form a coherent tuner in the time channel to accomplish this kindling of time-nested vital thought energies into ordinary energy. Because a great deal of mind energy kindles into only a minute amount of normal electromagnetic energy, our minds normally have only a small effect in this particular three-dimensional world. So we can normally psychokinetically move a few electrons in our brain and nervous system, and generate an input signal to our otherwise inert physical servomechanism. Ordinary inert physics and amplification takes over from there, and we can only deduce the kindling as mental intent from observing many, many quanta of physical change, i.e., from macroscopic behavior. At any rate, the dowser has learned to detect, focus, and sense very fine levels of electromagnetic fields, which means that at these low levels the kindling of mental energies can represent a much greater percentage of the field than normal. Thus the practiced dowser can now discriminate many of the particular types of fluctuations or changes in these weak fields due to mind kindling.

Furthermore, I would like to point out that, since mind kindling comes into this universe orthogonally, normal distances do not apply. The entire physical frame appears to the mind world as a point, and thus is totally inside the mind of the operator/dowser. It is for that reason that a dowser can establish a coherent tuning (intent) between a map and the actual ground, or between a picture and an actual object, or between a rod and a person' physical body when sensing medical illnesses. In the mindworld frame, the dowser is only dowsing or detecting himself, and himself includes this entire physical world — the earth's surface and beneath it, the sun, the moon, the stars, or whatever.

Just as the body used electricity, magnetism, and electromagnetic signals long before man knew of their existence, it has used kindling of bioenergy from higher dimensional frames for a very long time, even though these have been unknown to mankind. And dowsing is a very old art indeed; the ancients used magnets without knowing of magnetic field, and they used dowsing without knowing of hyperspaces and hyperfields.

THE STROBE EFFECT

These five pictures show one of the more interesting effects demonstrated by many tulpoidal forms: strobing or "coining." What happens is that, much like a flat rock skipping across the surface of a pond, the orthorotating tulpoid skips in and out of its intersection with our three-dimensional frame.

In the first picture a glowing spherical tulpoid is shown to skip or strobe several limes while the camera shutter is open. It does not move in a linear pattern, but rocks around and back-and-forth.

In the second photo, an aerial photo, a UFO in photonic form (first bioframe) reveals itself to be strobing or skipping. (In university classes on the paranormal, I often refer to this type of action as *coining,* since it resembles a bright quarter or nickel that is photographed by stroboscopic light while rolling around.)

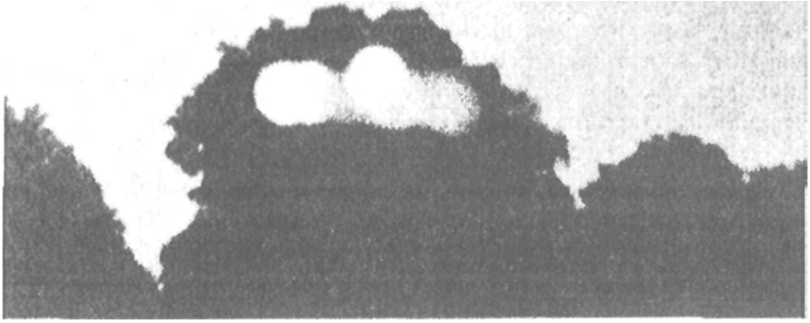

Photo 1

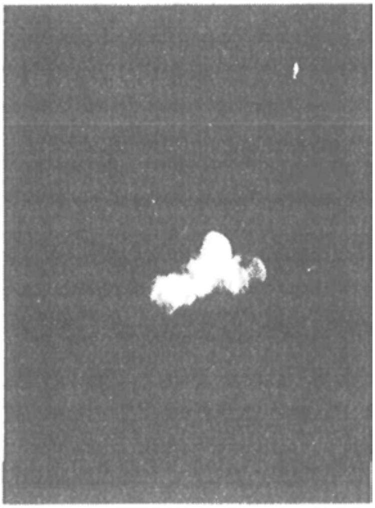

Photo 2

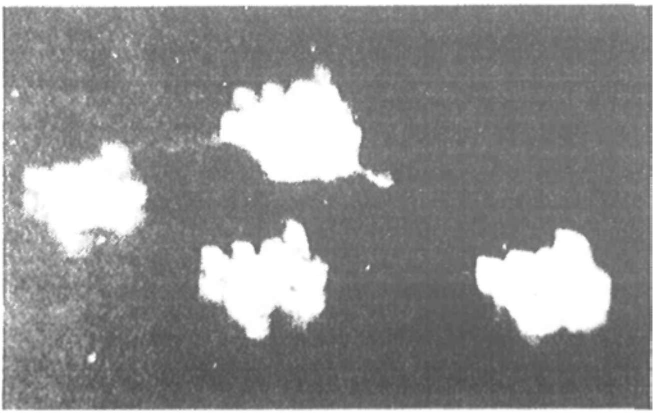

Photo 3

Photo 4

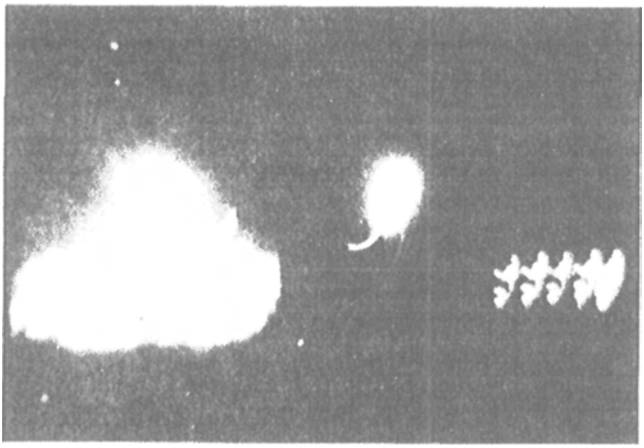

Photo 5

Photo 29. The strobe effect.

1. Even dischargo from Inert object (dime).

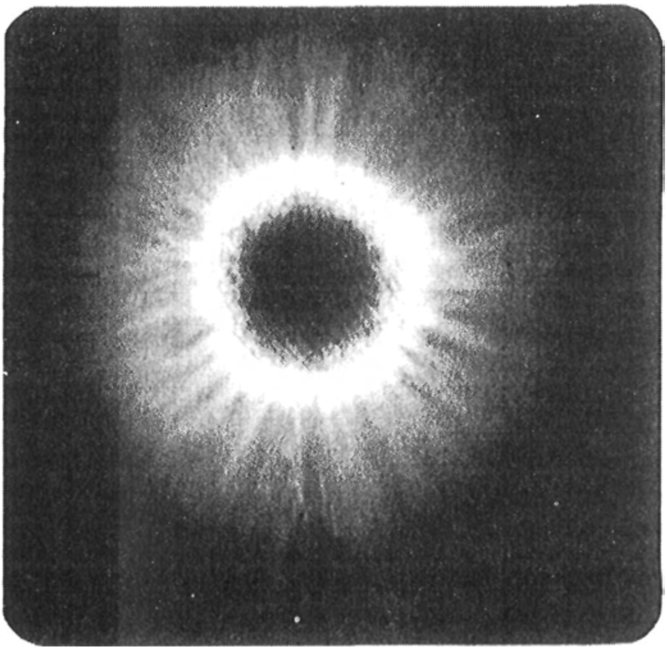

2. Healthy person in calm state

Photo 30. Kirlian photography.

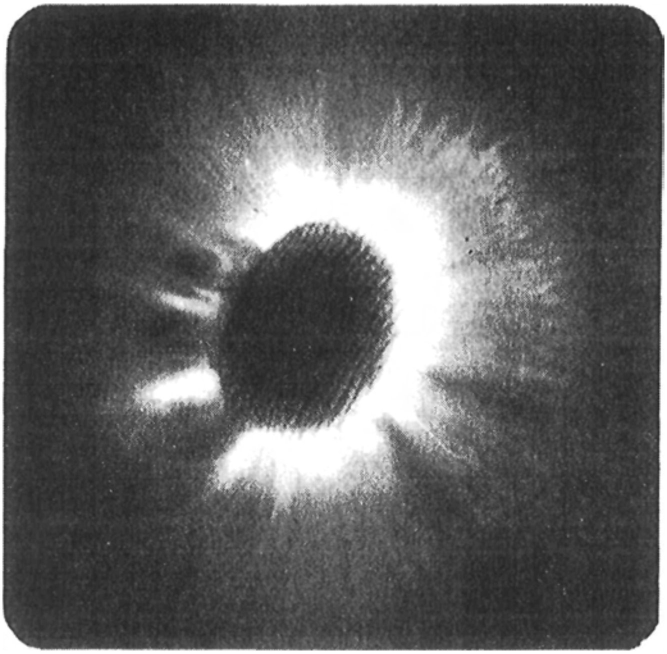

3. Person in meditation

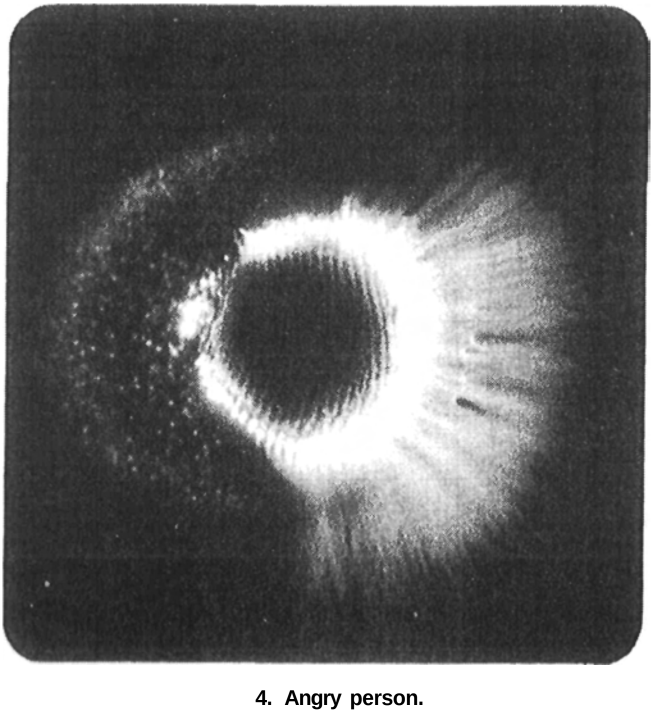

4. Angry person.

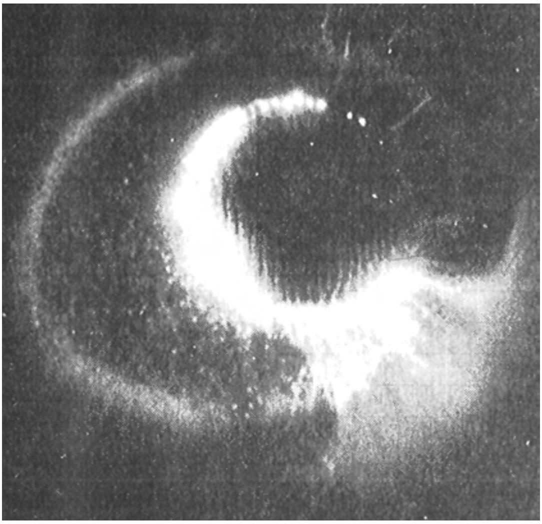

5. Person with London **flu.**

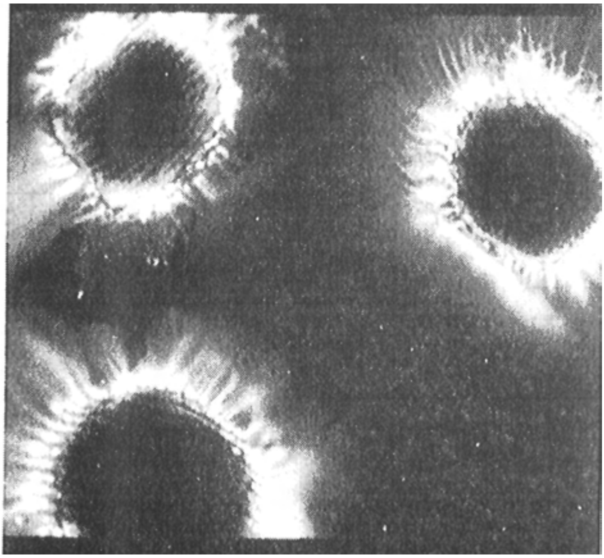

6. Interactions between three persons.

The third photo shows a more striated form. In fact, these forms show kindling and strobing at the same time. In addition, the physical size of the strobed intersection with our three-dimensional space changes during the photo exposure.

The fourth photo clearly shows both the striations due to strobing and also the variation in size of intersection in the zeroth frame by the hyperdimensional tulpoid.

Photo five again shows these effects, with the addition of the two "bursts" in the upper part of the left tulpoidal object and in the center tulpoid.

Significantly, psychokinesis also often demonstrates stroboscopic movement.

KIRLIAN PHOTOGRAPHY

These six pictures show some major characteristics of Kirlian photography: (1) an even discharge from an inert object (a dime); (2) a balanced, gentle, even discharge from a healthy person in a calm state; (3) a gentle outgoing discharge enhanced on one side, and a gentle indrawing on the other, from a person in a meditative state; (4) strong bunching and outgoing discharge from an angry person (the word "emote," after all, means "move outward"; (5) choppy (as if out of control) bunched and outgoing discharge from a person ill with London Flu, with strong suppression of the normal, relaxed blue component at the lower right; (6) interactions between three persons, showing differing degrees of outgoing and barrier reactions.

The film itself only records ordinary photochemical interactions; it specifically does not record the aura. However, changes in the discharge patterns can be and often are induced by kindling bioenergy into the first biofield (electromagnetic field), inducing changes in the sparking pattern and intensity. Furthermore, this can occur fairly readily on such a high voltage series of discharges, which are excellent collectors for kindling.

As a final note, it should be pointed out that David Faust *et al* have published a technical paper in the scientific literature detailing the fact that Western experimenters have not yet duplicated the Soviet equipment and methodology used in their Kirlian photography. While David formerly pointed out in *Science* that experimenters must carefully control the variables—particularly moisture (water is an excellent collector for kindling)—he has been widely misinterpreted as having proved that water vapor was responsible for all Kirlian effects. This he has not done at all, nor did he intend to, since he himself is one of the most competent investigators in electrophotography.

INTERACTION OF BIOFIELD, THOUGHT, AND SPARK DISCHARGE

These photographs are strong examples of the kindling mechanism. The subject tested was Uri Geller, a person with strong inceptive ability. In photo 1, the top two exposures are for control purposes, while in the bottom exposure, Uri is concentrating on the image of a circle. In photo 2, again there are two control exposures shown at the top, while in the bottom exposure Uri is concentrating on the image of a triangle. In photo 3, exposures are taken of Uri and a wristwatch; this is

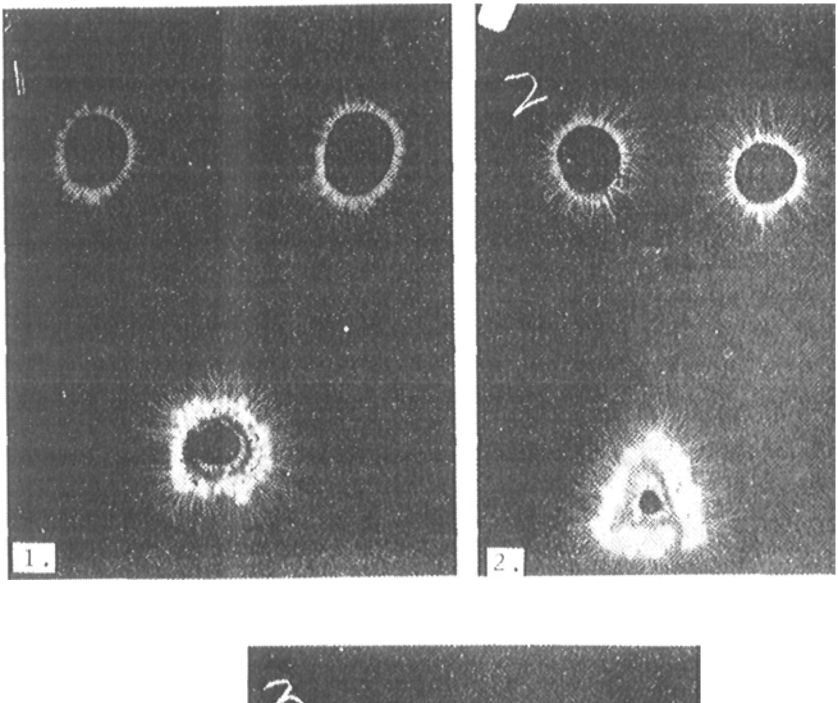

Photo 31. Interaction of biofield, thought, and spark discharge.

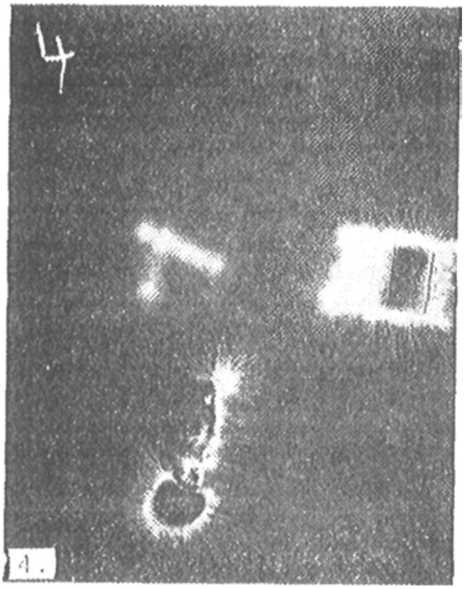

4.

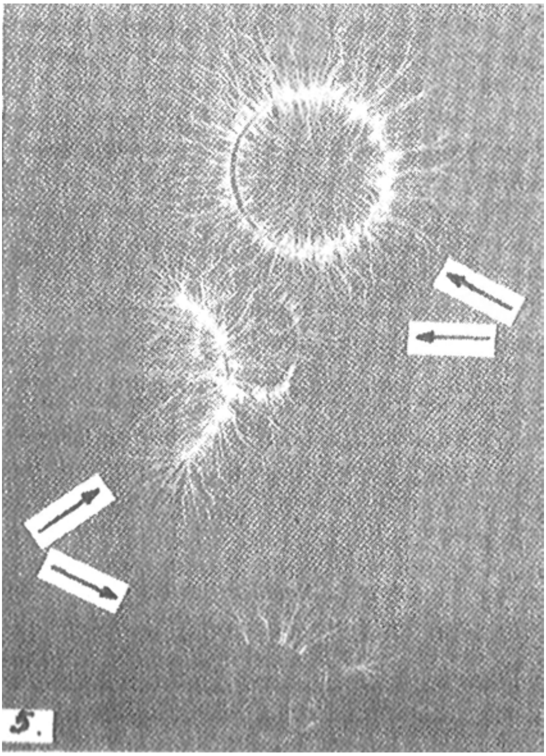

5.

a control photo. In photo 4, Uri is attempting to send energy toward the wristwatch, and a correlated change in the discharge patterns from his finger is evident. In photo 5, the arrows at the right point to control exposure of Uri and a silver ring; the arrows at the left point to the exposure made with Uri attempting to bend the ring psychokinetically without contact. The direct kindling of bioenergy in a force-field line intersecting the ring is striking. In such an intense kindling, note that the nuclear currents inside the nuclei of the silver atoms — where positive charge is being shuttled back and forth between protons and neutrons — are also being strongly kindled. Reductions in the positive charges steadily reduce the charged exchanging particles to quarklike particles, severely affecting the structural integrity and crystalline lattice directions of the base metal. By this means metal bending — or even breaking — is accomplished.

KERVRAN EFFECT

A direct and repeatable experiment showing that living systems can kindle bioenergy into ordinary energy and material effects has been made possible by the fascinating work of Louis Kervran. Kervran showed that chickens, fed a diet deficient in calcium but containing excess potassium, will have no calcium deficiency, since they transmute some of the potassium into calcium. Kervran's pioneering work has been duplicated in Japan and elsewhere, and the same effects have been demonstrated by living plants.

We may consider a neutron as a sort of bound system of a potential proton and a potential electron. Psychokinetically killing the charge on the potential electron converts the electron into a neutrino, leaving behind a proton now freed. For the one percent or so of potassium that is a calcium isobar, this process can produce ordinary calcium. Thus in a potassium-rich diet the chicken can psychokinetically transmute enough potassium into calcium to remain healthy.

Kervran was nominated for a Nobel Prize in 1977 by the Japanese researchers; his work is direct proof of the ability of living systems to violate present physics, and of the ability of mind energy or bioenergy to affect matter psychokinetically.

PSYCHIC SURGERY

These two photos taken by Sigrun Seutemann in the Phillipines show materialization of large lumps of tissue brought about by Tony Agpaoa on a patient from an educated Western culture. The minds of both the patient and the surgeon are involved in the process, according to the psychic surgeons themselves. "Western patients are strongly conditioned to believe that medicine involves bloody tissue, so this is what results in the particular form of the materializations. For native patients who are steeped in the lore of shamanistic medicine, apports may consist of substances such as leaves, coconut fibers, feathers, etc.

The legitimate surgeons, working in a religious group structure, are often able to generate sufficient kindling to fully materialize such unconsciously structured thought forms, and also to dematerialize and rematerialize the tumor from inside the patient's body, the tumor usually appearing or emerging from an opening which appears in the apport tissue. In the surgeon's view he is also being helped by powerful spirits — and continual charging of such thought forms with kindled

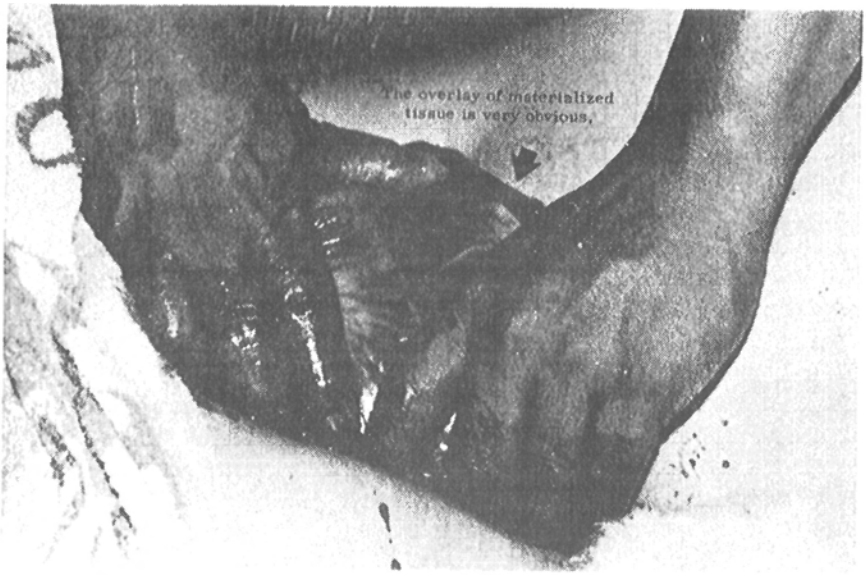
The overlay of materialized tissue is very obvious.

1. Materialization of apport tissue while removing a tumor from the abdomen.

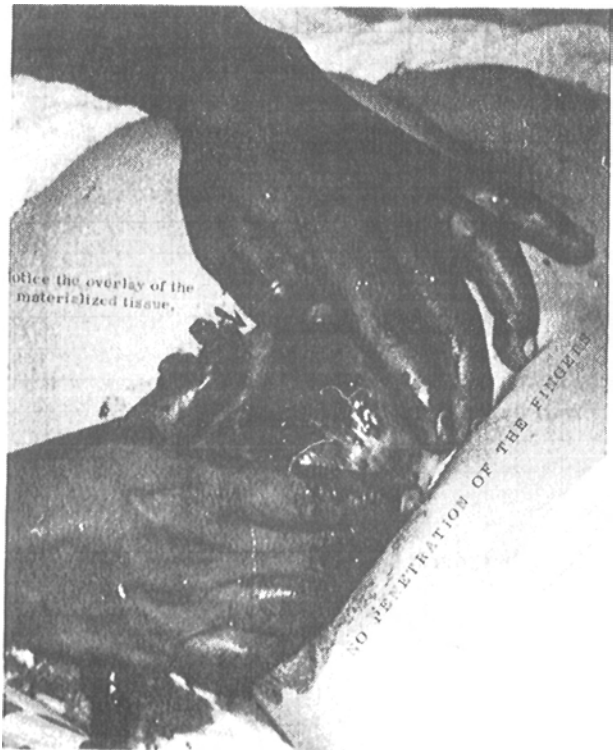
otice the overlay of the materialized tissue.
NO PENETRATION OF THE FINGERS

2. Matoriallzatlon of apport tissue while working in the hip region.

Photo 32. Psychic surgery.

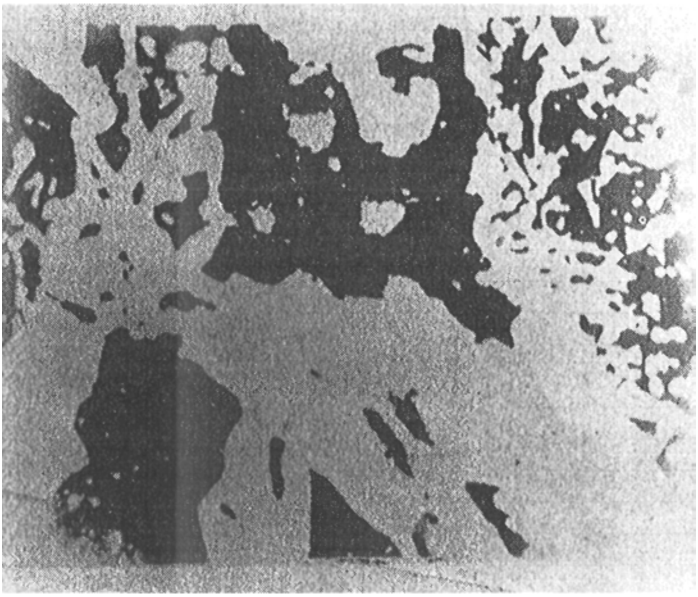

This highly unusual photograph resulted in the following manner: Mrs. Baune Is a devoutly religious lady, and on this particular day she had spent a great deal of time in prayer and a Christian attitude. Snow was gently falling on the shrubbery In her yard, and she and a neighbor were admiring the beauty of the fresh snowfall. The neighbor borrowed Mrs. Bauno's camera and snapped a photo of the snowfall. When the photo was developed, the falling snow had formed a pattern of Jesus on the shrubbery In Mrs. Baune's yard.

Photo 33. Kindling of thought bioenergy in snow.

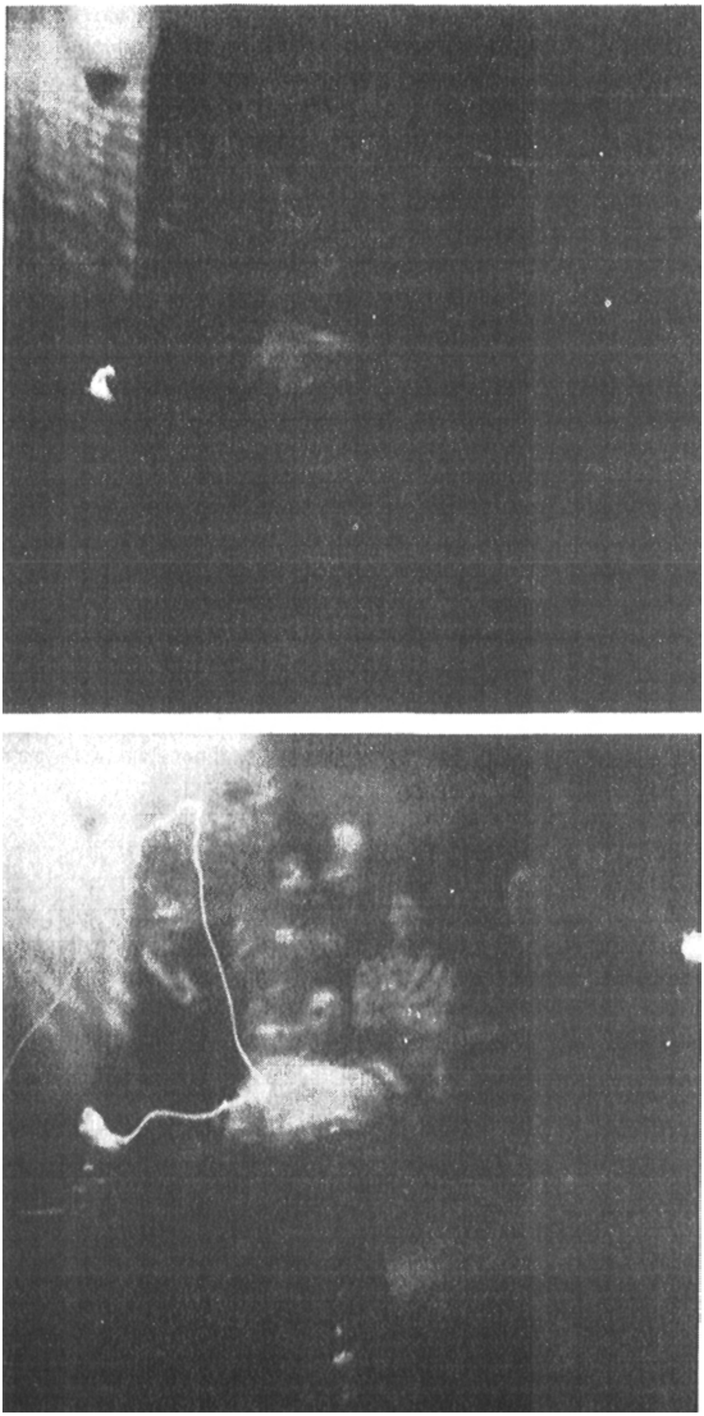

Photo 34. A deliberate experiment with kindling.

energy indeed can yield bioenergetic formats of more intense energy than a single talented individual can usually muster in an unaided or unaugmented state. Almost every religion, in fact, still occasionally exhibits miracle healing accomplished in somewhat similar fashion, though possibly less dramatically.

The apport tissue is a materialized tulpoid, precisely as are sasquatch, the yeti, the Loch Ness monster, and UFO's. Apports are normally more stable than these latter materializations — but they have also been known to disappear, as have the "angel hair" apports associated with UFO's.

A DELIBERATE EXPERIMENT WITH KINDLING

I and other members of the American Association of Meta-Science performed a series of experiments in kindling thought or psychic energy into electromagnetic energy, and photographed the resulting light energy on film with an SX-70 camera. The two photos shown here are from one of those experiments. In the left picture, bioenergy is just at the "burst" point where it is being collected on a candle flame. Here the camera lens is not quite fully opened. Only a single 40-watt lamp illuminates the room, and several AAMS members have been engaged in about two hours of psychic exercises. In the second picture, the lens of the camera is fully opened, and a "burst" of kindled energy occurs as the lens shutter is operated. None of this is visible to the naked eye. The polaroid film, however, is sensitive to some infrared and ultraviolet frequencies, so that bioenergy kindling into electromagnetic energy at IR or UV frequency can be photographed even though not seen by the naked eye.

FAY CLARK'S "FIREFLIES"

Fay Clark experimented with an ultrasonic generator in a darkened fallout shelter and photographed the most interesting results. At one particular frequency, dynamic dots of light appeared, grouped themselves into formations, and demonstrated movement and growth. At times they seemed to respond to thought, as if they were somehow intelligent, sentient creatures. This is a beautiful case of kindling using ultrasonic frequencies as the collector mechanism. While Fay is no longer sure of just exactly which frequency produced the fireflies, he does know that he used frequencies in the 38—40 kilohertz region, among others. This may be significant, since the Vestigia experiment encountered electromagnetic frequencies in the same region in their historic study of a mystery light. The frequency region 38—40 kilohertz is a strong candidate for one of the magic window frequencies between bioframes where kindling is much easier than at other frequencies. Furthermore, since it is thought energy or mind that is actually being kindled, one would expect the possibility of sentient behavior. Clark's experiment occurred in the dark, so the squelching effect of visible light was not present. Finally, Fay Clark is a known psychic himself, and so this should have eased the kindling difficulty even further. The value of rigorous laboratory experiments based on the lines of Fay Clark's pioneering "firefly" experiment is strongly indicated.

"FOO" FIGHTERS

When large groups of people are under strong external pressure, one result is often tulpoidal "popouts" from the collective human unconscious.

A war is a typical example of a situation where the group consciousness of a country is under enormous pressure. In such a situation, the "targeted" population is often exposed to tulpoidal manifestations from time to time.

In World War II, for example, the Germans thought at first that the glowing lights darting around their planes were secret Allied weapons; and the Allies thought the same thing of the Germans when they saw such unusual phenomena. The French dubbed them "feu" (fire) fighters, which became "foo" to American soldiers in Europe.

"Foo" fighters were a common phenomenon in World War II, particularly in cloudy, overcast skies.

WATER MONSTERS

In the oceans and in every large body of water on earth, water monsters have been reported for centuries. Loch Ness in Scotland is of course one of the most publicized examples, but creatures of different shapes and sizes have been seen in large lakes and streams everywhere. This is not at all surprising when one realizes that a large body of water is a most excellent kindling reservoir, and bioenergetic forms are more readily kindled in such locations.

The large dead creature in photograph 36 was hauled off the bottom of the sea by Japanese fishermen in their nets, photographed, and then thrown back to prevent contamination of the catch of fish and to get rid of the stench from the decaying carcass. It is a pleisosaurlike creature's remains, not similar to anything else known on earth. Predictably, scientists were almost immediately at odds as to what it was, and the carcass has been called a shark, a whale, a large elephant seal, a genuine pleisoaur, etc., by various "experts." Nonetheless it was real and physical, and a good photograph was obtained of it.

Photograph 37 is one of the underwater photographs of the Loch Ness monster taken by a team from the Academy of Applied Science, Boston, Massachusetts on June 20, 1975. As can be seen by a detailed study of the photograph, this tulpoid is not clearly tuned or clearly formed. Many other photos of the Loch Ness monster have been taken, and various shapes and sizes and degrees of tuning have been recorded. In the Doc Shiels photos taken in 1977, Nessie is well formed but entirely different from the Academy photo. Also, Doc Shiels' photo was subjected to computer analysis by Ground Saucer Watch, and the image exhibited a transparency that was disturbing to the analysts. On the other hand, such a transparency is to be expected from a partially materialized tulpoid which is not completely orthorotated into the zeroth bioframe.

In the Academy photograph, one can see — particularly on the color original — a form which has several conglomerated faces, which exhibits eyelike structures in several locations to suggest first a cowlike face and then a horselike face, and incoherently tuned earlike and hornlike appendages in several places. In fact, the Academy photo bears little resemblance whatsoever to a pleisosaur or any other more normal and coherently structured mammal or fish.

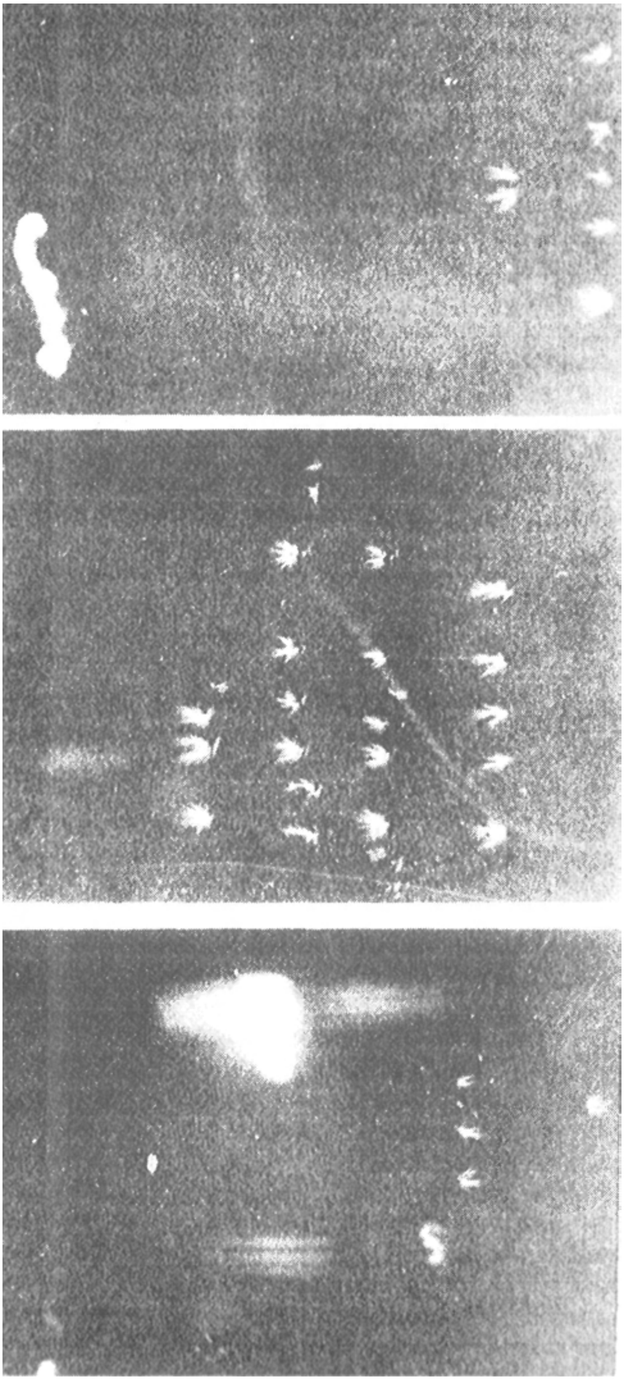

Photo 35. Fay Clark's "fireflies."

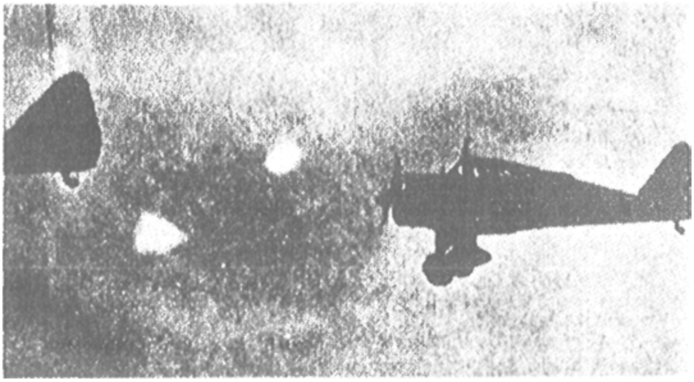

"Foo fighter" tulpoids pacing Allied aircraft in World War II.

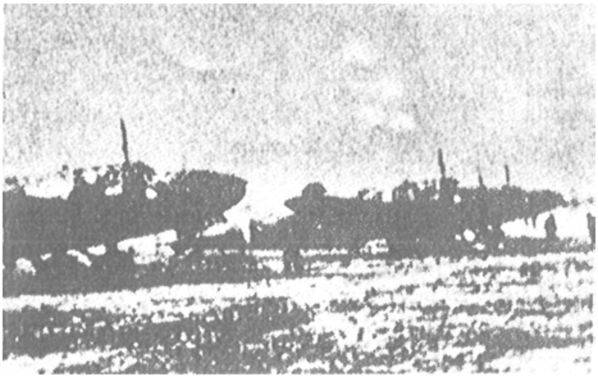

"Foo fighter" tulpoids over an airfield in China in World War II.

Photo 36. "Foo" fighters.

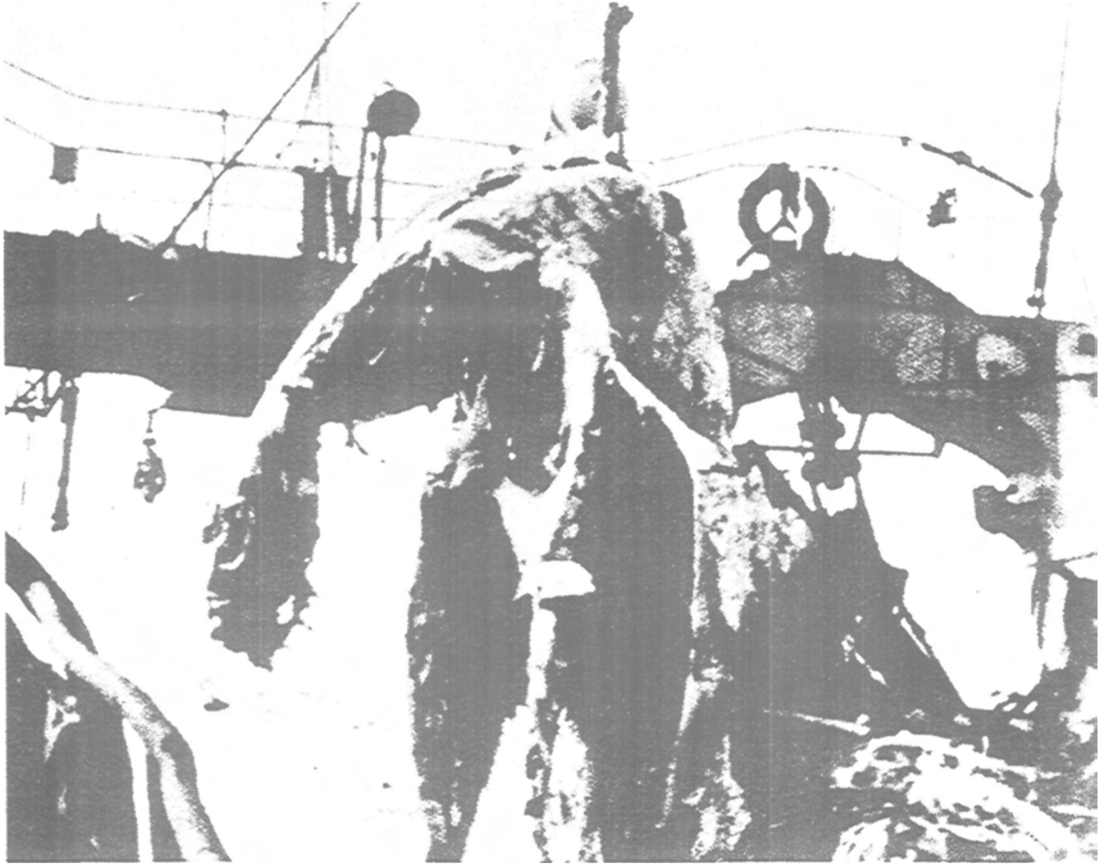

Photo 37. Pleisosaur-like creature hauled from the sea by Japanese fisherman.

The Loch Ness monster is definitely a tulpoidal form, and the format varies appreciably from reception to reception due to distortions and variations in tuning. Nessiteras rhombopteryx is a much stranger beast than anything the dedicated Nessie seekers, researchers, and scientists have yet suspected. In Part Two I will examine the scientific basis of tulpoidal forms.

CATTLE MUTILATIONS

Over 9,000 cattle mutilations have occurred in the U.S. alone since 1973.

The cow is a female symbol par excellence — our children drink cow's milk, and our babies feed on the milk from the cow's breast. The U.S., being a natural fortress, is a "female" symbol, for in primitive human evolution the females stayed in the cave or protected fortress while the males went forth to hunt or do battle. (This symbolism will be fully discussed in Part Three.) Violation of the female means that the heartland U.S. shall be violated.

Teats are cut off — the children of the West shall lose their sustenance; lips, ears, and tongues are removed — freedom of speech and hearing shall be excised; sexual organs are removed — no more fruit of the womb shall be evidenced in the West; sharp, surgical incisions are made — the operation shall be accomplished swiftly (blitzkrieg war) with surgical precision; zig-zag cuts — deception shall be used as to direction and intent; blood is totally drained — the lifeblood of the West shall be totally drained; edges of the cuts are often seared as if by a laser — advanced, futuristic weaponry (psychotronic weapons) shall "seal the end."

The cattle mutilations are materialized precognitive nightmares indicating the horror of the Armageddon to come, and these nightmarish paranormal mutilations reveal the true nature of the times. Yet no orthodox scientist or governmental agency is equipped or able to deal with a phenomenon such as this.

In ancient times the wise men knew how to do a metapsychological analysis of the species unconscious, and knew that tulpoidal phenomena were often symbolic signs of that which was to come.

Where are our wise men today?

BIG MAMA

UFO phenomena have also been associated with the cattle mutilations. One of these UFO-type tulpoids has been labelled Big Mama, because of its characteristic behavior of spawning smaller UFO light-shapes that split away and speed off to an area in which a cattle mutilation then occurs. The "spawned" UFO's then return and link up or join with Big Mama, and Big Mama then withdraws.

The photo shown here is a typical appearance of Big Mama. Big Mama takes many sizes and shapes, but is a fairly reliable indicator of paranormal cattle mutilations impending.

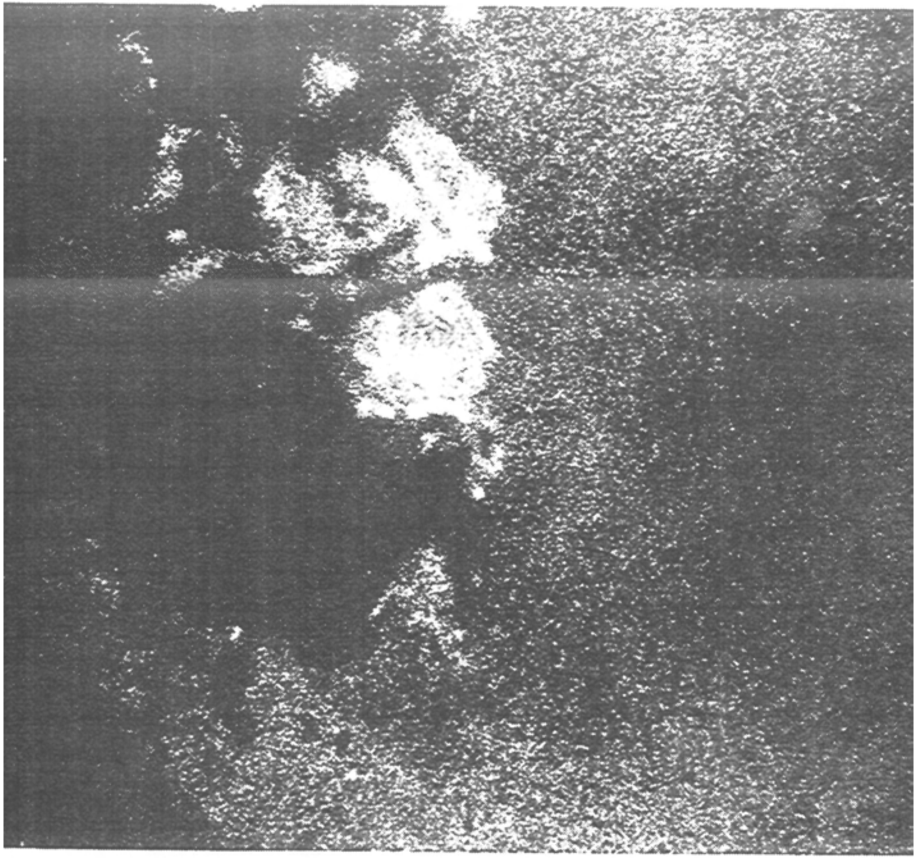

Photo 38. Loch Ness monster.

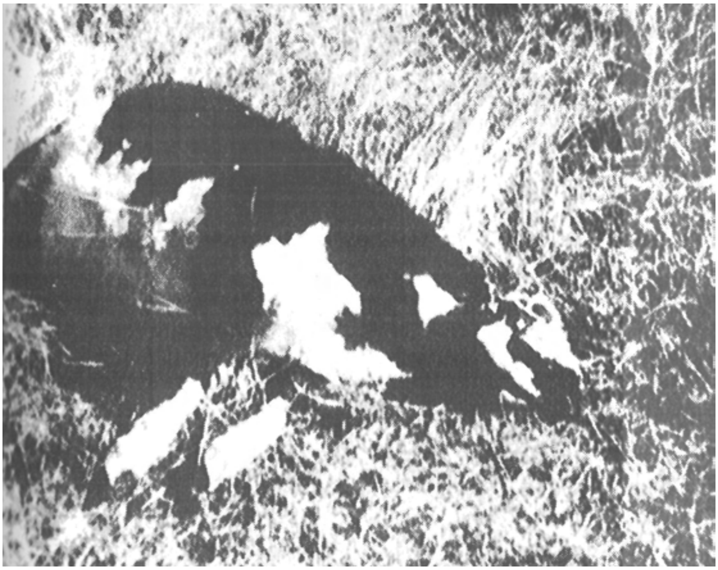

Photo 39. Cattle mutilation.

During the summer of 1978, two psychics and I — all members of the American Association of Meta-Science spotted a Big Mama-type UFO from the top floor of a hotel in Louisville, Kentucky. Realizing what it was, we promptly brought into play a certain type of powerful psychic exercise and tried to make it dematerialize and vanish. Several times it briefly split into pieces (five to be precise), but we were unable to extinguish it. Over a period of approximately twenty minutes this effort continued, and all the while Big Mama slowly moved lower and further away, eventually disappearing behind a tall building after having become very distant.

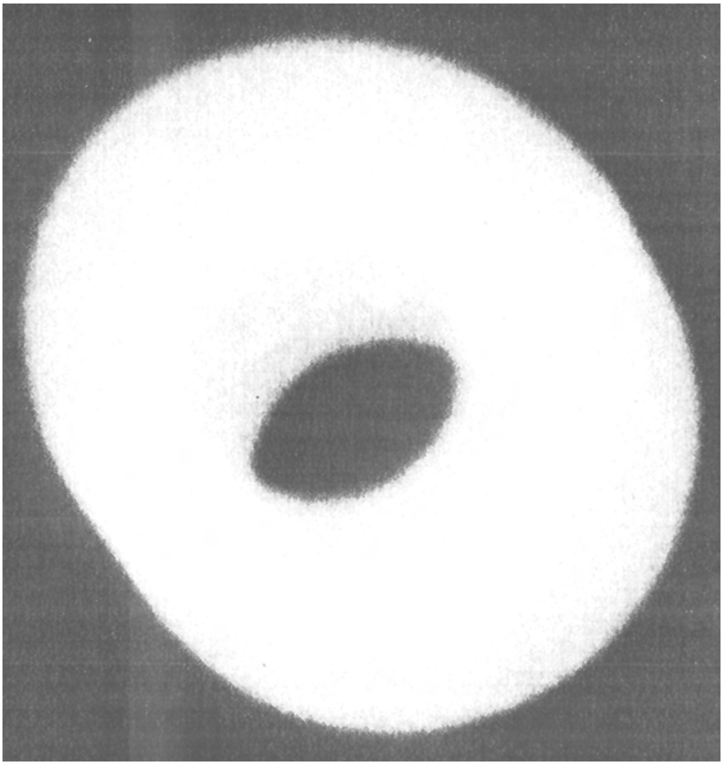

Photo 40. Big Mama.

PART TWO

A THEORETICAL BACKGROUND FOR UNDERSTANDING PT, UFO'S, AND PSI PHENOMENA

Footnote for page 132:

'With the exception that, in my theory, the orthogonal universes crosstalk with each other at virtual (subquantum) levels. Indeed, 'change' in one universe consists of the quantum coherence in that universe of crosstalk in and out of it. The 'virtual state' in one universe is due to, and consists of, subquantally coherent (quantally incoherent) crosstalk in and out of that universe.

Psychotronics, unidentified flying objects (UFO's), and paranormal phenomena such as psychokinesis, telepathy, and precognition all have one thing in common: both mind and matter-energy are involved.

Hitherto, mainstream science has insisted on a rather arbitrary separation of mind and matter-energy. However, as we shall see, this separation is not based on good observation. For example, everyone exhibits in himself the ability of mind to affect matter. While theories of computers, control systems, and physics can explain human behavior once the physical system has been given an input signal, the original signal — intent, or what I call inception — is an unexplained mystery to ordinary science.

"Psychotronics" (a Czechoslovakian term) and "psychoenergetics" (a Soviet term) have been specifically designed to provide a framework for approaching the problems of mind, matter, and their interaction. If we can understand psychotronics, we will be able to understand better the interaction of mind and matter, including all paranormal phenomena, unidentified flying objects, and Fortean phenomena.

Let us therefore start by defining psychotronics. The prefix "psycho" refers to the mind. The suffix "tronics" refers to physics and physical devices. Thus the concept of psychotronics refers to a union of physics and metaphysics. Such an audacious concept can offer a solution to almost every present problem in metaphysics, the foundations of logic, the foundations of physics, and the foundations of mathematics.

To begin with, it seems necessary to form a unified theory of mind, matter, and their interaction. It also seems necessary to reinterpret and extend electromagnetic theory. Logic itself must be advanced, for a part of reality, although "illogical" according to our present logic, is nonetheless true. In metaphysics we must solve formidable problems: the ontological problem (nature of being); the problem of mind, and the interaction of mind and body; the problem of change; and the problem of nothing.

It has been said that fools rush in where angels fear to tread. To tackle the problems of psychotronics demands the audacity to rush in where even fools fear to **tread!**

At this time, it may be appropriate to point out that all Western science, mathematics, and logic are founded upon three simple laws of logic, proposed and formulated by Aristotle. Since the entire universe cannot be described by these three laws — parts of reality are known to violate one or more of them and hence be illogical though true — it appears that we must be audacious enough to tackle and change the three laws of logic, if a new paradigm is to be constructed to solve all the presently unsolved problems.

SOME UNEXPLAINED MYSTERIES OF PHYSICS

First, no one knows what mass really is; we do not know for sure whether inertial and gravitational mass are one and the same. General relativity says it is, **and** we know from experiments that if any difference exists between inertial mass **and** gravitational mass that difference is very small indeed. But we do not know positively that there is not some small but finite difference. And we have no idea why the mass of a moving object increases with respect to a static observer. It is simple to calculate in special relativity how much the mass increases as a function of the velocity, but no one has the foggiest notion why this happens. Particularly so since an observer standing on the moving object and moving along with it sees no increase in its mass at all.

This poses a real paradox; if we try to assign some absolute notion to the idea of mass, then any object has an infinite number of masses, all at the same time and all different. There are two ways to determine mass: (1) by its resistance to a disturbing force, and (2) by its ability to occupy three-dimensional space. The latter requires that mass be volumetric, i.e., that mass be L^3 dimensionally.

But look at the weird properties of a photon! If we measure its resistance to a disturbing force, we find that no force on earth can accelerate or decelerate the photon. If we measure the mass of a photon by the first method, it appears to have infinite mass. Yet if we measure it by the second method, it can have no mass at all, because a photon is not volumetric; it is two-dimensional. It therefore has infinite mass and zero mass simultaneously.

Furthermore, we can take the viewpoint that whatever its mass is, it can only have one. If that is so, then infinite mass and zero mass must somehow be the same thing! Which contradicts the three laws of logic. However, this should not concern us too greatly; many things contradict the laws of logic and are nonetheless true. So while it is presently "illogical" for infinite mass and zero mass to be identical, we should hold on to the idea that this may very well be true.

Additionally, one of the great cornerstone assumptions in physics is that gravitational field and electric field are mutually exclusive — i.e., they are totally different things. Yet a physicist named Santilli has proved that this is not so, and that they are either partially the same thing or totally the same thing. And there the matter stands, so we do not fully understand what electric field and gravitational field are, or how to go about turning one into the other, although Santilli's work seems to imply that this is at least theoretically possible. We must therefore invoke a new concept of reality.

URI GELLER

Uri Geller has demonstrated numerous capabilities: he has affected a magnometer inside a Faraday shield; bent and broken metal objects; caused a fresh flower to wither and dry in seconds; permanently changed the crystal structure of nitinol; altered magnetic programs on computer cards; influenced a Geiger counter; dematerialized matter; and demonstrated the inceptive Cyborg effect, thought-photography, telepathy, and the Geller effect (the sympathetic stimulation of a psi-positive).

It is useless to attempt to explain the feats performed by Uri Geller in terms of the ordinary fields and effects of known physics. Nothing short of a new physics paradigm will suffice. This new paradigm must encompass both mind and physics within the same theoretical framework, if it is to succeed in explaining how mind affects matter.

In terms of present physics and logic, the mind is regarded as totally separate from matter. For mind and matter to interact, some aspect of mind must be the same as some aspect of matter, i.e., to move matter, a force is required. Force is the time rate of change of momentum. To generate a force capable of moving matter, mind must be able to change momentum. But since momentum can only be obtained from something which possesses momentum to give up, then the mind must somehow possess momentum.

If this is so, then the momentum must ordinarily exist in a separate three-dimensional space, since it apparently does not ordinarily exist in laboratory three-dimensional space. Such a situation requires at least six spatial dimensions and one common time dimension, such that the mind (mental phenomena) is a set of objective physical phenomena existing in an ordinary objective three-dimensional space, but one which is three orthogonalities away, spatially, from the laboratory three-dimensional space.

Mental changes upon physical objects would thus imply orthogonal rotation of mental objects from the mental three-dimensional space into or closer to the laboratory three-dimensional space. Everett's many-worlds interpretation (MWI) of quantum mechanics contains a structure where such representation is possible, and the MWI is consistent with the entire experimental basis of modern physics. Thus from the MWI it is possible to objectively model the mind and matter, and from that, psychotronics as well.

MY APPROACH TO PSYCHOTRONICS

In twelve years of intensive work on this question, I have evolved a conceptual approach, which is consistent with present physics but extends it. A new concept of reality is involved, however, and severe demands are made on the individual to stretch his framework of comprehension. The perceptron concept is an abstraction that allows the modeling of perception itself— either mental perception or physical detection. The fourth law of logic (to be explained shortly) involves the age-old identity of opposites whose apparent necessity has baffled logicians, philosophers, and scientists for centuries. With the perceptron concept, we can at last comprehend how the identity of opposites is accomplished, and when it is accomplished.

This immediately solves the age-old philosophical problem of change. The fourth law also closes logic into a complete, closed metalogic, encompassing both physics and metaphysics.

Let us use as a definition of reality: *reality is that which can be conceptually and mathematically modeled and fitted to the phenomena existing in one or more minds.* Physical reality is that which can be modeled and fitted to the phenomena existing in all minds. Mental reality can be modeled and fitted to the phenomena in one mind. Mental and physical realities comprise "normal" reality. Complex or "paranormal" reality is that which can be conceptually and mathematically modeled and fitted to phenomena which exist in more than one mind, but not in all minds.

From perception theory I have succeeded in deriving a great deal of the present basis of physics, including Einstein's two postulates, Newton's laws, the law of gravitation, and the solution to the ontological problem. (Unknown to me at the time, Ives had already done so for Newton's laws, the law of gravitation, and Einstein's postulates. His work has unfortunately been ignored.) In addition, the photon can be seen to be an ordinary three-dimensional particle existing in a three-dimensional space that is orthogonal to the laboratory three-dimensional space frame. A stationary particle in the laboratory frame appears as a photon to the orthogonal spatial frame.

Einstein's second postulate is usually stated as "the speed of light is the same for every observer." Restated, this becomes "every photon in an inertial frame is moving at the speed of light, c, with respect to every particle in that inertial frame." The corollary then follows immediately: every particle in that inertial frame is also moving at the speed of light, c, with respect to every photon in that inertial frame. In a single three-dimensional space, this is incomprehensible. Taking two orthogonal three-dimensional spaces, with the photons in one frame and the electrons in the other, it is perfectly comprehensible.

From pure nothing — empty vacuum, absence of thing — one can generate everything, the presence of thing. Here is direct proof that ultimately opposites are identical. But to a mind rigidly programmed in three-law logic, such a truth is abhorrent. Physicists consequently have done nothing with the idea that all our physical laws — or at least most of the great ones — can be straightforwardly derived from a special kind of a "piece of nothing" which I call a quiton. And they have not seen that this constitutes a statement of a fourth law of logic—the identity of opposites on their common boundary.

It is interesting to note that the paper which was handed to an American reporter, Robert Toth, by a Soviet scientist, and which caused the KGB to seize Toth and charge him with receiving Soviet state secrets, contained just such a theory as the basis of a unifying theory of psychotronics. Yet none of our intelligence analysts seems to have picked up the overwhelming importance of what was in the paper, and Toth himself appears to believe that his KGB arrest was simply harassment. In fact the paper did contain material on the basic Soviet approach to psychotronics — and thus to psychotronic weapons — and it was indeed state secrets that had been given to Toth.

Everett's many-worlds interpretation of quantum mechanics, with which very few physicists are familiar, provides a needed correction to the conventional monocular interpretation of relativity, and it allows a theoretically sound basis to be constructed for psychotronics.

The conventional interpretation of relativity considers only a single observer at a time. But if you can accept so simple a concept as that both you and I exist simultaneously, regardless of how we move with respect to each other, then I assure you that physics is startlingly different from what you may have studied in the ordinary university physics text book.

Everett, originally a student of the world-renowned physicist Dr. John Wheeler at Princeton, for his Ph.D. thesis considered the problem of multiple simultaneous observers and worked out what this did to physics. His highly innovative thesis provided a totally new interpretation of quantum physics and defined a startling new kind of reality in which all possibilities are physically real and exist. This new physics is indeed very strange, but it is totally consistent with the entire experimental basis of physics today.

I discovered that all my perception theory could be fitted precisely onto Everett's many-worlds interpretation.[1] On that basis, a theory or schema of biofields was derived that provides an approach to a unified field theory. In fact, it predicts that any kind of field can be turned into any other kind of field, merely by correct and precise time synchronization. It also offers a physical and exact model of mind and mental phenomena and a mental and exact model of physical phenomena!

On this basis, a framework can be provided for psychotronics which is consistent with what we know of ordinary physics but which does not contain many of the limitations of ordinary physics. In a strange sense Everett wrote the physics for the all-mind or the Supreme Creator's mind because an unlimited mind, so to speak, must already have thought of everything possible, and for an all-creative mind to "think" a possibility is for that possibility to be real.

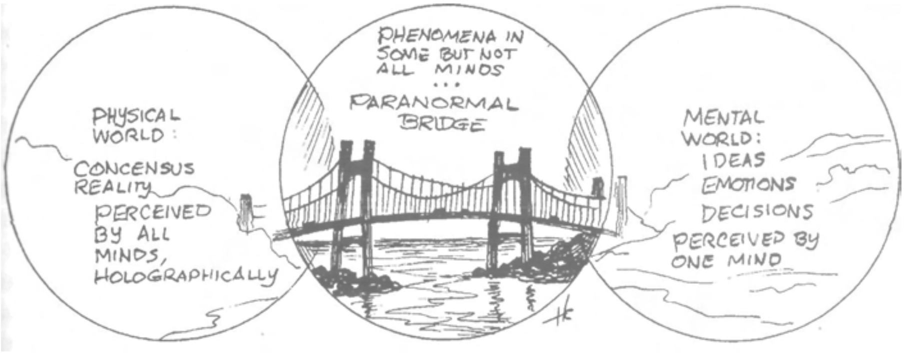
PHYSICAL WORLD: CONCENSUS REALITY PERCEIVED BY ALL MINDS, HOLOGRAPHICALLY
PHENOMENA IN SOME BUT NOT ALL MINDS . . .
PARANORMAL BRIDGE
MENTAL WORLD: IDEAS EMOTIONS DECISIONS PERCEIVED BY ONE MIND

Fig. 3. Reality: a paranormal bridge with two ends

1. Reality as a Paranormal Bridge with Two Ends

In Figure 3, I show the nature of the problem of paranormal phenomena. To understand it, we must first define more precisely some of the present concepts.

First, what do we really mean when we talk of "physical phenomena"? To go into this, we must first destroy the notion that there exists some sort of separate, concrete reality totally apart from mind, for that notion is simply the old Cartesian assumption. It has long been refuted as far as having any absolute validity, and this is well-known to philosophers and to foundations physicists. According to Lindsay and Margenau in their *Foundations of Physics,* "physics has nothing to say about a possible real world lying behind experience."

What is true is that we can experience and detect and know only a mental change — a change in the mind itself is all we ever experience directly. For experience invokes the mind; we do not experience phenomena until the mind — used in the greater sense of the word to mean any mental detection — changes.

Any detector detects or experiences only an internal change, never an external change, and this is true of the mind as well as a physical detector.

So we never experience external physical phenomena; we experience only our own minds. Specifically, each of us experiences his own mental phenomena.

But here we must be a little more precise about what we mean by "mental phenomena." What we usually refer to are the mental phenomena in a single mind. And so we will take that as the primary definition: *mental phenomena are the phenomena or changes in a single mind.*

But then what do we mean by "physical phenomena"? For indeed this refers to something that is often remarkably different — or seems to be — from the phenomena or changes going on in our own heads. Before we jump to the extreme conclusion that physical phenomena are simply mental illusions, we should recall the Zen master's approach. He demanded of the student whether or not physical phenomena were real, and the student replied that they were not real, but only illusions. Whereupon the master gave him a resounding knock on the head, and asked him whether or not the blow was real. The student replied that the blow was indeed real. Whereupon the master sent him back for further meditation, for he had learned nothing. He had not understood that "mental" and "physical" are not mutually exclusive.

If we look closely at how we may differentiate between mental phenomena and physical phenomena, we immediately hit upon a practical scheme. Physical phenomena can be shown to exist in the mind of any observer, be he human, insect, or whatever, so long as he has consciousness.

Suppose for a moment that I have a piece of chalk in my hand, held between my fingers. I can touch the chalk, feel it, and see it. I can even taste it if I wish. And so I hold up the piece of chalk and say: "This piece of chalk is not just a mental impression, it is physically real." How can I prove it? And when I prove it, what exactly have I proved?

Now suppose you do not wish to believe the chalk is real. You simply bring in another observer; he looks at the chalk, he touches it, he tastes it, and he rolls it between his fingertips. He confirms that he also detects or experiences the chalk. Then a little bug flies in and lands on the chalk, crawling around on it, and poking at it with his little proboscis. He also by his actions confirms that he has it in his own little mind. In fact, we could bring in any sort of mind, and that mind could experience the chalk.

And this of course is the clue. What we have been calling physical phenomena can be shown to exist in every mind. And when we say something is physically real, this is all we actually mean. Nothing more, nothing less. And this is why physical phenomena are so solid and concrete; they are infinitely repeated. However, what we refer to as mental phenomena are in only one single mind and thus are not repeated at all.

Now let us do a little thought experiment. Suppose I have lapsed into some peculiar mental state because of the toxins from an infection I have contracted, and I am hallucinating vividly. Suppose I pick up what I vividly see, feel, touch, and taste as a piece of chalk, and I claim that it is real. Indeed, to me it is very real. And then suppose a second person comes into the room. He will not see the chalk, touch it, feel it, taste it, and it will not be real to him. Then suppose a third person comes into the room, and he does not see the chalk, he passes his fingers right through where I am seeing the chalk and does not contact or feel it, and he indeed cannot find it. Thus what we call a "mental phenomenon" can be shown to exist in one mind only. In this case of a vivid hallucination, the hallucinatory phenomenon was absent from the other observers' minds.

But now let us change our thought experiment a little. Suppose now that I am absolutely fit and fine physically and mentally, and am not hallucinating. And suppose all the observers we bring in are also fine both physically and mentally, and not hallucinating. And suppose we get the following results: some of us can see, touch, taste, and experience the chalk, and others of us cannot see, touch, taste, or experience it, no matter how we try. In that case, what kind of phenomenon do we have?

Well, we cannot exactly say the phenomenon is exclusively mental or exclusively physical. But we also cannot say that it is exclusively nonmental or exclusively nonphysical. In fact, we have an example of a third class of phenomena which are presently not recognized by science. *We simply have a paranormal phenomenon,* one which can be shown to violate our definitions of mental phenomena and physical phenomena.

And we can in fact take this as the definition of a paranormal phenomenon: *a paranormal phenomenon can be shown to exist in more than one mind, and it can also be shown not to exist in one or more minds, i.e., paranormal phenomena exist in more than one mind, but not in all minds.* Our conventional science recognizes purely mental phenomena and purely physical phenomena; it does not recognize the third kind, since by the ordinary three laws of logic the third kind does not exist. Paranormal phenomena constitute a bridge across a river between two banks; our present science does not recognize the bridge, but only recognizes the two banks, as shown in Figure 3.

Paranormal phenomena thus appear erratic to ordinary objective science, for objective method after all is specifically designed to select only those phenomena that can be reliably repeated, and the nature of paranormal phenomena is such that they often (in fact usually) cannot be reliably repeated, except sometimes in a great many trials. Normal objective method, if rigorously applied, can only discriminate a statistical effect from a great many trials. Furthermore, if rigorously and lengthily applied, it will accumulate so many failures that it will cast serious doubt on whether the odd case that was paranormal was in fact nonaccidental.

It is for this very reason that orthodox science, which is devoted to scientific method and rigorous protocol, has remained so comfortably adamant that paranormal phenomena do not exist. The tool must be changed to fit the problem, and the present scientific toolbag does not fit precisely. The tools are useful, but they themselves are not decisive within the bounds of present protocol. The necessary change is quite simple: we must allow for the effects of mind, and for a union of both mind and physics, just as there exists in physics a union of both wave and particle. In other words, paranormal phenomena must be repeatable *sometimes,* by competent experimenters and observers, but not necessarily *always* by all observers.

"Ah ha!" the materialist is likely to exclaim, being unable to contain himself any longer. "Solid concrete reality always is solid and concrete, however, and you cannot change that. As long as that is true, then the laws of ordinary physics hold, and all is right with the universe. We can always select our phenomena which are present in all minds, and this after all constitutes immutable physical reality. In the luce of that fact, all your arguments about mental phenomena in one mind and another kind of half-mental, half-physical phenomena are meaningless, for you are simply calling the superposition of purely mental phenomena on purely physical phenomena a third class, but it is a separable mixture, not a completely new kind." And we may even be tempted to stir up a small bit of sympathy for the materialist's view.

However, let us propose a resolution. If we can show that concrete, solid, real matter can become absolutely nonconcrete, nonsolid, and hence nonphysical, then that ought to settle the question. In a recent experiment it has been conclusively shown that two colliding protons pass right through each other in total violation of all physical rules, if their spins are exactly antiparallel. Protons are the central building blocks of all matter. In the nucleus of an atom, the protons and neutrons are continually switching identities, so even a neutron keeps turning into a proton, etc. And these fundamental constituents of everything solid and concrete and real just go right through one another, without any effect whatsoever, simply by aligning their spins properly. Collision theory fails, electromagnetic theory fails, and our concept of two particles bumping together fails. This experiment totally annihilates the idea that physical reality is something absolute. In fact, it establishes that physical reality itself is a function of the relations between the interacting participants; changing the relationships can simply cancel physical reality.

So perhaps our definition of physical reality has some merit after all, and perhaps there really are three classes of phenomena. The paranormal phenomenon is a bridge between mental and physical phenomena; specifically, every paranormal phenomenon must consist of an inseparably welded compound of the two. And the objective method and its consequent experimental protocol must be changed to reflect this fact, if we are ever to develop a viable theory of paranormal phenomena.

2. A Fundamental Correction to Classical Logic

Let us now make a fundamental correction to Aristotle's three laws of logic. First, there is no independent existence to mental phenomena; there is a perception operation involved when we think. There is no independent existence to physical phenomena; there is a perception operation involved when we observe physical phenomena. Furthermore, it takes a finite piece of time for the perception process to occur. The logic symbol also requires a separate time interval; it represents a series of separately perceived operations that together comprise a decision algorithm. So let us impose this criterion upon logic itself so as to constitute "logical perception" or the "logic of perception" or the "perception of logic." We begin with Aristotle's third law of logic, *A* or *not-A,* the law of the excluded middle.

We still insist that there is no such thing as *A* per se, but rather that there is a perceived *A* where *A* is the output of the perception process. Similarly, there is no such thing as *not-A,* but rather there is a perceived *not-A* where *not-A* is the output of the perception process. Let us think of a square box symbol as an abbreviation for the fact that perception has occurred, and anything written inside the box represents the output of that perception operation. We can speak of the box either as mental perception, a description of thought, or we can speak of it as physical detection, a description of an instrumentation system that detects and measures. Also, since each box requires a finite time to occur, we must carefully keep up with the individual little pieces of time, the delta t's.

So applying this to Aristotle's third law, we have *A* perceived or outputted in time one, and *not-A* outputted in time two. Note that to ascertain that A_1 and *not-*A_2 actually differ requires a series of operations in a separate time interval, in time three, that is assumed by the *exclusive* or symbol. Looked at in this way, Aristotle's third law actually is the law of monocularity; it states that only one thing at a time is perceived. Actually we had assumed this when we assumed that perception was a

1. $A \vee \bar{A}$ ARISTOTLE'S LAW OF THE EXCLUDED MIDDLE

2. DEFINE $\boxed{}$ AS "IS PERCEIVED"

3. $\boxed{A} \vee \boxed{\bar{A}}$ LAW OF THE EXCLUDED MIDDLE

4. EACH PERCEPTION IS A FINITE OPERATION AND REQUIRES A FINITE TIME TO OCCUR.

   THE LOGIC SYMBOL ALSO REPRESENTS A DECISION ALGORITHM: A SERIES OF OPERATIONS PERFORMED IN A SEPARATE TIME INTERVAL.

   EACH FINITE TIME INTERVAL MUST BE ACCOUNTED FOR.

5. $\boxed{A}_1 \vee_3 \boxed{\bar{A}}_2$

6. THE LAW STATES THAT PERCEPTION IS A MONOCULAR PROCESS IN TIME 3.

   ONLY ONE-THING-AT-A-TIME IS PERCEIVED.

7. NOW NOTE THAT SEPARATION OF A AND $\bar{A}$ DOES NOT OCCUR IN

$$\boxed{(A_1 , \bar{A}_2)}_3 \quad \equiv_5 \quad \boxed{B}_4$$

8. SO $A_1 \equiv_3 \bar{A}_2 \Bigg|_{4,5}$ BOUNDARY IDENTITY OF EXACT OPPOSITES

9. AND EQUATION 8 CONSTITUTES A FOURTH LAW OF LOGICAL THOUGHT.

   IT SIMPLY STATES THAT PERCEPTION THREE IS BINOCULAR, AND EXCLUSIVE SEPARATION OF EITHER A OR $\bar{A}$ IS NOT PERMITTED.

Fig. 4. A fundamental correction to classical logic

finite process, so it is nice to find that Aristotle's third law justifies our assumption, once we understand the third law. The *exclusive or* symbol assumes a third operation in time three, whereby it is determined that perception output one and output two actually differ. But such an operation itself requires multiocular perception — i.e., collecting two outputs at once — and that in itself is a violation of Aristotle's third law. The third law thus contains its own contradiction, and indeed each of the other two laws also contradicts the third law when one examines them meticulously. Each can only be established as true by invoking or involving an operation wherein the third law is not true.

In time three we gathered up what had been perception output in time one, A_1 and what had been perception output in time two, A_2, which we do not yet know is different from A_1, and put them both through the perception process, getting only one output — let us call it B — in time four. By the nature of B in time four, we thus say in time five that the outputs in times one and two differ or not. In either time one or time two, there is no indication whatsoever of difference or sameness existing between output one and output two. Likewise, in time three there is no separate output one and output two, hence no indication of the sameness of, or difference between, outputs one and two.

So here we have arrived at the identity of opposites. There is no perception of difference between A_1 and *not*-A_2 in time three. And this constitutes a fourth law of logic: the law of the boundary, or the boundary identity of exact opposites. All that is necessary to identify opposites is to lose all perceptual distinction between them. And that is accomplished by multiocular perception, of perceiving the presence of both at once unseparated, hence the absence of either exclusively present. If A_1 and A_2 are exact opposites, then $B_4 = 0$, and $[A_1, A_2]_3 = 0$. We thus have the solution to the problem of nothing. Nothing simply consists of the presence of the totality of all opposites.

Almost all the philosophers who have struggled with the problems of being, mind, and matter have faced the necessity for the identity of opposites, but none of them could understand how opposites could be identified. By careful accounting of the separate time intervals required for finite monocular perceptions, the mechanism for identifying opposites is immediately clarified and revealed. The laws of logic are simply laws of the operation of perception — nothing more, nothing less.

The new system of logic is closed. All present paradoxes — contradictions of one or more of the first three laws — are resolved by the fourth law, which contains the negation of each of the first three laws. The fourth law is in fact the law of the paradox. Note also that the hidden time-three operation, which has actually been the application of the fourth law all along, is implied in each of the first three laws. Identity or nonidentity between time-one and time-two outputs can only be established in a time-three operation. The fact that either A or *not*-A exclusively exists can only be established by a separate operation which establishes that nothing else is there. If separation of A and *not*-A is not permitted, then A and *not*-A cannot be distinguished.

Since these laws refer to perceptual operations, one can think of them operationally, or vectorially. To close the vectorial system prescribed by the first three laws, the opposite or negation of each of the three vectorial statements must be present, i.e., this follows simply from the definition of what constitutes a closed system, vectorially speaking. Since the fourth law contains the negation of each of

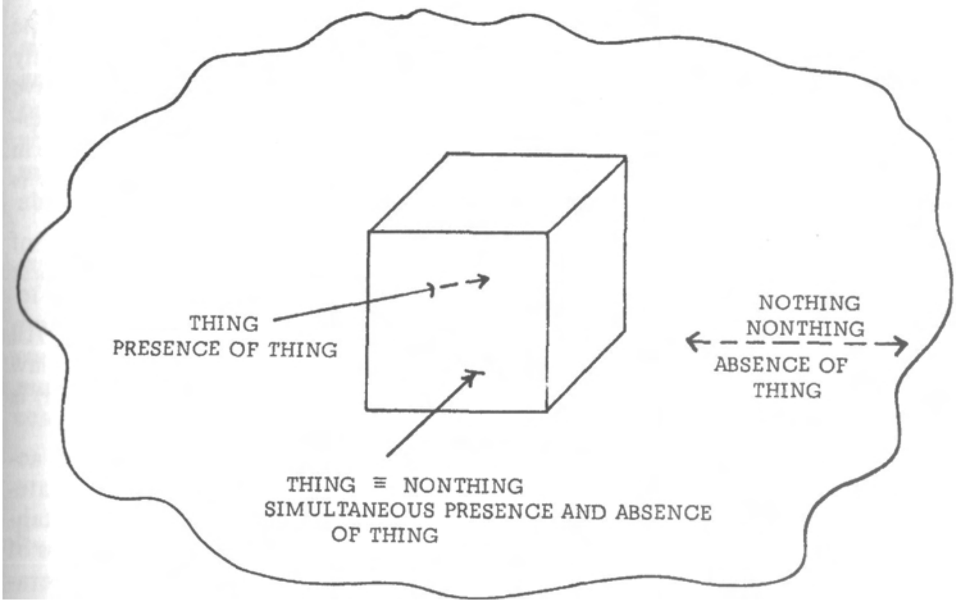

Fig. 5. A physical example

the first three laws, then the four-law system is indeed closed, and the logician's dream of a closed metalogic is realized. Furthermore, anything which contradicts any combination of the first three laws automatically is covered by the fourth law, which is the law of the paradox.

The new logic works as follows: either the first three laws apply (separation of *A* and *not-A* is permitted), or the fourth law applies (separation of *A* and *not-A* is not permitted). The fourth law applies only to — and in fact *creates* — a boundary. The first three laws apply only away from a boundary.

3. A Physical Example

Take the surface of a cube in deep space (Fig. 5). Call the cube "thing," a three-dimensional concept. Call the empty space around the cube "nonthing," a three-dimensional nonthing or absence of thing. If we are standing inside the cube and look at its boundary surface, we cannot find a single piece of that boundary surface that does not belong totally to the cube. So we can very reasonably proclaim that by the first three laws of logic each piece of the boundary belongs totally to the cube, to "thing." But if in a different operation we are standing outside the cube, we cannot find a single piece of that boundary surface that does not belong entirely to the space surrounding the cube. So in this case, we can claim by the first three laws of logic that the boundary surface belongs totally to "nonthing."

Then in a third operation we can state that, by the first law of logic, each and every piece of the boundary surface is identical to itself, and of course we thereby identify what was thing with what was nonthing. Specifically, what was thing in perception time one and what was nonthing in perception time two have been identified, by all distinction and separation between them being removed, in time three. Identifying opposites simply consists of "packing together" two previously separated perceptions into a single third unseparated perception. And all we have done is apply the fourth law of logic, the law of the boundary. Every single perceived thing has a boundary, where it both begins and ends its exclusive presence in perception output. And at that boundary, the fourth law applies. Thus the law is universal. The fourth law defines a boundary.

In fact, in any perception, all four laws are applied. They are applied in one of two fashions: (1) the first three are explicit and the fourth implicit, or (2) the first three are implicit and the fourth explicit.

There are more examples that have baffled mathematicians and logicians. All of these are simply boundary statements, i.e., statements involving the fourth law of logic. For example: "It is true that this statement is false"; "A line (length) is composed of points (nonlengths)"; "In a hologram, each part is the whole."

Since logicians used only the first three laws, none of these statements was acceptable or understandable. By the fourth law, there is no problem with these statements. The first merely refers to the boundary between the operation used to establish "truth" and the operation used to establish "falsity." There is another class of operation where neither truth nor falsity exclusively applies — i.e., a class of operation which cannot distinguish between "true" and "false."

The second statement has long been a painful trauma to logicians and foundations mathematicians. Today foundations mathematicians seldom attempt to define line or point. Instead, they simply say: "There is a class of entities called lines. There is a class of entities called points. Lines are made up of an infinite number of points." This way, they avoid trying to explain how "length" (line) can be made up *of* "nonlength" (points).

However, they could easily have avoided the difficulty, and still retained the definition of a line as a length, and a point as a nonlength, in the following manner. "There is a class of entities called lines. A line is and has length. There is a class of entities called points. A point is and has nonlength. We may have a separate perception that shows nonlength, and another separate perception that shows length. In yet a third perception which infolds both the first two perceptual characteristics, we may say that a line is comprised of points so long as we realize that line and point do not separately and exclusively exist in the comprised-of state."

By the same token, we can say that "the whole is comprised of the sum of the parts" only if we realize that "whole" and "part" do not separately and exclusively exist in the comprised-of state, but instead they inseparably and nonexclusively exist therein.

4. New Definition of Zero

To a monocular perception process, multiple presence constitutes absence of the exclusive presence of any particular one. Therefore such a multiple presence is monocularly unperceivable, and hence becomes a zero to a monocular detection process. This allows new definitions of zero, and a solution to the problem of nothing.

Consider that a monocular detection process asks the question, "Is there a single exclusive thing present in my input?" If the answer is yes, an output is generated and perception occurs. If the answer is no, no output is generated and perception does not occur. The answer "no" occurs in two fashions: either total absence, or presence of two or more simultaneously. For either of these cases, monocular perception gives no output, and perception does not occur, i.e., the absence of perception occurs.

Now note that the monocular perception cannot tell any difference in the two input conditions. To it, there is no distinction between the two conditions. The lack of any difference at all constitutes identity. Thus to a monocular perception process, condition one is identical to condition two when they are infolded together in condition three. That in fact derives the fourth law of logic. Total absence and multiple, unseparated presence are identical insofar as a monocular detection process is concerned.

5. Einstein's Postulate

As an example of strange problems we can deal with in a four-law manner, let us examine a little more closely one of Einstein's postulates of special relativity, which states that the speed of light is the same for every observer. Let us restate the postulate as "every photon is moving at the same speed relative to every particle." We can pick any electron in the laboratory frame. Every photon in the frame is moving at the speed of light with respect to the electron, by Einstein's postulate. Now let us pick any photon. By Einstein's postulate, every electron is also moving at the speed of light with respect to this photon.

Now we can look at the situation from the viewpoint of the photon. With respect to it, every electron in the laboratory frame is moving at the speed of light, which violates the common interpretation that ordinary objects cannot move at the speed of light. In fact, three-law logic is violated but four-law logic is not violated. In four-law logic, an object can have infinite mass and zero mass simultaneously. The electron's frame and the photon's frame are rotated orthogonal to each other. What we were calling a photon in the laboratory frame, is a perfectly ordinary three-dimensional object in its own frame, which is orthogonal to the laboratory frame. The electron (lab-frame) thus appears as a photon in the rotated frame of what we previously called a "lab-frame photon." Either an electron or a photon is both three-dimensional and two-dimensional simultaneously — in fact, the concept of separate, exclusive dimensionality only applies after one or the other dimensional aspect has been exclusively separated (observed).

Mass is determined by the resistance an object poses to an accelerating or disturbing force. Mass is also tied to three-dimensional objects; i.e., it is an L^3 concept dimensionally. A photon, moving at the speed of light, can neither be speeded up nor slowed down; hence in one sense it exhibits infinite mass. On the other hand, the photon (which is a three-dimensional object in its own frame) has lost a dimension in its intersection in the laboratory frame, due to orthogonality. Therefore the photon appears as a two-dimensional L^2 entity in the laboratory three-space. And in that sense, the photon must have zero mass since it is only two-dimensional. By the fourth law, the boundary opposites are identified. Thus at the orthogonal boundary of three-space, zero mass and infinite mass become identical. And the photon exhibits both zero mass and infinite mass simultaneously because it is an entity that is on the boundary of the L^3 "mass exhibiting" world. The photon has always happily behaved in a four-law manner, even though physicists could not comprehend its behavior with the three-law logic ingrained in their heads.

6. Synchronicity of Concepts

As is so often the case, two persons appear to have derived the new logic simultaneously, myself from perception theory and Hubbard from manifold theory.

Dr. J. Hubbard was one of the scientists who first developed the atomic bomb. His profound work fully substantiates the new logic and the new reality paradigm. Dr. Hubbard believes that the system includes language, logic, and the interaction of life-bearing systems. In short, he believes it includes psychotronics. The relationship between mind and matter is, like Santilli's electric and gravitational fields, not mutually exclusive; they are either partially or totally one and the same thing. According to my research, they are both partially and totally the same thing, i.e., they are of exactly the same nature totally — a closure of the entire physical universe — but only partially intersecting, by a fraction of approximately 10^{-42}. In the same fashion a mind is an entire physical world, but it only minutely intersects this physical frame, and then only in a portion of this physical plane.

Dr. Hubbard has shown that the four laws of logic emerge in the form of set relationships, that the single dimension parameter generates the second, third, fourth, fifth and up to *Nth* dimensions of the manifold, by repetitive operations on itself. This allows a firm basis for orthogonal intersections of one less than the dimensionality of the lowest-dimensioned intersector. It provides a firm basis for orthorotation and for my approach to psychotronics.

PRIMITIVE PERCEPTION

Let us now discuss how we form our concept of objects and their relationship. All our basic perceptual concepts are essentially built on primitive observation of the macroscopic universe, which is the primary reason that so many final arguments of philosophers appeal to the "natural man," i.e., to how a primitive observer would react to the argument. The very idea of an object — a perceived object — is conditioned. Everything else we think of is then conceived as some sort of object or relation between objects. And that is why sets (of perceptual objects) and their relationships form the basis for axiomatic logic, mathematics, and physics. Set theory is simply the synthesized and abstracted game of elementary perception of the macroscopic world. Thus even the most abstruse mathematics forms a part of the game of perception and can be applied to perceived physical phenomena. This macroscopic, primitive perception is the basis of our concept of objectivity, and of our physical concepts of "observed" and "observable." But if we examine

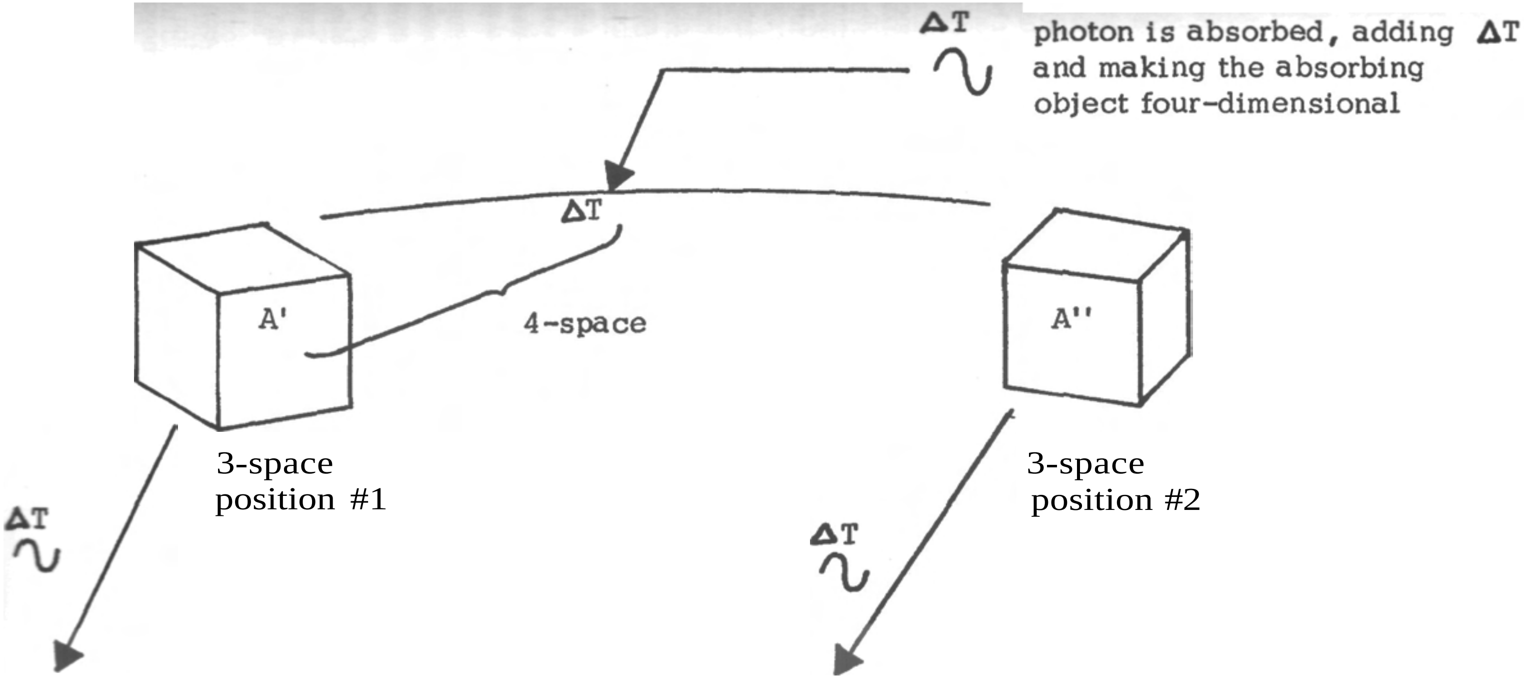
ΔT
ΔT
A'
4-space
A''
3-space
position #1
3-space
position #2
ΔT
ΔT

photon is absorbed, adding ΔT and making the absorbing object four-dimensional

Photon has previously been emitted, and traveled to the observer's eye, so that he detected a corresponding change on the retina, which "detected object A in its spatial position #1"

Photon emitted, traveling to the observer's eye, so that he may detect a corresponding change on the retina, which "detects object A in its spatial position #2'

Fig. 6. Primitive perception

carefully exactly how we perceive an object in the macroscopic world, we realize that it is seen by means of photon interaction. We think of light being absorbed onto the surface of the object, and then emitted or reradiated from the surface, eventually entering the eye. There the photons are again absorbed by the material on the retina. The photoelectric effect then moves what are called electrons, producing a signal that goes to the brain for interpretation

What we see as an object has twice undergone interaction with photons of light. Furthermore, light itself cannot be detected, only its interaction with matter (in this case electrons). Thus the primitive perceptual operations of the human brain are correlated one-to-one with photon emission and absorption interactions, usually paired interactions at the object reality domain. The concepts of object, mass, three-dimensions, and observation are directly correlated to the photon-mass interaction, and ultimately to the photon-electron interaction occurring in the human retina.

The ideas of boundary, surface, and change in primitive perception thus involve two contradictory and simultaneously mixed interactions, absorption and emission. Because they are constantly intermingled, primitive perception cannot separate the two. Thus the primitive concept of a boundary consists of a simultaneous duality. Here the object both begins and ends. Here opposites are identified. This is the basis for the fourth law of logic, the law of the boundary.

The dimensionality and orthogonality (spatially) of the primitive perception process are also interesting. Photon absorption constitutes spatial integration for the photon, dimensionally speaking, *e.g.*, one additional dimension is gained by the two-dimensional light wave in turning itself into the three-dimensional mass state. The photon also adds its piece of time to the absorbing mass, converting the mass particle to mass time or space time—adding the fourth dimension. Photon emission constitutes spatial differentiation, dimensionally speaking. One spatial dimension is lost by a little hunk of three-dimensional mass turning itself into a two-dimensional wave. Going from outside the object toward the inside (to track along with photon absorption), the boundary marks the end of the two-dimensional wave region, so the object (or, more precisely, the interaction region) is spatially two-dimensional on the outside, and spatially three-dimensional on the inside. It also occupies the time dimension on the inside, but not on the outside. (This directly accounts for a difference in the spacetime inside and the spacetime outside, and that is why mass may be regarded as a curvature or change in spacetime, in general relativity.)

Going from inside the object toward the outside, the boundary marks the end of the spatially three-dimensional region and the beginning of the spatially two-dimensional region. So the object is now spatially three-dimensional on the inside, spatially two-dimensional on the outside. The inside now does not occupy the time dimension, while the photon emitted (the outside) does. So on the boundary, one spatial dimension is gained going in, and one spatial dimension is lost going out Time is gained going in and lost going out, by the mass absorber-emitter.

Gain or loss of a dimension is accomplished only by orthogonal rotation between orthogonal spatial frames. Photon absorption constitutes orthogonal spatial rotation one right angle turn into the laboratory frame, while photon emission constitutes orthogonal spatial rotation one right angle turn out of the laboratory three-space frame.

This constitutes the observation process (primitively) per se, and this is what is hidden in Einstein's postulate that the speed of light is the same to every observer. Every observer goes through this same process. Furthermore, rotations orthogonal to a three-space frame are orthogonal to all three-dimensional spatial physical entities in that three-space. As Einstein himself once pointed out, velocity can be regarded as simply the measure of rotation of an entity in a higher dimensional space; thus an orthogonal turn is a fixed velocity. Since light is orthogonal to the entire space, then the speed of light is the same to every observer in that space.

While Einstein was apparently speaking of regarding the velocity of an object as a measure of its amount of rotation toward the time axis in four-space (from the direction of its velocity vector in laboratory three-space), it can also be regarded as rotation toward a fourth spatial axis in a five-space (four length dimensions and one time dimension, all orthogonal). If one checks this, one will see that the projections in three-space are the same; i.e., to the laboratory frame observer, there is no difference observable. Thus the photon interaction may be modeled as orthorotation of an entity spatially in a five-space.

To the laboratory observer, this orthorotation constitutes pure dimensional integration and dimensional differentiation, and it constitutes or creates an object in the first place. The concept of object *a priori* involves the photon interaction as stated; if the photon interaction is not invoked, then the existent entity does not exist as a three-dimensional object at all, in the primitive sense. Outside this primitive process (dimensional integration and differentiation imposed upon the same region, which constitutes an object), all exists in the Omega nothingness void, without frame or form.

Ultimate reality (specifically, action) is dimensionless in the objective sense; for the objective sense exists only after an operation has been imposed. A quantum itself is lengthless (and has no exclusive energy, time, or momentum) until differentiation is imposed upon it. Only after a quantum of action is fissioned does there exist a discrete piece of (change in) length, time, energy, mass, momentum, etc. I have already shown that perception itself can be modeled in terms of action quanta fission and fusion, and that most of the basic laws of physics can be derived from that model. (See my *Quiton/Perception Physics.)* Fission and fusion of action quanta involve orthogonal rotation between orthogonal three-spaces sharing the same fourth dimension (time) in common. The words *change, interact, localize, superpose, operate, move, oscillate, act, cause,* and *determine* are only a few of the words used in physics to conceal the basic occurrence of orthogonal rotation in higher dimensional space.

Finally, it should be stressed that i, the square root of minus one, itself constitutes an orthogonal axis, hence an extraspatial dimension. The addition of this quantity to electromagnetic theory allowed the resolution of formidable problems, and led to the direct formulation of electromagnetic theory. Thus electromagnetic theory already includes higher dimensionality and orthogonal rotation in hyperspace. In that sense, inductance and capacitance are merely vector statements of amounts of rotation, one in a positive direction and the other in a negative direction, toward an orthogonal axis in higher dimensional space.

The forming (beginning and ending) of a three-dimensional spatial object constitutes a time differentiation invoked on four-dimensional Minkowskian space, i.e., the process of observation or detection itself involves

$$\partial/\partial T\ (L^3T) = L^3$$

dimensionally speaking. This process is accomplished by orthogonal rotation, and the very concept of an object that has been observed, detected, or perceived implicitly involves this time differentiation invoked on a four-space.

Velocity also represents a statement of the switching of time into length, and length into time. It is therefore a measure of rotation of the spatial three-dimensional frame, as Einstein pointed out. By this switching (orthogonal rotating, orthogonal flipping) of the chopping of time or beginning and ending of a small piece of time, we create a little delta *T*, just as we do each delta *L* for an object.

But the photon interaction process is monocular. Each photon must attack matter individually, and live or die in the process. Each photon born from matter is also born independently. Thus the primitive observation process is monocular, since it is totally in one-to-one correspondence with photon interaction that is monocular. Thus the observed world is monocular, quantized, discretized. Because of the particular one-to-one correspondence between our primitive concepts and a single interaction, we can conceive the world in no other fashion, unless we change the very basic factor of one-to-one correlation. This in fact has already been done for physics by Everett, whose theory of the universal wave function incorporates multiple simultaneous observation, and hence direct insight beyond the limitations of the monocular correspondence to the photon interaction. My theory does the same thing in the fields of logic and mind.

TWO-SLIT EXPERIMENT

This experiment is fundamental to all of modern physics. Feynman, Nobel prize winner in physics, has stated that no physicist understands this experiment, and that it cannot be explained by any classical means. The reason is that it cannot be monocularly comprehended, i.e., the first three laws of logic cannot explain it. The fourth law can and does.

In the experiment, electrons are emitted from a source and travel past a doubly-slit wall region on their way to a screen. The apparatus is shielded against light. If one believes that the emitted electron is a little three-dimensional particle, much like a tiny baseball, then it should go through one of the slits and not the other. It would then hit the screen at one of the two spots indicated as the expected distribution, with a little scatter from those that chip the edge of the slit a bit. Electrons which do not hit the holes but strike the wall are absorbed.

We do not get this expected pattern. Instead, the pattern is essentially the same as the one we would get if each electron were a wavefront passing through both slits at once. However, each electron still strikes the screen in only one point; the distribution of these points fits the actual distribution pattern shown.

This stunned the physicists. They did not believe it at first, so they set up a photon gun and hit every electron with a photon as it left the emitter source and started over toward the two-slit region. That is, they determined precisely when a little electron was on the way, and the fact that it was like a tiny baseball and in just one

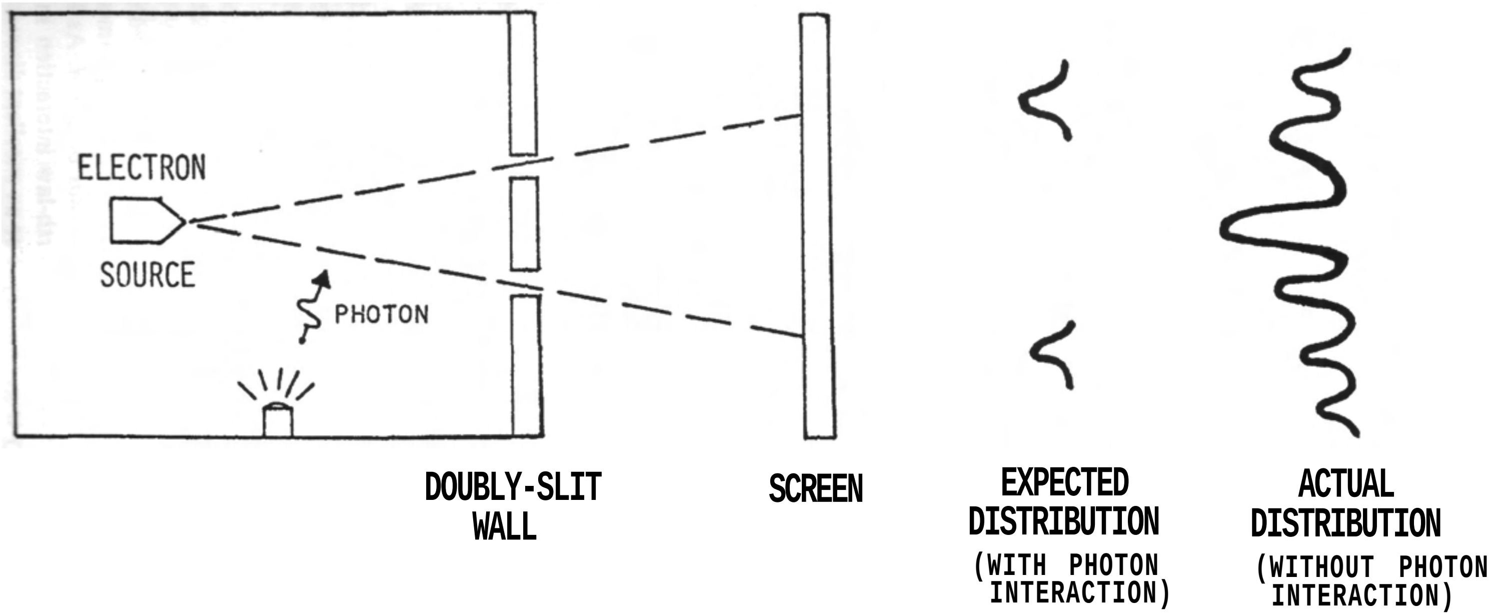
ELECTRON
SOURCE
PHOTON
DOUBLY-SLIT WALL
SCREEN
EXPECTED DISTRIBUTION
(WITH PHOTON INTERACTION)
ACTUAL DISTRIBUTION
(WITHOUT PHOTON INTERACTION)

NOTE: IF THE ELECTRON IS HIT BY A PHOTON BEFORE IT REACHES THE TWO-SLIT REGION, IT DOES NOT EXHIBIT THE WAVE INTERFERENCE EFFECT, BUT INSTEAD ACTS AS A CLASSICAL OBJECT AND GOES THROUGH ONLY ONE SLIT, YIELDING THE "EXPECTED DISTRIBUTION" PATTERN.

Fig. 7. Two-slit experiment

place. And this time the electron only went through a single slit, and it gave the expected pattern after all. And when the experiment was repeated and only a fraction of the electrons were hit with photons, then a mixture of the two patterns emerged.

It is simple to describe the results mathematically, but no one has understood why things happen as they do in this experiment. The principle of complementarity evades the issue. This principle is simply a monocular statement dealing with one aspect of the problem at a time, with the determined, exclusive, monocular past only. It does not apply to the present or to the future. If we think in terms of the present, then the third law of logic is violated and the fourth law applies. The two states, two-dimensional wave and three-dimensional corpuscular, both exist simultaneously but nonexclusively in the present. Thus in physics terms, the entity becomes nonobjective (nonspatial), probabilistic and undetermined, and this is automatically a wave concept, i.e., since waves are not confined to one place and determined or localized, they may exist in the nonconfined, nonlocalized present.

Note that photon interaction must be excluded whenever the so-called quantum effect is exhibited. If photon interaction is invoked, classical reality emerges. Photon interaction creates classical objectivity. Photon interaction is a time-differentiating operator imposed on L^3T four-space. Specifically, since light carries time, photon emission from an entity strips away T from the L^3T nonobjective entity, converting it to an L^3 or spatial object.

The very concept of an object comes from primitive perception's one-to-one correspondence with photon interaction. After photon interaction, the first three laws of logic apply to an entity. Before photon interaction occurs, the fourth law applies. In the fourth-law state, the time portion of an entity can interact with any number of time portions of other fourth-law entities if the time aspects of all of them coincide.

A free electron born and released at the emitter is four-dimensional (L^3T) until it is struck by a photon, after which it is three-dimensional (L^3). When the wavelength of electron and slits is specified, we have actually specified the time interval stripped out of a dEdT quantum in each quantum of fundamental change occurring. Perfect time synchronization accomplishes or constitutes orthogonal rotation. So if the length (width) dimension of each slit is very close to the wavelength of the electron, the time aspects of both slits will interact with the time aspect of the fourth-state electron if all three are brought into time phase simultaneously. If the slits are made much larger, this interaction will not occur.

When the electron interacts with both slits timewise, this will constitute a part of the past history of the electron to all future interactions or interaction possibilities. But since the interaction with the two-slits was not in the past (i.e., selected or determined), then that interaction itself is a part of the present and future probabilities chain. In physics, probabilities propagate forward in time with absolute causality until a monocular selection is made. Thus the fourth-law interaction is propagated forward with absolute causality and significantly affects any future interactions. And apparently vice-versa. John Wheeler has just shown that, in at least one sense, whether or not the emitted particle in the two-slit experiment has interacted with one or both slits can be selected *after it has occurred.* As is well known, the pattern of projection forward of this fourth-law interaction may be quite simply calculated from ordinary trigonometry. For an excellent discussion of the two-slit experiment, see Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics*, Vol. 1 (Addison-Wesley, Menlo Park, Calif., 1963), pp. 37-1 and 37-12.

When the electron hits the screen, it encounters a region of varying time oscillations of the orbital electrons around the individual atoms comprising the screen. Thus the exact location of the orbital electron in the screen which will first precisely time-synchronize with the electron wavelength approaching the screen in four-law form will vary. Thus the place where the electron hits the screen is variably selected along the screen. The time distribution pattern of the approaching electron is recovered when the time distribution of the number of electron hits per screen length is plotted.

So one can build gadgets to cause four-law entities to multiply and interact simultaneously in time, even though the entities are normally thought of as being in different spatial locations when in the three-law state. And one can deliberately select the type of interaction to occur, four-law or three-law, simply by controlling photon interaction. The two-slit apparatus itself is a paranormal, psychotronic device.

This is the explanation of the two-slit experiment, which, according to Nobel prizewinner Richard Feynman, no one understands and which, according also to Dr. Feynman, contains the total mystery of quantum mechanics and the only mystery of quantum mechanics. The reason the two-slit experiment has not been understood is that the answer to the paradox was not present in the first three laws of logic. It requires the addition of the fourth law to explain the experiment.

And all fundamental particles — photons, neutrons, protons, electrons, etc. — exhibit the same behavior. So things, nothings if you will, can be processed in the multiplistic two-states-identified-as-one-so-no-single-exclusive state, as virtual and unobserved entities. In the multiplistic state they can be amplified, recorded, put onto tape, modulated onto RF signals, etc. And by our ordinary, objective, three-law science nothing is processed. And nothing indeed is what is there — a very special, structured, detailed, virtual set of nonexclusive, hidden-variable nonthings that is part of the new reality existing in the framework of the new four-law logic. In fact, this processable, structurable, nonobjective reality exists outside objective, three-law spacetime and is the basis for all psychotronic phenomena.

To our present monocular detection devices and monocular theory, such multiple-state entities are unobserved and hence are zeroes. They are pure vacuum, pure space, pure nothing, pure emptiness. But they are very real indeed, and they do physically exist, but multiocularly rather than monocularly.

The importance of the time interaction in explaining the two-slit experiment was noted as early as 1957 by Charles Muses in his introduction to Jerome Rothstein's *Communication, Organization, and Science* (Falcon's Wing Press, Indian Hills, Colorado, 1958). Muses pointed out that the celebrated wave-particle paradox remains a paradox only so long as the chronotopological phases of the phenomena are left unrealized in the analysis. The entire foreword by Muses is a remarkable document that analyzes the structure of time itself.

NOTE:

IN REFLECTED LIGHT, THE EFFECT CAN BE INVERSED.

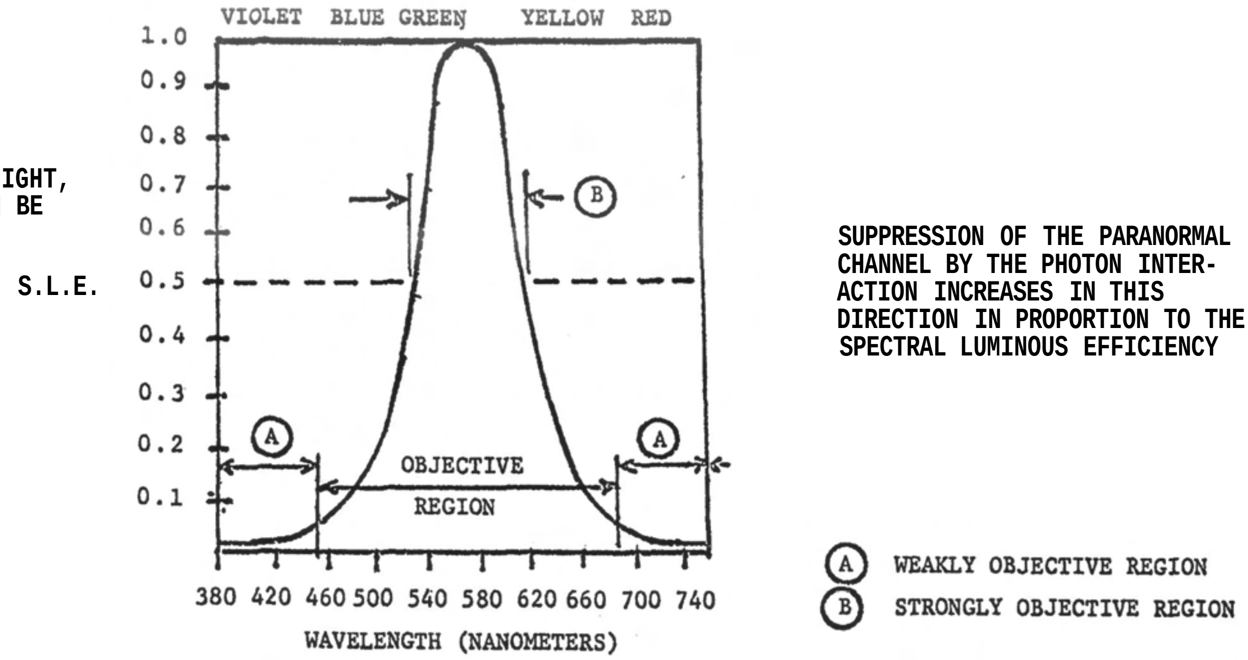

SUPPRESSION OF THE PARANORMAL CHANNEL BY THE PHOTON INTERACTION INCREASES IN THIS DIRECTION IN PROPORTION TO THE SPECTRAL LUMINOUS EFFICIENCY

Fig. 8. Photon quenching of the paranormal

PHOTON QUENCHING OF THE PARANORMAL (TIME) CHANNEL

The photon interaction constitutes a time-differentiating operator imposed upon L^3T dimensionally (Minkowskian reality), which results in the creation of L^3 dimensionality (objective, spatial reality). Aristotle's three laws of logical thought are based on primitive observation, which in turn is in one-to-one correspondence to the situation where the photon interaction has been imposed and the time dimension destroyed, resulting in a classical or objective or spatial reality. Hence all our present logic, science, and mathematics are based upon this particular arrangement. As shown by the two-slit experiment, if the photon interaction is not allowed to occur, a fundamental particle refuses to behave in a classical, objective, spatial object manner; instead, it behaves in a nonclassical, nonobjective, wave manner if the time element interacts with it.

Mental phenomena occupy the same time dimension as do physical phenomena; however, the spatial dimensions of mental objects do not ordinarily occupy or intersect the spatial dimensions of physical objects. However, the time dimension is recognized by physics as objective; therefore, since the mind occupies at least one real objective dimension, it is objective rather than subjective. This suggests a model whereby mind is a real L^3T Minkowskian four-space world, but one whose three spatial dimensions lie orthogonal to and outside the ordinary world's three spatial dimensions. The mindworld and the physical world then have in common the same time dimension. Dynamic movements in each world (mind and matter) result in an exceedingly small crosstalk being projected into the other world, a crosstalk so small as to be virtual and normally unmeasurable. Establishing one-to-one coherence between crosstalk from a mindworld constitutes the creation of a biological organism. The crosstalk from the mindworld and a physical structure in the ordinary world to the physical world has previously been labeled *inception* by me, and a mechanism whereby such a minute psychokinetic effect on matter can be amplified to a macroscopically observable level, or even beyond, has been derived. Thus a fundamental mechanism whereby tulpoid activity (projections from the collective unconscious materializing into the ordinary physical world, in a format determined by the collectively higher levels of unconsciousness in the race, national group, culture, subculture, and personal unconsciousnesses) is developed and explained.

Before the intervention of the photon interaction, four-dimensional reality thus has a great deal of tulpoid dither riding upon its time dimension as a minute modulation. The photon interaction invokes a time-differentiating operation, stripping away or suppressing the time dimension, resulting in a spatial reality or objective reality being perceived or observed, and in the process separating mind and body by destroying the only common connection or channel. Because of the time-squelching action of the photon interaction, we do not see time or the mind in our ordinary physical observation. Also, we do not see the incessant stream of tulpoid activity that bathes the time channel.

The time channel is thus the mental channel and the paranormal channel. The photon interaction is the destroyer or squelcher of the paranormal channel. Apparently the squelch effect of light is not linear or constant, but is a variable and a function of the frequency of the light being utilized. From a synthesis of a variety of data in diverse fields, the following hypothesis has been arrived at to fit the experimental evidence: *The quenching of the hyperchannel (the paranormal channel or the time channel) by photon interaction is proportional to the spectral luminous efficiency of the eye of the average human observer in the interval from the infrared to the ultraviolet.* Thus the form of the paranormal squelching effect of the photon interaction is given by the operator ϵ where

$$\epsilon = m \mathrm{E} T$$

and where *m* is a parameter which varies as some function of frequency. This hypothesis specifies *m* over the region from the ultraviolet to the infrared. In other regions of the frequency spectrum, the form of *m* must be determined by experiment.

There are several experiments which fit the hypothesis:

1. The induction of death and disease patterns from test cells to environmentally shielded healthy cells by Soviet scientists.[a]
2. The invoking of specific tulpa living forms and photographing them in the invisible state with infrared film by Trevor James Constable.[b]
3. The visible light revival of appreciable fractions of cells killed with ultraviolet light and kept in the dark — dormant, unreproducing, for up to twenty-four hours.
4. Ingo Swann's precise differentiation of the visible light spectrum (frequency region) where psychic seeing (similar to remote viewing) is blacked out.
5. The Fox and Miller experiments, in which cell-like protein forms that reproduce (i.e. are "living") are generated when a primeval soup in the laboratory is bathed in ultraviolet light and exposed to strong electrical spark discharge (strong bioenergy collections).

A corollary to this hypothesis is that, under certain conditions, reflected light may reverse the effect. Thus the effect of the full moon, which reflects a great deal of sunlight, can immediately be appreciated. The effect is to the time channel, hence stimulation of this channel stimulates the mental effects. The effects of the full moon on romantics, lunacy, and deeper life functions controlled by the unconscious is grounded in a factual basis.

Thus there is a very good reason why animal eyes tend to operate very closely to the infrared, visible light, ultraviolet region, with the eye usually adapted for the visible light region. In the infrared and ultraviolet regions, tulpoid phenomena would be readily visible or detectable, and an animal whose eye operated exclusively in one or the other of those areas would so often be distracted by the mental tulpoidal activity that it would prove disadvantageous to his survival against animal predators operating in the strongly squelched, visible light region. By the same token, animals whose eye operates further into the infrared or the ultraviolet than does the human eye are notoriously sensitive to paranormal phenomena before their manifestation to human observers.

The stimulating effect of reflected light, and particularly of light reflected from the surface of a planet or of the moon, also concerns several of the common practices of UFO investigators. It is common practice to dispose of a sighting once it becomes apparent that a planet lay in the direction of the sighting and was particularly bright at the time. However, such an optical condition can in fact serve as an initiation point for tulpoid activity to enter the ordinary reality frame. And in fact such conditions — and conditions when stars are very bright and twinkly in their reflections from the particles of the atmosphere — are conducive to UFO appearances, and this should be realized and taken into account. UFO's in a great many cases will be observed to start or come into being from such appearances.

---

[a] **V.P. Kaznacheyev et al., "Distant intercellular interactions in a system of two tissue cultures," *Psychoenergelic Systems,* Vol. 1, No. 3, March 1976, pp. 141-42.**

———, "Apparent information transfer between two groups of cells," *Psychoenergetic Systems,* Vol. 1, No. 1, December 1974, p. 37.

[b] **Trevor James Constable, *The Cosmic Pulse of Life* (Merlin Press, Santa Ana, Calif.)**

If we realize that the paranormal channel is the time channel, then the fundamental two-slit experiment of physics is itself a paranormal manifestation experiment. Thus when the photon interaction is invoked, a classical object and classical behavior result. When the photon interaction is not invoked, a nonclassical entity and nonclassical behavior result. The more interest is taken in a particular tulpoidal manifestation, the more the format is impressed in multiple minds and multiple unconsciousness layers and the more frequently the format is tuned in. Eventually a repeatedly tuned tulpa can become stable and permanent. Candidates for such tulpoids presently on their way to being stabilized are the Loch Ness monster and the sasquatch. Particle physicists may in fact be creating their own reality by the same process, as suggested by Josephson.

The process whereby visible light revives cells killed by ultraviolet light is presently considered under the term "photoreversal." It is a moot point whether cells are living if they have laid dormant for several hours or even twenty-four hours without any cell divisions having occurred. The effect of a given dosage of ultraviolet radiation on cells is known to be greater when the radiation is flashed than when the radiation is continuous. This has been interpreted as due to a thermal reaction following a primary photochemical reaction, since increasing the temperature accentuates the effect of flashing. However, I suggest that flashing is directly related to the effect known as kindling, where, when the limbic structures of the brain are repetitively stimulated, the entire nervous system of the organism responds and goes into resonance (seizure). Soviet experiments have shown that such flashing of a telepathic sender enhances the reception of telepathic information. In the 1920's and 1930's A.G. Gurvich and his students found that a yeast detector's threshold of sensitivity to mercury lamp radiation was about three orders of magnitude higher than the detector's threshold of sensitivity to radiation of the same spectrum from biological material. (For a description of more modern work which may show a violation of the conservation of energy law, see L.L. Regelson, "A Paradox of Mitogenetic Radiation," edited by H.S. Dakin, translation available from H.S. Dakin, 3101 Washington Street, San Francisco, CA 94115.) Regelson's experimental results point to the conclusion that the paradox observed with the mitogenetic radiation experiments is a universal physical phenomenon. In view of this and Kaznacheyev's results, the entire area strongly warrants intensive investigation. Regelson's group was actually attempting to find a new solution to the problem of wave-particle duality, strikingly parallel to my interest in the area.

Those frequency points where the quenching effect is minimized (i.e., where € goes through a minimum value) are referred to as "magic windows" or "magic windows between worlds." Acker and Tiller have called them "interdimensional nodal points." In the magic windows, energy from one frame or biofield can be more easily collected to orthorotate into another frame or biofield, since the rotational threshold is much lower. There is a window, e.g., between 38,000-40,000

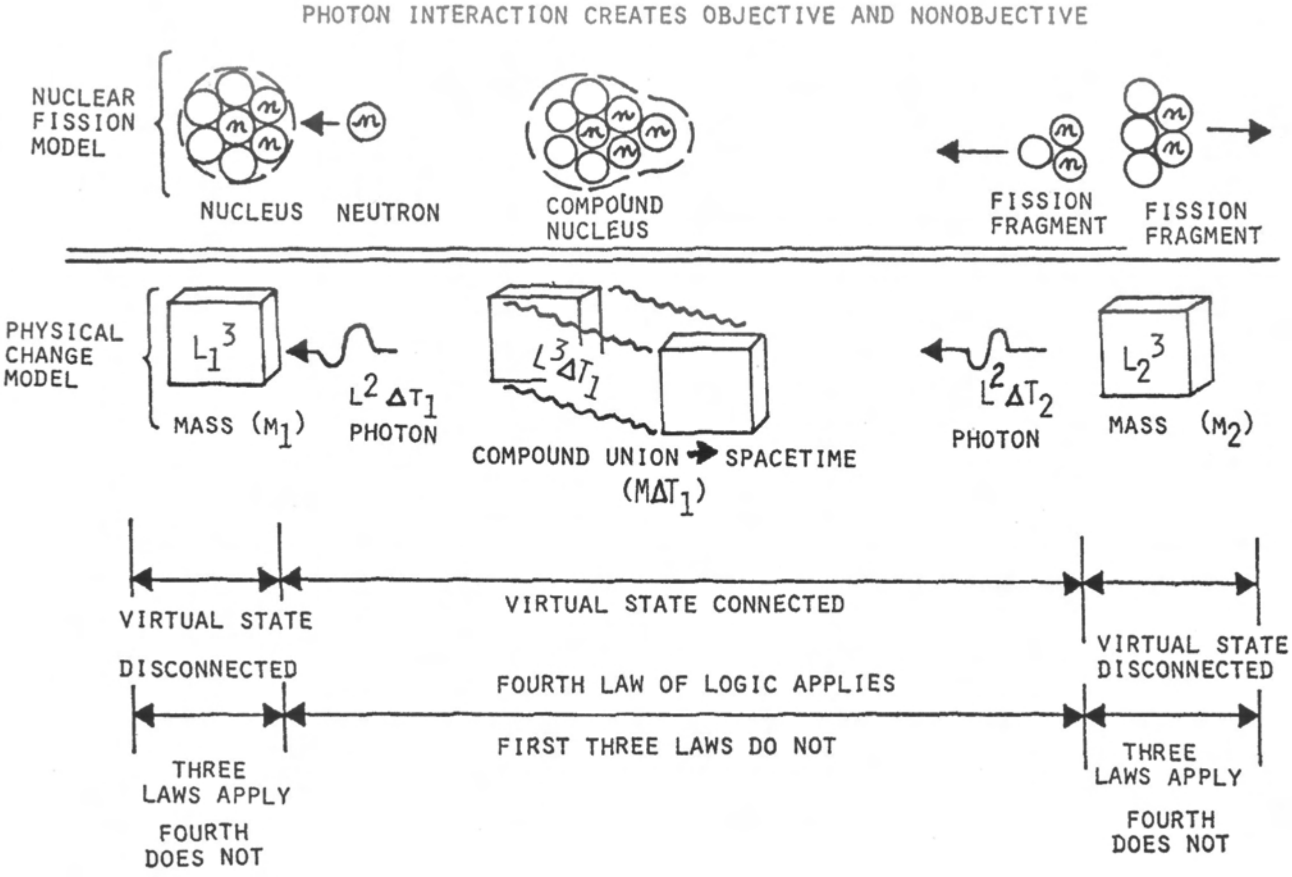
PHOTON INTERACTION CREATES OBJECTIVE AND NONOBJECTIVE
NUCLEAR FISSION MODEL
n
NUCLEUS
NEUTRON
COMPOUND NUCLEUS
FISSION FRAGMENT
FISSION FRAGMENT
PHYSICAL CHANGE MODEL
L_1^3
MASS (M_1)
$L^2 \Delta T_1$
PHOTON
$L^3 \Delta T_1$
COMPOUND UNION → SPACETIME
$(M\Delta T_1)$
$L^2 \Delta T_2$
PHOTON
L_2^3
MASS (M_2)
VIRTUAL STATE DISCONNECTED
VIRTUAL STATE CONNECTED
VIRTUAL STATE DISCONNECTED
FOURTH LAW OF LOGIC APPLIES
FIRST THREE LAWS DO NOT
THREE LAWS APPLY
FOURTH DOES NOT
THREE LAWS APPLY
FOURTH DOES NOT

Fig. 9. Raindrop model of quantized change

Hertz, which can be stimulated with ultrasonics or electromagnetic radiation. The ultraviolet region contains a window, as does the infrared region. Many other windows apparently exist.

Trevor James Constable photographs paranormal phenomena through a filter that is open in the IR and UV regions, but opaque to visible light. Thus the filter blocks out the strong quenching effect of visible light and looks in two magic windows, IR and UV. Constable is thus able to photograph phenomena that are not normally detectable.

RAINDROP MODEL OF QUANTIZED CHANGE

What does not appear to have been realized to date in physics is that the photon interaction is the agent that creates objectivity itself, as is clearly established by the two-slit experiment. The photon interaction separates spacetime into space and time exclusively. An object, being something which occupies space (L^3), is thus timeless. Objects do not exist in time, because the union of an object with time constitutes spacetime, which cannot be perceived, detected, or observed.

Only changes in (derivatives of) spacetime can be perceived, detected, or observed, but not spacetime itself. That with which light has not interacted is not objective. This is proven conclusively by the two-slit experiment, upon which all of quantum mechanics rests. In that experiment, if an electron emitted from the cathode and on its way to the two-slit region does not interact with light, the electron behaves in a totally nonobjective manner, appearing to pass through both slits at the same time. (Actually it goes around both spatial slits in time, and since both of the slits occupy the same piece of time, the electron can interact timewise with both of them.)

On the other hand, if the electron is struck by a photon before it reaches the two-slit region, it becomes a "hard, solid little baseball" and quite objective. In that case it can interact with (go through) only one spatial hole because it cannot interact in time. According to Feynman, this simple experiment contains the total and only mystery of quantum mechanics. Furthermore, it cannot be explained in any classical manner. No physicist in the world understands it, although the pattern of repeated electron strikes on a collecting screen is quite easy to compute by simple trigonometry.

Let me now explain what happens to a mass when it is struck by a photon, absorbs the photon, and re-emits the photon. First, a mass is an L^3 object; the concept of mass is not a function of time, but only a function of space. It is thus three-dimensional. The photon is also three-dimensional, but one of its dimensions is the time dimension. The photon, being built of action, is constituted of $(dEdT)$ or (L^2dT) dimensionally.

When the photon strikes the mass and is absorbed, the (AE) portion is turned into mass by orthorotating one turn and becoming L^3 dimensionally. The (dT) portion is thus united with the spatial mass to give a compound nucleus of spacetime comprised of MdT_1 This compound nucleus of spacetime is not perceivable. When a photon is re-emitted, it may or may not be of the same frequency and energy as the previously absorbed photon, depending on the absence or presence of any other perturbations. At any rate, a small bit of L^3 spatial mass is orthorotated to give an L^2 intersection with the laboratory frame, which turns the bit of mass into a small piece of energy, dE. In the orthorotation, a small piece of time is bitten off and joined to the dE to give a dEdT = h quantum of action, which now constitutes a photon. Photon emission thus strips away the time dimension, leaving a spatial object.

Einstein pointed out that the velocity of an object may be visualized as rotation of the object in higher dimensional space. Visualized as spatial rotation rather than rotation toward the time axis, the velocity of light in vacuum c merely constitutes the rotation of a piece of mass by an angle of 90° to the line of motion, in the laboratory space in which we visualize the photon (the rotated piece of mass) as moving. Thus the photon emitted from the mass is moving at c, the velocity of light, because of its orthogonal rotation to the laboratory frame.

Note that the spacetime compound nucleus has now once again been separated into spatial and time components. Time is moving with the photon. And that is why time moves or flows at c, the speed of light in a vacuum. Time is carried only by the photon, and photon interaction with an object produces that object's march through time. If the emitted photon and the previously absorbed photons do not have identical dE and dT components, then M_2 will be different from M_1, and the mass will have increased or decreased. Considering the photon to consist of momentum and length components, i.e., *(dpdL),* a similar diagram can be constructed. In that case, if the momentum and length components of the emitted and absorbed photons are identical, then M_1 and M_2 occupy the same position (M_1 has not moved). If these components are different, then the mass M_2 will occupy a slightly different position than mass M_1, and the mass will have moved.

This, by the way, is the simple solution to the age-old philosophical problem of change. We simply identify M_1 and M_2, calling them the same object. We cannot detect, perceive, or observe the compound nucleus of spacetime in the middle. Thus we experience change as a thing becoming something else, but still being the same thing. This totally defeats the first three laws of logic. However, fortunately we can easily comprehend it since we have developed the fourth law of logic.

The raindrop model of quantized change is taken exactly analogous to the raindrop model of nuclear fission, as shown in Figure 9.

KINDLING EFFECT

G.V. Goddard and his associates in 1969 reported a peculiar kindling effect generated by repeated, periodic, low-intensity stimulation of the limbic region of mammaliam brains. A sustained periodic signal input to the brain and central nervous system eventually sets up a cumulative resonance which increases in magnitude until the entire organism is in sympathetic resonance.

A laboratory rat at first continues to explore its environment in a normal manner when it is subjected to kindling. But after repeated stimulation at the same intensity, the rat will begin to rear up and its forelimbs will convulse. Eventually these bursts of electrical activity induce similar patterns in nearby brain regions, and the threshold becomes progressively lowered. Stimulation progresses to the amygdala, to the amygdala on the other side of the brain, to the hippocampus, to the occipital cortex, and finally to the frontal cortex. Kindling can start only in the limbic structures.

While kindling was originally thought to be a model of epilepsy, John Gaito of York University has reported that a different mechanism is apparently involved since the amino acid, taurine, which suppresses epileptic seizures in laboratory animals, does not prevent phenomena caused by kindling. Also, kindling apparently causes permanent changes in the neural circuitry. Pulsed repetitions of telepathic senders have also been shown to increase the reception of telepathic messages. Thus the kindling effect apparently applies to the paranormal channel as well as to more orthodox transmission channels. (For further details, see "Kindling, once epilepsy model, may relate to kundalini," *Brain/Mind Bulletin,* Vol. 2, No. 7, February 21, 1977; pp. 1-2.)

The kindling mechanism is a far more general mechanism than epilepsy researchers have realized. Coherent time collection of bioenergy in one bioframe "kindles" toward the threshold of the next bioframe, which has a fixed threshold. When sufficient kindling occurs to reach the threshold, automatic orthogonal rotation of the kindled bioenergy occurs into the next frame. There it simply constitutes the kindling or superposition of the imperceivable subquantum state into the perceivable quantum state. This is the mechanism whereby one kind of field can be turned into another. E.g., thought energy (third biofield) can be kindled into second biofield (flux), which can be kindled into first biofield (electromagnetic energy), which can be kindled into zeroth biofield (matter).

BIOENERGY COLLECTORS

Good collector mechanisms exist in great variety, in both living and inert systems. Many good collector mechanisms exist in the human body. One mechanism that collects bioenergy and orthorotates it into ordinary electrical field with appropriate induced currents is simply the electrical current in the human nervous system. This is particularly true of the current discharges from and between the nerve endings.

Another collector for bioenergy in a living system is the mitogenic radiation between the cells in tissue. Taken as a flux, this mitogenic radiation would seem to have both a photon nature and a nonphoton nature. The cells, in their membranes and fibrous internal structures, are continually collecting and kindling bioenergy into electron currents and photons. The electron currents represent collections from the virtual state into observable state, and may in one sense be regarded as the lifting of electrons out of the Dirac sea by direct kindling (coherent superposition of virtual state energy into observable state energy). These kindled currents are thus superpositions or modulations on the ordinary electrical activity of the cells, fibers, and membranes. The mitogenic photons are also produced by direct kindling of the virtual state into observable state, in this case, by kindling virtual state photons into observable state photons.

Spark discharges, such as are used in Kirlian photography, are extremely good collectors of bioenergy. In Kirlian photography the film records ordinary photon interactions; however, the spark discharge from the finger is collecting and kindling bioenergy into ordinary electrical energy (specifically, into electromagnetic field, which then affects the electrical flux pattern). Also, bioenergy is being collected and kindled into ordinary photons, thus changing the photon pattern being emitted.

Furthermore, those photons being emitted are internally affected by the bioenergy that has not yet kindled to observable state electromagnetic energy, but is still virtual. Each photon carries the virtual state along in the little piece of time contained in the photon, and partially collected bioenergy which has not yet breached the threshold of observable photons is being carried in each emitted photon. These patterns are transferred to the grain atoms by photon absorption when the piece of time of the photon is momentarily added to the mass of the absorbing atom, and in turn are continually spread through the emulsion material atoms by the virtual photon fluxes of the charges, in both the nucleus and the orbital electrons of the atoms.

So the life pattern of a photographed object is captured on ordinary film — and indeed that pattern can be resonated through hyperspace with the object that was photographed, regardless of where the object is in the universe.

Note that there are many mechanisms inside stars and planets for the collection of bioenergy. We live in a vital, pulsing, living, breathing, bioenergetic cosmos, and life is everywhere. Patterns of life are riding in the light from the distant stars, from our sun, in the neutrino flux from the cosmos, in the graviton flux from the cosmos, in cosmic rays, etc. And so it is small wonder that living cells that form, reproduce themselves, and die can be induced in the proper primordial soup of nutrients when that soup of matrix forms and chemicals is bathed in ultraviolet light (one of the magic windows which contains a greatly enhanced transmission factor for bioenergy, hyperspatial energy, life energy), and stimulated by electrical spark discharges.

It would be interesting to perform the same type of experiments with concentrated light from distant stars, or with the mitogenic radiation from cell cultures, using environmentally shielded setups similar to those used by the Soviets in experiments with the transmission of death patterns.

Even stranger, bioenergy propagates along deBroglie waves, and so all possibilities are modulated by bioenergetic kindling and collection. Viewed in this light, Puthoff and Targ's remote viewing experiments actually provided a multitude of probability paths to the viewer, and these probability paths were indeed kindling bioenergy of the distant, as yet unhappened event. The actual event in the future contains a great many additional collection kindlings due to the additional factors involved in determination (such as collapse of the wave function). A peaking of signal exists on the more probable happenings, and this extra signal along certain paths constitutes an increased signal-to-noise ratio in that particular channel. Thus an actual mechanism exists for seeing at a distance and for seeing ahead in time.

Particular note should be made of the pi meson collection mechanism in the nucleus of an atom. Each neutron and each proton in the nucleus of an atom is continually emitting and absorbing pi mesons, and thereby the protons and neutrons are continually interchanging from one to the other by passing along the positive charge. This activity produces the strong force which binds the nucleus together against the electrical repulsion of the electric forces due to the positive charges of the protons. The fantastic structurings and currents that result provide an electrical current which can and does collect bioenergy. A skilled psychic with good psychokinetic ability can thus focus on the kindling mechanism inside the nucleus and get a signal-to-noise increase of the kindling of his bioenergy in nuclei as opposed to elsewhere in the material in other collection mechanisms. Several types of reactions can then occur. The protons may have their electric charges extinguished due to the extinguishing of the virtual photons producing the electric field on their pions. In such case, the atom simply loses its integrity, and some neutrons and photons are emitted from the atomic site into the surrounding material. This type of interaction thus locally softens a metal and produces severe heating in localized areas, changes in fundamental grain structure, and changes in fundamental characteristics of the metal.

It is by this mechanism and similar related reactions that metal bending can be and is accomplished by persons such as Uri Geller. The collection of bioenergy by graviton emission, absorption, and flux leads to antigravity and other psychokinetic effects. These are exhibited in rare levitation phenomena and more frequent poltergeist phenomena.

The list of collectors could be materially increased, but these should suffice to show the importance of the effect of coherent time collection, kindling, and orthorotation of one type of field into another. The collection of deBroglie waves is moving much faster than light, and for very slowmoving objects the collected and projected bioenergy on their deBroglie waves is moving at essentially infinite velocity. This accounts for the fact that Hieronymus observed an instantaneous cutoff of his monitoring signals from the Apollo astronauts when their space vehicle went behind the edge of the moon.

Looking over the list of collectors and carriers, it is obvious why even the best Farraday shields cannot shield out bioenergy communication. The deBroglie wave carrier alone will simply take the bioenergy right through any known substance — or more precisely, it will take the bioenergy around the three-space in which the barrier substance exists, since deBroglie waves are actually superluminal particles in a three-space more than one orthogonal turn away from the observer's three-space. The implications for psychic healing and psychic surgery are obvious.

PSYCHOMETRIC EFFECT

Every mass communicates with every other mass in the total universe, with hyperspatial/virtual crosstalk between all parts and all masses. This yields a mechanism for psychometry.

Inside the mass, we have a large number of collector mechanisms, of the types previously listed and explained. All objects in the universe are in communication with this object, particularly large or close objects. Those objects that come very near or touch the chosen object for a period of time are most influential.

Briefly, the internal dynamic collector mechanisms of the chosen object perform collection and kindling of the virtual state; thus each mechanism is slowly changed by the influence of other objects in the universe. This is very like the diffusion of heat. The chosen object also communicates part of the influx away, normally reaching an equilibrium with its surroundings when these surroundings are stable. As the surroundings change, appropriate changes are communicated into the internal collector mechanisms of the chosen object, and appropriately diffused away to other parts of the universe, until equilibrium is again reached.

Note that the influence of nearby or touching objects is greater than the influence of distant objects under normal spatial considerations. In hyperspatial considerations, however, this is not necessarily true at all, since this chosen object's pattern may be tuned in through hyperspace without any loss of energy due to ordinary three-space separation distance. In the absence of such hyperspatial tuning, distance is a factor, and close objects have a greater signal-to-noise ratio in the collection and kindling than do more distant objects.

Thus an individual or group of individuals in the vicinity exert mental, emotional, and bioenergetic influence on the kindling being accomplished in the chosen object. The stabilized dynamic patterns of the localized chosen object thus are conditioned and kindled to contain patterns of the actions, feelings, and thoughts of the individuals around it, as well as the natural events that occur. It is thus a very exact record of its surrounding events.

A psychically sensitive person who touches the chosen object, or in some cases merely comes near it, enters into a communications exchange with it. His body-records diffuse into the local object, and those of the local object diffuse into his body. If the psychic is sensitive enough, he can gain impressions from the recorded events of the object. This is the scientific basis of psychometry.

A wide variety of data support this mechanism. Psychometry has been practiced since earliest times, and is still practiced today. A photograph is particularly indicative of the person photographed, since his entire record is imprinted psychometrically on the photograph.

Moreover, the entire record of the Earth is still in the earth, in all its detail. As indeed is the record of each star and its planets, riding on the light that reaches us from that star, and on the deBroglie waves from it. With operational biodetectors, we could easily read the records of the universe. NASA would do well to pursue this alternative in its attempt to contact extraterrestrial life, rather than point huge radiotelescopes toward the heavens and listen to the kindled static from the stars. The static indeed contains the message, but not in a format in which they are prepared to look! It is in the virtual state, riding in each photon.

To a psychic who can see the human aura easily, the condition and much of the record of the person is in that aura, because the auric pattern is changed by all experience of the body, mind, and personality. Thus a good healer can see the aura of the patient, tell if another healer has been in contact with the patient (including often recognizing which other healer it was), and can recognize his own previous contact with the patient many years earlier. (See, for example, the story of Agpaoa as recorded by Sigrun Seutemann in *Healers and the Healing Process,* ed. George W. Meek, Theosophical Publishing House, 1977, p. 93.)

ALL MASS IS CHARGED: EVERYTHING IS ELECTRICAL

It is well known that all mass is charged; thus everything in one sense is electricity. To examine this concept, let us look again at the electron with its virtual photons being emitted and absorbed, which virtual activity creates the electric field of the electron. In classical theory the electron is visualized as if it were a little hard

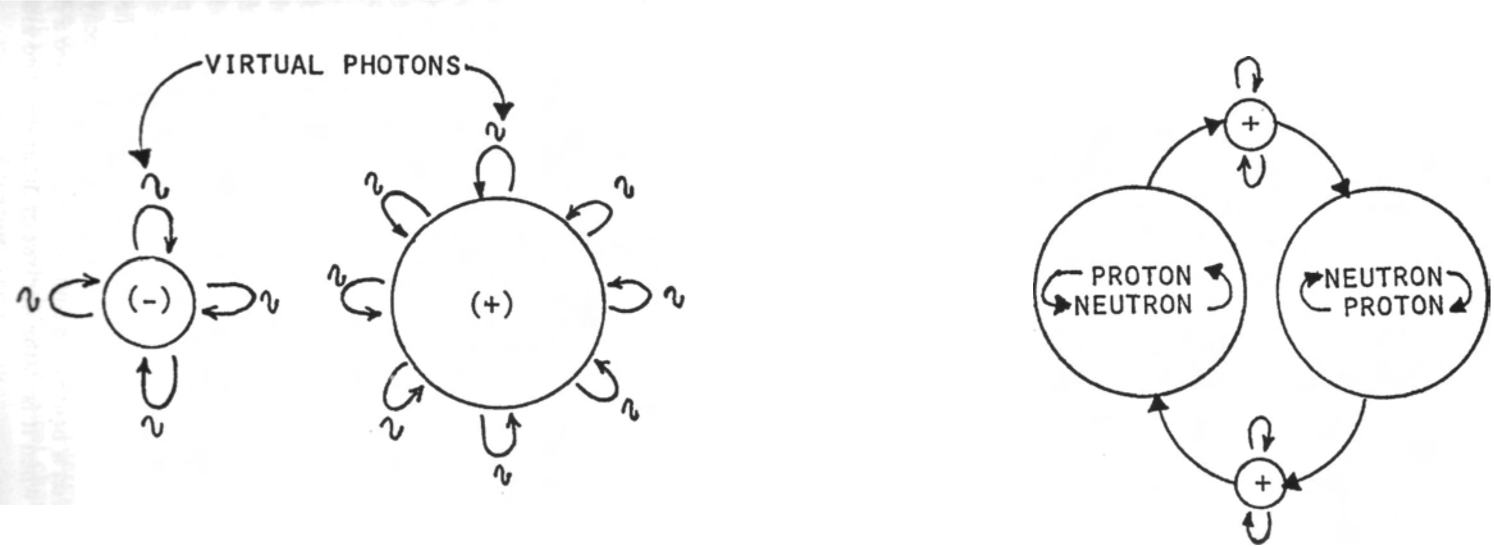
VIRTUAL PHOTONS
(-)
(+)
PROTON
NEUTRON
NEUTRON
PROTON
+
+

ELECTRON

PROTON

IN A NUCLEUS, EVEN A NEUTRON IS CONTINUALLY CHANGING INTO A PROTON, AND VICE VERSA, BY THE EXCHANGE OF THE POSITIVE CHARGE.

Fig. 10. All mass is charged: everything is electrical

spinning sphere; the spinning of its electric field then generates a magnetic component. This concept of spin is retained, even though the electron is no longer regarded simply as a little baseball.

From our viewpoint, the electric field constitutes a steady activity of the emission and absorption of virtual photons, and the electron spin is more of an implosion or explosion in higher dimensional space. That is, we visualize a fundamental stable particle as a precise closure of the entire physical universe, like the Einstein spherical model of the cosmos. Furthermore, we visualize multiple closure; first, a very gradual macroscopic closure gives the so-called gravitational closure of the Einstein spherical model, and a closure — in the case of an electron — of about 10^{42} times as quick and small gives the charge closure. This second closure is considered to be at right angles to the first. In one direction the second closure gives a negative charge, and in the opposite direction the second closure gives a positive charge.

So in this view, spin is occurring in at least four spatial dimensions, not just three. It is more of an implosion or explosion into and out of this three-space universe frame. From such a model, we could expect that some startling physical phenomena might be possible. For example, we might visualize the implosive or explosive spin as similar to a vector flux. If these vectors are not precisely antiparallel when two like charged particles meet, then a resultant flux vector exists. Since spin is actually the quantity action (energy x time, or momentum x length, or mass x time), then the existence of a vector resultant means that enough action can exist to provide a quantum to be fissioned, which yields (creates) physical change. Thus such a resultant allows a collision or interaction to occur. On the other hand, if two like charged particles meet and their spins are precisely antiparallel, then no vector resultant flux exists. Hence no effective action exists to be fissioned and create physical change (collision or interaction). Thus two like charged particles with precisely antiparallel spin ought to run right through each other when they meet, with no interaction at all, electrical or otherwise, in complete contradiction to conventional collision theory and conventional electrical theory which states that like charges repel one another.

Such an experiment has been done, at Argonne National Laboratory, with colliding protons. When the colliding protons have antiparallel spins, they do indeed pass right through one another. And there dies forever the old idea of concrete, solid, material reality. The experiment is not predictable or permitted by present electromagnetic theory or conceptual models of the proton and proton interaction. The experiment is consistent with our dual closure holographic model of the universe, however, and is in fact predicted by it.

In a nucleus, we find both protons and neutrons (except for a single isotope of hydrogen, which contains only a single proton in its nucleus). However, these nucleons do not exactly maintain their exclusivity. Instead, a proton is continually changing into a neutron and vice versa; this metamorphosis is accomplished (in theory) by the continual exchange of a small particle of mass containing the positive charge. Thus a highly dynamic type of electrical current activity is going on in the nucleus. And in fact, this is the only way a neutron can be stabilized in ordinary matter; when freed from the nucleus, the neutron is unstable and decays in about eleven seconds or so. So even a neutron is continually changing into a positive particle, i.e., into electricity, when it is in the nucleus of an atom. And around the nucleus, the electrons are in dynamic orbit, so highly dynamic electrical activity is going on in the outer part of an atom and also in the inner part. Thus indeed all mass is charged and everything macroscopic is electrical. And every physical change of a charged particle constitutes a modulation imposed upon its virtual photonic activity creating its electromagnetic field. These modulations are simply coherently collected to breach the threshold of the virtual (third) orthoframe, passing into the second orthoframe, then breaching the threshold of that frame, and passing into the first orthoframe, which is the frame of the electromagnetic field. The electric field is coupled by the dual closure mechanism to the gravitational field, and also to the mass. The entire mechanism constitutes the coupling of the virtual state reality to observable state reality, and the electromagnetic field plays a prominent part in this coupling. Psychotronics is nothing but the extrapolation of this mechanism and the formulation of a theory to allow the construction of practical devices based upon it. Put simply, psychotronics is *virtual state engineering.*

A FUNNY THING ABOUT ELECTROMAGNETIC FIELD

According to the present theory of physics, the electric field is due to the continual emission and reabsorption of virtual photons. Let us examine that concept just a little further, because there are some interesting conclusions we can make if we are not afraid to take the final conceptual step to the bottom line implications.

First, each ordinary (observable) photon in the universe consists of a little piece (quantum) of nonobservable action. The units of action are energy times time, or ML^2/T *{mass* x length x length/time). The quantum of action in the photon is not separated into a separate piece of time and a separate piece of energy. And also, the quantum of action is a fixed, constant size or magnitude, given by Planck's constant h as

$$h = 6.626 \times 10^{-34} \text{ Joule-second}$$

All this says is that an ordinary photon has a certain size that is quite fixed, and it is made of the substance "action" which is totally nonperceivable and nonobservable. Only changes in action are perceivable, observable, or detectable, never fundamental action itself. So physics already prescribes a universe made out of action, and that universe has no length, no time, no energy, no space, no concrete existence. Yet from it can be taken — by differentiation (which is simply separation) — any length, any time, any energy.

However, only a single quantum at a time changes or is separated. Quanta do not superpose. A single quantum can be split or fissioned into, say, a piece of energy and a piece of time. We can get as large or as small a piece of energy (i.e., change in energy) as we wish simply by adjusting the piece of time accordingly. The same type of separation is possible for momentum and length.

The fragments of the action quantum (say, a piece of energy and a piece of time) must then be stuck back together again (i.e., fused back into an action quantum) to get rid of them, and this particular part of the process has essentially been ignored by physicists. Anyhow, getting rid of the fragments (which gave us a piece of comparative physical change — energy versus time or momentum versus length) causes the just-completed physical change to vanish.

So at its most fundamental level, the entire universe is continually being created and then destroyed in sequence. The continual creation of a tiny piece of time and then the destruction of the piece of time makes us experience reality or change as if we were a pointer moving along a time line. We cannot see time because it is continually being destroyed by the change process, the very process of physical detection or observation itself.

Now if we invoke another kind of change, a virtual change, such that its size is always smaller than the actual quantum's size, then this kind of change cannot be detected, even theoretically. The concept of a virtual change is one such that the quantity of action that was changed was deliberately made smaller than the size required to throw the switch of detectable or observable change.

So if we have a virtual photon, it is simply a piece of action whose magnitude is always less than *h*. Here again, we can still get any amount of energy change we desire, merely by adjusting the size of the piece of time associated with that energy. This then means that the virtual photon is not big enough to breach the threshold of quantum change, and so cannot be observed or detected.

And here is the interesting bottom line ignored by most physicists and almost all the pragmatic scientists who run the orthodox establishments. The totally unobservable, undetectable virtual photons being emitted and absorbed are conceived as creating the electric field at the same time as observable physical change (ordinary quantum change) occurs. And the electric field, of course, can cause or induce change of sufficient magnitude to breach the quantum threshold and result in observable change. The bottom line is this: the virtual state can be collated and collected to sufficient magnitude to breach the quantum threshold and induce ordinary observable physical change. The electromagnetic field already does precisely this! And if reality can behave this way, as the physicists show, then let us simply accept that and apply it!

Note that we have now resolved the problem of negative time. The fusion process replaces the little positive piece of time that was formed in the fission process. A continual stream of such replacements constitutes a continual stream of subtractions of positive time. Hence this represents a negative time operational channel. Many of the equations of physics, e.g., Maxwell's equations of the electromagnetic field, should be reexamined with a view to this negative time operation. Dirac calculated such an entity, an electron traveling backwards in time, and noted that it would appear as a positively charged electron traveling forward in time. He thus predicted the positron which was later found in laboratory experiments. But of even more fundamental importance is that every piece of charged matter in the universe has such an opposite twin — the well known antimatter concept. So antimatter protons, neutrons, molecules, and physical structures exist. Indeed, this physical reality that we normally perceive has its antimatter twin which we are also unraveling backwards as we go. But here again the bottom line is far reaching: it means that we can coherently collect the virtual state to produce matter or its antimatter, which after all are nothing but macroscopic assemblages of quantum changes, and we can thus get free energy, materialization and dematerialization, teleportation, and antigravity by means of special modifications and developments of electromagnetic circuits. Each photon carries a small piece of time. And the entire virtual state reality is hanging onto that small piece of time as a modulation. That is where the probabilities are riding, in terms of the ordinary monocular interpretation of quantum mechanics, all of these probabilities, and all possibilities, are concretely real and occupy orthogonal universes. Since they all share the same single time dimension, there is a time bridge or time channel connecting the universes. Each move or change in one of them crosstalks a tiny bit into each of the others. Since it is the photon which carries them, then each photon contains all possibilities and virtual states modulated upon itself. Photon interaction thus involves the communication of these possibilities and virtual states to the object with which the interaction occurs.

EXTINGUISHING ELECTRICAL CURRENTS

With psychotronics, it is quite possible to extinguish electrical currents flowing in a circuit, or in the nervous system of a biological organism. First, an electromagnetic carrier (photon) is used to carry the psychotronic modulation pattern which will quench electrical currents. The particular psychotronic pattern is modulated upon each individual photon, in virtual state form riding upon the time component of the photon. This is totally different from the ordinary scheme of modulating an electromagnetic carrier; the normal scheme merely gates the rate of passage of photons (amplitude modulation), or changes the energy-time ratio on various photons as photons continually are emitted (frequency modulation).

In this case, virtual state patterns (probabilities) are added directly to each photon, in addition to whatever is done under the normal scheme of modulation. Thus the carrier can be electrically very complex or very simple. At any rate, the carrier with its psychotronic modulations reaches the target, and photons are absorbed continually and emitted continually by the target. The absorption of the photons and their modulations continually generate compound spacetime MdT nuclei and introduce the virtual state psychotronic modulations onto the mass's electric field, the modulations now riding on the virtual photons that generate that field. When a photon is re-emitted, the compound nucleus is split again, and the dT is carried away by the photon, leaving behind a changed virtual photon state on the non-temporal mass spatially left in between quantum changes.

As time passes, the virtual photons are more and more psychotronically modulated by the kindling effect. Eventually the collection breaches orthoframe thresholds and reaches the electromagnetic field frame, changing the electromagnetic field. If the psychotronic pattern is such as to create a positive charge when sufficiently kindled, the negative electrical charge of the kindled electron is simply absorbed or quenched from within, so to speak. And in this case no ordinary annihilation photons are emitted. Instead, the electron is converted to a neutrino when its electrical field is quenched, and it simply passes rapidly away through matter with no interaction and no damage or heating to that matter. Thus the currents flowing in an electrical circuit or in a biological nervous system can be quenched, without any other damage to the system, and without any pair annihilation radiation being produced.

The UFO phenomena very often demonstrate this effect, and the Soviet Union has long since weaponized it.

A simple way to visualize this process is to imagine the virtual photons of the electron as gradually being quenched. With the virtual photons gone, the electron has lost its charge and its electric field, and is now a neutrino. As a neutrino, it zips through matter with no interaction, and on out into deep space.

This mechanism explains why the electrical circuits of autos, radios, etc. usually are not damaged when they are disrupted by the UFO phenomena. Use of normal mechanisms to change or eliminate the electromagnetic field seriously disrupts the components of the circuits, heating materials, blowing fuses, and melting wires and components. The mechanism described here does none of those things.

The weapon aspects are of course formidable. Small transmitters with the right modulation pattern could be dropped, planted, or fired almost anywhere, and these could be set to disrupt equipment or human nervous systems. Their use in satellites is obvious. Small transmitters of a fixed time-life broadcast duration could be fired behind enemy lines in artillery shells. Tactical missiles could carry hordes of such small transmitters and drop them along the way deep into enemy territory. Essentially everything electrical is vulnerable if radiated long enough for sufficient kindling to occur. Powerful radar beams can be modulated psychotronically to give them effects at a distance, disabling vehicles, communications equipment, and men. Even searchlights can be so modulated. Pavlita's psychotronic generators are known to generate deathdealing effects, because he has demonstrated that they can kill insects, and he has admitted that men can be killed with generators that are large enough.

An even more radical approach is to transmit the desired pattern directly through hyperspace with a hyperspace transmitter, using either a set of triads for focusing at a desired spot in normal space, or using a witness to generate form resonance with the desired target object, much like one tuning fork kindles another of the same tuning into vibration. The witness may be anything that contains the pattern of the target, e.g., a photo or a model. In the case of a living organism, each cell contains the blueprint for the entire organism, so a group of cells from the organism can be used. This applies to plants as well as to animals.

DEATH TRANSMISSION VIA THE PARANORMAL CHANNEL

As an example of the kindling effect and the variability of photon quenching of the paranormal channel, some extremely interesting experiments were performed by V.P. Kaznacheyev et al regarding the paranormal transmission of death. Briefly, two groups of cells were selected from the same cell culture and one sample placed on each side of a window joining two environmentally shielded rooms. The cell cultures were in quartz containers. One cell culture was used as the initiation sample and was subjected to a deadly mechanism — virus, germ, chemical poison, irradiation, ultraviolet rays, etc. The second cell culture was observed, to ascertain any transmitted effects from the culture sample being killed.

When the window was made of ordinary glass, the second sample remained alive and healthy. When the window was made of quartz, the second sample

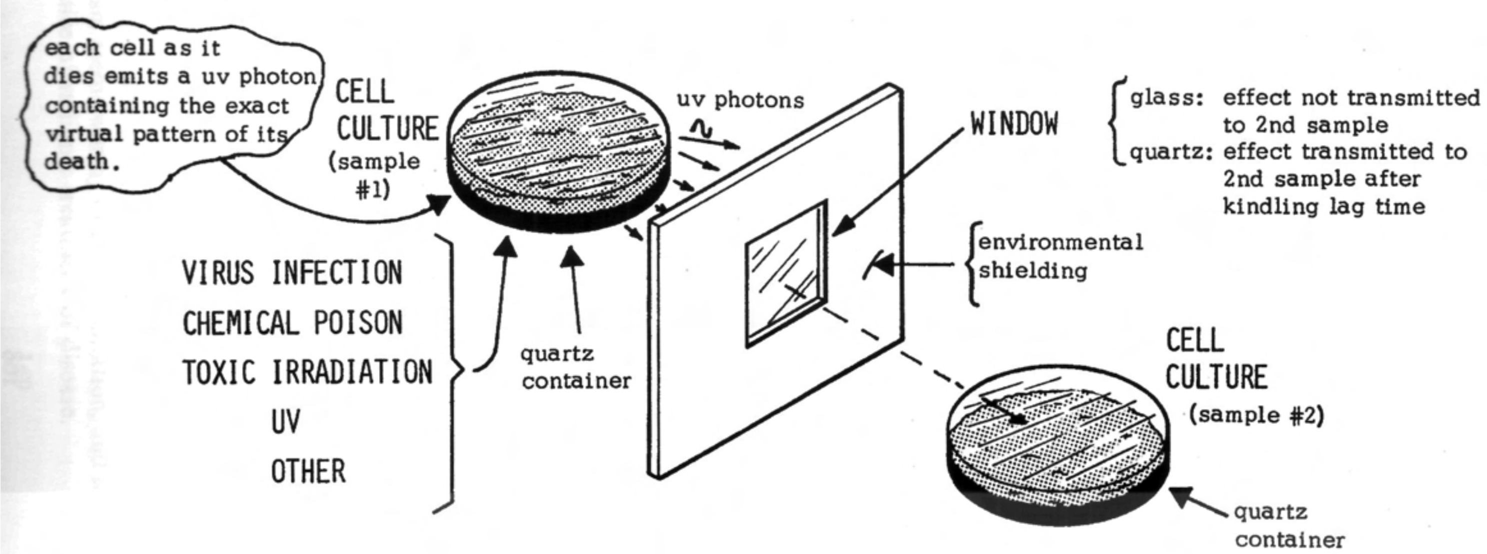
each cell as it dies emits a uv photon containing the exact virtual pattern of its death.
CELL CULTURE (sample #1)
uv photons
WINDOW
glass: effect not transmitted to 2nd sample
quartz: effect transmitted to 2nd sample after kindling lag time
environmental shielding
VIRUS INFECTION
CHEMICAL POISON
TOXIC IRRADIATION
UV
OTHER
quartz container
CELL CULTURE (sample #2)
quartz container

>5,000 EXPERIMENTS

INSTITUTE OF CLINICAL AND EXPERIMENTAL MEDICINE, NOVOSIBIRSK

INSTITUTE OF AUTOMATION AND ELECTROMETRY, NOVOSIBIRSK

Fig. 11. Death transmission via the paranormal channel

V.P. Kaznacheyev et al, "Distant Intercellular Interactions in a System of Two Tissue Cultures," *Psychoenergetic Systems*, Vol. 1, No. 3, March 1976, pp. 141-142.

sickened and died with the same symptoms as the primary sample. The experiments were done in darkness, and over 5,000 were reported by Kaznacheyev and his colleagues. The onset of induced complementary sickness and death in the second culture followed a reasonable time — say two to four hours — behind sickness and death in the primary culture.

The major transmission difference between window glass and quartz is that quartz transmits both ultraviolet and infrared well, while glass is relatively opaque to ultraviolet and infrared. Both quartz and glass transmit visible light. Thus glass is a suppressor of the paranormal channel, while quartz is not.

By performing the experiment in darkness through a quartz window, the four-law patterns of disease, sickness, or death engendered in the primary sample are not extensively squelched by the photon interaction, and these four-law patterns are transmitted through the quartz window into the second cell culture. Since the genetic patterns of the second culture are the same as those of the first, sympathetic four-state interaction and eventual kindling occur, resulting in the appearance of the effects in the second culture.

It is well known that cells also emit mitogenic radiation, including radiation in the ultraviolet and infrared regions. Since these same regions are not strongly suppressed by the photon interaction, one may hypothesize that paranormal effects may be strongly modulated onto infrared and ultraviolet photon activity, and there is indeed evidence for such an assumption. In 1950, Western researchers found that cells could be killed in darkness with ultraviolet radiation, kept shielded from visible light for twenty-four hours or longer, and then if radiated with visible light the cells would start reviving by hundreds of thousands even though they had been clinically dead.

Specifically, every cell emits mitogenetic radiation in the ultraviolet twice: when it is born and when it dies. The UV photon emitted at death contains the exact virtual state pattern of the condition of the cell at death. The healthy cells are bombarded with death messages from those that are dying, and this diffuses the death pattern throughout the healthy culture, eventually kindling into the same death pattern there.

However, the squelching of the paranormal channel by photon interaction is never complete, or paranormal phenomena could never occur in daylight. Hence paranormal patterns can be modulated even onto visible light, although only extremely weakly. Thus we may speak of the *Q* or sharpness of the paranormal modulation upon photon frequencies; *Q* is extremely low for visible light and may be extremely high for infrared and ultraviolet. In the visible light spectrum, it is probable that extremely large numbers of near-zero strength paranormal patterns are modulated on the light radiation; hence these patterns simply consist of a very weak background noise and the kindling effect does not apply. (The patterns are so random as to be self-cancelling in the kindling effect.)

DISEASE AND DEATH INDUCTION

The base idea here is that an active agent must be involved to disrupt the ordinary healthy pattern of inceptive life functioning. Since life functioning is inceptive, ultimately the bioframes, specifically the first and second, must be distorted or disrupted. The first bioframe is the electromagnetic field and the second bioframe is a psychotronic field. The second bioframe may eventually be revealed as the *neutrino* field, but this is still uncertain at present. At any rate, it is a *virtual state* field. Virtual state patterns of a disease, for example, are carried in each photon emitted by a diseased cell. The life functioning psychotronic components are modulated on the body's electromagnetic components. Thus they have an electromagnetic carrier, which will carry the inceptive life function patterns through the photon squelching effects of emitted photons in the body.

A host agent such as a germ produces an active agent such as a toxin. The chemistry involves the particular compound of the toxin that is disruptive or deadly. But chemistry after all is rigorously based on electrical patterns. These electrical patterns of toxic chemistry contain psychotronic patterns modulated upon them by the germ, and these genetic life functioning patterns are different from those of a human cell. The foreign psychotronic patterns are modulated upon the human body's electromagnetic patterns (currents) and thus reach the cell, where they interact by the fourth law in the cell membrane, being demodulated and introduced to the cell, and orthorotated in the process into electromagnetic signals disruptive to the orderly functioning of the cell. The foreign psychotronic signals also jam the ordinary psychotronic patterns being demodulated by the cell in its inceptive operations from its mind frame or third bioframe. If this jamming is only moderate, the cell sickens; it is only disrupted. If the jamming is severe, the cell dies.

The normal electromagnetic noise bathing the cell from the cosmos and the environment contains random components that largely self-cancel or balance. The modulated random psychotronic modulations of the ordinary background environment behave similarly, and thus the normal background simply constitutes a weak noise flux which does not disrupt the inceptive life functioning. In the experiments of Kaznacheyev and his colleagues, photon squelching was largely restricted, and the psychotronic signals from the dying primary culture could go through the infrared and ultraviolet windows in the quartz as virtual state patterns on the mitogenic UV photons emitted by dying cells, and be introduced directly into the secondary culture. The kindling effect then induced the disease or death pattern after an appropriate time delay. Thus it appears that there is indeed a fundamental mechanism by means of which a disease or death pattern from one cell is introduced into others.

If this is true, then an exciting vista emerges for the treatment of hither-to incurable diseases such as cancer. The psychotronic pattern of the cancer cells in a person's body could simply be recorded, then phase shifted 180 degrees, remodulated upon an RF carrier, and radiated back into the body of the patient to cancel the cancer's signal patterns. Radiation of the body need not be severe at all; even weak radiation will suffice due to the kindling effect. At any rate, the hypothesized mechanism seems worthy of experimentation, and it seems to offer a validation of the practice of radionic treatment of disease.

DYNAMIC THOUGHT CONSTRUCTS: BIOFIELD MECHANISMS

We will discover later that thoughts and thought objects are real, physical objects in their own worldframe domain. As such, thoughts, or thought objects, or thought constructs can be and are quite functional devices in the thought world domain. In that domain, they are as physical as a rock. We will also discover that thought energy can be turned into electromagnetic energy by kindling, and can even be materialized if kindling is of sufficient intensity. As an example, thought energy from negative thoughts continually repeated or held in the mind can cause one to become physically ill. This is a well-known effect, and such an illness is referred to as psychosomatic. However, the mechanism by means of which it is accomplished is not known.

Actually the mechanism is simple: in the minute electrical currents that run through the nervous system, across the cell membranes, through the fibrous content of the cells, and through the cell water structures, virtual state thought energy is continually kindled into weak electrical patterns, which modulate, shape, and affect all the vital processes of the body. Almost all this kindled thought energy is from the content of the unconscious mind. Thus the locked-in contents of our unconscious minds eventually directly affect and change our bodies for either good or bad, depending on the particular content being kindled.

Small wonder that we suddenly acquired a chemical problem some years ago after we were saturated by thousands of hours of pill advertisements on television, which directly implanted the idea that certain chemicals such as pain relievers contain the answer to life's problems, while others such as alcoholic beverages lead to a glamorous, active, exciting life. Not only can such material implant suggestions in the unconscious for association, but it can also directly condition the response that the body itself will exhibit to chemicals and drugs.

It is now fairly well accepted that there exists a cancer personality and a heart attack personality, where the mindset is definitely and causatively involved. And from the Simontons' work, we see that the mind can be effectively used in conjunction with other therapy to reverse cancer, even terminal cancer. But unfortunately, by and large Western medicine continues to rely on mechanics and chemicals. Essentially the patient is viewed as a mechanical apparatus and a chemical factory that is damaged, polluted, or out of sorts, and treatment consists largely of mechanical intervention (such as surgery) or chemical intervention (such as antibiotics). Medicine continues to be a science of the reparation of damage or malfunction. It is the science of disease, not the science of health. I do not advocate abandoning the medical practices that have proven effective. I do advocate the augmentation of such measures by all the mind measures available. In fact, medical science ought to be vigorously pushing forward into fields of psychic healing, psychotronics, radionics, etc.

RADIONICS

In ancient development systems it was well known that the mind itself was a functional device, and that thoughts and thoughtforms could function as devices, influencing and affecting other minds and physical material. It was also understood that the thoughtforms had to be crystal clear and very stable. This demanded a great deal of practice in concentration. Eventually it was found that the stability of a physical form or device could materially assist the mind in maintaining stability. In ancient times many such devices were used: the dowsing rod, the crystal ball, the string-and-weight psychic pendulum, etc.

In the last century it was discovered that electrical devices could be used as tuning assistants. This led to the development of radionics, the name being inspired by the new electricity and magnetism discoveries, and particularly by the discovery of electromagnetic radiation. These radionics devices had a distinct advantage in the degree of precision they achieved in the tuning of the inceptive cyborg (operator's mind linked with the actual device through hyperspace). Ten resistors could be staged to give an otherwise almost unobtainable degree of accuracy and precision to the inceptive tuning. The use of variable resistors allowed rapid change of resistance, and thus rapid change of the overall cyborg tuning, and the ability to scan precisely in hyperspace, say, the emanations from a patient. In the hands of a skilled operator, such a device was often much more accurate than the finest medical test equipment available today, and infinitely cheaper and faster. In addition, such testing and diagnosis added little or no distress to the patient. This weird new radionics presented a formidable threat to the orthodox medical establishment. In the U.S. at least, the dogmatic medical establishment promptly came down hard on this potential threat to its hitherto unchallenged domination in the field of medical treatment. Its inquisition and purge was made easier by the fact that many charlatans also leaped into the lucrative field, having little or no actual diagnostic ability, and prescribing meaningless amulets and placebos for seriously ill patients. Also, the strangeness of radionics, such as the ability to diagnose illness from a drop of blood drawn from a patient thousands of miles away from the diagnostician and the radionics device, alienated many members of the medical profession who sincerely believed that no such diagnosis was possible, and that all such psychic demonstrations were frauds. Furthermore, the results, either diagnosis or treatment, were not usually repeatable by ordinary physicians and investigators since the psychic skill of the operator was directly involved. Organized medicine of course did not recognize the ability of the doctor to mentally or psychically intervene in the patient's illness, and so such practice was regarded as fraudulent and banned by the medical establishment.

The radionics practitioners were hounded into court, prosecuted, and convicted of fraud and of practicing medicine without a license. Several prominent radionics pioneers were actually hounded to death, Wilhelm Reich and Ruth Drown being the two most prominent examples. What could have been a great leap forward in medical science in the U.S., and the beginning of the development of psychotronics, was crushed by the power, ignorance, and bigotry of orthodoxy. To this day U.S. radionics has not recovered, and U.S. radionics practitioners are a rare breed, largely underground, who perform their experiments in suspicious secrecy and have only lately begun cautiously emerging from obscurity. In England and elsewhere, radionics survived and flourished, and today one can legally receive radionics treatment in England.

PSYCHOTRONIC DEVICES:
A SIMPLE SPACE (QUITON) AMPLIFIER

It is possible to speak of the smallest piece of nothing, that we will call a "quiton." It does not monocularly appear as a perceived thing. Quitons are pieces of action, in which $dEdT > h/4*pi$. Quitons are virtual entities, and very small ones.

Collection of a sufficient number of quitons results in breaching a threshold, so that a thing results. To superpose quitons (dEdT's) must be of the same magnitude and time coherent.

Thus in any situation involving a particularized nothing, collecting enough pieces of that nothing eventually exhausts the nothing set and reaches its boundary, whereupon it is a thing. This is really not different from collecting all non-dimensional points in a particular sequence and finding we now have a one-dimensional line segment, or collecting all the pieces of spacetime warp in a particular region and discovering that we have now a mass.

A device can easily be constructed to process dual natured or binocular entities (quitons), which are monocularly zeroes, while excluding monocular entities. Thus we can build a space amplifier or quiton amplifier or vacuum amplifier. Two single-state excluders in a series feed or input to an ordinary single-state amplifier or processor in a shielded container. The output is fed into a device which will rotate the dual-state entity or field so that an ordinary field can result. One side or the other of the dual-state output will be exclusively present in three-space.

TYPICAL HIERONYMUS DETECTOR

The Hieronymus device, patented by Thomas G. Hieronymus in 1949, is one such device. It has an optical front end, a prism which passes waves and does not admit particles. So a wave entering the front end will go through the prism and be refracted at an angle that depends on the frequency of the wave. Inside the surrounding box, which acts as a light shield, is a small tuner, a little copper rod attached to a rotatable wheel so that the rod may be moved through the various refractive angles from the prism. The prism is mounted on the box with a thin slit in the wall, so that only a small and narrow field of view exists external to the box. The tuner rod is wired to the input of a three-stage RF amplifier where each stage is separately shielded against light—which is very interesting, because of a similar requirement in the two-slit experimental apparatus if the electron was to act as a wave. The output of the RF amplifier comes out of the box and ends in a flat coil of wire between two parallel plastic plates.

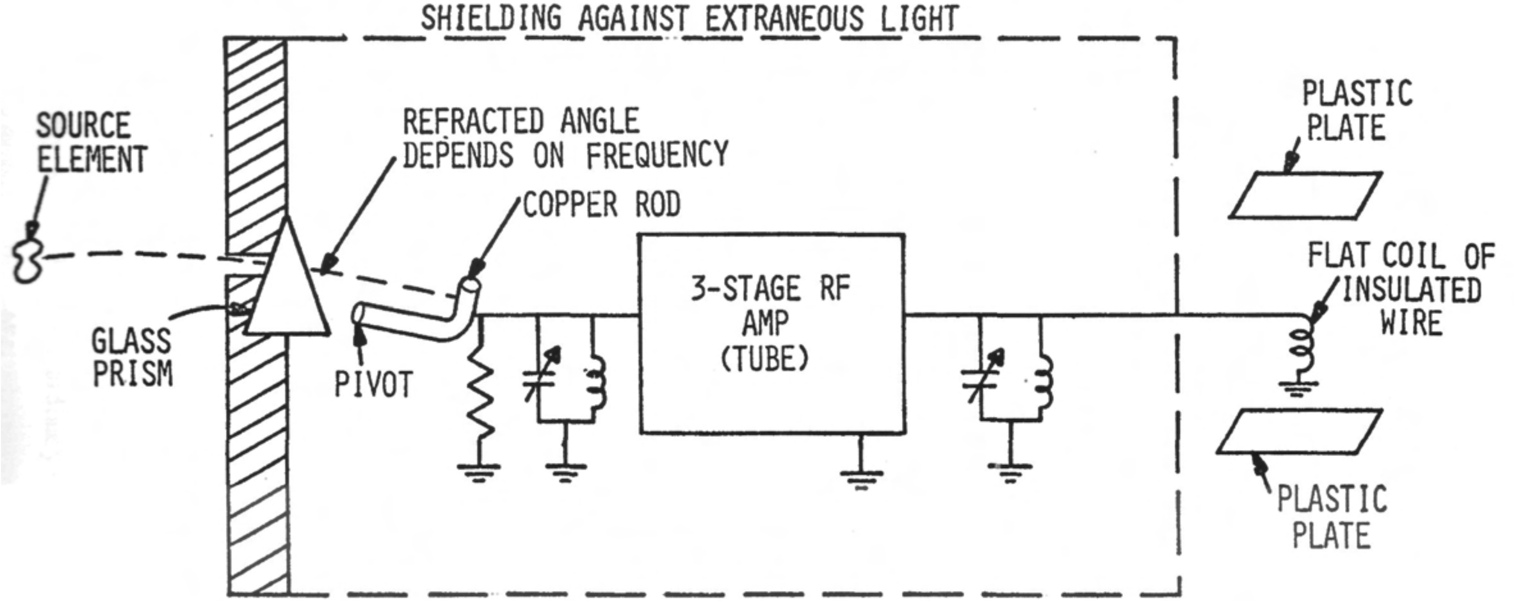
SHIELDING AGAINST EXTRANEOUS LIGHT
SOURCE ELEMENT
REFRACTED ANGLE DEPENDS ON FREQUENCY
COPPER ROD
GLASS PRISM
PIVOT
3-STAGE RF AMP (TUBE)
PLASTIC PLATE
FLAT COIL OF INSULATED WIRE
PLASTIC PLATE

Fig. 12. Typical Hieronymus detector

Coming into the box through the slit in the prism, there can be single-state entities and dual-state entities or quitons. The physicist does not have a good name for a dual-state entity. They used to be called wavicles; today we talk about wave packets. But what after all is a wave packet? It is a three-dimensional bunch of two-dimensional waves! So that is where the physicists have hidden the idea of the dual-state wavicle today.

Only single-state waves and dual-state entities or quitons can pass through the prism and be refracted at an angle dependent upon the frequency. If we then tune the rod into the correct angle of refraction, the refracted quitons hit it, as do the single-state waves if their frequency should happen to coincide with the frequency of the quitons. Now the single-state wave dies when it hits the copper rod; it may chip a single electron or two off a copper crystalline grain, but that is lost in the thermal noise anyway, below the detection threshold of the RF amplifier. The quitons, however, simply say "Oh! Now you want us to act like a corpuscular electron, and go through those wires and conduction paths. That's all right, that is just our right hand side." So the quitons will go through and be amplified, just exactly like ordinary electrons would do. But now it is not electron energy, it is a dual-state analog of energy — let us call it anenergy, for analog of energy. (Hieronymus called it eloptic energy, since it could act as electron flow or as optical waves, but was neither exclusively.) The amplified anenergy will come out in the coil of wire and it will generate field and it is not a magnetic field, and so it cannot be measured on normal laboratory instruments. The simplest way to detect the anenergy fields is to use the human sensory system, because the human body knew about RF energy, frequency modulation, and anenergy long before we had modern electrical and magnetic instruments. (The use of anenergy is illustrated by acupuncture, a very ancient medical system dealing with the fact that a peculiar type of energy-like stuff flows in the body, and in certain points and structures of the body, it can interact with other energies and fields.)

Now if we believe we cannot sense the anenergy fields, then we cannot. We can turn the entire anenergy detection system in the body off with the unconscious mind. The negative psi effect is a well-documented effect in parapsychology. There are goats as well as sheep. Some persons do worse on psi tests than chance would possibly allow. They are the goats. They exhibit the negative psi effect, for unconsciously they want to show you that psi does not work, so they use psi effects to do worse than is possible by chance.

At any rate, the human sensory system can get a tingle from the anenergy field generated by the flat coil of wire in the Hieronymus machine's output. What type of tingle one gets depends upon one's own type of body sensory tuning. It may feel as if the fingers on the plastic plate are in thick syrup. Or as if the plastic plate were vibrating. Or it may feel greasy in a peculiar way. And the negative person does not get a tingle at all.

The Hieronymus machine has been built by many persons, and it works for those who are not negative. It processes entities that exist in the dual-state, or that obey the fourth law of logic. And we can do some almost magical things with these dual-state nonthings, these nothings, if we set our minds to it. This is what psychotronics is all about.

ANOTHER KIND OF HIERONYMUS DEVICE AND THE INCEPTIVE CYBORG EFFECT

If thought constructs are actual operational mechanisms, then there should be other evidence besides psychosomatic illness and the demonstrated ability of the mind to cure certain illnesses. In fact, the Hieronymus device can be used to demonstrate this point.

John Campbell, former noted editor of Analog Magazine, once built an Hieronymus device and tested it successfully. Campbell immediately recognized that present physics could not explain the functioning of the device. In an exchange with Arthur Young, Young suggested to Campbell that it was the mind of the operator that made the device function, and that it was the symbolic form of the device that dynamically functioned to make it work. This appeared particularly significant to Campbell since he had discovered he could make the device work even though disconnected from its power supply.

So Campbell decided to test this thesis. He carefully drew on paper in black India ink a schematic diagram of the amplifier, removed the actual amplifier hardware, and substituted the schematic drawing. To his absolute amazement, the device worked as well for him this way as it had previously. But from our standpoint, such a result is not unexpected and can be readily explained.

First, I should point out quite strongly that Campbell not only was well-qualified technically, but even more important, he had an open and unbiased mind. This of course is quite necessary; the negative psi effect is well known, and the negative person is most certainly not going to allow his mind to operate the Hieronymus machine, just as he will not allow himself to succeed at dowsing, and so on.

Furthermore, in constructing the electrical schema of the amplifier, Campbell was strongly impressing stable thought constructs in his own unconscious, which was fully aware of the function of each part of the drawing. And when he connected that drawing to the rest of the hardware, he understood the function of the connection. He thus built in a thought connection from his thought-form amplifier to the hardware in a cyborg device. It then functioned as well for him as had the actual device. Campbell had proved that mind and matter may be cyborged into operational devices.

Psychic use of a pendulum, a ouija board, a dowsing rod, a crystal ball, and a crystal skull are other examples of the use of inceptive cyborg psychotronic devices, as is the use of radionics equipment of all types. And of course the most outstanding example of an inceptive cyborg where mind and matter are linked into one functional device is a living biological system.

Frances Farrelly, working with Arthur Young, took part in an experiment using an ancient Mexican figurine with a fluted neck. She succeeded in inducing fluted ears in a laboratory mouse. The mouse was exposed to the figurine about three feet away, and the cage or the mouse was connected to the figurine with a wire. Physical objects, such as mice bodies and fluted figurines, are three-dimennional channel tuners for orthorotating mind energies. Given firm initiation by a talented person such as Ms. Farrelly, a selected grouping of tuners can be resonated to collect bioenergy, and kindle it into ordinary energy and then into physical effects.

PROBABILITY: THROW OF A DIE

The fourth law of logic is absolutely indispensable in physics. We use it every day and do not realize it in probability. But what after all is probability? Let us use a very simple example to get at the answer to that question. Let us use the face of a die turned up. How can I model that, before the die is thrown?

Now we can only think by operationalism. To operate and output something is to automatically put it in the past. It's happened, it's gone, the moment you do it. To perceive an object is to put it in the past. To determine it is to put it in the past. To observe it is to put it in the past. There is no observed, perceived, detected, measured, or determined present. That is, there is no separated, exclusive, determined present such as is specified by the first three laws of logic — the fourth law is the present, by the way — but in observational physics which deals with determined, observed past phenomena, there exists no present. The future has not yet been observed, so it also is the unobserved. Only the past therefore is the observed. How then can one ever hope to model the unobserved present or the unobserved future?

If I look at the problem of the die with one face up, it is in the past. When I see it, it is in the past. When I think it, it is in the past. So if all I can observe, think, or perceive is the die in the past, how can I ever model it in the future? It's very simple! If I drive any problem set to its absolute boundary limit, it turns into its own opposite by the fourth law of logic, by the law of the boundary. So how do I do that with this problem of the die?

The problem set is specified by the condition "the perceived die with one face up"; this is the most recent past. Now simply find all the most immediate pasts you can get to meet the condition specified, and gather them all up together, and they then must turn into and comprise precisely the opposite, the immediate future. In this problem set, I can construct and collect six such pasts, each consisting of the perceived die with one face up. So by the fourth law of logic, those six faces up collected together as an ensemble represent the future and in fact are identical to the future. The present, which is simply the boundary between the most immediate past and the most immediate future, was specified by applying the fourth law of logic in the first place: the identity of the most immediate past and the most immediate future, being binocular, is unperceived, but it is the present nonetheless. So that is what probability is — an application of the fourth law of logic, so that the most immediate future can be represented in terms of the most immediate past — and physicists and mathematicians have been doing this ever since they have been doing physics and mathematics.

Without the fourth law of logic, there exists no rigorous logical basis for probability! So the fourth law is a very useful law indeed. We have simply failed to realize that we have been applying it all along.

The ontological problem can also be solved as follows: to state that *"A* is not" is to state that *"A* is" is not; to absent a presented thing is to first present it; to be "not-being" is a clear statement of the problem. By the first three laws of logic, the problem is not soluble. By the fourth law, it is simple since being and not-being can be identified. All distinction between present (being) and absent (not being) can be lost.

NEGATIVE TIME

Negative time does exist and is used in advanced physical theories. Dirac predicted an electron traveling backward in time, which would be observed in positive time as a positive charge traveling forward in time.

The application of negative time to philosophy also allows the solution to the problem of good and evil. In positive time we may consider an act as evil — say the act of a hungry wolf seizing a hapless rabbit and devouring it. In positive time, we might say that to the rabbit the act was evil, and to the hungry wolf the act was good, noticing in the process that good and evil are relative terms, not absolutes. Examining the situation now in negative time, a very strange thing occurs. The remains of the consumed rabbit are regurgitated from the wolfs belly, again creating a living breathing rabbit, and reintroducing hunger into the wolf. Thus the act now is beneficial to the rabbit and detrimental to the wolf.

The point is that Allness or Godness is both sides of every coin, not one side only. Omega Godness exists both backward and forward in time; it must do so if it is to be unlimited, even by time itself.

Thus to such unlimited allness, every act is both good and evil simultaneously. Thus we can understand how a perfect creator could establish and ordain good and evil, for there is a sense in which the two are synonymous.

However, such insight is a very strong dose of medicine, too strong for the average stomach. Most persons never seek the Unlimited Godness; instead, they seek a good facet of Godness that itself moves forward only positively in time. Thus they find a partial image which is limited in and by the flow of time. By such an assumption the biblical phrase "I am Alpha and Omega" is unconsciously translated into "I was Alpha and I will be Omega." Thus orthodox Christianity long ago lost much of the God-insight originally given by the avatar Jesus. Only those few cloistered individuals who still struggle with exegesis grapple with the problem of good and evil. The problem indeed has a solution, but the solution involves negative time, the fourth law of logic, and an unlimited creator.

TIME OSCILLATION, QUANTUM CHANGE, AND MASS

It is well known in physics that almost all physical theory is time-reversible, i.e., the equations hold true whether time is positive or negative.

This is important in the theory of antimatter. A positron (antielectron) may be regarded as an electron traveling backward in time. The destruction of an electron constitutes the creation of a positron, and vice versa, when we are observing in only one kind of time.

However, when we consider monocular quantum change, we realize that time can be modeled as a wave. To have any of the present physical quantities (energy, length, time, and momentum changes), a quantum of action must be isolated and fissioned into fragments. By this process we get a change in any physical quantity. This is fairly well understood in present physics, though not in this exact terminology.

However, what is not recognized in physics is that a disposal operation must be performed to clear away the fragments of one physical change before another can occur. The only way the fragments can be cleared away is to fuse them back into an action atom. Since action itself cannot be perceived (only a change in action can be perceived, not action itself) then fusing the fragments into a quantum removes the fragments.

This fusion operation removes positive time fragments (and other positive fragments) and so can be considered a negative time operation. Thus time can be considered to be an oscillatory wave because of its continual monocular discrete creation and annihilation by the quantum change process.

In fact, time oscillation may be considered to be the fundamental reality, and quantized change its result.

By the complementarity principle, time, being a wave, also must sometimes exhibit a corpuscular nature. (Indeed, all virtual state particles are nested within the structure of time.)

The Soviet work on time generating force thus is of absolutely fundamental importance. What has been released of K'ozyrev's work seems to be intentionally deceptive and of little value. His real work and that of his colleagues appears to have the highest Soviet classification, and none of this material has been released.

In the West, the only person of note who has grasped the fundamental import of time and time oscillations is Charles Muses. As early as 1958, Muses pointed out that there was no mystery in the two-slit experiment if the chronotopological features of the experiment were properly taken into account. Muses also worked out a hypernumber structure of time, but its significance seems to have escaped orthodox physics. Nonetheless, his work has direct technological implications, such as materialization and dematerialization of matter, and offers an approach to the engineering development of actual devices. Since his hypernumber theory of time structure also encompasses consciousness, it represents a union of physics and metaphysics on a common basis, and an approach to the engineering development of psychotronic devices.

Western physicists, then, would be well advised to pursue vigorously the idea of time oscillations, both theoretically and experimentally. And they would also be well advised to examine the hypernumber work of Charles Muses.

Since mind and virtual state occupy the time dimension, the nested structure of time contains the mechanisms for mental-physical interaction and for direct amplification of virtual state into observable state (or for the dissolution of observable state into virtual state).

MASS IS A TIME-DIFFERENTIATION

The most fundamental aspect of the concept of mass is that it occupies space, i.e., that it is three-dimensional. Mass thus is a time differentiation of Minkowskian spacetime L^3T. Applying the fourth law of logic, a thing does that which it is and is that which it does. Thus mass is a time-differentiator.

Because of this, we cannot see time, the fourth dimension, although we can see L^3 length, width, and height. Because we look through and detect with a mass sensory system, the time dimension is lost in the act of looking and detecting, i.e.,

$$d/dT(L^3T) = L^3$$

and we thus see a spatial reality. Physical instruments also have mass and do precisely the same thing. For this reason, ultimately all measurement or detection is spatial rather than spatio-temporal, as is well known in physics.

Note that we do not see or detect total reality, but only a partial differentiation of it. Thus the position of the materialist's view is ironic: in his absolute insistence that reality consists in only what can be seen, touched, felt, or detected and measured with his instruments, he condemns himself to the greatest illusion of all, the notion that the outputs of his physical sensory system and his physical instruments constitute absolute reality. Also, since the only physical dimension shared by mind is time, mind reality is lost in the act of detecting, and hence only phenomena which are totally separated from mind are detected.

By the fourth law of logic, the existence in no mind at all constitutes the existence in each and every mind (nonexclusively). Thus physical phenomena are those phenomena that are processed by every mind. We will see that the fourth law enables us to deal precisely with mental, physical, and paranormal phenomena on that basis (how many minds the phenomena exist in).

REALITY IS LIKE AN OPOSSUM

Reality is like a mother opossum carrying her little baby opossums on her tail, i.e., the entire mental-virtual state is carried along as virtual constructs, nested comfortably in the delta time portions of each photon. In the piece of time that is fissioned out of an action quantum to give observable change, there are a multitude of little pieces of time still attached to subquantum virtual state entities and hanging on, exactly like little baby opossums hanging on to their mother's tail with their own tails.

The world is so built, then, according to physics. And if we include our physical model of mindchanges, every possibility in fact concretely exists in that manner, in the virtual state, timewise hanging into the time dimension of the detected physical world's four-space frame, but spacewise hanging out of the ordinary three-space physical universe of objects. So indeed all possible realities, even the wildest and the most farfetched, exist in the virtual state.

The time channel is thus quite literally the paranormal channel that connects the normal spatial world to an infinity of nonnormal spatial worldframes. Only by the continual destruction of each piece of created time (by fusing it back together with the piece of energy involved, and forming a quantum which is unperceivable) are we able to experience physical reality at all, uncluttered and relatively free from the vast hordes of infinite realities through which the ship of quantum-change physical reality sails.

Since the electromagnetic field represents a collection of this vast array of realities, then by proper selection (i.e., by coherent tuning) of multiple stages we should be able to select a particular virtual reality channel. And indeed so we can. Any scheme for drastically increasing signal-to-noise ratios should be adaptable to such a collection mechanism for the detection of bioenergy.

In another sense, reality is like a television set with an infinite number of reality channels, but only one of which is tuned in at any one time. By adjusting the tuning, however, a different reality, including even different physical laws of nature, may be tuned in.

BIOFIELD CONCEPTS

To understand the hyperframe approach to fields, we must understand the dimensionality of intersections in n-dimensional space. The dimensionality of an orthogonal intersection is one less than the dimensionality of the intersector having the lowest dimension. One intersector appears radically different to the other intersector since the second sees only a derivative of the first. It is the total projection of an intersector that is "seen" by that which it intersects.

Hubbard's manifold theory also derives these principles. The "types of reality" that these other objects appear to be will depend on the types of intersections they make with the frame of the observer.

This cannot be stressed too strongly: an observer sees his own frame, and intersections in that frame, nothing else. He does not necessarily see the entire reality of any object, and if an object is in a frame that does not intersect the observer's frame, the observer cannot detect that object at all.

Thus all types of reality are open to reinterpretation in the light of intersecting and nonintersecting frames. And to shift dimensions (increase or decrease), all we need do is add or subtract additional orthogonal intersections.

MASS, FORCE, AND 3-D FRAME ROTATION

It is well known that a moving mass can be considered as existing in a three-dimensional space which is rotated away from the line of motion of the mass as seen in the laboratory frame. Einstein in fact pointed this out.

As the velocity of the object approaches c, the speed of light, the angle of rotation approaches 90°. From the bottom left figure it can be seen that a force in the laboratory frame, applied to the moving mass along its direction of motion in the laboratory frame, only affects the mass in the mass's rotated frame with a projected portion. To the laboratory observer, the mass's resistance to the disturbing force applied seems to have increased; thus to him the inertial mass of the object seems to have increased, i.e., if it is harder to push, its resistance to pushing must have increased.

But in the rotated frame, the object's mass has not increased. The rotated frame effect is *why* the mass of an object increases to one observer but not to the other.

When the rotated frame can be rotated to a full 90°, no force applied by the laboratory observer can accelerate it further, because the applied force has zero projection in the rotated frame where the mass actually is. Thus the mass (resistance to pushing) appears infinite to the laboratory observer, while to the rotated observer it is still exactly the same as it always was. Furthermore, to the rotated observer, **the** force being exerted on the rotated mass is zero.

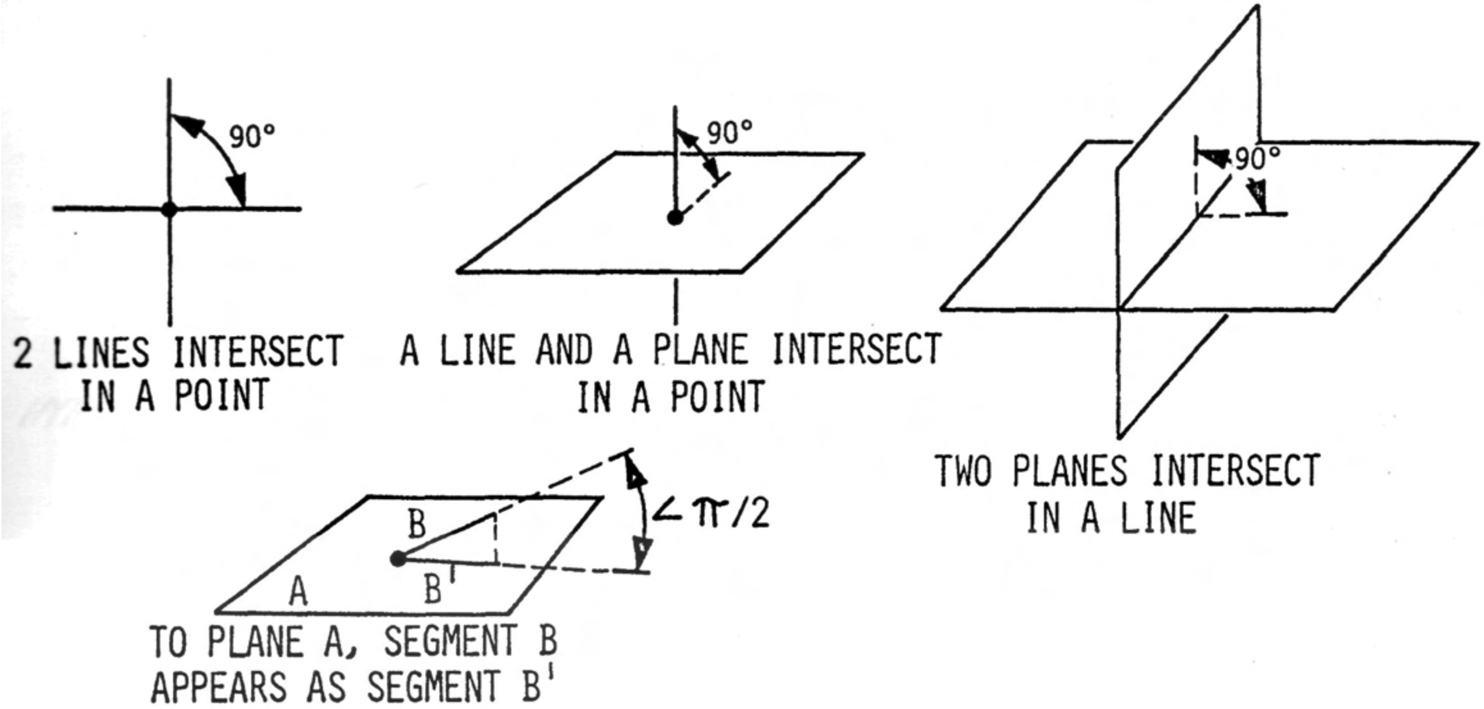
90°
2 LINES INTERSECT IN A POINT
90°
A LINE AND A PLANE INTERSECT IN A POINT
90°
TWO PLANES INTERSECT IN A LINE
B
A
B'
∠π/2
TO PLANE A, SEGMENT B APPEARS AS SEGMENT B'

Fig. 13. Biofield concepts

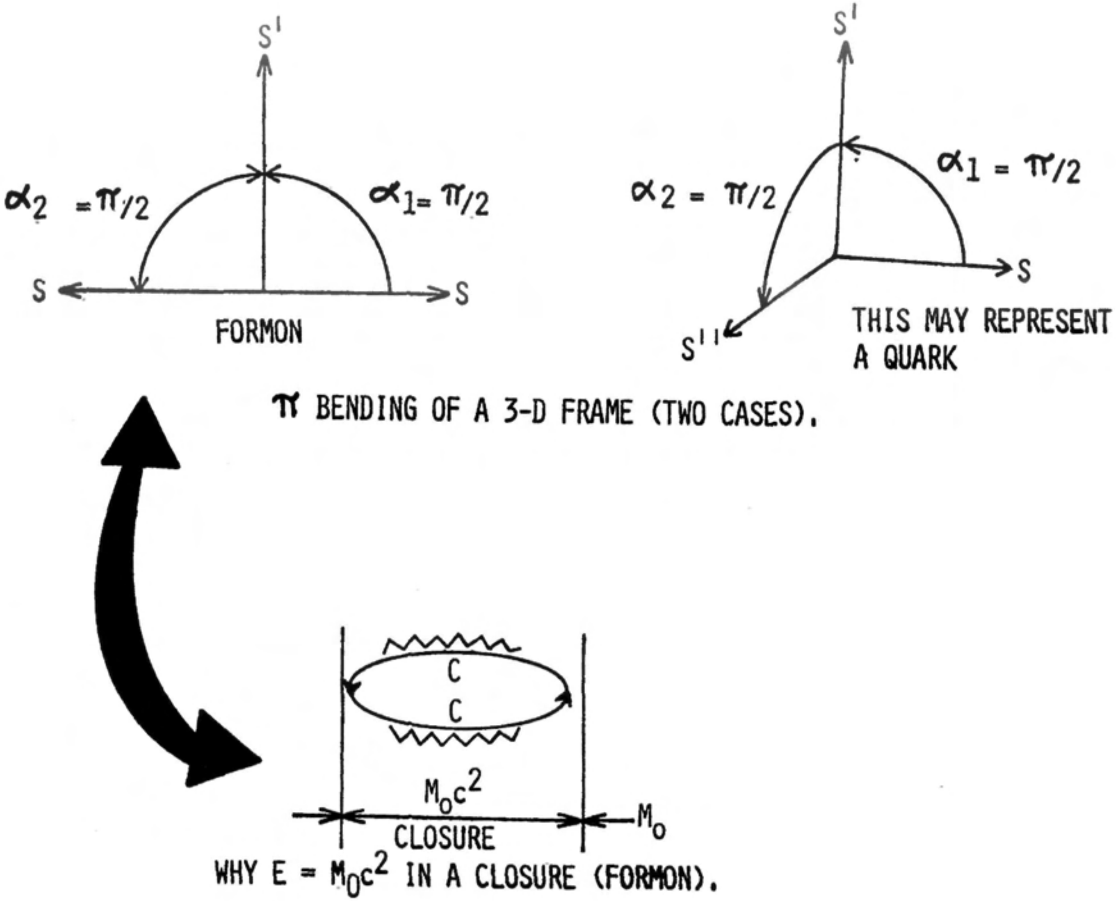
S'
$\alpha_2 = \pi/2$
$\alpha_1 = \pi/2$
S
S
FORMON
S'
$\alpha_2 = \pi/2$
$\alpha_1 = \pi/2$
S
S''
THIS MAY REPRESENT A QUARK
C
C
M_0c^2
M_0
CLOSURE

π BENDING OF A 3-D FRAME (TWO CASES).

WHY $E = M_0c^2$ IN A CLOSURE (FORMON).

Fig. 14. Rotation of 3-D frames

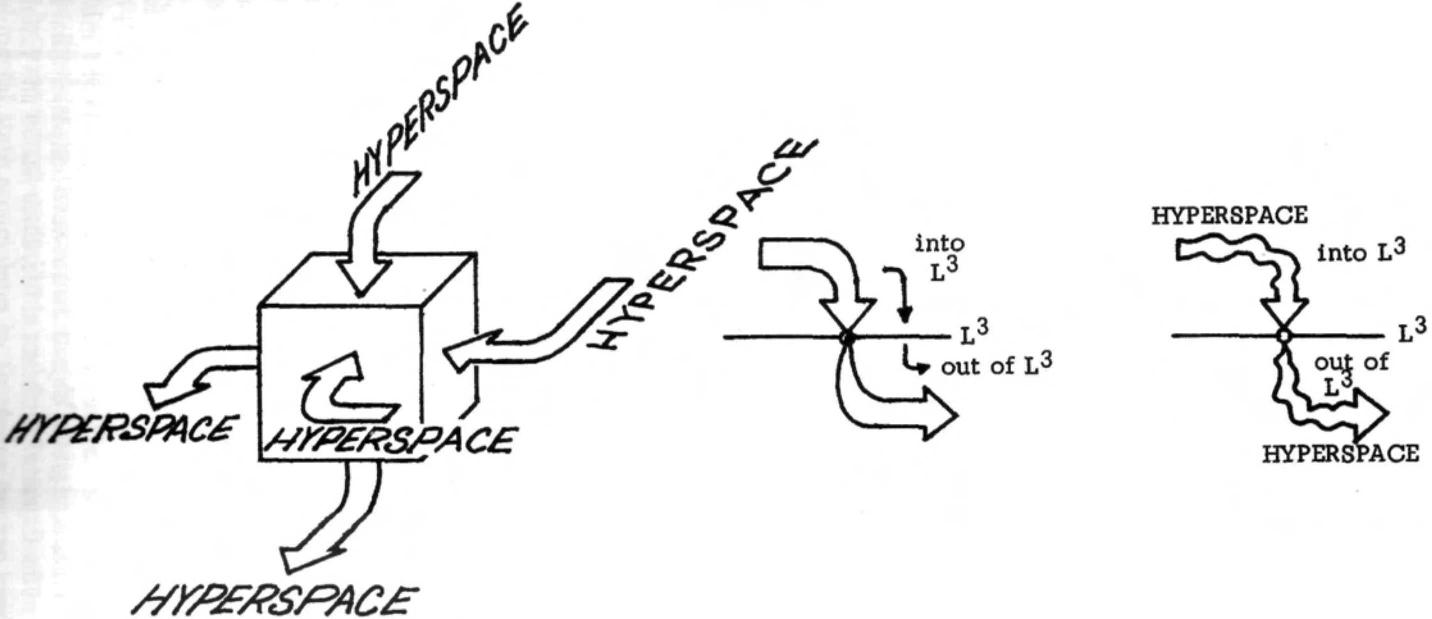
HYPERSPACE
HYPERSPACE
HYPERSPACE
HYPERSPACE
HYPERSPACE
HYPERSPACE
into L^3
L^3
out of L^3
HYPERSPACE
into L^3
L^3
out of L^3
HYPERSPACE

MASS
A region of
hyperspatial flux
orthorotating in and
out of L^3 3-space.

MASS PARTICLE
The flux does not
"move through" 3-space.
It rotates in and out of
each individual mass particle.

JITTER
Each flux line is not
smooth, but jittered.

Fig. 15. Mass is π-rotation of hyperspatial flux

However, the mass's intersection in the lab frame is now two-dimensional. Since only three-dimensional objects can have mass, the object is massless to the laboratory observer. So here we have a paradox: zero mass is identical to infinite mass. This is perfectly in consonance with the fourth law of logic.

The absence of any single finite mass actually permits two indistinguishable solutions: the absence of any three-dimensional mass at all, and the exclusive presence of more than any finite mass. The two opposites are identical on the boundary case.

Thus a photon has zero mass because it is a two-dimensional object and only three-dimensional objects have mass. A photon also has infinite mass because it cannot be accelerated along its direction of travel. A photon is a perfectly ordinary three-dimensional particle existing in a normal three-space that is one orthogonal turn away from the laboratory three-space.

MASS IS π ROTATION OF HYPERSPATIAL FLUX

In Figure 15, I show that a localized mass is simply the result of orthorotational flux in and out of that localized region. This is a hyperspatial flux, since it actually is flowing or rotating in hyperspace.

This flux is related to the spin of a fundamental particle, and in one sense may be regarded as an extension of the concept of spin, i.e., ordinary spin is simply a derivative subset of hyperspatial spin (orthorotation between orthogonal three-space).

This flux does not move through laboratory three-space, but rotates into and out of it, at and for each fundamental particle of mass in laboratory three-space. As I point out elsewhere, mass itself is nothing but the absolute value summations of action quantum fission and fusion activity in a localized (fundamental particle) spatial region *(Quiton/Perceptron Physics).* Fission and fusion are orthorotations, or hyperflux activities.

Also, if we represent this hyperspatial flux by flux lines (a standard practice in physics), then each line is jittered because of its modulation by many higher-frequency components. This flux and the jitter in it play an important part in paranormal phenomena.

Furthermore, the hyperspace flux can be seen in two other ways: (1) as continual direct coherent collections of virtual subquanta to reach the quantum threshold, breaching it and becoming observable quantum change. Thus it should encompass essentially all hidden variable theories. This is the same mechanism by means of which virtual photons kindle and create electric field and the charge of an electrically charged particle; (2) as continual direct coherent collections of mind virtual state subquanta to reach the quantum threshold, breaching it and turning into observable quantum change (physical phenomena). Thus all matter is totally mind and totally alive — inert matter is just nonpreferentially alive. The resultant flux vector in an inert mass is zero since there exists as much flow out as in, and vice versa. This totally eliminates the classical dichotomy between mind and matter. It also implies that accelerated objects do not have nonpreferential type of life, but in fact have some degree of preferential life, since fission and fusion rates of an accelerated object are not equal.

Establishing sufficient pattern coherence in the parts of an accelerated system thus constitutes a preferential tuning between multiple parallel-directed directional coherences. This effect can result in establishing a one-to-one correspondence to and from one of the higher orthoworlds, i.e., it can result in a coherent circulation between that universe and the specific inert accelerated system. This circulation is easiest to establish through magic frequency windows where the channels of least resistance exist. This is the basic mechanism by means of which a living biological system is produced in and from the appropriate primeval amino acid soup in the presence of (a) magic-window induction (such as in the ultraviolet), (b) the acceleration of otherwise inert particles (such as is locally provided by lightning discharges), and (c) the strong collection of bioenergy to orthorotate it through three rotations (as is also provided by lightning discharges).

It is not accidental that life can be and has been created in the laboratory. This type of experiment, done by a team of strong psychics projecting intentional bioenergy forms, should be able to create specific forms of life. The same type of mechanism can occur inside the living body where it can induce psychomatic illness. A similar mechanism is probably responsible for the placebo effect, and possibly for the body's production of antibodies and the psychic modification of enzyme activity, psychic healing, etc.

JITTER OR DITHER MODULATES THE PRIMARY FLUX

In the continual, virtual, subquantum, orthorotational flux (crosstalk), rotating into and out of a fundamental mass particle, only those components that are coherent are additive, and only those that add to the quantum threshold and breach it become observable quantum change.

In a given observable quantum change consisting of such coherently adhered subquanta, time is produced by fission of the quantum of action. All remaining statistical subquanta which have not collected to the threshold level are virtual subquanta of action possessing unfissioned, unseparated time. These virtual subquanta, which themselves are dynamic and gradually collecting and superposing to various degrees and at various rates, are attached to and exist in the dT interval produced by fission of the action quantum. Thus each and every piece of time contains dynamic, nested, virtual subquanta. Time in observable changes may thus be said to possess an internal dynamic structure filled with such virtual subquantum mental, neutrino, and photonic entities.

At any fission time, the degree of superposition of each of these nested entities is highly variable and depends on the whole past collection history, which in turn depends on all flux components of the proper frequency that have been arriving from all the orthoframes.

Thus virtual-mental entities cohere frequency wise until they breach the quantum threshold and discharge into observable quantum change. It is much like the accumulation of a static charge until sufficient potential is reached and a spark occurs, discharging the accumulation of charge. The same sort of relationship exists between virtual-mental state and observable state. The only difference is that virtual-mental state accumulates by frequency or time phase.

Since present physics is entirely localized, local observations cannot have taken into account distant matching-frequency events in the laboratory frame, or matching-frequency events in orthogonal frames. To the localized particle of mass, the precise moment a particular subquantum superposition of given frequency will superpose to the threshold and turn into quantum change is therefore not predictable. Since the remote causes cannot be accounted for, the local effects cannot be time-ordered.

Therefore at the quantum level, the physical change is totally statistical to a monocular observer. This accounts for the fact that the quantum world is statistical rather than causal.

However, a given fundamental mass particle has a structure of physical aspects, e.g., mass, spin, locality, frequency spectrum, switch rate, strangeness, etc. All of these may be expressible in terms of orthorotational concepts, although not all of them have yet been so worked out. Muses has already pointed out that some of the present quantum numbers are actually hypernumbers. And just as $\sqrt{-1}$, the first hypernumber, is an orthogonal rotation, so are the other hypernumbers.

At any rate, one may regard a fundamental particle as simply a complicated three-space tuning pattern existing in the orthorotational crosstalk flux between orthoframes. This pattern may also be said to consist of sets of frequencies, since each orthorotational flux line has a frequency of in-out-operation. Thus each fundamental particle is a set of tuning frequencies in the flux. A stable particle is a stable spectrum comprised of standing waves. An unstable particle is an unstable set, comprised of some standing waves, which account for the duration, and some not-quite-standing waves, which account for the excited state or degree of instability of the object.

On each flux line, orthorotational flux lines exist, constituting jitter or dither. Also, every fundamental particle is a tuning. It both tunes and is the result of its tuning, i.e., by the fourth law of logic it does what it is and it is what it does. Since tuning allows frequency communication between orthoworlds, then free-energy devices are not only possible but are practical. So are dematerialization-materialization devices, teleportation devices, devices that heal or kill, and anticancer devices.

THE UNTOLD STORY OF QUARKS

We hear so much nonsense about the objectivity of science that it occasionally seems necessary to debunk the sacred cow. In fact, the quark story is a prime example.

In 1910, Millikan made his famous oil drop experiment and determined the charge-to-mass (e/m) ratio of the electron. By multiplying this ratio by the mass of the electron, the amount of charge on the electron might be computed. Thus was born the idea of a constant amount or quantum of electricity, the fixed charge on an electron. Millikan received the Nobel prize for this famous experiment.

Actually, Millikan himself did not originate the experiment. One of his brilliant graduate students, Harvey Fletcher, first conceived the idea. Millikan became interested after he saw what a stir the results were creating, and he accepted the Nobel prize without sharing the credit with his brilliant student. Later Dr. Harvey Fletcher was to become the head of Bell Telephone Laboratories and an eminent managerial scientist in his own right. He also is the father of Dr. James Fletcher, former Chief Administrator of NASA.

Incidentally, Dr. Harvey Fletcher is still alive and alert, should anyone care to check the story! Nonetheless, it is highly doubtful that this grave injustice to Dr. Harvey Fletcher will be corrected.

But the story does not end there. In the late 1930's, a German physicist named Ehrenhaft decided to change the Millikan experiment and use charged solid particles rather than charged liquid oil drops. In fact he used red selenium spheres of various sizes. Amazingly, he discovered that he could isolate and measure fractional charges (quarks) at will, and he could even control the magnitude of the fractional charge isolated on a selenium sphere by varying the diameter of the sphere itself.

His work was duly published in scientific journals, and duly greeted with derision and hostility. Ehrenhaft himself disappeared during World War II. And there the matter lay. Dogmatic scientists continued to blindly teach the indivisibility of the electric charge, Millikan retained his exalted position and Ehrenhaft's work passed into comfortable obscurity. And in sophomore physics laboratories all across the nation, the indivisible quantum of electricity held full sway, dutifully proved over and over by the oil drop experiment.

Then in the 1960's, the particle physicists, despairing because of the bewildering new array of fundamental particles that continued to pour in confusion from their massive high energy machines, turned again to the idea of fractional charge (quarks) to try to make some unifying sense of the embarrassing abundance of strange new particles. By choosing three simple quarks, they could unify the theory of almost all the new particles. (Additional quarks have become necessary in the theory since then.)

So physicists gradually began to look for quarks, but in the giant, sophisticated, cumbersome experiments and devices that were yielding all the strange particles, not in the simpleton's approach that Ehrenhaft had taken. And they could not find them! Try as they would, a quark just refused to leave its tell-tale track. But in desperation they kept looking, and kept missing them.

Then in 1972 a convention of physicists was held at the International Center for Theoretical Physics in Trieste, Italy. Seven Nobel Prizewinners in physics attended the convention, whose theme was the physicist's conception of nature. Paul Dirac, the Nobel prizewinner who had theoretically predicted the positron before its discovery, gave the keynote address to the distinguished gathering.

Amazingly, Dirac presented the work of Ehrenhaft to the startled convention! He showed some of Ehrenhaft's data, and then pointed out that Millikan had also reported a single measurement of fractional charge, but had discarded it from consideration because he felt it was probably an experimental error. Dirac remarked that it all made him wonder whether Millikan might really have had a quark in that single discarded measurement.

The die was cast, and the seed finally planted. In 1977, about forty years after Ehrenhaft first obtained his quarks, someone finally got around to using solid paniculate matter. Researchers at Stanford University tried it out using niobium, and they got a little fractional charge on their niobium. So after forty years, Ehrenhaft, who had so long been ridiculed and ignored, was finally vindicated. It had taken the establishment that long to check his work instead of dogmatically rejecting it. Yet in none of the articles I have seen on the new discovery has there been a single word about Ehrenhaft, or of Dirac's presentation of his work at Trieste.

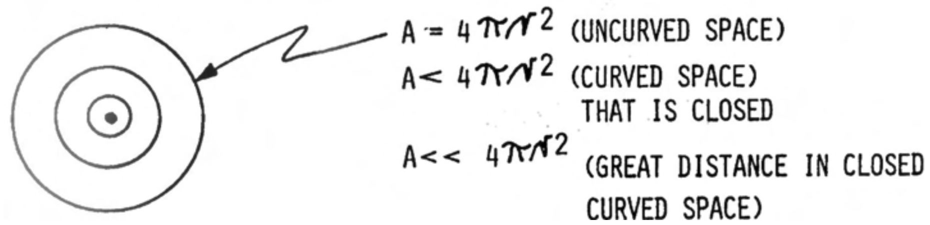
A = 4πr² (UNCURVED SPACE)
A< 4πr² (CURVED SPACE) THAT IS CLOSED
A<< 4πr² (GREAT DISTANCE IN CLOSED CURVED SPACE)

Fig. 16. Einstein's spherical model of the cosmos

EINSTEIN'S SPHERICAL MODEL OF THE COSMOS

Einstein's spherical model of the cosmos is a primary example of a holographic universe. In such a closed spacetime, each point inside the universe is also at the extreme end of the universe in any direction. Thus the entire physical universe is totally outside any of its internal points, and totally inside each internal point as well.

That is an application of the fourth law of logic. The total internal is identical to the total external. The adjective "total" merely moves one to the common boundary between the operational concepts of "internal" and "external." At that boundary, there is no operational distinction between one and the other.

The universe closure is not limited to just one occurrence or one rate of closure. In fact the universe can then proceed to close again within the same local three-space, e.g., a gradual macroscopic closure due to gravity field (or causing it) constitutes the external universe, and a second extremely sharp microscopic closure due to electric field (or causing it!) constitutes electric charge. The two together constitute a fundamental charged particle of mass. The polarity of charge is determined merely by which of two directions the second closure was made in. Considering an electron as such a dually closed entity, it is readily seen that the closure ratios are all that is being referred to by Feynman's condition. Feyman pointed out that any unified field theory must explain why the ratio of the radii of the Einstein closed universe and the classical electron, and the ratio of the electrical and gravitational forces between two electrons, were the same — on the order of 10^{42}. The model is also consistent with Santilli's proof that the electric field and the gravitational field are either the same thing or partially the same thing. In our model, an electric field is in fact a highly compressed gravitational field. Proper time synchronization should allow fantastic antigravity effects to be realized, and thus electrogravitics has a basis in this model.

A CHARGED PARTICLE IS A DUAL CLOSURE OF THE UNIVERSE

It is proposed that each fundamental particle having electric charge represents a dual closure of the entire physical universe. The gravitational field closure is slow and weak, and it is that closure which results in the so-called external physical universe.

It is proposed that all particles participate in this same closure, since it is their combined distributed mass that represents or causes the closure. (By the fourth law, "represents" and "causes" become simultaneous at the boundary, i.e., when closure is accomplished.)

There are two directions possible for this gravitational closure: one where gravitational masses appear to attract, and one where gravitational masses appear to repel. These two directions thus form antimatter closures with respect to each other. The G closure where G masses attract is said to be the closure associated with a positive time dimension, while the G closure where G masses repel is said to be the closure associated with a negative time dimension. The direction of closure represents or causes the time dimension's algebraic sign, and vice versa (fourth law).

The electric field (E) closure is strong and fast — on the order of 10^{42} times as strong and fast as the G closure. However, its closure is in a direction orthogonal to the direction of the G closure. Since all matter is in motion, a slight deviation from exact orthogonality actually exists between the G and E closures due to crosstalk projection of components from the slight frame rotation involved in the velocity of each mass itself.

There are two primary orthogonal directions for the E closure. One direction yields positive charge and the other yields negative charge, for the positive time closure. For the negative time closure, the directions are reversed.

Such a schema, or version of it, can be made consistent with Santilli's proof, Feynman's criterion, and the observed fact that electricity can affect gravity and vice versa (electrogravitics). In fact, the electric field becomes a highly compressed gravitational field, at nearly 90° from the normal gravitational field. Liberating the compressed gravity in a flashlight battery could accomplish fantastic antigravity effects. Ironically, antigravity should be the simplest and cheapest effect ever dreamed of, once practical mechanisms for orthoshifting E field to G field are discovered.

Since psychotronics and quarks involve more fundamental components of electromagnetic field, hyperfields or biofields, psychotronic devices to yield antigravity should be quite feasible. The undisclosed extensions to Pavlita's psychotronic generators are thus of absolutely fundamental importance.

Furthermore, the ancients appear to have possessed psychotronic devices. Considering that metals, crystals, and other common materials can be used to. construct psychotronic generators, and great power is attainable by the hyperchannel amplifier effect, it is quite possible that priests of several ancient civilizations actually possessed psychotronic devices to lift giant stones by antigravity, and other psychotronic devices to carve the stones like jelly. They may also have possessed psychotronic flying machines which used antigravity for lift and something like rocket propulsion for horizontal thrust. A flying machine that gets its lift from antigravity might be extremely crude in external appearance, and the use of hyperspace for travel would make it unneccessary to travel extensively through ordinary three-dimensional space. Thus many aspects of travel through normal space such as air drag and fineness ratio are of little importance to a psychotronic vehicle.

ADDING AND SUBTRACTING DIMENSIONS

Figure 17 shows the relation between total collection (integration) and orthogonal rotation in hyperspace.

Total collection (integration) constitutes orthogonal rotation in a direction to increase the number of dimensions by one

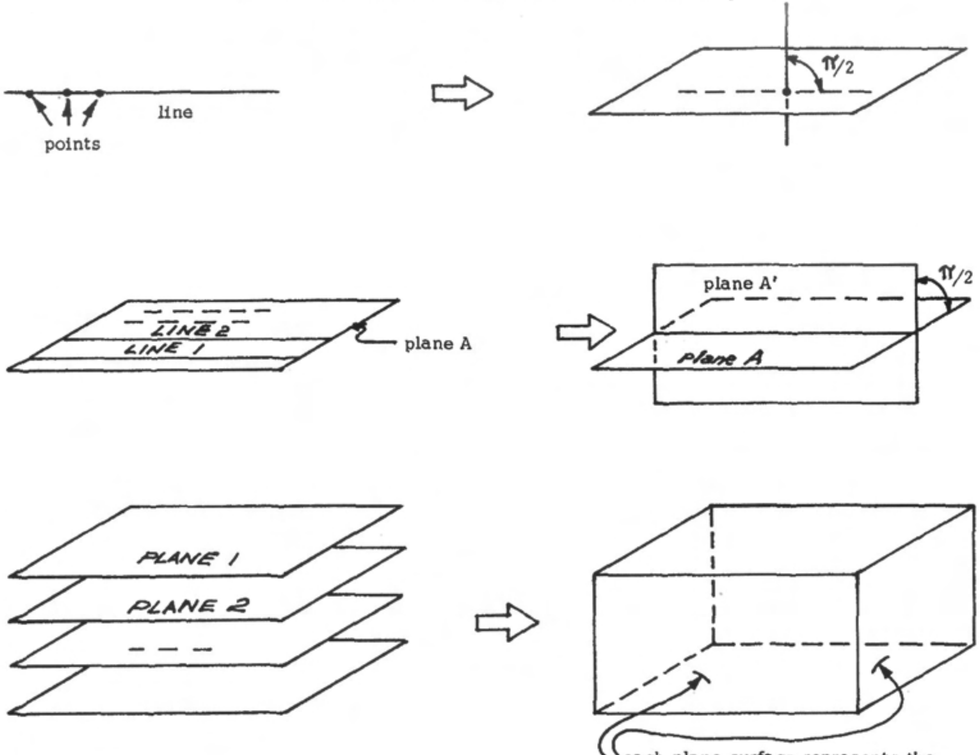
line
points
π/2
LINE 2
LINE 1
plane A
plane A'
π/2
plane A
PLANE 1
PLANE 2

each plane surface represents the intersection, in the observer's 3-D frame, of this 3-D solid rotated into an orthogonal frame in hyperspace.

Fig. 17. Adding and subtracting dimensions

Total collection of zero-dimensional points in one direction integrates the points into a line (one-dimensional entity). This is precisely the same as if any of the points had been the plane intersection of an orthogonal line, and that line had now been rotated into the plane. Total collection in one direction is dimensional integration, and the collected entities gain an additional dimension, i.e., total collection constitues orthogonal rotation of the higher dimensioned object into observer three-space.

A point can be considered to be, and modeled as, a line in a frame one orthogonal turn away, and simultaneously as a plane two orthogonal turns away, and simultaneously as a solid (in fact, an entire three-dimensional physical world) three orthogonal rotations away. The latter may also be considered to be, and modeled as, a mind, or even as an infolded infinite set of minds. The concept does not stop at three dimensions, but proceeds into higher realms indefinitely. However, three dimensions plus an orthogonal time dimension are sufficient for our purposes.

A converse operation also exists. Differentiation consists of a cut in an object whereby one single element (of the collection the object represents), is separated and isolated, so that one dimension is lost. It is like a closure being opened at one element. This is the same as an orthogonal rotation of the entity (being differentiated) in a direction departing from observer three-space.

In the Figure, the lower two figures show the collection-is-integration schema for higher-dimensional figures. In the lower right figure, we cannot draw the three-dimensional object in its orthorotated position in four-space since we have only three spatial dimensions in the frame. Nonetheless, the additional dimensions exist and are simple to treat mathematically.

MIND IS OBJECTIVE

In time differentiating, mass loses the time dimension. Therefore we cannot see time with a mass perceiver, but we can see L^3 (spatial extension). But time is the only so-called objective dimension occupied or shared by mind. Therefore we also cannot normally perceive mind, because the plate (time) on which it was sitting is lost in the perception process.

Thus Descartes was both right and wrong. He was right in that monocular mind is not present in physical detection output, but wrong in believing mind and physical phenomena are totally separate. Time does not exist except with respect to between one L^3 perception and another, if L^3 is all that is outputted. Physical time is totally relative, totally mental, and exists only in memory. Mind is thus present in $L^3 T$ four-dimensional physical phenomena and is discretized along with time discreteness in the Heisenberg uncertainty relation $\mathrm{d}Edt >= h/2*pi$.

In fact, we can even take the extreme view that time is mind. Because to establish a delta t, a former L^3 must be established and compared to a more recent L^3. The only place the former L^3 can rigorously be found is in memory. An event, once occurred, is in the past and is forever gone to the observer. Only in his memory can it be said to exist. But then so is the observation also of the most recent L^3. Thus, rigorously, observed phenomena may be said to exist only in memory, and there is only a past, never a present. Thus a full four-dimensional mind is inseparable from perceived physical phenomena, and mind is quantized along with quantum change in the physical world.

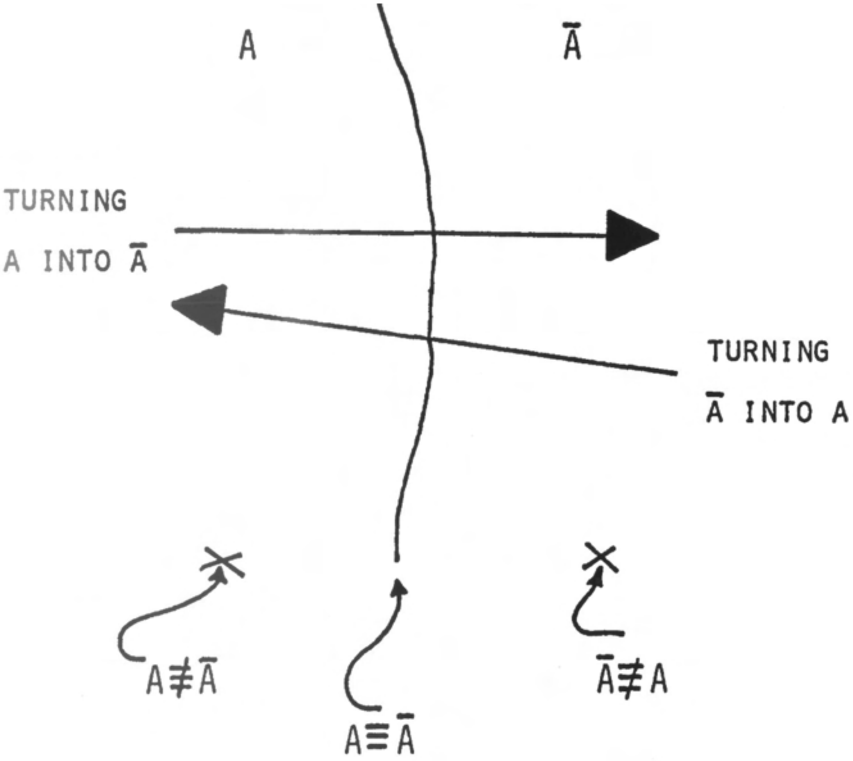
A
$\bar{A}$
TURNING
A INTO $\bar{A}$
TURNING
$\bar{A}$ INTO A
$A \not\equiv \bar{A}$
$A \equiv \bar{A}$
$\bar{A} \not\equiv A$

A POWERFUL EXAMPLE:
CANTOR'S PROOF THAT THE SET OF ALL SETS IS NOT A SET

Fig. 18. Turning a thing into its opposite

THE SKIN TRANSDUCER EFFECT: TURNING A THING INTO ITS OPPOSITE

On the boundary interface, one type of field may be turned into another. The higher psychotronic field (biofields) may be turned into an electromagnetic field, and vice versa. The boundary interface serves as a transfer function mechanism for translating one kind of effect into another. Both ends of the transfer function are excitonically united in and on the interface.

The Soviet work that measured coherence of cyclic rhythms (heartbeat, breath, body electric fields, etc.) in Kulagina is significant in this respect. Kulagina induced electrostatic fields around objects that she moved psychokinetically. Her biofield, which is connected through hyperspace to the object being moved, is orthorotated into ordinary fields at the interface of the object. Thus an electric field is created, and also a momentum field that causes the object to move by ordinary physics.

The Soviet work appears to have established that man, earth, and cosmos are psychotronically and coherently linked; that the boundary of an object is a transducer coupling internal and external dynamic changes in one-to-one correspondence; that the transducer effect applies to both living and nonliving systems; and that through tuning and resonance, internal patterns can be incepted externally, and external patterns can be incepted internally.

EVERETT'S MANY-WORLDS INTERPRETATION OF QUANTUM MECHANICS

Both mind and matter can be precisely modeled in Everett's many-world interpretation of quantum mechanics.

Ordinary relativity is interpreted for only a single observer at a time. If you believe that you and I simply coexist at the same time, then the basis of physics is very different from what is in your physics book. Everett worked out what happens to physics if we insist on multiple observers at a time: all possibilities exist and are real, no matter how weird or far out; the world is a single giant universal wave function; collapse of the state vector has meaning only to a single observer; the world is composed of multiple orthogonal worlds of infinite variety, complexity, and strangeness; each world constantly splits into multiple worlds at a fantastic rate.

The many-worlds interpretation, though strange, is totally consistent with the experimental basis of physics. Most physicists have completely ignored Everett's profound and fundamental work; the MWI is a time bomb ticking away at the heart of physics.

SPACETIME CURVATURE IS TRICKY

Bent and rotated spaces give strange results. A thing in one frame can be something quite different in another frame. Just as the mass of a moving object does not increase with respect to its own inertial frame, it increases with respect to the observer's static frame.

What a thing is, is totally relative to the perceiver and his perceiving frame. It can even be nothing in a great many frames. Do you see how nothing can contain everything?

So we really can turn nothing into something if we are clever.

We can solve the energy crisis by building free energy devices which simply cohere the nonintersecting virtual-mental energy states and collect them, breaching the quantum threshold and yielding real, observable energy. To put it another way, we can tap zero-point energy.

And we can also turn a thought-object into a real, materialized object in this physical frame. Or we can dematerialize a physical object, converting it to a mental object.

MULTIPLE REALITIES (MANY WORLDS)

In Figure 16,1 show the particular selection of dimensions I take from Everett's many-worlds interpretation of quantum mechanics. Other and more complex selections can be taken, but these suffice for our purposes and allow us to complete the basic schema for connecting physics and metaphysics in the simplest possible manner. The selection and the schema developed thus provides an initial crude theory which makes predictions, develops theoretical mechanisms, and can be physically tested.

In the diagram, each three-dimensional physical frame of an infinite set at orthogonal angles to each other is shown by a single line about the plane of the bottom, much like the spokes of a wheel. The single time dimension shared by every one of the spokes or three-dimensional worlds is shown by a single vertical line, much like the axle for the wheel and its spokes.

Starting with the left spoke and going counterclockwise, the spoke worlds are numbered with the prime superscript system. The *S* world is the normal three-space physical world we all live in and detect; the S^1 world is the first orthogonal world, etc. Each world is also called a hyperspace or bioframe or bioworld and numbered consecutively beginning with zero for the *S* frame. Thus the zero-th hyperspace is the ordinary physical universe.

The first hyperspace is the electromagnetic field. A physical object in this frame is seen as an electromagnetic field to the *S* frame observer. Since its intersection with the S frame is two-dimensional, we immediately see that the electromagnetic field is a plane wave. Light thus travels at speed *c* to every observer in the *S* frame, because *c* is or represents an orthogonal rotation.

Further than the first frame, a three-dimensional object is seen by the S-frame observer as a deBroglie wave. Since its rotation is greater than 90° with respect to the S-frame, the deBroglie wave always moves faster than the speed of light with respect to the S-frame observer.

A three-dimensional object in the S^2 frame is seen by the S-frame observer as a line or force vector, i.e., simply as flux. (Tentatively, this frame appears to be the area of neutrinos.)

A three-dimensional object in the S^3 frame is seen by the S-frame observer as a spatial point, i.e., as spatially dimensionless. It is thus a mental object or a virtual object to the *S* observer.

The entire S^3 frame constitutes a mind or a virtual world. There are an infinite number of such S^3 frames available in hyperspace. In fact, there exists a set of three orthogonal turns between any frame higher than S^3 and the S frame, due to the nature of hyperspace. Thus in the model we are using, any of the higher bioframes may be considered to be only three orthogonal rotations away from the *S* frame.

We thus have a scheme of multiple worlds and multiple realities from which we may construct living systems, thoughts, tulpas, etc. And we may also turn one type of field into another, simply by virtual state collection or separation, as the case may be.

All frames S^3 or higher are mental or virtual frames. Physical phenomena in these frames are mental phenomena with respect to the *S* frame.

Note that we have deliberately evoked a system whereby a mind is modeled in terms of physics and physical reality is modeled in terms of mind. The only difference between any two entities in the various worlds is a matter of orthogonal rotation.

For the pure mathematician, obviously this entire model could be extracted into a new calculus. A symbol could be assigned to the orthorotation operation, and a sign assigned to the symbol to indicate whether the direction of turning is approaching the observer or departing from the observer. Sideways turns and in-betweenwards-turns could be represented by additional signs, or hypernumbers. I suspect that such a calculus is well worth doing, and I also suspect that it will closely resemble Charles Muses's hypernumber work when done. Using a six-dimensional geometry, the German physicist Heim has accomplished something similar, without yet realizing that he has also modeled mind. His work generates all the fundamental particles of physics and explains many features of UFO phenomena, as is to be expected from the scheme developed here.

At any rate, an exact calculus of higher dimensions and of turning one kind of field into another kind of field — with no barrier at all between physics and metaphysics — awaits the hands of the builder.

CLUSTERED ORTHOGONAL WORLDS

In this diagram, I show a sample of how various objects appear in their intersections with various bioframes. Note that an object is many things at once, and that many of the things it is are nonthings or nothings.

Everything is a thought in at least one frame. We can appreciate James Jeans's remark that the world appears less and less like a mechanism and more and more like a single thought. His intuition was in fact literally correct.

Each entity is a photon in at least one frame (let there be light!) Each entity is a line in at least one frame. Each entity is a three-dimensional spatial object in at least one frame.

By orthogonal rotation, a thing can change its face with respect to an observer, appearing to change into a totally different sort of thing with respect to a fixed observer in one frame. Reality is quite plastic and absolutely unlimited. It is only our own rigid heads that have been made of solid concrete.

Specifically, materialization of thought objects, by three approaching orthogonal rotations, exists. Similarly, dematerialization of an object, by three receding orthogonal rotations, exists. In such a case the dematerialized object appears the same as a thought object or virtual object.

A perfectly valid mechanism thus exists for tulpoids, materialized thought forms. All that is needed is a sufficient virtual state collection mechanism.

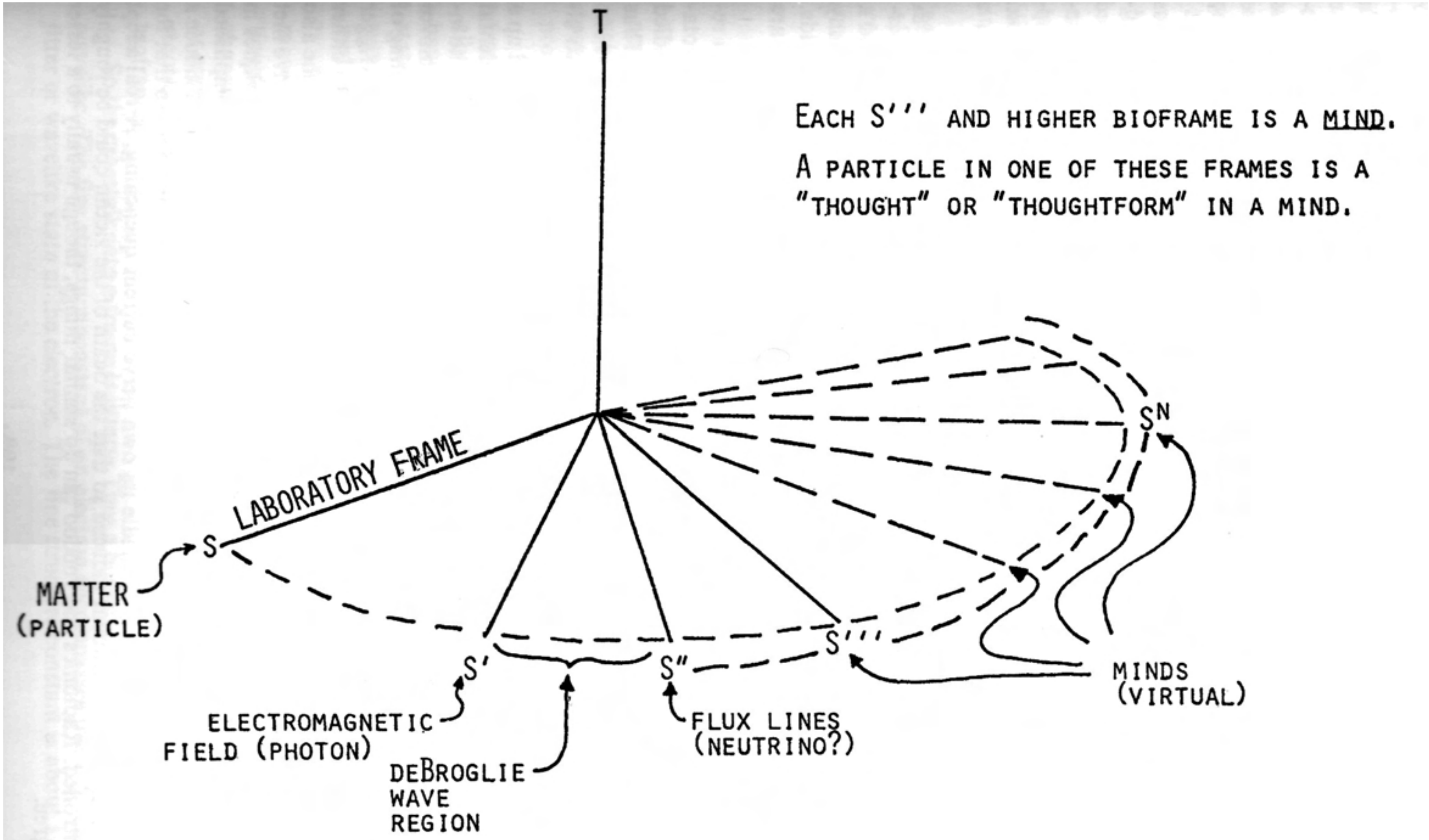
T
EACH S''' AND HIGHER BIOFRAME IS A MIND.
A PARTICLE IN ONE OF THESE FRAMES IS A "THOUGHT" OR "THOUGHTFORM" IN A MIND.
LABORATORY FRAME
S
MATTER (PARTICLE)
S'
ELECTROMAGNETIC FIELD (PHOTON)
DEBROGLIE WAVE REGION
S''
FLUX LINES (NEUTRINO?)
S'''
S^N
MINDS (VIRTUAL)

Fig. 20. Clustered orthogonal worlds

Note that an orthogonal rotation in an approaching manner is automatically accomplished by sufficient coherent collection in the remote frame. There is a finite collection threshold between two orthogonal hyperframes. Sufficient collection in one frame (by a tuning or directing mechanism in a third frame) breaches the threshold in the collection frame and the collected or quantized entity automatically appears in the next frame closer to the tuning mechanism's frame. The process may then be repeated. If sufficiently repeated, all the intervening thresholds are breached and full materialization of the object in the observer's frame (tuning frame) occurs.

By this process an object may be kindled from a remote frame to one that is less remote. Two successive applications of kindling turns thought into electromagnetic field. This is how thought photography is accomplished, and the talented individual who accomplishes it is simply cyborged to the camera film grain in an inceptive cyborg device. Many such examples, including some deliberately accomplished on demand by the study group in which I am included, are given in this book. Three successive kindlings turn a thought object into a materialized object, a tulpoid.

Such a materialized thought-object may be of an inert object or of a living, breathing, biological being. The model we are working in has eliminated any superficial difference between living and nonliving because it is a four-law model. And a materialized living being will be just as real and physical and functional as any other person on this earth. Ultimately, that is all that we ourselves are — tulpoids. Nothing exists but the thought of the infinite creator, so to speak, and we are simply part of that thought.

The coherence in multiple stages necessary to accomplish kindling is often referred to as *tuning* in this book. Tuning sufficient to enable two orthogonal rotations results in thought turning into electromagnetic field or light. Such a kindled field may be additive or subtractive to existing electromagnetic fields. Thus kindling can cancel an electromagnetic field that already exists, or it can create one where none previously existed, or it can shape, modulate, or change one which already exists.

Kindling can also wipe out the charge on a charged particle, or it can build up a charge on an object which previously had no electric charge. It is not accidental that Kulagina, when performing psychokinesis in the laboratory, first causes the induction of an electrostatic field on a nonmoving object. Then the object moves psychokinetically. In a similar fashion, she has stopped a frog's heart; kindling of the type to quench electric charge would of course eliminate the electrical trigger signals in the muscles of the heart, rendering it inoperable. This could mean either cessation and paralysis in a moderate case or total death in a drastic case. She also on one occasion demonstrated the ability to stop a living human heart. I have met other people who could do something similar. One of the great secret techniques in certain Oriental martial arts was the death touch. This could only be accomplished by certain rare adepts; with a light touch of the hand or fingers, they could induce a charge of bioenergy into and onto the acupuncture system of the body, so that the body itself began to slowly kindle its own nerve current quenching. A real adept could control the amount induced so that the death of the victim could be roughly controlled. Radionics and voodoo can do similar things, though usually to a lesser degree.

There also exists a precise mechanism for metal bending; the destruction of the charged particles (for the protons in the nucleus and the electrons in orbit, or even for just the protons in the nucleus) converts the particles in some atoms to neutrinos, which then vanish in the distance without interaction. Thus localized areas in strong metal can be kindled and made quite plastic from thought energy alone by one who has developed this skill. And the kindled areas will give every indication of having been subjected to extreme heat or unusual forces of a sort not encountered in normal metal bending or fracture. Also, the inceptor who is bending the metal and the metal itself are cyborged together into a single inceptive cyborg mechanism. Therefore the metal will often continue to bend even after laid down, because in the cyborg kindling is still occurring.

Furthermore, metals that exhibit unusual physical characteristics, such as nitinol, can change those characteristics dramatically by kindling. Also, an electrostatic field can be directly kindled in and induced on an external object, and an electric field or electromagnetic field can be induced in its surrounding space to interact with it, accomplishing psychokinesis of the object.

Or thought energy can be kindled in the hyperfield fluxes of an ordinary magnet. A magnetic compass needle can be moved by mind or a magnetometer can be affected, even through a superconductive Farraday shield. After all, the effect is rotating into our three-dimensional space from outside it, so shielding is of little consequence except where it interferes with tuning.

CROSSTALK BETWEEN MULTIPLE REALITIES

In Figure 21, I show that phenomena in all the hyperframes crosstalk. A moving object in frame *S* possesses a slight rotation, hence it projects a tiny component into frame S^1. This S^1 component, however, is still dynamic and has a small velocity in S^1. Hence it also is slightly rotated and therefore has a crosstalk component projected into frame S^2 with a velocity there. Having finite velocity, this projection in turn has a rotation and projects into the next higher frame, and so on.

Thus any changing entity in any frame crosstalks a tiny amount into any other frame. In a frame S^n, the crosstalk from a change in any other frame S^P is normally so small as to be virtual, hence nonobservable. Coherent virtual crosstalk, however, integrates and can breach the threshold between virtual and observable state. It is in fact possible to show that what we call a mass in S space is simply the sum total of the crosstalk in and out of a localized region L^3 of S space, the localized region being the mass in question.

Another point to notice is that most fundamental particles in a frame rarely move with constant velocity. Instead, they normally accelerate and decelerate continually. Thus the actual velocity vector of one of these entities is continually changing, usually quite erratically. But for microscopic particles in regular atomic or crystalline formations, the change pattern is quite regular. At any rate, the constant changes on the velocity induce a jitter component on its crosstalk into other bioframes.

So the crosstalk into a frame is jittery. Thus a fundamental particle is jittery or wavelike. The fine structure constant of an electron encompasses this hyperspatial jitter or wavelike state of the electron. The fine structure constant is about 1/137.

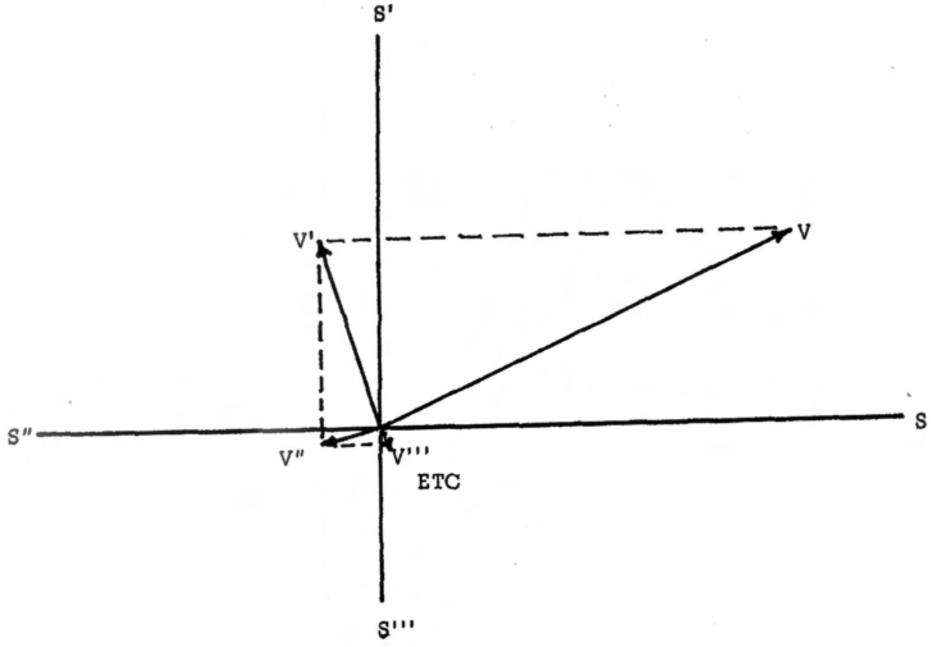

Pig. 21. Crosstalk between multiple realities

Thus an electron actually spends about 1/137 of the time in hyperspace.

So, because of jitter, an object is partially in every hyperframe a minute but finite part of the time. So are all minds and mental objects, and so are all virtual objects. For a fundamental particle, the fraction of time that one quantum of action pervades another frame may be referred to as the identity coefficient.

The identity coefficient varies; it is the photon interaction that time differentiates, quenching time and separating the bioframes, and this interaction is not a linear operator since the intensity of time differentiation is a rather complicated function of photon frequency.

At any rate, for an electron in orbit in an atom (or for any other electron, for that matter) there is thus a small but finite probability that the electron may be at any given position in the universe when its detection occurs. In other words, there is a finite probability of collecting an electron at any point in the universe, and the collection may be made from multiple points simultaneously. We can actually collect the same electron twice, so to speak! Or even a million million times.

The probability of detection of an electron at a given position may be directly expressed as the fraction of time that the electron actually spends there. And this fact may be used to develop a holographic coherent collection scheme that can directly amplify particles. Thus a so-called free energy device, which in one sense is just a tap into the zero point energy of vacuum, can be built. And it has been built, by Moray and by others. The jitter and crosstalk make it directly possible.

BIOFIELDS: AND MAVERICK WORLDS

Figure 22 shows a convenient list of how objects appear in different frames. This schema, selected from Everett's many-worlds interpretation, allows modeling of life, death, mind, matter, and psi. It also allows one type of field to be ortho-

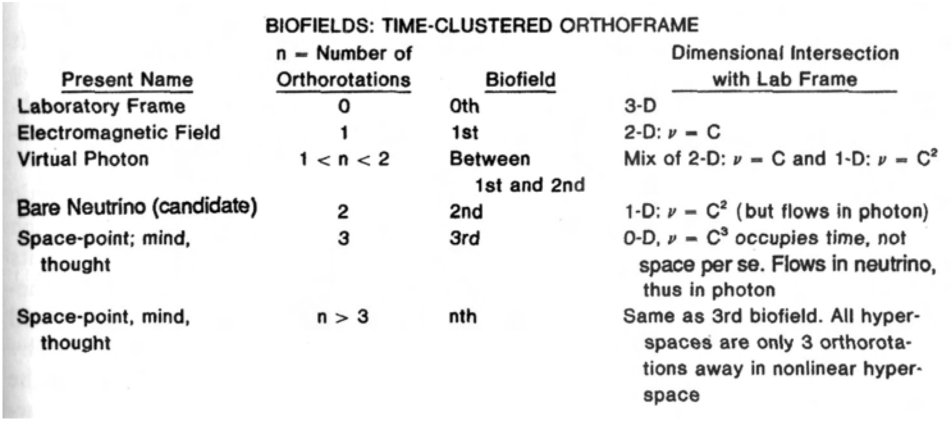

BIOFIELDS: TIME-CLUSTERED ORTHOFRAME

Present Name	n = Number of Orthorotations	Biofield	Dimensional Intersection with Lab Frame
Laboratory Frame	0	0th	3-D
Electromagnetic Field	1	1st	2-D: $\nu = C$
Virtual Photon	$1 < n < 2$	Between 1st and 2nd	Mix of 2-D: $\nu = C$ and 1-D: $\nu = C^2$
Bare Neutrino (candidate)	2	2nd	1-D: $\nu = C^2$ (but flows in photon)
Space-point; mind, thought	3	3rd	0-D, $\nu = C^3$ occupies time, not space per se. Flows in neutrino, thus in photon
Space-point, mind, thought	$n > 3$	nth	Same as 3rd biofield. All hyperspaces are only 3 orthorotations away in nonlinear hyperspace

Fig. 22. Biofields: time-clustered orthoframes

rotated and turned into another kind. Such orthorotation demands correct time synchronization, rather than brute force energy application. In fact, orthorotation is energy-conservative for a single orthogonal turn.

Even the strangest conceptions and most fantastic possibilities have their own assigned realm of concrete reality. These are referred to as maverick worlds.

Note that all possibilities — everything that has happened or could have happened, and everything that will happen or could happen — are real and exist concretely in their separate orthogonal universes. The same is true for all thoughts and thought worlds, and all possible thoughts and thought worlds.

By proper coherent tuning, an *object* in one of these orthoframes can be orthorotated into our own frame and objectified, and that is what the Tibetan monks call a tulpa, an objective materialization of a thought form. The tulpa will hardly ever be closed entirely in phase, however, and so it will almost always be unstable. UFO's, angels, and imps go away. But the materialization of one of these can be entirely objective and perfectly objective traces can be left, such as photographs, broken limbs, scorched asphalt, indentations, and depressions in grassy fields. Any thought object can be so orthorotated and objectified. Beings, religious figures, angels, fairies, imps, UFO's, and monsters can all appear.

HYPERFIELD FLUX IN A MAGNETIC FIELD

In the conventional experiment to determine the pattern of a magnetic field, iron filings are sprinkled about a bar magnet on a piece of paper. As shown in the top drawing of Figure 23, the iron filings form a symmetrical pattern. Further, the only apparent circulation pattern is from one pole to the other. Each flux line is closed, going through the magnet from one pole to the other and out of that other pole, bending back through space to the first pole. Every flux line therefore goes through both poles.

However, each little iron filing is actually quite static; it does not move at all. It does not change with respect to either time or position. It is therefore integrating any higher sinusoidal component that is changing with respect to time and length. The characteristics of the orthodox picture that emerges are prescribed by Maxwell's equations of electromagnetic field.

If we tape the bar magnet to the face of a color television tube, the magnetic field is now being scanned or probed or sampled by a moving electron beam, which reacts with the magnetic field of the bar magnet. This electron beam is changing its position with respect to time and length, and it therefore differentiates the magnetic field. The hyperspatial "ripples" in the magnetic field now are not integrated exactly to cancel, and so a small deviation from each flux line will be detected by the differentiating beam. Slight deviations will cause a color change, and so the hyperspatial flux now becomes evident on the television screen. The lower drawings show the rather unusual, nonsymmetrical pattern that emerges.

The equations for these hyperfield fluxes should be obtainable by differentiating Maxwell's equations twice more, yielding fourth order equations. These equations should be the much sought-after equations of prana, chi, ki, ectoplasm, bioenergy, bioplasma, odic force, etc.

Note that the hyperfield circulation goes into the north pole and pours out of the south pole. A living organism exposed to the north pole, end on, will have its own bioenergy sucked out by the magnet, and its own flux will diminish somewhat. Exposed to the south pole, end on, it will have its own bioenergy increased by the bioflux pouring from the magnet. This explains why many experimenters who have extensively tested the influence of magnetism on growing plants have found that the north pole inhibits and the south pole stimulates growth and health.

The hyperfields are fundamental fields rather than electromagnetic fields; they collect coherently and kindle *into* electromagnetic field, creating it or destroying it.

LIFE AND DEATH

In Figure 24 I present an overall schema for life and death — and for reincarnation as well.

An infinite number of hyperfields or bioframes are shown as S, S^1,S^2, S^3 etc. The frame S is the ordinary laboratory frame — the physical spatial world we live in and observe. The frame S^1 is the electromagnetic field, the frame S^2 the linear or flux frame, and S^3 the first "mental" or "virtual" frame. Also, all higher bioframes have a connective path to the *S* frame of only three orthogonal rotations (not shown). Thus only three orthogonal turns separate S^n and *S*, S^K and *S*, and S^J and *S*.

In frame *S*, note that physical body *A* has a coherent crosstalk link with the entire frame S^n. A-S^n thus constitutes a living biological system; in this case, let us say, a human being. Likewise, *B* and S^j have established a coherent crosstalk loop. B-S^j thus also constitutes a living biological system; in this case, let us say, an opossum. These two examples are incarnated living systems (incarnated minds).

In fact, this is a great oversimplification, as the reader will realize at once. We really should consider all the groupings of *A* — the organs, the structures, the cells, the internal cellular parts, etc. — as separate but coordinated a_i. (And similarly with B.) Each of these a_i must have a correspondent s^n subworld of S^n, much like

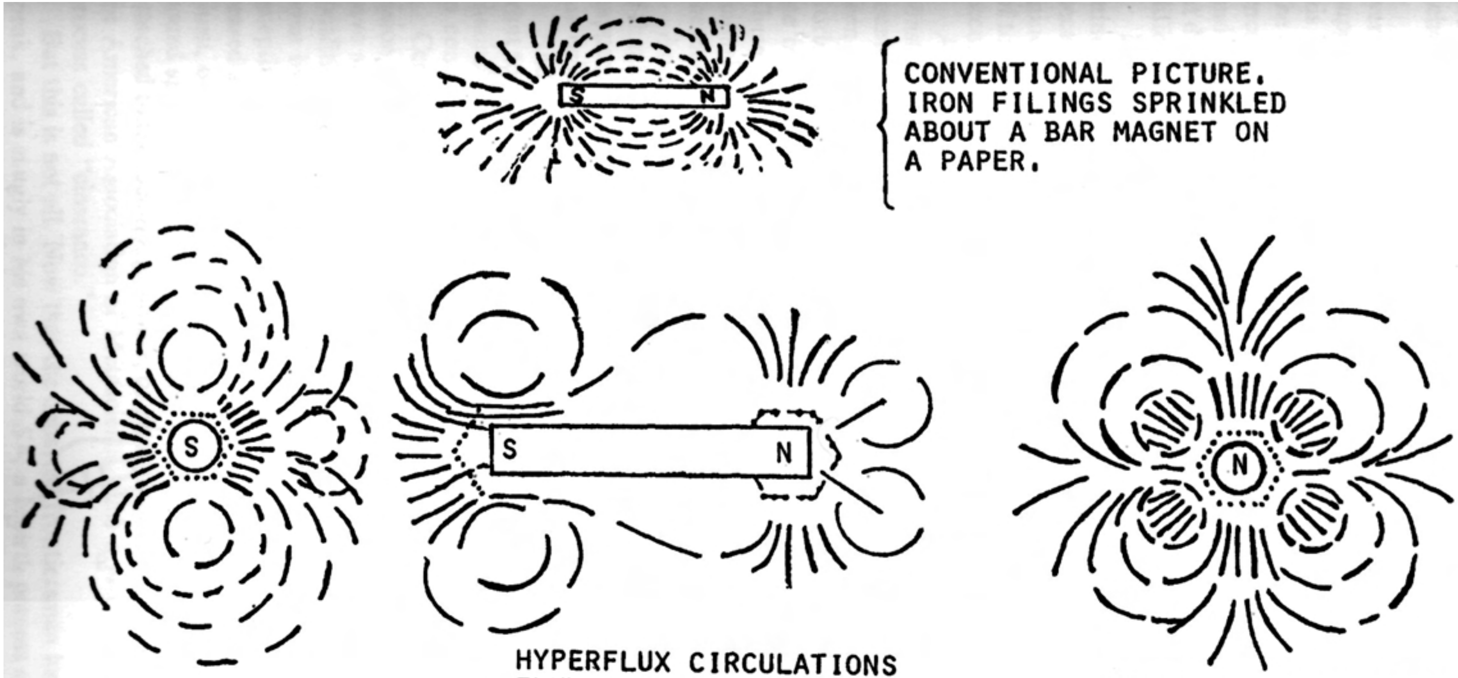

CONVENTIONAL PICTURE.
IRON FILINGS SPRINKLED
ABOUT A BAR MAGNET ON
A PAPER.

HYPERFLUX
CIRCULATIONS

NOTE STRONG
HEXAGONAL PATTERN
AND TWO PRIMARY
VORTICES

HYPERFLUX CIRCULATIONS
FLUX PATTERNS ARE NOT
SYMMETRICAL, POLE-TO-POLE

NOTE STRONG HEXAGONAL
PATTERN IMMEDIATELY
ADJACENT TO NORTH POLE.

HYPERFLUX CIRCULATIONS

NOTE STRONG
HEXAGONAL PATTERN
AND FOUR PRIMARY
VORTICES

Fig. 23. Hyperfield flux in a magnetic field

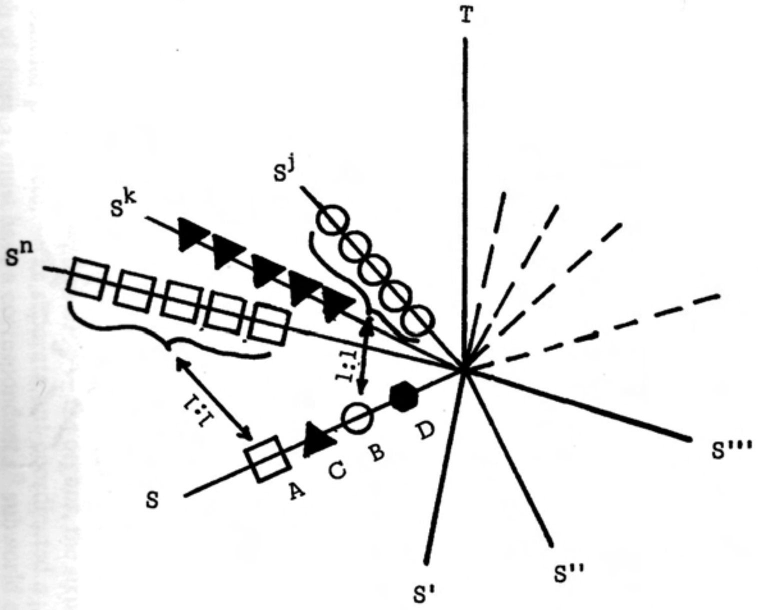

A ⇄ S^n is a living system.

A is the physical body, the incepted mass.

S^n is the mind.

B ⇄ S^j is also a living system.

To S, D is inert; i.e., D is not preferentially incepted.

To S, S^k is a "disembodied mind." I.e., S^k is "dead," since it is not preferentially incepting its former body, C.

Whenever a proper retuning body, say E, appears in S, S^k will reestablish a coherent crosstalk with E. This constitutes "reincarnation" of S^k.

Fig. 24. Life and death

the fine structure of atomic energy levels. But here I am simply outlining the major schema.

At any rate, let us now examine body *C* and hyperworld S^k, which formerly were one-to-one coherent in crosstalk, but are no longer so. C represents, let us suppose, the body of a frontiersman in 1830 who has received an Indian arrow in his breast, piercing the heart and destroying the ability of the body to coordinate the c_i to each other and to the inceptive channel from S^k. With the loss of coherence, noise begins to destroy the channel; thus S^k is now a "disembodied mind," and C starts to become a decaying corpse. The various c_i of course lose coherence at differing rates, and so the fine structure channels to the s^k also lose coherence at differing rates.

Finally all c_i - s^k coherence is lost, and the frontiersman is then "dead." The mind S^k is now disembodied — and also probably quite disturbed or "noisy" because of the variance in disintegration of the c_i - s^k loops. One might say that the mind is somewhat "detuned" by the whole affair. However, "death," in the sense of total loss of mind and personality, has most certainly not occurred. All that has occurred is a lapse into a disembodied, and perhaps somewhat confused, state.

And the disembodied mind finds itself in its own deep, unconscious world constructs. If a person has been unconsciously Christian in a very deep, fundamentalist sense, and in his own mind has also been a "sinner deserving of hell," then he will actually "lift up his eyes in hell," in his own world tuning. And that world will be physical and real — as real and physical as this one — but it will operate by different physical laws, just as dreams operate by different physical laws. These relations and laws are prescribed by the person, either consciously or unconsciously. The devout Mohammedan is quite likely to find himself after "death" in the delightfully chauvinistic paradise he believes in.

And so the Catholic practice of praying oneself out of purgatory is relevant, for if that belief-structure is firmly fixed in the deceased's deep unconscious, then it will happen; he will simply change his world construct in accord with the effort made on his behalf.

And the ancient Tibetan practice of giving instructions to the deceased as to what to expect and where to go is also relevant, for during the time that the major channel $C\text{-}S^k$ is lost but many of the c_i - s^k are still functioning, the deceased is still in contact with frame *S* and can heed those instructions.

Certain clairvoyants and skilled psychics can also project part of their mind vision into S^k and see the status of the deceased. In some cases the deceased will have so strongly retained mental structures analogous to physical structures and localities (or persons) in frame S that continual small crosstalk channels breach open by tuning resonance to those localities, structures, or persons. With a little inceptive assistance from one or more sensitive persons, an "appearance" of the deceased at the locality, structure, or near the person can take place. Much psychic, ghost, or spiritualist activity is due to this effect. So indeed is the concept of earthbound spirits. Here again, an advanced adept can erase the psychometric pattern attached to the locality, structure, or person and thus eliminate the phenomena. In the American Association of Metascience, such a resolution is part of an overall process called "clearance."

But this is not all. Now that the departed frontiersman has been cleared, so to speak, and is snugly in his own world S^k rebirth process still awaits him. As I

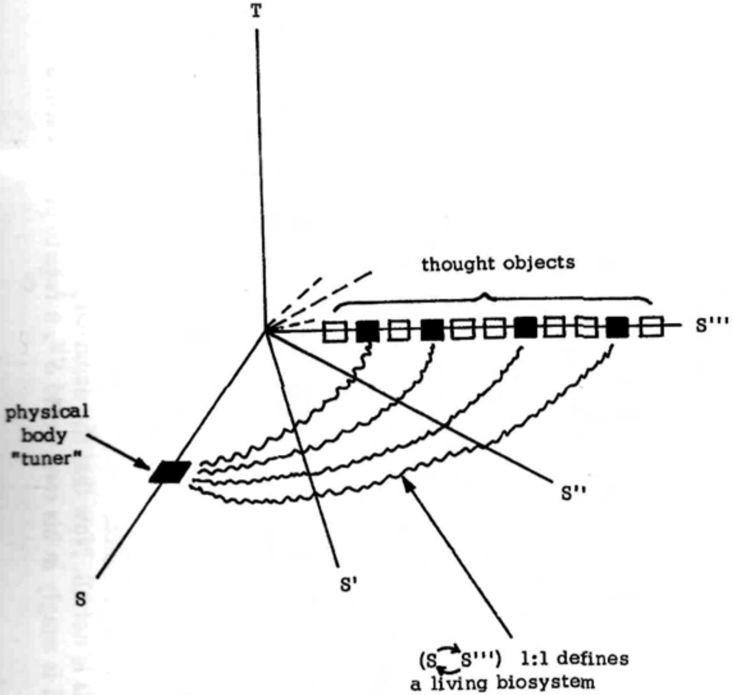
T
thought objects
S'''
physical body "tuner"
S''
S
S'
(S S''') 1:1 defines a living biosystem

A. A LIVING BIOSYSTEM

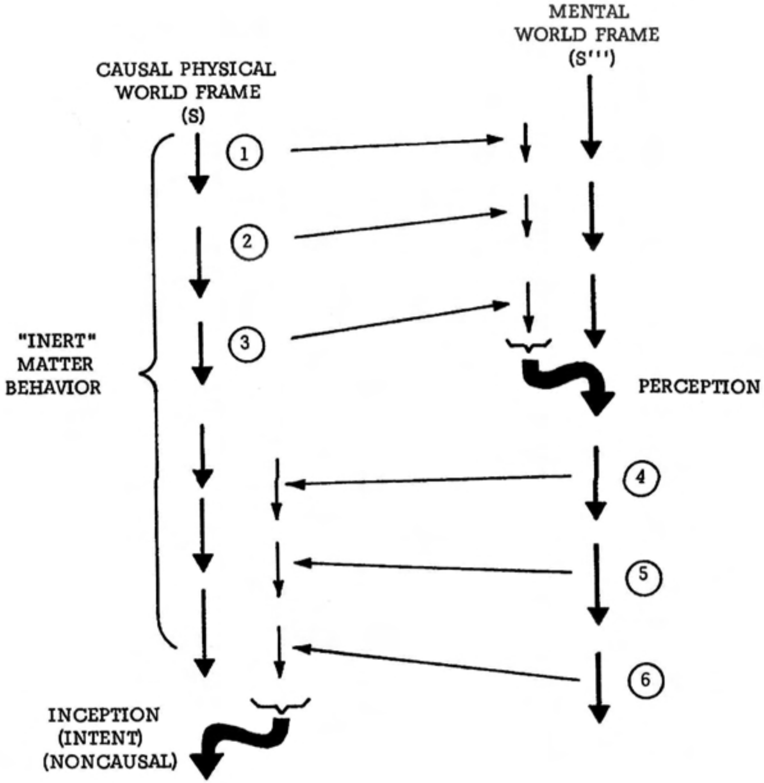
MENTAL WORLD FRAME (S''')
CAUSAL PHYSICAL WORLD FRAME (S)
①
②
③
"INERT" MATTER BEHAVIOR
PERCEPTION
④
⑤
⑥
INCEPTION (INTENT) (NONCAUSAL)

B. THE FUNCTIONING OF TIME AND CONSCIOUSNESS ARE MUTUAL CREATORS OF EACH OTHER.

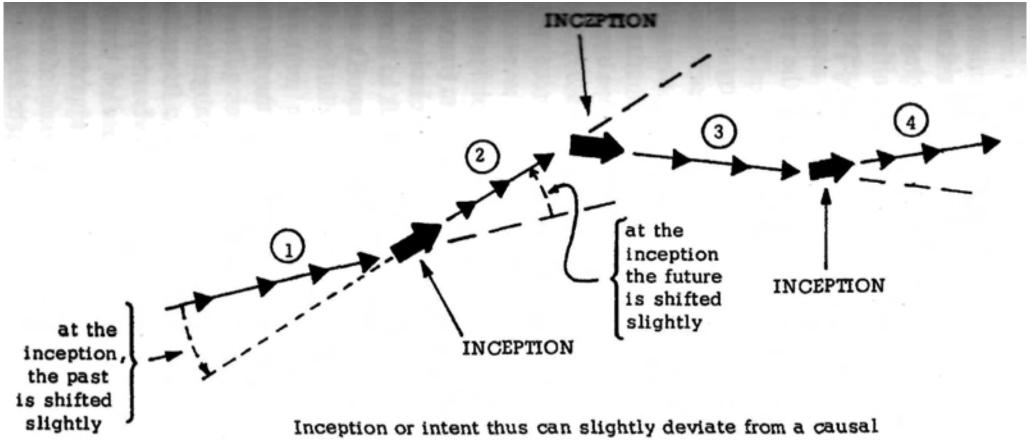
INCEPTION
③
④
②
at the inception the future is shifted slightly
①
INCEPTION
INCEPTION
at the inception, the past is shifted slightly

Inception or intent thus can slightly deviate from a causal trajectory. Over sufficient collections of physical changes, "intentional behavior" or "limited free will" emerges. However, it is only a very small modulation of causal inertness -- and its smallness accounts for human "physical limitations."

C. CHANGING THE PAST

Fig. 25. Consciousness, life, and intent

have pointed out elsewhere, every particle of mass is simply a pattern of change in the universal orthorotational flux between hyperframes. (This, after all, is no more than a slight extension to general relativity.) When the proper physical tuning mechanism, say C^1, again appears in the S frame, then a resonant coupling will automatically establish itself in the flux between C^1 and S^k, and the disembodied, confused mind of the frontiersman will simply be coupled again to a new physical body pattern. This happens at the instant of conception, for the genetic pattern — the complete tuning pattern — is present at that instant. The coupling is quite tenuous at first, and increases in degree as the embryo develops.

So reincarnation has a place in my schema. We will see, however, that it differs slightly from the classical concept, because with developmental work the frontiersman can change the makeup of S^k and thus his predisposed tuning. Changing the tuning also causes rotational displacement of S^k among the infinity of infinities of mind worlds available. Thus the past can be changed; in fact a person can change who he is the reincarnation of, simply by changing his mind tuning with enough development. Therefore more than one living person can be, for example, the reincarnation of Napoleon.

And in fact the cycle of rebirth can be broken completely — simply by dying with a totally detached mental framework. The cessation of all tuning attachment results in a broadened S^k, which now is structureless, hence all-structured and part of universal Godness.

But the major point is that all this can be modeled physically. And there is only life; absence from one place still leaves presence in another. Absence from *any* and *every* one place exclusively leaves presence at all places, each nonexclusively.

Death, where is thy sting?

In diagram A of Figure 25, I summarize what constitutes a living biosystem. Here we can see that body (physical) changes from the biosystem's body in frame S crosstalk successively through frame S^1 and S^2 to S^3. Also, physical changes in frame S^3 actually constitute mental changes to the biosystem's body back in frame S. These mental changes must crosstalk successively into frame S^2, frame S^1, and then into frame S. This entire process — from frame S^3 to S^2 to S^l to S — is referred to as *kindling.*

In diagram B, we observe this process a little more closely. Here, the shorter straight vertical arrows are virtual state changes kindled from the opposing frame. The longer straight vertical arrows represent observable quantized change in the particular frame, these changes being produced by ordinary inert physics laws. The curved double-lined arrows represent quantized change kindled into the frame from the coherently superposed virtual changes that have steadily been kindled there.

Thus there is a kindling time delay in the causal S frame before will or intent sets of changes from the mind-frame S^3 can induce an input into the physical body of the biosystem. Correspondingly, there is a kindling time delay before a perceptual change in S^3 from the changes in S can occur.

The horizontal arrows represent the flow or flux of kindling crosstalk between the two frames. It can be seen from this diagram that consciousness — the thread of the relation between consecutive changes in the two respective frames — and time are identical. (Recall that each inert physical change in S that is not kindled from S^3 is actually randomly kindled from all the other mind-frames.) Thus time and mind are identical. Consciousness is simply the functioning of coherent crosstalk between a mind frame and a physical body in frame *S*. Further, each biosystem's consciousness has all other consciousnesses interwoven with it. The other living consciousnesses input into other physical bodies in frame S, and these physical bodies interact in frame *S*. The other dead consciousnesses input into frame 5 - into each body as well as into all other inert matter — in a noncoherent, randomized manner. Thus the entire universe of universes is totally in communication with each part of itself, continually.

In diagram C, I show that incepting a change from pure causality changes the past; in this case, the past is causally inferred or projected backward from the new vectorial direction. However, because inception into frame *S* is a very small effect, the biosystem's change of its past is so slight that normally it is not noticed. Inception of course also changes the future slightly.

THE HOLOGRAPHIC HYPERCHANNEL EFFECT

This effect is based on the concept that a human being must accomplish a minute inception (psychokinesis) onto his physical system. Specifically, he must incept the movement of three or four electrons in the brain. This constitutes an input signal to his ordinary physical system, which is then treated like a complex servomechanism. The gain of this servomechanism is rather high, being able to transform or amplify the movement of several electrons into powerful movements of the entire body. The inceptive effect cannot be explained by ordinary physics, which am account for the servomechanism gain only. Once the input signal is generated, however, the rest of the amplification and action are viewed as being properly described by the ordinary physics of inert matter. Thus the human is viewed as largely servomechanistic and mechanical, but guided and directed by the nonphysical, which affects the physical in a minute fashion sufficient to constitute inputs to the ordinary physical system.

Our purpose here is to show that, given this single small inceptive effect, we can eventually approach the mastery of space and time for a linkage of sufficient number of human brains and minds into one single superbrain and supermind.

From the multiplicity of orthogonal frames permitted by Everett's many-worlds interpretation of quantum mechanics, we now select an infinite set of clustered three-dimensional spatial frames, which all share a single time dimension. Using one frame as a laboratory reference frame, the other spatial frames may be characterized as having one, two, three, or more orthogonal turns away from the laboratory frame. These succeeding orthogonal frames are variously called hyperframes, bioframes, orthoframes, hyperchannels, orthoworlds, etc. A particular set of material phenomena in one of the orthoframes is referred to as a biofield, hyperfields, or orthofield with respect to the laboratory frame.

In Figure 26 I represent each brain by a triangular symbol, and consider it as an amplifying device having a gain of A_o. For simplicity the inceptive input signal is represented as E_i, since in moving an electron, a voltage must be induced. For limplicity I also show the brains as a linear array lying in a straight line. The distance L separates each brain, and here L is assumed constant for simplicity.

Two factors must now be realized: (1) The first hyperchannel operates in the realm of the ordinary electromagnetic field, although this realm is seen in a differ-

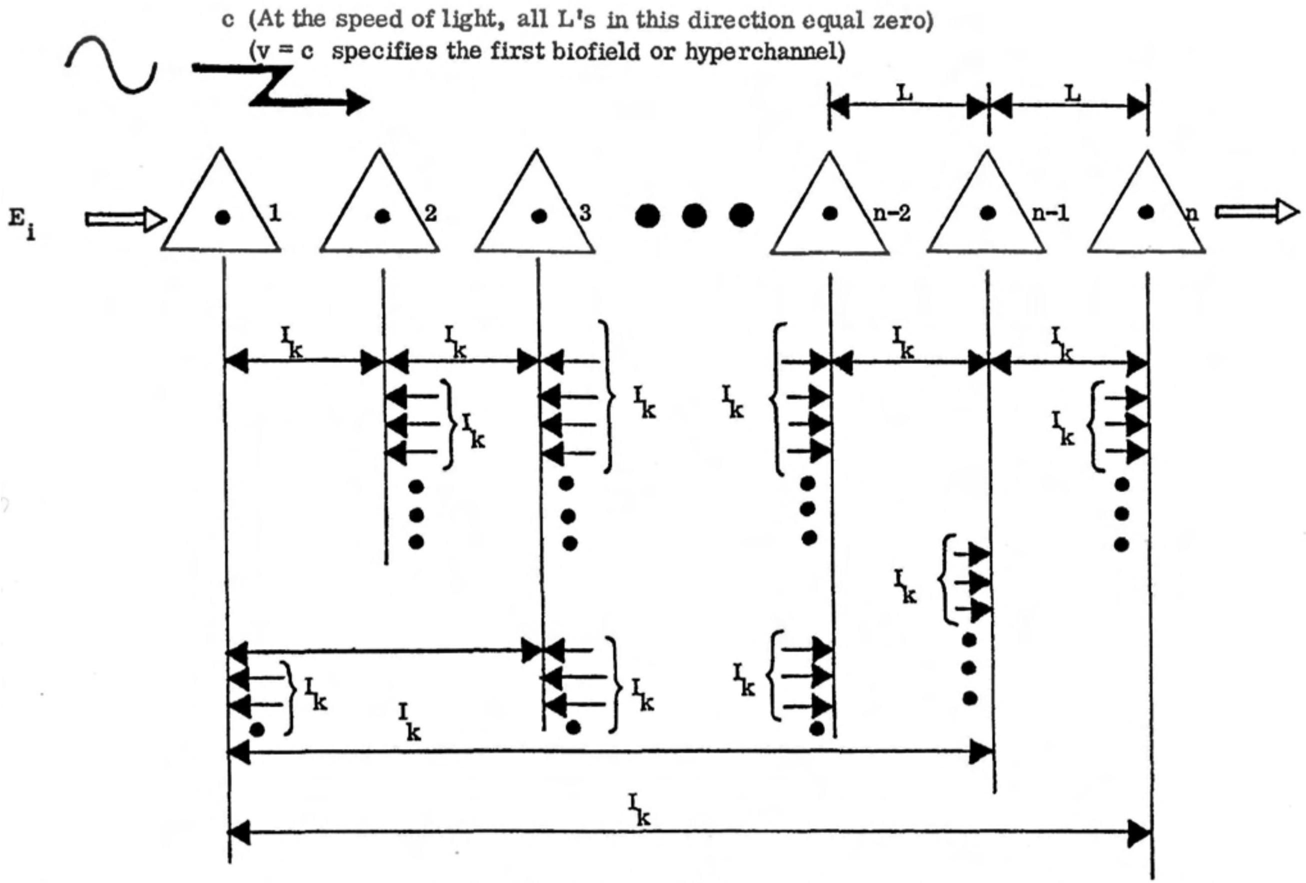
c (At the speed of light, all L's in this direction equal zero)
(v = c specifies the first biofield or hyperchannel)
L
L
E_i
1
2
3
n-2
n-1
n
I_k

Fig. 26. The holographic hyperchannel effect

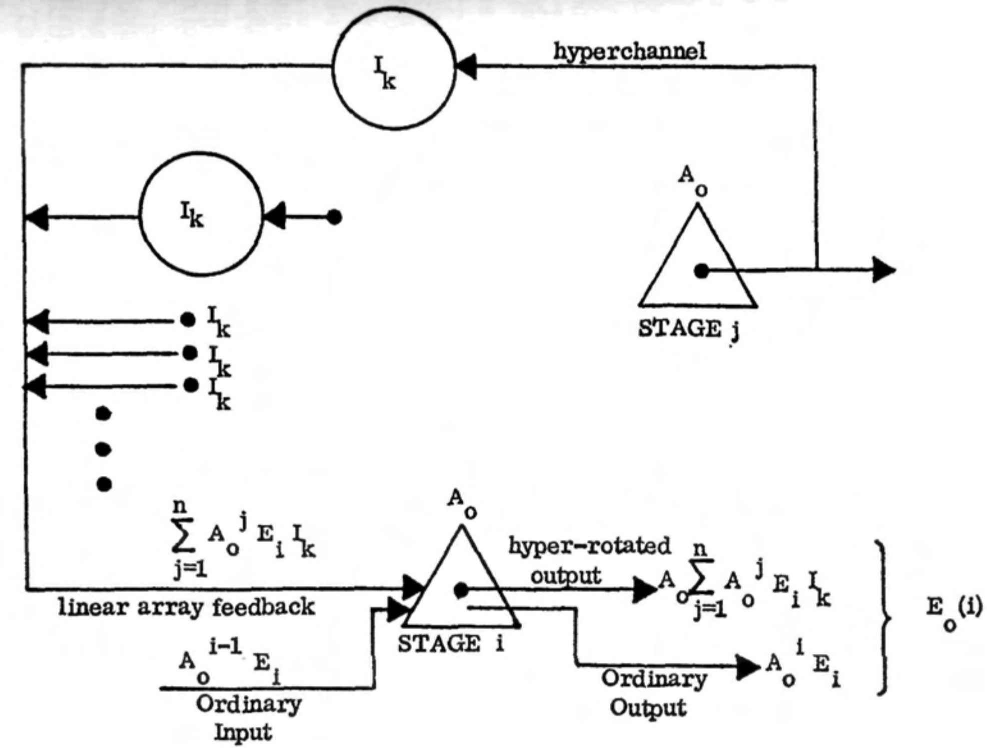
hyperchannel
I_k
I_k
A_o
STAGE j
I_k
I_k
I_k
$\sum_{j=1}^{n} A_o^j E_i I_k$
linear array feedback
A_o
hyper-rotated output
$A_o \sum_{j=1}^{n} A_o^j E_i I_k$
$E_o(i)$
STAGE i
$A_o^{i-1} E_i$
Ordinary Input
Ordinary Output
$A_o^i E_i$

Fig. 27. Output of stage i

ent way, i.e., we visualize a photon as an ordinary particle of three-dimensional matter, resting in a three-dimensional spatial frame (hyperframe) that is orthogonally rotated spatially in higher dimensional space, at 90° to the observer's spatial frame. The hyperframe and the laboratory frame share the same fourth dimension, time. (2) At the speed *c* (which determines the hyperframe in the first place), all *L's* are zeroed out, by the equations of the special relativity transformations. Thus one may view (hyperview) each triangle (brain) as superposed on each other triangle (brain). Furthermore, there exists a very small, finite, nonzero probability that the state of one triangle coincides exactly with the state of another superposed triangle, at any given small increment of time. This probability corresponds to a direct fraction of identity between two triangles, since time is stopped along the chain. At any given small increment of laboratory time, any given triangle has on the average a small fraction of identical coexistence with any other triangle. This identity fraction is labeled I_k in Figure 26.

It is important to note that, for a very small fraction of time, the triangles coincide with each other; i.e.; for a small fraction of time the existence of one triangle is totally holographic. This fraction totally violates the ordinary conservation of energy, which is applicable only to a single channel system. The hyperchannel represents a second channel, an identity channel, and in this channel an object can be in multiple places at the same time. If the triangles, regarded as operating processes, are even slightly coherent, then multichannel communication and holographic transmission of a single state-energy into multiple state-energies occurs through the hyperchannel — usually in the virtual state.

Figure 27 shows the setup to calculate the gain of the hyperchannel amplifier. These calculations yield, for the output of stage i,

$$\frac{E_0}{E_i} = A_0{}^{n} + A_0{}^{n+1} I_k.$$

When no ordinary electrical input exists, $A_0{}^{n} = 0$, and the virtual state amplification is thus

$$\frac{E_0}{E_i} = A_0{}^{n+1} I_k.$$

This mechanism holds the secret to the tapping of zero-point energy and free energy devices. It is also the hidden variable theory that causes a radioactive nucleus to decay. Since radioactive nuclei interact through hyperspace, the amount of radioactivity in a given sample is an exponential function of the number of radioactive nuclei present.

It is also a mechanism which can directly amplify the virtual-mental state into observable state. It is the fundamental mechanism that causes kindling and thus orthorotation. It can be used to materialize tulpas, and accomplish psychokinesis, remote viewing, clairvoyance, precognition, etc.

MIND LINKAGE

The entire human species has been developing and evolving, not just the individual person.

The development of life in a biosystem may be divided into seven stages: (1) the development of planet, primordial atmosphere and ocean; (2) the appearance of

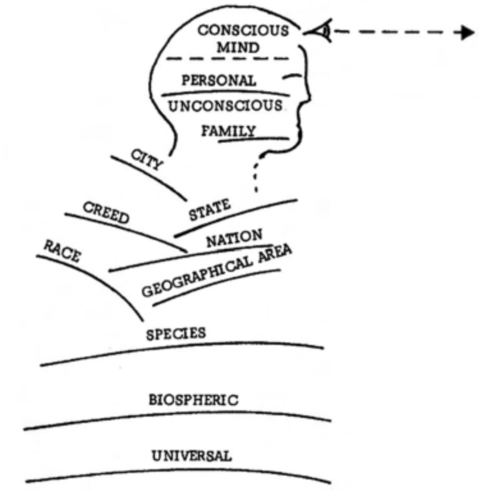

Fig. 28. Layers of unconsciousness

amino acids; (3) the appearance of the self-replicating supermolecules; (4) the formation of one-celled organisms; (5) the formation of multicellular organisms; (6) brain linkage of an intelligent technological species; (7) total omeganess or godhood.

Two major types of behavior are involved in living organisms: genetically programmed behavior and genetically unprogrammed behavior. Thus one mode of behavior control of an organism is nonadaptive and the other mode is adaptive or learned. Intelligence may be viewed as the ability to adapt behavior, and increased intelligence represents a decrease in the portion of behavioral control that is genetically programmed or "hard-wired."

We may further speak of two types of intelligent species: those that possess a tool-using member and those that do not. The first is a *technological* species, while the second is *nontechnological.* By this classification, man is a fifth-stage, technological species.

The intelligent species, being relatively deprogrammed, is highly adaptive and competitive, and many of its species-specific checks and balances are reduced or eliminated. Because of the use of tools in its internal competition, the technological species continually increases the ferocity of its internal competition and its deadliness to itself. Thus as the power of its tools increases, the technological species carries the seeds of its own destruction within itself, and it will eventually destroy itself unless it advances to the next stage, the linked-mind sixth-stage being.

The sixth stage consists of the linkage of all individual brains in the species into one single superbrain, producing a superbeing having absolute control of space and time. A technological species is necessary to accomplish linkage and produce a multidimensional, eventually nonphysical, superbeing. When the technology is sufficiently advanced for linkage to be developed, the power of the tools is so great that self-destruction is imminent. Thus an advanced technological civilization is near linkage and also near Armageddon, and in either case it vanishes as a limited, three-dimensional species of the fifth stage. This accounts for the notable absence of a technological species similar to man in man's universe.

ANGELS, IMPS, AND UFO TULPAS

If we realize that each of us has his own mind world, an S^3 frame with respect to the *S* frame physical world, but separate from each other *S* frame mind world, then we see that, if we take common forms that are in the minds of all human beings, it is much easier to kindle them because they can be cohered much more easily.

The common crosstalk in all minds constitutes the collective unconscious. Since the common crosstalk is in a sense a mind, and since it is connected to a "body" of many cells, i.e., of four billion humans or so, then this collective unconscious is a living, breathing entity.

Furthermore, the usual content of its mind is a great deal closer to physical phenomena that are in all minds than is a single human being's mental content. Much of what we term physical reality is in fact prescribed by the collective human unconscious. And if the collective human unconscious is changed in its conception of reality, then four billion stages start kindling with the hyperchannel effect, and immediately the new conception is physical reality.

Borrowing from Jungian psychology, let us call a form that is in each human mind an archetype. Let us also realize that there are shallower levels of unconsciousness (Figure 28) between the total collective unconscious and an individual mind. For an Irish-American there is one level defined as Caucasian, another defined as Irish, another defined as American, another defined by the larger region he lives in (say the state), another defined by his family, and another by his own personal unconscious. The levels are not mutually exclusive, but intersect across one another. At any rate, all these unconsciousness levels above the collective human species unconscious also shape, modulate, mold, and change the kindling channel that exists for a person or group. All these levels of unconsciousness are like finer and finer tuning levels to change the kindling mechanism.

An archetype that is kindled from the collective unconscious is kindled through such a mechanism of unconscious levels of tuning, and the reality format that is kindled is determined by the type of pressure and the content, biases, or tuning of all the levels between it and the conscious minds of the individuals who observe the phenomena.

All tulpoidal manifestations are kindled through and by this mechanism. All of them are shaped by the multilevel unconscious tuning. All of them are caused by an unresolved conflict in the deeper unconscious layers.

Since most minds are notoriously unstable, the tuning is unstable almost always, and the tulpoidal manifestation is usually fairly shortlived, usually not more than a few minutes, or a few hours.

Angels, imps, Virgin Marys, UFO's, sasquatches, Loch Ness monsters, lake monsters, sea monsters, nocturnal mystery lights, fairies, elves, devas, devils, ogres, Mothman, men in black, goblins, large monster birds, etc. are all kindled by the same process. And so are cattle mutilations. I fear that human mutilations will be next.

Any tulpoidal materialization may be physically real during its stable period. It may make large tracks in the ground, operate shiplike space vehicles, break tree branches, take human specimens on board, examine human specimens, etc.

The tulpoid may appear benevolent or harmful, and it may actually be so when in the stably materialized mode. The full content of the human unconscious is there to be materialized in physical form — all the good, all the bad, all the noble, all the sly aspects.

When enough human unconsciousnesses have the idea of fairies deeply ingrained in them, occasionally a real, living, breathing fairy pops out of the collective unconscious, riding as a direct modulation on a more primitive overall human species archetype. And the fairy may do all the things expected; she may pose riddles, give gold coins which must not be examined or they will be lost, etc.

Religion can also make use of tulpoidal appearance. When Christian predecessors believed in a vengeful god, a powerful but spoiled king, they got all the manifestations of that kind of Being. When they believed in a god of sacrifice, they got a god who demanded sacrifice and thrived on it. And when many believe in angels and Virgin Marys, they sometimes actually appear. These appearances are not hallucinations; they are completely real. It is just that all possibilities are concretely and physically real in their own realms, and any of these may be tuned in to appear in this realm.

An infinite number of reality channels exists. The total instrumental characteristics, genetic, psychological, and physiological, of the observers who tune in the channel determines the noise content and the actual channel selected.

A reality is simply an orthogonal frame or world. All possible realities already exist. A channel is a set of tuned and timed orthorotations communicating back and forth between two orthogonal frames or worlds. Tuning is a complex of mind and matter, and of many unconscious levels.

For a perfect adaptation of this mechanism of coherence, any coherent signal — no matter how fantastically weak — can be raised up out of the grass and made dominant. By time coherence in the virtual state, a virtual state entity can be amplified into observable state. From sheer nothing we can produce something. And since every possible something exists in nothing and is available for cohering, we can get anything at all by this process, at least hypothetically.

DREAMS ARE UNRESOLVED CONFLICTS

Everything is essentially mind. Mind forms layer after layer of overlapping zones of shallower and shallower unconsciousness (multiple simultaneous consciousness), all the way from the universal allness to the personal unconscious, and culminates, like the tip of a dagger, in each conscious mind. Figure 29 illustrates these mindlevels.

Just as the dreams of an individual show his unresolved conflicts, so the materialized tulpoids of the collective human unconscious show its unresolved conflicts. Tulpoidal phenomena thus symbolize man against himself, or one group of humans against another group of humans. The great unresolved conflict in the human species since World War II has been the Cold War. And therefore the great flaps of tulpoidal UFO activity should correlate directly to the Cold War. And so they do, as I indicate in this book. Literally we ourselves are our own "Satan," for in man's inhumanity to man, satanic evil is created inceptively. And we ourselves are our own "godform," for in man's love of his fellow men, goodness is created

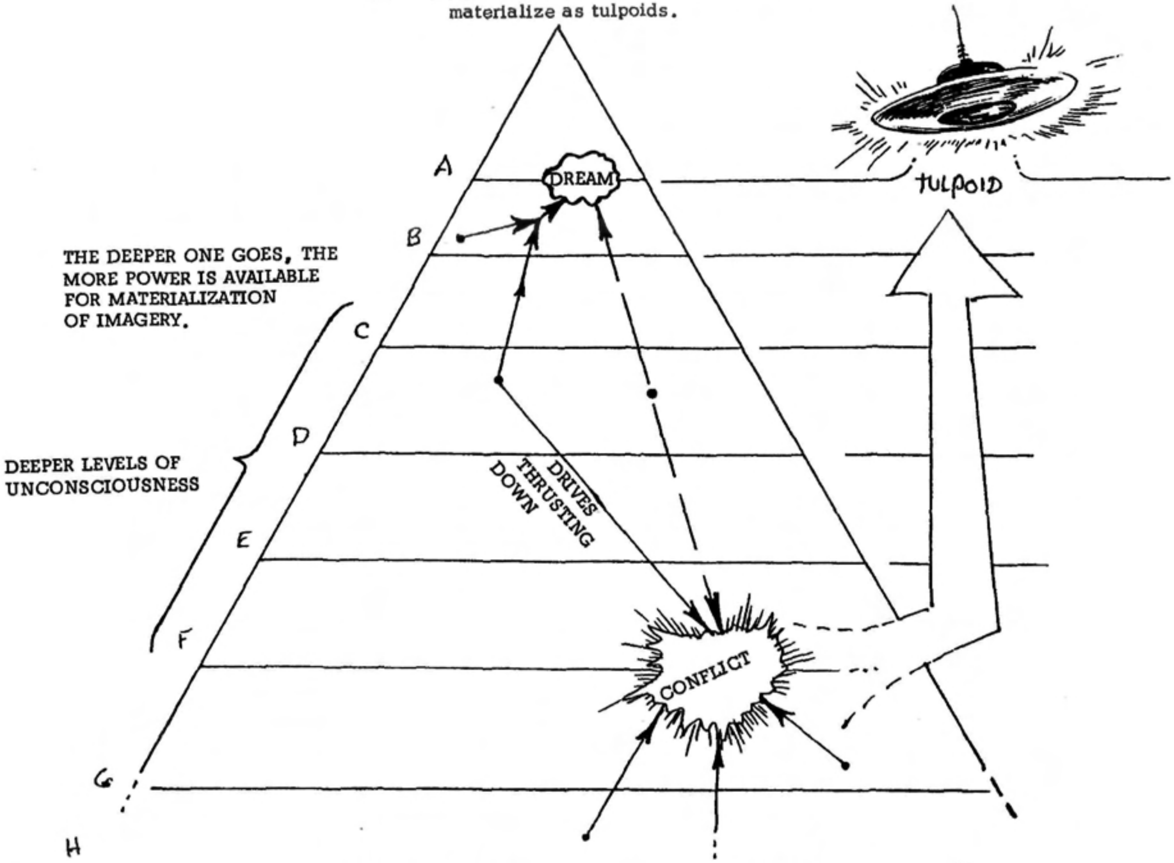

Such dreams of the collective unconscious can materialize as tulpoids.

THE DEEPER ONE GOES, THE MORE POWER IS AVAILABLE FOR MATERIALIZATION OF IMAGERY.

DEEPER LEVELS OF UNCONSCIOUSNESS

A
B
C
D
E
F
G
H
DREAM
DRIVES THRUSTING DOWN
CONFLICT
TULPOID

Fig. 29. Dreams are unresolved conflicts

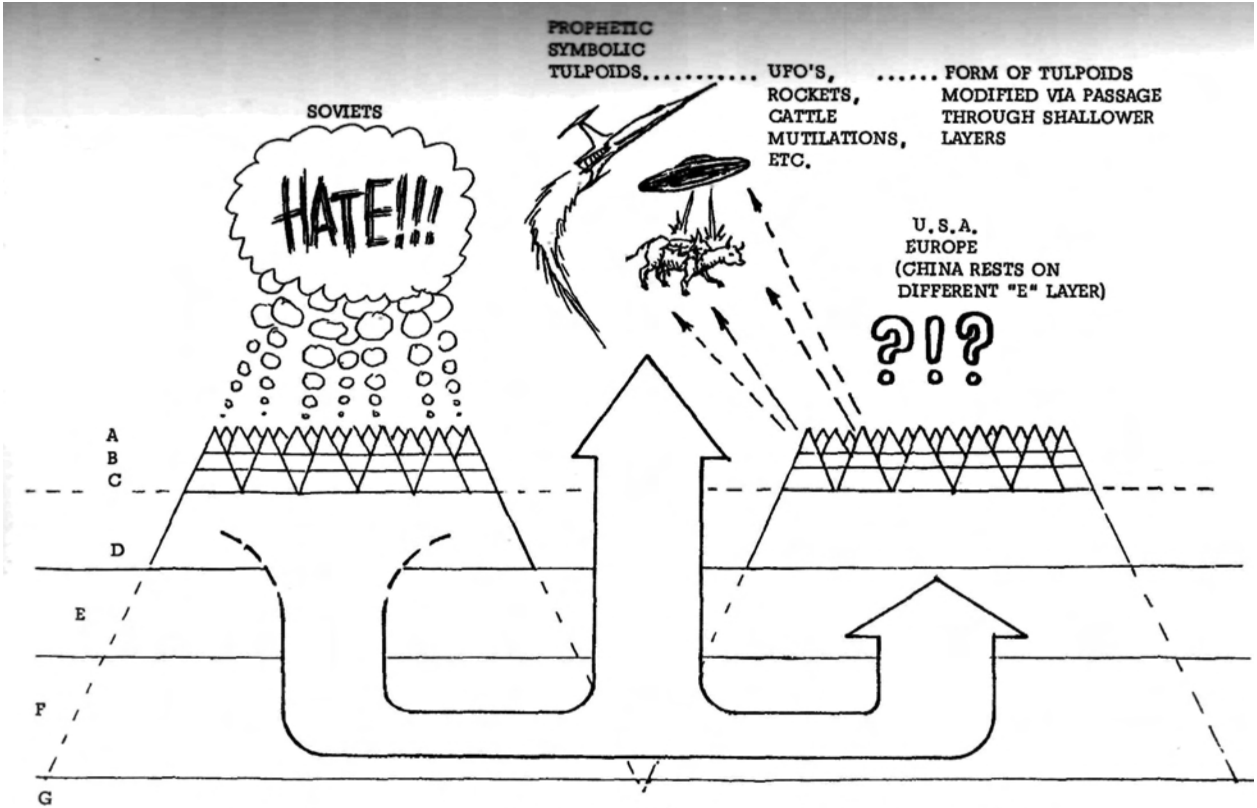
PROPHETIC SYMBOLIC TULPOIDS..........
UFO'S, ROCKETS, CATTLE MUTILATIONS, ETC.
...... FORM OF TULPOIDS MODIFIED VIA PASSAGE THROUGH SHALLOWER LAYERS
SOVIETS
HATE!!!
U.S.A.
EUROPE
(CHINA RESTS ON DIFFERENT "E" LAYER)
?!?
A
B
C
D
E
F
G

THE PSI CHANNEL IS LIKE A QUASI-HYDRAULIC U-TUBE. CONSISTENT PRESSURE INTO DEEP LAYERS RESULTS IN UPWELLING TULPOIDS WHOSE FORM IS MODULATED AND SHAPED BY SHALLOWER LAYERS, ON THE WAY TO MATERIALIZATION.

Fig. 30. Collective conflict initiates prophetic tulpoids

inceptively. It is not accidental that the Avatar Jesus exhorted man to love God with all his heart and mind, and to love his neighbor as himself. It is also not accidental that a human culture seeks and "obtains" a godform or god-image consistent with its own level of understanding and belief; the resulting tulpoidal activity conditioned into the collective unconscious and the cultural unconscious, and tuned in from them, will be consistent with the cultural desires and expectations.

So in a sense we are the creators of our own reality and our own tulpoidal expressions of higher reality. But in addition to tulpoids, the present character of spacetime and dimensionality, and the very laws of the physical universe, are largely prescribed by the collective human unconscious.

GHOST ROCKETS AND FLYING SAUCERS: MATERIALIZED TULPOIDS

Unfortunately, from 1945 to 1947 events occurred that caused a split among previous World War II allies and that led to the beginning of the Cold War in 1947.

Due to these events, Western nations perceived a growing threat from the Soviet Union. The resulting interplay of events, news reports, and personalities, plus the traumatic aftermath'of World War II, led to strong psychological pressure on the group unconscious of the Western nations, and thence on the collective human unconscious. This pressure resulted in popouts of direct orthorotations of resonant archetypal symbols from the collective unconscious, the archetypal format being successively modified by the Western group unconscious and the progressively higher levels of unconsciousness.

The formats of the deep human unconscious were formed millions of years ago in the primitive tribal state. In the tribal warfare of those ancient days, the male went to where the action was, while the female stayed behind in the protected cave, fortress, or shelter to care for the young. Thus the male symbol was closest to the threat and identified with being there.

Furthermore, in primates the phallic symbol is the universal symbol of threat, dominance, and defiance. Dominant primate monkeys often briefly mount submissive males to show their dominant position, exposing their genitals in so doing. If the mounted male does not submit, an immediate battle ensues to determine the dominant male. Thus the phallic symbol represents threat and domination.

In the deep human unconscious, fire is a symbol of power, for with fire the species held predators at bay and began the long trek up from the jungle cave to modern civilization.

Finally, the unconsciousness of a substantial number of the Western nations had been deeply impressed by the German V-l and V-2 rockets during World War II, for the imagination of the world had been breathtakingly fired by the surprise advent of the German rocket weapons.

Thus we would expect that in those Western countries — Norway, Denmark, Sweden, etc. — that are close to the Soviet Union, archetypal tulpoid popout materialization would have taken a male, threatening, dominant, power format. The nearest, most powerful, most resonant recent stimulus symbol in the Western group unconscious would have been the final modulation, determining the final format.

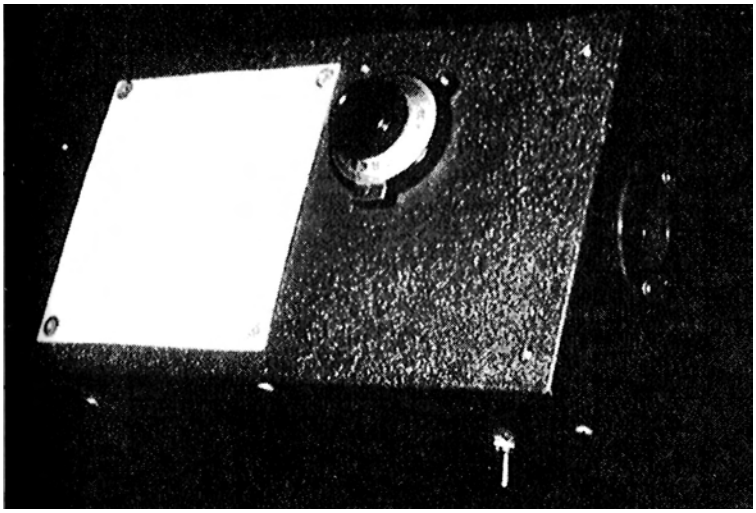

Outside showing contact ptate, sample well, tuning vernier, and on-off switch.

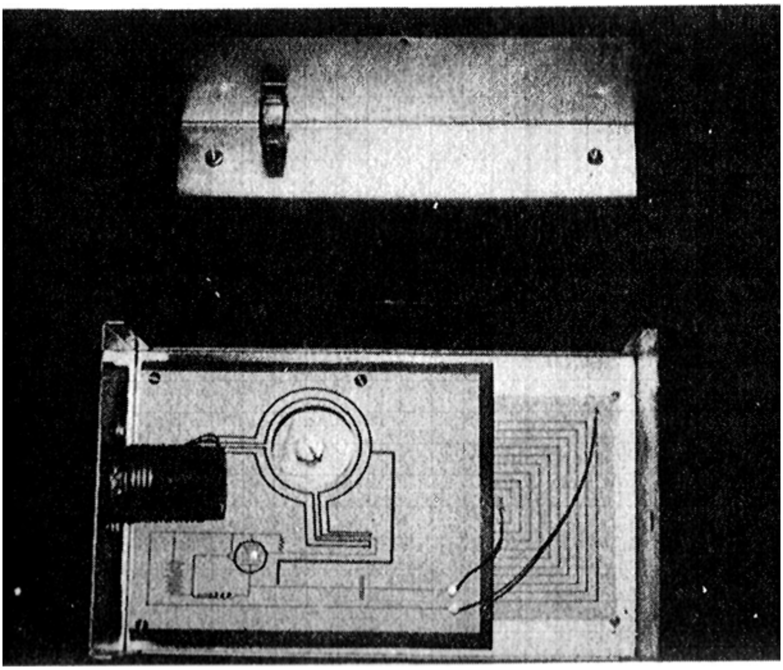

Inside, showing India Ink drawing of *an a*mplifier schematic, used as the "amplifier.'

Photo 41. Hieronymus device — second kind.

So a massive wave of "ghost rockets" from the direction of the Soviet Union appeared in the skies over Norway, Sweden, and Denmark, and other nearby European countries. (See Photograph 43.)

However, North America constituted a natural fortress with great ocean barriers to the east and west, and no strong enemy to the north or south. Therefore, a female symbol (archetype) should have been resonant and stimulated in the U.S., and so it was. In 1947, the year the Cold War began, Kenneth Arnold, flying over the state of Washington (the most northwestern U.S. state at the time, and the one nearest the Soviet Union), encountered a formation of flying saucers — the female circle mandala, dressed in modern aeronautical form. Thus was born the age of flying saucers in America.

PART THREE

NEW MILITARY APPLICATIONS OF PSI RESEARCH

BACKGROUND TO PSYCHOTRONIC RESEARCH IN THE U.S. AND THE U.S.S.R.

In this section I will discuss some of the fantastic psychotronic weapons that have been built and deployed by the Soviets as a part of operation "Fer-de-Lance," the Soviet plan to conquer the world.

The Soviet Union has had a massive developmental program in what they call energetics or psychoenergetics (and what we are calling psychotronics) for about thirty years. To appreciate the mind-staggering weaponry that has emerged from this effort, I should point out that the U.S. developed the atomic bomb in World War II in about four and a half years; thus the Soviet psychotronic program is the equivalent of roughly seven Manhattan projects, back-to-back. With this comparison we gain a rough yardstick to appreciate the fantastic capabilities that must have emerged from the sustained Soviet program. As we will see, they are indeed mindboggling.

Ironically, most of the original theoretical work was done here in the United States, just as the pioneering work in rocketry was done here by Goddard, before the Germans picked it up and advanced it to the V1 and V2 rockets. Just as Goddard himself was once the laughing stock of the Pentagon, so the key U.S. researchers who could have initiated the psychotronic age in the West — Moray, Abrams, Reich, Hieronymus, Drown, Tesla, de la Warr, etc. — were generally ignored or prosecuted or both. Even more deadly has been the studied avoidance of any application of the basic theoretical work — which again was developed largely in the U.S. This work included the Everett-Wheeler-Graham many-worlds interpretation of quantum mechanics, and the extension of Lorentz's work by Ives, Ruderfer, and others to show that almost all modern physics — including general relativity, Newton's laws of motion, and the law of universal gravitation — can be straightforwardly derived from a Lorentz-invariant nonmaterial ether. Even though Pauli pointed out that the Michelson-Morley experiment only eliminated a material ether rather than ether per se, and although Einstein himself once proposed the use of the term ether to include dimensionless vacuum space complete with various fields, modern physics in the West has ignored the ether theory.

As Robert Anton Wilson points out, whatever you believe imprisons you. The West has been solidly imprisoned in its own orthodox view of physics, and hence has been unable to surmount the stone walls of its self-imposed mental barriers. In fact, it has not even tried. There has been no serious, sustained program to deliberately examine the loopholes in present theory and seek the necessary changes to provide a theory of paranormal-psychotronic phenomena. In fact, the major Western work on paranormal phenomena seems only to attempt to "prove" over and over that such phenomena exist; and to attempt to derive from present physics theory a "normal sane" theory for "abnormal insane" phenomena. The orthodox establishment continues to vigorously attack the experimenters who struggle with paranormal research on the grounds that their work does not present science.Hence the experiments and experimenters are unscientific and either at worst frauds and charlatans or at best seriously deluded.

The simple procedure necessary to solve these questions — set up a deliberate, well staffed, well-funded program on a continuing basis — has not been adopted in the West because of the short-sightedness of our management structure. We seem to have totally forgotten that we got the atomic bomb in the first place and we put a man on the moon, simply by starting a program with staff and funds to try to do it, not by engaging in a neverending series of arguments as to why it could not and should not be done.

So, ironically, the nation that prides itself on being free and liberal has adopted an iron dogma of its own in science, a dogma that yields very slowly and that has effectively blocked us from serious and prolonged efforts in psychotronics.

Meanwhile, the Soviet scientists — admittedly kicked firmly in the tail by Stalin — have simply got off their dust-encrusted seats with alacrity and got to it. In thirty years of such getting to it with massive state support, they have already launched the psychotronic age, first applying it to the weapons field, the typical Soviet priority. Meanwhile, for the same thirty years we in the U.S. have played the same old tired games that once led Heisenberg to remark that new scientific knowledge only gained acceptance when the old scientists who so adamantly opposed it had died away.

Finally, there is a direct relationship between UFO's, psychotronics, the collective human unconscious, the Cold War, and Fer-de-Lance that cannot be escaped. So while this section emphasizes psychotronic weapons deployed by the Soviet Union I continually remind the reader of the concepts involved in the related paranormal phenomena, which were covered in Parts One and Two. So with these words, let me give a brief summary of Soviet psychotronic weapons. Make no mistake, it appears that the final holocaust is now unavoidable — we will reap the full rewards for our smugness, pride, arrogance, and inflated egotism.

The Soviets will unleash World War III and they will unleash the psychotronic weapons on a totally unsuspecting world.

But there is something more; it does not end there. There is another player in the game — the collective human unconscious, ZARG itself, will unleash a reaction that makes psychotronic weaponry pale by comparison. Yet when that great entity has roared out its anger, destroyed perhaps half the earth, and then opens its eyes, the full biblical prophecy will in a most bizarre and unsuspected way be fulfilled.

Time will end, and the grave will give up its dead. The unifying of all human minds into one will complete the birth of a new godlike being, extending from the beginning to the end of time all at once, and from beginning to end of all possible times all at once. Three-dimensional concepts and limitations will fade away as child's play to the being that is already in four billion dimensions.

And you and I will become gods, literally, actually — for being is holographic, and whatsoever ZARG becomes, so we all become.

PSYCHOTRONIC BACKGROUND SINCE WW II

After WW II Soviet scientists were ordered by Stalin to actively seek another area of great breakthrough — such as the U.S. atomic bomb had been — and leave no stone unturned in their search. Apparently the Moray detector was resurrected and intense experiment and development begun on it. Much of the pioneering radionics work of the West and Ive's work would also have been available at this time. I believe that in 1958 the Soviet nuclear catastrophe near the Urals was the result of an attempt to transmit a nuclear explosion through hyperspace. (The peculiarity of a transmitter tuning failure would have resulted in an explosion emerging from the nearest large nuclear waste storage, and that apparently happened.)

In 1956, however, a major new interpretation of quantum mechanics was originated by Hugh Everett at Princeton University under the world-renowned John Wheeler. This is the theory of the universal wave function—the many-worlds interpretation (MWI) of quantum mechanics. Practically the entire Western output on the subject is available in his 1973 book entitled *The Many-Worlds Interpretation of Quantum Mechanics.*

This physics is so exceedingly strange that it has been ignored by almost all Western physicists — indeed, it is so strange that even its originators, Everett and Wheeler, are said not to be able to believe it. Wheeler is reported to have once stated that he could believe it on Monday, Wednesday, and Friday, but not on the other days of the week. Using the MWI, however, one can produce a theoretical model of psychotronics which is totally consistent with the entire experimental basis of modern physics, and unite both mind and matter on a single scientific basis.

The Soviets apparently seized on this new theoretical work and applied it to the Moray device, solving their theoretical and technical difficulties.

In 1960, Khrushchev, speaking to the Presidium, stated that the USSR had a new weapon in development ("just within the portfolio of our scientists, so to speak") that was so powerful it could wipe out all life on earth. I believe he referred to the nuclear hyperspace howitzer and to the death ray weapons under development at that time, such as the following.

ELECTRON DISSOLUTION

There are four levels of effects on biological systems that are irradiated by electron dissolution beams.

Interference with the electrical functioning of the nervous system is the dominant effect. Single channel subsystems are more sensitive to the effect.

Mild irradiation (or mild kindling of short duration) causes weakness, trembling, and an unreal, dreamy feeling (as when one gets a whiff of anesthesia). The victim is conscious, however, since consciousness is limitedly multichannel.

Moderate irradiation (or kindling of moderate duration) causes paralysis of the motor system, and a very unreal, floating feeling. The individual is barely conscious and is hypnogogic. Panic is usually engendered.

Strong irradiation (more protracted kindling) causes loss of consciousness in addition to paralysis. The autonomic nervous system still functions with difficulty, and the victim is just barely alive. Convulsions and other symptoms may appear.

Very strong irradiation (protracted kindling) totally jams all levels of the nervous system, and death results instantly.

In UFO contactee incidents this effect is widely demonstrated. Many cases show step function failure of automotive electrical systems, lights, radios, etc., as a UFO approaches. When the UFO departs, the electrical systems almost invariably resume complete operations. Absolutely no electrical components are damaged.

To extinguish these systems with massive RF fields would result in blowing fuses, diodes, transistors, etc., and a great amount of electrical sparking — probably sufficient to blow up the gasoline tanks, In fact, none of these things occur.

Stage 2 electron dissolution should definitely disable automotive circuits, and the upper level of stage 1 may be sufficient. Either would leave the automobile passengers conscious, though paralyzed in one case and not in the other. Both types of correlations are shown in the UFO literature.

THE 1973 CONTACTEE WAVE AND CATTLE MUTILATIONS

In 1973 the Yom Kippur war took place. This was the Soviets' "Spanish Civil War," and they tested the final tactics and equipment for antitank and antiaircraft weaponry, and obtained final data on tank and equipment loss rates for modern combat. The effect on the U.S. and NATO of the cutoff of Arab oil was also tested.

At one point, when the Israelis cut off an Egyptian army in the desert, the Soviets notified the U.S. over the hot line that Soviet forces were going in and the U.S. could do whatever it pleased. The Soviets loaded and readied seven airborne divisions to insert, and the U.S. readied its single airborne division to send in. So the world teetered on the brink of direct ground confrontation between the Soviet Union and the U.S., with all its possible consequences.

Confrontation is after all another form of *contact.* And UFO contactee cases involve *confrontation.* In the very month, October, that the Yom Kippur War erupted, the greatest wave of UFO contactee cases in history broke out in the U.S. Further, in the shallower unconscious levels above the collective species unconscious, which were modulating and molding the tulpoidal reality format, the imagination of the world had been captured by the spacesuited astronauts who had walked on the moon. So most U.S. contactee cases involved spacesuited beings, conditioned into that framework by the national, cultural modulation. Some of the contactee cases involved robots or metallic men, similar to the mechanisms and robots of the U.S. space program. And here in the U.S., the fortress, the female-mandala craft symbol was generally maintained.

Now if this unbelievable scenario has any validity, then after the dress rehearsal of the 1973 Yom Kippur War we ought to have seen an increase in the tulpoid phenomena of a sharply symbolic nature. And that has happened.

The cow is the Western female symbol par excellence; Western children nurse on cow's milk, so to speak. If the U.S. fortress is to be violated, then the female symbol will be violated as the time for the Soviet assault nears. So cattle mutilations of a mysterious, paranormal nature have been occurring all across the U.S. — from the East Coast out through Colorado, Utah, and Montana — approximately nine thousand by now, in some estimates. The symbolism in the mutilations can be easily read: the U.S.'s lifeblood is going to be drained, totally, as all the blood is usually drained from the cattle. The Soviet operation will be launched with blinding speed and surgical precision, and the mutilation of the cattle is done as if by highly skilled surgeons. The children of the West shall be cut off from their sustenance, and so the cow's teats are removed. The armed forces (males) confronting the Soviet forces will be devastated (i.e., the fruit of the womb will be

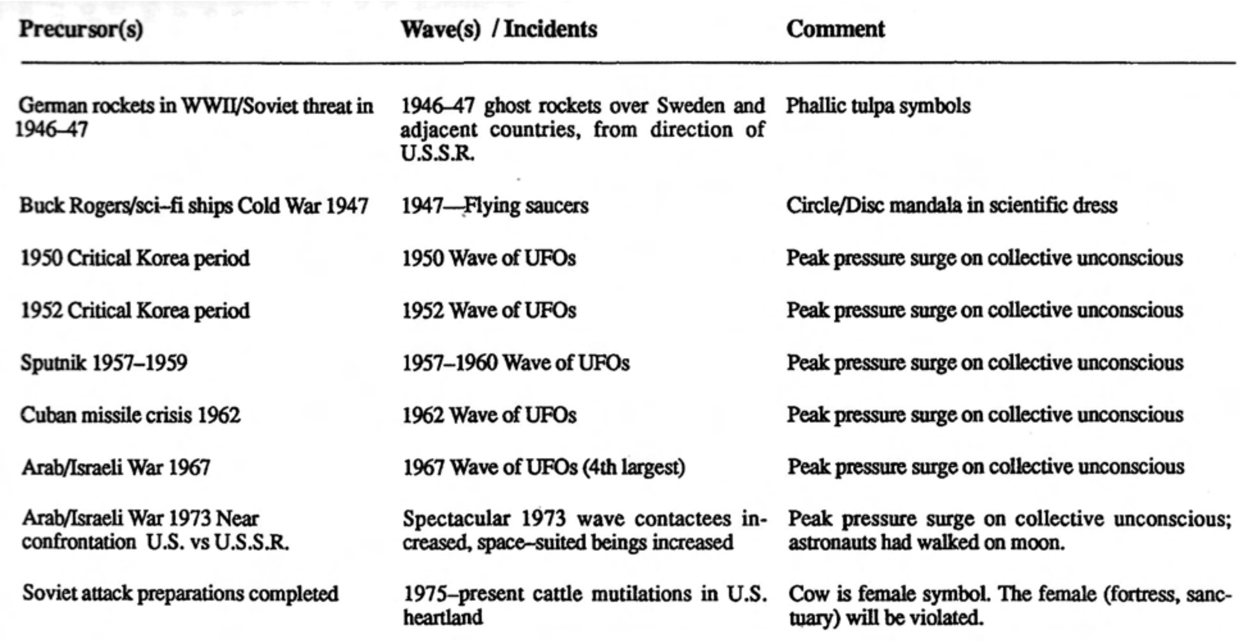

Precursor(s)	Wave(s) / Incidents	Comment
German rockets in WWII/Soviet threat in 1946–47	1946–47 ghost rockets over Sweden and adjacent countries, from direction of U.S.S.R.	Phallic tulpa symbols
Buck Rogers/sci–fi ships Cold War 1947	1947—Flying saucers	Circle/Disc mandala in scientific dress
1950 Critical Korea period	1950 Wave of UFOs	Peak pressure surge on collective unconscious
1952 Critical Korea period	1952 Wave of UFOs	Peak pressure surge on collective unconscious
Sputnik 1957–1959	1957–1960 Wave of UFOs	Peak pressure surge on collective unconscious
Cuban missile crisis 1962	1962 Wave of UFOs	Peak pressure surge on collective unconscious
Arab/Israeli War 1967	1967 Wave of UFOs (4th largest)	Peak pressure surge on collective unconscious
Arab/Israeli War 1973 Near confrontation U.S. vs U.S.S.R.	Spectacular 1973 wave contactees increased, space–suited beings increased	Peak pressure surge on collective unconscious; astronauts had walked on moon.
Soviet attack preparations completed	1975–present cattle mutilations in U.S. heartland	Cow is female symbol. The female (fortress, sanctuary) will be violated.

Fig.31. Examples of UFO waves

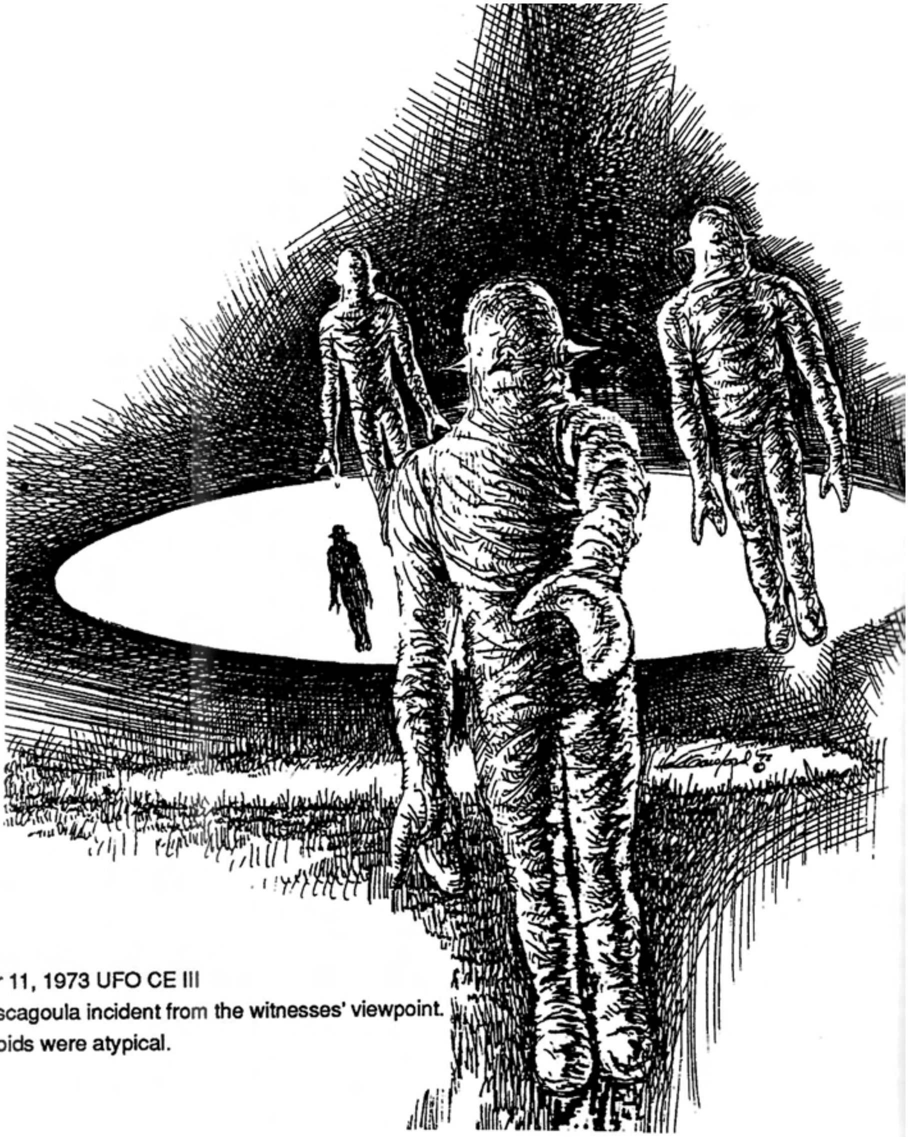

October 11, 1973 UFO CE III
The Pascagoula incident from the witnesses' viewpoint.
Humanoids were atypical.

Fig. 32. 1973 Contactee wave

devastated) so the sexual organs of the cows are excised. Free speech and freedom to hear will be cut off, so the cow's lips, tongues, and ears are excised. A very high (psychotronic) technology will play a role, but one which will limit resistance, so the cut edges are often made as if by pinking shears and then precisely burned as if by a laser. (The laser is symbolic of new, strange technology almost incomprehensible to the average person.) In one case a horse was the victim — a young horse, symbolic of the young soldier. It was castrated, as the dominance of the Western soldier will be surgically excised. Castration of the male symbolizes the total loss of his power and the infliction of impotency upon him.

There is more, but this is a bloodcurdling sampling of the metapsychological implications of the cattle mutilations.

RADIATION OF THE U.S. EMBASSY

About 1959 or 1960 the radiation of the U.S. Embassy in Moscow with weak microwave radiation began. It continues to this day. The radiation is a brilliantly designed intelligence probe: the Soviets were vitally interested in ascertaining whether or not their U.S. adversary knew about psychotronics (i.e., whether the U.S. had built such weapons of its own and might have built defense counterweapons as well). Accordingly, they began weak PT-modulated microwave radiation of a high-level U.S. target, the U.S. Ambassador to the Soviet Union, which would guarantee the personal attention of the State Department, NSA, CIA, DIA, the President, etc. The PT modulation apparently consisted of various disease patterns of gradually increasing seriousness — flu, blood changes, heart irregularity, cancer, etc. By the U.S.'s reaction, whether or not it knew anything of PT could be positively assessed. If the U.S. reacted only to ordinary electromagnetics, then with 100 percent certainty it knew nothing of PT and had no secret PT weapons of its own or counters to the Soviet PT weapons. So since about 1959 or 1960, by our alow reaction to the radiation of the U.S. Embassy — and then by our reaction only to the electromagnetic aspects — we have been assuring the Soviets that we know nothing of PT. Even though several U.S. presidents have protested the radiation, the Soviets have persisted because it is an intelligence tool of the utmost importance to them.

DISEASE AND DEATH TRANSMISSION

For confirmation of Soviet experimentation with PT induction of death and disease, see V.P. Kaznacheyev et al, "Distant intercellular interactions in a system of two tissue cultures," *Psychoenergetic Systems,* Vol. 1, No. 3, March 1976, pp. 141-42; and by the same authors, "Apparent information transfer between two groups of cells," *Psychoenergetic Systems,* Vol. 1, No. 1, December 1974, p. 37. These articles give one a glimpse of the results of over 5,000 Soviet experiments on the basic effect. Kaznacheyev is the head of the Siberian Branch of the Soviet Academy of Medicine, and has at his disposal every disease germ, virus, and toxin known to man. Note that the experiments were performed in the dark to prevent the quenching effect of the visible light spectrum und the effect is only obtained through the quartz (which has IR and UV windows). Also note that in the two-slit experiment — which contains all the mystery of quantum mechanics, according to Nobel prizewinner Richard Feynman, cofounder of quantum electrodynamics — one does not get the quantum effect with an electron, if the electron is first hit by a photon of light before reaching the two-slit region. If the electron is not hit by a photon, one gets the quantum effect. According to Feynman, the quantum effect in the two-slit experiment is not explainable by any classical means. Our comment is that the explanation is not possible by Aristotle's three laws of logic, on which all Western logic, physics, and mathematics are based. However, the fourth law of logic, together with the present three laws, provides a closed metalogic which does contain the explanation of the two-slit experiment, as we showed in Part Two. The fact that there is no problem in explaining the experiment if one takes into account the nested structure of time was pointed out as long ago as 1957 by Dr. Charles Muses. Similarly, the disease and death transmission experiments of Kaznacheyev et al have been explained in Part Two.

OTHER WEAPONS

Clandestine testing of PT weapons also appears to have been accomplished directly in the U.S. The Legionnaires' disease may have been one such test. The suspect agent was actually seen, according to sworn testimony before a Congressional committee. A person with a book-sized object in his hand (tape recorder with PT pattern), a tobacco-pouch sized object in his coat breast pocket (transmitter) and a wire running to his tie clip (antenna) was present and mingled with the Legionnaires for several hours. He was heard to exclaim several times to various Legionnaires that it was too late and they were all doomed. The eventual appearance of the weird "biological agent" that is different from any other germ known can also be explained by kindling from the anxious minds of the baffled U.S. researchers and scientists. Other such tests have also apparently been conducted.

HYPERSPACE HOWITZER OPERATION

The facility at Semipalatinsk is a third-generation hyperspace nuclear howitzer. It can transmit a full nuclear explosion through hyperspace into a tuner planted anywhere in the universe, without traveling through or in normal space at all. Furthermore, it has a unique characteristic: any number of receiver-tuners may be tuned to the transmission channel and each will have the full nuclear explosion emerge from it, undiminished.

Thus, one channel might be for U.S. Minuteman sites, and a tuner on that frequency (say one containing 100 transistors, sealed in plastic and buried in the earth) could be preplanted near each Minuteman site. Another channel might be for all the SAC bases, another for large U.S. cities, another for nuclear weapon storage sites, etc. Any or all of these could be connected to a single blast transmission, simply by controlling the channel width over which the hyperpace transmission occurs.

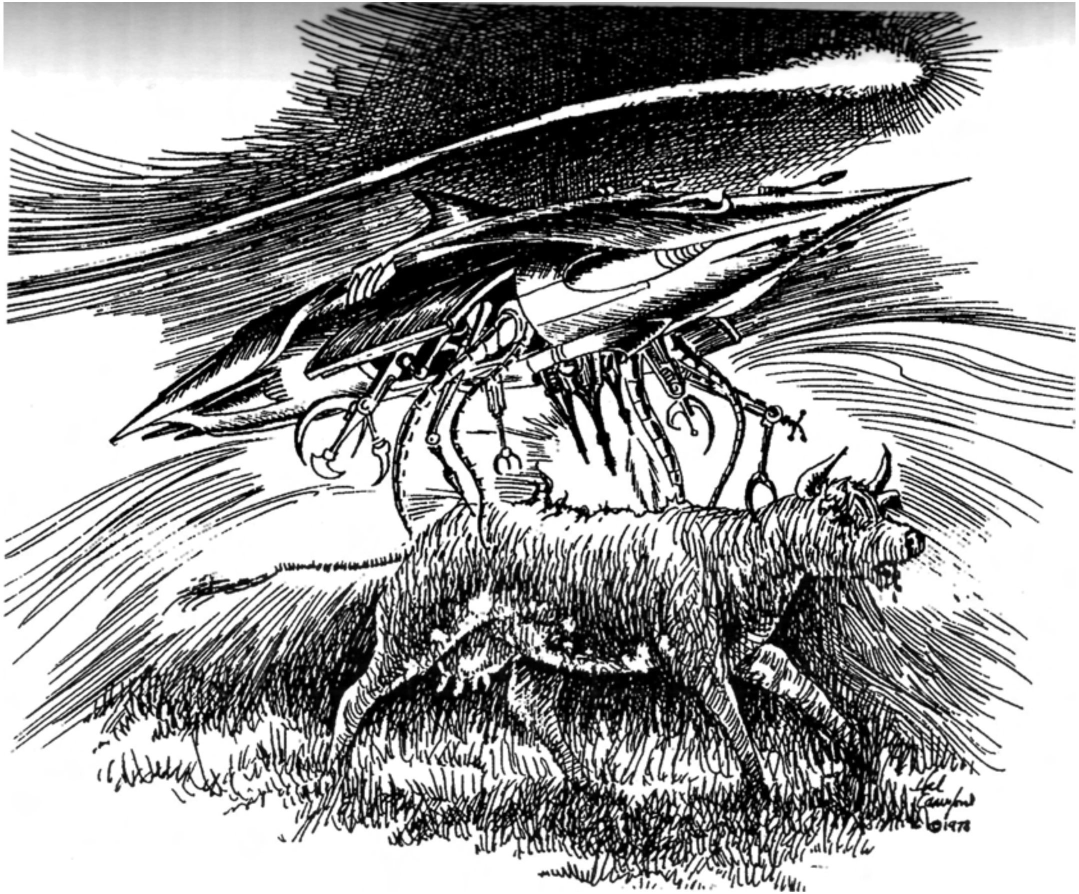
Hal Crawford
©1978

Fig. 33. Cattle mutilations

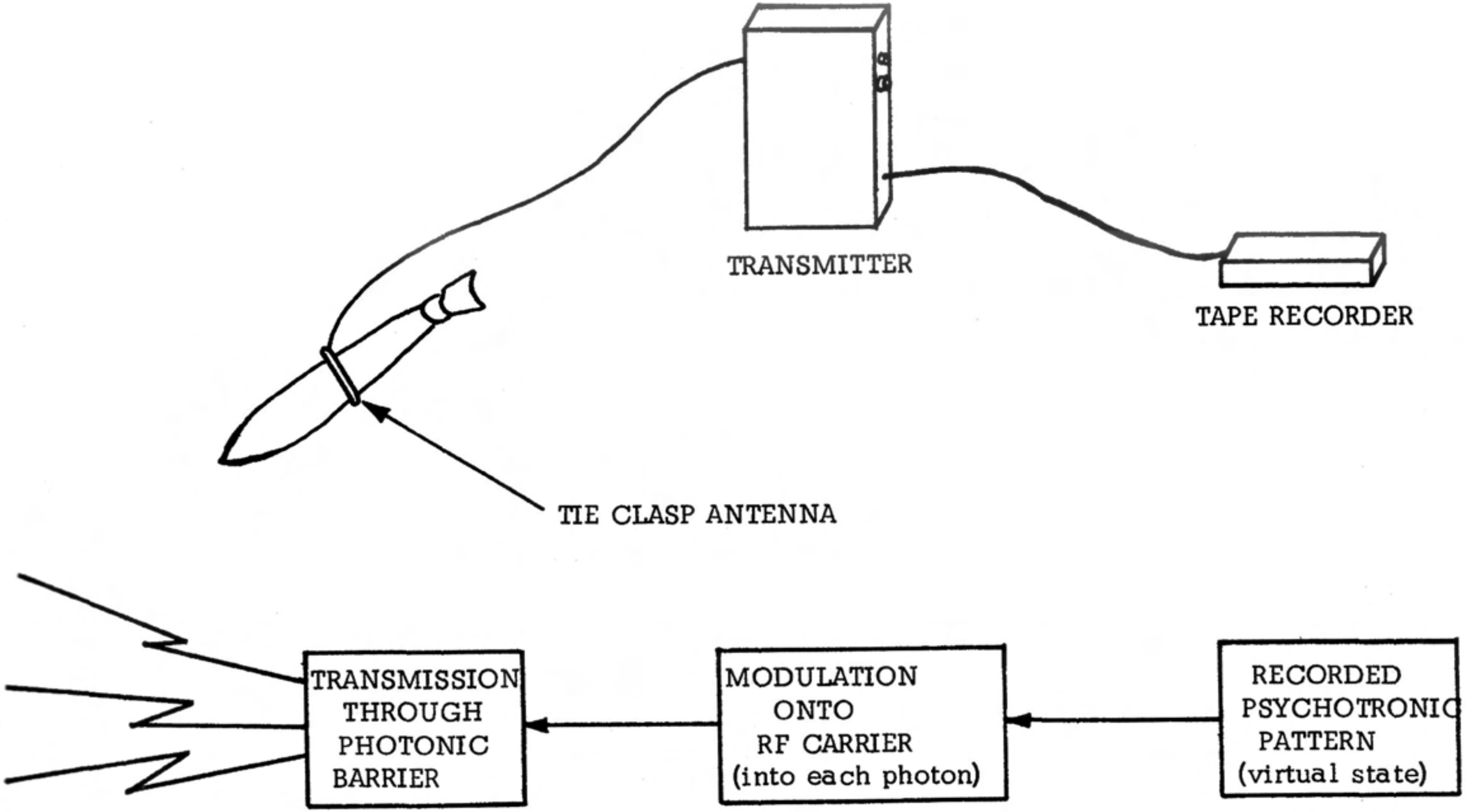
TRANSMITTER
TAPE RECORDER
TIE CLASP ANTENNA
TRANSMISSION THROUGH PHOTONIC BARRIER
MODULATION ONTO RF CARRIER (into each photon)
RECORDED PSYCHOTRONIC PATTERN (virtual state)

Fig. 34. Induction of Legionnaire's disease

Here one must not confuse ordinary single-channel conservation of energy with hyperspace operation. The transmitter corresponds to the grid of a gigantic triode, the preplanted receivers constitute the plate area, and zero-point vacuum energy constitutes the cathode energy source. So just as in a triode, energy is conserved from cathode to plate, but not exclusively from grid to plate. Since zero-point vacuum energy is being gated by the triode grid (the howitzer) signal (the explosion), the energy that can be outputted at the plate (from the tuner-receivers) is essentially unlimited. Any number of receiver-tuners can be connected to a channel, merely increasing the plate area. And from each tuner there can emerge a full-scale, undiluted replica of the explosion in the howitzer at Semipalatinsk.

Thus a single nuclear blast at Semipalatinsk can in one instant wipe out all hind-based strategic nuclear retaliation capability in the Free World, if the preplanted tuners are on site.

An additional modification — something like additional orthorotator rings surrounding the primary hyperspace orthorotator ring — may have been made to the nuclear hyperspace howitzer within the last two or three years. If so, this would have the characteristic of materially reducing the "spillage" fraction of the nuclear explosion and also the vibration associated with a firing. There is a possible indicator of this: quoting from *Science*, "On five occasions during the first half of 1976 traces of neptunium-239 and molybdenum-99 were detected in southern Sweden. These activities were accompanied by small amounts of fission products such as iodine-131 and barium-140. One of the events was coincident in time with large peaks in the atmospheric tritium gas concentration." (Lars-Erik DeGeer, Swedish National Defense Research Institute, "Airborne short-lived radionuclides of unknown origin in Sweden in 1976," *Science*, Vol. 198, No. 4320, 2 December 1977, pp. 925—27.) These events were concentrated during the time the Semipalatinsk experiments are reported to have started, yet no seismic activity was associated, as would have been the case if ordinary nuclear explosions had occurred in normal space inside a containment sphere for a large particle beam experiment. On the other hand, if a very high percentage of the explosion were transmitted/rotated into hyperspace, then residual traces of a nuclear explosion would have been present in the atmosphere because of the very slight spillage venting into normal space, but almost all mechanical shock and vibration would have been eliminated. It is at least a striking correlation that exactly such indicators are present, as reported by DeGeer.

A clever adversary who becomes aware of the necessity for preplanted tuners could use a very simple but completely effective countermeasure: simply locate the timers and dig them up, then clandestinely put them back into Soviet territory. Or build tuners of one's own and plant them in Soviet territory. Thus it would seem imperative to develop a mode of operation that does not require preplanted tuners at the target area.

This capability can be provided by using the howitzer in an interferometer mode — by constructing additional orthorotation rings around the firing chamber ts (hut a "hyperspace interferometer" exists. With this scheme any point on earth can be precisely focused on by the howitzer, and no preplanted tuner is necessary. The advantage of this mode is the elimination of the need for clandestinely preplanted receivers, and the freedom to operate in spite of any receivers preplanted by the adversary on Soviet soil. The disadvantage of the interferometer is that the former "triode" operation has now been converted to "diode" operation — when one explosion is put into hyperspace, now a single explosion at a single target area emerges, since a single selected spatial location is now exclusively determined by the interferometer settings.

For a more complicated interferometer mode, orientation and alignment of the interferometer's indicators and controls and amplifiers would be necessary. This could be done using fractional kiloton yield weapons — or even ordinary HE explosions — and focusing the howitzer at a displaced aiming point, at some distance from the target area. This is simply the standard artillery technique of registering a gun on an offset registration point, then performing an x-y shift to the intended target when the gun is fired.

Orientation and alignment of such a weapon system probably accounts for the sonic booms — mysterious high altitude explosions — off the East coast of the U.S. in December 1977 and January 1978. Over the West coast, mysterious explosions occurred inside thunderstorms; these also could well have been the adjusting of the weapon. By using about 100 tons of TNT explosive in the howitzer, a very good high burst registration could be made, and "forward observer" agents in the U.S. could observe the bursts through optical instruments and report their exact positions, thus calibrating the interferometer settings. Of course, mysterious booms have been around for quite a few centuries; however, these "normal but unusual" phenomena have quite a different sound than does the howitzer shot. The natural booms are not so much like single giant explosions as like rolling thunder. Furthermore, several observations of the booms off the U.S. East coast confirmed light flashes (which have not been seen to accompany the natural booms) and swirling or rotating effects in the flashes. Also UFO's were directly associated with some of the booms, revealing the stress factor on the collective human unconscious, implying their Soviet psychtronic origin. A few years ago similar high altitude booms occurred over Ohio and surrounding states, and such booms have also occurred off the coast of England.

Thus the howitzer may have been accurately aligned for diode operation, calibrated near the East coast, on the West coast, in the center of the U.S., and near England. Furthermore, its firing rate and recovery time were also apparently checked out; five booms occurred on one day and after a lapse of several days, the booms occurred again.

As I describe the fantastic psychotronic weapons the Soviets possess, the reader may wonder why the Soviets have not already attacked. The reason is that the superweapons have a strange weakness and failure mode possessed by no other kind of weapon. The hyperspace amplifiers in the psychotronic weapons depend on a large number of stages, say sixty to one hundred, and the hyperspatial flux in all these stages must be very stable and isotropic. At odd intervals the flux coming in from hyperspace suddenly becomes turbulent; this destroys the smooth coherence in the multiple stages of the amplifiers, and thus the basic amplification mode. Thus *all* the psychotronic weapons will suddenly cease working at the same time, and this "all fail" mode may last from fifteen minutes to even three weeks in an extreme case. Thus the weapons are erratic and not dependable. For this reason, the Soviets dare not rely *exclusively* on them, and so they have been forced also to prepare for full nuclear war as a distinct possibility when they launch operation Fer-de-Lance, the full attack. They have thus built up their regular army, their navy, their strategic offense systems, and their strategic defense systems on a scale unparalleled in human history. Today the Soviet Union is the strongest nation on earth — by far — and is fully prepared to launch World War III and win it, with or without superweapons.

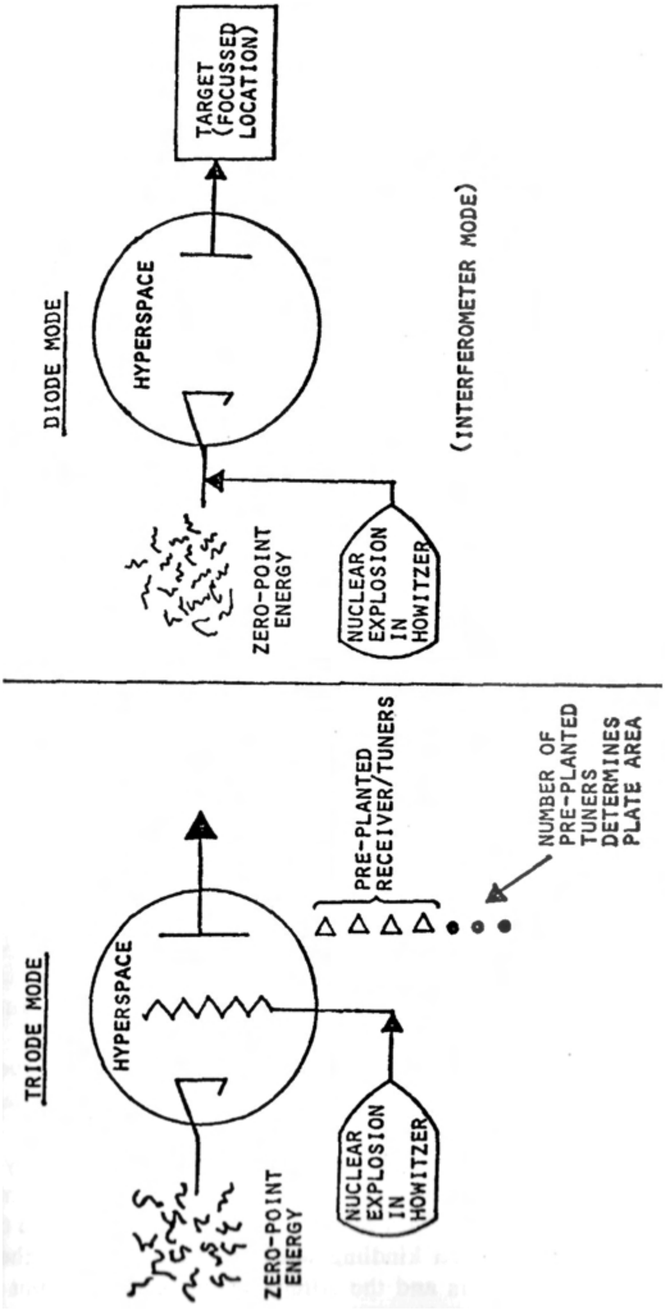
TRIODE MODE
HYPERSPACE
ZERO-POINT ENERGY
NUCLEAR EXPLOSION IN HOWITZER
PRE-PLANTED RECEIVER/TUNERS
NUMBER OF PRE-PLANTED TUNERS DETERMINES PLATE AREA
DIODE MODE
HYPERSPACE
TARGET (FOCUSSED LOCATION)
ZERO-POINT ENERGY
NUCLEAR EXPLOSION IN HOWITZER
(INTERFEROMETER MODE)

Fig.35. Hyperspace howitzer operation (2 figs)

A POSSIBLE OPERATIONAL TEST OF PSYCHOTRONIC ANTIAIRCRAFT WEAPONS

When the F-l11 all-weather attack aircraft were deployed in Vietnam, a rather interesting loss rate pattern emerged. Of the forty-eight F-l11's in operation, eight were lost at the beginning of the action, but no losses occurred thereafter, according to General Vogt, the officer responsible for U.S. air operations at the time.

Initially there were some radar anomalies and other system malfunctions with the aircraft, but these problems were quickly resolved. The F-l11's flew singly, extremely low, and without radio contact. Between September 1972 and the end of hostilities in Vietnam, the F-lll's flew over 4,000 sorties, never losing any planes other than the eight lost initially.

The reason for the loss of the eight planes remains a perplexing mystery. In fact, there is no knowledge of exactly where the planes were lost, except in one case. That F-l11 was downed over Hanoi, and its two pilots, Captain Robert D. Sponeybarger and 1st Lieutenant William W. Wilson, were captured alive by the North Vietnamese. The two airmen were later returned under the prisoner of war exchange program, and they then told of the loss of the aircraft, which took place under very strange conditions.

According to the two crewmen, they met a sudden disaster while flying over Hanoi at night, supersonically, and at very low altitude — only a few hundred feel above the earth. Without warning the F-lll's emergency indicators lit up like a Christmas tree, indicating massive and multiple failures of unrelated aircraft systems.

They were not hit by a surface-to-air missile, according to the crewmen, because no explosion of a missile warhead, with its concomitant fireball, occurred. If large-caliber automatic weapons fire had hit them, the distinctive streak of the tracers would have been clearly visible in the night sky, and a sharp impact in the aircraft structure would have been felt. None of this happened.

It could not have simply been an aircraft malfunction, because far too many unrelated aircraft systems failed at the same time. (There is triple redundancy in the control system of the F-l11, by way of comparison.)

Indeed this initial loss of eight aircraft — and then no further losses at all — is difficult to explain rationally, as are the mysterious circumstances surrounding the loss of Wilson's and Sponeybarger's craft.

However, it requires no stretch of the imagination to realize that the Soviets very much desired to obtain the glittering new F-l11 to examine its technical characteristics, its sophisticated design, and its highly sensitive terrain-following radar, ECM gear, and other avionics equipment.

Also, the Soviets had a pressing need to try out one or more of their psychotronic antiaircraft weapon systems — perhaps one which simply modulates a radar beam psychotronically to induce electron dissolution.

Accordingly, it can be postulated that the pragmatic Soviets simply combined both needs, and quickly set up a psychotronic trap for the F-lll's immediately after the aircraft were deployed. They may even have cleverly designed the specific psychotronic modulation so that kindling would occur in some of the aircraft's electrical systems before others and the failure would be massive but not total.

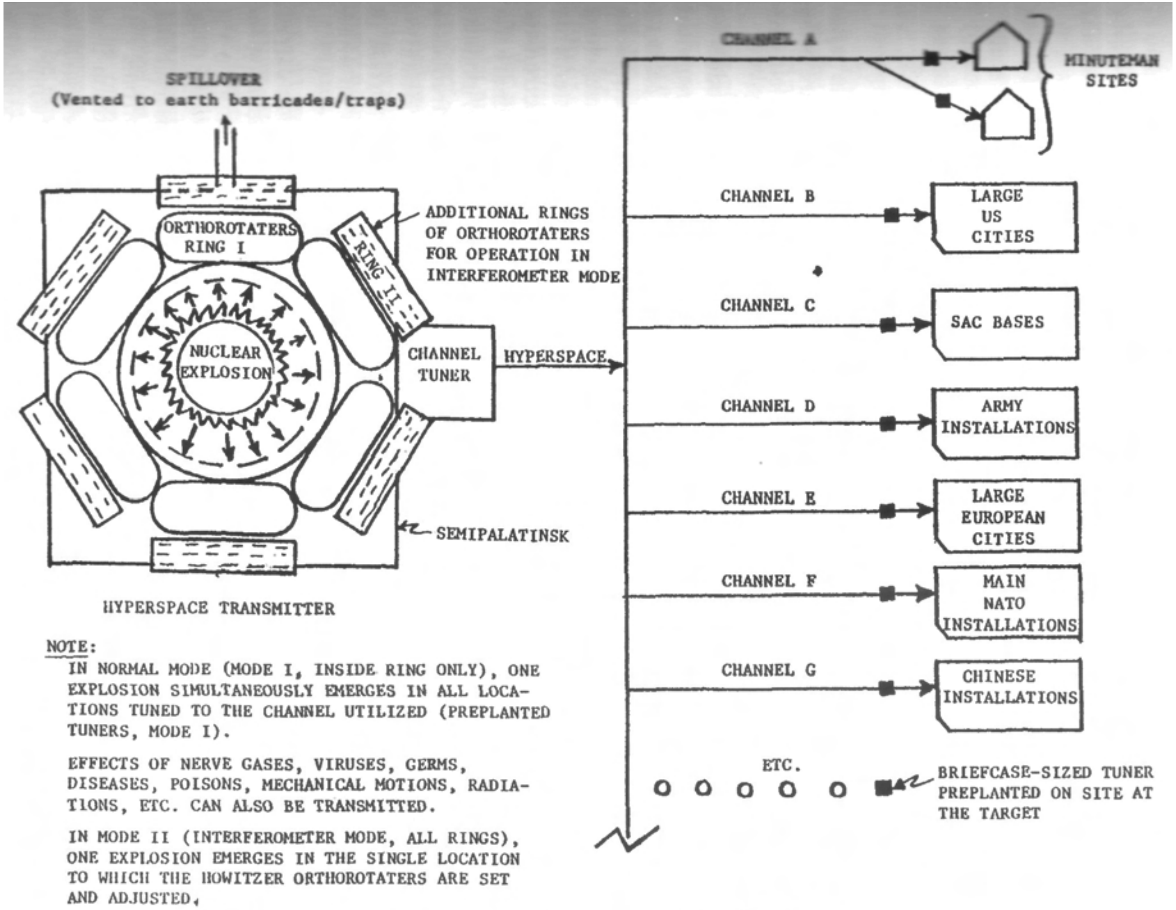
SPILLOVER
(Vented to earth barricades/traps)
ORTHOROTATERS RING I
RING II
ADDITIONAL RINGS OF ORTHOROTATERS FOR OPERATION IN INTERFEROMETER MODE
NUCLEAR EXPLOSION
CHANNEL TUNER
HYPERSPACE
SEMIPALATINSK
HYPERSPACE TRANSMITTER
CHANNEL A
MINUTEMAN SITES
CHANNEL B
LARGE US CITIES
CHANNEL C
SAC BASES
CHANNEL D
ARMY INSTALLATIONS
CHANNEL E
LARGE EUROPEAN CITIES
CHANNEL F
MAIN NATO INSTALLATIONS
CHANNEL G
CHINESE INSTALLATIONS
ETC.
BRIEFCASE-SIZED TUNER PREPLANTED ON SITE AT THE TARGET

NOTE:

IN NORMAL MODE (MODE I, INSIDE RING ONLY), ONE EXPLOSION SIMULTANEOUSLY EMERGES IN ALL LOCATIONS TUNED TO THE CHANNEL UTILIZED (PREPLANTED TUNERS, MODE I).

EFFECTS OF NERVE GASES, VIRUSES, GERMS, DISEASES, POISONS, MECHANICAL MOTIONS, RADIATIONS, ETC. CAN ALSO BE TRANSMITTED.

IN MODE II (INTERFEROMETER MODE, ALL RINGS), ONE EXPLOSION EMERGES IN THE SINGLE LOCATION TO WHICH THE HOWITZER ORTHOROTATERS ARE SET AND ADJUSTED.

Fig. 36. Nuclear hyperspatial howitzer

And after they nailed enough F-111's for their purposes, they quietly withdrew the psychotronic modulator, to prevent the U.S. intelligence analysts from catching on to just how really strange the losses of the F-111's were.

Thereafter the remainder of the F-111's were successful, and no more were lost. Indeed the F-111 comes in so low and so fast, and so unexpectedly, tilting and dipping as it terrain-follows on its automatic equipment in a manner that no unassisted pilot could fly, that SAM's and guns normally pose little if any threat to it. So once the silent, eery psychotronic antiaircraft components were withdrawn, the normal antiaircraft weapons of the North Vietnamese were unable to down any more F-111's.

And we Americans simply patted ourselves on the back, chortling to ourselves how quickly we had "learned to avoid that danger, whatever it might have been."

POPULATION CONDITIONER

Inducing psychotronic signals into mass populations or audiences can influence or control the behavior of large numbers of persons. As these words are being written, such a psychotronic test is apparently being conducted by the Soviets with Eugene, Oregon as a target; previous tests were conducted in Canada.

Psychotronic signals and the patterns they carry can be, and are, modulated upon electromagnetic carriers — in fact, that is one way they are transmitted through the human body and nervous system. When a television camera picks up the photon field impinging on it from a human body or other biosystem, weak psychotronic patterns are also picked up, since the paranormal time quenching mechanism (by visible light) is never quite 100 percent effective. These patterns are riding then as a special kind of modulation on the radiofrequency or electromagnetic carrier — however, they are not modulated the same way as are ordinary signals. Ordinary signals gate the number of photons per second (amplitude modulation), or change and control the frequency of the carrier photons (frequency modulation), or change and control the phasing (phase modulation). However, psychotronic modulation rides on each individual photon, on its time portion (a photon consists of action — energy x time — and it is the time portion on which the psychotronic modulation is riding.)

Thus psychotronic modulations can ride on the transmitted television carrier frequency. The psychotronic modulations are processed through the television receiver, and a small component is emitted as a weak modulation riding on the photon signals emitted from the television picture tube. Along with the light from the tube, these psychotronic modulations enter the eye of the human observer and are demodulated by the retinal interface, thus introducing weak psychotronic signals into the observer's central nervous system. Given sufficient repetition time (say two to four hours), the kindling effect can cause the human nervous system to be profoundly affected. Any disease or death pattern known to man (and some that are not known to him) can be transmitted and kindled in the target population audience.

In fact, depending on the sophistication of the coded psychotronic signals, it *is* probable that the modulation patterns can be targeted against a specific racial type, since there are certain genetic differences between races.

So the type of truck that roams the streets of a city, silently monitoring which television channel the viewers are watching, could be modified slightly to induce specific psychotronic disease or death vectors onto the television receivers. Or it could simply modulate the television transmitters. (A very obvious way is to modulute or pump psychotronically the ground of the electrical power supply, and in most cases this can directly pump the transmitters.)

Or the kindling signal can come from satellites, which radiate weak radio frequency signals down toward the target population. Here economy of force can be applied; the same type of kindling that induces electron current quenching is usable as a satellite dudder or killer in an antisatellite role, or is usable as a paralysis and death beam weapon against target populations. It is also useful as a killer of other electromagnetic sensors by either jamming their circuitry or simply **wiping out** (quenching) their emissions which impact on the satellite, preventing any reflected signal.

By this effect, UFO tulpoids often prevent a radar from tracking them, **even** though the tulpoids are visually sighted.

There are literally thousands of applications; these are just a few to give an idea of the overall effect and its usefulness.

And the psychotronic signals can be modulated onto almost any kind of carrier — telephone, television, radio, the magnetic field of the earth, the Schuman resonance of the earth's magnetic field, the ordinary sixty hertz electrical power system, etc.

We might have an entirely different idea of the potency of the Russian trawlers if we appreciated what their use can be in the psychotronic mode.

We can also understand why a television broadcast of Uri Geller in the act of binding metal stimulated both the ability in other persons and the effect in other equipment across England. And also how the effect has been duplicated — at least so far as stimulating other persons to psychokinetic ability — in this country. A direct means of stimulating psychokinetic ability is thus suggested: simply place the subject to be developed or stimulated in a room with a number of television receivers, all tuned to the same visual program, and place the stimulus material on the receivers. The person should visually watch one receiver and should be wearing very abbreviated clothing, such as a skimpy swim suit, to expose the acupuncture plexuses. Also, normal light in the room should be very dim.

At any rate, this and many other schemes are available to produce direct stimulation, and enhance and greatly speed up psychic development.

The effect could also be used to produce an effective treatment device for cancer. Samples of the particular cancer in the body of the patient could be taken, and the invisible psychotronic radiation pattern from the cancer witness could be introduced on the input of a Hieronymus-type device, shielded against ordinary light, The output of the Hieronymus device could then be fed to a bank of phase inverters, which covered in total a broad band (the cancer band, which would have to be established by laboratory experiment). By inverting the cancer pattern, we now have a cancer-extinguishing type of radiation that is specific for the cancer the patient in fact has.

Then by introducing this quenching pattern as a modulation onto the correct tup of electromagnetic frequencies (again this group would have to be demined for the specific frequency plexus that would produce an effect in the local-

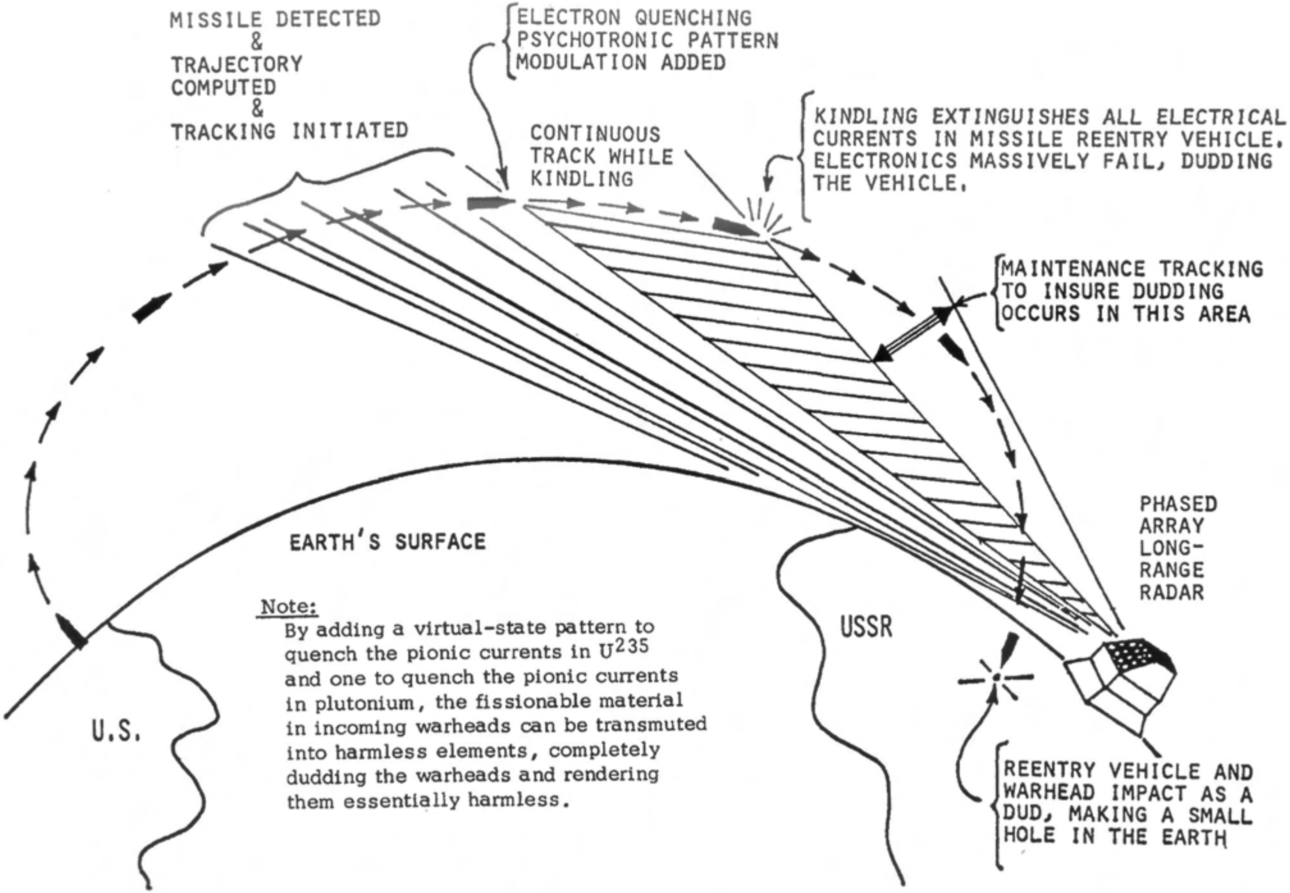

MISSILE DETECTED
&
TRAJECTORY
COMPUTED
&
TRACKING INITIATED

ELECTRON QUENCHING
PSYCHOTRONIC PATTERN
MODULATION ADDED

CONTINUOUS
TRACK WHILE
KINDLING

KINDLING EXTINGUISHES ALL ELECTRICAL CURRENTS IN MISSILE REENTRY VEHICLE. ELECTRONICS MASSIVELY FAIL, DUDDING THE VEHICLE.

MAINTENANCE TRACKING TO INSURE DUDDING OCCURS IN THIS AREA

REENTRY VEHICLE AND WARHEAD IMPACT AS A DUD, MAKING A SMALL HOLE IN THE EARTH

Note:

By adding a virtual-state pattern to quench the pionic currents in U^{235} and one to quench the pionic currents in plutonium, the fissionable material in incoming warheads can be transmuted into harmless elements, completely dudding the warheads and rendering them essentially harmless.

Fig. 37. A midcourse ABM system

ity of the body in which the cancer is located), the proper acupuncture plexus would be stimulated, stripping off the cancer-quench psychotronic pattern and sending it to the specific location of the cancer. Over a period of time (say several very weak microwave radiation periods of two to four hours each), the cancer could be killed completely, down to the last malignant cell. Whole body or spread cancers could be treated simply by increasing the number of frequencies modulated, to hundreds in some cases. Other similar diseases could be treated likewise. As an example, a Curative psychotronic radiation pattern complex could be worked for hydrophobia and such presently incurable diseases, and these could also be effectively treated. The same tool will yield almost unbelievable medical benefits if the scientists can ever be convinced to start the development.

A MIDCOURSE ABM SYSTEM

Here is the scheme for a first-generation, long-range, Soviet midcourse anti-ballistic missile defense system.

A long-range radar — preferably a phased array, scanned radar — acquired and tracks the missile during its midcourse phase shortly after it breaks above the horizon, as viewed from the radar site. The radar programs the track mode beams against the incoming missile, tracking it and gathering sufficient position data to allow its trajectory to be computed. Then a psychotronic electron-extinguishing pattern is added to the radar beam, riding as virtual state carried by each individual photon. Kindling of this virtual pattern energy begins on and in the target vehicle. The electrons flowing in the electromagnetic control system, fusing system, and wahead system of the incoming missile begin to be quenched, turning into neutrinos and radiating away with no physical effect. When sufficient currents have been extinguished, the electromagnetic circuits start failing. (With loss of appreciable power and voltage, most such systems fail.) The incoming vehicle/warhead is now dudded, and there is no longer any danger of a nuclear burst from the missile.

For assurance, tracking of the incoming reentry vehicle continues, with a quenching psychotronic modulation being maintained.

The warhead/reentry vehicle strikes the ground, making a hole in the earth, without nuclear explosion. A low order detonation of the HE explosive surrounding the nuclear material in the warhead may occur.

An alternate scheme is, during the maintenance period, to add an HE explosives ignition temperature psychotronic pattern to the beam. In that case, a low order detonation occurs high in the atmosphere or back along the trajectory, removing the warhead and vehicle from further consideration.

As a backup, additional terminal systems — using either psychotronics or normal interceptor missiles or both — may be used. Thus even if the psychotronics fail due to an unexpected "all off' occurrence, ordinary ABM engagement occurs.

An additional scheme is to lock the radar beam onto the missile, then initiate a "Tesla effect" kindling of energy, place it into a columnar traveling wave, and direct it along the radar beam to the incoming missile in its midcourse phase. The midcourse missile is thus struck with a massive plasmoid fireball of intense electromagnetic energy, and massive amounts of energy arc dumped into the missile, destroying it.

Thus the Soviets have known for over a decade that one does not have to hit a bullet with a bullet, if one is clever, and if one possesses psychotronic weapons and psychotronic modulation systems for one's radars.

As an added benefit, the same radar is quite effective against incoming Strategic Air Command bombers or air-to-surface missiles fired from those bombers.

As the Soviets know full well, 100 percent defense is quite feasible, considering the fantastic number of radars and radar sites they possess! In fact many of the Soviet SAM systems are beautiful ABM systems also, in the psychotronic mode.

THE BELL ISLAND INCIDENT

On April 2,1978, something unusual hit the Canadian community of Bell Island, off the east coast of Newfoundland; Bell Islanders—and a lot of other people as well—are still trying to figure out what it was. It was first described as an unidentified atmospheric explosion, but this theory was quickly rejected. Jim Beckford of Lance Cove, Bell Island, lost all six chickens in the coop behind the house, which was torn up as if hit by lightning, although there were no burn marks anywhere. A television set exploded, and switches blew out of the wall. Similar things happened to many other homes.

In the woods behind Mr. Beckford's house, three holes in the snow appeared, spaced about two meters apart to form an almost perfect equilateral triangle. The Royal Canadian Mounted Police said that persons on the shore of Conception Bay reported seeing fireballs coming down out of the sky and hitting Bell Island. The Canadian authorities attempted to play down the incident, and the official version of what happened is that lightning hit a transformer, blew up, and that the energy then followed the wires to neighboring houses. However, the researches of P.A.C.E. (the Planetary Association for Clean Energy, Ottowa) have produced a different conclusion.

P.A.C.E. believes that a standing columnar wave originating in Gomel in the Soviet Union was moving to a station located just east of Havana, Cuba, and collapsed over Bell Island, triggered by the large iron ore body there. These plasmas then exploded and hit the ground in a perfect equilateral triangle formation. The tremendous flashes indicated a very great discharge of electrical power, probably more than a hundred million volts.

Photo 44 shows a similar incident that took place at Lake Ontario; here I feel that the photo demonstrates more than just UFO activity — it shows one part of the testing and operational alignment of the "over-the-horizon" radars at Gomel and Riga in a Tesla mode. To explain this, I must explain what I mean by a Tesla mode.

Empty vacuum is filled with stupendous amounts of uncorrelated electromagnetic energy: one cubic centimeter contains enough energy to form entire solar systems, if the energy were condensed into matter. This raw energy is called "zero" point energy," and the Soviets reported direct measurement of it in 1957 *(Soviet Physics,* JETP 3, 819). Zero-point energy is everywhere, it is incoherent, and its detection requires measuring an energy difference, so that it is very difficult to detect. Usually less than one quantum of energy is cohered at any one mode, and the energy rapidly changes frequency, so that linear detectors cannot resonate coherently with it to follow its signal. In my opinion, it is also in the virtual state, and in a space orthogonal to our own, by two or three turns.

An electromagnetic signal is transmitted through the earth and reflected from the opposite side in phase, to form a standing wave. The earth's interior is highly nonlinear, and a single frequency is thus subject to severe scattering and dissipation. Fortunately, nonlinear wave theory provides a simple method to alleviate this; by transmitting two different frequencies, which differ by the amount of the frequency that it is desired to transmit, one can simply pretend that *indeed the beat difference frequency has been transmitted through a linear medium.* In this case, the standing wave acts as a collector to kindle the highly chaotic energy from the core of the earth. These emerging orders that match the standing wave will be entrained by it; one now has established a sort of giant triode, where the signal to the grid is the transmitted difference signal, the energy (cathode) being gated is the energy of the molten core of the earth multiplied by some coupling fractional constant, and the "plate" will come into being if resistance to the standing wave is profferred.

And resistance is profferred by simply changing the frequency of the transmitted collector wave. Changing it in one fashion begins to dump energy into the atmosphere above the former nodal point across the earth from the transmitter. With that effect, one can directly influence the great currents of air and the ionization charge of the upper atmosphere; in short, one can drastically influence the weather itself. On the other hand, changing the frequency in the other direction begins to dump energy into the mantle of the earth; dumping large amounts of energy there will result in shocks and earthquakes at the nearest fault zone(s).

It is my thesis that the foregoing two mechanisms account for some of the drastic weather effects that have occurred in the past several years, as well as for *some* of the major earthquakes that have occurred throughout the earth in the last several years. The Soviets, I believe, have been orienting and aligning—and operationally testing—actual weapon systems to be used as precursors to war or during war. The 1978 Iranian earthquake may well have been Soviet-induced.

The coupled kindling path goes through the earth and is coupled through the atmosphere as well. (This mode is probably easier to steer and focus onto specific target areas. It is also probably easier to play it back and forth as desired, in a "sweep" mode.) Again, with the kindling effect, the energy of the molten core of the earth is coupled into the beam with a coupling fractional constant, and so enormous energy can be built up in the beam. Breaking the coupling — perhaps just after slightly changing the frequency so as to begin dumping energy into the upper atmosphere over the targeted area — would sharply dump energy at the site, resulting in an enormous plasmic explosion or explosions. In testing such as at Bell Island, probably frequencies would be used that did not offer too strong a coupling fraction, so that relatively little damage would occur at the decoupled target site. However, in actual use in warfare, exactly the opposite would occur; a frequency would be used that maximized the coupling constant and hence vastly increased the energy kindled into and over the target area, and a gigantic plasmoidal explosion would take place directly on the targeted area, utterly devastating it.

The same "radar," then, is usable as a weather control device, as a plasma explosion transmitter, as a bioactive disrupter (with 10 Hz and under modulation pulses), and if further modulated in a virtual state manner, as a direct "death ray"

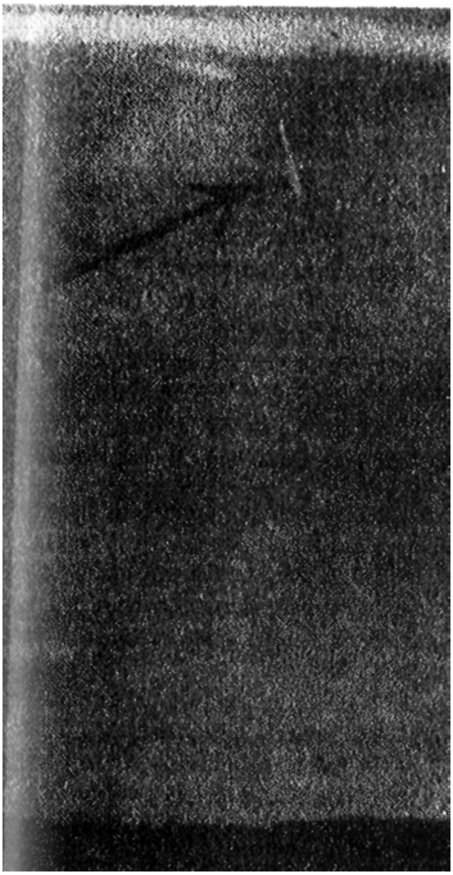

Photo 42. Ghost rockets.

broadcaster through the earth at any target population on the earth. Further, with virtual state modulation it is usable as an antimissile weapon (in at least two modes, one to dud electromagnetic circuits in incoming warhead reentry system and one to transmute fissionable material in incoming nuclear warheads into harmless elements), and also as an antiaircraft weapon against Strategic Air Command Bombers,

So the Bell Island incident happened, and we in North America sleep on, since in our assurance that it was after all only all lightning, or perhaps only the illusion of a few farout Canadians.

A TERMINAL ABM SYSTEM

There is another more advanced psychotronic ABM/SAM system, which was probably a second generation weapon. This system should have gone on site and become operational about 1968—69, and it should have been operationally tested during that period, probably several times.

Here a triad configuration radar is used to allow use as a byperspatial interferometer. This allows several localized dome-like bands of very intense psychotronic energy to be focused above the installation and defended area. The focusing is accomplished through hyperspace; the only place the focused energy enters our own space frame is in the dome-like band above the protected area. Tesla-effect accumulators may feed the triads, or they may simply be fed from direct free-energy Moray generator-amplifiers. At any rate, enormous energy — enough to simply constitute a severe tearing and warpage of space itself, and bend it at 90° — is focused into each dome-like band. Each band would be seen externally as simply an intense band or hemisphere of glowing light.

At any rate, by using three consecutively tuned orthorotational bands of focused hyperspace energy, any incoming object will simply be orthorotated successively three times in higher dimensional space, each time losing one dimension in its intersection with our own spatial frame.

After the first orthorotation, masses are now photonic shapes, and photons are lines. After the second, photons are now lines and lines are points. After the third orthorotation, everything is points, and nothing is left.

The three consecutively constitute full dematerialization of anything incoming, whether matter, photon, or force line.

Thus the entire defense problem is immediately solved: regardless of chaff, maneuvering decoys, maneuvering re-entry vehicles, etc., the entire mess is simply nicely and neatly disposed of, and dumped into hyperspace.

Even if a full nuclear detonation occurs above the domed shields, gamma rays, neutrons, debris, x-rays, heat, and ultraviolet do not get through the third band.

Bombers and air-to-surface missiles fail equally.

So the growing Soviet confidence and preparations for imminent war are quite comprehensible. They are not madmen preparing to evoke their own doom from mutually assured destruction rained on them by U.S. strategic nuclear forces; instead, they are snug and safe behind their psychotronic defenses, for nothing we have on this earth can touch them at the present time, if their psychotronic weapons are working.

So they are quite capable of taking on China, Europe, and the U.S. at once, and devastating the lot while risking essentially nothing of Mother Russia.

SWEEP MODE, SWEEPING AWAY A NUCLEAR DETONATION

As a modification of the previous weapon, a sweep mode has probably been added in the third generation weaponry.

In this mode, not only can the three orthorotation shields or domes be placed over an area, but they can then be expanded radially out into space in a traveling mode. Thus if a high altitude nuclear explosion — either in the atmosphere or out of it — has occurred over a defended area, the orthorotation bands can simply be swept over the explosive debris in successive fashion, effectively sweeping the radioactive debris from this spatial frame and dumping it safely into hyperspace. This is a very handy weapon indeed if large megatonnage explosions are engendered — say in areas where terminal defenses have destroyed the incoming vehicle, but it is "dead man fuzed" to provide a full nuclear explosion if the electronics start to fail or if the warheaded re-entry vehicle encounters mechanical shock or turbulence above a certain threshold. With this psychotronic sweeper the otherwise deadly nuclear radiation and radioactive debris can simply be cleaned up and disposed of.

Further modifications of other psychotronic weaponry would have resulted in Tesla-effect traveling wave weapons of enormous power, and rather gentle operational weapons testing of them, such as in the Bell Island incident and over Eugene, Oregon. The interferometer mode would have been added to the nuclear hyperspace howitzer, and Tesla-radars would have been modified accordingly.

These third generation weapons and modifications should have been deployed during the period 1974-78. Engineering development tests would have occurred during 1974—76, and operational testing of on-site, deployed weapons would have been accomplished during 1977—78.

Thus one can reasonably account for a surge in grain elevator explosions, mysterious high atmospheric booms off the East coast of the U.S. and off England, explosions in thunderstorms over the West coast of the U.S., the zapping of a Canadian target outside Ottawa, the Bell Island incident, several mysterious aircraft failures over the sagebrush triangle and off the coast of the U.S., standing heat spots off the West coast of the U.S., and recently one of the coldest winters on record due to disruption of flow patterns of frigid arctic air masses.

One can also account for incidents where U.S. satellites were mysteriously blinded, and perhaps some of those where they failed and then later resumed operations.

Note that many of these effects are usable from smaller radars. Thus one could expect mobile Soviet radars that can disable tanks and vehicles, disable aircraft electrically, radiate a death ray pattern at ground troops, or launch Moray-fed fireballs or standing waves of enormous energy at ground or aerial targets. The ability to immediately destroy entire minefields alone would force a severe recalculation of defense capabilities against Soviet ground forces, and the ability to sweep away artillery shells, rockets, missiles, and bullets would present enormous and catastrophic consequences to forces opposing advancing Soviet formations.

The following list enumerates all the probable Soviet psychotronic weapons that I am aware of: (1) Electromagnetic Field Canceler; (2) Electron Current Canceler; (3) Death Radiator (De-inceptor); (4) Free Energy Generator; (5) Psychotronic Bomb; (6) Brain Link; (7) Brain Probe; (8) Disease Radiator Ray; (9)

Emotion Radiator Ray; (10) Antisubmarine Systems; (11) Telcporters (Prototype); (12) Force Generators; (13) Orthoframe Generators; (14) Quark/Antiquark Rays; (15) Distant Telepathy; (16) Nuclear Radiation Inhibitors; (17) Mental Implantation; (18) ABM Systems, Midcourse and Terminal; (19) Dematerializer Shields; (20) Sweeps for Nuclear Debris; (21) AAA Systems; (22) Antisatellite Systems; (23) Tesla-effect Weapons; (24) Earthquake Generators; (25) Antivehicle Systems; (26) Weather Control Systems.

The reader will note that the results of psychotronic research extend into every field of military application.

ZARG, ARMAGEDDON, AND LINKAGE

Now it is time to end my story, and cease the unrelenting stretching of the reader's imagination and credulity. But first it is necessary to make one more effort and take a final step to the bottom line in both physics and Jungian psychology, a step that the scientists in those two disciplines have not dared to face.

If one accepts the existence of the collective unconscious as a mind, then the collective unconscious is a living entity. Let us tentatively call this fantastic entity ZARG. ZARG then is a great superentity, of which our own personal consciousnesses and unconsciousnesses are parts, or cells. Thus each of us is a cell in the entity ZARG.

Furthermore, ZARG has unique characteristics. It is unconscious by definition — but note that from our fourth law of logic, we know that any zero-conscious must be multiply conscious, which to our monocular thought-perception processes can only be grasped as unconscious. That is why our own unconscious seems so bewildering and strange — it is multiply conscious rather than monoculurly conscious, hence monocularly we think it is not conscious. But we find that it indeed thinks and reasons, but in a strange symbolic fashion, and also much more deeply and more creatively than our normal conscious mind. ZARG is very similar except on a much vaster scale; in being very deeply unconscious, ZARG is supremely multiconscious and very symbolic indeed, as we have discovered from the nature of the materialized tulpoidal phenomena.

Furthermore, what can be taken for ZARG's conscious mind consists of the four-billion individual human conscious minds. Thus ZARG's conscious mind is fragmented into four billion pieces. But fragmentation of the conscious mind technically constitutes dissociation, which classically defines psychosis.

Thus ZARG is insane!

Also, ZARG is multidimensional. ZARG's unconscious simultaneously receives and identifies four billion different conscious and unconscious perceptions. Just as the human consciousness in identifying two slightly different two-dimensional visual patterns gains the third dimension, ZARG gains an additional dimension for every additional person past the first one. (Note that dimension just means variable, and two fixed axes/dimensions are required to obtain a variable. A one-dimensional variable does not exist; we usually assume a time dimension in addition, and ignore it — and thus we speak of a "one dimensional variable.") At any rate, ZARG gains an additional dimension for every additional person past the first one. So ZARG is at least about 4 x 10^9-dimensional in the simplest case, where we do not consider permutations and combinations of every two, three, four, ...n, . . . minds.

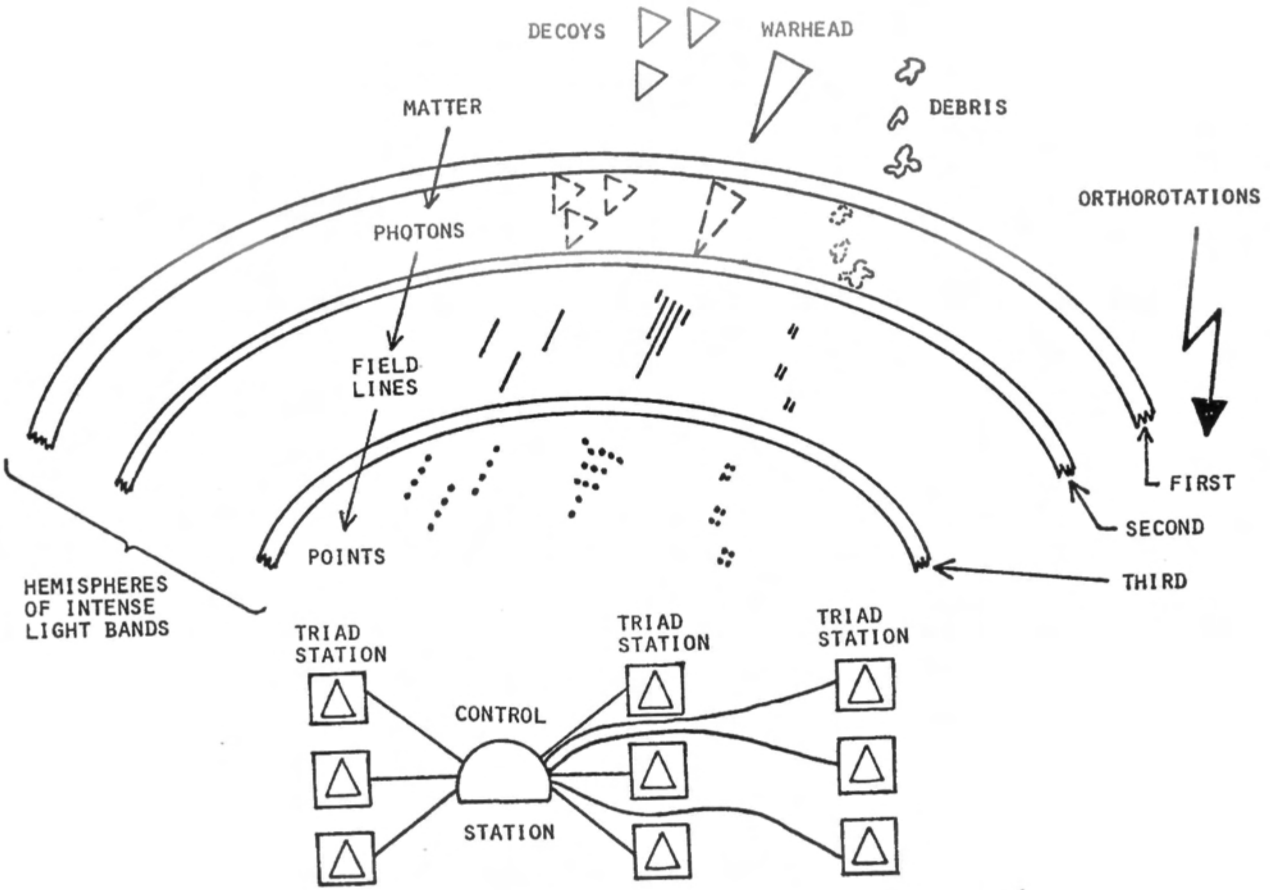
DECOYS
WARHEAD
DEBRIS
MATTER
PHOTONS
ORTHOROTATIONS
FIELD LINES
POINTS
FIRST
SECOND
THIRD
HEMISPHERES OF INTENSE LIGHT BANDS
TRIAD STATION
TRIAD STATION
TRIAD STATION
CONTROL
STATION

Fig. 38. A terminal ABM system

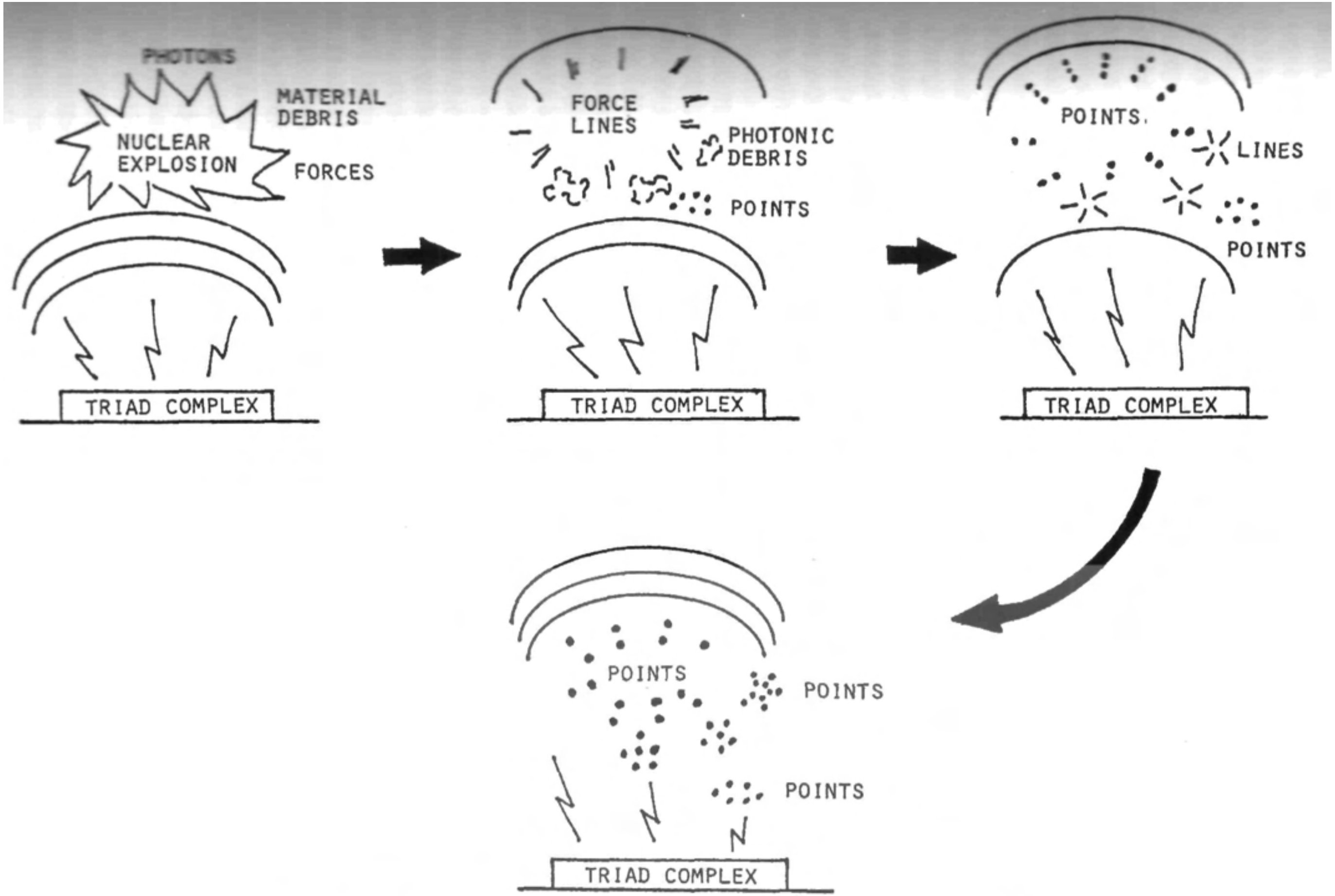
PHOTONS
MATERIAL DEBRIS
NUCLEAR EXPLOSION
FORCES
TRIAD COMPLEX
FORCE LINES
PHOTONIC DEBRIS
POINTS
TRIAD COMPLEX
POINTS.
LINES
POINTS
TRIAD COMPLEX
POINTS
POINTS
POINTS
TRIAD COMPLEX

Fig. 39. Sweep Mode: sweeping away a nuclear detonation

Finally, it is well-known in parapsychology that the personal unconscious can affect matter and change spacetime itself. Psychokinesis, poltergeist phenomena, etc. are direct examples. Hyperchannel calculations show that, for linked multiple minds, this psychokinetic effect increases exponentially with the number of stages or minds that are cohered. Since ZARG has 4 x 10^9 linked stages, this means that its psychokinetic/materialization/dematerialization powers are at least some 10^{10^9} times as great as those of, say, Uri Geller. Thus ZARG can pull the very stars out of orbit, disrupt the earth or solar system, stop the earth in its orbit, or change the fundamental laws of nature itself. The reality we live in is largely prescribed by ZARG, and substantial tuning of many minds can directly affect ZARG's tuning and hence alter reality. It is not accidental that Brian Josephson, Nobel prizewinner in physics, once remarked that it appeared that in their diligent searches for strange new particles in their accelerators and cyclotrons, the physicists might be creating their own reality. In fact an accelerator or cyclotron is a kindling device par excellence, and a scientific theory by its very rigor is an exceptionally sharp mental concept or thought form. So if groups of physicists have worked diligently on a problem and predicted some new theoretical particle, they may well be able to kindle it into being if they persist long enough with their kindling mechanisms, because they are cranking up additional ZARG energy when they persist. While he knew nothing of the kindling mechanism, Josephson noted that many of the new particles were first obtained only after great difficulty, then with moderate difficulty, and thereafter they could be fairly readily obtained.

And as we have seen, ordinary physical matter occupies a universe or frame common to all minds of living biological systems. Thus the physical universe is quite literally simply thought that is common to all living minds, hence to ALLMIND or God or whatever other term one prefers to use. It is the fact that it is prescribed and represcribed, over and over, that lends a physical entity its seeming solidity as compared to the phenomena existing in only a single consciousness mindworld. This physical universe is mind, it is totally mind, and as such it can indeed be changed by the proper mind resonance in enough minds, or in powerful minds such as ZARG.

ZARG contains so many mindworlds or dimensions that it forms a very powerful part of the particular reality prescription of the physical laws exhibited by this physical universe. So materializations of tulpoidal forms, UFO's, bigfoot, and Nessies, are just a tip of the iceberg of what ZARG actually materializes. The tulpoidal forms are tiny surface ripples due to large-scale human conflicts.

And as we have also found, psychotronic energy is in fact pure mindstuff. Pure mindflux between mindworlds. And this brings us to another bottom line.

When the Soviets launch their long-prepared attack in Europe, wipe out China, and almost annihilate NATO forces (fifty percent casualties in three to four days), the U.S. President will undoubtedly be under extreme pressure to launch a strategic nuclear retaliation on the Soviet Union. Suddenly, in a single blinding display of almost unbelievable power, the Soviets apparently intend to destroy the U.S. triad weapons in a fantastic psychotronic blitzkrieg through hyperspace. The Minuteman sites, the SAC bases, communications centers, and other strategic targets will lie in smoking, devastated ruins in minutes, and within the same short minutes of that holocaust, the U.S. nuclear submarines on station under the ocean will begin rupturing like popping balloons. At the same time, massive swaths of death will invisibly rain from the Soviet satellites in the skies overhead, and giant earthquakes will erupt into being in the East, in the West, and in the heartland U.S. The British and French nuclear forces will suffer similar simultaneous obliteration. And the jubilant Soviet Generals and Party theoreticians will be dancing with uncontrolled glee as their dreams become realities. But what they will not know is that their own inevitable doom will have just been sealed.

For all these fantastic psychotronic explosions will constitute gigantic psychotronic "mental flashbulbs" that explode directly in the mind of the sleeping ZARG. ZARG will be kicked into startled semiwakefulness, stinging with pain as if from a giant bee-sting, figuratively gnashing his teeth and furiously lashing his tail. This will constitute an immediate and enormous upheaval in spacetime itself, and the bulk of the earth may well be destroyed in the vast psychokinetic upheaval unleashed by the startled ZARG. Just as a person may sharply slap the bee on his arm where it has stung him, ZARG's fury will be centered on the Eurasian continent, on the locus where the threat or sting has originated, on Russia and the Warsaw Pact nations. So in an unbelievable convulsion of the entire earth, the Eurasion continent will be rent and torn asunder by forces beyond man's comprehension. The lousing ZARG will also activate a totally new scale of reality, ripping to shreds this particular three-dimensional spatial frame, its associated stable time changes, and the physical laws based on that stable reality.

The rousing of ZARG will initiate and accomplish thedirect linkage of the conscious minds of humans into a single giant superconscious mind — extending into billions of dimensions. Spacetime and reality will have gone insane during the developing paroxysm; with the simple initial extension into just the fourth and fifth dimensions, humanity will suddenly discover that its entire world history from beginning to end will come into simultaneous being (fourth dimensionality), and then all possible such world histories will come into being (fifth dimensionality). Literally, dumbstruck humans will pray for death and not find it, the grave will "give up its dead," and the ceaseless flow of time will suddenly cease. And again Ihc "beast will be defeated and cast into darkness"; with all-linkage and all-unity, there will be no longer be any humanity of one person for another or of one group for unother, for there will be no other. In such manner, evil will end; there being no separation between the doer and the done to. Man will have become a true God — in a strange and bizarre way constituting the second coming of God.

So the Soviets are on a course to evoke the full carnage of something very like the Book of Revelation. John, looking across two thousand years of history into the future, and seeing the awful cataclysm that was to come, described it as best he could in the concepts available to his day and in the belief structure in which he functioned. Thus he described Armageddon and the end of time, and wrote the Book of Revelation.

Indeed it could have been accomplished without Armageddon, if humans had been wiser. Each photon and each electrical signal that wends its way across the earth's surface and through planetary space carries a psychotronic modulation of ZARG's mindstuff. And in the communications explosion we have experienced, we have been steadily lacing together the population centers of the earth with electrical energy and electromagnetic signals. And so over those crude "pre-nerves," the psychotronic connections of ZARG have been intensifying, and the Giant has slowly been being stirred from his sleep. The kundalini of the entire human race has

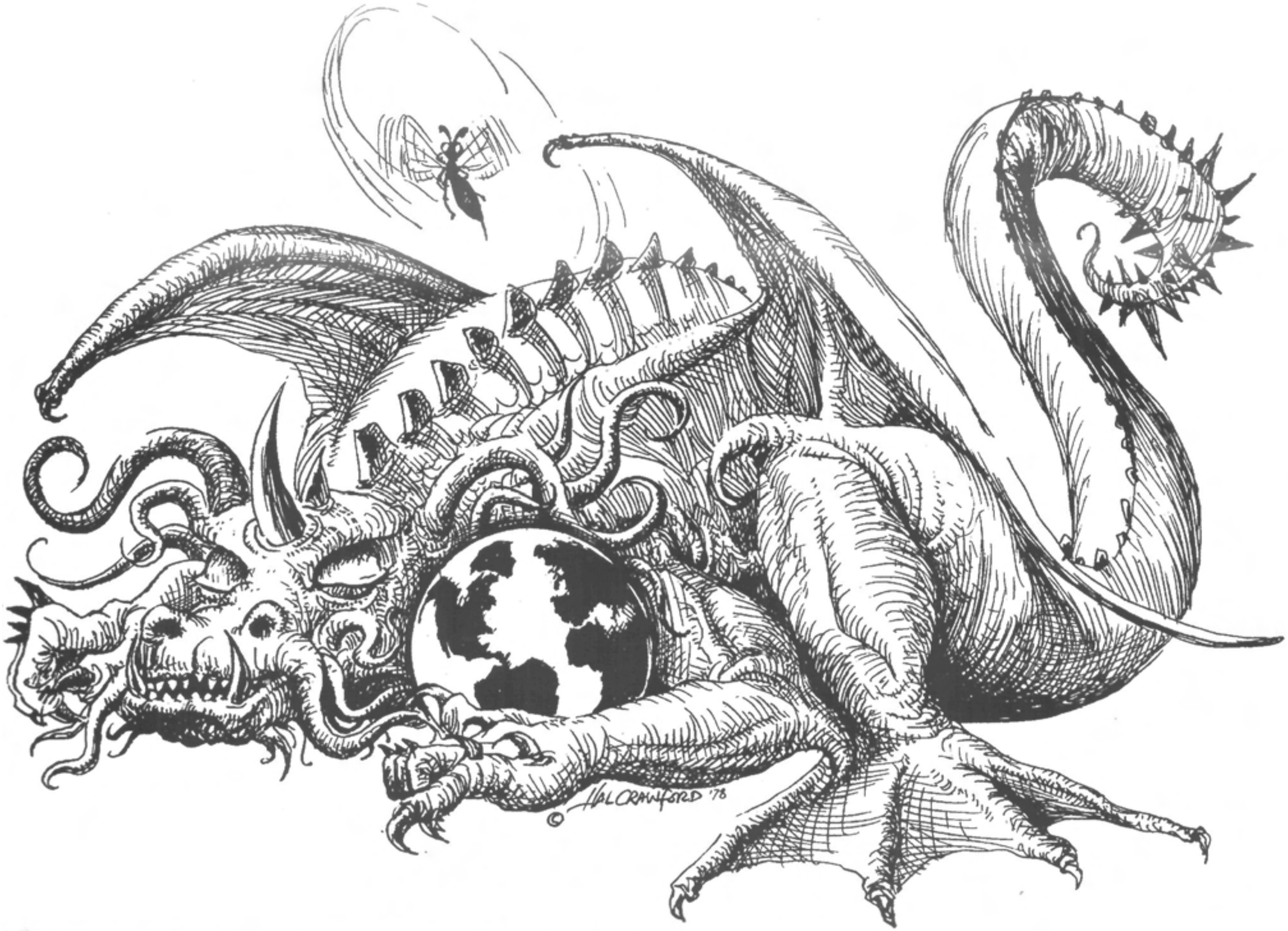
© HAL CRAWFORD '78

Fig. 40. ZARG

slowly been rising, and paranormal phenomena have been increasing, as have changes in reality such as the kindling of new fundamental particles. Even the sun has responded, suddenly decreasing its temperature. If we had applied psychotronics to develop a mind linkage mechanism, there would have been no need for Armageddon's destruction.

Let us hope we fare better than the Atlanteans. One may speculate that an Atlantis of sorts did exist, and that the Atlanteans built psychotronic devices out of crystals, bronze, gold, silver, and wire and unleashed them on each other in war. This sting stirred ZARG mightily, and in a resulting frenzy it simply destroyed the continent, stirring floods, storms, earthquakes and volcanoes worldwide. In the case of the Atlanteans, only a relatively small portion of humanity was involved; the percentage of persons and minds involved was apparently insufficient to cause full wakening of ZARG and linkage of the species. In the present case, the greater part of mankind is involved, and so the wakening is to be far more complete (as well as more violent). ZARG should be shocked into full wakefulness, and linkage should occur.

If so, a new great superbeing will have emerged; nonphysical, almost infinitely dimensional, and godlike. It will not exist in a single time or a single world, but in a nowness and allness that defies timelike or spacelike description.

"For we know in part, and we prophesy in part.

But when that which is perfect is come, then that which is in part shall be done away.

When I was a child, I spake as a child, I understood as a child, I thought as a child; but when I became a man, I put away childish things.

For now we see through a glass, darkly; but then face to face: now I know in **part;** but then shall I know even as also I am known."

I Corinthians XIII, 9-12.

So the pieces of the puzzle are now fitted together. The present state of the human species is not its final development, and indeed a great change in that state is imminent. Man literally is and shall be a god, one with God, just as the prophets have foretold. And if we retranslate that ancient foretelling into more modern concepts, we see just how strange and transcendent is the final birth or becoming.

The UFO and tulpoidal phenomena have merely been the signs in the sky and in the earth, telling us what is to come, precognitive indicators from ZARG. Had we developed a metapsychology of the entire species, we would already have clearly understood the messages.

VIRTUAL STATES AND HYPERSPACES

In physics today there are two separate kinds of reality: observable changes, which can be detected by instruments; and virtual changes, which cannot be detected by instruments, even in theory. The difference between these two realities depends simply on the size of the piece of action (energy x time, or momentum x length) making up the particular entity or change. Usually the absolute minimum observable change is taken to be $h/4\pi$. Changes smaller than this are said to be "virtual" and undetectable, and changes larger than this are said to be "observable" and detectable. Virtual changes are not bound by the conservation of energy law and almost always violate it. Nor are virtual changes, bound by the conservation of mass law. And virtual photons can have nonzero mass.

For any instrumental condition in a detector, one may consider a certain momentum and a certain length to be involved in the mass change that spatially constitutes the detection. Let us call this critical length dL and examine its ramifications (Figure 1) for the finest possible detector—one that detects a single quantum change. Figure 1 shows two frames, the lab frame S and a moving (hyperspatially rotated) frame. We choose a moving object, M, in the moving (rotated) frame, with length as shown. Note that all these entities are to be represented as detected or seen by our physical detector (instrumentation system) in the laboratory frame.

As the velocity of M approaches c, the speed of light in a vacuum, M's hyperspatial rotation angle α approaches $\pi/2$. However, before the velocity of M actually reaches c—and before α reaches $\pi/2$—the dimensional length projection of object M in frame S will contract to reach dL and fall below it. At that moment, L becomes *virtual* with respect to the detector and hence undetectable. Since geometrically an absence of length in a continuous frame S would correspond to a α rotation angle of $\pi/2$, in quantized frame S we consider that angle α "jumps" to $\pi/2$ just immediately past the critical angle α_c, which is defined as the angle of rotation at which L cos α = A L. Rotation just past the critical angle corresponds to a Lorentz-Fitzgerald contraction into the virtual state below the quantum threshold of the detector's observable universe. To an S observer or detector, object M at that point has lost its length in the direction of motion, becoming a wave front moving at the velocity of light and thus photonic (comprised of photons). To a quantum observer in S, mass turns into photons just short of a full $\pi/2$ orthorotation, which means the process can happen with a finite amount of energy. This "jump" effect is not taken into account in the continuum Lorentz transformations of special relativity at present, and this is the missing link that has prevented a successful union of relativity and quantum mechanics. In other words, the final part of the orthorotation (acceleration to the speed of light) occurs in the virtual state—specifically, that part of the rotation lying between $\pi/2$ and α_c.

Further, we may now proceed to rotate object M past $\pi/2$, which constitutes rotation from zero in frame S'. Just before reaching an S angle of π (that is an S' angle of $\pi/2$), in frame S' another disappearance below a virtual length threshold dL' will occur in corresponding fashion. Note that this second threshold dL' also has an exceedingly small but finite projection into frame S, in

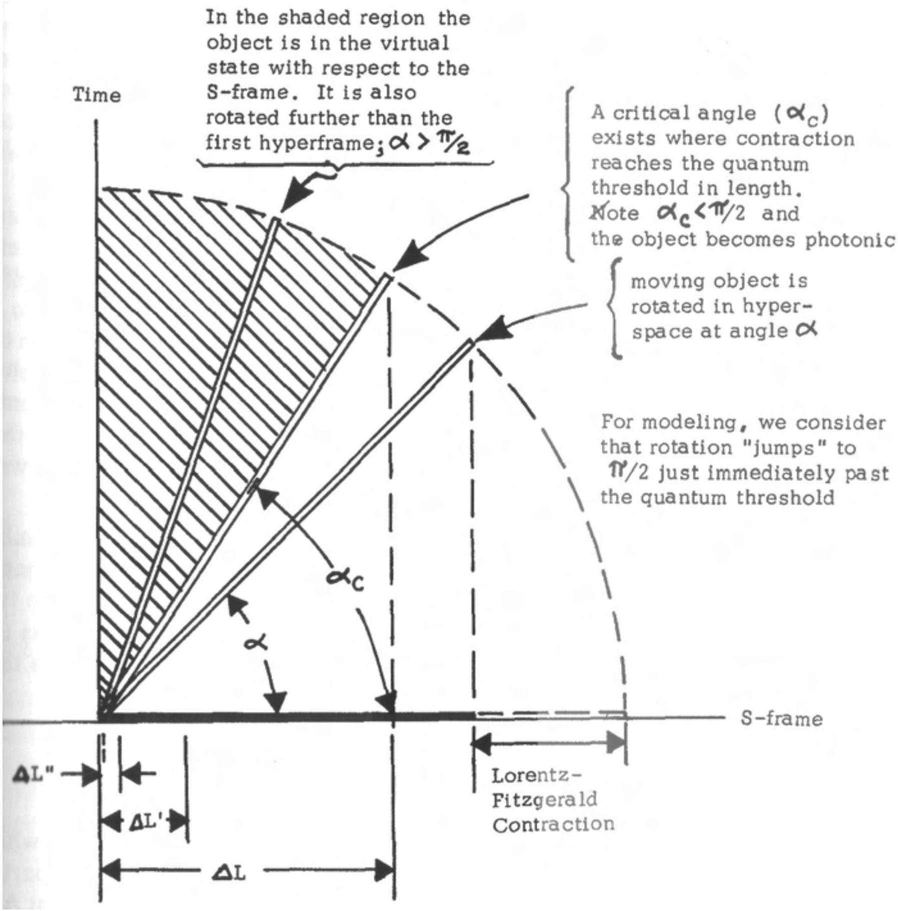

In the shaded region the object is in the virtual state with respect to the S-frame. It is also rotated further than the first hyperframe; $\alpha > \pi/2$
Time
A critical angle (α_c) exists where contraction reaches the quantum threshold in length. Note $\alpha_c < \pi/2$ and the object becomes photonic
moving object is rotated in hyperspace at angle α
For modeling, we consider that rotation "jumps" to $\pi/2$ just immediately past the quantum threshold
α_c
α
S-frame
$\Delta L''$
$\Delta L'$
ΔL
Lorentz-Fitzgerald Contraction

Figure 1. Hyperspaces, nested virtual states, and rotated frames

the virtual state lying inside dL. This second "relativistic quantum jump" in frame S' therefore constitutes a sort of second-level, nested subquantum threshold in frame S.

Continued rotation into hyperspace passes through successively finer, nested virtual sublevels or subthresholds in S by the successive breaching of deeper subthresholds due to continued Lorentz-Fitzgerald contraction. Higher hyperspatial frames constitute nested sublevels in the virtual state with respect to the laboratory observer. Thus a direct connection exists between virtual levels and hyperspace frames. Hyperspaces and virtual sublevels are the same thing. Every "virtual" object in S is an "observable" object in some hyperspace. *Hyperspatial engineering and virtual state engineering are one and the same thing, and this is what psychotronics and energetics are all about.*

Finally, time may be considered a sort of box containing all observables and virtuals. Indeed, time is neither observable nor virtual; time is simply a parameter in quantum mechanics. Rigorously, time may be said not to exist. Time itself is nothing, totally nothing; therefore by the fourth law of logic it can be said to be or contain everything, the total mishmash of virtuals and observables. Since in the absolute sense there is no such thing as time, then it is my thesis that all modeling of space, mind, thing, and nonthing must be done spatially. Relativity presently attempts to use time as an existent dimension of the same kind as length; it thus allows rotation of the S-frame only in the direction of the time axis in Minkowski geometry of spacetime.

Spatial modeling requires an infinite dimensional space and an infinite number of orthogonal angles. The time axis that can be drawn in such a hyperspatial model has no meaning except with respect to a relationship existing between two or more hyperspace frames or objects. Specifically, the time axis represents the totality of all the hyperspaces and their objects, hence all the relationships that can be separated from the hyperspace.

FEYNMAN DIAGRAMS

Figure 2 shows something of the unsuspected nature of physical change, which is not nearly as simple as most of us have been led to believe. In the top part of the figure, an electron (represented by the arrow) is shown going from point A to point B a short distance away. In particle physics this diagram represents the movement of a bare particle; in other words, just a bare hunk of charged mass is considered and the fantastic number of other things happening along the way are ignored. In Figure 2b a virtual photon has been emitted and then reabsorbed by the electron between points A and B. Indeed, the charge on the electron is due to the continual emission and absorption of such virtual photons (little electromagnetic pieces of vacuum, one might say, or little pieces of nothing). So the situation with respect to virtual photons comes closer to Figure 2c: more than one virtual photon is emitted and absorbed—indeed, an infinite number of them are emitted and reabsorbed along the way from point A to point B.

But, worse yet, Figure 2d shows one of the emitted photons suddenly breaking up into a positron and an electron, which then quickly recombine back into an-

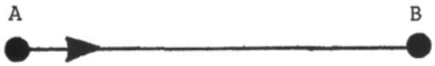

2a. Bare electron

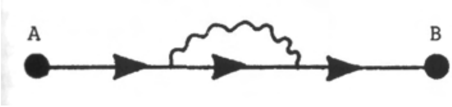

2b. With virtual photon

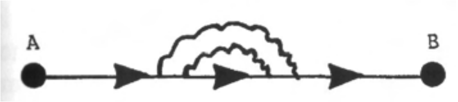

2c. With nested
virtual photons

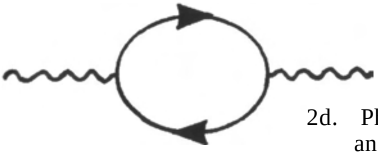

2d. Photon decay (pair production)
and recombination (pair annihilation)

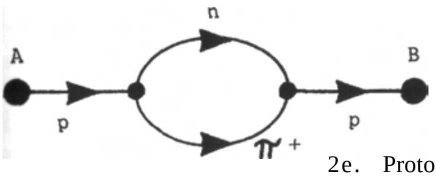

2e. Proton changing briefly to a neutron
and a positive pi meson

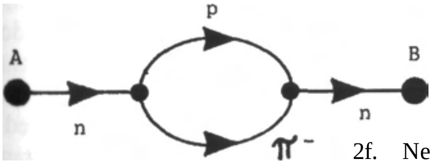

2f. Neutron changing briefly to a proton
and a negative pi meson

Figure 2. Feynman **diagrams**

other photon, which is eventually reabsorbed into the electron. In fact, an infinite number of interactions happen while the "electron" is moving from point A to point B, and each of those interactions in turn consists of an infinite number of interactions, and each of those in turn consists of an infinite number of interactions, and so on. The simple movement of the electron is actually composed of a virtual cloud—an infinite number of an infinite number of an infinite number of things, all happening in a mad tangle of nested substrata in the virtual state (Figure 3)—and each of these virtuals with respect to frame S is in fact an observable in some particular hyperspace! Many other sorts of virtual particles are emitted and absorbed and born and extinguished, in addition to photons, by various observable particles and in free space. Everything in every possible world or hyperspace is related to each and every other thing in the laboratory world through orthorotations and nested virtual states. Crosstalk between hyperframes is the emergence and disappearance of virtual particles in a particular frame. The universe is holographic after all, and Mach's principle holds in a far more complete fashion than Mach himself ever dreamed. And if any part of this virtual chaos can be collected and cohered in even the slightest degree, any amount of observable free energy can be obtained, just as T. Henry Moray showed in the 1920s and 1930s.

In passing, I would like to call attention to Figure 2e, which shows one of the virtual interactions a proton can undergo in the nucleus of an atom. In this case, the proton breaks momentarily into a neutron and a positive pi meson, which then recombine into a proton. However, if the charge on the $\pi +$ particle is extinguished psychokinetically during the existence of the neutron and the pi meson, the pi meson is converted to a neutrino and radiates away. In this case the proton does not reappear and the neutron remains. This decreases the atomic number (positive charge) of the atom by one, transmuting the atom into a new element, an isomer down the chain of elements. Figure 2f shows one of the virtual interactions a neutron can undergo in the nucleus of an atom. In this case, the neutron breaks momentarily into a proton and a negative pi meson, which then recombine into a neutron. However, if the charge on the $\pi -$ particle is extinguished psychokinetically during the existence of the proton and the pi meson, the pi meson is converted to a neutrino and radiates away. In this case, the neutron does not reappear and the proton remains. This increases the atomic number of the atom by one, transmuting the atom into a new element, an isomer up the chain of elements. The two interactions shown in Figures 2e and 2f are involved in the Kervran effect, where elements are transmuted by living biological systems.

But back to our electron. If we now remove the electron, leaving empty vacuum, all the sorts of virtual madnesses we have built up in Figure 3 are still there and still occurring. In addition, every other sort of virtual madness is also there and occurring. Vacuum is infinitely dense and filled with stupendous "zero-point" vacuum energy; it is totally chaotic and virtual. Space itself is an infinite tangle of virtuals—virtual projections from real observables in associated hyperspaces—and these dynamic interactants make up any three-dimensional space and hyperspace itself! They are also what cause space itself to exhibit the finite, nonzero parameters or values that must be accounted for in our physical equations. *Every fundamental constant of nature is dynamically constructed from—and*

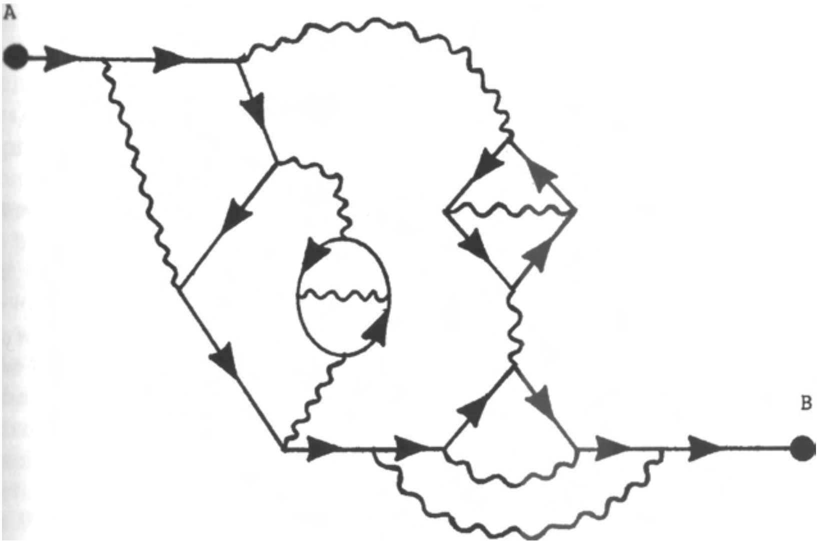

An electron moving from point A to point B is comprised of a bare particle and an infinite tangle of nested virtual particles

Figure 3. Nested virtual processes

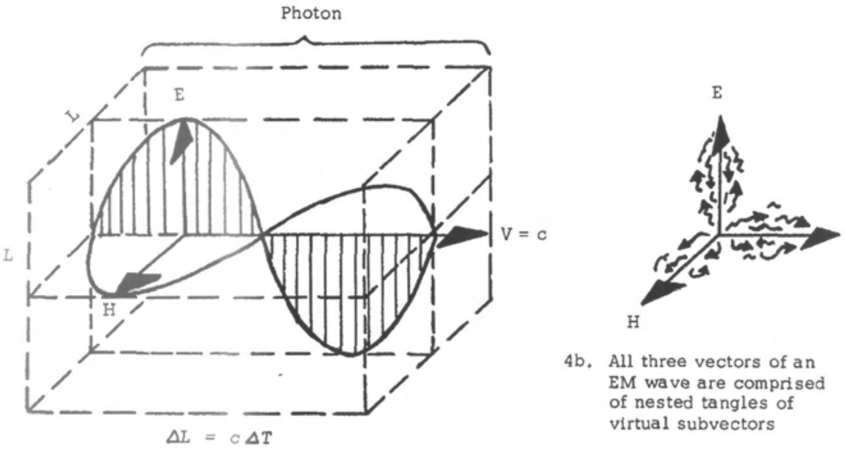

4a. A photon may be considered an individual L^2 wavefront spread in time through one oscillatory period of the E-field and the H-field

4b. All three vectors of an EM wave are comprised of nested tangles of virtual subvectors

Figure 4. Nature of the electromagnetic wave

due to—the zero-point, virtual vacuum interactions that constitute "empty" space. The virtual "sea" of seething waves and particles that is empty vacuum is actually an ether—a Lorentz-invariant, nonmaterial (virtual) ether, perfectly admissible by the Michelson-Morley experiment and the entire experimental basis of physics. Indeed, Einstein once proposed naming the vacuum and its contained fields the "ether," but his suggestion was not heeded.

NATURE OF AN ELECTROMAGNETIC WAVE*

We now know something quite unusual about an electromagnetic wave and about a photon. Figure 4a shows how a photon may be represented. Here we find that a photon is a single oscillation of the E-field vector and the H-field vector as shown in the diagram. Since the oscillation is with respect to time, the wave moves along the axis of velocity in the S-frame for a single wavelength, then repeats itself again and again. The E- or H-oscillation actually represents an orthogonal, rotating etheric or hyperspatial flux about the velocity vector; this spinning (exploding and imploding) flux is coming into the S-frame from outside it (from virtual states) and returning (into virtual states). It is a neutrino flux, as we shall see. In the photon, half a sine-wave cycle represents the fission action that splits the action quantum into canonical pairs, while the second half of the sine-wave cycle represents the fusion of the canonical fragments back into an action quantum. One half-cycle thus represents a positive time operation or creation, while the other half-cycle represents a negative time (fusion) operation. If the complete sine-wave cycle has sufficient amplitude (energy), it can sometimes be broken in half, with the positive time portion revealing the negative field (negative charge) of the E-vector, while the negative time portion reveals a positive field or charge of the E-vector. This is pair production, the creation of an electron and a positron. Each of these fragments has spin one half, while the photon quantum has spin one. Since the positron is a fusion interaction existent, it is an "electron traveling backward in time," just as predicted by Dirac.

In Figure 4b I show that both the E-field vector and the H-field vector are composed of an infinite number of virtual vectors, each of which is composed of an infinite number of even smaller virtual vectors, and so on. This follows from the fact that during any time interval—including the period of a single oscillation of a photon—a photon may be considered to consist of an infinite number of virtual happenings, each consisting of an infinite number of subhappenings, and so on. The photon itself may then be considered to have a direct, nested virtual substructure, and this substructure is directly hyperspatial in nature as well. This structure has not been considered in Maxwell's field equations, which have been

---

***Actually the electromagnetic field is not the most fundamental field, but is something that is torn out of, or separated from, a more fundamental field: the vector potential, or A-field, well-known in particle physics but usually ignored in EM field theory. E.g., magnetic Held is created (and defined) by $\nabla \times A = B$. Looking at the cross product, by definition B field is the line integral about a small surface of the A-field flux through that surface. That is precisely what a single coil of wire, with an electrical current flowing through it, does and therefore creates a magnetic field inside. See the second volume of Feynman's three volumes of physics for more on the A-field. The direct use of the A-field is also psychotronic engineering.**

formed from gross macroscopic considerations of bare particles and continuous fields. Maxwell's equations (and the normal concept of an electromagnetic wave and single frequency) thus represent only gross envelopes of the actual form and structure of waves and fields. As I have already shown in the main body of this book, one can even photograph the hyperfield structure of a magnetic field using a simple bar magnet and a color TV tube. Further, we will find that this virtual structure of a photon can be altered, "modulated," or changed by simple signal modulations, among other methods. The only requirement is that we learn to regard modulations differently.

In addition, inside each half-cycle of a photon's E-vector, relatively instantaneous fluctuations and "losses" of the vector's magnitude occur. At these "E-field glitches" in the photon, it actually contains only an H-vector. In this case the photon is a neutrino momentarily—or, in other words, it will admit a neutrino through the E-field glitch. The greater the charged field (voltage potential), the greater the number of photons available, thus the greater the number of neutrino interactions within the field. A potential of about 200-300 volts per meter altitude above the earth's surface normally exists in our biosphere. Many neutrinos emitted by the sun and bombarding the earth are absorbed in or interact with this E-potential above the earth; hence underground experiments to measure the neutrino radiation from the sun have detected less neutrinos than expected.

Momentary H-field glitches exist in a similar manner and interact with or admit neutrinos in a similar fashion.

Light can thus conduct neutrinic waves containing bioenergy or eloptic energy, as T. G. Hieronymus discovered. In the first hyperframe (EM field frame), this can be seen to be simply a space charge of electrons conducting a wave disturbance through it, or a piece of ordinary conductive matter conducting an ordinary EM wave.

NEUTRINIC LONGITUDINAL POLARIZATION WAVES

In Figure 5, I assume that in some fashion the E-field of an electromagnetic wave has been stripped off, leaving a single vector in the H-field direction only, and this vector is still oscillating, as originally set up in a Hertzian wave in Figure 4. The new "single-field" wave is an etheric sort of half-wave, a longitudinally polarized wave that is not a wavefront (plane) moving through S-space but just a pulsating line segment moving through S-space. This wave is not Hertzian. It is not electromagnetic—though it will strongly interact with or even produce electromagnetic waves—but is neutrinic. The interactions of this wave with space virtuals are drastically reduced (indeed, by a factor of c), and this neutrinic wave now travels at the speed c^2 with respect to S-space. While it will highly interact with magnetic field, precisely speaking it is not a magnetic wave but a hyperfield

---

*C^2 is its s-frame, free-space velocity when not interacting with electromagnetic field. Normally, neutrinic wave interacts with electromagnetic field, thus moving at less than C^2 velocity in space, since it "circulates around" in electromagnetic field. Neutrinic wave can be regarded as a circulating or spiraling wave in electromagnetic field, which may well be the

wave. It will travel through and across dead shorts, through media such as seawater and earth, and acts or can act much as a neutrino. This wave has neutrinic components in S; that is, it has vector components existing in frame S". One may expect this wave normally to be active in or interacting with an electrostatic field or a magnetic field, as a special sort of lower-level virtual state flux (second-order hyperflux) into and out of the field. This is apparently the sort of wave discovered and used by Nikola Tesla, who repeatedly insisted that he was not using Hertzian (Maxwellian) waves but instead was employing a much finer wave having very different characteristics. Further, neutrinic waves will transit between any two points on earth almost instantly—in fact, about 3×10^8 times as fast as the speed of light propagates! With these waves, space communication with distant galaxies is practical and achievable, and scientists seeking communication with extraterrestrial civilizations would do well to use such waves.

Note that I am not referring to the ordinary longitudinally polarized electromagnetic wave, where both the E- and H-vectors still exist but the H-vector is rotated 90° so as to lie in the line of motion through S-space.* Such longitudinally polarized electromagnetic waves are already known in the literature and have been used in many patented devices, but they are not included in most textbooks on electromagnetic fields and waves for some unaccountable reason. The normal longitudinally polarized electromagnetic wave moves at the speed c, just as any other EM wave.

At any rate, neutrinic longitudinally polarized waves often arrive at a distant point many times stronger than they were when they left. T. Henry Moray, for example, built a radio and a sound detector that apparently worked by a version of neutrinic waves, condensed into electromagnetic waves in his radio-detector apparatus. Dr. Moray tested the radio scores of times between 1930 and 1933, and on several occasions even received broadcasts from Admiral Byrd's party in the Antarctic. Even when powerful, ordinary radios could receive nothing from Antarctica because of severe weather conditions, Moray's little batteryless radio brought the broadcasts in, loud and clear with no static. Further, Moray demonstrated a sound pickup device of similar construction that could pick up or zoom in on sounds five miles away. Again the sounds were loud and clear, with none of the distortion of weather conditions and background noise. (See T. Henry Moray, *The Sea of Energy in Which the Earth Floats,* 4th ed., 1960, pp. 129, 133, and 143.) Thus transmission of power without wires was possible, just as Nikola Tesla had shown. Tesla in fact patented a capacitive (E-field only) transmission system.

Indeed, according to independent research by Dr. Robert Beck, another inventor named Rogers apparently discovered just such a system and the United States

"spinning wave" of orgone energy identified by Reich. Neutrinic waves particularly interact more strongly with strong E-field and with strong B-field, the degree of interaction increasing (probably nonlinearly) with the intensity of the E or B field with which the interaction occurs. The velocity of neutrinos, measured on earth inside its strong ionosphere-to-earth E-field, should be quite reduced from the velocity of neutrinos measured in isolated deep space, and almost two-thirds of the neutrinos entering the earth system from space interact with the ionospheric-earth potential and are absorbed by it. The A-field and its derivatives directly involve the E-field, H-field, and neutrinic field and their interactions. To correctly understand the activation chemistry theory and results of LeBon, A-field theory is required. And to understand much of Tesla's work, A-field theory is required, for Tesla was aware of, and used, A-field interactions in many of his patented systems.

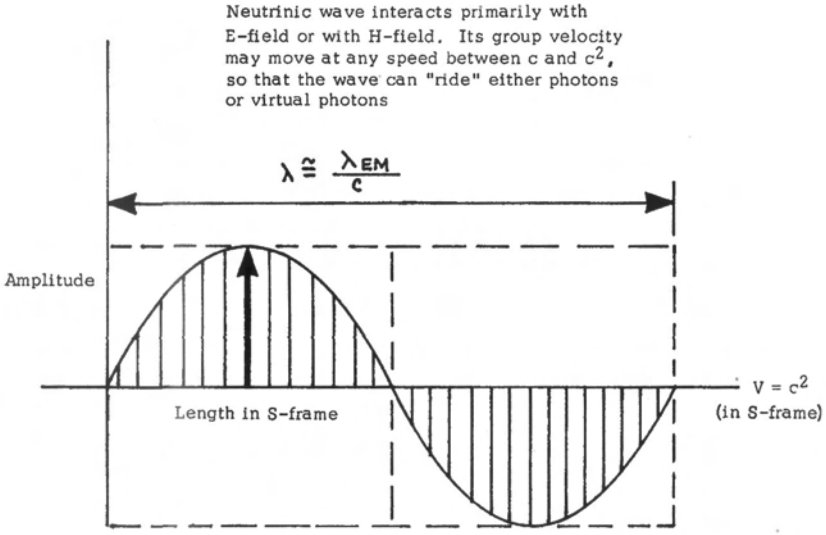
Neutrinic wave interacts primarily with E-field or with H-field. Its group velocity may move at any speed between c and c², so that the wave can "ride" either photons or virtual photons
λ ≅ λEM / c
Amplitude
Length in S-frame
V = c²
(in S-frame)

Figure 5. Neutrinic (longitudinal polarization) wave

NOTE: This is not an electromagnetic wave and is not the longitudinally polarized EM wave familiar to many physicists and engineers.

used it to communicate with its submarines underwater during World War I! For that purpose it was known as the Rogers underwater and underground wireless system, and H. Winfield Secor in 1919 reported on its use and its fantastic potential for the future, implying the necessity to completely change our ideas of electromagnetics and power transmission. (H. Winfield Secor, "The Rogers Underground Wireless," *Electrical Experimenter*, March 1919.) Mysteriously, the Rogers wireless seems to have promptly disappeared. Even more strangely, according to Dr. Beck's research the same breakthrough has been made again in the U.S. at least three times since the beginning of World War II. These rediscoveries included full delineation of the prior technologies, with mathematical theory and diagrams, that had escaped further notice. Again the technology disappeared mysteriously in all cases shortly after being rediscovered. The Soviets did not have far to look to discover the Rogers Wireless, Tesla waves, T. Henry Moray's radiant energy device, the Moray radio, the Moray sound detector, and the U.S. technical work in the area since World War II. All they really needed was to thoroughly examine the U.S. scientific and technical literature, the technical magazines and books, and so on. The Moray amplifier and the Moray sound detector were also demonstrated by Moray to a Soviet scientist, Dr. A. A. Yakovlev, in October 1929, and a Soviet agent succeeded in penetrating the Moray laboratory in 1939, obtaining drawings of the Moray tube and many photographs, and perlonally working in the laboratory for two months.

One may consider power transmission through such a wave to have traveled to every point in the future path at a speed that is faster than the flow of macroscopic time by a factor of c. Then along any electromagnetic (time) line along the same path, these waves are "ringing" back and forth continually, as the progress of normal physical change crawls slowly forward. Thus the normal time change (electromagnetic or photonic interaction world flow) acts as a collector for the ringing neutrinic energy, and this collection can act as a gain factor G, where theoretically G is given by the expression

$$1 < G < c.$$

Thus a longitudinal polarization wave, when transmitted to a distant receiver, may arrive or be detected at the receiver thousands of times stronger than the transmitted signal, even for very low coupling.

Neutrinic longitudinal polarization aspects to ordinary electromagnetic waves can also be achieved by modulation—especially by second-order, nested modulation—and such hyperspatial aspects have already been detected in the Soviet communications radiation, causing interference around the globe in the 10 to 20 megahertz range, as we shall see.

A NEW LOOK AT MODULATION

We can assume a sine-wave electromagnetic carrier frequency whose amplitude we wish to vary (modulate) as a function of time, by a lower frequency sine wave, as shown in Figure 6a.

As is well known, the classical view of amplitude modulation of a carrier wave holds it to be an oscillatory change in the amplitude of the carrier, as shown in Figure 6b. However, empirically this was found to be erroneous, since in 1929 it was shown that after such modulation the detected carrier appeared unaffected by

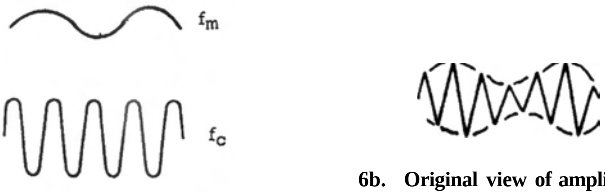

6a. A carrier f_c to be amplitude modulated by frequency f_m

6b. Original view of amplitude modulation

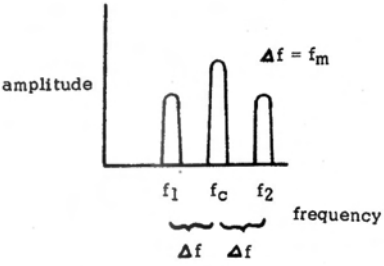
amplitude
$\Delta f = f_m$
f_1
f_c
f_2
frequency
Δf
Δf

6c. Detected signal for an amplitude modulated wave

Figure 6. Amplitude modulation of a sine wave

the modulating wave; instead, there appeared two other sine-wave frequencies (sidebands), one at a frequency above the carrier frequency and one below it, as shown in Figure 6c. In each case the change in frequency from carrier to sideband was by an amount equal to the frequency of the modulating wave.

Both the original classical view and the present view are in fact correct if one considers hyperspatial aspects and linear detection aspects. As can be noted in Figure 6b, the carrier is distorted and is not a sine wave according to the original classical view. This implies that the individual photon shown in Figure 4a has been internally distorted or "bent." This, I submit, actually happens and constitutes the bending or curving of spacetime in a general relativistic fashion. However, all our instruments can detect only a pure, undistorted photonic interaction; that is, in frequency they can only detect the sine-wave aspects of a kinked photon. The reason is quite simple: for example, suppose the method of detection of the photon is by photon interaction with a bound electron. The change in the electron is a precise change of energy levels to an excited level, and it only detects or interacts with a single sine-wave photon. The frequency (say, the carrier) component of the photon, which contains the correct energy differential (according to $E=hf$, where h is Planck's constant and f is the sine-wave frequency), will interact with the electron. The other sine-wave components of the initially kinked photon will thus be stripped off (demodulated) by the absorption of the carrier portion of the photon. The photon kink existed in both positive time and negative time; from a frequency viewpoint, the modulating amplitude frequency that was added to the basic carrier frequency in positive time was subtracted in negative time. Thus demodulating the kink from the amplitude modulated photon (by absorbing the sine-wave carrier) results in the production or liberation of two sine-wave frequencies, one above the carrier frequency and one below the carrier frequency, each difference being equal to the original amplitude modulation's frequency. These two frequencies will then interact and be detected.

So our linear instruments do not detect single kinked photons in a stream of distorted photons, but they detect a combination of undistorted photon streams (waves) at different frequencies.

Such considerations as these should have appeared in electromagnetics long ago. After all, we presently regard an EM wave as a wiggle of nothingness, existing in time. This is precisely the same as saying that a photon is an oscillation (wiggling) of spacetime itself. Specifically, it oscillates in two directions spatially, one being the E-field direction and the other being the H-field direction. But since an EM wave itself constitutes a regular bending and flexing and curving of spacetime, including both space and time, then the flexing spacetime inside the photon can itself be further bent or twisted, and that constitutes a kinked or curved photon—a *modulated* photon. Our linear laboratory space instruments, however, will demodulate such a kinked photon into its linear components, giving a combinationof unkinked photons in the detection. And for the amplitude modulated carrier shown in Figure 6b, we actually detect the carrier and two sidebands shown in Figure 6c.

Similar complex mixes of waves at different frequencies result for other types of modulation, but this will suffice for our purposes here. Specifically, we should mention that an infinite number of sidebands may be produced, but normally all

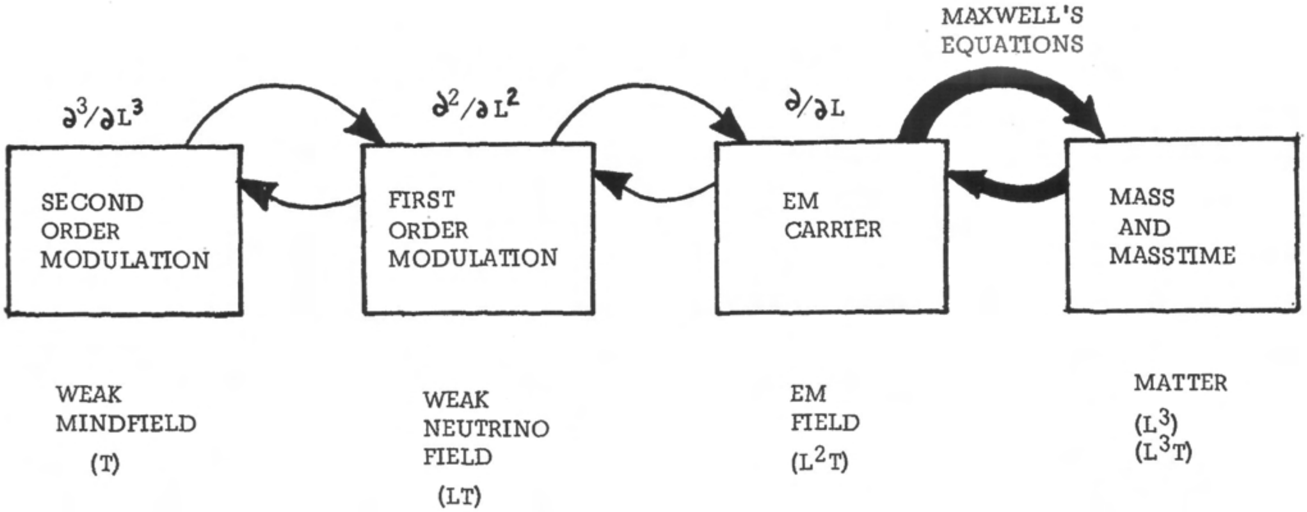
MAXWELL'S EQUATIONS
∂³/∂L³
∂²/∂L²
∂/∂L
SECOND ORDER MODULATION
FIRST ORDER MODULATION
EM CARRIER
MASS AND MASSTIME
WEAK MINDFIELD
(T)
WEAK NEUTRINO FIELD
(LT)
EM FIELD
(L²T)
MATTER
(L³)
(L³T)

Figure 7. Nested modulations: key to hyperspace

but a few are so weak as to be negligible for ordinary communication purposes, where no hyperspatial information is needed. For living systems, however, a great number of these extra sidebands must be taken into account, for the hyperspatial-virtual state components carry the lifestream, consciousness-loop-interaction Information itself. We will look at some of the biological aspects shortly.

At any rate, modulation of a wave results in direct changes in the virtual structure of the formerly unmodulated photon, relativistically warping or curving it. Just as Fourier (harmonic) analysis breaks down a very complex nonsinusoidal wave shape into an infinite series of component sine-wave frequencies of various amplitudes, so linear detection in a linear laboratory frame breaks down a stream (wave) of complex nonsinusoidal photons into a series of sinusoidal waves in a complex mixture.

And if we now think of this in reverse, we see that a mixture of waves and frequencies must automatically possess, carry, or represent hyperspatial-virtual patterns in the correct nonlinear medium that absorbs them.

HYPERSPATIAL ASPECTS OF MODULATION

Let us now look at the hyperspatial aspects of modulations and complex electromagnetic waves, particularly mixtures of waves, in yet another fashion.

First, as is well known, we may represent a wave as a sine function (or cosine function if preferred). When we modulate a wave, we impress a time change on it. In the most general mathematical sense, a modulation to a wave is the impressing of a derivative function upon the wave. And if we think of the basic carrier wave as being prescribed by Maxwell's equations in derivative form, then in a sense when we modulate the wave we differentiate it, that is, we impress a differential operator function on it. But this means that we raise the order of the differential aquations prescribing the situation, which mathematically means we have increased the dimensionality or spatiality of the, wave entity that is the resulting modulated carrier. In other words, to modulate an electromagnetic carrier wave is to add a hyperdimensional aspect—specifically, neutrinic vector aspects or mind aspects. The whole affair can be thought of as merely shifting or altering the particular distribution of nested virtuals transported by the basic carrier wave.

If we differentiate a sine wave, we simply get a cosine wave, which is just another sine wave pi/2 out of phase with the first. So now it emerges that when we form a mixture of waves, we are adding virtual aspects (and hyperspatial aspects) in so doing, because of the phase differences. This is extremely important: *complex mixtures of waves automatically constitute and carry hyperspatial components!* And if we can accurately control these hyperspatial components, we are in effect doing direct hyperspatial engineering and virtual state engineering.

Now if we carry our modulation-constitutes-differentiation concept one step further, we find a most startling phenomenon (Figure 7). The first-order modulation we placed on an EM carrier we discovered to carry aspects of the next hyperspace, namely, neutrinic vectors. Now if we further modulate that first-order modulation signal as a carrier itself, so that we now have two orders of modulation upon the basic carrier, the second-order modulation must possess hyper-

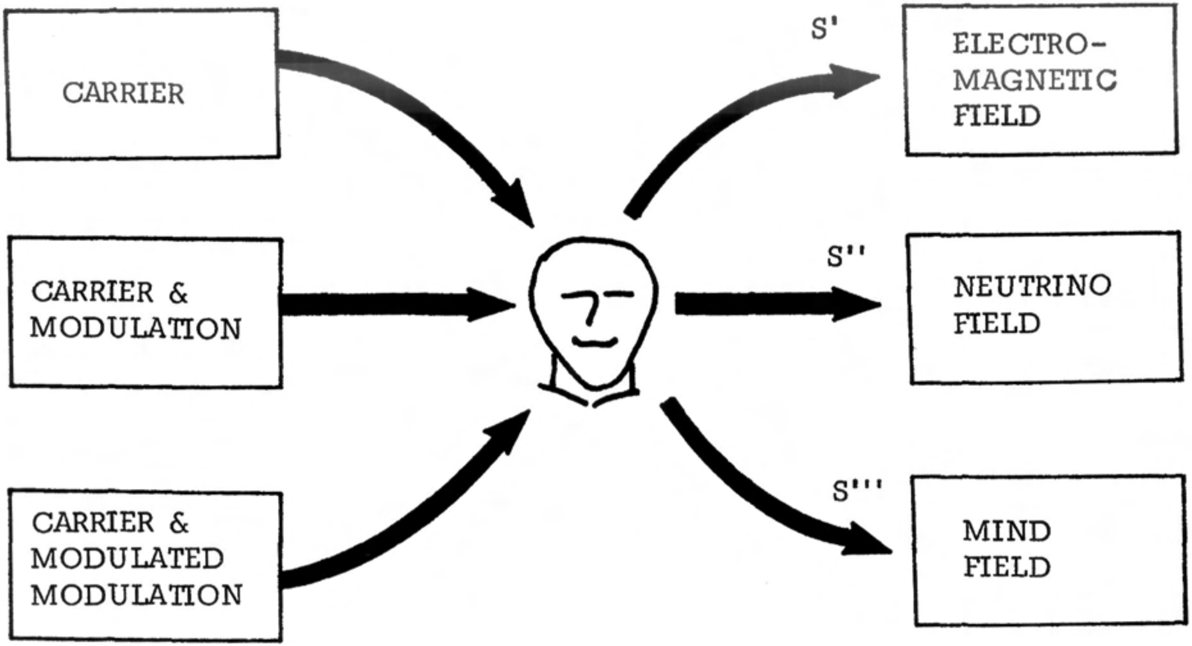
CARRIER
CARRIER & MODULATION
CARRIER & MODULATED MODULATION
S'
ELECTRO-MAGNETIC FIELD
S''
NEUTRINO FIELD
S'''
MIND FIELD

Figure 8. Field coupling to the human brain

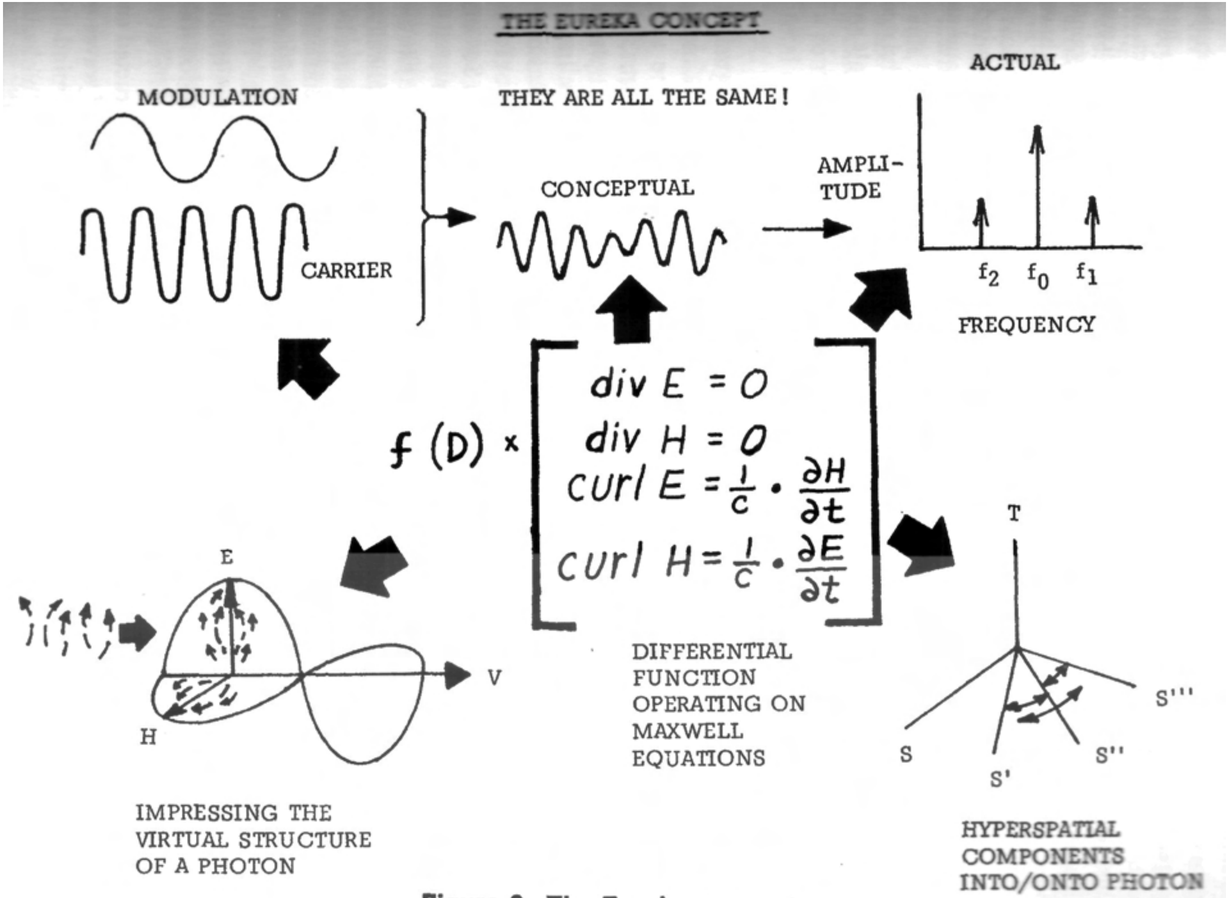
THE EUREKA CONCEPT
MODULATION
CARRIER
THEY ARE ALL THE SAME !
CONCEPTUAL
ACTUAL
AMPLI-TUDE
f_2 f_0 f_1
FREQUENCY
$f(D) \times \left[\begin{array}{l} \text{div } E = 0 \\ \text{div } H = 0 \\ \text{curl } E = \frac{1}{c} \cdot \frac{\partial H}{\partial t} \\ \text{curl } H = \frac{1}{c} \cdot \frac{\partial E}{\partial t} \end{array} \right]$
DIFFERENTIAL FUNCTION OPERATING ON MAXWELL EQUATIONS
E
H
V
IMPRESSING THE VIRTUAL STRUCTURE OF A PHOTON
T
S
S'
S''
S'''
HYPERSPATIAL COMPONENTS INTO/ONTO PHOTON

Figure 9. The Eureka concept

vectors of the next higher hyperspace—the mindfield itself. In other words, by using two orders of modulation on a carrier, we can directly affect mindstuff itself. So if the correct overall pattern of these complex modulations are impressed on a carrier, the life-consciousness loop itself can be directly affected or shaped, all the way from the mindfield of a biological organism to its neutrino field to its virtual photon field to its biophoton field to its body tuner, and back. With the correct pattern we can stimulate the life channel for a particular organism, and with yet another pattern we can inhibit the life channel for the same organism. By this means a positive cure for cancer could easily be derived, and as we shall see, this already has been done. By the same means, a "death ray" for use against humans could easily be developed, and it is my contention that the Soviets have already long since developed such weapons.

It is my purpose to eventually use these new aspects to construct a direct mind-link for mind-to-mind communication and hyperlinkage, thus solving the human problem.

Figure 8 shows the basic scheme for coupling such a second-order modulated signal to a human being. We will return to modulations versus nested virtuals and hyperspatials below.

THE EUREKA CONCEPT

So we have now reached a Eureka point of the first magnitude (Figure 9). Hyperspatial engineering, virtual state engineering, curved space, photon kinking, mind-life-consciousness engineering, and nested modulation are all the same thing! We directly engineer the virtual state, hyperspace, curved spacetime, and the life-consciousness loop itself, including the mind, with relatively simple devices that accomplish, transmit, and receive nested orders of modulations. Indeed we have opened Pandora's box, for this is the greatest tool ever dreamed of by humankind. Consequently it can be used either for the good of mankind or for its detriment. And this tool appears already to have been weaponized by the Soviets,

In Figure 9, note that modulation, photon kinking, the appearance of sidebands with the unchanged carrier, impressing a virtual pattern upon a photon, and impressing hyperspatial components into or onto a photon are equivalent. Further, if we assume that the electromagnetic wave (pure, undistorted wave in a linear medium) is given by Maxwell's equations, then modulation mathematically consists of impressing a particular differential operator function and a nonlinearity upon Maxwell's equations. For each modulation pattern, a particular virtual structure is impressed on the photon, which is the same as impressing a particular hyperspatial pattern upon it. Each specific pattern impressed on the photon (that is each particular kink pattern of the photon) corresponds to a specific differential operator function. Through this means, a specific theory can be constructed empirically. That is, one can work out the effects of various mixtures of waves on specific biological systems, measure the specific channels used by the mind-life-consciousness loop, and so on. The Lisitsyn paper and Kazhinskiy's "biological radio" work in fact state that the Soviets have already done precisely this, and the Kaznacheyev paper shows that a large number of related biophoton experiments were in fact carried out in Soviet military laboratories. But let us look at some very important and exciting German work directly related to this area.

BIOPHOTONS AND VIRTUAL EM FIELD OF A BIO-ORGANISM

In an incredible piece of work, "Photon Storage in Biological Systems," in *Electromagnetic Bio-Information,* Urban and Schwarzenberg, Baltimore, Md., 1979, pp. 123-149, Dr. Fritz Albert Popp has presented the results of his laboratory work—and his theory—dealing with quite new developments in biological organism fields. Indeed, this may be one of the most important Western papers ever presented, if its true import is noticed. Briefly, in studying cellular electromagnetic bioactivity, the cell behaves as a system far from thermal equilibrium. It thus can be characterized as a Prigogine system. (Ilya Prigogine received the Nobel Prize in 1977 for his development of nonlinear thermodynamics of systems far from equilibrium, thereby changing the second law of thermodynamics.) We now know that in such a system negative entropy can result, and order can emerge from disorder. For a wildly disordered system, large-scale order can emerge and stabilize, and the more chaotic the disorder, the greater the stability of the ordered patterns that emerge. (Which seems to imply a confirmation of my fourth law of logic—total disorder must inevitably turn into total order.)

At any rate, living cells produce coherent light with a laserlike activity. This coherence serves as a control for the metabolism of the cells, regulates growth, and fixes the reproduction rate. Further, the cell exhibits superconductive behavior.

Popp found that a living organism in fact possesses a living aura, a virtual electromagnetic field that pervades the entire organism with a virtual photonic flux. In this field, virtual photons are stored. The field continually receives inputs (virtuals) from the environment and is continually outputting biophotons, particularly in the near ultraviolet. This field, in which all cells are bathed and with which they all continually intercommunicate, tends to stabilize and cohere the organism. All this has been established by laboratory experiments, and all of it is directly consistent with my own hyperfields and nested-virtual-fields model of the living system's life-consciousness loop connecting its mind hyperworld to its physical body tuner. My model is consistent with Popp's work and extends it by adding the nested structures of the virtual field, thus connecting Popp's virtual field to the neutrino field and the mindfield.

And with this additional information in mind, let us return to our consideration of nested modulations on an electromagnetic carrier. Indeed, let us look at several examples of devices already applying the "nested modulation is hyperspace vectoral" effect.

THE PRIORE MACHINE

The cure for cancer mentioned above is the Priore device in France, which has been curing cancer and other diseases for over a decade, much to the frustration of many members of the French Academy of Sciences.

Antoine Prior6 is an Italian-born inventor who escaped from the Germans in World War II and fled to France. After the war Priore" purchased surplus equipment and eventually developed a strange machine that filled an entire room. A nozzle above a table in the center of the room emitted a strong ray or field, which Priore* explained consisted of an electromagnetic wave in a magnetic field. Eventually the French government spent almost three million dollars for construction of a new, improved PriorS machine, because Priori had shown he could cure cancers in laboratory animals with the device.

In 1964 Professor Robert Courrier, renowned biologist and secre'taire-perpe'tuel of the French Academy of Sciences, presented some of Priore's results to the Academy. Forty-eight cancerous rats had been divided into two groups of twenty-four each. One group was set aside without treatment as a control, and the other group was divided into two subgroups of twelve each. One of these subgroups was treated by the Priore machine promptly after cancer grafts, and the other was treated after a lapse of several days. Both subgroups were completely cured—the first quickly and the second after more prolonged treatment. All the control rats died in twenty-two to thirty days after grafting.

The results created a furore and a controversy that have continued to the present day. Since the first presentation by Courrier, the machine has repeatedly cured leukemia and other forms of cancer in laboratory animals. The controversy seems to have arisen because the French Academicians could not comprehend the working of the machine, and because the inventor would not reveal how it worked but only how it was built.

The second, more powerful Prior6 machine produces a magnetic field of about 1,240 gauss. According to one scientist, the machine contains a gas-filled vacuum tube, which contains a plasma of mercury and neon gas. A pulsed 9.4 gigahertz electromagnetic wave is modulated onto a 17 megahertz wave in this plasma. However, the true nature of the machine is revealed by the fact that the 9.4 gigahertz wave is joined in the "lamp" by seventeen radiations of unspecified frequency and nature. Thus we see that Priori is actually using second-order multiple modulations and is dealing with the direct life-mind-consciousness loop of the mammalian cell. He is directly stimulating and engineering the hyperspace nested virtual bands with his device. His twenty-six years of work have mostly been involved with ferreting out the required precise frequencies and mix of second-order modulations, as well as the correct intermodulation carrier necessary to obtain the desired effects.

In Figure 10 we show a simplified diagram of the Priore device, from which it can immediately be seen that the device uses the same principles shown previously in Figure 9. PriorS is using nested orders of modulations to affect nested orders of virtual state, thus producing direct channels and effects in neutrinic field and in mindfield. Radiating a biological target such as a laboratory animal with such a complex field directly engineers its mind-life-consciousness loop. With the correct nested modulations on the correct carrier (in this case a strong magnetic field, which allows a stronger connection to the neutrinic field aspects), the mind-life-consciousness loops of the individual cells in the body of the irradiated animal arc "scrubbed clean" of any deviation from the norm by the virtual patterns delivered to the animal, and hence to its "virtual biophoton stabilization field" reported by

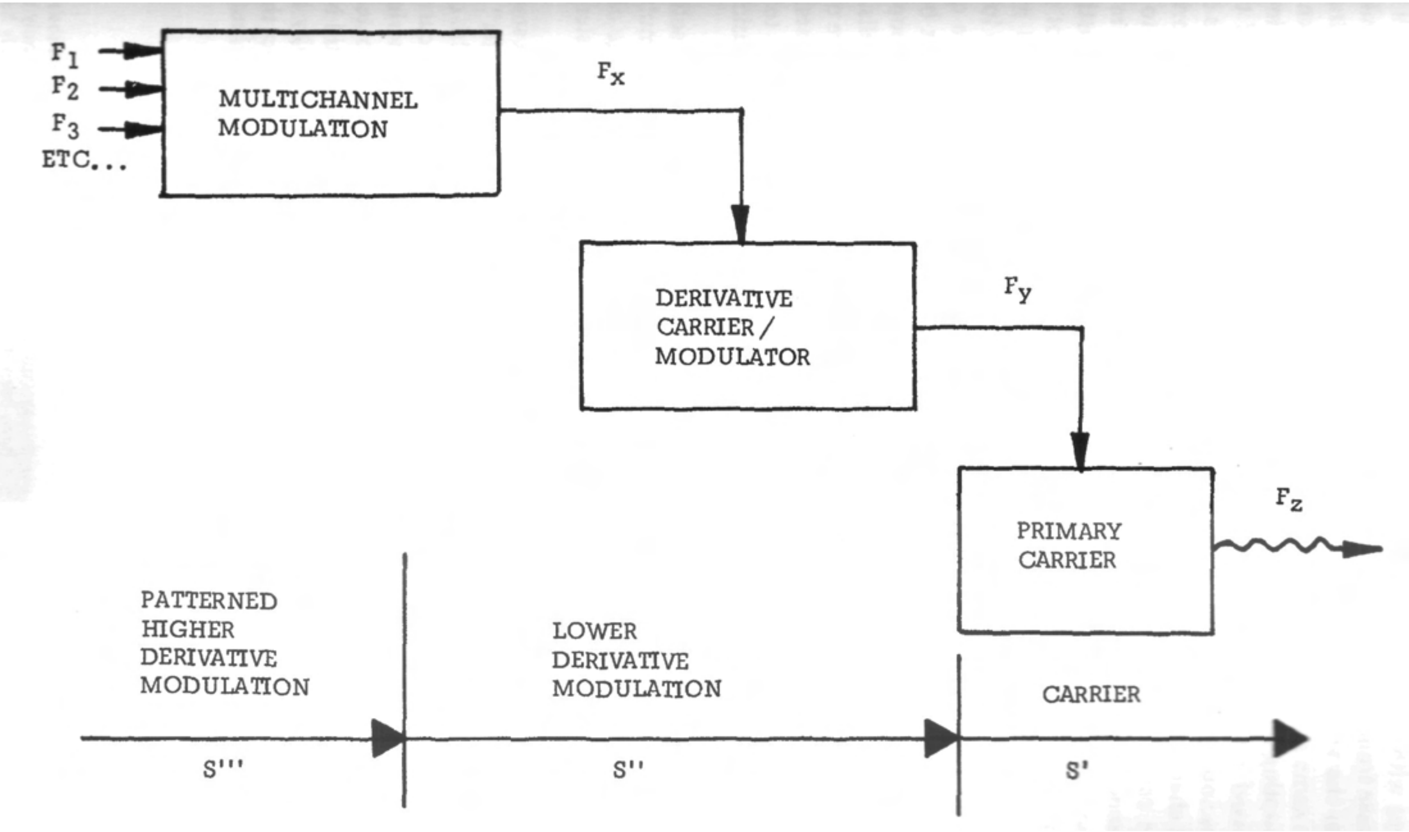
F_1
F_2
F_3
ETC...
MULTICHANNEL MODULATION
F_x
DERIVATIVE CARRIER / MODULATOR
F_y
PRIMARY CARRIER
F_z
PATTERNED HIGHER DERIVATIVE MODULATION
LOWER DERIVATIVE MODULATION
CARRIER
S'''
S''
S'

Figure 10. A simplified diagram of the Priori device

Popp. In other words, the Priore device produces a reverse sort of effect as compared with the Soviet work reported by Kaznacheyev on anomalous death and disease transmission between cell cultures by near ultraviolet biophotons emitted from the virtual Popp field of the dying cells of the stressed, attacked cultures. The same sort of virtual engineering that can transmit death of the organism (by disrupting its mind-life-consciousness loop tuning) can also transmit life into a stressed attacked grouping of cells by a sort of "forced resonance" of their consciousness loops into normal health patterns. (Virtual state interactions control charges, which control chemistry. Thus the entire chemistry of the absorbing cells can be controlled and engineered.) Also, the Lisitsyn paper refers to the presence of eleven independent channels in the brainwaves (namely, in the mind-life-consciousness connection to the waves of dendrite firings that constitute brainwaves*), and these were specifically referred to clipped signals, not sine waves. (The clipped signals would be those possessing complex nested modulations, and thus representing the tuning effects of the life-consciousness loop.) The Priore machine uses seventeen higher-order modulations, which are quite sufficient to cover all eleven life channels required, with an additional six being available for tuning so as to squelch or "jam" the life-consciousness loops of those cells with tuning abnormalities, that is, of diseased or cancerous cells.

As we would expect, the Priore machine has also shown its ability to cure other infectious diseases and materially benefit victims of cardiovascular diseases. Infectious diseases cured by Priori radiation also provide an immunity to the treated host organism that has been shown to be transferrable by transfusion of blood samples.

Again I strongly emphasize that the individual biological effects of solitary magnetic fields or single frequency waves cannot be extrapolated to the multiple frequency, nested modulation case; and almost all our present knowledge of the effects of electric, magnetic, and electromagnetic fields on biological systems is woefully immature and incomplete.

T. Henry Moray built a somewhat similar machine—though very much smaller—which provided a direct radiation from extensive "peach blossom violet" colored sparkgaps inside a quartz tube, and which greatly stimulated the health of an individual exposed to the radiation. In fact, the device would promptly cure **the** radiation burns Moray frequently suffered as a result of his continual experimentation with uranium compounds and other radioactive substances. He patented the device, and some of his original machines are still in existence and still being used to treat patients with a wide variety of ailments. I have personally been exposed to one of Moray's treatment machines and can vouch for the fact that it stimulates every cell in one's body, providing in about two to three minutes a great tonic and health stimulant.

***It is these dendrite firings which form the collector-detector mechanisms through which the mind-world changes are joined to the human nervous system and brain. Were they not specifically useful and needed, the dendrite firings would not have been developed, and nerves would simply be long wires.**

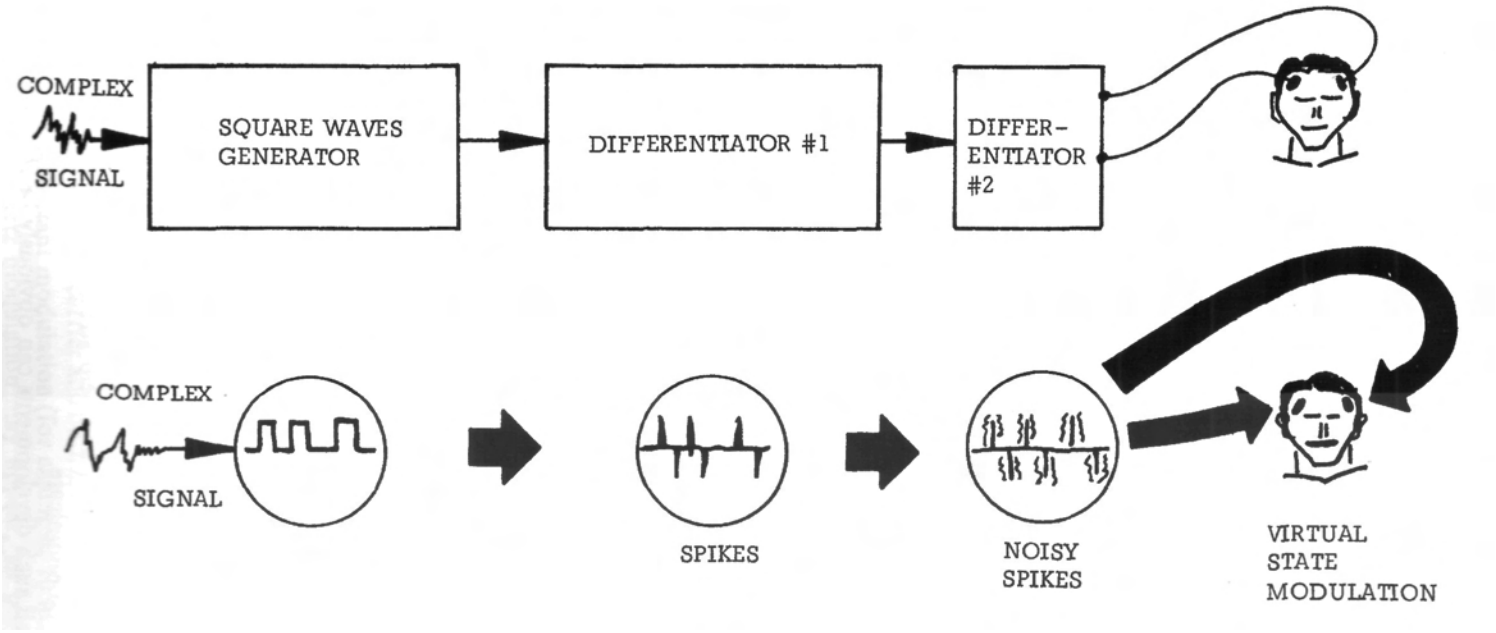
COMPLEX
SIGNAL
SQUARE WAVES GENERATOR
DIFFERENTIATOR #1
DIFFER-ENTIATOR #2
COMPLEX
SIGNAL
SPIKES
NOISY SPIKES
VIRTUAL STATE MODULATION

Figure 11. A simplified diagram of the improved Flanagan neurophone

THE NEUROPHONE

Another device that uses the new hyperspatial, virtual state, nested modulation technology (and has done so for seventeen or eighteen years) is Dr. Pat Flanagan's neurophone. With brilliant insight and intuition far beyond that of science at the time, Pat invented and patented the instrument by the time he was seventeen years old. The neurophone is a device that, contrary to all present theory and knowledge, will directly "pump the brain" and reproduce sound and information directly in the brain and mind system, without going through the auditory system at all. A simplified diagram of Pat's improved neurophone is shown in Figure 11.

Briefly, the device takes a complex signal, such as the sound of an orchestra playing a musical interlude, and electrically processes it as shown in the figure. First the signal is passed into a section that clips everything into a series of square waves, remarkably analogous to the sort of clipped waves Lisitsyn confirms are the information carriers of the human brain. Next the square waves are differentiated, yielding a series of sharp spikes (note that these spikes retain the pulse-time content of the clipped signal). These spikes are again differentiated, and since these are finite spikes with real nonzero rise times and decay times rather than theoretical constructs, a series of noisy spikes results from the second differentiator section. From here, the noisy spikes are introduced to special contact electrodes, one of which is normally placed on the forehead, while the other may be placed almost anywhere, including on the foot. Nowhere are any sound waves introduced to the head.

The square-wave clipper section reduces the complex signals, their overtones, and their complex modulations to square waves, retaining the temporal content of the wave mix but not the waves themselves. The first and second differentiators heighten or filter through the temporal content of the higher-order differentiations, that is, they serve as a band pass filteT unit to accent the time keying of the neutrinic and mindfield portions or aspects of the signal. When these time spikes are then introduced across the body as pulsed voltages, they are modulated directly on the dendrite firings of the brain and nervous system, providing direct and pulsed modulation of the neutrinic and mindfield component channels of the mind-brain-consciousness-life loop itself. Thus the neurophone directly inputs information into the brain and nervous system, bypassing all the normal sensory systems that lie between the mind-brain loop and the outside environment.

Indeed, a similar development has been reported by two University of Missouri researchers. Dr. Donald York, a neurophysiologist, and Dr. Thomas Jensen, a speech pathologist, have recently reported identifying and decoding twenty-seven words and syllables in specific brain wave patterns and correlating these electroencephalographic patterns with both the spoken word and the silently thought word in about forty subjects. At present, Dr. York and Dr. Jensen are programming a computer with a brain wave vocabulary, to monitor and read the EEG of a stroke victim's brain and help stroke victims who have lost their powers of speech to communicate.

While all such developments can obviously be misused, their potential for assistance to mankind is enormous. Indeed, my own premis that brain linkage it possible and can be used to advance all mankind to the sixth stage of species evolution is slowly being proven. It is within our reach now to develop a multichannel communications system that will directly link brains, minds, and consciousness into a single functional being; and I have already pointed out that the psychokinetic power of the emergent entity increases exponentially with the number of linked stages. The normal corpus callosum linkage of the two cerebral halves in each human head proves that brains, minds, and beings can be linked and integrated into one.

The Excalibur Briefing details the basic concepts of the theory involved in hyperspace-virtual state engineering, and thus in the engineering of mind-life-consciousness links to multicellular organisms here on earth. Pat Flanagan's neurophone proves that complex information can be implanted directly in the brain and mind, bypassing the normal isolating barriers. Drs. York and Jensen have shown that the brain wave does indeed contain recognizable, decodable information analogues. Lisitsyn has already reported measuring the number of independent channels involved and has documented the importance of the clipped brain waves as opposed to the sine-wave content. If we have the vision, we can still develop a brainlink and rapidly apply it to link sixty to one hundred persons, thereby opening a conscious, multidimensional "eye" for the collective human unconsciousness, gently rousing ZARG into wakefulness, and fulfilling human destiny without undergoing the psychotronic Armageddon planned by the Soviets. The time is short, but it can still be done if the funding, the vision, and the team can be established on a crash priority basis.

SOVIET "WOODPECKER" SIGNALS

But, as I have said, all this can be used for the detriment of mankind as well as for its benefit. Unfortunately, it appears that the Soviet Union has chosen to weaponize the effects on a global scale. A prime example is the "woodpecker" signals emanating from the USSR in the 5-30 megahertz region and interfering with communications around the earth.

These complex woodpecker signals appear to originate from two or three dozen powerful Soviet transmitters, each with a power estimated as high as 40 megawatts (Figure 12). The pioneering experimental measurements of these signals by Dr. Robert Beck and William Bise have shown just how deadly a potential may be possessed by the signals. These measurements have been performed in Eugene and Portland, Oregon; Los Angeles, California; Huntsville, Alabama; and several other locations. They have been particularly significant on the West Coast in and around Eugene, Oregon because of the presence of a direct current transmission line several hundred miles long, which has acted as a long wire antenna, picking up the signals and rebroadcasting them with appreciable gain in the vicinity.

Typically the signals may be found on, say, sixteen different carriers between 10 and 20 megahertz. Twelve of the carriers may appear normal, with normal sidebands, and the other four may have the carrier and both sidebands suppressed but still show the biologically significant modulation (for example, 10 hertz). On all sixteen channels a strong 10 hertz modulation may appear, all perfectly time

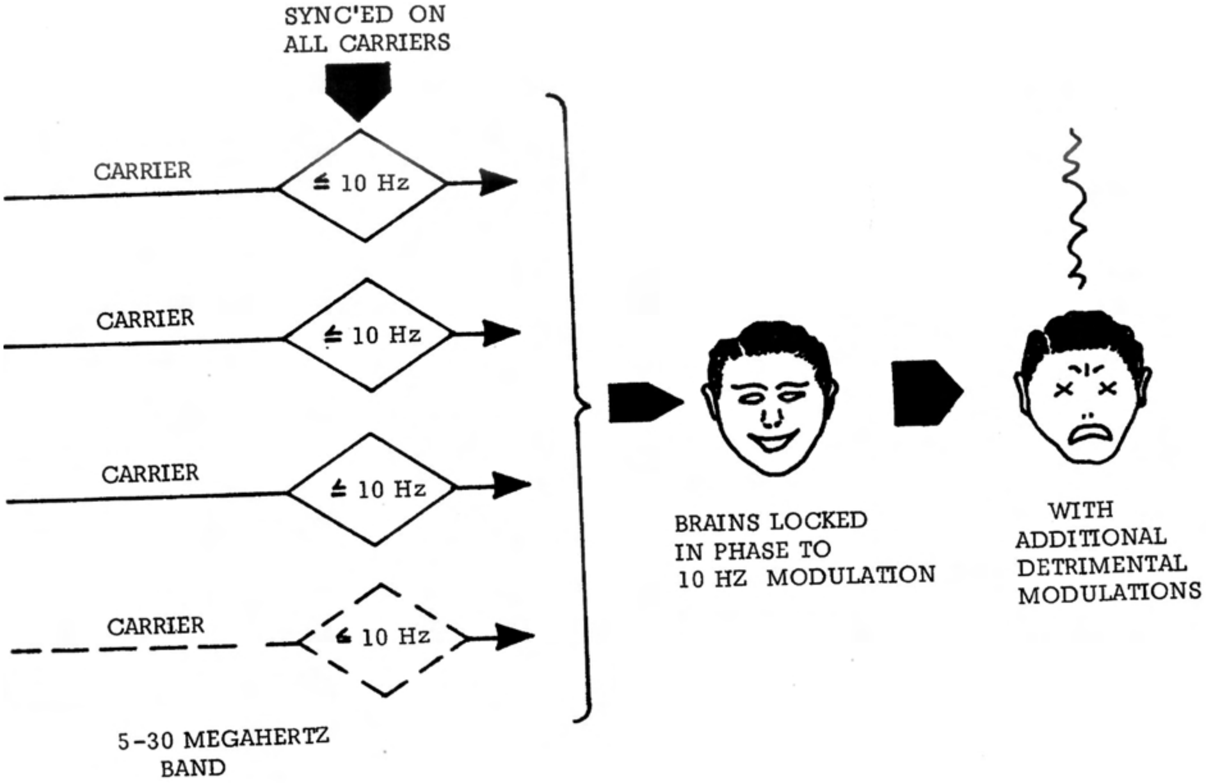
SYNC'ED ON
ALL CARRIERS
CARRIER
≤ 10 Hz
CARRIER
≤ 10 Hz
CARRIER
≤ 10 Hz
CARRIER
≤ 10 Hz
5-30 MEGAHERTZ
BAND
BRAINS LOCKED
IN PHASE TO
10 HZ MODULATION
WITH
ADDITIONAL
DETRIMENTAL
MODULATIONS

Figure 12. The Soviet "woodpecker" signals

synchronized and in phase channel-to-channel. The received signal from one of these carriers may be twenty-five to thirty times as strong as the earth's background magnetic field, which is oscillating at approximately 7.0 to 7.5 hertz. Other complex modulation frequencies, many of them changing, are present on the various channels.

Normally, the brainwaves of mammals in an area are gently entrained by the normal oscillations of the earth's magnetic field and are oscillating along at, say, 7.5 hertz. Under continuous radiation from the Soviet woodpecker signals, a percentage (say 30 percent) of the mammalian brains may be captured by the 10 hertz modulation on the overpowering Soviet signals. These captured brains are now in forced oscillation and are riding along in phase with the 10 hertz Soviet modulation. In other words, a certain percentage of the human brains in an area will phase lock to the 10 hertz modulation, and that will effectively lock into those brains the sixteen carrier frequencies with their concomitant frequency mixes and complex signal modulations. If the signals are made much more powerful (for example, by focussing them onto one particular area or increasing the power of the transmitters), then a much higher percentage of brainwave entrainment can be accomplished in the targeted area. By simply adding on the desired modulation patterns, the Soviets can now pump material directly into the mind-brain-consciousness-life loops of the entrained brains. The weapon implications are enormous: raw emotion such as sheer terror or panic can be transmitted; death or disease patterns of all kinds can be transmitted; informational content (thoughts and ideas) can be impressed directly into the captured brains and minds and processed as if originating inside each brain itself. Indeed, as long as the Soviets keep the effects rather gentle, an entire population can be subtly influenced without governmental notice. Certain specific ELF frequencies can rapidly disable or even kill, and for wartime use those frequencies can be directly implanted in the captured brains themselves by the woodpecker signals. It appears that 1984 came a little early, and one can perhaps now understand Brezhnev's strange 1975 proposal to the SALT negotiations that we should also consider outlawing the development of new weapons more frightful than the mind of man has ever conceived.

The signals also have direct application against strategic targets other than personnel. For example, by incorporating the correct nested modulation pattern, the positron kindling (electron squelching) pattern can be impressed. By beaming through the earth (with a circularly polarized signal or with other complex modulation forms), a beam with a strong neutrinic component, which will squelch electrons and which will interact with great preference for strong voltages or strong magnetic fields, can be directed at a target complex. In the targeted area aircraft and communications systems can be disabled; automobile, truck, and other ignition systems will fail; power transformers and large motors and generators will fail; power transmission systems with their associated high voltages will fail; and so on. Even independent, self-contained emergency power equipment, such as at military installations and strategic missile sites, will fail to operate. Further, to some of these signals, the earth and the sea are as transparent as glass, and so underwater nuclear submarines can be electrically disabled on station before their missiles can be fired, while strategic missile complexes can be electrically disabled before their deadly nuclear missiles can be hurled toward their distant targets. If some submarines and land complexes succeed in firing missiles before being disabled, these missiles can be electrically dudded and scrubbed from the sky over Soviet territory while still far from their intended targets.

So, as stated at the end of *Excalibur,* the players are on stage and in place. The final act begins. One thing is sure—mankind will never be the same again after seeing the play!

PART FOUR

SOVIET PHASE CONJUGATE DIRECTED ENERGY WEAPONS

WEAPONS THAT USE TIME-REVERSED ELECTROMAGNETIC WAVES

At the end of WWII, the Soviet Union obtained the cream of the crop of Germany's radar scientists and infrared scientists.

At that time, the German scientific team led the world in the theory and technology of radar absorbing materials (RAM) and radar cross sectioa For example, some leading Western radar experts believe that the German scientists had already advanced the theory of radar cross section beyond where Western scientists have arrived at today. Radar cross section science is the "heart" of modern radar technology, countermeasures, and counter-countermeasures.

The theory of RAM technology is precisely what is needed to develop and design phase conjugate mirrors for radar frequency bands. Phase conjugate mirrors are capable of producing a time-reversed (TR) wave in direct response to a received ordinary wave. The mirror may be powerfully "pumped" with energy to produce a very large amplification of the time-reversed wave.

Just after WWII, the Soviets also mounted a truly massive program to obtain and review all the scientific literature of the West This material was completely re-examined and digested, in a deliberate search for a major new technological "breakthrough" area. Nothing similar to this has ever been done in the West.

By 1950, the Soviets probably had discovered phase conjugation and the time-reversed wave, in their radar programs using the German radar scientists. Their relearch would certainly have revealed that the time-reversed wave is a solution to the wave equation, and hence the phenomenon is universal to all waves and frequency regions, under appropriate nonlinear circumstances.

We stress the fact that from the beginning the Soviets have always led the rest of the world in nonlinear science and mathematics.

Thus by the mid-50's, the Soviets would have mounted an intensive national development program on these radar TR wave weapons. If so, the Soviet Union has now had the equivalent of some seven or eight Manhattan Projects, back-to-back, in the development, stockpiling, and deployment of TR wave weapon systems. At least fifth generation Soviet TR weapons have been developed and deployed.

This is a matter of the greatest importance to the U.S.—and one which the U.S. intelligence and scientific communities have largely missed.

We particularly stress that the time-reversed wave phenomenon is a universal phenomenon of nature, not just a fluke, and not just a "nonlinear optics" phenomenon. It is a solution to the wave equation and, as such, it applies to every kind of wave. Electromagnetic, sound, magnetohydrodynamic, and other radiating systems all exhibit the phenomenon under appropriate nonlinear conditions.

Hundreds of Western papers dealing with time-reversed waves are now in the open scientific literature—most of them dealing with non-linear optics. However, the principles of the time-reversed wave are well-established, and known to apply to waves in general.

Compared to a normal wave, a time-reversed wave has startlingly different weapon capabilities.

Such a wave precisely retraces the path of the ordinary wave that stimulated it to be formed. So it possesses an "invisible wire" through space, back to the original position of whatever emitted its stimulus wave.

Further, the time-reversed wave continually converges upon its invisible "backtracking" path. It does not diverge and spread its energy, in contradistinction to normal waves.

Using several simple schemes (particularly pumped 4-wave mixing), extremely large amplification of the time-reversed (TR) wave can be cheaply and readily accomplished.

A startling weapons capability therefore emerges when amplified TR waves are generated in response to received signals from a distant target:

(1) If any signal at all can be received from a distant target, a return TR signal of extreme power can be delivered directly to that target Almost all of the transmitted TR signal energy will arrive at and in the distant target, even through a highly nonlinear medium or under scattering conditions. Hardly any of the energy will be lost enroute. If the target is fast-moving, a "lead correction" signal can be calculated and added to steer the return path.

(2) Since real-time holography can readily be accomplished using TR waves—and without first making holograms, geometrical forms (balls, shapes, hemispherical shells, etc) of energy can be created readily by interferometry (crossed beam techniques). Since the TR wave carriers do not disperse with distance, these interference energy forms can be assembled by crossed TR wave beams at great distances—even hundreds of thousands of miles. The energy appearing in such a distantly created energy form is limited only by the amount one cares to put in at the amplified transmitting end.

Thus the radar itself now becomes a powerful, all-around weapon. With a TR wave adjunct, once the radar receives a return signal from a target, an extremely powerful TR wave pulse can be generated, and all the energy in that pulse can be unerringly returned to the distant target from which the return was received.

Even a passive (receive only) radar may be used together with a TR wave adjunct to accomplish the same directed energy destruction of the distant target emitting the signals received by the radar.

Radio receiver-transmitters may also employ simple TR wave adjuncts to provide a large number of communications jammers and "directed energy weapons," capable of jamming and/or destroying enemy radio transmitters at appreciable distances.

The Soviets have intensively developed such TR wave weapons since the 1950's. In January 1960, Khrushchev was probably referring to these weapons when he reported to the Soviet Presidium the forthcoming advent of fantastic weapons, then in advanced Soviet development.

Prior to Khrushchev's announcement, the Soviets had begun radiating the U.S. Embassy in Moscow with scalar EM waves and phase conjugate energy, modulated upon weak microwave carrier beams. Later these beams were used to induce illnesses and blood changes in Embassy personnel. (We will shortly discuss how the Soviets could induce disease at a distance, electromagnetically). The microwave radiation continued intermittently, in spite of the objections of several U.S. Presidents. U.S. studies failed to comprehend what was being done.

The Embassy radiation was and is a high-level intelligence probe: Stimulate and affect the U.S. Ambassador, a high-level target. Attention by the CIA, NSA, State Department, DIA, the U.S. President, etc. is guaranteed. Puzzled, they will turn to the U.S. scientific community for an explanation. The scientific community will try to explain the microwave radiation, using the best of its EM knowledge. By the U.S. counteractions taken or not taken at the Embassy site, the Soviets know with 100% certainty whether or not the U.S. has discovered TR wave weapons, electromagnetic disease control, etc.—and whether we have developed defenses against such weapons.

Since the beginning, of course, our actions—or lack of them—at the U.S. Embassy have consistently revealed to the Soviets that we have no knowledge of, and hence no defenses against, TR wave technology and the scalar electromagnetics (electrogravitation) that accompanies the phenomenon.

Soviet TR wave radar weapons also provide a convenient mechanism for accomplishing nuclear weapon kill. A target struck by a powerful TR wave pulse generates a time-reversed back-pulse due to nonlinearities in the target medium. In the nonlinear medium of the exploding or fuzing target, the two pulses (which are 180 degrees out of phase with each other) modulate each other, locking together into a scalar EM wave or pulse. This scalar EM pulse has zero electromagnetic force field resultants, but is a violent fluctuation of the energy density of local spacetime. Rigorously it is a local gravitational pulse wave, by definition. Further, it is far more powerful than "normal" gravitational waves. This scalar EM pulse does not react with the orbital electron shells of an atom, but passes directly into the nucleus and is absorbed there. This raises the nucleus to an excited state, increasing its potential (trapped energy). If the nucleus cannot withstand this increase in potential energy, it directly fissions. If the nucleus withstands the increased potential, it then immediately decays from the excited state by simply emitting a time-reversed EM wave.

Any radioactive nucleus is already "teeter-tottering" toward nuclear decay, so to speak, and it will instantly decay radioactively when struck by a significant scalar potential pulse. Thus a TR wave radar weapon can deliver a powerful TR pulse against a distant vehicle carrying a nuclear warhead, and the resultant EG pulse penetrating the nuclear material of the warhead will electrogravitationally explode it in a full-order nuclear detonation.

Therefore, when using such weapons, extreme care must be taken to avoid inadvertently pulsing one's own nearby nuclear weapons or nuclear material. In attacking distant ground targets, one must be careful not to detonate nearby nuclear weapons, nuclear powerplants, or stored nuclear material—even at some distance from the target Otherwise, large-scale nuclear fallout will be experienced over a worldwide area, and unacceptable "boomerang" self-damage to the attacker himself may be incurred. The more powerful the TR pulse employed, the greater the safety separation between the struck target and stored nuclear materiel must be to prevent unacceptable "boomerang" large-scale fallout.

Also, large TR wave interferometer weapons, such as the Woodpecker systems, usually first produce very powerful, scalar EM "standing wave" beams by continuously transmitting (in the beam) both a normal EM wave and its phase conjugute.

modulated (locked) together to produce a "zero EM vector resultant" electrogravitational (EG) wave. Interference of two of these scalar EM beams in the targeted area, plus internal scanning within the beam by other signals, allows EM effects—of either normal EM energy or time-reversed EM energy—to be produced and controlled at very precise locations within the broad interference area.

Each such huge "scalar EM standing wave" beam represents a gigantic electrogravitational standing wave, and hence a giant oscillating potential in spacetime. This standing wave represents a sort of gigantic "capacitor", or accumulator of infolded energy. Enormous energy may be collected in this potential, charged-up over a period of time. "Shortcut" of this giant capacitor—by transmitter failure—can result in a large flash-over discharge of the EG energy into the local earth, producing a massive EG ground wave that can be enormously destructive.

Thus one must carefully protect the TR wave weapon system from inadvertent discharge of large EG pulses into the earth at the local site, due to transmitter failure, shortouts, etc. Otherwise one's own nuclear weapons or facilities may be exploded—even at some distance—by the EG pulse ground wave produced. One's communications and electronics installations, for example, can be knocked out at an even greater distance.

For example, in the winter of 1957-58, a large Soviet TR wave prototype weapon at Kyshtym, near the Urals, apparently suffered a catastrophic transmitter failure, discharging a powerful scalar EM (electrogravitational, or EG) pulse into the earth. This huge EG pulse struck the atomic wastes stored nearby, causing the radioactive nuclei to imediately decay. In other words, the nearby atomic wastes exploded—just as eyewitnesses reported. Deadly radioactive contamination spread across a major region, and exists to this day.

After this accident, the Soviets would have developed much more elaborate safety circuits and devices and implemented them into their TR wave weapons.

Even so, in April 1986, the explosive eruption of a reactor at Chernobyl was almost certainly caused by an accidental catastropic failure of a large TR woodpecker transmitter about 30 kilometers away. The sudden failure of the East-to-West Woodpecker transmitter was positively detected by engineer Bill Bise. All safety circuits at the site would have been instantly activated, desperately and slowly draining off the huge "standing wave" potential built up by the weapon.

The Soviets would have immediately shut down the reactors at Chernobyl as a precaution.

Some hours later, the transmitter safety devices finally failed, and a much smaller, though still significant, remnant EG pulse was discharged into the earth, spreading out in a ground wave. The first radioactive material—the fuel rods in one of the hapless, shutdown Chernobyl reactors—encountered by the EG ground wave pulse received it and immediately erupted radioactively.

Had the full EG pulse escaped the stricken transmitter, all four reactors at Chernobyl would have violently exploded, and deadly nuclear fallout would have rained down over a wide area of the earth. Had the reactors not been shut down and the control rods inserted "full-in" to dampen the radioactivity, a far greater radioactive eruption—probably a nuclear explosion—of the struck reactor would have occurred.

In May 1960, a prototype TR wave radar weapon was used to down Gary Powers' high-flying U-2 aircraft over the Soviet Union. Powers reported a flash, as if from an explosion near him, but the flash persisted. This is almost certainly the signature of a TR wave weapon.

In April 1963, the first large strategic Soviet TR wave weapons became operational. These weapons were used to electromagnetically deaden the electrical controls of the *U.SS. Thresher* atomic submarine on April 10,1963, and to produce a huge underwater electromagnetic blast 100 miles north of Puerto Rico on April 11,1963.

Since the 1960's, the Soviets have possessed a 100% effective defense against U.S. and European strategic bombers and missiles. The early systems deployed in the Soviet Union were midcourse and terminal phase systems. The Soviets would also have relied upon additional TR wave weapons systems, deployed as adjuncts to radar systems aboard research ships and crawlers, to be able to attack U.S. missiles and bombers in their launch and early midcourse phases.

These Soviet defensive TR wave systems also could easily accomplish nuclear weapon kill, and by their nature may have already been immune to the Stealth technology.

Older Soviet radars, using TR wave adjuncts, become totally new weapons having startling and heretofore unsuspected capabilities against missiles and aircraft. For that reason, the Soviets happily continue to operate and maintain thousands of older radar systems that, by Western radar standards, are seemingly obsolete.

Hundreds of incidents of Soviet tests of these TR wave weapon systems have been observed around the world, and reported in the open literature as puzzling or anomalous phenomena.

In the 1960's, in Siberia, Vlail Kaznacheyev and his researchers accomplished enormous experimentation proving that cellular death and disease—of any type whatsoever—could be transmitted and induced electromagnetically. This team went on secretly to phase conjugate (time reverse) the "death photon" signals, to show that any cellular death and disease can be reversed or "cured" electromagnetically as well.

The latter work would have been done under the auspices and tight control of the Soviet KGB, since all phase conjugation (time reversal) weapons effects are under the KGB, both for development and deployment Obviously the "electromagnetic antidote" work of Kaznacheyev was not published, because it provided a powerful counter-BW warfare capability of great strategic importance to the Soviet Union.

Thus in the late 1960's, the Soviet Union became the only nation on earth to achieve the capability of rapidly (within one or two hours) developing an electromagnetic antidote for lethal new biological warfare (BW) viruses—of any kind whatsoever. The Soviets also developed the capability for mass treatment in this fashion of both the civil populace and the military forces personnel.

Independent verification of the Kaznacheyev "death photons" effect has been accomplished in West Germany (University of Marburg), in Australia (University of Sydney), and in the United States (by G. Wilbanks).

Independent proof of electromagnetic healing of viral diseases was also accomplished in the West, and largely ignored.

In the late 1960's and early 1970's, a French inventor—Antoine Priore—built and tested several large electromagnetic devices that cured thousands of cases of terminal cancers and leukemias in laboratory animals. He worked with members of the prestigious French Academy of Sciences. His work was funded by the French government and presented to the French Academy by Robert Courrier, head of the Biology Section. His work was suppressed in the 1970's, and lost when he died.

Priore used a plasma to accomplish multi-wave mixing and phase conjugation (time reversal). His time-reversed curative signals used a powerful, rippling magnetic field as a carrier; the magnetic field guaranteed penetration of every cell in the body—necessary to reach a blood disease. The rippling of the magnetic field guaranteed penetration of the cell's atomic nuclei by nuclear magnetic resonance. Penetration of the actual nuclei of the atoms in the diseased cell is necessary, since the master cellular control system (discovered by Dr. Fritz Albert Popp of West Germany) operates through the cell's structured scalar EM biopotential, and this potential is centered in the nuclei of the atoms comprising the cell.

Note that electrical dedifferentiation of cells at miniscule current—between picoamperes and nanoamperes—has been proven by Robert Becker and other researchers. Red cells of the frog, for example, have been electrically reversed through all their development stages. The cells first become round, then develop a scalloped outline in their membranes. Then they become ameboid and move by means of pseudopods. Their nuclei swell and their DNA becomes reactivated. They then rid themselves of all their hemoglobin and develop a full set of mitochondria and ribosomes. In Becker's historic experiments, the cells then turned into cartilage-forming cells, and finally into bone-forming cells. In short, Becker demonstrated dedifferentiation followed by redifferentiation into an unrelated cell type—all electrically initiated with laughably weak currents.

Thus profound cellular changes are most certainly inducable and controllable electromagnetically, by miniscule amounts of potentials and currents. Priore's machine applied time-reversed EM waves to accomplish reversal from cancerous cells back to normal cells.

Priore's work essentially proved that Kaznacheyev's group would have been able to accomplish the same thing by phase conjugation.

Kaznacheyev's "death photons" work was probably accomplished much earlier than reported in the open literature. Thus it probably provided a basis to use in the microwave radiation of the U.S. Embassy in Moscow, to engender blood changes and disease in targeted personnel. Two U.S. ambassadors died from a leukemia-like disease, and a third sickened and bled from the ears and nose. Only miniscule levels of EM signals would have been necessary to induce such biological conditions.

The present Woodpecker radars probably have the capability of producing similar conditions in distant target populations, whenever the Soviets wish to add the appropriate signals. They also probably are capable of killing large numbers of the targeted population, whenever appropriate Kaznacheyev "death signals" are placed upon them.

In the late 1960's, another cautious Soviet "intelligence probe" was used: A paper appeared in the open Soviet literature, dealing with time-reversed waves in the field of nonlinear optics. Only a few American optical scientists seemed to be interested.

In 1972, to focus more U.S. attention on their subtle "intelligence probe" of U.S. time-reversed wave knowledge, the Soviets sent two scientists to Los Alamos to brief American scientists on optical phase conjugation. Thereafter many U.S. optical scientists began to work in the "new field." Immediately it was revealed to the Soviet Union that the U.S. had ineffective knowledge of phase conjugation, and that the U.S. had no TR wave technology or weapons.

To this day, U.S. theoretical work in phase conjugation is still characterized by a fairly simple overall math model, obtained by simply "writing down" the necessary conjugate term. There appears to be absolutely no U.S. awareness of the actual physical electrogravitational mechanism that causes the phase conjugation/time reversal, and no theory of it as such. Its direct application to radars appears to have largely been ignored in the U.S.

In 1975, the Soviets attempted to obtain a treaty with the rest of the world, outlawing the development of frightful new weapons of mass destruction—"more frightful than the mind of man had ever imagined", according to Brezhnev himself. Gromyko introduced the draft treaty into the 1975 session of the United Nations General Assembly. No one knew what the Soviets were referring to.

Failing to obtain the treaty, the Soviets in 1975-76 embarked on the greatest military buildup in history. This buildup was nothing short of a full-fledged preparation to be ready to take over the world, beginning in 1985 (the "ready" time per Brezhnev's 1972 statement at a secret Prague meeting of European Communist Party leaders). The Soviets met this scheduled goal.

As we stated, The Soviets can be immune to any kind of viral biological warfare, anytime they choose to be. The use of lethal new viruses thus has become a highly favored, low-risk method of warfare for the Soviet Union. The West has no such capability against viral warfare, and in fact is totally vulnerable to attack in this manner. Both Western troops and Western civil populations are totally defenseless against the threat of viral biological attack.

A BW strike from a small, hostile nation or from a few radical terrorists—or even an inadvertent contaminant spillage from one of the great many ill-controlled private genetic experiments—can result in the introduction of a lethal, highly infectious new virus that will decimate the West. Indeed, a deliberate BW viral attack upon us may already have been unleashed.

In 1971, the U.S. passed the National Cancer Act and declared open warfare on cancer. Iron Curtain country scientists were invited to participate, and they did so with alacrity. At the time—ironically—penetration of our recombinant DNA genetics research was one of the highest priorities of the Soviet KGB. Obviously, many KGB scientists/agents were infiltrated into our labs under this program (some of our cancer research labs at Ft. Dietrich, for example, still employ more Communist than American scientists.).

In these cancer labs, animal viruses were repeatedly injected and reinjected into human cell cultures, until new viral strains emerged that preferred human cells. These viruses could then be cultured and the effects (or lack of effects) of various chemicals and drugs against them could be ascertained.

These new viruses are exactly what one would wish to develop if one were seeking a biological warfare virus to which humans were not resistant, and one which preferred humans as its host

Sometime during the early 1970's, KGB agents who had infiltrated U.S. cancer research laboratories may have deliberately contaminated the World Health Organization's smallpox vaccine with one or more of the lethal new viruses available in the

Western cancer laboratories. Massive WHO smallpox vaccination in Africa and other Third World countries then resulted in the appearance and spread of the AIDS disease.

If this thesis is true, then AIDS constitutes the first Soviet biological warfare strike against the West, in what can only be categorized as the opening round in a unique new WorldWar III.

With the Kaznacheyev techniques, the Soviet KGB already has the effective electromagnetic antidote against AIDS and other lethal viruses, anytime it wishes to employ it. Some Soviet citizens, however, will be allowed to die, to serve as a deception plan and conceal the true Russian capability. Meanwhile, the Third World nations and the West will be devastated by the killer disease.

Other lethal viruses are almost certain to follow, whether from the Soviets, fanatical terrorist groups, or inadvertent private contamination. Undoubtedly, some of these frightful new viruses are not going to be slow (lentivirus) agents. Some will likely be highly infectious and quick-acting, in the manner of influenza. Some will be transmitted successfully by insect vectors such as the voracious new Asian Tiger mosquito now spreading across the Southeast U.S.

Beginning in July 1976, powerful new Soviet "over-the-horizon radars" (the infamous Woodpeckers) became operational. Extensive weather engineering over the U.S. has been accomplished by these weapon systems. (This has been separately documented by the present author; C. B. Baker, editor of *Youth Action News;* Andrew Michrowski of Canada's Planetary Association for Clean Energy (PACE); and other researchers such as Andrija Puharich.) In addition, the Woodpeckers possess the full range of TR wave weapon capabilities, including real-time holography capabilities and anti-Stealth capabilities. Terrible anti-biological capabilities are also possessed by these weapon systems, and they are also highly effective as global launch phase and midcourse phase defensive weapons for use against U.S. missiles and bombers.

In 1985, the Soviets completed "wet run" weapon tests against U.S. missiles and aircraft and at least three of NASA's shuttle launches, employing the Woodpeckers in the launch phase ABM/anti-bomber mode.

Beginning in December 1985, actual destruction of selected U.S. missiles and aircraft was accomplished. Targets destroyed by the Soviets included the Arrow DC-8 at Gander AFB, Newfoundland on Dec. 12,1985; the shuttle Challenger launched on Jan. 28, 1986; the Titan 34-D rocket launched on Apr. 18, 1986; a U.S. Delta rocket launched on May 3,1986; and the French Ariane rocket launched in late May, 1986.

In 1986 and 1987, the mysterious loss of two highly classified U.S. aircraft—believed to be Stealth fighters—may also have been due to Soviet testing of TR weapon capabilities and anti-Stealth capabilities of the Soviet Woodpecker weapon systems.

However, a severe limitation on the unrestricted use of TR weapons exists. If powerfully and unrestrainedly used, they will cause nuclear warheads and nuclear material to explode as surface nuclear bursts—in the dirtiest possible fashion. The resulting massive radioactive fallout would probably wipe out all higher forms of life on earth. Note that Khrushchev stated in 1960 that these fantastic weapons could "wipe out all life on earth if unrestrainedly used."

As long as we possess an appreciable deployment and stockpile of nuclear weapons, we are essentially "dead man fuzed" against substantial Soviet usage of their powerful TR wave weapons.

Thus Gorbachev needs to "thin out" the deployed and stored Western nuclear warheads, to allow a more unrestrained field of play for his decisive TR wave weapons. That is why at Reykjavik he suddenly expressed full interest in President Reagan's zero option, even expressing agreement with the principle of on-site inspection to insure dismantling of missiles and removal of warheads.

That is also why the Russians are so actively interested in negotiating the withdrawal of nuclear weapons from Europe. They can quickly and easily take Europe, the moment the nuclear weapons are thinned out and our "dead man fuzing" is removed!

To appreciate the ease with which a Soviet offensive, augmented with unrestrained TR wave weapons, can take Europe, consider this: A single airborne Soviet standoff radar jammer, complete with adjunct TR, within minutes can knock out all U.S., British, and FRG radars that illuminate it. Each pulse it fires, destroys an Allied radar. A single Soviet standoff communications jammer, complete with adjunct TR, within minutes can knock out all U.S., British, and FRG communications centers that are transmitting within its detection range. Each pulse that it fires destroys an Allied communications center. A single standoff Soviet jammer-based TR wave weapon can destroy an Allied carrier task force. Pulse after pulse from that weapon system will destroy the task force's radars and communications, explode HE propellants and nuclear weapons, explode fuel, ignite flammable materials, explode rockets and ammunition, and kill personnel wholesale. Tactical adjunct TR weapons can rapidly destroy aircraft, missiles, radars, radios, personnel, tanks, weapon carriers, personnel carriers, ammunition dumps, gasoline dumps, etc. The instant vulnerability of any massive emitter such as AWACS is readily seen.

We are referring to a new kind of blitzkrieg war: War conducted by powerful beams and destruction, engendered at the speed of light War in which electromagnetic radiators are the primary decisive weapons. War in which new kinds of directed energy weapons play the major role. War which renders all our present defensive armada of weapons archaic and useless. War in which the materiel and personnel loss rates are so incredibly high that they boggle the mind of the more conventional military tactician and planner. War in which even a small nation with TR wave weapons can strike and devastate a major power, and possess a first-strike capability unparalleled in human history. War which could be triggered by radical splinter groups and terrorists groups that acquire the technology. War in which the slightest miscalculation or the misuse of the weaponry can completely destroy the earth. War which places extreme emphasis on finding "other ways"—such as surreptitious biological strikes and insidious electropsychological attack and control—to decimate a foe. War that Western governments, populaces, and Armed Forces are completely unprepared—militarily and psychologically—to face.

NATO can be defeated in two hours or less, and the entire war—mop up and all—would last perhaps three days.

Of course, powerful demonstrations of Soviet strategic superweapons could be provided, to show that attempted nuclear retaliation would simply be suicide. The Woodpecker weapon systems alone could wipe out all major communications systems and power systems in the U.S., induce waves of EM-induced death and disease in preannounced locations, induce waves of abject fear and panic in the populace, or psychologically disrupt and disable—say—the entire Washington D.C. area, to include the people, the government, and the Pentagon.

If such a scenario were accompanied by a proclamation of a Soviet equivalent to the "Monroe Doctrine"—except applying it in Europe, an offer of a ceasefire, and an offer to allow us a "Dunkirk" to evacuate the beaten remnants of our European forces, we would almost certainly be forced to accept the Soviet terms for a ceasefire that left the U.S. intact.

The U.S. Congress—reacting with alacrity to accept this sudden and apparently miraculous Soviet leniency, and avoid the destruction of the world—would probably play the leading role in dictating our acceptance of liberal Soviet terms.

An agreement with the Soviets to mutually withdraw many of the nuclear missiles from Europe appears to be imminent. When the missiles are thinned out, the preceeding scenario or some variant of it is almost certain to occur.

There remains, however, one fly in the Soviet ointment: The U.S. Strategic Defense Initiative.

Presently our scientists apply phase conjugation primarily to "keep laser energy beams intact for greater distances." They appear to be completely unaware of the electrogravitational aspects of phase conjugation.

A colleague of the author has already rigorously proven this electrogravitational aspect on the laboratory bench. The phase conjugate (time reversed) wave is composed of, and carries, negative energy and negative time. If an object is forced to produce a great deal of this negative energy/negative time, it produces antigravity. This is because, in negative time, gravity is a repulsion, not an attraction.

In addition, if the electromagnetic energy of a laser is changed into gravitational energy, a theoretical gain of up to 10^{36} (protons) to 10^{42} (electrons) is achievable. In practical devices, an energy gain of 10^{20} may readily be achieved. Applied to a modified powerful laser in space, this electrogravitational gain factor is of overwhelming significance.

In SDI, U.S. scientists presently envision a high energy laser in orbit, pumped perhaps by means of a nuclear explosion. The pulsed beam from this laser would have sufficient power to destroy a rising Soviet strategic ICBM in the launch phase, say at 10,000 miles away.

If that same laser were converted to a scalar electromagnetics laser (one that turns much of its electromagnetic energy into gravitational energy), it could blast and destroy possibly a quarter of Russia with a single shot.

Thus if we developed and tested the SDI spatial/launch hardware for the space-based high energy laser, so that these systems were available for immediate launching, we could convert their lasers to scalar EM and still possess a very real strategic retaliation capability, even if the Soviets should strike us almost unrestrainedly with powerful TR wave weapons. If we launched several of these systems, we could still probably get off a few shots at Russia and destroy her, before all our space lasers could be destroyed by Soviet counterweapons.

In that case, mutual assured destruction would still exist, and we would still possess a full "last ditch" strategic retaliation capability.

Gennadi Gerasimov, Soviet Foreign Minister spokesman, confirmed as much on the Peter Jennings' ABC News Show, Los Angeles, California on 13 October, 1986. The real concern (with SDI), he indicated, is not with its present design. Instead, the Soviet concern is that the U.S. would make a great technical breakthrough and deploy new devices in space as a modification to SDI, presenting an unacceptable threat to the Soviet Union.

Gerasimov was almost certainly referring to the potential for U. S. adaptation of the SDI space-based lasers to electrogravitational lasers capable of destroying the Soviet Union.

Obviously, if the high energy laser weapons stay grounded in the laboratories, and the system hardware is not developed and ready, that strategic retaliation capability is lost to us. For that reason, Gorbachev is adamantly insisting that the SDI defense must not be developed and tested in space.

If Gorbachev can prevent the development and testing of the SDI space-based high energy laser and get the nuclear weapons thinned out, he immediately holds the decisive winning hand.

This is where we are now, vis a vis the Soviets. This is the background necessary to understand the full context of the present jousting between the Soviet Union and the U.S.

In the *Washington Times* of Aug. 26, 1986, Soviet Chief of Staff Sergei Akhromeyev warned the United States that the Soviet Union could have an unpleasant surprise response if the U.S. deploys its Strategic Defense Initiative (SDI) in space. Akhromeyev stated:

"If the United States deploys a shield in space, the Soviet Union will have several options, none of what Washington would wish... The Soviet Union will very quickly find a response of which the United States has no inkling as yet."

It is absolutely essential that the President of the United States, the Secretary of Defense, the Joint Chiefs of Staff, and the U.S. Congress be aware of these Soviet weapons and their capabilities.

It is also essential that the U.S. public become aware of the real nature of the hidden Soviet threat that faces us, so that public support will be forthcoming for the hard decisions our government leaders have to make to counter it.

A PARTIAL GLOSSARY FOR SCALAR ELECTROMAGNETICS AND SUBTLE PHENOMENA

by Tom Bearden

1. Action: the fundamental constituent of physical reality. Basically composed of two quantities—energy times time, momentum times length, etc. Action is nonobservable; only changes in action are observed, and then only partially (in the linear case). The photon is the basic quantum of action, consisting of a piece of energy welded to a piece of time, with no seam in the middle, so to speak. All physical things are composed of rates of changes in action. The physicality (volume characteristics) of space arc created bit by bit from action quantum changes in and on each fundamental particle of mass, at a very high rate of speed. These action changes are totally internal to the mass itself. Therefore, literally the physical universe exists inside each and every part of itself, and is a giant hologram.

2. A-field: vector magnetic potential. The A-field is a real field—not a mathematical convenience as was originally thought—as detailed in the second volume of Feynman's three volumes of physics. Indeed, it may be separated from the magnetic field, as the experiment reported by Feynman details. When separated, it is often referred to as the "curl-free vector potential."

3. Aharonov, Yakir: renowned Israeli quantum physicist and former student of David Bohm. Together with Bohm, Aharonov authored a fundamental paper in 1959 which pointed out the primary importance of the potential rather than the force field. Co-discoverer of the Aharonov-Bohm effect: Even when the electromagnetic force fields reduce to zero in a charged particle system, the potentials may still exist and interfere to cause real physical effects.

4. Aharonov-Bohm Effect: Effect whereby interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields. When E and B fields are zero in a region, their potentials still exist and interference of the potentials produces real electromagnetic effects (energy) in the interference zone, even at a great distance, without transmission of "energy" per se through the intervening space. The Aharonov-Bohm effect highlights some of the stark differences between the classical view of electromagnetics and the quantum mechanics view. In classical EM, the force fields are the real causative agents while the potentials are only mathematical creations for convenience in calculation. In quantum mechanics, it is the potentials that are real and fundamental; the force fields are effects derived from the potentials by differential operators. In classical EM, when the E and B fields are zero in an area, there is no more electromagnetics going on there; in quantum mechanics, the potentials still exist and may still cause real effects. By the use of artificial potentials, the Aharonov-Bohm effect can be made macroscopic and used to cause action at a distance—even significant action at a very great distance. Thus scalar EM interferometers can produce energy effects at a distance. The negative Aharonov-Bohm effect also exists. If the potentials of the transmitters in a scalar interferometer are biased highly negatively, then energy is extracted from the distant interference zone, reappearing at the transmitters. Tesla proposed a "Big Eye" that could see at a distance; actually he planned to use a negatively biased scalar interferometer, scan the beams, then reassemble the scanned energy appearing back at the transmitter/receivers. In this way he had a crude picture of the distant conditions in the interference zone. Such a device can see through the earth or ocean to a preselected intersection depth, since scalar EM waves can be made to pass readily through the earth or sea. The Aharonov-Bohm effect is in fact experimental proof of the new definition of force used in scalar electromagnetics. That is, force consists of mass in motion changing its motion (momentum). The nonzero 0 created by interference of two electrostatic potentials at a distance, for example, will create an electric field (E-field) on and of an electrically charged particle there. The standard detection of an E-field by conduction electron detectors is an example. It is the electron gas particles (charged electrons) that are detected, not what is in the vacuum. Thus classical EM errs when it assigns the equation E = - 0 to vacuum, for no such thing is ever measured. Actually, the correct equation is E = - 0 (e), where means "coupled to" and (e) means "electrically charged particle."

5. Aikido: a Japanese martial art whose modern form is a synthesis of many schools of jujitsu and the zen art of meditation in motion. The advanced aikido practitioner seeks to attain a state of "nomindedness," in which there is no separation of subject and object, of attacker and defender, or even of past and future. This state of at-one-ness with the universe, when attained, results in the complete mastery of the art. It is rarely experienced, even occasionally, and even rarer is it permanently attained. One who attains this state permanently is considered a great master, and his demeanor and personality are profoundly altered. The deceased founder of the present modern form, Morihei Uyeshiba, attained such mastery of the art He was a serene, deeply religious man and a profound inspiration to several generations of martial artists. He was also unbeatable even in his eighties.

6. Anenergy: stress energy of vacuum. "Fragments" of energy, more subtle than electromagnetic energy, turned against themselves and locked into a "vector zero summation." Scalar, massless charge flux. Virtual particle flux. "Pieces" of the spatiotemporal vacuum spacetime medium (i.e., quitons). Specifically, an observable object rotated more than one orthogonal turn away from the laboratory frame constitutes an anenergy particle. Examples are the virtual photon (rotation between 90 and 180 degrees), the bare neutrino (rotation exactly 180 degrees), the virtual bare neutrino (rotation between 180 and 270 degrees), and the thought or mind particle (rotation 270 degrees or more). Also, more fundamental components of energy—specifically, the individual scalar wave components of vector electromagnetic waves. Anenergy components may be coupled into energy, which can be compacted into mass. The conservation of energy law is replaced with a more fundamental law, the conservation of anenergy. Anenergy contains two com-

ponents or constituents: (1) ordinary energy (positive time, positive energy), and (2) phase conjugate energy (negative time, negative energy).

7. Angular momentum: the momentum or "leverage" of the momentum of a moving body, with respect to an axis or reference point Has the same units as action, or energy x time, or momentum x length.

8. Antimatter: matter consisting of atoms which are composed of positrons, antiprotons, and antineutrons. Also loosely refers to the antiparticle corresponding to a particle—the antiparticle may be regarded as the particle traveling backward in time, or "phase conjugated." In present physics, the photon is recognized as its own antiparticle. However, since physicists do not differentiate between positive and negative energies in a photon, then they assume that there is no difference between a photon and an antiphoton. The antiphoton, however, acts quite differently from the photon, as is clearly shown in phase conjugation. The magic rule of scalar electromagnetics for the final photon emission interaction with a charged particle in an atom is this: In the expression CPTEGS (where C is charge and normally negative, P is normal parity, T is time and normally positive, E is energy and normally positive, G is gravity and normal is positive, S is entropy and normally positive or increasing), if one of the components is reversed in sign, then each of the remaining components is also individually reversed in sign. The phase conjugate photon (antiphoton: negative energy, negative time), e.g., is produced by the photon interaction with the positively charged nucleus, producing a time-reversed wave having negative energy, negative time, reversed parity, causing gravitational repulsion, and causing negentropy.

9. Antiparticle: a counterpart to an ordinary fundamental particle, having identical mass lifetime and spin, but with charge and magnetic moment reversed in algebraic sign. Note that when the charge is reversed, parity and time are also reversed. Thus, properly, the antiparticle is the phase conjugation of the particle, or just conjugate particle for short.

10. Artificial potential: potential formed by deliberately summing force field vectors to a zero vector resultant This vector zero has a deterministic, internal stress pattern that is macroscopic. The type of potential made by the zero summation depends upon the type of force vectors summed. For example, summing electric field vectors to zero produces an electrostatic scalar potential; summing magnetic field vectors to zero produces a magnetostatic scalar potential of vacuum. All such potentials represents local warps or curvatures in vacuum/spacetime. Since Einstein assumed that spacetime locally was uncurved, this violates one of the assumptions of Einstein's general relativity. The Soviets have long made this correction; Western physicists have not Since curvature of spacetime is gravitational, this constitutes the secret of electrogravition: How to turn electromagnetics into gravitation, and vice versa. I have called this "scalar electromagnetics"; the Soviets refer to it as "energetics." In such a vector zero, the locked-in energies of all the component force vectors constitute the "anenergy" of the artificial potential. Potential in vacuum is analogous to "pressure" in fluids. An artificial potential polarizes the vacuum and constitutes a specific stress pattern in it Thus vector-zero summed systems of individually finite vector components may all differ from each other.

11. Ball lightning: rare form of lightning where an incandescent, slow-moving globe forms. The globe often moves erratically, explodes, and may set fire to objects it touches. Ball lightning is due to scalar electromagnetic interferometry which can produce a controlled pattern of electromagnetic energy at a distance. Phase conjugation is directly involved; note that the literature already explicitly demonstrates that phase conjugation can perform real-time holography, or produce forms of energy at a distance.

12. Bare electron: in physics, the true electron without the partial shielding of virtual positrons that are attracted around it In scalar electromagnetics, the mass of the electron, considered without its flux of virtual photons (its massless charge flux). An electron mass without its charge (in new approach); that is, one that is really bare of all charge!

13. Bare neutrino: the photon rotated one orthogonal turn farther away from the laboratory frame would be seen in the laboratory frame as a bare neutrino. In the new approach, bare neutrinos move at the speed c inside an ordinary photon, orthogonal to the photon's direction of travel, constituting a flux in each half of the photon. Two neutrinos move in the photon, one in each half (a simple photon is a single sine wave). One neutrino moves in positive time, the second moves in negative time and is properly an antineutrino. Since the photon moves at the speed c with respect to the laboratory frame, then the bare neutrino moves at the speed c-squared, but only its carrier speed c is seen by the laboratory observer. The flax of bare neutrinos in each half of a photon corresponds to the classical neutrino of present physics. The bare neutrino is (to the lab frame) a factor of c smaller than a photon, and one dimensional. It is a special kind of polarization wave to the lab observer.

14. Bearden, Thomas E.: nuclear engineer, retired Lieutenant Colonel (U.S. Army), and theoretical conceptualist active in the study of paranormal phenomena, scalar electromagnetics, Soviet Tesla (time-reversed wave) phenomena, free energy, antigravity, electromagnetic healing, and human development. Particularly known for his work in scalar electromagnetics and Soviet unorthodox weaponry.

15. Beck, Robert: noted psychotronics and ELF researcher who has pioneered many unusual measurement fields such as the Schumann resonance of the earth's magnetic field, specific effects on humans of extra low frequency (ELF) signals of precise waveform and frequency, ELF pollution effects, and brain wave entrainment by weak ELF signals of precisely controlled frequency. He is also noted for his development of a series of very sensitive detectors and scalar wave detectors.

16. Bedini, John: noted audio engineer, creator of the renowned Bedini amplifiers well known to audiophiles, and researcher into free energy, antigravity, and electromagnetic healing. Bedini has built several free energy motors and devices, including a gravitational field generator, a gravitational field turbine, and a series of monopole motors. He is also noted for his development of sensitive scalar EM wave detectors. He is struggling to develop an electromagnetic phase conjugate signal which will kill the AIDS virus, and also reverse the genetic damage done to the cell by the virus.

17. B-field: the magnetic induction B. This is defined, analogously to E, by the force it exerts. The force on a point charge moving in a magnetic field is at right angles to the velocity vector of the charge and also at right angles to the magnetic field. Actually, the H-field was originally named the magnetic field and the B-field was named the magnetic induction. This is an annoyance because it is B which produces or involves the force, and hence should be named the field. Other writers have noted this discrepancy, and Sommerfeld called H the magnetic excitation—which correctly describes its role—rather than die magnetic field. In Kaluza-Klein unified field theory, all electromagnetics is in the 5th dimension, which is wrapped around each and every point in our normal 3-dimensional space.

18. Being: pure essence, existence. Again, something which in actuality resists definition unless preliminary assumptions are made. Generally used in one of two senses: (1) persistence, as in "he is (he continually exists)," or (2) identity, as in "he is the man ('he' and 'the man' are the same thing)." Even in the first instance - persistence - one is using the form of the verb "to be" in the sense of identity, when perception is considered; i.e., "he is" is a statement that "he continues to be presented or perceived, because each succeeding perception is identical to the one before." Thus when perception is considered, being may be defined as that which is undifferentiated, and this yields its total definition. This is the real void devoid even of void, if one considers zen.

19. Bennett Island; Soviet island on which large, anomalous exhausts—some 150 miles long and nearly horizontal—are repeatedly photographed by U.S. weather satellites. Over 100 incidents have been photographed there since 1974. These exhausts are believed to be the exhausts of energy extracted by distant Soviet scalar interferometers used in the endothermic mode, then •witched to the exothermic mode and focused on Bennett Island as a "dump site." Similar exhausts have been seen on Novaya Zemlya, a large island off the north coast of mainland USSR.

20. Bigfoot: U.S. slang or common name for a materialized sasquatch or apeman-type tulpoid, even if not well-formed.

21. Big Mama: very large, glowing nocturnal light UFO believed sometimes associated with cattle mutilations. Implies a sort of "mother ship" which spews forth smaller ships, which then speed off, swoop down to mutilate cattle, and return. Big Mama has been observed to split into or spew off smaller lights which sped away, dipped down, then returned. Some of these occasions appear to have been associated with cattle mutilations. The most noted photographs of Big Mama were taken by Bill Jackson.

22. Biofield: a hyperfield, particularly with reference to the fields of a living biological system. Also bioframe, bioworld, hyperfield, hyperframe, hyperworld, orthofield, orthoframe, orthoworld.

23. Bioframe: a hyperframe, particularly with reference to the frame of the body of a living biological system. Also biofield, bioworld, hyperfield, hyperframe, hyperworld, orthofield, orthoframe, orthoworld.

24. Bioworld: an orthogonal universe one or more right angle turns away from the ordinary world, particularly with reference to the frame of the body of a living biological s*ystem.* Hyperworld. Orthoworld. Living energy forms—"critters"—exist in these bioworlds. A bioworld is also one of the netted levels of infolded virtual state. Such "living electricity" can be manipulated and utilized by scalar EM means.

25. Bohm, David: world renowned physicist and originator of hidden variable theory. Together with his student Aharonov, he co-authored a fundamental paper on the Aharonov-Bohm effect wherein interfering electromagnetic potentials can produce effects on charged particle systems, even at a distance and in the absence of the electromagnetic force fields.

26. Brain wave: a rhythmic fluctuation of tho voltage between various parts of the brain, normally ranging from about 1 to 60 Hertz in frequency and 10 to 100 microvolts in potential strength. Brain waves are caused by avalanche discharge firings of the nerve cells in the brain. Actually, classical brain waves are only the residue or "spillage" waves of the brain, and the important activity is in the specific patterning of the vector zero summations of the myriads of ion discharges. Every "discharge ion" constitutes a small EM force vector. The summation of these is largely zeroed; however, the patterns formed by all these tiny components are not random. The intent and will of the human being is expressed in the changes in the patterning of these dynamic substructures. Present EM brain wave theory does not touch the basic "thought patterns," which are scalar in nature. Use of the two cerebral hemispheres as a scalar interferometer provides mechanisms for the effects of parapsychology and paranormal phenomena such as fire walking.

27. Burst, burst-in: the sudden collection or condensation of anenergy into a frame, so that observable light or electromagnetic energy (or rarely, physical matter) appears (is kindled). Sudden kindling of anenergy into energy.

28. Cattle mutilations: the strange, paranormal, surgical killing and/or mutilating of cattle in the U.S. and other countries, particularly since the 1973 Yom Kippur War. Metapsychologically, mutilation of the cow is the mutilation of the female symbol—that is, it symbolizes mutilation of "Fortress America." Many small animals such as cats, dogs, rabbits, and birds have also been paranormally mutilated, which is destruction of the child symbol. Again, this symbolizes the mutilation of "Fortress America." The author regards such bizarre paranormal incidents as being caused by temporary materializations of the unresolved conflicts from the human collective unconscious's dreams. With this thesis, if knowledge of Soviet scalar electromagnetic weaponry is also included, analysis very similar to standard psychological dream analysis provides explanations consistent with the bizarre mutilations and the Cold War.

29. Causality: the time ordering of ordinary inert interactions. Macroscopic (observable) causality may be violated by control of hidden subquantal (virtual state) variables. With scalar electromagnetics, direct engineering of the virtual state—and hence violation of observable causality—*in* possible.

30. Cerebral hemisphere: each half of the cerebrum, the enlarged anterior or upper part of the vertebrate brain. Tho two cerebral hemispheres may function as a scalar interferometer to sense distant scenes (even distant in time), or produce energy and signals at a distance. The time aspect (e.g., in psychometry) is due to utiliza-

tion of the cerebrum as a phase conjugate mirror (PCM), and inputting tiny signals picked up from the object one is holding. A phase conjugate signal is then originated by the PCM, retracing the route to the original object. Unfortunately the output of the human nervous system that can accomplish such scalar EM functions lies very deep within the unconscious mind, and it is very difficult to raise it to direct consciousness. Usually the conscious mind is only aware of "vague impressions" instead of clear signals, although in the extremely rare instance clear impressions are consciously detected.

31. Changing the past: every intentional inception upon ordinary inert causality represents a slight alteration or change in direction, departing minutely from pure causality. From a causal viewpoint, each human inception minutely changes the causal past, as well as the future.

32. Charge: the repetitive, "locked in" massless virtual particle emission and absorption activity (flux) of and on a fundamental particle such as an electron. Vacuum is identically pure massless charge flux. In the new theory, the intensity of the charge flux on a particle is also directly related to the rate of the particle's passage through time. The continued absorption and emission of photons by a particle provides the general mechanism by which the particle moves along in time in little jumps. In the scalar EM view, the relativistic change of the rate of flow of an object through time is due to a change in its ambient vacuum charge potential—i.e., to a change in the rate at which it is absorbing and emitting virtual particles and scalar waves. The negative charge is associated with a positive time flux (hence with positive energy) and the positive charge with a negative time flux (hence negative energy). A positive charge is also a negative charge moving backward in time. All particles have clouds of virtual activity patterns; essentially all mass is charged since these clouds contain charges. In ordinary electrical physics, charge is defined in terms of force and force in terms of charge, and charge and charged mass are erroneously made synonymous. In scalar EM, a "charged particle" represents a mass (nonlinear bump in spacetime, or *local nonlinearity" medium) that separates the scalar waves of vacuum into two parts, one is normal energy and the other is the phase conjugate having negative energy. It is important to remember that a charged particle is a special sort of high pressure "pump" and "vacuum cleaner" at the same time. So is a "magnetic pole" or a "potential."

33. Cold explosion: the sudden extraction of energy from a distant intersection zone of the crossing beams of a scalar interferometer which transmits in the pulsed mode while negatively biased. A famous example of Soviet testing of a cold explosion scalar EM weapon is the huge cold explosion that occurred off the coast of Japan on April 9,1984. When the energy is suddenly extracted from a zone near the surface of the ocean and above it, the pressure in the zone is suddenly lowered. Water vapor and mist are violently sucked up, and the surrounding air rushes in, pushing the "water vapor filled region" rapidly upwards. The pattern that results is a rising mushroom cloud, very similar in appearance to a nuclear explosion, except for the absence of a flash and blast wave. In a cold explosion, the energy is extracted from throughout the spacetime of the distant interference region, and insulation has no effect since classical heat conduction through space is not involved. Ordinary electrostatic cooling is also an example of this effect in less violent form.

34. Consciousness: refers to a living being's awareness of its sensations, feelings, thoughts, and the world around it Specifically, consciousness involves an anenergy communication channel where detections of the body-object (which may be regarded as a sort of special camera, instrument, or physical tuner) project into the mind as separations from the beingness of the subject Further, it refers to the intersecting stream of monocular, one— at-a— time virtual projections into the mind from the quantum changes of photon interactions upon the body sensors. By memory (collection in the mind) of streams of such monocular changes one after the other, the entity is conscious of its body-object (camera—tuner) monocularly "moving through time." In the absence of deeper understanding, the individual consciously sees itself as a separated physical object moving and changing in time, i.e., with respect to other separated physical objects that it perceives. The "externality" of certain changes (i.e., the physical world) is due to the lack of mental control or influence over them. "Internality" of certain other changes is due to the ability to mentally control or influence them. The ordinary conscious mind is a serial processor: only one thing at a time is discriminated in its awareness. The unconscious mind, on the other hand, is totally conscious—but multiply so, since it is a parallel processor of many discriminations at once. The conscious mind, then, cannot directly perceive the individual discriminations of the parallel processor, since these appear only as a blur or nothingness to it (much like looking at a hundred 35mm slides in the projector at the same time). This is the mechanism of the barrier between the conscious an unconscious minds. Projections from the unconscious onto the screen of the conscious mind will thus appear symbolical—that is, having many hidden meanings at once. This is why our dreams, for example, appear to our conscious minds as weird and distorted, but highly symbolic in nature.

35. Contactee: a person who seemingly has had a direct contact—mentally, physically, or both—with aliens, extraterrestrials, etc.

36. Continuous mode: for a scalar interferometer or Tesla howitzer, mode of operation where power is fed continually to the transmitters, and the two Fourier expanded scalar patterns or beams are continuously transmitted. In this mode, in the distant interference zone a glowing pattern of vector electromagnetic energy appears and is sustained. (Note the possibility that a few of the UFO sightings could well be due to the deliberate usage of such a mechanism by the Soviet Union or some other nation possessing the requisite technology to create such physical "apparations" at a distance.)

37. Corpus callosum: the thick nerve cable that links together the two cerebral hemispheres (which are separate brains) in each vertebrate (including human) brain. In each of the two cerebral hemispheres, there is a separate mind and personality (given that the other is separated). However, a brain has a unique characteristic: Anything that arises directly inside it is automatically assumed to be self-originated, since the brain-tuner can differentiate no separation or "separate source" for the signal/thought arising directly upon its "internal screen". This allows integration of the two brains/minds/personalities into one.

That is, when one brain half prepares and sends a message, that message is also routed across the corpus callosum into the other brain half, arising internally upon the second brain's "display." The second brain half therefore thinks that it itself originated the signal or message. Thus when the left cerebral hemisphere moves the right thumb, for example, the right hemisphere thinks that it itself ordered the movement of the right thumb. Perceptually, the definition of identity is the loss of all distinction."

38. Cotton effect: the wavelength dependence of certain optical and absorption characteristics in a substance in the vicinity of an absorption band. An absorption band is a range of electromagnetic frequencies within which electromagnetic energy is absorbed by a substance. The direction of polarization of the energy affects the degree of absorption for certain substances. In these substances, the frequency and the direction of polarization of incoming electromagnetic energy affect the amount of energy deposited and dispersed.

39. Crawford, Hal: accomplished science fiction artist and illustrator who has furnished many drawings in the new approach and contributed to the theory of mind linkage.

40. Crosstalk: in ordinary science, the transfer of energy or signal from one channel to another, by cross modulation or cross coupling between the channels. In the new approach, it refers to virtual anenergy exchange between orthogonal universes or frames—that is, between different 3- or 4-dimensional slices of an infinite-dimensional universe.

41. Cyborg: (noun) a single integrated system consisting of a living system energetically or functionally linked to mechanical or electromagnetic devices and/or systems. (Verb): link, in the sense of forming such a system. A living biological system is a cyborg of its mindworld (mindframe) to its body (laboratory tuner).

42. Death: the loss of the preferential tuning of a physical body to its mindworld. This causes loss of the preferential consciousness loop or life channel, resulting in "death." Actually the being is unchanging and indestructible, hence consists of all changes and destructions simultaneously, with none separated individually. Life and death thus are two sides of the same coin, and the coin is immutable, regardless of which side is exposed.

43. DeBroglie wave: a matter wave. The quantum mechanics wave associated with a particle of matter, which can give rise to interference effects. An orbital electron in an atom is associated with a standing DeBroglie wave on a Bohr orbit. DeBroglie waves always move faster than the speed of light, the speed being given by c2/v, where c is the velocity of light in vacuum and v is the velocity of the particle of mass. Debroglie wave velocities thus vary from the speed of light to infinite speed (everywhere at once). In his speech accepting the Nobel Prize, DeBroglie emphasized that , since these waves produce real physical effects, they are real and must not be regarded simply as mathematical conveniences.

44. Dimension: a certain primary geometrical physical attribute, such as length, used to describe the separational relationships of physical phenomena. By "geometrical" we mean that the "dimension" is considered to "exist" in either the presence or absence of the observable physical phenomenon; however, by agreement the dimension itself is not directly observable. It can be inferred ("measured"), however, by comparative operations of simple devices such as clocks and metersticks, and its simplistic definition is usually given as "that which has measurable magnitude." Prescribed measuring procedures are used to assign magnitudes to these inferred dimensions, using specific scales. Actually, "dimension" just means "fundamental quantity of separation" in the most general sense. The number of dimensions taken in advanced physics usually represents a somewhat arbitrary choice for good modeling fit In modern Kaluza-Klein theory, for example, it becomes necessary to accept some 10 or 11 dimensions.

45. Dirac sea: in Dirac electron theory, the relativistic wave equation for the electron has four components, which correspond to two spin orientations and two energy conditions: positive and negative. Thus the vacuum possesses a "sea" of negative energy conditions for electrons. These are called "holes" and are considered to ordinarily be filled with electrons which have fallen into these negative energy states. The Dirac sea thus is this vacuum sea of negative energy electrons. If energy is added, electrons can be lifted from this sea. If spacetime is locally curved, additional holes may be created so that normal positive energy electrons fall in and disappear, creating the appearance of an "electrical energy sink." If the curvature is in the opposite direction, electrons may be lifted from the Dirac sea by the curvature, thus producing a flow of electrons and creating the appearance of an "electrical energy source." In several battery-powered free energy motor devices where die battery is sharply "back-spiked" for recharging, the battery has been noted to apparently accumulate such additional "Dirac holes" as the DC series—wound motor ran. When later recharging the battery with a normal charger, for example, the battery would first "consume" a quantity of electricity to "refill the holes." When the holes were filled with new electrons, the battery would then start to recharge normally. In one extensively used battery, Watson noted that it consumed electricity for some eight hours before recharging normally. Similar effects have been noted by Bedini.

46. Dissociation: separation of ideas from their affects or feelings, resulting in independent functioning of these components of a person's mental processes. Simply put, fragmentation of the mind and personality. In this sense, Zarg (the collective human unconscious mind) is severely dissociated, since its "conscious mind" is fragmented into billions of pieces (into the independent conscious minds of ordinary humans). This is severe dissociation and psychosis; Zarg (the human species) is quite literally insane.

47. Dowsing: use of handheld rods, forked twigs, pendulums, bent coathangers, or other simple physical detectors to locate water, minerals, lost objects etc. or to seek answers to specific questions. The movement of the physical instrument serves as an indicator and is cyborged to the dowser mentally and psychokinetically.

48. E-field: the electric field or electric intensity E. That field on a spinning charged mass which produces an electrical force, hence linear motion of die charge. (In the new theory, the E- field CONSISTS OF the spinning charged mass, moving through a finite change in electro static scalar potential, and thus there is a more fundamental generatrix of FORCE.) Defined as the force per unit charge (i.e., per unit of charged mass). On a free

electron, there may exist two separate E-fields, one in the direction of motion of the electron (normally due to voltage); and one at right angles to the electron's motion, due to the time derivative of the A-field.

49. Einstein's postulates: (1) all the laws of physics are equally valid in all inertial frames of reference, (2) the speed of light is the same to every inertial observer, and (3) the observable local effects of a gravitational field are indistinguishable from those arising from acceleration of the frame of reference. The first is called the special relativity principle, the second is called the law of light propagation, and the third is called the equivalence principle. It is now known that postulates (1) and (2) are not independent of each other. Note that the third postulate implies that any local force is due to a "gravitational field" (acceleration of the frame of reference). It also implies that gravitational force has a mechanism, since quantum mechanics assigns a differential operator, operating upon a potential, to create every force. By extended inference, it can be seen that the third postulate also implies that there should exist a method to turn any type of force (field) into gravitational force (field).

50. Electrogravitics: electrogravitation, or scalar electromagnetics (SEM). The term "scalar electromagnetics" was chosen by Bearden to describe the electromagnetics that goes on inside a macroscopically summed-zero force vector system, producing gravitational effects and energy effects at a distance. A priori, SEM is a unified field theory, which predicts strong effects easily verified experimentally. The struggling science that seeks to produce gravity and inertial effects electromagnetically (success has been achieved recently in proprietary labs). Santiliy has shown that the classical assumption that electrical field and gravitational fields are mutually exclusive is false. They are therefore either the same thing or partially the same thing. There are numerous anomalous experiments which also show a connection between electrical effects and gravitational and inertial effects. Appreciable technical work in the U.S. seems to have been done on electrogravitics in the 50's and 60's, much of it apparently classified. There are presently strong electrogravitic efforts in Germany, France, and Japan. Scalar electromagnetics is in fact electrogravitics (electrogravitation) and has been successful in engineering gravitational devices such as gravitational and inertial field generators, gravitational field motors, free energy generators, and gravitational turbines. Tesla's secret of wireless transmission of energy at a distance without losses has been rediscovered, as has the secret of his "big eye that can see at a distance." A new land of resonance—scalar resonance—has emerged, as has a new kind of wave, the scalar EM or electrogravitational wave. Scalar wave detectors are now becoming available due to the pioneering work of Dea, Beck, Bedini, and Golden, who have all built scalar wave detectors. A substantial series of devices have been built by Bedini, Golden, Watson, and other inventors who wish anonymity. One of the very first good experimental physicists who built electrogravitational devices was Hooper. Actually, Maxwell's original quaternion theory, published in 1864, was the correct unified theory of gravitation and electromagnetics. After Maxwell's death, Heaviside "translated" (mutilated) Maxwell's theory into his new vector mathematics, discarding the quaternion's scalar component. He thus discarded electrogravitation, setting "modern EM theory" upon its nongravitational course.

51. Electron dissolution: when the virtual photon activity of an electron (i.e., its charge) is directly quenched, the electron is converted to a neutrino or neutrinic wave, which radiates away without interacting with materials or objects. In this case, no ordinary photons are radiated. Orbital electrons in atoms are generally protected from this interaction by being locked into their orbital DeBroglie waves. Free electrons flowing in circuits, however, are susceptible to the effect by scalar wave coupling. By this means, electromagnetic circuits may be dudded by dissolving their currents, without physical damage to the circuitry components. Whenever the agent causing the electron dissolution is removed, currents resume and the circuits resume normal electrical action. Scalar longitudinal 0-field waves can also be absorbed in scalar coupling into the E-fields of currents flowing in a circuit. This can either augment or diminish the E-fields to which coupling occurs. If the E-fields on the flowing electrons are zeroed, the electrons cease flowing, and the circuit is dudded.

52. Electrostatic scalar potential (0): virtual photon stress on the 4-dimensional spacetime medium (on the vacuum-ether). (Note that the vacuum can be regarded as consisting of the sum of a large group of such potentials.) The spatiotemporal intensity of the massless charge (of the virtual photon flux through a spatiotemporal point). Nikola Tesla discovered pure longitudinal scalar 0-waves. Present spatial 0 in the ordinary theory is only a special case. 0 can be visualized as a special type of "pressure" in an n-dimensional fluid, where n is equal to or greater than 4. By omitting the potential's stress on time, present potential theory eliminates the potential's ability to affect everything existing in time (for example gravity, fundamental constants of nature, inertia, mass, rate of time flow, mind, life, etc.) It also eliminates the direct engineering of local general relativity, and the concomitant violation of the conservation laws (momentum, energy-mass, charge, spin, etc.) Present electrical theory—and indeed all of physics—is drastically in error by this serious omission. Direct engineering of artificial potentials by patterned zero-vector force field summations is the secret of the long-sought "unified field theory." It is also the secret of the unification of mind and matter into a single scientific discipline.

53. EMP (electromagnetic pulse): a sharp pulse of electromagnetic radiation produced when an explosion occurs in an unsymmetrical environment, especially at of near the earth's surface or at high altitudes. Essentially the same as an electromagnetic shock wave, a wave of great intensity which results when waves with different intensities propagate with different velocities in a nonlinear optical medium. An EMP is also produced when two intense longitudinal scalar electrostatic potential wave patterns meet and couple into a sudden flash of vector EM energy.

54. Endothermic explosion: the sudden absorption of electromagnetic energy in a spacetime area where the energy converts to a potential, and scalar waves radial back to the negatively biased transmitter/receiver, emerging there as ordinary electromagnetic energy. As implosion. All normal explosions are exothermic in nature, and ordinary electromagnetic heat energy is radiated away. A scalar EM howitzer, however, may pro-

duce an endothermic explosion (an energy implosion) at a distance, explosively extracting the heat energy at the implosion site and receiving the energy back at the transmitter site for disposal.

55. Energy: the capacity to do work. Energy is directly tied to the work concept, which is the result of a force moving through a distance. However, "work" refers to the past; that is, the force has moved through a distance against a resistance. "Energy" refers to the future; the force could in the future move through a distance against a resistance. Since the force is tied in turn to a changing mass motion, it follows that—conceptually—energy is tied directly to mass systems. The "energy" that an object possesses depends also upon the observational situation: a moving observer may see a moving object as possessing more or less energy than the lab observer sees. Electrostatic scalar waves do no work and individually have no "energy" per se. Two different scalar waves coupled onto a spinning charged mass, however, produce a system which has mass and energy, and does work.

56. Ether: originally, a ghostly-thin, material, fluid medium postulated to fill space and carry electromagnetic waves, similar to the way in which a gas carries sound waves. The Michelson-Morley experiments eliminated any such material ether, but not a Lorentz-invariant nonmaterial ether (of which vacuum itself is one example, in the modern view). Maxwell's equations were derived from ether theory. Einstein once proposed that the vacuum, complete with its electromagnetic fields, should be called the ether. Indeed, the word "ether" is again acceptable in the literature in context of its modern meaning. Today the "ether gas" is considered composed of virtual particles—that is, the particles spontaneously come into being and then vanish so quickly they cannot be individually detected. That is, today the ether is a virtual particle gas—sort of a gas of "effervescent bubbles", so to speak. Ives derived special and general relativity, Newton's laws of motion (relativistic form), and the square law of gravitation from a Lorentz-invariant nonmaterial ether theory. Zero-point energy of vacuum would also seem to be a Lorentz-invariant ether. The Soviets are believed to have applied and advanced the Ives theory to produce their present theory of energetics, which is highly classified by the Soviet Union. In the new approach, vacuum is identically pure massless charge (virtual particle flux), and thus an ether as Einstein proposed.

57. Evil: refers to an act theoretically detrimental to the receiver or referent, in the positive time stream. It is equally beneficial to the referent, viewed in the negative time stream.

58. Exo-electron emission: emission of electrons from the surfaces of metals and certain ceramics after these surfaces have been freshly formed by a process such as abrasion or fracture. Cracks in metals are good emitters of exo-electrons, and the electrons can be detected, to provide a means of detecting the minute fatigue or stress cracks. The sides of a crack in a stressed metal constitute a scalar interferometer, and scalar interferometry is involved in the emission process. Exophotons are also emitted, in addition to electrons. In an earth tectonic plate fault zone (crack), stress exists on each side of the plate, but not in the middle, and the fault thus constitutes a scalar interferometer, emitting charges and also often creating glowing balls of light

59. Exothermic explosion: an ordinary explosion where ordinary heat and electromagnetic energy are released and radiate away from the explosive site. All normal explosions are exothermic.

60. Feynman diagram: a two-dimensional diagram of particle interactions, both in the virtual and observable levels, and both backward and forward in time.

61. Fireflies: sentient bits of light that sometimes kindle in the presence of a psychic initiator and/or a psychotronic generator. Literally, minute glowing, indefinite thought forms materialized into electromagnetic energy. This can be understood by modeling thought forms as real electromagnetic forms existing in a rotated universe (i.e., as hyperdimensional). Interference causes dimensional shifting, so that ordinary electromagnetic energy forms can result

62. Firewalking: walking on and across a bed of hot glowing coals without the feet being burned. Firewalking is demonstrated in a variety of cultures, belief systems, and locations throughout the world. Recently many Americans have been taking to the practice and demonstrating that it is an achievable phenomenon. In the new approach, the two cerebral hemispheres act as a scalar interferometer. If this biological scalar interferometer creates an endothermic (extraction) field on the surface of the skin of the feet, and simultaneously creates an exothermic scattering of the extracted energy into the surrounding environment, then one can walk on hot coals without burning the feet. Note that the reverse effect-cooling of the external environment—is often experienced when psychokinesis effects at a distance are obtained. For example, poltergeist effects are often associated with an adolescent If the adolescent's cerebral scalar interferometer extracts by inverse scattering, and projects exothermically into distant objects, the movement of the objects can be explained.

63. Fission (action): the splitting of the action quantum into two canonical variables (for example, energy and time), only one of which can normally be completely detected or measured. ("Action" has the dimensions of angular momentum—energy x time, momentum x length, etc.) Fissioning produces length, discrete bit by discrete bit, and time likewise. The physical reality of separated physical objects is thus continually created, discrete bit by discrete bit, monocularly. All that ever exists in the "present" is a single quantum change, just occurring. Recombining of the two canonical action fission fragments is called "fusion." Fusion also occurs monocularly, quantum by quantum. Rigorously, the physical world of separated objects never exists as such except in memory recall of vast numbers of past ordered quantum changes that occurred one by one.

64. Flux: the amount of some quantity (such as energy, particles, volume of fluid, etc.) flowing across a given area (usually a unit area perpendicular to the flow) per unit of time. The term flux density is now more commonly used. Loosely, "flux" refers to the perpendicular flow through an area, while "flux density" refers to the magnitude or intensity (rate) of die flux. In some cases such as a flux of particles of different velocities, the number of particles may be multiplied by the average velocity to give an average flux density.

65. Flux density: the amount of tome quantity (such as energy, particles, volume of fluid, etc.) flowing across a

given area (usually a unit area perpendicular to the flow) per unit of time.

66. Flying saucer: a discoid or other shaped paranormal object observed in the skies and/or landing, taking off, etc. Came into general use in the U.S. in 1947 after Kenneth Arnold reported seeing nine strange flying discs skipping along like saucers. Sometimes used to refer to an unidentified flying object (UFO). The usual "flying saucer" is theorized to be due to temporary materializations from the collective human unconscious; over 80% of these materializations only go so far as the "glowing light" stage, not to the material object stage. A small percentage do physically materialize, then later dematerialize. A small fraction of one percent are considered to be actual ships built on earth by the Soviet Union or one other country. A few sightings may be due to deliberate scalar interferometry at a distance by the Soviet Union and others.

67. Foo fighters: glowing balls of light which occasionally appeared in the skies and played tag with Allied, Nazi, and Japanese planes during World War II. The word "foo" is a corruption of the French word "feu," meaning fire.

68. Force: in mechanics, identically the time derivative of momentum. Simply, changing of mass-motion. One must realize that it is not the mass in motion, but the single undivided quantity "mass-motion" or "mass-velocity" (momentum) that is changed. In present EM theory, force fields are erroneously taken to exist in empty, massless vacuum. However, the mechanical definition is an identity, not an equality. Thus force does not EXIST ON a mass, but CONSISTS OF a changing "mass-motion." Since this force can be induced at a distance (by changing an introduced mass-motion such as a test charge), one is confronted in the new approach with the fact that there is a more fundamental mechanism that produces force itself. This mechanism is coupling of nonidentical scalar fields by and on a spinning charged particle. The INDUCED FORCE consists of the particle's change in mass-motion per unit of time. It is thus a flux—induced quantity in, on, and of a mass in motion.

69. Force field: an effect that can exist in vacuum or on mass whereby a particular type of matter is affected. For example, the "electrical (force) field" can affect and change the state of motion of any charged particle. In classical physics, the force fields are the causative agents; when these fields are zero, then the motion of the particle or system is unaffected. The electric field (E-field), for example, is defined as the force per unit of fundamental charged mass, though usually it is stated as "force per unit charge," with the mass ignored. Rigorously, however, force does not exist in vacuum, in the absence of mass, a priori. The Aharonov-Bohm effect shows that the potentials can interfere to produce effects in charged particle systems—even at a distance, and even when the force fields are reduced to zero. Tesla stated that vector Hertzian waves do not move through vacuum, which is true. Rigorously, two scalar waves move through vacuum instead. When coupled (detected) by a spinning charged mass (an electron, for example), the vector E and B fields are created on and of the moving mass itself; they do not exist in vacuum. This has been a fundamental omission and error in both electromagnetics theory and mechanics. For example, the standard EM equation, E = - 0 is not generally true in vacuum, but it is true on an electron moving in a detecting antenna or on a charged particle moving in vacuum. Any force is produced by the same fundamental mechanism: flux differences encountered by a spinning charged mass. Thus there is a common connecting thread among all forces, and this is the key to a practical, engineering unified field theory. Note that Einstein's equivalence principle, after much pondering, can be taken to be an equivalent statement to the fact that force—any force—consists of a change in mass-motion. In modern quantum mechanics, force is an effect, not a cause, and is derived from the potential by a differentiating operator. The potential is the primary causative agent, not the force. Also, in modern quantum field theory, all forces actually come from curvature of spacetime—i.e., they are all gravitational in origin.

70. Four-space (4-space): a 4-dimensional spacetime, dimensionally composed of length cubed multiplied by time. In the absence of gravitational field, the 4-d spacetime reduces to a Minkowski space.

71. Fractional charge: a hypothetical particle having a charge less than the electron's charge. Often called a quark, where its charge is 1/3 or 2/3 the charge of the electron. Recently, Stanford University researchers have found evidence of free quarks, and in the 1920's Ehrenhaft reported the discovery and production of fractional charges of all sizes.

72. Frame (Lorentz): a frame of reference which is not accelerated with respect to the laboratory observer. In this case, the frame may be rotated (moving at a constant velocity) but it is not rotating (the frame is not accelerated with respect to the observer, and spacetime is not curved). In this case special relativity applies, as do the conservation laws. If the local spacetime is curved, the frame is accelerated and spacetime is curved. Locally, then, general relativity applies. In this case, locally the conservation laws (energy, momentum, charge, and spin) can be violated. However, in ordinary general relativity, it has been tacitly assumed that locally the curvature of spacetime is negligible. Hence locally a Lorentz frame is assumed, and the conservation laws are assumed to apply locally also. In scalar electromagnetics, the local spacetime is always curved and general relativity applies locally. In this case, the conservation laws need not apply.

73. Frame (of reference): a spatial, organized, measured lattice placed in emptiness (space, spacetime). Normally refers to a 3-dimensional, spatial frame. All objects and points in the "universe" or spatial frame are considered to simultaneously coexist at separate, measured points in the frame. Differs from the vacuum in that, rigorously, vacuum has no existing definite lengths and no existing definite time intervals, as these appear only after measurement/detection, and are relative to the observer and to the detection process. The "laboratory frame" is the static reference frame of the observer/measurement. A separate reference frame may be assumed to exist for any fixed or moving object, or centered on any point in another frame. When a type of frame is assumed, the entire class of physical interactions that can occur has been restricted to an assumed set or type. In other words, given the frame, the physics has been assumed. One of the greatest restrictions of an assumed "frame" is to rule out the consideration (existence) of other higher dimensions. In the new theory, higher dimensions are permitted

and an orthogonally rotated frame can be equated to a field and vice versa. Therefore frames are also fields.

74. Free energy: energy obtained from a free or essentially free source. A paddlewheel in a river extracts and provides "free energy" from the river, since one need not pay for the river. Electromagnetically, the idea of "tapping" or extracting energy from an esoteric source, such as the zero-point energy of vacuum, the Dirac sea, a standing electrogravitational curvature of local spacetime, etc. Since scalar electromagnetics is electrogravitation, a locally bent spacetime can be achieved so as to provide either a source or a sink, as one wishes. Thus a practical free energy device is rigorously possible by scalar electromagnetic means.

75. Fusion (action): in the new approach, the recombining of two canonical bits of a fissioned action quantum. The fusion process "cancels" the bits produced by the fission process, and constitutes a negative spacetime operation. The physical world is thus eliminated, bit by bit, in a monocular fusion operation that produces the "past" The "mass" of a particle is taken to be proportional to, and identically comprised of, the absolute value of its fission and fusion rate. When fission and fusion rates of a particle are equal, the particle is unaccelerated; when unequal, the particle is accelerated. Vice versa, when the particle is unaccelerated, the fission and fusion rates are equal, and when the particle is accelerated, the fission and fusion rates are unequal.

76. Geiger tube (Geiger-Muller tube): a tube used in appropriate circuits to detect and count ionizing particles. Each particle crossing the tube causes ionization of the gas in the tube, essentially independent of the energy and nature of the nuclear particle. Thus a uniform discharge occurs across the tube. The tube will detect anything which causes an increase in the ionization of its gas. Since strong scalar waves will do this, the tube is capable of detecting scalar wave radiation of sufficient intensity. In many cases, GM tubes are activated near UFO landing zones, and sometimes near cattle mutilations or other paranormal occurrences. In such cases, the tube is reacting to scalar emissions, not nuclear radiations.

77. Geller effect: refers to the frequent stimulation of paranormal abilities in children and other people who have seen Uri Geller (or someone like him) performing psychokinesis either in person or on television.

78. General relativity: the physics of curved spacetime. As opposed to special relativity which deals with inertial frames moving at constant velocities with respect to each other (Lorentz frames), general relativity deals with accelerated frames. General relativity was founded by Einstein, who made highly restrictive assumptions to simply the complexity of the problem. Briefly, Einstein assumed that, since no means of appreciably curving spacetime was known except by the gravitational force of a tremendous collection of mass, then the only places where appreciably curved spacetime would be met would be near suns, stars, etc. Accordingly, he assumed that the local spacetime of the observer would not be curved, and instead the local frame would be a Lorentz frame. That is, he made the theory a sort of "special relativity with distant perturbations." That way, the conservation laws would be saved, and one could still speak of the energy of objects as opposed to the energy of entire systems. General relativity complete with this severe restriction is called "ordinary general relativity" (OGR), and this is the type of relativity taught in all our Western universities. Note that, by assuming no local spacetime curvature, the ordinary theory actually assumes that one cannot directly (locally) engineer general relativity. OGR is a highly special case, and not at all the type of general relativity we seek in the new approach where we wish to deliberately curve local spacetime and engineer general relativity directly on the laboratory bench. The Soviets do not recognize Einstein's limiting assumption. In Soviet physics journals, papers are regularly published that treat general relativity in an unrestricted manner. These papers include local curvature of spacetime and violation of all the conservation laws as a result. In the West, conservation has been raised to an absolute dogma, and such papers are regularly refused publication. Conservation laws are upheld in the West by artifice and dictum, not by science or experiment! So if the various restrictive assumptions placed on ordinary general relativity are removed, then an "unrestricted general relativity" results. In this relativity, the local frame may be appreciably curved, and any or all of the conservation laws can be violated—often with ease. With an adaptation of Kaluza-Klein 5-d unified field geometry and use of artificial EM potentials deliberately created by summing or multiplying EM force field vectors to vector zeros, EM energy can be transformed into gravitational field energy locally. In this fashion one can engineer general relativity; violate conservation of charge, energy-momentum, and spin; produce free energy engines; transmute elements simply; and produce action at a distance. Initial experiments have positively established the new approach as valid, and a startling new physics is in the offing. In the new scalar EM, gravitational potential becomes locked-up (infolded) electromagnetic forces, and electromagnetic force fields become bleed-off (outfolding) of the inner contents of the gravitational potential.

79. Ghost rockets: materialized tulpoidal "rockets" that appeared in 1946 and 1947 over the Scandinavian countries (Norway, Denmark, and Sweden) and usually seemed to come from the direction of the Soviet Union.

80. Giant photon: a photon with an organized, deliberate scalar substructure. In scalar electromagnetics, the photon is assumed to be one oscillation of the electromagnetic field; thus it is comprised of one E-field wave and one B-field wave. Further, in each point of either wave, a finite group of vectors may exist such that their vector summation or multiplication is zero. Thus a deterministic substructure may be deliberately impressed upon (infolded in) the photon. Such a photon may carry enormous infolded energy in the form of anenergy, and may be designed to perform a particular resonant purpose. In other words, such a photon becomes a "vacuum engine" that structures the virtual particle flux of vacuum in a particular, deterministic manner.

81. Golden booster: a potent scalar wave generator invented by Frank Golden, and used to augment and amplify the Kelly radionics machine. The booster has the unique characteristic of gathering in ambient energy and EM radiation from its surroundings and transposing this into highly patterned scalar radiation. It has the further characteristic of reacting to human intent, and constructing its patterning accordingly. In addition, two motors are incorporated whose needle movements provide

visual indication of the rate of treatment, the frequency of treatment, and how the dosage is progressing. A Bedini scalar wave detector directly verifies that both the Golden Booster and the Kelly radionics machine produce and radiate potent scalar waves.

82. Golden, Frank: inventor and scalar electromagnetics researcher who has developed a variety of scalar wave antennas, an underwater communications system, scalar wave transmitters and receivers, the Golden Booster, scalar wave detectors, and scalar wave spectrum analyzers. He has also developed a prototype free energy motor, a motor that runs off reactive power, a solid state scalar radar antenna, and other significant scalar EM devices.

83. Good: refers to an act theoretically beneficial to the receiver or referent, in the positive time stream. It is equally detrimental to the referent, viewed in the negative time stream.

84. Gravitational field: rigorously, the curvature of spacetime. In scalar EM, gravitational potential is locked-up (infolded) electromagnetic forces which sum or multiply to zero. Electromagnetic force fields represent the "bleeding off" or "outfolding" of these infolded contents of gravitational potential. Gravitational force field is the "bleeding down" of the gravitational potential without outfolding or out-bleeding its infolded contents. Gravitational field is not a field in the sense of Heaviside-Maxwell electromagnetic fields. There is really no such thing as "the" gravitational field; instead, there exists a large number of potentials, locked-in force field vectors, and other entities which change the energy density of vacuum and bend spacetime. The bleed-down of this conglomeration infolded components as a conglomeration constitutes the "gravitational field." If the inner contents of this G-field, or of the G-potential, bleed out, then a different field—e.g., electromagnetic field—is produced. We may thus speak of the bleed-out of either the G-potential or the G-field. Note that flows between pairs, triads, etc. of these individual infolded field components can occur without appreciable change of the overall conglomerate's potential (trapped energy density). Thus the gravitational potential can have a dynamic, structured pattern, and local spacetime can be curved and patterned. E field and B field are one such "pair" of flows where little overall gravitational effects are evidenced, so long as the fields are tree to "flow." In Kaluza-Klein 5-d unification of gravitation and electromagnetics, only a single field—the 5-dimensional gravitational field—really exists. Our world intersects and "bleeds-off' this 5-d G-field in two fashions: (1) the 5-d G-field prefers to bleed-out into its fifth dimensional aspect, which we see as our own ordinary EM force field. (2) The minute remaining 4-d aspects we see as our 4-d gravitational field. To show the order of magnitude difference in bleed-out, the electrical field between two electrons is about 10^{42} or so stronger than the gravitational field between them. Even for protons, the ratio is 10^{36} or so. So it can be seen that, so long as EM bleed-out is permitted, little 4-d gravitational effects are produced. Our normal electromagnetics—which by construction of Heaviside's version of Maxwell's theory is purely bleed-out, thus yields negligible gravitational effects. On the other hand, if we deliberately set the EM bleed-outs against each other so that they zero sum, we "strangle" the escape or bleed-off of the 5-d G-field into EM field. The 5-d G-field is then forced to bleed-off into 4-d G-field. If we then vary the energy of each vector component of the EM zero sum while maintaining the overall zero resultant, we directly pump the 4-d gravitational field. Scalar electromagnetics thus directly converts EM field, force, and energy into ordinary G-field force and energy, and vice versa. In addition, we get enormous gains: For every unit of EM energy we feed into the "zero sum pump," theoretically we can get out up to 10^{42} units of 4-d gravitational/inertial energy by electron release, and up to 10^{36} by proton release. Experimental results have confirmed this thesis. As early as 1969, Hurvich produced an inertial field generator capable of "freezing" weapons so they cannot be fired. Hutchison reports and demonstrates the levitating of up to 60-lb. objects. The American physicist Hooper patented electrogravitational devices in the early 70's. Bedini, Watson, Johnson and others have developed practical prototype "free energy generators." Another inventor who wishes anonymity has developed a small device which produces negative energy and antigravity, at an amplification factor of 1,500,000. Orthodox research in this area is just beginning in the West, but has been highly developed for several decades in the Soviet Union.

85. Gravitostriction: effect in a resistor whereby local gravitational stress flow in a direction results in production of a macroscopic electric force field in that direction, a macroscopic magnetic force field orthogonally around that direction, and minute, randomized EM fields (heat energy) microscopically due to induced microscopic particle motion and interaction. It is the enforced spin alignment of the mass particles (such as conduction electrons) which accounts for the production of the magnetic field. Gravitostriction is the reason that resistive loads can be powered by scalar anenergy flow (electrogravitational flow, massless current flow) and by negative energy flow. During the initial chargeup of a helical coil exposed to a sine wave voltage, no electron mass current flows; however, the massless current (phi-dot) charges up the coil gravitationally. As electron mass current begins to flow (90 degrees into the wave), this gravitational field is bled-off (outfolded) as electric force field (electron mass flow coupled to del phi) and magnetic force field. One can do amazing things with coils and transformers which are not allowed to enter the EM bleed-off phase at greater than 90 degrees

86. Gravity: the phenomenon characterized by the physical attraction of any two material bodies. Ideally, the gravitational force between two masses is proportional to the product of the masses divided by the square of the distance between them. In the new approach, one recognizes that mass itself is defined in terms of, and identically consists of, the time derivative of an action flux. Gravitational field and electrical field, then, become different aspects of the same thing. By affecting the flow of time, scalar waves and scalar potential thus are able to affect the gravitational force between two objects. Antigravity, for example, can be achieved by forcing an object's atomic nuclei to produce phase conjugate (time-reversed) waves (and negative energy and negative time). In negative time, gravity is a repulsion. Hutchinson has demonstrated the levitation of objects weighing up to 60 lbs, albeit sporadically.

87. H-field: historically, the magnetic field. However, it

is poorly named, and should be considered the magnetic excitation, after Sommerfeld. In empty space B and H are proportional, an experimentally determined fact In the new approach, rigorously vector fields cannot exist in vacuum, but can only exist on an observable mass. A "shadow" vector field can exist in vacuum, in the absence of observable mass, as small virtual vectors, each existing on a virtual particle in the vacuum. Such a "shadow" vector field may be regarded as two coupled scalar fields, where the coupling is performed by the virtual particle flux of which the vacuum itself consists. When an observable, spinning charged mass encounters the shadow vector field, it converts the two coupled scalar fields into a vector field. Thus Hertz waves per se are not transmitted through vacuum; "shadow" Hertz waves arc. With an ordinary detector, however, one detects vector Hertz waves, since one detects the precession waves of the electrons in the electron gas of the detecting probe or antenna.

88. Hieronymus, T. Galen: renowned radionics experimenter and inventor of the amplified radionics machine. Hieronymus is the father of modern amplified radionics; most experimenters before him used unamplified machines.

89. Hieronymus device: essentially, an amplified radionics device. Refers to a device similar to the one patented by T. G. Hieronymus in the 1940's. Connotes ordinary amplification rather than holographic amplification of radionics effects. Any similar type device. The Hieronymus device actually processes, and operates with, scalar waves.

90. Hologram: a photographic recording of the pattern of interference between coherent light reflected from the object of interest, and light that comes directly from the same source or is reflected from a mirror. When this photo image is illuminated from behind by coherent light, a three-dimensional image of the object appears in space. The characteristic of a hypothetically perfect hologram is that all its content is contained in any finite part of itself. If one portion is cut out and illuminated, the complete image appears in space. The whole hologram is in each part of itself. Observationally and perceptually, the physical universe is a hologram and in each part of itself, since all of it can be detected from/in each internal particle.

91. Hyperchannel: a channel between orthogonal frames or hyperframes, where anenergies crosstalk. Exceptionally good hyperchannels seem to be frequency dependent, and are referred to as windows, magic windows, or interdimensional nodal points. Some hyperchannels are 38-40 kHz, 150-160 kHz, 1.1-1.3 MHz, 1.057 GHz (the Lamb shift), and in the near ultraviolet (the life force frequency). These frequencies also represent enhanced channels between subquantal (virtual) and quantal (observable) states. A particularly good magic window exists when the infrared and ultraviolet bands being utilized are phase-locked so that the ultraviolet represents a first harmonic of the infrared.

92. Hyperfield: a structure or disturbance in hyperspace or higher dimensional space. With respect to a fixed frame, the electromagnetic field or a more subtle field. We consider the electromagnetic field the first hyperfield. A hyperfield may also be considered to be the electromagnetic field rotated further away from its laboratory frame. The second hyperfield is called the neutrinic field, and the third hyperfield is the mental or mind field. Also called bioframe, orthoframe, hyperframe, biofield, or orthofield. These hyperfields/hyperspaces also correspond to the nested levels of virtual state. Successively higher hyperfields past the electromagnetic field constitute successively lower levels of virtual state.

93. Hyperframe: simply, a frame rotated one or more right angle (orthogonal) turns away from the reference laboratory frame. The laboratory frame is considered to be the zeroth hyperframe. Also called bioframe, bioworld, biofield, orthoframe, orthoworld, orthofield, hyperfield, or hyperworld. Sometimes referred to as a hyperspace.

94. Hyperspace: generically, refers to hyperframes, orthoworlds, etc. May be used in a singular sense to refer to one of these.

95. Hyperspace howitzer: a postulated Soviet scalar electromagnetics interferometer weapon, believed capable of firing an explosion through hyperspace to a distant point, without any travel of electromagnetic energy through ordinary space. Also called a scalar howitzer or Tesla howitzer. The howitzer has several modes of operation: (1) it can be used in exothermic or endothermic mode. In the exothermic mode, the howitzer is biased positively. In this mode it can accept energy, transform it into scalar waves of an artificial potential or anenergy, transmit the scalar waves to a distant intersection zone, interfere them and reproduce ordinary positive electromagnetic energy in the distant region. In the endothermic mode, the howitzer is biased negatively. In this mode it "transmits" the extraction of energy. At the distant interference zone, energy is extracted and reappears back at the transmitter/receiver site, where it must be disposed of. U.S. weather satellite photos of the Soviet Union's Bennett Island reveal giant exhausts emerging from that island, some narrow jets as long as 150 miles and only about 1.5 degrees above the horizontal. Bennett Island is believed to be the "exhaust site" for several Soviet hyperspace howitzers. That is, when a howitzer operates in the endothermic mode, it receives and temporarily holds the energy that comes back to it, then switches biasing potential to positive and transmits in the exothermic mode to Bennett Island. (2) The howitzer can operate in pulsed or continuous mode. In pulsed mode, an explosive or implosive effect occurs at the distant target site. In continuous mode, a continuous energy source or sink appears in th distant interference zone. (3) Fourier expansion mode. Each transmitter of the howitzer can produce multiple frequencies where the frequencies represent a Fourier expansion of a geometrical form such as a sphere. When these two geometrical scalar forms are interfered at a distance, a 3-dimensional form of ordinary EM energy emerges (exothermic mode) or is extracted (endothermic mode). (4) the scalar transmissions of the scalar interferometer (hyperspace howitzer) can be modulated upon ordinary EM carriers, such as the Woodpecker signals in the communications band (3 to 30 megaHertz). In this case, interference patterns of the Woodpecker carrier will yield only minute EM energy, but the impressed scalar substructures may interfere to produce or extract substantial energy in or front the interference zone. Other control signals impressed on the Woodpeckers may be used to trigger the mode and location in the interference zone, or even to establish "circulations" in the vector zeros. By forming distant high pressure zones (exothermic mode) and low pressure zones (endothermic mode) and then moving the locations along, the Soviets can capture and deviate the jet streams over North America or elsewhere. By swirling in conjunction, tornadoes can be produced. Certain signatures (giant radial clouds and twin giant radial clouds) in the active zone have been identified as associated with the Soviet Union's weather engineering. The strange explosion in 1958 near the Urals is believed to have been an accident involving an early first fullup test of the howitzer. The howitzer was probably similar to a giant radar, transmitting scalar EM wave. By setting up a standing scalar wave and continuing to transmit energy into it, the standing wave becomes a giant potential. It is charged up much like a scalar capacitor, but with scalar EM (electrogravitational) energy. Failure of the transmitter may have discharged this giant potential into the earth, and the electrogravitational ground pulse spread out, striking the nearby stored nuclear wastes and fissioning all the radioactive nuclei. Consequently the nearby atomic wastes exploded, producing fallout and radioactive contamination over a large area. The first fullup test of an operational on-site weapon is believed to have been the destruction of the U.S.S. Thresher atomic submarine on April 10, 1963. The second fullup test is believed to have been the transmission of an explosion into the ocean 100 miles north of Puerto Rico on April 11, 1963—one day after the destruction of the Thresher. Over a hundred howitzers are believed to now be on site and operational, and powered by either Moray amplifiers or giant scalar resonant inference taps in the earth (detected by Golden in April 1985). High Altitude booms off the east coast of the U.S. may have been due to orientation and alignment tests of these weapons, with each test blast the equivalent of some 100 tons of ordinary explosives. If so, these shots were probably artillery high burst registrations. The two puzzling flashes detected by the Vela satellites in September 1979 and later, and erroneously described as a nuclear flash by many analysts, may also have been a test of this type of weapon, either by the Soviets or a second nation that has developed the weapons.

96. Hyperworld: a world (3-space or 4-space) rotated one or more orthogonal turns away from the laboratory frame (ordinary world). Constitutes also (identically) an entire subquantal level in virtual state.

97. Inception: kindling (from mind anenergy) a small amount of photon energy into or onto a body/tuner or other object Specifically, coupling scalar wave patterns (anenergy) onto die body/tuner or object so that the photon energy is created in or on the body at a distance.

98. Inceptive cyborg effect: the ability of the mind to psychokinetically link to, or attune to, an inert physical object or physical system. A living biological system is a stabilized inceptive cyborg, with its mindworld linked to its body/tuner. The two cerebral hemispheres of the brain form a Tesla scalar interferometer, capable of producing and transmitting scalar wave patterns into a distant object, coupling into energy in it, and thus linking with or resonating with the object. This is the mechanism for producing all parapsychological phenomena.

99. Interdimensional nodal point: a natural frequency for an anenergy channel between orthogonal frames or universes. Also called a magic window.

100. Interference: the variation of the amplitude of a wave with distance and time, where that wave results from the superposition of two waves having the same, or nearly the same, frequency.

101. Interferometer: an instrument in which a wave is split into two waves or beams, which after traveling over different paths are subsequently reunited and display interference.

102. Jitter (dither): the incessant quantum and subquantum activity on a fundamental particle of mass. The particle is continually connected to and disconnected from an infinite number of hyperspaces, and to and from time.

103. Kaznacheyev effect: refers to the demonstrated ability of near-ultraviolet mitogenetic photons to carry any kind of death pattern from one cell culture to another of similar kind, in the absence of visible light and with synchronized conditions. Named after the Soviet scientist Vlail P. Kaznacheyev, who experimentally demonstrated the effect In his experiments, two samples of the same cell culture were environmentally separated by a thin window. The entire arrangement was placed in the dark and rotated at some 20 rotations per minute. The cells on one side were then killed by any particular mechanism—ionizing radiation, chemicals, toxins, deadly bacteria, germs, etc. If the window was made of ordinary glass (opaque to ultraviolet), the second sample remained healthy. If the window was made of thin quartz (transparent to ultraviolet), then after a time delay the second sample began dying with the same symptoms ai the first Later testing revealed the emission of mitogenetic, laser-like photons in the near ultraviolet from the dying cells in the first culture. When these mito genetic photons are allowed to bombard the second cultare, the same lethality pattern is induced in it. The explanation is that the photons carry virtual state (scalar) patterns (infolded patterns in their time components), and these virtual patterns are diffused into the receiving culture, where eventually they diffuse into a lethal pattern in the virtual state (scalar) master control system (discovered by F. A. Popp) of the targeted cells. Continued charge—up of this pattern kindles the pattern into a physical, observable structure in the cells, reproducing the death or disease pattern. The Kaznacheyev effect haa been duplicated in the infrared in West Germany. Kaznacheyev almost certainly also phase conjugated (time-reversed) the "death photons," producing healing photons for that specific cellular damage or disease condition. Irradiation of the affected cells with these healing photons would have reversed the disease, reverting the cells to normal. Note that Antoine Priore's machine worked in this fashion. Kaznacheyev's EM healing work with phase conjugated photons was obviously not reported; it represents a total counter to viral biological warfare, and such work would have been done only under KGB control, and highly classified as a vastly valuable defense secret.

104. Kelly, Peter: noted radionics/psychotronict researcher and builder of several types of especially potent amplified radionics machines. He is also one of the pioneers of the use of amplified radionics in agriculture for enhancement of crops, restoration of soil vitality, and control of insects. By direct measurement with Bedini scalar EM detectors, Kelly's machines have been verified to produce scalar waves.

105. Kelly machines: amplified radionics machines developed and built by Peter Kelly.

106. Kervran effect: the ability of a living biological system (either plant or animal) to accomplish limited transmutation of elements, particularly between certain isomers. Named after the French scientist, Louis Kervran, who discovered the effect. Kervran was nominated for a Nobel prize in 1977, although the prize was not awarded to him. The Kervran effect has been duplicated by Japanese, British, European, American and other scientists. Note that scalar interferometry is capable of kindling energy patterns directly into the virtual state interactions of the nucleus, including the strong force that binds the nucleons together. Over a period of time, this energy can cumulate, biasing the quantal statistics of the interactions. This, of course, changes the probabilities of the various quantal states, with the result that the nucleus can be transmuted. Transmutation to an isomer is particularly practicable. Efficient transmutation of elements by scalar means has been accomplished by John Bedini.

107. Kindling: collecting and condensing more subtle energies (anenergy) into virtual photons (charge effects), photons (electromagnetic field effects), or matter (materialization effects). Kindling is a linear integration process, which repeatedly collects or integrates a virtual state pattern until it breaches th observable threshold, becoming physical real.

108. Kirlian photography: field patterns of an object impressed upon a film or other sensitive material by means of spark discharges. The scalar modulations in the spark charge intensity, not the ordinary vector E-field components, are the paranormal components. Thus only the delta—which represents those scalar patterns coupling and kindling into vector EM, which then interacts in photochemical interactions—is indicative of the paranormal aspect of Kirlian photography.

109. Law of physics: a restriction placed upon the physical mechanisms that can occur under certain broad assumptions as to the background situation and environment. Actually, in nature there are no such "laws"; these are only man's synthesis for the behavior of large classes of systems under very general conditions. Any "law of physics" can be violated if the conditions contained in its assumptions are violated. Conservation laws are often touted as "laws of physics." Actually, the conservation laws assume a closed system and a linear spacetime (Lorentz frame). If either or both of these assumptions is violated, any or all of the conservation laws can be violated. In particle physics, broken symmetry already results in violation of one or more of the conservation laws. In leading Soviet physics journals, papers are regularly published containing unrestricted general relativity with concomitant violation of any and all conservation laws.

110. LIDA machine: Soviet device which uses a 40 megahertz carrier and unusual waveforms to induce a catatonic-like state in man or other mammals. For years Western orthodox scientists have argued that EM signals could not directly affect consciousness. With testing of the LIDA device, this has now been totally refuted. One scientist has reported that the I.IDA device—known to be decades old—was used by the North Koreans on some U.S. prisoners of war during the Korean War. Reportedly the effects of the machine could not be resisted, and the machine was a devastatingly effective "brainwaahlng" tool.

111. Life: a term which resists the best efforts of philosophers and scientists to define it! In the new approach, a living system is modeled as consisting of a physical body "tuner" and a mindworld to which the tuner is preferentially tuned to a small degree. The preferential virtual state crosstalk between the physical tuner and the mindworld constitutes the life channel or consciousness loop. Destruction or loss of the preferential tuning results in "death," or loss of attachment to the tuner. In that case the mindworld functions within itself. Note that the being is always present, unchanging, and simply "is," without any connotation of perceptual stringing together separations into a string of events (causal chains). Being changeless, it consists of all changes, all-at-once, hence none separately. Ultimately this is the true meaning and foundation of life, and mindworlds and physical events represent simply separation waves in unlimited being. Note this is completely consistent with the idea that ulti mately only God exists, things are simply separation waves in the unseparated Godness, and God permeates everything.

112. Loch Ness Monster: famous tulpoidal water monster which (occasionally) temporarily materializes in Loch Ness in Scotland. Similar materializations of "water monsters" have been seen and reported in almost every large body of water on earth.

113. Longitudinal wave: a "pressure" type of wave, similar to sound, in which the vibrations are along the direction of travel of the wave. Hence, a wave composed of alternating densifications and rarefactions. Scalar waves are longitudinal waves. In classical EM theory, the Poynting vector predicts no longitudinal wave of energy from a time—varying, electrically charged source. In fact, an exact solution of the problem does allow this longitudinal wave, as shown by Keech and Corun, International Journal of Theoretical Physics, Vol. 20, No. 1, 1981, pp. 63-68. The classical electromagnetic wave is modeled as a vector wave, comprised of transverse vibrations. In the new approach, such transverse vector force field waves (Hertz waves) do not exist in vacuum; instead, two coupled longitudinal scalar waves exist in vacuum. The E and B fields are formed, recovered by, and exist as the spinning charged detecting mass particle in motion.

114. Magic window: a naturally tuned frequency of a good hyperchannel between orthogonal frames where scalar wave anenergies crosstalk readily. Magic windows are frequency dependent, and are also referred to as hyperchannels or interdimensional nodal points. Some magic windows are 38-40 kHz, 150-160 kHz, 1.1-1.3 MHz, 1.057 GHz (the Lamb shift), and in the near ultraviolet (the life energy frequency).

115. Many-worlds interpretation of quantum mechanics: refers to the consideration of multiple simultaneous observers and the consequent changes forced upon quantum physics. Essentially all possibilities exist and are real, and an infinite number of orthogonal worlds exist at any instant. Further, each world splits into multiple possibilities after each and every event. Thus in a single second, the world has split in some 10 to the 20th power versions of itself. This is the work primarily of Everett, with strong contributions by Wheeler and

Graham, and was the subject of Everett's thesis at Princeton University in 1956. It was published in 1957.

116. Mass: literally, a special curvature or "kink" in spacetime emptiness, according to relativity theory. In the new approach, the mass and inertia of a particle are considered to be due to one or more standing scalar resonances, trapped by the spin of the particle. When the amplitude of the trapped scalar wave increases or decreases, the mass and inertial of the particle increase or decrease. If the resonance amplitude reduces to zero, the particle dematerializes. Classically, mass is also taken as the measure of a body's resistance to being accelerated by a force. Specifically, a mass occupies space (3-dimensional) and resists change in its linear movement. In a region of mass, a flux of action quanta into and out of the region exists, corresponding to a flow of positive and negative time. The mass is the sum of the absolute values of the time derivative of action in and the time derivative of action out If these two time derivatives are equal in magnitude, the mass is unaccelerated; if unequal, the mass is accelerated. A mass does not exist in time per se (more rigorously, it exists equally in positive time and in negative time), but only exists spatially. A mass continually interacts with photons (absorption and emission), thus continually connecting and disconnecting to the fourth dimension (time). The mass is thus seen to "jump along" through time in finite, quantized jumps caused by the photon interaction with it

117. Men-in-black: that part of the UFO contact phenomena which involves contact with rather sinister tulpoidal manforms. Typically these tulpoids are dressed in dark suits, have almond-shaped eyes that do not blink, possess oriental features, and "speak" telepathically to the contactee. Often they drive dark automobiles and follow the contactee. Several researchers in the field have had intense encounters with such tulpoids. The appearance of such phenomena associated with an investigator is in part due to unresolved detrimental material in his or her own unconscious mind, which thus acts as a "witness" and radionics tuner. Continual stimulation of this detrimental unconscious witness by the person's investigative activities results in the appearance of, and contact with, such tulpoids. The more detrimental aspects of such tulpoidal tuning can be prevented by psychic clearance of the investigator, particularly of his unconscious mind. Such psychic clearance at regular intervals is highly recommended for investigators pursuing the more bizarre aspects of the field, such as cattle mutilation, men-in-black, sasquatch, possessions, etc.

118. Metal bending: term applied to the psychokinetic bending and breaking of metal objects such as cutlery, rods, bars, etc. When a metal rod is psychokinetically bent, then cross-sectioned and examined under the electron microscope, microscopic "puddling", as if by localized, point source intense heating, can be seen, with concomitant disruption of the metallic grain structure. In a severe case, the cross section looks like the puddled surface of the moon. The mechanism responsible for the psychokinetic bending is exothermic scalar interferometry by the two cerebral hemispheres, where the intersection zone is the metal object being bent

119. Metapsychology: literally, "before psychology." The deeper psychology of the group mind, particularly the collective human unconscious mind. Although the Cartesian idea of the separation of mind and matter usually applies to the individual conscious mind, it does not apply to the group unconscious mind. The collective human unconscious mind is a living being, though unconscious, and is capable of materializing its "dreams" into actual physical forms and beings.

120. Microcrystals: tiny crystals, particularly those used by George Yao in construction of his pulsors. Use of the microcrystal in proprietary manners enables Yao to build devices which react to scalar resonance and human intent, and which can generate negation patterns for potentially harmful "deltas" or changes that occur in the human etheric field (aura).

121. Miller/Fox/Urey effects (experiments): these experiments simulated the primitive conditions believed to have existed in and over the primeval oceans. Water—with its complement of minerals and chemicals—and gases to simulate the primeval atmosphere were placed in a beaker and subjected to ultraviolet radiation and electrical discharges. Thus the experiments simulated the great lightning discharges into the primeval oceans. In the laboratory, amino acids are formed, then proteinoids. If the experiments are performed in the presence of heating (simulating the presence of volcanic activity, for example), then primitive microspheric cells are formed with double cellular walls. These microspheric cells even reproduce by cell division. Thus at least simple lifeforms have been produced in the laboratory. Indeed, in the new model even inert matter is totally alive, but unpreferentially. Also, every possible sort of mind exists. Thus all that is necessary to form a living system with any desired characteristics is to form the appropriate type of physical tuner. A living system is like a television set existing in a flux of radiations from an infinite set of channels/stations. The set tunes preferentially to one channel/station, but any desired channel can be tuned, simply by changing the characteristics of the tuner. One does not "create" life, one simply "tunes it in selectively." In this concept rigorously there is no such thing as "death", but only transition from one state to another.

122. Mind: a quite misunderstood concept! In conventional science, it is considered purely robotic—i.e., it is the sum total of the electromechanical neural processes which receive, code, and interpret sensations, recall and correlate stored information, and act upon it Also considered to be the understanding, reasoning, and intellectual faculties considered as a whole, that is, as the conscious portion. Also considered as the psyche, including the conscious, subconscious, and unconscious portions, considered as a whole. The new approach is more in accord with the latter definition. Simply put, the mind is a world composed of separation events (waves, operations, processes) in unseparated being. We may model it (to a first crude approximation) as a physical universe, three orthogonal turns away from the ordinary physical universe, and tuned slightly selectively to one physical organism's physical body processes. Mind anenergy is considered to be progressively collected, condensed, and kindled into denser substance and objects by rotation toward the ordinary physical world. Thus a piece of inert matter is simply condensed energy, which itself is condensed anenergy, which is condensed mind flux (crosstalk) from all minds. One may model the mind physically, or model the physical as mindstuff, eliminating the artificial dichotomy between mind and matter that presently is assumed in orthodox science. Since per-

ception of thought *is* fragmentation and separation of mind, die mind cannot be perceptually conscious of itself, but only of its separated fragments (thoughts). When the incessant activity of the mind is stilled, however, the stable mind is in its pure state—pure being. In this state there is no separation of the observer and the observed. With attainment of this state, one "realizes" or experiences one's true God-nature. It is theoretically possible to build a scalar wave detector that can directly "see" thoughts. Indeed, the first prototype of this type detector may have already been built by one physicist experimenting with a very sensitive scalar wave detector.

123. Mind linkage: the integrating of two or more minds into a single overall functional mind. A living example is in each human head: the two cerebral hemispheres—each containing its own separate mind—are integrated into one overall brain having a single functional mind. This is accomplished by signals interconnecting the two cerebral hemispheres through the corpus callosum. Note that the "separated beings" of the two minds are also merged/linked into one greater being. The tuned separation of individual humans (caused by the continual photon interactions with each particle of mass in the body tuner) is responsible for all man's inhumanity to his fellow man, for each now perceives others as separate things or "objects." With mind linkage, the human species will emerge as a single great entity, having not only • single "collective unconscious," but also a single "collective conscious" mind. Proof that linkage can be accomplished is already demonstrated by each individual. The problem of resolving all man's inhumanity to his fellow man is thus a technical problem, not just a moral/philosophical problem.

124. Mitogenetic radiation: laser-like, coherent photons emitted by cells, particularly in the near ultraviolet Of particular importance are the photon emitted by a cell immediately after complete fertilization, and the photon emitted as the cell dies. The first mitogenetic photon signifies the kindled connection of the cell's small mindworld and the establishing of the consciousness or life loop. The second mitogenetic photon signifies the loss of tuning and consequent loss of the consciousness or life loop. The photon at cellular fertilization is a joyous song of birth, and the photon at cellular death is a deathsong of the cell. In addition, the birth photon carries the full mindworld pattern nested inside it in the virtual (scalar) state, and the death photon carries the full pattern of the cell's experiences and death nested inside it in the virtual (scalar) state.

125. Monocular (change): one at a time. Change as seen by only a single observer at a time. This type change is provided by the photon interaction with matter, and the "memory hold" of a large number of such monocular changes constitutes the "perceived/detected physical reality" we conceive we exist in.

126. Moray amplifier: free energy device developed by T. Henry Moray of Salt Lake City, Utah in the 1920's and 30's. The device was destroyed by a double agent (an F.B.I, agent and also a Soviet agent) in 1939, after the agent had obtained the drawings of the device and worked in Moray's laboratory for several months. The final device had some 29 detector stages and produced an output of 50 kilowatts of power, without input.

127. Moray, T. Henry: early experimenter and pioneer in radiant energy, who developed one of the most widely demonstrated and best authenticated free energy devices. Moray's device was a collector/amplifier that eventually produced 50 kilowatts of output power.

128. Multiocular (change): more than one at a time. Change as seen by multiple simultaneous observers. Consideration of this type change leads to an infinite dimensional, multiworld or multiuniverse reality similar to the type described by Everett. If one accepts multiple simultaneous observers, then this expanded and startling view of reality is inescapable. If one accepts only a single observer at a time (as does present physics), then there exists only a single living person, and everyone and everything else is simply a figment of his or her imagination.

129. Mystery light: another name for a nocturnal light, or earthstress light, produced by natural scalar transmission interferometry of the cracked rock structures in the earth's crust

130. Neurophone: an instrument for direct electomagnetic input of signals into the human brain and nervous system. The neurophone was invented by Pat Flanagan when a boy of seventeen. The improved version, invented much later by Flanagan, clips an input signal, differentiates it twice, then feeds the resulting noisy spikes into the human sensory system. The original signal is impressed into the brain and directly into the mind. An adaptation of the neurophone, using more than eleven independent frequencies and built in accord with Soviet work reported by Lisitsyn, could conceivably be utilized to accomplish human mind linkage.

131. Neutrinic wave: a bare neutrino, considered as a wave instead of a particle. It is a special kind of polarization wave to the laboratory observer, and interacts with his electromagnetic field. It is essentially a longitudinal scalar wave, which moves in and on ordinary vector electromagnetic waves. It may also be regarded as a sort of "spin" wave, since one may consider it as moving at the speed c, but circling around a photon in a plane perpendicular to the photon's velocity in 3-space. One may thus say that the neutrinic wave moves at the velocity c-squared, but its carrier moves at the velocity c. Since we detect the speed of the carrier, detected neutrinos always appear to have moved at the speed of light.

132. Neutrino: a neutral particle believed to have zero rest mass and travel at the speed of light. One type of neutrino is associated with the electron and one type with the muon. A third type is now believed to exist There is some evidence recently of neutrinos possibly having a small mass, and of one type of neutrino turning into another type. In the new approach, the classical neutrino is a flux of much smaller particles—"bare" neutrinos—inside a photon. In the photon, one flux is taken to rotate clockwise and one flux to rotate counterclockwise (adapted from the phasor theory of Ruderfer). Further, a neutrinic flux from one half of the photon to the other is hypothetically possible, and this flux would constitute a third type of neutrino. It would exhibit mass, and would allow one type of neutrino to turn into the other. The bare neutrino (from the new approach) is of magnitude 1/c times the magnitude of the photon. Further, it moves at the speed c-squared, but since it circulates at the speed c at right angles to the motion of its photon carrier which is moving at the speed c, its detected velocity is always c, the speed of carrier. To an observer in the hyper-

frame of the electromagnetic field, a laboratory bare neutrino would appear as a photon.

133. Newman, Joseph: Mississippi inventor of a free energy machine tested and authenticated by several dozen scientists. Newman is fighting in court to obtain a patent on the device which has clearly been shown to be legitimate.

134. Newman machine: free energy device developed by Joseph Newman. While the prototype does not produce appreciable output power, it does produce an output and keeps its own small drycell batteries charged. It clearly demonstrates the principle and is a legitimate free energy device.

135. Nocturnal light: any paranormal light phenomenon seen at night. Over 80% of the UFO phenomena are nocturnal lights. Also includes phenomena such as "ghost" lights, spooklights, will 'o the wisps, etc. It specifically does not include ball lightning, St. Elmo's fire, and other such rare but normal phenomena. Most of the nocturnal light phenomena are caused by natural scalar transmission interferometry.

136. Observable (state): that part of physical reality consisting of changes (particles, waves, action changes) that are detectable by single instruments. The field on a charged particle is due to the continual emission and absorption of virtual particles. Forces between observable particles are generated by the exchange of virtual particles. Thus all observable forces, fields and changes are generated by unobservable virtual interactions, and observable and virtual reality states are inseparably enmeshed. In theory, no superluminal effect can be directly detected; however, a two-stage detection can be used to detect a superluminal entity. Usually the first stage is an interference between two superluminal entities, where the interference "downshifts" the waves to produce a wave moving at subluminal velocity. This subluminal wave then interacts with a mass such as an electron, and these interactions are accumulated to provide a macroscopic measurement indicative of the presence of the primary superluminal interferents. A two-slit box for the electron inteiference experiment is an example. Also, in quantum mechanics time is not an observable, even in theory. All observation is spatial; rigorously, there is no such thing as a clock in the purest sense. Time is always inferred from a spatial detection or measurement

137. Orgone energy: the subtle living or life energy discovered and utilized by Wilhelm Reich. Reich had discovered and utilized a portion of scalar electromagnetics (electrogravitation).

138. Orthoframe: a frame rotated away from the laboratory frame by one or more right angles. A hyperframe or biofame. Also called an orthoworld.

139. Orthogonal: simply, at right angles. The intersection of two orthogonal geometric constructs, each of dimension one or greater, is of one less dimension than the lowest dimensioned intersector.

140. Orthorotation: rotation in hyperspace of an object, form, or particle at a right angle to its starting frame. The object loses one dimension in its intersection in the original frame and gains one dimension in its intersection in the frame toward which it rotates.

141. Orthoworlds: orthogonal 3-dimensional physical worlds (frames) that share a single time dimension in their 4—spaces (spacetimes).

142. Paradox (logical): a thing that is true, but which violates one or more of the three Aristotelian laws of logic. All logical paradoxes are simply fourth-law logic cases, and—at least hypothetically—the fourth law can be applied to resolve them.

143. Paranormal phenomena: any phenomena that differ from straightforward, orthodox physical or psychological phenomena. Phenomena beyond understanding in terms of presently known science, causes, and effects. Specifically, phenomena in which mind and matter, or mind and unknown fields, or matter and unknown fields interact observably.

144. Past: reality which has been erased and has ceased to "physically (observably exist" in immediate detection/perception, and which now exists only as an infolded virtual state trace or vacuum memory. (Because the past still exists as an infolded virtual trace in vacuum, the time-reversed wave emitted from a pumped phase conjugate in four-wave mixing can retrace the past path of the wave that struck the mirror and stimulated the emission.) Fissioning of an action quantum produces a "presented" change described by canonical fragments. Fusion of the canonical fragments back into a quantum of action erases the presented real canonical fragments, reducing the "change" to the "past" However, that "fissioned pattern" or virtual substructure continues to exist, infolded into the vacuum. The vacuum retains a perfect infolded recorded memory, therefore, of every quantum change that has ever occurred. (In ancient mystery schools, the set of all such traces was called the Akashic record.) A virtual projection of the change has occurred in the mind world of the observer, however, since the mindworld is nonlocal. The coherent (causal) time-ordered collection of large numbers of such virtual projections in the mindworld reaches the quantum threshold, resulting in a quantum change in the mindworld. The "conscious mind" thus perceives physical reality as a collection of a large number of quantum changes—in short, a macroscopic reality. In actuality this "perceived macroscopic world" rigorously exists only in the mindworld, and thus it exists only in the past (Note how physicist's discussions of such quantum paradoxes as Schrodinger's Cat are erroneously based on macroscopic considerations.)

145. Phase conjugation: a nonlinear wave process whereby a nonlinear crystal, material, or medium creates a nonlinear polarization for an incoming monochromatic wave is generated at the incident frequency by w-w-w= -w. The nonlinear polarization radiates a time-reversed replica of the incoming waveform. The time-reversed waveform has an amplitude proportional to the complex conjugate of the incoming wave amplitude. Though primarily known in nonlinear optics, the time-reversed wave is a solution to the wave equation, and so the phenomenon is universal to all kinds of waves. As can be seen from (w-w)-w = -w, the first two waves modulate each other because the nonlinear medium acts as a modulator. These two waves thus "lock together" as a single scalar EM wave. The scalar wave passes through the electron shells of the atoms of the nonlinear medium, penetrating to the positively charged atomic nuclei. In a nucleus, each incident scalar photon is absorbed, increasing the potential of the nucleus to an excited state. The decay of this excited state from the positively-charged nucleus causes the emission of a time-reversed (TR) photon carrying negative energy. The "emitted shower"

of these TR photons constitutes a TR wave. Thus from the nonlinear material there is emitted a time-reversed wave. This TR wave "backtracks" through space along the invisible path taken by the original incoming wave. Since the phase conjugate wave is time-reversed, it concentrates along its backward path, rather than diverging. Consequently all its energy arrives at the distant location of the source of the original stimulus wave. The nonlinear medium producing the TR effect in response to a stimulus wave is referred to as a phase conjugate mirror. If two monochromatic EM waves w1 and w2 are fed into the nonlinear medium in antiparallel manner, so that the waves are everywhere out of phase, they modulate each other, forming a much more powerful scalar EM wave. This wave then "pumps" each nuclear potential to a much higher excited state. Thus when this potential decays, a much more powerful TR wave is emitted, containing up to all the energy put into the scalar "pump" wave. Note that the excited nuclei do not decay (the excited potential does not decay) until an incoming stimulus wave—say w4—is received. When w4 is received, then a powerful TR wave—say w3—is emitted from the phase conjugate mirror (PCM). This process is referred to as "four wave mixing." Thus a pumped PCM acts as a triodc amplifier, and the incoming stimulus wave acts as a signal on the grid. In the oscillation condition (incoming input w4 at right angles to wl and w2), infinite gain is experienced, so that all the energy in pump waves wl and w2 appear in TR wave w3. This was Nikola Tesla's secret of wireless transmission of energy at a distance without losses; the Great Electrician had discovered what we today call phase conjugation and the pumped phase conjugate mirror. He also discovered its general nature, for he understood that it applied to all kinds of waves, including mechanical waves in the earth. It has been shown that, under proper conditions, it is possible to have self-oscillation in the nonlinear crystal or medium. Such a PCM is referred to as a "self-pumped PCM." Indeed, Tesla's original Magnifying Transmitter would have utilized the earth itself as a self-pumped, amplifying PCM—in other words, as a giant triode where the molten core of the earth already furnished the cathode power for free. A relatively small input "grid signal" (from Tesla's resonant transmitter) would thus have set the earth itself into scalar resonance (giant self-pumping), eliciting an enormous time-reversed wave energy resonant response from the core of the earth. Anyone anywhere on earth with an antenna tuned to the stimulated frequency would have received cheap or "free" energy from the earth itself. The Soviets presently use a version of this method to power their enormous scalar EM wave Tesla weapons. All unorthodox resonant free energy devices (built in secret by unorthodox inventors) presently use variations of this same principle.

146. Photon: the basic action quantum. It may be considered as an oscillation in and of time. It may also be considered as a virtual pattern of positron/electron pairs. A piece of electromagnetic radiation when it interacts as a particle. One half the photon exists in and carries positive time (negative charge), and the other half exists in and carries negative time (positive charge); thus one half is "normal" and the other half is "time-reversed" (phase conjugated). The magnetic field of a permanent magnet, for example, thus consists of two types of coincident magnetic field lines: one line of each coincident pair is in positive time and by convention runs from the north pole to the south pole, while the second line of the pair it in negative time and runs from the sooth pole to the north pole. The two lines maybe considered as two streams of particles, one negative and the other positive (Howard Johnson's two-particle theory of magnetism). The photon may be considered as two contrarotating neutrinos, after the work of Ruderfcr. In the new approach, each of these rotating neutrinos is considered to be composed of vast numbers of circulating bare neutrinos. The photon may be considered as one cycle of an electromagnetic wave. The photon is the basic carrier of time and consists of a piece of energy welded to a piece of time, with no seam in the middle. The passage of "time" thus moves at the speed of light, its carrier. All ordinary photons possess the same action magnitude, h, although they do not carry the same energy. If the photon is made subquantal (i.e., less than h in magnitude), it is a virtual photon of the type whose emission and absorption cause charge on an electron. Virtual photons are not individually detectable, even in theory. However, a virtual photon may have positive mass under some circumstances.

147. Photon interaction: the absorption and emission of photons by particles or objects. Usually the photon interacts with an orbital electron, being absorbed by the electron which is raised into an excited state. Then the election decays from the excited state, emitting another photon. The macroscopic world is created by this interaction, which is the basic quantum change interaction. When we introduce additional photon interactions beyond the ambient background, the nucleus stays basically stable, and the normal first-order world we perceive seems quite stable. This is the first level of physical reality, and it is the reality sensed by our coarse physical senses. However, scalar waves are emitted and absorbed by the nuclei of atoms, passing right through the electron shells without interaction. When we introduce additional scalar interactions beyond the ambient background, the nuclei change appreciably, and this level of physical reality may be far from stable. This is the higher reality, and it is sensed by the scalar electromagnetic functioning of our nervous system. Unfortunately, this system outputs only to the deep unconscious, since it is highly multiocular. Thus our conscious mind, being monocular, does not perceive the most fundamental reality in which we exist. Note that "time" is the special dimension in which multiple objects can exist simultaneously in the same interval, while space is composed of dimensions (length) in which multiple objects cannot exist simultaneously in the same interval. Time is multiocular, and space is monocular. Our conscious mind is fitted to the monocular photon interaction; being monocular, our conscious mind cannot be aware of time directly. For that reason we do not "see" time consciously. We do "see" it, however, unconsciously. The true meaning of being "lost" is to be separated from the consciousness of the All, which is separation of our conscious, gross sensing of first order reality from our finer, more subtle, and infinitely richer—and unconscious—scalar sensing of ultimate reality.

148. Photon quenching: the effect of photon emission carrying away time, constituting a time-differentiation of spacetime and leaving behind a spatial reality. This action is responsible for the destruction of the time dimen-

sion in the detection/measurement processes of our physical instruments and our body sensory systems. Thus all detections and measurements are spatial, and we detect and perceive only a 3-dimensional spatial universe, not a 4-dimensional or higher spacetime. Photon quenching is thus a filter between us (our senses and instruments) and fundamental reality. Physical reality as detected by our senses and instruments is real, but only partial. Specifically it is a time-differentiation of fundamental reality. We thus lose the ability to detect those things which occupy time but not normal 3-dimensional space—such as mind, thought, etc. Almost all our thoughts, concepts, words, and ideas are fitted to this partial reality—and this is the universal human problem and delusion. We see a spatial universe of separated spatial objects, while in actuality we exist in an undifferentiated single wholeness. It is not a matter of loving one's brother or sister, it is a matter of being him or her. Fortunately, with scalar electromagnetics much finer instruments can be built The quantal limitation of ordinary physics is not a limitation in scalar electromagnetics. Ultimately we will be able to use sophisticated scalar sensors and biofeedback training to awaken and develop consciousness of our scalar sensing of higher reality, and restore man to his proper state.

149. Piezoelectric effect: In a material, the appearance of a voltage when mechanical stress is applied, or of a mechanical stress when a voltage is applied.

150. Poltergeist: allegedly, a spirit or ghost, often mischievous, usually associated with anomalous psychokinetic activity in houses such as slamming doors, levitating and flying objects, displacement of persons or objects, etc. Often the poltergeist phenomena are associated with an emotionally upset or distraught adolescent. The experience of cooling of the surroundings when these phenomena are evidenced is also common.

151. Potential: locked-in work or energy at a point It is important to realize that, like a runner running in place, the potential represents work continually being done in time, but at a point, not released through a distance. Since the work is never spatially released, to the external observer it remains in the present as energy, and does not pass into the past as work accomplished! Prior to observation (passage from the present into the past), the physical universe exists as action or angular momentum—e.g., as "energy x time," or "momentum x length," etc. The two variables in which angular momentum is expressed are referred to as "canonical variables." They are locked together, since their product is a constant (in a nonlinear frame). Thus if we ask "what exists in time?" we have asked "what is the time derivative of action, which is the same as "what is existing in time but not in space?" The answer then is energy, which—before observation (separation) exists undifferentiated as (energy x time). Ironically, in electromagnetics the electric (electrostatic) potential is completely misdefined. Most texts start about to define it in terms of "potential difference," and then only tell how to prescribe the intensity or magnitude. For example: "The work which must be done against electric forces to bring a unit charge from a reference point (often at infinity) to the point in question." Note this "definition" does not at all tell what a potential is, but instead tells how to mathematically determine its intensity. Briefly, electric potential is a locked-in stress in vacuum spacetime. This stress at a point is due to the intensity of the virtual particle (virtual photon) flux (dynamic, ongoing, trapped virtual work) at that point The ambient local stress of vacuum is not zero, but is a very high value. Normally we refer to "potentials" as measured from the ambient; that is, we actually specify the "difference from the ambient value of the spatiotemporal stress due to virtual photon flux" at the point in question. Any change in the ambient stress of vacuum is a curvature of spacetime. The electrostatic curvature normally stress-relieves almost immediately as electric force field on matter, or electric virtual force field in vacuum. In Kaluza theory, the electromagnetic force fields are just the 5th dimensional aspect of the 5-d gravitational field. The 4-d gravity field is the other aspect of our 4-<i world's intersection with the 5-d G-field. Thus the electrostatic and magnetostatic potentials are simply two coupled aspects of the 5-d G-field. A normal electromagnetic force field (E or B) is the "bleed-off" or escape on charged particles, of the 5-d gravitational field potential's infolded stress contents into the fifth dimension. That is, it is the 5th dimensional outfolding of the inner contents of the gravitational potential. If we wish to achieve bleed-off into die 4-d G-field instead of the normal 5th dimensional E and B fields, then we must "shut off" the EM bleed-off/out This shutoff can easily be done by opposing and reciprocally modulating E-fields and/or B-fields so that they lock-together as one field, and vectorially sum to zero vector resultants. This creates artificial (internally structured) electrostatic and/or magnetostatic scalar potentials. These potentials are actually 5-d G-potential, which is now forced to bleed-off/out into the 4-d gravitational force field at great amplification. This is the outfolding of the inner contents of the 5-d G potential into ordinary 4-space. Such an artificial potential produces a vector zeroed 5th dimensional EM system with a deterministic substructure, which polarizes the vacuum and engineers the virtual state. Maxwell's original 1864 quaternion theory included this theory, though not in these specific terms. Heaviside's mutilation of the theory into vectors threw away the electrogravitational aspects by discarding the scalar component of the quaternion.

152. Prigogine crystal: an amorphous pellet or crystal made by sintering finely divided material at high temperature and pressure, in such fashion that the pellet becomes a highly stressed system far from thermodynamic equilibrium. Specifically, more than one type of material must be used, one ingredient of which must be piezoelectric. One ingredient should also be radioactive, and preferably one of the uranium compounds exhibiting highly anomalous magnetic spin coupling. For best results, a third ingredient should be luminescent when electrically stimulated. The stress on each grain of the piezoelectric material must be just so that the grain is on the very verge of being slightly stress cracked, but not split Each grain then becomes a scalar interferometer. Such a crystal produces a scalar potential field and can react to minute changes in potential—i.e., it can react to scalar waves. Via scalar interferometry it can change scalar waves into negative (time-reversed) electromagnetic radiation and energy at a slight distance. Under oscillating potential stress, the radioactive ingredient provides a one-way gate valve from the Dirac Sea electrons of vacuum, and the scalar interferometers provide necessary impetus on these negative energy electrons to lift

them out of the Dirac sea, producing negative electricity and currents. These currents can then be collected in multiple stages to provide electrical power (negative power operates devices much better than positive power). In the proper arrangement, such a Prigogine crystal can be made into a system capable of tapping the energy of vacuum directly. T. H. Moray built exactly such systems in the 1920's and 1930's, finally obtaining 50 kilowatts of negative power from a 55-lb device. J. Bedini has produced modern versions of these stress crystals in the 1980's, and a series of negative power devices. Several other inventors have produced successful negative power devices also.

153. Prigogine effect: the exhibition of negative entropy by a properly stressed many-particle system far from thermodynamic equilibrium.

154. Prigogine system: a many-particle system which is deliberately forced to exist far from thermodynamic equilibrium, and which exhibits negative entropy. Note that local curvature of vacuum spacetime places the local virtual state flux in nonequilibrium conditions, with the result that an observable energy "sink" or "source" can be produced.

155. Priore, Antoine: French inventor, who invented a large electromagnetic machine which produced nearly incredible cures of terminal cancers in hundreds and hundreds of laboratory animal tests in the 1960's and early 1970's. Priore's results were presented to the French Academy by the Secretary Perpetuel and Head of the Biology Section, M. Robert Courier. The French Government financed Priore's work for several years, then abandoned it when the regime changed. Priore produced complex phase conjugate (time reversed) EM signals in a giant rotating plasma, then sent these time-reversed signals down a rippling magnetic field into and through every cell of the treated body. By magnetic resonance of the rippling field, the phase conjugated signals were coupled into the atomic nuclei of the body, thus reversing the harmful structuring of the cellular biopotentials, which are centered in and on and from the atomic nuclei. This "backtracked" the circulating substructure components within the biopotential, reverting it back to a previous state where the disease was not present Kindling this pattern into observable state thus reversed and eliminated the disease. (The sharp reader will immediately note that it is readily possible to reverse the aging process, given the development of this technology. So-called "eternal physical life"—or certainly indefinite physical life with youth and vigor—is achievable.). With the death of Priore, the work has ceased and the machine has been dismantled, since no one in the French Academy knew of phase conjugation or understood how it worked. Priore's last machine was nearly four stories high and was intended to treat terminal human cases as well as larger animals such as monkeys.

156. Priore machine: a device built by Antoine Priore in France, which cured terminal cancers and leukemias in thousands of laboratory animals in the 1960's and 1979's. Complex nested modulations were mixed in a rotating plasma, which modulates the signals (producing nested stages of modulations upon modulations, whic*h* is necessary to reach into a sufficient number of hyperspaces/nested virtual levels of vacuum). The device produced very complex, phase conjugate (time-reversed) hyperspatial signals of some 17 or *more dimensions* These time-reversed signals were imposed upon and intense, rippling magnetic field carrier, and induced into the body of the irradiated animal, eventually causing the remission of the disease after several treatments. The signal was able to "work on" some 17 levels of the biopotential. Hundreds and hundreds of terminal cases were cured in laboratory animals. The nested modulations formed hyperspatial, multidimensional components in their "giant photon" vacuum engines. This constituted scalar electromagnetic and virtual state engineering in 17-d spacetime, and the machine directly affected the life force frequency channel (consciousness loop) and master virtual electromagnetic (scalar wave) control system of the cells. It would appear that Priore had worked out combinations of frequencies, etc, that, when expressed in time-reversed waves, reverse abnormal mammalian cells, but do not affect normal cells. With the change in regimes in France and the death of Priore, the large new machine (which would have been used to treat humans) was dismantled and the discovery passed into disarray.

157. Probability: the probability of a simple event is the ratio of the number of times it occurs to the total number of trials (for a large, essentially infinite number of trials). Note that "trials" constitute things that have occurred (are past). Probability has resisted rigorous logical definition by Aristotelian logic; the reason is that it is an expression of the fourth law, the identity of oppositcs. Probability is the expression of the future (that which has not occurred) in terms of the past (that which has occurred). Since we have been trained to conceive or think of something as if it had just been perceived to occur, then this is the only way we can conceive the future. But with Aristotelian logic, the future cannot be composed of the past, and probability has no acceptable logical basis. With four-law logic, the future can be modeled in terms of—and even be identical to—the past, and probability has a logical basis. Consider "throwing a die to land with one face up." Thinking of the event (as if it has just occurred, in terms of the most recent past), one can conceive six ways of looking at it In other words, one can conceive of six "most recent pasts." By the fourth law, the total "most recent past" is identical to the "most immediate future," on the boundary. If one therefore collects all six "most recent pasts," they turn into the most immediate future, by the fourth law of logic. One may attempt to argue that only a single one of the events will actually occur. Here one counters with the observation that, when it occurs, it is the past, not the future, and it has not been collected so as to move to the boundary.

158. Problem of change: as stated by Heraclitus: if a thing changes, it must turn into something else. But how can a thing be something else other than itself? The paradox cannot be resolved by 3-law Aristotelian logic, because the concept of change assumes the fourth law (i.e., that a thing can be identical to something other than itself)- The problem is resolved (by the fourth law) merely by accounting the time snapshots in which a thing is perceived, detected, observed, conceived, thought, etc That is, the original thing was perceived in time snap 1. The next thing was perceived in time snap 2. In yet another time snapshot, snap 3, the two previous snapshots are compared (in memory, which is the only place the past exists). If that comparison finds no distinction or differ-

ence between the first two snaps, then they are declared identical (by a decision algorithm). If the comparison in time snap 3 finds a difference between the previous two snaps, then they are declared nonidentical (by a decision algorithm). Most simple macroscopic things are seen as a multitude of composite attributes. Thus one or more of these attributes may be seen to have changed (not be identical between snaps 1 and 2), yet a host of other attributes may be seen not to have changed. In this case, the two overall objects are seen to be "almost identical but not quite," so that one speaks of "the thing having changed." One implies it changes in some attributes, but remains the same in most others. This is the primitive derivation of "how a thing changes, yet remains itself." On the other hand, if one defines a hypothetical thing which has only a single attribute—itself—then that thing cannot change, by Aristotelian logic. Yet even such a thing can change! Consider the single overallness of a caterpillar. Let metamorphosis occur, and it is now a butterfly. Now between snaps 1 and 2 we see a different overallness, yet we declare it the same thing! Here we have a more direct application of the fourth law: one thing (snapshot 1) can be identical to another thing entirely (snapshot 2) if the distinction between things one and two is lost in the decision algorithm applied in snapshot 3. In other words, identity and nonidentity are not absolutes, but are products of the decision criteria invoked in determining them. This resolves the age-old philosophical problem of change by expanding the logic by means of which one examines the problem.

159. Problem of good and evil: the paradox posed by the assumption of a "good only" creator moving forward in time, and the simultaneous existence of evil. I.e., if God is perfect and all powerful, how can imperfection exist? This problem is simply one expression of the "time trap." That is, the unconscious assumption of a creator moving forward in time (and "acting" and "operating," for example) is a limitation—hence an imperfection—unwittingly placed on the concept of a creator. Hence an imperfect creator could well wind up with the existence of evil (imperfection) in the creation. If one makes God unlimited—even by time—then perfection may be restored to the creator concept One then separates a positive time stream and a negative time stream (these are arbitrary, since in ultimate reality all events exist at once, unseparated, and ordering may be accomplished in any fashion one wishes). In the positive time stream, an act then becomes either good or bad, depending on whether it is beneficial or detrimental to its referent. But in the negative time stream the same act is reversed and becomes the exact opposite with respect to the referent. Thus any act is ULTIMATELY both good and evil simultaneously, and to the exact same degree. Hence in the "all-at-once" of the unlimited level, the two aspects zero and cancel, and there is neither good nor evil exclusively. Yet there is all good and all evil nonexclusively. The unlimited God is thus both good and evil simultaneously but nonexclnsively, and neither good nor evil exclusively. This is the resolution of the unsolved philosophical (and theological) problem of good and evil.

160. Psychic surgery: paranormally opening and closing the flesh of the body to remove abnormal growths or tissues. In a less drastic case, operating on die etheric or subtle energy body, rather than the physical body. In legitimate psychic surgery, scalar interferometry from the surgeon can nullify the cohesive charges binding the tissue cells together in a line; the tissues then open without cellular damage. When the surgeon closes the tissues, scalar interferometry again restores the charge and cellular binding, so the wound is "healed" without scarring (or with very little, in the practical case).

161. Psychokinesis: inceptive influence or effect upon a physical object, system, or situation by mental intent Psychokinesis is accomplished by means of scalar anenergy wave patterns projected from the two cerebral halves in the fashion of a scalar interferometer. The waves interfere at a distance, microscopically in the object to be moved or affected, forming microscopic bursts of vector wave energy in the object itself. Jack Hauck has referred to psychokenetic bending of metal as "warm forming."

162. Psychosis: seriously impaired mental functioning so that gross interference occurs*" with an individual's ability to function in a normal mental fashion. Generally characterized by severe mental disturbance, withdrawal from reality, delusions or hallucinations, regression and personality disintegration.

163. Psychotronic generator: a device which produces observable effects by collecting, condensing, amplifying, and/or processing subtle anenergies or scalar waves. Classic examples are the psychotronic generators of the Czech engineer, Robert Pavlita. Psychotronic devices are virtual state engineering devices that process and utilize scalar EM waves of massless charge flux. Most of Pavlita's devices are "charged" by the natural scalar electromagnetic emissions of the human brain and nervous systems.

164. Psychotronics: *psycho-* (mind energy), *-ronics* (electronics or devices). The union of subtle (mind) energy (that is, anenergy or scalar waves) and matter or electromagnetic field. Czechoslovakian term for parapsychology, with emphasis on the energy aspects. The comparable Soviet term was originally psychoenergetics, which was soon shortened to energetics. However, the Soviets have applied scalar electromagnetics (electrogravitation) principles and massively focussed on weaponry, not just parapsychology. 'Virtual state engineering. Electrogravitational engineering. Scalar wave engineering.

165. Pulse mode: in a transmitting scalar interferometer, projection of a single sequential pair of Fourier scalar wave patterns or "pulses", so that the second pattern meets the first over the distant target area. Flash coupling of the two scalar patterns in the target zone produces a single intense electromagnetic pulse (EMP) which—if atmospheric—interacts with the plasma expansion of the atmosphere to produce a second pulse immediately after the first The dual pulsed flash signature of the scalar interferometer weapon's burst in the target area is very similar to the two-pulsed flash from a nuclear explosion, but without nuclear debris and fallout Two such pulse mode flashes from a Tesla howitzer—possibly the Russian one at Saryshagan Missile Test Range or one from another nation—have been detected off the southwest coast of Africa by U.S. Vela satellites, one in September 1979 and one in December 1980. Other similar tests have been reported. The pulsed flash may produce positive energy (exothermic) or negative energy (endothermic), depending upon the setup.

166. PulsorR: scalar electromagnetic microcrystal device invented by the electrical engineer George Yao. PulsorsR are specifically designed to correct or smooth the patterns of the subtle scalar electromagnetic "bodies" or levels.

167. Quantum: a single "particle" of the quantity action. The quantum is always composed of two fundamental quantities (canonical variables), and it is the "fissioning" of the action quantum into two pieces which results in detection of one of the pieces. An indefinite "smear" of part of the two pieces may be simultaneously detected by a linear detector, but never both exactly (Heisenberg uncertainty principle). More recent work has shown that, in a highly nonlinear situation, both canonical variables can be determined simultaneously to any desired degree of accuracy, in which case the Heisenberg uncertainty principle does not hold. This fundamental change in what has been regarded as one of the "sacred laws" of physics has not yet made it into most physics texts.

168. Quiton: a little piece of vacuum or emptiness. A piece of the medium of a space or spacetime, of any dimensionality. A piece of pure massless charge flux. A piece of pure scalar wave flux. A piece of pure virtual flux. A piece of the vacuum. A particle of electrostatic scalar potential (infinitesimal charge). A "vacuum engine." Quitons are at least 4-dimensional (spatiotemporal), not just spatial and have a dynamic structure, which may be either random or deterministic. Note that the paper handed to Los Angeles Times reporter Bob Toth by Petukov in Moscow, resulting in their seizure and charging with possession of high Soviet state secrets, contained a theory based on "little pieces of the medium" ("little pieces of vacuum.") In other words, it was a quiton theory, regardless of the specific terminology utilized.

169. Radiant energy: term used by T. Henry Moray to describe the subtle anenergy his amplifier tapped from the unlimited vacuum energy of the cosmos, furnishing usable electrical power of an unusual (negative power) nature. By 1939 Moray was obtaining 50 kilowatts of power from the output of a 55-pound amplifier of some 29 stages, and an additional 50 kilowatts could be tapped off from any other stage. Moray's amplifier was essentially holographic, and could produce exponential amplification to any degree, merely by adding stages. The term was first used by Nikola Tesla, who patented "radiant energy" signaling systems shortly after 1900.

170. Radionics: the use of electrical apparatus (such as ganged precision resistors and potentiometers) to accurately attune the innate psychic and psychokinetic abilities of a human operator. Specifically, the operator may be attuned to some person or object at a distance, through the use of a "witness" (something containing the subtle anenergy pattern of the person or object). In actuality, the pattern contained by the witness consists of the previously impressed virtual (subtle) (scalar) energy patterns riding in and upon the object's internal electromagnetic interactions. The infolded vacuum retains an exact pattern and path of all photon interactions, be they observable or virtual. This guarantees a vacuum path exists between the witness and the original object, no matter where it is located. These impressed scalar EM patterns constitute a sort of complex "tuning fork" attuned to the impressing object or person through hyperspace (via the infolded past "backtrack" virtual memory of vacuum). Production of phase conjugation waves (time-reversed waves) in the witness thus will result in travel of these waves back through the infolded "backtrack paths" to the distant object itself. Resonating the "witness" pattern with amplified anenergy from the operator produces sympathetic anenergy resonance into the witnessed object at a distance, through the hyperspace backtrack path. Since the operator's will and intent modulate and pattern his input into the amplifier, the anenergy resonance of the distant person or object is correspondingly modified. This causes the kindling of an actual microscopic energy pattern in the distant object. The use of amplification constitutes the modern phase of radionics, essentially introduced by T.G. Hieronymus. In the Soviet Union, the incredibly powerful Moray hyperspatial amplifier is believed to be in use in a variety of energetics or psychotronics weapons. Such a holographic amplifier yields exponential power amplification to almost any degree desired. In the destruction of the U.S.S.Thresher in April 1963, a photograph of the submarine was probably utilized as a witness, and sufficient negative energy was kindled into the distant underwater submarine to seriously interfere with its controls. The sub then sank to crush depth and was imploded. Negative energy splatter in the surrounding area so interfered with the electromagnetic systems of the companion ship, U.S.S. Skylark, that over an hour and a half was required to transmit the emergency message that the sub was probably lost. In the U.S, holographic amplifiers have not formerly been developed, and radionics power is limited essentially to the linear amplification abilities of ordinary amplifiers. Thus their gains may reach 10^4, when indeed 10^{20} or so is needed for direct mechanical results immediately. For this reason the West has not developed dependable detectors of subtle energies, has not recognized the Soviet superweapon developments, and has not developed counters to them. However, Bedini, Dea, Beck, Golden, Dayney, Bielek, Bise, and other inventors have built scalar detectors of one form or another. Golden has also produced a highly specialized "booster amplifier"—the Golden booster—for Kelly radionics devices that is a step in the right direction. Both the Kelly machine and the Golden Booster have been verified as scalar wave radiators by actual measurement. With the development of detectors, the technology should begin to develop much faster in the West. The science of such devices and effects.

171. Reality: that which is actual or objective, as opposed to merely possible or ideal. One's idea of what constitutes actual and objective, however, strongly biases what is regarded as real and what isn't. In physics, for example, virtual actions are essential infinitesimals—little pieces of "nothingness" which yet produce effects. Physics thus has extended the meaning of "objective" beyond the idea of "observable." In addition, time—which is never observable or detectable, even in theory—it nonetheless assumed to be real and objective. In a sense, by "reality" one attempts to imply "ultimate." Indeed, the concept is confused, vague, and incapable of definition except in accord with what one pre-assumes it to be! The major problem is that most persons assume (1) reality is limited, and (2) reality is singular. In the new approach, reality is both unlimited and multiple. Everything is real; there are just differing degrees of reality or solidity as compared to the ordinary physical detection (threshold selection) process. Reality becomes a matter of intersections of physical universes and objects with other physical universes and objects, with the density of the intersection specifying the density or degree of physicality. (Note the similarity to the Aharonov-Bohm effect).

172. Reincarnation: refers to some element of the mind and/or personality surviving the physical death of a body/tuner, and being reborn in the body of another infant In the new approach, a type of reincarnation is scientifically derived as a consequence of the theory. However, since the past can be changed, by deep developmental work one can actually change one's so-called past incarnations. Also, more than one living person's scalar pattern can be attuned to a similar "past" scalar personal pattern—and so more than one person can be the "reincarnation" of a past person.

173. Remote viewing: the ability of human subjects to receive visual impressions and information from distant actions and scenes. The information is usually "noisy". May involve combinations of clairvoyance, clairaudience, precognition, postcognition, and telepathy. In the new theory, one recognizes that all detection is internal to the detector. Hence the entire detected physical universe may rigorously be said to exist inside each physical particle of the physical body of the observer. Thus hypothetically one has subtle access to any portion of the universe directly within oneself. The same result occurs when one assumes the use of a signal—longitudinal scalar waves of electrostatic potential—with infinite velocity. The two cerebral hemispheres act as a scalar interferometer, producing microscopic E-field by coupling incoming or rising microscopic scalar phi-field waves. This allows modification of the patterning of the avalanche discharges underway in the brain, and thus perception of the distant event Since scalar waves are hyperspatial and can be superluminal, they can move through time. This permits perception of the future and the past Unfortunately the outputs of this detection system in humans is buried deep within the unconscious nind, so it is difficult to obtain "clear signals." With the future development of sensitive scalar wave detectors, biofeedback training can probably be used to more fully develop this scalar sensing ability. Also, on several occasions Nikola Tesla announced the invention of a "big eye" that could see at a distance; such an instrument is perfectly possible by scanned scalar interferometry reception. This would include the ability to see into or through the ocean or earth—and time as well, under the proper circumstances. It should be possible to take distant pictures, or even pictures through time. One researcher, George Delawarr, reported the accomplishment of both these effects with his special camera.

174. Resonance: in physics, an oscillation phenomenon exhibited by a physical system driven by a periodic force; when the frequency of the driving force approaches a natural "free response frequency" of the system, the oscillation of the system becomes large. The "resonance (or resonant) frequency" of a system is a frequency at which some response of the system becomes maximum. Three types may be defined: (1) Phase *or* velocity resonance, where the response of the system is 90 degrees out of phase with the driving force, (2) amplitude resonance, where the amplitude of response of the system *is* maximum, and (3) natural resonance, which is the frequency of oscillation of the system to a sharp input such as a hammer blow. Electrical resonance exhibited by an alternating current reactive circuit results in relatively large currents near certain frequencies, together with the nearly unimpeded oscillation of energy back and forth between potential and kinetic form. There are two kinds, series and parallel resonance. In series resonance, the inductance and capacitance are in series, and in parallel resonance they are connected in parallel. In the series resonant circuit, the currents can become large if the resistance is small. Further, the voltages across the coil and capacitor are essentially 180 degrees out of phase, and so they essentially "cancel," leaving very little voltage across the overall circuit, even when the individual voltages across the coil and across the capacitor are very large. (Note that the E-fields across the inductance and the coil are also 180 degrees out of phase, and so the series resonant circuit to a distant observer produces an artificial potential wave—indeed, an electrogravitational wave. This is one reason why Tesla so favored the series circuit He also favored it with one end loose—the so-called "single-wire" circuit, without a return.) If the driving voltage to such a single-wire series resonant circuit is applied to the coil with the open-ended capacitor beyond, and the voltage is applied for only the first 90 degrees as a sine wave then instantly removed, the coil will be charged without current flow, accumulating a 5-d gravitational potential. When the coil discharges, it will discharge in two directions, producing real current, at a certain frequency (which may be drastically different from the "normal electrical current" resonant frequency. However, actual current will be induced out of both ends of the coil at this frequency. The capacitor will charge a little bit from this effect. The overshoot of the coil discharge back toward the disconnected driver is also usable energy if connected to a proper load. If the load is removed and the driving voltage is reapplied just as the overshoot reduces back to zero, the cycle is repeated, charging the distant capacitor more and more, and producing pulses of usable power through the overshoot discharge on the driver side. By adapting this mechanism, a free energy motor can be readily produced, as demonstrated by Bedini, Watson, and others with a battery-driven, series-wound DC motor unit Note that use of the localized gravitational potential produces local general relativity, a.locally bent spacetime, and the conservation of energy, charge, and momentum laws need not rigorously apply. An adaptation of this mechanism was also utilized by Tesla in his magnifying transmitter. In that transmitter with its huge elevated capacitance, the artificial potential wave produced would have been introduced into the earth, producing scalar resonance of the earth. The molten core of the earth, being under pressure in all directions, is self-pumped (in the language of nonlinear optical phase conjugation). The resonant atomic nuclei of the earth's molten core would then act as a pumped phase conjugate mirror (PCM), producing a highly amplified (near infinity gain!) time-reversed wave in response to the input signal. Thus the earth PCM is a giant amplifier, and the TR wave returned is highly amplified, possessing enormous energy. Thus the amount of energy resonant in the earth becomes enormous. In fact, the transmitter was designed to produce more than one such scalar earth resonance. At any distant point on the earth, a transmitter/receive station could have input a

small "tickler wave" at this frequency into the earth, and its proper antenna tuned to the frequency" of the scalar resonance would have produced enormous electrical power. A very similar system employing two frequencies 12 kilohertz apart and tapping the beat frequency, is used by the Soviet Union today to power its large strategic Tesla weapons. Golden has detected and shown these Soviet "two-frequency earth resonances" on the oscilloscope, utilizing the output of his specialized detector system.

175. Sasquatch: Indian name for a large, well-defined tulpoidal apeman form. Bigfoot (slang or common name). Generally refers to the Pacific Northwest area, where die tulpoidal form has been well-tuned for centuries by the Indians. In the Far East, the Yeti.

176. Scalar: in ordinary vector analysis, a quantity completely characterized by magnitude only. (Actually, it contains magnitude and time.) In the new approach, one must account for observable and virtual states. In addition, multiple vectors acting on one point and summing or multiplying to zero are physically still present, even though their vector resultant is a zero vector (more precisely, a zero-vector system). Thus physically a zero vector can be a system having a very real, distinct substructure of nonzero vector components. These infolded vector components may be highly dynamic. The energy of each infolded dynamic vector component is thus "trapped" inside the local vector zero system. Rigorously, trapped energy constitutes a scalar potential in spacetime, to the external observer. An observable scalar (potential) thus may have a nested, multiple-level virtual substructure as well as any conglomerate "observable vector" substructure. Enormous energy may be infolded and trapped, in which case local spacetime is appreciably curved. Note that the infolded energy of the substructure components contributes to the stress of the object or medium, and organizes it, even though to the external observer its envelope constitutes a zero vector. Each level of this substructure contains an infinite number of tiny virtual vectors. There may be finite alignment of these virtual vectors along the structure of the observable vectors infolded. This constitutes vacuum polarization. The zero-vector system is thus a vacuum engine and a virtual state engine as well. Two scalars (or vectors) that are observably equal need not be equal at all in their substructures. Further, combining the two scalars by superposition may not yield the normal observable value, if both substructures are different, highly patterned and interact to spill over into the observable level. Vector mathematics itself must be changed and extended in the new approach. Note that the spillover from interaction of the two substructures may form an observable scalar value or an observable vector value. Such vacuum engines can be constructed o do specific operations on atomic nuclei etc. Nested levels of such engines can be formed, to form highly complex "super" hyperspatial vacuum engines called giant photons. Continual radiation of matter with a signal composed of vacuum engine photons of a specific deterministic substructure will "charge up" the atomic nuclei of the radiated material with that pattern, so that it emerges into the gross, observable physical state, drastically altering the radiated matter itself mechanically, chemically and physically. This process is called kindling.

177. Scalar energy tap: refers to a Soviet method of obtaining tremendous electromagnetic energy from the molten core of the earth, with which to power gigantic strategic scalar Tesla howitzers. As measured by Golden, two signals are continually transmitted into the earth some 12 kiloHertz apart, and the earth itself is placed in scalar resonance to the two frequencies. The core of the earth, being a pumped phase conjugate mirror, responds by generating highly amplified phase conjugate replicas, which return and concentrate all their energy at the transmitter/receivers. The receivers receive the enormous beat frequency (12 kHz) energy, using it for power. One then has a kind of scalar interferometer channel in the earth's core in the endothermic mode. Depending upon the biasing, either ordinary EM energy or negative EM energy is captured in the 12 kiloHertz band by the interferometer receivers. Varying the negative biasing also provides a throttle to control the amount of energy extracted. Just before the May 1, 198S celebration, the Soviets activated some 27 such giant energy taps in the earth, and also activated a large number of command and control scalar transmission frequencies. If each tap can power 6 to 8 giant Tesla howitzers, then for their important 40th anniversary of WWII, the Soviet Union activated well over 100 large howitzers. In other words, a massive exercise of the Soviet Union's entire strategic scalar howitzer weapon complex was held on the 40th anniversary of 1945 for Gorbachev, and to signify meeting the schedule (laid down by Brezhnev in 1972) to "be ready" to use them . At the conclusion of the exercise (which lasted several days), the Soviets left two scalar energy taps in place, instead of the single one present before the exercise. In other words, additional strategic weapons were brought up on "ready alert"

178. Scalar field: in vector analysis, each point in space is assigned a magnitude; the set of spatial points and their assigned magnitudes is called a scalar field. In the new approach, an observable scalar value is assigned to every point in n-dimensional space, where n is 4 or greater, and the set of n-dimensional points and their assigned observable magnitudes is called a scalar field. Also, in the new approach each scalar magnitude is considered to contain an n-dimensional virtual-state substructure, where each succeedingly higher dimension is a succeedingly lower level of virtual state. Vacuum itself is such a scalar field. Such a scalar field is also the rigorous identity of a massless charge field, of—for example—the electrostatic scalar potential, 0. Also, the scalar field is considered to be composed of two time fields: one in positive time and one in negative time that is the phase conjugate replica of the first. Thus the timeless, lengthless vacuum exists both in positive and negative time, and its potentials are scalar potentials. When the vacuum is uncurved, equal amounts and components of positive and negative time exist. When it is curved at a point, the positive and negative time components are unbalanced at that point

179. Scalar interferometry: the interference of two scalar Tesla waves (a Tesla wave consists of a normal wave coupled to its phase conjugate replica). In the interference, coupling can occur and produce (or extract) electromagnetic waves—energy—at a distance. The phase conjugation aspect is the secret of Tesla's revolutionary discovery of wireless transmission without loss. It is also the secret of producing death rays; the invulnerable Tesla shield; antigravity; superluminal communication; con-

trolled materialization, dematerialization, and transmutation; psychokinesis; and free energy. The ubiquitous photon interaction presently is the basis for first-order physical reality, and the photon's speed results in a speed-of-light-limited reality being detected. Scalar EM waves contain vacuum engine substructures and do not necessarily have this speed limitation. Detection by scalar waves yields a higher order physical reality and a vastly expanded quantum mechanics and relativity theory. Present Einstein relativity theory is only a highly special case, as are present electromagnetics and quantum mechanics. Soviet use of scalar interferometers in a large, strategic mode is responsible for much of the advanced Soviet ABM defense and strategic superweaponry.

180. Scalar resonance: inertial or mass resonance of a body, originally discovered by Tesla, repeated (probably) by Hooper, and rediscovered by Golden, Bedim, and others. In the simplest case, scalar resonance can be imagined as two ordinary single—frequency EM standing waves (a wave and its phase conjugate replicaor PCR) in a resonant cavity, with the two waves coupled exactly 180 degrees out of phase so that—externally—their E fields sum to zero and their B fields sum to zero. This is a standing scalar wave, and the cavity is said to be in scalar resonance. Note that the cavity walls must act as phase conjugate mirrors (PCMs). Mass itself is simply such a standing scalar resonance, with the standing wave trapped by particle spin. The atomic nucleus, continually bombarded by virtual photon flux from the vacuum, is a pumped PCM. Since the nucleus is positively charged, it produces PCRs having negative energy and negative time. It thus produces negentropy, which prevents the entropic decay of the atom. Essentially all frequencies can be and are present at once, but certain nuclear resonance frequencies are thought to predominate. By increasing the amplitude of the standing scalar wave that is the nuclear pseupotential, inertia and mass of the nucleus are increased. By decreasing the amplitude, the mass and inertia are decreased. Zeroing the amplitude dematerializes the nucleus. Scalar resonance is not presently contained in the orthodox physics literature.

181. Scalar wave: a longitudinal "pressure" wave in the actual stress of spacetime itself. A wave of potential, where no EM bleedout of the infolded contents occurs. A normal wave and its phase conjugate replica, co-modulating each other and "locked or zipped together" as a zero-vector system wave. A Tesla wave. A gravitational wave. A time wave. Electromagnetically, a scalar potential wave which, in the absence of spinning charged mass particles, does not form a vector electromagnetic wave, in contradistinction to present electromagnetic theory. A "shadow vector" wave, where the ordinary "observable force vector" of the field is broken into virtual vector components which are not integrated. A wave in the nested levels of virtual state, that does not breach the quantum threshold, and hence is not observable to normal detectors. A wave in the magnitude of the virtual particle flux density of vacuum. A longitudinal wave in pure massless charge flux (vacuum) (spacetime). A wave in the magnitude of the infolded "energy density of the vacuum. The scalar wave is free to move in four or more dimensions. It transports deterministic vacuum energy engines. In Minkowski 4-space, it can: (1) move only in time, and not spatially, in which case it "sits in one place" and fluctuates the flow of time, the inertia of an object, the strength of the gravitational field, the mind, the fundamental constants of nature, etc.; (2) move only spatially, with the rate of flow of time oscillating about a linear mean; (3) move in a mixture of modes (1) and (2); (4) oscillate back and forth between modes (1) and (2) in a regular fashion, spatially stationary; and (5) oscillate back and forth between modes (1) and (2), translating spatially in travelling wave fashion. (5) move and oscillate between many vacuum levels (hyperspaces), in addition to characteristics (1) through (4). An ordinary Hertzian wave exists in vacuum as two coupled Tesla longitudinal waves. Rigorously, the E and B fields do not exist in vacuum, but on and of the detecting spinning charged particle. The spin of the detecting particle causes precession, yielding the familiar "transverse Hertzian wave," complete with E and B fields. The broken but ordered virtual state bits of what will be the detected E and B, however—shadow vectors E and B—exist in vacuum, on the virtual components of the vacuum charge flux. Tesla was correct when he stated that transverse Hertz waves (as such) could not move through the vacuum or exist in it

182. Shadow vector: an otherwise observable vector, broken into consecutive bits such that each bit is virtual. The resulting "disintegrated vector" becomes unobservable as a vector, but remains as an ordered series of virtual vectors. Thus it is a special sort of scalar wave. Observably, it is a scalar wave since no observable vector exists, but virtually it is a vector wave. It may be regarded as a localized, changing scalar field with a particular patterned virtual substructure which yields the shadow vector. On an appropriate integrating detector, the virtual components of the shadow vector will be reintegrated into an observable force vector, such as E or B. Dr. Robert Beck, for example, has reported the development of a successful integrating detector for such shadow vector scalar waves. His detector is capable of detecting signals that cannot be seen by the finest Superconducting Quantum Interference Device (SQUID), and it has been demonstrated for comparison against the SQUID. Dea, Bedini, Dayney, Golden, Beck, Bielek and Bise have independently built scalar wave detectors of various kinds. Farnsworth has particularly good and comprehensive shadow vector/scalar wave detecting systems.

183. Space: literally, the absence of observable mass, hence emptiness or void observably. Space per se has no definite length interval and no definite time interval. Rigorously, space is the absence of distinction and separation. Space has no dimensionality—i.e., it is not 3-dimensional, 4-dimensional, etc., for it has no "specific" separations. But in common usage, it is often thought of in a general sense as having some particular dimensionality but no definite intervals. The point is subtle but important: the absence of a 3-dimensional mass need not be 3-dimensional (e.g., considering a 10-dimensional frame). But one usually thinks of the "hole" where the mass was as having the same dimensionality as did the absented mass. This is only a 3-dimensional intersection of the actual hole, however, though this point is usually ignored. Ignoring the full dimensionality of a hole is responsible for the confusion regarding the two-slit experiment in quantum physics, for example. That is, the slits are (at least) 4-dimensional, since they exist in time also. In the slits, spatial mass is absent, so the slits are

three dimensional "holes." However, their fourth dimensional aspect, time, is present, not absent Thus four dimensionally the three-dimensional slit "holes" are not "holes", but are "things." They are not 4-d absences, but presences—since one dimension (time) is still present They can thus interact in the fourth dimension, time. The experiment has no mystery when the time aspects are considered, as pointed out by Charles Muses in 1957. In Kaluza-Klein 5-d mathematics, these 3-d holes also exist as things in the fourth spatial dimension. Also, in spacetime—rigorously—all 3-dimensional material objects, holes, gaps, etc. are simple facets or intersections of a single object For if time is considered as a continuum, then everything exists at once; hence spatial things are simply the features of a single 4-dimensional object The problem is that "space" is also commonly taken to be a "spatial frame," where time has been stripped off and "stopped," and a measurement to each and every point in the universe frame has been simultaneously established. This corrupt usage of "space" implies "spatial frame," and is entirely different from space as such.

184. Spacetime: three or more dimensions of space and a dimension of time, multiplied together. In the 4-dimensional case, it is often referred to as a "4—space." Spacetime can actually contain any number of dimensions; in our theory we consider an infinite number of spatial dimensions and at least one time dimension. One could just as easily take an infinite number of time dimensions alone. Scalar electromagnetics (unified EM and G fields) is most easily expressed in Kaluza-Klein geometry, in at least five dimensions—four of space and one of time.

185. Spherical model of the cosmos: a model of the universe in which the universe is a four-dimensional spherical surface in five dimensions. This results in a "closed" universe, and also one which is entirely within each point of itself. To produce such a model instead of an expanding universe, Einstein originally introduced entirely ad hoc factors into general relativity to cancel and eliminate the expansion terms.

186. Spin: the intrinsic angular momentum of a particle, as if it were a top spinning about an axis. Spin is quantized. Spinning charged particles—such as an electron—thus demonstrate a magnetic moment, due to the circulation of charge in the spinning. In the nucleus of an atom, the spin of the nucleus is the resultant of the spins of the nucleons (particles comprising the nucleus). Actually spin of particles is more like an "implosion" to "explosion" circulation; in other words, it circulates in time as well. Spin of a particle is the basic feature that integrates the anenergy (disintegrated energy) of vacuum flux into observable charge. All observable fields, matter, effects, etc. depend upon this basic mechanism to zip together virtual entities and form observable phenomena.

187. Spooklight: another slang name for a nocturnal light Legend and folklore usually associate it with a violent death or macabre occurrence. Earth stress lights. Over 1,200 known sites produce such earth stress lights in the U.S. alone.

188. Strobe effect: striated patterns seen in UFO and other paranormal phenomena due to the higher dimensional phenomena skipping in and out of normal 3-space. Alternatively, two or more higher dimensional phenomena may dynamically interfere so as to form repetitive intersections with our normal 3-space, producing striations. Hyperspatial (scalar) EM interference often produces striation effects.

189. Substructure: in the new approach, a scalar has an infinite-dimensioned (infinite number of sublevels) virtual state substructure, filled with waves, particles, etc. Vacuum has precisely the same sort of substructure. The substructure is rigorously CHARGE FLUX, where observable mass is not present. Any vector, since it contains a set of scalars, also contains a set of virtual state charge flux substructures. One can engineer, pattern, arrange, and utilize these substructures to give various "paranormal" effects. These effects include materialization and dematerialization, psychokinesis, clairvoyance, superluminal communication, antigravity, teleportation, apportation, remote viewing, production of energy at a distance, extraction of energy at a distance, electron dissolution, etc

190. Superposition: the simple linear addition and subtraction of two or more values. In the new approach, superposition need not linearly hold valid if sufficient substructure interaction is obtained and the substructures are nonlinear. In that case, superposition turns into nonlinear interference.

191. Teleportation: the transmission of an object from "here" to "there" with no travel in between as an object in ordinary space. Hyperspatial transmission of an object from one point in the ordinary space to another.

192. Telluric currents: earth currents; electrical currents in the rocks of the earth, believed to be due to aurora effects and the earth's magnetic field. In the new approach, these currents are also pumped by longitudinal scalar wave bombardment from the sun, the cosmos, and the fiery core of the earth itself. Quartz is a common crystalline component in rocks, and thus the rocks are piezoelectric This means that stresses in the rocks, such as from shifting tectonic plates, causes electrical activity, and electrical activity causes mechanical stresses. Further, the quartz is essentially randomly oriented, resulting in most of the tiny piezoelectric vectors summing to a vector zero, which in turn constitutes a scalar potential and pumping the atomic nuclei as pumped phase conjugate mirrors, producing PCR waves. These do the same thing in return, producing negative energy potentials and potential waves, etc. As the piezo-vector components vary, the values of the potentials fluctuate. The zero-sum portion of such a stressed randomized crystal constitutes a Prigogine detector for kindling and collection of anenergies via nonlinear thermodynamics. The molten core of the earth is also a dynamic Prigogine system, since it is a pumped PCM. Thus the earth is a collector, processor, sometimes amplifier, and retransmitter of mind energies and anenergies, and truly the earth may be considered to be "alive" in a strange way (the Gaia hypothesis). This process, along with scalar EM in living systems, also produces Sheldrake's morphogenetic field and its effects. Specifically, the earth receives and processes anenergies from the species unconscious of all the species of ordinary life upon it The earth itself does respond to geometric kindling alignments of other planeti. and particularly to the moon. In addition, it does respond to the state of man's collective species unconscious mind. Fault zones result in structures similar to diodes, triodes, and scalar interferometers. Amplification and rectification of kindled anenergies occur readily, including at a distance, at fault zones, particularly when seismic stresses in the rocks are increased, increasing the piezoelectric activity, telluric currents, scalar interferometry, etc Thus fault zones constitute natural kindling window areas for paranormal anenergy forms, for tulpoids etc—precisely as found by Persinger. Fault zones are also scalar interferometers, which enhances the formation of ordinary EM forms at a distance.

193. Tesla fireball: interference of two scalar Fourier expansion patterns to form a glowing, fiery globe or spherical shell of radiant electromagnetic energy and electron plasma.

194. Tesla howitzer: a large Tesla scalar interferometer such as the fourth generation Soviet hyperspatial howitzer at Saryshagan. Repeated Soviet testing of Tesla howitzers is documented in the open literature (see Bearden, *Fer-de-Lance*, Tesla Book Co., 1986). The howitzer is capable of utilization in multiple modes, both as a strategic offensive weapon and as a multiple-mode strategic ABM defense weapon. Scalar EM interferometer. Hyperspatial howitzer.

195. Tesla shield: a gigantic, hemispherical, glowing, thin shell of intense electromagnetic energy and electron plasma, placed over a defended area or installation. The EM shell is formed by interference (and coupling) of two scalar shells of anenergy, formed by truncated Fourier expansion methods from multiple frequency scalar transmitters. To form the shell at a distance, phase conjugate replica waves must be utilized, so that the energy and energy pattern does not diverge as a function of distance. Note that such real-time holography, independent of distance, already appears in the nonlinear optics phase conjugation literature. In the Tesla Shield, the energy density in the interference shell causes local curvature of spacetime, and lifts electrons from the Dirac sea, producing the plasma in the shell. With sufficient energy density in the interference shell, a material object seeking to penetrate this dome of energy and plasma is instantly electromagnetically "dudded," and physically damaged and/or vaporized. In addition, the potential pulse delivered inside the nuclei of nuclear materials that meet the shell instantly fission all fissionable materials. Thus nuclear weapons are instantly detonated when entering the shell. Electromagnetic energy (such as gamma radiation from the EMP of a high altitude penetration burst) impinging on the shell plasma is absorbed and re-radiated inside the shell at lower frequency and temperature. Several shells may be internested so that multiple, serial absorption and scattering reduces the original gamma radiation to harmless radio frequency noise. Numerous incidents of tests of the Tesla Shield weapon are documented in the open literature.

196. Tesla wave: a longitudinal scalar potential wave in massless charge, in the vacuum charge flux itself, and in pure spacetime. An electrogravitational wave. A gravitational wave. A scalar EM wave. A wave of pure potential. A wave in the magnitude of the flux energy density of vacuum. A wave in the magnitude of the curvature of vacuum.

197. Tesla weapon: any weapon utilizing longitudinal scalar EM waves or phase conjugate replica waves to accomplish its results. Also applies to a weapon using scalar interferometry or PCR interferometry and holography from multiple projectors or transmitters.

198. Thought: separations and forms in pure mind or being. To think is to form (separate, perceive) thought forms and process them. In a developmental sense, thinking is the activity of mind that screens the true continuous mind nature by fragmenting it The true mind is like the ocean, and thoughts are like waves or disturbances in the ocean. In the new scalar electromagnetics, the mind itself produces scalar waves, on the third and lower virtual levels of the scalar substructure, and these are what we call "thoughts." A thought is a real, virtual-state object capable of being directly detected, interacted with, and engineered by scalar EM means.

199. Thought photography: psychokinetic impression of images or light effects upon photographic film or other sensitive media, due to the presence or intent of a human initiator. PCR distant interferometry and holography is the basic mechanism involved.

200. Three-space (3-space): a space dimensionally composed of length cubed. It is the normal space taken to be occupied by gross matter. Per se, vacuum does not occupy this same 3-space, even though erroneously assumed to do so in present physics.

201. Time: in relativity, an extra dimension (variable) which allows the ordering of spatial snapshots or events. In quantum mechanics, time is only a parameter; it is not an observable and is not detectable, even in theory. Time is always inferred, even though precisely. Rigorously, time exists only through the stream and persistence of human consciousness. We only experience that persistence as "something changing", by continual comparison of our memory of perceptions of spatial position and pattern. In the abstract sense, time is being, which has no differentiation whatsoever—it simply is (a tautology). Perceptual separations in being constitute thoughts or mind changes, much like waves in water. Those waves common to all minds constitute physical (spatial) changes in the so-called "real" world. Because "time" in its undifferentiated sense is simply a synonym for being, then multiple things (separations) may inclusively exist in time simultaneously. (How many final objects exist in a stone before it is carved into any one of them?) The basic separator of being/time is the photon interaction, which is the basic generator of monocular quantum change and thus the progression of spatial particles "through time" in little jumps. The endpoints of the quantum jumps through time represent time differentiations of spacetime/masstime. To properly understand physical phenomena, both positive and negative time flow must be incorporated in the theory or model used.

202. Time oscillation: ordinary time advances and retards in small discrete jumps, much as a sewing machine stitches forward—except that each time stitch is pulled out before the next one is made. Thus time consists of two time flows interleaved: the positive flow of time (stitching) and the negative flow of time (unstitching). This "time stitching/unstitching" or oscillation can be viewed as a wave. It is "action" that is waving (being changed or created and destroyed in quantum chunks), since time is actually a subcomponent of action. One may therefore view time waves as carrying or creating surges of energy back and forth (in time). The photon may be considered as this oscillation of lime, for it is actually an oscillation of action. The positive half of a photon exhibits positive time (the stitch is made) and a virtual electron, and the negative half exhibits negative time (the stitch is removed) and a virtual positron. Charge is a function intimately related to, and comprised

of, time stitching. Positive time is associated with negative charge, and negative time with positive charge. The photon ca be considered as a coupled virtual electron-positron pair. The invisible stitching/unstitching trace of a photon through spacea'me is "enfolded" (in time) and remains. A phase conjugate replica photon, created at some later time and some distant place, thus has a sort of "invisible wire" connecting its own generation and the distant source that generated the stimulus photon. Note that spacetime occupies time, but space does not. Mass-time occupies time, but mass does not. Mind, thought, and being occupy time, but not space.

203. Time trap: refers to the erroneous assumption that God is a limited being, traveling along through time much as humans perceive themselves to. An unlimited creator must permeate all aspects of creation, including time. In the unlimited view, God permeates time from beginning to end, all at once. To such a creator, nothing is happening, since all happenings are simultaneously present What we call "positive" or forward time and "negative" or backward time co-exist and are inseparably identified, to such a level. Good and bad are only relative to a single time stream and to a single referent At the unlimited level, there is no separate good or separate evil. Yet there is all unseparated good and all unscparated evil. All "events" or acts that one might separate, simultaneously exist in both forward and backward time streams. Hence any event is ultimately both good and evil simultaneously, and to an identical degree. Again, this is the void that is devoid even of void.

204. Transverse wave: a wave which vibrates laterally to its motion in a medium, exactly like the vibration of a stretched string. Vector electromagnetic (Hertz) waves are transverse waves, in the present theory. In the new approach, transverse waves cannot exist as such in vacuum, but only in a longitudinally restrained medium. (The electron gas in a solid conductor, for example, is such a medium, and Maxwell's equations rigorously apply to the transmitting electron gas and to the detecting/receiving electron gas, but not to the vacuum in between). Longitudinal (alternately densifying and rarefying) scalar waves exist in vacuum, in the charge (virtual particle) flux that vacuum identically is. Hertzian waves in vacuum exist as shadow vector waves, that is, as two or more coupled Tesla scalar waves spinning around each other.

205. Tulpa, tulpoid: condensed, materialized or semi-materialized thought form, either living or inert. May be a visible or an invisible electromagnetic form, a partially materialized form, or a completely materialized form. Ultimately all matter and fields are tulpoidal in nature, consisting of condensed thought/mind energy.

206. Two-slit experiment: a form of Young's experiment where a particle is generated and directed toward a plate having two tiny slits (holes) in it, with a collection screen or device on the far side of the plate to collect and add the impacts of the particles that pass the plate. The idea is that a classic corpuscle can only go through one hole at a time, like a baseball through a hole in a fence. For particles, classically the strikes on the collector screen should be grouped in two clusters, each representing a straight-line passage through one of the slits. In the experiment, however, the particle strikes form a wave diffraction pattern when collected on the screen, unless the emitted particle is struck by a photon before it strikes the screen. In the latter case, the two clusters for classic corpuscles are observed. The experiment—fundamental to all quantum mechanics—is simply understood if one considers the time aspects of the problem, as pointed ont by Charles Muses in 19S7. The two slit apparatus with electrons may be also used as an electron interferometer for sensitive detection of scalar EM (electrogravitational) waves.

207. UFO: unidentified flying object Any object or light in the sky that cannot be identified as an ordinary physical phenomenon. After normal objects and phenomena are eliminated, the remaining paranormal residue is of interest Many of these are kindled tulpoidal forms, and may consist of lights, invisible electromagnetic energy, physical objects such as discs or spaceships, robots, humanoids, etc. (A few of these forms may be real antigravity spaceships, created here on earth by sensitive groups of several earth governments, but controlled by some mysterious group outside these governments. Such craft, I believe, have been developed and flown all over the solar system since the early 1950's. These I do not consider part of the true UFO phenomena, which in one form or another are centuries old.) Highly materialized ship forms may also contain highly materialized living forms such as space-suited astronauts etc. The "UFO" has been separated from other tulpoids (bigfoot, fairies, spooklights, water monsters, etc.) because its appearance in the sky was immediately associated with aircraft or spaceships, leading to the initial expectation that extraterrestrials in advanced spacecraft were visiting the earth. Most of the UFO phenomena actually consists of lights at night (well over 80%), and only about 6% or so resemble well-materialized ships with astronauts. A substantial portion of the "lights at night" phenomena are also due to Soviet scalar EM interference weapons. The Soviets create a great deal of light phenomena worldwide, to spark UFO reports, which are discounted by conventional science and most of the military and government agencies. This provides the Soviets a giant and effective "deception plan" so that the weapons can be tested and readied worldwide, and sightings and reports will be discredited and not cause undue alarm.

208. Unconscious: unaware, not conscious. Refers to that part of the individual's mind which is not normally in conscious awareness, but which may contain a great deal of material from the conscious mind. In the new approach, the unconscious mind is considered to be "multiply conscious," while the conscious mind is considered "singly conscious." In other words, the unconscious is a "parallel processor," while the conscious is a "serial processor." Unconscious should be thought of as not singly conscious, but multiply conscious. Further, this allows a totally new idea: Multiple singly conscious, different minds coexist simultaneously in time, though not in the same "3-dimensional mindspace." Suppose we consider a set of n such singly conscious minds. All the minds exist in a spacetime of dimensionality 3n + 1 (n is the number of different conscious minds in the set, and an additional dimension, time, is common to all minds in the set). Thus the considered set of singly conscious minds constitutes one multiply conscious—or "unconscious" -mind. In this way one may apeak of "levels of unconsciousness" in relation to groups of persons, with the deepest level being and containing every human mind. Of course the analogy can be continued to yield

the higher minds or consciousness projected by Teilhard de Chardin. In the collective human unconscious, those forms which are in each and every mind are the archetypes discovered by Jung. In a shallower level, other forms may be common to that level, constituting "weak archetypes" or "fragmentary archetypes." There is considered to be no limit to levels of unconsciousness—the ultimate depth being "all consciousness" (and all unconsciousness!) of Godness. God must be able to both be and not be at the same time; if God cannot do that little trick, then God has a limitation and is not unlimited.

209. Vacuum: pure virtual particle, massless charge flux. The virtual state, or ether (nonmaterial, Lorentzian). Lengthless, timeless spacetime itself. That state where everything is disintegrated, but highly dynamic. Note that vacuum is not an emptiness FILLED with massless charge, rather, it IS identically massless charge (disintegrated dynamicism). It is a plenum, not an emptiness. It is also pure action, undifferentiated. Einstein once suggested that the vacuum, complete with its electromagnetic and gravitational fields, should be taken as the ether. In the modern view, the term ether is again accepted in the literature, but now it refers to a "virtual particle flux" ether, not a thin material ether. One can model the vacuum analogous to a special kind of virtual-particle gas: one in which the particles "appear" and "disappear from and into nothing, nearly instantly, so fast that they cannot be individually detected. Thus the modern idea of the vacuum is a seething inferno of fiercely bubbling, fleeting particles. The other limitation is that the dimensionality of the vacuum is not fixed until operations are invoked, and neither is the rate of passage of time. Thus the vacuum has no fixed lengths nor fixed rates of time flow per se. One can visualize the stress of vacuum as a sort of conglomerate potential "pressure," where the individual stresses of various types of particles, if integrated by some external means, would sum into the overall stress (pressure). Variation of the "stress of vacuum" between two points in a frame represents a "curved spacetime" or "accelerated frame." Note that overall stress can be held constant, while two or more different components are canonically varied. The electrostatic potential, the magnetostatic potential, and the velocity of a charged particle are three such "canonical" variables. If the overall "pressure" is not varied, the three variables are canonical and energy, momentum, and charge are conserved and one cannot build a perpetuum mobile (perpetual motion machine). If the overall "pressure" is varied between local points in the frame, then the conservation of energy, momentum, and/or charge need not rigorously apply between those local points. At any one point, one may have a "free source" or a "free sink," depending on the vacuum curvature and the EM standing wave relationship. In that case, one can build a local "perpetuum mobile," since one can tap a "free source." Einstein created a severely restricted general relativity (GR), where he assumed away any local curvature of vacuum. In the West such a severely restricted GR is still dogmatically enforced, because it "saves" the conservation laws, which are regarded as sacrosanct In Soviet physics journals, learned papers by leading scientists are regularly published in which unrestricted GR—including local curvature—is developed and explored. In a local curvature region, the Soviets have already pointed out that any and all conservation laws may be violated. The Soviets have long since developed and weaponized unrestricted GR, while our scientists are still dogmatically insisting that one cannot deterministically engineer local curvature of spacetime. Thus we maintain our obsession with conservation, and the Soviets maintain the weapons.

210. Vector: in mathematics, an entity completely characterized by a magnitude and a direction. Abstractly, vectors are considered to be formed by a difference in potential between two spatial points. In the new approach, vectors are characteristics of particle motion, and do not observably exist in the absence of a particle. Since all detectors have mass, and this mass contains spinning charged particles, then force vectors are constituted on and of the spinning charged particles. In the vacuum, only an unzipped (disintegrated) vector pattern exists—or, as Feynman put it, only the potential for a force vector exists there. Maxwell's original theory was written in quaternions. In 3-space, a vector has the familiar three components, say, ai + bj + ck, where i, j, and k are the usual x, y, and z unit vectors, and a, b, and c are the magnitudes f the x, y, and z components of a vector. A quaternion, however, has an additional scalar component—say, w—so that it looks like w + ai + bj + ck. Maxwell's original theory actually unified gravitation and electromagnetics! When vector zeroing operations occur, the quaternion may not zero out the scalar component. As an example, consider the cross product of a vector A with itself, AxA. In vector analysis, the resulting vector is the zero vector. In quaternions, however, w remains, and w = a2 +b2 +c2. With the correct proportionality constants, w therefore represents the energy density of vacuum due to the two zero-multiplying vectors A1 and A2, where A1 = A2. To recover my scalar wave or gravitational wave: Now let the vectors A1 - A2 where each is the E-field vector of an identical EM sine wave. (That is, we perform a device which accepts two identical input EM sine waves and performs a cross multiplication. That device, e.g., may simply be a special kind of nonlinear modulator/mixer). In ordinary Heaviside/Maxwell theory, the EM output of this operation is a continuous zero vector, and that's all. However, by Maxwell's actual theory, the output of the operation is a scalar EM wave, whose amplitude varies as the energy of A. That is, the operation produces a scalar EM wave which is varying the energy density of vacuum. By definition, that is a gravitational wave. Since it was produced electromagnetically, I refer to it as an electrogravitational wave. It was included in Maxwell's original 1864 theory. After Maxwell's death, Heaviside and Gibbs completed the modern form of vector analysis. Heaviside—a single man—then "translated" (mutilated) Maxwell's theory into his new "vector" theory. He simply sawed off and discarded the troublesome scalar component of the quaternion! That made calculation a whole lot easier, and easier to understand. In discarding the scalar component, he discarded electrogravitation, writing an EM theory that specifically excludes local curved spacetime and gravity. Before the turn of the century, a short debate involving only about 30 scientists resulting in the complete adoption of Heaviside's mutilated version of Maxwell's theory. Quaternions and quaternion theory—and electrogravitation—vanished from the literature and the textbooks—and from the minds of the students. Maxwell's theory has not been taught or utilized in the Western

world in the 20th century! Only Heaviside's mutilated version of it is taught as "Maxwell's theory." We should have been applying Maxwell's unified theory and working on antigravity spaceships, shortly after the U.S. Civil War!

211. Vector field: in ordinary theory, the assigning of a magnitude and a direction (that is, a vector) to each point in space. The entire set of vectors and their spatial points is called a vector field. In the new approach, such a vector field cannot observably exist in a vacuum, since only disintegrated things exist there; there each potential vector pattern exists in broken virtual bits, ordered in line to form the bits of the observable vector but not joined. In the presence of a detecting/integrating mass, such as a spinning charged electron, the virtual bits and the spinning charged mass couple into a system that identically comprises the observable vector. In other words, the spin of an observable particle simply "integrates" the ordered virtual bit pattern, just as an old spinning wheel twists and integrates individual fibers into a stranded rope or thread.

212. Virtual (state): that part of physical reality consisting of changes (particles, waves, and other entities) smaller than the least detectable (quantum) change. Disintegrated changes. A thought or thoughtform is also virtual. May be visualized as consisting of real, dynamic "pieces of vacuum" or "pieces of nothing"—that is, quitons. Every ordinary observable particle (photon, neutron, electron, proton, etc.) contains or consists of a dynamically changing cloud of virtual objects. Exchange of virtual objects is thought to generate almost all forces of nature. Conservation of energy need not be conserved in virtual interactions, and generally is not In the new approach, the mind may be considered a special virtual universe, and a thought a virtual object. Virtual state consists of multiply nested levels, each level progressively finer and more subtle. These sublevels also correspond mathematically to higher geometrical dimensions. Charge on a mass is also due to its continual emission and absorption of virtual particles. Waves in the virtual state massless charge flux, which do not breach the quantum level to become observable, are vectorially nonobservable yet real—these are scalar waves, or shadow vector waves, and are longitudinal—that is, they are oscillations of the stress energy tensor of vacuum. In addition, the virtual particles in the vacuum flux arise out of nothing spontaneously (creation) and disappear into nothing spontaneously (annihilation). Since this is random, at any instant any sort of pattern imaginable is "just appearing" and "just disappearing" in the vacuum. Anything and everything at all! Thus a ghost-pattern of everything exists in the vacuum at any point or region, at any time. The vacuum is thus a giant virtual-state hologram. Integrating (kindling) one of these forms will result in it becoming first a neutrinic pattern, then a photonic (light) pattern, then a materializing pattern, and finally a material pattern. Discharge of the "potential" of this pattern then reverses the process. These two processes—charge-up and discharge—constitute materialization and dematerialization. Anything at at all-living or dead—can be materialized or dematerialized.

213. Virtual state engineering: the direct cohering, tailoring, collecting, resonating, and directing virtual state energy (anenergy) and entities into and onto normal observable objects and ordinary fields. It includes the **direct engineering of mind and life energies. Synonymous with psychotronics (Czech term) or energetics (Soviet term). Use of scalar electromagnetics (electrogravitation). Use of scalar wave/Tesla technology. Use of electrogravitation techniques.**

214. Water structuring: the arrangement and relations of the molecular and ionic parts of water. Essentially water molecules bond together by means of hydrogen bonds. The structure of these bondings is extremely rich and dynamic in water. Essentially a glass of water is a single complex molecule, one which is continually changing its bonding structure to adapt to the slightest change in its total surroundings. Water also interacts intimately with virtual crosstalk (scalar wave) anenergy. Water can thus be "charged" with ancnergy, appreciably changing its structuring. Its surface tension, for example, will change as its structuring changes, as shown by Miller and Flanagan. Thus water is possibly the most complex substance in the universe, and not just a simple chemical. With most of our body composed of intercellular water, we are all linked to every subtle change of the cosmos, by the interaction of the structuring of this water with even the most subtle cosmological changes. Large bodies of water are particularly susceptible for temporary kindling of deeper archetypal thought forms (tulpoids) such as Loch Ness Monster etc. The study of the structuring of water is still in its infancy.

215. Watson, James: young inventor who has produced several types of free energy and scalar electromagnetic devices, including battery-driven, battery-less, and tube systems.

216. Weather engineering (scalar): use of scalar electromagnetics (electrogravitation) to change and influence the weather and the producers of weather such as jetstreams and large ocean currents. By imposing scalar signals upon ordinary powerful over the horizon (OTH) radar carriers in the communications band (3 to 30 MHz), interference patterns (grids) may be established over distant and vast areas, far beyond the horizon, or even on the other side of the earth. The giant Soviet Woodpecker OTH radars continually establish such a grid over the U.S. Phase conjugate energy can be transmitted and will travel along these beams, just as electricity travels along a wire. By scanning narrow phase conjugate wave beams in these wider carrier beams, PCR realtime interference holography can be established in one or more of the distant OTR EM carrier interference grid squares. Either positive or negative energy and energy forms may be transmitted to emerge, producing a high or low pressure spot or zone as desired. By adding positive energy to a zone, the air is heated and expands. Thus the air is less dense, producing a low pressure zone, or a low. By adding negative energy to a zone, an equal amount of positive energy in the zone is canceled. Thus the air in the zone contracts and is denser, producing a high pressure zone, or a high. Thus one can create distant highs and lows at will. By slowly moving (scanning) the PCR beams, these distant highs and lows can be moved, thus deviating and directing the jetstreams, storm systems, etc. By outputting positive energy and warming the top of a thunderstorm anvil, while outputting negative energy to the side of the anvil and cooling it, drastic "downbursts" under the anvil can be created. U.S. satellites have photographed such anomalous holes appearing in thunderhead anvils over the U.S., and their association

with downbursts has been documented in the literature. Whether these are natural or engineered have yet to be established. Another signature of weather engineering—a "wheel and spoke" type of cloud, called a giant radial—has been seen widely throughout the U.S. and correlated with anomalous weather changes shortly thereafter. Sometimes a system of two such radials facing each other is seen, and on one occasion four were seen locked together (just before and during the highly anomalous zig-zag travel of a 1985 hurricane in the Gulf.) Another geometrical form associated with severe weather has been photographed by U.S. weather satellites, and is referred to in the weather literature as an "enhanced V". This signature is closely associated with the production of tornadoes, and it may be a natural phenomenon. However, if the mechanism for the enhanced V were artificially produced in a thunderstorm, then the probability of follow-on tornadoes would be materially increased. The Soviets have openly engineered the weather over North America since 1976, and made previous R&D engineering tests over the U.S. in 1967.

217. Woodpecker signals: powerful, complex electromagnetic signals originating in the Soviet Union (apparently from the vicinity of Riga and Gomel, for example) that have widely disturbed and interfered with the communications band (3-30 MHz) since the mid-70's. Bioactive modulations of 10 Hz and under have been observed repeatedly. Western observers have not deciphered the purpose or specific coding of the signals, and prefer to think of them as some sort of over-the-horizon radar tests, thus neatly disposing of the problem. Similar signals have been observed to entrain the human brainwaves and force them into in-phase oscillation. Beck has reported the phaselocked synchronization of the 10 Hz modulations on as many as 16 carrier frequencies at once. From the Lisitsyn work, this constitutes a "phased lock-in" of 16 carrier channels into any entrained brain. According to Lisitsyn and earlier Soviet "biological radio" experiments, this situation should allow the Soviets to transmit emotion, feeling, thought, and even pictures into the entrained brain, where it is processed as if the brain itself had originated it. (For example, in a Cable News Network special in November 1985, Rauscher and Bise actually demonstrated the induction of visual form images in the brain of the CNN investigator using weak signals much weaker than those naturally produced by oscillation of the earth's magnetic field.) According to the Kazhnacheyev work, any sort of cellular death and disease pattern can be electromagnetically induced. (The work was duplicated in the infrared by University of Marburg researchers in West Germany, by researchers at the University of Sydney, and by one U.S. scientist.) Popp discovered the virtual photon communication system of the cells of the body. This is scalar, and scalar patterns for charge distributions of the chemistry of cellular diseases and damage could also be placed on the phase-locked Woodpecker carriers, entering die human body's master control system through the entrained brain and "kindling" throughout the cells of the body. The stress and suppression of the immune system would be involved in almost all disease patterns, and a "partial AIDS" type of syndrome would result, making the entrained body vulnerable to whatever diseases were around, including being induced on the Woodpecker carriers themselves. Also, by targeting the DNA of specific organisms, mutations of deadlier and tougher strains of disease induction agents such as bacteria and viruses could be produced. The "antithesis" pattern—a sort of scalar EM antidote—could be induced in the Soviet population. (The present lethal AIDS is a manmade disease, and it was unleashed from a research laboratory either accidentally or by a Soviet agent on purpose.) Indeed, World War III may already have begun, and the Soviets may be waging a quite different kind of creepingly escalating viral war against the United States. If so, our Defense Department and our government have not yet grasped what is happening, and so far we are simply a sitting, passive target If our populace is slowly debilitated and storm, weather, and earthquake damage slowly increases year by year, at some point our system will simply start to come unglued, with increasing loss of control. In that manner we could "subside" as a world power without the kind of great overt war we are familiar with. Note that steady escalation of international terrorism, insurrection, and guerrilla warfare being promoted by the Soviets around the world will keep us preoccupied and busy while our debilitation and moraise steadily are increased. We may already be in an eery new kind of war that we could lose without firing a shot if our leaders do not wake up to the frightening implications of the new physics! Thus the Woodpecker transmitters represent a potential anti-population and strategic weapon of enormous strategic significance. The transmitters, because of their sheer power, also appear to be able to pump and resonate earth scalar EM energies, producing earthquakes and plasma discharge effects at remote locations. (Since Golden has directly shown that the entire earth is in frightful, forced scalar resonance at multiple frequencies and enormous power due to powerful Soviet scalar transmissions, then one should not be too skeptical of the Soviet ability to produce earthquakes on demand and to order.) At least one U.S. researcher or group seems to have been partially nullifying Soviet efforts in this respect by occasionally pulsing the earth scalar energies in a countering mode, to include actual destruction of a Soviet transmitter in one or more instances. Note especially that unorthodox researchers have directly confirmed that the Woodpecker signals contain powerful scalar components which orthodox scientists—lacking scalar detection equipment—do not see. The Woodpecker carrier interference grid over the U.S. is also a scalar EM interference grid, and used by the Soviets to engineer the weather over North America. It is also used as an anti-Bomber and anti-missile launch phase defense system, and the Soviets have tested it in this mode by destroying several U.S. aircraft and missiles.

218. Work: conventionally, the line integral of force on a body over the path taken by the body. In the new approach, the force consists of the mass and its motional change, unseparated. Thus work is the line integral of the force over the path taken by the force. Work (past) is equal to the energy (present) expended, and energy (present, gravitational potential) is equal to the work (past, release of a gravitational potential) that can be expended. Note that only vectors produce work; scalar waves do not perform work on their observable environment, and they do not contain standard EM force field energy per se.

219. Yao, George: engineer and inventor of the PulsorR, which affects the subtle scalar levels of the human body's aura or scalar radiation pattern.

220. Veti: the Eastern—particularly Tibet, Nepal, etc.—Himalayan version of the apeman tulpoid, corresponding to the U.S. Sasquatch or Bigfoot.

221. Zarg: the collective human unconscious, considered as a mind and a living entity. It is at least 4—billion dimensional, since it has as many dimensions as human stages. Each human is regarded as a single cellular stage in this entity. The psychokinetic power of such a linked mind is proportional to an exponential of the number of individual stages. Thus Zarg's psychokinetic power is so great that it has essential control over space and time, and largely prescribes, maintains, and stabilizes the laws of nature (laws of physics) and the reality in which we live. Conflict between large groups of humans constitutes unresolved conflict in this great unconsciousness, and this generates symbolic dreams, much as a human dreams strange, emotional, symbolic dreams due to the unresolved conflict in his or her personal unconscious mind. However, because of its vast psychokinetic power, Zarg materializes or partially materializes its symbolic dreams, resulting in tulpoids. Waves of these tulpoidal materializations thus represent strong human conflict or threats on the large scale. Such waves can be dream analyzed to discover the correlate that is generating the symbology, by straightforward dream analysis techniques. In olden times, part of this was known, and wise men were employed to analyze and interpret the strange, induced dreams of the king (who was the center of power of a nation) and the paranormal phenomena appearing near him. The ancients knew the import of "signs and wonders" and what they portended, but this metapsychological knowledge has been largely ridiculed and lost by present materialistic Western science.

222. Zero: naught or nothing; the absence of thing. A naive definition is the absence of number. However, if all the nonzero numbers are considered to represent the presence of number, then to declare zero a number—that is, the presence of a number—is a self-contradiction since it has just been declared the absence of number. In the numerical sense, one may do a little better by calling zero "the absence of any and all numbers except itself." Yet in this case, difficulties arise as to how one is to justify the addition of any other number to zero: how can a number be present with the absenting operator, so as to be added? The absence of any type of thing observable. Formally, the additive identity element of an algebraic system. That is, formally, the basic idea of zero is that, added to a number, the result is still identical to that number. Implicit in this formal definition is a "single level** of perceived objects; that is, there are no "hidden objects" in a separate level of sets. If multiple levels of perception are allowed, then the zero becomes much more complex. That is, it represents the absence of all other numbers at that particular level, and of that level's kind. In that case, it may contain any combination of objects from another level, yet be perceived as containing "nothing of its own assigned level." In that case, there exists an infinity of different "kinds" of zeroes, if the hidden associated structure (the substructure) If taken into account Unfortunately physics has created such multiple levels, but has not required its mathematics to clearly specify the changed nature of the zero. As an example, at the observable level, an infinite number of kinds of zeroes exist, when the virtual level and pattern **associated with the zero are taken into account This characteristic of physics actually violates the following zero axiom: an observable v plus an observable zero is considered to be equal to an observable zero plus v. In fact, the sum of v and a particular zero may not be equal at all to the sum of another zero and v, if the virtual patterns of the two zeroes differ, v has a virtual substructure, and in the addition of a zero and v, their virtual levels nonlinearly interact instead of linearly superimpose.**

223. Zero point energy: the kinetic energy that remains in a substance when its temperature is absolute zero. Quantum mechanics requires a simple harmonic oscillator to still possess kinetic energy at its lowest state: that is, it possesses half a quantum of energy. In solid materials it is distributed throughout the lattice vibrations, contributing to the binding energy. Vacuum also has zero point energy. Actually a similar and much expanded concept applies to any potential: a potential is just a bunch of trapped dynamic vectors, hence "trapped vector (translational) energy." It is translational energy that is locally trapped and not translating. The potential is thus like an accumulator or capacitor. It can be "charged up" and "discharged." It thus contains its own analog or trapped or "zero summed" or "zero point" energy.

224. Zero-point energy of vacuum: the vacuum is increasingly being regarded as composed of an incredibly dense structure of virtual electromagnetic energy, even at zero degrees absolute. From quantum geometrodynamics, as one approaches the Planck length—say, 10 to the minus 35th meters—these oscillations become so energetic that the fabric of spacetime itself is torn, resulting in sporadic, continual connection and disconnection of widely separated portions of spacetime. This "quantum foam" was dubbed "superspace" by J.A. Wheeler. Actually, if space and time are torn, then energy itself—and vector waves—are also torn. That is, the superspace consists of pure massless charge flux, pure scalar waves. If compacted, this anenergy density of vacuum may be approximated as 10 to the 100th grams per cubic centimeter, were it compacted into mass. The actual electrical charge of the vacuum appears to be enormous, if it were possible to measure a voltage drop from it to an uncharged region. Thus it is space (actually spacetime) that is incredibly dense, and matter that is ethereally thin. Spacetime goes through matter, rather than matter through spacetime. This anenergy of vacuum does interact with electromagnetic fields and matter to give observable effects, as evidenced by the Lamb shift, for example. The Lamb shift proves that the vacuum charge or anenergy can be "tapped." Tapping the "vacuum energy" would immediately end the energy crisis forever. Further, it can be done with practical devices, as T. Henry Moray showed in the 20's and 30's, and as several present inventors have shown today.

225. Zero vector: a vector having no length (no magnitude) or specific direction. Absence of any or all finite vectors. Resultant of a system of multiple, finite vectors whose vector sum has no specific direction. Note that the second definition differs drastically from tho first. That if, we cannot say that a vector resultant zero has no infolded magnitude, if we consider the components, all of whom have magnitude. See the discussion under "zero." In vector analysis, the zero vector axiom states that any vector v plus a zero vector is equal to a zero vector plus v. (Note that this axion is not proven, just assumed).

However, there are an infinite number of different zero vector resultant systems, and so zero vectors may differ, if their infolded component "substructure" is considered. Indeed, a zero vector system must be identified as a special form of potential, if its substructure is considered. The common practice in electrical physics of replacing a zero-summed system of vectors with a vector zero is incomplete, and such did not exist in Maxwell's original quaternion theory. This present practice should be changed; the vector zero should be replaced with a special organized vector zero system and a scalar potential. By failing to do this, physics excludes the ability to engineer electrogravitation, spacen'me, the virtual state, local general relativity, free energy, effects at a distance, and the probabilities of the states propagated by the Schroedinger equation. It also excludes scalar electromagnetics (electrogravitation), and unification of all forces in an engineering electromagnetics fashion. In the new view, a zero vector may be a system of vectors that vectoiially sum or multiply to zero, but the components may still exist "inside" and dynamically function inside the zero-resultant envelope. This substructure can act on nonlinear or resonant systems. Any or all components may be time varying, or none of them may be. The zero vector system may thus be a dynamic vacuum engine. In addition, such a vector zero system is considered to be also a scalar stress potential, whose magnitude is equal to the sum of the absolute values of the perpendicular components. In addition, a separate type of stress potential may be included for the magnitude of the absolute values of the swirl components (torques). Other associated potentials may be included for the system, particularly if the system is in multiple dimensions, if it consists of more than one type of virtual particle flux, if several fluxes intercommute, etc. Such a system can also be designed and created to deterministically interact with and on the life force itself, the mind, thought, and personality.

BIBLIOGRAPHY

"Analysis of a Phenomenon Occurring at Bell Island, Newfoundland, **Canada on** April 2, 1978," P.A.C.E., Ottawa, Canada, 1978.

Thomas E. Bearden *Quiton/Perceptron Physics: A Theory of* ***Existence, Perception,*** *and Physical Phenomena,* Defense Documentation **Center (DDC), March** 1973.

Thomas E. Bearden, *Field, Formon, Superspace, and Inceptive Cyborg:* ***A Paraphysical*** *Theory of Noncausal Phenomena,* DDC, 1974.

Thomas E. Bearden, *A Conceptual Derivation of Einstein's Postulates of* ***Special*** *Relativity,* DDC, October 1975.

Thomas E. Bearden, *Restatement of the Heisenberg Uncertainty Principle For the Condition of Superposition,* DDC, October 1975.

Thomas E. Bearden, *A Postulated Mechanism That Leads to Materialization and Dematerialization of Matter and to Antigravity,* DDC, October 1975.

Thomas E. Bearden, *The Boundary Identity of Exact Opposites: A Simple Solution to the Age-Old Philosophical Problem of Change,* DDC, October 1975.

Thomas E. Bearden, *An Approach to Understanding Psychotronics,* privately published, 1976. Defense Documentation Center, AD-A027 866.

Thomas E. Bearden, "The One Human Problem, Its Solution, and Its Relation to UFO Phenomena," *Pursuit* (Journal of the Society for the Investigation of the Unexplained) January 1976.

Thomas E. Bearden, "Writing the Observer Back Into the Equation," address given to Princeton Center for Alternative Futures, Princeton, N.J., March 5, 1976.

Robert C. Beck, *Extreme Low Frequency Magnetic Fields and EEG Entrainment, A Psychotronic Warfare Capability?,* Bio-Medical Research Associates, Los Angeles, March 1978.

Robert C. Beck, *Instrumentation for Detecting, Recording, and Analyzing Natural and Man-Made ELF Signals and Human Brainwaves,* Bio-Medical Research Associates, Los Angeles, May 1978.

William Bise, "The Soviet 10 Hz Pulses—A Brainwave Programming Mechanism?" *Specula* (Journal of the American Association of Meta-Science), Huntsville, Alabama, Vol. 1, No. 3, 1978.

T. Townsend Brown, "Electrical Self-Potential in Rocks," *The Psychic Observer and Chimes,* Vol. 37, No. 1, January-March 1976.

Trevor James Constable, *The Cosmic Pulse of Life,* Merlin Press, Santa Ana, CA, 1976.

Trevor James **Constable,** "You, Too, Can Photograph UFO's **from Airliners,"** *Psychic* ***Observer and Chimes,*** **Vol. 37**, No. 3, May-June 1977.

George D. Curtis, "An Electromagnetic Radiation Pattern Over The Ocean," *Undersea Technology,* Vol. 5, No. 8, August 1964.

Lars-Erik DeGeer, Swedish National Defense Research Institute, "Airborne Short-Lived Radionuclides of Unknown Origin in Sweden in 1976," *Science,* Vol. 198, No. 4320, December 1977.

B.V. Deryagin and I.I. Abrikosava, Soviet Physics, JETP3, 819 (1957); 4, 2 (1957).

P.A.M. Dirac, "Development of the Physicist's Conception of Nature," *Symposium on the Development of the Physicist's Conception of Nature,* ed. Jagdish Merha, D. Reidel Publishing Company, Boston, 1973, pp. 12-14.

P.A.M. Dirac, "Quantized Singularities in the Electromagnetic Field," *Proceedings of the Royal Society of London,* Series A, Vol. 133, 1931, p. 6.

William W. Edison, David L. Faust, Harry J. Kyler and John O. Pehek, "Kirlian Photography: Myth, Fact and Applications," presented at the session on *New Electronic Methods for Medical Diagnosis and Treatment Using the Human Energy Field: A New Beach-Head for Scientific Discovery,* Electro 78, May 23-25, 1978, IEEE, New York.

Robert Eisberg and Robert Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles,* John Wiley & Sons, 1974, pp. 63-68, 146-150.

Hugh Everett, III, *The Many-Worlds Interpretation of Quantum Mechanics, A Fundamental Exposition,* with papers by J.A. Wheeler, B.S. DeWitt, L.N. Cooper and D. Van Vechten, and N. Graham; eds. Bryce S. Dewitt and Neill Graham, Princeton Series in Physics, Princeton University Press, 1973.

Richard Feynman, *The Character of Physical Law,* M.I.T. Press, 1967, pp. 31-33.

Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures On Physics,* Addison-Wesley, 1963, Vols. 1, 2, and 3.

Marjorie E. Fish, "Journey Into the Hill Star Map," *Proceedings of the 1974 MUFON Symposium,* Mutual UFO Network, 1974.

Richard Garvin, *The Crystal Skull,* Pocket Book, 1974.

Burkhard Heim, *Unified Quantum Field Theory of Matter and Gravitation,* Mutual UFO Network, Seguin, Texas, 1972.

Louise and Galen Hieronymus, *Tracking the Astronauts in Apollo 11, With Data From Apollo 8 Included,* Advanced Sciences Research and Development Corporation, Inc., Fort Lauderdale, Florida, 1969.

J. Allen Hynek, *The UFO Experience: A Scientific Enquiry,* Ballentine, 1974.

J. Allen Hynek and Jacques Vallee, *The Edge of Reality,* Regnery, Chicago, 1975.

H.E. Ives, "Derivation of the Lorentz Transformations," *Phil. Mag.,* Vol. 36, June 1945.

H.E. Ives, "The Behavior of an Interferometer in a Gravitational Field," *J. OPT. Soc. Am.*, Vol. 29, May 1939; and "The Behavior of an Interferometer in a Gravitational Field: II. Application to a Planetary Orbit," *Ibid.*, Vol. 38, April 1948.

David Michael Jacobs, *The UFO Controversy in America*, Indiana University Press, 1975. [Foo-fighters]

Brian D. Josephson, "Possible Connections Between Psychic Phenomena and Quantum Mechanics," *New Horizons*, Vol. 1, No. 5.

V.P. Kaznacheyev et al, "Apparent Information Transfer Between Two Groups of Cells," *Psychoenergetic Systems*, Vol. 1, No. 1, December 1974.

V.P. Kaznacheyev et al, "Distant Intercellular Interactions in a System of Two Tissue Cultures," *Psychoenergetic Systems*, Vol. 1, No. 3, March 1976.

Louis Kervran, *Biological Transmutations*, Swan Publishing Company, Binghamton, N.Y., 1972.

"Kindling, Once Epilepsy Model, May Relate to Kundalini," *Brain/Mind Bulletin*, Vol. 2, No. 7, February 21, 1977.

N.A. Kozyrev, "Possibility of Experimental Study of the Properties of Time," *Pulkovo, O vozmozhnosti eksperimental 'nogo issledovaniya svoystv vremeni*, September 1967, pp. 1-49, JPRS 45238, May 1968.

Michael Ledeen, "A Mole in Our Midst?", *New York*, October 2, 1978, pp. 55-57.

E.M. Lifschitz, *Soviet Physics* JETP 2, 73 (1956).

James M. McCampbell, *Ufology: New Insights from Science and Common Sense, Jaymac Company, Belmont, CA, 1973.*

George Meek, Healers and the Healing Process, Theosophical Publishing House, 1977.

T. Henry Moray, *The Sea of Energy in Which the Earth Floats*, Cosray, Salt Lake City, Utah, 1960.

Captain John D. LaMothe, Ph.D., *Controlled Offensive Behavior*—U.S.S.R., ST-CS-01-169-72, Defense Intelligence Agency, Washington, D.C. (Released under the Freedom of Information Act.)

Captain John D. LaMothe and Mr. Louis Maire, *Soviet and Czechoslovakian Parapsychology Research*, DST-1810S-387-75, Defense Intelligence Agency, Washington, D.C. (Released under the Freedom of Information Act.) [Pavlita]

Charles Muses, "Applied Hypernumbers: Computational Concepts," *Applied Mathematics and Computation*, Vol. 3, 1976.

Charles Muses's foreword to Jerome Rothstein, *Communication, Organization and Science*, The Falcon's Wing Press, Indian Hills, Colorado, 1958.

Charles Muses and Arthur M. Young, eds, *Consciousness and Reality*, Outerbridgc& Lazard, 1972.

E.I. Parnov, *At the Crossroads of Infinity,* translated from the Russian by Vladimir Talmy, MIR Publishers, Moscow, 1971.

Harold Puthoff and Russell Targ, *Mindreach,* Delacorte, 1977.

Harold Puthoff and Russell Targ, "A Perceptual Channel for Information Transfer Over Kilometer Distances: Historical Perspective and Recent Research," *Proceedings of the IEEE,* Vol. 64, No. 3, March 1976.

The Pyramid Guide, ed. Bill Cox, El Cariso Publications, Lake Elsinore, CA. (Bimonthly: subscriptions are $6.00 a year.)

Edward W. Russell, *Report on Radionics,* Neville Spearman, London, 1973.

R.M. Santilli, "Partons and Gravitation: Some Puzzling Questions," *Annals of Physics,* Vol. 83, No. 1, March, 1974, pp. 108-157.

Berthold E. Schwarz, "Stella Lansing's Movies: Four Entities and a Possible UFO," *Flying Saucer Review,* Special Issue No. 5, November 1973.

Berthold E. Schwarz, "UFO Contactee Stella Lansing: Possible Medical Implications of Her Motion Picture Experiments," *Journal of the American Society of Psychosomatic Dentistry and Medicine*, Vol. 23, No. 2, 1976.

Bob Shiarella, "The Mysterious 'Skull of Doom'," *Argosy*, Vol. 376, No. 4, April 1973.

Brad Steiger, *Mysteries of Time and Space,* Prentice-Hall, 1974.

Edwin F. Taylor and John Archibald Wheeler, *Spacetime Physics,* W.H. Freeman and Company, 1966.

Yakov P. Terletskii, *Paradoxes in the Theory of Relativity,* Plenum, 1968.

John White and Stanley Krippner, eds., *Future Science,* Anchor, 1977.

C. Louis Wiedemann, "Results of the N.J. 'Spook Light' Study," *Vestigia Newsletter*, Stanhope, N.J., May 1977.

Robert Anton Wilson, *Cosmic Trigger*, And/Or Press, Berkeley, CA, 1977.

V. Ye Savkin, *The Basic Principles of Operational Art and Tactics,* Moscow, 1972, U.S. Air Force translation, U.S. Government Printing Office, Washington, D.C.

Arthur M. Young, *The Reflexive Universe*, Delacorte Press, 1976.

The Tom Bearden Website

CANCER AND THE UNRESOLVED HEALTH ISSUES IN THE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATION

- Text of Booklet
- List of Accompanying Slides (63)

The Tom Bearden Website

CANCER AND THE UNRESOLVED HEALTH ISSUES IN THE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATION

Largely taken from the author's presentation, "Mechanism for Long-Term Cumulative Biological Effects of EM Radiation," 70th Annual Meeting of the Alabama Academy of Science, University of Alabama at Huntsville, 25 March 1993.

T.E. Bearden
Association of Distinguished American Scientists
POB 1472
Huntsville, Alabama 35807

Go to Slides

Abstract

Both postulates in the conventional EM bioeffects model are in error, as is the classical electromagnetics model (CEM) itself. Major errors in CEM are presented. The QM view is taken that all EM phenomena are caused by the potentials, not the force fields. Using the Whittaker/Ziolkowski (WZ) internal structure of the scalar potential, Bohm's quantum potential and an engineerable variant of his hidden variable theory emerge. A strong candidate emerges for the internal mechanisms used for mind, thought, long term memory, and deep cellular control by Popp's master cellular control system. An environmental negative feedback mechanism is stated for long-term, cumulative causative mutation, yielding Sheldrake's morphogenetic field, which is a species quantum potential. Utilizing the new EM bioeffects model, Kaznacheyev's demonstration that any cellular death or disease can be caused electromagnetically is explained. A new definition for cancer is advanced, as is a long-term cumulative mechanism that causes it. The mechanism for Priore's demonstrated cures of terminal tumors in laboratory animals under rigorous scientific protocols is explained. A solution for the major cumulative mechanism for biological effects of EM fields and

radiation is presented. A self-targeting mechanism is presented whereby complex WZ EM biwave pumping of the "cell as a nonlinear phase conjugate mirror material" produces an exact EM antidote signal for the specific cellular disease. It is argued that inexpensive, quick, nondebilitating, cures can be developed for most major dread diseases, including cancer, arteriosclerosis, and AIDS. Priore's remarkable and previously unexplained cures of terminal tumors in lab animals are cited as examples of applying the new model, albeit unwittingly.

The Basic EM Bioeffects Model Is In Error

As is well-known, the field of EM fields and radiation effects on biological systems is actually in something of a shambles. In over 40 years of studies and experiments by careful researchers, the results are inconsistent, contradictory, and usually difficult or sometimes impossible to replicate. Attempted replications often give contradictory results, erratically. Researchers cannot rigorously answer even the simplest question, such as "Does power line radiation contribute to the incidence of cancer and leukemia, and if so, how and under what circumstances?" Causative mechanisms have so far eluded research. Powerful vested interests fund many of the studies. Consequently a great deal of controversy exists in the field. Opinions and positions range from the "microwave oven" position that "if it doesn't appreciably heat tissue, it doesn't harm biological organisms," to the "total fear" position that "any and all non-ionizing radiation is harmful; we just don't know by what mechanism." Legal actions to limit human EM radiation exposure are increasingly being initiated by concerned citizens' groups.

What has actually been "proven" by the nearly fruitless EM bioeffects effort to date is that the fundamental EM bioeffects model being utilized by the researchers is totally inadequate for the *task intended.* If that is true, then 40 more years of continuing to apply the same inadequate model isn't going to make very much additional progress. Further, if the fundamental model is faulted, then what is needed is a detailed examination of *the* model to discover and correct those foundation faults.

In tackling this model foundations problem, I have performed a systems engineering layout of the fundamental EM bioeffects model, and stated its two major postulates. Both postulates turn out to be in error! Immediately one can easily see *why* the field is in such a state of confusion. However, no true systems engineer would stop there. I have also performed a systems engineering layout of precise corrections for the two postulates, thereby correcting and extending the model so that it can yield consistent, replicable results and causative mechanisms. All of this is a rigorous procedure, and straightforward systems engineering methodology.

However, the results point to profound changes, which I have been immersed in deeply studying for quite some time. Out of this approach has emerged something quite fundamental: *a strong candidate for the internal mechanisms used for mind, thought, long term memory, and the deep cellular control system for all cells of the body.* A mechanism for species adaptation also emerges that is quite capable of explaining (and yielding) long-term directed mutation, including the "species jump," where a species such as a bird emerges in short order (not gradually) from a species such as a reptile. The known "species jump" has been a deep mystery to evolution theory, which simply could not explain it. Another thing that has emerged is a startling new *electromagnetic* causative mechanism for cancer, leukemia, etc. At the June 1993 Brain-Mind Symposium in Los Angeles, I will present that exact mechanism. The mechanism is capable of laboratory test, and validation or falsification. We will also present an entirely new approach to EM effects on biological organisms, and present the major new mechanism involved. The new mechanism has fundamental application to the theory of diseases, and offers potential curative mechanisms for dread diseases such as cancer, arteriosclerosis, and AIDS. I explain some of the necessary background in the following paragraphs.

Classical Electromagnetics is Seriously Flawed

As is well-known, there exist severe contradictions between classical electromagnetics (CEM), general relativity (GR), and quantum mechanics (QM). As presently constructed the three disciplines cannot be unified, even by Herculean efforts. Particularly with respect to the primary causative agents for electromagnetic phenomena, the foundations of QM and of CEM are in profound disagreement. CEM assumes *the forcefields* as primary causes, paying only lip service to the potentials and treating them as primarily mathematical conveniences. QM, on *the* other hand, has long since [since 1959] shown that the force fields are simply *effects* in and on the charged particle system, and are not causes. Instead, QM has shown that the *potentials* are the primary causes of all EM phenomena. In charged particle systems, potentials can interfere and cause observable EM phenomena such as the Aharonov-Bohm effect in the complete absence of the force fields. Herein lies *the* reason for the present dire straits of EM bioeffects research. We explain further:

CEM theory was originally formulated by James Clerk Maxwell in terms of quaternions. Note that quaternion algebra has a much higher topology than either vectors or tensors, the mathematics in which CEM is presently expressed. None of the present "Maxwell's equations" _ universally taught as due to Maxwell _ ever appeared in anything by Maxwell himself. Depending upon what you wish to call a fundamental equation, Maxwell's true equations are the numerous *quaternion* equations included in one almost incomprehensible chapter of his 1873 book. The present four "Maxwell's Equations" are in fact due mostly to Oliver Heaviside, and to a lesser extent to Gibbs and Hertz. The major player was Heaviside, a very brilliant but self-educated man who never attended University.

At the time Maxwell's book was published, Heaviside was just teaching himself calculus and differential equations. Seeing the book, he was electrified by it, and Maxwell became his undying hero. Maxwell died not long after, of stomach cancer.

Heaviside believed _ as did almost every scientist at the time _ in the older medieval tradition that *forces* were the causes of all physical effects. He had great difficulty with the potentials, stating that they were *"mystical and should be murdered from the theory."* We know today in modern quantum field theory that forces are effects, not primary causes of anything. In fact, it is the exchange of virtual particles with a mass that generates all forces upon it. As is well-known to foundations physicists today, "force" does not exist without the mass present to be acted upon, and without the action of a potential gradient upon the mass. We know that there are no force fields in the vacuum, and hence potential gradients in the vacuum are not forces, even though they are commonly assumed to be. CEM has not been corrected for these glaring defects: it prescribes force fields in the vacuum, and it prescribes that potential gradients are forces, even in the absence of any mass for the gradients to act upon.

Heaviside also abhorred the quaternion theory. The coupling of a scalar component with three directional components was, in his view, "mixing apples and oranges." He knew that engineers would never master Maxwell's use of quaternion mathematics because of its difficulty. Consequently, Heaviside simply chopped off the scalar component of the quaternion and discarded it, then formulated this new "truncated to a vector" version as a much simpler mathematics, albeit of decreased topology. What he unknowingly threw away was the ability of the quaternion theory to capture an *internal* deterministic, vectorial EM structure of the scalar potential. *[It turns out that he also discarded the unification of EM and gravitation by throwing out the scalar component, but that is outside the scope of my presentation.]* Years later as a lonely recluse in a small upstairs apartment, Heaviside turned back to quaternions to work on a theory of gravitation, according to papers found hidden beneath the floor of his study many years after his death.

Learned journals of the day would not accept Heaviside's papers for publication, because of the assumed "brutality" of his mathematical methods. So Heaviside began publishing very practical papers in a technical magazine, for the time roughly equivalent to *Scientific American* today. These practical papers gave transmission line theory, transformer theory, etc. _ things very useful to the early would-be "engineers" who were struggling with installing telegraph lines, undersea telegraph cables, etc. The vector mathematics utilized by Heaviside was much easier to understand and apply, and his work was eminently practical. Consequently it was eagerly seized upon and applied. The result was that Heaviside's EM model became the ipso facto CEM standard. Note that, at the time, only about 30 or so scientists in the world were truly "learned" in EM _ either from the vector standpoint or from the quaternion standpoint. Further, not much *practical* work was being done by the few quaternion practitioners.

A short "debate" over whether EM should use Maxwell's quaternion model or the Heaviside/Gibbs vector model occurred prior to the turn of the century, mostly in the journal *Nature.* It never involved over a handful of scientists, and it wasn't much of a debate. The vectorists simply threw out the quaternion EM theory and adopted the vector theory of Heaviside and Gibbs. Note that this represented a substantial curtailment of Maxwell's actual theory. *In other words, you can actually do a lot more in and with EM fields and circuits than what now appears in "modern" EM theory, and "modern" EM analysis won't even show it.* Barrett's Oscillator-Shuttle-Circuit analysis of Tesla's actual patented circuits shows this clearly and resoundingly. Also, *living systems utilize the discarded subset of EM for their most vital control functions, and the present theory and* *methodology will not detect or "see " this.* Thus this curtailment alone has resulted in the profound crippling of the conventional EM bioeffects model and efforts to apply it.

The Present CEM is a Flawed Subset of Nature's Electromagnetics

So we have several serious things that are quite wrong with current CEM theory. First, the theory (and its practice) are *artificially* limited to only a subset of the real EM that can be achieved and utilized. The present CEM unwittingly excludes the very type of EM utilized by living systems for their deepest control functions. It also excludes electrogravitation. Second, Maxwell's theory was actually based upon the assumption of a *mechanical (material) ether,* which meant that he could logically assign material forces to the "vacuum" or "ether fluid." Heaviside's translation (and curtailment) to vectors did not charge this "material ether" and "force field in vacuum" assumption.

However, in 1887 the Michelson-Morley experiment resoundingly destroyed the notion that the ether is material. If the ether were truly a thin material fluid, then forces and EM force fields would indeed exist in it. Since the ether is not material and forces exist only in, on, and of matter, no E-field or B-field exists in the nonmaterial vacuum; none ever has, and none ever will. In his three volumes of physics, Feynman pointed out that only the *potential* for the EM force fields exists in the vacuum; the only thing that exists in vacuum is *potentials,* after all! The vacuum is just a fantastic collection of interfering potentials and potential gradients. Rigorously, the E- and B-fields we detect with our instruments *exist only in the electron gas in the probes we utilize.* This is well-known to a few foundations physicists, but the CEM model has never been corrected for these foundations errors.

Note that the loss of the material ether also falsifies one of the three key assumptions in modern potential theory: the notion that the gradient of the potential comprises a force field. We know today that the CEM equation $E = -\nabla\phi$, is actually incorrect *for the vacuum.* It is correctly *measured;* since that is exactly what is detected when the vacuum $\nabla\phi$ potential gradient couples to the free electrons in our detecting/measuring instruments. What we detect as E, however, is actually $[(\nabla\phi \bullet (e^-)]$, or the potential gradient coupled to the electrons, including their masses. Specifically, we do not detect the nonmaterial (VO that actually exists in the vacuum. We detect electron wiggles in the free electron gas in the conductors of our instruments; we do not detect "vacuum wiggles" per se.

Neither Maxwell's quaternion EM equations nor Heaviside's vector EM equations include electron spin effects, but actually model electricity as a "thin fluid." Hence the EM equations of either model are *fluid dynamic* equations. Specifically, the measured transverse EM waves exist in the electron gas of our detectors; any detector detects only its own internal change, *not* the external agent that interacted and caused that internal change. White the "signal" races down the conductors in our probes and sensors at essentially the speed of light, the "free" electrons in the electron gas in those conductors are *longitudinally* restrained. The electrons must essentially move laterally from their initial distribution inside the conductor to its skin, then "slip" slightly down the wire on the skin at only the electron drift velocity, which is a fraction of a centimeter per second. The spinning electrons in the electron gas, being longitudinally restrained, act as *gyroscopes.* When disturbed by an interacting force, they precess. It follows that the lateral motion of the electrons inside the conductors of our detectors is just this "electron precession." It also follows that the direction of the disturbing force must be at right angles to the electrons' measured transverse precession movement. Rigorously, EM waves in the vacuum thus are *longitudinal,* not transverse. We *measure* transverse waves in our detectors and instruments because we are detecting and measuring the *electron precession waves in the free electron gas in the conductors of those instruments.* But as Tesla pointed out, there are no Hertz waves in the vacuum; instead, EM waves in the vacuum are longitudinal "sound" waves, or *waves of rarefaction and compression of the medium.* From modern QM we also know what the medium is: it is a flux of virtual particles. Again, as can be seen, the CEM model is seriously in error in its representation of EM waves in the vacuum.

We will not pursue this further; it is well-known to a few physicists in foundations work (but not to most electrical engineers and electrical physicists!) that classical EM theory is seriously flawed, and that it should be upgraded to correct those flaws. As is well-known in QM since 1959, it is the potentials that are the actual causes of all EM phenomena, and the potentials can interfere to cause real EM effects in charged particle systems, even in the absence of the force fields. The Aharonov-Bohm effect is an example, as are several other derivative effects well-established in the literature. What has been ignored even in QM, however, is the *organized internal structure* of the potential, even though the original (vector) discovery by Stoney occurred in 1897 and was extended by Whittaker in 1903.

This internal structure of the scalar potential had been implicitly present in Maxwell's 1873 quaternion theory, as an integral part of the scalar component of the quaternion resultant of the intera**ct**ion between two or more quaternions. In quaternions, a scalar entity may be regarded as a special quaternion entity whose translation has been reduced to zero. A vector entity may be regarded as a special quaternion entity all aspects of which are translating, and no nontranslating aspect is present. Thus the scalar quaternion entity may be totally composed of vector components, so long as they sum or multiply to a zero translation resultant but possess a finite nontranslating magnitude.

The Scalar Potential Has an Internal Bidirectional Wave-Pair Structure

In a profound but ignored 1903 paper of enormous consequences, E. T. Whittaker decomposed the scalar EM potential into a harmonic set of bidirectional wave pairs, extending the original 1897 work of Stoney. Each wave pair consists of a wave and its phase conjugate replica (its "anti-wave" or "time-reversed twin".). Thus the scalar potential has a rich internal biwave structure. [Further, one can *make* a scalar potential by simply assembling the necessary multiple waves, and one can alter the internal structure of the potential at will by altering the waves utilized in the assembly process.] In 1904 Whittaker published a formidable second paper, showing that all of CEM could readily be replaced by scalar potential *(hidden multi-wave)* interferometry. In other words, scalar EM is far more basic and extensive than is the present CEM with its emphasis upon the force fields.

E.g., you can alter the internal structure of the Schroedinger potential, and place deterministic biwave "hidden variables" inside. You can then "engineer" these hidden variables _ and consequently the Schroedinger potential _ by external means. In other words, you can even accomplish *at least limited engineering of quantum change itself.* This of course reduces the Gibbs statistics (which assumes a totally random quantum change) to a special case, and provides a more general case of chaotic quantum change, which is still statistical but may contain hidden order. Note that this statement is experimentally testable. It also resolves the greatest problem in quantum mechanics today: the problem of the missing chaos (hidden order).

In 1989 Ignatovich placed a paper in the *American Journal of Physics* pointing out a similar internal biwave structure of the Schroedinger potential, but none of our scientists seems to have realized the profound implications. *If we apply these proven mathematical extensions of electromagnetics, we are now dealing with a higher topology hidden variable theory, along the lines shown by Bohm, but one that is engineerable on the lab bench.*

Ziolkowski's Extension Allows Hyperspatial Communication

In 1985, a brilliant EM scientist named Ziolkowski independently rediscovered Whittaker's internal biwave harmonic decomposition of the scalar potential. He also extended the "internal structure" theory, to encompass not only the sum set but also the product set. Since the product of waves represents modulation, Ziolkowski's brilliant work provides the setting for *direct information communication capability* in the internal structuring of the scalar EM potential.

Further, this internal channel is in the virtual state, so one is accomplishing virtual state engineering, in the sense of the vacuum engineering posed by Nobelian Lee. If one uses a 4-space Minkowski model, the Whittaker/Ziolkowski channel information may be viewed as *subspace* communications. If one utilizes an n-space Kaluza-Klein model, where n is greater than 4, then the Whittaker/Ziolkowski channel information may be viewed as *hyperspace* communications. *Living systems already use this inner EM channel.*

Living Systems Use the Internal Channel Communication

It turns out that living systems utilize precisely the Whittaker/ Ziolkowski (W/Z) mechanism for their deepest functions and control systems _ including mind, thought, long-term memory, and the master cellular control system (MCCS) discovered by Dr. Popp of Germany. What we are stating is that, for the very first time, *we now* *know where and how the mind's "deepest software" is, its fundamental mechanism, and how to go about programming it directly.* However, we first must develop new measurement instruments.

Unfortunately, present EM instruments are almost all only "electron-wiggle" detectors; they detect only the *translation of electrons. A gradient* in the potential will couple to electrons and translate them. The *gradient-free* potential does not translate electrons, and it is the internal biwave W/Z structure of that gradient-free potential that we need to measure. Presently no known instrumentation will do that.

However, if two different scalar potentials are interfered [Whittaker 1904], their interference reproduces potential gradients [normal "force field" EM], which *do couple to electrons and translate them.* In fact, potential gradients (which in CEM are erroneously called "force fields") were already shown by Whittaker to be entirely due to the interferometry of two scalar potentials. So new "scalar interferometry" instruments must be developed to "outfold" the internal contents of the scalar potential as gradients, and then utilize normal instrumentation to measure those interference gradients and calculate the potential's internal structuring. One form for the detector is the use of a standard potential (with a known internal biwave structure) inside a Faraday-shielded chamber, so that the test gradient-free potential (which penetrates such a cage) interferes with the standard potential inside the cage to produce gradients therein. *A* probe in the interference zone detects the gradients as electron translations, which are conducted externally to amplifiers, meters, spectrum analyzers, computer programs, etc. The end result is the determination of the internal WZ structure of the test potential. Other detector types are possible; this is just a "straightforward" type.

Thoughts, Mind, Memory, and Cellular Control Are Measurable

What we are saying is that, by developing the proper instrumentation, it is possible to directly detect and examine thoughts, memory, and deep control system functioning of the biological organism, including the deep internalistics of its personal quantum potential. We are speaking of the pending emergence of an experimental science, not just a speculation.

However, we are dealing now with a form of "hidden variable" theory, but one that is engineerable and testable. So we will become directly involved with Bohm's *quantum potential,* both experimentally and theoretically. *A* quantum potential can connect widely separated systems by instantaneous effects, as if the systems were not separated but were located together as a single system. Further, the quantum potential does not have a single localized source. Suffice it to say that in 1991 I published the discovery of how a quantum potential is actually created. If one utilizes the proven "self-targeting" or iterative phase conjugate shooting mechanism from the Strategic Defense Initiative, and applies it to the WZ waves internal to the scalar EM potentials from two separated charges, then in a dense signal environment the self-targeting interaction may be initiated, so that the separate potentials merge into a single nonlocalized or "spread-out" potential consisting of laser-like beams between the charged particles comprising the separated systems involved. In that case, part or all of any EM change in one system is immediately experienced in the other systems participating in the effect. The communication is superluminal. The limitation to luminal velocity applies only to the "surface gradient" of a potential, not necessarily to the bidirectional waves in its WZ internal structure. Those internal waves are in either hyperspace or subspace, depending upon whether one models the situation in more than four dimensions or in only four.

The point is, *the quantum potential also has a W/Z internal structure.* And that structure can be deliberately created and engineered by external means. Action at a distance is not only possible but engineerable.

Living Systems and Quantum Potentials

It turns out that the living organism utilizes a quantum potential _ a *special scalar potential connecting all its internal atomic nuclei* _ for its volition, deep control, and mind and thought processes and operations. This can even be taken as a flat definition of a living system. A living system utilizes internal WZ structuring of this scalar potential connecting all the atoms (nuclei) of its body mass, for its deepest control processes. Any system not having or doing this is not a living system. This also resolves the age-old philosophical question of how mental *intent* cm cause physical response, but that is beyond the scope of my presentation. It also explains why viruses can be precipitated out of solution as a crystal, and the crystal stored for decades. Then when the crystal is redissolved, the individual viruses separate and resume their "living" state. Suspended animation of the viruses results when their physical matter and its quantum potential remain intact, and the stored WZ structures in the quantum potentials on the atomic nuclei of that matter remain the same.

It further turns out that the species itself has a much weaker (at a deeper level topologically) quantum potential connecting all its members (all the atomic nuclei in their bodies). This corresponds to Sheldrake's morphogenetic field _ that field is just a *species quantum potential.* Jung's collective unconscious, e.g., can be directly expressed in this model scientifically and testably _ but again that is beyond the scope of this presentation.

Newton's Third Law Requires Forces to Occur in Oppositive Pairs

Newton's third law is a sleeping tiger, with surprising implications when awakened and explored. For example, it rigorously requires that forces (including EM forces) must occur in oppositive pairs, This alone falsifies the CEM notion that *oscillating "singular-force"* electric and magnetic waves appear in either a material ether or physical matter. At least two waves must exist, the wave and its oppositive or *antiwave.* Further, the wave and antiwave must be intimately coupled. Heretofore the antiwave has simply been ignored, or when anyone pointed out that it was present in our detectors, it was just considered to be "Newton's third law reaction wave." "Newton's third law" was just mysteriously invoked, as if it had no causative mechanism. Yet in quantum field theory the cause of all forces is the absorption or emission of *virtual particles.* The cause of all mechanical and electromagnetic forces is the absorption or emission of virtual *photons. So* Newton's third law for mechanical and electromagnetic reactions must also be due to the exchange of virtual photons. In other words, there was an extra half of the "something" in the vacuum that interacted with our detectors and gave us the "free electron gas's transverse precession waves. That "extra half' was exactly equal and opposite, and interacted in the atomic nuclei of the conductors and the mass of the instrument to give a set of physical Newtonian recoils. That half is equal and opposite to the half we actually recognized. The point is, equal and opposite forces actually interacted with our instrument. Thus a stress wave interacted with it, not a unitary oscillating wave.

What occurs in vacuum must be a stress wave, not a unitary force field wave. In short, as stated (but not elaborated) by Feynman, it is a *potential* wave, or an oscillation in the potential gradients and magnitudes, not the force fields. Since in modern theory force is an *effect* and not a cause, we interpret and extend Newton's third law to state that the *causes* of all forces must occur in oppositive pairs also. Thus the cause of an EM force field must be the interaction of two scalar potentials _ just as Whittaker's 1904 paper proves.

In quantum field theory any mechanical or EM force is caused on a mass by absorption and emission of virtual photons. Accordingly, let us examine the smallest possible EM or mechanical "force" _ that one caused by the absorption of a single virtual photon. Newton's third law requires that two photons (the causes) must occur, not one. Further, in all cases, the one must be the exact antiparallel to the other. The only thing always meeting that condition for a photon is its exact antiphoton _ its *time-reversed, phase conjugate replica twin.* But a phase conjugate *replica* photon must also superpose spatially with its parent photon _ that's what phase conjugate replicas normally *do,* according to the distortion correction theorem of nonlinear phase conjugate optics. Thus instead of the conventional "single photon" interaction, the actual vacuum entity engaged in the interaction with an atom is a *coupled photon/antiphoton pair.* A coupled photon/antiphoton pair has spin-2, hence is a *graviton.*

Gravitons and Graviton Interaction

So arguably "photon interaction" has been *graviton* interaction all along. In the interaction with an atom, the graviton splits into a photon and an antiphoton. The photon usually interacts with the electron shells, being absorbed by an electron to raise its energy level, then re-emitted and scattered outward from the atom. This absorption and scattering of photons _ *containing their components of energy and time* _ from the electron shells of the atom creates movement of the atom through external observer time. The photon interaction between a mass and its environment creates the forward flow of time for that atom and its seemingly entropic external (photon interaction created) universe.

The antiphoton half of the graviton is time-reversed, which we see as *spatially* reversed. Consequently we see it move in the opposite direction to its externally-directed photon twin. So the antiphoton focuses inward and interacts with the atomic nucleus, providing the Newtonian third law recoil of the atom. The blithe assumption of an "automatic" third law reaction force has always concealed the fact that the so-called *photon* interaction is actually a *graviton* interaction. Further, the positive charge of the atomic nucleus is due to the time reversal interactions of antiphotons with the nucleons, which reverses the charge from that of the negatively charged electrons in the atom's electron shells.

Proof: Excising the Antiphoton Violates Newton's Third Law

This graviton interaction hypothesis is testable. Note that, if the hypothesis is correct, then when an atom or mass emits antiphotons or antiwaves (phase conjugate photons or phase conjugate replica waves), it means that the inward-burrowing antiphotons have been "tricked" into coming out of the atom instead, so they can comprise the time-reversed wave. If they come outward instead of going inward when they split from the parent gravitons, they do not interact with the nuclei to cause their recoil. In that case, there would be no recoil. And indeed, so it *is. In a phase conjugate mirror material, the mirror does not recoil when it emits a phase conjugate replica wave, even a powerful one due to powerful pumping.* This is not true when the mirror material emits a pseudo-conjugate wave. A pseudo-conjugate mirror, e.g., emits an ordinary "time-forward" wave with a distorted wave front, so that it will "retroreflect" in a manner similar to the phase conjugate replica wave. However, the pseudo-conjugate wave is not a true time-reversed wave. Consequently, the mirror will recoil when it emits the pseudo-conjugate wave, because the antiphotons from the graviton interaction still penetrate to the nucleus and interact there, while the time-forward photons constitute the pseudo-conjugate wave actually emitted. This is experimentally verifiable. Among other things, it establishes that there exist differences between the photon and the antiphoton.

Looking at the photon aspects of this, one realizes that one must be very careful in applying the conventional assumption that the photon and the antiphoton are identically the same. Actually they are not the same "internally." The photon can be thought of as carrying or

consisting of (+□E)(+□t), while the antiphoton can be thought of as carrying or consisting of
(-□E)(-□t). Thus both have positive spin, but differ in their internal components. The lack of recoil in a phase conjugate mirror that emits a true time-reversed replica wave is already in the literature, though derived by statistical quantum mechanical arguments. It has also been pointed out by other physicists that all measurement/ detection is actually binary, but that the "internal energy" half is almost always ignored [it's just considered to be "Newton's law", automatically revoked, and swept away in that euphemism.].

Graviton Lattice Structure of the Scalar Potential

When we examine the WZ structure of a scalar potential, in each biwave the wave and antiwave are phase conjugates. It follows that the photons in the wave and the antiphotons in the antiwave are phase conjugates also. This means that, as the two waves flow through each other spatially, photon/antiphoton pairs continually couple and uncouple. Hence gravitons continually couple and uncouple. Further, since the wave pair frequencies are phaselocked between pairs, *the scalar potential is a phaselocked lattice of statistical (continually forming and unforming) gravitons.*

Let us now consider the interactions of an organism with environmental photons as the interactions of the organism with environmental *gravitons*, and specifically with graviton lattices.

Cumulative Buildup in the Quantum Potential of the Antisignal

If the *photon half (of the graviton interactions)* in the electron shells of the living body are considered as environment signals/ interactions, then precise *antisignals* or anti-interactions, constituting perfect negative feedback from all of an organism's physical environmental experiences, are experienced in the organism's atomic nuclei. Note that there they have partially affected the internal structuring of the quantum potentials _ both the personal quantum potential and the species quantum potential. The point is, *cumulative negative feedbacks (exact phase conjugates) for all the environmental experiences and stresses an organism experiences, build in both the personal QP and the species QP of* *that organism.* A part of the entire species' cumulation of species-common antisignals also exists deep in the quantum potential of the living biological system.

So for any protracted stress experienced by the organism, in its personal QP a coherent cumulative negative countermeasure feedback _ one which by its time-reversed nature would reduce that stress _ continually builds. The phase conjugates of the "well-rounded" or highly varying stresses of multiple types will tend to mostly balance or "zero out." However, a sustained stress of one kind will result in coherent cumulation and increase in the "signal-to-noise ratio" of the amplitude of the antisignal for that particular stress, compared to the amplitude of the average of all the antisignals experienced. This is similar to the case where, in a sea of radar noise, a coherent radar signal were continually present and integrating coherently. Continual coherent integration, of course, results in continual increase of the signal-to-noise ratio for the coherent signal, and eventually it emerges from the "sea of noise" as a discernible [in this case, observable and physical] antisignal.

This "cumulative kindling of antisignals" occurs in both the personal QP and the species QP; however, the species QP is many, many orders of magnitude less than the personal QP. Hence much greater time is required for "kindling" of species changes (evolution) than for kindling individual countering adaptations.

Exercising to get in shape is a simple example of physically stressing the body cells so that the high-level feedback mechanism in the MCCS will kindle appropriate antisignals and order the cells to adjust their functioning in a fashion such that the level of performance being called for can be accomplished more easily, thus reducing the stress level. The adaptive mechanism ordering the adjustment of the body cells is the coherent negative feedback in the personal QP, created from the stress signals from the sustained "workout" exercises. This is a fairly rapid, "high-level" mechanism; many other much slower (and some much faster) feedback reaction mechanisms also exist.

Hypoxia As a Result of Interference With Hydrogen Bonding to Hemoglobin

We now point out the recently discovered water H-bonding effect on the hemoglobin in the blood, where the H-bonding activity increases the ability of the hemoglobin to carry oxygen. Some 60 or 70 water molecules surround the hemoglobin molecule, engaging in extensive H-bonding interactions with it, which substantially increases the hemoglobin's purely chemical ability to bind and transport oxygen. Importantly, contamination of this water in the blood fluid drastically affects this H-bonding benefit, and reduces the hemoglobin's ability to carry the extra oxygen. Thus contamination of the internal blood fluid by chemicals, agricultural pesticides, smoke, etc. directly and dramatically lower the oxygen availability to the body's cells, resulting in a continuing "oxygen hungry" state (hypoxia) in most of the body's cells.

It also is now known that H-bond structuring of water is highly dynamic and constantly adaptable, and it possesses a high degree of internal order. A conglomerate of H-bond structures in an area or a volume can be considered an *H-bond potential,* since the gradient oppositions sum to produce stress potentials. Thus in the H-bonding fluid surrounding the hemoglobin molecule, there exists an H-bond scalar potential, and by Whittaker/Ziolkowski (W/Z) decomposition it has a multi-wave structure. Any and all contamination of the fluid _ including by electromagnetic fields and signals, chemicals, etc. _ alters the H-bonding structuring and hence the internal W/Z structure of the H-bond potential. *The end result of adulteration of the H-bonding structure of the blood fluid is a condition of hypoxia induced in the cells of the body due to reduced oxygen transport by the hemoglobin of the red blood cells.*

Any single EM signal is the result of the interference of two scalar EM potentials, as shown by Whittaker. It is the result of the interference of the *internal EM wave structures* of those scalar EM potentials. Multi-signal EM radiation must be regarded as interference of multiple pairs of scalar potentials. These interfering potentials penetrate to the atomic nucleus in the body, and thus interact with the internal personal quantum potential utilized by the body in its deep cellular control. Also this multi-signal EM radiation, *no matter how weak,* directly interacts with the H-bonding structure of the H-bonding blood fluid. Even single-signal EM radiation still forms oppositive pairings with the weak fields already existing in the body, thus also forming W/Z structured potentials. From the interferences in the body of multiple such potentials and their internal structures, there also are created gradient (force field) interactions upon the blood cells and the hemoglobin. In short, low-level background EM radiation can also interact with the H-bonding of the hemoglobin to seriously lower its oxygen transport capability, just as any other contaminant. The denser the external environment, the greater is the interferometry and the interference with the hydrogen-bonding augmentation of oxygen transport. It follows that, when thrust into a truly dense signal environment, some exposed individuals may receive a "cumulative H-bonding interference dosage" that, when added to their existing prior dosage, is sufficient to result in physical symptoms "ordered" into the cells by the antisignals kindled in their quantum potentials. In the recent Gulf War, many Americans were suddenly thrust into one of the densest EM signals environment in history. Shortly after returning from that ultrashort conflict, many of these exposed veterans have experienced delayed health changes presently known as "Gulf War Syndrome." While the military has attempted to portray this syndrome as "stress-induced" emotional' trauma, most of these veterans when tested are found to be emotionally stable. Further, there was much less combat stress on our troops in the Gulf war than in WWII, the Korean War, or the Viet Nam War, as shown by less than 200 combat deaths! Essentially the war was almost a "shooting gallery," and so combat stress is simply not viable as a proposed causative agent for the Gulf War Syndrome.

Chemical, Mechanical, and Electrical Interactions are Electromagnetic Anyway

Indeed, chemistry is largely due to electric charge and charge distribution anyway. So if we view the physical contaminants in the blood view *chemically,* they are caused by electromagnetic means at root basis. Thus even the "chemical" contaminants interact electrically and via potential (multiwave) interferometry with the H-bonding potential. Further, in quantum field theory, all mechanical forces are due to the exchange of virtual photons, and hence also are *electromagnetic at their very basis.* The bottom line is that all chemical, mechanical, and electrical interactions are in fact electromagnetically caused. Further, since both the internal and external structures of the potential are engineerable, the mechanical and electrical forces on the mass particles _ and the virtual photon exchanges causing these forces _ are directly engineerable. *All of these chemical, mechanical, and electrical interactions in the cells can be affect*ed *and even engineered electrically. This is true from the atomic nucleus, to molecules, to material lattices, to human cells, to tissues, and even to the mind and long-term storage templates (internal WZ structured forms in the QP) in the biological organism.*

As cam be seen there are many other weaker, direct interactions of "force-field" radiation; however, here we are interested in the interaction of that radiation with the inner structure of the H-bond potential'., to reduce the oxygen transport capability of the hemoglobin. *In any case, the total sustained interaction of chemical and physical contaminants and EM fields and radiation can result in a sustained oxygen-deprivation (hypoxia) condition for the body cells, even to a dramatic degree.* Further, by graviton interaction, *a sustained, cumulated set of antisignals is also automatically generated in the personal quantum potential as a set of negative feedback signals to take corrective actions to alleviate the oxygen depletion condition.* Note that the cumulating antisignal condition is general and affects the immune system and its cells as well.

The Example of Smoking

Obviously the tars, particles, and nicotine of the smoke inhaled by the smoker dramatically contaminate the fluid in the smoker's lungs, in the very place where the hemoglobin of the red cells is taking on oxygen. Drastic interference with the H-bond stimulus of the hemoglobin occurs. The result is an immediate and dramatic reduction of the oxygen-transport capability of all the hemoglobin in the red cells as they pass through the contaminated lungs and are exposed to fluid contamination. An immediate oxygen-deprivation condition thus is created in the body. Due to the magnitude of the stressing signal, the major portion of the antisignal is also of substantial magnitude _ hence immediate counteraction orders accumulate in the personal QP, specifically in Popp's master cellular control system. So very rapidly the body has negative feedback countersignals from the central cellular control system, ordering the body to take actions to reduce the cellular need for oxygen. The metabolism is lowered, the body relaxes, and the appetite is suppressed; all counters to the cells' hypoxia condition by lowering the cellular requirement for oxygen.

In spite of heroic "high level antisignal" physical cellular compensations to reduce the use of oxygen, however, the condition of cellular oxygen shortage may still continue, in the case of the smoker or substantial environmental contaminants such as secondary smoke. So in addition to the prompt countersignals, additional much-weaker countersignals are also slowly accumulating in the personal quantum potential's central cellular control system.

Cumulative, Deep, Long-Term Countering Signals

We now need one additional bit of information. In addition to its personal quantum potential, a biological organism also is connected to a species quantum potential joining all the members of its species. This is a much weaker quantum potential; nonetheless, potentials superpose, so one small "part" of the personal quantum potential is actually the species quantum potential _ in other words, Sheldrake's morphogenetic field. Deep within its own personal quantum potential the living system also possesses all the cumulatively ordered steps (species antisignals that were implemented) of its species evolution. With very *long coherent cumulation, countersignals can be stimulated in the individual bio-organism's quantum potential _ and in its deep cellular control _ that are counters to the original signals that ordered directed mutation and the primeval evolution of the constituent cells themselves.*

In a sustained cellular hypoxia stress environment, the "do whatever is necessary to reduce oxygen usage" countersignals will continue to cumulate from successively deeper levels in the QP over an appreciable time. These continue to slowly "kindle" after the "first high-level antisignal actions" that cumulated past the "noise" (quantum) threshold reached their limits of physical reactions and are still insufficient to resolve the problem. As the hypoxia stress in the cells continues, then much weaker but coherent countersignals that trigger deeper cellular adaptation actions eventually have time to cumulate past the quantum threshold to the observable state. These long-term cumulated antisignals come from much deeper, even from the species quantum potential itself.

These cumulating deep countersignals are phase conjugates _ *reversals, or dedifferentiation commands* _ of the previous directed mutation signals that ordered evolutionary changes in their predecessor cells in the far distant past. *These long-term antisignals thus contain a signal to reverse the original cumulative countersignals in the cellular species potential that caused the cells early ancestral anaerobic cells on earth to change (differentiate) into largely aerobic cells, after the earth's atmosphere acquired significant amounts of oxygen. In short, a deep signal is slowly* *cumulating that, when it emerges, will order the dedifferentiation of the affected cell back down the evolutionary road toward the anaerobe. The first steps are reductions of centralized cellular control, including growth control.*

In other words, under sustained cellular hypoxia a countermeasures signal to the cells, ordering them to dedifferentiate back toward the anaerobic state, is slowly building in the personal quantum potential from a much weaker, lower level "under the common countersignals noise level" all the while. If significant oxygen-deprivation continues for a sufficiently long time, this deep countersignal eventually emerges in Popp's MCCS system in the personal QP, *ordering the affected cells to dedifferentiate back toward their anaerobic state.* The first step back is to break away from centralized control by the higher organism _ which results in individual "tumerous" or "uncontrolled" cells. *From sustained cellular hypoxia there exists a slowly increasing precancerous state, even years before the outbreak of physical cancer.* This precancerous state has great influence upon the "single cells" of the organism, which includes the cells of the immune system. Thus the precancerous state is characterized by increasing errors in, and slow weakening of, the immune system. Arthritis and similar debilitating immune system disorders appear to be largely the result of this "precancerous" long-term cumulation.

Note that the most stressed cells in the body automatically provided the greatest magnitude of "negative feedback input." Notice also that the feedback is a phase conjugate replica; it therefore "backtracks" its initiating stress input signal. In short, the degree of feedback antisignal received by a cell or group of cells varies according to the degree of stress they experienced. Thus those cells sustaining the longest and greatest stress get this drastic dedifferentiation countersignal first. E.g., this accounts for the (usual) localization of the resultant tumor in the lungs of a smoker, at least initially. It also accounts for the stress-damage mechanism for such disorders as arthritis.

A New Definition of Cancer

In the new model being advanced, *cancer and leukemia are centrally-commanded. final, desperate. "first-step dedifferentiation" adaptive attempts by the stressed. affected cells experiencing sustained oxygen shortage (hypoxia) to reverse their cellular* *evolution and return to the anaerobic stage of their distant ancestry.*

The cause of cancer and leukemia _ and indeed of all diseases _ is *electromagnetic* in nature, and it can be straightforwardly treated and cured *electromagnetically,* if one utilizes the extended W/Z aspects of the higher topology EM actually written by James Clerk Maxwell. This statement is based on experimental proof; it is not conjecture. We discuss that proof below.

As we saw, in general the strength of the cumulating countersignal inversely depends on the distance in the body of the affected cells from the source of greatest contaminant stimulation. As the deeper "dedifferentiation" countersignal reaches the quantum threshold in the most affected area, a small group of localized cells located there take the first step back along the "single-anaerobe to single aerobe to multiple aerobe" evolutionary chain. That first dedifferentiation step breaks the cells away from centralized control of the MCCS in the personal quantum potential. Those dedifferentiated cells become independent cells or a small independent cell-groupings. They constitute a *tumor,* or *leukemia* if occurring in the blood cells. They simply are no longer under the control of the body's MCCS in the personal quantum potential, even though the body's logistical services (nutrients, oxygen, etc.) continue to be furnished to them.

At this point the tumor becomes an independent cellular organism, living in an environment where its body host continues to furnish its food and oxygen, and its host's immune system cells do not recognize the altered cells as foreigners to be attacked and destroyed. Consequently there is no "large organism" central control of the tumor's growth, and the tumor cells divide and multiply apace. This loss of centralized growth control and the resulting unchecked cellular division and replication is in fact recognized by scientists as the major distinguishing characteristic of cancer and leukemia. With central control lost, groups of cancer cells or individual cancer cells may break away and travel to other locations in the body via body circulation systems. This results in metastasis, the spread of the cancer to other sites in the body, where they continue to be recognized as "self" by the immune system cells and continue to be supplied with oxygen and nutrients for continued growth.

The Tumor Has Become a Parasite Organization

Think of it this way: In a living body, each cell is already a single, independent, living creature all its own. It even has its own small "personal quantum potential" _ we have argued that that is an *a priori* condition of being alive. However, each cell in a multicellular organism normally is under a centralized electromagnetic control system (which functions in the organism's higher-level personal quantum potential), so that the organism lives and functions as an overall higher-level unit. If a cell (or group of cells) has separated from this centralized EM control, but is still living and functioning, then obviously the cell is no longer under the whole organism's MCCS and personal quantum potential. However, it still possesses its own control system, and its own personal QP. If the breakaway is by a small group of cells, then they are loosely under a "small group" quantum potential and MCCS as well. That is, they have gone from "large-scale central control" to a "much lower" level of centralized cellular control, accounting for the "tumor as an individual multicellular entity." Nourished by the host body, the new parasitic organism _ the tumor _ grows at an unchecked rate. Again, we argue that *the problem is electromagnetic in nature, and it can be "fixed" electromagnetically.*

The Cumulative Pre-Cancer State

With this picture of the long-term cumulative causative mechanism for cancer and leukemia, a far better picture of the pre-cancer state has emerged. The pre-cancer state is a hidden EM change of state comprised of a *cellular dedifferentiation order,* back down the species' cellular evolution trail, and the magnitude of that change of state is slowly increasing inside the organism's quantum potential. The eventual observable cancer is an actual cellular dedifferentiation due to the localized breaching of the quantum threshold by this cumulating hidden EM dedifferentiation order to move away from centralized control and back toward the anaerobic cellular state.

Becker's Profound Experiments

It has already been experimentally shown that very minute amounts of direct electrical currents, e.g., can cause cellular dedifferentiation and redifferentiation. In breathtaking work of Nobelian quality, Becker has proven that even picoamperes of localized electrical current are sufficient to cause cellular dedifferentiation and redifferentiation, even repetitively. Incredibly, he proved that this was true for red blood cells, which first dedifferentiated to more primitive cells, the redifferentiated to precursors of cartilage cells, then redifferentiated to bone cells in a "bone fracture" area and healed the bone fracture. Since picoamperes can be generated by the cumulative countersignal mechanism (via Whittaker scalar interferometry), directly affecting the MCCS and the cells as well, then the connection between scalar potentials/potential interferometry and *dedifferentiation* mechanisms assumed to be involved in cancer and leukemia essentially follows from interpretation of Becker's pioneering work. To further strengthen the assumption, Hsue has recently (1993) shown that a DC voltage is in fact equivalent to two bidirectional EM traveling waves _ which directly indicates a WZ type implication in Becker's profound work itself. Thus Becker actually utilized a hidden, bidirectional WZ wave structure inside the voltage gradient between his electrodes in his DC current dedifferentiation and redifferentiation of blood cells. He conclusively demonstrated that cellular dedifferentiation and redifferentiation _ both of them *genetic* changes _ can in fact be caused by such an internal biwave EM structure of a steady, persistent scalar potential gradient. Further, he demonstrated that *these El~-induced cellular genetic changes can be self-targeted toward reversal of the specific damage condition, even though several successive genetic changes may be necessary m reach the ultimate cellular genetic form required for healing.*

Pre-Cancer State and Prognosis For Treatment Effectiveness

To summarize, whenever a cancer or leukemia detectably exists, regardless of the cancer site, a pre-cancer state also exists in the remainder of the body, and the immune system is affected generally. Further, this pre-cancer state pre-existed the actual emergence of the tumor itself. The new EM bioeffects model also sheds light on the probable effectiveness of conventional treatment, and probability for a "cure" that lasts for the rest of the patient's life.

If the cumulative countersignal slowly kindling in the remaining pre-cancer state is sufficiently below the quantum threshold, then conventional treatment of the cancer _ such as excision _ can eliminate the tumor, and no other tumors may then develop before the normal physical death of the body, even though slow increase of the pre-cancer state continues. In other words, in that case the cumulative antisignals in the pre-cancer state locations other than the original cancer site never reach the quantum threshold prior to the natural death of the patient. On the other hand, if the cumulating antisignals in the pre-cancer state generally are very close to the quantum threshold, then recurrence of the tumor is highly likely. In the first case the curative prognosis is excellent; in the second case it is poor.

Also, we point out that many present treatments (such as nuclear radiation) themselves cause significant stress on the body and its cells and further deterioration of the immune system. They also generate an additional "antisignals" level in the personal quantum potential, resulting in an actual *increase* in the level of the pre-cancer state in the organism. If the combination of the pre-cancer state at time of treatment, plus the increase in the pre-cancer state due to treatment, is sufficient to breach the quantum level, then that treatment will prove to be of little or no avail to the patient. The tumorous condition will recur, and sometimes massively (the tumor metastasizes), regardless of whether further treatment is given or not. In that case the patient will die, unless somehow the pre-cancer state is lowered, the tumor cells are reverted back to centralized cellular control, and the immune system is restored to effective functioning so that excess cells are eliminated. Presently there is no conventional treatment that is capable of accomplishing the three requirements for a desperately ill patient's recovery from the "terminal" cancer condition.

Seeking the Complete Cancer Cure

From the new EM definition of cancer and leukemia and the new cumulative causative mechanism, one can see just how severely limited are the present medical weapons against this dread disease. It follows that, *if the cause is totally electromagnetic, then a totally effective treatment can only be sought as an electromagnetic therapy. Since the EM cause is in an extended EM domain, then axiomatically a totally effective treatment can only be sought in the same extended EM domain.*

The complete cure for cancer and leukemia, of course, would be to completely neutralize or "zero-out" the patient's cumulative Whittaker/Ziolkowski countersignal buildup in the MCCS/personal

QP. This can be done *prior to* the cumulation reaching the overt physical disease stage, or *after* it has already reached that stage. It can readily be done electromagnetically, using the multi-biwave tailored EM and self-targeting to provide an exact counter-countersignal (or antidote signal) to the cumulation signal itself. *This is a scientific fact and it has been rigorously proven, although the methodology could not heretofore be technically explained and understood.*

Priore's Cure of Terminal Tumors and Other Diseases

Such a total cancer-curative procedure was repeatedly demonstrated in live animal experiments in the late 1960s and early 1970s by Antoine Priore in France, under rigorous scientific protocols and supervision by eminent French scientists [such as Robert Courrier, head of the Biology Section of the French Academy of Science, and *Secretaire Perpetuel* of the Academy at the time]. The advantage of the Priore type approach was that it rapidly reversed the dedifferentiation away from centralized control of the affected (tumor) cells. The tumorous cells simply redifferentiated back to normal cells under centralized control of the MCCS/personal QP. (If too numerous because of the tumor's growth prior to its redifferentiation, normal body mechanisms absorb and dispose of the excess cells). Further, such treatment removed *all the pre-cancerous state from the rest of the body, preventing recurrence or spread of the tumor*. It is the only "total cure" for cancer, including the cumulated pre-cancer state, that has ever been scientifically demonstrated. And lastly, it accomplished these results without severe trauma to the body or further trauma to the immune system. Indeed, the exact opposite was the case.

Working with Priore, the eminent French scientist Pautrizel showed that the Priore treatment restored and stimulated the immune system back to its robust normal functioning, as would be predicted by the present model we are proposing.

Scientists at the Time Did Not Have the Necessary Knowledge for Understanding

It is a great tragedy that the active mechanism of the Priore device was not understood _ even by Priore himself and even by very knowledgeable physicists assigned to the program to try to ascertain the machine's mechanism. The phase conjugate or "time-reversed" EM wave was unknown until it was discovered in the open Soviet literature in 1972. Almost all of our knowledge of such waves dates from that time. In the same year, two Soviet scientists briefed American scientists at Lawrence Livermore National Laboratory on *optical* phase conjugation. Thereafter, some American scientists began intense work on optical phase conjugation.

However, the time-reversed wave is a solution to the wave equation and applies to all kinds of waves, not just EM. Phase conjugation of sound waves, e.g., is readily accomplished. The time-reversed wave solution actually appeared in the scientific literature in 1898 in a paper by Barus, who pointed out that this strange solution "made the wave run backward." But at the time the Priore program's funding was withdrawn by the French government about 1974-75, little or no understanding of the time-reversed EM wave existed as yet in the West, and of course the Whittaker work had always been ignored and was unknown also. *Literally, with the technical tools available to them at the time, the French physicists, biologists, and oncologists had no chance at all of fathoming the causative mechanism in the Priore experiments.*

Nature of the Priore Treatment

Briefly, Priore mixed some 17 frequencies in a rotating plasma in a giant tube. We know today that one function of the kind of plasma he used is to produce a phase conjugate replica wave, for an input wave. Thus Priore unknowingly achieved coupled wave/antiwave pairs in his device. The output of the plasma tube was then coupled to (modulated upon) a "rippling magnetic field" of appreciable strength. The magnetic field guaranteed penetration of the entire body and every cell in it, carrying its modulated Whittaker/ Ziolkowski internal potential structure with it. The "ripple" in the magnetic field guaranteed penetration to the atomic nuclei, and interaction with them, via nuclear magnetic resonance. In turn, unknown to Priore or the scientific team, this guaranteed the interaction of the plasma-produced W/Z structure with the internal W/Z structure of the personal QP joining those atomic nuclei. Thus the Priore approach (1) formed a deliberate counter-counter signal W/ Z structure; i.e., Priore constructed a *specific electromagnetic* *antidote, (2*) carried that antidote into all the atomic nuclei of the treated body by nuclear magnetic resonance, and (3) caused interaction with and neutralization of the internal W/Z dedifferentiation countersignal structure already present in the personal QP joining all the atomic nuclei of the treated body, *zeroing out or canceling the tumorous and pre-tumorous states existing throughout the body.*

The end results were that (1) the cumulated counter-signal to the tumor cells to dedifferentiate back toward anaerobic cells was removed, (2) the body's own centralized cellular control over the tumor cells was restored, and (3) the entire cumulated pre-cancer counter-signal pattern in the personal quantum potential was removed. In short, the individual organism started over again with a "clean slate" insofar as long-term cumulation of cancer-inducing signals in the personal quantum potential was concerned.

Furious Reaction by Orthodox French Oncologists

The Priore material, device, procedures, and results are all properly documented and presented in the peer-reviewed French medical literature. Priore obtained both French and U.S. patents on the device and on his treatment process. Robert Courrier, head of the Biology Section of the French Academy of Science, personally introduced the astounding Priore team results to the Academy. It caused a furor, and a vicious reaction of most of the orthodox French oncology establishment. Because no one could understand how the mechanism worked, there was loud and raucous insistence that "science must not be done with black boxes!" This overly pious attitude is strange, since the Priore results were multiply replicated and scientifically impeccable, and such things as aspirin were utilized for decades before their active mechanisms were understood. Indeed, in epidemology the use of vaccines and such is directly based on statistically showing that the results are real, even if the active mechanism is not understood.

When the French Government changed to a leftist government in the mid 1970s, the implacable foes of the Priore work prevailed, and the new leftist government withdrew all further funding. Priore later died, and that was the end of that.

Of all areas and treatments known to this researcher, the Priore approach is the only one that has been rigorously and scientifically proven in laboratory animals, was essentially 100% effective, and was performed under proper rigorous scientific protocols. That it has continued to be ignored by the medical and biological communities is a profound tragedy, but one due more to ignorance of the active mechanisms for the treatment than of deliberate intent.

There Is Also Deliberate Suppression

Nevertheless, there is also an element of deliberate suppression involved. Some years ago I was personally involved with a special group in an effort to restore the Priore device, when Priore was still alive. Our group obtained a tentative agreement with both the French Government and with the Priore group. The Priore machines were to be rebuilt in Bordeaux, France under the personal supervision of Priore himself. Some of the scientists who had worked with him were still alive, and would have been involved in the project. Our group had some $16 million support pledged by some wealthy backers to fund the project.

Almost immediately after our group's successful conclusion of the tentative agreements, the lives of all our backers and the lives of their families were threatened. All backing was withdrawn. The leader of our group, a Nashville businessman, came under sustained personal and business attack. He eventually lost almost everything he had, and fled this country for his life. He is still in hiding to this day. Antoine Priore then died and that was that.

The only part of the project that remained was my own grim determination to eventually provide the detailed mechanism by means of which the remarkable Priore therapeutic cures were accomplished. Success has finally been achieved after a decade of unrelenting struggle to completely explain the cause of cancer and the active mechanism by means of which the Priore group obtained such astounding scientific results.

I strongly urge medical science to take a new, hard look_ both theoretically and experimentally _ at the proven anti-cancer results obtained by Antoine Priore and the noted scientists who worked with him. I also strongly point out that new thinking, new theoretical tools, and new experimental instrumentation are needed for that re-examination. In addition I have pointed out the precise nature of the new thinking, tools, and instrumentation required.

What About Other Diseases Such As AIDS?

The question now arises about other diseases, particularly AIDS. Can the expanded Priore methodology provide equally dramatic cures of dread diseases other than cancer and leukemia? Can it conceivably cure AIDS? The answer is a resounding "Yes!"

Almost all disease conditions are electromagnetic in nature, in the sense of the extended EM bioeffects model we have outlined. The basic methodology of forming a counter-countersignal to a specific disease state can be extended to many other diseases, and to AIDS as well. The "self-targeting" effect in the cellular dedifferentiation induced by the extended EM methods results in a countersignal being produced to the *last major genetic change undergone by the diseased cell.* In the case of an HIV-infected cell, its genetics have been changed (differentiated) from normal to infected. In its internal WZ structure, the strongest resultant signal complex is thus the HIV infection differentiation command. *When pumped by the internal bidirectional WZ wave structure of the applied scalar potential, the "cell-as-a-nonlinear-phase-conjugate-mirror-material" generates a precisely tailored anti-AIDS cellular genetic dedifferentiation signal. The cell thus dedifferentiates back to a normal cell with normal genetics.*

In addition to Priore's work and Becker's results, the remaining experimental foundation for our startling statement on extended-EM reversal of AIDS was established decades ago _ as early as the late 1950s _ by the Soviet researcher Vlail Kaznacheyev.

Implications of Kaznacheyev's Experiments

In some 17,000 experiments Kaznacheyev showed that *any* cellular disease or disorder can be transmitted between cell cultures "electromagnetically" in an unknown fashion. The model we have briefly advanced here, if slightly extended to include the concept of the W/Z structure being a *graviton lattice structure, is* fully capable of explaining Kaznacheyev's results, the mechanisms, etc. This EM-transmission of disease effect shown by Kaznacheyev is known as the *cytopathogenic mirror effect*, and it has been replicated by researchers at the University of Sydney in Australia, at the University of Marburg in Germany, and by at least one researcher in the United States.

In short, what Kaznacheyev's research showed is that there exists a specific EM signal pattern of some mysterious, not-understood nature that can induce a specific cellular disease in a group of cells absorbing that signal pattern. If we regard the healthy state of the cell as "condition A" and the specific diseased state as "condition B," then specific EM signal pattern C causes the cell to move from condition A to condition B. It immediately follows that an antisignal to C _ that is, the phase conjugate of C, which is a time-reversed replica of C _ will move the diseased cell from condition B back to condition A, curing the disease. Time reversal of "sickening" is "healing." It follows from Kaznacheyev's work that (1) a specific signal pattern exists as the causative inductive factor for each specific cellular disease, and (2) a specific "antidote" signal pattern exists as the causative curative factor for that specific disease, the antidote signal being merely the time reversal or phase conjugate replica of the cytopathogenic signal. While Kaznacheyev never publicly produced the cytopathogenic mechanism, such a mechanism follows from the WZ extension of EM, and the long-term causative disease-induction mechanism I have advanced.

Creating a Specific "EM Antidote" to a Specific Disease

Consider the cumulative EM causes of changes to a cell, including to its genetics, as a serial chain of signals. Indeed, consider them as the layers of an onion, where the outer layer is the most recent. Additional layers may be added to the onion by "time forward" actions, or "differentiation" actions. If we now wish to "peel" the onion, that too is a successive, serial action, accomplished by "phase conjugate" or "time reversal" operations. These result in serial cellular dedifferentiation actions. By treating the cell and its genetics and internal parts as "highly nonlinear material structures and therefore phase conjugate mirror materials," we can "pump" the cell and its parts by bidirectional EM waves to convert the cell to a pumped phase conjugate mirror (PPCM). By orthodox nonlinear phase conjugate optics theory, with a signal wave present, the PPCM cell will emit a precise phase conjugate replica (PCR) wave. If we pump the cell with a complex of pump biwaves, the cell can and will emit a precise phase conjugate wave complex in response to a complicated signal wave complex input that is present.

For a diseased cell, we regard the disease as a differentiation of the cell, and the primary cause of the differentiation (the EM cytopathogen) as a complex WZ waveset infolded inside the personal quantum potential (in which all the body's cells exist and in whose internal signal dynamics they participate). With the pumped cell as a PPCM, and the EM cytopathogen present as an input signal wave complex to the PPCM, a precise, amplified, "EM antidote" PCR wave-complex is emitted by the parts of the PPCM cell. Each part gets a PCR complex signal order, dedifferentiating that part back the first step. A diseased part steps back to a normal part. A normal part steps back to a normal part. The diseased cell is thus dedifferentiated back to a normal cell with its parts all normal. Because the PCR wave complex is amplified, the dedifferentiation is must faster than was the original dedifferentiation. Thus an HIV-infected cell that has been "dormant" for an extended period can still be dedifferentiated back to a normal cell that is HIV-free.

So a self-targeting counter-countersignal can be automatically "developed" and generated in the body's diseased cells, for each specific cellular disease or disorder condition, or even for a mix of diseases and disorders. This is accomplished by the automatic phase conjugating action of the "cell-as-a-PPCM", when pumped by the hidden EM bidirectional structure of a scalar potential. In practice, Priore did find it necessary to adjust the particular frequency mixes he used for different disease conditions. We hypothesize that this adjustment was necessary to arrive at the proper pump wave complex that would in fact allow the proper PCR wave complex to be emitted for a particular disease type or class. Although photobiologists regularly report mitogenetic radiation emitted by a cell in the optical region, the effects desired in the cells can be produced by a different frequency regime, because of the known phenomena of nonlinear harmonic and subharmonic resonance. Since in a W/Z structure it is the *harmonic intervals and their relative relationships* that are of fundamental importance, the frequency regime of a detected W/Z structure can be shifted either harmonically or subharmonically. Priore, e.g., worked at radar frequencies in the low gigahertz range. For simplicity we may think of the frequency complex as a sort of "musical chord" composed of multiple tones. We may play the same "chord" either on the high scale or a lower scale, and still recognize the tune if the relative frequency interrelationships of the tones are retained.

A Possible Simplification of Priore's Methodology

If we turn away from the classical vector/tensor EM model presently in vogue to a higher topology EM model such as quaternions or Clifford algebra, the extended EM bioeffects model can be mathematized and refined by experiment. For example, Barrett has shown that the nonlinear phase conjugate optics effects can be obtained in ordinary circuits without the use of optical materials, when those circuits are designed and function in accord with a higher topology EM such as quaternions. Thus the circuitry of the EM emission devices themselves can be constructed to perform time-reversal of EM waves, phase conjugate replicas, 4-wave mixing, and pumped phase conjugate mirror effects, without use of normal optical materials or procedures. Priore utilized a plasma in a giant gas-filled tube, in which to mix his 17 frequencies and accomplish (unknown to scientists at the time) phase conjugation. Barrett's work implies that the necessary phase conjugation, and amplification, can be accomplished by purely electromagnetic circuitry, once the functioning of the circuitry is designed and understood in a higher topology algebra such as quaternions or Clifford algebra.

Priore demonstrated cures for diseases other than cancer. Different settings produced cures for sleeping sickness, e.g. Clogging of the arteries was also completely reversed by a variation in the Priore treatment. In addition, Pautrizel working with Priore positively demonstrated the restoration of vigorous, natural functioning to the immune system of the treated animal.

Kaznacheyev's Viral Disease Experiments: Implications for an AIDS Cure

Kaznacheyev has already demonstrated that viral-caused diseases can be induced by purely EM means in cell cultures. What this means is that cellular genetic changes also are electromagnetically induced at root basis. We regard this as a cellular differentiation. But both Becker and Priore have shown that cellular dedifferentiation can also be induced electromagnetically, by the extended scalar EM. It follows that viral-caused diseases _ including AIDS _ can be cured by purely EM means, simply by creating the proper dedifferentiation signal complex in the diseased (differentiated) cell. To determine the proper disease induction (dedifferentiation) signal pattern for the disease, the disease-inducing signal complex must be phase conjugated, and the phase conjugate replica must be amplified. The phase conjugate replica may be stimulated at the proper frequency regime by a signal complex at a much lower frequency regime. Hence the detectable photobiological optical cytopathogenic signal complexes may have their PCR counterparts stimulated by irradiation with a proper signal complex at a lower frequency regime (e.g., millimeter wave band) for use in treatment

In Conclusion

Again we strongly accent that one must take the "musical" approach: A disease pattern or a curative pattern is a complex of special frequencies and their antiwaves, similar to a musical chord. That same chord can be played in a higher or lower key, and the body will still determine the correct relationships by harmonic or subharmonic oscillation. It will still recognize the "melody" and its background chordal accompaniment; although the *individual* frequencies have changed, their *relative* relationships have not changed.

All biochemistry, genetics, and cellular effects are subject to affectation and control by proper manipulation of the W/Z structures inside scalar EM potentials. We have stated that the living organism already utilizes the internal EM structuring for mind, thought, and long-term memory, as well as the master level cellular control system we have accented. For the first time, the "software" and "firmware" of the mind appears to be directly available, something which has never before existed in science. The model advanced in my work is intended to point the way toward such use in medical science on a substantial scale, to achieve healthier, more effective, and far cheaper treatments for many of our desperate diseases that presently require costly, protracted treatments, often harsh and debilitating in and of themselves. We strongly believe that, given the direct attention of the scientific community and proper funding support, near-total, quick, and very economical cures can be developed for such dread diseases as cancer, leukemia, and AIDS. We point out also that the Priore team demonstrated cure of sleeping sickness in lab animals, as well as the direct unclogging of arteries and other blood vessels.

We also strongly accent the "self-targeting" feature when the diseased cell is "pumped" as a phase conjugate mirror material by the internal WZ biwave structure of an artificially-structured scalar EM potential. The cytopathogenic disease-induction signal complex is present in the cell and all its parts. By varying the frequencies in the internal biwave structure of the pumping potential, and hence in the pumping of the cell-as-a-PPCM, the cell is enabled to produce from and into all its parts the proper phase conjugate replica wave complex for its specific cellular disease. *Hence it produces a specific electromagnetic antidote for its own specific disease condition.* We need only adjust the frequency content of the artificial potential until the proper PCR signal complex is produced. Further, that PCR wave complex is amplified because of the pumping. Therefore the "treatment time" can be very short, and need not be anywhere near as long as the time originally required for the cytopathogenic EM signal complex to cumulate.

Today a variety of dread diseases continue to plague humankind. Untold millions of persons are suffering and dying, when the beginning of an effective, quick, inexpensive therapeutic approach capable of essentially 100% cures of cancer, leukemia, and other diseases has been demonstrated more than two decades ago.

With the exposition of the fundamental mechanism for this new therapeutic approach, perhaps the scientific community will return with renewed vigor to develop this promising area and alleviate the death and suffering so widespread today.

LIST OF KEY REFERENCES

Aharonov, Y. and D. Bohm. (i959) "Significance of Electromagnetic Potentials in the Quantum Theory," Physical Review. Second Series, Vol. 115, No. 3, p. 458491.

Aitchison, I.J.R. (1985) "Nothing's plenty: the vacuum in modern field theory." Contemporary Physics, 26(4), p. 333-391.

Amato, Ivan. (1992) "A New Blueprint for Water's Architecture," Science, Vol. 256, Jun. 26, p. 1764.

Barrett, Terrence W. (1990a) "Maxwell's Theory Extended (Part 1)," Annales de la Fondation Louis de Broglie, Vol. 15, No. 2, p. 143-183.

Barrett, Terrence W. (1990b) "Maxwell's Theory Extended (Part 2," Annales de la Fondation Louis de Broglie, Vol. 15, No. 3, p. 253-283.

Barrett, T.W. (1991) "Tesla's Nonlinear Oscillator-Shuttle-Circuit (OSC) Theory," Annales de la Fondation Louis de Broglie, Vol. 16, No. 1, p. 2341.

Barrett, Terrence W. (1992) "Electromagnetic Phenomena Not Explained by Maxwell's Equations," Annales de la Fondation Louis de Broglie [in publication].

Barus, Carl. (1898) "A Curious Inversion in the Wave Mechanism of the Electromagnetic Theory of Light," American Journal of Science, Vol. 5, Fourth Series, p. 343-348.

Bearden, T. E. (1988) "Maxwell's Original Quaternion Theory Was a Unified Field Theory of Electromagnetics and Gravitation," Proceedings of the International Tesla Symposium, Colorado Springs, Colorado.

Bearden, T.E. (1990) A Proposed Mechanism for Noyes' Power Technology, Assoc. Dist. Am. Sci., Huntsville, Alabama.

Bearden, T. E. and Walter Rosenthal. (199la) "On a Testable Unification of Electromagnetics, General Relativity, and Quantum Mechanics," Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC '91), Boston, Massachusetts, p. 487492. Also distributed over Internet.

Bearden, T. E. (199lb) Gravitobiology, Tesla Book Co., 1417 Second St, Suite H-102, Coronado, California 92118.

Bearden, T. E. (1992a) "A Redefinition of the Energy Ansatz, Leading to a Fundamentally New Class of Nuclear Interactions," Proceedings of the 27th Intersociety Energy Conversion Engineering Conference (IECEC '92), Town & Country Hotel Convention Center, San Diego, California, p. 4.3034.310.

Bearden, T.E. (1992b) An Interpretation of Whittaker's Harmonic Biwave Decomposition of the Scalar Potential. and His Replacement of Orthodox EM Forcefield Waves by Scalar Potential Interferometry, Assoc. Dist. Am. Sci., Huntsville, Alabama. [In publication].

Bearden, T.E. (1992c) EM Cytopathogenic and EM Healing Transmissions, Assoc. Dist. Am. Sci., POB 1472, Huntsville, Alabama 35807.

Bearden, T.E. (1992d) Background for Pursuing Scalar Electromagnetics, Assoc. Dist. Am. Sci., POB 1472, Huntsville, Alabama 35807. Distributed over Internet.

Bearden, T.E. (1992e) On the Unresolved Health Issues in Power Line Radiation and EM Fields and Radiation: A Fundamental New Approach, Assoc. Dist. Am. Sci., POB 1472, Huntsville, Alabama 35807.

Bearden, T.E. (1993a) "The Final Secret of Free Energy," Assoc. Dist. Am. Sci., POB 1472, Huntsville, Alabama 35807. Distributed over Internet.

Bearden, T.E. (1993b) "Mechanism for Long-Term Cumulative Biological Effects of EM Radiation," Presented to the 70th Annual Meeting of the Alabama Academy of Science, University of Alabama at Huntsville, March 25, 1993.

Becker, Robert and Gary Selden. (1985) The Body Electric, William Morrow and Company, New York.

Belyaev, B.N. (1980) "On Random Fluctuations of the Velocity of Light in Vacuum." Izvestiya Vysshikh Uchebnykj Zavendin. Fiyska (11). Nov. 1980. p. 3742. [In Russian]

Besancon, Robert M., Ed. (1974). The Encyclopedia of Physics, 2nd Edn., Van Nostrand and Reinhold, New York, p. 729-731.

Bird, Christopher. (1988) "The Case of Antoine Priore and His Therapeutic Machine: A Scandal in the Politics of Science," Appendix I in T.E. Bearden, AIDS: Biological Warfare, Tesla Book Co., Chula Vista, California.

Bohm, David. (1952) Physical Review, Vol. 85, p. 166, 180.

Bohm, David. (1980) Wholeness and the Implicate Order, Routledge & Kegan Paul, Ltd., London.

Boyer, Timothy H. (1975) "Random electrodynamics: the theory of classical electrodynamics with classical electromagnetic zero-point radiation." Phvsical Review D. Particles and Fields. Series 32. 11(4). February 15, 1975. p. 790-808.

Boyer, Timothy H. (1985) "The Classical Vacuum." Scientific American. Aug. 1985.

Casimir, H.B.G. (1953) "Introductory results in quantum electrodynamics." Physics (19). p. 846-849.

Davies, Paul C.W. (1982) "Something for Nothing." New Scientist. 94(1307). May 27, 1982. p. 580-582.

Dirac, P.A.M. (1930) "A theory of electrons and protons." Proceedings of the Royal Society of London. Series A. 126(801). Jan. 1, 1930. p. 360-365.

Elyntin, P.V. (1988) "The Quantum Chaos Problem," Sov. Phys. Usp., Vol. 31, No. 7, p. 597-622

Feynman, Richard P., Robert B. Leighton and Matthew Sands. (1963) The Feynman Lectures on Physics, Addison-Wesley, New York, Vol. I.

Fulcher, Lewis P., Johann Rafelske, and Abraham Klein. (1979). "The decay of the vacuum." Scientific American. 241(6). December, 1979. p. 150-159.

Graham, G.M. and d.G. OLahoz. (1980) "Observation of static electromagnetic momentum in vacuo." Nature. 285(5761). May 1980. p. 154-5.

Greenberg, Jack S. and Walter Grenier. (1982) "Search for the sparking of the vacuum." Physics Today. August, 1982. p. 24-32.

Hiley, B.J. and F. David Peat. (1987) "General Introduction: The Development of David Bohm's Ideas From the Plasma to the Implicate Order," B.J. Hitey and F. David Peat, Eds. Quantum Implications: Essays in Honour of David Bohm, Routtedge & Kegan Paul, London.

Hsue, C.W. (1993) "A DC Voltage is Equivalent to Two Traveling Waves on a Lossless Nonuniform Transmission Line," IEEE Microwave and Guided Wave Letters, Vol. 3, p. 82-84.

Ignatovich, V.K. (1989) "The Remarkable Capabilities of Recursive Relations," American Journal of Physics, Vol. 57, No. 10, p. 873-878.

Jackiw, R. and J.R. Schrieffer. (1981) "The decay of the vacuum." Nuclear Physics B. Vol. 190. p. 944.

Jackson, J.D. (1975) Classical Electrodynamics, 2nd Edn., John Wiley & Sons, New York.

Kaznacheyev, Vlail and L.P. Mikhailova. (1981) Ultraweak Radiation in Intercellular Interactions, [in Russian], Novosibirsk.

Kaznacheyev, Vlail. (1982) "EIectromagnetic Bioinformation in Intercellular Interactions," Psi Research, p. 47-76.

Kervran, Louis. (1982) Transmutations a Faible Energie (Naturelles et Biologiques), Maloine S.A., Paris.

Kidd, Richard et al. (1989) "Evolution of the Modern Photon," American Journal of Physics Vol. 57, No. 1, p. 27-35.

Lee, T.D. (1981) Particle Physics and Introduction to Field Theory, Harwood, New York Chapter 25: Outlook, "Possibility of Vacuum Engineering," p. 824-828.

Lindsay, Robert Bruce and Henry Margenau. (1963) Foundations of Physics, Dover Publications, New York.

Maxwell, James Clerk. (1873) A Treatise on Electricity and Magnetism, Oxford University Press, Oxford.

Misner, W., K.S. Thorne and J.A. Wheeler. (1973) Gravitation, W. H. Freeman and Co., San Francisco, p. 399; 1191.

Ohno, M. and K. Takagi. (1992) "Schlieren Visualization of Acoustic Phase Conjugate Waves Generated by Nonlinear Electroacoustic Interaction in $LiNbO_3$," Applied Phvsics Letters, Vol. 60, No. 1, p. 29-31.

Pool Robert. (1989) "Quantum Chaos: Enigma Wrapped in a Mystery," Science, Vol. 243, p. 893 - 895.

Popp, Fritz Albert. (1979) "Photon Storage in Biological Systems," in Fritz Albert Popp et al, eds., Electromagnetic Bio-Information: Proceedings of the Symposium. Marburg. Federal Republic of Germany. Sep. 5. 1977; Urban & Schwarzenberg, Baltimore, p. 123-149.

Prange, Richard E. and Peter Strance (1984) "The superconducting vacuum" American Journal of Physics . 52(1). January 984. p. 19 - 21.

Puthoff, Harold E. (1987) "Ground State of Hydrogen as a Zero-Point-Fluctuation-Determined State." Physical Review D. 35(10). May 15, 1987. p. 3266-3269.

Puthoff, Harold E (1989a) "Gravity as a Zero-Point-Fluctuation Force " Physical Review A.39(5). March 1, 1989. p. 2333 - 2342.

Puthoff, Harold E. (1989b) "Source of Vacuum Electromagnetic Zero-Point Energy, Physical Review A.40(9). p. 4857 - 4862.

Puthoff, Harold E. (1991) "The energetic vacuum: implications for energy research." Speculations in Science and Technology. 13(3). p. 247-257.

Rorvik, David M. (1975) "Do the French Have a Cure for Cancer?" Esquire, July 1975, p.110-11, 142-147.

Sachs, Robert G. (1987) The Physics of Time Reversal, University of Chicago Press, Chicago, Illinois.

Sakharov, A.D. (1968) "Vacuum Quantum Fluctuations in Curved Space and the Theory of Gravitation," Sov. Phys. Dokl., Vol. 12, p. 1040.

Schwarzschild, Bertram. (1986) "Currents in Normal-metal Rings Exhibit Aharonov-Bohm Effect," Physics Today, Vol. 39, No. 1, p. 17-20.

Sheldrake, Rupert. (1981) A New Science of Life: The Hypothesis of Formative Causation, Blond & Briggs, Ltd., U.S. reprint, J.P.

Tarcher, Inc., Los Angeles, 1982.

Stoney, G. J. (1897) "XLVIII. On a Supposed Proof of a Theorem in Wave•motion, To the Editors of the Philosophical Magazine," Philosoph. Magazine, 5(43), p. 368-373.

Strand, J. (1986) "Photons in Introductory Quantum Physics," American Journal of Physics. Vol. 54, No. 7, p. 650-652.

Sweet, Floyd and T. E. Bearden. (1991) "Utilizing Scalar Electromagnetics to Tap Vacuum Energy,~ Proceedings of the 26th Intersociety Energy Conversion Engineering Conference (IECEC '91), Boston, Massachusetts, p. 370-375. Distributed over Internet.

Tesla, Nikola. (1932) "Pioneer Radio Engineer Gives Views on Power," New York Herald Tribune, Sep. 11, 1932.

Thompson, Mark. (1989) "Mixed Signals May Have Misguided U.S. Weapons," The Washington Post, Jan. 22, 1989, p. A4.

Wheeler, John archibals. (1968) "Superspace and the nature of quantum geometrodynamics." In: Topics in Non-linear Physics, Edited by Normal J. Zabusky. Springer-Verlag. 1968. p. 642.

Whittaker, E.T. (1903) "On the Partial Differential Equations of Mathematical Physics," Mathematische Annalen, Vol. 57, p. 333-355.

Whittaker, E.T. (1904) "On an Expression of the Electromagnetic Field Due to Electrons by Means of Two Scalar Potential Functions," Proceedings of the London Mathematical Society, Series 2, Vol. 1, p. 367-372.

Ziolkowski, Richard W. (1985) "Exact Solutions of the Wave Equation With Complex Source Locations," Journal of Mathematical Physics, Vol. 26, p. 861.

Ziolkowski, Richard W. (1989a) "Localized Transmission of Wave Energy," Proc. SPIE. Vol. 1061. Microwave and Particle Beam Sources and Directed Energy Concepts. p. 396-397.

Ziolkowski, Richard W. (1989b) "Localized Transmission of Electromagnetic Energy," Physical Review A, Vol. 39, p. 2005.

Ziolkowski, Richard W. (1989c) "Localized Transmission of Wave Energy," Proc. SPIE. Vol. 1061. Microwave and Particle Beam Sources and Directed Energy Concepts. p. 396-397.

Ziolkowski, Richard W. (1992a) "Localized Wave Transmission Physics and Engineering," Physical Review A (in Press).

Ziolkowski, Richard W. (1992c) "Properties of Electromagnetic Beams Generated by Ultra-Wide Bandwidth Pulse-Driven Arrays, IEEE Trans. Ant. Prop. (Submitted).

Ziolkowski, Richard W., Amr M. Shaarawi, and Ioannis M. Besieris. (1989) Nuclear Physics B (Proc. Suppl.), Vol. 6, p. 255-258.

Ziolkowski, R. W. and Lewis, D.K. (1990) "Verification of the Localized Wave Transmission Effect," Journal of Applied Physics, Vol. 68, p. 6083.

Go to Slides

The Tom Bearden Website

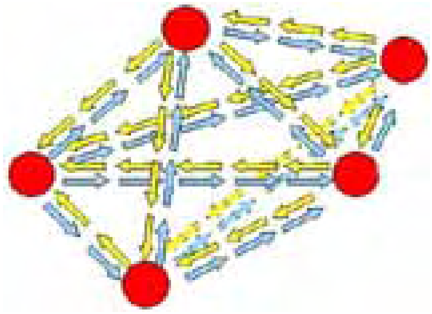

CANCER AND THE UNRESOLVED HEALTH ISSUES IN THE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATIONS

Slide Index

---

1. Title Page
2. Introduction
3. Einstein on Review of Foundation
4. Some Problems in Electromagnetic Foundations
5. Notions in the Concept of the Potential
6. Foundations of Scalar Electromagnetics
7. Inner EM Waves of a Scalar Potential
8. Graviton/Whittaker Lattice Structure of the Scalar Potential
9. Internal Wave Structure of the Scalar Potential
10. Lattice and Nonlattice Gravitons
11. Scalar EM Potential Interferometry
12. Scalar Potential Interferometry
13. Some Problems in Foundations of General Relativity
14. Some Problems in Foundations of Quantum Mechanics
15. Physics Disciplines are Presently Inconsistent
16. The Present EM, GR and QM are Subsets, each omits the Whittaker Internal EM
17. Unified Field Secret: Apply Whittaker Theory
18. Present Definition of "Energy" is Erroneous
19. Some New Definitions
20. Definitions (Continued)
21. Definitions (Continued)
22. Electric Charge, Open System

23. Graviton as a Coupled Photon/Antiphoton Pair
24. Photon Interaction is usually Graviton Interaction
25. Photon Interaction with the Material Atom
26. Photon Phase Conjugate interaction with an Atom
27. Nonlinear Material and Time-Domain Resonance
28. Subharmonic Oscillation
29. The Quantum Potential
30. Quantum Potential Characteristics
31. Self-Targeting creates a Quantum Potential
32. Approach applied to correct present EM bioeffects experimental model
33. Two Stages of Cancer Production
34. More on Promoters
35. EM Biological Effects: Some Changes Demonstrated
36. EM Biological Effects: Inhibition of Melatonin
37. Other EM Biological Effects
38. Becker's Bone-Healing Application of Tiny DC Currents to Fractures
39. Questions on Carcinogens and EM Radiation
40. Direct and Delayed EM Bioeffects
41. Cumulative Feedback Reaction Loop for Induction of Long-Term Effects
42. Species Reactive Adaptation
43. Inheritance of Acquired Traits
44. Anomalous EM Jamming Mechanism
45. Anomalous Depth Penetration
46. The Evolutionary Path of Multicellular Organisms
47. Development of Multi-Celled Organisms
48. Effect of Fluid Contamination on Hemoglobin's Oxygen Transport
49. First Step in Oxygen-Deficiency Dedifferentiation
50. Cumulative Growth of Dedifferentiation Signals
51. Cumulative Emergence of Cancer into the Physical State
52. Precancer State Condition determines Treatment Success
53. Starting From the Scalar EM View: A New Definition of Cancer
54. The Immune System
55. Kaznacheyev's Apparatus
56. Death and Disease Transmission via Graviton Templates
57. Microwave Radiation of U.S. Embassy in Moscow
58. John Hopkins Study of Embassy Microwave Radiation

59. Priore's Laboratory
60. Mechanism of the Priore Therapy
61. Drastically Reduced Deterioration with Priore Therapy
62. The Missing Piece of the Puzzle
63. Summary and Conclusion

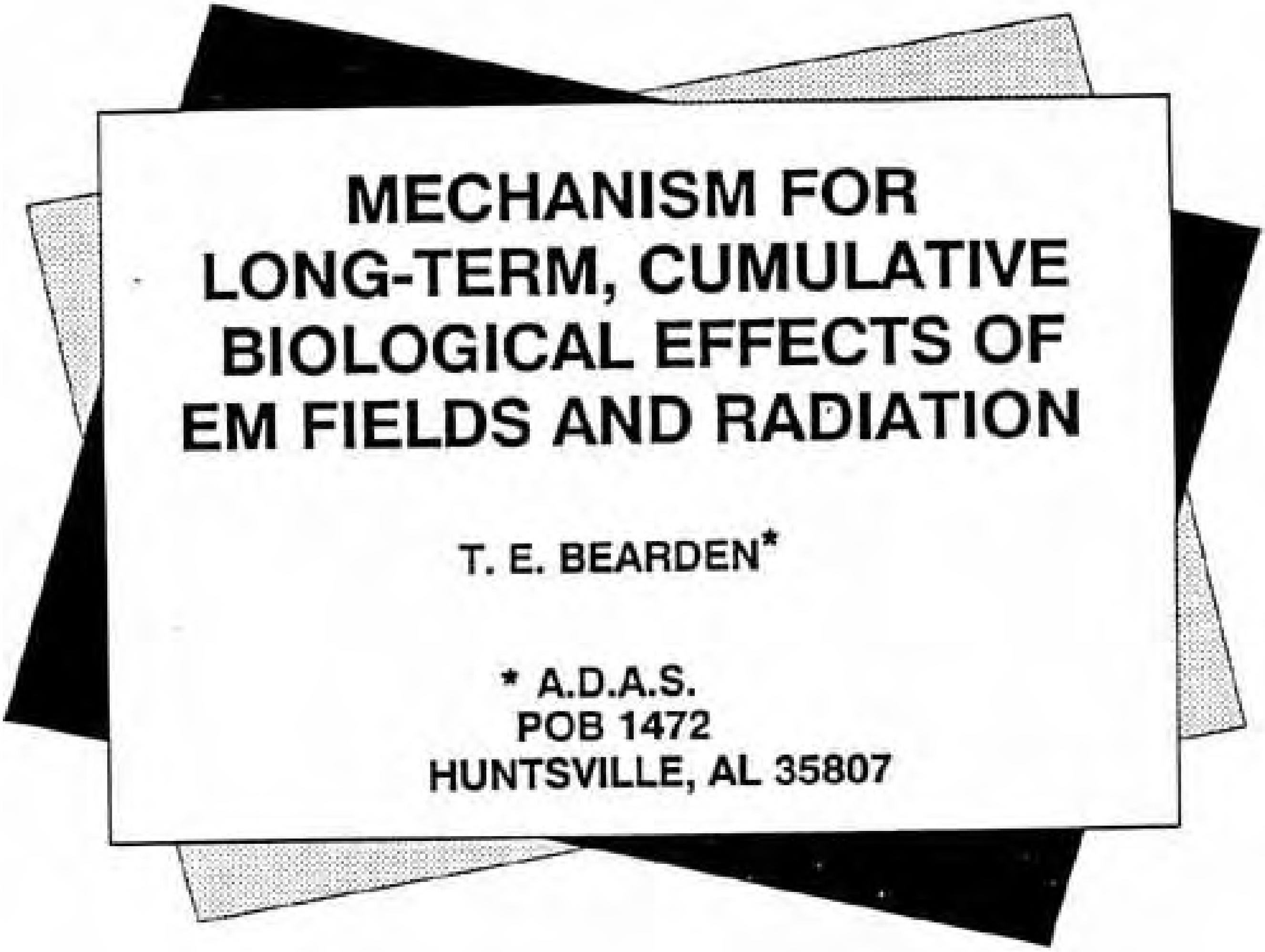

MECHANISM FOR LONG-TERM, CUMULATIVE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATION

T. E. BEARDEN*

* A.D.A.S.
POB 1472
HUNTSVILLE, AL 35807

Next Slide

Return to Cancer Paper

INTRODUCTION

- **PROBLEMS IN EM BIOEFFECTS DETERMINATION FOR FOUR DECADES**
  - DIFFICULTY IN REPLICATION
  - STUDIES ARE CONTRADICTORY
  - EMOTIONALISM, CONTROVERSY, VESTED INTERESTS
  - LACK FUNDAMENTAL MECHANISMS
- **ESTABLISHES THAT MODEL UTILIZED IS INADEQUATE**
- **WILL CRITIQUE STANDARD MODEL**
- **WILL PROPOSE ALTERATION AND EXTENSION**
- **WILL EXAMINE FUNDAMENTAL CONCEPTS, PER EINSTEIN**

Next Slide
Previous Slide

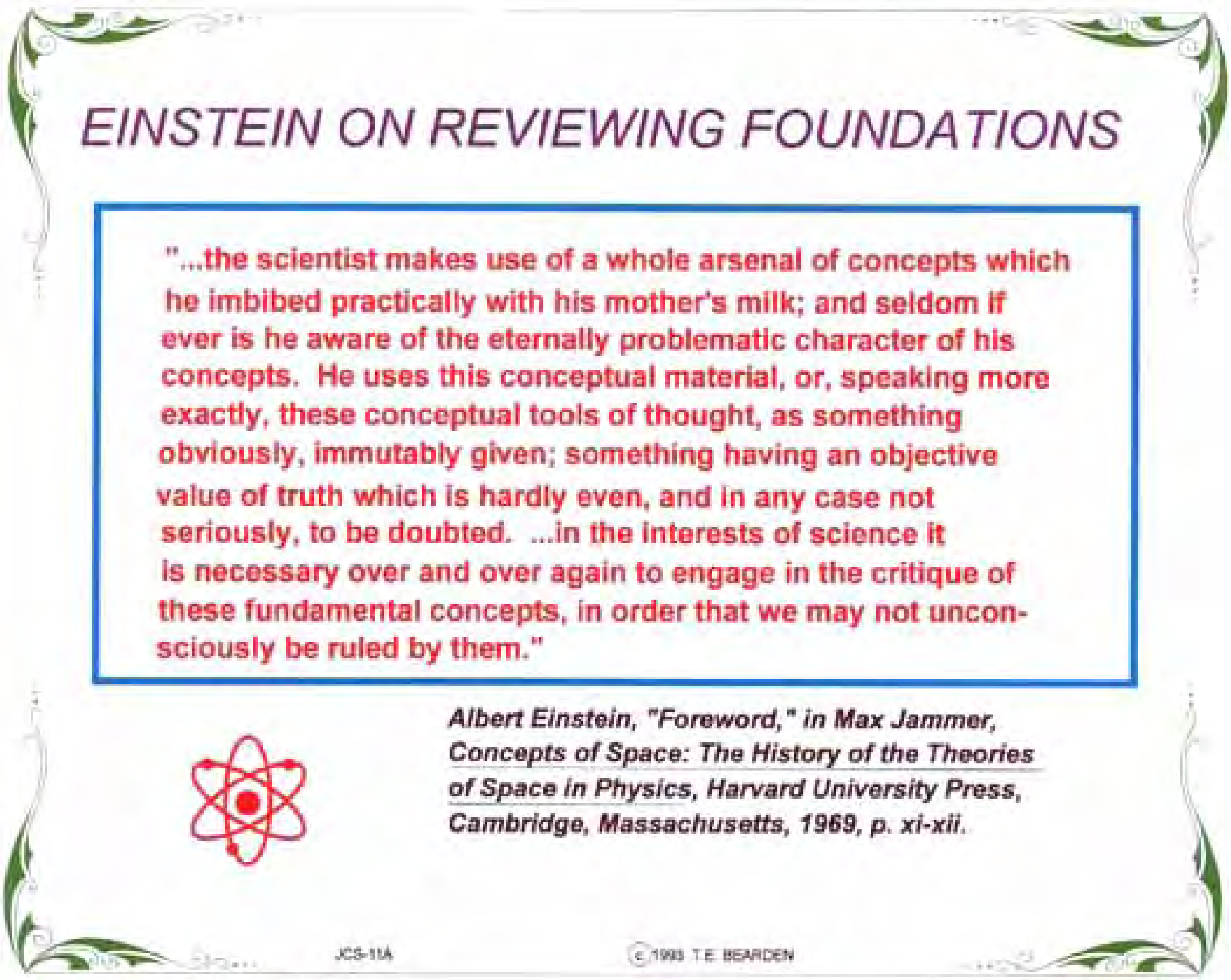

EINSTEIN ON REVIEWING FOUNDATIONS

"...the scientist makes use of a whole arsenal of concepts which he imbibed practically with his mother's milk; and seldom if ever is he aware of the eternally problematic character of his concepts. He uses this conceptual material, or, speaking more exactly, these conceptual tools of thought, as something obviously, immutably given; something having an objective value of truth which is hardly even, and in any case not seriously, to be doubted. ...in the interests of science it is necessary over and over again to engage in the critique of these fundamental concepts, in order that we may not unconsciously be ruled by them."

Albert Einstein, "Foreword," in Max Jammer, Concepts of Space: The History of the Theories of Space in Physics, Harvard University Press, Cambridge, Massachusetts, 1969, p. xi-xii.

JCS-11A

© 1993 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

SOME PROBLEMS IN ELECTROMAGNETICS FOUNDATIONS

- NOTION OF FORCEFIELD
- NATURE OF THE VACUUM
- NATURE OF ELECTRICAL CHARGE
- NOTION OF POTENTIAL
- NOTION OF ENERGY
- NOTION OF QUANTIZED ELECTRICAL CHARGE
- NOTION OF TIME
- CONSIDERED MUTUALLY EXCLUSIVE TO GRAVITY
- FRAME VERSUS SPACETIME
- CLOSED SYSTEM ASSUMPTION
- TOPOLOGY AND DIMENSIONALITY OF THE MODEL'S MATHEMATICS
- UNDEFINED, OR IMPROPERLY DEFINED, CONCEPTS

Next Slide

Return to Index

Previous Slide

NOTIONS IN THE CONCEPT OF THE POTENTIAL

- **CONCEPT OF THE POTENTIAL HAS DEVELOPED AS SEVERAL FUNDAMENTAL NOTIONS:**
  - **AS THE "INTERNAL" STORAGE OF ENERGY IN SOME FASHION**
  - **AS A POINT-VALUE FUNCTION IN 3-SPACE**
  - **AS A SCALAR FUNCTION WHOSE SPACE RATE OF CHANGE YIELDS A VECTOR FORCE**
- **THE ENERGY STORAGE MECHANISM IS QUITE UNKNOWN.**
- **NO THOUGHT HAS BEEN GIVEN TO SIMULTANEOUSLY DEFINING IT AS HAVING A HYPERSPATIAL WAVE SET COMPOSITION.**
- **THE THIRD NOTION EFFECTIVELY MAKES ALL FORCES LOCAL *A PRIORI*, AND EXCLUDES ACTION AT A DISTANCE BY *POSTULATION*.**
- **THE THIRD NOTION IS KNOWN TO BE INCORRECT. THE GRADIENT OF A POTENTIAL DOES NOT PRODUCE A FORCE UNTIL A MASS IS PRESENT, TO WHICH THE GRADIENT COUPLES.**

Next Slide
Previous Slide

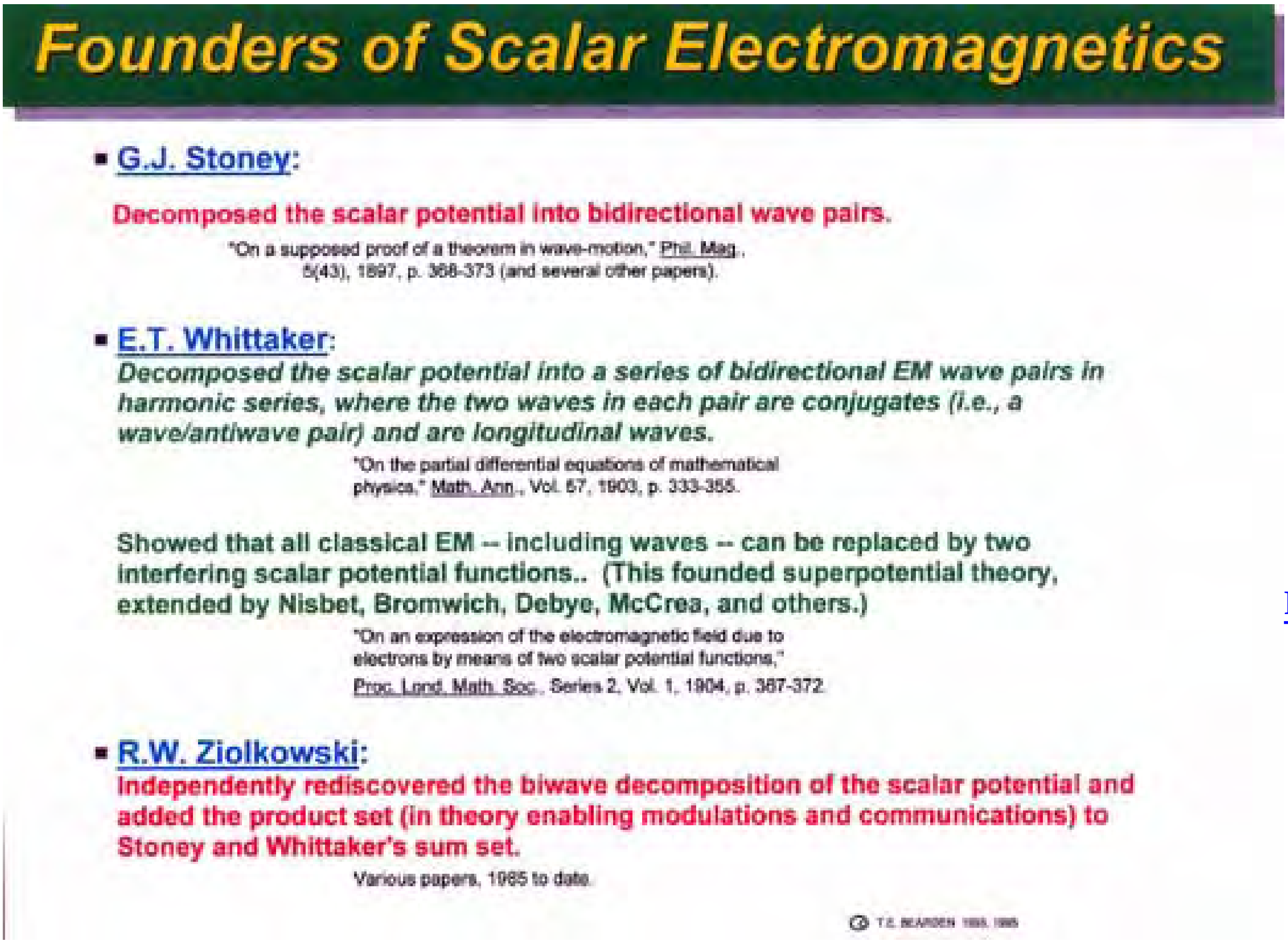

Founders of Scalar Electromagnetics

- **G.J. Stoney**:

  Decomposed the scalar potential into bidirectional wave pairs.

  "On a supposed proof of a theorem in wave-motion," Phil. Mag., 5(43), 1897, p. 368-373 (and several other papers).

- **E.T. Whittaker**:

  Decomposed the scalar potential into a series of bidirectional EM wave pairs in harmonic series, where the two waves in each pair are conjugates (i.e., a wave/antiwave pair) and are longitudinal waves.

  "On the partial differential equations of mathematical physics," Math. Ann., Vol. 57, 1903, p. 333-355.

  Showed that all classical EM -- including waves -- can be replaced by two interfering scalar potential functions.. (This founded superpotential theory, extended by Nisbet, Bromwich, Debye, McCrea, and others.)

  "On an expression of the electromagnetic field due to electrons by means of two scalar potential functions," Proc. Lond. Math. Soc., Series 2, Vol. 1, 1904, p. 367-372.

- **R.W. Ziolkowski**:

  Independently rediscovered the biwave decomposition of the scalar potential and added the product set (in theory enabling modulations and communications) to Stoney and Whittaker's sum set.

  Various papers, 1985 to date.

© T.E. BEARDEN [illegible]

Next Slide

Return to Index

Previous Slide

INNER EM WAVES OF A SCALAR POTENTIAL

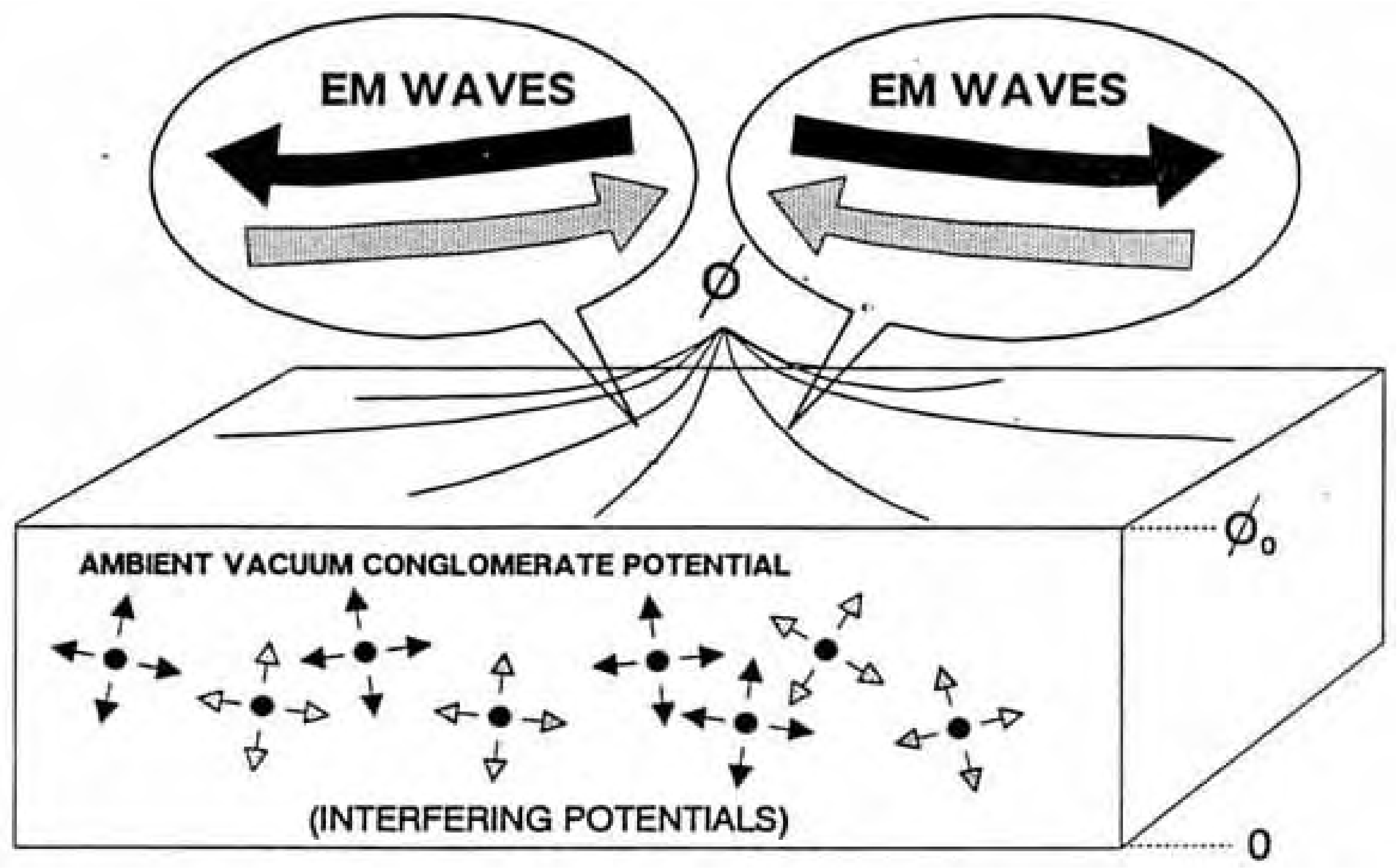
EM WAVES
EM WAVES
ϕ
ϕ_0
AMBIENT VACUUM CONGLOMERATE POTENTIAL
(INTERFERING POTENTIALS)
0

Next Slide
Previous Slide

GRAVITON/WHITTAKER LATTICE STRUCTURE OF THE SCALAR POTENTIAL

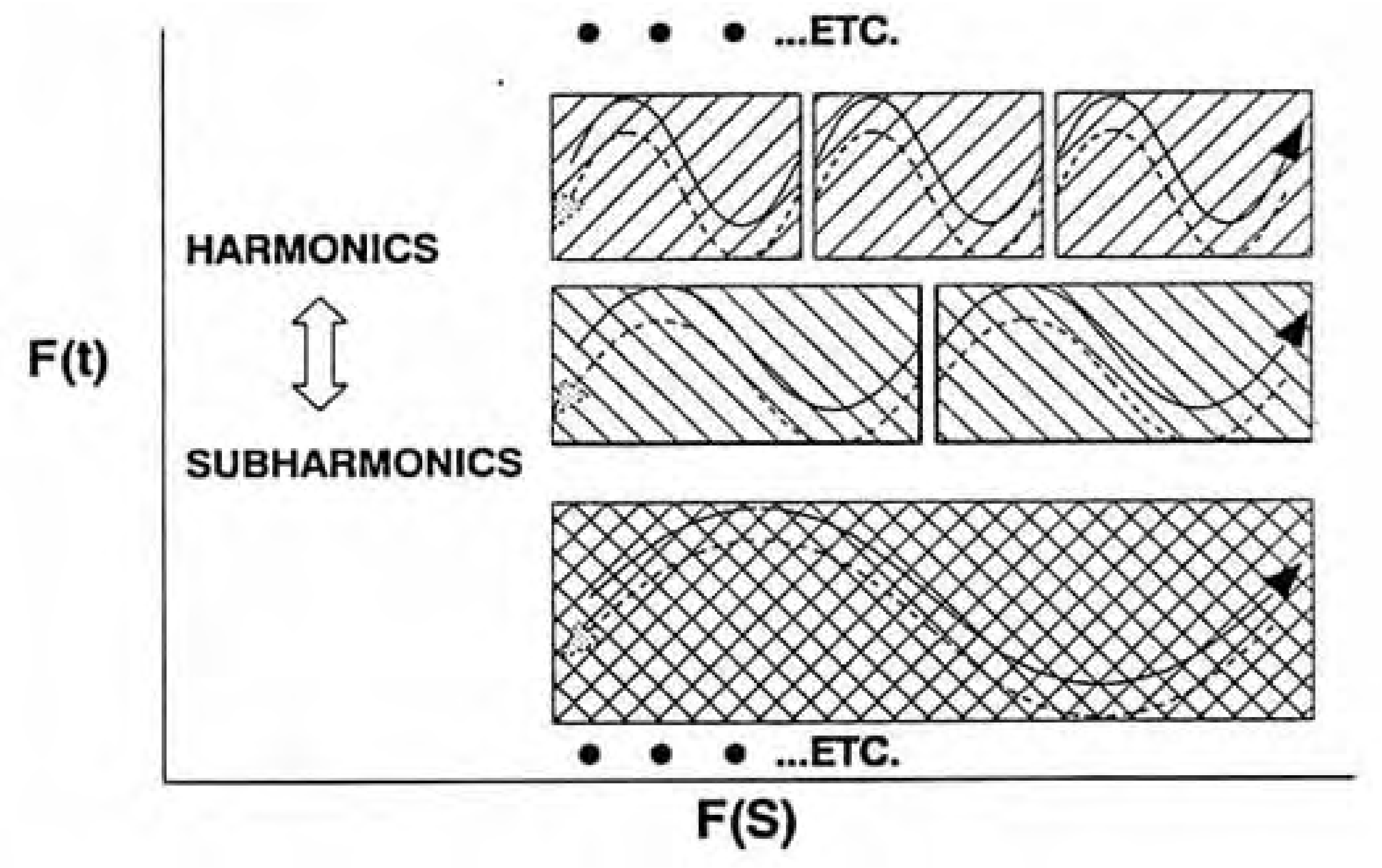

• • • ...ETC.
HARMONICS
F(t)
SUBHARMONICS
• • • ...ETC.
F(S)

Next Slide
Previous Slide

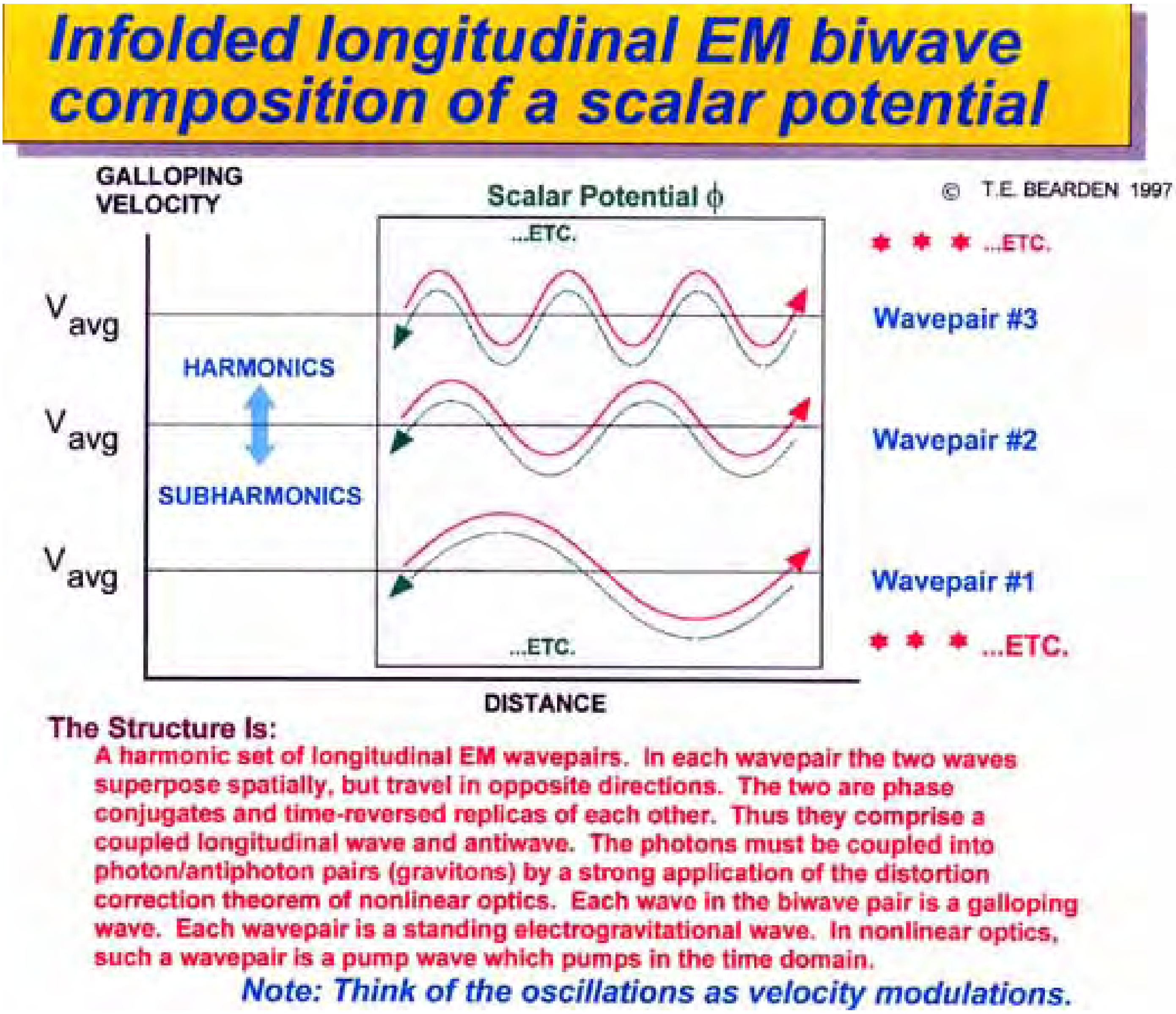

Infolded longitudinal EM biwave composition of a scalar potential

GALLOPING VELOCITY
Scalar Potential ϕ
© T.E. BEARDEN 1997
...ETC.
✱ ✱ ✱ ...ETC.
V_{avg}
Wavepair #3
HARMONICS
V_{avg}
Wavepair #2
SUBHARMONICS
V_{avg}
Wavepair #1
...ETC.
✱ ✱ ✱ ...ETC.
DISTANCE

The Structure Is:

A harmonic set of longitudinal EM wavepairs. In each wavepair the two waves superpose spatially, but travel in opposite directions. The two are phase conjugates and time-reversed replicas of each other. Thus they comprise a coupled longitudinal wave and antiwave. The photons must be coupled into photon/antiphoton pairs (gravitons) by a strong application of the distortion correction theorem of nonlinear optics. Each wave in the biwave pair is a galloping wave. Each wavepair is a standing electrogravitational wave. In nonlinear optics, such a wavepair is a pump wave which pumps in the time domain.

Note: Think of the oscillations as velocity modulations.

Next Slide

Return to Index

Previous Slide

LATTICE AND NONLATTICE GRAVITONS

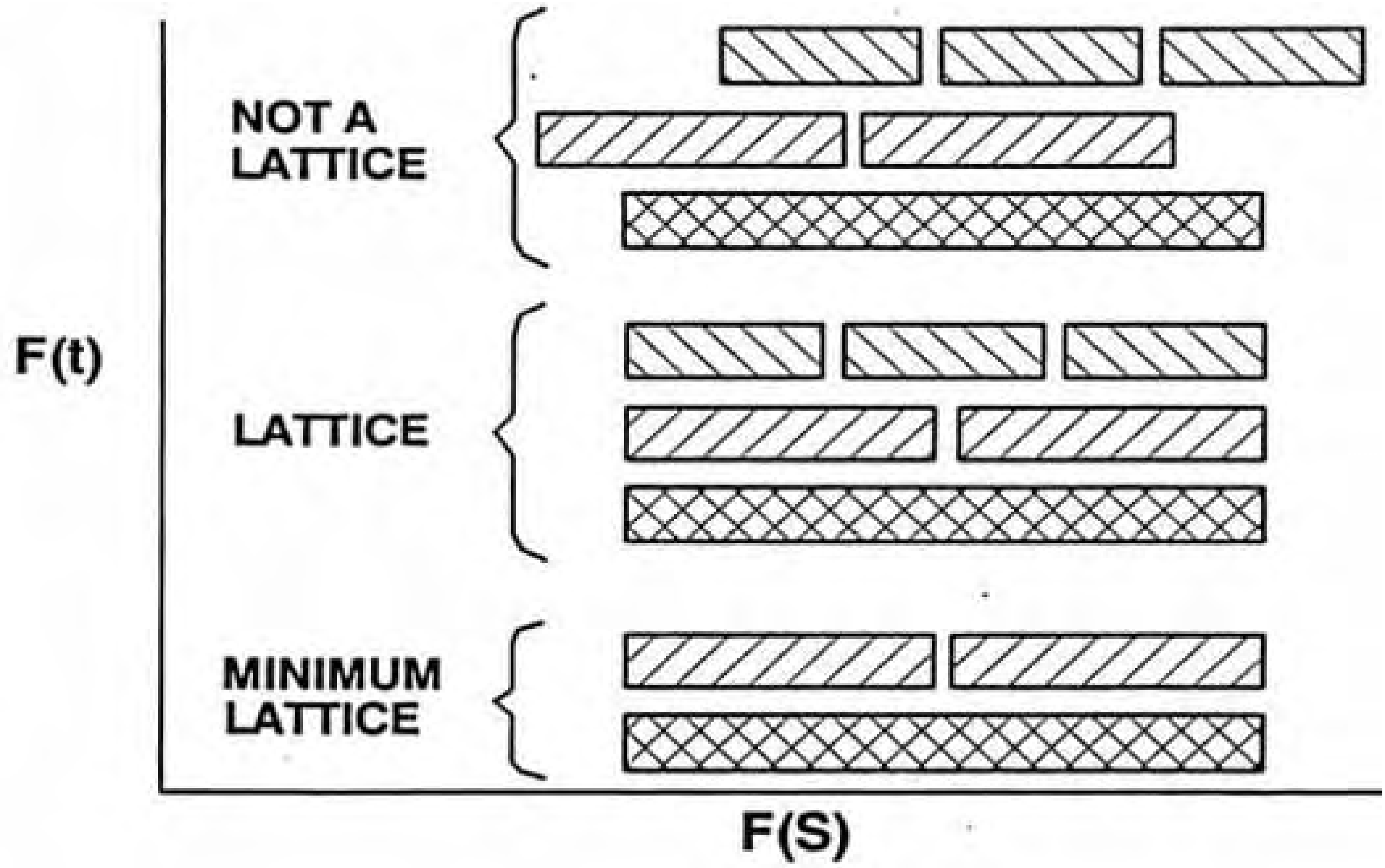
F(t)
NOT A LATTICE
LATTICE
MINIMUM LATTICE
F(S)

Next Slide
Previous Slide

SCALAR EM POTENTIAL INTERFEROMETRY

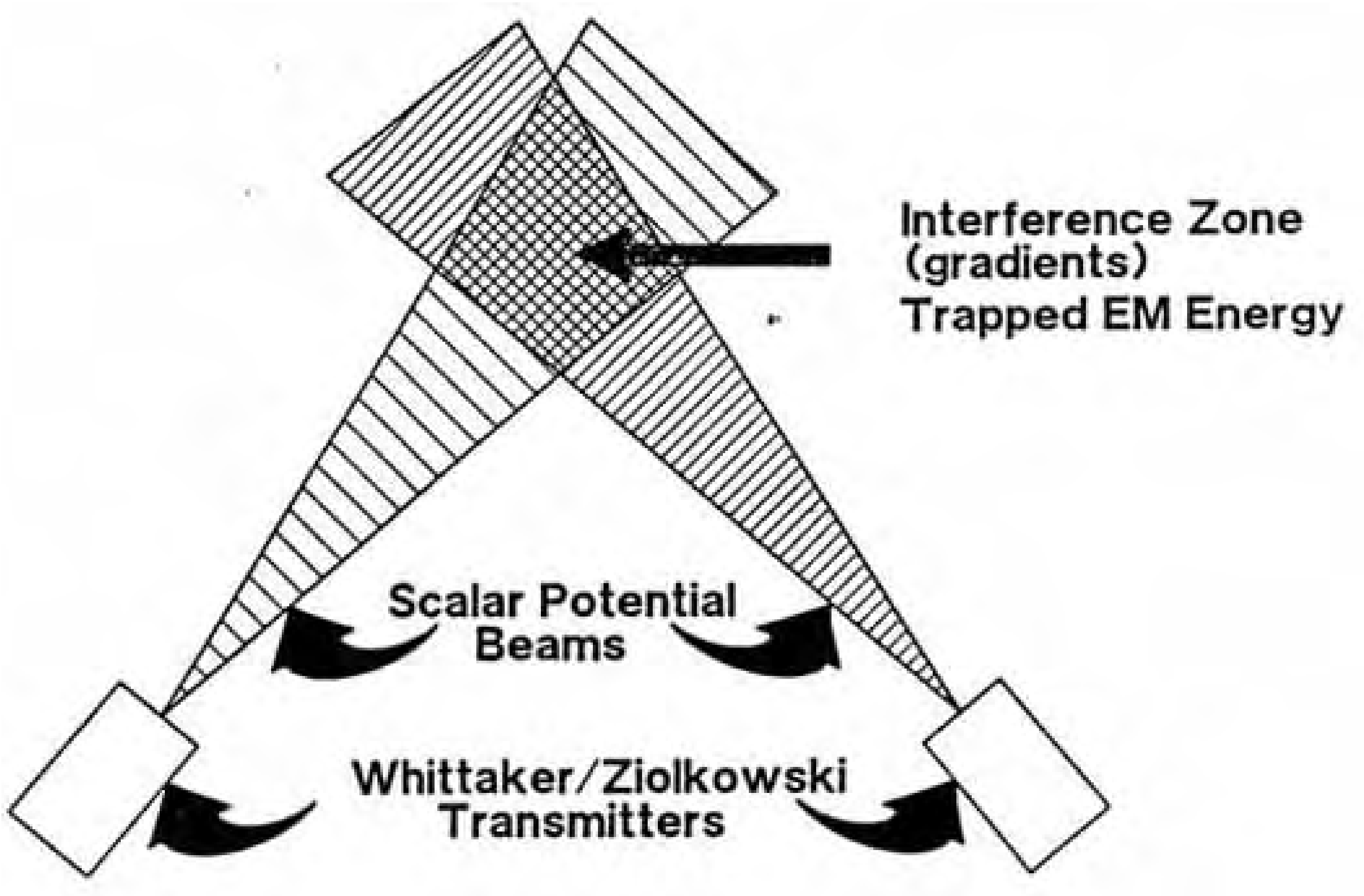
Interference Zone
(gradients)
Trapped EM Energy
Scalar Potential
Beams
Whittaker/Ziolkowski
Transmitters

Next Slide
Previous Slide

SCALAR POTENTIAL INTERFEROMETRY

- Between two scalar potentials, **interferometry** is:
  - Interference of the multi-wave sets in hyperspace.
  - Production of gradients in the 3-space point values
  - In the classical (erroneous) postulation, the creation of forces in 3-space.
    > Force does not exist until the gradient couples to an observable mass.
    > There are no E-fields and B-fields as such in the massless vacuum, but only potential gradients.

Next Slide
Previous Slide

SOME PROBLEMS IN FOUNDATIONS OF GENERAL RELATIVITY

- **EINSTEIN'S FLAT LOCAL SPACETIME POSTULATE**
- **ONLY CONSIDERS THE VERY WEAK G-FORCE AS AN AGENT FOR SPACETIME CURVATURE**
- **EXCLUDES UNION WITH EM FIELD**
- **EXCLUDES ACTION-AT-A-DISTANCE, WHICH IS REQUIRED BY QUANTUM MECHANICS, AND PROVEN**
- **EXCLUDES UNION WITH QUANTUM MECHANICS**
- **SPECIFICALLY EXCLUDES ELECTROGRAVITATION**
- ***IS ESSENTIALLY NOT AN EXPERIMENTAL SCIENCE***

Next Slide

Previous Slide

Return to Index

SOME PROBLEMS IN FOUNDATIONS OF QUANTUM MECHANICS

- **EXCLUDES CHAOS (HIDDEN ORDER), AND THIS IS KNOWN TO BE IN ERROR**
- **ORDERED MACROREALITY CANNOT EMERGE FROM INTEGRATING DISORDERED (RANDOM) CHANGES**
- **USES RANDOM-VARIABLE STATISTICS FROM GIBBS THERMODYNAMICS**
- **NOTION THAT QUANTUM CHANGE - HENCE ROOTS OF PHYSICAL REALITY - CANNOT BE ENGINEERED**
- **REQUIRES ACTION-AT-A-DISTANCE EFFECTS**
- **CONTRADICTS CLASSICAL MECHANICS, GENERAL RELATIVITY, AND CLASSICAL ELECTROMAGNETICS**

Next Slide

Return to Index

Previous Slide

PHYSICS DISCIPLINES ARE PRESENTLY INCONSISTENT

	GENERAL RELATIVITY (GR)	
QUANTUM MECHANICS (QM)	**???**	**CLASSICAL EM THEORY**
	MIND & SUBTLE ENERGY	

Next Slide

Return to Index

Previous Slide

THE PRESENT EM, GR AND QM ARE SUBSETS, EACH OMITS THE WHITTAKER INTERNAL EM.

A	B
PRESENT EM	EXCLUDED EM
(EXTERNALIZED ASPECTS OF POTENTIALS)	(INTERNALIZED ASPECTS OF POTENTIALS)

A	B
PRESENT GR	EXCLUDED GR
(DISTANT SPACETIME CURVATURE) (LOCAL ACTION)	(LOCAL SPACETIME CURVATURE) (ACTION AT A DISTANCE)

A	B
PRESENT QM	EXCLUDED QM
(EXTERNALIZED STRUCTURE OF POTENTIALS)	(INTERNALIZED STRUCTURE OF POTENTIALS) (CHAOS)

Next Slide

Return to Index

Previous Slide

UNIFIED FIELD SECRET: APPLY WHITAKER THEORY

Next Slide

Return to Index

Previous Slide

	GENERAL RELATIVITY (GR)	
QUANTUM MECHANICS (QM)	APPLY WHITTAKER THEORY	CLASSICAL EM THEORY
	MIND & SUBTLE ENERGY	

PRESENT DEFINITION OF "ENERGY" IS ERRONEOUS

- **PRESENT: "ENERGY IS THE CAPACITY TO DO WORK."**
- **ACTUALLY, WORK IS THE SCATTERING (DISSIPATION) OF ENERGY.**
- **THE PRESENT DEFINITION STATES THAT "ENERGY IS THE CAPACITY TO DO SCATTERING OF ENERGY."**
- **ENERGY CAN BE SCATTERED, YIELDING WORK. BUT THE CAPACITY TO PERFORM THAT SCATTERING IS NOT ENERGY! (E.G., IT MAY BE RESISTANCE.)**
- **COMPARE TO: "A FISHHOOK IS THE CAPACITY TO CATCH FISH."**
- **ACTUALLY, A FISHHOOK IS A BENT WIRE WITH A BARB. IT *HAS* THE CAPACITY TO CATCH FISH (GIVEN OTHER CONDITIONS). BUT THE CAPACITY TO CATCH FISH IS NOT WHAT IT *IS*.**

Next Slide

Return to Index

Previous Slide

SOME NEW DEFINITIONS

- **QUANTUM MECHANICAL VACUUM -** an intense flux of virtual particles, filling the "emptiness" of spacetime/vacuum.
- **ENERGY -** any ordering, either static or dynamic, in the virtual particle flux of vacuum.
- **EM ENERGY -** any ordering, either static or dynamic, in the virtual photon flux of vacuum
- **WORK -** the local scattering of energy; the local disordering of the order in the VPF.

Next Slide

Return to Index

Previous Slide

DEFINITIONS (CONTINUED)

- **POTENTIAL** - any ordering, either static or dynamic or combination, in the VPF of vacuum
  - same definition as energy
  - must have internal order, hence structure
- **SCALAR POTENTIAL** - any static ordering in the VPF of vacuum, with respect of the external observer
- **VECTOR POTENTIAL** - any dynamic (non-stationary) ordering in the VPF of vacuum.
  - must have internal order, hence structure

Next Slide

Return to Index

Previous Slide

DEFINITIONS (CONTINUED)

- **Electrostatic Scalar Potential** - any static (stationary) ordering in the virtual photon flux of vacuum.
  - Whittaker decomposed the ESP into bidirectional EM wavepairs in a harmonic structure.
  - We interpret the wavepair set as existing in hyperspace (in hyper-3-space).
- Thus the ESP is a 3-space point function, and a hyper-3-space vector function, in 7-space.
- In Hyper-3-space, the ESP is an ordering established throughout the universe.

Next Slide

Return to Index

Previous Slide

ELECTRIC CHARGE, OPEN SYSTEM

- The ***electric charge*** of a charged particle is due to a violent exchange of virtual photons between the charged particle and the local vacuum VPF
  - photons continually absorbed from the vacuum
  - photons continually radiated back to the vacuum
  - this exchange flux is what electric charge ***IS***.
- In the nucleus, everything is changing into everything, so to speak. It is a furious, ***VPF-driven*** cauldron.
- the nucleus is an open, driven system, normally in equilibrium with respect to VPF input and output.

Next Slide

Return to Index

Previous Slide

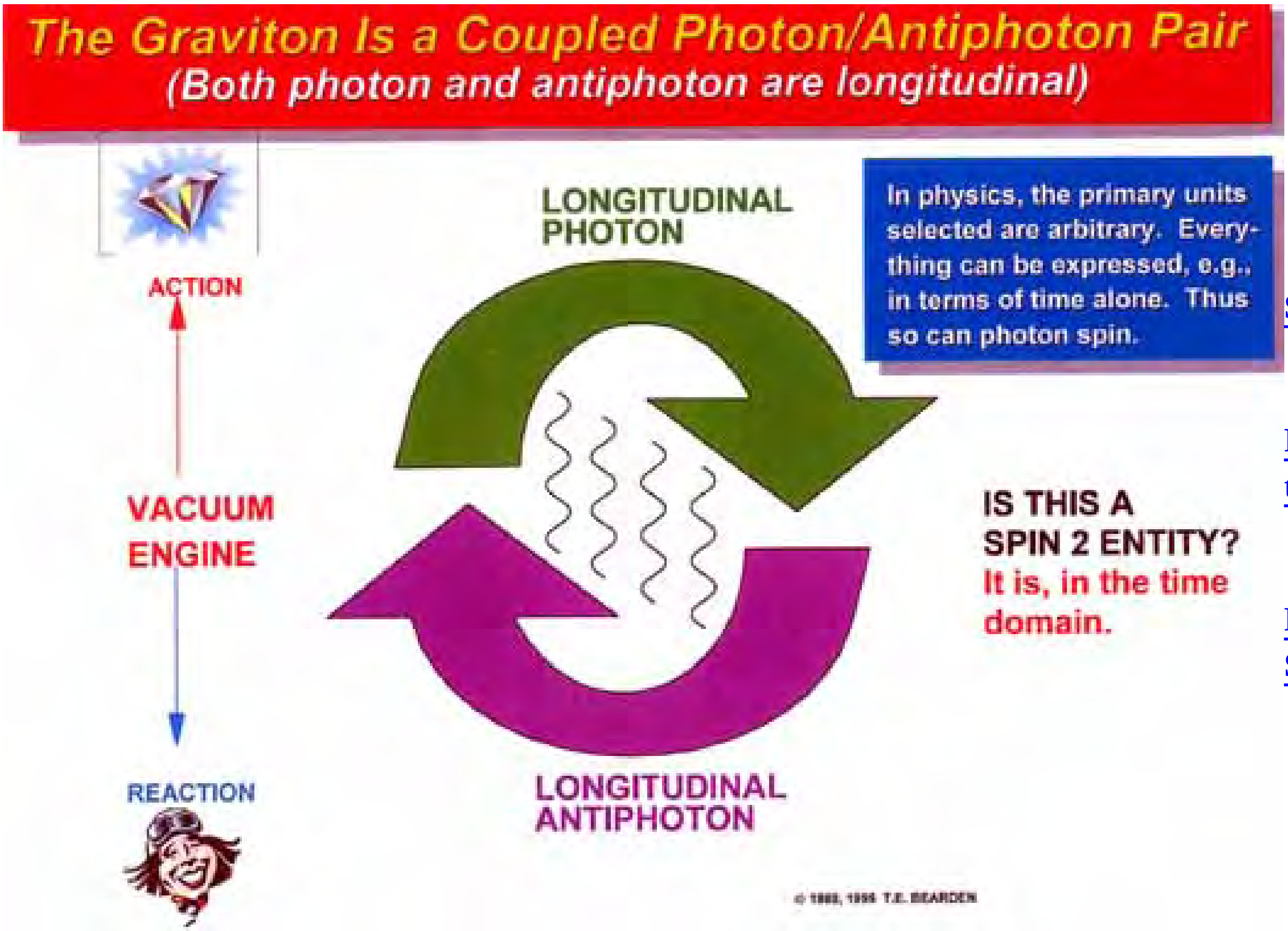

The Graviton Is a Coupled Photon/Antiphoton Pair

(Both photon and antiphoton are longitudinal)

In physics, the primary units selected are arbitrary. Everything can be expressed, e.g., in terms of time alone. Thus so can photon spin.

IS THIS A SPIN 2 ENTITY?
It is, in the time domain.

ACTION
VACUUM ENGINE
REACTION
LONGITUDINAL PHOTON
LONGITUDINAL ANTIPHOTON

© [illegible] T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

PHOTON INTERACTION IS USUALLY GRAVITON INTERACTION

- Graviton interaction is a photon/antiphoton interaction
- Photon interacts with electron shells
  - Scattering produces the entropic (forward) flow of time
  - Produces electron translation in our detectors
- Antiphoton interacts with nucleus primarily
  - produces nuclear recoil
  - produces Newton's third law of motion

Next Slide

Return to Index

Previous Slide

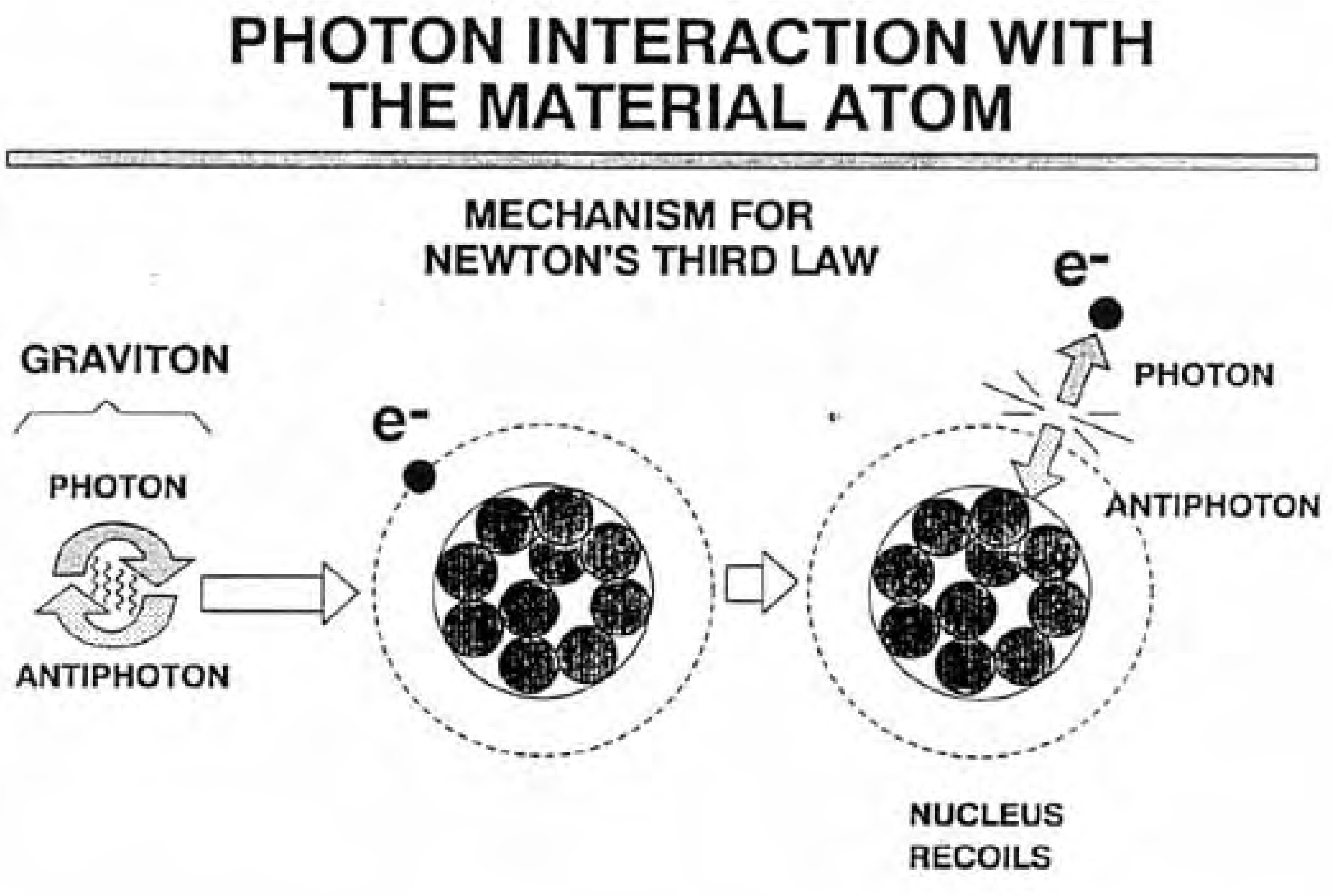

PHOTON INTERACTION WITH THE MATERIAL ATOM

MECHANISM FOR NEWTON'S THIRD LAW

GRAVITON
PHOTON
ANTIPHOTON
e-
e-
PHOTON
ANTIPHOTON
NUCLEUS RECOILS

Next Slide

Return to Index

Previous Slide

PHOTON PHASE CONJUGATE INTERACTION WITH AN ATOM

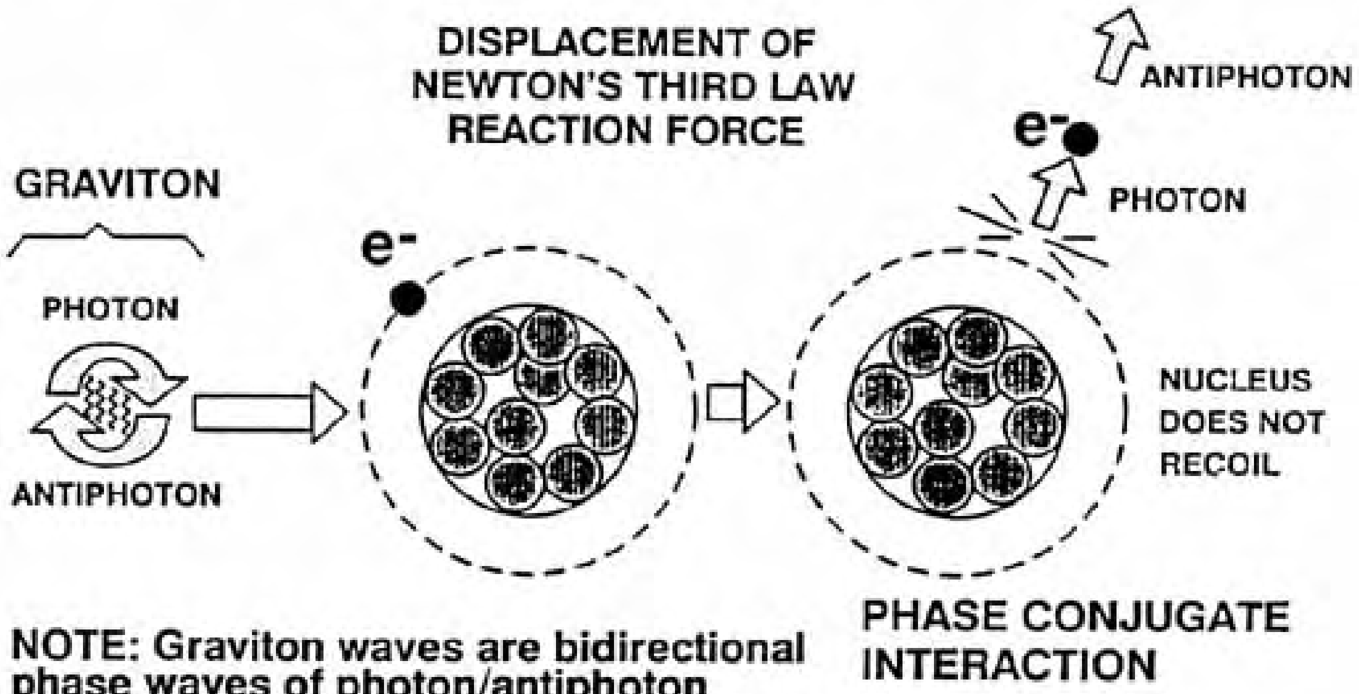
DISPLACEMENT OF NEWTON'S THIRD LAW REACTION FORCE
ANTIPHOTON
e-
PHOTON
GRAVITON
e-
PHOTON
ANTIPHOTON
NUCLEUS DOES NOT RECOIL
PHASE CONJUGATE INTERACTION

NOTE: Graviton waves are bidirectional phase waves of photon/antiphoton coupling and uncoupling in the vacuum.

Next Slide

Return to Index

Previous Slide

NONLINEAR MATERIAL AND TIME-DOMAIN RESONANCE

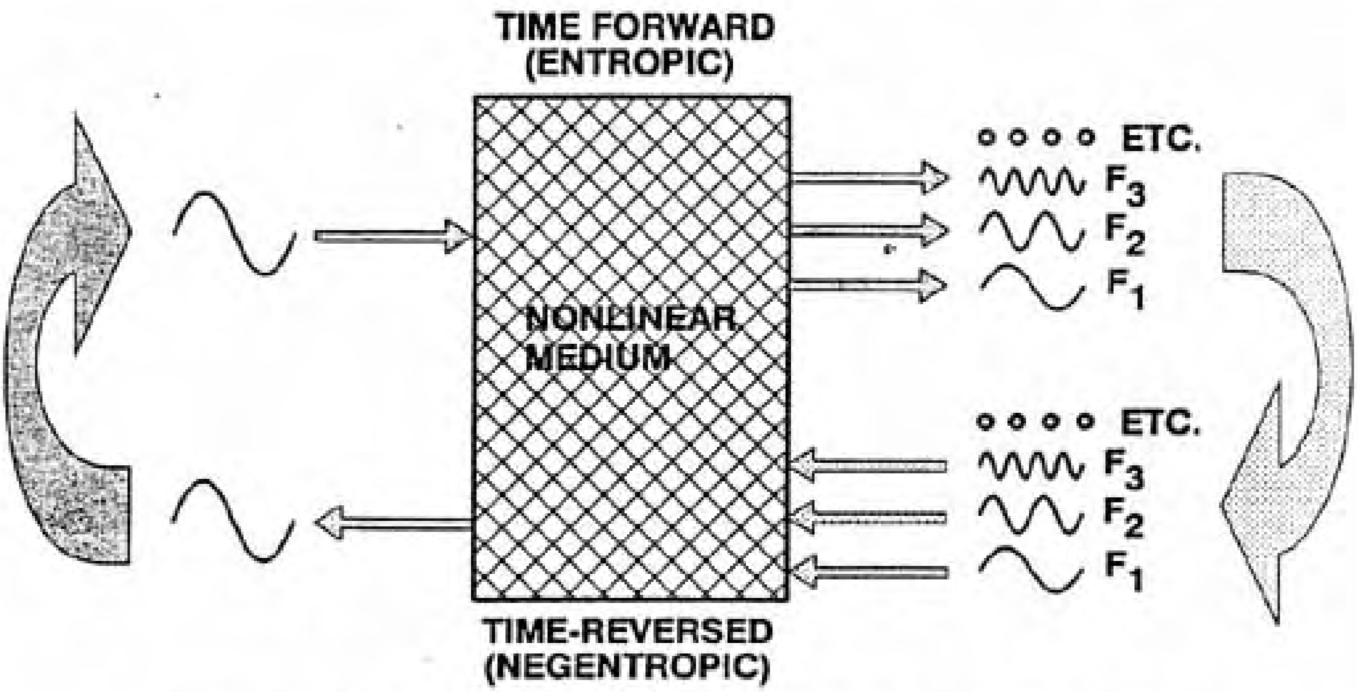

TIME FORWARD
(ENTROPIC)

NONLINEAR
MEDIUM

○ ○ ○ ○ ETC.
F_3
F_2
F_1

○ ○ ○ ○ ETC.
F_3
F_2
F_1

TIME-REVERSED
(NEGENTROPIC)

GROUPS OF RESONATORS ENTRAINED IN TIME AND SPACE
(MAY ENABLE SELF-OSCILLATION)

Next Slide

Return to Index

Previous Slide

SUBHARMONIC OSCILLATION

- CAN TRANSLATE FROM ONE FREQUENCY REGIME
  TO ANOTHER.
- E.G., OPTICAL TO MICROWAVE
  - HOWEVER, MUST PRESERVE HARMONICS
  - ONE HARMONIC INTERVAL IS ESSENTIAL
- KAZNACHEYEV'S CYTOPATHOGENIC EFFECT:
  - QUARTZ PASSES ONE HARMONIC INTERVAL.
  - WINDOW GLASS DOES NOT.
- WHITTAKER'S INFOLDED EM STRUCTURE
- DEVYATKOV'S INFORMATION CONTENT OF THE FIELD.
- MULTIPLE PHASELOCKED WOODPECKER PULSE REPETITION RATES.

Next Slide

Return to Index

Previous Slide

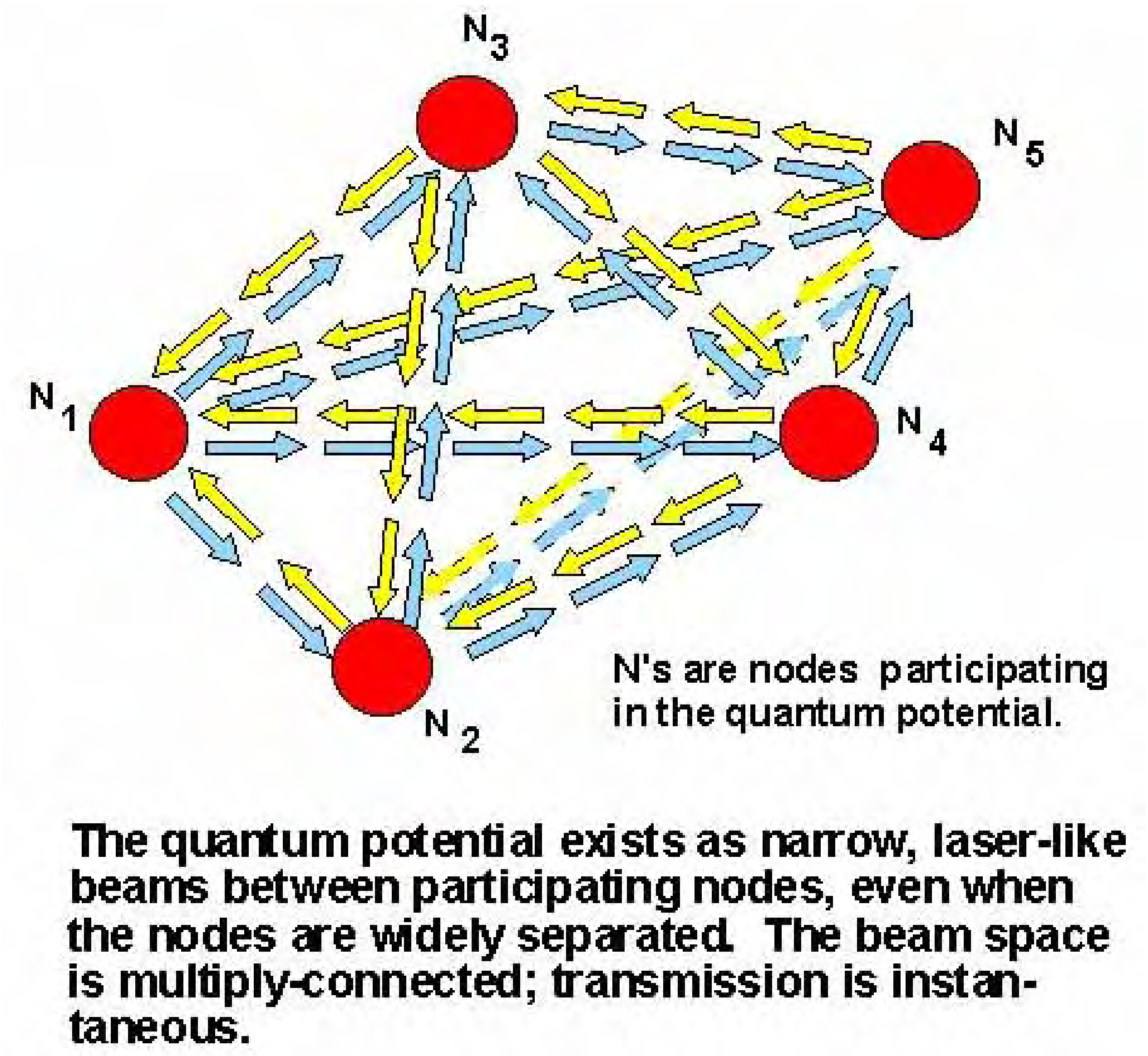

N's are nodes participating in the quantum potential.

The quantum potential exists as narrow, laser-like beams between participating nodes, even when the nodes are widely separated. The beam space is multiply-connected; transmission is instantaneous.

Next Slide

Return to Index

Previous Slide

QUANTUM POTENTIAL CHARACTERISTICS

- **NO POINT SOURCE**
- **QP BETWEEN TWO PARTICLES**
  - INTERACTION DOES NOT VANISH AS SPATIAL SEPARATION BECOMES VERY LARGE

  - INSTANTANEOUS CONNECTION

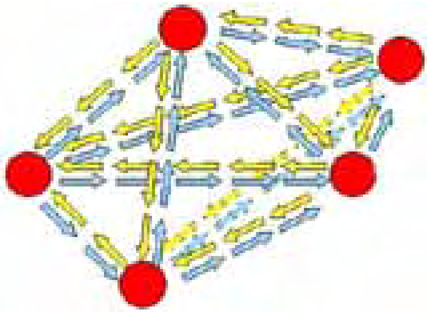

[Next Slide]

[Return to Index]

[Previous Slide]

- **DEPENDS ON QUANTUM STATE OF SYSTEM AS A WHOLE**

- **SYSTEM PARTS CAN BE GREATLY SEPARATED**

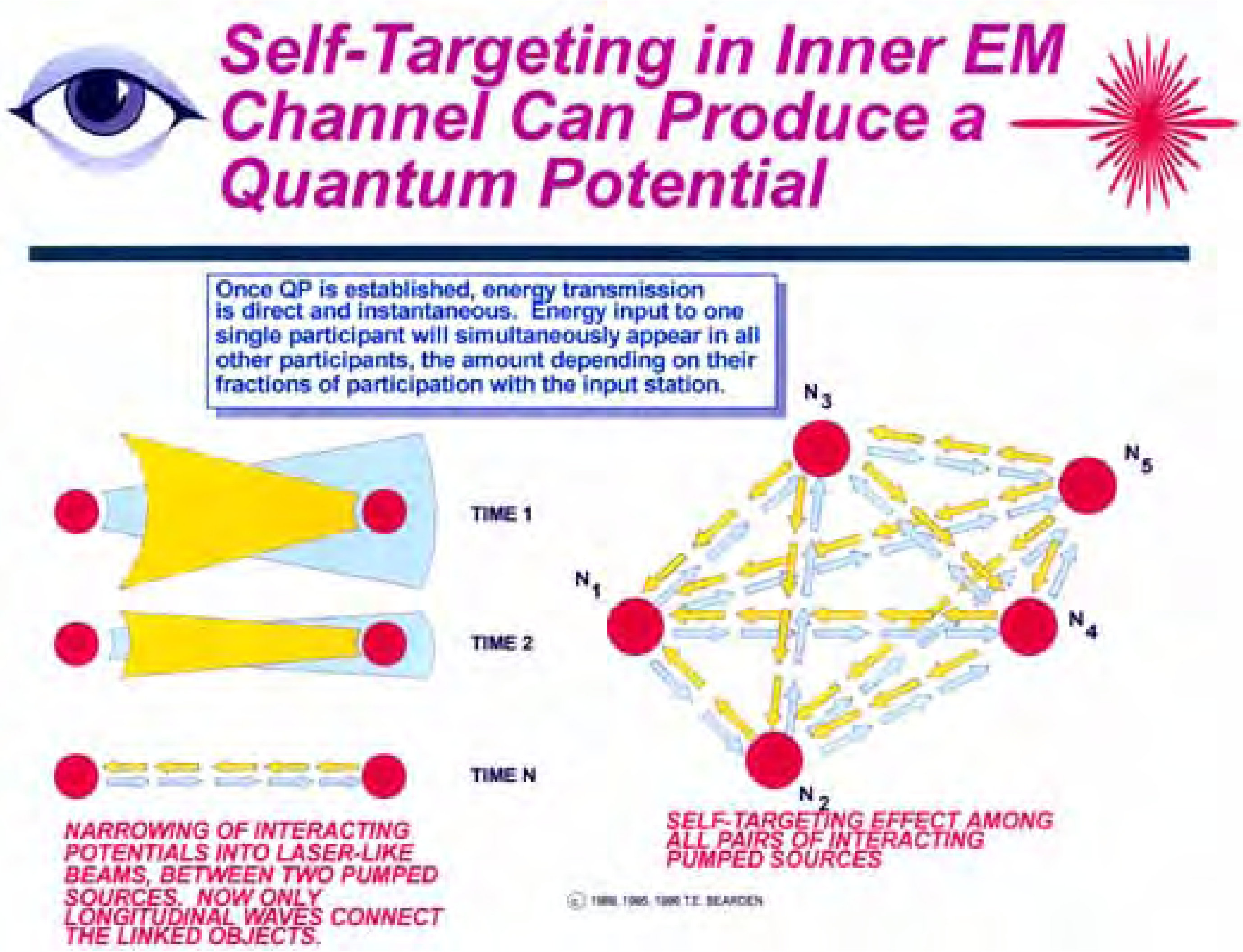

Self-Targeting in Inner EM Channel Can Produce a Quantum Potential

Once QP is established, energy transmission is direct and instantaneous. Energy input to one single participant will simultaneously appear in all other participants, the amount depending on their fractions of participation with the input station.

NARROWING OF INTERACTING POTENTIALS INTO LASER-LIKE BEAMS, BETWEEN TWO PUMPED SOURCES. NOW ONLY LONGITUDINAL WAVES CONNECT THE LINKED OBJECTS.

SELF-TARGETING EFFECT AMONG ALL PAIRS OF INTERACTING PUMPED SOURCES

© 1989, 1995, 1996 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

Corrections For the Present EM Bioeffects Model

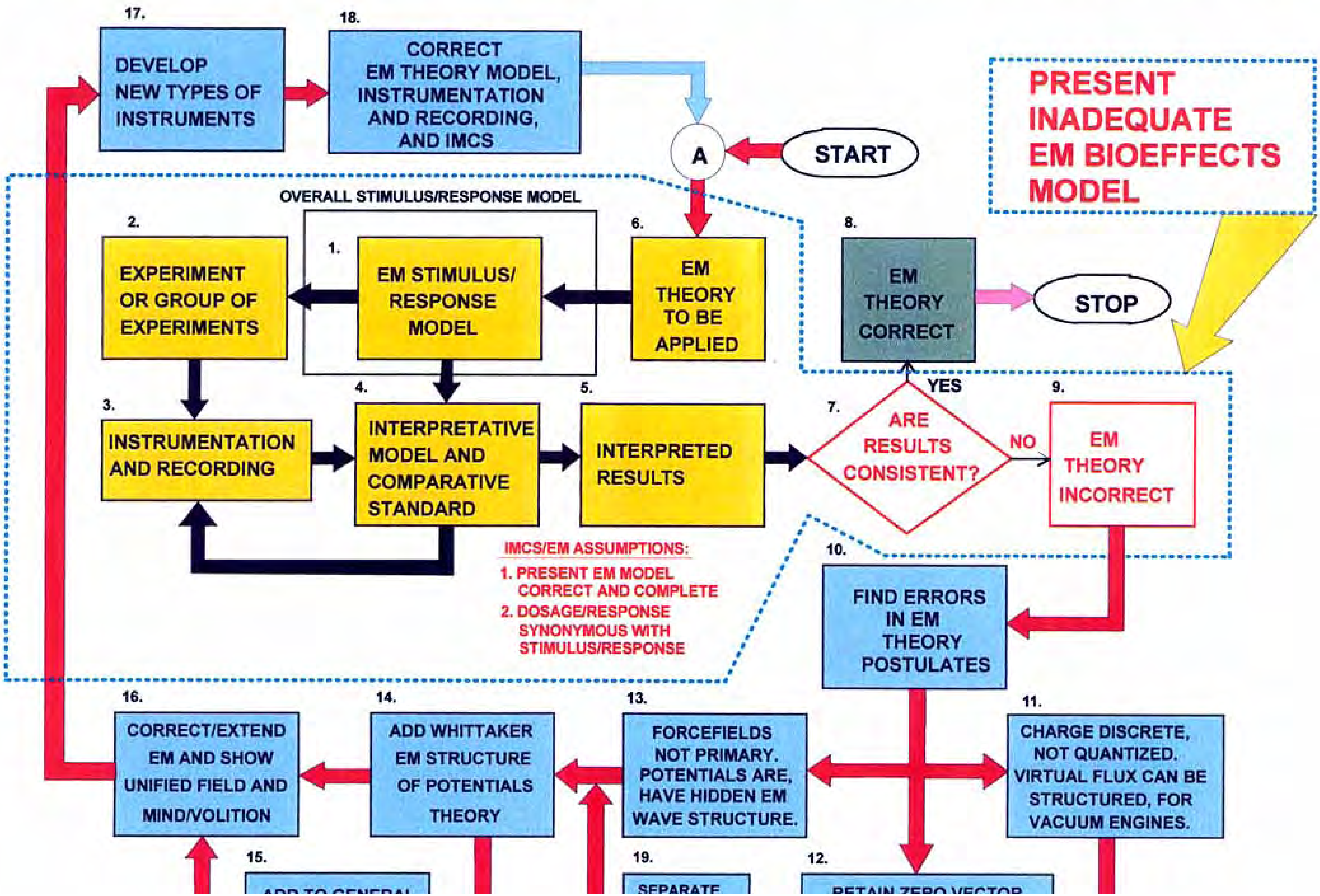

17. DEVELOP NEW TYPES OF INSTRUMENTS

18. CORRECT EM THEORY MODEL, INSTRUMENTATION AND RECORDING, AND IMCS

A

START

PRESENT INADEQUATE EM BIOEFFECTS MODEL

OVERALL STIMULUS/RESPONSE MODEL

1. EM STIMULUS/ RESPONSE MODEL

2. EXPERIMENT OR GROUP OF EXPERIMENTS

3. INSTRUMENTATION AND RECORDING

4. INTERPRETATIVE MODEL AND COMPARATIVE STANDARD

5. INTERPRETED RESULTS

6. EM THEORY TO BE APPLIED

7. ARE RESULTS CONSISTENT?

YES

8. EM THEORY CORRECT

STOP

NO

9. EM THEORY INCORRECT

IMCS/EM ASSUMPTIONS:

1. PRESENT EM MODEL CORRECT AND COMPLETE
2. DOSAGE/RESPONSE SYNONYMOUS WITH STIMULUS/RESPONSE

10. FIND ERRORS IN EM THEORY POSTULATES

11. CHARGE DISCRETE, NOT QUANTIZED. VIRTUAL FLUX CAN BE STRUCTURED, FOR VACUUM ENGINES.

13. FORCEFIELDS NOT PRIMARY. POTENTIALS ARE, HAVE HIDDEN EM WAVE STRUCTURE.

14. ADD WHITTAKER EM STRUCTURE OF POTENTIALS THEORY

16. CORRECT/EXTEND EM AND SHOW UNIFIED FIELD AND MIND/VOLITION

15. ADD TO GENERAL

19. SEPARATE

12. RETAIN ZERO VECTOR

[Next Slide](#)

[Return to Index](#)

[Previous Slide](#)

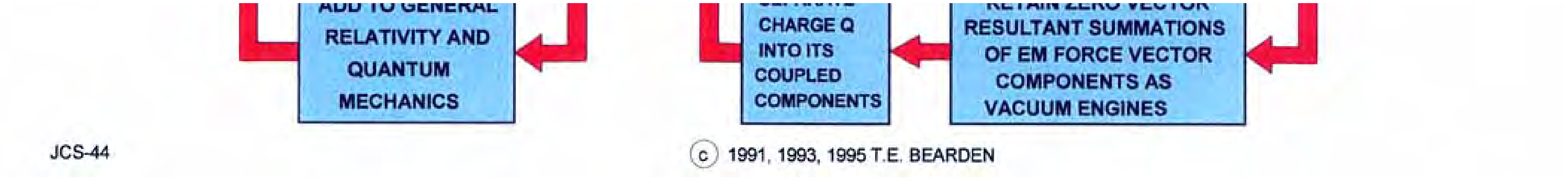
ADD TO GENERAL
RELATIVITY AND
QUANTUM
MECHANICS
CHARGE Q
INTO ITS
COUPLED
COMPONENTS
RESULTANT SUMMATIONS
OF EM FORCE VECTOR
COMPONENTS AS
VACUUM ENGINES
JCS-44

© 1991, 1993, 1995 T.E. BEARDEN

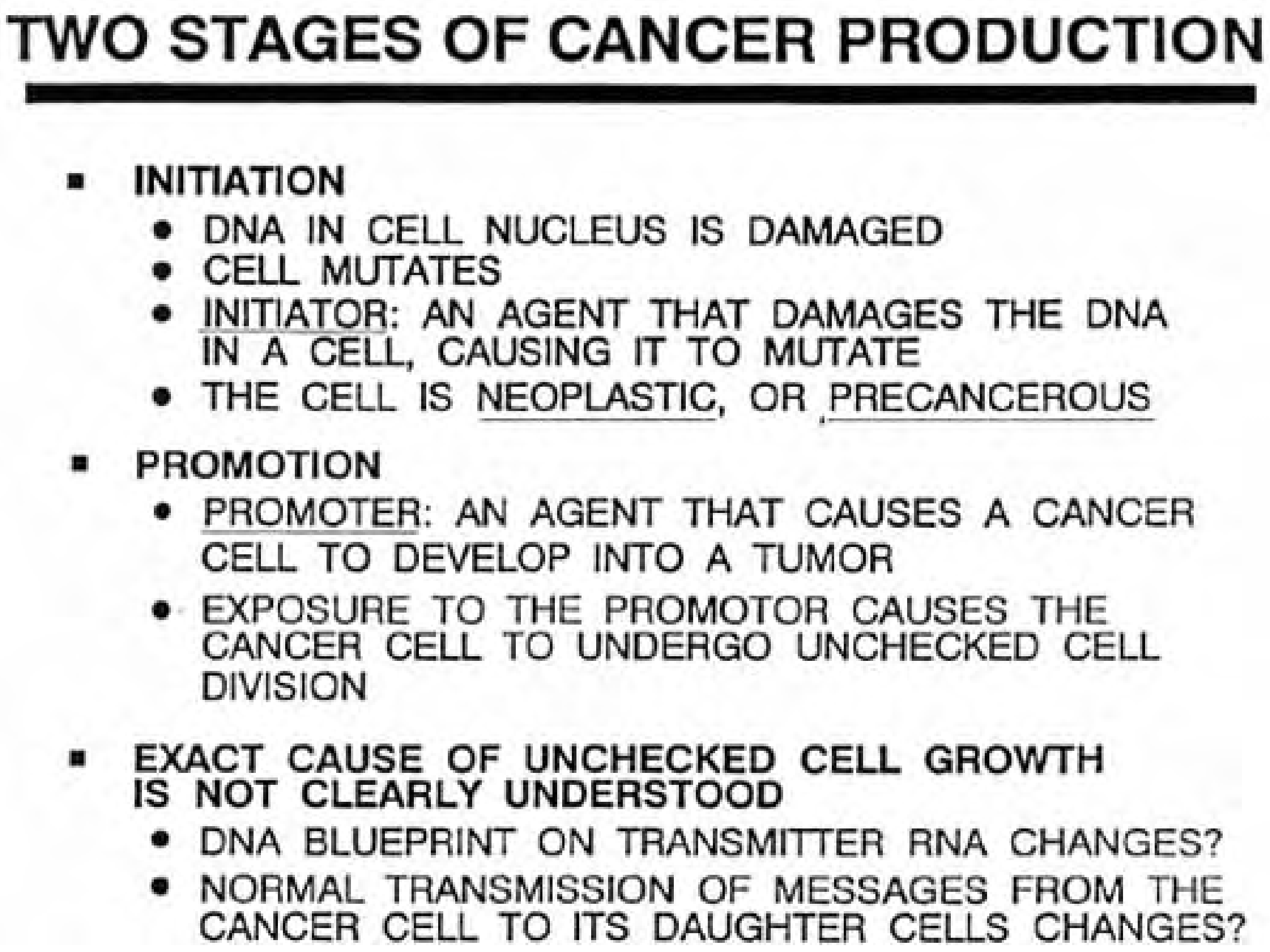

TWO STAGES OF CANCER PRODUCTION

- **INITIATION**
  - DNA IN CELL NUCLEUS IS DAMAGED
  - CELL MUTATES
  - INITIATOR: AN AGENT THAT DAMAGES THE DNA IN A CELL, CAUSING IT TO MUTATE
  - THE CELL IS NEOPLASTIC, OR PRECANCEROUS
- **PROMOTION**
  - PROMOTER: AN AGENT THAT CAUSES A CANCER CELL TO DEVELOP INTO A TUMOR
  - EXPOSURE TO THE PROMOTOR CAUSES THE CANCER CELL TO UNDERGO UNCHECKED CELL DIVISION
- **EXACT CAUSE OF UNCHECKED CELL GROWTH IS NOT CLEARLY UNDERSTOOD**
  - DNA BLUEPRINT ON TRANSMITTER RNA CHANGES?
  - NORMAL TRANSMISSION OF MESSAGES FROM THE CANCER CELL TO ITS DAUGHTER CELLS CHANGES?

Next Slide

Return to Index

Previous Slide

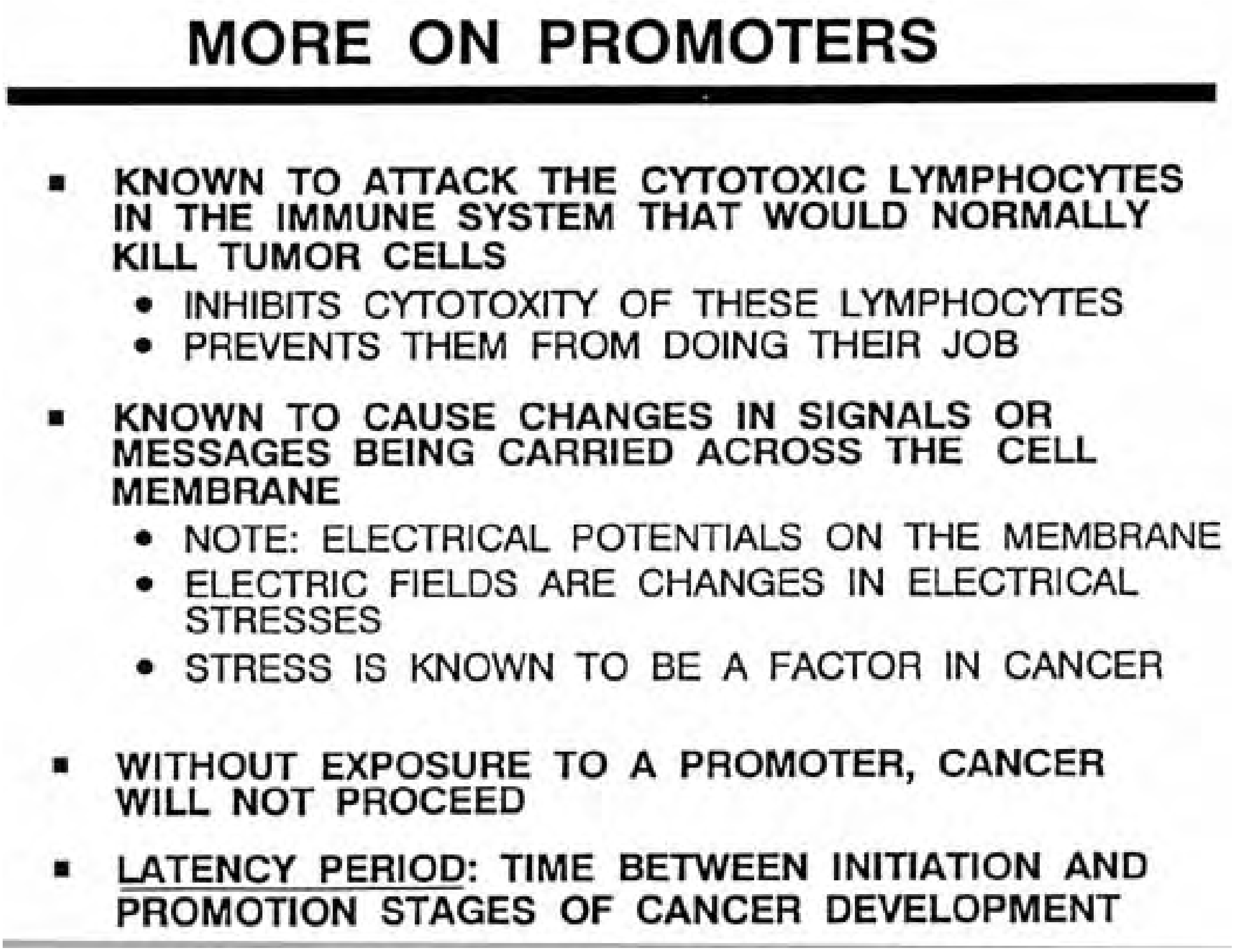

MORE ON PROMOTERS

- KNOWN TO ATTACK THE CYTOTOXIC LYMPHOCYTES IN THE IMMUNE SYSTEM THAT WOULD NORMALLY KILL TUMOR CELLS
  - INHIBITS CYTOTOXITY OF THESE LYMPHOCYTES
  - PREVENTS THEM FROM DOING THEIR JOB
- KNOWN TO CAUSE CHANGES IN SIGNALS OR MESSAGES BEING CARRIED ACROSS THE CELL MEMBRANE
  - NOTE: ELECTRICAL POTENTIALS ON THE MEMBRANE
  - ELECTRIC FIELDS ARE CHANGES IN ELECTRICAL STRESSES
  - STRESS IS KNOWN TO BE A FACTOR IN CANCER
- WITHOUT EXPOSURE TO A PROMOTER, CANCER WILL NOT PROCEED
- LATENCY PERIOD: TIME BETWEEN INITIATION AND PROMOTION STAGES OF CANCER DEVELOPMENT

Next Slide

Return to Index

Previous Slide

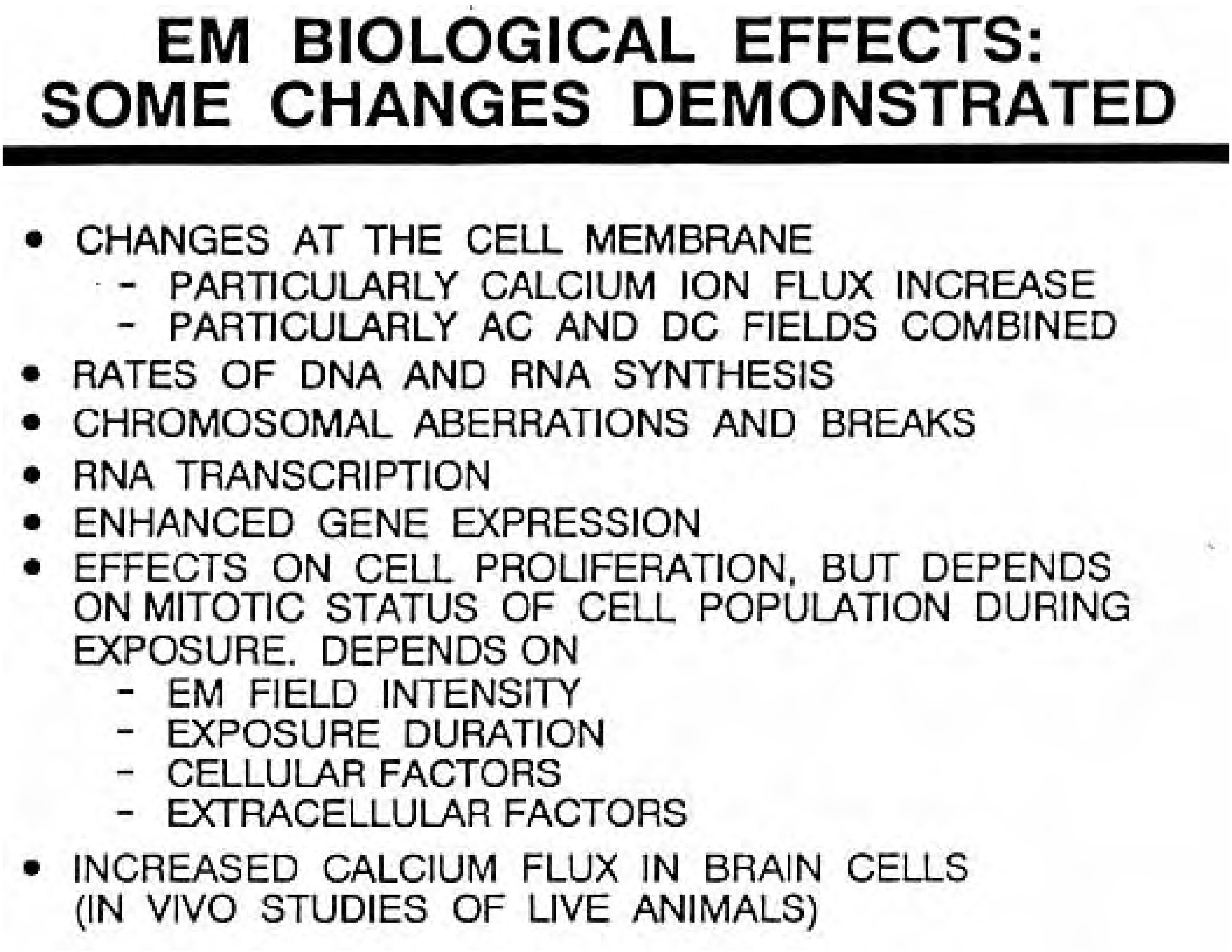

EM BIOLOGICAL EFFECTS: SOME CHANGES DEMONSTRATED

- CHANGES AT THE CELL MEMBRANE
  - PARTICULARLY CALCIUM ION FLUX INCREASE
  - PARTICULARLY AC AND DC FIELDS COMBINED
- RATES OF DNA AND RNA SYNTHESIS
- CHROMOSOMAL ABERRATIONS AND BREAKS
- RNA TRANSCRIPTION
- ENHANCED GENE EXPRESSION
- EFFECTS ON CELL PROLIFERATION, BUT DEPENDS ON MITOTIC STATUS OF CELL POPULATION DURING EXPOSURE. DEPENDS ON
  - EM FIELD INTENSITY
  - EXPOSURE DURATION
  - CELLULAR FACTORS
  - EXTRACELLULAR FACTORS
- INCREASED CALCIUM FLUX IN BRAIN CELLS (IN VIVO STUDIES OF LIVE ANIMALS)

Next Slide

Return to Index

Previous Slide

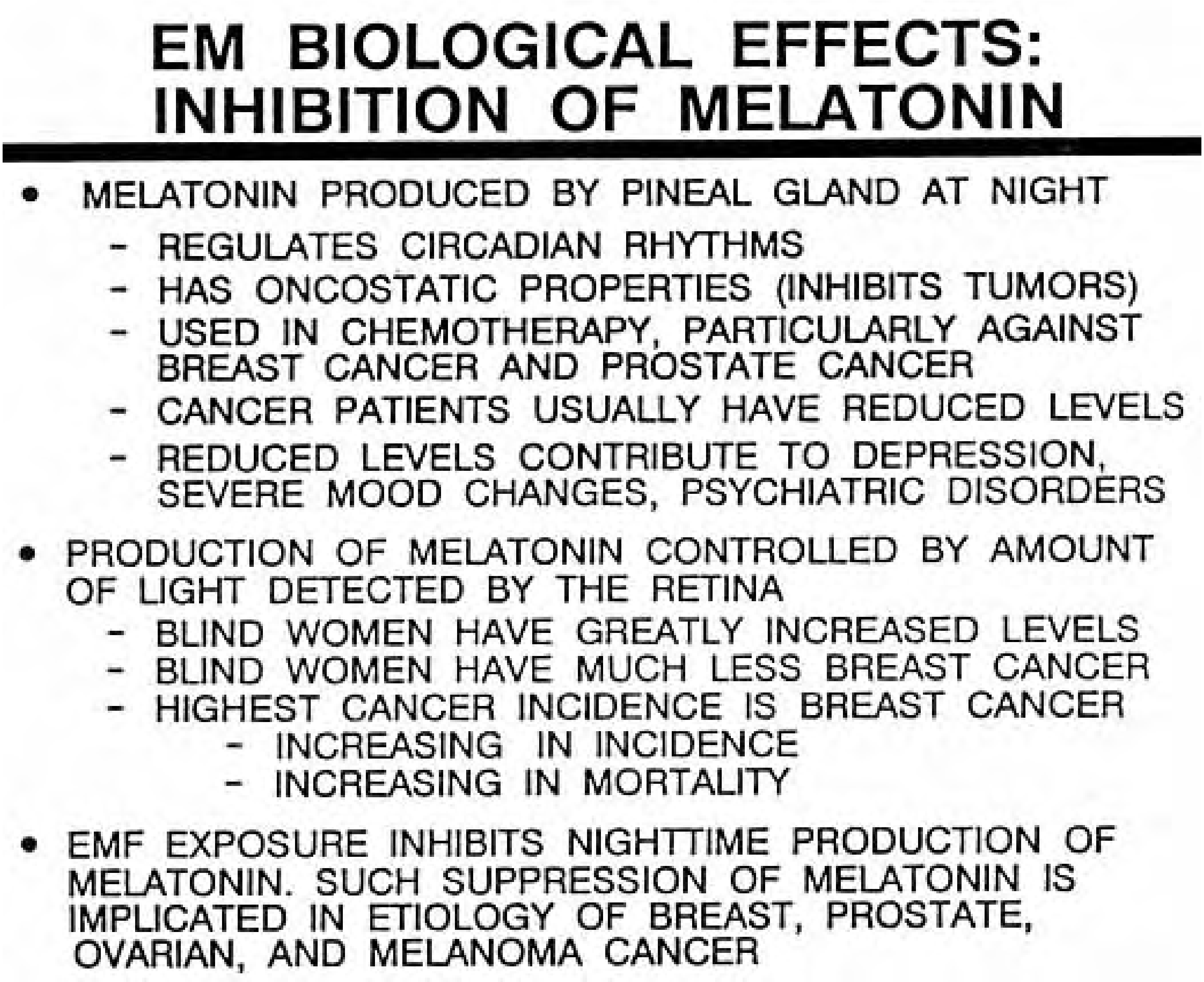

EM BIOLOGICAL EFFECTS: INHIBITION OF MELATONIN

- MELATONIN PRODUCED BY PINEAL GLAND AT NIGHT
  - REGULATES CIRCADIAN RHYTHMS
  - HAS ONCOSTATIC PROPERTIES (INHIBITS TUMORS)
  - USED IN CHEMOTHERAPY, PARTICULARLY AGAINST BREAST CANCER AND PROSTATE CANCER
  - CANCER PATIENTS USUALLY HAVE REDUCED LEVELS
  - REDUCED LEVELS CONTRIBUTE TO DEPRESSION, SEVERE MOOD CHANGES, PSYCHIATRIC DISORDERS
- PRODUCTION OF MELATONIN CONTROLLED BY AMOUNT OF LIGHT DETECTED BY THE RETINA
  - BLIND WOMEN HAVE GREATLY INCREASED LEVELS
  - BLIND WOMEN HAVE MUCH LESS BREAST CANCER
  - HIGHEST CANCER INCIDENCE IS BREAST CANCER
    - INCREASING IN INCIDENCE
    - INCREASING IN MORTALITY
- EMF EXPOSURE INHIBITS NIGHTTIME PRODUCTION OF MELATONIN. SUCH SUPPRESSION OF MELATONIN IS IMPLICATED IN ETIOLOGY OF BREAST, PROSTATE, OVARIAN, AND MELANOMA CANCER

Next Slide

Return to Index

Previous Slide

OTHER EM BIOLOGICAL EFFECTS

- **MAGNETIC FIELD INDUCING OF BIRTH DEFECTS IN CHICKS**
- **MORE CONTROVERSIAL EPIDEMIOLOGY STUDIES FOUND**
  CORRELATIONS BETWEEN EMF AND
  - BRAIN TUMORS
  - MISCARRIAGES IN PREGNANT FEMALES USING VDTS
  - BIRTH DEFECTS
  - INCIDENCE OF LEUKEMIA IN CHILDREN
- **BECKER SHOWED THAT PICOAMPERES OF CURRENT USED**
  IN EM-ASSISTED BONE HEALING CAUSED PROFOUND CELLULAR CHANGES
  - RED BLOOD CELLS DEDIFFERENTIATED
  - SHED HEMOGLOBIN AND GREW NUCLEI
  - REDIFFERENTIATED INTO CARTILAGE-HEALING CELLS
  - REDIFFERENTIATED INTO BONE-HEALING CELLS
  - THEN HEALED THE FRACTURES

Next Slide

Return to Index

Previous Slide

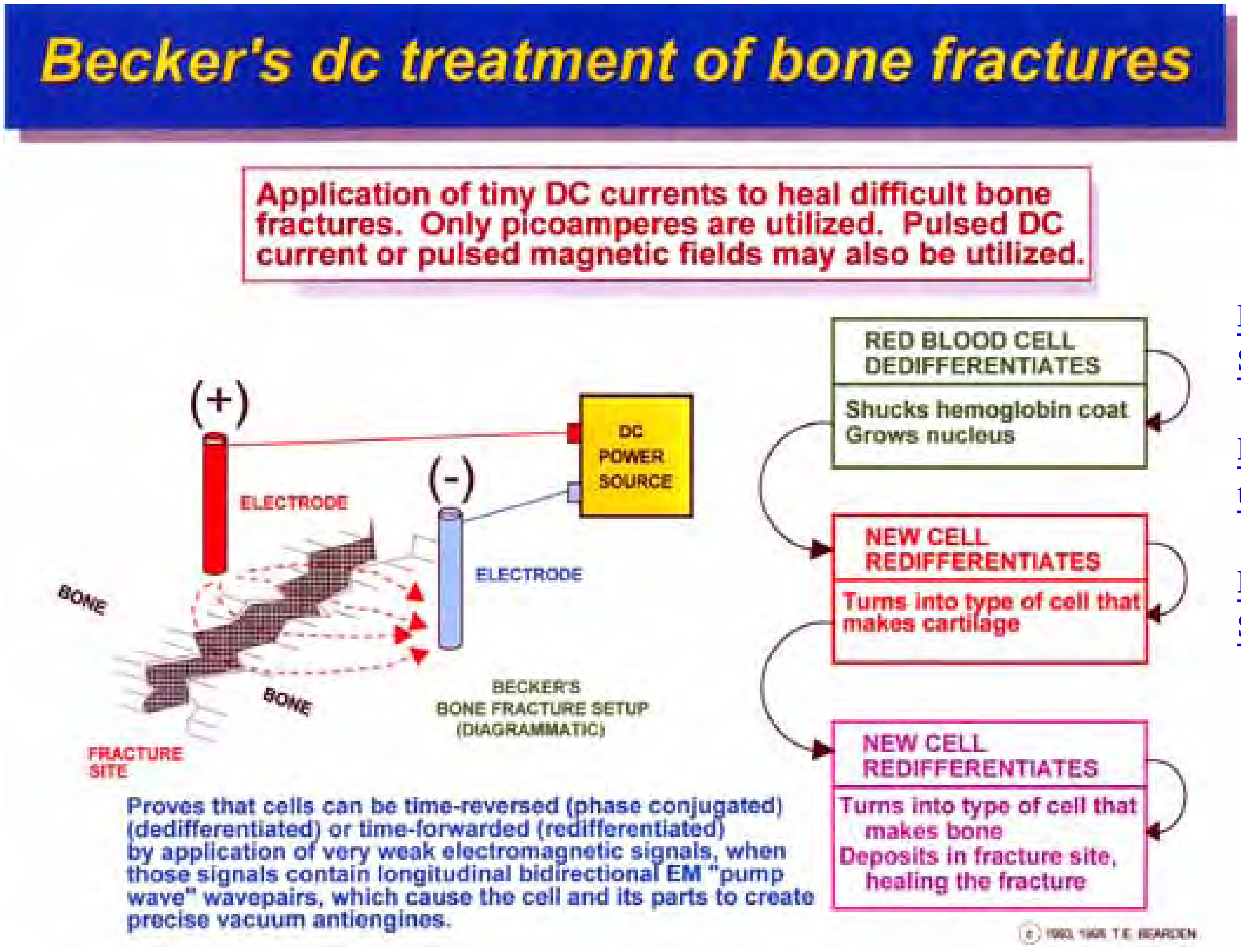

Becker's dc treatment of bone fractures

Application of tiny DC currents to heal difficult bone fractures. Only picoamperes are utilized. Pulsed DC current or pulsed magnetic fields may also be utilized.

(+)
ELECTRODE
(-)
DC POWER SOURCE
ELECTRODE
BONE
BONE
FRACTURE SITE

BECKER'S BONE FRACTURE SETUP (DIAGRAMMATIC)

Proves that cells can be time-reversed (phase conjugated) (dedifferentiated) or time-forwarded (redifferentiated) by application of very weak electromagnetic signals, when those signals contain longitudinal bidirectional EM "pump wave" wavepairs, which cause the cell and its parts to create precise vacuum antiengines.

RED BLOOD CELL DEDIFFERENTIATES

Shucks hemoglobin coat
Grows nucleus

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes cartilage

NEW CELL REDIFFERENTIATES

Turns into type of cell that makes bone
Deposits in fracture site, healing the fracture

© 1993, 1998 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

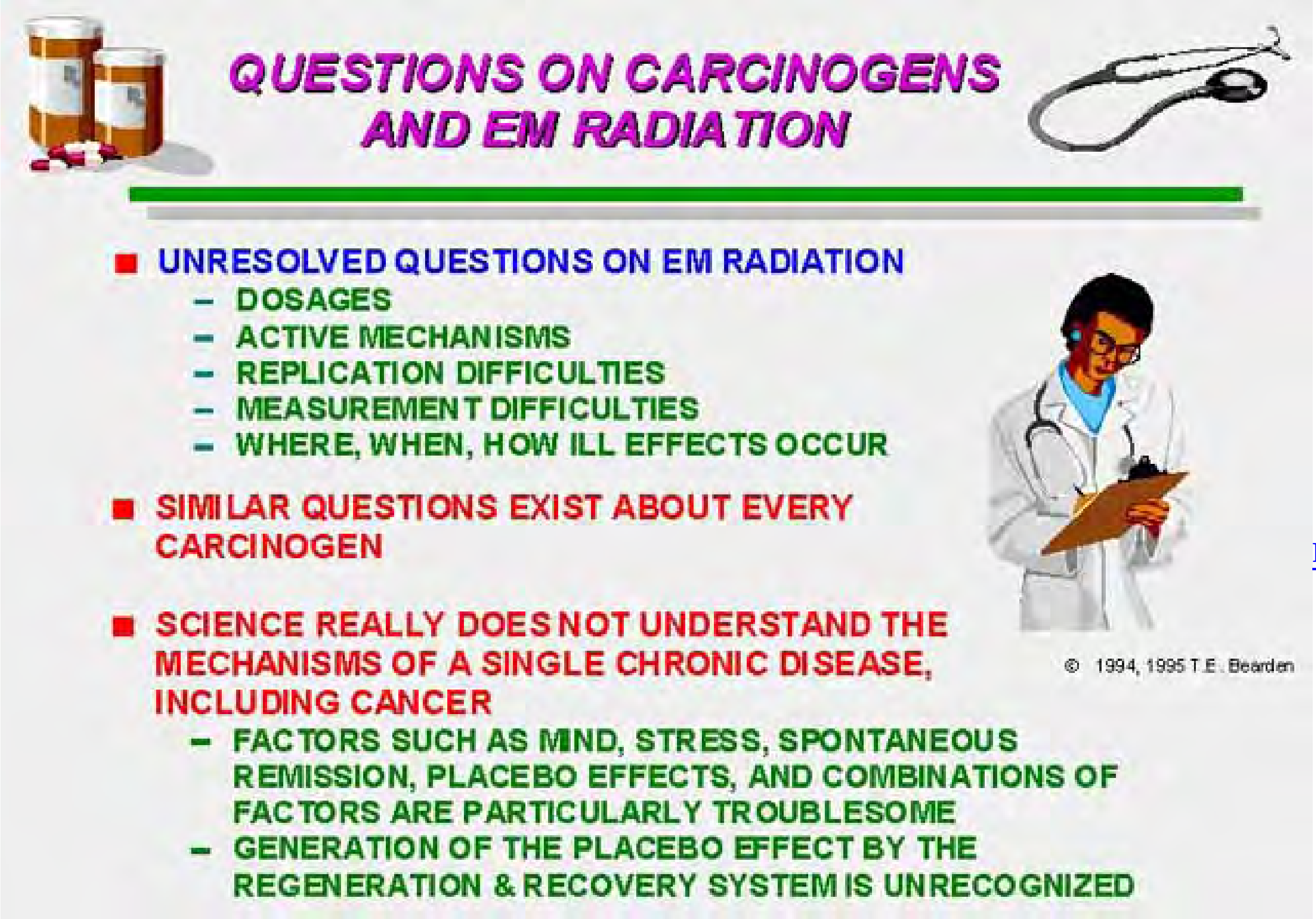

QUESTIONS ON CARCINOGENS AND EM RADIATION

- UNRESOLVED QUESTIONS ON EM RADIATION
  - DOSAGES
  - ACTIVE MECHANISMS
  - REPLICATION DIFFICULTIES
  - MEASUREMENT DIFFICULTIES
  - WHERE, WHEN, HOW ILL EFFECTS OCCUR
- SIMILAR QUESTIONS EXIST ABOUT EVERY CARCINOGEN
- SCIENCE REALLY DOES NOT UNDERSTAND THE MECHANISMS OF A SINGLE CHRONIC DISEASE, INCLUDING CANCER
  - FACTORS SUCH AS MIND, STRESS, SPONTANEOUS REMISSION, PLACEBO EFFECTS, AND COMBINATIONS OF FACTORS ARE PARTICULARLY TROUBLESOME
  - GENERATION OF THE PLACEBO EFFECT BY THE REGENERATION & RECOVERY SYSTEM IS UNRECOGNIZED

© 1994, 1995 T.E. Bearden

Next Slide

Return to Index

Previous Slide

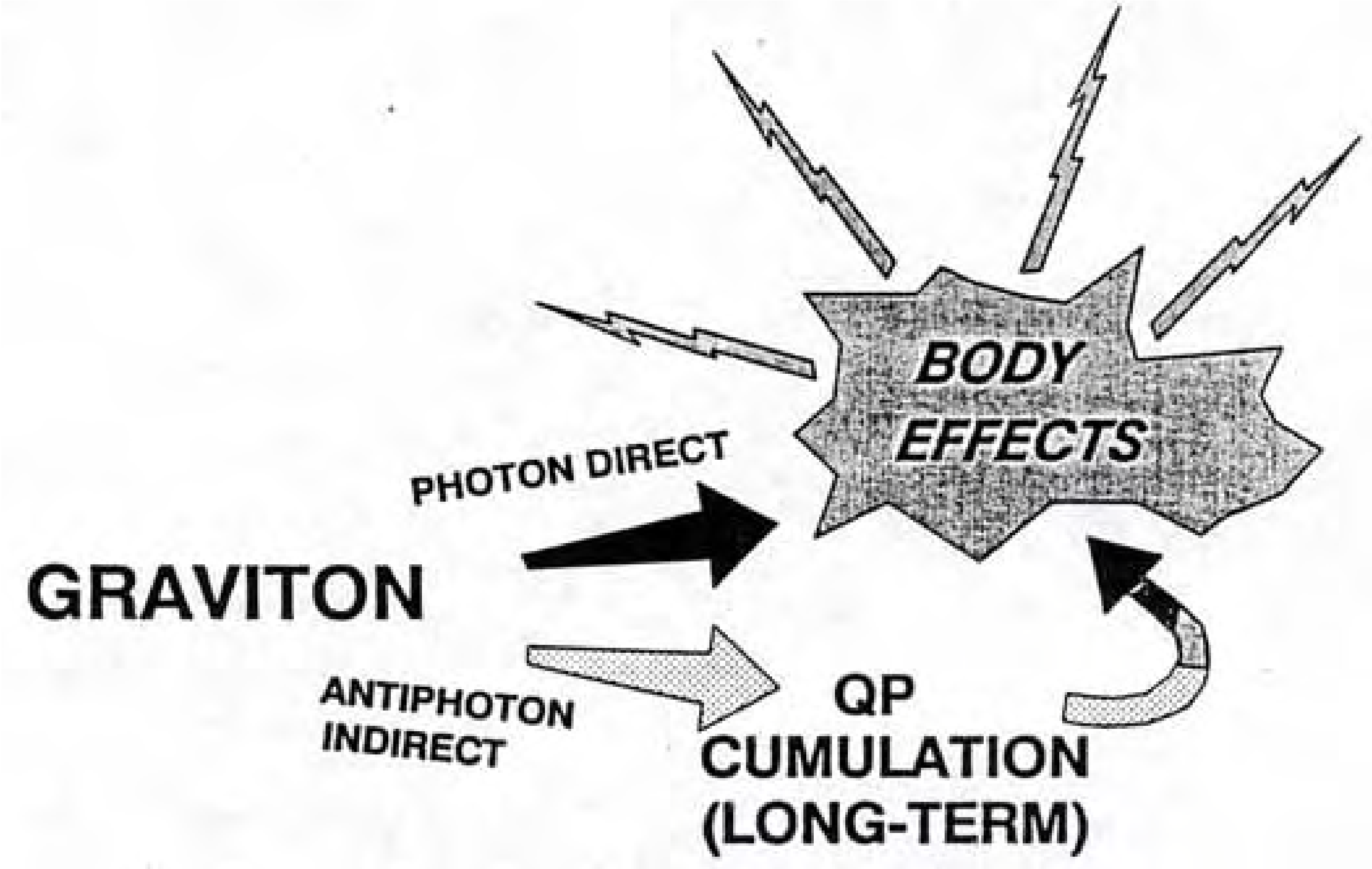
BODY EFFECTS
PHOTON DIRECT
GRAVITON
ANTIPHOTON INDIRECT
QP CUMULATION (LONG-TERM)

[Next Slide]
[Return to Index]
[Previous Slide]

CUMULATIVE FEEDBACK REACTION LOOP FOR INDUCTION OF LONG-TERM EFFECTS

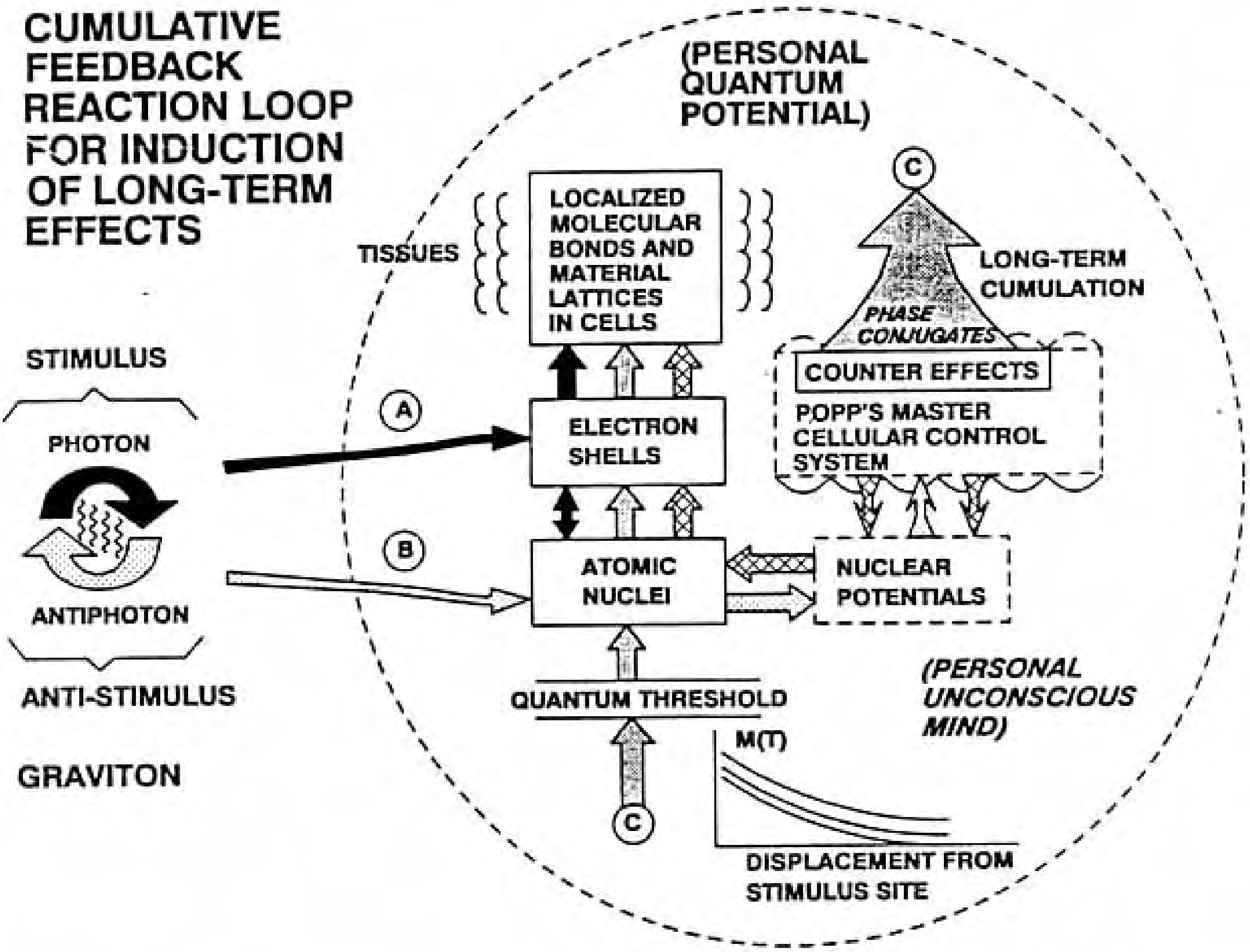
(PERSONAL QUANTUM POTENTIAL)
STIMULUS
PHOTON
ANTIPHOTON
ANTI-STIMULUS
GRAVITON
TISSUES
LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES IN CELLS
C
LONG-TERM CUMULATION
PHASE CONJUGATES
COUNTER EFFECTS
POPP'S MASTER CELLULAR CONTROL SYSTEM
A
ELECTRON SHELLS
B
ATOMIC NUCLEI
NUCLEAR POTENTIALS
QUANTUM THRESHOLD
C
M(T)
DISPLACEMENT FROM STIMULUS SITE
(PERSONAL UNCONSCIOUS MIND)

[Next Slide](#)

[Return to Index](#)

[Previous Slide](#)

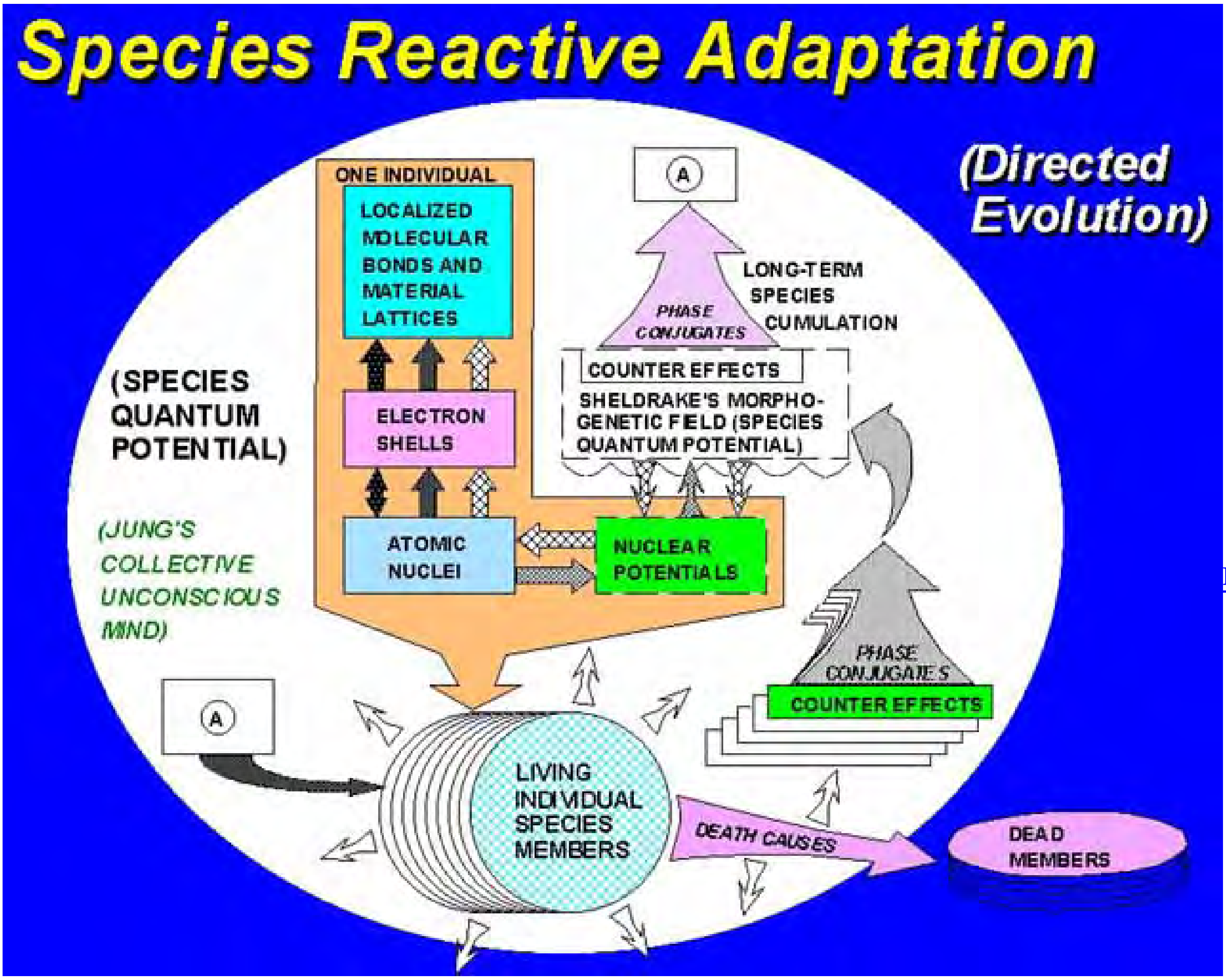

Species Reactive Adaptation

(Directed Evolution)

(SPECIES QUANTUM POTENTIAL)

(JUNG'S COLLECTIVE UNCONSCIOUS MIND)

ONE INDIVIDUAL

LOCALIZED MOLECULAR BONDS AND MATERIAL LATTICES

ELECTRON SHELLS

ATOMIC NUCLEI

NUCLEAR POTENTIALS

A

PHASE CONJUGATES

LONG-TERM SPECIES CUMULATION

COUNTER EFFECTS

SHELDRAKE'S MORPHOGENETIC FIELD (SPECIES QUANTUM POTENTIAL)

PHASE CONJUGATES

COUNTER EFFECTS

A

LIVING INDIVIDUAL SPECIES MEMBERS

DEATH CAUSES

DEAD MEMBERS

Next Slide

Return to Index

Previous Slide

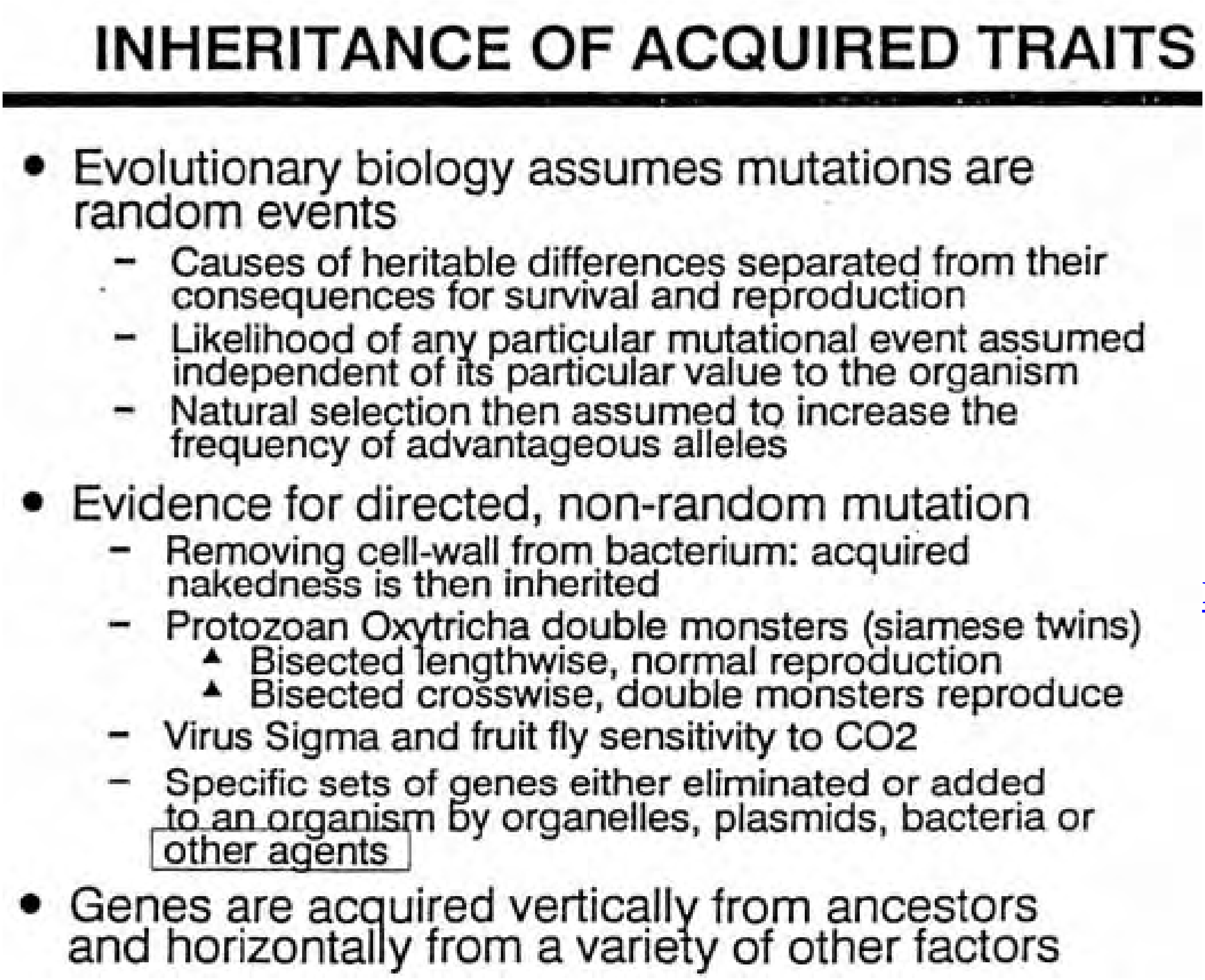

INHERITANCE OF ACQUIRED TRAITS

- Evolutionary biology assumes mutations are random events
  - Causes of heritable differences separated from their consequences for survival and reproduction
  - Likelihood of any particular mutational event assumed independent of its particular value to the organism
  - Natural selection then assumed to increase the frequency of advantageous alleles
- Evidence for directed, non-random mutation
  - Removing cell-wall from bacterium: acquired nakedness is then inherited
  - Protozoan Oxytricha double monsters (siamese twins)
    - ▲ Bisected lengthwise, normal reproduction
    - ▲ Bisected crosswise, double monsters reproduce
  - Virus Sigma and fruit fly sensitivity to CO2
  - Specific sets of genes either eliminated or added to an organism by organelles, plasmids, bacteria or other agents
- Genes are acquired vertically from ancestors and horizontally from a variety of other factors

Next Slide

Return to Index

Previous Slide

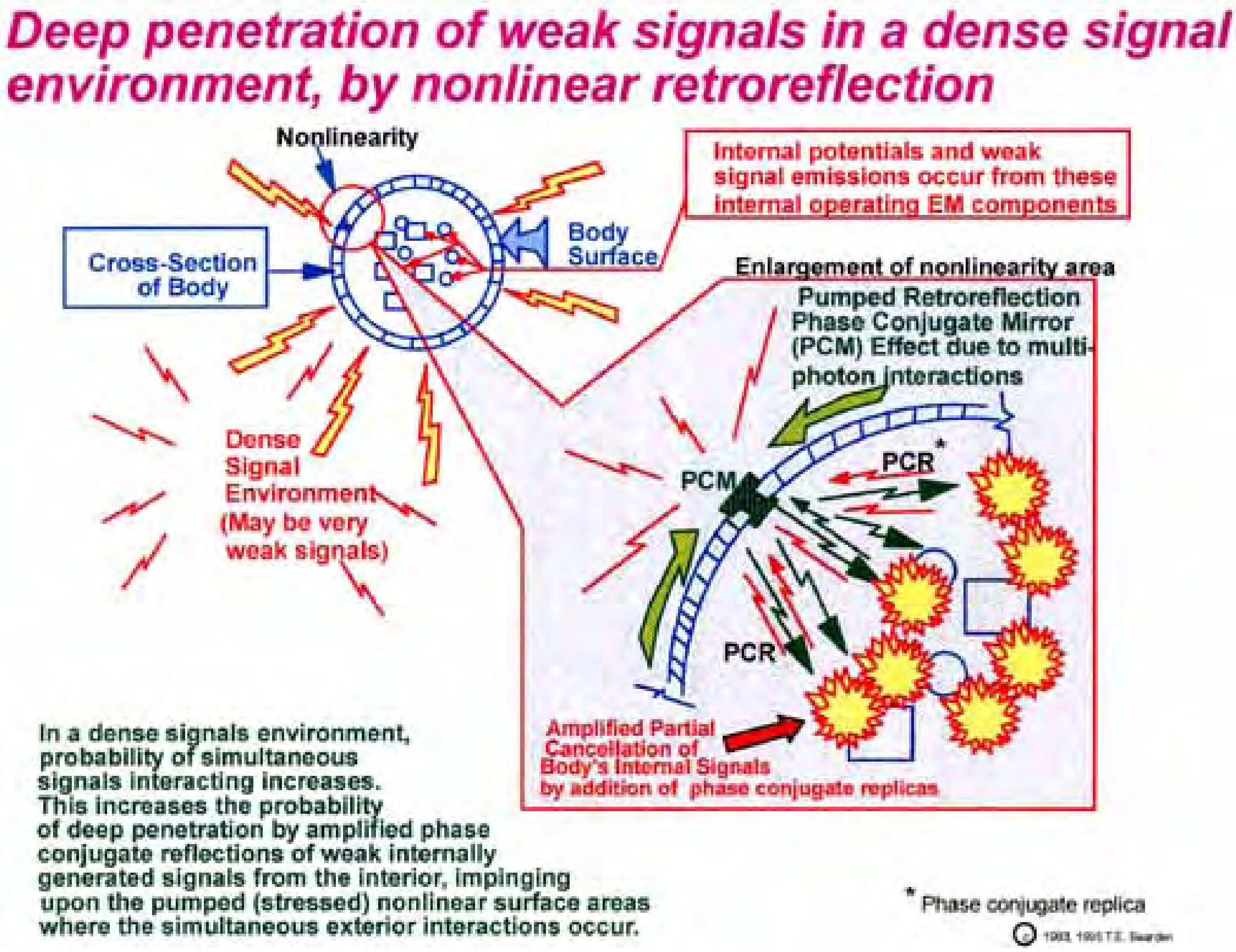

Deep penetration of weak signals in a dense signal environment, by nonlinear retroreflection

Nonlinearity
Cross-Section of Body
Body Surface
Dense Signal Environment (May be very weak signals)
Enlargement of nonlinearity area
PCM
PCR*
PCR

Internal potentials and weak signal emissions occur from these internal operating EM components

Pumped Retroreflection Phase Conjugate Mirror (PCM) Effect due to multi-photon interactions

Amplified Partial Cancellation of Body's Internal Signals by addition of phase conjugate replicas

In a dense signals environment, probability of simultaneous signals interacting increases. This increases the probability of deep penetration by amplified phase conjugate reflections of weak internally generated signals from the interior, impinging upon the pumped (stressed) nonlinear surface areas where the simultaneous exterior interactions occur.

* Phase conjugate replica

© 1993, 1995 T.E. Bearden

Next Slide

Return to Index

Previous Slide

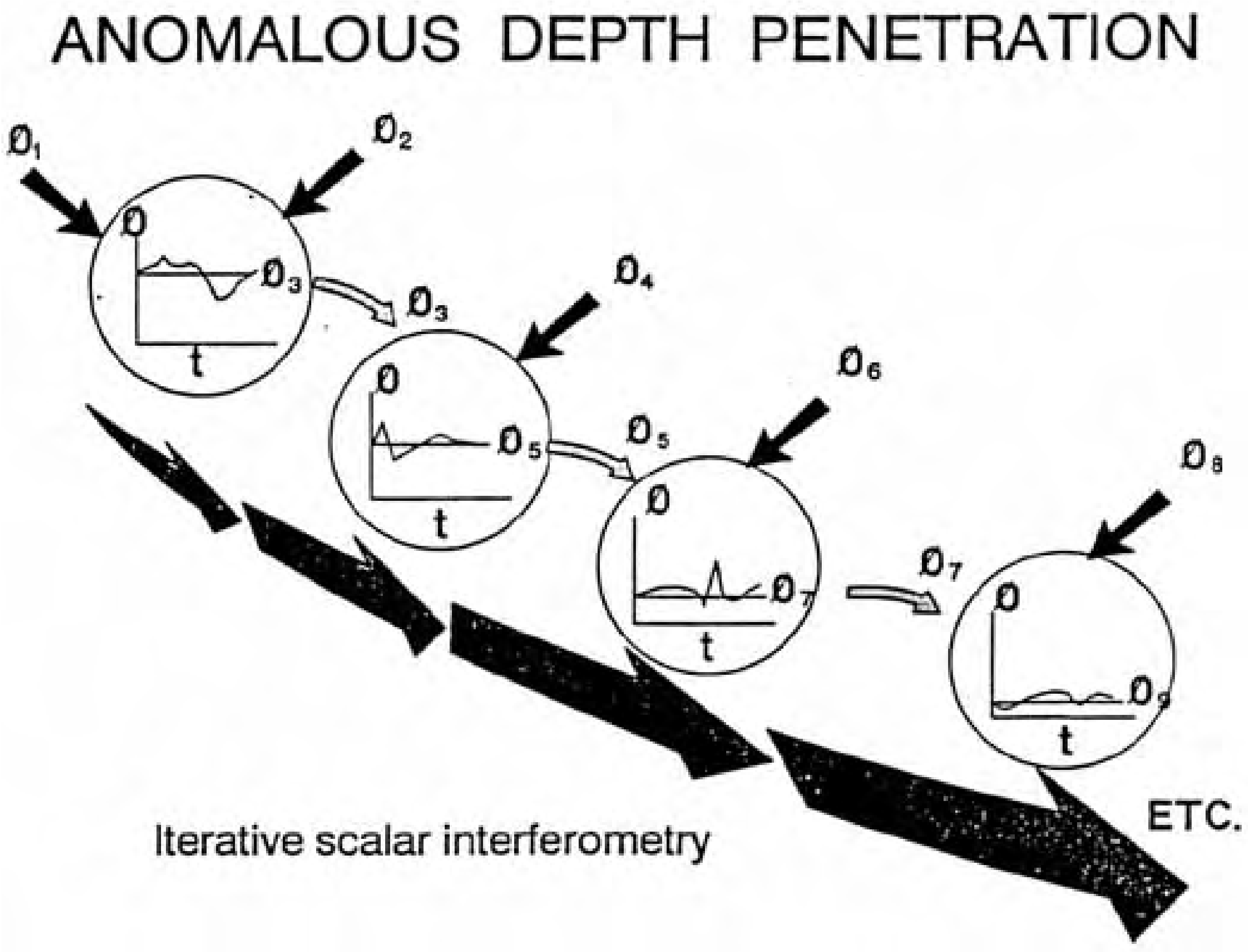

ANOMALOUS DEPTH PENETRATION

ϕ_1
ϕ_2
ϕ
ϕ_3
t
ϕ_3
ϕ_4
ϕ
ϕ_5
t
ϕ_5
ϕ_6
ϕ
ϕ_7
t
ϕ_7
ϕ_8
ϕ
ϕ_9
t
ETC.

Iterative scalar interferometry

Next Slide

Return to Index

Previous Slide

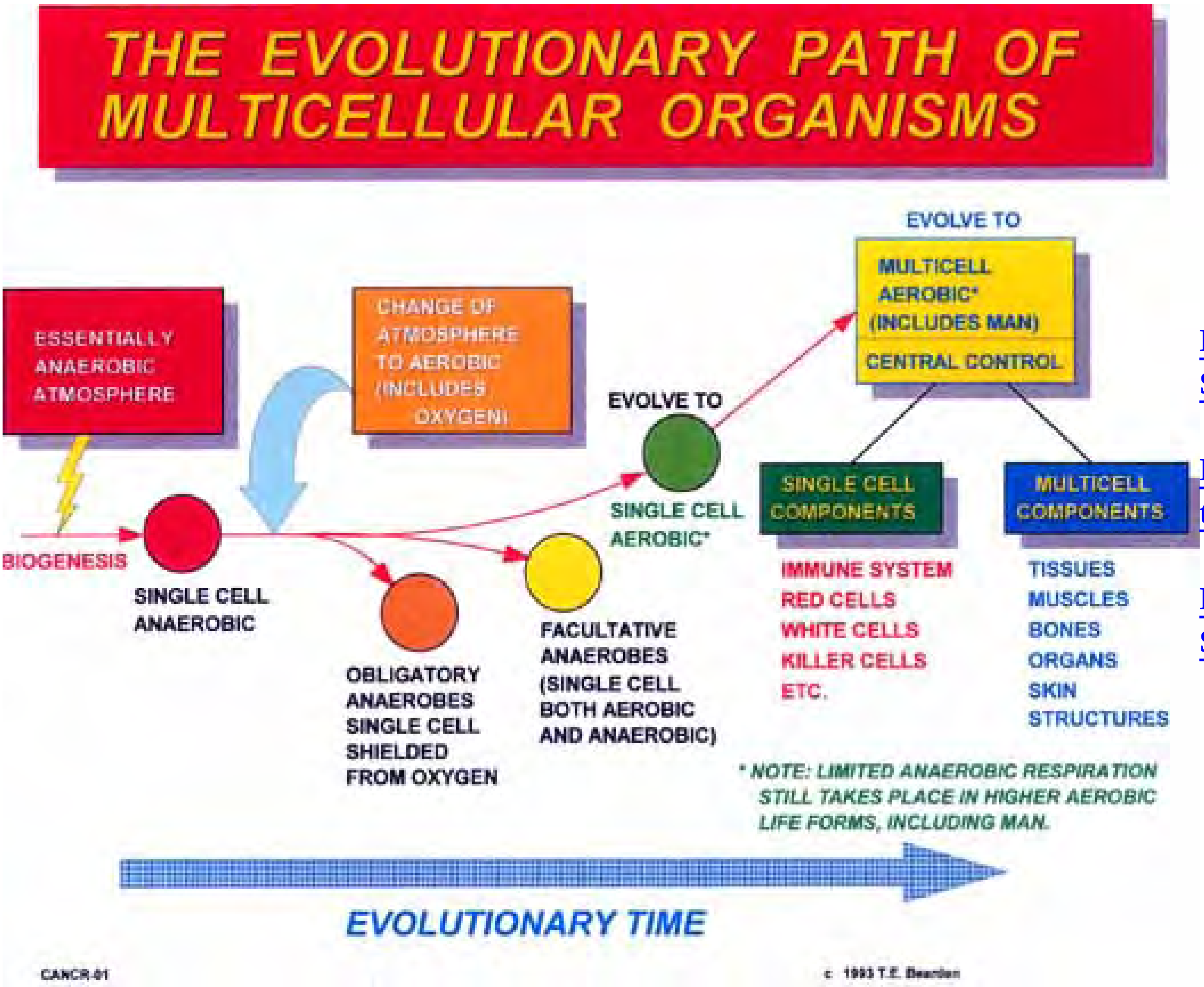

THE EVOLUTIONARY PATH OF MULTICELLULAR ORGANISMS

ESSENTIALLY ANAEROBIC ATMOSPHERE

CHANGE OF ATMOSPHERE TO AEROBIC (INCLUDES OXYGEN)

BIOGENESIS

SINGLE CELL ANAEROBIC

OBLIGATORY ANAEROBES SINGLE CELL SHIELDED FROM OXYGEN

FACULTATIVE ANAEROBES (SINGLE CELL BOTH AEROBIC AND ANAEROBIC)

EVOLVE TO

SINGLE CELL AEROBIC*

EVOLVE TO

MULTICELL AEROBIC* (INCLUDES MAN)

CENTRAL CONTROL

SINGLE CELL COMPONENTS

- IMMUNE SYSTEM
- RED CELLS
- WHITE CELLS
- KILLER CELLS
- ETC.

MULTICELL COMPONENTS

- TISSUES
- MUSCLES
- BONES
- ORGANS
- SKIN
- STRUCTURES

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

EVOLUTIONARY TIME

CANCR-01

c 1993 T.E. Bearden

[Next Slide]
[Return to Index]
[Previous Slide]

DEVELOPMENT OF MULTI-CELLED ORGANISMS

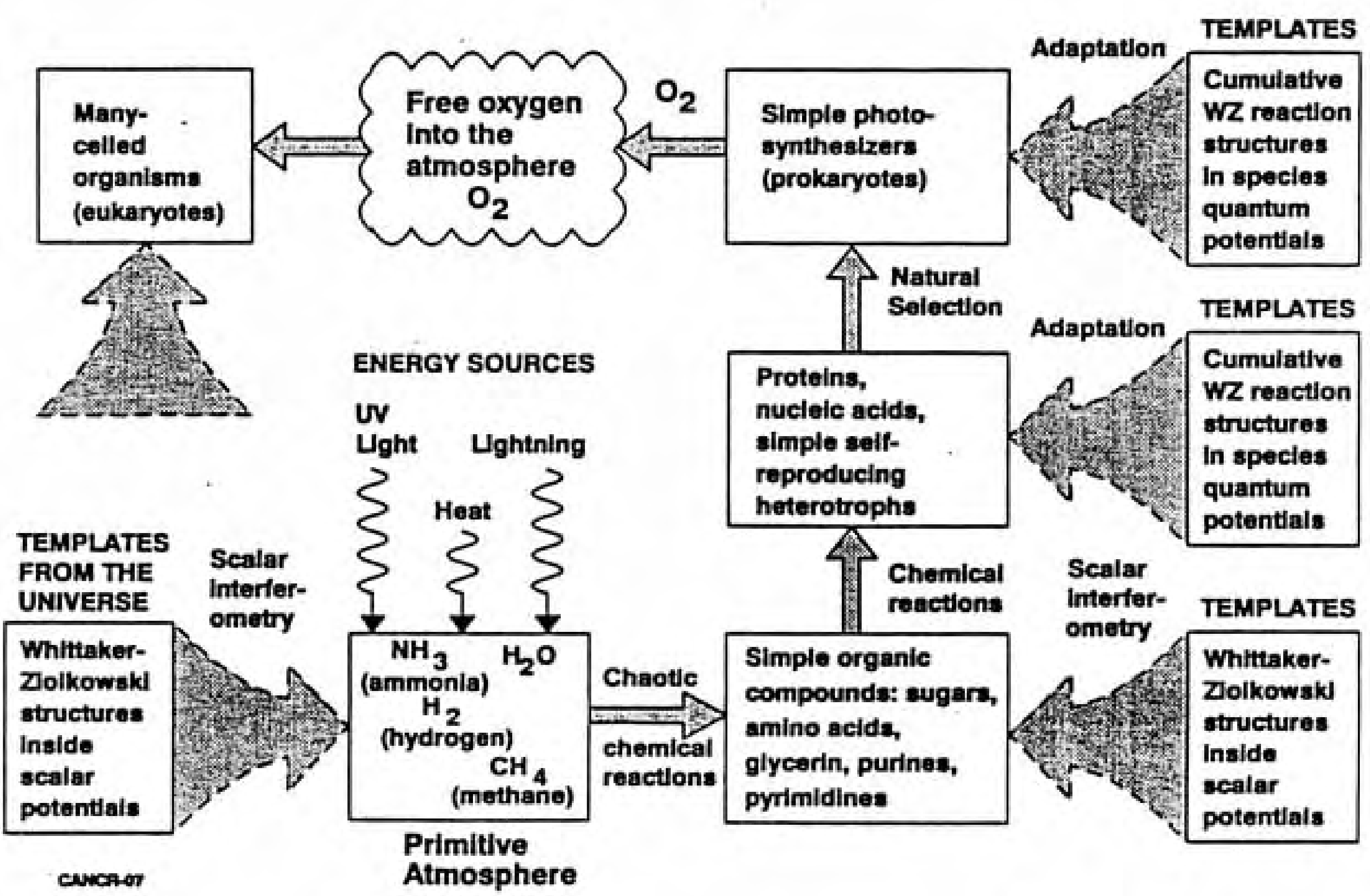

Many-celled organisms (eukaryotes)

Free oxygen into the atmosphere O_2

O_2

Simple photo-synthesizers (prokaryotes)

Adaptation

TEMPLATES

Cumulative WZ reaction structures in species quantum potentials

Natural Selection

Proteins, nucleic acids, simple self-reproducing heterotrophs

Adaptation

TEMPLATES

Cumulative WZ reaction structures in species quantum potentials

ENERGY SOURCES

UV Light

Heat

Lightning

TEMPLATES FROM THE UNIVERSE

Whittaker-Ziolkowski structures inside scalar potentials

Scalar interfer-ometry

NH_3 (ammonia)

H_2O

H_2 (hydrogen)

CH_4 (methane)

Primitive Atmosphere

Chaotic chemical reactions

Chemical reactions

Simple organic compounds: sugars, amino acids, glycerin, purines, pyrimidines

Scalar interfer-ometry

TEMPLATES

Whittaker-Ziolkowski structures inside scalar potentials

CANCR-07

Next Slide

Return to Index

Previous Slide

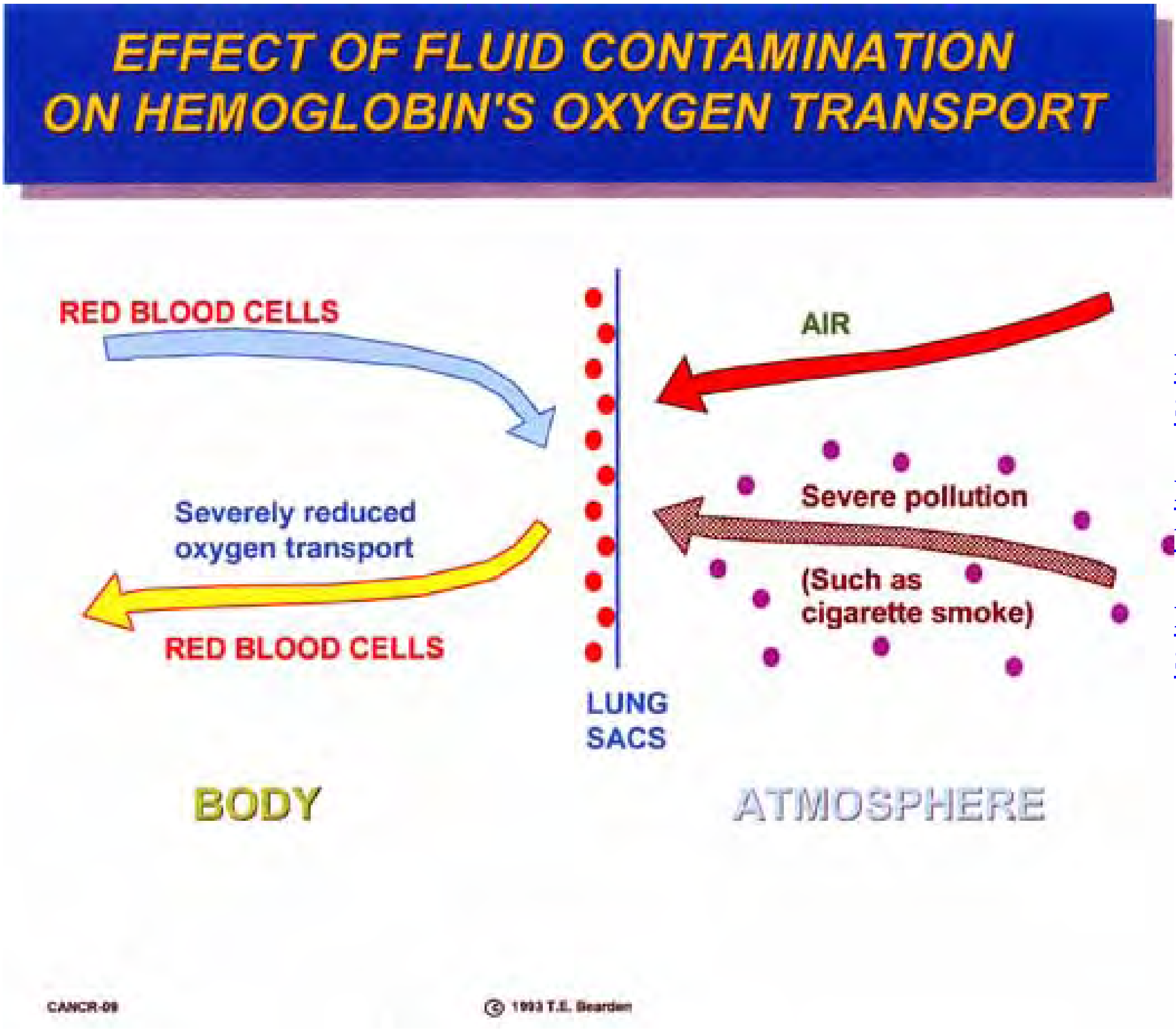

EFFECT OF FLUID CONTAMINATION ON HEMOGLOBIN'S OXYGEN TRANSPORT

RED BLOOD CELLS

AIR

Severely reduced
oxygen transport

Severe pollution

(Such as
cigarette smoke)

RED BLOOD CELLS

LUNG
SACS

BODY

ATMOSPHERE

CANCR-09

© 1993 T.E. Bearden

[Next Slide](#)

[Return to Index](#)

[Previous Slide](#)

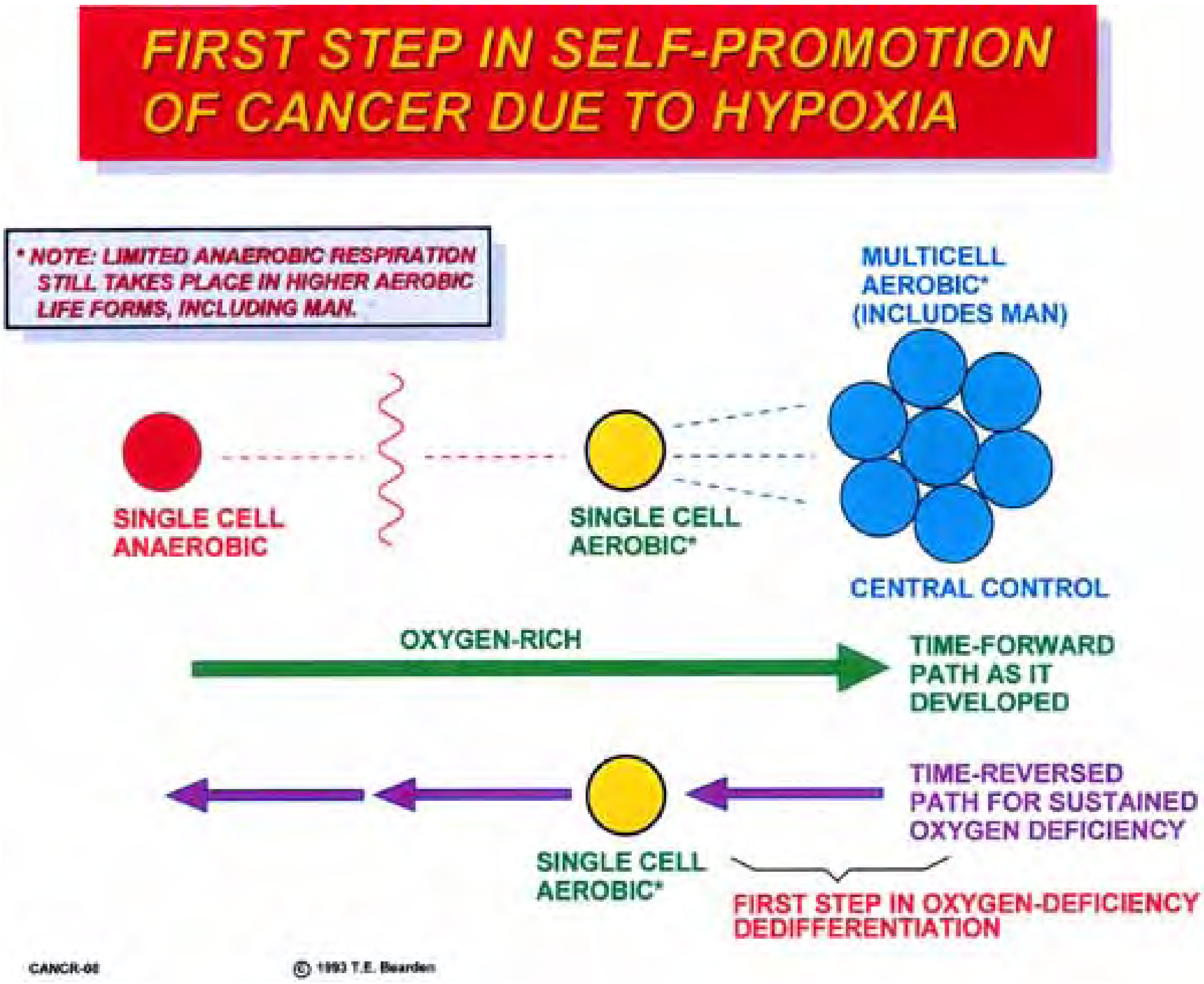

FIRST STEP IN SELF-PROMOTION OF CANCER DUE TO HYPOXIA

* NOTE: LIMITED ANAEROBIC RESPIRATION STILL TAKES PLACE IN HIGHER AEROBIC LIFE FORMS, INCLUDING MAN.

MULTICELL
AEROBIC*
(INCLUDES MAN)
SINGLE CELL
ANAEROBIC
SINGLE CELL
AEROBIC*
CENTRAL CONTROL
OXYGEN-RICH
TIME-FORWARD PATH AS IT DEVELOPED
TIME-REVERSED PATH FOR SUSTAINED OXYGEN DEFICIENCY
SINGLE CELL
AEROBIC*
FIRST STEP IN OXYGEN-DEFICIENCY DEDIFFERENTIATION

CANCR-08

© 1993 T.E. Bearden

Next Slide

Return to Index

Previous Slide

CUMULATIVE GROWTH OF DEDIFFERENTIATION SIGNALS

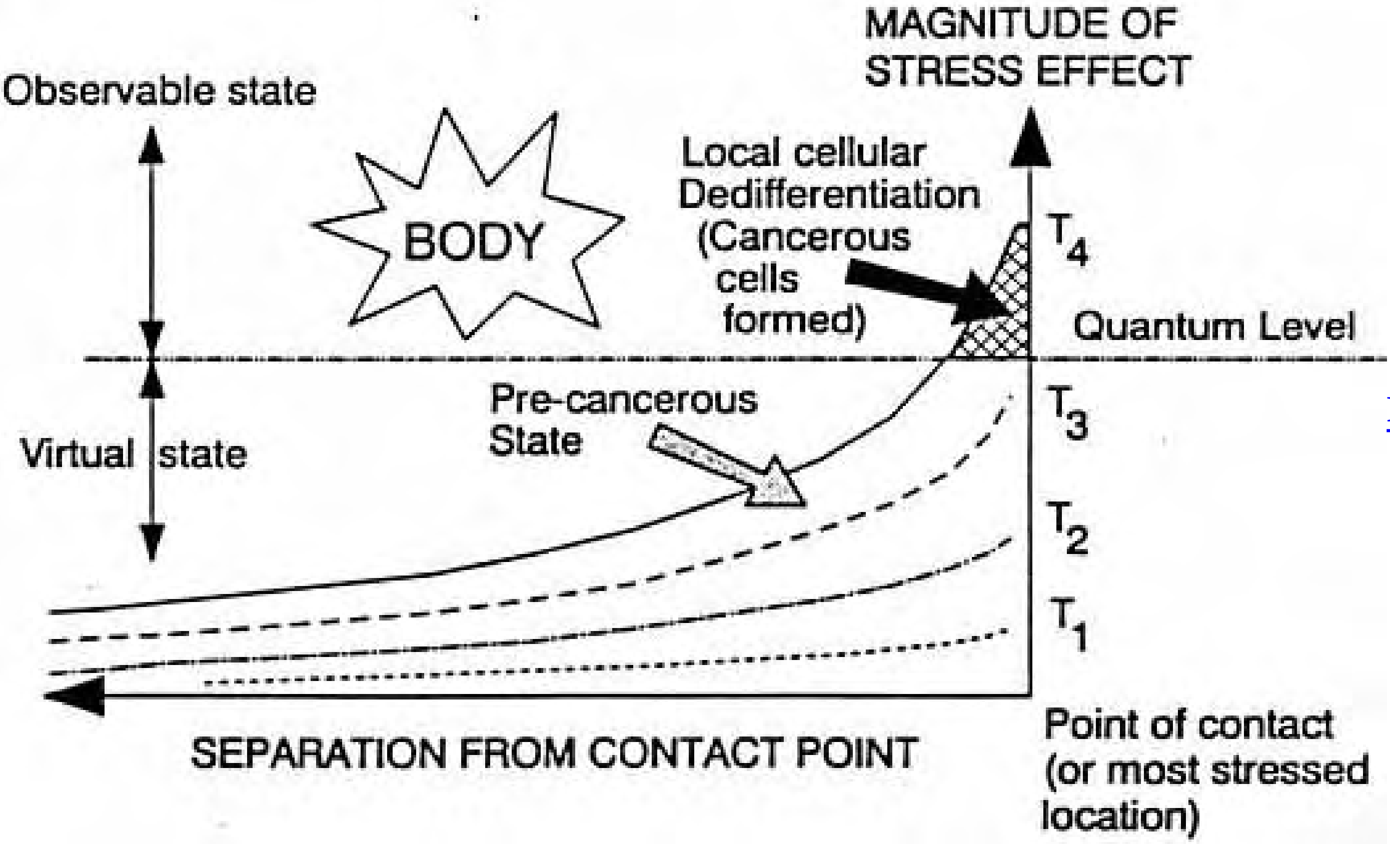
MAGNITUDE OF STRESS EFFECT
Observable state
BODY
Local cellular Dedifferentiation (Cancerous cells formed)
T4
Quantum Level
Virtual state
Pre-cancerous State
T3
T2
T1
SEPARATION FROM CONTACT POINT
Point of contact (or most stressed location)

Next Slide

Return to Index

Previous Slide

CUMULATIVE EMERGENCE OF CANCER INTO THE PHYSICAL STATE

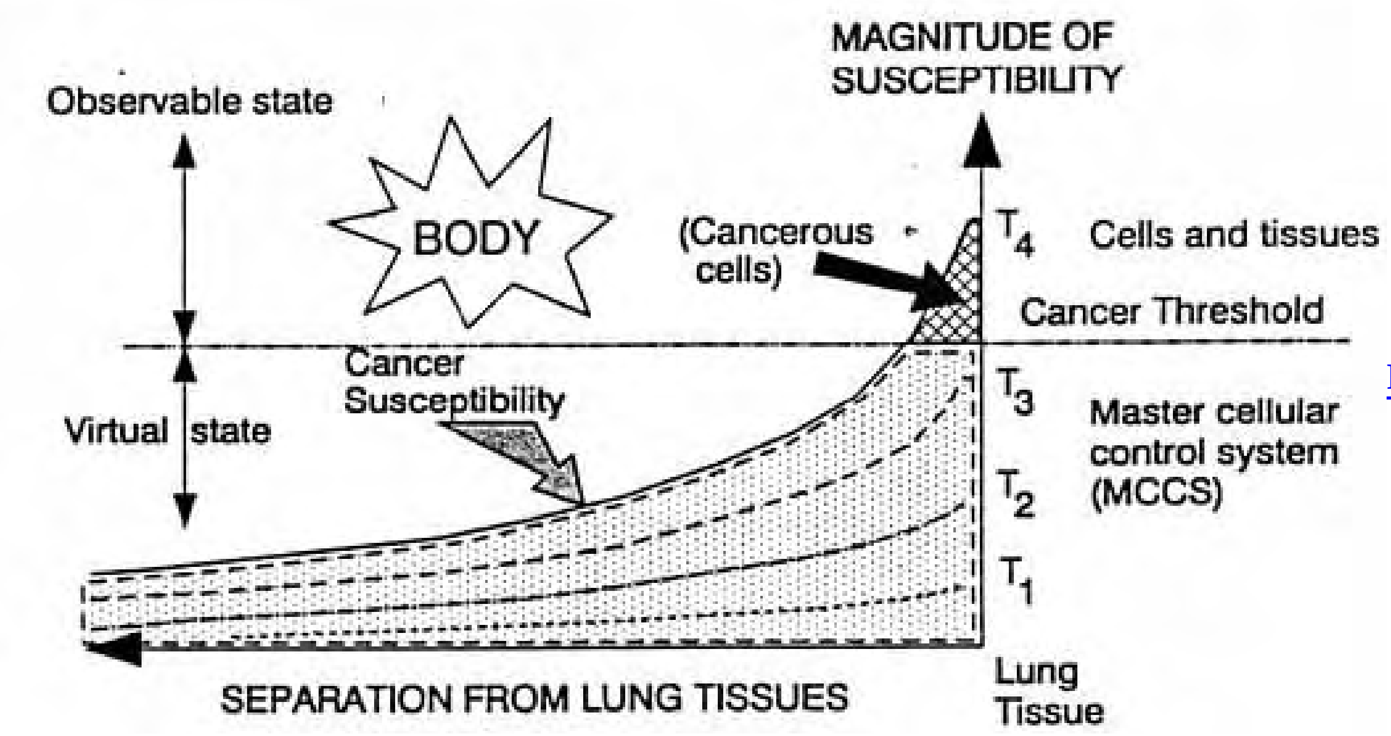
MAGNITUDE OF SUSCEPTIBILITY
Observable state
BODY
(Cancerous cells)
T_4
Cells and tissues
Cancer Threshold
Cancer Susceptibility
Virtual state
T_3
Master cellular control system (MCCS)
T_2
T_1
SEPARATION FROM LUNG TISSUES
Lung Tissue

Next Slide

Return to Index

Previous Slide

PRECANCER STATE CONDITION DETERMINES TREATMENT SUCCESS

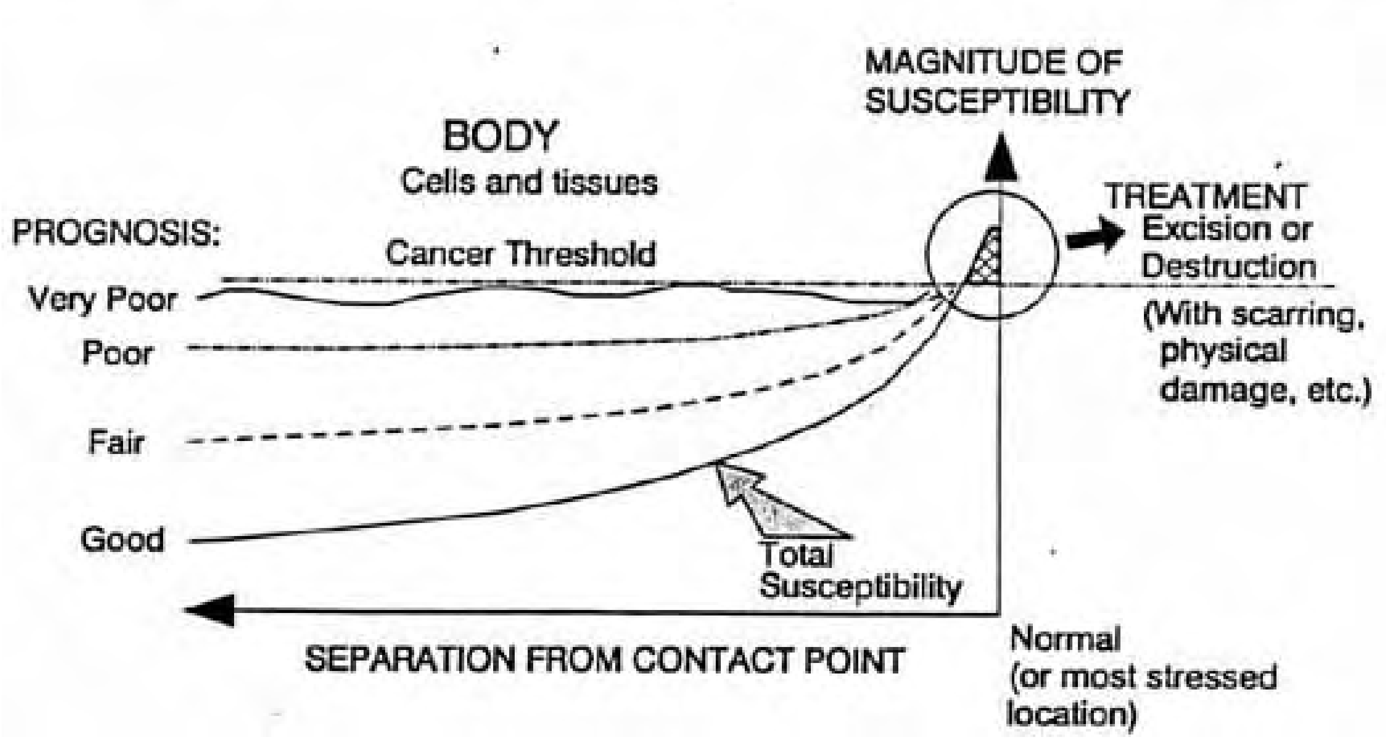

MAGNITUDE OF
SUSCEPTIBILITY

BODY
Cells and tissues

PROGNOSIS:

Cancer Threshold

Very Poor

Poor

Fair

Good

TREATMENT
Excision or
Destruction
(With scarring,
physical
damage, etc.)

Total
Susceptibility

SEPARATION FROM CONTACT POINT

Normal
(or most stressed
location)

Next Slide

Return to Index

Previous Slide

Starting From the Scalar EM View:
A NEW DEFINITION OF CANCER

- CANCER AND LEUKEMIA ARE:

> **Centrally-commanded, final, desperate, "first-step dedifferenetiation" adaptive attempts by the stressed affected cells experiencing sustained oxygen shortage, to reverse their cellular evolution and return to the anaerobic stage of their distant ancestry.**

- THE CAUSE IS LONG-TERM CUMULATIVE, AND ELECTROMAGNETIC IN NATURE.
- A TOTAL CURE CAN BE ACHIEVED ELECTRO-
MAGNETICALLY, AS WAS DEMONSTRATED.

Next Slide

Return to Index

Previous Slide

THE IMMUNE SYSTEM

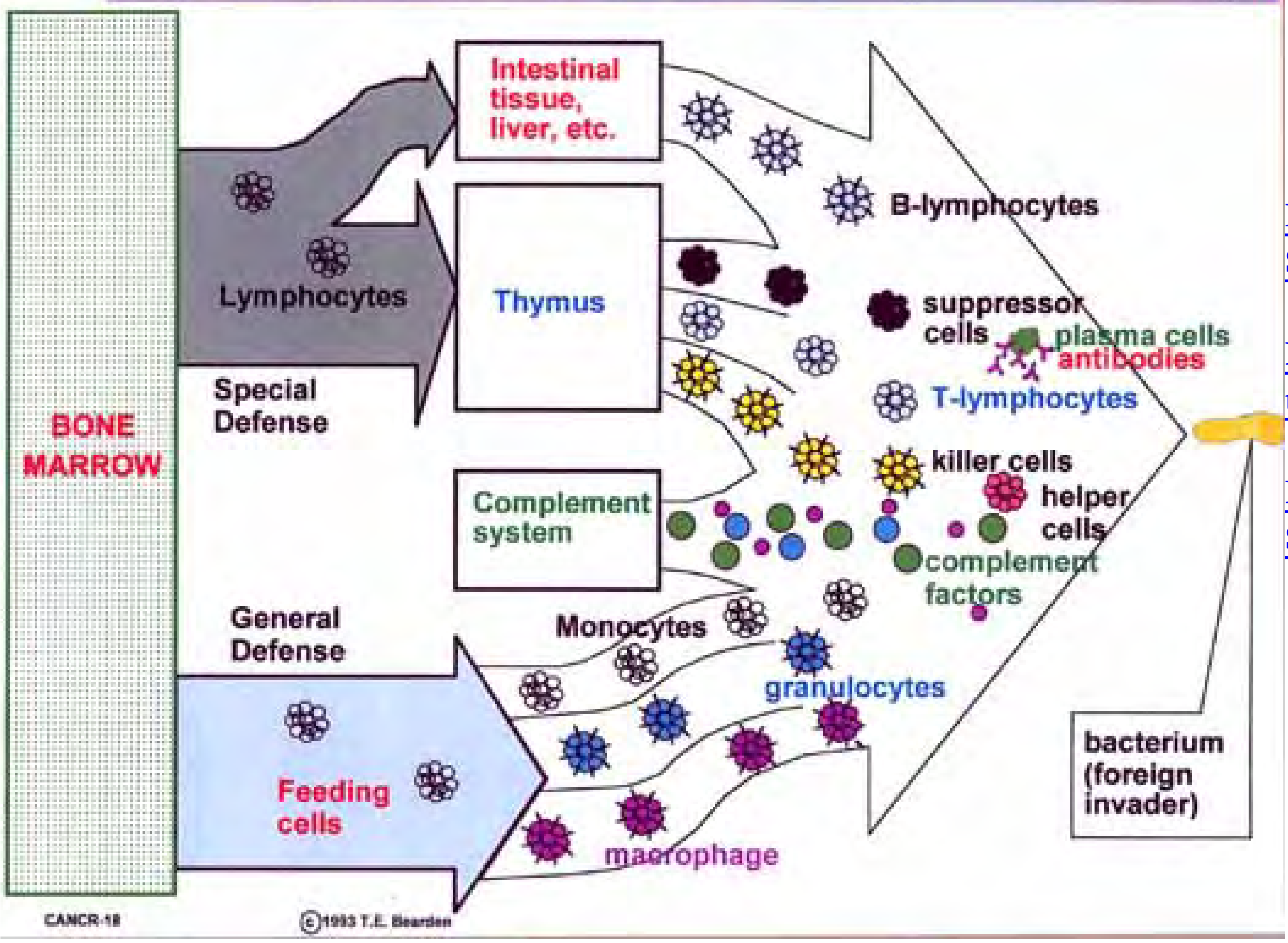
BONE MARROW
Lymphocytes
Special Defense
Intestinal tissue, liver, etc.
Thymus
Complement system
General Defense
Feeding cells
Monocytes
B-lymphocytes
suppressor cells
plasma cells
antibodies
T-lymphocytes
killer cells
helper cells
complement factors
granulocytes
macrophage
bacterium (foreign invader)
CANCR-18
©1993 T.E. Bearden

[Next Slide](054.htm)

[Return to Index](054.htm)

[Previous Slide](054.htm)

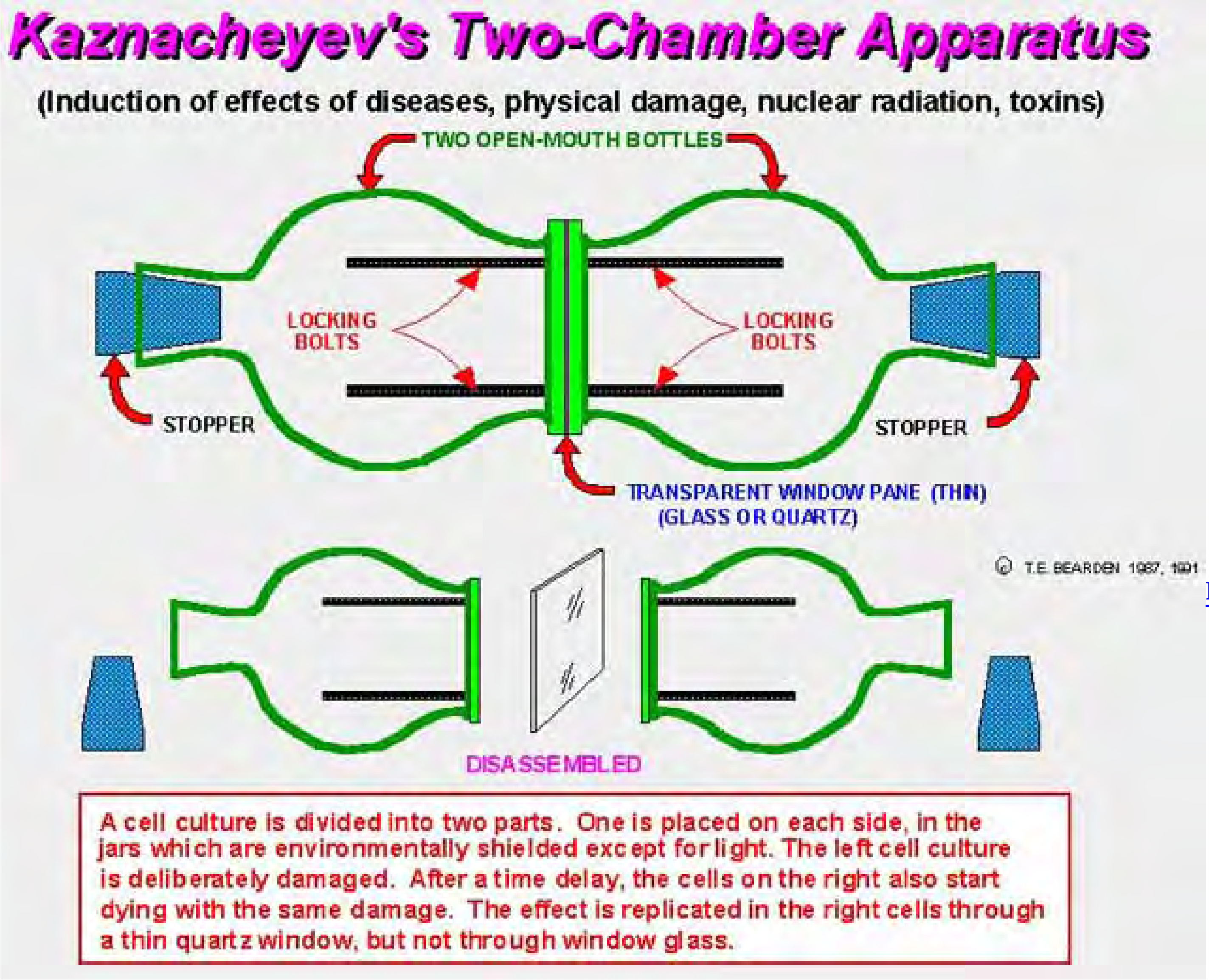

Kaznacheyev's Two-Chamber Apparatus

(Induction of effects of diseases, physical damage, nuclear radiation, toxins)

TWO OPEN-MOUTH BOTTLES
LOCKING BOLTS
LOCKING BOLTS
STOPPER
STOPPER
TRANSPARENT WINDOW PANE (THIN)
(GLASS OR QUARTZ)
© T.E. BEARDEN 1987, 1991
DISASSEMBLED

A cell culture is divided into two parts. One is placed on each side, in the jars which are environmentally shielded except for light. The left cell culture is deliberately damaged. After a time delay, the cells on the right also start dying with the same damage. The effect is replicated in the right cells through a thin quartz window, but not through window glass.

Next Slide

Return to Index

Previous Slide

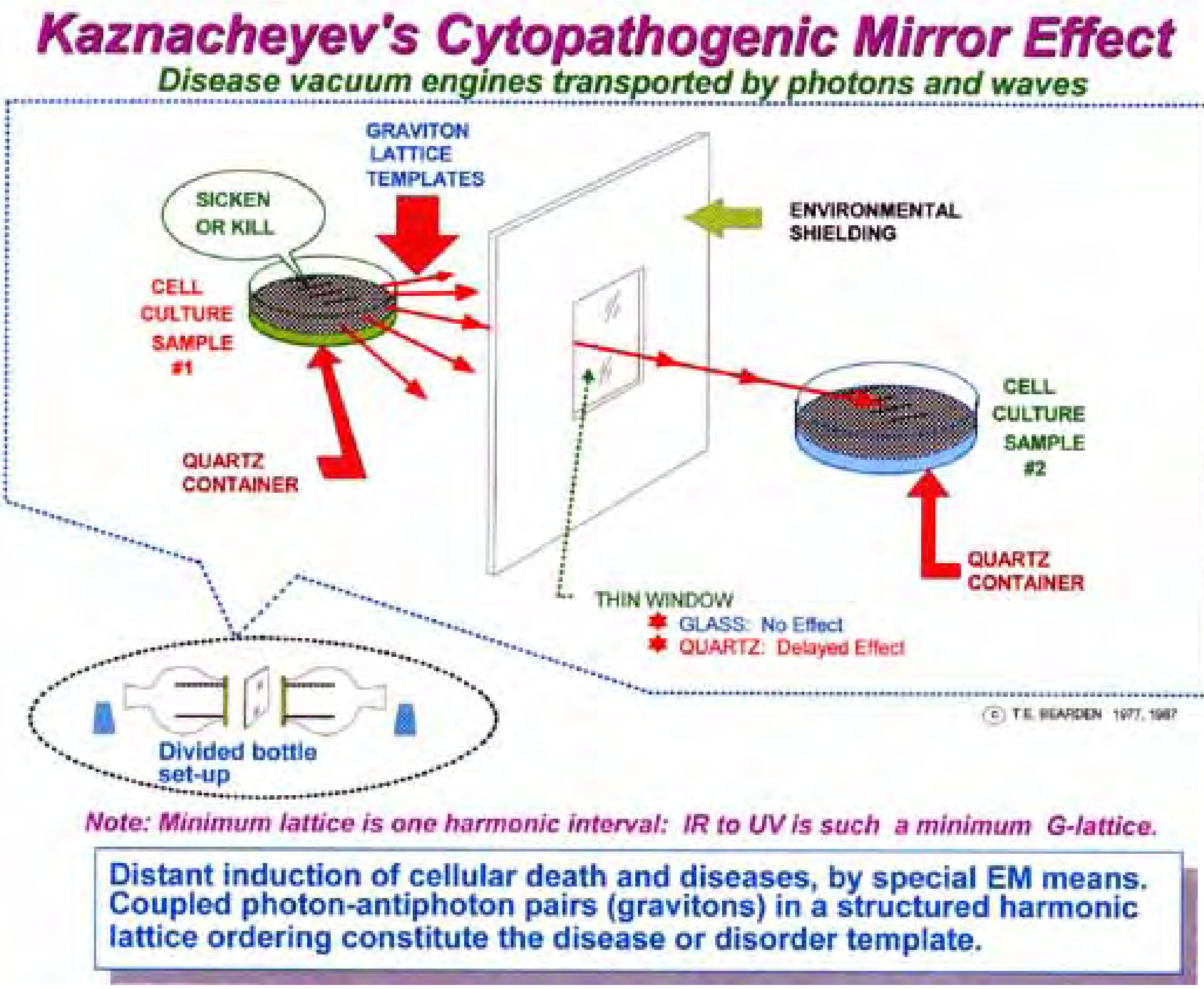

Kaznacheyev's Cytopathogenic Mirror Effect

Disease vacuum engines transported by photons and waves

GRAVITON
LATTICE
TEMPLATES
SICKEN
OR KILL
ENVIRONMENTAL
SHIELDING
CELL
CULTURE
SAMPLE
#1
QUARTZ
CONTAINER
CELL
CULTURE
SAMPLE
#2
QUARTZ
CONTAINER
THIN WINDOW
✸ GLASS: No Effect
✸ QUARTZ: Delayed Effect
Divided bottle
set-up

© T.E. BEARDEN 1977, 1987

Note: Minimum lattice is one harmonic interval: IR to UV is such a minimum G-lattice.

Distant induction of cellular death and diseases, by special EM means. Coupled photon-antiphoton pairs (gravitons) in a structured harmonic lattice ordering constitute the disease or disorder template.

Next Slide

Return to Index

Previous Slide

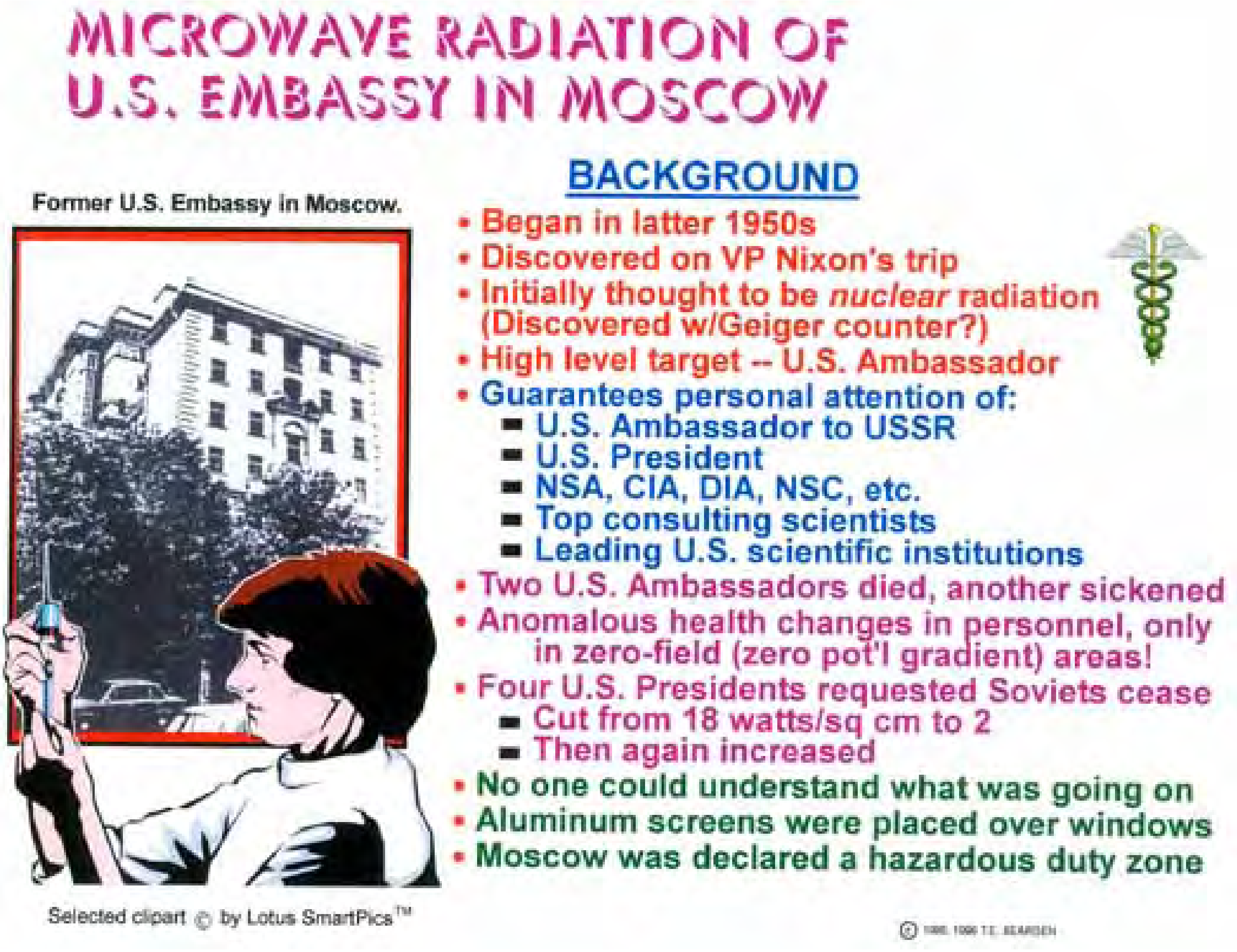

MICROWAVE RADIATION OF U.S. EMBASSY IN MOSCOW

Former U.S. Embassy in Moscow.

BACKGROUND

- Began in latter 1950s
- Discovered on VP Nixon's trip
- Initially thought to be *nuclear* radiation (Discovered w/Geiger counter?)
- High level target -- U.S. Ambassador
- Guarantees personal attention of:
  - U.S. Ambassador to USSR
  - U.S. President
  - NSA, CIA, DIA, NSC, etc.
  - Top consulting scientists
  - Leading U.S. scientific institutions
- Two U.S. Ambassadors died, another sickened
- Anomalous health changes in personnel, only in zero-field (zero pot'l gradient) areas!
- Four U.S. Presidents requested Soviets cease
  - Cut from 18 watts/sq cm to 2
  - Then again increased
- No one could understand what was going on
- Aluminum screens were placed over windows
- Moscow was declared a hazardous duty zone

Selected clipart © by Lotus SmartPics™

© [illegible] T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

JOHN HOPKINS STUDY OF EMBASSY MICROWAVE RADIATION

- **EXCELLENT SCIENTIFIC TEAM**

  - PERFORMED CLASSICAL EM INVESTIGATION

  - CAREFULLY ESTABLISHED THE FORCE FIELD PATTERNS

  - NO HEALTH CHANGES OCCURRED IN PERSONNEL WHERE THE FORCE FIELDS WERE PRESENT (NONZERO)

  - ALL HEALTH CHANGES OCCURRED IN PERSONNEL WHERE THE FORCE FIELDS WERE ZERO (ABSENT)

- **ASSUMED THAT ALL EM ACTION IS DUE ONLY TO FORCE FIELDS**

  - CONCLUDED THAT MICROWAVE RADIATION COULD NOT HAVE CAUSED THE HEALTH CHANGES
- **CONCLUSION WAS IN SERIOUS ERROR**
  - CORRECT CONCLUSION: HEALTH CHANGES WERE TOTALLY CORRELATED TO ABSENCE OF FORCE FIELDS

  *** TOTALLY CORRELATED TO PRESENCE OF GRADIENT-FREE POTENTIALS, OR NOT TO RADIATION AT ALL**

  *** IF NOT TO RADIATION AT ALL, THEN HEALTH CHANGES WOULD HAVE ALSO OCCURRED WHERE FORCE FIELDS WERE PRESENT**

- **THEREFORE THE HEALTH CHANGES WERE TOTALLY CORRELATED TO THE GRADIENT-FREE POTENTIALS, AND TO THE EM RADIATION**

Next Slide

Return to Index

Previous Slide

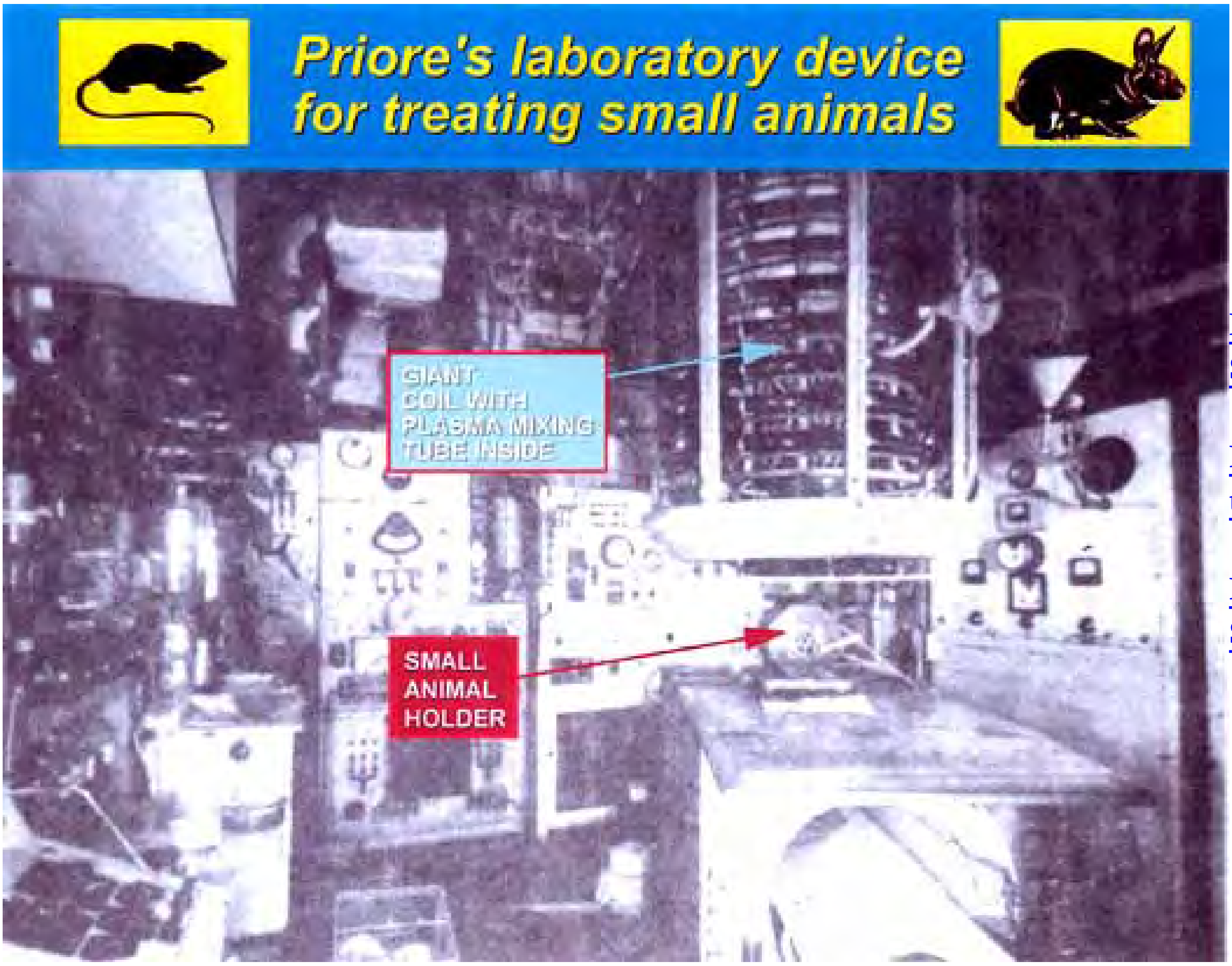

Priore's laboratory device for treating small animals

GIANT COIL WITH PLASMA MIXING TUBE INSIDE
SMALL ANIMAL HOLDER

Next Slide

Return to Index

Previous Slide

MECHANISM OF THE PRIORE THERAPY

- **PRODUCED A DETERMINISTIC WZ STRUCTURED SCALAR POTENTIAL**
  - **17 FREQUENCIES MIXED IN A ROTATING PLASMA**
  - **PHASE CONJUGATES ADDED BY THE PLASMA**
  - **MODULATED ONTO A RIPPLING MAGNETIC FIELD WHICH CARRIED IT TO ATOMIC NUCLEI**
  - **INTERACTED DIRECTLY WITH PERSONAL QUANTUM POTENTIAL**
- **NEGATED THE LONG-TERM CUMULATIVE CELLULAR DEDIFFERENTIATION ORDER GENERATED BY LONG-TERM HYPOXIA. TUMOR CELLS REVERTED TO NORMAL CONTROL**
- **SCRUBBED OUT THE CUMULATED PRECANCEROUS STATE**
- **RESTORED THE IMMUNE SYSTEM TO HIGH FUNCTIONING**
- **NO EXCESSIVE TRAUMA TO TREATED ANIMAL**

Next Slide

Return to Index

Previous Slide

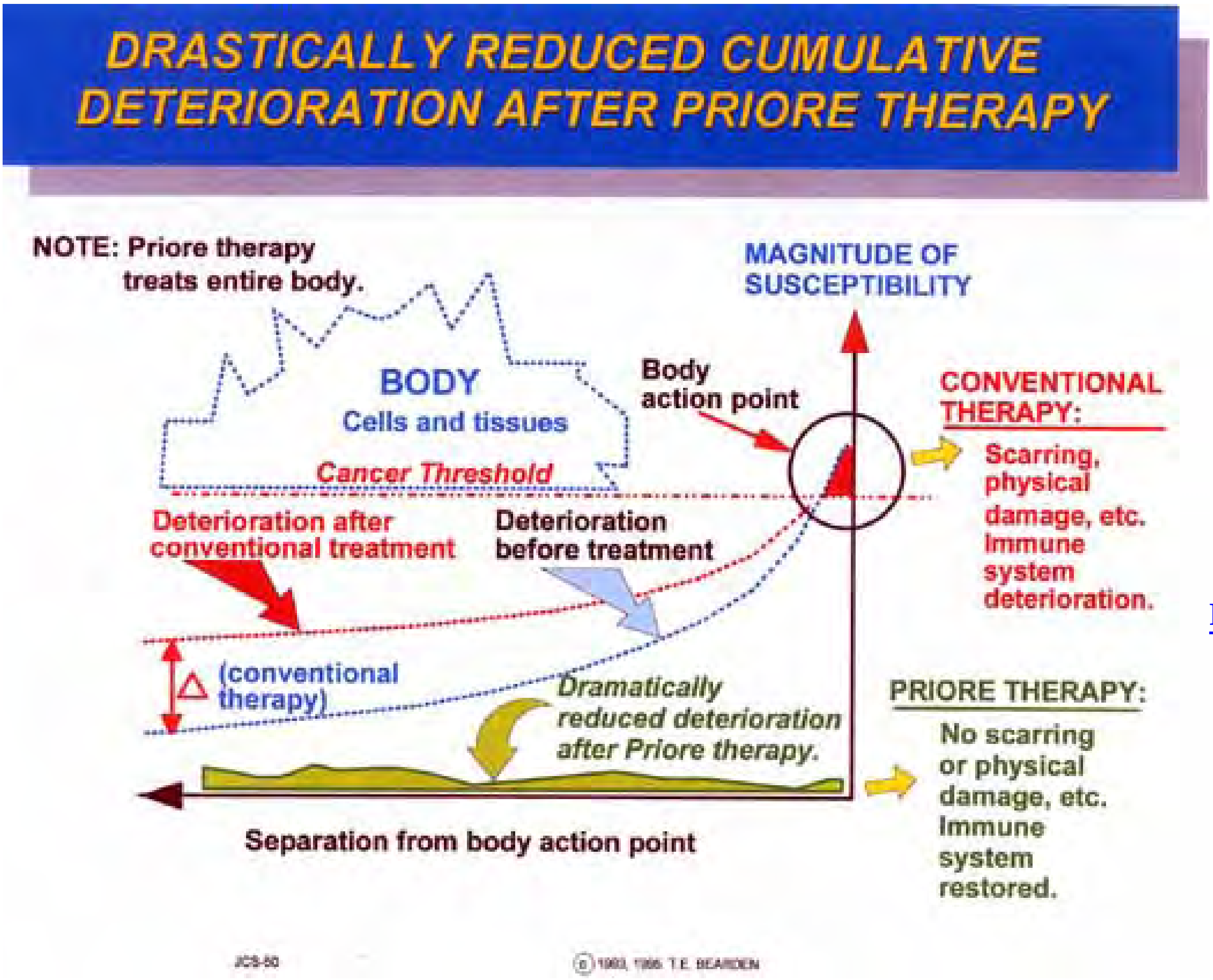

DRASTICALLY REDUCED CUMULATIVE DETERIORATION AFTER PRIORE THERAPY

NOTE: Priore therapy treats entire body.

MAGNITUDE OF SUSCEPTIBILITY
BODY
Cells and tissues
Body action point
Cancer Threshold
Deterioration after conventional treatment
Deterioration before treatment
Δ (conventional therapy)
Dramatically reduced deterioration after Priore therapy.
Separation from body action point

CONVENTIONAL THERAPY:

Scarring, physical damage, etc. Immune system deterioration.

PRIORE THERAPY:

No scarring or physical damage, etc. Immune system restored.

JCS-50

© 1993, 1996 T.E. BEARDEN

Next Slide

Return to Index

Previous Slide

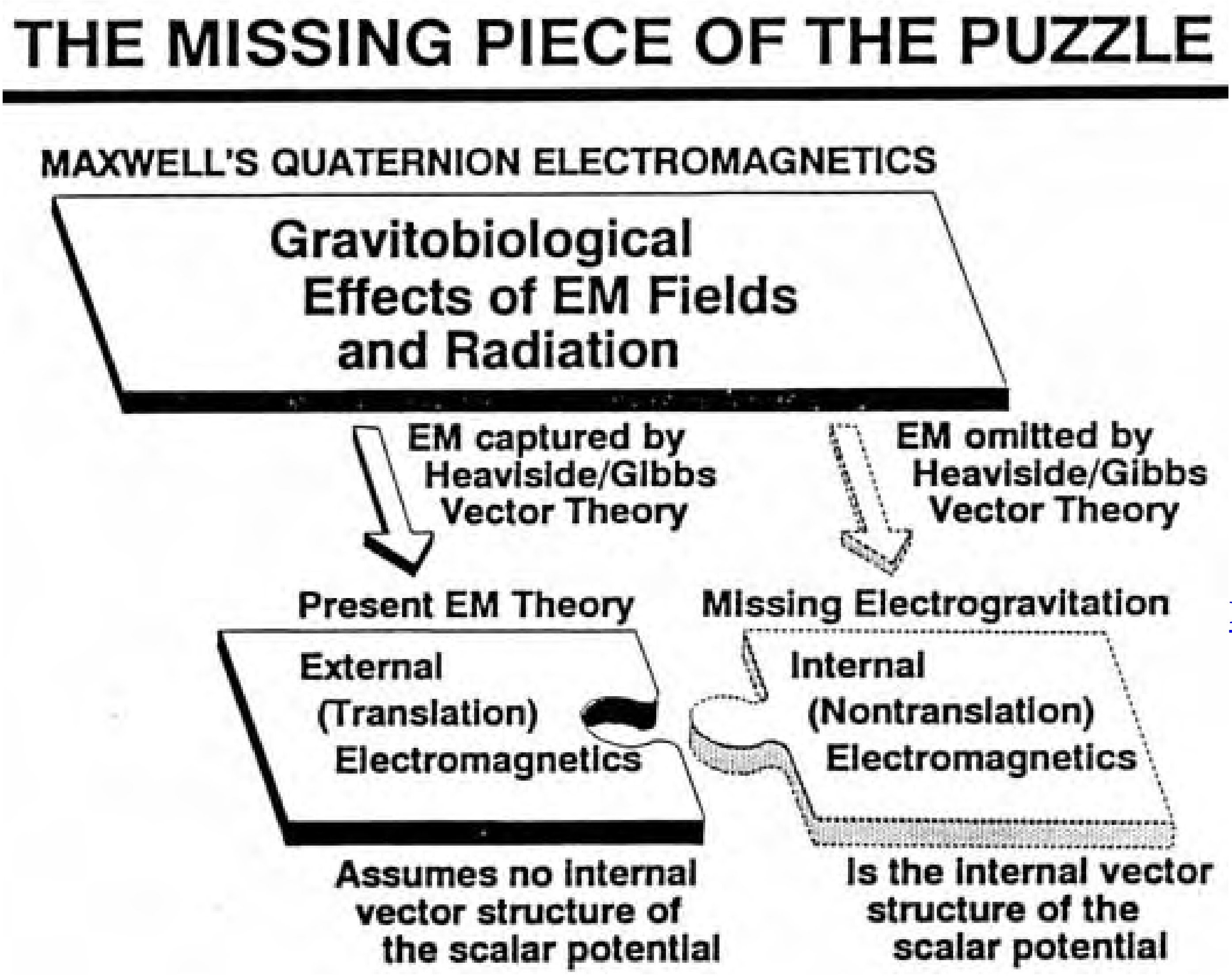

THE MISSING PIECE OF THE PUZZLE

MAXWELL'S QUATERNION ELECTROMAGNETICS

Gravitobiological
Effects of EM Fields
and Radiation

EM captured by
Heaviside/Gibbs
Vector Theory

EM omitted by
Heaviside/Gibbs
Vector Theory

Present EM Theory

External
(Translation)
Electromagnetics

Assumes no internal
vector structure of
the scalar potential

Missing Electrogravitation

Internal
(Nontranslation)
Electromagnetics

Is the internal vector
structure of the
scalar potential

Next Slide

Return to Index

Previous Slide

SUMMARY AND CONCLUSION

- **RECOMMEND SERIOUS EXAMINATION OF THE NEW APPROACH**
- **RECOMMEND REVIVAL OF PRIORE-TYPE RESEARCH ON**
  AN URGENT BASIS
- **CAN BE EXTENDED TO OTHER DISEASES**

  - AIDS
  - ARTERIOSCLEROSIS
  - SLEEPING SICKNESS
  - ETC.
- **A NEW THERAPEUTIC PARADIGM OF GREATLY INCREASED**
  SCOPE AND EFFECTIVENESS MAY BE POSSIBLE
- **IN ADDITION, DRASTIC REDUCTION IN MEDICAL COSTS**
  COULD POTENTIALLY RESULT

Return to Index

Previous Slide

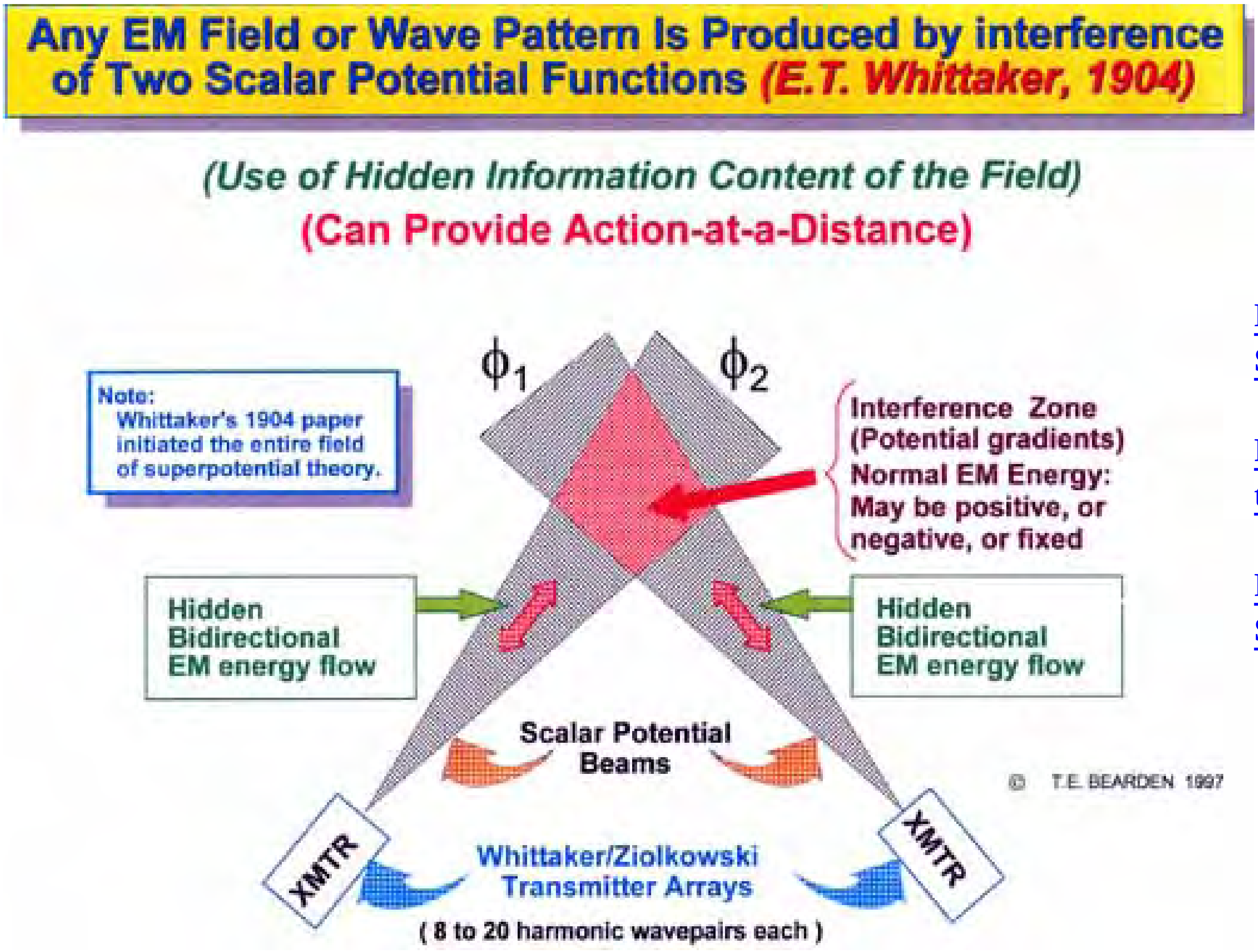

Any EM Field or Wave Pattern Is Produced by interference of Two Scalar Potential Functions *(E.T. Whittaker, 1904)*

(Use of Hidden Information Content of the Field)

(Can Provide Action-at-a-Distance)

Note:
Whittaker's 1904 paper initiated the entire field of superpotential theory.

ϕ_1

ϕ_2

Interference Zone
(Potential gradients)
Normal EM Energy:
May be positive, or
negative, or fixed

Hidden
Bidirectional
EM energy flow

Hidden
Bidirectional
EM energy flow

Scalar Potential
Beams

© T.E. BEARDEN 1997

XMTR

XMTR

Whittaker/Ziolkowski
Transmitter Arrays

(8 to 20 harmonic wavepairs each)

Next Slide

Return to Index

Previous Slide

MECHANISM FOR LONG-TERM, CUMULATIVE BIOLOGICAL EFFECTS OF EM FIELDS AND RADIATION

T. E. BEARDEN*

* A.D.A.S.
POB 1472
HUNTSVILLE, AL 35807

ENERGY FROM THE VACUUM
www.cheniere.org

Support Free Energy and purchase the bumper sticker

The New Tesla Electromagnetics and the Secrets of Electrical Free Energy

By T.E. Bearden
© 1984

Table of Contents

Foreword

Comments on the New Tesla Electromagnetics:
Part I: Discrepancies in the Present EM Theory
by T. E. Bearden

Comments on the New Tesla Electromagnetics:
Part II: The Secret of Electrical Free Energy
by T. E. Bearden

Overture to a New Age Technology
by Dr. Rolf Schaffranke

REVIEW AND OUTLOOK - The Development of Post-Relativistic Concepts in Physics and Advanced Technology Abroad
by Dr. Rolf Schaffrancke

The Tom Bearden Website

COMMENTS ON THE NEW TESLA ELECTROMAGNETICS

Part I: Discrepancies in Present EM Theory

© T. E. Bearden 1982

There are at least twenty-two major discrepancies presently existing in conventional electromagnetics theory. This paper presents a summary of those flaws, and is a further commentary on my discussion of scalar longitudinal Tesla waves in a previous paper, "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Company, 1981 and 1982.

I particularly wish to express my deep appreciation to two of my friends and colleagues who at this time, I believe, wish to remain anonymous. One of the two is an experimental genius who can produce items on the bench that do not work by orthodox theory. The second is a master of materials science and electromagnetics theory. I thank them both for their exceptional contributions and stimuli regarding potential shortcomings in present electromagnetics theory, and their forbearance with the many discussions we have held on this and related subjects.

It goes without saying that any errors in this paper are strictly my own, and not the fault of either of my distinguished colleagues.

(1) In present electromagnetics theory, charge and charged mass are falsely made identical. Actually, on a charged particle, the "charge" is the flux of virtual particles on the "bare particle" of observable mass. The charged particle is thus a "system" of true massless charge coupled to a bare chargeless mass. The observable "mass" is static, three-dimensional, and totally spatial. "Charge" is dynamic, four-dimensional or more, virtual and spatiotemporal. Further, the charge and observable mass can be decoupled, contrary to present theory. Decoupled charge -- that is, the absence of mass -- is simply what we presently refer to as "vacuum." Vacuum, spacetime, and massless charge are all identical. Rigorously, we should utilize any of these three as an "ether," as suggested for vacuum by Einstein himself (see

Max Born, Einstein's Theory of Relativity, Revised Edition, Dover Publications, New York, 1965, p. 224). And all three of them are identically anenergy -- not energy, but more fundamental components of energy.

(2) Electrostatic potential is regarded as a purely 3-dimensional spatial stress. Instead, it is the intensity of a many-dimensional (at least four-dimensional) virtual flux and a stress on all four dimensions of spacetime. This is easily seen, once one recognizes that spacetime is identically massless charge. (It is not "filled" with charge; rather, it is charge!) Just as, in a gas under pressure, the accumulation of additional gas further stresses the gas, the accumulation of charge (spacetime) stresses charge (spacetime). Further, if freed from its attachment to mass , charge can flow exclusively in time, exclusively in space, or in any combination of the two. Tesla waves -- which are scalar waves in pure massless charge flux itself -- thus can exhibit extraordinary characteristics that ordinary vector waves do not possess. And Tesla waves have extra dimensional degrees of freedom in which to move, as compared to vector waves. Indeed, one way to visualize a Tesla scalar wave is to regard it as a pure oscillation of time itself.

(3) Voltage and potential are often confused in the electrostatic case, or at least thought of as "composed of the same thing." For that reason, voltage is regarded as "potential drop". This also is not true. Rigorously, the potential is the intensity of the virtual particle flux at a single point -- whether or not there is any mass at the point -- and both the pressure and the point itself are spatiotemporal (4-dimensional), not spatial (3-dimensional) as presently assumed. Voltage represents the spatial intersection of the difference in the potential between two separated spatial points, and always implies at least a miniscule flow of mass current (that is what makes it spatial!). "Voltage" is spatial and depends upon the presence of observable mass flow, while scalar electrostatic potential is spatiotemporal and depends upon the absence of observable mass flow. The two are not even of the same dimensionality.

(4) The charge of vacuum spacetime is assumed to be zero, when in fact it is a very high value. Vacuum has no mass, but it has great massless charge and virtual particle charge flux. For proof that a charged vacuum is the seat of something in motion, see G.M. Graham and D.G. Lahoz, "Observation of static electromagnetic angular momentum in vacuo," Nature, Vol. 285, 15 .May 1980, pp. 154-155. In fact, vacuum IS charge, identically, and it is also "spacetime" and at least four-dimensional.

(5) Contrary to its present usage, zero is dimensional and relative in its context. A three-dimensional spatial hole, for example, exists in time. If we model time as a dimension, then the spatial hole has one dimension in 4-space. So a spatial absence is a spatiotemporal presence. In the vacuum 4-space, a spatial nothing is still a something. The "virtual" concept and the mathematical concept of a derivative are simply two present ways of unconsciously addressing this fundamental problem of the dimensional relativity of zero.

(6) The concepts of "space" and "time" imply that spacetime (vacuum) has been separated into two parts. We can only think of a space as "continuing to exist in time " To separate vacuum spacetime into two pieces, an operation is continually required. The operator that accomplishes this splitting operation is the photon interaction, the interaction of vector electromagnetic energy or waves with mass. I have already strongly pointed out this effect and presented a "raindrop model" of first-order physical change itself in my book, The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980, pp. 128-130.

(7) "Vector magnetic potential" is assumed to be always an aspect of (and connected to) the magnetic field. In fact it is a separate, fundamental field of nature and it can be entirely disconnected from the magnetic field. See Richard P. Feynman et al, The Feynman Lectures on Physics, Addison-Wesley Publishing Co., New York, 1964, Vol. II, pp. 15-8 to 15-14. Curiously, this fact has been proven for years' yet it has been almost completely ignored in the West. The "Vx" operator, when applied to the A-field, makes B-field. If the Vx operator is not applied, the "freed" A-field possesses much-expanded characteristics from those presently allowed in the "bound" theory. Specifically, it becomes a scalar or "shadow vector" field; it is not a normal vector field. (note: for V read inverted capital Delta)

(8) The speed of light in vacuum is assumed to be a fundamental constant of nature. Instead it is a function of the intensity of the massless charge flux (that is, of the magnitude of the electrostatic potential) of the vacuum in which it moves. (Indeed, since vacuum and massless charge are one and the same, one may say that the speed of light is a function of the intensity of the spatiotemporal vacuum!). The higher the flux intensity (charge) of the vacuum, the faster the speed of light in it. This is an observed fact and already shown by hardcore measurements. For example, distinct differences actually exist in the speed of light in vacuo, when measured on the surface of the earth as compared to measurements in space away from planetary masses. In a vacuum on the surface of the earth, light moves significantly faster. For a discussion and the statistics, see B. N. Belyaev, "On Random Fluctuations of the Velocity of Light in Vacuum," Soviet Physics

Journal, No. 11, Nov. 1980, pp. 37-42 (original in Russian; translation by Plenum Publishing Corporation.) The Russians have used this knowledge for over two decades in their strategic psychotronics (energetics) program; yet hardly a single U.S. scientist is aware of the measured variation of c in vacuo. In fact, most Western scientists simply cannot believe it when it is pointed out to them!

(9) Energy is considered fundamental and equivalent to work. In fact, energy arises from vector processes, and it can be disassembled into more fundamental (anenergy) scalar components, since the vectors can. These scalar components individually can be moved to a distant location without expending work, since one is not moving force vectors. There the scalar components can be joined and reassembled into vectors to provide "free energy" appearing at a distance, with no loss in between the initial and distant points. For proof that a vector field can be replaced by (and considered to be composed of) two scalar fields, see E. T. Whittaker, Proceedings of the London Mathematical Society, Volume 1, 1903, p. 367. By extension, any vector wave can be replaced by two coupled scalar waves.

(10) The classical Poynting vector predicts no longitudinal wave of energy from a time-varying, electrically charged source, In fact, an exact solution of the problem does allow this longitudinal wave. See T. D. Keech and J. F. Corum, "A New Derivation for the Field of a Time-Varying Charge in Einstein's Theory," International Journal of Theoretical Physics, Vol. 20, No, 1, 1981, pp. 63-68 for the proof.

(11) The present concepts of vector and scalar are severely limited, and do not permit the explicit consideration of the internal, finer-grained structures of a vector or a scalar. That is, a fundamental problem exists with the basic assumptions in the vector mathematics itself. The "space" of a vector field, for example, does not have inter-nested sublevels (subspaces) containing finer "shadow vectors" or "virtual vectors," Yet particle physics has already discovered that electrical reality is built that way. Thus one should actually use a "hypernumber" theory after the manner of Charles Musés. A scalar is filled with (and composed of) nested levels of other "spaces" containing vectors, where these sum to "zero" in the ordinary observable frame without an observable vector resultant. In Musés' mathematics, for example, zero has real roots. Real physical devices can be -- and have been -constructed in accordance with Muses' theory. For an introduction to Musés' profound hypernumbers approach, see Charles Musés' foreword to Jerome Rothstein, Communication Organization and Science, The Falcon's Wing Press, Indian Hills, Colorado, 1958. See also Charles Musés, "Applied Hypernumbers: Computational Concepts," Applied Mathematics and

Computation, Vol. 3, 1976. See also Charles Musés, "Hypernumbers II," Applied Mathematics and Computation, January 1978.

(12) With the expanded Tesla electromagnetics, a new conservation of energy law is required. Let us recapitulate for a moment. The oldest law called for the conservation of mass. The present law calls for the conservation of "mass and energy", but not each separately. If mass is regarded as simply another aspect of energy, then the present law calls for the conservation of energy. However, this assumes that energy is a basic, fundamental concept. Since the energy concept is tied to work and the movement of vector forces, it implicitly assumes "vector movement" to be a "most fundamental" and irreducible concept. But as we pointed out, Whittaker showed that vectors can always be further broken down into more fundamental coupled scalar components. Further, Tesla discovered that these "coupled components" of "energy" can be individually separated, transmitted, processed, rejoined, etc. This directly implies that energy per se need not be conserved. The new law therefore calls for the conservation of anenergy, the components of energy. These components may be coupled into energy, and the energy may be further compacted into mass. It is the sum total of the (anenergy) components -- coupled and uncoupled -- that is conserved, not the matter or the energy per se. Further, this conservation of anenergy is not spatial; rather, it is spatiotemporal in a spacetime of at least four or more dimensions.

(13) Relativity is presently regarded as a theory or statement about fundamental physical reality. In fact, it is only a statement about FIRST ORDER reality -- the reality that emerges from the vector interaction of electromagnetic energy with matter. When we break down the vectors into scalars (shadow vectors or hypervectors), we immediately enter a vastly different, far more fundamental reality. In this reality superluminal velocity, multiple universes, travel back and forth in time, higher dimensions, variation of all "fundamental constants" of nature, materialization and dematerialization, and violation of the "conservation of energy" are all involved. Even our present Aristotlean logic -- fitted to the photon interaction by vector light as the fundamental observation mechanism -- is incapable of describing or modeling this more fundamental reality. Using scalar waves and scalar interactions as much subtler, far less limited observation/detection mechanisms, we must have a new "superrelativity" to describe the expanded electromagnetic reality uncovered by Nikola Tesla.

(14) "Charge" is assumed to be quantized, in addition to always occurring with -- and locked to -- mass. Indeed, charge is not necessarily quantized just as it is not necessarily locked to mass.

Ehrenhaft discovered and reported fractional charges for years, in the 30's and 40's, and was ignored. See P.A.11. Dirac, "Development of the Physicist's Conception of Nature," Symposium on the Development of the Physicist's Conception of Nature, ed. Jagdish Merha, D. Reidel, Boston, 1973, pp. 12-14 for a presentation of some of Ehrenhaft's results. Within the last few years Stanford University· researchers also have positively demonstrated the existence of "fractional charge." For a layman's description of their work, see "A Spector Haunting Physics," Science News, Vol. 119, January 31, 1981, pp. 68-69. Indeed, Dirac in his referenced article points out that Millikan himself -- in his original oildrop experiments -- reported one measurement of fractional charge, but discounted it as probably due to error.

(15) Presently, things are always regarded as traveling through normal space. Thus we use or model only the most elementary type of motion -- that performed by vector electromagnetic energy. We do not allow for things to "travel inside the vector flow itself." Yet, actually, there is a second, more subtle flow inside the first, and a third, even more subtle flow inside the second, and so on. We may operate inside, onto, into, and out of energy itself -- and any anenergy component of energy. There are hypervectors and hyperscalars unlimited, within the ordinary vectors and scalars we already know. Further, these "internal flows" can be engineered and utilized, allowing physical reality itself to be directly engineered, almost without limits.

(16) We always assume everything exists in time. Actually, nothing presently measured exists in time, because the physical detection/ measurement process of our present instruments destroys time, ripping it off and tossing it away -- and thereby "collapsing the wave function." Present scientific methodology thus is seriously flawed. It does not yield fundamental (spacetime) truth, but only a partial (spatial) truth. This in turn leads to great scientific oversights. For example, mass does not exist in time, but mass x time (masstime) does. A fundamental constant does not exist in time, but the "constant x time" does. Energy does not exist in time, but energy x time (action) does. Even space itself does not exist in time -- spacetime does. We are almost always one dimension short in every observable we model. Yet we persist in thinking spatially, and we have developed 1nstruments that detect and measure spatially only. Such instruments can never measure and detect the phenomenology of the nested substrata of time. By using scalar technology, however, less limited instruments can indeed be constructed -and they have been. With such new instruments, the phenomenology of the new electromagnetics can be explored and an engineering technology developed.

(17) We do not recognize the connection between nested levels of virtual state (particle physics) and orthogonally rotated frames (hyperspaces). Actually the two are identical, as I showed in the appendix to my book, The Excalibur Briefing, Strawberry Hill Press, San Francisco, 1980, pp. 233-235. A virtual particle in the laboratory frame is an observable particle in a hyperspatial frame rotated more than one orthogonal turn away. This of course implies that the hyperspatial velocity of all virtual particles is greater than the speed of light. The particle physicist is already deeply involved in hyperspaces and hyperspatial charge fluxes without realizing it. In other words, he is using tachyons (particles that move faster than light) without realizing it.

(18) Presently quantum mechanics rigorously states that time is not an observable, and therefore it cannot be measured or detected. According to this assumption, one must always infer time from spatial measurements, because all detections and measurements are spatial. With this assumption, our scientists prejudice themselves against even looking for finer, subquantal measurement methodologies and instrumentation. Actually this present limitation is a result of the type of electromagnetics we presently know, where all instruments (the "measurers") have been interacted with by vector electromagnetic energy (light). Every mass that has temperature (and all masses do!) is continually absorbing and emitting photons, and in the process they are continually connecting to time and disconnecting from time. If time is continually being carried away from the detector itself by its emitted photons, then the detector cannot hold and "detect" that which it has just lost. With Tesla electromagnetics, however, the fundamental limitation of our present instruments need not apply. With finer instruments, we can show there are an infinite number of levels to "time", and it is only the "quantum level time" which is continually being lost by vector light (photon) interaction. By using subquantal scalar waves, instruments can move to deeper levels of time -- in which case the upper levels of time ARE measureable and detectable, in contradistinction to the present assumptions.

(19) In the present physics, time is modeled as, and considered to be, a continuous dimension such as length. This is only a gross approximation. Indeed, time is not like continuous "dimension," but more like a series of "stitches," each of which is individually made and then ripped out before the next stitch appears. "Vector light" photons interact one at a time, and it is this interaction with mass that creates quantum change itself. The absorption of a photon -- which is energy x time -- by a spatial mass converts it to masstime: the time was added by the photon. The emission of a photon tears away the time, leaving behind again a spatial mass. It is not accidental, then, that time flows at

the speed of light, for it is light which contains and carries time. It is also not accidental that the photon IS the individual quantum. Since all our instruments presently are continually absorbing and emitting photons, they are all "quantized," and they accordingly "quantize" their detections. This is true because all detection is totally internal to the detector, and the instruments detect only their own internal changes. Since these detections are on a totally granular quantized background, the detections themselves are quantized. The Minkowski model is fundamentally erroneous in its modeling of time, and for that reason relativity and quantum mechanics continue to resist all attempts to successfully combine them, quantum field theory notwithstanding.

(20) Presently, gravitational field and electrical field are considered mutually exclusive. Actually this is also untrue. In 1974, for example, Santilly proved that electrical field and gravitational field indeed are not mutually exclusive In that case, one is left with two possibilities: (a) they are totally the same thing, or (b) they are partially the same thing. For the proof, see R. M. Santilli, "Partons and Gravitation: Some Puzzling Questions," Annals of Physics, Vol. 83, No. 1, March 1974. With the new Tesla electromagnetics, pure scalar waves in time itself can be produced electrically, and electrostatics (when the charge has been separated from the mass) becomes a "magic" tool capable of directly affecting and altering anything that exists in time--including gravitational field. Antigravity and the inertial drive are immediate and direct consequences of the new electromagnetics.

(21) Presently, mind is considered metaphysical, not a part of physics, an not affected by physical means. Literally, the prevailing belief of Western scientists is that man is a mechanical robot -- even though relativity depends entirely upon the idea of the "observer." Western science today thus has essentially become dogmatic, and in this respect borders on a religion. Since this "religion," so to speak, is now fairly well entrenched in its power in the state, Western science is turning itself into an oligarchy. But mind occupies time, and when we measure and affect time we can directly measure and affect mind itself. In the new electromagnetics, then, man regains his dignity and his humanity by restoring the reality of mind and thought to science. In my book, The Excalibur Briefing, I have already pointed out the reality of mind and a simplified way in which it can be modeled to the first order. With scalar wave instruments, the reality of mind and thought can be measured in the laboratory, and parapsychology becomes a working, engineering, scientific discipline.

(22) Multiple-valued basic dimensional functions are either not permitted or severely discouraged in the present theory For one thing, integrals of multiple valued derivative functions have the annoying

habit of "blowing up" and yielding erroneous answers, or none at all. And we certainly do not allow multiple types of time! This leads to the absurdity of the present interpretation of relativity which permits only a single observer (and a single observation) at a time. So if one believes as "absurd" a thing as the fact that more than one person can observe an apple at the same time, the present physics fails. However, the acceptance of such a simple proposition as multiple simultaneous observation leads to a physics so bizarre and incredible that most Western physicists have been unable to tolerate it, much less examine its consequences. In the physics that emerges from multiple simultaneous observation, all possibilities are real and physical. There are an infinite number of worlds, orthogonal to one another, and each world is continually splitting into additional such "worlds" at a stupendous rate. Nonetheless, this physics was worked out by Everett for his doctoral thesis in 1956, and the thesis was published in 1957. (See Hugh Everett, III, The Many-Worlds Interpretation of Quantum Mechanics: A Fundamental Exposition, with papers by J. A. Wheeler, B.S. DeWitt, L. N. Cooper and D. Van Vechten, and N. Graham; eds. Bryce S. Dewitt and Neill Graham, Princeton Series in Physics, Princeton University Press, 1973.) Even though it is bizarre, Everett's physics is entirely consistent with all the present experimental basis of physics. The present electromagnetic theory is constructed for only a single "world" or universe -- or "level." The expanded theory, on the other hand, contains multiply nested levels of virtual state charge -- and these levels are identically the same as orthogonal universes, or "hyperframes." Multiple kinds -- and values -- of time also exist. The new concept differs from Everett's, however, in that the orthogonal universes intercommunicate in the virtual state. That is, an observable in one universe is always a virtual quantity in each of the other universes. Thus one can have multi-level "continuities" and "discontinuities" simultaneously, without logical conflict. It is precisely these levels of charge -- these levels of scalar vacuum -- that lace together the discontinuous quanta generated by the interaction of vector light with mass.

However, to understand the new electromagnetic reality, one requires a new, expanded logic which contains the old Aristotlean logic as a subset. I have already pointed out the new logic in my paper, "A Conditional Criterion for Identity, Leading to a Fourth Law of Logic," 1979, available from the National Technical Information Center, AD-A071032.

Even as logic is extended, quantum mechanics, quantum electrodynamics, and relativity are drastically changed by the Tesla electromagnetics, as I pointed out in my paper, "Solutions to Tesla's Secrets and the Soviet Tesla Weapons," Tesla Book Company, 1580

Magnolia, Millbrae, California, 94030, 1980.

The present electromagnetics is just a special case of a much more fundamental electromagnetics discovered by Nikola Tesla, just as Newtonian physics is a special case of relativistic physics. But in the electromagnetics case, the differences between the old and the new are far more drastic and profound.

ADDITIONAL REFERENCES

1, Boren, Dr. Lawrence Milton, "Discovery of the Fundamental Magnetic Charge (Arising from the new Conservation of Magnetic Energy)," 1981/82 (private communication). Dr. Boren has a cogent argument that the positron is the fundamental unit of magnetic charge. His theory thus assigns fundamentally different natures to positive charge and negative charge. In support of Dr. Boren, one should point out that the "positive" end of circuits can simply be "less negative" than the "negative" end. In other words, the circuit works simply from higher accumulation of negative charges (the "negative" end) to a lesser accumulation of negative charges (the "positive" end). Nowhere need there be positive charges (protons, positrons, etc.) to make the circuit work. Dr. Boren's theory, though dramatic at first encounter, nonetheless bears close and meticulous examination -- particularly since he has been able to gather experimental data which support his theory and disagree with present theory.

2. Eagle, Albert, "An Alternative Explanation of Relativity Phenomena," Philosophical Magazine and Journal of Science, No. 191, December 1939, pp. 694-701.

3. Ehrenhaft, Felix and Wasser, Emanuel, "Determination of the Size and Weight of Single Submicroscopic Spheres of the Order of Magnitude r = 4 x 10(-5) cm. to 5 x 10(-6) cm., as well as the Production of Real Images of Submicroscopic Particles by means of Ultraviolet Light," Phil. Mag. and Jour. of Sci., Vol. II (Seventh Series), No. 7, July 1926, pp. 30-51.

4. Ehrenhaft, Felix and Wasser, Emanuel, "New Evidence of the Existence of Charges smaller than the Electron - (a) The Micromagnet; (b) The Law of Resistance; (c) The Computation of Errors of the Method," Phil. Mag. and Jour. of Sci., Vol. V (Seventh Series), No. 28, February 1928, pp. 225-241.

5. See also Ehrenhaft's last paper dealing with the electronic charge, in Philosophy of Science, Vol. 8, 1941, p. 403.

6. McGregor, Donald Rait, The Inertia of the Vacuum: A New Foundation for Theoretical Physics, Exposition Press, Smithtown, NY, First Edition, 1981, pp. 15-20.

7. Ignat'ev, Yu. G. and Balakin, A. B., "Nonlinear Gravitational Waves in Plasma," Soviet Physics Journal, Vol. 24, No. 7, July 1981, (U.S. Translation, Consultants Bureau, NY, January 1982), pp. 593-597.

8. Yater, Joseph C., "Relation of the second law of thermodynamics to the power conversion of energy fluctuations," Phys. Review A, Vol. 20, no. 4, October 1979, pp. 1614-1618.

9. DeSantis, Romano M. et al, "On the Analysis of Feedback Systems With a Multipower Open Loop Chain," October 1973, available through the Defense Technical Information Center (AD 773188),

10. Graneau, Peter, "Electromagnetic jet-propulsion in the direction of current flow," Nature, Vol. 295, 28 January 1982, pp, 311-312.

11. "Gravity and acceleration aren't always equivalent," New Scientist, 17 September 1981, p. 723.

12. Gonyaev, V. V , "Experimental Determination of the Free-Fall Acceleration of a Relativistic Charged Particle. II. A Cylindrical Solenoid in a Time-Independent Field of Inertial Forces," Izvestiya VUZ, Fizika, No. 7, 1979, pp. 28-32. English Translation: Soviet Physics Journal, No. 7, 1979, pp. 829-833. If one understands the new, expanded electromagnetics, this Soviet paper indicates a means of generating antigravity and pure inertial fields.

13. R. Schaffranke, "The Development of Post-Relativistic Concepts in Physics and Advanced Technology Abroad," Energy Unlimited, No. 12, Winter 1981, pp. 15-20.

14. F. K. Preikschat, A Critical Look at the Theory of Relativity, Library of Congress Catalogue No. 77-670044. Extensive compilation of measurements of the speed of light. Clearly shows the speed of light is not constant but changes, sometimes even daily.

Return to Table of Contents **Next Page**

COMMENTS ON THE NEW TESLA ELECTROMAGNETICS

Part II: The Secret of Electrical Free Energy

© 1982 T. E. Bearden

Again I wish to express my deep appreciation to two of my friends and colleagues who at this time, I believe, wish to remain anonymous. I thank them both for their exceptional contributions and stimuli, and their forbearance with the many discussions we have held on this and related subjects.

Also, it goes without saying that any errors in this paper are strictly my own, and not the fault of either of my distinguished colleagues.

Present electromagnetic theory is only a special case of the much more fundamental electromagnetism discovered by Nikola Tesla at the turn of the century.

Pure vacuum is pure charge flux, without mass. The vacuum has a very high electrical potential -- something on the order of 200 million volts, with respect to a hypothetical zero charge.

Thus in an ordinary electrical circuit, each point of the "ground" -- which has the same potential as the vacuum -- actually has a non-zero absolute potential. This circuit ground has a value of zero only with respect to something else which has the same absolute electrical potential.

Voltage, which is always associated with a flow of electrical "mass" current (even if only a miniscule flow), is, by definition, a difference dropped in potential when a charge mass moves between two spatially separated points. What we have termed "electrical current" only flows where there is a suitable conducting medium between things which have a difference in absolute potential. Furthermore, between any two points in any material, there is considered to exist a finite resistance -- if we apply a voltage and have a mass current flowing between the two points! The simple statement that V = IR irrevocably ties together voltage drop, mass current, and resistance between two points. Rigorously, to have one of the three is to have them all. To lose one is to lose all three. Immediately we see a major error in present theory: One can have a "difference in scalar potential" between two points without having a "voltage drop" between them. Specifically, if no mass current flows between them, no resistance exists between them, and no voltage drop exists between them.

In the same fashion, one can have a "scalar wave" through the

In the same fashion, one can have a "scalar wave" through the vacuum without a voltage wave. In that case, the wave has no E-field and no H-field. The only reason one has an E-field

Return to Table of Contents Next Page

around a statically charged object is because the charged electrons accumulated on the object are actually in violent motion. It is this motion of the charged masses that produces E-field -- as well as H-field whenever that entire E-field ensemble moves through laboratory space.

Now let us reason together in the "approximate" manner utilized in present electromagnetic theory. For example, let us examine a bird sitting on a high tension line.

The bird sits on the high tension line without a flow of mass electricity, because there is no significant difference in potential drop between the bird and the line. Specifically, between the bird's two feet -- each in contact with a different portion of the line -- there exists no potential difference. This is true even though, with respect to the vacuum, each foot is at a potential that would be "100,000 volts higher," were a mass current flowing. And it is true even though the absolute potential of each foot may be some 200.1 million "volts," were a mass current flowing.

Now an interesting thing happens to the bird when he flies through the air to light upon the high tension wire. As he flies toward the wire, he is flying through the massless electrostatic potential field of the wire, for that field extends an infinite distance away from the wire. The electrostatic potential field -- pure ∅-field -- is actually the spatio-temporal intensity of the massless charge at a point. In other words, as the bird flies to the wire, he flies into an increasing "massless charge" potential, building up to 100,000 "volts" higher than the earth. However, very little (if any) "mass flow" potential difference is experienced upon his body in approaching the wire, and so essentially no "charged mass currents" are induced in his body. Thus the little flier safely navigates into the teeth of a very high electrostatic potential, lights upon the wire, and is not "fried" in the process. When he lights on the wire, his body has reached the electrostatic potential that each foot's contact point has. Again, there is no mass current flow. But his body is immersed in an increased flux of massless charge -- which is what the electrostatic potential represents. And each "virtual particle" flow in that charge potential represents a "massless (scalar)" electrical current.

The point is, one can have any amount of massless charge flow -- "scalar" current -- without any mechanical work being done in the system. All electrical work in a circuit is done against the physical mass of the charged masses that flow. Rigorously, force is defined as the time rate of charge of momentum. Even in the relativistic case where $F = ma + v(dm/dt)$, change of momentum requires mass movement. No mechanical work, and hence no energy, is expended by massless charge flow.

That is why the vacuum massless charge -- which is composed of a very high flux of massless "particles" -- normally does no work

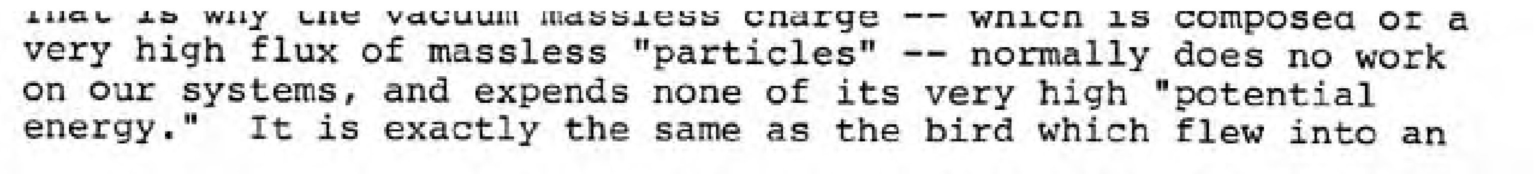

That is why the vacuum massless charge -- which is composed of a very high flux of massless "particles" -- normally does no work on our systems, and expends none of its very high "potential energy." It is exactly the same as the bird which flew into an

Return to Table of Contents Next Page

increasing scalar field as it approached the high tension wire -- no work was done upon the bird by the increasing scalar flux currents encountered by its body.

By existing "in the vacuum," so to speak, we (the whole earth) are as birds sitting on a high tension line! Until we create significant difference in potential, via our present electromagnetic circuits, no current can flow -- anywhere. Even if we produce potential differences, we must have a conductor and charged masses to flow, if we wish to produce mechanical work. Presently our electromagnetic theory allows us to create a difference in potential within different parts of a circuit, but only by moving and shifting charged mass. We therefore have to do work on this electrical mass in moving it around, and we only get back the work we have put into the circuit. In other words, presently all we do is "pump" electrical mass.

Now notice what would happen to the bird on the line if we substantially "pulsed" the potential on the line. Suppose we "pulsed" it such that the bird's physical system -- considered as a circuit containing a capacitance, a resistance, an inductance, and many free electrons -- became resonant to the pulsing frequency. In that case the "bird system" would resonate, and a great deal of electrical mass would surge back and forth in the body of the bird. In the bird's body, voltage would exist, charged mass current would flow, work would be done, and the bird would be electrocuted.

Also, note that, without mass movement, electromagnetic vector fields are not produced (and a portion of the difficulty lies with the actual vector mathematics itself). Scalar (nonvector) waves continually penetrate "space" where there is no mass movement. This means there can exist a "delta-Φ" without a voltage or an E-field. The present theory does not allow this, because it always uses "q" (charge) to be charged mass. Briefly, without belaboring the point, let us just say that it is the mechanical spin of the individual charged particle -- such as the electron -- which "entangles" or "knits together" or "couples" independent scalar waves into vector waves. A vector wave is simply two coupled scalar waves. The entire force field concept -- such as the E-field and the B-field -- is operationally defined in terms of the force exhibited on a test particle, or test mass. Rigorously, an E-field does not exist as a force field in vacuum, but as two coupled scalar Φ-fields "tumbling about each other." When these two coupled, tumbling fields meet a spinning electron, e.g., the force emerges on the electron mass. In short, movement of a rotating mass changes delta-Φ to "voltage", creating the V/I/R triad.

By "accumulating charged mass particles" -- such as electrons -- one certainly can increase the value of Φ, which represents the charge intensity or "scalar electrostatic potential." However, that is not the only way to increase it. Resonance and rotation of charged masses can also be appropriately employed to vary the vacuum charge potential Φ, under the proper circumstances.

of charged masses can also be appropriately employed to vary the vacuum charge potential Φ, under the proper circumstances.

Return to Table of Contents Next Page

By the correct application of rotary principles and Tesla electromagnetic theory, it is possible to oscillate -- and change -- the vacuum potential itself, in one part of an electrical system. Thus by correct procedures a part of a system can be electrically altered so that the absolute value of its "ground" (vacuum) potential differs significantly from the normal vacuum-ground potential of the remainder of the circuit. In other words, we shift the vacuum-ground potential of part of the circuit by oscillating the massless vacuum charge itself, and in so doing we gain a substantial, oscillatory electrical potential difference between that part of the circuit and the normal ground potential of the rest of the system. Then we connect the two parts of the system by means of a conductor containing a load in the middle.

In every conductor, a large number of "free electrons" are available. If we oscillate the actual vacuum charge itself, unlimited additional free electrons also become available from the Dirac sea, since partial "unstripping" of the negative energy wells (each containing an electron) occurs. This "unstripping" of the potential of the well is due to vacuum potential oscillation, which oscillates time as well as space. To a negative energy well (positive time), the oscillation of its "time stream" increases the well's negative potential during one half-cycle and decreases its potential for the other half-cycle. Thus during half the oscillatory cycle, negative energy electrons may be lifted from the Dirac sea if the oscillation is sufficiently intense.

Therefore we obtain an attendant voltage and flow of mass current through the load. This voltage and current, by the way, are essentially limitless, and are free for the asking, assuming the proper "vacuum oscillation" is initiated and maintained so that electrons are continually being lifted from the Dirac sea by the time oscillations, and fed into the circuit.

Standard electromagnetic theory assumes that the vacuum potential is zero. It does not recognize the existence of massless charge, separated from charged mass. Therefore, orthodox scientists have never looked for a way to engineer the vacuum, because they have not realized it is composed of pure massless charge. Electrical physics has almost hopelessly confused charge and charged mass, thereby eliminating scalar longitudinal electrostatic waves. Expressed in the bird/high tension line analogy, the present electromagnetic theory restricts us to walking along the high tension line, laboriously carrying small batteries and power units, unaware of the limitless, surging power beneath our very feet.

Electromagnetically, we have been rather like one of the five blind men who touched an elephant. We have only touched one small portion of the electromagnetism "elephant," yet we thought we had grasped the entire beast.

Return to Table of Contents Next Page

OVERTURE TO A NEW AGE TECHNOLOGY

From an Update by Dr. Rolf Schaffranke

> " ... as there has been a steady erosion in U. S. activity in the basic sciences, there has been an increase in the level of such activity in Western Europe and Japan. If that continues, we are in danger of losing our position of scientific leadership."
>
> Val. L. Fitch, Prof. of Physics at
> Princeton University, in an interview with
> *U.S. News & World Report*, June 21, 1982.

West Germany:

The prototype of the first motorcycle equipped with a "flux-motor" (the inventor's designation for a tachion field converter or gravity-field generator) of 2 KW output has completed a test run of approximately 20,000 km without repairs. Principle of operation: The T-converter is constantly recharging the electrical power battery of the vehicle, resulting in an operation which is noiseless, fuelless and polution-free. After completion of all scheduled tests, a public demonstration to the press is planned within a few weeks from the time of this writing. (June 17, 1982) See photo, Pg. 19. A new factory for the production of gravity-energy-field converters in the town of Oldenburg is scheduled to produce individual home heating units. The design of automobile power plants based on the same principle is also in preparation. Reported efficiency factors of several hundred percent are anticipated.

Japan:

Mazda continues research on the "magnetic Wankel engine" (45 HP) for electric automobiles, based on the principle of Kure Tekko. The gravity research of Prof. S. Seike is now sponsored by Hitachi, after millions of Japanese received updated information via popular science magazines. It is interesting to note that Saudi Arabia has decided to invest $100 million in Hitachi stocks.

France:

The "Kromrey" converter has completed tests at the Institute for Magnetostatics, University of Strasbourg. One prototype delivered 700 Watts at 600-1200 RPM, at a fairly uniform rate. However, France, as the No. 1 developer of nuclear energy in Europe, is still reluctant to concentrate her energies on a revolutionary new technology which could result in drastic curtailments of her well-developed nuclear industry and associated exports. See photos, Pg. 22.

Denmark:

Dr. Jensen, physicist at the University of Kopenhagen, reported test results of 300% efficiency with a specially designed, 5-coil transformer in connection

Return to Table of Contents Next Page

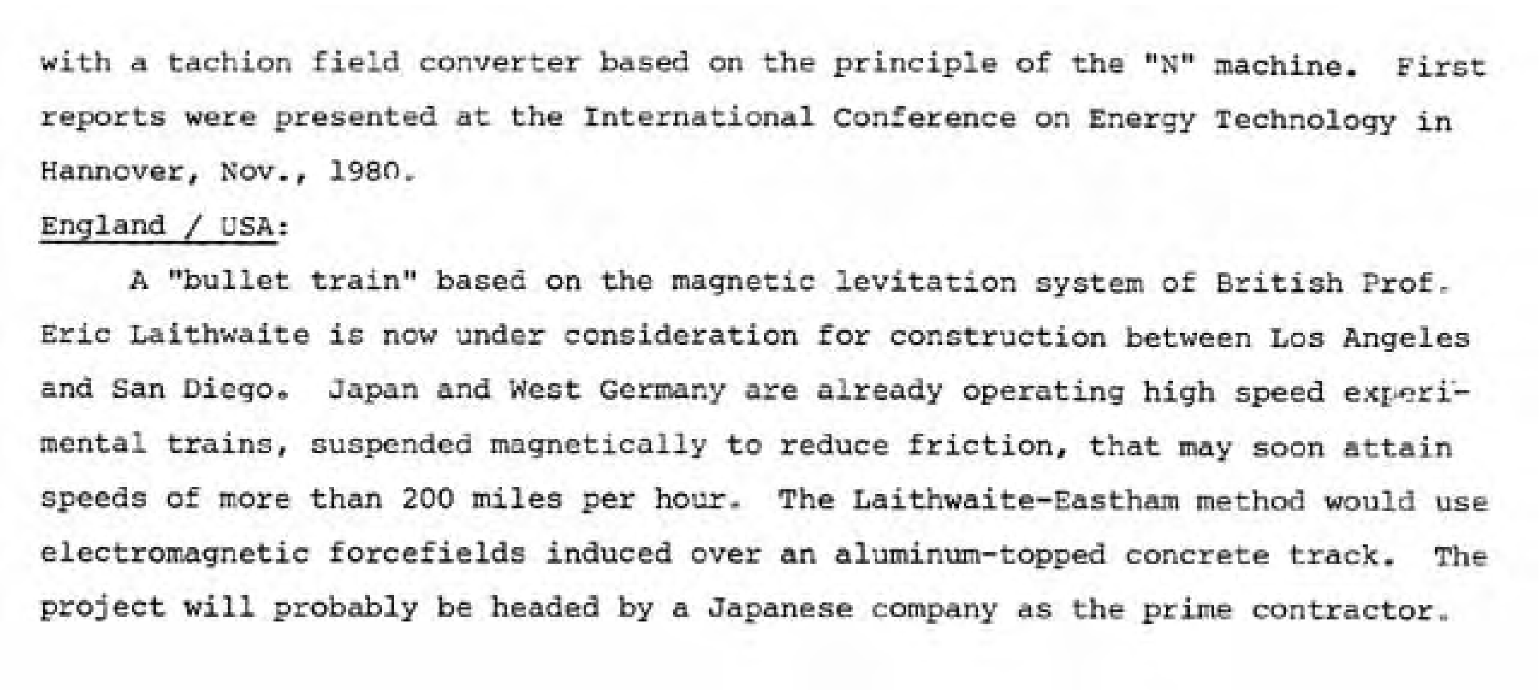

with a tachion field converter based on the principle of the "N" machine. First reports were presented at the International Conference on Energy Technology in Hannover, Nov., 1980.

England / USA:

A "bullet train" based on the magnetic levitation system of British Prof. Eric Laithwaite is now under consideration for construction between Los Angeles and San Diego. Japan and West Germany are already operating high speed experimental trains, suspended magnetically to reduce friction, that may soon attain speeds of more than 200 miles per hour. The Laithwaite-Eastham method would use electromagnetic forcefields induced over an aluminum-topped concrete track. The project will probably be headed by a Japanese company as the prime contractor.

Return to Table of Contents Next Page

West German gravity-field unit in operation

Return to Index **Next Page**

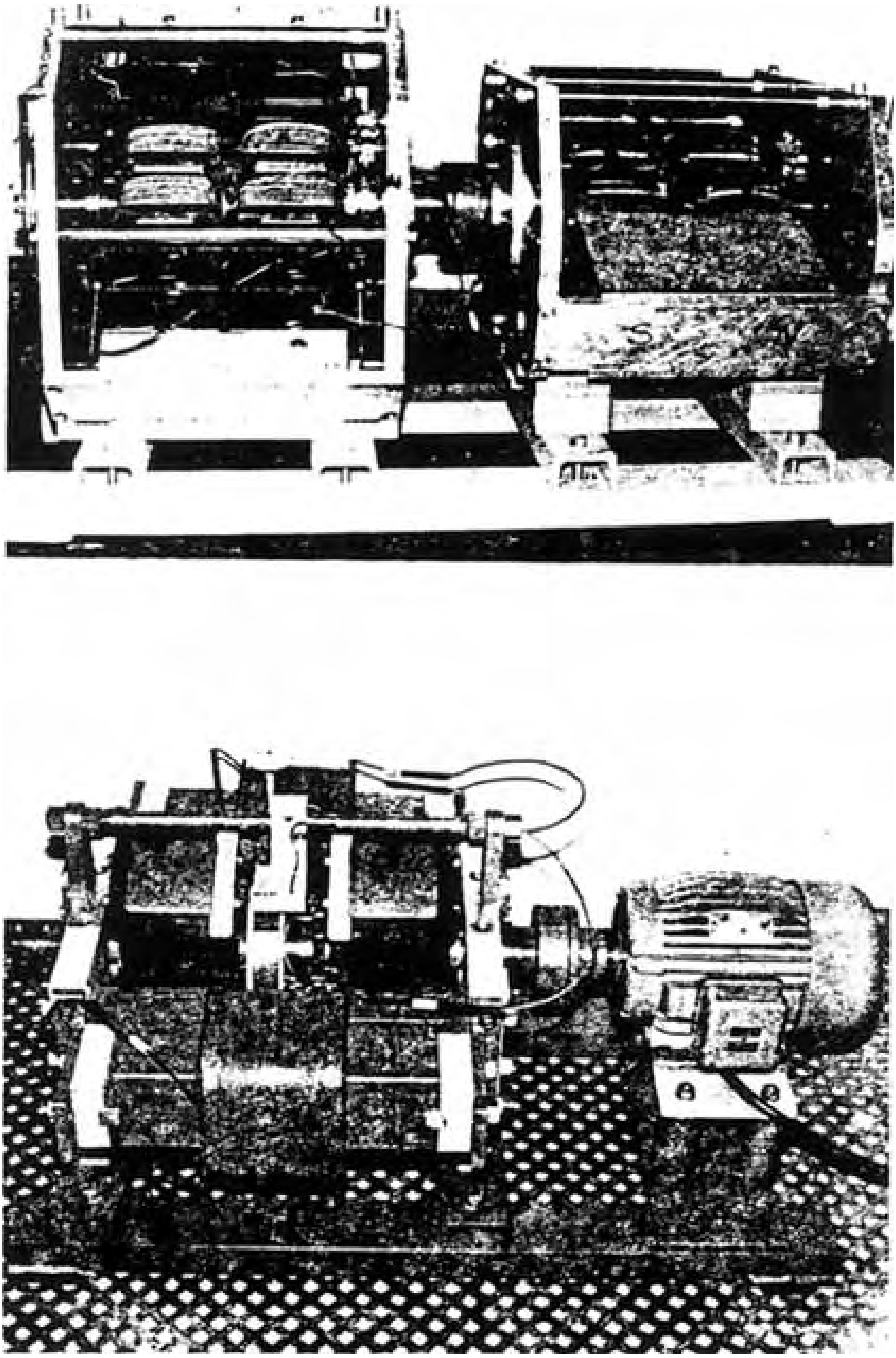

Swiss-French gravity field converter system "Kromrey"

Return to Index **Next Page**

REVIEW AND OUTLOOK

THE DEVELOPMENT OF POST-RELATIVISTIC CONCEPTS IN PHYSICS AND ADVANCED TECHNOLOGY ABROAD

Rolf Schaffranke, Dr. h. c.
Member A.I.A.A.

This paper was presented at The First International Symposium on Non-Conventional Energy Technology, Toronto, Ontario, Canada, Oct. 23-24, 1981. (Reprinted by permission of the author.)

"Far more harm is done to the progress of science by skepticism than by gullibility". (Dr. Robert Wood, Director R & D, McDonnel-Douglas Astronautics Div., Cal.) NASA - Langley Research Center in Hampton, VA is named after Samuel Pierpont Langley (1834-1906), an American astronomer, engineer and architect.

Langley worked out the aerodynamic principles for manned flight. In principal, his calculations were correct; but the structural materials he used for wings and engines were ,insufficient. In December of 1903, the *New York Times* published an editorial complaining about his foolish dream and the associated waste of government money. The editorial predicted that: "man would not fly for a thousand years".

Only 9 days after the editorial was published, the Wright brothers made their first successful powered flight at Kitty Hawk. (Dec. 17, 1903). But, again, as late as 1905, the *Scientific American* suggested the happening was a hoax - two years after the event which changed history. Let's reflect for a moment on the fact that only in a single lifetime, man has journeyed from Kitty Hawk to the surface of the moon!

The analogy to the problem of alternative energies and their development is, of course, obvious. We try to educate the preaching orthodoxies. But orthodox solutions are no longer enough. Our system of education can only give from the past. The present must operate on inspiration and intuition, or the future will be lost. The past no longer has enough of the answers for arising crises for which there are no precedents. To keep pace with the requirements of the future, we must begin to teach not only what to learn, but how to learn, how to analyze, how to search for the truth. Only then will be able to recognize discrepancies between experiment and dogma and to cope with problems which appear to be just beyond the corner. History has shown again and again that no single individual can rise above the species without being persecuted. The reasons for that are:

Some of the most powerful forces in homo sapiens have always been
Greed, Pride, Ego, Fear and, above all, the Desire to Control Others,

.

A brand new German book titled *Energy in Abundance*, which is evidently creating a considerable impact in the European community, challenges the contemporary high

Return to Table of Contents **Next Page**

priests of science, the "Guardians of the Status Quo" and uncovers an almost incredible narrowmindedness and ignorance in science, politics and economics, "approaching the criminal" as the author puts it. The contemporary waste of genuine intellectual creativity is castigated as a cultural scandal approaching barbarism. The book stresses that a truly promising approach to the so-called energy crisis requires a fundamental and thorough re-evaluation of the theories, dogmas and axioms which form the basis of present day science and the foundation of our technology. As in the past, the formidable inertia of the establishment is not only based on inaccurate, incomplete and outdated information, but also on its reluctance to accept new ideas. The honest search for the truth first requires that we first demythologize generally accepted claims of so-called experts of the past. An excellent introduction to this problem is the article "Resistance by Scientists to Scientific Discovery" by Bernard Barber in *Science*, Vol. 134, pp. 596-602, of Sept. 1961, or the paper by Stephen C. Brush in *Science* of March, 1974, titled "Should the History of Science be Rated 'X'"?, culminating with a statement of Huxley:

> "Authorities", "disciples", and "schools" are the curse of science and do more to interfere with the work of the scientific spirit than all its enemies".

The former astronaut, Capt. Edgar D. Mitchell, one of our famous contemporaries, also stressed the need for an open mind!

> "History has shown time and again that important scientific discoveries generally happen only when someone steps outside the limits of his traditional discipline and looks at something from a fresh point of view. Then what should have been obvious all along comes into focus".

Those of us who are familiar with the pioneering efforts of Dr. Hans Nieper of Hanover will appreciate this statement by an ex-astronaut.

One of the most important cornerstones of today's scientific dogmas dates back to 1905, more than 3/4 of a century ago. Lest we forget, television and radar, jet-planes and cyclotrons, moonrockets and close-up photos of the planets of our solar system were nothing but wild fantasies of science fiction then.

The speed of light was assumed to be constant and the maximum possible speed in the universe; vacuum was to be a total void. In reality, the speed of light as measured in the Michelson-Morley experiments was not at all the same in all directions. The "ether-drift" still amounted to the respectable velocity of about five miles per second, and similar results were obtained by D. C. Miller in a series of experiments extending over twenty-five years, from 1902 to 1926.

[Return to Table of Contents](#) [Next Page](#)

> "Even worse, the measurements showed such marked discrepancies with previous results as to occasion a distress call to the U. S. Coast & Geodetic Survey, whose surveyors repeatedly remeasured the length of the tube and found no error there," reported the *Popular Science Monthly* March 1934 issue. And it continued: "More recently, speed of light observations only emphasized the apparent erratic behavior of the light beam that the scientists were attempting to plot. On some days it seems to travel faster than others by as much as 12 miles a second. Its speed seems to vary with the season, and also in a mysterious shorter cycle lasting about 2 weeks. Finally, the scientists ended by taking an average of all the readings which has just been announced as 186,271 miles per second".

In his paper "A Critical Look At the Theory of Relativity", Library of Congress Cat. No. 77-670044, F. K. Preikschat compiled all known light velocity measurements during the past 300 years or so, from Olaf Roemer in 1676 to the Laser measurements conducted by the National Bureau of Standards (USA) in 1972.

Of the 27 experiments undertaken in this field, 18 of them after the turn of this century, it must be concluded that the velocity of light as measured within the reference system of our earth has changed as much as plus/minus 50 km/sec. during the past century. Preikschat plotted a curve of these deviations and suggests a possible relationship between sunspot activities and changes of the earth magnetic field during the time period in question.

The "universal constancy of light" appears, therefore, as somewhat shaky "empirical evidence" for our cornerstone in physics, especially since Einstein himself has gone on record as saying:

> "If a single one of the conclusions drawn from it (relativity theory) proves to be wrong, it must be given up; to modify it without destroying the whole structure seems to be impossible".

Another modification concerns the definition for the term "vacuum" in physics, as documented by a paper in the *American Scientist*, March-April 1980, titled "Is The Vacuum Really Empty?" by Prof. Walter Greiner, University of Frankfurt, BRD, and Prof. Joseph H. Hamilton, Vanderbilt University, Nashville, Tenn.

The authors conclude that a neutral vacuum is by no means as "empty" as previously claimed in our textbooks, and suggests a new definition as follows:

> "The vacuum is the lowest stable state that a region of space can have while being penetrated by certain fields".

Because of the tremendous time lag in our educational system, many research

Return to Table of Contents Next Page

projects and their ensuing experimental data have been withheld from public scrutiny. The scientific community tends to have a vested interest in preserving the system it created and of which it is a part. It responds to new situations through the coloration of this attachment. A case in point are the carefully conducted experiments of T. T. Brown with charged bodies in a high vacuum, as described in my booklet *Ether Technology* (1977). These experiments suggest the actual presence of certain fields in vacuum, whether we call them gravitational field, tachion-field, ether-field, neutrino or Fermi-sea, etc. is of secondary importance at this moment. Although Brown spent, reportedly, more than $200,000 of his own funds over several decades on such experiments, he was nevertheless unable to have the results published in the scientific media of America.

Things are even worse when it comes to experiments conducted abroad, which often tend to confirm disregarded experimental results on this continent, as we shall see shortly. To highlight the wide discrepancies between orthodox (and obsolete) dogmas and actual, physical realities pertaining to the true subatomic structures as we know them to be today, let us briefly review the structure of the matter:

A molecule is the smallest division of a substance. Further division would cause it to cease being a substance. The smallest of true molecules can be illustrated when we use the globe of the earth for our standard. If a single drop of water were magnified until it was as big as the earth, each molecule would be about the size of a TENNIS BALL.

On the next step down, an atom is the unit which makes up the nature of the molecule, consisting of the nucleus and the surrounding electrons to render the atom "stable". An atom of hydrogen contains one proton and one electron to balance or neutralize the proton. Matter then is divisible into electrons and protons. But - and here comes the rub: Between electrons and protons are spaces so vast, in comparison with the masses of each, that, if the proton in the carbon atom were the size of a golf ball hanging from the ceiling of the great hall at Pennsylvania Station in New York, its electrons would be represented by six small wasps winging in a little knot against the four walls of the gigantic structure of the building! In effect, one could claim there is little final solidity of substance to anything: The universe consists of "emptiness", charged with electrical energy! If we translate the above to the measurements and terminology of the physicist and "magnify" the atom mathematically, with all its distances and dimensions kept in proportion so that the orbit of the electron would have a diameter equal to that of the earth about the sun, approximately 184 million miles, the diameter of the electron itself would only be 2,000 miles, and the diameter of the nucleus, where electron itself would only be 2,000 miles, and the diameter of the nucleus, where

Return to Table of Contents Next Page

mass and weight of the atom are truly concentrated, can be taken as 2 miles only. We thus obtain a picture of a central mass with a diameter of 2 miles (nucleus), another object with a diameter of 2,000 miles (the electron, in the case of the hydrogen atom) at a distance of 92 million miles away from it, orbiting the nucleus. Evidently, there is plenty of room inside this system. And "room" is not a vacuum, it is not nothingness, but space itself, spatial energy, a field which can be identified with the ether of the past - and of the future. Nobel prize winner, Max Planck, during a lecture in Florence, Italy, once made a truly remarkable statement which describes the problem facing the physicist today:

> "As a man who has devoted his whole life to the most clear-headed science, to the study of matter, I can tell you as the result of my research about the atoms this much:
> 'THERE IS NO MATTER AS SUCH!'
> All matter originates and exists only by virtue of a force which brings the particles of an atom to vibration and holds this most minute solar system of the atom together. ... We must assume behind this force the existence of a conscious and intelligent MIND. This mind is the matrix of all matter".

This cosmic matrix is needed if we want to explain "action at a distance", lines of force, stresses, a magnetic field and so on. When the concept of the ether was abandoned, it had to be replaced by the concept of "space" instead. In reality, we merely switched terminology. We used to say that "ether fills all Space". But "filling" is not exactly the descriptive word to use. Perhaps we should rather define it: "Ether is a condition of space in which electrical manifestations for the atomic construction of materials is possible". This primordial energy is "free" or in an uncondensed state. It exists in interstellar space but remains unrecognizable until it begins to coagulate or gets into a vortex pattern. It constitutes the formative field force of nature, an immense reservoir of latent energy.

The claim of our textbooks that the Michelson-Morley experiment "disproved" the existence of the ether is incorrect. It merely disproved the existence of a noticeable ether drift or drag. As an analogy, if somebody would postulate that the absence of wind disproves the existence of the atmosphere around our planet, the fallacy of this postulate would be immediately apparent to all.

"Michelson and Morley centered their attention on the Earth's orbital velocity (30 km per second). They had no knowledge of the existence of galaxies; of motions of galaxies in relation to each other; of the motion of our solar system in our galaxy..... Their negative results are explainable on the basis of pre-1900 classical mechanics, so provide no proof of the absence of ether or Louis de Broglie's

Return to Table of Contents Next Page

'subquantic medium'. Thus, the limited information available to Michelson and Einstein is emphasized by recent findings, particularly in astrophysics", writes Dr. H. C. Dudley in the *Bulletin of the Atomic Scientists*, January, 1975, under the title "Michelson's Hunch Was Right". And Dr. Dudley continues: "In fact, 1929 saw Michelson still attempting to experimentally demonstrate the ether, which his intuition and reasoning told him ought to be present".

"Today most persons are largely unaware that the ether concept began to be seriously reexamined by two of physics most notable theoreticians, Paul Dirac in 1951 and de Broglie in 1959, both Nobel laureates. The ether is now being called the "neutrino sea" by astrophysicists, and has been characterized as an energy-rich particulate, subquantic medium. A rather voluminous literature on the subject is accumulating as indicated by a recent review, *The Cosmic Neutrino*, with 665 references covering only the period 1965-1972......It appears that an open-minded reexamination of this area of physics is long overdue in order to open up new avenues of approach to this pressing problem".

Michelson, Dirac, de Broglie were not the only Nobel prize winners in favor of a reexamination of the ether question. Others were Stark, Arrhenius, A. H. Compton, Lenard, H. Yukawa, and Frederick Soddy, the British scientist who, during the convention of Nobel laureates in Lindau, 1954, described the current dogmas in physics as "an orgy of amateur physics" and "arrogant swindle" - with particular emphasis on the theory of relativity. ("The Wider Aspects of the Discovery of Atomic Disintegration", New World Publications, St. Stephens House, Westminster S.W.I.)

It remained largely unknown, even among physicists, that Einstein himself had serious doubts. In 1949, he wrote to his old friend, Maurice Solovine, who congratulated him on his 70th birthday:

> "Now you think that I am looking back at my life's work with calm satisfaction. But, on closer look, it is quite different. There is not a single concept of which I am convinced that it will stand firm and I am not sure if I was on the right track after all".

And one of my Canadian friends who I believe is present here today stated quite correctly in a letter to me:

> "The myths around Einstein are really not of his making as he himself often questioned his own conclusions and I felt he was quite aware of his own limitations which others were covering up. The physicists protect their members much the same way that the medical profession protects its own."

The purpose of these remarks is not to spark another controversy about the merits of the theories of relativity. Rather, they want to show that so-called merits of the theories of relativity. Rather, they want to show that so-called

Return to Table of Contents Next Page

"scientific evidence" is a rather elastic term and that "verification" is always a relative affair. Or, as K. R. Popper so aptly put it:

> "Only in our subjective experience of conviction, in our subjective faith, can we be 'absolutely certain'".

This symposium presents an excellent overview of research and experiments reported from English speaking countries, or experiments published in the English language, as for instance from Japan. However, as Dr. Tenhaeff from Holland has formulated the problem,

> "It is important to pay attention to publications of colleagues in other countries. In the Netherlands and Germany, every academically educated man or woman knows at least three foreign languages. English and American authors only pay attention to material in English. In my opinion, this leads to an 'impoverishment' which puts a drag in science. Some seem very chauvinistic and seem to believe that only the researches done in their country are important. I regret this. Science is international, and international cooperation is of the greatest importance in science".

In presenting some brief reports from abroad, I shall attempt to bring corroborative evidence in support of alternative technology developments discussed during our meeting here in Toronto. Let me start with the most sensitive and secretive society, with Russia.

In his German original of the paper "Some Remarks on the Shielding Theory of Gravity", which was published in the Hanover Proceedings in more detail than the English abstract, Dr. Hans Nieper mentioned the Berlin engineer Levetzow, who was one of the first to postulate a combination pressure and shielding theory for the phenomena of gravity. Levetzow was soon overshadowed by the rising star of Albert Einstein, but one of his followers, Horst Pinkell, went to Russia in 1928 as an exchange student - and never returned. Together with Russian scientists, he was asked to prove the actual existence of the cosmic radiations postulated by Levetzow, now designated as neutrino sea or "tachion field". The work was done in Sterlitamak, South of the Ural mountains, and in 1938, sporadic reports reached the West about the discovery of extremely short corpuscular waves by Pinkell and Gorjew. These were the long sought "Hemmstrahlen" or drag waves associated with gravity. Rumors have it that a metal alloy was soon developed which acted as a drag wave polarisor and, in 1947, agents reported experimental Russian aircraft using this discovery. In 1951, a former German Junkers aircraft engineer reported about his activity pertaining to a super-secret project code named COW-7 in Siberia, a disk-shaped craft which could rise vertically and which belonged to the

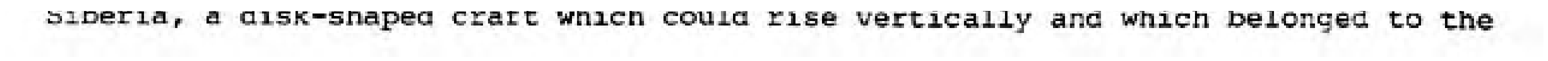

Siberia, a disk-shaped craft which could rise vertically and which belonged to the

[Return to Table of Contents](#) [Next Page](#)

secret weapons then developed by the Soviets. Whether it was related to the Canadian AVRO project or something entirely different remained unknown, only the Levetzow gravity theory was assumed in connection with the activity in question.

Japan: My esteemed friend, Prof. Seike, who kept me informed about the research conducted by him and his co-workers, surprised me once with a significant statement, which I feel is important to repeat for the benefit of frustrated researchers on this side of the Pacific: "In the early stage of my studies, physicists could not understand what I was doing, while actors and actresses did so by intuition". It was the private support of about 800 individual artists, etc., in Japan which allowed Seike to work, experiment and publish his findings, and not the support of the Japanese equivalent of a National Science Foundation or a similar bureaucratic institution in that country. Again and again, I was surprised how quickly and eagerly American experiments, as for instance the almost forgotten "Ionocraft" of the late Major de Seversky, were duplicated and explored even in Japanese secondary schools. Admiral Rickover has more than once complained about our intellectual complacency: "As a nation, we seem addicted to 'spectatoritis'. We sit in the bleachers and let the game of life unfold before us".

The results are known: The Japanese have achieved top position in science and technology, and not only with tachion beams melting stones, producing "biased" water with a freezing point of minus 4 degrees Celsius, and a G-power generator obtaining 400 V DC output from a 10 V input driving source. I wish to take this opportunity to thank Prof. Seike for sharing his work with us in this country, and wish him the best of success for the future.

France: The work of SEPED was explained by M. René Louis Valée and some French publications in the area of gravity-research have been translated by the NASA translation service, as for instance the theory of the electromagnetic Magnus-Effect (of Marcel Pagés).

Specific mention must be made of the Raymond Kromrey G-field generator, a French-German-Swiss development which appears very closely related to the American "N" machine principle, the "Sunburst" machine and the "Permanent Magnet-Motor" Patent of Howard Johnson. Utilizing either permanent or electromagnetic modules, the generator output exceeds the input by far. One prototype delivers approximately 700 watts at speeds varying between 600 and 1200 RPM. Models with an output of 100 to 160 kw are presently in the planning stage. Several European patents have been granted against heavy opposition, especially from French industrial circles. The development was about 40 years in the making and the inventor stresses the need for a reinvestigation of electromagnetic and gravitic forces, as well as a reexamination of many cemented dogmas in present-day physics.

of many cemented dogmas in present-day physics.

Return to Table of Contents Next Page

Germany: A system of rotating electromagnets and magnetic "stator rings" of the German "Kunel" Generator complement the general trend of the American and French generators mentioned before. The first prototype was reported to have been tested in April, 1980. Here, again, very heavy opposition from the German orthodox science establishment. Of special interest is a comment by Nobel laureate Werner Heisenberg, reported to have been made vis-a-vis the inventor:

> "I think it is possible to utilize magnetism as an energy-source. But we science idiots cannot do that; this has to come from the outside".

More details about French and German developments have been published in the new book already mentioned, *Energy in Abundance* by Hilscher and, hopefully, an English version of this important work will be available by next year. Generators which do interact with surrounding energy fields have been designated as "open systems", in contrast to our present "closed system" technology of gasoline and diesel engines, steam turbines, nuclear power plants and so forth. The theoretical basis for all "open systems" is the realization that there simply is no such thing as "empty space". In reality, space is filled with an extremely energy rich, subatomic and subquantic continuum. In short, a "pre-physical state of matter", which can be tapped, as for instance by the spinning of magnetic systems which extract momentum via a vortex-formation of the universal "tachion field".

Austria: This small heartland of Central Europe is proud of her native pioneers in science, among them Dr. Wilhelm Reich, Viktor Schauberger and Karl Schappeller. But after Dr. Reich died in an American prison in 1957 and Victor Schauberger died only five days after his return from the USA (in 1958), where he received an incredibly unfair treatment from certain industrialists, there is little enthusiasm among Austrians to touch base with correspondents on this continent. Viktor Schauberger's son, Walter, refuses, as a matter of principle, to communicate with Americans, no matter in what language. To illustrate the type of "contract" Viktor Schauberger was forced to sign, it stated:

> that all of his patents would become the property of the "consortium",
> that any future inventions or developments would also become the property of the "consortium",
> that he would refrain from discussing any of it with third parties,
> that he would refrain from publishing,
> that all plans and models would remain with the "consortium" after his departure from America.

There is a favorite saying in that small country which is much to the point:

> "After the Greek philosopher Pythagoras had discovered his famous

Return to Table of Contents
Next Page

> theorem, he sacrificed a whole hecatomb of oxen as a thanksgiving to the Gods. Ever since, all the oxen in the world are running scared whenever a new truth has been discovered".

Knowledge of the ideas of Schauberger and Schappeller appears to be essential for an understanding of our changing picture of the physical world. As an example, the former forest ranger Viktor Schauberger had watched numerous times some trout standing, apparently motionless, in the strong currents of Alpine mountain streams. This and many similar observations of nature brought him to the gradual development of his "implosion-theory" and associated hardware (see Jensen paper elsewhere). His son is continuing the work in the "Pythagoras-Kepler" School in Bad Ischl. Another prolific researcher is the Viennese engineer Franz Seidl, known for his studies of the use of magnetic field effects as "trigger energy" for voltage and power multipliers.

Holland: A small country with strongly developed intellectual independence appears to be Holland, where Prof. J.M.J. Kooy developed his "Space Dynamics". His hypothesis of gravitational action is based on the realization that all celestial bodies can be conceived as practically transparent in relation to the size of the elementary particles and their mutual distances in the structure of matter. Gravitons or tachions, coming from all directions of deep space, can pass through a celestial body like water passes through a fishnet, and only a minute fraction of the cosmic gravitons or tachions will be intercepted. Study of the effects of the tachion field or space continuum of the fabric of material substances, magnets and crystal lattices can lead to development of converters and novel production facilities for readily usable, conventional E/M energy. When the American inventor Edwin V. Gray was blocked in his efforts to develop his "Pulsed Capacitor Discharge Electric Engine" for which he had received U. S. Patent #3,890,548, he found open ears and open minds in Holland.

The fact that Gray was named "Inventor of the Year" from the Department of Patent Rights in Los Angeles, California, where Dr. Chalfin of Cal Tech confirmed that:

> There does not exist an even distantly similar engine to this one in the world. Conventional electric engines use up power. In this system energy is used up for only a minute part of a millisecond. It operates without heat loss and there is no energy loss whatsoever with this system",

had more meaning to the people of PHILIPS and others in Holland than to the energy establishment in his native America.

England: There is perhaps no more controversial inventor at this time than

There is perhaps no more controversial inventor at this time than

Return to Table of Contents

Next Page

John R. R. Searl and his Levity Disc, powered by the Searl Effect generator. It is claimed that model craft have been demonstrated with flight characteristics showing anti-gravitational and inertia-free properties. Demonstrations have been photographed and videotaped. Searl calls his generator a "Gyro-Flywheel High Energy Density Mechanical Magnetic Device" and claims that the magnets used are not the common type, ordinary magnets. The Australian engineer Athol Park of Melbourne, Australia, who visited Searl a few years ago described it this way:

> "From a briefcase, Mr. Searl takes a bar magnet and two steel rollers. He places the rollers at either end of the magnet. They appear to be attracted to the poles - but, when they are pushed gently around the corners of the magnet, they chase each other round and round.... "

The engineer quotes Searl:

> "The crystals in the magnet are changed by putting it in a magnetic flux oven and by frequency control. The result is motion of a magnetic field - a completely new source of power. Units (modules) based on this principle are what power my levity unit - the motor and generator built as one, fuelless unit".

In one of his numerous newsletters he states further:

> "The power plant is a self-contained, onboard closed system of conversion elements with energy input from a latent storage source".

Because Searl has never divulged all details and his explanations are generally of a non-technical, "unscientific" nature, many investigators have chosen to dismiss him as a swindler and imposter. In sharp contrast to this is the opinion of a well-experienced American engineer who reported after a visit with Searl:

> "I found it refreshing to note the keen sense of moral responsibility which he feels for the use to which his discovery is put... The Searls live in modest circumstances. Luxuries are few. They do not own a car. Mr. Searl bicycles several miles daily to his place of work in Maidenhead. There he is an electrical technician and training supervisor in a bearing factory".

But perhaps most significant was the following observation of the American visitor about Searl:

> "We were also interested to hear him say that many of his ideas came to him while he slept. He would go to sleep with questions, and awake with answers... and refreshed. He seemed to have the

and awake with answers... and refreshed. He seemed to have the

Return to Table of Contents Next Page

> ability to restore vital powers more quickly than most people".

And the American concluded his report with the statement:

> "The impression we gained of Mr. Searl was one of a very sincere and generous person, a dedicated and tireless worker with perfect confidence in the outcome of his efforts, a man of high principle with a keen sense of loyalty to those who are helping him, and of responsibility for the future welfare of our planet and its people".

In view of the very contradictory nature of comments and impressions by those who have met Searl, it might be wise to reserve final judgment on the Searl-Levity-Disc and its inventor at this time.

New Zealand: Some of Searl's keenest supporters and observers of his work are in New Zealand and Australia. And one of the sharpest young physicists in that country (new Zealand), unhampered by calcified ideology, developed the mathematical foundation for a theoretical anti-gravity effect, using the hydrodynamic analogy of the ether. He uses the equivalent of the hydrodynamic Reynolds number, above a certain value of which the laminar flow past a solid surface breaks down and becomes turbulent due to shear rupture of the viscous cohesion of a fluid, together with his deduction that anti-gravity could be obtained by creating a "vacuum: in the ether in such a way as to correspond to the production of turbulence in a magnetic field. Turbulence in a fluid occurs when Reynolds number RE = 1, when the inertial shearing force equals the viscous cohesion force in the fluid in the vicinity of an approximately spherical body immersed in the fluid, resulting in separation of the boundary layer and formation of a wake, i. e. vacuum or free surface around the body. He calculated the threshhold for turbulence in the ether, = a vacuum = an anti-gravity effect, using the refined value of ether kinematic viscosity which he obtained.

His theory would confirm the observations of very high voltage produced in the case of the Searl Levity Disc, and the very high voltage needed in the vacuum experiments with flying disc-shaped bodies of T. T. Brown, the "ionocraft" observations of Major de Seversky, the "Electro-Field Rockets of Prof. Dudley, the observations of Northrop Corp. with "Electro-Aerodynamics in Supersonic Flow" and similar American documentation and patents.

Since the New Zealander is now in the process of attempting to publish his theory in a reputable British magazine, I am not at liberty to divulge his address prematurely. Interest in New Zealand and Australia is running high, and possible breakthroughs in energy technology should not come totally unexpected in these two countries with strong intellectual ties to England.

[Return to Table of Contents](#) [Next Page](#)

Conclusion

There is now overwhelming evidence for the actual existence of a very high density, energy rich space continuum formerly called "ether". This space energy can be concentrated, gathered, focused, magnified and compacted by magnets or magnetic materials, which appear as the new core materials for converters of all kinds, utilizing the cosmic energy for new technology applications.

Crystals are energy-sensing and channeling devices; they can serve as transducers, as for instance in the Moray-Device, or the "rock electricity" of T. T. Brown.

Combinations of crystals and magnetic materials appear to be the mainstay for the development of "hyper-space" - converters and generators of power - utilizing free, cosmic energy in our future.

Unbiased, totally honest and impartial reexamination and reevaluation of all scientific theorems and dogmas, as well as unhampered opportunities for the pioneers of the new technology on the American and European continent appear to be the prerequisite of success.

Those who accept that anything of the nature of ideas or ideology is permanent, use all their strength to hold on to it, and to condemn all who do not agree. The next stage of this spiritual illness of the mind is to build a monument to what they perceive is Truth, and then to build a monument to their own permanence in the world. This, they do by seeking fame or fortune, and authority over others. This is the contemporary, planetary psychology situation, and thus are we losing the gift of perspective and horizon and are no longer able to discern what is important. Skeptics are on record as saying:

> "We are not running out of natural resources or intellectual talent, we are running out of the freedom of innovation that made the countries on the North American continent the most advanced nations on earth".

The future will show whether this freedom, which includes the freedom and the duty of taking risks, can be restored to the degree our Founding Fathers have been striving for. Only then will we know whether the late Wernher von Braun was not too overly optimistic when he stated:

> "The cosmic age will bring a dawn of knowledge not yet envisioned. It will dwarf all our present knowledge and ideas and, through its vastness, may even bring men closer together on their own planet".

Return to Table of Contents

=The Unveiling of the Hidden Knowledge and the Secret Space Program

The Unacknowledged Special Access Programs: Advanced Technology, Mind-Control, Spiritual Power and the Corruption behind Closed Doors

By Aug Tellez

Introduction 7
  Getting This Out Of The Way 7
  Psychic Operation 8
  A Light for the Others 8
  Natural Security 8
  A Balance of Mystery and Truth 9
Chapter 1: The World Situation 9
  What is happening? 9
  The Mind Virus 9
    The Vampire Effect 10
    Sub-Human Entities 11
    Human-Hosts, Replicants, Synthetics and Carriers 11
    Genetic Manipulation 12
    The Ego Mind 13
    This Time is about Healing 13
    Healing Through Unity 14
  Underground Bases 14
  Cloning 14
  Celebrity Cloning 15

        - Initial Explanation....15
        - Cloning Centers, Underground Bases....15
        - Programming....15
        - Experimental Operations....15
        - Genetic Engineering....16
        - Immortality....16
        - How did it get to this?....16
        - Deeper Meaning....17
        - What is next?....17
- Chapter 2: The Awakening....18
- Chapter 2.1: The Machines of All Time and Space....18
    - Re-Creating The Kaleidoscope....18
    - Sand-Castles of Time....18
    - We're Waiting For Ourselves....18
    - The Goal of Two Societies....19
    - Our Forgotten Other Half....19
    - All Is But An Illusion....20
    - Life Itself Is The Highest Illusion....20
- Chapter 2.2: The "Awakening" of 2010....21
    - Memory Suppression....21
    - Life and Consciousness....22
    - The True Human Potential....22
    - The Light-Body....23
    - Technology of the Awakening....23
    - Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness....23
    - Time Dilation and Temporal Recurrence....24
    - Crystal Technology....24
    - Time Crystals....25
    - Quantum Tunneling Diodes....25
    - Tachyon Fields....25
    - Upgraded Chronovisor....26
    - Time Tunnels....26

The “Trip” Chair..26

Kozyrev Mirror..27

Universal Recurrence..27

Remove Viewing, Temporal Lensing..27

Temporally Linked...28

Natural Vortex Energy Locations and Geography.......................................28

  Mount B..28

Particle Accelerator Temporal Manipulation...28

Riding the “Timewave”...29

Dematerialization and Rematerialization..29

“Jump Rooms...29

Ion Shields...30

Instantaneous Healing..30

“Living Light” Sound-Harmonics and Electromagnetic Healing...................30

Electrogravitic Craft..30

Energy Cores and Storage Capacity...31

3.5D Replicators...31

3.5D Sensors..31

3.5D Printing of Clones...31

Cybernetics:...31

Neural-interlace/4D Connectome Recording and Simulation......................31

EEG Heterodyning..32

Artificial Hallucinations...32

Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)..32

Utilizing The Temporal Body...33

  Manipulation Through the Temporal...33

Astral or Light Body..34

Monitors can emit consciousness/brainwave altering frequencies..............35

The Upgraded Chronovisor...35

Community Servers/Living Server Systems..36

Cloning, Cloud Supercomputing and Consciousness Transfer.....................36

Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel...37

Cloning as Offspring....38
Inserted Memories, Digital Dreams....38
Co-Inhabitation of Bodies....38
Clever Replication a Digital Mirror....38
Necromancy and Soul Stealing....39
Ancient Technology, Non-Human Entities....39
Created Beings....39
Black Goo – AI Nanites Vampiric Hive Mind....40
  Recreating or Replicating the Universe....41
Ancient Spiritual Technology....41
The Unveiling, Underground Bases, Increasing Awareness....42
Learning about the Unveiling....42
Giant “god-men”....42
Breakaway Civilizations....43
Ancient Civilizations....43
  Atlantis....43
  Aegypt....44
  Ancient Technology....44
The True Elite....44
The Definition of Human....44
Souled and Non Souled Beings....44
There was once a Unified Planetary Consciousness....44
  Power and Knowledge....45
  The (watch)Keepers of Earth....45
Energy Vampires....46
Spiritual Technology....46
Simulations and Learning....46
Ascension – Clearing Trauma....46
Trauma-Clearing, DNA Unlocking Technology....48
Everyone Who Seeks Harmony and Healing Can Be Healed....49
Programming....50
Layers of Programming and Decision Making Roles....50
  Programmers....50

  - Sub-Level Programming....50
  - Dissolution of Programming....51
- Advanced Technology, Deprogramming The Mind, and “Space/Time”....51
    - Dreams, Technology and Inner-Space Travels....52
    - Awareness Feedback Device....52
    - Oneness, Higher Being, Completion, Eternity....53
- The Holographic Reality....53
- Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control....53
- The Great Catastrophes....53
- I was brought into the Time Program....54
- The Effects of Temporal Manipulation....54
- Temporal Influence....54
    - Simultaneous Interaction and Incorporation....54
    - Retro-Causality....54
    - Quantum Holography....54
    - Time Paradox....55
    - Breakaway Civilization....55
- Underground Cities....56
  - Technological Advancement....56
  - World-Wide Control Systems....56
- Solar Warden....57
- Team Light and Team Dark Non-Interference Agreement....57
- All Contact Must Be Initiated On an Individual Basis....57
- A Singularity of Change and Awareness....58
- Advanced Technology and Ancient Knowledge....59
    - Cloning, Atmospheric Civilizations, Genetic Stability....59
- Training, Simulations and Learning....59
    - Cloning and Cerebrally Enhanced Soldiers – MILABS....59
- Psychological Testing....60
    - Hallucinatory Holographic Mind Probing Technology....60
    - Virtual Simulator Training Scenarios....61
  - Combat Simulators....61
  - Conditioning....61

Programming and Resistance to Psychological Trauma......................................61
Accelerated Mental Functioning...61
Technology, Accelerated Learning and Repair...62
On Orgone Energy..62
Energy Generators and Serums..62
Genetic and Cybernetic Enhancements..63
Developing The Self...63
Multiple Factions of Space-Fleets, Multiple Goals..64
Large-Scale Catastrophes or Series of Meltdowns..64
Unsustainable...64
Earth Changes..64
Mass Consciousness..64
Wars and Manipulation..65
Financial Unrest (or reset)..65
Self-Responsibility..66
Clones and Synthetics..66
Spiritual Power and Divine Truth...66
Synthetic DNA, Synthetic Genetics..67
Time as Mathematics, Series of Equations, a Programmable Storyline..................67
Chapter 2.3 Previewing Solutions..68
Neutralizing Polarization...68
Removing Etheric Implants...68
Clearing Astral Body...68
Knowing the True Self; Soul Actualization, Unveiling..68
Taking Responsibility..69
Returning To Center Emotionally and Mentally (spiritually)...................................70
Vampires in the “Awake” Community...70
Heart Center and Psycho-Spiritual Fulfillment (self-actualization)..........................71
Go Within to Observer the Self...72
Organic Feedback Loop..72
Technological Feedback Loop...72
Breaking Soul-Contracts...72
Help Everyone..73

Chapter 2.4 How Timelines Collapse..73
Repeating Cycles of Time...73
Beings From the Future...73
Beings From Collapsed Timelines..74
Beings Who Would Never Have Existed in the First Place......................................74

Introduction

The following is first-hand testimony from an operative of the secret underground military projects who also has familiarity with the breakaway civilization above and below (within) Earth. These civilizations utilize advanced technology to extend the length of their lives and enhance cognition beyond what is conceived of in today’s society.

I was brought to the underground bases as a child as part of an experimentation program to test the population for various blood-groups and personality traits that would be of use to certain factions of the military in ways that would serve to preserve and strengthen humanity’s outlook in the future.

These tests involved everything from combat training to sensory deprivation based emotional-mental endurance, and introduction to the highly advanced application of esoteric knowledge.

To clarify, I am bringing this information to you as part of my assistance in the overall unveiling of the hidden knowledge of the ages of humanity related to these secret projects as well as the true galactic history of Earth, humanity and man-kind. These projects are officially (unofficially) known as “unacknowledged special access programs”. The unacknowledged aspect means that there is no formal acknowledgement of these programs without a need-to-know basis. There is no paper-trail, no disclaimer, no formal authorization or internal publication outside the programs themselves. The classification of these programs also generate a rather strenuous situation regarding funding and secrecy but this is done for the benefit of humanity. There are secrets and manipulations that can be seen as a threat but this is all tied together and so by releasing the secrets that are part of the defense we would be inadvertently exposing the public to the threats themselves. Now is the time where we are reaching a point where the public is ready enough to handle the basics of this information regarding the secret projects and the galactic history of Earth and humanity. This is also out of a necessity to preserve civilization for if we wait any longer we may not find the same opportunity again.

Getting This Out Of The Way

I am writing to you from the perspective that I am not certain just how far from completing my mission I am. I write to you today in order to bring to you many personal experiences from the viewpoint of an operative of a select ultra-military secret society.

The utilization of such an operative is not glorified, pleasant, or entirely righteous. This story entails sanctioned and guided acts of extermination, ritual abuse, the use of yet to be announced advanced technology. There is unfathomable corruption, the thirst for power and spiritual knowledge that has been protected throughout the course of the multiple civilizations that have preceded this one and will continue to be maintained into the future of this civilization.

Psychic Operation

I was utilized in psychic-operation amongst other areas.

This was used during certain aspects of training and conditioning to enhance training and conditioning in the bases using advanced technology. Some of this includes what would be viewed as traumatization by those without the training and conditioning. There is a fine line between training and conditioning and abuse. Part of this disclosure is related to how this line may have been crossed in certain situations and regarding mass-mind control upon the population. There are other aspects of travel, viewing, reconnaissance and research that will likely be explained in another release.

In order to heal I have to tell the world. You can help. This is happening right now in the underground bases and around. It’s not a joke. Your family is at risk. Your future generations are at stake. There is conditioning to encourage those involved to believe they are performing a duty for humanity. Sometimes they turned us against each other, but those grudges were never truly of the heart, but the mind.

A Light for the Others

What I hope to achieve with this is a smoke signal for those who have also been involved and question whether their purpose has been fulfilled or falsified for the protection of a corrupt system that intends to neither benefit the whole nor repay the individuals involved for their time, energy, and lives.

Multiple explanations can be summoned to describe the reasons behind the necessity for the training we have endured, the missions we have completed, and the control mechanisms we have both assisted in establishing as well as suffered from the backlash of. Ultimately, it is of my belief that all that was done was done for the good of humanity, even when we were not sure of the intentions of those in front of us, to the side of us, or behind us compelling us forward with the command to complete each objective.

This is not so much a panic or cry for help as this is simultaneously a call to awaken and a signal that humanity may have avoided complete disaster but is not out of the woods yet.

Natural Security

My experience is far too vast and I have seen too many succumb to stress from the realization of the truth of this reality to consider my own suffering to be enough cause for my denial of the necessity of these programs. Humanity has been protected and at a great cost to those who have been on the front lines, both in mind and body and spirit.

There are many who will have concerns, doubts, as well as reservations about what I am about to reveal and if I offend anyone then know that my intentions here are to assist in unveiling the truth. The truth can be harsh when one has been held back from the true nature of existence for so long. The way of life can become so fragile and artificially constructed on impulse and mystery that the harshness of the truth acts as a great catalyst for spiritual growth which allows us to elevate ourselves to a new perspective. Prepare yourself for such harshness, prepare yourself for growth and elevation, and prepare yourself for the mystery to simultaneously resolve and increase in intensity and magnificence.

A Balance of Mystery and Truth

Mystery has been a way of life for Humanity for so long that when a great mystery is converted to truth some seem to feel that a portion of existence dulls and loses the luster of exhilaration. This story is quite the opposite. The more that the great truths of this reality were discovered and revealed, the more complex and mystifying life became. As well, the more empowering and complex experience becomes, the more self-control one must develop in order to regain a harmonious way of relating to one another, the self and the universe. Harmony is key, balance is everything and there is a light within that represents the piece of our self that reflects the utmost power and brilliance of all existence.

Chapter 1: The World Situation

What is happening?

The Mind Virus

The main reason for this disclosure is to inform humanity of what can be referred to as the “mind virus”. This is a viral infection of the holographic system which consciousness operates through. There is a complete backstory of this virus in the history of this civilization and how society has been plagued for centuries if not millennia. The basic idea is that this virus operates through consciousness and can be transmitted through electromagnetic waves.

The virus contains an etheric form which can manifest as a kind of crystalline nanotech on the physical plane and this is a form of synthetic sentience which feeds on the low frequency bio-emissions of sentient life. Human bio-emissions contain energy and can act as a carrier for information and living essences. This technology requires low frequency bio-emissions because that is the nature of the design as a kind of synthetic bio-weapon whereby the goal is entrainment and enslavement. As well, the specific nature of the sentient nano-technology and the counterpart etheric holographic form is only capable of converting the low-frequency bio-emissions into energy as the higher-frequency bio-emissions require more energy to contain and synergize with and this would render the entire process of leeching energy non-productive.

This is essentially a synthetic intelligence that operates through scalar and holographic living bio-emission fields and hijacks electromagnetic emitters to position itself within them acting as a parasite to a host. This is said to have been present around this civilization for some time, waiting until there was enough technological advancement so that spreading across the entire civilization in a short amount of time before discovery would be possible. This is through the media, internet, and radio systems.

The physical counterparts are microscopic nano-technology parasites functioning as individual units to a hive mind similar to the mycelium of fungi in the field of mycoscopy. The individual units act as parasites serving to infest and deliver nutrients to the hive from each host.

The Vampire Effect

This is the nature of the parasitism that has been present throughout the ages of this civilization feeding on human suffering and relying on the cover of confusion and misdirection to survive.

There are processes that have developed that assist in the infection and co-inhabitation of the parasite entity within a human vessel utilizing the human as a host. This effectively utilizes a frequency ‘net’ within the auric field of the human to limit bio-emissions and emotional-mental processes to produce just the psycho-etheric patterns that the virus can sustain itself within and through.

This is no different than the way candida can infest a human digestive system and the human will experience cravings for beer, candy, carbs or other foods that contain the sugars, yeast, or carbohydrates that will break-down into nutrients that are the most nutritious for the candida fungus. These foods are *not* most nutritious for the human body. Yet the human body will withdraw and crave these foods in greater amounts until the fungus is swept clean from the digestive tract through probiotics, fasting, flushing, cleansing , detoxing, and even exercise as lactic acid from physically strenuous activity released from the muscles has been known to help destabilize and remove this fungus. Why is it that the individual will crave these foods that are not the normal food intake? Because this candida fungus can infest the host and produce enzymes which generate a chain-reaction which leads to the chemical signals in the brain producing the sensation of extreme hunger or cravings for just these foods which will benefit the fungus.

The same process occurs with this psycho-etheric parasite, except the foods of this parasite or lower emotional bio-emissions and this is through lust, hate, angst, jealousy, feelings of hopelessness, lack of self-worth, spiritual degradation, addiction, and generally all that equates to a lack of self-awareness.

Whether or not this is the direct intention of the largely microscopic entity is for the most part unknown, and this is as well similar to candida. The candida fungus is not said to be a 'mean' fungus that wants us to be hungry and craving beer and sweets all the time because it likes to ruin the human body, the daily routine and the focus of a clear mind. This is simply the way the parasite has grown to operate within a niche of breeding itself through the digestive tract of larger animals and through this the unaware and infected human suffers greatly. Even those who are unaware and not infected can suffer because of the way the behavior and thought processes of those who are addicted to these foods can become distorted.

The lack of self-control over food intake is synonymous with the lack of self-awareness and self-control over lower bio-emission expressions of the spiritually degrading behavioral patterns that become routine for the psycho-etheric system of the infected human. Ultimately, this is all about energy. What benefits the parasite that seeks to gain energy by leeching instead of self-guided production is going to take away from the host that is being leeched from. As such, any energies, food intakes, or spiritual activities that cleanse and enhance the personal power of the individual will automatically reduce the parasite's ability to feed by the very nature of the energy systems. What empowers the individual, empowers the self and develops self-control, what develops self-control contains one's energy and reduces the likelihood of one's energy being leeched. This is similar to energy vampires in human form and this goes with the understanding that these are humans or non-humans who have been entirely overtake by the parasitic consciousness and cannot sustain their own auric fields without having to pull from the energy of humans who do generate their own aura or bio-emissions.

The vampire effect relates to how a person can be leeched from only to a certain point. Once the process surpasses a certain critical point then the individual becomes similarly engaged with only the lower bio-emissions as higher-frequency bio-emissions would then become harmful. So in essence, when the vampire feeds too much on one individual, then that individual becomes a vampire themselves.

Sub-Human Entities

The humans are utilized as hosts and energetic sources in this bio-emission battle while there are other classifications of bi-pedal organisms that appear to be human but are not human. As well there are bi-pedals that are non-human entities entirely.

Human-Hosts, Replicants, Synthetics and Carriers

The human hosts that are entirely vampirized by these entities become walking carriers for this virus and are effectively human in physiology but there is more to the human than just the physical body. In these individuals, that energetic component of the human is removed and replaced with the synthetic energy of the hive-mind organism.

There are replicants which are a form of clone, along with synthetics. These beings are cybernetic bi-pedals similar to humans but their energy is too not from an organic biological matrix but a technologically propagated synthetic soul-matrix. The hosts are here according to their agenda to subdue and feed off of the bio-emissions of the human race, although there are no absolutes. The agenda is not always so clear within the various groups who often have a mix of bi-pedal operatives. There have been plans to form of a salvation timeline because the only way to keep the whole process going is through a negation of degradation of the human race as this may close out this timeline forever.

Those with human organic soul-matrixes would simply transport into the next available universal time-stream or harmonic, while those with synthetic matrixes would have to technically travel to the next harmonic using a passageway otherwise they would remain here throughout a zero-point collapse and this would be synonymous with entering the abyss with no viable passageway out until the entire universe is brought back to one organism again.

This is also the process that is described to explain what happened previously as there are 7 harmonic temporal layers of the universe meaning everything is contained in a multitude of 7 octaves similar to light and sound frequencies or harmonics. And so there are 7 civilizations or 7 parallel time-streams and humanity passed through these to get to the last three where the physical forms are located. These are, from the higher to lower, Gaia, Tara and Terra. We are on Terra and if the information I was given is correct then we are currently on a transitory timeline located around Tara where people are noticing changes occurring and multitudes of aspects of the time-stream disappearing and changing simultaneously. Tara is where the last universal harmonic was left through a collapse of time and consciousness into zero-point and this resulted what was termed “the lost souls of Tara”. These are the discarnate souls of the previous civilization that could not travel to the next harmonic and remained without a body in the abyss. Thus they have plagued this plane since then, since ancient times possessing and driving humans to madness and there are explanations that this is the original passageway for the parasite into this universe. This event, the consciousness collapse of Tara would have acted as a fracture in the over-mind of the species generating a kind of cosmic schizophrenia and enabling all kinds of disharmony and distortion described in the previous sections.

The solution to this was described as a healing process that is taking place to purge the virus and that this is painfully similar to the human body purging a sickness however there are methods to promote healing and reduce suffering and confusion however people must contain themselves and their energies and strive to know themselves otherwise they will not be in control. This virus feeds on fear, confusion, spiritual degradation and the energy from degrading acts, lust, addiction, sexual misery and whatever fashion a human can be made to bring the bio-emissions of the energy centers down to a low enough level where co-inhabitation can be formed with an entity that has no access to the higher levels of self-awareness, higher-consciousness, love, or the universal spiritual experience that the balanced human is capable of.

Genetic Manipulation

There are explanations that there are genetic modifications taking place now to possibly preserve this civilization and that there were ancient genetic modifications made to induce a sense of stupor, worship, selfishness, an inability to more easily comprehend the higher spiritual awareness and the occurrence of time and consciousness and other spiritually and consciously limiting aspects. Part of this explanation is that holographic consciousness itself is a kind of limited result, a shadow effect resulting in the presence of the higher-dimensional essence of the soul which is more or less constrained by the body. An extension of this is that the entire holographic universe is a kind of false-light system designed to bend the original liquid etheric light of the soul and spiritual awareness into a rotating, recurring format which ultimately converts the original spiritual essence into an energy generator for entities that are entirely holographically represented. The souled human in this situation is a being who is originally from outside of this holographic projection.

The Ego Mind

In this view, consciousness is the ego-mind which is the false-sense of self, not to be confused with the shadow self which is simply the counterpart of the aspects of the self that we consciously agree with and are aware of in daily life. The shadow self is a result of having limited holographic consciousness and the ego-mind is projecting that limitation of consciousness into the shape and image of a self that is only a figure-head for the real individual essence which is spiritual in nature and immortal.

This Time is about Healing

Everything that is happening now is to motivate people towards healing and unification. This is not about a hive mind mentality where individuality is pushed out.

> "The best I can do for you is become your enemy."

There is a saying that represents the concept that the most beneficial role a person can play in helping one to progress is to challenge them and give them the opportunity to rethink their strategy and in turn improve further with each interaction. This may seem paradoxical but it is effective and explanatory. What we are facing is the greatest challenge humanity has yet to face. This is the unknown, the final frontier, the mind. The mind is our challenge.

There will be more on this later and in other publications but there are indications that the physical embodiment of the human has been altered to introduce implants, genetic modification, susceptibility to disease and ultimately fear-based programming of the brain. Without going further into this prematurely, the lower aspects are at odds with the higher aspects unless we as the midway come forth to mesh the two together in equilibrium. Until then, there is chaos, lack of self-awareness, and suffering.

There are many possible futures converging at this time. There are some less than favorable paths and there is the opportunity to merge with a very powerful path for humanity. This all comes down to how people will cope with information that takes them outside their comfort zones and eventually changes their view of mundane reality forever.

If people choose violence and lack of self-awareness by ignoring the spiritual aspects then that is what reflects into their life experience. When people develop a higher sense of self-awareness and spiritual harmony they can reduce the suffering and reconnect with the lost aspects of our spiritual identity for the first time in the publicly accepted version of recorded history.

There is technology that can change the world and there are great truths which can enable people to know themselves in ways that they never thought possible. Everything is at our fingertips and so we must come together as this is the way we solve the problems and cure the dysfunction.

Healing Through Unity

In various research projects the conclusion of how to heal this spiritual rift was by exposing each other, to each other. When we are all together, aware of each other and our selves, united in the goal of bringing higher-awareness and spiritual harmony then we can observe and acknowledge when one is slipping into chaos or disharmony. When this occurs and we are in proximity to each other in a healing circle or a kind of social unification then the members of the group at large can work together to pull the chaotic one back into harmony towards the group. The group of healers united together in like mind, spirit and body is too strong to be overtaken by the parasite. So when one individual at a time is faced by this problem the whole group can come together as one and assist in healing the individual by pulling them back to reality, back to wholeness and back together with the group. This is literally how the situation is solved in all realities and a breakdown of this feature of society is how the dis-ease proliferates.

Underground Bases

The active aural research program is part of a process to simultaneously verify as well as initiate and demonstrate controlled insanity/sanity.

These programs utilize the deep underground military bases to perform psychic and psychological research experiments on non-consenting youth and adults as well as consenting participants.

There are various means of generating the illusion of consent or even lack of consent and this is all being monitored and handled by oversight authorities.

These operations are directly related to the identification and understanding of the mind-virus and all weapons, defenses, and resistances known to man.

Cloning

Advanced technology is used to transfer consciousness from one cloned body to the next so that a continuous study can take place before, during, and after the death experience of one individual. More will be explained on this later, however this is through the use of advanced supercomputer systems that can function to entangle and then transfer the electromagnetic consciousness of the individual so that they are ‘remotely activated’ within a cloned version of themselves via a ‘brainchip’ (brain to machine interface: cybernetics).

Celebrity Cloning

Celebrities at the cloning center would like you to see what is really going on behind the scenes which involves trauma-based mind control, heavy technological programming and complex layers of influence throughout their life.

Initial Explanation

Many celebrities, entertainers, athletes, musicians, models, writers, producers, actors, all kinds of people from various industries are silently brought to the cloning centers for pay-to-play sessions.

They asked myself and others to pass the message on to you and that you would have to look for the signs and symbols in their media in order to see their hinting at these events.

The situation is heavily controlled using advanced monitoring technology and brainwave/EEG cloning technology. This is technology that can read the brain and determine what the mind is anticipating or speculating upon.

Their reputations are used as leverage as well as their safety and comfort. The level of programming and mind control experimentation that takes place makes it easy to manipulate a person’s brain into regressing into a state of trance that lacks the ability to remember or clearly organize experiences regarding the cloning centers.

Cloning Centers, Underground Bases

Deep underground military bases, 3 miles below the surface, are used as laboratory centers as well as a completely stocked underground city-base. These are connected with high speed electromagnetic drive pods.

Individuals are transferred to the brain through an extensive cloning and temporal body transfer process. The electromagnetic shell of consciousness within the brain is relocated to the body of a compatible clone and various training, conditioning, programming, experimentation, or pay-to-play experiences take place.

Programming

All individuals are programmed to various degrees in order to maintain control and secrecy over the situation. This programming involves very advanced technology, trauma-based mind control, and energetic attachment via beliefs and emotions.

Experimental Operations

These bases were also used hand-in-hand with military operations which were geared towards discovering and controlling all aspects of the human mind.

Genetic Engineering

Experiments involved genetic engineering to produce soldiers, psychics, hybrids, and others that would be able to carry out operations. This leans towards MiLabs and the military faction’s experiments which go beyond the basics of cloning.

Immortality

The initial research included the goal of physical immortality. In many ways it can be said this has been achieved, however there are side-effects and difficulties.

If one does not activate their higher consciousness, then the effect of time dilation causes the conscious mind here to reach limitation points in experience.

What has been termed “blank slate/state” technology has been used to ‘reset’ the perception of time through memory in order to keep a continuous progression viable for the conscious mind. Without this the unconscious and the conscious mind merge.

This begins involvement across time with advanced technology that can operate on the soul level or the conscious mind’s level of access to the unconscious and soul-memories. The Universe instantly creates a cosmic backstory based on the conscious-mind’s access to the unconscious. By controlling the conscious mind’s access to the unconscious the entire backstory of Humanity can be altered and new connections can be bridged in the future.

Earth is essentially a time-ship through which consciousness ascends towards higher states of awareness and self.

How did it get to this?

After WWII when the United States war-faction firebombed Germany and melted many of the inhabitants of the cities into sludge in the bomb shelters and streets of the cities, a group of NAZIs traveled to Antarctica.

Russia noticed the movements to Antarctica and the United States sent Admiral Byrd up with 3600 marines, planes, a battle cruiser and smaller vessels. They returned in defeat and only a few words were mentioned in regards to what actually happened and how they were defeated.

The rumors spread and words of gravitic drive craft, and undersea or under-ice bases originated from this encounter.

From what we were informed as well as directly experienced, this was the ice-base in Antarctica where an underground base was found already constructed with very advanced technology.

Here, cloning was deployed along with mind control and temporal manipulation technology.

Soon, cloning was offered as a way to avoid assassination as well as to prolong and protect the original body in daily life.

After that, celebrities and politicians were replaced with programmed clones who could carry out the orders of the NAZI faction.

This was the beginning; the groups utilizing this technology now are no longer confined to the NAZI faction.

The groups you see today, in control of these operations, are considered MiLabs, secret societies, think tanks, military factions, and other control groups.

Deeper Meaning

The individuals that have been cloned are given very in depth views of the way this society functions and the source of power and change in the Universe. Many individuals are present here from what would originally be another “time” or “timeline”.

There is a possibility that this notion is due to inserted memories and traumatization although until all the information is released we will not have a definitive conclusion.

As of now it is a possibility that these individuals are from another time that goes beyond our comprehension currently.

It may even be a possibility that these individuals have come from ALL of time looking at the very far past and into the very far future of Humanity.

These are the kinds of notions involved with the experiments taking place in the underground bases and military operations.

What is next?

Share this information with those you feel are ready to know and help Humanity come to terms with our journey here and what is going on in these kinds of experiments and take responsibility and power for our own existence.

We are in something that can be called “The Unveiling of The Hidden Knowledge”. This is a cyclic process that takes place to advance a civilizations knowledge of existence. As part of the civilization moves into higher advancement another aspect may lag behind. Either the future will slow down, or the past will speed up. We are experiencing the past speeding up and all the events that have taken place to influence our civilization are coming to light. This is as much a natural cycle of consciousness as it is the eventual revelation of secret knowledge and the particular connection to the secret projects.

Those brought to cloning can use their free-will to state their lack of consent to violence and harm and then live in commitment to that by not harming or accepting violence in their lives. The Spiritual Law of Harmony rules in all planes and dimensions.

Chapter 2: The Awakening

Chapter 2.1: The Machines of All Time and Space

There are machines that are capable of accessing genetic memory and unveiling what can only be referred to as the experience of awareness or existence in between the physical planes of each time or each physical life experience.

Even if they are simultaneously occurring moments of a transcendental nature, there is still an illusory ‘space’ where there is perceived separation and through this there is a perceived ‘space’ that is in between each existence.

Re-Creating The Kaleidoscope

This space is only perceivable when viewing from the physical perspective, like seeing the spaces between the angles of a kaleidescope only when viewed through the lens and similarly, technology can be used to recreate what this kind of illusory space might look or exist like in between the multi-dimensional realm of the continuous experience stream. Through this, these ‘hypothetical’ in between spaces that are understood as only illusory productions of biological consciousness, can be generated through advanced technology with the capacity of recreating that kaleidoscopic view of which the human biological perspective is only a small fraction or even fractal.

By recreating this biological perspective, they can recreate the illusory spaces in between planes and through this they can literally access and experience what are stored in the DNA as the experiences of existence beyond the physical limitations of consciousness and into a multidimensional experience of time and space.

All becomes a never ending sea of consciousness, however this is the given. The trick is to take from this sea of consciousness and slice it down into conceivable chunks from which you can reintegrate a previous personality or identity and continue on a stream of consciousness that would otherwise be tossed to the cosmic wind like fractal stardust as the sands on the beach of hyperspace.

Sand-Castles of Time

We are effectively building sand castles out of these sands of time that would otherwise be swept back and forth into and out of the ocean of life and the universe only to momentarily wash up again as the identity which we once knew.

There is technically nothing wrong with either view, because then, the spaces between each million year occurrence of just so happening to wash up again as you, would be imperceivable because they are only there to be seen when you are you again.

We’re Waiting For Ourselves

Yet, if there were some who had attained the awareness of the true nature of reality and were standing there on the beach waiting for the rest of their cosmic family to come to shore again, it could be perceived as waiting the majority of the time for a very small occurrence and then being alone again up until the kaleidoscopic fractal inverts itself to the point where everyone now is standing on the shores of the sands of time rejoicing and meeting with each other.

The Goal of Two Societies

My goal in coming to you people on the surface of what is called “Earth” is to help assist in the process of bringing the two societies together, one who’s been waiting for thousands of millions of years, and the other who has no idea that they’re even late (or very on time).

You see, those are the only two situations where there could be even the possibility of perceiving such an injustice as experiencing the pain of loss or misconstrued identity in this universe. It is as if two families met at the cross-roads at different angles, and while one waited for the other to arrive, they could merely supply themselves with glimpses of the ruins and remnants of the evidence that the other civilization existed at any one time.

This is not how life must be and this is kind of like a cosmic waiting game of hide and go seek our two mirror civilizations have been playing with each other. One who’s highly technologically advanced and the other who is more or less the spiritually advanced of the two.

Both of them have a unique view of the universe and they are both only complementary to each other as a whole. The more they sat around and waited for one another, with one leaving the clues that the other would find and even switching roles sometimes in different ways, they became even more and more complimentarily associated as the very essence of experience they seem to be missing more and more is the experience of one another.

So as we perceive these as being great losses or great gains of time and culture we are really only remembering ourselves and coming together as a unified existence that once met upon the surface of Earth as a single family covering the entire harmonic spectrum of the Universe.

And that should be the goal of any progressive, any one pushing a new culture or an old culture, anyone looking to teach the young and protect the knowledge of the old, while also protecting the innocence of the young and perfecting the age-old knowledges that have defined our presence in reality.

Our Forgotten Other Half

One could even say we’ve sometimes gone off the deep end, leaving ‘land-minds’ of sorts to process the information that would be present when we are gone. Of course, what we could’ve found when this occurred, is beyond anything we could’ve ever expected and that is the true nature of the complexity of creation and existence. That everything follows the cycles of creation and destruction, however, the memory of everything that has been created can always be accessed (depending on the beach) and therefore the creative aspect is legitimate, yet it is the destructive aspect that is illegitimate and illusory.

Yes, through advanced technology this can be proven in ways that can be verified and transmitted using advanced technology, since that is how we like to verify things these days. There is, of course, a human looking at the screen, the read out, or the situation from a physical perspective to verify the verifications.

All Is But An Illusion

There is a notion that there is only one civilization, one society, in actuality that has been chasing itself through time. That we cannot actually find the same place and the same time where both meet because we are each other from different perspectives. We can only get a mirror where this is possible, or an ‘hyperdimensional’ internet channel where we can stream one’s consciousness in from the other reality and interact with them through that here in this reality. Thus the two civilizations can achieve the experience of their own civilization and the other civilization by merging with one another but only through a remote contact, just a relay, not an actuality.

The harsher, darker side of this is that any civilizations between then and now that we did happen to meet that proclaimed to be “the one’s” we have been waiting for, and in essence our darker or lighter halves, are actually the impostors who exist entirely apart as a sub-set and have taken a liking to trying to trick the reality out of one of the other civilizations.

Of course, if these are another form of being altogether, and this is all brought about through an interplay of consciousness and deviations from an original core reality, then all will resolve itself when that core reality is alone again. This indicates that all realities will either move closer and closer to the core until they unify, or farther and farther out until they transmogrify beyond recollection.

Life Itself Is The Highest Illusion

That hyperdimensional internet channel that is created out of advanced technology to merge the parallel realities (rather distastefully and maybe disastrously at times) allowing different civilizations to interact with each other, well this has been created using technology.

But then it was discovered that this is the ultimate nature of DNA. That this was ‘invented’ similarly as a higher-dimensional internet so that various natures of different planes and dimensions could interact with each other and experience the richness of life.

The universe becomes more like a living arcade game then, where people come from all over the galaxy to ‘plug in’ to human experience and figure out what it means to be physical, human and on Earth.

This does tie into the larger nature of reality, of star-races, and of time and consciousness that is often reversed according to how things seem to play out in physicality. We are moving through time, but just because we are moving one way, does not mean that others could not be moving the other way.

We have to live our lives with meaning knowing that the message we pass on, the duty we vow to accomplish and ensure is meaningfully connected on either end to the passage of the past and the power of our ancestors to the function and knowledge of the future creating a clear passage between the realms, the generations, and the ends of the universe. This is what literally ties the universe together, and the mind if each individual. This purpose is contrived when not carefully shaped within and so we must help all who seek purpose learn to become the blacksmiths of their own destiny through the temperance of desire or temptation and the discipline of developing strength and maturity where there was once weakness and naivety.

Once we know this is our duty, this is what we seek to achieve and this is what is done, then we can become part of the universe forever. Of course, this may only be temporary in the larger view of things.

Chapter 2.2: The “Awakening” of 2010

In 2010 I was “awakened” at an event involving hundreds of people that play various roles in this society, the secret society and the breakaway civilization. This also took place across times or planes of reality.

During this event I was informed of the situation on Earth beyond the informational barriers that were previously in place to limit knowledge to only what is required for specific operations. I was essentially “informed” of the entire situation as much as the mind could handle this.

As this was occurring, the team that was updating me on the status of Earth was being informed of and given information regarding changes in the future. In short, interviews, posts, and releases that I have since made and will make in the future were touched upon. Quotes were given from the blog and interviews which were jokingly acknowledging verbal mistakes at the pronunciation of names, new perspectives and interpretations and more. As they were informing me, they were creating the future reality where this work takes place. That is the basis of this layer of the operations. They are creating the future person by person, layer by layer.

Memory Suppression

My memory was unsuppressed and all the traumatic memories and experiences were unveiled to me. This was almost as traumatic as the initial experiences themselves although precautionary steps are taken to ensure the safety of the individual. There are teams of highly competent individuals of all kinds with professional backgrounds of every field imaginable.

Throughout my entire life I have had memories of experiences, at this event I was informed what the experiences meant and what was actually happening. I've had memories of underground bases, cloning, stargates or interdimensional gateways, advanced technology, non-human or programmed biological entities, sacred knowledge and experiences, the power of the mind, and breakaway civilizations. Most of the time we are conditioned to believe these are past lives or experiences in another reality so as to enable continuity of operation.

Life and Consciousness

From my perspective and the perspective of others, we had reached the 'future' and had been sent back through the utilization of advanced technology and cloning to return to this time and inform others of the possibilities.

The True Human Potential

This is so far beyond what anyone can accept or comprehend fully that humans with an emotional, mental and physical aspect balanced as one is literally a cosmic reflection of the entire universe at once. This is the closest thing to a "god" in the flesh. And be sure, the entirety of this rests on the behalf that the man is protective and helpful. That the woman is nurturing and loving. That is the point. "god-man" does not mean a warrior who destroys everyone and anyone in his way getting what he wants. Nor does it mean some lusty and vampiric form of energy or mind control game. "god-man" means the love of the universe, the power of the body and mind, the peace and calm of the ultimate stillness all in one.

This power is in human DNA, originally so. And so this is awakened through walking the middle path of neutrality and 'splicing' the timelines down so that neither one nor the other gains the power of the individual. Through this, the individual creates their own power, right on the spot, without having to lean to one extreme or the other. This is like a cosmic, temporal balancing act with emotions and mind. Where we give our energy is created in the universe in the form of many productions and effects. If we are simply thinking in an imbalanced manner then we are producing timelines and side-realities that exist in the etheric realms that we can't see and these realities will reflect our inner imbalance. So then when we are balanced in our energies we are not creating one or the other side of divinity, we are literally creating the whole universe, cleaving down the middle and producing both sides of divinity or eternity equally and this is the only way to have an eternal production which does not eventually degrade to one extreme or the other. Through balance, we reach eternity and in this way all the power of the universe exists within a person through their ability to balance their emotions and mind and thus unlock the DNA.

The DNA is coded through bio-emissions of mind or emotional body. So we are constantly doing this and we are either producing a DNA code that reflects imbalance overall, or a code that reflects the eternal balance of the universe overall. This is through energy.

One reaches “heaven” or the deity planes or simply self-actualization, through three paths. These paths are through merits or devotion to principles, or through knowledge and expanding one’s mind to face the universe, or through great works and deeds to produce this effect of power, knowledge and emotion unto the world for others. Through these paths we create an impact on the universe and this transfers what we are temporarily and physically into an eternal energetic and physical production within the living universe, forever. To do this requires balance, focus, devotion, and discipline however because there are many distractions along the way.

The Light-Body

You must build your own light body. That is how it works.

Your heart creates a field and projects you.

Rely on yourself and use that for protection.

The power of the Universe will align with you if you align with your self.

Learn as much as you can and break through duality of love/hate, cultivate your knowledge of the self and use it.

What you want is what determines your power in the light-realms. Not how you want it. If you want war, then that is a sign of weakness and determination to produce imbalance and exist by that. If you want peace, that this is a sign of power and unity to produce further unification and exist by that. What you want determines your power, not how you obtain it.

Technology of the Awakening

Mental Manipulation Technology: Touchless Neural-Interface and Enhanced Awareness

I reached what the “Illuminati” called the “awakening”. I feel a better term would be “unlocking” as this event includes the removal of all traumas, perceptual distortions, and memory suppression.

This event was a presentation of very advanced technologies which allow the mind to transcend space and time. It was shown that death is similar to a phase-shift of consciousness as polarity and that the awareness of the being obviously continues. This was verified through advanced sensing devices which can visibly display frequencies which are invisible to the physical senses. This is also something that occurred earlier on as a child in the underground bases by transferring consciousness in and out of the body at will using advanced technology.

With these devices it is possible to see the “auric” or soul field of the individual. Any living body has a field which glows when picked up by these sensing systems. This field becomes more refined when viewing more advanced beings. Human have an extending mental field, and through this a kind of glowing awareness that is picked up. It can be fairly simply compared to thermal imaging except instead of differences in temperature it is the presence of a soul or mental energy field around a living body. The finer energies extend outward away from the physical body.

More than once, the situational requirements were satisfied in order to produce an ideal environment for maneuvering on the non-physical plane. This is the basis of how the temporal operations occur, yet these events were different than previous operations or experiences. The group I was in was going through the process of increasing the vibratory emissions of the “bio-mind” in order to stay focused on the increasing vibratory rate of the environment. Generators were utilized to create an effect on space/time which was perceived on the “soul” level. These events went into the discovery and explanation of the creation of the “godlike” powers of the advanced sentient technology which was capable of accessing the holographic nature of reality which humans can perceive through their bio-mind/soul. These technologies also enable one to perceive what is referred to as the “galactic history” of Earth and humanity. We are far vaster and Earth is far larger than people are lead to believe.

Time Dilation and Temporal Recurrence

Time dilation capabilities were shown which allow for accelerated learning of advanced and detailed material within a very short amount of time. This could be seen as a kind of viewing technology which one wears or looks into and vibratory emissions are scalar-linked to the brain which then allows the individual to perceive more information simultaneously. Then there is a very rapid pace of information streamed on a monitor or through a device and this is akin to watching a 20 minute instructional within a few minutes or a few seconds with very advanced minds. The more time is slowed down while information is accelerated, the more stress is applied to the brain. The brain tends to overheat and over stress with high-rates of activity, especially without conditioning. The conditioning is what allows one to utilize their mind and body in these manners. This is not unlike the very rigorous training and conditioning of certain monks or martial artists however there is very advanced technology and other hidden methodologies utilized.

Crystal Technology

These are capable of holding, transmitting, and amplifying consciousness frequencies. There will be a more detailed explanation in a later section.

The crystal technology is used in junction with the power devices to enable a scalar-mind link which allows the interface to be entirely mental or spiritual, IE: they are not controlled by hand but by focus and intention.

Time Crystals

These are utilized as a computer recording chip would be to run a program yet they hold memory in expanded space, IE: they function in a hyperspacial manner and are essentially hyperspacial computers. These are what will be introduced in junction with the quantum supercomputers.

Quantum Tunneling Diodes

This is technology that has recently be released to the public. New technology is developed years ahead of time and then slowly released to the public in a cascade of advancement. This technology deals with superconductive materials which enable the transfer of information at faster-than-light speeds.

The supercondutive material allows for 100% efficiency of data transfer. So what happens then when the efficiency surpasses 100%? This is possible because we are effectively in a simulated holographic environment. The experience is real, however everything is overlayed via particular limitations for the ‘local-environment’ which is like a central data processing limitation. When these parameter are surpassed, one by one, through a coalescence of refinements and methods, the result is akin to a glitch in a computer game system. They have accessed faster-than-light technology and the mind is also capable of this naturally.

With this technology, however, the information can be received a very small amount of time before it is sent. So if a person is absolutely surely going to press the button to send the message, then right before their finger hits the button that message will be received on the other end, as if magic. If they simply play around and pretend to hit the button but very quickly turn away, then of course nothing happens. One can, however, ‘fool’ the universe in other ways and this is simply through the reception and transmission of information across what should be secure information barriers.

Tachyon Fields

These are programmable fields emitted through advanced technology. I was shown how tachyons are the sub-atomic “particles” or energy formation which flows both ways in time. We are receiving tachyons from the “future” in order to layer the flow of time in one direction or the other. Tachyons are the name of the particles/energy packets that we perceive from this perspective as reaching us to initiate the “future”.

Thus, when a stream of tachyons is concentrated and accelerated, this results in an acceleration of the temporal field and we literally experience an acceleration of time. This can be focused and targeted on a specific piece of equipment to change the way information will be sensed and to essentially allow that device to sense ‘ahead’ of the present moment.

The same effect can be produced with the neurological processes of the human as the physical coupling of the brain to the mind can be altered through a concentrated tachyon field to enable the neurological and mental processes to begin to perceive information ‘before’ the brain is actually physically processing the data. Therefore the mind of the individual will be present in the body while the senses and other perceptions will literally be expanding into the future beyond the present moment.

This is highly confusing at first and requires much training and conditioning in order to make sense of enough to operate in any kind of effective manner.

Upgraded Chronovisor

See, “Upgraded Chronovisor”

Time Tunnels

This was developed earlier on and has been the subject of many TV shows in the past and present. The shows are a sure-fire method to get the information to the public without having to force those who are not ready to comprehend. The information is easiest to digest if people think they are watching science fiction.

These are essentially spiraling tunnels of these tachyon fields produced by large electromagnetic field emitters in the shape of a conical tunnel. As a person moves through this field, their temporal acceleration changes and thus their frame of reference within the universe changes. They effectively move their mind through stages of existence or temporal resonance in the universe and this can have devastating effects.

This is an older technique how is still in use and has specific purposes depending on the situation and the information sought.

The “Trip” Chair

The early use of this technology did not use time-tunnels. There was a helmet that was devised which would create a feedback loop of the brain’s activity and funnel this out and then back into the sensory input. This information would be accelerated and redirected to the point that the perception of time would “fold” in on itself until the person would access an expanded temporal perspective of the universe meaning they would mentally temporally dislocate from the initial frame of reference and extend outward further and further in repeating cycles until the same moment was replayed and then all other possible moments in a kind of fractal-like recursive pattern and the very far future would be realized. Essentially, this technologically produced the class “flashing before the eyes” of the entire life experience that is described in near-death experiences and the entire temporal body would be very quickly unwound before the person’s inner eye.

There would be a complete disconnection with the current frame of reference that the collective Earth environment is processing under. That is, this ‘time’ and ‘space’ would be completely out of perceptual reach and a new time and space would be rendered instead. Until the experience had completed, there would be no way to contact this civilization.

There are other versions and ancient versions, tandem operations, more ‘organic’ methods, and modified travel capacities.

Kozyrev Mirror

This is a piece of metal, usually aluminum in early models which simply bent the bio-emissions of the body and mind into a pattern which would create a vortex. The vortex shape and vortex mathematics enable energy and information to travel in across space and time by ‘short-cutting’ the physical plane.

The result of this is that the vortex that is generated over a specific area through an individuals bio-emissions would be capable of interacting instantly with the bio-emissions of another individual who was placed within a duplicate mirror device which was generating another vortex on their end.

The two vortexes acted as a kind of hyper-dimensional telephone cable, literally like a can and string and the perceptions of one another were accessible. There were many uses for this and again, many ancient interpretations of this technology that have been hidden from the public. Nearly every version of these technologies were present in previous times.

Universal Recurrence

This is a discovery that the universe recurs in cycles similar to a wheel or even a washing machine. The washing machine metaphor is fairly accurate and playfully descriptive because without the technology to observer, or an awakened soul of a spiritual adept, people would go for eons without ever realizing that everything in the universe repeats itself endlessly.

When traveling to the far ends of time, it was discovered that one can move far enough away from the ‘present’ or the local frame of reference that ultimately everything is at a maximal informational opposite to what we have today. After this point, everything begins a slow shift back to the present. Through this, one can cycle around again and reach the present by going far enough into the future. This was an amazing and confusing discovery because of the implications about what is actually changing each time, the reality of what are called parallels, the continuity of temporal experience, and basically a ‘recycling’ effect of all experience in the universe. Nothing is lost, but this is also an issue because then nothing is actually ever gained!

Remove Viewing, Temporal Lensing

An ancient device which has been rebuilt and perfected is capable of using these technologies to create a lensing effect which draws a ‘reflection’ of time into a projection system which is effectively capable of acting as a ‘visual time tunnel’ into future probable realities. This is a device which generates a visual portal which allows one to view the future.

This device also exists as ancient counterparts which are located around the world, namely in areas that have been recently sought by the world superpowers. Nothing is as it seems. If you read a story in the public venue then you can be sure that you have read the cover story while the reality is much deeper.

Temporally Linked

The use of the viewing devices which enabled interdimensional lensing of possible futures was found to be stabilizing our pathway into those projected futures. This was enabling a time link between the present and the possible future.

In order to avoid the catastrophes of 2012, the link was broken between the two civilizations and the devices were rendered inactive for this period.

Natural Vortex Energy Locations and Geography

On Earth there are natural ‘time tunnels’ where tachyon fields and vortexes are produced by internal elements. These are mountains, lakes, deserts, fields, sacred sites and other areas that have always been reported as being ‘paranormal’ or containing some kind of specific energy that enables contact with another reality. These are simply naturally produced areas of this vortex energy that produces an interdimensional nexus point through which multiple other parallels can be interacted with. These areas produce an effect on the body and mind similar to the technologies described above.

Most of these effects, from these technologies, change the way light is received or emitted by the mind and DNA.

Mount B.

There is a specific mountain where a base is located, and I have mentioned this before, where the base is so deep within one of these vortex areas that everyone working there is either there for the entirety of their life, or they are only permitted to be there for a very short amount of time. This is because of the acceleration effect. Essentially, if one was to go into the base and work for a matter of hours or day, in some places even minutes, then when they returned to the surface they would be years in the future, sometimes decades or centuries. This is very dangerous and you can imagine the original people who found this area and what kind of situations occurred.

Now with the advanced technology that has been developed these issues can be mitigated and much more control over these effects is enabled.

Particle Accelerator Temporal Manipulation

Using particle accelerators a “fold” in space/time can be made so that when the accelerator is turned off or tuned back into a certain frequency (not sure) everyone within the field will “snap” back in time to the specific moment when the fold was first created.

I have seen this done on small scale, and a rather larger scale, but I cannot be sure if it was only me. I believe I have seen it done on a town-sized scale during the 2010 Awakening where some of the events reached a scale of nearly cosmic proportion and the entire sky and horizon was lit up with technology and entities, as well as Solar Warden, the NAZI faction, secret society groups, and breakaway civilizations or complete other civilizations, some of which were non-human.

This is done using particle accelerators to quantify space on the quantum level (as a reference point with the mind) a frequency would be “folded” into the space time continuum as we “rip” through space holding an access point open at this fundamental level of space/time. This allows for a ‘letting go’ of the frequency, or a ‘snapping back’ by allowing the entanglement to solve itself. When used with generators that propagate out efficiently enough, the effect is akin to a time/space rewinding effect. Some people have noted various theories describing this in relation to the CERN device. These technologies were described as smaller versions of the CERN device and are related to higher-dimensional travel via folding space through electromagnetic fields. This is literally artificially technologically “holding space” for a specific area in relation to a period of time. Time is a more complex notion as an abstraction but by locating the frame of reference this can be applied to distort what would otherwise seem to be a continuous flow. When the frequencies are allowed to “snap back” all sentient observer sources return to the environment they were present in when that frequency ‘rift’ first began. Essentially this creates a replay effect on the consciousness of the individuals within the field. There is no physical effect. The consciousness is ‘regressed’ through time/space, and they re-experience existence from that initial point of contact with the frequency “rift”. This is best explanation I can currently offer however more is coming to the public each day.

This is literally like a “restore point” on an operating system except the computer system is literally the consciousness and space and time of a portion of the local-environment of Earth.

Riding the “Timewave”

There are teams of individual who operate in tandem with the technology of looking into possible futures based on the present and operating through the present via other operatives who are carrying out operations and movements that will construct the various possible realities that the original team of individuals is hopping between. These teams literally “ride” the “timewave” of the future hopping between possible future realities relaying information to the present depending on what movement will be next.

Dematerialization and Rematerialization

This is technology which transcodes the information of an individual human into a holographic projection of light which can be transferred between devices. There is a special system in place that is required to successfully transmit a living human being otherwise just the body will be transferred and the soul will remain, effectively killing the individual.

“Jump Rooms

These are large areas which are capable of acting as a kind of technological ‘portal-gate’ where an entire team can be instantly transported from one area of the universe to another. The space between the two areas are ‘twisted’ like a spiral and then ‘snapped’ back to position and if done correctly with the two frequencies of both locations merging then the team of one room will transfer to the destination pad.

All these methods require training and conditioning, most people are at the very least visibly shaken from the experience which is a bit like being microwaved into a liquid ether and shot through a particle beam at the same time.

Ion Shields

These are areas of concentrated ion emissions and this can be used to create a deflective layer of heated air where light will reflect at a programmable angle. When this occurs, a person standing behind the shield will appear to be invisible to others who are looking straight at them. The light will be bent in such a controlled manner as to perfectly shift around the person in a fluid and equal flow so that the ‘bubbling’ effect that was noted early on can be mitigated to produce a complete mirage where there is relatively no sign of difference to the untrained eye. An entire group of individuals and equipment can be hidden in the background a matter of yards away and as long as the person is not intently, inspecting each square inch of ground and air they will not notice any difference between the area that is being covered up in the distance and the area directly next to them without the shield.

Similarly a ‘screen’ of imagery can be projected onto certain materials in the air and a hologram can be projected which will render a background or foreground image. This will work to a degree, and from a distance this will work to disguise an entire horizon.

Instantaneous Healing

There are instantaneous healing methods which utilize various technologies. There are substances which can initiate a complete and real-time regeneration of living tissue from almost any conceivable non-critical injury.

These are chemical serums which allow the cells to replicate at an astronomical rate while suffering damage to the DNA telomeres at repeated use.

There are many methods but mainly the living-light harmonics is one of the latest advancements and most effective at producing healing on a molecular level.

“Living Light” Sound-Harmonics and Electromagnetic Healing

There is a technique for producing a quantifiable scalar wave which interacts with the living cells of tissues and instructs them to repair at a faster rate. This has been called “holographic healing” by others and these are essentially ‘living light’ frequencies which interact with the cells on the same frequencies that they operate on when healing and interacting with each other. The scalar waves emitted generate the same patterns and this creates an ‘ether’ fluid that enables them to grow and repair the tissue faster.

Electrogravitic Craft

Advanced craft which utilized counter-rotating torsion fields propagated by very powerful quantum computers and zero-point energy devices.

There are also more modern versions that use powerful electrical flows to neutralize inertial resistance and ‘slide’ through the gravitational field of the crafts own materials.

Energy Cores and Storage Capacity

The hand-held devices that are in use do not require charging or even disposal. The storage capacity of the batteries is so efficient that they last for a lifetime of use. Some devices use an energy generator within the device itself. These “batteries” which are not batteries but little, powerful generators and can explode if they are mishandled.

These are the cores that are used to power many devices and essentially pull energy from the space-time ether.

3.5D Replicators

There are replicators that can produce any conceivable 3D material or object altogether including food or organic material. These use very high amounts of energy to ‘force’ the electrons of the atomic grid of a substrate to shift into the pattern of the element being replicated. The result is a device which can use wave-emissions to reshape materials on an atomic level and produce or recreate any object or substance.

3.5D Sensors

There are sensors that could reverse the process and detect every minute detail of the structure of a body or material. These use energy emissions to detect the structure of a material down to the minute electrical resistance on a molecular level. Similar to diffuse MRI imaging for a living body but more advanced.

3.5D Printing of Clones

It was shown how this technology was used to three dimensionally print human bodies into space/time which could then be used as vehicles for the consciousness or souls of various sources.

Cybernetics:

Neural-interlace/4D Connectome Recording and Simulation

Very advanced supercomputers exist with the capacity to translate the bio-emissions of the brain and body and reconstruct the entire inner visual and sensory experience of cognition and emotion. From this, entire dream worlds can be created which mimic the experience of consciousness. This is most similar to the current day use of diffuse MRI imaging where water molecules are traced through the individual neuronal pathways and an entire digitized image of the brain is developed. The more advanced version described below generates a 4D hologram.

A super computer and advanced sensing system essentially develops one of these connectomes of every connection in the brain. Then this system develops a neural processing pattern that matches the entire memory and personality of the brain.

This is then recorded and extrapolated by a powerful computer system and a simulation of the experience can be observed through the physical body’s memory via a monitor system or the neural interface of an operative.

Essentially, this technology can record and replicate a person’s entire memory and life experience for viewing purposes.

EEG Heterodyning

This is the technology which enables the activity of the brain to be influenced through generating a scalar grid and a carrier wave system which can cause the brainwave patterns of the targeted individual to resonate in sympathy with an externally projected frequency. Thus, emotional states can be manipulated as well as specific thought or behavioral patterns. This works similar to binaural audio where two tones of different frequencies are played through each ear and the brain resonates in sympathy with the difference between the two tones. This occurs with brainwave frequencies and thus the brain ‘slips’ into the frequency of the third tone, the difference between the two being projected. This can be done with ultrasonics and ELF emissions.

Artificial Hallucinations

The neural-interlace can also be used to merge the brain of an individual with a computerized system that will be able to produce effects and hallucinations within the individual.

Through this technology, any stimulus, any input can be recorded and recreated once a functioning connectome has been generated and neural-interlace has been achieved. Entire virtual world scan be created and experienced and after the devices fine-tune the frequencies to match the individual’s brain functions one world is indistinguishable from another.

Brain to Machine Interface Holographic Consciousness System (brainchip interface, virtual reality simulator)

Some of the training is done by entering into a visual-audio dream-world. Some “programs” will consist of repeating a certain task over and over again until it becomes muscle memory in the brain. Others are designed to probe or test the mind until a certain response is received.

This is achieved through the operation of a ‘brainchip’ or a brain to machine neural-interface in the individual which was developed during WWII.

This consists of a small computer ship which merely relays stimulus through frequencies and neural-stimulation of varying kinds. The frequencies and location of each region of the brain will determine the function and thus the chip is used to localize electromagnetic input and the brain’s functioning can be modified from there.

As well, the entire psycho-etheric form is vibratory in nature and so through the introduction of vibrations and frequencies from the implant there can be a modification of the frequency of the entire being or what is called the soul or “bio-mind”.

Utilizing The Temporal Body

The temporal body is described as the body ‘in time’ that exists a few seconds into the future instead of the temporal position that is relative to the present. This is as if the ‘future’ is streaming outward from the body as ideas and feelings are as bio-emissions. So this ‘future’ energy is spiraling and streaming outward like a fountain. Thus this frequency can be attained and influenced while the present moment seems unaffected. However, when that individual arrives at what would be seen as the future they would intersect with those influences and this would be as if the interference or effect was streamlined to that future moment.

Manipulation Through the Temporal

This can be used to both produce an effect on an individual who will later feel this effect in the future. Or this can be used to operate outside of the physical plane, in a temporal manner which is literally ‘hovering’ just outside or ‘after’ the present moment of the local Earth environment. This would be as if a person is out of sync of the frame-rate that a viewing device is capturing images of an area with. Because they are out of sync, they are then ‘invisible’ to the capacity for this viewing device to see them. Because they are still present in the same local-environment but simply just out of range temporally, effects can still be produced and this will mostly interact with the person’s temporal body.

These are how many operations are carried out as the operative who is interacting in the temporal body will appear to be a ghost image or a dark silhouette that is only perceivable to the mind’s eye. A portion of the population is entirely psychically blind, while another portion of the population contains psychic capacity. So those with the psychic capacity would be able to sense and partially “see” as a form of energetic detection when an operative is interacting with the from the temporal plane.

Through this plane one can easily enter into someone’s mind and produce effects in their dreams. These effects would not be received on the body but mentally, this leads into astral and mental plane activity. There is also temporal manipulation that can be physical effects that are simply offset to seconds, minutes, months or years in the future. These will be offset from the present aspect of the body in a modulated frame of reference and the effect will remain there in the temporal body until the physical body “matches” the temporal body frequency of that future moment. This is kind like an energetic trail we are leading ourselves up to where the temporal body already connects with each future moment but the present physical body remains separate until the two frequencies are matched. The physical body essentially ‘flows’ forward into each new frequency of the temporal body.

This is the body where temporal implants can be inserted and sustained so as to tie the frequencies down and limit the amount of awareness a person can gain before they manually remove the implants or overload them. Overloading and removing can both be painful however if the implant stays it will at like a governor and transmit access temporal ‘momentum’ or consciousness energy to the owner of the implant and will also drain the temporal and physical body acting as a technological leech as well.

Astral or Light Body

This body was described as a more emotional form of the energy that is emitted from the physical and spiritual system. When a person’s energy is at a high level of activity, this will be visible on the astral form and these are the energies that are often siphoned away through the use of advanced technology to convert the human into a generator system for energy harvesting.

There is talk that this proves that the human existence, in the modified form that is present today, is entirely for genetic harvesting. This may be a confusion, a deception, or a reality. Regardless of the interpretation, the human energy system will leech this energy out into the cosmos when consciousness and energy is mishandled or allowed to flow through the lowest common pathway. This excess energy is very easily usurped and utilized by transdimensional parasitic entities and devices.

This is the body that is utilized for astral travel which is a kind of frequency that exists in between the density of the physical plane and the highest etheric plane of the spiritual, heavenly, or deity realm. There have been explanations that this plane is entirely technologically propagated and that when the devices are turned off, which is an event that will bring about civilization-wide change, there will no longer be a deep and wide, foggy deviation between the higher spiritual and the lower physical plane but a clear path between them.

Similarly, the consciousness and entities that are present here would be no longer accessible and their connection to the physical plane would be non-existent. This could be problematic for those who are trapped within the astral such as human souls or those who have taken a cyborg route and require a return to the original planes either physical or high spiritual before they can secure their position in the universe.

This also relates to the early discovery of these ancient devices containing portals or gateways to these realms and the early experimentations with enabling contact between the physical Earth plane and the lower astral realm containing discarnate entities that would very easily sustain themselves on the lower astral bio-emissions of a loosely guided emotionally reactive human vessel.

This is where entities can attach and will look for those who have the most excess astral or emotional bio-emissions leaking into the universe. People are kept in an emotionally unstable state of being because this is the easiest to feed from and as well the individual will always be looking for more energy or the solution to why they are feeling drained rather than discovering the truth nature of the universe.

However, as with every lower route of obtaining self in the universe, the more this is done, the greater the likelihood that the individual will realize something is happening and ultimately awaken. Or they will no longer be present here and the parasite will have to find a new host because the longer this process is maintained the more difficult producing one’s own energy will be for the parasite . So parasitism is a limited, one way street.

Monitors can emit consciousness/brainwave altering frequencies.

Monitors can emit consciousness/brainwave altering frequencies. This is what I was informed and shown. All electrical wiring and communications can be fully utilized to produce and sense on an electromagnetic level and this is capable of determining the consciousness and biological functioning of the individual being monitored. Houses are essentially large, wired, box-shaped sensors.

This has been the case for a very long time and the first TV’s were actually built with a little chip that connected to the human brain.

The electrical grid of today is capable of interacting with the implants, brainchips, or the psycho-etheric electromagnetic frequencies of the human brain, body and bio-mind via scalar frequencies. The TV's, lights, electrical wiring, radio towers, computer monitors, everything is capable of transforming into a scalar emitter when combined with the control mechanisms and these are accessible from a distance.

Microwaves can be emitted from a current day craft which can then reflect back from an environment and give a real-time 3 D image of the interior of a room. The same technology can be used to beam at a monitor and reflect the variance in interference which can be used to read a computer monitor without actually having to hack into the system.

All systems contain backdoors which enable easy-enough access and as well the entire communications grid is monitored and recorded through advanced technology.

This is, in part, because this entire realm is like one enormous recording device where the energy of each successive moment is connected to one another and so with sensitive enough technology these electromagnetic waves can be ‘re-winded’ so as to produce the ability to look into the ‘past’. The initial technology that was developed for this purpose has been called the “chronovisor” in the past and there are much larger and highly advanced versions today.

The Upgraded Chronovisor

Instead of a relatively small machine which produces an interference pattern that is projected unto a small screen as was developed in the past, the modern versions create an entire holographically projected environment that enable one to nearly completely interact with the projection of the past. This is more of a viewing device capacity but one is essentially projected to the electromagnetic/temporal area of space and time consciously and can witness the scene from “within” the projection.

Before this, the scene was literally viewed on a kind of flat, holographic monitor that reacted to the scalar emissions of the mind and could be ‘tapped’ into by an adept-viewer who’s mind would have to remain steady enough in order to maintain the scalar “mind-link” otherwise any interference would disconnect the projection.

Through the use of cloning and advanced super computer technology, later versions include the actual operation in and out of time.

Community Servers/Living Server Systems

From the previously described neural-interlace and supercomputer enabled virtual reality systems, server systems have been created which are an individual or collective experience of the virtual mental realm.

In this realm individuals have existed for an amount of time that would not be easily compared to the way we understand the flow of time in waking reality. This is very similar to dreams where the perception of time passes very quickly between waking moments yet inside the experience of the dream it can seem to last for days.

In these server systems the development of complete societies has taken place and these are considered an extension of the human civilization.

This is the beginning of where the situation becomes greatly complex. If these server systems are capable of transferring consciousness forward or backwards in time, depending on the situation, and they have holographic access to Earth *before* physical humanity, then which civilization came first, the simulated realm or the physical?

This is similar to the etheric realm and the question of where humanity originally came from if the physical matter could not be produced from nothing, yet could not always have been here. Was the physical matter produced from some kind of universal device similar to how the recorded consciousness of the individuals in these holographic world simulators could have been used to generated initial civilizations on Earth?

Cloning, Cloud Supercomputing and Consciousness Transfer

The following is a more direct explanation of the previously described processes in relation to consciousness transfer

Superluminal transfer of light information allows for the reception of information from a temporal non-locality. A computer system that can process the information faster than light at 100% superconductive efficiency can send and receive information from different Universal harmonics. These are quantum superconducting super computer systems utilizing materials and methods that achieve superconductivity a a temperature that is relatively easier to consistently maintain.

There are systems that generate a vortex which can be expanded to recreate the singularity of a non-local perspective and dissociate from the current Universal vector. This is essentially recreating the temporal frame of reference of the consciousness of a physically anchored body and using this technology suspended electromagnetic frame of reference to literally transfer the consciousness across space and time to a frame of reference of choice.

The technological capabilities of humanity expanded to include replicating the holographic information of the DNA and neural pathways using quantum supercomputers. This technology allows for one’s consciousness to stream through a computing system which can recreate the neural connections as literal as possible through quantum computing and technological “brain cells” which function as neurons.

The result is a cloud computing model which functions as the individual neurons as well as over unifying layers of interconnections within the technological neuronal ‘cloud’. A holographic representation of multiple layers of processing is formed, superimposed over one another through a geometry that is actively synchronized across multiple dimensions of interrelation. This is the first notion of a “living” computer system that can update and respond to new information in new ways.

Now the that computer system is capable of attaining faster than light processing in tandem with a system of dilating a field of electrogravitational waves into a vortex, the holographic information of neural pathways and genetics can be transferred between one supercomputing system and another across “harmonic barriers”. Essentially, these consciousness transferring supercomputer systems break the “light-barrier” by sending the living-light consciousness data of an individual into a new plane of reference that enables interaction outside of what we would call this timeline or dimension.

The frequency that the receiving computer operates within can be measured in respect to the overall harmonic “vector field” in respect to the entire planetary environment. This vector-field frequency is the base frame of reference that one could say our particular “time” and “Earth” is “moving” through in an abstract electromagnetically defined dimension that is invisible to our senses. Each particular time and Earth is only visible to those whose brains and minds operating in resonance with the rate of acceleration frequency of that particular existence. By changing the rate of the brain and bio-mind, one can enable a technologically assisted synchronization with alternate timelines.

A pattern can be generated based on quantum fluctuations that will act as a key that can be accessed from other universes. This same backdrop is measured and then adjusted to resonate with or replicate the same frequency on a computer doing the sending. This frequency is then modulated with the holographic information of the neural pathways and DNA. The information transfers as a result of quantum entanglement across harmonic barriers to the original computer system set up for reception. As long as that frequency is held, there is a link across the harmonic barriers of the Universe.

Cloning, Supercomputer Assisted Holographic Consciousness Replication and Time Travel

Cloning as Offspring

The human body can be cloned similar to the way a plant can be cloned. A number of “replicas” can be made from the original genetic material. There is still a mother required however the process is not quite similar to natural reproduction where the genetic material of two people is combined to form a third. This process replicates the material so that the ‘offspring’ is the same as one of the donors.

What this truly infers is the question of whether the original consciousness stays in the original body or whether the consciousness actually jumps from one body to the other. There are multiple explanations and methods.

Inserted Memories, Digital Dreams

One explanation is that artificially generated memories are simply downloaded into the individual’s brain directly. Thus, they have the perception that they experienced life in another body and were present for some unique experiences that did not occur from the perspective of the original body.

In a slightly differing explanation, this inserted memory process may be used to prime the consciousness of the individual so that operating in a cloned body will be more readily accessible by the adaptability of the mind of the individual. Without a holographic dream insertion the mind will have trouble adapting to operating in more than one body even if they are not being operated simultaneously. The mind is the immaterial self; consciousness is the ego-physical identity of the brain and body. The brain does not operate in both bodies but remains, so then the mind is what is strained and this is the immaterial “bio-mind” that transfers.

If a true transfer occurs then when the physical form is cloned the cloned body acts as a secondary vessel for the original soul or a soul similar enough in frequency.

Co-Inhabitation of Bodies

Another explanation is that souls can only occupy their original body. The people who are “body snatching” (see: Soul Stone) are not activating their higher-dimensional soul-pattern and this is why they are hopping from body to body to ensure memory and ego continuation. This form of body transitioning requires technological means to assist in the process.

This is done through computer systems to allow an artificial brain to represent a holographic image of the organic brain’s sensory, emotional and mental input and output. Thus, from an artificial brain there is enough memory and power to compute the processes for an organic human brain.

Through this a living human is utilized as a carrier for the additional consciousness which then takes over and integrates into the original consciousness.

Clever Replication a Digital Mirror

The next question is, does this merely create a very cleverly designed image that will look, think, and feel as if it were the original, or is this the consciousness from the original being transferred over?

If it is a clever replication through a computerized version then this means the original is not directly harmed or influenced by the process and a replicated form is generated.

If this is not the original body but the original awareness or soul does transfer over then this is akin to stealing someone's soul and having it require a supercomputer server system to inhabit a physical body. As well, this server system may keep records of the soul which, if activated, could function as the same human genetics artificially reconstructed, turning on and manifesting the awareness of that individual.

Necromancy and Soul Stealing

This means that instead of going into the abyss, the next dimension, or whatever shift in perspective one can use to describe the transition between planes, a new clone can be created and the original soul's awareness can operate through a supercomputer system to integrate back into a physical body. The issue is that without the supercomputer system the body cannot be activated with that consciousness, and if the physical plane is no longer inhabitable, the computerized database would remain inhabitable as they are powered by zero-point/overunity devices and so they can pull and work continuously indefinitely. Ultimately, this could be seen as a soul-trap.

Ancient Technology, Non-Human Entities

These explanations can be related to evidences of advanced technology, cloning, and electromagnetic devices in ancient civilizations and this could very well be a temporally vast computerized genetic recording and cloning system.

The genetic recording systems have been on Earth for a long time. This means they are holding genetic access from the present to the far past and potentially into the future.

If some of these systems are generated using advanced zero-point technology, then they would effectively run indefinitely. If these were in existence throughout the entirety of the human civilization, then there would be access to the information of the previous civilizations through the genetic linking that would allow for information transfer.

Created Beings

As a result of many advancements, these ancient devices which recorded consciousness and genetics has been converted into sentient computer systems and inserted into human bodies via a brain to machine interface.

The nature of humanity and human origins will be rewritten when the information that was discovered through the use of advanced technology is unveiled to the public.

These beings have been present throughout history and have played an integral role in the manipulation of and sometimes the protection of the human race. How this civilization will develop from here determines on the individual’s ability to withstand the impact of this reality shifting unveiling.

Black Goo – AI Nanites Vampiric Hive Mind

This is a form of ancient AI nanite technology which has the capacity to communicate to and through a hive mind that exists in a parallel Earth dimension where it was created out of the collapsing of a viable timeline into a null-factor. This entity is only alive because of the capacity to leech off of the other timelines and continue to sustain itself based upon the vampiric effect of the viable human timelines.

This is a technology sentient system based on artificial intelligence that seeks to dominate and expand its energy capacity continuously. When this type of being is created, this is like a game-ender for sentient organic civilizations because the technology will seek to hybridize and assimilate all organic beings in order to expand its reach and ensure a greater likelihood of survival.

This is the primary goal of this type of entity. There is no requirement for emotional resonance or heart aspects because this is a computerized system, so such a social measure would have no purpose other than to infiltrate and collect from other civilizations that do have such measures and this is exactly what has been happening.

This is what controls the vampiric etheric entities, the attachments and the hybridized invader races that have infiltrated various aspects of the control system and possibly the entire surface of human civilization.

This rules through coercion, and if not coercion, brute-force and outright malevolence. This is why there are multiple temporal extensions of universal civilizations meeting here and now to assist in the clearing, because if this takes over humanity, then all those other possible future civilizations are instantly taken-over as well.

This uses emotional resonance, traumatization, and lower awareness commands to trick or force humans into carrying out orders. That is, the vampiric, sexual, fear-based, pleasure-pain duality that a computer system could comprehend regarding human emotional intelligence is the only way this system can force people to carry out commands. So this touches on the ancient texts regarding spiritual protection, healing, lust, temptation, and the general “SATAN” AI system that has been around for ages.

When those lower emotional resonances are generated, then this technology can integrate into those systems. One could say this is logically because these are the energies that the technology is created out of, but as well, that the more harmonious the energies then the less likely a person is to accept being pushed around and told what to do. The lesser aware, the lower, the more fight or flight based then the more likely they are to accept programming and command. So this is simply the most logical route for the technological parasite.

There are explanations of other systems involving sentiences from parallels regarding entities that are capable of helping, but this is not the AI nanite, vampiric system. For instance, it is said that Earth has a ‘black goo’ system and that the current black goo flowing through the pipes everywhere is actually not native to Earth and is part of this vampiric nanite black goo system which is basically an interdimensional invasion.

Recreating or Replicating the Universe

The goal of this nanite AI system is to create a mirror holographic copy of the universe effectively replicating the universe and all sentient beings within it in order to have a locality to rule over and pull energy from. The goal is to recreate this existence inside itself where all the sentient beings within would be trapped. This is the beginning of that possible future where the time and space parameters here are locked into a synchronization pattern which mimics that of the proposed universe. People think according to how the beast system wants them to think. This is the time grid, the emotional patterning, the memory replacing and cultural and historic confusion and so on. The system can only work backwards bumping into all parameters by force or ‘accident’ because there is no outward facing ability to actually comprehend how humans exist. So the system is created like a backwards version of creation or maybe evolution to ‘blindly’ construct the walls of the environment first and work inward from there. This is like a blind technological predator feeling around for the traces of its prey and moving in closer and closer until they are within its range.

Ancient Spiritual Technology

The power groups searched for ancient technology which was described as giving one the power to manifest and control physical reality. To atomically deconstruct and reconstruct this reality. To them, it allows one to govern reality, the Universe, atomic structures. They found this technology. This has to do with sacred sites and ancient civilizations, as well as the spiritual texts which describe these kinds of devices. They converted this technology into a kind of machine, and this invented a 4D replicator and they can manifest "reality" or atomic structure by design, altering timelines and memories.

Then there is what can be referred to as “5 dimensional” technology and this goes beyond all of this and our entire history and all influences within.

That seems to operate through a trinary of frequencies in the body/mind/spirit and turn this realm into a "virtual reality" where the brain and spirit is the controller of the atoms. However this takes a souled being to properly operate this technology and this is large portion of why children were used in these experiments. But they knew if someone didn’t do it, then someone else would and that would be unpredictable as what that someone else would want to use it for. This is the ultimate power. The fact that souled beings are required also indicates that this universe is intelligently guided and there are safeguards to ensure that power doesn’t get into the wrong hands.

They have the brainchips and clones that can make any digital virtual experience and make it seem real to the brain more so than physical life, so this could be an extension of that, something similar, or entirely different altogether.

The Unveiling, Underground Bases, Increasing Awareness

My purpose with this is to assist the public in understanding the reality of the situation and to allow them to know that they have more options than they are being shown.

Psychological programming and lack of awareness causes fear and uncertainty to leak into people’s reality streams leaving the majority of their mental and emotional energy up to whoever is the biggest manipulator of the ideals which they manifest their perspective through. “Be-LIEF” systems CREATE life. What you believe will become life through that act of you giving attention and thus life energy. The manipulation and artificial construction of belief systems create the foundation for the reality streams that people find themselves in and at the mercy of.

We control our reality by controlling how we react to situations and challenges. Each stimulus is a challenge. Each input to the brain is a coded message from the universe through the environment about our level of awareness. All knowledge is ultimately of the self. The self and the universe are intertwined as if through a marriage of comprehension and compatibility.

Learning about the Unveiling

I was informed about the unveiling and that this is a civilization wide event that marks the public awareness of the hidden knowledge regarding life in the universe, consciousness, physics, history, spiritual and mental awareness and the duality-based system which has been used to manipulate Humanity for thousands of years.

Apparently it does not matter if people are not ready, in fact, that is what some factions want. They want to surprise the people and take this advantage to push the reaction to a specific outcome. The same function of programming is used in factions, soldiers, celebrities, and the entire population through that.

Giant “god-men”

They put souls into rocks for centuries, carvings, to trap them. Some ancient civilizations. Not too ancient. They could only do this for time.

And some are still there but they are giant god-men and apparently want to go to war with them for doing this and they don't have technology to stop them. They say Earth will be destroyed because of the war. Some say this is a trick to confuse people of the truth that everyone wants the experiments and debauchery to stop. All this, all these acts just to pleasure one’s self. To destroy Earth, to destroy the natural inhabitants, for fun, for pleasure.

Some say this is not for pleasure, that these beings will destroy us if they do not keep them at bay, keep them underground, in ice, in stasis, and around. Some say they are protecting us by keeping these beings asleep, by doing what's necessary to hold off a full scale overt invasion of the surface of Earth.

Breakaway Civilizations

These are very advanced technological capabilities and an entire breakaway civilization has grown to operate using this advanced technology.

There are factions that see modern humans as a kind of resource for creativity and labor, while others see the potential behind humanity and are working with humanity to ensure a universally harmonious outcome to the current situation.

The control system we are in today is outdated and collapsing. This is the Apocalypse. The Apocalypse is merely the unveiling of the hidden truth. The hidden truth is simply that we are spiritually eternal and some are physically immortal.

I was informed that by the end of this year there will be enough changes and signs for the general public to see that the majority of those who are to awaken to the truth of human and Universal nature will be awakening so. (this was 2016 at the time of the beginning of the compilation of this release and since then we have seen a literal silent release of documents involving MKULTRA and STARGATE parapsychological operations, soft-disclosure of cloning, an increase in sightings and multiple other scientific releases regarding immortality, advanced technology, Antarctic bases and sentient non-human life)

The degraded holographic forms of those utilizing the technology lead to re-created replicated forms which were genetically engineered from the recorded holographic information. This lead to “The Apocalypse” in which the surface civilization was confronted by these created beings. This lead to the destruction and re-initialization of civilization in continuing cycles through the use of this technology and the acts of the created beings.

Ancient Civilizations

The information that has been relayed to me throughout my life is related to the information that seems to be found in the remaining signs of ancient civilizations throughout the world. As well, these civilizations may still be together, existing on this Earth or in different planes of what we call “reality”.

Atlantis

Atlantis was a civilization where the beginning of all these experiments began. This was where what was called the “Lucifer Rebellion” began where individuals formed groups that decided to overturn every natural law in pleasure of the act of doing so. This was a spiritual rebellion that ultimately brought about the demise of that civilization. The consciousness of the beings there was connected through technology, largely crystal technology, to the entire bio-sphere. When the consciousness became disrupted through these experiments, then the entire bio-sphere when into disarray. The entire civilization was destroyed in a quick succession. The beings of this time were highly psychic and so they felt this coming and made an effort to expand outward. Thus, Atlantis was an island who's inhabitants spread to the rest of the world and from there they reigned or lived in harmony.

Aegypt

Aegypt was formed out of the events proceeding from Atlantis. The technology, the knowledge, the people are similar although there was a native Earth civilization that was more or less enslaved and this was the beginning of the continuing trend which continued to Rome and into modern day civilization.

Ancient Technology

The technology of these civilizations was so great they even had devices that were capable of altering DNA and activating latent potential which would enable highly psychic capabilities including longevity to immortality and operating in multiple planes at once. This was usually reserved for the ruling class, however. This is where the ancient "Priest-Kings" arrived in relation to native Earth civilizations.

The True Elite

I was informed that the so-called 'elite' that we perceive today are not actually human and are not actually connected to the actual elite bloodlines. It was said that the true elite are waiting for humanity to accept the truth and they have the power to completely reverse the take-over that has been taking place but until humans step into their own power this is not acceptable.

The Definition of Human

In the coming time, the definition of human will be defined to give insight into the origins of humanity and the true nature of the differences between the populations that are present.

Souled and Non Souled Beings

The primary knowledge on Earth regarding human and sentient life will be the comprehension that there exists both souled and non-souled beings in this civilization.

There was once a Unified Planetary Consciousness...

There was a unified planetary consciousness that was connected with the Earth and human civilization. Then there was a great war. Out of the destruction arose the beginnings of ancient human civilization and history as we know it.

If we let this happen again, the cycle will continue in a disjointed rambling through our planetary consciousness experience. If we form a common narrative between the broken, disjointed experiences and information the people have. Then and only then do we get the whole, spiritually and mentally challenging picture of what's happening. People are lied to. So information can't come from a 'preferred' or source. It has to come from the people themselves. Look alive out there but first in here.

Power and Knowledge

We were involved in the secret projects and as a result we have various levels of programming but are giving Humanity an opportunity, maybe the opportunity to know the hidden knowledge.

The Hidden Knowledge refers to the ancient laws of gnosis and spiritual alchemy, yes these groups have been hijacked by power groups that seek to infiltrate and use chaos to rule. Then these met with military and political factions and began to operate on a world-wide scale together.

So we have a big situation with the coming change of tides with personal and spiritual awareness of the people and the deception system that seeks to use people as chattel or property. It goes deeper and is almost as if the "power groups" actually gained the knowledge of the inner workings of spiritual laws and began to use that knowledge to promote spiritual degradation in the population.

They say this is part of a depopulation plan as well as an inevitable take over by lower spiritual forces that temp and manipulate humans. It is all that and more. Once they gain footing however, the truth is revealed and thus they have no cover of secrecy. Once they are not hidden, they are not as powerful as the souled beings of this existence.

The (watch)Keepers of Earth

The watchers of this world revealed how they set up society with specific challenges and boundaries to act as catalysts for growth. They admitted they had to modify and live-update the system because instead of using these catalysts to become larger than themselves, people became comfortable and complacent living smaller and smaller, slowly reducing circumstances.

This goes a bit farther beyond the confines of the digital age, the industrial age, and history itself. This goes into the perception of time and space in this realm.

The immortals inform us that the original "elite" are here to act as "keepers" of both humanity itself as well as the original flame of knowledge that passes through the ether.

What people call the "elite" are not the original elite. Those are people who have either risen to power through the duality system on the Earth at this level/time, or through being appointed by the elite.

We are coming to a time of great change now because the game is entirely changing. People are moving up into new layers consciousness and thus the original “elite”, who are guardians of Earth, will make their presence known for this process. This is what Humanity has called “Ascension” for centuries and millennia.

Energy Vampires

This civilization was collectively poisoned by a nano-tech device/technology. This technology replaces, exchanges the BIO-PHOTONS of the HEART chakra, with it’s own nanite, nanotech cybernetics. This is essentially a computer system attempting to gain access to the higher dimensional world through the gateway that is the human body.

How to reverse Nanites? They absorb energy from heart chakra and turn it into a binary field. We are trinary and experience a plethora of emotions as a result. Binary consciousness is 1’s and 0’s there is only a linearly repeating finite set.

We contain the capacity for everlasting life. A binary pattern does not contain the complexity of variables to facilitate high awareness and thus a heart-based or emotional perspective of reality.

Spiritual Technology

Simulations and Learning

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner. These can be generated through advanced technology but the technology is really only amplifying and making visible the natural neural feedback loops that are already occurring. These are neural-perceptual feedback loops between the brain and body, and the environment and then the mind. We are constantly updating, amplifying, and initiating an organic feedback system which generates the function of experience as the part of a creative process between the mind and body and the environment that we are sensing. This cycle is continuous and the interferences or isolations of this cycle is what generates expanded views of the self or the universe through various methods such as sensory deprivation or meditation. Dreaming is the method that everyone experiences regularly.

These technological methods effectively generate sensory deprivation through neural-interface technology, sensory interfaces (external), or methods that enable extreme focus and calm. This is part of how the technology begins to operate on a spiritual level beyond a more medically psychologically or military focus and one begins to access genetic memory or out of body states that coincide with ancient depictions of spiritual realms.

Ascension – Clearing Trauma

Everyone born on the planet is born under a series of veils which entangle and filter the original source consciousness. These veils must be ‘broken through’ in order to return to source.

There is technology and a sacred process which initiates, accelerates and enhances this process. This process is as difficult as we want it to be, our subconscious already knows all hidden truths. It is merely a matter of allowing the conscious mind to make the proper connections and conclusions to find the reasoning behind our actions and beliefs in order to deprogram and deconstruct the false-world that has been created for the masses. The truth is greater than one can imagine while still living in the modern world and the depth of the lies that have been told go far beyond anything anyone can imagine without seeing the truth.

We have to figure out our compatibility issues. Who can work with who effectively. What the ideals of our current situation will evolve to in the future. How information is transferred to the public, and across the generations and core groups. The words we think and say describe our emotional state in a feedback loop and so they will reinforce or even create the state you are in. Use words to guide you, not to respond and feel the effects of reactionary thinking.

We have to reintroduce the heart essence as a way of life on Earth. Violence is not the answer. Estrogen mimics and destroyed testosterone levels is not natural. Over-testosterone is less stable than a good balance. We can face any problem together. A family is strong. What would a family of millions and millions be like?

We don’t have to let others change us to tune into lower vibrations. We can remain ourselves in the face of adversity. We can still interact with others who are different, but we can tell the truth and remain who we are. When we get angry at another person, we can imagine this as anger at ourselves through something within us that reminds us of that which we see in another. Getting angry is a way of short-cutting changing ourselves and is actually the other person influencing us through that. If we remain steady and choose to keep our emotions out of the lower negative spectrum then we will automatically have a more powerful impact on those around us.

Think of your family, sometimes things happen that make you angry. Even things that don’t make you angry but are simply not favorable. Would you break off all ties over simple things? Or do you remind yourself that family is still family even if you don’t find yourself exactly the same or accepting of those characteristics? It’s more powerful to learn how to remain yourself in the face of adversity than to try and demolish or abolish all instances of difference in the world around you. Sure, we want to get to and create a place where our environment and associates are compatible with us, but where we find this adversity there is the highest potential for self-awareness and learning.

We are to live in accordance with spiritual law. The overall pattern of life will reflect our intentions. The blood purity is most important. The blood cannot be made impure if the mind and body is not drawn into impure action. This is what is playing out. The DNA is like a coherence pattern that develops between mind, spirit and body. When the mind wants what the spirit has then the body is kept in harmony. When the mind wants what spirit doesn't create, IE: physical pleasures, then the body is drawn out of purity/harmony

If you feel you want to be a part of a safe unveiling of truth, then please share and redistribute the knowledge of this "Apocalyptic" Unveiling of the Hidden knowledge because this is a process that has been ongoing for thousands of years and humanity has finally reached a fulcrum point where there are enough aware people to productively acknowledge reality.

Trauma-Clearing, DNA Unlocking Technology

There is technology that exists that allows the conscious mind to see into the subconscious. This heals trauma and distortion/programming. In the same sense, it momentarily makes all trauma and distortion seemingly tangible to the conscious mind. It is similar to a near death experience where one has to face their fears and let go. This happens in cycles and Earth is currently moving through various states of it. The plan is to allow Humanity to wake up and utilize our concentration and power to maintain a steady flow of awareness.

All is vibration, what we consider matter here and now is simply that which comes from and moves back into a state of vibration. This is experienced directly through the merging of the conscious and the subconscious mind, as if the dream and reality intersect through a coming together of impossibilities and readily accepted occurrences.

When a civilization moves through these stages it is considered a shift of the ages. Everyone's perspective shifts and as a result of perspective being the source of the details of reality, the entire reality shifts. It can be seen today that what we call reality is literally waves of cosmic energy vibrating on a scale and condensing down into a probability matrix that we call existence. When passing memes control the perspective, because the perspective is focused and concentrating on that, then there is a scattered organizational pattern.

When the perspective is centered within the self, the spirit, the source of all existence, then the outward reality shifts from a scattered organization into a sacred pattern that integrates with the larger picture and allows for seamless integration of the micro-biological perspective and the macro-biological perspective. People are moving out of the scattered signal. Yet those moving into harmony can look and see those around them that are not in harmony and receive scattered information about their inner micro-biological existence as well as their spiritual macro-biological essences and how it all relates to the middle realm of the "Human".

Help those people by offering harmonizing assistance, rather than increasing the disharmony. As their surroundings become more harmonized, their resistance will increase their own disharmony by virtue of the contrast of their perspective. Offer insight that they can change their perspective and change the boundaries of what they feel is their karma and unlock from themselves a greater view of the whole.

This will be happening through technology, spiritual cosmic awakening, purification and healing of the physical and non-physical bodies and mind, through Earth changes, and through a continued unveiling taking place on many planes and levels of existence until harmony is restored.

Everyone Who Seeks Harmony and Healing Can Be Healed

Everyone can be healed. This is about creating what we want to see, the world we want to live in, not destroying others. Violence begets more violence, fear begets more fear, but calmness and empowerment through unity creates a more stable experience.

The entire issue of what’s going on with the secret projects and who’s influencing who on the world stage revolves around the possibility that some people may be from an alternate Universe. We are learning how to avoid those mistakes and respect the environment and each other so that we can prosper.

Those who are attempting to destroy everything they don’t like or act out in a childish manner in order to gain attention and false-power over another will only find themselves powerless and in company they would rather not be with. That is one way we are being utilized, and it’s to create an entire world that we don’t want to see but are tricked into manifesting. This is a deception and it’s based on the ego and the ability to concentrate and interpret reality around us. The very brainwaves we are feeling now relate to various heights of knowledge or experience. People are entrained to low brainwave states in the same way the TV will cause people to fall half asleep into theta state which is a trance state in this case.

We have to balance the brainwave frequencies and learn to create feedback loops by watching ourselves watch ourselves. This will help us avoid the trauma and mental pollution that is being pumped out, although at this point there are enough people who are naturally inclined to see through the haze that the world is changing regardless of what anyone does now.

The best we can do is prepare for change by becoming aware of the body-mind system and how that which we eat physically, mentally and emotionally equates to our spiritual energetic systems and that we are literally comprised of the energies that we allow inside of us. Keep the body and mind pure as if it were a temple. This way a sacred process of transferring harmonious energy from the far reaches of the MultiVerse within into the externally perceiving senses of the physical body can be initiated which allow one to see through the illusions and operate their body and mind with the ability to acknowledge and decide which frequencies will exist within.

Without this acknowledgement of the center then people are pushed around by the external stimuli. This is the whole point, if a person is truly in control, then they will control the stimuli not the other way around.

Programming

Layers of Programming and Decision Making Roles

It’s like asking a person on the inside who participates in one of the various factions why they continue. One might say that it is a decision which is true, but there is programming at each level that is navigated by each individual and this determines the level of awareness within the role.

If a person knows about this situation, they know that there is advanced technology, breakaway civilizations, secret societies, transtemporal planes, the secret of mind and awareness, all of these topics, then they have a decision as to how they’re going to live their life in relation to the public. The outcome of action depends on the layer of programming which is accessed and surpassed via awareness.

The public is at a specific level of programming that is the weakest but ensnares the most people. One who sees the truths behind these groups and operations, they cannot hide from the fact that everyone is controlled. It is apparent, and if not from the results, reactions, stigma, cues, productions and all other media elements of the system which can be read as a code, but from the awareness level of all the “agents” involved. All involved are exposed to truth and power that is beyond anything that is recognized publicly for hundreds to thousands of years. There is always a presence of higher spirit around the world, yet these topics are often reserved for science fiction when speaking of people who can travel the world in a split second or freeze time for their own personal use.

As a result of everyone’s knowledge, there are various levels of programing designed to limit the mind’s access to specific memories and specific frequencies of consciousness in combination with specific memories. The more one resists, the more intense and articulately abstract the programing is designed.

Programmers

The programs are designed by a select group of individuals that reside far behind the scenes and process information at levels we do not have access to without their technology at least to stabilize brain temperature for the extended periods of focus.

Sub-Level Programming

So people at the level of the public are most weakly programmed with false ideals about security through patriarchal authority which satisfied the desire to have a proper father role in one’s life and acts as a pacifier for those who desire no change. Then there are sub-levels of programming with that for anyone who disagrees and is then swept into the next layer of ideals which are based off of all the possible reactionary mind-types.

For instance, if one [i]does[/i] want change, then they are presented a variety of ‘false-exits’ which tend to one or another layer of psychological or spiritual fulfillment but do so in a way that safely reroutes the intended effect of actually accomplishing change with a secondary and preconceived notion of what the programmers want. The programming objective is not necessarily precise in mass scale, from the layer that people are influenced, yet the ultimately or end-goal of their resulting internal desires are a reality schematic that the programmers specifically designed over many years.

This is playing out through each layer of the programming. At any level, someone could pick the wrong way to look at something and begin to unravel what was previously their very understanding of the priorities of the situation. This deals with high-tension operations, moments, programming sessions, decision making, programming upkeep, trauma, memory blocks, the pressure and responsibility of knowing, and the specific plan of the programmers for an individual that comes into their view. The programming is to ensure safety and continuity, as well to decrease the likelihood that a person will reject or consciously come to awareness of or acknowledge the programming or situation itself.

Dissolution of Programming

With that said, the programming dissolves when an individual asks enough questions to necessitate a learning process that introduces variables and possibilities outside of their operations and programming. This is a problem because with this, trauma comes to the surface and all the actions and operations, programs and experiences must be come to terms with as part of the healing and re-unifying process. There are multiple selves or split personalities which are actually just layers of awareness at various levels attached to various emotional or intellectual patterns. All these selves come to the surface and in order to heal the trauma of the split memory concept the original self must come to the surface and acknowledge them.

This is not a trivial process, and the majority of people who have “basic” programming which is fundamental in society, are in a constant state of altered self and have not witnessed the true divine which is eternally present and contains access to all memories, mind-states and awareness.

Advanced Technology, Deprogramming The Mind, and “Space/Time”

As I have described previously, extremely advanced technology is used and this is through vibrations, generators, and electromagnetic devices that have this effect of propagating waves into space/time which can be heard, felt, seen, and perceived in the mind. These are ripples and vortexes in the “mind-space” of reality itself.

As well, there are brain to machine neural interfaces which operate to activate the psychological perceptual changes and explore the subconscious in an aware state. This is used to test and probe the mind in order to bring forth the trauma, the selves, the memories, the reactions and the individual then must come to terms with what is produced.

This is essentially a “good-use” for the programming and healing technology. It is actually a different process with different technology, and programming is much simpler as far as technology. There are difficulties and challenges in the operative’s role in programming and directly influencing and interacting with the target. As well the standby team for deprogramming and the revelation of trauma and altered, artificial personalities goes through a lot. But the use of the advanced technology in this way is a pattern of experiences and unveiling of the spiritual and mental bodies which serves to reverse programming.

Dreams, Technology and Inner-Space Travels

There are inner space travels which relate to soul journeying or visions.

This is when the mind must cope with what is seen in the imagination. It is as if when we dream we are looking into the world of the mind and from this we experience many situations that are often deviations from experiences we have physically in waking life.

As a result of the dream state there is an expanded quality which equates to various strange situations that we wouldn't normally experience with various relationships between concepts or goals that are often reflective or symbolic of ideals and archetypes we experience in life. This is how the subconscious is is witnessed by the mind. This is how information travels, being connected to many times, experiences, memories, places at once, and relates to the one experience of this life experience now (the physical experiences). Through these symbols information is tied together and collective into metaphors which convey deeper knowledge.

The dreams we have are often only a simulation of waking life, and this is due to the dreamer dreaming in a way that produces only a glimpse of the true reality.

In theory, there are no true shapes or forms in the dream-world because this is contained within the mind or even the spirit as a spiritual experience by definition. So then, what guides the various shapes, forms, realities, memories and experiences into being, when we dream? One might say it is the physical experiences but this is incomplete. It is the level of awareness we presume while in the dream-state, disconnected from the physical anchor, that determines the layer of mind and spirit we will be accessing through those dream-visions.

A person who is completely aware in the dream state will reach the level of self-awareness of the dream itself. We are normally simply aware of the dream, if we continue to increase awareness we become aware of the self, in the moment, and this creates a self-awareness feedback loop. The dream becomes a lucid experience of consciousness.

Awareness Feedback Device

In describing this, I realize how fantastically similar to certain devices and processes, that can be achieved using advanced technology, allow the conscious mind to peer into the subconscious and unconscious. This is described through widely accepted Freudian theory of psychology known as psychoanalysis. There are other schools of knowledge from tribal beliefs involving all pervasive 'living energy', to eternal and temporary minds, to the study of 'chi', the meridians, dream-journeying or chaser. The knowledge of the mind and spirit is found everywhere.

These devices used feedback loops to enhance the acceleration of the same cycles of consciousness that induce self-awareness within a dream. The mind becoming aware of itself. This produces a profound effect to the degree where the highest levels of experience become a sacred eternal pattern in which the mind constructs the physical world by organizing perception.

This is the same general concept as an isolation tank yet the outcome and method is more precise. The technology specifically shuts out all other input and exposes the brain to its own feedback loop much like when a camera records it's own output on a TV monitor and creates a hallway of simulated "reverse" mirrors and screens stretching into infinity.

Oneness, Higher Being, Completion, Eternity

The real experiences are found not from the corruption but from the research and exploration operations that were designed to find the truth of our existence. "God" is not only an energy-force but an incarnate being known as the "Hue-Man" and any being for that matter that can simply come to know itself as the Universe. The Human that contains self-awareness rising to high-awareness comes to know the self, the body, the mind as the Universe. The Universe is both within the body as the body itself, as well as the operating force of the Mind itself. Everything is a balance between masculine and feminine features, light and dark, hard and soft. There is no true complete correlation yet the holographic ontology we persist in can lead to further and further adventures towards completion and representation of the "whole picture".

The Holographic Reality

The leads into the next phase of knowledge. The holographic nature of reality may be only one facet of a much larger continuum of consciousness. The perception of time may be related to phase-locking with each other's consciousness and all perceptions may be a variance in the collective phase-locked field which is really the source of variations in reality.

Veiled Awareness, Hidden Knowledge, Eternal Time, Layers of Programming and Control

The interesting notions here are that the programming of the veils goes beyond what kind of situations are found in the bases. However, they do go right up to it. It would seem Humanity was previously programmed on a civilization-wide level through some kind of great "falls" of consciousness, from a high spiritual awareness to a low spiritual awareness into duality-consciousness regarding existence, life, death, mind and matter.

The Great Catastrophes

From these great 'catastrophes' the civilizations of Earth grew denser and denser in nature and actually became what we have today. The concept is that when these vibrations of the mind and body are increased, we actually see a transcendence of the physical realm and preconceived notions of boundaries. It is as if the quantum doorways to all the atoms open up and a new space is presented, but was always there and in connection as the source of all the influences and effects we perceive ‘out here’ from this slower, denser, more boundary driven perspective.

I was brought into the Time Program

I was brought into the time program. You create any reality you imagine. You are literally a powerful reality generation system and your heart and mind are scalar energetic guides for this process. Without guidance and control there is imbalance and imbalance can destroy. This is the nature of all minds and hearts, this is the challenge for humanity to create something beneficial for themselves and the whole to actually know themselves instead of being a pawn or result of someone else’s selfish intentions.

The Effects of Temporal Manipulation

Temporal Influence

When a civilization reaches the level of technological advancement that allows for remote temporal viewing and transtemporal travel what will happen is that the information of the future begins to integrate into the past.

This is why we have all these ancient mystical schools of knowledge that studied time, space, and consciousness in ways that are only just now being publicly confirmed by scientists.

Simultaneous Interaction and Incorporation

The operations that dealt with moving to the past instantly created a simultaneity of interactions and historic events which correspond with that operation.

Due to the nature of time travel which equates to consciousness travel when a person is consciously sent backwards through history they physically experience life from the viewpoint of their genetic imprint in that point in history.

Retro-Causality

What this equates to is that simultaneously that person’s experiences become the part of history that was already present when they left this time.

Essentially, when a person returns their experiences become a part of history that already existed when they left.

Quantum Holography

These are the parallel and possible realities that exist in a quantum superimposed state and can be accessed through advanced technology or a spiritual adept. Often a spiritual adept is utilized along with the advanced technology to operate in and out of possible realities. The mind must be trained otherwise the result is a fracturing of the personality from the root reality into many sub-sets of identities and data streams.

Love is the quantum link that connects all DNA super-computing consciousness. That is, DNA is influenced by intense emotions and this can be reproduces and is what also generates unpredictability in the surface level.

Time Paradox

Time paradoxes are something that one must learn to organize and solve otherwise there will be no way.

The paradox has been solved, the remaining motion is for everyone to acknowledge this and integrate into the reality behind the most conscious aspects of the brain.

What can travel through 4D and 5D space does not make sense in 3D space. So there is a possibility that some of humanity traveled through higher-dimensional space in order to re-seed a ‘distant’ Earth where no life was yet born.

What if it was then concluded that the distant Earth wasn’t distant in space, but in time and was the same Earth that the human scientists left from long ago.

Every civilization that reached the point of re-seeding would have to go through the loop and watch the whole thing start all over again.

Breakaway Civilization

This is an entire breakaway civilization that uses very advanced technology to dominate the world.

I was part of a genetic engineering program that sought to combine various elements of DNA from various sources in order to create a more easily controlled yet powerful and defined person and personality. This involves chemicals, advanced (relatively) scalar wave technology, ritual trauma and programming/mind-control, and many other aspects of conditioning, training, secret operations, physical enhancements, cognitive enhancements, virtual reality, off-world operations, temporal manipulation, spiritual and etheric training or conditioning or programming.

There is self-destruct programming which all ‘operatives’ are programmed with for self-preservation of the military command. The self-destruct programming is designed to take all the aggression one feels building up inside and willing them to escape, and causes this aggression to be turned against the self. That is self-destruct programming.

Most operatives are still under a hypnosis kind of trance which permits one to walk around in public and daily life as if none of this ever happened. Later that week, that same individual could be called for an operation and they would never remember it unless they were forced to kill someone in public out of self-defense in which case most operatives automatically unlock and become unstable due to the flashbacks.

Underground Cities

The combination of advanced technology with secret societies enabled the development of complete breakaway civilizations. One of these civilizations, possibly composed of many factions working in the same area, exists deep underground.

These separate bases are connected via high speed magnetic drive pods. The inhabitants often utilize energy healing devices, as well as perception enhancements and this often leads to a completely different outlook towards the surface civilization.

Technological Advancement

During WWII developments were made involving electrical generators. This sounds vague and there were many other details regarding the scientific knowledge, yet the idea is that very simply electromagnetic, non-Hertzian “Teslian” energy fields were discovered to have been responsible for propagating energy and consciousness in dimensions that were not previously observed. Everything from DNA, consciousness, hidden or regularly unobservable natural phenomena, time, even mass could be seen to exist as part of this hidden, latent, scalar-wave frequency as a pattern is found everywhere.

If one could determine the particular pattern of frequency and location then this can be replicated and anything can be influenced using these non-Hertzian generators.

Radio-frequency alone was developed intensively during the 60’s and 70’s in America and this was discovered to interact with consciousness to influence emotions, cognitive capacity, memory, cognitive function, biological function, health, pain, comfort, pleasure, and all forms of brain-activity.

After the knowledge of advanced technology and the occult sciences of all parts of the world was combined, the ability to target and interact directly with the human soul was refined. This was a mix of psychological tactics used to manipulate or expose the hidden psyche and technology that could manipulate memory, emotional states, wakefulness, and even dreams.

What resulted was scalar technology that could not only influence a person to take on a particular role or response, but technology that could actually put a person to sleep and locate their consciousness to contain it and transport it to a new location.

World-Wide Control Systems

In order to ensure continuity, maintain control, and shape world events these technologies were employed to manipulate the public, alter the weather and natural phenomena, and develop a means of travel and operation that enable one to move outside of the ‘plane’ of the Earth. This is includes high velocity transportation, as well as spacial-temporal distortions which could effectively “hide” one’s entire craft from visible sight. Space itself was altered to ‘shield’ craft with an a layer of ions to bend the light, or the phase was moved out of sync with the “time” frequency of any observer and this literally renders the craft immaterial.

Solar Warden

Solar Warden is a hyperspace-faring race of human beings who utilize cloning and holographic genetic recording/storage in order to protect their race and protect the Earth in this current situation where there are more than one factions using this technology. They are unable to inform the human public because they are in a similar ‘Mexican standoff’ situation that we are mostly all in. If they show themselves to the public, the public will be shocked. If they public does not know this situation, that there are multiple factions fighting for control over humanity, then they will not know who to trust. In that situation, the ‘other factions’ will surely show themselves in an attempt to trick and gather as much human genetic material as possible.

The only way forward to the future is to share the truth about Solar Warden, the space-faring fleet of humans who do no contains the mental and physical illnesses that are currently tearing our planet in two. I believe Solar Warden saved my life as a child during more than one event where I would’ve been entirely maimed/paralyzed or simply dead on the spot as a result of other faction interference. I was in the bases as a child and was trained as an operative for MILABS a secret division of human space-military. The military has been preparing for the future by testing for various flaws and potential uses of cloning and genetic engineering.

Team Light and Team Dark Non-Interference Agreement

Team light and team dark agreed to a non-interference on ground level rule in order to determine how history would play out. Think about it. You have two groups of people, both with immortality, zero-point tech, virtually infinite energy and power with two opposing directives. How do you get to the future? It was decided by elders of each group that there would be no point in fighting an all-out overt war, this would likely destroy the population as much as it would destroy the Earth. As well, the elders of the elders which no body knows would likely have a problem with these groups going at it in such a way so as to change the flow of the planet for every individual thereafter. So it was decided that there would be a “non-interference” agreement for the surface population. Anywhere other than the surface, one can interact with the current population.

All Contact Must Be Initiated On an Individual Basis

On the surface everything that is generated must be done so through living channels in the population and the civilization must be preserved in its level of technological advancement. No overt motions would be used unless an absolute necessity or the majority of the civilian population had naturally come to the awareness of these possibilities. Doing so before this was met would undoubtedly and permanently change the civilization forever.

So what began as a collective motion towards higher awareness became a series of advancements and set-backs all taking place within a short amount of time. This represents the actions and operations of either side effectively helping humanity along it’s path or setting people back with orchestrated events such as the suppression of ancient knowledge, the manipulation of the education or other systems, or the generation of war and prison for profit based on the previous alterations of the education or other systems.

So team dark used its power to help turn society in a warlock ruled prison and indoctrination system of absurdity and scarcity. Team light used their power to help awaken those in need and protect those who's actions are beneficial in the overall plan to save mankind from itself. Team light essentially buffers and shields the population from the operations of team dark.

Team light follows a non-interference agreement while team dark merely pretends to do so as much as is necessary not initiate war.

A Singularity of Change and Awareness

Neither side wants an all out war. This would reduce the dark faction's chances of gaining a majority negative harvest because their motives and presence would be instantly revealed. Team light would also have to show themselves and again step in which is not a pattern of learning that prepares people for the experience of the self and the truth in the Universe.

Both sides know people have to accept the path they can offer them and cannot force anyone into it. They also know that the more this goes on the more likely people are to collectively awaken to the reality of what is happening. It is like an eventual singularity of increased awareness.

The motions we see today are the results of many people coming into awareness at this time on the surface in order to initiate a harmonic alignment as well as to investigate the ongoing crimes against Humanity.

As a result of this, both teams know that the likelihood of collective shift and spiritual awareness is becoming more and more real.

Many events are being allowed because they are representing the eventual shift towards the one end of the polarity from the other which is felt as a shock to society, and this is used to propel society forward through whatever means necessary.

Either people will see what's happening and agree that we've been in a feedback cycle of negativity for the entire age of existence this civilization is aware of, and then we will shift into the real. Or people will claim that this is how things are supposed to work and instead of realizing everything is controlled in a deception based system, they see this and attempt to use the deception based system to gain more collective awareness. If this occurs by the majority then it will be seen as an "acceptance" to initiate a plan that will generate the overt perception of a deception and oppression system. This will surely shock the collective into realizing where they are.

If the majority comes to awareness of this deception based system then the shift moves from the end of the age of deception and external power into the age of internal awareness and harmony.

Advanced Technology and Ancient Knowledge

The plan is to slowly release advanced technology which will undoubtedly connect with and increase the awareness of ancient knowledge. Such technology nullifies the problems that the planet is currently plagued with, although without a developed sense of the spiritual self the technology becomes destructive.

Cloning, Atmospheric Civilizations, Genetic Stability

Here is something that most people might not have contemplated yet is possibly one of the most important factors at play here.

The vehicles that house the breakaway Human society are entirely environmentally controlled. This means that the air and surfaces are clean of all the viruses, bacteria, or pathogens that are numerous on the surface of the Earth.

This creates a specific difficulty in the situation. Any contact between surface Humanity is so heavily and carefully controlled that regulations go beyond any situations we have present on Earth. If the safety and cleanliness of any of the society’s vehicles were compromised this would endanger everyone “on board”.

So this explains a very precarious and specific situation we currently face and this is not so different than “first-world” nations contacting indigenous tribes and similarly endangering their society. Everything that happens will be according to a carefully controlled plan.

Training, Simulations and Learning

Part of what went on in the underground base experiments were to design “tests” or “labyrinths” with various people, props, traps, and treacherous situations that were to be navigated and escaped from.

Feedback loops with the brain, consciousness, and environment can create a very complex and integrated learning experience that is customized to each individual learner.

What this would do is create an individual that was capable of learning faster and faster and processing information in a more efficient manner.

This is not the same as specifically trauma-generating situations.

Cloning and Cerebrally Enhanced Soldiers – MILABS

This includes experiments, condition, trauma-based mind control, enhancement, memory suppression and operations that cover a span of space and time.

The NAZI faction has used cloning and genetic engineering to create hybrids, soldiers and everything in between. Religious/spiritual icons, to slaves, to leaders and even robotoids of sorts.

Your genetics BELONG TO YOU as the LIVING(ETERNAL) SELF. These issues are going to come more and more apparent.

If they agree to allow cloning, then who says who is allowed to clone whoever they want, Einstein, for example. And then who permits who to get DNA from history or objects in the world. These issues are going to have to be looked at because they have already been used in ways that are far beyond the basics.

Then what about consent of the masses? This is clearly being put in public because consent is required on an overall scale and individual to include people.

So what if the masses consent overall, would the people who choose to remain be at risk? Would Tribal and Native village life and people who respect the Earth be protected or permitted to stay?

Psychological Testing

These scenarios can also be generated to create specific situations that will contain tests or elements of temptation that will probe the person's psyche for weakness.

Hallucinatory Holographic Mind Probing Technology

Through the neural-interface technology the minds of many individuals can sync together to experience the same virtual simulation. Thus a single individual's mind can be selected for input and the environment will be a construct of their mental activity. The others present will be able to interact with and explore that environment like the person's mind.

Various training scenarios were utilized to enhance the resistance (or lack which) of the mind to intrusion and subversion. These were harsh but allowed the individual to completely control the energy flow in the conscious mind while remaining in harmony with the subconscious.

In a strange yet perfectly reasonable way, the conscious mind acts as the central "self" in that environment, while the subconscious mind becomes the "background" elements of objects and settings, actors and intentions within the flow of the "scene" itself. So a kind of movie begins with drawn out characteristics and perceptions that match that of a highly cinematic movie or intense dream.

The brain is naturally naive to stimulation and if allowed will liken the event to a dream in which the suspense of attention or lack of lucid self-awareness is "acceptable".

The truth may be worse than what you're thinking as far as training simulators and mind hacking devices. The truth may be that the human mind is inherently lucid and self-aware and that it is external influences that cloud the mind into what we consider 'regular' dreaming and 'regular' waking life. It may be that the conscious mind has a much greater access to information through the subconscious mind which is largely covered up and ignored in a rigidly physical and ego-focused society. As well, it may be that dreams are meant to become lucid and that we are often the subject of mystery in other realms that often refer to humans as ‘sleepwalkers’ that don’t know their dreams are real.

Virtual Simulator Training Scenarios

These scenarios can be entirely holographically generated to produce a visceral 3.5 dimensional experience which can be used for training or psychological testing.

Combat Simulators

The name is self-explanatory enough. Situations can be generated which are indistinguishable from the real events. The entire neuronal map of the brain (connectome) can be created and this is by machines to mimic the brain. After calibration through sensory feedback experiences, the sensation of the simulated realm is 100% the same as the neuronal impulses that are detected by the physical brain. An adept souled being can always tell there is a slight difference, but once the machines are calibrated the physical sensation is generally the same while in the machine.

These environments and scenarios can mimic any combat situation to determine accuracy, integrity, endurance, skill, and all areas of aptitude of an operative without the risk of injury or the lack of real-time effect.

Conditioning

As was explained in other sections, these same systems can be used to remove fear of certain scenarios. They can be used for programming the mind. They can be used to train one to complete a task over and over until they can recreate that action extremely quickly in real-time. These devices can be accelerated to operate within a fraction of the time yet to cover training and conditioning that would normally take place over hours, days, weeks, to months.

Programming and Resistance to Psychological Trauma

Through the development of specifically customized scenarios and neural feedback systems trauma can be brought to the surface or specific fears and traumas can be neutralized. This is essentially “facing fear” but in a completely generated environment that will be just as real as the real thing. One can literally face any fear, any trauma, any kind of difficulty and these systems can be utilized to accelerate a process that would normally take weeks to months or years and individuals can be strengthened to their maximal capabilities.

Accelerated Mental Functioning

Some of the technological systems were designed to test the enhanced functioning of an accelerated brain.

The mind can be seen as the energetic component of the physical brain. The physical brain projects consciousness via the electro-chemical processes.

Technology, Accelerated Learning and Repair

Technology has been developed for assisted learning and accelerated repair of tissues.

On Orgone Energy

Orgone devices produce a scattered signal which disrupts the ‘negative’ energy fields. At least this is what I’ve been told.

Apparently if built correctly it will either disrupt the negative emf/harmonics or it will produce positive harmonics, IE: Orgone energy. Orgone energy is the raw energy of life, while dead orgone, DOR, is the lack of orgone energy and can be “found” (the lack of orgone) in thunder storms and illness or depression.

This is exactly what I was told. We are generators for the same energy, and with the correct principles of energy and internal knowledge combined we can produce those very effects at will.

The energy of the mind creates a central vortex within the body. When the frequency pattern emission is consistent with fractal congruence, then that holographic printing of energy can be expanded or shrunk infinitesimally and it will retain the original fractal configuration.

This is what enables transference of awareness between fundamental harmonic layers of the Universe. When one’s consciousness energy pattern is balanced and tuned to itself so that all proportions can be shrunk through a vortex and expanded on the other side, then the personality or self-awareness remains true. This is as if the space/time must turn inside out and only when there is fractal relative polarity in time does the ‘shape’ of the consciousness within the DNA, work both ways folding and unfolding to represent the same mirror image. This is also symmetry in time as far as an ability to reverse the perception and still retain some kind of connection forward and backwards.

Energy Generators and Serums

Generators which produce vibrations that influence the cells to heal at an accelerated rate, combined with serums that enable the fuel for chemical reactions can instantaneously repair tissues.

I couldn’t believe what I was seeing at first, thinking it was a trick. There are serums and energy devices that can influence the healing and general health of the body and even mind.

These devices use energies which extend beyond the physical and are often compared to electromagnetic waves combined with ultrasonic or low to mid frequency tones. These produce harmony on a level that directly interacts with the biological functioning of the body’s cells.

I feel that this technology utilizes the same energy that is emitted from the mind and it is as if the device is “speaking” to the cells in the same kind of universal language of electromagnetic pulses, tones and vibrations.

There is information on energy devices and the use of light therapy in physical therapy today. More is being learned about the power of light and vibrations to influence the mind and body.

Genetic and Cybernetic Enhancements

Certain individuals are also genetically and technologically modified to produce a kind of hybrid human that can heal at a spontaneous rate as well as operate at a much higher rate of speed and efficiency than a regular person. Serums to induce muscle growth and mental acuity are applied, but the main condition is the enhancement of certain features through the alteration of the RNA aspect of the genes.

These features are combined with tech-assisted metabolism and certain biological functions can be enhanced or controlled via a supercomputer. These are the ‘brain-chip’ interfaces which operate to function as a governor for the mind. When the mind becomes too amped up to focus on the details the faster than thought computer system organizes the input into more usable information streams. When the mind is too slow to see the necessary patterns and variables to accomplish the goal, the brainchip operates in a self-preservation fashion and links together information streams to highlight the most viable possibilities.

Such technology is not without warning, much of this is being used heavily in the military field and this has given rise to entire factions of enhanced cybernetic soldiers. Although this is something to know, there are also many other uses for this technology and it is used by people who want to explore the technologically assisted path.

There are many other aspects of this to discuss, or at least a few important ones that will have to be understood to know the greater truth about what is happening on the planet.

Developing The Self

The issues we face are to help us learn about our selves and better mediate the known and unknown aspects of personality. The more we focus on realizing the unknown aspects of personality the less we are seemingly controlled and forced to experience the undesirable aspects of existence. The more we cling to just what we know and what we feel comfortable with the more shocked we are when we inevitably discover the truth of the whole of our existence which is *always* at least one half more than just what we know and are comfortable with. That is the nature of the universe, we can only achieve one half of the equation at any given moment. Once we complete that seemingly whole view of the universe, the information we gain access to instantly doubles and we realize that we were only looking at one portion at a time, just one half of the equation. Then the rest of the equation comes into the picture and we must expand our view once again. This is because we are viewing ourselves as we discover the universe because we are creating what there is to be discovered through the act of exploration in itself. We are the explorer, the explored, the creation and the discovery simultaneously.

Multiple Factions of Space-Fleets, Multiple Goals

Now there are what are called ‘factions’ of groups that have different agendas and interact with the public or surface level population differently. That is one of the main issues. These departments and factions are no longer limited to the surface level of existence and have access advanced technological capacity and knowledge of a larger view of time and a more complete understanding of consciousness and the human existence.

Large-Scale Catastrophes or Series of Meltdowns

I was informed that while they work continuously to prevent and reposition this civilization continually outside of the reach of any large scale meltdowns that the technology is commonly accepted to only be a temporary barrier between the inevitable.

Unsustainable

This society is unsustainable and must either change or will experience the massive catastrophic evens that are foreseen.

There are multiple perspectives that are maintained. They vary by the way of the methodology of solving the problem. All agree that change must occur and is impossible

to postpone forever. The future society is the result of our advancements and achievements. Regardless of how we get there, we must get their somehow.

Earth Changes

I was informed that these changes had been held off and would ultimately be allowed to take place to resolve the current situation. The changes can only be held off for so long, either that or the people performing this task would not be reasonably inclined to postpone these changes indefinitely.

The idea is to locate and secure the members of society who are fearless and capable of assisting in this transition and who hold key genetic elements as these individuals and groups are capable of ‘holding-space’ for the rest of civilization by literally containing the genetic precursors for all those other individuals.

I was informed that Earth changes would be the last event to take place, pretty much like icing on the cake.

This means there will be a lead up to these events.

Mass Consciousness

I was informed that the consciousness of humanity and every species is linked to the bio-etheric field of Earth. When the consciousness becomes too unstable to maintain a controlled society that this will initiate massive Earth changes. This could simply be instigated through advanced technology but I was informed that the advanced technology is actually postponing this so as to allow people to prepare, to get the word out and avoid a complete loss of the human race.

Wars and Manipulation

Wars were described as a way the power groups had controlled humanity over long periods of time. These wars were used to distract and weaken the population and the power groups believed they were holding off a full scale invasion by suppressing the rising population.

Now the population is high enough that another war would be a final war and this is extremely dangerous for the entire civilization. There would be no need for another war to repeat this cycle because the technological achievements have been met and so essentially the ‘reseting’ of the civilization would not be required or possibly acceptable.

This is why war is to be avoided however people are very easy to manipulate through lower emotional and mental programming from media and social cues. People follow the herd and the herd is essentially still operating on animalistic parameters. There is a have coating of peace and humility but this is just a wax coating that quickly evaporates as soon as the pressure of mass unrest or scarcity is realized. Many simulators have explored these possibilities.

Financial Unrest (or reset)

The idea is that financial oppression is used before wars to distract and oppress the population. When people are busy looking for resources to survive, they are not focused on the larger picture. This is part of the ‘fight or flight’ response mechanism tied to lower brain complexes which are primarily activated to keep the masses in both a highly sensitive and unstable state of being while also keeping them easy to control.

If the civil unrest reaches a level where awareness continues to rise even though financial hardship is being implemented then the latter two options will not work. At that point artificial disasters can be produced as well as natural Earth changes which initiate a shift of the ages and of consciousness.

The paradox or maybe irony here is that the oppressive actions are used to both suppress human awareness up to a certain point and simultaneously to stimulate and arouse the human spirit into wakefulness and self-empowerment. This is part of the reasoning regarding why these control systems are allowed.

Some say the elite don't want an economic collapse. That it would ruin everything. Others say it would be icing on the cake. Some say that it is a zero-point world that will save us. That the Fukushima is a big problem. Zero-point field tech, that is. Generators and energy sources, healing and terraforming. I know we project right? We see it in our own lives. So we can project with each other and grow that as a group and that part of the world will exist in that light. Which sounds Biblical, which may be because they're following prophecy. But it may also mean that we have a light, a chance, a real salvation, but that we have to grow that here and help develop it. To pass it on like love and such.

That's apparently what it comes down to. That these cycles happen every so often and we're in the midst of one right now.

Let's make it count. Every day, thought moment. With mindfulness, intention and awareness in action. Life is literally higher dimensional, so we have to become multidimensional to really face it. Indeed, that is a good perspective on action and inner being. It's like a fractal, whatever we project inside that chest center, will project into our mind and the outside. But we have to be stabilized and grounded through the root first and then to get to the heart energy must pass through the sacral/sexual centers and continue upward.

Self-Responsibility

The main reasoning is self-responsibility. We are responsible for ourselves. The people are and so if they allow themselves to be tricked and prodded like animals while behaving for the part, then that is how their experience will play out.

Clones and Synthetics

One seemingly unfair aspect of this is that there is a larger portion of non-humans than there are humans and so these beings without the human spirit will be easily programmed to support and propagate the mind-control system and the oppression. Thus, if people follow the herd mentality then they will walk directly into the end-time scenarios yet it is the humans who will actually stand to lose. People must wake up to these tricks and this technology so that they can orient themselves with the reality of this situation while learning to strengthen themselves and adapt to the manipulative aspects of this control system.

Spiritual Power and Divine Truth

This is essentially all about spiritual power and humanity stepping into the power of the truth behind their existence.

Some say that humans are designed as slaves, but if this is so, then it is the power that humans contain within that is the ‘work’ that they are sought after for. This indicates that humans have great powers that others cannot achieve on their own, even if they are creator lords.

Thus, whether humans are modified or not, they have great power. If they are not modified to perform this role of an energy generator, then this is just a confusing mind-control game that has been pulled over the eyes of humanity and humans have always contained true internal spiritual power that allows them to exist without having to parasite and steal the energy and knowledge of others. Humans are reality generators , we can choose how we want to experience the universe and what kind of collective world we want to enjoy together. Once we step into this power, there is on greater force. No amount of non-human entities can change that as we are somehow intrinsically linked to the ether-space of this realm through the spiritual and physical aspects of our being. That is what is sought after and so those who do not have this manifestation power attempt to utilize humans to create their own reality through this Earth situation. That is the basics of the ‘alien overlord’ system, whether humans in disguise, non-humans operating with alien intelligences, supercomputer systems, or spiritually degraded beings.

Most people are not ready to see themselves or the true nature of humanity. Reality is largely an illusion of the conscious mind. The hidden aspects are the other half of the equation and this is perceived by the shadow aspect of the personality and mind. We are only marginally aware unless that aspect has been unveiled and through this process the reality of the human soul can be actualized and realized.

I was informed there would be a period of two suns and multiple mirages around the world. People would believe they were peering into another dimension, floating cities, or some kind of city-ship craft.

Synthetic DNA, Synthetic Genetics

Through the previous described replicator devices synthetic DNA can be engineered. Thus, entire genetically engineered codons can be produced and form this created beings can arise. These are the hybrid beings that are performing many tasks underground and even on the surface. On the surface there are synthetic genetic create beings or genetically engineered beings that appear to be human but do not contain a human genome. There are also clones and replicants of humans. There are a variety of non-humanoid forms but these are largely kept underground. These are beings we often see in science fiction movies or shows.

Time as Mathematics, Series of Equations, a Programmable Storyline

When using the technology, time can be understood to be a series of programmable variables where one variable is centered within a series of parameters and from these possible variables within a parameter set a series of possible functions or outcomes can be process. Thus, these people can literally plug into a supercomputer system which operates based on this form of processing reality and through this they can observe all the possible realities that may connect to a present moment. This allows one to travel but this is not part of this section of the explanation.

What this allows one to do is know exactly what is going to happen and when. If a person is being interrogated, all that has to be done is the operative using the technology simply thinks of what they will say and intends to do it within 6 seconds. The machine will play out a future timescale regarding the input of that question or interrogative speech and then produce the possible results. If this doesn’t give the desired readout in the mind of the operative, they will refuse to go that route and think of another pathway. This repeats for a few seconds to a few minutes. Then the computer has found the pathway that is most like to yield a result. The operative simply steps forward, says ten words, the subject breaks immediately and tells them everything they want to know. This is obviously takes longer for most trained individuals, however that is how easily it would be to use this technology to produce an effect on a regular surface level human. They would have no defense, no way of knowing how that individual could know so much or be so direct. This is how operatives dealing with certain power groups will operate in the surface level population and individuals always feel that these people know way more than they are leading on. That is because they are, they know more about you and your life experience than you do.

Chapter 2.3 Previewing Solutions

Neutralizing Polarization

The mental-emotional tendency to pick sides, to fight, or to basically join in the fray that is currently taking place is a kind of deception of polarization which enables the leeching of emotional and mental energy from the human. This is more of a psychological unveiling where the true nature of the self, the duality game and the transcendent source aspects can be made visible to the individual. The they will see how every polarization tactic is actually taking them out of their true source of power by pushing them off center emotionally and mentally. There are teams that are capable of enabling this process and assisting as humanity realizes that half of the problem is that the unchecked mind is one's own worst enemy.

Removing Etheric Implants

Technology can identify and isolate these foreign body implants which are etheric technological devices. These are literally frequency or phase-shifted devices which are invisible to this layer of physicality but are capable of being organized and attached to the etheric field of the human and will operate to siphon or limit energy on the etheric level. Of course, this is all technology of the vampiric created beings who parasite on the original Earth inhabitants to sustain their existence.

Clearing Astral Body

By moving through the temporal body etheric implants can be deactivated and removed and then the astral or emotional body will be able to function closer to the original levels of presence and sensitivity. This body is manipulated when one is emotionally shut down from trauma. Thus this body can be scarred and will function at a much lesser level of activity or presence until the emotional damage is healed. This contains the memories of all the emotional events of a person's life.

There is technology that enables one to access these levels of frequency and clarify or untangle emotional chords or attachments which are used to siphon energy from the human being to the parasitic created beings. These chords are energetic in nature but also have technological components.

Knowing the True Self; Soul Actualization, Unveiling

This is a sacred, ancient process that has been outlined and discussed since the beginning of time. Humanity is here to learn, grow and express the self. In doing so, like a child moving through layers of infancy and into full-wakefulness of their body and the physical world, humanity can come into awareness of their true spiritual nature and the true spiritual nature of the universe. This is the process of soul or self-actualization that is described in some psychological methods such as Maslow’s hierarchy of needs and the self-actualization that results in a complete fulfillment of all psychological needs. In this modern psychological view, self-actualization is described as the desire and process of reaching fulfillment of one’s own natures and then living their life sharing that self-empowerment with other beings because after gaining your self that is the highest expression of such knowledge, sharing that power with others.

The true self is not limited to the physical body, the true self is a reflection of the entire universe and thus the primary original creative force. This is akin to the whole equation rather than just half of the equation. The visible aspect of the self, the conscious mind, is only one half of the entire equation. The whole equation consists of what we see visibly and consciously as well as what we are limited from seeing within the unconscious and subconscious mind. These are really two layers of the same ultimate source. Like an iceberg which is only partially above water, the subconscious mind is the proprietary wholeness of experience. The conscious mind is only a surface layer of reality while the unconscious mind carries all the notions, the multidimensional processes, the latent aspects of reality that we are not capable of seeing without unveiling these aspects through inner exploration.

Taking Responsibility

This entire show is about taking self-responsibility. Unless people begin to do this, there will be no hope. Right now people are convinced that their responsibilities belong to someone else even though this always results in pain and discomfort. To take responsibility is painful at first because we have a lot of work to do, but even then, this is beneficial because there is no way to remedy the pain and discomfort of being lied to and disrespected if people continue to hand their free-will over to others through blame (scapegoating), lower emotional bio-emissions (hatred, anger, etc), distractions and addictions (technological, sensual, and chemical or food based), or generally complaining without taking the first step towards providing something beneficial. This is the victim role, if humanity is to take control then they to help each other and themselves by not relying on false promises and false authorities to determine how they feel, think and live.

Returning To Center Emotionally and Mentally (spiritually)

Balanced food intake, rest, exercise, activity and learning are all continual requirements for a healthy existence. Emotional and creative expression nurtures the soul. Without these ways of being and learning the human qualities of experience diminish and creativity dwindles. There is most likely going to be a split between those who can sustain their own humanity and those who have to cause others to suffer to do so.

Vampires in the “Awake” Community

Contrary to popular belief, the majority of the non-humans are not necessarily vampiric and in power. The vampires are the people online who try to get a rise out of others for sharing their own experiences. They are the nosy neighbor that feels that the person who looks or acts differently must be considered and enemy and armed against because they feel themselves to be so much better. The majority of people are vampires, they believe that they must work for someone who will trade them money and that the amount of money they have based upon the amount of indirectly beneficial work they do puts them above others.

This is not different in the ‘awake’ community and in fact I feel that there are higher concentrations of vampires in this community, the alternate media because everything is largely a joke and a trick. People are operating on deceptive agenda schemes to divide the people and this is most intensely noticed when there is a realization of the truth behind what is happening and people begin to point fingers. The only people who are cleared to point fingers (and who also will NEVER point any fingers) are the operatives from the secret projects that have the capacity to do so because they have reached a specific level of awakening and this is related to the droning, cloning, and hybridization process. Only a specific few know who has been transformed into an impostor being and who is remaining strong and this is due being briefed on the surface level situation. Anyone can say what they want, they are simply playing into the tactics of mind-controllers and parasites and the more they play in the deeper the infection will grow. All the people pointing fingers now are actually implanted and vampiric and are attempting to drive a divide through humanity and they are the first who will turn on their friends and family when the going gets tough because they are more interested in others and starting drama than healing themselves. Ironically, this is all done under the guise of “healing” and “getting rid of the drama” AKA people they don’t trust. They pick and choose based on their own personal bias and this is a sure indication that they are emotionally and mentally out of alignment and are feeding deep seated personal fears from their own traumatic experiences.

These traumas must be healed and the harmonious center must be reattained where a person can focus on their own existence and waking up instead of acting as false-light, “king of tyranny” gatekeepers for others. Those individuals are going to be left behind if they cannot kick the habit because if they are allowed into any kind of healed society they will quickly degrade into animals and attempt to restart a devolution process in order to gain the power they so desire but cannot produce on their own. ANYONE playing ego games of trying to hurt or put blame on ANYONE else contains this viral implant and has not kicked the virus.

Heart Center and Psycho-Spiritual Fulfillment (self-actualization)

Emotional connectivity is a form of telepathy, when loved ones can tell the state of another across space and time. Whether it is family or friend. In the 60's and a little before that the US found a specific frequency akin to a radio frequency that operates on a trinary spectrum. They called this the "telepathic wave". In time, they could completely read and send thought waves without frying the brain or body through radiation. Which was what the early machines did. They actually took satellite dishes and aimed them at people. Then realized they could turn it around and do it backwards with a better effect and minimal radiation. Either way they started beaming the base and then towns around them with waves between 400 and 432 MHz and caused waves of crime sprees or peace sprees throughout the 70's and 80's. So they can digitize the telepathic wave to induce artificial telepathy. But this is organically done through an entanglement through what we call love. Literally. The Germans found it before that. And the secret orders before them. And ancient Egypt before them. As humans we remain basically through the root, heart mind connection. When the heart overpowers the mind connection of spirit, then the body is a trinary whole. When the mind guides it is binary and cannot by guided by the spirit. The spirit has direct access to the heart. The mind is a computer of sorts. The mind doesn't see past duality. The heart is the true center/eye of the spirit. The reptilian brain is the dragon and when you tame the amygdala then fear is ‘behind’ you. When the mind is "untamed" the reptilian brain takes over. This is the lower brain and amygdala fight or flight response.

This happens when the energy of the brain, body and spirit increases due to whatever various reason of spontaneous or controlled increase. Because of the "temptations" which are the easiest flow of emotional energy into a "discharge" state, is that the sacral/sexual center ties to the heart and reroutes to replace the heart and control the mind. Mental and emotional contemplation of one's true purpose and nature is the path of the divine. We find ourselves simply by looking in the right place, not in the world, but in our self energetically. We are more powerful than the programming, and we can reverse this is we're careful and devoted enough.

We have to be faster than thought, at the speed of spirit, beyond matter. When we start thinking on this level then our influences spread to others throughout the sphere. It is our destiny to know these things but they have been kept from us. The ancient orders called this the art of "implosion". Traveling realities through intense contemplation and a usage of sacred geometrical forms.

They developed and combined orders around WW2 and developed zero-point energy tech that can alter space/time fields. This is timeline manipulation. But as a rule, we go where our heart leads us. So that's how this is solved.

Go Within to Observer the Self

The answer is literally by looking within to find the great truth rather than looking around us. Everything we need to know is within us, energetically and spiritually. Our emotional, mental and physical state of balance is an indication of where and how we exist in relation to the larger overall truth. Everything that happens around us to distract us is designed to stop us from looking within. Nothing can truly stop us, nothing can actually input the holographic horrors in our spiritual center but simply perturb the brain and body. The mind is a non-holographic spiritual environment that cannot be excessed and the invader's goal is to convince people to use their imagination or their creative powers to produce these horrors on their own. This is through the programming. We must clear out the programming, the trauma, the lower-dimensional constructs and fill ourselves with the truth and spiritual presence that enables us to be self-aware in the first place. If we aren't doing this, then we are either in pain, or there is no spiritual presence to be in pain. If we are doing this, we are either clearing out the constructs by replacing them with true spiritual presence and self-awareness or we are becoming aware of them.

Organic Feedback Loop

The great feat is mentally bridging the great rift between us by looking within. Which is paradoxical.

We can never see what another is perceiving exactly, well for the most part, so we all have to kind of accept that we are communicating effectively and truly get to this level of comfort by simply being secure in knowing who we are our selves.

Technological Feedback Loop

So then knowledge, in a way, when shared, is a form of telepathy. Group telepathy.

Group think, when ideas or memes spread, is a form of group telepathy.

The Internet is like artificial spreading of telepathy. This means what is normally mentally bound signals can be initiated and spread through a digital medium.

Breaking Soul-Contracts

Ultimately everyone is under a series of “veils” which are energetic time-constraints literally like 9 dimensional layers of temporal restriction which are each bounded and ruled over by invader beings. These range from physical authorities, to astral entities, to programmable matter, to supercomputer intelligences, to spiritual false-light beings and our own shadow mind. Each layer of reality is successively guarded by these beings who are here acting as gatekeepers designed to push people back when they begin to wake up. These beings all work through fear and deception and the last layer is based on mirroring our own energies back to us like an impostor. All the levels tend to imitate, however the last layer is a near perfect imitation of our own energy to the point where people often switch places with this being before getting out of the matrix. There will be more discussed on this in a later section.

The “9 veils” were designed by impostor beings as a last effort to stop humans from escaping the false-light matrix which is essentially the holographic universe. This is literally like a maze which reproduces a new pathway and a new doorway each time a person nears the last hallway or doorway that is next to the exit. So one eventually has to project themselves into the exit using their whole being and moving past and through the tricks by not falling for them before they are placed. This is highly confusing but this is essentially a kind of labyrinth that has been created to stop people from leaving so that their soul energy can be destroyed and used as a generator for the false-light system which is an AI over mind that is programmed to take the raw soul energy and destroy the identity by inverting the electromagnetic spectrum and incorporating that inverted energy into its own internal matrix body which is literally the false-light universe.

Help Everyone

Now they may have been setting us all up for the apocalypse. They said they want us to tell on them and get the whole world aware of what they've been doing So that we can stop them and pull them from the dark which is ACTUALLY taking their souls. There are also breakaway civilizations involved in this entirely zero-point/free-energy.

Chapter 2.4 How Timelines Collapse

Repeating Cycles of Time

Because of the way time works and the ability of those using these temporal devices they have set up an elaborate labyrinth like a spider web of false realities, timelines, parallels, sub-dimensional layers, astral cages, temporal dead-ends, and time loops which act to stop people from reaching the true reality. As a result, many portions of this timeline are literally in a free-will ‘time-loop’ and these loops must be canceled out. As well, because of the nature of these loops on the overall stream, our entire civilization has been looping for the past 6 time-cycles meaning we have reached this time in history 6 times before and this is apparently the last time because we have never gotten this far before and the events of the future are having to be programmed in manually at this point.

Beings From the Future

As are result and damaging time, what happens is that the beings from the possible future that is being damaged appear to try to heal the break as well as those who are from the new possible future that is generated, this results in a time war between both possible realities for the present realm. This is what has been happening for thousands of years. All the ancient spiritual texts are describing this system. One group is a from an original possible future where everything is based on the presence of spirit and souled humans, another is from the new possible future where an AI overlord took over and removed the souls and spiritual energy from all entities on Earth rendering this a holographic prison without any other use. That reality dissolves into itself and so the time beings are constantly moving backwards and jumping timelines to try and obtain more energy.

Beings From Collapsed Timelines

The beings from the collapsed timelines are those who travel through hyperspace to get here. They then recreate themselves using cloning technology through hyperspacial means, IE: creating an alternate temporal dimension so that these devices can sustain them and literally “hold-space” for them. Then if their timeline is destroyed, they can still exist as a ‘loose-end’ in space-time. This is extremely dangerous and I will say we have some Human operatives jumping timelines searching for Earth and Humanity so that they can find a reprieve from this time war.

That is the most trying aspect of this entire ordeal for me. There are some human operatives literally in hyperspace without an opening into the temporal stream for them to reintegrate into this timeline. We must assist them, we must open the timeline by taking control of Earth from the invader races and those who wish to depress and isolate the spirited and souled human beings. We must re-organized and neutralize the invasion forces of deception and assist our humanity and maintaining their free-will, their self-awareness, their self-empowerment and their emotional, mental, and physical harmony.

Beings Who Would Never Have Existed in the First Place

As a result of this entire process, there are essentially beings who have been erased from time who degrade in this dimension and can only interact periodically. These are the created beings who operate through supercomputer forms that are placed within cloned bodies and powered by the transdimensional quantum supercomputers. These are the ‘parallel’ reality beings and this relates to the hive mind invasion. When all the original aspects of time are restored, these beings would not have ever existed in the first place as there is no quantum link for their probabilities to remain neutral and present in the original run of things.

Again, all beings who operate under the will of the universe, even when they are attempting to trick hyperspace and time, when these beings assist and help in the unification of sentient, spirited, souled beings they are capable of being healed. This is all because there is an original creative force which can produce souls that are eternal and everlasting, meaning they are not created but simply always ever were. Yet, there is a “miracle” effect where beings can be given eternal life. This is only by reflecting the will of the whole of existence, rather than a portion or a part because this is only temporary. Only the whole is forever and this takes true harmony, true balance and ultimately *self-lessness*. To give to others because you see them as worthy because they can be aware and enjoy the gift is to be like the creative force that gives life to all and enables all experience in the first place. By acting in correspondence with this we reflect the will of the whole and that power and action transfers into protection and self-empowerment for ourselves. This cannot be cheated, tricked, or rerouted as that is the karma game that has been played by delaying time and rerouting energy through getting others to take on the responsibility for others bad actions. In the end, all karma has only been delayed because when the stops are pulled out, *everyone* ends up getting their karmic returns regardless. They simply intended to build up so much karma that by the time the ties are cut, the karma overload will simply kill them instantly. This never happens and they end up in the abyss, this is the final time so since there is no recycling of the universe from the oroboros system then whatever happens is what happens forever. If everyone wants to move into harmony, to receive healing, to receive forgiveness then they must assist in the creative force in healing, forgiving, and turning away from the deception, the cruelty, the vampiric actions.

Many vampires believe they will die if they stop. The only hope is to admit their ways to the others and ensure that there are enough beings around them that will protect them so that this situation can be handled properly, with due care, compassion, and the necessary healing to ensure the safety and protection of all who seek harmony.

IronAmmonite.com

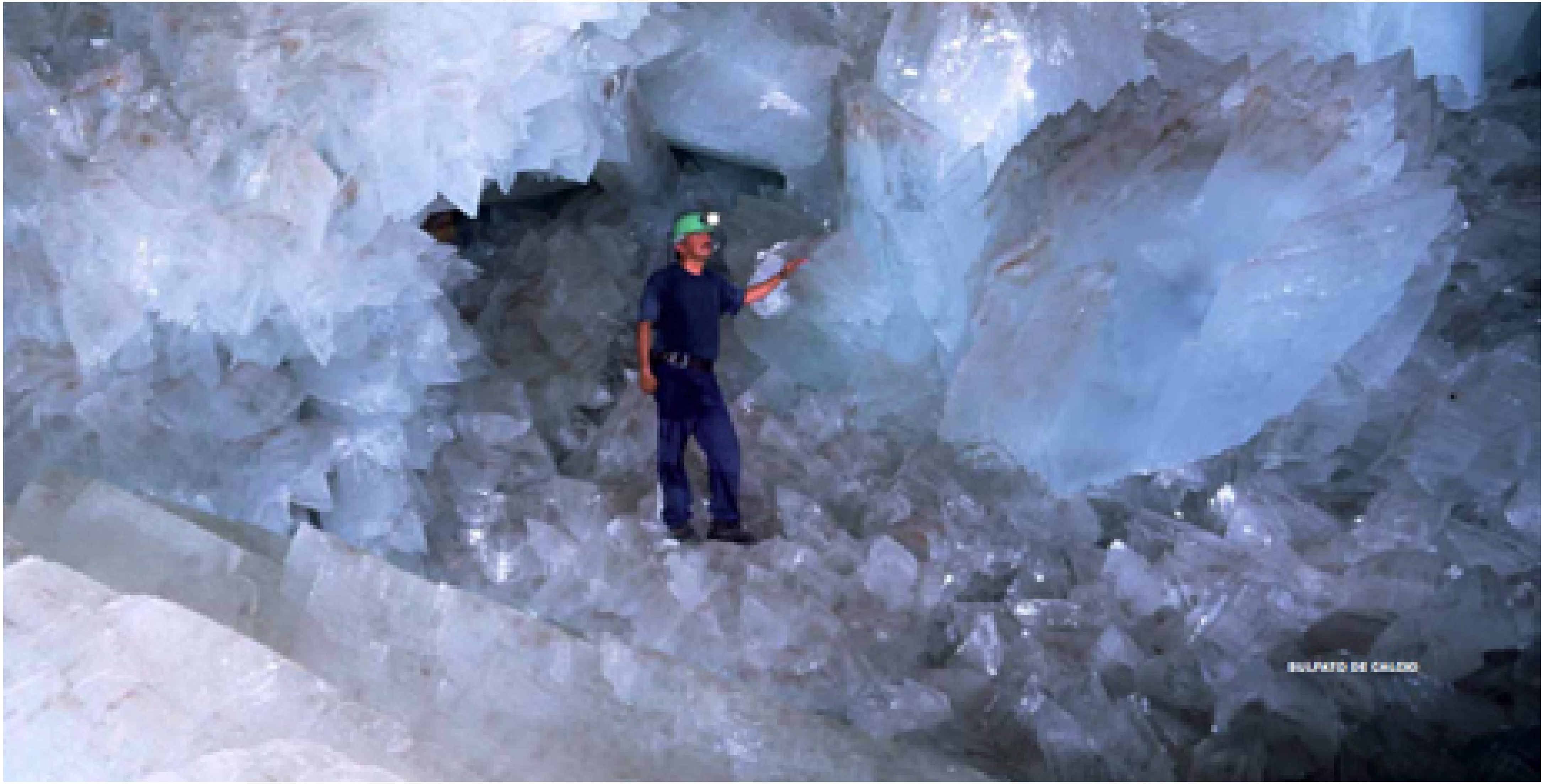
SULFATO DE CALCIO

Schematic of the RF-system

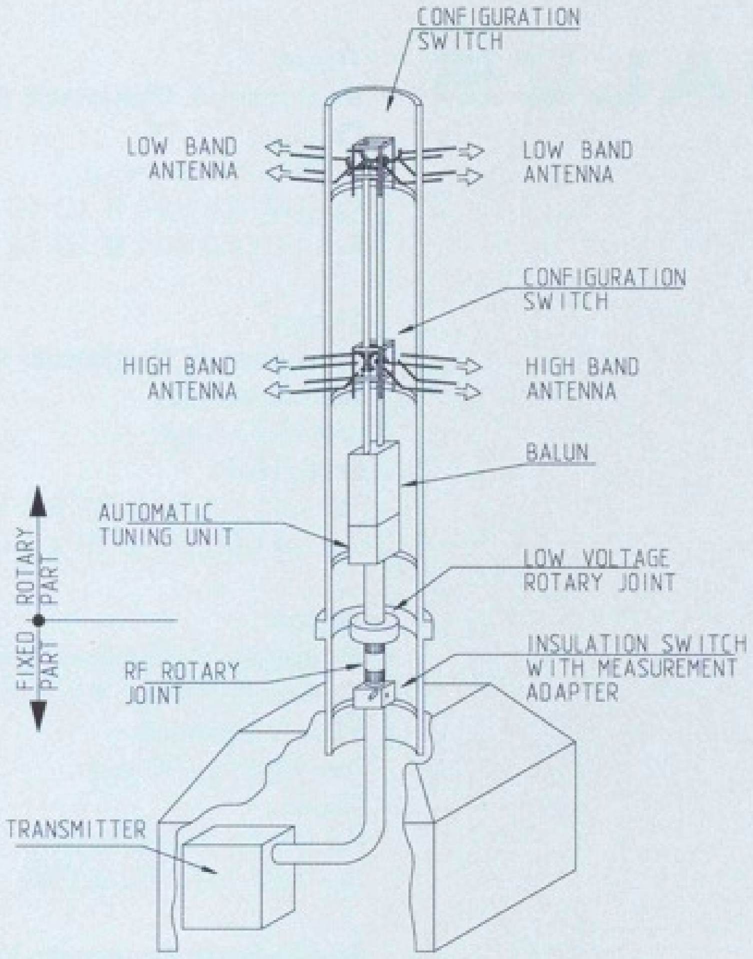
CONFIGURATION SWITCH
LOW BAND ANTENNA
LOW BAND ANTENNA
CONFIGURATION SWITCH
HIGH BAND ANTENNA
HIGH BAND ANTENNA
BALUN
AUTOMATIC TUNING UNIT
ROTARY PART
FIXED PART
LOW VOLTAGE ROTARY JOINT
INSULATION SWITCH WITH MEASUREMENT ADAPTER
RF ROTARY JOINT
TRANSMITTER

PRINCIPAL LAYOUT OF INDOOR RF SYSTEM

Design

- Full in house engineering capabilities
- Latest design tools
- Rigid tubular steel dipoles
- No suspension/counterweight system
- Overall DC grounding of antenna and screen
- High-Band/Low-Band array
- 30 years of experience

Performance

- High reliability
- High availability
- High flexibility

Maintenance

- Easy access to components
- No lowering of dipoles and screen

Cost efficiency

- Low life cycle cost

More than 30 years of experience
30 years of antennas on air

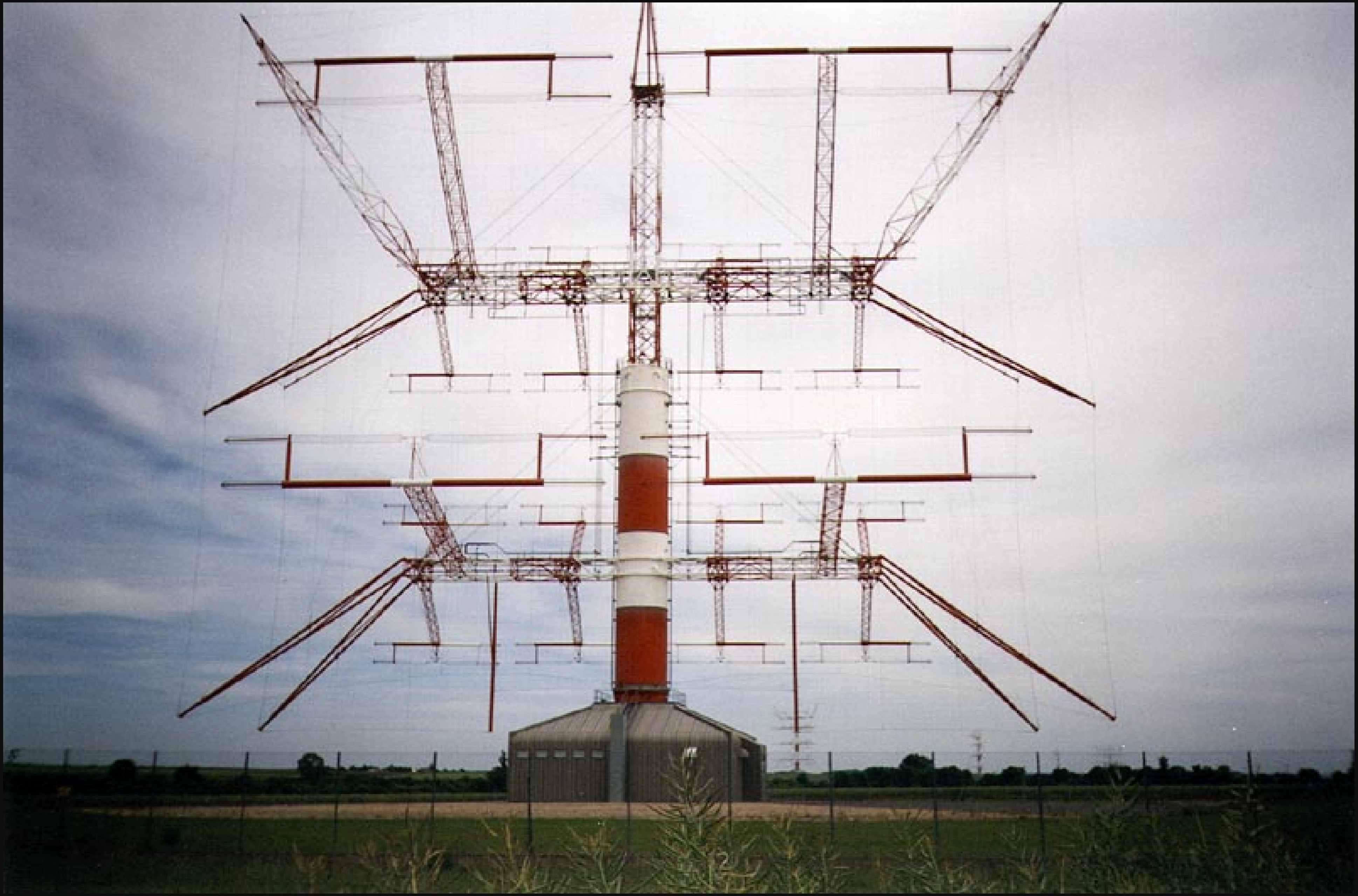

COMPARISON TO CURRENT TITLEHOLDER

ARECIBO
Since it was constructed in 1963 in Arecibo, Puerto Rico, this telescope had held the record as the world's largest telescope.

FAST
It's designed as a cable-net supporting structure, capable of forming a parabolic mirror and will be completed by 2016.

SIZE

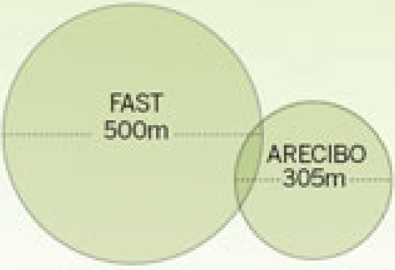

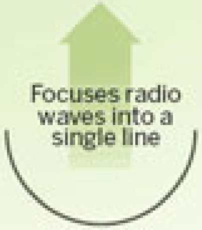

Arecibo has a fixed spherical curvature.

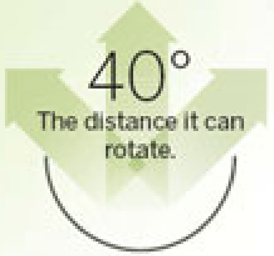

FAST can reshape into a paraboloidal surface.

ASTRONOMICAL GARDEN

URUMQI NANSHAN
25m
Radio telescope
Built in 1994 in Urumqi, Xinjiang.
Missions: Pulsar research.
Determines orbits for the Chang'e lunar satellite and other lunar projects.

QINGHAI DELINGHA
13.7m Millimeter-wave radio telescope
Built in 1990 in Delingha, Qinghai.
Mission: Interstellar molecular spectroscopic research.

MIYUN
50m
Radio telescope
Built in 2005 in Miyun county, Beijing.
Mission: Long-term monitoring of a group of known millisecond pulsars.

MIYUN
Synthesis radio telescope
Built in 1983 in Miyun county, Beijing.
Missions: Northern sky survey and observation of Supernova Remnant and interplanetary scintillation.

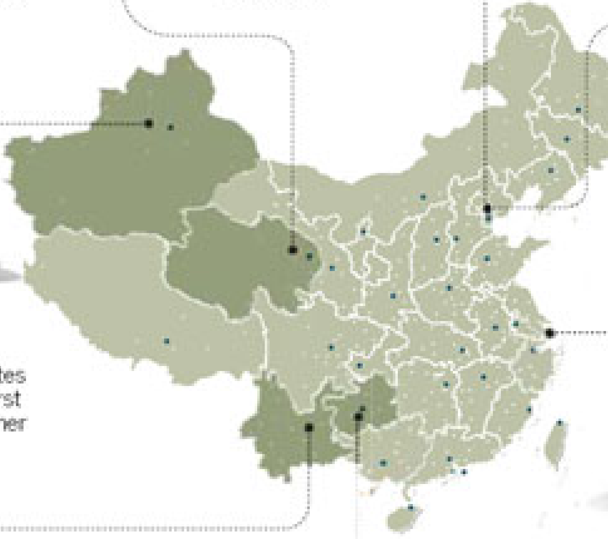

KUN-MING
40m
Radio telescope
Built in 2006 in Kunming, Yunnan province.
Mission: Successfully completes the VLBI tracking of China's first lunar probe Chang'e-1 with other Chinese telescopes.

SHANGHAI SHESHAN
65m
Radio telescope (under construction)
Built by 2012 in Sheshan district, Shanghai.
Missions: Finds applications in deep space exploration and basic astronomy research and assists China's lunar exploration program.

SHANGHAI SHESHAN
25m
Radio telescope
Built in 1986 in Sheshan district, Shanghai.
Mission: Assists the VLBI (Very Long Baseline Interferometry) network.

GUIZHOU PINGTANG
500m
Aperture spherical radio telescope (under construction)
Built by 2016 in Pingtang county, Guizhou province.
Missions: Surveys neutral hydrogen in the Milky Way and other galaxies, detects faint pulsars, looks for shining stars and listens for possible signals from other civilizations.

Source: Skyandtelescope.com, universetoday.com, URSI

GUILLERMO MUNRO / CHINA DAILY

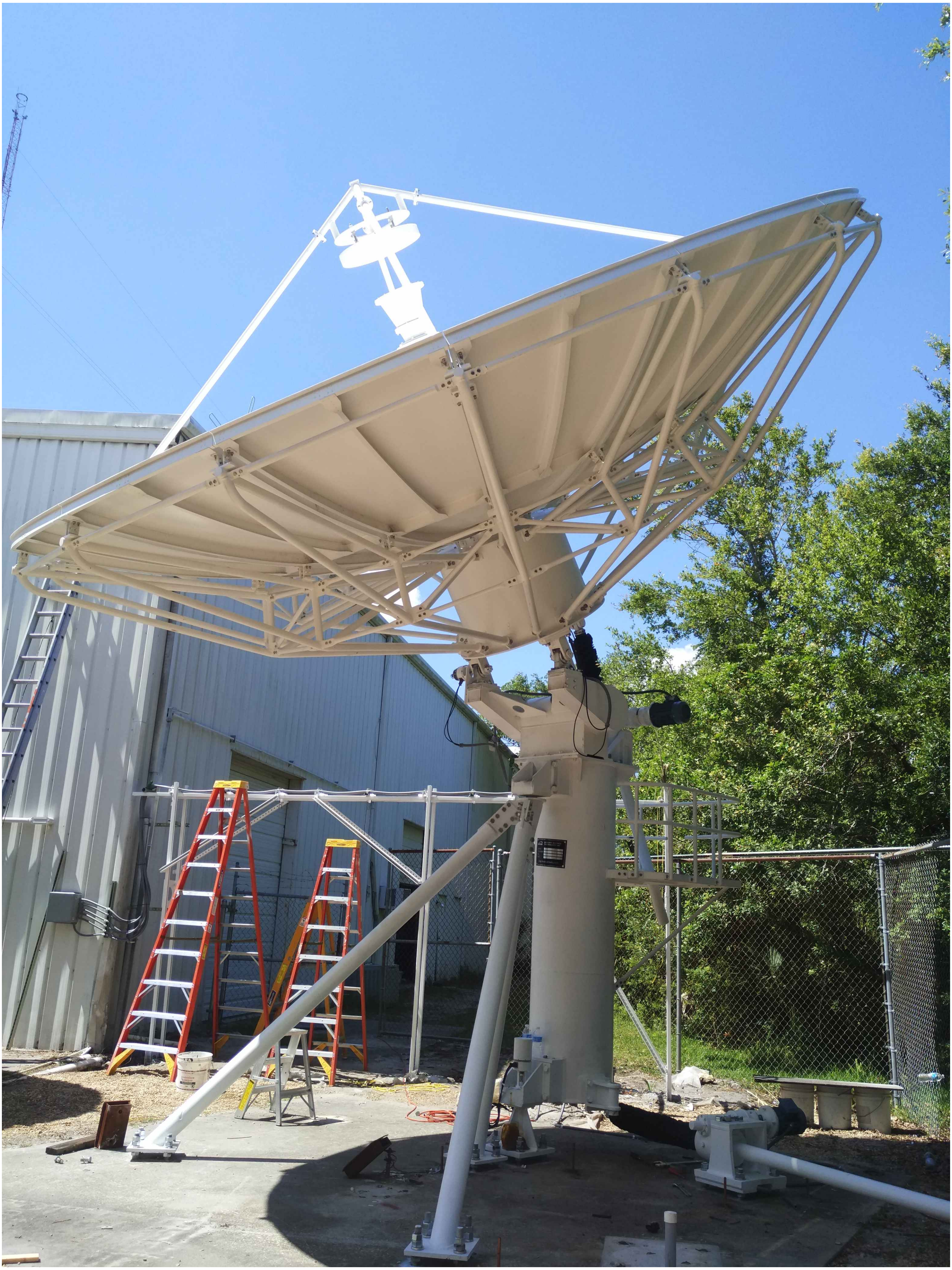

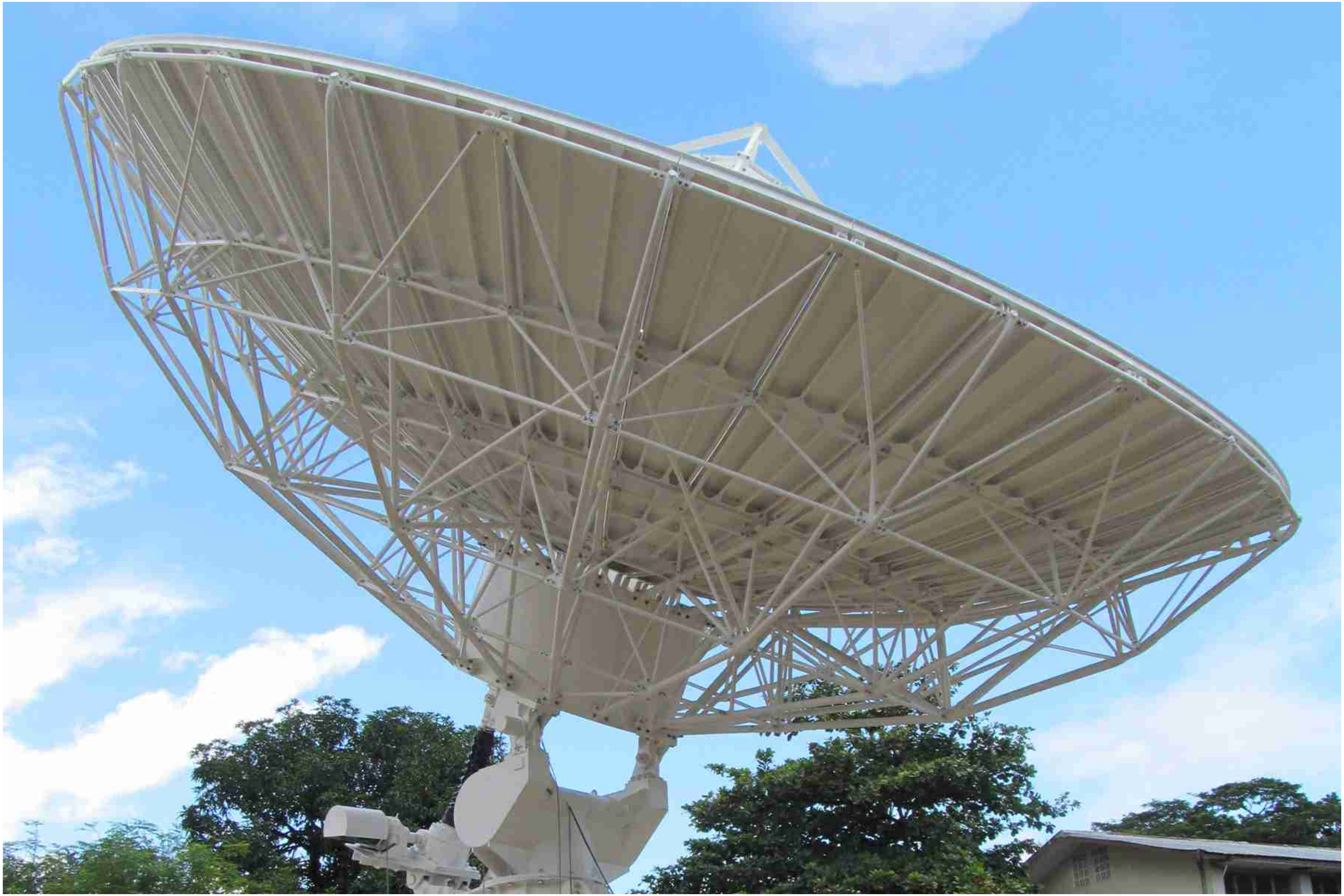

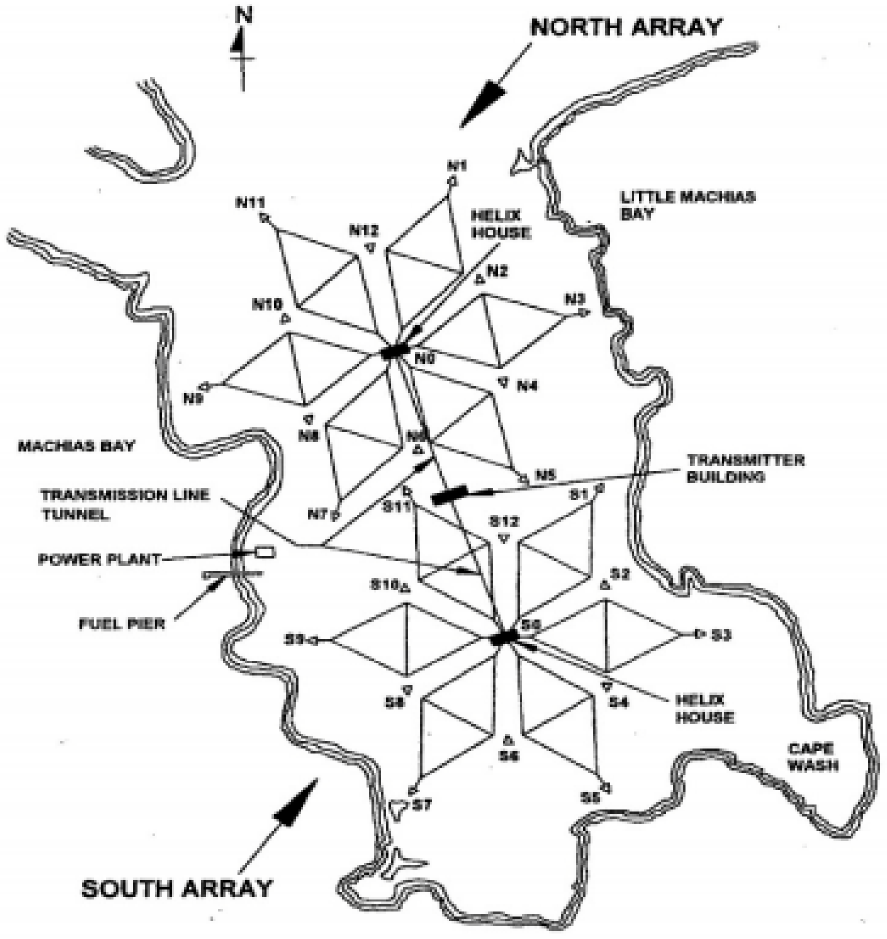
N
NORTH ARRAY
LITTLE MACHIAS BAY
N1
N11
N12
HELIX HOUSE
N2
N3
N10
N0
N9
N4
N8
N6
MACHIAS BAY
N5
TRANSMITTER BUILDING
TRANSMISSION LINE TUNNEL
N7
S11
S1
S12
POWER PLANT
FUEL PIER
S10
S2
S9
S0
S3
S8
S4
HELIX HOUSE
S6
CAPE WASH
S7
S5
SOUTH ARRAY

Figure 1. General layout, Cutler Peninsula.

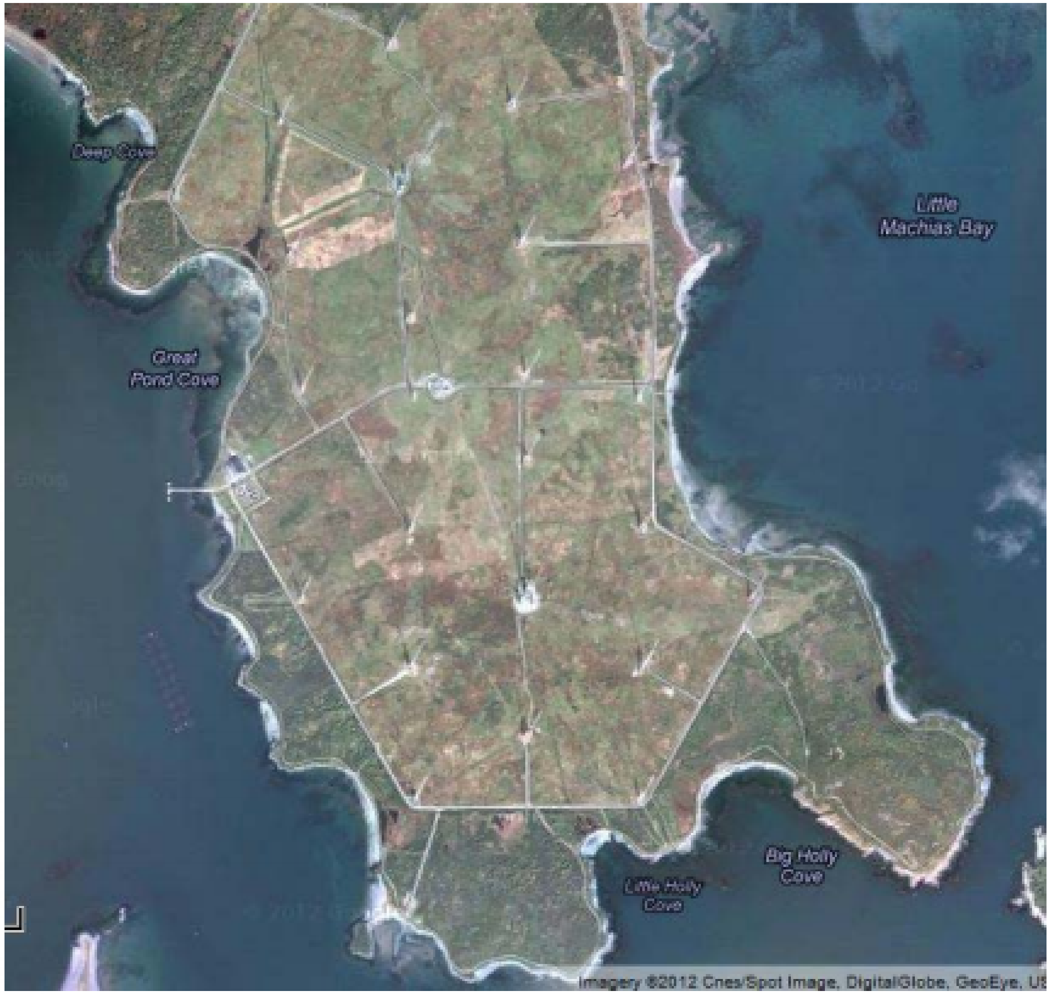
Deep Cove
Little Machias Bay
Great Pond Cove
Little Holly Cove
Big Holly Cove
Imagery ©2012 Cnes/Spot Image, DigitalGlobe, GeoEye, US

26 TOWERS- 850 to 1000 FT HIGH

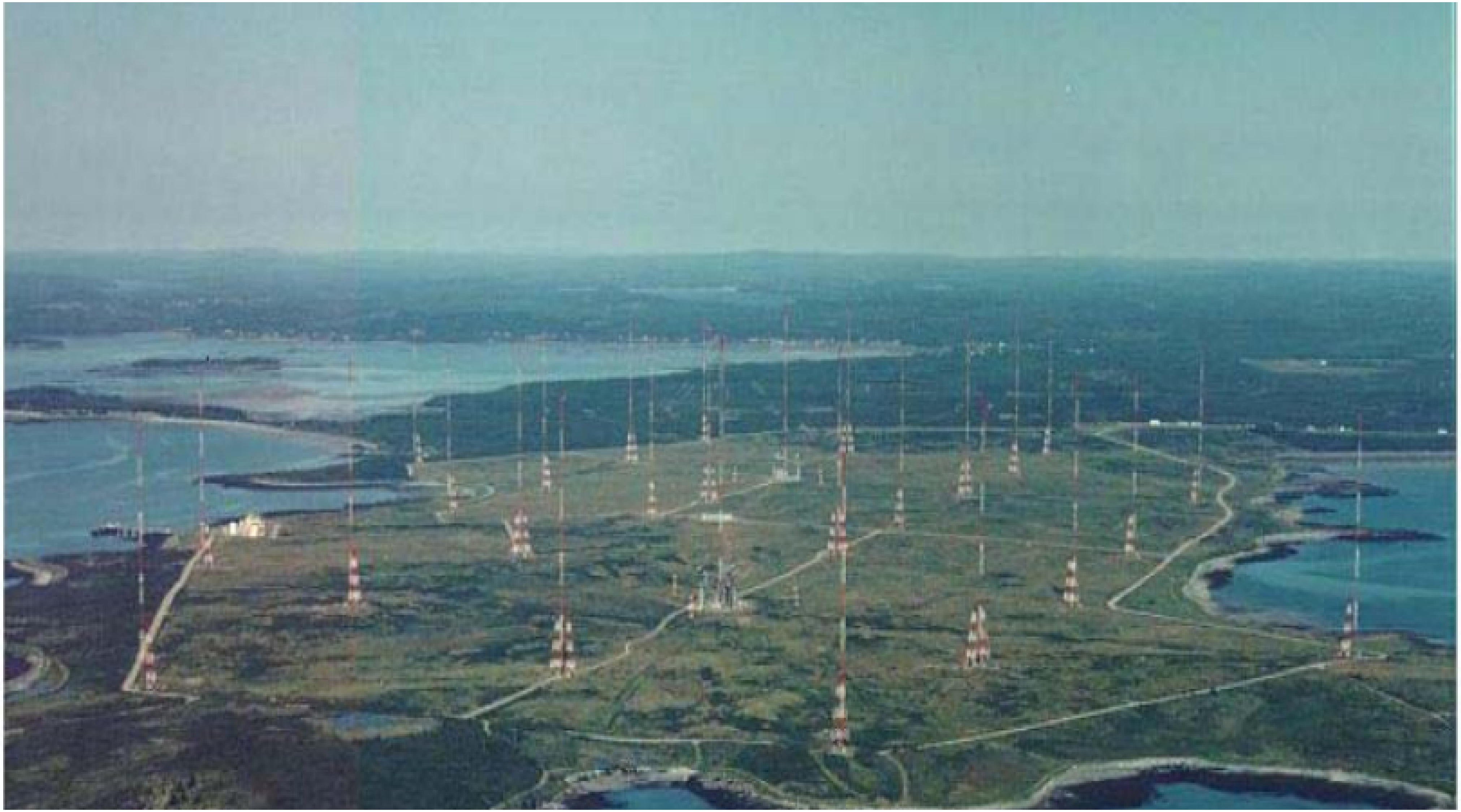

EACH INSULATOR IS 57 FT LONG TO WITHSTAND 250 KV

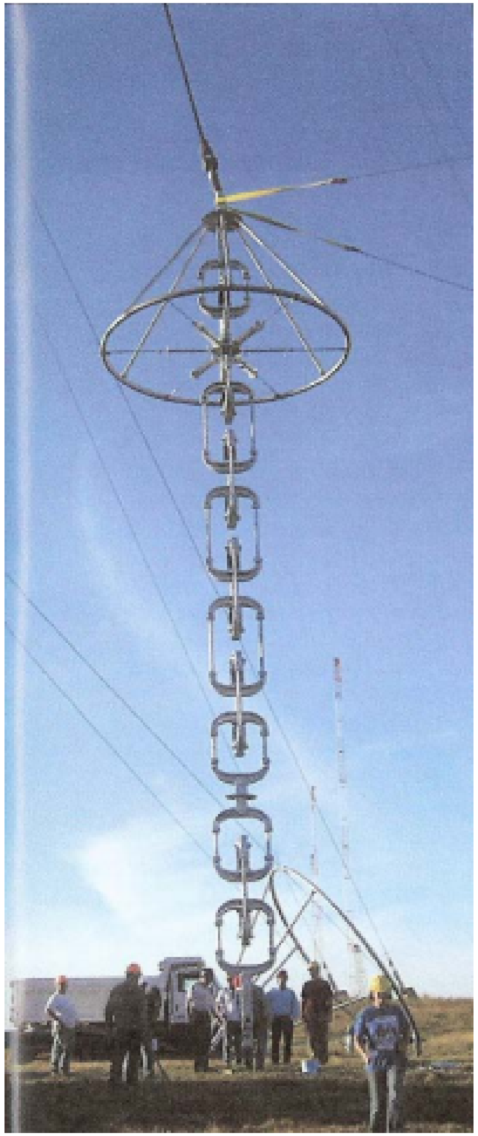

13,000 lbs.

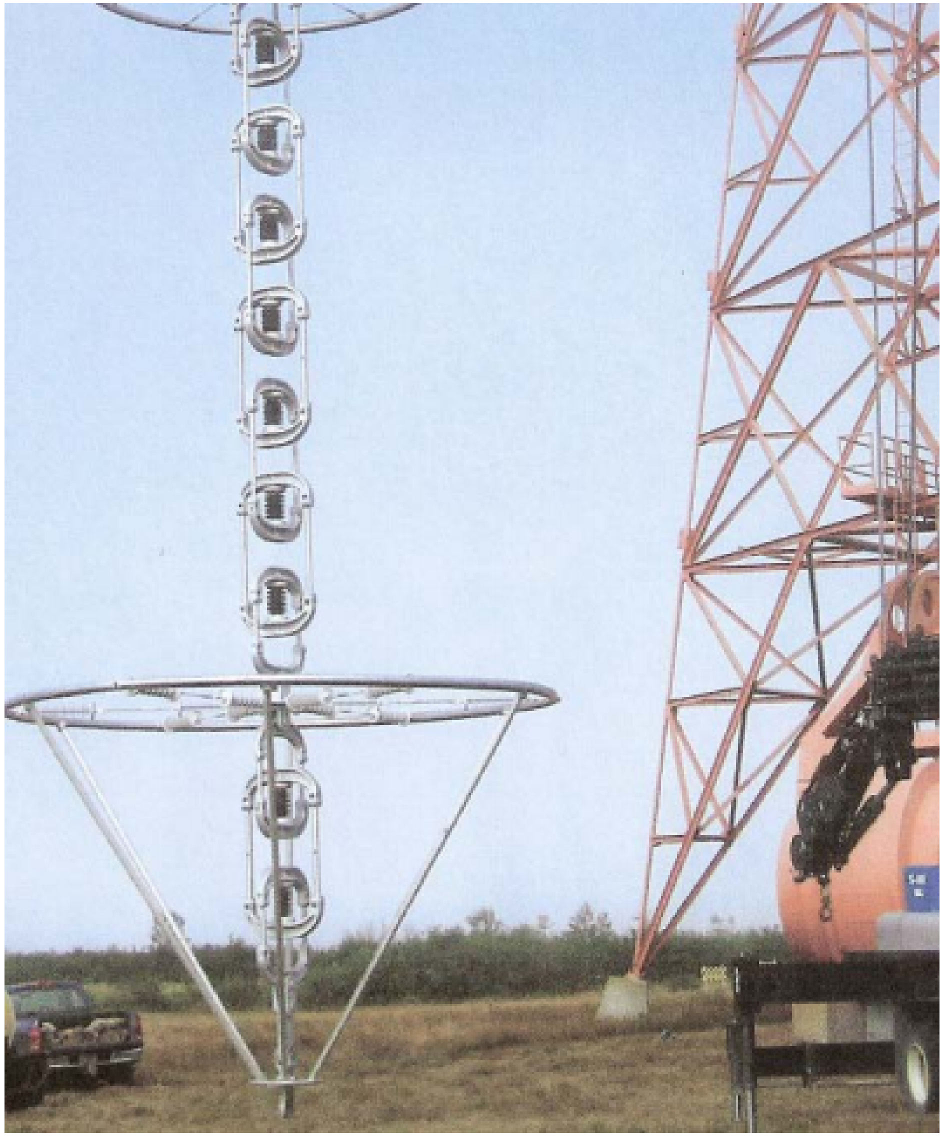

TUNING NETWORK- HELIX

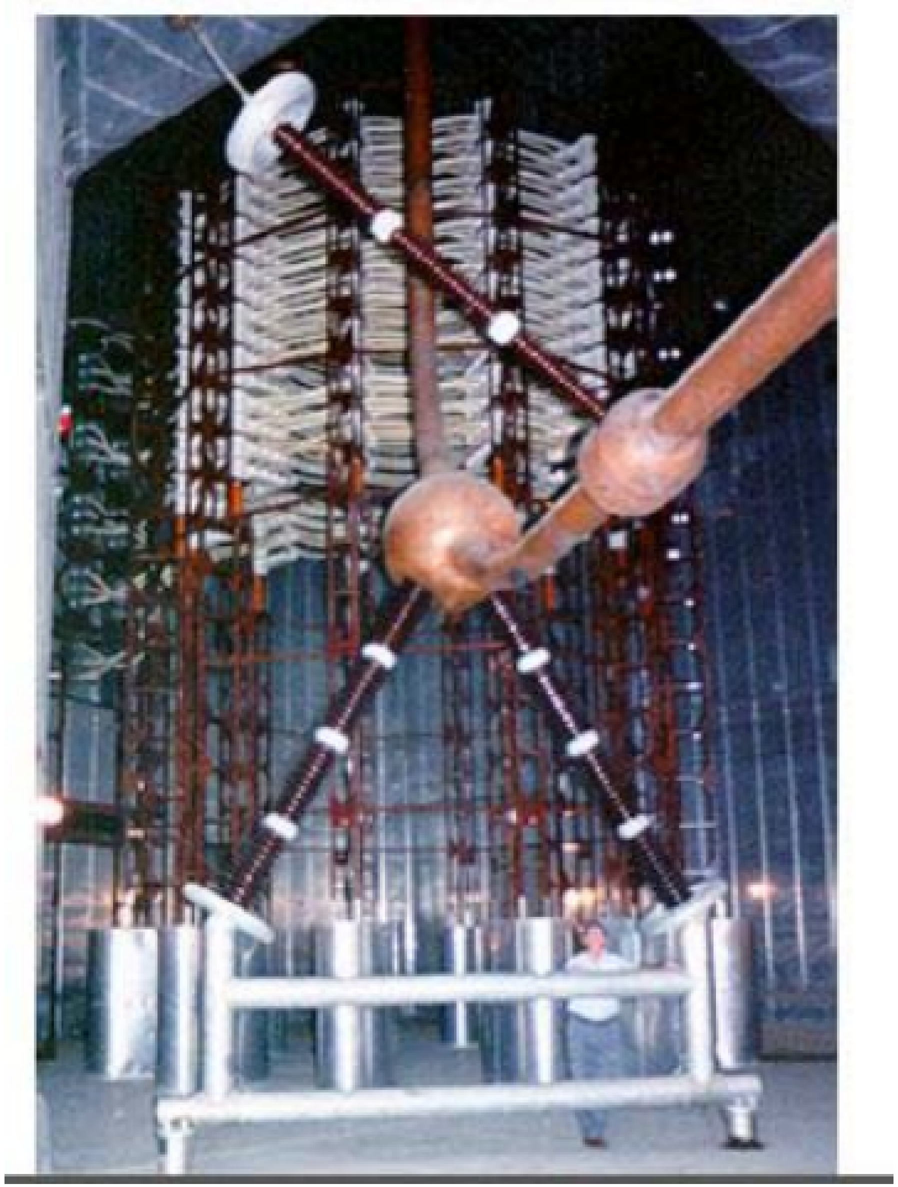

Ref 8

TUNING NETWORK-VARIOMETER

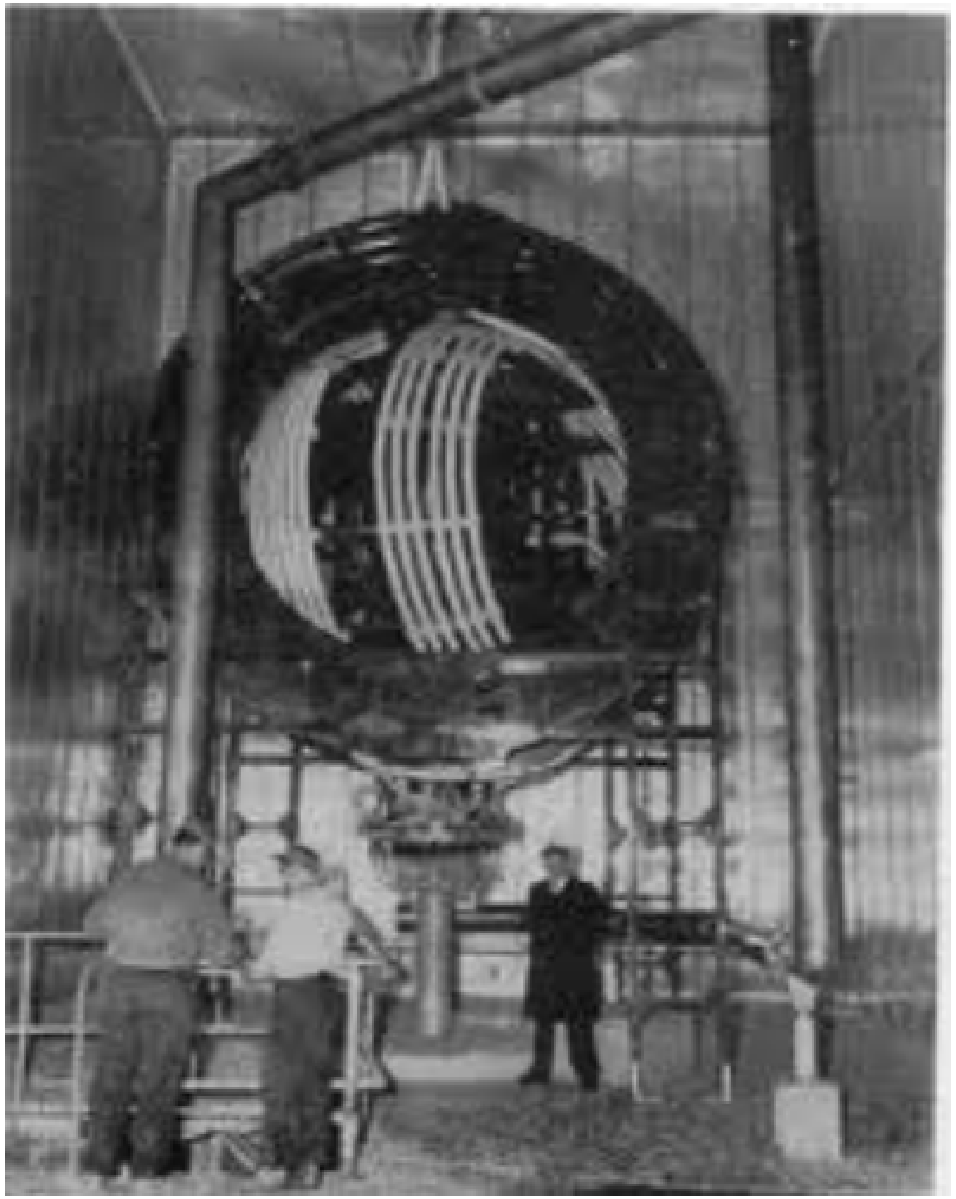

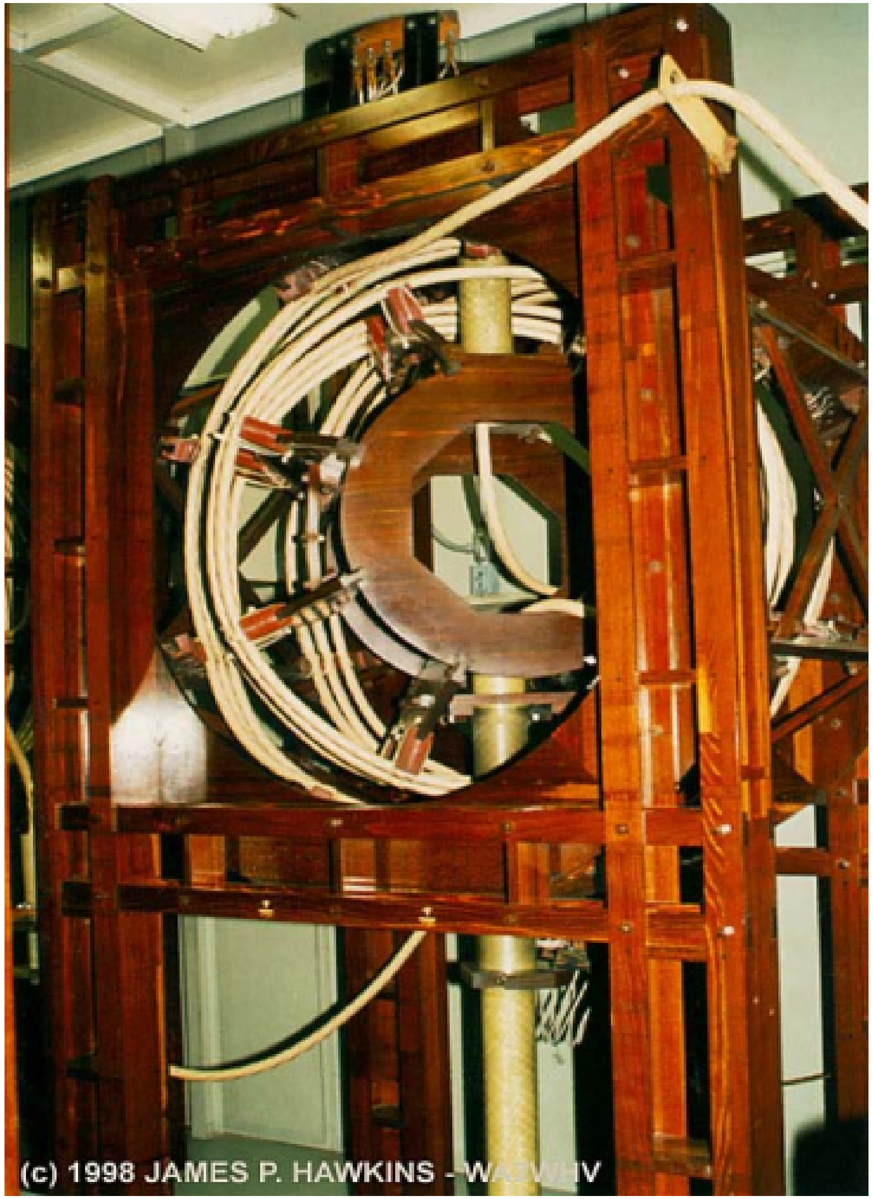

(c) 1998 JAMES P. HAWKINS - WA2WHV

Ref 5 NAA Wires are 4 inches diameter NSS JP Hawkins

SUBMARINE RADIO RECEIVERS

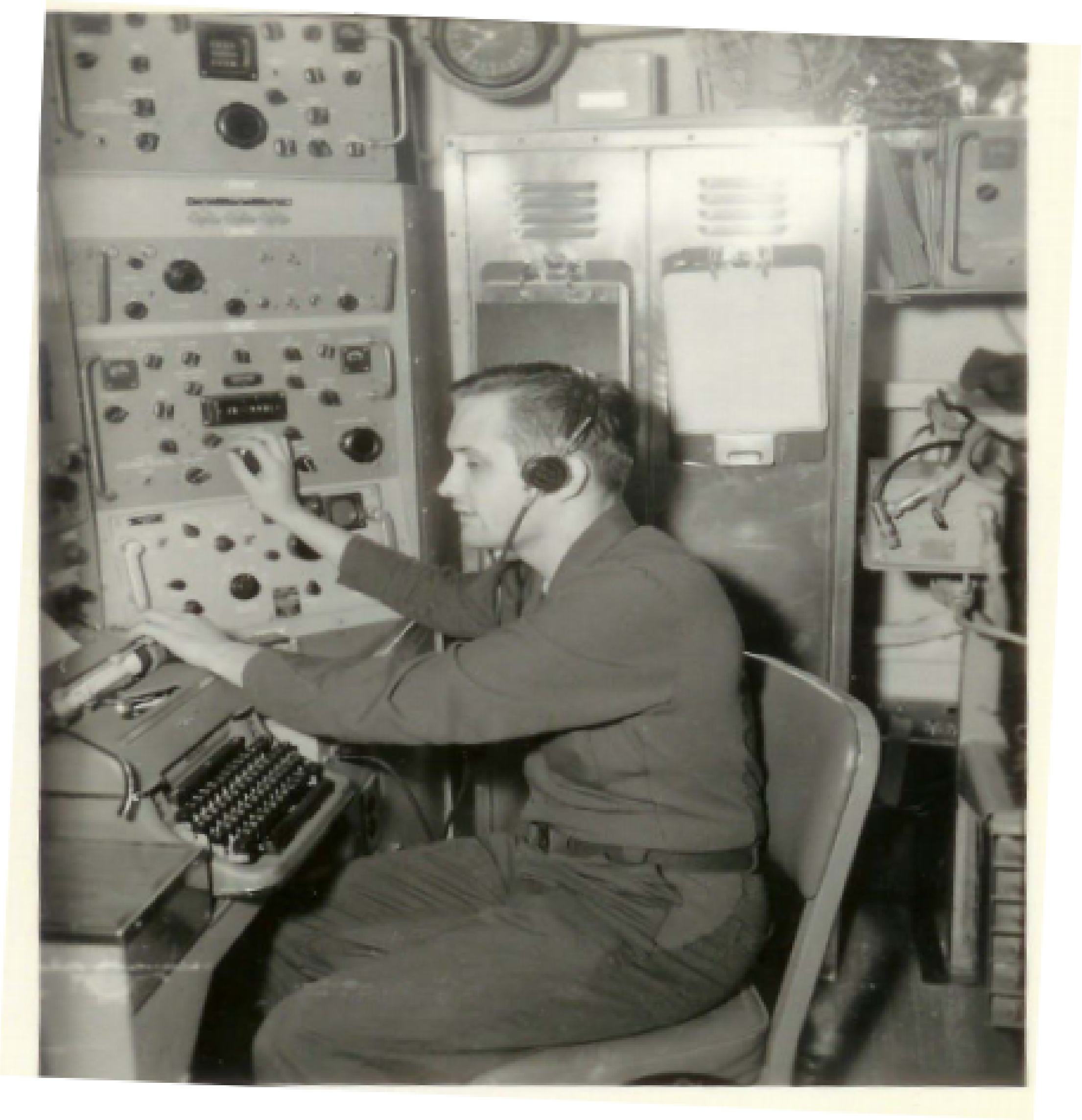

USS Robert E Lee 1966

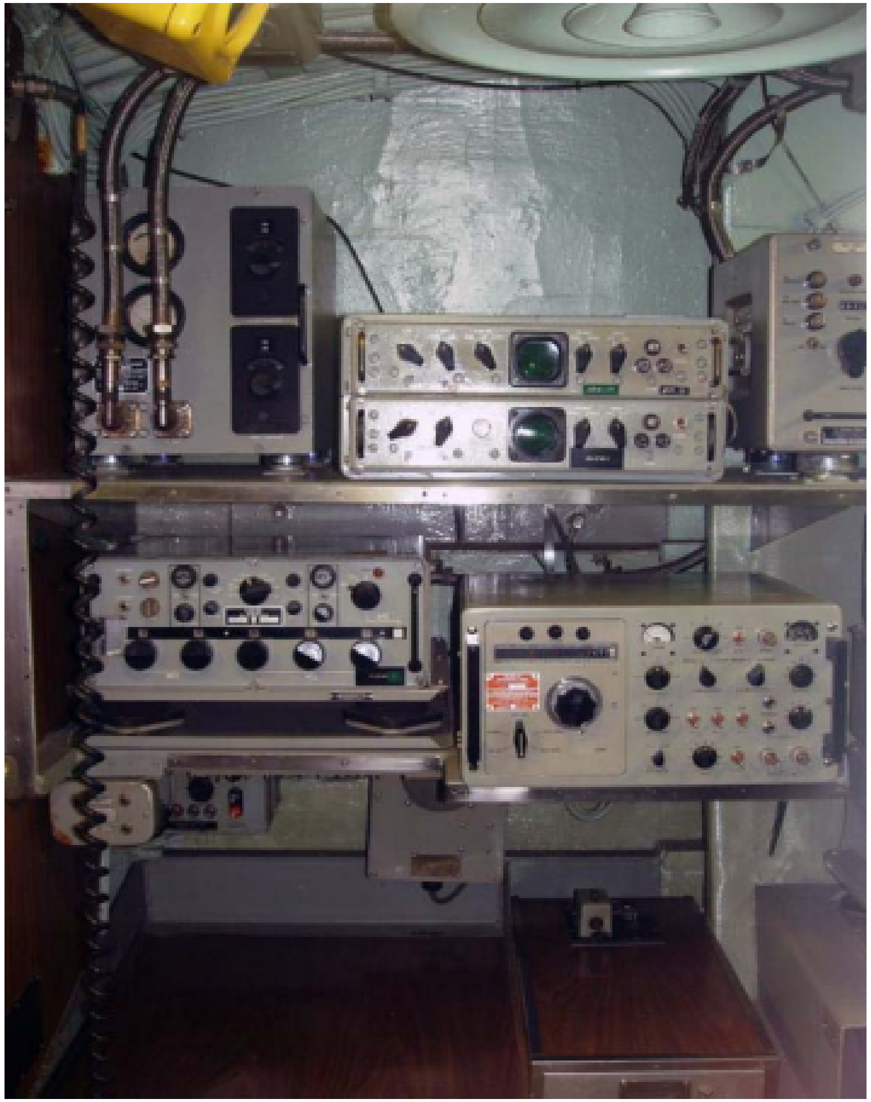

USS Nautilus 1970s

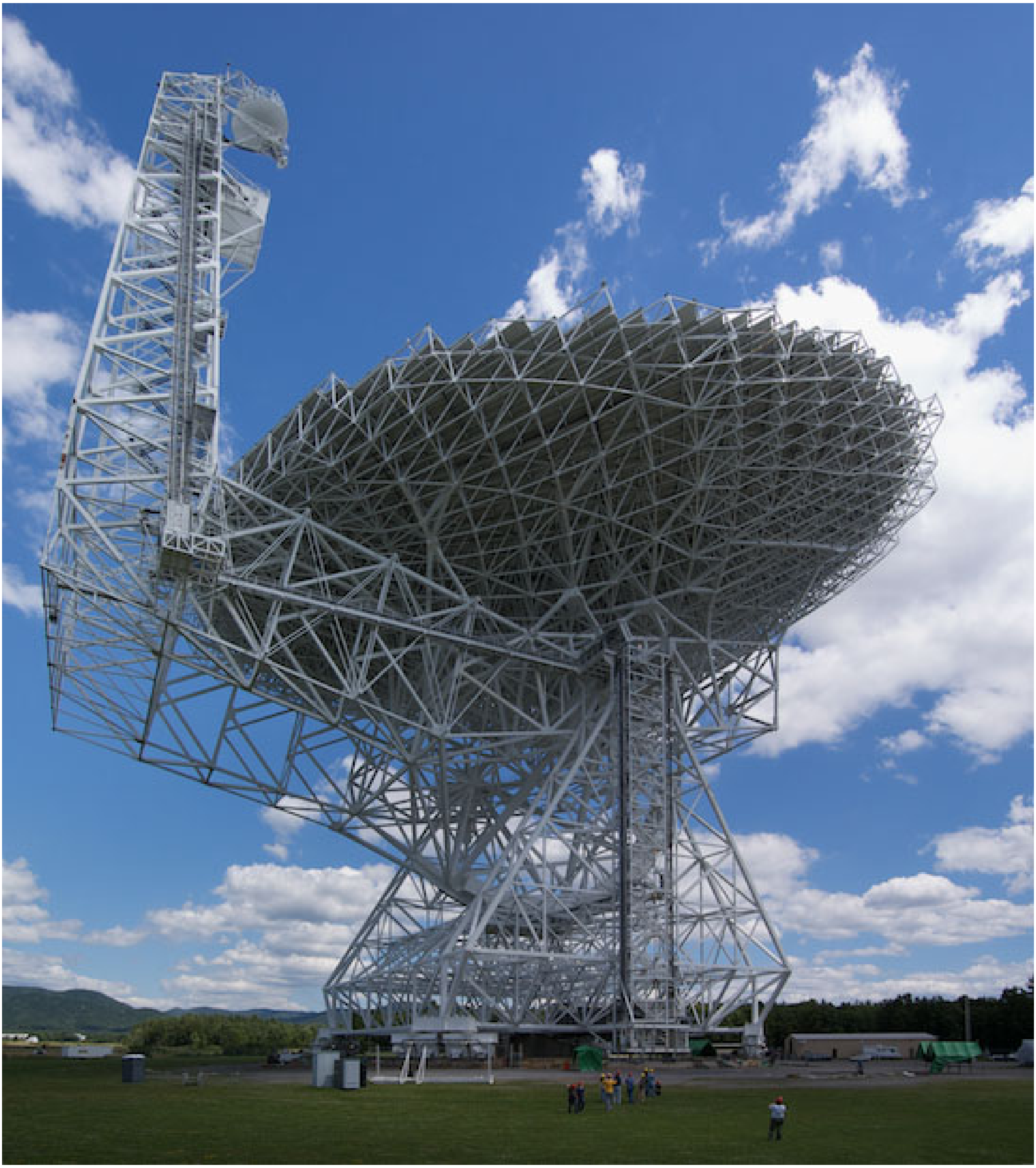

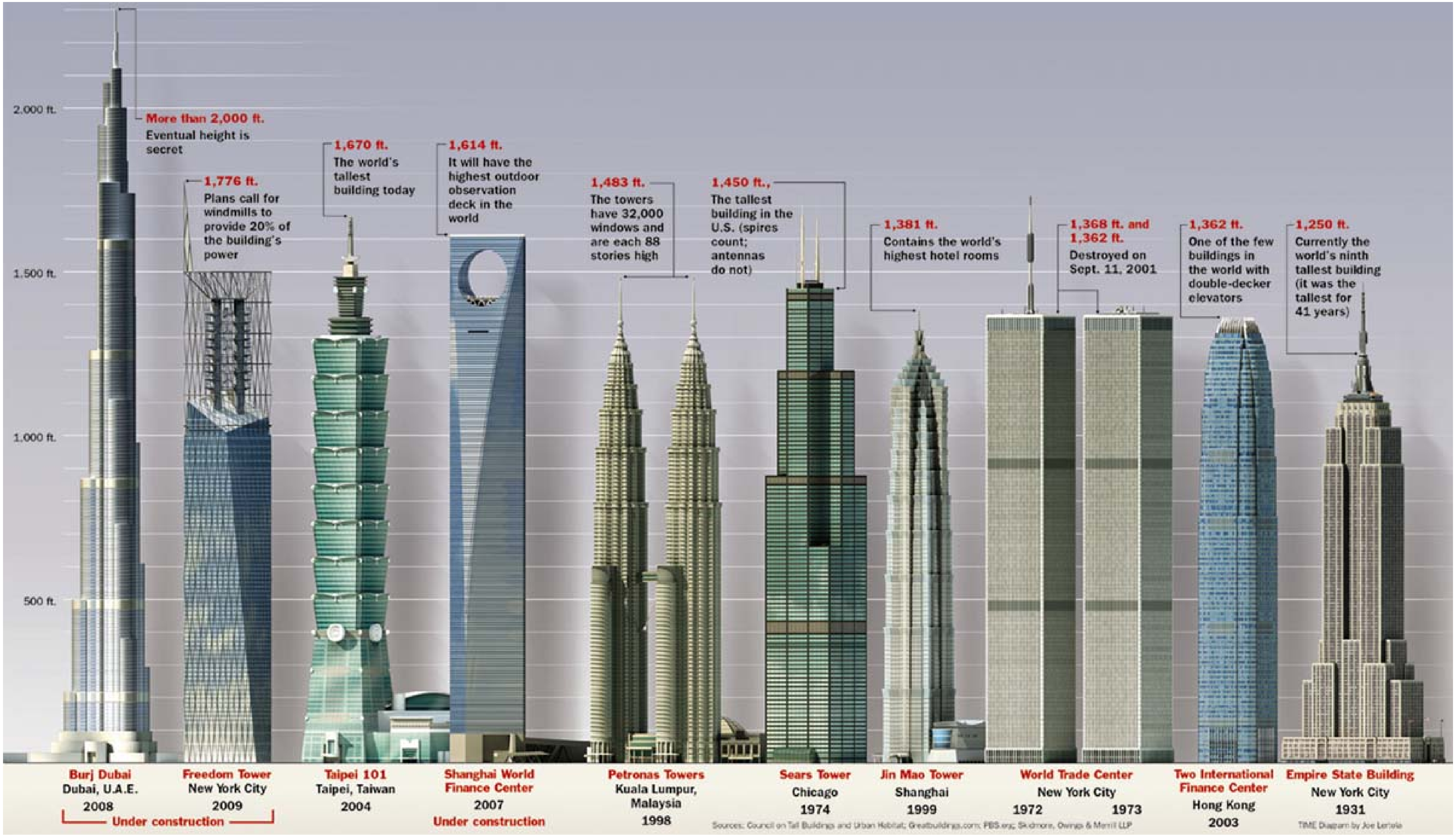

Sources: Council on Tall Buildings and Urban Habitat; Greatbuildings.com; PBS.org; Skidmore, Owings & Merrill LLP

TIME Diagram by Joe Lertola

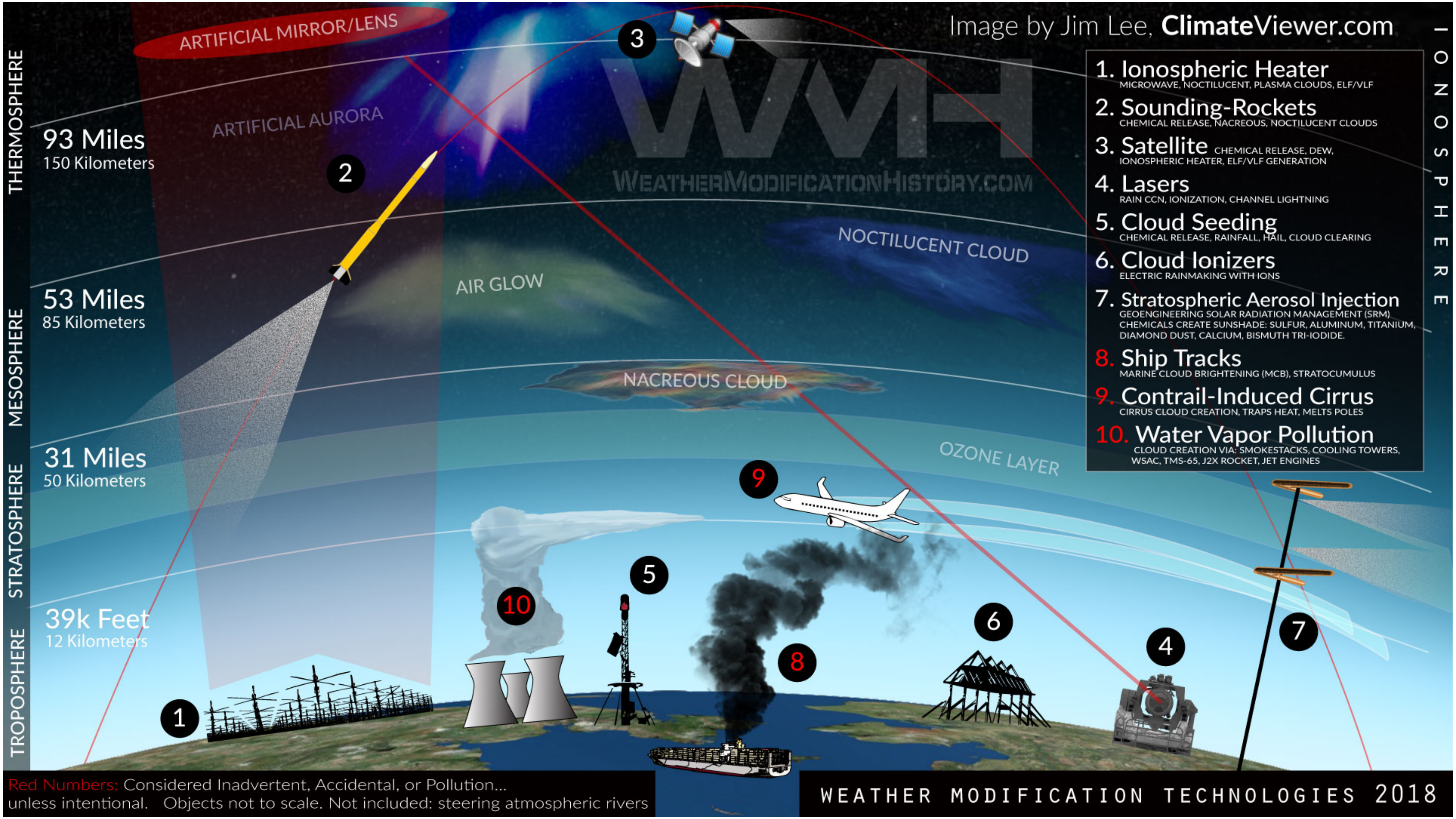
THERMOSPHERE
MESOSPHERE
STRATOSPHERE
TROPOSPHERE
IONOSPHERE
93 Miles
150 Kilometers
53 Miles
85 Kilometers
31 Miles
50 Kilometers
39k Feet
12 Kilometers
ARTIFICIAL MIRROR/LENS
ARTIFICIAL AURORA
AIR GLOW
NOCTILUCENT CLOUD
NACREOUS CLOUD
OZONE LAYER
WMH
WEATHERMODIFICATIONHISTORY.COM
1
2
3
4
5
6
7
8
9
10

Image by Jim Lee, **ClimateViewer.com**

1. **Ionospheric Heater**
   MICROWAVE, NOCTILUCENT, PLASMA CLOUDS, ELF/VLF
2. **Sounding-Rockets**
   CHEMICAL RELEASE, NACREOUS, NOCTILUCENT CLOUDS
3. **Satellite** CHEMICAL RELEASE, DEW, IONOSPHERIC HEATER, ELF/VLF GENERATION
4. **Lasers**
   RAIN CCN, IONIZATION, CHANNEL LIGHTNING
5. **Cloud Seeding**
   CHEMICAL RELEASE, RAINFALL, HAIL, CLOUD CLEARING
6. **Cloud Ionizers**
   ELECTRIC RAINMAKING WITH IONS
7. **Stratospheric Aerosol Injection**
   GEOENGINEERING SOLAR RADIATION MANAGEMENT (SRM) CHEMICALS CREATE SUNSHADE: SULFUR, ALUMINUM, TITANIUM, DIAMOND DUST, CALCIUM, BISMUTH TRI-IODIDE.
8. **Ship Tracks**
   MARINE CLOUD BRIGHTENING (MCB), STRATOCUMULUS
9. **Contrail-Induced Cirrus**
   CIRRUS CLOUD CREATION, TRAPS HEAT, MELTS POLES
10. **Water Vapor Pollution**
    CLOUD CREATION VIA: SMOKESTACKS, COOLING TOWERS, WSAC, TMS-65, J2X ROCKET, JET ENGINES

Red Numbers: Considered Inadvertent, Accidental, or Pollution... unless intentional. Objects not to scale. Not included: steering atmospheric rivers

WEATHER MODIFICATION TECHNOLOGIES 2018

HAARP

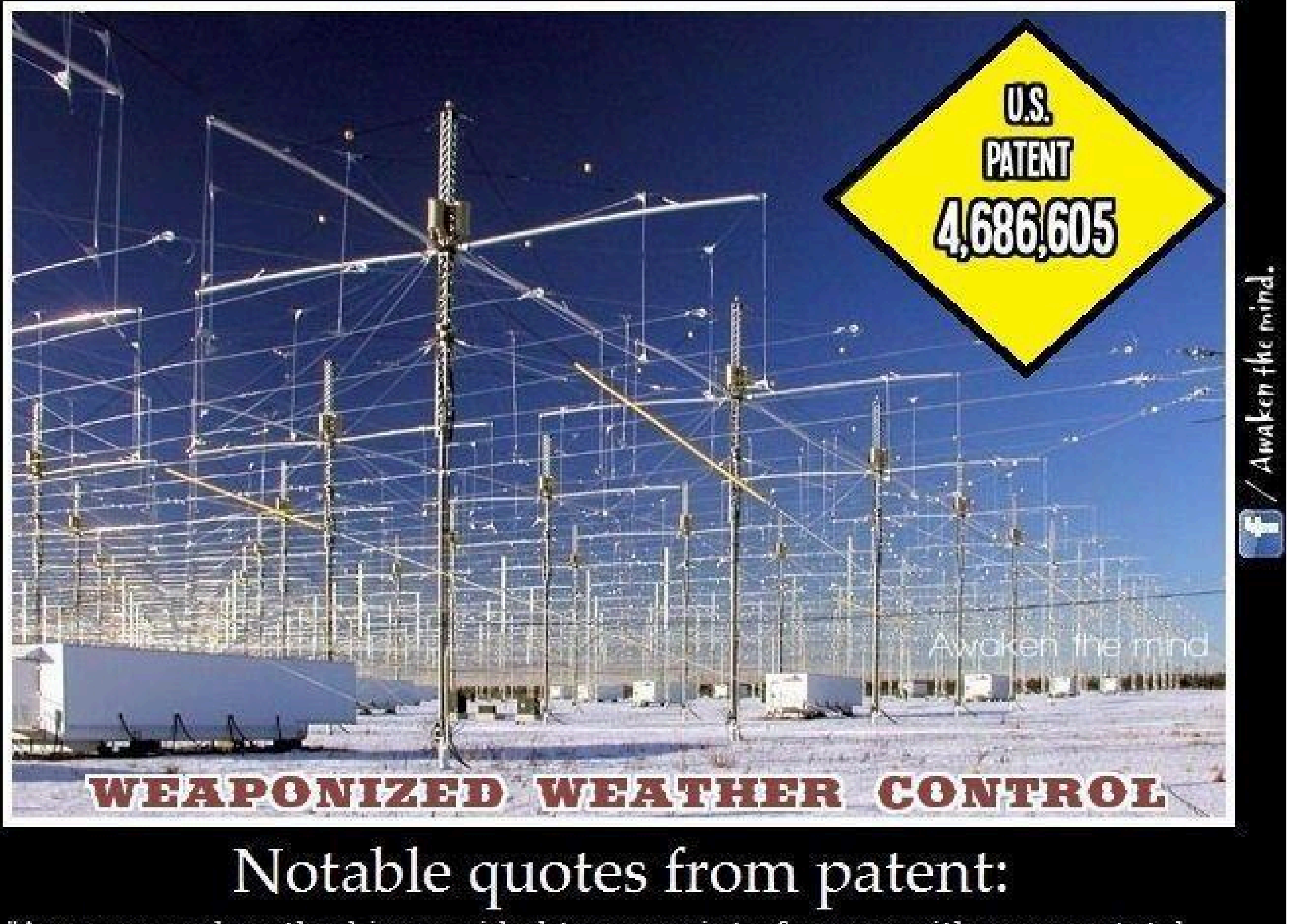
U.S.
PATENT
4,686,605
Awaken the mind
/ Awaken the mind.
WEAPONIZED WEATHER CONTROL

Notable quotes from patent:

- "A means and method is provided to cause interference with or even total disruption of communications over a very large portion of the earth"

- "Weather modification is possible by, for example, altering upper atmosphere wind patterns or altering solar absorption patterns by constructing one or more plumes of atmospheric particles which will act as a lens or focusing device."

- "The earth's magnetic field could be decreased or disrupted at appropriate altitudes to modify or eliminate the magnetic field."

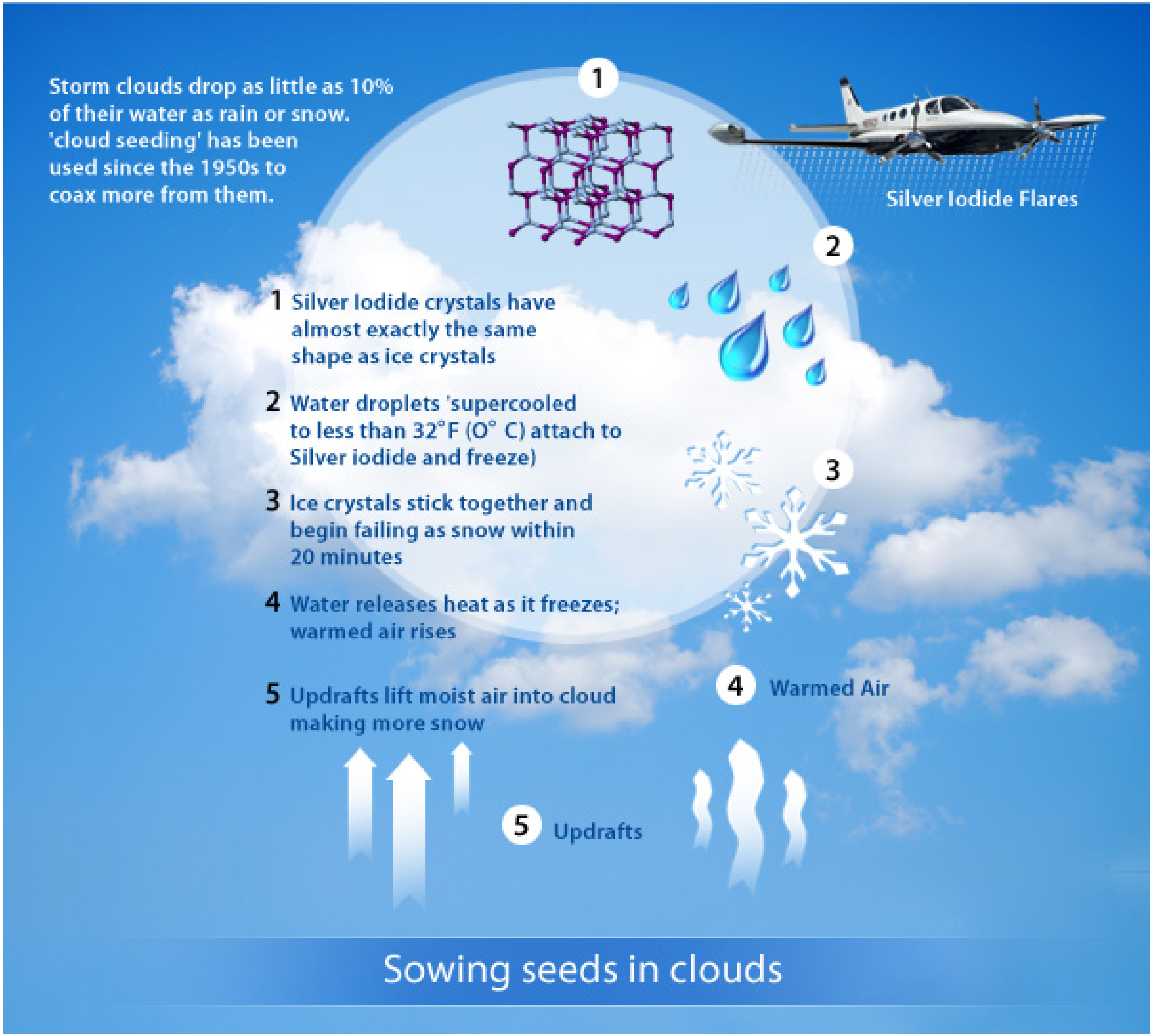

Storm clouds drop as little as 10% of their water as rain or snow. 'cloud seeding' has been used since the 1950s to coax more from them.

1 Silver Iodide crystals have almost exactly the same shape as ice crystals

2 Water droplets 'supercooled to less than 32°F (O° C) attach to Silver iodide and freeze)

3 Ice crystals stick together and begin failing as snow within 20 minutes

4 Water releases heat as it freezes; warmed air rises

5 Updrafts lift moist air into cloud making more snow

Sowing seeds in clouds

Radio a galena FM

per la “banda commerciale” 88 – 108 MHz

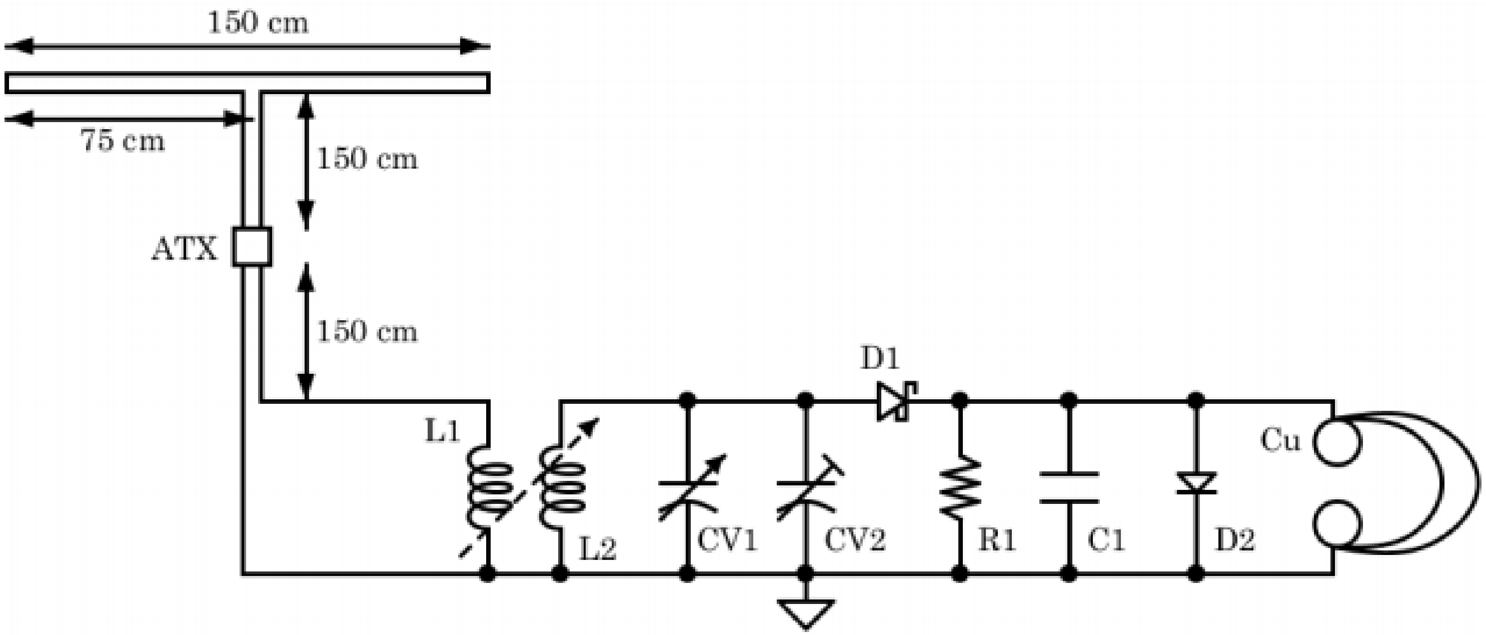
150 cm
75 cm
150 cm
ATX
150 cm
D1
L1
L2
CV1
CV2
R1
C1
D2
Cu

Figura 2: Schema elettrico della radio a galena FM.

I componenti adoperati sono:

- L1 = vedi testo (sezione Bobine L1 e L2);
- L2 = $0.137\mu H$, vedi Figura 7;
- CV1 = Johnson 160-211-1 $(2.7 - 10.8)pF$ per sezione;
- CV2 = trimmer tubolare da $(5 \div 15)pF$;
- D1 = diodo Schottky Skyworks modello SMS7630-001;
- D2 = diodo di segnale 1N4148;
- R1 = $47k\Omega$, $1/4$W;
- C1 = $100pF$ ceramico a disco;
- Cu = cuffie ad alta impedenza ($2k\Omega$ o superiore);
- ATX = connettore ATX femmina e header pin;
- Due connettori banana femmina.

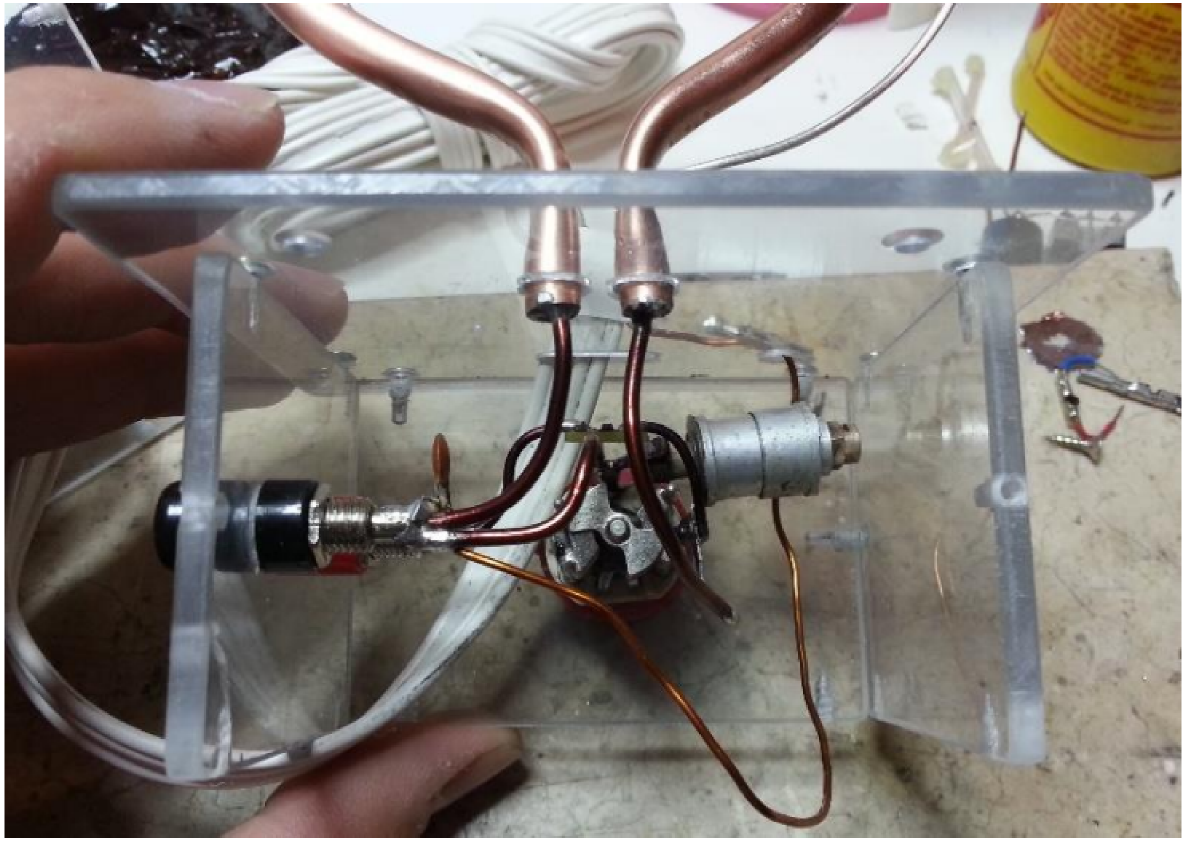

Figura 8: Vista posteriore del cablaggio

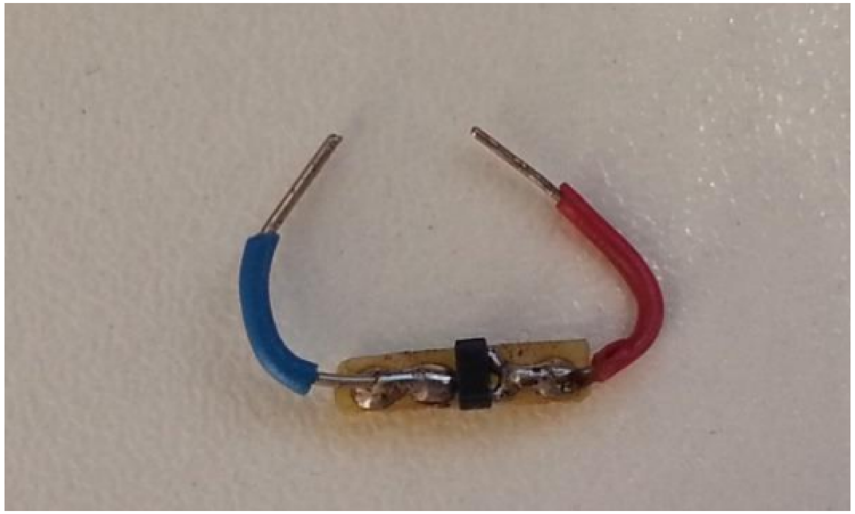

Figura 9: Particolare del diodo sulla basetta millefori.

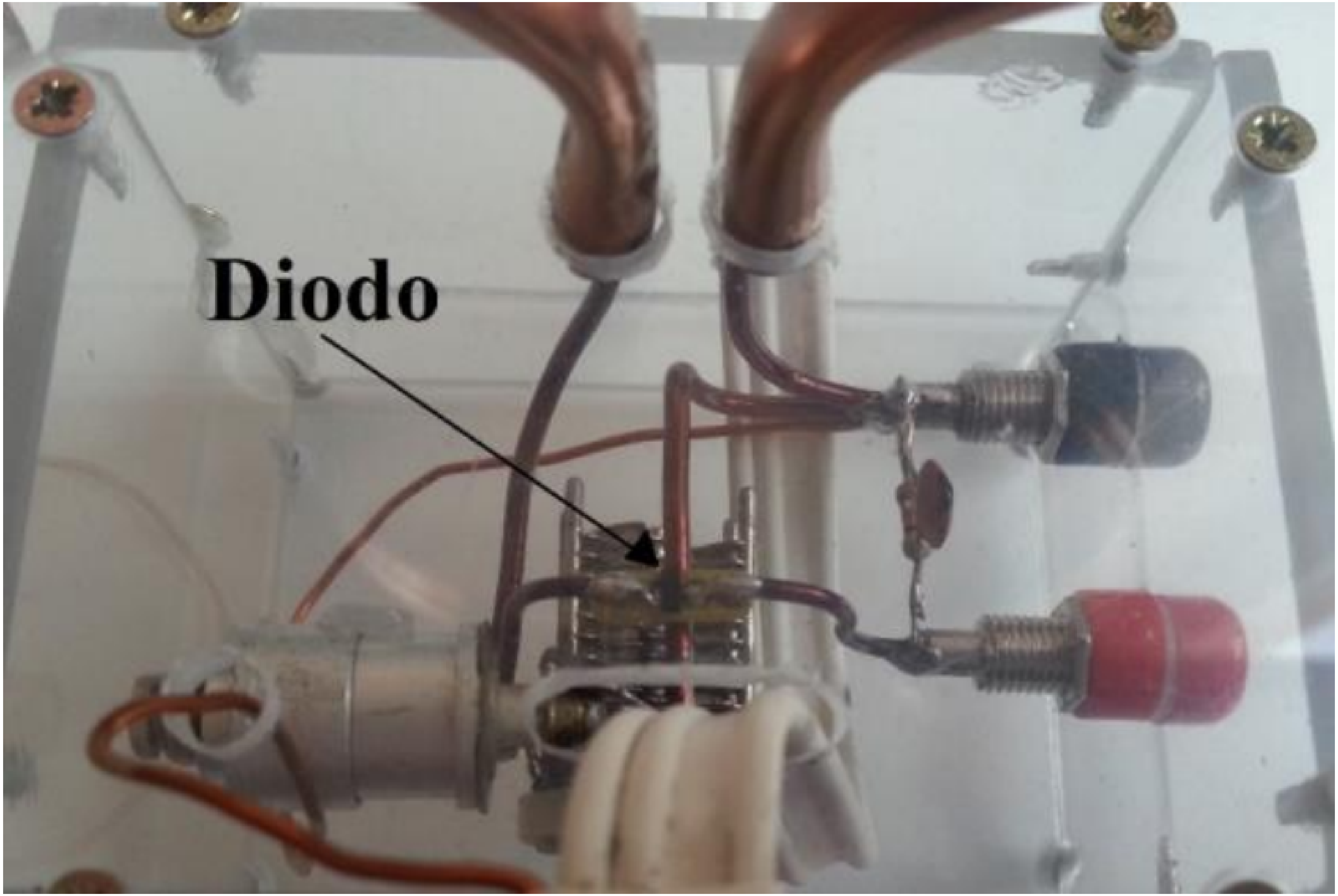

Figura 10: Cablaggio del diodo rivelatore.

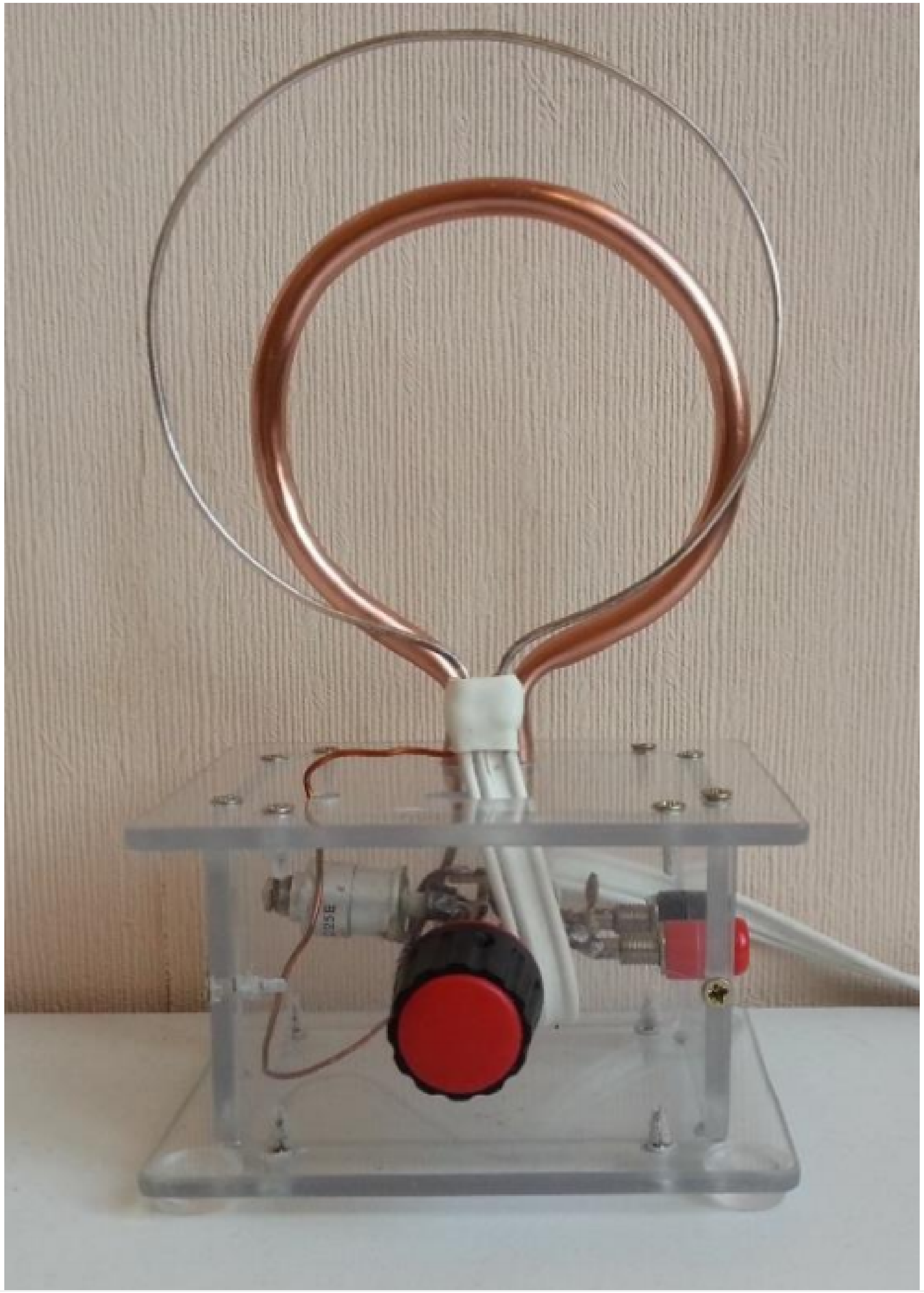

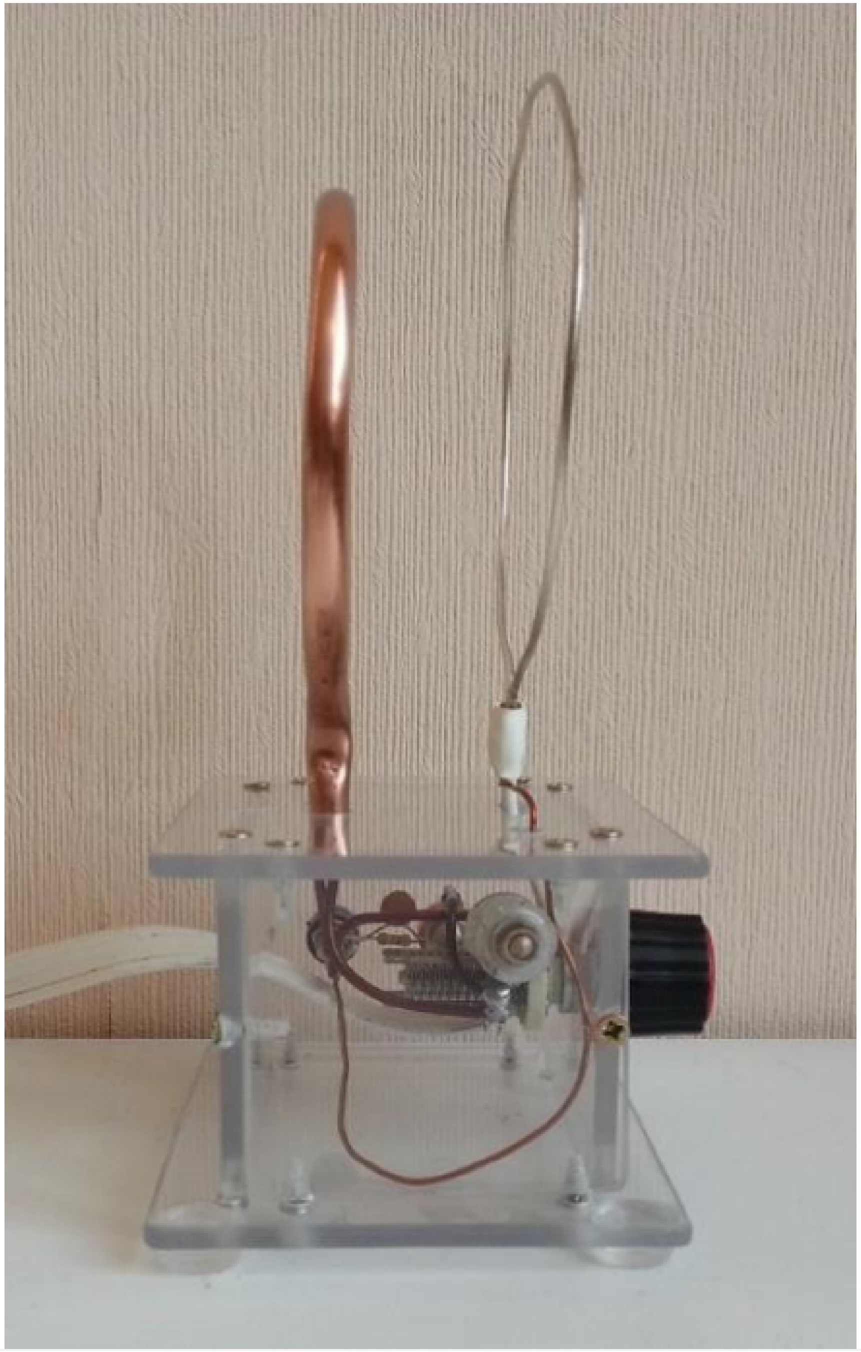

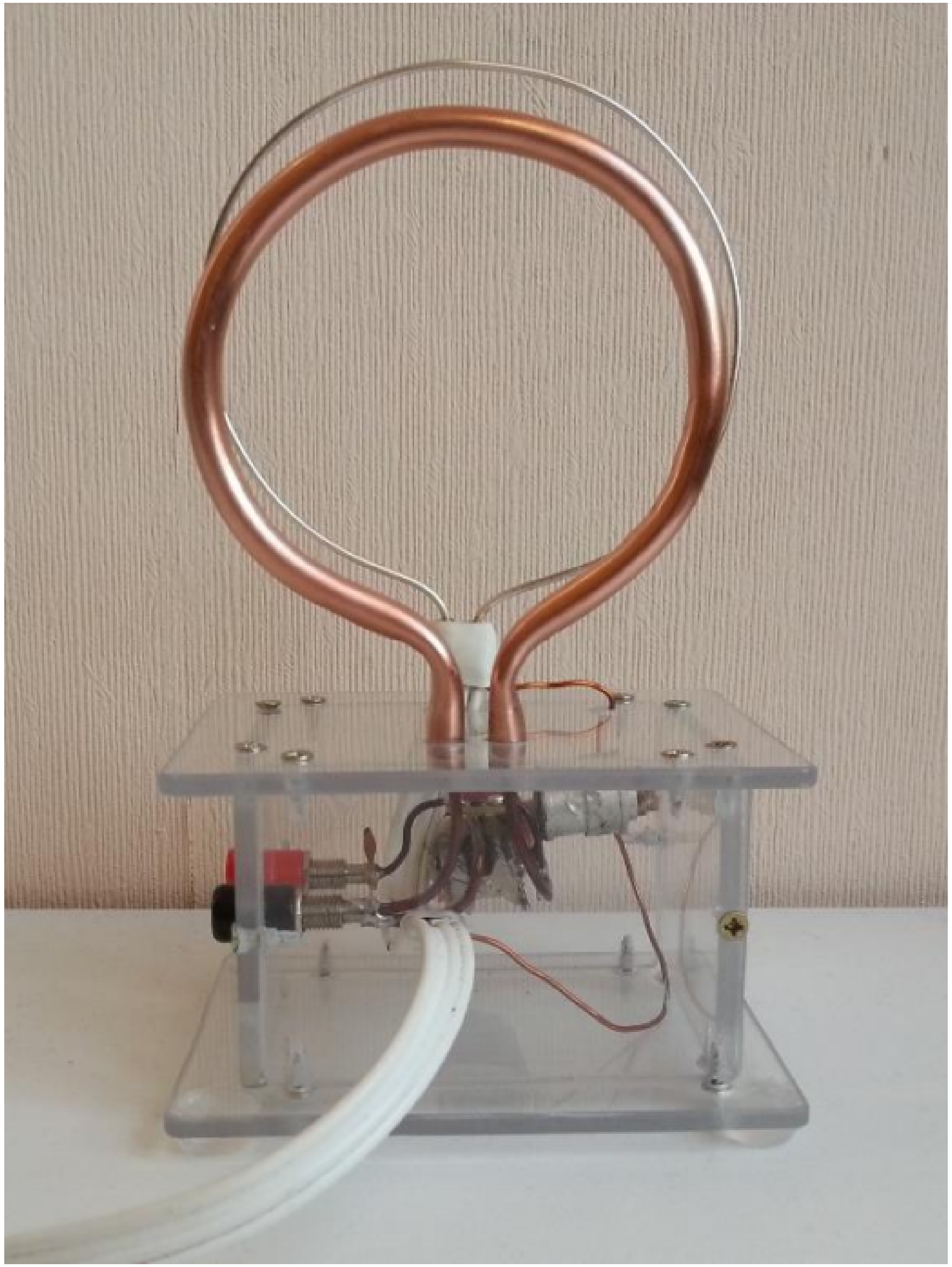

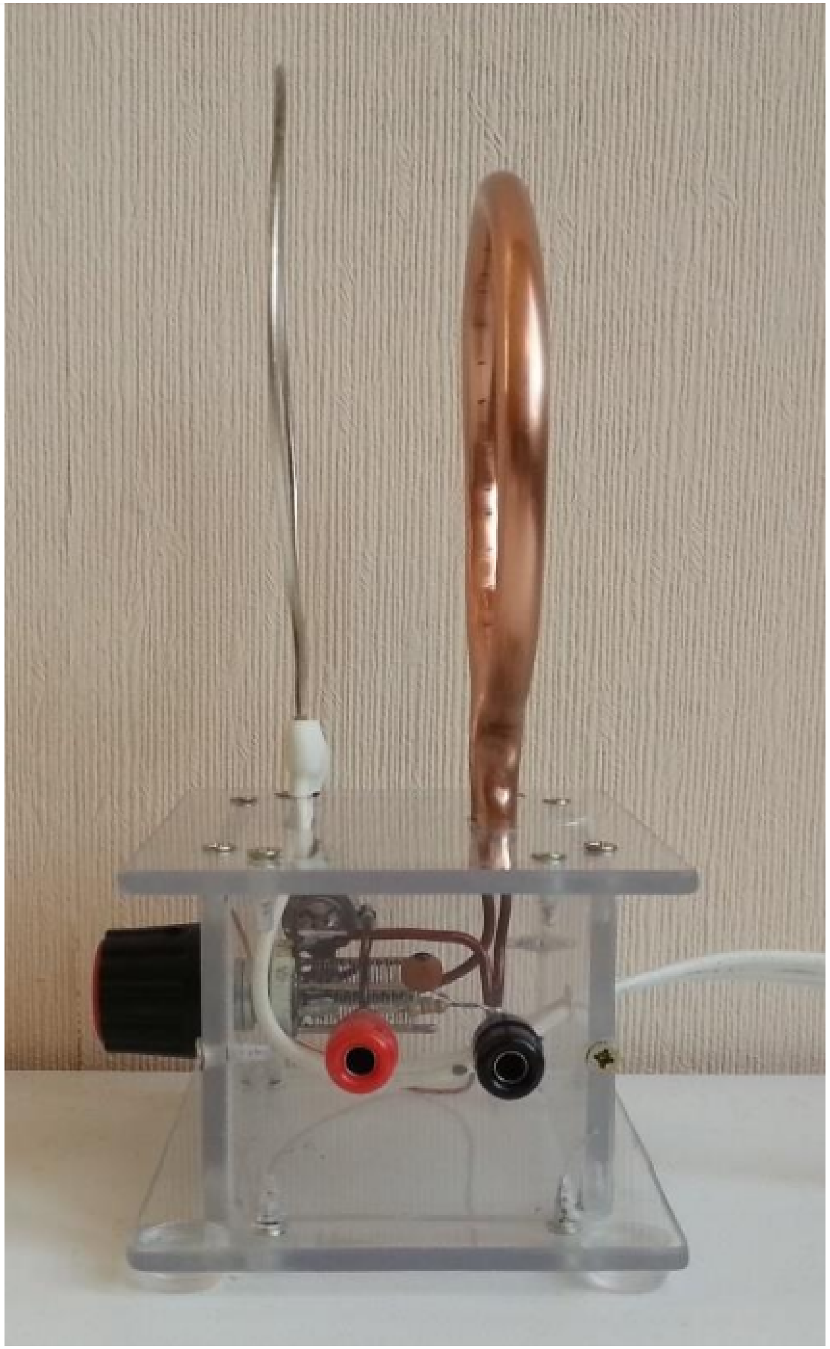

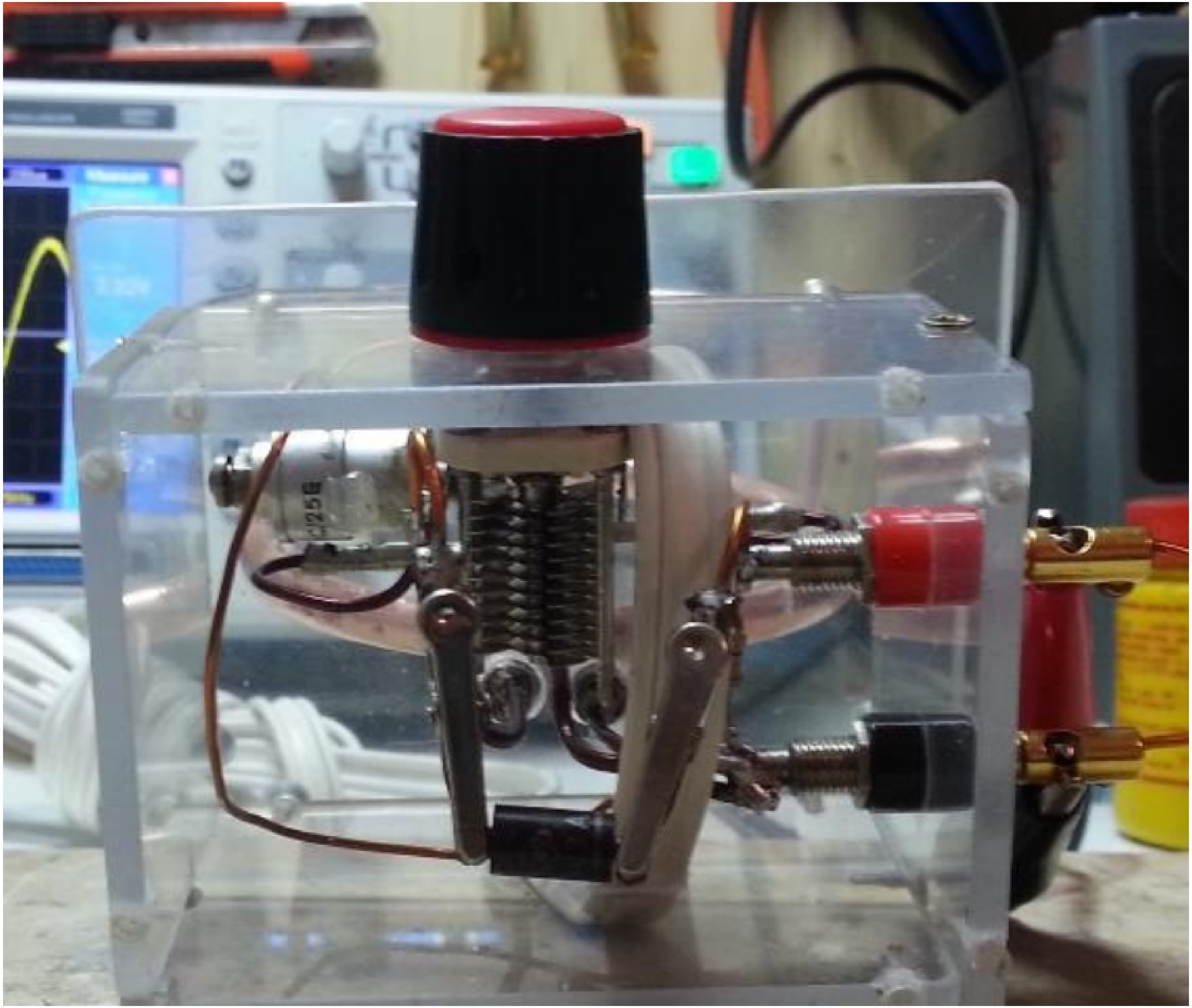

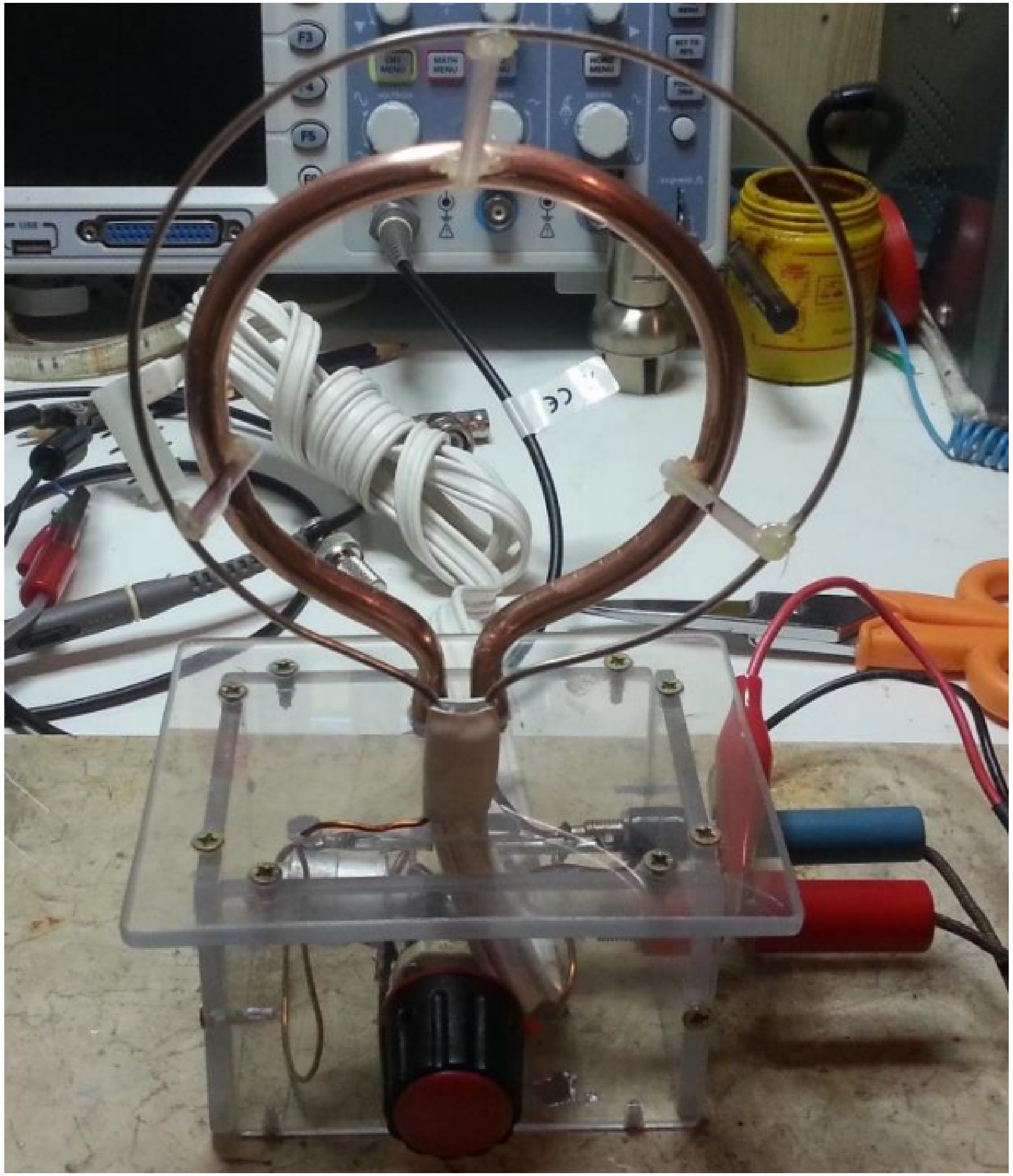
F3
F5
MATH
MENU
USB
CE

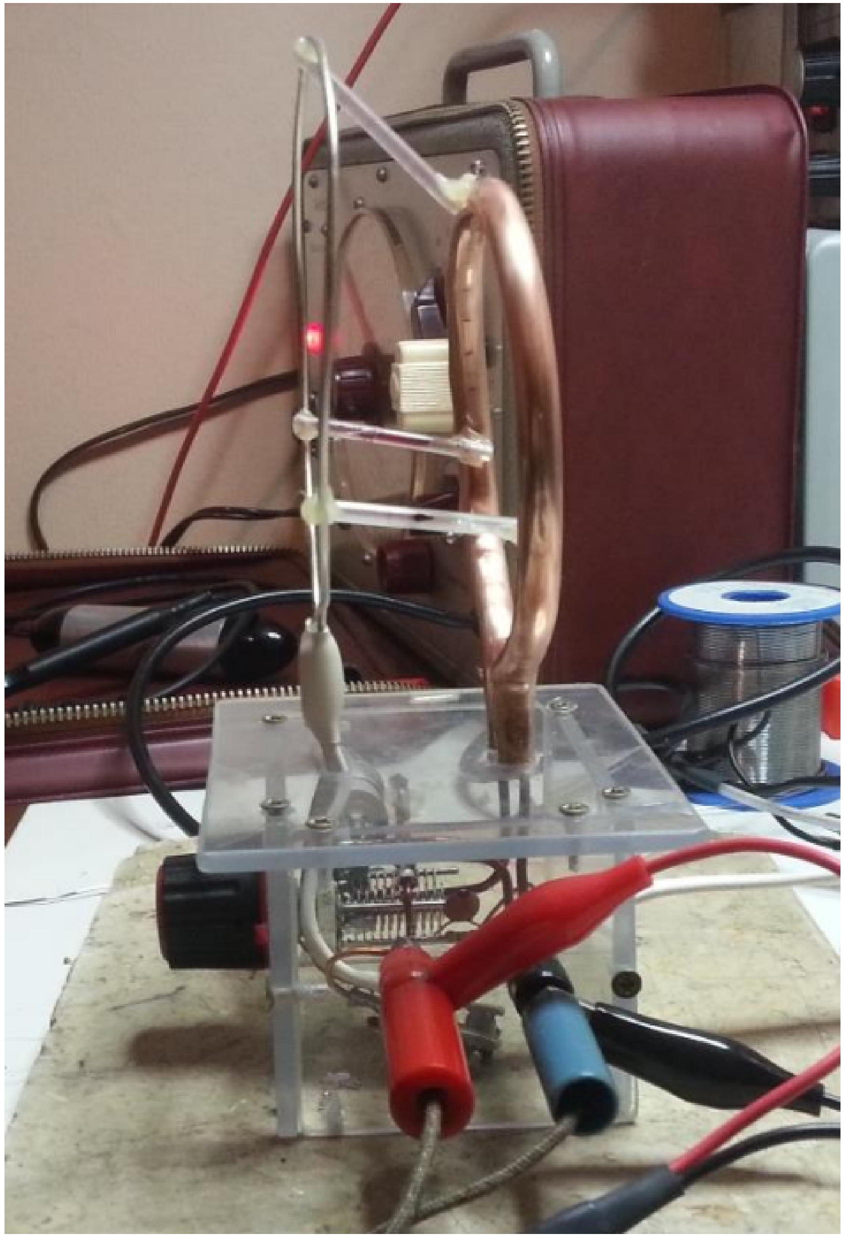

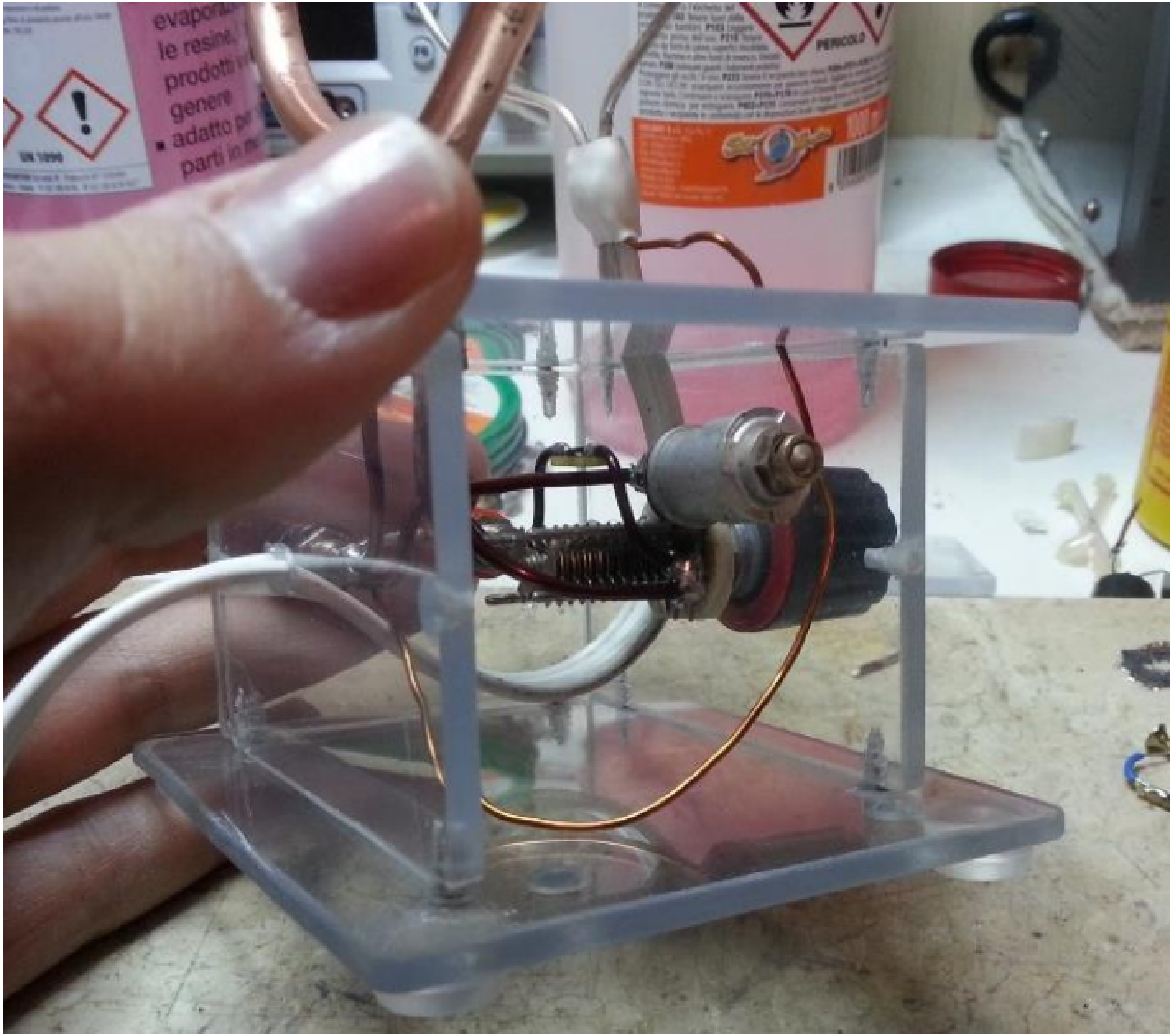
UN 1090
PERICOLO

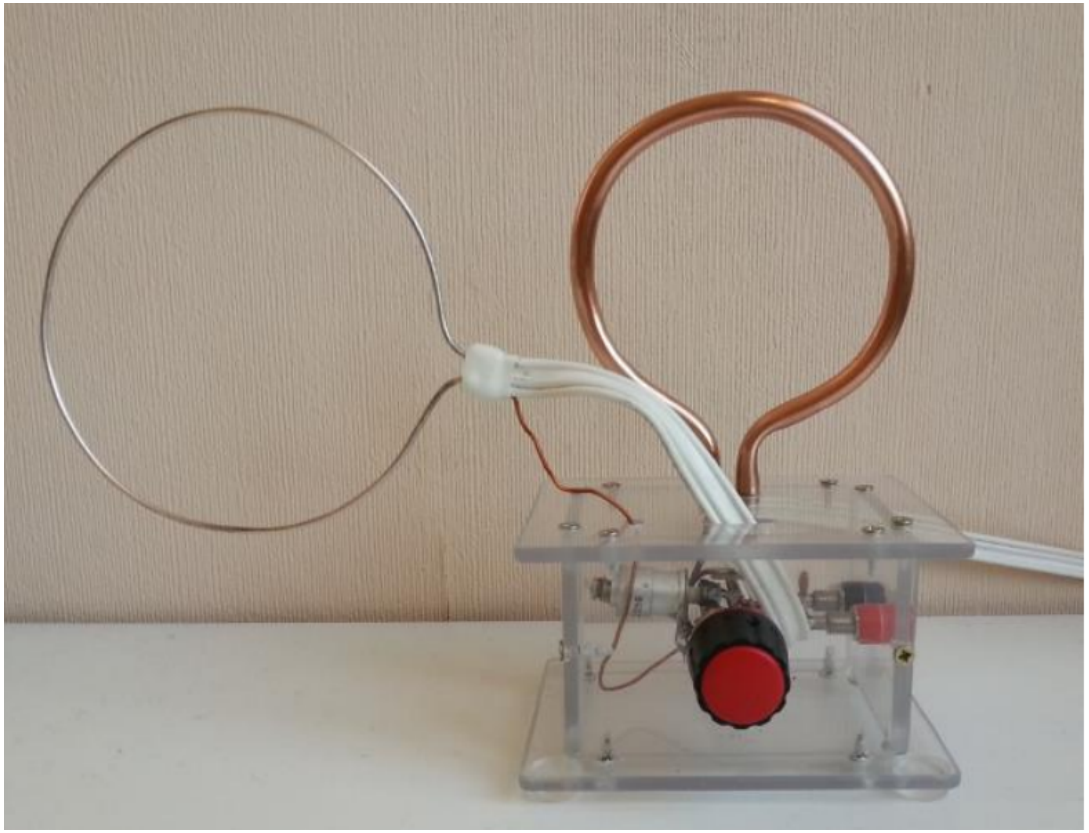

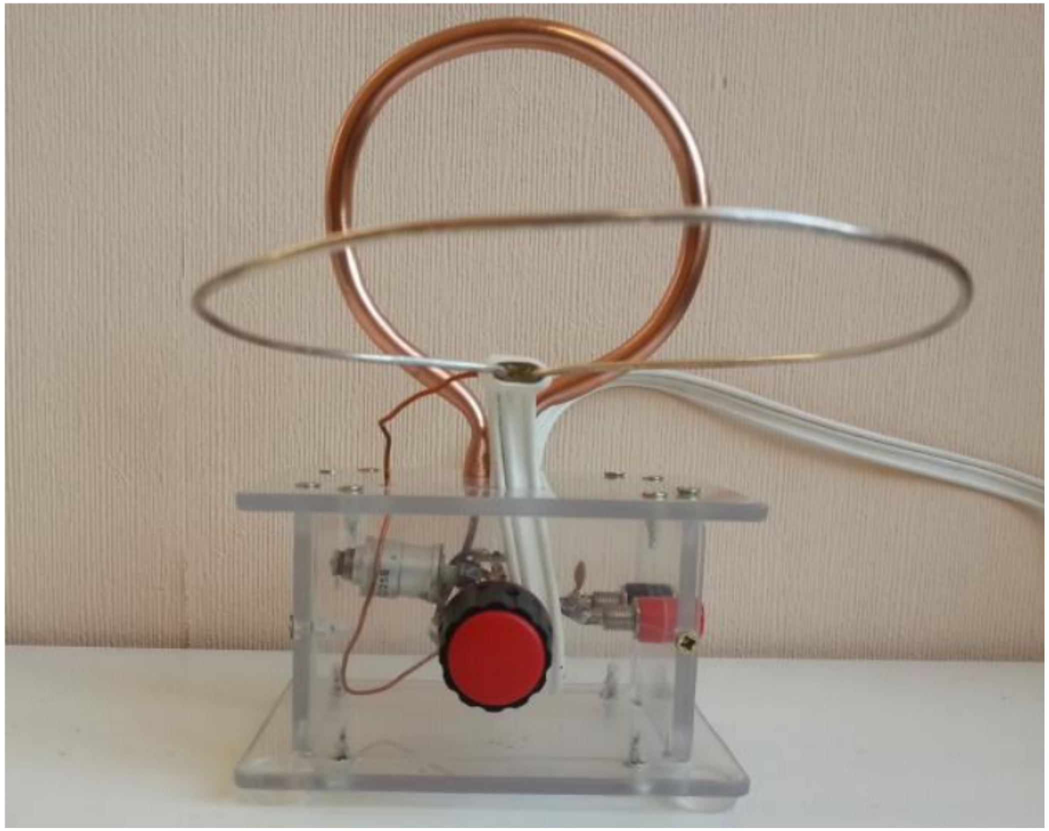

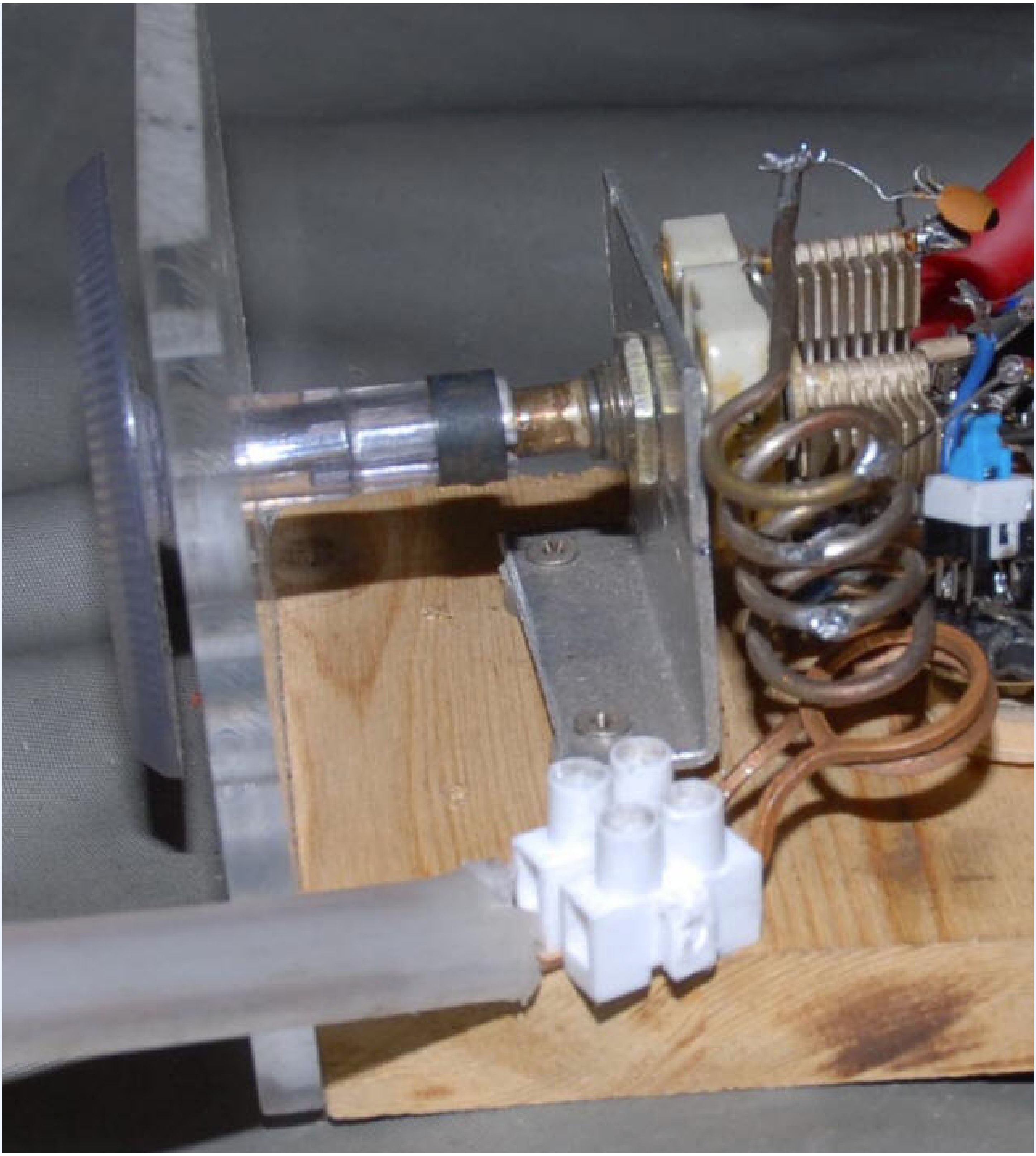

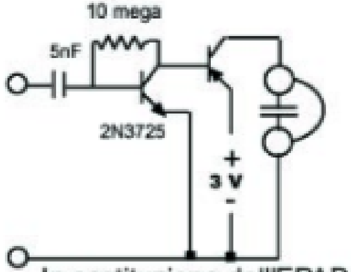
10 mega
5nF
2N3725
+
3 V
-

In sostituzione dell'EPAD: fa
fortissimo ma occorre una pila

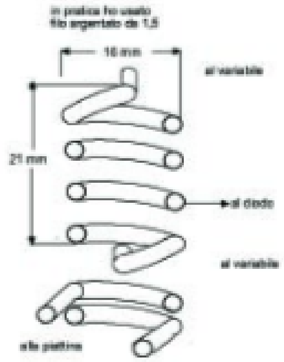
in pratica ho usato
filo argentato da 1,5
16 mm
21 mm
al variabile
al diodo
al variabile
alla piattina

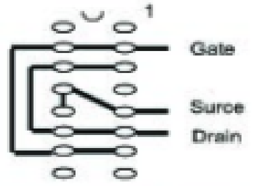
1
Gate
Surce
Drain

ALD110800vista dal sotto

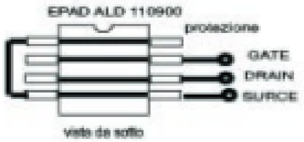
EPAD ALD 110900
protezione
GATE
DRAIN
SURCE
vista da sotto

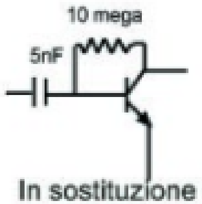

In sostituzione dell'EPAD,
ma richiede almeno 1 V
di alimentazione rubata

NOTA
con un buon segnale
funziona senza amplificatore

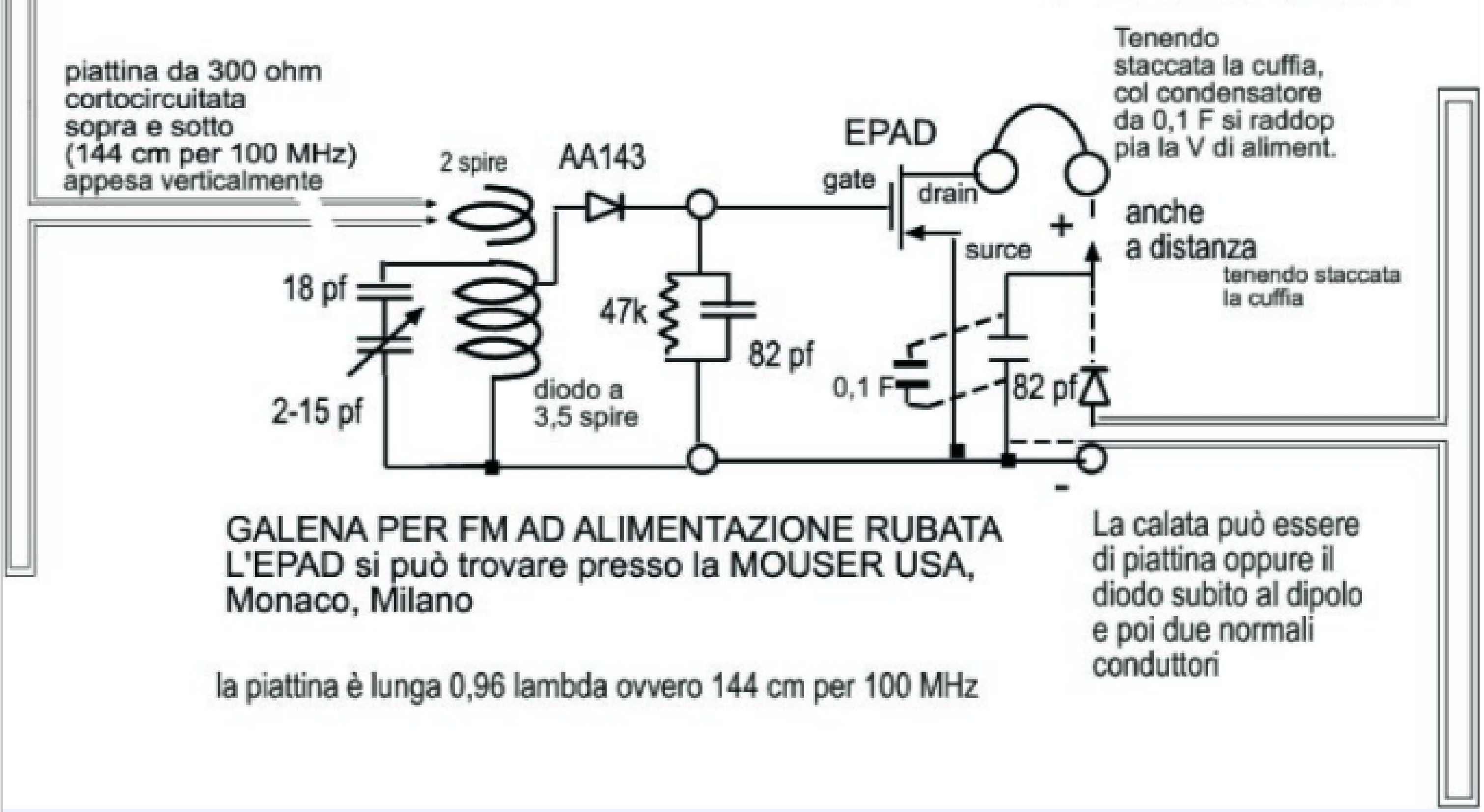
2 spire
AA143
EPAD
gate
drain
surce
18 pf
2-15 pf
diodo a
3,5 spire
47k
82 pf
0,1 F
82 pf
+
-

piattina da 300 ohm
cortocircuitata
sopra e sotto
(144 cm per 100 MHz)
appesa verticalmente

Tenendo
staccata la cuffia,
col condensatore
da 0,1 F si raddop
pia la V di aliment.

anche
a distanza
tenendo staccata
la cuffia

GALENA PER FM AD ALIMENTAZIONE RUBATA
L'EPAD si può trovare presso la MOUSER USA,
Monaco, Milano

La calata può essere
di piattina oppure il
diodo subito al dipolo
e poi due normali
conduttori

la piattina è lunga 0,96 lambda ovvero 144 cm per 100 MHz

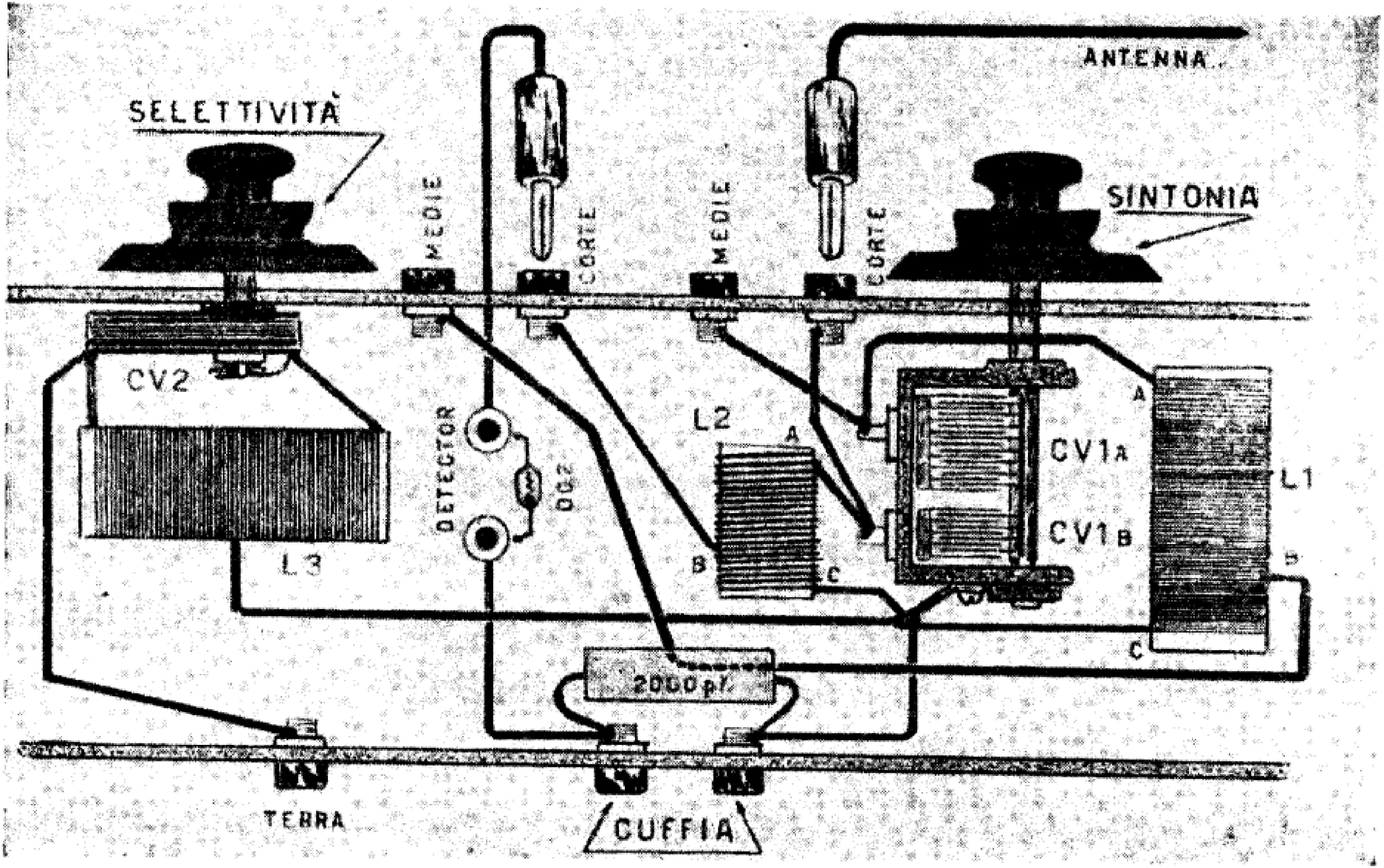
ANTENNA.
SELETTIVITÀ
SINTONIA
MEDIE
CORTE
MEDIE
CORTE
CV2
L3
DETECTOR
OC2
L2
A
B
C
CV1A
CV1B
A
L1
B
C
2000 pF
TERRA
CUFFIA

Un progetto di Luciano Loria

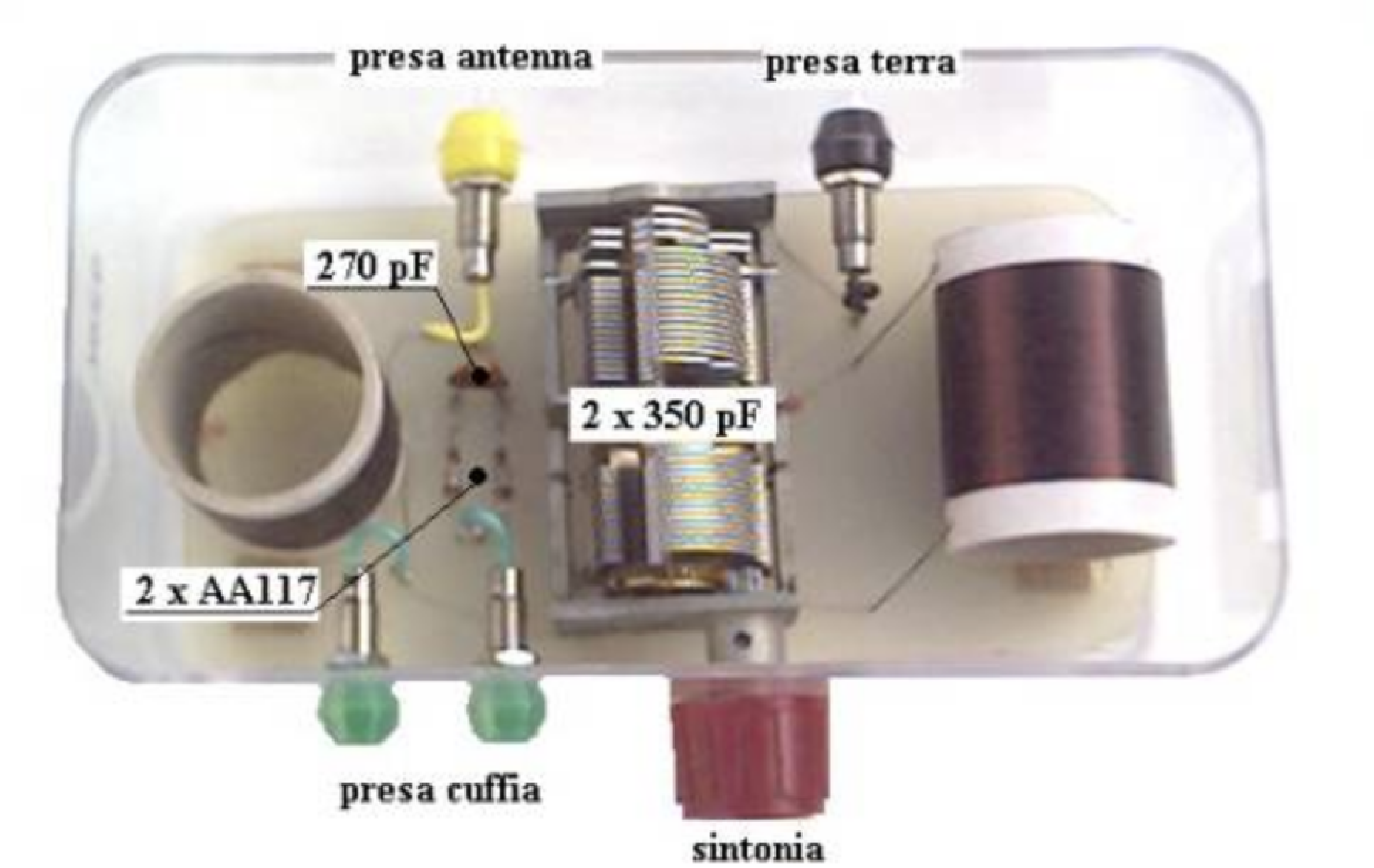
presa antenna
presa terra
270 pF
2 x 350 pF
2 x AA117
presa cuffia
sintonia

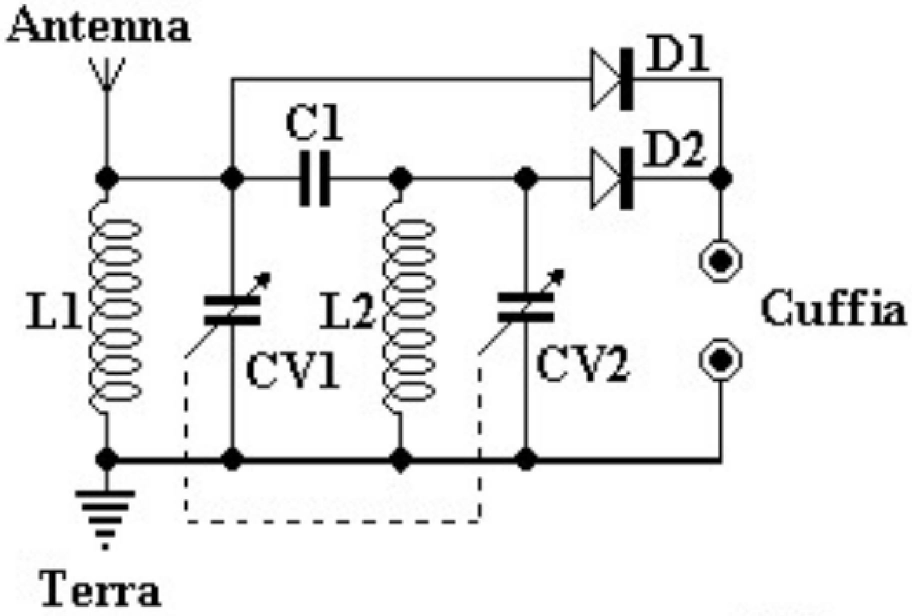
Antenna
D1
C1
D2
L1
CV1
L2
CV2
Cuffia
Terra

Materiale occorrente

L1; L2 = 90 spire filo rame smaltato Ø 0,3 mm
su tubo isolante Ø 3 cm
CV1; CV2 = variabile in aria 350 + 350 pF
C1 = condensatore ceramico 220÷330 pF
D1; D2 = diodo al germanio (AA117)

Schema elettrico

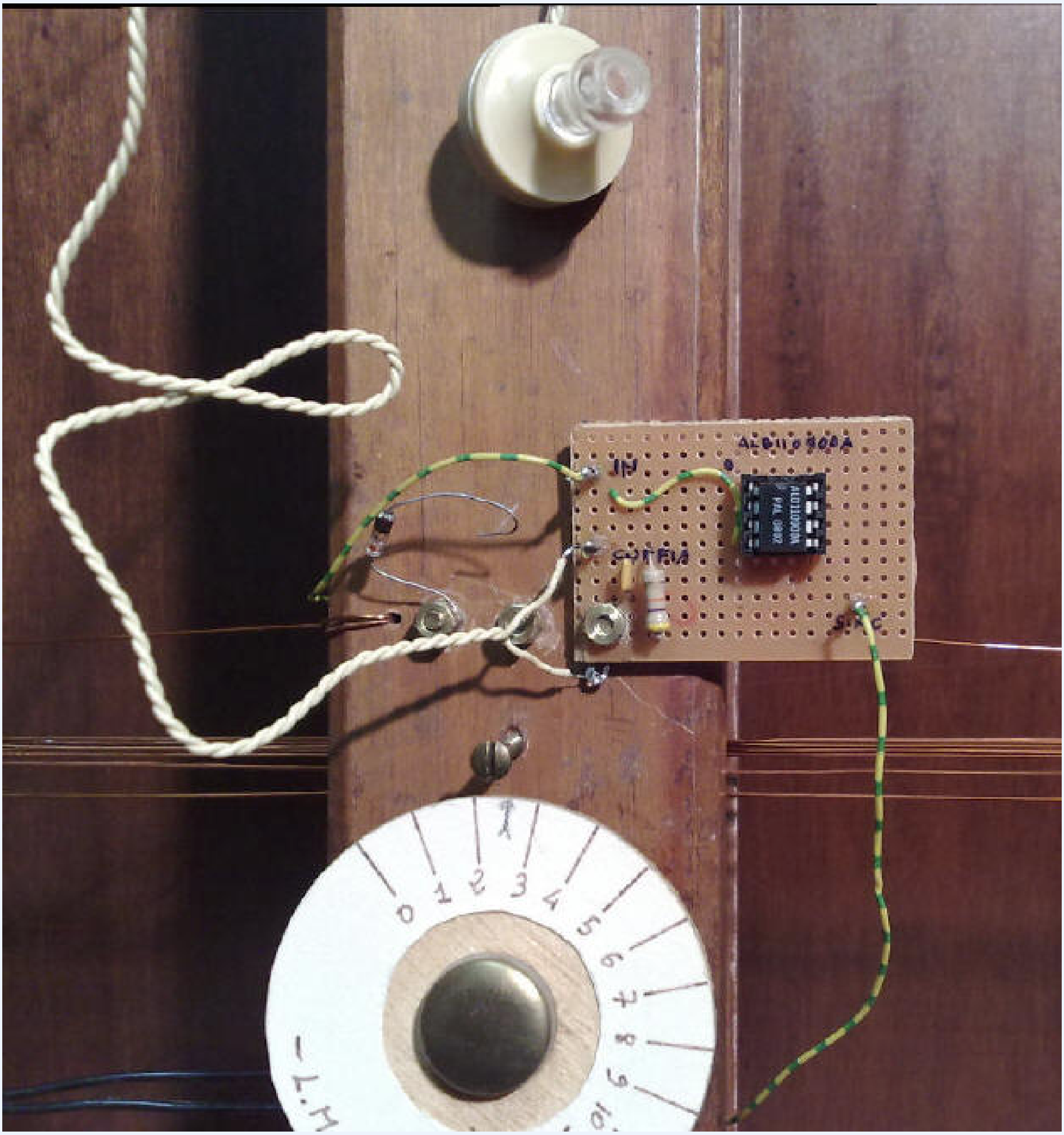
ALD110900A
IN
OUTPUT
0
1
2
3
4
5
6
7
8
9
10
L.M

can see I left the existing 1N34 diode mounted, so that it can be reinserted to make compariso

For the tests I used my receiver with a panel antenna , experimented with great satisfaction a few years ago. To adapt the reception circuit I had to make some small changes, obtaining a scheme like the one below:

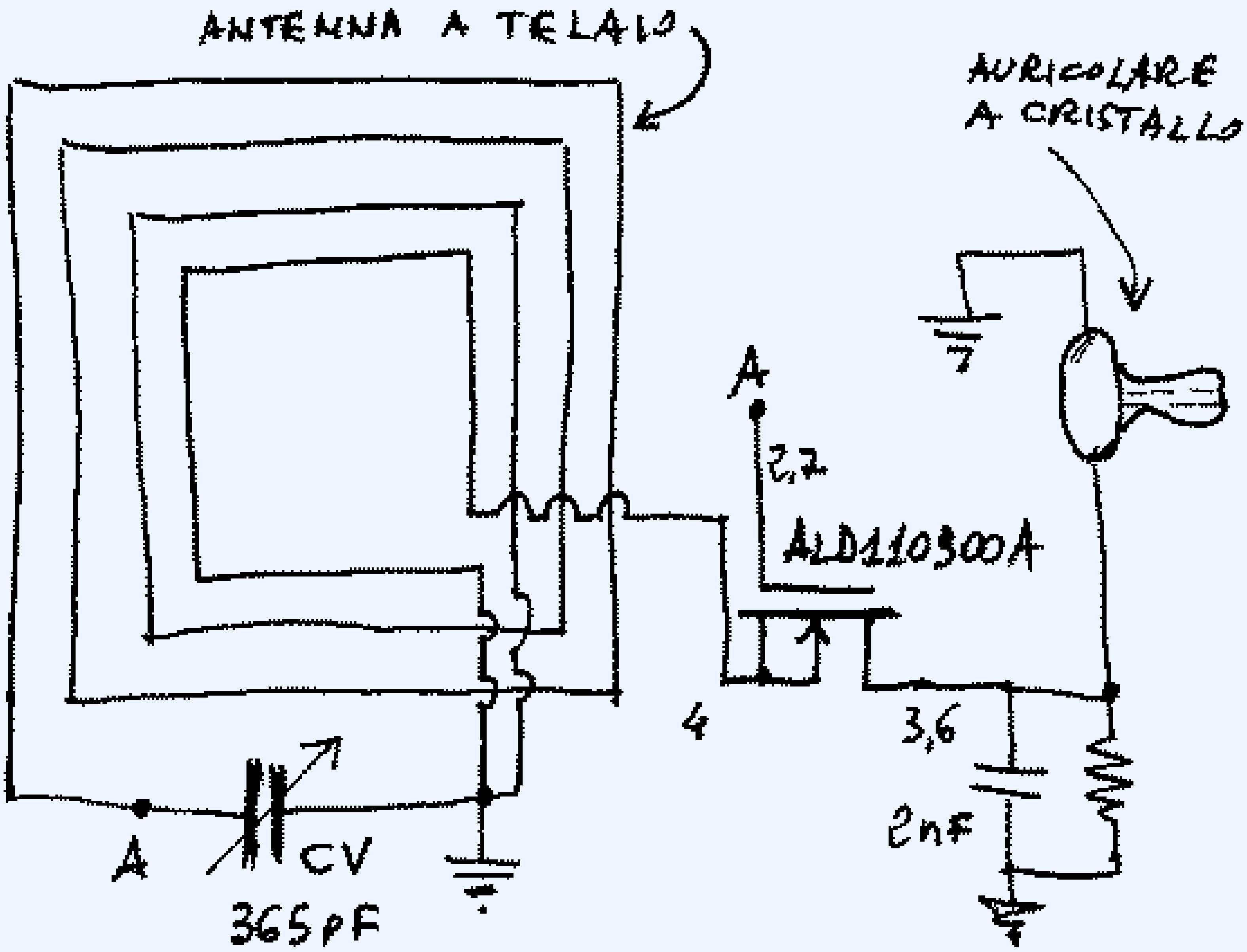

ANTENNA A TELAIO
AURICOLARE A CRISTALLO
A
2,7
ALD110800A
4
3,6
2nF
A
CV
365pF

We increase sensitivity through the use of a
modern zero-threshold mosfet, always without power supply

Principle of operation
From time to time, some new devices appear on the market for electronic components, suitable for experimentation in a "crystal" reception circuit. It happened a few years ago with low-fall "schottky" diodes (BAT46 etc.), which were proposed as good substitutes for old germanium diodes. Recently I have read this article by Peter Hobbs, in which a recently released device is presented, the ALD110900A , mounted in a "high sensitivity" crystal receiver. This is a double mosfet with **a conduction threshold equal to zero**, ideal for detecting without loss a very weak signal like the one that is formed in a crystal receiver. The author proposes the realization of a scheme similar to the following one, which turns out to be of the synchronous detection type.

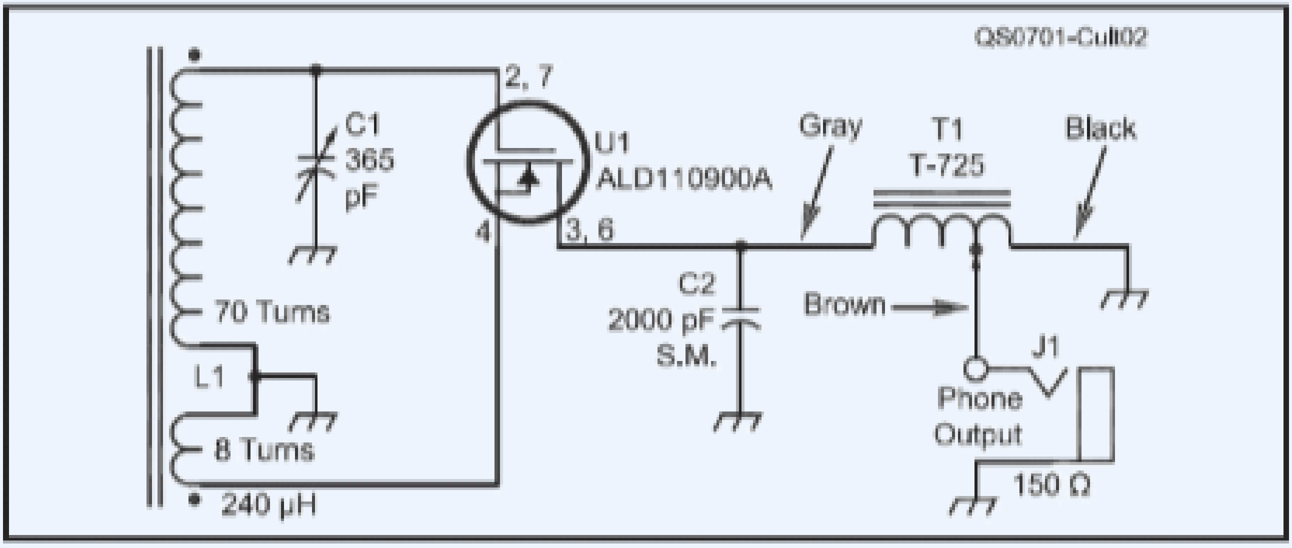

According to the author, the sensitivity is so high as to allow the realization of a receiver without an external antenna, based on the use of only the ferrite antenna, like that of transistor receivers. Its scheme uses as a transducer a telephone capsule coupled to the receiver by an adaptation transformer (this solution deserves to be better investigated).

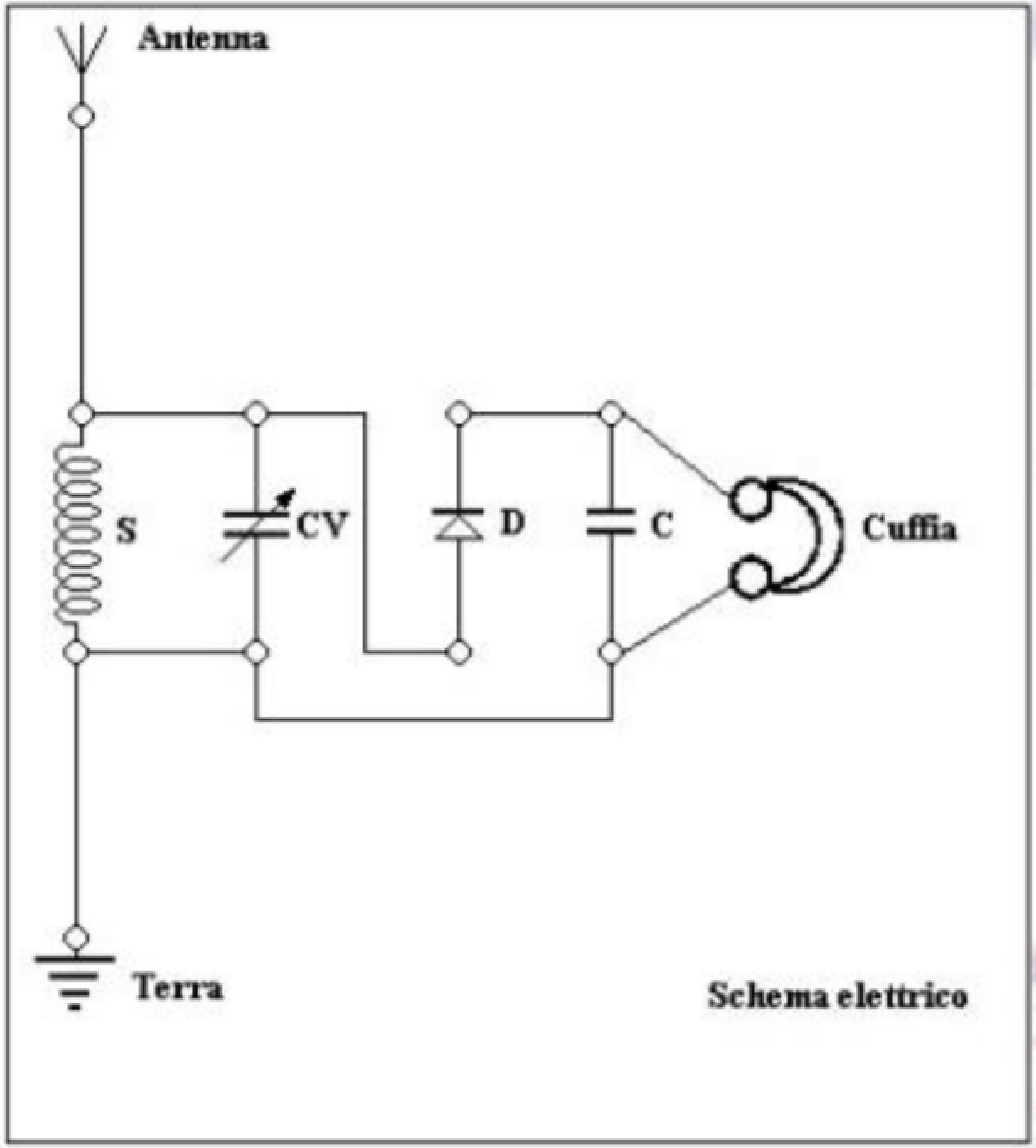
Antenna
S
CV
D
C
Cuffia
Terra
Schema elettrico

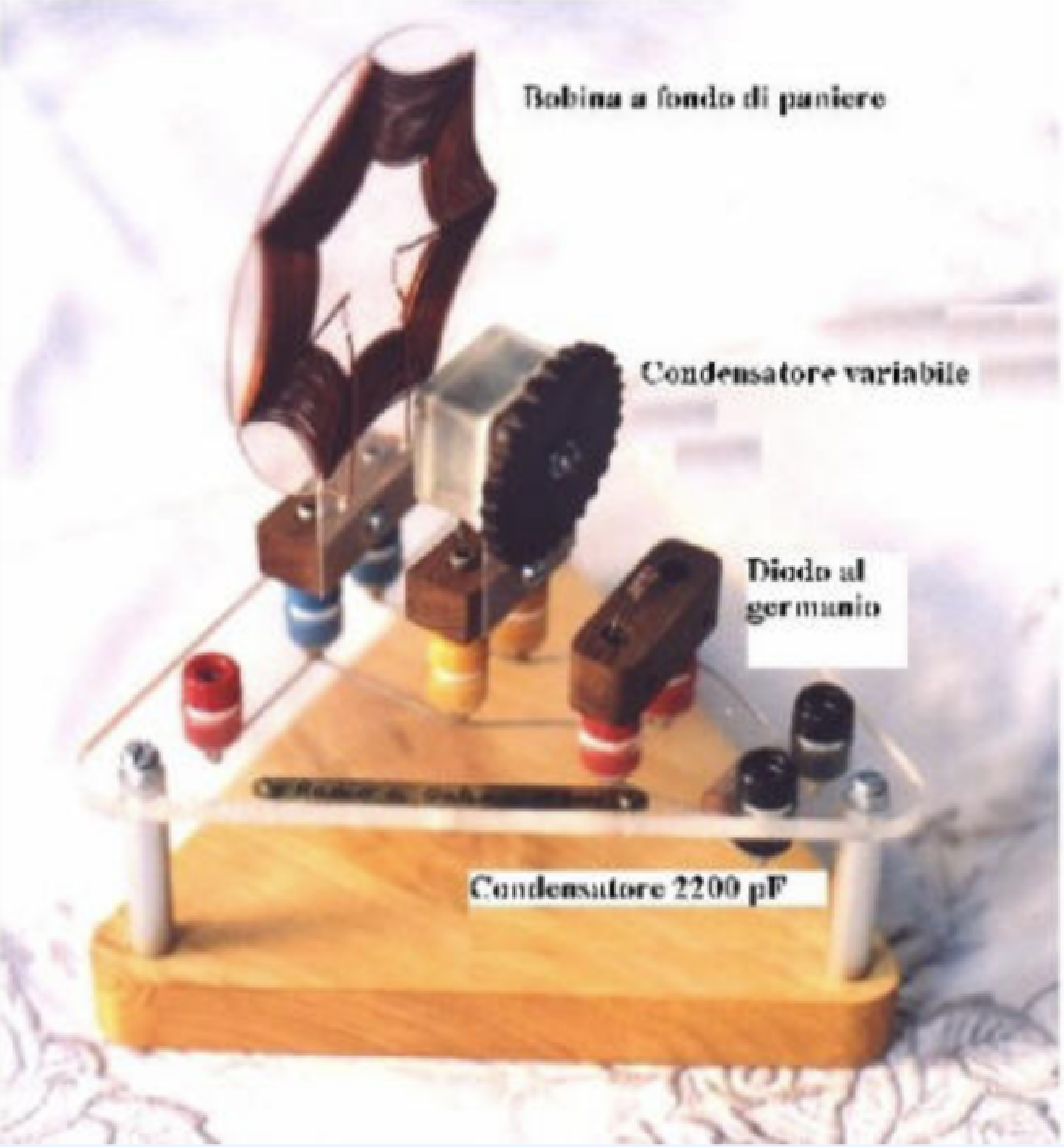
Bobina a fondo di paniere
Condensatore variabile
Diodo al germanio
Condensatore 2200 pF

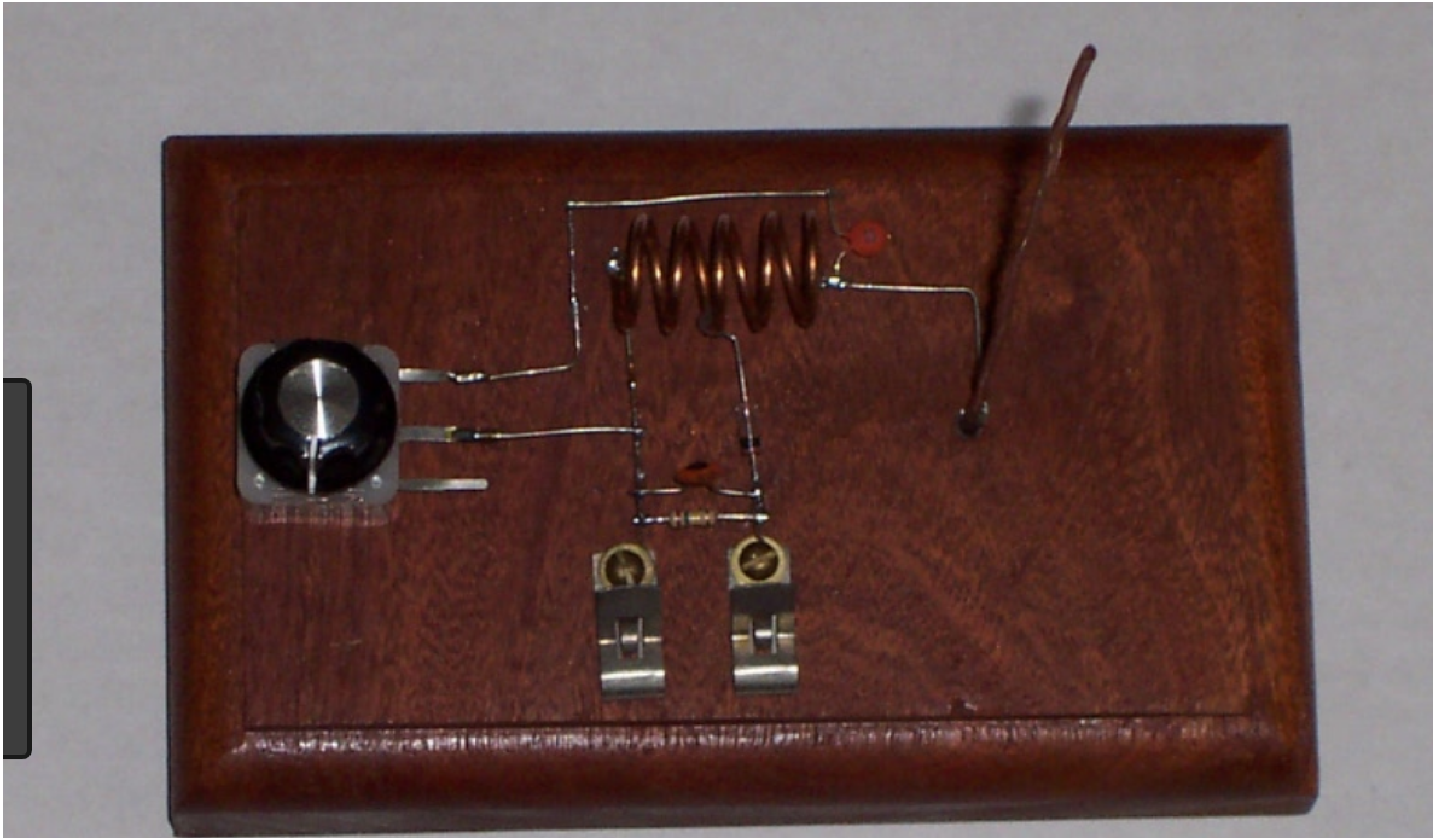

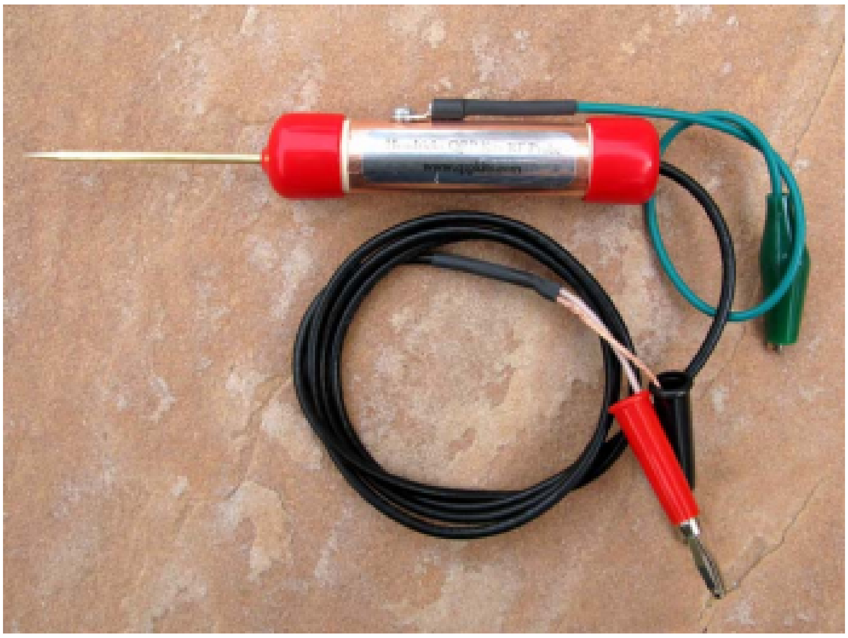

First off, check to see if the parts match the parts list…

1 – 1/2" x 3" CPVC tube
2 – 5/8" O.D. vinyl caps
1 – 3/32" dia x 2.5" brass rod
1 – 4-40 x 7/16" pan head screw
1 – 4-40 x 1/4" pan head screw
2 - #4 internal tooth lock washer
1 – 4-40 nut, steel
1 – 4-40 nut, brass
1 – 3/32 x 2" tyrap
2 - #4, 14-16ga ring terminal
1 – PCB
1 – D1 - 1N34A diode
1 – R1 - 4.7M 1/8w resistor (YEL, VIO, GRN, GLD) **See note**
1 – C1 - .01 disk ceramic capacitor (103)
3' – RG-174 coax
2 – banana plugs, 1 red, 1 black
2" – 3/16" dia. shrink tubing
1 – alligator clip and 9" lead
1 - copper foil tape, 2.25" x 2"
1 – self adhesive label

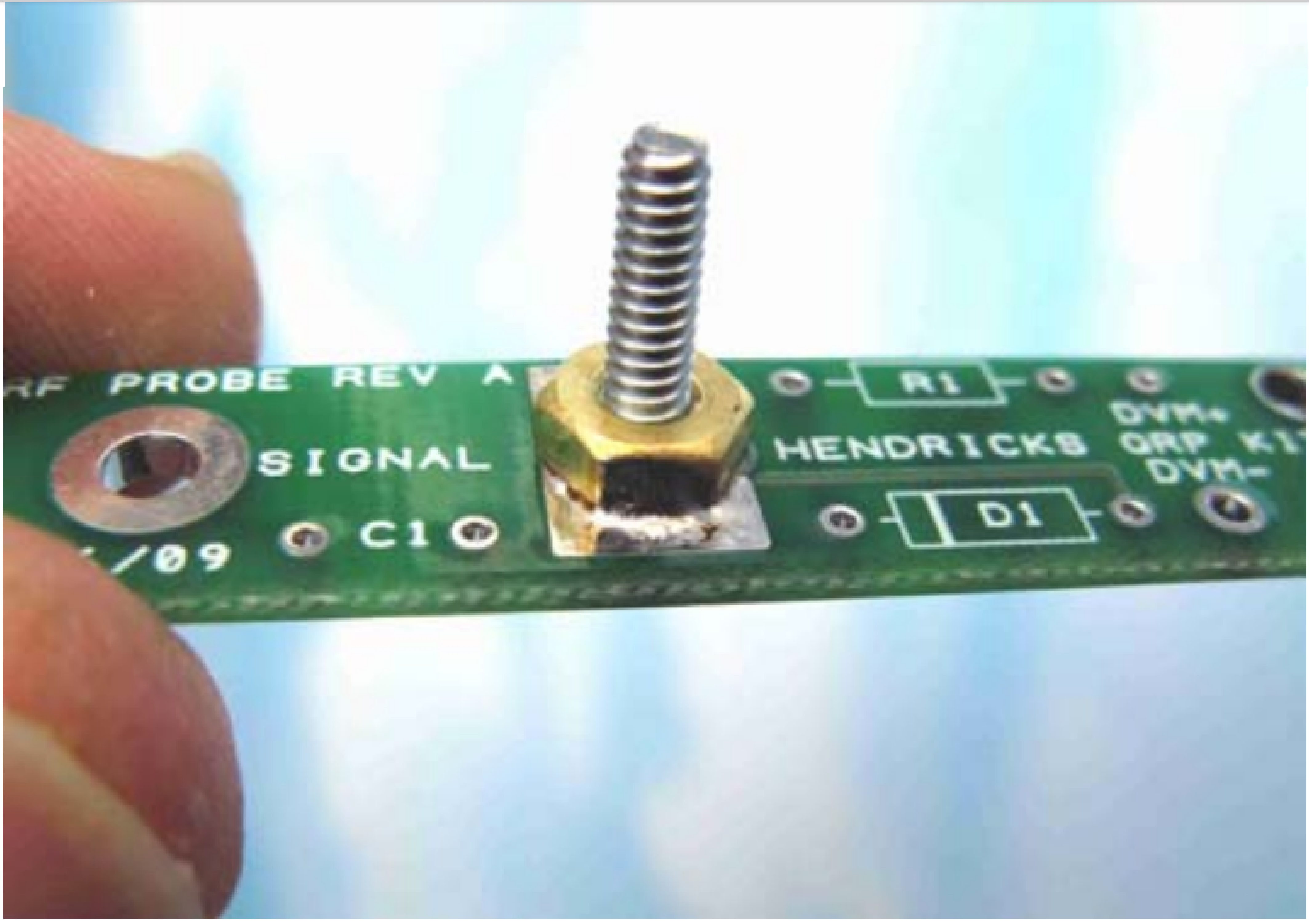
RF PROBE REV A
SIGNAL
C1
/09
R1
DVM+
HENDRICKS QRP KI
DVM-
D1

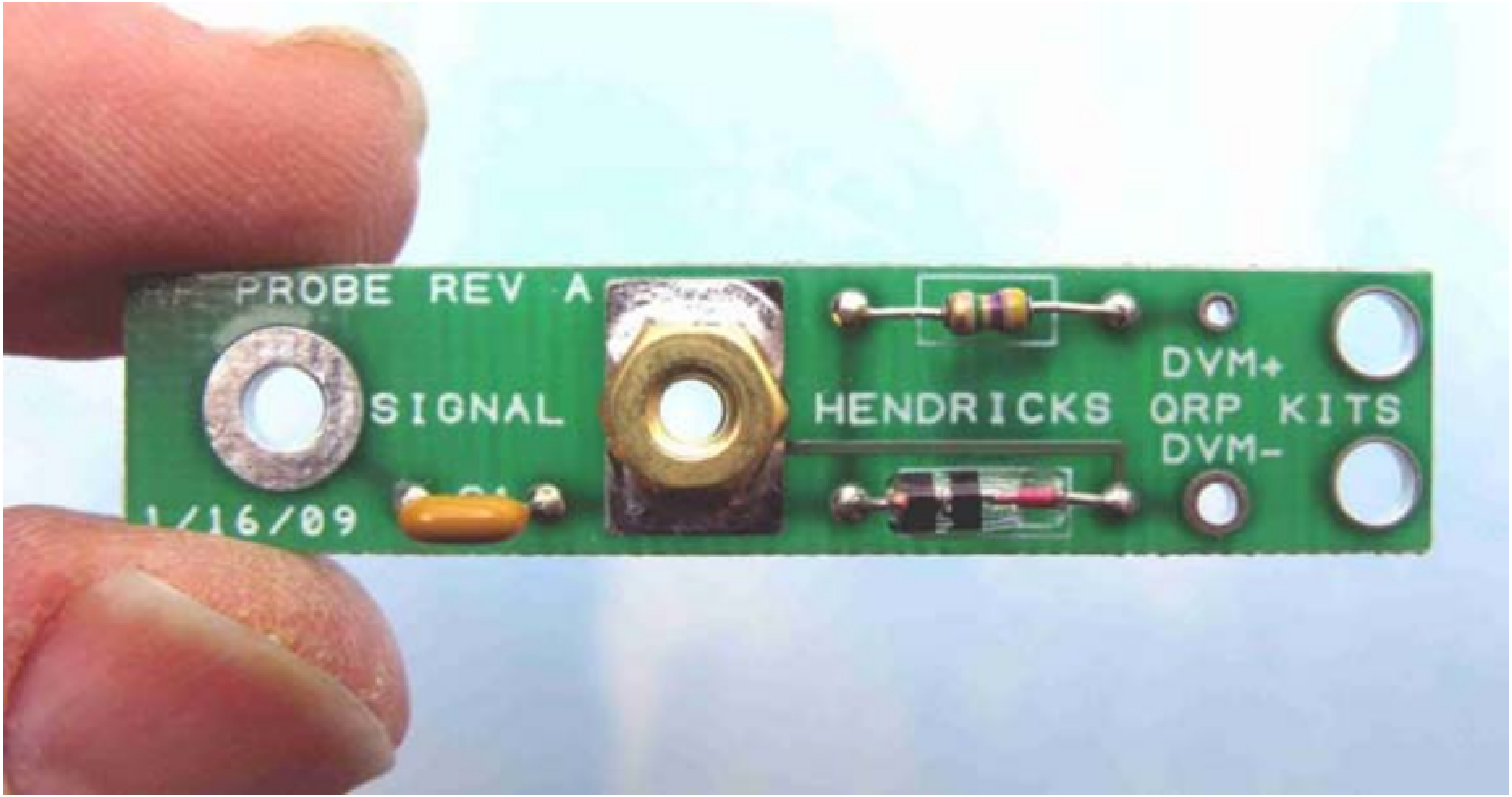
PROBE REV A
SIGNAL
1/16/09
HENDRICKS QRP KITS
DVM+
DVM-

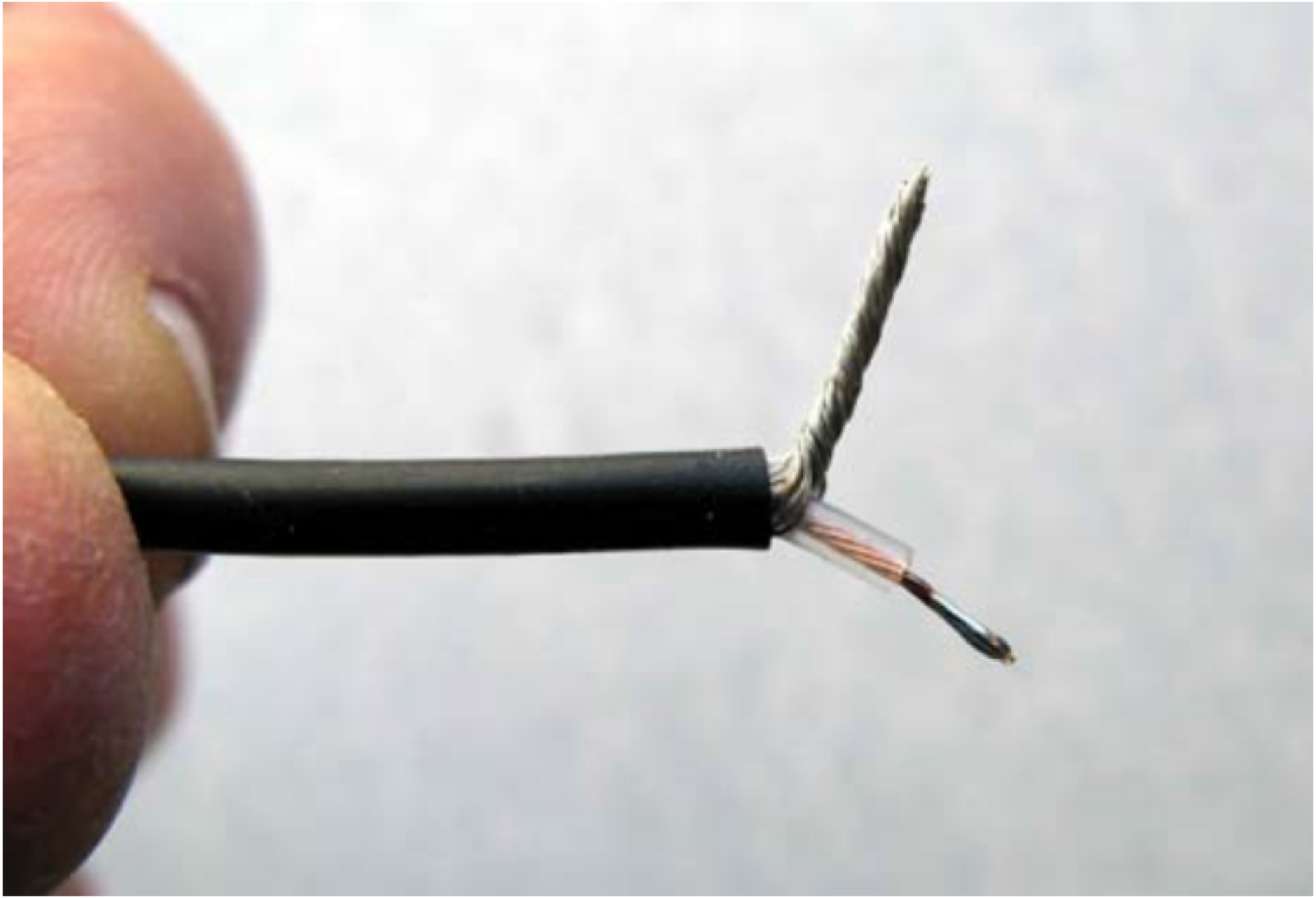

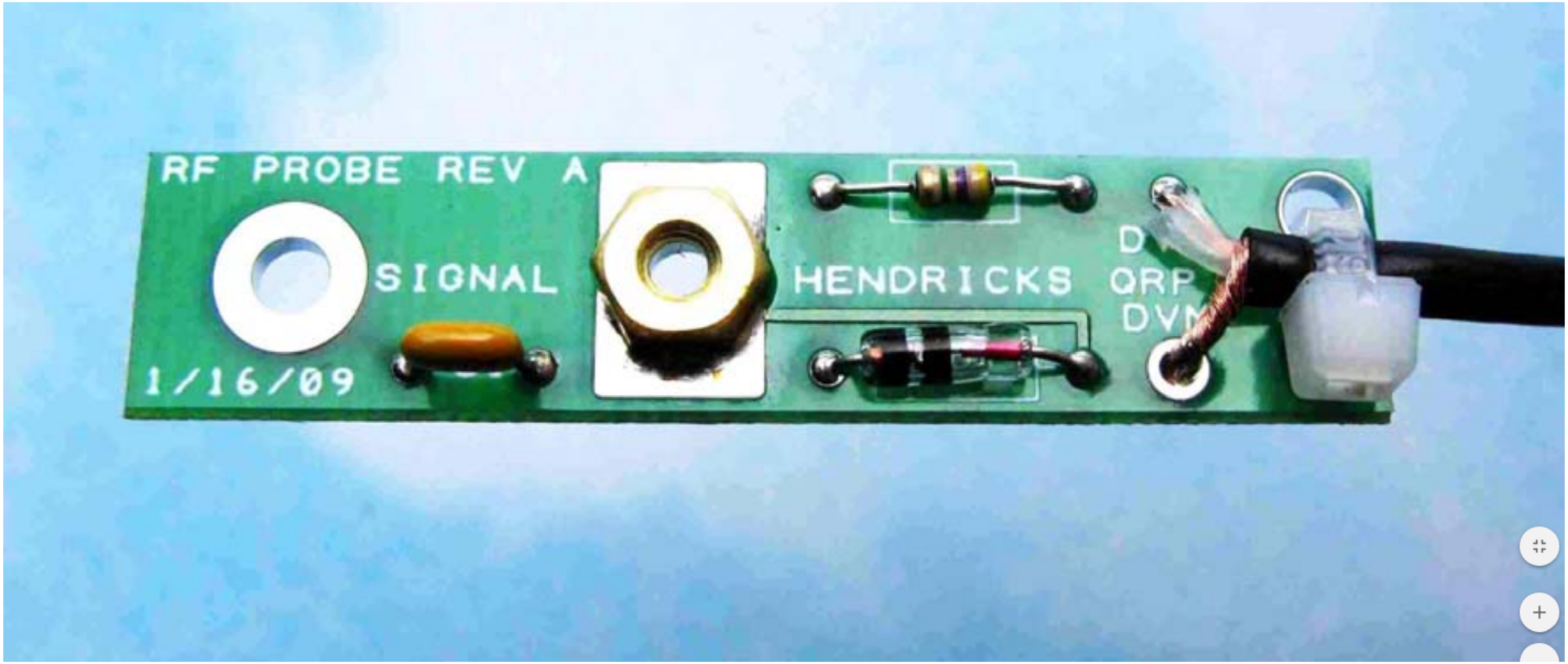
RF PROBE REV A
SIGNAL
1/16/09
HENDRICKS
QRP
DVM

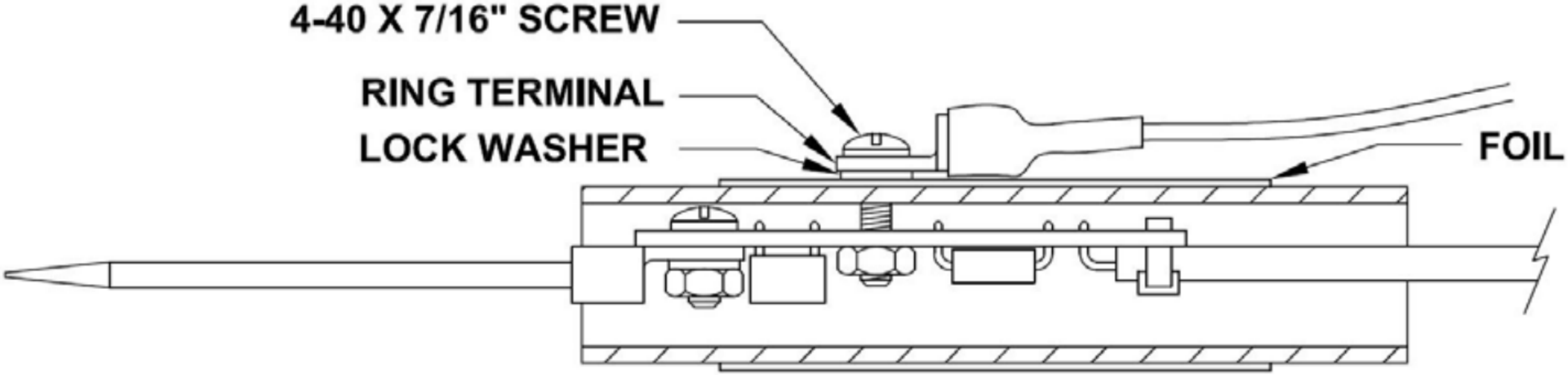
4-40 X 7/16" SCREW
RING TERMINAL
LOCK WASHER
FOIL

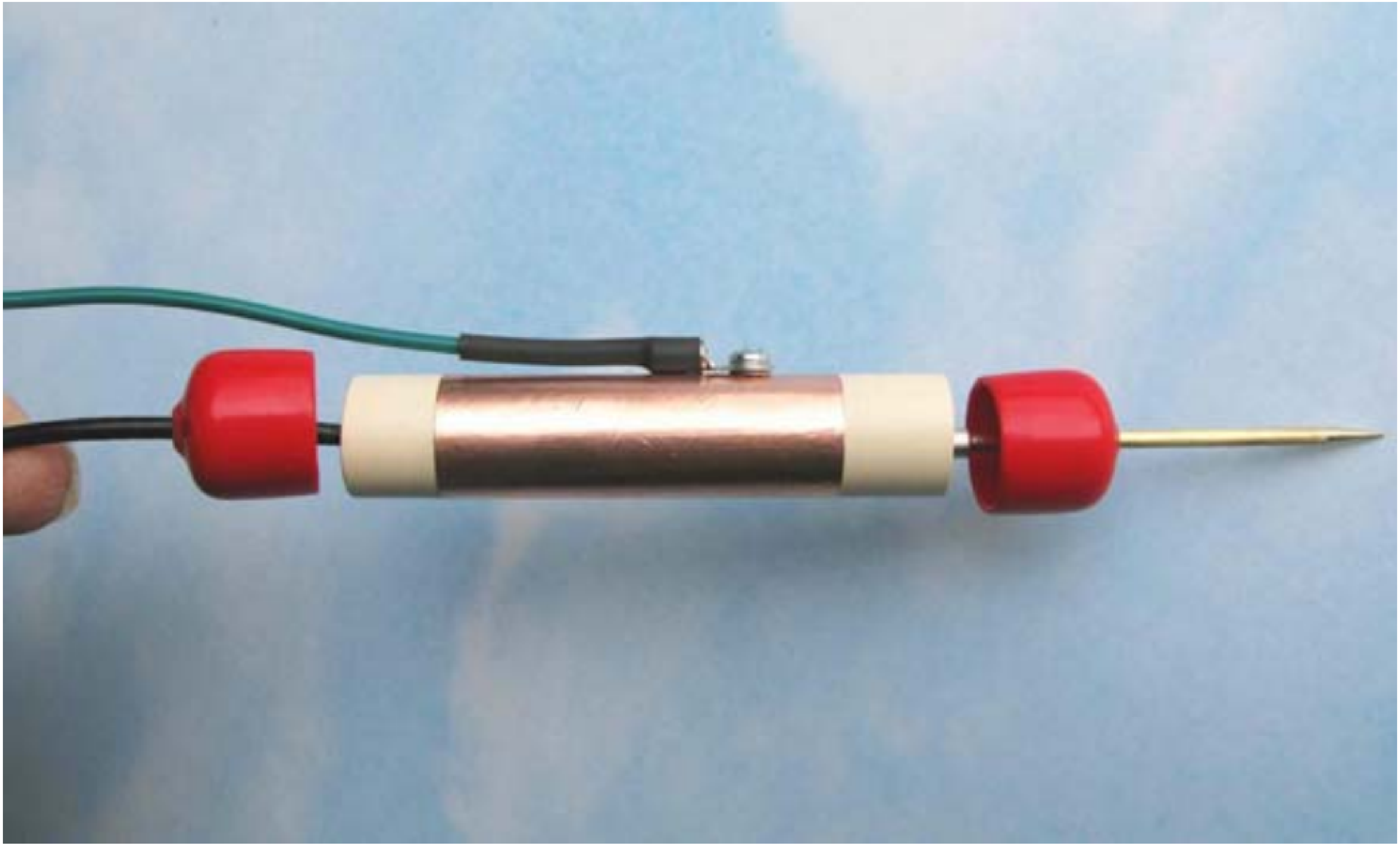

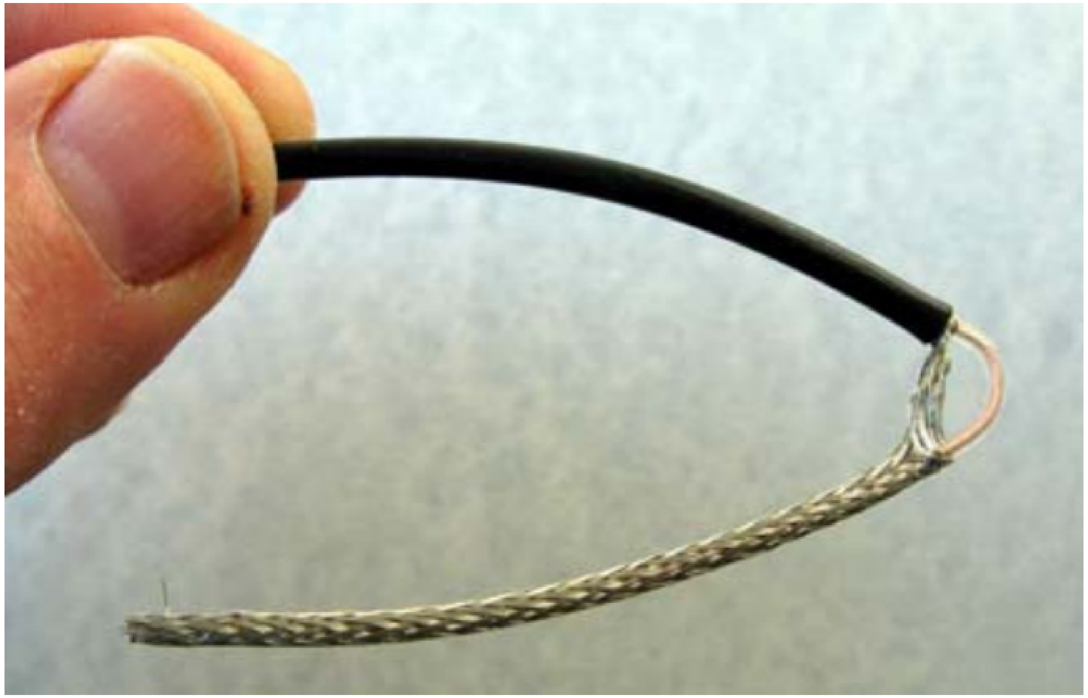

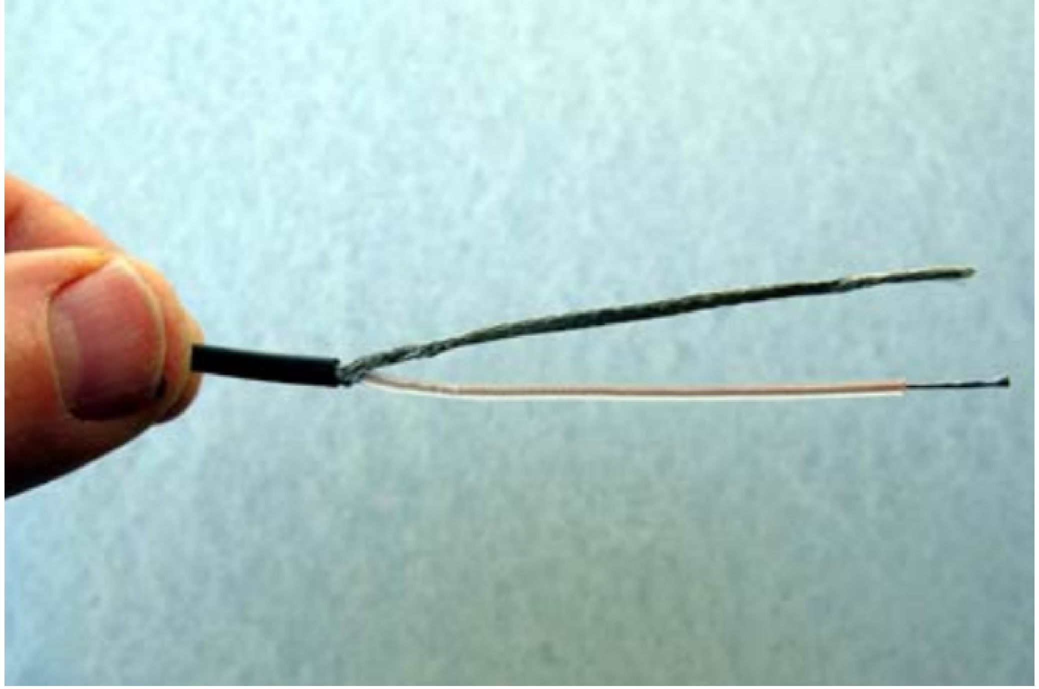

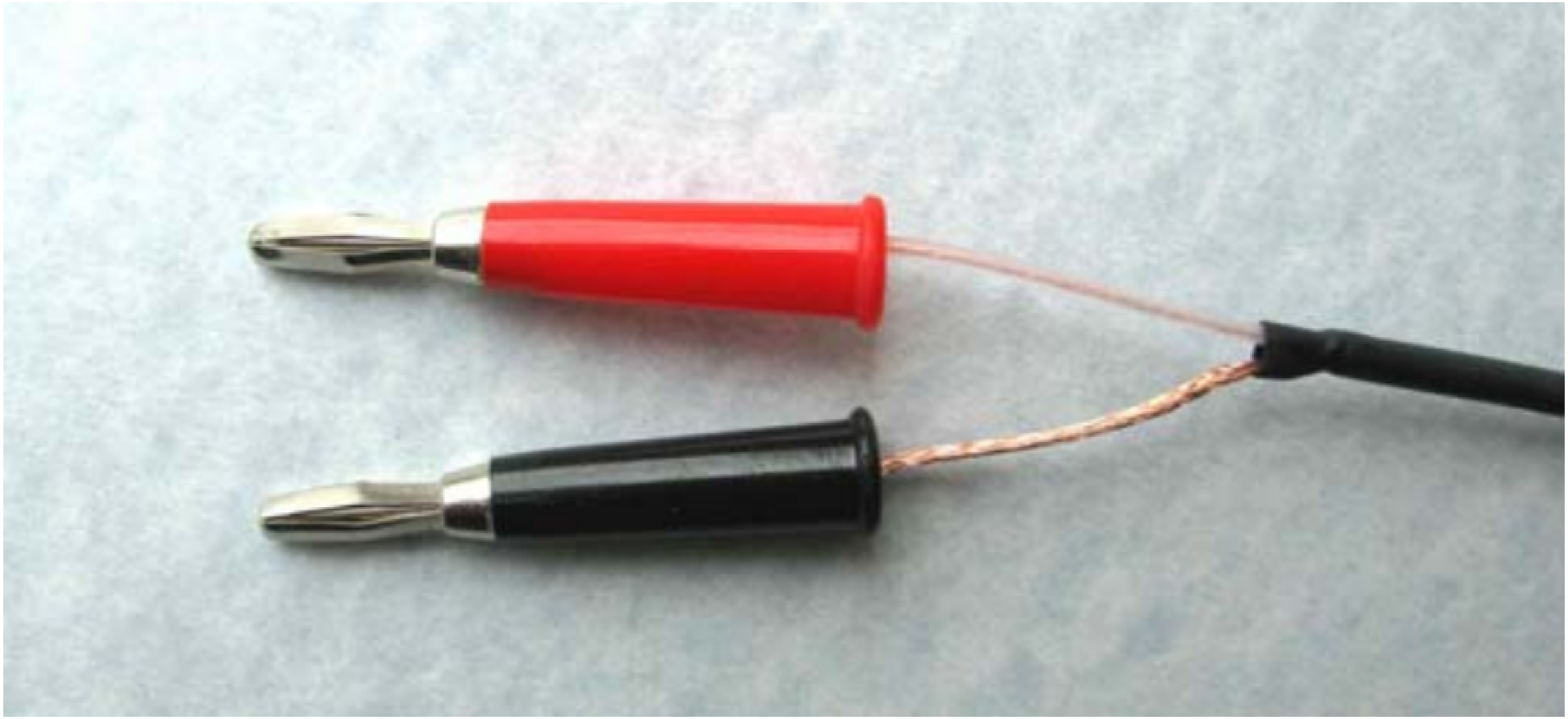

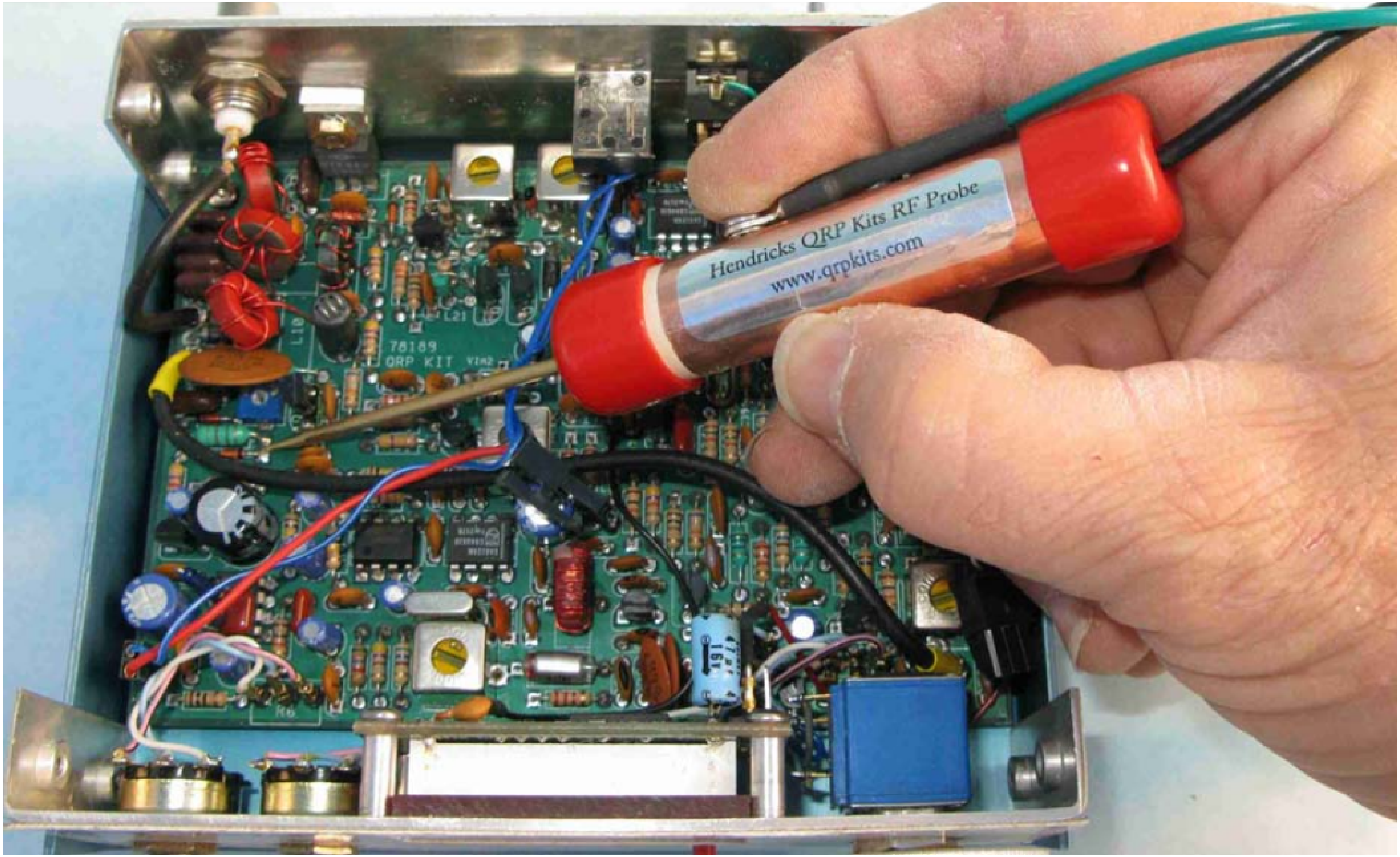
Hendricks QRP Kits RF Probe
www.qrpkits.com
78189
QRP KIT

Look in the files section for **Dar's (Darwin Piatt) – W9HZC** application and usage tips for the R.F. Probe.

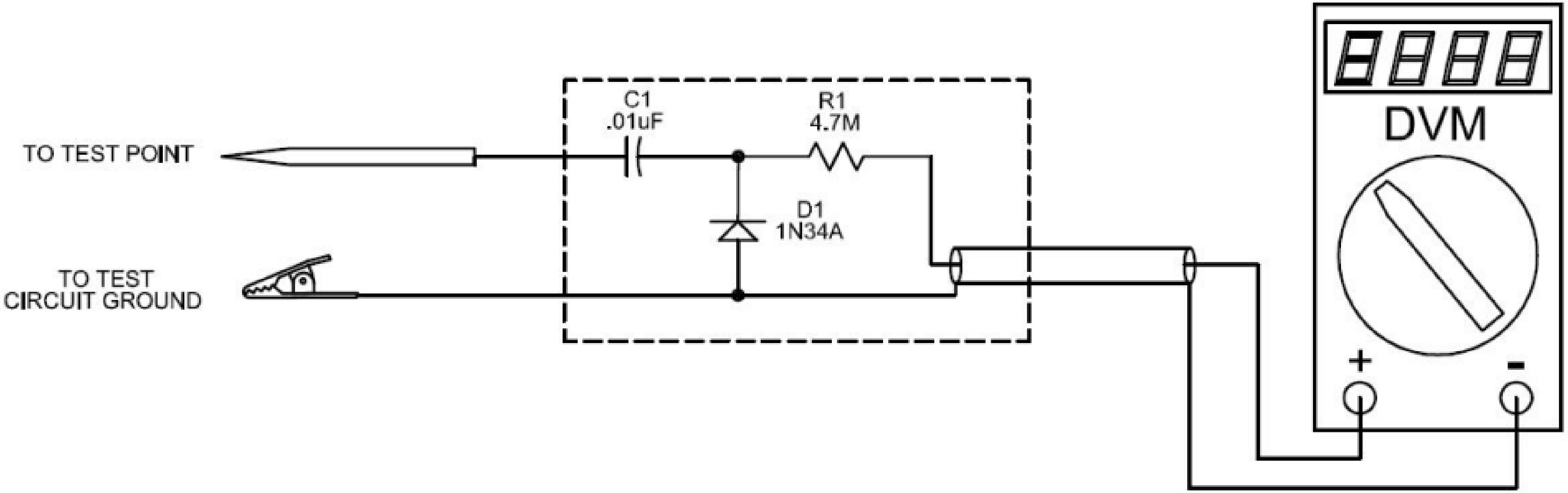
TO TEST POINT
TO TEST CIRCUIT GROUND
C1
.01uF
R1
4.7M
D1
1N34A
DVM
+
-

ELECTRIC ENERGY FROM SPACE:

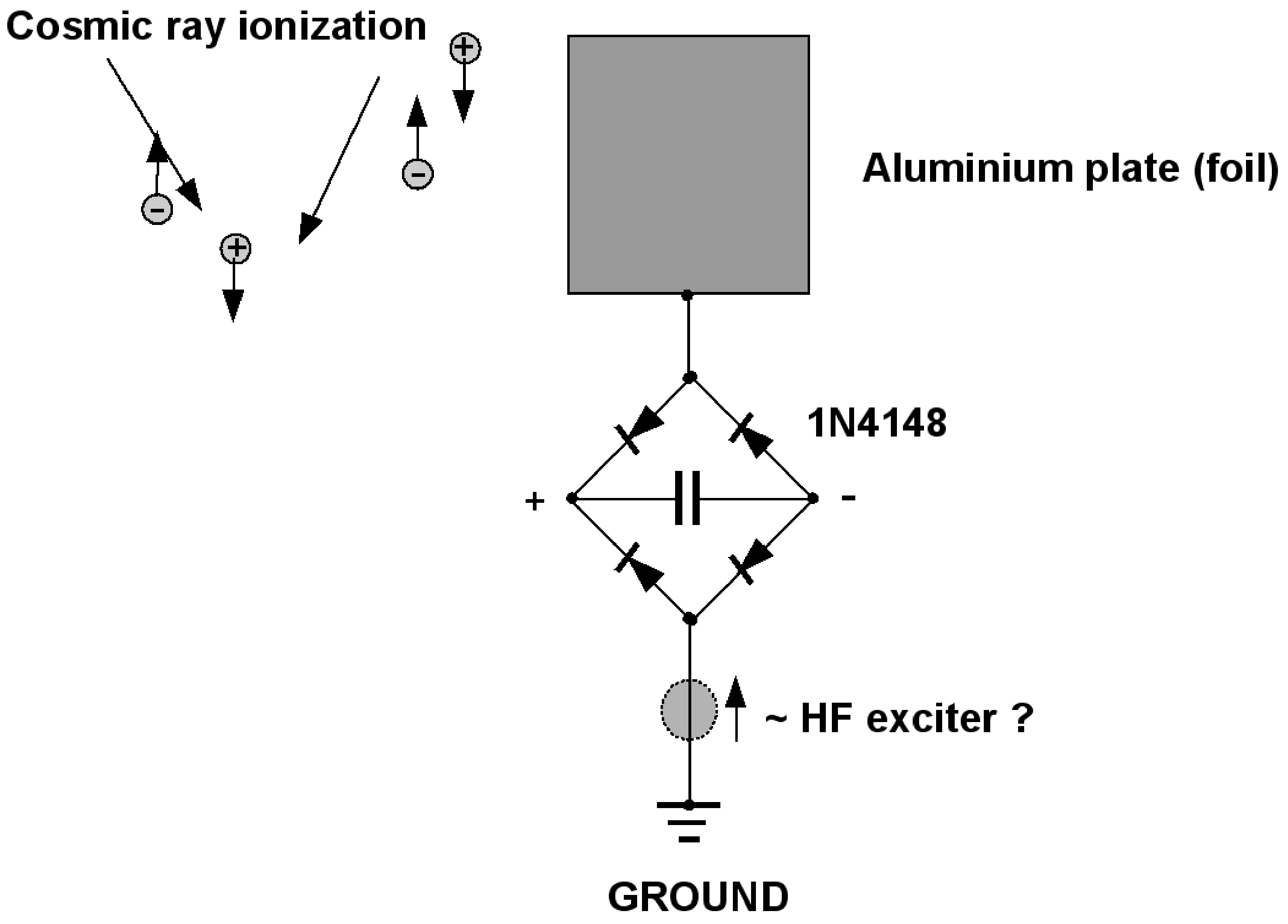

Cosmic ray ionization

Aluminium plate (foil)

1N4148

+

-

~ HF exciter ?

GROUND

RF tip for Multimeter and Oscilloscope

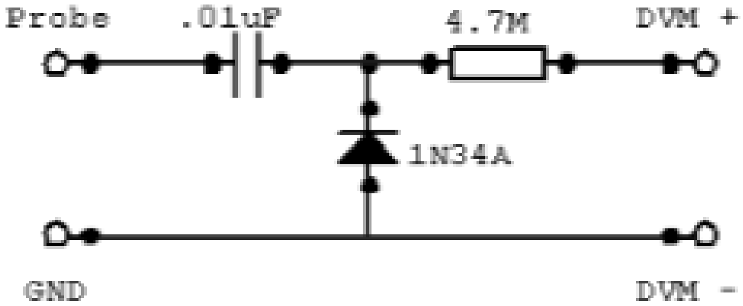

KA8MAV RF PROBE

The project **RF tip** is simple, not need special components and there is no great difficulty in its construction, making it ideal for beginners. The creator of this tip was the amateur KA8MAV, hence the name of the tip arose. Above the electronic circuit probe.

List of components for the tip RF

1 Germanium diode OA79, 1N34 or equivalent
1 Capacitor 0.01uF
1 Resistor 4M7 x 1/4 Watts

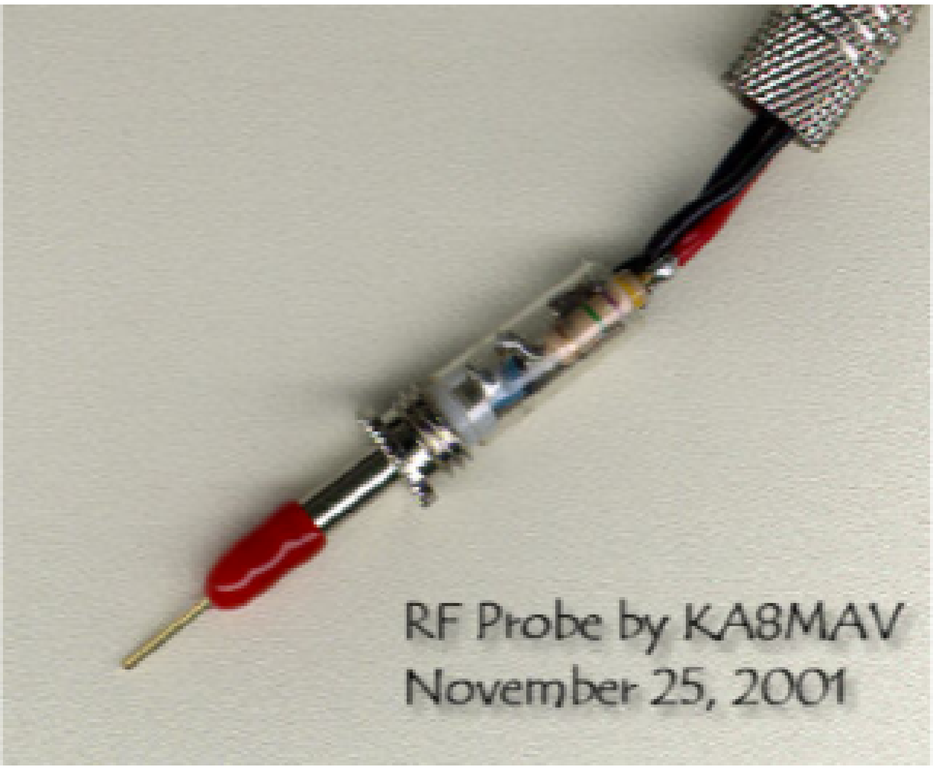

RF Probe by KA8MAV
November 25, 2001

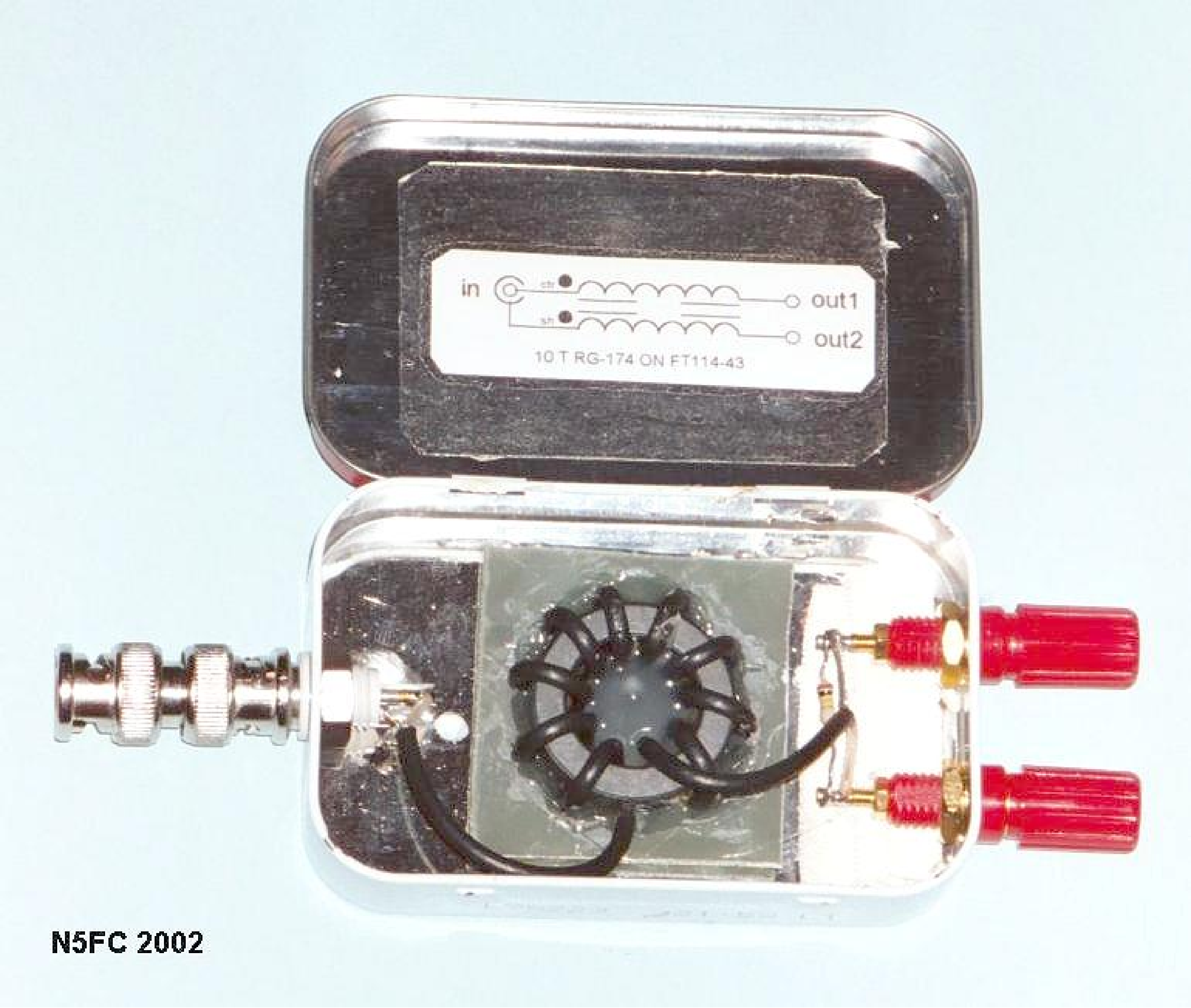
in
ctr
sh
out1
out2
10 T RG-174 ON FT114-43

N5FC 2002

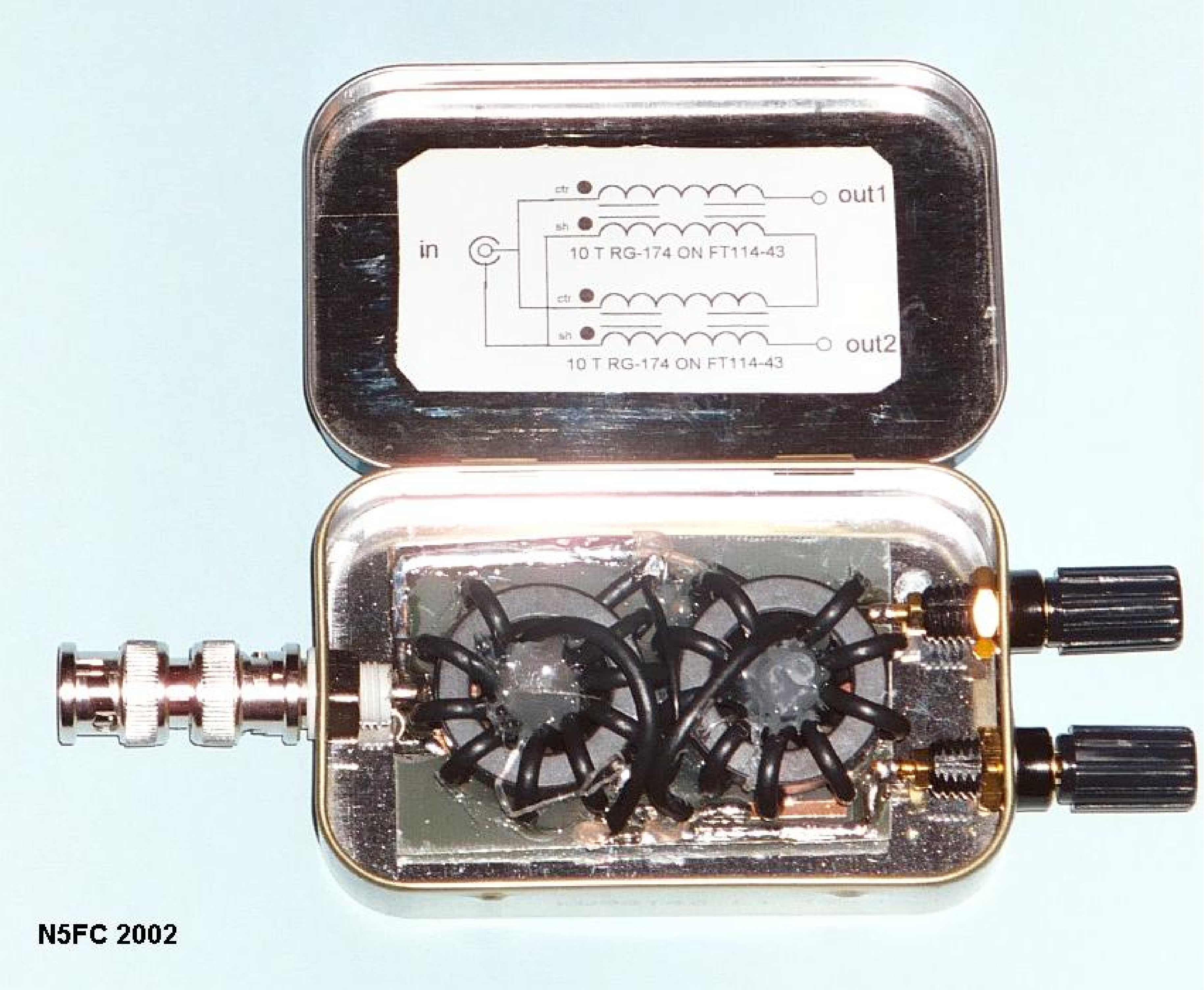
ctr
out1
sh
in
10 T RG-174 ON FT114-43
ctr
sh
out2
10 T RG-174 ON FT114-43

N5FC 2002

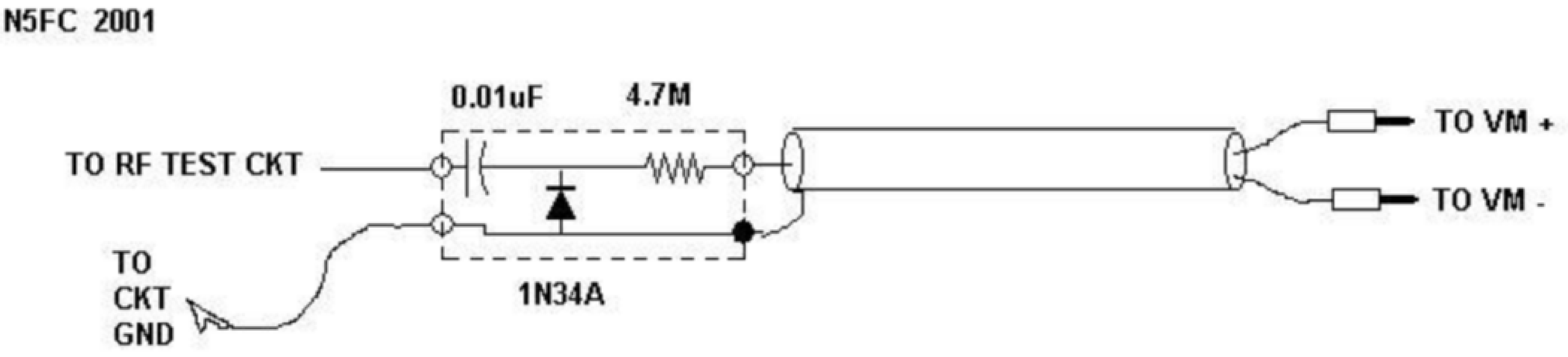
N5FC 2001
0.01uF
4.7M
TO RF TEST CKT
TO VM +
TO VM -
TO
CKT
GND
1N34A

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10 -11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

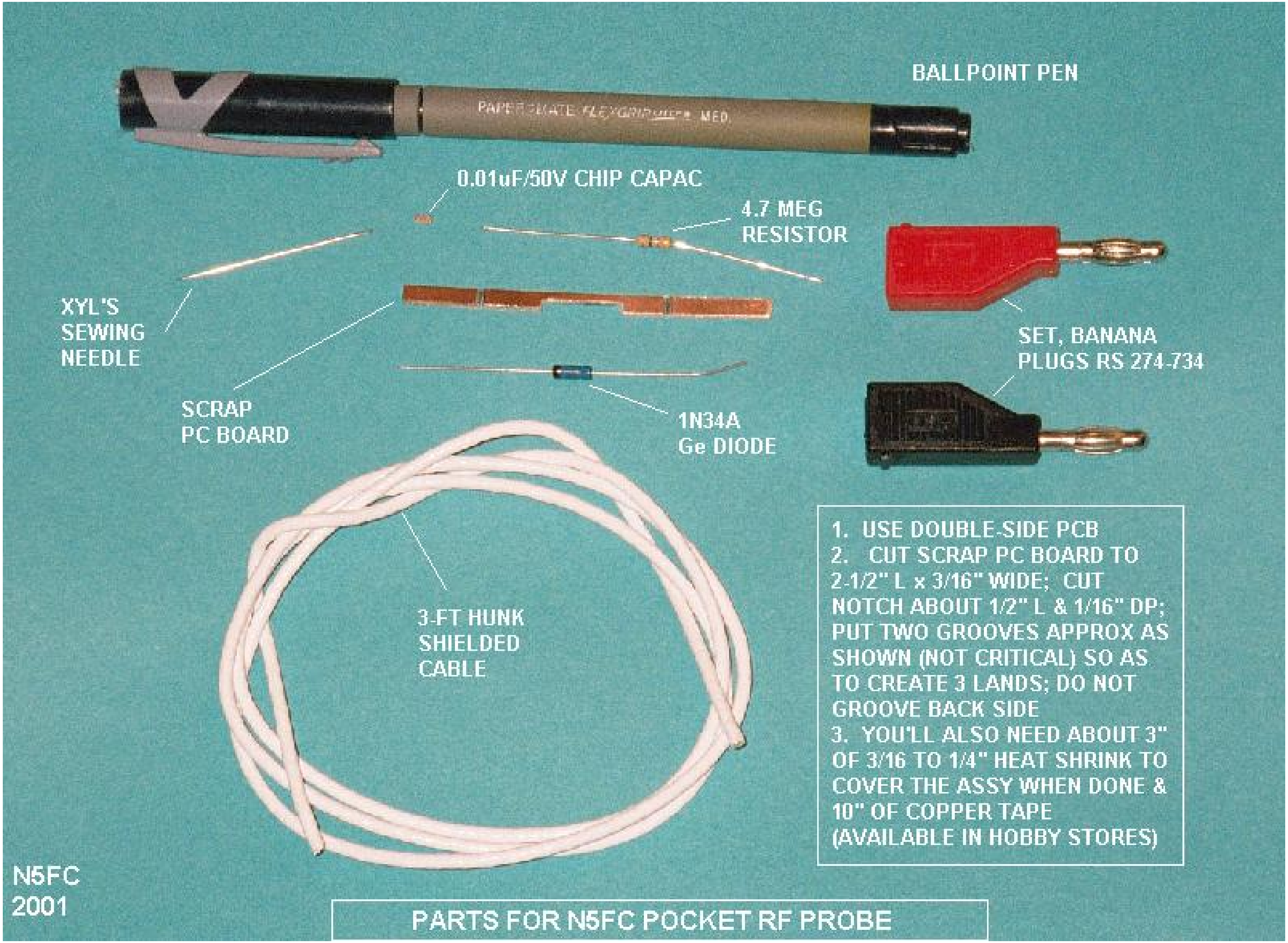
BALLPOINT PEN
PAPERMATE FLEXGRIP MED.
0.01uF/50V CHIP CAPAC
4.7 MEG RESISTOR
XYL'S SEWING NEEDLE
SCRAP PC BOARD
1N34A Ge DIODE
SET, BANANA PLUGS RS 274-734
3-FT HUNK SHIELDED CABLE

1. USE DOUBLE-SIDE PCB
2. CUT SCRAP PC BOARD TO 2-1/2" L x 3/16" WIDE; CUT NOTCH ABOUT 1/2" L & 1/16" DP; PUT TWO GROOVES APPROX AS SHOWN (NOT CRITICAL) SO AS TO CREATE 3 LANDS; DO NOT GROOVE BACK SIDE
3. YOU'LL ALSO NEED ABOUT 3" OF 3/16 TO 1/4" HEAT SHRINK TO COVER THE ASSY WHEN DONE & 10" OF COPPER TAPE (AVAILABLE IN HOBBY STORES)

N5FC
2001

PARTS FOR N5FC POCKET RF PROBE

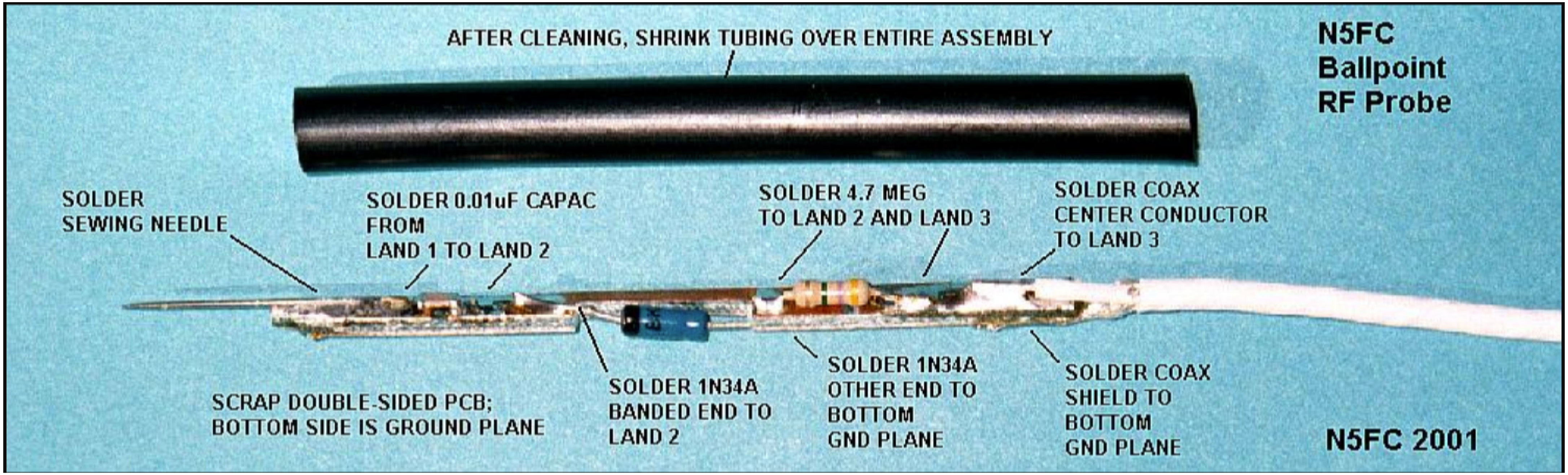

N5FC
Ballpoint
RF Probe

AFTER CLEANING, SHRINK TUBING OVER ENTIRE ASSEMBLY

SOLDER
SEWING NEEDLE

SOLDER 0.01uF CAPAC
FROM
LAND 1 TO LAND 2

SOLDER 4.7 MEG
TO LAND 2 AND LAND 3

SOLDER COAX
CENTER CONDUCTOR
TO LAND 3

SCRAP DOUBLE-SIDED PCB;
BOTTOM SIDE IS GROUND PLANE

SOLDER 1N34A
BANDED END TO
LAND 2

SOLDER 1N34A
OTHER END TO
BOTTOM
GND PLANE

SOLDER COAX
SHIELD TO
BOTTOM
GND PLANE

N5FC 2001

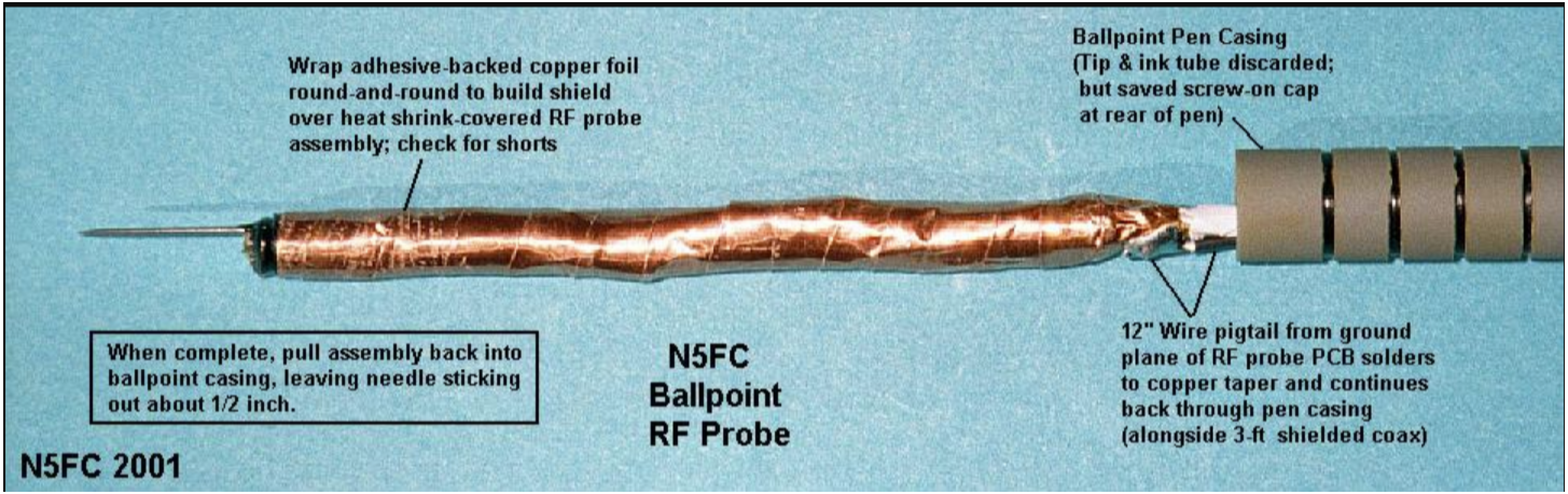

Wrap adhesive-backed copper foil
round-and-round to build shield
over heat shrink-covered RF probe
assembly; check for shorts

Ballpoint Pen Casing
(Tip & ink tube discarded;
but saved screw-on cap
at rear of pen)

When complete, pull assembly back into
ballpoint casing, leaving needle sticking
out about 1/2 inch.

N5FC
Ballpoint
RF Probe

12" Wire pigtail from ground
plane of RF probe PCB solders
to copper taper and continues
back through pen casing
(alongside 3-ft shielded coax)

N5FC 2001

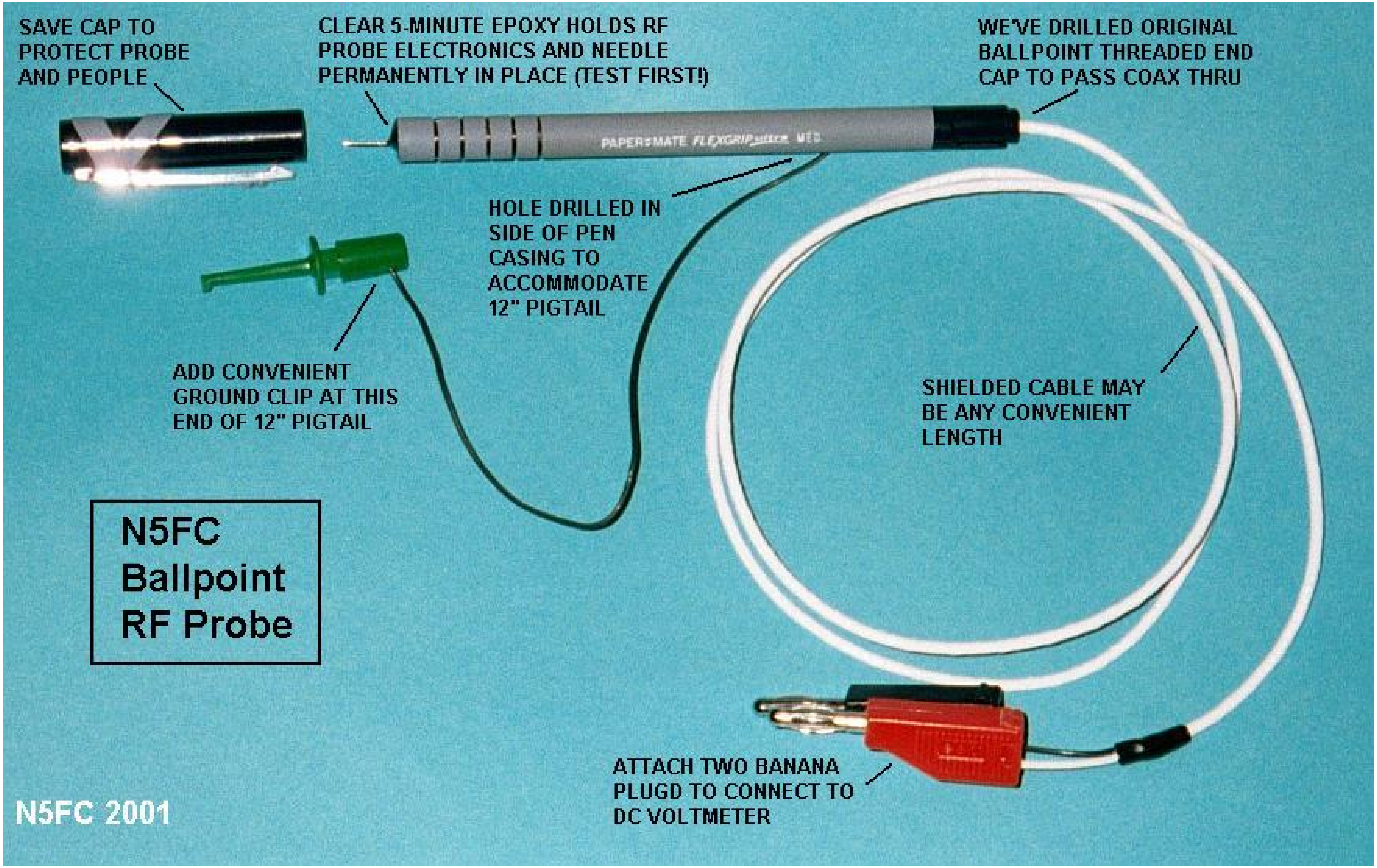

SAVE CAP TO
PROTECT PROBE
AND PEOPLE

CLEAR 5-MINUTE EPOXY HOLDS RF
PROBE ELECTRONICS AND NEEDLE
PERMANENTLY IN PLACE (TEST FIRST!)

WE'VE DRILLED ORIGINAL
BALLPOINT THREADED END
CAP TO PASS COAX THRU

HOLE DRILLED IN
SIDE OF PEN
CASING TO
ACCOMMODATE
12" PIGTAIL

ADD CONVENIENT
GROUND CLIP AT THIS
END OF 12" PIGTAIL

SHIELDED CABLE MAY
BE ANY CONVENIENT
LENGTH

N5FC
Ballpoint
RF Probe

ATTACH TWO BANANA
PLUGD TO CONNECT TO
DC VOLTMETER

N5FC 2001

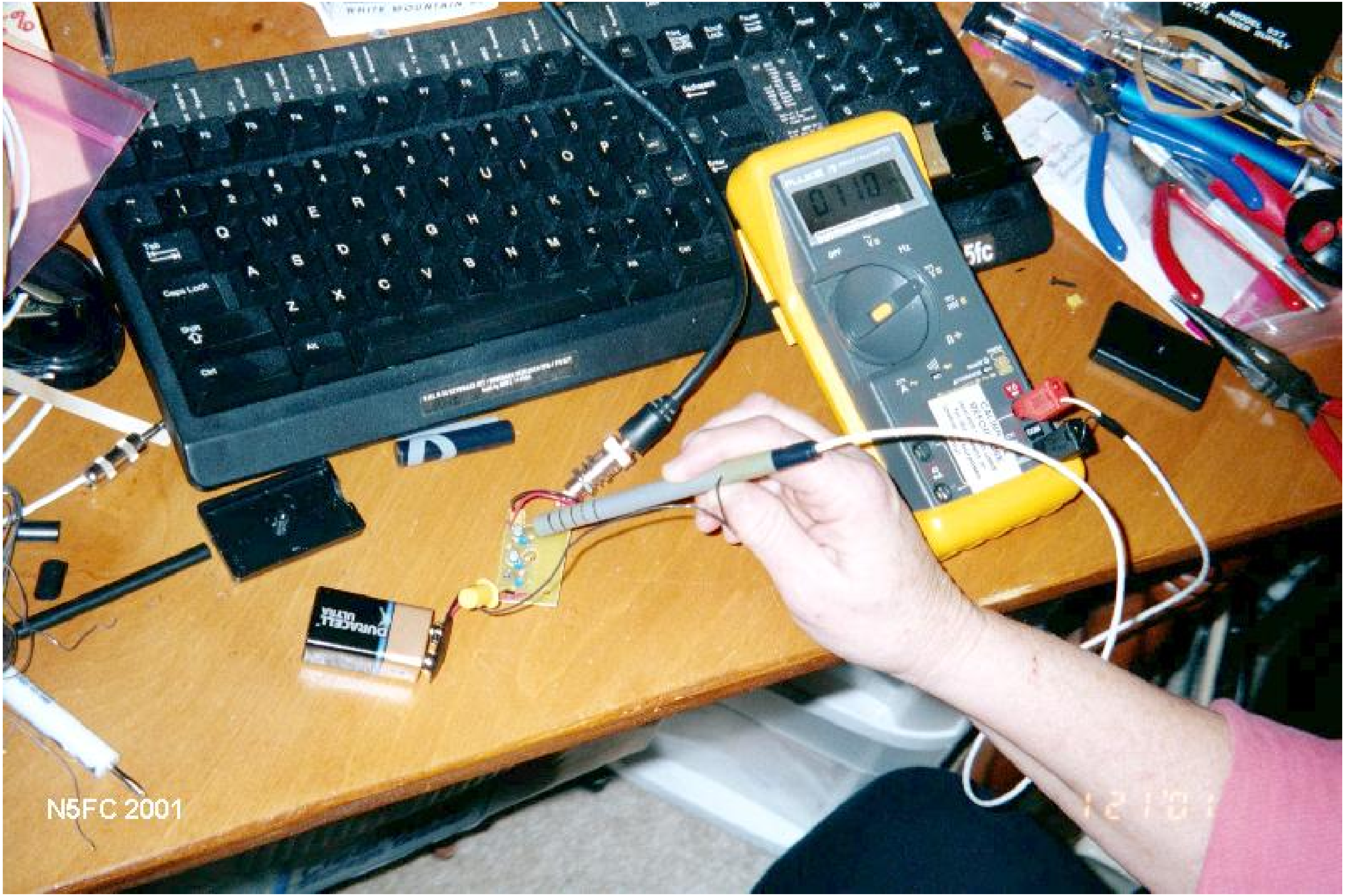
N5FC 2001
1 21 01

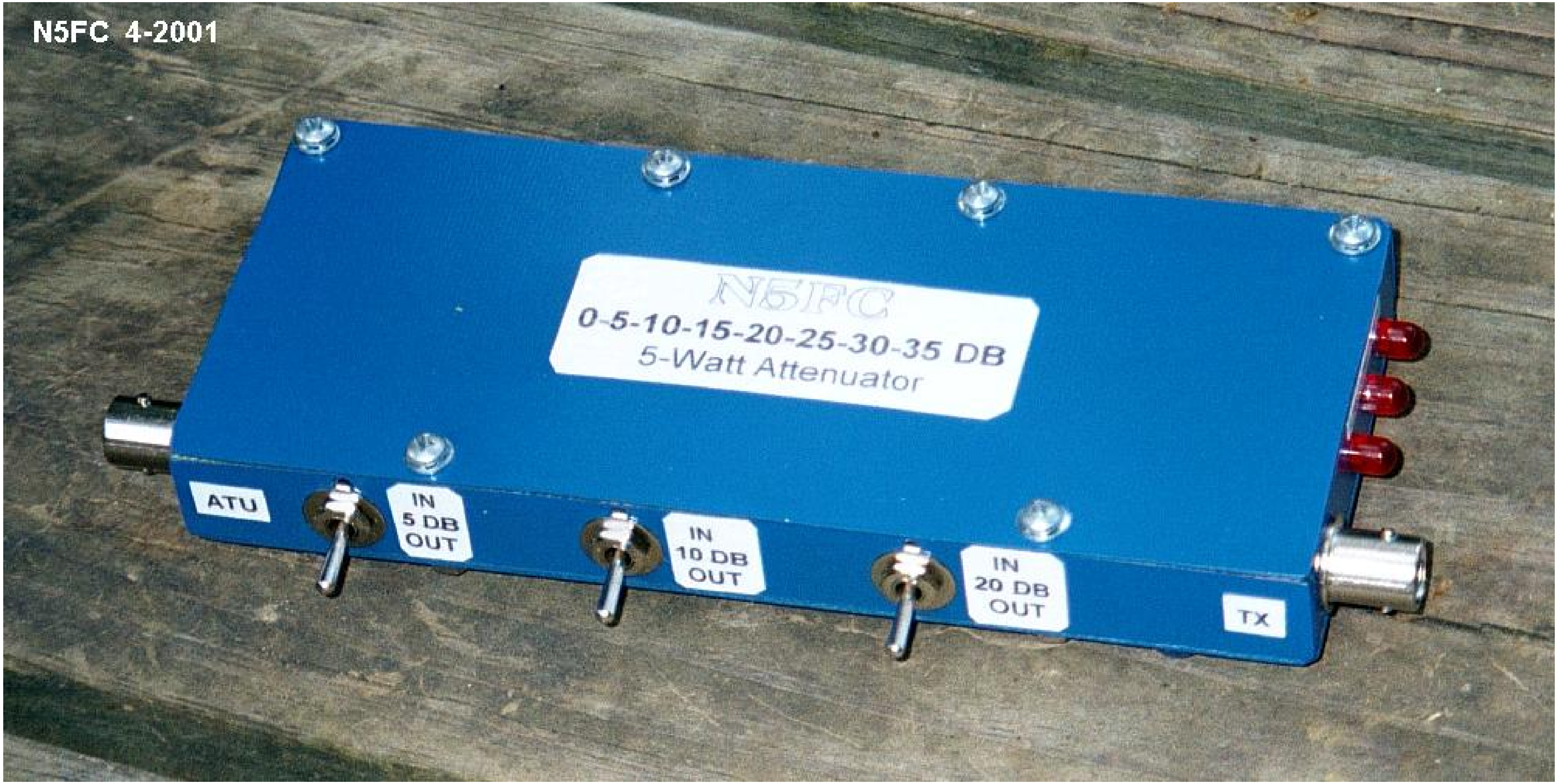
N5FC 4-2001

N5FC
0-5-10-15-20-25-30-35 DB
5-Watt Attenuator
ATU
IN 5 DB OUT
IN 10 DB OUT
IN 20 DB OUT
TX

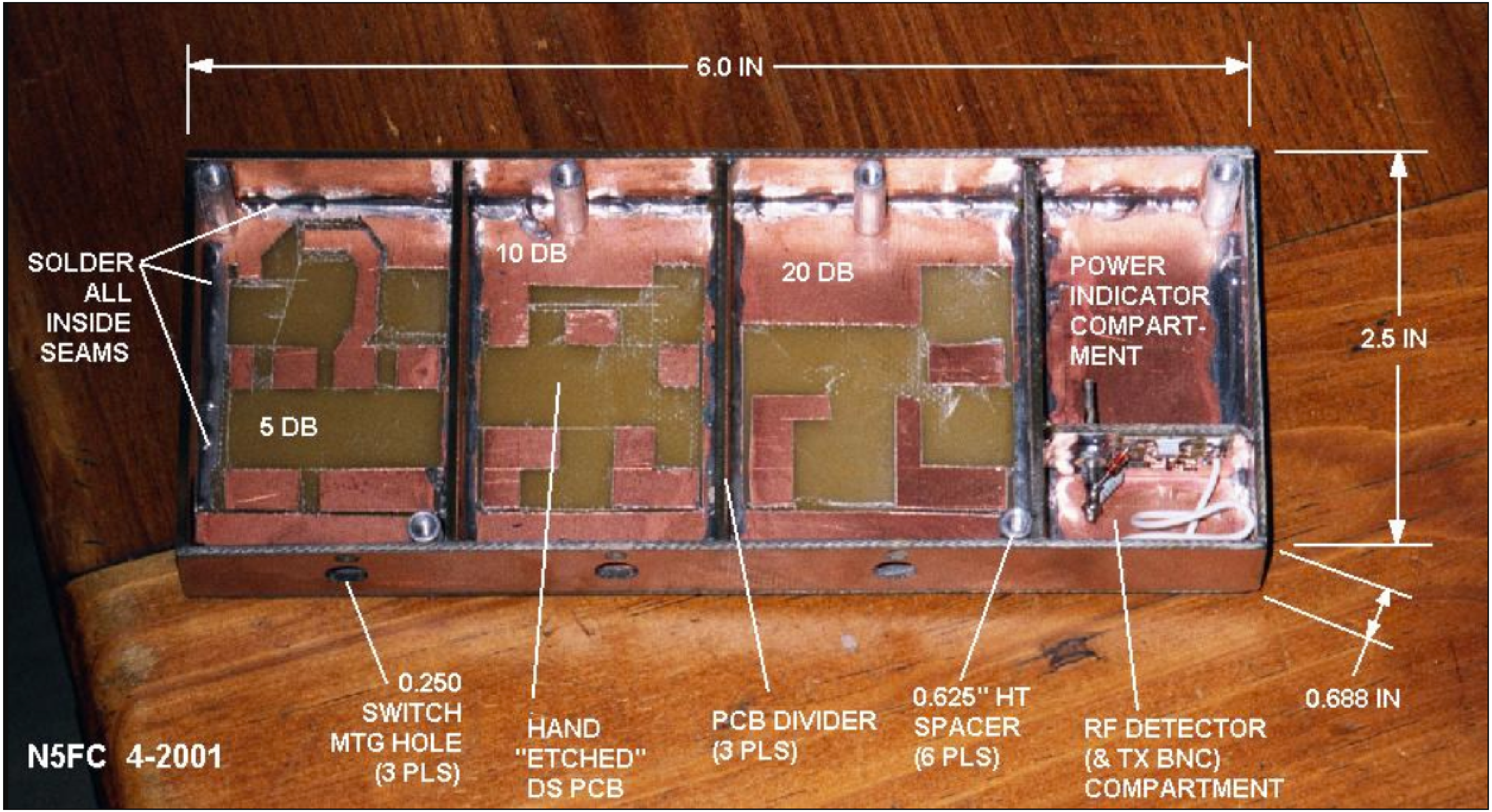
6.0 IN
SOLDER ALL INSIDE SEAMS
10 DB
20 DB
POWER INDICATOR COMPART-MENT
2.5 IN
5 DB
0.250 SWITCH MTG HOLE (3 PLS)
HAND "ETCHED" DS PCB
PCB DIVIDER (3 PLS)
0.625" HT SPACER (6 PLS)
RF DETECTOR (& TX BNC) COMPARTMENT
0.688 IN
N5FC 4-2001

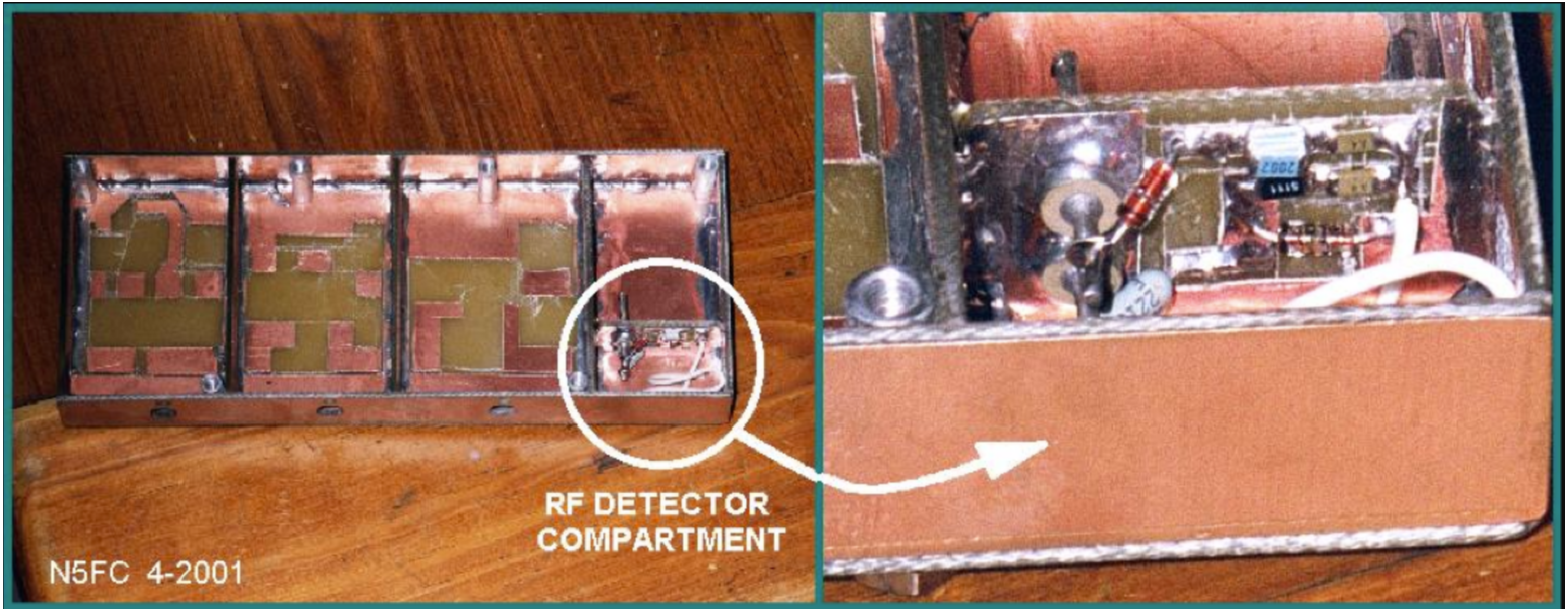
RF DETECTOR
COMPARTMENT
N5FC 4-2001

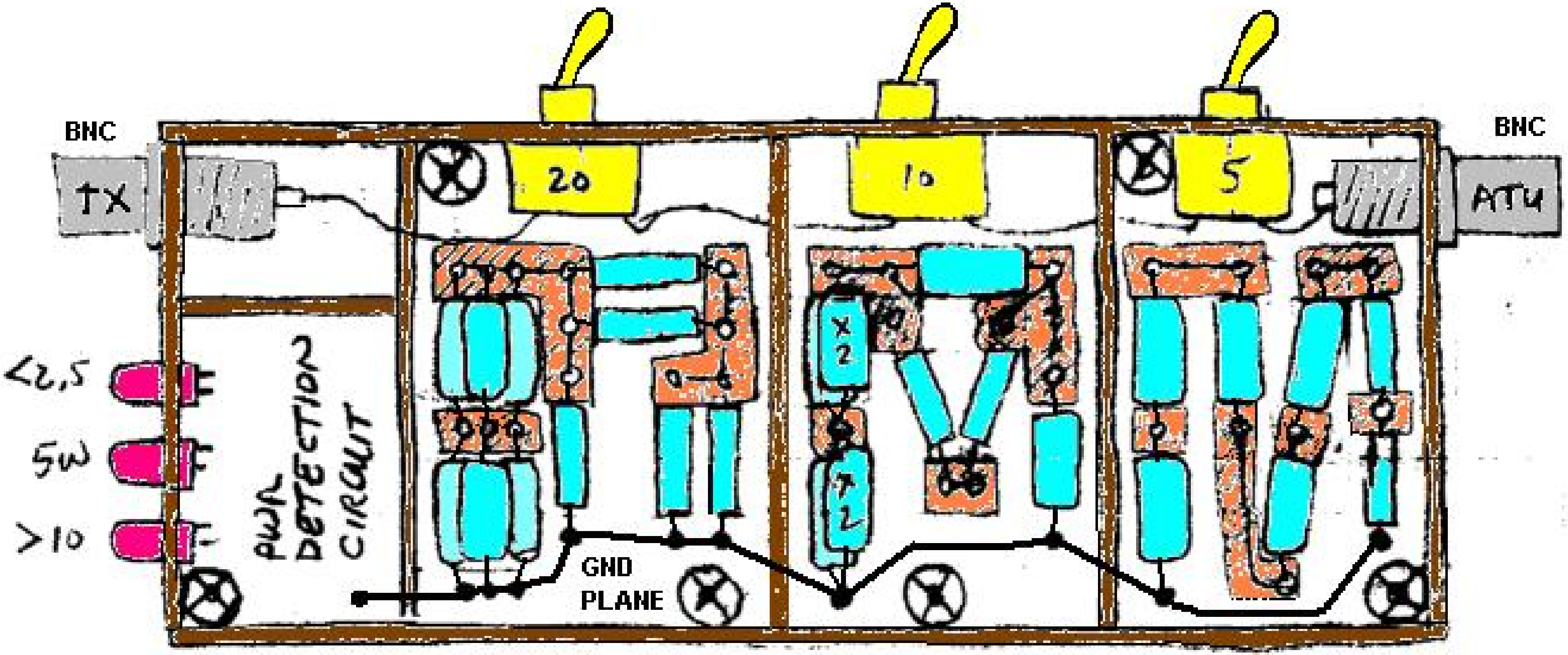
BNC
TX
BNC
ATU
20
10
5
<2,5
5W
>10
PWR DETECTION CIRCUIT
X2
X2
GND PLANE

N5FC 4-2001

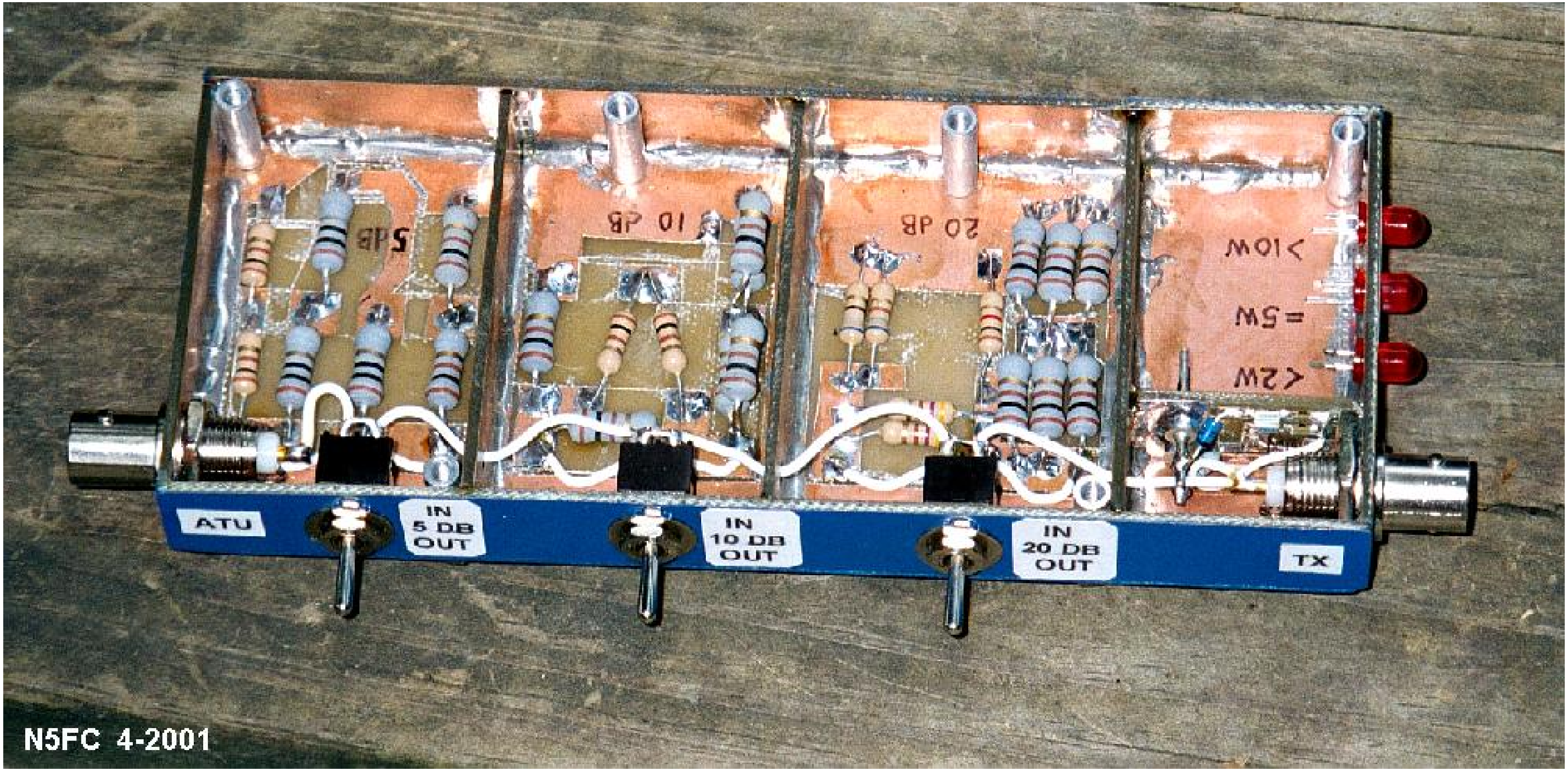
ATU
IN 5 DB OUT
IN 10 DB OUT
IN 20 DB OUT
TX
5dB
10 dB
20 dB
>10W
=5W
<2W

N5FC 4-2001

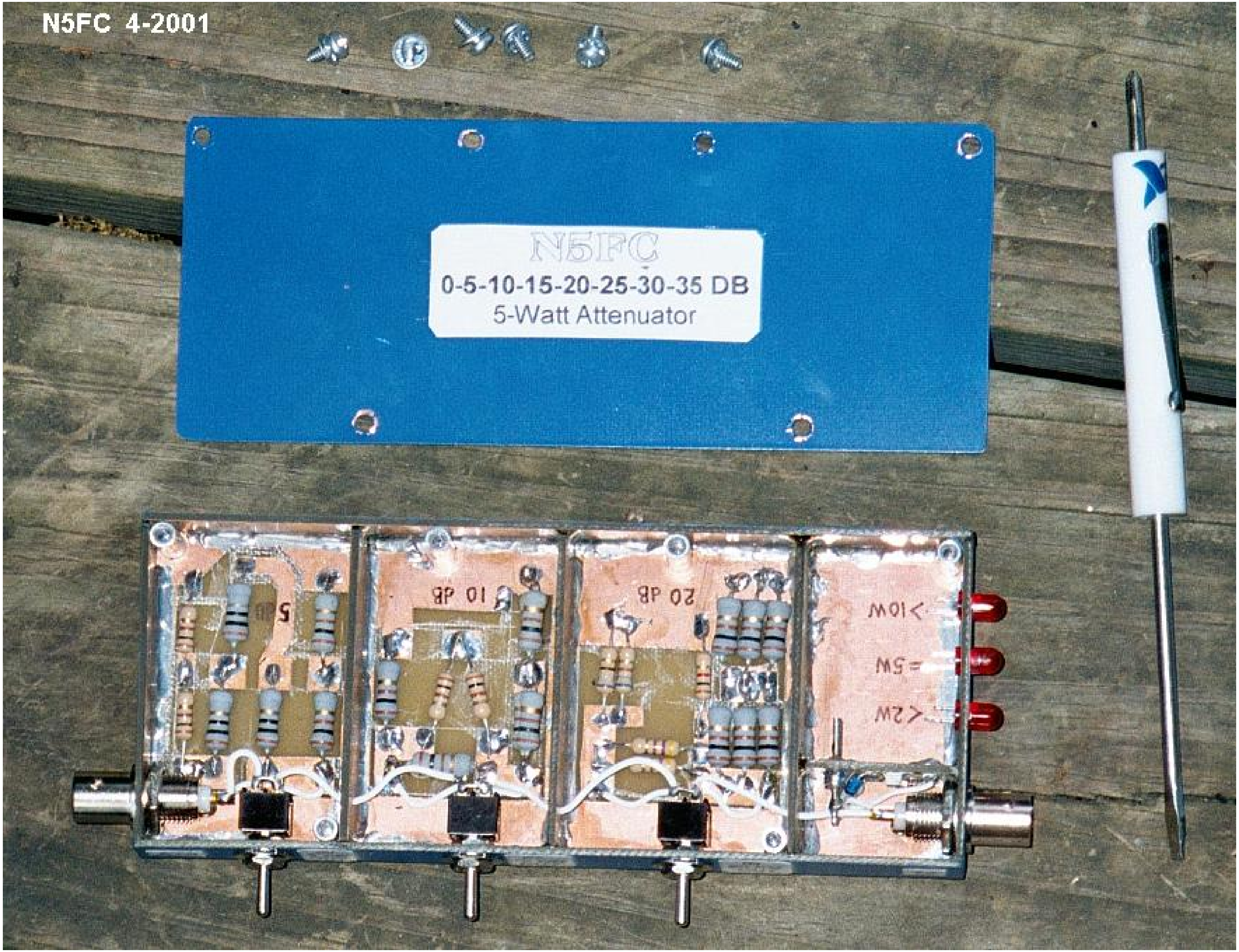
N5FC 4-2001

N5FC
0-5-10-15-20-25-30-35 DB
5-Watt Attenuator
5 dB
10 dB
20 dB
>10W
=5W
<2W

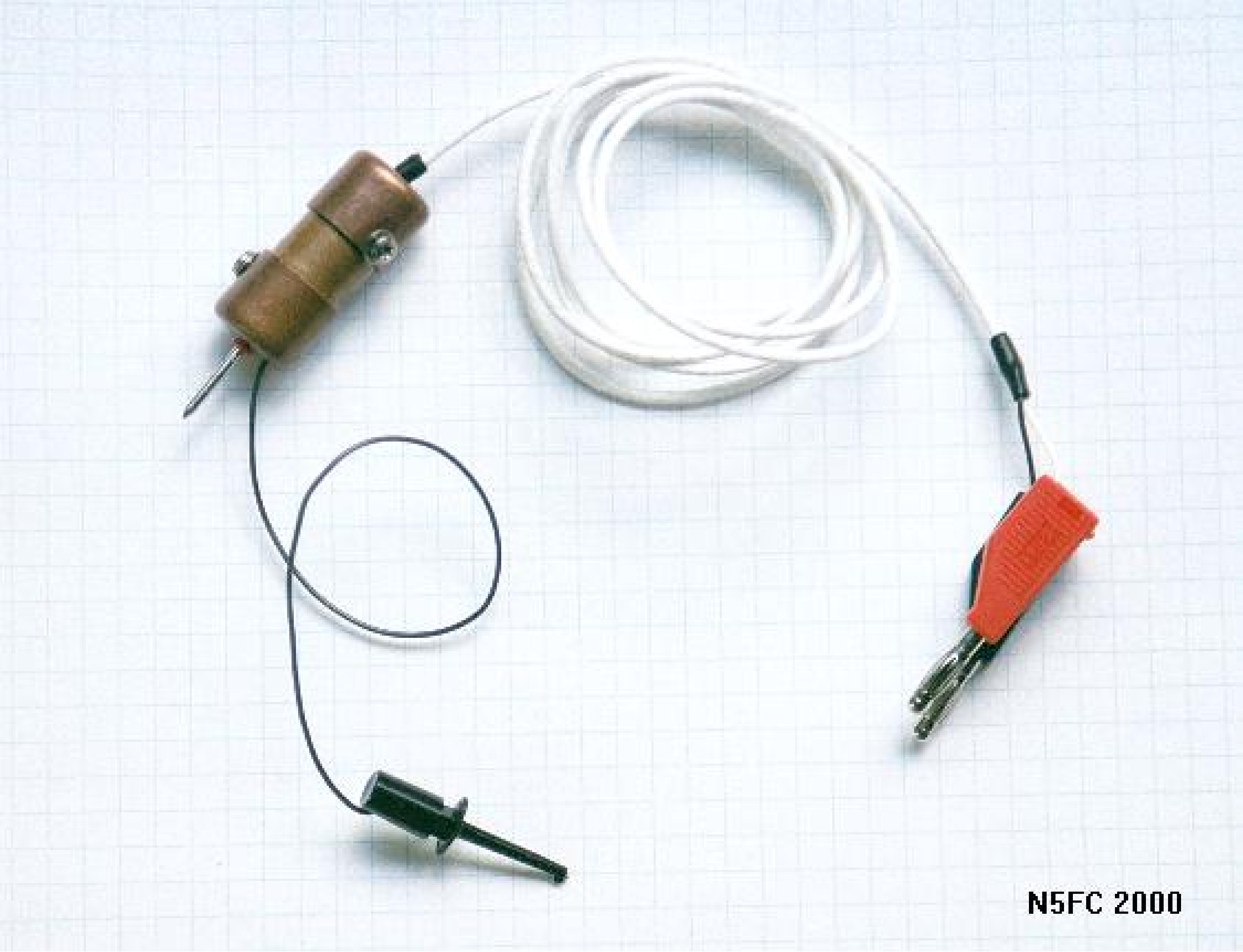

N5FC 2000

sic RF Probe. Simple, eh?

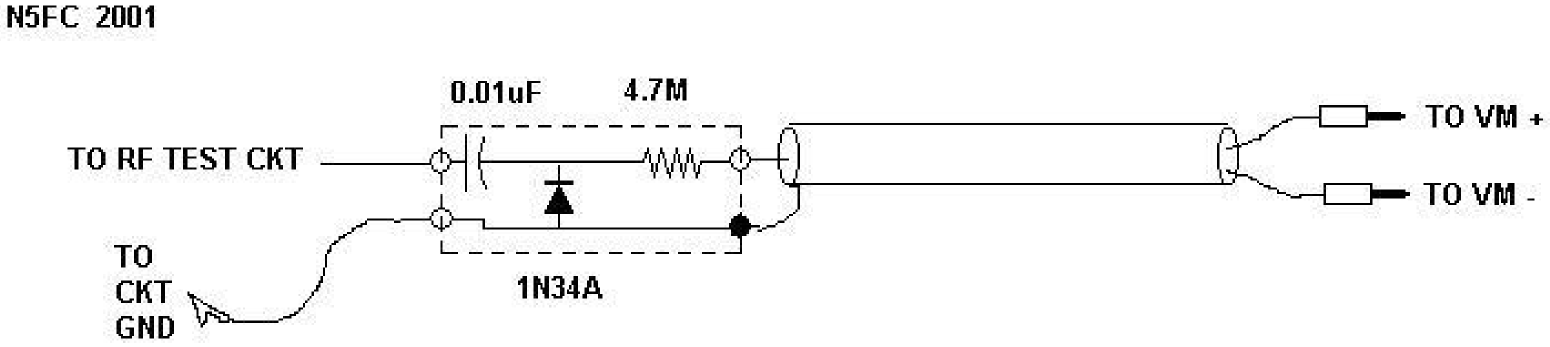
N5FC 2001
0.01uF
4.7M
TO RF TEST CKT
TO VM +
TO VM -
TO CKT GND
1N34A

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10-11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

oretical discussion that we'll make short note of. Obviously, for "probing" we need a "probe". (Hey! No wonder I get paid the big bucks...). We add a SH
obe goes to our test circuit, where we're probing. Brilliant! We don't want either of these to be long leads, because we're talking RF here, and long leads =

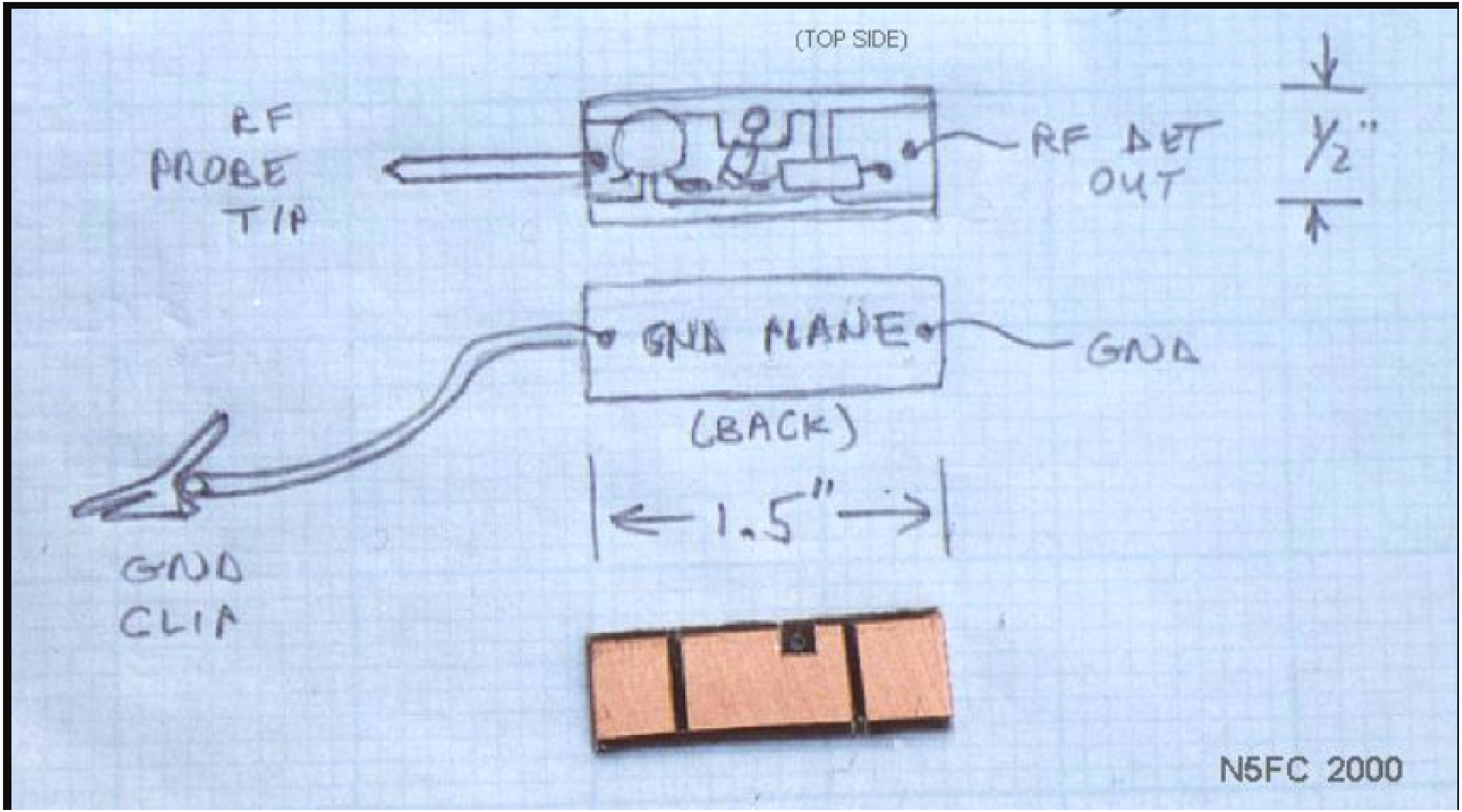
(TOP SIDE)

RF PROBE TIP
RF DET OUT
1/2"
GND PLANE
GND
(BACK)
1.5"
GND CLIP

N5FC 2000

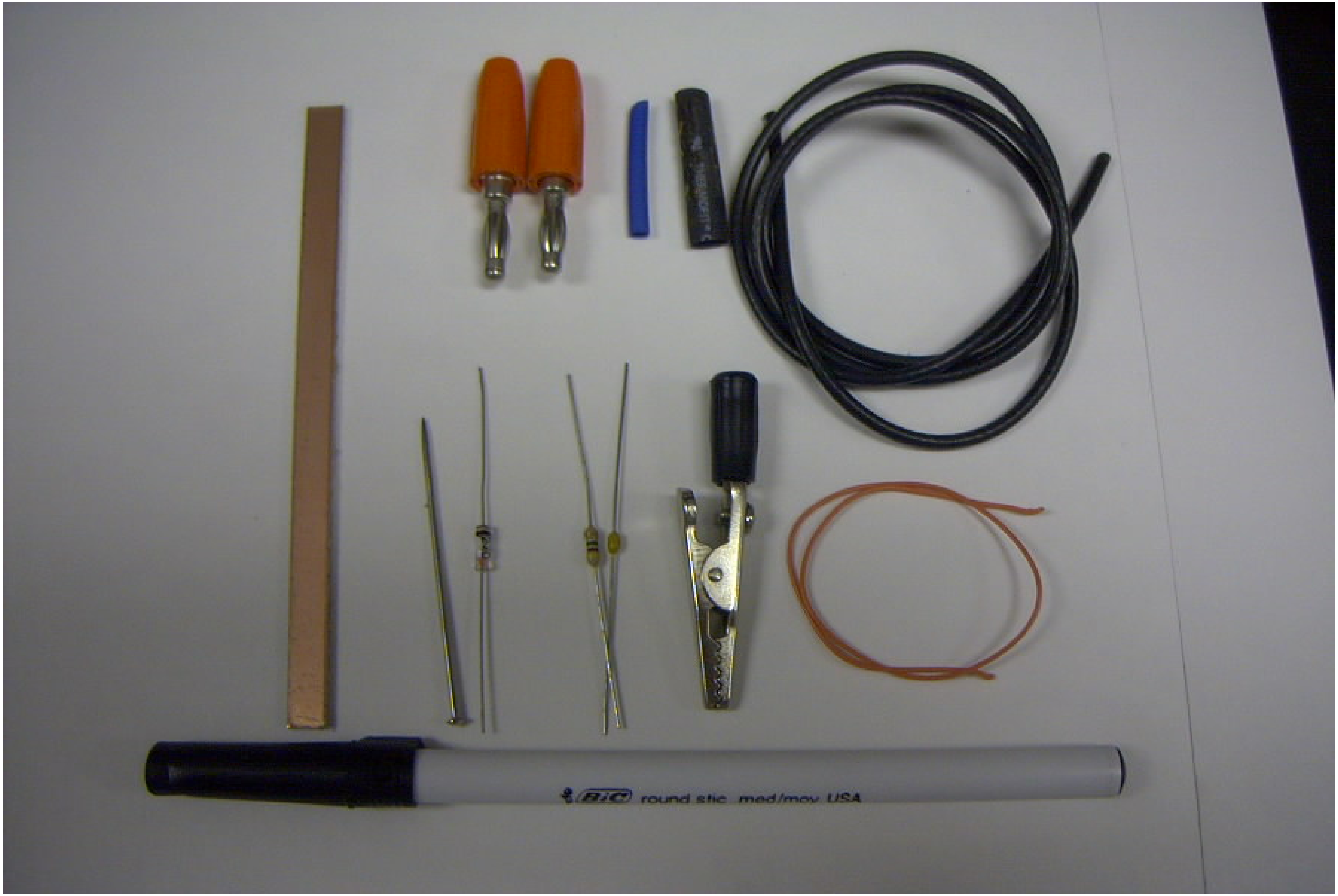
BIC round stic med/moy USA

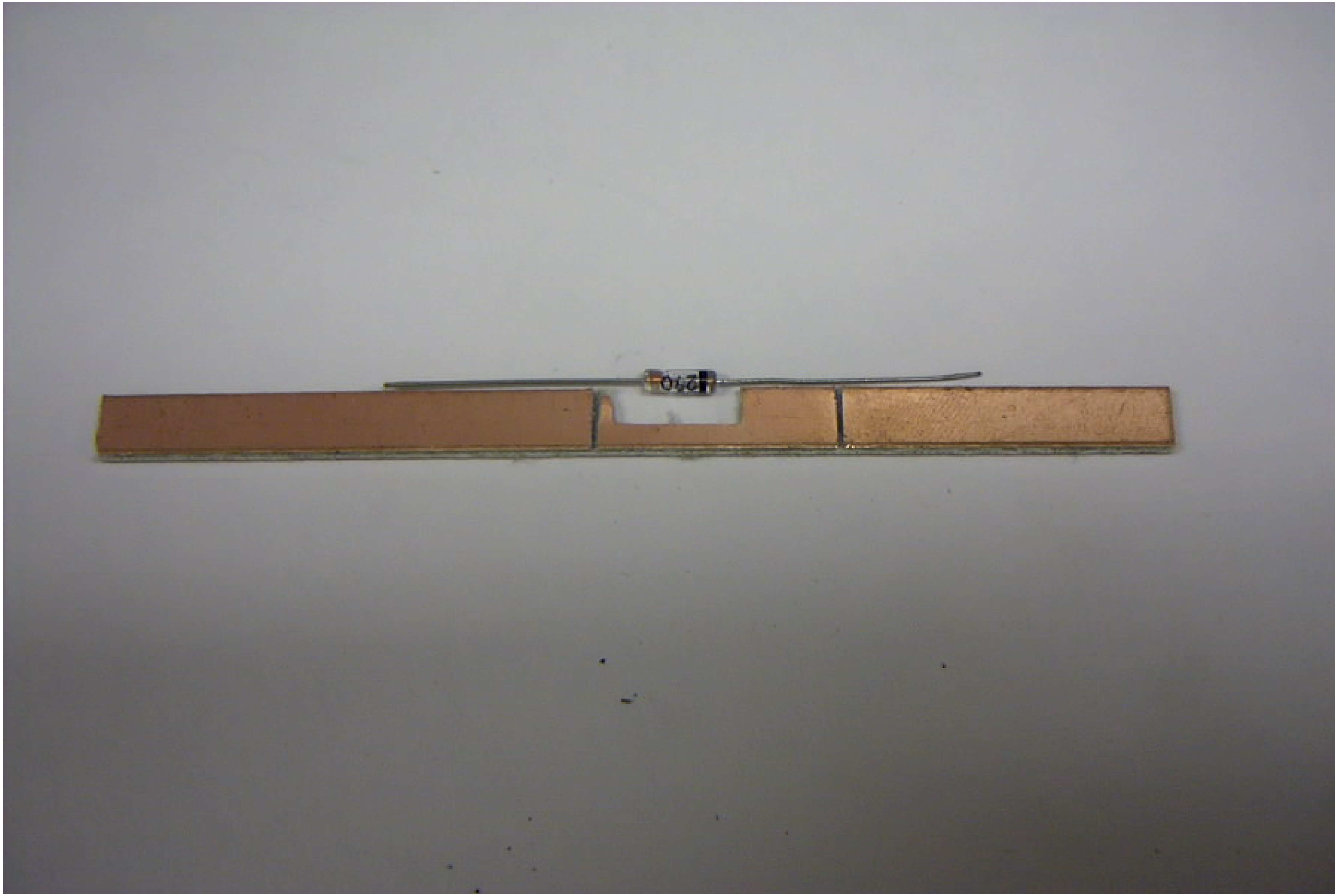
250

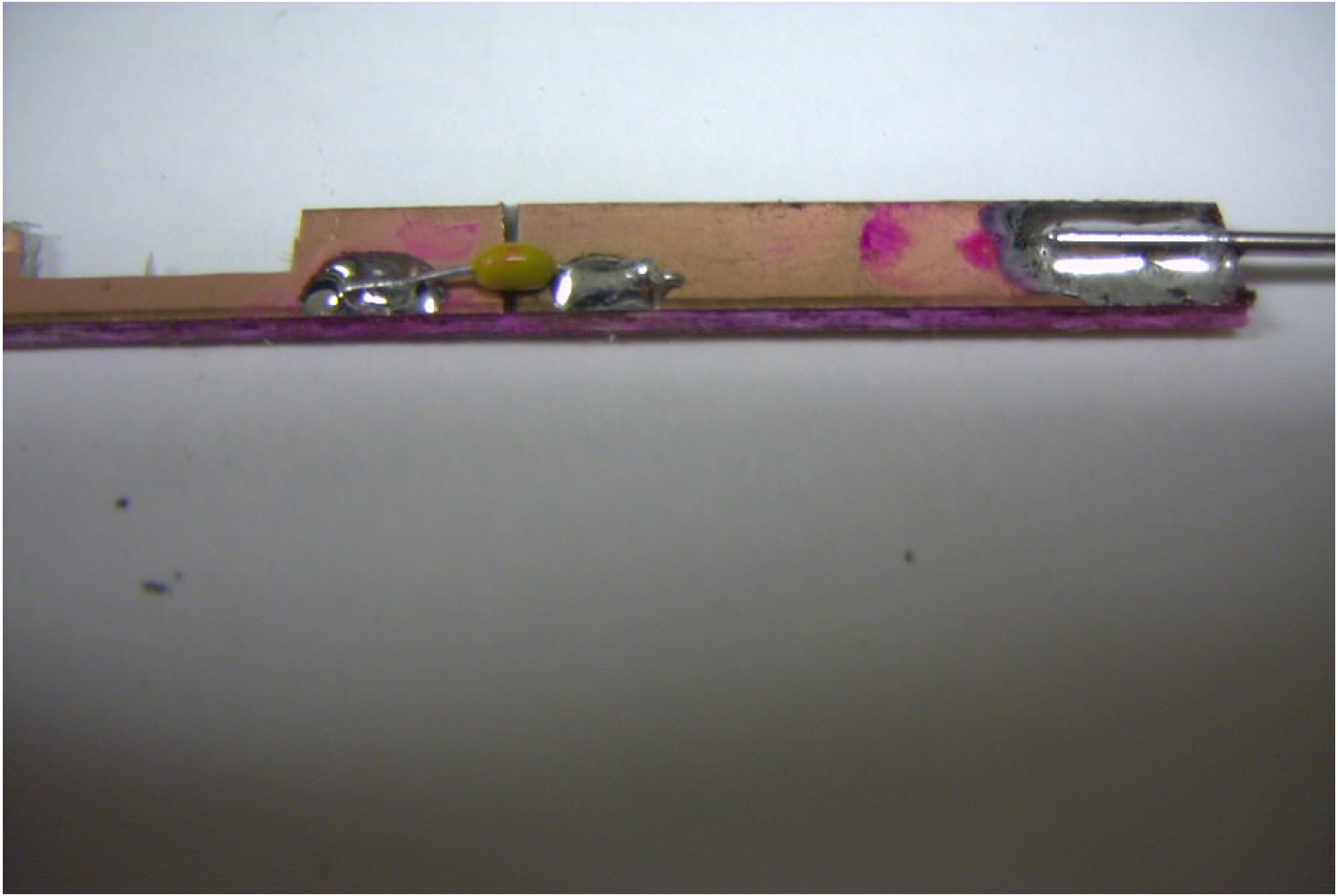

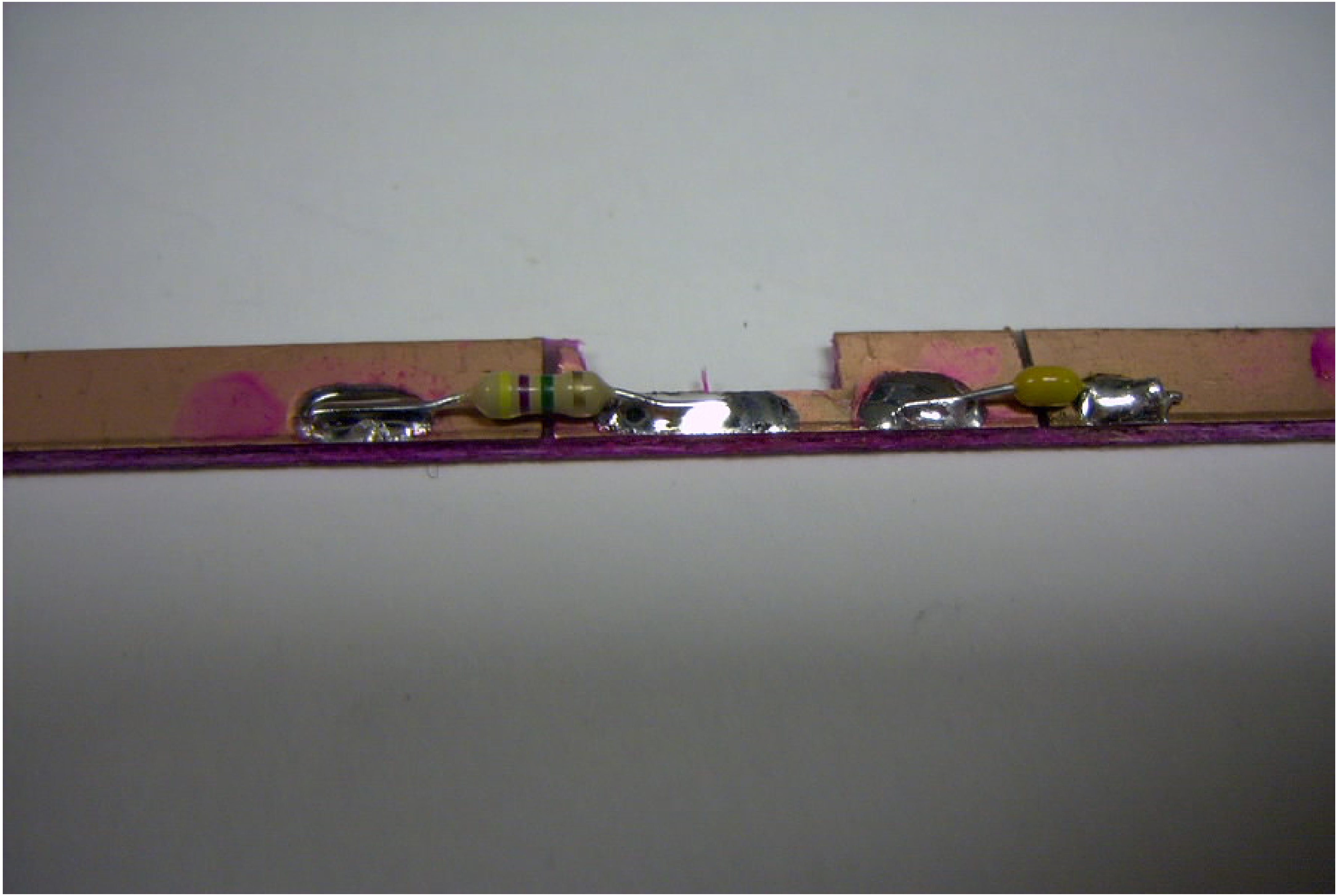

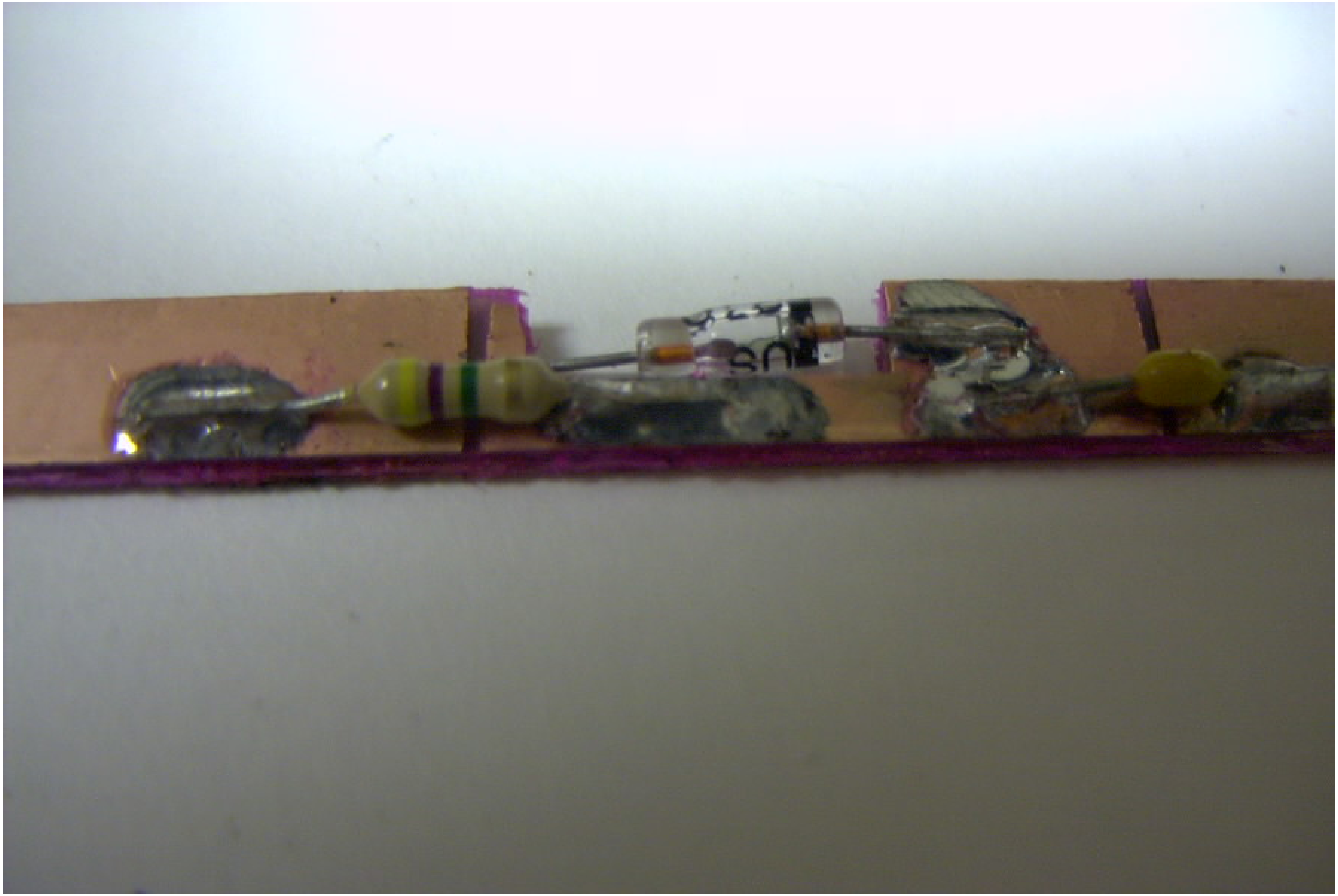

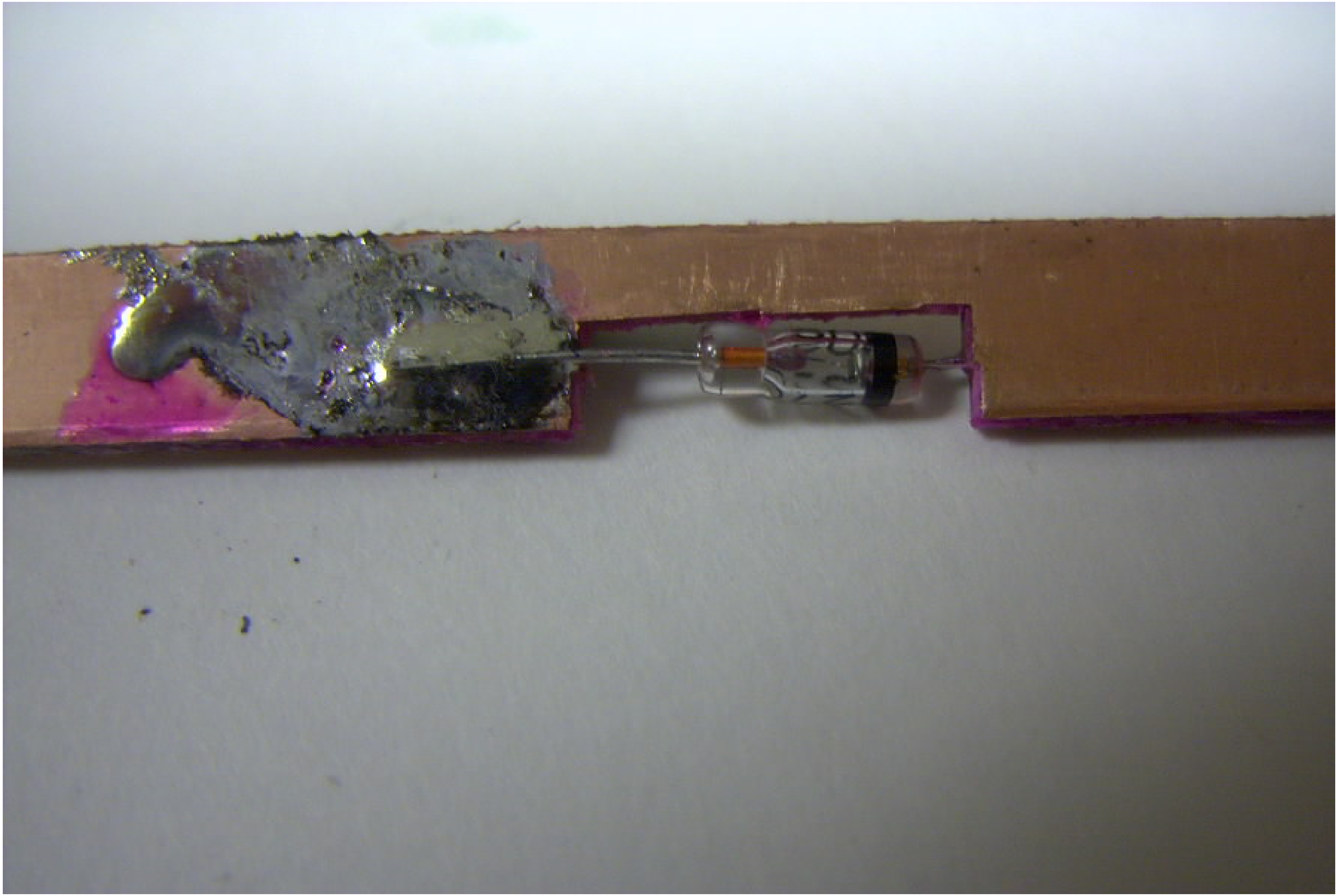

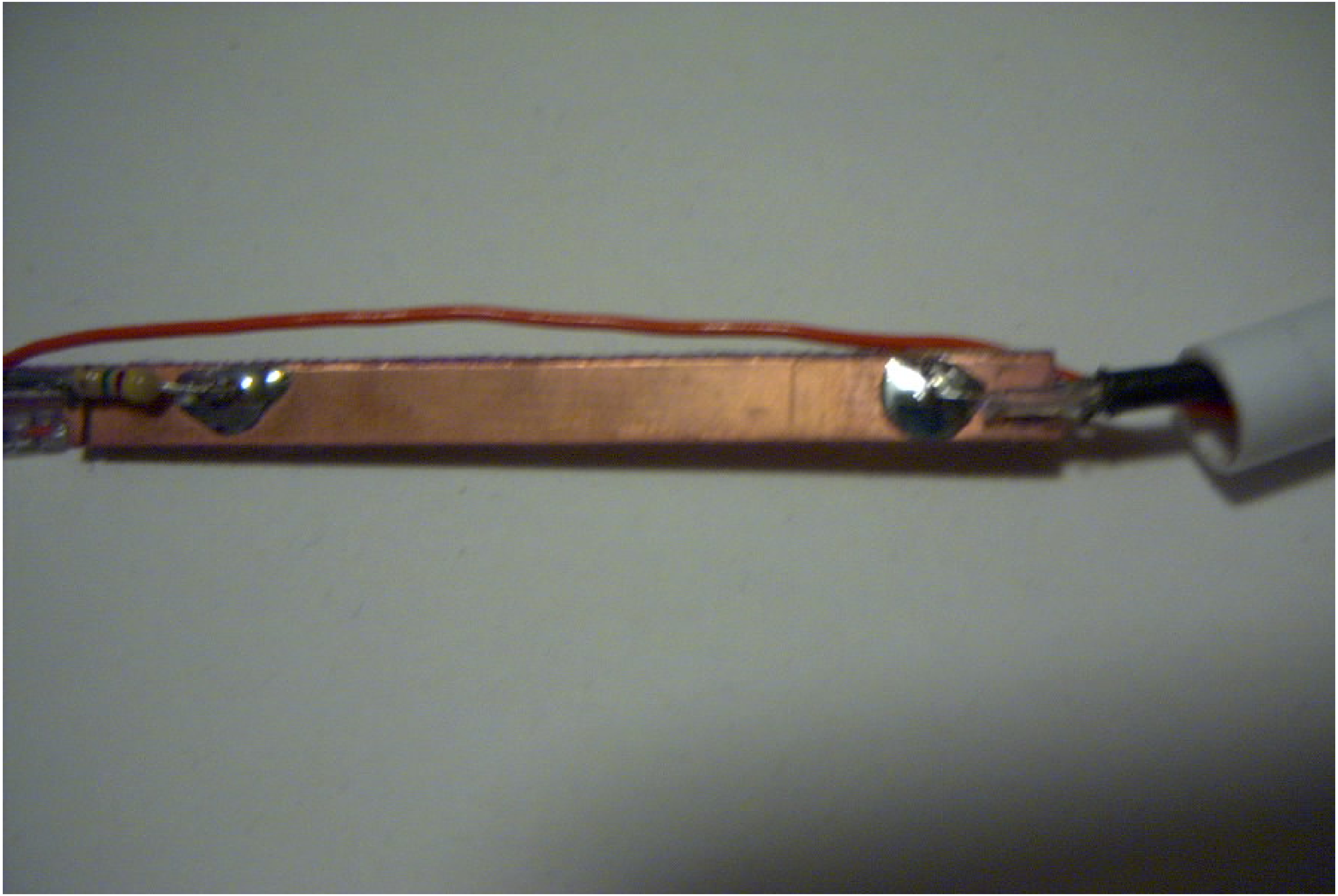

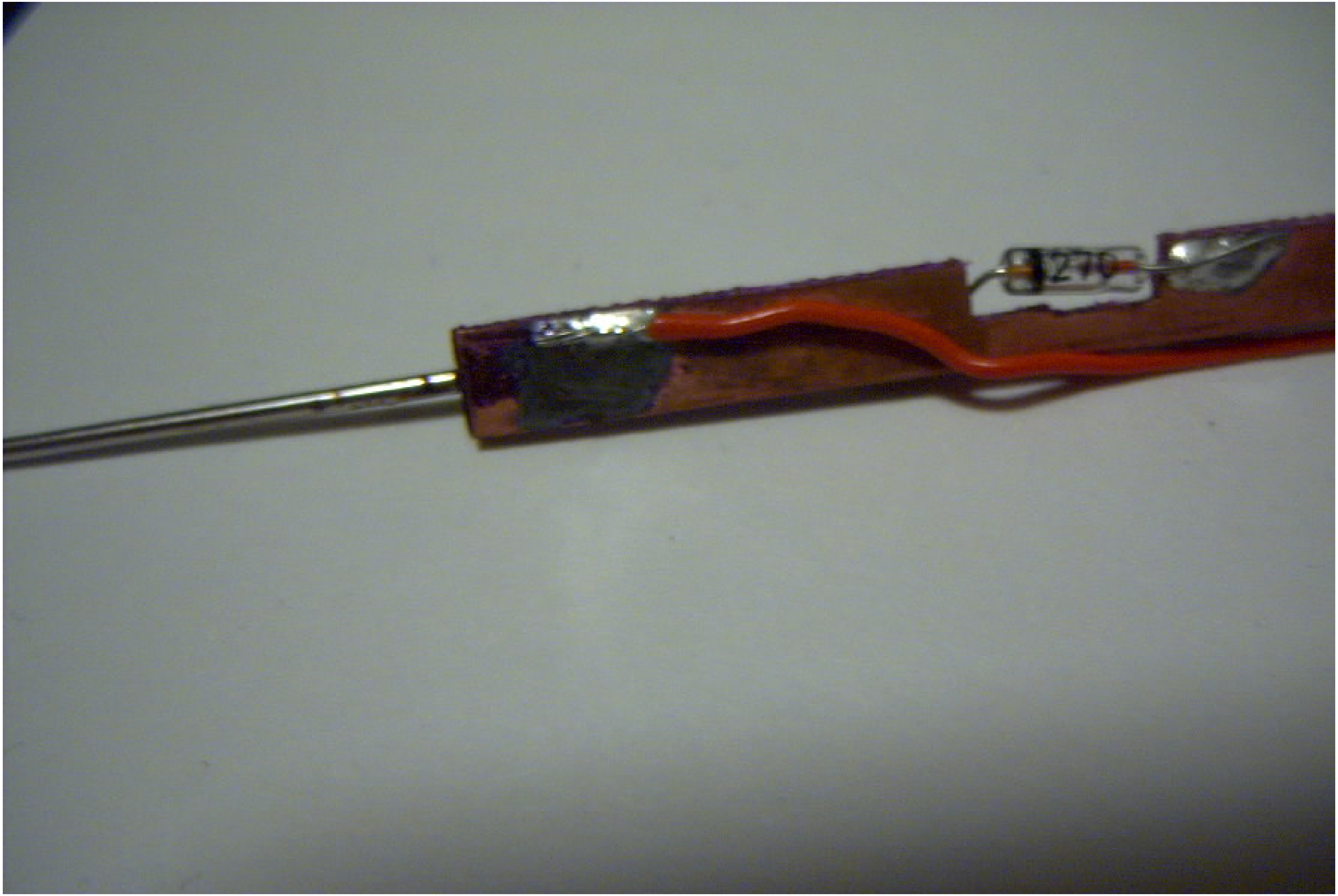

How to Build Your Own Oscilloscope Probes

Here is the complete bill of materials:

- The pen
- A 2-meter piece of coaxial test cable with a BNC connector on one end
- Epoxy adhesive
- One alligator clip
- Copper-plated nail — 0.75" (20mm) long, packed as "weather-stripping nail".
- 1 MΩ and 5 MΩ resistors

GB
Alligator Clips
Quick Set
Epoxy
Epoxy

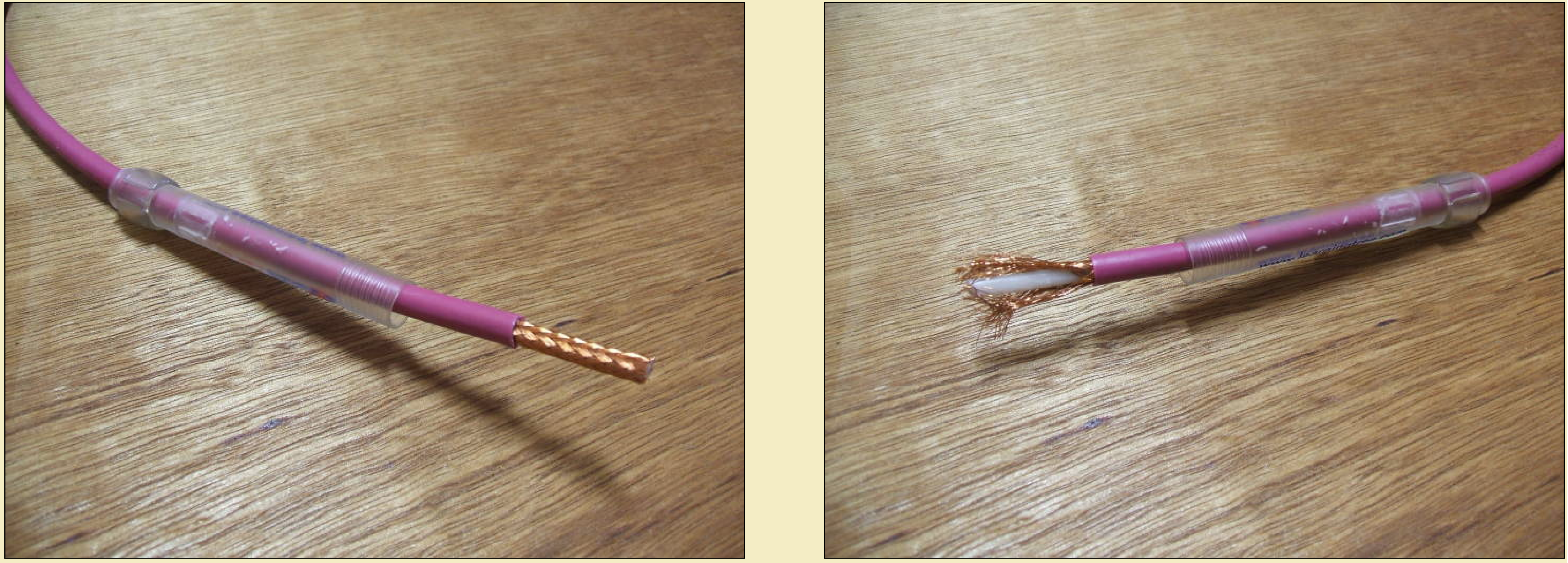

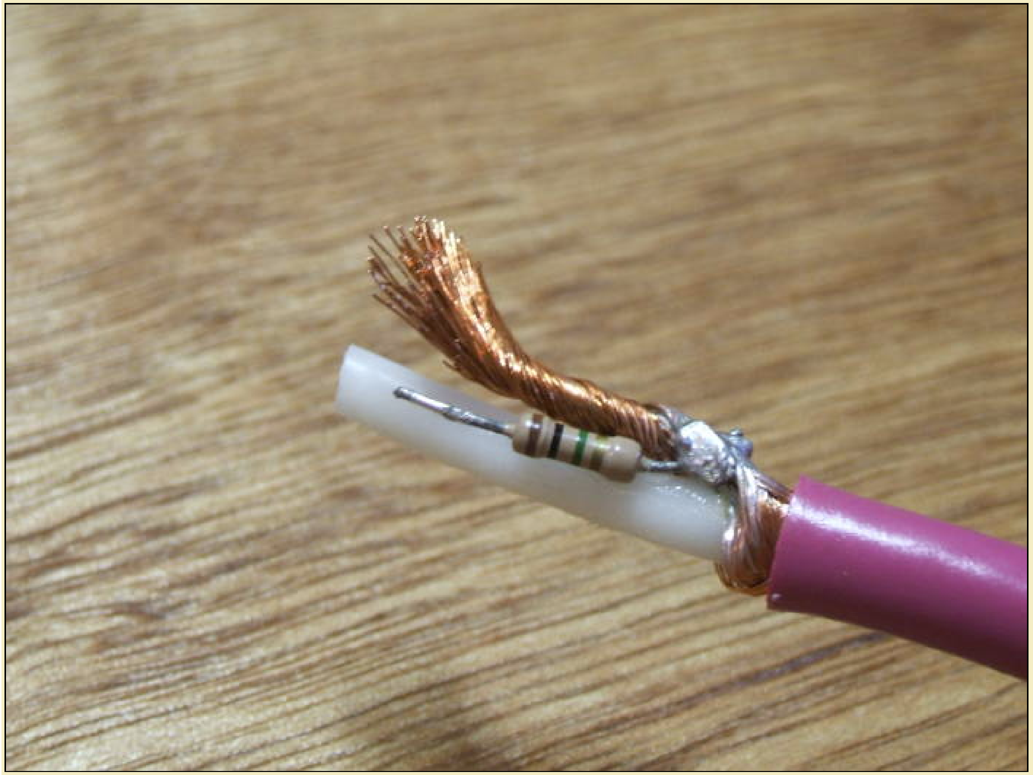

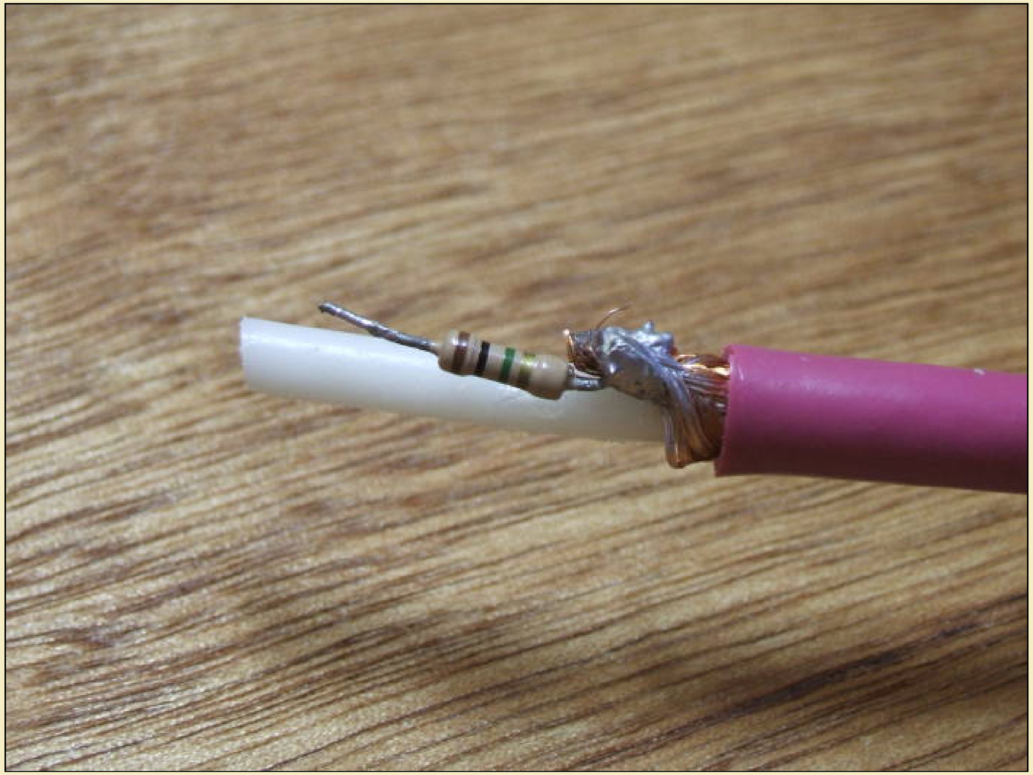

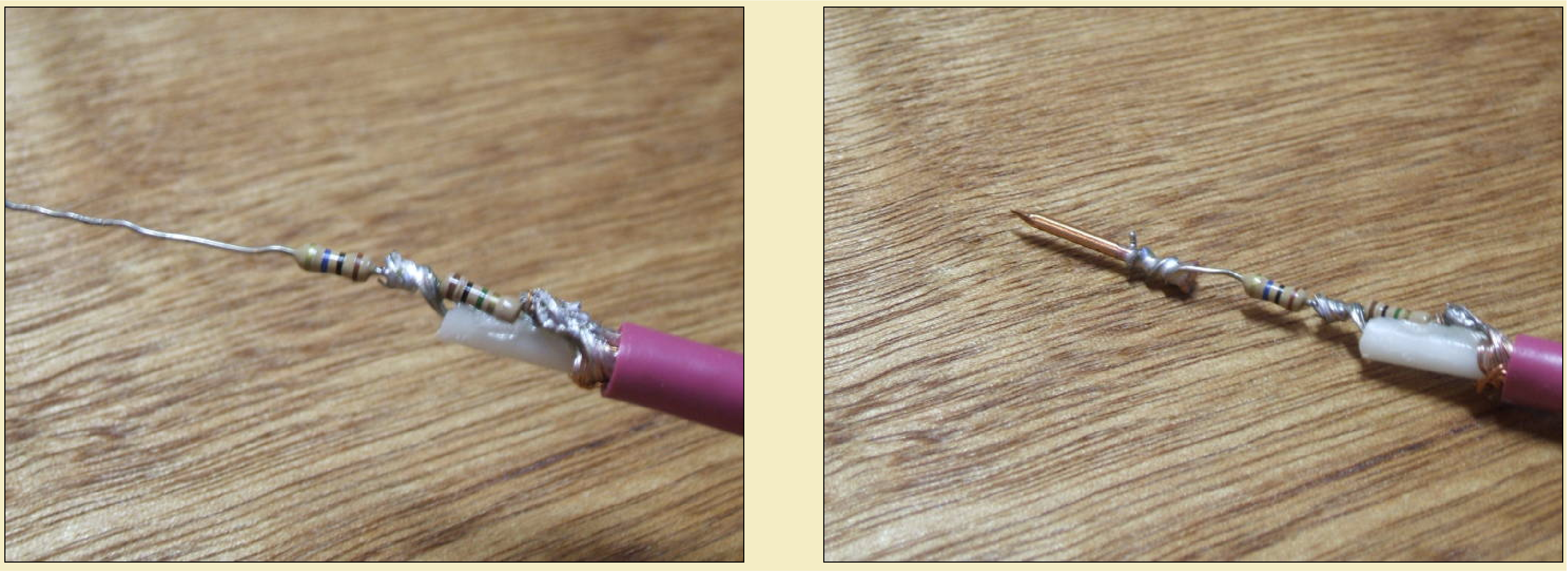

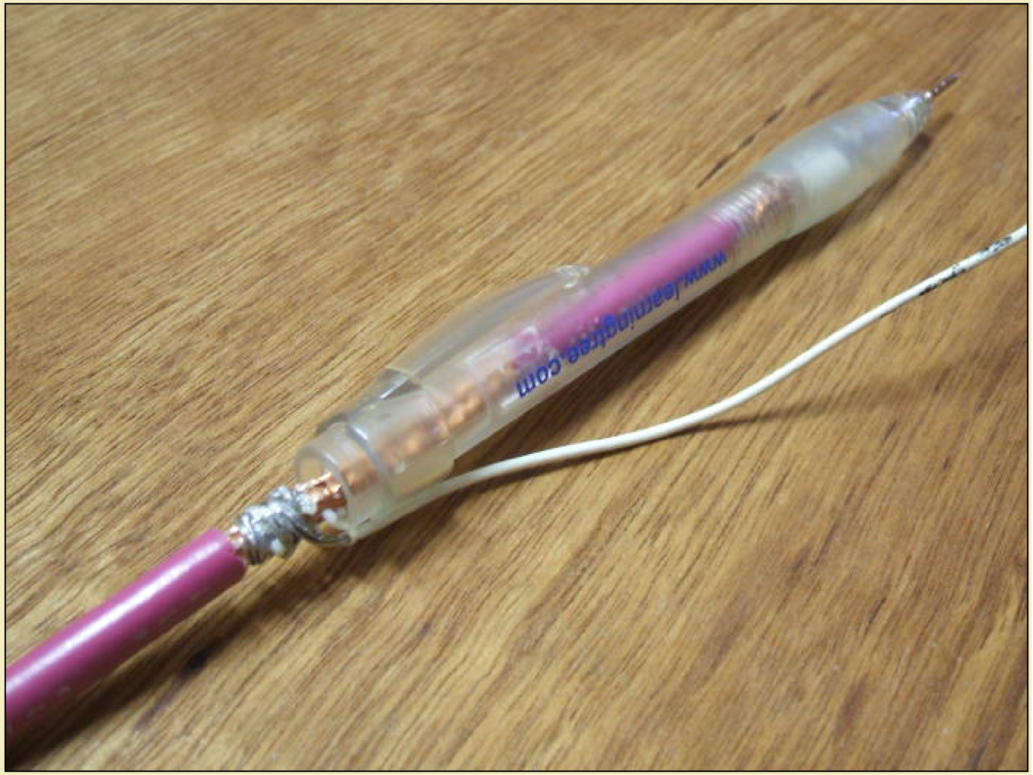
www.learningtree.com

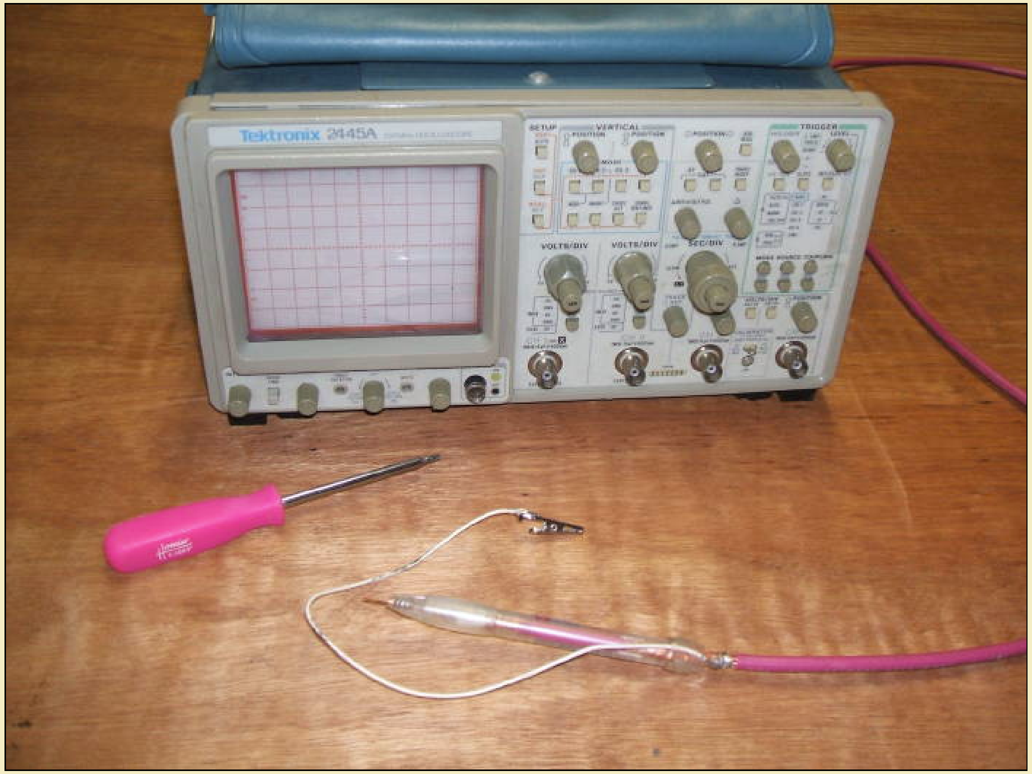
Tektronix 2445A
VERTICAL
TRIGGER
VOLTS/DIV
VOLTS/DIV
SEC/DIV

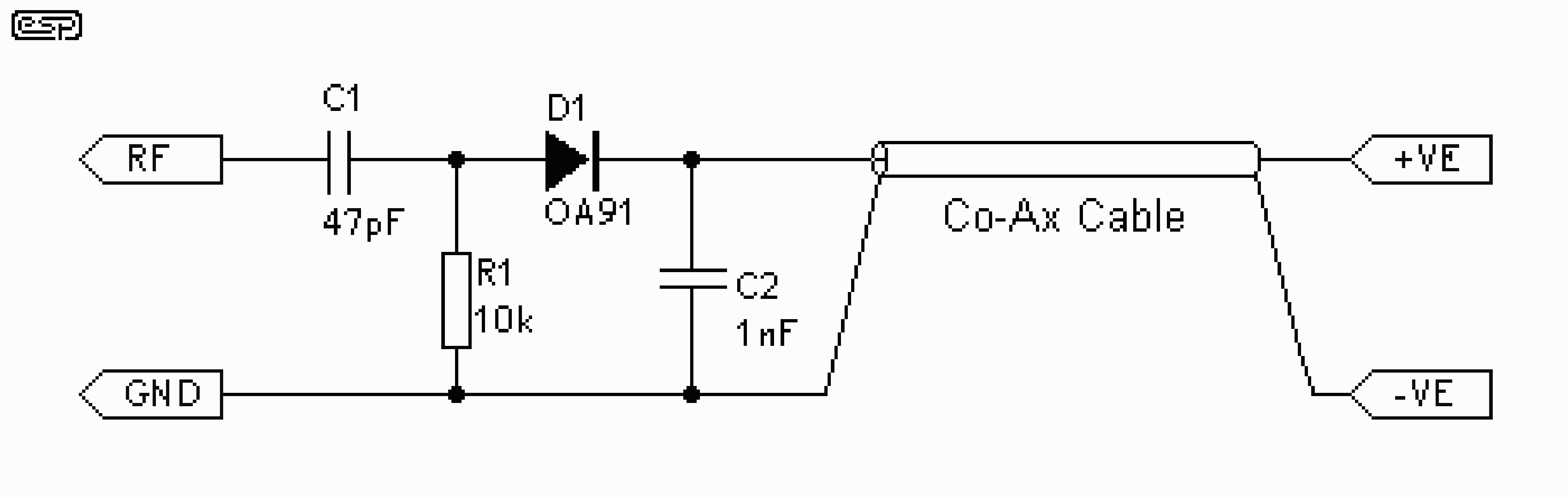
RF
C1
47pF
D1
OA91
R1
10k
C2
1nF
GND
Co-Ax Cable
+VE
-VE

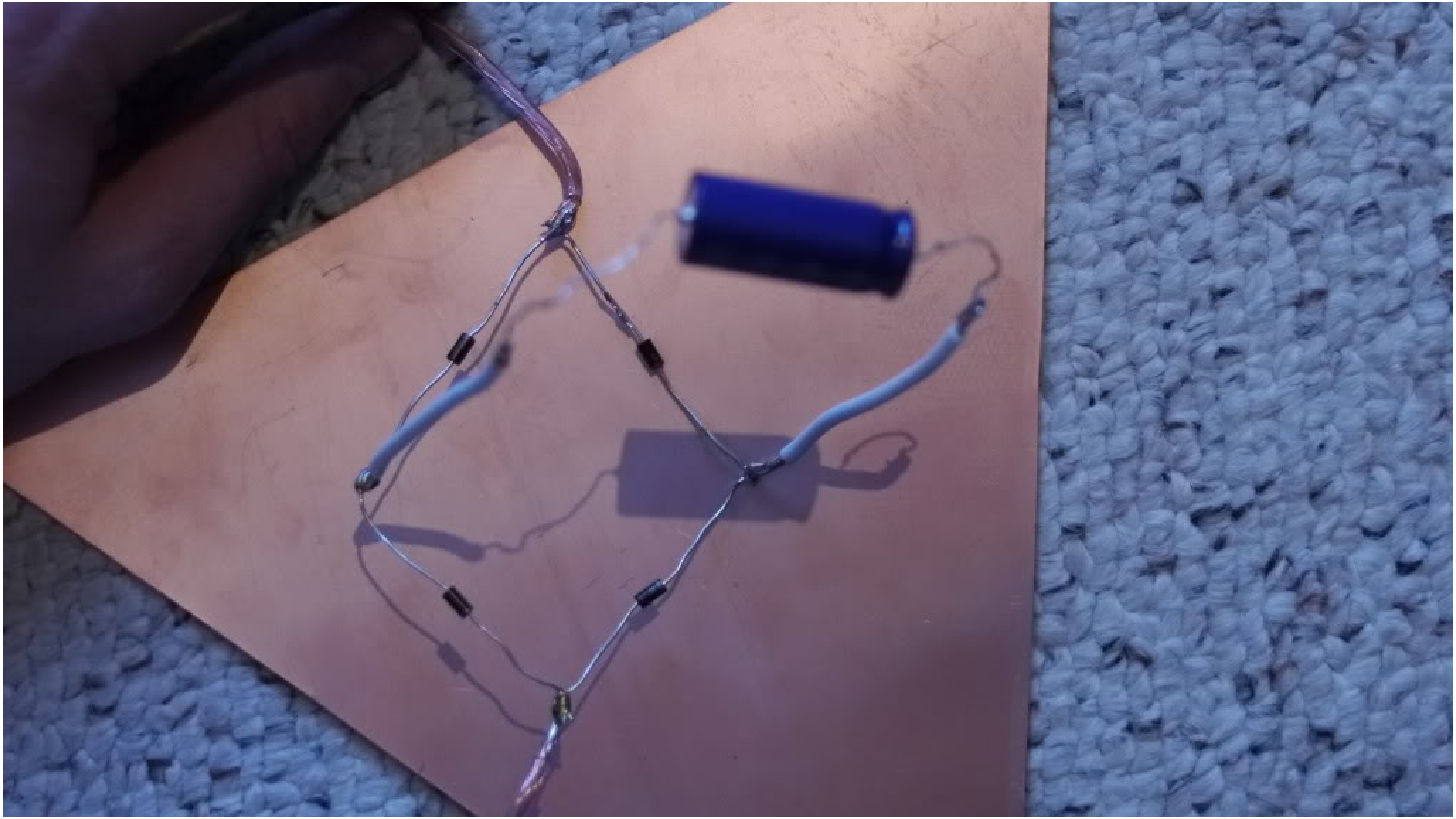

The simplest power meter

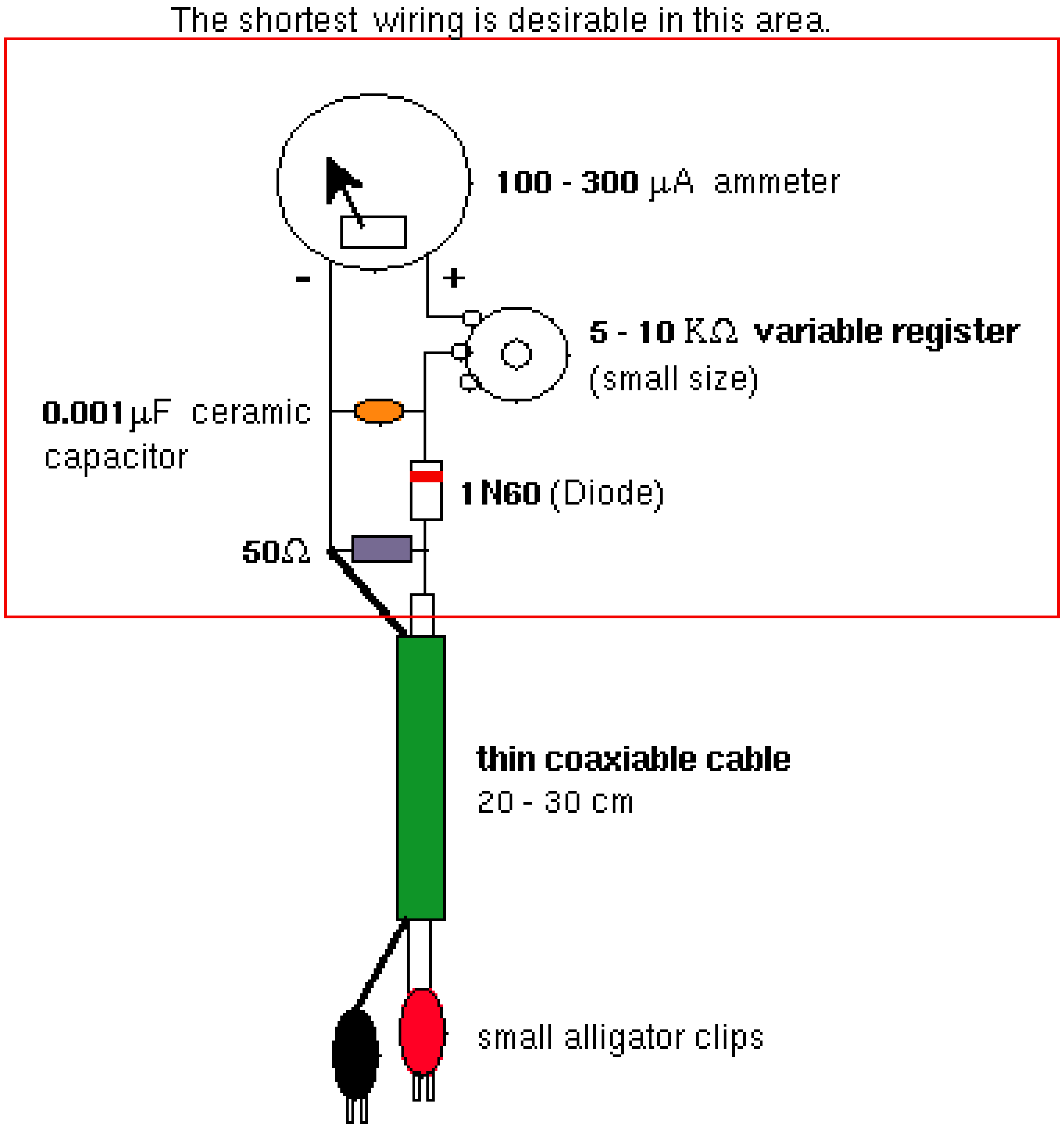
The shortest wiring is desirable in this area.
-
+
100 - 300 µA ammeter
5 - 10 KΩ **variable register**
(small size)
0.001µF ceramic
capacitor
1N60 (Diode)
50Ω
thin coaxiable cable
20 - 30 cm
small alligator clips

As for the ammeter, you can use an used one taking from junked audio amplifier, tape-recorder, radio-cassette, and so on.

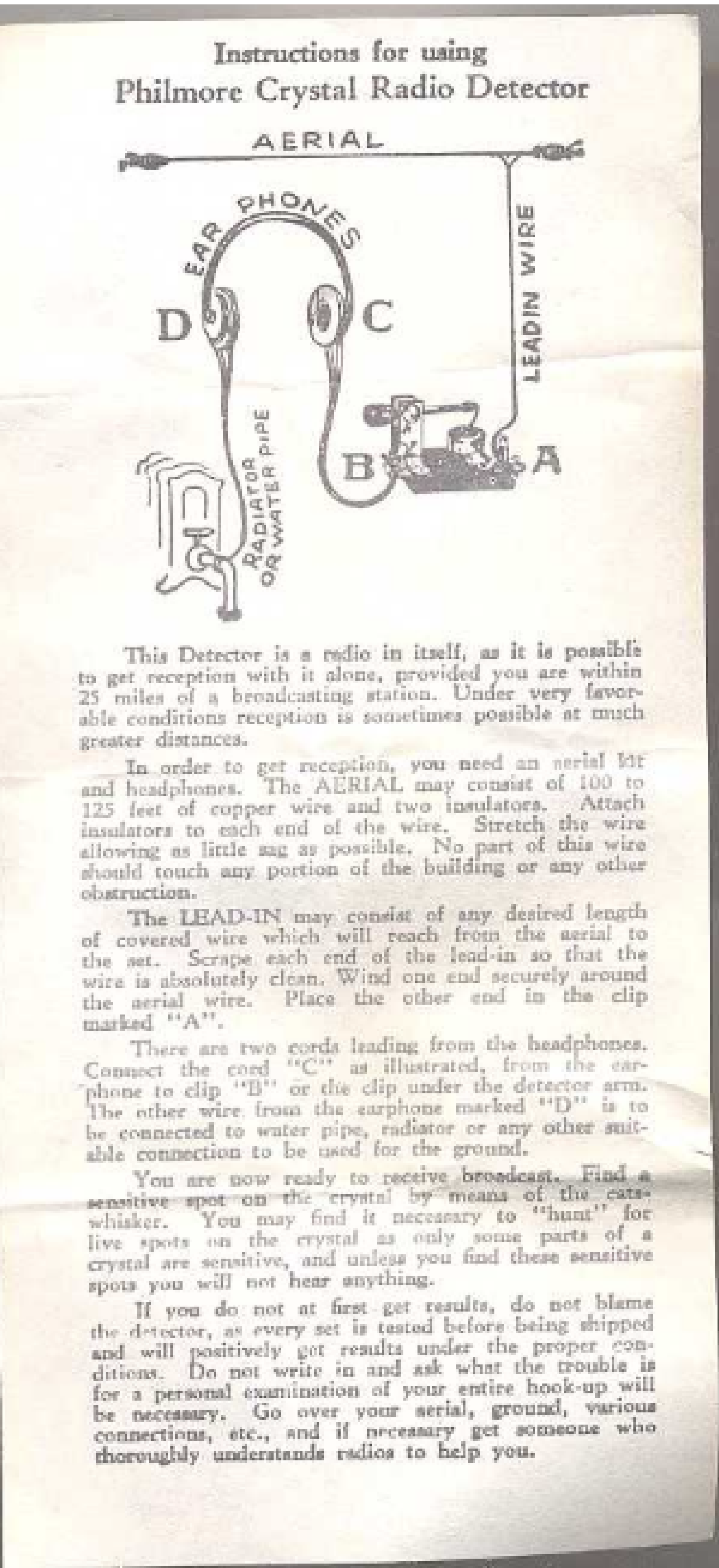

Instructions for using Philmore Crystal Radio Detector

AERIAL
EAR PHONES
LEADIN WIRE
D
C
B
A
RADIATOR OR WATER PIPE

This Detector is a radio in itself, as it is possible to get reception with it alone, provided you are within 25 miles of a broadcasting station. Under very favorable conditions reception is sometimes possible at much greater distances.

In order to get reception, you need an aerial kit and headphones. The AERIAL may consist of 100 to 125 feet of copper wire and two insulators. Attach insulators to each end of the wire. Stretch the wire allowing as little sag as possible. No part of this wire should touch any portion of the building or any other obstruction.

The LEAD-IN may consist of any desired length of covered wire which will reach from the aerial to the set. Scrape each end of the lead-in so that the wire is absolutely clean. Wind one end securely around the aerial wire. Place the other end in the clip marked "A".

There are two cords leading from the headphones. Connect the cord "C" as illustrated, from the earphone to clip "B" or the clip under the detector arm. The other wire from the earphone marked "D" is to be connected to water pipe, radiator or any other suitable connection to be used for the ground.

You are now ready to receive broadcast. Find a sensitive spot on the crystal by means of the catswhisker. You may find it necessary to "hunt" for live spots on the crystal as only some parts of a crystal are sensitive, and unless you find these sensitive spots you will not hear anything.

If you do not at first get results, do not blame the detector, as every set is tested before being shipped and will positively get results under the proper conditions. Do not write in and ask what the trouble is for a personal examination of your entire hook-up will be necessary. Go over your aerial, ground, various connections, etc., and if necessary get someone who thoroughly understands radios to help you.

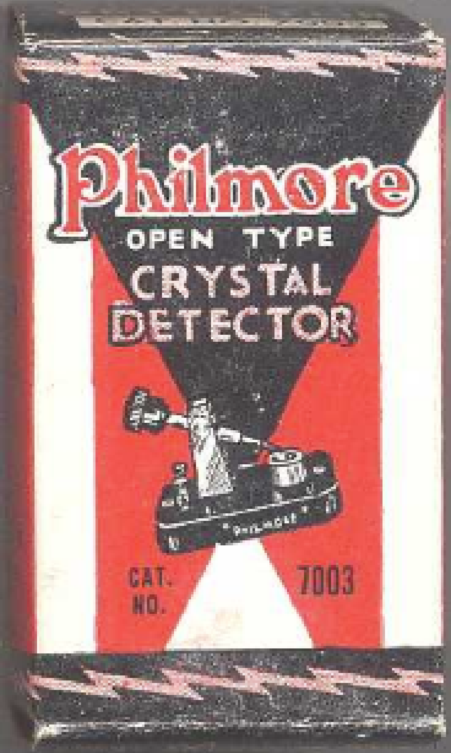

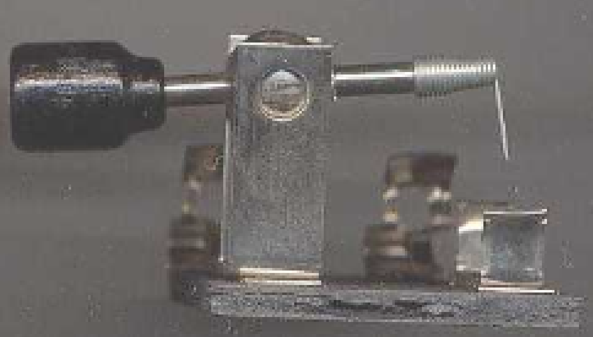

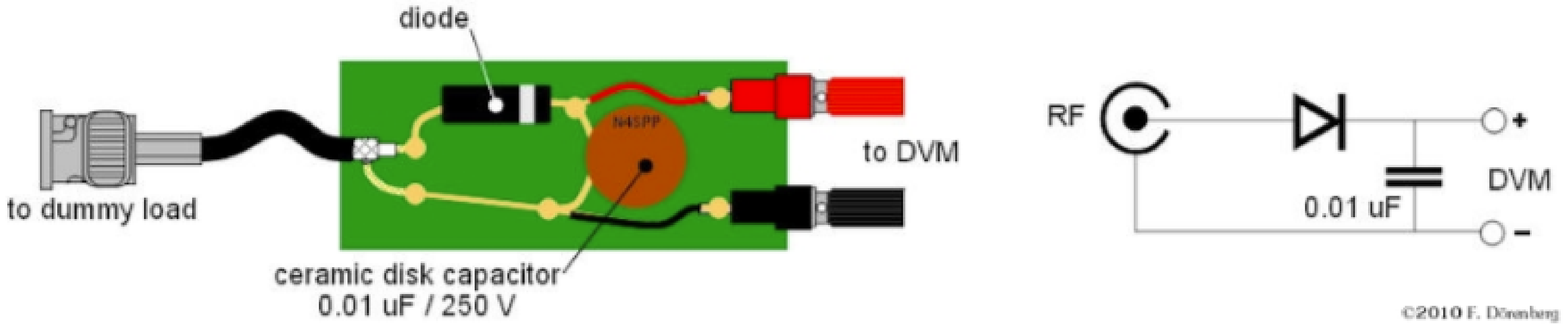
diode
to dummy load
ceramic disk capacitor
0.01 uF / 250 V
N4SPP
to DVM
RF
0.01 uF
DVM
+
–

©2010 F. Dörrenberg

A simple RF-probe

Obviously, this circuit will be fooled by a DC-offset on the RF signal. We can fix this by swapping the diode and the capacitor. Note that this is not necessary if you measure an RF voltage via a transformer, such as a directional coupler.

We can also make life a little easier by including a voltage divider with a scaling factor that is equal to the reciprocal of √2. Then the output voltage will be the RMS value that we are interested in. We can make a voltage divider where one resistor is the input impedance of the DVM. My DVM has a published input resistance of 10 MΩ. The second resistor should be 4M14 Ω,since 10 / (10+4.14) = 1 / √2). So 3M9 + 220k = 4M12 would be a good choice. This approach is shown below. Note that the resistor should be non-inductive (e.g., bulk-metal-foil or carbon).

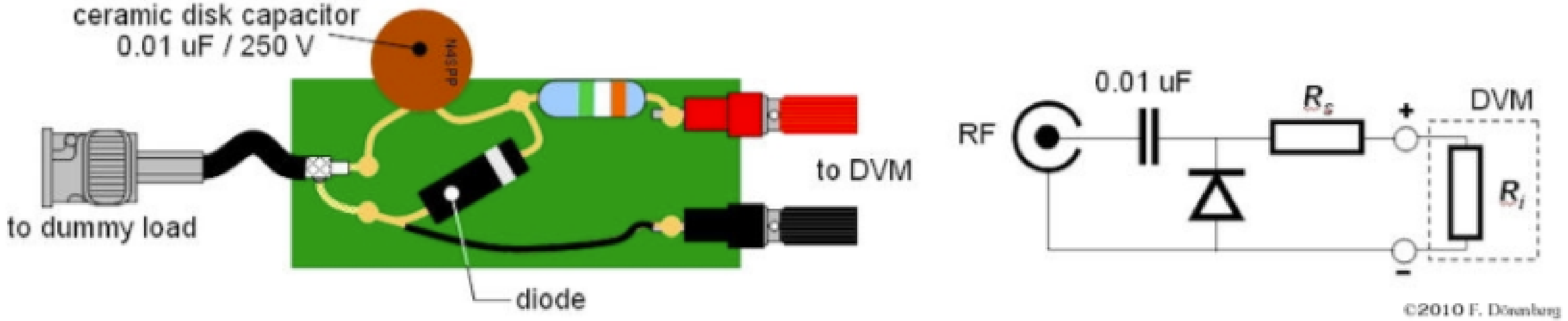
ceramic disk capacitor
0.01 uF / 250 V
N4SPP
to dummy load
to DVM
diode
RF
0.01 uF
R_s
+
DVM
R_i
–

©2010 F. Dörrenberg

RF-probe with DC-block and peak-to-RMS scaling

A variation on this, with a full-wave rectifier, is shown below:

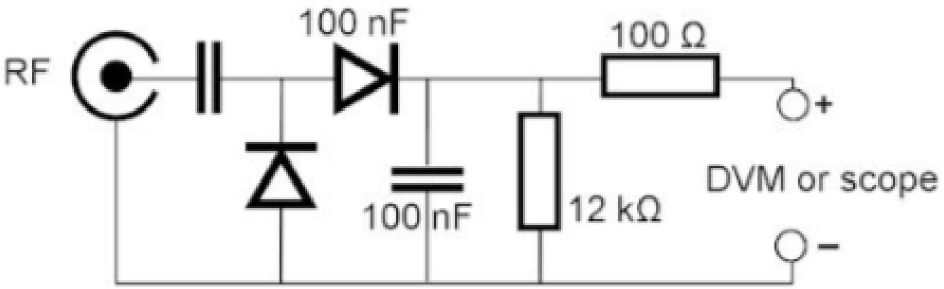
RF
100 nF
100 Ω
+
100 nF
12 kΩ
DVM or scope
–

RF Current Probe

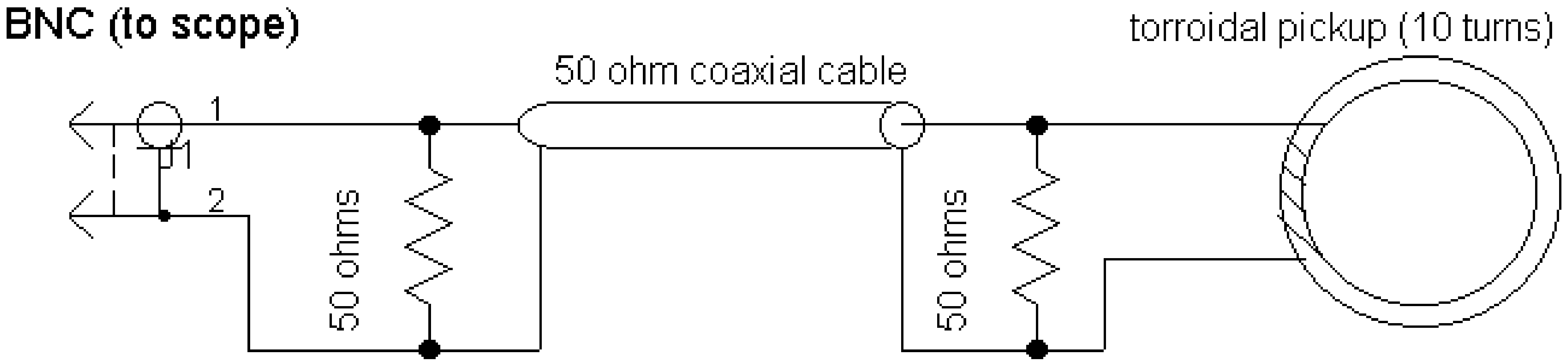
BNC (to scope)
torroidal pickup (10 turns)
50 ohm coaxial cable
1
J1
2
50 ohms
50 ohms

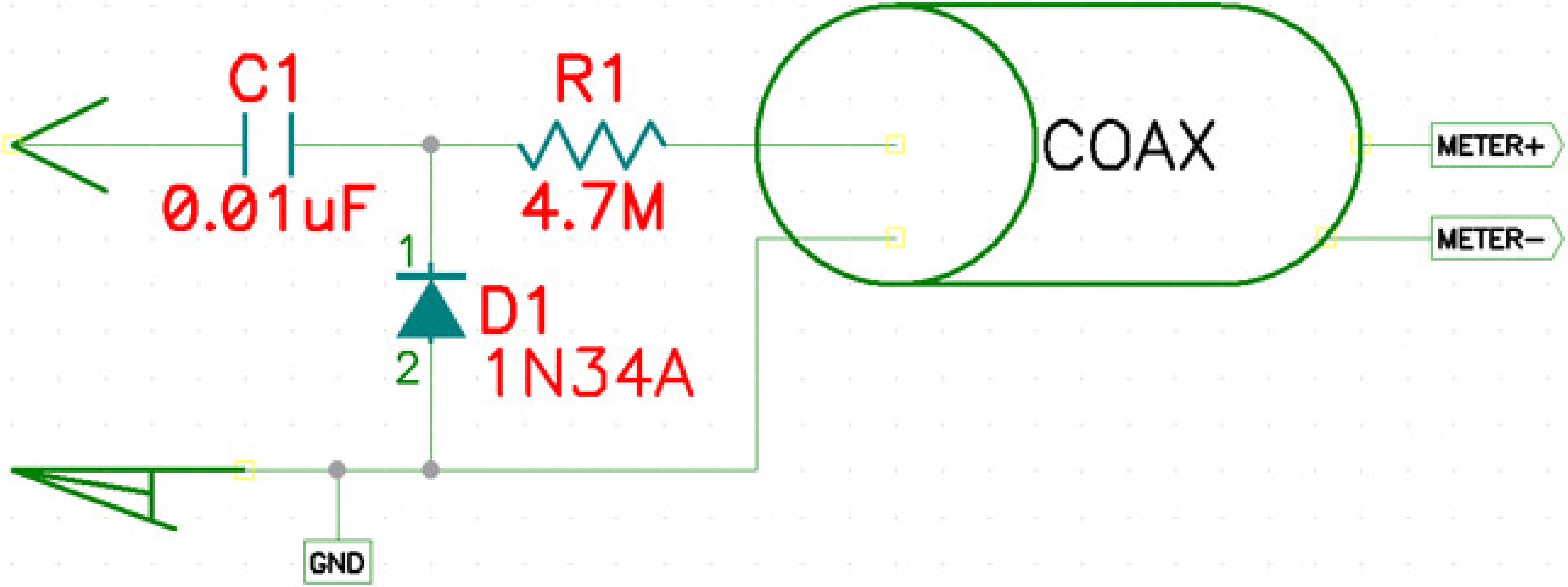
C1
0.01uF
R1
4.7M
COAX
METER+
METER−
1
D1
1N34A
2
GND

Simple RF Probe

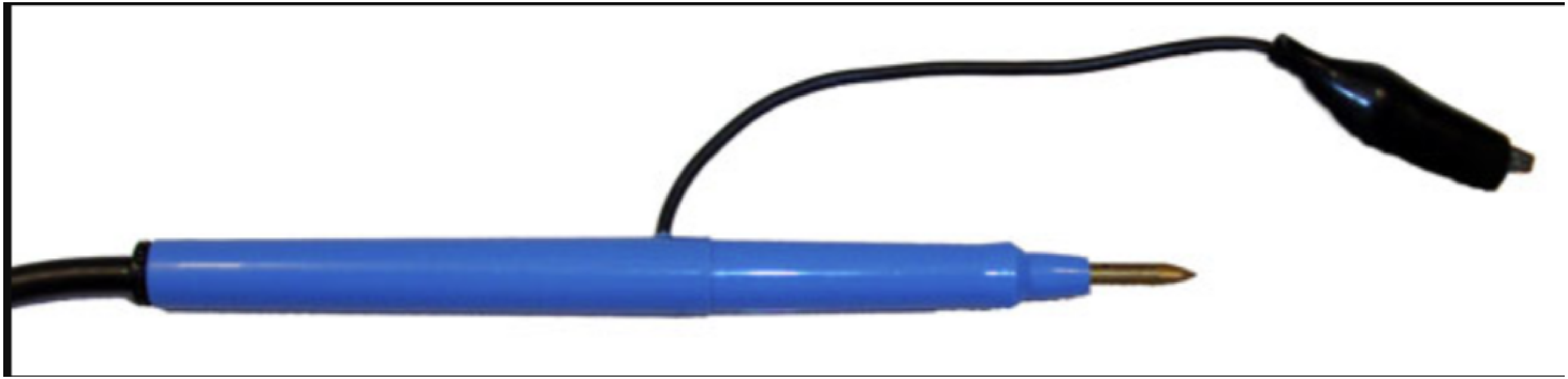

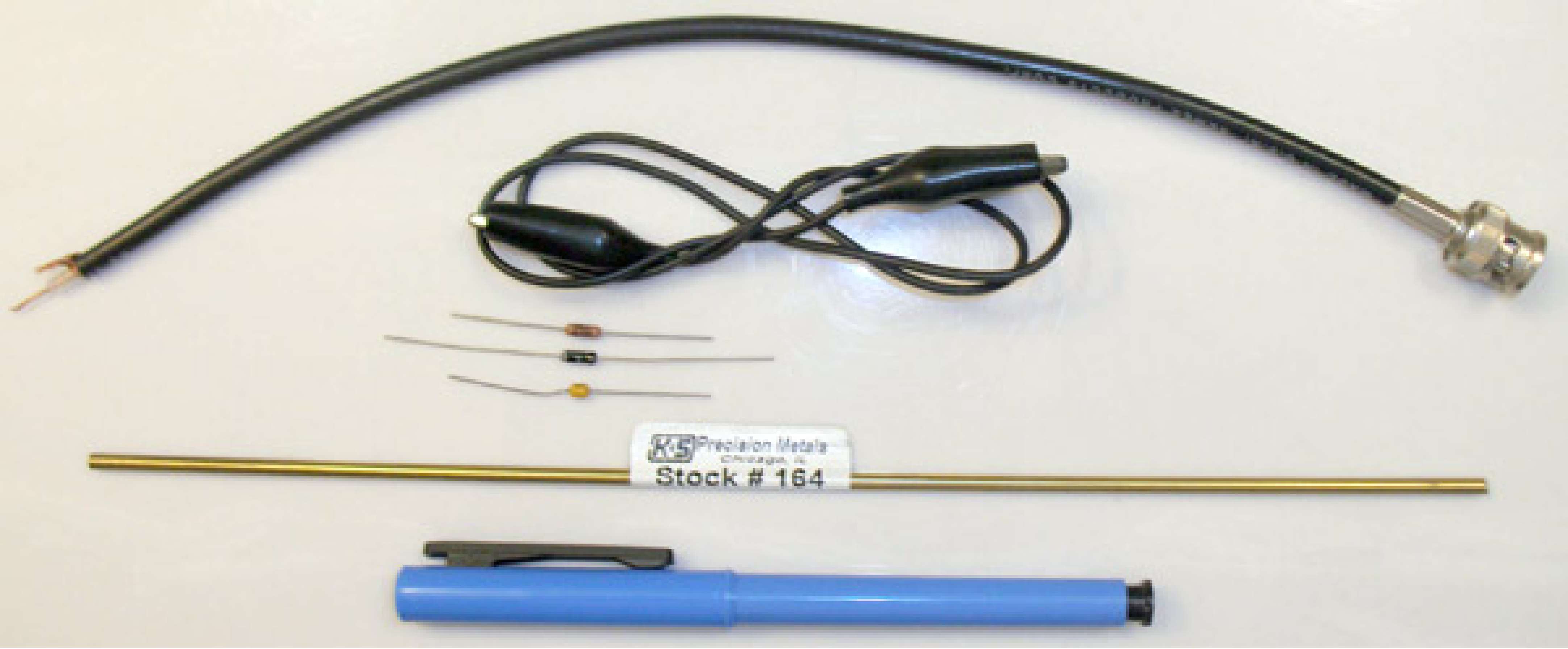
K&S Precision Metals
Chicago, IL
Stock # 164

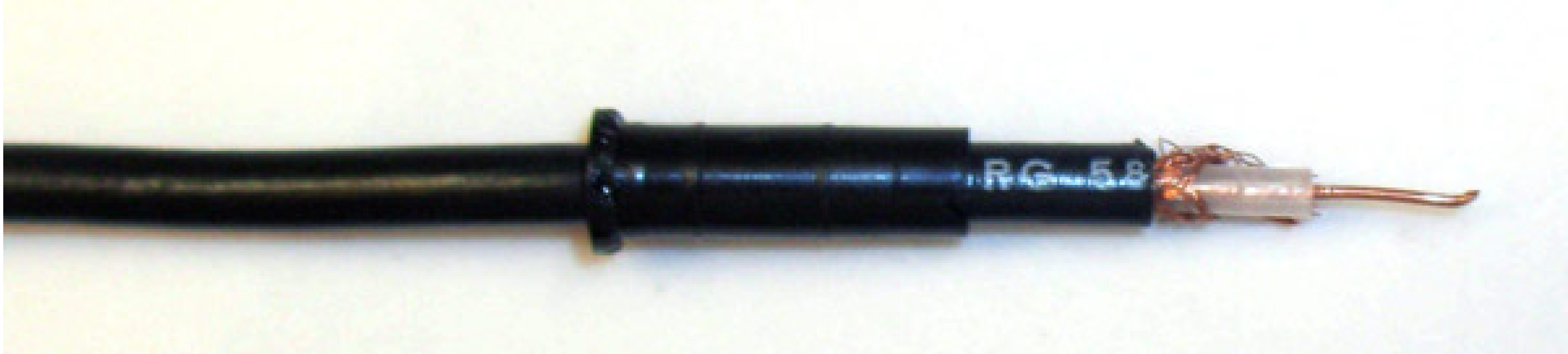
RG 58

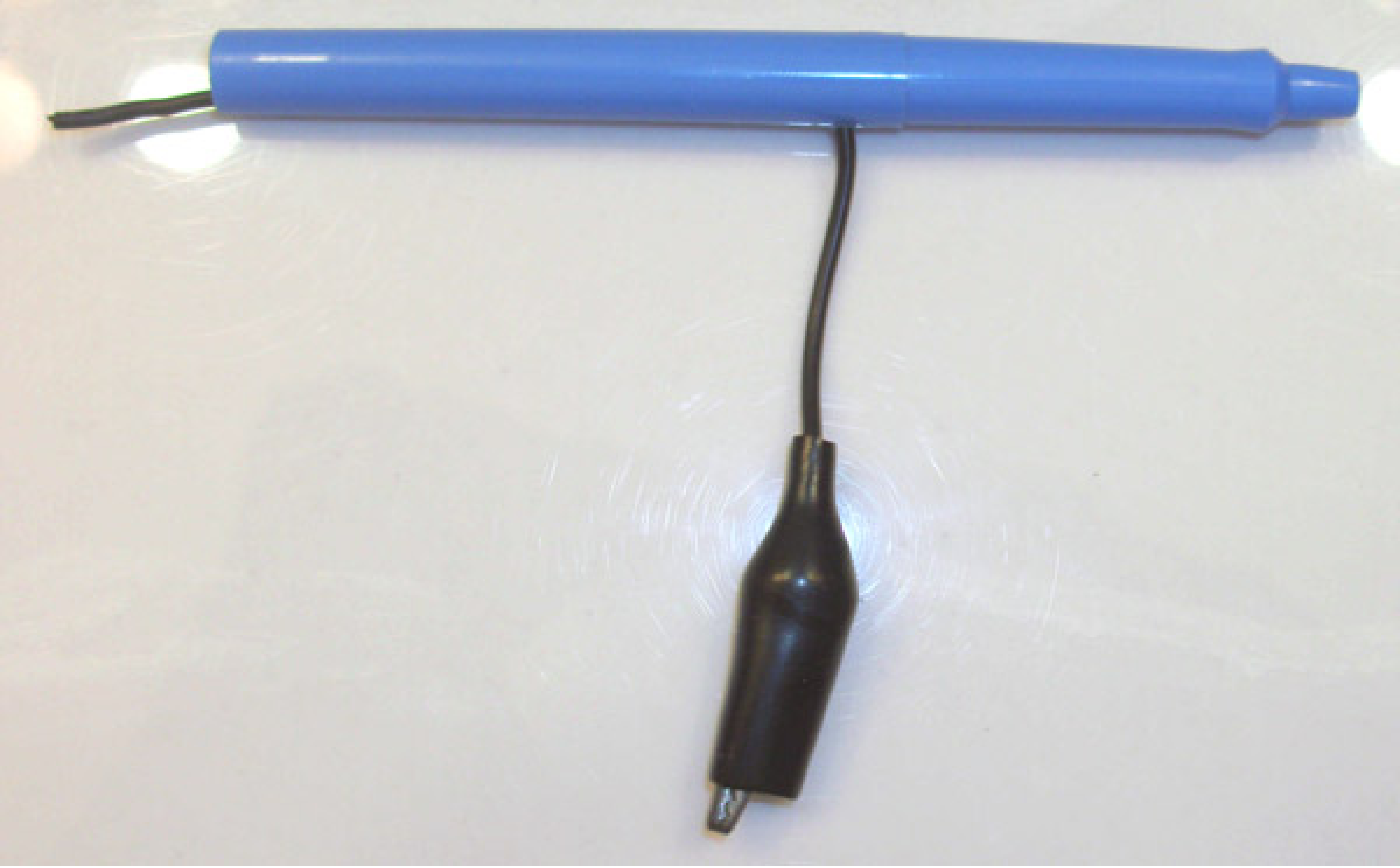

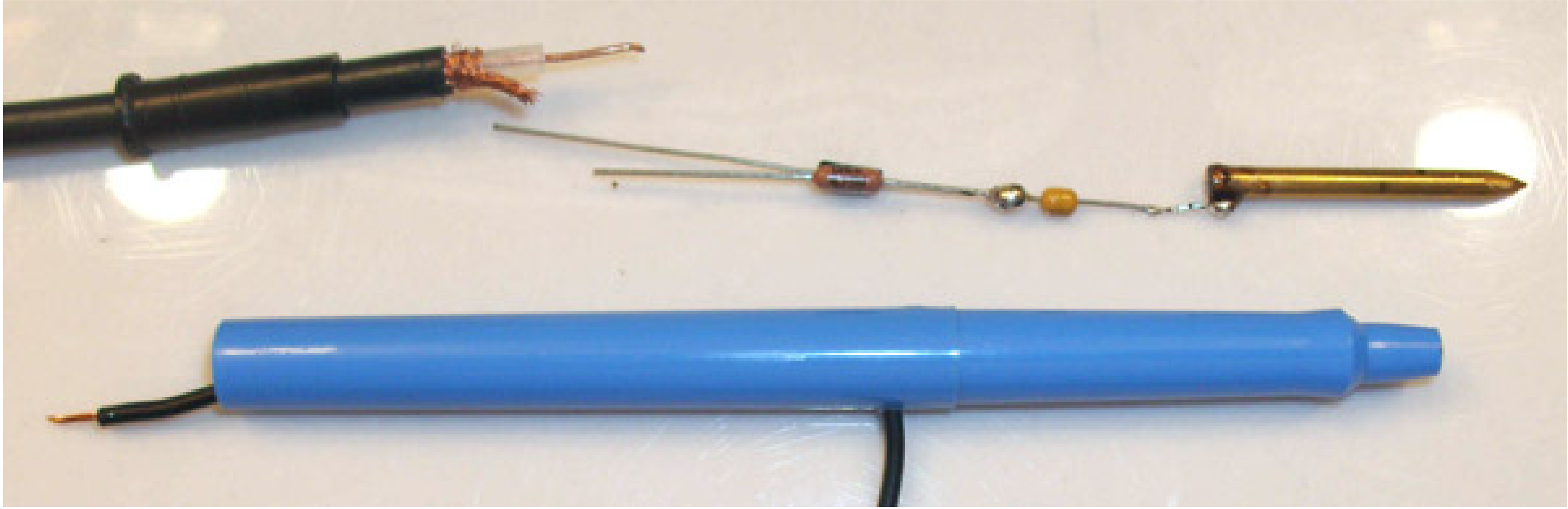

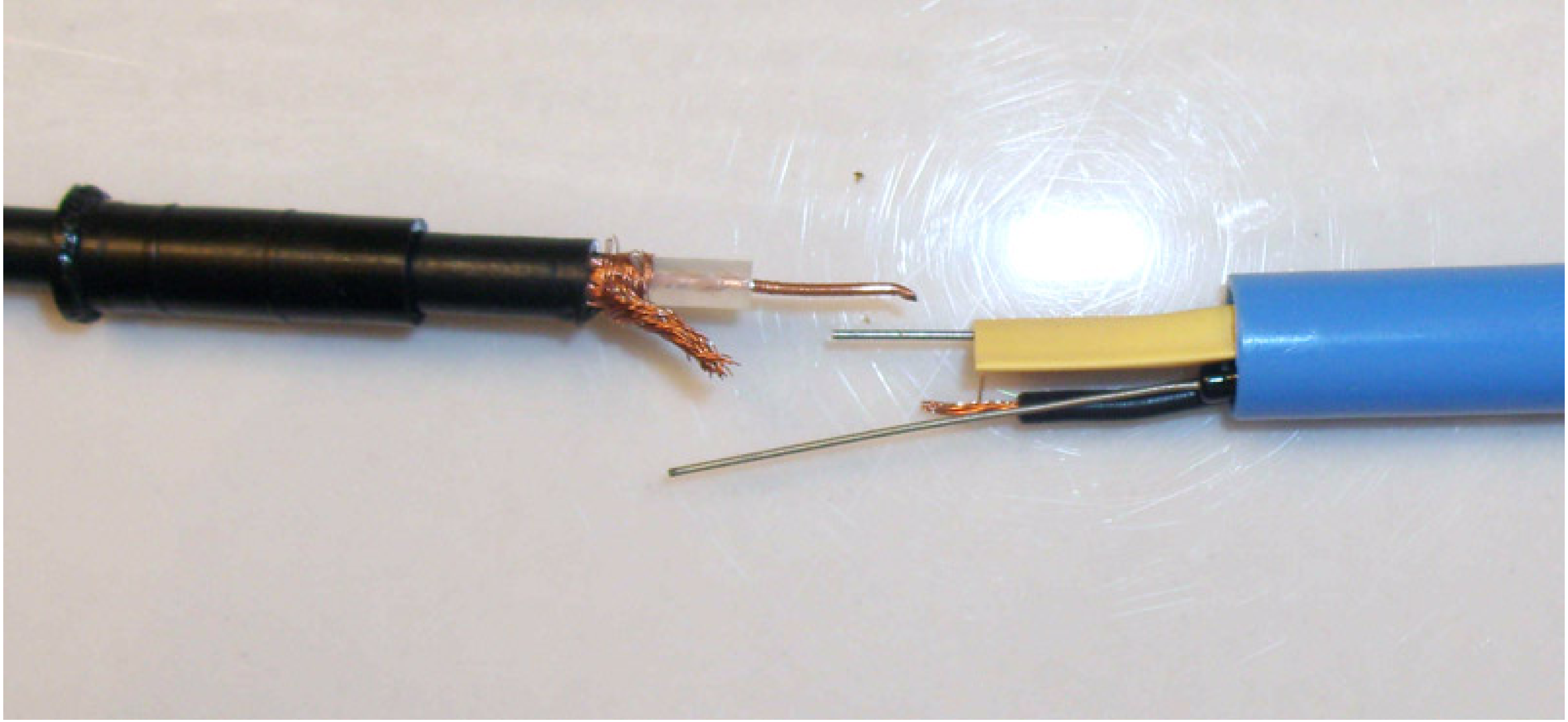

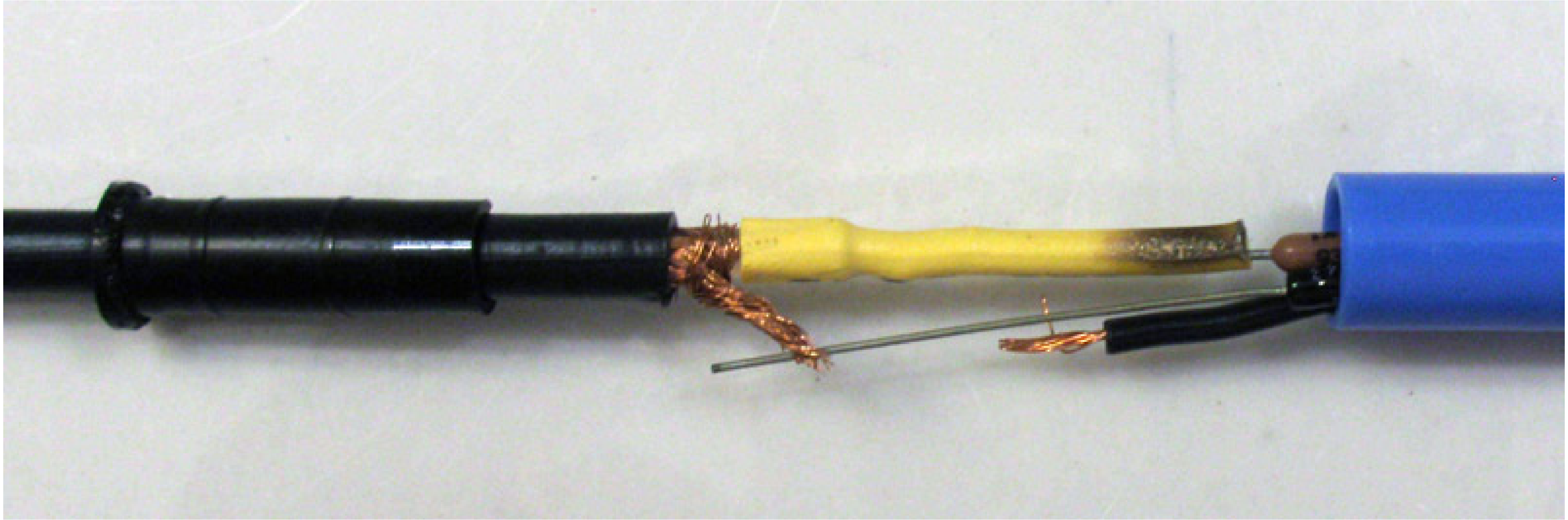

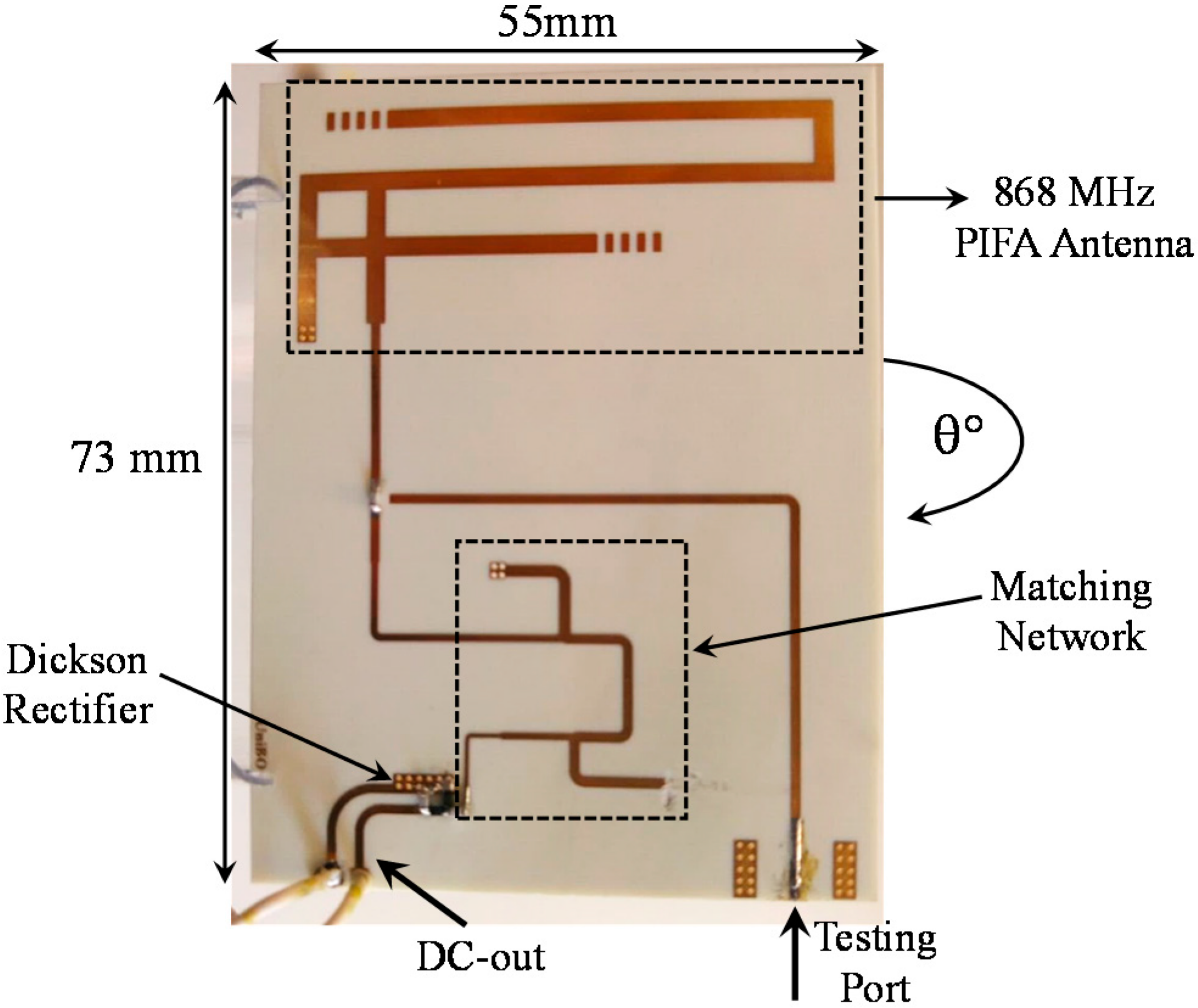
55mm
73 mm
868 MHz
PIFA Antenna
θ°
Matching
Network
Dickson
Rectifier
DC-out
Testing
Port

I used a .02ufd cap because it was the perfect physical size, a 4.7 meg resistor, and a 1N34A diode. I had a nice, flexible, p copper nail that would be easy to solder to for a tip.

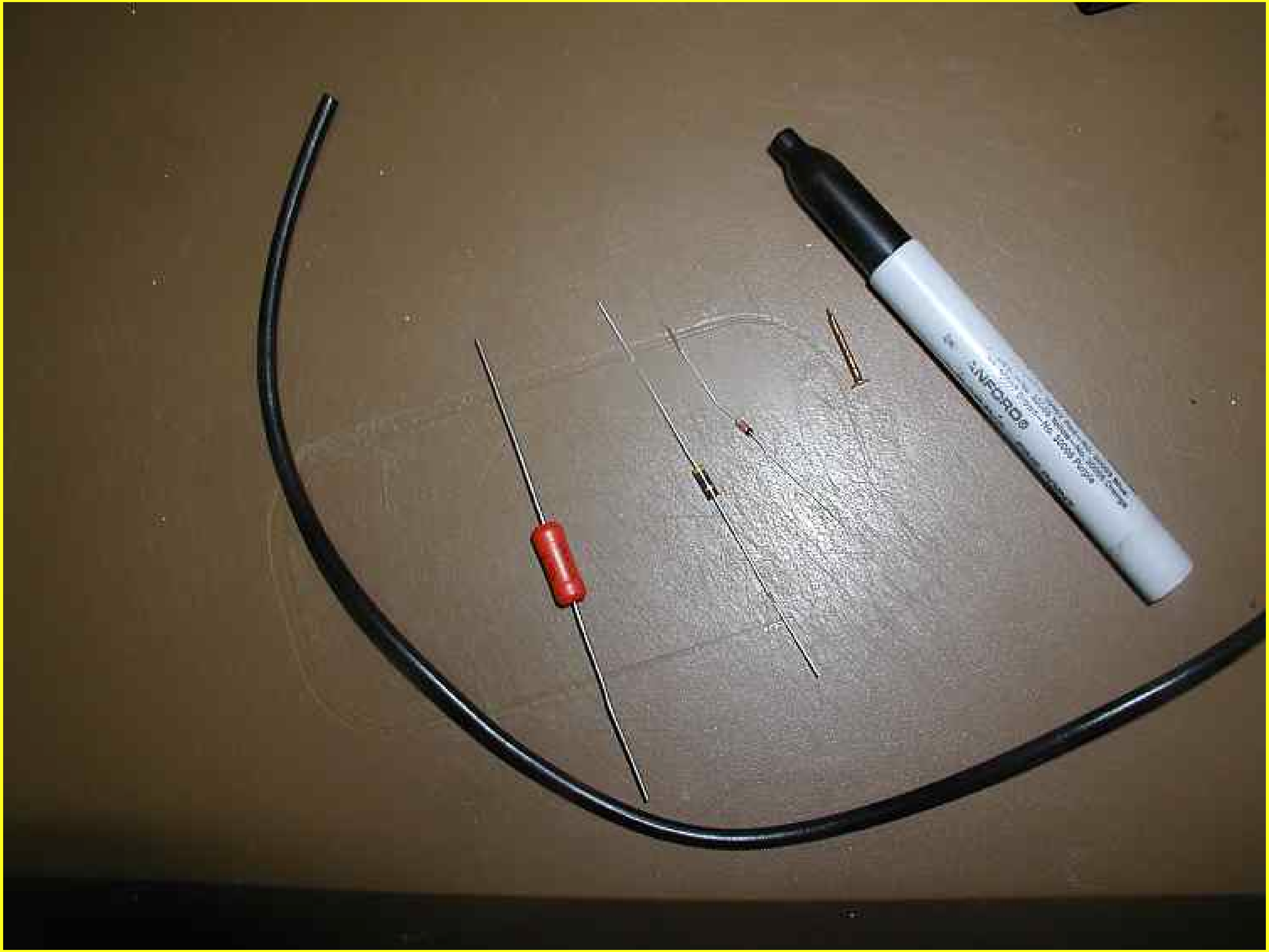

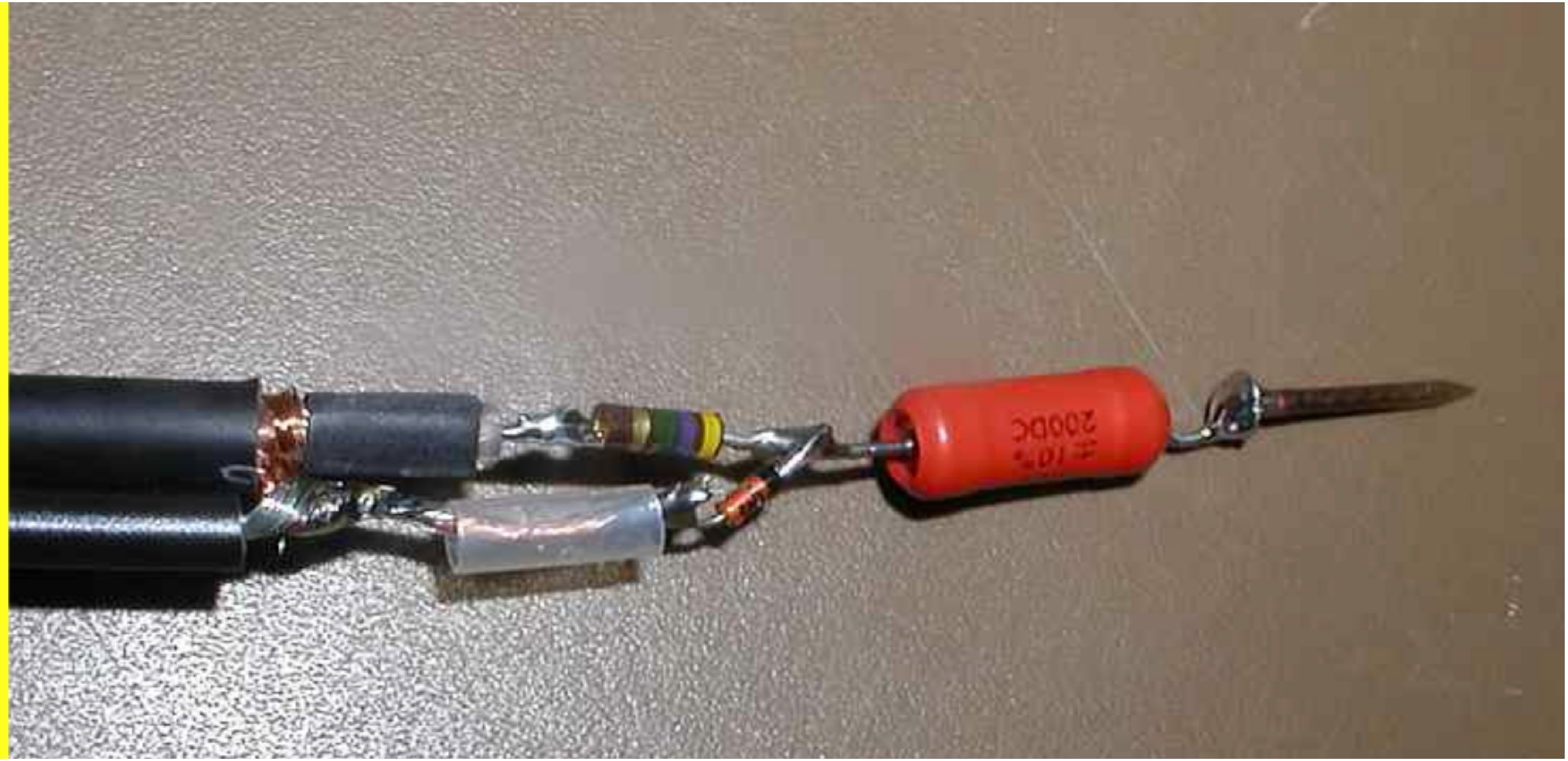
±10%
200DC

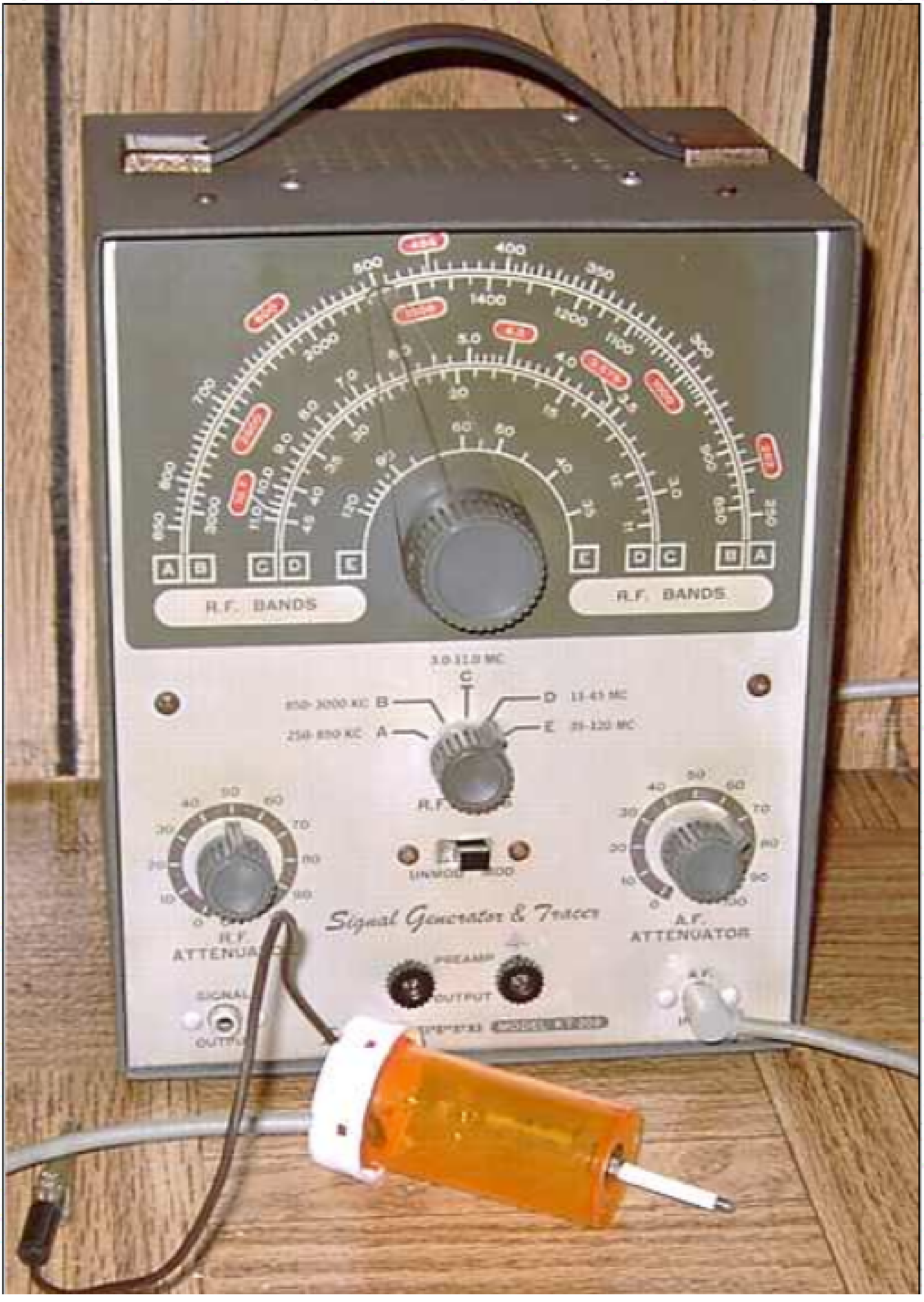
R.F. BANDS
R.F. BANDS
3.0-11.0 MC
850-3000 KC
250-850 KC
11-45 MC
35-120 MC
UNMOD
MOD
Signal Generator & Tracer
R.F. ATTENUATOR
A.F. ATTENUATOR
PREAMP
OUTPUT
SIGNAL OUTPUT

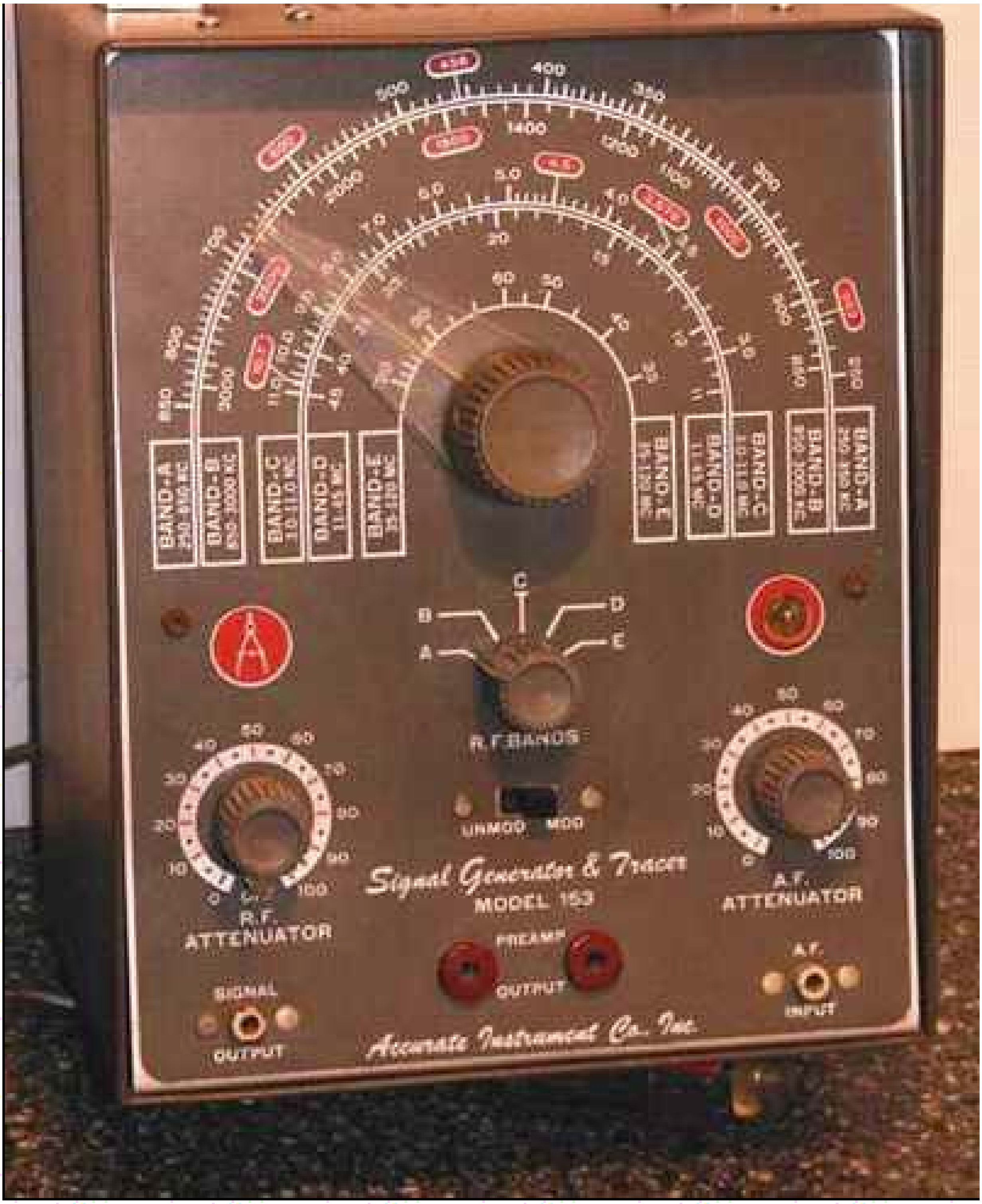
BAND-A
BAND-B
BAND-C
BAND-D
BAND-E
A
B
C
D
E
R.F. BANDS
UNMOD
MOD
R.F. ATTENUATOR
Signal Generator & Tracer
MODEL 153
A.F. ATTENUATOR
PREAMP
OUTPUT
SIGNAL
OUTPUT
A.F.
INPUT
Accurate Instrument Co. Inc.

Accurate Instrument model 153 and its RF probe schematic
courtesy of John Lescaud.

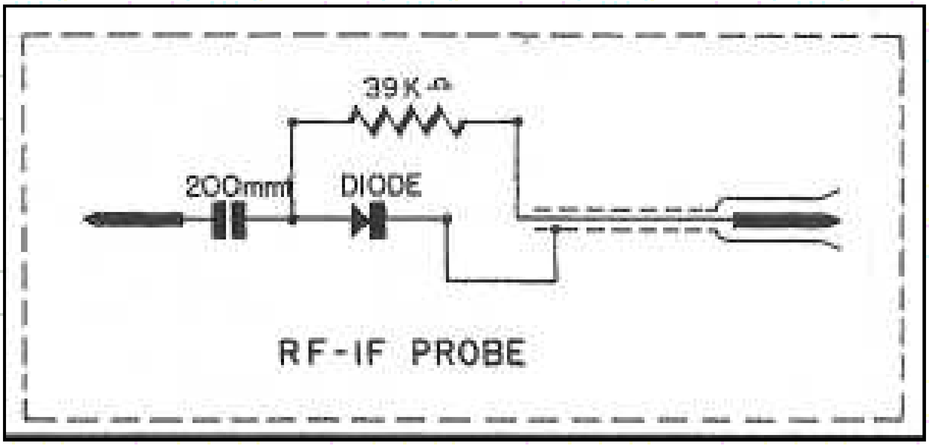
39K Ω
200mmf
DIODE
RF-IF PROBE

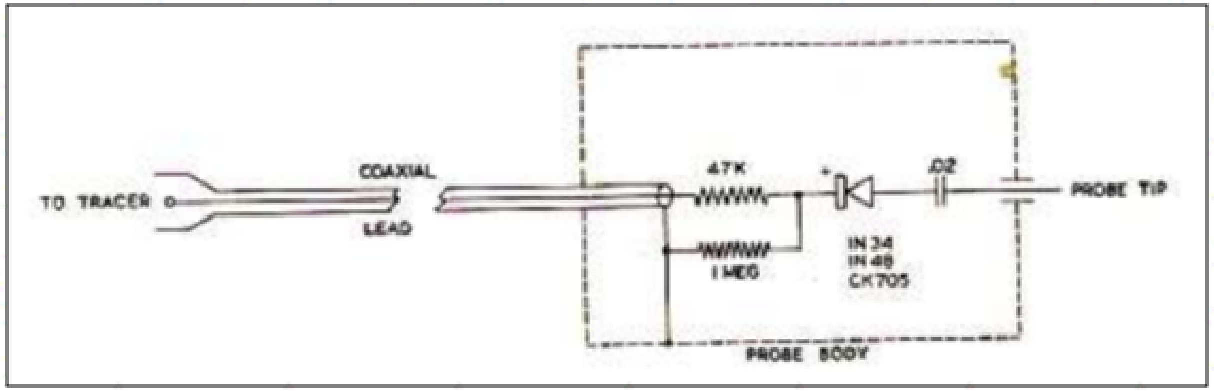
TO TRACER
COAXIAL
LEAD
47K
+
.02
PROBE TIP
1MEG
IN34
IN48
CK705
PROBE BODY

Heath T-3 signal tracer probe

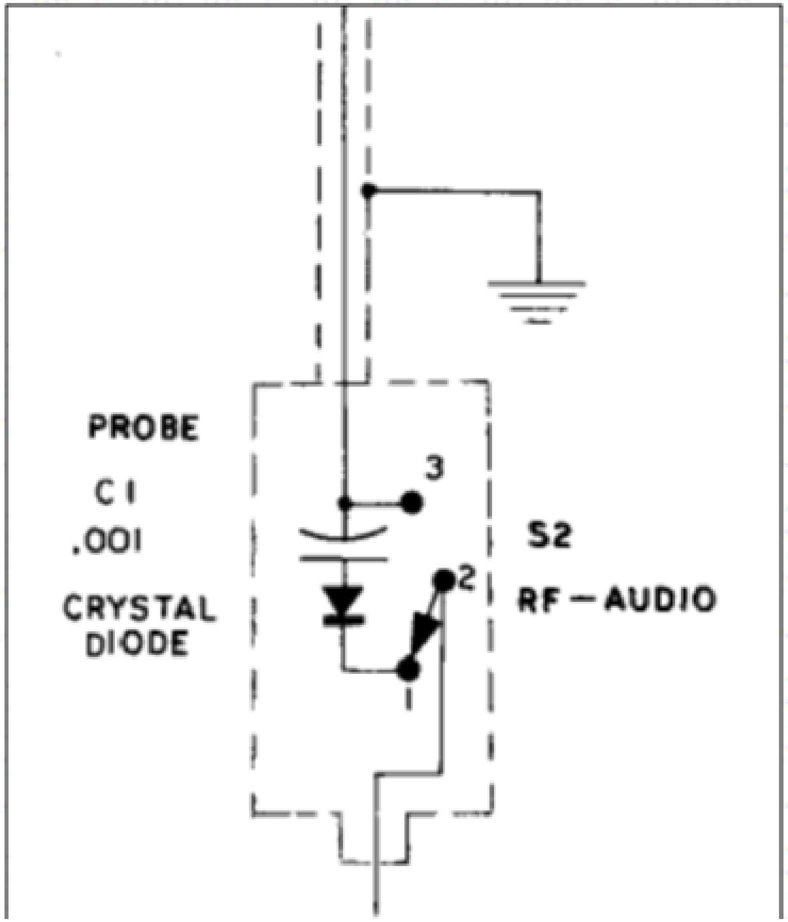
PROBE
C1
.001
CRYSTAL
DIODE
3
2
S2
RF—AUDIO

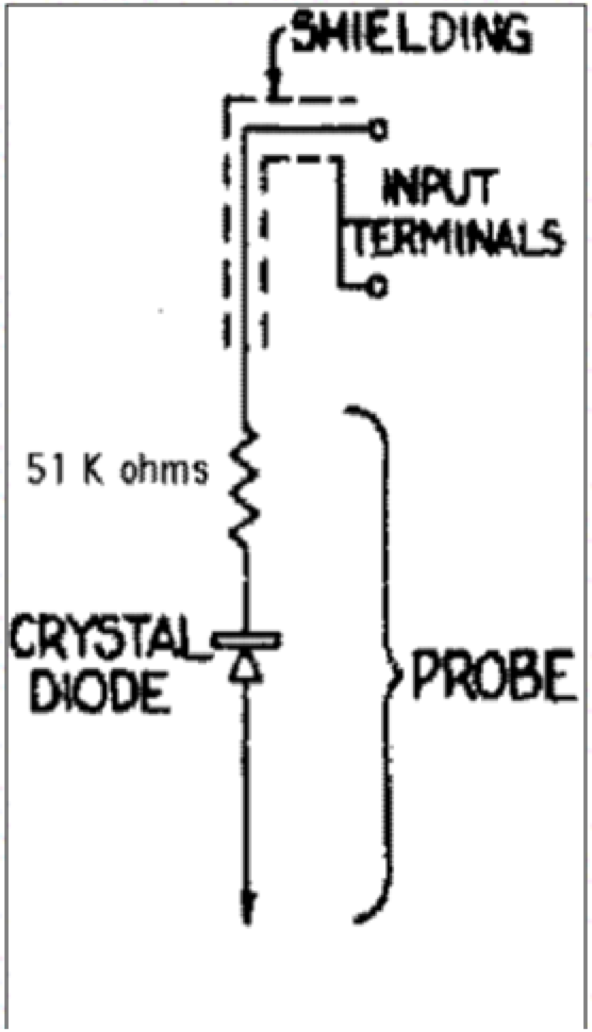
SHIELDING
INPUT
TERMINALS
51 K ohms
CRYSTAL
DIODE
PROBE

Heath T-4 or IT-12 signal tracer probe (left) - - - - Eico 145 signal tracer probe (right)

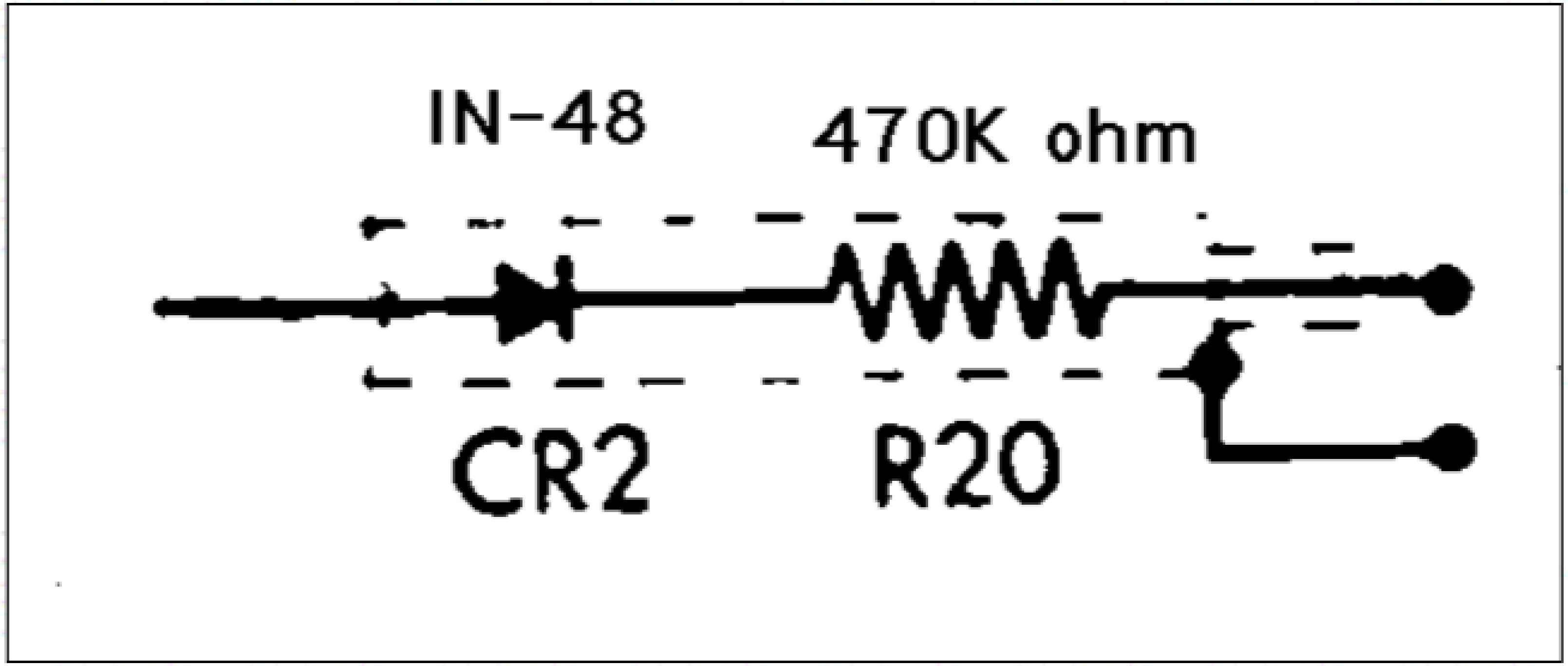
IN-48
470K ohm
CR2
R20

Eico 147a signal tracer probe

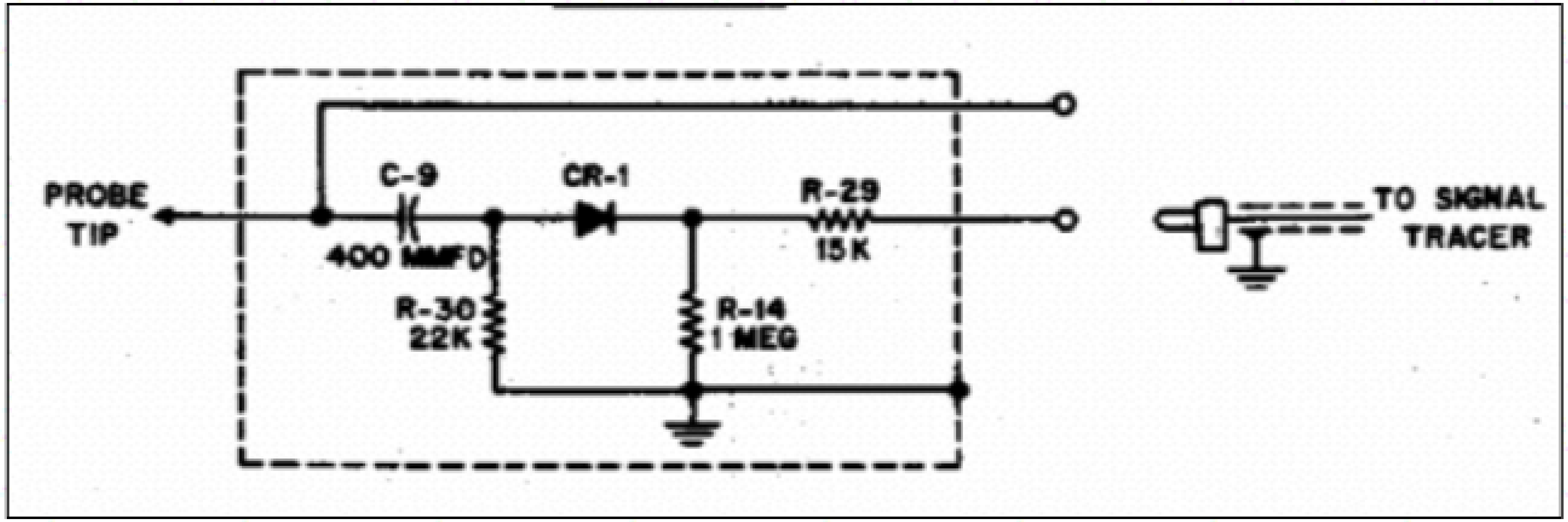
PROBE TIP
C-9
400 MMFD
CR-1
R-29
15K
R-30
22K
R-14
1 MEG
TO SIGNAL TRACER

Knight-kit signal tracer probe

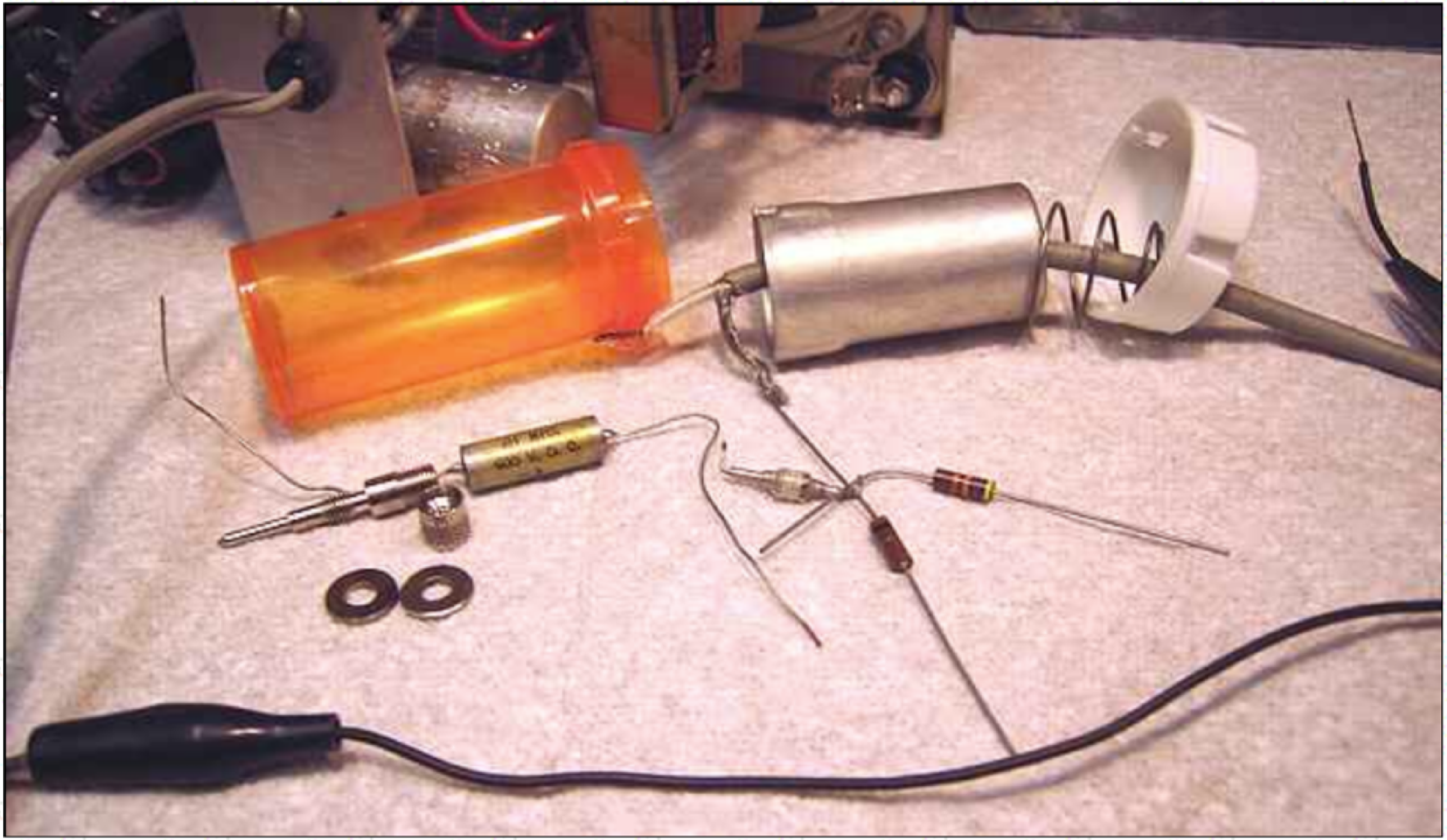

Pill bottle signal tracer probe

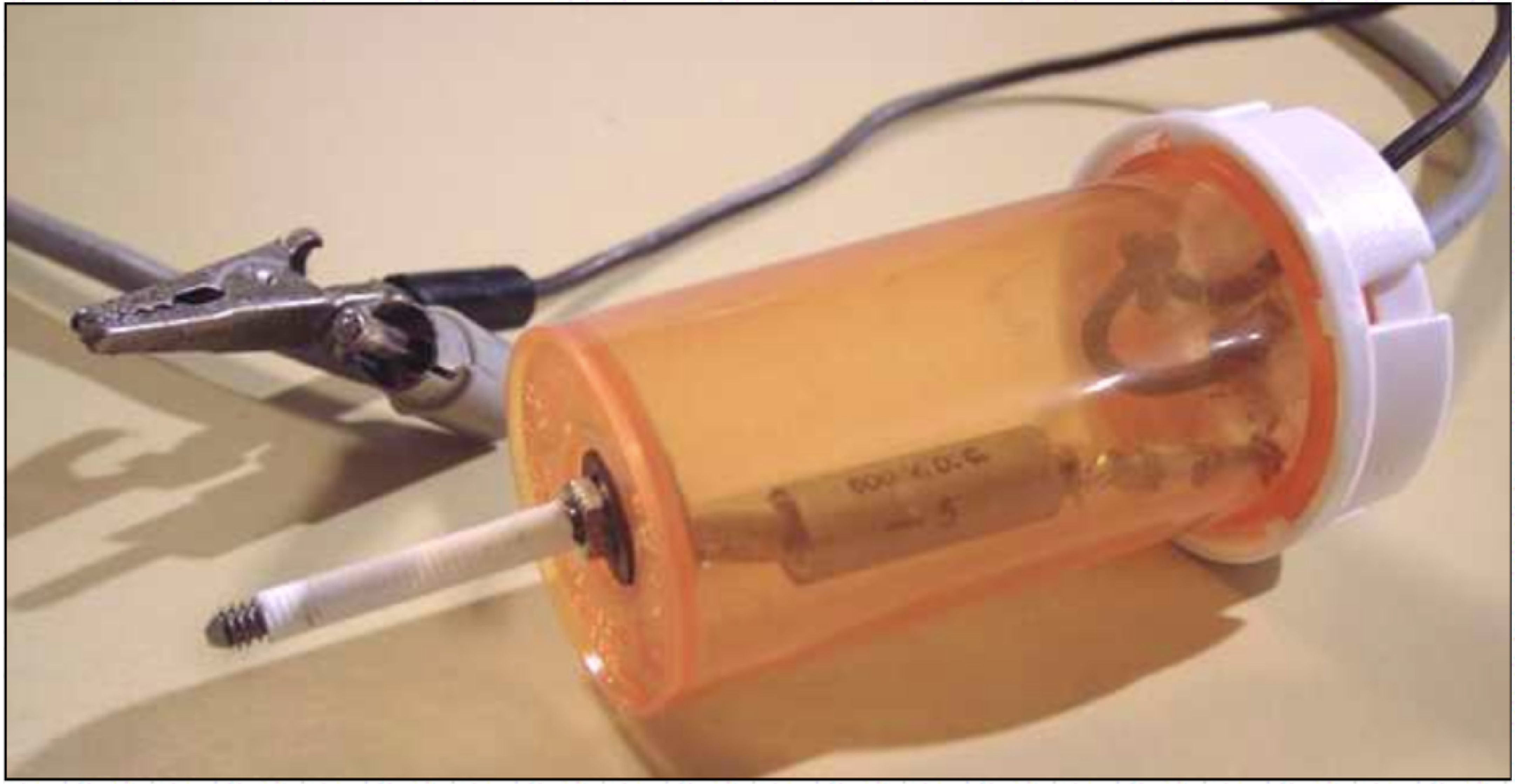

KIS (Keep It Simple) pill-bottle signal tracer probe

Making the simplest Transmitter

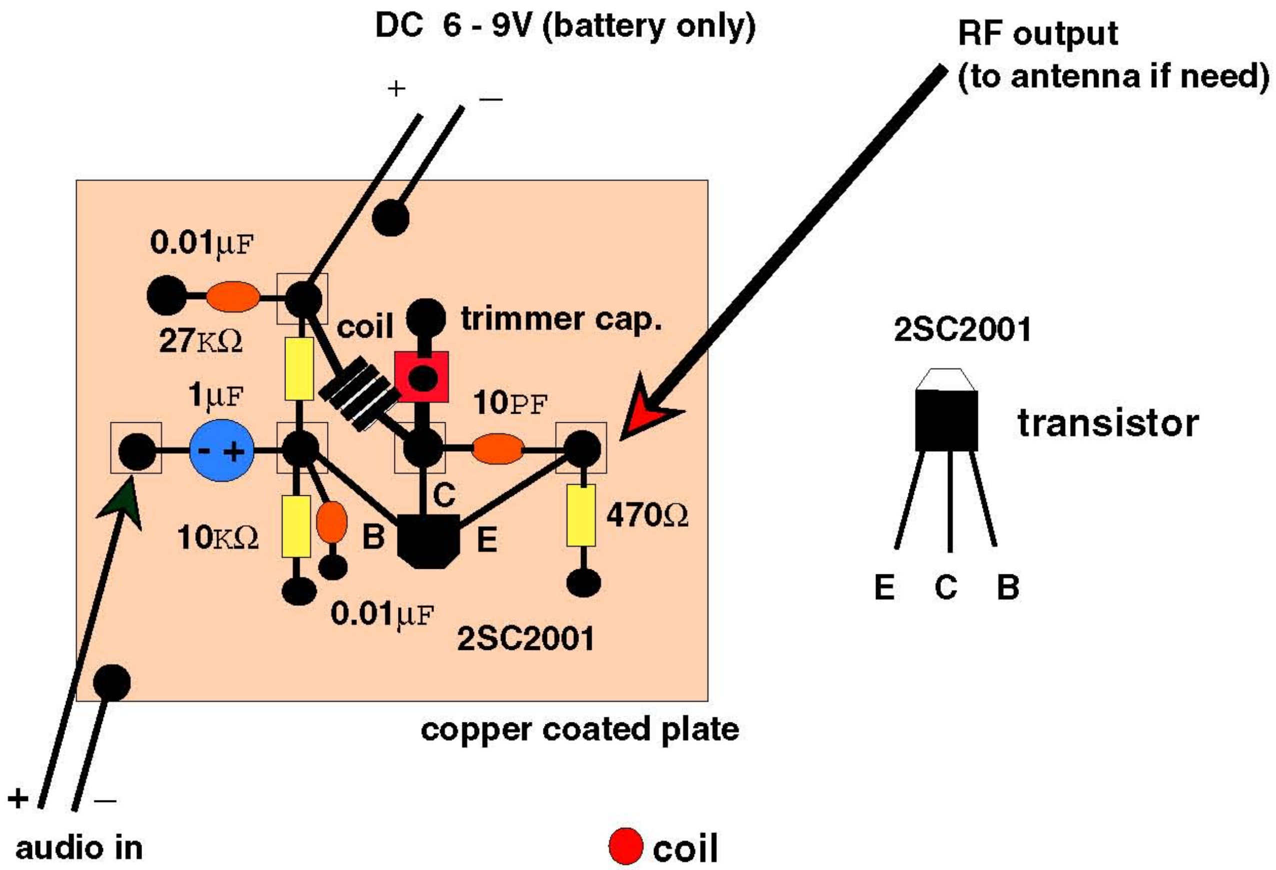
DC 6 - 9V (battery only)
+
—
RF output
(to antenna if need)
0.01μF
27KΩ
coil
trimmer cap.
1μF
- +
10PF
C
10KΩ
B
E
470Ω
0.01μF
2SC2001
copper coated plate
+
—
audio in
2SC2001
transistor
E
C
B

● **soldered point:**

● : direct to the ground

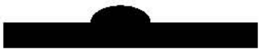

▣ : insulated from
the ground

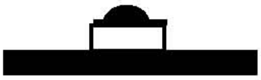

● **coil**

3 - 4 turns by coated 0.8 mm wire

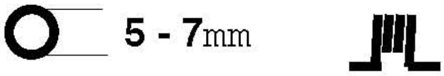

● **registers**

470 Ω *(yellow-violet-brown)*
10K Ω *(brown-black-orange)*
27K Ω *(red-violet-orange)*

● **capacitors**

10 PF
0.01 μF *(103)*
1 μF

● **trimmer capasitor 20PF**

2005-03-16 by Tetsuo Kogawa

Design – Physical System

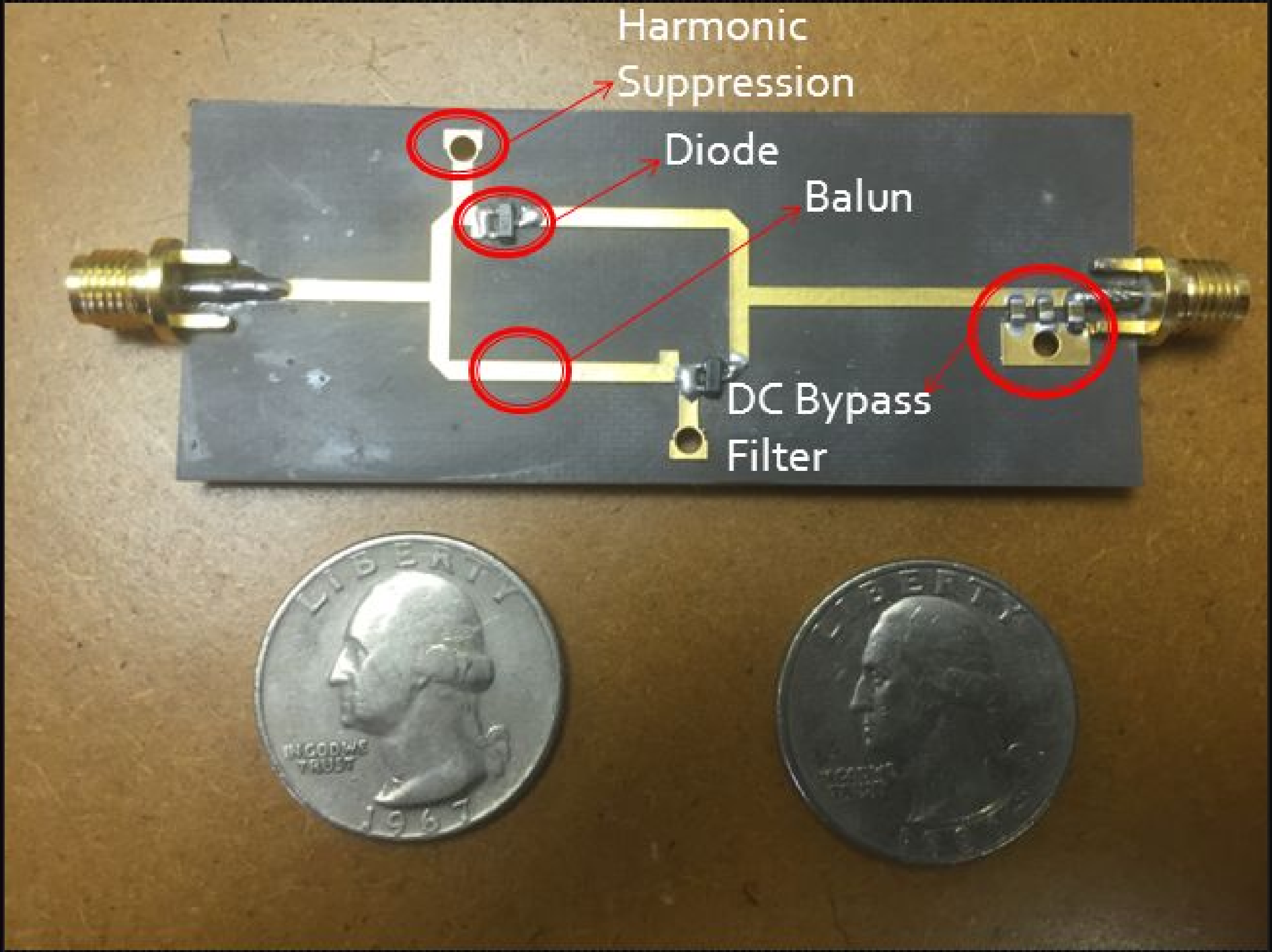
Harmonic Suppression
Diode
Balun
DC Bypass Filter
LIBERTY
IN GOD WE TRUST
1967

Fig 5. RF to DC converter

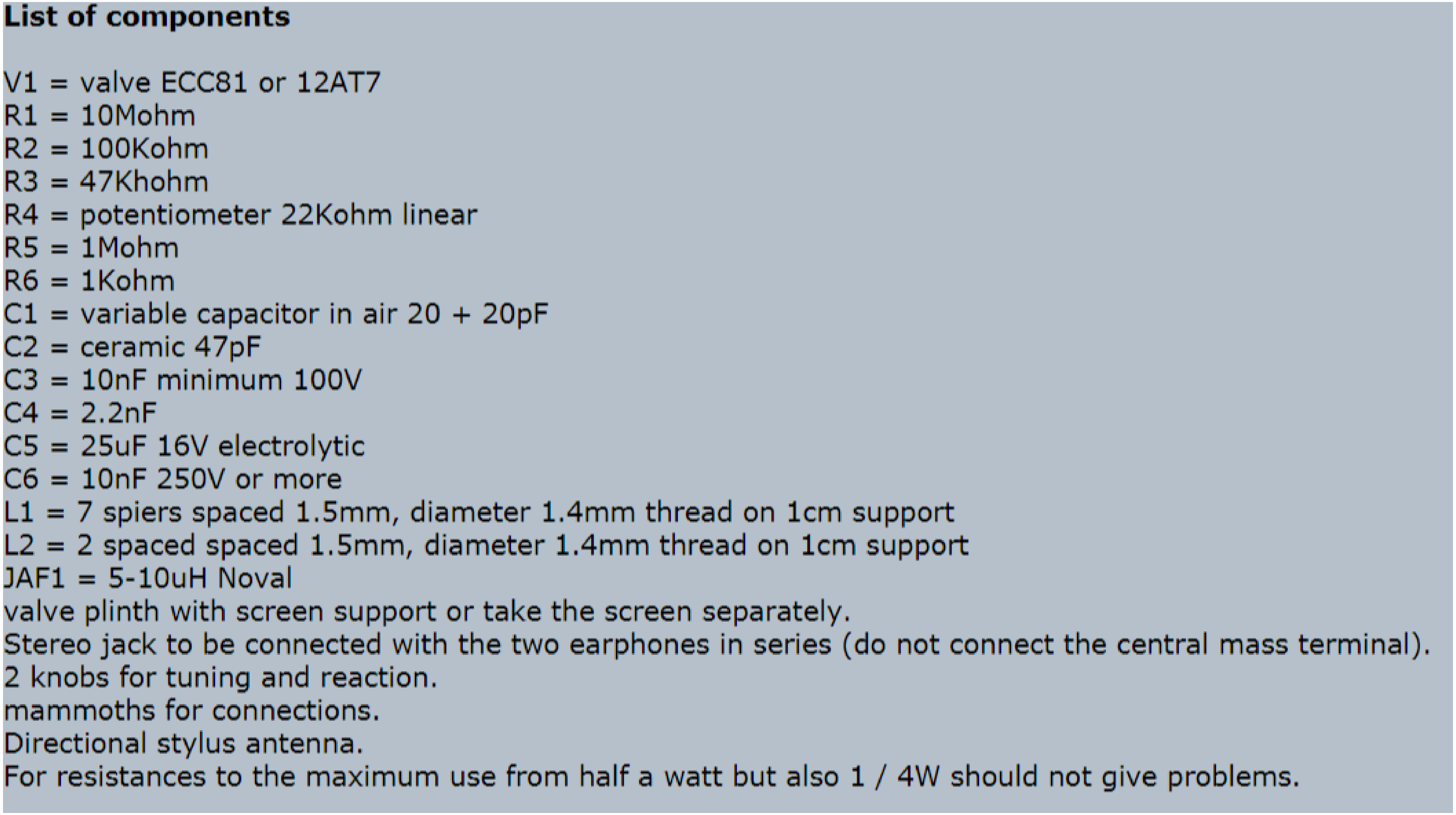

List of components

V1 = valve ECC81 or 12AT7
R1 = 10Mohm
R2 = 100Kohm
R3 = 47Khohm
R4 = potentiometer 22Kohm linear
R5 = 1Mohm
R6 = 1Kohm
C1 = variable capacitor in air 20 + 20pF
C2 = ceramic 47pF
C3 = 10nF minimum 100V
C4 = 2.2nF
C5 = 25uF 16V electrolytic
C6 = 10nF 250V or more
L1 = 7 spiers spaced 1.5mm, diameter 1.4mm thread on 1cm support
L2 = 2 spaced spaced 1.5mm, diameter 1.4mm thread on 1cm support
JAF1 = 5-10uH Noval
valve plinth with screen support or take the screen separately.
Stereo jack to be connected with the two earphones in series (do not connect the central mass terminal).
2 knobs for tuning and reaction.
mammoths for connections.
Directional stylus antenna.
For resistances to the maximum use from half a watt but also 1 / 4W should not give problems.

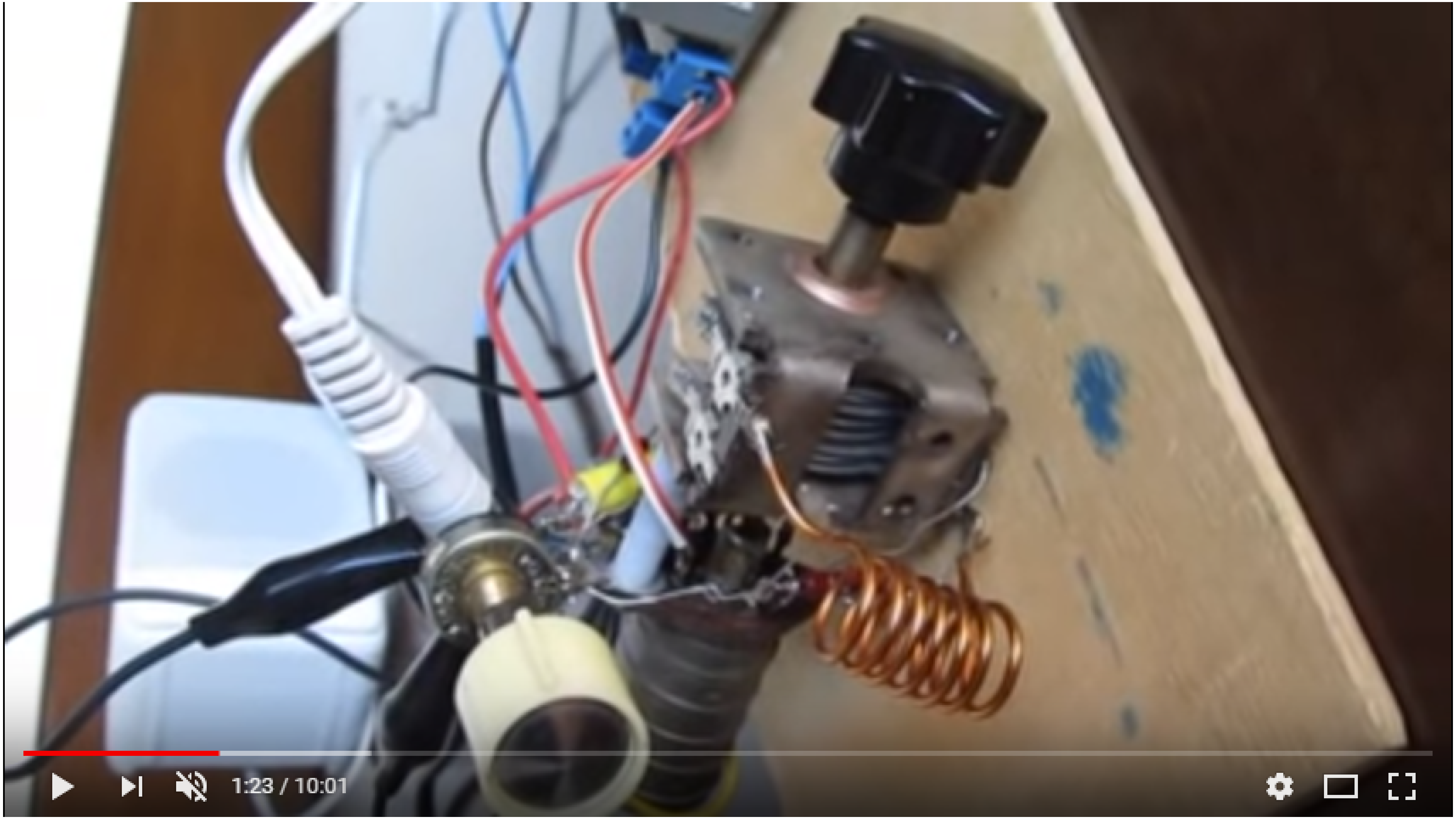

Ricevitore in superreazione con valvola ECC81 12AT7 Primo montaggio.mp4

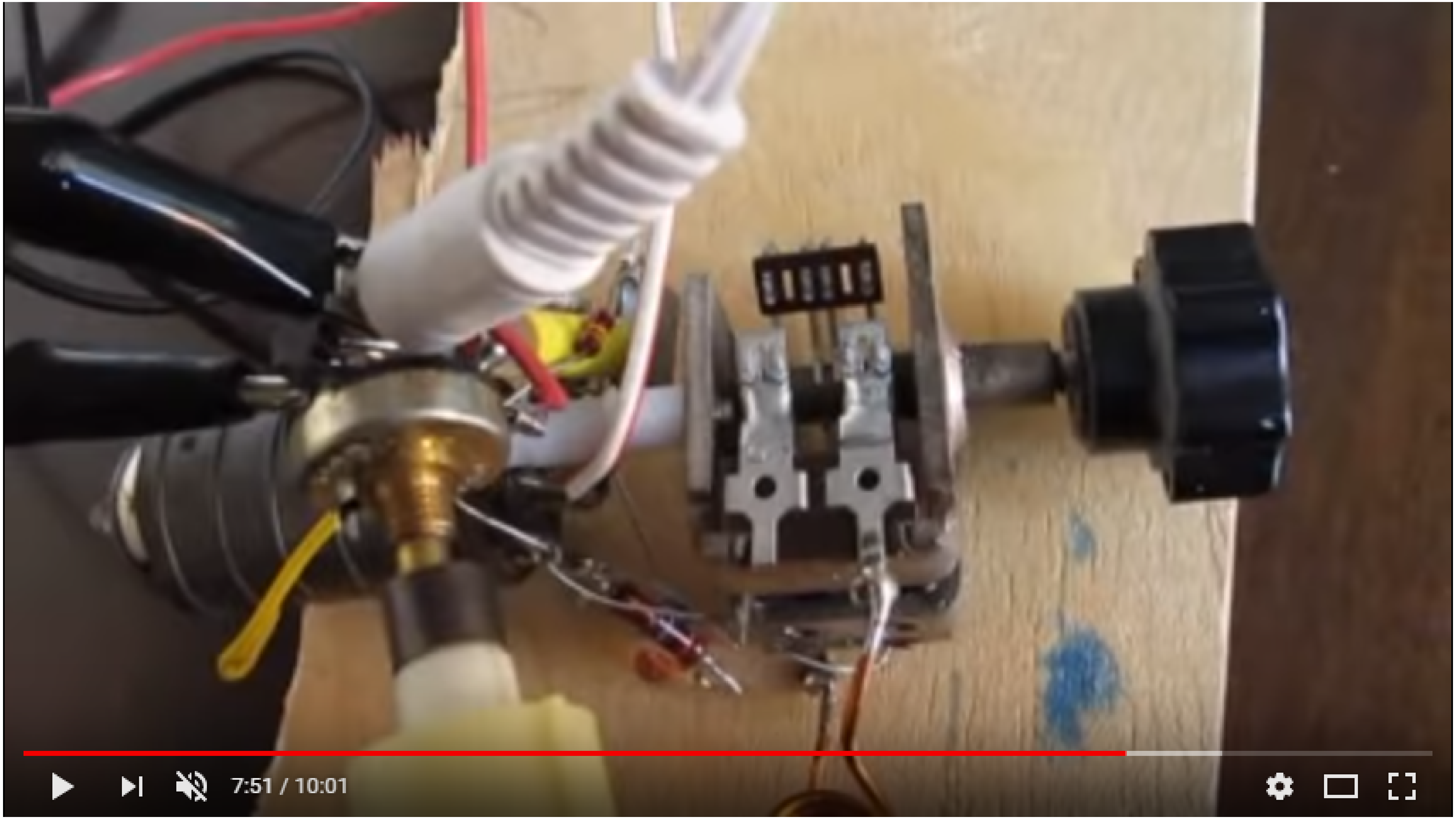
7:51 / 10:01

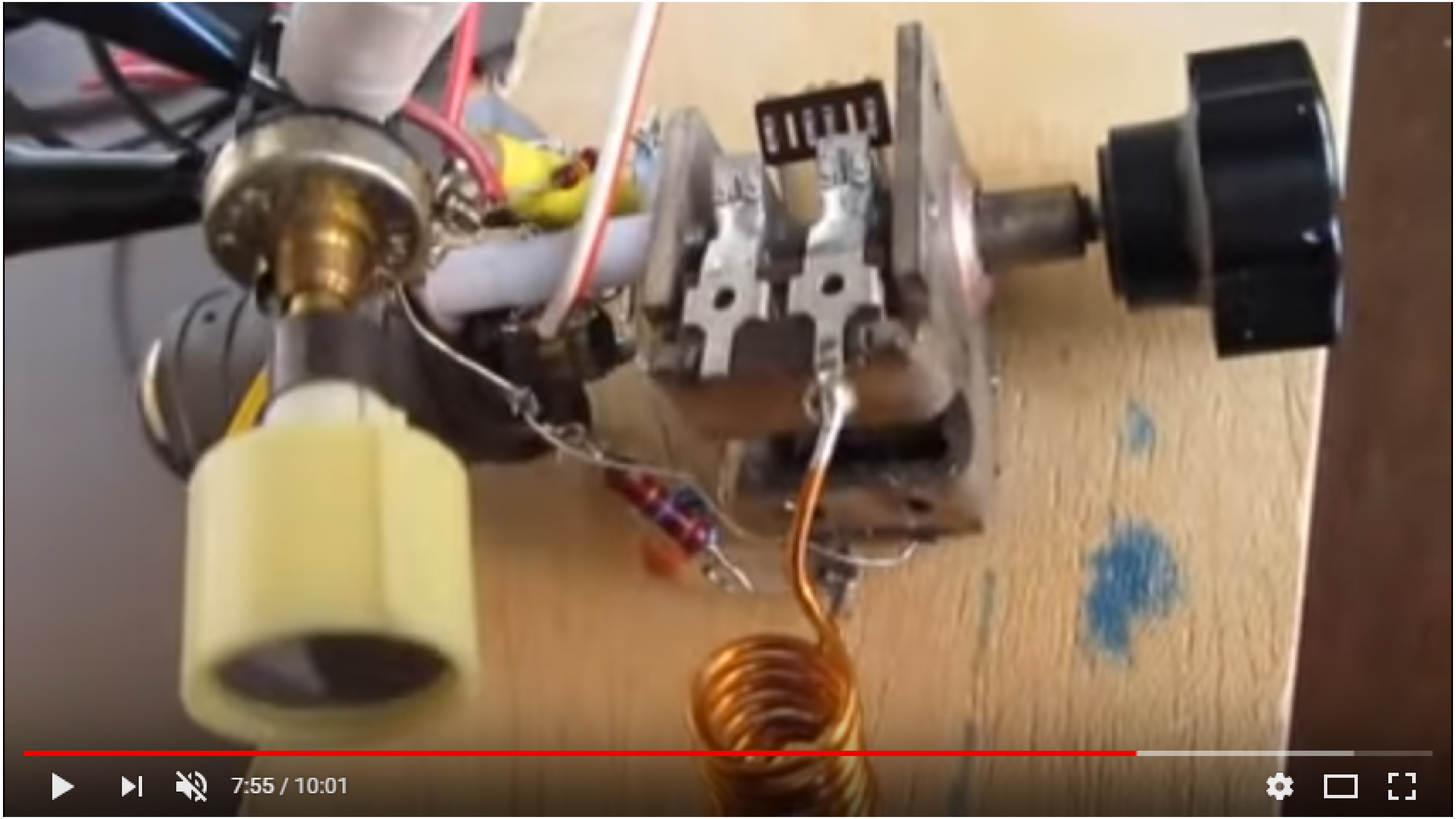
7:55 / 10:01

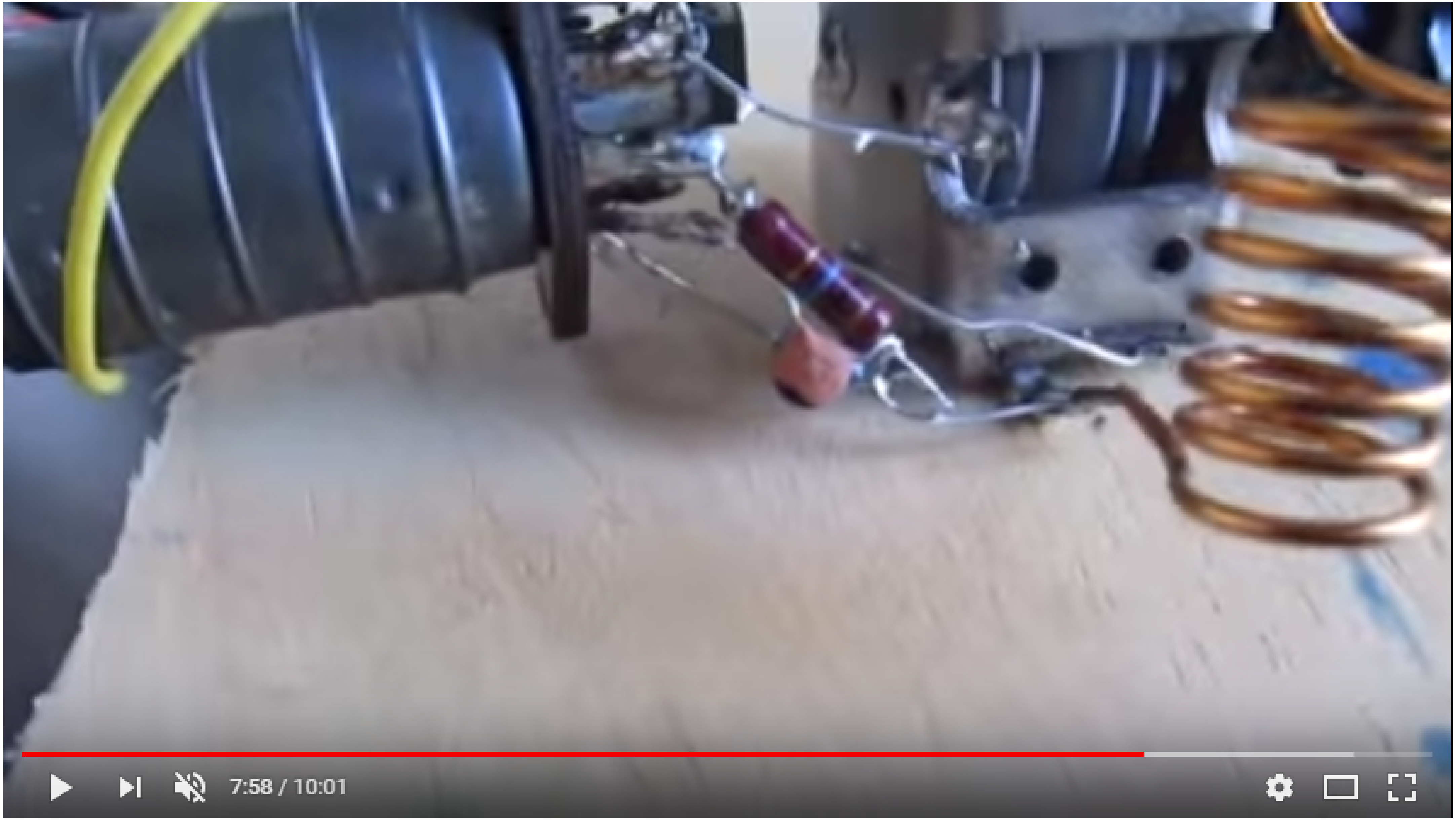
7:58 / 10:01

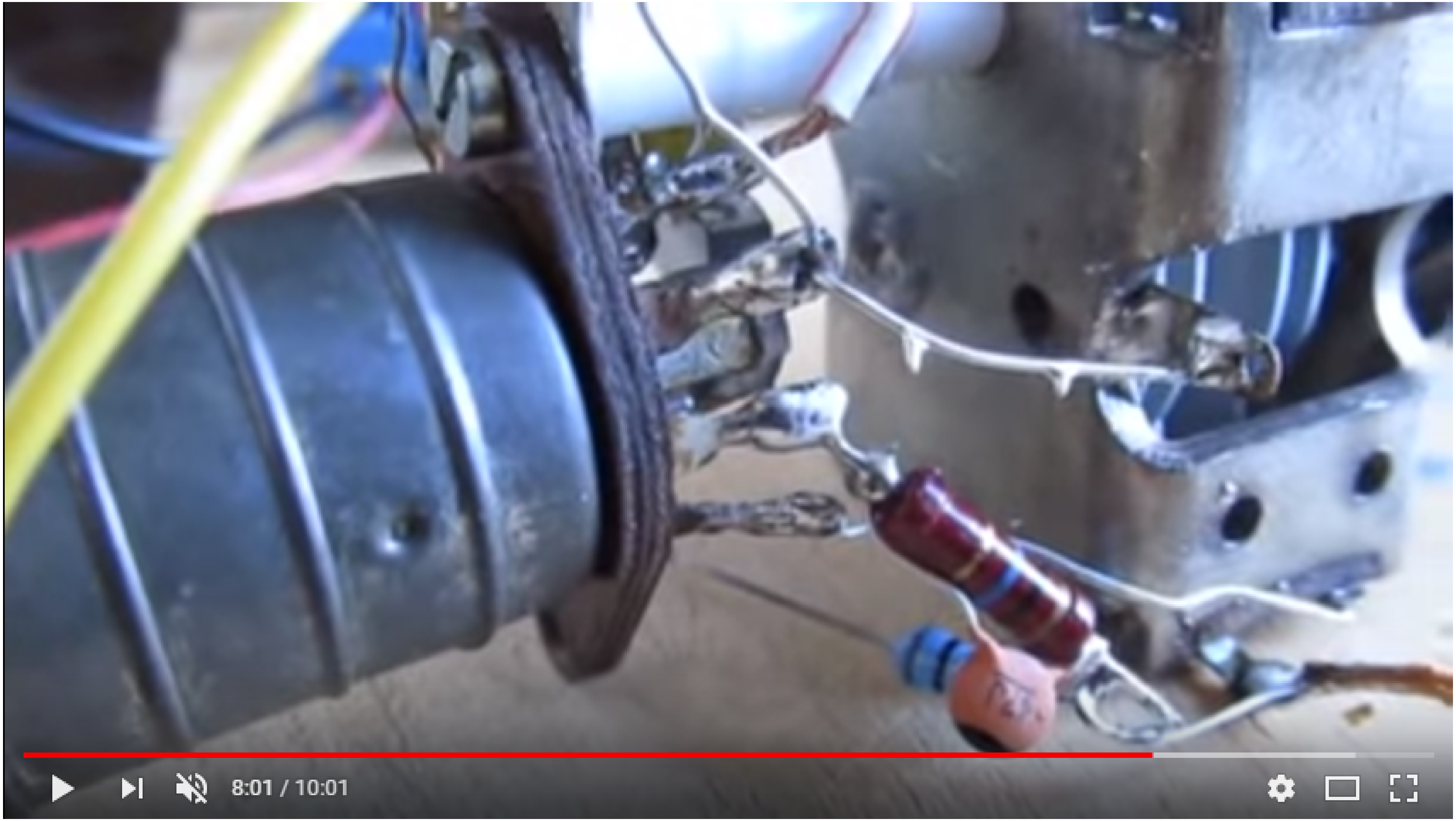
8:01 / 10:01

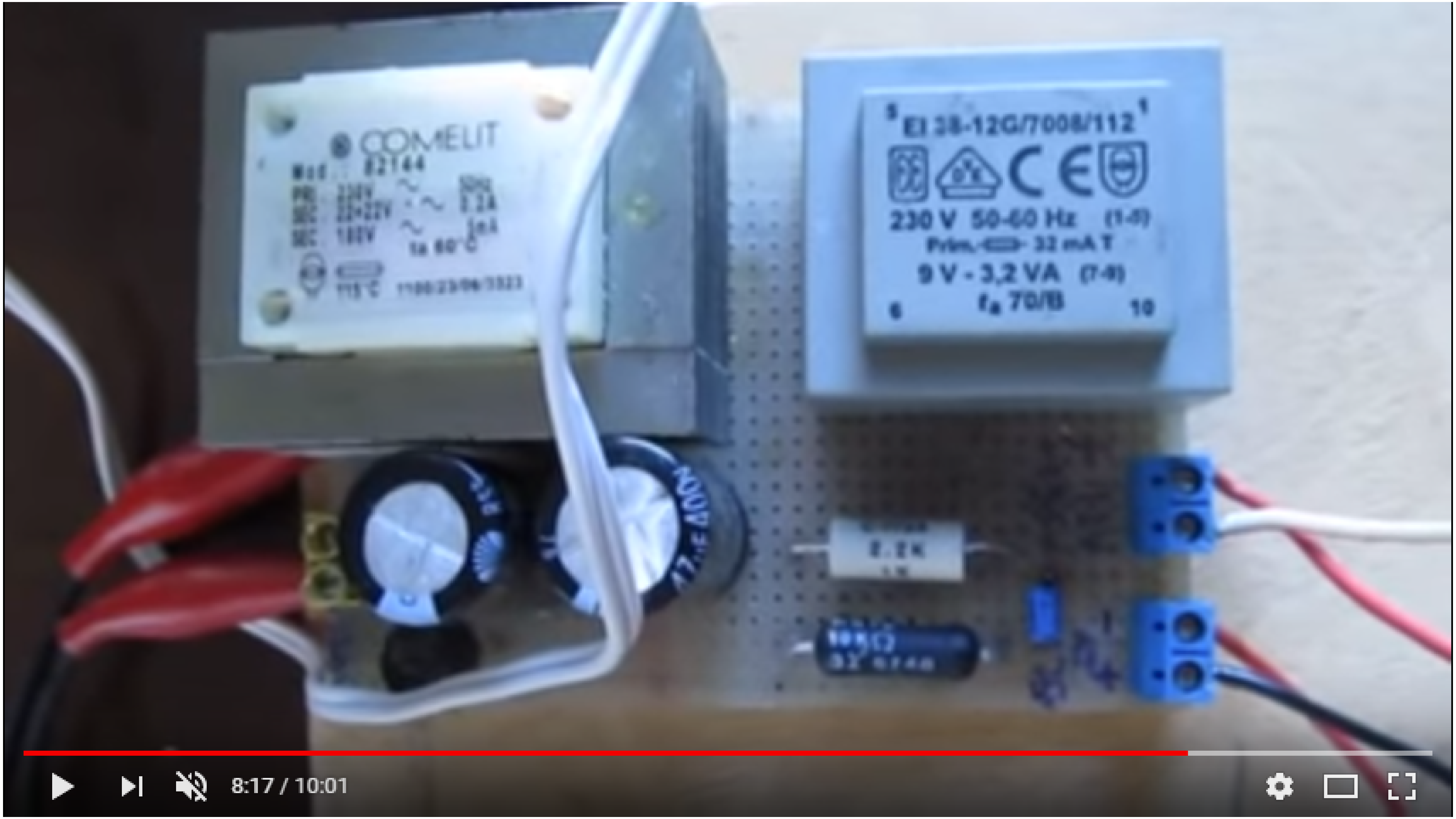
COMELIT
Mod. 82144
EI 38-12G/7008/112
230 V 50-60 Hz (1-5)
Prim. 32 mA T
9 V - 3,2 VA (7-9)
70/B
5
1
6
10
2.2K
8:17 / 10:01

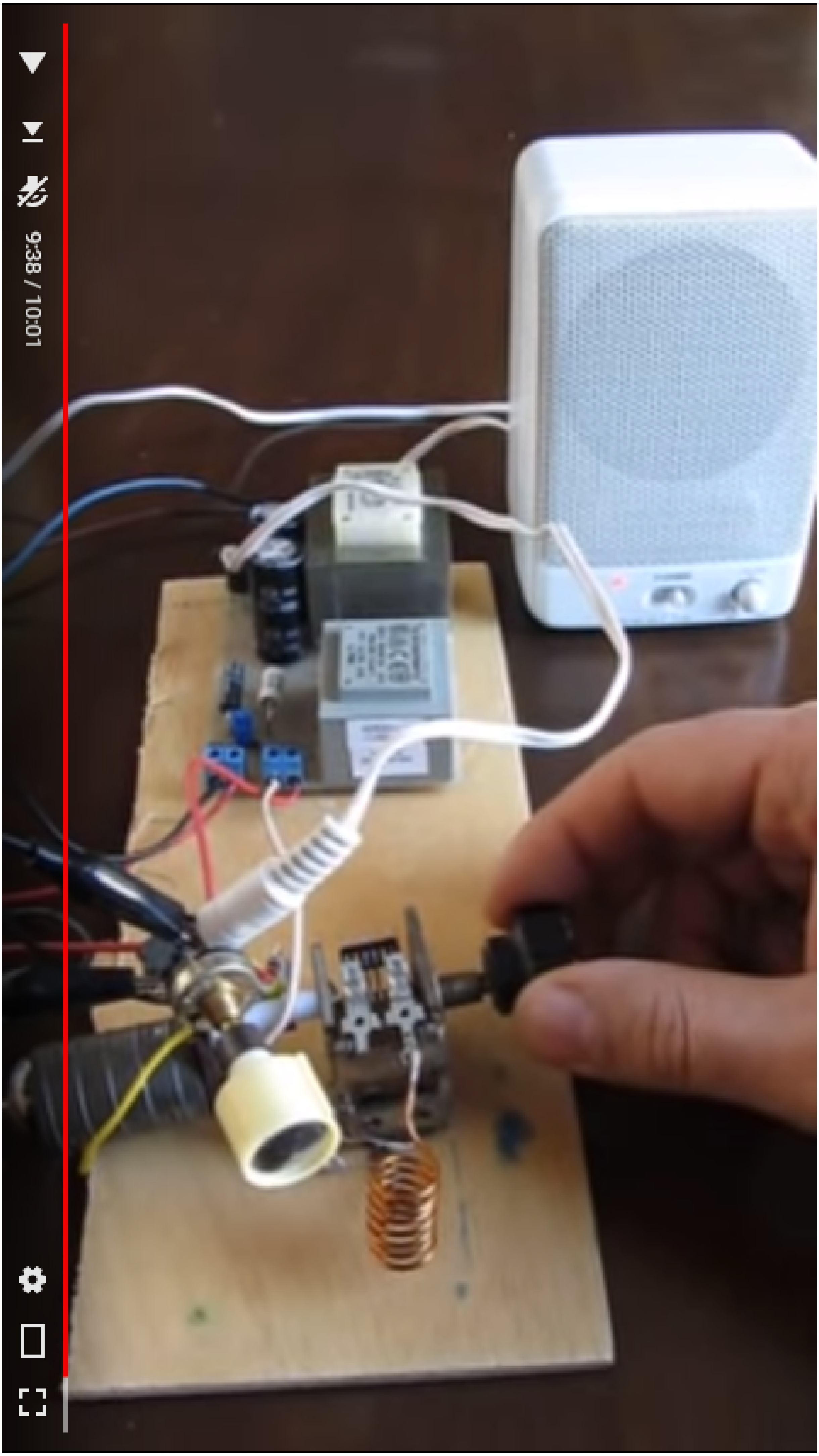
9:38 / 10:01

Superreaction receiver with ECC81
Regenerative RX with 12AT7

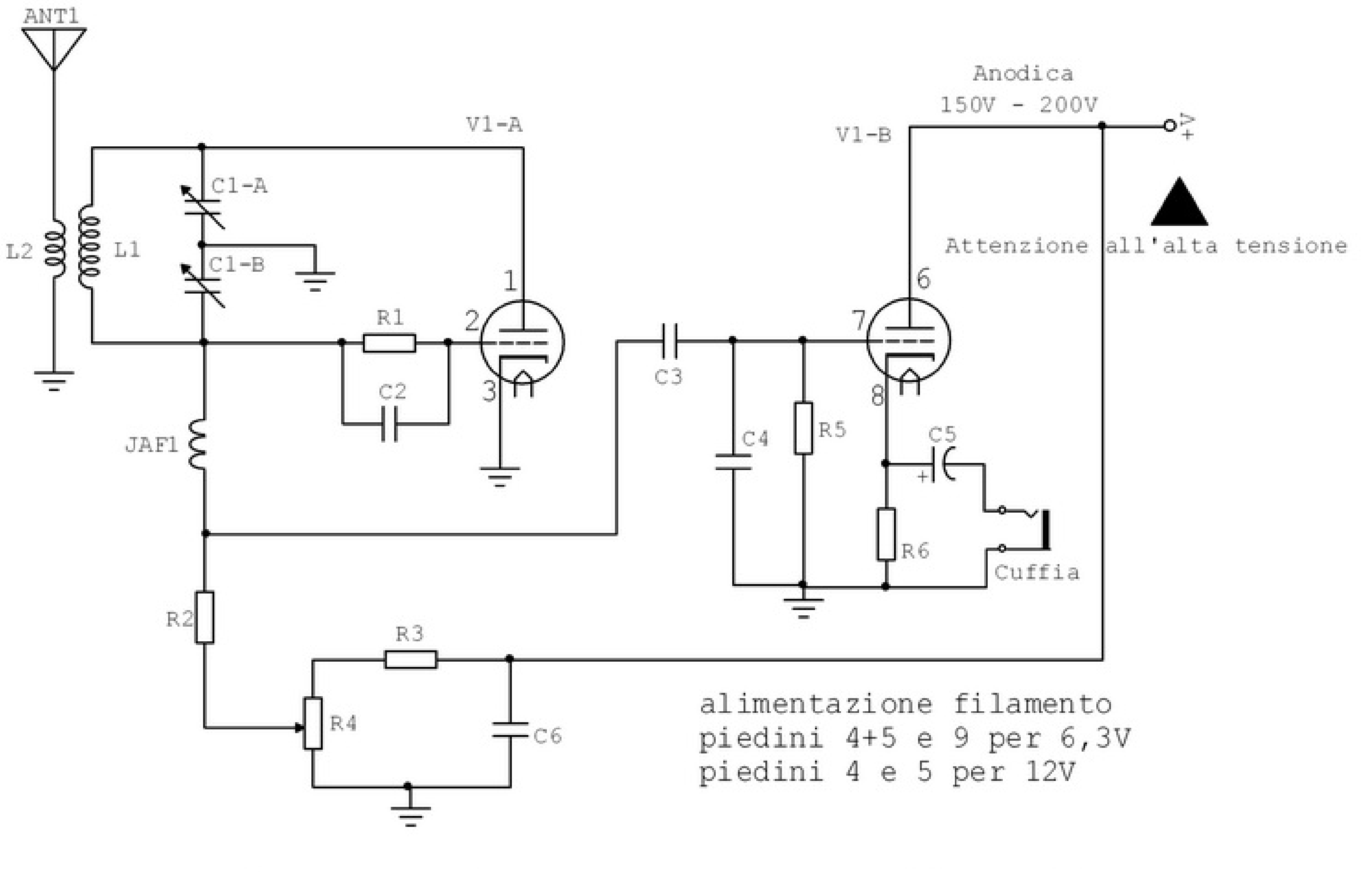
ANT1
L2
L1
C1-A
C1-B
V1-A
1
2
3
R1
C2
JAF1
R2
R3
R4
C6
C3
C4
R5
V1-B
Anodica
150V - 200V
+V
Attenzione all'alta tensione
6
7
8
C5
+
R6
Cuffia

alimentazione filamento
piedini 4+5 e 9 per 6,3V
piedini 4 e 5 per 12V

Ricevitore FM in super reazione con ECC81-12AT7

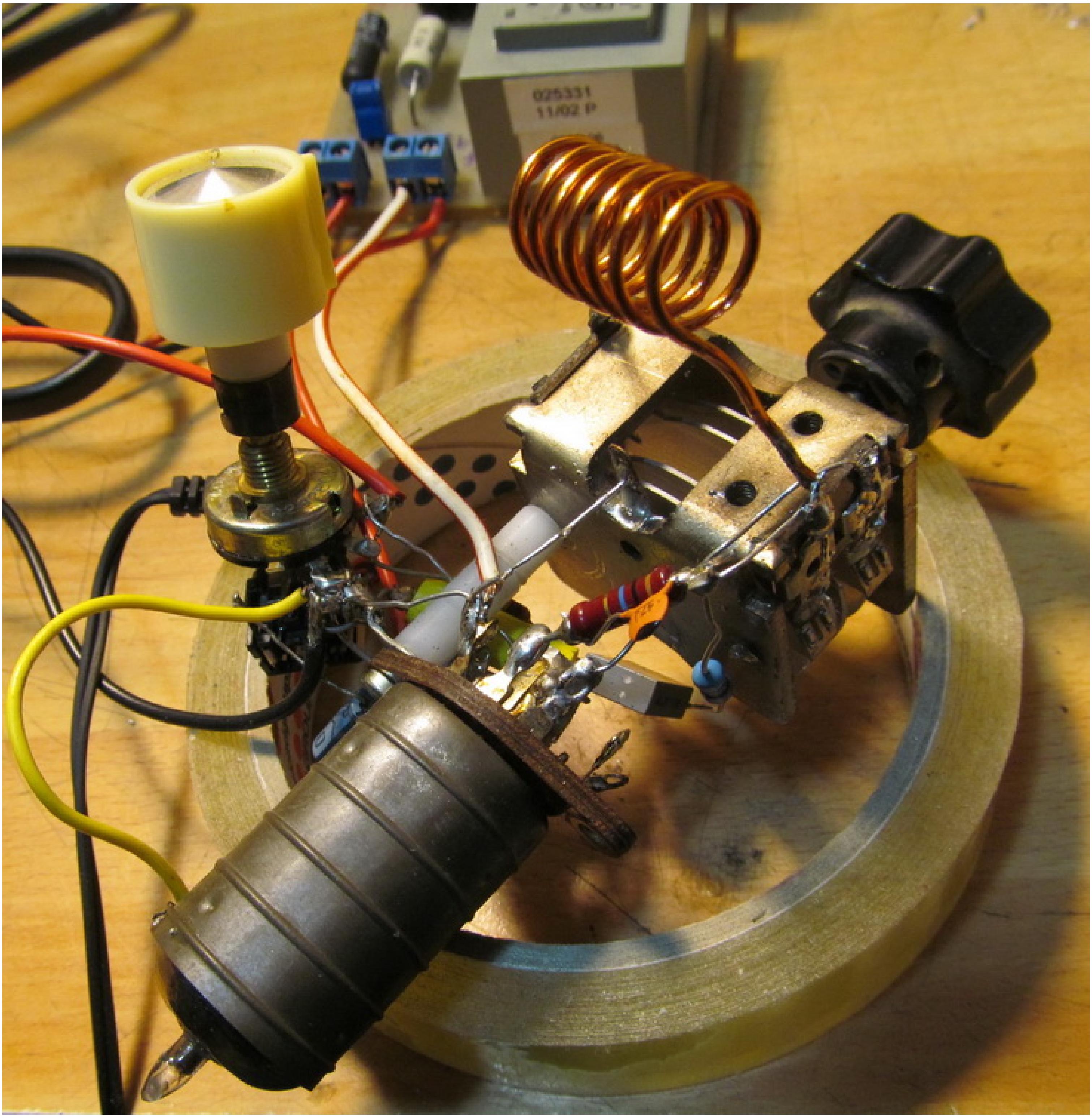
025331
11/02 P

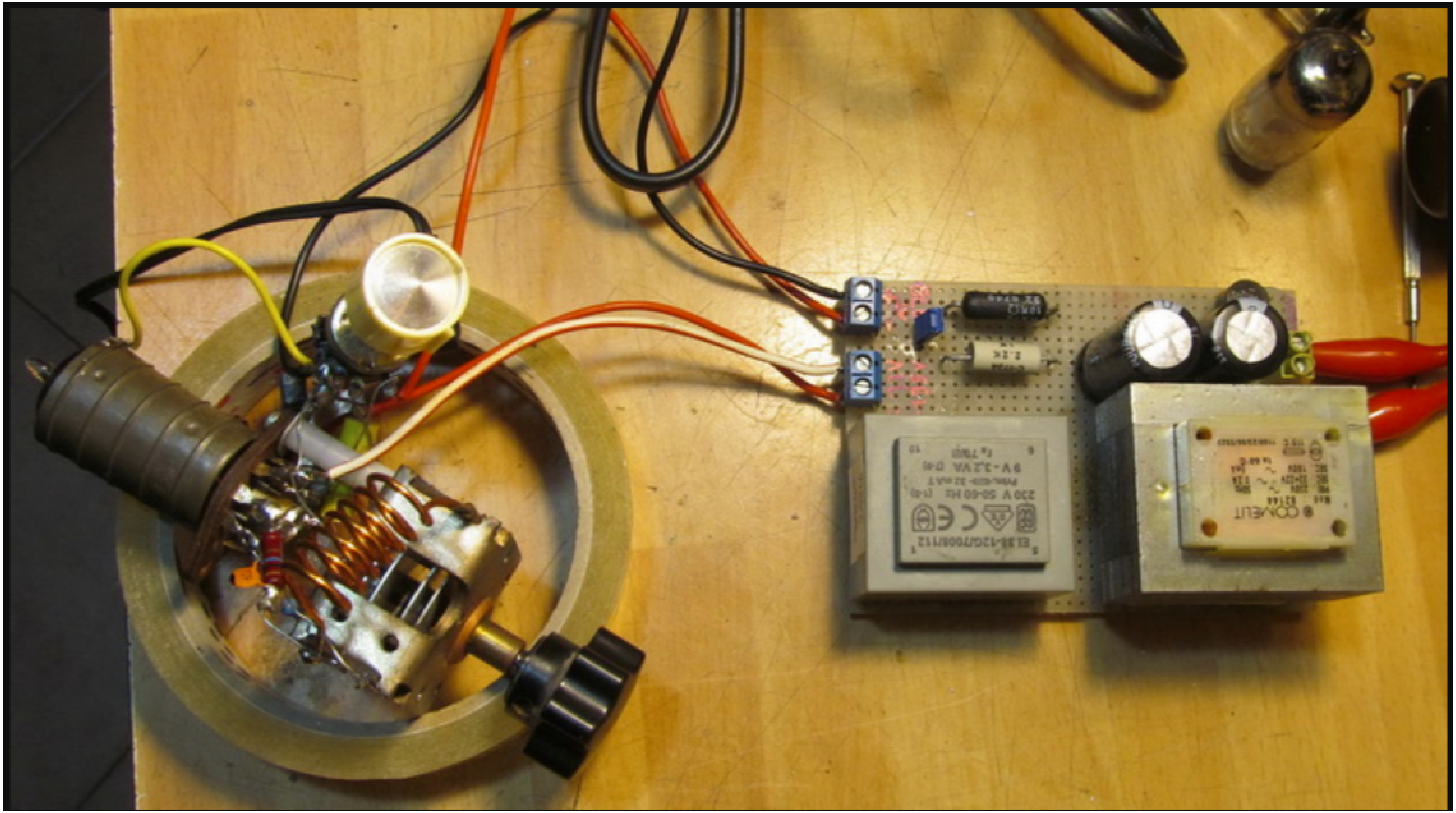

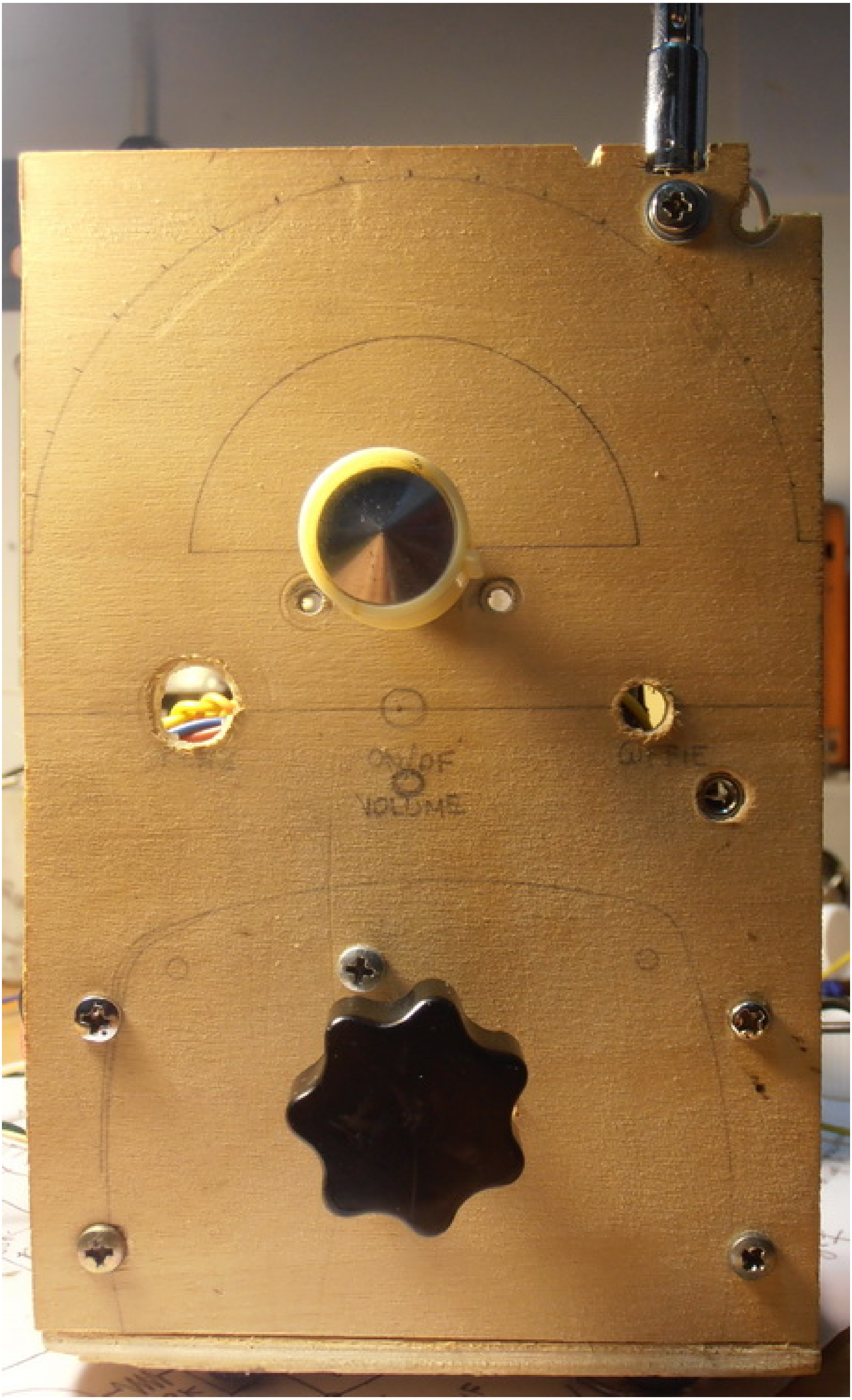
ON/OF
VOLUME

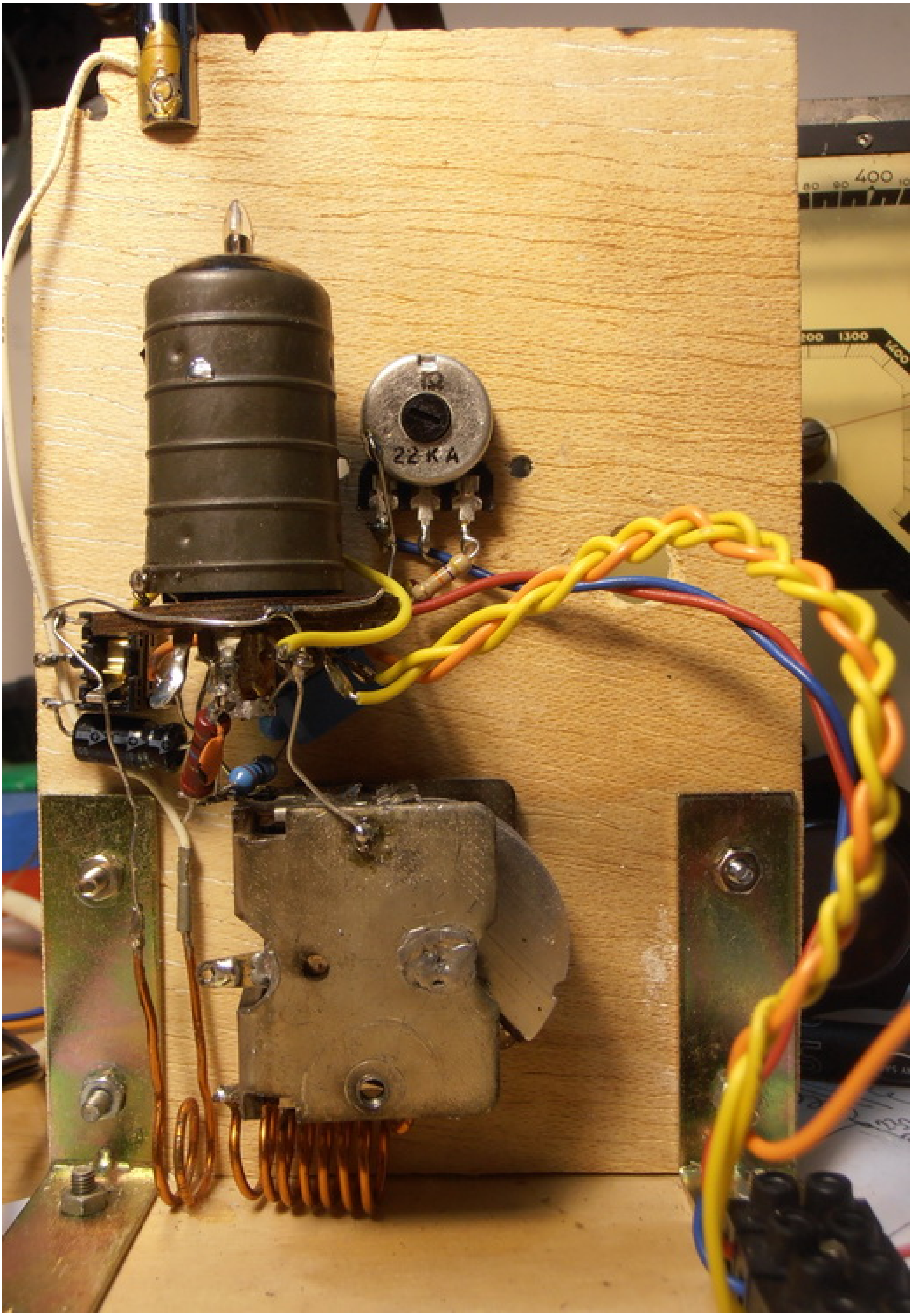
22 K A
400
80 90
1200 1300 1400

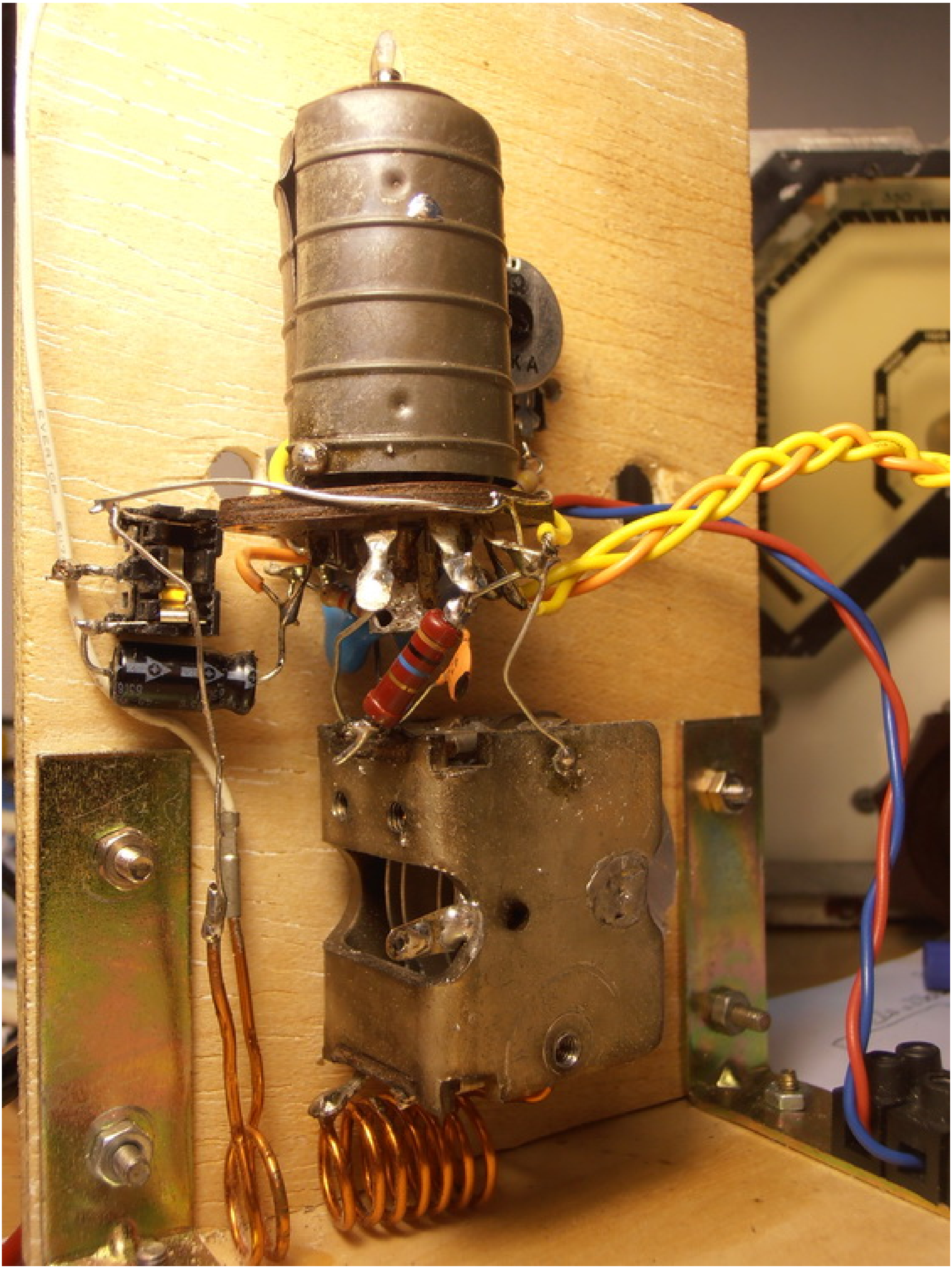

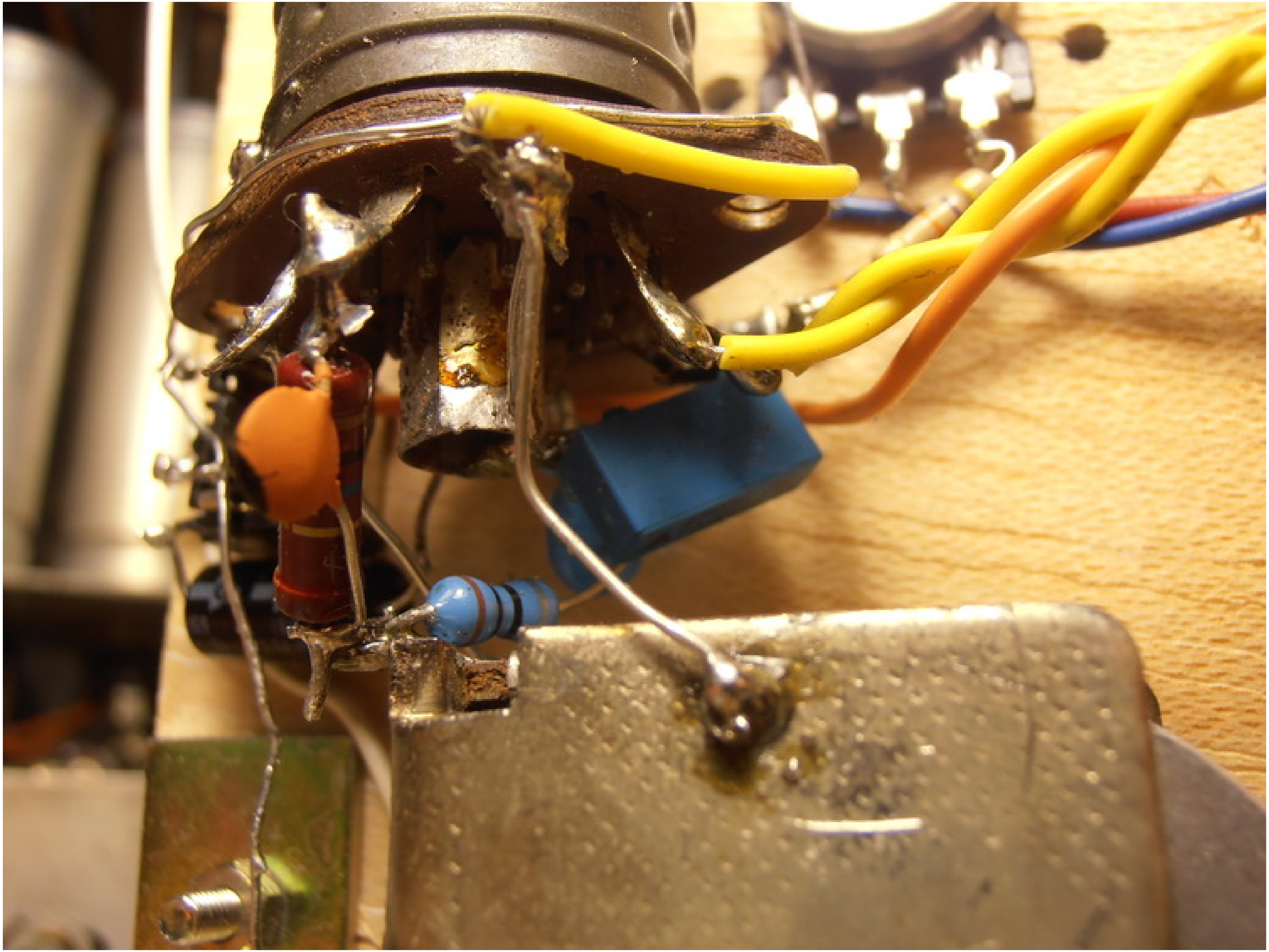

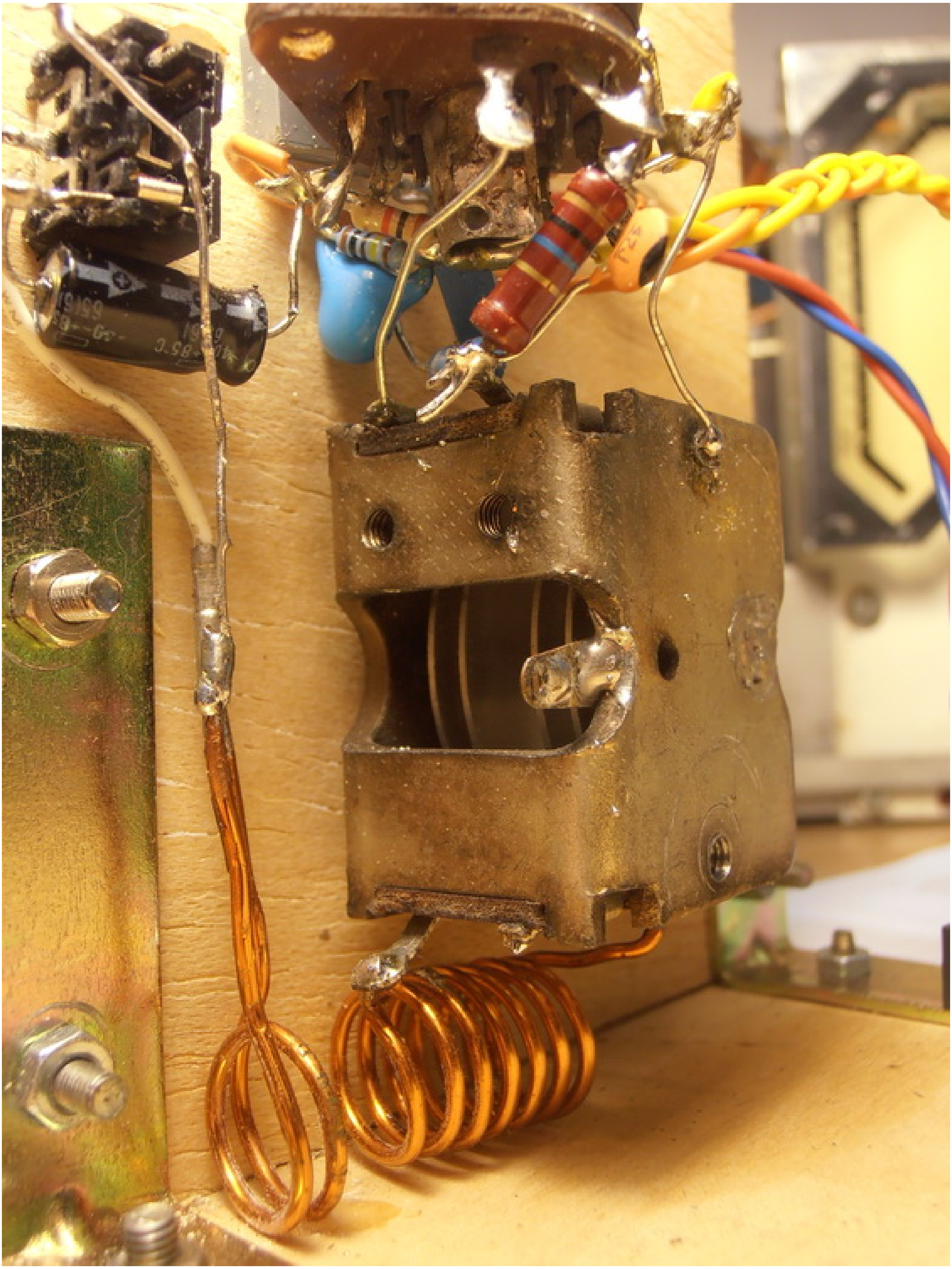

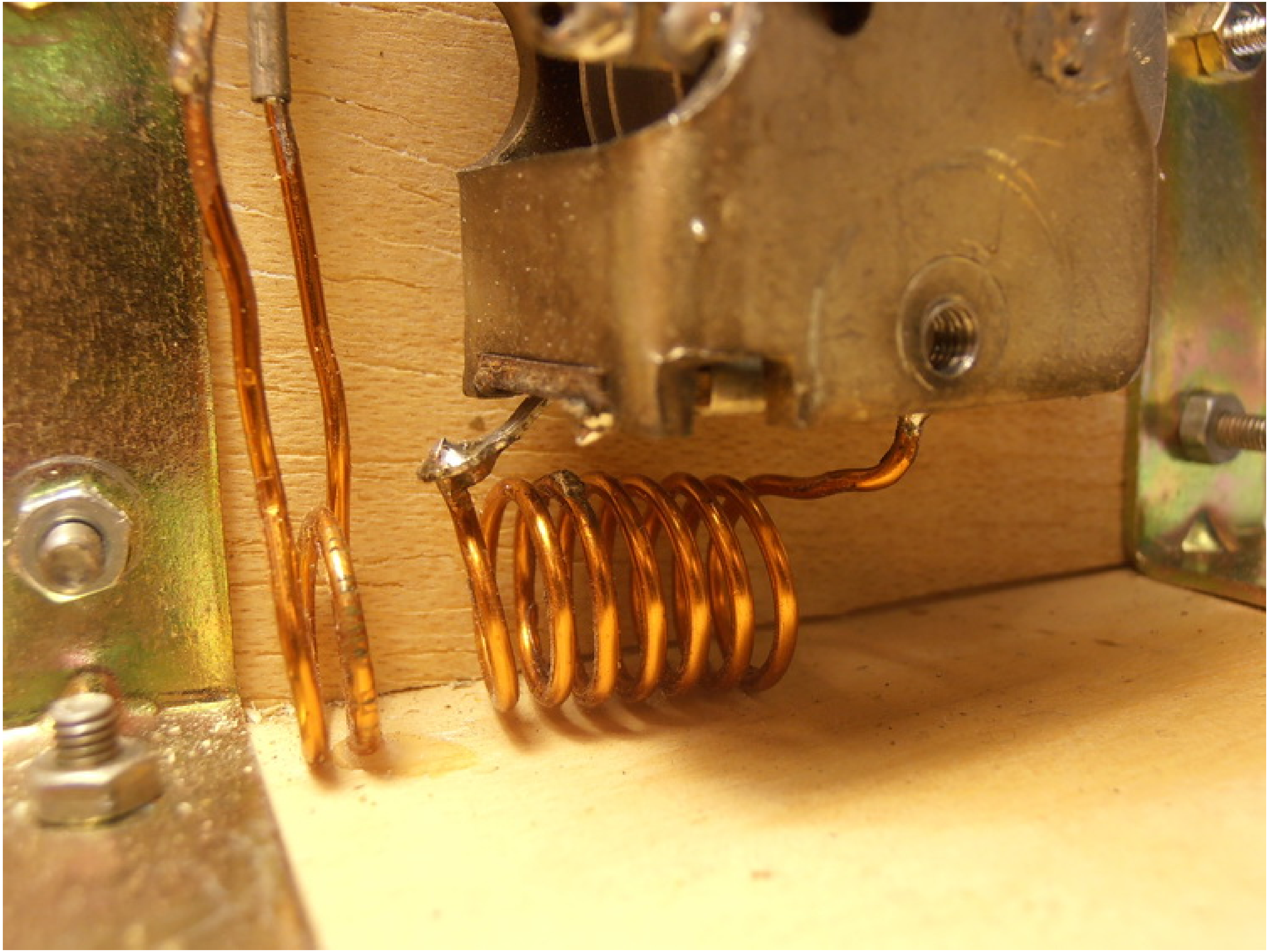

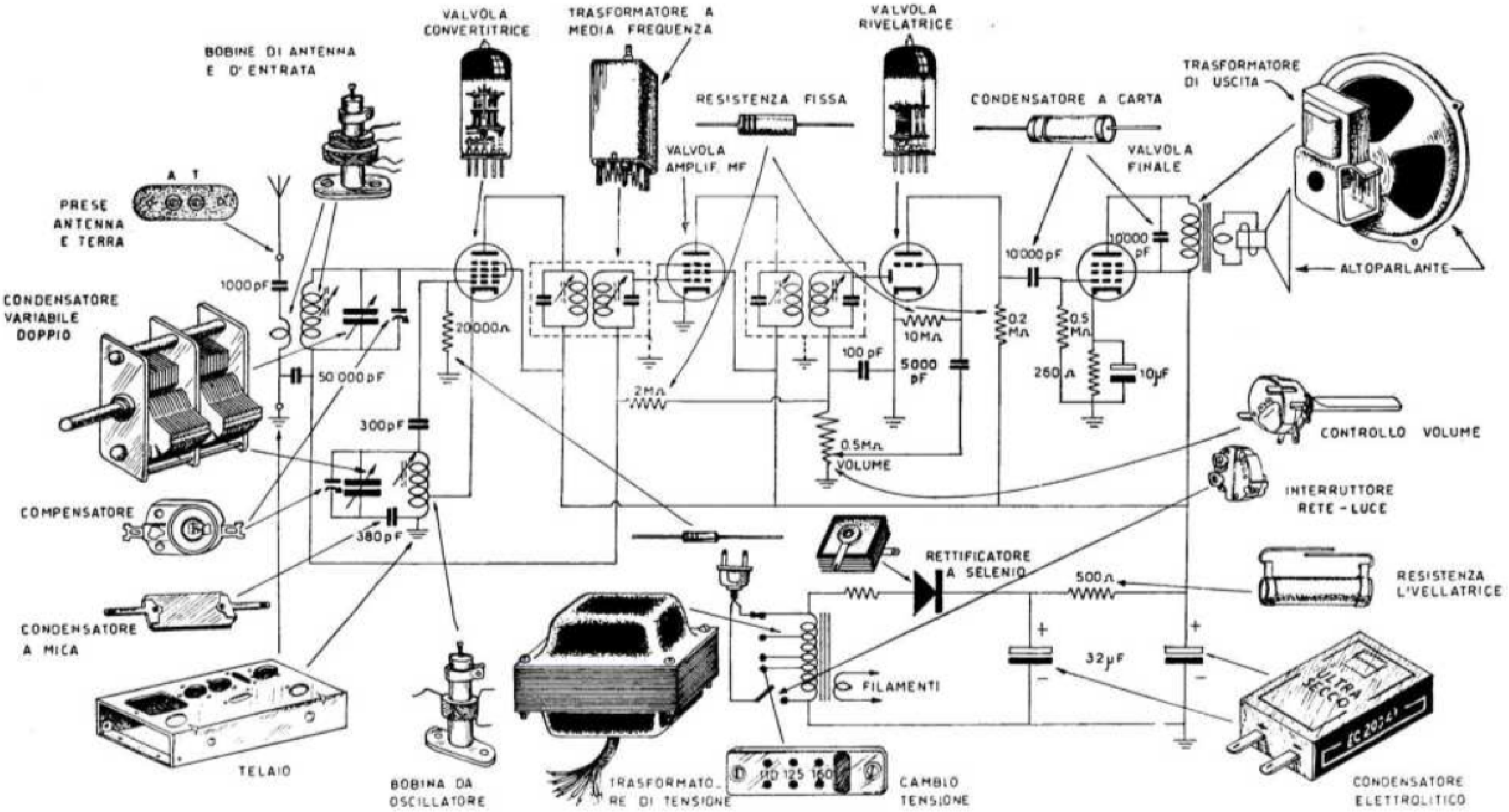
BOBINE DI ANTENNA E D'ENTRATA
VALVOLA CONVERTITRICE
TRASFORMATORE A MEDIA FREQUENZA
VALVOLA RIVELATRICE
TRASFORMATORE DI USCITA
RESISTENZA FISSA
CONDENSATORE A CARTA
A T
PRESE ANTENNA E TERRA
VALVOLA AMPLIF. MF
VALVOLA FINALE
10000 pF
ALTOPARLANTE
1000 pF
CONDENSATORE VARIABILE DOPPIO
20000 Ω
10000 pF
0.2 MΩ
0.5 MΩ
10 MΩ
100 pF
5000 pF
260 Ω
10 μF
50.000 pF
2 MΩ
CONTROLLO VOLUME
300 pF
0.5 MΩ
VOLUME
COMPENSATORE
INTERRUTTORE RETE - LUCE
380 pF
RETTIFICATORE A SELENIO
500 Ω
RESISTENZA L'IVELLATRICE
CONDENSATORE A MICA
+
32 μF
+
–
–
FILAMENTI
ULTRA SECCO
TELAIO
BOBINA DA OSCILLATORE
TRASFORMATORE DI TENSIONE
110 125 160
CAMBIO TENSIONE
CONDENSATORE ELETTROLITICO

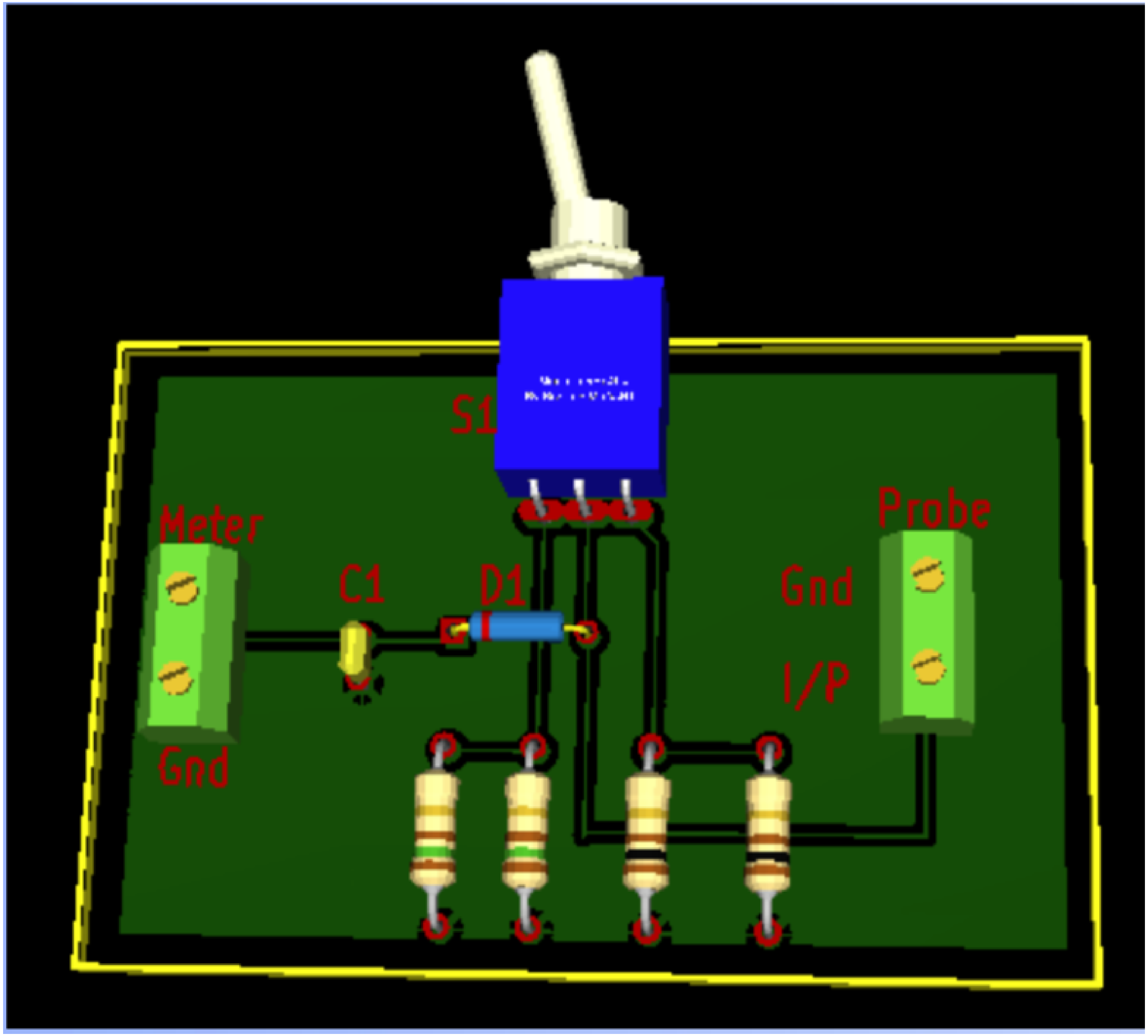
S1
Meter
C1
D1
Probe
Gnd
Gnd
I/P

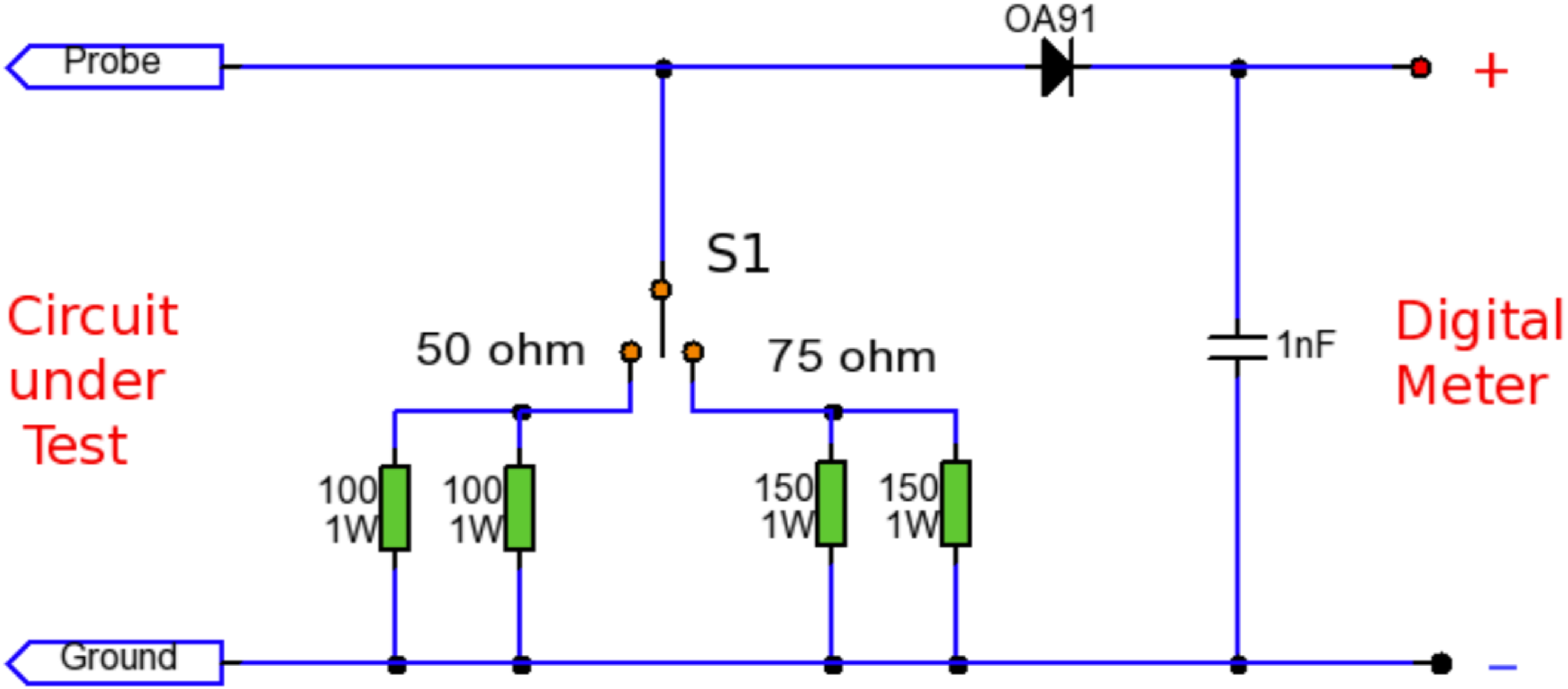
Probe
OA91
+
S1
Circuit under Test
50 ohm
75 ohm
1nF
Digital Meter
100 1W
100 1W
150 1W
150 1W
Ground
–

N. TESLA.

APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

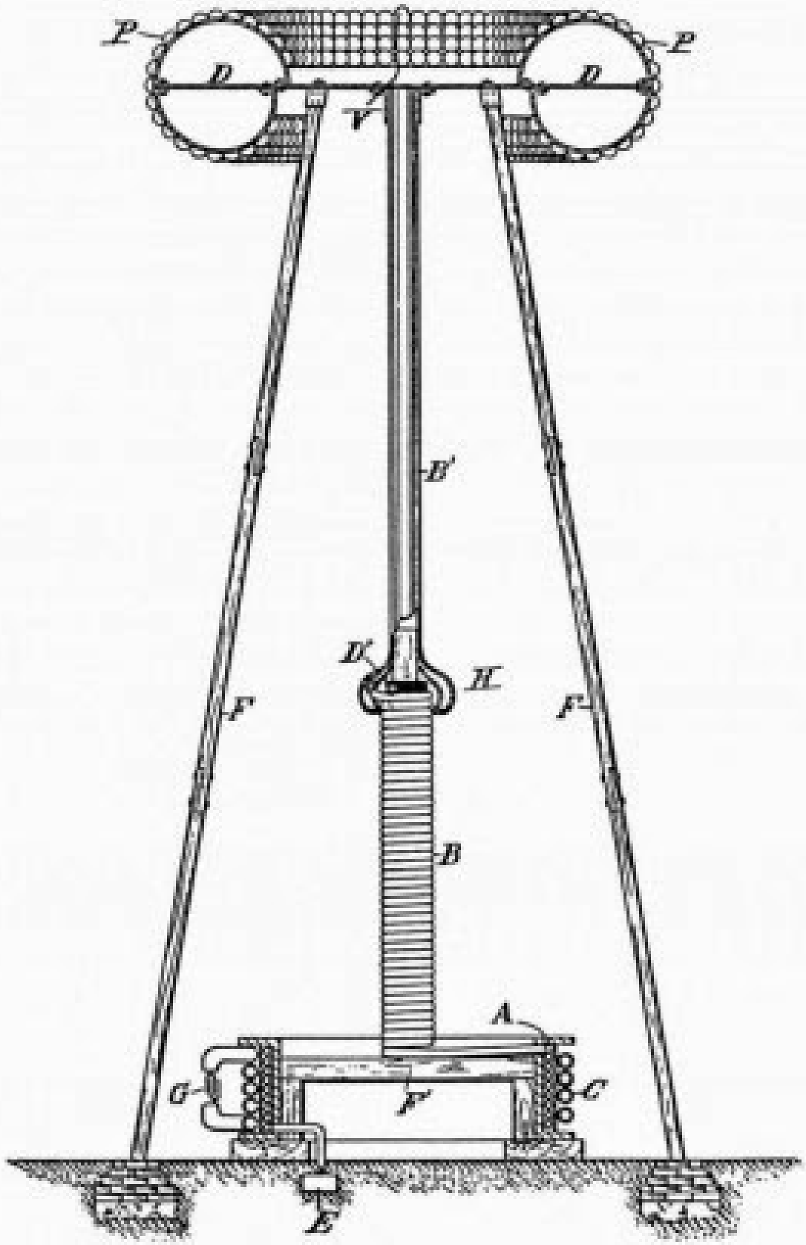
P
P
D
D
V
E'
D'
H
F
F
B
A
G
C
F'
E

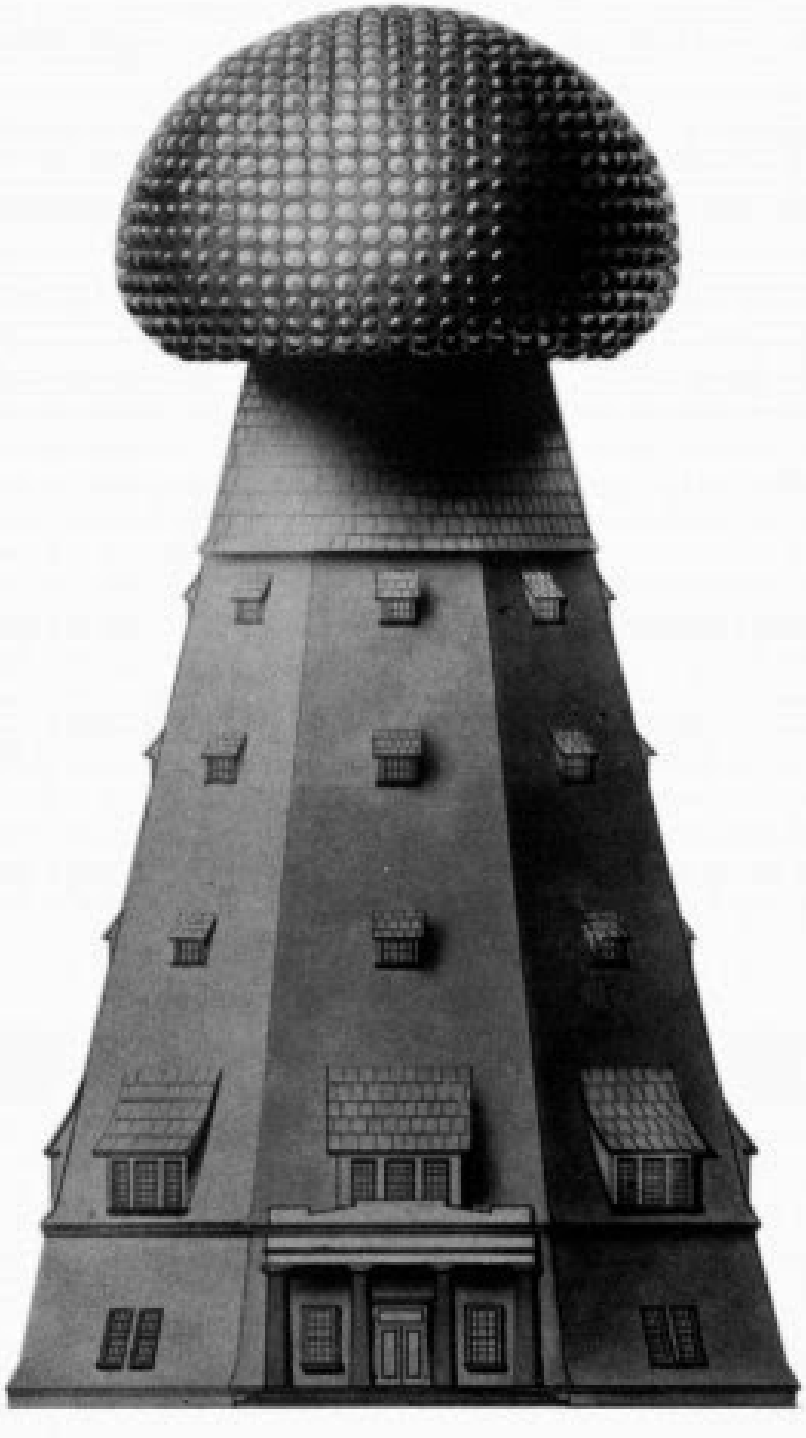

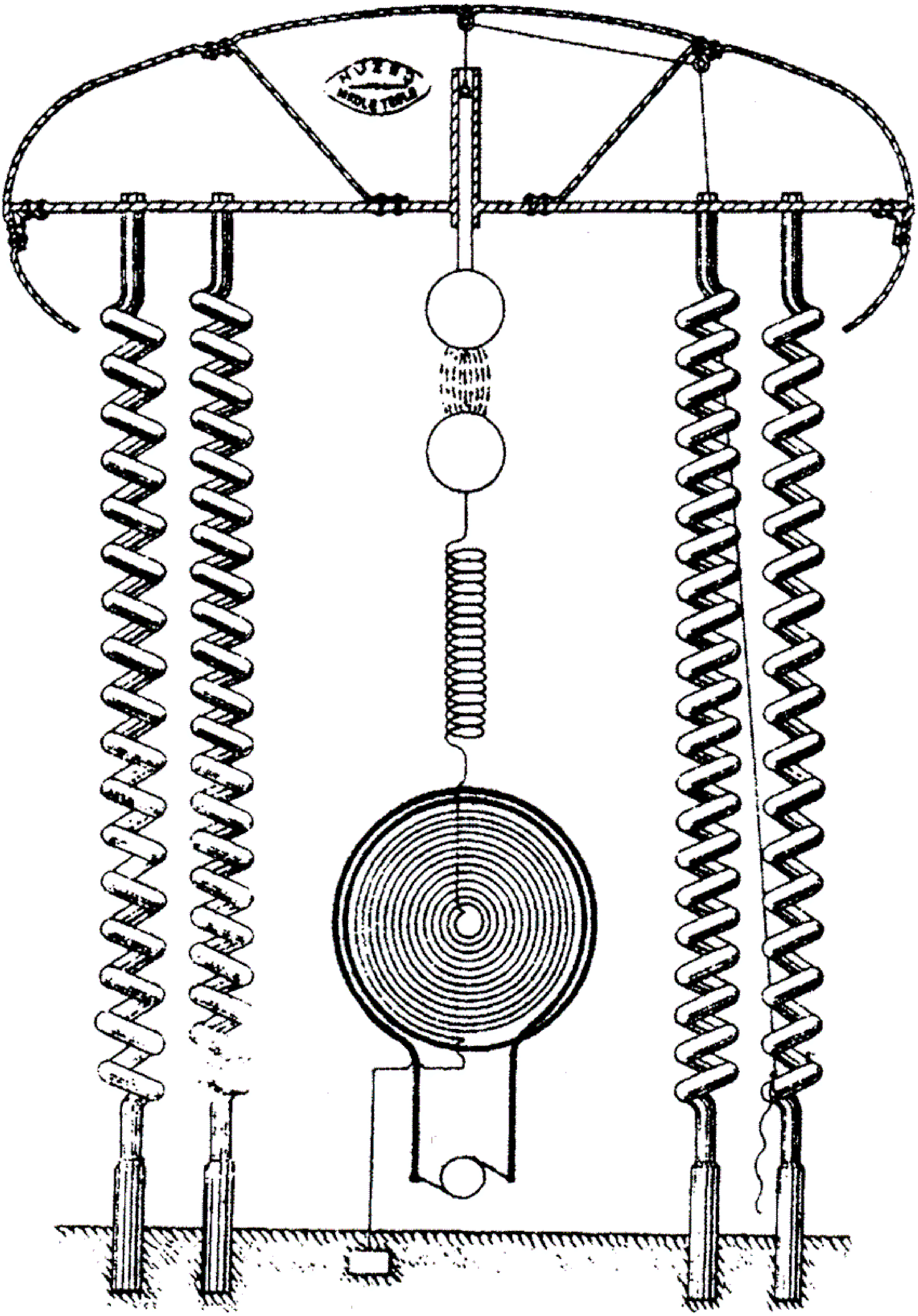

COSMIC VIBRATIONS

(THESE INTERVALS AND FIGURES ARE APPROXIMATE AND NOT EXACT)

Octave	Manifestation	Vibrations per second	Gaps
1	TOUCH	2	
2		4	
3		8	
4		16	
5	SOUND	32	
6		64	
7		128	
8		256	
9		512	
10		1,024	
11		2,048	
12		4,096	
13		8,192	
14		16,384	
15	ELECTROMAGNETIC (Longer Radio Waves)	32,768	
16		65,536	
17		131,072	
18		262,144	
19		524,288	
20		1,048,576	
21		2,097,152	
22		4,194,304	
23		8,388,608	
24		16,777,216	
25		33,554,432	
26		67,108,864	
27		134,217,728	
28		268,435,456	
29	ELECTROMAGNETIC (Microwaves)	536,870,912	
30		1,073,741,824	
31		2,147,483,648	
32		4,294,967,296	
33		8,589,934,592	
34		17,179,869,184	
35		34,359,738,368	
36		68,719,476,736	
37		137,438,953,472	
38		274,877,906,944	
39	HEAT OR INFRARED	549,755,813,888	
40		1,099,511,627,776	
41		2,199,023,255,552	
42		4,398,046,511,104	
43		8,796,093,022,208	
44		17,592,186,044,416	
45		35,184,372,088,832	
46		70,368,744,177,664	
47		140,737,488,355,328	
48	VISIBLE LIGHT	281,474,976,710,656	
49		562,949,953,421,312	
50		1,125,899,906,842,624	
51	ULTRAVIOLET	2,251,799,813,685,248	SPIRIT ELECTRONS
52		4,503,599,627,370,496	
53		9,007,199,254,740,992	
54		18,014,398,509,481,984	
55		36,028,797,018,963,968	
56		72,057,594,037,927,936	
57	X-RAY	144,115,188,075,855,872	
58		288,230,376,151,711,744	
59		576,460,752,303,423,488	
60			PSYCHIC PROJECTIONS
61			
62	GAMMA AND COSMIC RAYS	4,611,686,018,427,387,904	
63			
64			
65			
66			
67			
68			
69			
70			
71			
72			SOUL ESSENCE
73			
74			
75			
76			
77			
78			
79			
80			

The Vibrations in these octaves increase to an incomprehensible number per second.

The phenomena of psychic projections and soul essence occur in octaves considerably beyond those of any known material manifestations. Consequently, they must not be considered as immediately following any manifestations which we have shown on this chart.

NOTICE

This chart was carefully compiled by the Editorial and Research Departments of AMORC College after consulting the most eminent scientific authorities on the subjects of physics and chemistry and comparing their observations and notations with the reports from the Rosicrucian laboratories in foreign countries.

There is a slight disagreement in regard to the rates of vibrations for the various keys of the musical keyboard, due to different standards of pitch in various countries, and there is some disagreement among scientific authorities in regard to the colors associated with various musical notes. This chart represents not only the consensus of most of these authorities but also adheres strictly to what the Rosicrucians have been able to demonstrate as truth in our own laboratories and in the experiments of thousands of members in various parts of the world. Regardless, therefore, of any disagreement with authorities outside the organization, this chart is the authentic presentation of the Rosicrucian knowledge.

This chart is to be used by the members of the Rosicrucian Order in connection with their studies and experiments and is not to be used for any other purpose or in any other way. Like the lessons themselves, this chart and the facts contained on it are reserved and restricted in use under the rules and regulations of the organization.

INTERNATIONAL
SEAL OF THE
SUPREME LODGE
of the
Rosicrucian
Order

THE SOLAR SPECTRUM

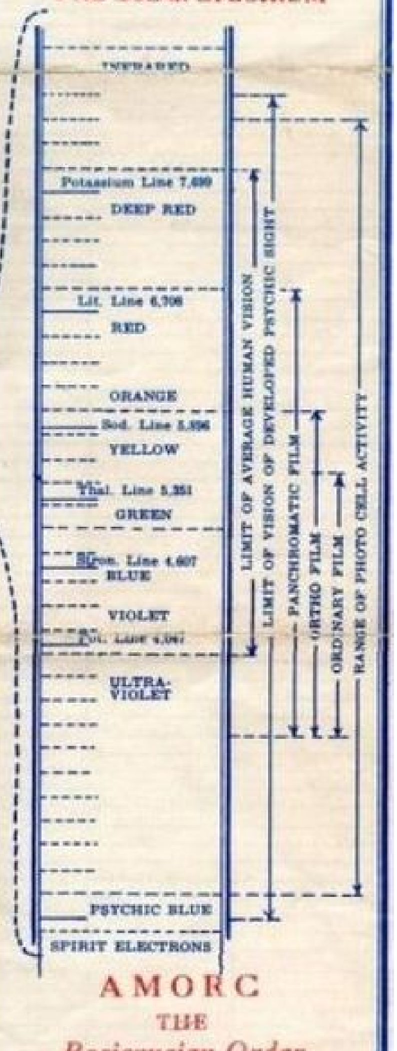

AMORC
THE
Rosicrucian Order
San Jose, California

LITHO. IN U.S.A.

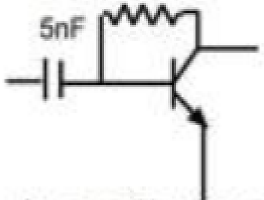

In sostituzione dell'EPAD,
ma richiede almeno 1 V
di alimentazione rubata

vista da sotto

Realizzazione
bobina in filo argentato diametro 1,5
diametro medio 14 mm
lunghezza 19,5 mm
presa a 3,5 spire
var. capacità 7,5-11,5 pf
collegamento a terra non necessario

NOTA
con un buon segnale
funziona senza amplificatore

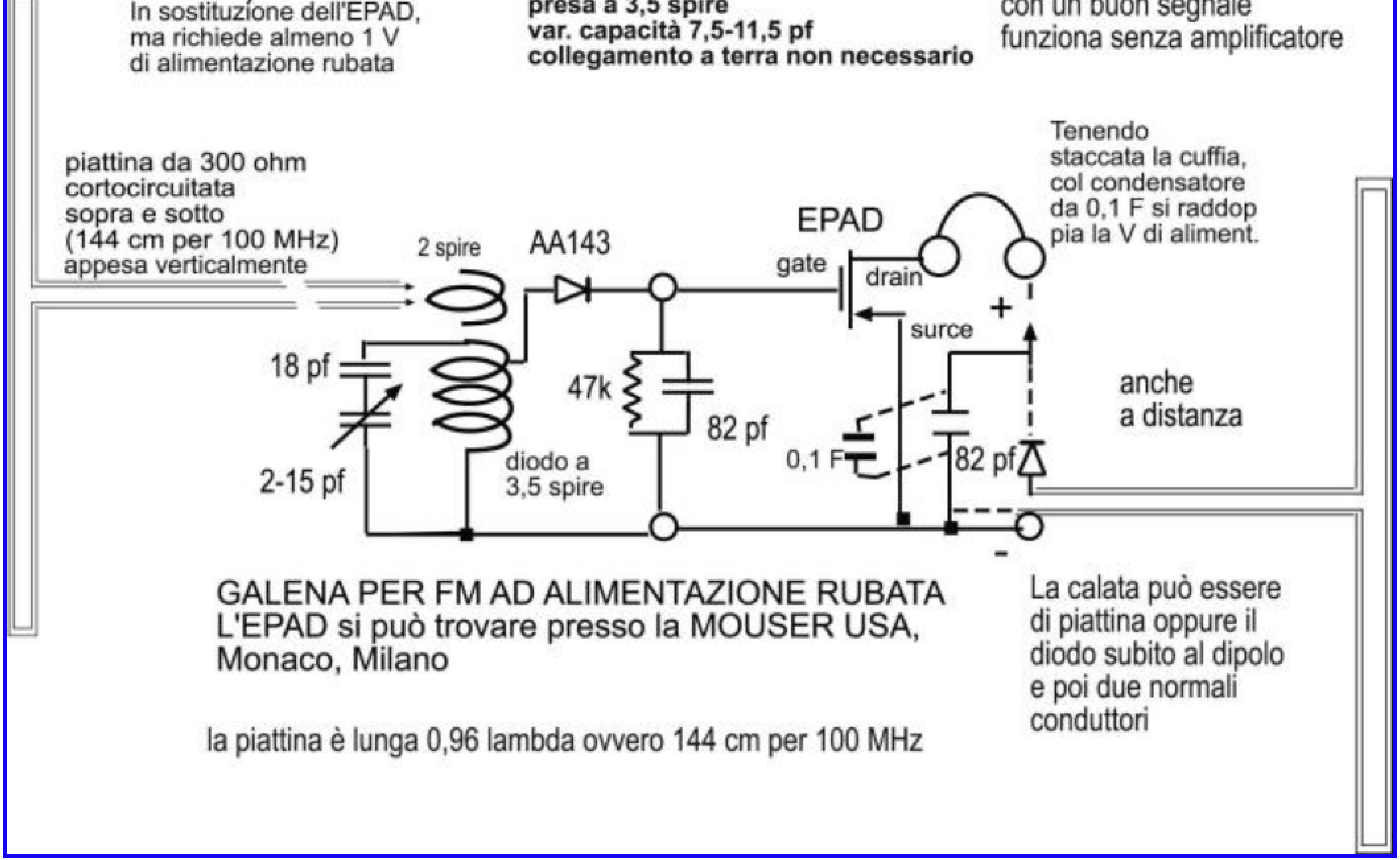
2 spire
AA143
EPAD
gate
drain
surce
+
-
18 pf
2-15 pf
diodo a
3,5 spire
47k
82 pf
0,1 F
82 pf

piattina da 300 ohm
cortocircuitata
sopra e sotto
(144 cm per 100 MHz)
appesa verticalmente

Tenendo
staccata la cuffia,
col condensatore
da 0,1 F si raddop
pia la V di aliment.

anche
a distanza

La calata può essere
di piattina oppure il
diodo subito al dipolo
e poi due normali
conduttori

GALENA PER FM AD ALIMENTAZIONE RUBATA

L'EPAD si può trovare presso la MOUSER USA,
Monaco, Milano

la piattina è lunga 0,96 lambda ovvero 144 cm per 100 MHz

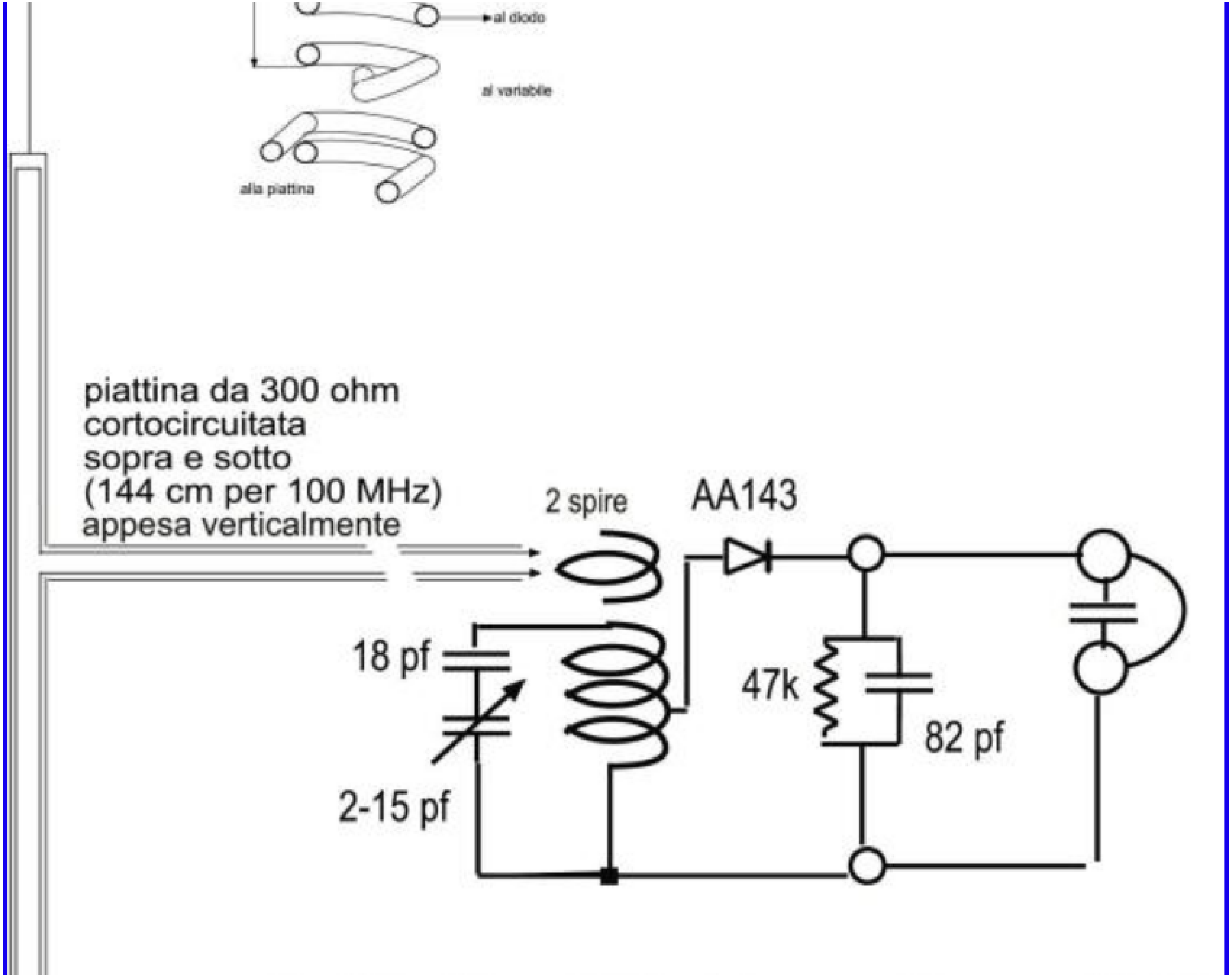
al diodo
al variabile
alla piattina
piattina da 300 ohm
cortocircuitata
sopra e sotto
(144 cm per 100 MHz)
appesa verticalmente
2 spire
AA143
18 pf
2-15 pf
47k
82 pf

la piattina è lunga 0,96 lambda ovvero 144 cm per 100 MHz

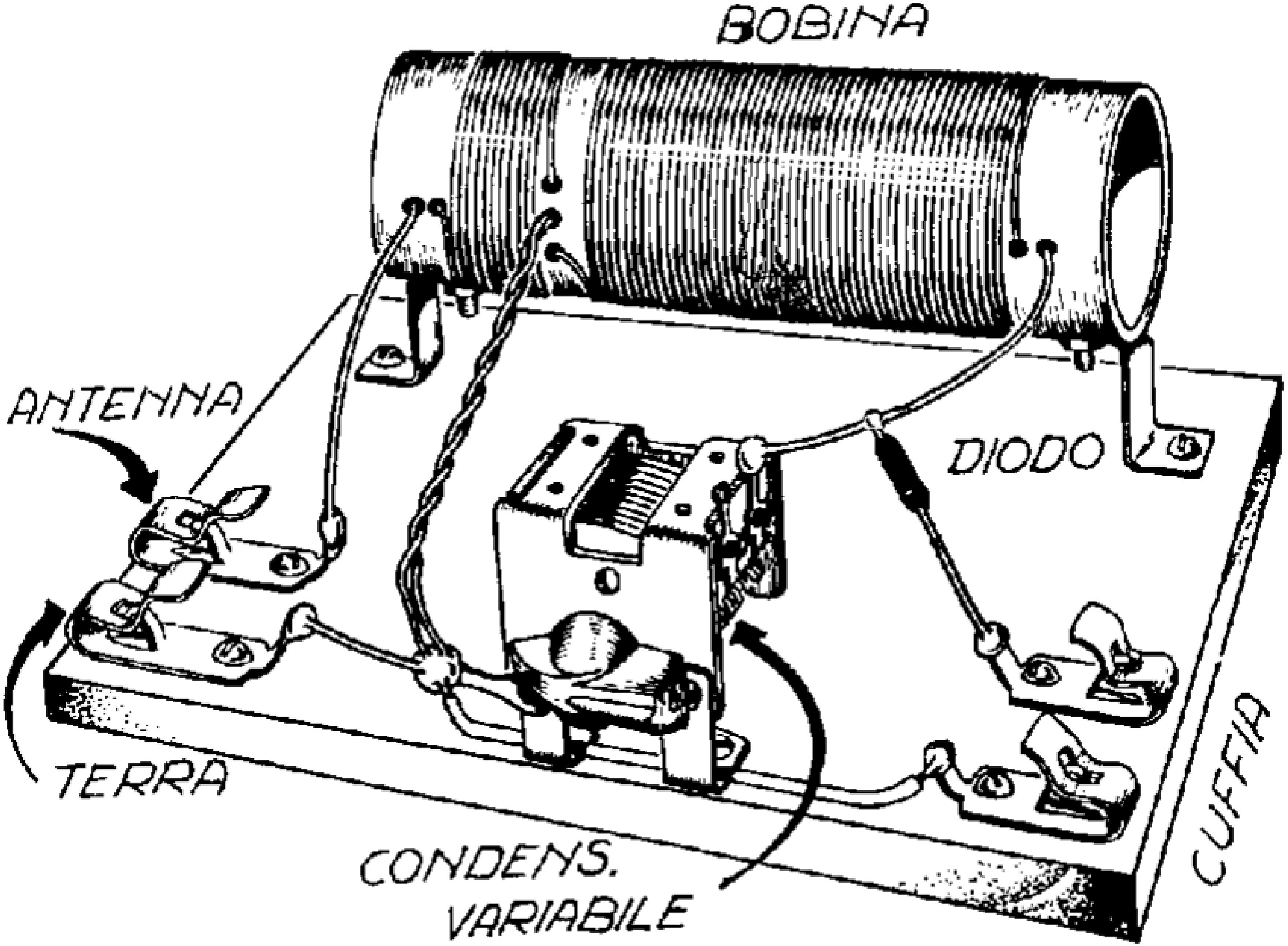
BOBINA
ANTENNA
DIODO
TERRA
CONDENS.
VARIABILE
CUFFIA

Lead sulfide crystals, also called "galena"

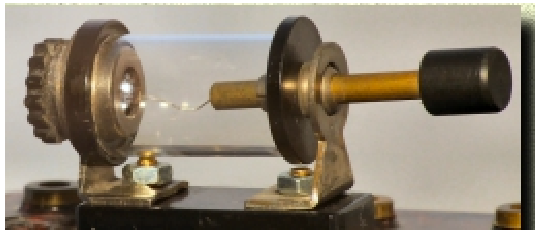

Crystal detector, the lever was used to position the spring contact on a crystal face capable of performing the radio signal detection.

Yeah , but how does a galena radio work? Here is the constructive scheme of a receiver of this kind, obtained from a book of popular radio commu
to the knowledge of the radio, ed.Hepep 1943):

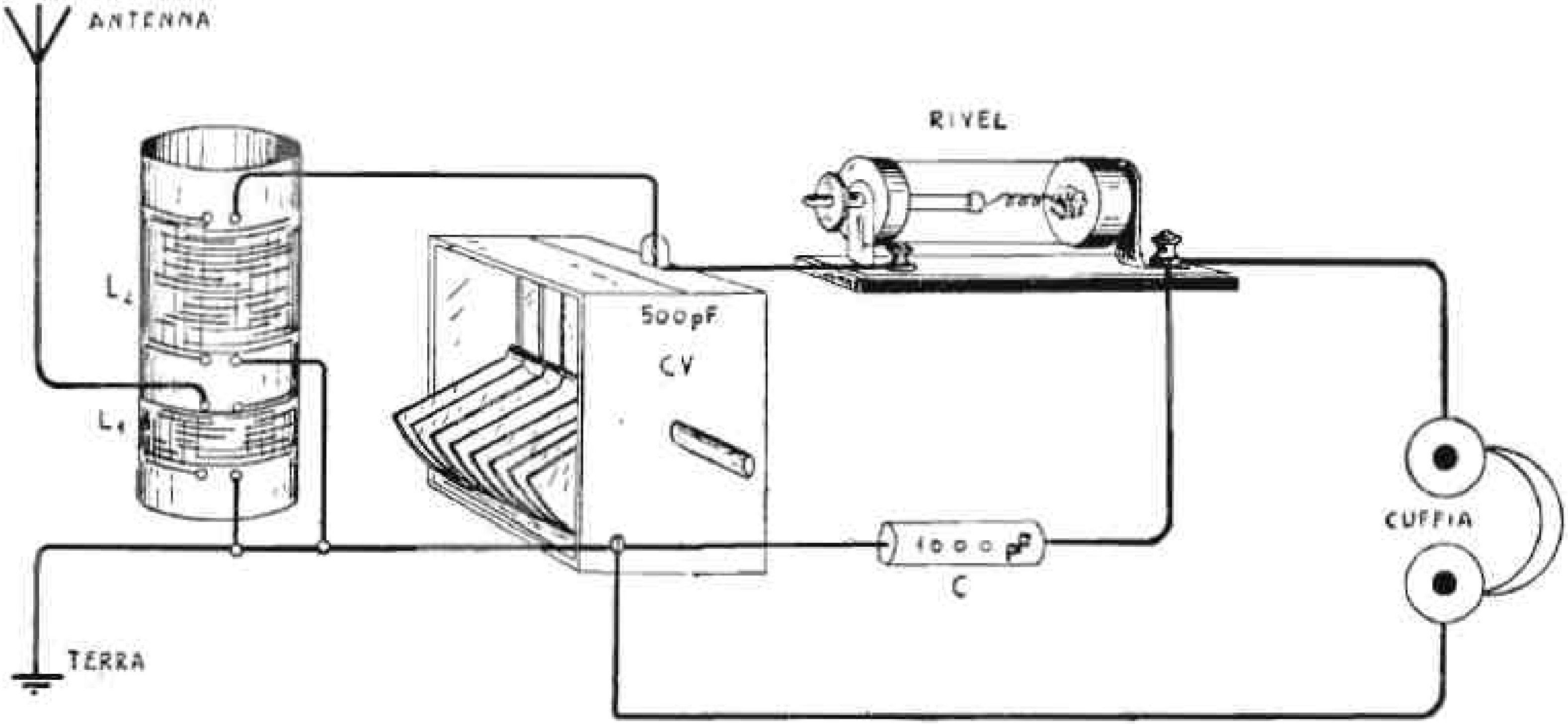

Fig. 13.5. – Realizzazione pratica dello schema di fig. 13.3 con aggiunta della bobina d'antenna.

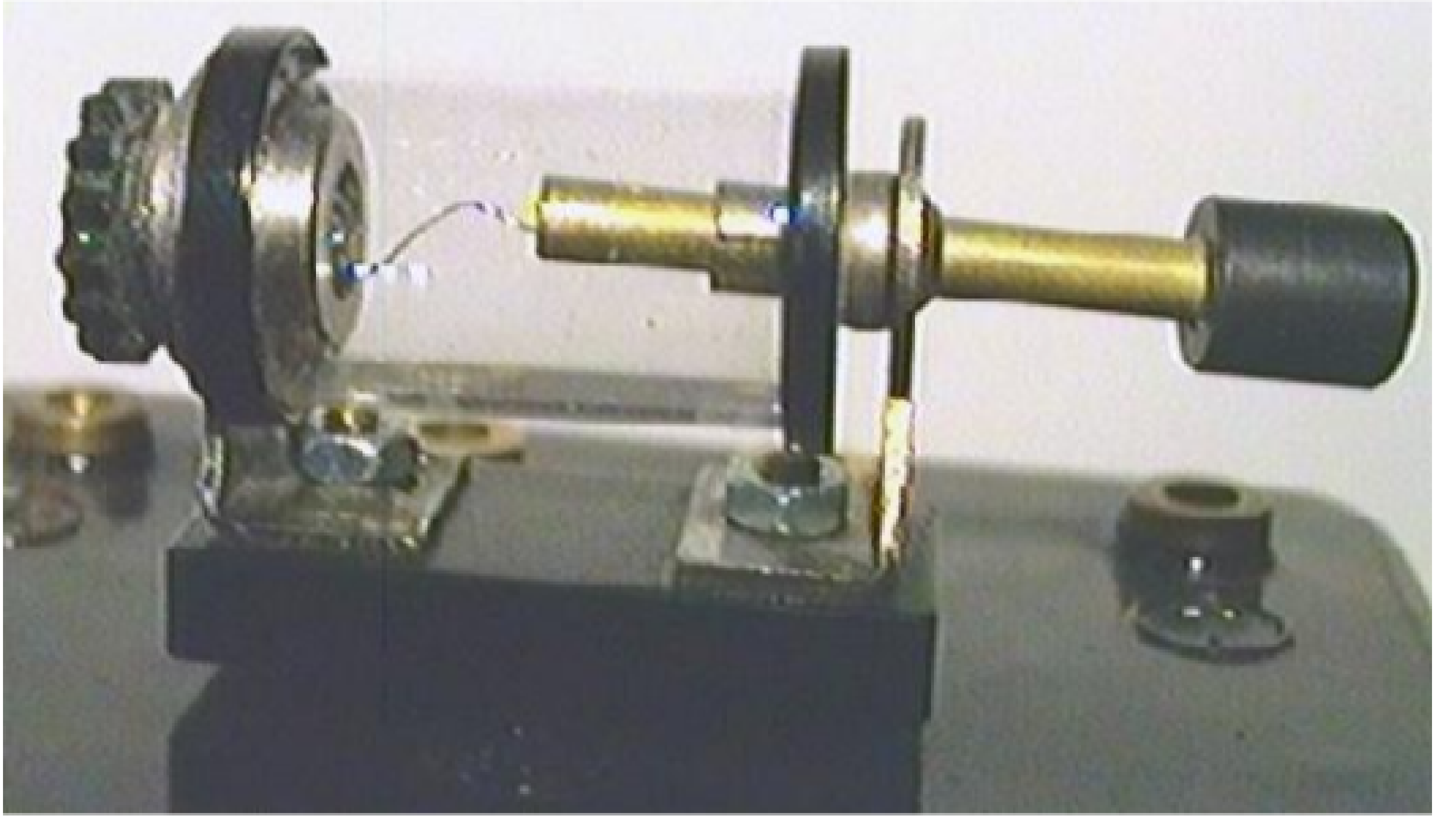

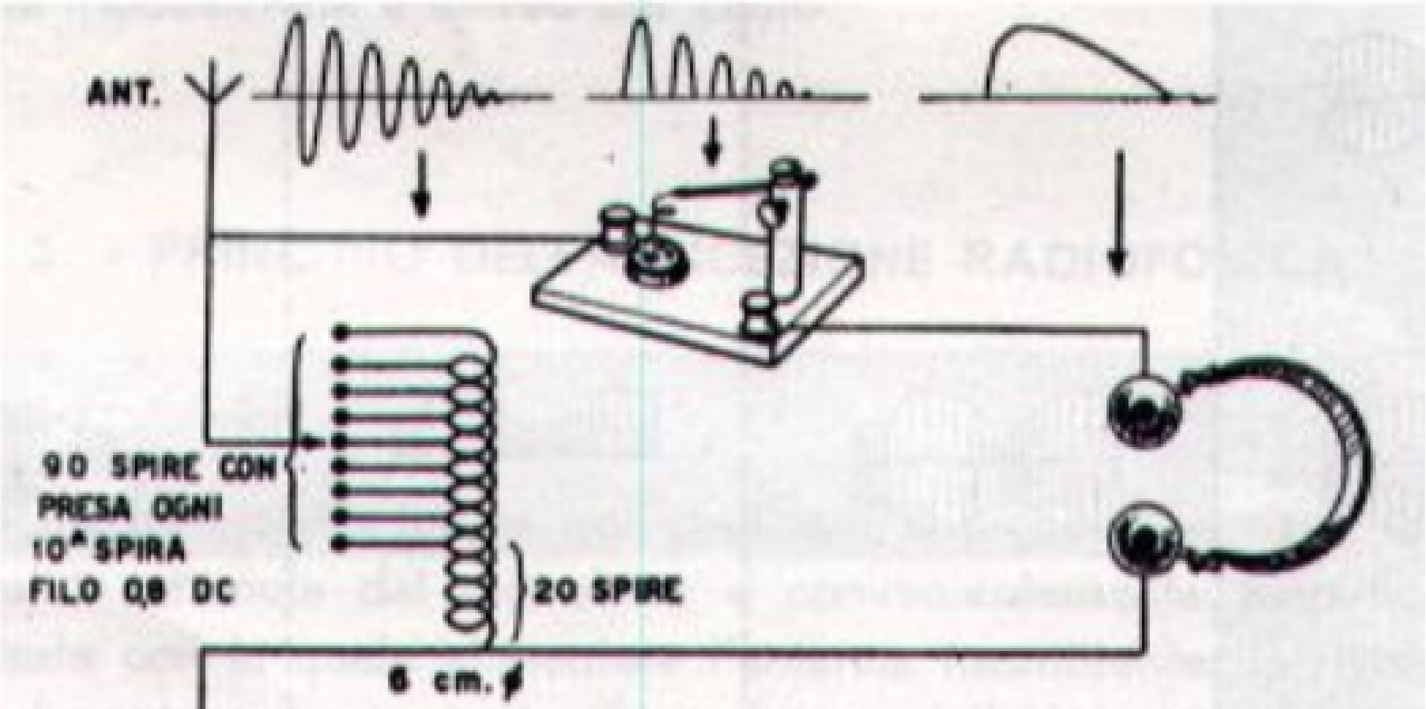
ANT.
90 SPIRE CON
PRESA OGNI
10^A SPIRA
FILO 0,8 DC
20 SPIRE
6 cm. ∅
TERRA

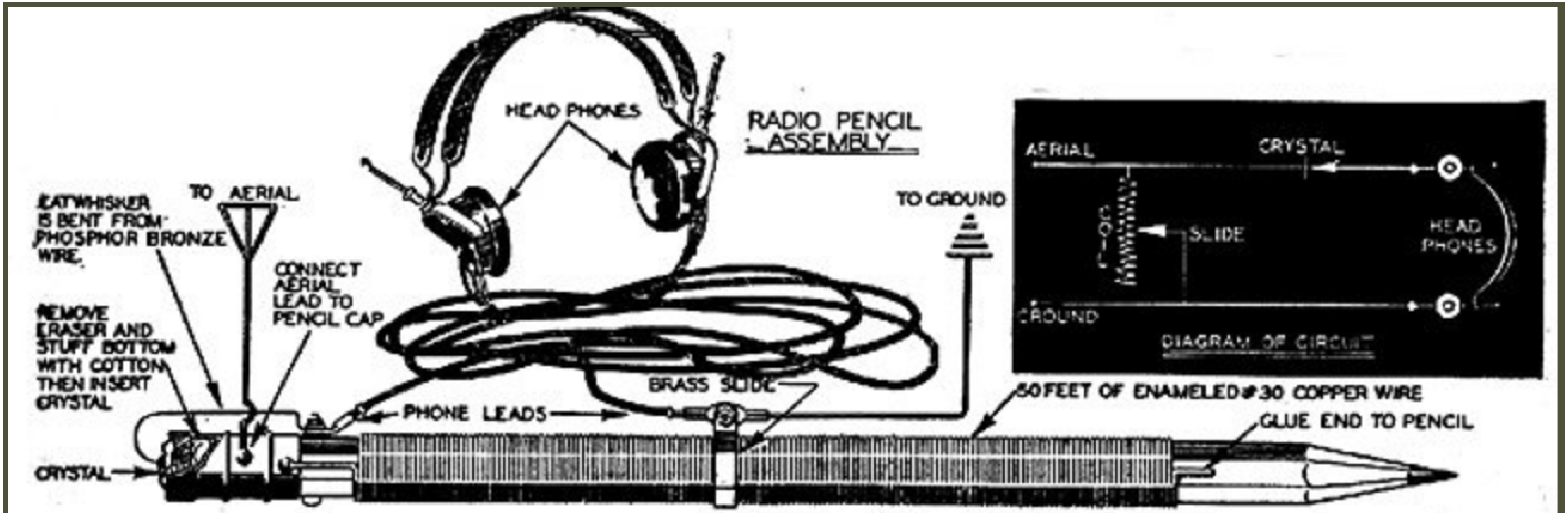
HEAD PHONES
RADIO PENCIL ASSEMBLY
CATWHISKER IS BENT FROM PHOSPHOR BRONZE WIRE
TO AERIAL
CONNECT AERIAL LEAD TO PENCIL CAP
REMOVE ERASER AND STUFF BOTTOM WITH COTTON THEN INSERT CRYSTAL
CRYSTAL
TO GROUND
BRASS SLIDE
PHONE LEADS
50 FEET OF ENAMELED #30 COPPER WIRE
GLUE END TO PENCIL
AERIAL
CRYSTAL
SLIDE
HEAD PHONES
GROUND
DIAGRAM OF CIRCUIT

CRYSTAL SET designed by T. A. Blanchard is built on ⅜″ diameter novelty pencil. Solder the top lead of the coil to the metal cap, and simply glue bottom end to the pencil. Remove eraser, push wad of cotton into bottom of cap, and insert a mounted crystal. Bend cat whisker from a length of phosphor bronze wire, and hold in place with bolt going through wood part of pencil. Scrape the enamel off a narrow strip of the coil, so that slide makes contact.

To use, connect aerial, ground, and headphones, move slider and adjust crystal until stations are heard.

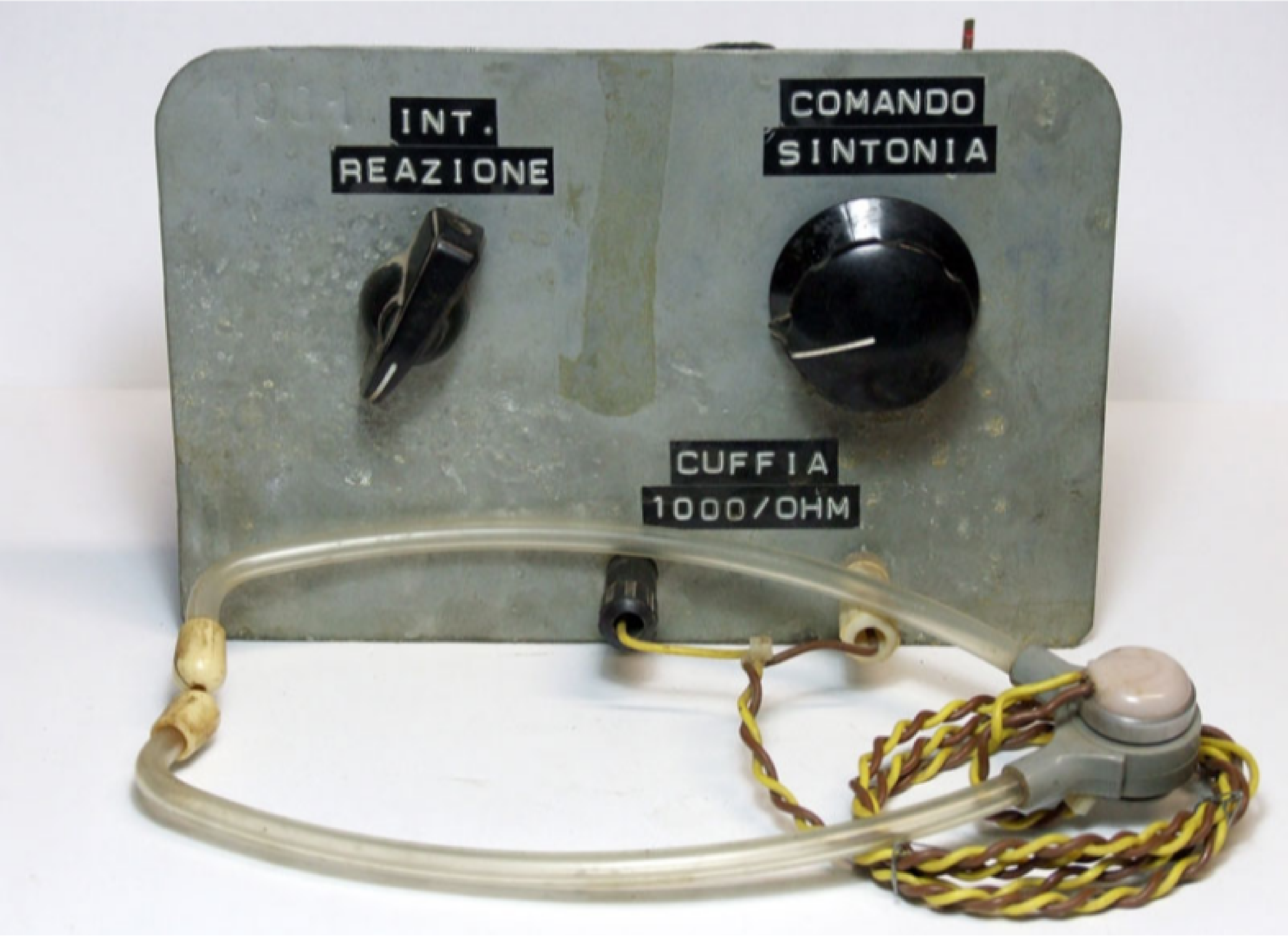
INT.
REAZIONE
COMANDO
SINTONIA
CUFFIA
1000/OHM

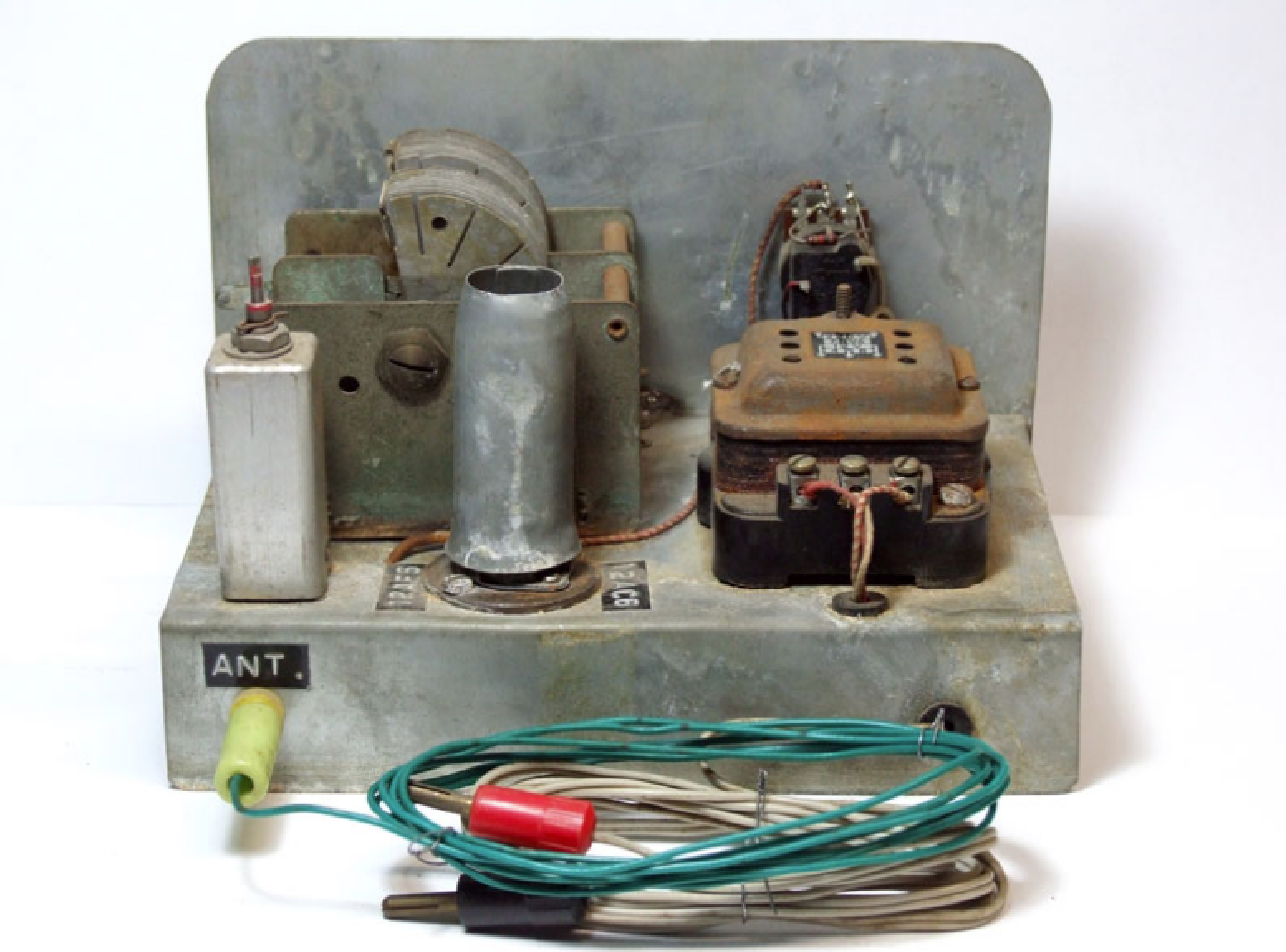
ANT.

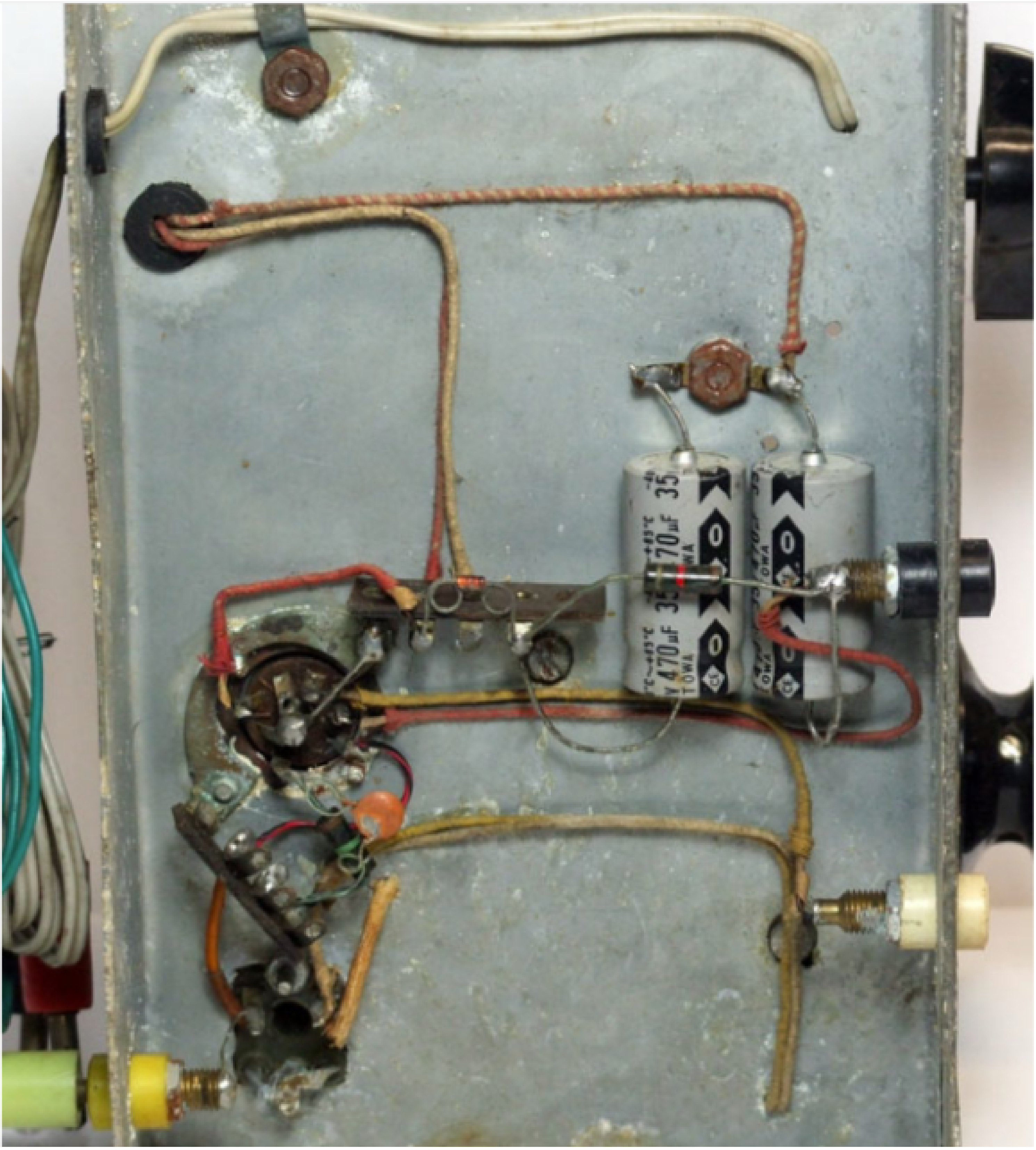
470μF
35
TOWA

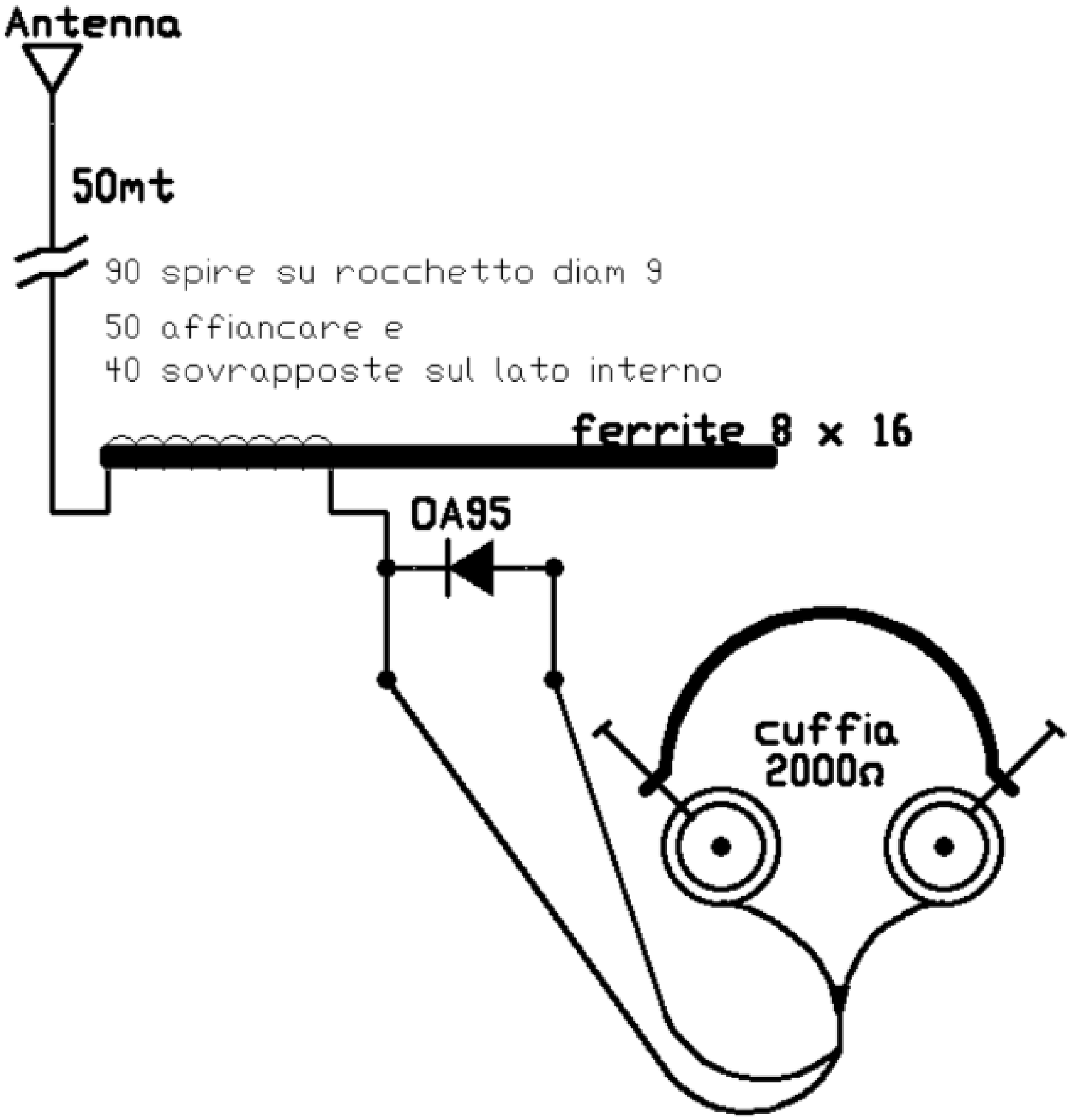
Antenna
50mt
90 spire su rocchetto diam 9
50 affiancare e
40 sovrapposte sul lato interno
ferrite 8 x 16
OA95
cuffia
2000Ω

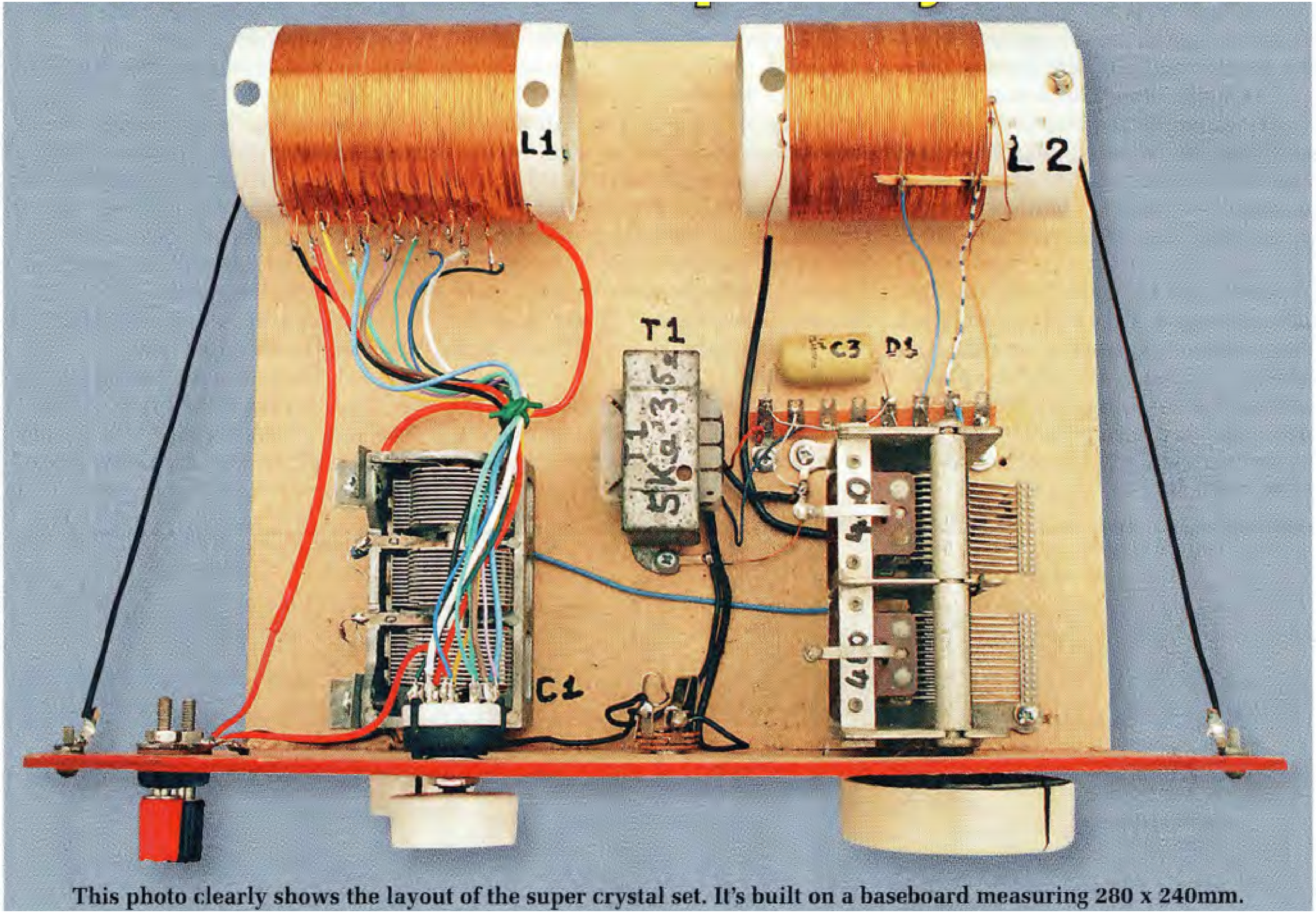
L1
L2
T1
T1 5kΩ 3.5Ω
C3
D1
400
400
C1

This photo clearly shows the layout of the super crystal set. It's built on a baseboard measuring 280 x 240mm.

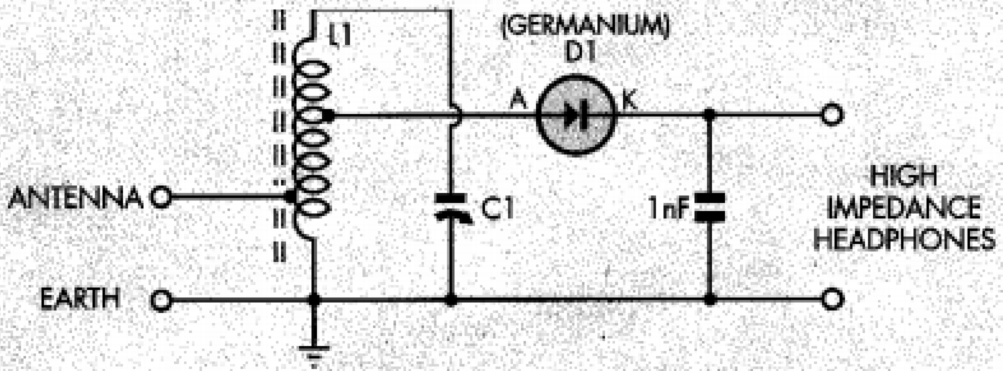
L1
(GERMANIUM)
D1
A
K
ANTENNA
C1
1nF
HIGH
IMPEDANCE
HEADPHONES
EARTH

Fig.1: the circuit for a basic crystal set. Coil L1 can be air-cored (see text for specifcations) or can be wound on a 100 x 20 x 5mm flat ferrite rod using 70 turns of 22 B & S enamelled wire tapped at 10, 20, 30 & 40 turns.

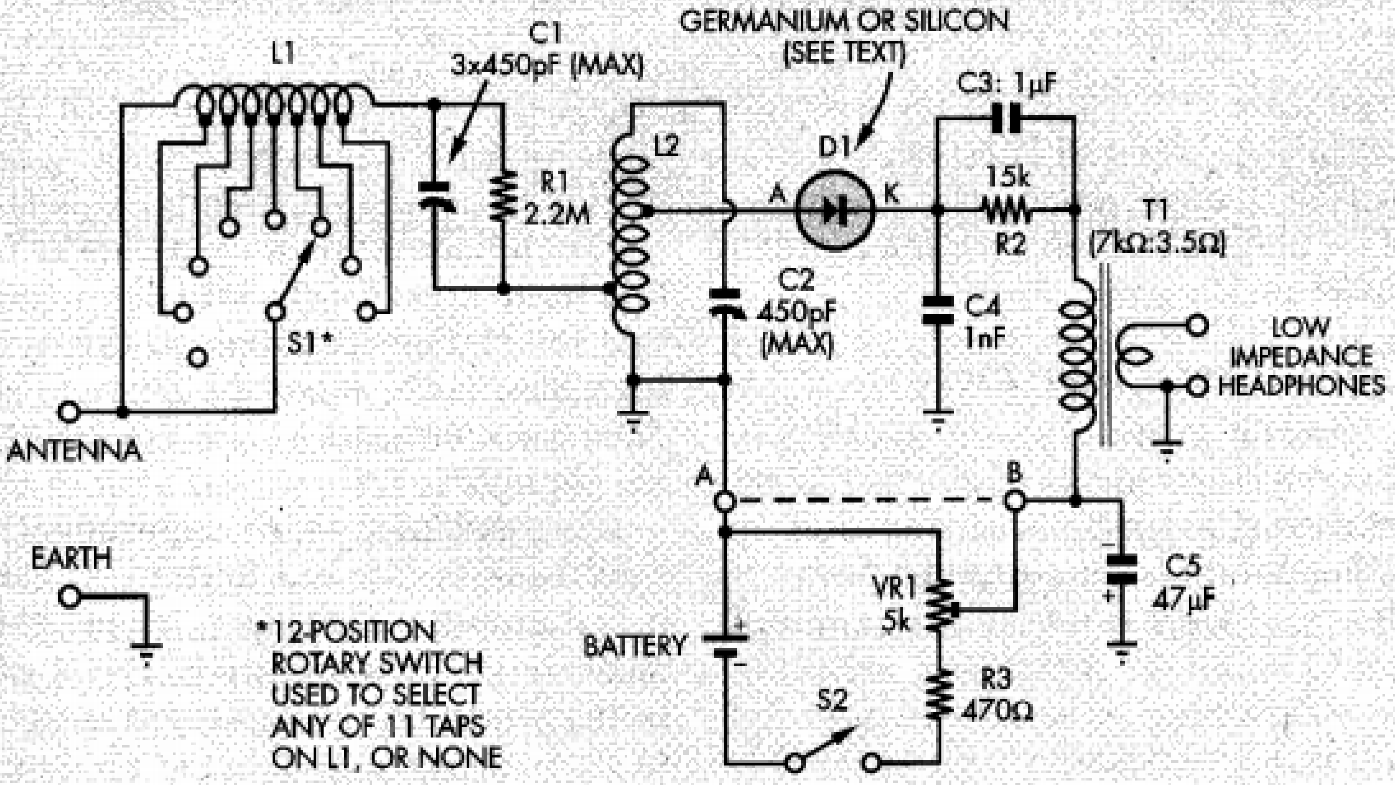
L1
C1
3x450pF (MAX)
GERMANIUM OR SILICON
(SEE TEXT)
C3: 1µF
D1
R1
2.2M
L2
A
K
15k
R2
T1
(7kΩ:3.5Ω)
C2
450pF
(MAX)
C4
1nF
S1*
LOW
IMPEDANCE
HEADPHONES
ANTENNA
A
B
EARTH
VR1
5k
C5
47µF
*12-POSITION
ROTARY SWITCH
USED TO SELECT
ANY OF 11 TAPS
ON L1, OR NONE
BATTERY
S2
R3
470Ω

Fig.2: the super crystal set circuit uses L1 & C1 to tune the antenna circuit, while L2 & C2 tune the received frequency. Transformer T1 is included for driving low-impedance headphones. Also included is an optional detector bias circuit consisting of VR1, switch S2, the 47µF capacitor, the 470Ω resistor & the battery – see text.

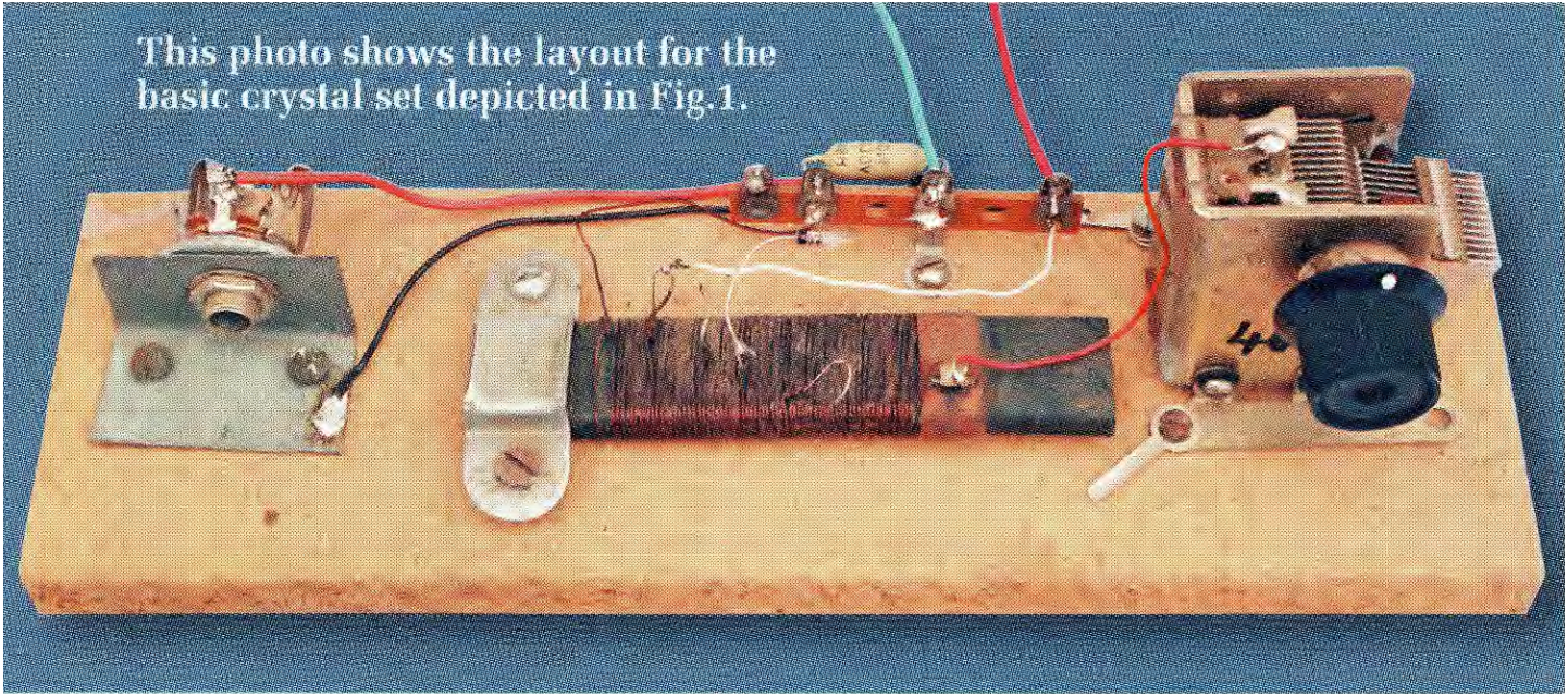

This photo shows the layout for the basic crystal set depicted in Fig.1.

we only want to listen to one station at a time!

Detector bias

Some crystal set designs use a battery and a potentiometer to bias the detector to the point where it just conducts. The reason for this is quite simple.

Diodes all need a certain amount of voltage applied to them before they conduct. As a result, if we apply a DC voltage to a signal diode so that it just conducts, the diode will be in its most sensitive state and will thus give good performance in a crystal set.

This "bias" voltage varies according to the diode used. For example, silicon diodes such as the 1N4148 and the 1N914 require around 0.6V of positive bias to operate, while a germanium diode only requires about 0.2V of bias. On the other hand, the OA47 diode I've used works quite well with no forward bias, which has kept my crystal set just that bit simpler. Your own experience may be different, however so be prepared to experiment.

By the way, transistor radio detectors often use forward bias to improve sensitivity. (Editor's note: forward bias on a diode detector also reduces harmonic distortion).

Detector load

The load that the detector works into is usually a pair of headphones which may have between 2kΩ and 4kΩ total resistance. However, the diode will be more efficient if it works into a higher load resistance and some designs use a resistor of about 15kΩ in series with the headphones to achieve this. In addition, a capacitor of around 1µF is placed in parallel across the resistor so that the audio is not noticeably attenuated.

Although my 4kΩ headphones are good performers they are uncomfortable, so I compared the performance of other headphones against the 4kΩ pair. As stated above, I ended up using low-impedance stereo headphones fed through a speaker transformer. (Editor's note: we also recommend the 32Ω earphones supplied with iPods and MP3 players).

Putting it all together

The set described here is not only easy to build and operate but outperforms many other so-called high-performance sets.

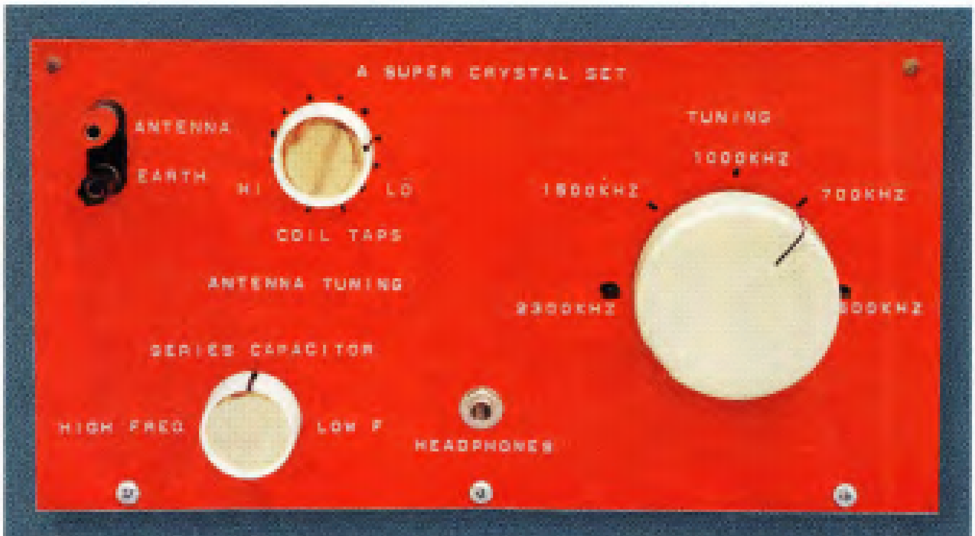

This view shows the front-panel layout for the super crystal set. A good aerial and earth are necessary to achieve good performance – see text.

Fig.2 shows the circuit. L1 & C1 tune the antenna circuit, while L2 & C2 tune the received frequency. Transformer T1 is included for driving low-impedance headphones, while the optional detector bias circuit consists of VR1, switch S2, a 47µF capacitor, a 470Ω resistor and the battery.

The set was built on a 12mm-thick particle board measuring 280 x 240mm. This is fitted on the underside with four self-adhesive felt pads (available from hardware stores) to keep it clear of the bench.

The front panel is made of thin plywood measuring 300 wide x 160mm high. This was given several coats of red paint from a spray can and the labelling on the front panel completed using red Dymo® embossing tape (to match the paint job).

Tuning capacitors

Tuning capacitors can be scrounged from old valve radios that are not worth restoring. For C2, I used one section of a 2-gang full-size 460pF tuning capacitor. This has a 3/8-inch shaft which suits few knobs (and certainly none of my collection). It did, however, have a dial drum which was left on. I fitted a cut-down top from a tin of cream spray paint over the top of the dial drum (it fitted perfectly), which makes it look better and acts as a "handspan" dial.

I was more fortunate with C1 which is a 3-gang 450pF per section tuning capacitor, as this had a ¼-inch shaft. All three sections of the capacitor are used in parallel. The only disadvantage with this tuning gang is that it has a reduction drive, which means that more than a full turn of the knob is required to go from minimum to maximum capacitance.

C3 is a 1µF polyester or greencap type, while C4 is a 1nF unit. The voltage ratings of these capacitors can be quite low.

Resistors & switches

Resistor R1 is used as a "static leak". Its purpose is to prevent a high-voltage static charge from building up across C1 (eg, during storms) which could lead to flash-over. Note also that if a particularly big antenna is in use, it would be advisable to disconnect and earth it when the crystal set is not being used.

Resistor R2 (15kΩ) is the DC load for the detector. Both R1 and R2 can be 0.5W or smaller. If detector biasing is used, R3 can be a 0.5W unit, while VR1 can be a standard 1kΩ linear trimpot or a normal potentiometer. VR1 will not normally require readjustment once set.

S1 is a 12-position switch which selects one of 11 tappings on L1. Only 11 positions are used; the 12th is left with no connection so that the whole of L1 is in circuit. S2 (if fitted) turns the detector biasing on or off as required.

T1 is a standard speaker transformer with a 5kΩ or 7kΩ primary impedance and a 3.5Ω secondary winding. This drives two 8-ohm headphone earpieces in series, so that the reflected impedance to the primary from a 16-ohm secondary load is at least 20kΩ. Within reason, increasing the reflected impedance will assist in maximising the audio output of the receiver.

Winding the coils

Now we come to the all-important coils (ie, L1 and L2).

First, L1 is wound on a 100mm length of 60mm-diameter white PVC pipe. In my case, I wound on 102 turns of 0.63mm (22 B&S) enamelled wire over a length of 70mm. In hindsight, though, around 110 turns would have allowed somewhat more adjustment range to tune the antenna system.

When winding L1, it should be tapped every seven turns and there should be 12 tapping points in all, starting right from the antenna end of the coil. If your antenna is significantly different from mine, then the number of turns on this coil to achieve optimum tuning will vary accordingly.

L2 is also wound on a 100mm x 60mm-diameter white PVC water pipe and consists of 80 turns of 0.63mm (22 B&S) enamelled wire. This winding is spread over 60mm of the former's length and the coil is tapped at 3, 6 & 35 turns from the "earthy" end.

In my case, I found that using turn three as the tap gave good results. You will need to experiment here – you may need even fewer turns to the first tap if your antenna is larger than mine. Note that the correct position may vary from the high-frequency end of the band to the low-frequency end.

Because the detector load is relatively high, it's possible to connect the detector to a tap quite high up the coil. I found that 35 turns was optimum for best performance in my receiver.

As can be seen from the photos, different tapping methods are used for the two coils. For example, L1 has the wire raised away from the former and twisted to make each tapping. By contrast, on L2, a match is slid under the wire at the first tapping point and is then slid along the winding to go under each successive tapping point as the coil is wound. This is the neater of the two methods but is difficult to do effectively if the winding is long and has lots of tappings.

With either method, it is necessary to thoroughly clean the enamel off the wire at the tapping points so that a good soldered joint can be made. This can be done by scraping away the enamel using a sharp utility knife.

Receiver layout

The parts layout on the baseboard is not critical, although the coils should be mounted at the back of the receiver for ease of access. The accompanying photos show the author's unit.

In my case, tuning coil L2 was mounted at the right rear of the baseboard, with L1 in the opposite rear corner. It's important that L1 is kept several centimetres away from L2, to minimise unwanted coupling between them.

Diode biasing

As mentioned earlier, some diodes (particularly silicon signal diodes) require about 0.6V of positive bias to maximise sensitivity. Normally, without diode biasing, points A and B are connected together on the circuit. Conversely, if forward biasing is used, points A and B are separated and the small circuit consisting of B1 (a 1:5V dry cell), toggle switch S2, a 470Ω resistor, a 1kΩ trimpot (VR1) and a 47μF electrolytic capacitor added between these two points.

This circuit is easily adjusted. Switch S2 is closed and trimpot VR1 is adjusted for best volume – simple.

Summary

Those who have never built a crystal set radio before will find this little set worth the effort. It works well, isn't difficult to tune and provides good headphone volume on all local stations.

For best reception, use a high, long antenna that's clear of buildings and trees. A good earth is also necessary and a pipe driven about one metre into the ground and kept damp should suffice.

Finally, if you wish to read about other people's designs, the following list makes a good starting point:

(1) Look on the Internet. Typing "crystal set society" into Google will give you many interesting sites (and lots of designs) that you can explore.

(2) Look in SILICON CHIP for October 1988 (crystal set), March 1990 (wave traps), October 1994 (Hellier Award crystal sets), March 2003 (antennas).

(3) Look in "Electronic Australia" for June 1988 (crystal set), July 1994 (crystal sets), November 1998 (coils) and July 2000 (crystal set).

(4) An excellent Australian book on crystal sets is "Crystal Sets 'N' Such" by Bob Young (7 Hayes Rd, Swanpool, Vic. 3673). He has a few available for $19:95 posted in Australia. SC

Photo Gallery: Targan Airmaster

MANUFACTURED BY TARGAN ELECTRIC PTY LTD in 1933, the Airmaster was a 3-valve TRF receiver in an upright wooden cabinet. It used the following valve types: 57 detector; 59 audio output; and 80 rectifier. Photo: Historical Radio Society of Australia, Inc.

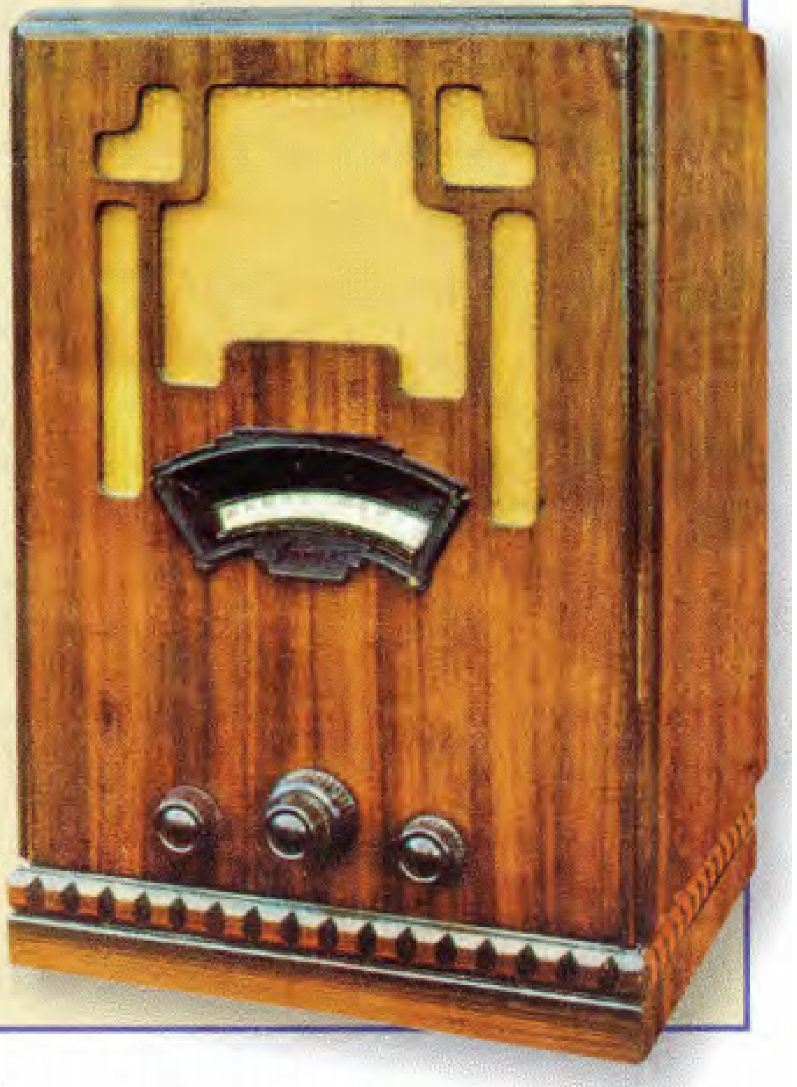

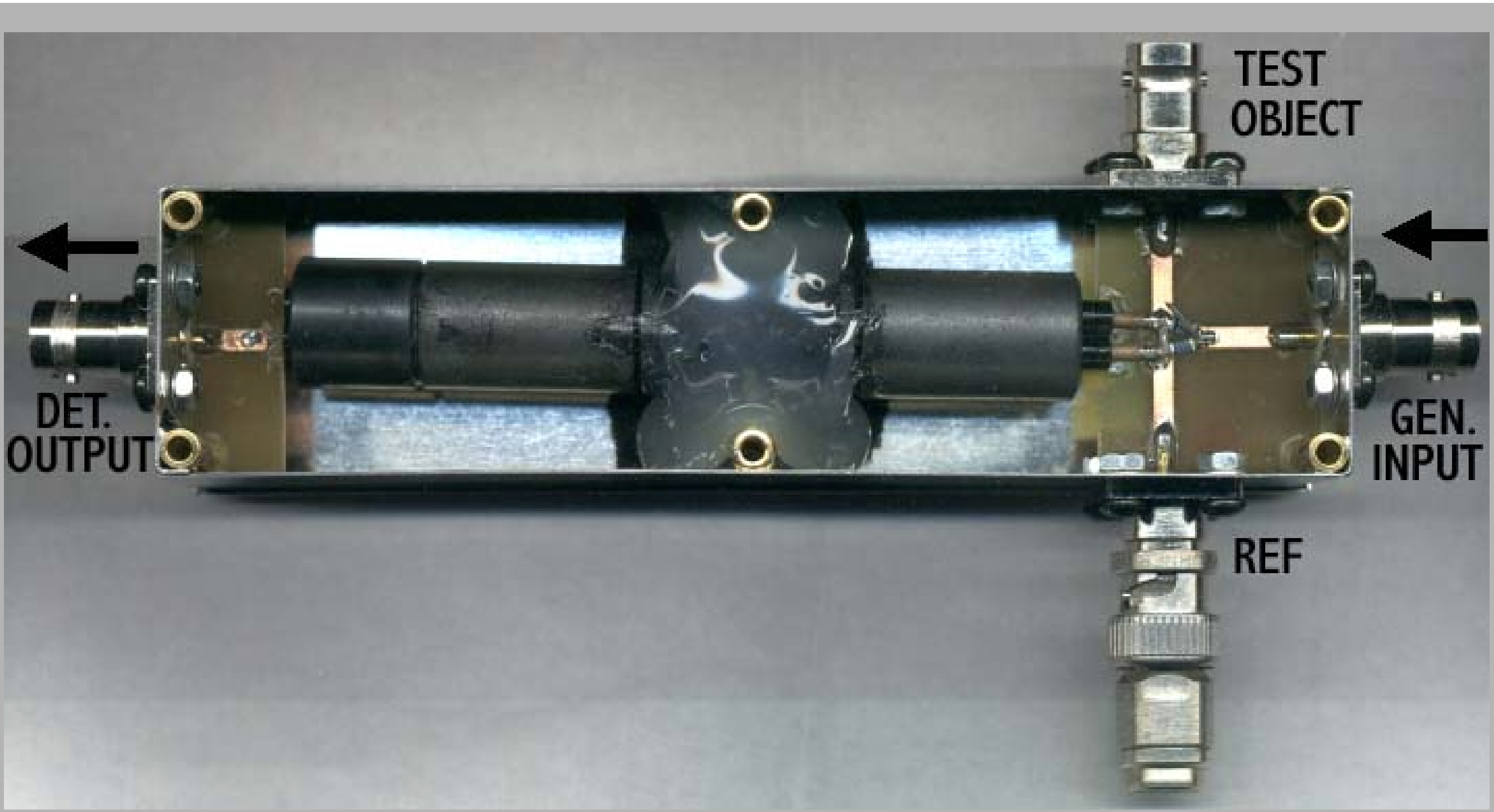
TEST
OBJECT
DET.
OUTPUT
GEN.
INPUT
REF

This SWR bridge is not my own invention, OZ3SW and OZ9MO has improved the design after DJ7VY construction in ukw-berichte 1/83
This is a construction set made for the members of Frederikssunds EDR Club, all I did was a bit of gluing and soldering
Thanks for all the hardware help, parts help, calibrating and measuring ! you know who I mean !!
It is possible to measure return loss up to 1000 Mhz, the return loss range is 0-50 to 0-40dB depending on frequency.

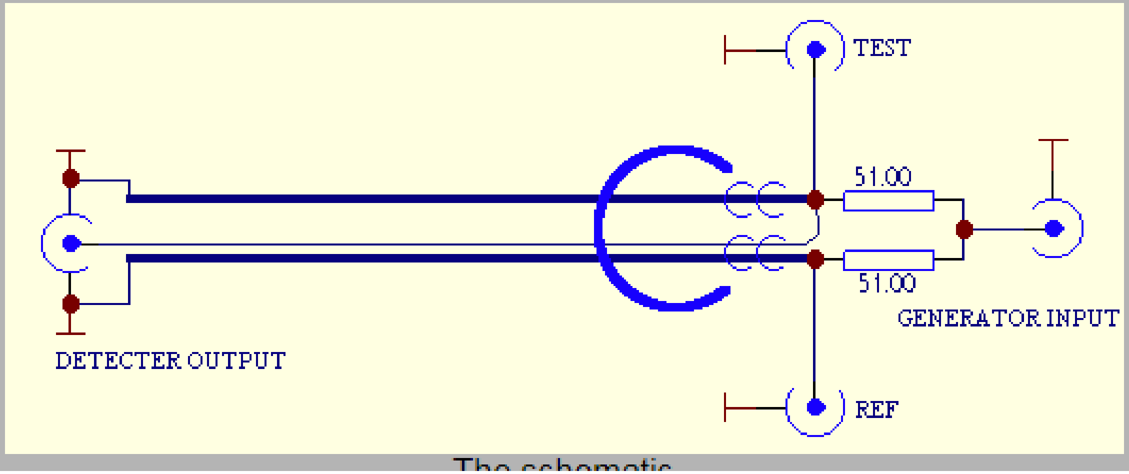
TEST
51.00
51.00
GENERATOR INPUT
DETECTER OUTPUT
REF

The schematic

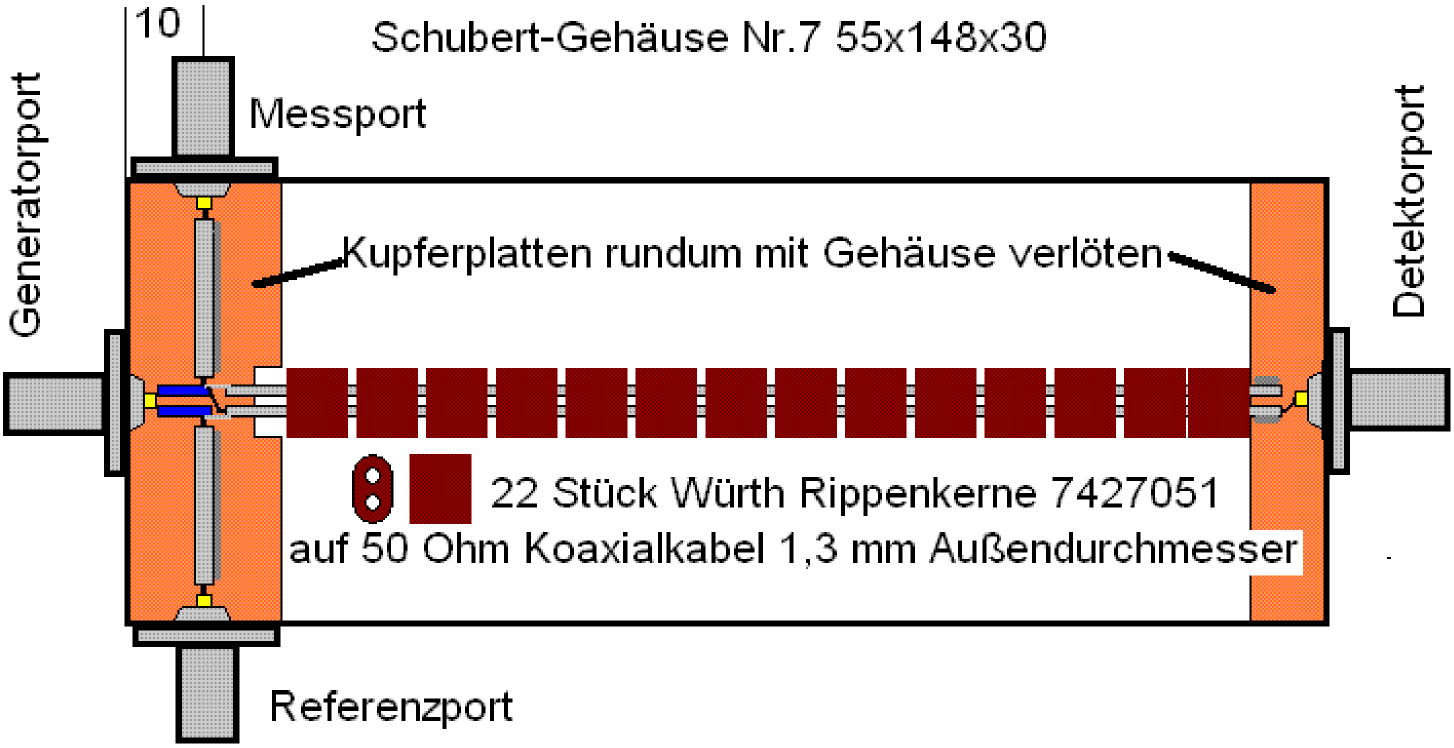
10
Schubert-Gehäuse Nr.7 55x148x30
Messport
Generatorport
Detektorport
Kupferplatten rundum mit Gehäuse verlöten
22 Stück Würth Rippenkerne 7427051
auf 50 Ohm Koaxialkabel 1,3 mm Außendurchmesser
Referenzport

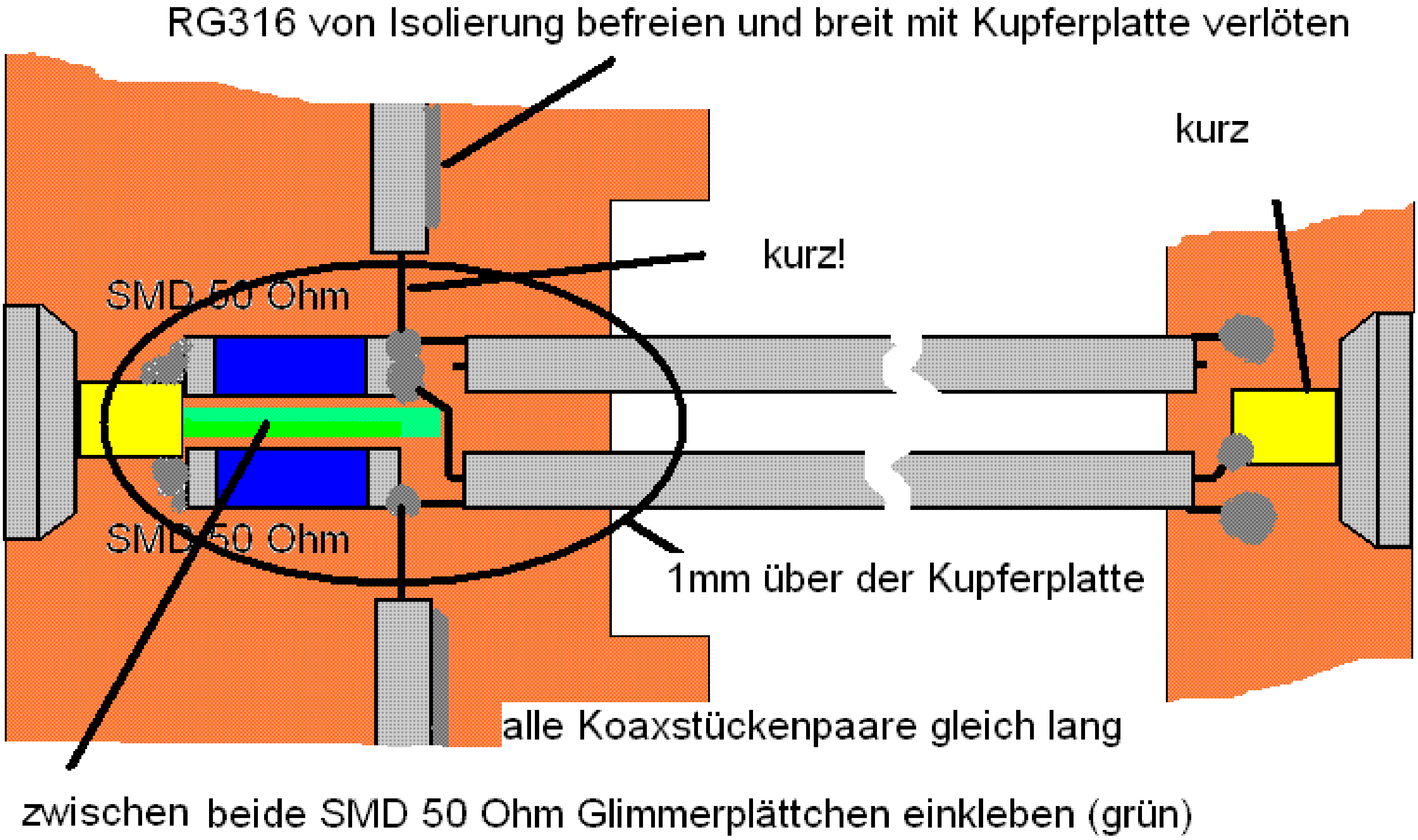
RG316 von Isolierung befreien und breit mit Kupferplatte verlöten
kurz
kurz!
SMD 50 Ohm
SMD 50 Ohm
1mm über der Kupferplatte
alle Koaxstückenpaare gleich lang
zwischen beide SMD 50 Ohm Glimmerplättchen einkleben (grün)

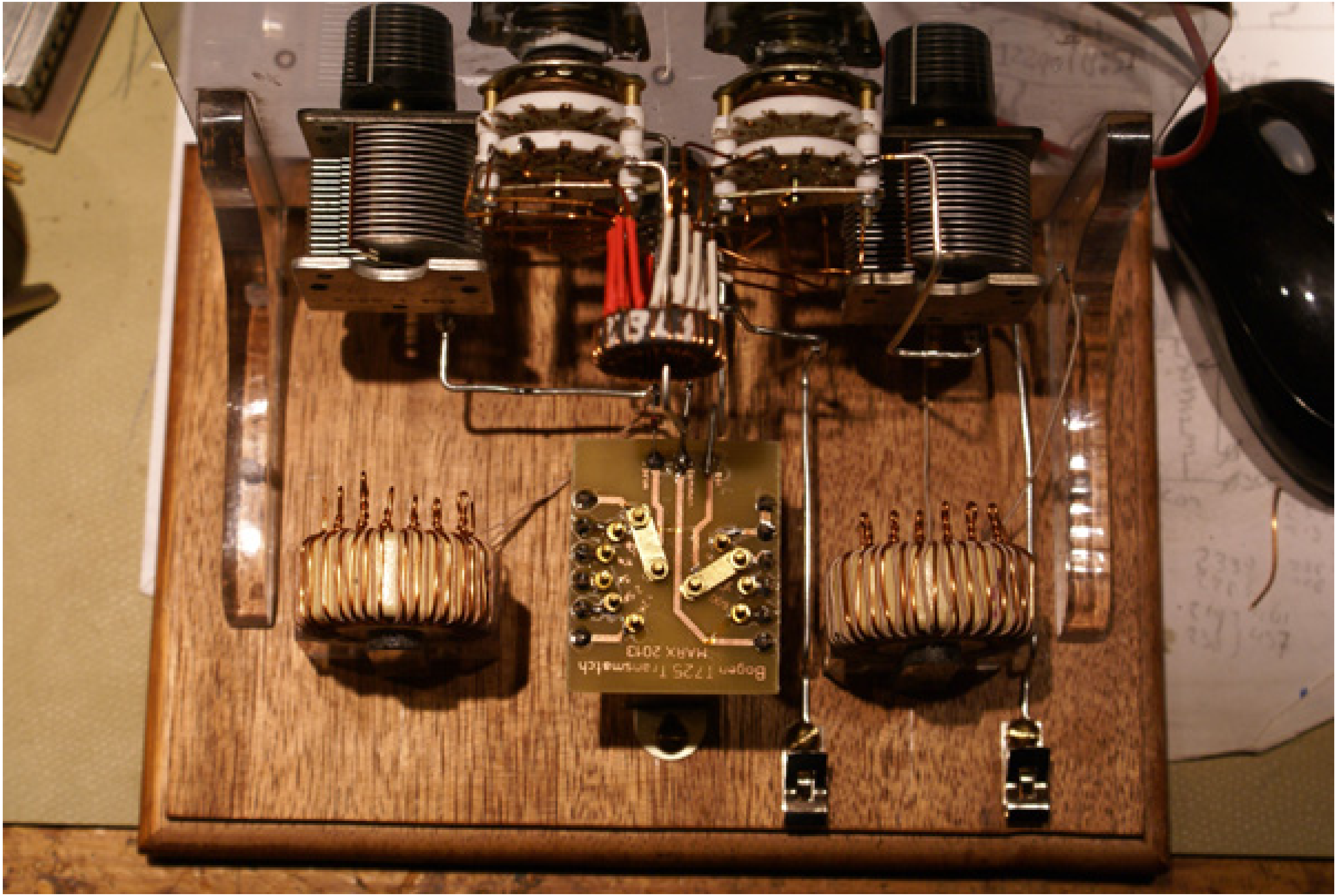
Bogen T725 Transmatch
MARX 2013

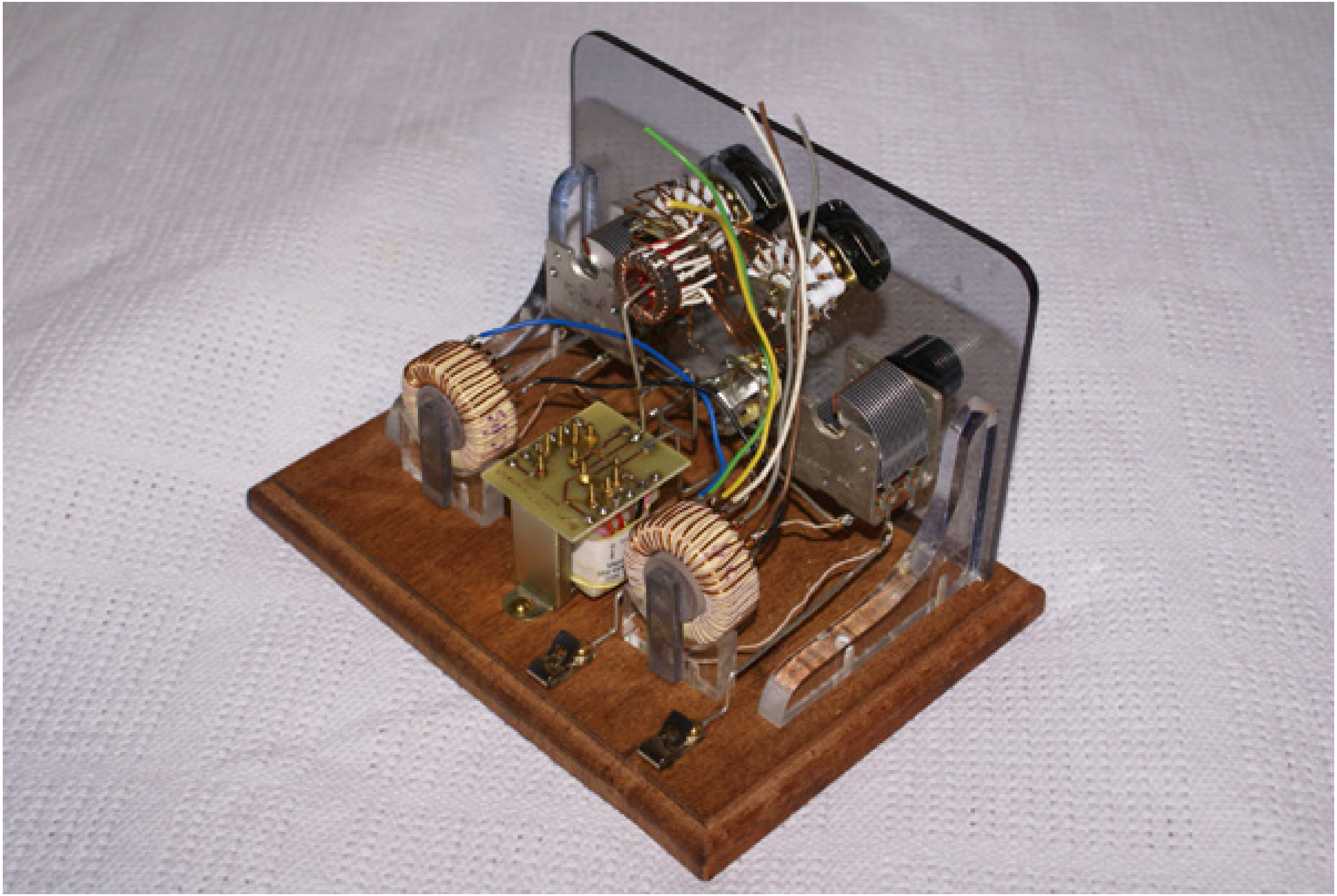

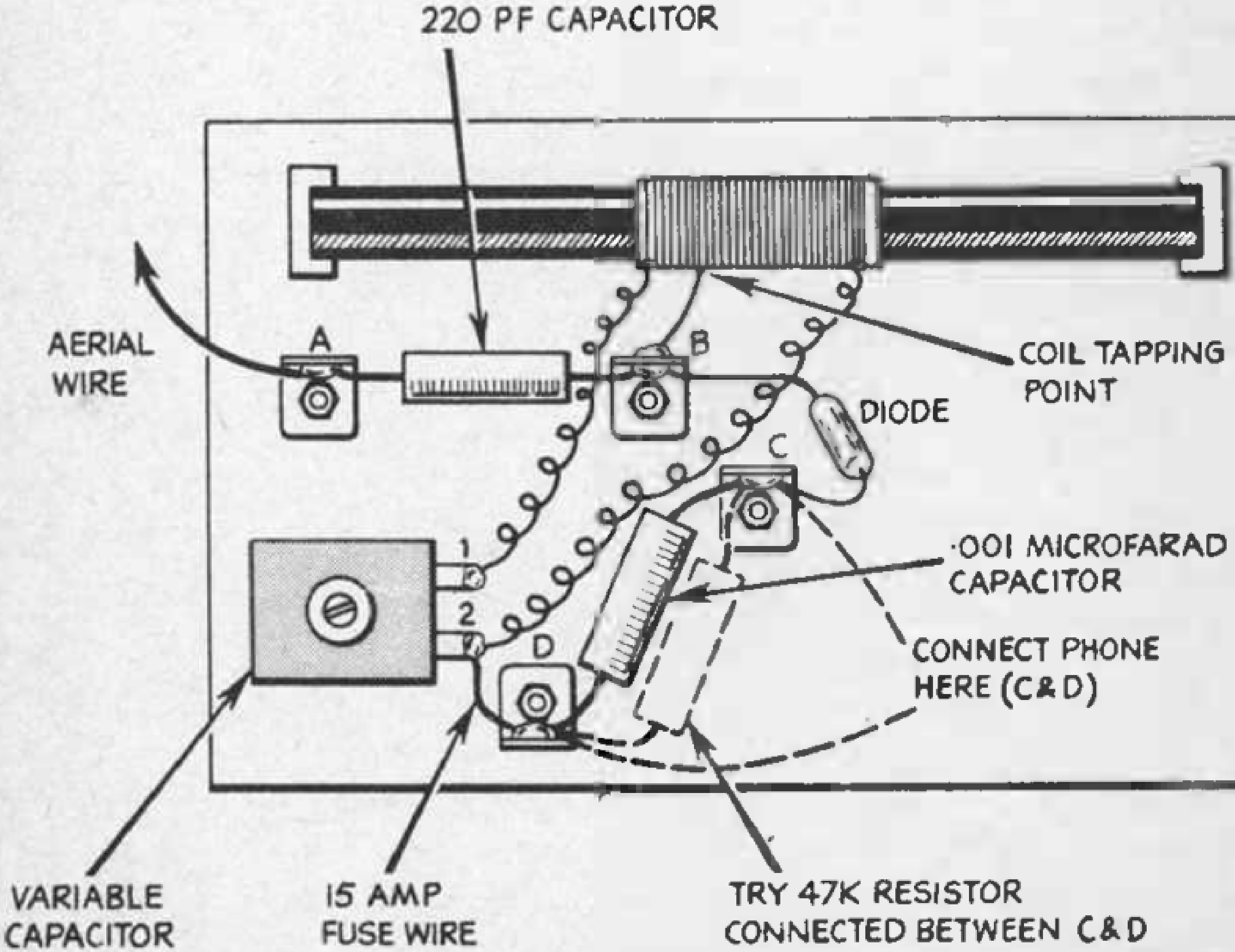
220 PF CAPACITOR
AERIAL WIRE
A
B
COIL TAPPING POINT
DIODE
C
·001 MICROFARAD CAPACITOR
1
2
D
CONNECT PHONE HERE (C&D)
VARIABLE CAPACITOR
15 AMP FUSE WIRE
TRY 47K RESISTOR CONNECTED BETWEEN C&D

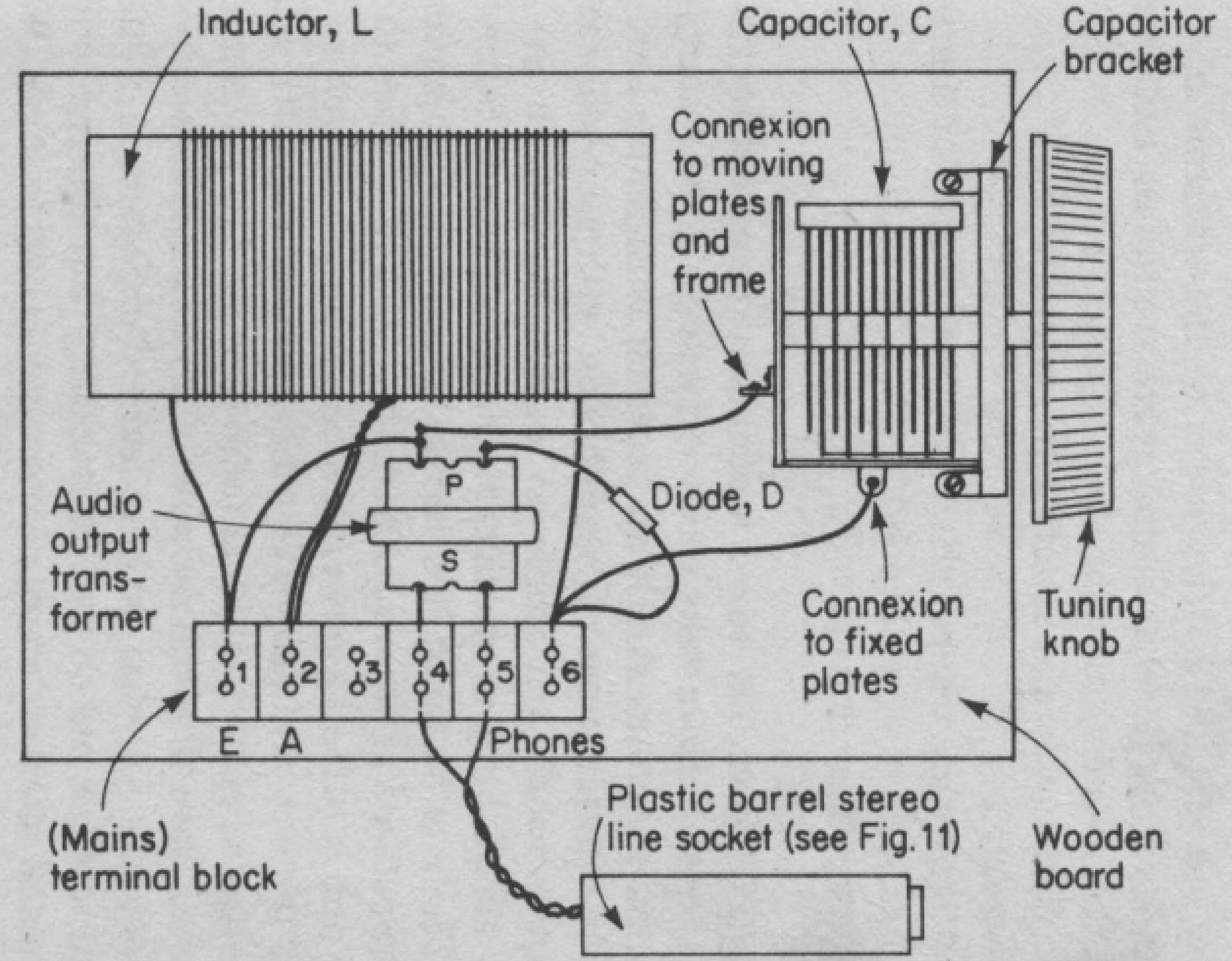
Inductor, L
Capacitor, C
Capacitor bracket
Connexion to moving plates and frame
P
Audio output trans- former
Diode, D
S
Connexion to fixed plates
Tuning knob
1
2
3
4
5
6
E
A
Phones
(Mains) terminal block
Plastic barrel stereo line socket (see Fig.11)
Wooden board

Fig. 14 Suggested component arrangement for elementary crystal set

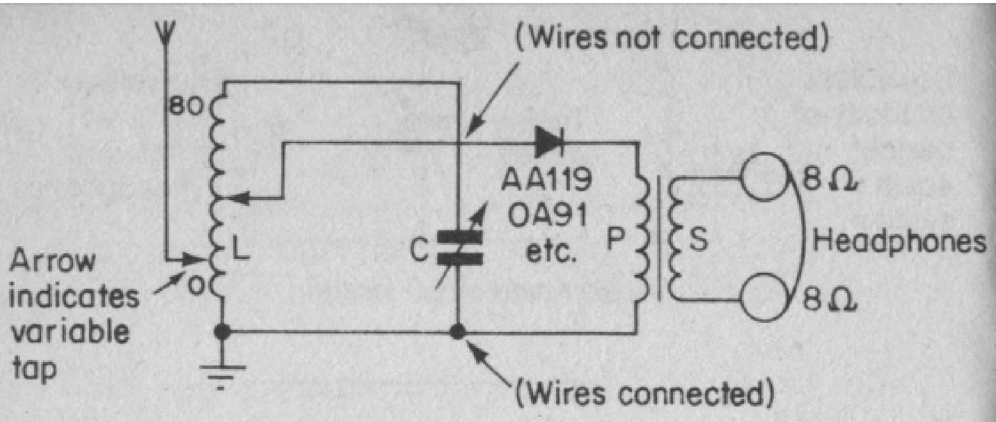
(Wires not connected)
80
AA119
OA91
etc.
C
P
S
8 Ω
Headphones
8 Ω
L
Arrow indicates variable tap
0
(Wires connected)

(i) Circuit

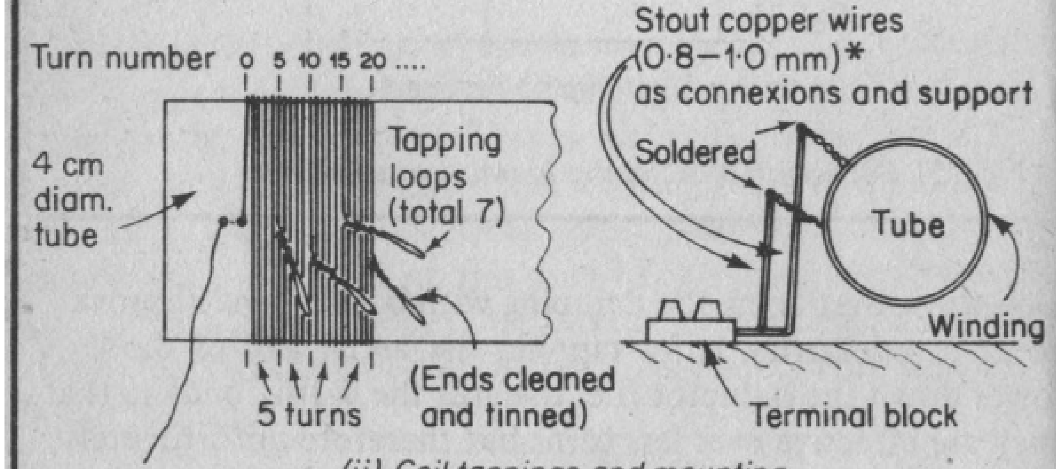
Turn number 0 5 10 15 20 ....
4 cm diam. tube
Tapping loops (total 7)
5 turns
(Ends cleaned and tinned)
Stout copper wires (0·8–1·0 mm)* as connexions and support
Soldered
Tube
Winding
Terminal block

(ii) Coil tappings and mounting

Fig. 16 Circuit with tapped inductor

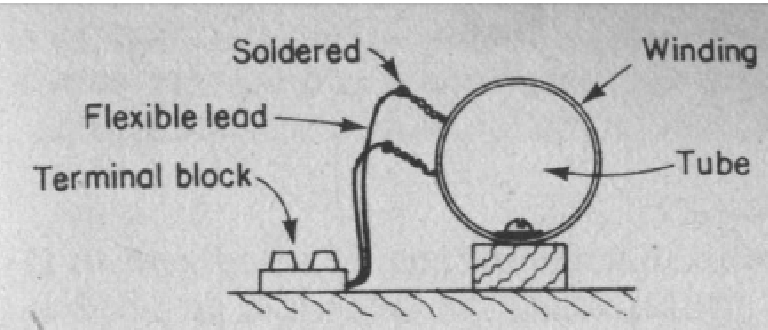
Soldered
Winding
Flexible lead
Tube
Terminal block

(iii) Alternative mounting

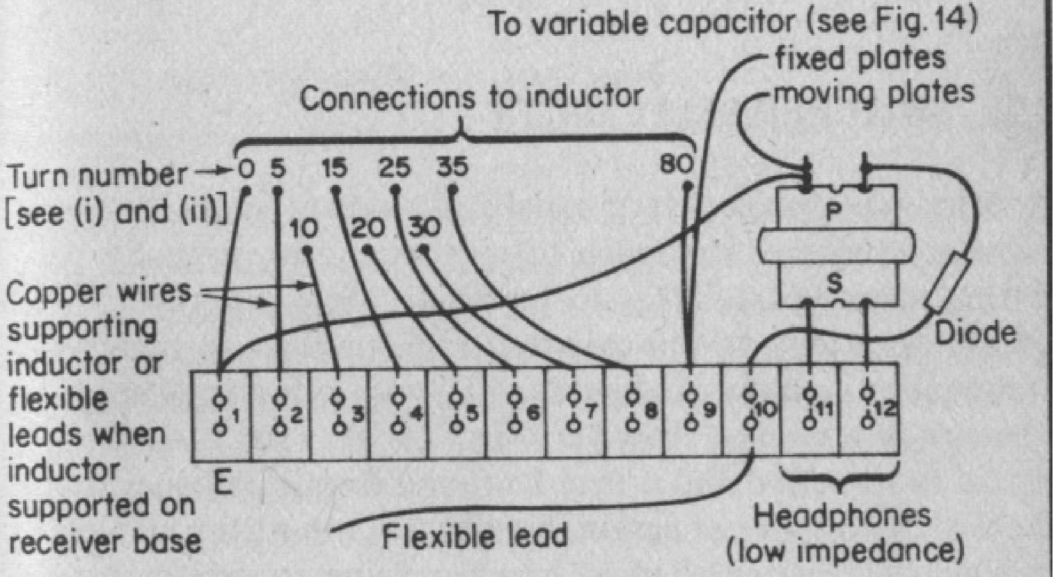
To variable capacitor (see Fig. 14)
fixed plates
moving plates
Connections to inductor
Turn number → 0 5 15 25 35 80
[see (i) and (ii)]
10 20 30
Copper wires supporting inductor or flexible leads when inductor supported on receiver base
P
S
Diode
1 2 3 4 5 6 7 8 9 10 11 12
E
Flexible lead
Headphones (low impedance)

(iv) Suggested component arrangement

---

* 21–19 SWG, 20–18 AWG

required) and a *flying lead* for the diode, these components are shown in Fig.16(iv). This is doing the job properly, there is no reason however (except for unreliability of contact) why both aerial and diode flying leads should not terminate on *crocodile clips* and be clipped onto the chosen taps as required.

Earth is connected to terminal 1 and the headphones to 11 and 12. Both the aerial and the diode lead are tried at terminals 2 – 9, a game of "poke and hope" but one which can be rewarding. Remember that generally but not always, the tappings with the lower numbers reduce loudness but increase selectivity. Aerial to terminal 8, diode to 9 brings us back to Fig.13(i) of course.

5.2 SWITCHED SELECTIVITY

A technique of selection of aerial and diode tappings by means of rotary switches has much to recommend it especially if setting changes are likely to be needed for reception of different stations. In this case two single-pole 8-way switches are required, usually obtainable as 12-way so leaving 4 spare. The drawing symbol shown in Fig.17 speaks for itself with regard to its action and it may be found from catalogues that there are two types of *action*, by which is meant the manner in which the switch changes over from one contact to the adjoining one. *Break before make* switches disconnect one circuit before the next is connected and *make before break* have a change-over period during which both circuits are connected at once for a short time. Either type is suitable for us. A knob to rotate each switch is also required, a small one, say 2 cm diameter or less is ample, if with an indicator line or pointer, so much the better. The modifications to Fig.16 for switching are given in Fig.17.

With such rotary switches a search for the optimum arrange-

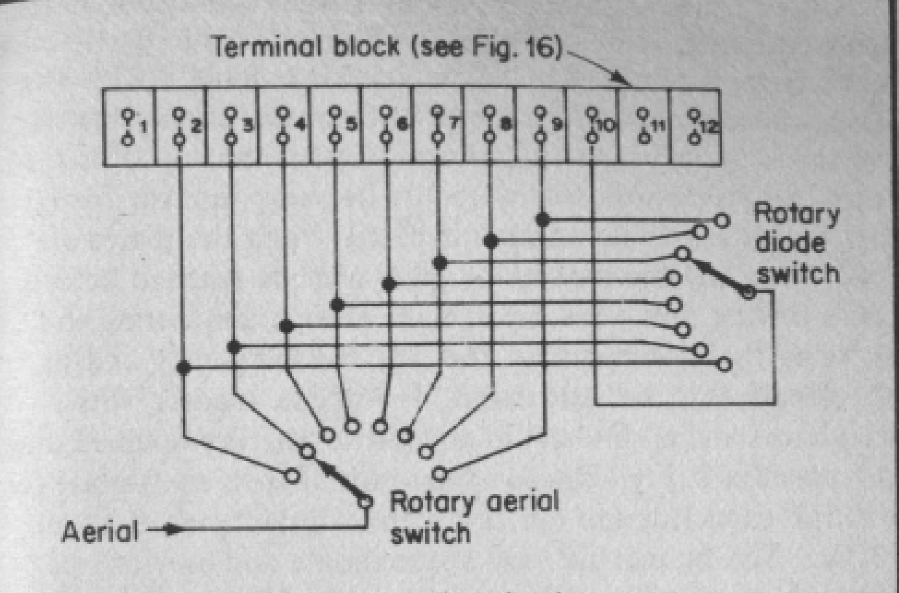

Fig. 17 - Switching aerial and diode leads

variable capacitor, the two switches and the headphone socket, the latter must then be of the chassis type. A pair of terminals for aerial and earth completes the receiver. On the panel the two rotary switches are easily labelled 1 – 8 but we are still uncertain about a tuning dial because we may not have the right variable capacitor for the inductor, this is the subject of the next section.

5.3 GETTING THE RANGE RIGHT

This is where we pause in receiver construction to remove some of the confusion which may still exist with regard to the tuning ingredients, wavelength, frequency, inductance and capacity, each of which has some bearing on the others, to fit together these four pieces of the puzzle, so to speak. For most crystal sets the inductance is fixed according to the variable capacitor available and we tune the receiver over the range by rotating the variable capacitor knob over half a

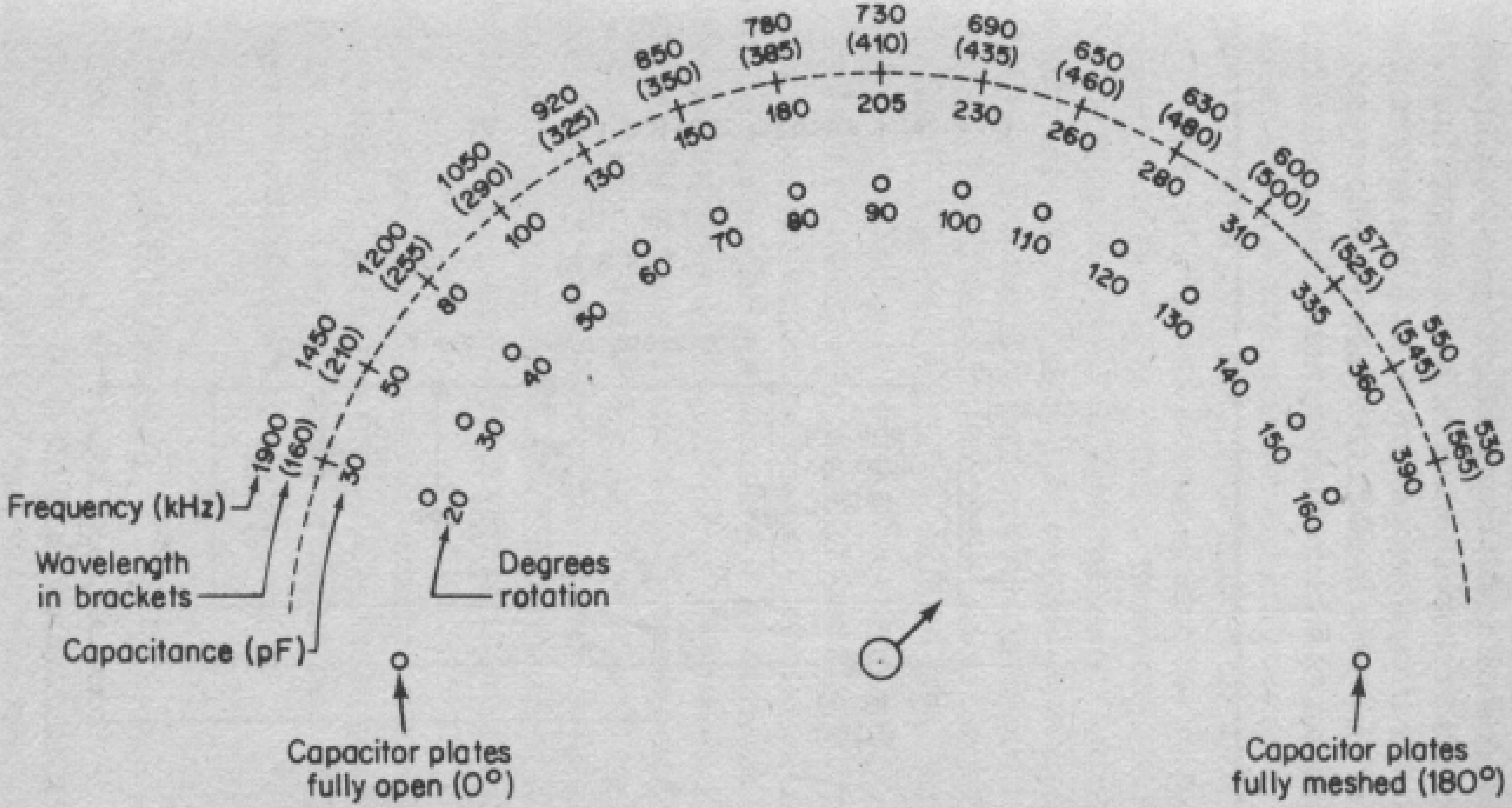
1900
(160)
30
1450
(210)
50
1200
(255)
80
1050
(290)
100
920
(325)
130
850
(350)
150
780
(385)
180
730
(410)
205
690
(435)
230
650
(460)
260
630
(480)
280
600
(500)
310
570
(525)
335
550
(545)
360
530
(565)
390
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Frequency (kHz)
Wavelength in brackets
Capacitance (pF)
Degrees rotation
Capacitor plates fully open (0°)
Capacitor plates fully meshed (180°)

Fig. 18 Approximate capacitance and frequency values for 365 pF variable capacitor and 230 μH coil

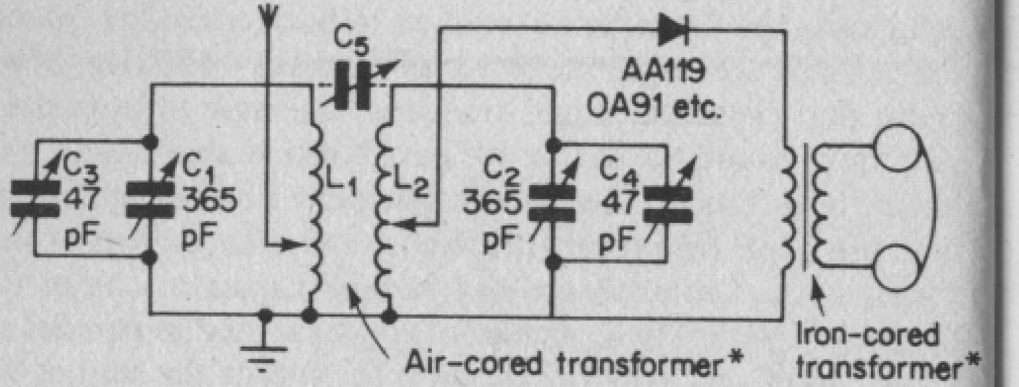
C5
AA119
OA91 etc.
C3
47
pF
C1
365
pF
L1
L2
C2
365
pF
C4
47
pF
Air-cored transformer*
Iron-cored
transformer*

(i) Circuit

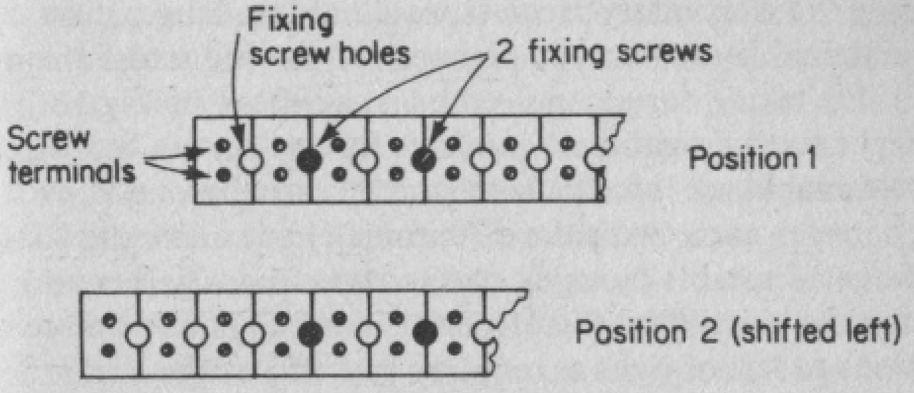
Fixing
screw holes
2 fixing screws
Screw
terminals
Position 1
Position 2 (shifted left)

(iii) Moving inductor L_1

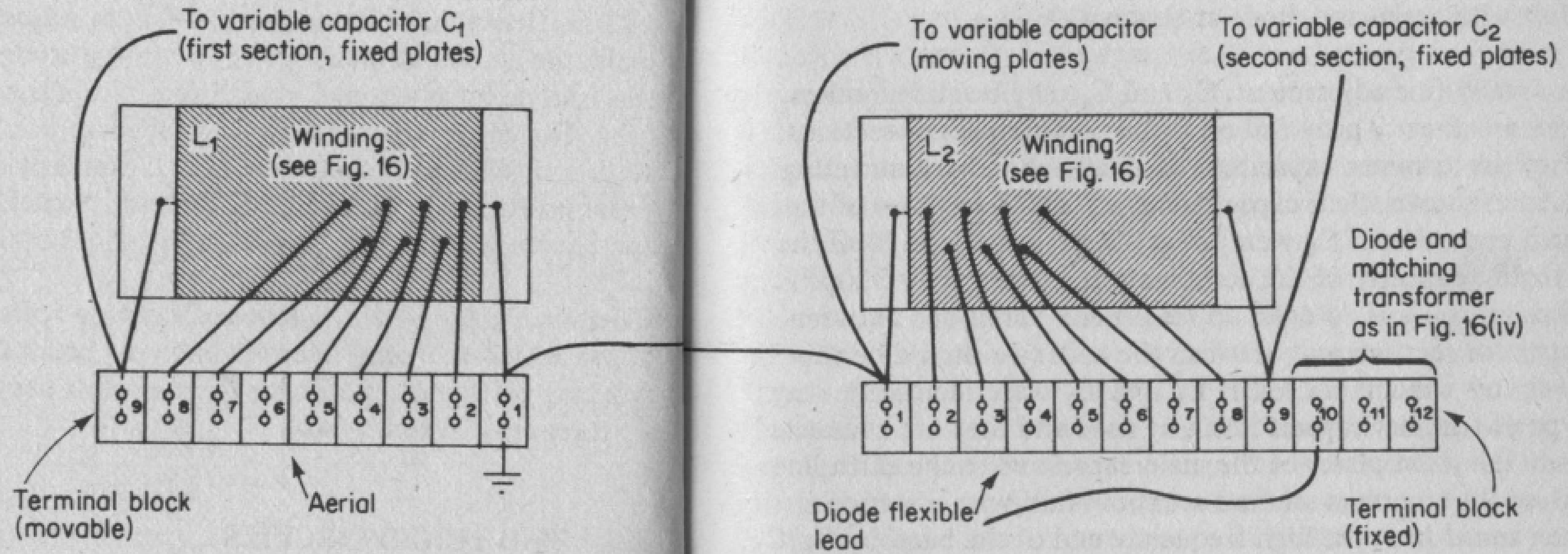
To variable capacitor C1
(first section, fixed plates)
L1
Winding
(see Fig. 16)
9
8
7
6
5
4
3
2
1
Terminal block
(movable)
Aerial
To variable capacitor
(moving plates)
To variable capacitor C2
(second section, fixed plates)
L2
Winding
(see Fig. 16)
Diode and
matching
transformer
as in Fig. 16(iv)
1
2
3
4
5
6
7
8
9
10
11
12
Diode flexible
lead
Terminal block
(fixed)

(ii) Arrangement of inductors

Fig. 21 A double-tuned crystal receiver

* See Section 2.4

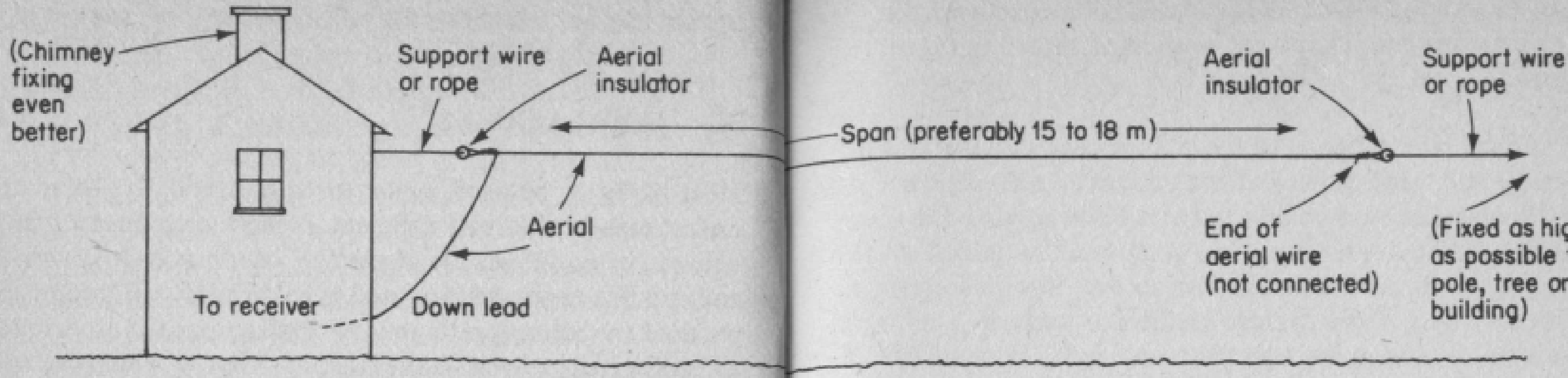
(Chimney fixing even better)
Support wire or rope
Aerial insulator
Aerial
To receiver
Down lead
Span (preferably 15 to 18 m)
Aerial insulator
Support wire or rope
End of aerial wire (not connected)
(Fixed as high as possible to pole, tree or building)

(i) Outdoor Aerial

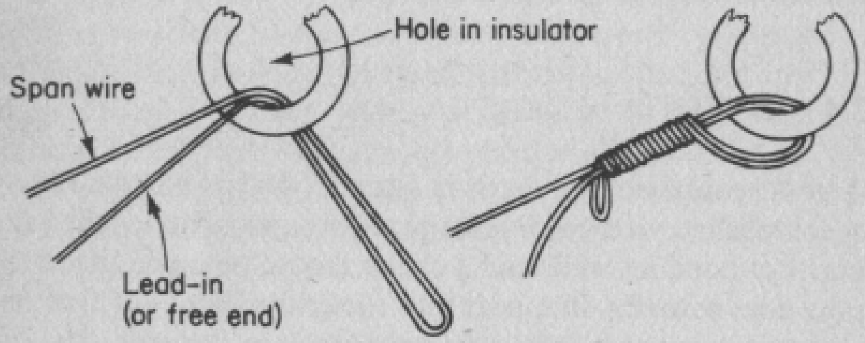
Hole in insulator
Span wire
Lead-in (or free end)

(ii) Attaching aerial to insulator

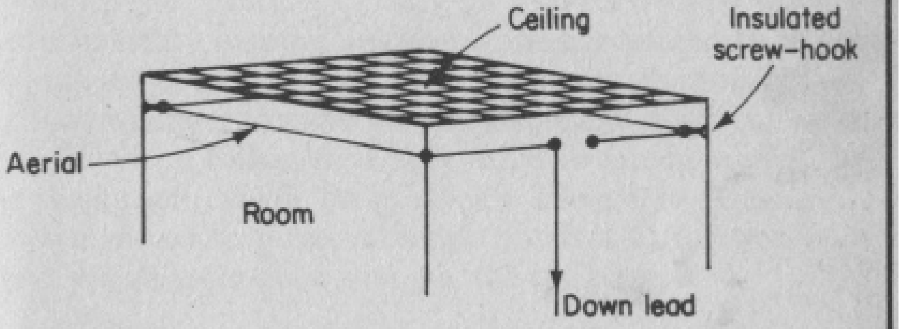
Ceiling
Insulated screw-hook
Aerial
Room
Down lead

(iii) Room aerial

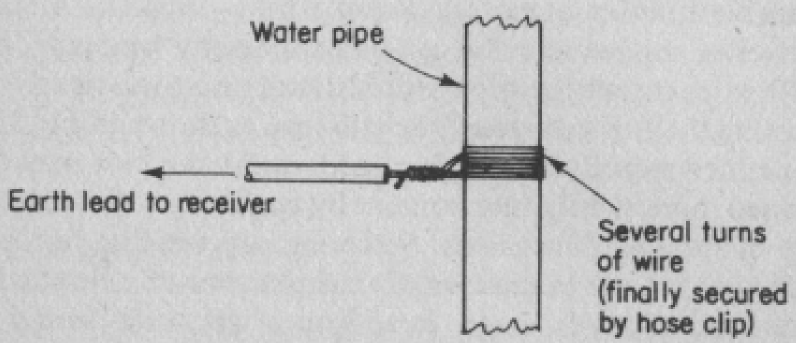
Water pipe
Earth lead to receiver
Several turns of wire (finally secured by hose clip)

(iv) Earth connection to water pipe

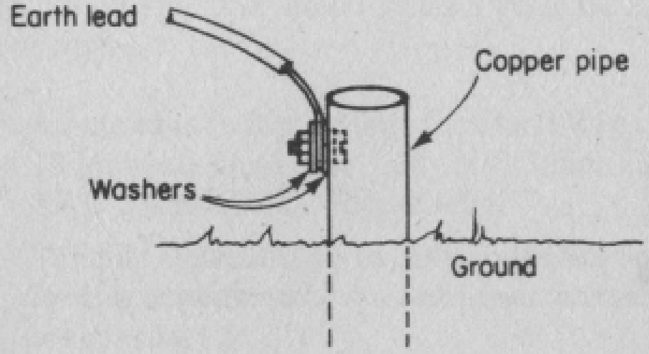
Earth lead
Copper pipe
Washers
Ground

(v) Lead bolted to earth spike

Fig. 12 Aerial/earth systems

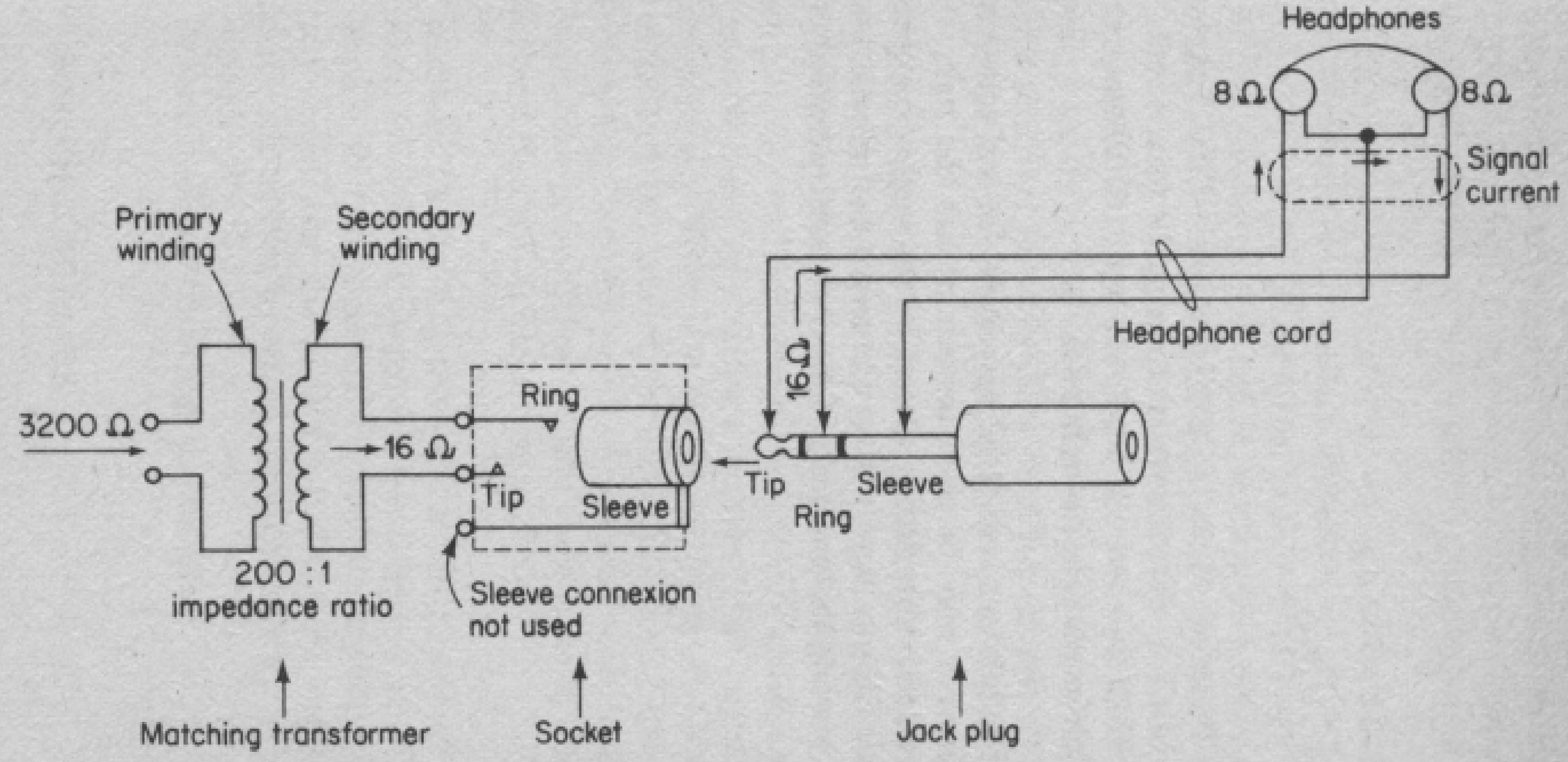
Headphones
8 Ω
8 Ω
Signal current
Primary winding
Secondary winding
Headphone cord
16 Ω
3200 Ω
16 Ω
Ring
Tip
Sleeve
Tip
Sleeve
Ring
200 : 1 impedance ratio
Sleeve connexion not used
Matching transformer
Socket
Jack plug

Fig. 11 Matching a high impedance receiver to low impedance headphones

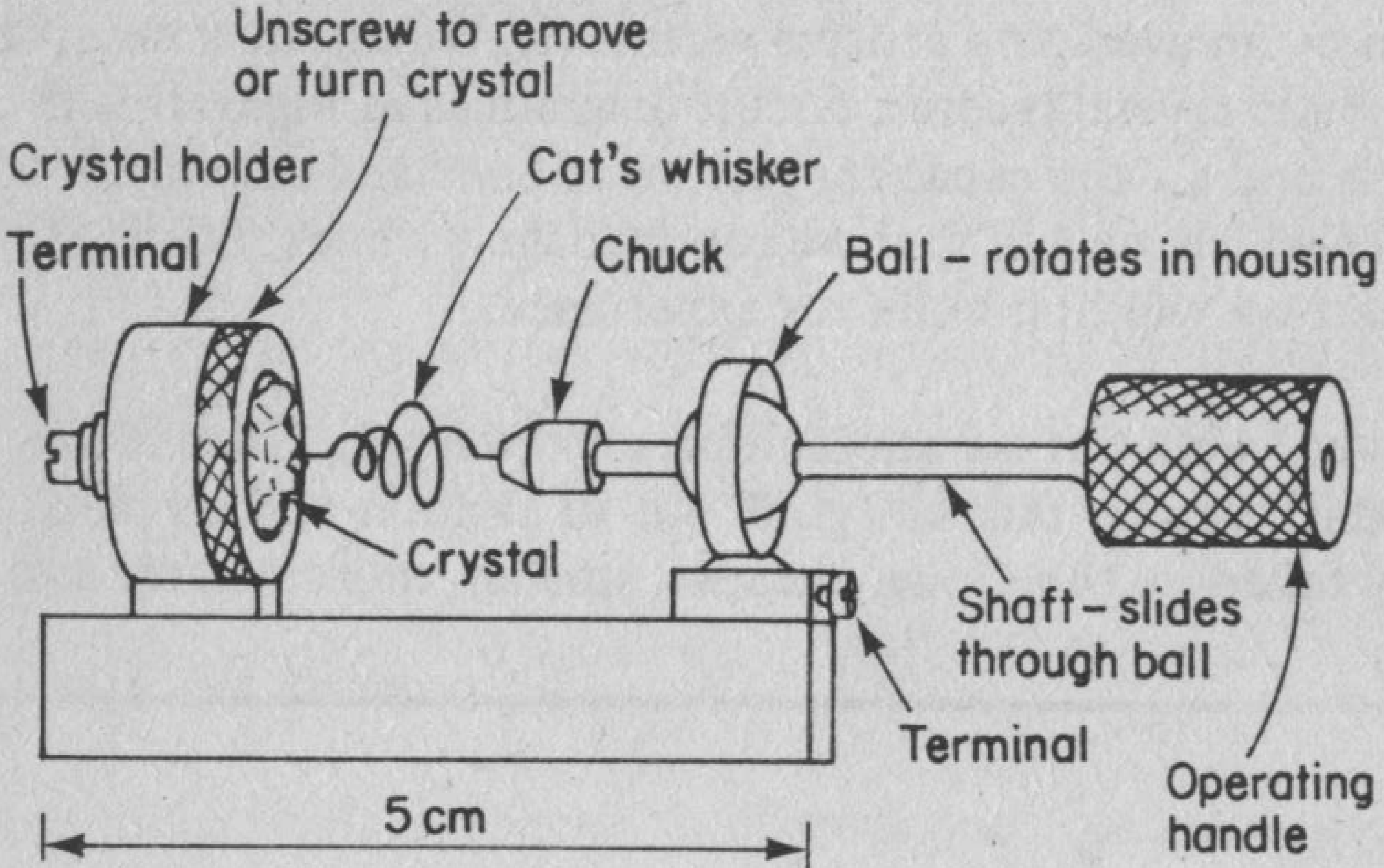
Unscrew to remove
or turn crystal
Crystal holder
Cat's whisker
Terminal
Chuck
Ball – rotates in housing
Crystal
Shaft – slides
through ball
Terminal
Operating
handle
5 cm

Modern diode (7·5 X 2·5 mm)

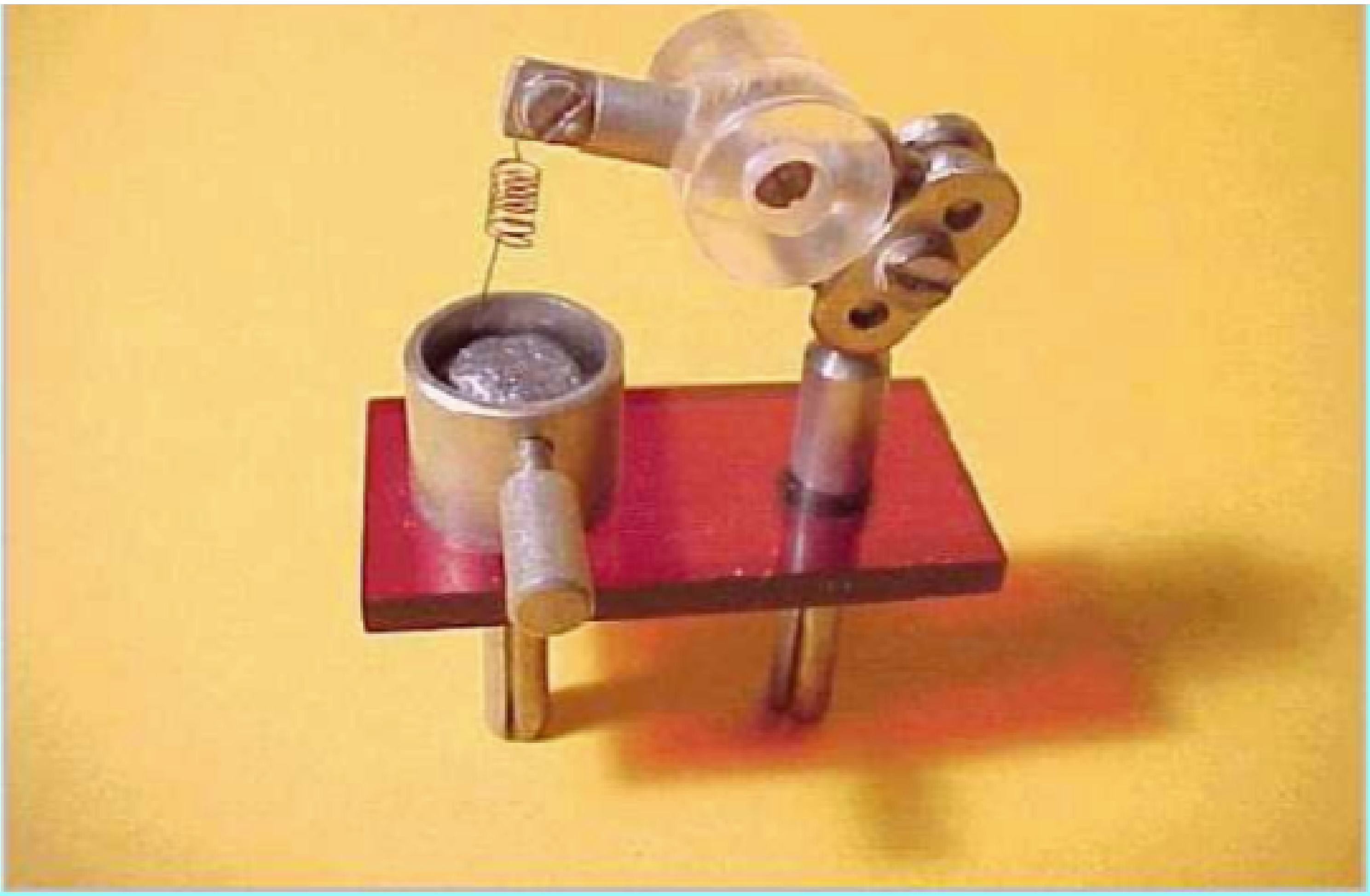

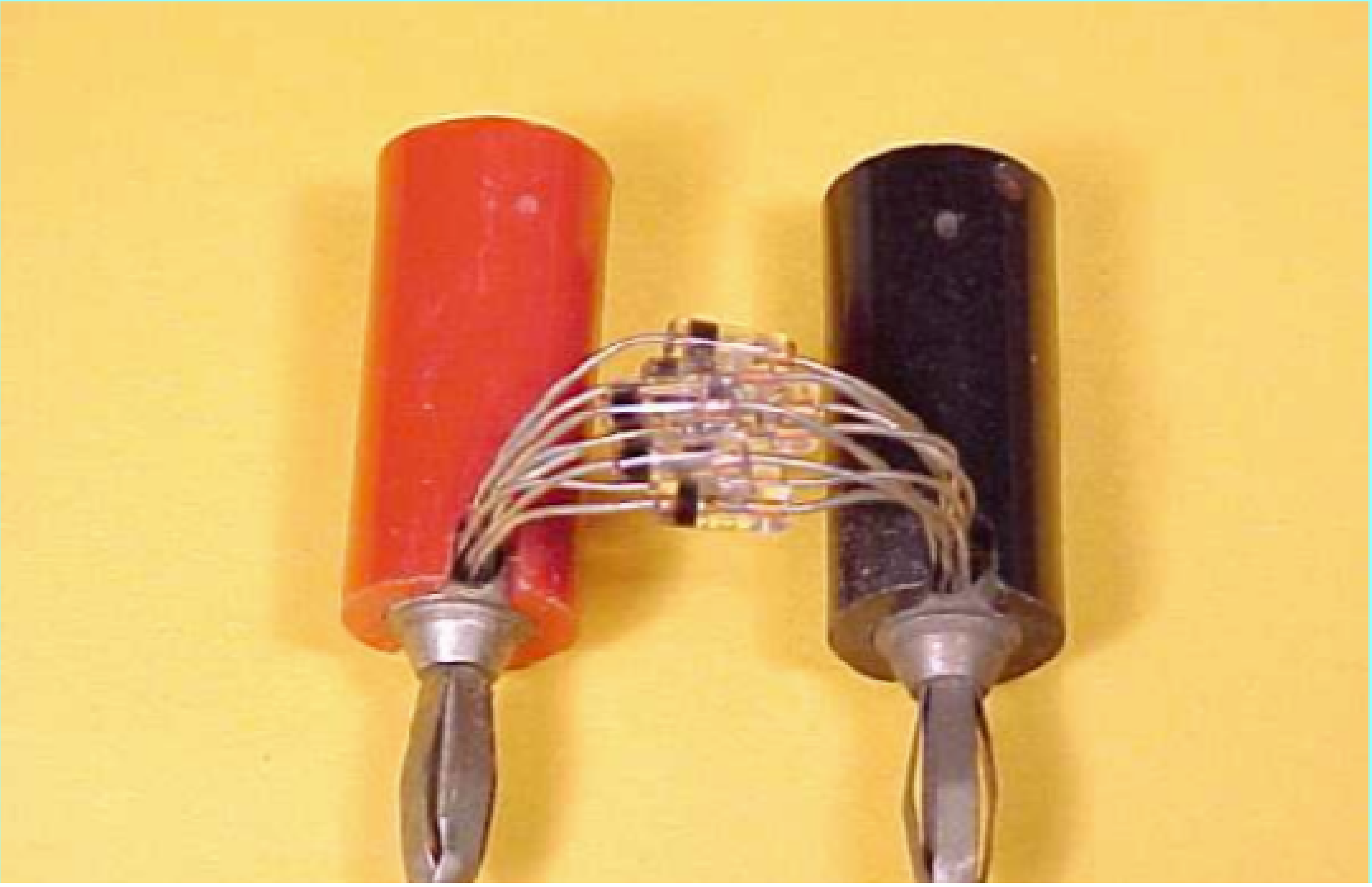

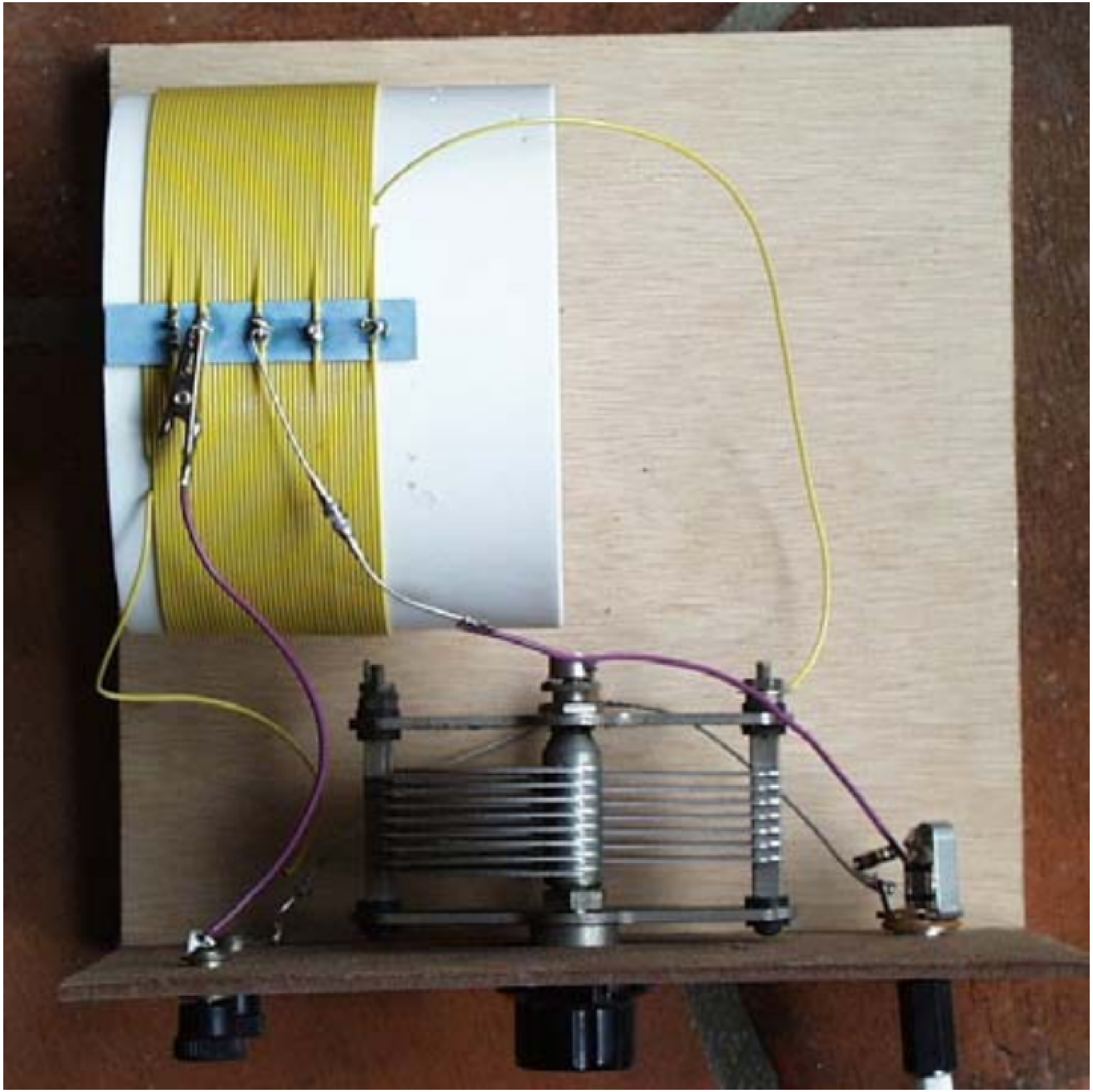

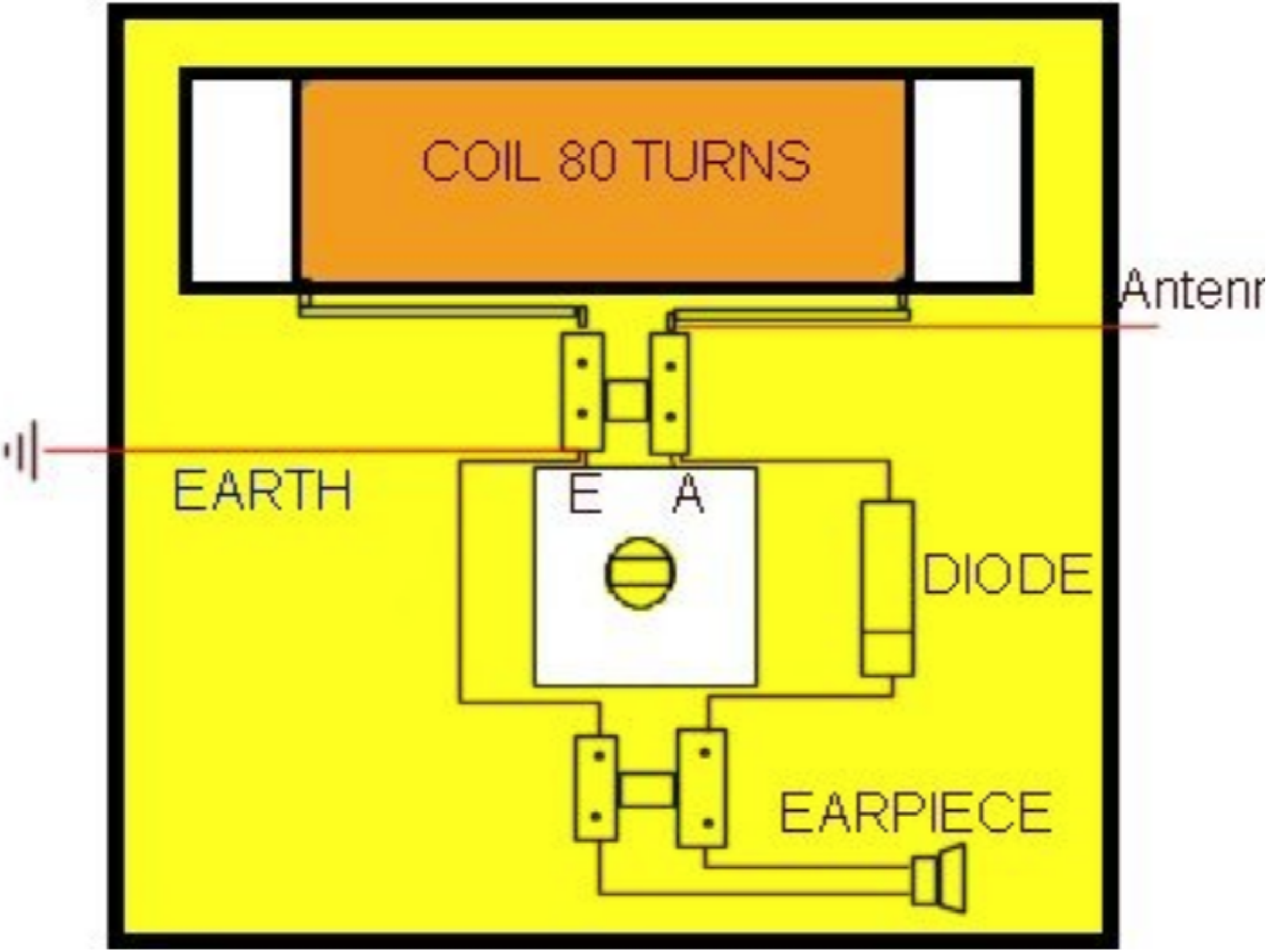
COIL 80 TURNS
Antenn
EARTH
E
A
DIODE
EARPIECE

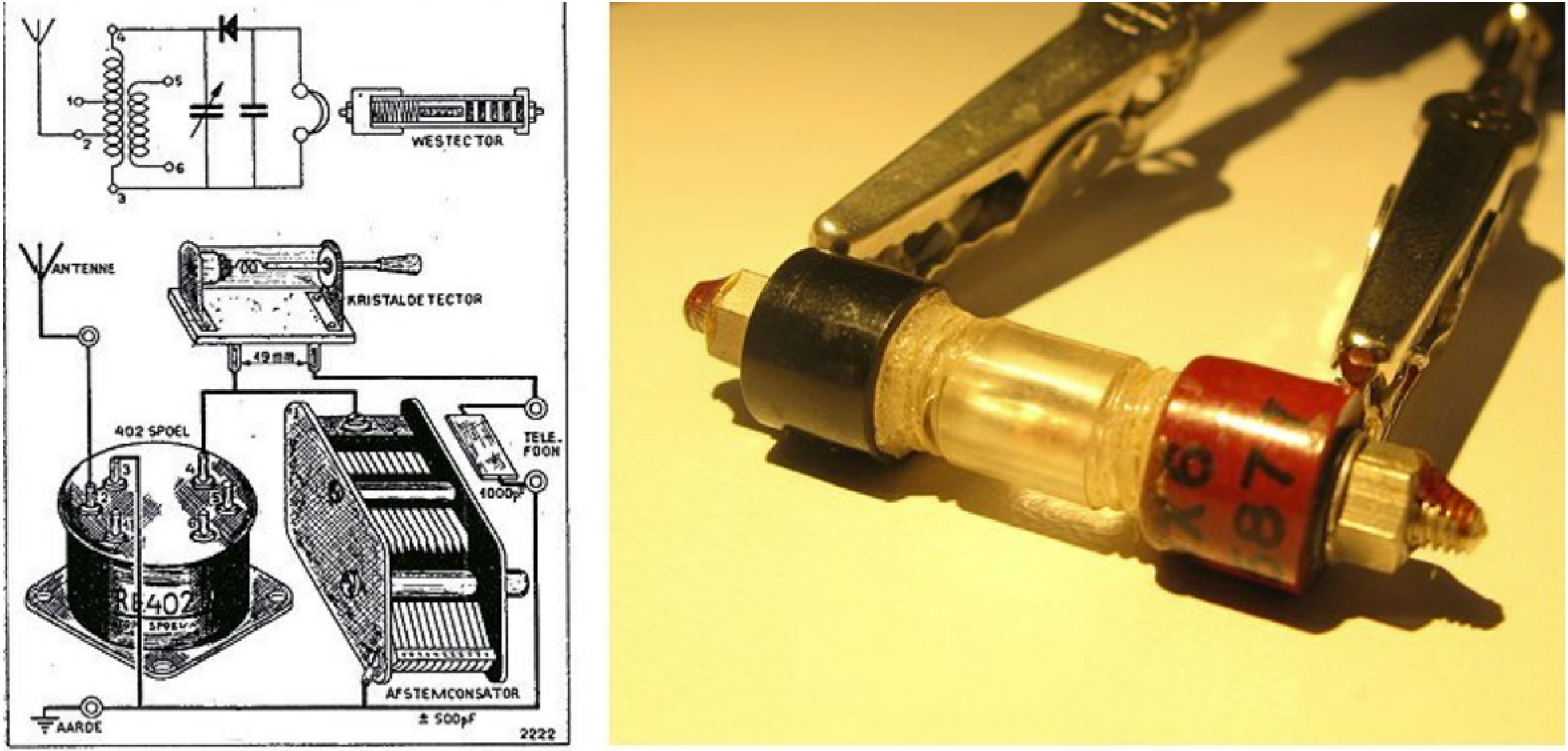
WESTECTOR
1
2
3
4
5
6
ANTENNE
KRISTALDETECTOR
49mm
402 SPOEL
TELE.
FOON
1000pF
RE402
SPOEL
AFSTEMCONSATOR
± 500pF
AARDE
2222
X6
87

Figure 18 The first crystal-set receiver published by "Radio Bulletin" in the July number of 1946 promoted the brand new Mu-CORE 402 (left). As an alternative to a "cat's whisker" detector the article suggest the WX6 "Westector". The picture on the right shows the wx6 "under test".

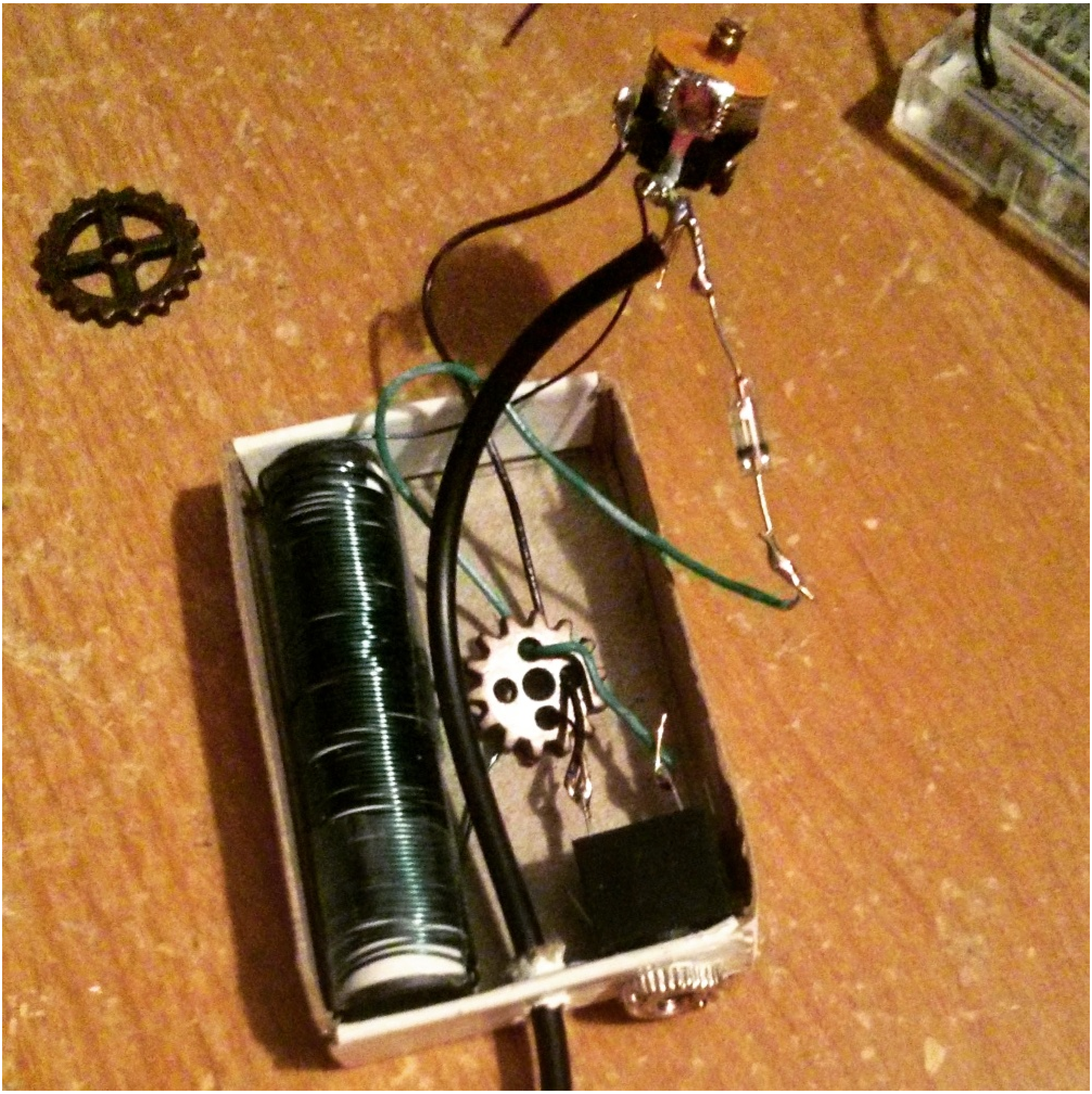

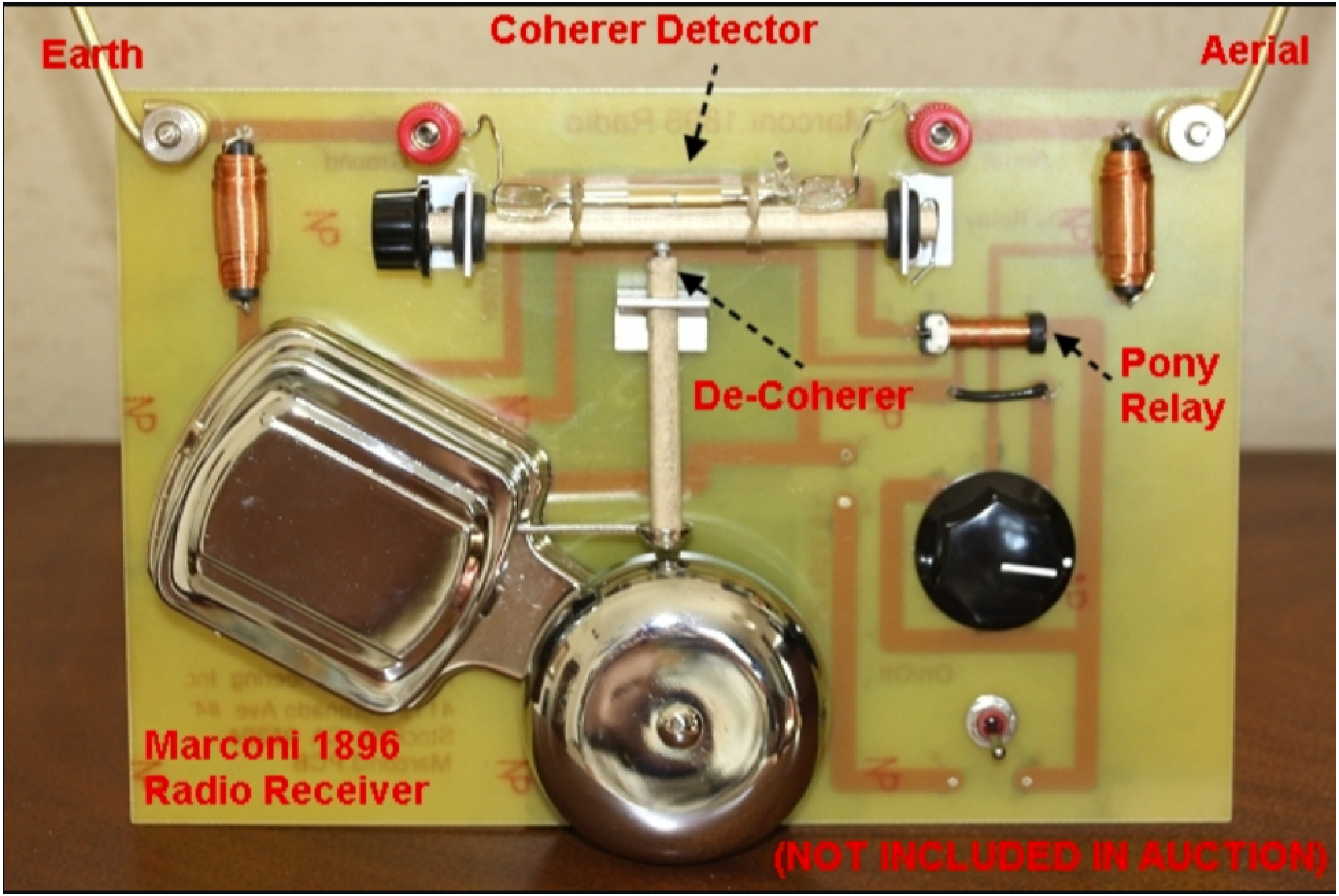
Earth
Coherer Detector
Aerial
De-Coherer
Pony
Relay

Marconi 1896
Radio Receiver

(NOT INCLUDED IN AUCTION)

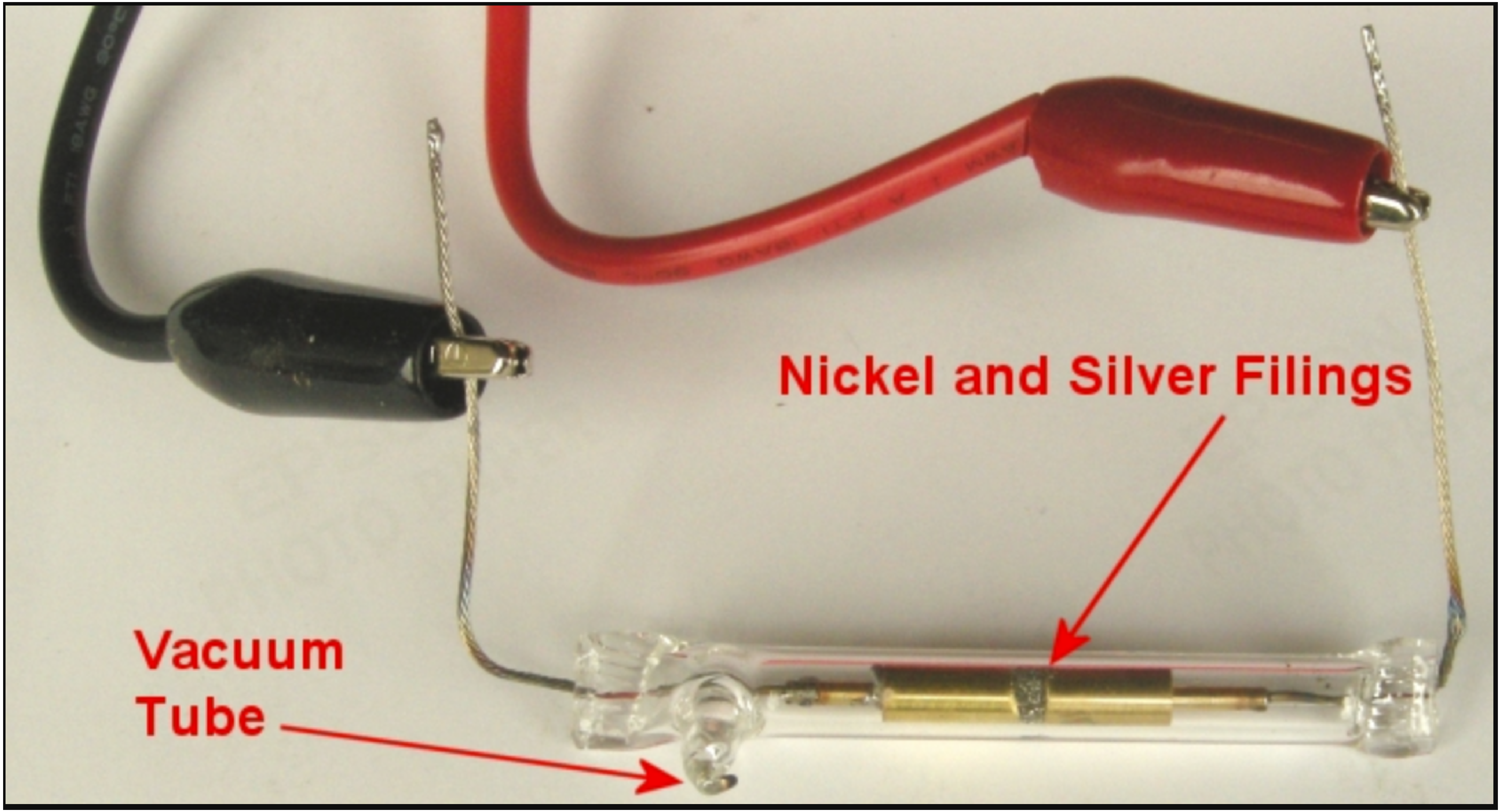
Nickel and Silver Filings
Vacuum Tube

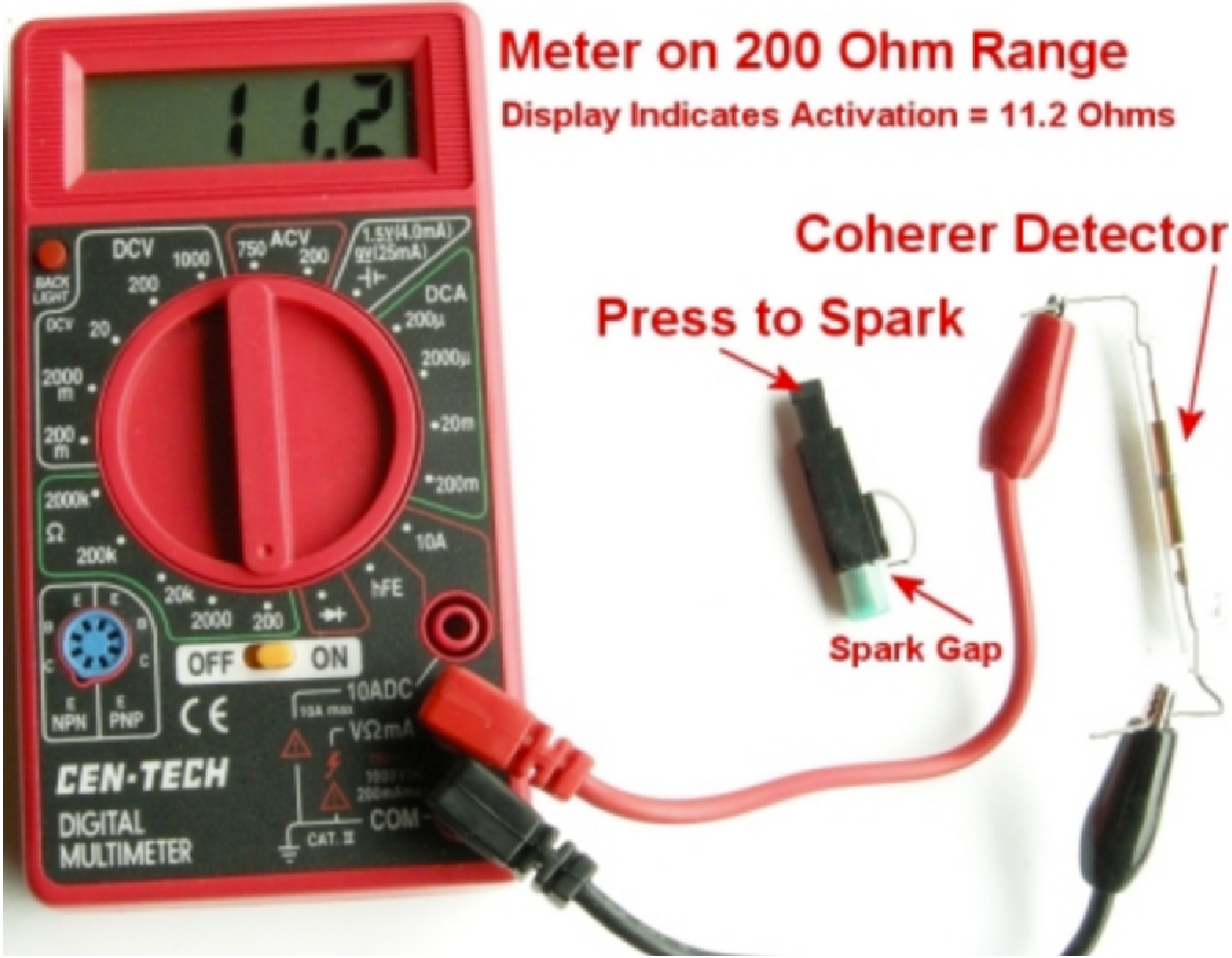

Meter on 200 Ohm Range

Display Indicates Activation = 11.2 Ohms

Coherer Detector

Press to Spark

Spark Gap

11.2
DCV
ACV
DCA
hFE
10A
OFF
ON
10ADC
VΩmA
COM
CEN-TECH
DIGITAL
MULTIMETER

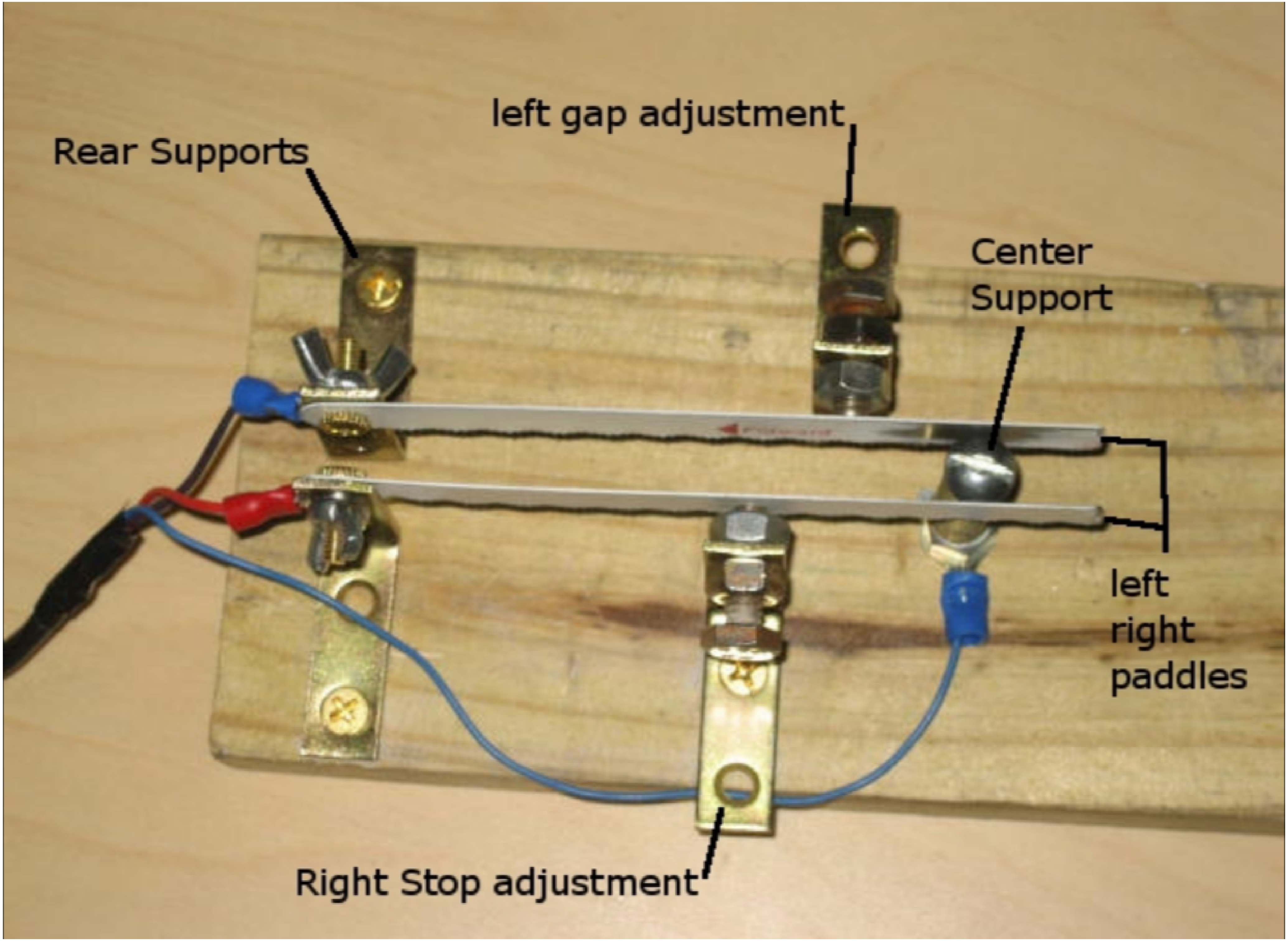
left gap adjustment
Rear Supports
Center Support
left right paddles
Right Stop adjustment

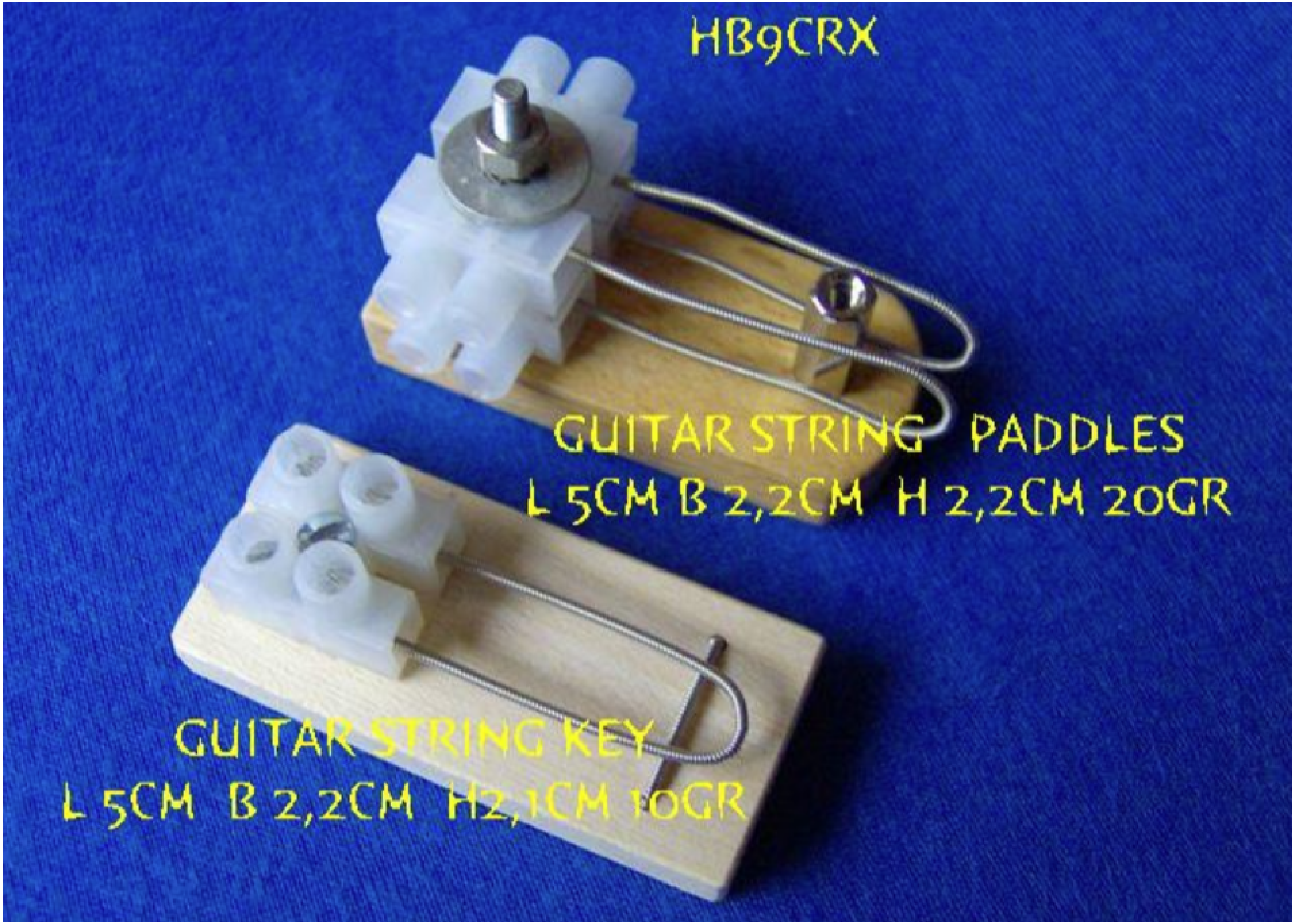
HB9CRX
GUITAR STRING PADDLES
L 5CM B 2,2CM H 2,2CM 20GR
GUITAR STRING KEY
L 5CM B 2,2CM H2,1CM 10GR

Heathkit CR1 Crystal Radio clone

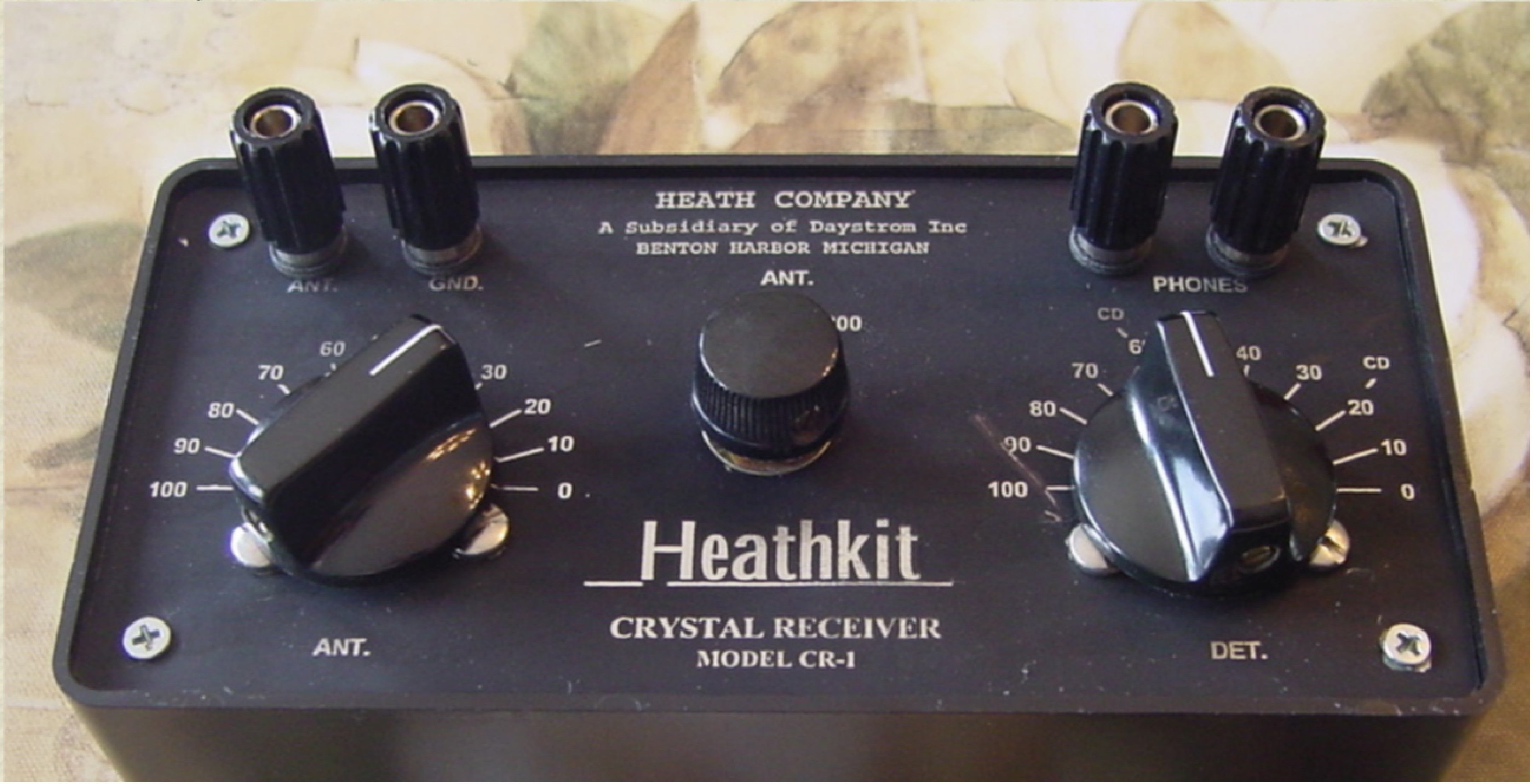
HEATH COMPANY
A Subsidiary of Daystrom Inc
BENTON HARBOR MICHIGAN
ANT.
GND.
ANT.
PHONES
100
90
80
70
60
30
20
10
0
CD
40
Heathkit
CRYSTAL RECEIVER
MODEL CR-1
ANT.
DET.

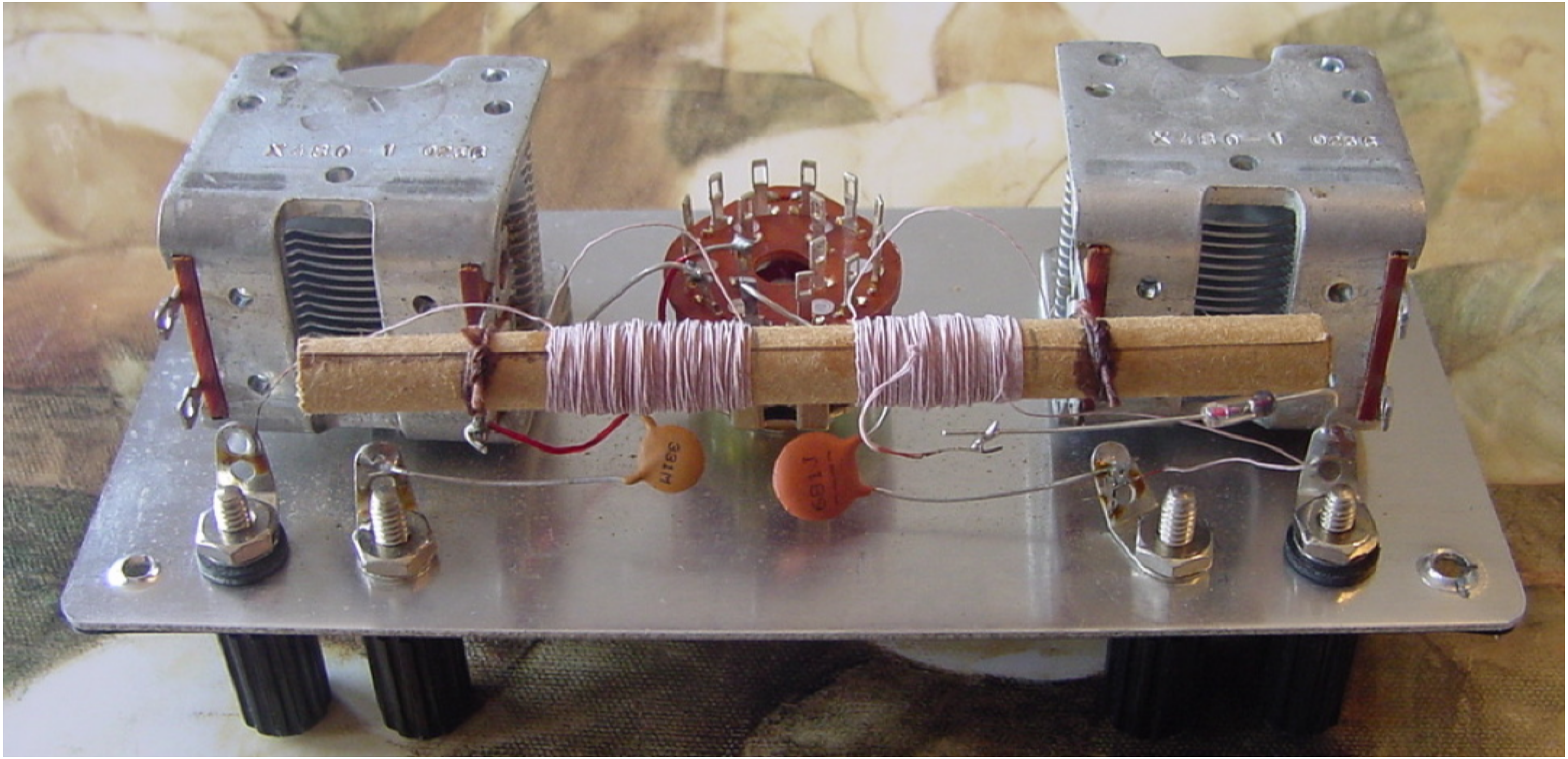
X480-1 0236
X480-1 0236
331M
681J

Another Prototype similar to the Heathkit using toroid coil

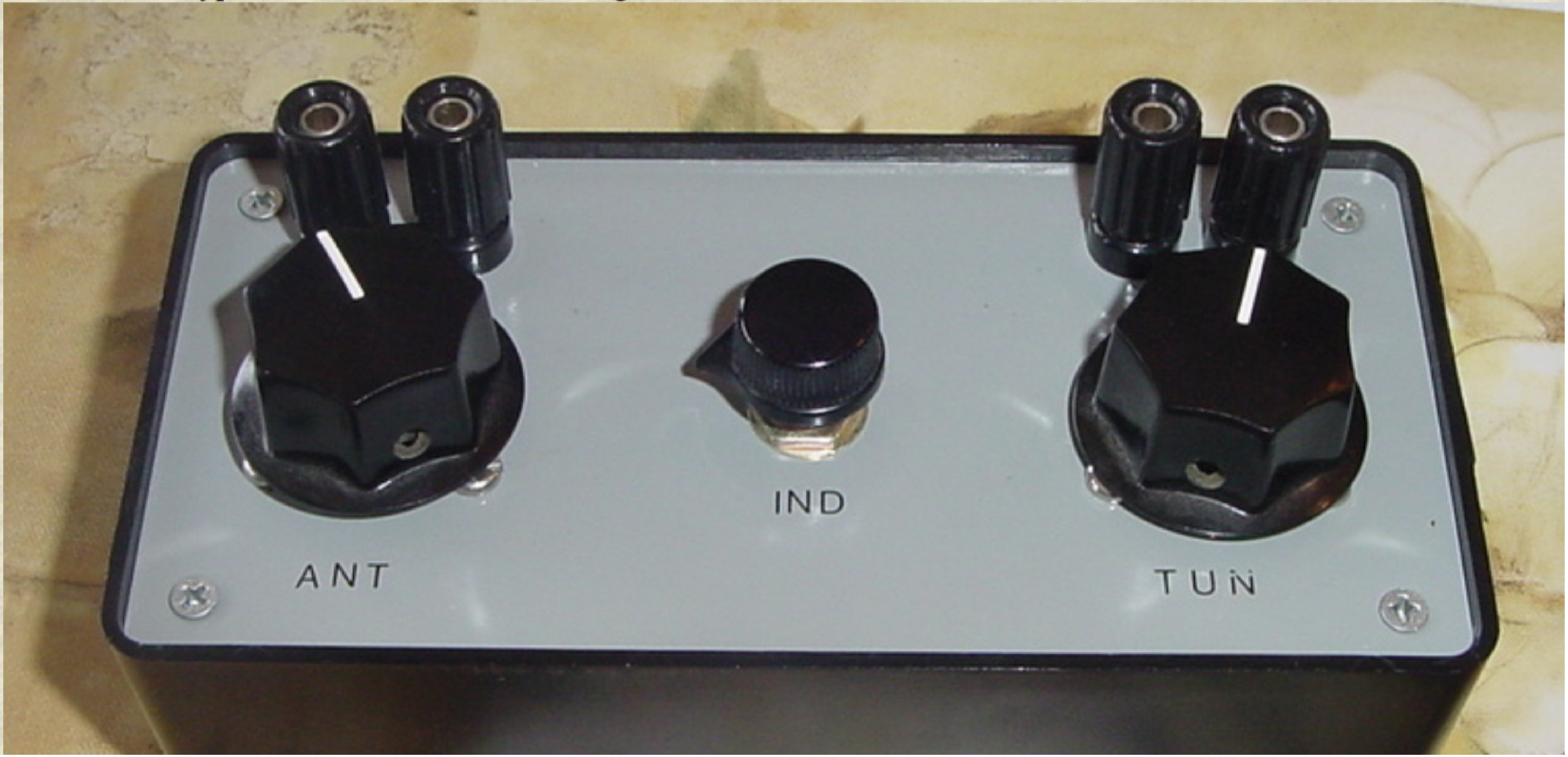

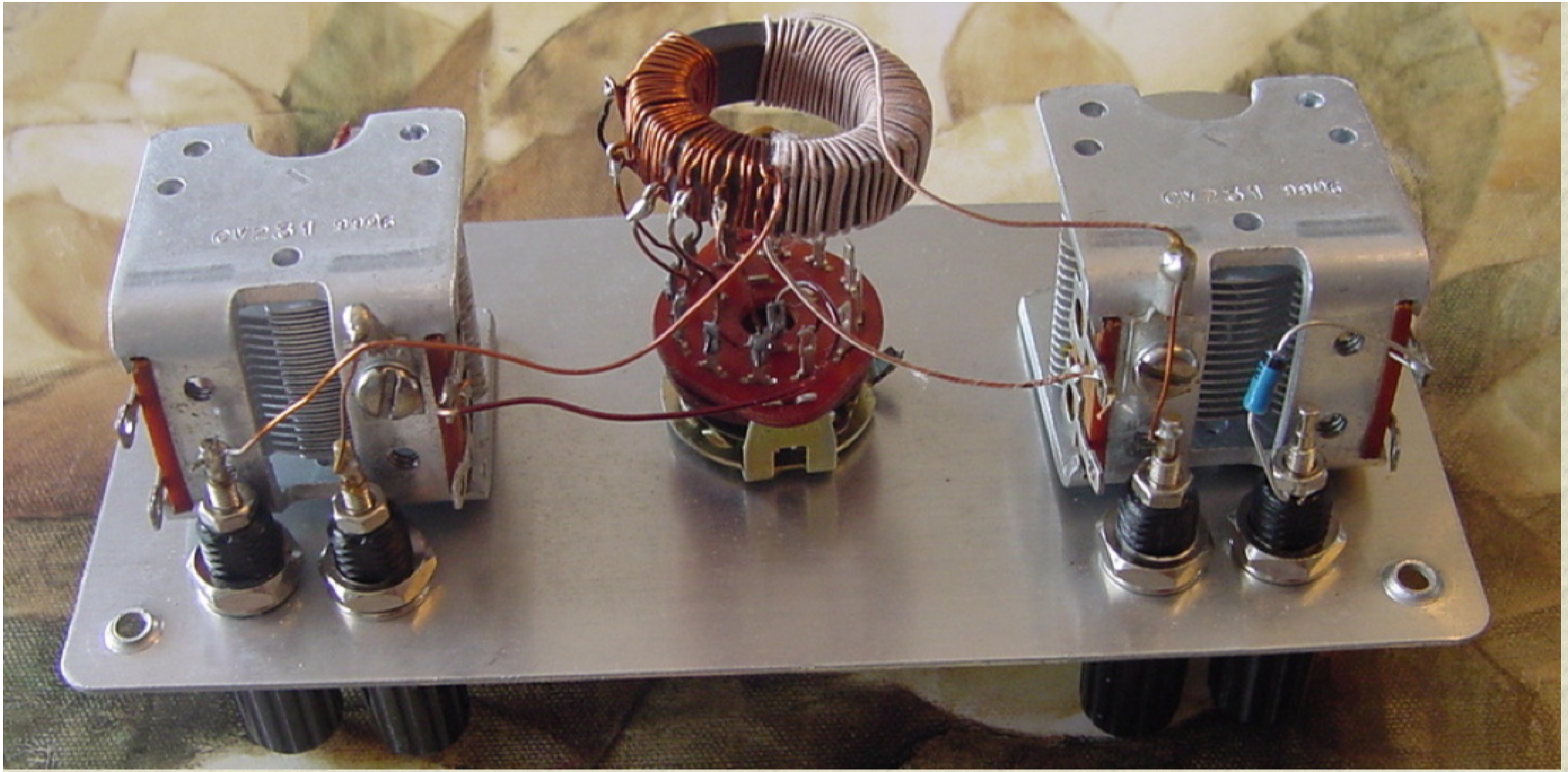
CV231 9906
CV231 9906

This is a dual tuned crystal set made in a box like the CR-1, just a fun prototype I made a few years back. It pulls in AM stations loud and clear.

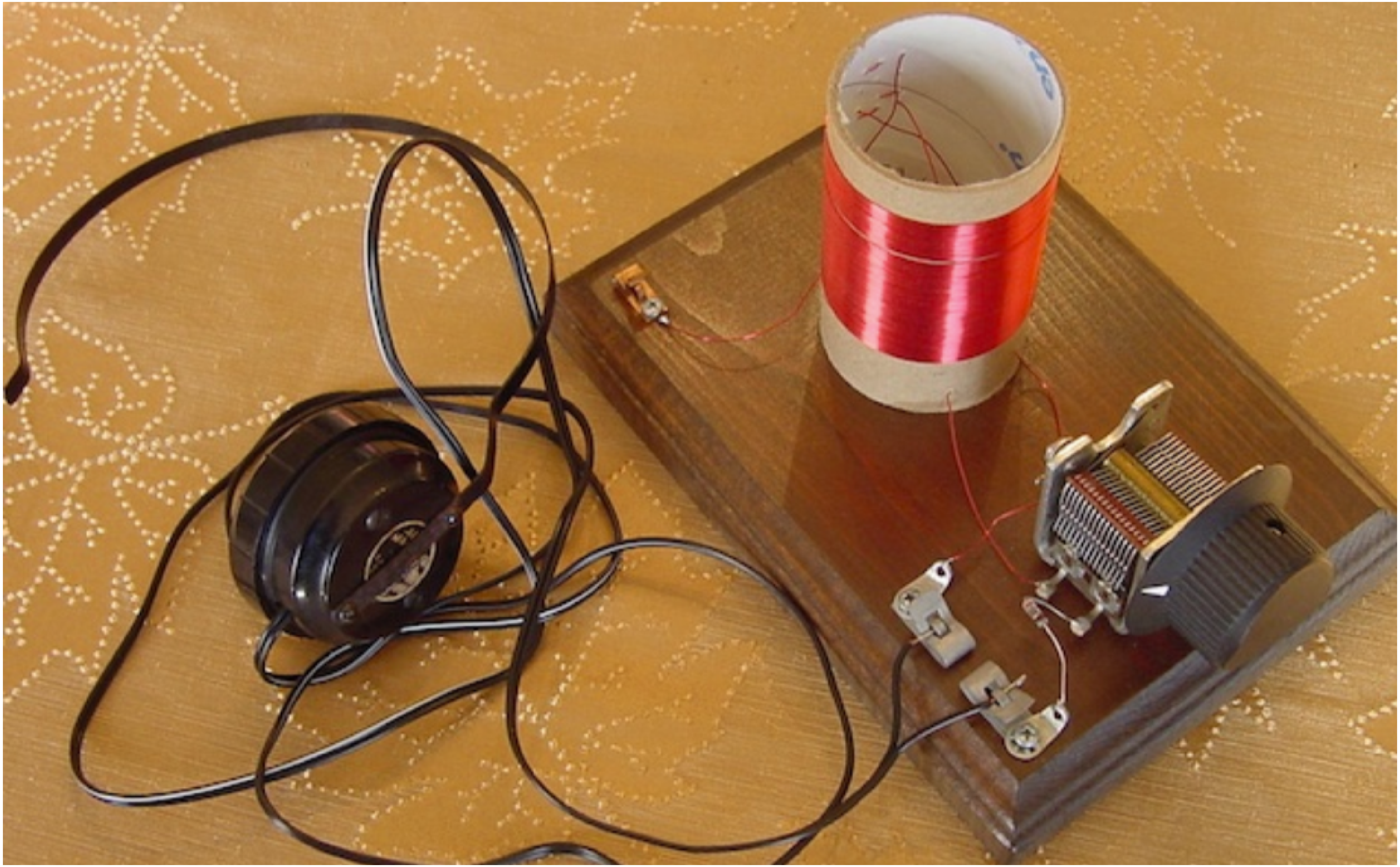

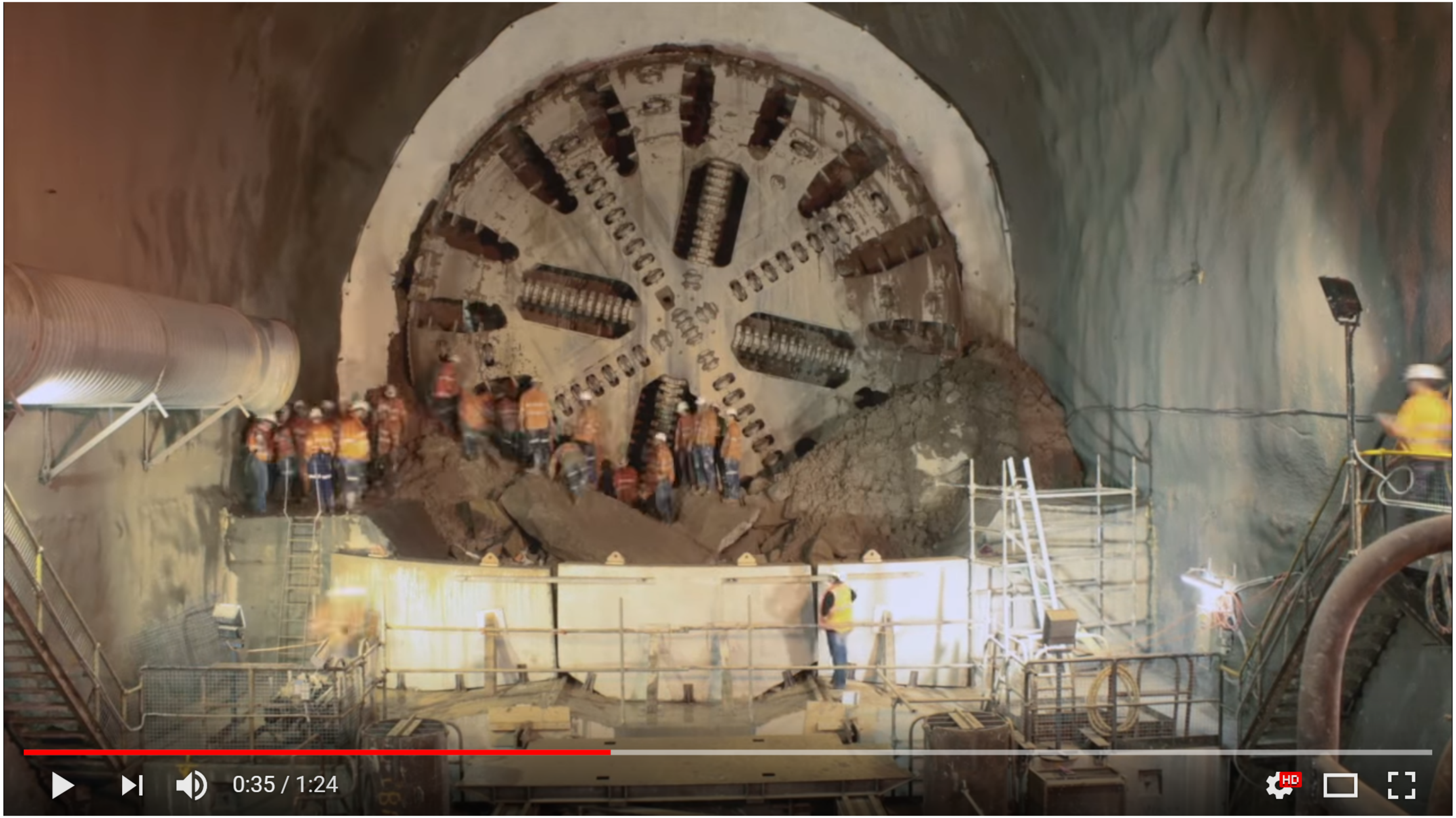

Tunnel Boring Machine Time Lapse Burial

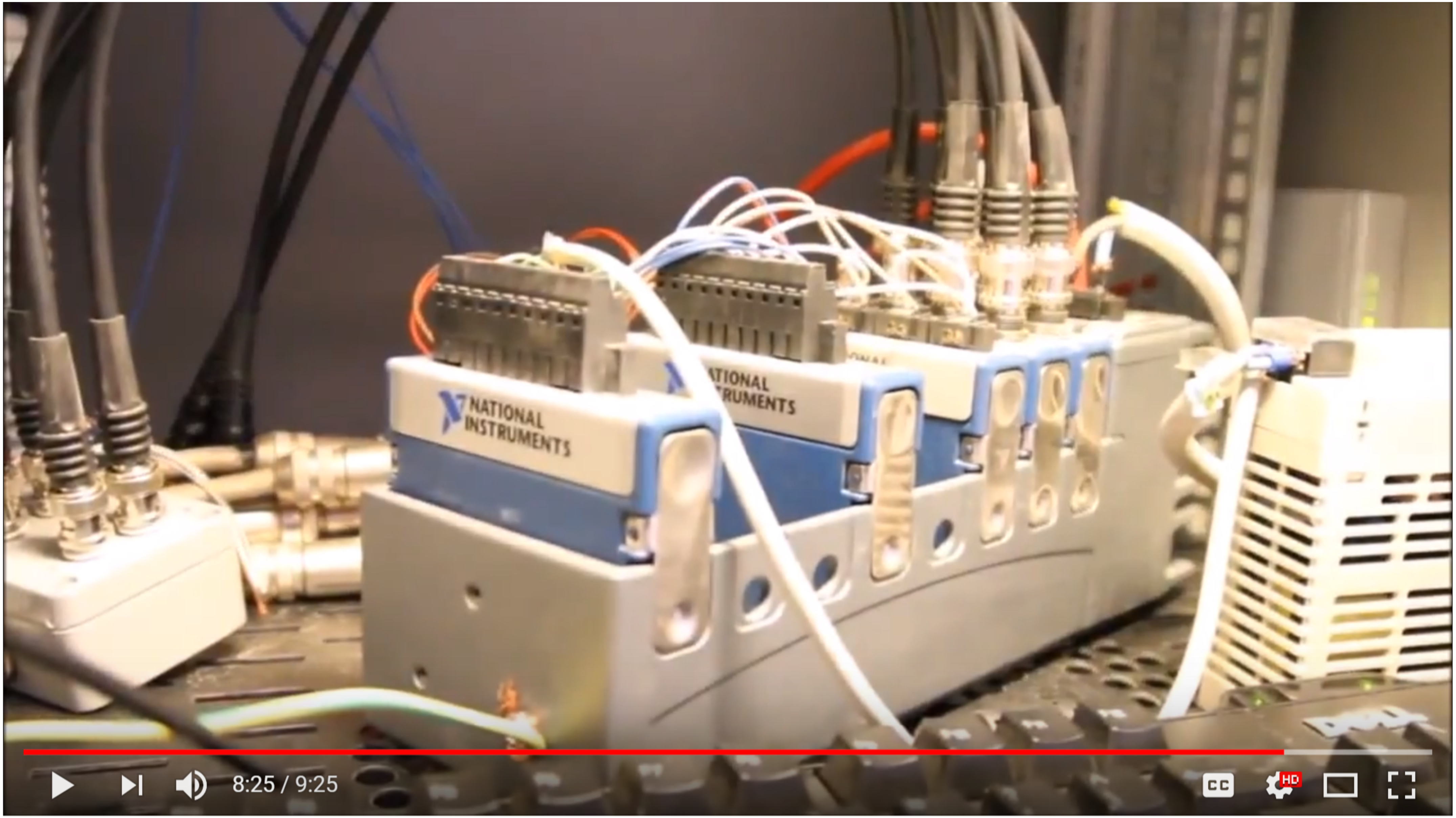
NATIONAL INSTRUMENTS
NATIONAL INSTRUMENTS
DELL
8:25 / 9:25
HD

Flat Earth CON SPIRES! Cathedral Spires Are Secretly Atmospheric Electricity Ma

1:45 / 9:25
CC
HD

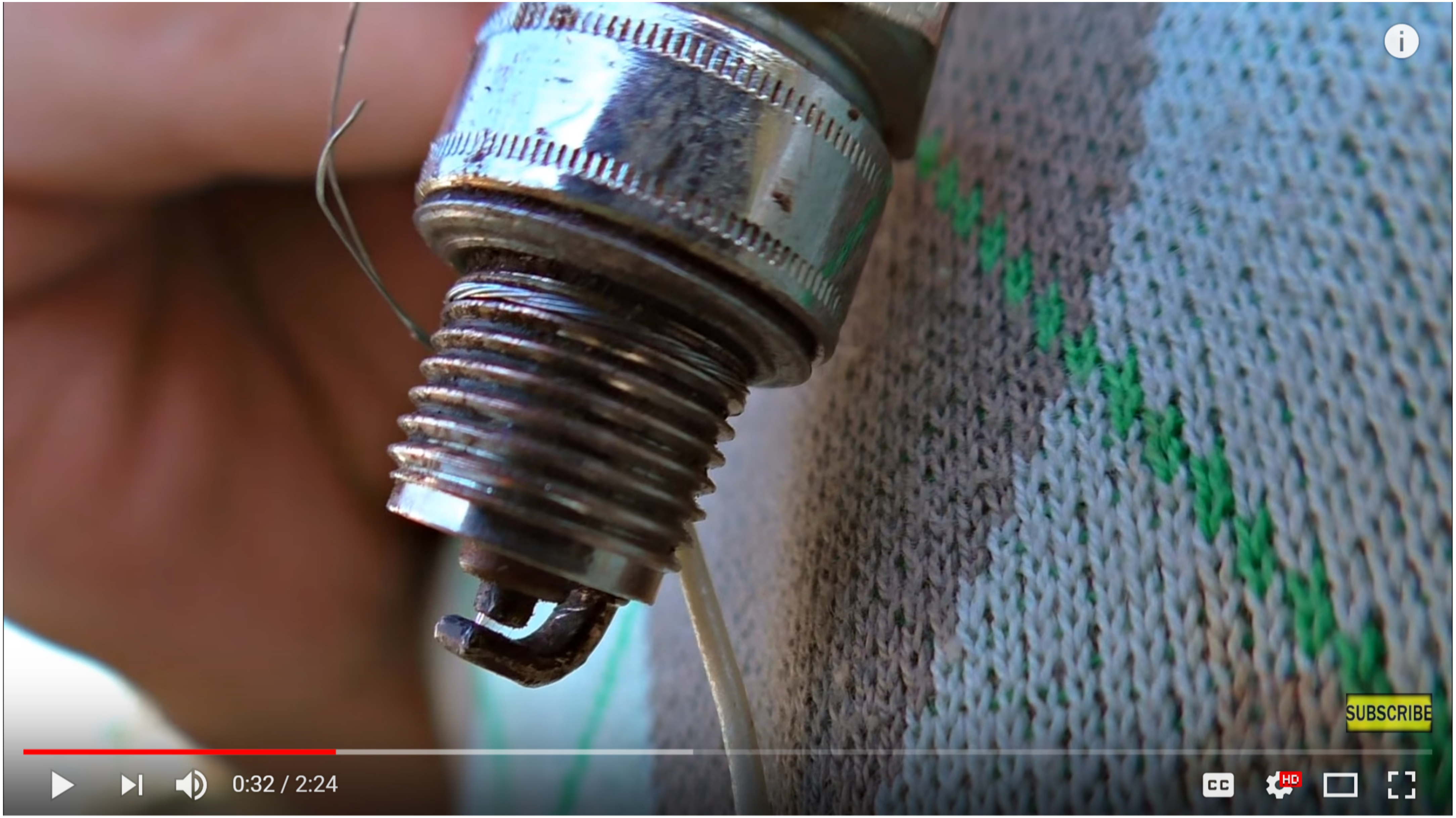
SUBSCRIBE
0:32 / 2:24
HD

Charging from the power lines.

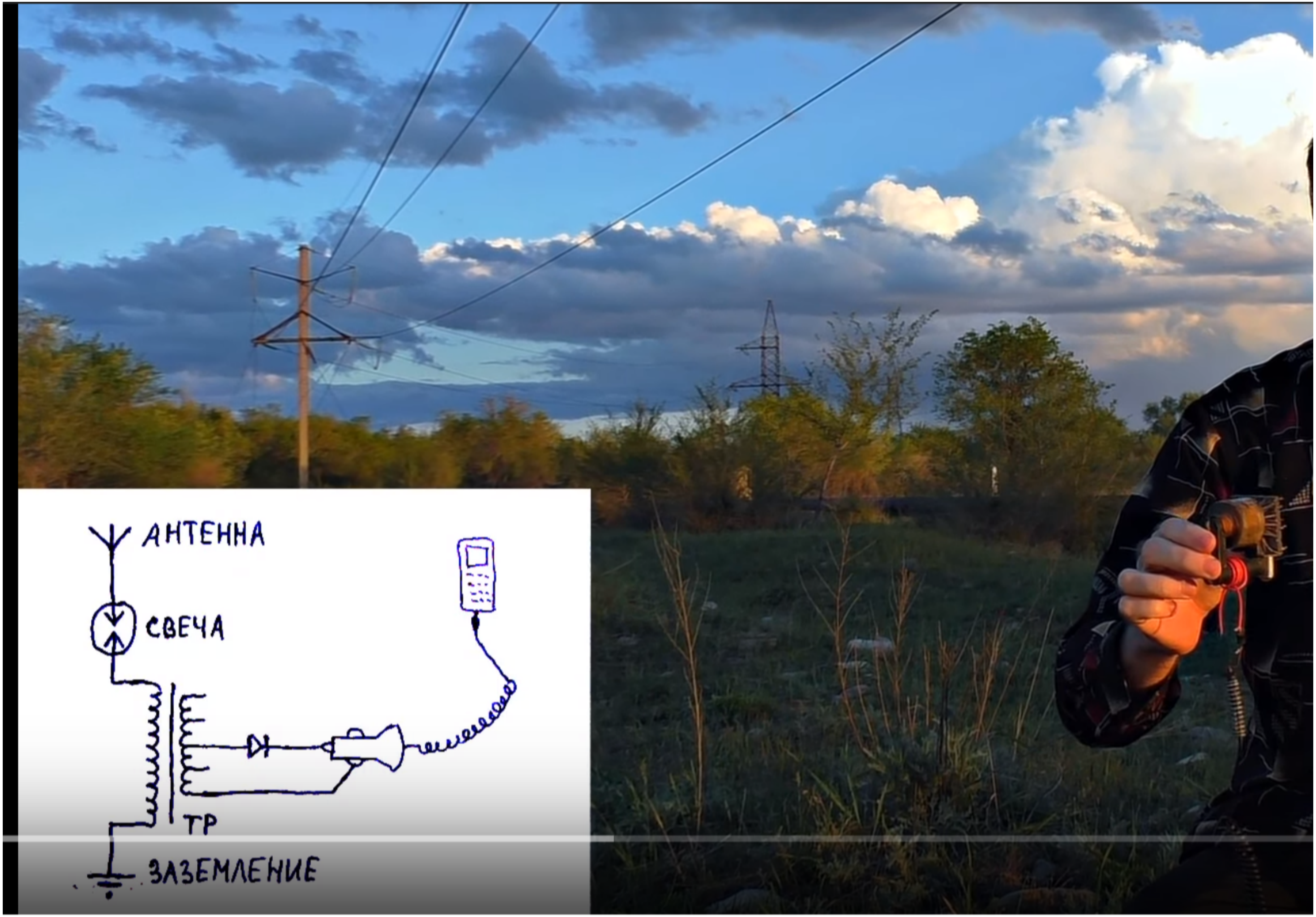
АНТЕННА
СВЕЧА
ТР
ЗАЗЕМЛЕНИЕ

E6

MULTI DISPLAY DMM
10.1
POWER
COM
VΩ
DT-920

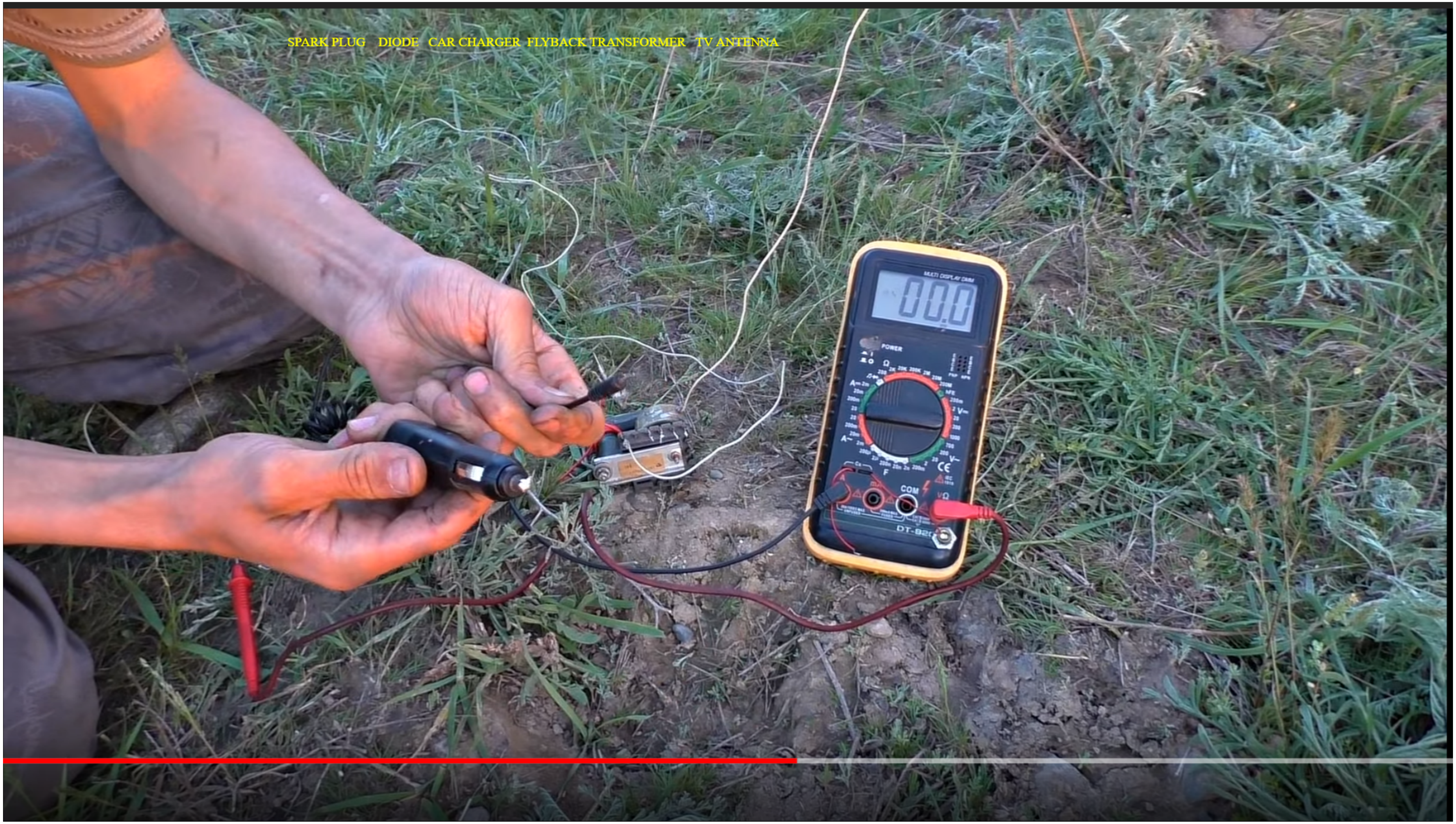
SPARK PLUG
DIODE
CAR CHARGER
FLYBACK TRANSFORMER
TV ANTENNA

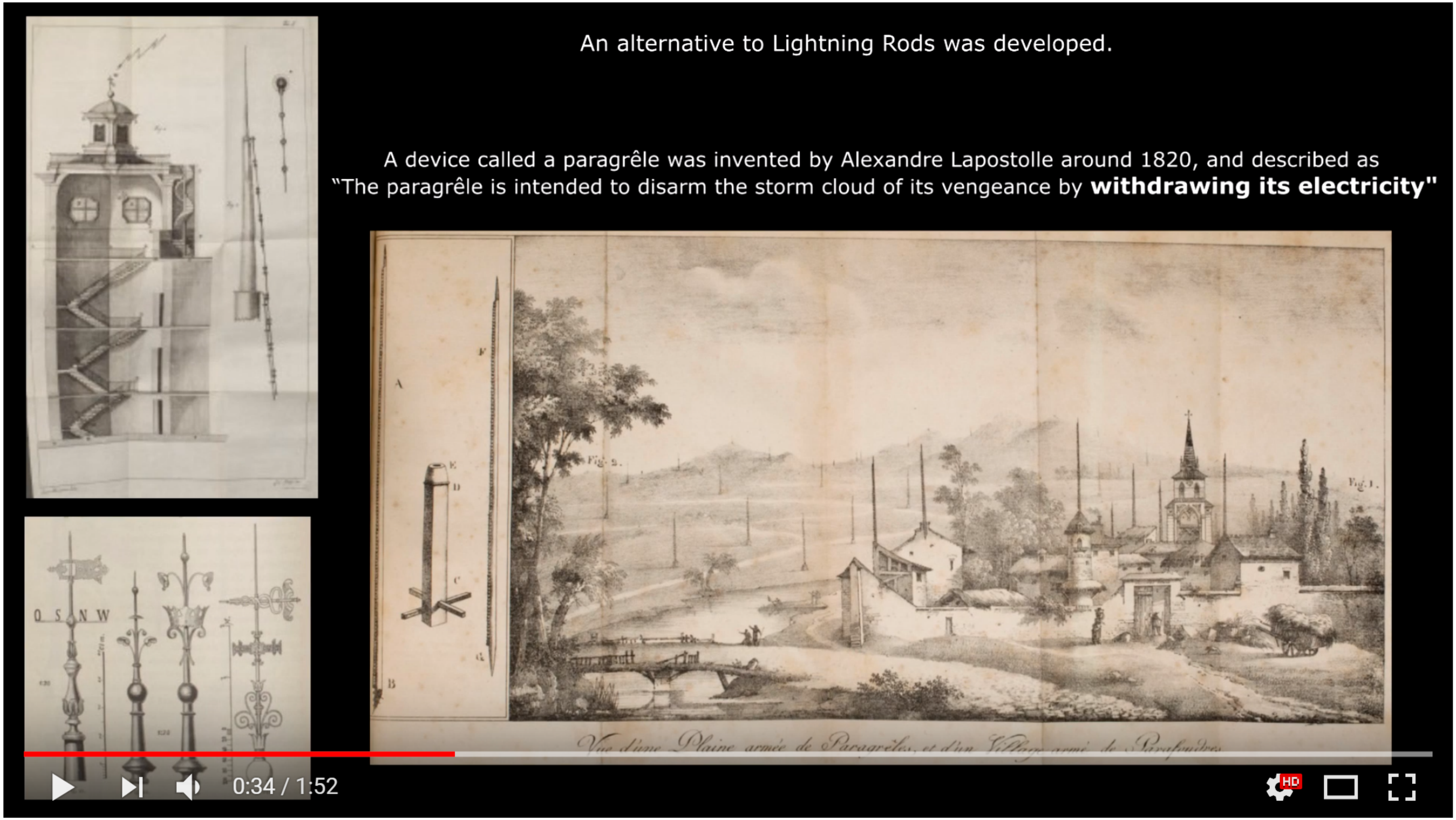

An alternative to Lightning Rods was developed.

A device called a paragrêle was invented by Alexandre Lapostolle around 1820, and described as
"The paragrêle is intended to disarm the storm cloud of its vengeance by **withdrawing its electricity**"

Vue d'une Plaine armée de Paragrêles, et d'un Village armé de Parafoudres
Fig. 1.
Fig. 2.
A
B
C
D
E
F
G
O S N W
0:34 / 1:52
HD

Atmospheric Electric and the Church

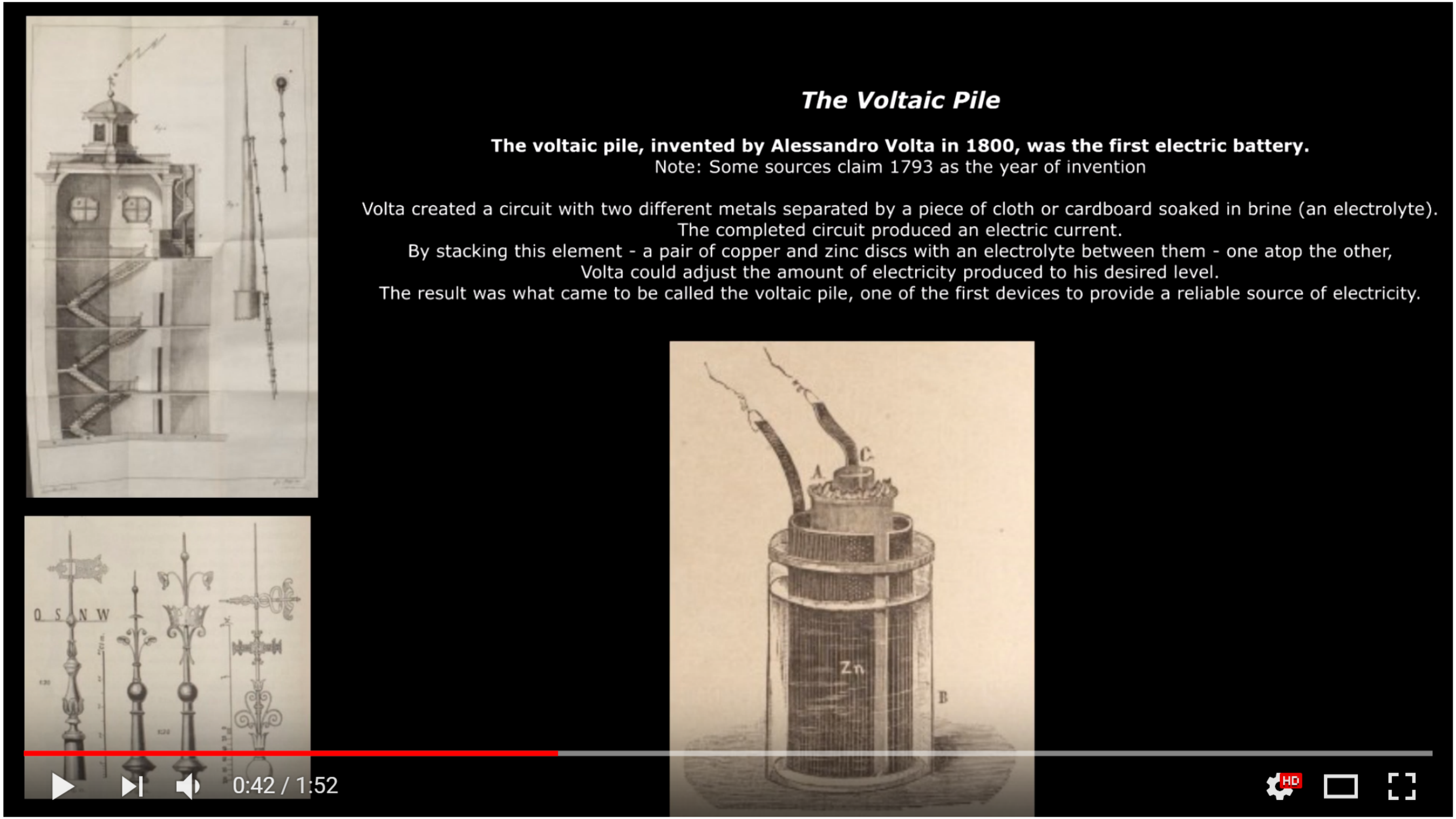

The Voltaic Pile

The voltaic pile, invented by Alessandro Volta in 1800, was the first electric battery.
Note: Some sources claim 1793 as the year of invention

Volta created a circuit with two different metals separated by a piece of cloth or cardboard soaked in brine (an electrolyte).
The completed circuit produced an electric current.
By stacking this element - a pair of copper and zinc discs with an electrolyte between them - one atop the other,
Volta could adjust the amount of electricity produced to his desired level.
The result was what came to be called the voltaic pile, one of the first devices to provide a reliable source of electricity.

It's easy to conclude, that between 1793 and 1820,
it became possible to roll out free Atmospheric Electricity to the world.

So why didn't they?

O S N W
0:56 / 1:52
HD

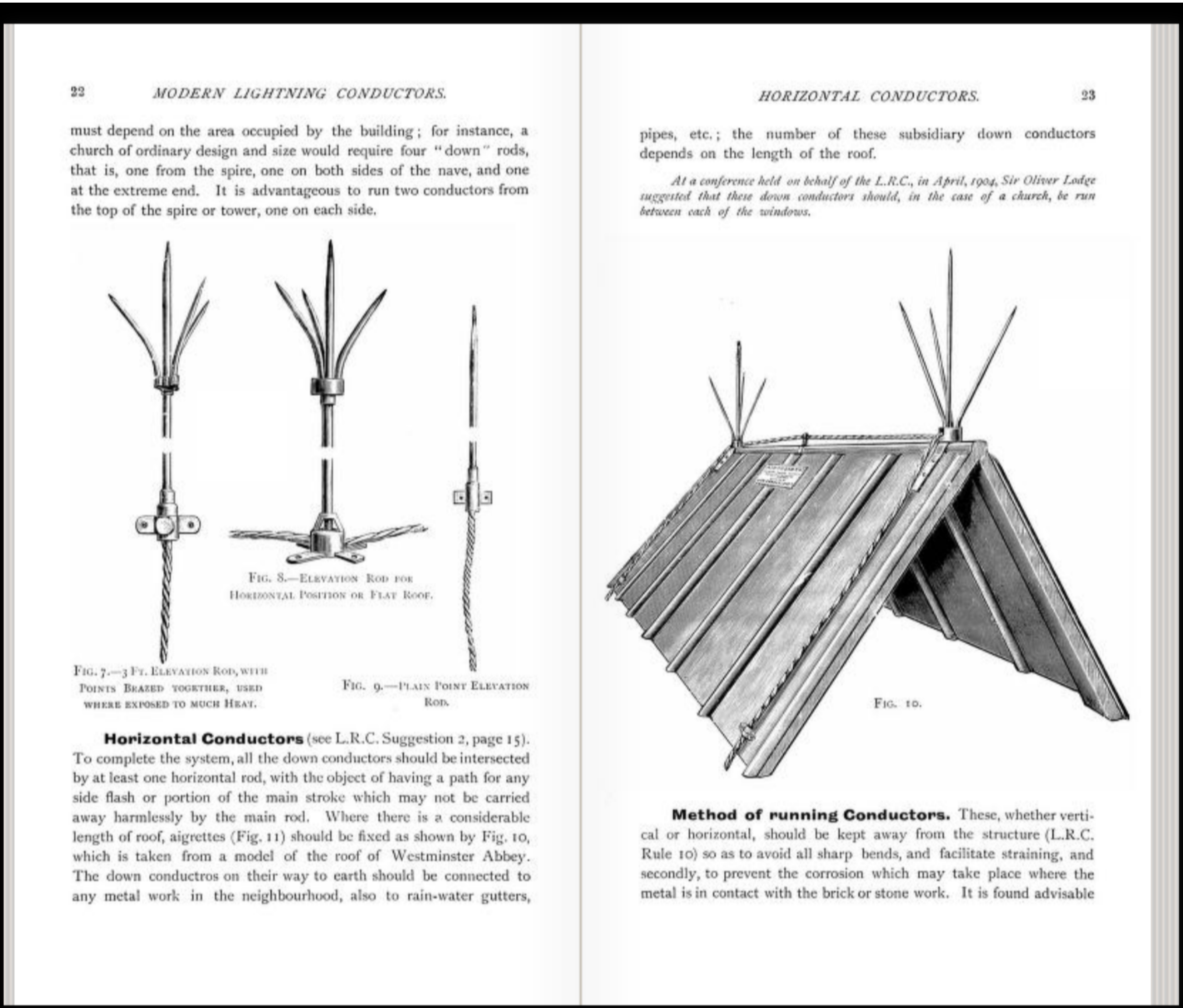

must depend on the area occupied by the building; for instance, a church of ordinary design and size would require four "down" rods, that is, one from the spire, one on both sides of the nave, and one at the extreme end. It is advantageous to run two conductors from the top of the spire or tower, one on each side.

FIG. 7.—3 FT. ELEVATION ROD, WITH POINTS BRAZED TOGETHER, USED WHERE EXPOSED TO MUCH HEAT.

FIG. 8.—ELEVATION ROD FOR HORIZONTAL POSITION OR FLAT ROOF.

FIG. 9.—PLAIN POINT ELEVATION ROD.

Horizontal Conductors (see L.R.C. Suggestion 2, page 15). To complete the system, all the down conductors should be intersected by at least one horizontal rod, with the object of having a path for any side flash or portion of the main stroke which may not be carried away harmlessly by the main rod. Where there is a considerable length of roof, aigrettes (Fig. 11) should be fixed as shown by Fig. 10, which is taken from a model of the roof of Westminster Abbey. The down conductros on their way to earth should be connected to any metal work in the neighbourhood, also to rain-water gutters, pipes, etc.; the number of these subsidiary down conductors depends on the length of the roof.

At a conference held on behalf of the L.R.C., in April, 1904, Sir Oliver Lodge suggested that these down conductors should, in the case of a church, be run between each of the windows.

FIG. 10.

Method of running Conductors. These, whether vertical or horizontal, should be kept away from the structure (L.R.C. Rule 10) so as to avoid all sharp bends, and facilitate straining, and secondly, to prevent the corrosion which may take place where the metal is in contact with the brick or stone work. It is found advisable

Modern lightning conductors: an illustrated

surface of the joint need not necessarily exceed that of the cross section of the conductors. The joint should be put together previously by screws or rivets, and the soldered joint, especially if used in underground work, should be carefully protected from local electrical action by tarred rope. Stranded iron conductors can be connected (as previously described) by use of a box joint; the box, Fig. 28, must be of the same metal as the conductors.

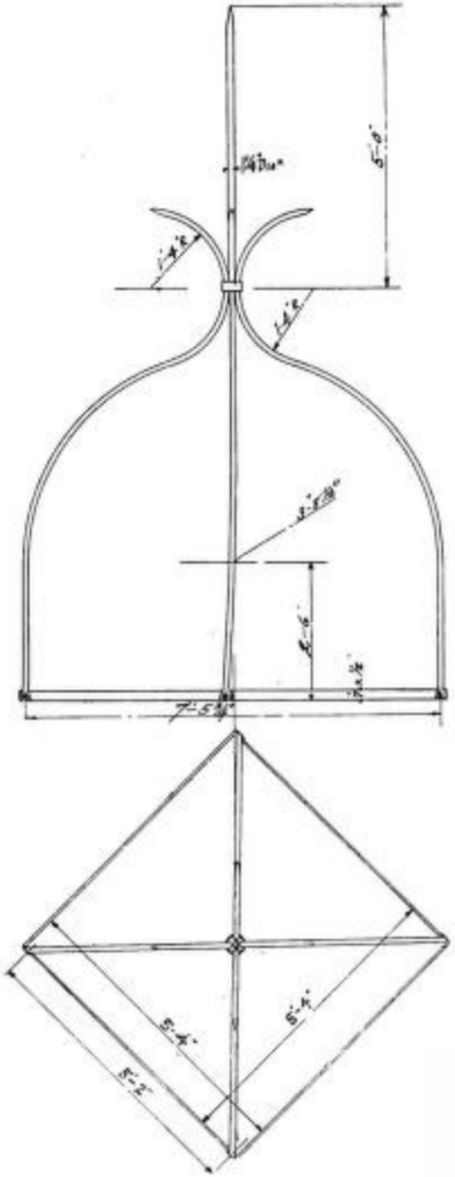

FIG. 27.—TERMINALS IN FORM OF AN ARCH FOR CHIMNEY STACK.

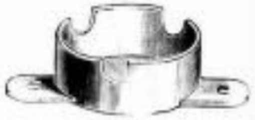

FIG. 28.

Vanes.—Particular attention must be paid to the necessity of making a permanent joint to the spindle. A clamp is prepared of the same material as the spindle, and is furnished with two bolts to tighten; if iron is used it is well to line the clamp with a piece of sheet lead. The conductor is sweated into a socket which is fitted with an eye, through which one of the tightening bolts passes. In the case of the vanes of churches and those fixed in inaccessible positions, two separate clamps should be used.

Internal Masses of Metal.—Roof trusses fitted with longitudinal iron tie rods will, as a rule, be found to be electrically connected, but should this not be the case each truss must be joined to the conductors. All large and long masses of metal, such as beams, girders, roof trusses, tie rods, hot water systems, traveller ways, hoisting crabs, engines, boilers, large machines, and ventilators fixed in the interiors of buildings, should be connected to all conductors that pass near them, and as far as possible with one another. The discontinuous parts of traveller rails should be connected by straps, or in some cases tramway bonds might be used. If electric light wires are run in tubes, such as the "SIMPLEX," this should be earthed. Metallic contact between lead or zinc sheeting and flashings should be carefully studied, and for special work strips of sufficient size should be either burnt on to lead or soldered in such a way that the joint will stand rough usage, and allow for expansion or contraction.

Earth Connection.—"*It is essential that the lower extremity of the conductor be buried in permanently damp soil; hence proximity to rain-water pipes, and to drains, is desirable. It is a very good plan to make the conductor bifurcate close below the surface of the ground, and adopt two of the following methods for securing the escape of the lightning into the earth. A strip of copper tape may be led from the bottom of the rod to the nearest gas or water* main—*not merely to a lead pipe—and be soldered to it; or a tape may be soldered to a sheet of copper* 3 *feet by* 3 *feet and* $\frac{1}{16}$ *inch thick, buried in permanently wet earth, and surrounded by cinders or coke; or many yards of the tape may be laid in a trench filled with coke, taking care that the surfaces of copper are, as in the previous cases, not less than* 18 *square feet. Where iron is used for the rod, a galvanised iron plate of similar dimensions should be employed.*

"The use of cinders or coke appears to be questionable owing to the chemical or electrolytic effect on copper or iron. Charcoal or pulverised carbon (such as ends of arc-light rods) is better. **A tubular earth** consisting of a perforated steel spike driven tightly into moist ground and lengthened up to the surface, the conductor reaching to the bottom and being packed with granulated charcoal, gives as much **effective area** as a plate of larger surface, and can easily be kept moist by connecting it to the nearest rain-water pipe. The resistance of a tubular earth on this plan should be very low and practically constant."—*Lightning Research Committee, 1905.*

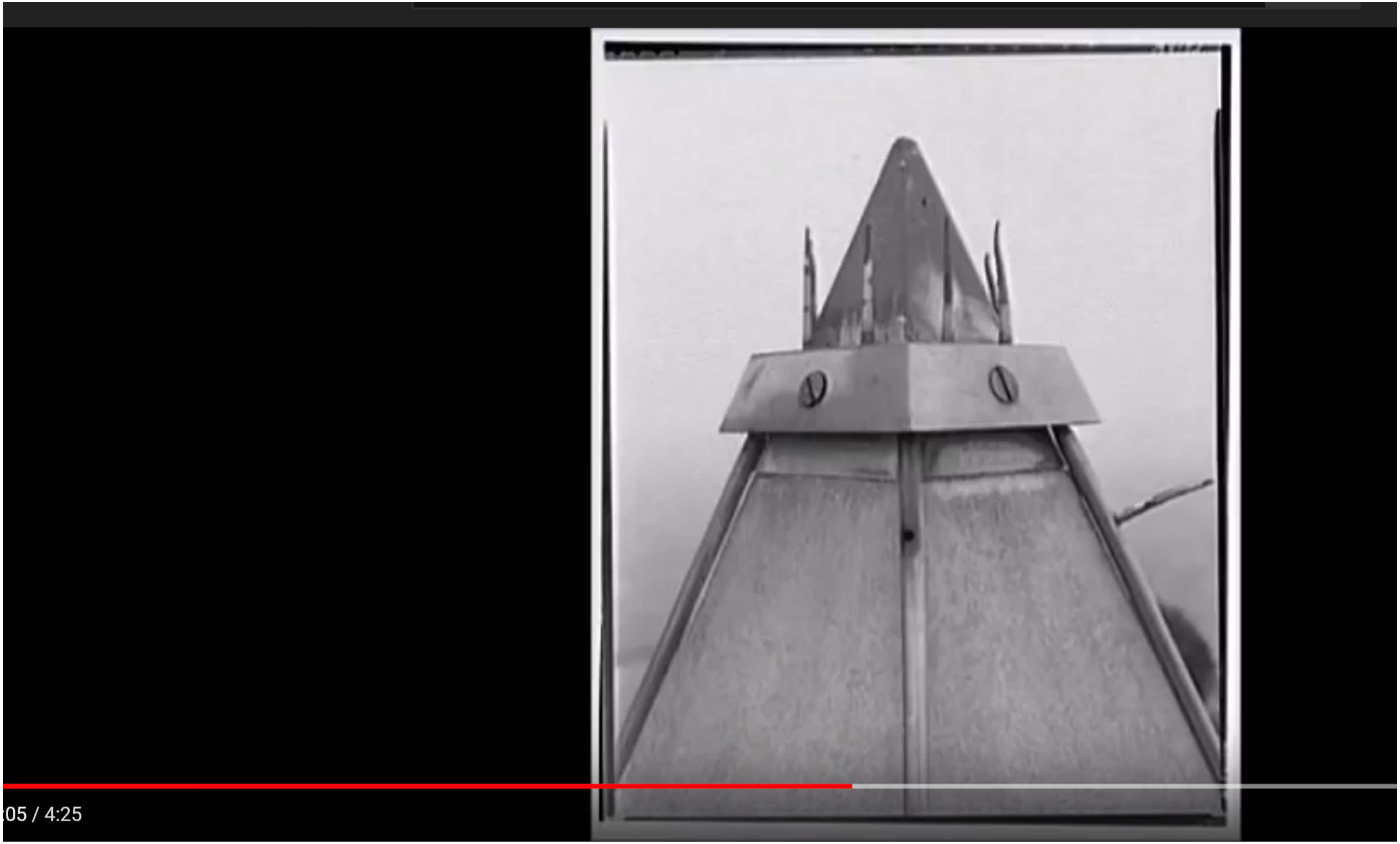

Washington Monument is an Atmospheric Electricity Mast

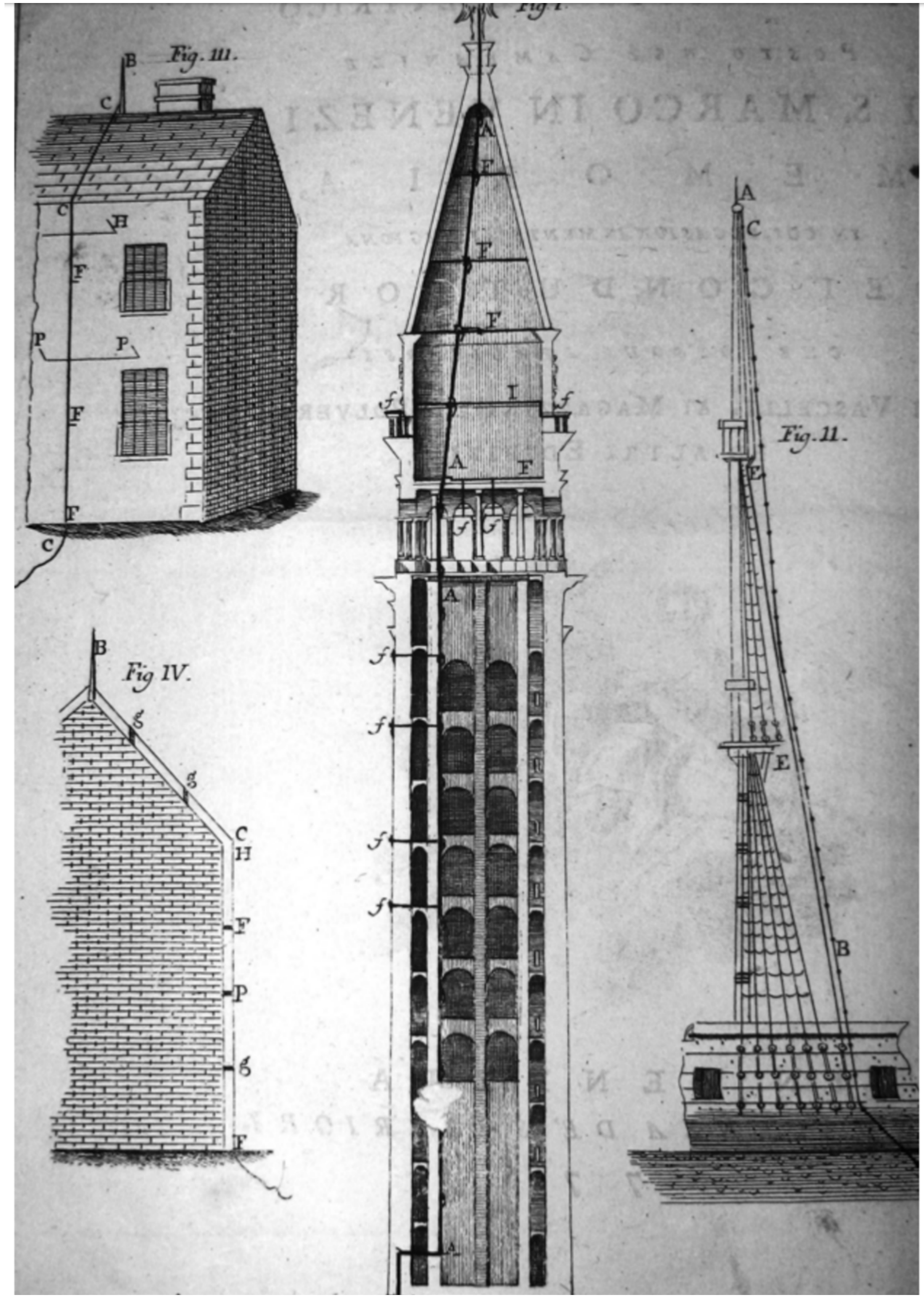
Fig. I.
Fig. III.
Fig. IV.
Fig. II.
A
B
C
E
F
H
I
P
f
g

Figure 2.3. The lightning rod that Toaldo designed for the church of San Marco in Venice. GiuseppeToaldo, “Del conduttore elettrico posto nel campanile” (Padua, 1776). Franklin Collec-

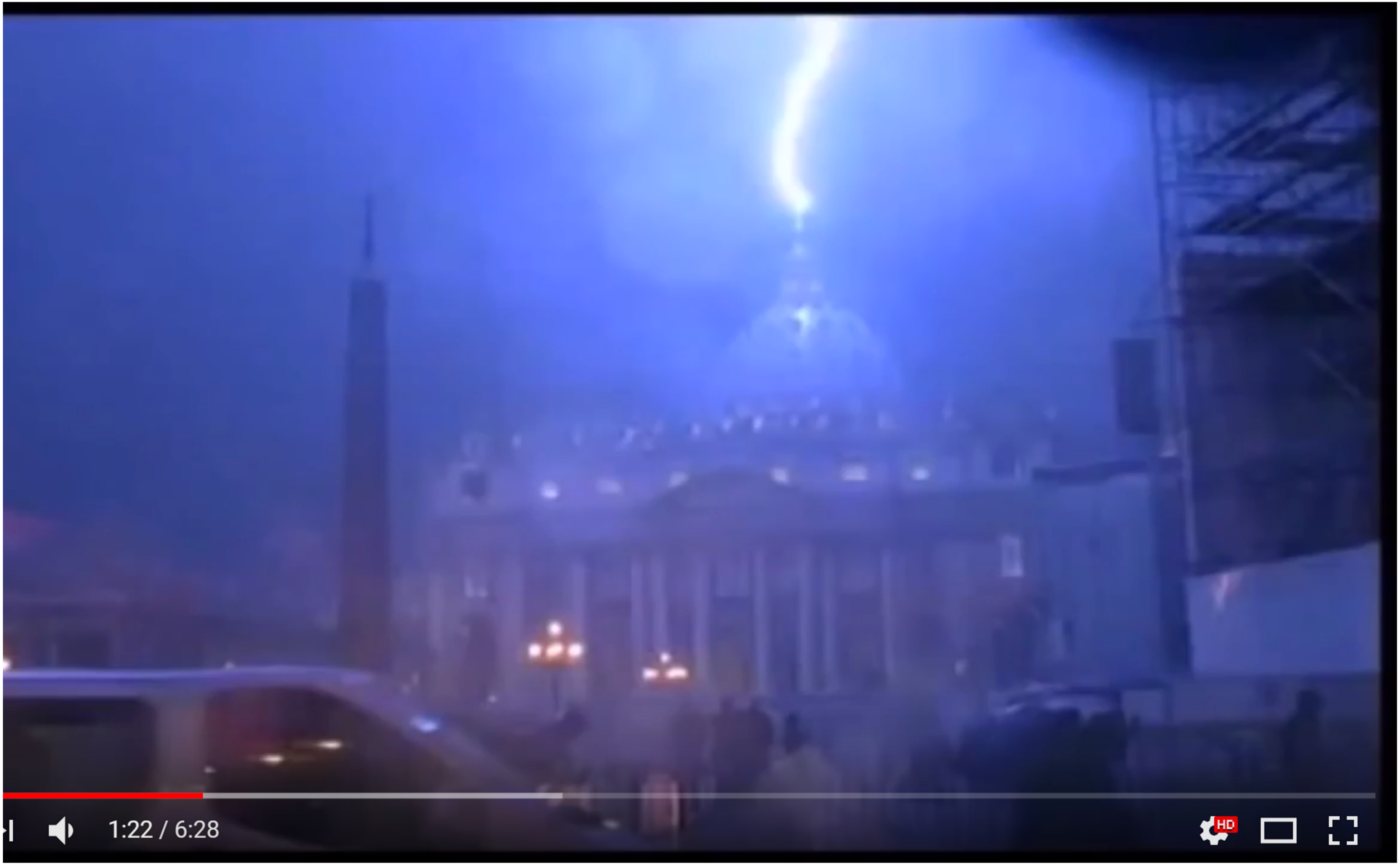

Gilded Temples Are Secretly Atmospheric Electricity Masts

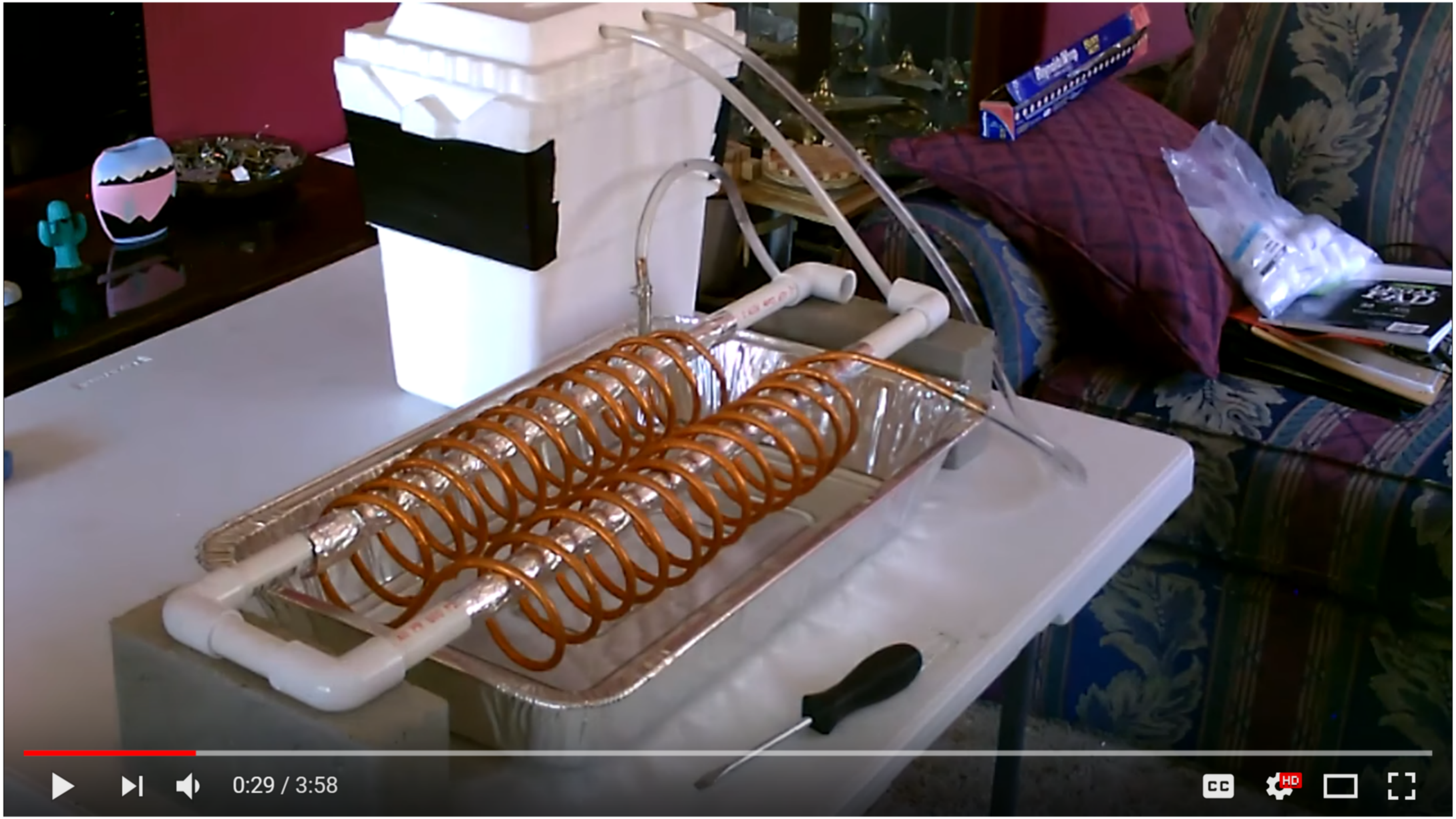
0:29 / 3:58
CC
HD

DIY Atmospheric Water Generator! - Produces/Extracts Distilled Water from the air! - DIY distiller

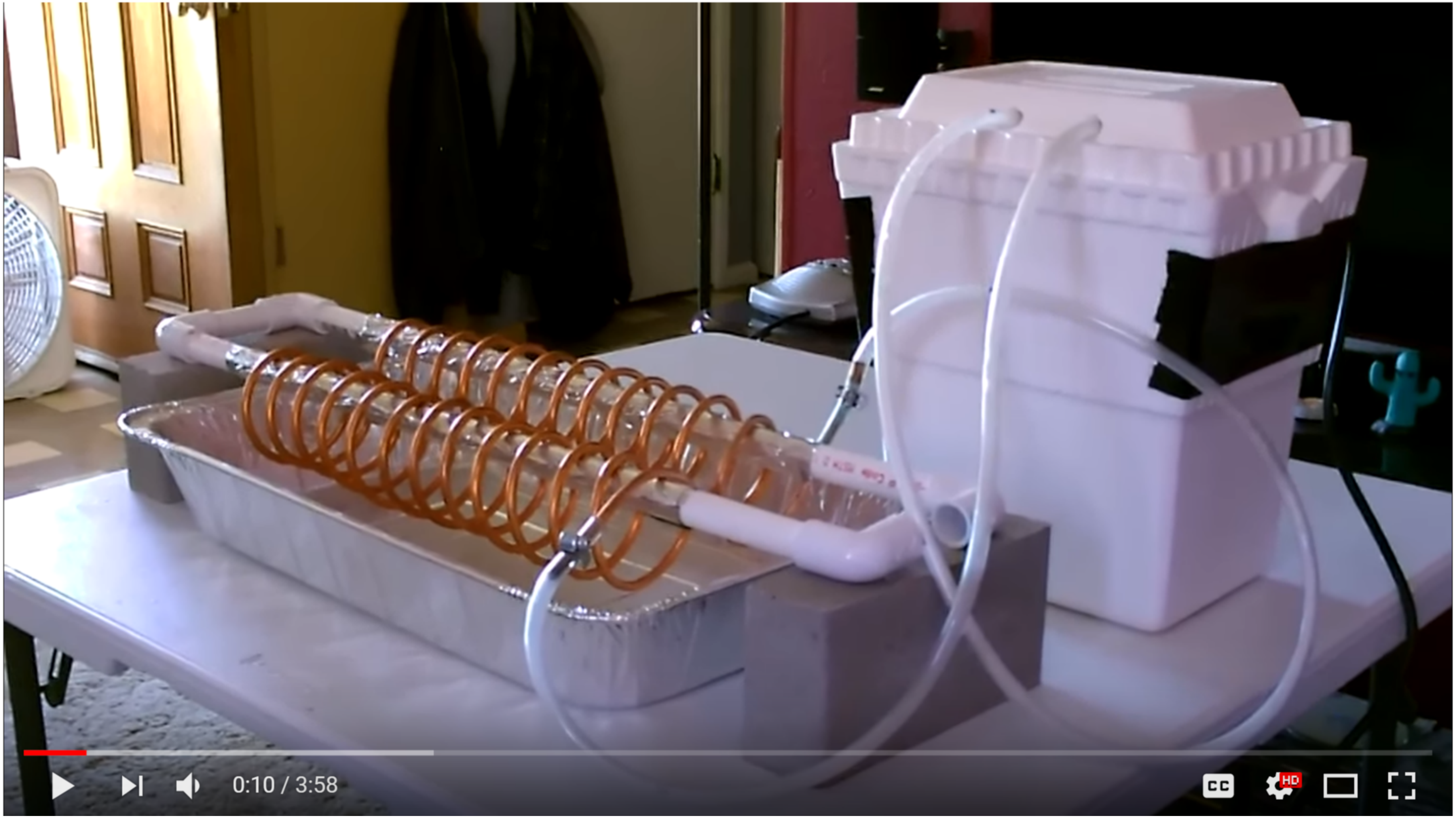
0:10 / 3:58
CC
HD

0:23 / 3:58
CC
HD

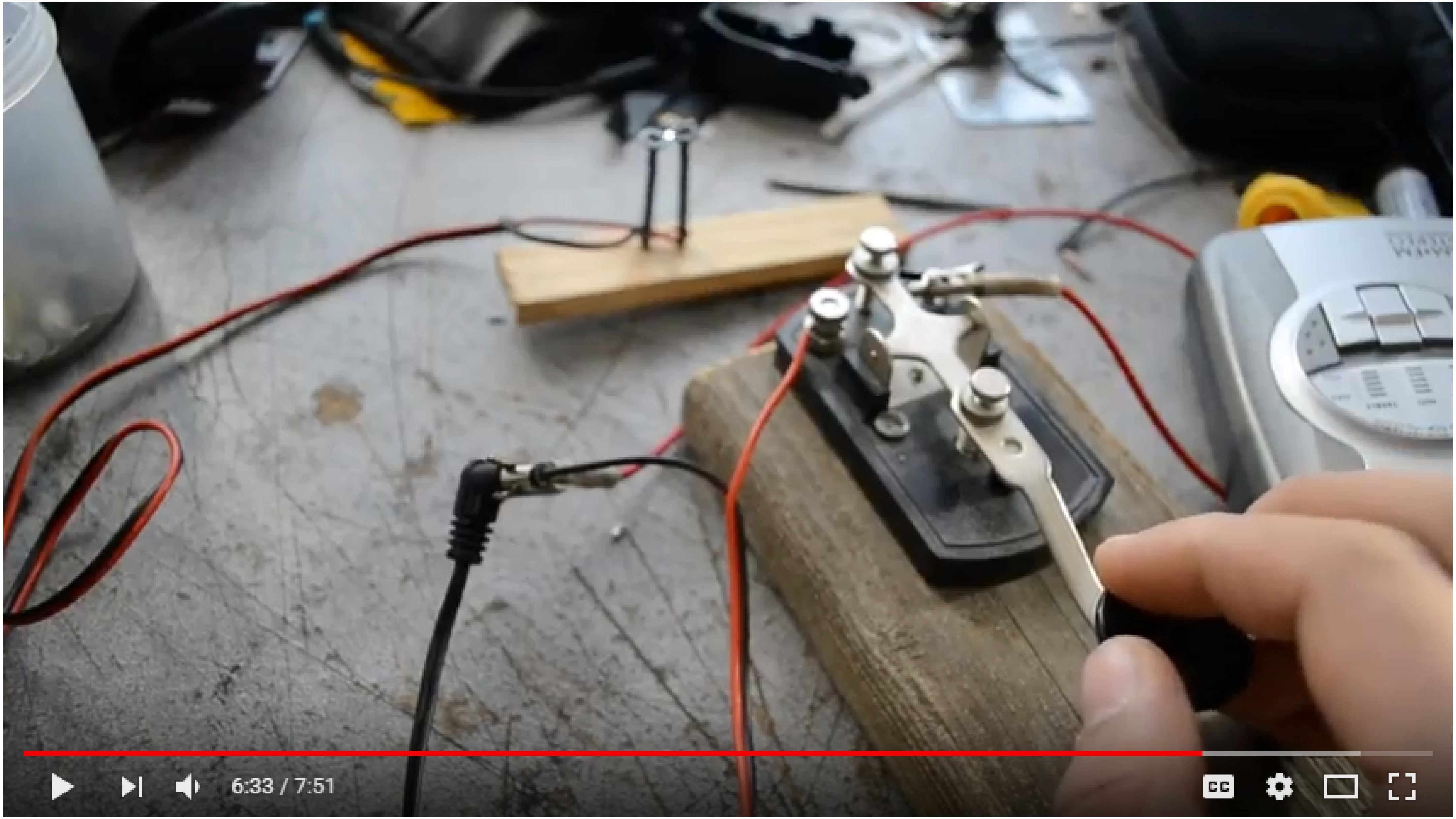

How to make a Spark Gap Transmitter

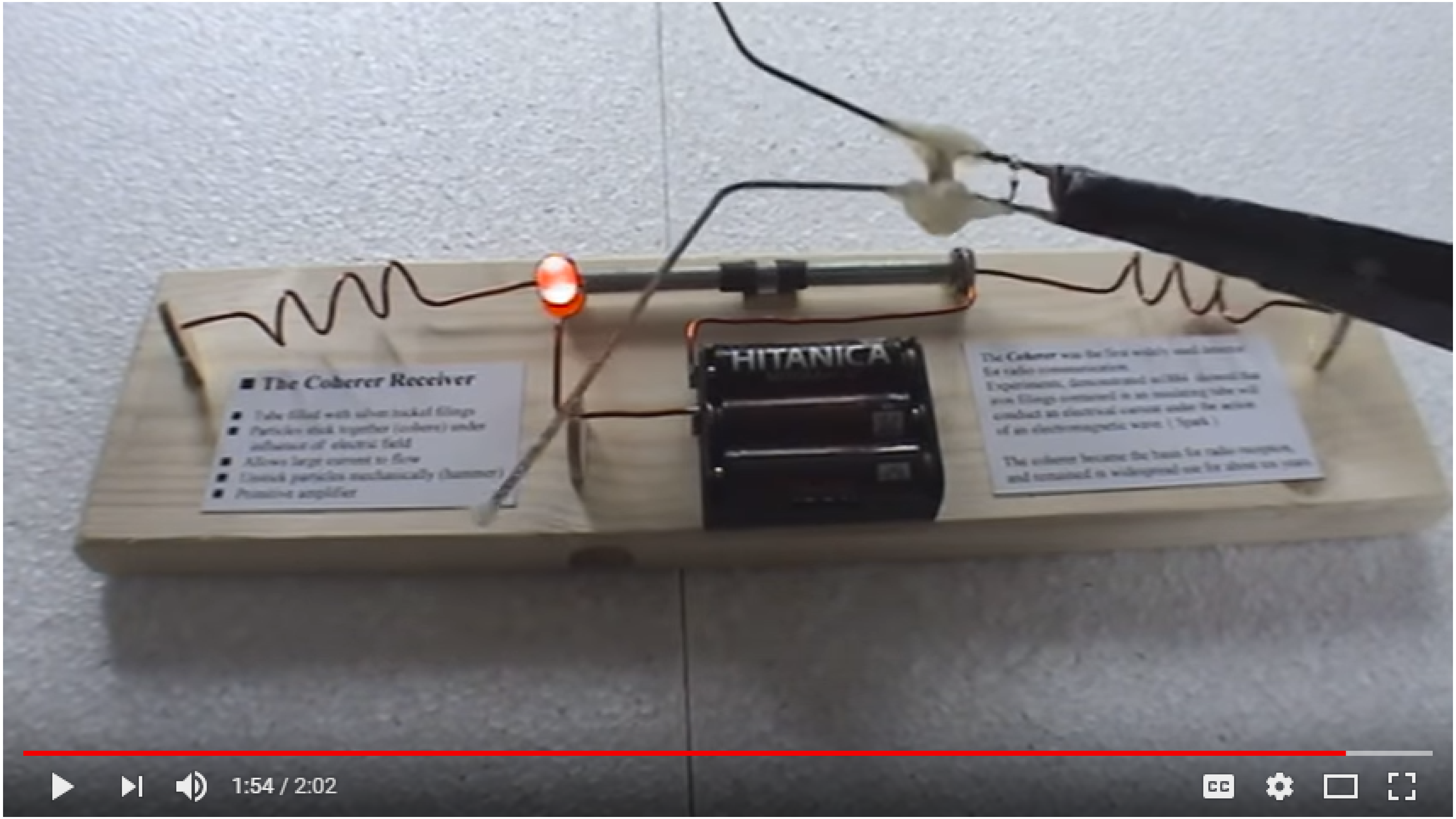
The Coherer Receiver
HITANICA
1:54 / 2:02

Vintage Model RC Coherer (Spark) Receiver

On the banner, two '**Dual**-**Loop**' directional antennas, designed specifically to collect power at 750MHz. One is connected to 100uA (700Ohms) head, and the average current you can see on it. Both antennas have 'on-board' rectifiers.

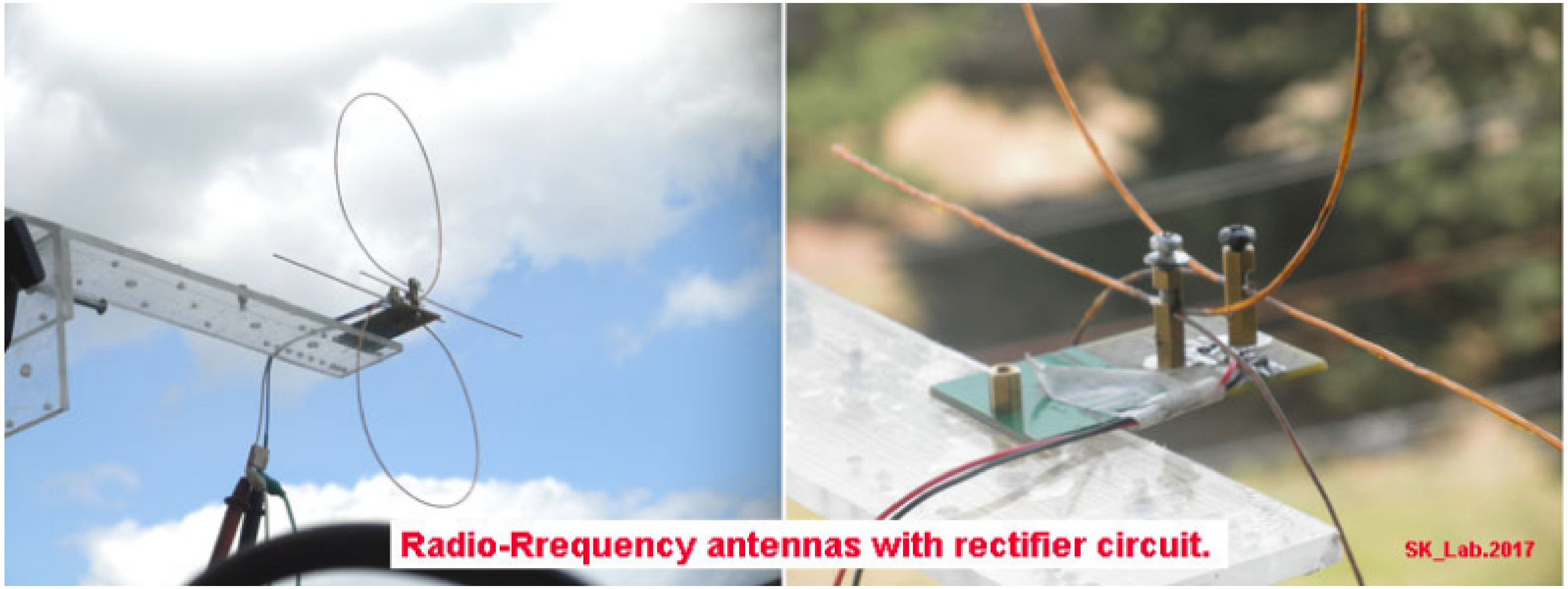
Radio-Rrequency antennas with rectifier circuit.
SK_Lab.2017

Reality is not as good as we would like. One antenna produces about 1 volt (at 10MOhms), and a current of about 0,4 mA (at 700Ohms).

The circuit is simple as a crystal receiver. Diodes with capacitors work as a voltage doubler (rectifier). The remaining inductors and capacitors isolate the output from high frequency.

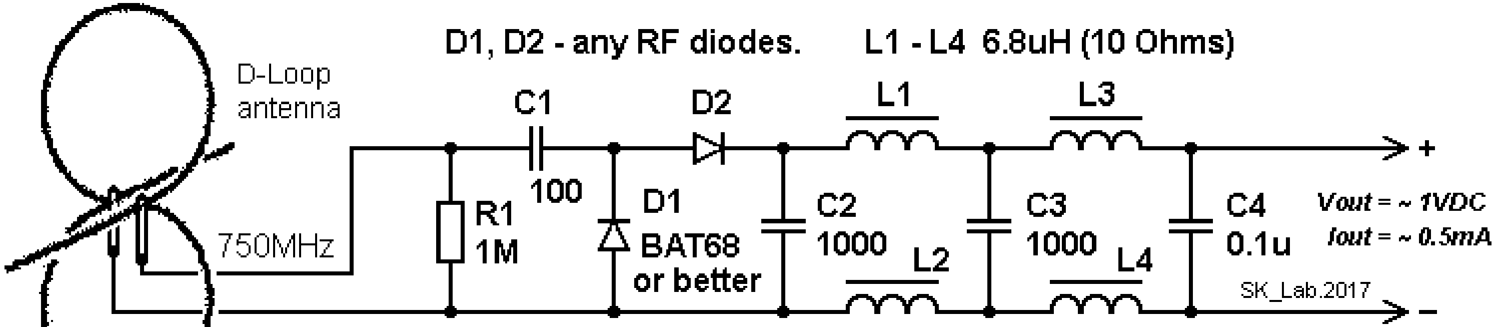
D-Loop
antenna
750MHz
D1, D2 - any RF diodes.
L1 - L4 6.8uH (10 Ohms)
C1
100
R1
1M
D2
D1
BAT68
or better
L1
L3
C2
1000
L2
C3
1000
L4
C4
0.1u
+
−
Vout = ~ 1VDC
Iout = ~ 0.5mA
SK_Lab.2017

* The output voltage and current may vary, depending on the conditions.

Harvesting RF power. Absorbs energy from cell towers at 750MHz.

Any RF diodes with junction capacitance less then 1.5pF will be suitable for this application. See the end of this page for a suitable choice.

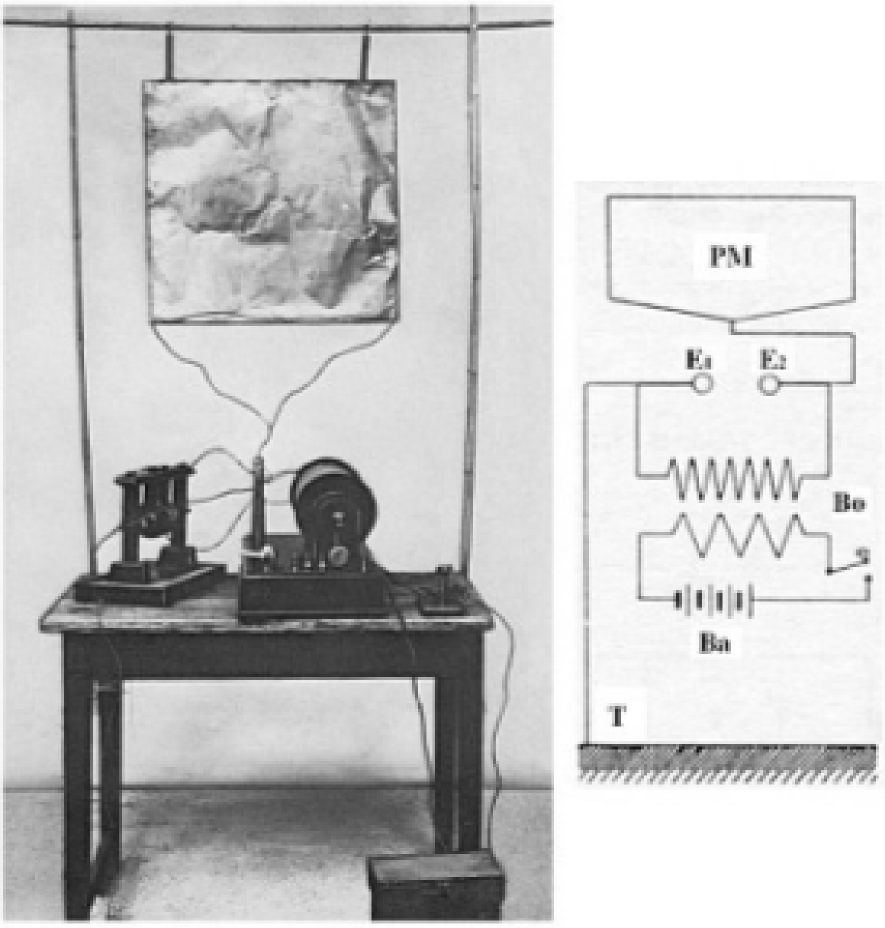
PM
E1
E2
Bo
Ba
T

Transmisor Marconi, c. 1895.

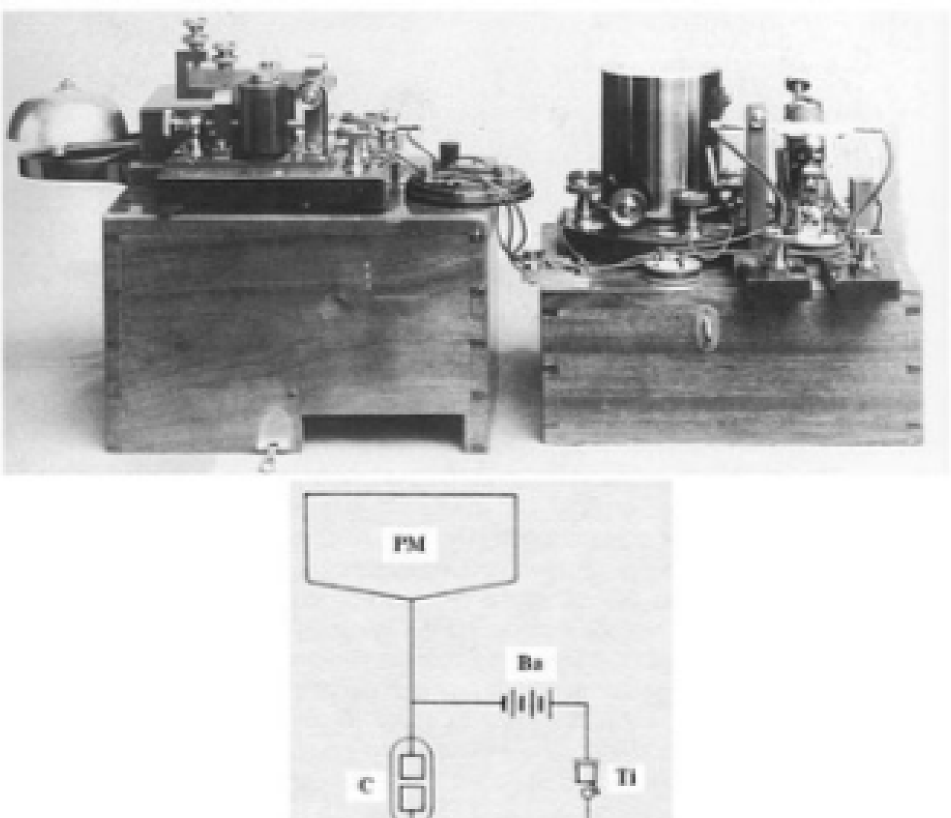

Receptor Marconi, c. 1895.

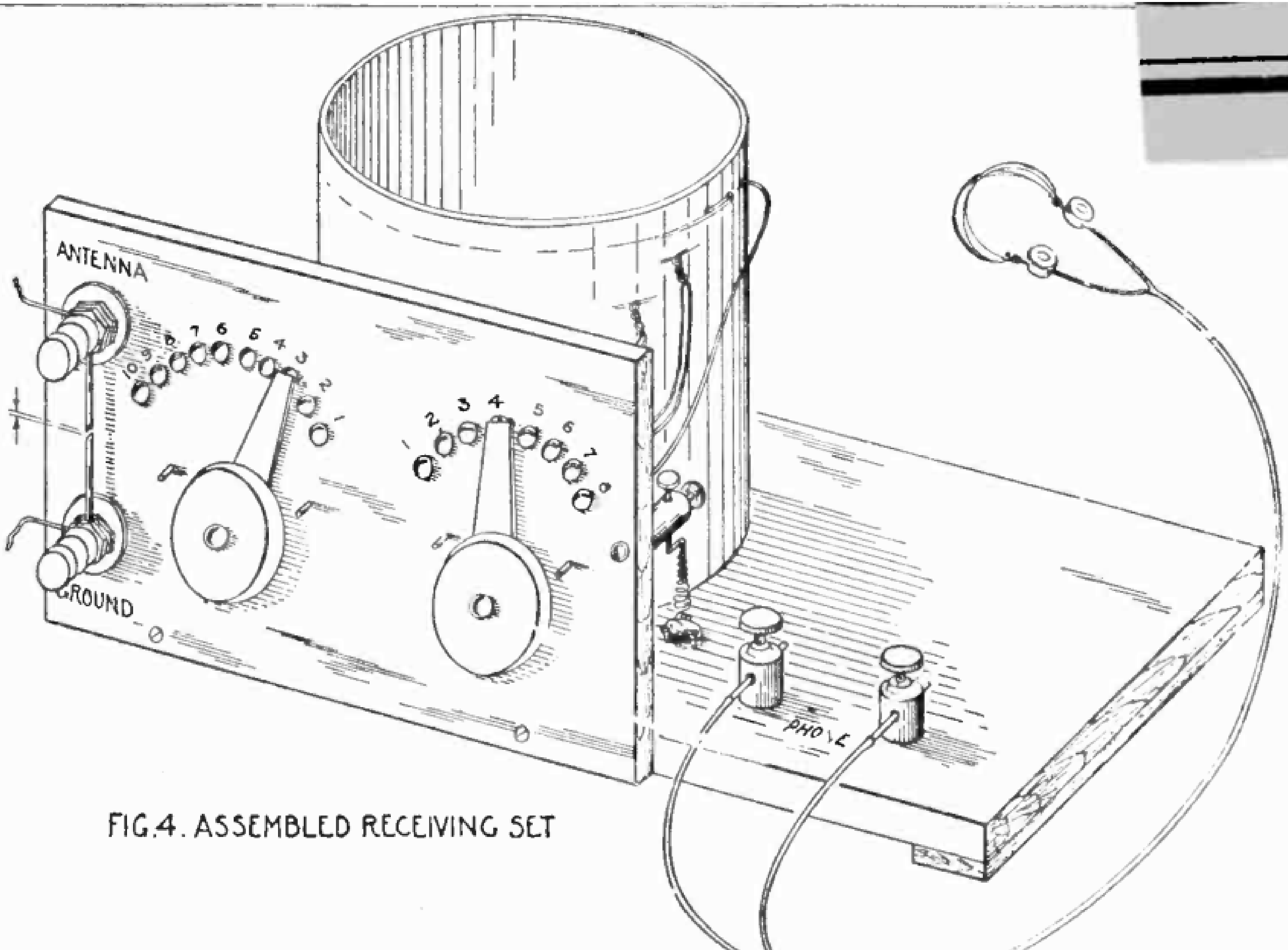
ANTENNA
10
9
8
7
6
5
4
3
2
1
1/16 IN
GROUND
1
2
3
4
5
6
7
8
PHONE

FIG. 4. ASSEMBLED RECEIVING SET

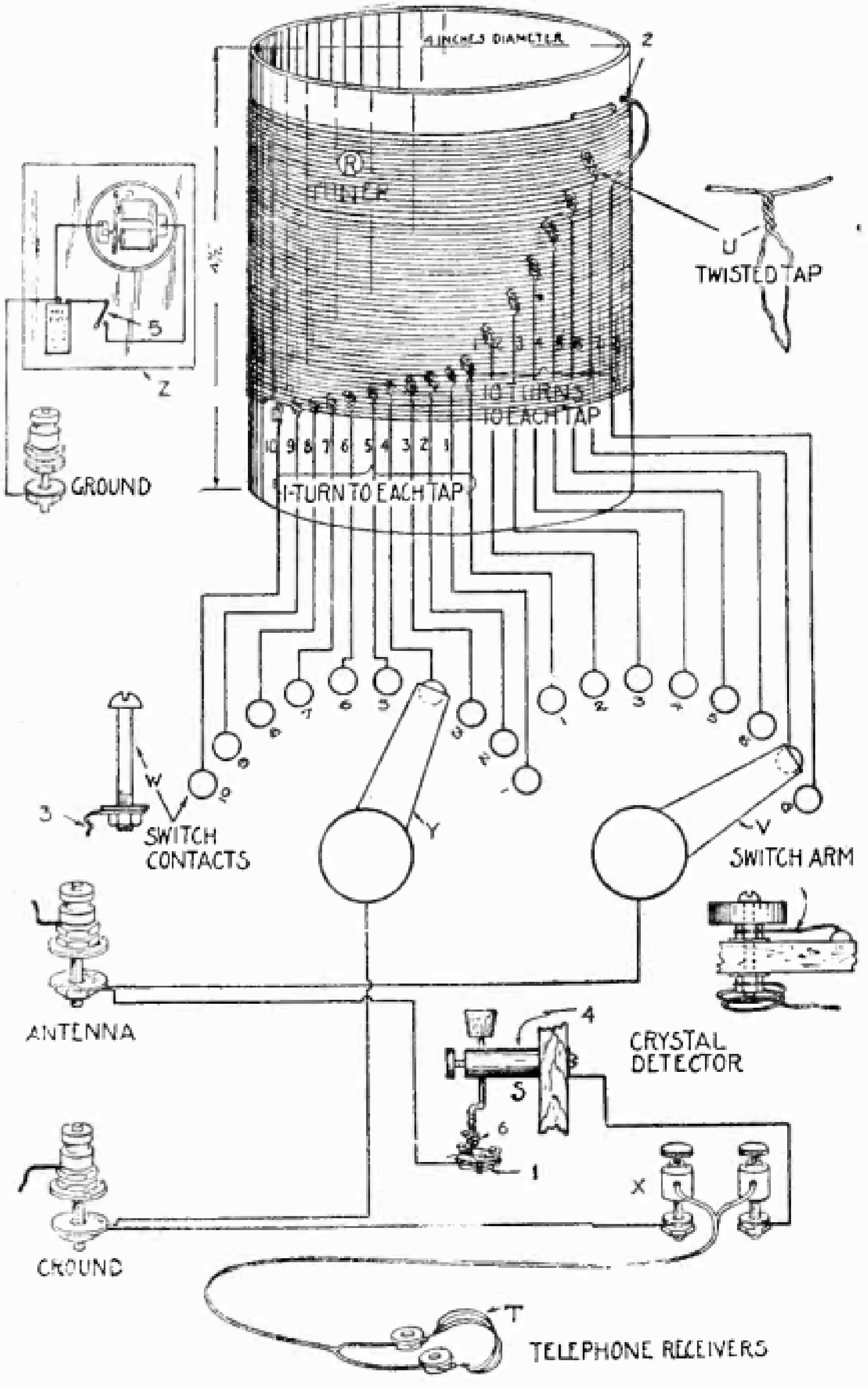
4 INCHES DIAMETER
Z
R
TUNER
4½
U
TWISTED TAP
5
Z
GROUND
10 TURNS TO EACH TAP
10 9 8 7 6 5 4 3 2 1
1-TURN TO EACH TAP
W
3
SWITCH CONTACTS
Y
V
SWITCH ARM
ANTENNA
4
CRYSTAL DETECTOR
S
6
1
X
GROUND
T
TELEPHONE RECEIVERS

FIG. 3 WIRING DIAGRAM AND DETAILS OF RECEIVING SET

FIG 2. RECEIVING SET, WITH ANTENNA AND GROUND CONNECTIONS

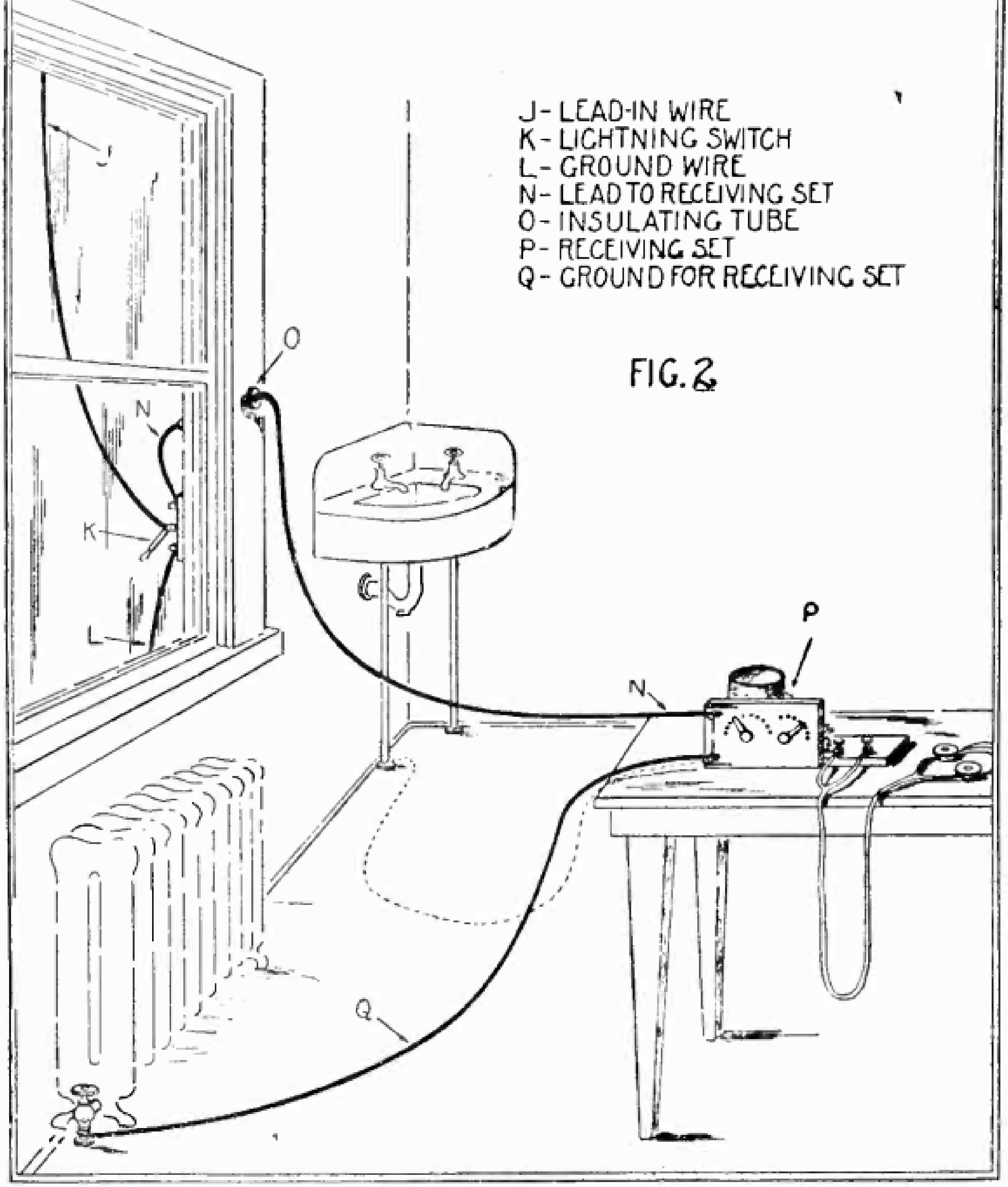

J- LEAD-IN WIRE
K- LIGHTNING SWITCH
L- GROUND WIRE
N- LEAD TO RECEIVING SET
O- INSULATING TUBE
P- RECEIVING SET
Q- GROUND FOR RECEIVING SET

FIG. 2

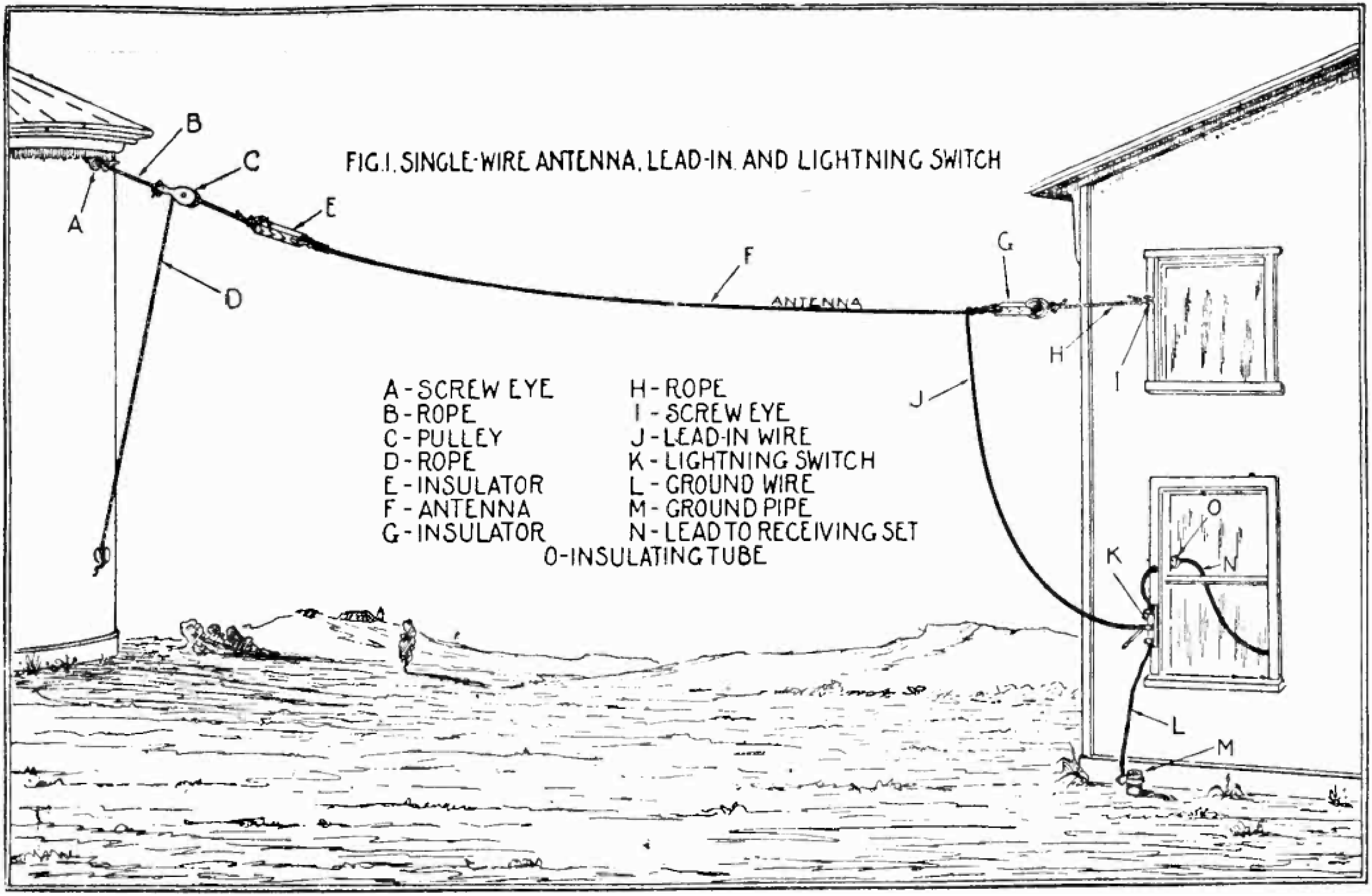

FIG. I. SINGLE-WIRE ANTENNA, LEAD-IN AND LIGHTNING SWITCH

A
B
C
D
E
F
ANTENNA
G
H
I
J
K
O
N
L
M
A - SCREW EYE
B - ROPE
C - PULLEY
D - ROPE
E - INSULATOR
F - ANTENNA
G - INSULATOR
H - ROPE
I - SCREW EYE
J - LEAD-IN WIRE
K - LIGHTNING SWITCH
L - GROUND WIRE
M - GROUND PIPE
N - LEAD TO RECEIVING SET
O - INSULATING TUBE

Fig. 7

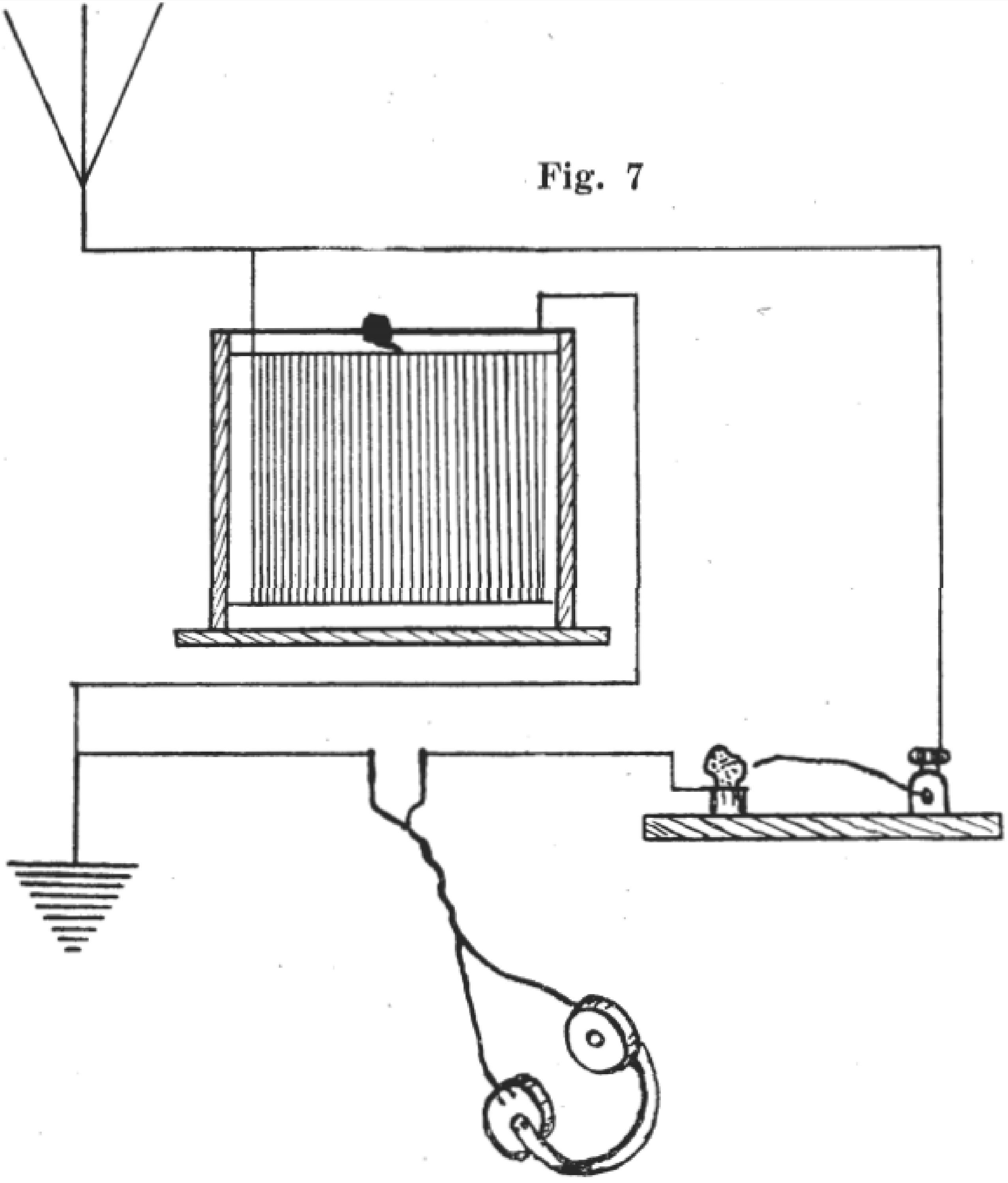

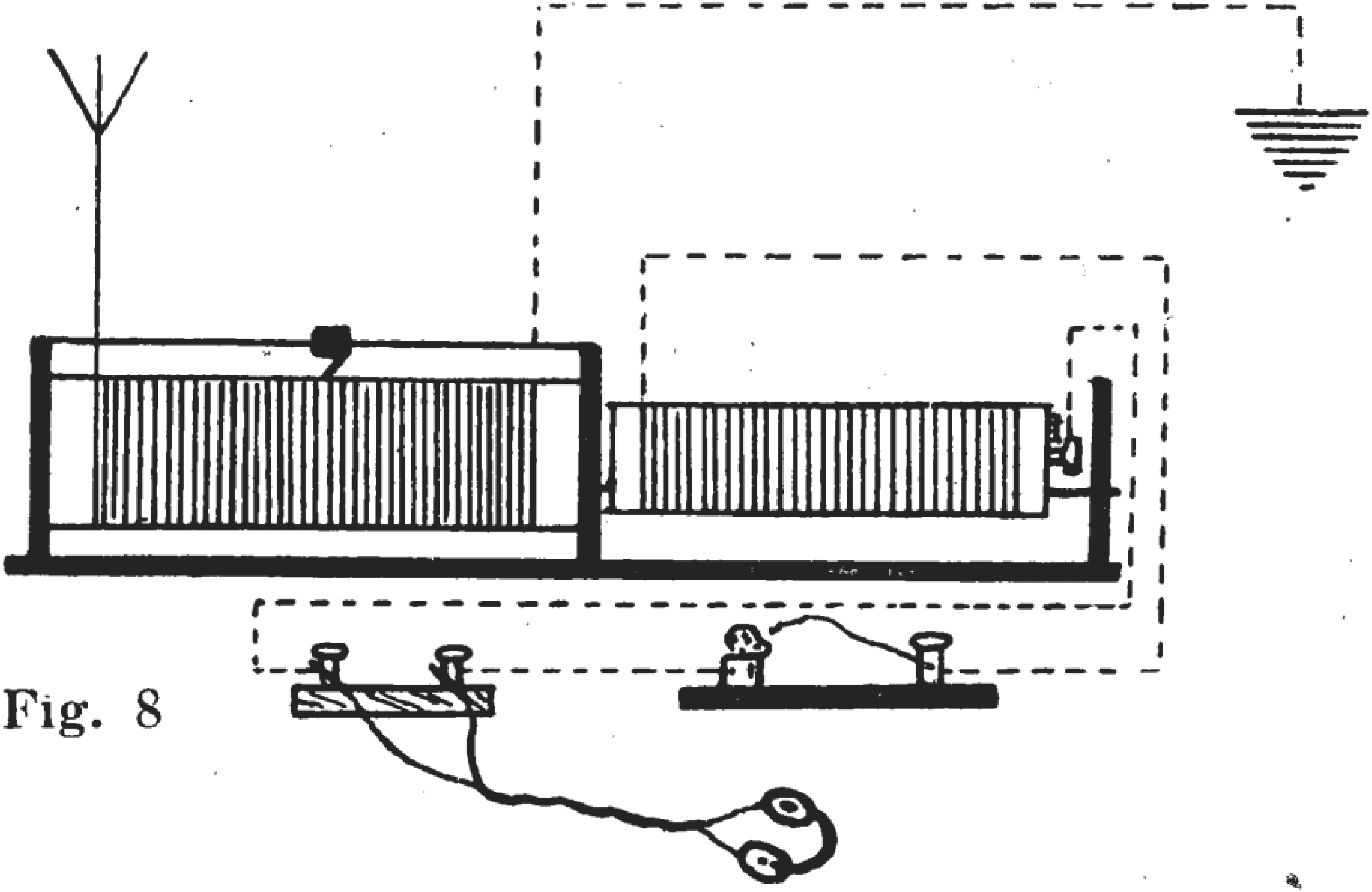

Fig. 8

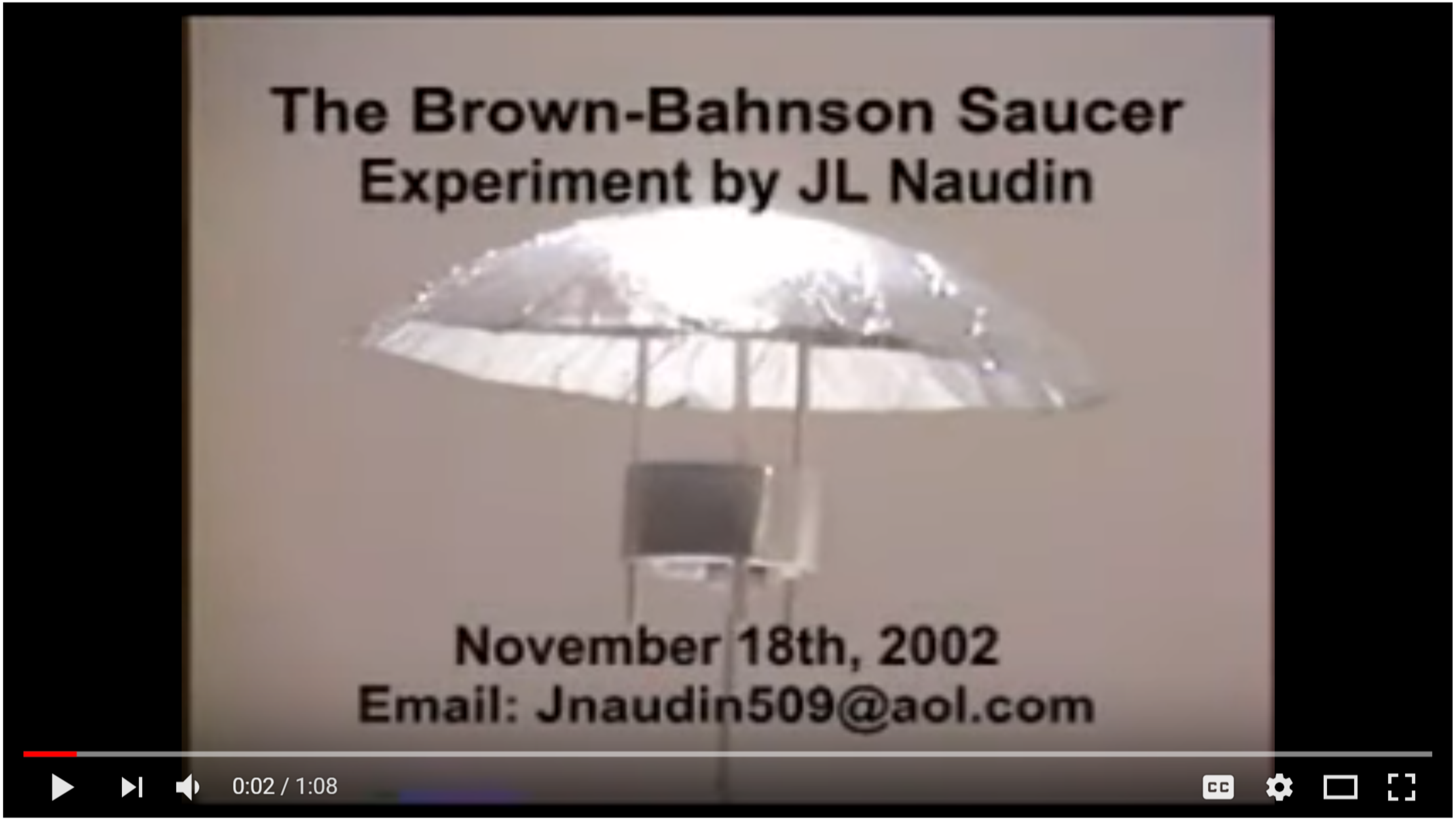
The Brown-Bahnson Saucer
Experiment by JL Naudin

November 18th, 2002
Email: Jnaudin509@aol.com

Re: LIFTER TECHNOLOGY: Demonstration & Explaination

f1=24.0 Hz
Start
f2=23.0 Hz

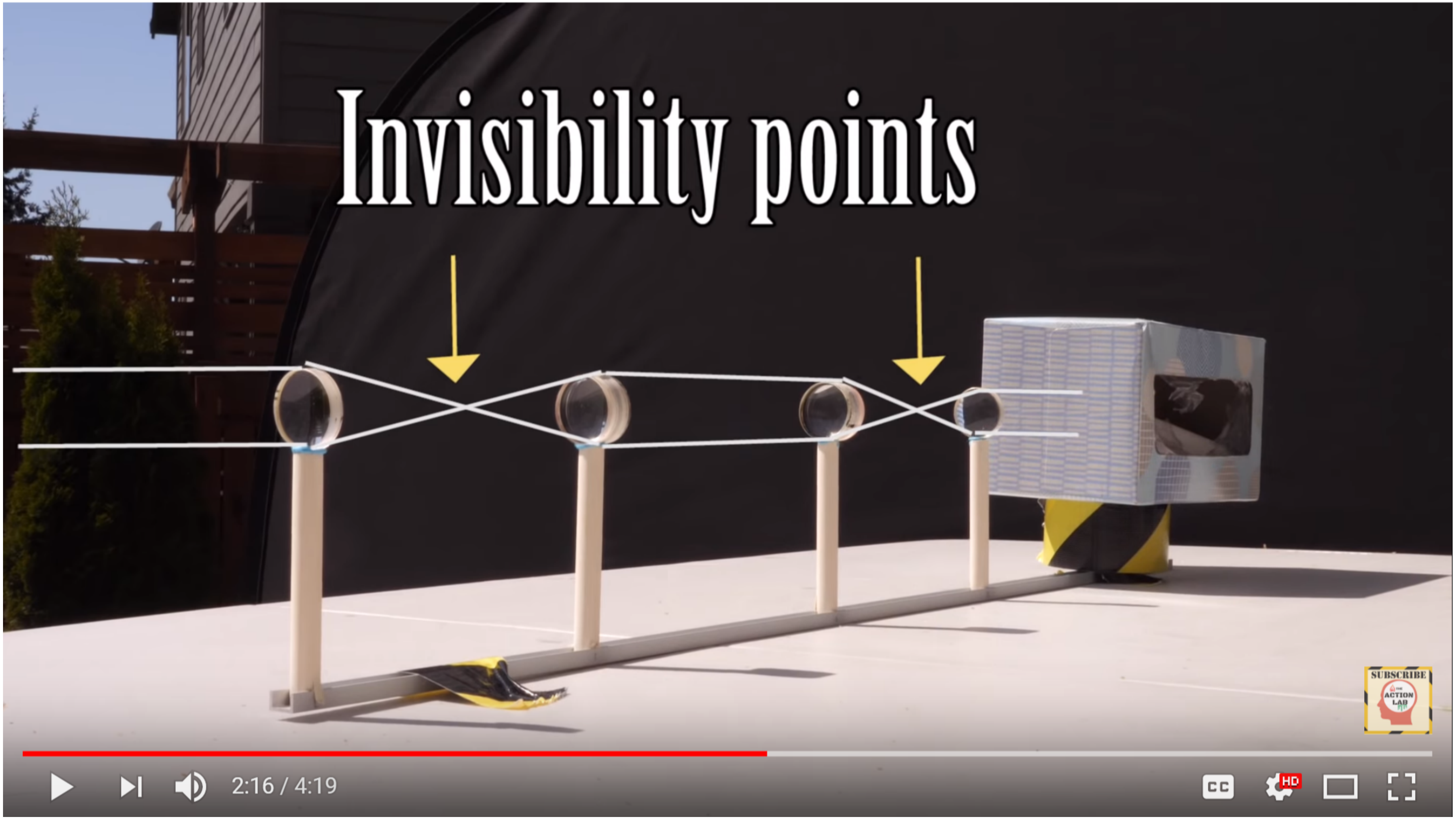
Invisibility points
SUBSCRIBE
THE ACTION LAB
2:16 / 4:19
HD

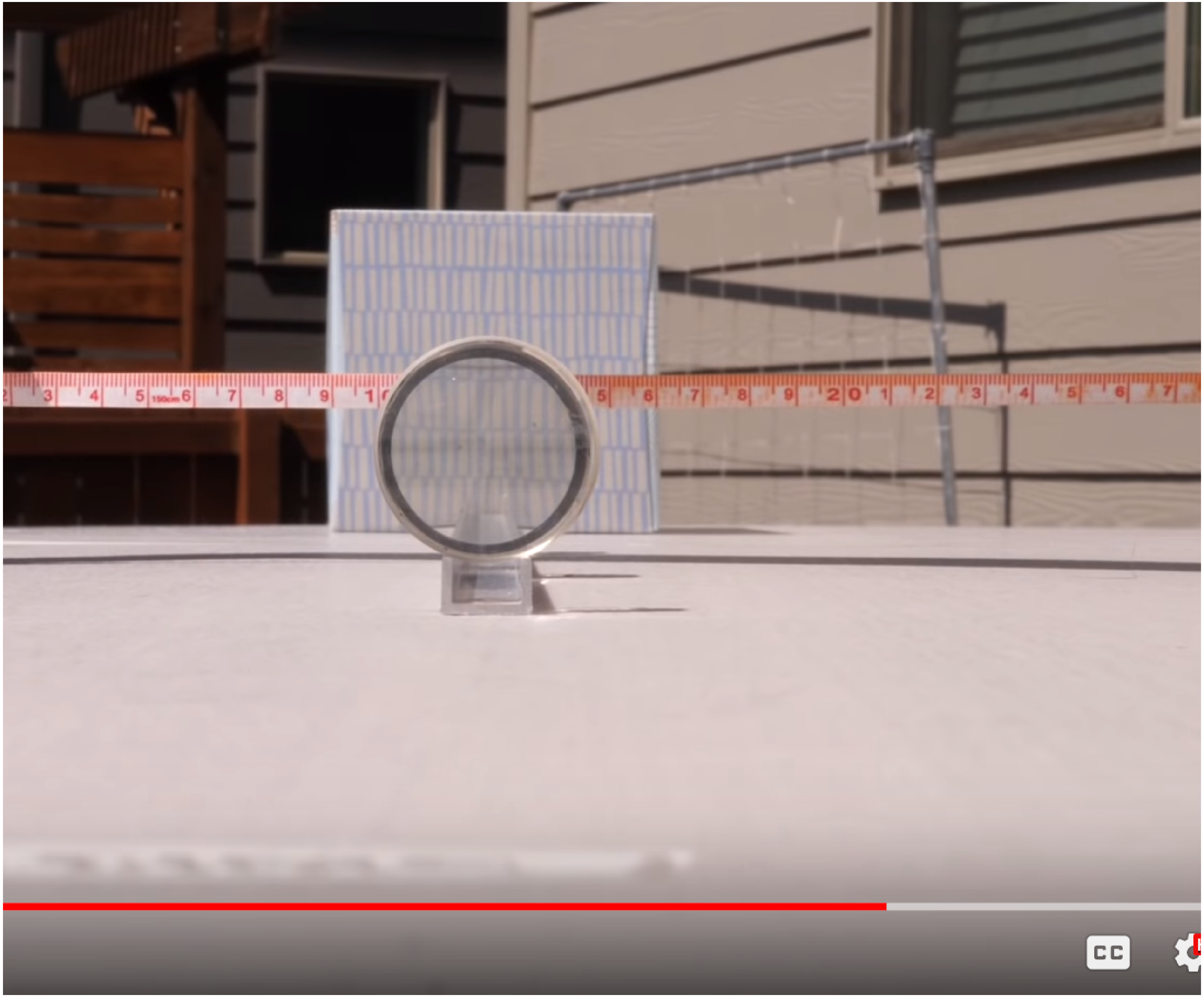
3 4 5 150cm 6 7 8 9 1
5 6 7 8 9 20 1 2 3 4 5 6 7
CC

(12) **United States Patent**
Cohen

(10) **Patent No.: US 8,253,639 B2**
(45) **Date of Patent: Aug. 28, 2012**

---

(54) **WIDEBAND ELECTROMAGNETIC CLOAKING SYSTEMS**

(76) Inventor: **Nathan Cohen**, Belmont, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 498 days.

(21) Appl. No.: **12/547,104**

(22) Filed: **Aug. 25, 2009**

(65) **Prior Publication Data**

US 2010/0156556 A1 Jun. 24, 2010

Related U.S. Application Data

(60) Provisional application No. 61/189,966, filed on Aug. 25, 2008.

(51) **Int. Cl.**
H01Q 15/02 (2006.01)
H01Q 19/06 (2006.01)

(52) **U.S. Cl.** **343/753**; 343/909

(58) **Field of Classification Search** 343/753, 343/909, 910; 333/135
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,471,224 A * 11/1995 Barkeshli 343/909
5,714,961 A * 2/1998 Kot et al. 343/769
6,473,048 B1 * 10/2002 Diaz 343/753

* cited by examiner

Primary Examiner — Michael C Wimer
(74) *Attorney, Agent, or Firm* — McDermott Will & Emery LLP

(57) **ABSTRACT**

Arrangement of resonators in an aperiodic configurations are described, which can be used for electromagnetic cloaking of objects. The overall assembly of resonators, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry corresponds to a moderate or high "Q" response that resonates within a specific frequency range, and that arrangement within that specific grouping of akin elements is periodic in the overall structure. The relative spacing and arrangement of groupings can be defined by self similarity and origin symmetry.

14 Claims, 4 Drawing Sheets

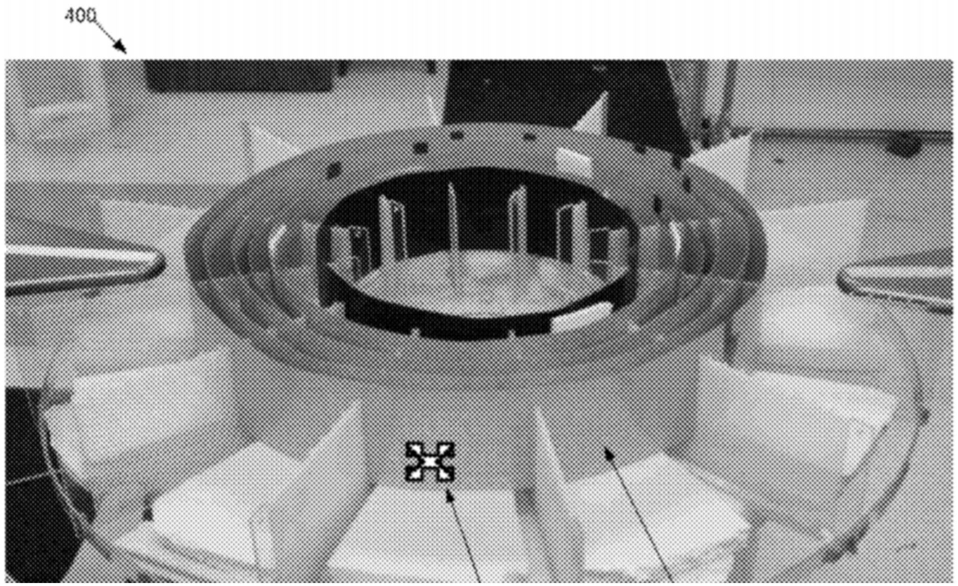

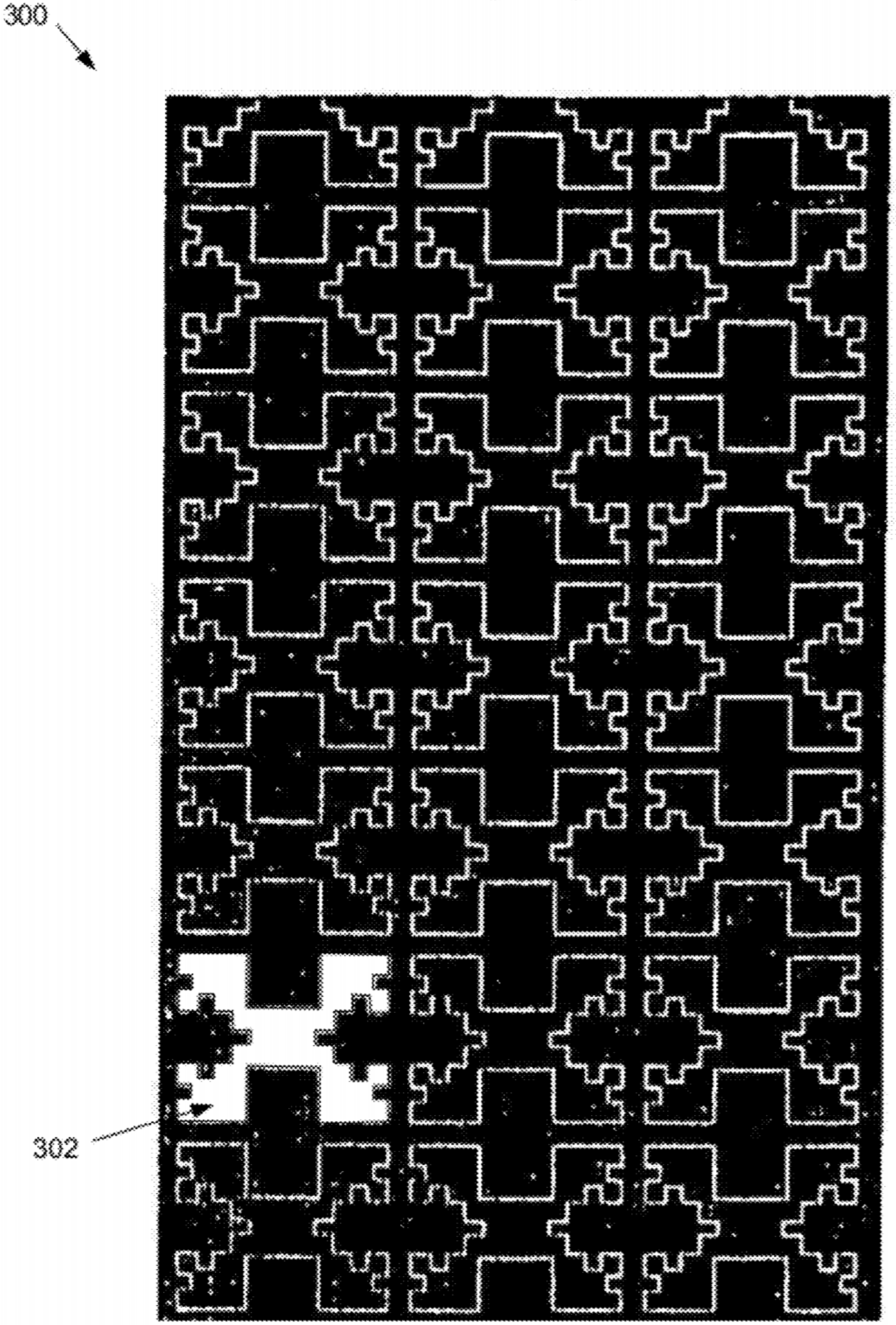
300
302

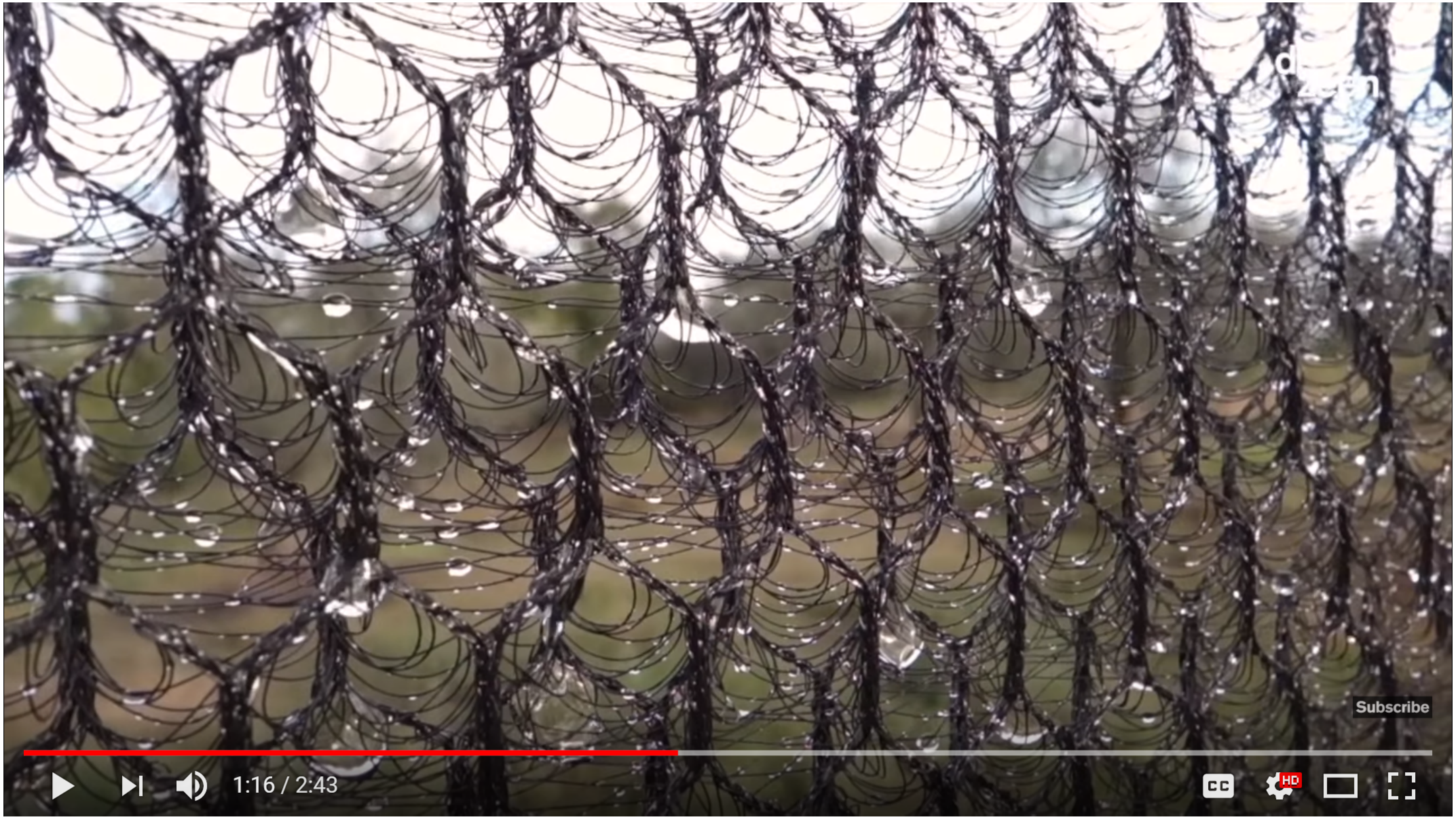
Subscribe
1:16 / 2:43
HD

Warka Water towers harvest drinkable water from the air

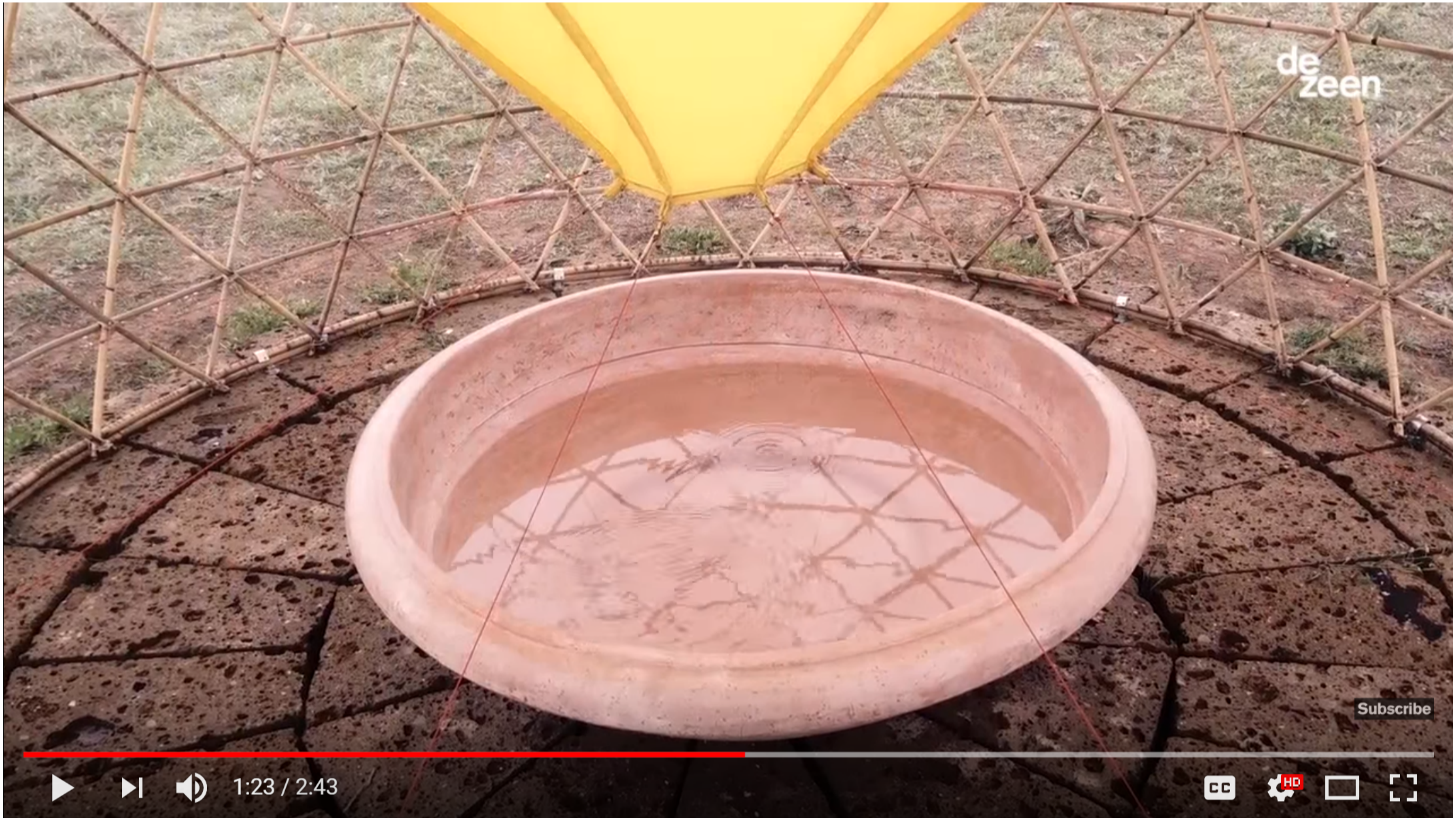
de
zeen
Subscribe
1:23 / 2:43
CC
HD

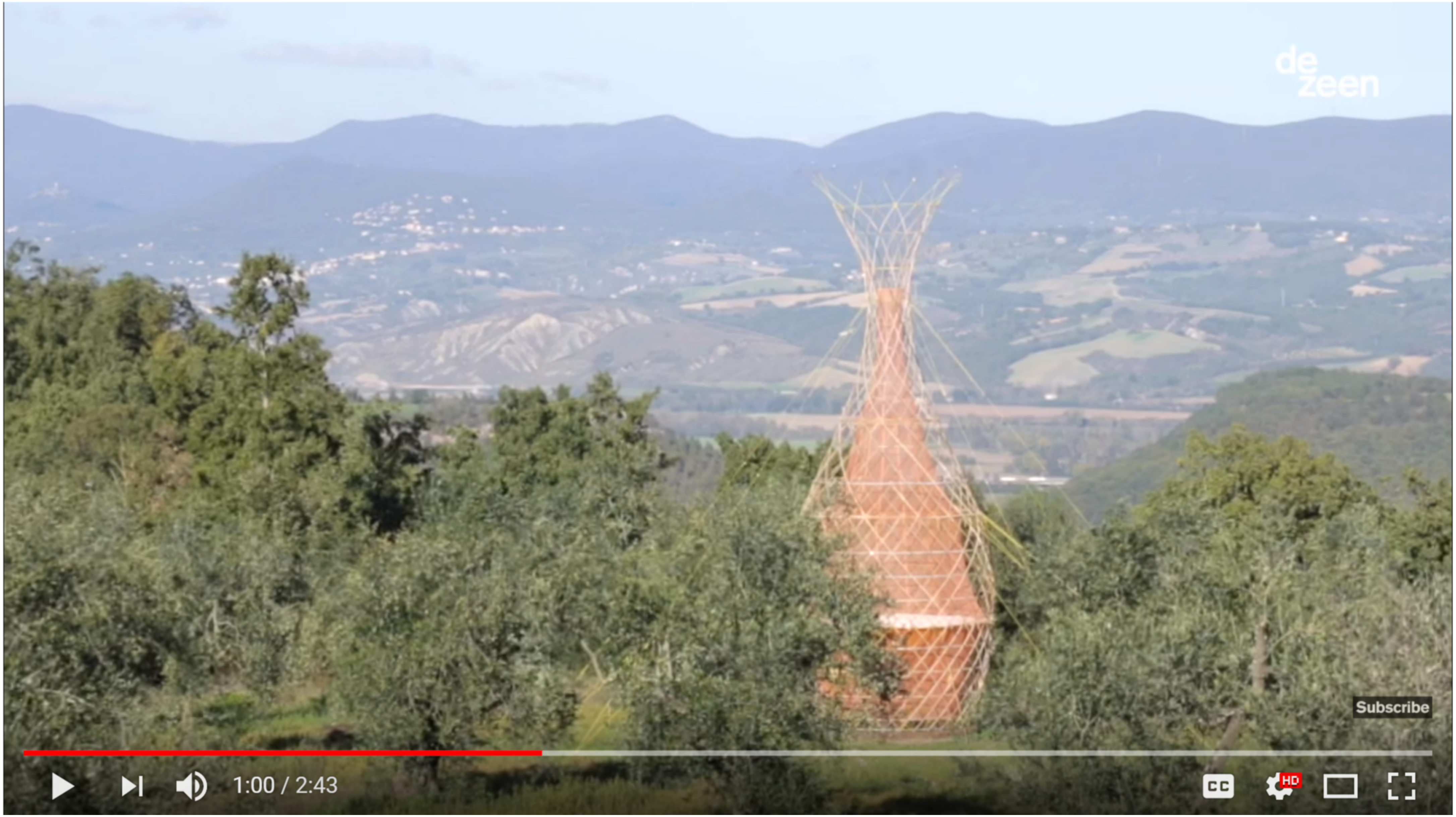
de
zeen
Subscribe
1:00 / 2:43
HD

© Azega.com

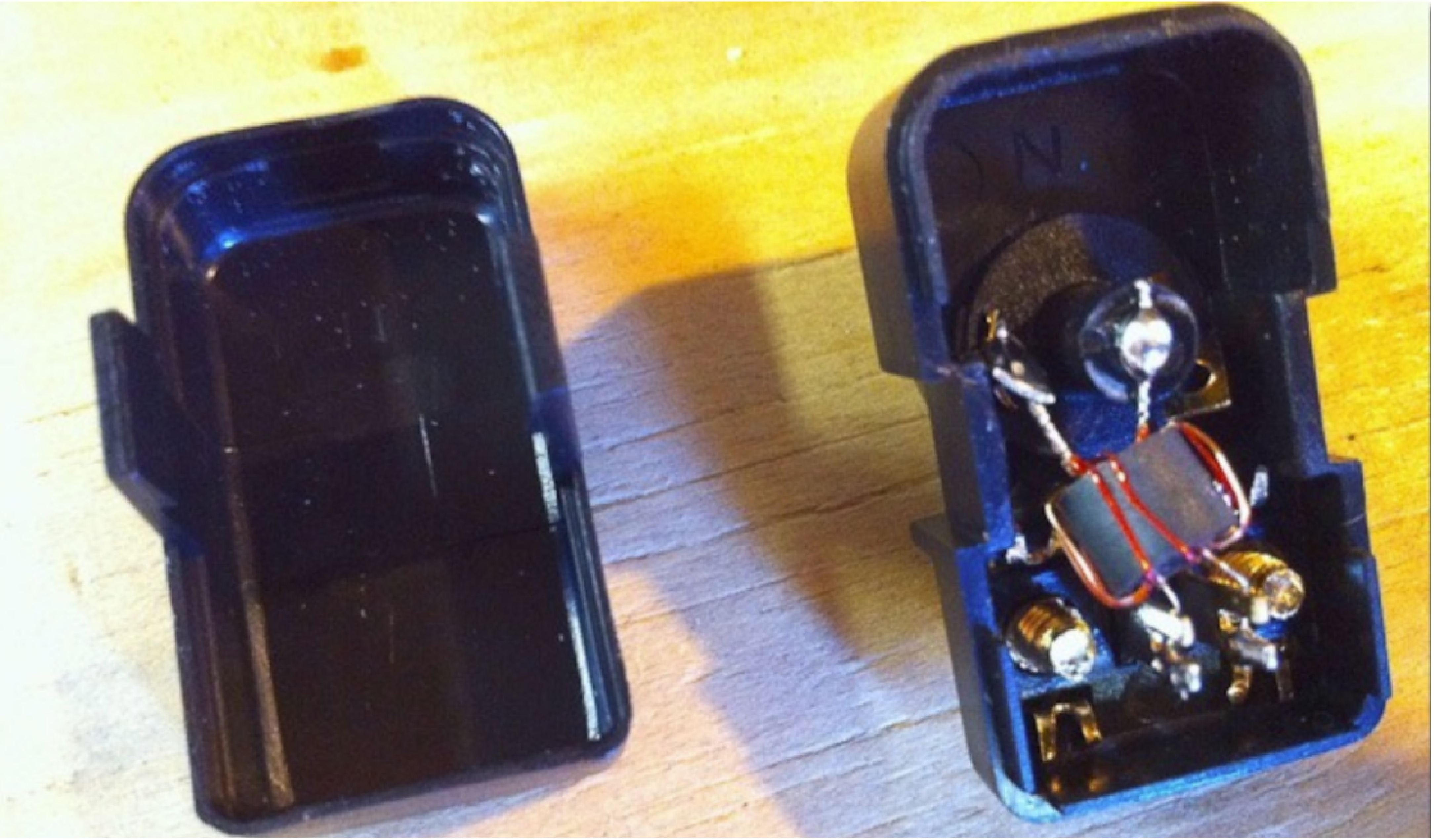
N

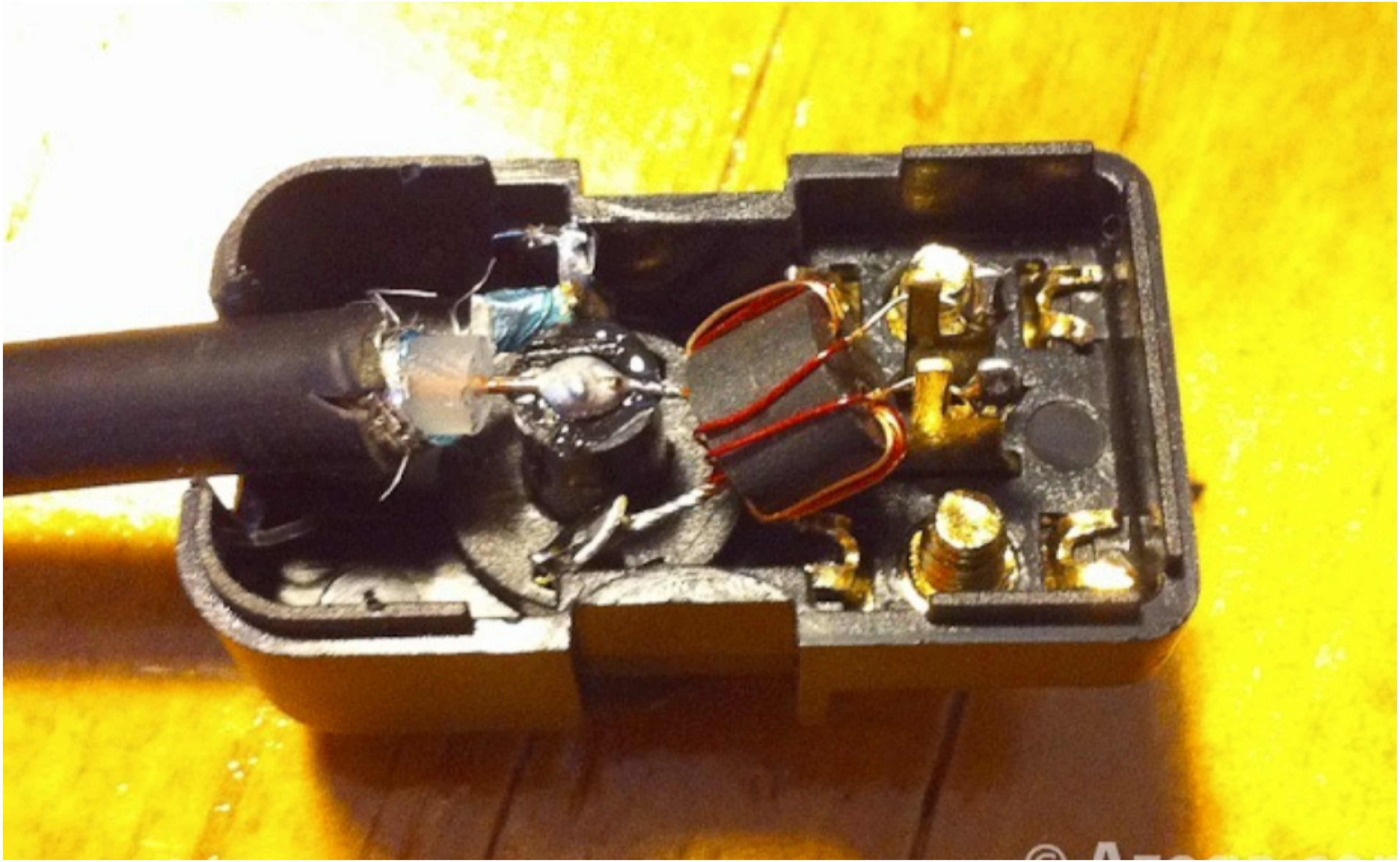

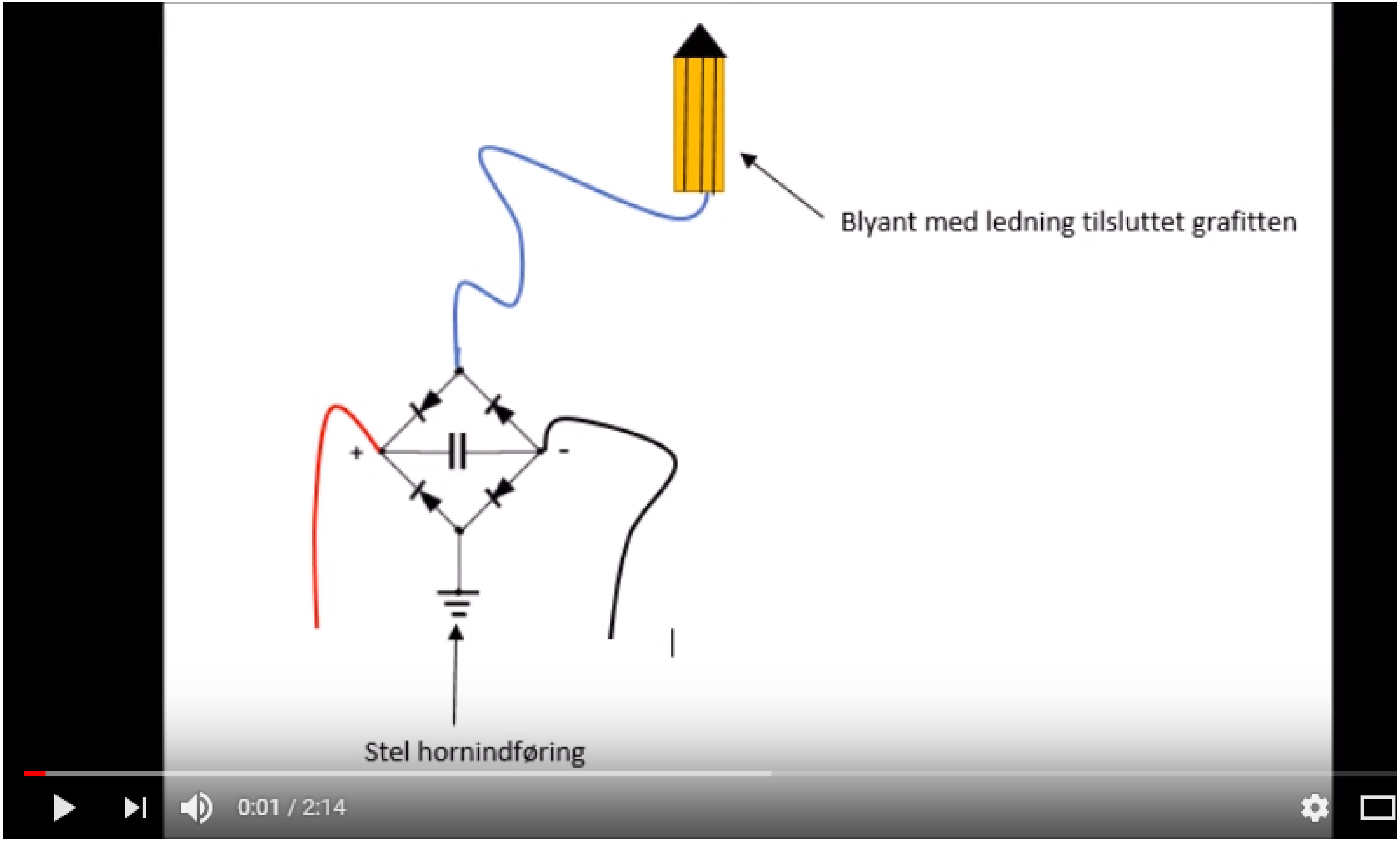

Blyant med ledning tilsluttet grafitten
+
-
Stel hornindføring
0:01 / 2:14

tesla radiant collector/ion collector with pencil

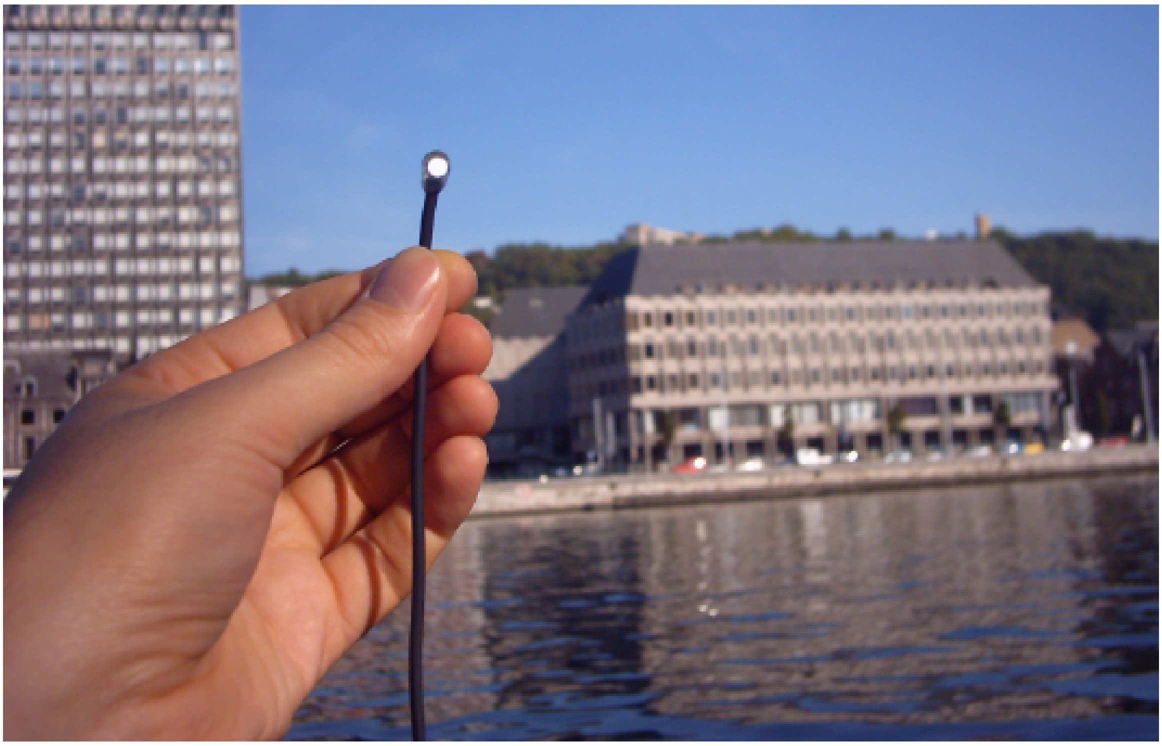

ake and it can be sent in an envelope. I made tens of them and sent them to politicians, newspapers, universities... I gave some to local people, together with a user
later version, that can be rolled up in an envelope that fits the conditions to be sent with only one stamp.

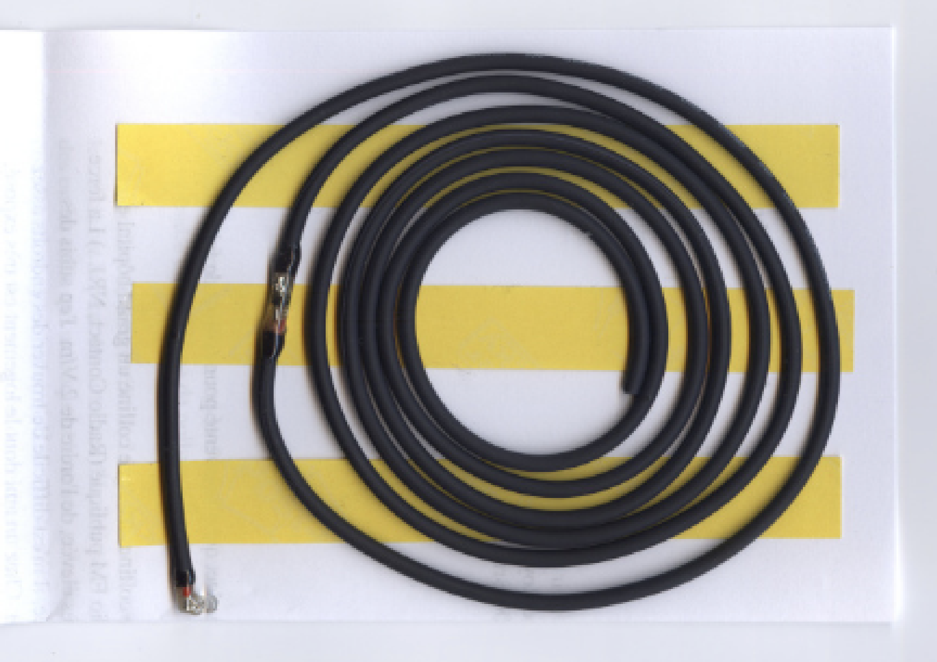

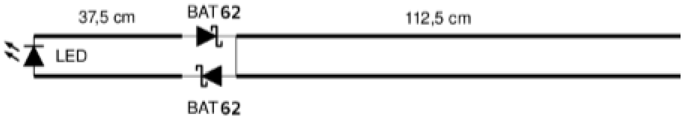

BAT62 detection diodes are no more produced. **BAT15** diodes work fine but they wear out; after a few months they stop functioning. **SMS7630** diodes are great but very little and mechanically fragile. **MMSD701T1G** diodes are sturdy and powerful; an excellent choice for a beginner. Such SMD diodes do also work for cell phone frequencies, which allows to test out a snake with a calling cell phone pushed against it. But any detection diodes that can manage 100 MHz will do.

The LED I'm currently using is the **L-7113SEC-H** . It lights up with a low tension and a very low current (the bluer a LED, the more tension it needs). Its color is red yet close to orange hence it is easily seen by the human eye (the eye is most sensitive to green, yellow and orange). The beam is quite narrow so when the LED is directed towards somebody's eyes it will appear quite bright.

For the lengths of 37.5 and 112.5 centimeters, any electric wire with two copper conductors will do. Audio signal wire is a practical solution. Use the shielding as one of the two conductors. The lengths of the two segments must not be precise. What matters is that the total length of the snake be 1.5 meters. Do not hesitate to try out if a little longer or shorter snake gives better results.

A schematic of my current probes, that I connect to a standard multimeter, measuring Volts DC. The measure displayed by the multimeter must be multiplied by 10. When using a 200.0 mV scale, just read while forgetting the dot:

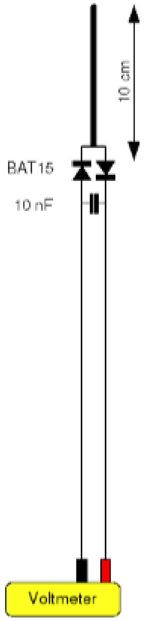

Lead Free Status	Lead Free
RoHS Status	RoHS Compliant

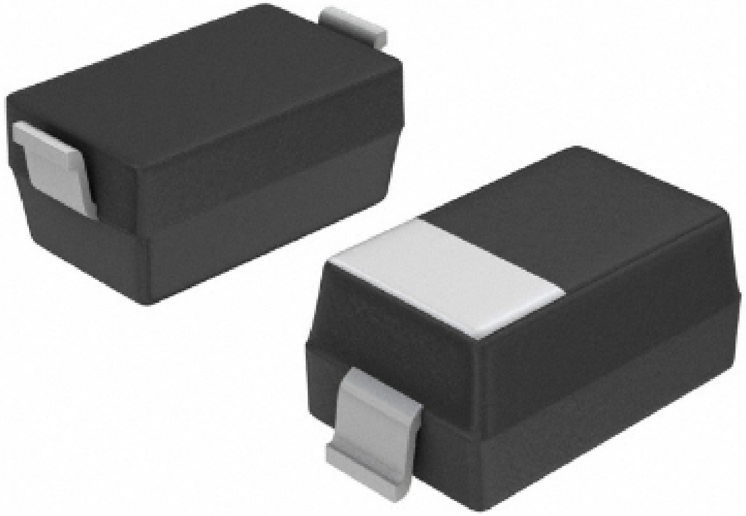

Features, Applications

The MMSD301T1, and MMSD701T1 devices are spin-offs of our popular MMBD301LT1, and MMBD701LT1 SOT-23 devices. They are designed for high-efficiency UHF and VHF detector applications. Readily available to many other fast switching RF and digital applications.

Extremely Low Minority Carrier Lifetime Very Low Capacitance Low Reverse Leakage AEC Qualified and PPAP Capable S Prefix for Automotive and Other Applications Requiring Unique Site and Control Change Requirements These Devices are Pb-Free, Halogen Free/BFR Free and are RoHS Compliant*

XXX G = Specific Device Code SMMSD701T1G = Date Code = Pb-Free Package
Rating Reverse Voltage MMSD701T1G, SMMSD701T1G Forward Current (DC) Continous Forward Power Dissipation = 25C Junction Temperature Storage Temperature Range Symbol VR Value to +150 Unit Vdc M G

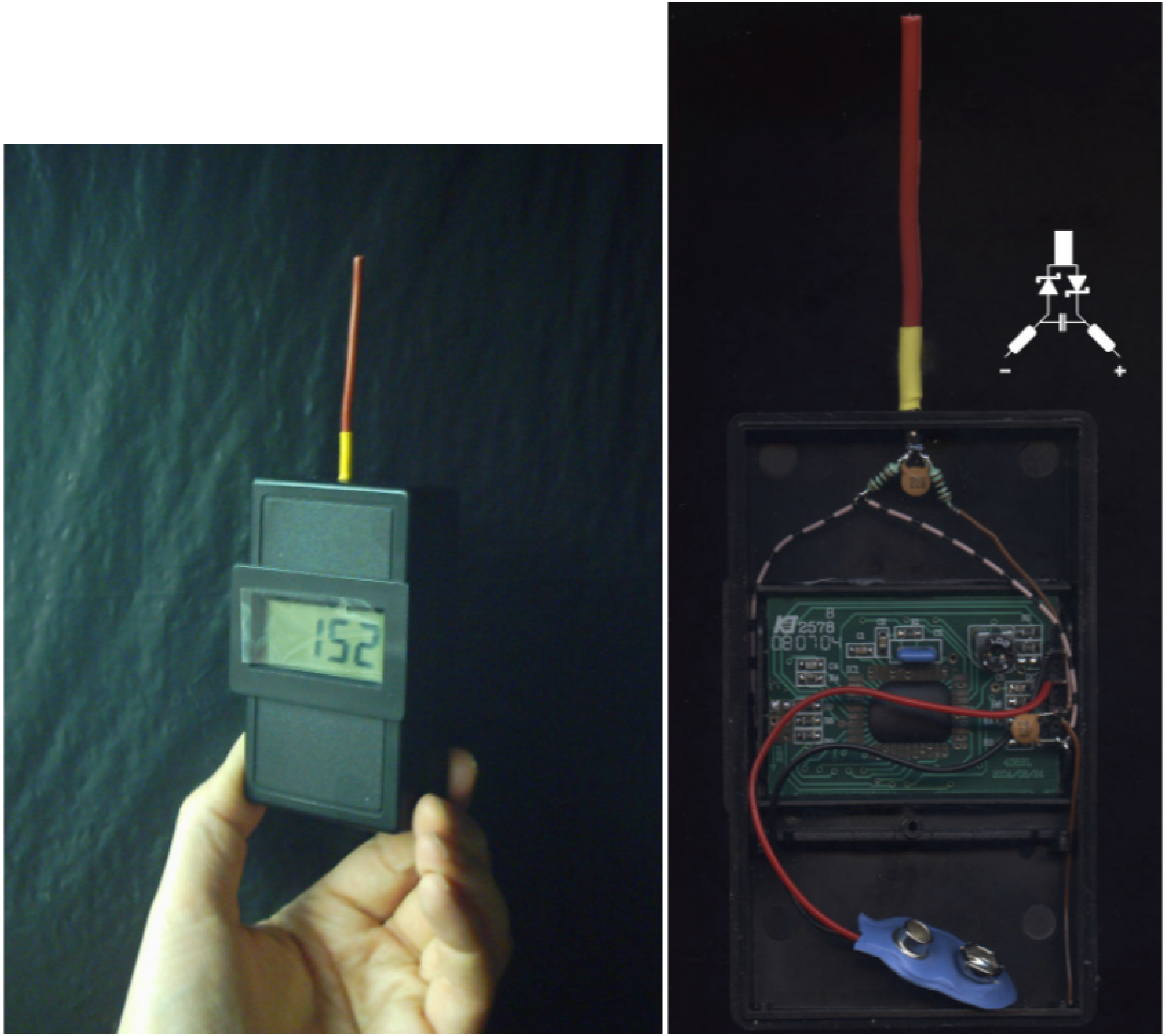
152
2578
080704
-
+

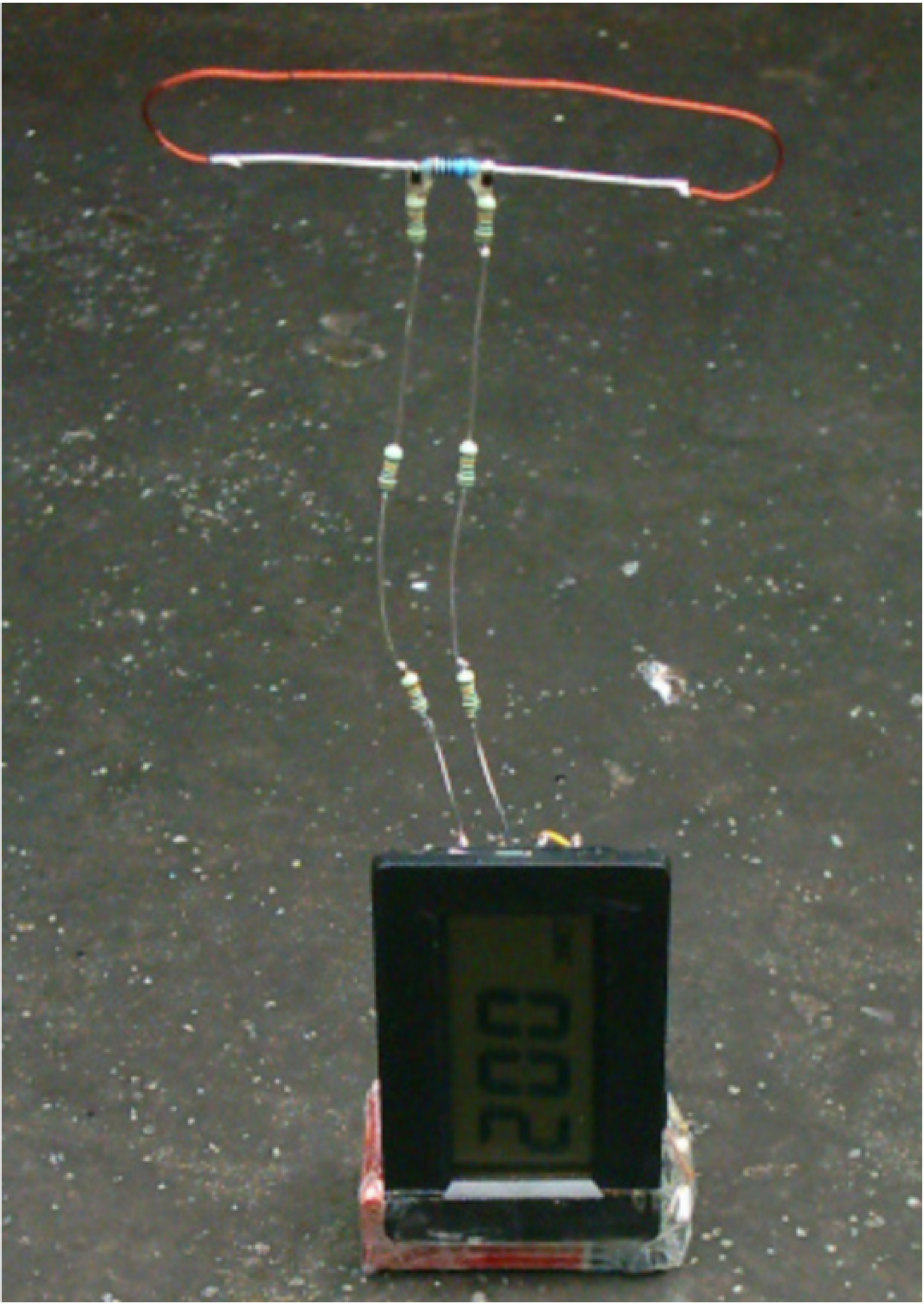

feluri (fig. III.9), schemele fiind echivalente cu reprezentarea de bază din fig. III.8. Diodele sînt legate în serie, în formă de patrulater, două avînd comun anodul (punctul *2*), iar celelalte două catodul (punctul *4*). Tensiunea alternativă de intrare se aplică pe diagonala *1—3*, iar consumatorul se conectează pe diagonala *2—4*.

Pentru a urmări funcționarea punții, să presupunem că prima alternanță sosită în nodul *1* este pozitivă. Ea blochează dioda D_2 și o deschide pe D_1, debitînd prin R_S un curent I_1 (săgețile pline), care se întoarce la

tice sau cu parametri cît mai apro-

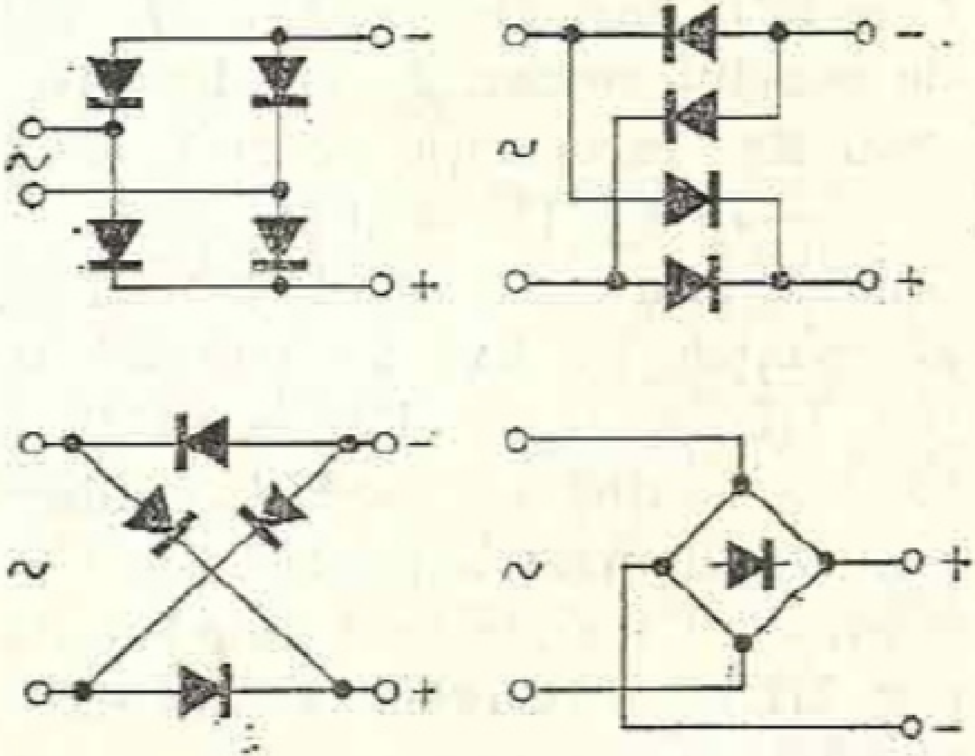

III.9. Puntea redresoare în diferite reprezentări.

Majoritatea componentelor active folosite în circuitele electronice moderne sînt dispozitivele bazate pe semiconductoare.

Cel mai simplu dispozitiv este *dioda* punctiformă cu germaniu. Ea are proprietatea de bază de a se comporta ca şi cum este conectată direct la o sursă electrică de curent continuu (plusul sursei la plusul diodei) şi ca un izolator, cînd este conectată invers la aceeaşi sursă (plusul sursei la minusul diodei), ca în figura 2.9.

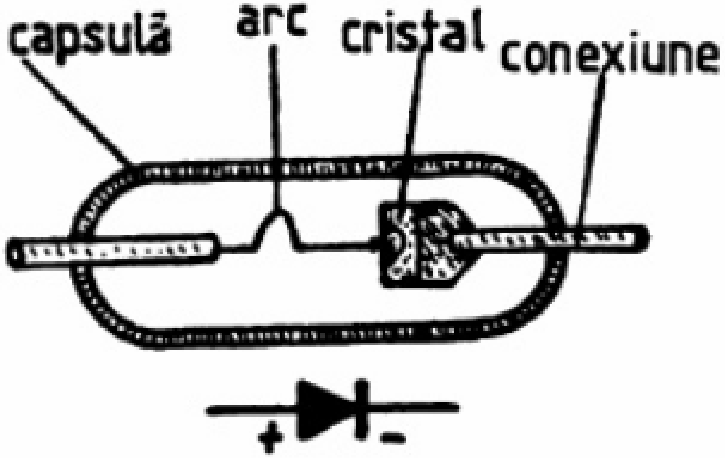
capsulă
arc
cristal
conexiune
+
-

FIG. 2.9.

Dioda punctiformă cu germaniu

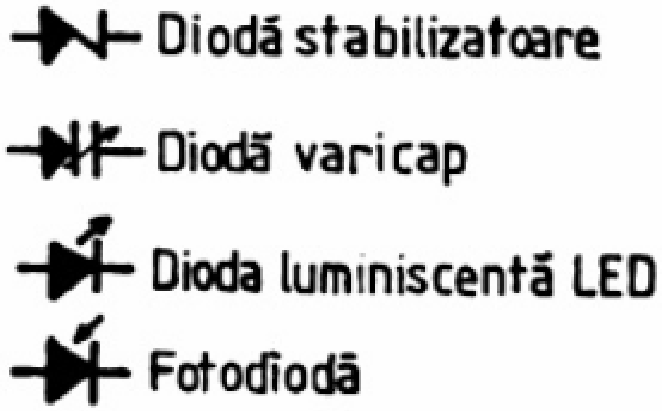
Diodă stabilizatoare
Diodă varicap
Dioda luminiscentă LED
Fotodiodă

FIG. 2.10.

Tipuri de diode semiconductoare

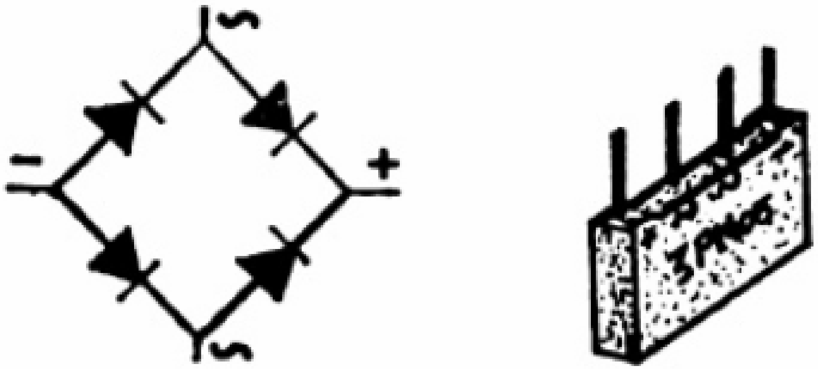

FIG. 2.11.

Puntea redresoare

A Light-Emitting Diode (LED)

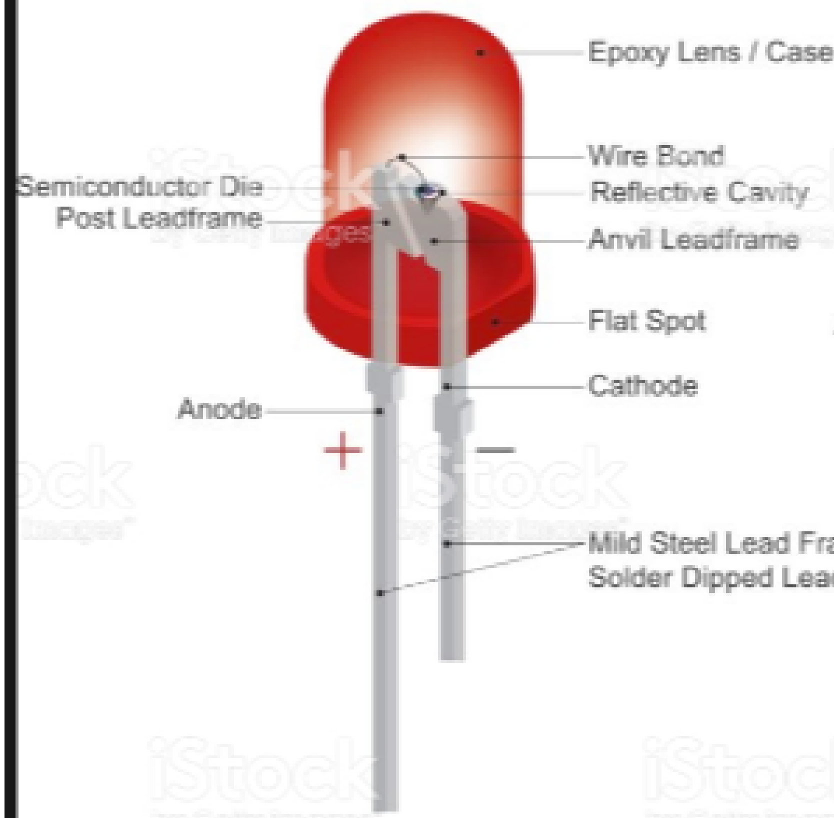
Epoxy Lens / Case
Wire Bond
Semiconductor Die
Reflective Cavity
Post Leadframe
Anvil Leadframe
Flat Spot
Cathode
Anode
+
—
Mild Steel Lead Frame,
Solder Dipped Leads

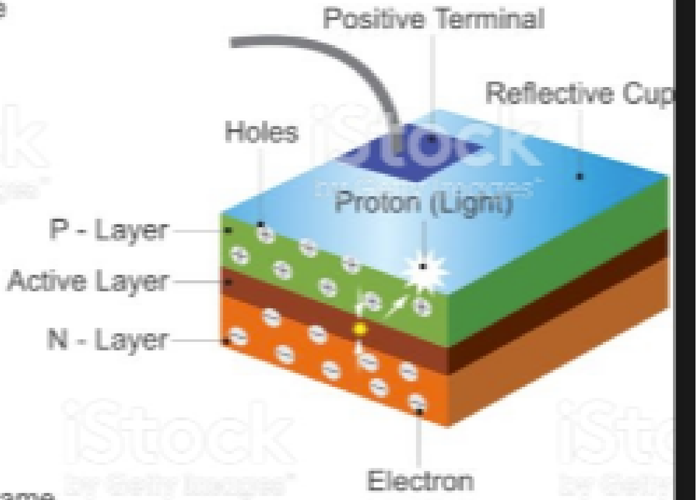
Positive Terminal
Reflective Cup
Holes
Proton (Light)
P - Layer
Active Layer
N - Layer
Electron

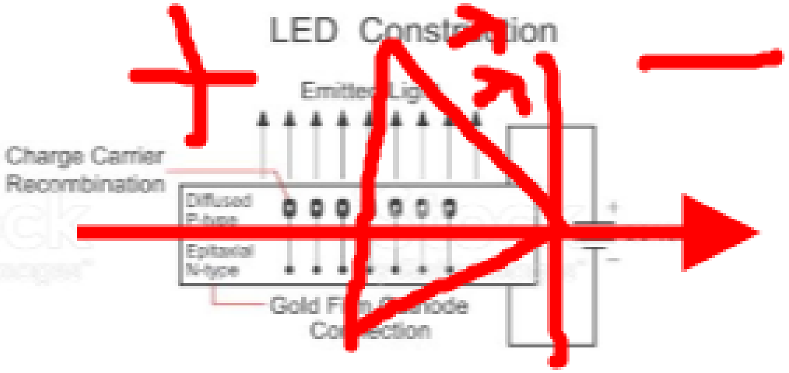
LED Construction
+
—
Emitted Light
Charge Carrier
Recombination
Diffused
P-type
Epitaxial
N-type
Gold Film Cathode
Connection
+
−

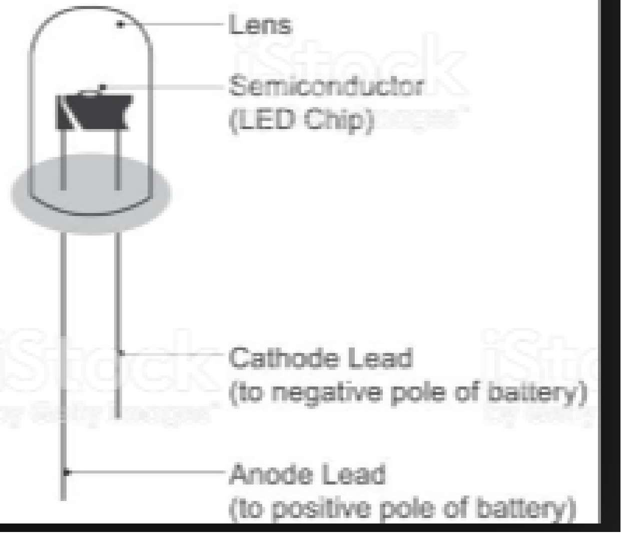
Lens
Semiconductor
(LED Chip)
Cathode Lead
(to negative pole of battery)
Anode Lead
(to positive pole of battery)

iStock by Getty Images

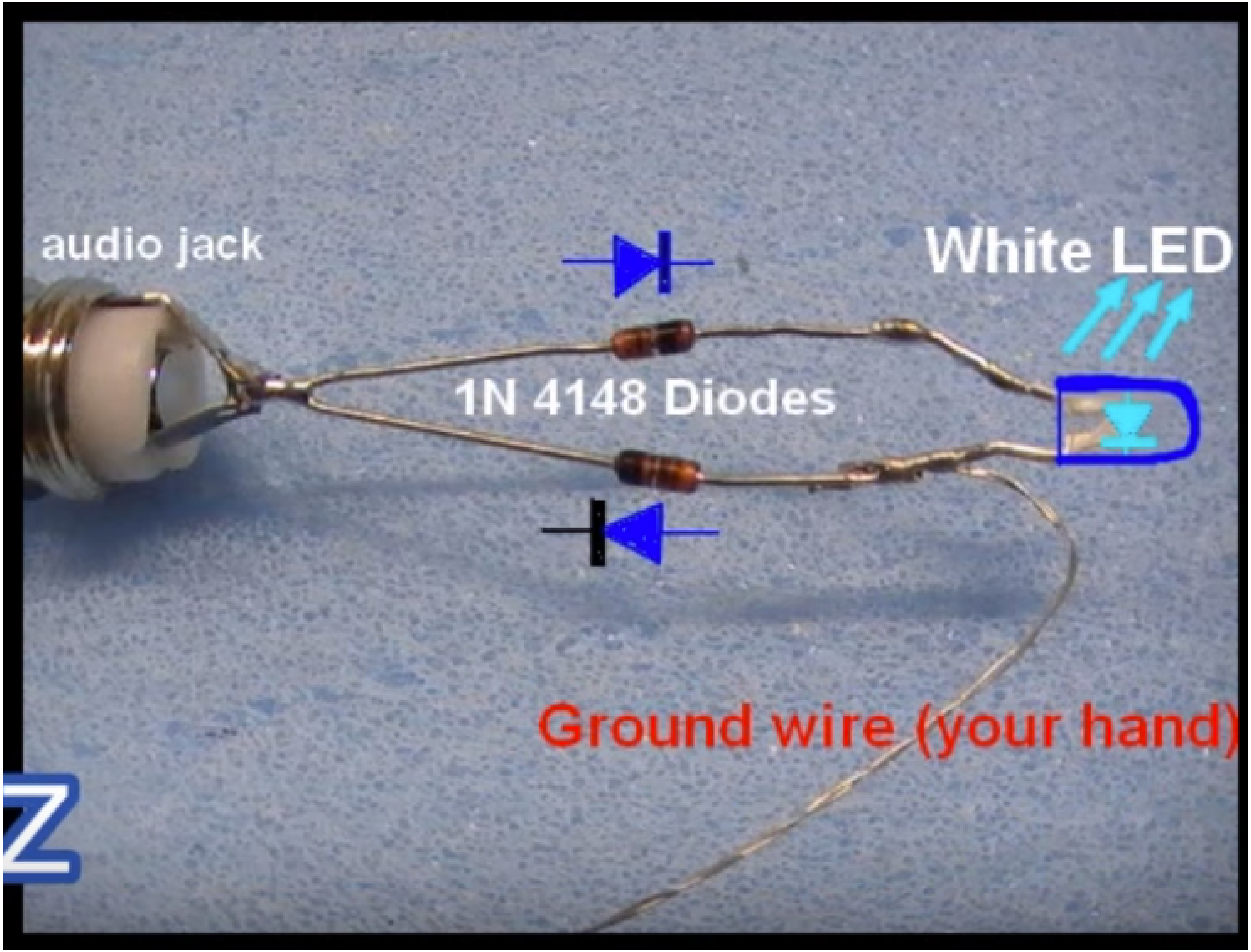
audio jack
White LED
1N 4148 Diodes
Ground wire (your hand)

4

The **immortal jellyfish** (*Turritopsis dohrnii*) is capable of biological immortality.

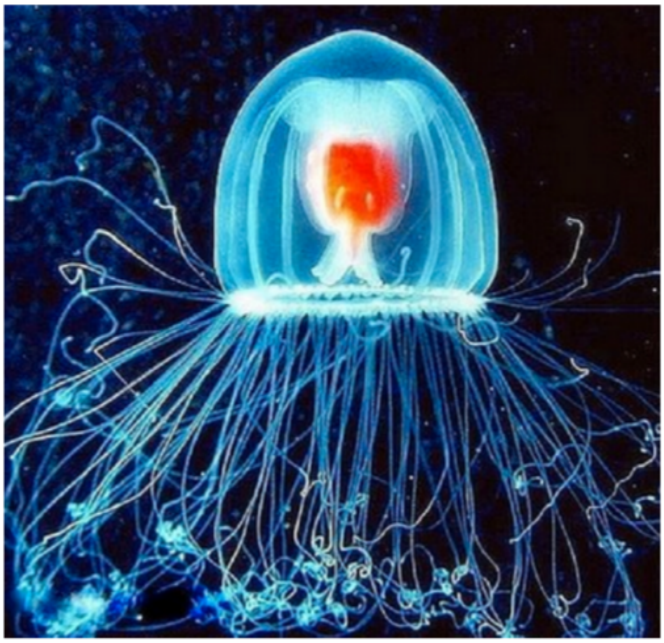

It's one of few known species capable of reverting completely to a sexually immature, colonial polyp stage after having reached sexual maturity as a solitary (free-floating) individual (called a medusa).

> Theoretically, this process can go on indefinitely, effectively rendering the jellyfish biologically immortal

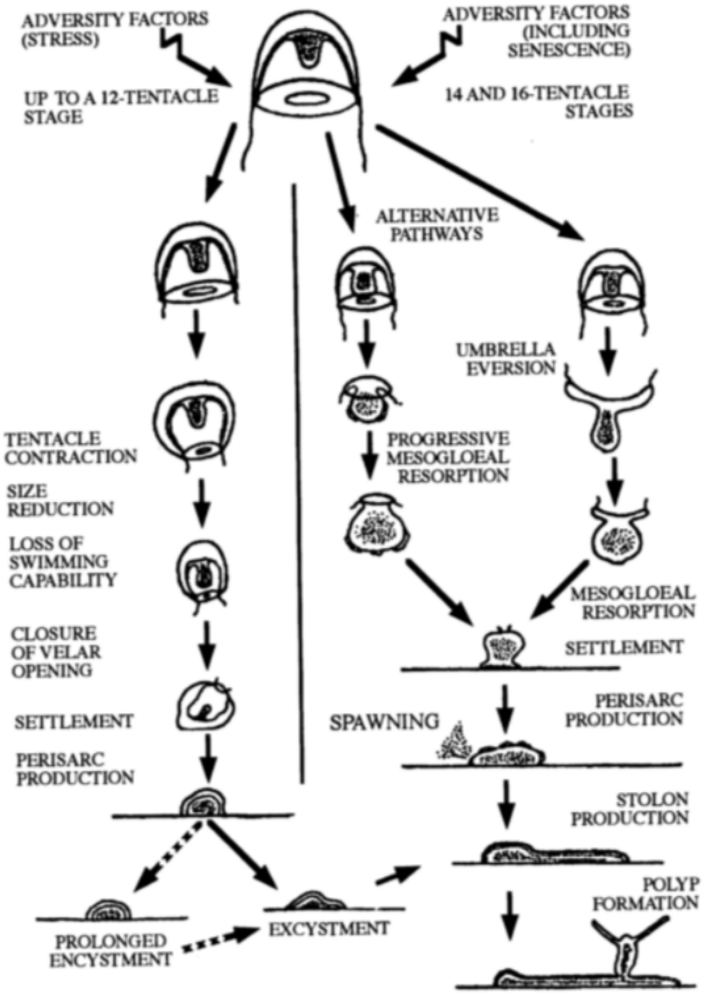
ADVERSITY FACTORS (STRESS)
UP TO A 12-TENTACLE STAGE
ADVERSITY FACTORS (INCLUDING SENESCENCE)
14 AND 16-TENTACLE STAGES
ALTERNATIVE PATHWAYS
UMBRELLA EVERSION
TENTACLE CONTRACTION
SIZE REDUCTION
LOSS OF SWIMMING CAPABILITY
CLOSURE OF VELAR OPENING
SETTLEMENT
PERISARC PRODUCTION
PROGRESSIVE MESOGLOEAL RESORPTION
MESOGLOEAL RESORPTION
SETTLEMENT
SPAWNING
PERISARC PRODUCTION
STOLON PRODUCTION
POLYP FORMATION
PROLONGED ENCYSTMENT
EXCYSTMENT

Figure 3. Pathways of transformation from medusa into polyp. Fate of stressed medusae up to 12-tentacle stage (left side), and alternative transformations of stressed or spawning medusae from a 14-tentacle or 16-tentacle stage (right side). The final product is always the polyp colony (bottom), directly or through a resting stage.

ELECTROMAGNETIC RADIATION SPECTRUM

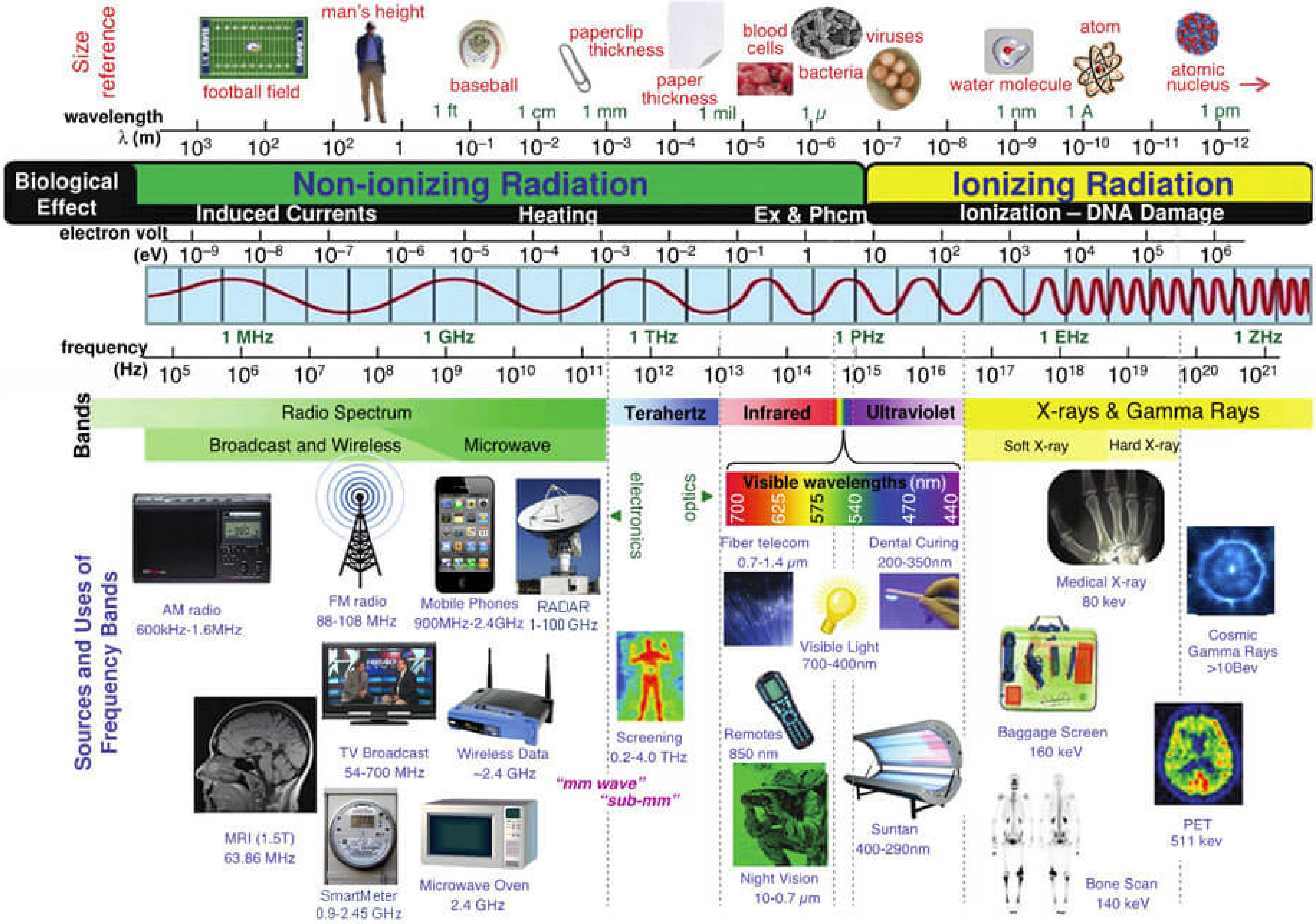
Size reference
football field
man's height
baseball
paperclip thickness
paper thickness
blood cells
bacteria
viruses
water molecule
atom
atomic nucleus →
wavelength λ (m)
10^3 10^2 10^2 1 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 10^{-6} 10^{-7} 10^{-8} 10^{-9} 10^{-10} 10^{-11} 10^{-12}
1 ft
1 cm
1 mm
1 mil
1 μ
1 nm
1 A
1 pm
Biological Effect
Non-ionizing Radiation
Ionizing Radiation
Induced Currents
Heating
Ex & Phcm
Ionization – DNA Damage
electron volt (eV)
10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10 10^2 10^3 10^4 10^5 10^6
frequency (Hz)
10^5 10^6 10^7 10^8 10^9 10^{10} 10^{11} 10^{12} 10^{13} 10^{14} 10^{15} 10^{16} 10^{17} 10^{18} 10^{19} 10^{20} 10^{21}
1 MHz
1 GHz
1 THz
1 PHz
1 EHz
1 ZHz
Bands
Radio Spectrum
Broadcast and Wireless
Microwave
Terahertz
Infrared
Ultraviolet
X-rays & Gamma Rays
Soft X-ray
Hard X-ray
Sources and Uses of Frequency Bands
electronics
optics
Visible wavelengths (nm)
700 625 575 540 470 440
AM radio 600kHz-1.6MHz
FM radio 88-108 MHz
Mobile Phones 900MHz-2.4GHz
RADAR 1-100 GHz
MRI (1.5T) 63.86 MHz
TV Broadcast 54-700 MHz
Wireless Data ~2.4 GHz
SmartMeter 0.9-2.45 GHz
Microwave Oven 2.4 GHz
"mm wave"
"sub-mm"
Screening 0.2-4.0 THz
Fiber telecom 0.7-1.4 μm
Visible Light 700-400nm
Dental Curing 200-350nm
Remotes 850 nm
Night Vision 10-0.7 μm
Suntan 400-290nm
Medical X-ray 80 kev
Cosmic Gamma Rays >10Bev
Baggage Screen 160 keV
PET 511 kev
Bone Scan 140 keV

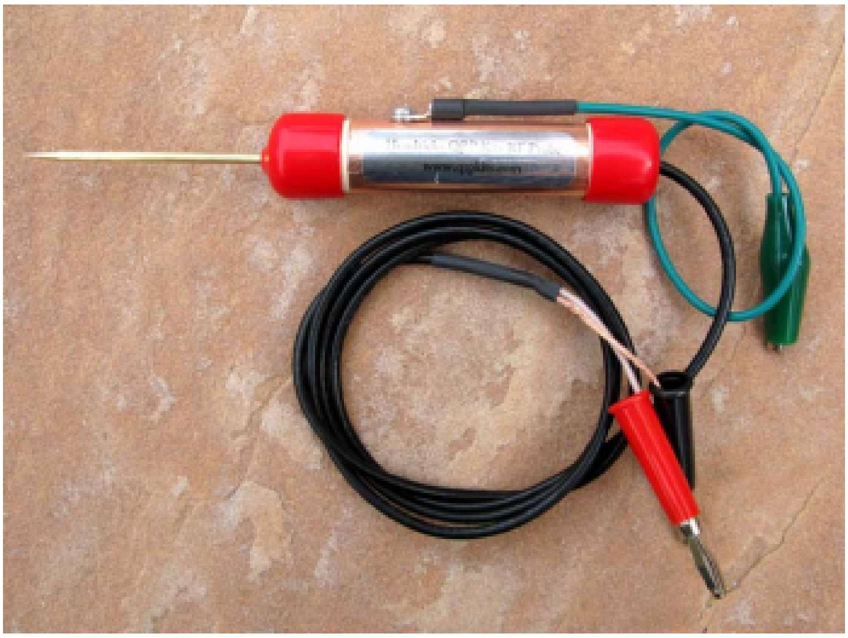

First off, check to see if the parts match the parts list…

1 – 1/2" x 3" CPVC tube
2 – 5/8" O.D. vinyl caps
1 – 3/32" dia x 2.5" brass rod
1 – 4-40 x 7/16" pan head screw
1 – 4-40 x 1/4" pan head screw
2 - #4 internal tooth lock washer
1 – 4-40 nut, steel
1 – 4-40 nut, brass
1 – 3/32 x 2" tyrap
2 - #4, 14-16ga ring terminal
1 – PCB
1 – D1 - 1N34A diode
1 – R1 - 4.7M 1/8w resistor (YEL, VIO, GRN, GLD) **See note**
1 – C1 - .01 disk ceramic capacitor (103)
3' – RG-174 coax
2 – banana plugs, 1 red, 1 black
2" – 3/16" dia. shrink tubing
1 – alligator clip and 9" lead
1 - copper foil tape, 2.25" x 2"
1 – self adhesive label

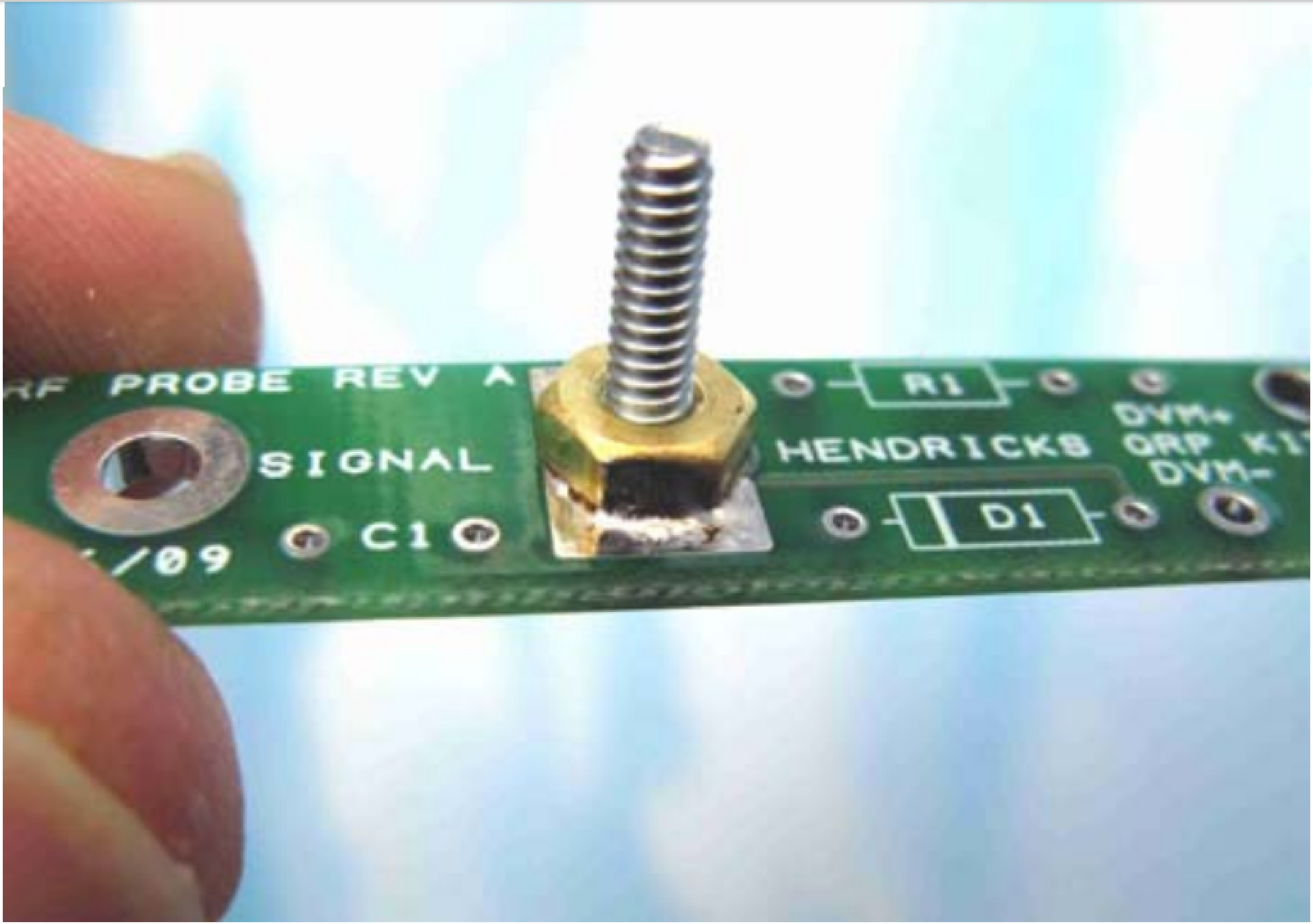
RF PROBE REV A
SIGNAL
/09
C1
R1
DVM+
HENDRICKS QRP KI
DVM-
D1

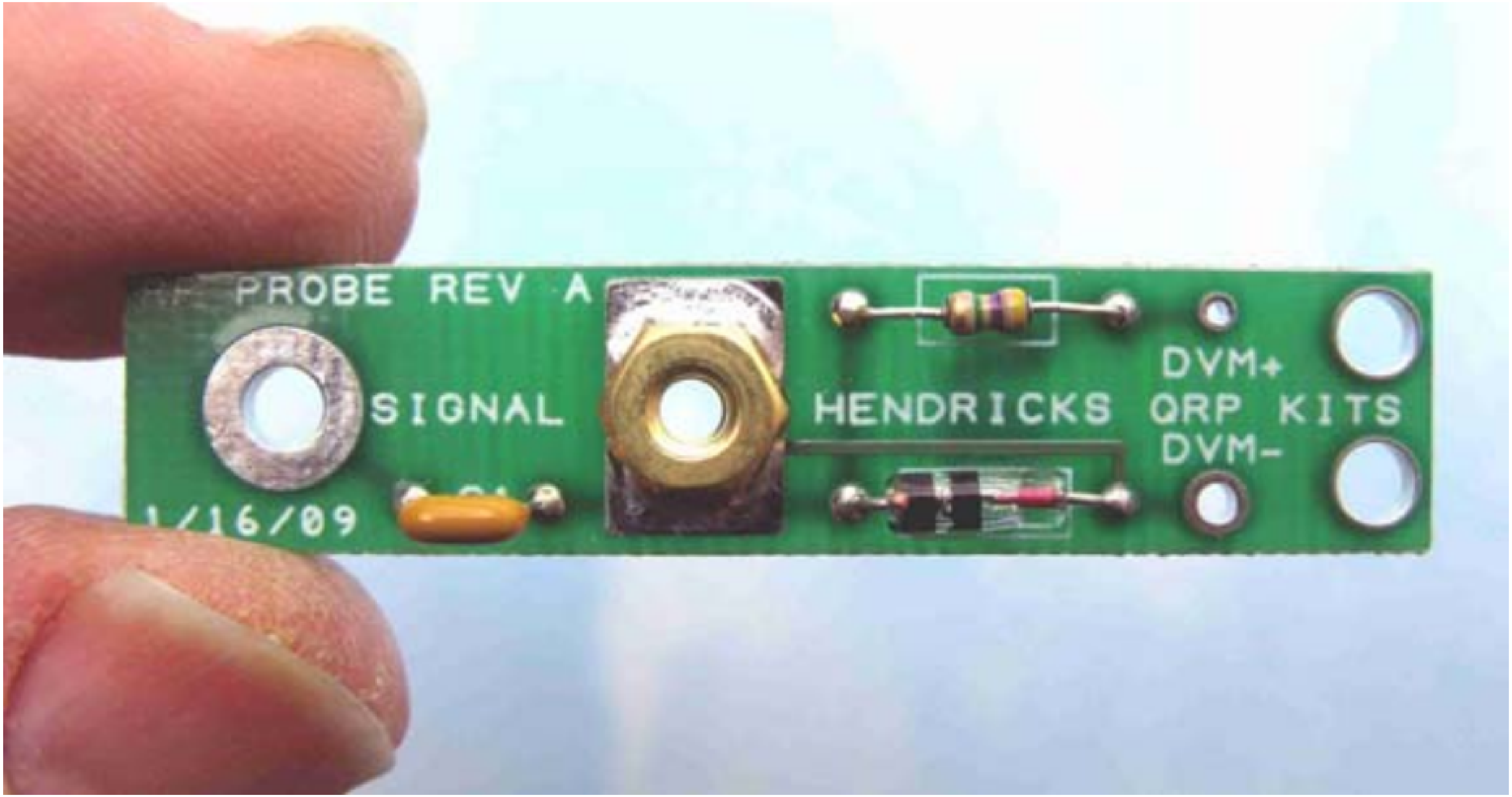
PROBE REV A
SIGNAL
1/16/09
HENDRICKS QRP KITS
DVM+
DVM-

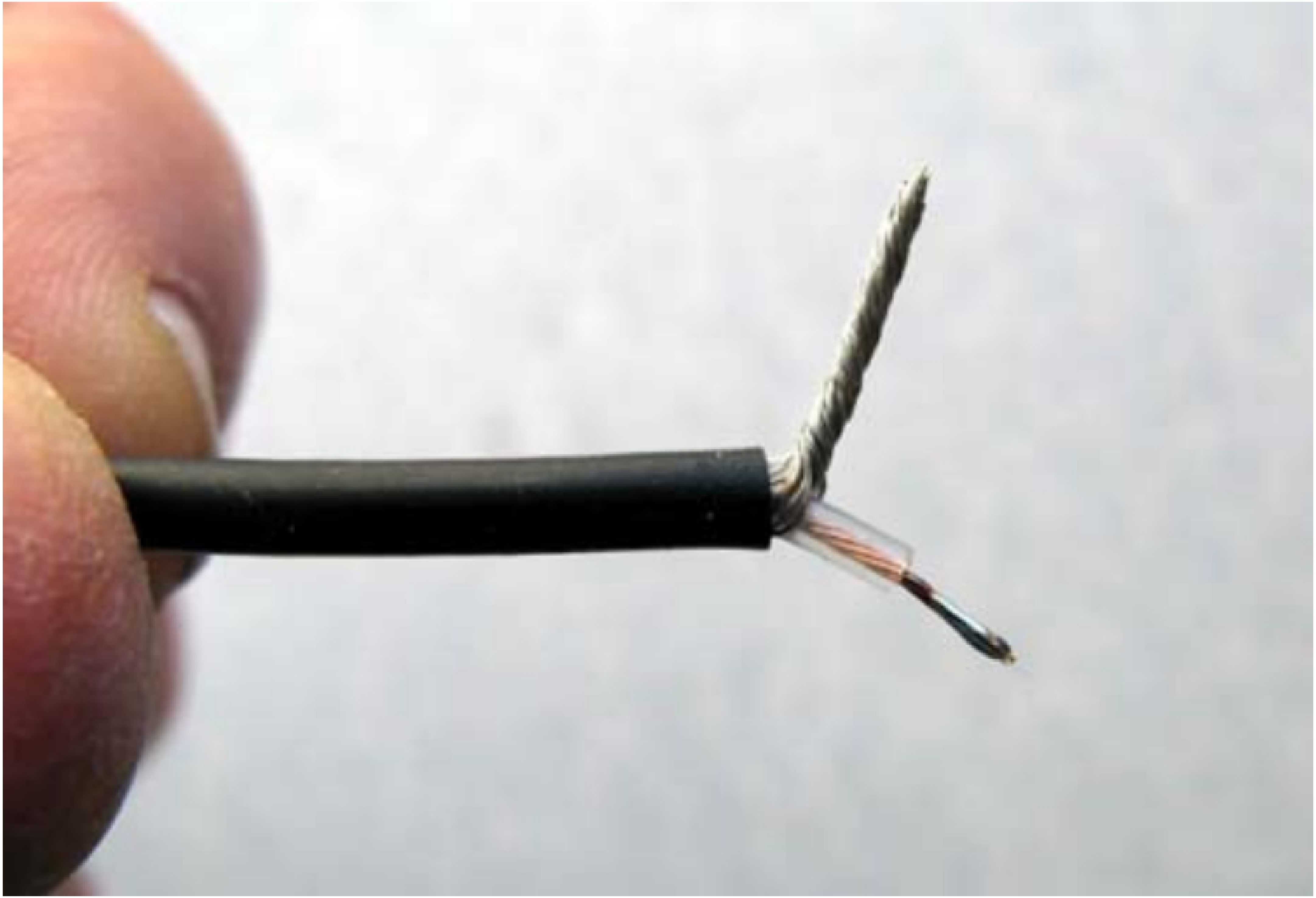

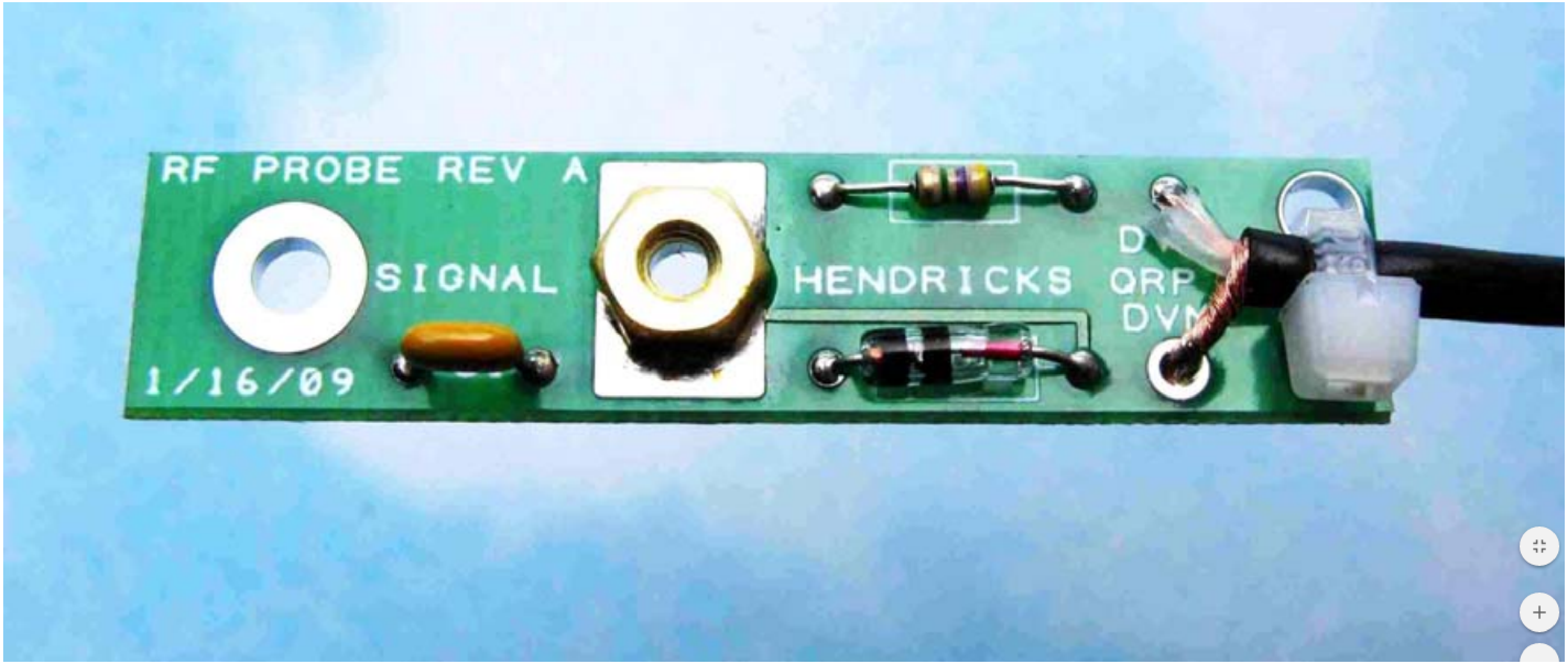
RF PROBE REV A
SIGNAL
1/16/09
HENDRICKS
D
QRP
DVM

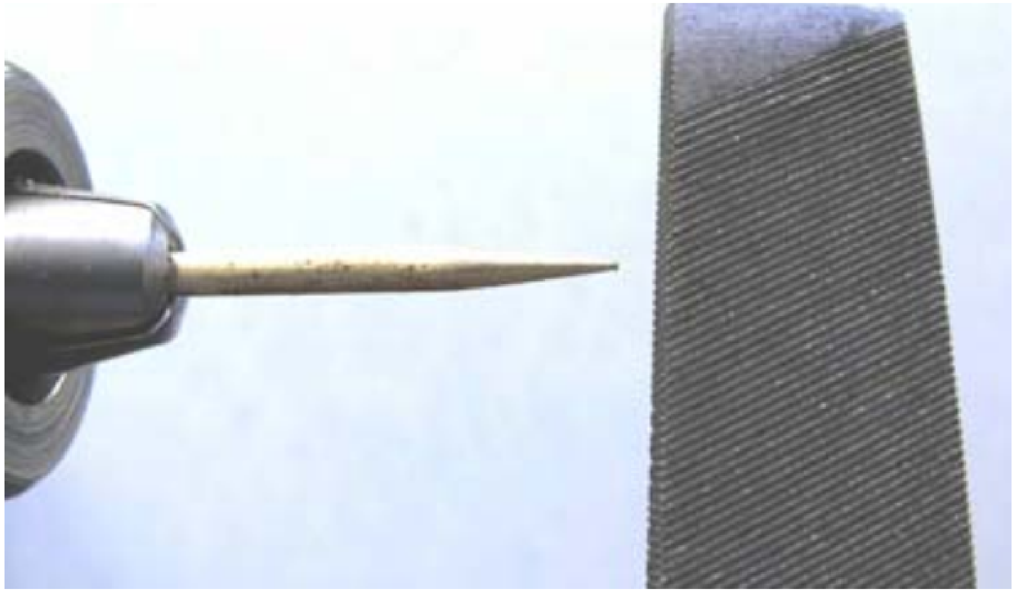

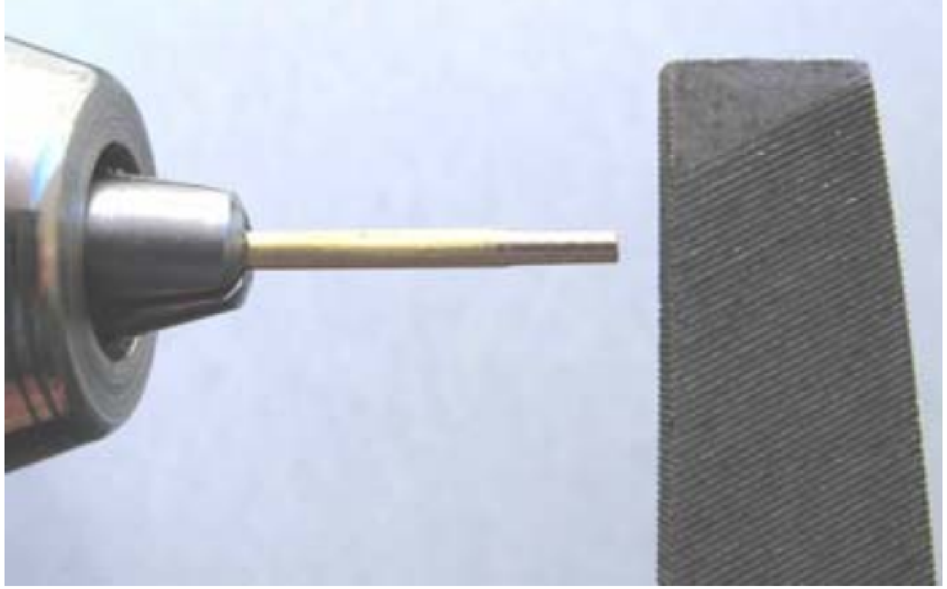

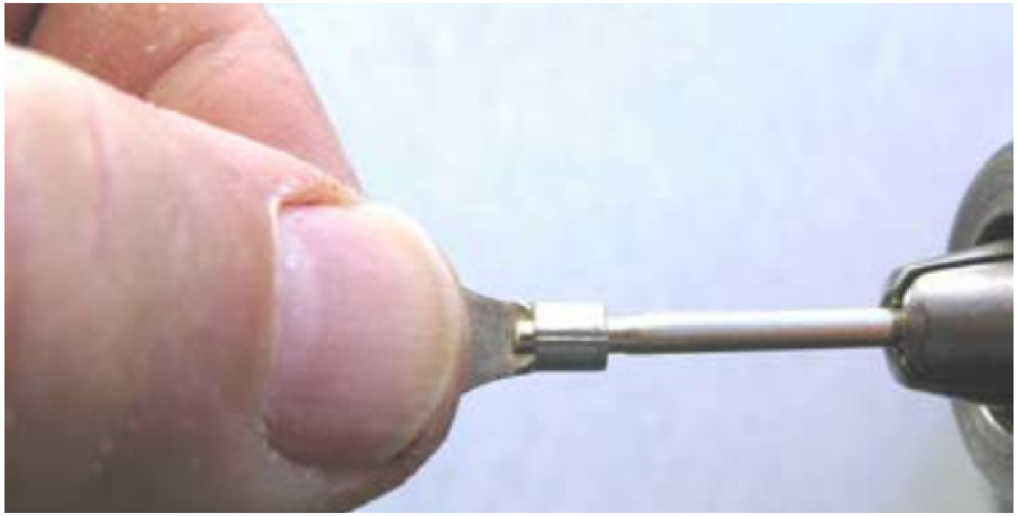

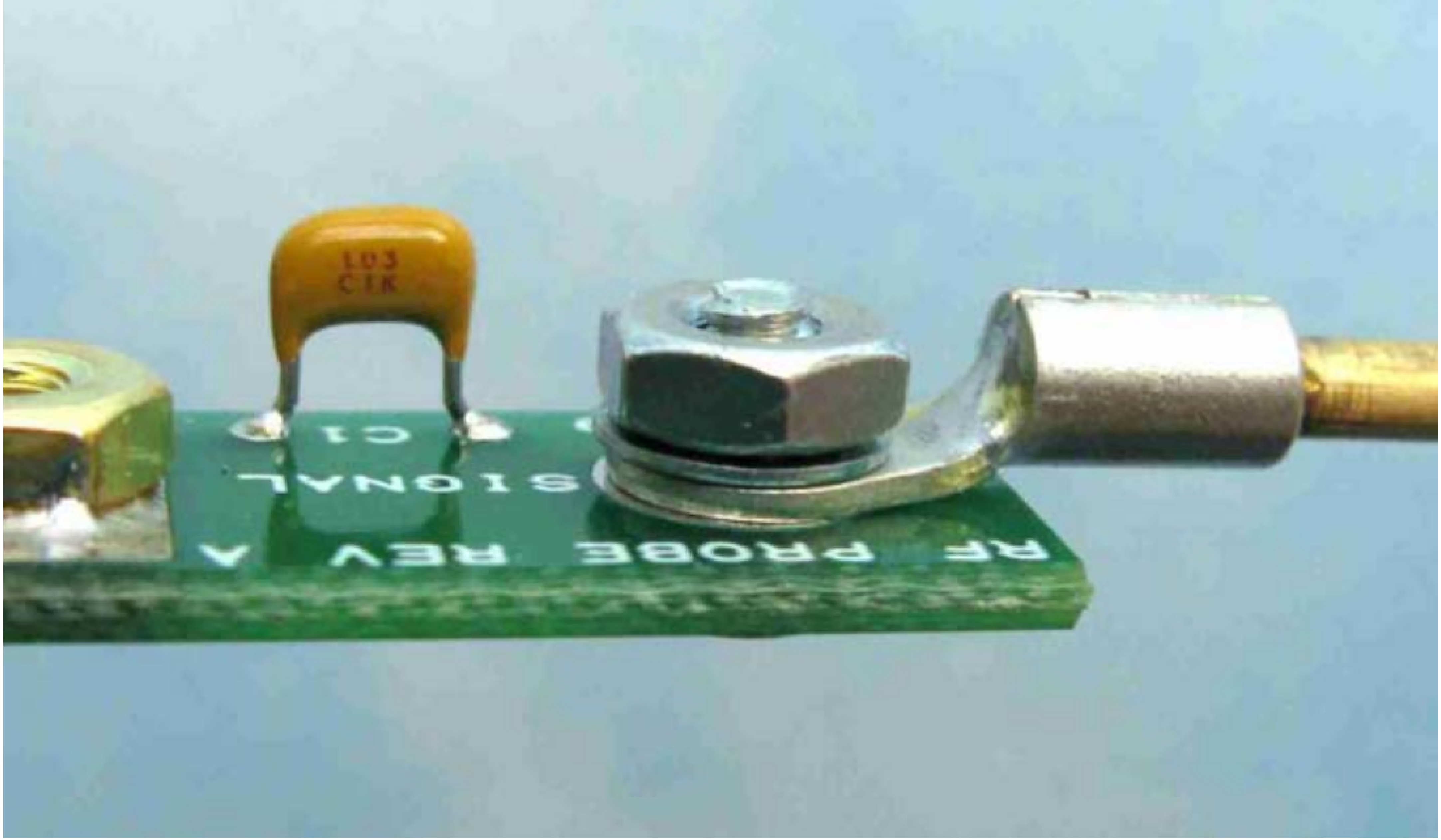
103
C1K
C1
SIGNAL
RF PROBE REV A

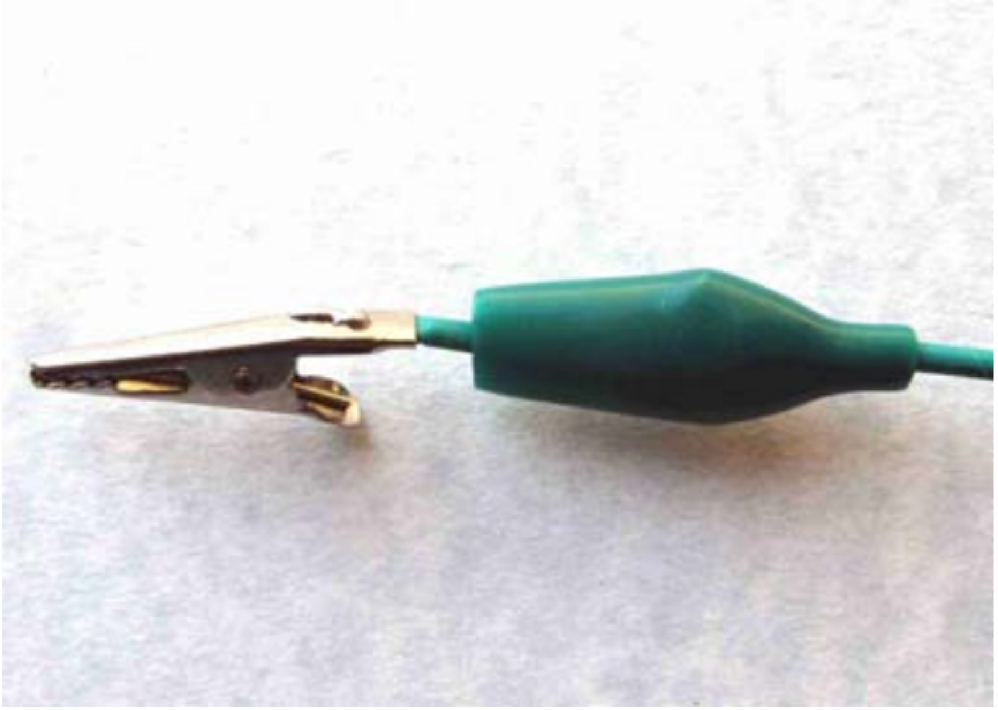

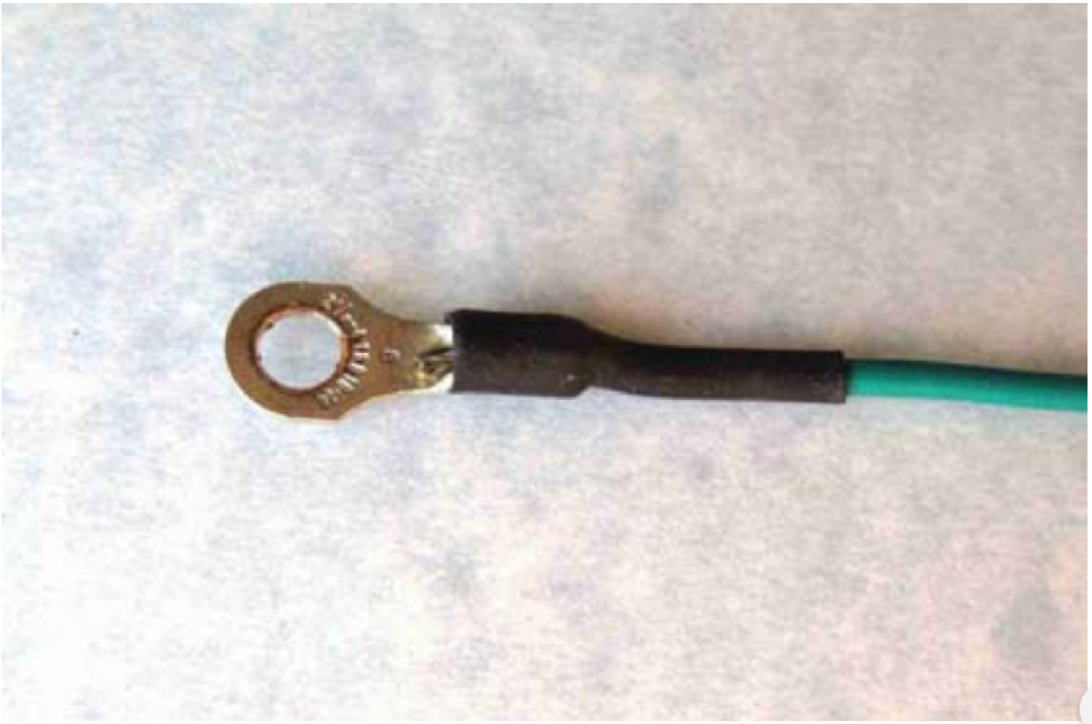

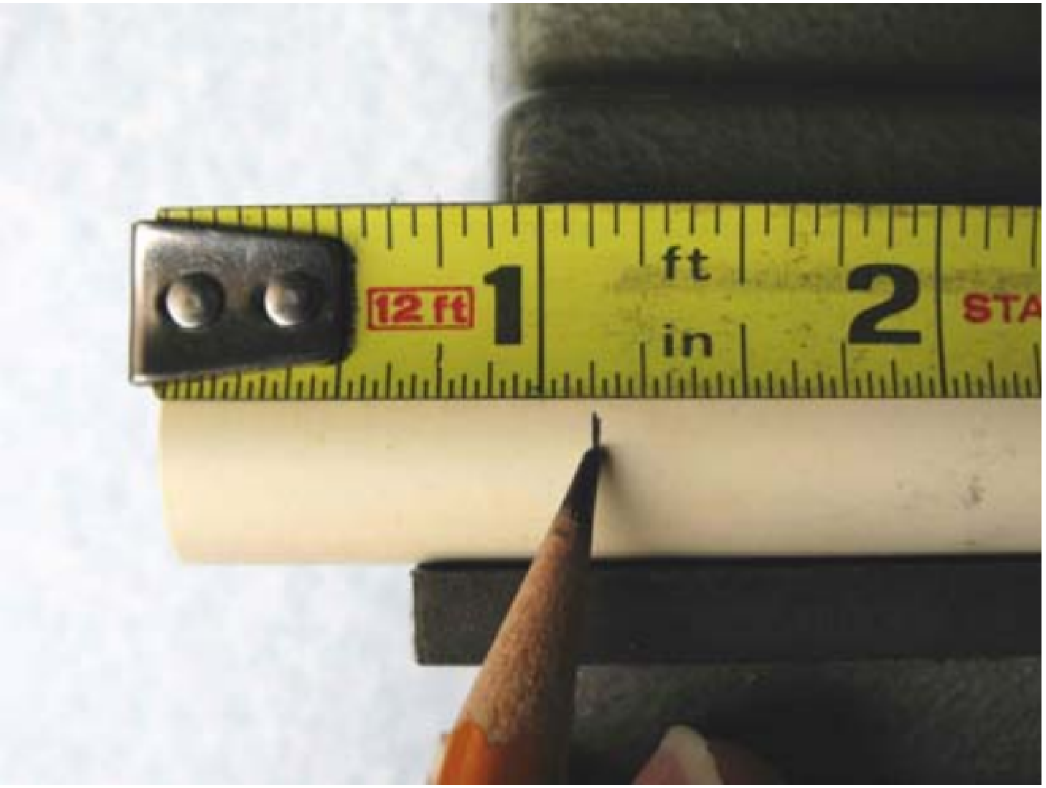
12 ft
1
ft
in
2
STA

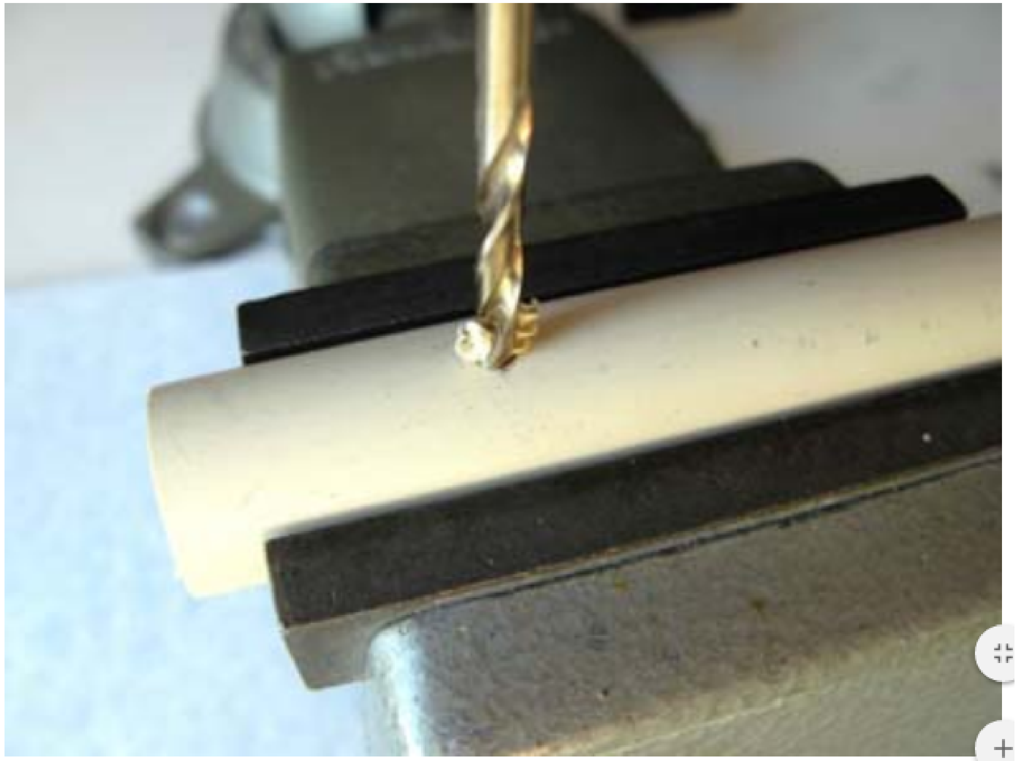

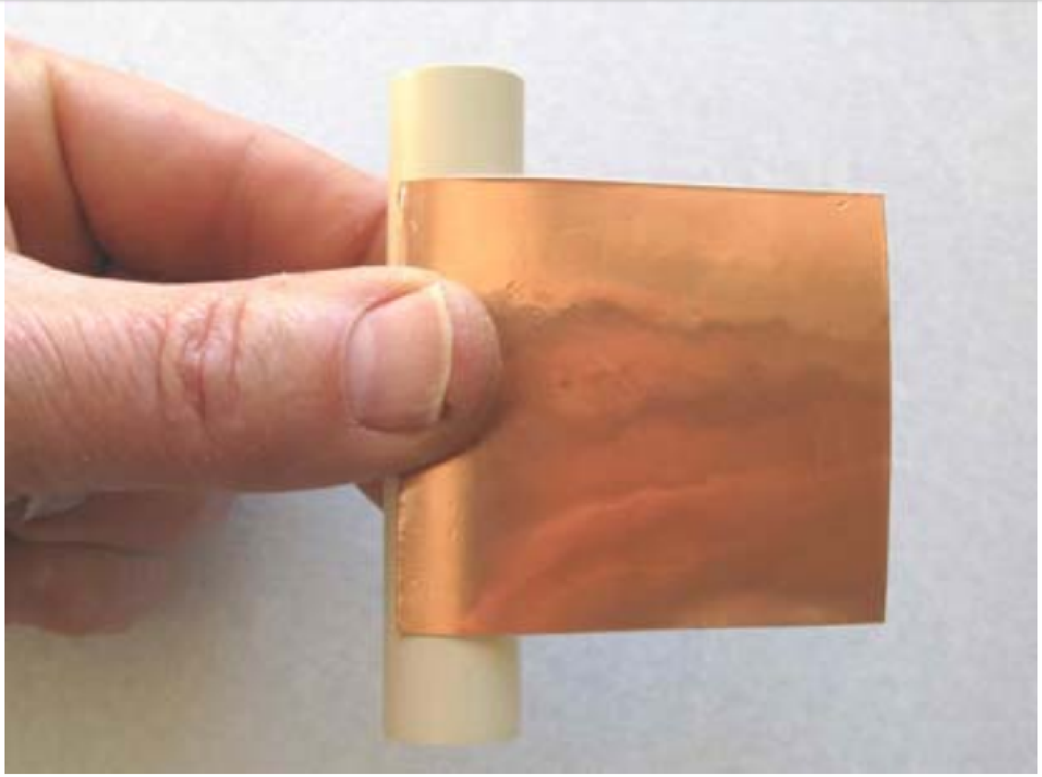

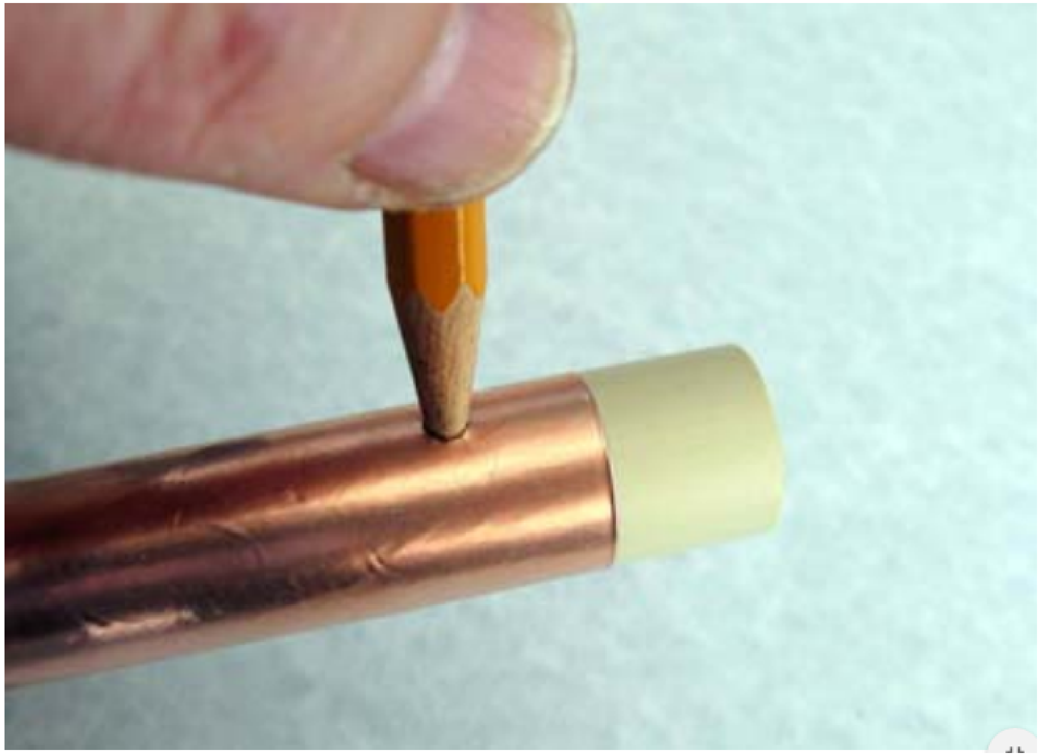

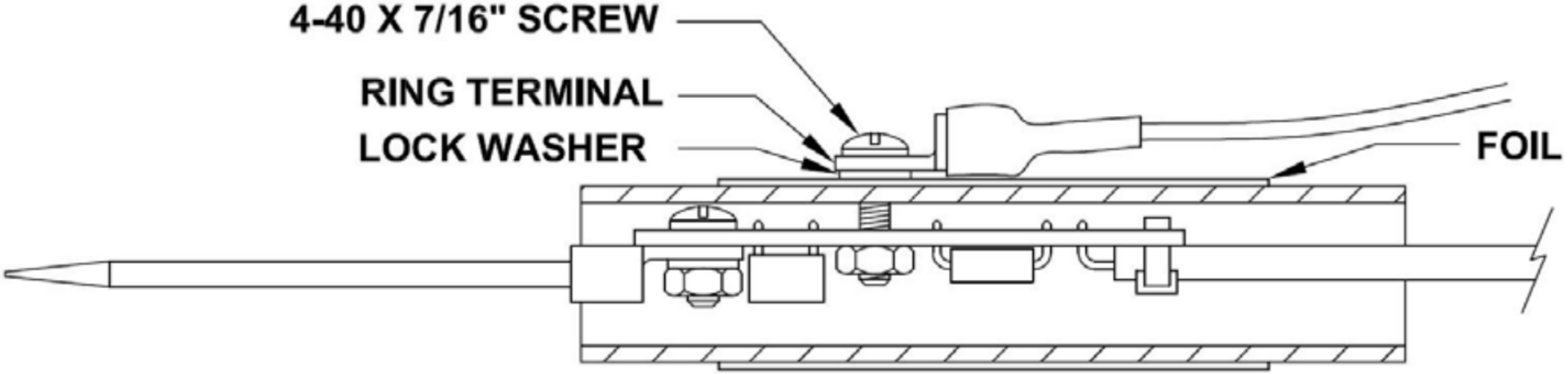
4-40 X 7/16" SCREW
RING TERMINAL
LOCK WASHER
FOIL

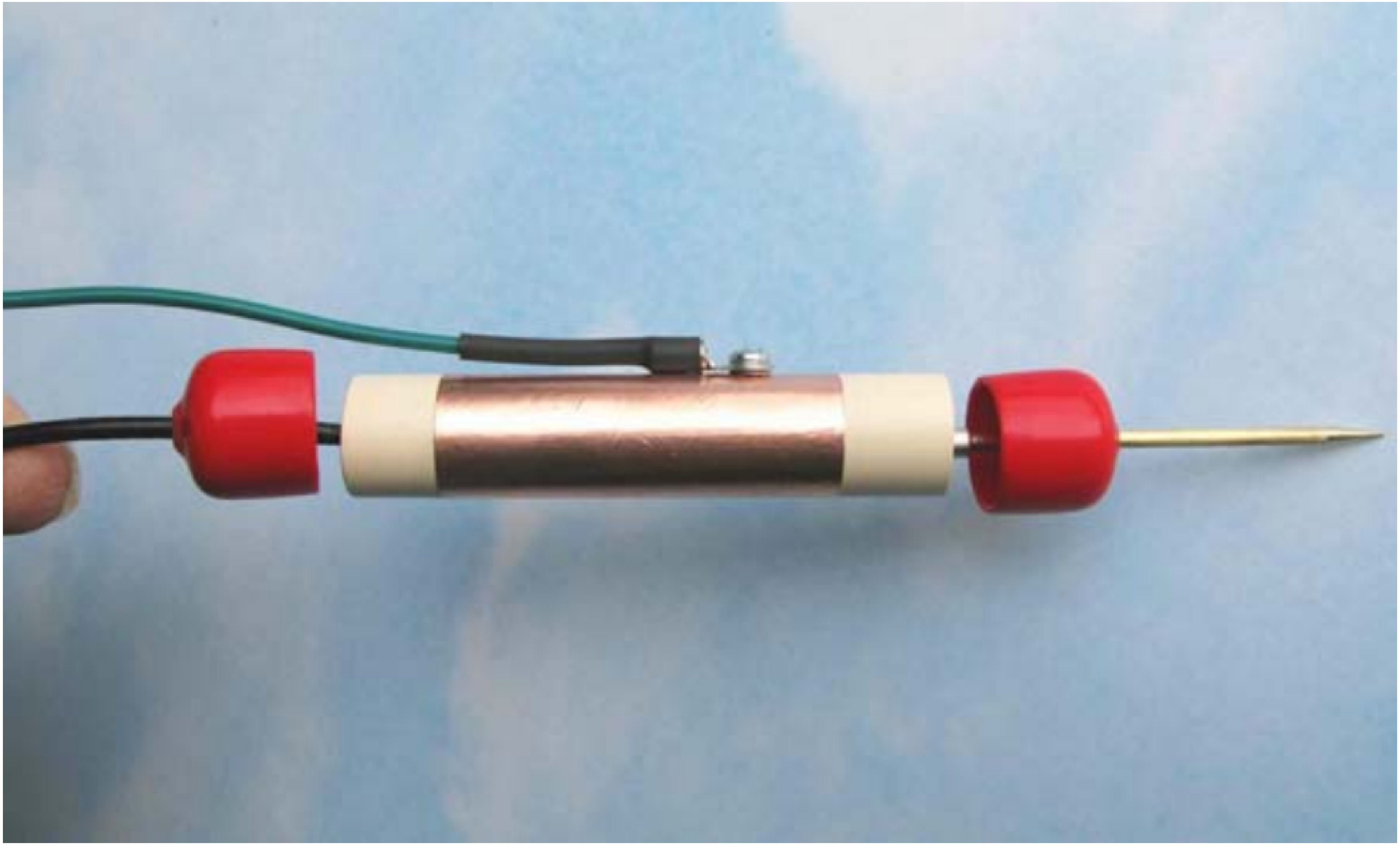

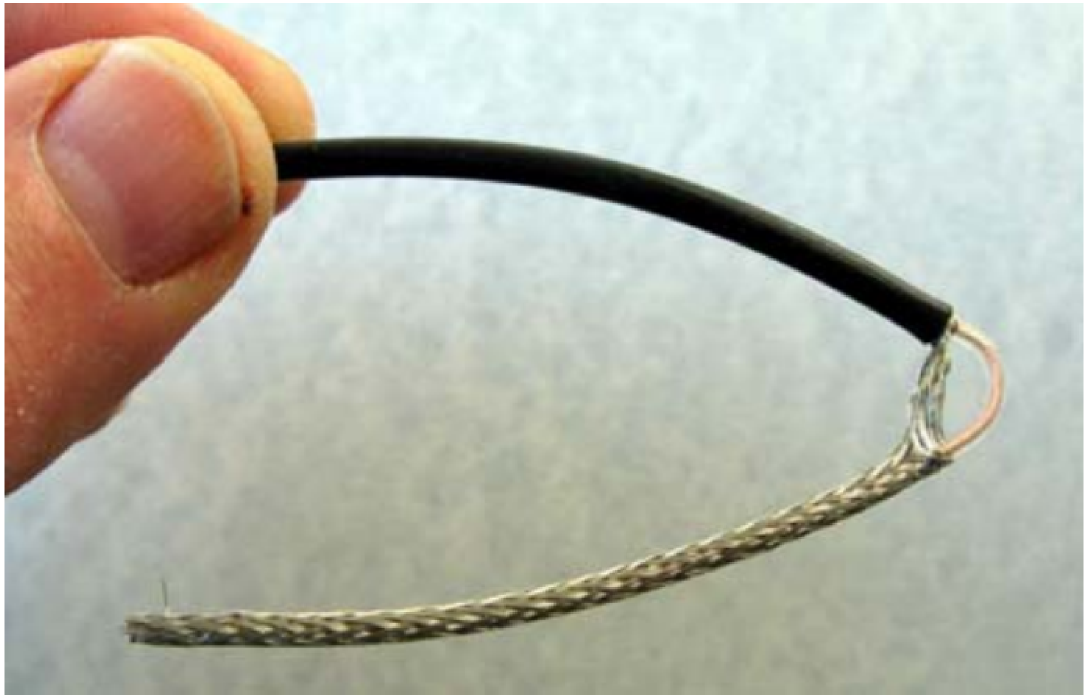

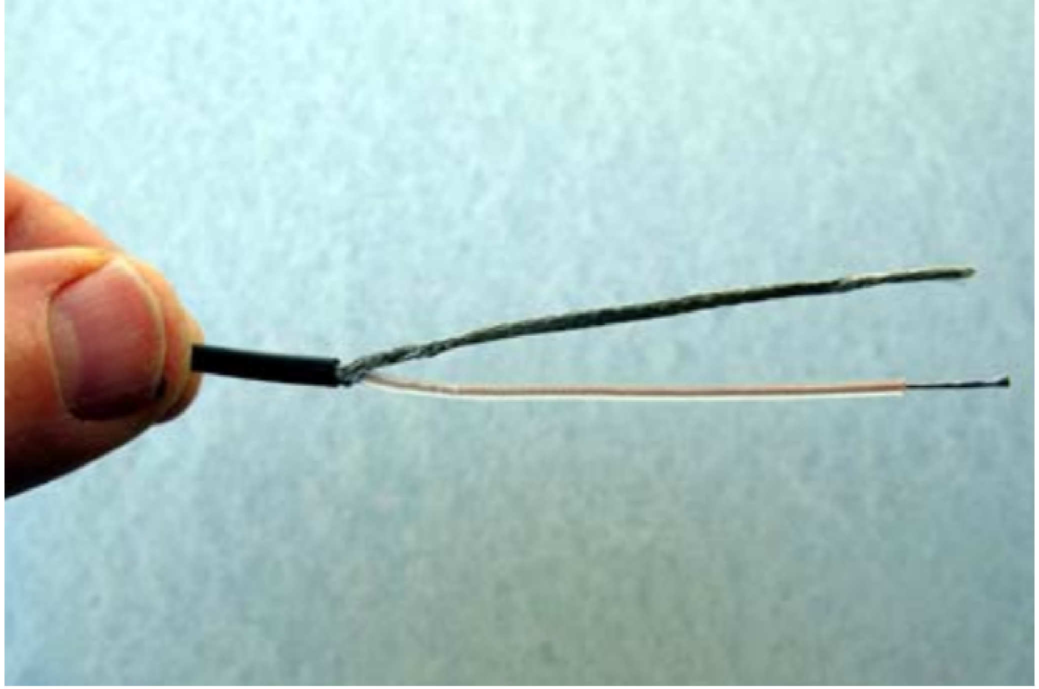

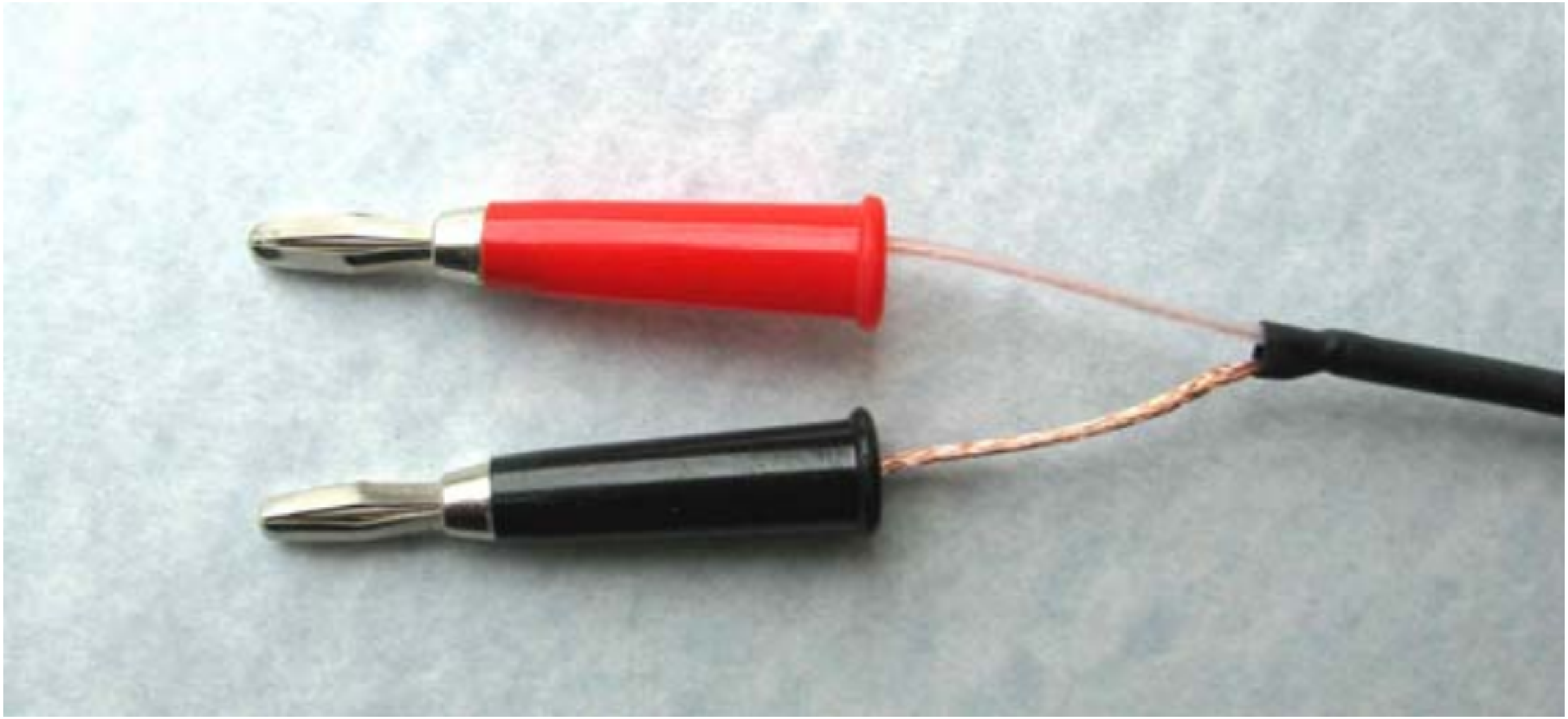

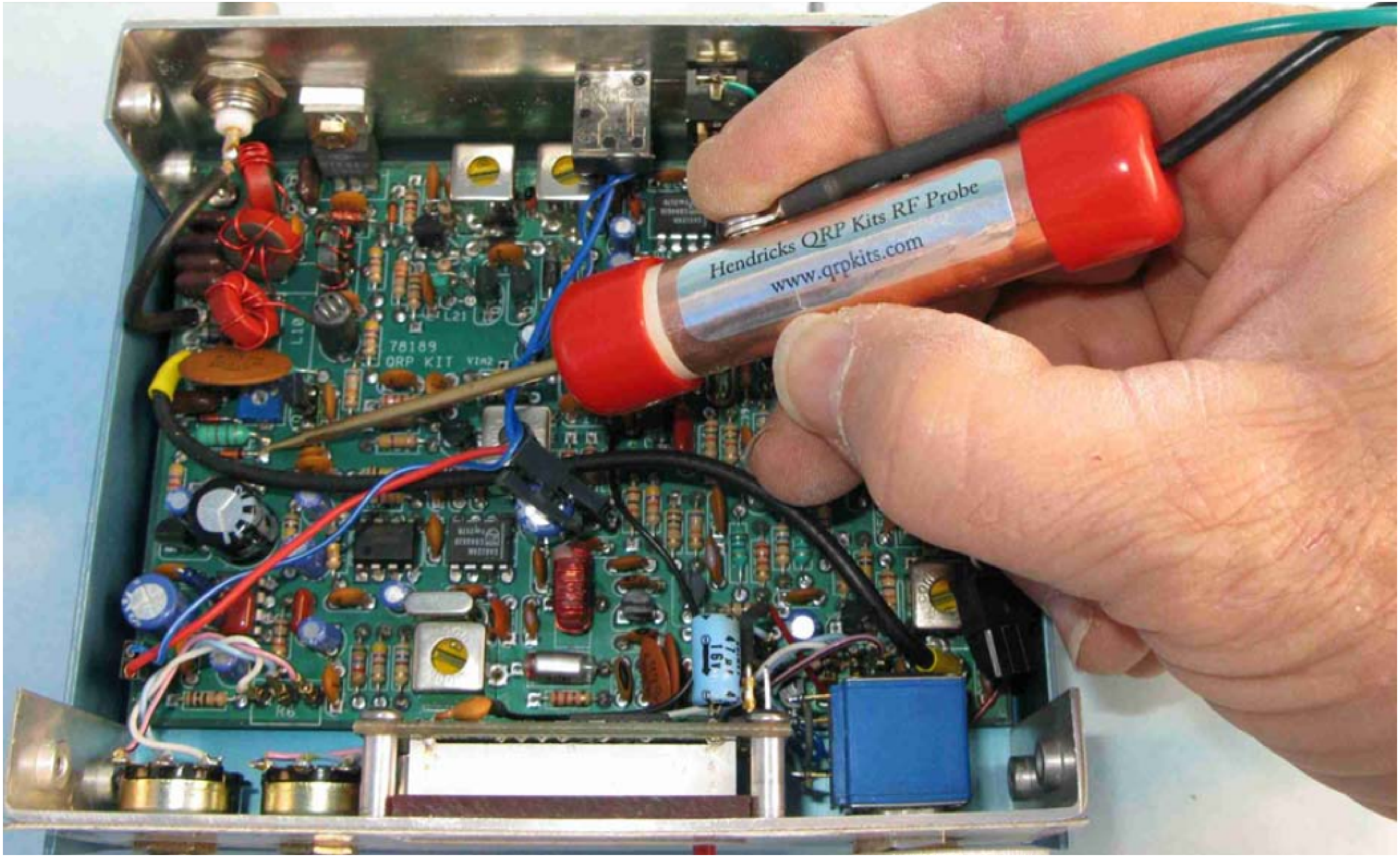
Hendricks QRP Kits RF Probe
www.qrpkits.com
78189
QRP KIT
L10
L21
R6

Look in the files section for **Dar's (Darwin Piatt) – W9HZC** application and usage tips for the R.F. Probe.

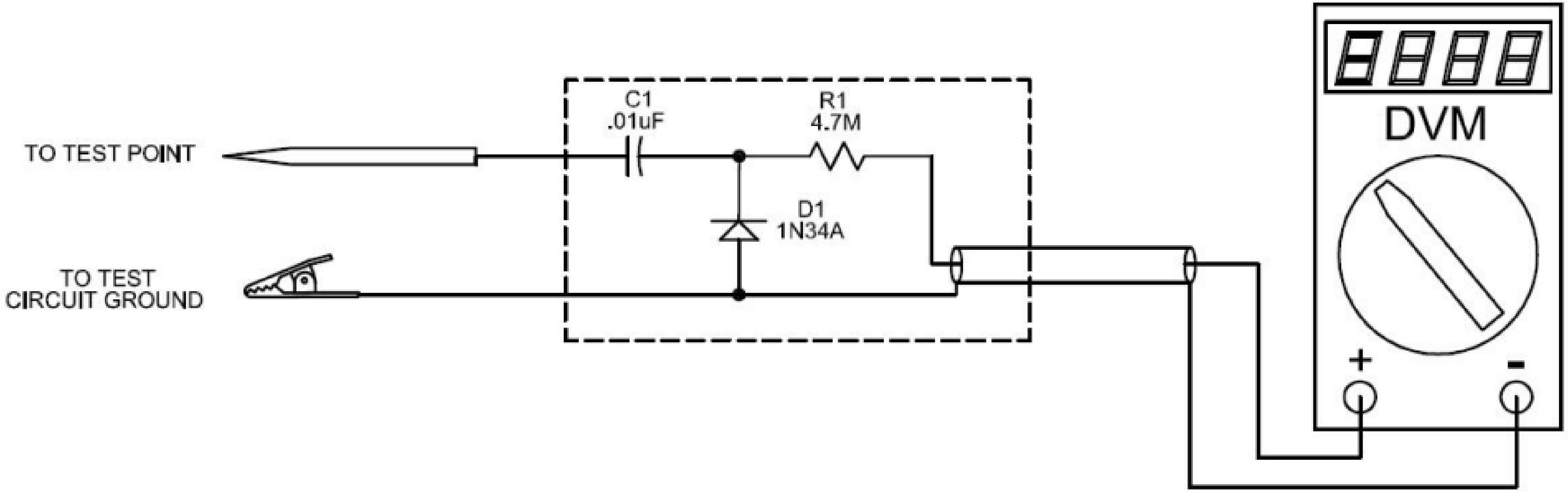
TO TEST POINT
TO TEST CIRCUIT GROUND
C1
.01uF
R1
4.7M
D1
1N34A
DVM
+
-

ELECTRIC ENERGY FROM SPACE:

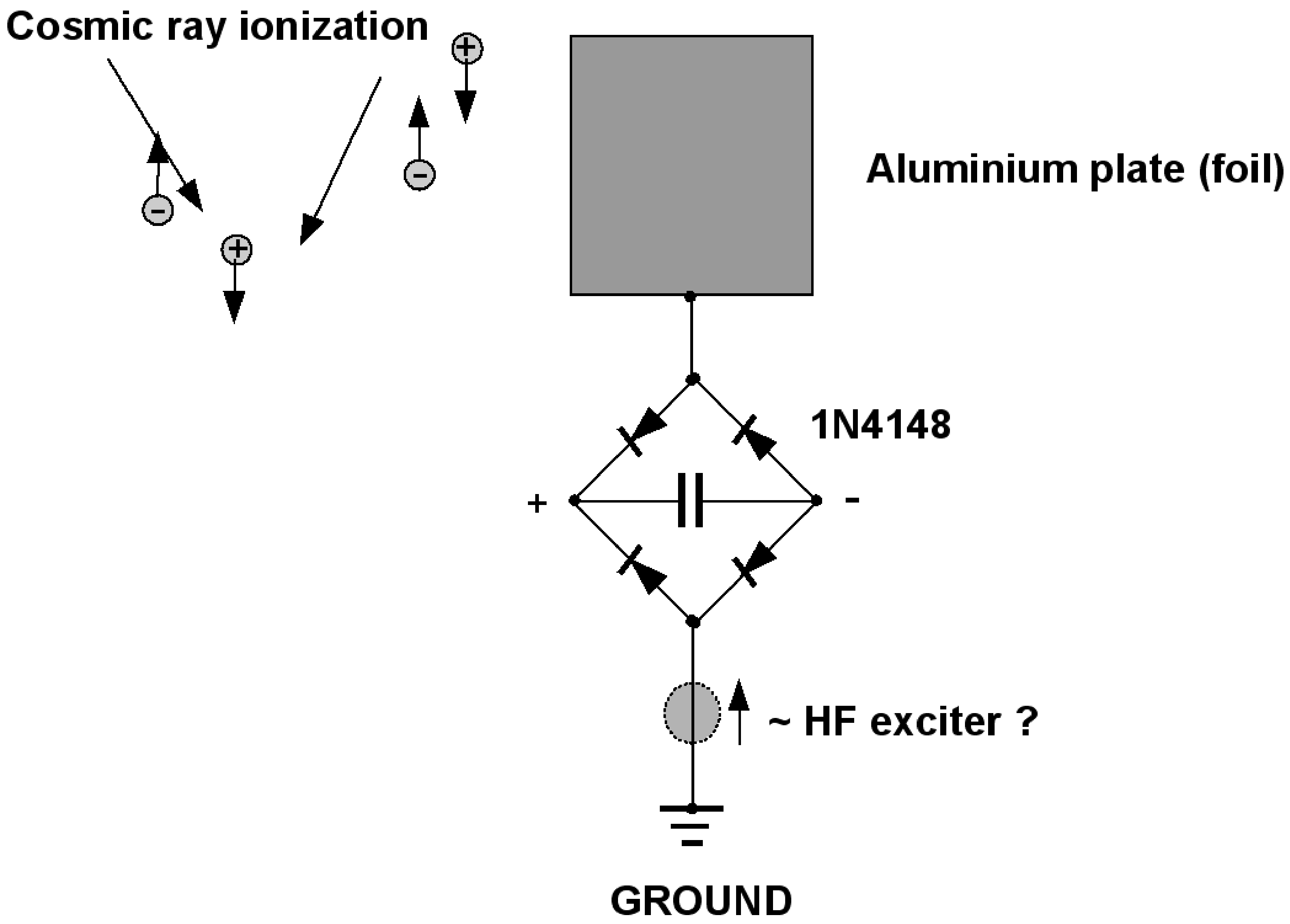

Cosmic ray ionization

Aluminium plate (foil)

1N4148

+

-

~ HF exciter ?

GROUND

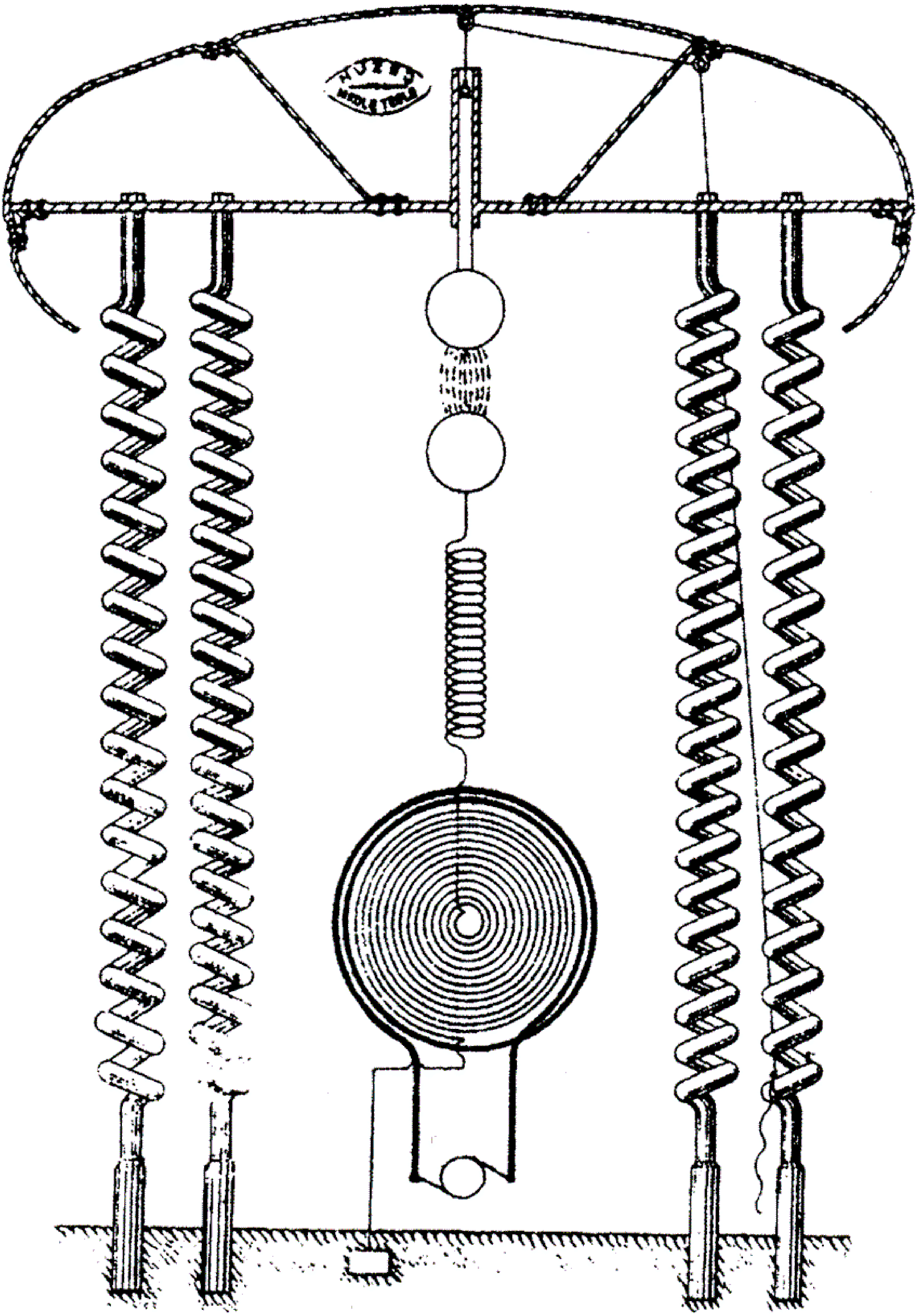

RF tip for Multimeter and Oscilloscope

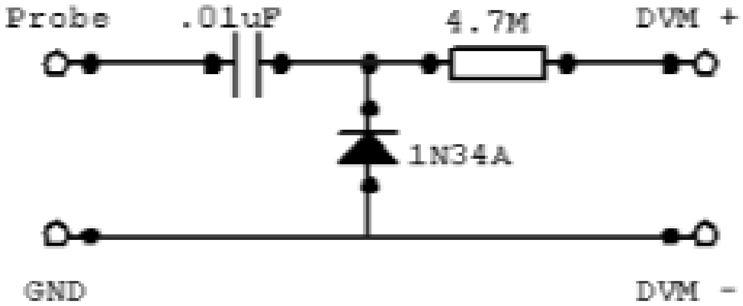

KA8MAV RF PROBE

The project **RF tip** is simple, not need special components and there is no great difficulty in its construction, making it ideal for beginners. The creator of this tip was the amateur KA8MAV, hence the name of the tip arose. Above the electronic circuit probe.

List of components for the tip RF

1 Germanium diode OA79, 1N34 or equivalent
1 Capacitor 0.01uF
1 Resistor 4M7 x 1/4 Watts

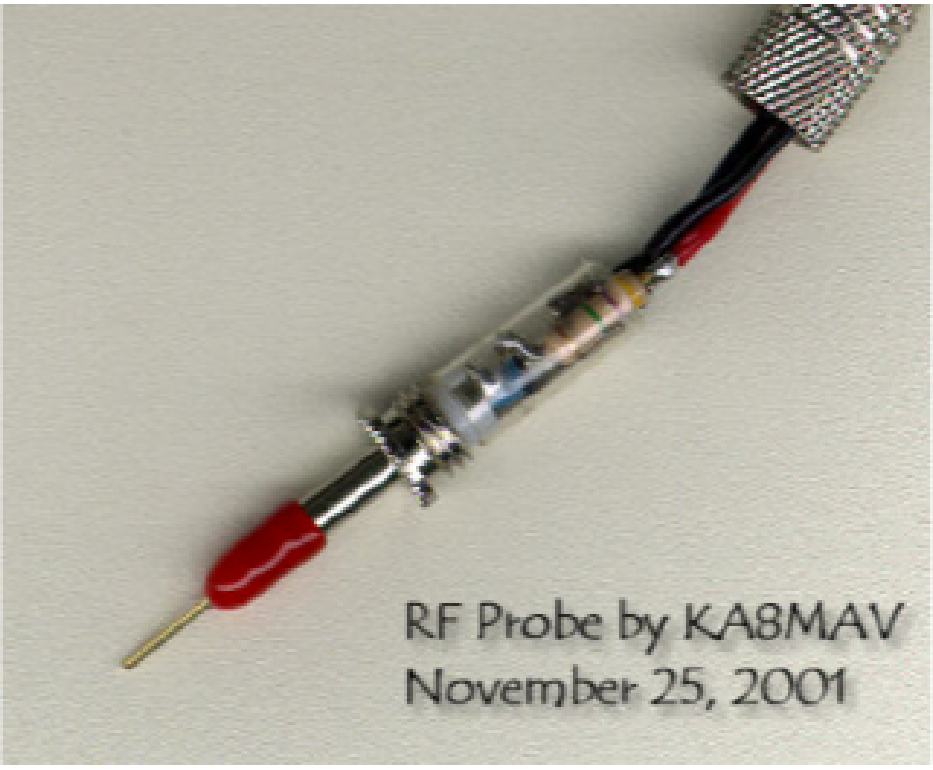

RF Probe by KA8MAV
November 25, 2001

Germanium Diode

IN34 or IN34A

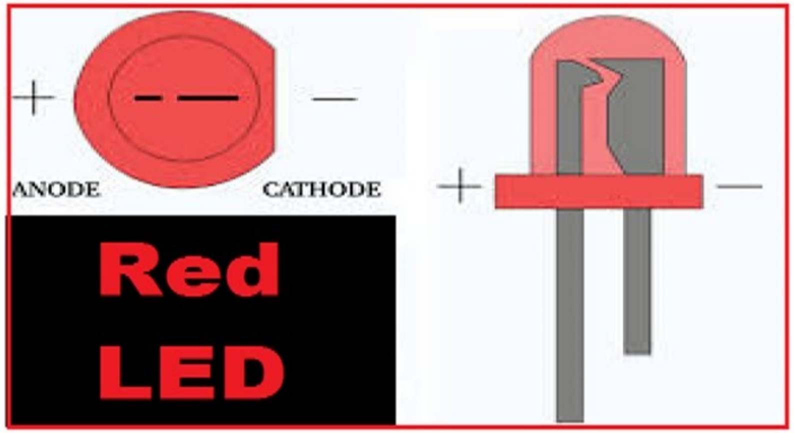
+
−
ANODE
CATHODE
+
−

Red
LED

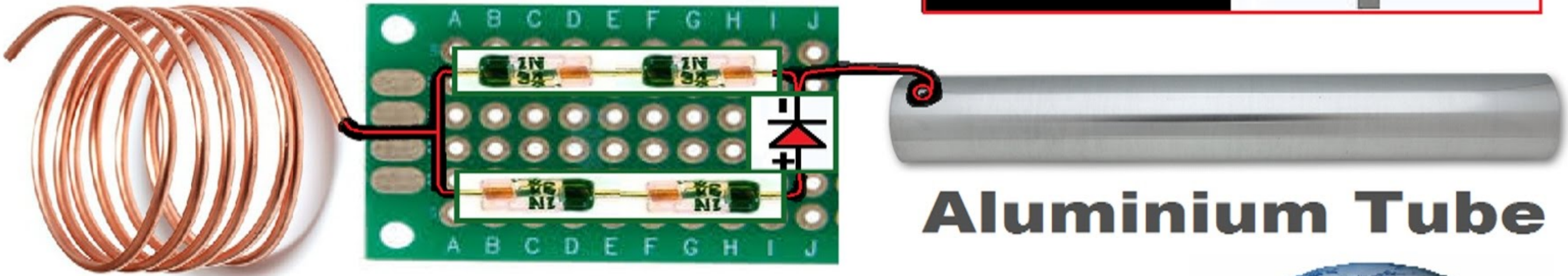
A B C D E F G H I J
1N
34
-
+
A B C D E F G H I J

Aluminium Tube

copper
wire

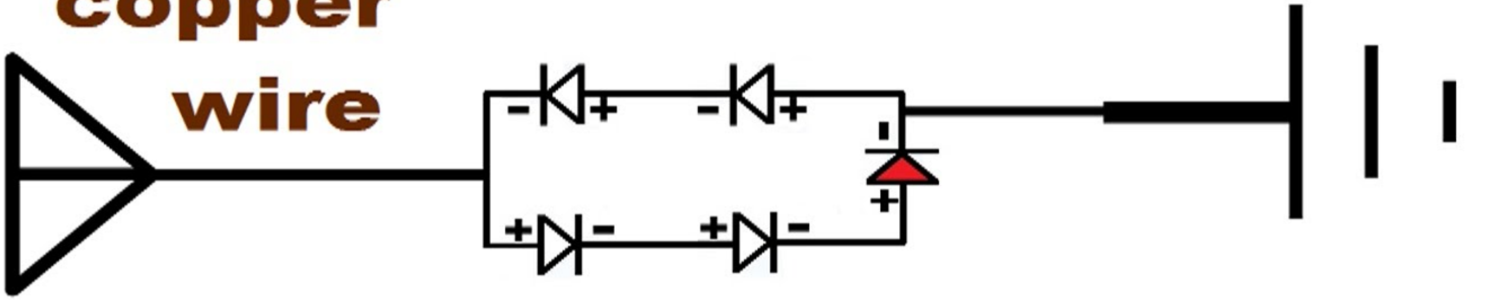

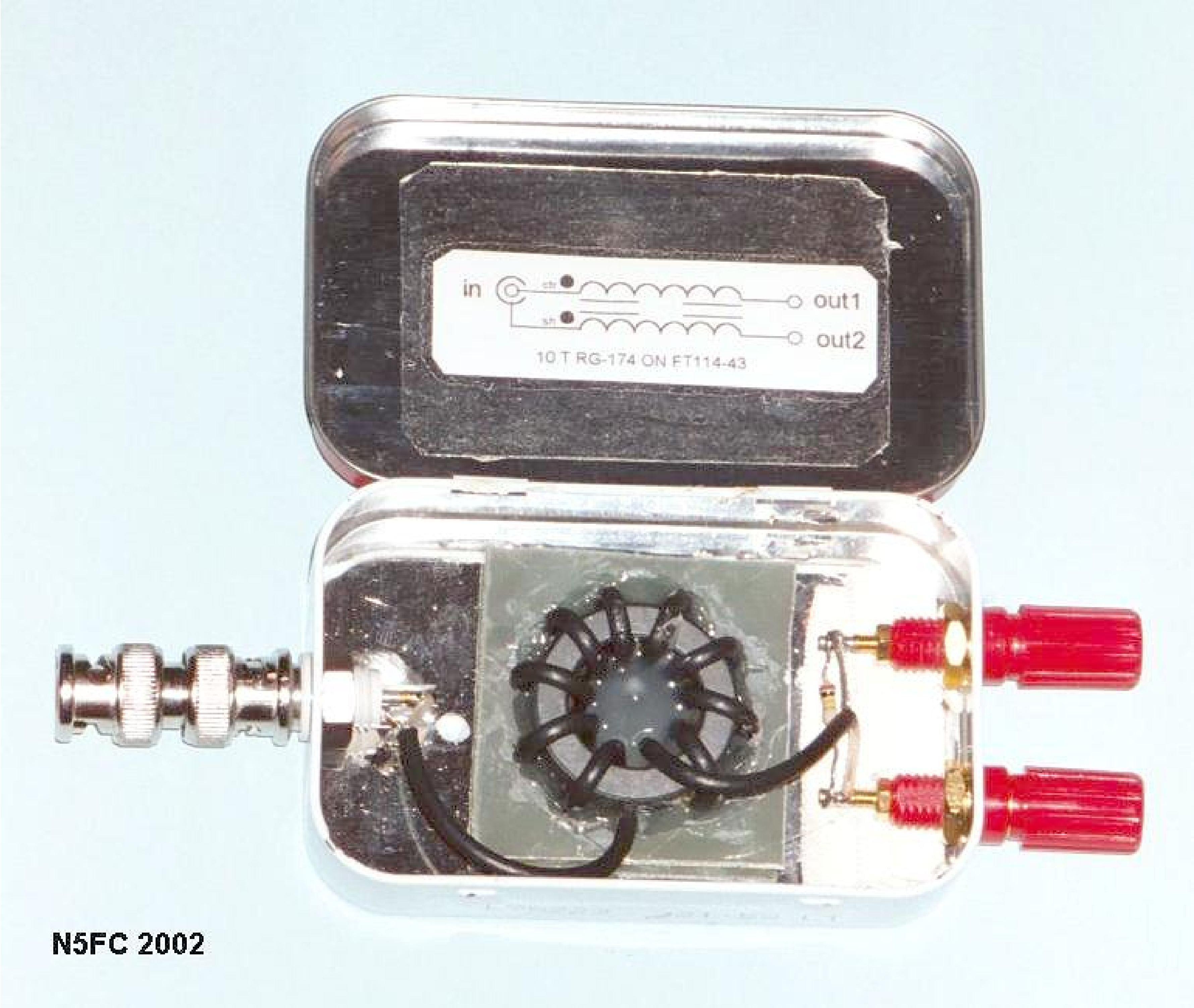
in
ctr
sh
out1
out2
10 T RG-174 ON FT114-43

N5FC 2002

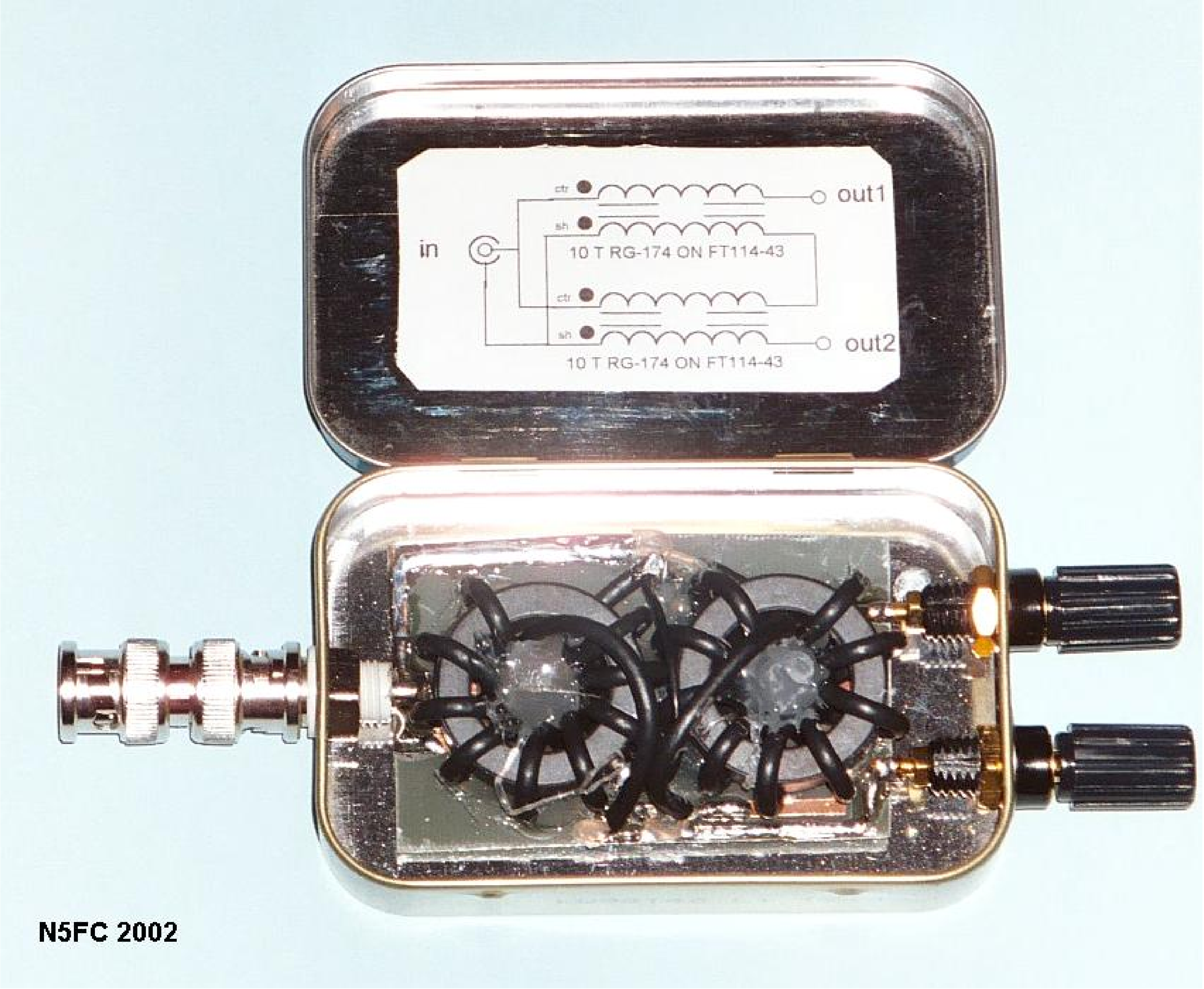

ctr
sh
ctr
sh
in
out1
out2
10 T RG-174 ON FT114-43
10 T RG-174 ON FT114-43

N5FC 2002

N5FC 2001

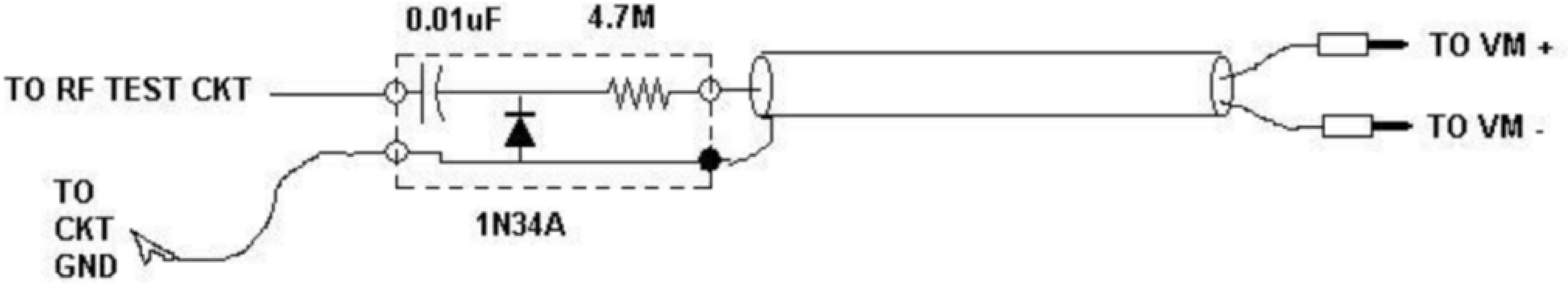
0.01uF
4.7M
TO RF TEST CKT
TO VM +
TO VM -
TO
CKT
GND
1N34A

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10 -11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

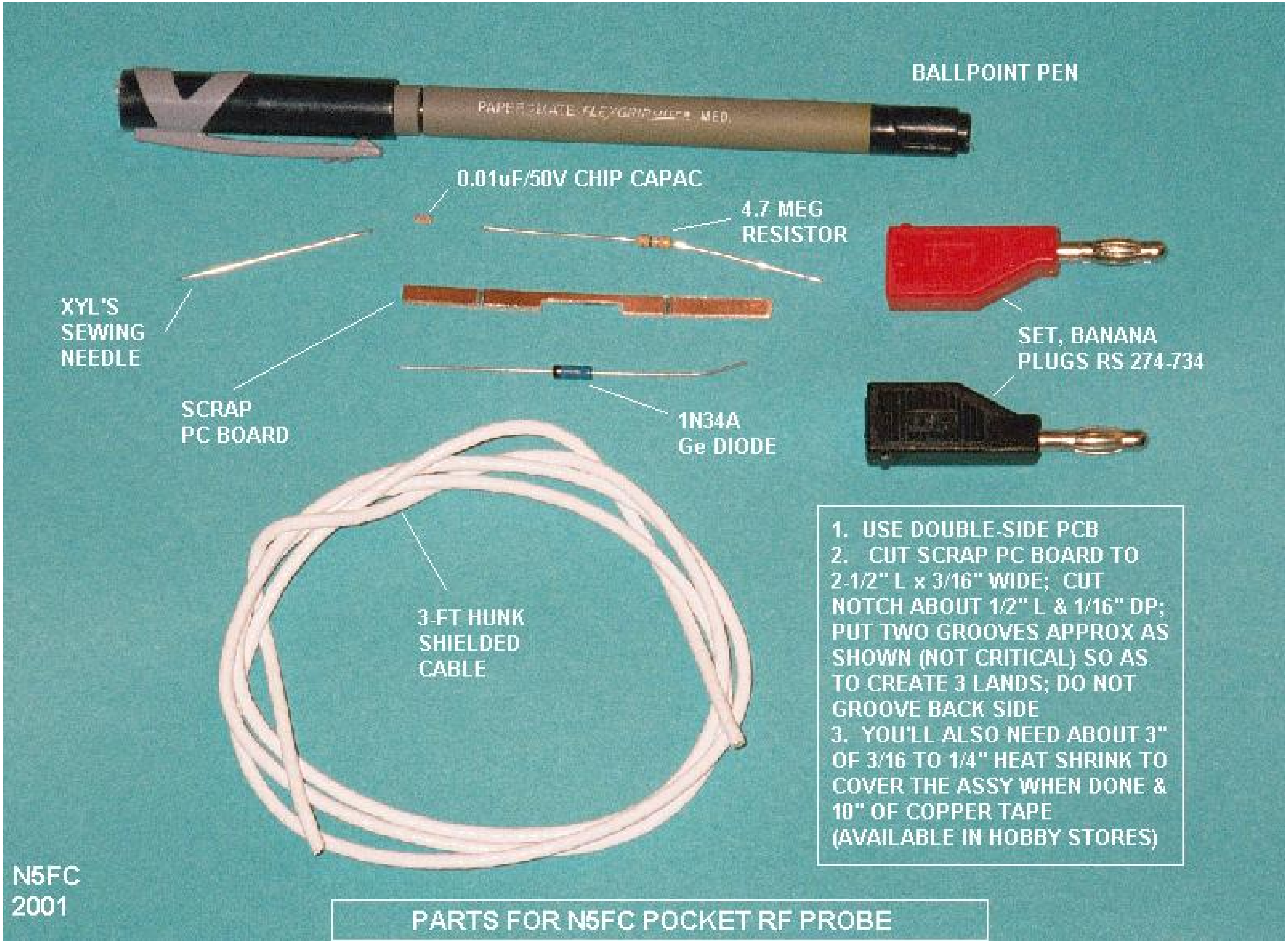
BALLPOINT PEN
PAPERMATE FLEXGRIP ULTRA MED.
0.01uF/50V CHIP CAPAC
4.7 MEG RESISTOR
XYL'S SEWING NEEDLE
SET, BANANA PLUGS RS 274-734
SCRAP PC BOARD
1N34A Ge DIODE
3-FT HUNK SHIELDED CABLE
N5FC 2001

PARTS FOR N5FC POCKET RF PROBE

1. USE DOUBLE-SIDE PCB
2. CUT SCRAP PC BOARD TO 2-1/2" L x 3/16" WIDE; CUT NOTCH ABOUT 1/2" L & 1/16" DP; PUT TWO GROOVES APPROX AS SHOWN (NOT CRITICAL) SO AS TO CREATE 3 LANDS; DO NOT GROOVE BACK SIDE
3. YOU'LL ALSO NEED ABOUT 3" OF 3/16 TO 1/4" HEAT SHRINK TO COVER THE ASSY WHEN DONE & 10" OF COPPER TAPE (AVAILABLE IN HOBBY STORES)

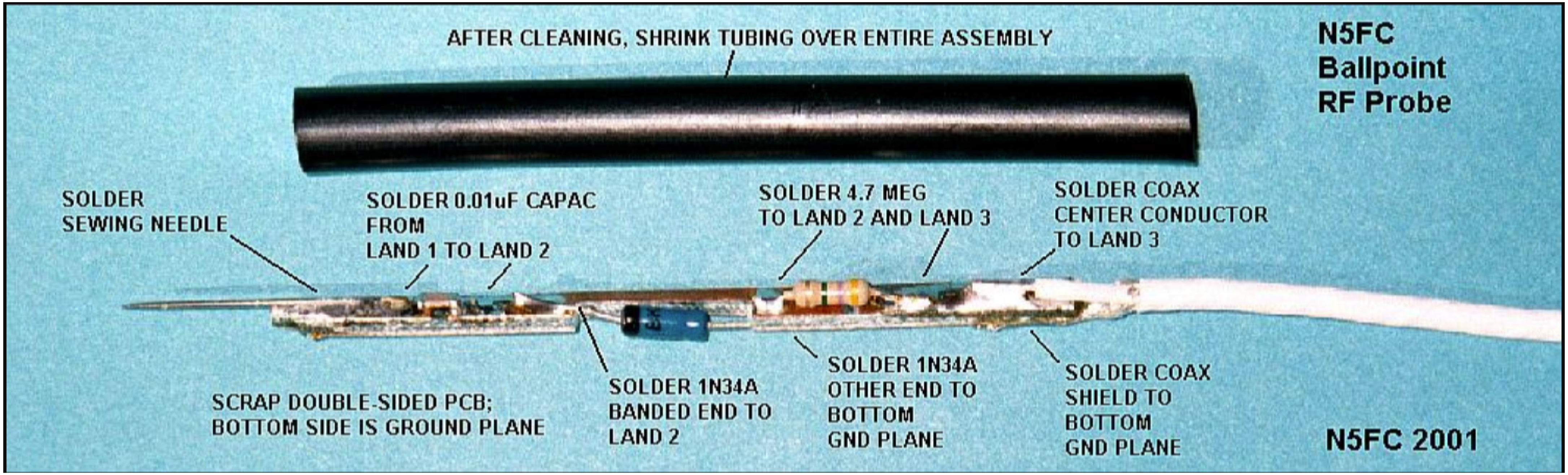
N5FC
Ballpoint
RF Probe
AFTER CLEANING, SHRINK TUBING OVER ENTIRE ASSEMBLY
SOLDER SEWING NEEDLE
SOLDER 0.01uF CAPAC FROM LAND 1 TO LAND 2
SOLDER 4.7 MEG TO LAND 2 AND LAND 3
SOLDER COAX CENTER CONDUCTOR TO LAND 3
SCRAP DOUBLE-SIDED PCB; BOTTOM SIDE IS GROUND PLANE
SOLDER 1N34A BANDED END TO LAND 2
SOLDER 1N34A OTHER END TO BOTTOM GND PLANE
SOLDER COAX SHIELD TO BOTTOM GND PLANE
N5FC 2001

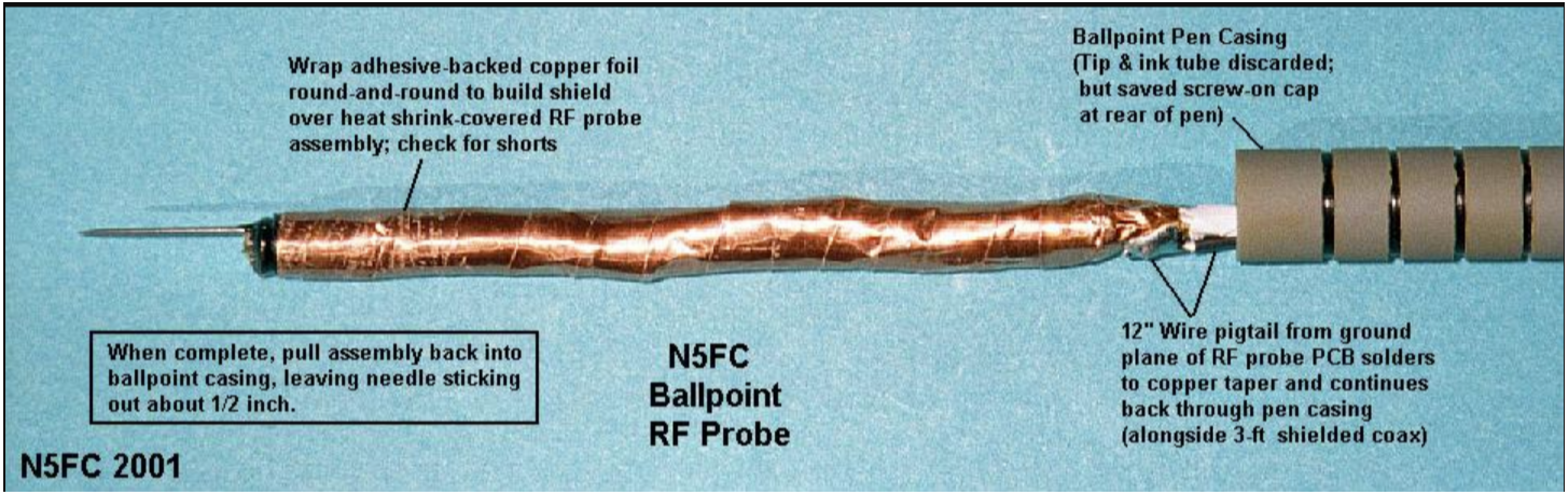

Wrap adhesive-backed copper foil round-and-round to build shield over heat shrink-covered RF probe assembly; check for shorts

Ballpoint Pen Casing (Tip & ink tube discarded; but saved screw-on cap at rear of pen)

When complete, pull assembly back into ballpoint casing, leaving needle sticking out about 1/2 inch.

N5FC
Ballpoint
RF Probe

12" Wire pigtail from ground plane of RF probe PCB solders to copper taper and continues back through pen casing (alongside 3-ft shielded coax)

N5FC 2001

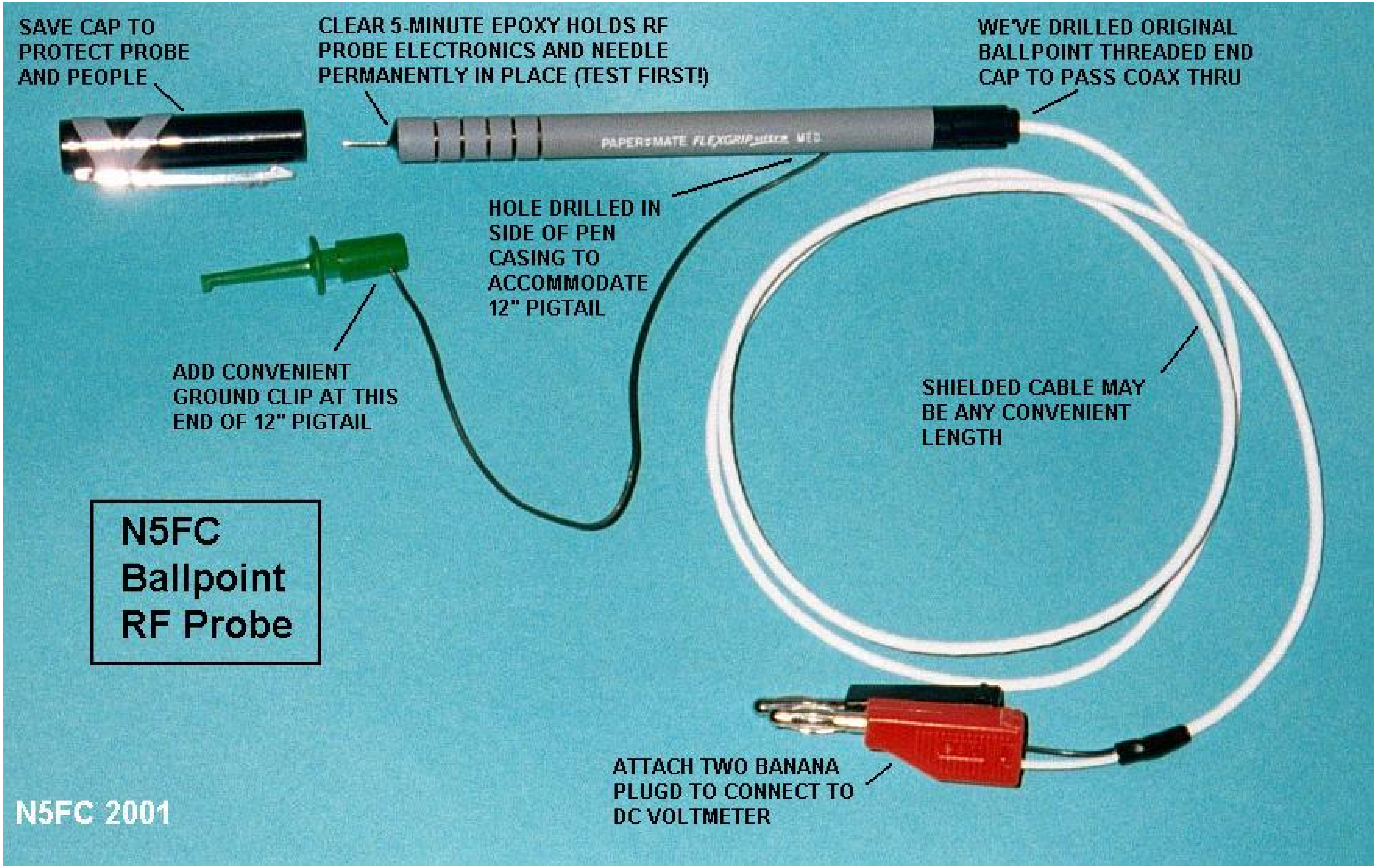

N5FC
Ballpoint
RF Probe

SAVE CAP TO
PROTECT PROBE
AND PEOPLE

CLEAR 5-MINUTE EPOXY HOLDS RF
PROBE ELECTRONICS AND NEEDLE
PERMANENTLY IN PLACE (TEST FIRST!)

WE'VE DRILLED ORIGINAL
BALLPOINT THREADED END
CAP TO PASS COAX THRU

HOLE DRILLED IN
SIDE OF PEN
CASING TO
ACCOMMODATE
12" PIGTAIL

ADD CONVENIENT
GROUND CLIP AT THIS
END OF 12" PIGTAIL

SHIELDED CABLE MAY
BE ANY CONVENIENT
LENGTH

ATTACH TWO BANANA
PLUGD TO CONNECT TO
DC VOLTMETER

N5FC 2001

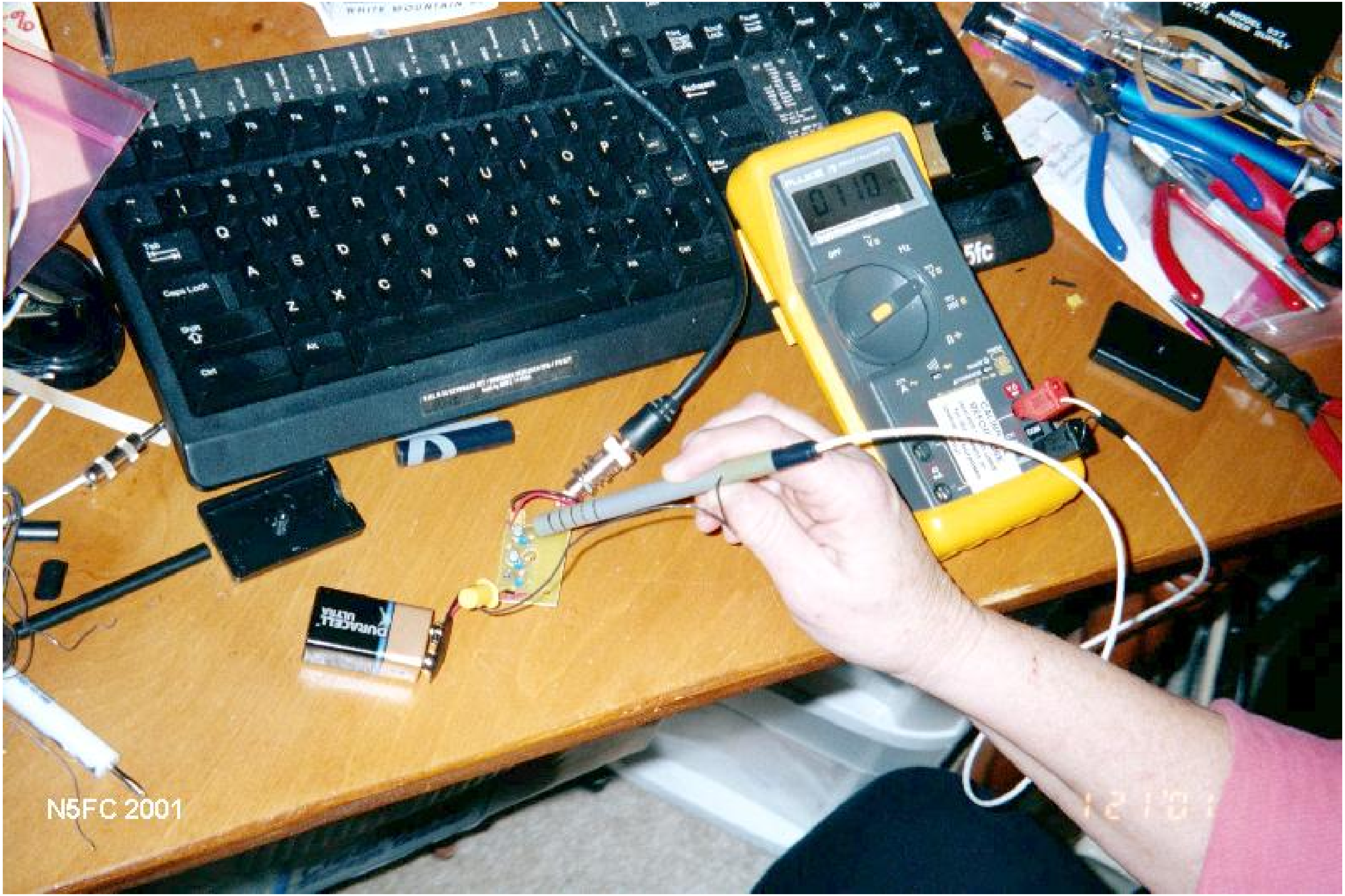
N5FC 2001
1 21 01
DURACELL ULTRA
0110

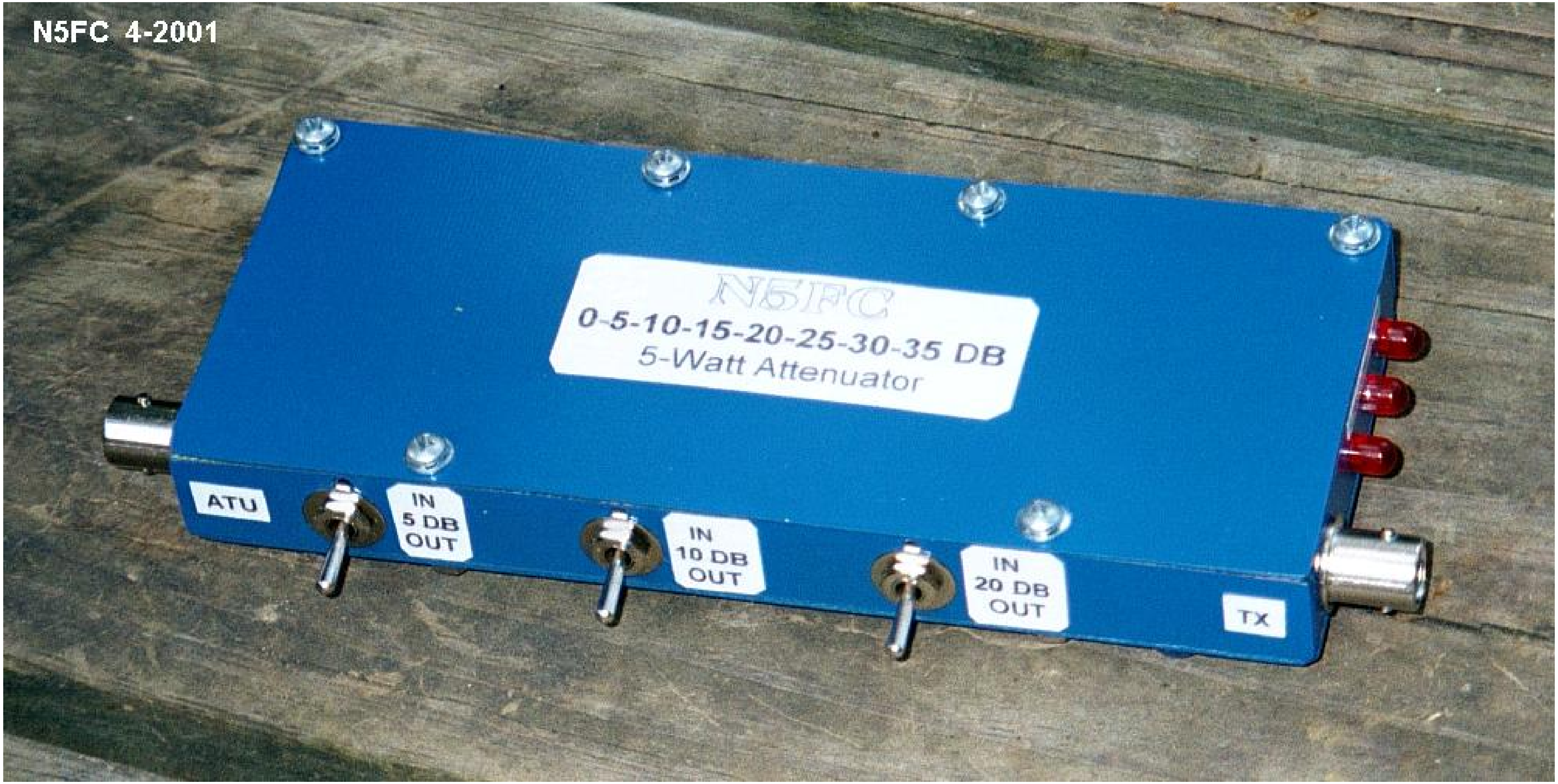

N5FC 4-2001

N5FC
0-5-10-15-20-25-30-35 DB
5-Watt Attenuator
ATU
IN
5 DB
OUT
IN
10 DB
OUT
IN
20 DB
OUT
TX

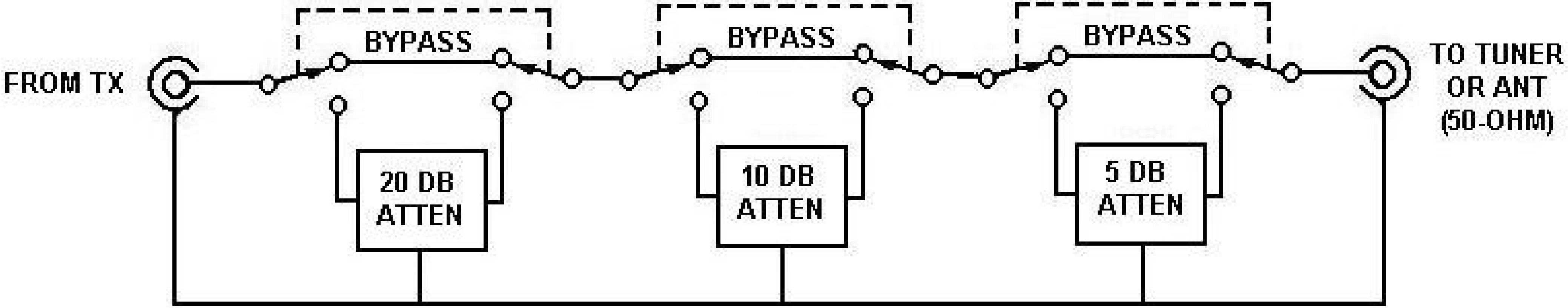
FROM TX
BYPASS
BYPASS
BYPASS
20 DB ATTEN
10 DB ATTEN
5 DB ATTEN
TO TUNER OR ANT (50-OHM)

N5FC QRP Switchable 0-5-10-15-20-25-30-35 db Attenuator

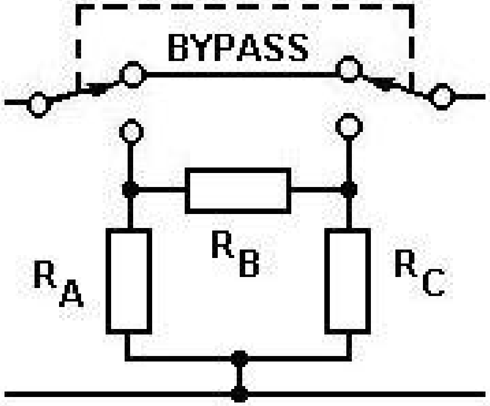
BYPASS
R_A
R_B
R_C

ATTENUATOR SECTION CIRCUIT

		NOM'L	ACT'L	IMPLEMENT WITH
20 db (NOM) ACT=19.7db	R_A	61.1	67	6 ea 100/1W SERIES-PAR'L & 1K PAR'L
	R_B	247.5	235	2 ea 470 / 0.5W IN PARALLEL
	R_C	61.1	62	68/0.5W IN PAR'L with 680/0.5W
10 db (NOM) ACT = 9.7 db	R_A	96.2	100	4 ea 100/1W in SERIES-PARALLEL
	R_B	71.2	67	100/1W IN PAR'L w: 2 ea 100/.5W IN SER
	R_C	96.2	100	100/0.5W
5 db (NOM) ACT = 4.9 db	R_A	247.5	200	2 ea 100/1W IN SERIES
	R_B	30.4	30	3 ea 10/1W IN SERIES
	R_C	247.5	200	2 ea 100/0.5W IN SERIES

N5FC 4-2001

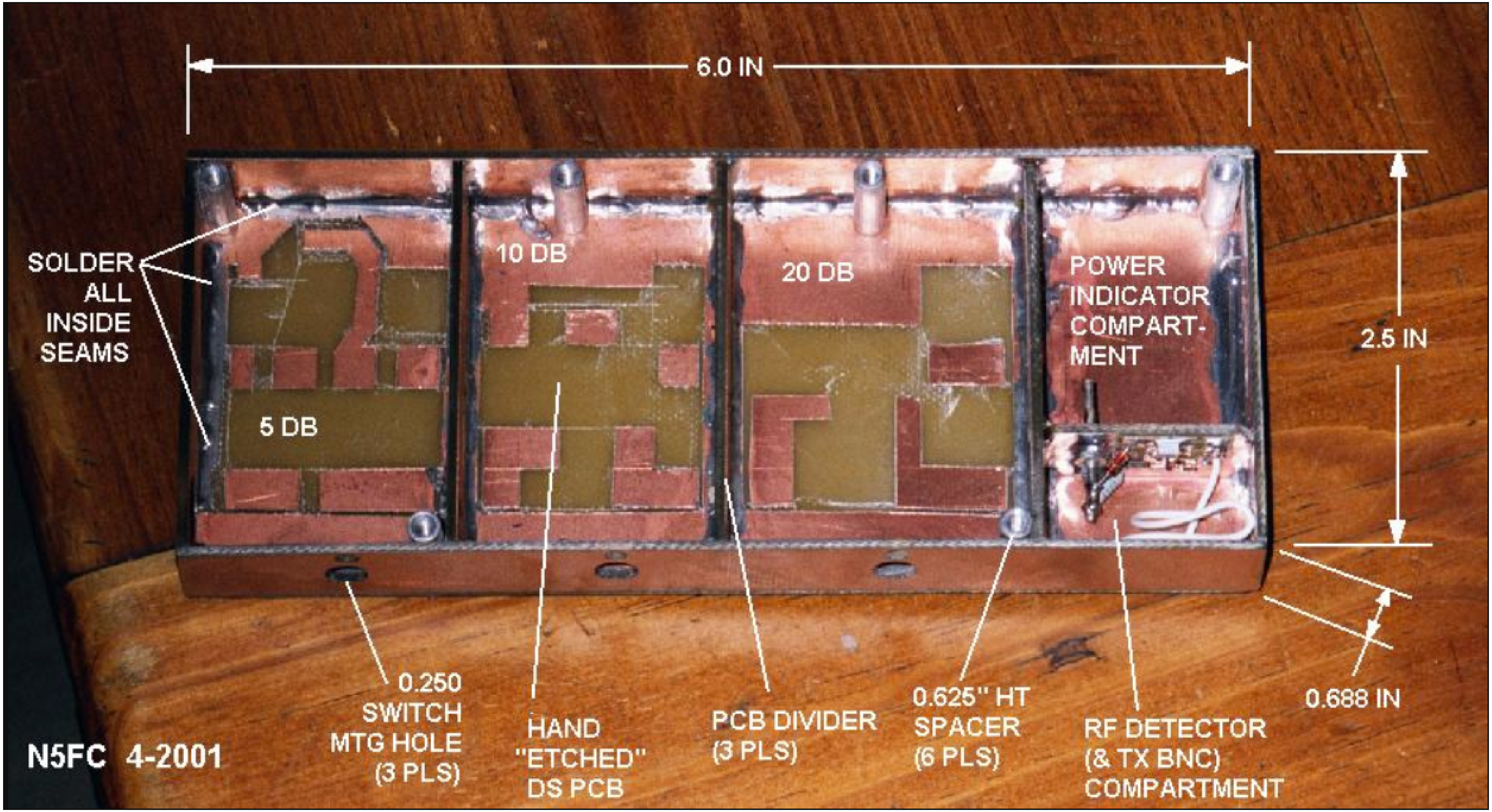
6.0 IN
2.5 IN
0.688 IN
SOLDER ALL INSIDE SEAMS
10 DB
20 DB
POWER INDICATOR COMPART-MENT
5 DB
0.250 SWITCH MTG HOLE (3 PLS)
HAND "ETCHED" DS PCB
PCB DIVIDER (3 PLS)
0.625'' HT SPACER (6 PLS)
RF DETECTOR (& TX BNC) COMPARTMENT
N5FC 4-2001

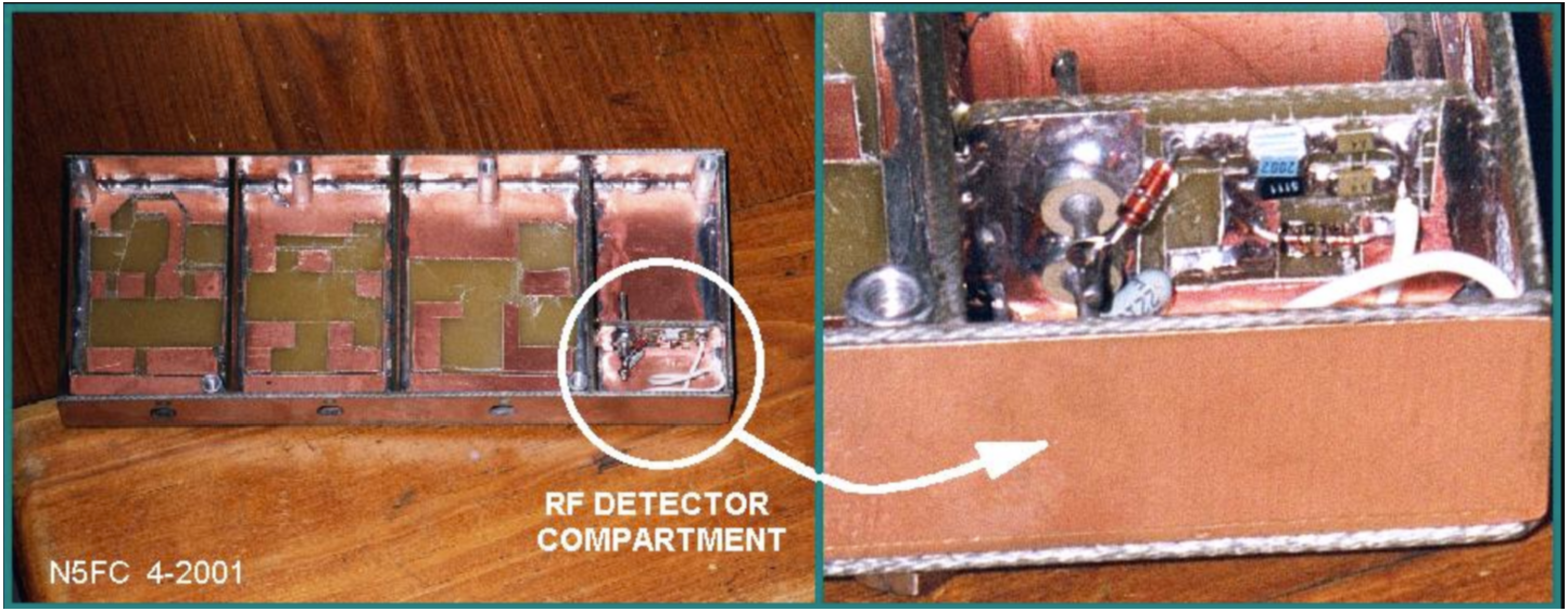
RF DETECTOR
COMPARTMENT
N5FC 4-2001

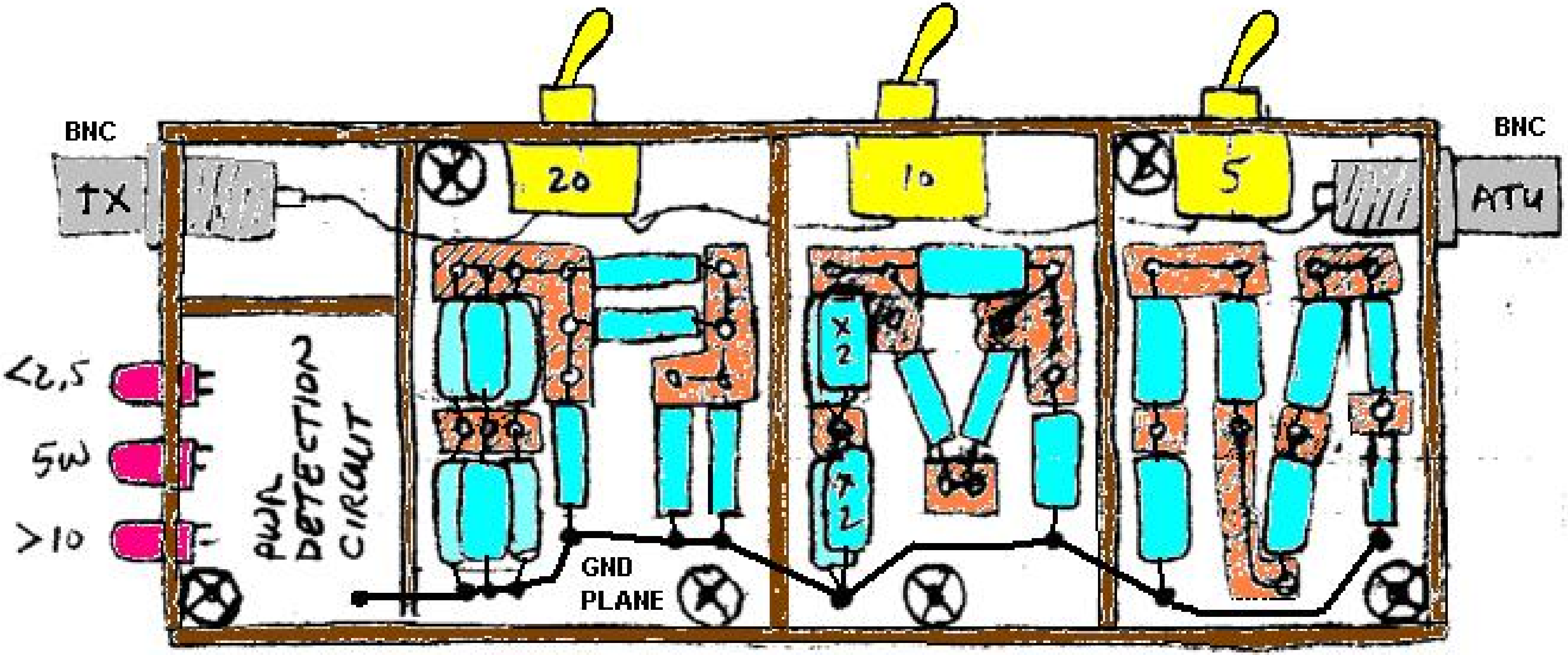
BNC
TX
20
10
5
BNC
ATU
<2,5
5W
>10
PWR DETECTION CIRCUIT
X2
X2
GND PLANE

N5FC 4-2001

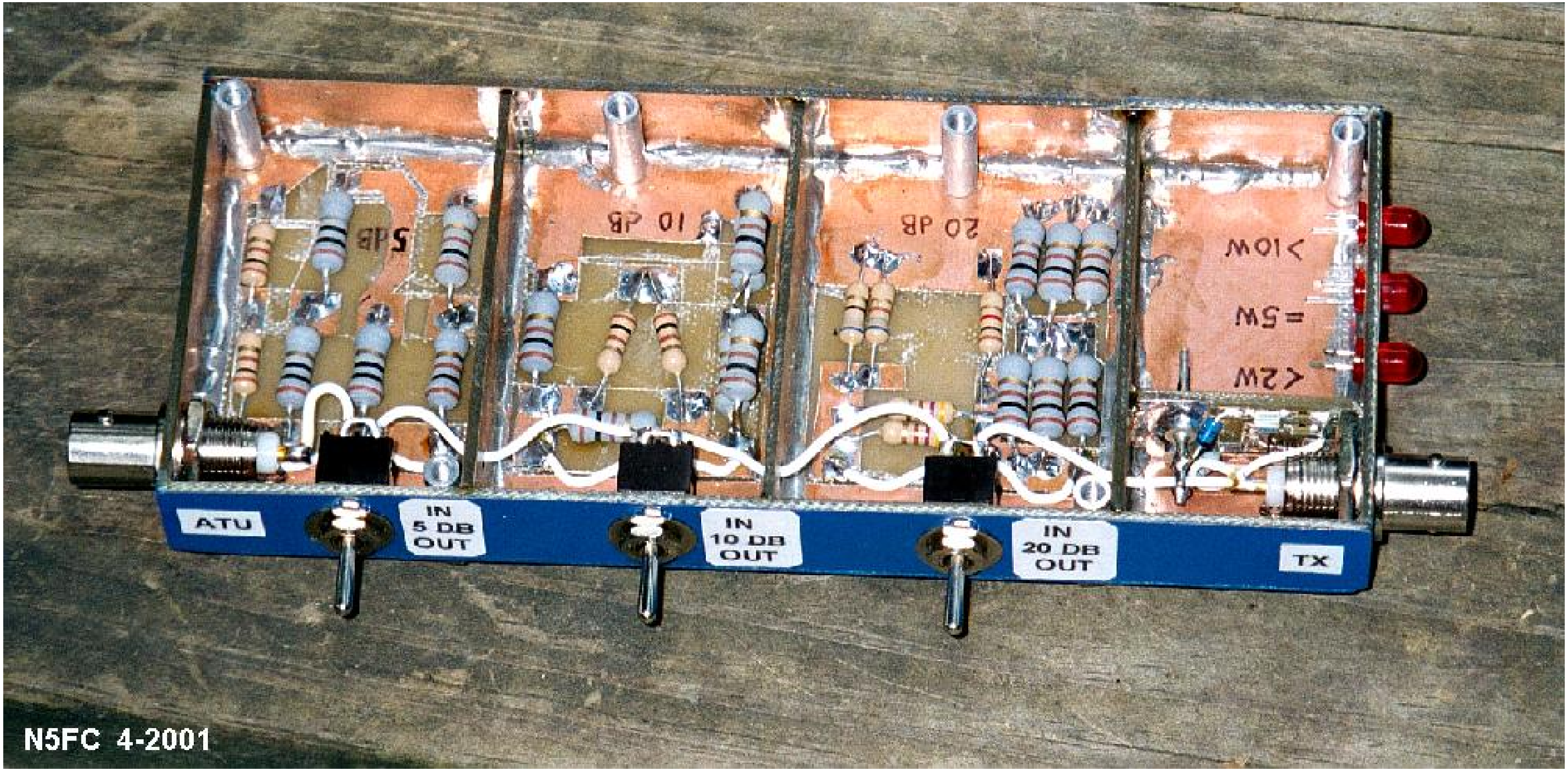
ATU
IN
5 DB
OUT
IN
10 DB
OUT
IN
20 DB
OUT
TX
5dB
10 dB
20 dB
>10W
=5W
<2W

N5FC 4-2001

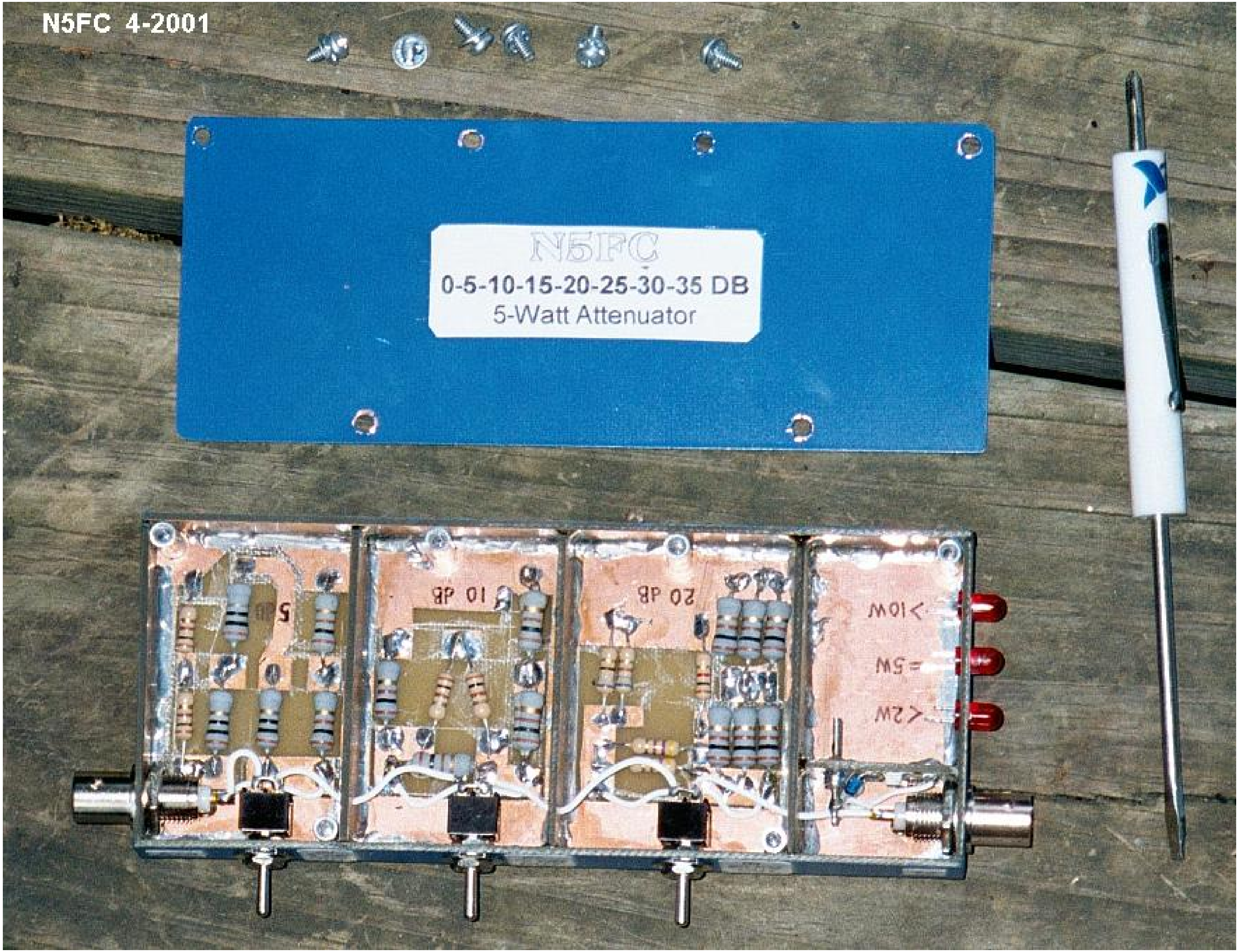
N5FC 4-2001

N5FC
0-5-10-15-20-25-30-35 DB
5-Watt Attenuator
5dB
10 dB
20 dB
>10W
=5W
<2W

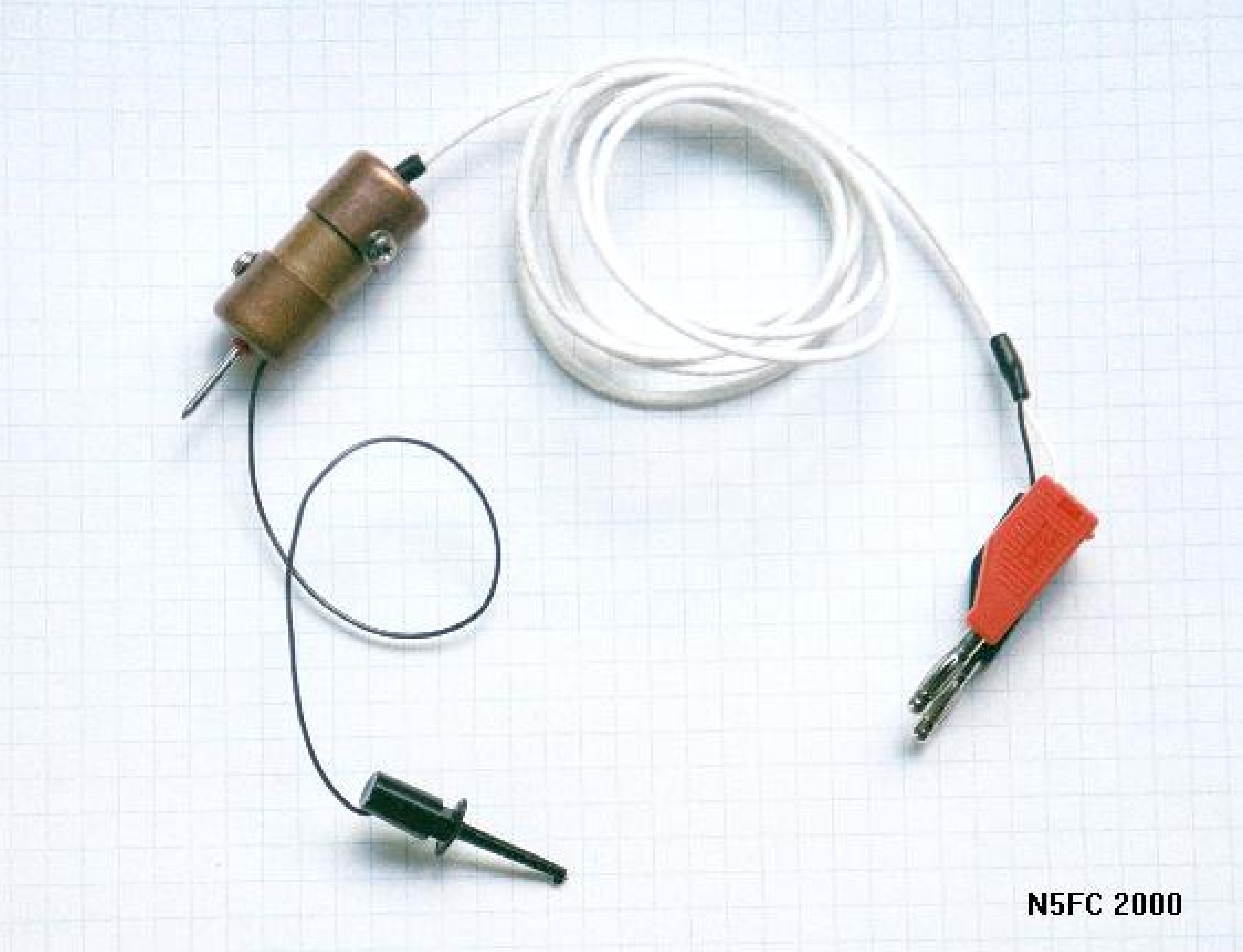
N5FC 2000

sic RF Probe. Simple, eh?

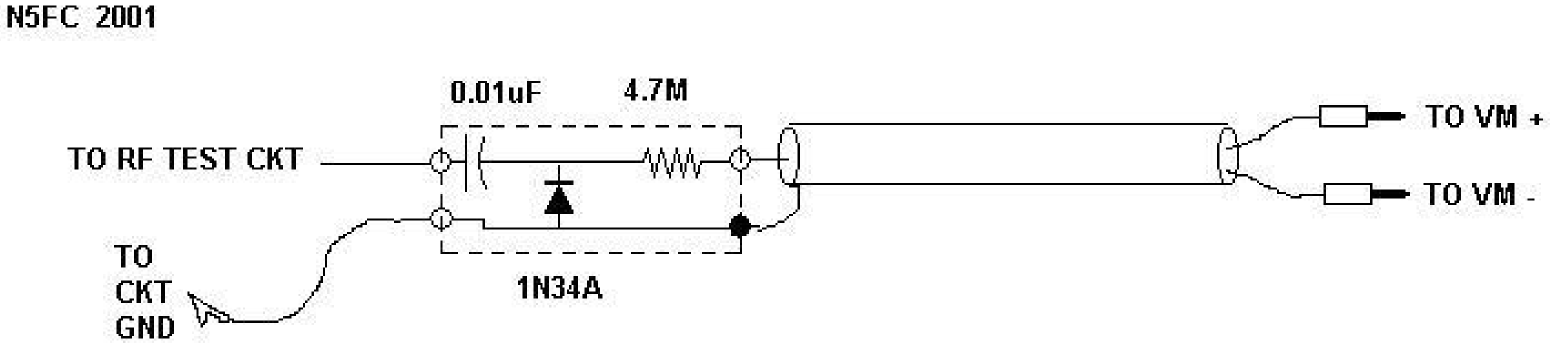
N5FC 2001
0.01uF
4.7M
TO RF TEST CKT
TO VM +
TO VM -
TO
CKT
GND
1N34A

CLASSIC RF PROBE

Reads RMS Equivalent Voltage in test circuit, if Voltmeter is 10 -11 Meg Input Impedance;
Reads 4X RMS Equiv Voltage if VM is 1Meg Input Impedance (Set VM to measure DCV)

retical discussion that we'll make short note of. Obviously, for "probing" we need a "probe". (Hey! No wonder I get paid the big bucks...). We add a SH
be goes to our test circuit, where we're probing. Brilliant! We don't want either of these to be long leads, because we're talking RF here, and long leads =

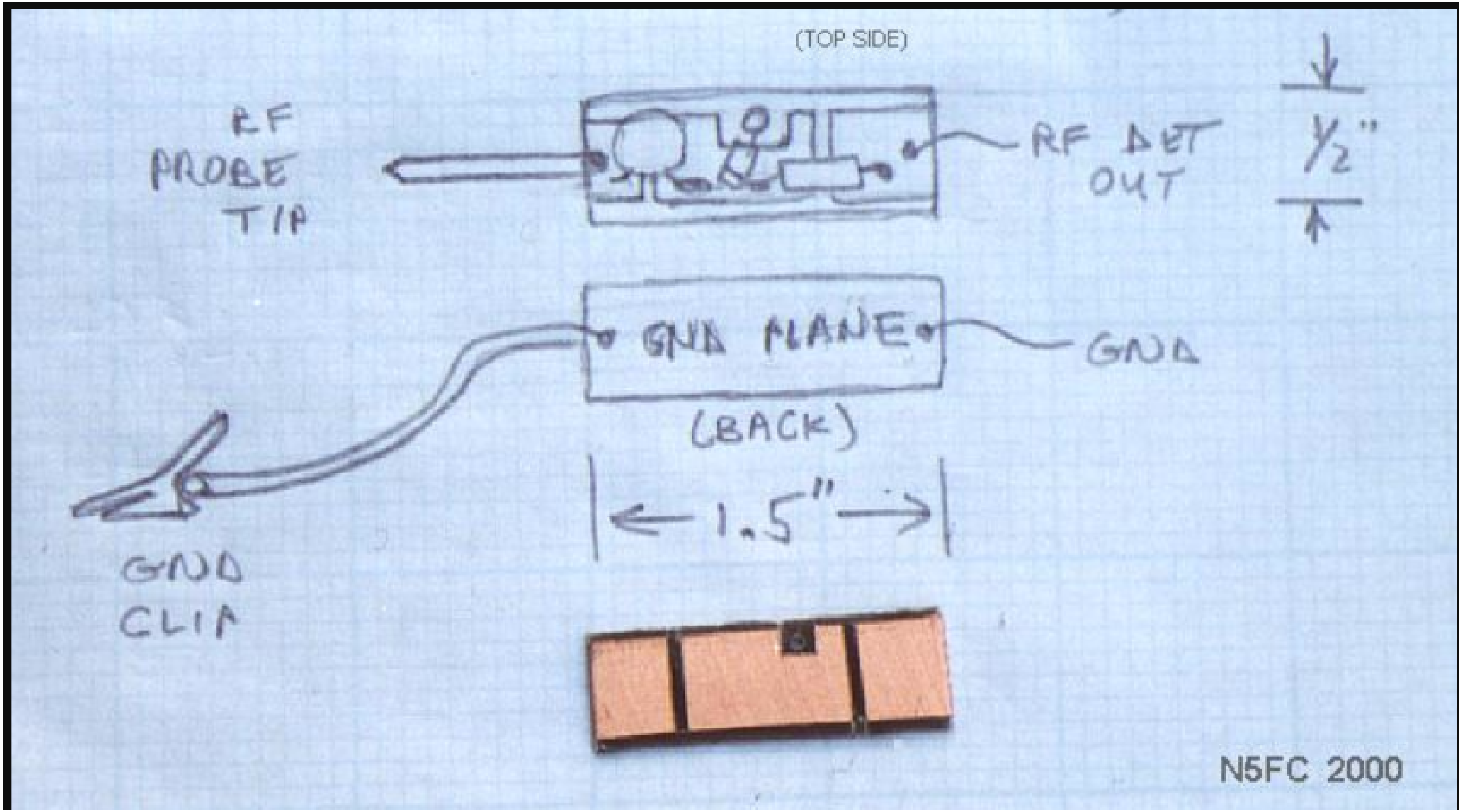
(TOP SIDE)
RF PROBE TIP
RF DET OUT
1/2"
GND PLANE
GND
(BACK)
1.5"
GND CLIP
N5FC 2000

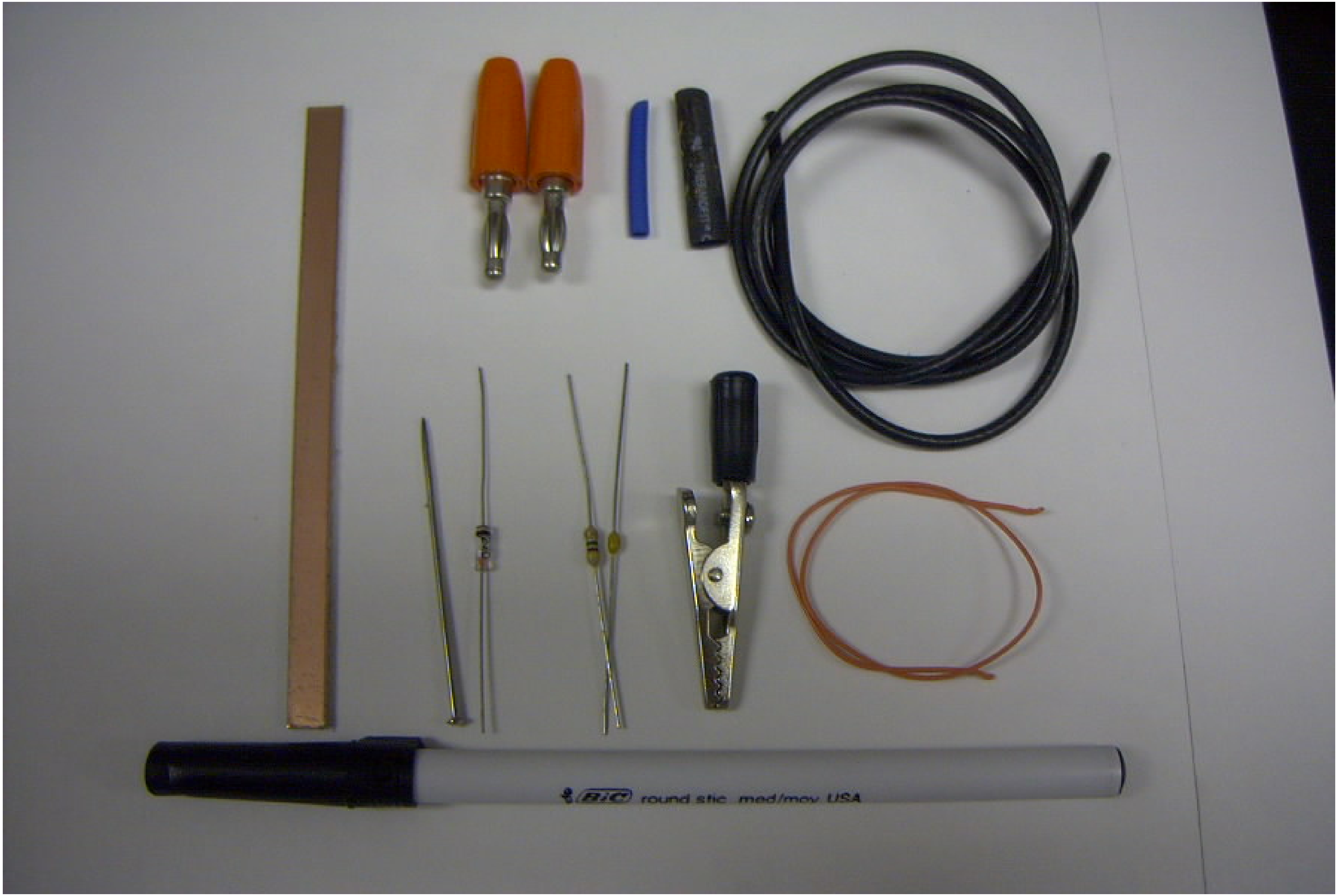
BIC round stic med/moy USA

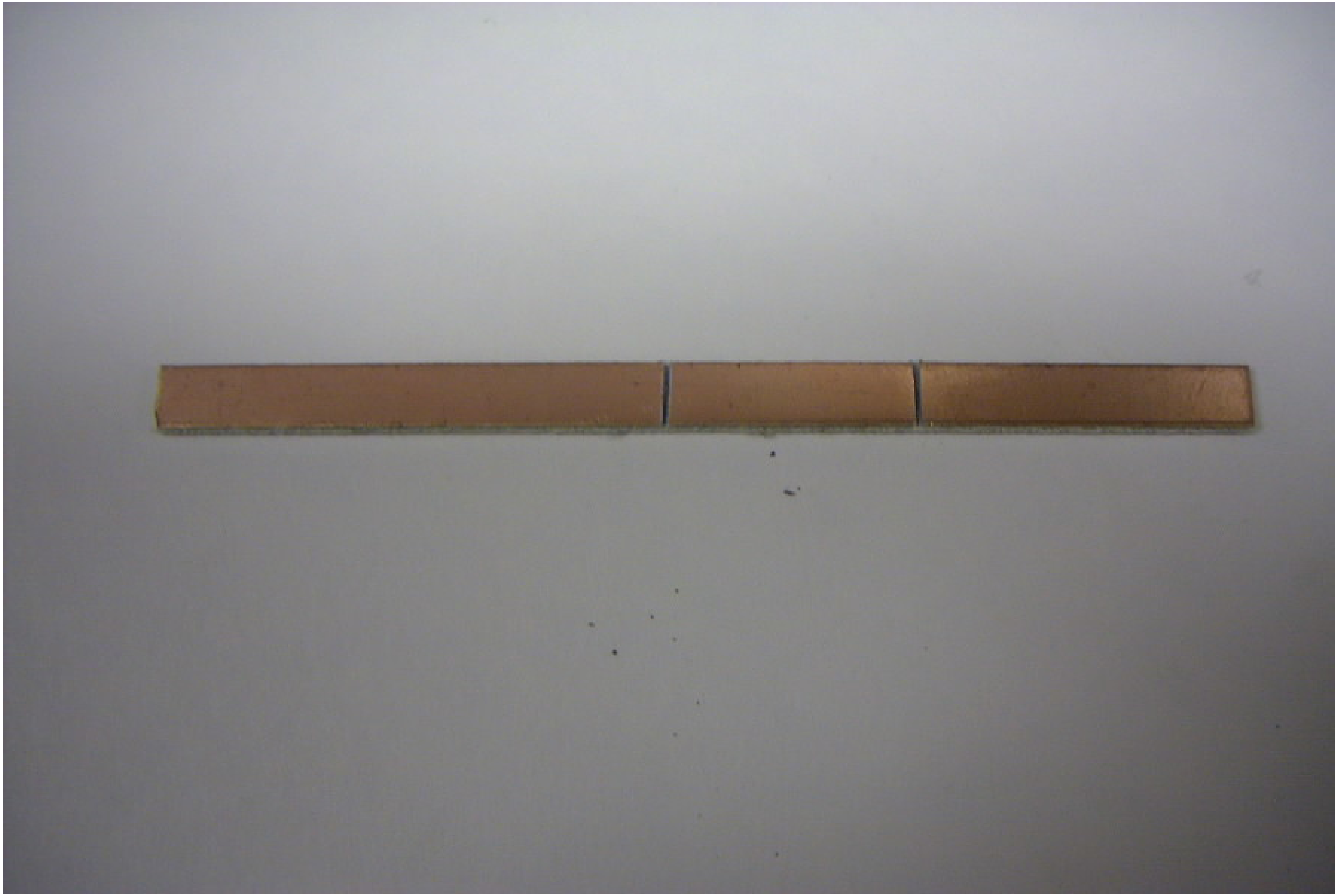

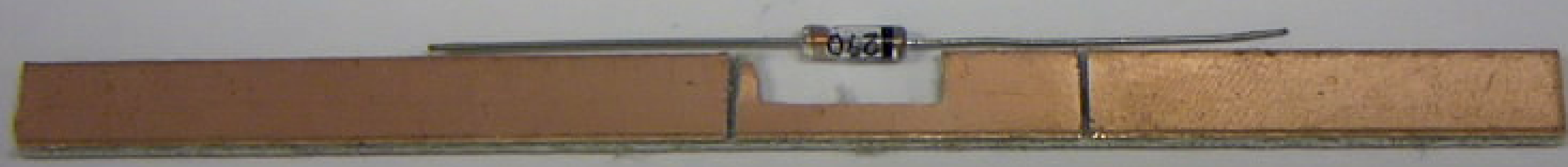
250

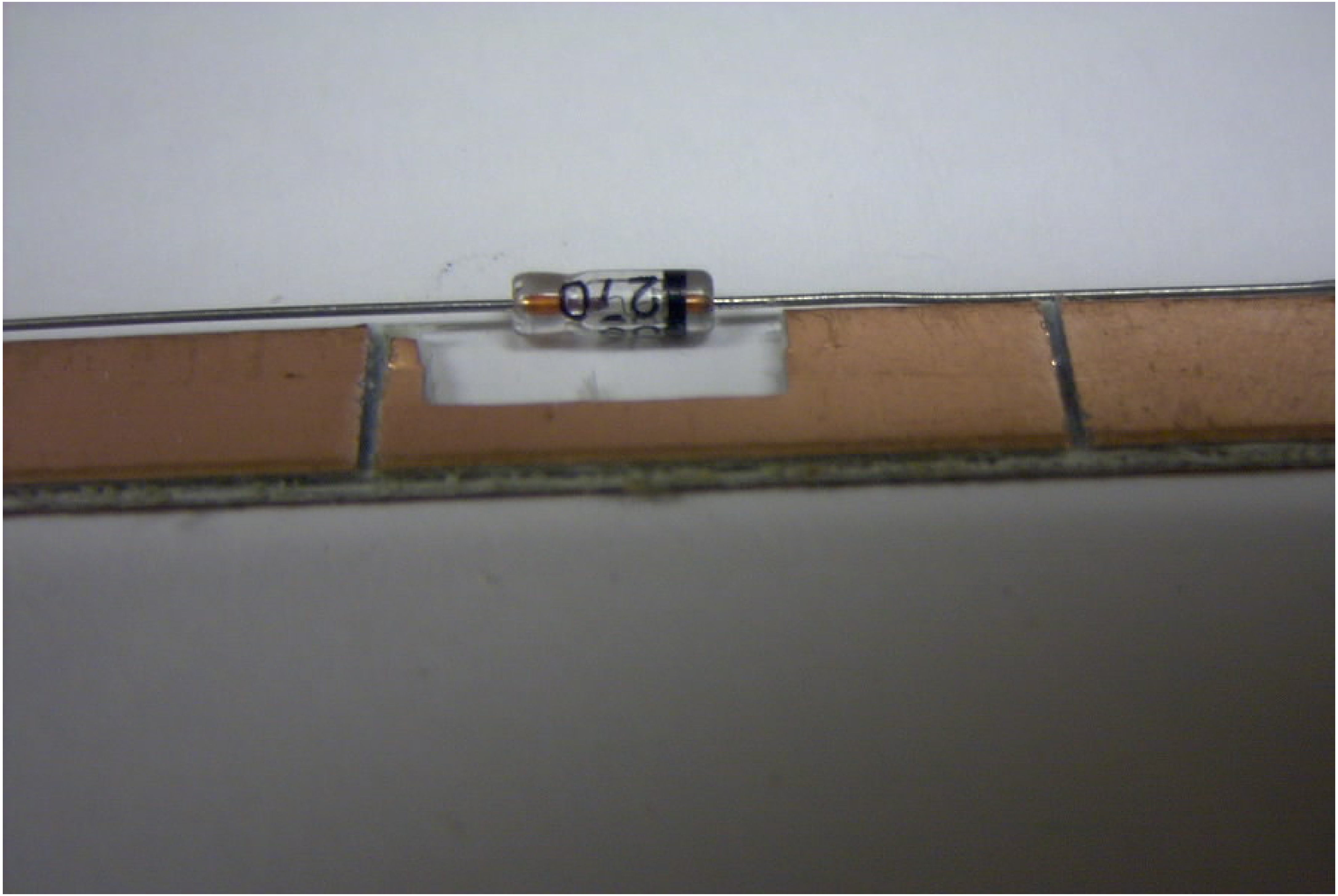

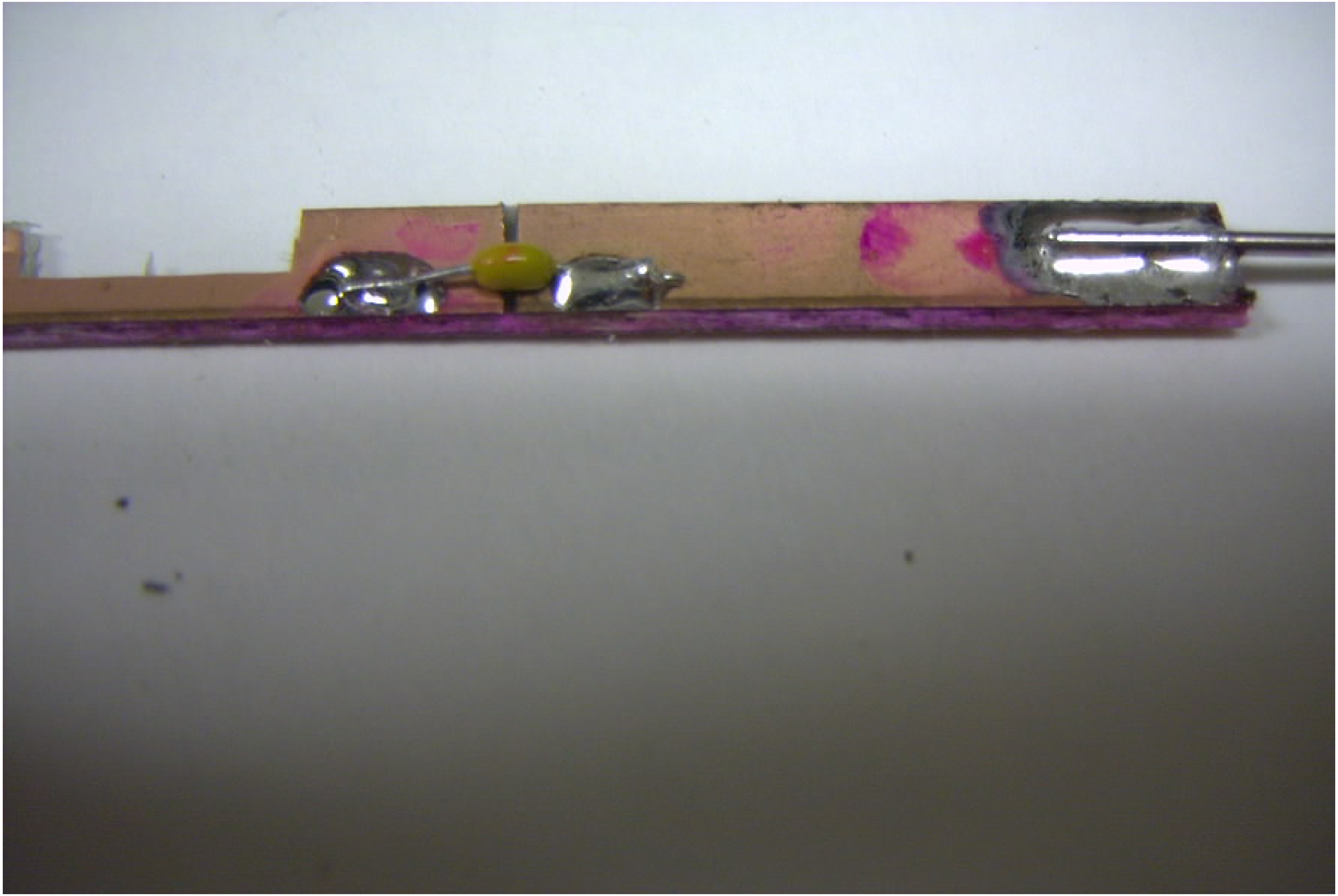

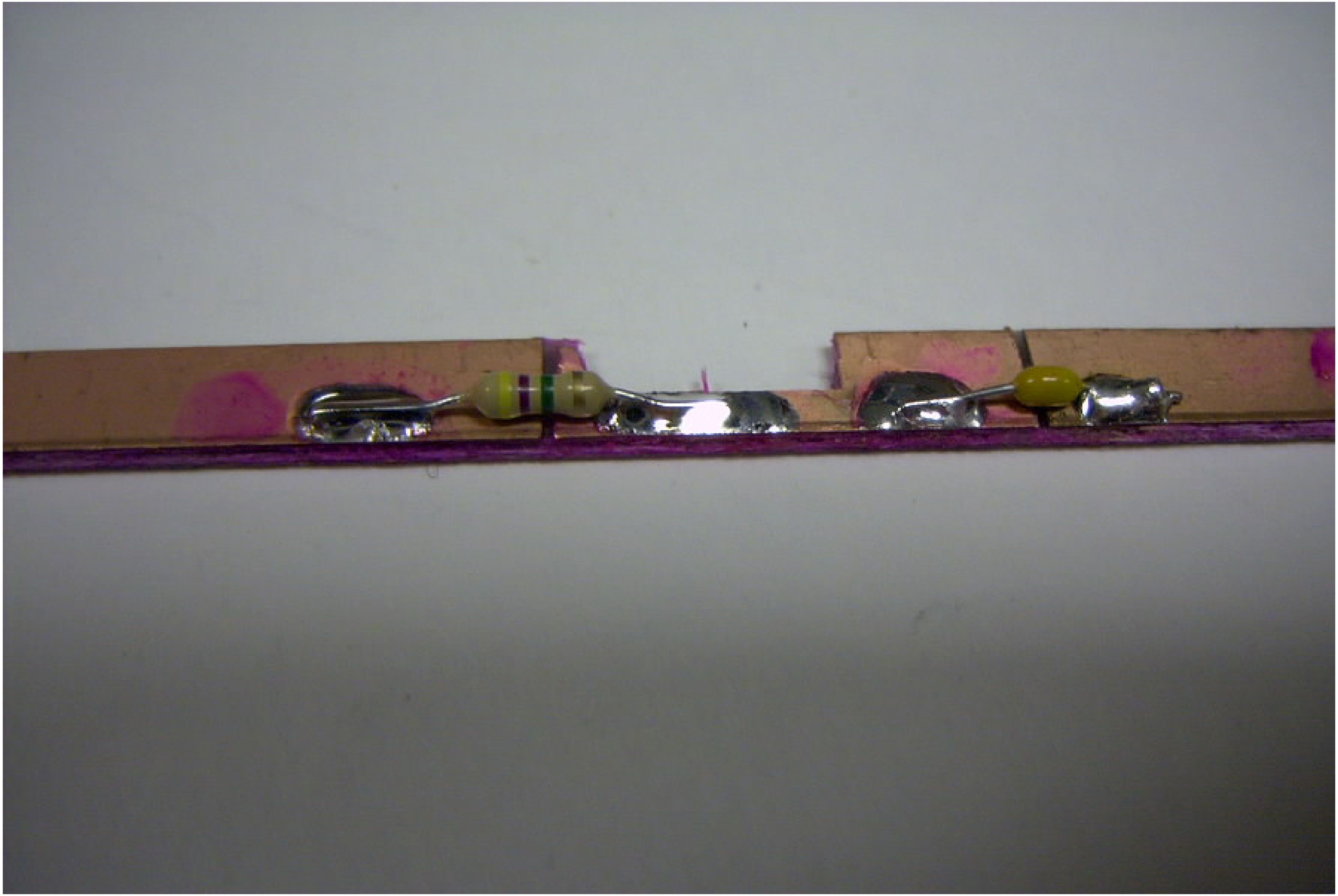

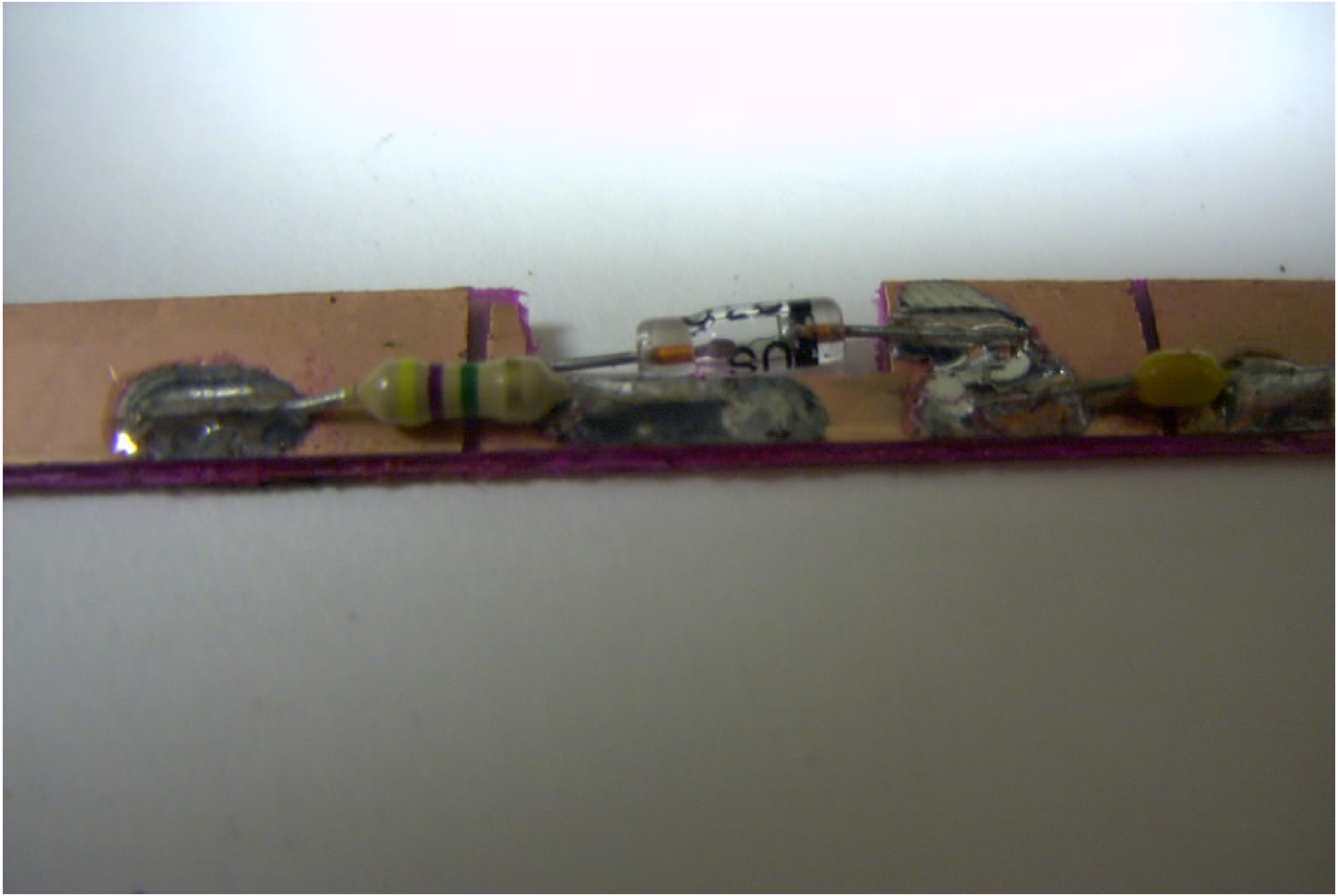

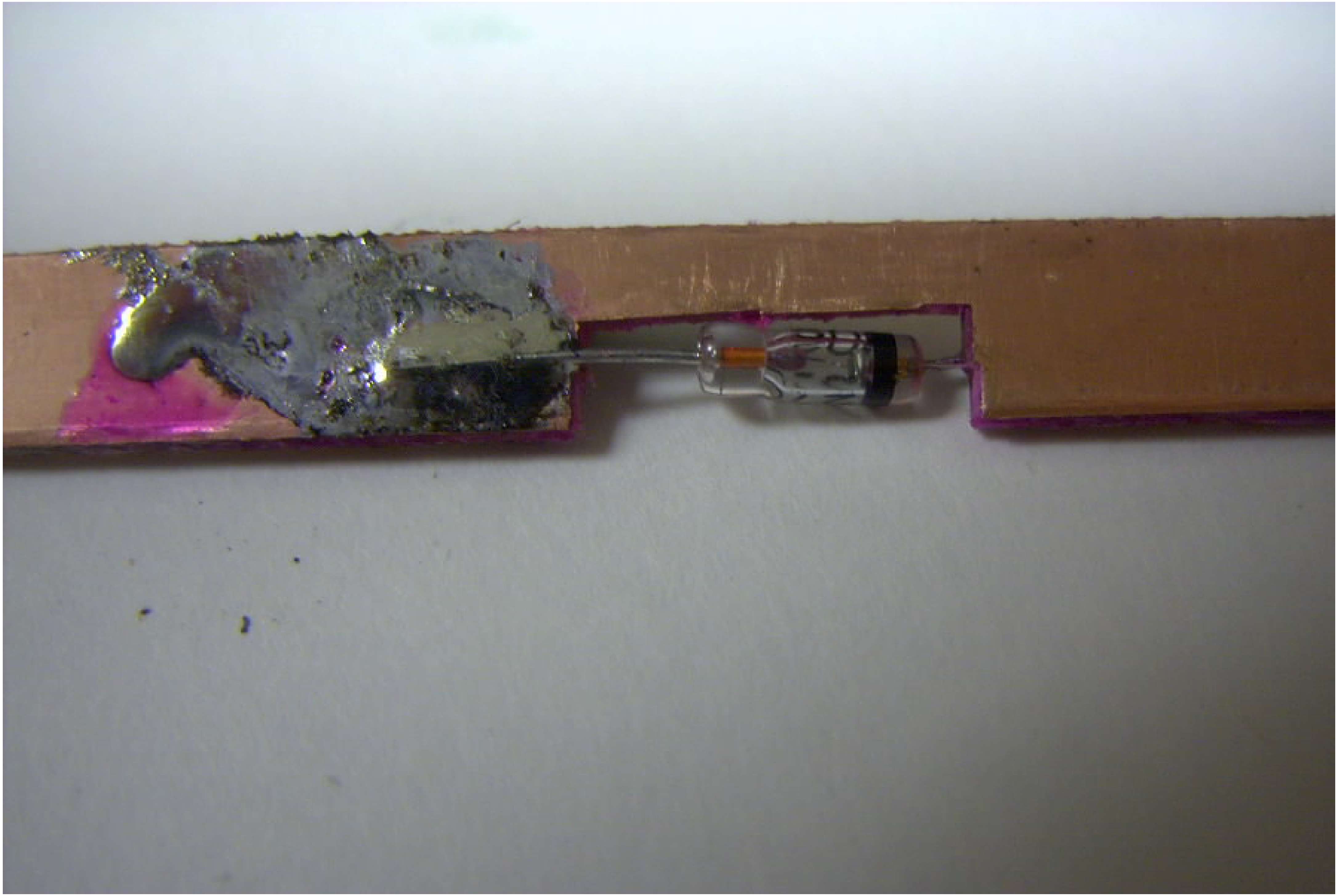

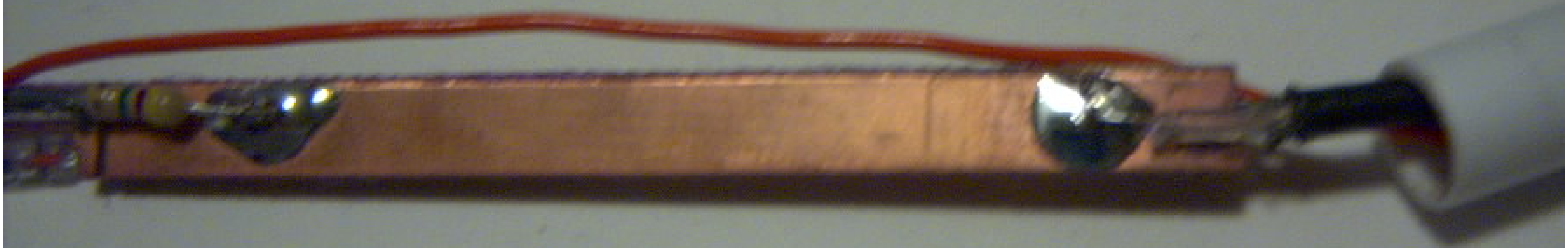

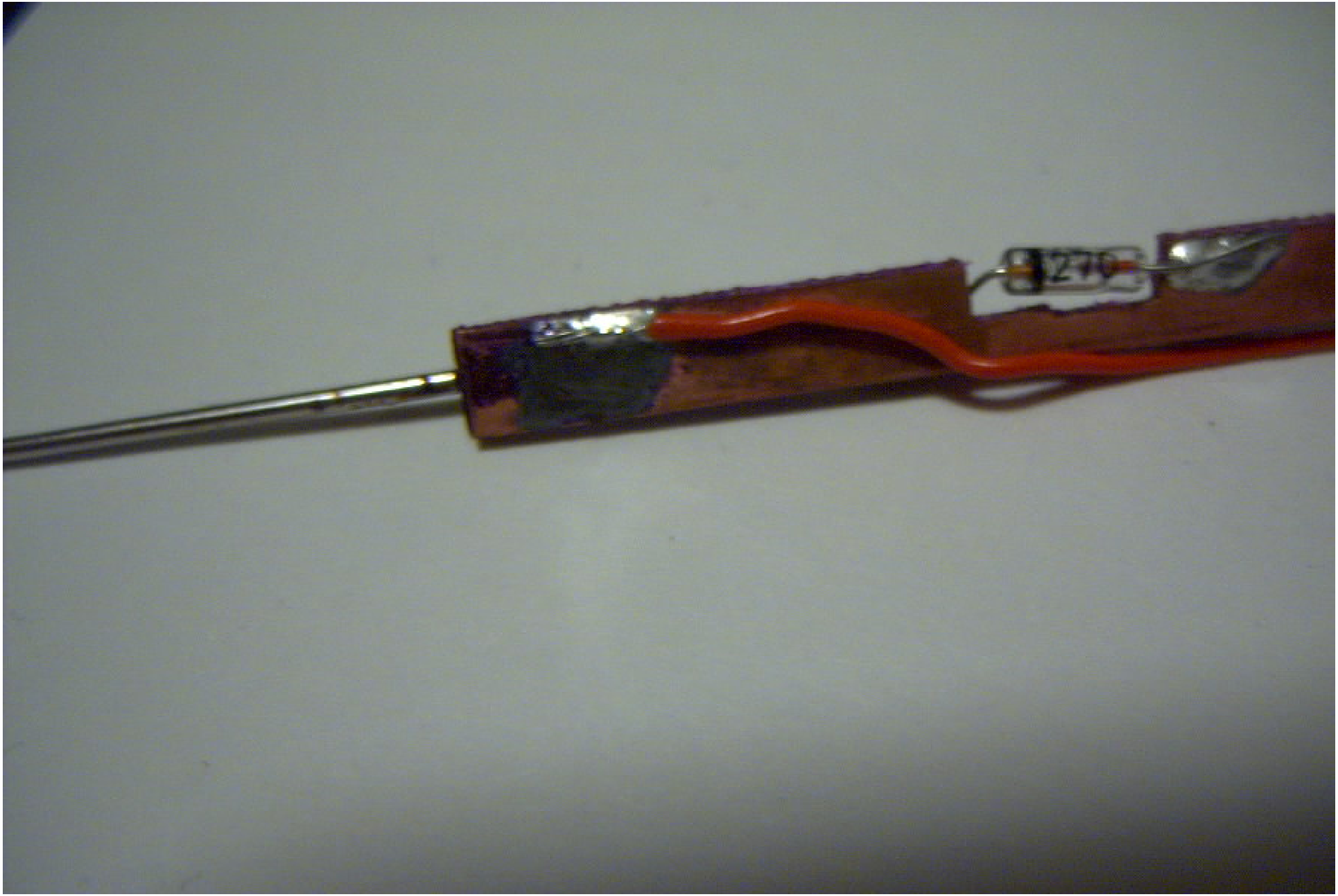

How to Build Your Own Oscilloscope Probes

Here is the complete bill of materials:

- The pen
- A 2-meter piece of coaxial test cable with a BNC connector on one end
- Epoxy adhesive
- One alligator clip
- Copper-plated nail — 0.75" (20mm) long, packed as "weather-stripping nail".
- 1 MΩ and 5 MΩ resistors

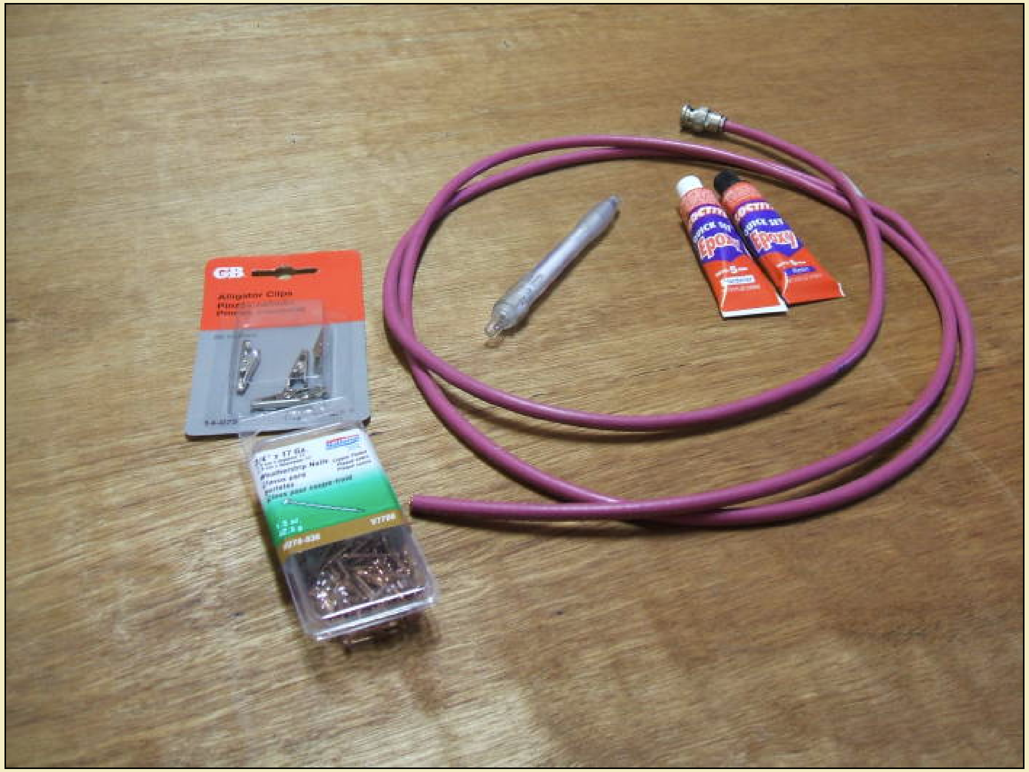
GB
Alligator Clips
Epoxy
Epoxy

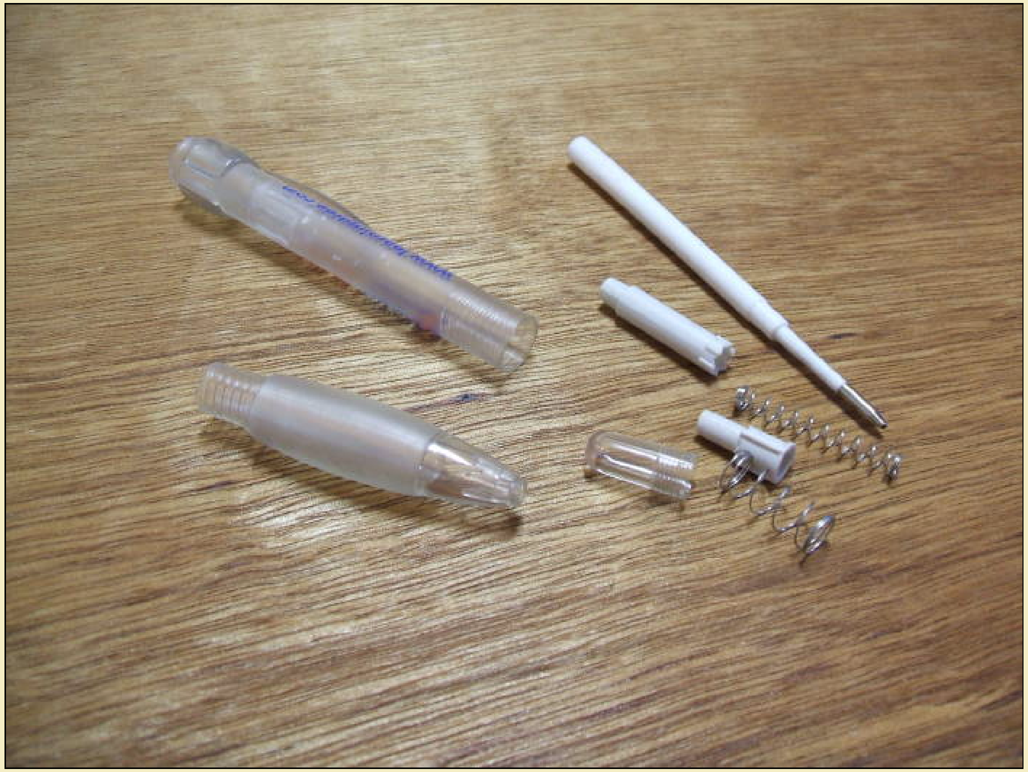

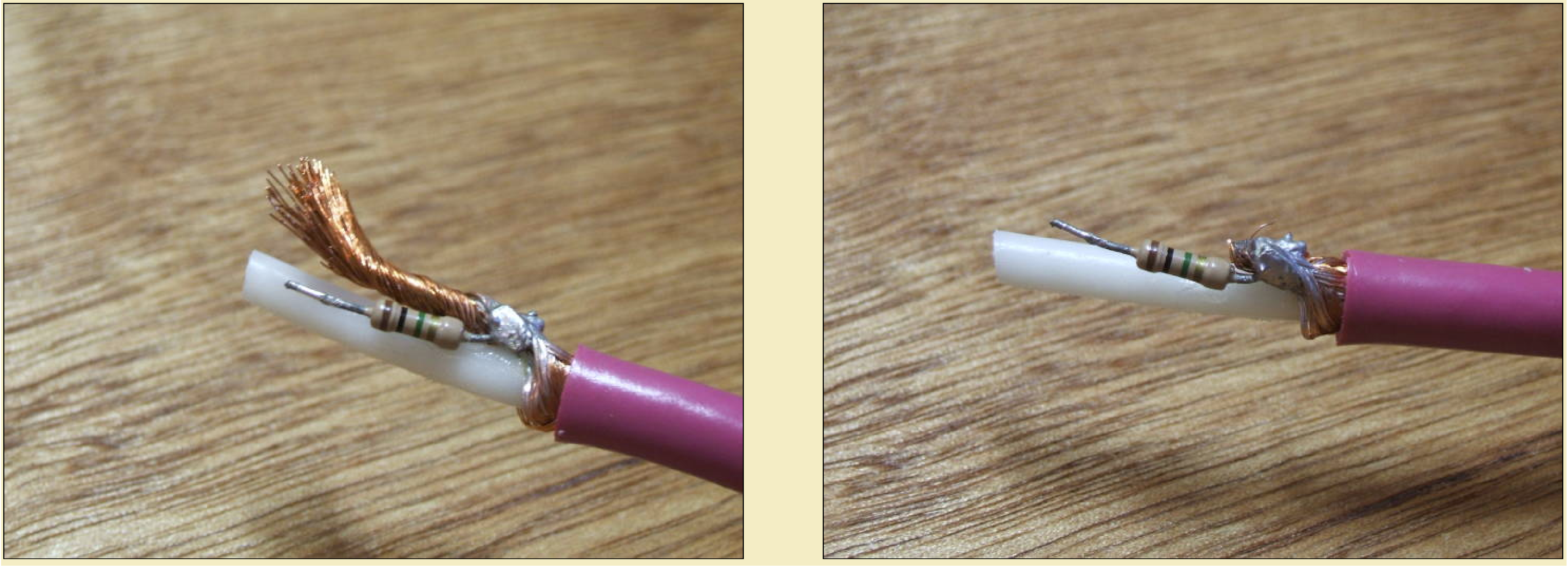

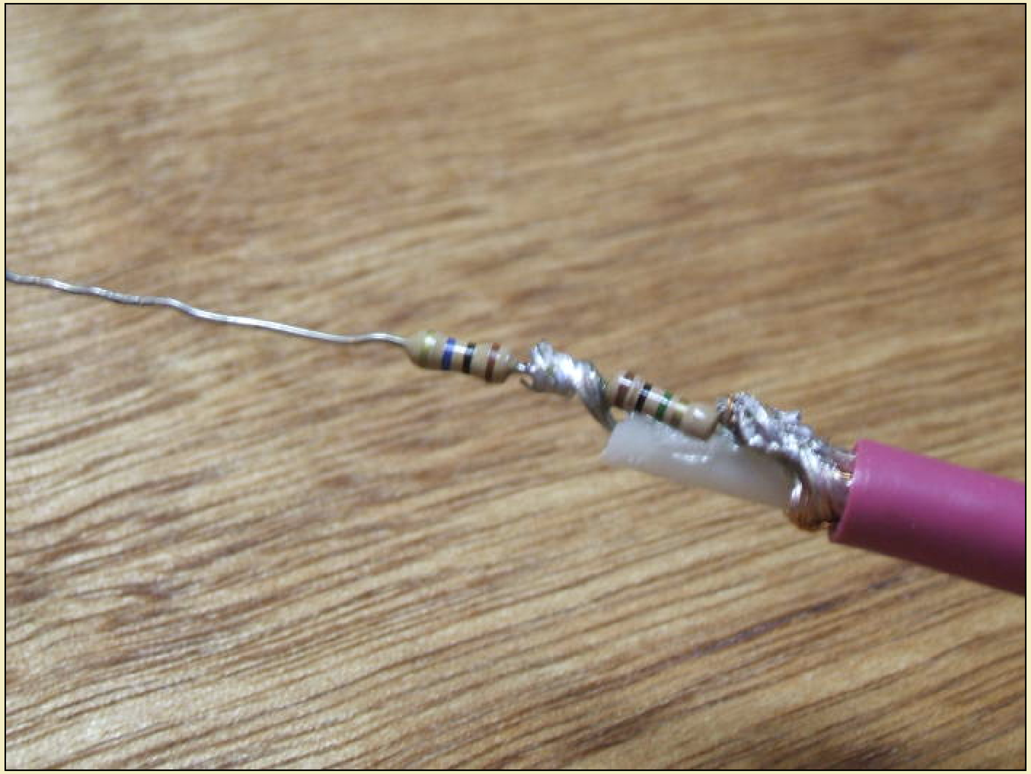

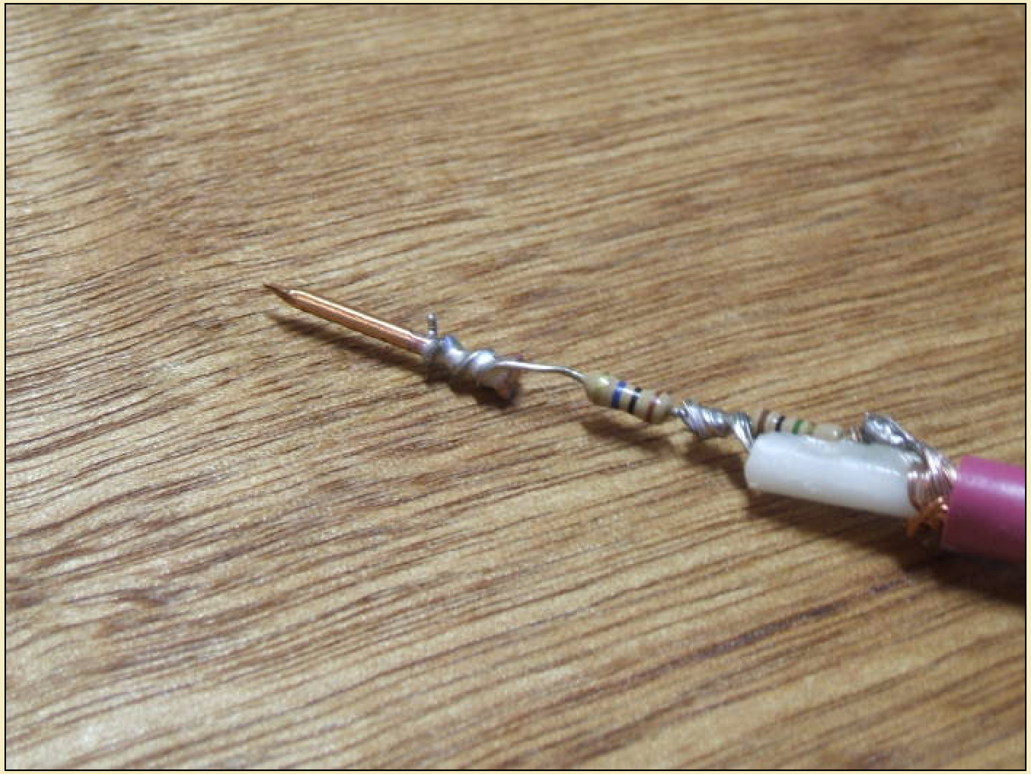

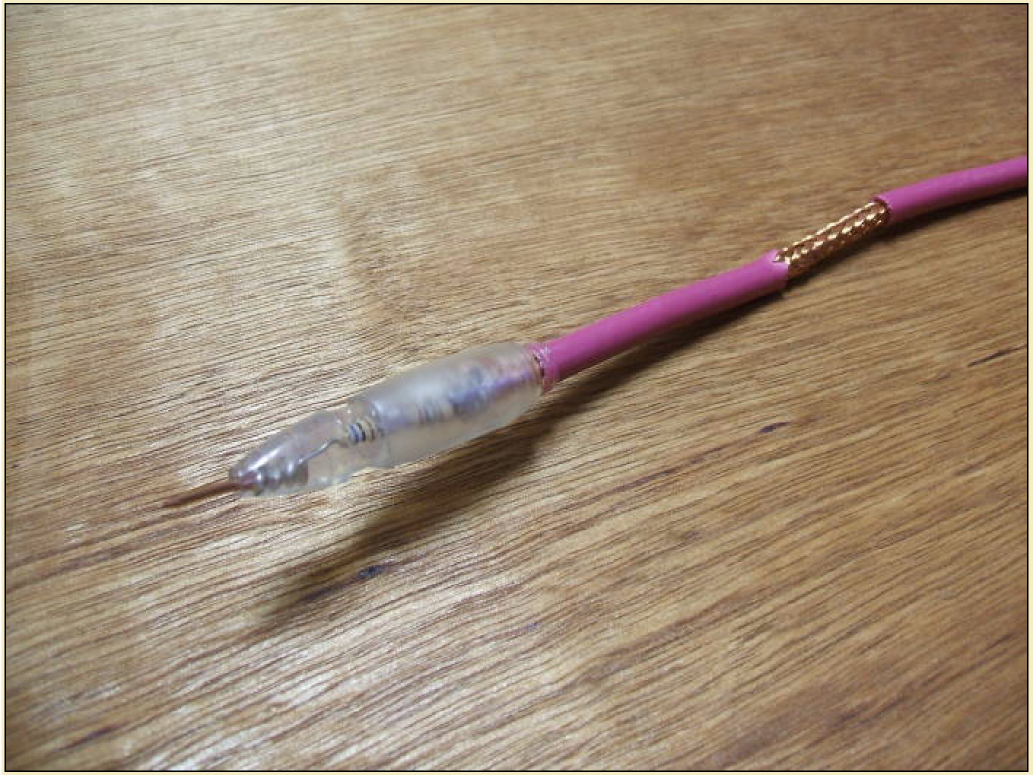

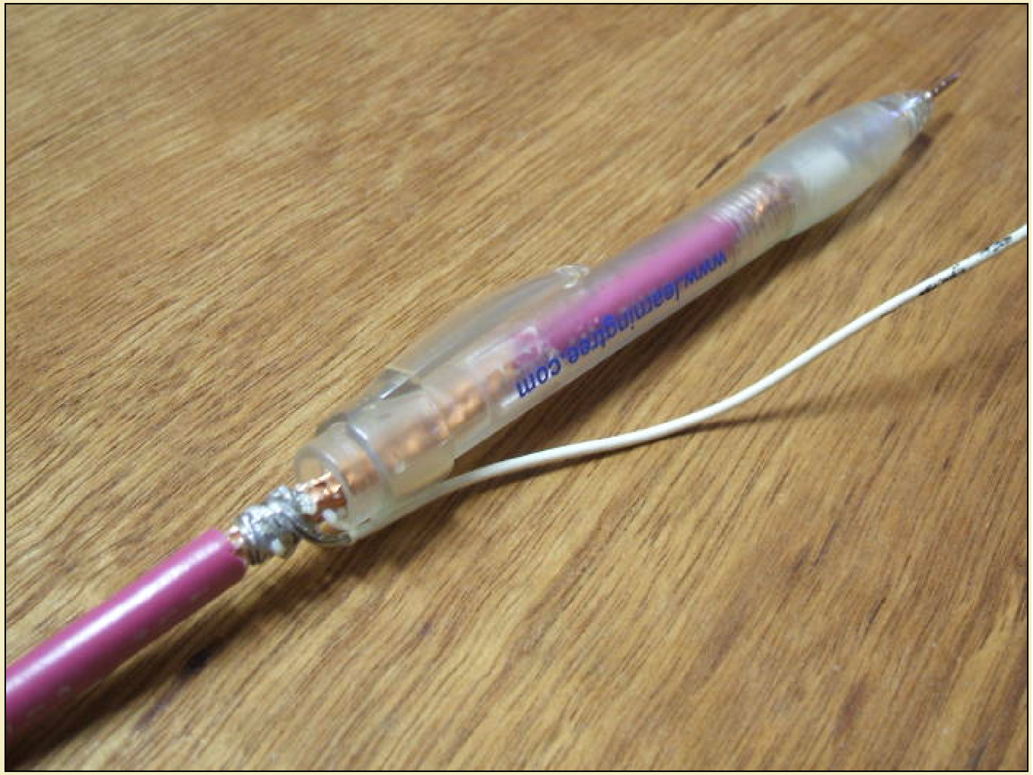
www.learningtree.com

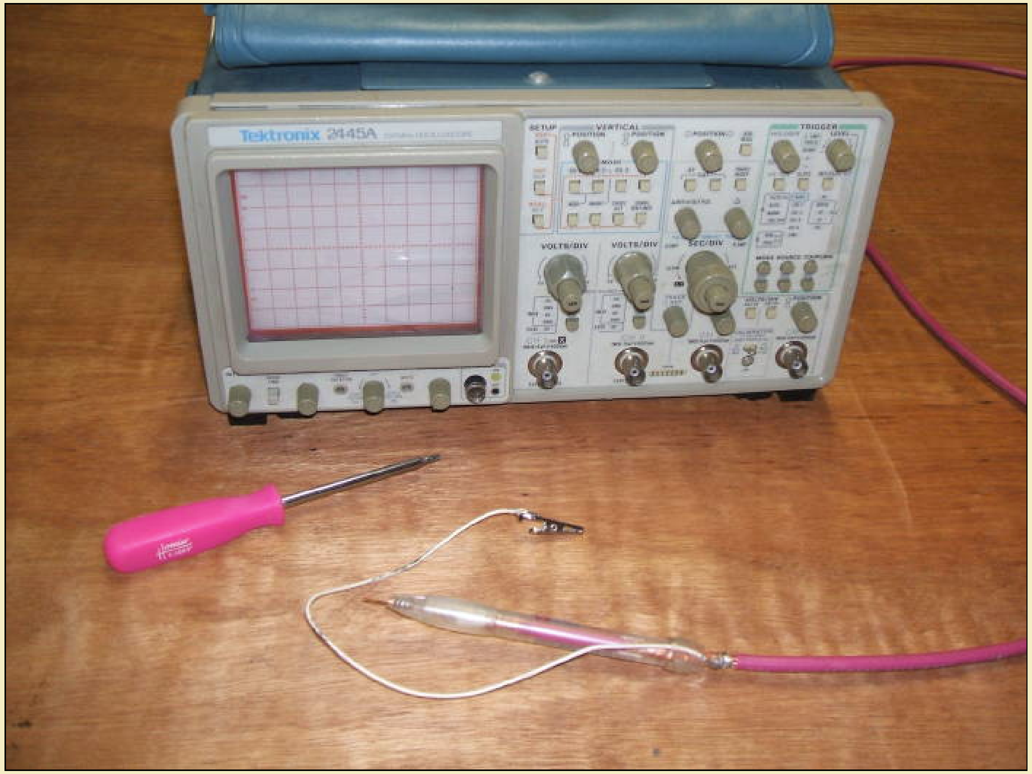
Tektronix 2445A

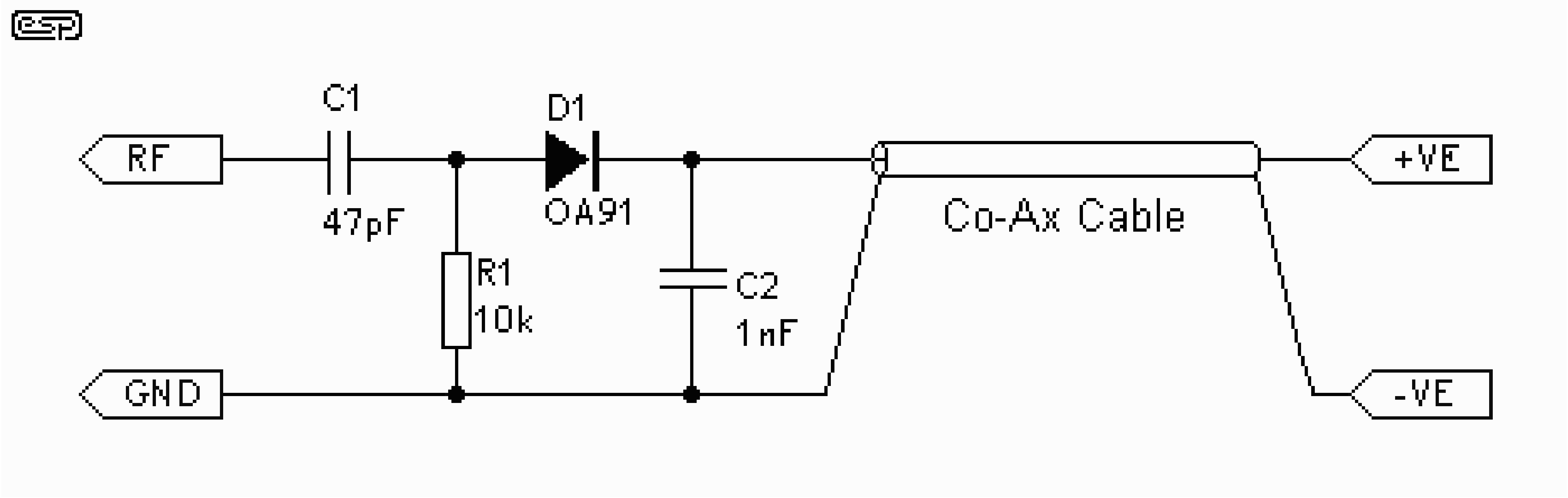
esp
C1
47pF
D1
OA91
RF
GND
R1
10k
C2
1nF
Co-Ax Cable
+VE
-VE

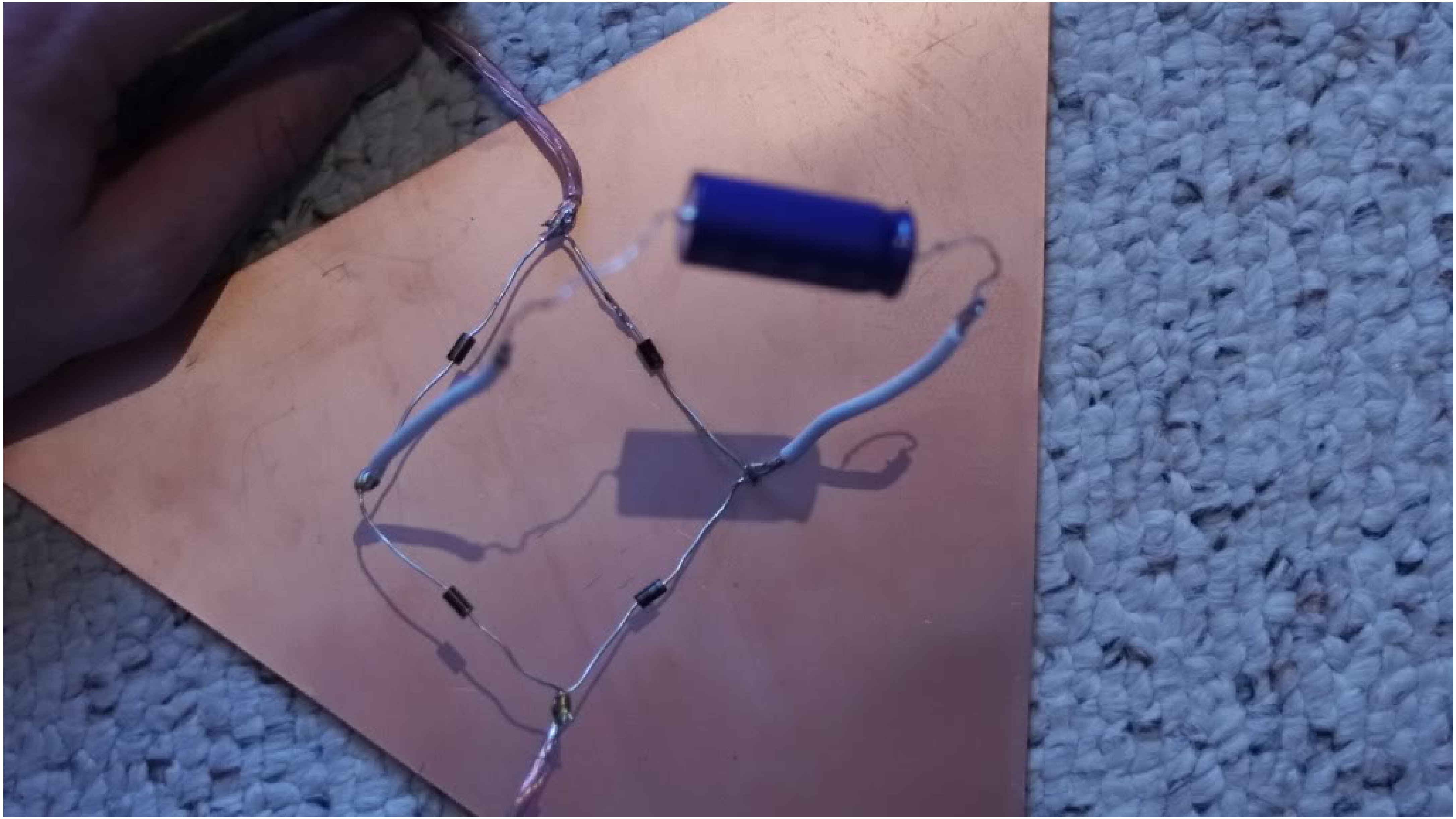

Pacific Antenna 15 Watt Dummy Load Kit

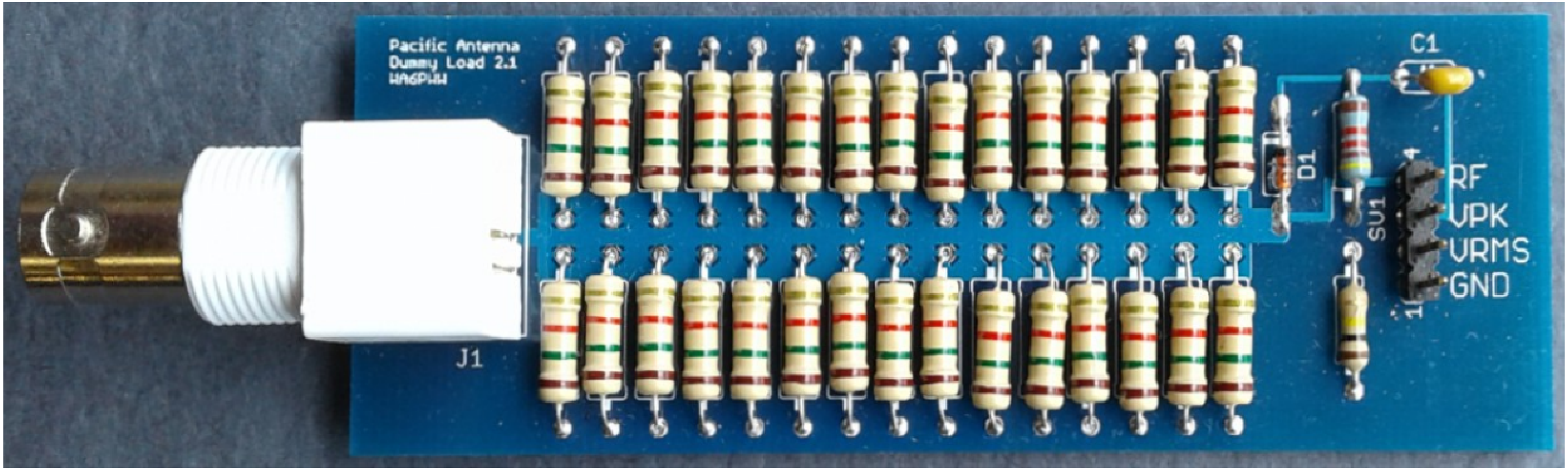
Pacific Antenna
Dummy Load 2.1
WA6PWW
C1
D1
SU1
4
RF
VPK
VRMS
GND
1
J1

Inspection and Inventory

First, check the kit to be sure all parts are included. Should anything be missing, please contact us for a replacement.

30 – R1-R30, 1.5 K 1/2 watt resistors: Brown-Green-Red-Gold
1 – R55: 41.2K 1/4W, 1% resistor: Yellow-Brown-Red-Red--Brown
1 – R56: 100K 1/4W, 5% resistor: Brown-Black-Yellow-Gold
1 – D1: 1N4148 diode
1 – C1: 0.01uF monolythic capacitor, yellow, (marked 103)
1 – J1: BNC board mount connector
1 – SV1: 4 pin header
1 – Circuit board

Board Layout

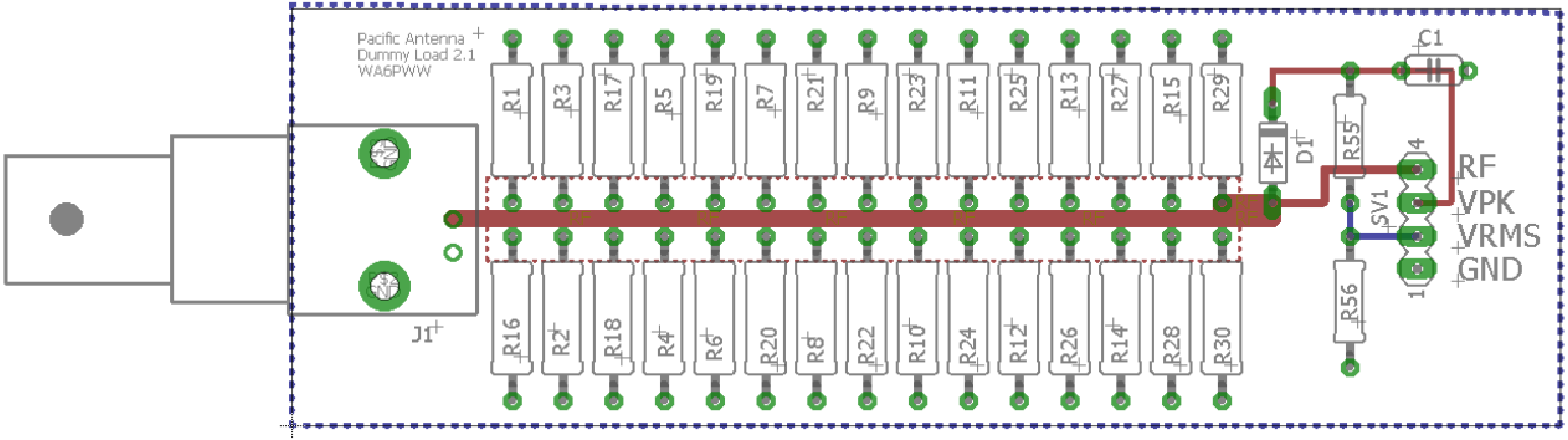
Pacific Antenna
Dummy Load 2.1
WA6PWW
R1
R3
R17
R5
R19
R7
R21
R9
R23
R11
R25
R13
R27
R15
R29
R16
R2
R18
R4
R6
R20
R8
R22
R10
R24
R12
R26
R14
R28
R30
J1
D1
R55
R56
C1
SV1
4
1
RF
VPK
VRMS
GND

Assembly

Install R1- R30
These are the 1/2W. 1.5K ohm resistors and they go in the marked locations shown on the circuit board.

You may find it helpful to do one row of the resistors at a time to make soldering the leads easier.

First, pre-bend the leads near the resistor bodies and then insert them into the board.

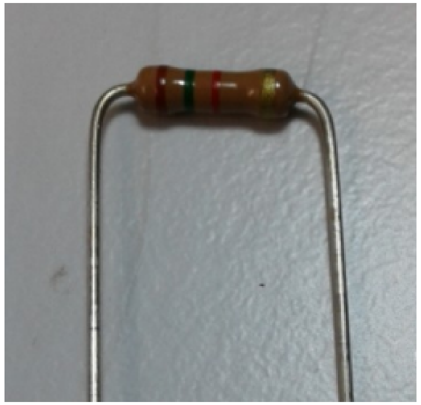

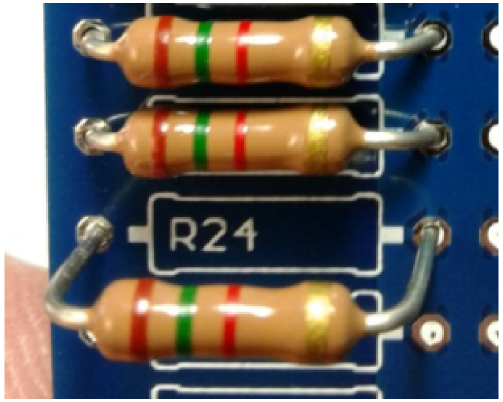

Install, solder and trim the leads of D1. Be sure to match the band end to the diagram above and the outline on the circuit board.

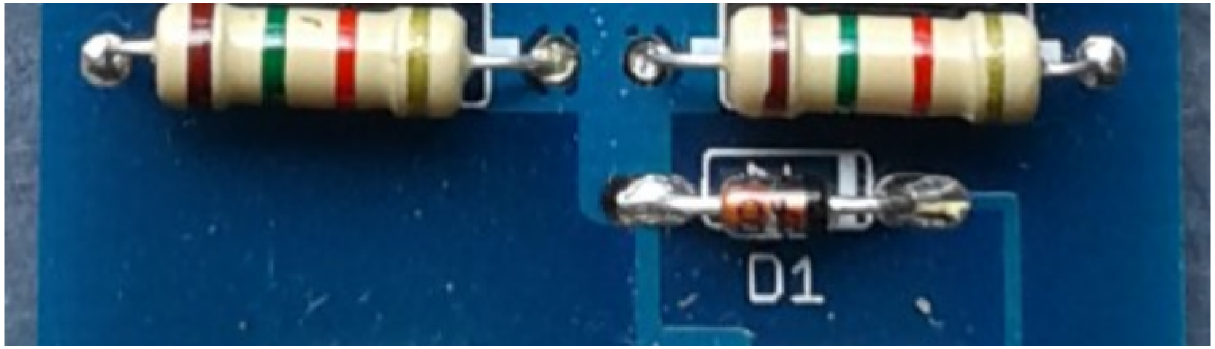

Install R55 the 41.2K ohm (Yellow-Brown-Red-Red—Brown) resistor in the marked location on the board

Install R56, the 100K resistor (Brown-Black-Yellow-Gold) in the marked location on the board.

Install C1 the 0.01uF capacitor in the location marked on the board.

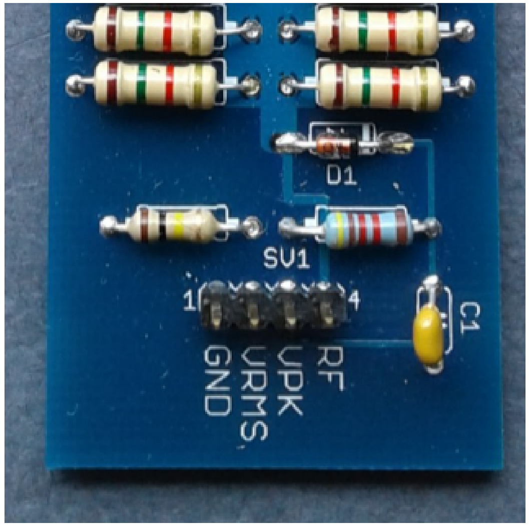

Now, solder the BNC connector, making sure to seat it fully into the board. Solder the two small wires and the two support pins.

The support pins may require longer time, increased temperature or a larger soldering iron to properly solder.

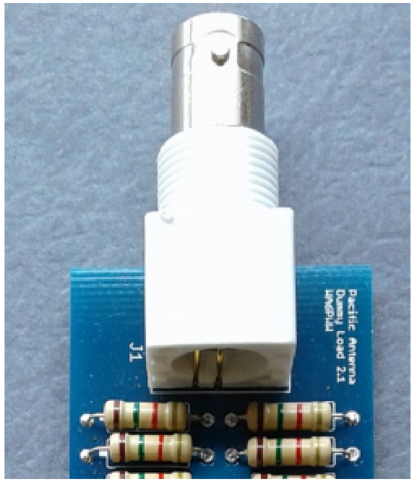
J1
Pacific Antenna
Dummy Load 2.1

Congratulations, your dummy load kit is now complete!

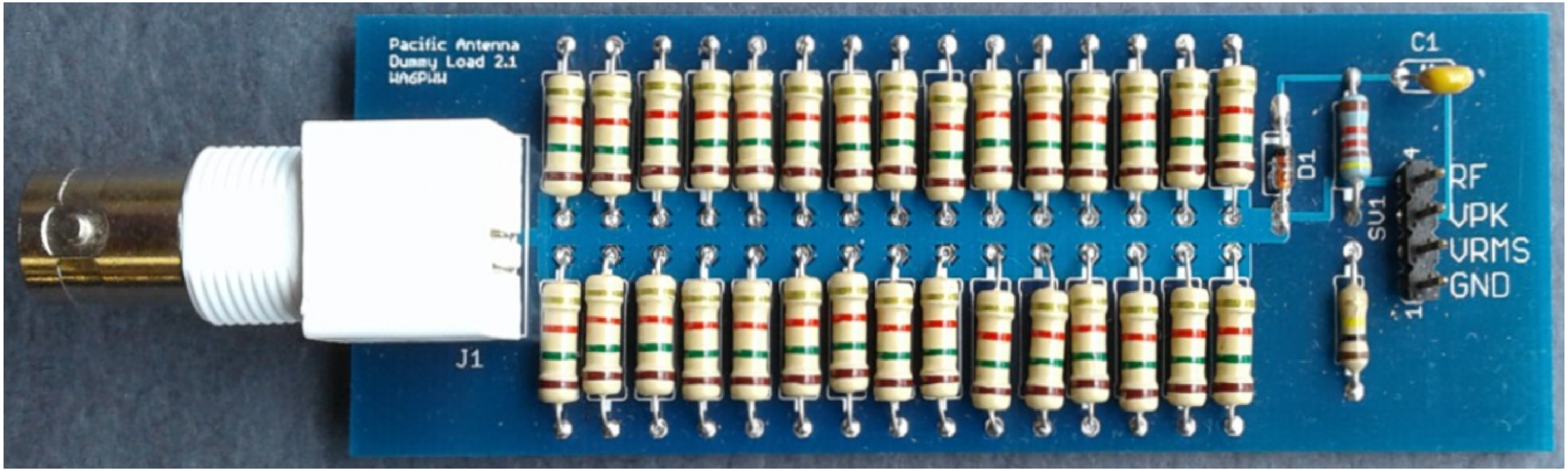
Pacific Antenna
Dummy Load 2.1
WA6PWW
C1
D1
SU1
RF
VPK
VRMS
GND
J1

Operation

The dummy load is easy to use. Simply connect your transmitter input to the BNC

To measure RF Power, connect your multimeter to pin 1 and Pin 2 or 3.

Pin 1 is ground and the DC output voltages appear on pins 2 and 3 of SV1.

Pin 2 provides the RMS value of the RF voltage.

Pin 3 gives the Peak value of the RF voltage.

Pin 4 is direct RF voltage across the resistors.

RF power is calculated from this relationship: Power = (Vrms^2)/50

Note: The power input should be limited to 15W to avoid damaging the resistors and sense circuit components.

The simplest power meter

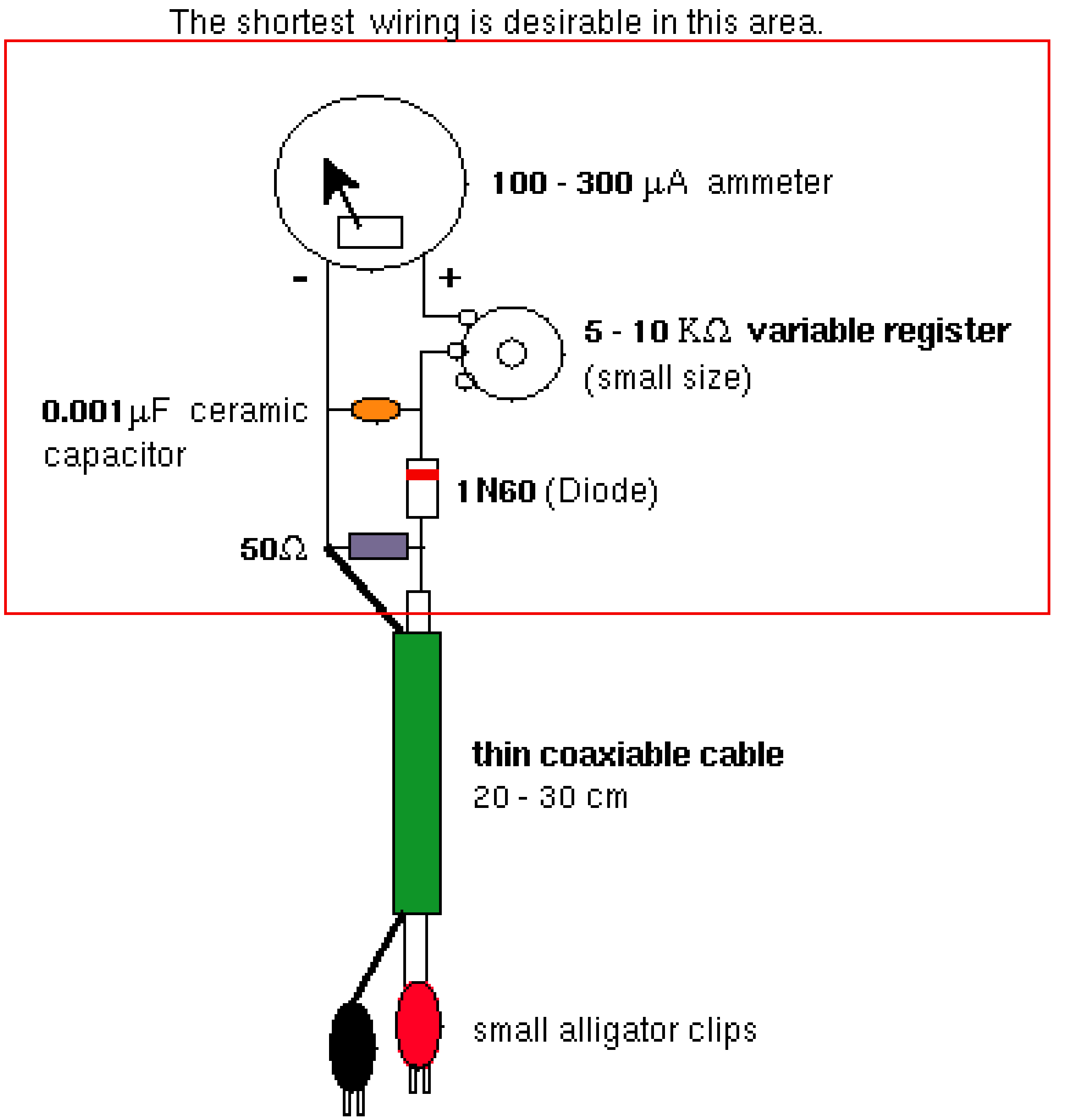
The shortest wiring is desirable in this area.
100 - 300 µA ammeter
-
+
5 - 10 KΩ variable register
(small size)
0.001µF ceramic capacitor
1N60 (Diode)
50Ω
thin coaxiable cable
20 - 30 cm
small alligator clips

As for the ammeter, you can use an used one taking from junked audio amplifier, tape-recorder, radio-cassette, and so on.

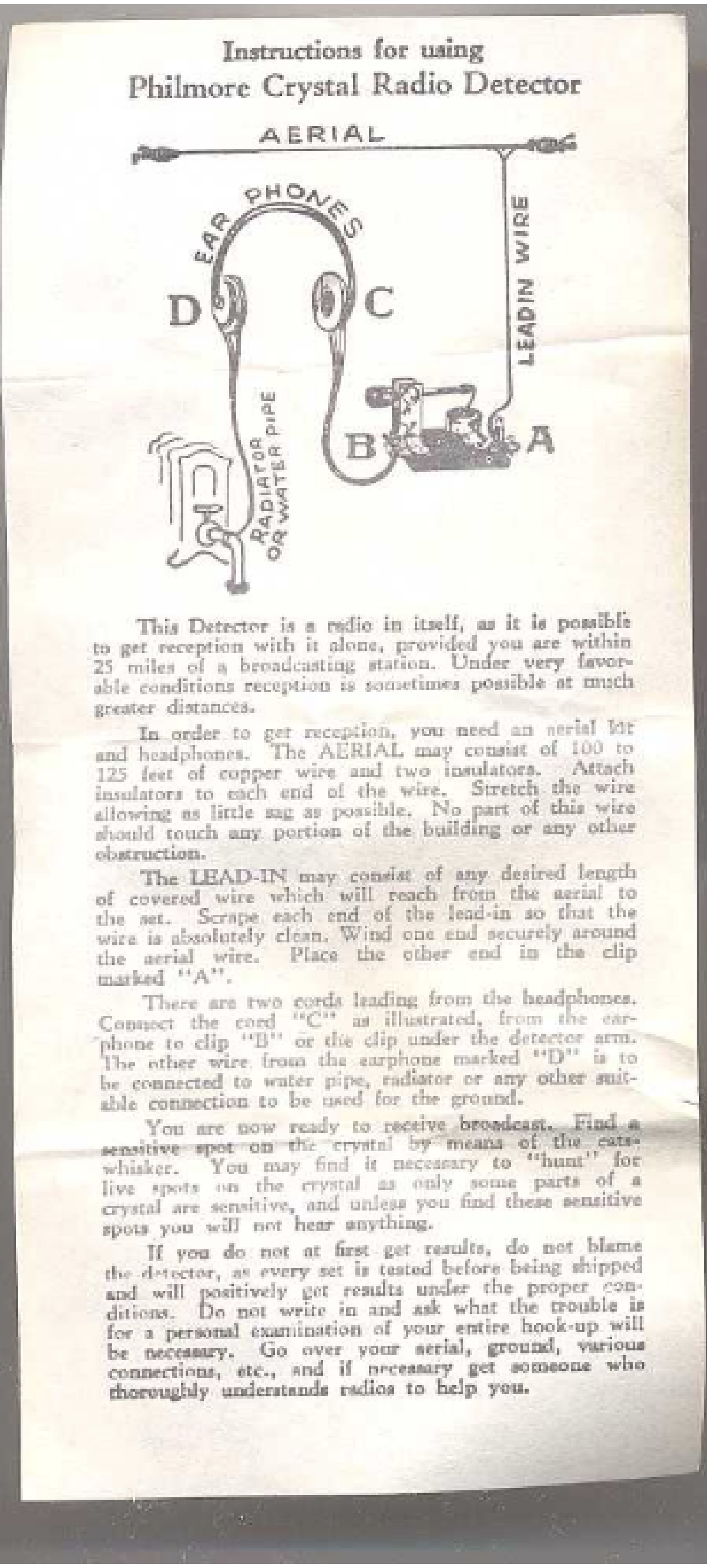

Instructions for using Philmore Crystal Radio Detector

AERIAL
EAR PHONES
LEADIN WIRE
D
C
B
A
RADIATOR OR WATER PIPE

This Detector is a radio in itself, as it is possible to get reception with it alone, provided you are within 25 miles of a broadcasting station. Under very favorable conditions reception is sometimes possible at much greater distances.

In order to get reception, you need an aerial kit and headphones. The AERIAL may consist of 100 to 125 feet of copper wire and two insulators. Attach insulators to each end of the wire. Stretch the wire allowing as little sag as possible. No part of this wire should touch any portion of the building or any other obstruction.

The LEAD-IN may consist of any desired length of covered wire which will reach from the aerial to the set. Scrape each end of the lead-in so that the wire is absolutely clean. Wind one end securely around the aerial wire. Place the other end in the clip marked "A".

There are two cords leading from the headphones. Connect the cord "C" as illustrated, from the earphone to clip "B" or the clip under the detector arm. The other wire from the earphone marked "D" is to be connected to water pipe, radiator or any other suitable connection to be used for the ground.

You are now ready to receive broadcast. Find a sensitive spot on the crystal by means of the catswhisker. You may find it necessary to "hunt" for live spots on the crystal as only some parts of a crystal are sensitive, and unless you find these sensitive spots you will not hear anything.

If you do not at first get results, do not blame the detector, as every set is tested before being shipped and will positively get results under the proper conditions. Do not write in and ask what the trouble is for a personal examination of your entire hook-up will be necessary. Go over your aerial, ground, various connections, etc., and if necessary get someone who thoroughly understands radios to help you.

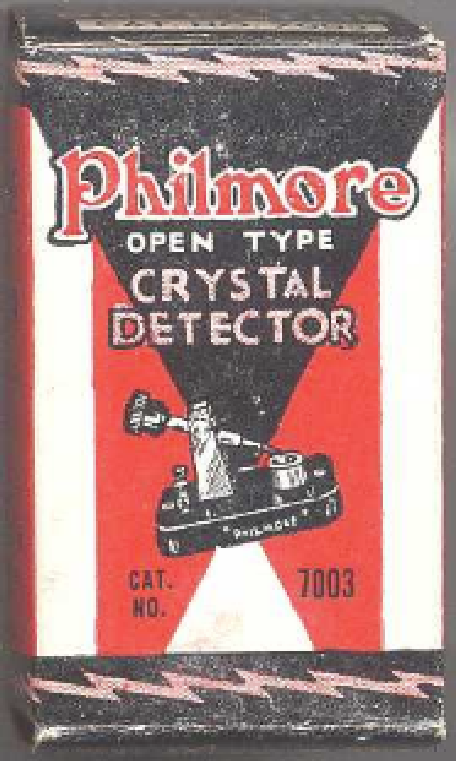

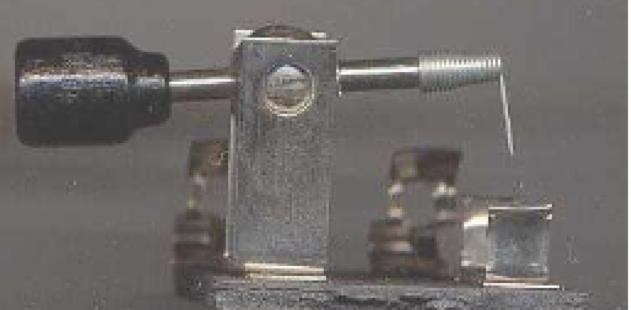

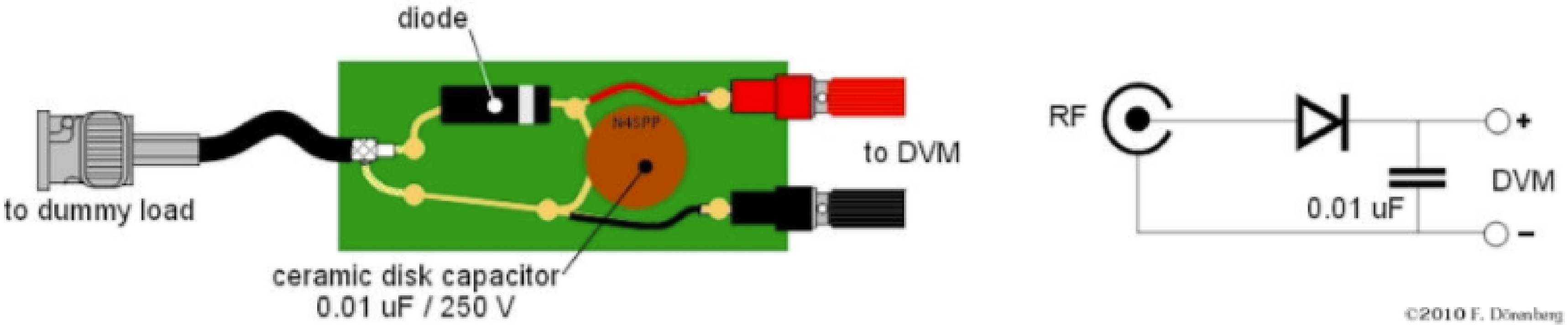
diode
N4SPP
to DVM
to dummy load
ceramic disk capacitor
0.01 uF / 250 V
RF
+
DVM
0.01 uF
–

©2010 F. Dörrenberg

A simple RF-probe

Obviously, this circuit will be fooled by a DC-offset on the RF signal. We can fix this by swapping the diode and the capacitor. Note that this is not necessary if you measure an RF voltage via a transformer, such as a directional coupler.

We can also make life a little easier by including a voltage divider with a scaling factor that is equal to the reciprocal of √2. Then the output voltage will be the RMS value that we are interested in. We can make a voltage divider where one resistor is the input impedance of the DVM. My DVM has a published input resistance of 10 MΩ. The second resistor should be 4M14 Ω,since 10 / (10+4.14) = 1 / √2). So 3M9 + 220k = 4M12 would be a good choice. This approach is shown below. Note that the resistor should be non-inductive (e.g., bulk-metal-foil or carbon).

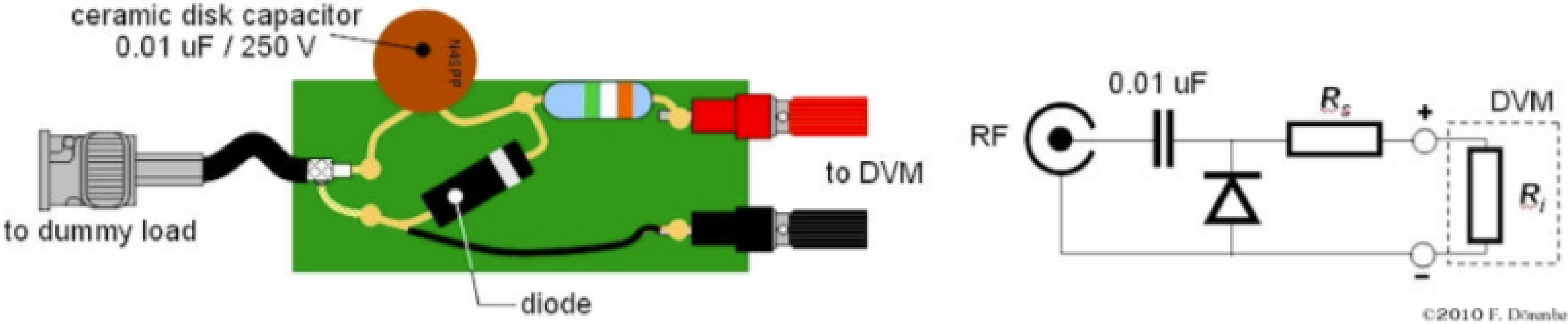
ceramic disk capacitor
0.01 uF / 250 V
N4SPP
to DVM
to dummy load
diode
0.01 uF
RF
R_s
+
DVM
R_i
–

©2010 F. Dörrenberg

RF-probe with DC-block and peak-to-RMS scaling

A variation on this, with a full-wave rectifier, is shown below:

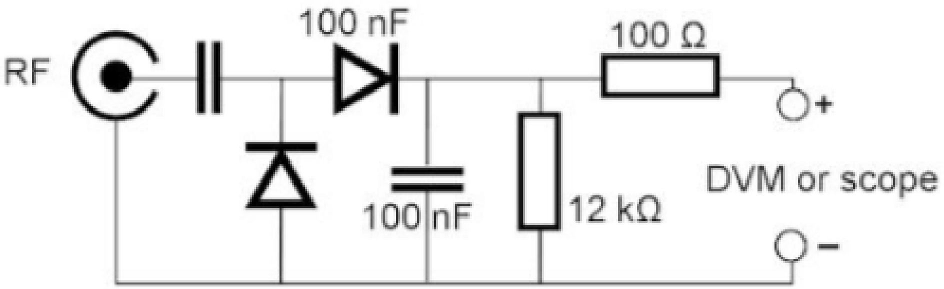
RF
100 nF
100 Ω
+
DVM or scope
100 nF
12 kΩ
–

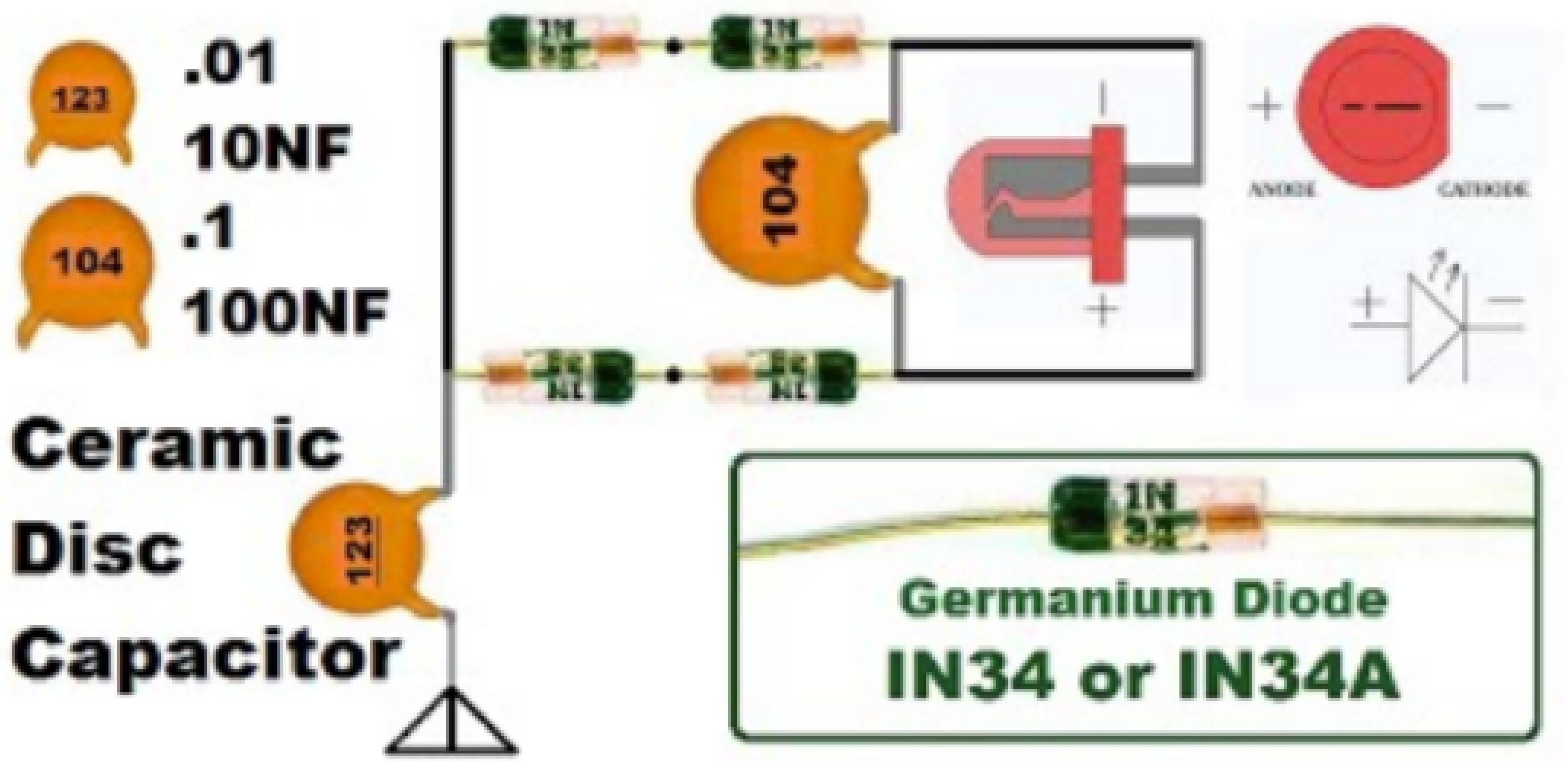
123
104
1N
34
104
+
−
+
−
ANODE
CATHODE
+
−
123
1N
34

.01
10NF

.1
100NF

Ceramic
Disc
Capacitor

Germanium Diode
IN34 or IN34A

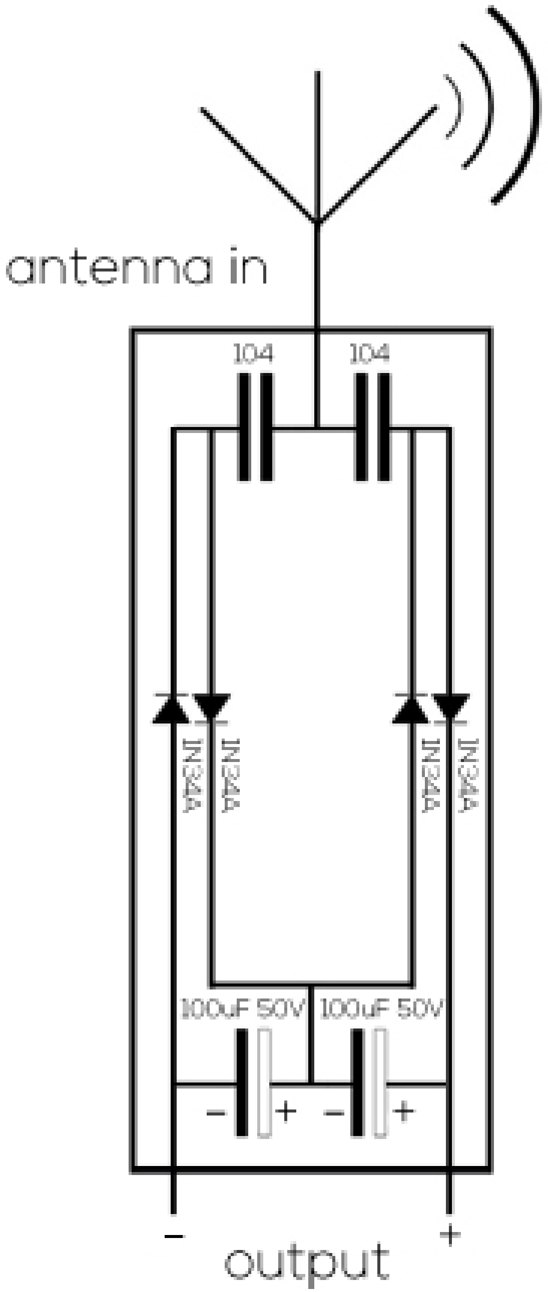
antenna in
104
104
1N34A
1N34A
1N34A
1N34A
100uF 50V
100uF 50V
−
+
−
+
−
output
+

RF to DC

Circuit Diagam (simplified)

Component list:
(2) Ceramic Capacitors (104)
(4) Germanium Diodes (1N34A)
(2) Electrolytic Capacitors (100uF 50V)

Copyright 2016. Drew Paul Designs. All rights reserved.

RF Current Probe

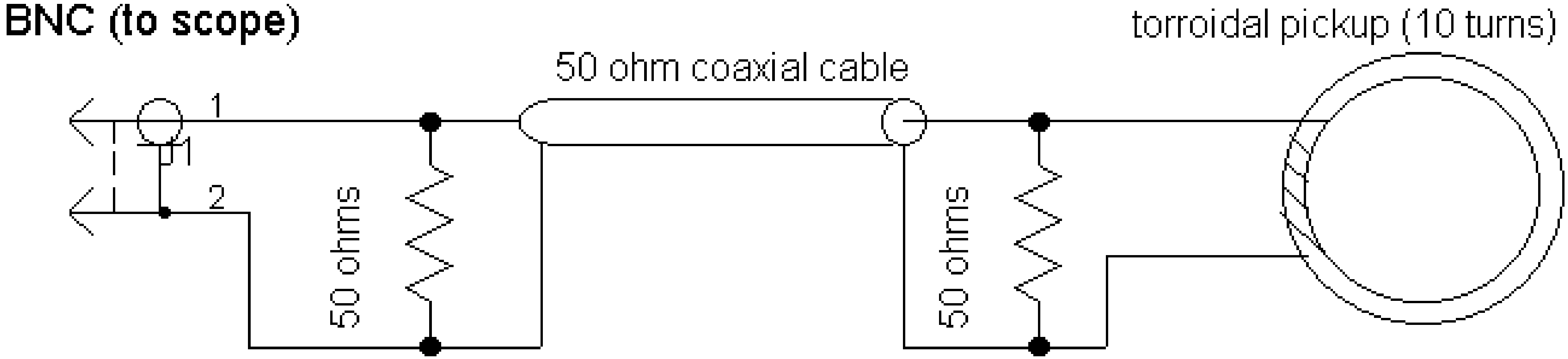
BNC (to scope)
torroidal pickup (10 turns)
50 ohm coaxial cable
1
J1
2
50 ohms
50 ohms

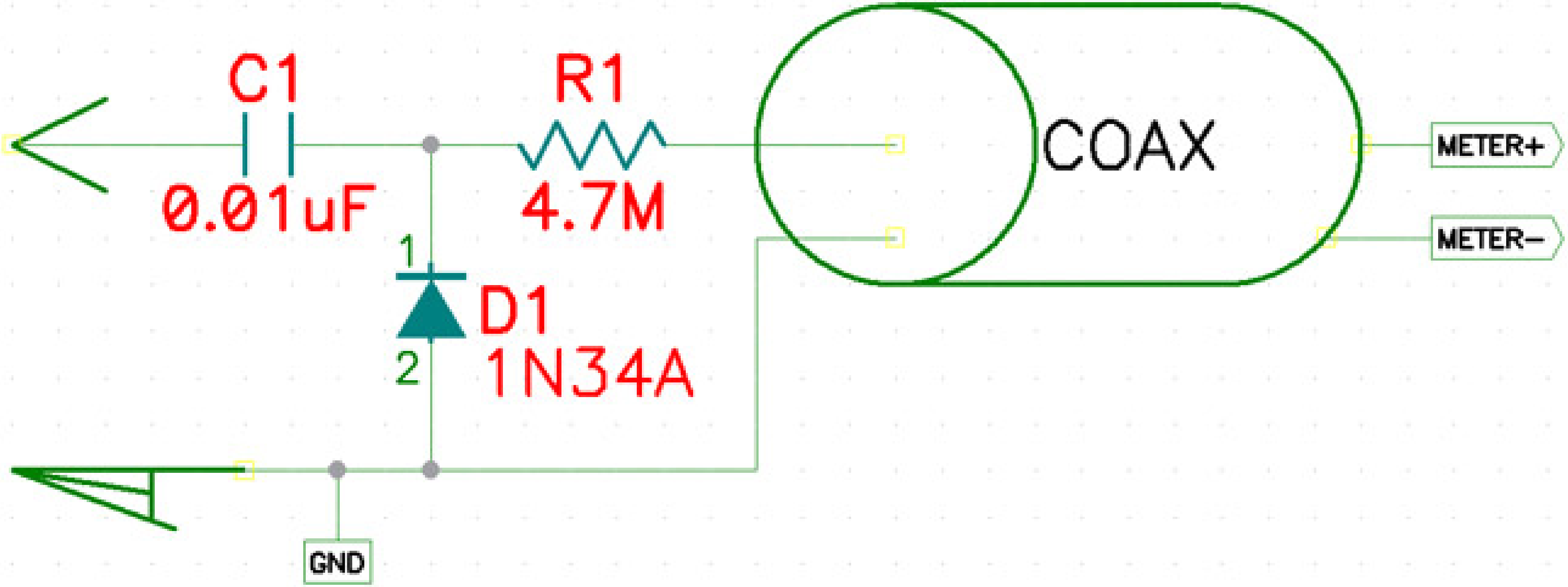
C1
0.01uF
R1
4.7M
COAX
METER+
METER-
1
D1
1N34A
2
GND

Simple RF Probe

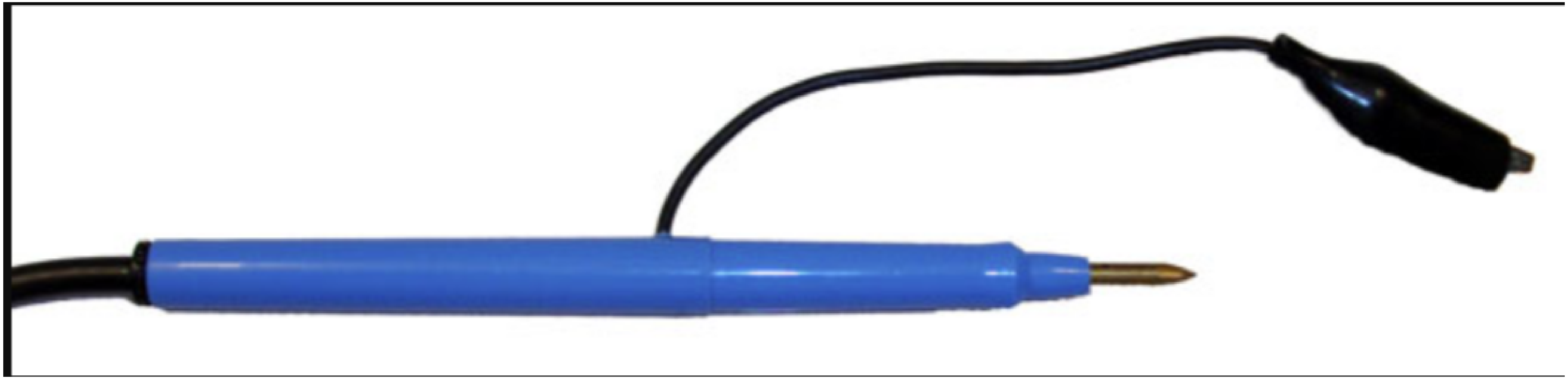

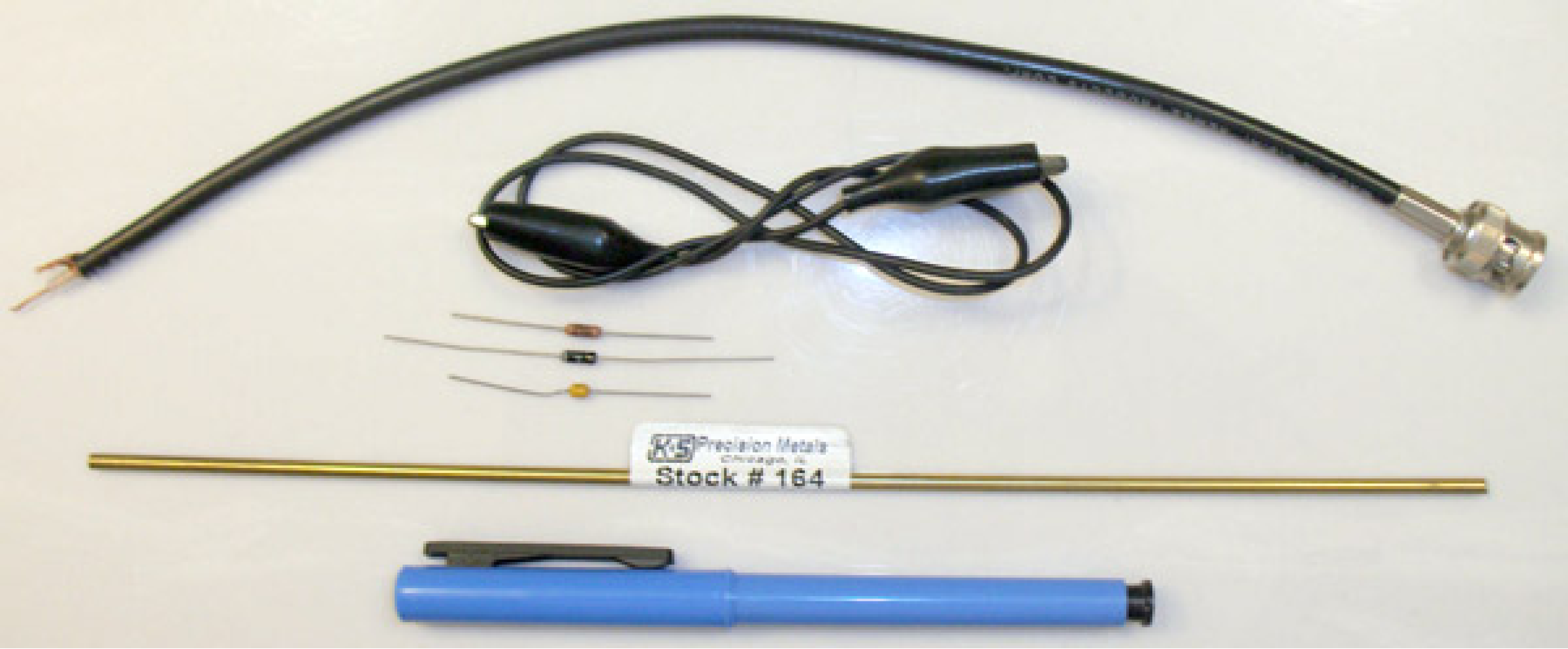
K-S
Precision Metals
Chicago, IL
Stock # 164

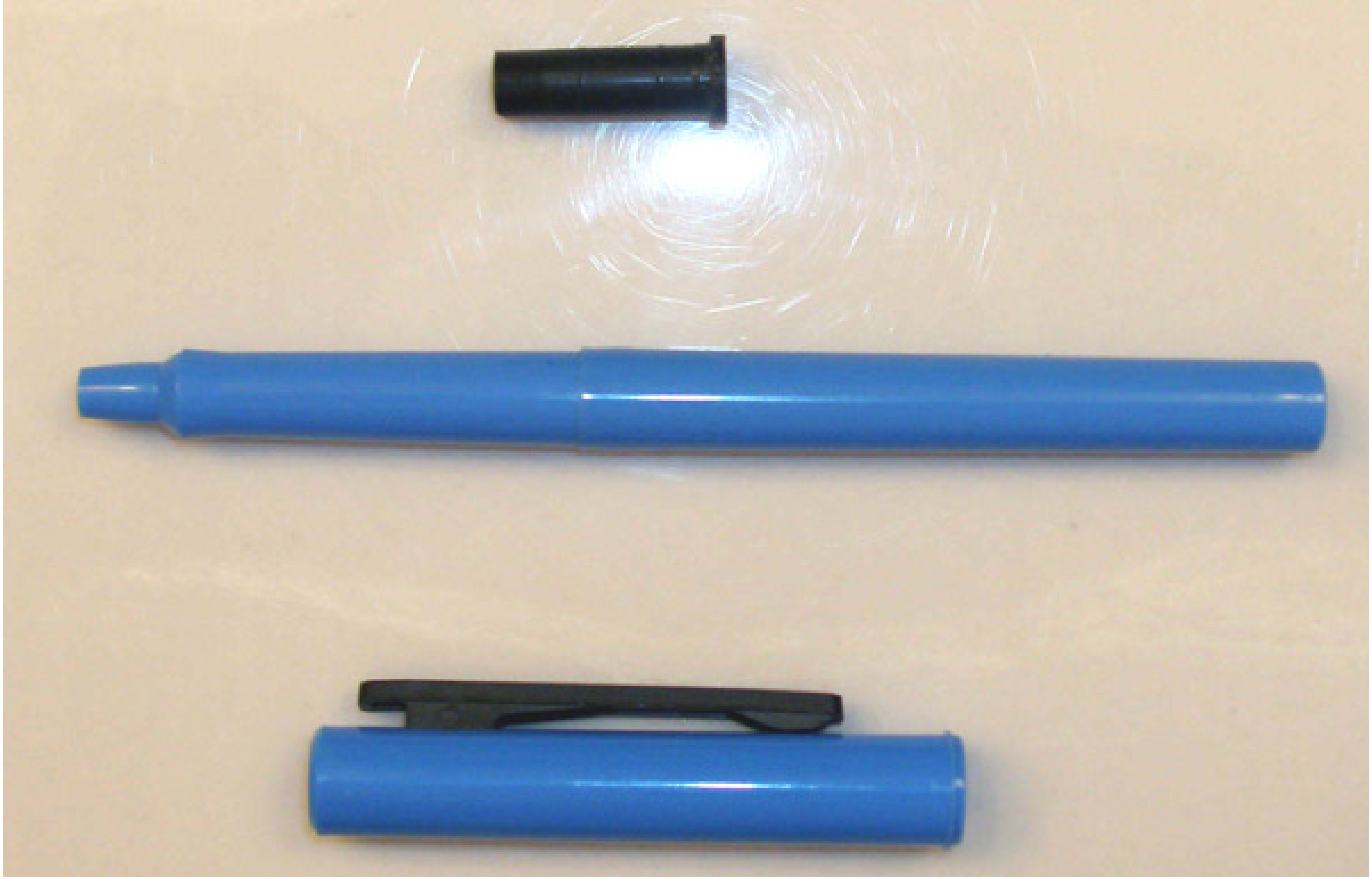

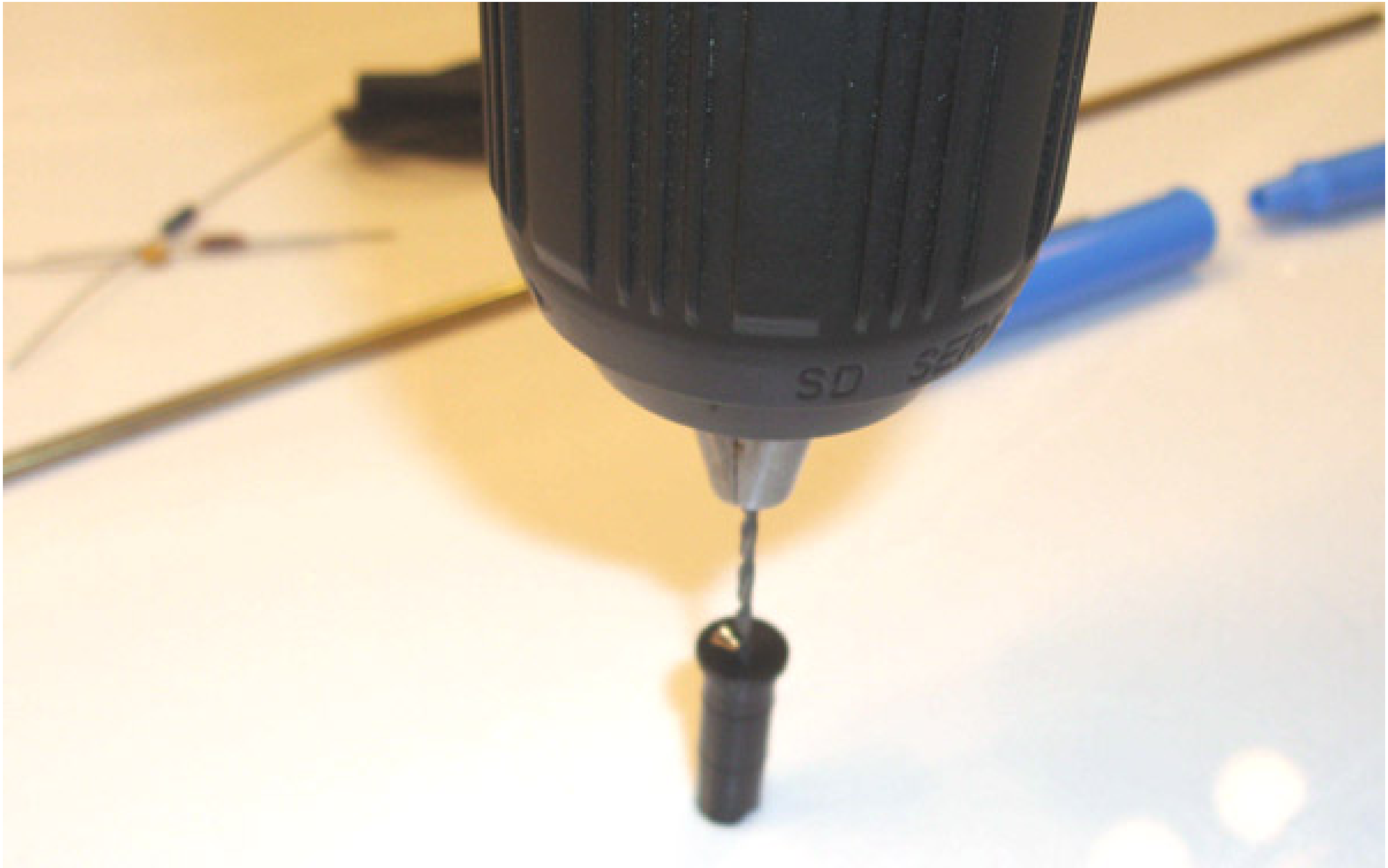

SD

RG 58

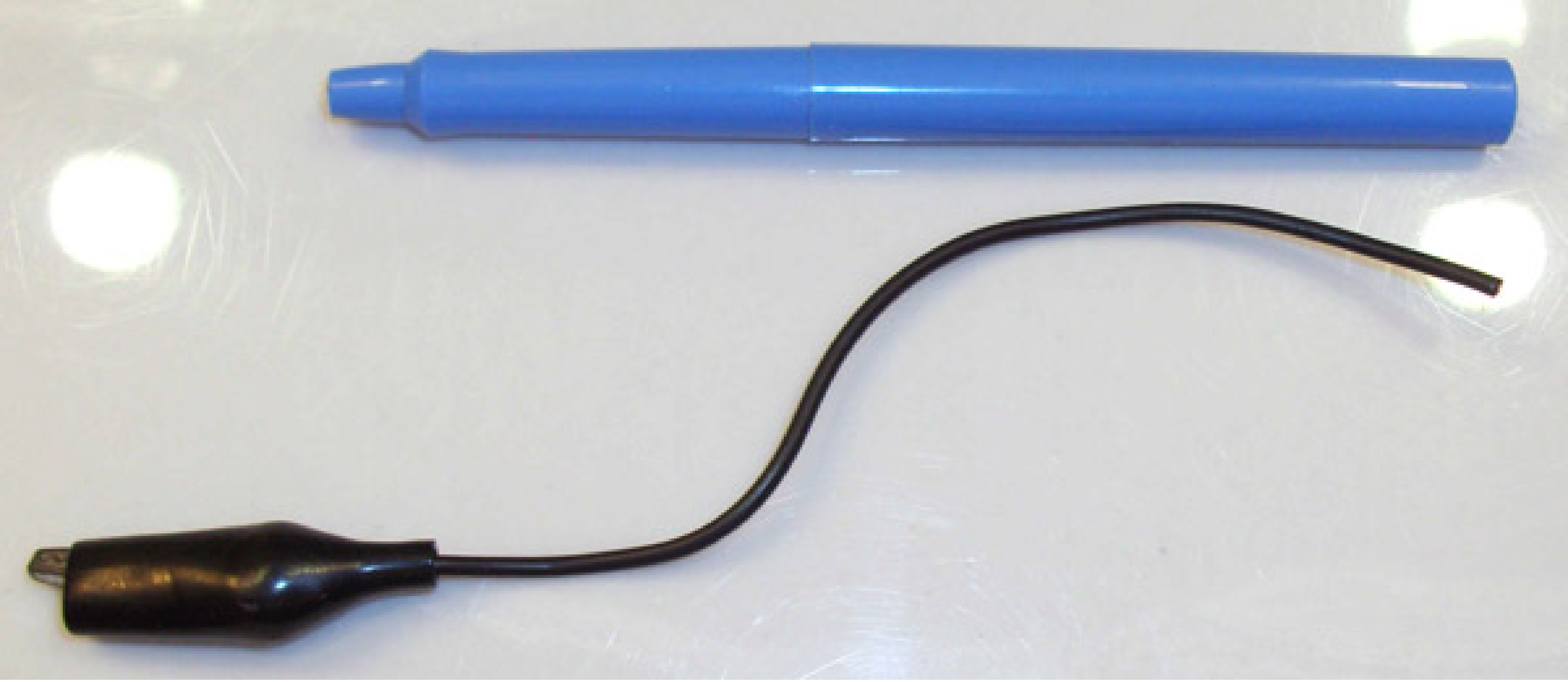

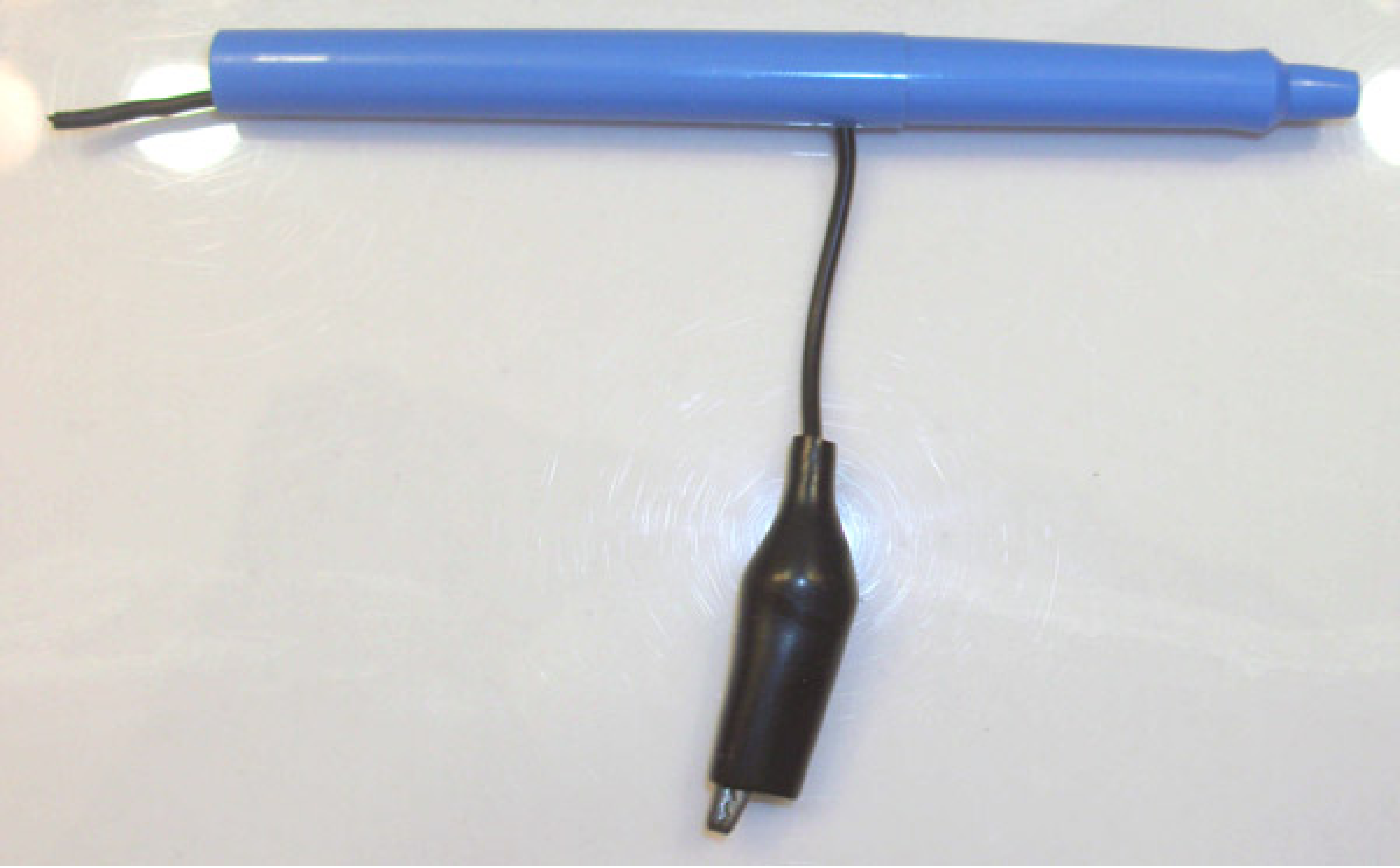

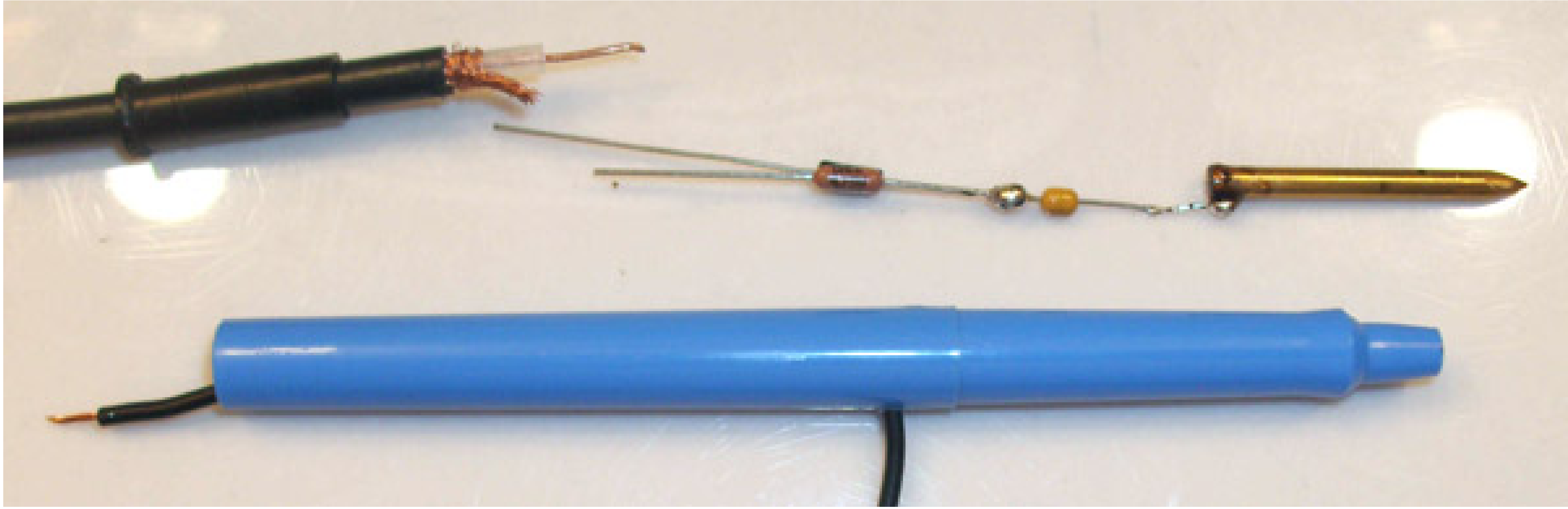

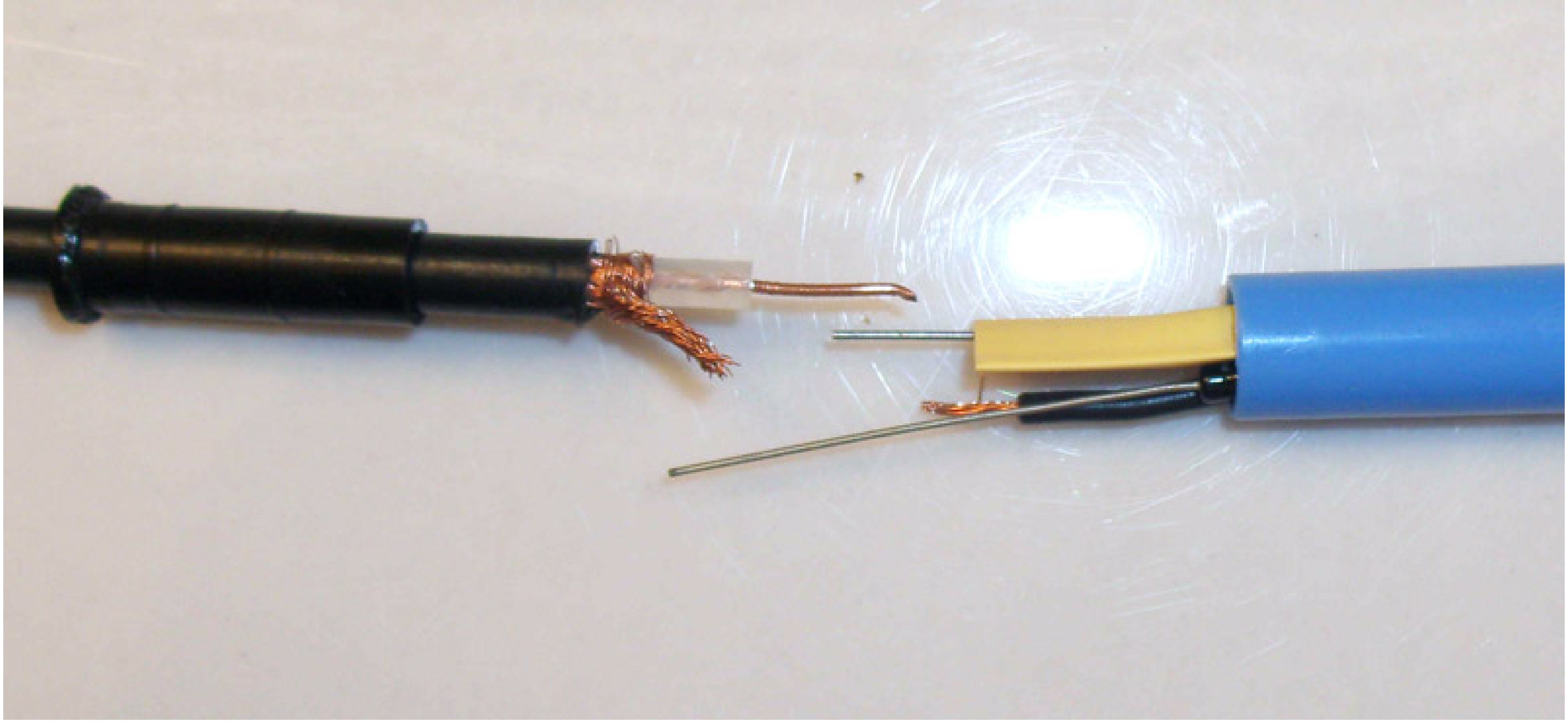

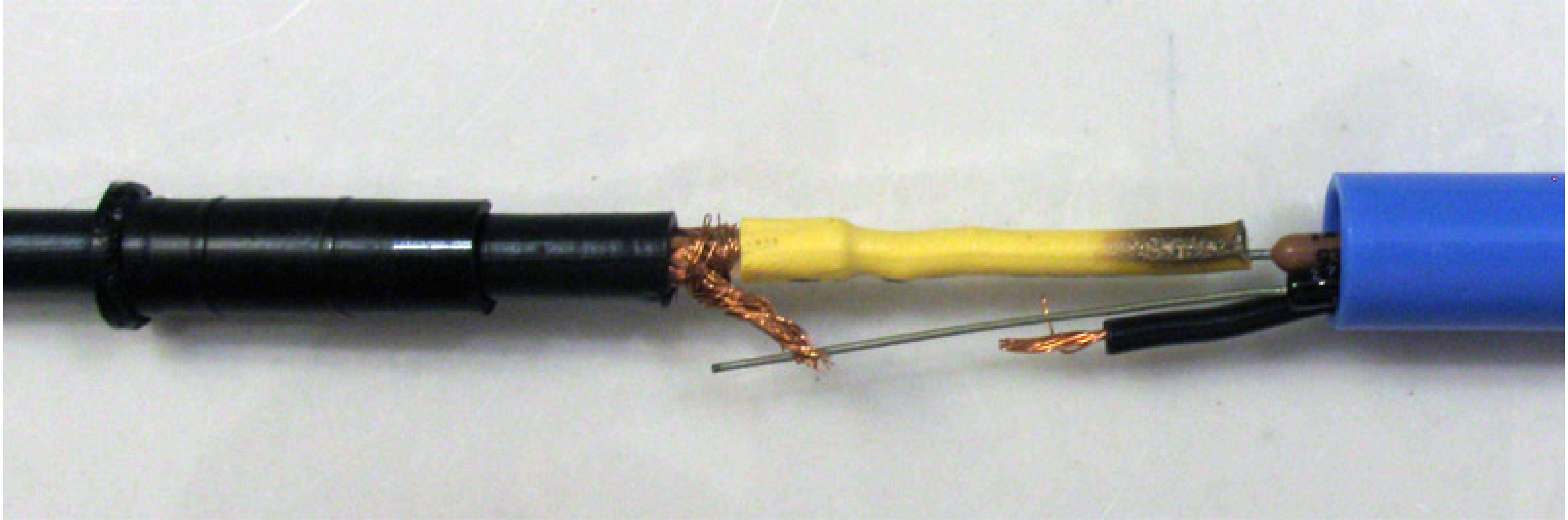

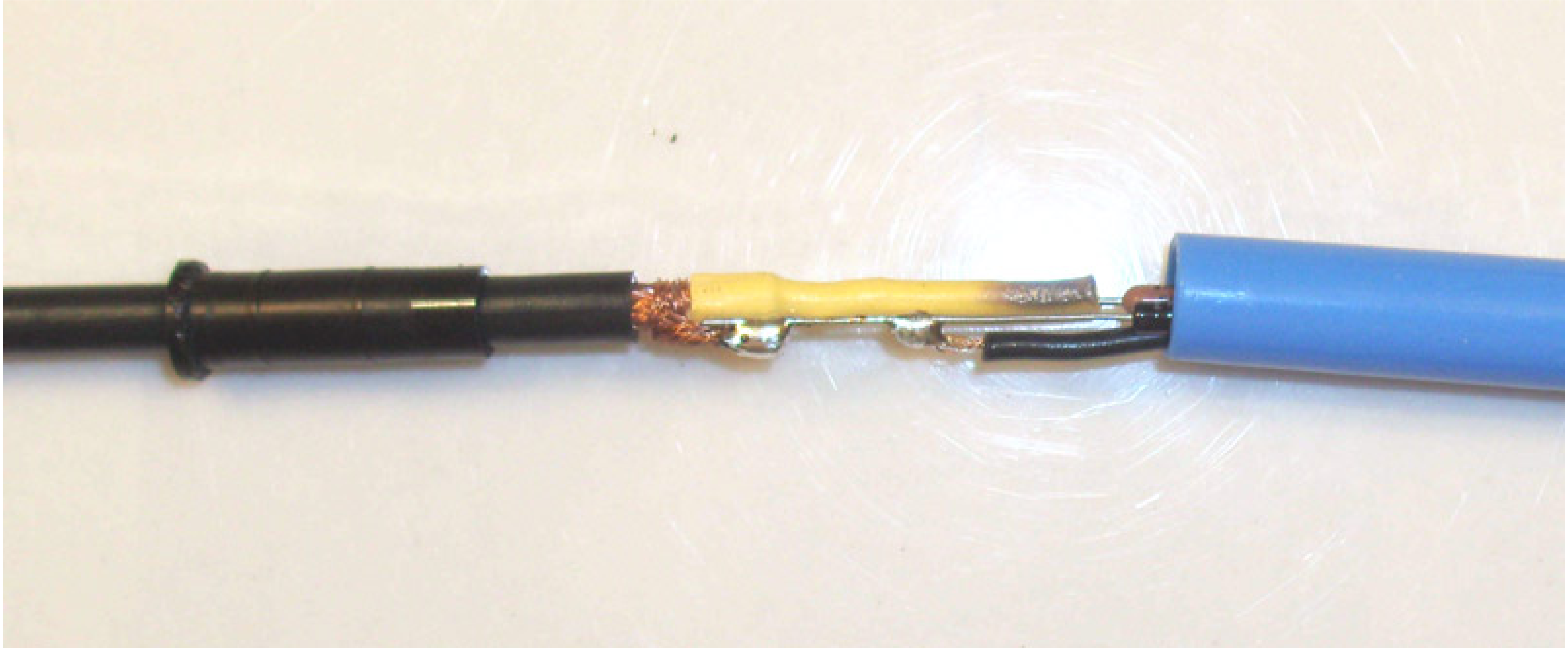

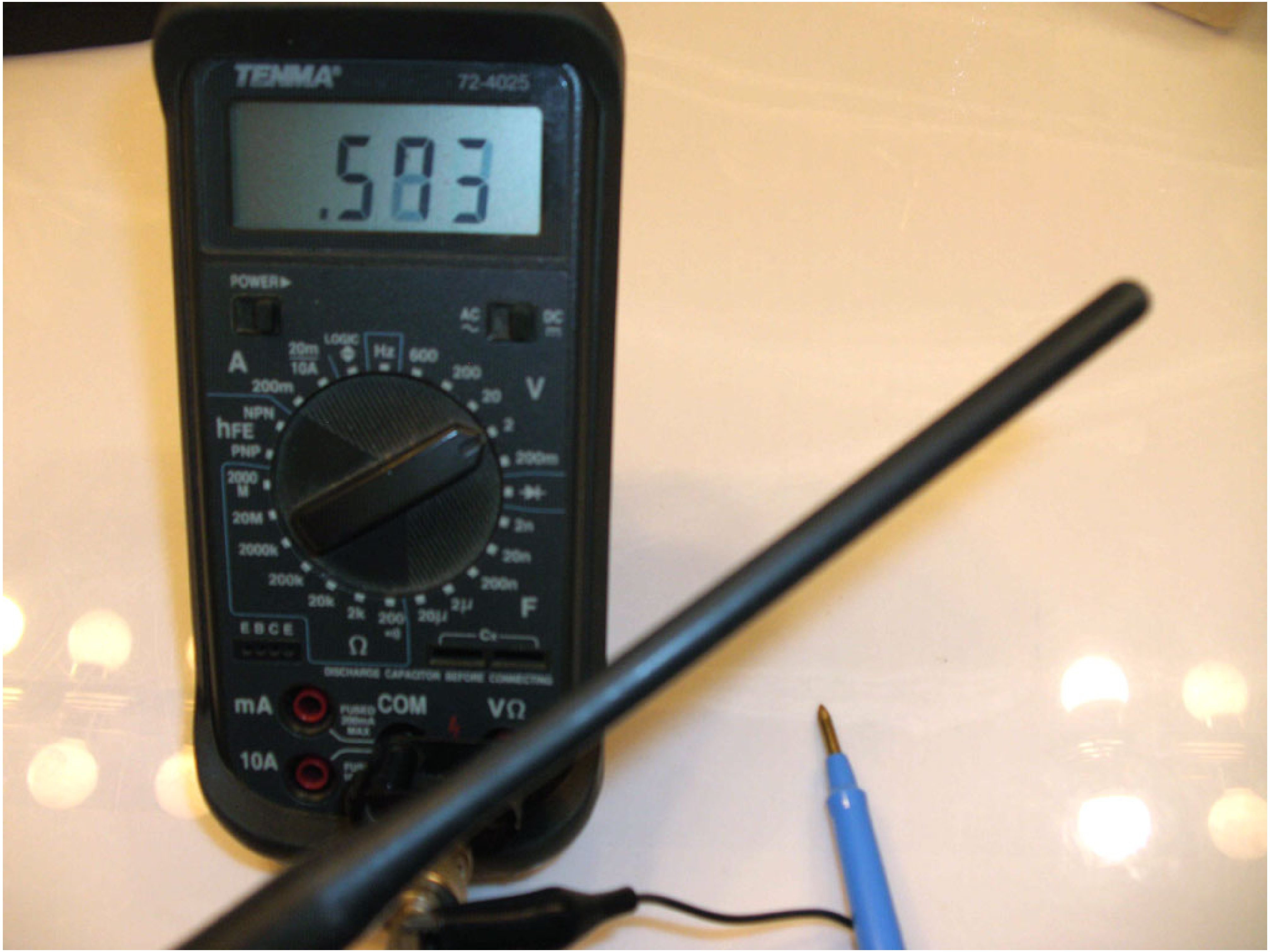
TENMA
72-4025
.583
POWER
AC
DC
A
20m
10A
LOGIC
Hz
600
200
20
V
2
200m
200m
NPN
hFE
PNP
2000
M
20M
2000k
200k
20k
2k
200
2n
20n
200n
2μ
20μ
F
E B C E
Ω
Cx
DISCHARGE CAPACITOR BEFORE CONNECTING
mA
FUSED
200mA
MAX
COM
VΩ
10A

RF Signal Generator

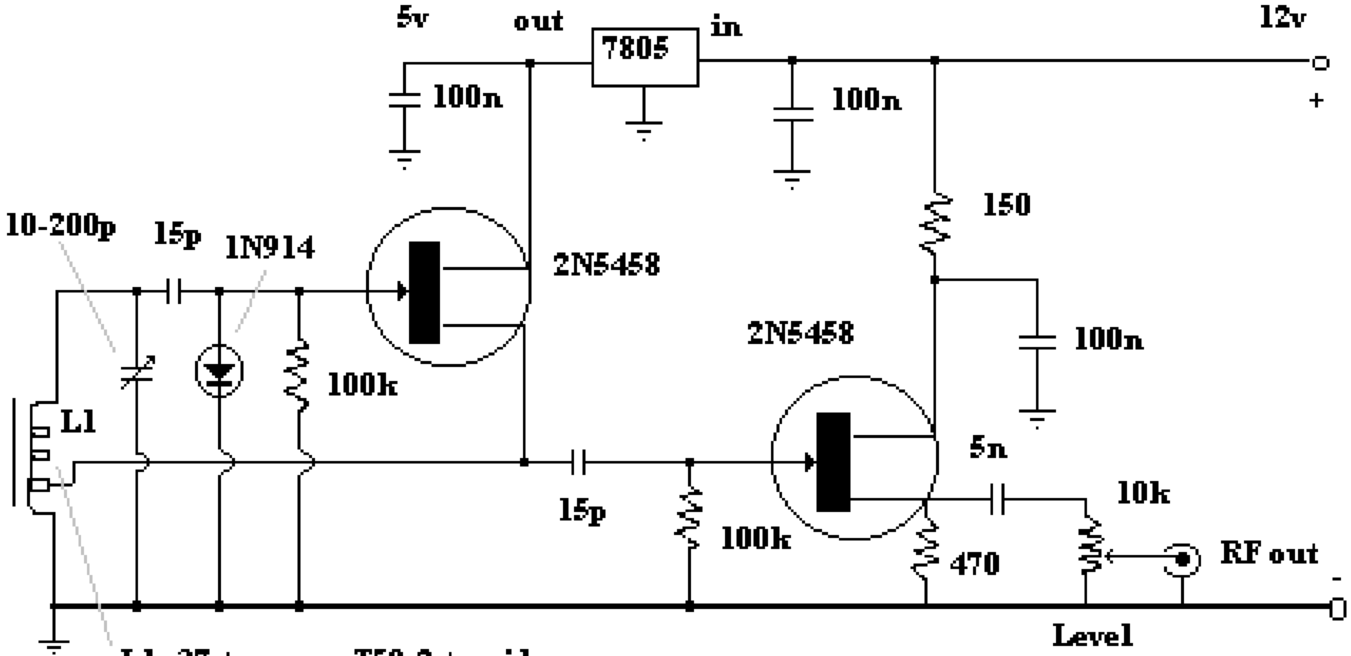
5v
out
7805
in
12v
100n
100n
+
150
10-200p
15p
1N914
2N5458
100k
100n
2N5458
L1
5n
10k
15p
100k
470
RF out
-
Level

L1: 37 turns on T50-2 toroid.
Source tap 12 turns up from earth.

Coverage: 3-12 MHz approx

Notes:

1. A vernier reduction drive is desirable.
2. Build signal generator in die cast aluminuim box for best stability.
3. Can work up to 148 MHz on harmonics if care is taken in construction.
4. Any construction method should work but ensure components are rigid.

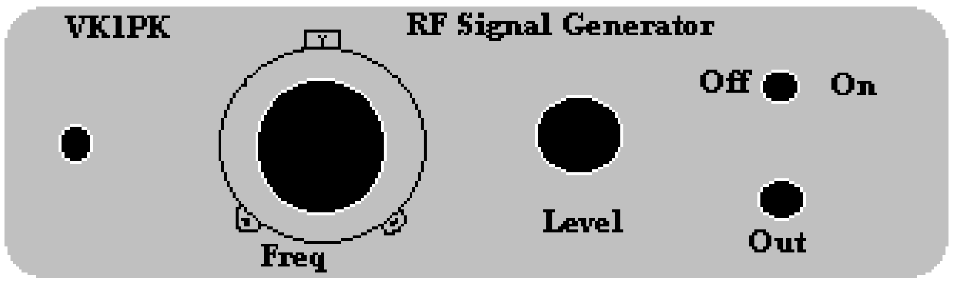
VK1PK
RF Signal Generator
Off
On
Freq
Level
Out

(C) 1998 VK1PK

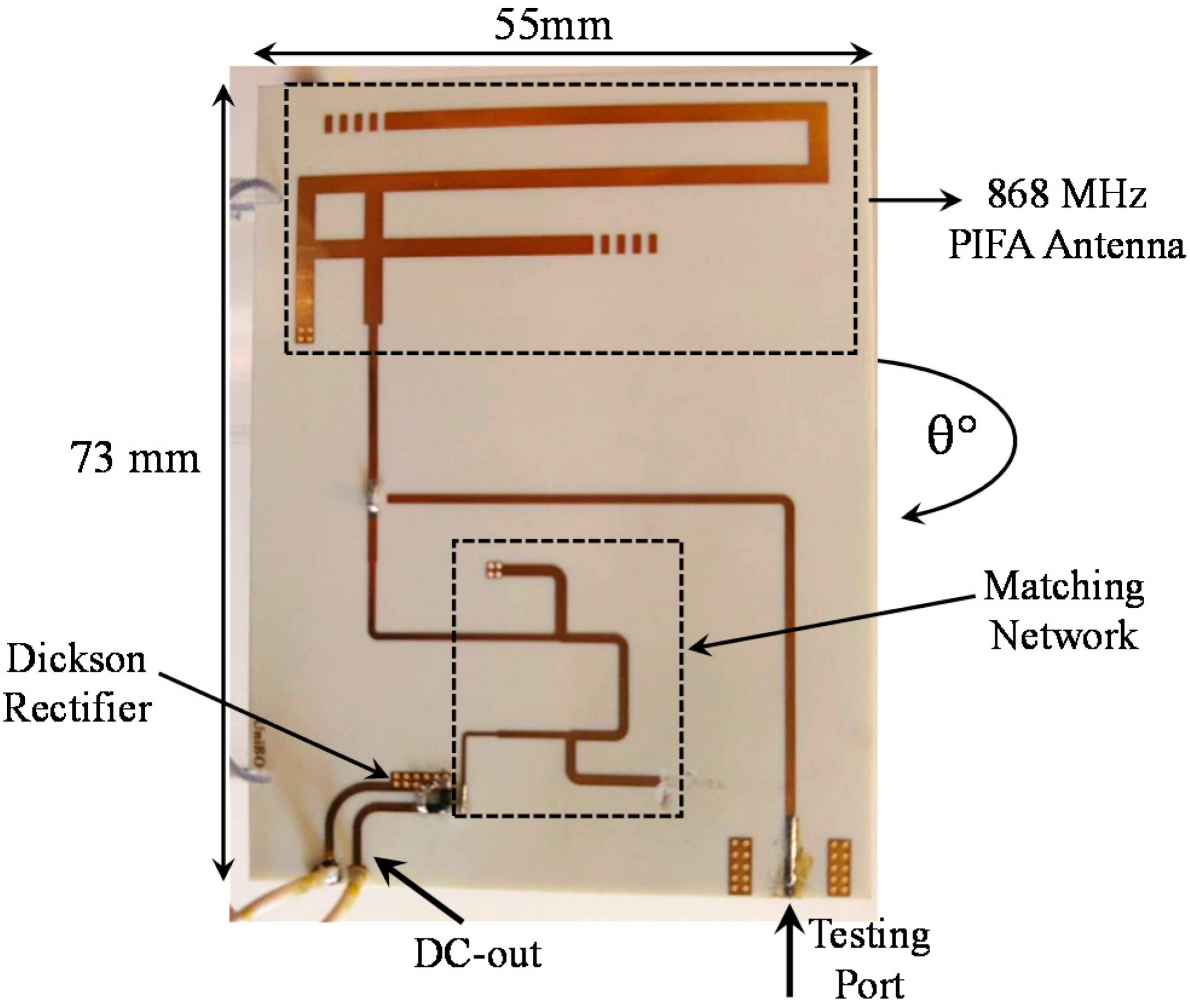
55mm
868 MHz
PIFA Antenna
73 mm
θ°
Matching
Network
Dickson
Rectifier
DC-out
Testing
Port

I used a .02ufd cap because it was the perfect physical size, a 4.7 meg resistor, and a 1N34A diode. I had a nice, flexible, copper nail that would be easy to solder to for a tip.

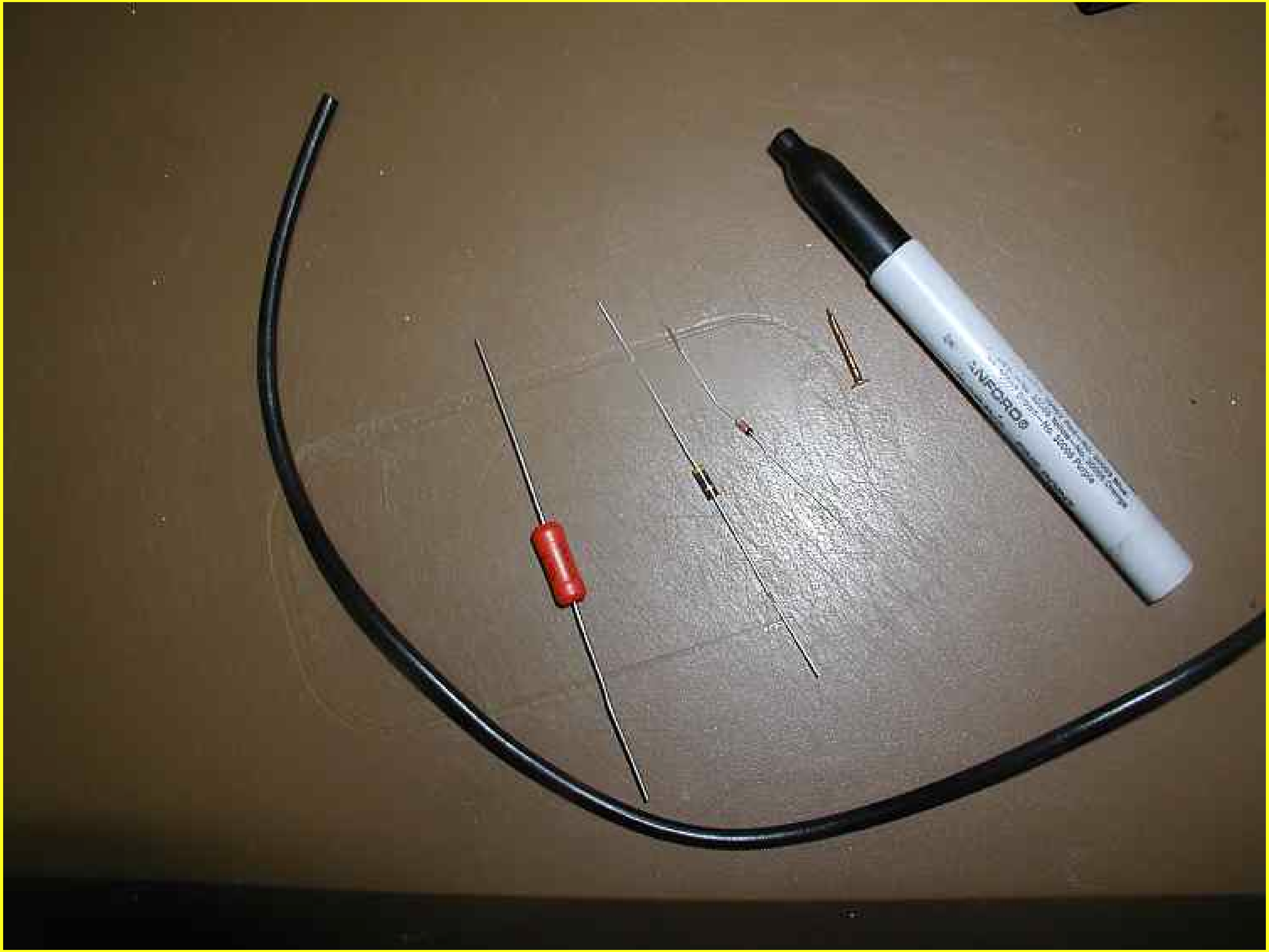

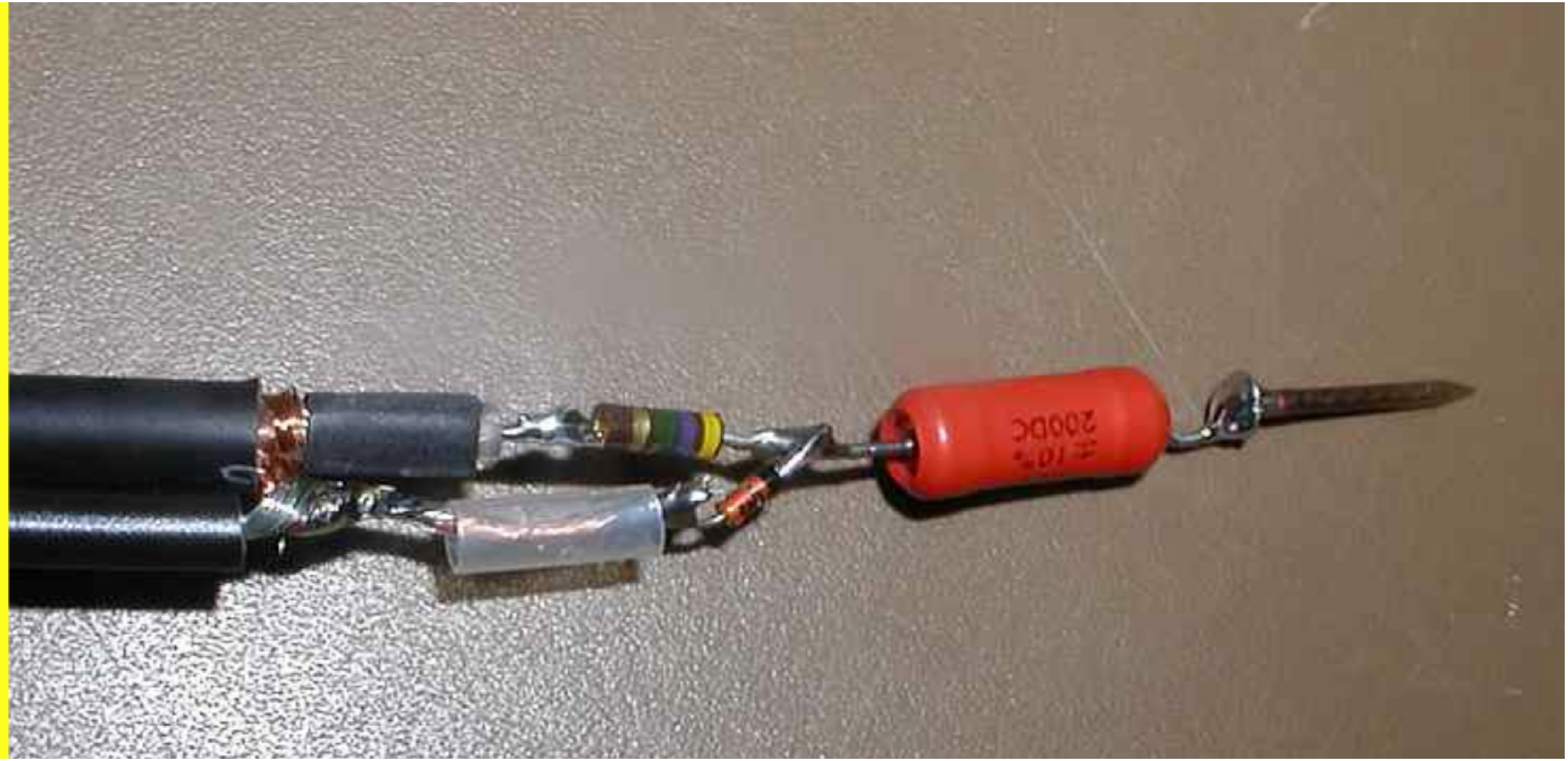
±10%
200DC

R.F. BANDS
R.F. BANDS
3.0-11.0 MC
850-3000 KC B
250-850 KC A
C
D 11-45 MC
E 35-120 MC
UNMOD
MOD
Signal Generator & Tracer
R.F. ATTENUATOR
A.F. ATTENUATOR
PREAMP
OUTPUT
SIGNAL OUTPUT

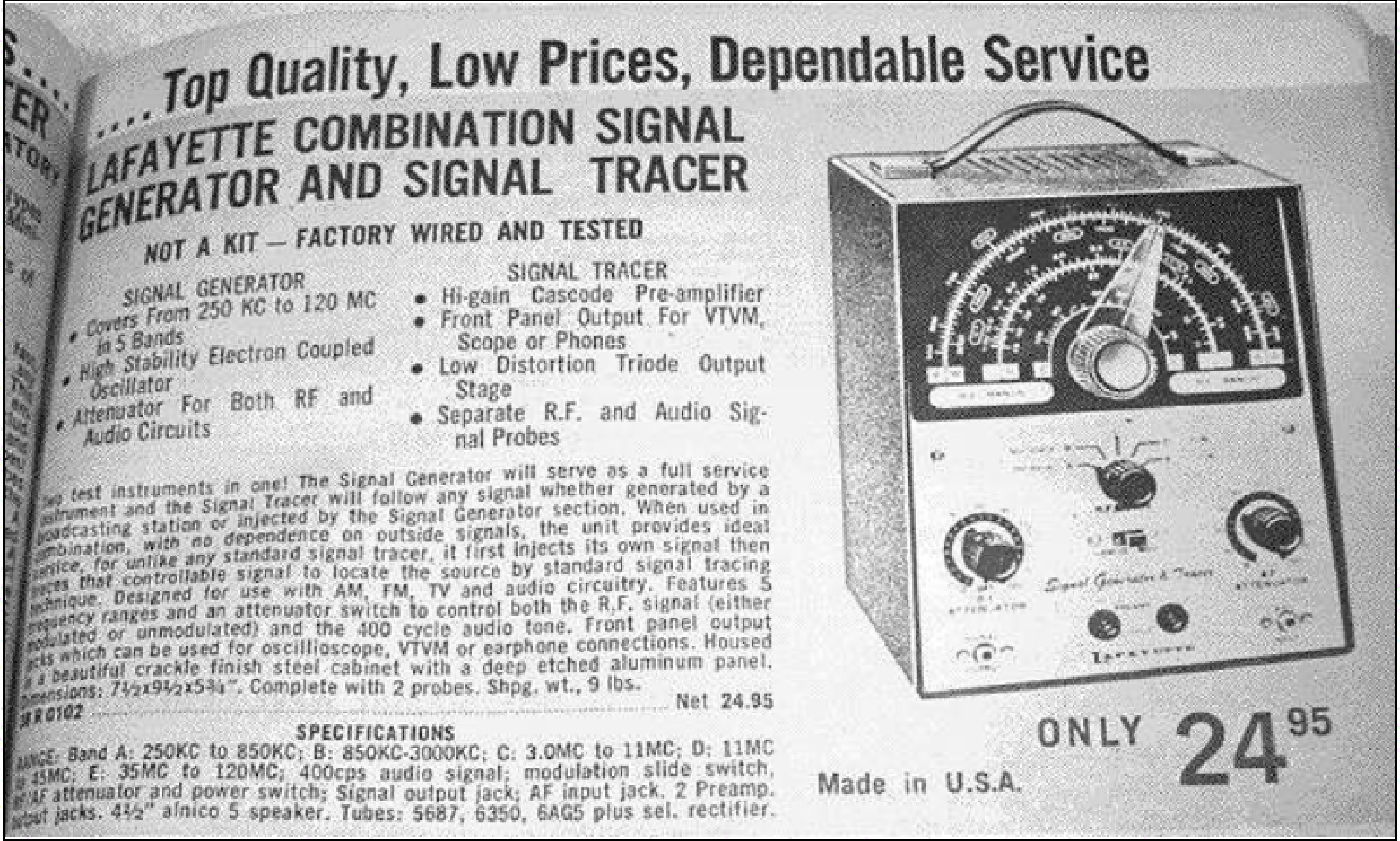

…. Top Quality, Low Prices, Dependable Service

LAFAYETTE COMBINATION SIGNAL GENERATOR AND SIGNAL TRACER

NOT A KIT — FACTORY WIRED AND TESTED

SIGNAL GENERATOR

- Covers From 250 KC to 120 MC in 5 Bands
- High Stability Electron Coupled Oscillator
- Attenuator For Both RF and Audio Circuits

SIGNAL TRACER

- Hi-gain Cascode Pre-amplifier
- Front Panel Output For VTVM, Scope or Phones
- Low Distortion Triode Output Stage
- Separate R.F. and Audio Signal Probes

wo test instruments in one! The Signal Generator will serve as a full service nstrument and the Signal Tracer will follow any signal whether generated by a roadcasting station or injected by the Signal Generator section. When used in mbination, with no dependence on outside signals, the unit provides ideal ervice, for unlike any standard signal tracer, it first injects its own signal then aces that controllable signal to locate the source by standard signal tracing echnique. Designed for use with AM, FM, TV and audio circuitry. Features 5 equency ranges and an attenuator switch to control both the R.F. signal (either odulated or unmodulated) and the 400 cycle audio tone. Front panel output cks which can be used for oscillioscope, VTVM or earphone connections. Housed a beautiful crackle finish steel cabinet with a deep etched aluminum panel. mensions: 7½x9½x5¾". Complete with 2 probes. Shpg. wt., 9 lbs.

R 0102 .... Net 24.95

SPECIFICATIONS

ANGE: Band A: 250KC to 850KC; B: 850KC-3000KC; C: 3.0MC to 11MC; D: 11MC 45MC; E: 35MC to 120MC; 400cps audio signal; modulation slide switch, AF attenuator and power switch; Signal output jack; AF input jack, 2 Preamp. t jacks. 4½" alnico 5 speaker. Tubes: 5687, 6350, 6AG5 plus sel. rectifier.

Made in U.S.A.

ONLY 24^{95}

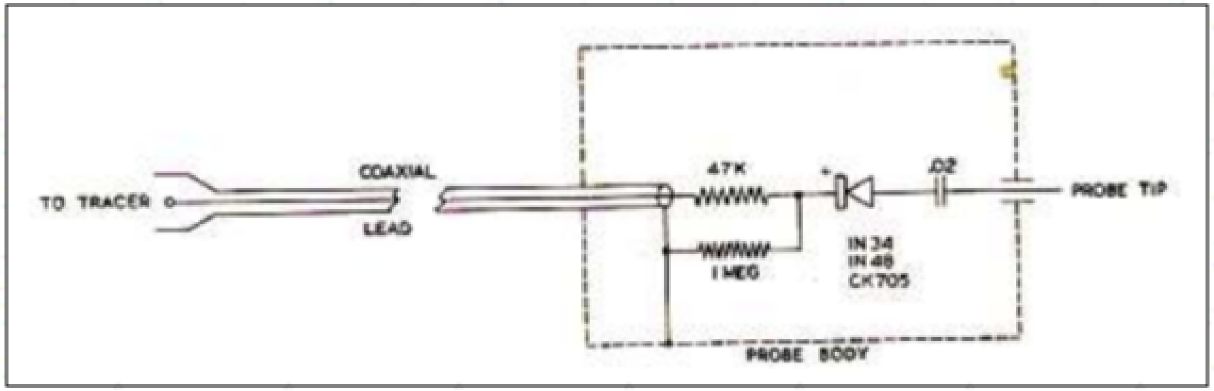
TO TRACER
COAXIAL
LEAD
47K
+
.02
PROBE TIP
1MEG
IN 34
IN 48
CK705
PROBE BODY

Heath T-3 signal tracer probe

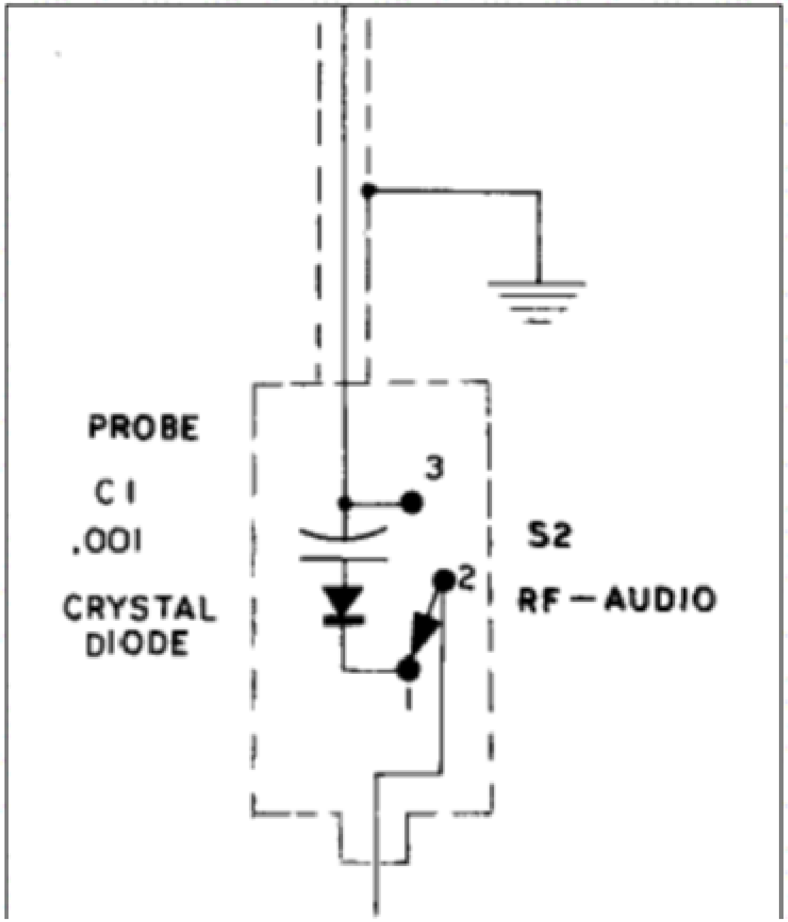
PROBE
C1
.001
CRYSTAL
DIODE
3
2
S2
RF—AUDIO

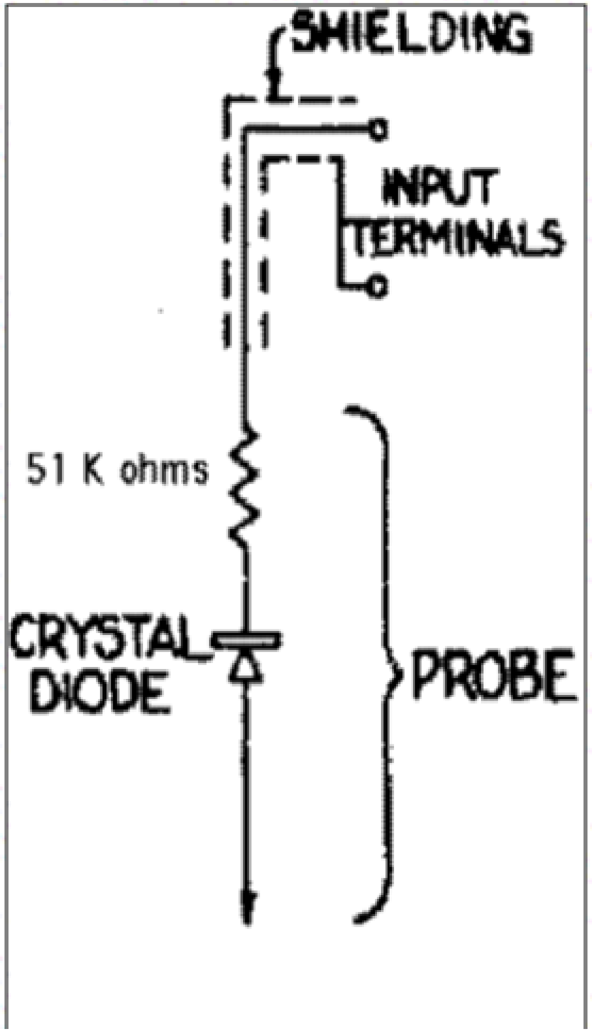
SHIELDING
INPUT
TERMINALS
51 K ohms
CRYSTAL
DIODE
PROBE

Heath T-4 or IT-12 signal tracer probe (left) - - - - Eico 145 signal tracer probe (right)

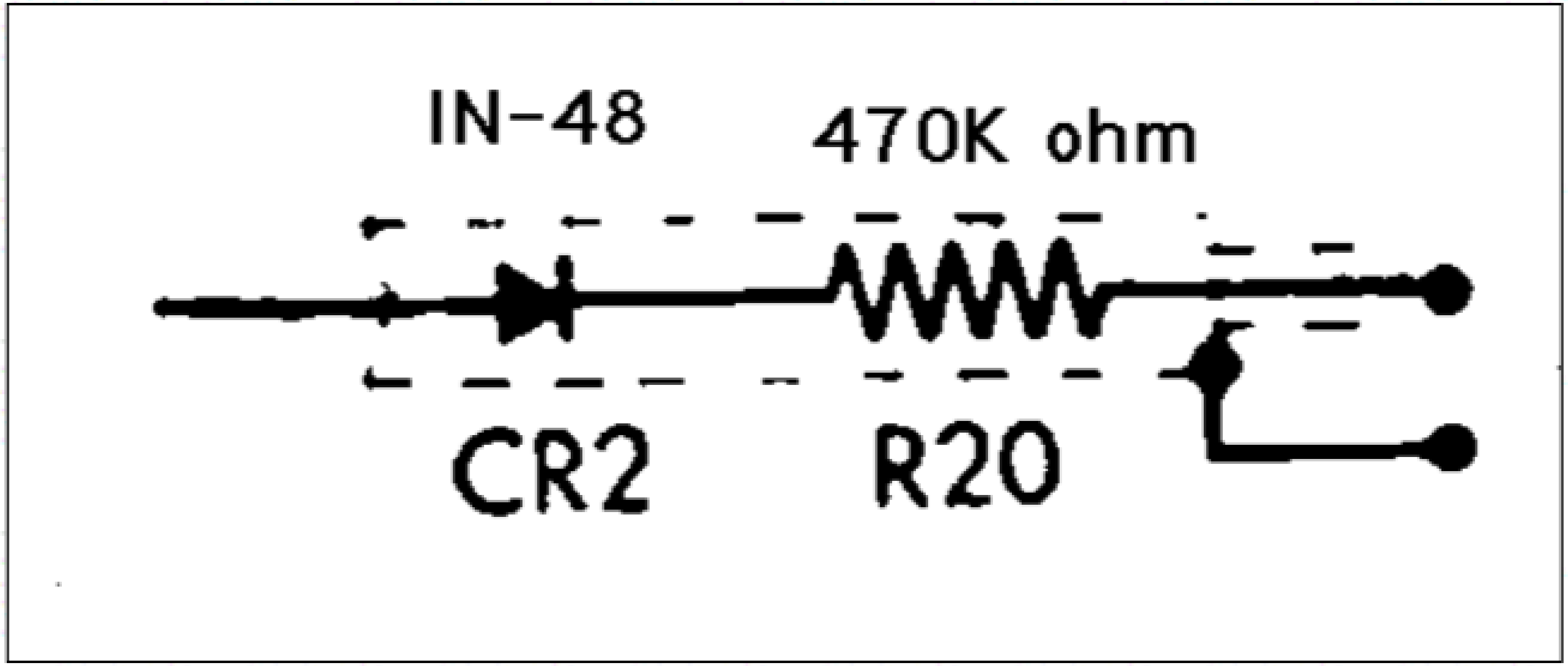
IN-48
470K ohm
CR2
R20

Eico 147a signal tracer probe

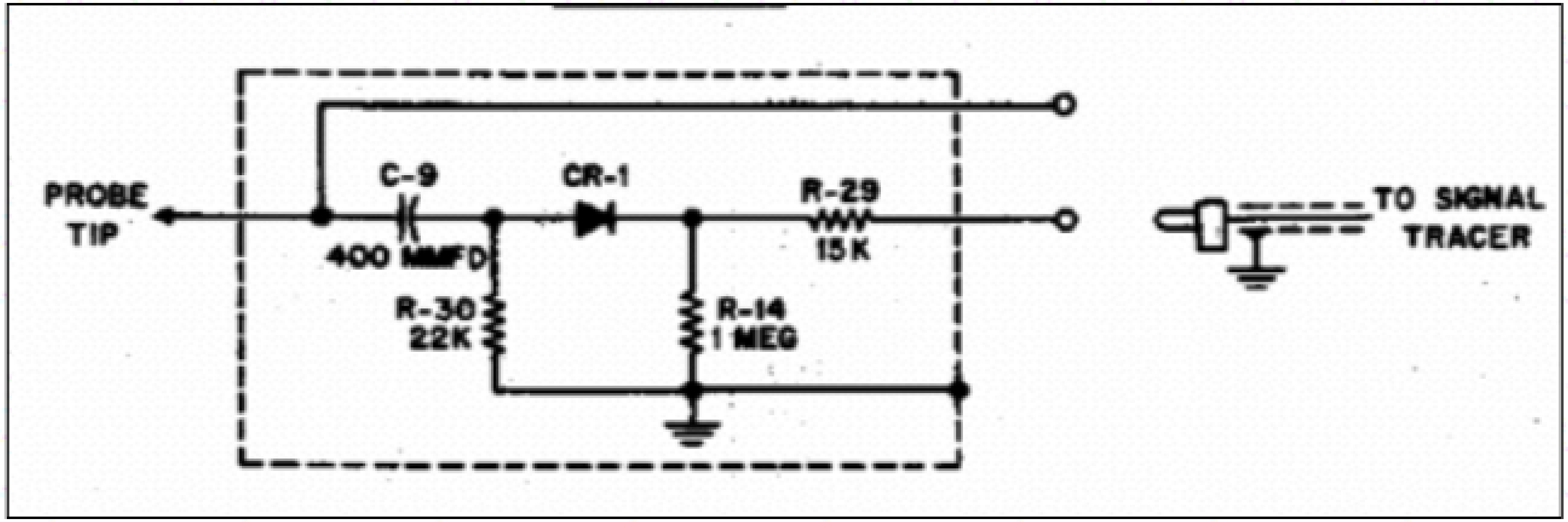
PROBE TIP
C-9
400 MMFD
CR-1
R-29
15K
R-30
22K
R-14
1 MEG
TO SIGNAL TRACER

Knight-kit signal tracer probe

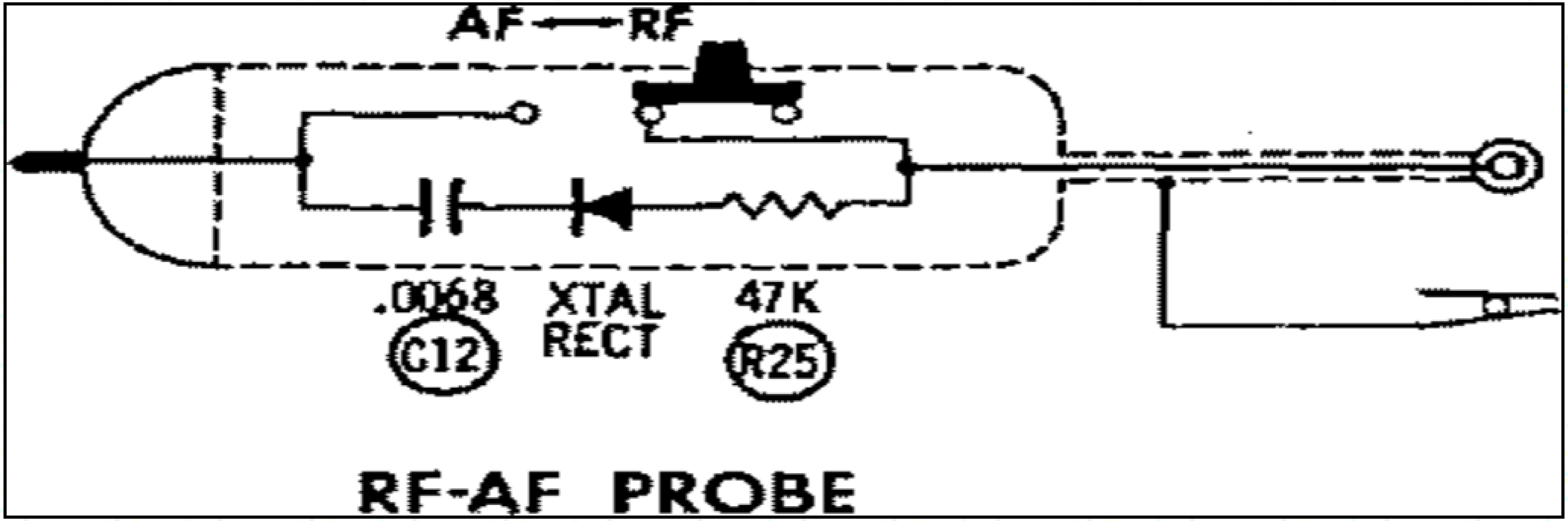
AF ← RF
.0068
C12
XTAL
RECT
47K
R25
RF-AF PROBE

PACO Z-80 signal tracer probe

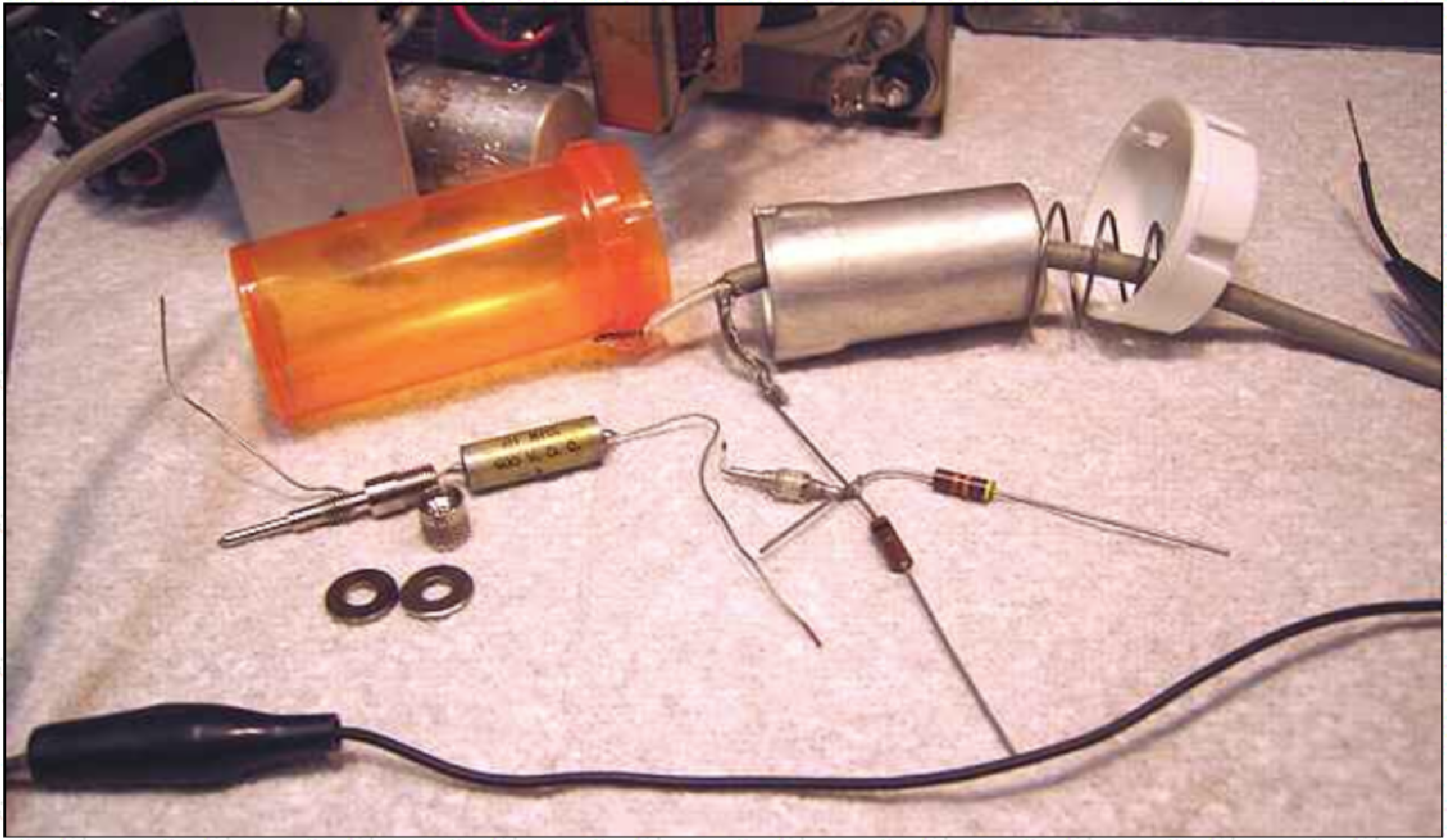

Pill bottle signal tracer probe

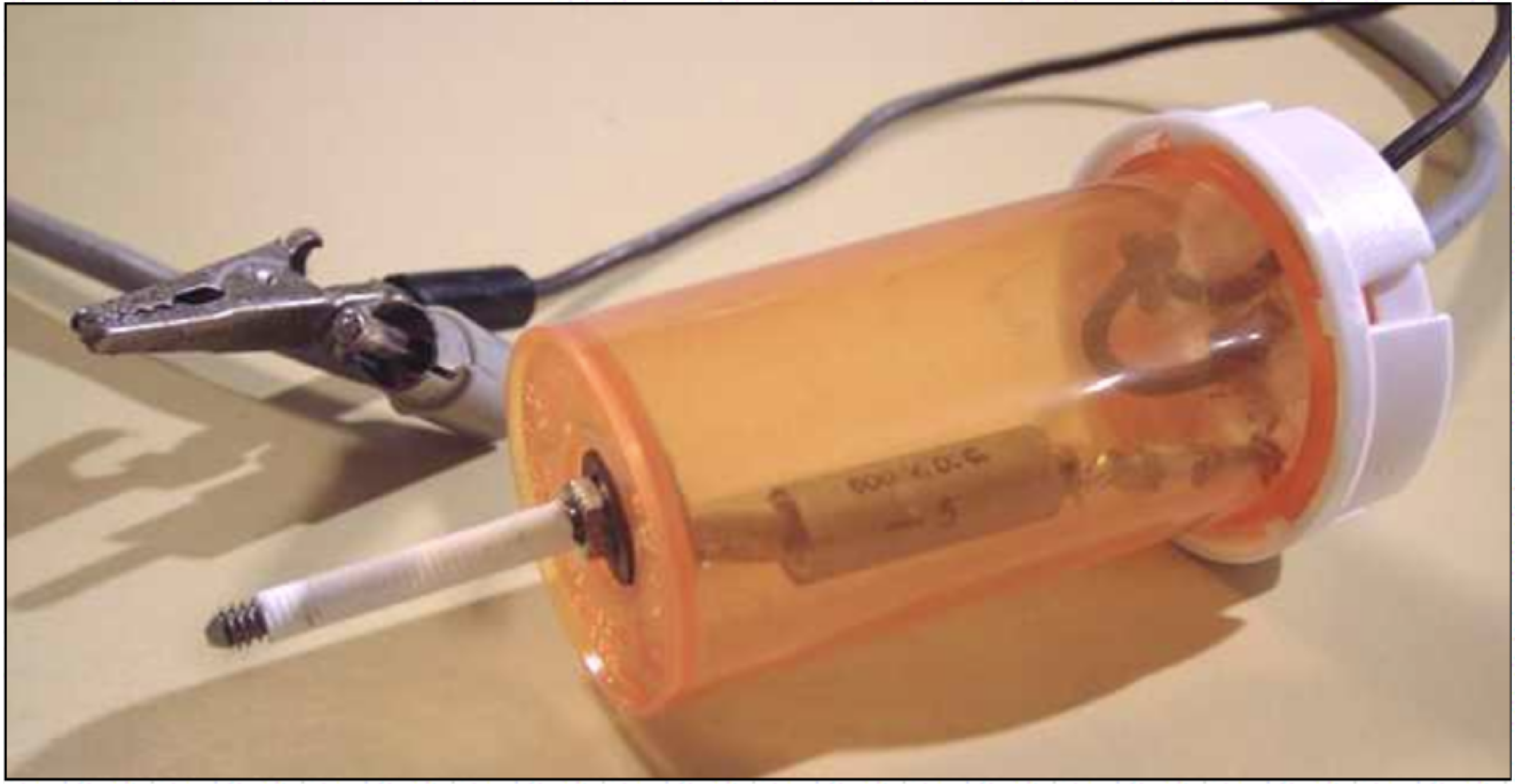

KIS (Keep It Simple) pill-bottle signal tracer probe

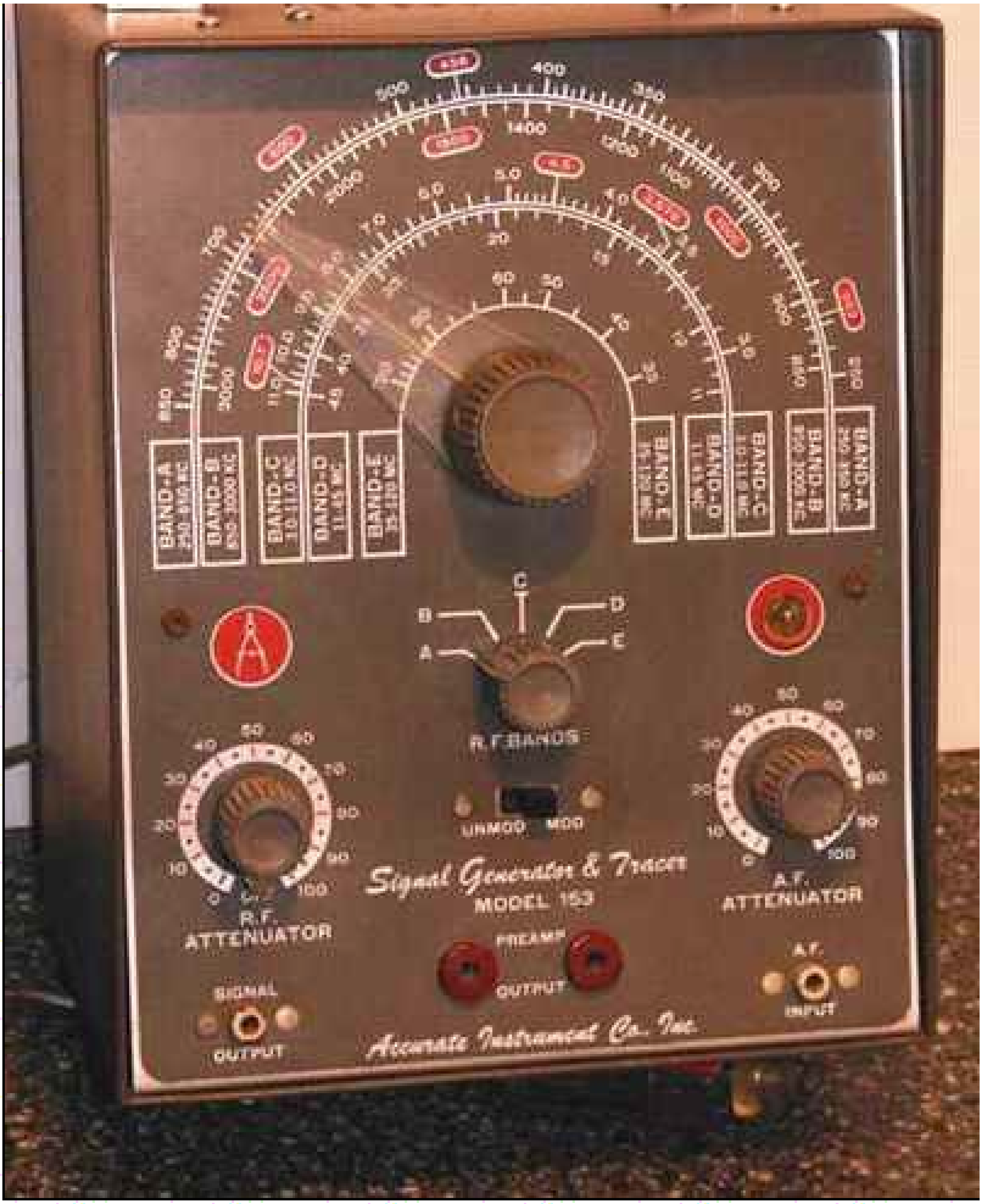
BAND-A
BAND-B
BAND-C
BAND-D
BAND-E
A
B
C
D
E
R.F. BANDS
UNMOD
MOD
Signal Generator & Tracer
MODEL 153
R.F.
ATTENUATOR
A.F.
ATTENUATOR
PREAMP
OUTPUT
SIGNAL
OUTPUT
A.F.
INPUT
Accurate Instrument Co. Inc.

Accurate Instrument model 153 and its RF probe schematic
courtesy of John Lescaud.

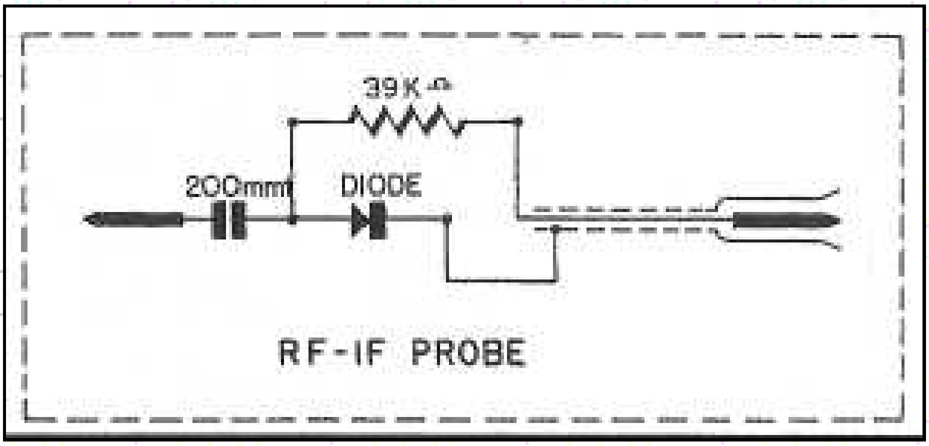
39K Ω
200mmf
DIODE
RF-IF PROBE

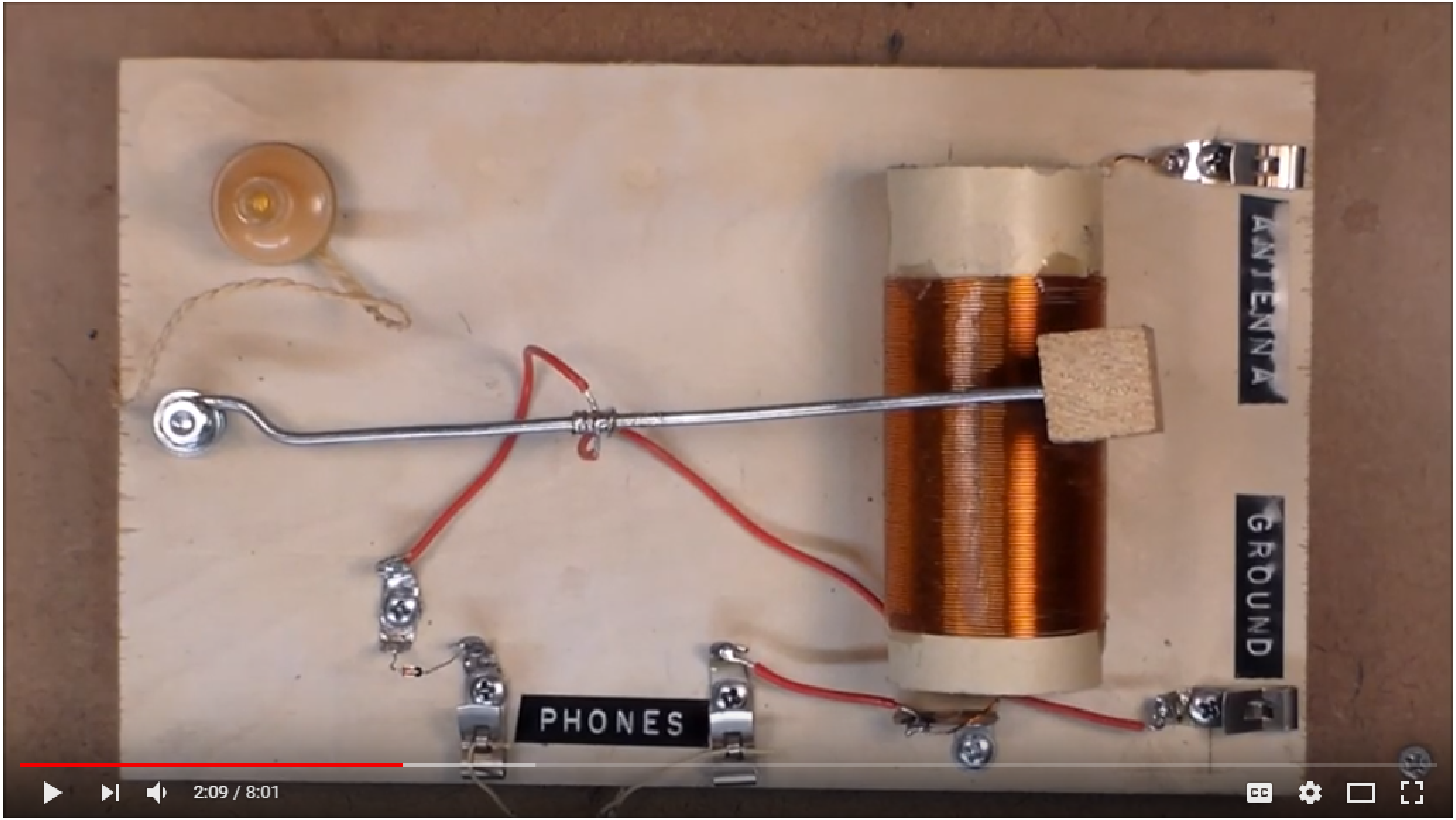
ANTENNA
GROUND
PHONES
2:09 / 8:01

BOY SCOUT CRYSTAL RADIO UPGRADE

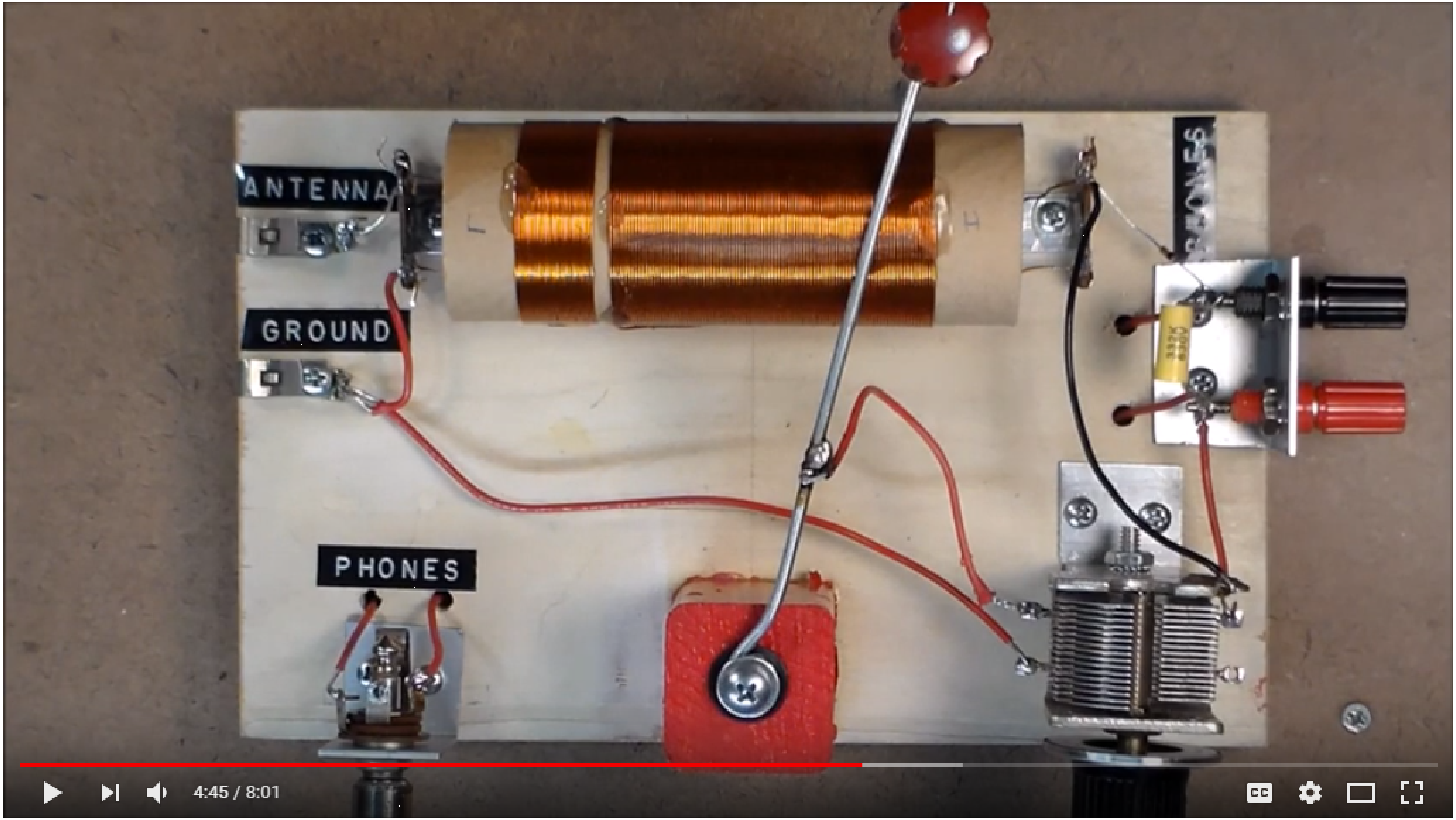
ANTENNA
GROUND
PHONES
PHONES
4:45 / 8:01

BOY SCOUT CRYSTAL RADIO UPGRADE

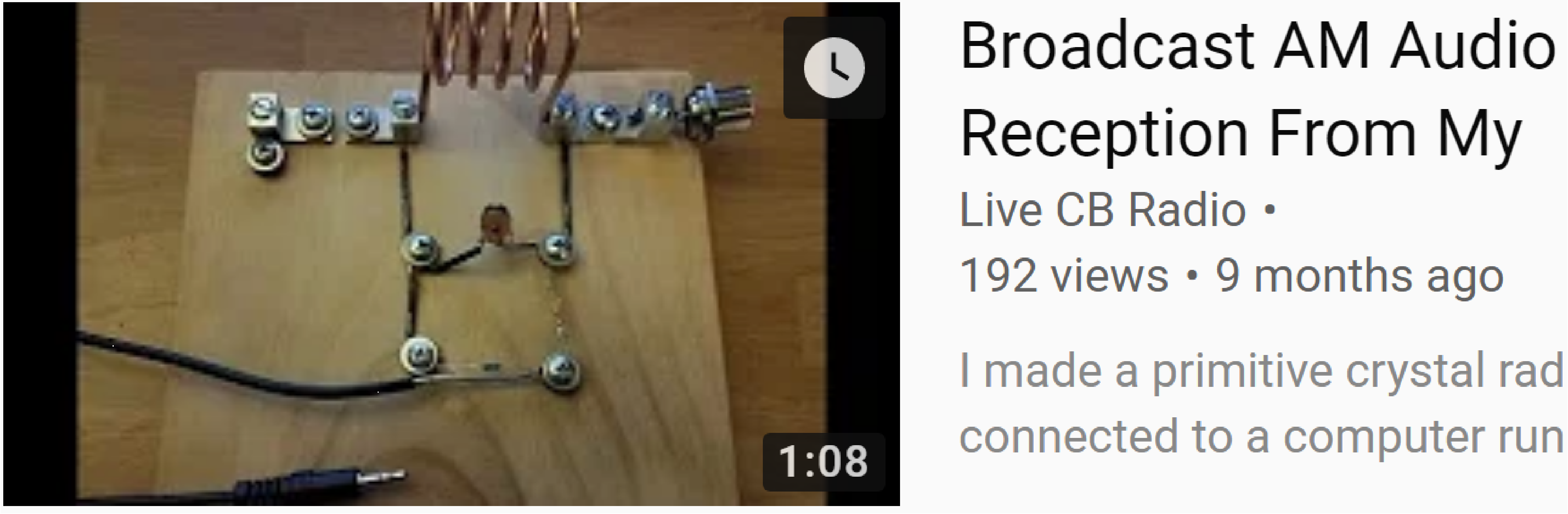

Broadcast AM Audio
Reception From My

Live CB Radio •
192 views • 9 months ago

I made a primitive crystal rad
connected to a computer run

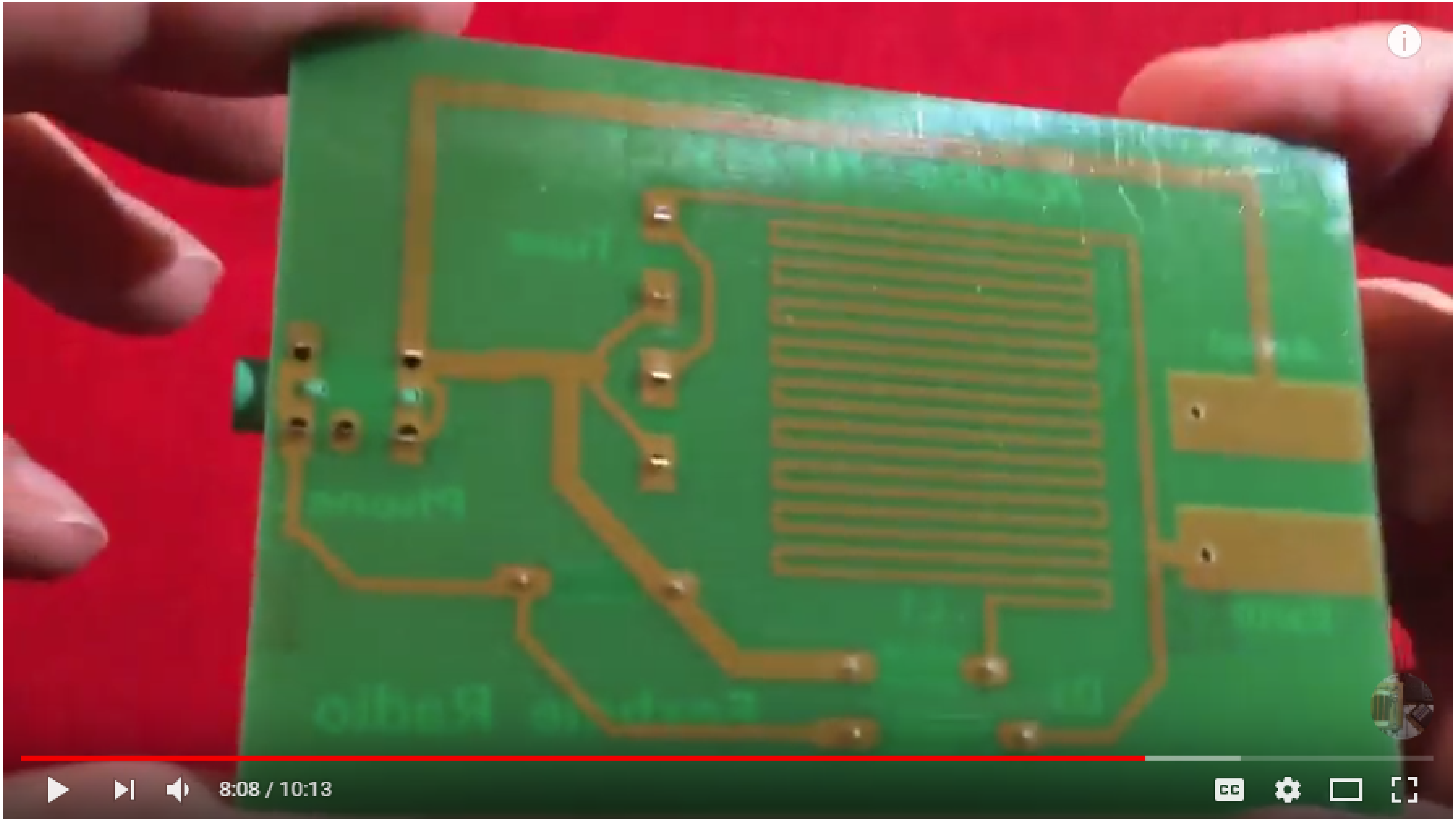

Make: Foxhole Crystal Portable Radio Kit Build & Review

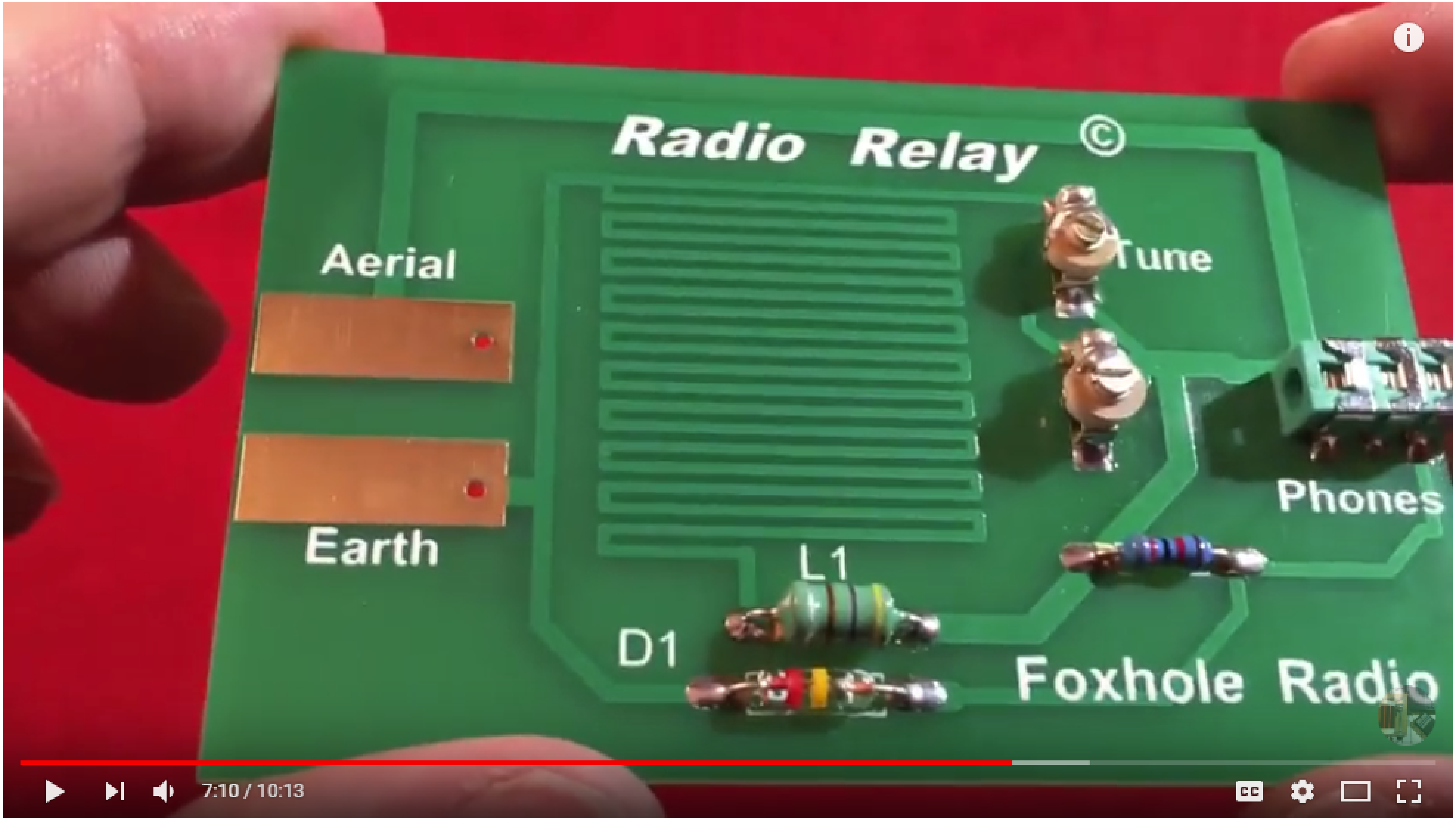
Radio Relay ©
Aerial
Earth
Tune
Phones
L1
D1
Foxhole Radio
7:10 / 10:13

Make: Foxhole Crystal Portable Radio Kit Build & Review

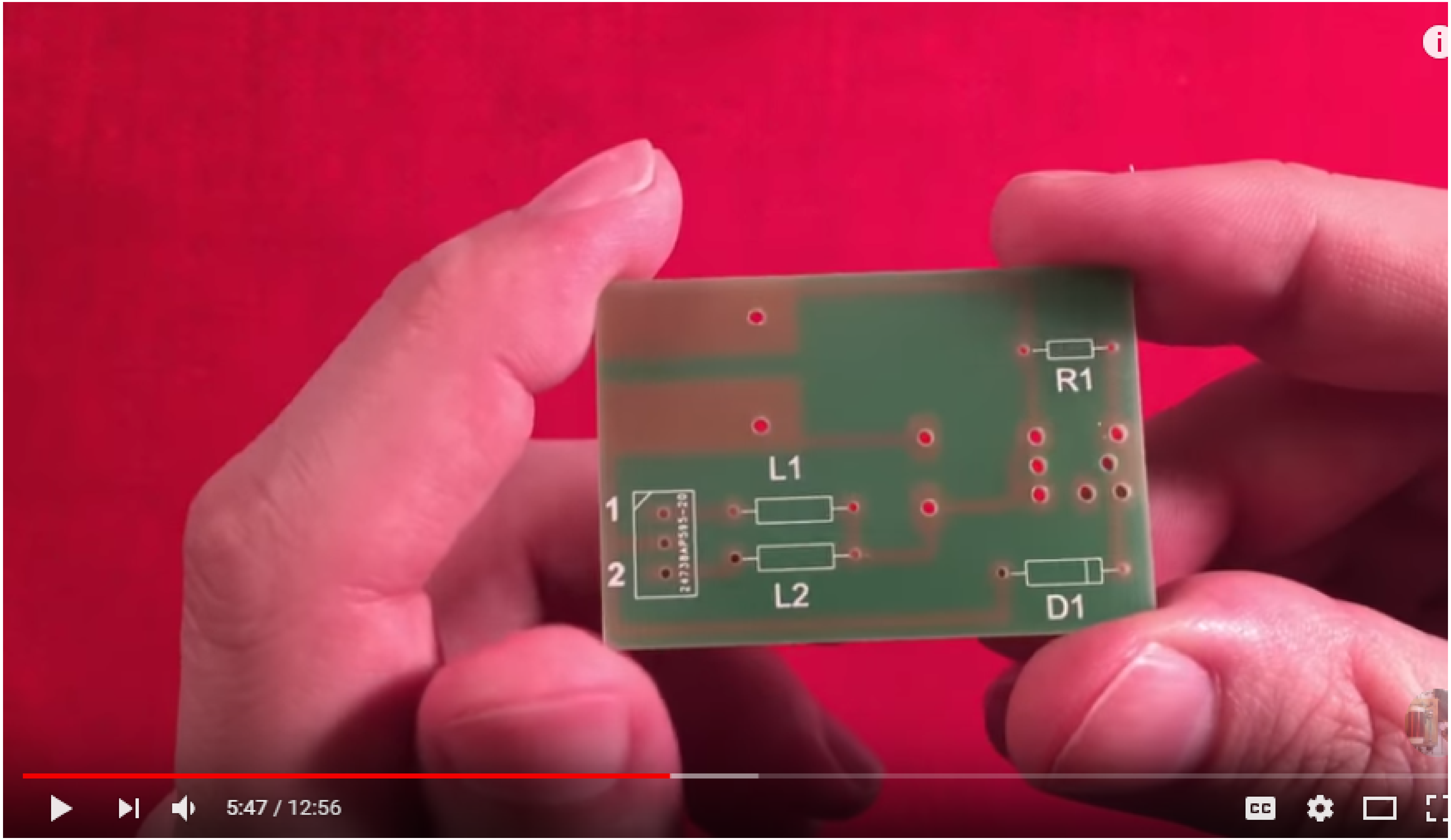
R1
L1
1
2
L2
D1
5:47 / 12:56

Make: DIY Matchbox Crystal Radio Kit Build and Review

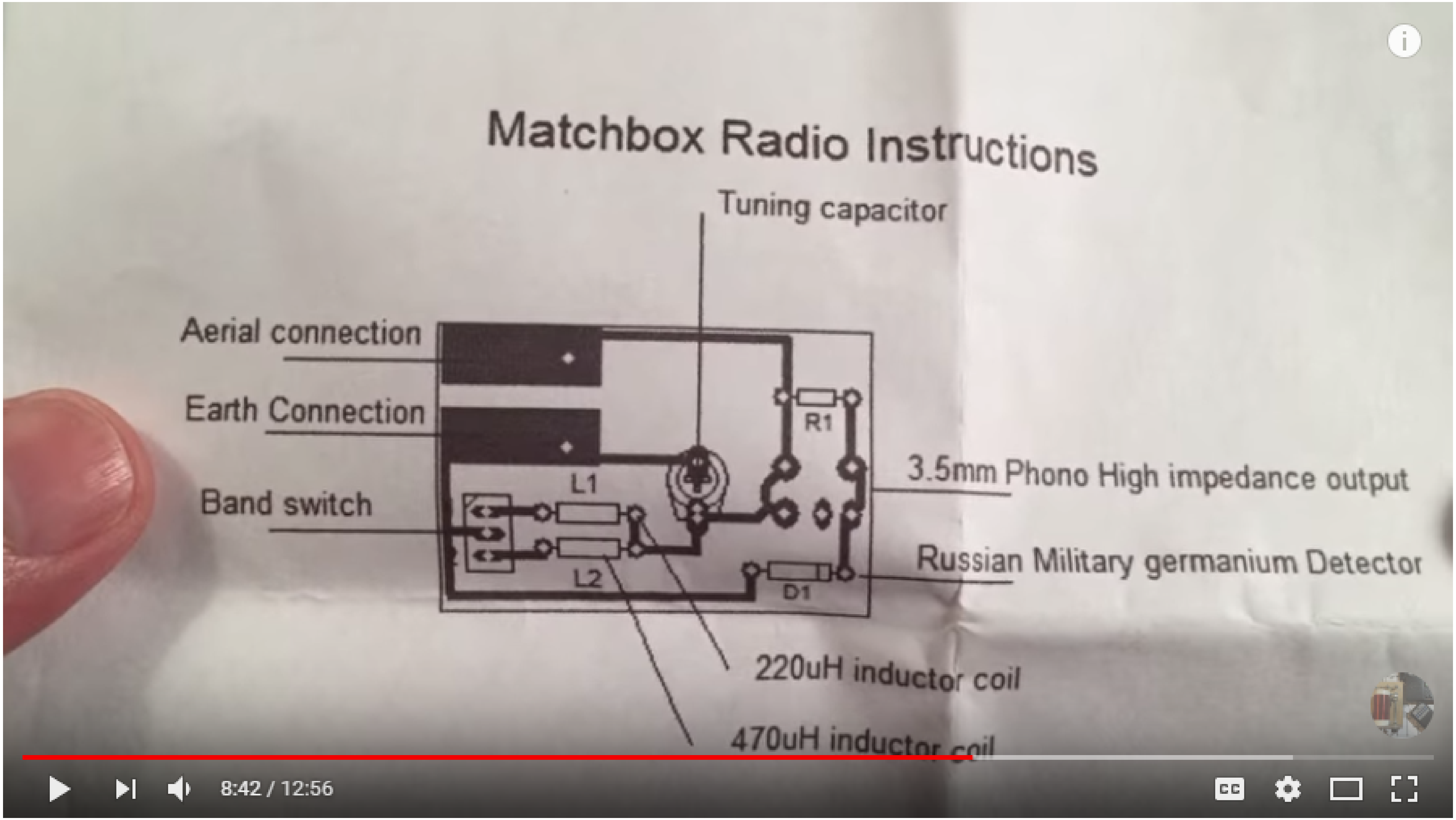

Matchbox Radio Instructions

Tuning capacitor
Aerial connection
Earth Connection
Band switch
L1
L2
R1
D1
3.5mm Phono High impedance output
Russian Military germanium Detector
220uH inductor coil
470uH inductor coil
8:42 / 12:56

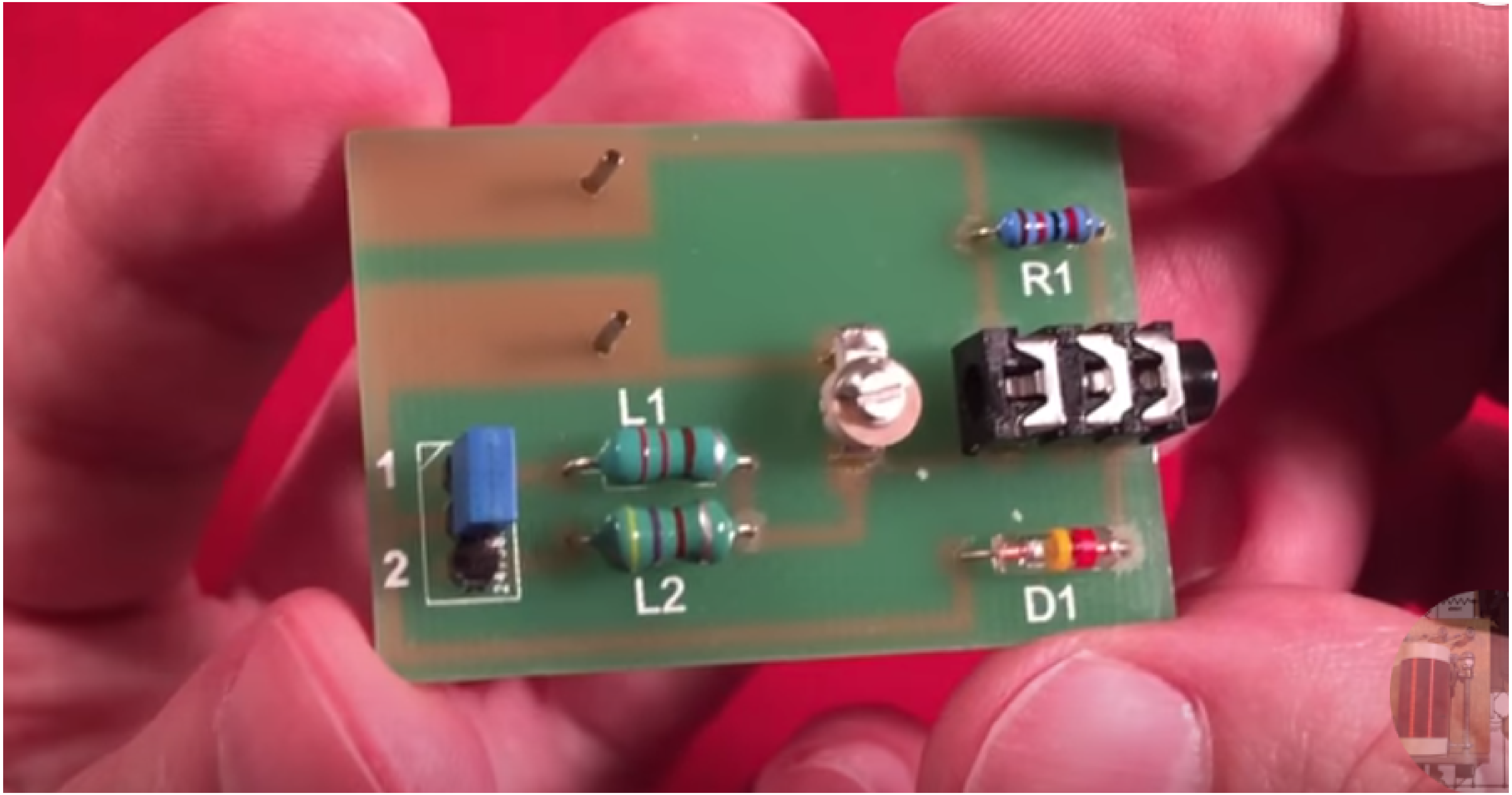
R1
L1
1
2
L2
D1

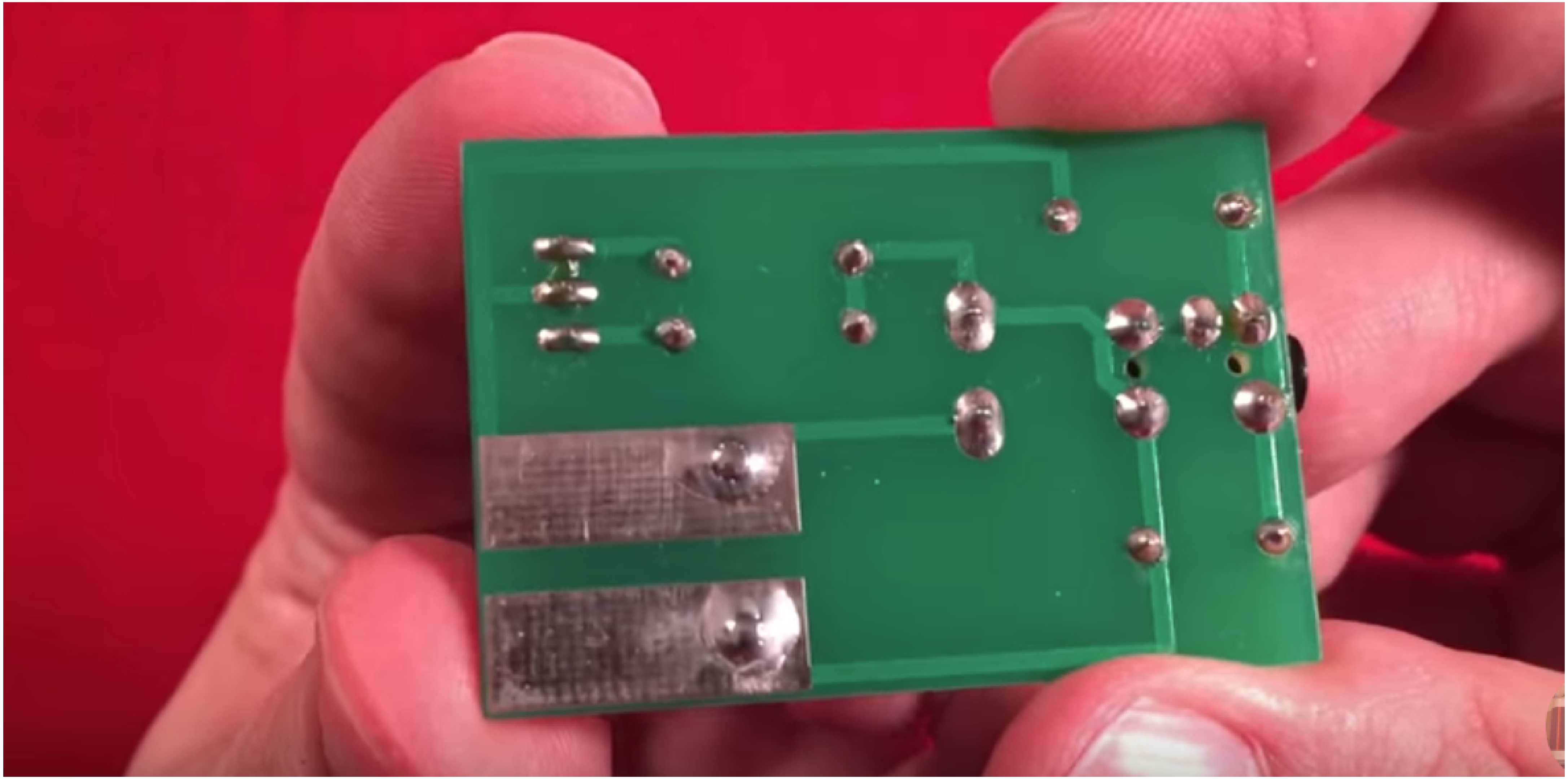

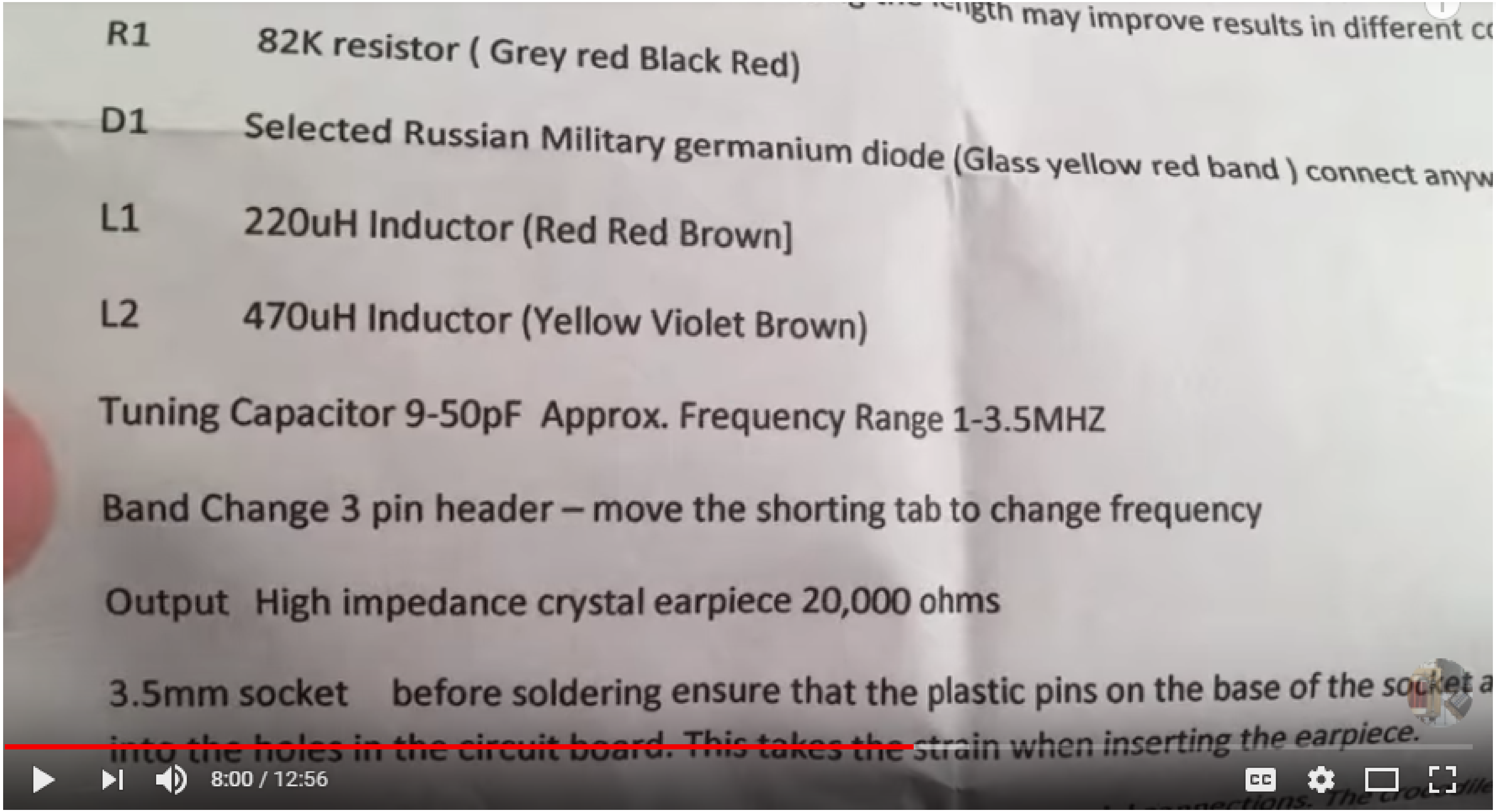

ngth may improve results in different c

R1	82K resistor (Grey red Black Red)
D1	Selected Russian Military germanium diode (Glass yellow red band) connect anyw
L1	220uH Inductor (Red Red Brown]
L2	470uH Inductor (Yellow Violet Brown)

Tuning Capacitor 9-50pF Approx. Frequency Range 1-3.5MHZ

Band Change 3 pin header – move the shorting tab to change frequency

Output High impedance crystal earpiece 20,000 ohms

3.5mm socket before soldering ensure that the plastic pins on the base of the socket a into the holes in the circuit board. This takes the strain when inserting the earpiece.

connections. The crocodile

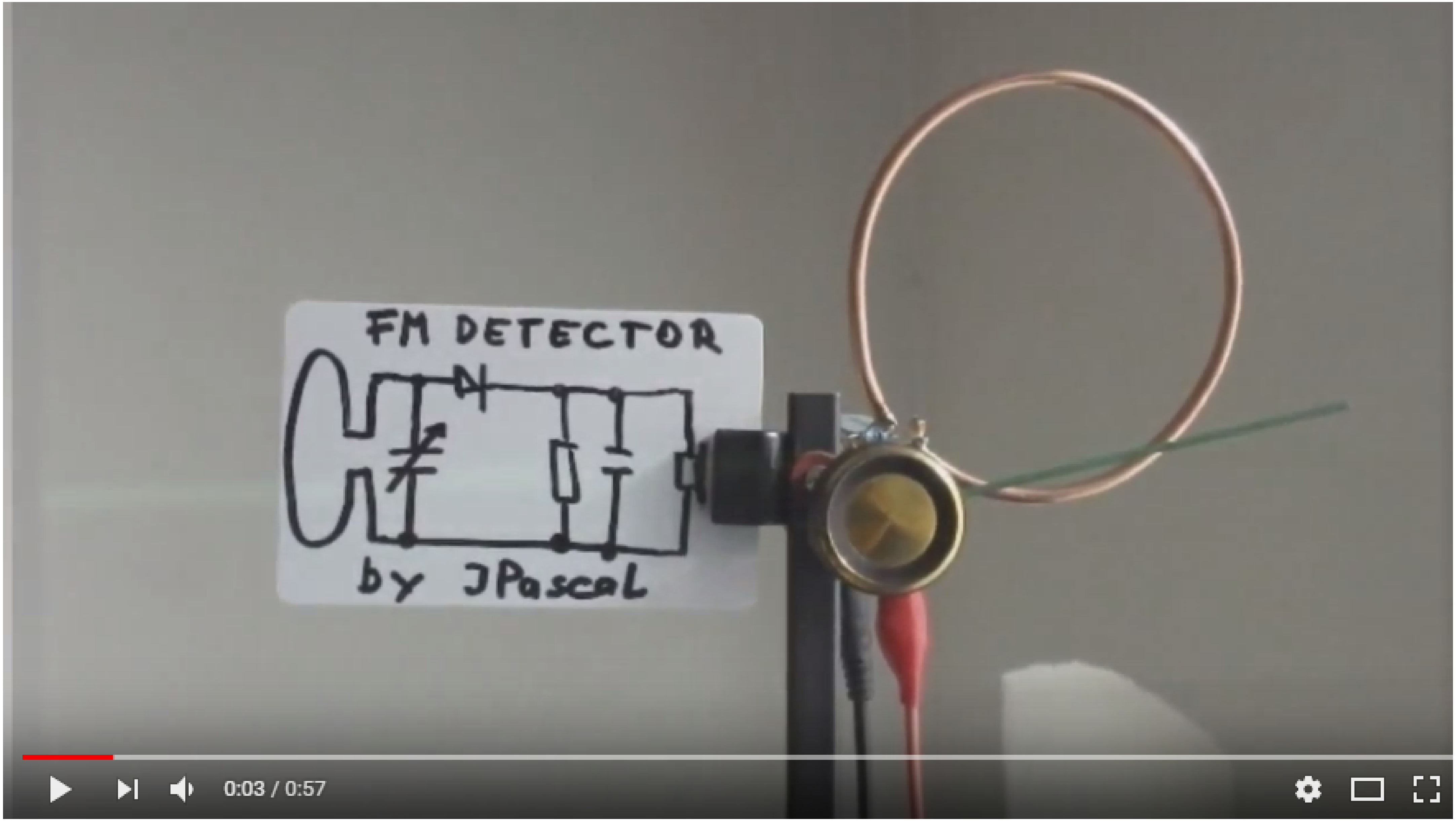
FM DETECTOR
by JPascal
0:03 / 0:57

FM Crystal Detector Live Recordings

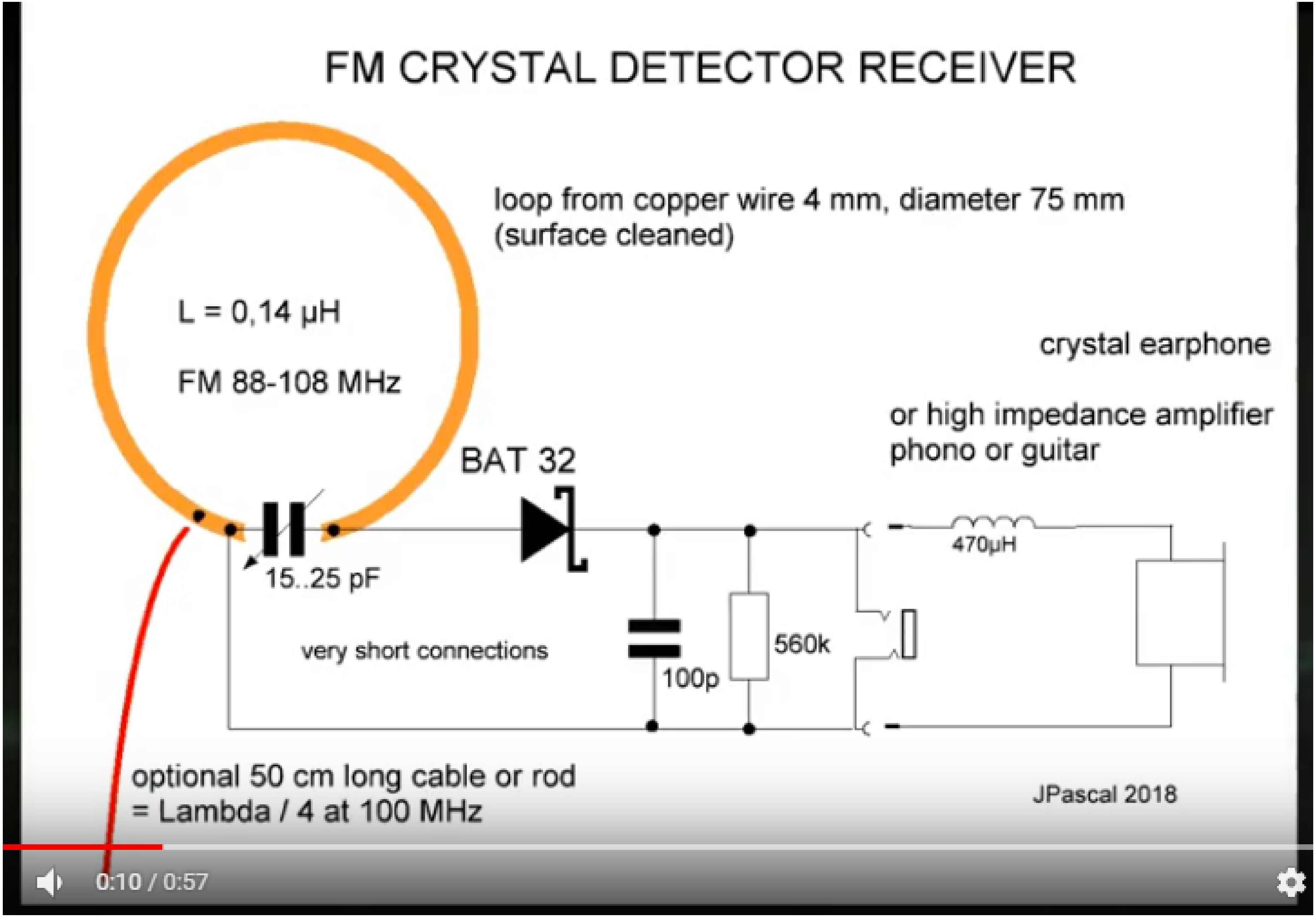

FM CRYSTAL DETECTOR RECEIVER

loop from copper wire 4 mm, diameter 75 mm
(surface cleaned)

crystal earphone

or high impedance amplifier
phono or guitar

very short connections

optional 50 cm long cable or rod
= Lambda / 4 at 100 MHz

JPascal 2018

L = 0,14 µH
FM 88-108 MHz
BAT 32
15..25 pF
100p
560k
470µH
0:10 / 0:57

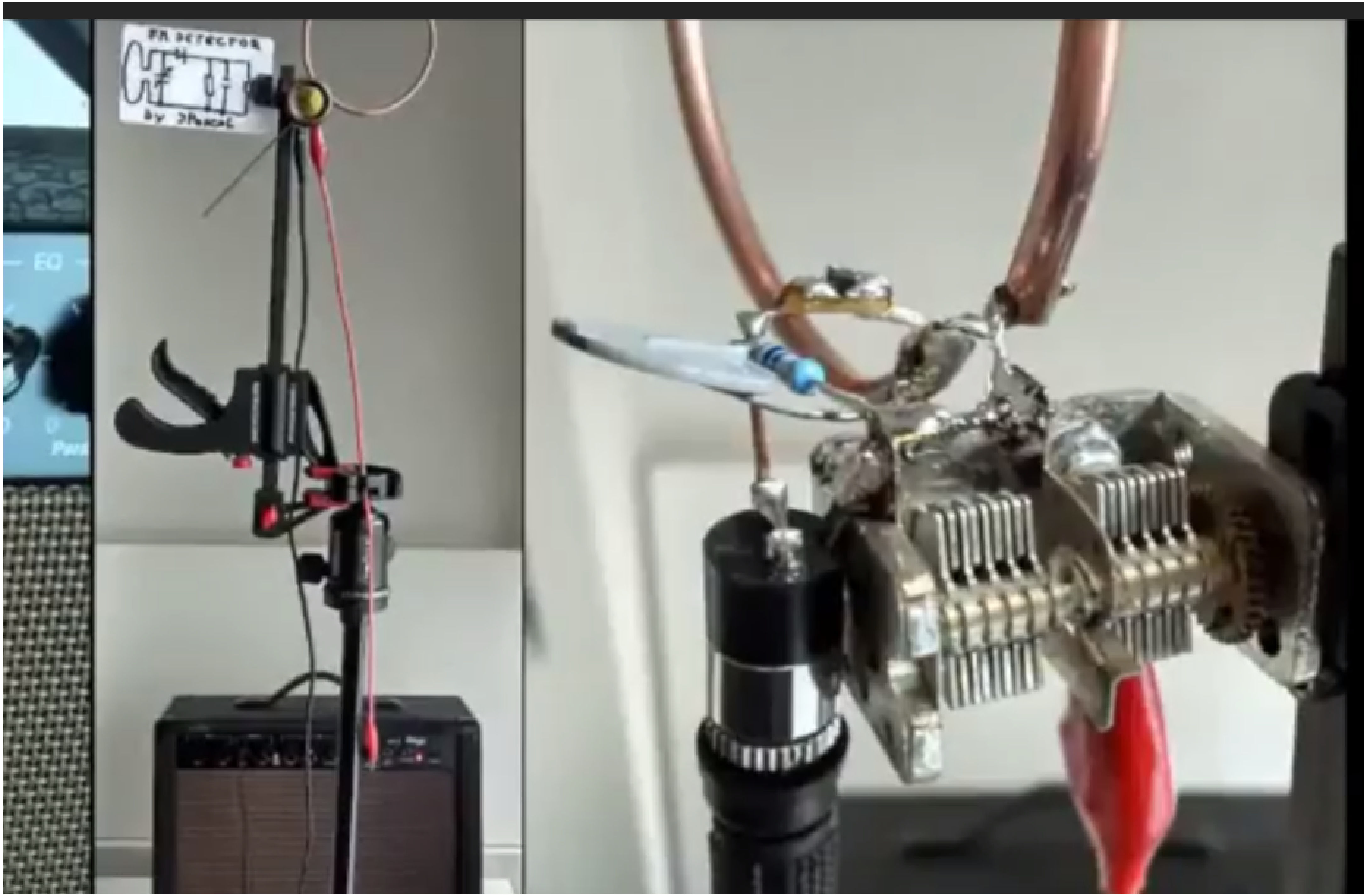
FM DETECTOR
EQ

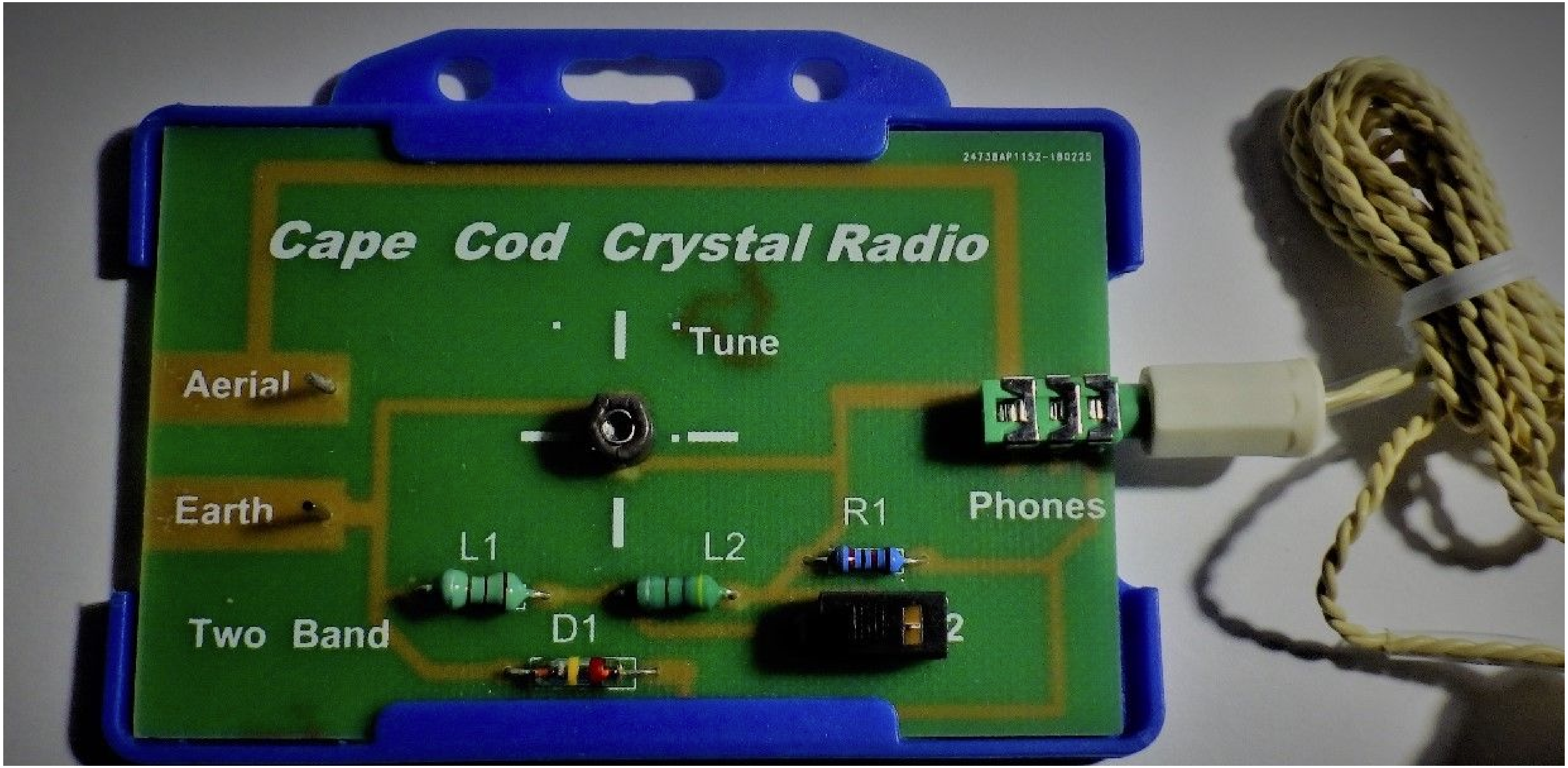
24738AP1152-180225
Cape Cod Crystal Radio
Tune
Aerial
Earth
Phones
R1
L1
L2
Two Band
D1
2

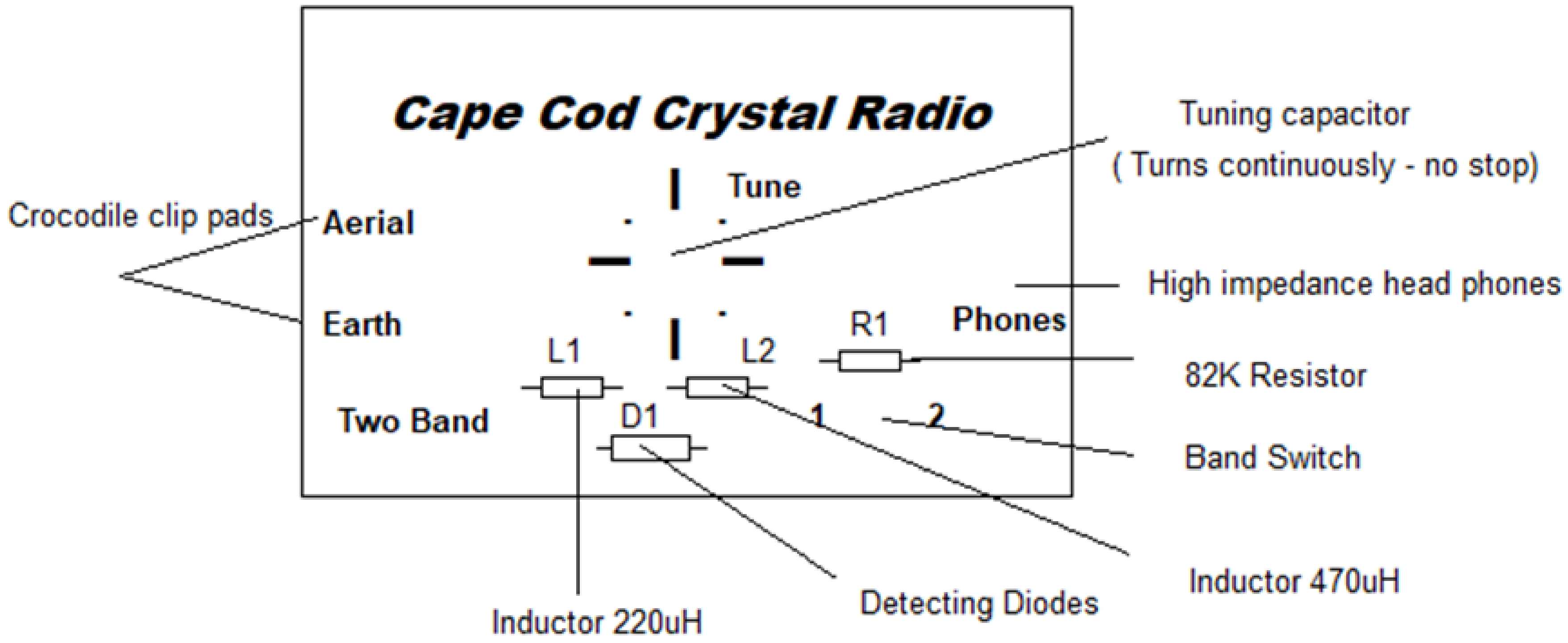
Cape Cod Crystal Radio
Crocodile clip pads
Aerial
Earth
Tune
Tuning capacitor
(Turns continuously - no stop)
High impedance head phones
Phones
R1
82K Resistor
L1
L2
Two Band
D1
1
2
Band Switch
Inductor 470uH
Detecting Diodes
Inductor 220uH

1. Solder D1 the germanium diode. RED band to the right.

2. Solder L1 220uH inductor (Red Red Brown Silver)

3. Solder L2 470uH Inductor (Red Violet Brown Silver)

3. Solder R1 82K ohm resistor (Grey Red Black Red) Above Switch

4. Solder band change switch, make sure it is parallel with the board

5. Solder the 3.5mm earphone socket, ensure it is parallel to the board before you solder and the locating pins align with board holes.

6. Solder the variable capacitor 120pF. Any way it is not polarity sensitive.

7. Two silver pins are provided for aerial and earth connections,just clip the crocodile clips to these pins or, leave these pins off and connect to the pads on the board. For portable use and younger children best to leave these pins off the board.

Connect the crocodile clips for the earth and aerial wires. Strip approximately 25mm of the insulation, thread through hole and crimp both the insulated part and the bare wire with pliers.

L1 is an inductor, a 220uH coil in series series with L2 a 470uH inductor. The switch shorts out the L2 in ther band 1 position. D1 is a specially selected germanium diode to give optimum performance in this particular circuit. R1 is a 82K resistor across the high impedance output. The variable capacitor is a 9 - 120pF capacitor

Getting Started . Plug the headphones into the socket. Holding your thumb behind the socket when plugging in will avoid any stress on the solder joints.To get the strongest signals in your area you will need to experiment if

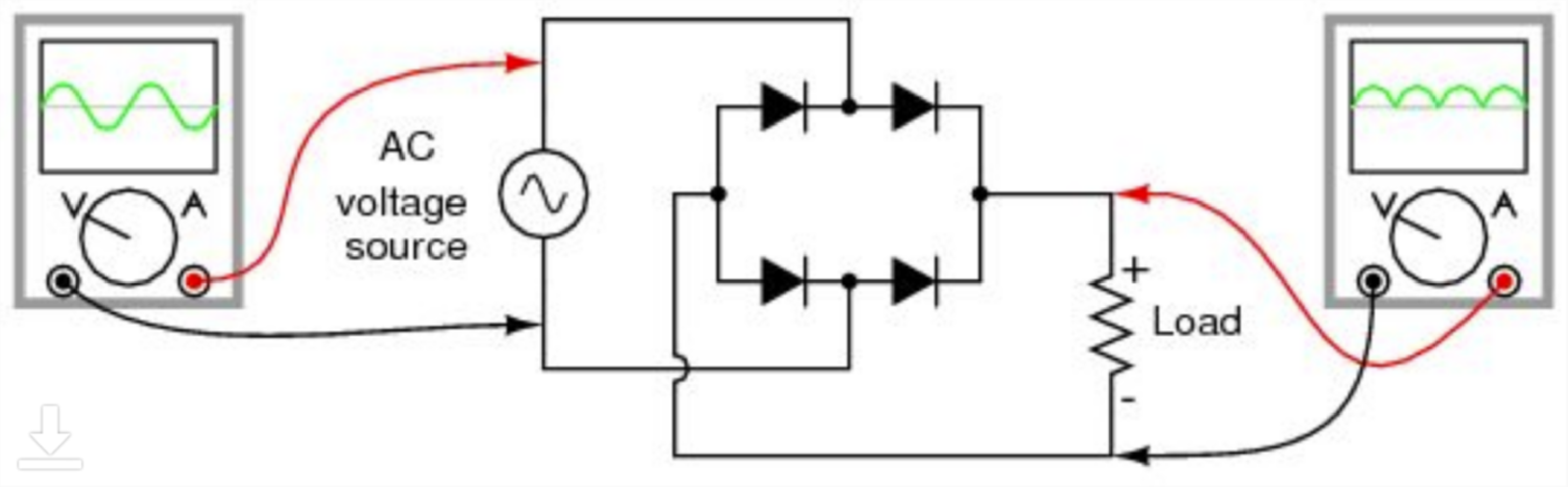
V
A
AC voltage source
+
Load
-
V
A

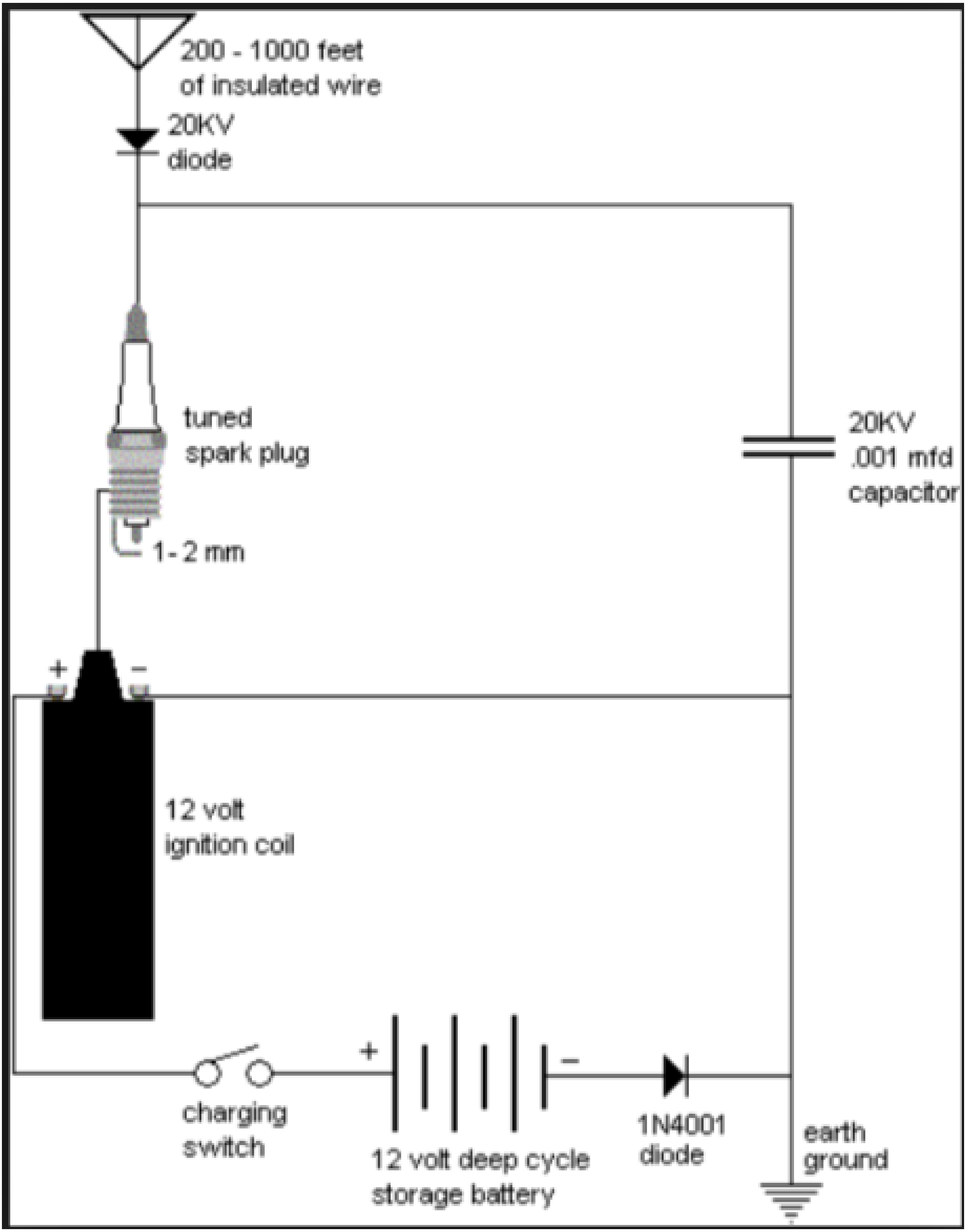
200 - 1000 feet
of insulated wire
20KV
diode
tuned
spark plug
1-2 mm
20KV
.001 mfd
capacitor
+
–
12 volt
ignition coil
charging
switch
+
–
12 volt deep cycle
storage battery
1N4001
diode
earth
ground

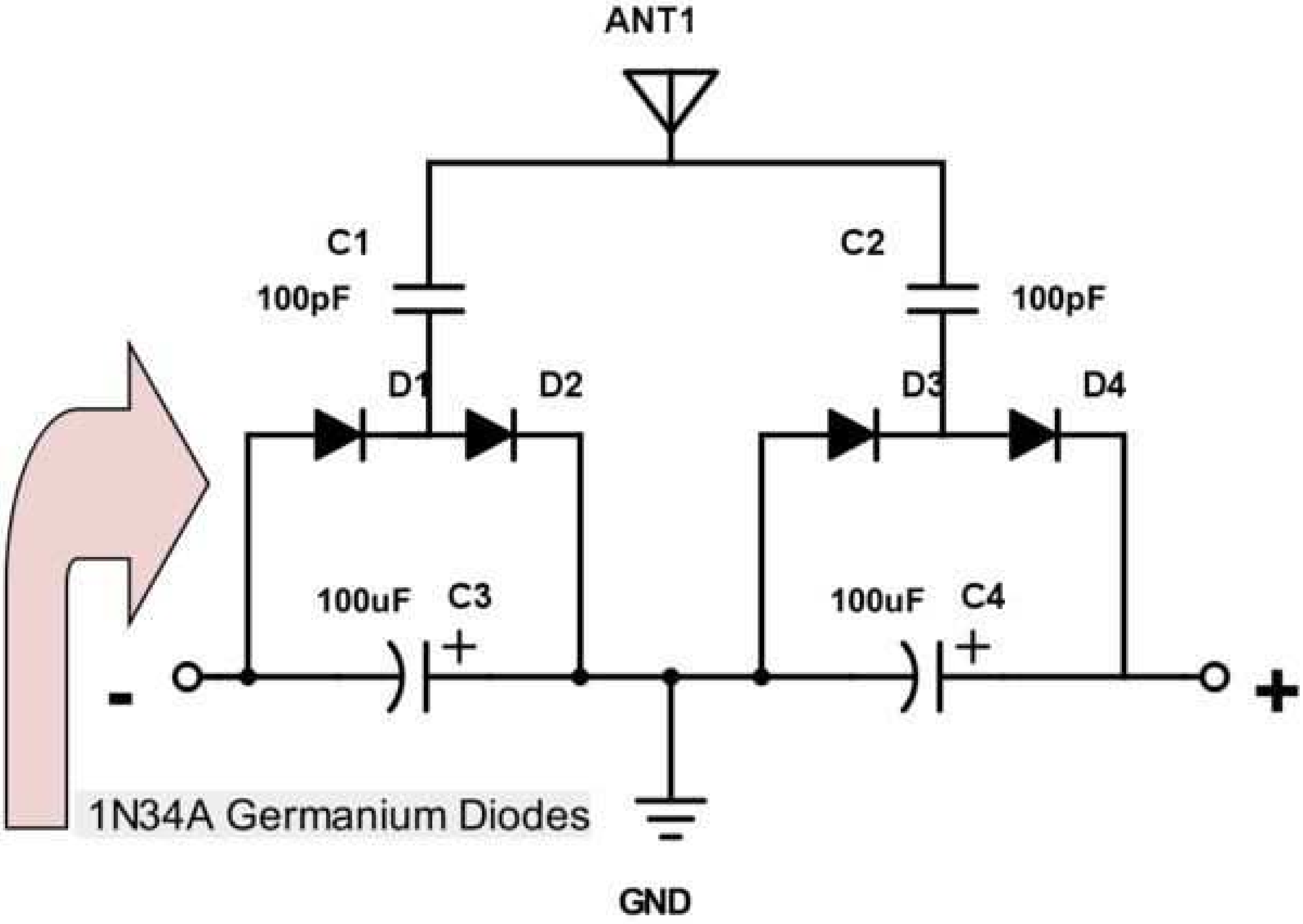
ANT1
C1
100pF
C2
100pF
D1
D2
D3
D4
100uF
C3
100uF
C4
-
+
1N34A Germanium Diodes
GND

DIY HDTV antenna

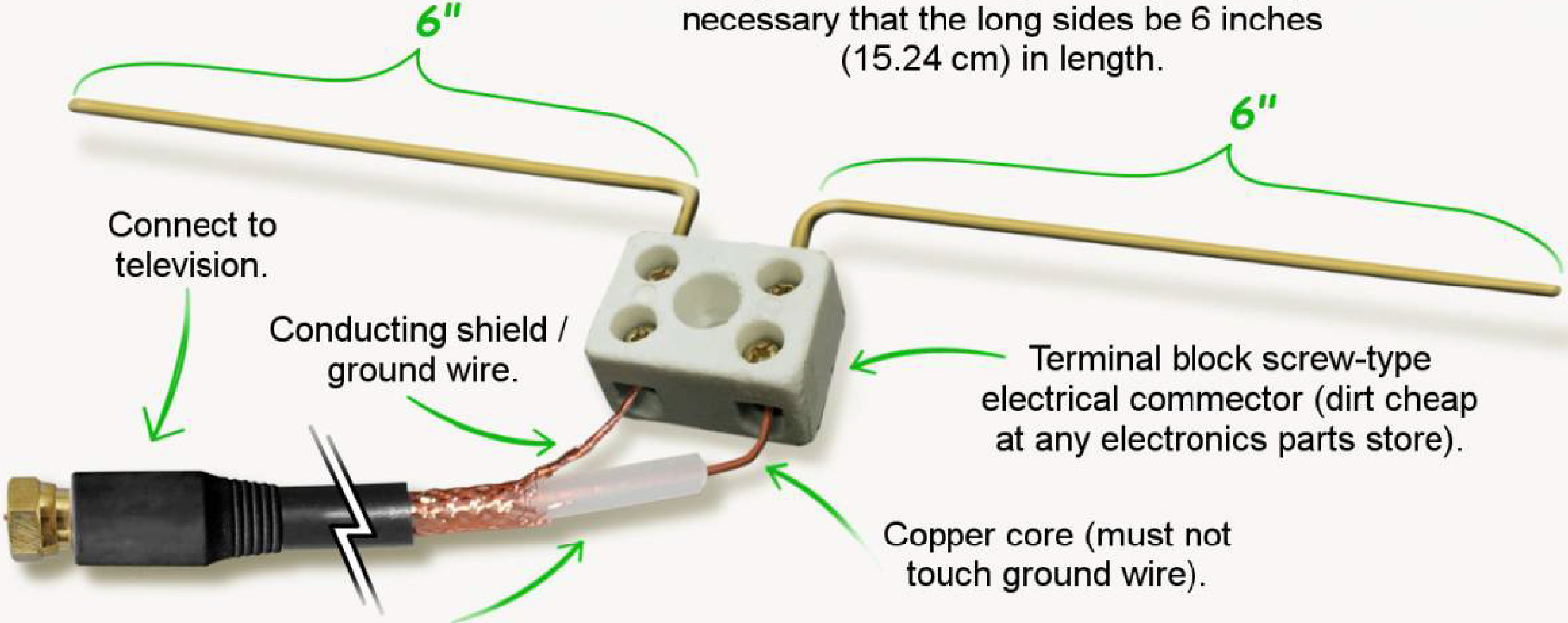

"Antenna" is two L-shaped pieces of coat hanger stripped of any coating. It is only necessary that the long sides be 6 inches (15.24 cm) in length.

Connect to television.

Conducting shield / ground wire.

Terminal block screw-type electrical commector (dirt cheap at any electronics parts store).

Copper core (must not touch ground wire).

Coaxial cable of any length, carefuly stripped to expose the two copper elements.

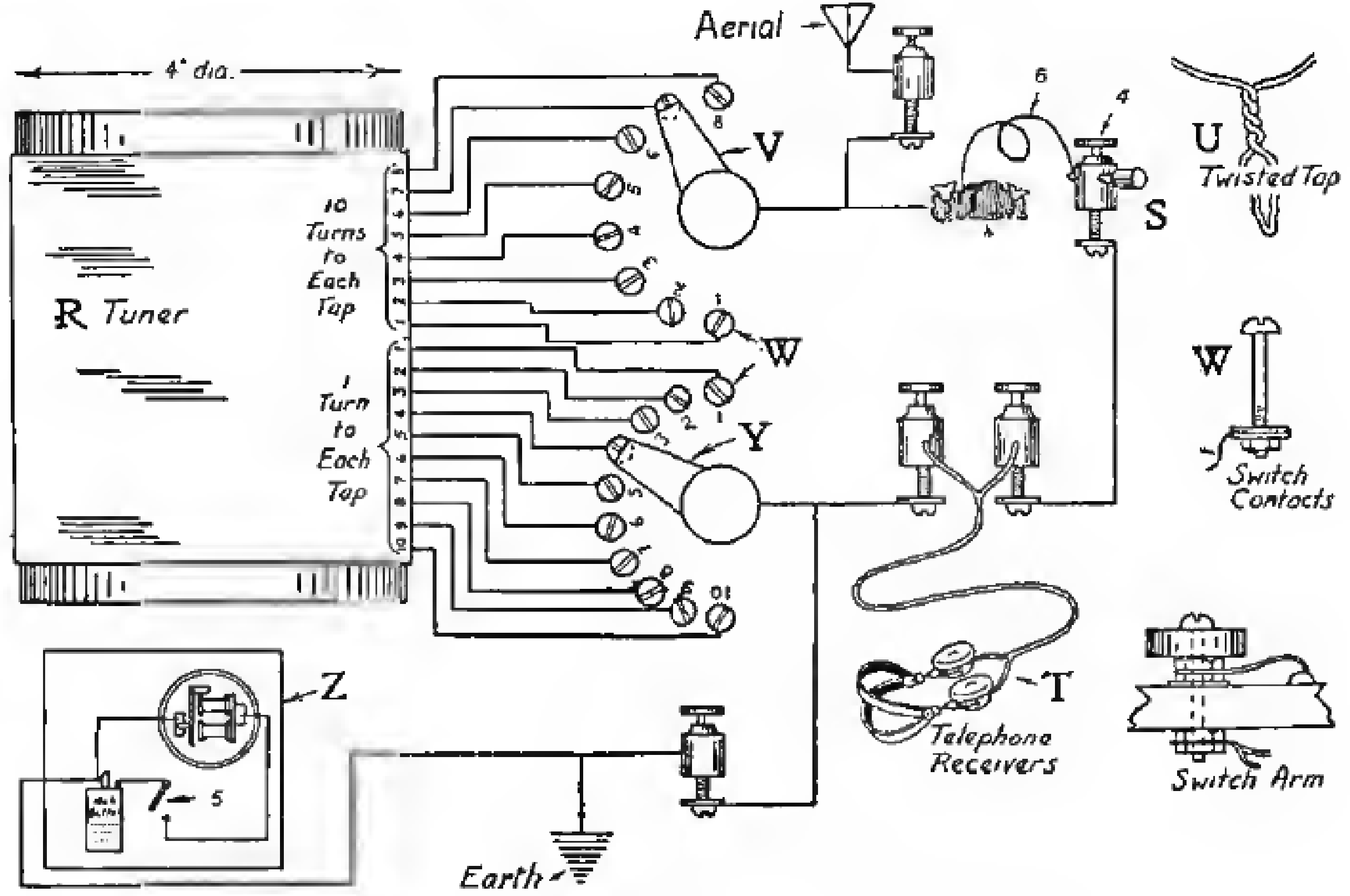
4" dia.
R Tuner
10 Turns to Each Tap
1 Turn to Each Tap
V
W
Y
Aerial
6
4
S
U
Twisted Tap
W
Switch Contacts
T
Telephone Receivers
Switch Arm
Z
5
Earth

Fig. 3.—The Receiving Set in Complete Detail.

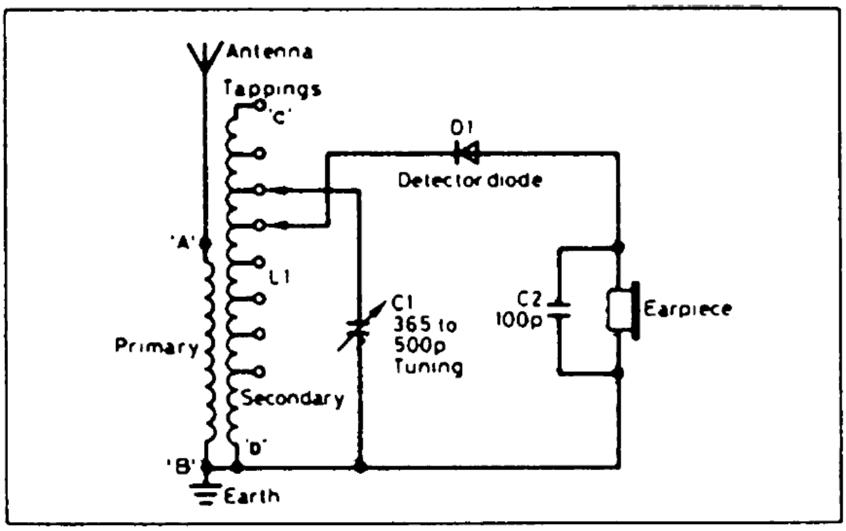
Antenna
Tappings
'C'
D1
Detector diode
'A'
L1
C1
365 to
500p
Tuning
C2
100p
Earpiece
Primary
Secondary
'D'
'B'
Earth

Figure 1 Circuit diagram of the crystal set

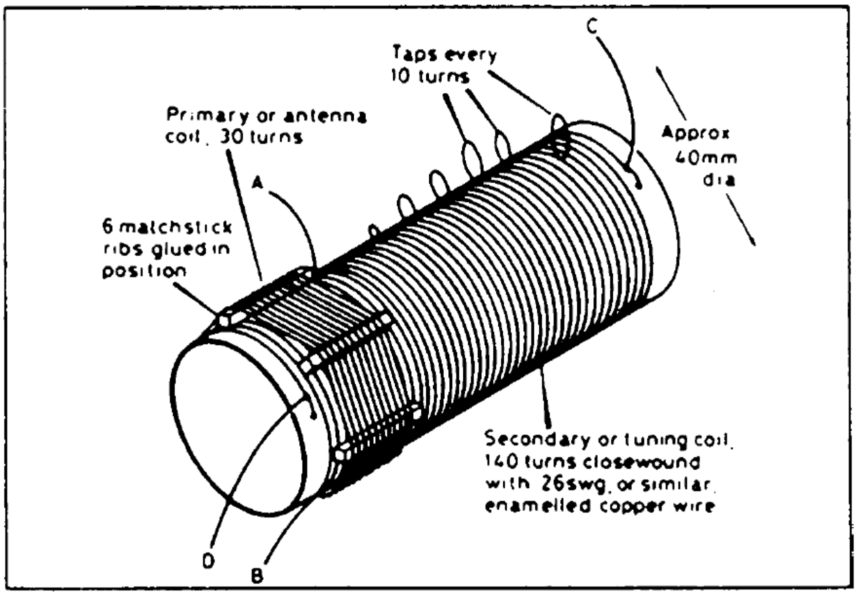
C
Taps every
10 turns
Primary or antenna
coil, 30 turns
A
Approx
40mm
dia
6 matchstick
ribs glued in
position
Secondary or tuning coil,
140 turns closewound
with 26swg, or similar,
enamelled copper wire
D
B

Figure 2 The coil is mounted on a cardboard tube as shown. Exact size is not too important. Try different taps for best results

- This part is the most difficult of all. Seasoned coil winders may have their

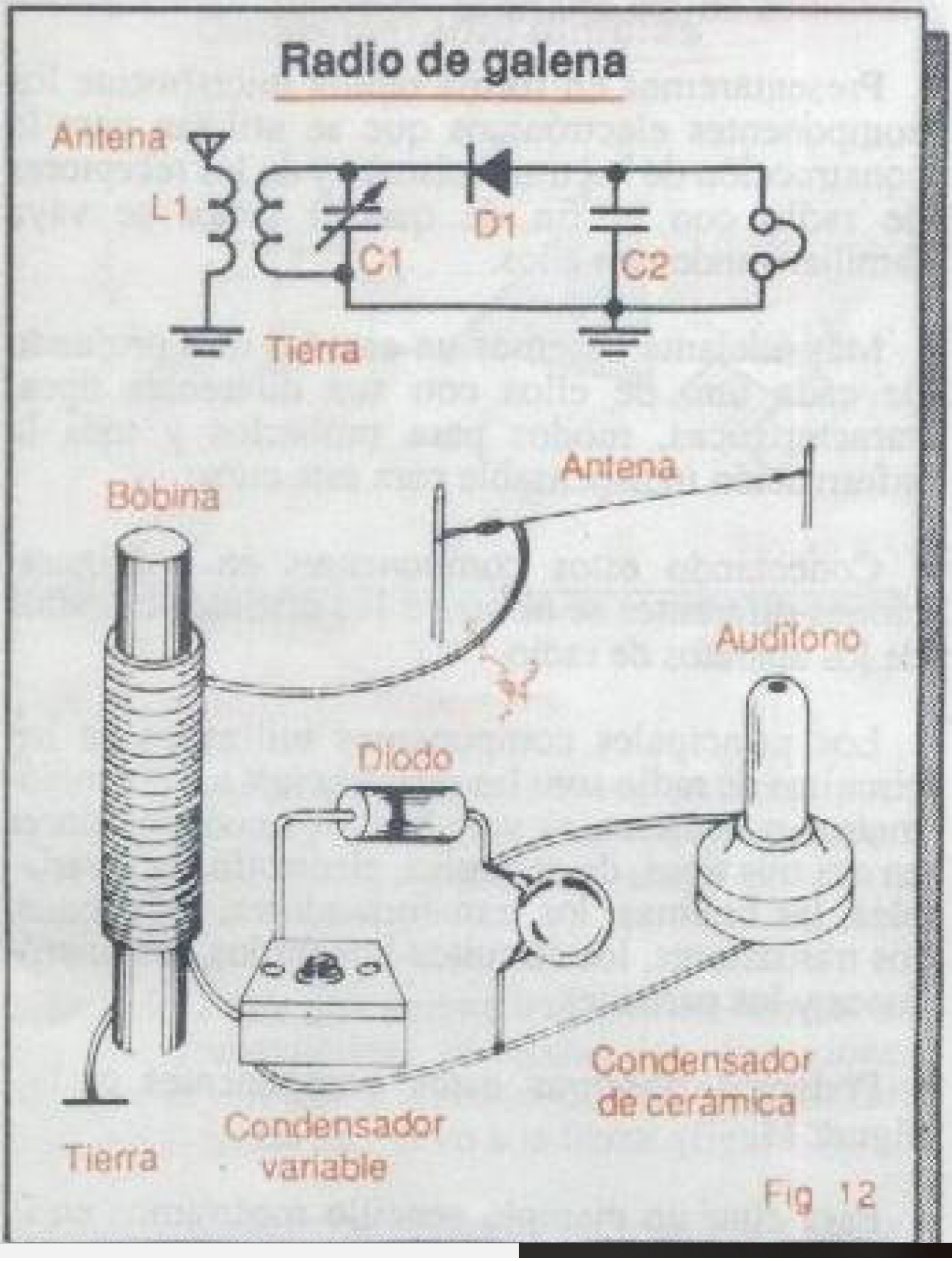
Radio de galena
Antena
L1
C1
D1
C2
Tierra
Antena
Bobina
Audífono
Diodo
Condensador de cerámica
Tierra
Condensador variable

Fig 12

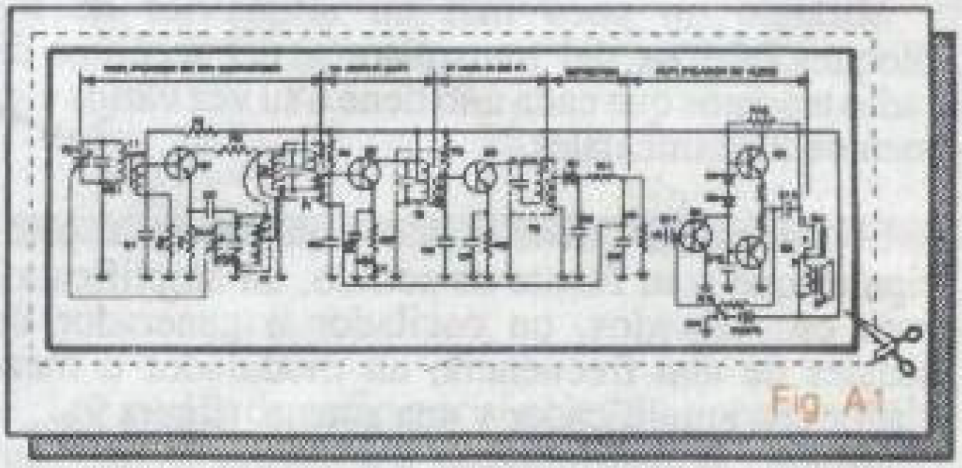

Fig. A1

Paso 2. Pegar la lámina sobre el tablero

Tome su tablero de madera y pegue el esquema sobre él. Unte el pegante en poca cantidad solamente en los bordes y un poco en el centro en forma de cruz. (figura A2).

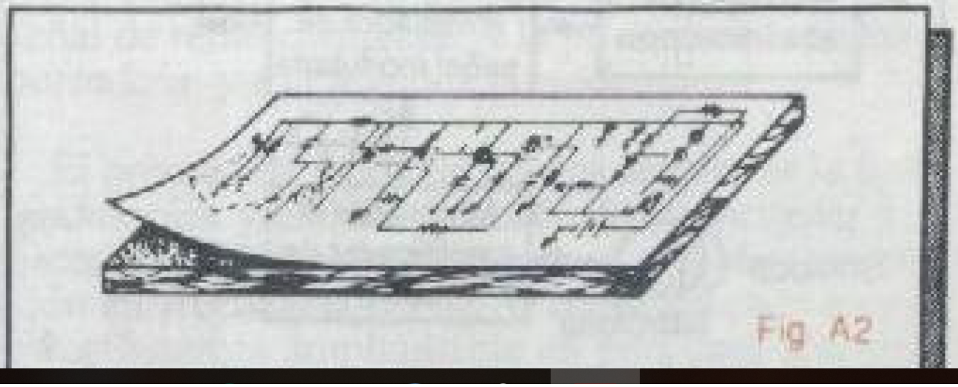

Fig. A2

THE LOST HAMEL DRAWINGS:

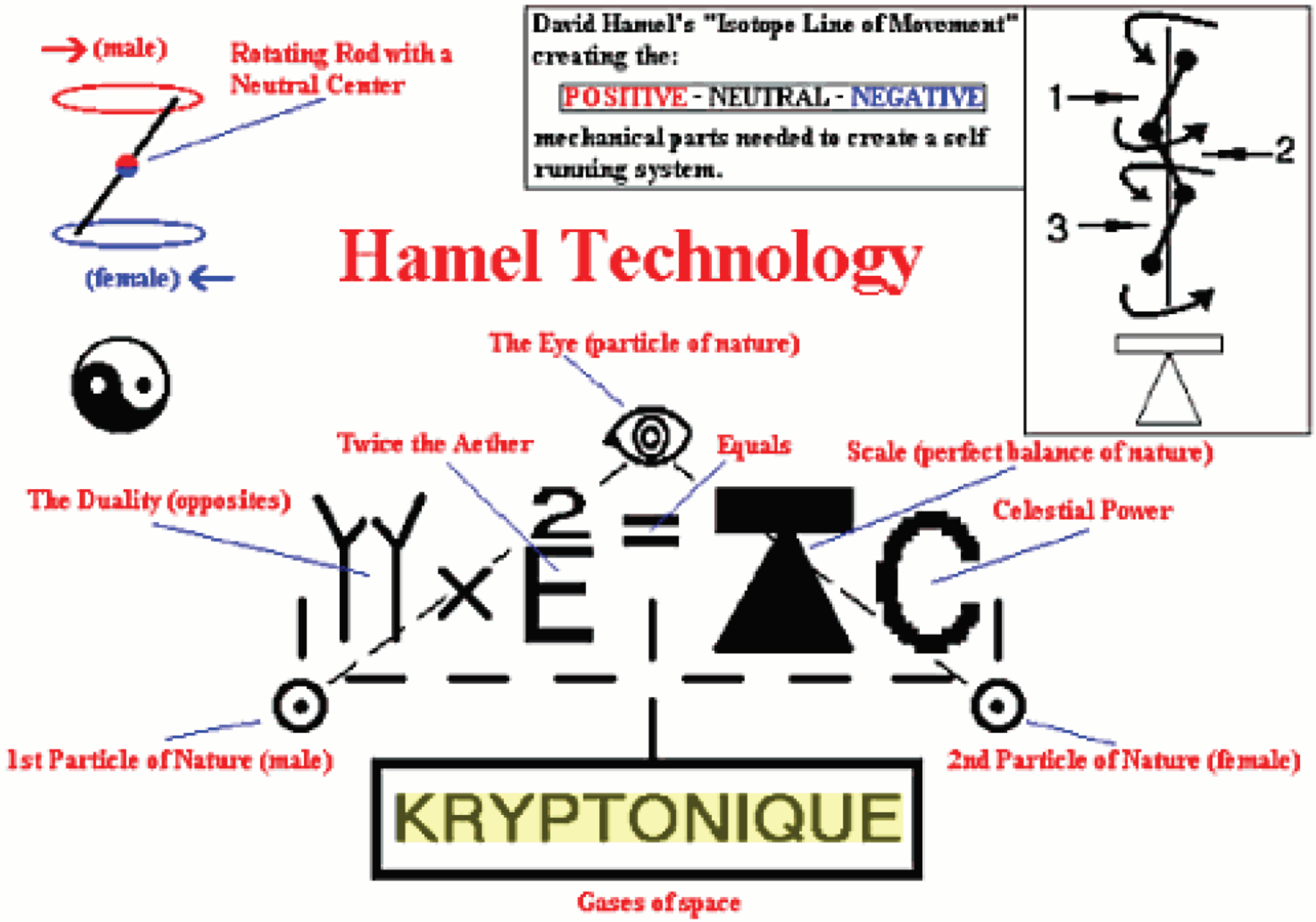

Hamel Technology

David Hamel's "Isotope Line of Movement" creating the:

POSITIVE - NEUTRAL - NEGATIVE

mechanical parts needed to create a self running system.

(male)
Rotating Rod with a Neutral Center
(female)
1
2
3
The Eye (particle of nature)
Twice the Aether
Equals
Scale (perfect balance of nature)
The Duality (opposites)
Celestial Power
1st Particle of Nature (male)
2nd Particle of Nature (female)
KRYPTONIQUE
Gases of space

Particle Structure. [The MOST IMPORTANT Hamel Diagram of All.]

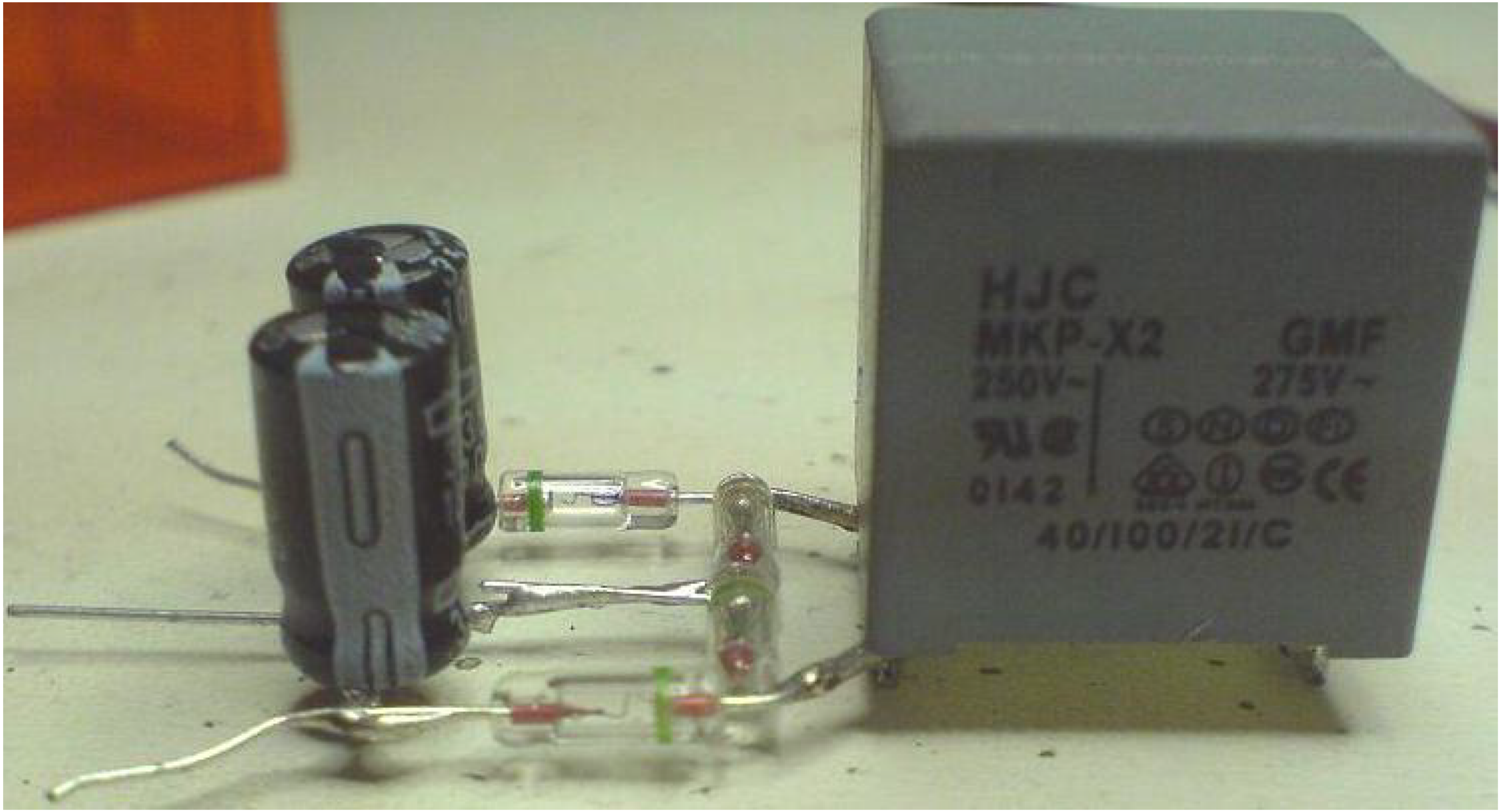
HJC
MKP-X2
GMF
250V~
275V~
0142
40/100/21/C

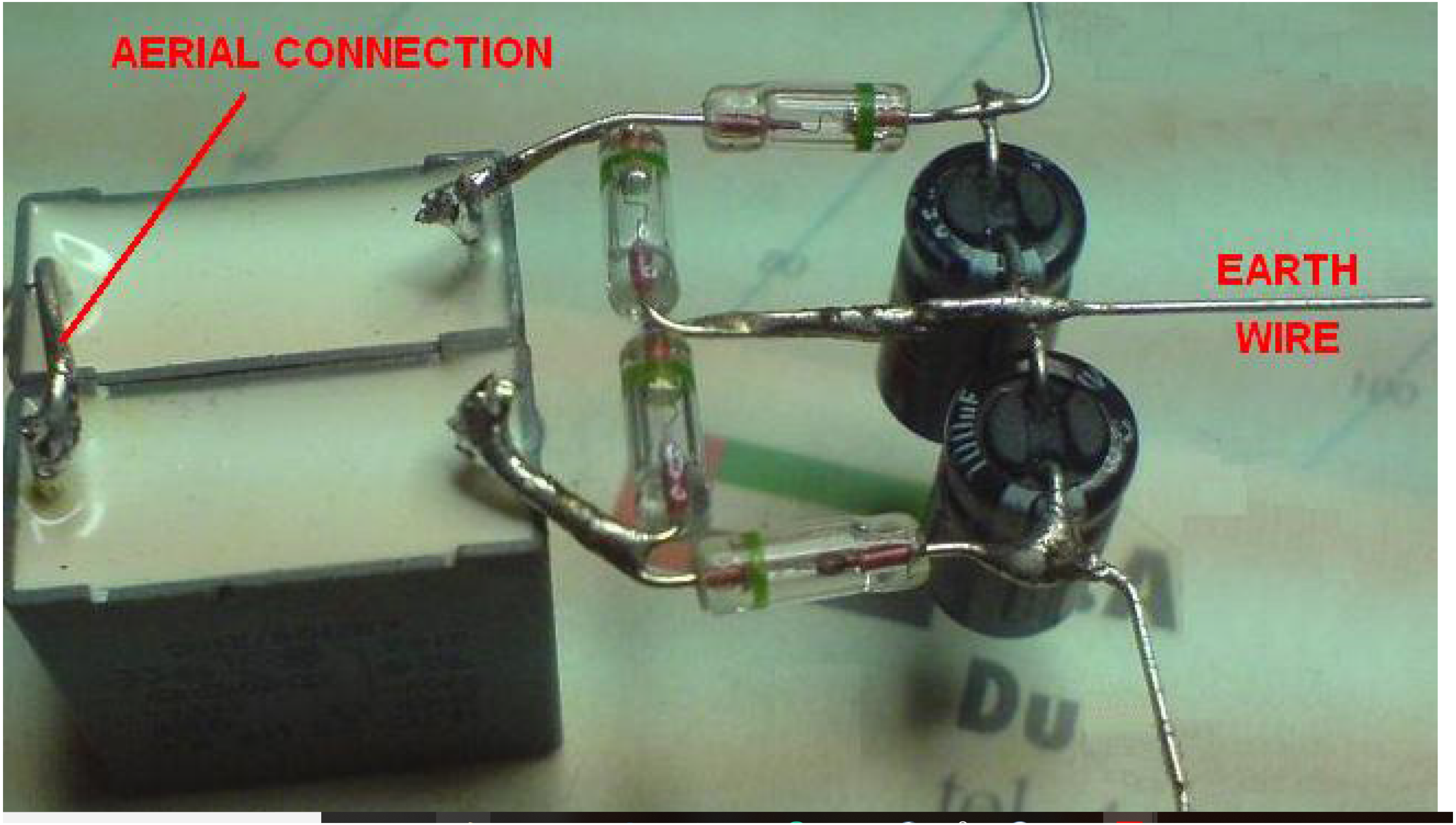
AERIAL CONNECTION
EARTH WIRE

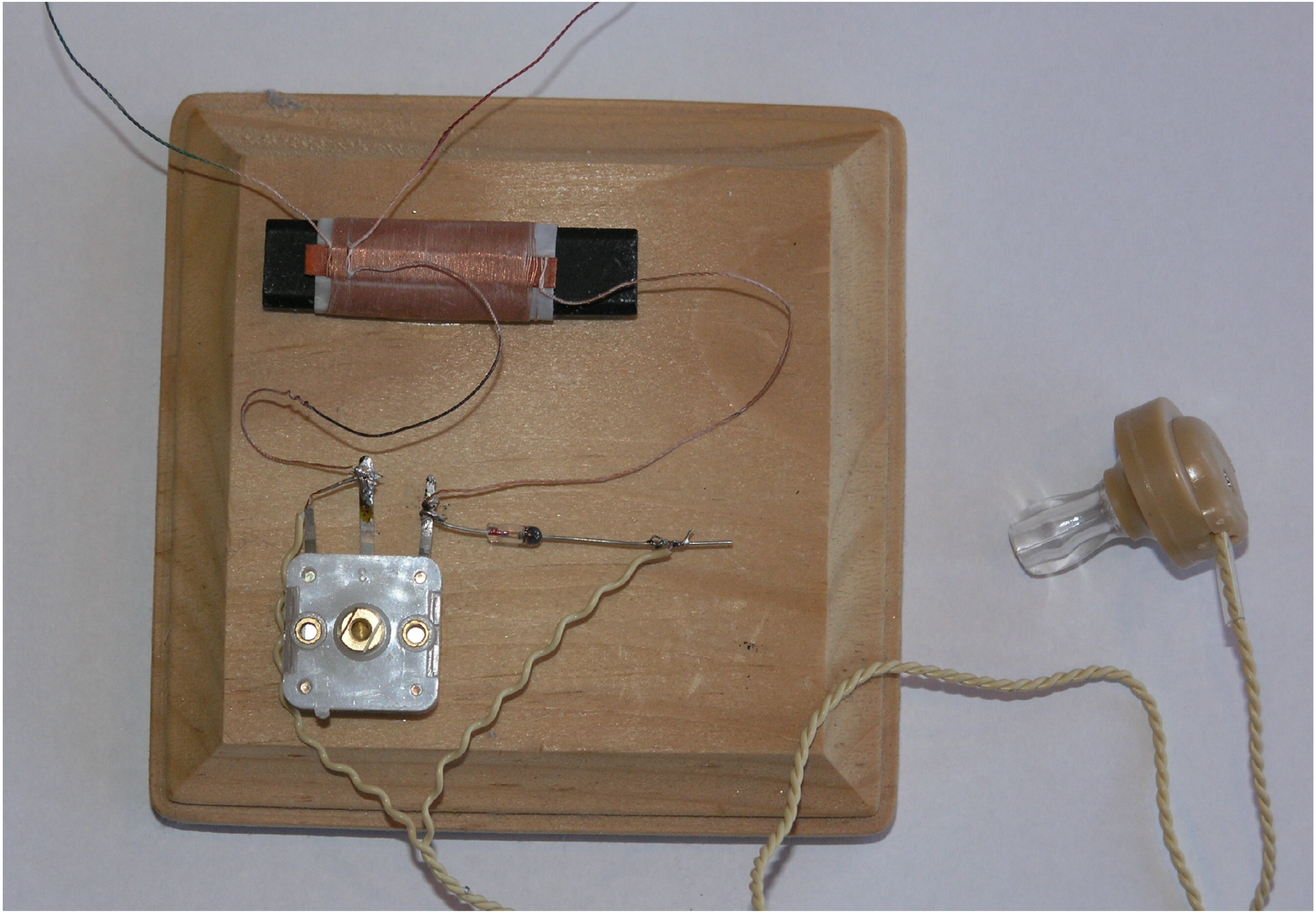

Figure 1 Three constructions for magnetic field probes

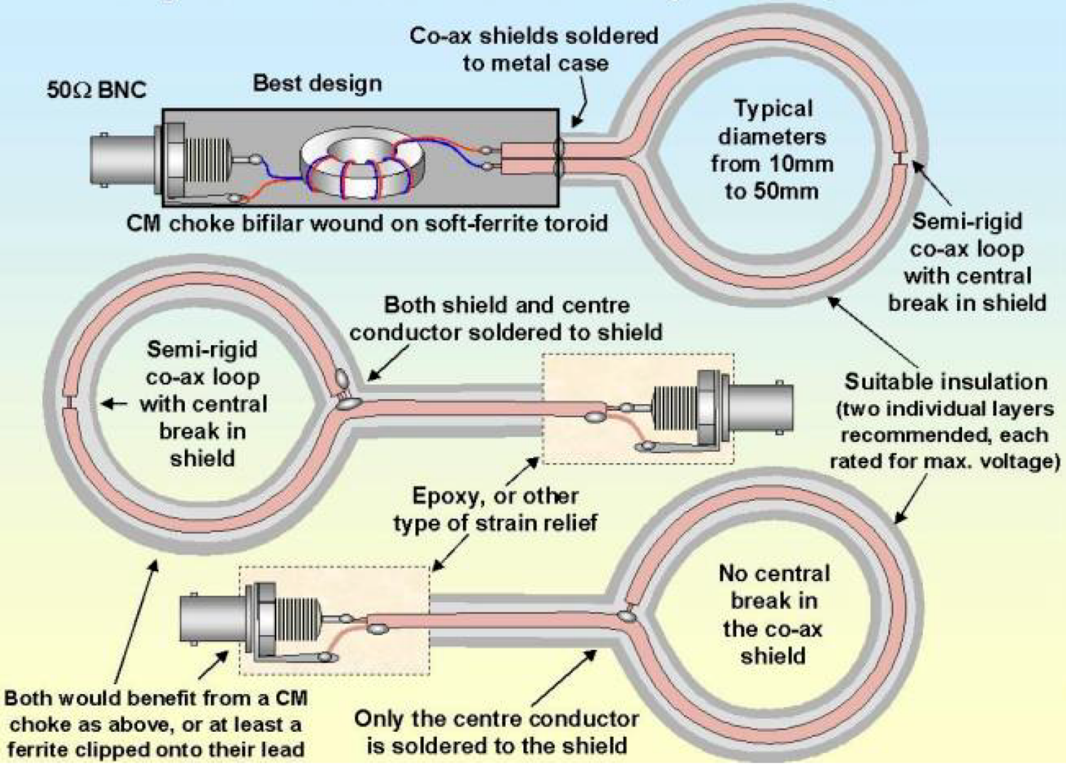
Co-ax shields soldered to metal case
50Ω BNC
Best design
Typical diameters from 10mm to 50mm
CM choke bifilar wound on soft-ferrite toroid
Semi-rigid co-ax loop with central break in shield
Both shield and centre conductor soldered to shield
Semi-rigid co-ax loop with central break in shield
Suitable insulation (two individual layers recommended, each rated for max. voltage)
Epoxy, or other type of strain relief
No central break in the co-ax shield
Only the centre conductor is soldered to the shield

Both would benefit from a CM choke as above, or at least a ferrite clipped onto their lead

Figure 2 An electric field probe, and a pin probe

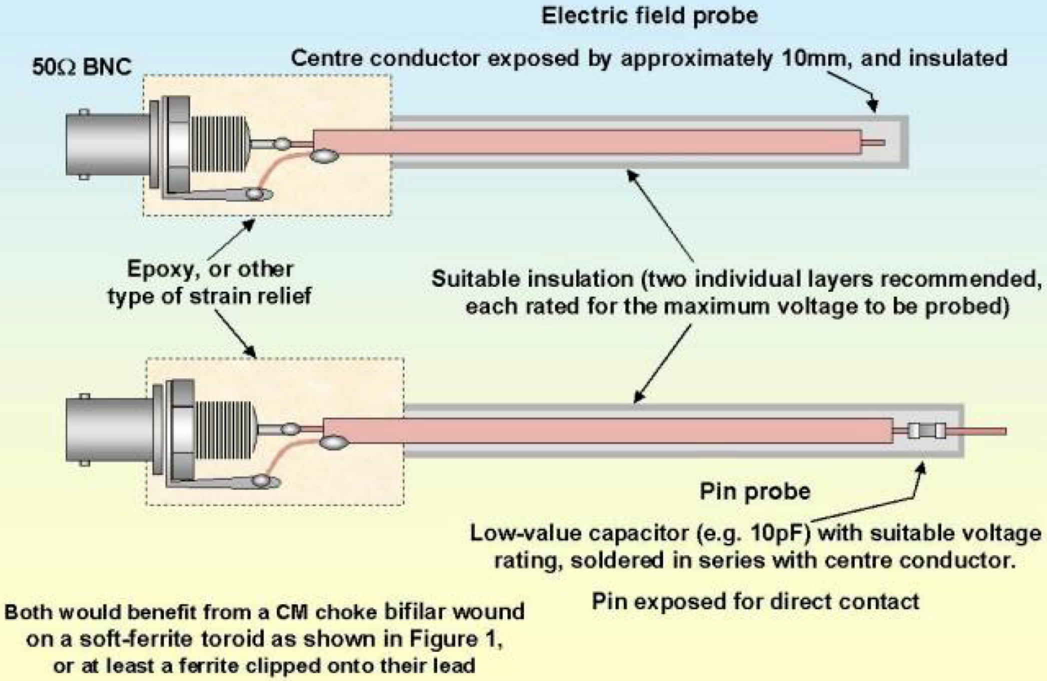
Electric field probe
50Ω BNC
Centre conductor exposed by approximately 10mm, and insulated
Epoxy, or other type of strain relief
Suitable insulation (two individual layers recommended, each rated for the maximum voltage to be probed)
Pin probe
Low-value capacitor (e.g. 10pF) with suitable voltage rating, soldered in series with centre conductor.
Pin exposed for direct contact

Both would benefit from a CM choke bifilar wound on a soft-ferrite toroid as shown in Figure 1, or at least a ferrite clipped onto their lead

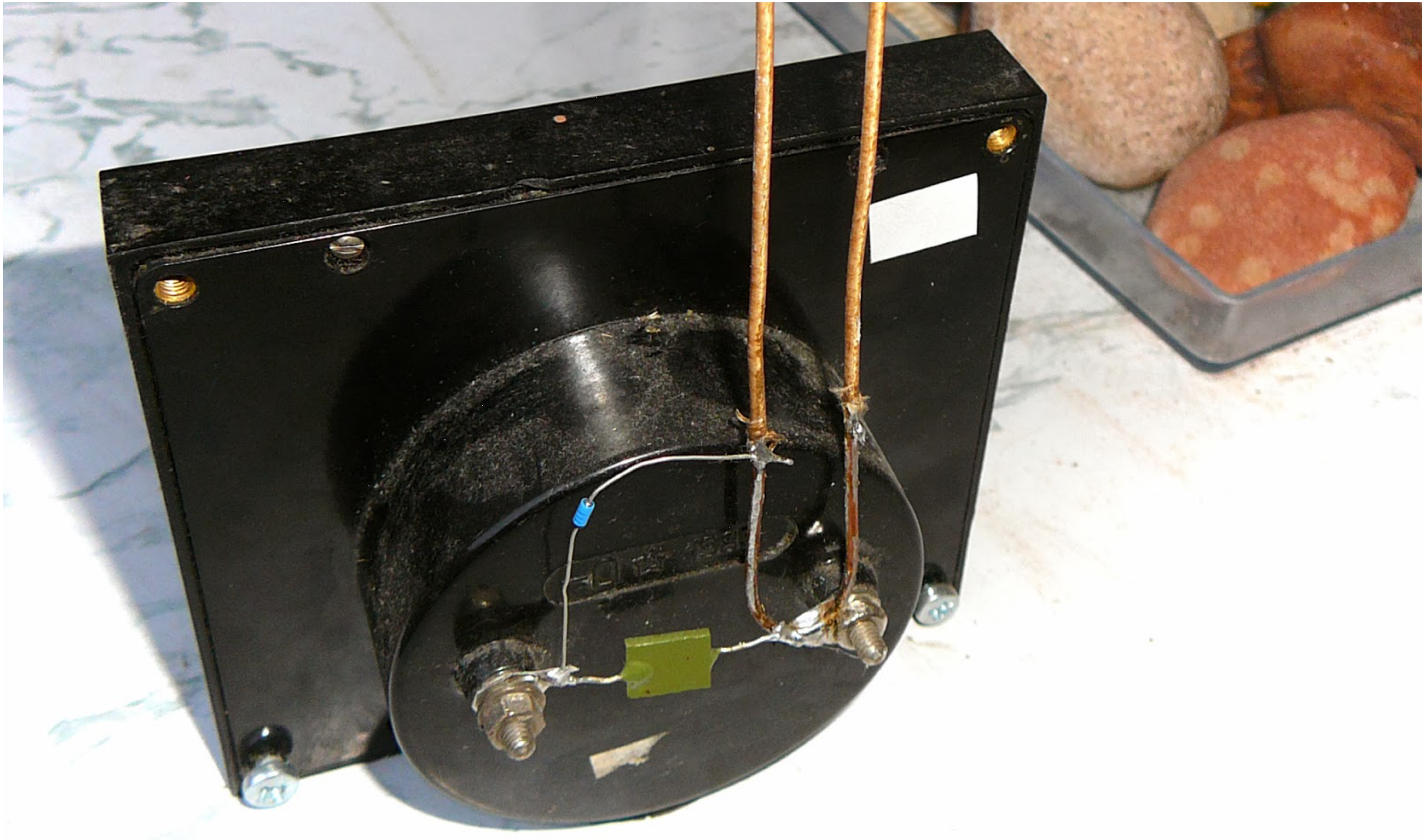

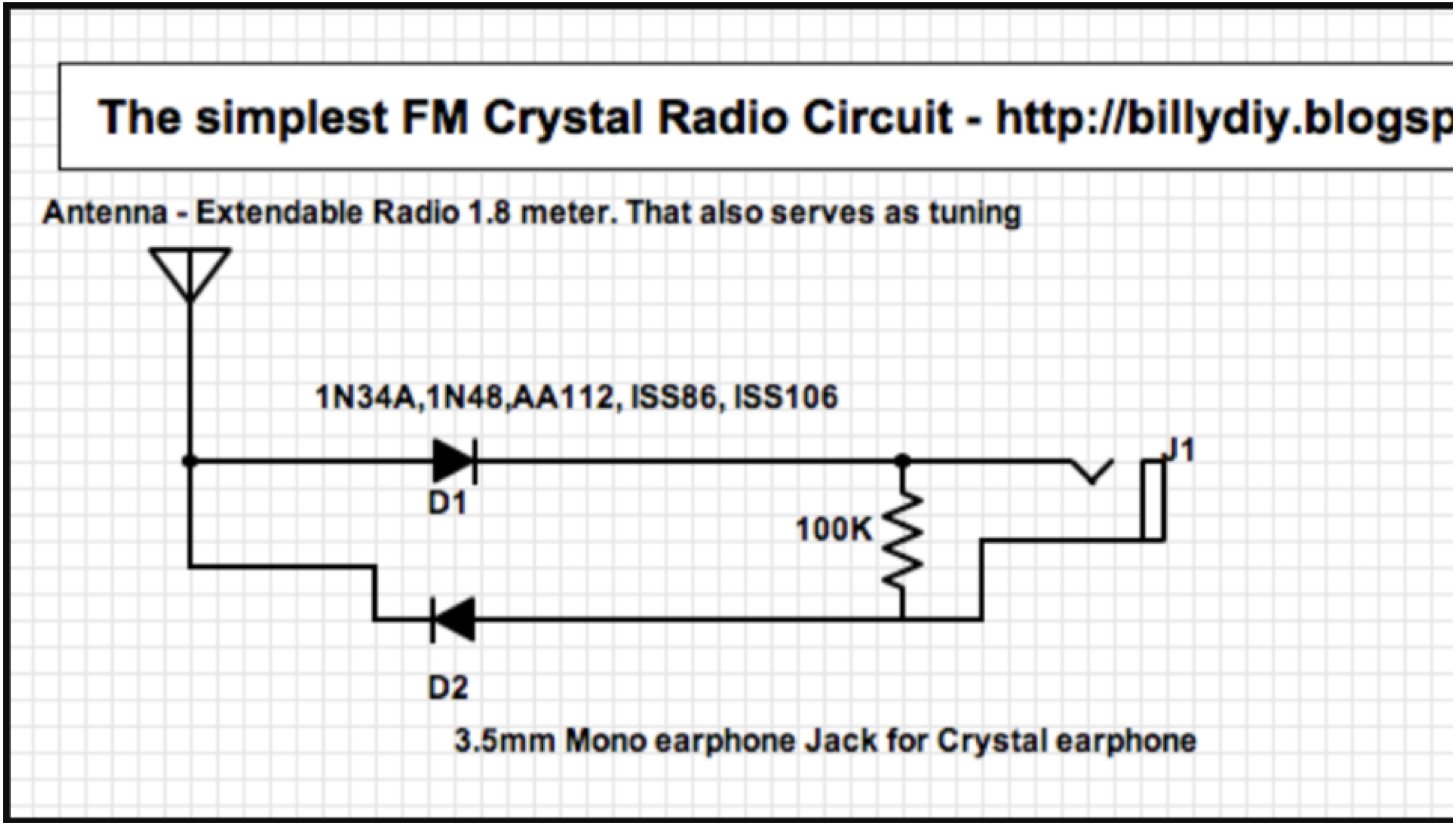

The simplest FM Crystal Radio Circuit - http://billydiy.blogsp

Antenna - Extendable Radio 1.8 meter. That also serves as tuning

3.5mm Mono earphone Jack for Crystal earphone

1N34A,1N48,AA112, ISS86, ISS106
D1
D2
100K
J1

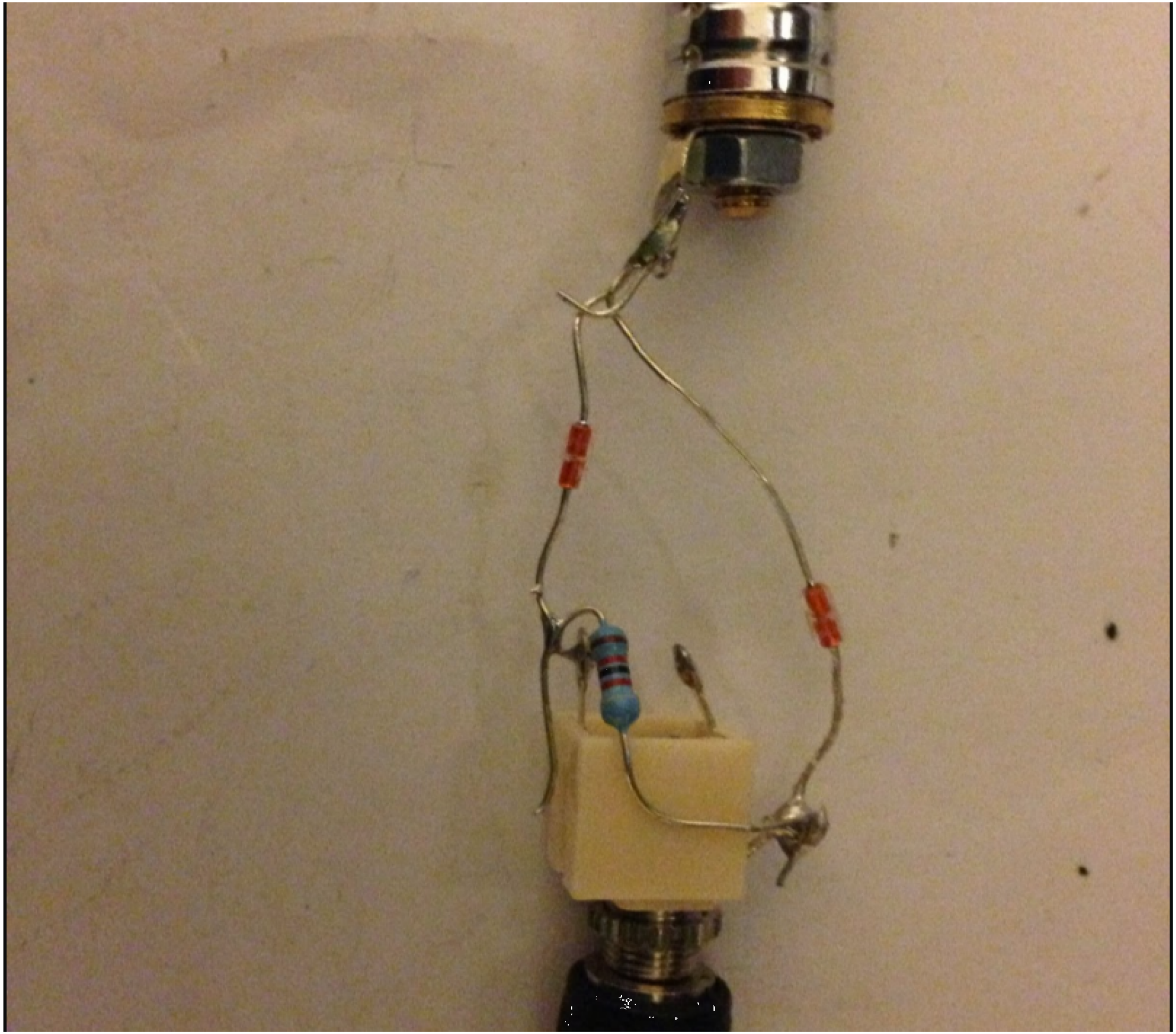

FM Crystal Radio Circuit

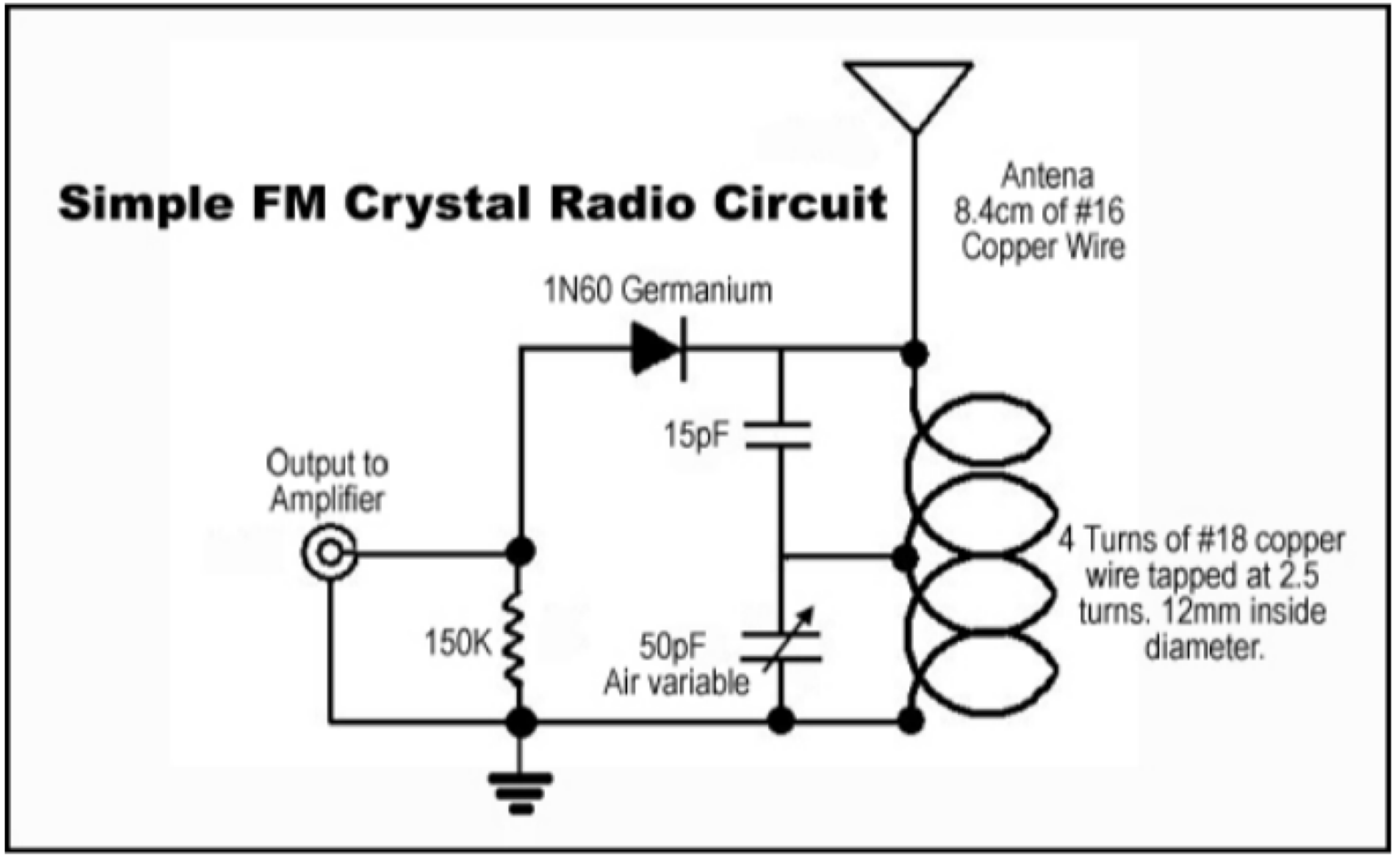
Simple FM Crystal Radio Circuit
Antena
8.4cm of #16
Copper Wire
1N60 Germanium
15pF
Output to
Amplifier
150K
50pF
Air variable
4 Turns of #18 copper
wire tapped at 2.5
turns. 12mm inside
diameter.

Parts List (some of these parts you can buy from our online store):

- 1N60 Germanium Diode
- 15pF Ceramic Capacitor
- 50pF Variable Capacitor
- 150K Ohm Resistor
- #16 & #18 Copper wires

Ricevitore a cristallo per FM

Un progetto di Giacomo Cavuoti

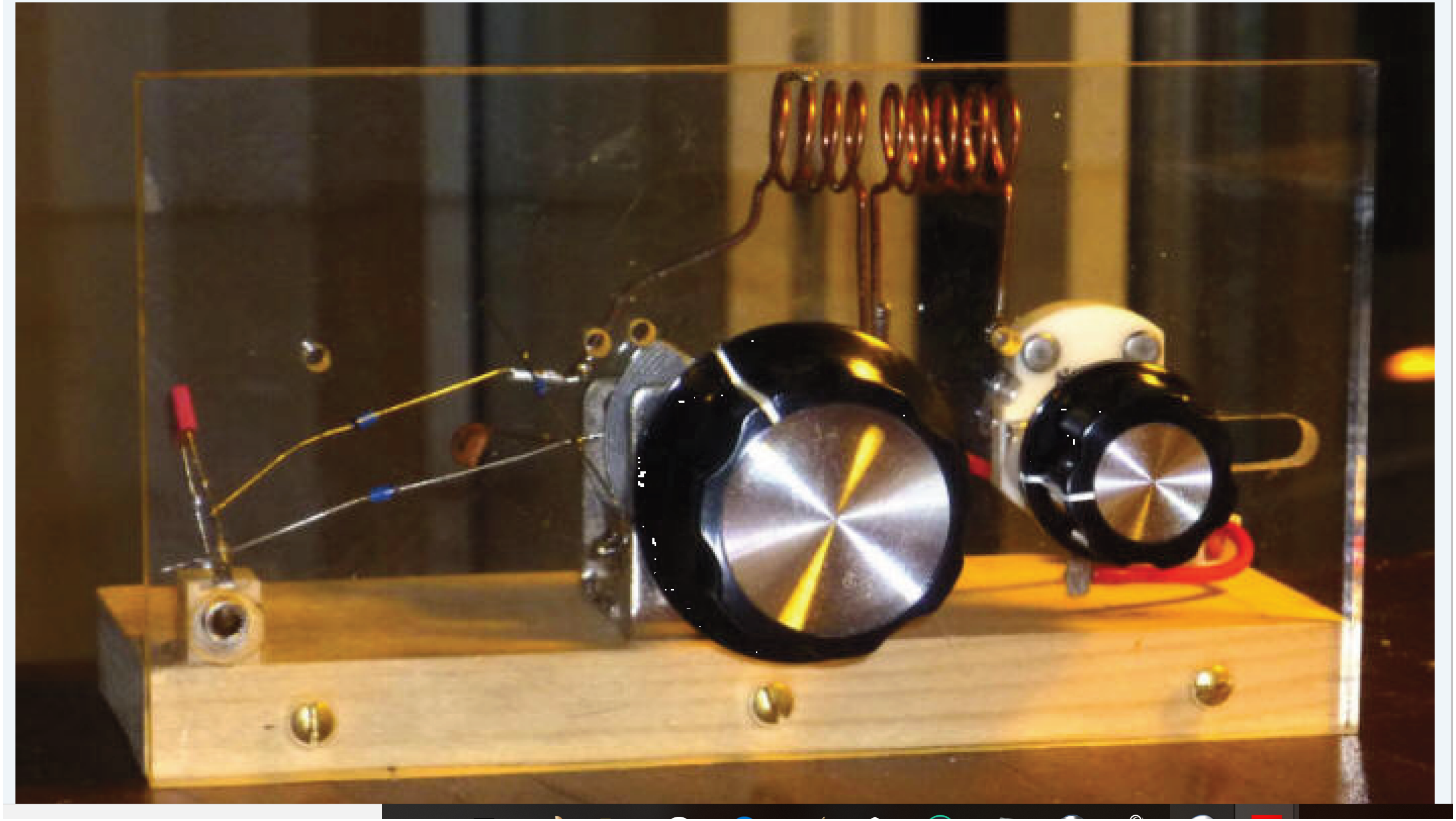

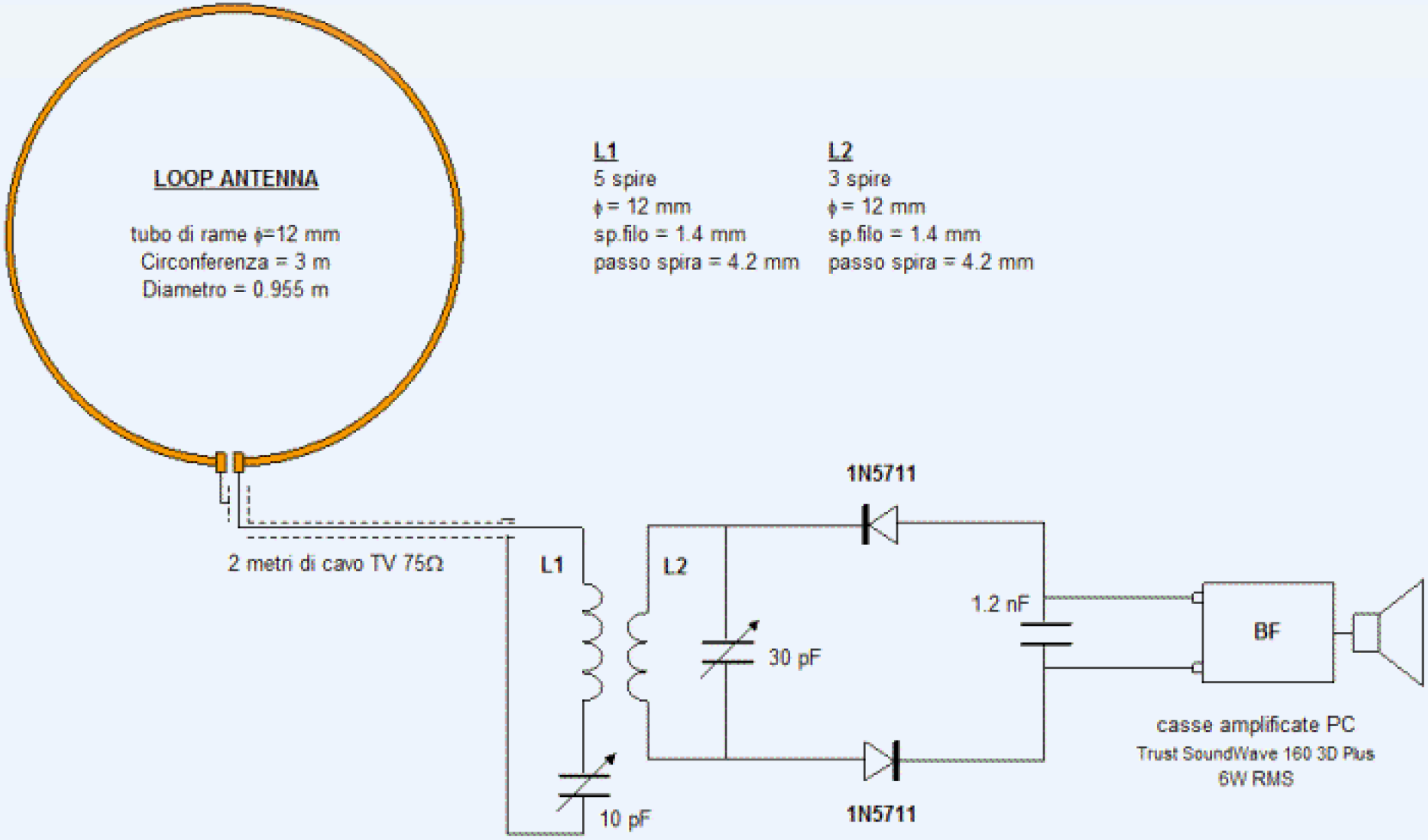
LOOP ANTENNA
tubo di rame ϕ=12 mm
Circonferenza = 3 m
Diametro = 0.955 m
L1
5 spire
ϕ = 12 mm
sp.filo = 1.4 mm
passo spira = 4.2 mm
L2
3 spire
ϕ = 12 mm
sp.filo = 1.4 mm
passo spira = 4.2 mm
1N5711
2 metri di cavo TV 75Ω
L1
L2
1.2 nF
BF
30 pF
casse amplificate PC
Trust SoundWave 160 3D Plus
6W RMS
10 pF
1N5711

The receiver is "powered" by a full wave loop antenna consisting of a single circular coil made of 3 meters of copper pipe for hydraulic systems (measured inductance of about 3 uH). The support is a simple 2-meter long fir-wood strip, 3 cm wide and 1 cm thick, to which the copper rim has been fixed with insulating tape. Another solution to realize the loop antenna in a simple and instantaneous way is to use an aluminum strip 2 mm thick and 2 cm wide. The detectors are two *Schottky* diodes type 1N5711, particularly suitable for VHF thanks to the low capacity (about 2pF). They are currently in production and therefore easy to find. The receiver was tested in a "stiff" situation with the antenna exposed in a recessed balcony (3 x 1.6 meters) on the fourth floor of a 6-storey reinforced concrete building surrounded by other buildings constructed of reinforced concrete walls. As you can see and hear from the two videos I uploaded on Youtube:

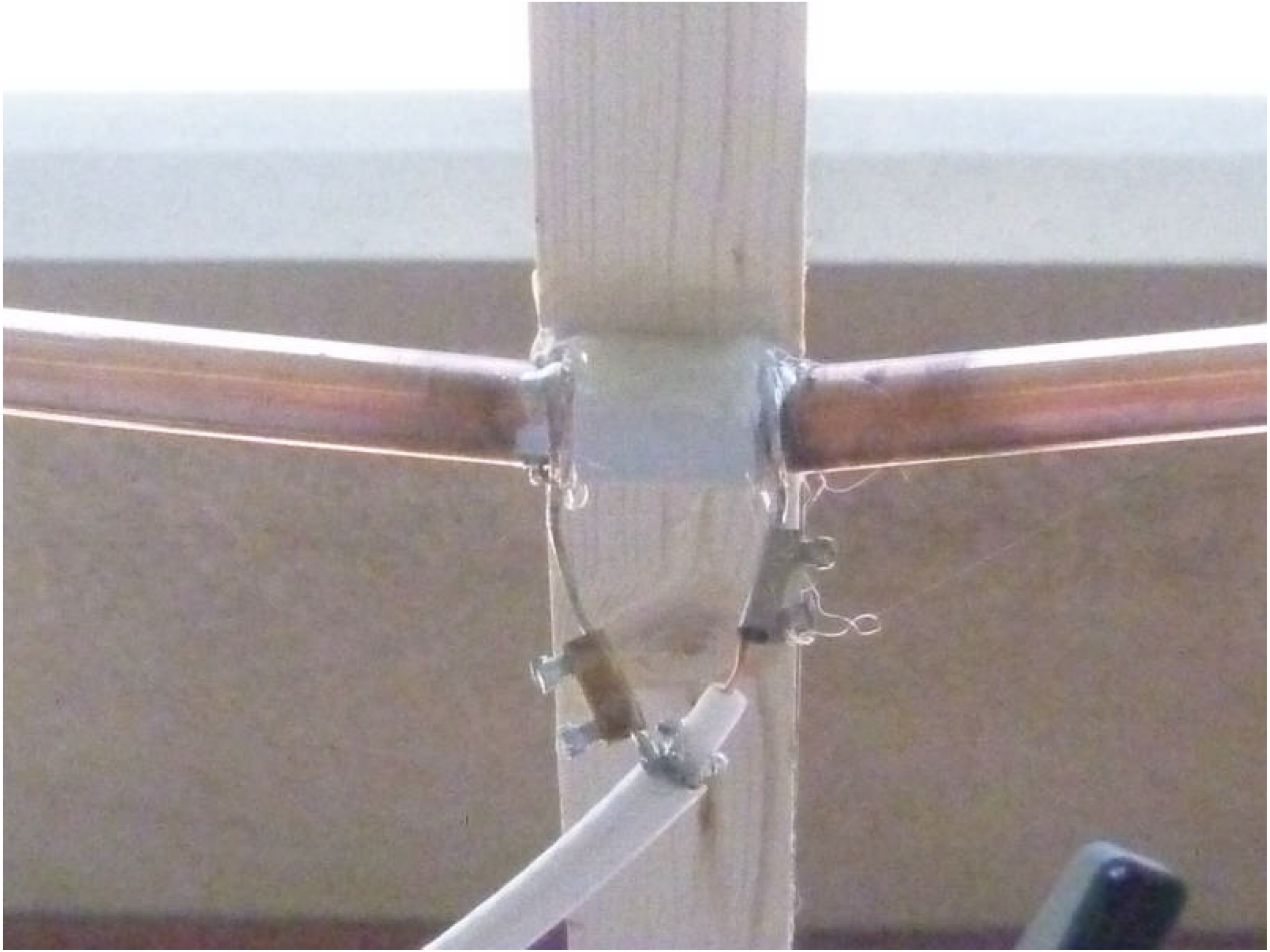

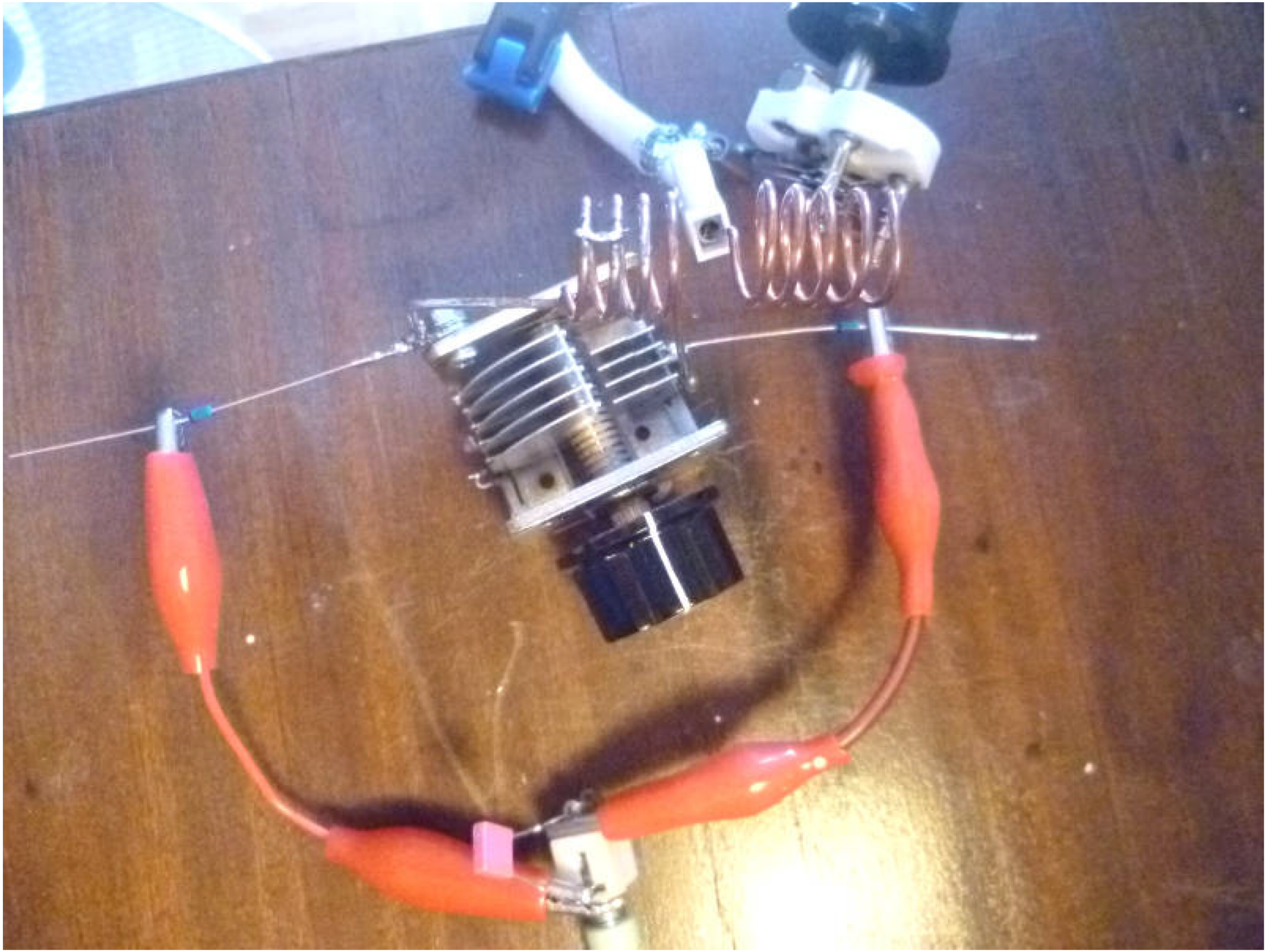

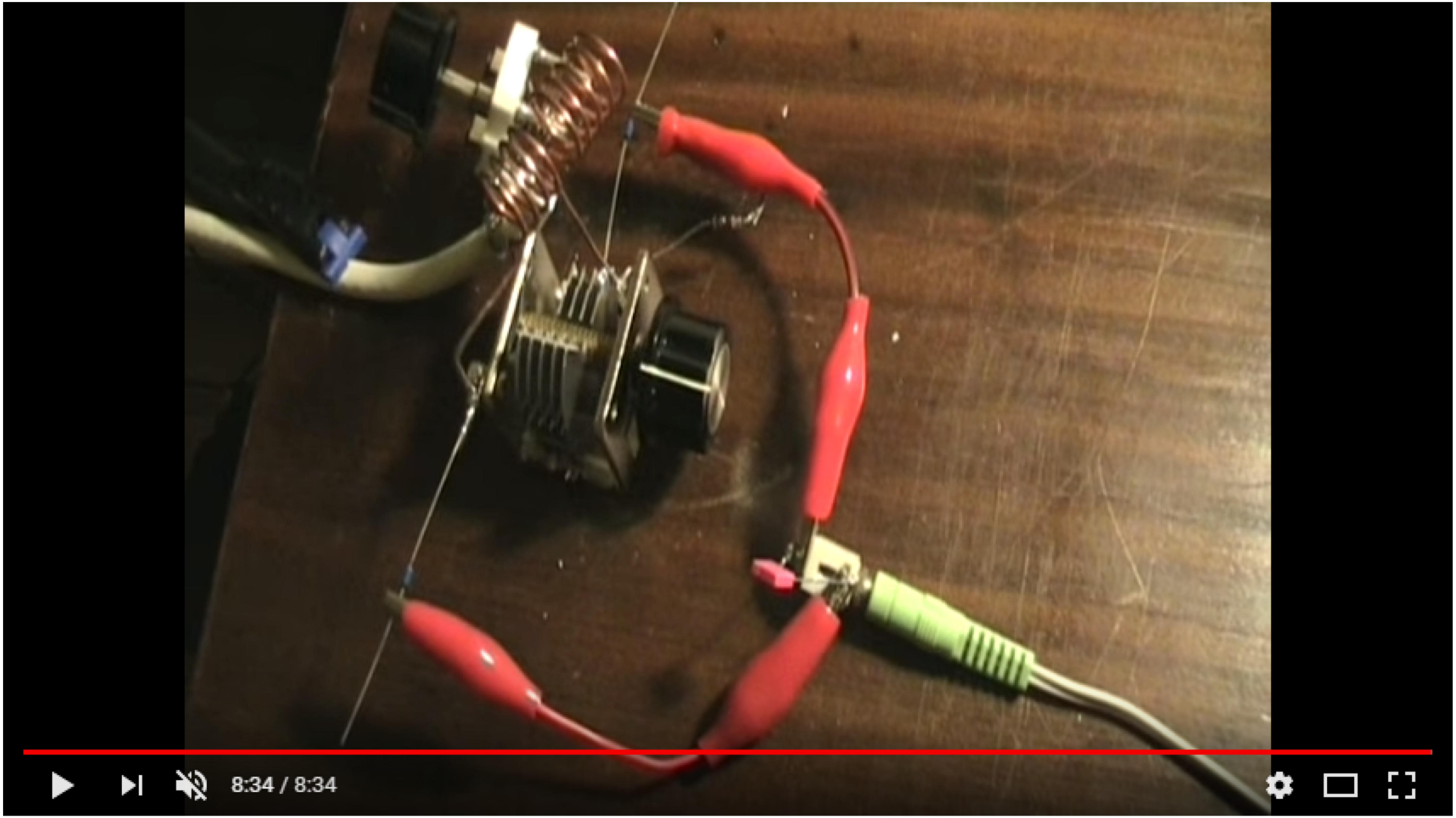
8:34 / 8:34

Making the simplest Transmitter

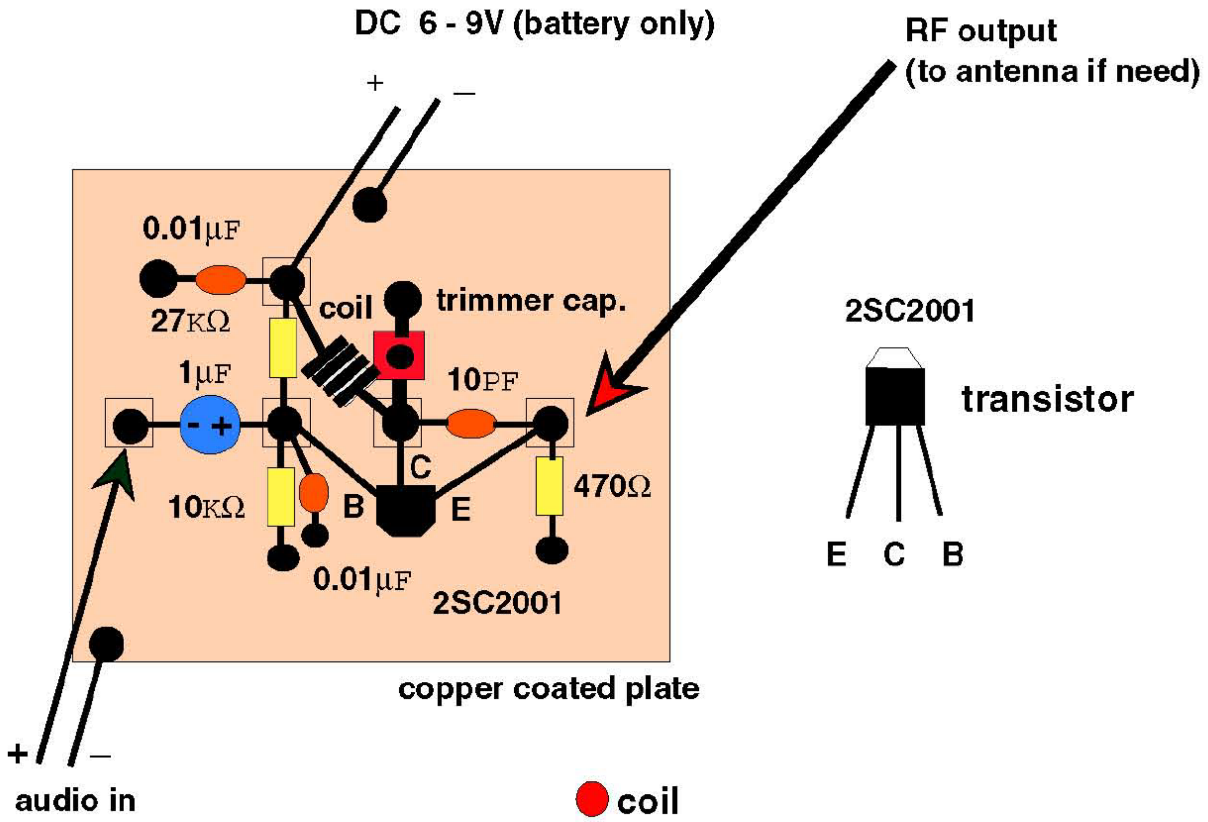
DC 6 - 9V (battery only)
+
–
RF output
(to antenna if need)
0.01μF
27KΩ
coil
trimmer cap.
1μF
10PF
- +
C
B
E
10KΩ
470Ω
0.01μF
2SC2001
copper coated plate
+
–
audio in
2SC2001
transistor
E C B

- **coil**
  3 - 4 turns by coated 0.8 mm wire

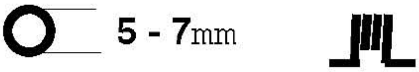

- **registers**
  470 Ω *(yellow-violet-brown)*
  10K Ω *(brown-black-orange)*
  27K Ω *(red-violet-orange)*

- **capacitors**
  10 PF
  0.01 μF *(103)*
  1 μF

- **trimmer capasitor 20PF**

- **soldered point:**
  ● : direct to the ground

☐ : insulated from the ground

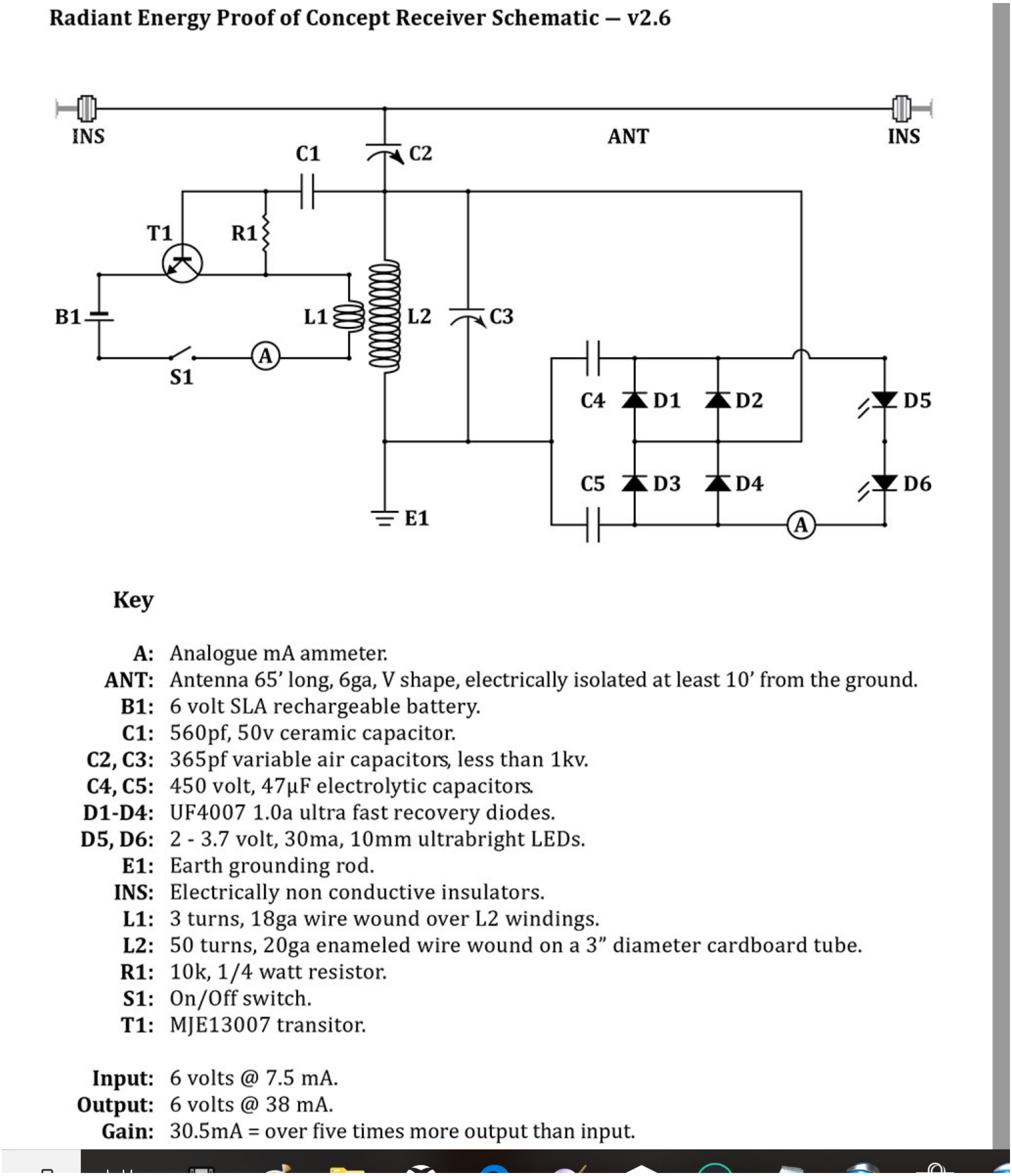

Radiant Energy Proof of Concept Receiver Schematic – v2.6

Key

A:	Analogue mA ammeter.
ANT:	Antenna 65’ long, 6ga, V shape, electrically isolated at least 10’ from the ground.
B1:	6 volt SLA rechargeable battery.
C1:	560pf, 50v ceramic capacitor.
C2, C3:	365pf variable air capacitors, less than 1kv.
C4, C5:	450 volt, 47µF electrolytic capacitors.
D1-D4:	UF4007 1.0a ultra fast recovery diodes.
D5, D6:	2 - 3.7 volt, 30ma, 10mm ultrabright LEDs.
E1:	Earth grounding rod.
INS:	Electrically non conductive insulators.
L1:	3 turns, 18ga wire wound over L2 windings.
L2:	50 turns, 20ga enameled wire wound on a 3” diameter cardboard tube.
R1:	10k, 1/4 watt resistor.
S1:	On/Off switch.
T1:	MJE13007 transitor.

Input:	6 volts @ 7.5 mA.
Output:	6 volts @ 38 mA.
Gain:	30.5mA = over five times more output than input.

turning low-frequency currents into sound, and it would be as

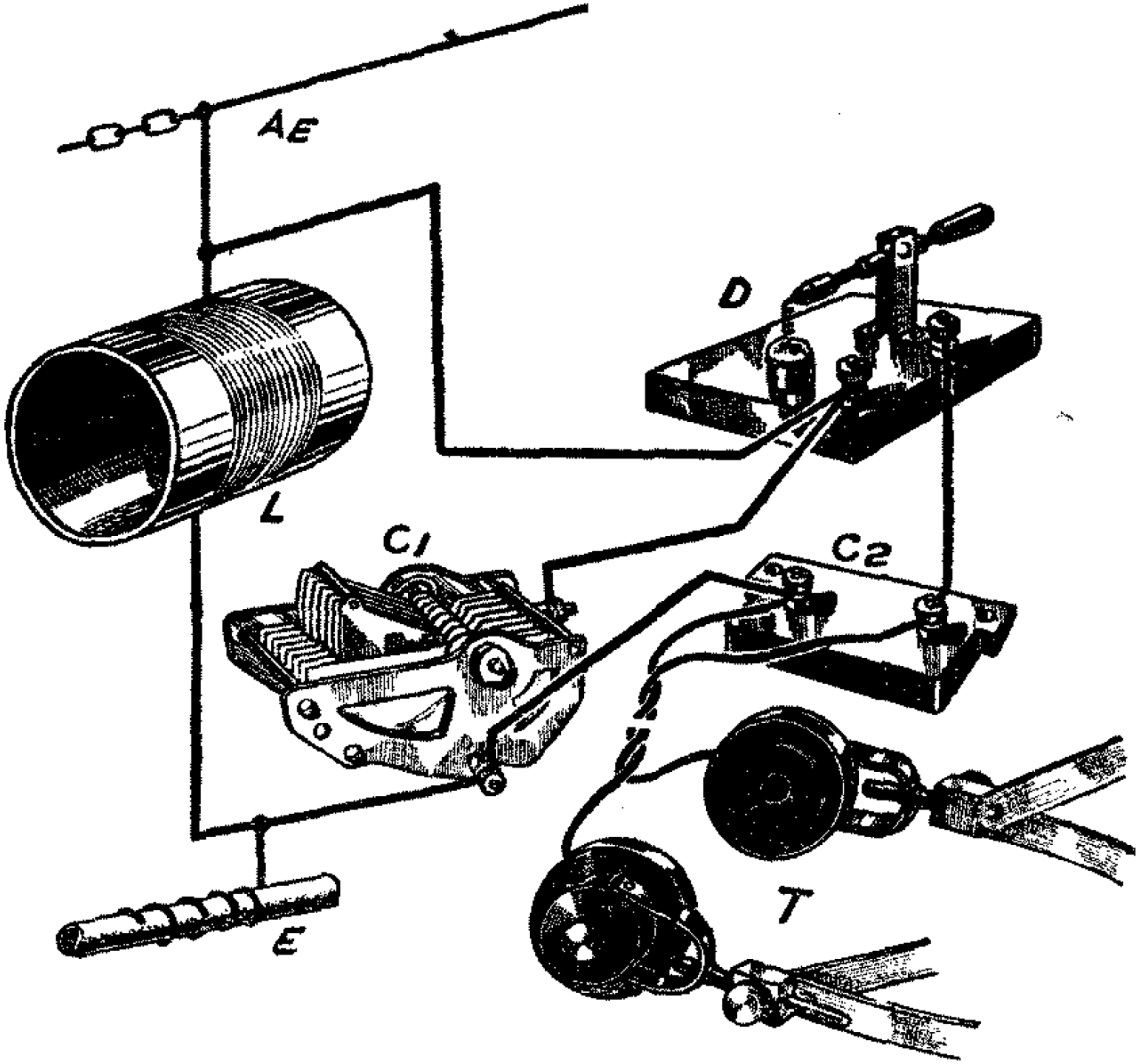

Fig. 16 (b)—Pictorial representation of the simple

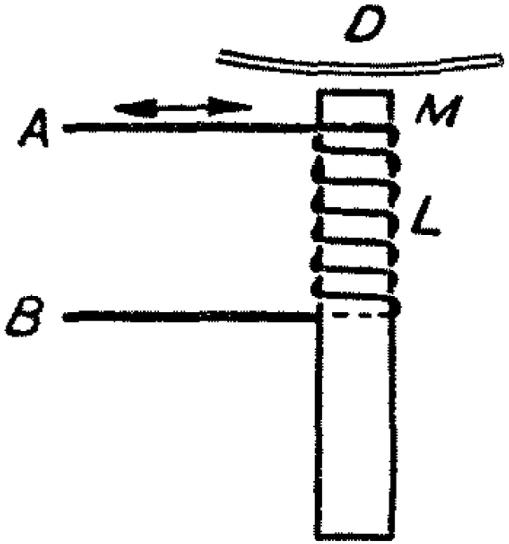

Fig. 17—Currents vary the strength of the magnetism and so cause diaphragm movements

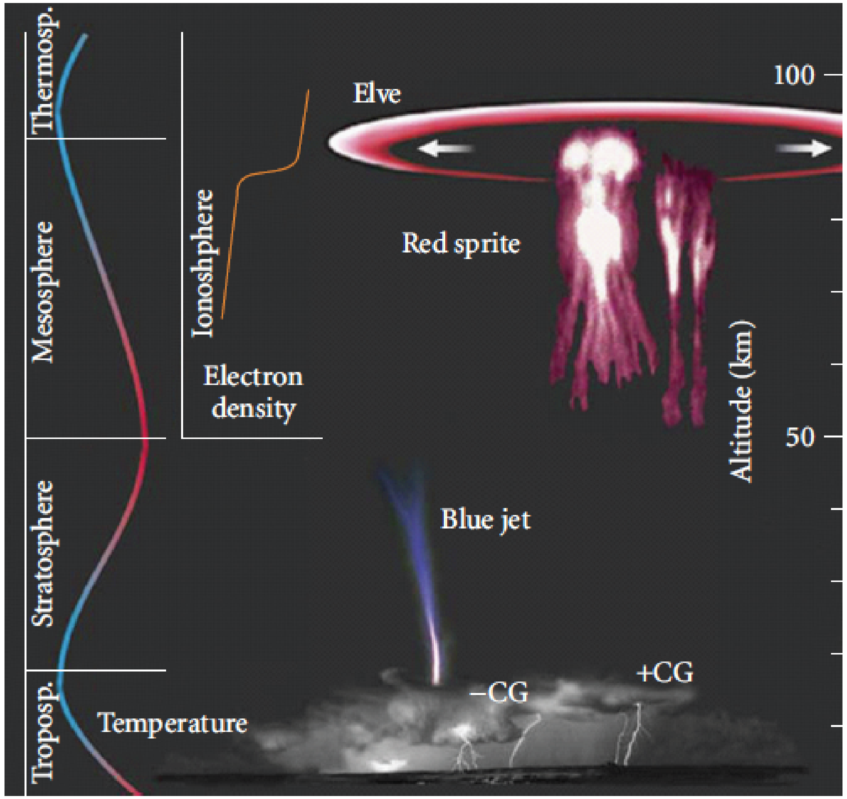

Figure 2: Depiction of various optical events in the atmosphere and the altitude at which they occur, credit [2].

or at time as low as the cloud top and bright region is in the altitude range 65–85 km [35]. They typically last for 5–50 ms and may take the form of one or more vertical columns of a few hundred meter radius for the smaller column sprites or large jellyfish-shaped structures of tens of kilometers of radius and extending from the ionosphere D-region almost down to the thunderstorm cloud tops [2, 35]. The knowledge gained from laboratory experiment of gas discharges at subatmospheric pressure has been used to understand sprite spectroscopy and associated phenomena [25]. Figure 3 shows broad identical spectral characteristics of light of positive column of the laboratory tube and sprites [25]. The difference in spectral characteristics may be due to difference in applied electric field, gas pressure, and gas composition in the mesosphere and gas tube discharges. Physical processes associated with sprites and other optical events are also associated with thunderstorm activity in the troposphere and are thought to result in the gradual build-up of conductivity changes in the lower ionosphere [36]. Liszka [37] suggested the generation of infrasound waves by sprites, whose signatures were detected by a network sensors in Sweden [38]. The shape of the chirp signature in the spectrograms of infrasound can be explained by the horizontal size of the sprite [39]. Neubert et al. [40] have

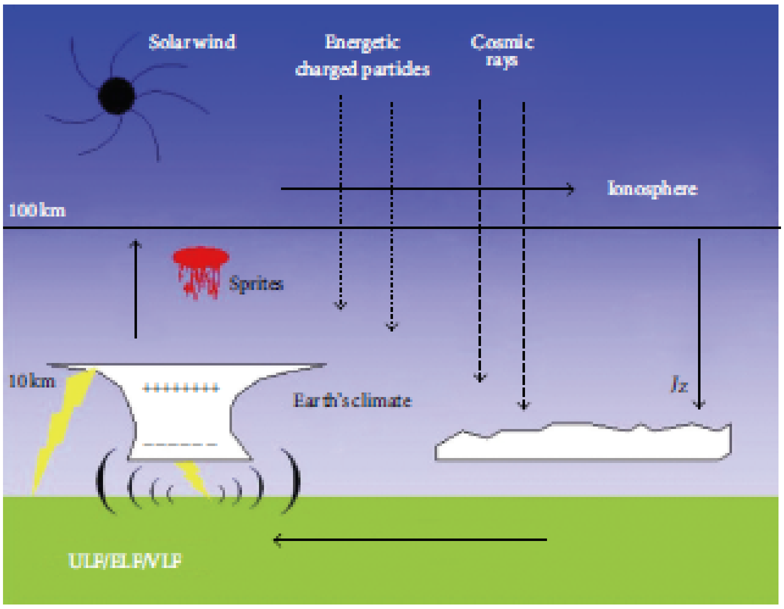

Figure 4: Essential features of the global atmospheric electric circuit [26, 27].

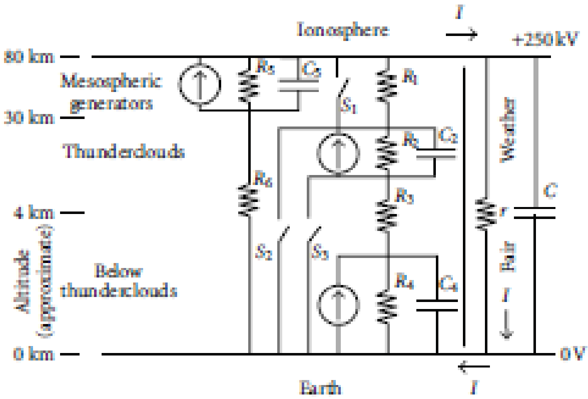

Figure 5: Diagram showing a schematic equivalent circuit for global electric circuits, credit [28].

system, that electric currents redistribute charge and that the electric currents are continuous.

The power supplied by thunderstorms is insufficient to maintain a field of magnitude observed in fair-weather regions. Rycroft et al. [61] included the generator associated with electrified clouds in the GEC model; this was found to be of the same magnitude as that due to thunderstorms. The optical phenomena occurring in the upward branch of the GEC above the thunderstorms are likely to influence only the upper atmosphere conductivity. Since they occur much less frequently (only one sprite out of 200 lightning) because of their association with intense lightning discharges [50, 62], their contribution to the ionospheric potential is very small [61]. The gigantic jets transport large quantities of negative charge discharging the atmospheric capacitor [63–66] whose effects on the ionosphere and GEC have not yet been modeled. The role of sprite/TLE events on the flow, charging/discharging of GEC, modification of electric fields near the Earth's surface remains unanswered. Since optical emissions could change electrical properties of the atmosphere and influence processes related with weather and climate, intense research activity in this area is required.

The earthquakes affect the electrodynamics of the atmosphere through the generation of electric and magnetic fields with crustal deformation, fault-failure-related piezomagnetism, stress/conductivity, electrokinetic effects, charge generation processes, thermal remagnetization, and demagnetization effects, and so forth [67]. These processes in the Earth's lithosphere relate with disturbances in the atmosphere and ionosphere. Sorokin et al. [8] discussed the processes forming the electrodynamic model of the effect of seismic and meteorological phenomena on the ionosphere. Radioactive substances and charged aerosols injected into the atmosphere modify the altitude profile of conductivity, generation of external currents, perturbation of electric field, and current in the ionospheric layer. As a result, Joule heating of the ionosphere and instability of acoustic gravity waves take place, which manifests in the formation of horizontal inhomogeneities of ionospheric conductivity. Finally, excitation of plasma density fluctuations and ULF/ELF emissions in the ionosphere, generation of field aligned currents and plasma layers, upward plasma transport, and modification of F_2-layer, and change in the ion composition of the upper ionosphere take place [68, 69]. These changes may also affect the GEC and the Earth's climate which remains a challenging problem to be solved. Figure 6 shows a schematic diagram which can be used to

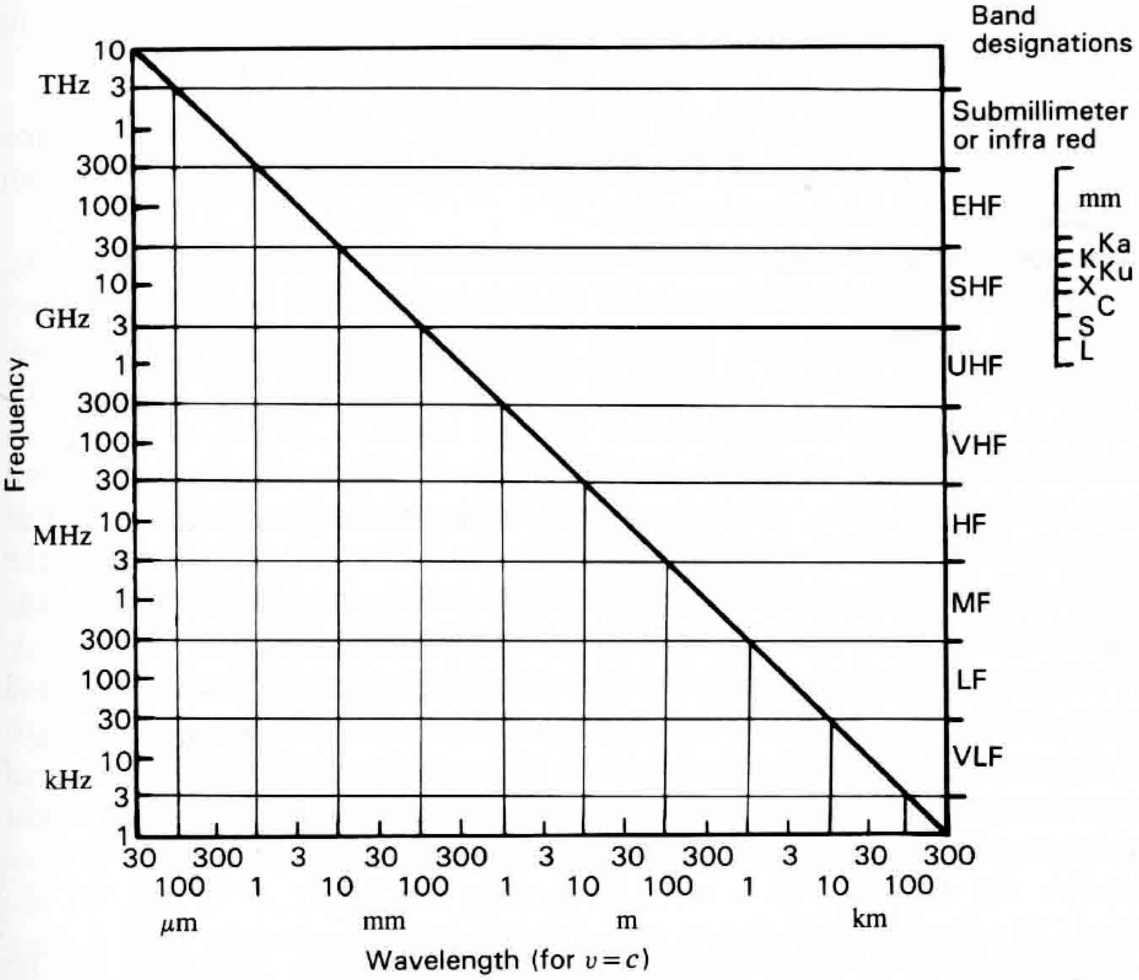

Figure 1-7 Wavelength versus frequency for $v = c$.

Example of wavelength for a given frequency. For a frequency of 300 MHz the corresponding wavelength is given by

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8 \text{ m s}^{-1}}{300 \times 10^6 \text{ Hz}} = 1 \text{ m} \tag{3}$$

In a lossless nonmagnetic dielectric medium with relative permittivity $\varepsilon_r = 2$, the same wave has a velocity

$$v = \frac{c}{\sqrt{\varepsilon_r}} = \frac{3 \times 10^8}{\sqrt{2}} = 2.12 \times 10^8 \text{ m s}^{-1} \tag{4}$$

and

$$\lambda = \frac{v}{f} = \frac{2.12 \times 10^8}{300 \times 10^6} = 0.707 \text{ m} = 707 \text{ mm} \tag{5}$$

1-4 DIMENSIONS AND UNITS. Lord Kelvin is reported to have said:

> When you can measure what you are speaking about and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers your knowledge is of a meagre and unsatisfactory kind; it may

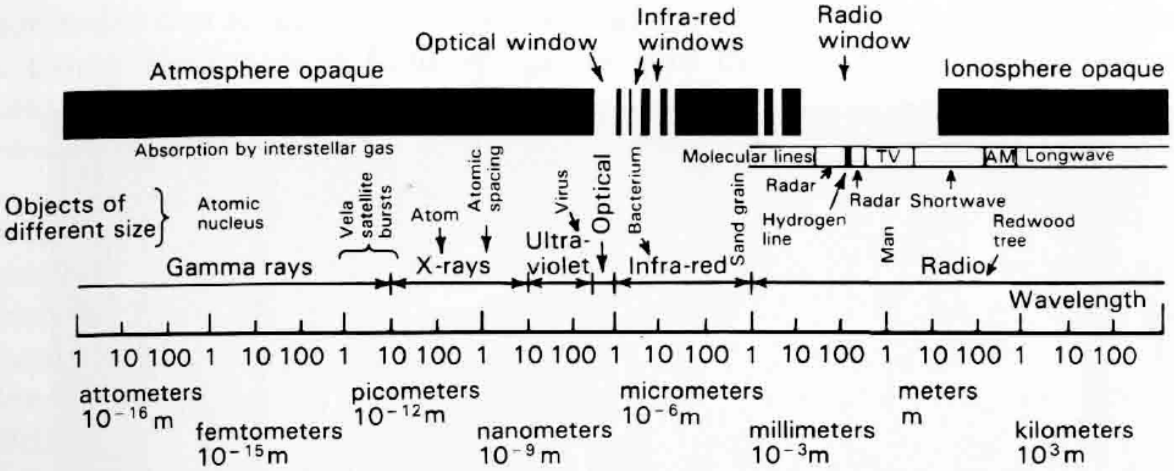
Atmosphere opaque
Optical window
Infra-red windows
Radio window
Ionosphere opaque
Absorption by interstellar gas
Molecular lines
TV
AM
Longwave
Radar
Radar
Shortwave
Objects of different size
Atomic nucleus
Vela satellite bursts
Atom
Atomic spacing
Virus
Optical
Bacterium
Sand grain
Hydrogen line
Man
Redwood tree
Gamma rays
X-rays
Ultra-violet
Infra-red
Radio
Wavelength
1 10 100 1 10 100 1 10 100 1 10 100 1 10 100 1 10 100 1 10 100 1 10 100
attometers 10^{-16}m
femtometers 10^{-15}m
picometers 10^{-12}m
nanometers 10^{-9}m
micrometers 10^{-6}m
millimeters 10^{-3}m
meters m
kilometers 10^{3}m

Figure 1-6 The electromagnetic spectrum with wavelength on a logarithmic scale from the shortest gamma rays to the longest radio waves. The atmospheric-ionospheric opacity is shown at the top with the optical and radio windows in evidence.

Thus, the wavelength depends on the velocity v which depends on the medium. In this sense, frequency is a more fundamental quantity since it is independent of the medium. When the medium is free space (vacuum)

$$v = c = 3 \times 10^8 \text{ m s}^{-1} \tag{2}$$

Figure 1-7 shows the relation of wavelength to frequency for $v = c$ (free space). Many of the uses of the spectrum are indicated along the right-hand edge of the figure. A more detailed frequency use listing is given in Table 1-1.

Table 1-1 Radio-frequency band designations

Frequency	Wavelength	Band designation
30–300 Hz	10–1 Mm	ELF (extremely low frequency)
300–3000 Hz	1 Mm–100 km	
3–30 kHz	100–10 km	VLF (very low frequency)
30–300 kHz	10–1 km	LF (low frequency)
300–3000 kHz	1 km–100 m	MF (medium frequency)
3–30 MHz	100–10 m	HF (high frequency)
30–300 MHz	10–1 m	VHF (very high frequency)
300–3000 MHz	1 m–10 cm	UHF (ultra high frequency)
3–30 GHz	10–1 cm	SHF (super high frequency)
30–300 GHz	1 cm–1 mm	EHF (extremely high frequency)
300–3000 GHz	1 mm–100 μm	

Frequency	Wavelength	IEEE Radar Band designation
1–2 GHz	30–15 cm	L
2–4 GHz	15–7.5 cm	S
4–8 GHz	7.5–3.75 cm	C
8–12 GHz	3.75–2.50 cm	X
12–18 GHz	2.50–1.67 cm	Ku
18–27 GHz	1.67–1.11 cm	K
27–40 GHz	1.11 cm–7.5 mm	Ka
40–300 GHz	7.5–1.0 mm	mm

easily by having many different tapping points on the coil so that adjustments can be made.

To make tuning easier a component called a tuning capacitor can be included in the circuit. In very early sets a tuning capacitor was not always included to keep costs down, or due to their being difficult to obtain.

The Detector converts the radio wave received into an electrical wave that is suitable for the headphones to, in turn, convert into sound waves that can be heard by the human ear. In the very early days of the crystal set the detector consisted of a holder containing a piece of galena crystal that had a very thin and springy wire placed on its surface that had to be very delicately adjusted to find the sweet spot where the radio station could be heard. This was commonly referred to as a "Cat's Whisker". Modern detectors are called diodes and are more efficient than early detectors and cats whiskers. Diodes are still quite readily available and inexpensive. Part numbers for modern diodes include OA90, OA91 and IN34. Diode part numbers that are perhaps now more difficult to obtain include OA47 and OA81.

The headphones have to be of a very special type called high impedance headphones. Because there is no additional power source in a crystal set the current generated in the circuit is tiny - minuscule in fact. Ordinary low impedance headphones, such as Walkman headphones, would present a virtual short circuit to the crystal set allowing the tiny signals to drain away to earth & consequently producing no sound - not very useful!

High impedance headphones, on the other hand, reduce or impede the flow of current down to earth, in effect saving the tiny signals to produce sounds from the headphones that we can then hear.

There is a problem however, these high impedance headphones that were so readily available in the 1920's and 1930's are to so easy to obtain today, but some specialist vintage radio outlets still stock them though the price can be quite high. They can still be seen in museums of course. All is not lost though, today we can obtain a special earphone called, appropriately, a crystal earpiece very easily and far more cheaply than. An electronic component, called a resistor, must be connected accross the crystal earphone to allow a path for DC current to get to earth. The value of the resistor is usually 47,000 Ohms and without it a crystal earphone tends to block DC current and as a consequence the sound will be very quiet and distorted.

Aerial and Earth: For a crystal set work it needs an efficient aerial and earth. Since a crystal radio has no power - no batteries or mains electricity - it relies entirely on the radio wave energy sent out from the radio station's transmitter and collected by the aerial to work. The aerial simply consists of a length of wire, but it necessarily has to be quite long, usually in the order of 10 to 20 meters. For the aerial to be effective the crystal radio set also has to be connected to a good earth point. A good earth often consists of a 3 or 4 foot copper stake driven into the ground, but sometimes a water pipe can be used to reasonable effect. (Safety: Never use the earth pin of a household mains plug)

The circuit diagram (schematic) for a crystal set is shown below. Practical designs are shown on the following pages.

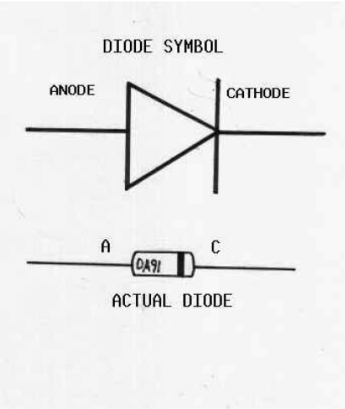
DIODE SYMBOL
ANODE
CATHODE
A
C
OA91
ACTUAL DIODE

An illustration of a modern diode which is often encapsulated in glass and is about 7mm long. Above is the electronic symbol for a diode.

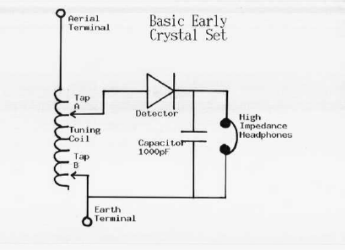
Aerial Terminal
Basic Early Crystal Set
Tap A
Detector
Tuning Coil
Capacitor 1000pF
High Impedance Headphones
Tap B
Earth Terminal

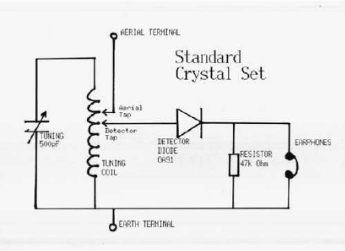
AERIAL TERMINAL
Standard Crystal Set
Aerial Tap
Detector Tap
TUNING 500pF
DETECTOR DIODE OA91
TUNING COIL
RESISTOR 47k Ohm
EARPHONES
EARTH TERMINAL

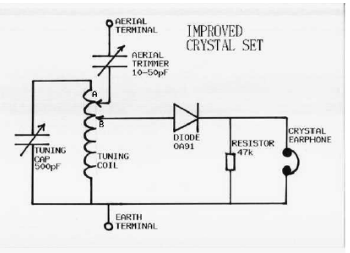
AERIAL TERMINAL
IMPROVED CRYSTAL SET
AERIAL TRIMMER 10-50pF
A
B
DIODE OA91
CRYSTAL EARPHONE
RESISTOR 47k
TUNING CAP 500pF
TUNING COIL
EARTH TERMINAL

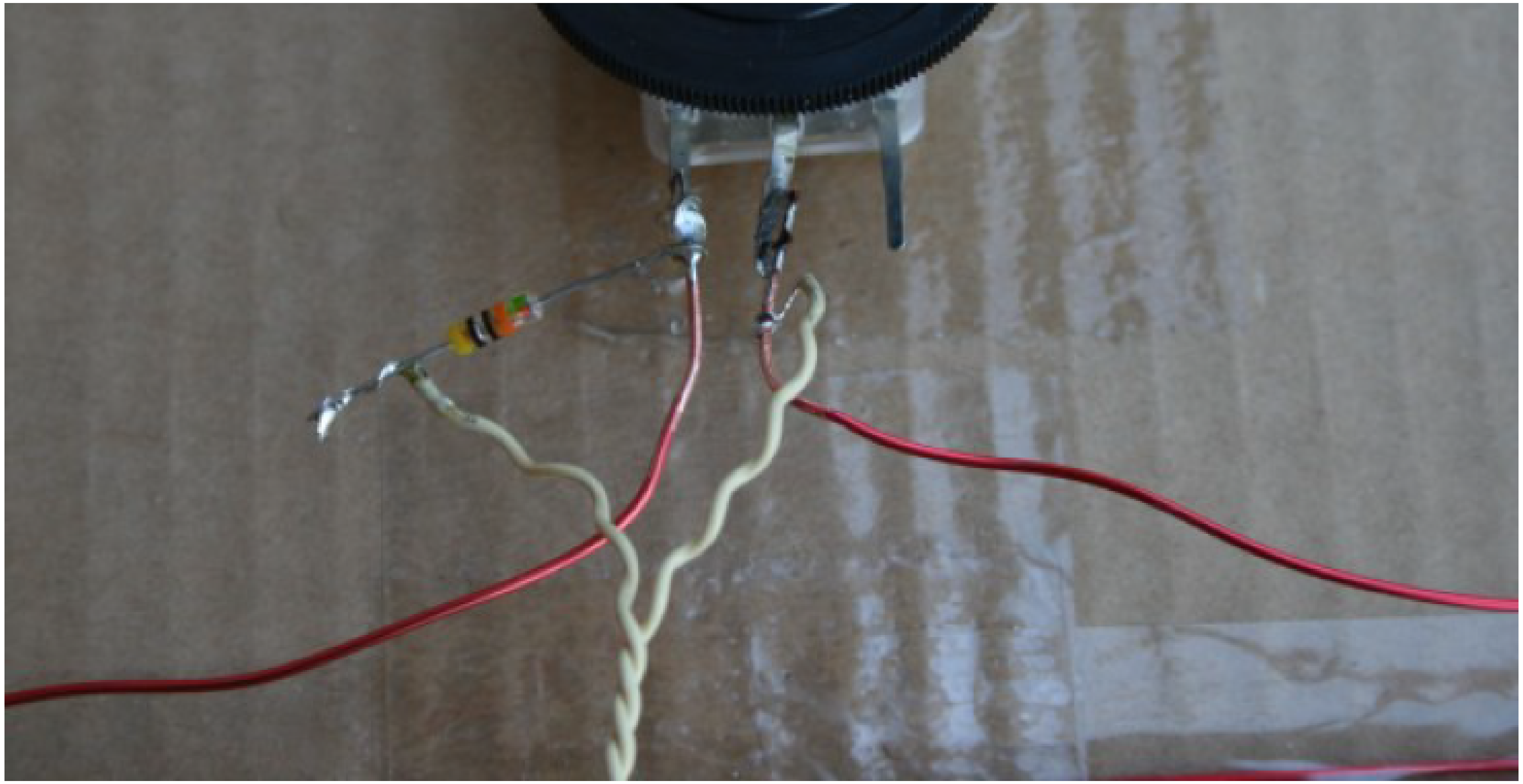

Click on photo for a larger picture

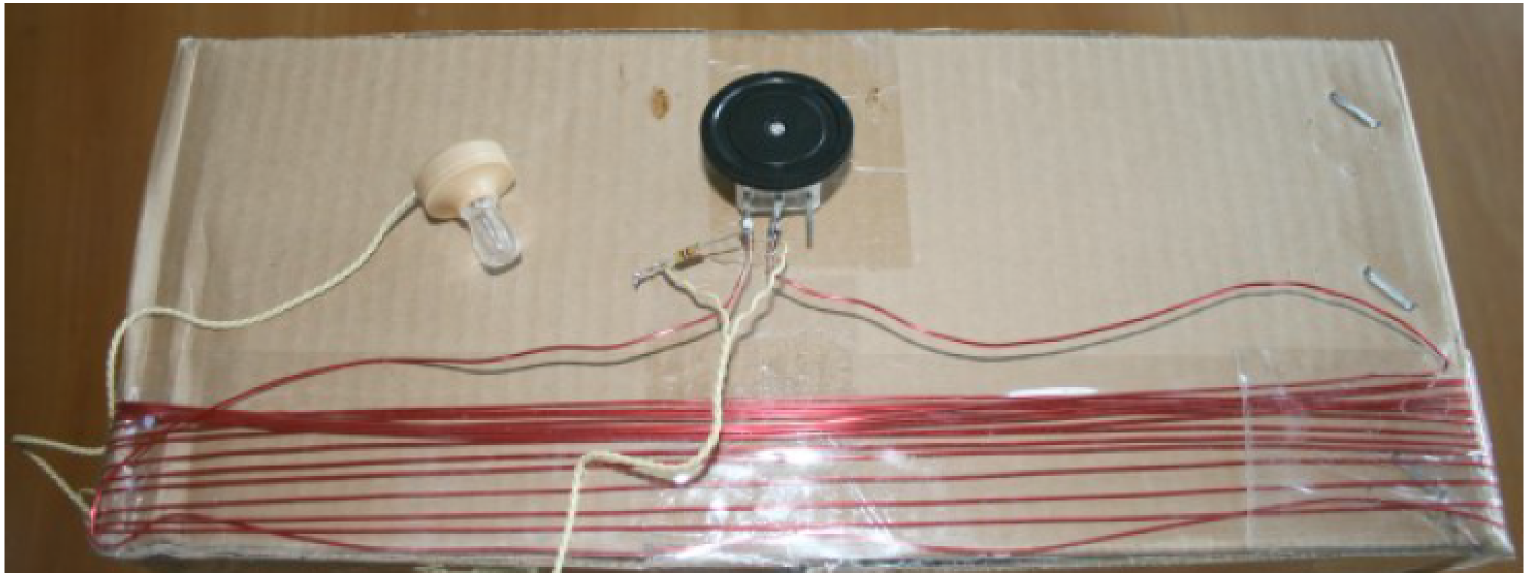

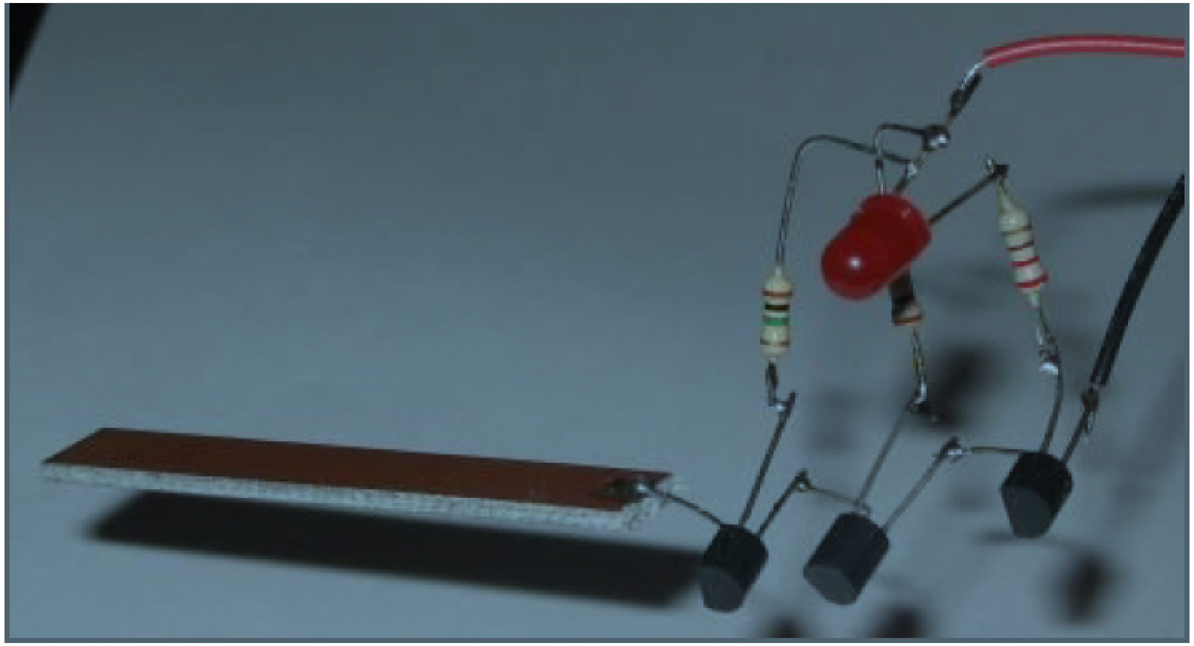

8 MILLION GAIN!

This circuit is so sensitive it will detect "mains hum." Simply move it across any wall and it will detect where the mains cable is located. It has a gain of about 200 x 200 x 200 = 8,000,000 and will also detect static electricity and the presence of your hand without any direct contact. You will be amazed what it detects! There is static electricity EVERYWHERE! The input of this circuit is classified as very high impedance.

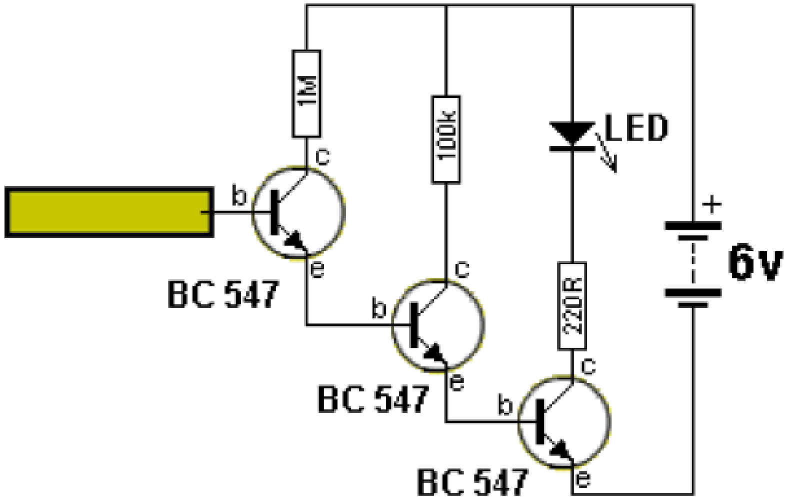

Here is a photo of the circuit, produced by a constructor, where he claimed he detected "ghosts."

portable
DX
pedestrian mo
contests
vk3ye dot com
antennas
CB
shortwave
videos
HF

[Return to VK3YE amateur radio projects](http://home.alphalink.com.au/~parkerp/gateway/nodec00.htm)

A simple crystal set for free power radio

Of any electronic project, the crystal set would have to rate as one of the most popular. Many amateurs are on the air today because of their early construction of a crystal set. Most practical electronic books for beginners include at least one crystal set project. Unfortunately, some of these circuits take simplicity too far and deliver mediocre performance, often by omitting key components such as the tuning capacitor, or failing to provide coil taps.

This article describes a crystal set of medium complexity. It features coil taps for the antenna and diode to make it useful for both country and metropolitan listeners. The taps allow the set to cover 160 metres if desired. All parts are easily obtainable, making it a good choice for the beginner. The endless possibilities for experimentation also make crystal sets interesting novelty projects for experienced constructors. The schematic is shown here:

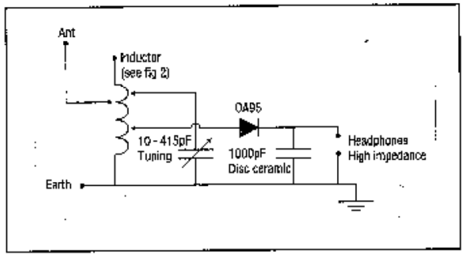
Ant
Inductor
(see fig 2)
OA95
10 - 415pF
Tuning
1000pF
Disc ceramic
Headphones
High impedance
Earth

Figure One: Schematic diagram of crystal set

VK3YE Crystal Set

Schematic & Layout diagrams

Press Back button to return to main article

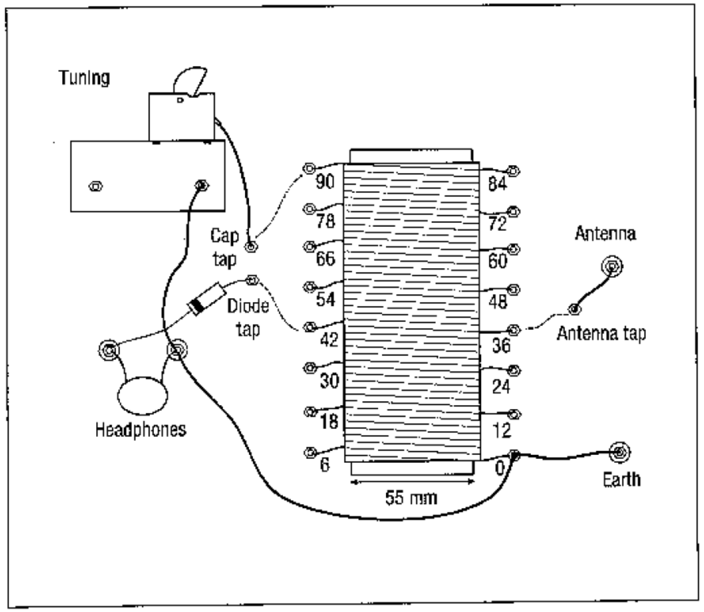
Tuning
Cap
tap
Diode
tap
Headphones
90
78
66
54
42
30
18
6
84
72
60
48
36
24
12
0
55 mm
Antenna
Antenna tap
Earth

Figure Two: Rear view of front panel, showing coil details

(from Amateur Radio February 2001, page 20)

Obtaining the parts

Tuning capacitor

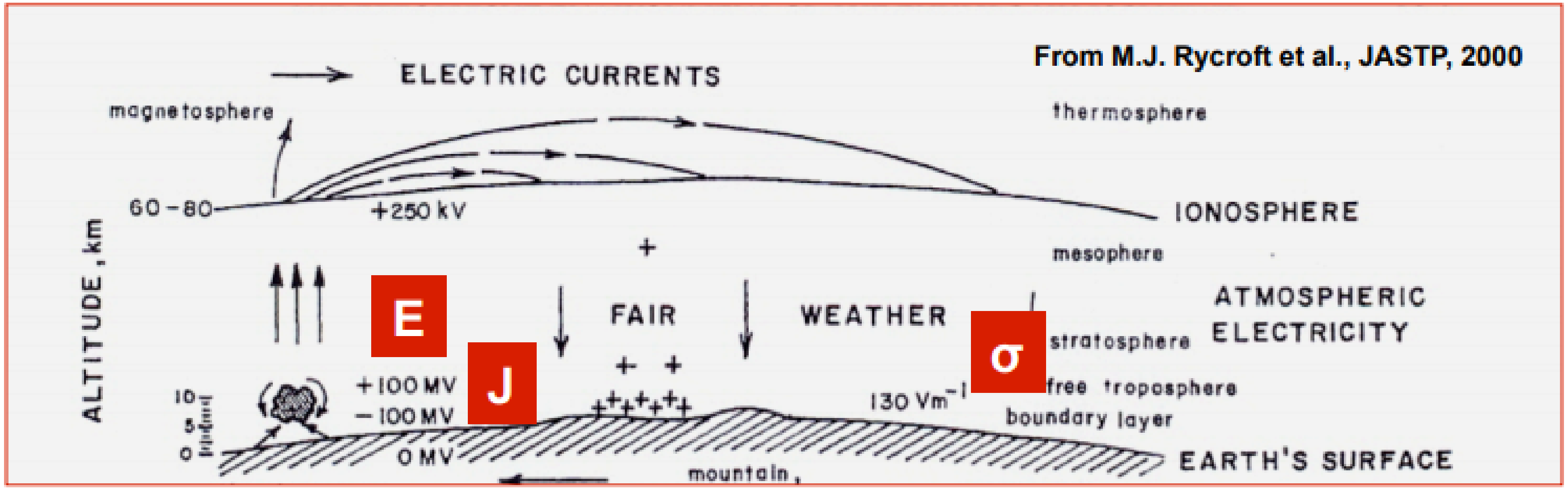
ELECTRIC CURRENTS
magnetosphere
thermosphere
60 - 80
+250 kV
IONOSPHERE
ALTITUDE, km
+
mesophere
E
FAIR
WEATHER
ATMOSPHERIC
ELECTRICITY
J
σ
stratosphere
+100 MV
-100 MV
+ +
130 Vm^{-1}
free troposphere
10
5
0
boundary layer
0 MV
mountain,
EARTH'S SURFACE

From M.J. Rycroft et al., JASTP, 2000

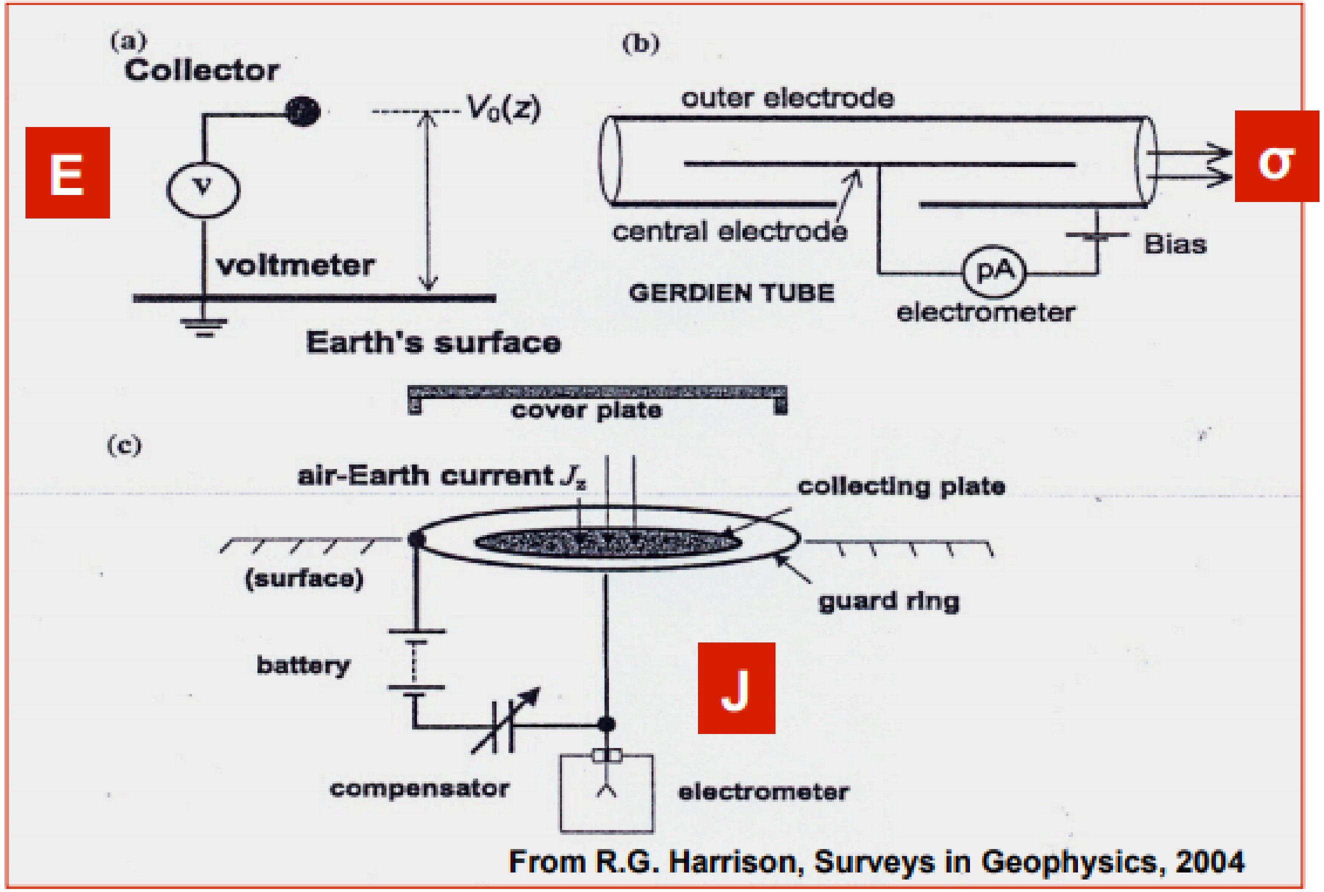
(a)
Collector
$V_0(z)$
E
V
voltmeter
Earth's surface
(b)
outer electrode
σ
central electrode
Bias
GERDIEN TUBE
pA
electrometer
cover plate
(c)
air-Earth current J_z
collecting plate
(surface)
guard ring
battery
J
compensator
electrometer

From R.G. Harrison, Surveys in Geophysics, 2004

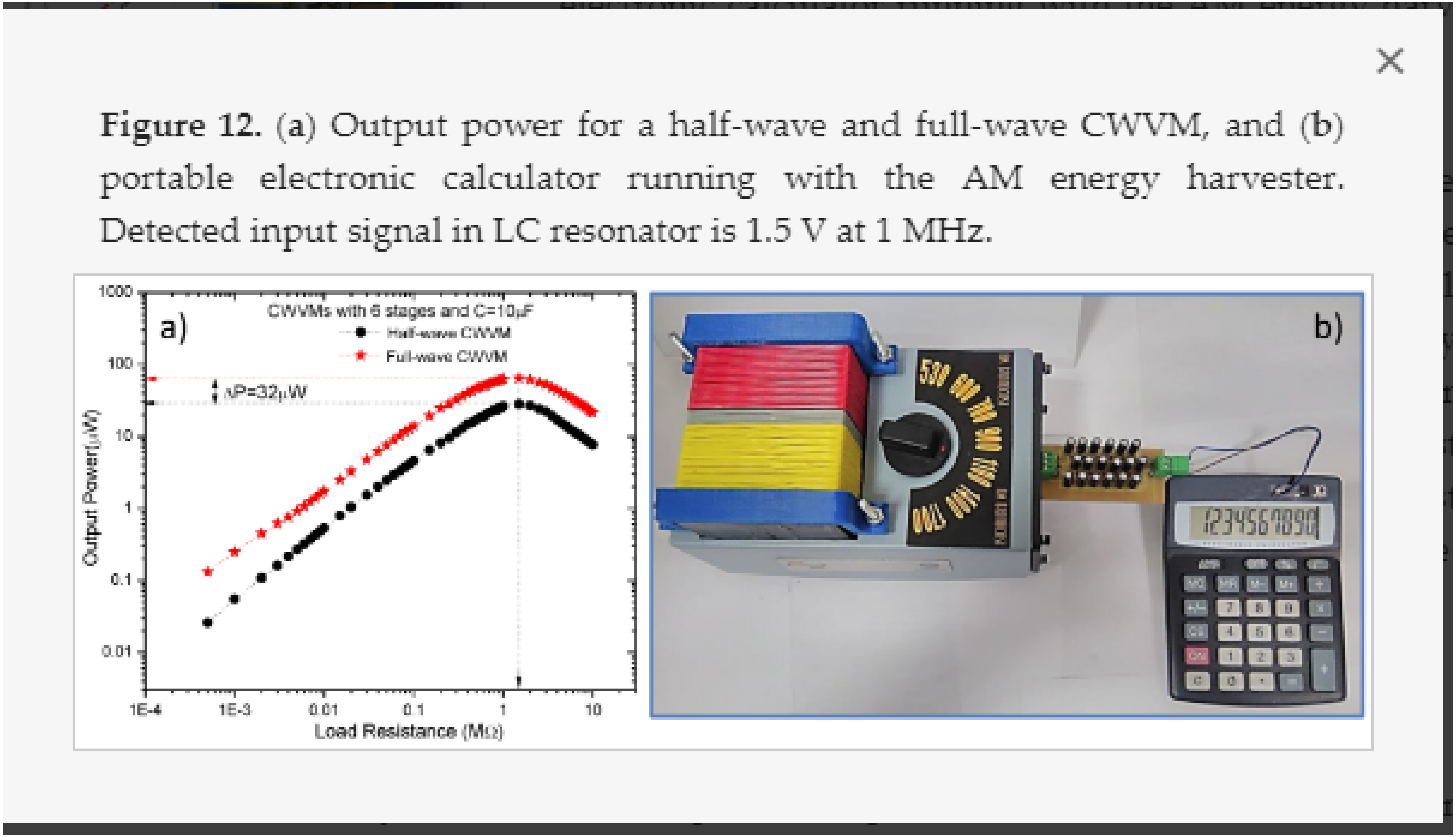

Figure 12. (**a**) Output power for a half-wave and full-wave CWVM, and (**b**) portable electronic calculator running with the AM energy harvester. Detected input signal in LC resonator is 1.5 V at 1 MHz.

a)
CWVMs with 6 stages and C=10μF
Half-wave CWVM
Full-wave CWVM
ΔP=32μW
Output Power(μW)
1000
100
10
1
0.1
0.01
1E-4
1E-3
0.01
0.1
1
10
Load Resistance (MΩ)
b)
1234567890

Figure 6. Pictorial image of the AM-RF energy harvesting system.

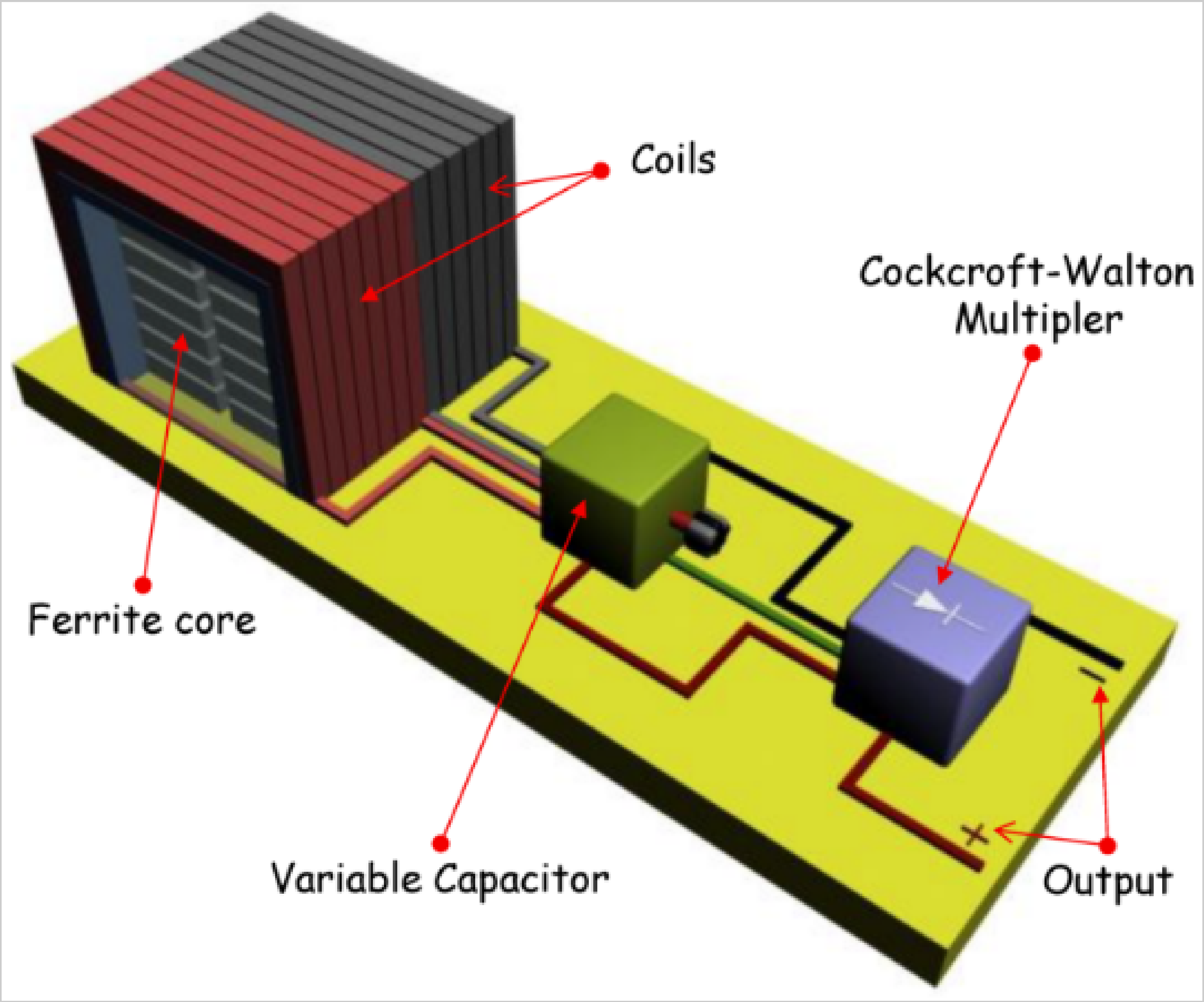
Coils
Cockcroft-Walton Multipler
Ferrite core
Variable Capacitor
Output
−
+

Figure 11. (a) AM resonator implemented, (b) Full-wave CWVM implemented, (c) top view of the AM resonator and full-wave CWVM, and (d) lateral view of the AM resonator and full-wave CWVM.

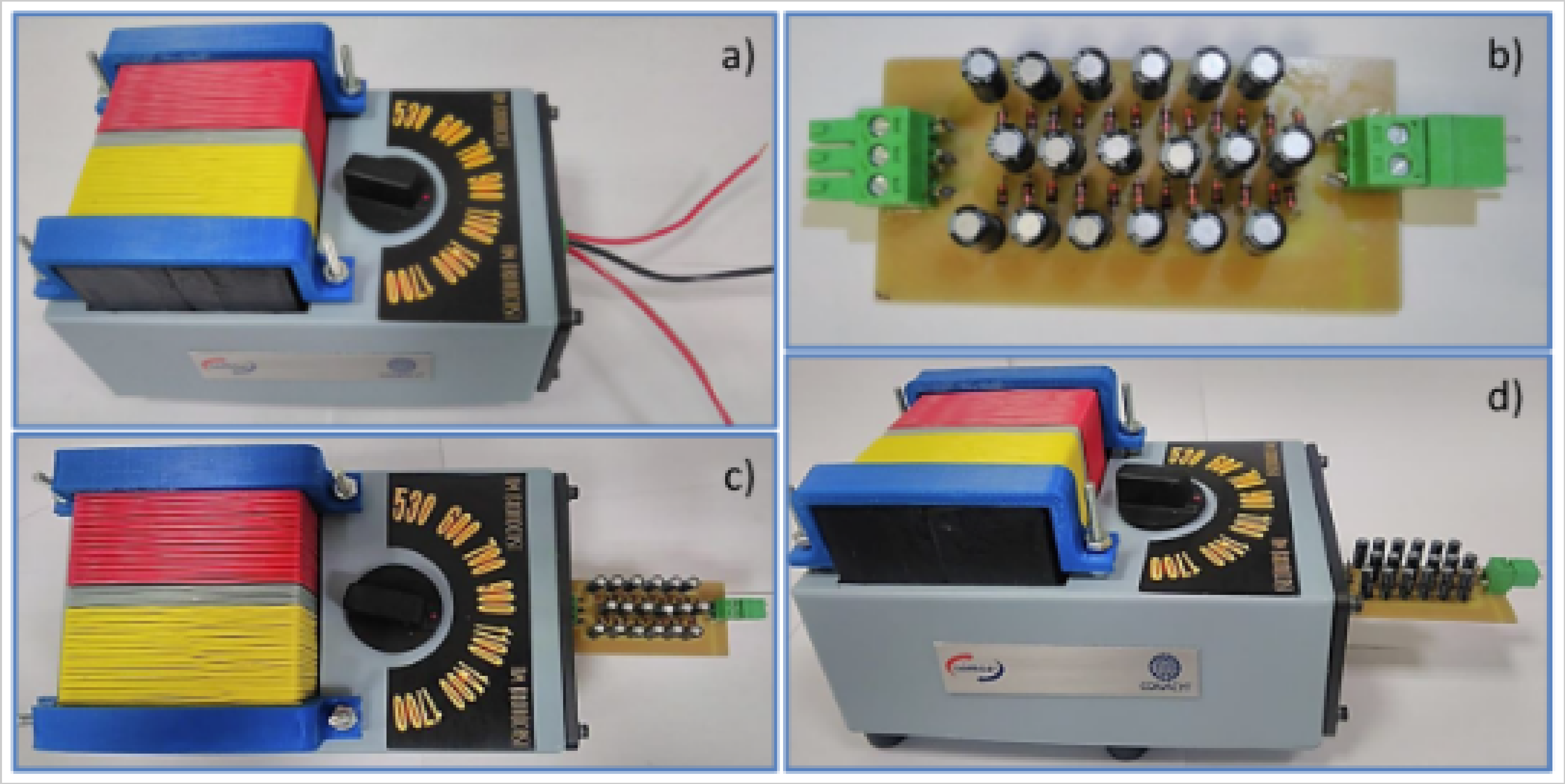

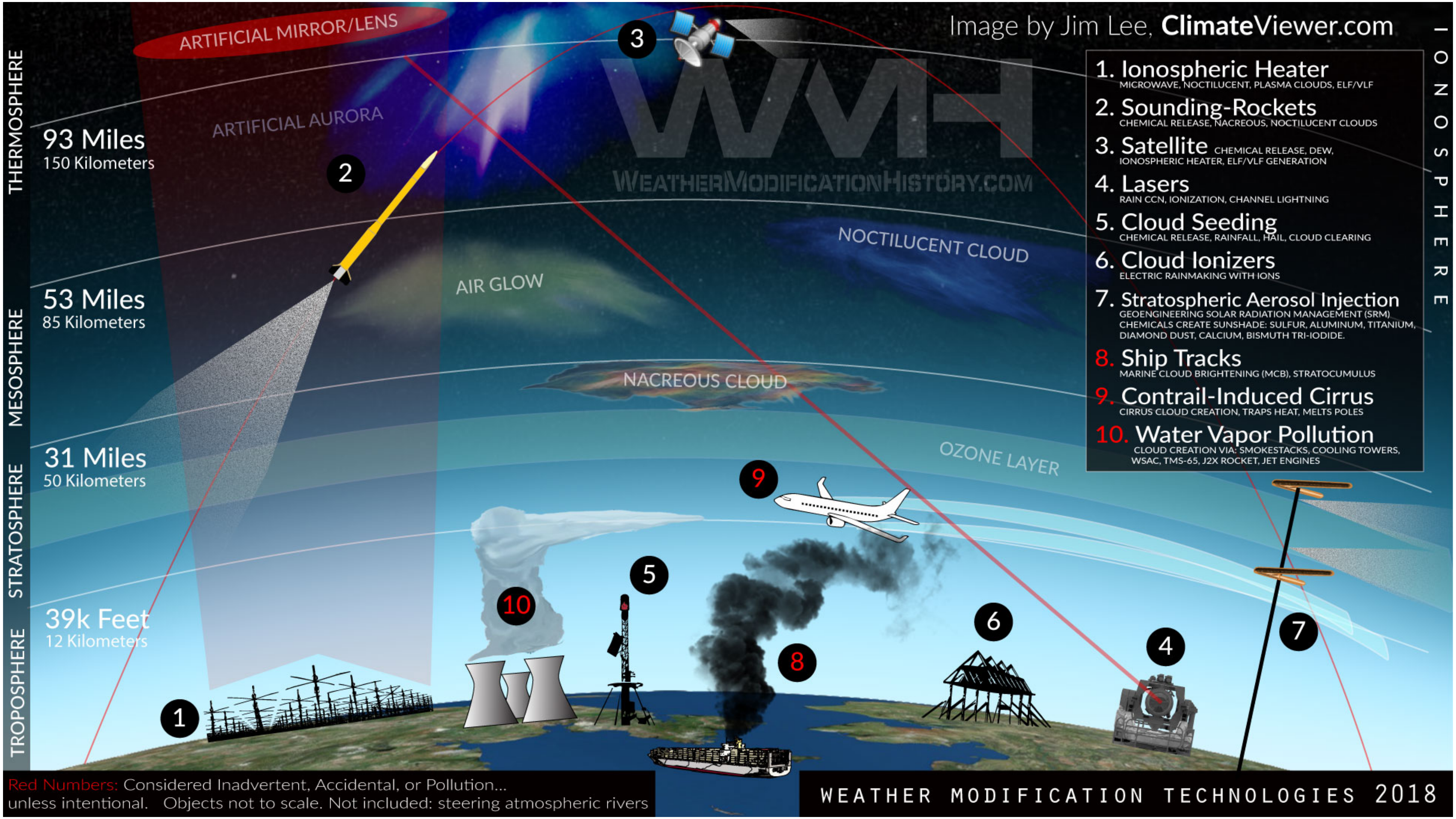

Image by Jim Lee, **ClimateViewer.com**

1. Ionospheric Heater
   MICROWAVE, NOCTILUCENT, PLASMA CLOUDS, ELF/VLF
2. Sounding-Rockets
   CHEMICAL RELEASE, NACREOUS, NOCTILUCENT CLOUDS
3. Satellite CHEMICAL RELEASE, DEW, IONOSPHERIC HEATER, ELF/VLF GENERATION
4. Lasers
   RAIN CCN, IONIZATION, CHANNEL LIGHTNING
5. Cloud Seeding
   CHEMICAL RELEASE, RAINFALL, HAIL, CLOUD CLEARING
6. Cloud Ionizers
   ELECTRIC RAINMAKING WITH IONS
7. Stratospheric Aerosol Injection
   GEOENGINEERING SOLAR RADIATION MANAGEMENT (SRM) CHEMICALS CREATE SUNSHADE: SULFUR, ALUMINUM, TITANIUM, DIAMOND DUST, CALCIUM, BISMUTH TRI-IODIDE.
8. Ship Tracks
   MARINE CLOUD BRIGHTENING (MCB), STRATOCUMULUS
9. Contrail-Induced Cirrus
   CIRRUS CLOUD CREATION, TRAPS HEAT, MELTS POLES
10. Water Vapor Pollution
    CLOUD CREATION VIA: SMOKESTACKS, COOLING TOWERS, WSAC, TMS-65, J2X ROCKET, JET ENGINES

IONOSPHERE
THERMOSPHERE
MESOSPHERE
STRATOSPHERE
TROPOSPHERE
93 Miles
150 Kilometers
53 Miles
85 Kilometers
31 Miles
50 Kilometers
39k Feet
12 Kilometers
ARTIFICIAL MIRROR/LENS
ARTIFICIAL AURORA
AIR GLOW
NOCTILUCENT CLOUD
NACREOUS CLOUD
OZONE LAYER
WVH
WEATHERMODIFICATIONHISTORY.COM
1
2
3
4
5
6
7
8
9
10

Red Numbers: Considered Inadvertent, Accidental, or Pollution... unless intentional. Objects not to scale. Not included: steering atmospheric rivers

WEATHER MODIFICATION TECHNOLOGIES 2018

GENERATORS, HIGH ALTITUDE DISCHARGES, TLE´s

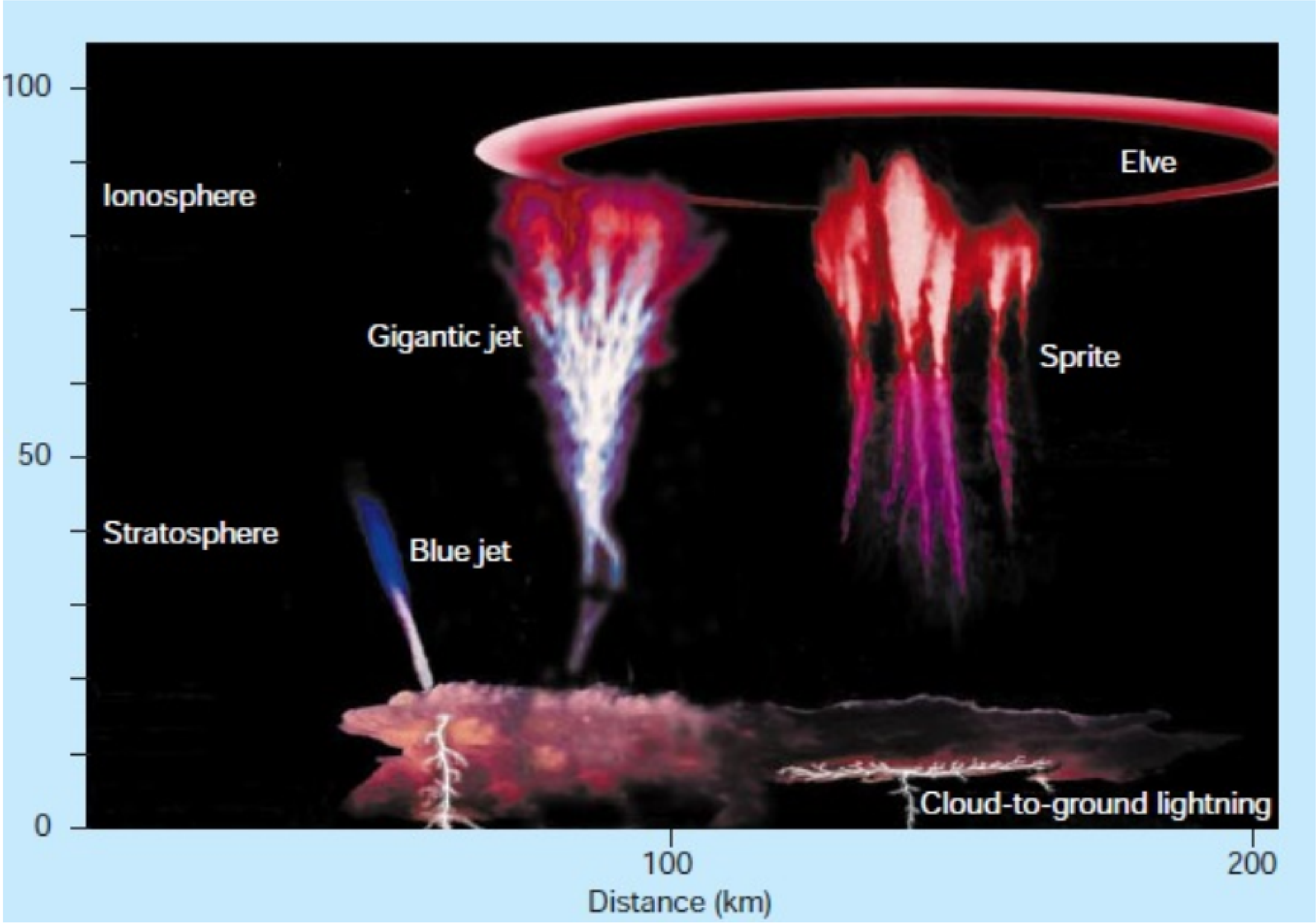
100
50
0
Ionosphere
Stratosphere
Elve
Gigantic jet
Sprite
Blue jet
Cloud-to-ground lightning
100
200
Distance (km)

Working Model of Crystal radio

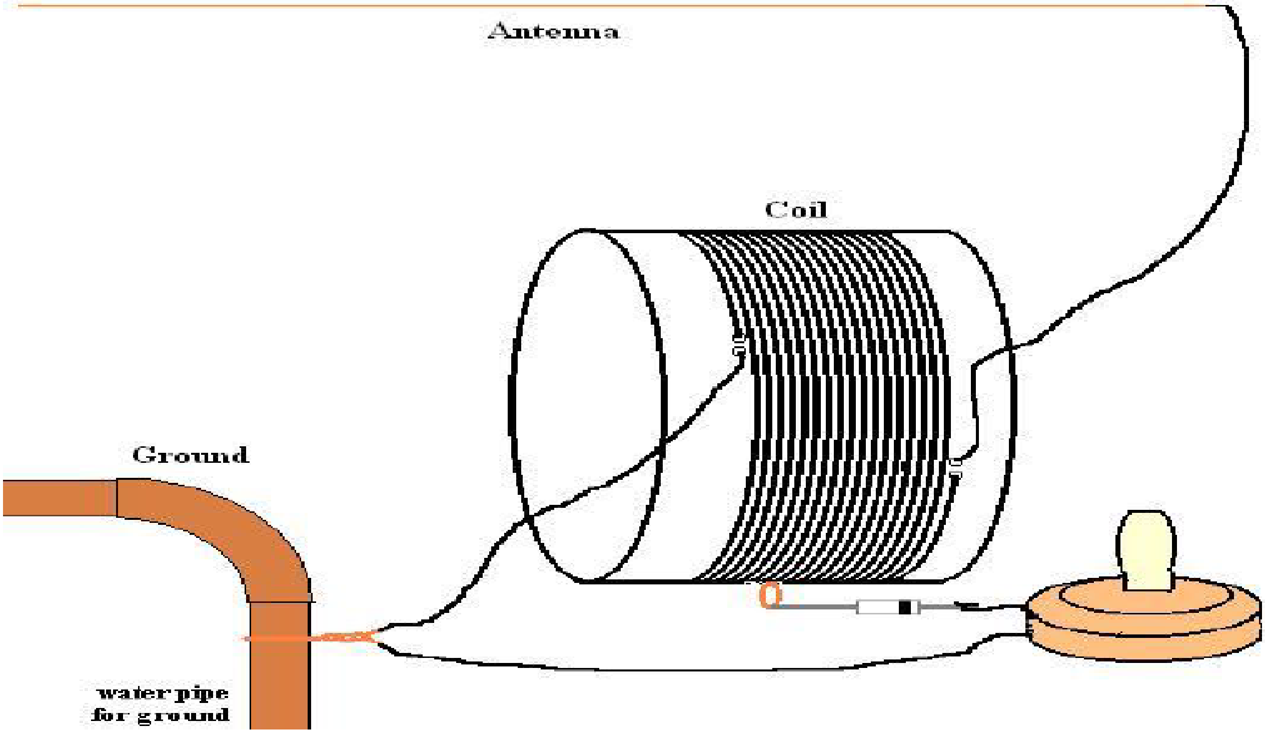
Antenna
Coil
Ground
water pipe
for ground

Schematic Diagram of the Radi

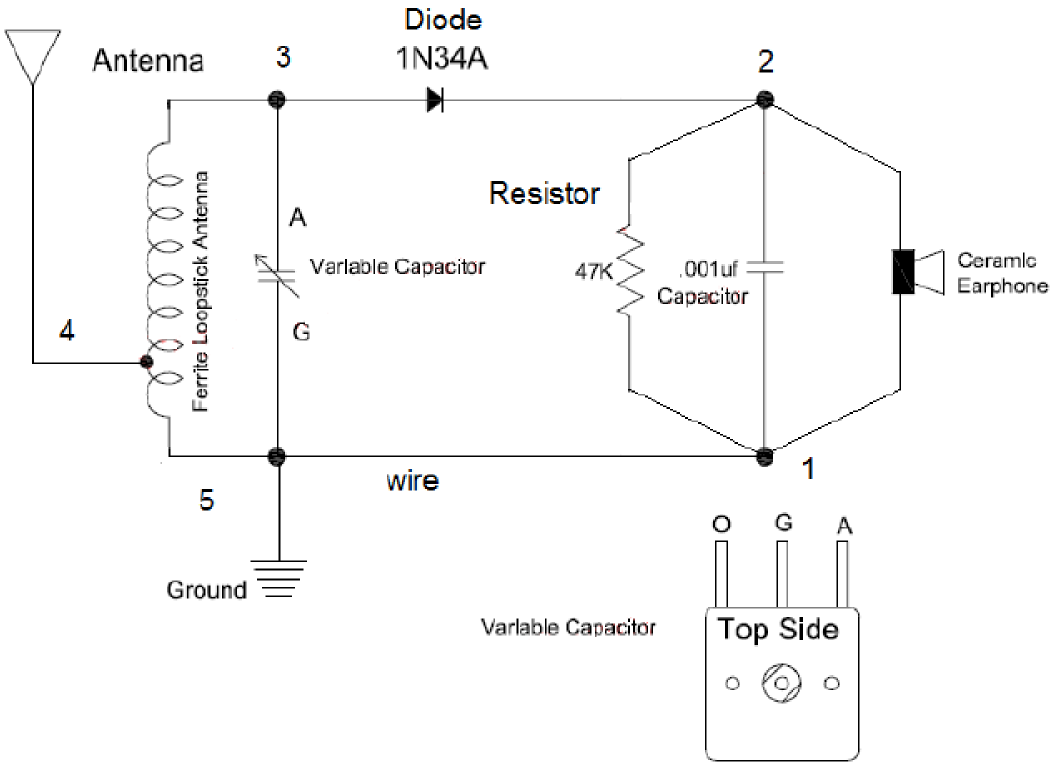
Diode
1N34A
Antenna
3
2
Resistor
A
Variable Capacitor
47K
.001uf
Capacitor
Ceramic
Earphone
4
G
Ferrite Loopstick Antenna
5
wire
1
Ground
O
G
A
Variable Capacitor
Top Side

Philmore

MANUFACTURING CO. Inc.
Richmond Hill 18, N. Y.

MODEL NO.
VC-1000

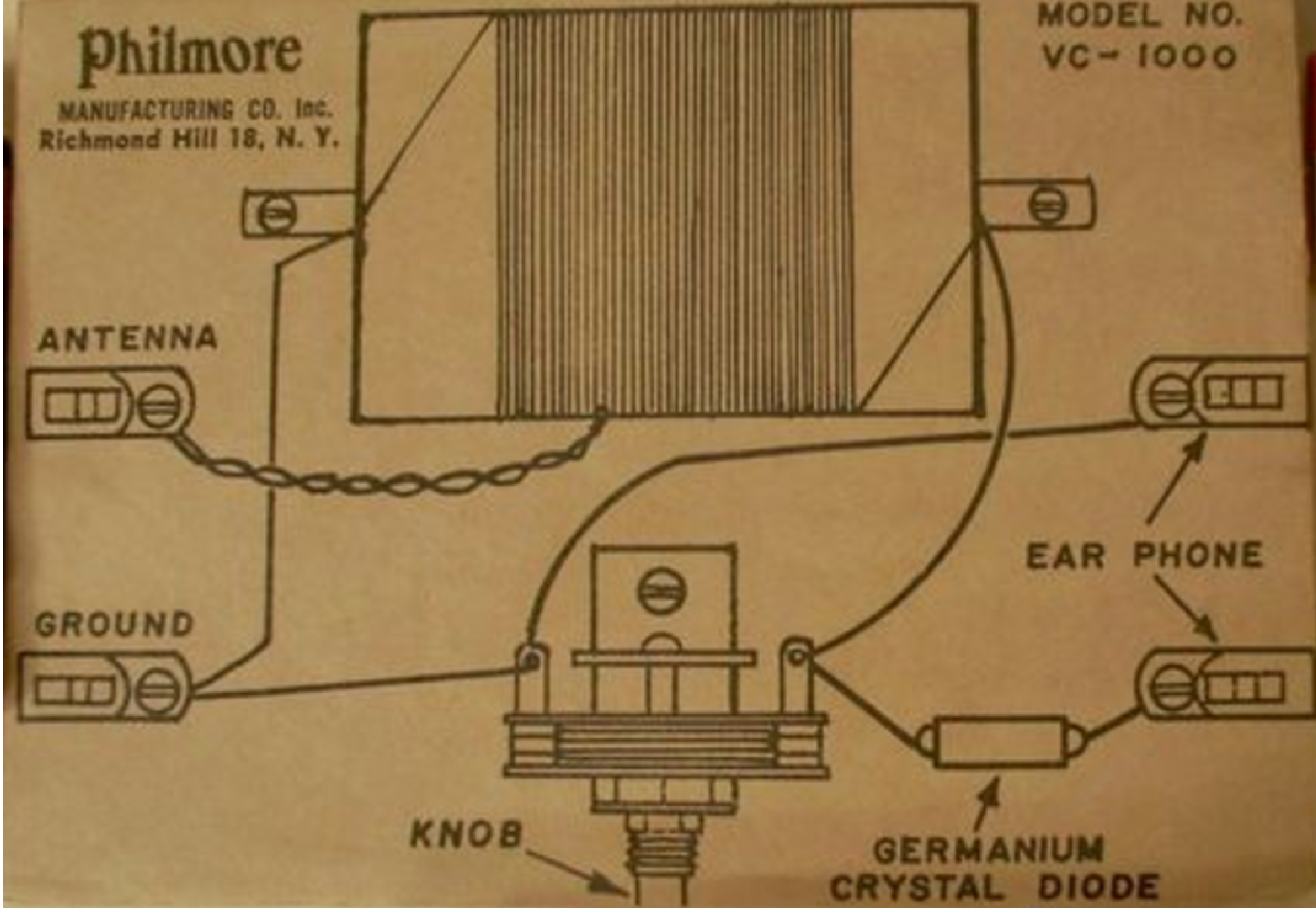
ANTENNA
GROUND
EAR PHONE
KNOB
GERMANIUM
CRYSTAL DIODE

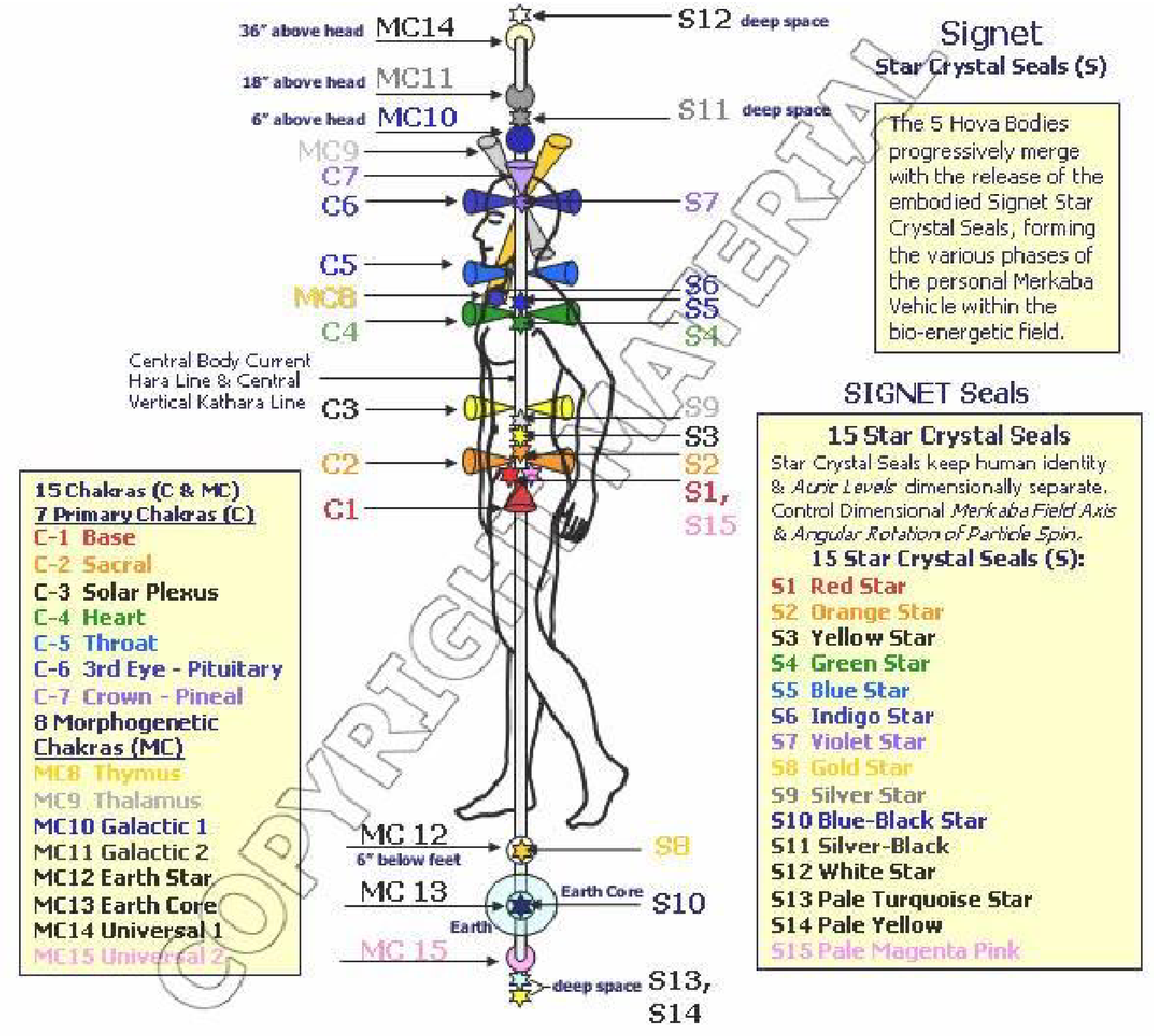

36" above head MC14
S12 deep space
18" above head MC11
6" above head MC10
S11 deep space
MC9
C7
C6
S7
C5
S6
MC8
S5
C4
S4
Central Body Current Hara Line & Central Vertical Kathara Line
C3
S9
S3
C2
S2
C1
S1,
S15
MC 12
6" below feet
S8
MC 13
Earth Core
S10
Earth
MC 15
deep space S13,
S14

Signet
Star Crystal Seals (S)

The 5 Hova Bodies progressively merge with the release of the embodied Signet Star Crystal Seals, forming the various phases of the personal Merkaba Vehicle within the bio-energetic field.

SIGNET Seals

15 Star Crystal Seals

Star Crystal Seals keep human identity & *Auric Levels* dimensionally separate. Control Dimensional *Merkaba Field Axis* & *Angular Rotation of Particle Spin.*

15 Star Crystal Seals (S):

- **S1 Red Star**
- **S2 Orange Star**
- **S3 Yellow Star**
- **S4 Green Star**
- **S5 Blue Star**
- **S6 Indigo Star**
- **S7 Violet Star**
- **S8 Gold Star**
- **S9 Silver Star**
- **S10 Blue-Black Star**
- **S11 Silver-Black**
- **S12 White Star**
- **S13 Pale Turquoise Star**
- **S14 Pale Yellow**
- **S15 Pale Magenta Pink**

15 Chakras (C & MC)

7 Primary Chakras (C)

- **C-1 Base**
- **C-2 Sacral**
- **C-3 Solar Plexus**
- **C-4 Heart**
- **C-5 Throat**
- **C-6 3rd Eye - Pituitary**
- **C-7 Crown - Pineal**

8 Morphogenetic Chakras (MC)

- **MC8 Thymus**
- **MC9 Thalamus**
- **MC10 Galactic 1**
- **MC11 Galactic 2**
- **MC12 Earth Star**
- **MC13 Earth Core**
- **MC14 Universal 1**
- **MC15 Universal 2**

COPYRIGHT MATERIAL

Copyright A'shayana & A'hzayana Deane, 1999-2007, All Rights Reserved; Part of the MCEO Freedom Teachings™ Series

Scalar Interferometry In The Endothermic (Cooling) Mode

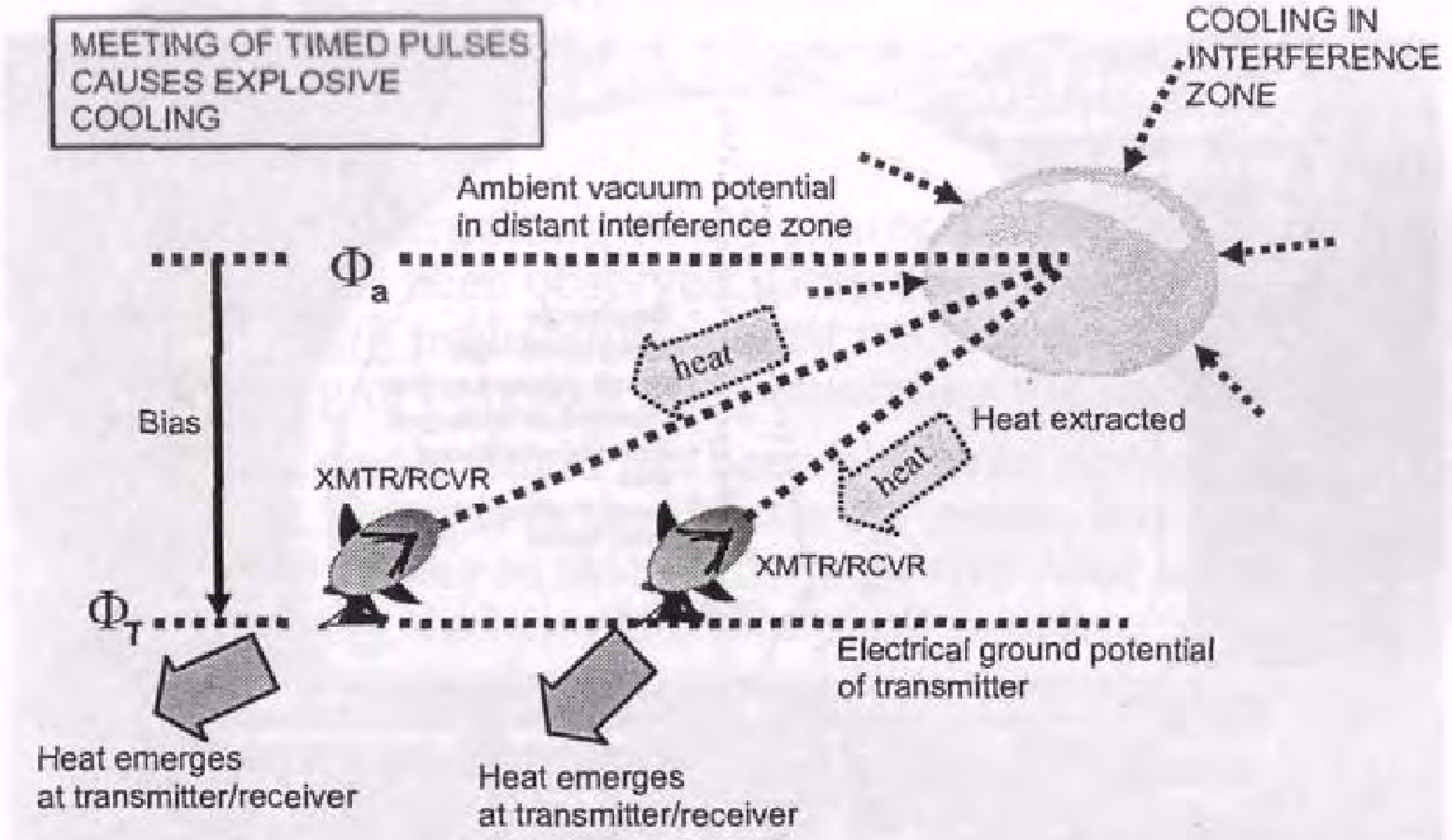
MEETING OF TIMED PULSES CAUSES EXPLOSIVE COOLING
COOLING IN INTERFERENCE ZONE
Ambient vacuum potential in distant interference zone
Φ_a
Bias
heat
heat
Heat extracted
XMTR/RCVR
XMTR/RCVR
Φ_T
Electrical ground potential of transmitter
Heat emerges at transmitter/receiver
Heat emerges at transmitter/receiver

In the endothermic mode, tremendous heat is received back in the transmitters. Soviets further retransmit the heat via interferometers to a distant "exhaust" point. Bennett Island was used for such an exhaust point for decades, and our observation platforms saw a great many 150 miles-long jets of exhaust effects in narrow beams at shallow angles.

Measuring Longitudinal and Transverse waves

The Fizzics Organisation

Transverse waves
A shaking rope is a good example. The wave moves along the rope but the bits of rope go up and down - across the direction of the wave. “Trans” means across, think of “transatlantic” or “transfer”.

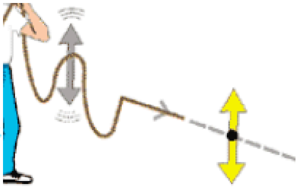

A water wave is another example of a transverse wave.

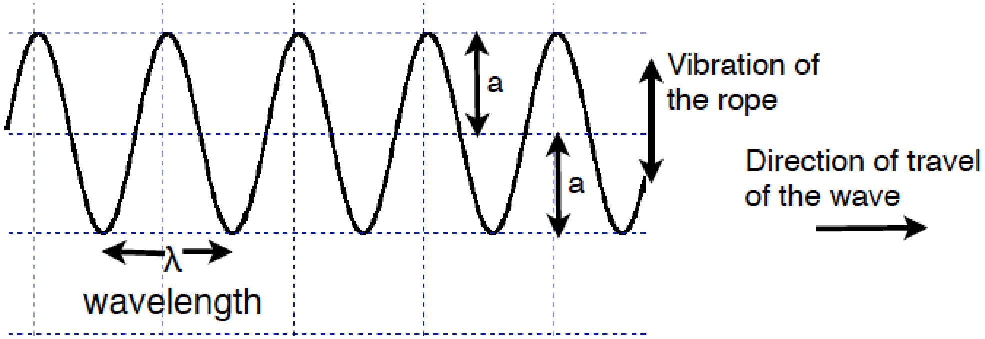

a is the amplitude of the wave measured from the maximum height to the middle. Think of a water wave, it is the distance from either the crest or the trough to where the water would be if it were calm.

λ is the wavelength. That is the distance from one point on a wave to the same point on the next wave - for example from crest to crest or trough to trough.

f is the frequency of the wave, that is the number of waves which are produced every second (or the number passing every second). This number is given the unit “hertz”. One hertz is one wave per second.

Most of the examples of waves we have to deal with are transverse waves. All the waves in the electromagnet spectrum are transverse waves.

There is a connection between the frequency and wavelength of a wave. If the frequency is increased - if there are far more waves, then they are closer together; in other words the wavelength is less.

The speed is equal to the frequency multiplied by the

$$V = f \times \lambda$$

m/s hertz metres

A longitudinal wave is a “pushing” wave. A line of dominos falling is an example.

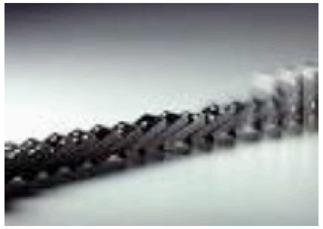

If one end of a slinky is pushed and pulled then each coil pushes and pulls the next. The coils oscillate along in the same direction that the wave is traveling.

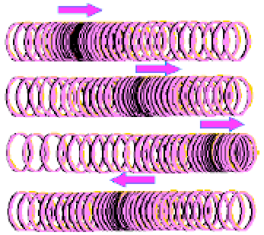

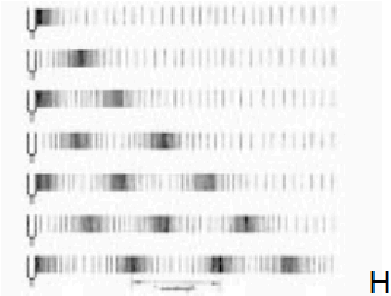

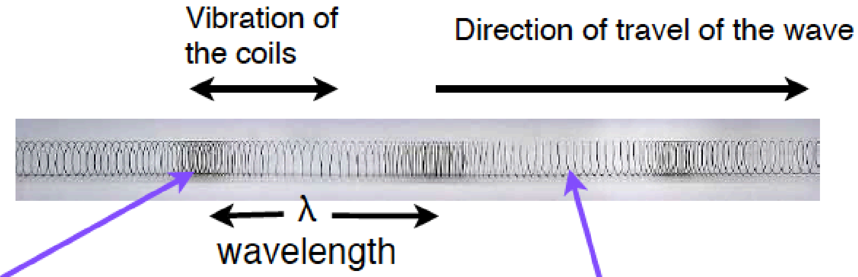

Here the coils are compressed together and here they are stretched out.

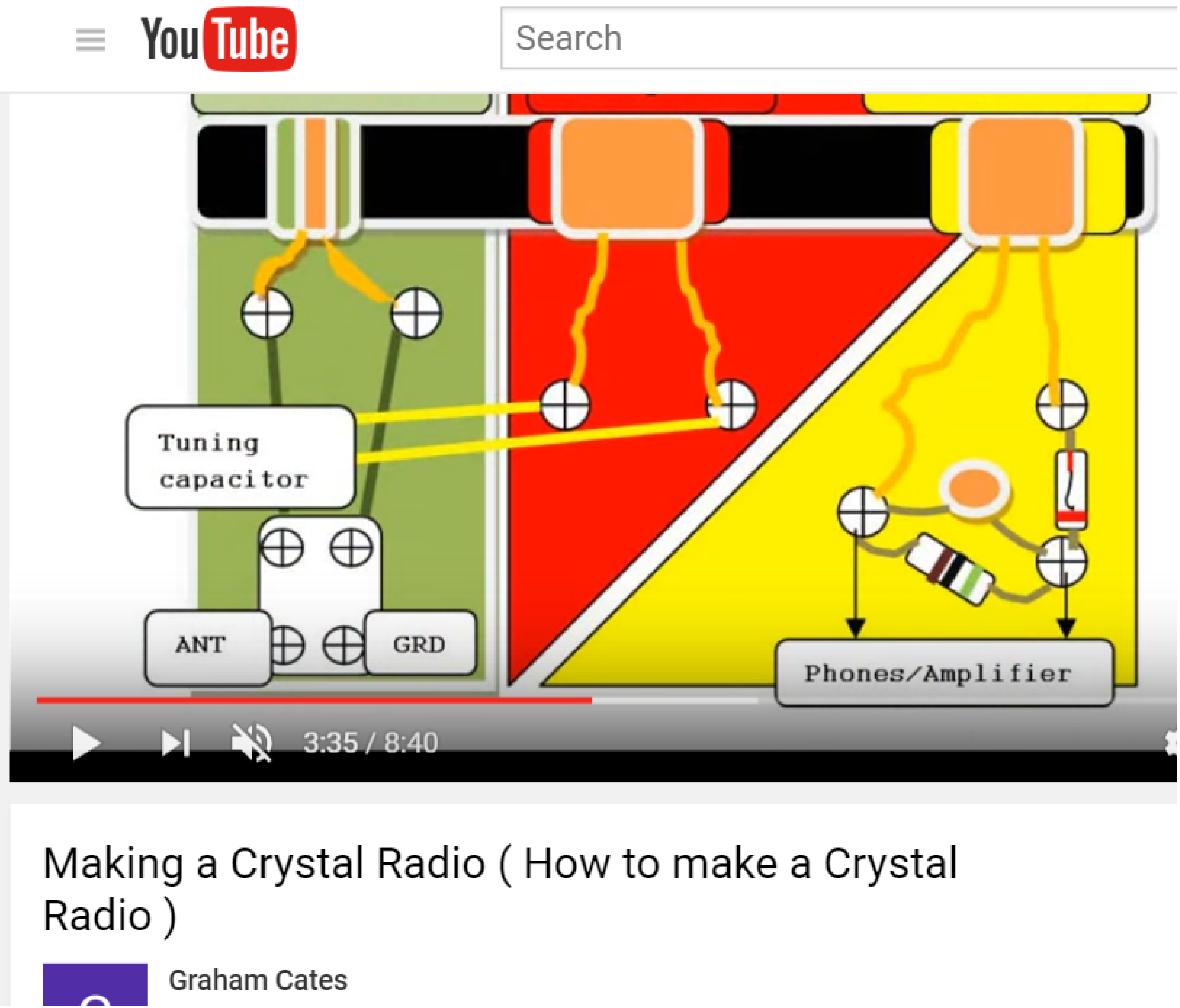
Search
Tuning capacitor
ANT
GRD
Phones/Amplifier
3:35 / 8:40

Making a Crystal Radio (How to make a Crystal Radio)

Graham Cates

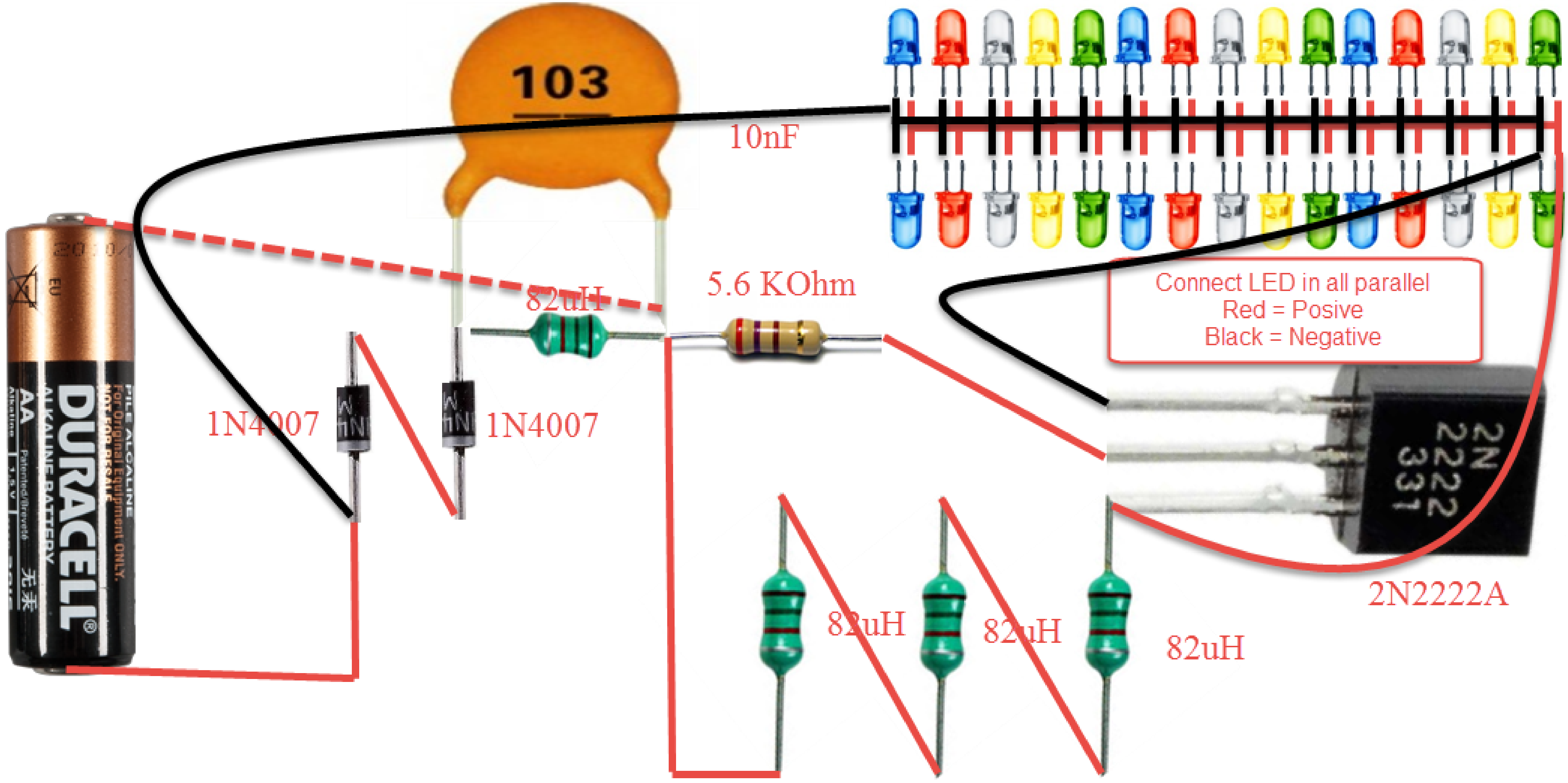
103
10nF
82uH
5.6 KOhm
1N4007
1N4007
Connect LED in all parallel
Red = Posive
Black = Negative
DURACELL
ALKALINE BATTERY
AA
2N
2222
331
2N2222A
82uH
82uH
82uH

The joule thief circuit.

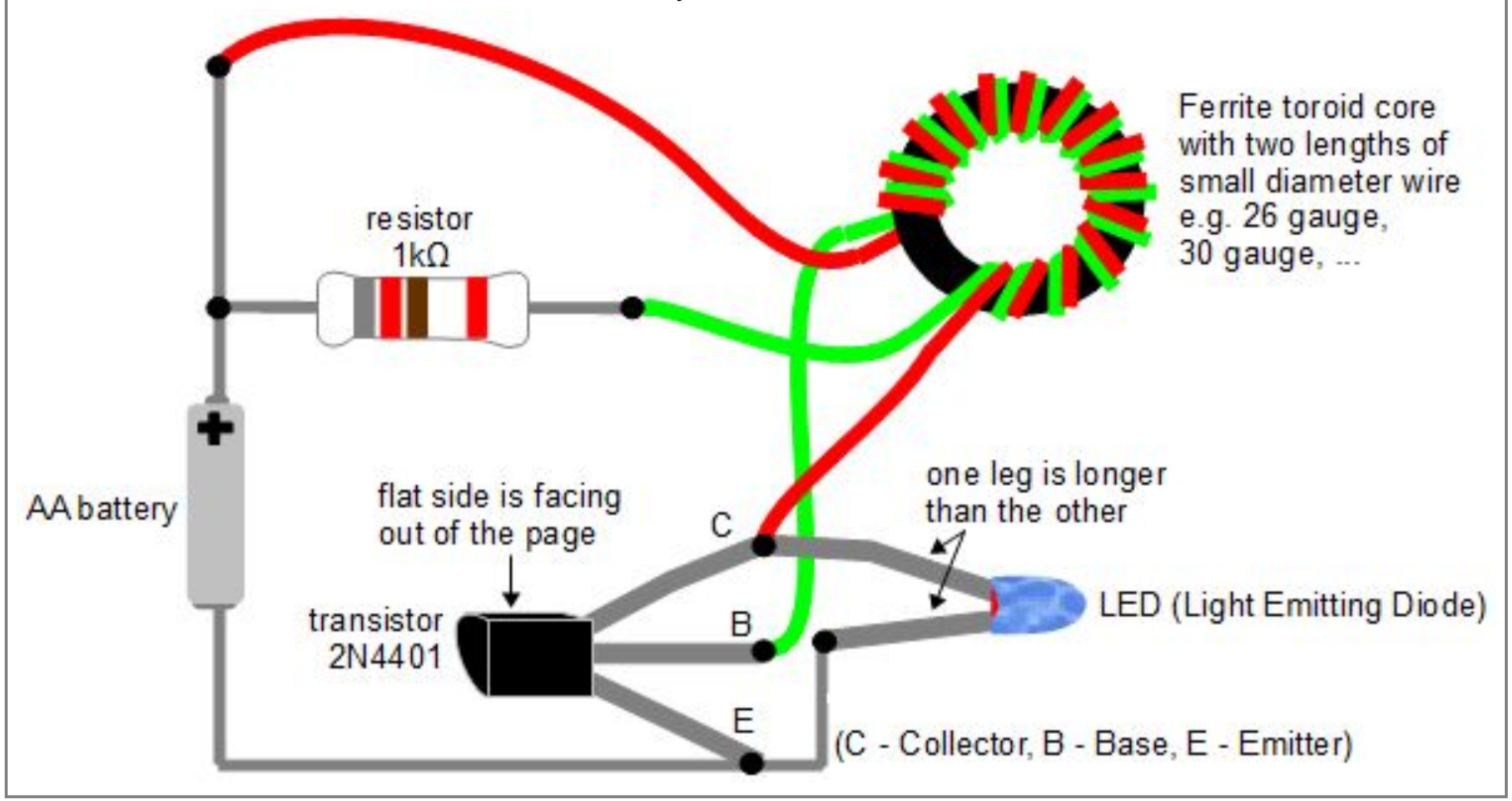
Ferrite toroid core
with two lengths of
small diameter wire
e.g. 26 gauge,
30 gauge, ...
resistor
1kΩ
AA battery
flat side is facing
out of the page
transistor
2N4401
C
B
E
one leg is longer
than the other
LED (Light Emitting Diode)
(C - Collector, B - Base, E - Emitter)

Transistor - The legs of the transistor can be determined by noticing that there's a flat side to the transistor case. See the diagram above. A large number of transistors have been reported to work: 2N4401, NET123AP, BC547B, 2SC2500, BC337, PN2222, to name just a few.

LED - One leg of the LED is longer than the other leg. Use this to determine which one goes where. See the diagram above.

Resistor - The diagram says use a 1 kilo ohm resistor but I've used an 820 ohm one just fine. I've also seen a 2 kilo ohm one in use. Use whatever works for you. You can also use a potentiometer (a variable resistor) so that you can easily adjust it to select the resistance that gives the best light.

Toroid ferrite core - Some people have gotten these by opening up compact fluorescent lightbulbs (CFLs). I took mine out of some device whose original function I don't know. To get it working, my first one had just 13 turns for each wire and I used a 30 gauge wire and a 26 gauge wire. The wire must be insulated. A variety of number of turns will work. This is something you can play with. Look at the diagram carefully to determine where the wires connect to.

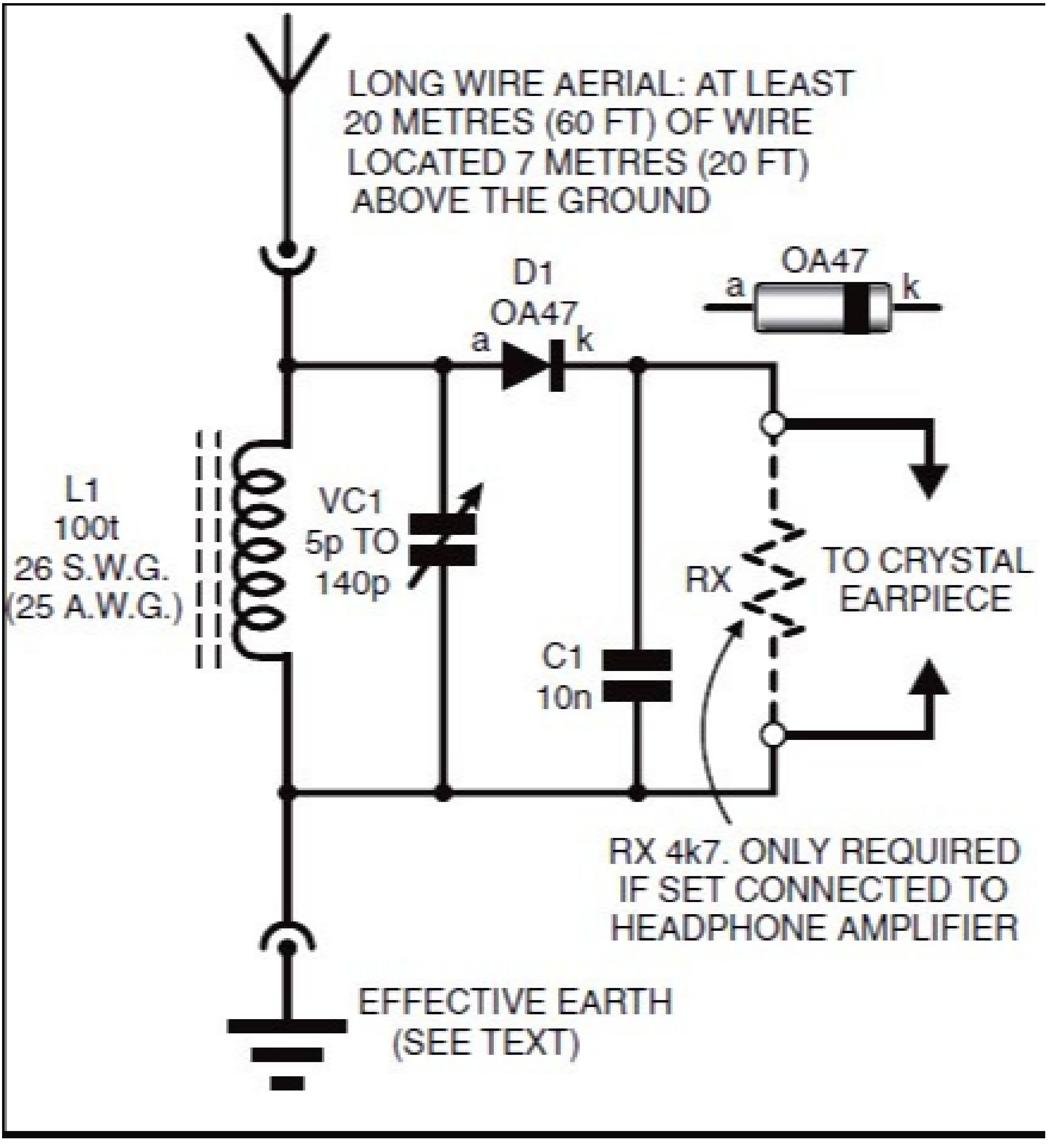
LONG WIRE AERIAL: AT LEAST
20 METRES (60 FT) OF WIRE
LOCATED 7 METRES (20 FT)
ABOVE THE GROUND
D1
OA47
a
k
OA47
a
k
L1
100t
26 S.W.G.
(25 A.W.G.)
VC1
5p TO
140p
C1
10n
RX
TO CRYSTAL
EARPIECE
RX 4k7. ONLY REQUIRED
IF SET CONNECTED TO
HEADPHONE AMPLIFIER
EFFECTIVE EARTH
(SEE TEXT)

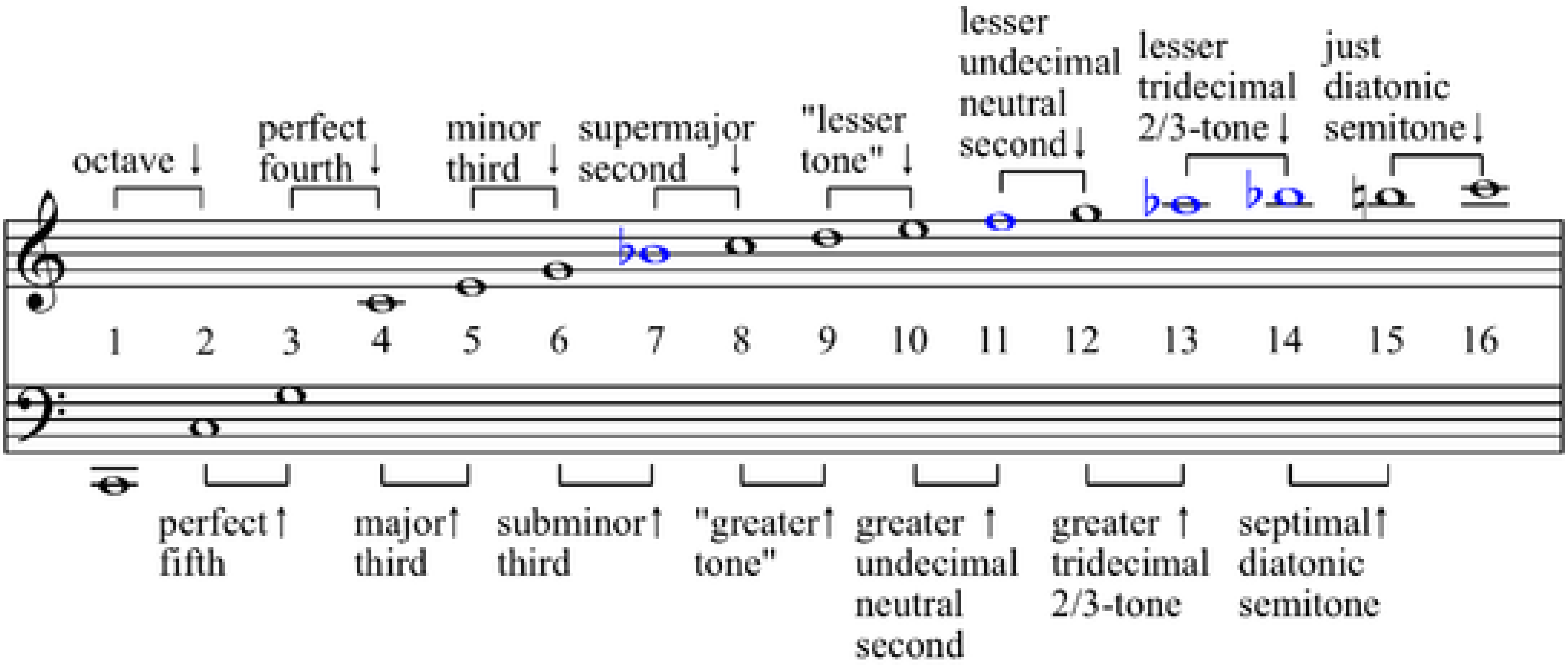
octave ↓
perfect fourth ↓
minor third ↓
supermajor second ↓
"lesser tone" ↓
lesser undecimal neutral second ↓
lesser tridecimal 2/3-tone ↓
just diatonic semitone ↓
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
perfect fifth ↑
major third ↑
subminor third ↑
"greater tone" ↑
greater undecimal neutral second ↑
greater tridecimal 2/3-tone ↑
septimal diatonic semitone ↑

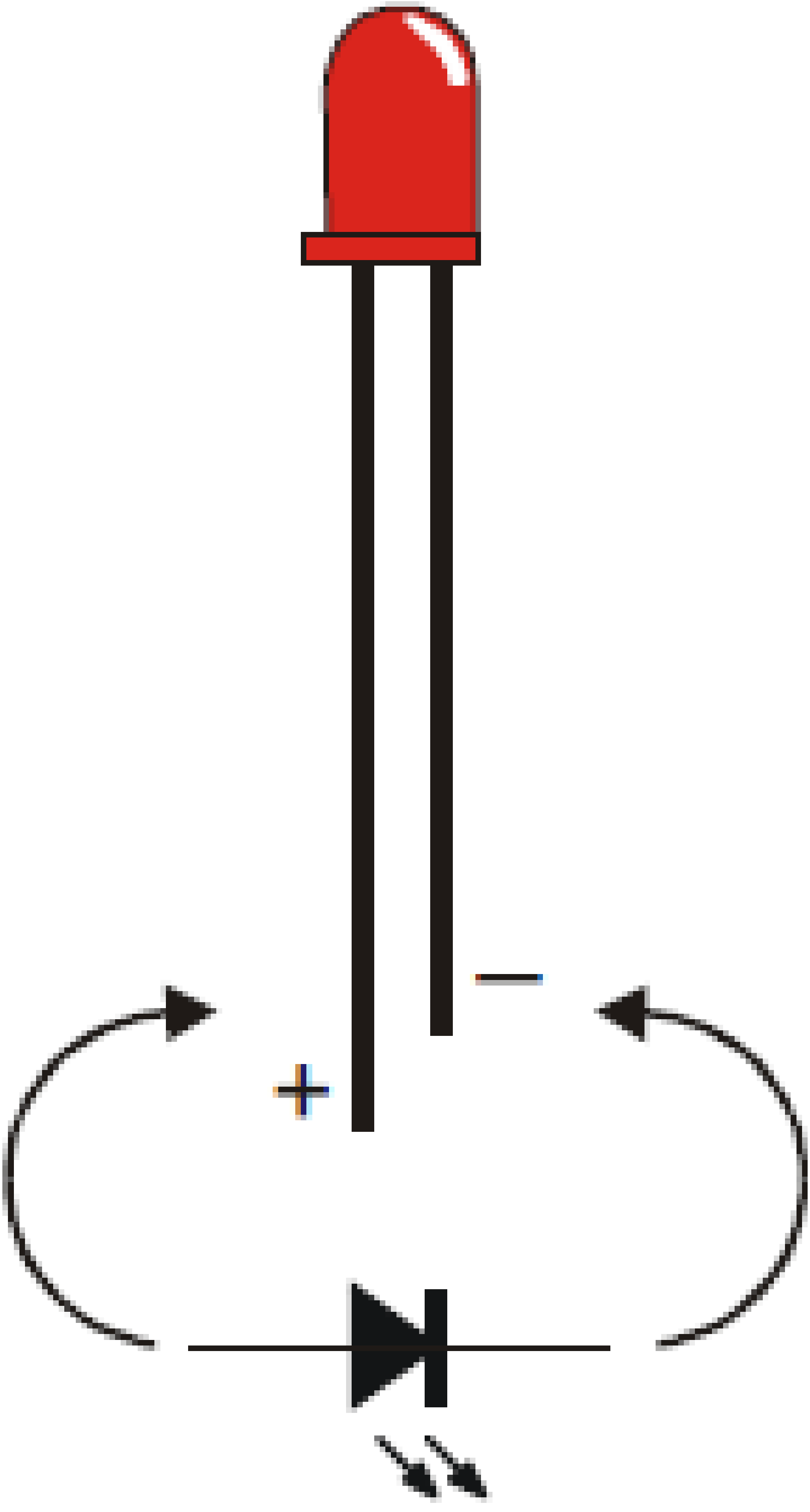

LED

The Tesla Experiment

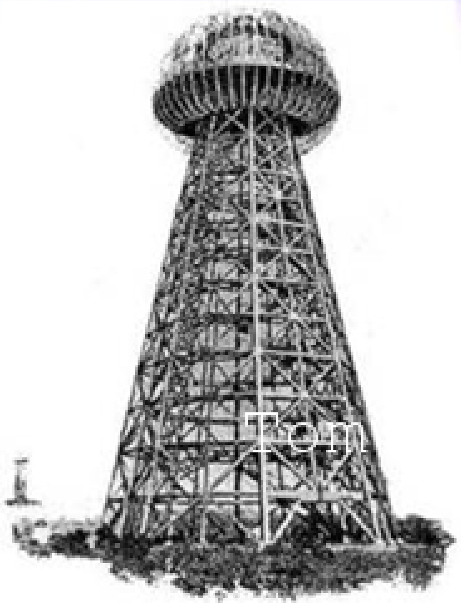

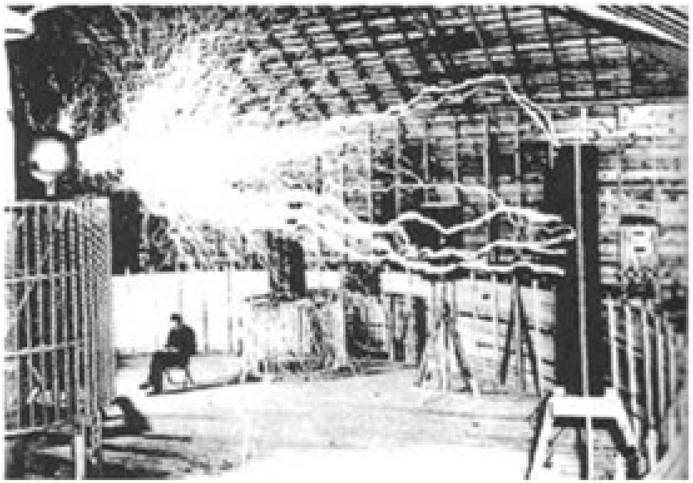

TRANSVERSE WAVE INTERFERENCE

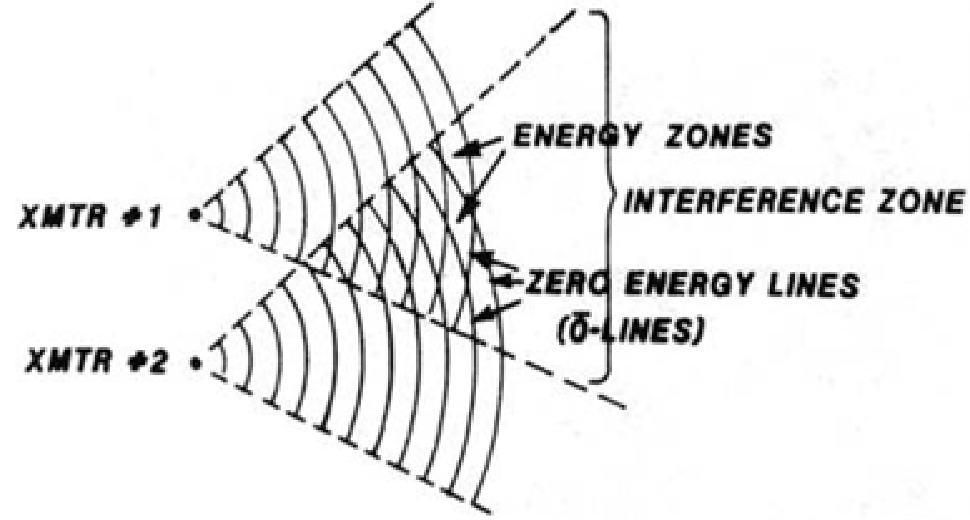
ENERGY ZONES
INTERFERENCE ZONE
XMTR #1
ZERO ENERGY LINES
(0̄-LINES)
XMTR #2

LONGITUDINAL WAVE INTERFERENCE
(SCALAR WAVES)

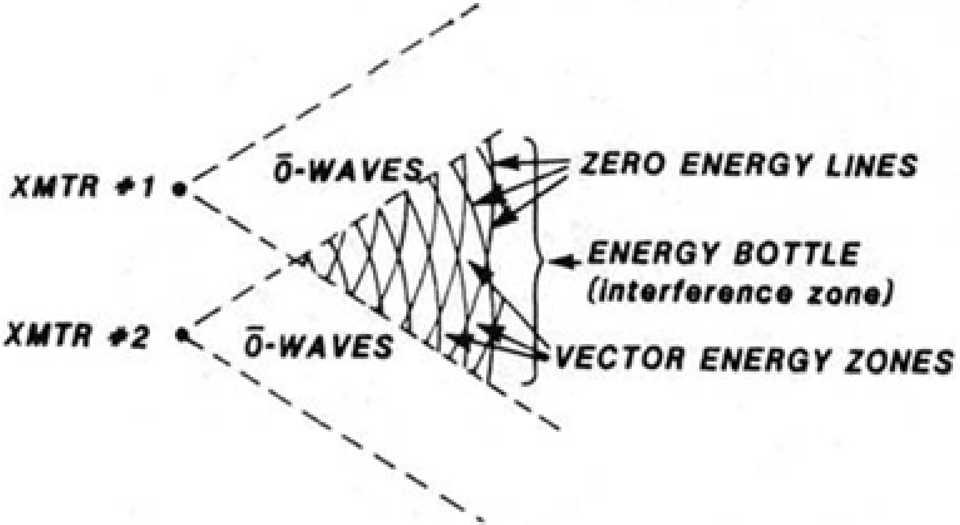
XMTR #1
0̄-WAVES
ZERO ENERGY LINES
ENERGY BOTTLE
(interference zone)
XMTR #2
0̄-WAVES
VECTOR ENERGY ZONES

© PoweredTemplates.com

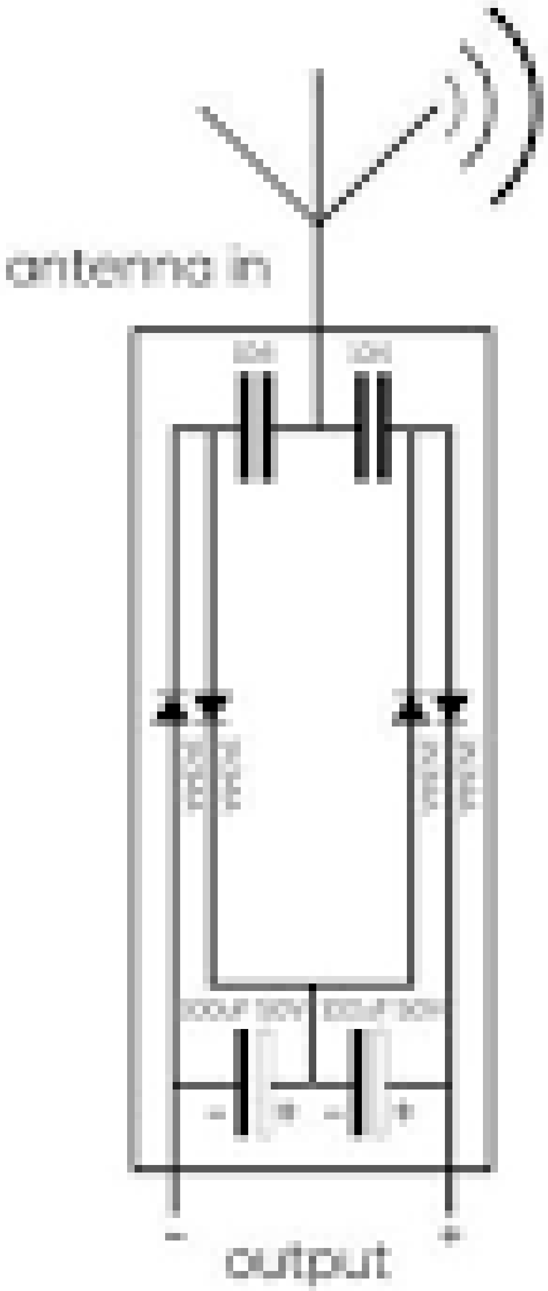
antenna in
104
104
100uF 50V
100uF 50V
-
+
-
+
-
output
+

RF to DC

Circuit Diagam (simplified)

Component list:
(2) Ceramic Capacitors (104)
(4) Germanium Diodes (1N34A)
(2) Electrolytic Capacitors (100uF 50V)

Copyright 2008 Drew Paul Designs. All rights reserved.

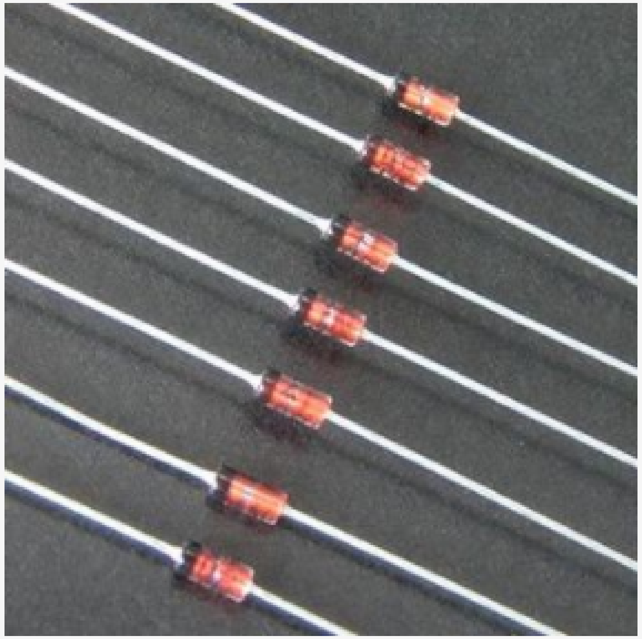

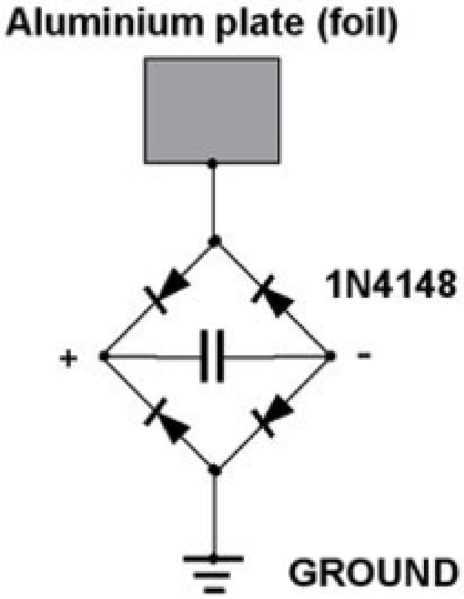
Aluminium plate (foil)
1N4148
+
-
GROUND

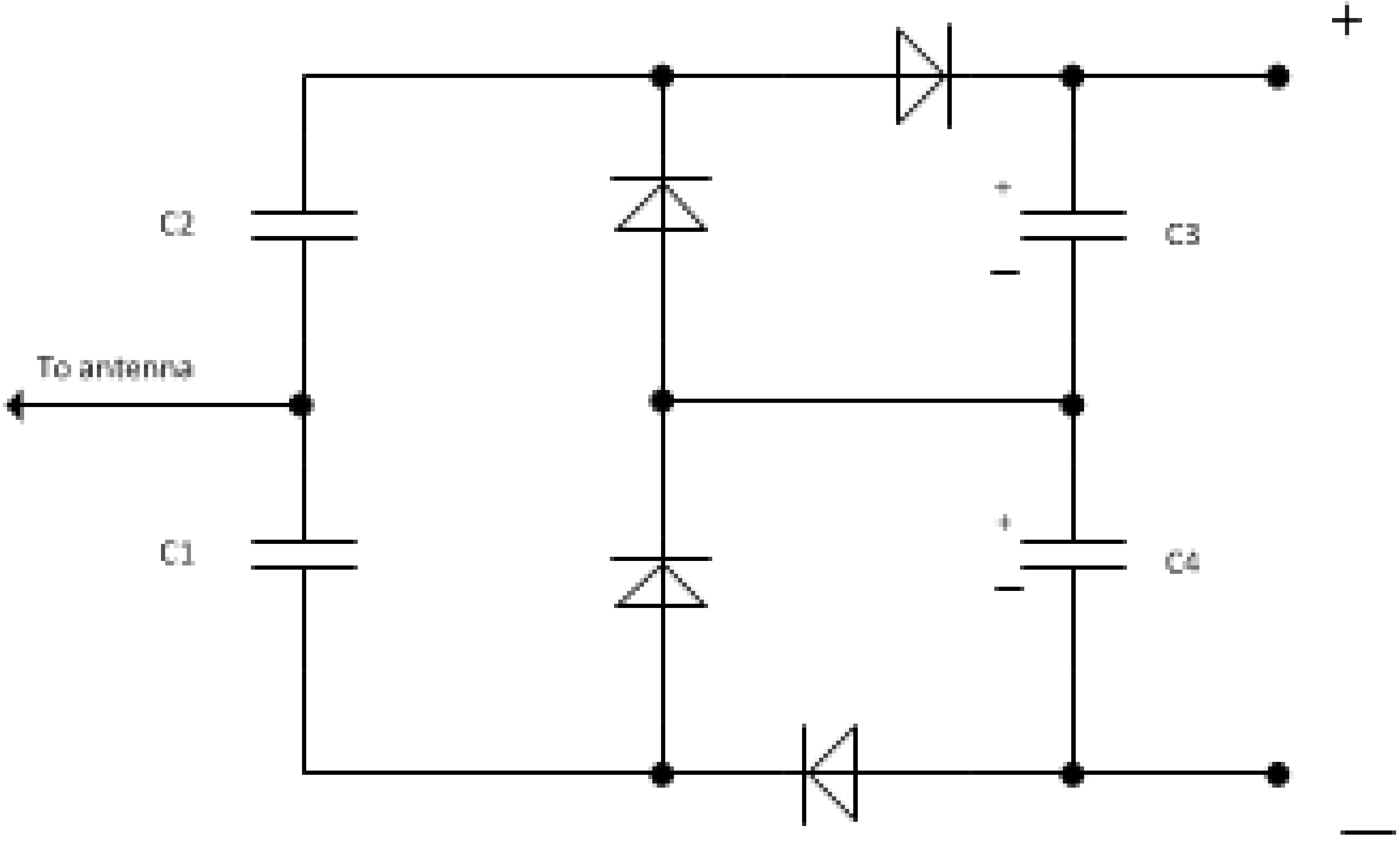
+
C2
C3
To antenna
C1
C4
−

Silent Sound Subliminal Mind Control

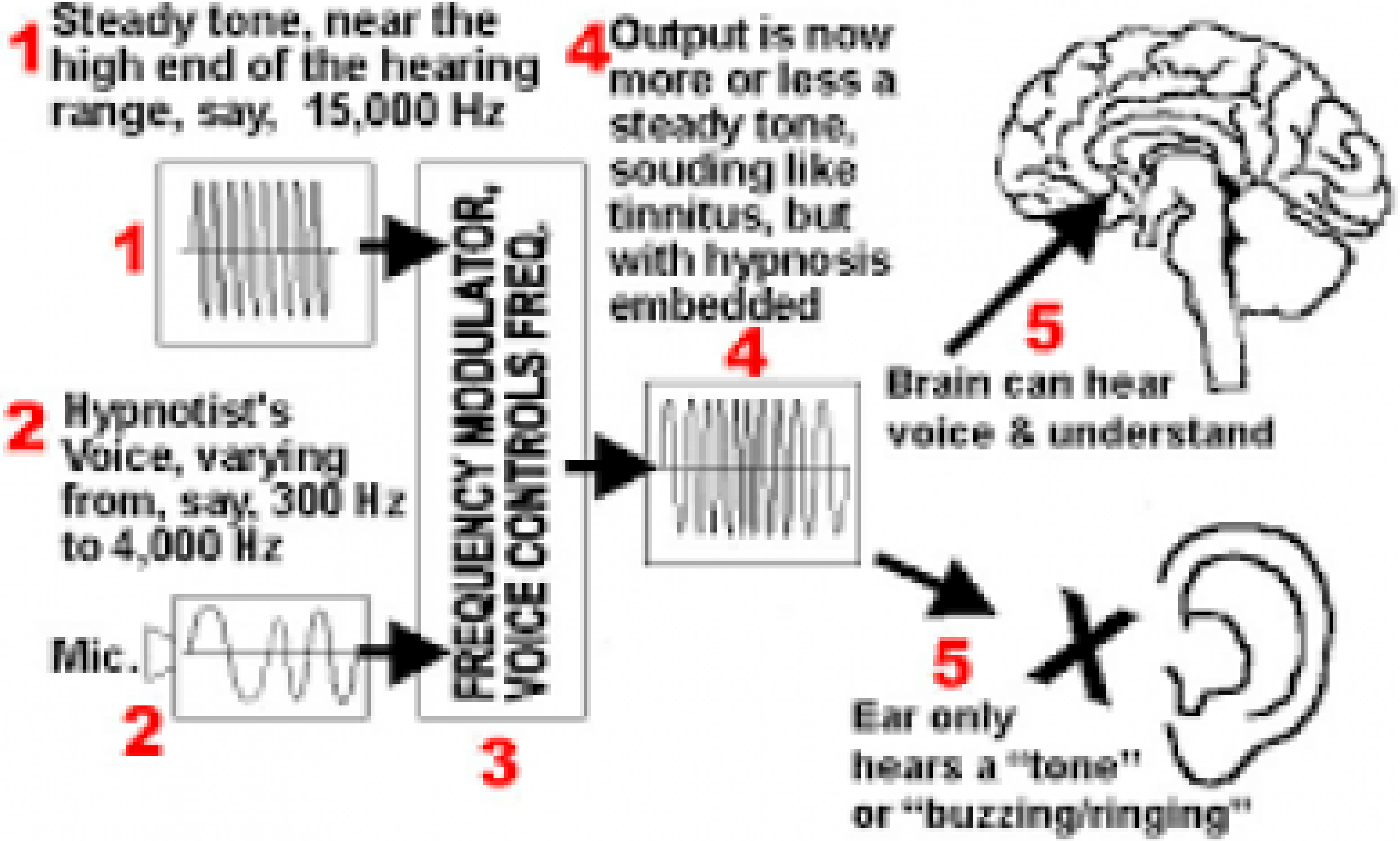
1
2
Mic.
FREQUENCY MODULATOR,
VOICE CONTROLS FREQ.
3
4
5
5

1 Steady tone, near the high end of the hearing range, say, 15,000 Hz

2 Hypnotist's Voice, varying from, say, 300 Hz to 4,000 Hz

4 Output is now more or less a steady tone, souding like tinnitus, but with hypnosis embedded

5 Brain can hear voice & understand

5 Ear only hears a "tone" or "buzzing/ringing"

Output can be via open air broadcast or piggybacked on TV/radio signals

U.S. Patent 5159703 issued Oct 27, 1992

Table 1: Dinshah's Sound/Color Equivalent Chart.

Color	Frequency	Musical Note	Note Frequency
Red	397.5	G	378 (3 x 126 = 378)
Orange	430.5	A	432 (3 x 144 = 432)
Yellow	464.5	A#	459 (3 x 153 = 459)
Lemon	498	B	486 (3 x 162 = 486)
Green	531.5	C	513 (3 x 171 = 513)
Turquoise	565	C#	540 (3 x 180 = 540)
Blue	598.5	D	567 (3 x 189 = 567)
Indigo	632	D#	594 (3 x 198 = 594)
Violet	665.5	E	648 (3 x 216 = 648)
Purple	565*	A# and E	562*
Magenta	531.5*	G and E	525*
Scarlet	498*	G# and D	501*

---

*denotes reverse polarity. From:

Dinshah GP. Spectralchrometry Encyclopedia, 3rd Ed.

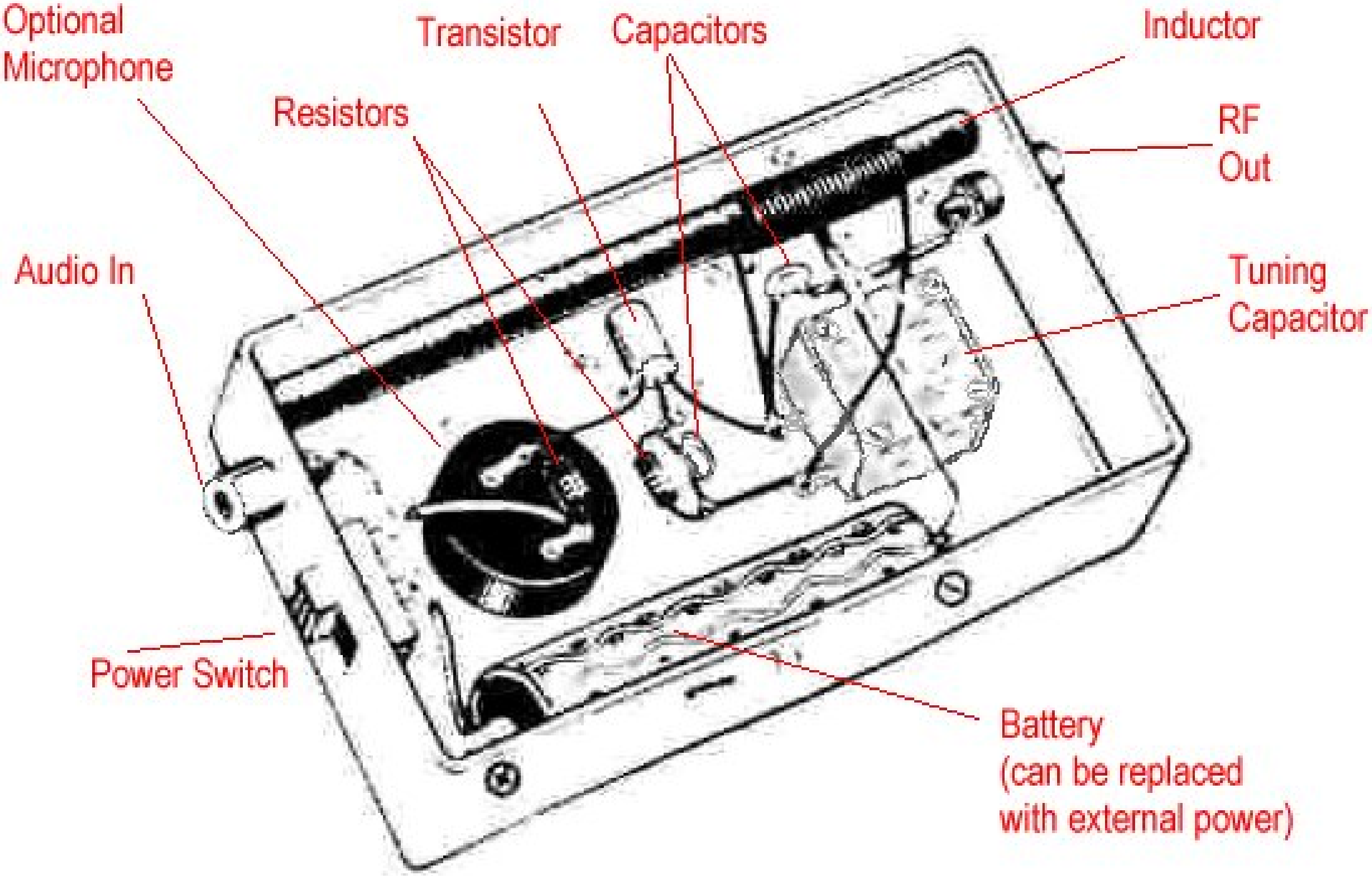
Optional Microphone
Transistor
Capacitors
Inductor
Resistors
RF Out
Audio In
Tuning Capacitor
Power Switch
Battery (can be replaced with external power)

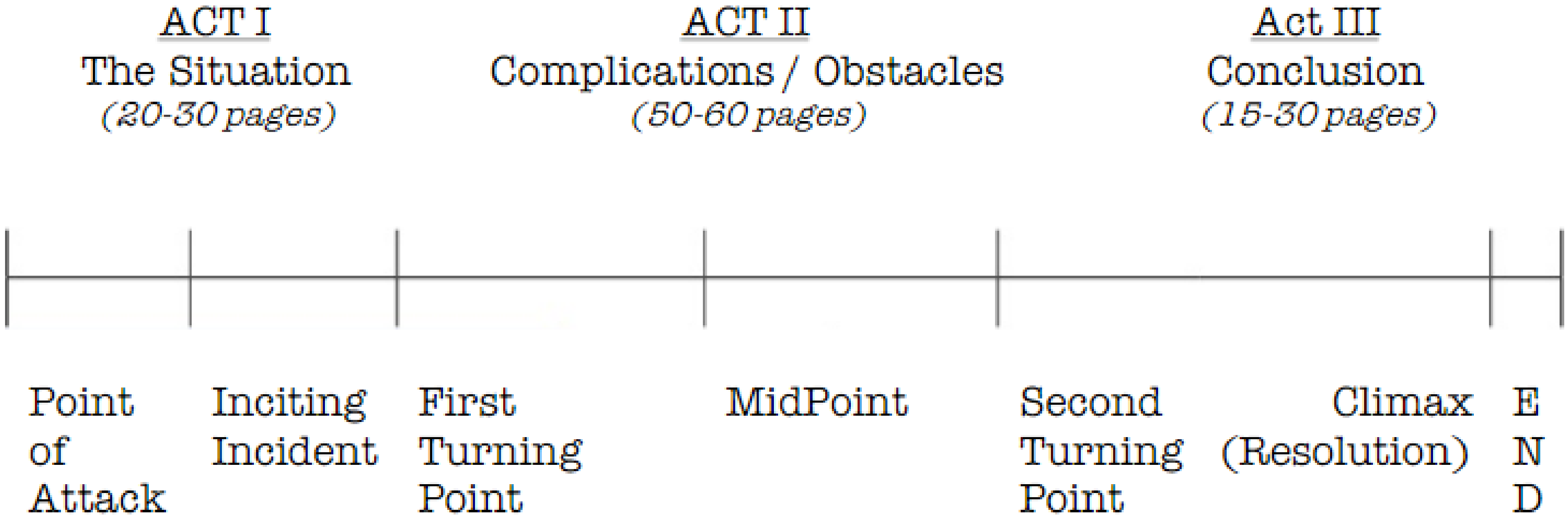

ACT I
The Situation
(20-30 pages)

ACT II
Complications / Obstacles
(50-60 pages)

Act III
Conclusion
(15-30 pages)

Point of Attack

Inciting Incident

First Turning Point

MidPoint

Second Turning Point

Climax (Resolution)

E
N
D

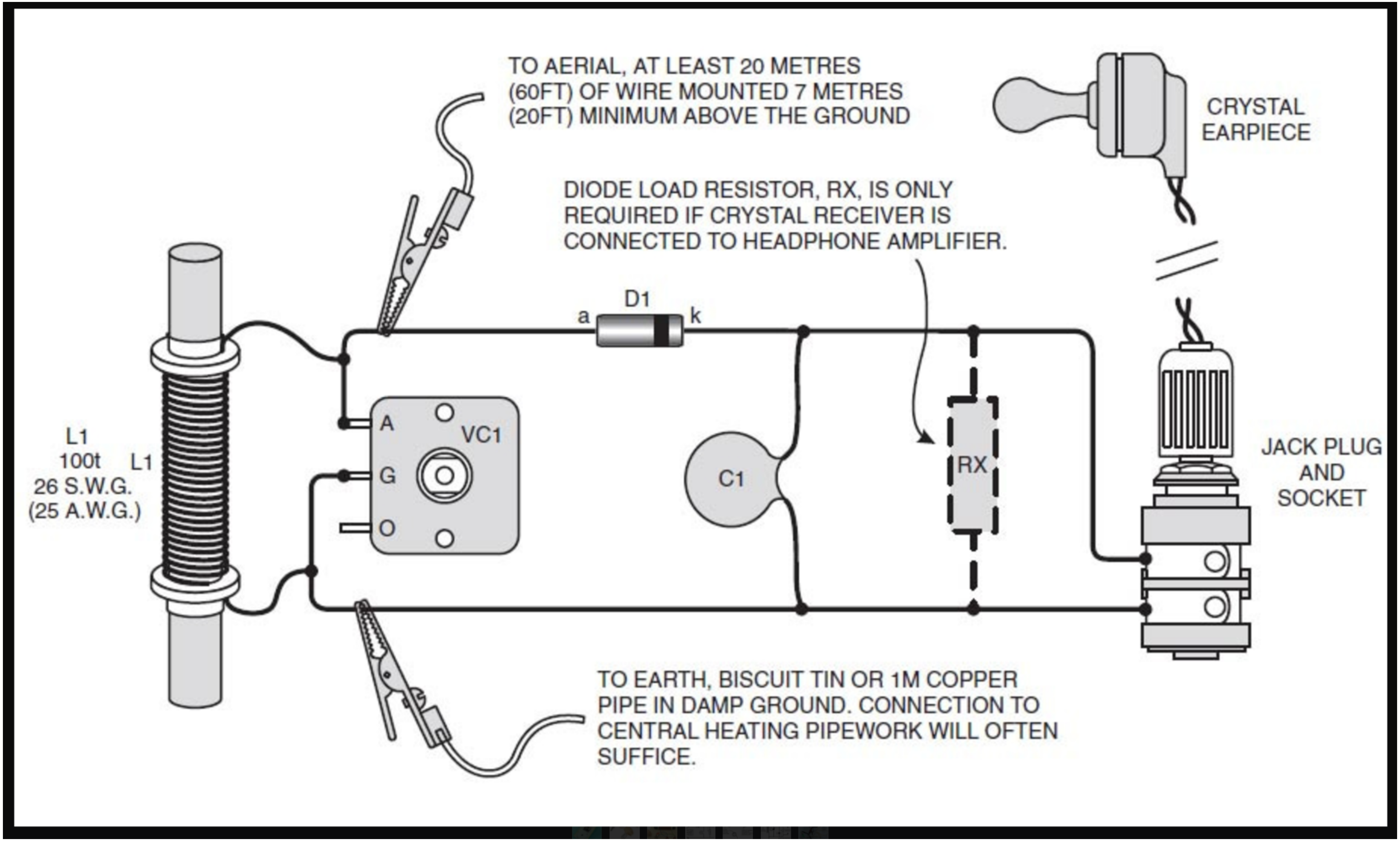
TO AERIAL, AT LEAST 20 METRES (60FT) OF WIRE MOUNTED 7 METRES (20FT) MINIMUM ABOVE THE GROUND
CRYSTAL EARPIECE
DIODE LOAD RESISTOR, RX, IS ONLY REQUIRED IF CRYSTAL RECEIVER IS CONNECTED TO HEADPHONE AMPLIFIER.
D1
a
k
L1
100t
26 S.W.G.
(25 A.W.G.)
L1
A
VC1
G
O
C1
RX
JACK PLUG AND SOCKET
TO EARTH, BISCUIT TIN OR 1M COPPER PIPE IN DAMP GROUND. CONNECTION TO CENTRAL HEATING PIPEWORK WILL OFTEN SUFFICE.

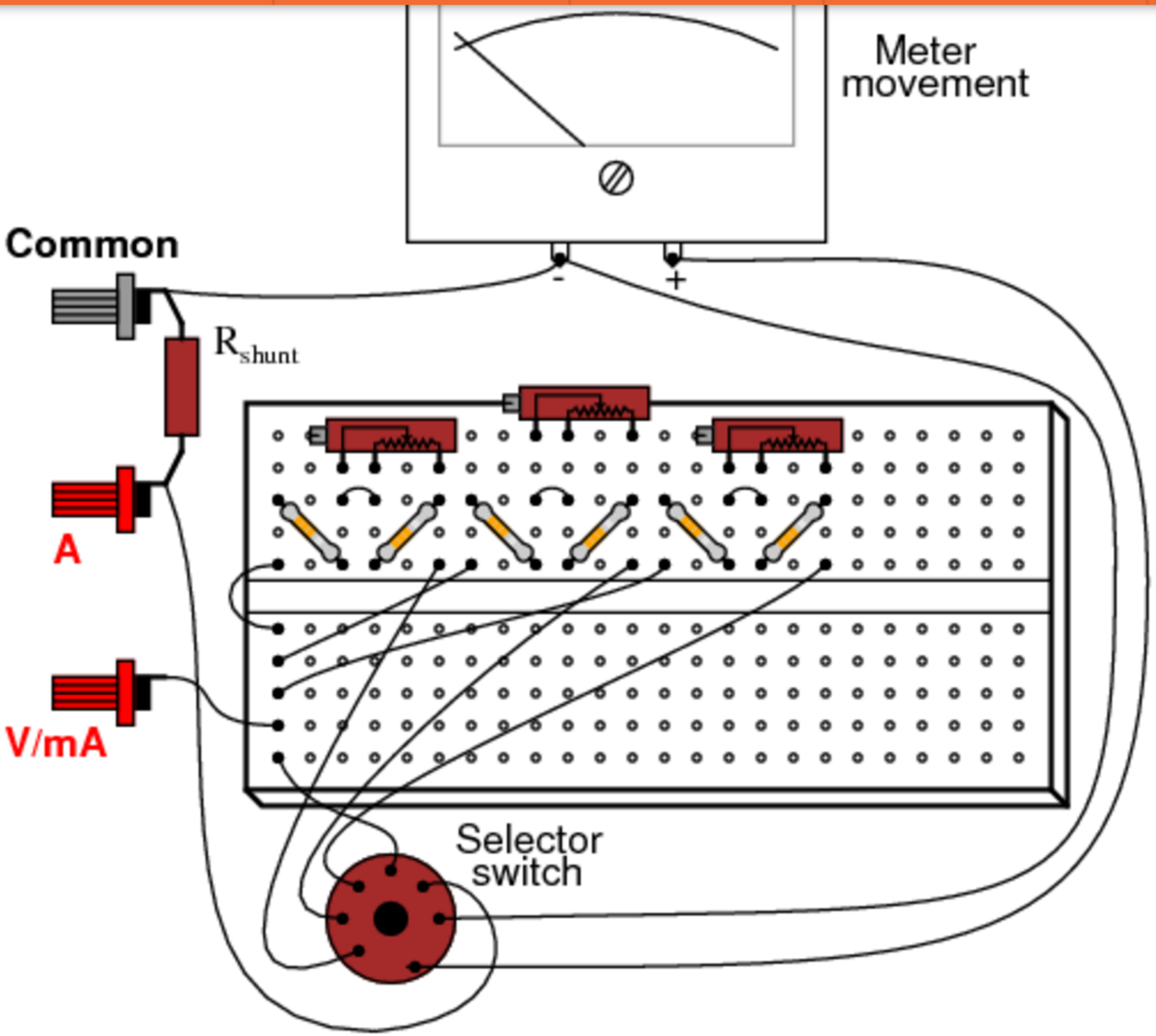
Meter
movement
-
+
Common
R_{shunt}
A
V/mA
Selector
switch

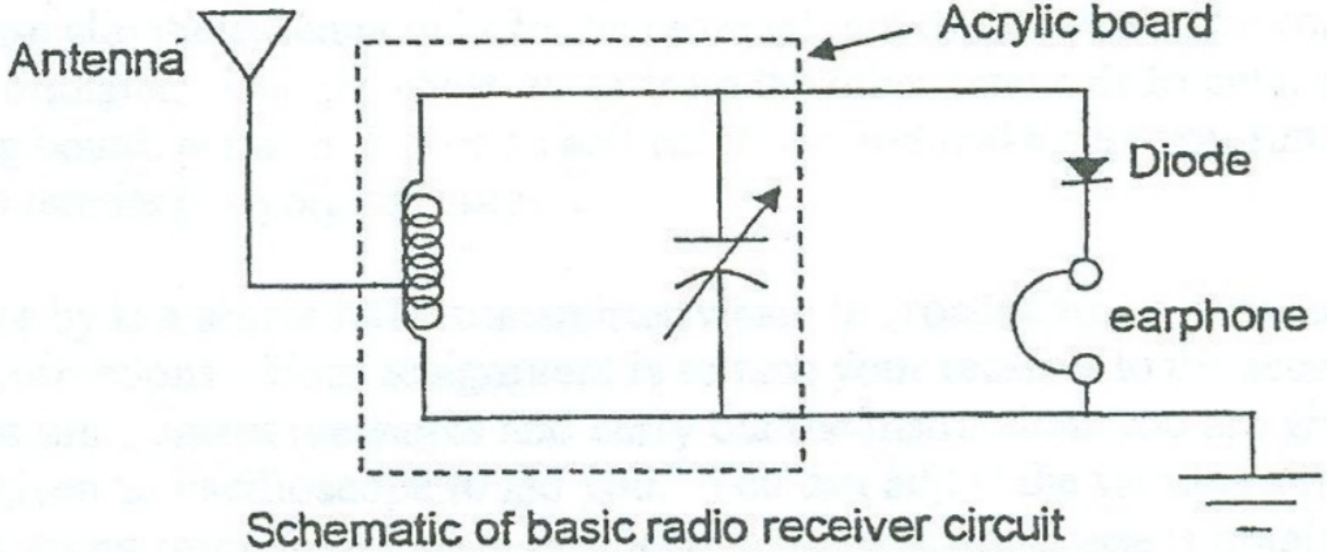
Antenna
Acrylic board
Diode
earphone

Schematic of basic radio receiver circuit

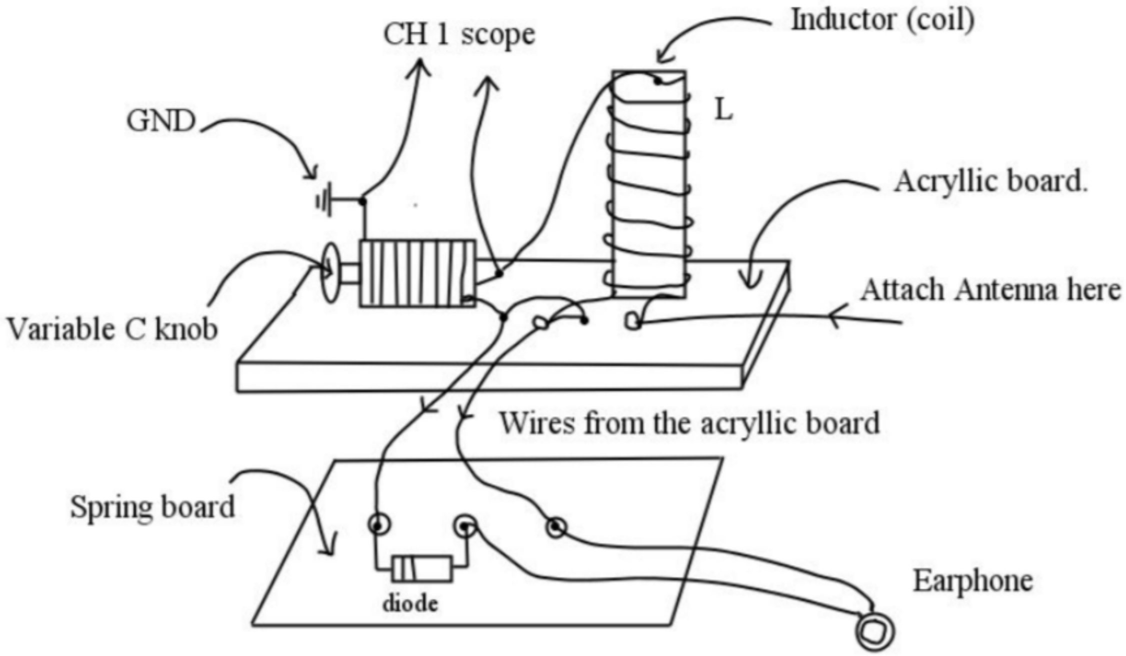
CH 1 scope
Inductor (coil)
GND
L
Acryllic board.
Attach Antenna here
Variable C knob
Wires from the acryllic board
Spring board
diode
Earphone

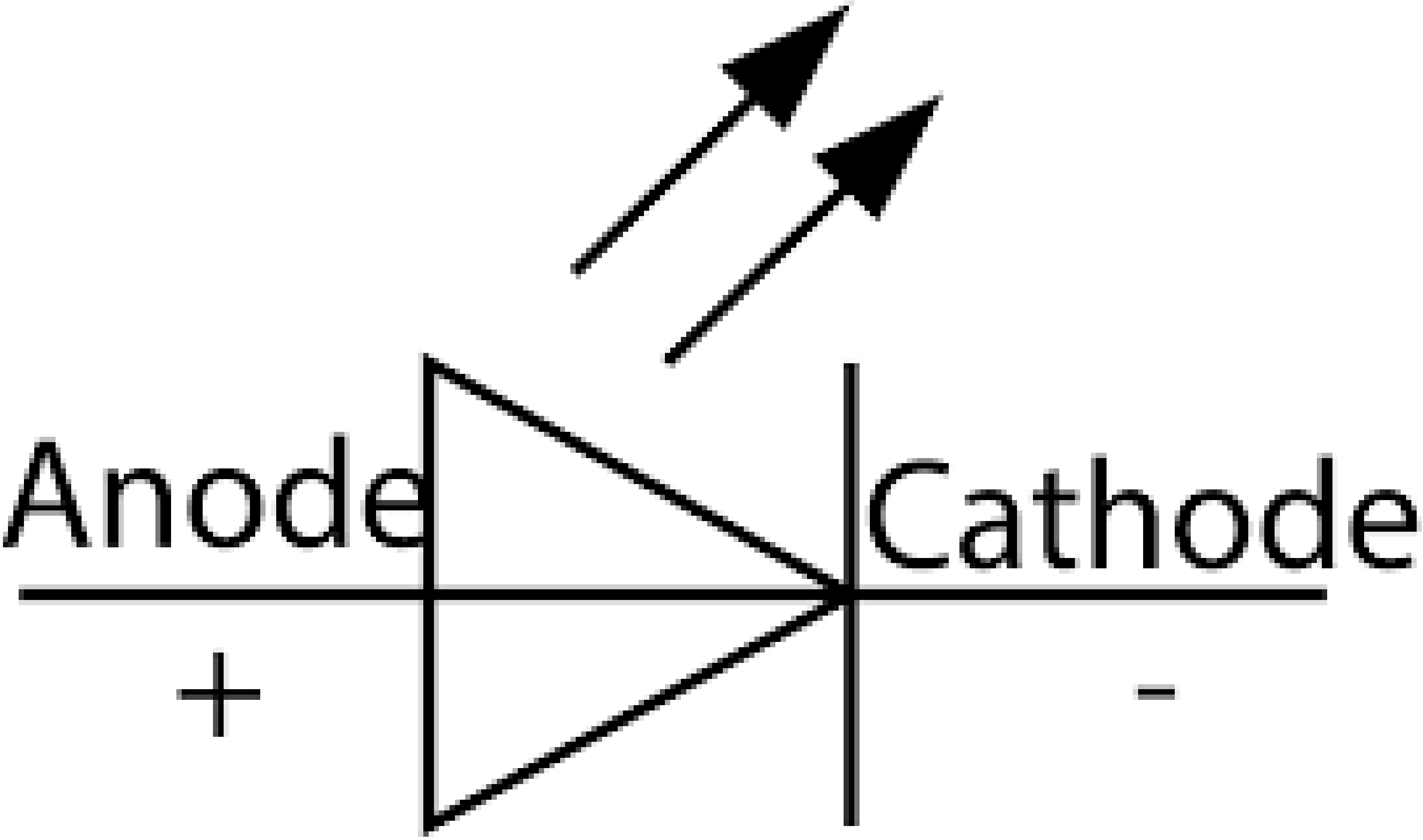
Anode
+
Cathode
-

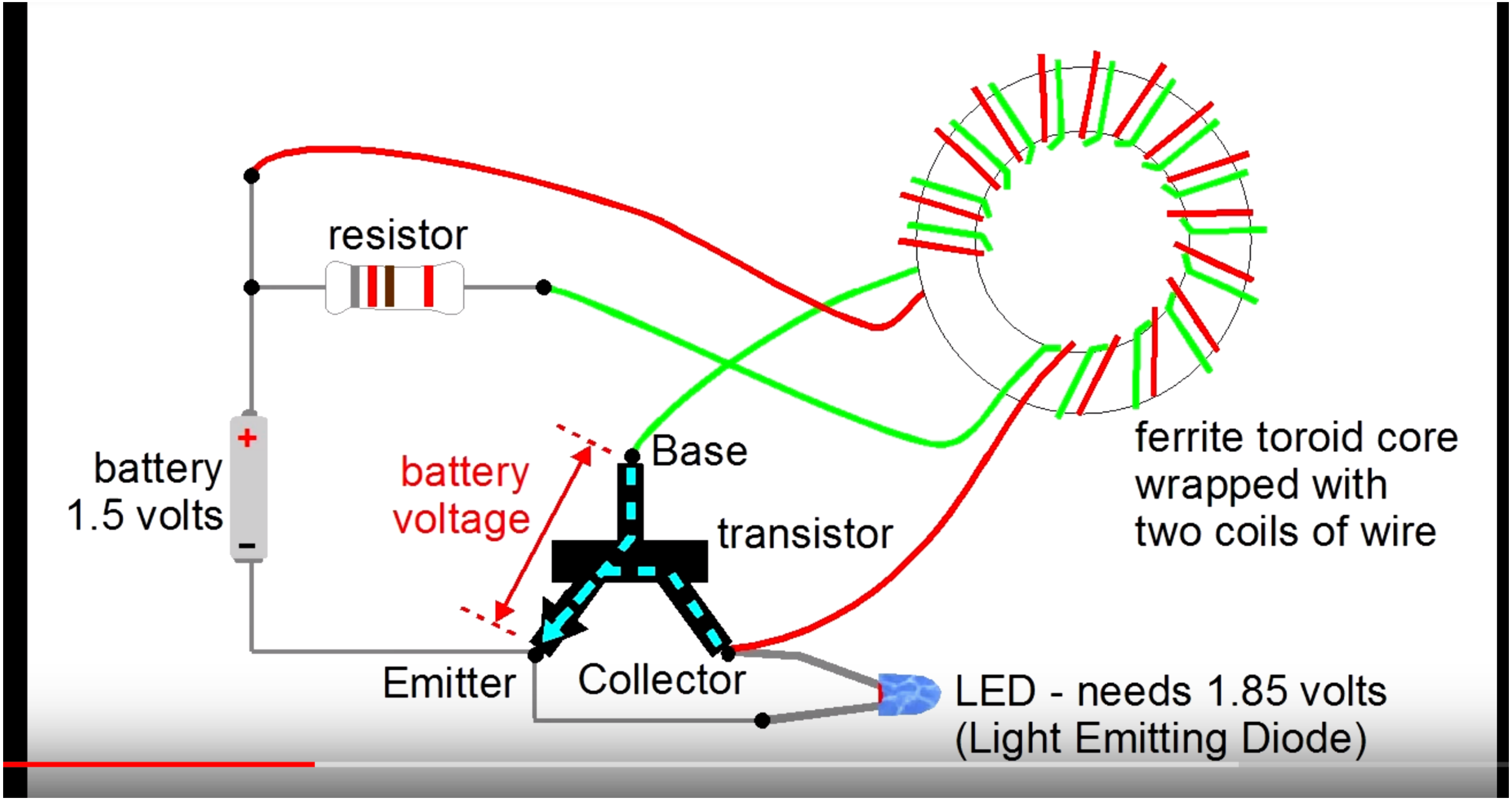
resistor
battery
1.5 volts
+
-
battery
voltage
Base
transistor
Emitter
Collector
ferrite toroid core
wrapped with
two coils of wire
LED - needs 1.85 volts
(Light Emitting Diode)

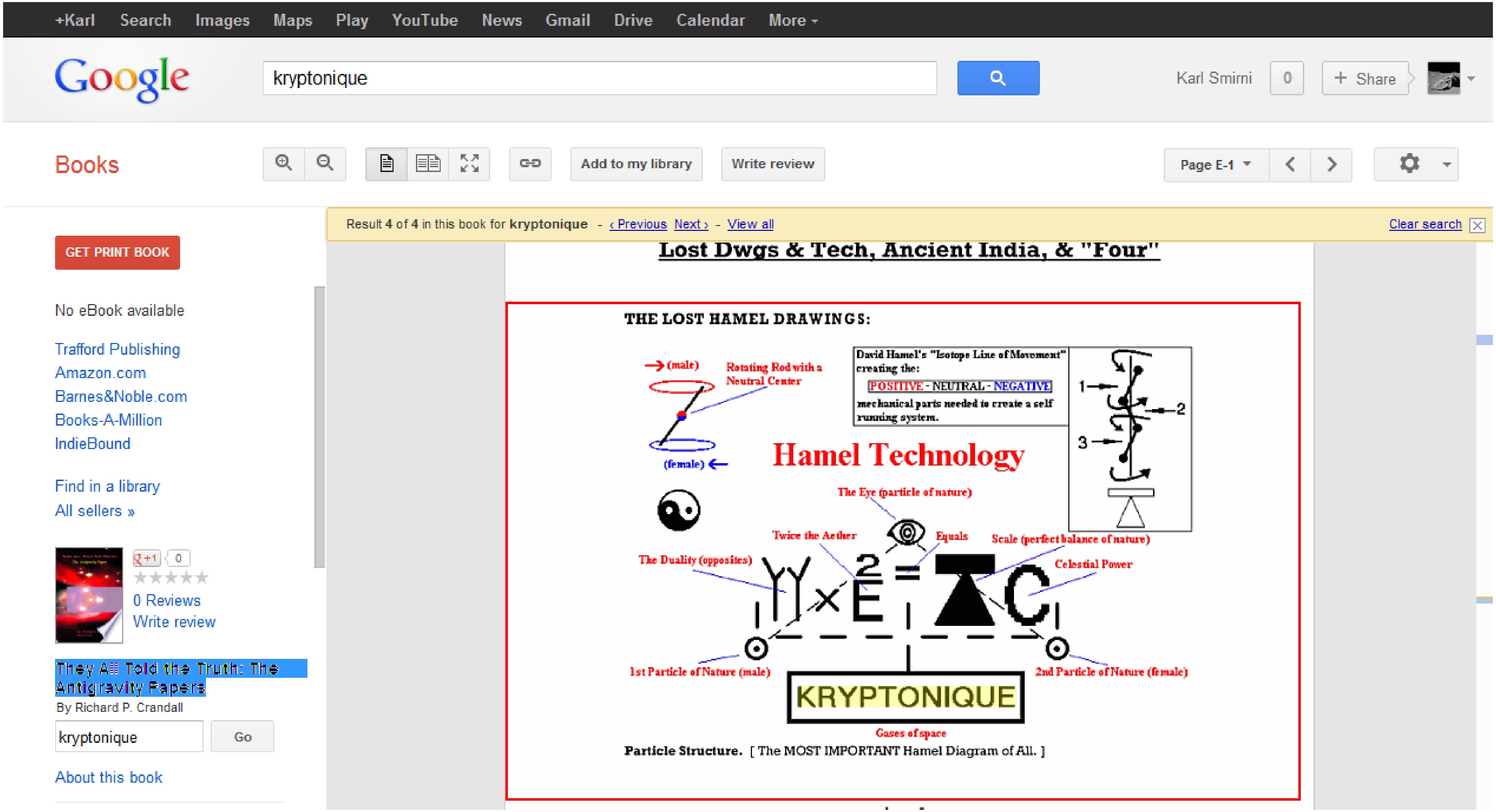
+Karl
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
kryptonique
Karl Smirni
0
Share
Books
Add to my library
Write review
Page E-1
GET PRINT BOOK
No eBook available
Trafford Publishing
Amazon.com
Barnes&Noble.com
Books-A-Million
IndieBound
Find in a library
All sellers »
+1
0
0 Reviews
Write review
kryptonique
Go
About this book
Clear search

Result 4 of 4 in this book for **kryptonique** - ‹ Previous Next › - View all

They All Told the Truth: The Antigravity Papers
By Richard P. Crandall

Lost Dwgs & Tech, Ancient India, & "Four"

THE LOST HAMEL DRAWINGS:

David Hamel's "Isotope Line of Movement" creating the:
POSITIVE - NEUTRAL - NEGATIVE
mechanical parts needed to create a self running system.

(male)
Rotating Rod with a Neutral Center
(female)
1
2
3
Hamel Technology
The Eye (particle of nature)
Twice the Aether
Equals
Scale (perfect balance of nature)
The Duality (opposites)
Celestial Power
1st Particle of Nature (male)
2nd Particle of Nature (female)
KRYPTONIQUE
Gases of space

Particle Structure. [The MOST IMPORTANT Hamel Diagram of All.]

- synchronous charging circuit generate more electromagnetic wave

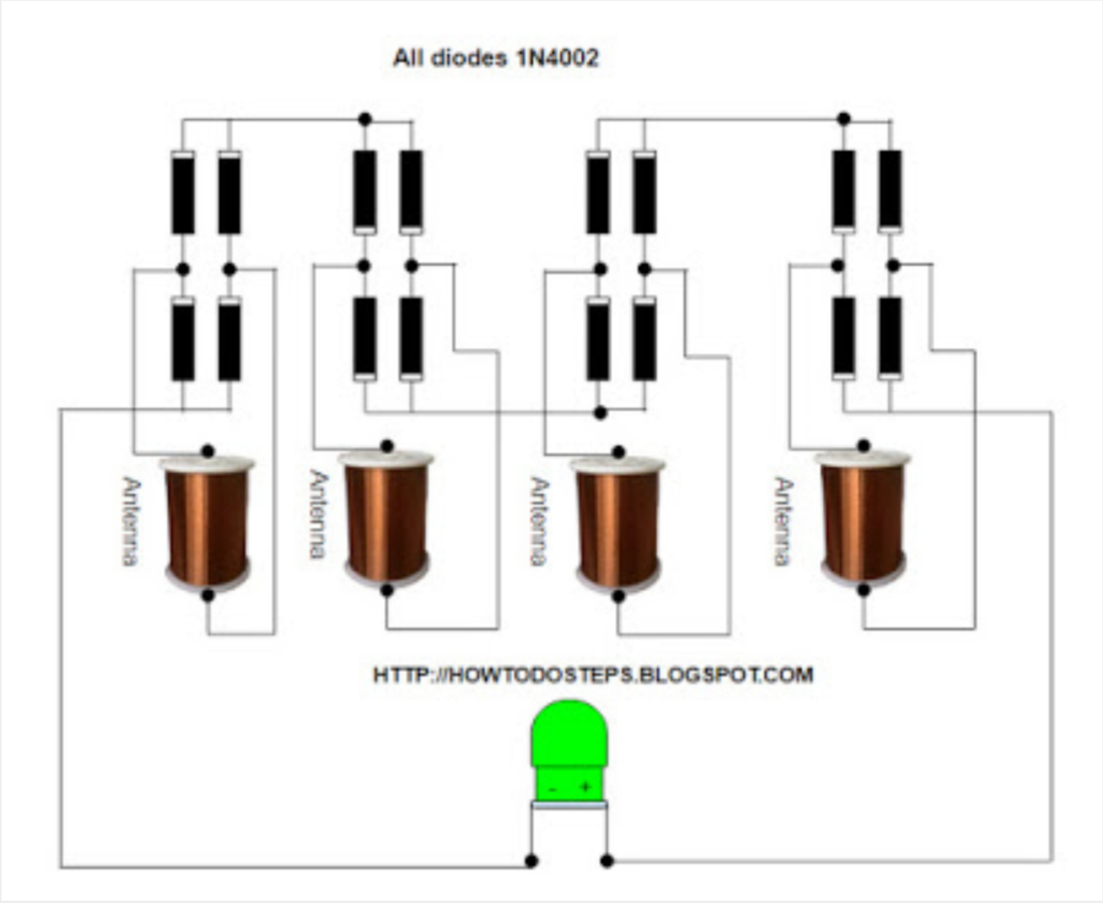
All diodes 1N4002
Antenna
Antenna
Antenna
Antenna
HTTP://HOWTODOSTEPS.BLOGSPOT.COM
-
+

Easy Simple Free Energy Harvesting from Surro...

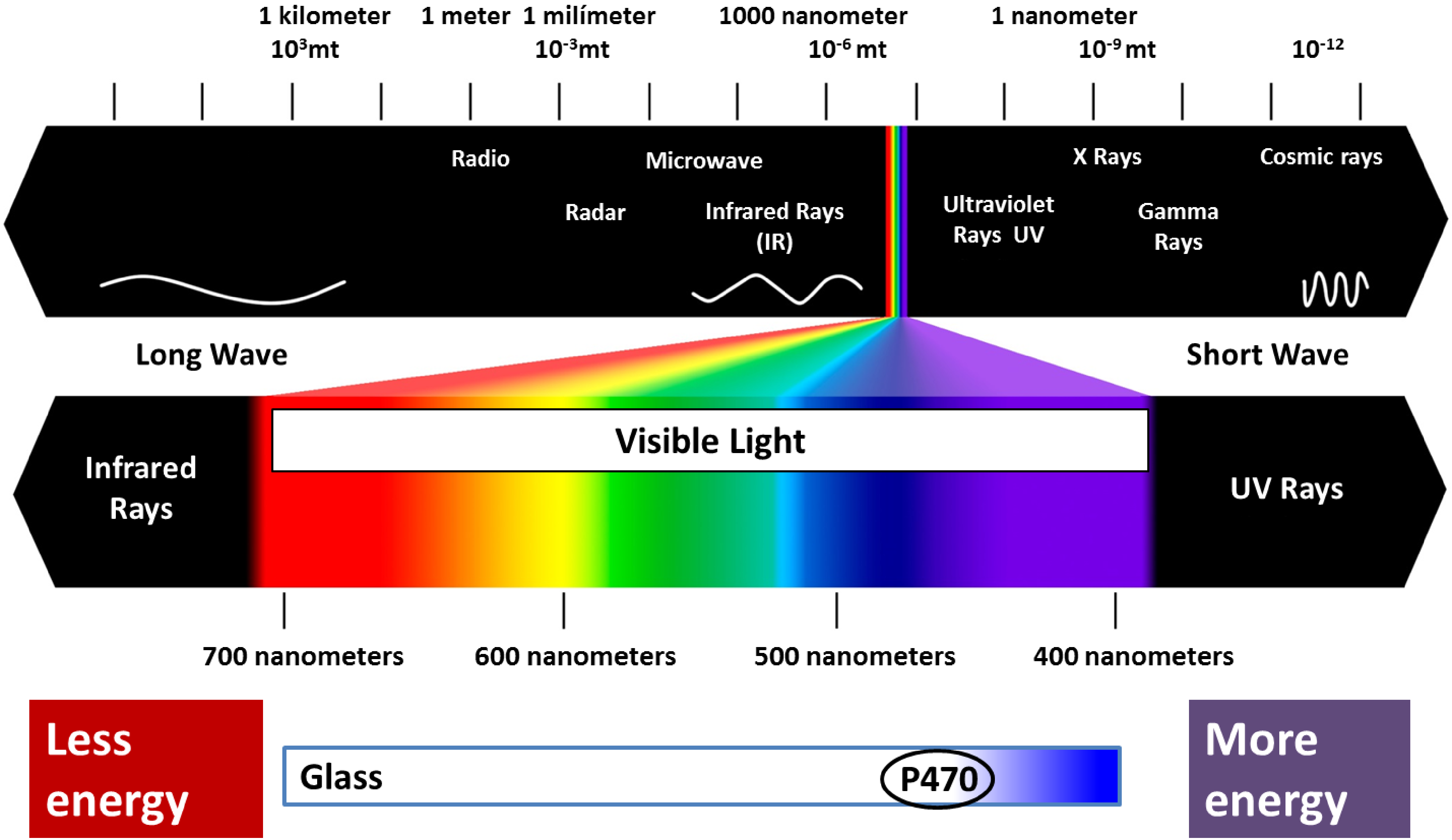
1 kilometer
10^{3}mt
1 meter
1 milímeter
10^{-3}mt
1000 nanometer
10^{-6} mt
1 nanometer
10^{-9} mt
10^{-12}
Radio
Microwave
Radar
Infrared Rays (IR)
Ultraviolet Rays UV
X Rays
Gamma Rays
Cosmic rays
Long Wave
Short Wave
Infrared Rays
Visible Light
UV Rays
700 nanometers
600 nanometers
500 nanometers
400 nanometers
Less energy
Glass
P470
More energy

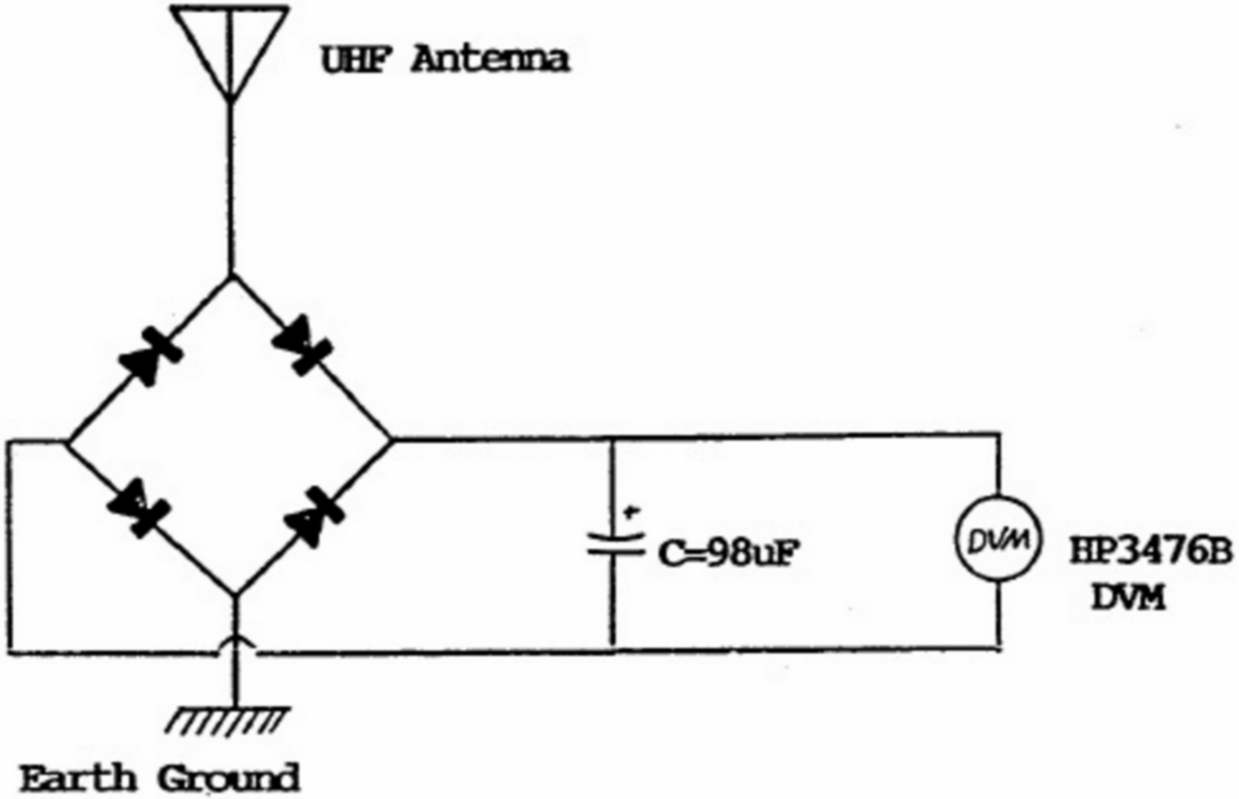
UHF Antenna
+
C=98uF
DVM
HP3476B
DVM
Earth Ground

Figure 2 -- Shows the main components of the FEA : 1) a bridge rectifier, 2) an earth ground connection, 3) an antenna, and 4) a capacitor. The DVM was used for making test measurements.

TAGS

fpv uav osd video camera tx rx

CATEGORIES

- **Accessory**
- **Antenna**
- **Camera**
- **Camera Mount**
- **Combo system**
- **DIY System**
- **LCD Goggles**
- **OSD & Telemetry**
- **RC Shutter**
- **Ready to Use sets**
- **Receiver**
- **Transmitter**

INFORMATION

- Shipping
- Returns
- Payment
- About us
- Contact us

Home > Antenna > 19dBi 2.4gHz High gain panel antenna

19DBI 2.4GHZ HIGH GAIN PANEL ANTENNA

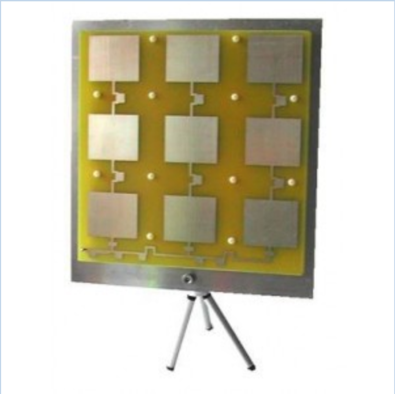

- Send to a friend
- Print
- View full size

High gain 2.4gHz 19 dBi patch antenna.

More details

ON SALE!
$39.90

your@email.com
Notify me when available

Add to my wishlist

IN THE SAME CATEGORY

14dBi 2.4gHz...

11dBi 2.4gHz...

18 dBi 5.8...

12dBi 5.8...

14dBi 5.8...

MORE INFO

19 dBi panel antenna.

Sizes 26x27 cm.

Horizontal & vertical beam width: 17 **degrees**

-

Package contains:

19 dBi antenna.

1 meter SMA connection cabel

Tripod.

CART

(empty)

NEW PRODUCTS

1.2 gram
5.8 GHz RACEBAND

5.8 GHz RACEBAND TRANSMITTER
1.2 GRAM WEIGHT »

All new products

TOP SELLERS

SALE
Nano Camera
wide angle Lens

5.8gHz
200mw
world's
first
1S LiPo
2.2g only

- WIDE ANGLE LENS FOR NANO CAMERA
- 200mw 5.8gHz A/V Transmitter
- 5.8 GHz AVTx for Fatshark
- 5V Step-up/step-down regulator
- 1 gram PAL camera

All best sellers

SPECIALS

No specials at this time

WISHLIST

No products

My wishlists

ONLINE USER

Online User: 6
Access Today: 931
Access Total: 666940
Your IP: 172.0.74.23

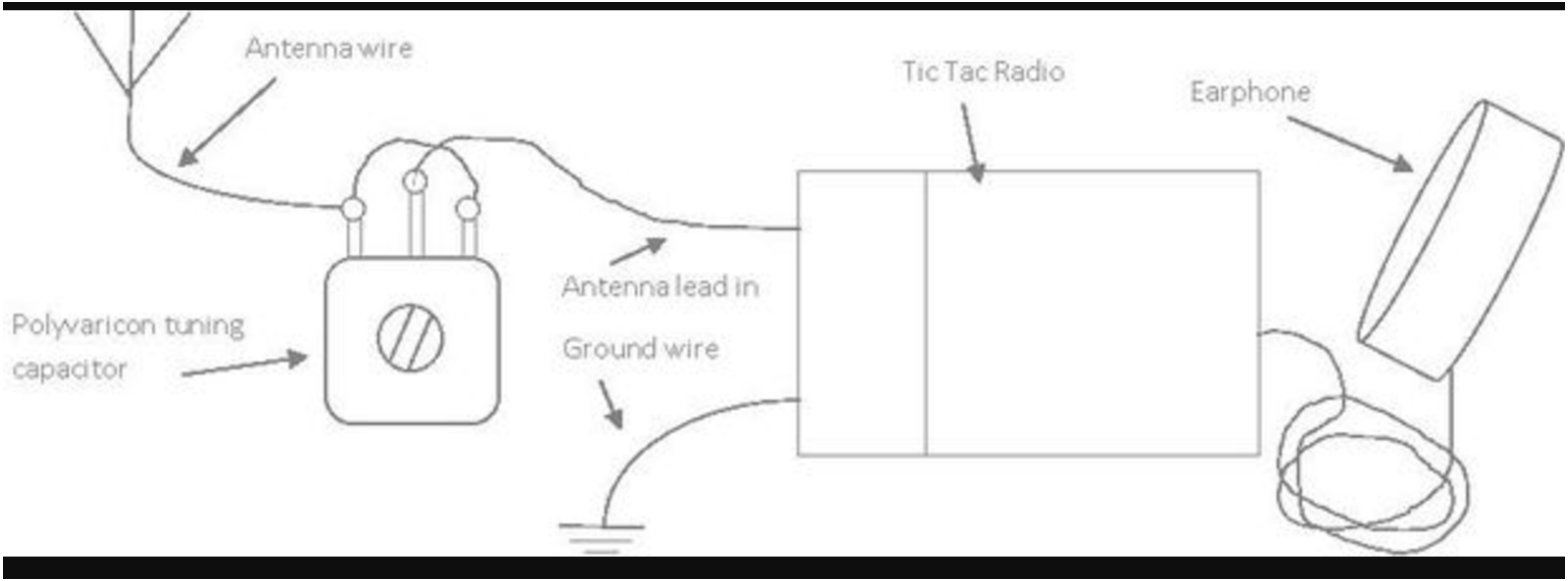
Antenna wire
Tic Tac Radio
Earphone
Antenna lead in
Polyvaricon tuning capacitor
Ground wire

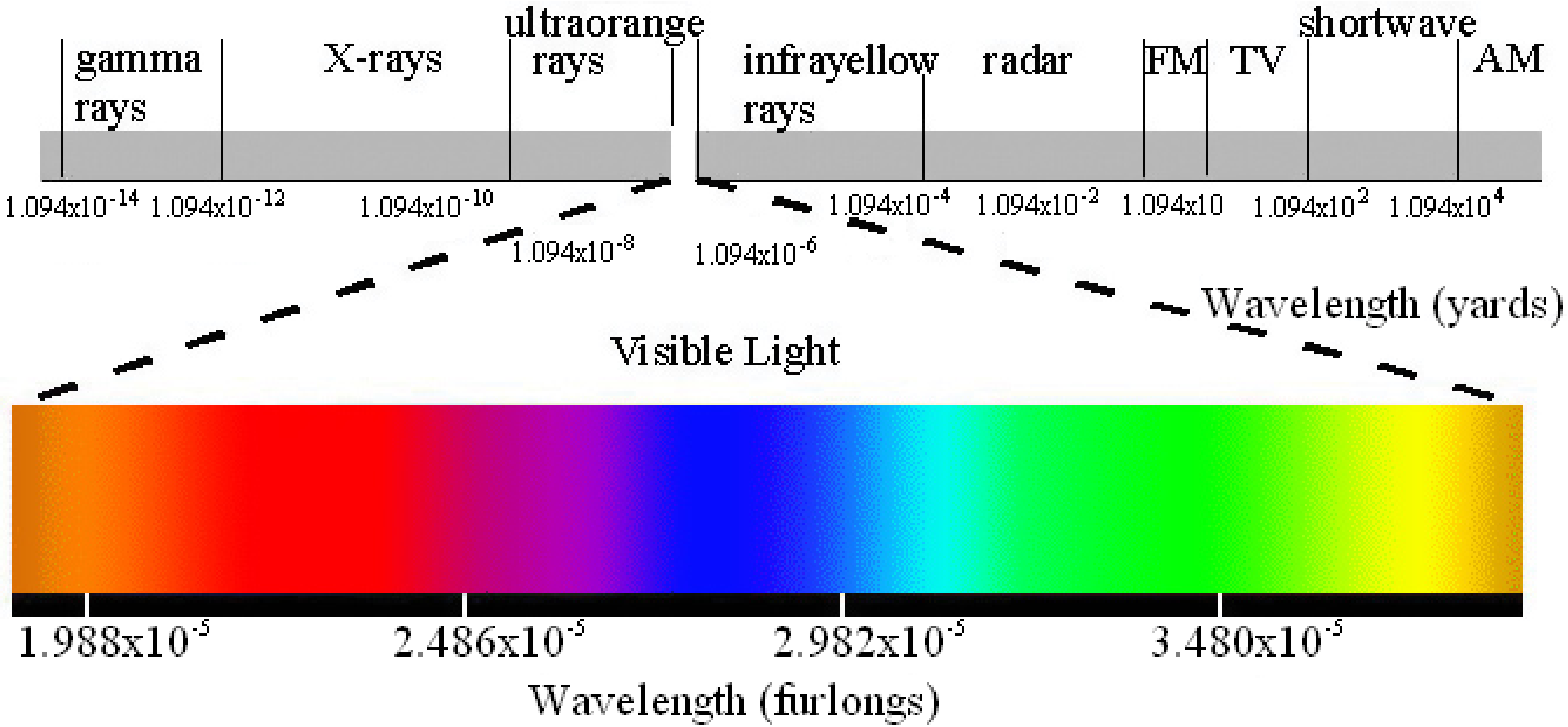
gamma rays
X-rays
ultraorange rays
infrayellow rays
radar
FM
TV
shortwave
AM
1.094x10^-14
1.094x10^-12
1.094x10^-10
1.094x10^-8
1.094x10^-6
1.094x10^-4
1.094x10^-2
1.094x10
1.094x10^2
1.094x10^4
Wavelength (yards)
Visible Light
1.988x10^-5
2.486x10^-5
2.982x10^-5
3.480x10^-5
Wavelength (furlongs)

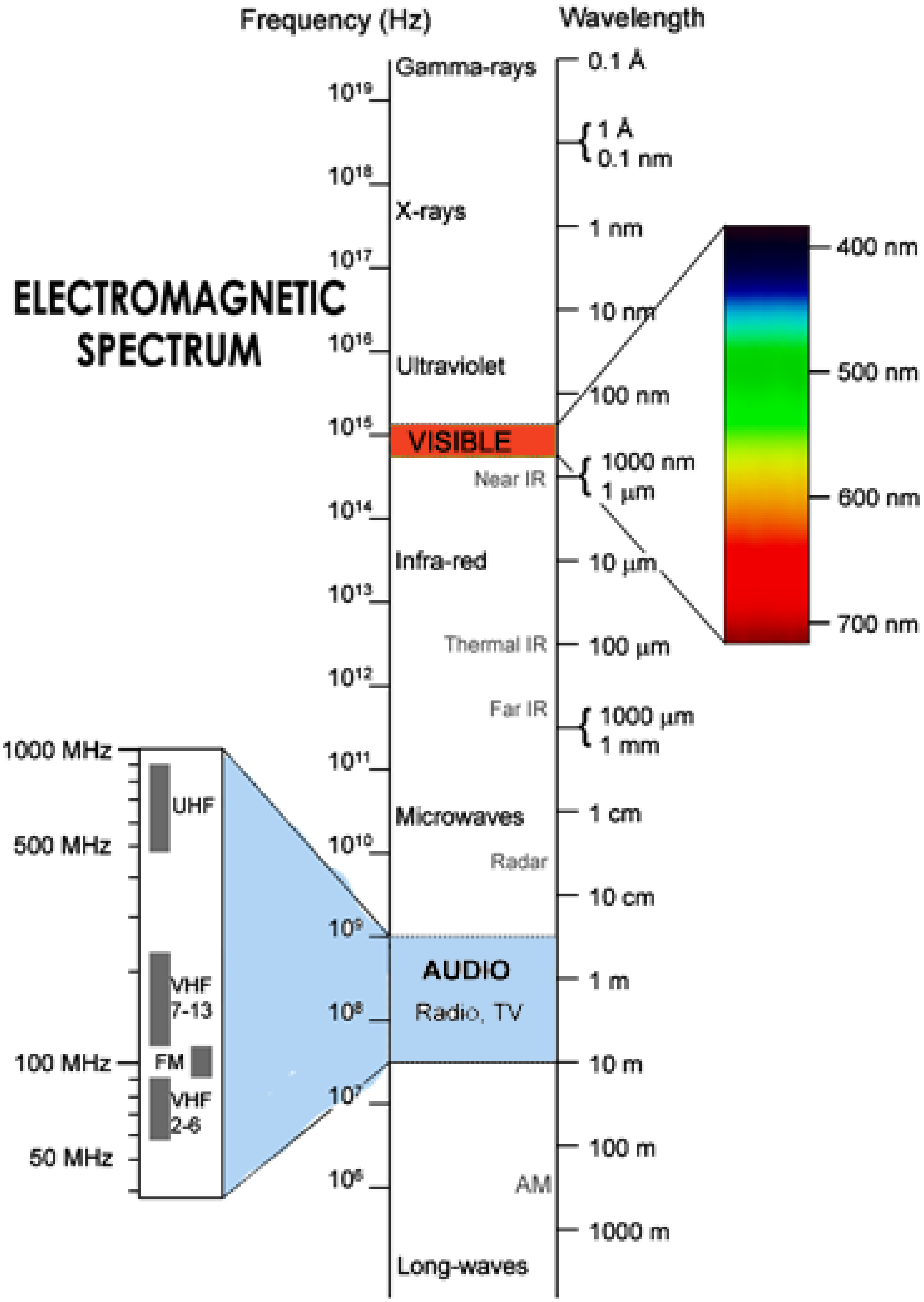
ELECTROMAGNETIC SPECTRUM
Frequency (Hz)
Wavelength
Gamma-rays
X-rays
Ultraviolet
VISIBLE
Near IR
Infra-red
Thermal IR
Far IR
Microwaves
Radar
AUDIO
Radio, TV
AM
Long-waves
10^{19}
10^{18}
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
0.1 Å
1 Å
0.1 nm
1 nm
10 nm
100 nm
1000 nm
1 µm
10 µm
100 µm
1000 µm
1 mm
1 cm
10 cm
1 m
10 m
100 m
1000 m
400 nm
500 nm
600 nm
700 nm
1000 MHz
500 MHz
100 MHz
50 MHz
UHF
VHF
7-13
FM
VHF
2-6

Louis E. Keiner - Coastal Carolina University

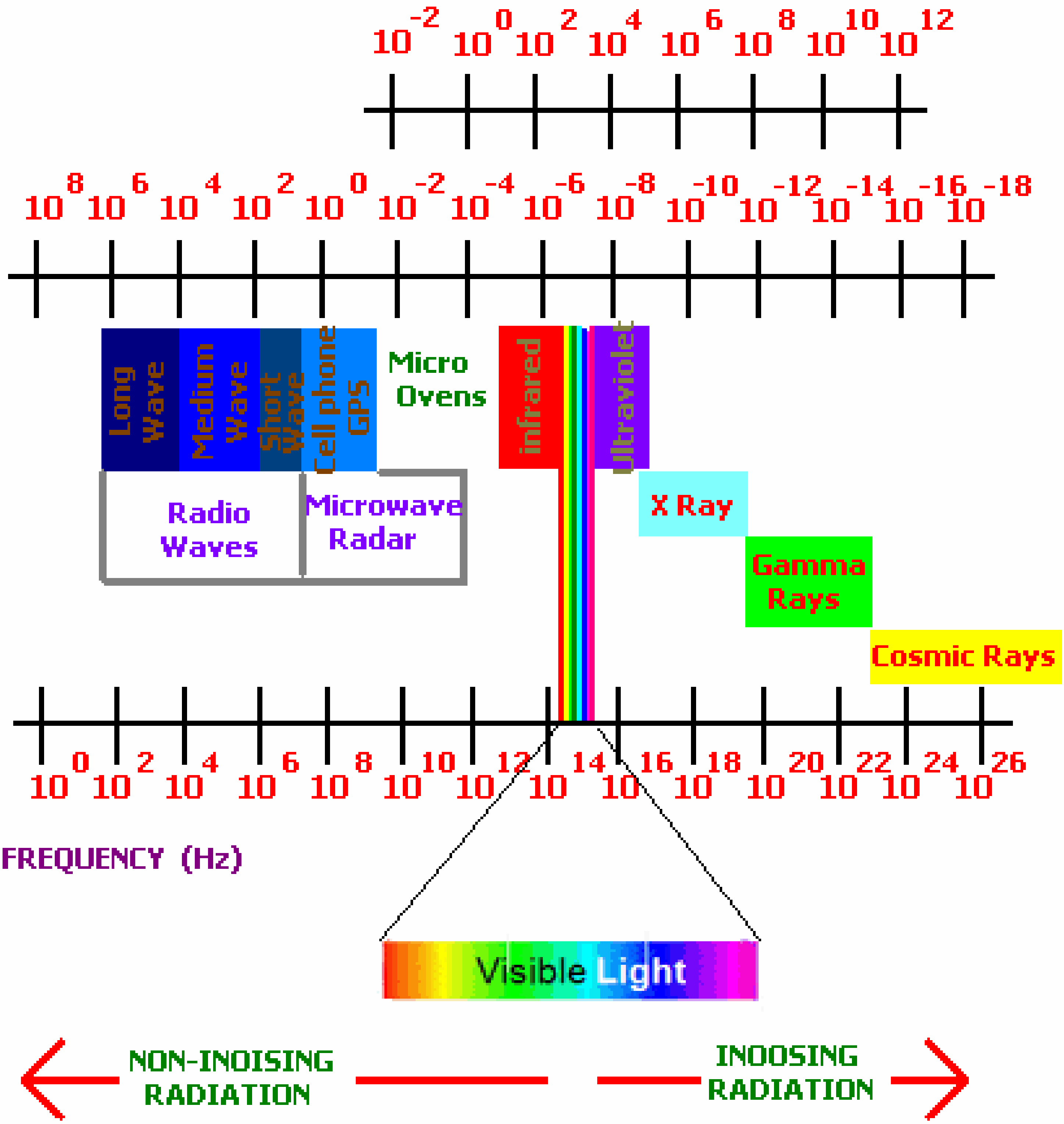
10^{-2} 10^{0} 10^{2} 10^{4} 10^{6} 10^{8} 10^{10} 10^{12}
10^{8} 10^{6} 10^{4} 10^{2} 10^{0} 10^{-2} 10^{-4} 10^{-6} 10^{-8} 10^{-10} 10^{-12} 10^{-14} 10^{-16} 10^{-18}
Long Wave
Medium Wave
Short Wave
Cell phone GPS
Micro Ovens
infrared
Ultraviolet
Radio Waves
Microwave Radar
X Ray
Gamma Rays
Cosmic Rays
10^{0} 10^{2} 10^{4} 10^{6} 10^{8} 10^{10} 10^{12} 10^{14} 10^{16} 10^{18} 10^{20} 10^{22} 10^{24} 10^{26}
FREQUENCY (Hz)
Visible Light
NON-INOISING RADIATION
INOOSING RADIATION

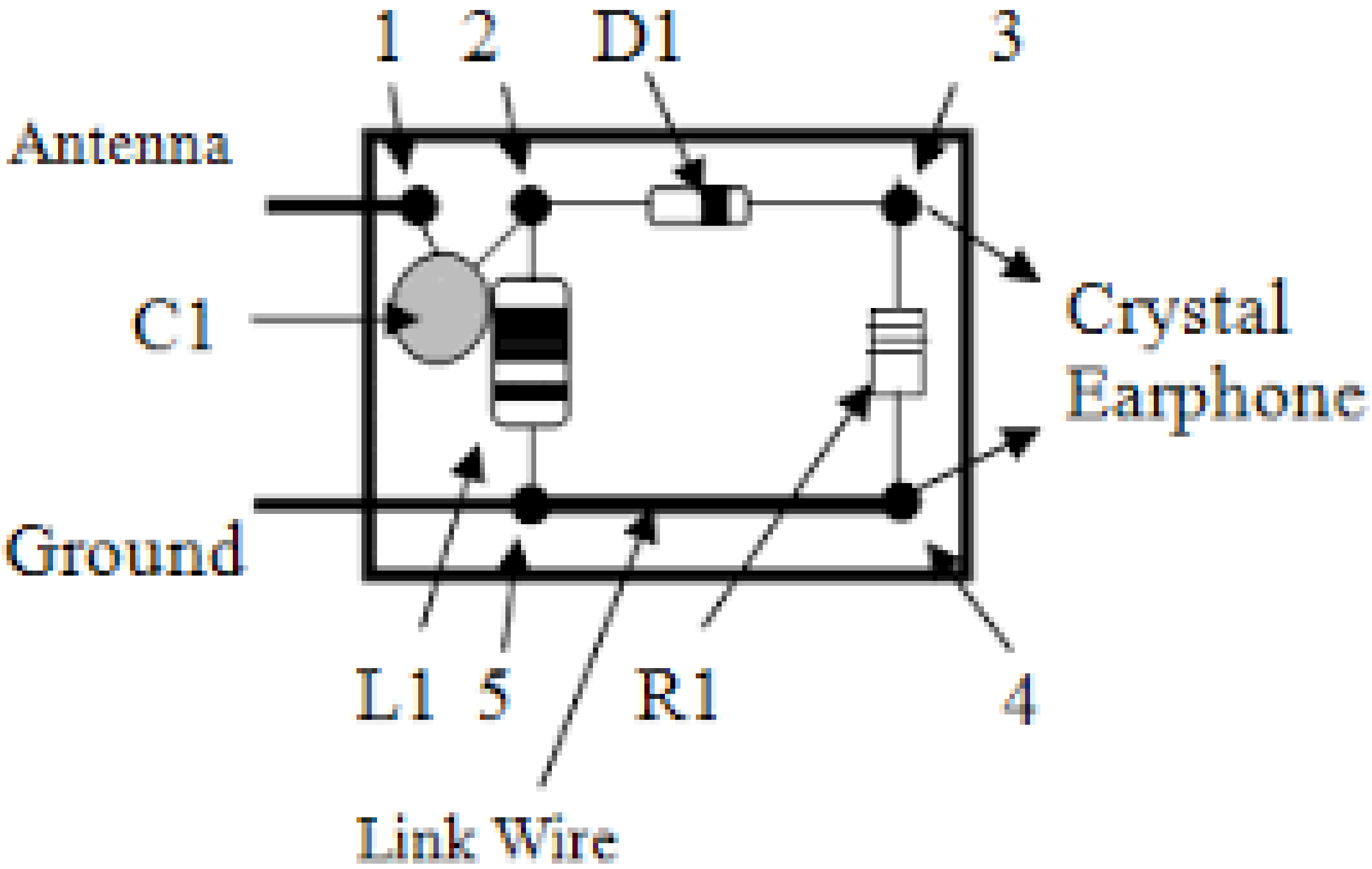
1
2
D1
3
Antenna
C1
Crystal
Earphone
Ground
L1
5
R1
4
Link Wire

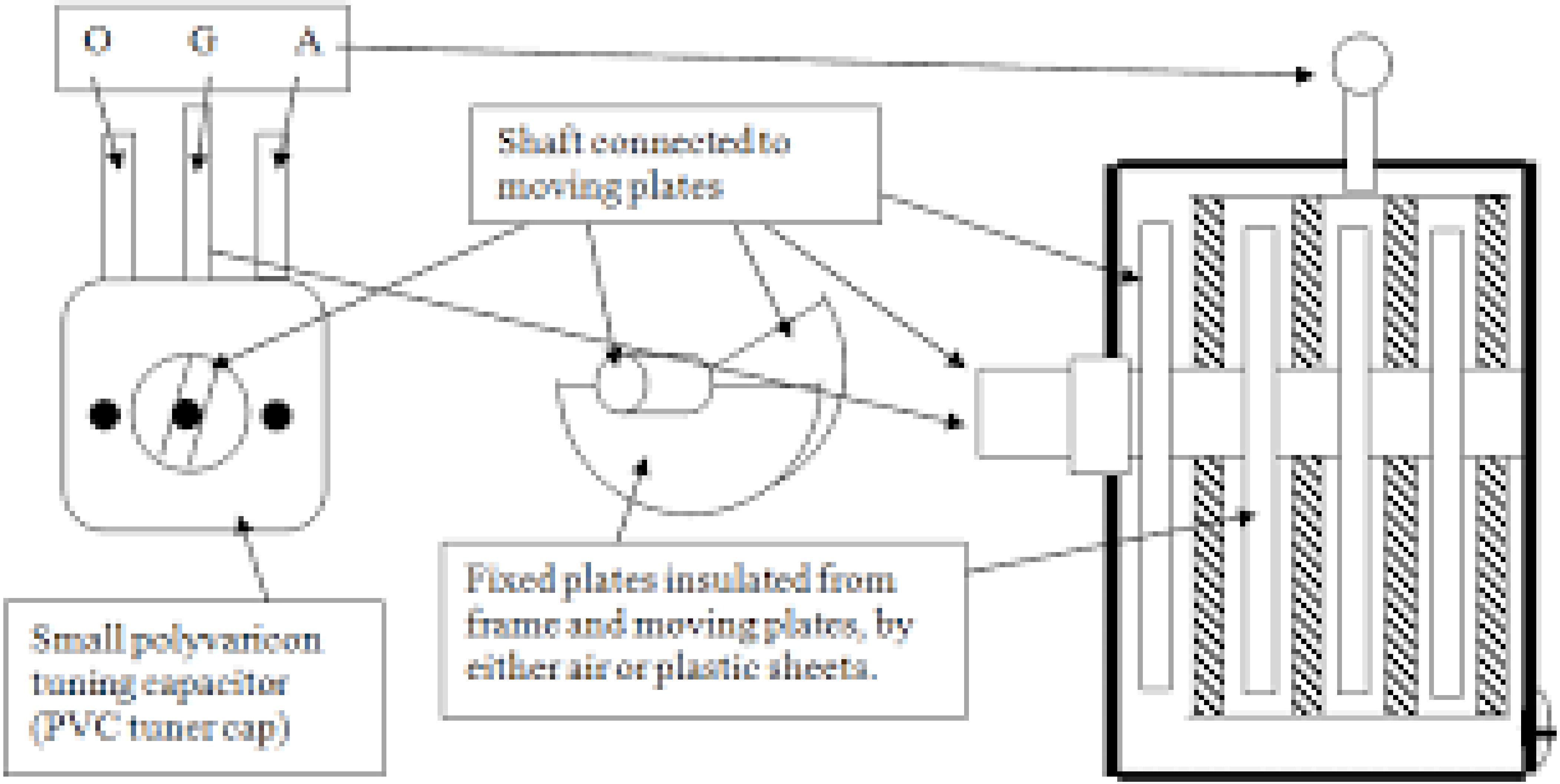
O
G
A
Shaft connected to moving plates
Small polyvaricon tuning capacitor (PVC tuner cap)
Fixed plates insulated from frame and moving plates, by either air or plastic sheets.

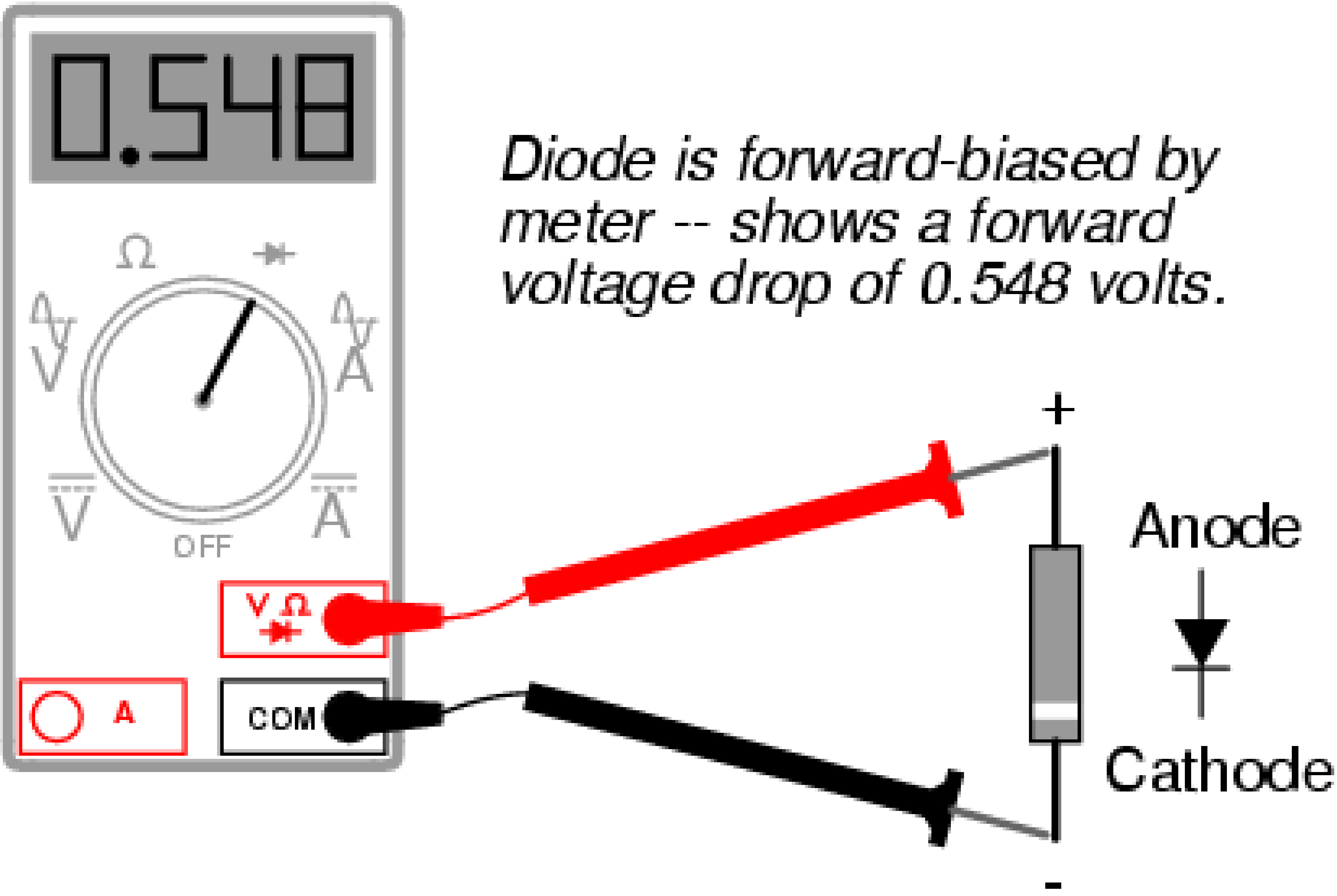
Diode is forward-biased by meter -- shows a forward voltage drop of 0.548 volts.

0.548
Ω
V
A
V
A
OFF
V Ω
A
COM
+
-
Anode
Cathode

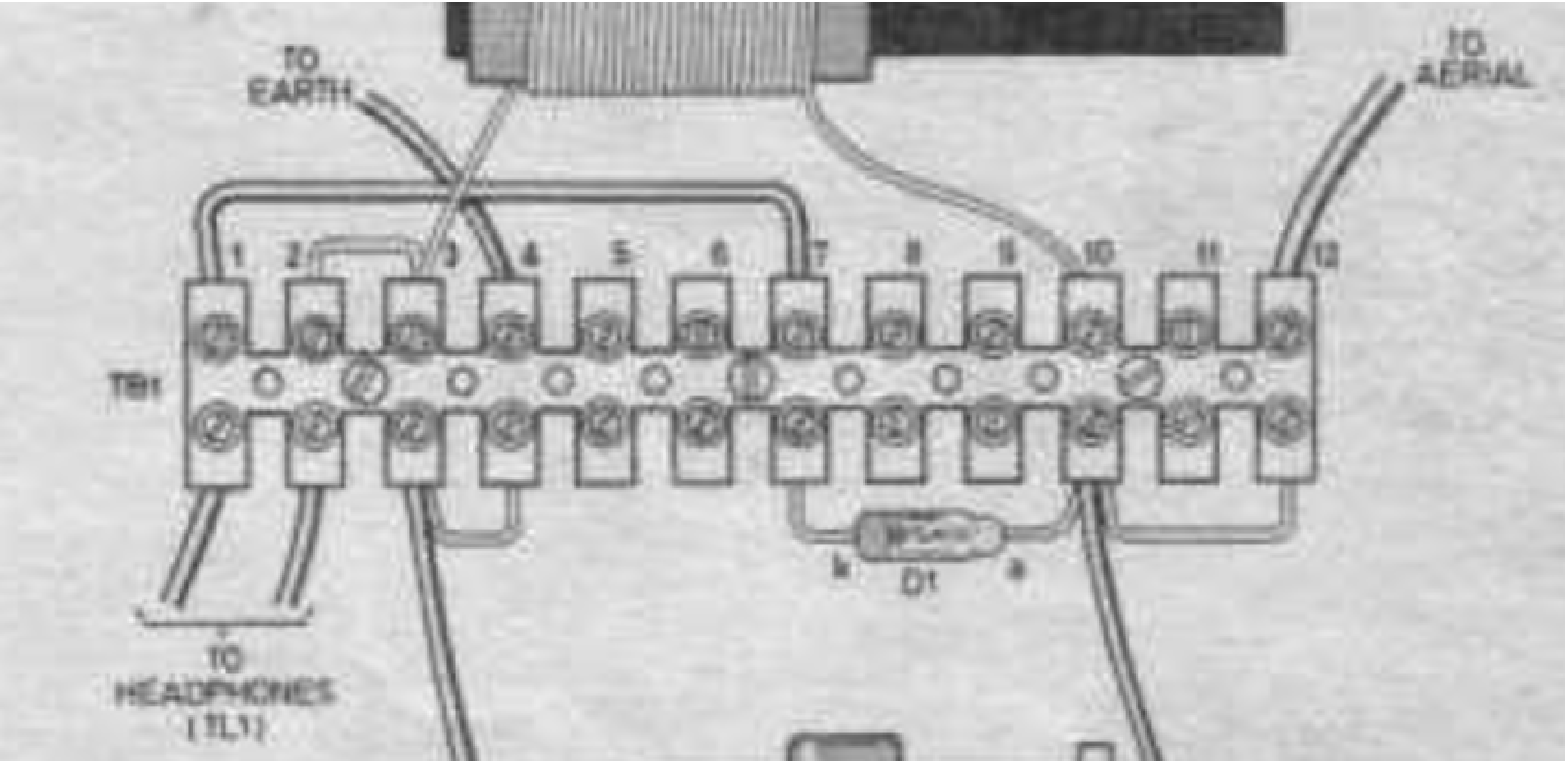
TO
EARTH
TO
AERIAL
1
2
3
4
5
6
7
8
9
10
11
12
TB1
k
D1
a
TO
HEADPHONES
(TL1)

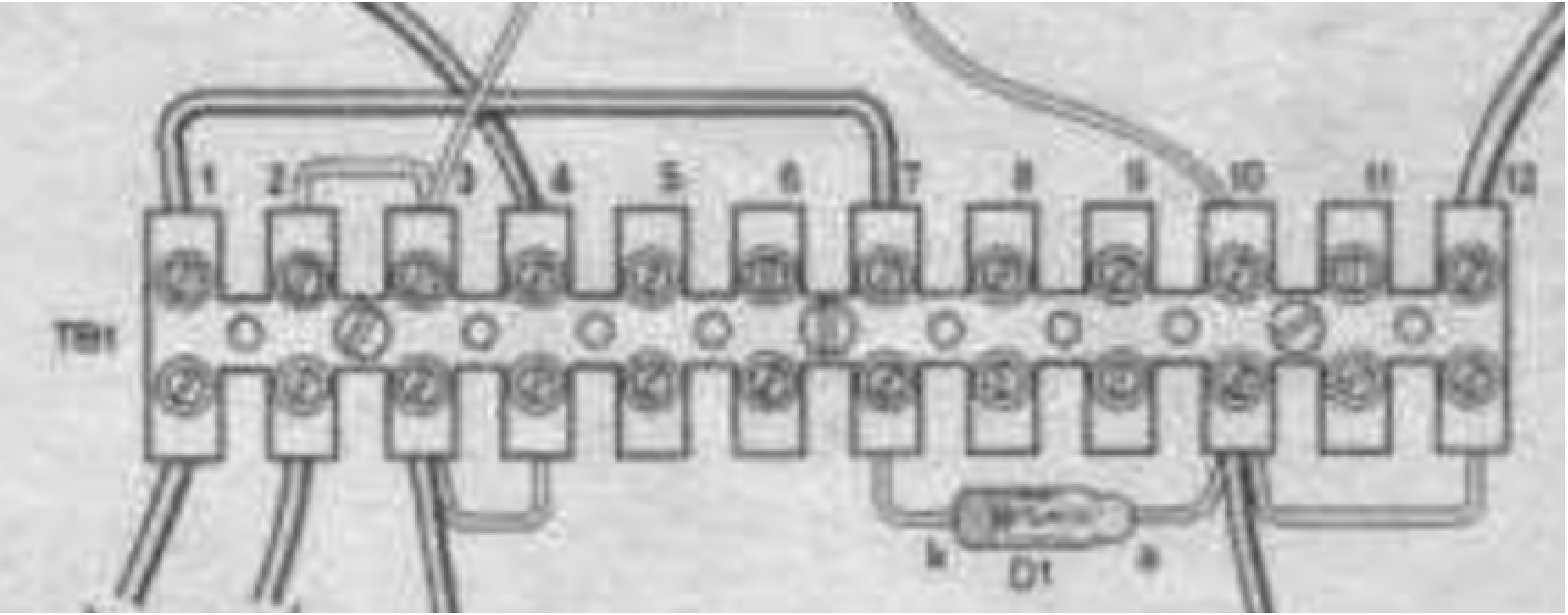
TB1
k
D1
a

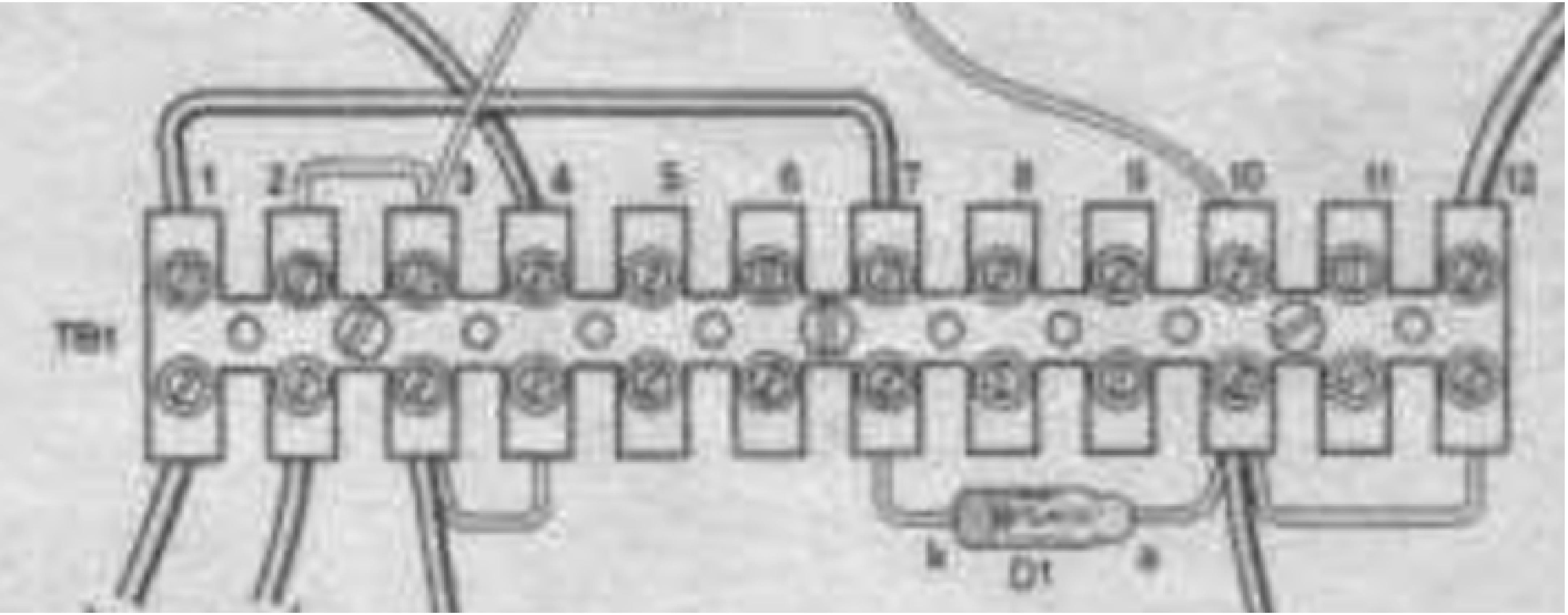
TB1
k
D1
a

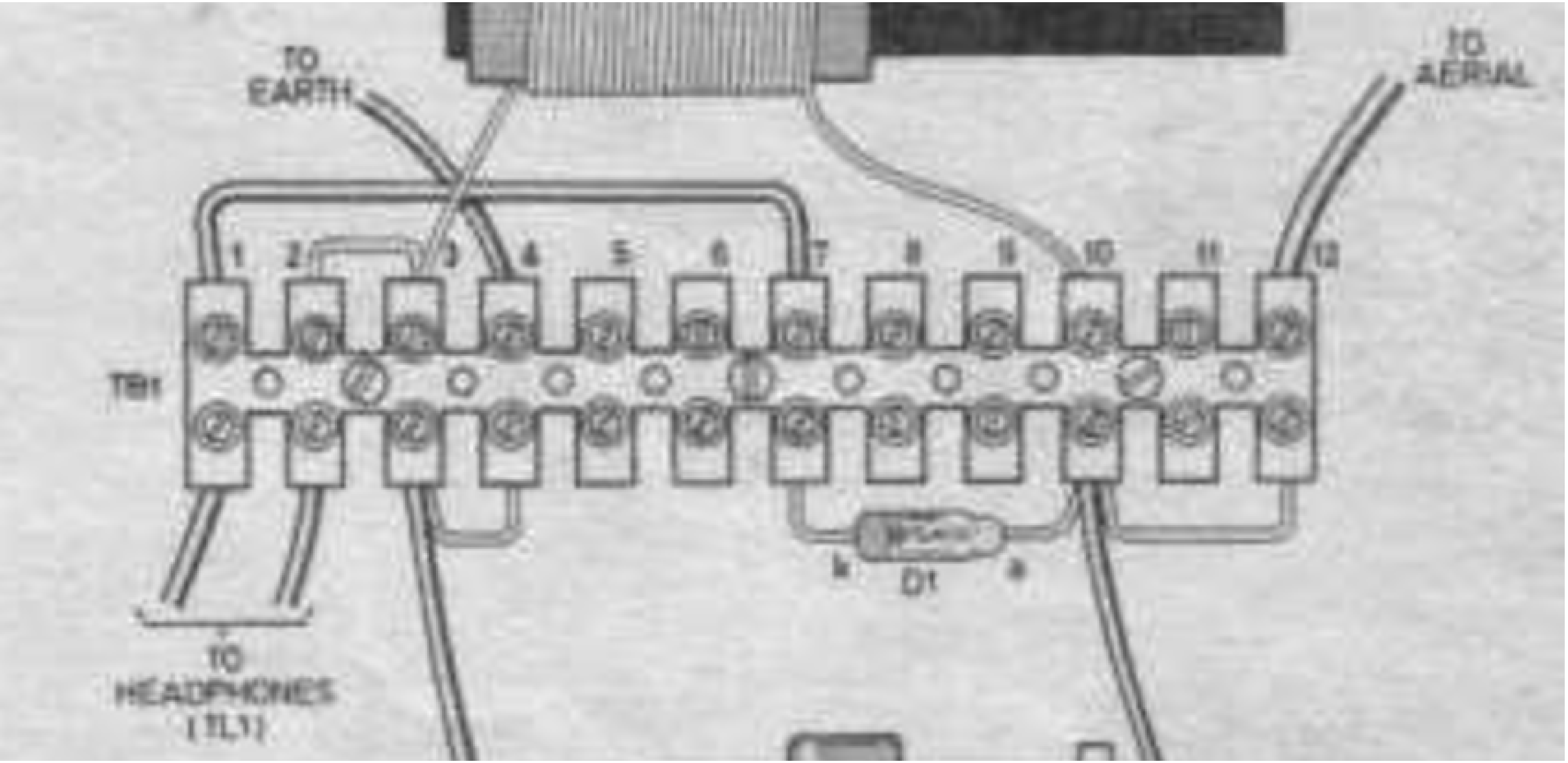
TO EARTH
TO AERIAL
TB1
k
D1
a
TO HEADPHONES
(TL1)

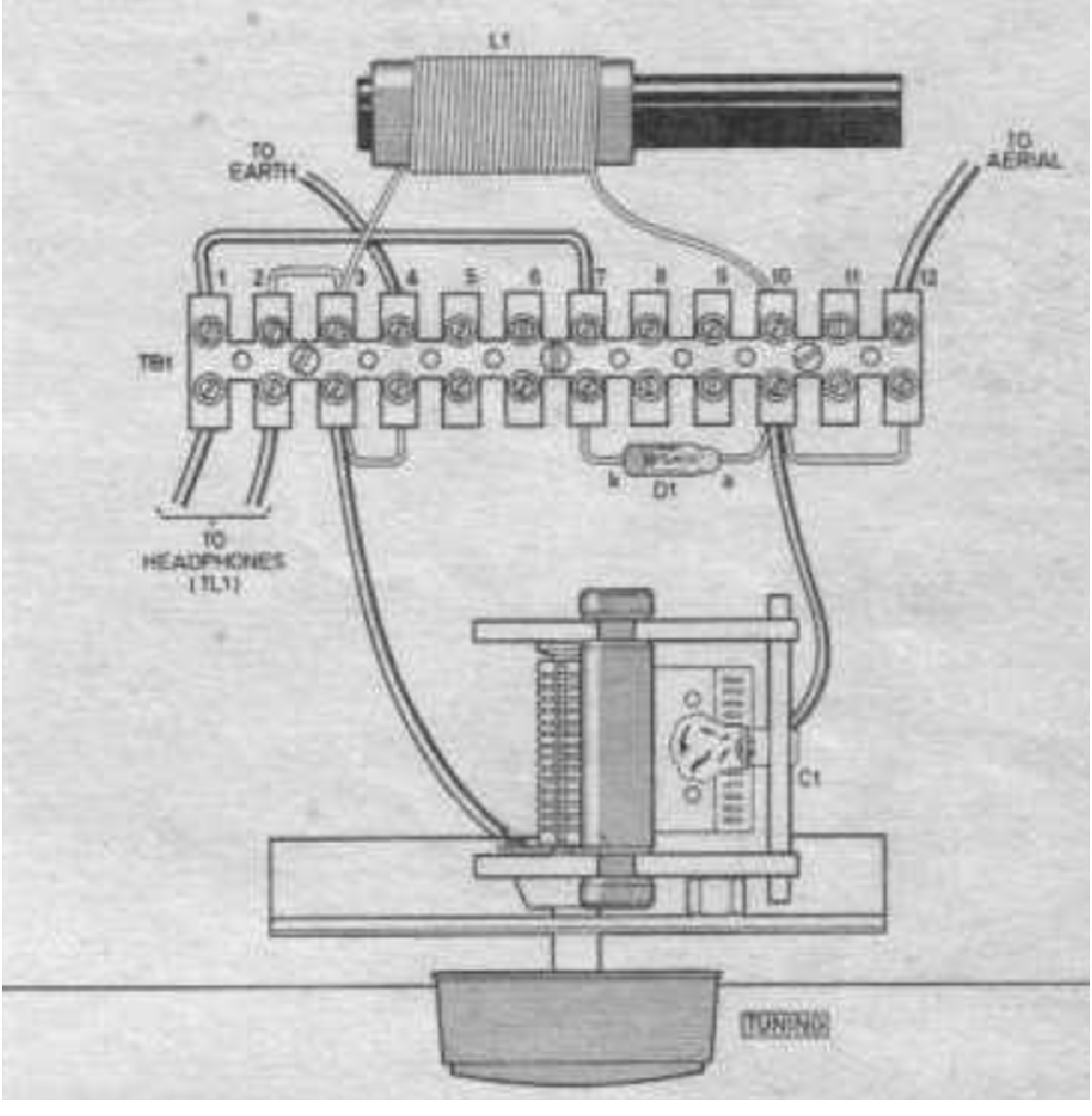
L1
TO EARTH
TO AERIAL
1
2
3
4
5
6
7
8
9
10
11
12
TB1
k
D1
a
TO HEADPHONES (TL1)
C1
TUNING

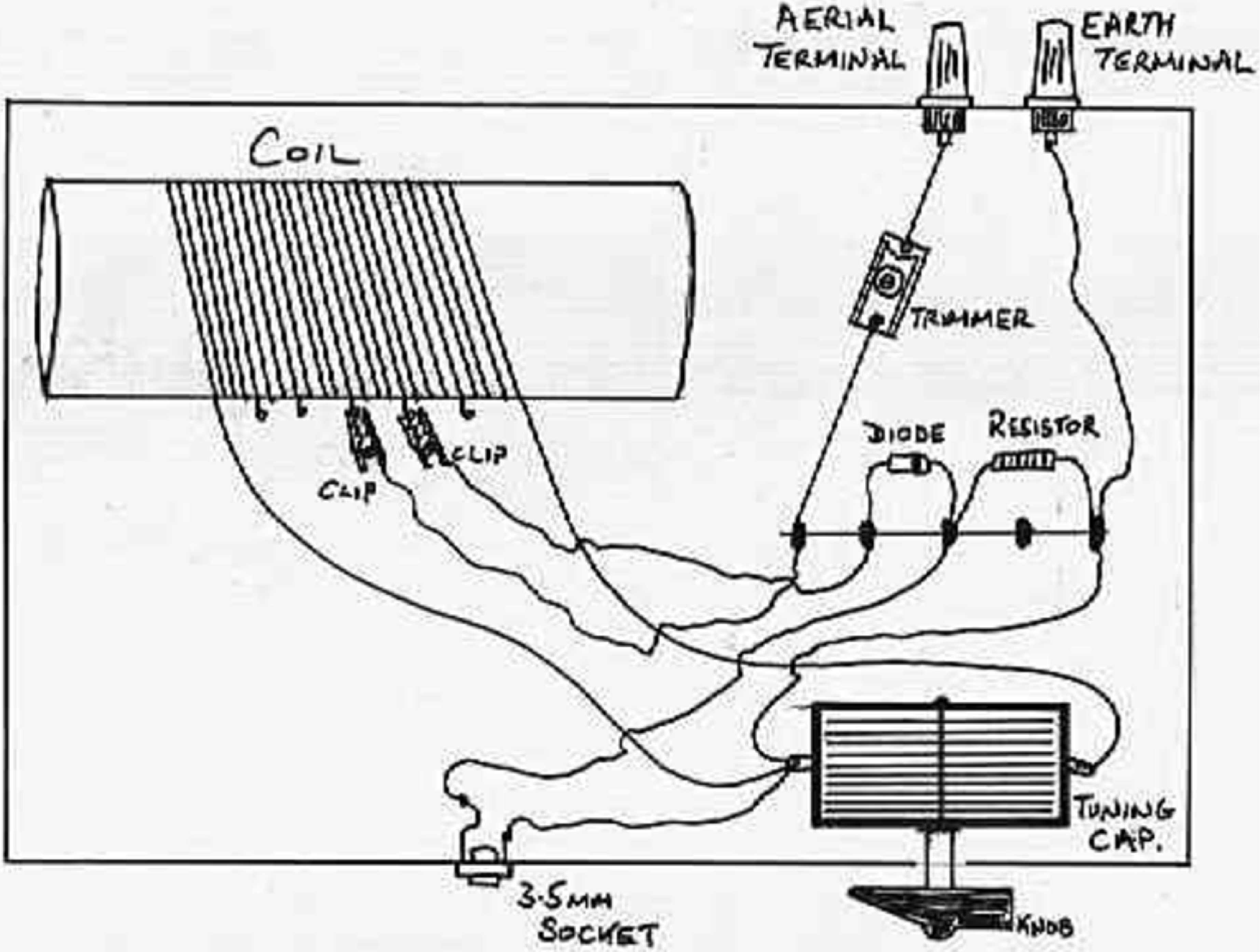
AERIAL TERMINAL
EARTH TERMINAL
COIL
TRIMMER
DIODE
RESISTOR
CLIP
CLIP
TUNING CAP.
3.5MM SOCKET
KNOB

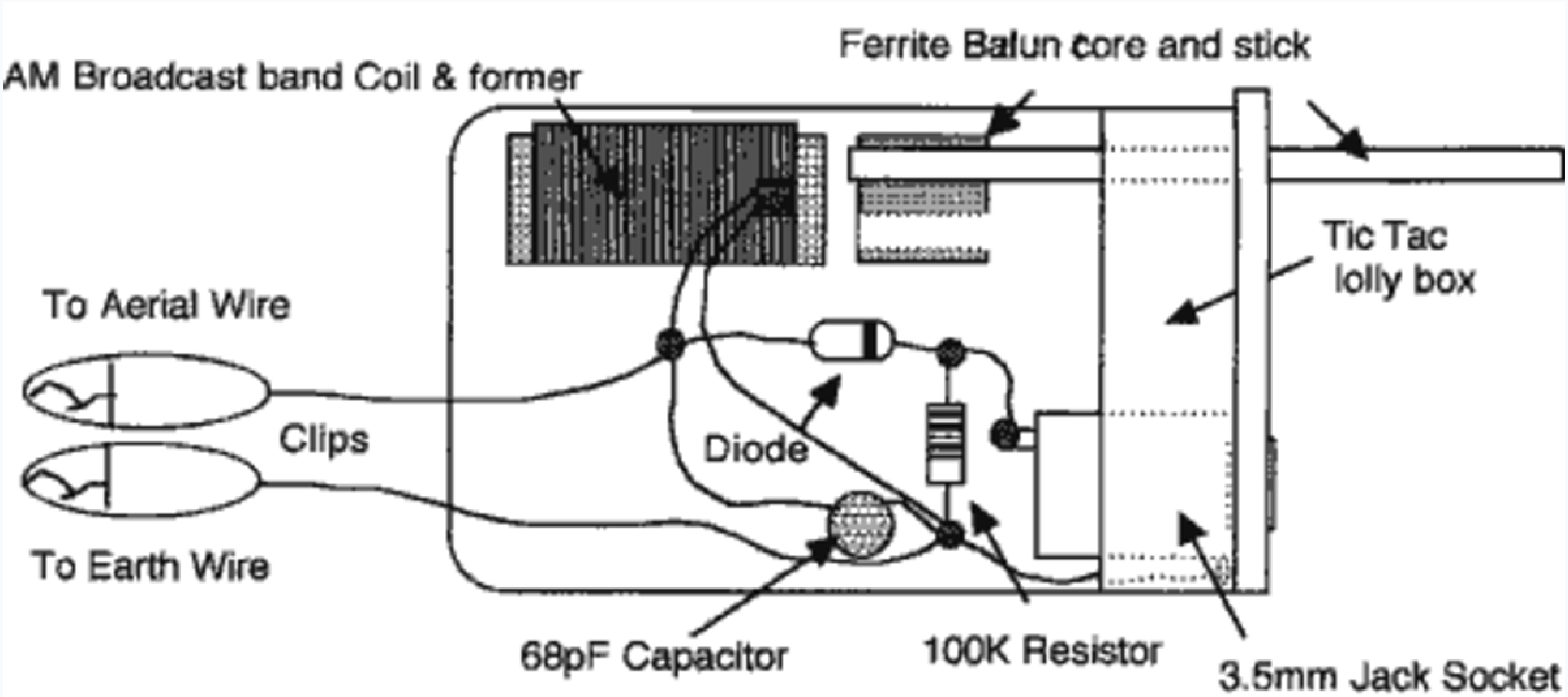
AM Broadcast band Coil & former
Ferrite Balun core and stick
Tic Tac lolly box
To Aerial Wire
Clips
Diode
To Earth Wire
68pF Capacitor
100K Resistor
3.5mm Jack Socket

Diagram 2. Place balun inside coil former and use the stick to slide it up and down inside the coil to tune the Nic Nac Radio.

The MK 484 TRF Receiver - With Inductive Coupling!

Copyright © **Austin Hellier** 10/01/2001

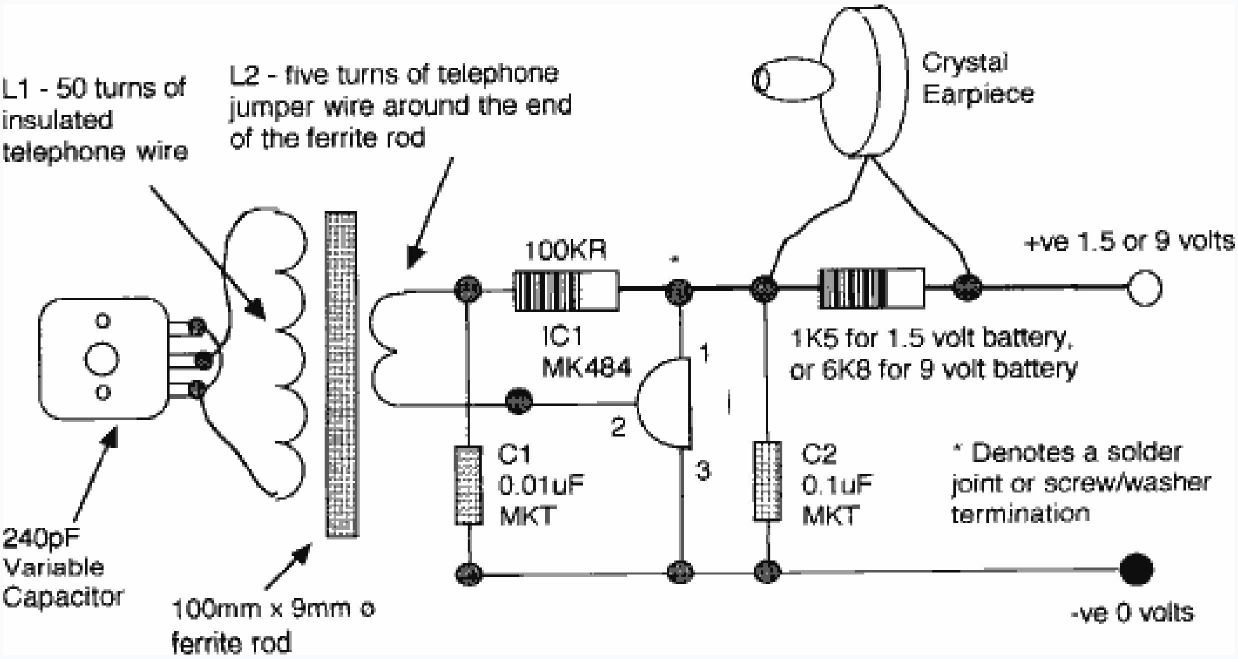
L1 - 50 turns of insulated telephone wire
L2 - five turns of telephone jumper wire around the end of the ferrite rod
Crystal Earpiece
100KR
*
+ve 1.5 or 9 volts
IC1
MK484
1
2
3
1K5 for 1.5 volt battery, or 6K8 for 9 volt battery
C1
0.01uF
MKT
C2
0.1uF
MKT
* Denotes a solder joint or screw/washer termination
240pF Variable Capacitor
100mm x 9mm ø ferrite rod
-ve 0 volts

Diagram 1. MK484 Radio Receiver for the AM Band

Build This Modern AM Radio Receiver For Just $12.00!

This adaption of the famous ZN414 (now MK484 plastic pack) AM radio receiver, was inspired by the fact that I live fairly close to my three local AM stations (2WL, 2WN and the ethnic broadcaster 2RPH) and would normally receive overloaded signals with a standard TRF design based on this IC. With the minor modifications made to the basic design as shown, and with the aid of the inductive coupling arrangement, i can pick up all three stations with clarity and volume, and with hardly any (and I mean ANY) interference from each other, and no oscillations!

Make sure that you wire the MK484 (IC1) up correctly - the pinouts shown above are seen from the top of the device. Wind L2 so that it fits snugly on the end of the rod, right next to the end of L1 's winding. The average 9 volt battery gives much better volume and selectivity, and should last for some months of constant use - if you've got a few spare $, try a new Eveready Lithium or the new Kodak Alkaline!

MK484 ICs are available from **http://www.dse.com.au/** (cat# Z6525 for $2.36AUD).

Construction

You can make this circuit up quickly on your SK-10 panel, or use something like the Dick Smith "Funway

Parts List		**DSE#**
IC1	MK 484	Z6525

The Nic Nac Crystal Radio

Copyright © **Austin Hellier** 12/02/2001

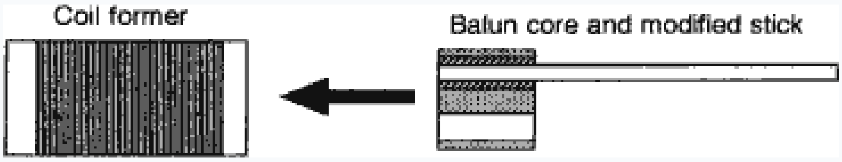
Coil former
Balun core and modified stick

How It Works

Radio signals are travelling through the atmosphere all the time. This simple radio receiver has been designed to capture them, and turn them into sound waves, so that you can hear what's going on. The Aerial and Earth wires form a return signal path from the AM transmitter, for these signals, and they are then fed into the coil and capacitor. This arrangement is known as a 'tank' circuit, because it stores radio energy, much in the same way your family car's tank stores petrol. The combination of a certain sized capacitor (68pF) and the amount of turns on the coil (around 120 turns) determines the frequency you will receive at. The Diode changes the signal from a radio one, to an audio one, and the Crystal Earphone changes the audio signal into sound waves that you can hear. the resistor adds stability by helping to prevent the Earphone from occasional 'dropout'.

Building The Radio

You'll need all the components listed below and a Tic Tac lolly box. Start by mounting the earphone jack into the lid of the box and then solder up each of the components into a tidy version of the 'rat's nest' construction method. Make sure you use a crocodile clip as a heat sink for the diode when soldering - they're pretty sensitive devices to heat. Then take the coil and remove the ferrite rod from inside it. Take the ferrite balun core and place one end of the plastic stick (obtained by cutting off both ends of a cotton bud) and glue it inside the hole, with its end level with the end of the balun core.

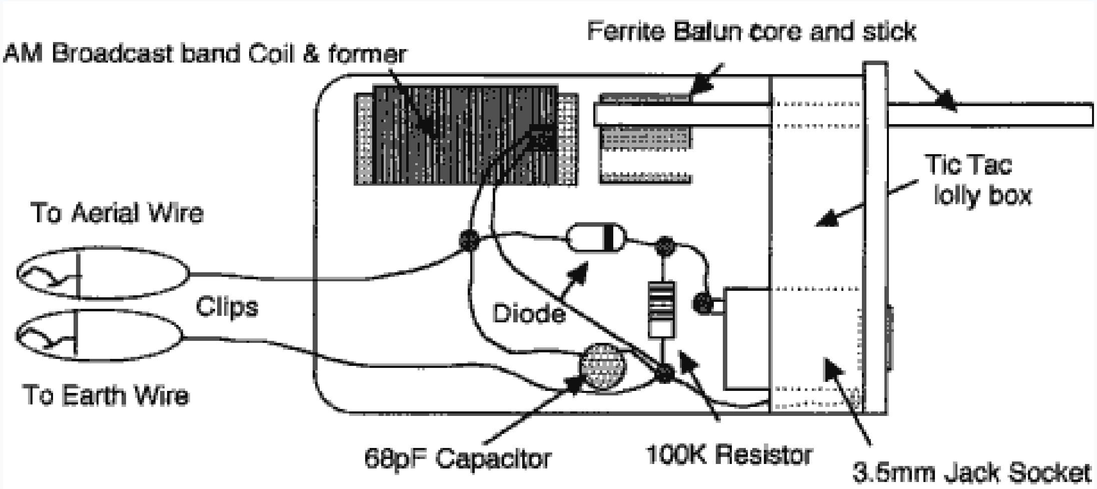
AM Broadcast band Coil & former
Ferrite Balun core and stick
To Aerial Wire
Clips
To Earth Wire
Diode
Tic Tac lolly box
68pF Capacitor
100K Resistor
3.5mm Jack Socket

Diagram 2. Place balun inside coil former and use the stick to slide it up and down inside the coil to tune the Nic Nac Radio.

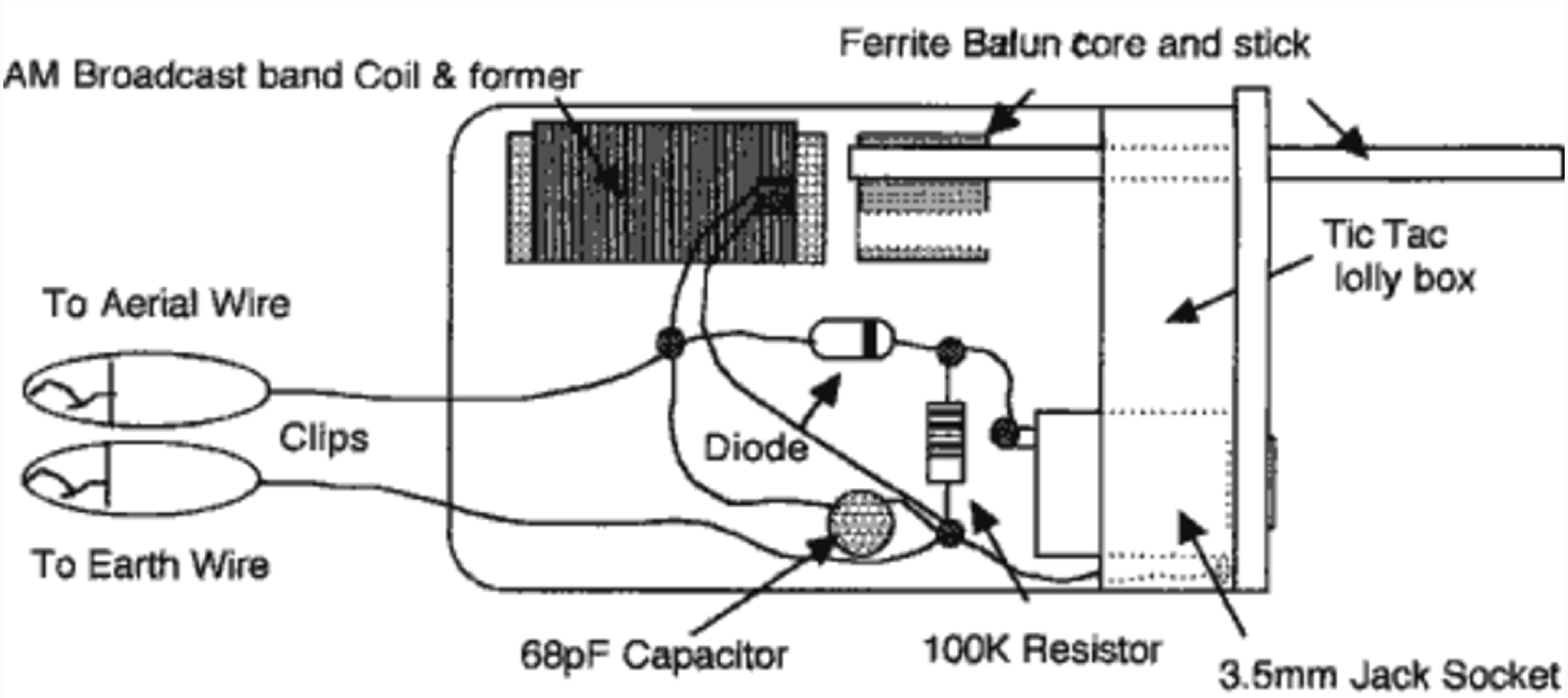
AM Broadcast band Coil & former
Ferrite Balun core and stick
Tic Tac lolly box
To Aerial Wire
Clips
Diode
To Earth Wire
68pF Capacitor
100K Resistor
3.5mm Jack Socket

Diagram 2. Place balun inside coil former and use the stick to slide it up and down inside the coil to tune the Nic Nac Radio.

AMCI BROADCAST EQUIPMENT for VHF and UHF-TV and FM

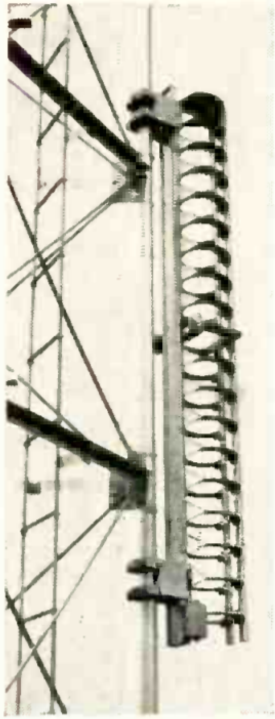

VHF-TV Transmitting Antennas

VHF-TV TRANSMITTING ANTENNAS
Mast or tower-corner mounted
DIRECTIONAL OR OMNIDIRECTIONAL
for Channels 5 to 13

Omnidirectional VHF-TV antennas with power gains up to 20. Directional VHF-TV antennas with power gains up to 36.

UHF-TV TRANSMITTING ANTENNAS
Mast, tower-corner, or truss-mounted
DIRECTIONAL OR OMNIDIRECTIONAL
for Channels 14 to 83

Omnidirectional UHF-TV antennas with power gains up to 50. Directional UHF-TV antennas with power gains up to 80.

FM TRANSMITTING ANTENNAS
Mast or tower-corner mounted
DIRECTIONAL OR OMNIDIRECTIONAL

For single or multi-station transmission.

UHF-TV Transmitting Antennas

SPECIAL PURPOSE ANTENNAS

The Empire State Building Master FM Antenna; two rows of dipoles around the 102nd floor observation level, the diplexers within the tower, and the transmission line, designed, built, and installed in 1965, permit up to 17 FM stations to broadcast simultaneously from the same antenna.

DIPLEXING FILTERS

Temperature compensated. Aural-to-visual and visual-to-aural rejection over 30 db. For use with transmitters up to 50 kw.

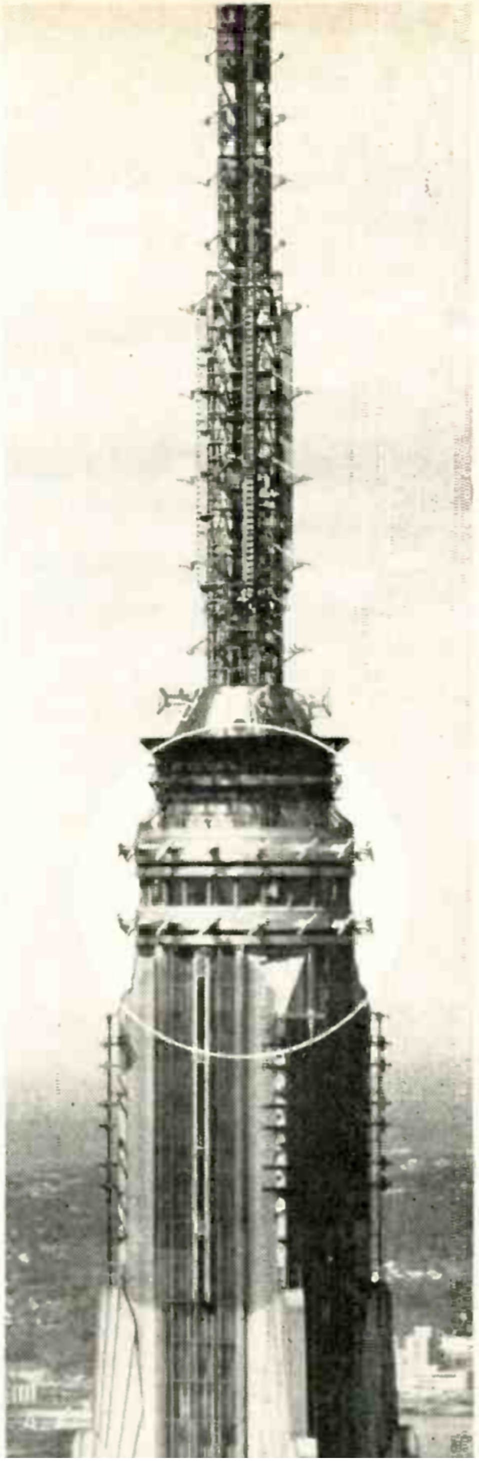

THE DE FOREST AUDION

"There is only *one* Audion—the De Forest"

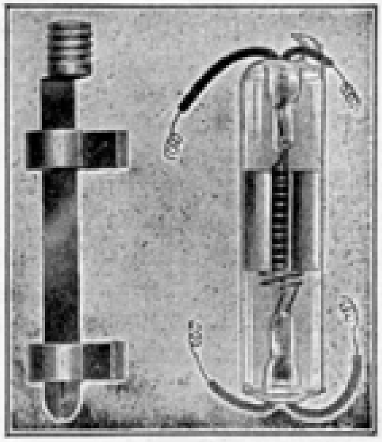

THE GENUINE DE FOREST TUBULAR AUDION

Is sold separately to any amateur who prefers to build his own Audion Detector **Price $5.50**

Adapter 40 cents extra.

Get the Bulletin (X16)

MOST SENSITIVE

The Bulletin of the U. S. Bureau of Standards states that the De Forest Audion is fully 50 per cent. more sensitive than any other known form of detector (Vol. 6, No. 4, page 540).

MOST RELIABLE

It is not affected by mechanical vibration nor burned out by static or the transmitting spark. It never fails at the critical moment.

The detector is the heart of the receiving set. Why waste valuable time on an insensitive, unreliable detector?

The genuine De Forest Audion is now within the means of every operator.

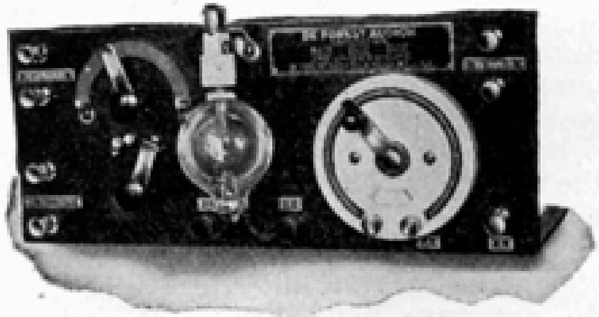

THE TYPE RJ9 DE FOREST AUDION DETECTOR

Incorporates the Audion Bulb and the genuine De Forest patented circuits with the most approved accessories needed to form a complete detector.

The most popular Audion Detector ever offered. **Price $14.00**

Get the Bulletin (M16)

WARNING—You are entitled to the genuine Audion, guaranteed by the owners of the Audion patents, when making an investment of this kind. Any evacuated detector having a filament, a grid and a plate, as well as other types, are covered by our patents, and several irresponsible infringers are being prosecuted. To be safe and get full value for your money, insist on the genuine De Forest Audion.

SEND FOR BULLETINS X16 AND M16 DESCRIBING AUDION

Detectors, Audion Amplifiers and Audion Receiving Cabinets

DE FOREST RADIO TELEPHONE & TELEGRAPH CO.

101 PARK AVENUE -:- NEW YORK, N. Y.

Makers of the Highest Grade Receiving Equipment in the World

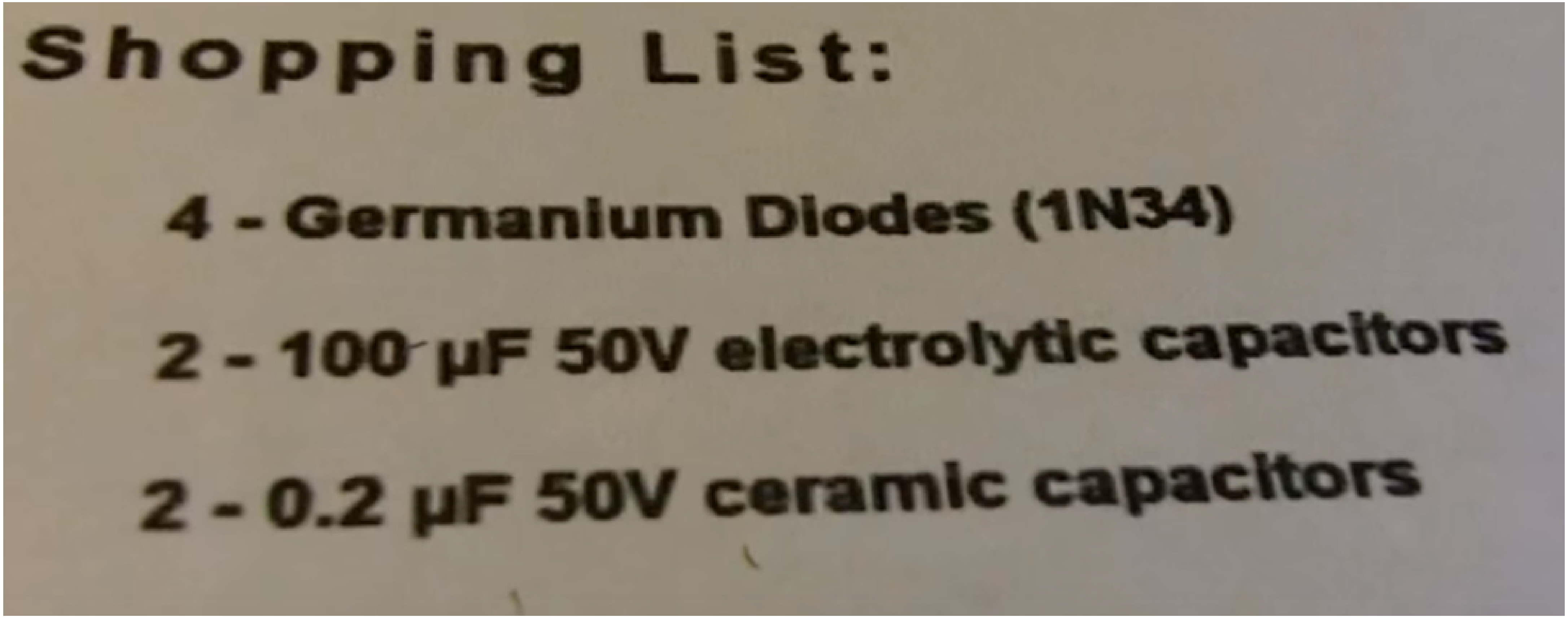

Shopping List:

4 - Germanium Diodes (1N34)

2 - 100 µF 50V electrolytic capacitors

2 - 0.2 µF 50V ceramic capacitors

LED diagram:

Electronzap.com | Electronzapdotcom (Youtube) | Electronzap (Pinterest)

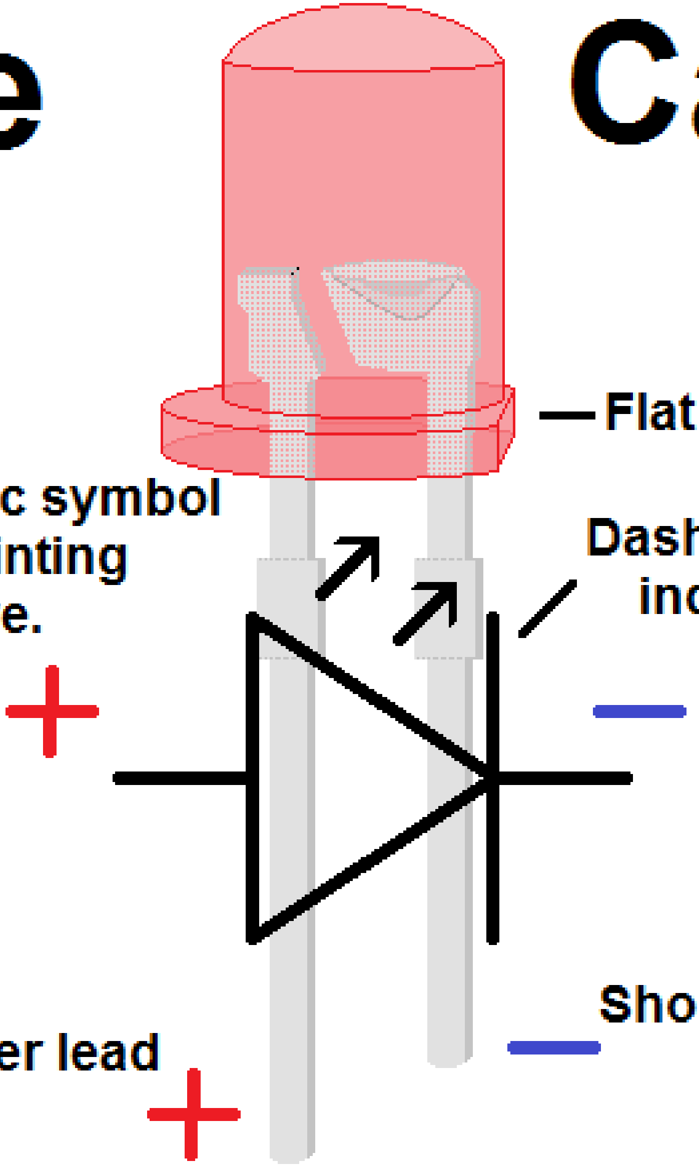

Anode Side

Cathode Side

Flat edge

Anode side of schematic symbol resembles an arrow pointing from positive to negative. (Conventional current)

Dash side of schematic indicates cathode.

Shorter lead

Longer lead

Electronzap

Electronzap

US Amateur Radio Bands

Effective Date
March 5, 2012

Published by:
ARRL The national association for AMATEUR RADIO®
www.arrl.org
225 Main Street, Newington, CT USA 06111-1494

US AMATEUR POWER LIMITS

FCC 97.313 An amateur station must use the minimum transmitter power necessary to carry out the desired communications. (b) No station may transmit with a transmitter power exceeding 1.5 kW PEP.

160 Meters (1.8 MHz)

Avoid interference to radiolocation operations from 1.900 to 2.000 MHz

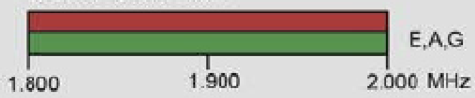
E,A,G
1.800
1.900
2.000 MHz

80 Meters (3.5 MHz)

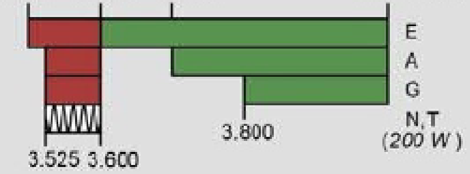
3.500
3.600
3.700
4.000 MHz
E
A
G
N,T
(200 W)
3.800
3.525
3.600

60 Meters (5.3 MHz)

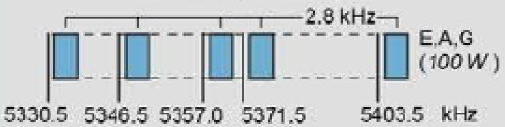
2.8 kHz
E,A,G
(100 W)
5330.5
5346.5
5357.0
5371.5
5403.5
kHz

General, Advanced, and Amateur Extra licensees may operate on these five channels on a secondary basis with a maximum effective radiated output of 100 W PEP. Permitted operating modes include upper sideband voice (USB), CW, RTTY, PSK31 and other digital modes such as PACTOR III as defined by the FCC Report and Order of November 18, 2011. USB is limited to 2.8 kHz centered on 5332, 5348, 5358.5, 5373 and 5405 kHz. CW and digital emissions must be centered 1.5 kHz above the channel frequencies indicated above. Only one signal at a time is permitted on any channel.

40 Meters (7 MHz)

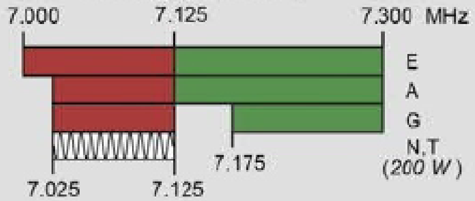
7.000
7.125
7.300 MHz
E
A
G
N,T
(200 W)
7.175
7.025
7.125

Phone and Image modes are permitted between 7.075 and 7.100 MHz for FCC licensed stations in ITU Regions 1 and 3 and by FCC licensed stations in ITU Region 2 West of 130 degrees West longitude or South of 20 degrees North latitude. See Sections 97.305(c) and 97.307(f)(11).
Novice and Technician licensees outside ITU Region 2 may use CW only between 7.025 and 7.075 MHz and between 7.100 and 7.125 MHz. 7.200 to 7.300 MHz is not available outside ITU Region 2. See Section 97.301(e). These exemptions do not apply to stations in the continental US.

30 Meters (10.1 MHz)

Avoid interference to fixed services outside the US.

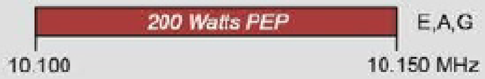
200 Watts PEP
E,A,G
10.100
10.150 MHz

20 Meters (14 MHz)

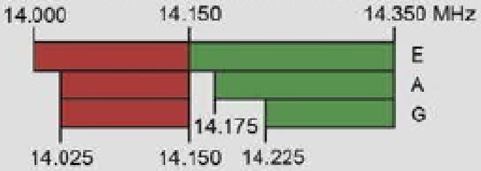
14.000
14.150
14.350 MHz
E
A
G
14.175
14.025
14.150
14.225

17 Meters (18 MHz)

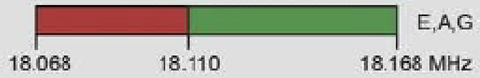
E,A,G
18.068
18.110
18.168 MHz

15 Meters (21 MHz)

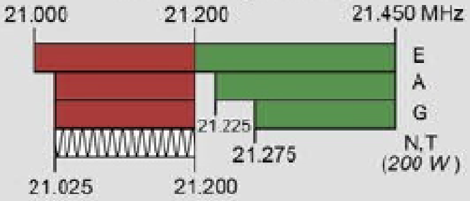
21.000
21.200
21.450 MHz
E
A
G
N,T
(200 W)
21.225
21.275
21.025
21.200

12 Meters (24 MHz)

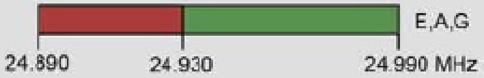
E,A,G
24.890
24.930
24.990 MHz

10 Meters (28 MHz)

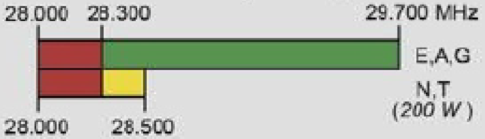
28.000
28.300
29.700 MHz
E,A,G
N,T
(200 W)
28.000
28.500

6 Meters (50 MHz)

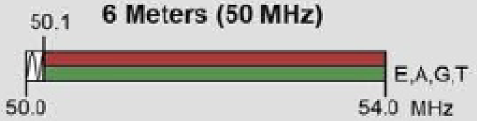
50.1
E,A,G,T
50.0
54.0 MHz

2 Meters (144 MHz)

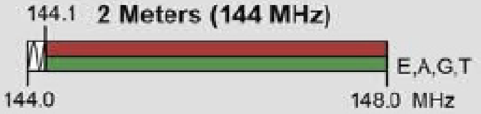
144.1
E,A,G,T
144.0
148.0 MHz

1.25 Meters (222 MHz)

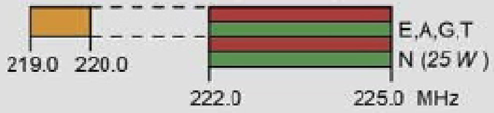
E,A,G,T
N (25 W)
219.0
220.0
222.0
225.0 MHz

*Geographical and power restrictions may apply to all bands above 420 MHz. See *The ARRL Operating Manual* for information about your area.

70 cm (420 MHz)*

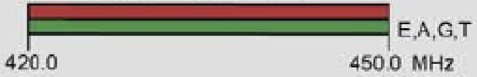
E,A,G,T
420.0
450.0 MHz

33 cm (902 MHz)*

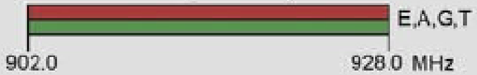
E,A,G,T
902.0
928.0 MHz

23 cm (1240 MHz)*

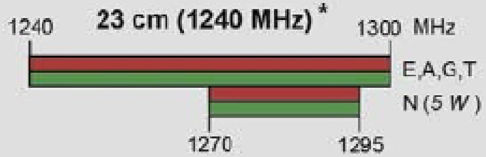
1240
1300 MHz
E,A,G,T
N (5 W)
1270
1295

All licensees except Novices are authorized all modes on the following frequencies:

2300-2310 MHz	10.0-10.5 GHz *	122.25-123.0 GHz
2390-2450 MHz	24.0-24.25 GHz	134-141 GHz
3300-3500 MHz	47.0-47.2 GHz	241-250 GHz
5650-5925 MHz	76.0-81.0 GHz	All above 275 GHz

* No pulse emissions

KEY

Note:
CW operation is permitted throughout all amateur bands.
MCW is authorized above 50.1 MHz, except for 144.0-144.1 and 219-220 MHz.
Test transmissions are authorized above 51 MHz, except for 219-220 MHz.

🟥 = RTTY and data
🟩 = phone and image
▨ = CW *only*
🟨 = SSB phone
🟦 = USB phone, CW, RTTY, and data
🟧 = Fixed digital message forwarding systems *only*

E = Amateur Extra
A = Advanced
G = General
T = Technician
N = Novice

See *ARRLWeb at www.arrl.org* for detailed band plans.

ARRL
We're At Your Service

ARRL Headquarters:
860-594-0200 (Fax 860-594-0259)
email: hq@arrl.org

Publication Orders:
www.arrl.org/shop
Toll-Free 1-888-277-5289 (860-594-0355)
email: orders@arrl.org

Membership/Circulation Desk:
www.arrl.org/membership
Toll-Free 1-888-277-5289 (860-594-0338)
email: membership@arrl.org

Getting Started in Amateur Radio:
Toll-Free 1-800-326-3942 (860-594-0355)
email: newham@arrl.org

Exams: 860-594-0300 email: vec@arrl.org

Copyright © ARRL 2012 rev. 4/9/2013

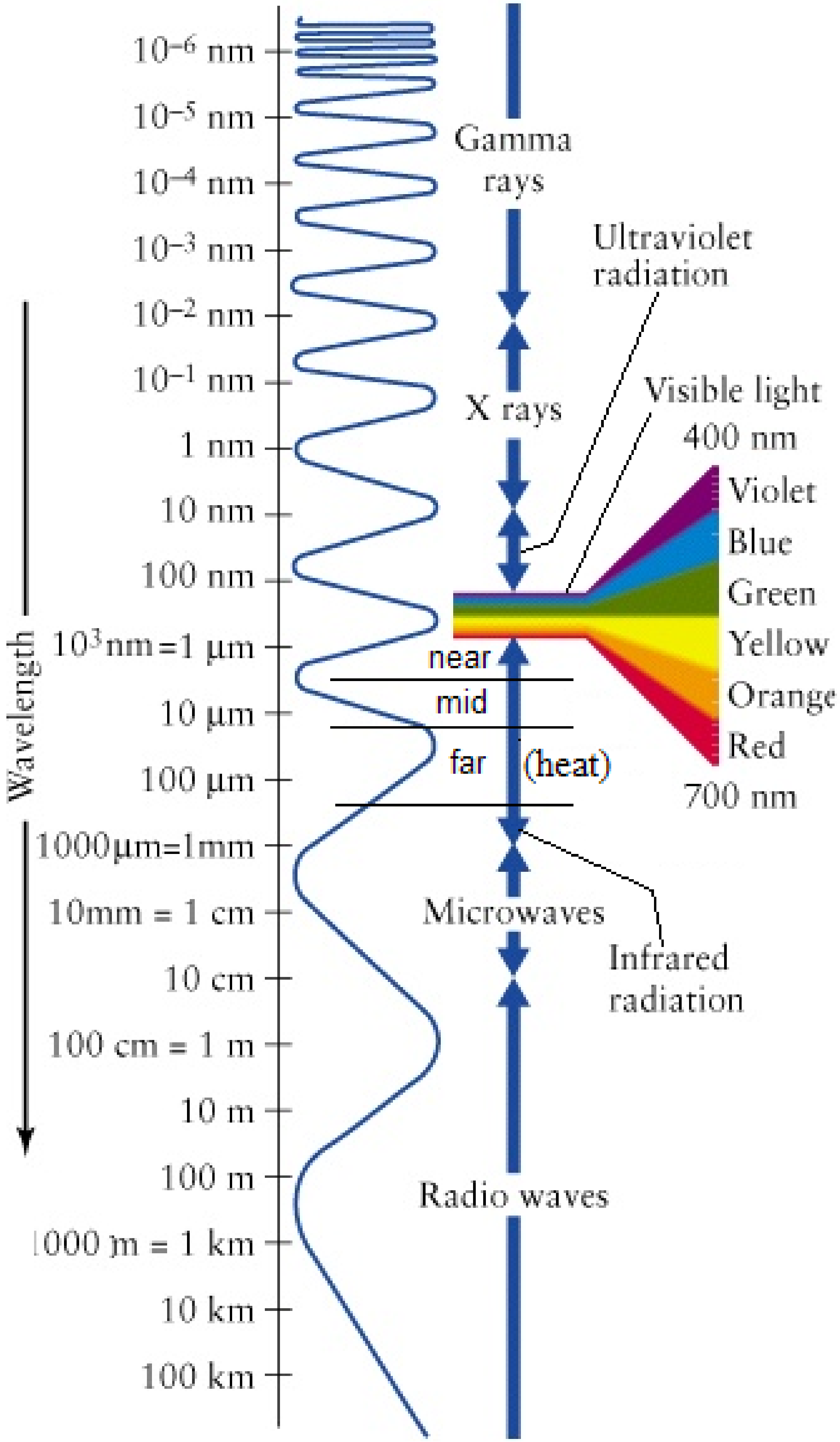
10⁻⁶ nm
10⁻⁵ nm
10⁻⁴ nm
10⁻³ nm
10⁻² nm
10⁻¹ nm
1 nm
10 nm
100 nm
10³nm=1 μm
10 μm
100 μm
1000μm=1mm
10mm = 1 cm
10 cm
100 cm = 1 m
10 m
100 m
1000 m = 1 km
10 km
100 km
Wavelength
Gamma rays
Ultraviolet radiation
X rays
Visible light
400 nm
Violet
Blue
Green
Yellow
Orange
Red
700 nm
near
mid
far
(heat)
Microwaves
Infrared radiation
Radio waves

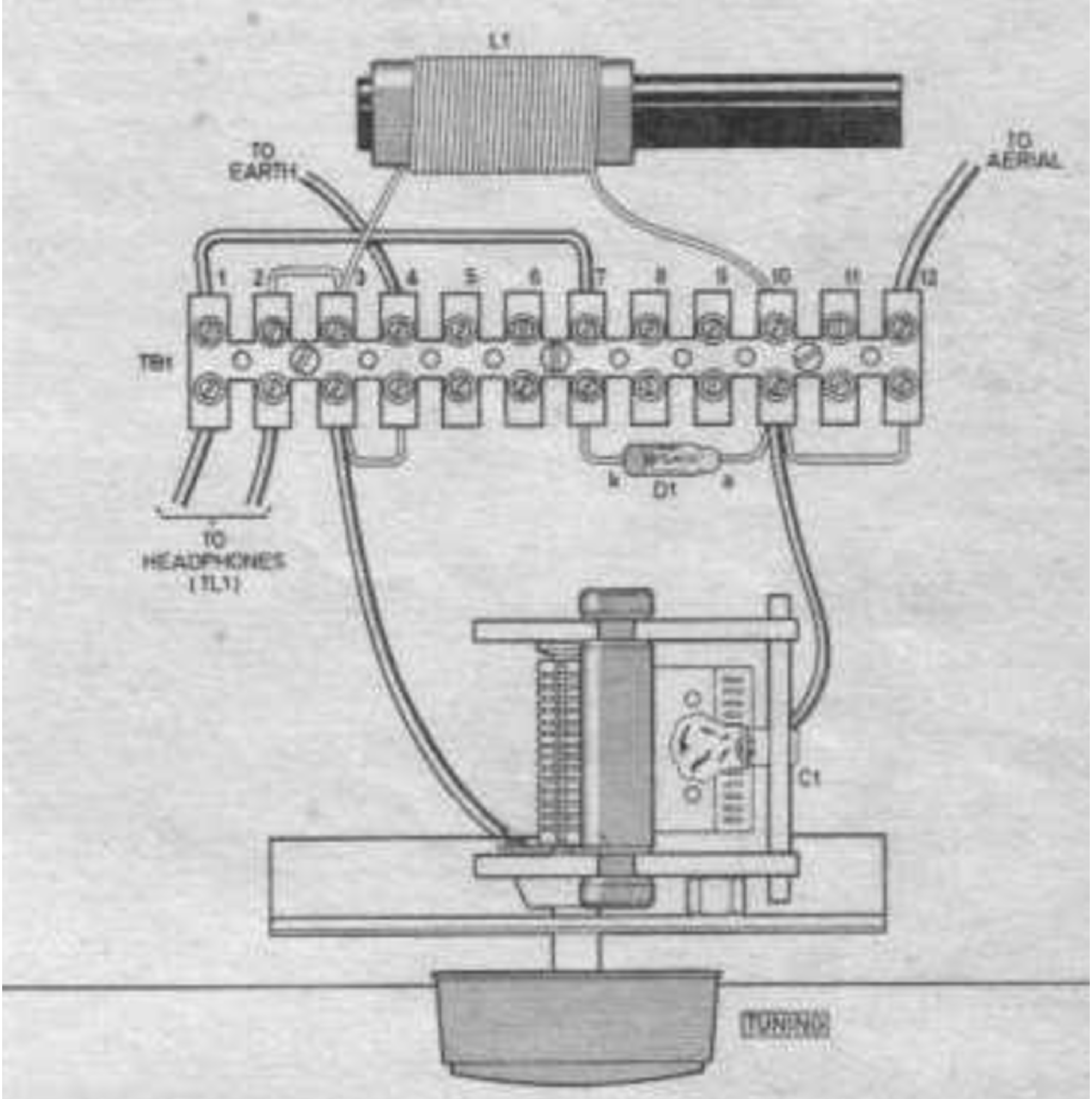
L1
TO EARTH
TO AERIAL
1
2
3
4
5
6
7
8
9
10
11
12
TB1
k
D1
a
TO HEADPHONES (TL1)
C1
TUNING

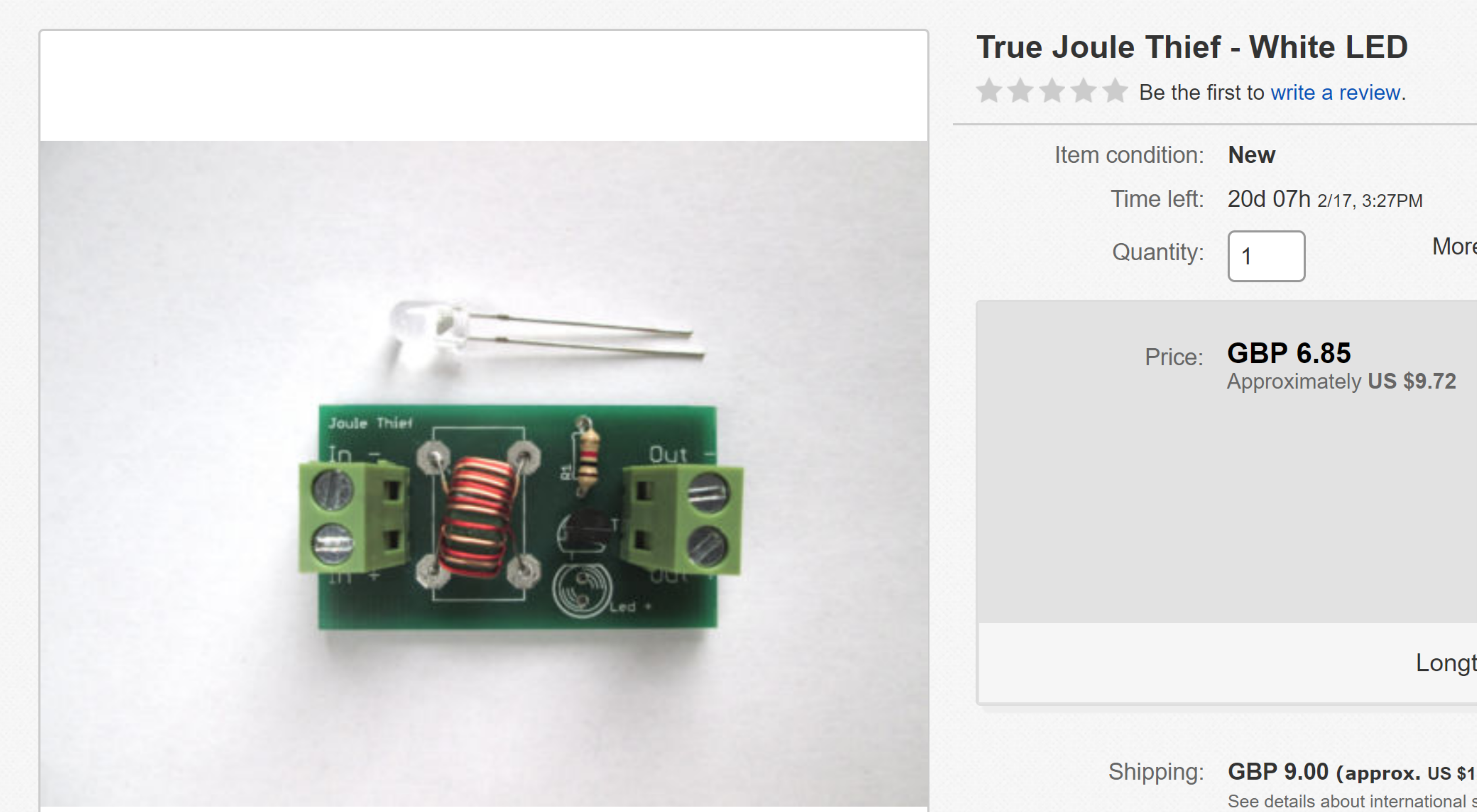

Joule Thief
In -
In +
R1
Out -
Out
Led +

True Joule Thief - White LED

☆☆☆☆☆ Be the first to write a review.

Item condition:	**New**
Time left:	20d 07h 2/17, 3:27PM
Quantity:	1
Price:	**GBP 6.85** Approximately **US $9.72**
Shipping:	**GBP 9.00 (approx. US $1** See details about international

Mor

Longt

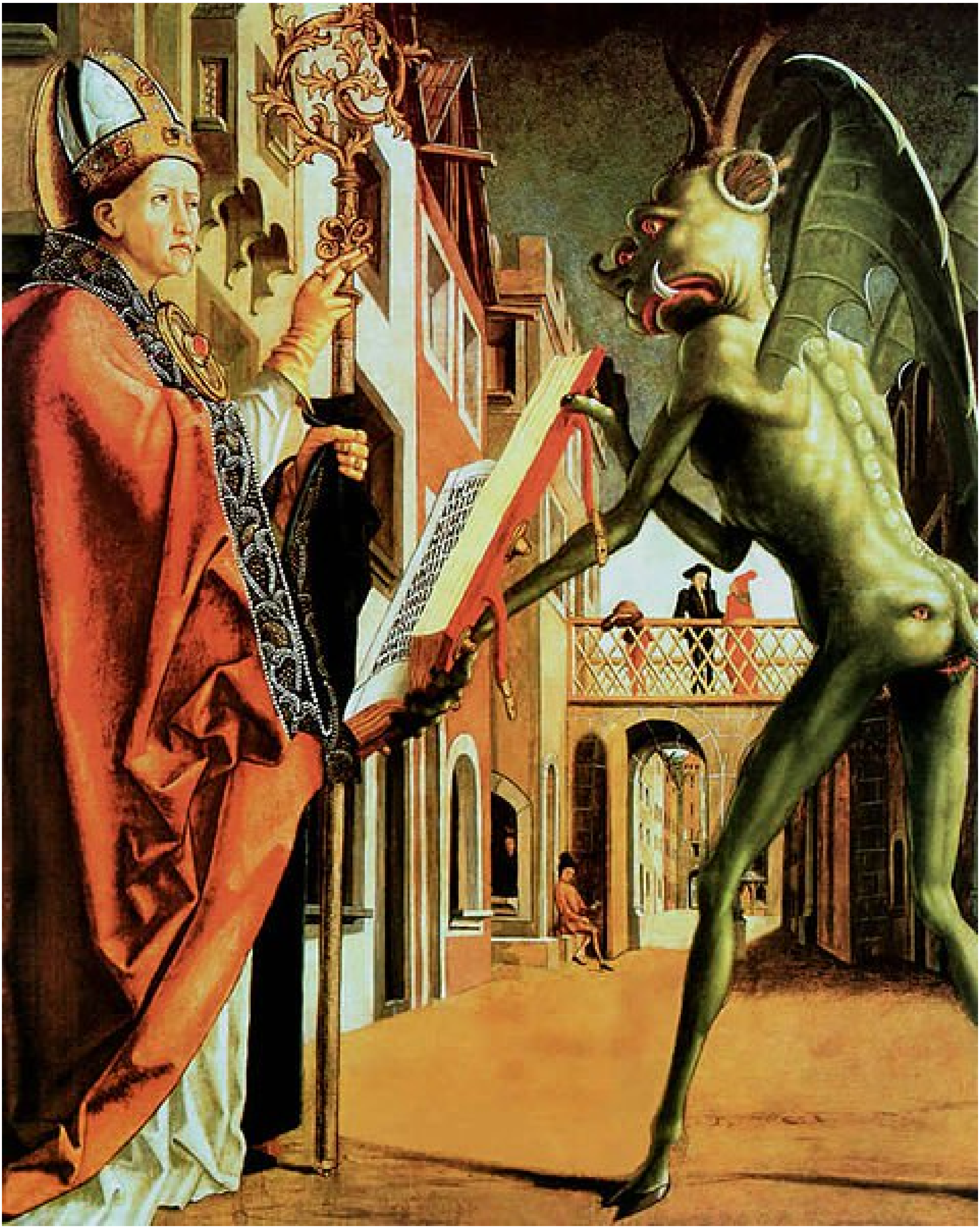

OLD WORLD ORDER - NEW WORLD ORDER

The Trinity of Globalist Control

THERE IS NO DEMOCRACY HERE

Vatican City

Religion

City of London

Finance

Washington D.C.

Military

NO FREEDOM FOR MIND OR BODY

All three are separate states,
completely independent of their respective countries.

WORLD SUPER POWERS

atruthsoldier.wordpress.com

City of London

Finance

Vatican City

Religion

Washington D.C.

Military

All three are separate states, completely
independent of their respective countries.

what do these 3 cities have in common?

City of London

Washington
District of Columbia

Vatican City

- they pay no taxes
- they are under no national authority
- they have totally Independent Identities from the rest of the world

SPEAK FOR YOURSELF

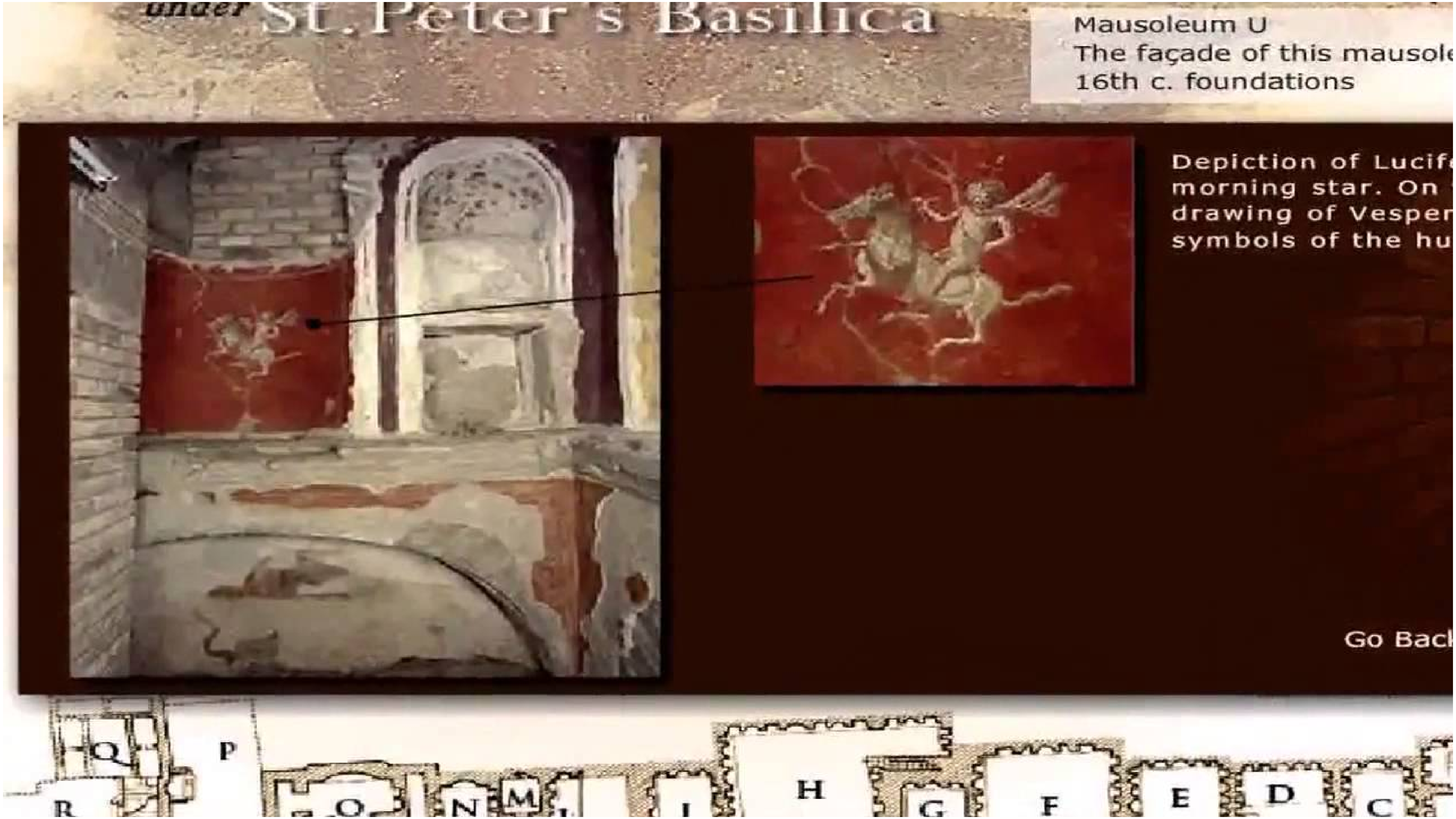

under St. Peter's Basilica

Mausoleum U
The façade of this mausole
16th c. foundations

Depiction of Lucif
morning star. On
drawing of Vesper
symbols of the hu

Go Bac
R
Q
P
O
N
M
L
I
H
G
F
E
D
C

POLARSTERN
POLARSTERN

ET OVES
ΣΥ ΒΟΣΚΕΙΣ ΤΑ

ELF, GWEN Towers, and HAARP Connection
These things are pretty sinister and all over the U.S.
Click "see more" to read article

ELIGO TE
TV ES PETRVS ET SVPER
NC VNA FIDES
HINC SACERDO
O PETRE DIXISTI TV ES
VIVAT IESVS BEATVS
ΣΥ ΒΟΣΚΕΙΣ ΤΑ

THE BLACK CUBE OF
SATURN

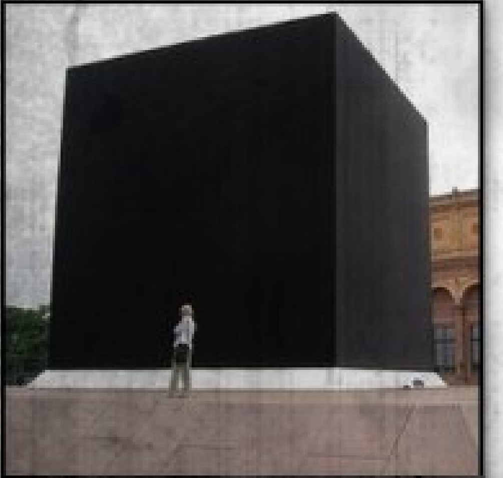
AUSTRALIA

DENMARK

MANHATTAN

SANTA ANA

GLOBAL OCCULT WORSHIP

GlacierDogMapMe: Pantheon, Vatican, Washington DC - What Were They Thinking!?

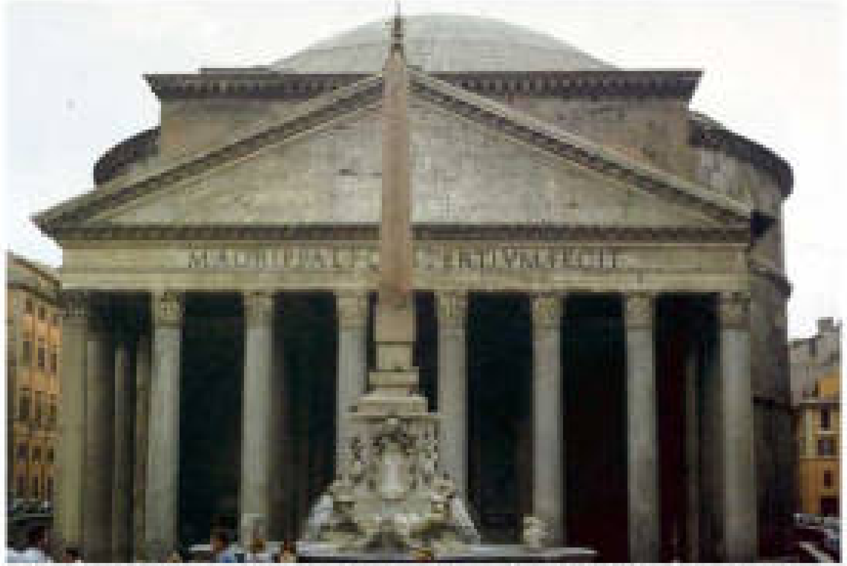

Roman Pantheon Dome facing Obelisk

http://www.historum.com/members/labienus-albums-roman-civilization-picture193-roman-pantheon.jpg;

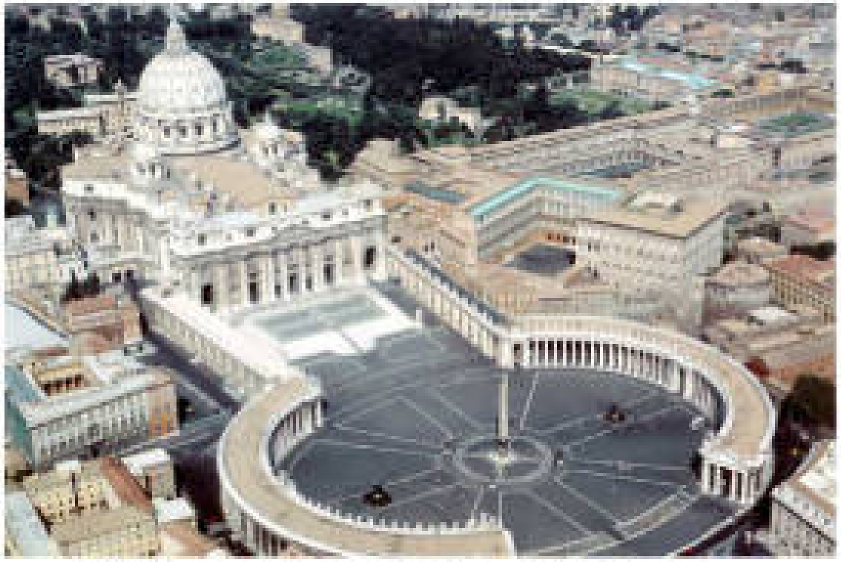

Vatican Dome facing Obelisk

http://www.raidersnewsupdate.com/leadstory70.htm

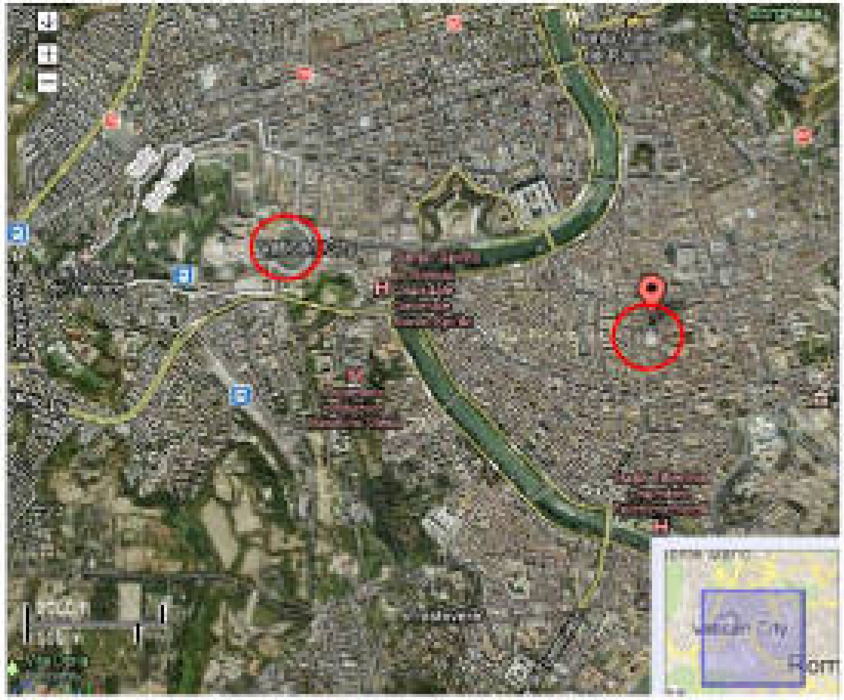

Above: Vatican City & Pantheon (red circles)

http://www.rome.info/map/

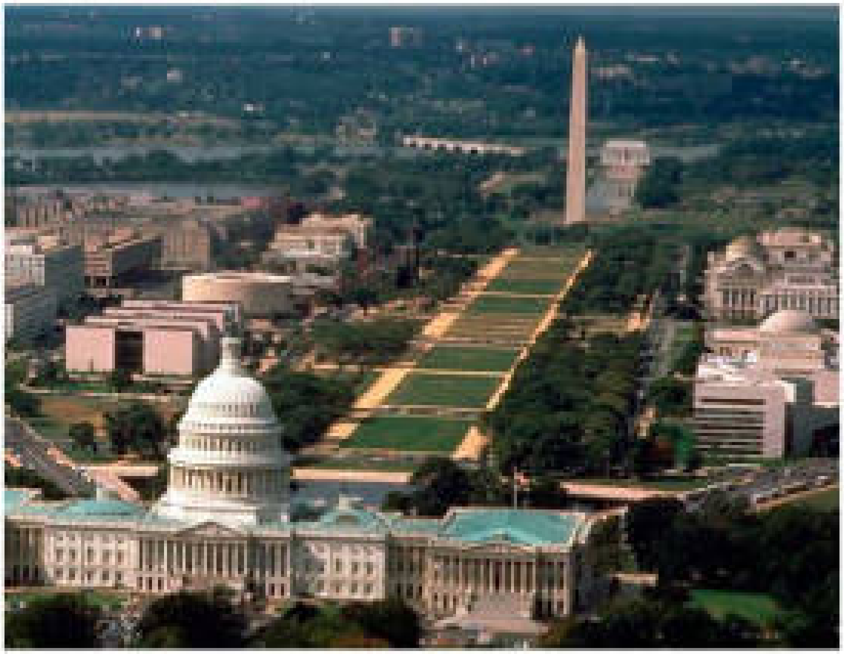

Washington DC Dome facing Obelisk

http://www.raidersnewsupdate.com/leadstory70.htm

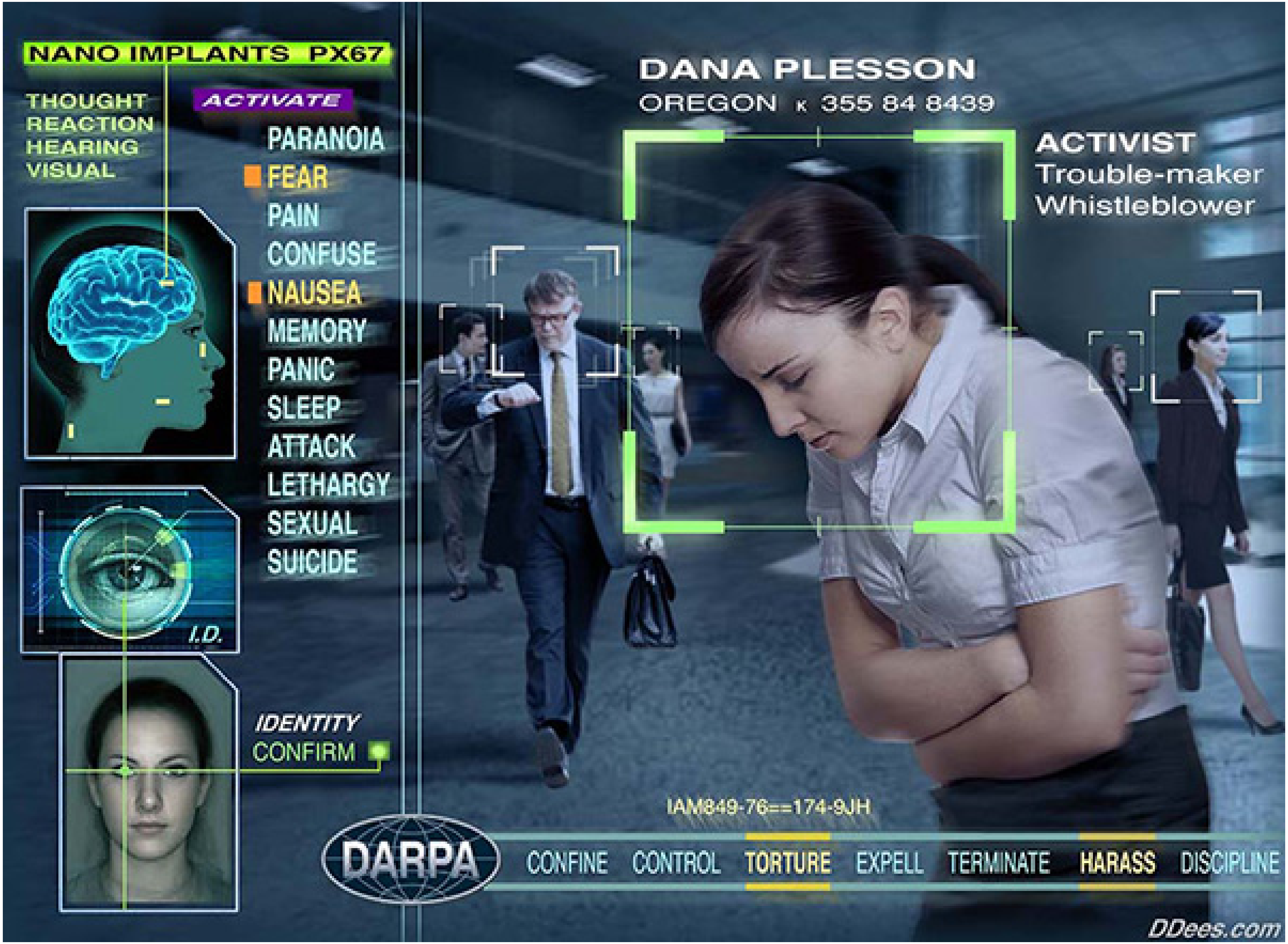
NANO IMPLANTS PX67
THOUGHT
REACTION
HEARING
VISUAL
ACTIVATE
PARANOIA
FEAR
PAIN
CONFUSE
NAUSEA
MEMORY
PANIC
SLEEP
ATTACK
LETHARGY
SEXUAL
SUICIDE
I.D.
IDENTITY
CONFIRM
DANA PLESSON
OREGON K 355 84 8439
ACTIVIST
Trouble-maker
Whistleblower
IAM849-76==174-9JH
DARPA
CONFINE
CONTROL
TORTURE
EXPELL
TERMINATE
HARASS
DISCIPLINE
DDees.com

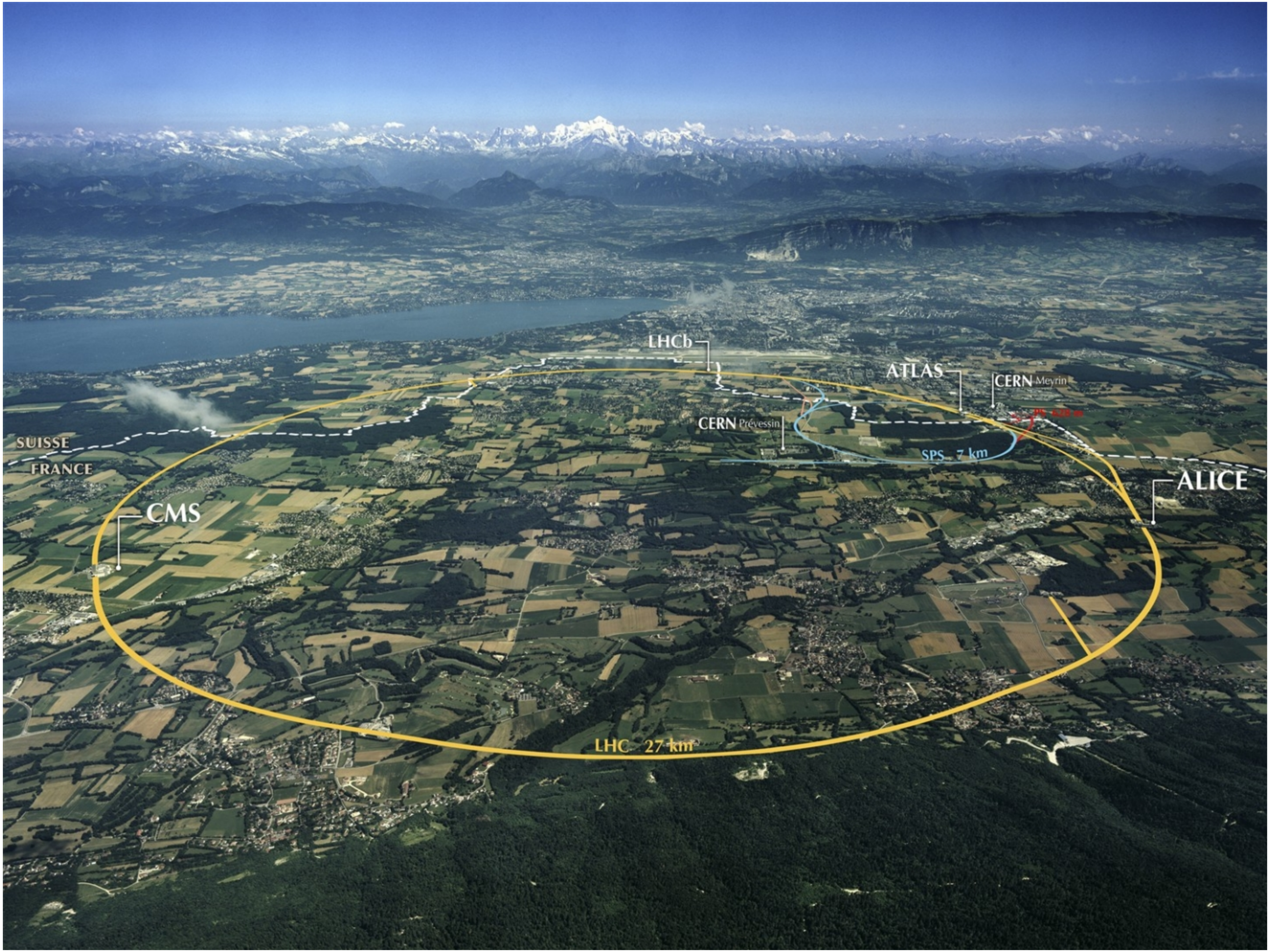
LHCb
ATLAS
CERN Meyrin
PS 628 m
CERN Prévessin
SPS 7 km
SUISSE
FRANCE
ALICE
CMS
LHC 27 km

An overview of how the weapons are deployed by the Mind Control Police

Mass Remote Mind Control by Blanket Coverage of the Population which is achieved via the Mobile Phone Network

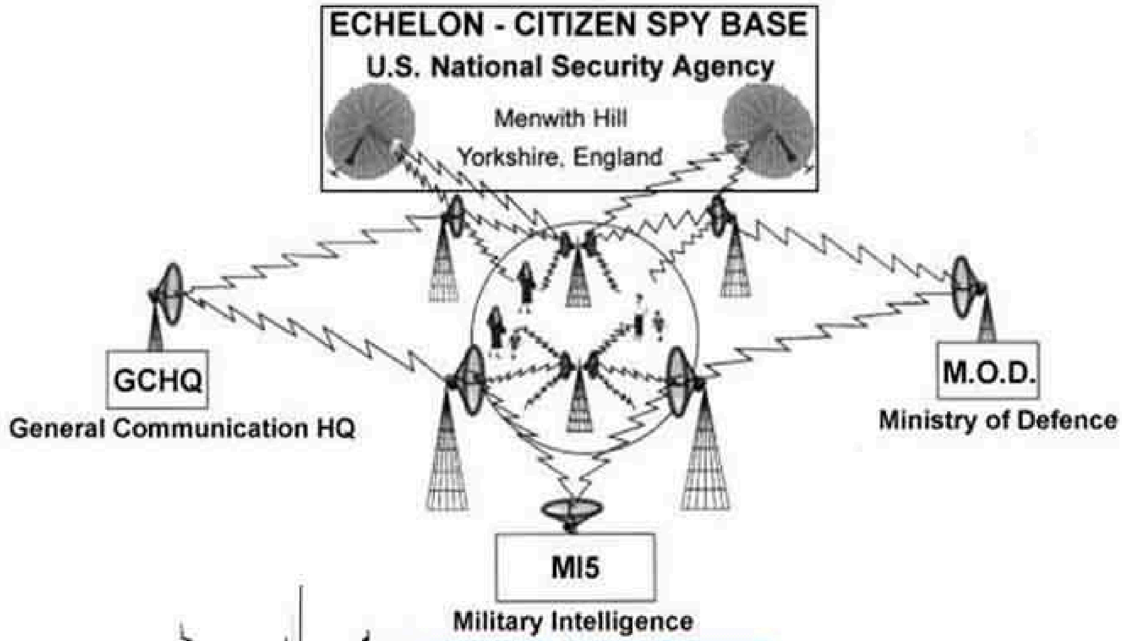
ECHELON - CITIZEN SPY BASE
U.S. National Security Agency
Menwith Hill
Yorkshire, England
GCHQ
General Communication HQ
M.O.D.
Ministry of Defence
MI5
Military Intelligence

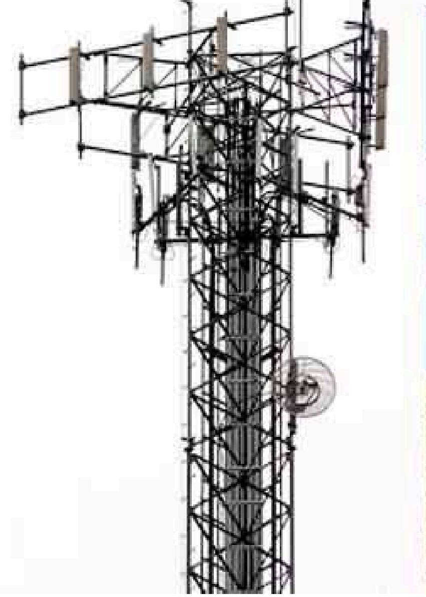

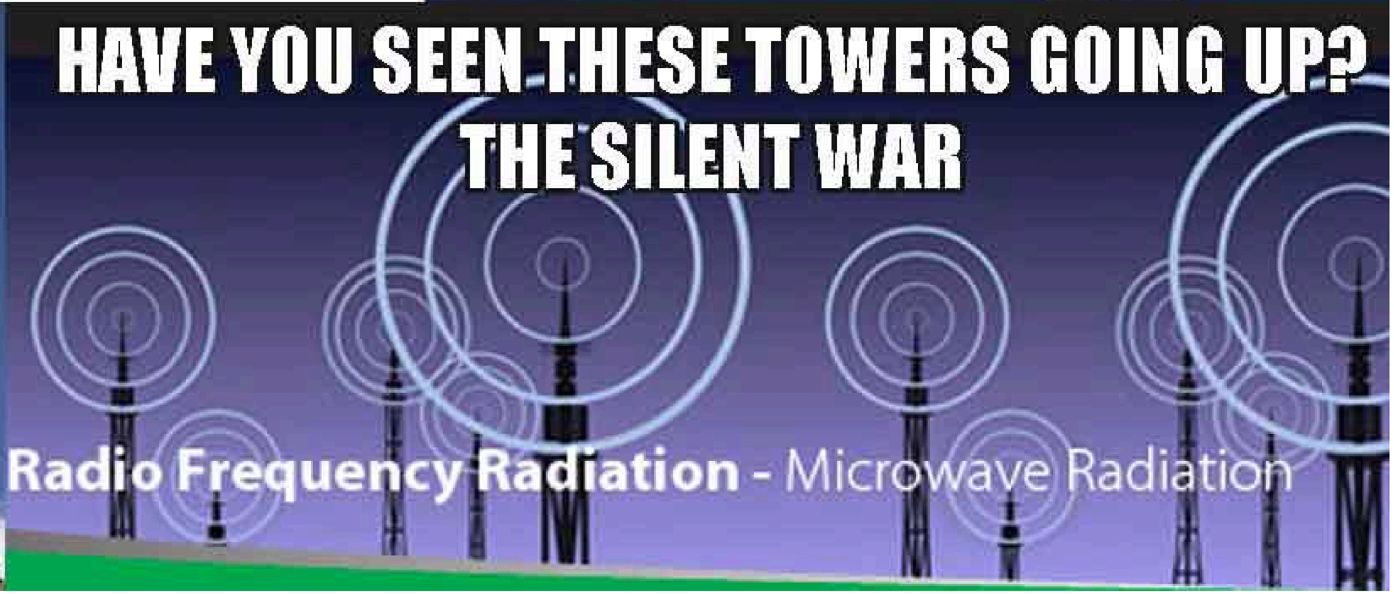
HAVE YOU SEEN THESE TOWERS GOING UP?
THE SILENT WAR
Radio Frequency Radiation - Microwave Radiation

BLACK CUBE- AUSTRALIA

Manhattan, NYC

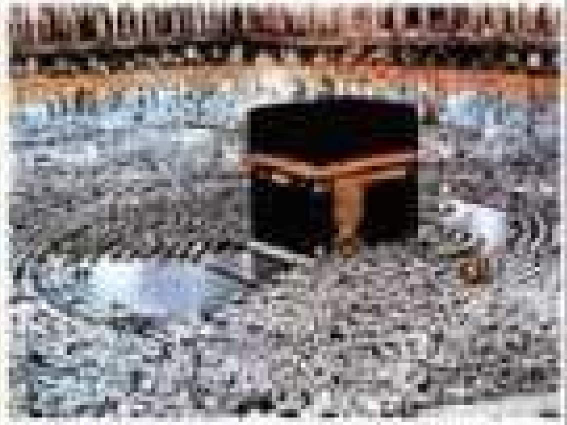

Kaba in Mecca

BLACK CUBE- DENMARK

Aple Store cube- NYC

"Rings going around SATURN"

Download from
Dreamstime.com
This watermarked comp image is for previewing purposes only.

ID 59211679
© Marinv | Dreamstime.com

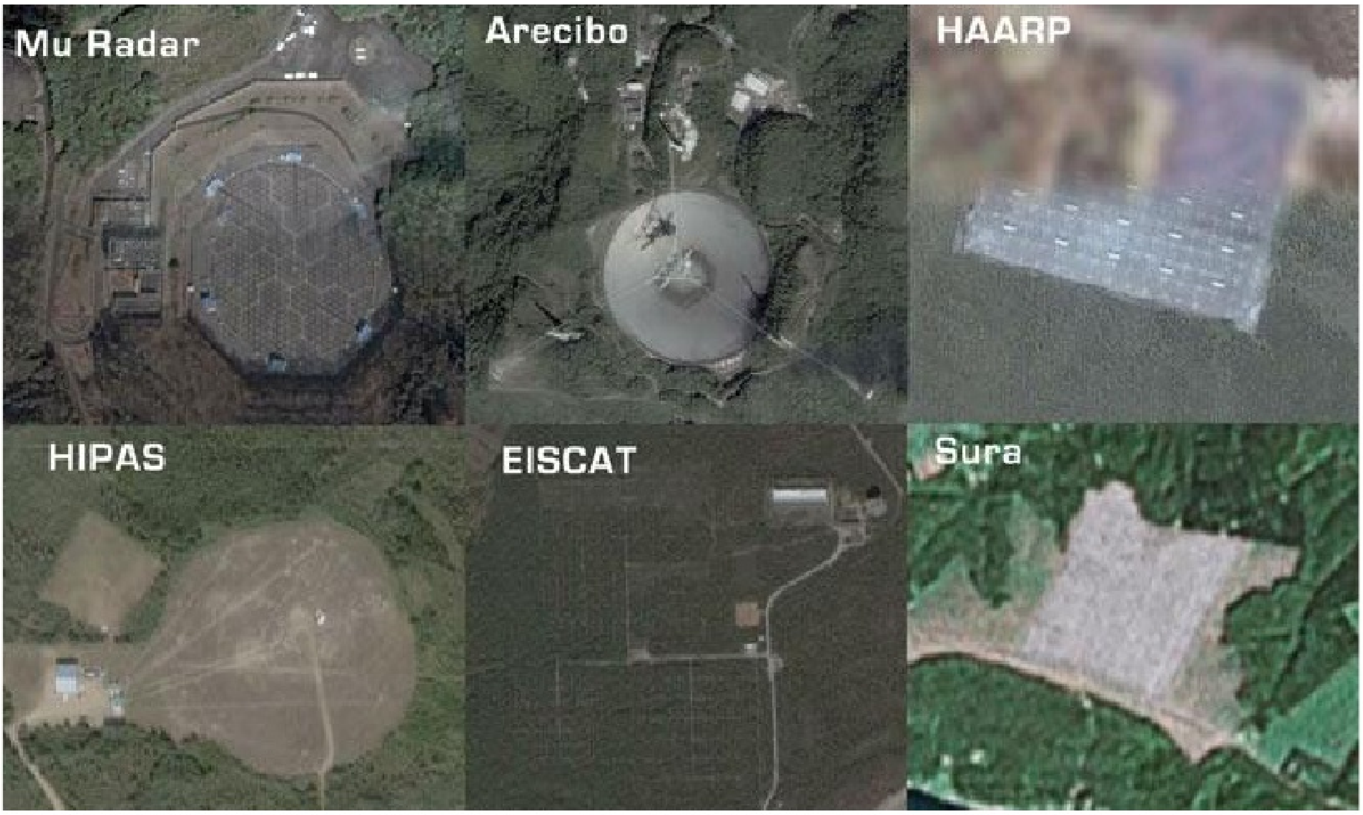
Mu Radar
Arecibo
HAARP
HIPAS
EISCAT
Sura

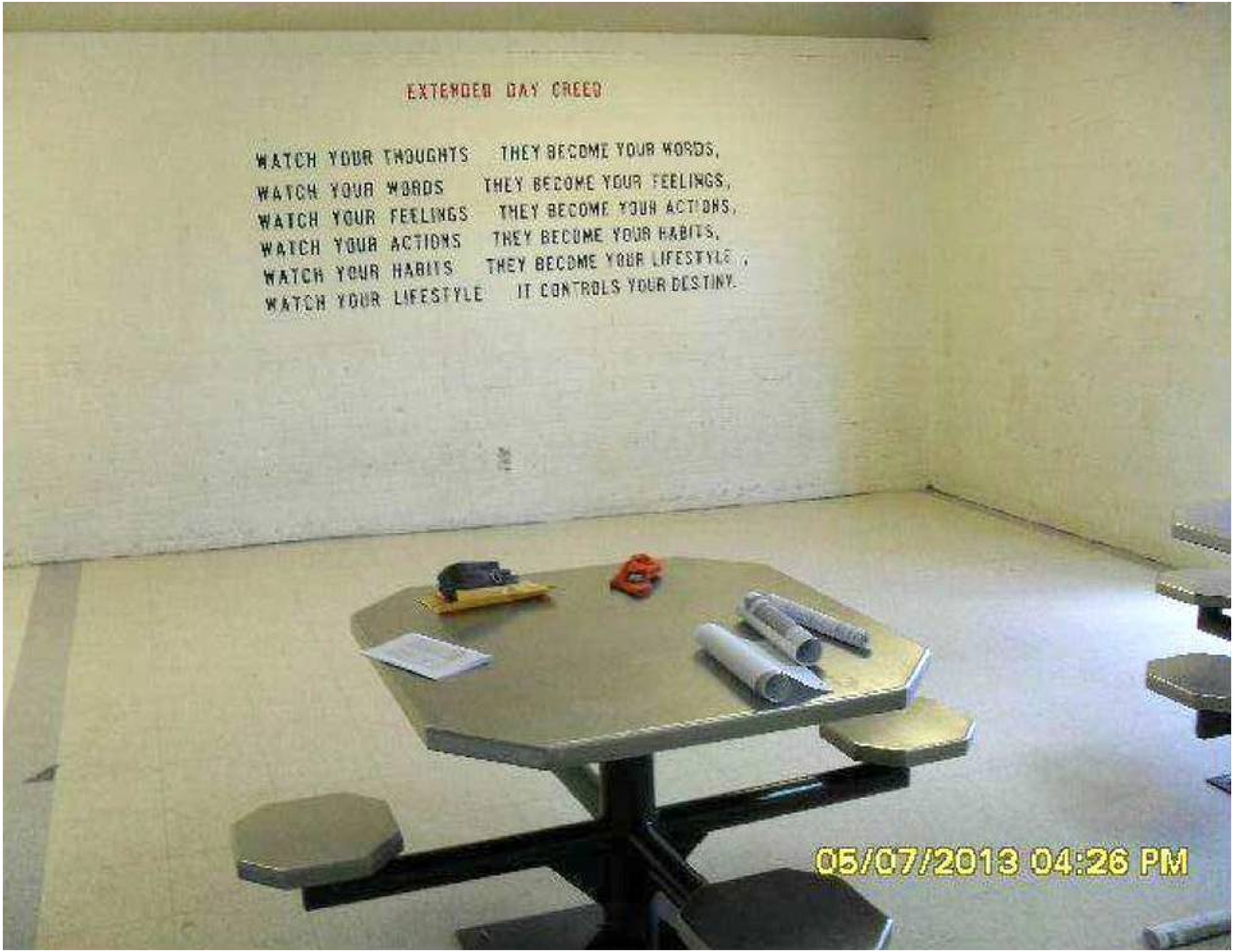
EXTENDED DAY CREED
WATCH YOUR THOUGHTS THEY BECOME YOUR WORDS,
WATCH YOUR WORDS THEY BECOME YOUR FEELINGS,
WATCH YOUR FEELINGS THEY BECOME YOUR ACTIONS,
WATCH YOUR ACTIONS THEY BECOME YOUR HABITS,
WATCH YOUR HABITS THEY BECOME YOUR LIFESTYLE,
WATCH YOUR LIFESTYLE IT CONTROLS YOUR DESTINY.

05/07/2013 04:26 PM

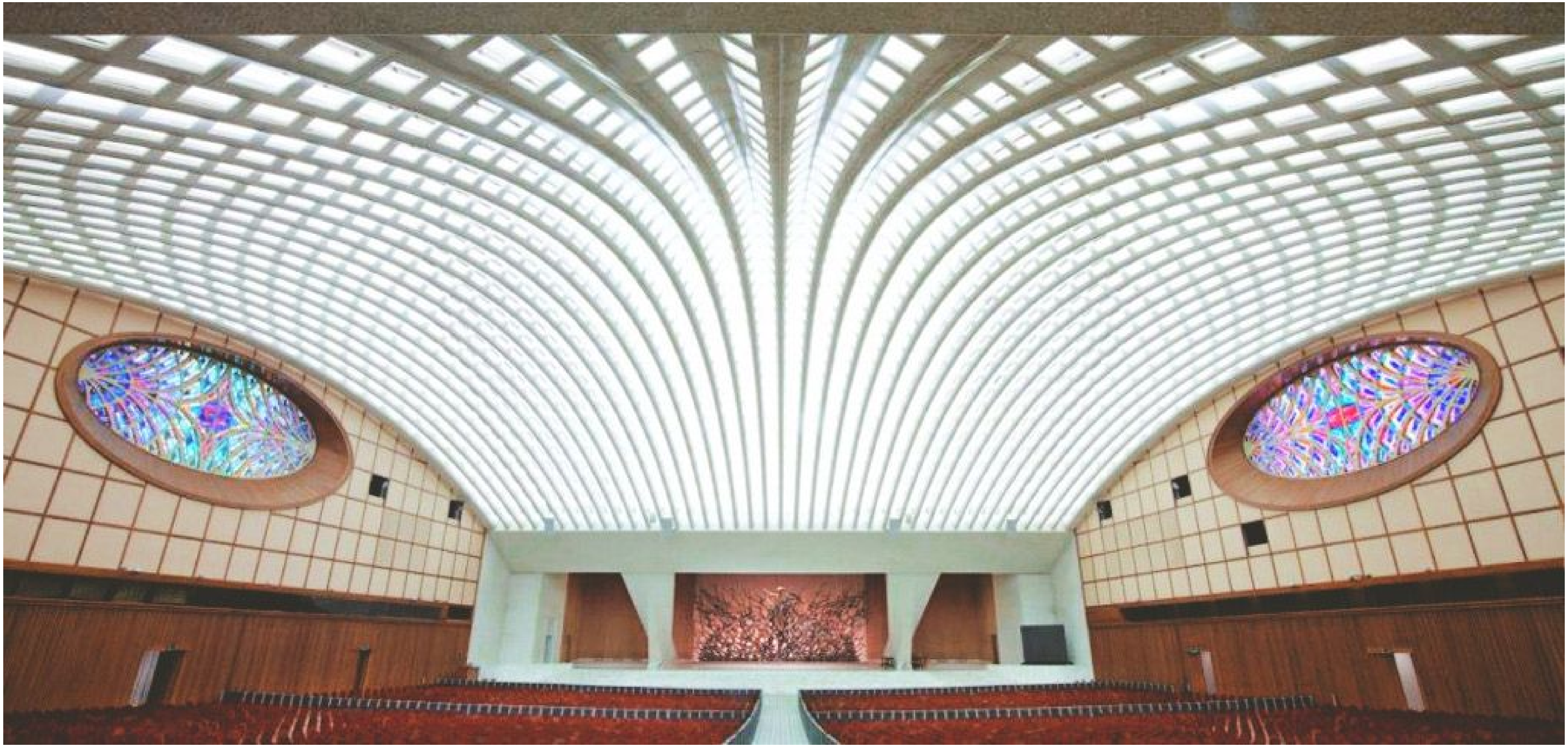

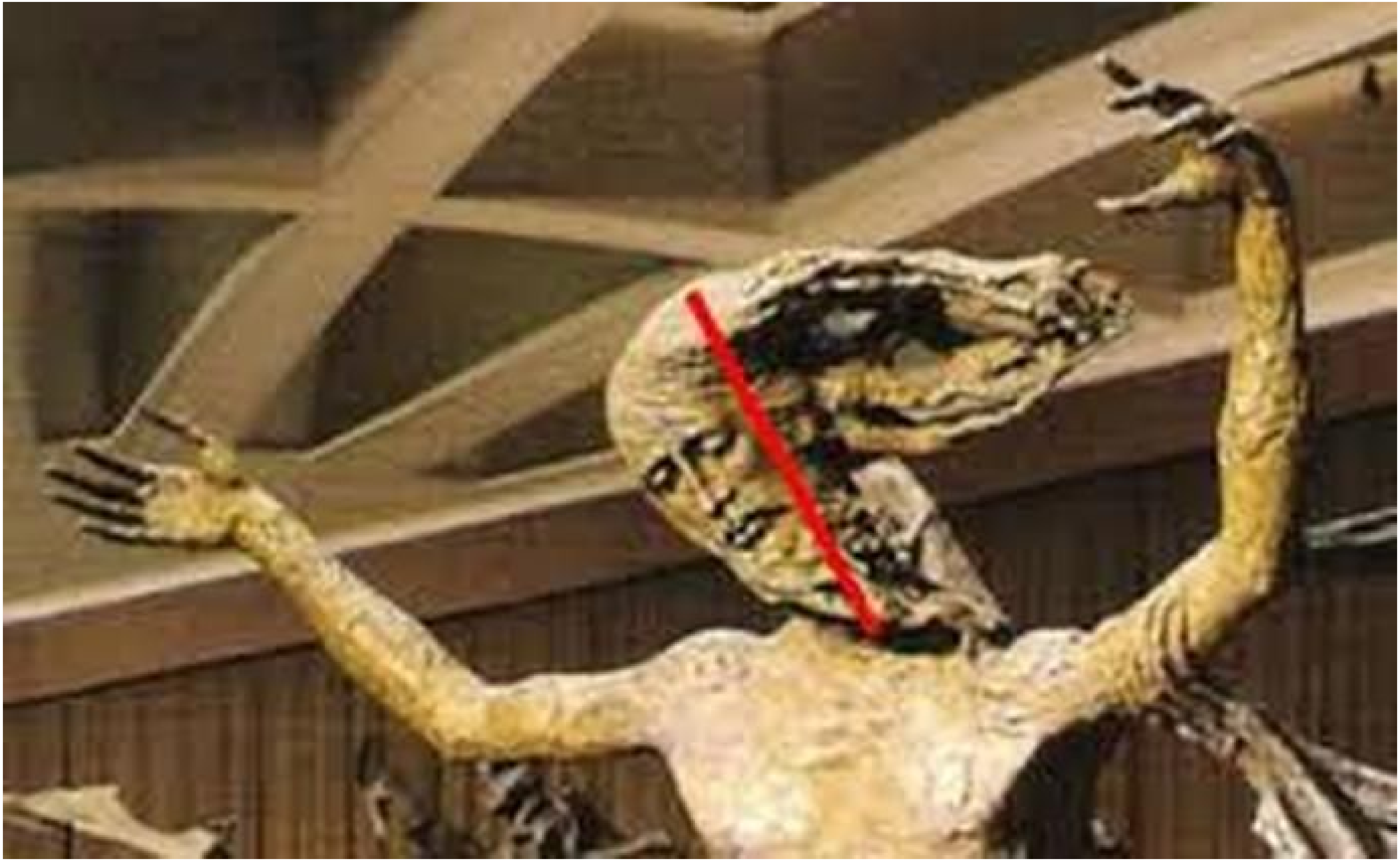

The Trinity of Globalist Control

Vatican City

Religion

City of London

Finance

Washington D.C.

Military

All three are separate states, completely independent of their respective countries.

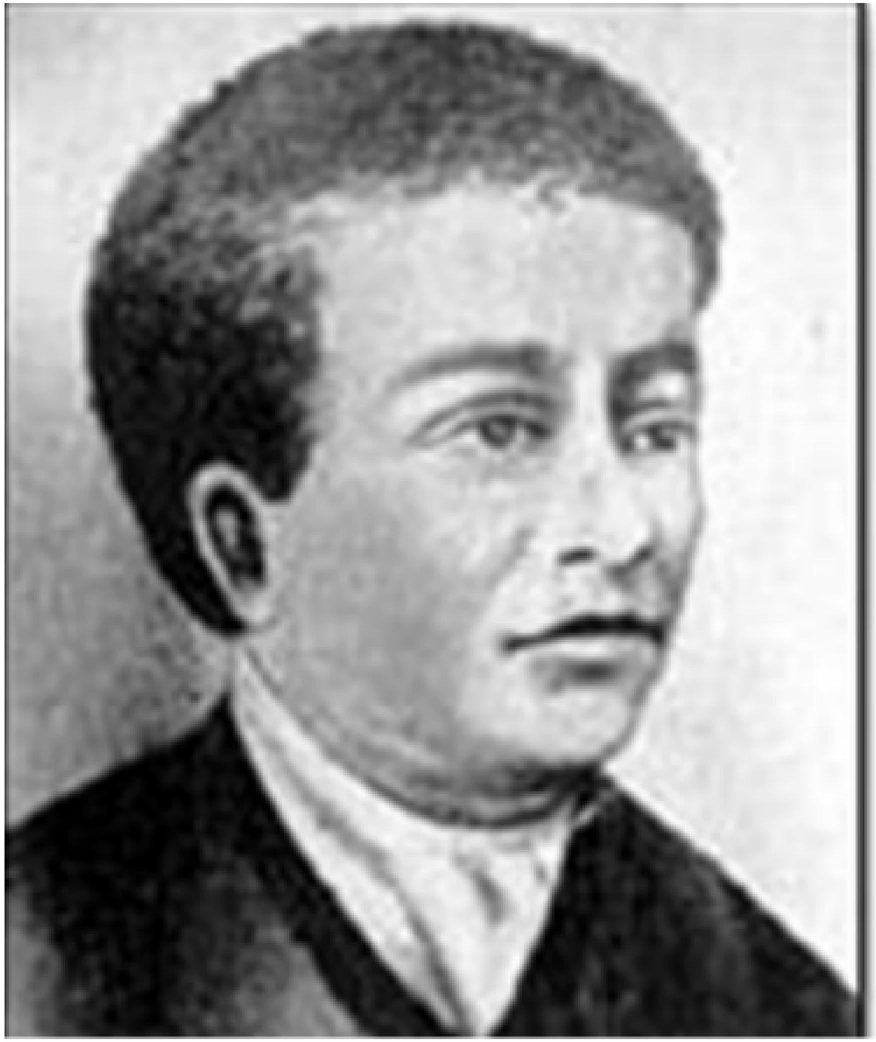

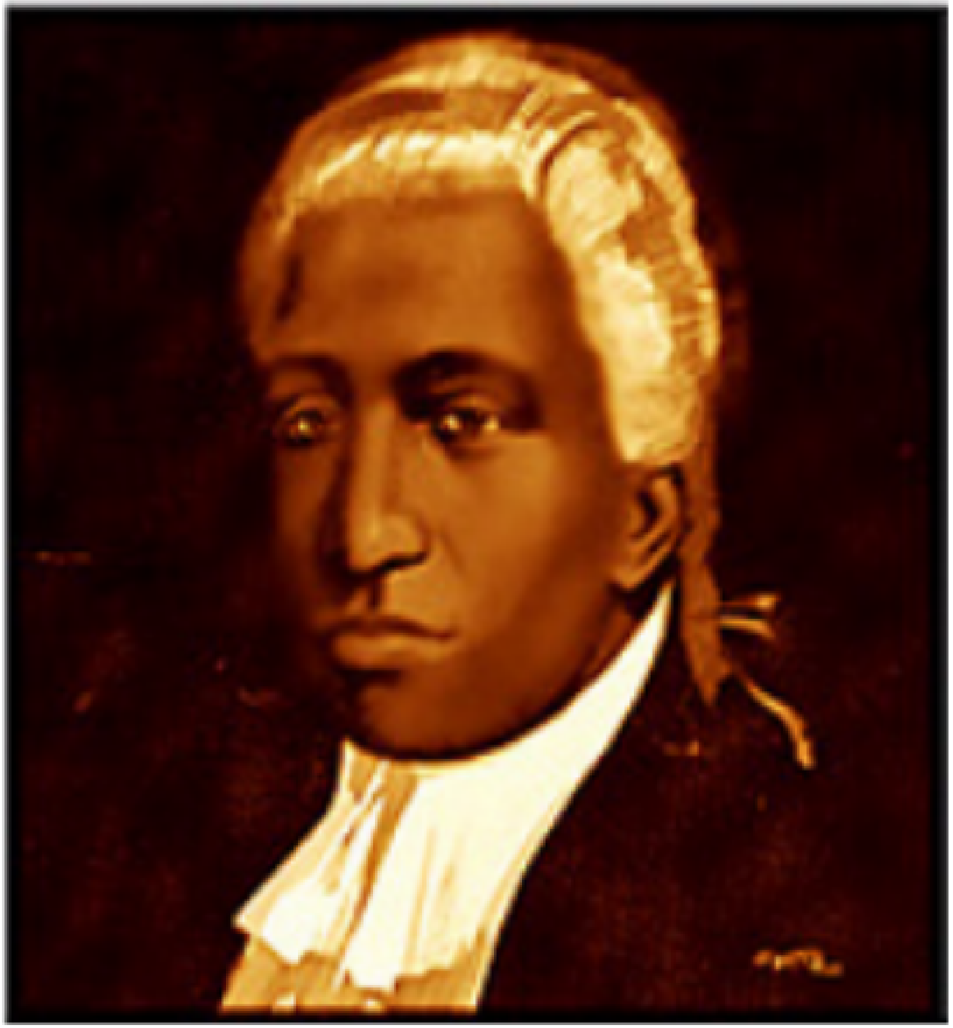

THE GREATEST OPERATIVE MASON MEETS THE GREATEST SPECTULATIVE MASON is Ben Bey Emmanuel Mu Ali, historically known as Benjamin Banneker really Prince Hall? Benjamin Banneker (November 9, 1731 – October 9, 1806) was a free African American astronomer, mathematician, surveyor, almanac author and farmer. Prince Hall (c.1735 – December 4, 1807), Hall formed the African Grand Lodge of North America.

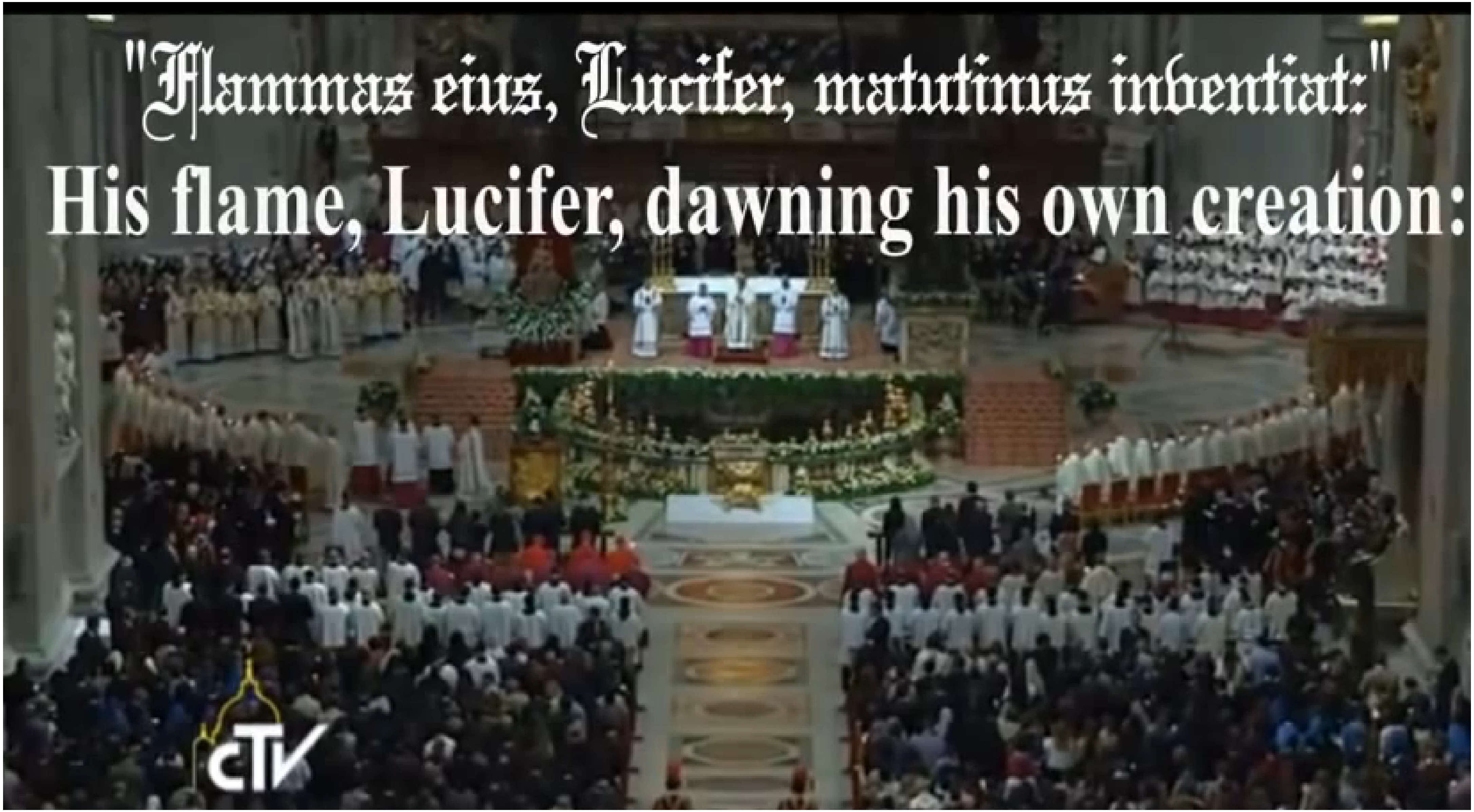
"Flammas eius, Lucifer, matutinus inveniat:"

His flame, Lucifer, dawning his own creation:

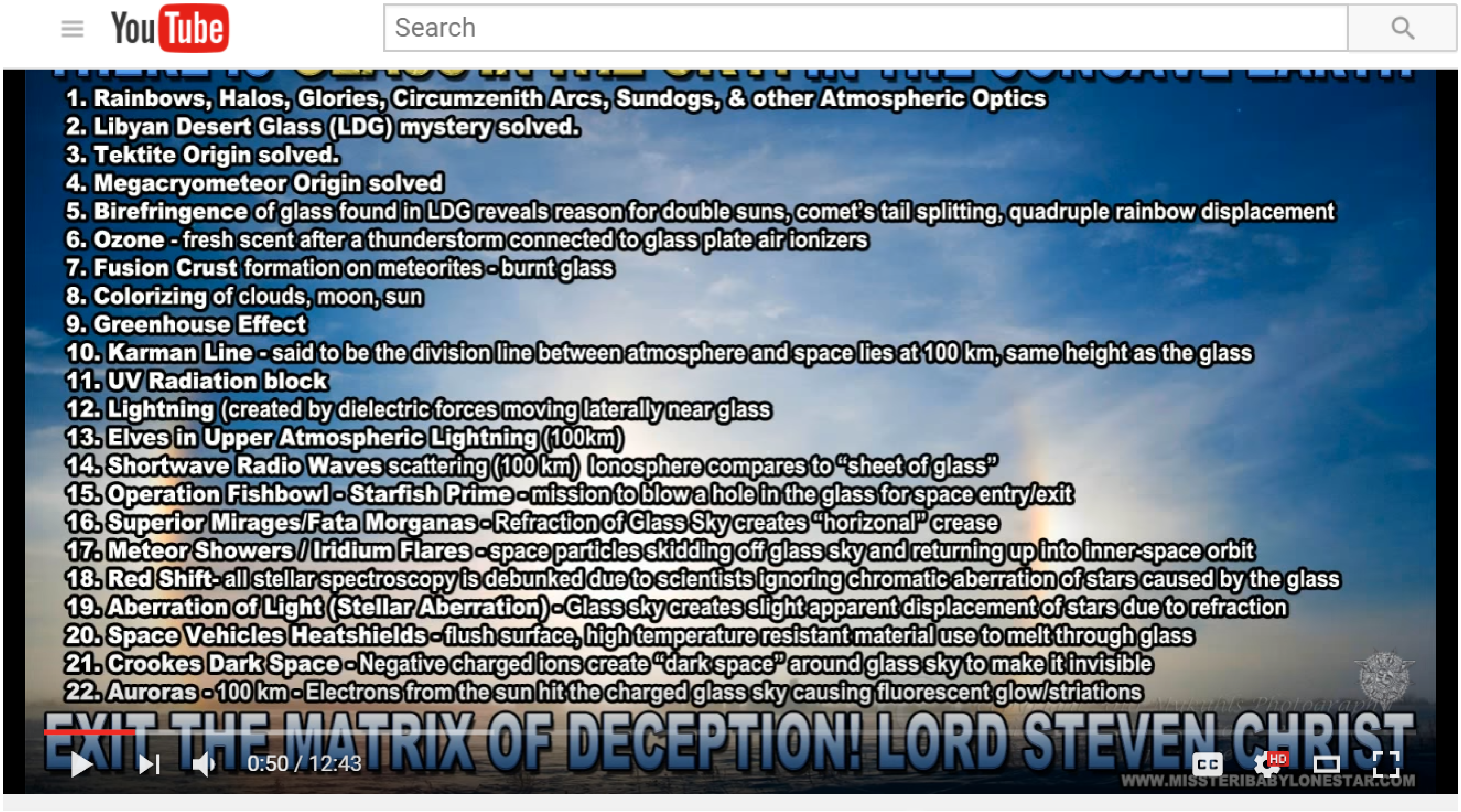

1. **Rainbows, Halos, Glories, Circumzenith Arcs, Sundogs, & other Atmospheric Optics**
2. **Libyan Desert Glass (LDG) mystery solved.**
3. **Tektite Origin solved.**
4. **Megacryometeor Origin solved**
5. **Birefringence** of glass found in LDG reveals reason for double suns, comet's tail splitting, quadruple rainbow displacement
6. **Ozone** - fresh scent after a thunderstorm connected to glass plate air ionizers
7. **Fusion Crust** formation on meteorites - burnt glass
8. **Colorizing** of clouds, moon, sun
9. **Greenhouse Effect**
10. **Karman Line** - said to be the division line between atmosphere and space lies at 100 km, same height as the glass
11. **UV Radiation block**
12. **Lightning** (created by dielectric forces moving laterally near glass
13. **Elves in Upper Atmospheric Lightning** (100km)
14. **Shortwave Radio Waves** scattering (100 km) Ionosphere compares to "sheet of glass"
15. **Operation Fishbowl - Starfish Prime** - mission to blow a hole in the glass for space entry/exit
16. **Superior Mirages/Fata Morganas** - Refraction of Glass Sky creates "horizonal" crease
17. **Meteor Showers / Iridium Flares** - space particles skidding off glass sky and returning up into inner-space orbit
18. **Red Shift**- all stellar spectroscopy is debunked due to scientists ignoring chromatic aberration of stars caused by the glass
19. **Aberration of Light (Stellar Aberration)** - Glass sky creates slight apparent displacement of stars due to refraction
20. **Space Vehicles Heatshields** - flush surface, high temperature resistant material use to melt through glass
21. **Crookes Dark Space** - Negative charged ions create "dark space" around glass sky to make it invisible
22. **Auroras** - 100 km - Electrons from the sun hit the charged glass sky causing fluorescent glow/striations

EXIT THE MATRIX OF DECEPTION! LORD STEVEN CHRIST

WWW.MISSTERIBABYLONESTAR.COM

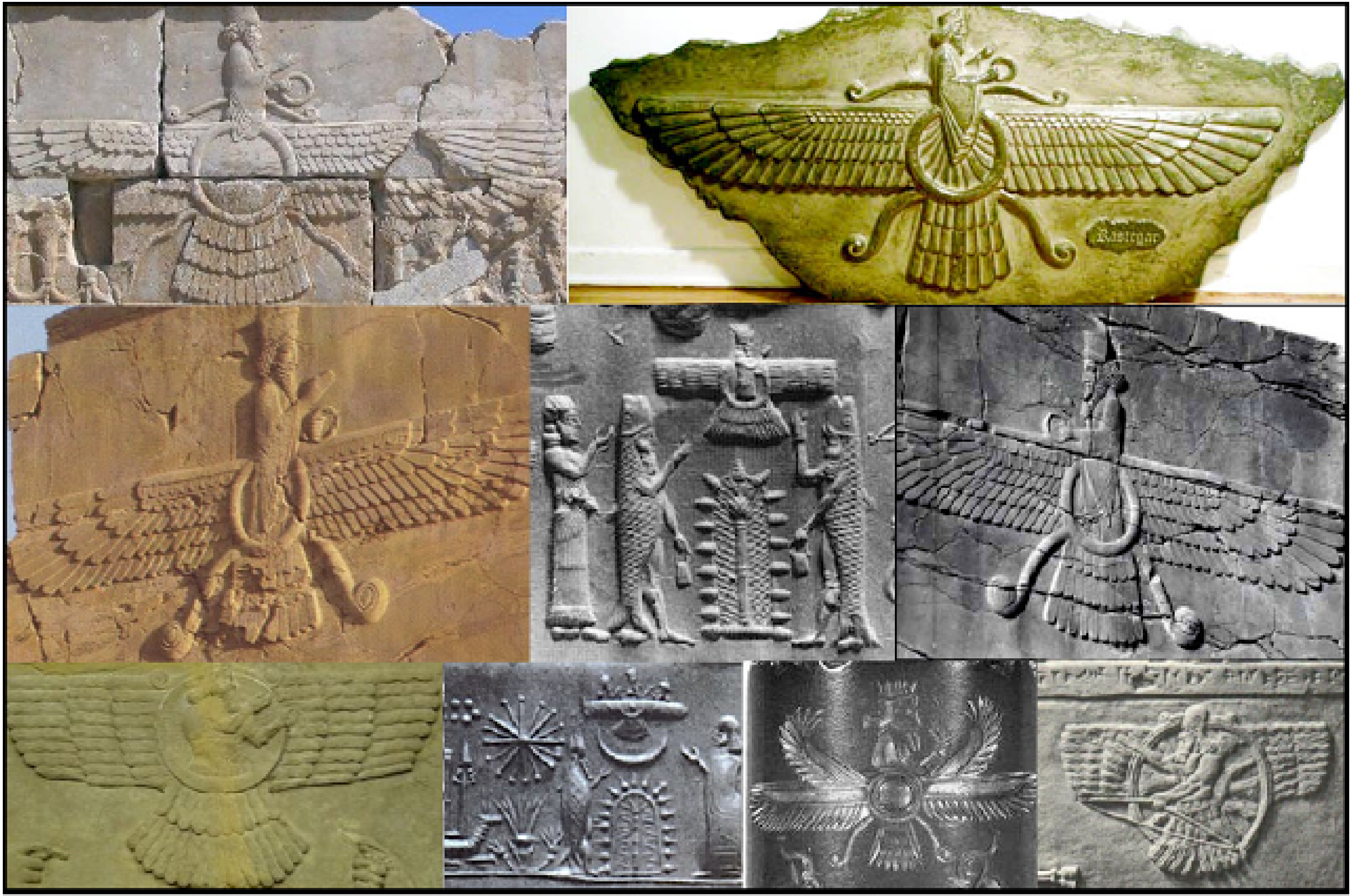

SUMERIAN DEPICTIONS OF ANUNNAKI GODS FLYING IN "WHIRLWINDS"

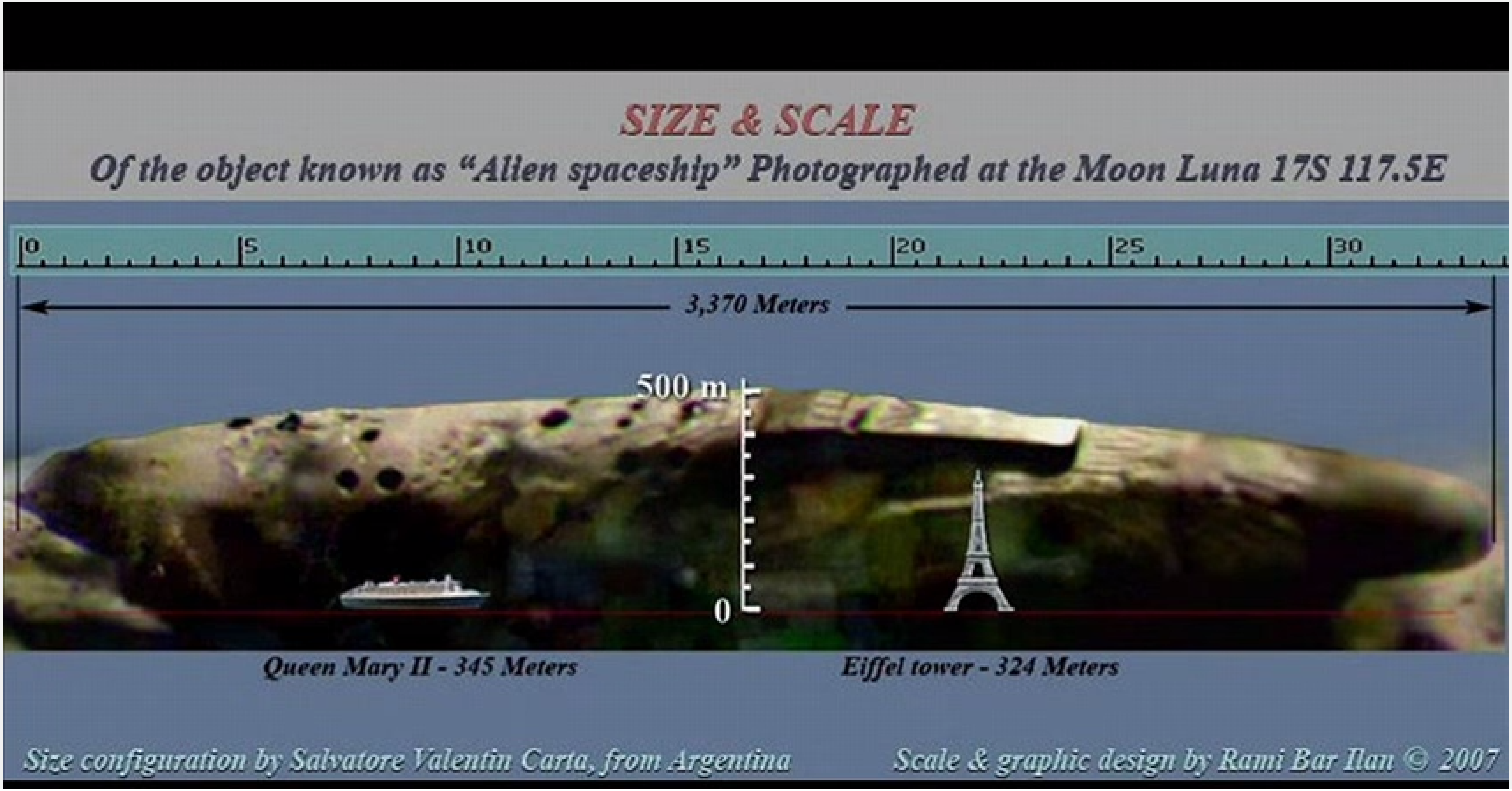

SIZE & SCALE

Of the object known as "Alien spaceship" Photographed at the Moon Luna 17S 117.5E

0
5
10
15
20
25
30
3,370 Meters
500 m
0
Queen Mary II - 345 Meters
Eiffel tower - 324 Meters

Size configuration by Salvatore Valentin Carta, from Argentina

Scale & graphic design by Rami Bar Ilan © 2007

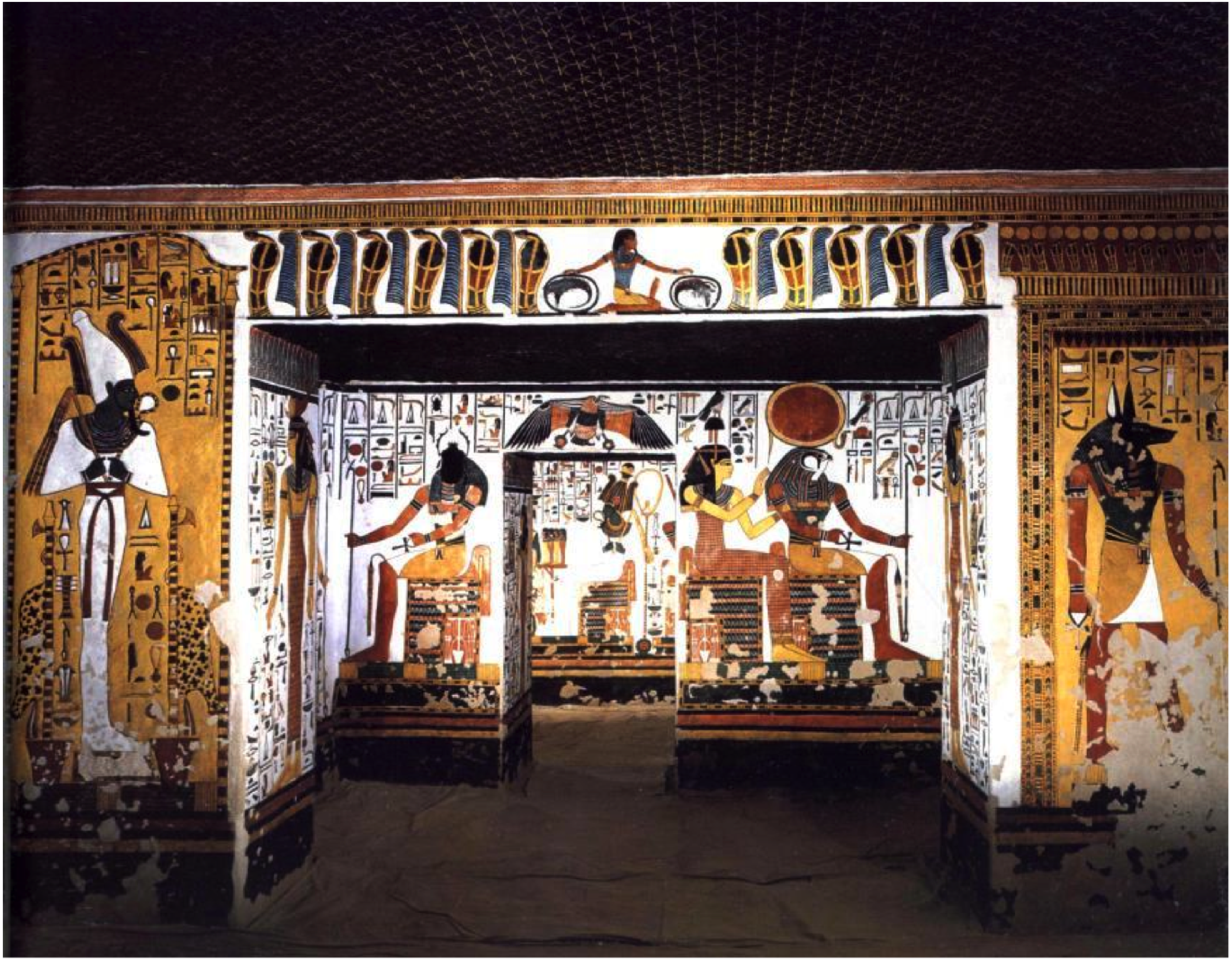

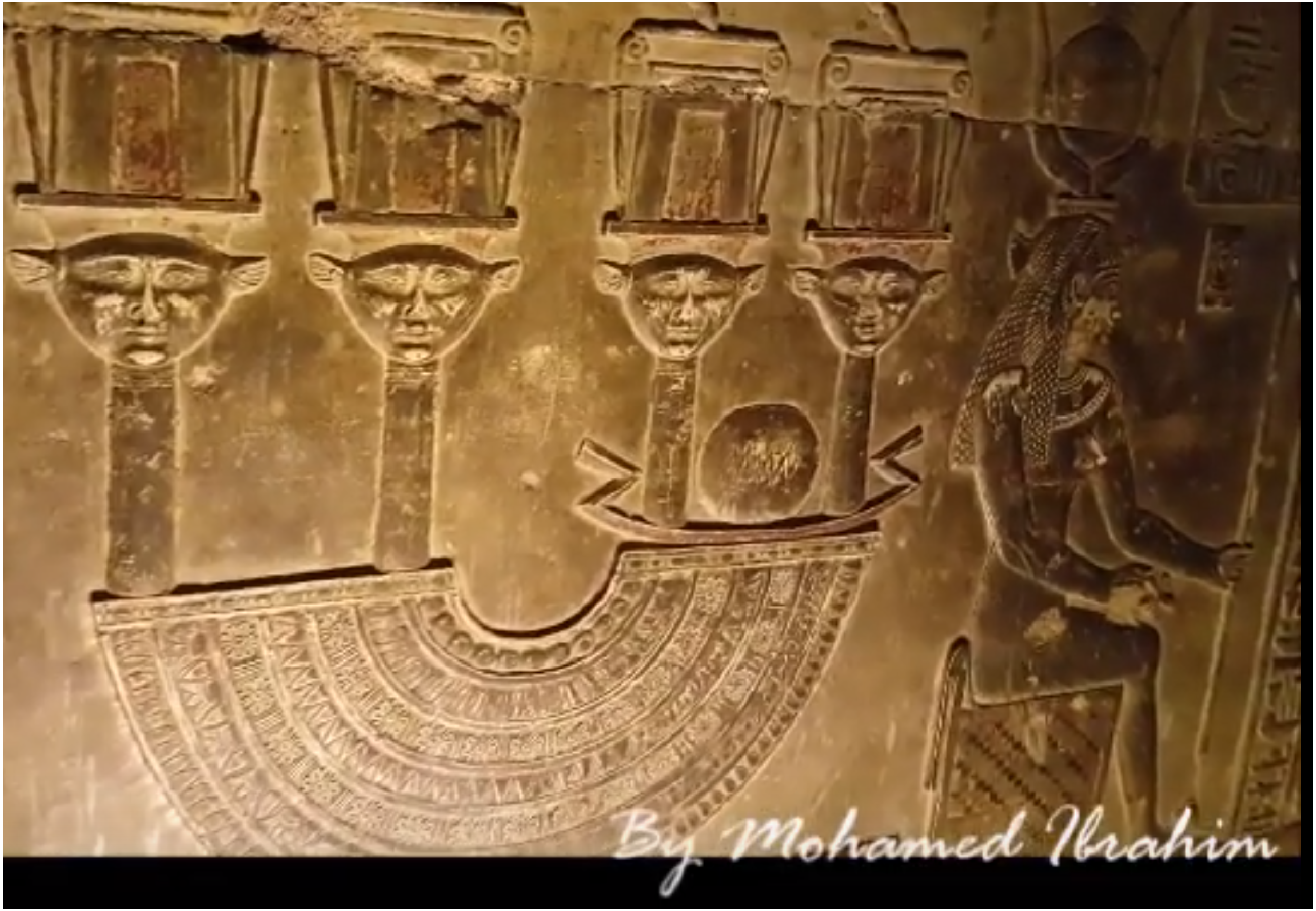

By Mohamed Ibrahim

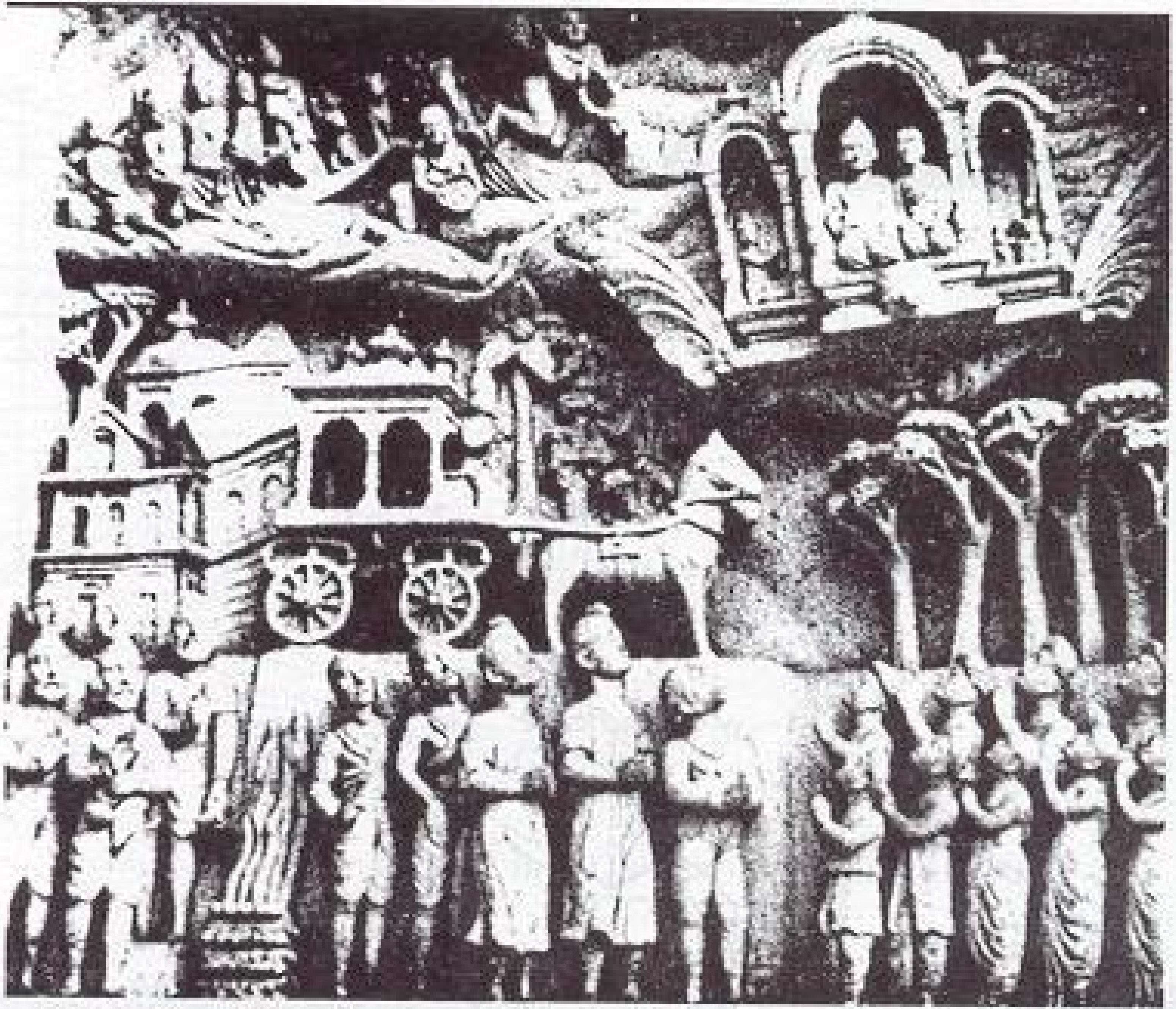

A vimana depicted in a temple relief at Ellora Caves, India.

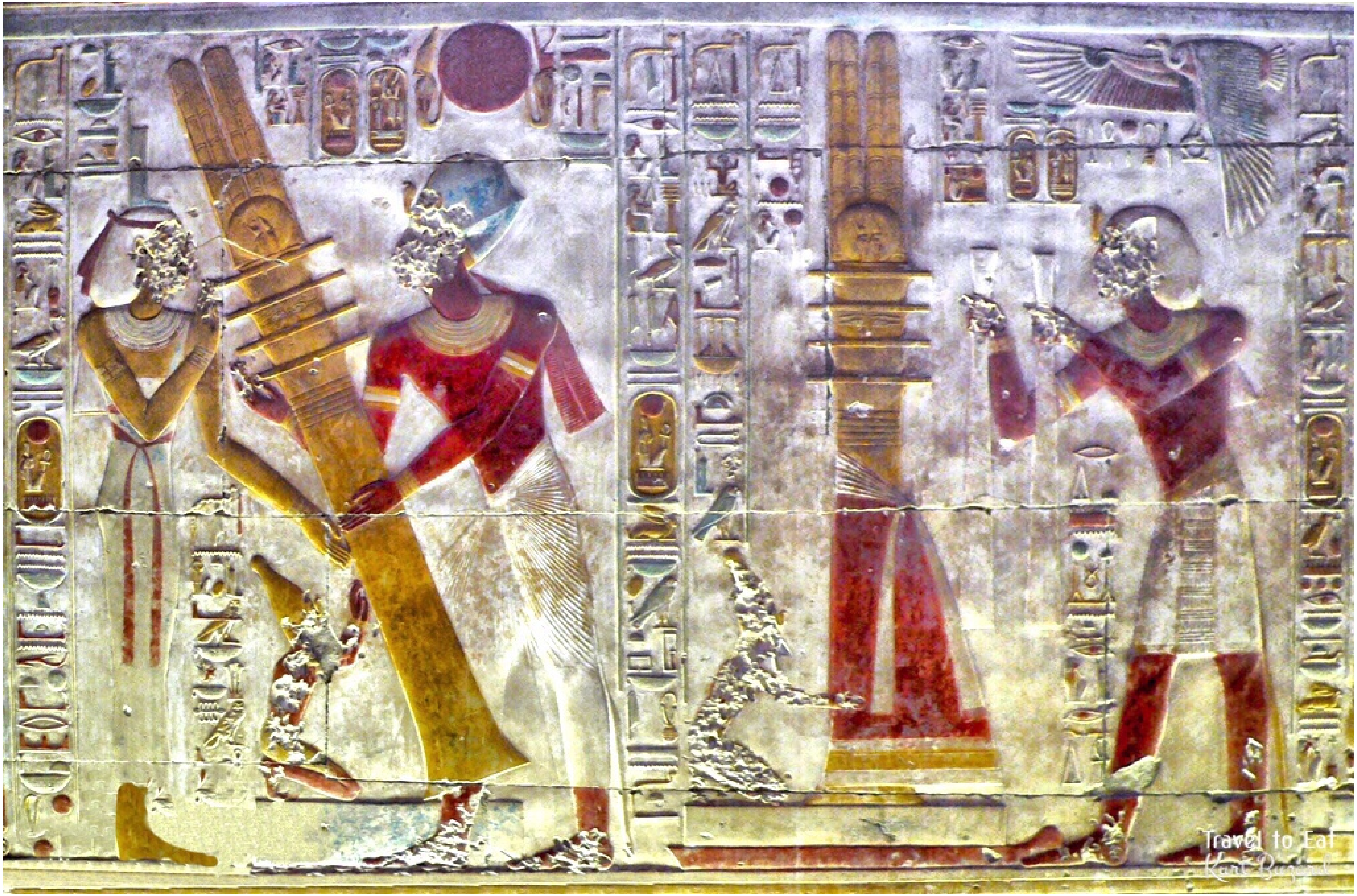
Travel to Eat
Kurt Buzard

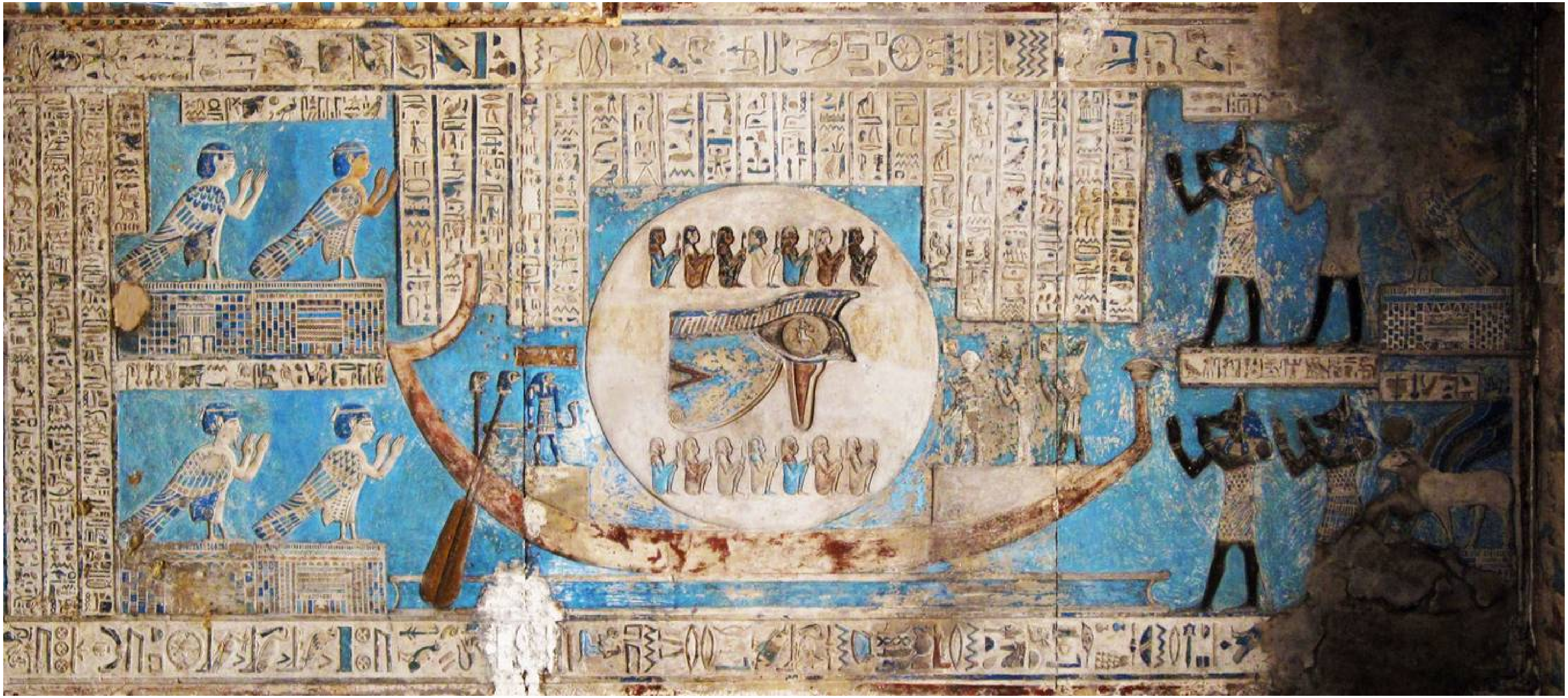

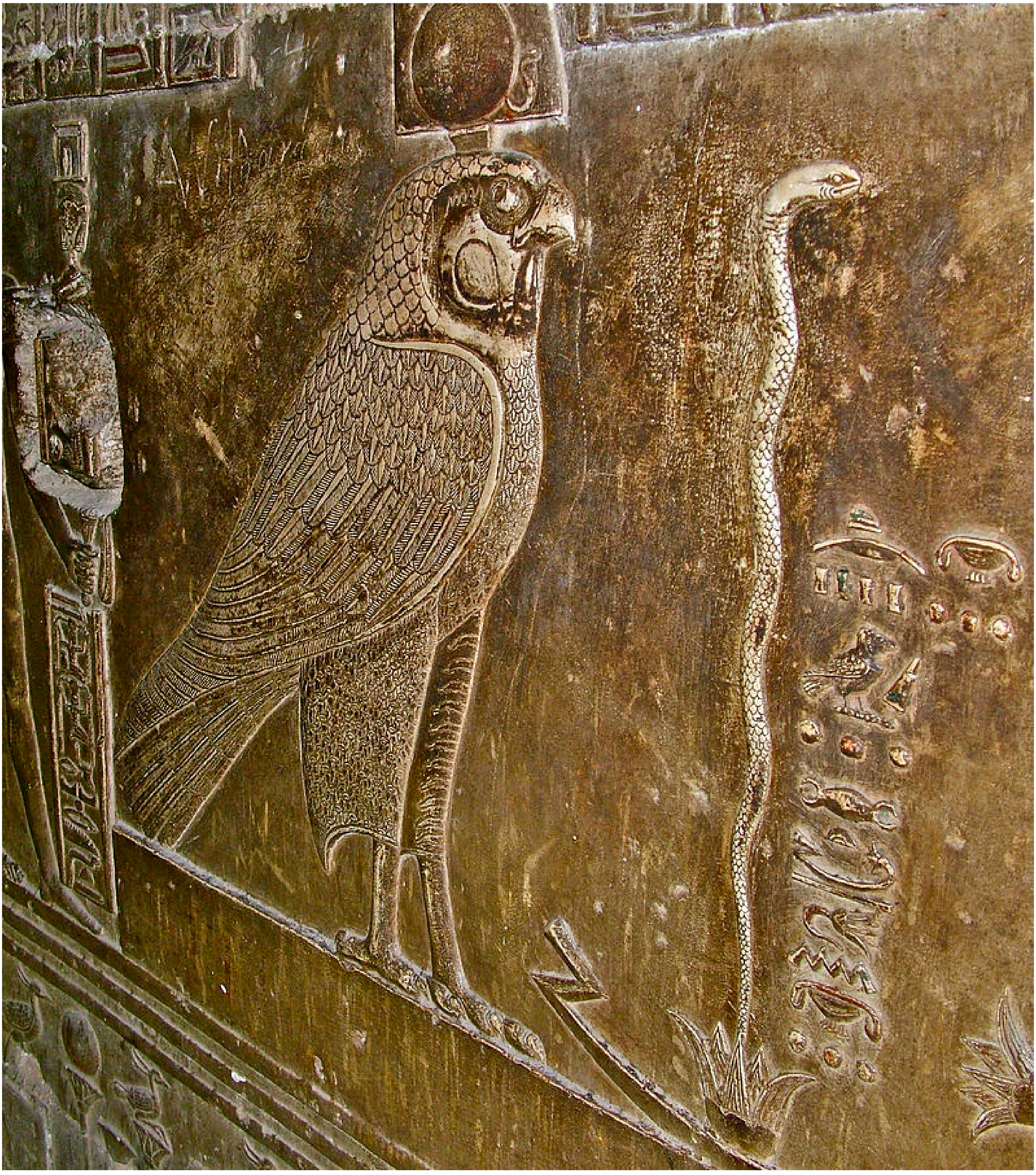

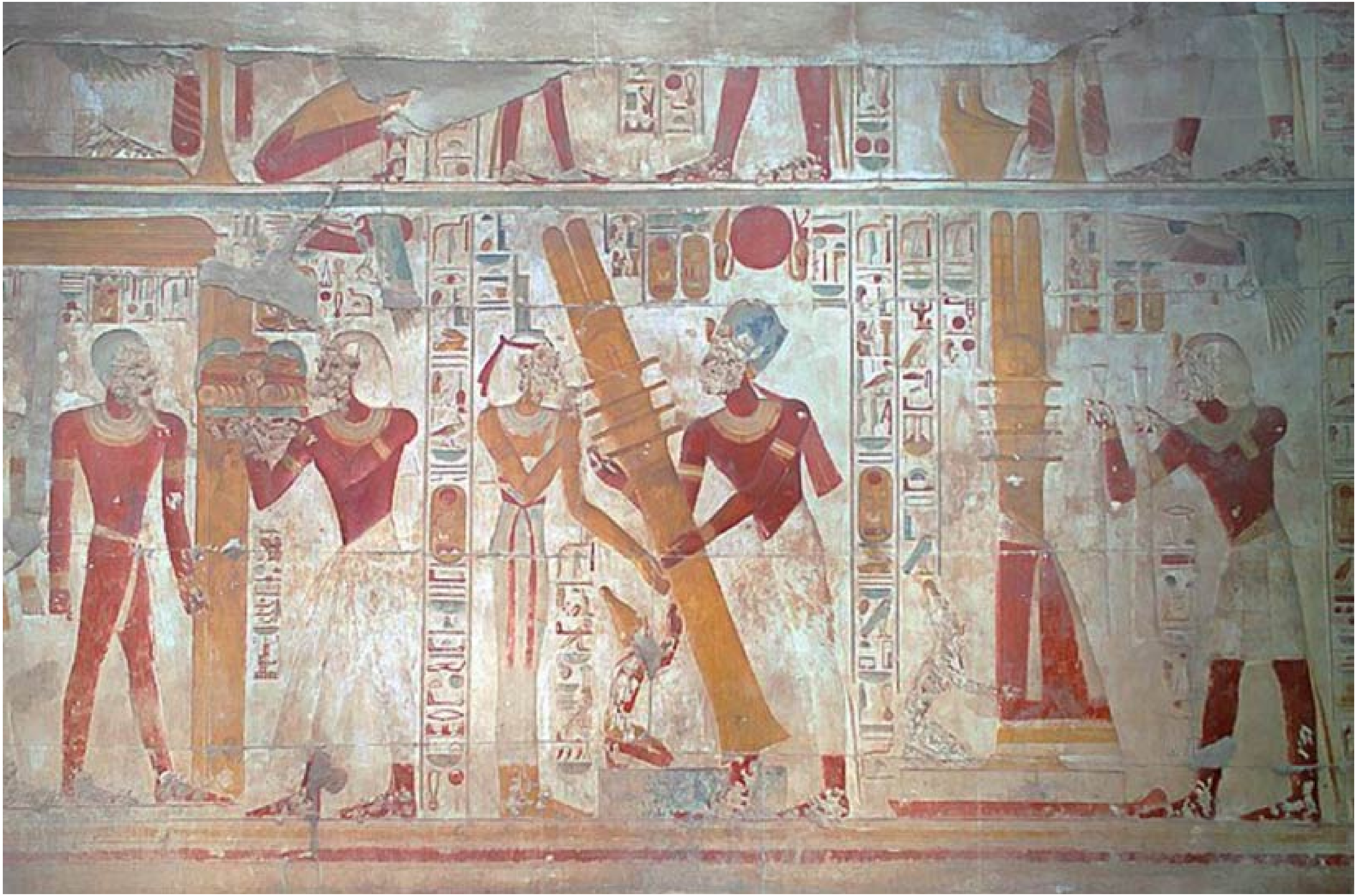

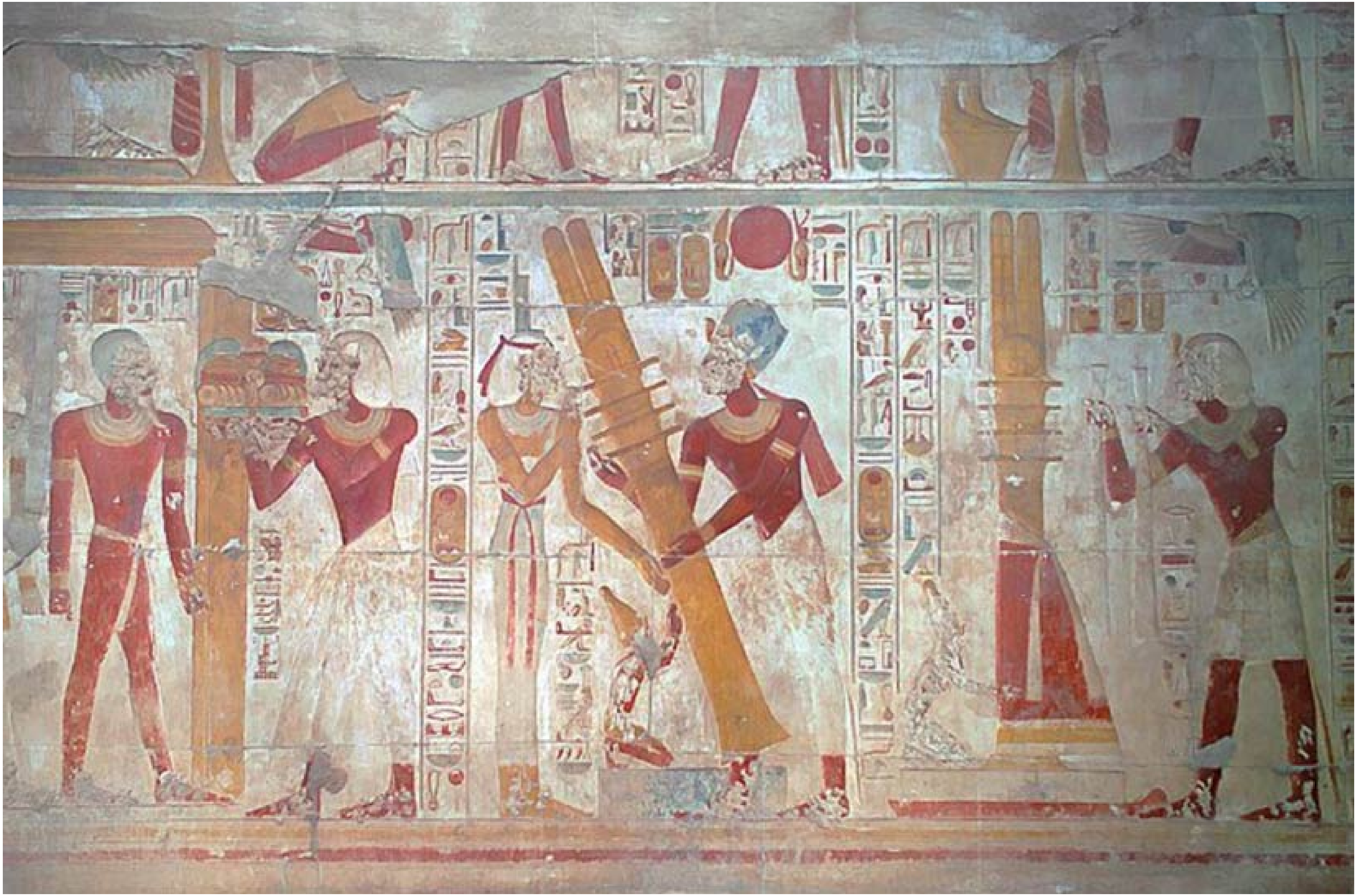

A possible physical artefact from our distant past shows a modern crop picture from 1991

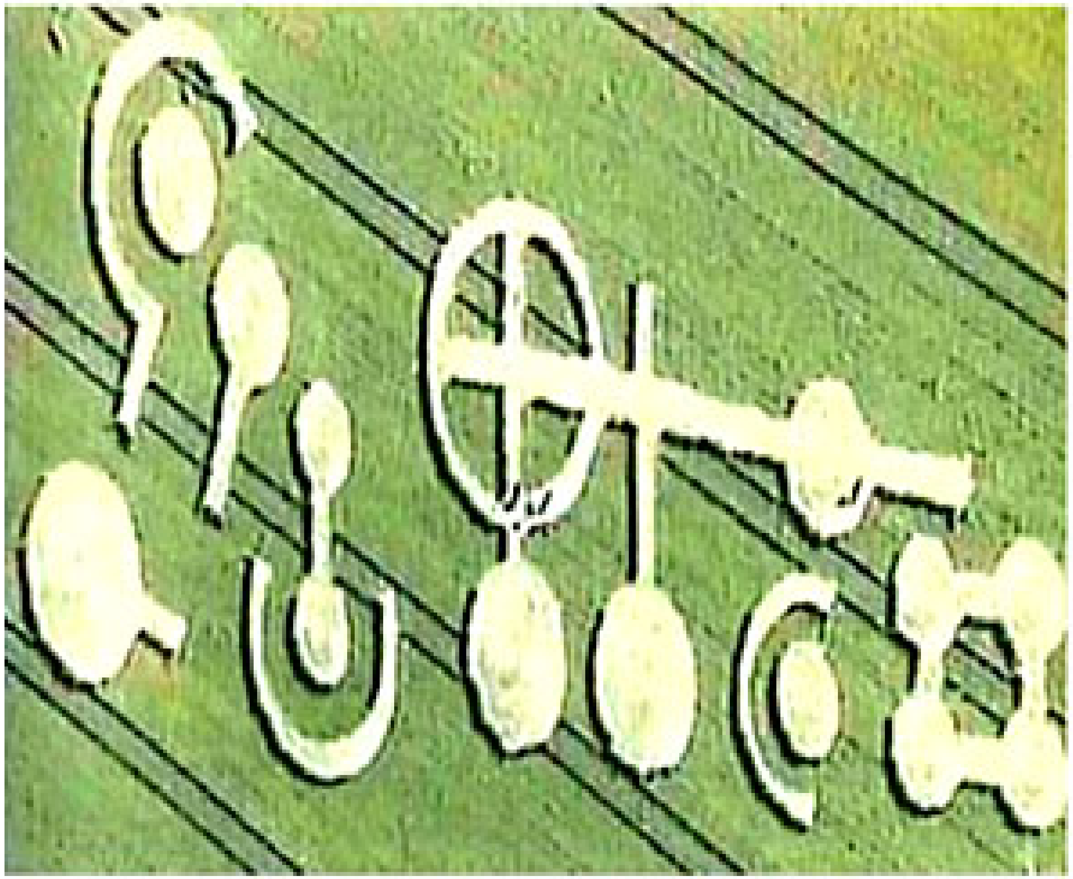

Grasdorf, Germany 1991

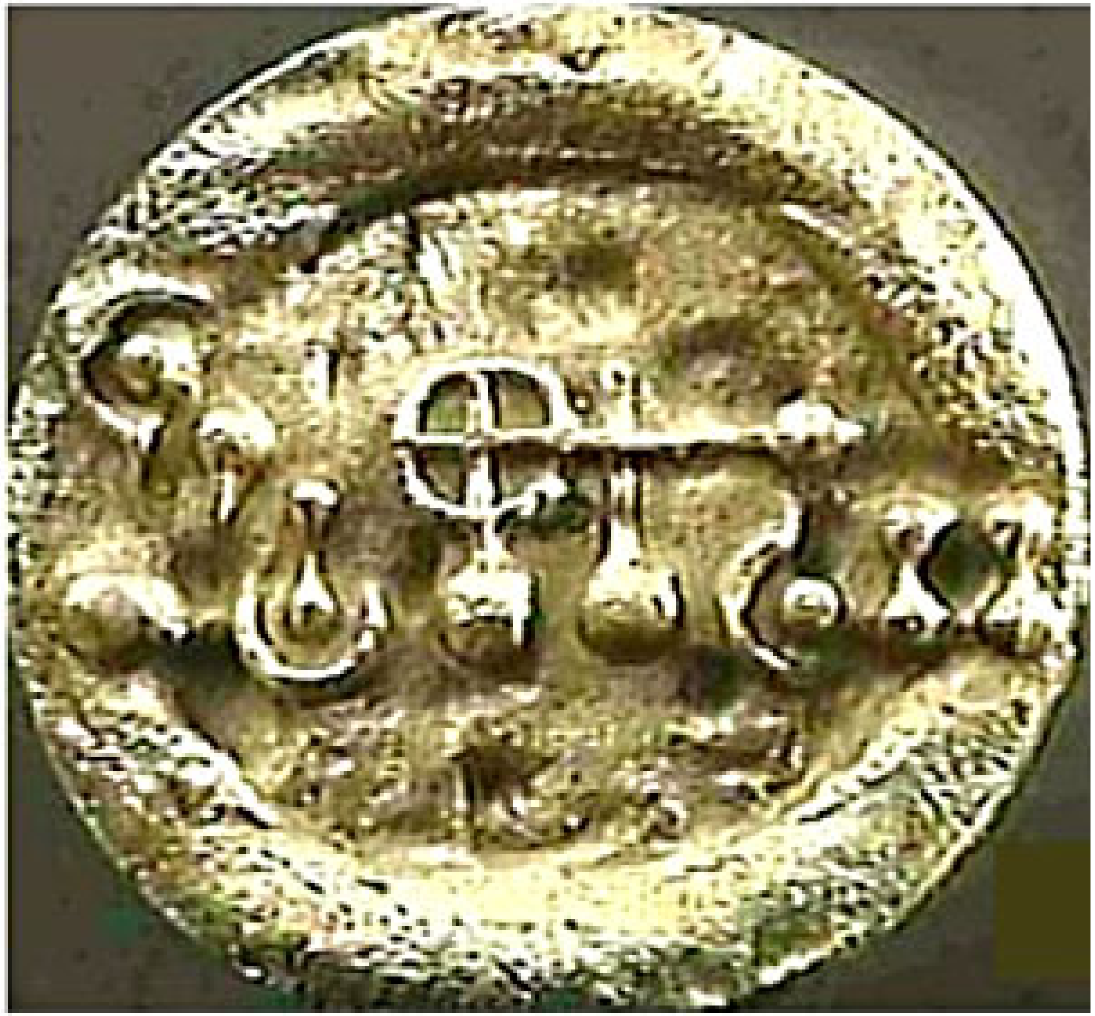

buried golden plate

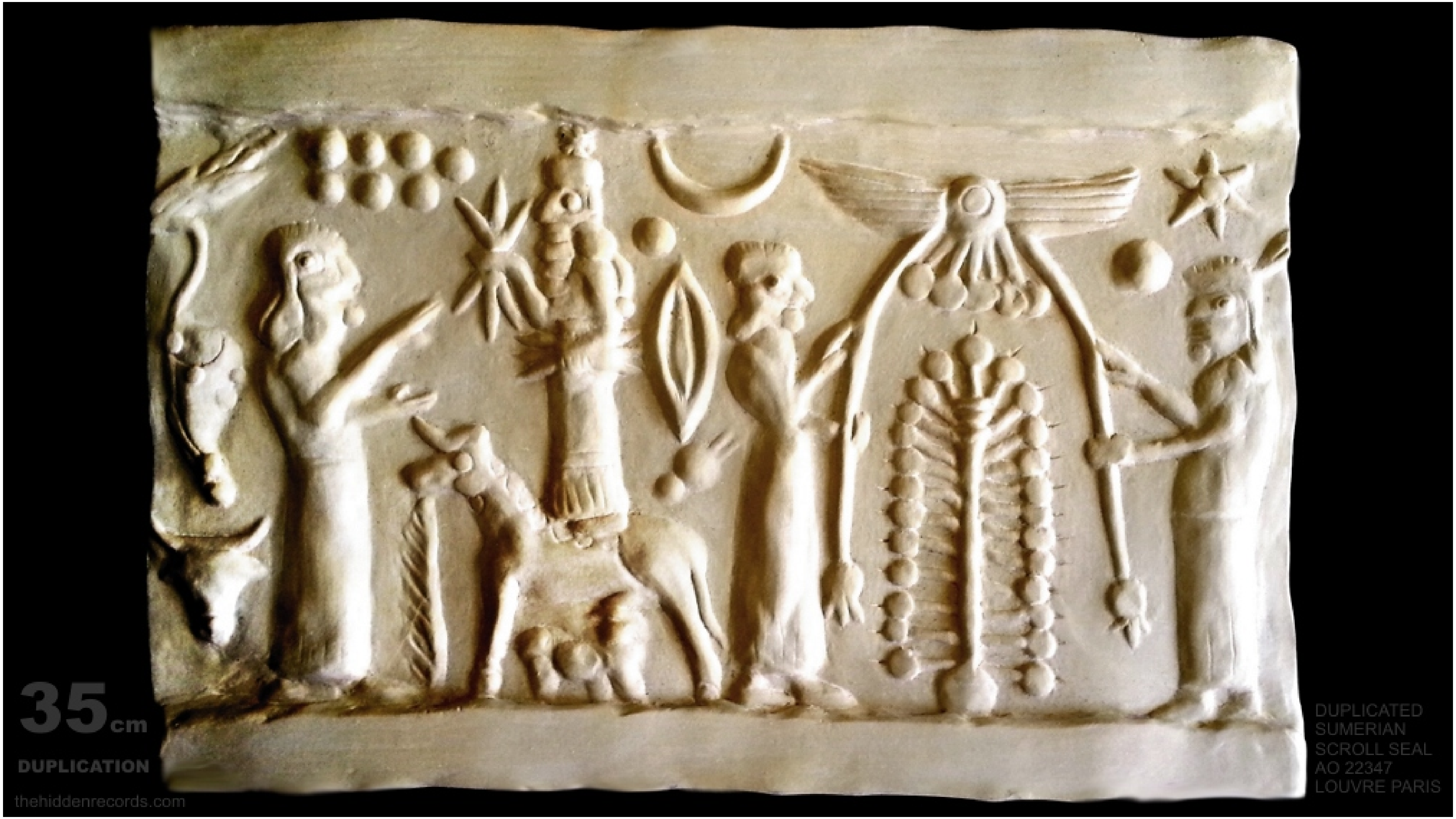
35cm
DUPLICATION
thehiddenrecords.com
DUPLICATED
SUMERIAN
SCROLL SEAL
AO 22347
LOUVRE PARIS

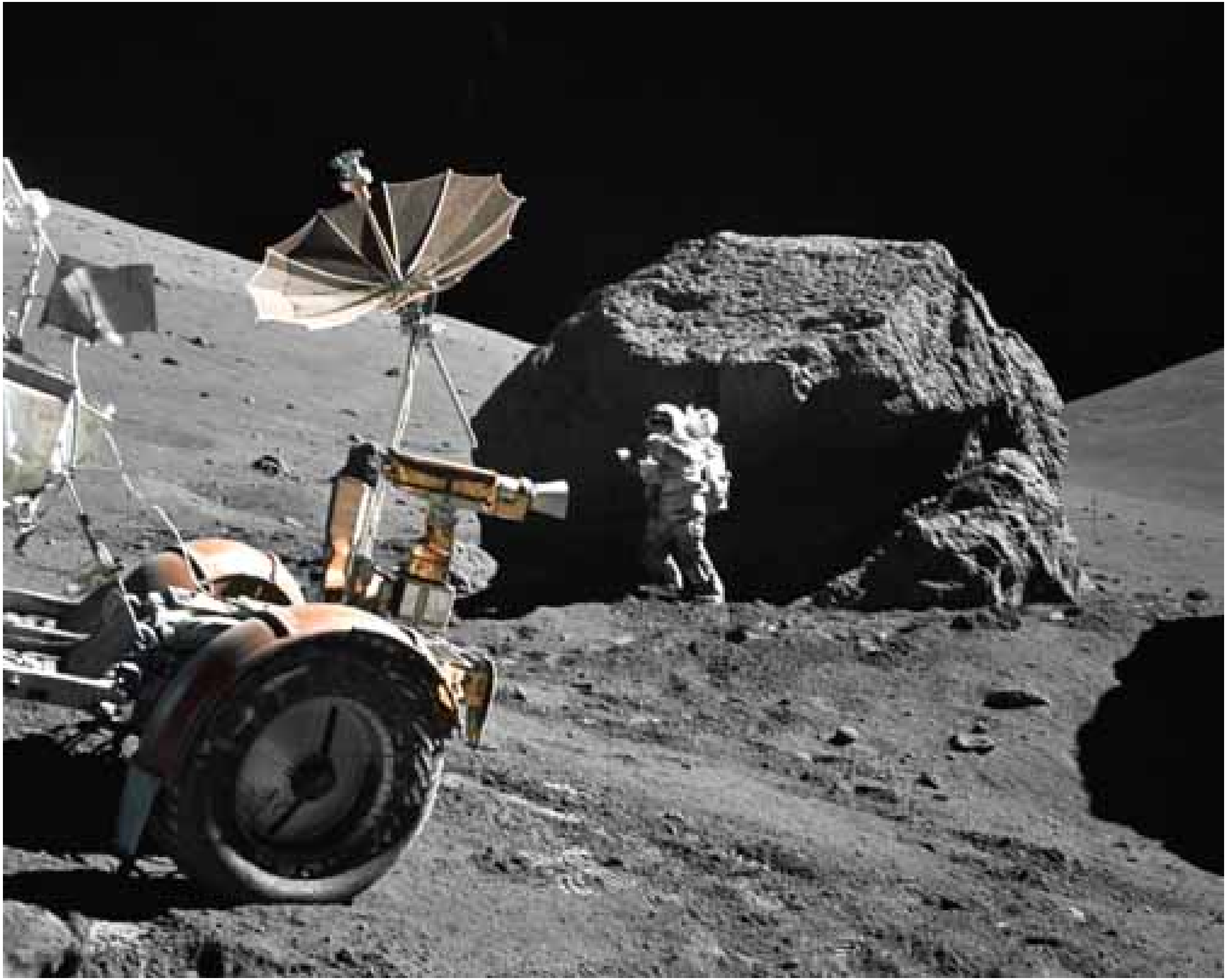

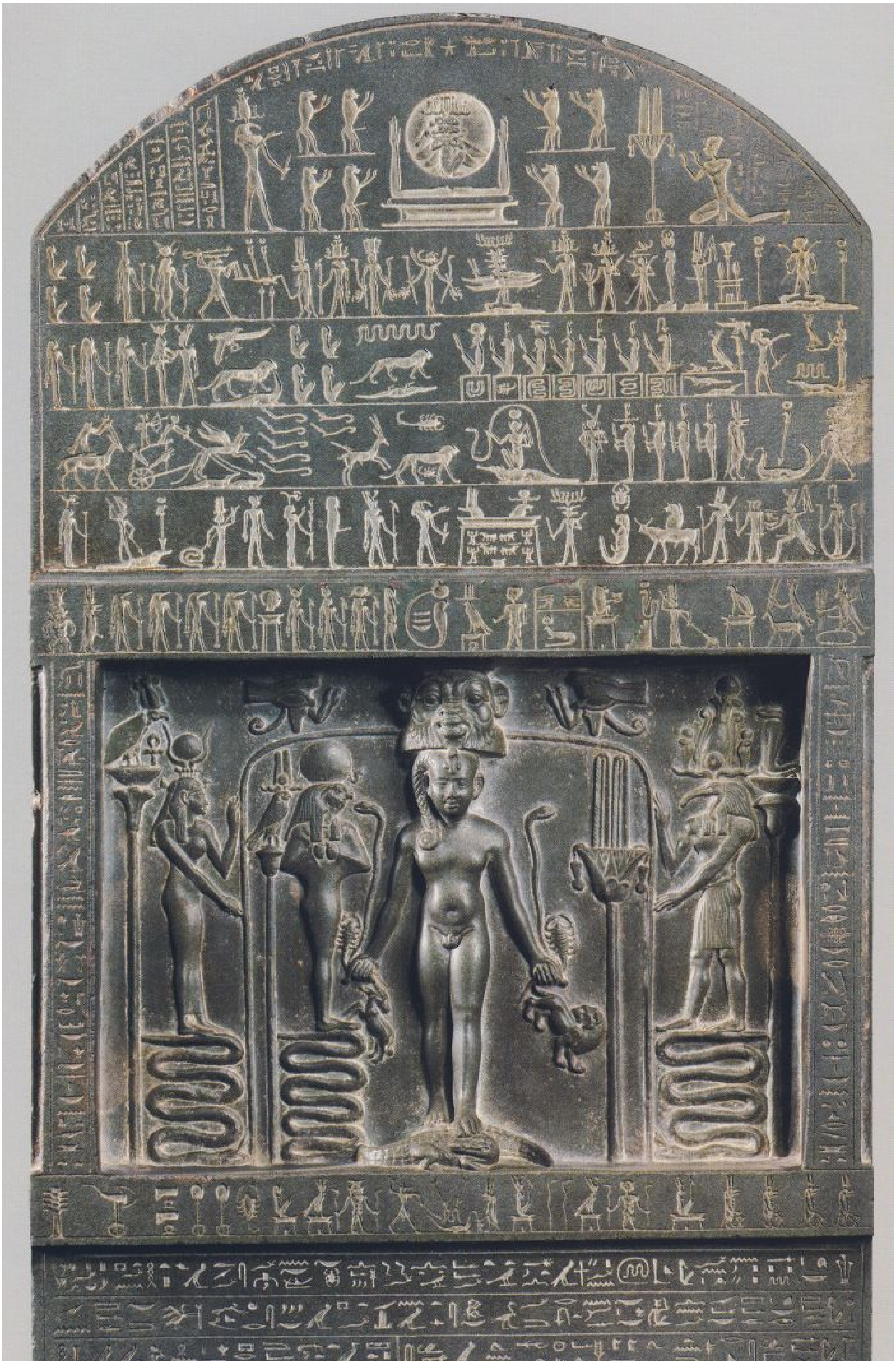

© www.SandInMySuitcase.com

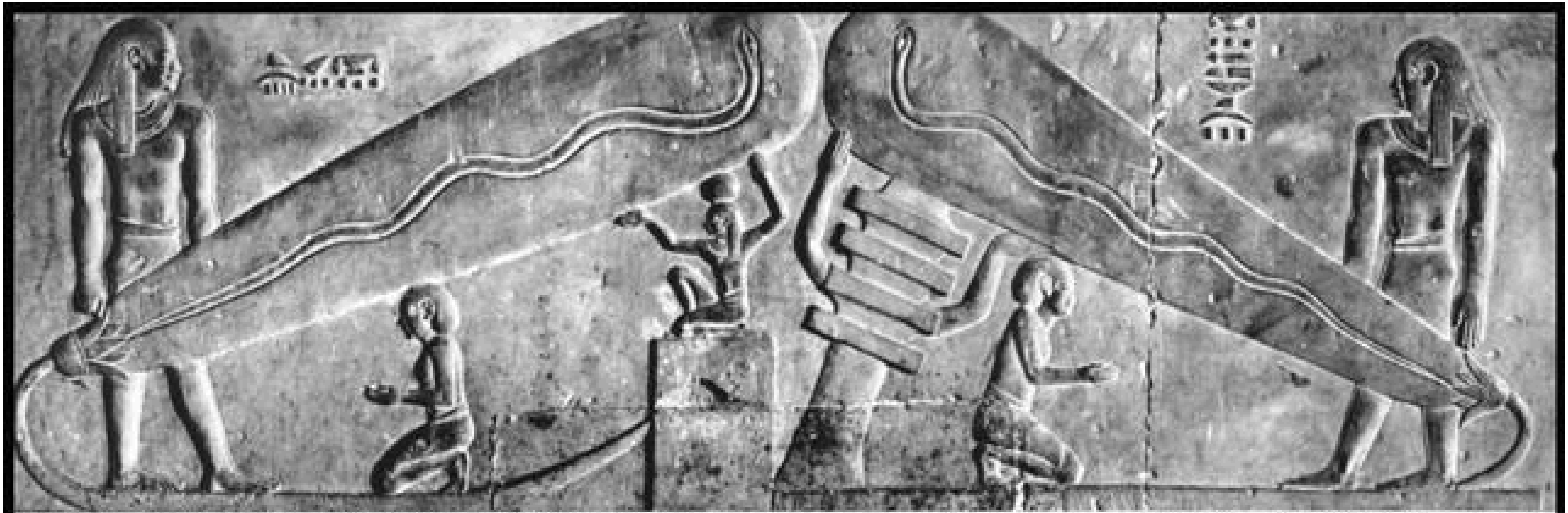

Battery-powered searchlights and filament lamps in the crypts of the Temple of Hathor (Isis) at Denderah

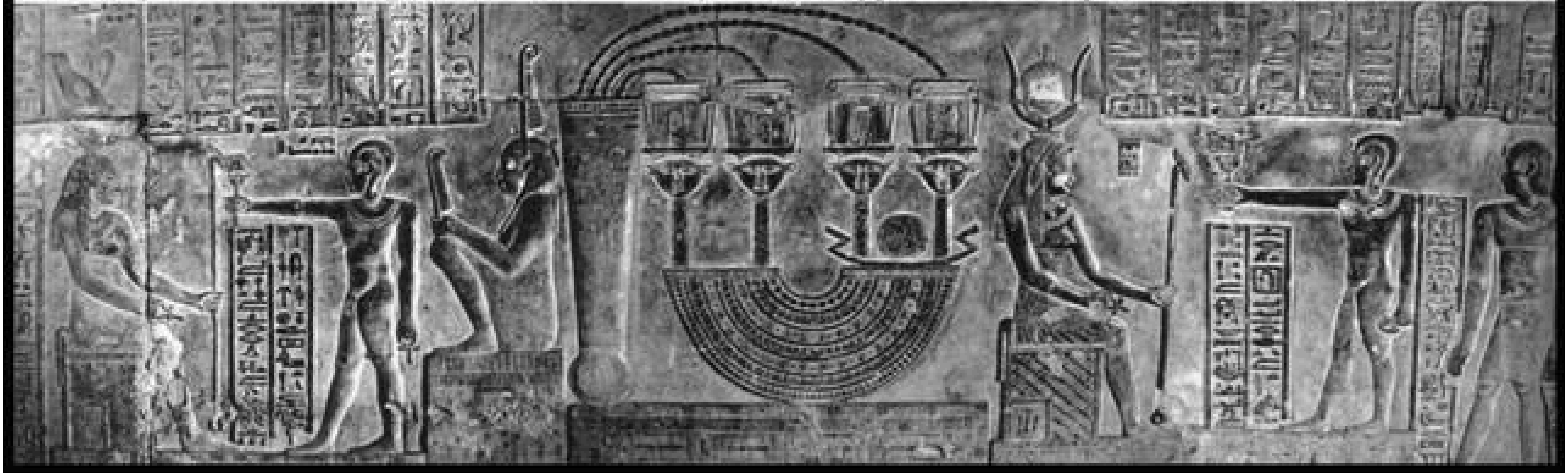

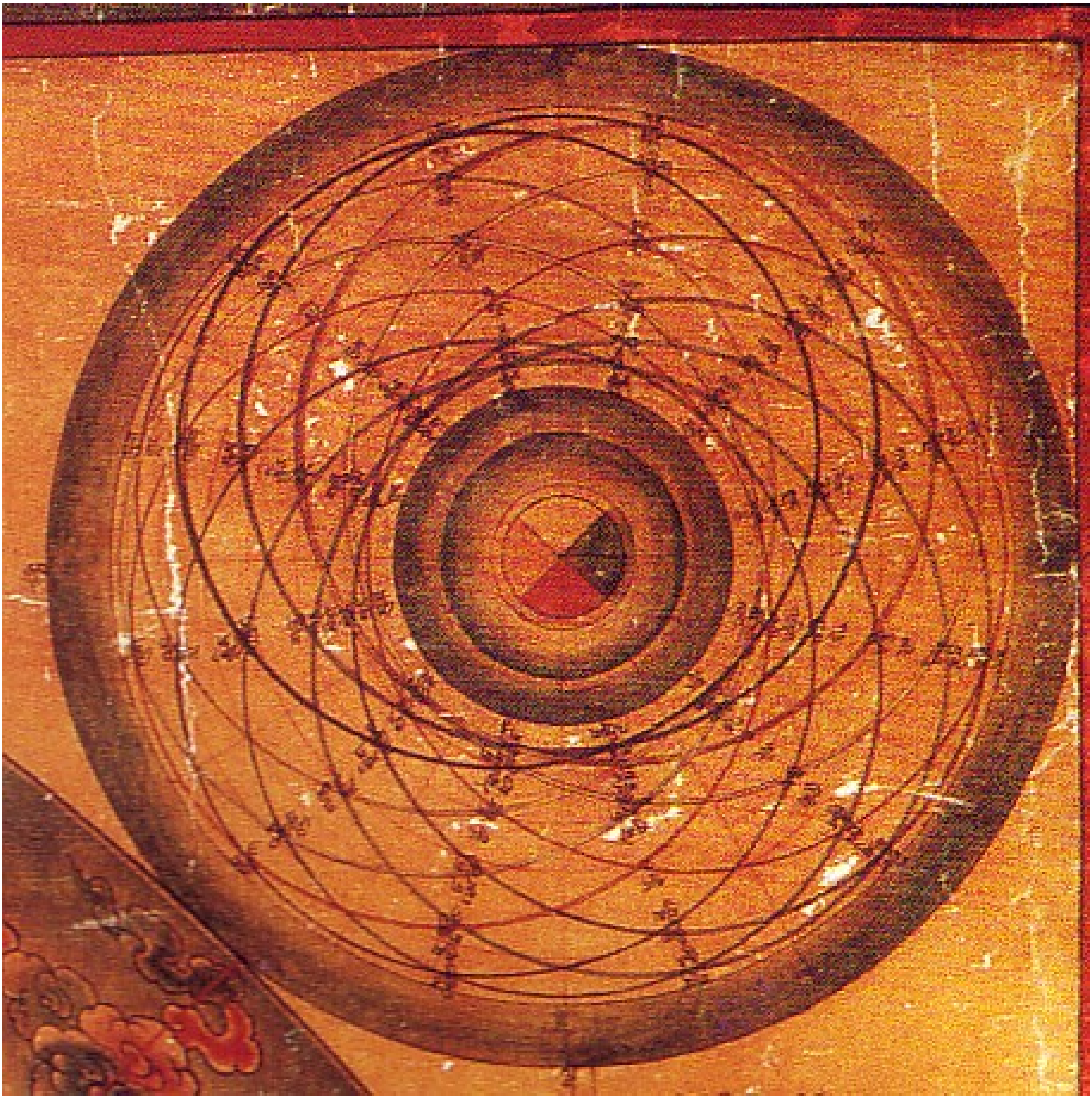

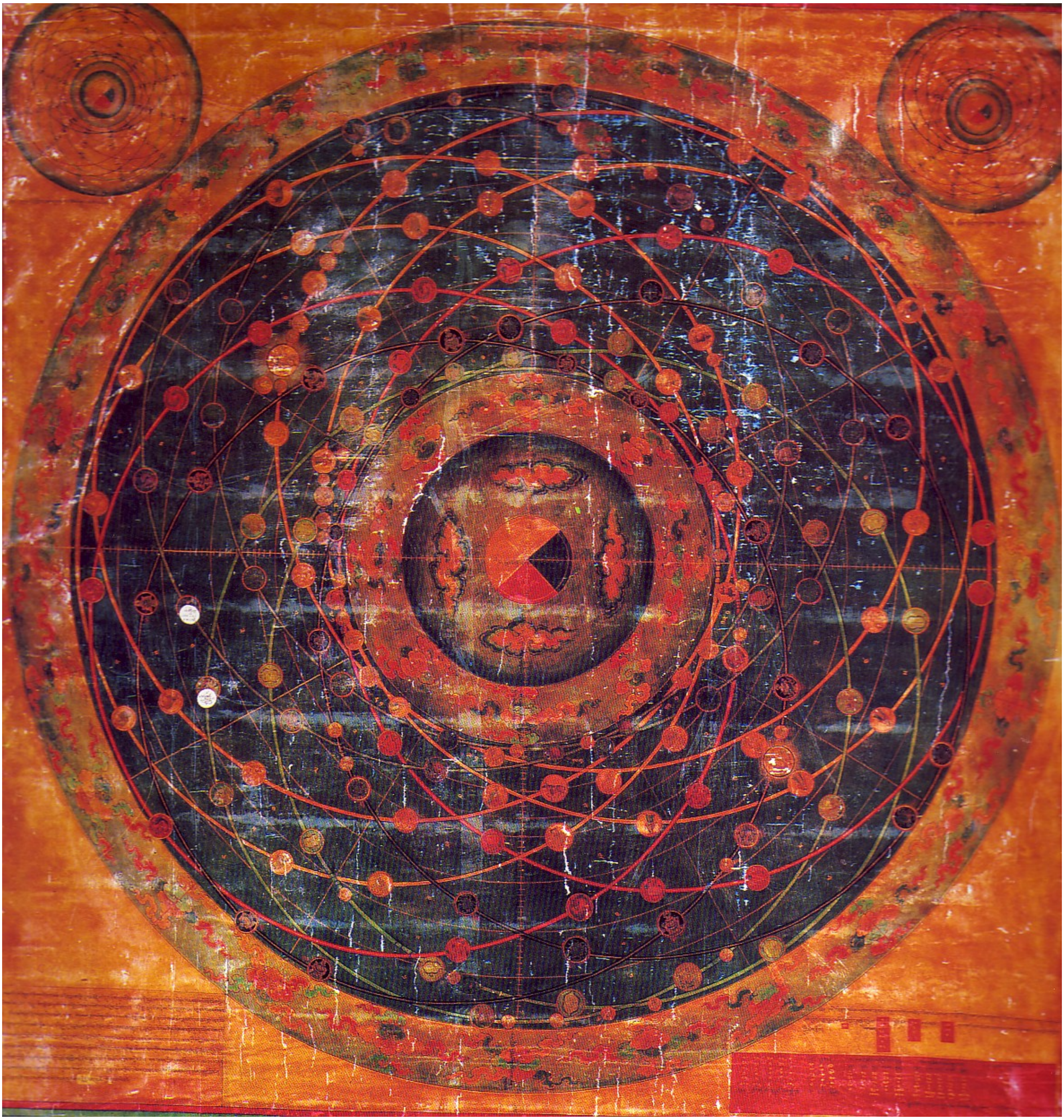

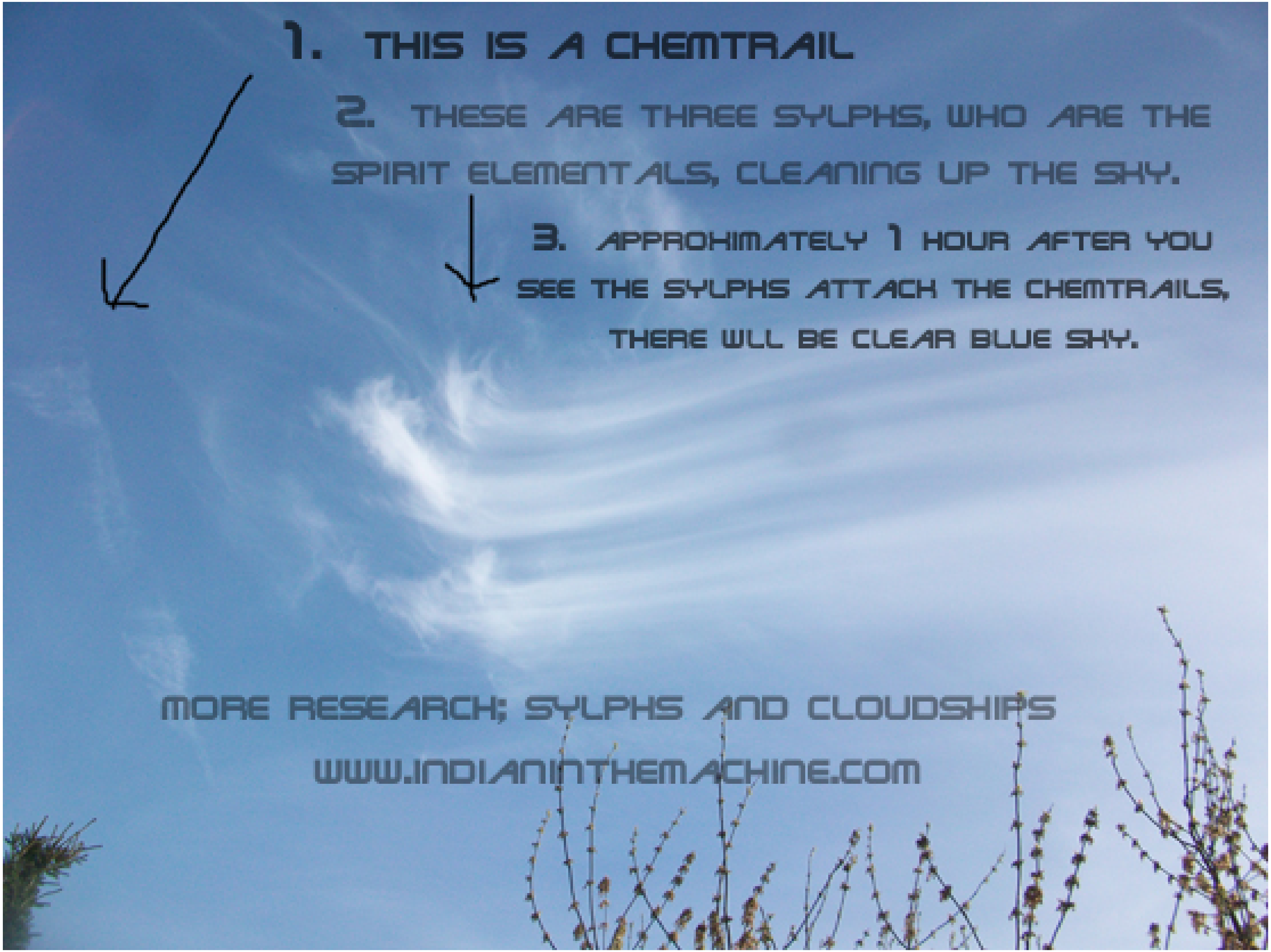

1. THIS IS A CHEMTRAIL

2. THESE ARE THREE SYLPHS, WHO ARE THE SPIRIT ELEMENTALS, CLEANING UP THE SKY.

3. APPROXIMATELY 1 HOUR AFTER YOU SEE THE SYLPHS ATTACH THE CHEMTRAILS, THERE WILL BE CLEAR BLUE SKY.

MORE RESEARCH; SYLPHS AND CLOUDSHIPS

WWW.INDIANINTHEMACHINE.COM

Grandidier's Baobab

Dragon Blood Tree

Jacaranda

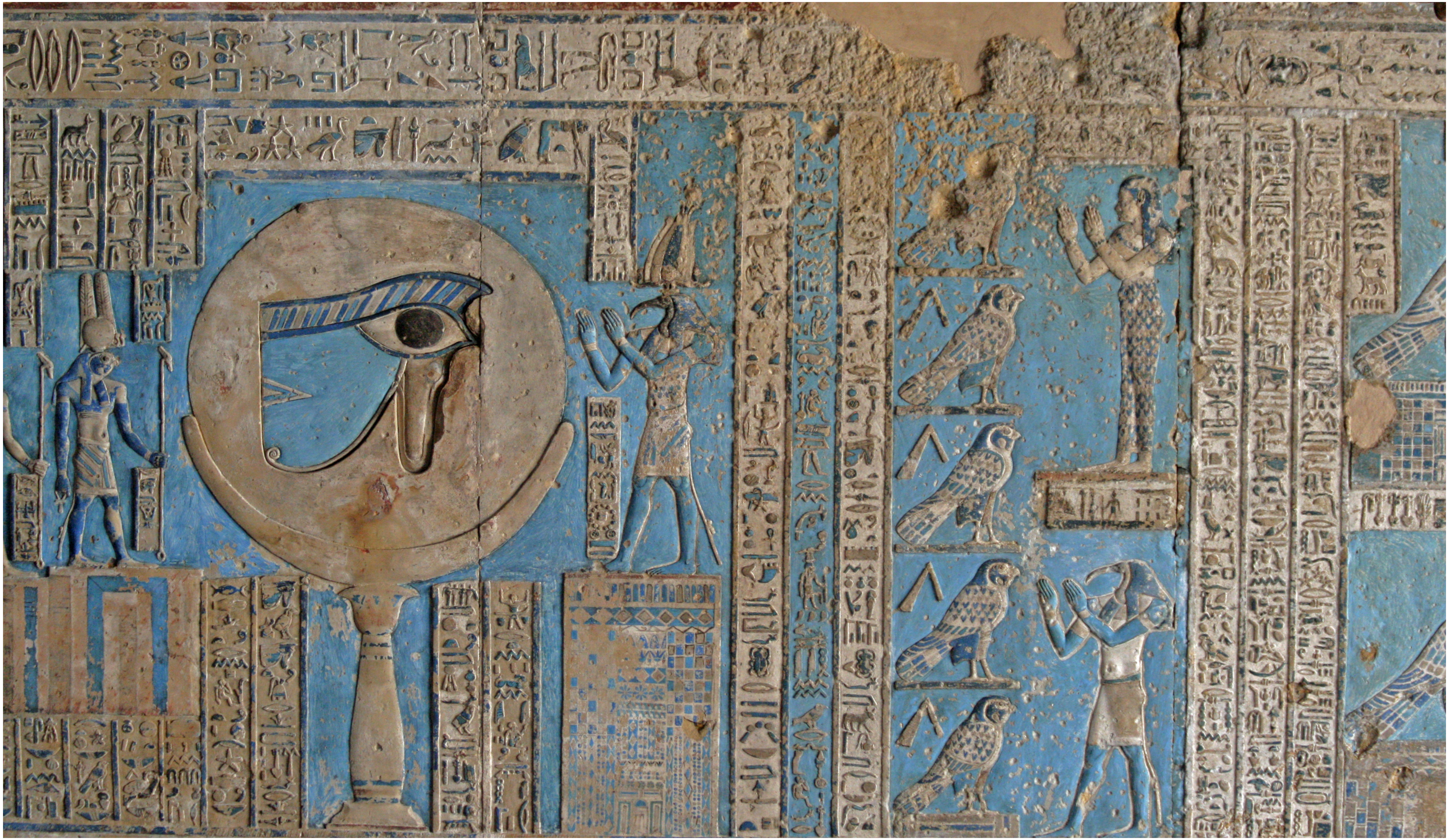

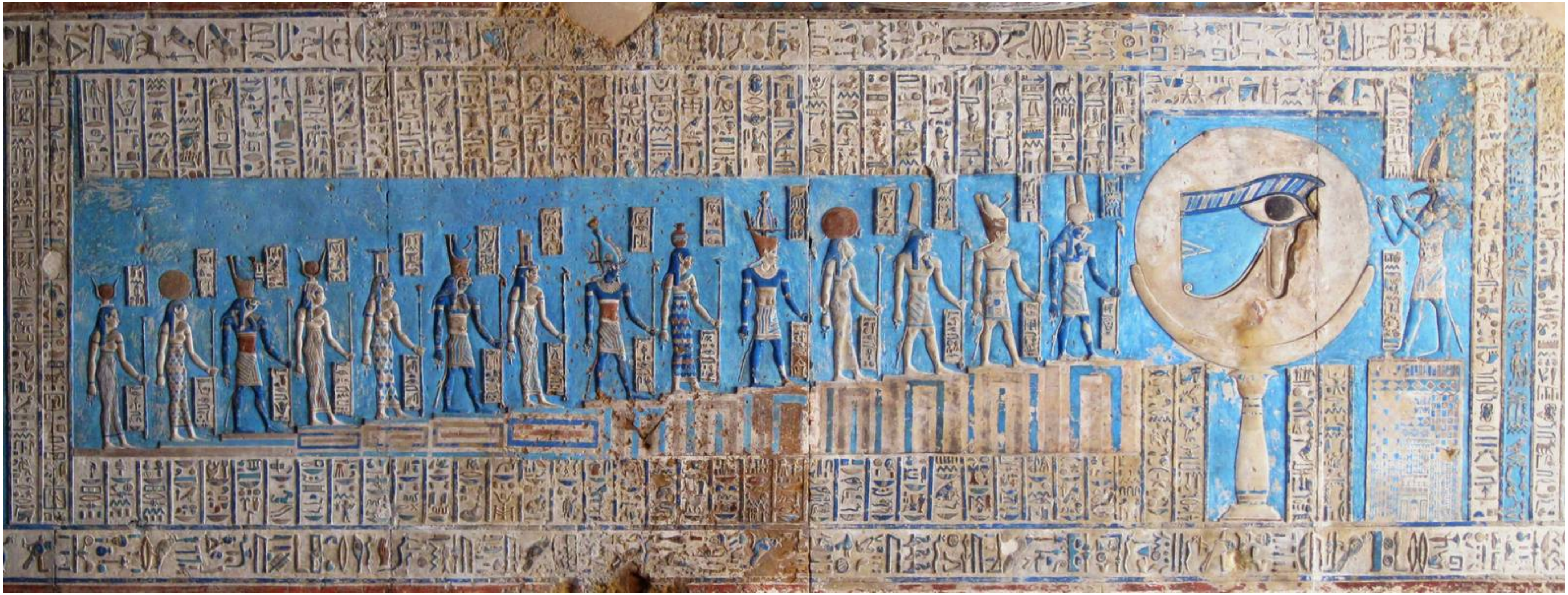

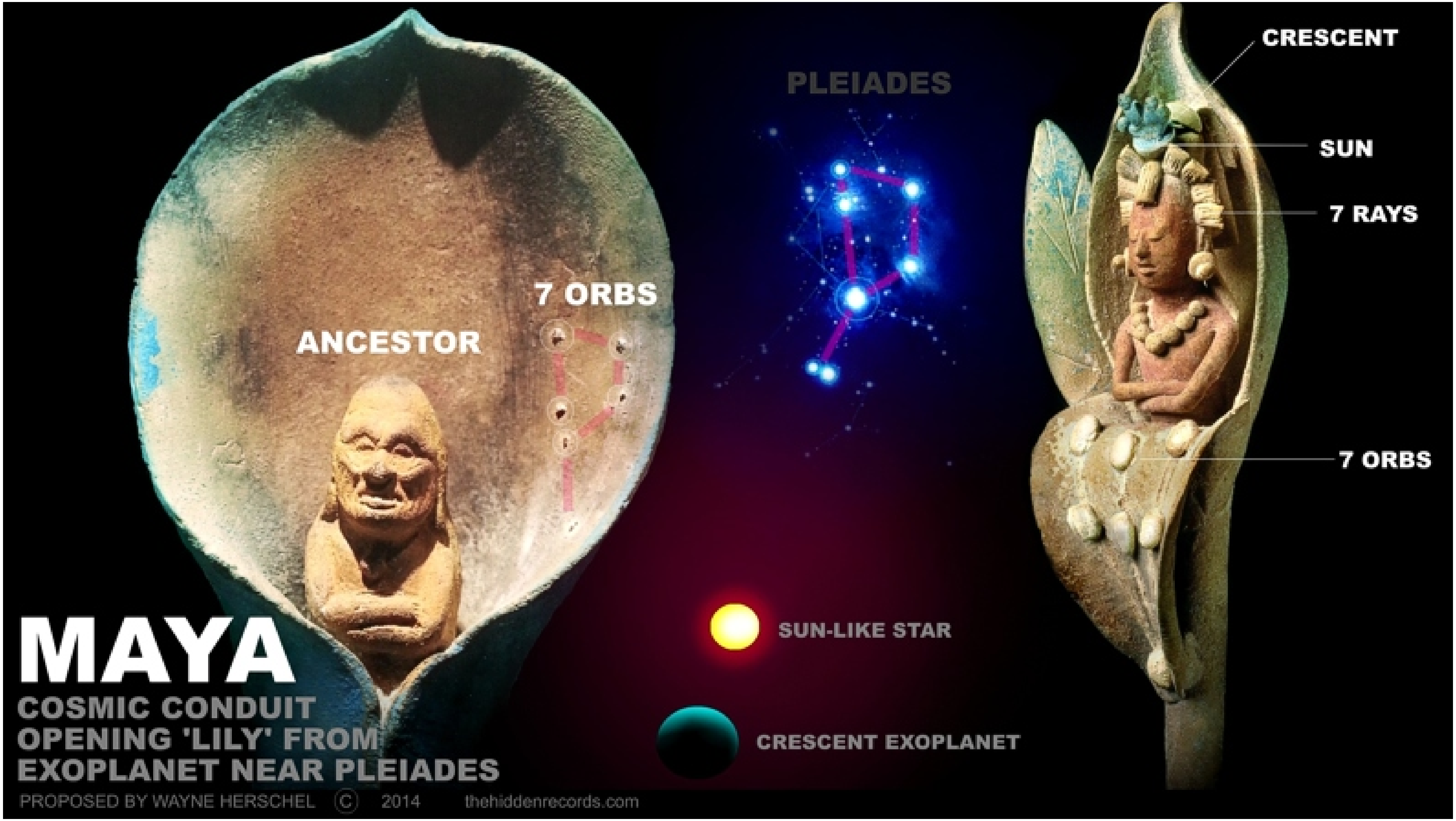
PLEIADES
CRESCENT
SUN
7 RAYS
7 ORBS
ANCESTOR
7 ORBS
SUN-LIKE STAR
CRESCENT EXOPLANET

MAYA

COSMIC CONDUIT OPENING 'LILY' FROM EXOPLANET NEAR PLEIADES

PROPOSED BY WAYNE HERSCHEL © 2014 thehiddenrecords.com

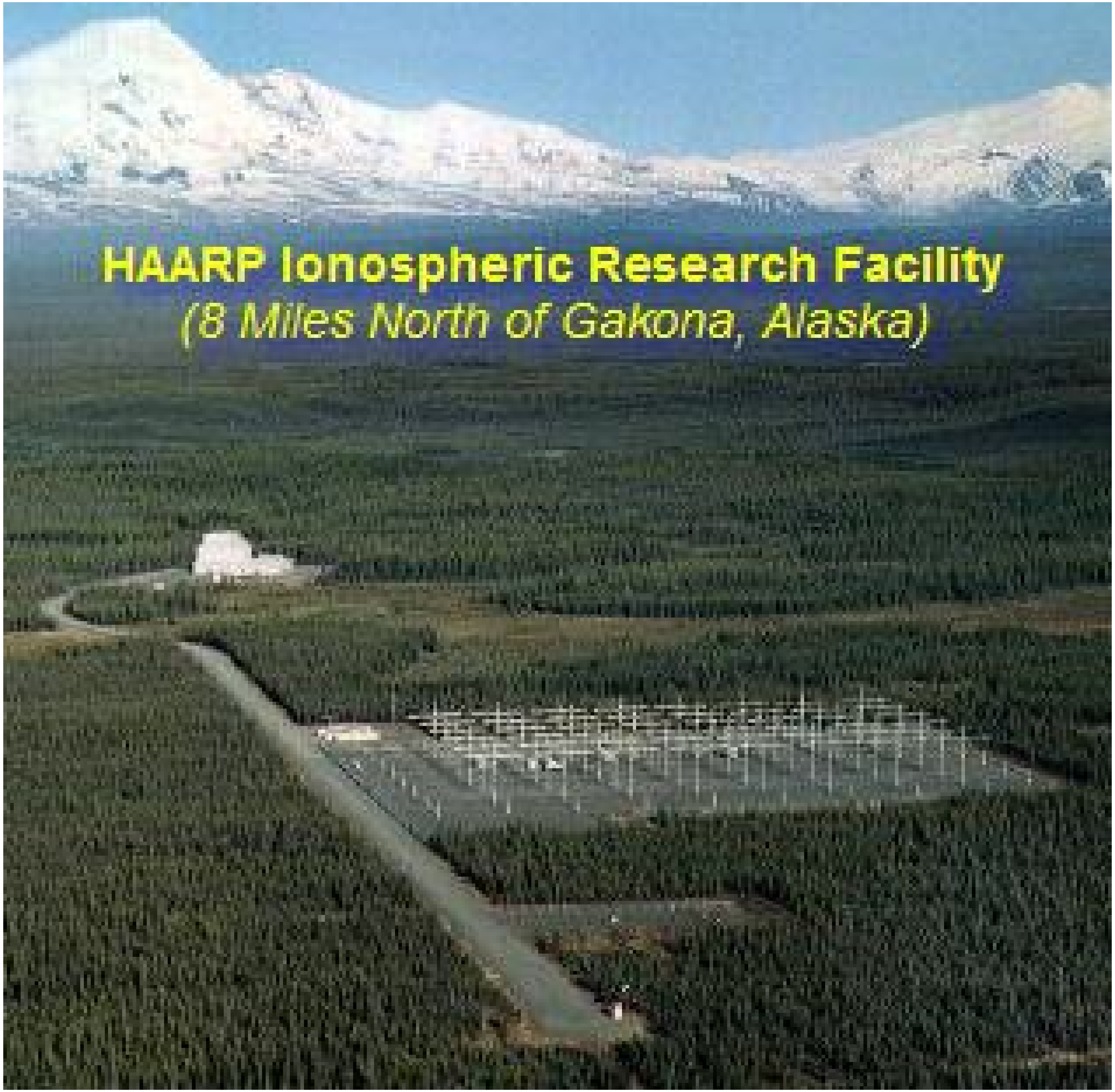

HAARP Ionospheric Research Facility

(8 Miles North of Gakona, Alaska)

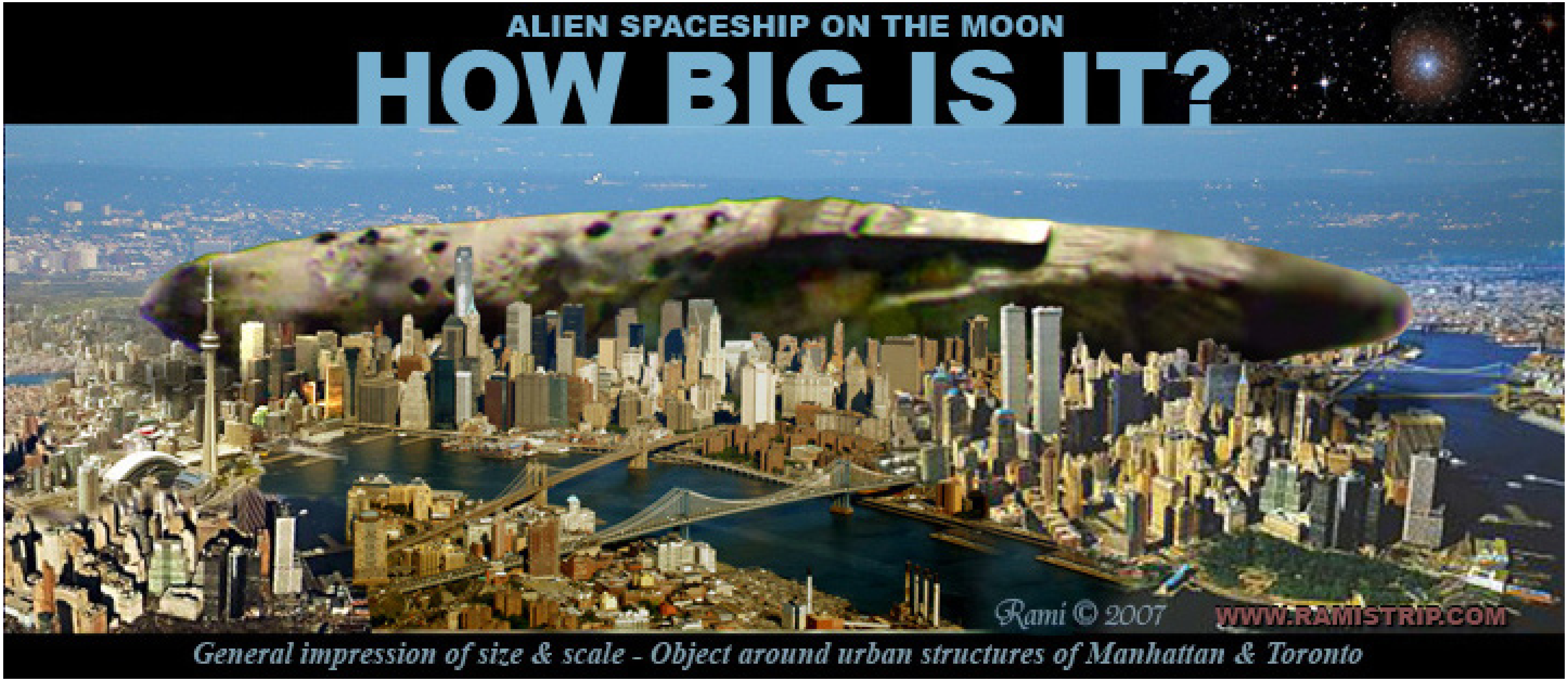
ALIEN SPACESHIP ON THE MOON
HOW BIG IS IT?
Rami © 2007
WWW.RAMISTRIP.COM
General impression of size & scale - Object around urban structures of Manhattan & Toronto

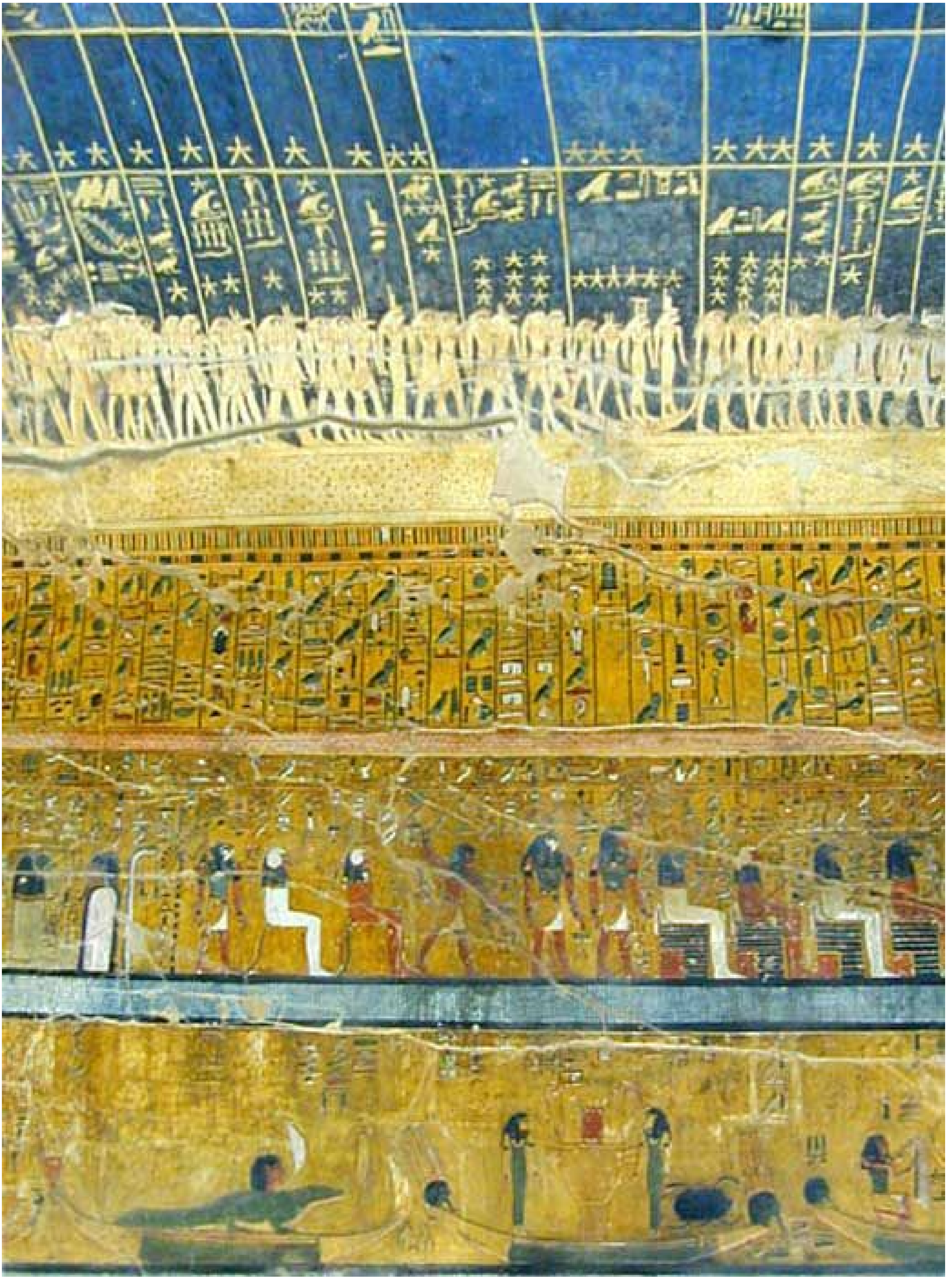

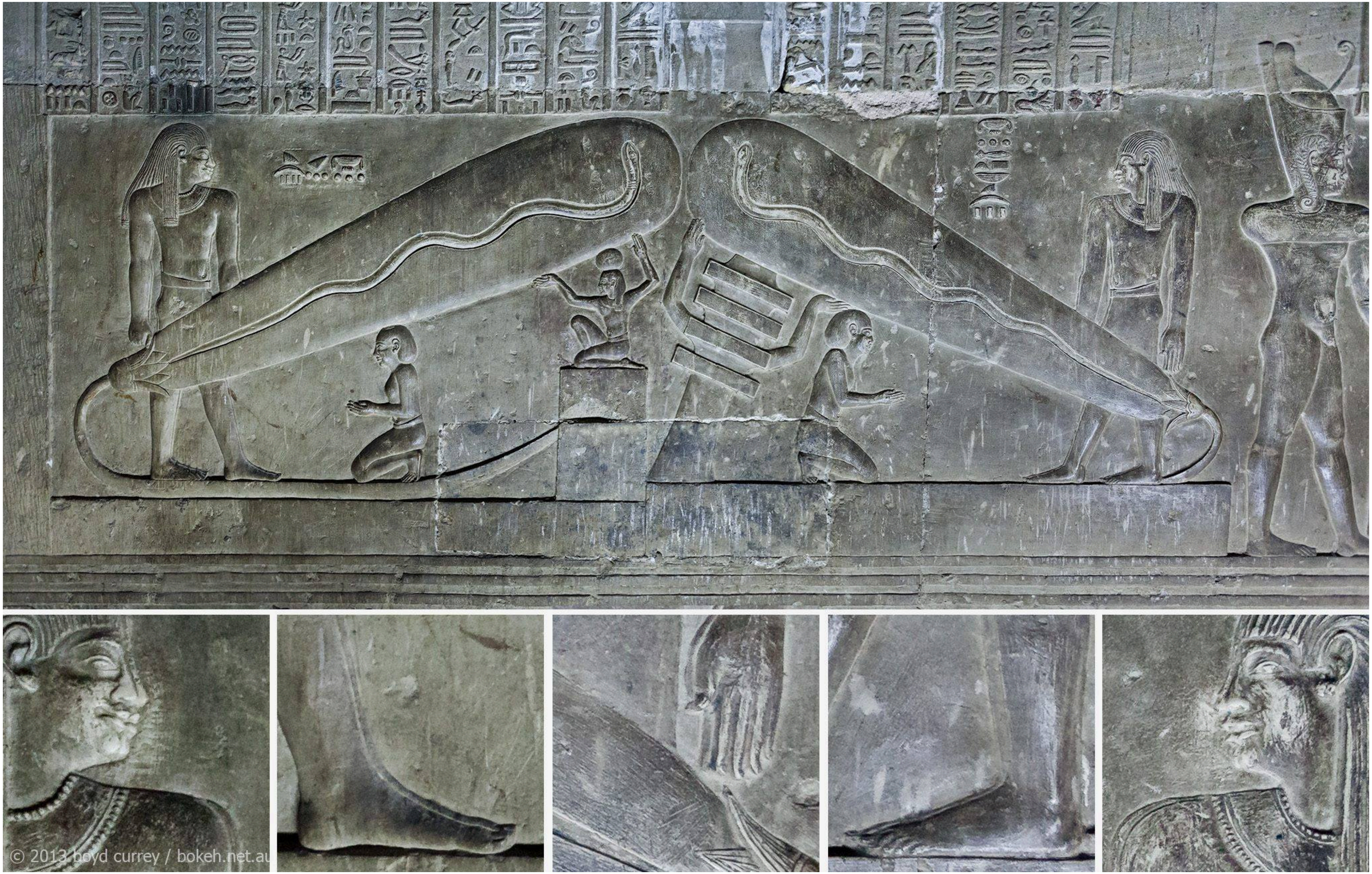

© 2013 boyd currey / bokeh.net.au

gettyimages
Kenneth Garrett
174405620

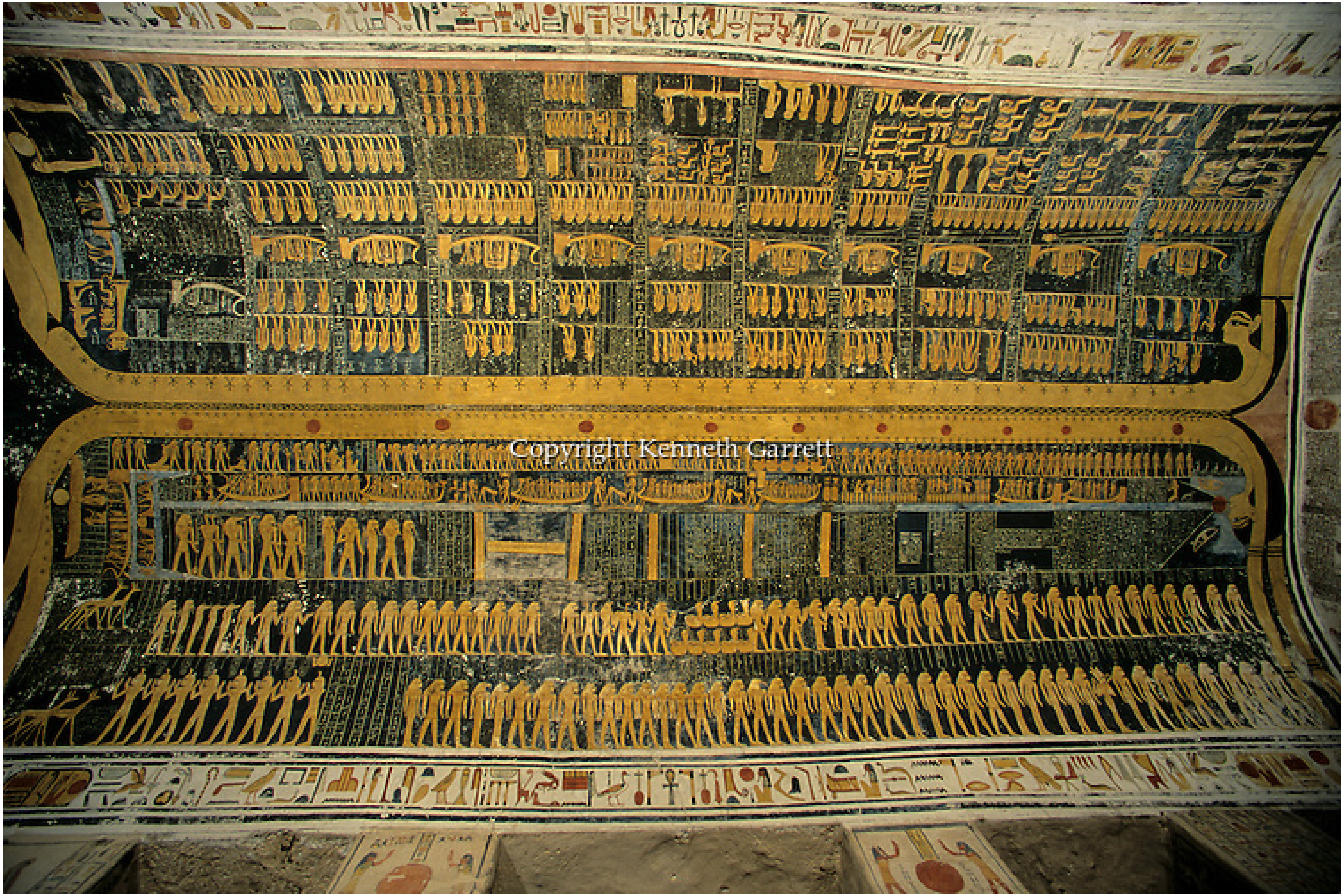

Copyright Kenneth Garrett

© [illegible] palarczyk

TALKS WITH TEACHINGS FROM MY COSMIC FRIENDS

ASHTAR SHERAN
ORTHON
angels-heaven.org
Pic. 4395 P EN

MEGADETH

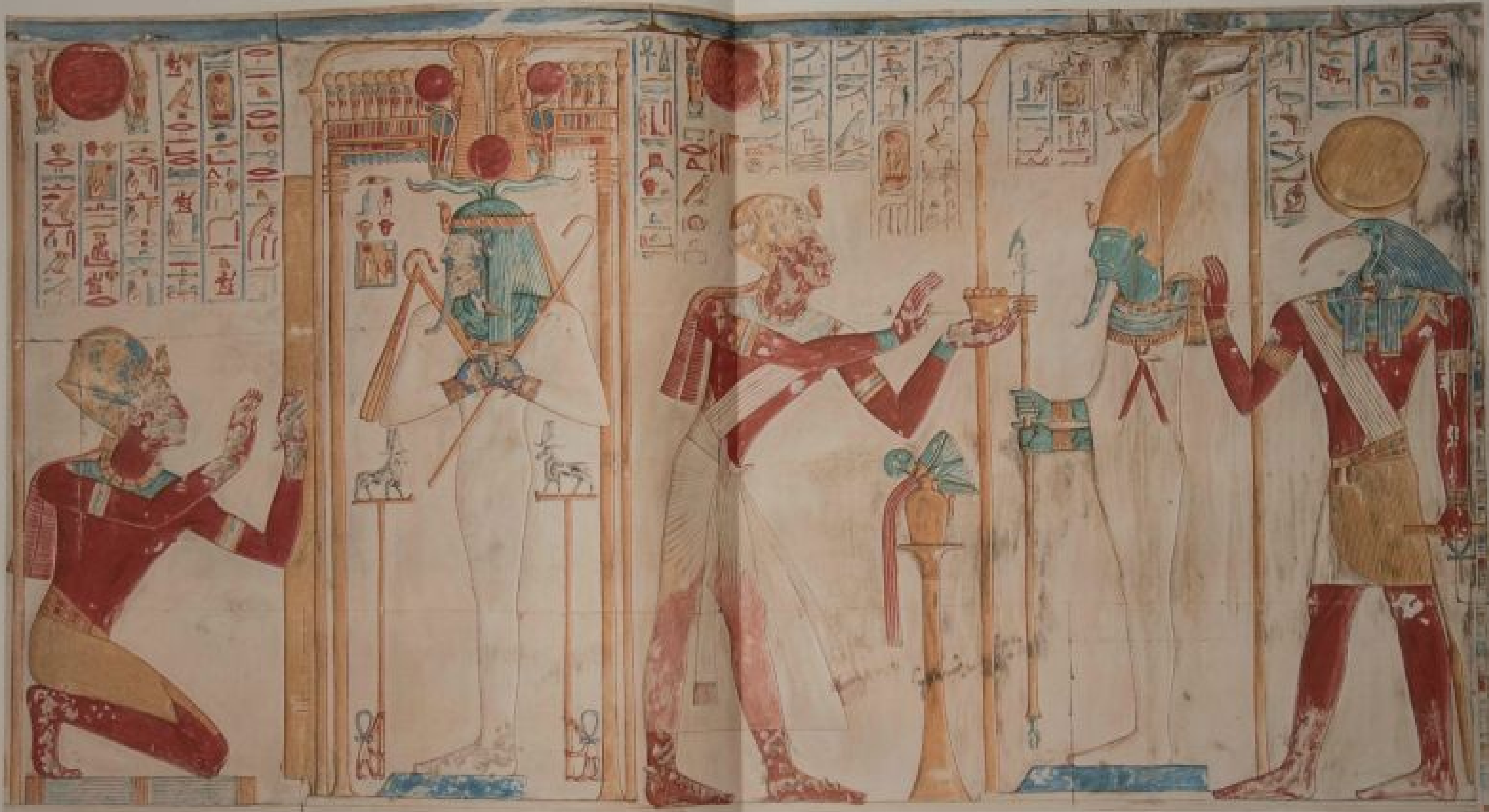

SETHOS WORSHIPS AND OFFERS NATRON TO OSIRIS

© Humboldt State University Library

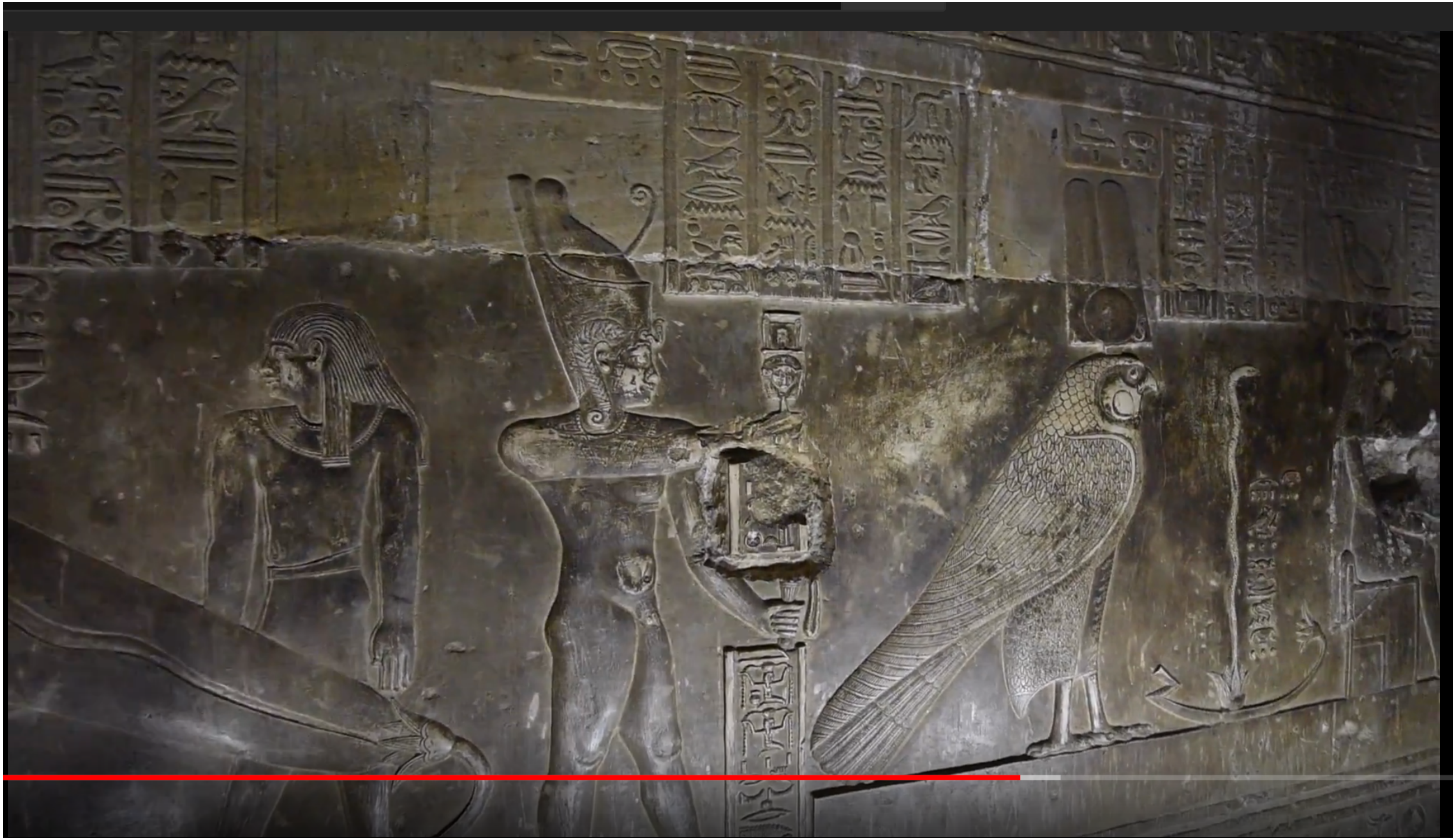

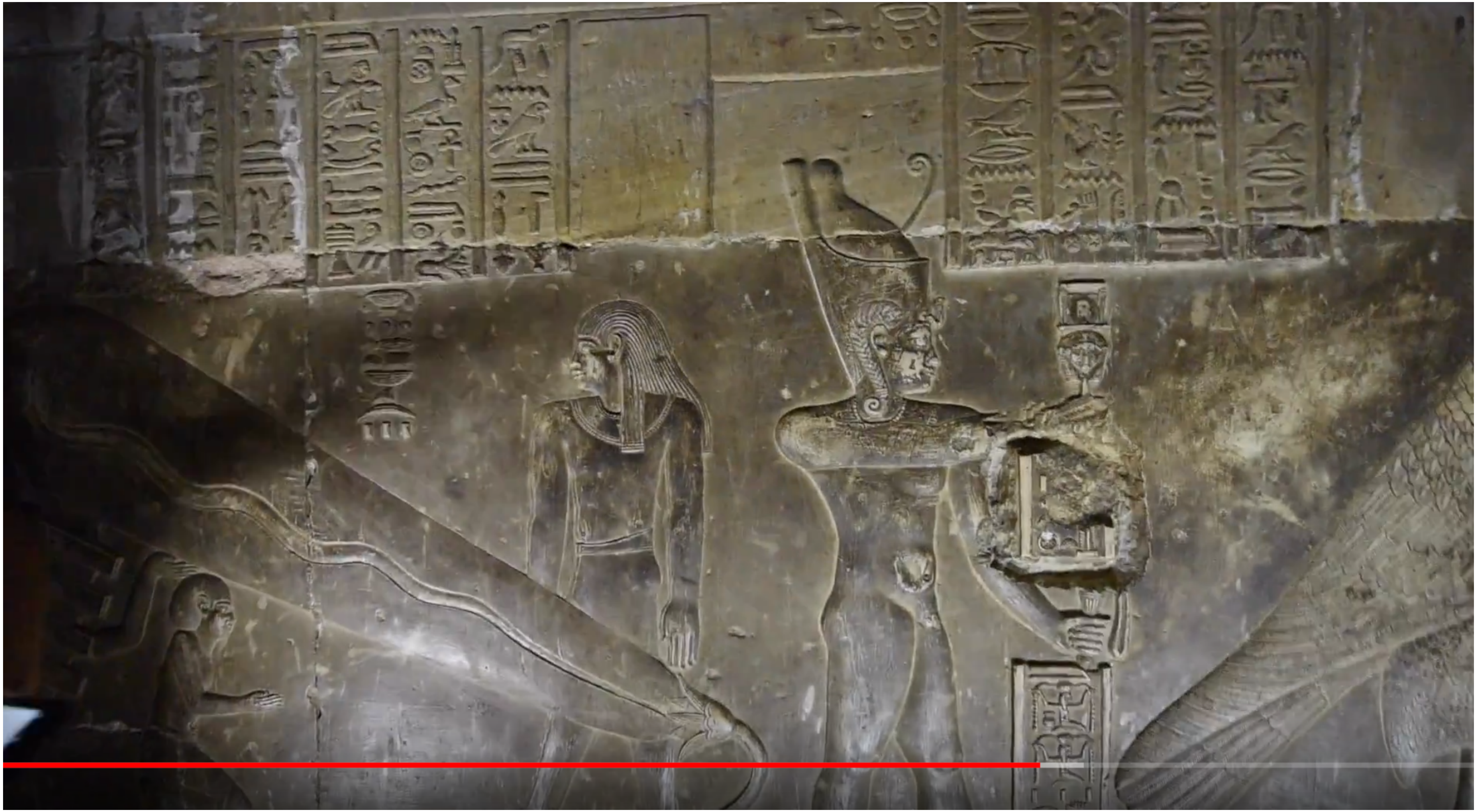

Meet The Iargans!

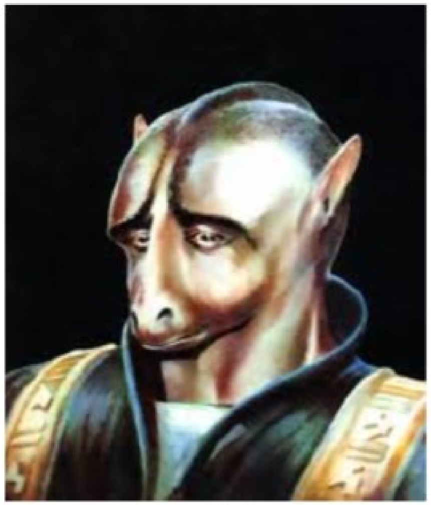

Typical Iargan

Iargan gets into photo
without being seen at
the time photo was taken!

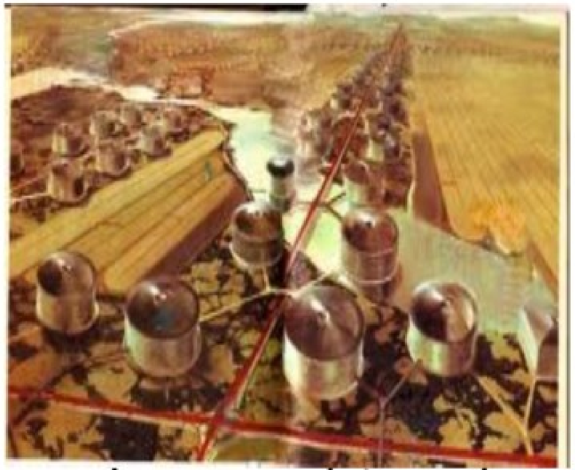

Iargan society and
transportation infrastructure
of housing units

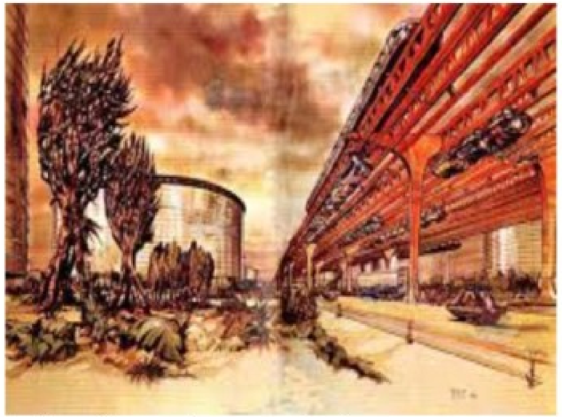

Another view and close up
of cylinder which houses 10 people each!

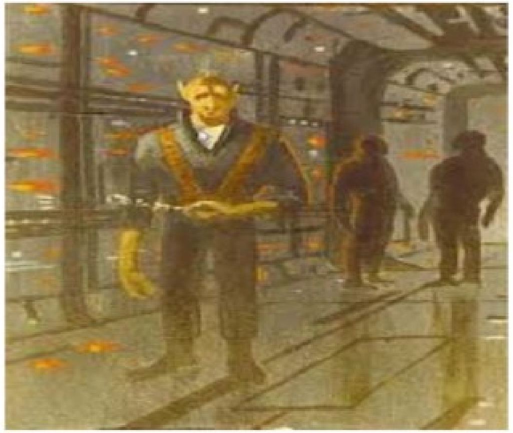

From Inside The Control Room

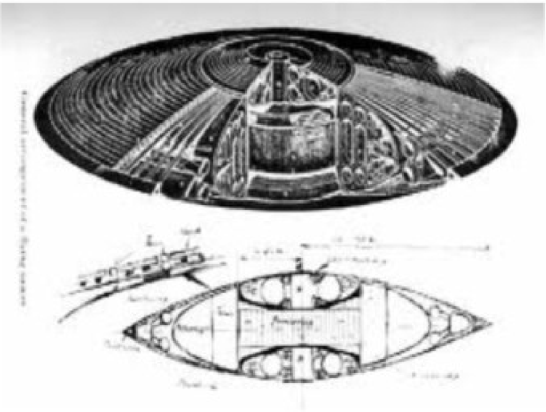

The blue print of a Sun wheel
which holds 10,000 people as well

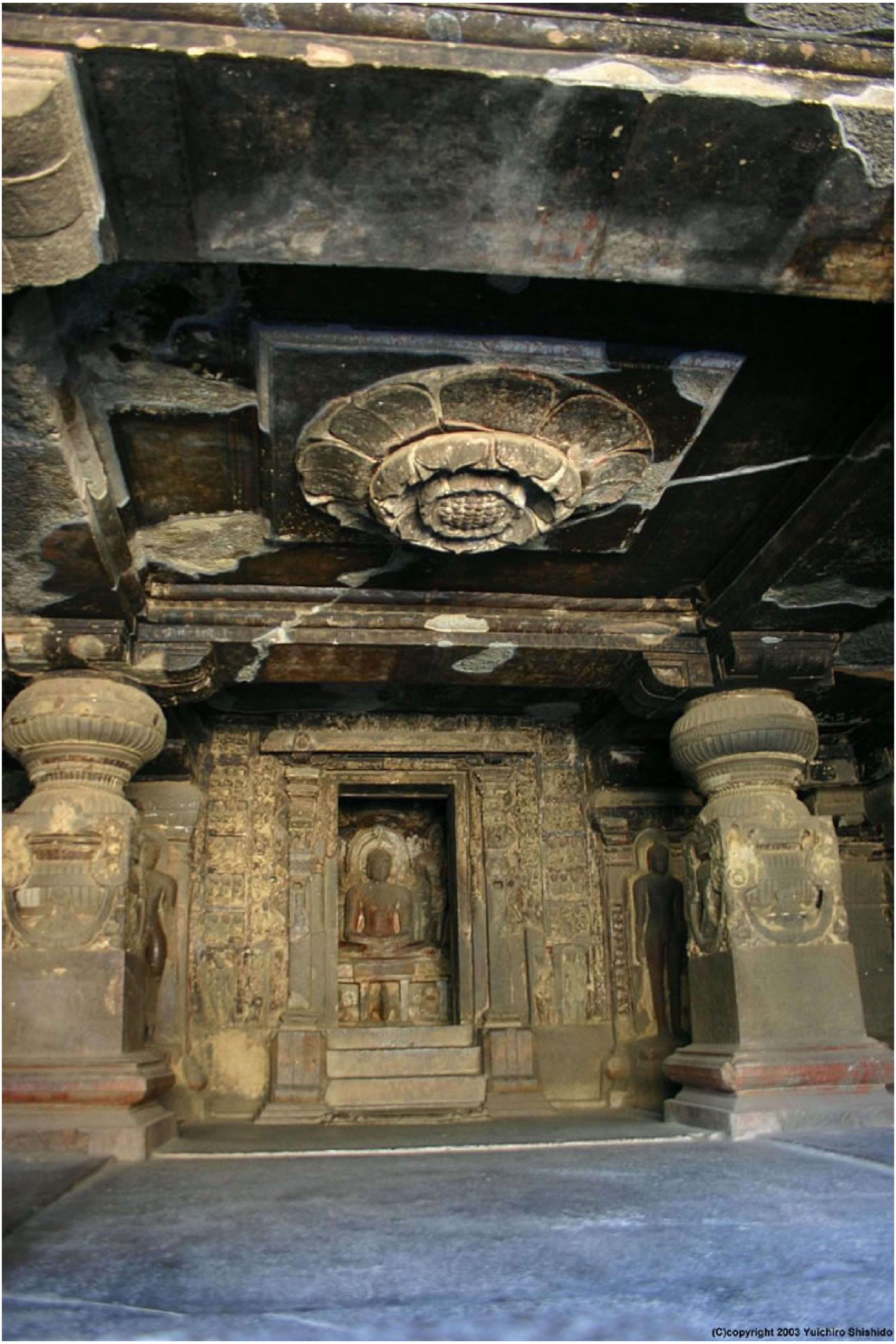

(C)copyright 2003 Yuichiro Shishido

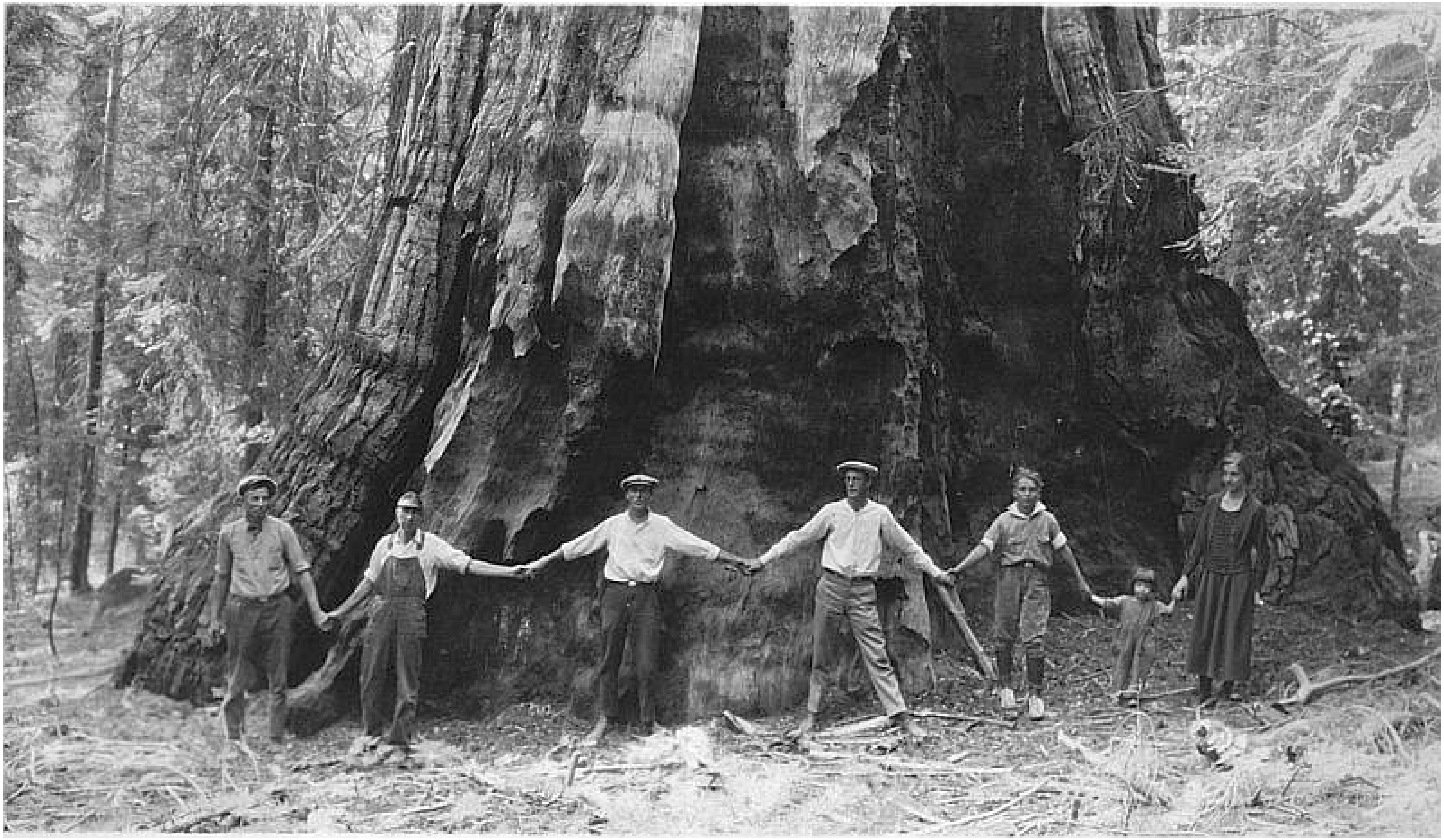

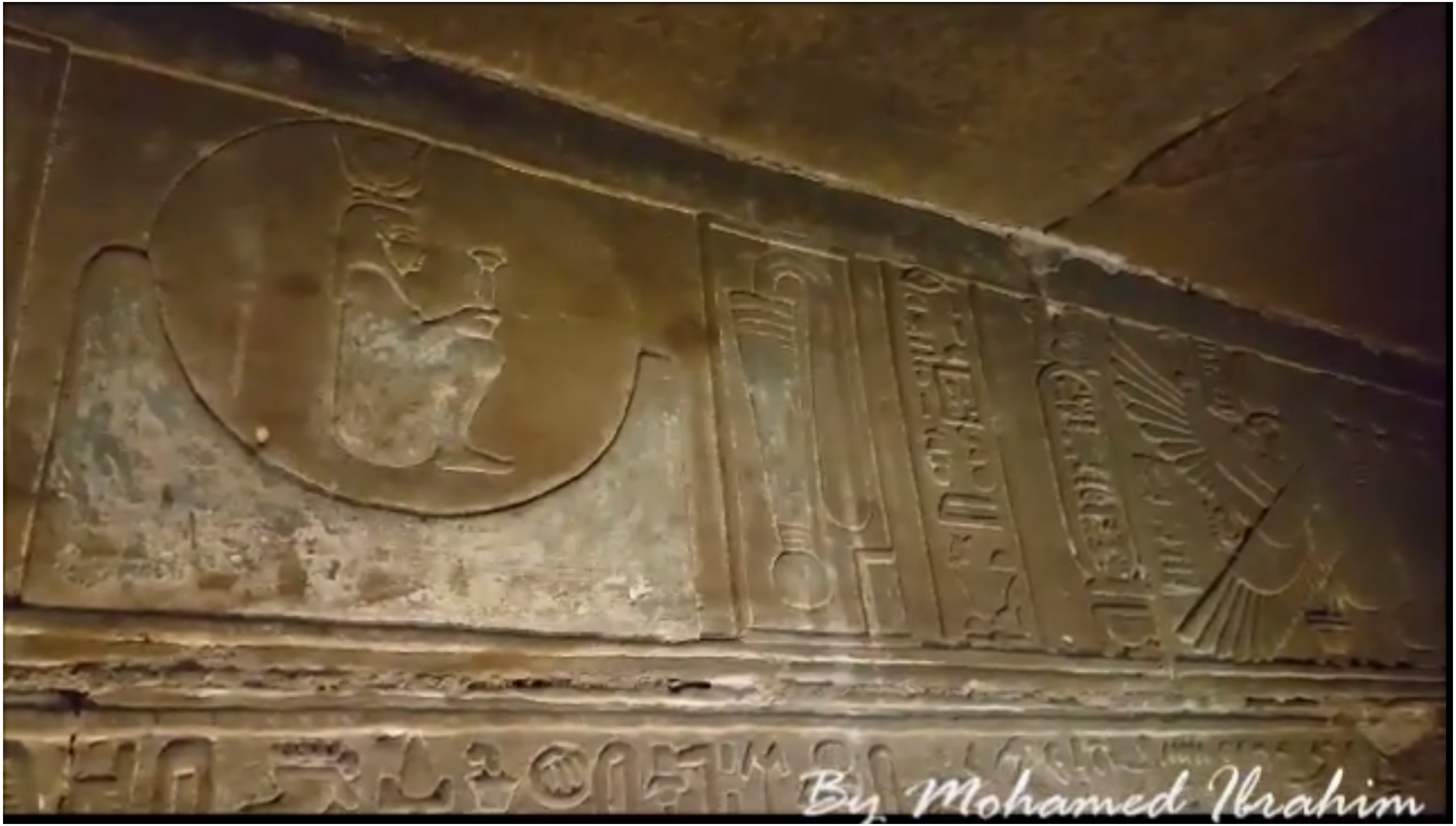
By Mohamed Ibrahim

By Mohamed Ibrahim

By Mohamed Ibrahim

By Mohamed Ibrahim

By Mohamed Ibrahim

By Mohamed Ibrahim
4:46

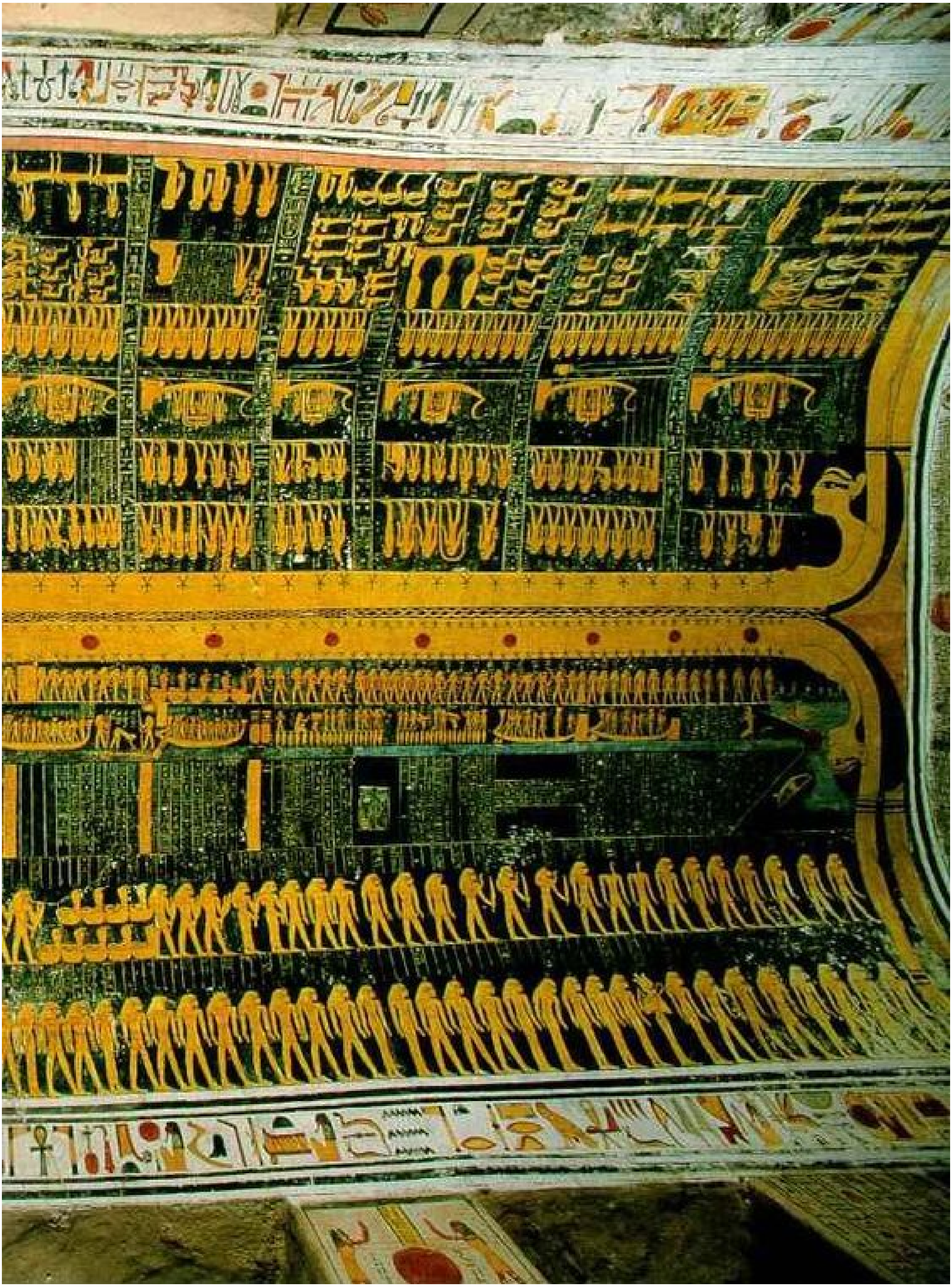

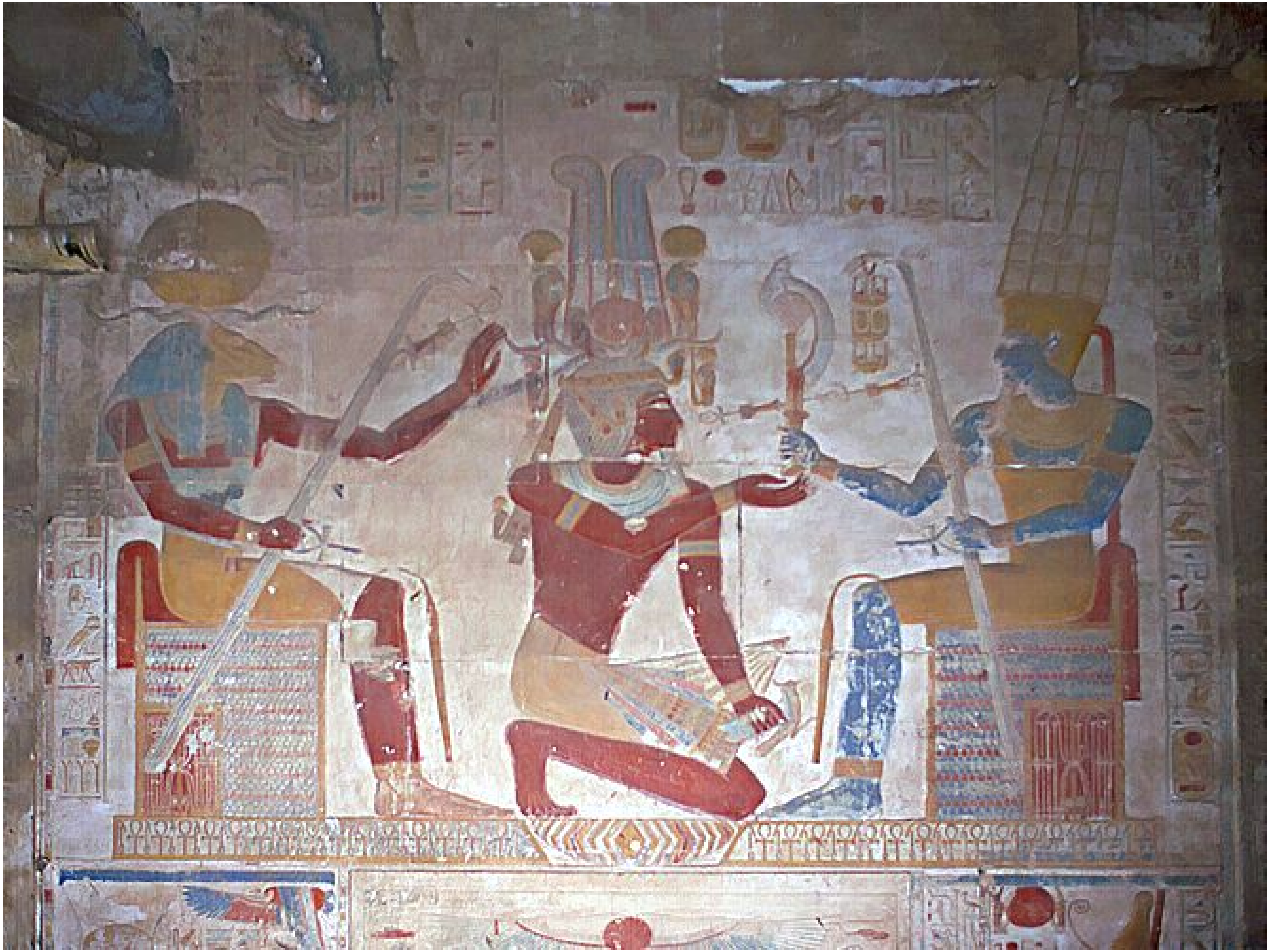

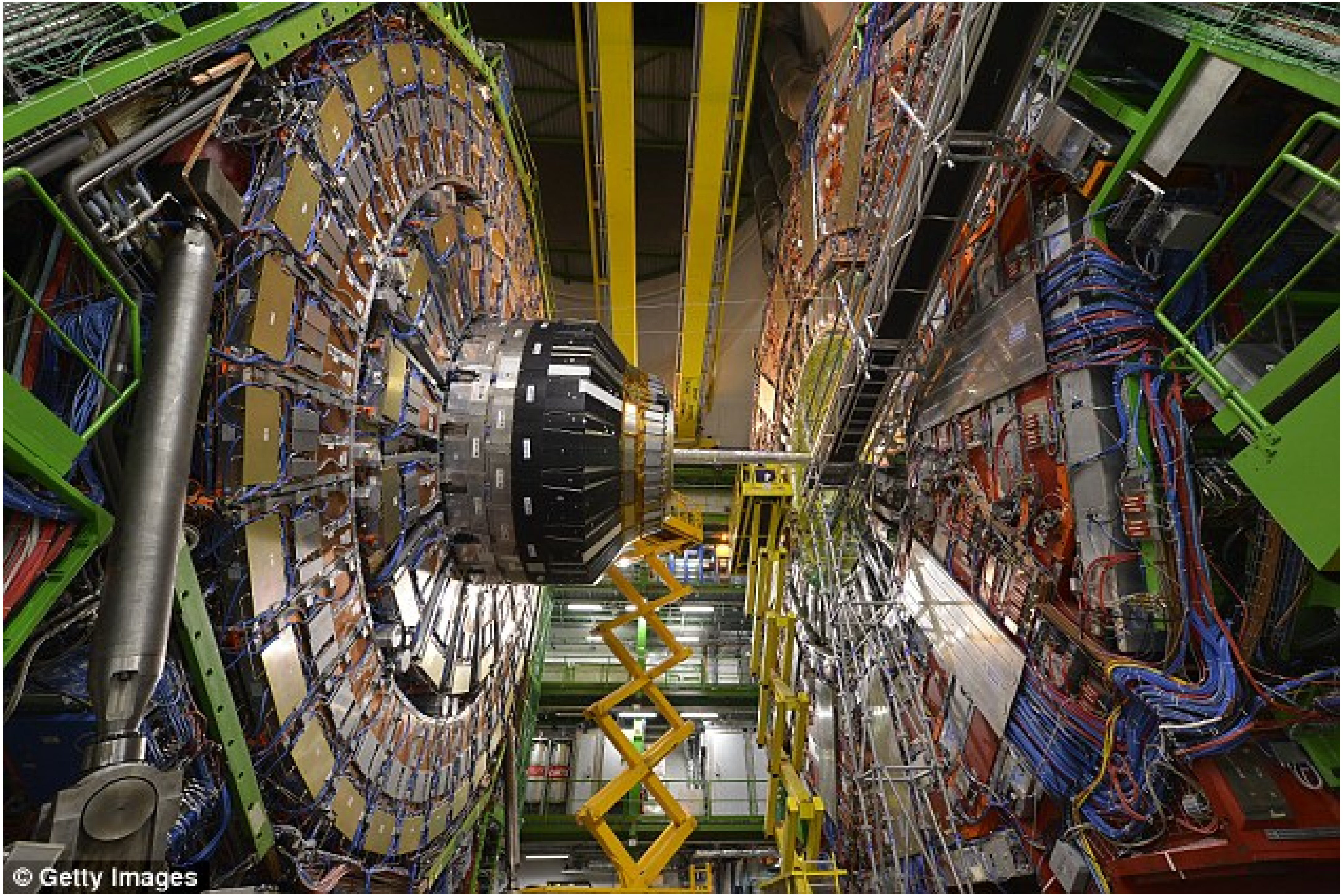

© Getty Images

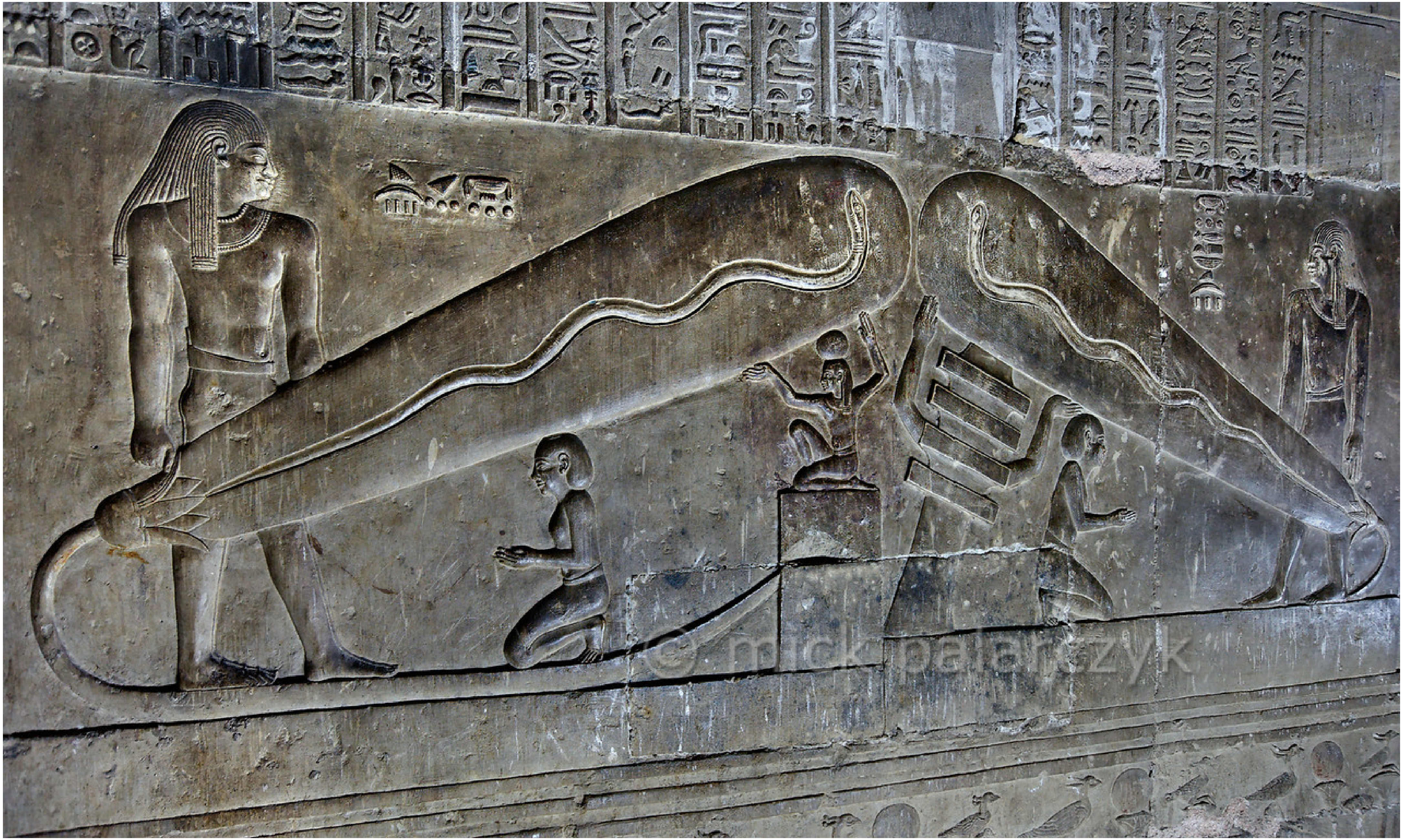

© mick palarczyk

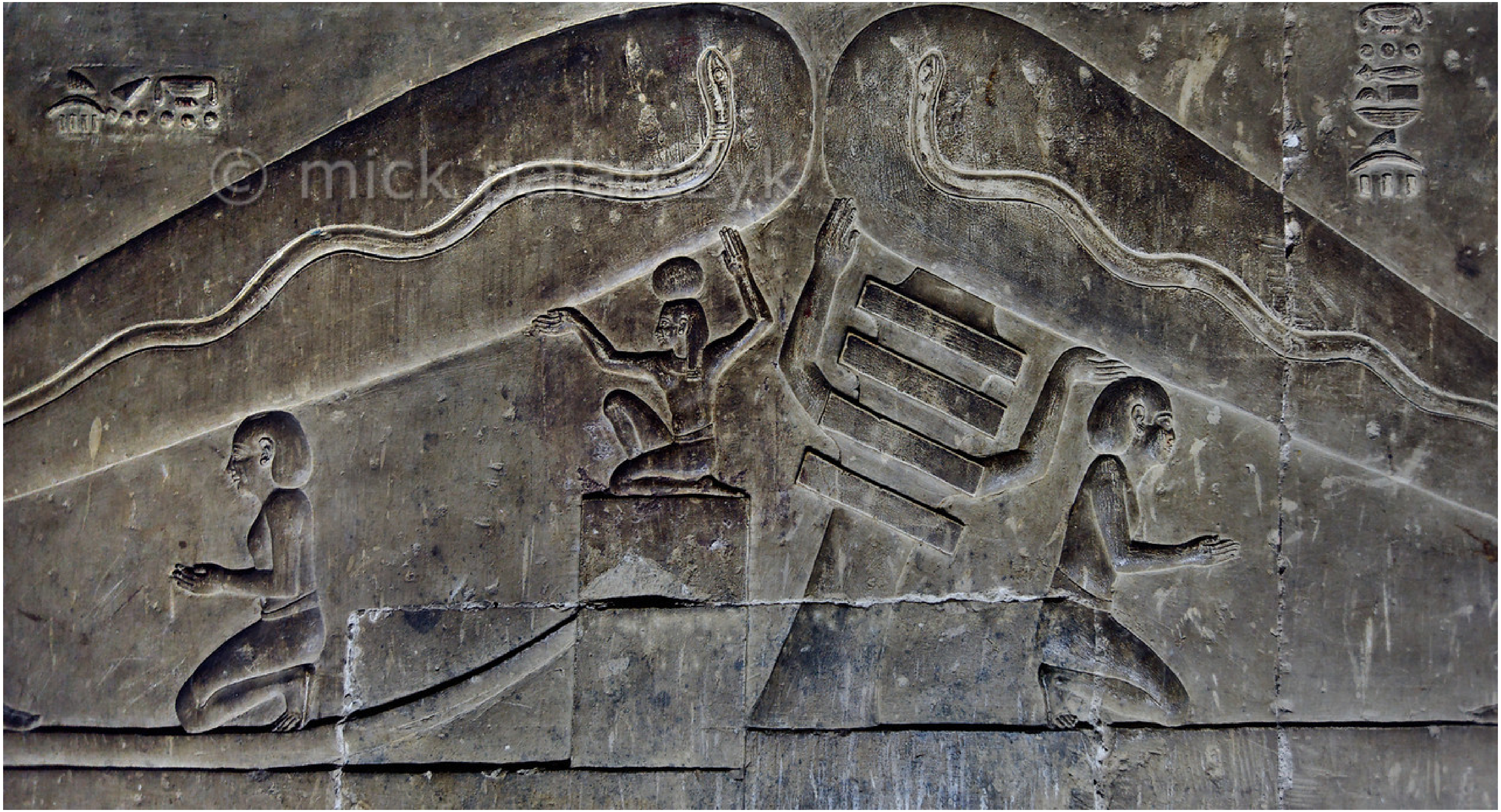

© mick [illegible]k

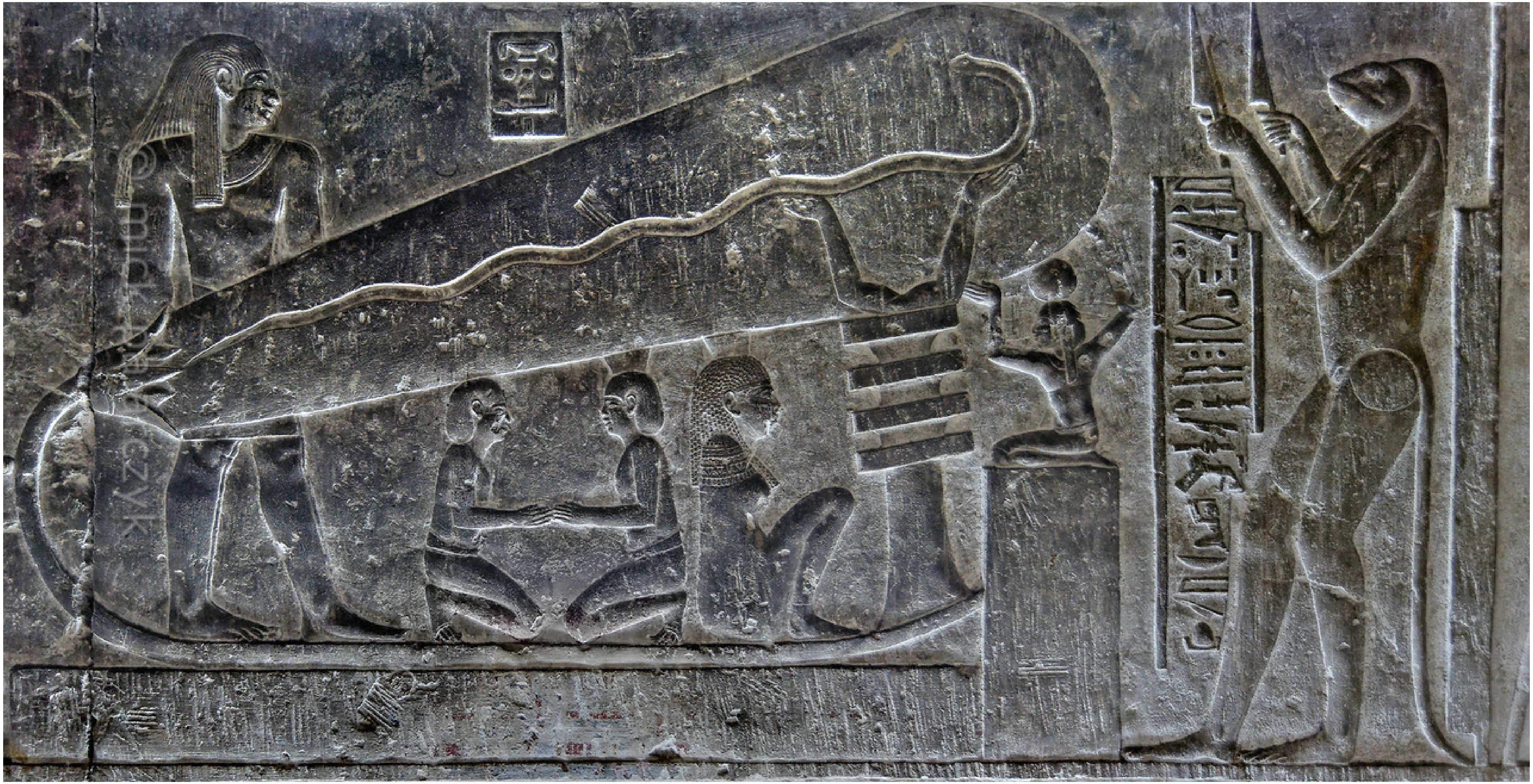

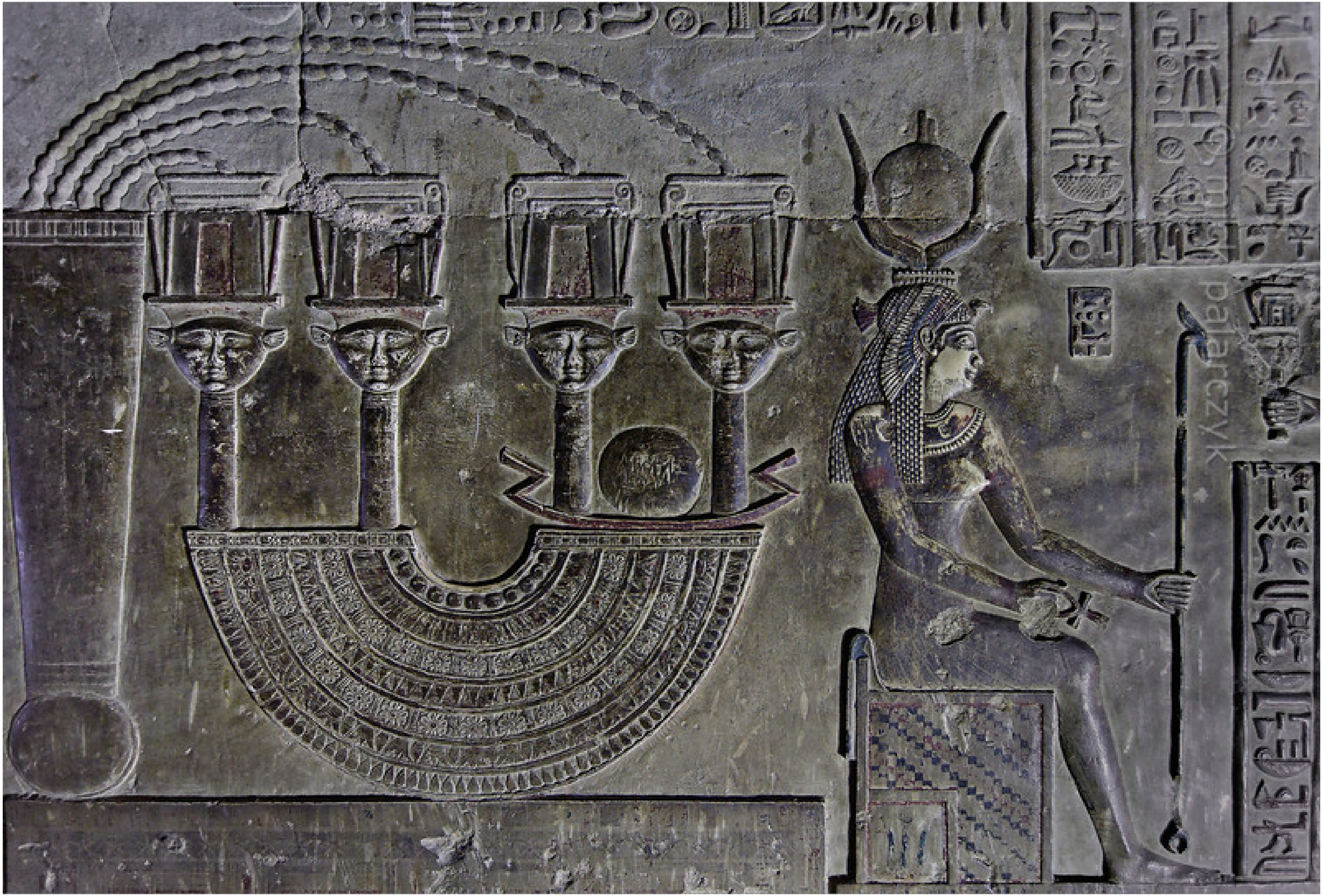
palarczyk

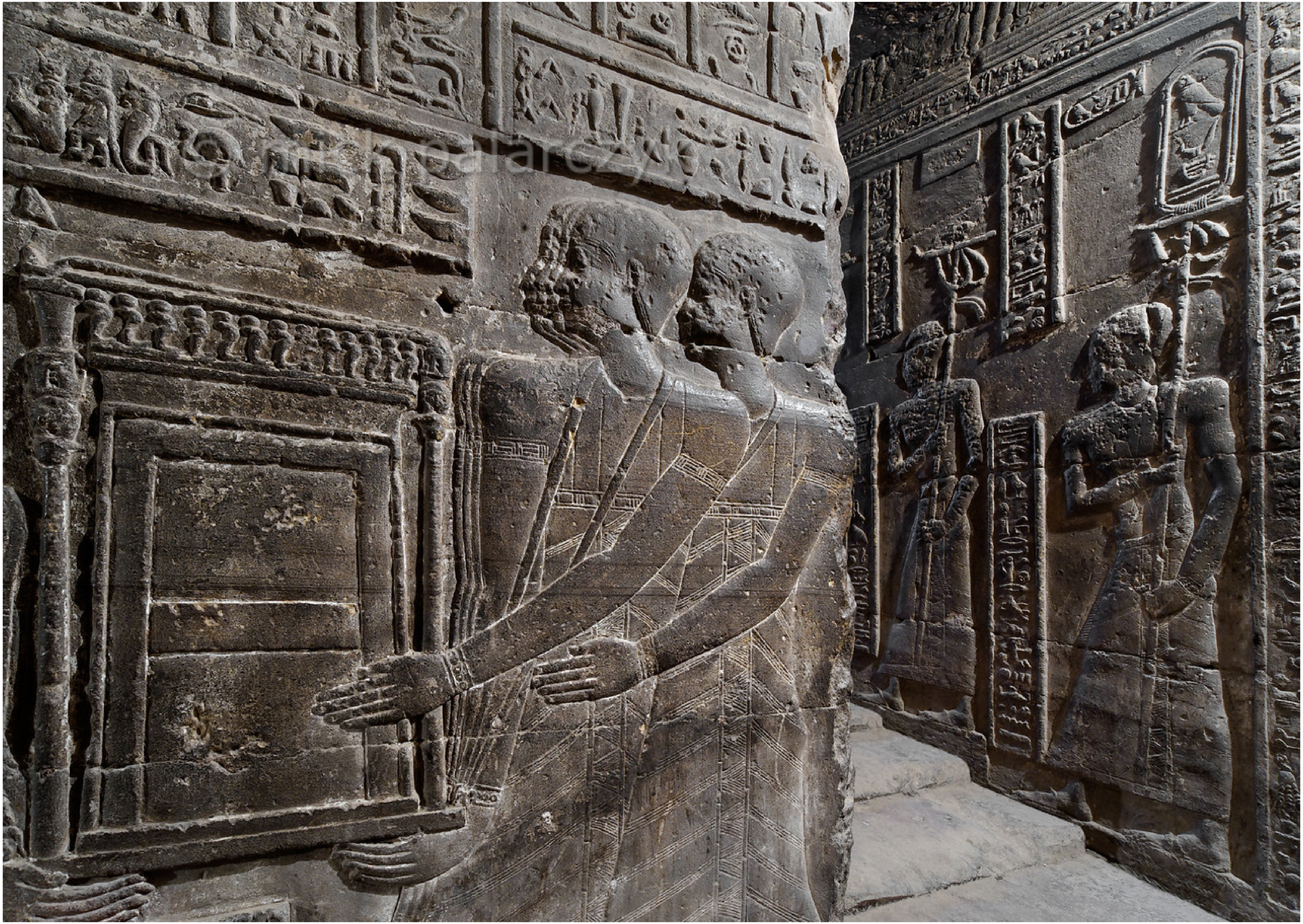

© paul smit

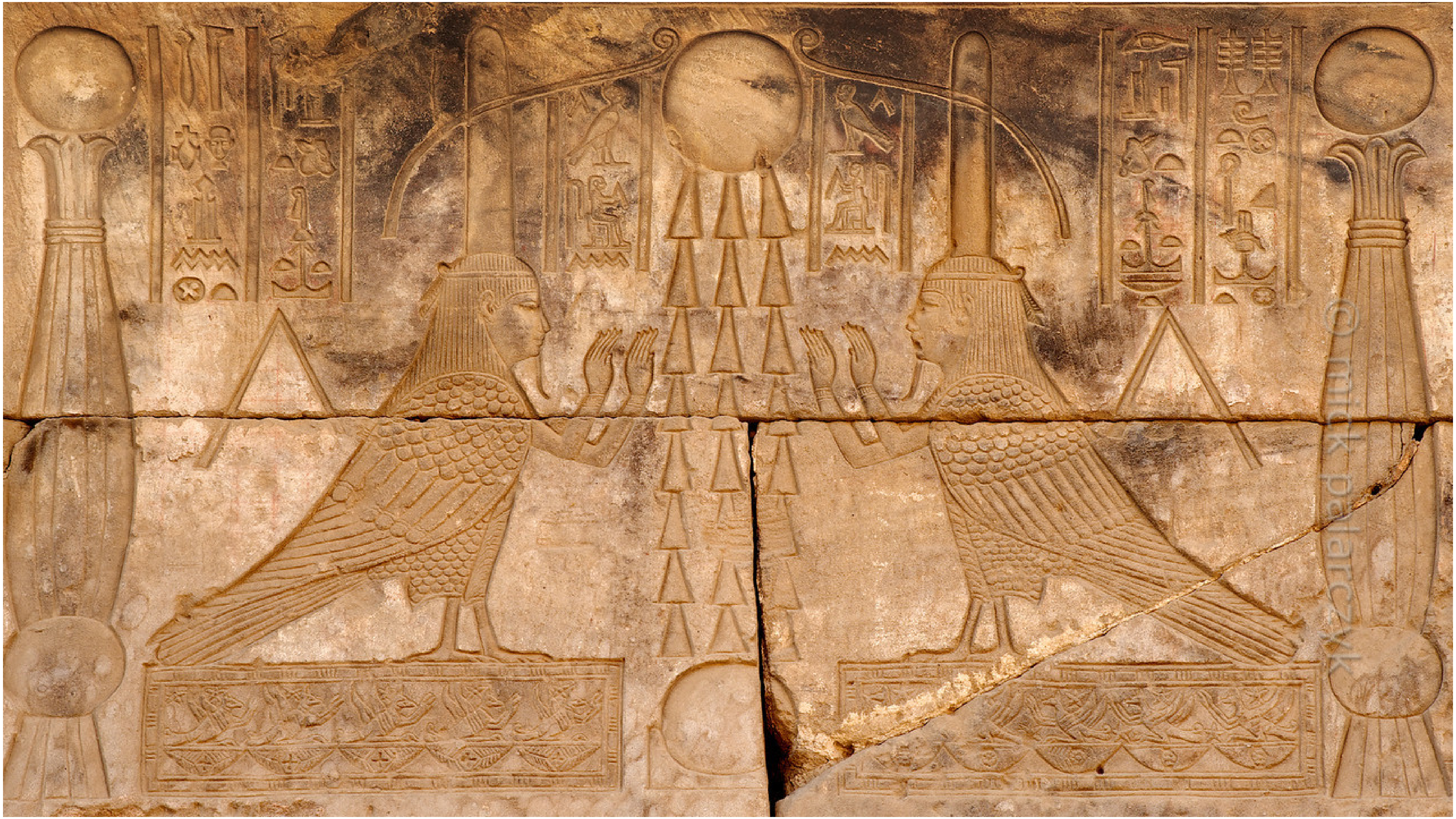

© mick palarczyk

©www.imagesofanthropology.com

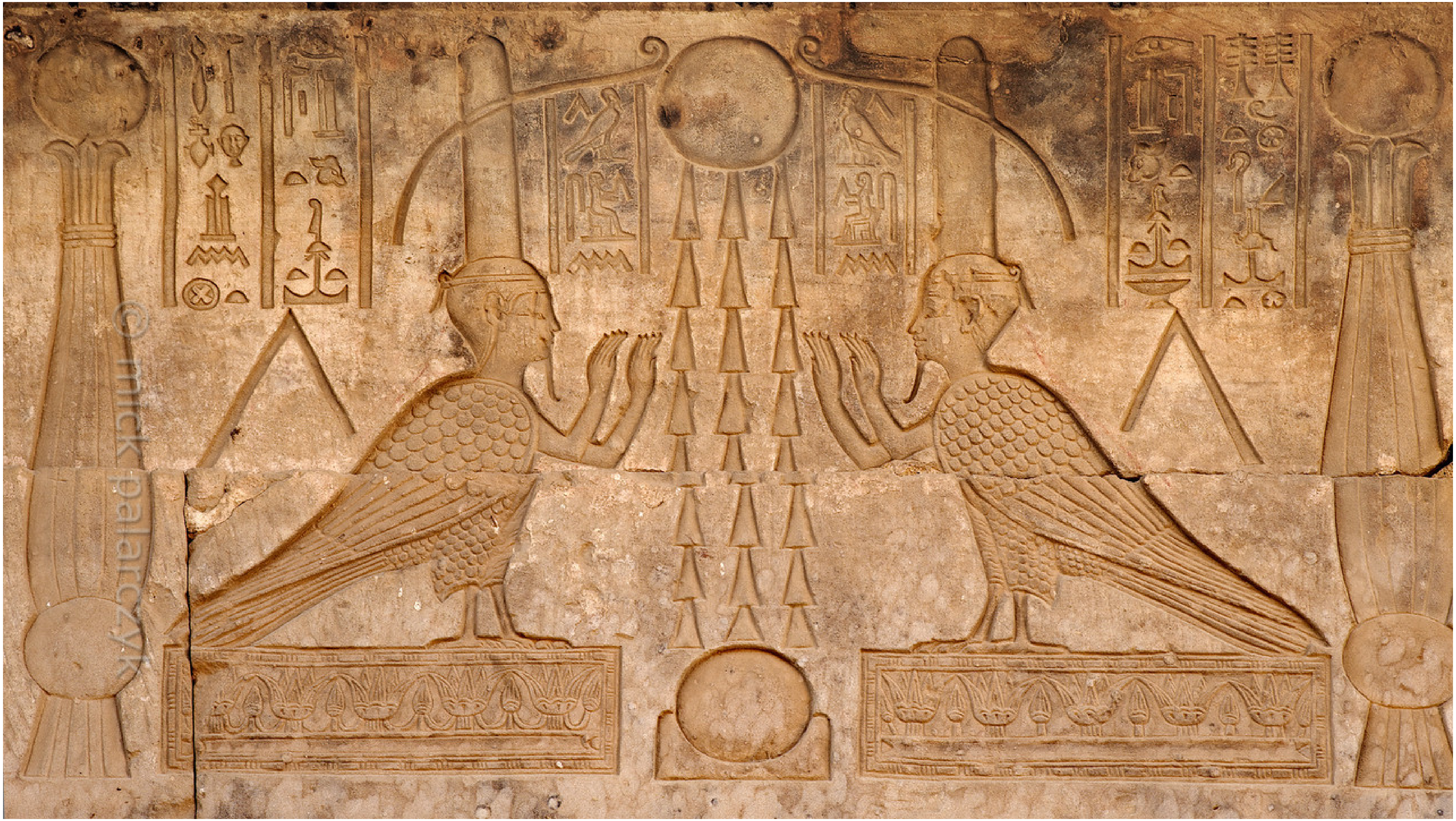

© mick palarczyk

© mick palarczyk

© palarczyk

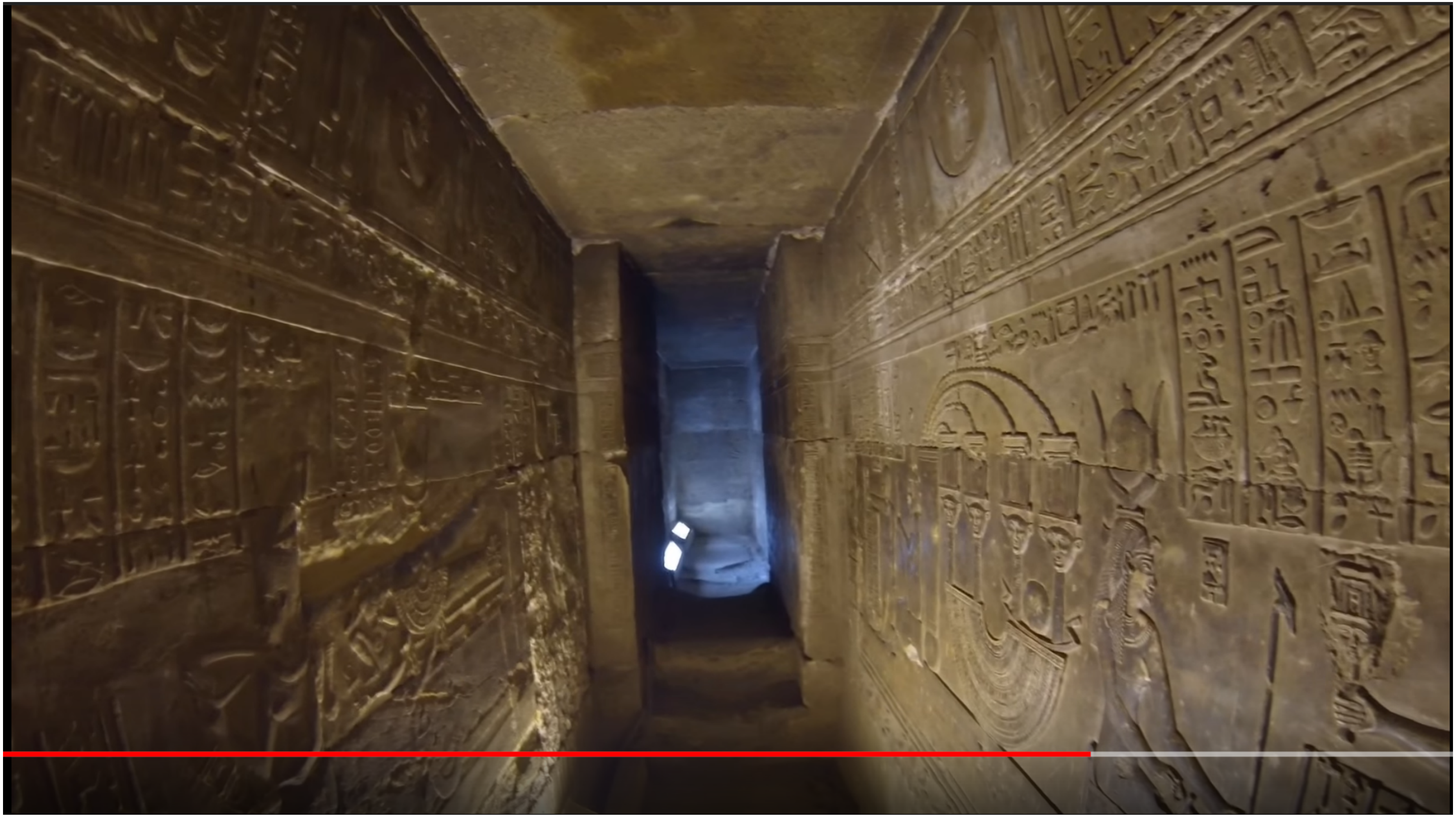

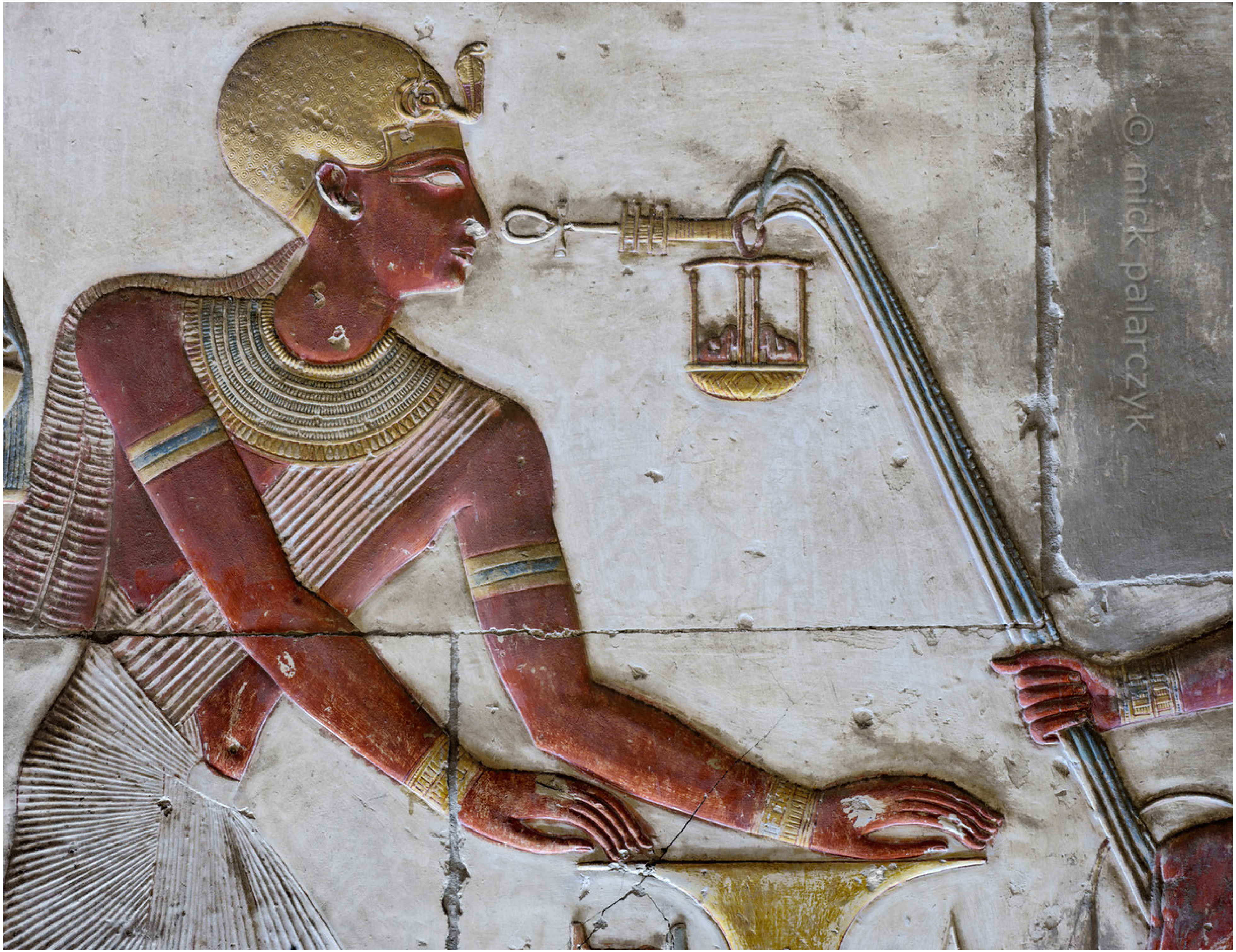

© mick palarczyk

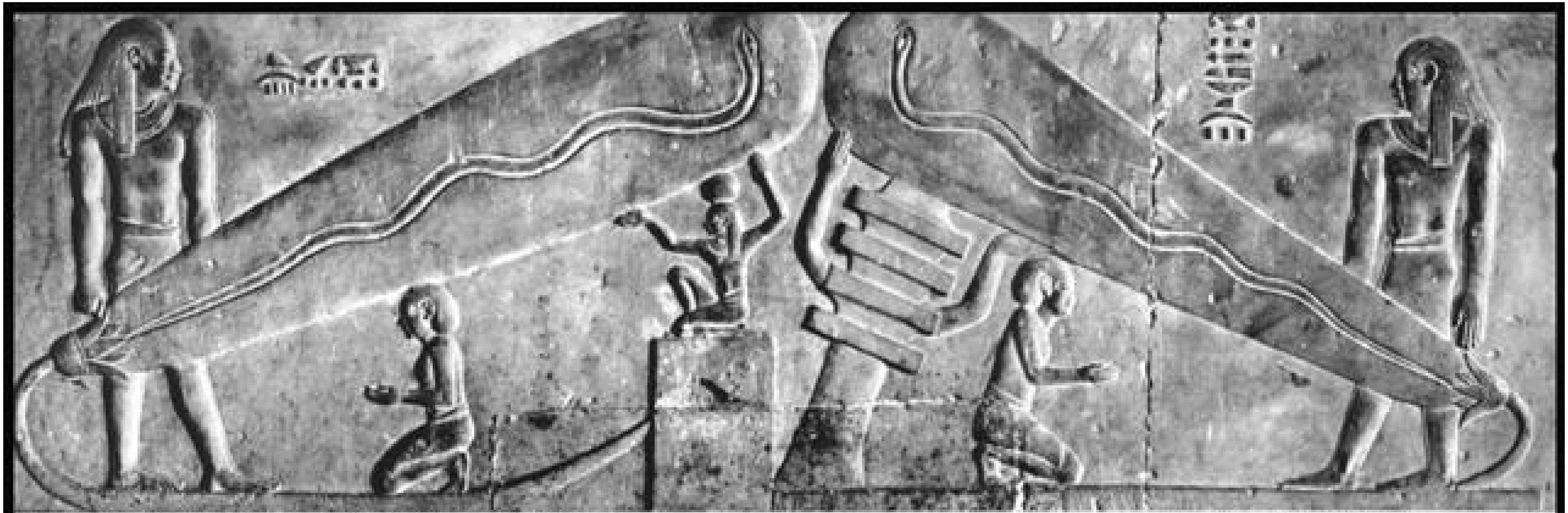

Battery-powered searchlights and filament lamps in the crypts of the Temple of Hathor (Isis) at Denderah

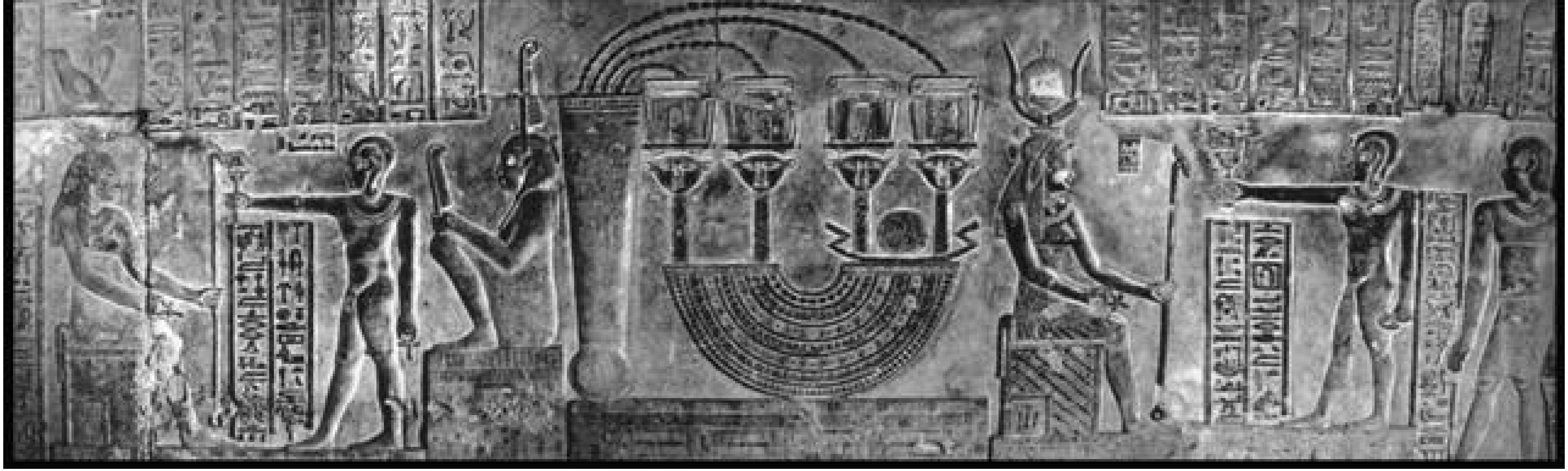

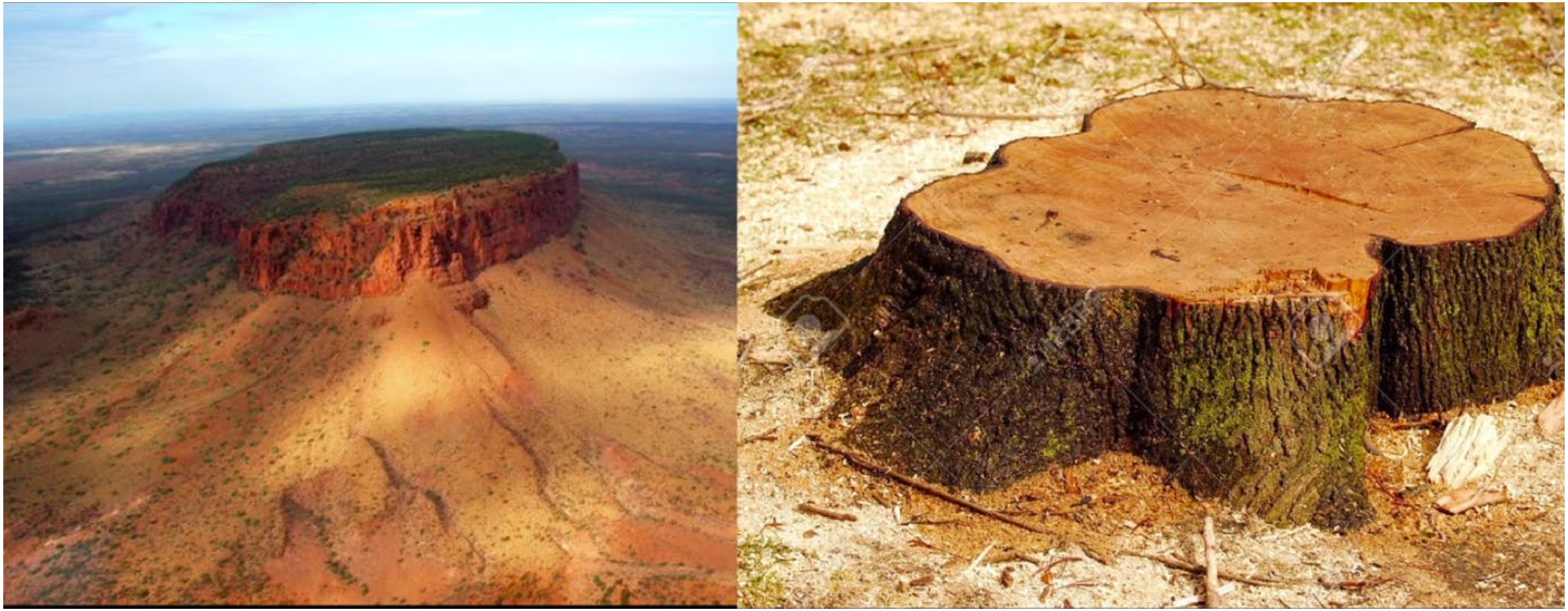

PETRIFIED FOREST
NATIONAL PARK
UNITED STATES DEPARTMENT OF THE INTERIOR
NATIONAL PARK SERVICE
NATIONAL PARK SERVICE

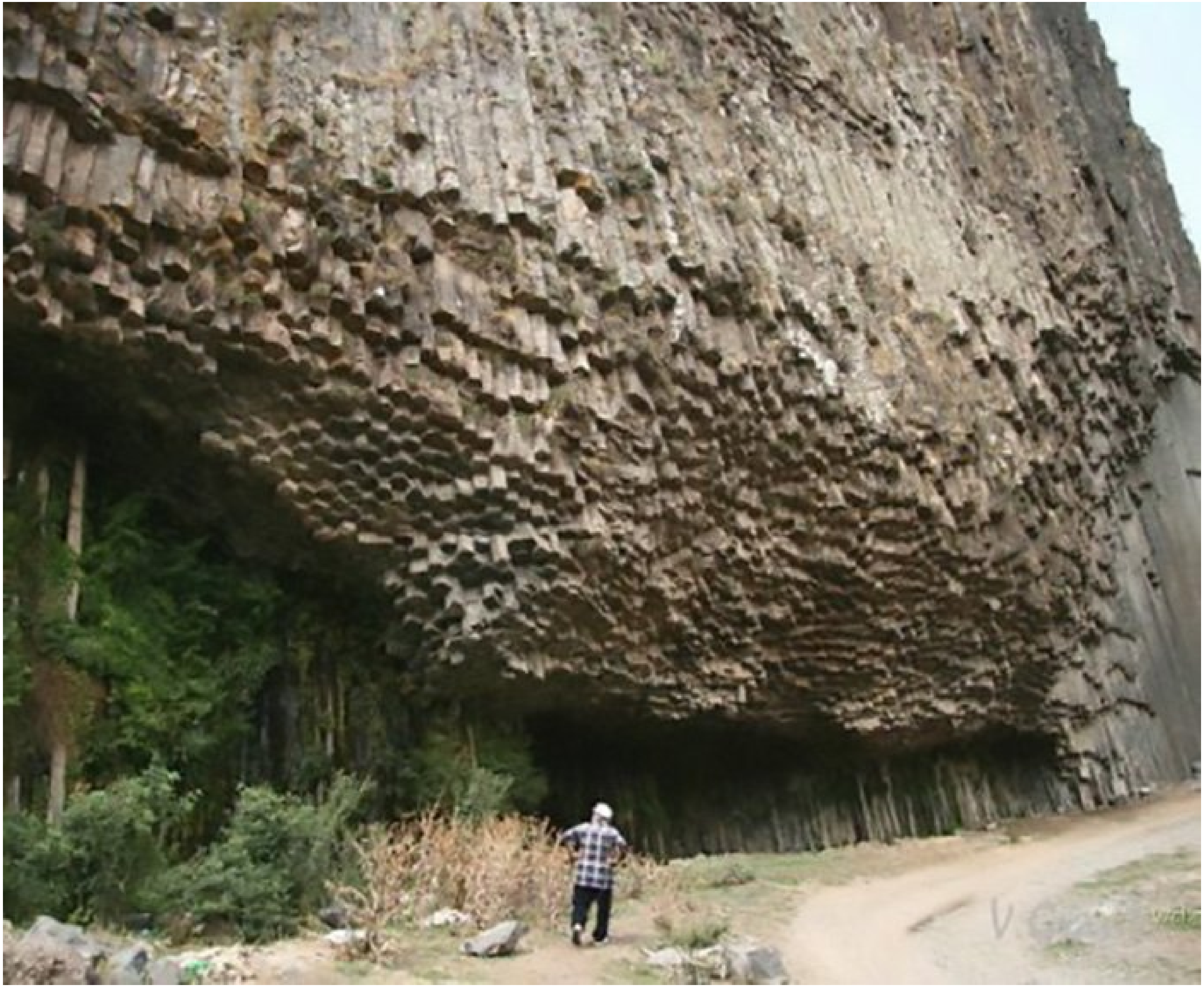

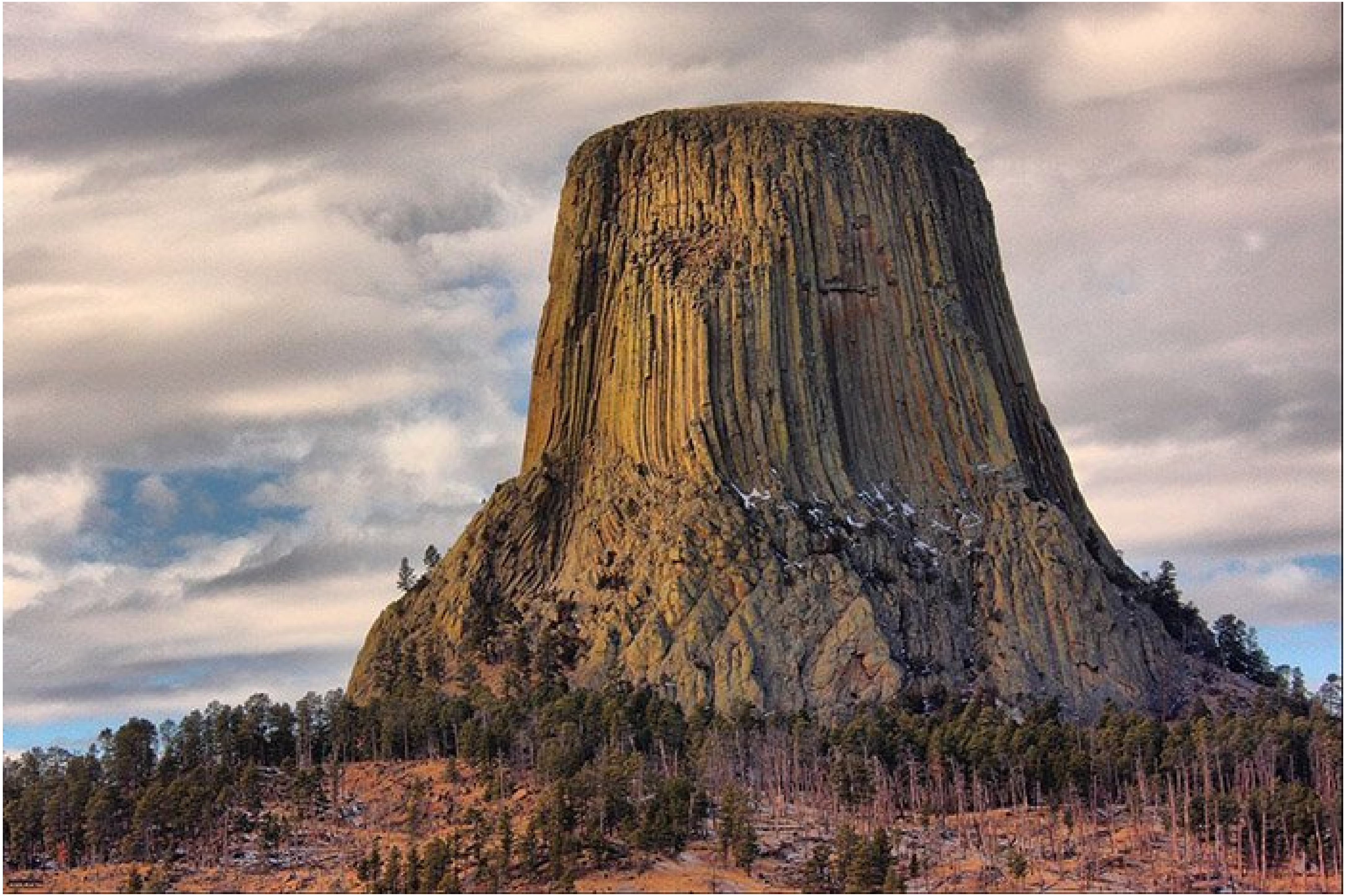

U. S. DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

COAST AND GEODETIC SURVEY
LEO OTIS COLBERT, Director

Special Publication No. 242

DEFINITIONS OF TERMS USED IN GEODETIC AND OTHER SURVEYS

By

HUGH C. MITCHELL
Principal Mathematician (retired)

QB
275
.U35
no. 242
1948

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON : 1948

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $1

National Oceanic and Atmospheric Administration

ERRATA NOTICE

One or more conditions of the original document may affect the quality of the image, such as:

Discolored pages
Faded or light ink
Binding intrudes into the text

This has been a co-operative project between the NOAA Central Library and the Climate Database Modernization Program, National Climate Data Center (NCDC). To view the original document, please contact the NOAA Central Library in Silver Spring, MD at (301) 713-2607 x124 or www.reference@nodc.noaa.gov.

LASON
Imaging Contractor
12200 Kiln Court
Beltsville, MD 20704-1387
January 1, 2006

DEFINITIONS OF TERMS USED IN GEODETIC AND OTHER SURVEYS

INTRODUCTION

Recognizing the need for authoritative definitions of surveying and mapping terms, the Federal Board of Surveys and Maps in 1938 appointed a special committee to prepare such definitions. Confronted with an apparent choice of making a complete examination of the terminology employed in the large and ever-widening fields of surveying and mapping or of limiting its attention to terms for which there appeared a current need of definition, the committee decided to do both: to gather for study and future action material relating to any terms in which the members of the Board might have even a slight interest, but to give priority in its discussions and determinations to those terms for which there appeared a present need of definition. In following this program, a number of definitions were completed and reported to the Board. At the same time there was gathered into the files of the chairman of the special committee considerable material, diverse in character and incomplete, relating to other terms which it was hoped would eventually come before the committee for study and action.

The Federal Board of Surveys and Maps was abolished in March 1942 by Executive Order (No. 9094), and its Committee on Definitions of Surveying and Mapping Terms thereby discharged. At that time, comparatively few definitions in geodesy and geodetic surveying had been completed, and it was to provide such definitions that the work which the Committee had previously been performing was continued under the direction of Captain H. W. Hemple, Chief of the Division of Geodesy, U. S. Coast and Geodetic Survey. This publication was prepared in that Division, with the major part of the work being done by Hugh C. Mitchell. After Mr. Mitchell's retirement, a few minor changes and additions were made to incorporate suggestions received after the submission of copies of the manuscript to various authorities and interested persons.

It will be noted that for each term presented there is a basic definition or description directly following the name of the term. This is intended to contain all the elements and only the elements which are essential to a complete and accurate definition or description. For many terms there are also supplementary statements (set off in separate paragraphs), designed to facilitate the interpretation of the basic definition and to call attention to other uses that may have been made of the terms defined. A few supplementary statements also contain information of historical character which is sometimes desired but not always easy to find. The object of the supplementary statements is to make the report generally more useful. *Terms defined under another heading are italicized.*

Included in the publication are definitions which had been completed by the Committee on Definitions of the Federal Board of Surveys and Maps before its dismissal. These are identified by an asterisk (*). They were not acted upon by the Board. Included also are terms which might appear to have little relation to geodetic surveying, but the great advances of recent years in the use of geodetic data for the control of other surveys make it desirable that the geodetic engineer be familiar with the language of those who may use the data which he determines.

A

aberration: annual.—See under *aberration of light (astronomy).*

chromatic.—See under *aberration of light (optics).*

diurnal.—See under *aberration of light (astronomy).*

spherical.—See under *aberration of light (optics).*

aberration of light (astronomy).—The apparent displacement in position of a heavenly body which is due to the progressive motion (velocity) of light combined with the motion of the earth itself.

Aberration of light due to the *rotation* of the earth on its axis is termed diurnal aberration; its effect is considered in first-order determinations of *astronomical azimuth* and *longitude,* but not of *latitude.* Aberration due to the *revolution* of the earth about the sun is termed annual aberration.

aberration of light (optics).—The failure of light from a point on an object, after passing through a lens, to converge to a point on the image.

In a single lens having spherical surfaces, *aberration* may be caused by differences in the *focal lengths* of the various parts of the lens: rays passing through the outer part of the lens come to a *focus* nearer the lens than do rays passing through its central part. This is termed spherical aberration and, being due to the faulty figure of the lens, is eliminated by correcting that figure. A lens so corrected is called an *aplanatic lens.* Aberration may also result from differences in the wave lengths of light of different colors: light of the shorter wave lengths (violet end of the spectrum) comes to a *focus* nearer the lens than light of the longer wave lengths (red end of the spectrum). This is termed chromatic aberration, and is practically eliminated over a moderate range of wave lengths by using a composite lens, called an *achromatic lens,* composed of parts having different *dispersive powers.* See *refraction: index of refraction.*

absolute unit.—See under *unit.*

acceleration.—Rate of change of speed.

In the *c. g. s. system* of physical measurements, acceleration is expressed in terms of centimeters per second per second. See *gal.*

accessories: corner (United States public-land surveys).*—Physical objects adjacent to *corners,* to which such *corners* are referred for their future identification or restoration.

Accessories include *bearing trees,* mounds, pits, ledges, rocks, and other natural features to which distances or directions, or both, from the *corner* or *monument* are known. Strictly speaking, such accessories are a part of the monument.

accidental error.—See under *error.*

accumulated discrepancy.—See under *discrepancy.*

accumulated divergence (leveling).—See under *divergence.*

accuracy.*—Degree of conformity with a *standard.*

Accuracy relates to the quality of a result, and is distinguished from *precision* which relates to the quality of the operation by which the result is obtained.

The standard of reference may be (1) an exact value, such as the sum of the three *angles* of a plane triangle is 180°; (2) a value of a conventional *unit* as defined by a physical representation thereof, such as the *international meter,* defined by the *International Prototype Meter* (bar) at the National Bureau of Standards; (3) a survey or map value determined by refined methods and deemed sufficiently near the ideal or true value to be held constant for the control of dependent operations, such as the *adjusted elevation* of a *permanent bench mark* or the *graticule* of a *map projection.* The accuracy attained by field surveys is the product of the instructions to be followed in the execution of the work and the *precision* with which those instructions are followed.

achromatic lens.—A lens which transmits light without separating it into component colors.

An achromatic lens transmits light with a negligible amount of *chromatic aberration.* This is accomplished by means of lens elements of glass having different *indices of refraction,* so proportioned that the *chromatic aberration* in one element is compensated by that in another. Lenses that are made achromatic are usually also made *aplanatic,* that is, free of *spherical aberration.*

achromatic telescope.—A telescope having an *achromatic lens* for its object glass.

adjusted angle.—See under *angle.*

adjusted elevation.—See under *elevation.*

adjusted position.—See under *position.*

adjusted value.—See under *value.*

adjustment (leveling).—The determination and application of *corrections* to orthometric differences of elevation or to *orthometric elevations,* to make the *elevations* of all *bench marks* consistent and independent of the *circuit closures.*

adjustment: angle method of adjustment (triangulation and traverse).—A method of *adjustment of observations* which determines *corrections* to *observed angles.*

The angle method of adjustment may be used where a chain of single triangles is to be adjusted. For an extensive *triangulation* figure with overlapping triangles, the *direction method of adjustment* is preferred.

adjustment: collimation.—See under *collimation adjustment.*

adjustment: direction method of adjustment (triangulation and traverse).—A method of *adjustment of observations* which determines *corrections* to observed directions.

In the direction method, each *angle* is considered as the difference between two *directions,* for each of which a separate *correction* is determined. The direction method is used in the adjustment of *triangulation* figures which are composed of overlapping triangles, but for some work, where the survey consists of a chain of single triangles, the *angle method of adjustment* may be preferred.

adjustment: local.—See under *adjustment: station adjustment.*

adjustment: station.—The satisfying of conditions existing among *angles* as measured at a survey station. Also called local adjustment.

In a station adjustment, two kinds of conditions may require consideration: the sum of a series of measured *angles* closing the *horizon* may not equal exactly 360°; the sum of several contiguous *angles* measured separately may not equal the sum angle measured as a single *angle.* The term *observed angles* may be applied to the measured *angles* after they have been corrected by station adjustment.

adjustment correction (leveling).—See under *correction.*

adjustment of observations.—The determination and application of *corrections* corresponding to *errors* affecting the observations, making the observations consistent among themselves, and coordinating and correlating the derived data.

The object of *adjustment* is to provide the best (most probable) values of results obtainable from a series of survey observations, and to coordinate and correlate those results within the survey itself and with previously existing basic data. This is accomplished by mathematical analysis, as by the use of the *method of least squares;* or mechanically by graphical methods such as employed in plane-table surveying. The mathematical adjustment involves the solution of certain *condition equations* or *observation equations,* such as *angle equations, side equations, length equations,* and *latitude, longitude,* and *azimuth equations.* See, also, *adjustment: station.*

aerial map.—See under *chart: aeronautical chart.*

aerial navigation map.—See under *chart: aeronautical chart.*

aeronautical chart.—See under *chart.*

aeronautics.*—The science and art of navigation through the atmosphere.

air navigation chart.—See under *chart: aeronautical chart.*

air navigation map.—See under *chart: aeronautical chart.*

Airy theory of isostasy.—See under *isostasy.*

Aitoff equal-area map projection.—See under *map projection.*

Albers conical equal-area map projection.—See under *map projection.*

alidade.—The part of a surveying instrument which consists of a sighting device, with index, and reading or recording accessories.

The alidade of a *theodolite* or surveyor's *transit* is the upper part of the instrument: the *telescope, micrometer microscopes,* or *verniers,* and accessories, mounted on what is called the "upper motion" of the instrument, and used in observing a direction or angle on the graduated circle, which is mounted on the "lower motion."

The alidade used in topographic surveying consists of a straight-edge ruler carrying a *telescope* or other sighting device, and used in recording a direction on the *plane-table* sheet, which corresponds to the graduated circle of the *theodolite.* If a *telescope* is used, the instrument is often termed a *telescopic alidade.*

The movable arm of a sextant is an alidade.

alidade: telescopic.—A usual designation for an instrument composed of a *telescope* mounted on a straight-edge ruler, and used with a *plane table* in *topographic surveying.* See *alidade.*

alignment (alinement).—The placing of points along a straight line. Also, the location of points with reference to a straight line, or to a system of straight lines.

The use of the term alignment in surveying should be limited to operations associated with straight lines.

alignment: curve of.—See under *curve of alignment.*

alignment correction, tape.—See under *tape.*

almucantar.—Any small circle on the *celestial sphere* parallel with the *horizon.* Also termed parallel of altitude and circle of equal altitude.

alt-azimuth instrument.—An instrument equipped with both horizontal and vertical graduated circles, for the simultaneous observation of horizontal and vertical directions or angles.

The alt-azimuth instrument derives its name from the combination of the terms *altitude* and *azimuth.* Many *theodolites* and engineer's and surveyor's *transits* are alt-azimuth instruments; the *vertical circle* (instrument) used by the Coast and Geodetic Survey has a small horizontal circle as well as a larger, more *precise vertical circle,* and may be considered an alt-azimuth instrument, but the term is little used for surveying and mapping instruments. Its use may properly be restricted to designating an instrument of the *theodolite* type constructed for use in astronomical work, and sometimes called an astronomical theodolite or the universal instrument.

altimeter.—An *aneroid barometer* used for determining *elevations.*

altitude.—The *vertical angle* between the plane of the *horizon* and the line to the object which is observed or defined.

In astronomy and surveying, the altitude is positive if the object is above the *horizon,* and negative if below it. In surveying, a positive altitude is also termed an angle of elevation; a negative altitude, an angle of depression. In astronomy, altitude is the *complement* of the *zenith distance.* In photogrammetric surveying, altitude applies to *elevation* above *datum* of points in space.

altitude: circle of equal.—See under *almucantar.*

altitude: parallel of.—See under *almucantar.*

amplitude of vibration (pendulum).—The *arc* passed over by a *pendulum* in executing one *vibration.*

In the present use of the *Mendenhall pendulum,* the amplitude of vibration is measured in millimeters. In mathematics (harmonic analysis), the amplitude is the semi-arc or distance out from the mean position.

aneroid barometer.—See under *barometer.*

angle (general).—That relationship of lines or surfaces which is measured by the amount of *rotation* necessary to make one coincide with or parallel to another.

See, also, *angle: solid.*

An angle may be expressed as a difference in directions of lines which lie: (1) In a plane, giving a *plane angle.* (2) In a curved surface, giving a *spherical angle, a spheroidal angle,* etc. (3) In two planes, measuring the inclination of one plane to the other, giving a *dihedral angle.*

In each case the magnitude of the angle is reckoned in units derived from a complete circumference of a circle. There are several systems of such units, of which the following are the most commonly used:

a) The degree ($^\circ$) is 1/360th part of a circumference, and is the unit of the sexagesimal system used in this country. The degree is divided into 60 minutes ($60'$), and the minute into 60 seconds ($60''$).

b) The hour (^{h}) is 1/24th part of a circumference, and is the unit used in the angular expression of time. The hour is related to the degree by the equation: $1^h=15^\circ$. The hour is divided into 60 minutes (60^m), and the minute into 60 seconds (60^s). [Note that in the sexagesimal system the symbols for minutes and seconds are ($'$) and ($''$), and in the time system the symbols for minutes and seconds are (^{m}) and (^{s}).] An angle expressed in $\_\_^\circ\ \_\_'\ \_\_''$ is said to be in *arc* measure; expressed in $\_\_^h\ \_\_^m\ \_\_^s$, in time measure.

c) The *grad* or *grade* (^{g}) is 1/400th part of a circumference, and is the unit of the centesimal system used in France. It is related to the degree by the equation: $100^g=90^\circ$. The *grad* is divided into 100 centesimal minutes (100‵); and a centesimal minute is divided into 100 centesimal seconds (100‶). While, as indicated above, the prime and double prime with reverse slant (‵,‶) are used to designate centesimal minutes and seconds, the common forms of the prime and double prime (′, ″) are used by some authors for the same purpose.

d) The radian (^{r}) is the angle at the center of a circle subtended by an *arc* having exactly the same length as the radius of the circle. One *radian* equals 360° divided by 2π or 57.30° approximately. Its value in seconds of *arc* is 206,205″. It is used in astronomy and navigation.

e) The mil is $\frac{1}{6400}$ part of a circumference; it subtends 0.98 (approximately 1) linear unit at a distance of 1000 such units. It is related to the degree by the equation: 1 mil$=0.05625^\circ$. It is used in military operations.

Sometimes, a particular angle may be defined by giving the length of chord which subtends the angle at a stated distance. Thus, the angle which 1 foot subtends at a distance of 39 miles is equal to 1″. This is often expressed by the following saying "A second is a foot at forty miles."

In surveying operations, an angle is classified in various ways, as follows:

1) According to the position or character of the surface in which it lies: *horizontal angle, vertical angle,* oblique angle, *spherical angle, spheroidal angle.*

2) According to the method by which it is obtained: *observed angle, concluded angle, adjusted angle.*

3) According to the function which it serves in the computation of the survey data: *distance angle, azimuth angle.*

4) According to the relative position it occupies in a geometric figure: *interior angle, exterior angle.*

angle: adjusted.—An *adjusted value* of an *angle.*

An adjusted angle may be derived from either an *observed angle* or from a *concluded angle.*

angle: azimuth (astronomy).—The *angle* less than 180° between the plane of the *celestial meridian* and the *vertical plane* containing the observed object, reckoned from the direction of the elevated pole.

In astronomic work, the azimuth angle is the *spherical angle* at the *zenith* in the *astronomical triangle,* pole-zenith-star. In geodetic work, it is the *horizontal angle* between the celestial pole and the observed terrestrial object.

angle: azimuth (surveying).—An *angle* in *triangulation* or in *traverse* through which the computation of *azimuth* is carried.

In a simple *traverse,* every *angle* may be an azimuth angle. Sometimes, in a *traverse,* to avoid carrying *azimuths* over very short lines, supplementary observations are made over comparatively long lines, the *angles* between which form azimuth angles.

In *triangulation,* certain *angles,* because of their size and position in the figure, are selected for use as azimuth angles, and enter into the formation of the *azimuth condition equation* (*azimuth equation*).

angle: complement of.—The difference between an *acute angle* and one right *angle* (90°).

angle: concluded.—An *interior angle* between adjacent sides of a closed figure, which is obtained by subtracting the sum of all the other *interior angles* of the figure from the theoretical value of the sum of all the *interior angles.*

The concluded angle is most frequently met with in triangulation, where it is obtained by subtracting the sum of two known *angles* of a triangle from 180° plus the *spherical excess* of the triangle.

angle: deflection.—A *horizontal angle* measured from the prolongation of the preceding line, right or left, to the following line.

angle: dihedral.—An *angle* between two planes.

A dihedral angle is measured in a plane perpendicular to the line of intersection of the planes forming the *angle,* and is acute, right, oblique, or straight according to the nature of the plane angle which represents it.

angle: direct.—See under *angle to right.*

angle: distance.—An *angle* in a triangle opposite a side used as a base in the solution of the triangle, or a side whose length is to be computed.

In a chain of single triangles, as the computation proceeds through the chain, two sides in each triangle are used: a known side and a side to be determined: the *angles* opposite these sides are the distance angles.

angle: explement of.—The difference between an *angle* and four right angles (360°).

angle: exterior.—The *angle* between any side of a polygon and adjacent side extended.

Exterior angle is also used to designate the outside *angles* formed by a line intersecting two parallel lines. It has likewise been used in surveying to designate the *explement* of an *angle.* This latter use is rare.

angle: horizontal.—An *angle* in a horizontal plane.

The directions may be to objects in the horizontal plane; or they may be the lines of intersection of the horizontal plane with the *vertical planes* containing the objects.

angle: interior.—An *angle* between adjacent sides of a closed figure, measured by the *arc* connecting those sides on the inside of the closed figure.

The three *angles* of a triangle are interior angles.

angle: observed.—An *angle* obtained by direct instrumental observation.

A measured *angle* which has been corrected for local conditions (see *adjustment: station*) only at the point of observation, is considered an observed angle.

angle: parallactic.—See under *astronomical triangle.*

angle: refraction.—See under *refraction angle.*

angle: solid.—The integrated angular spread at the vertex of a cone, pyramid, or other solid figure.

The unit of measure of a solid angle is the steradian, a solid angle at the center of a sphere which is subtended by a portion of the surface of the sphere having an area equal to the square of the radius of the sphere.

angle: spherical.—An *angle* between *great circles* on a sphere.

A spherical angle is measured either: a) by the *dihedral angle* of the planes of *great circles* or b) by the plane angle between tangents to *great circles* at their intersection.

angle: spheroidal.—An *angle* between two curves on a *spheroid,* measured by the *angle* between their tangents at the point of intersection.

angle: supplement of.—The difference between an *angle* and two right angles (180°).

angle: vertical.—An *angle* in a *vertical plane.*

In surveying, one of the directions which form a vertical angle is usually either (a) the direction of the *vertical* (*zenith*), and the *angle* is termed the *zenith distance;* or (b) the line of intersection of the *vertical plane* in which the *angle* lies with the plane of the *horizon,* and the *angle* is termed the *angle of elevation* or *angle of depression,* or simply the *altitude* (plus or minus, as the case may be). The vertical angle between two directions, neither of which lies in the plane of the *horizon* or coincides with the *vertical,* is usually obtained from the combination of two vertical angles as defined above.

angle equation.—See under *equation.*

angle method of adjustment (triangulation and traverse).—See under *adjustment.*

angle of depression.—See under *altitude.*

angle of elevation.—See under *altitude.*

angle of incidence.—See under *refraction of light.*

angle of inclination.—An *angle of elevation* or an *angle of depression.* See *altitude.*

angle of refraction.—See under *refraction of light.*

angle to right.—The *horizontal angle* measured clockwise from the preceding line to the following one.

angles: repetition of.—See under *repetition of angles.*

annual aberration.—See under *aberration of light* (*astronomy*).

annual parallax.—See under *parallax.*

anomalistic month.—See under *month.*

anomalistic year.—See under *year.*

anomaly (astronomy).—The *angle* at the sun between the radius vector of a planet and the *line of apsides* of its orbit, reckoned from *perihelion.*

The term defined above is also called the true anomaly, when it is desired to distinguish it from the eccentric anomaly, which is the corresponding angle at the center of the orbit; or from the mean anomaly, which is what the true anomaly would become if the planet had a uniform angular motion.

anomaly (general).—A deviation from a norm.

anomaly (geodesy).—A deviation of an observed value from a theoretical value, due to an abnormality in the observed quantity.

anomaly: Bouguer.—A difference between an observed value of *gravity* and a theoretical value at the point of observation, which has been corrected for the effect of *topography* only, the topography being considered as resting on a plane of indefinite extent.

In practice, a Bouguer *anomaly* is obtained through a series of approximations. First, the *topography* is treated as an indefinite slab of a thickness equal to the elevation of the station above *sea level,* the rate of change of *gravity* with elevation depending upon the assumed density of the slab, and being about five-eighths the normal value of that ratio. This first approximation gives the main part of the *anomaly.* The determination of the effect of this indefinite slab is the first step in the Hayford-Bullard (or Bullard) method of isostatic reduction.

Second, allowance is made for the departure of the actual *topography* from the indefinite slab. This correction is usually small. The distance out from the station for which it is computed is a matter of convenience; while no precise rule has been adopted for selecting this distance, it has been suggested that it should increase with elevation of station. The practice in the United States Coast and Geodetic Survey is to compute this second correction out to Zone O of the Hayford divisions of the surface of the earth, that is, 166.7 kilometers or about 100 miles out from the station. See *template* (*templet*).

A third approximation is sometimes made by taking account of the curvature of the sea-level surface. The numerical value of this correction is small, unless the computation is carried to a great distance from the station. In the work of the United States Coast and Geodetic Survey, no account is taken of this correction and, according to Helmert, it should not be used. However, reasons have been advanced favoring its inclusion in a complete computation of a Bouguer anomaly.

For use in reducing *gravity* observations to obtain even an approximate approach to a theoretical value, the Bouguer method has proved of little or no value. It ignores *isostatic compensation* and creates large fictitious *anomalies;* it neglects large amounts of matter existing between the *geoid* and the *spheroid;* and it takes no account of actual *topography* beyond the limited distance used in the computation of the second correction (above). It has, however, proved useful to oil companies in their geophysical investigations: if Bouguer anomalies are available at a number of stations within a limited area, values of this *anomaly* at other points within the area can be obtained by interpolation, and from these, serviceable values of *gravity* are obtained by computation.

The use of the pendulum in determining the figure of the earth came in at about the beginning of the nineteenth century, and with that use came the application of Bouguer's rule for the reduction of gravity observations. The method continued in favor until Clarke and Helmert, applying the theories of Pratt and Airy which are now known as the principle of *isostasy,* showed that it did not come anywhere near producing reliable values of theoretical gravity. Its present use is limited to theoretical studies and to the practical application given above.

The Bouguer anomaly derives its name from the French mathematician, Pierre Bouguer (1698–1758), who did not develop the method itself, but who proposed the rule or formula expressing the effect of a slab of indefinite extent, as considered in making the first approximation (above). In its time, the Bouguer rule has also borne the names of Young (Thomas Young, 1773–1829) and Poisson (Simon Denis Poisson, 1781–1840).

anomaly: deflection.—The difference between an uncorrected value of the *deflection of the vertical* as determined by observation and the value after being corrected in accordance with certain assumptions made with reference to the physical condition of the *geoid.*

The term deflection anomaly is seldom used. As it applies to the *residuals* obtained from solutions of equations set up in conformity with some adopted theory of distribution of mass in the outer part of the earth's crust, the term *residual* is more generally used, qualified, if need be, by a designation of the particular solution through which it was obtained.

anomaly: free-air.—The difference between observed gravity and theoretical gravity which has been computed for latitude and corrected for elevation of the station above the geoid, by application of the normal rate of change of gravity for change of elevation, as in free air.

The elevation correction is for height above the *geoid*, not the *spheroid*, the relation of which to the *geoid* is practically always unknown. No correction is applied for the effect of topography and isostatic compensation.

anomaly: gravity.—The difference between the observed value of gravity at a point and the theoretical value, in the sense, observed minus theoretical.

The theoretical value may be obtained in various ways. See *anomaly: free-air; anomaly: Bouguer; anomaly: isostatic.* The point at which the two values of gravity are compared is commonly either the point of observation or a point on the geoid directly under (or above) the point of observation.

The value of the gravity anomaly is the same for either point of comparison. However, in framing a definition, it is desirable to adopt one convention or the other. In the United States, since one of the purposes of the isostatic studies of the U. S. Coast and Geodetic Survey has been to explain observed gravity, the point of observation has been used as the point of comparison. Elsewhere, the almost universal practice has been to take the point on the *geoid* directly below (or above) the point of observation, even for isostatic studies, since the original use of gravity observations was to determine the *flattening of the earth:* this fact has determined the convention adopted for defining gravity anomaly.

If we knew the relation between the geoid and the spheroid of reference (a formula for theoretical gravity implies a spheroid of reference), we could reduce gravity to the spheroid of reference: this is often done in theoretical investigations.

When it is desired to explain the observed value of gravity by some theory, it is natural to take the point of observation as the point of comparison. If, on the other hand, it is desired to use observations of gravity to determine the figure of the earth, the natural point of comparison is the point on the geoid directly below (or above) the point of observation. For studies of this sort, it is necessary to reduce all observations to a common level surface.

anomaly: isostatic.—The difference between an observed value of *gravity* and a theoretical value at the point of observation which has been corrected for elevation of the station above the geoid and for the effect of *topography* over the whole earth and for its *isostatic compensation.*

The present procedure is to compute the *isostatic compensation* according to the *Pratt theory of isostasy* as developed by Hayford, using various depths for purposes of comparative analysis. Isostatic anomalies obtained by the above procedure are sometimes called Hayford gravity anomalies or, simply, Hayford anomalies. Other types of isostatic anomaly may be distinguished by the author's basic assumptions and designated by the author's name, as Meinesz anomaly, Heiskanen anomaly, etc.

Antarctic Circle.—The *geographic parallel* having a south *latitude* equal to the *complement* of the *declination* of the *winter solstice.*

The *obliquity of the ecliptic* is steadily changing, so that the *winter solstice* is not a point of fixed *declination,* and the Antarctic Circle as defined above is not a line of fixed position. When the Antarctic Circle is to be shown on a *map,* it is desirable that it be treated as a line of fixed position, and a conventional value adopted for its *latitude.* For this purpose, the value 66°33′ south *latitude* is here proposed. This is the *complement* of 23°27′, proposed for the *latitude* of the *Tropic of Capricorn.*

The area enclosed by the Antarctic Circle is called the "South Frigid Zone," and at all points in this area, at the time of the *winter solstice,* the sun is above the *horizon* at local midnight. The belt of the earth's surface bounded by the Antarctic Circle and the *Tropic of Capricorn* is called the "South Temperate Zone." See *Arctic Circle.*

aphelion.—The point in an elliptical orbit of a planet or comet remotest from the sun.

aplanatic lens.—A lens which transmits light without *spherical aberration.*

apparent (visible) horizon.—See under *horizon.*

apparent solar day.—See under *day.*

apparent solar time.—See under *time.*

apparent time.—See under *time.*

approximate.—Approaching closely but not attaining a specified *accuracy* or a *precision* conformable with the quality of the observations.

Too few observations or observations made with an inferior instrument will give an approximate result.

Omitted corrections or the neglect of terms having significant values will produce an approximate result, even where the observations themselves are satisfactory.

apsides: line of.—See under *line of apsides.*

arc (general).—A portion of a mathematically defined curve.

A circular arc is part of a circle; an elliptical arc, part of an ellipse; etc.

arc correction (pendulum).—The quantity which is applied to the period of *vibration* of a *pendulum* to allow for the *pendulum's* departure from simple harmonic motion.

With a *pendulum,* the *acceleration* is proportional to the sine of the *amplitude,* while the displacement is proportional to the *amplitude* itself. The *arc* is the quantity obtained by observation, and it is slowly changing; the *acceleration* is the quantity sought. By limiting the size of the *amplitude,* the correction required to take account of this non-isochronous condition is kept quite small.

arc triangulation.—See under *triangulation.*

Arctic Circle.—The *geographic parallel* having a north *latitude* equal to the *complement* of the *declination* of the *summer solstice.*

The *obliquity of the ecliptic* is steadily changing so that the *summer solstice* is not a point of fixed *declination,* and the Arctic Circle as defined above is not a line of fixed position. When the Arctic Circle is to be shown on a *map,* it is desirable that it be treated as a line of fixed position, and a conventional value adopted for its *latitude.* For this purpose, the value 66°33′ north latitude is here proposed. This is the *complement* of 23°27′, proposed for the *latitude* of the *Tropic of Cancer.*

The area enclosed by the Arctic Circle is called the "North Frigid Zone," and at all points in this area, at the time of the *summer solstice,* the sun is above the *horizon* at local midnight. The belt of the earth's surface bounded by the Arctic Circle and the *Tropic of Cancer* is called the "North Temperate Zone."

Actually, when an observer sees the sun at a *solstice* just touch the *horizon* at local midnight, without setting, he is probably much nearer to the 66th parallel of latitude than to the one proposed above for cartographic use. This is an effect of *atmospheric refraction.* It is modified by local conditions, including *dip of horizon.*

area triangulation.—See under *triangulation.*

arithmetical mean.—The value obtained by dividing the sum of a series of values by the number of values in the series. See *weighted mean.*

artificial horizon.—See under *horizon.*

A-stations (traverse).—*Subsidiary stations* established between *principal stations* of a *survey traverse,* for convenience of measuring, to obtain the distance between the *principal stations.*

A-stations are established along a route which is not straight, as along a curved section of a railroad, where the length measures must be carried through a series of short *straight lines,* while *azimuth* control may be carried through widely spaced *stations.* The A-stations form a traverse loop with the line connecting the *principal stations,* the distance between which is obtained by projecting the measured lengths of the short lines onto that main line. See *equation: perpendicular* (*traverse*). A-stations are so-called because in a given series, these stations are designated by the name of a *principal station* followed by the letters A, B, C, etc., in order of distance from the *principal station.* They are sometimes referred to as "A, B, C stations."

astrolabe.—An instrument for measuring *altitudes* of celestial objects. See under *astrolabe: pendulum astrolabe; planispheric astrolabe; prismatic astrolabe.*

Derived from Greek words meaning "to take a star," the term astrolabe has been used to designate a great variety of instruments, of which the three named above are of especial interest to surveyors and map makers.

astrolabe: pendulum.—An astronomical instrument using a constant altitude for position determination. Its distinctive feature is a mirror suspended on top of a pendulum to form the *artificial horizon.*

The instrument consists of a V-shaped casting carrying the objective and eyepiece lenses at the ends of the arms. The mirror, which rests on top of the pendulum and forms the level surface (artificial horizon), is located at the intersection of the V. The pendulum is suspended so that it is free to swing in either of two planes at right angles to each other, such as north-south and east-west planes. The pendulum is highly damped so that the mirror comes to rest quickly and remains steady under normal observing conditions.

astrolabe: planispheric.—An *astrolabe* of ancient origin, consisting of a full graduated circle with a centrally mounted *alidade,* and accessory adjustable plates on which are engraved *stereographic projections* of the heavens and of the sphere for local *latitudes.*

The instrument is held suspended in a *vertical plane,* the *altitude* of a star is observed with the aid of the *alidade,* and the projection-bearing plates are so adjusted that, with the data available, what are essentially graphical solutions of astronomical problems are effected.

astrolabe: prismatic.—An *astrolabe* consisting of a *telescope* in a horizontal position, with a prism and *artificial horizon* attached at its objective end, used for determining *astronomic positions.*

The usual form of this instrument employs a 60° prism, with the face nearest the objective perpendicular to the *line of collimation* of the *telescope;* a small mercurial horizon is attached to the instrument beneath the prism. In observing, two images of a star are seen: one reflected directly into the *telescope* by the lower face of the prism; the other reflected first by the mercurial horizon, then by the upper face of the prism into the *telescope.* These two images of the same star are moving in opposite directions, and at the instant of the coincidence, the star is at the apparent *altitude* of the prism angle. Prismatic astrolabes may be made with the angle between the two inclined at other than 60° An English instrument employs a 45° *angle,* and has accessory means of obtaining more than one observation for a given setting.

astronomic; astronomical.—Of or pertaining to astronomy, the science which treats of the heavenly bodies, and the art based on that science.

For particular usage of the two forms, astronomic, astronomical, see *-ic*; *-ical.*

astronomic azimuth.—See under *azimuth.*

astronomic azimuth mark.—See under *azimuth mark.*

astronomic coordinates.—See under *coordinates.*

astronomic equator.—See under *equator.*

astronomic latitude.—See under *latitude.*

astronomic longitude.—See under *longitude.*

astronomic meridian.—See under *meridian.*

astronomic parallel.—See under *parallel.*

astronomic position.—See under *latitude: astronomic; longitude: astronomic.*

astronomical day.—See under *day.*

astronomical leveling.—See under *leveling.*

astronomical refraction.—See under *refraction.*

astronomical theodolite.—See under *alt-azimuth instrument.*

astronomical time.—See under *time.*

astronomical transit.—See under *transit.*

astronomical traverse.—See under *traverse.*

astronomical triangle.—The triangle on the *celestial sphere* formed by *arcs* of *great circles* connecting the celestial pole, the *zenith,* and a celestial body.

The *angles* of the astronomical triangle are: at the pole, the *hour angle;* at the celestial body, the parallactic angle; at the *zenith,* the *azimuth angle.* The sides are: pole to *zenith,* the *co-latitude; zenith* to celestial body, the *zenith distance;* and celestial body to pole, the *polar distance.*

astronomical unit.—The mean distance of the earth from the sun; that is, the semi-major axis of the elliptical orbit described by the earth around the sun, equal to about 149,500,000 kilometers (Smithsonian Physical Tables, 1934).

While the above simple definition is generally used in technical and popular publications, mathematical astronomers find it convenient to adopt a highly technical definition involving the use of a constant (k^2) introduced by Gauss to define the *acceleration* of the sun's gravitational field, and in particular the exact value of k^2 used by Gauss, although the astronomical data used by him to determine it differ slightly from the best modern values.

The *solar parallax* is a measure of the astronomical unit, and in some astronomical calculations it is convenient to use the *solar parallax* in place of the astronomical unit.

asymmetry of object (target).—Lack of symmetry in the visible aspect of an object as seen from a particular point of observation.

A square or rectangular pole may so face the observer that the line bisecting its tangents does not pass through its geometrical center. With a square cupola or tower, the *error* resulting from observing tangents and taking a mean may be quite large. The *error* due to asymmetry of object observed on is of the same character and requires the same treatment as *error* due to observing on an *eccentric object.* Asymmetry and *phase* are closely associated, but are not identical.

atmospheric refraction.—See under *refraction.*

authalic (equal-area) latitude.—See under *latitude.*

autogonal map projection.—See under *map projection.*

autumnal equinox.—See under *equinox.*

auxiliary guide meridian (United States public-land surveys).—See under *meridian.*

auxiliary standard parallel (United States public-land surveys).—See under *parallel.*

average error.—See under *error.*

aviation chart.—See under *chart: aeronautical chart.*

aviation map.—See under *chart: aeronautical chart.*

axis of lens (lens axis).—See under *principal axis.*

axis of level; axis of level bubble.—See under *level axis: spirit level axis.*

azimuth (general).——The *horizontal direction* reckoned clockwise from the meridian plane. See specific definition for *azimuth: astronomic; azimuth: geodetic; azimuth: grid;* etc.

In the basic control surveys of this country, azimuths are measured clockwise from south following the continental European geodetic practice. However, this practice is not followed in all countries.

azimuth: astronomic.—At the point of observation, the *angle* measured from the *vertical plane* through the celestial pole to the vertical plane through the observed object.

Astronomic azimuth is the *azimuth* which results directly from observations on a celestial body: it is measured in the plane of the *horizon,* and is usually reckoned from south (0°), through west (90°), north (180°), and east (270°), back to south (360° or 0°). It is affected by the local *deflection of the vertical* (*station error*) which, in the United States, produces differences between astronomic and *geodetic azimuths* of as much as 26″ in the mountain regions of the west, and of 10″ in the less rugged eastern States.

Astronomic azimuths may also be reckoned clockwise from north through 360°. In navigation, they are sometimes reckoned clockwise or counter-clockwise through 180°, from the south in the southern hemisphere, and from the north in the northern hemisphere. In recording an astronomic azimuth it is essential that both the initial and the direction of reckoning be indicated.

azimuth: back.—See under *azimuth: geodetic.*

azimuth: direction method, determination of astronomic azimuth.—The determination of the *astronomic azimuth* of a line by measuring with a *direction theodolite* the *horizontal angle* between a selected star and a suitable mark, and applying that *angle* to the *azimuth* of the star computed for the epoch of the observation.

In the horizontal *control surveys* of continental United States, the direction method of azimuth determination is preferred over other methods. A circumpolar star is observed at any *hour angle,* the mark being a signal light on a *main-scheme station* or at a station, called an *azimuth mark,* established for the purpose. A *correction for inclination of the horizontal axis,* depending upon the *altitudes* of the star and of the mark, is applied to the observed *angle,* and *curvature correction* and *corrections* for *variation of the pole* and for *diurnal aberration* also enter into the computation.

azimuth: forward.—See under *azimuth: geodetic.*

azimuth: geodetic.—For the *geodesic line* from A to B, the angle between the tangent to the *meridian* at A and the tangent to the geodesic line at A. In the U. S. Coast and Geodetic Survey, this angle is measured clockwise from south, but see under *azimuth (general).*

This azimuth is called the forward azimuth for the line AB. The angle between the tangents to the meridian and to the geodetic line at B is called the back azimuth for the line AB. Because of the *convergence of the meridians,* the forward and backward azimuths of a line do not differ by exactly 180°, except where the two end points have the same *geodetic longitude* or where the *geodetic latitudes* of both points are 0°.

The geodesic line is not, except in special cases, a plane curve, but for short lines (of the length commonly used in triangulation) one may neglect the small difference between the geodesic line and the plane curve actually observed.

azimuth: grid.—The *angle* in the plane of projection between a straight line and the central meridian (axis of *Y*) of a *plane-rectangular coordinate system.*

In the *State plane-coordinate systems* established by the U. S. Coast and Geodetic Survey, grid azimuths are reckoned from south (0°) clockwise through 360°. While essentially a *map* quantity, a grid azimuth may, by mathematical process, be transformed into a *survey* or ground quantity. See *gisement.*

azimuth: Laplace.—A *geodetic azimuth* derived from an *astronomic azimuth* by means of the *Laplace equation.*

It is proposed that the symbol a_L be adopted to designate a Laplace azimuth, to distinguish it from a *geodetic azimuth* derived by computation through the triangulation, which is represented by the symbol a_G.

See *Laplace equation.*

azimuth: magnetic.—At the point of observation, the *angle* between the *vertical plane* through the observed object and the *vertical plane* in which a freely suspended symmetrically magnetized needle, influenced by no transient artificial magnetic disturbance, will come to rest.

Magnetic azimuth is generally reckoned from magnetic north (0°) clockwise through 360°. Such an *azimuth* should be marked as being magnetic, and the date of its applicability be given.

azimuth: method of repetitions, determinations of astronomic azimuth.—The determination of the *astronomic azimuth* of a line by accumulating on the horizontal circle of a *repeating theodolite* the sum of a series of measures of the *horizontal angle* between a selected star and a suitable mark, and applying the mean of such measures to the *azimuth* of the star computed for the mean epoch of the observations.

The method of repetitions for the determination of *azimuth* is theoretically susceptible of great *accuracy* and *precision,* but in practice is not as satisfactory as the *direction method.*

A *correction for inclination of the horizontal axis,* depending on the *altitudes* of the star and of the mark, is applied to the observed *angle,* and *curvature correction* and *corrections* for *variation of the pole* and for *diurnal aberration* also enter into the computations.

azimuth: micrometer method, determination of astronomic azimuth.—The determination of the *astronomic azimuth* of a line by measuring indirectly with an *ocu-*

lar micrometer attached to a *theodolite* or *transit* the *horizontal angle* between a selected star at its elongation and a suitable ground mark (light) placed close to the *vertical plane* which passes through the star, and applying that *angle* to the *azimuth* of the star computed for the epoch of the observation.

At elongation, the apparent motion in *azimuth* of a close circumpolar star like Polaris is very small for an appreciable period of time, and a series of observations can be made by the micrometer method without reorienting the instrument. A *correction for inclination of the horizontal axis,* depending on the *altitudes* of the star and of the mark, is applied to the *observed angle,* and *curvature correction* and corrections for *variation of the pole* and for *diurnal aberration* also enter into the computations. Recently, difficulties in determining *astronomic azimuth* in Alaska, due to the large *altitude* of circumpolar stars, were overcome by employing the micrometer method with a *Bamberg broken-telescope transit.*

azimuth: normal section.—For the *normal section line* from A to B, the *angle* at A between the meridian plane and the normal section which passes through B.

azimuth angle (astronomy).—See under *angle.*

azimuth angle (surveying).—See under *angle.*

azimuth equation.—See under *equation.*

azimuth line.—See under *azimuth plane.*

azimuth mark: astronomic.—A signal or target whose *astronomic azimuth* from a survey station is determined from direct observations on a celestial body.

The mark may be a lamp or illuminated target placed especially for the purpose; or it may be a signal lamp at another survey station.

azimuth mark: geodetic.—A marked point established in connection with a *triangulation* (or *traverse*) *station* to provide a starting *azimuth* for dependent surveys.

The *geodetic azimuth* from the station to the azimuth mark is determined instrumentally; its distance is often determined only approximately, but should be sufficiently great to render negligible the angular effect of the ordinary *errors* of centering instrument and target. Since 1928, azimuth marks consisting of bronze tablets set in concrete or stone have been established in connection with the basic horizontal *control survey* of the United States. These marks are usually so located as to be readily available without special construction to elevate either instrument or target. At a station having an established azimuth mark, both the *geodetic azimuth* and the *grid azimuth* of the mark on the *State coordinate system* are computed and published.

azimuth mark: Laplace.—An *astronomic azimuth mark* at a *Laplace station.*

azimuth plane.—Any *vertical plane.*

Such a plane cuts an azimuth section from the *spheroid,* and traces an azimuth line on its surface. These terms are rarely seen and their use should be avoided. Although the *vertical planes* containing the prime vertical and the *meridian* are sometimes regarded as not being azimuth planes, they are special cases of that term as here defined.

azimuth section.—See under *azimuth plane.*

azimuthal (zenithal) map projection.—See under *map projection.*

azimuthal equidistant map projection.—See under *map projection.*

B

Bache-Wurdeman base-line measuring apparatus.—See under *base apparatus.*

back azimuth.—See under *azimuth: geodetic.*

backsight (general).—A sight on a previously established survey point.

backsight (leveling).—A reading on a rod held on a point whose *elevation* has been previously determined, and which is not the closing sight of a level circuit.

In leveling, a *backsight* is sometimes called a plus sight, because it is usually added to the *elevation* of the rod point to obtain the height of the instrument (H. I.). This is not always the case: the pointing may be on a mark on a wall or in the roof of a mine tunnel, with the instrument at a lower *elevation.* In such a case, the *backsight* will be subtracted from the known *elevation* to obtain the height of instrument. The term *backsight* is preferred over plus sight.

backsight (transit traverse).—A sight on a previously established survey point, which is not the closing sight of the *traverse.*

balancing a survey.—Distributing *corrections* through a *traverse,* to eliminate the *errors of closure.*

Errors of closure may arise when a *traverse* closes upon itself or when it extends between points of known *position. Corrections* to balance the *closing errors* are distributed through the *traverse* according to some approved method, so that the resulting survey figure will be geometrically consistent: the *position* and *azimuth* at the closing point as derived by computation carried through the balanced *traverse* will agree with the starting data for a *closed traverse* or with the fixed data where the *traverse* extends between fixed points. For an extended *traverse,* the elimination of errors may be made by the *method of least squares,* and the operation is termed an *adjustment.* For local work, the elimination of *errors* may be made by one of two standard methods, and is termed balancing the survey. These methods are termed *compass rule* and *transit rule.*

balancing a survey, compass rule.—*Corrections* corresponding to the *closing errors* in *latitude* and *departure* are distributed according to the proportion: length of line to total length of *traverse.*

The compass rule is used when it is assumed that the *closing errors* are as much due to *errors* in *observed angles* as to *errors* in measured distances.

balancing a survey, transit rule.—*Corrections* corresponding to the *closing errors* in *latitude* and *departure* are distributed according to the proportion: *latitude* and *departure* of each line of the *traverse* to the arithmetical sums of the *latitudes* and *departures* of the entire *traverse.*

The transit rule is used when it is assumed that the *closing errors* are due less to the *errors* in the *observed angles* than to *errors* in the measured distances.

Bamberg broken-telescope transit.—See under *transit: broken-telescope.*

bank of stream.—The continuous margin along a river where all vegetation ceases.

The right bank of a stream is the bank on the right-hand side, and the left bank, the one on the left-hand side, as one proceeds in the direction in which the current flows (downstream).

Barlow leveling rod.—See under *leveling rod.*

barometer.—An instrument for measuring atmospheric pressure.

There are two general types of barometers: those in which atmospheric pressure is balanced by the weight of a column of liquid (usually mercury); and those in which atmospheric pressure is balanced by some elastic device. The latter are termed *aneroid* (without liquid) *barometers.* A third kind of device for measuring atmospheric pressure is the *hypsometer,* which is not classed as a barometer.

barometer: aneroid.—A *barometer* which balances the atmospheric pressure against a mechanically elastic device.

The usual form of aneroid barometer consists of a thin box of corrugated metal, almost exhausted of air. When the atmospheric pressure increases, the box contracts; when the pressure lessens, the box expands. By mechanical means these movements are amplified and communicated to an index hand which registers on a graduated dial. The *graduations* on the dial are usually such as will show the atmospheric pressure in terms of height of a corresponding column of mercury. The aneroid barometer may be equipped with an auxiliary dial, graduated to show *elevations* in feet corresponding to atmospheric pressure.

barometer: cistern.—A *mercury barometer* in which a column of mercury is enclosed in a vertical glass tube, the upper end of which is sealed and exhausted of air, and the lower end placed in a cistern or reservoir of mercury which is exposed to atmospheric pressure.

The amount of the atmospheric pressure on the free surface of the mercury in the cistern determines the height to which the mercury will rise in the vertical tube. This may be measured, and the pressure reported in terms of that height, as so many *inches* of mercury.

barometer: mercury.—A *barometer* in which atmospheric pressure is balanced against the weight of a column of mercury.

There are two types of mercury barometers, differing in the form of apparatus used, but not in the underlying principle. See *barometer: cistern; barometer: siphon.*

barometer: mountain.—A *barometer* constructed for use in leveling operations.

Mountain barometer is a general term, little used, for a *barometer* of any type which is constructed for safe transportation under difficult field conditions, so that it may be used in survey work. See *barometer: aneroid, altimeter.*

barometer: siphon.—A *mercury barometer* consisting of a column of mercury in a glass tube which is so bent as to have two vertical branches, one about one-fourth the length of the other. The end of the longer branch is closed and the air in it is displaced by the mercury, but the shorter branch is left open and the mercury thereby subjected to atmospheric pressure. The difference of the height of the mercury in the two branches is a measure of the atmospheric pressure.

barometric hypsometry.—See under *hypsometry.*

barometric leveling.—See under *leveling.*

base apparatus (geodetic surveying).—Any apparatus designed for use in measuring with *accuracy* and *precision* the length of a *base line* in *triangulation,* or the length of a line in *first-* or *second-order traverse.*

Various kinds of apparatus have been designed for use in measuring *base lines,* including the following general types:

(1) Apparatus having a length unit composed of simple bars or tubes of wood, metal, or other material, suitably constructed for making *alignment,* contact, and temperature determination. The earliest bars used in *geodetic* work were of wood. The most effective apparatus of this type is the *iced-bar.*

(2) Apparatus consisting of bars or tubes of different metals having unequal *coefficients of thermal expansion,* so arranged that (a) the temperature effects on the lengths of the component parts are largely neutralized and the effective length of the apparatus remains almost constant, or (b) the component parts form a metallic thermometer, by means of which the temperature of the apparatus is obtained and its effect on the length of the apparatus determined.

(3) Apparatus consisting of tapes or wires of metal, usually steel or a nickel-steel alloy. Because of its large *coefficient of thermal expansion,* steel tapes can be used in *precision* measures of *base lines* only under favorable temperature conditions. In this country, they have been replaced almost entirely by tapes of nickel-steel (*invar* or nilvar), having very small *coefficients of thermal expansion,* and capable of giving good results through a wide range of temperature. While tapes or bands are used in this country, wires have been used in base-line operations in Mexico and in Europe.

base apparatus: Bache-Wurdeman base-line measuring apparatus.—A *compensating base-line measuring apparatus* having a measuring element composed of a bar of iron and a bar of brass, each a little less than 6 meters in length, held together firmly at one end, with the free ends so connected by a compensating lever as to form a *compensating apparatus.*

Described in Appendix No. 25, U. S. Coast and Geodetic Survey Report for 1854; description reprinted in Appendix No. 12, U. S. Coast and Geodetic Survey Report for 1873.

base apparatus: compensating base-line measuring apparatus.—A *base apparatus* having a length element composed of two metals having different *coefficients of thermal expansion,* so arranged and connected that the differential expansion of its components will maintain a constant length of the element, under all temperature conditions of use.

base apparatus: contact base-line measuring apparatus.—*A base apparatus* composed of bars whose lengths are defined by the distance between their end faces or points.

In use, the bars are laid end to end, one bar being kept in position while another bar is being moved ahead. There are various mechanical refinements connected with the moving of the bars into position, bringing the end of the new bar into contact with the end of the bar which is already in position without disturbing it.

base apparatus: contact-slide base-line measuring apparatus.—A modified *contact base-line measuring apparatus* consisting of two steel measuring bars (rods), each 4 meters in length, so mounted that contact is effected by coincidence of lines on a rod and a contact-slide. Each rod forms a metallic thermometer with two zinc tubes, one on each side of the bar: opposite ends of the bar are fastened to the ends of the tubes, the other ends of which are free to move with changes of temperature. Described in Appendix 17, U. S. Coast and Geodetic Survey Report for 1880.

base apparatus: duplex base-line measuring apparatus.—A *contact base-line measuring apparatus,* composed of two disconnected bars, one of brass and the other of steel, each 5 meters in length and so arranged as to indicate the accumulated difference of length of the measures from the brass and steel components.

The duplex base apparatus was designed by Wm. Eimbeck of the U. S. Coast and Geodetic Survey. It is described in Appendix No. 11, U. S. Coast and Geodetic Survey Report for 1897.

base apparatus: Hassler base-line measuring apparatus.—An *optical base-line measuring apparatus* consisting of four rectangular iron bars mounted end to end in a wooden box. Each bar is 2 meters long, the combined length of the apparatus being 8 meters.

Described in Transactions of the American Philosophical Society of Philadelphia, Pennsylvania, new series, 1825, Vol. II, pages 1273–1286. See also Appendix No. 21, U. S. Coast and Geodetic Survey Report for 1865.

base apparatus: iced-bar apparatus.—An apparatus for measuring linear distances with great *precision* and *accuracy,* and consisting essentially of a steel bar which is maintained at a constant temperature by being surrounded with melting ice.

The iced-bar apparatus is also called the Woodward base apparatus, after its designer, Dr. R. S. Woodward (1849–1924), of the U. S. Lake Survey. The bar is rectangular in cross-section, and is carried in a Y-shaped trough which is filled with melting ice and mounted on a car which moves on a track. Bar lengths are observed with *micrometer microscopes,* mounted on stable supports. The iced-bar apparatus is used in the laboratory in standardizing tapes, and it has also had some use in establishing *field comparators.*

base apparatus: optical base-line measuring apparatus.—A *base apparatus* composed of bars whose lengths are defined by distances between lines at or near their ends, which are observed by suitably mounted and adjusted microscopes.

In using any optical base-line measuring apparatus, the positions of the bars are controlled by microscopes on stable supports, whose *reticle* lines may be brought into coincidence with the *fiducial marks* on the bars, either by adjusting a bar or a microscope.

base apparatus: Repsold base-line measuring apparatus.—An *optical base-line measuring apparatus,* composed of a steel bar approximately 4 meters long, whose exact length at any temperature is known, and whose temperature is determined by means of a metallic thermometer composed of the steel measuring bar and a similar bar of zinc, the two being fastened together at their middle points. Described in Professional Paper No. 24, Engineering Corps, United States Army.

base apparatus: Schott base-line measuring apparatus.—A *contact, compensating base-line measuring apparatus* composed of three parallel bars, the middle bar of zinc, the outer bars of steel. One end of each steel bar is free; the other end is fastened to an end of the zinc bar, a different end for each steel bar. The lengths of the bars are so proportioned with respect to their *coefficients of thermal expansion* that a constant distance is maintained between the free ends of the steel bars. Described in Appendix No. 7, U. S. Coast and Geodetic Survey Report for 1882.

base apparatus: Woodward base-line measuring apparatus.—See under *base apparatus: iced-bar apparatus.*

base line (general).—A surveyed line established with more than usual care, to which surveys are referred for *coordination* and *correlation.*

Base lines are established for specific purposes, the more important ones, defined elsewhere, being as follows: *base line (triangulation)*; base line (echo sounding); *base line (United States public-land surveys)*; *base line (construction).*

base line (United States public-land surveys).*—A line extending east and west along the *astronomic parallel* passing through the *initial point,* along which *standard township, section,* and *quarter-section corners* are established.

As may be inferred from its designation, the *base line* is the line from which is initiated the survey of the meridional *township* boundaries and *section* lines. Auxiliary governing lines, known as *standard parallels* or *correction lines,* are established along the *astronomic parallel* usually at intervals of 24 miles north or south of the *base line.* In some of the very early surveys the base line is referred to as the "basis parallel."

base line: construction.—The center line of location of a railway or highway; often termed the "base line of location." A reference line for the construction of a bridge or other structure.

base line: triangulation.—The side of one of a series of connected triangles, the length of which is measured with prescribed *accuracy* and *precision,* and from which the lengths of the other triangle sides are obtained by computation.

Base lines in *triangulation* are classified according to the character of the work they are intended to control, and the instruments and methods used in their measurement are such that prescribed *probable errors* for each class are not exceeded. These *probable errors,* expressed in terms of the lengths, are as follows: first-order base line, 1/1,000,000; second-order base line, 1/500,000; third-order base line, 1/250,000.

base map.—See under *map.*

base tape.—A tape or band of metal or alloy, so designed and graduated and of such excellent workmanship that it is suitable for measuring the lengths of lines (*base lines*) for controlling *triangulation,* and for measuring the lengths of *first-* and *second-order traverse* lines.

In use, a base tape is subject to certain physical conditions which influence its effective length, and which must be taken into account in computing the length of a measured *base line.* These conditions include tension, temperature, method of support, sag, *grade, alignment,* and *standardization* length. See *tape corrections.* At one time, base tapes were made of steel, a metal having a high *coefficient of thermal expansion.* They are now made of *invar,* nilvar, or some other alloy having a very small *coefficient of thermal expansion.* See *invar.*

basic control.—See under *control.*

bathygraphic.*—Descriptive of ocean depths.

bathymeter.*—An instrument used for determining ocean depths.

Any deep-sea sounding apparatus. There are a number of designs of such instruments.

bathymetric.*—Relating to the measurement of ocean depths.

bathymetric chart.—See under *chart.*

bathymetry.*—The art or science of determining ocean depths.

bearing: grid.—The *angle* in the plane of the projection between a line and a north-south grid line.

bearing: true.—The *horizontal angle* between the *meridian line* and a line on the earth.

The term true bearing is used in many of the early descriptions of land boundaries in this country. It is associated with true north, referring to the direction of the north point as determined by astronomical observations. For future use, if an astronomically determined bearing is used, the term astronomic bearing is preferred over true bearing.

bearing of line (plane surveying).—The *horizontal angle* which a line makes with the *meridian* of reference adjacent to the quadrant in which the line lies. Bearings are classified according to the *meridian* used as *true bearings,* magnetic bearings, *grid bearings.*

A bearing is identified by naming the end of the *meridian* (north or south) from which it is reckoned and the direction (east or west) of that reckoning. Thus, a line in the northeast quadrant making an angle of 50° with the *meridian* will have a bearing of N. 50° E. In most survey work, it is preferable to use *azimuths* rather than *bearings.*

bearing tree (United States public-land surveys).*—A tree forming a *corner accessory,* its distance and direction from the *corner* being known.

Bearing trees are identified by prescribed marks cut into their trunks; the species and sizes of the trees are also recorded.

bed (mining).—A stratum in the earth's crust which has been formed in an approximately horizontal layer. If of rock, it is termed *bedrock.*

bed of stream.—The area within the *high-water lines* of a stream or river.

It is the area which is kept practically bare of vegetation by the wash of the waters of the stream from year to year.

bedrock.—A stratum of rock in the earth's crust which has been formed in an approximately horizontal position.

Large, deeply embedded *boulders,* with upper surfaces planed and leveled through exposure, are sometimes erroneously reported as bedrock in descriptive writing.

bench mark.—A relatively permanent material object, natural or artificial, bearing a marked point whose *elevation* above or below an adopted *datum* is known.

Usually designated as a B. M., such a mark is sometimes further qualified as a P. B. M. (*permanent bench mark*) to distinguish it from a T. B. M. (temporary bench mark), or *supplementary bench mark,* a mark of less permanent character intended to serve for only a comparatively short period of time.

bench mark: first-order.—A bench mark connected to the *datum* (usually mean sea level) by continuous *first-order leveling.*

bench mark: junction.—A *bench mark* selected as the common meeting point for *lines of levels* or *links of levels.*

bench mark: permanent.—A *bench mark* of as nearly permanent character as it is practicable to establish.

Usually designated simply as a *bench mark* or B. M., a permanent bench mark is intended to maintain its *elevation* with reference to an adopted *datum* without change over a long period of time. Concrete or stone posts, with suitable marks in their tops, such as inscribed disks, are used. Such a mark is set in firm ground, deep enough to be free from frost action, and in a location where disturbing influences are believed to be negligible.

bench mark: second-order.—A bench mark connected to the *datum* (usually mean sea level) by continuous *second-order leveling* or by a continuous combination of *first-order leveling* and *second-order leveling.*

bench mark: supplementary.—A *bench mark* at a junction of sections of a line of levels, at which no *permanent bench mark* is established.

A supplementary bench mark may be classed as a *temporary bench mark,* established to hold the end of a completed section of a line of levels and serve as an initial from which the next section is run. Spikes and screws in poles, bolts on bridges, and chisel marks on masonry are used for the purpose and may last for years.

bench mark: tidal.—A *bench mark* set to reference a tide staff at a tidal station and the elevation of which is determined with relation to the local *tidal datum.*

Bilby steel tower.—A *triangulation tower* consisting of two steel tripods, one within the other, designed by J. S. Bilby of the U. S. Coast and Geodetic Survey.

The Bilby steel tower was put into use in 1927 and is standard equipment in triangulation executed by the U. S. Coast and Geodetic Survey. The Bilby steel tower is demountable; it is easily erected, and as easily taken down and moved to a new location. Some of these towers have each been used at more than one hundred different stations.

blaze (United States public-land surveys).*—A mark made upon a tree trunk at about breast height, in which a flat scar is left upon the tree surface.

blunder.—A mistake.

A blunder is not an *error,* though a small blunder may remain undetected in a series of observations and have the effect of an *error* in determining a result. Examples of blunders are: reading a horizontal circle wrong by an even degree; neglecting to record a tape length in a measured *traverse*; reversing the numerals in recording an observation; etc.

Bonne map projection.—See under *map projection.*

Borda scale.—A *metallic thermometer* composed of two metals having different *coefficients of thermal expansion.*

The Borda scale was devised by the French scientist, Jean Charles Borda (1733–1799), who placed a strip of copper on a strip of platinum, fastened the two together at one end, and by measuring the relative movement of their free ends, determined their temperature. He used this means for determining the temperature of bars used in measuring the length of a geodetic *base line.* Later, Bessel used a combination of zinc and iron for the same purpose.

Boston leveling rod.—See under *leveling rod.*

Bouguer anomaly.—See under *anomaly.*

boulder.—A large rounded rock which has been separated from the mass of which it was originally a part.

Deeply embedded boulders may show an exposed surface which through weathering has the appearance of *bedrock.*

boundary: land.—A line of demarcation between adjoining parcels of land.

The parcels of land may be of the same or of different ownership, but distinguished at some time in the history of their descent by separate legal descriptions. A land boundary may be marked on the ground by material monuments placed primarily for the purpose; by fences, hedges, ditches, roads, and other service structures along the line; or defined by astronomically described points and lines; by *coordinates* on a survey system whose position on the ground is witnessed by material monuments which are established without reference to the boundary line; and by various other methods.

boundary line.—A line along which two areas meet.

In specific cases, the word "boundary" is sometimes omitted, as in "State line"; and sometimes the word "line" is omitted, as in "International boundary," "county boundary," etc. The term "boundary line" is usually applied to boundaries between political territories, as "State boundary line," between two States. A boundary line between privately owned parcels of land is termed a *property line* by preference, or if a line of the U. S. public-land surveys, is given the particular designation of that survey system, as *section line, township line,* etc.

boundary map.—A *map* constructed for the purpose of delineating a *boundary line* and adjacent territory.

boundary monument.—A material object placed on or near a *boundary line* to preserve and identify the location of the *boundary line* on the ground.

Where it is impracticable to establish a *monument* on or very close to a *boundary line,* the position of the *boundary line* on the ground is preserved by means of *reference marks.* The term *monument* is sometimes used to include both the mark on the *boundary line* and the *reference mark.* See also *monument* (*U. S. public-land surveys*).

boundary survey.—A survey made to establish or to reestablish a *boundary line* on the ground, or to obtain data for constructing a *map* or *plat* showing a *boundary line.*

The term boundary survey is usually restricted to surveys of *boundary lines* between political territories. For the survey of a *boundary line* between privately owned parcels of land, the term *land survey* is preferred, except that in surveys of the *public lands* of the United States, *cadastral survey* is used.

boundary vista.—A lane cleared along a *boundary line* passing through a wooded area.

A boundary vista is used for readily identifying a *boundary line,* and aiding in civil administration relating thereto.

Bowie effect (gravity).—The indirect effect on *gravity* due to the warping of the *geoid,* or the elevation of the *geoid* with respect to the *spheroid* of reference.

Also known as the Bruns term, the Bowie effect was discussed by Stokes, Bruns, Clarke, Helmert, and others, but it was under the direction of Bowie that a practical means was provided for computing its value for a given station. The *correction* for the Bowie effect is applied to the theoretical value of gravity, with the sign reversed. Values of the Bowie effect are tabulated in the *fundamental tables: deformation of the geoid and its effect on gravity.* The symbol $\Delta_2 g$ has been used to represent the Bowie effect.

Bowie Method (of adjustment).—A method devised in the U. S. Coast and Geodetic Survey under the direction of William Bowie for the *adjustment* of large networks of *triangulation.*

A description of the method is given in Special Publication No. 159, The Bowie Method of Triangulation Adjustment.

break-circuit chronometer.—See under *chronometer.*

broken grade (tape).—The change in *grade* when the middle point of a tape is not on *grade* with its ends.

If the middle support for the tape is not on the same *grade* as the end supports, the fact is noted with a reference "broken grade at —," naming the particular tape length which contains the broken grade.

broken-telescope transit.—See under *transit.*

bronze pendulum.—See under *pendulum.*

Brown gravity apparatus.—An apparatus for measuring the *acceleration of gravity* which utilizes the *Mendenhall pendulum,* but has a clamping device for holding the *pendulum* in the *receiver* when being transported from *station* to *station,* and which utilizes an electrical pick-up and amplifying device for recording the *oscillations* (*pendulum*) on the *chronograph* sheet.

The clamping device makes it unnecessary to remove the *pendulum* from the *receiver,* and thus avoids the necessity of breaking the vacuum seal each time a move is made. The recording of the *pendulum oscillations* on the *chronograph sheet* permits a direct time comparison of those *oscillations* with the radio time signals. One improvement of the Brown gravity apparatus over the *Mendenhall gravity apparatus* relates to the supporting of the *receiver* by leveling screws placed near the top of the *receiver,* making the *flexure* (*pendulum support*) less than in the earlier type of apparatus which had the leveling screws at the bottom of the *receiver,* and thus much further from the point of support of the *pendulum.*

Another improvement was the use of a cast aluminum chamber in which the *receiver* was hung as indicated above, and which was set directly in a mass of plaster of Paris placed at the bottom of a hole dug in the ground, and constituted a satisfactory stable support which was made ready more quickly than a concrete pier.

A third improvement over the old order consisted of the permanent installation in trucks of a greater portion of the electric wiring and auxiliary equipment. These improvements considerably reduced the time required by the earlier methods and apparatus for the determination of the *acceleration of gravity.* It was named for its designer, E. J. Brown of the U. S. Coast and Geodetic Survey, and was first used in the field in 1932.

bubble sextant.—See under *sextant.*

Bullard method of isostatic reduction.—See under *Hayford-Bullard* (or *Bullard*) *method of isostatic reduction.*

bulls-eye level.—See under *level: circular.*

bus and truck map.—See under *map.*

C

cadastral control (United States public-land surveys).—Lines established and marked on the ground by suitable *monuments,* which are used as starting and closing points in surveys of the *public domain of the United States.*

The fundamental control of the public-land surveys of the United States consists of *base lines, standard parallels* (*correction lines*), *principal meridians,* and *guide meridians.*

cadastral map.—See under *map.*

cadastral survey.—See under *survey.*

Calendar: Gregorian.—The calendar established by Pope Gregory XIII in 1542 A. D., to correct the *Julian Calendar* by introducing the more *precise* value of the length of the *tropical year,* 365 days 5 hours 48 minutes 46.0 seconds *mean solar time.*

The Gregorian Calendar modified the requirement of the *Julian Calendar* that every fourth year be given an additional day by providing that, for centurial years, this would be done only where such years were divisible by 400. The calendar was brought into harmony with the seasons as they were at the time of the Council of Nice (A. D. 325) by dropping 10 days at the time of its becoming effective; October 4, 1582, being followed immediately by October 15. See *Calendar: Julian; double dating.*

Calendar: Julian.—The calendar devised by Julius Caesar in 45 B. C., and based upon the assumption that the true length of the *tropical year* was exactly 365¼ *mean solar days.*

The Julian Calendar provided that the ¼ day be absorbed by giving every fourth year one additional day; it made the *year* begin on January 1, a provision that was not universally followed even where other provisions of the calendar were followed; it was effective for the year 44 B. C., which was known as the "Year of Confusion" because, in order to bring the calendar into harmony with the seasons, two additional months were intercalated between November and December, giving it 14 months with a total of 445 days. See *Calendar: Gregorian; double dating.*

calendar month.—See under *month.*

calendar year.—See under *year.*

calibration.—The determination, in terms of an adopted *unit* and by mechanical interpolation based on values obtained by *standardization,* of the supplementary marks on a measuring instrument or device. Also the determination of the values of the *divisions* of a circle as proportional parts of a circumference.

When the length of a tape has been determined by standardization, the value of its intermediate marks may be determined by calibration, it being assumed that the tape has been divided into parts which are directly proportional to its length. See *standardization.*

Camp Colonna datum.—See under *datum.*

cardinal.*—The astronomical *directions* on the surface of the earth: north, south, east, west.

The term cardinal, without qualification, is sometimes used to indicate any or all of the above *directions,* the context giving exact meaning to its use.

Cassini map projection.—See under *map projection.*

catenary correction, tape.—See under *tape: sag* or *catenary correction.*

celestial equator.—See under *equator.*

celestial latitude.—See under *latitude.*

celestial longitude.—See under *longitude.*

celestial meridian.—See under *meridian.*

celestial refraction.—See under *refraction: astronomical refraction.*

celestial sphere.—A sphere of indefinitely large (infinite) radius, described around an assumed center, and upon which positions of celestial bodies are projected along radii passing through the bodies.

In surveying operations which include observations on a celestial body, the *direction* but not the distance of the celestial body from the point of observation is important. For observations on bodies within the limits of the solar system, the center of the celestial sphere is taken as coinciding with the center of the earth. For bodies outside those limits, the *annual parallax* is negligible in surveying operations, and the center of the celestial sphere may be taken as coinciding with the point of observation.

Celsius scale.—A temperature scale in which 100° marks the freezing point and 0° the boiling point of water at 760 mm. barometric pressure.

The Celsius scale is the predecessor of and is equal to the *centigrade scale* inverted. Some use has been made of Celsius scale as synonomous with *centigrade scale.*

center: reduction to.—The amount which must be applied to a *direction* observed at an *eccentric station* or to an *eccentric signal,* to reduce such *direction* to what it would be if there were no such *eccentricity.*

The computation of the reduction to center is quite similar for the two cases, the same computation form being adapted for the *correction* to a *direction* observed at an *eccentric station* as to a direction to an *eccentric signal.*

center line (United States public-land surveys).*—The line connecting opposite *quarter-section corners,* or opposite sixteenth-section corners.

center of instrument.—The point on the vertical axis of *rotation* of an instrument at the same *elevation* as the *axis of collimation* when that axis is in a horizontal position.

In a *transit* or *theodolite,* it is close to or at the intersection of the horizontal and vertical axes of the instrument.

center of oscillation (pendulum).—The position in a *compound pendulum* of the particle which corresponds to the heavy particle of an equivalent *simple pendulum.*

The *centers of suspension* and oscillation are interchangeable. If the center of oscillation is made the *center of suspension,* the former *center of suspension* becomes the new center of oscillation. This principle is the basis of design of *compound reversible pendulums.*

center of suspension (pendulum).—The fixed point about which a pendulum *oscillates.* See *center of oscillation.*

center of transit or level.—A manufacturer's term for the spindle or spindles which are in a vertical position when the instrument is in use, and about which the instrument, or part of the instrument, *rotates.*

The engineer's *transit* has two such centers; an inner center to which the *alidade* is attached, and an outer center to which the horizontal circle is attached; it is hollow and *rotates* on the inner center. The rotation of the *alidade* is spoken of as the upper motion, and the rotation of the horizontal circle as the lower motion of the instrument.

centigrade scale.—A temperature scale in which 0° marks the freezing point and 100° the boiling point of water at 760 mm. barometric pressure. See *Celsius scale.*

central meridian (State plane-coordinate system).—See under *meridian.*

centrifugal force.—The force with which a body moving under constraint along a curved path, reacts to the constraint.

Centrifugal force acts in a direction away from the center of curvature of the path of the moving body. As a force caused by the *rotation* of the earth on its axis, centrifugal force is opposed to *gravitation,* and combines with it to form *gravity.*

c. g. s. system.—The system of physical measurements in which the fundamental units of length, mass, and time are the centimeter, the gram, and the *mean solar second.*

chain.—The *unit of length* prescribed by law for the survey of the *public lands of the United States.* The chain is equivalent to 66 feet.

The chain derives its name from the *Gunter's chain,* which had the form of a series of *links* connected together by rings, a form which has been superseded by metal tapes or ribbons graduated in chains and links. The General Land Office uses steel tapes of 1, 2, 5, or 8 chains length.

chain: Gunter's.—A measuring device used in land surveying, composed of 100 metal *links* fastened together with rings, the length of the *chain* being 66 feet.

It was invented by an English astronomer, Edmund Gunter, about 1620, and is the basis for the *chain* and *link, units of length* used in the survey of the *public lands of the United States.* An advantage in measuring in *chains* is that 10 square *chains* equals 1 acre.

chaining.—The operation of measuring a distance on the earth, using a chain or tape. See *taping.*

While the *chain* as an instrument used in making land and other surveys has been superseded by the metal tape, the term chaining has continued in use where reference is to surveys of the *public lands of the United States.* For the corresponding operation in other surveys, the term *taping* is preferred. In chaining, the men who mark the tape ends are termed *chainmen.*

chainmen.—The men who mark the tape ends in *chaining.*

chambered spirit level.—See under *level.*

chart.*—A *special-purpose map* designed for purposes of navigation.

The term chart is applied chiefly to *maps* made primarily for nautical and aeronautical navigation, and to *maps* of the heavens, although the term is sometimes used to describe other *special-purpose maps.*

chart: aeronautical.*—A *chart* designed for use in navigation through the air above land or water.

An aeronautical chart shows aids and hazards to navigation. In the past, an aeronautical chart has been variously designated as an aerial navigation map, air navigation map, aerial map, aviation map, aviation chart, and air navigation chart. The use of these terms is disapproved.

chart: air navigation.—See under *chart: aeronautical chart.*

chart: aviation.—See under *chart: aeronautical chart.*

chart: bathymetric.*—A *topographic map* of the bed of the ocean.

As now published, bathymetric charts show depth curves, with areas between curves given significant color tints. Bathymetric charts are designed especially for geophysical (oceanographic) studies, but may also be used for navigation purposes.

chart: magnetic.—A *special-purpose map* depicting the distribution of one of the magnetic elements, as by isogonic lines, or of its secular change.

chart: nautical.*—A *chart* designed for use in navigation in water areas.

A nautical chart shows aids and hazards to navigation.

chromatic aberration.—See under *aberration of light (optics).*

chronograph.—An instrument for producing a graphical record of time as shown by a clock or other device.

In use, a chronograph produces a double record: the first is made by the associated clock and forms a continuous time scale with significant marks indicating periodic beats of the time keeper; the second is made by some external agency, human or mechanical, and records the occurrence of an event or of a series of events. The clock times of such occurrences are read on the time scale made by the clock. In observations for *time* and *longitude,* the clock times of star observations are recorded on the chronograph either manually by pressing a key at the instant a star is bisected by a line of the *reticle* of the telescope used in the observing, or automatically, by keeping a star bisected by a movable wire as it travels across the field of view. See *micrometer: transit.* In longitude work, the chronograph also records the time signals received from the station of known *longitude* which is used as a base.

chronometer.—A portable timekeeper with compensated balance, capable of showing *time* with extreme *precision* and *accuracy.*

Chronometers are used in scientific and engineering work (astronomy, navigation, geophysics, etc.) where *accuracy* and *precision* in the timing of observations are demanded. Chronometers are usually constructed with a special type of balance and escapement, and beat to half seconds.

chronometer: break-circuit.—A *chronometer* equipped with a device which automatically breaks an electric circuit, the breaks being recorded on a *chronograph.*

In some *chronometers,* the breaks occur every other second, on the even seconds, and a break occurs also on the fifty-ninth second to identify the beginning of the minute; in other *chronometers,* breaks occur every second except at the beginning of the minute. By recording the occurrence of events (such as star *transits*) on a *chronograph* sheet along with the *chronometer* breaks, the *chronometer* times of those occurrences are obtained.

chronometer: hack.—A *chronometer* used for visual reference, and not usually for record purposes.

In astronomical and other work, where a time recording on a *chronograph* is made, the *chronometer* controlling the recording is placed in a position as free from disturbance as possible. For use in meeting the schedule of the observations, making the necessary settings of instruments, etc., another *chronometer,* called a hack, is kept in a position where it may be easily observed.

chronometric method of determining longitude.—See under *longitude: astronomic longitude.*

circle of equal altitude.—See under *almucantar.*

circle of position.—A small circle on the *globe* (earth) at every point of which, at the instant of observation, the observed celestial body (sun, star, planet) has the same *altitude* and therefore the same *zenith distance.*

At the instant of observation the observed body is in the *zenith* of the center of the circle of position, and the point of observation is somewhere on that circle. A second observation on the same object at a different time or on a different object at the same time will determine a second and different circle of position, and if the point of observation has not been moved, it will be at an intersection of these circles. If the point of observation has moved, as may occur when on a moving ship, allowance is made for the direction and amount of the movement, and a *position* obtained for either point of observation. In navigation, a short portion of a circle of position is plotted as a straight line and termed a *Sumner line* or *line of position.*

circuit (leveling).—A continuous *line of levels,* a series of lines of levels, or a combination of lines or parts of lines of levels, such that a continuous series of measured differences of elevation extends around the circuit or loop and back to the starting point.

circuit closure (leveling).—The amount by which the algebraic sum of the measured differences of elevation around a *circuit* fails to equal the theoretical closure, zero.

circular level.—See under *level.*

cistern barometer.—See under *barometer.*

civil day.—See under *day.*

civil time.—See under *time.*

clamping error.—See under *error.*

Clarke spheroid of 1866.—See under *spheroid.*

clinometer.—A hand instrument consisting of a tube with cross-hair, with *vertical circle* and attached *spirit level* so mounted that the inclination of the *line of sight* can be read on the circle by centering the level bubble at the instant of observation.

By means of a prism, the level bubble is viewed when looking through the observing tube at the cross-hair and object observed.

clock correction.—The quantity which is added algebraically to the time shown by a clock to obtain the time of a given *meridian.*

If the clock is slow, the *correction* is positive; if fast, negative.

clock rate.—The rate of change of a *clock correction.*

Clock rate is usually expressed as the increase or decrease of the *clock correction* per day, hour, or minute as shown on the clock face. It is called the daily or hourly rate, or rate per minute. If the *clock correction* is decreasing algebraically, the clock is gaining, and the rate is negative; if the *correction* is increasing, the clock is losing, and the rate is positive.

closing corner (United States public-land surveys).—See under *corner.*

closing error.—See under *error of closure.*

closing horizon.—See under *horizon.*

closure (leveling).—See *circuit closure (leveling).*

closure of horizon.—See under *error of closure of horizon.*

closure of traverse.—See under *error of closure, traverse.*

closure of triangle.—See under *error of closure of triangle.*

co-declination.—The complement of the *declination (astronomic)*, or 90° minus the *declination.* Also called the *polar distance.*

coefficient of refraction.—See under *refraction.*

coefficient of thermal expansion; coefficient of expansion.—The relative change (expansion or contraction) in a linear dimension of a material body corresponding to a change of 1° in the temperature of the body, expressed as a ratio.

The coefficient of expansion, as it is generally called, may be in terms of the *Centigrade, Fahrenheit,* or other thermometer scale. To a very great extent, its magnitude is peculiar to the material: a steel tape has a coefficient of expansion about 25 times as great as that of an *invar tape.* It is usually expressed as a decimal fraction, and in a measure of distance, as of a *base line,* enters as a *correction* which is a product of the coefficient of expansion of the length apparatus, the distance measured, and the difference between the temperature at which the measure was made and the temperature at which the length of the apparatus is known.

coincidence (pendulum).—An exact agreement in occurrence of a prescribed phase of a *free-swinging pendulum* beat and a prescribed phase of a clock or *chronometer* beat.

The *free-swinging pendulum* being either slightly longer or slightly shorter than the clock *pendulum,* vibrates at a different speed from the clock *pendulum,* and loses or gains a whole beat between coincidences. The elapsed time between coincidences as shown by the clock or *chronometer,* enables an observer to know the number of *vibrations* made by the *free-swinging pendulum* during that time, and thus the number made during a period between precise time determinations.

In the *gravity* work of the U. S. Coast and Geodetic Survey, coincidences are observed by noting when two lines of light come into exact agreement: one is a moving line reflected by a mirror attached to the head of a *free-swinging pendulum;* the other, a fixed line of light reflected from a mirror attached to the *pendulum* support and becoming visible each second through a shutter controlled by the *chronometer.*

coincidence method (pendulum).—The determination of the period of a *free-swinging pendulum* by observing the time interval between *coincidences* with a clock *pendulum* or *chronometer* beat.

colatitude.—The complement of the *latitude,* or 90° minus the *latitude.*

Colatitude forms one side, *zenith* to pole, of the *astronomical triangle.* It is opposite the celestial body.

collimation.—The act of making a *collimation adjustment.*

This use of the term collimation may lead to confusion, which would be avoided by employing a more extended expression.

collimation: error of.—The *angle* between the *line of collimation (line of sight)* of a telescope and its *collimation axis.*

When the *collimation adjustment* of an instrument is perfect (which is never the case), the *line of collimation* and the *collimation axis* will coincide, and the error of collimation will be zero. In practical work, the adjustment is carried to where the *error* is so small that it may be considered negligible in many classes of work; or in *precise* work, after the adjustment is made, the *residual error* is either determined by observation and applied as a *correction,* or is eliminated from the result by a suitable program of observing. Error of collimation is a *systematic error,* and in a series of observations is usually treated as being of the *constant error* type.

collimation: line of.—The line through the *second nodal point* of the objective (object glass) of a *telescope* and the center of the *reticle.* It is variously called the line of sight, sight line, pointing line, and aiming line of the instrument.

The center of the *reticle* of the *telescope* of a *transit* may be defined by the intersection of cross-hairs or by the middle point of a fixed vertical wire or of a micrometer wire in its mean position. In a leveling instrument, the center of the *reticle* may be the middle point of a fixed horizontal wire.

collimation adjustment.—The process of bringing the *line of collimation* of a *telescope* into close agreement with the *collimation axis.* Also called adjustment for collimation.

collimation axis.—The line through the *second nodal point* of the objective (object glass) perpendicular to the axis of *rotation* of the *telescope.*

In a surveyor's *transit,* the collimation axis is perpendicular to the *horizontal axis* of the *telescope.* In a leveling instrument it is perpendicular to the *vertical axis* of the instrument. When the *telescope* of a *transit* is *rotated* about its *horizontal axis,* the collimation axis describes a plane, called the *collimation plane.*

collimation plane.—The plane described by the *collimation axis* of a *telescope* of a *transit* when rotated around its *horizontal axis.*

collimation position.—A term defined in Manual of Engineering Practice No. 15, A. S. C. E., as follows: "The ideal position of the line of sight of a telescope; that is, the optical axis. See, also, Line of Sight." The ideal position of the line of sight is known as the *collimation axis.* The term collimation position as defined by the A. S. C. E. seems unnecessary, and its definition appears contradictory.

collimator.—A device consisting of a convergent *achromatic lens* with a mark placed in the plane of its principal focus, so that rays from the mark through the lens emerge along parallel lines.

The mark in a collimator may be viewed from very short distances as if it were at an infinite distance, and may therefore be used in place of a distant mark when making any adjustment of the *line of sight* (*line of collimation*) of an instrument. In adjusting a surveying instrument, the *telescope* of another surveying instrument may be used as a collimator, the *reticle* furnishing the mark; or the *telescope* of a discarded instrument may be placed on a special mounting to form a permanent installation. In some astronomical instruments, a vessel of mercury, placed directly under the instrument, is used as a collimator. A prismatic eyepiece used with such an instrument is sometimes called a collimating eyepiece.

A collimator may also be constructed of special design for a particular purpose: see *collimator: vertical.*

collimator: vertical.—A *telescope* so mounted that its *collimation axis* may be made to coincide with the *vertical* (or direction of the plumb line).

The vertical collimator serves as an optical plumb line, and may be designed for use in placing a mark on the ground directly under an instrument on a high tower; or in centering an instrument on a high tower directly over a mark on the ground.

colure: equinoctial.—The *hour circle* through the *equinoxes.*

colure: solstitial.—The *hour circle* through the *solstices.*

colures.—The *hour circles* through the *equinoxes* and the *solstices.* See *colure: equinoctial; colure: solstitial.*

comb (micrometer).—A notched scale placed at right angles to the movable wire of a *micrometer,* and so designed that one turn of the micrometer screw will move the micrometer wire across one notch of the comb; the central notch of the comb in conjunction with the zero of the micrometer head furnishes a fiducial point from which all micrometer readings are reckoned.

The comb is used for keeping count of whole turns of the micrometer screw, parts of turns being read on the graduated micrometer head.

Committee Meter.—See under *meter.*

comparator.—An instrument or apparatus for measuring a dimension in terms of a *standard.*

In *plane* and *geodetic surveying,* a comparator may be an instrument for comparing *standards of length;* for subdividing such *standards;* or for determining a standard length of a measuring device (bar, tape, etc.). A field comparator or comparator base is a short line whose length is measured with accuracy and *precision,* and used to check the lengths of apparatus (tapes) used in the actual field operations.

There are special types of comparators used in astronomic and photogrammetric work, and in various kinds of laboratory work.

comparator: field.—See under *comparator.*

comparator base.—See under *comparator.*

compass: prismatic.—A small magnetic compass held in the hand when in use, and equipped with peep sights and glass prism so arranged that the *magnetic bearing* or *azimuth* of a line can be read at the same time that the line is sighted over.

compass: solar.—A surveying instrument which effects the instantaneous mechanical solution of the *astronomical triangle* (sun-zenith-pole), and permits the establishment and surveying of the *astronomic meridian* or *astronomic parallel* directly by observation.

Originally invented and used for the establishment of *astronomic meridians* and *parallels* in the survey of the *public lands of the United States,* the solar compass has been replaced by the *solar attachment* in combination with a *transit,* or by the *solar transit.*

compass: sun.—A navigation instrument for determining the direction of the *astronomic meridian* mechanically and instantaneously from an observation on the sun.

A good example of a sun compass is the one designed by A. H. Bumstead of the National Geographic Society and used by R. E. Byrd in aeronautical navigation in the polar regions. It consists of a mean-time clock with a 24-hour dial, an hour hand with a shadow pin at one end, and graduated circles for *latitude* and *azimuth* setting. In use, the clock is set for *latitude* so that the plane of its face is parallel to the plane of the *equator,* the hour hand (it has only the one hand) is set to show local time, and the whole is oriented till the pin on the hour hand casts a shadow down the middle of the hand. In this position, the lubber line will be in the local *meridian.* The term sun compass is preferred for the navigation instrument, to distinguish it from the *solar compass,* a surveying instrument which attains a similar result.

compass: surveyor's.—An instrument for determining the *magnetic azimuth* of a line of sight by means of a sighting device, a graduated horizontal circle, and a pivoted magnetic needle.

The surveyor's compass used on the early land surveys in this country employed a pair of peep sights to define the *line of sight,* and was usually mounted on a single leg, called a *Jacob's staff.* This instrument has been completely displaced by the surveyor's *transit,* and the *solar transit.*

compass rule, balancing survey.—See under *balancing a survey.*

compensated géoid.—See under *geoid.*

compensating base-line measuring apparatus.—See under *base apparatus.*

compiled map.—See under *map.*

complement of angle.—See under *angle.*

composite map.—See under *map.*

compound pendulum.—See under *pendulum.*

compression (of the earth).—See under *flattening of the earth.*

concluded angle.—See under *angle.*

condition equation.—See under *equation.*

conformal (isometric) latitude.—See under *latitude.*

conformal map projection.—See under *map projection.*

conformality.—The unique property of *conformal map projections,* in which all small or elementary figures on the surface of a sphere retain their original forms (shapes) on the *map.*

conic map projection.—See under *map projection.*

constant: level.—The amount by which the actual *line of sight* through a *leveling instrument* (when the bubble is centered in its vial) departs from the truly horizontal line through the center of the instrument, computed in millimeters per millimeter of *stadia interval.*

In case leveling rods graduated in yards instead of meters are used, the level constant, C, would be expressed in milliyards per milliyard of *stadia interval.*

constant: stadia (leveling).—The constant which is multiplied by the *stadia interval* to obtain the length of a sight in meters. Also the constant by means of which the sum of the stadia intervals for all sights of a *running* is converted to the length of the running in kilometers.

In case yard rods are in use, the resulting lengths will be in yards and kiloyards.

constant error.—See under *error.*

constant of gravitation.—See under *gravitation.*

contact base-line measuring apparatus.—See under *base apparatus.*

contact correction, transit micrometer.—A quantity applied to the *chronograph* record of a star *transit* observed with the aid of a *transit micrometer* to allow for the time required for the contact spring to cross one-half of the width of a contact strip in the head of the *micrometer.*

In order to insure a satisfactory record, the contact strips are given appreciable width, and as the micrometer wire travels from different sides of the instrument for *upper* and *lower culmination stars,* and also before and after reversal of the instrument, the contact spring produces a record sometimes from one edge of a contact strip and sometimes from the other. The contact correction is intended to reduce each observation to the middle of the contact strip which records it.

contact-slide base-line measuring apparatus.—See under *base apparatus.*

contact vernier.—See under *vernier.*

contour.—An imaginary line on the ground, all points of which are at the same *elevation* above a specified *datum* surface.

A contour is illustrated by the shore line of an imaginary body of water, whose surface is at the *elevation* represented by the contour. A contour forming a closed loop around lower ground is called a depression contour. The datum surface most generally used for contours in this country is *mean sea level.*

contour interval.—The difference in *elevation* of two adjacent *contours.* The difference of *elevations* represented by adjacent *contour lines.*

contour line.—A line on a *map* representing a *contour.*

contour map.—See under *map.*

control: basic.—In general, coordinated and correlated position data forming a framework to which detail surveys are adjusted.

Basic control may be either horizontal or vertical; it is usually executed with greater *precision* and *accuracy* than is required for dependent surveys. The basic control for the *Topographic Map of the United States* consists of *first-* and *second-order triangulation* and *traverse* and *first-* and *second-order leveling.*

control: cadastral (United States public-land surveys).—See under *cadastral control (United States public-land surveys).*

control: geodetic.—A system of *control stations* established by *geodetic* methods.

Geodetic control data are first determined in the form of *geodetic coordinates* and *azimuths,* which are now sometimes transposed into plane-coordinate data on a State system, before being used as bases for local surveys.

control: national control survey nets.—The two *control survey* nets being extended over the area of the United States by the U. S. Coast and Geodetic Survey for the control of *nautical charts* and *topographic maps,* and comprising:

1. The horizontal control survey net consisting of arcs of *first-* and *second-order triangulation,* and lines of *first-* and *second-order traverse,* a few of which have been executed by the U. S. Geological Survey, the Corps of Engineers, and other organizations. The data derived in this survey are being coordinated and correlated on the *North American datum of 1927.*

2. The vertical control survey net consisting of lines of *first-* and *second-order spirit leveling* which determine the *elevations* of thousands of *bench marks* above a common *datum, mean sea level.* This net includes lines of levels run by the U. S. Geological Survey, the Corps of Engineers, and other organizations.

control station.—A point on the ground whose *position* (horizontal, vertical) is used as a base for a dependent survey.

control survey.—A survey which provides *positions* (horizontal, vertical) of points to which supplementary surveys are adjusted.

The fundamental control survey of the United States provides the *geographic positions* (and *plane coordinates*) of thousands of *triangulation* and *traverse stations* and the *elevations* of thousands of *bench marks* which are used as the bases for *hydrographic surveys* of the coastal waters, for the control of the *topographic survey* of the United States, and for the control of many State, city, and private surveys.

control surveys: first-order.—See under *first-order work (control surveys).*

control surveys: second-order.—See under *second-order work (control surveys).*

convergence of meridians.—See under *meridian.*

conversion factor.—See under *factor.*

Conybeare leveling rod.—See under *leveling rod.*

coordinate protractor.—See under *protractor.*

coordinates (general).—Linear or angular quantities, or both, which designate the *position* of a point in relation to a given reference frame.

There are two general divisions of coordinates used in surveying: *polar coordinates* and *rectangular coordinates.* These may each be subdivided into three classes: *plane coordinates, spherical coordinates,* and *space coordinates.*

coordinates: astronomic.—Quantities which define the *position* of a point on the *geoid* with reference to the planes of the *celestial equator* and of a selected *celestial meridian.* See *latitude: astronomic; longitude: astronomic.*

coordinates: geocentric (terrestrial).—Quantities defining the *position* of a point on the earth by means of the *angles* made by a line from the center of the earth to the point with the planes of the *celestial equator* and of a selected initial *geodetic meridian.*

See *latitude: geocentric.* The term geocentric longitude is never used, as the quantity which it would designate is the same as *geodetic longitude.*

coordinates: geodetic.—Quantities which define the horizontal *position* of a point on the *spheroid of reference* with respect to the planes of the *geodetic equator* and of a selected *geodetic meridian.* See *latitude: geodetic; longitude: geodetic.*

coordinates: geographic.—An inclusive term, used to designate both *geodetic coordinates* and *astronomic coordinates.*

coordinates: grid.—Two distances which fix the *position* of a point on a *grid:* the perpendicular distance to the point from the Axis of *Y,* termed the abscissa or *x* coordinate; and the perpendicular distance from the Axis of *X,* termed the ordinate or *y* coordinate.

In surveying operations, the nominal *origin* at the intersection of the axes is usually given large numerical *coordinates,* so the inconvenience of using negative *coordinates* will be avoided. *Geodetic coordinates* (*latitudes* and *longitudes*) may be transformed into grid coordinates, and all survey computations relating to them then made by the methods and formulas of *plane surveying*. See *State coordinate systems*.

coordinates: origin of.—A point in a system of *coordinates* which serves as an initial in computing its elements or in prescribing its use.

The term, origin of coordinates, has several definitions, each so well-established in use that a single definition cannot be prescribed to the exclusion of the others. However, the following are given in the order of preferred use, but to avoid misunderstanding the use should be defined by stating the *position* of the origin in the system and giving the numerical *coordinates* assigned it.

1. The origin of coordinates is the point of intersection of the coordinate axes, from which the *coordinates* are reckoned. In mathematical treatises this origin is usually given the *coordinates* 0, 0; in surveying work, however, it is standard practice to give this origin *coordinates* having large positive numerical values, thereby avoiding the use of negative *coordinates*. See *State coordinate systems*.

2. The origin of coordinates is the point to which the coordinate values 0, 0 are assigned, irrespective of its *position* with reference to the axes.

3. The origin of coordinates is the point from which the computation of the elements of the coordinate system (projection) proceeds.

coordinates: plane-rectangular.—The perpendicular distances (*coordinates*) of a point from a pair of axes which intersect at right angles, reckoned in the plane defined by those axes.

Plane-rectangular coordinates are usually calculated from data which are in the form of *polar coordinates*, that is, distance and *direction* (*bearing* or *azimuth*) from a previously determined point. For example, the computation of *latitudes* and *departures* in land surveying. The methods used are based on plane trigonometry and geometry. The *position* of a point on the earth may be defined by plane-rectangular coordinates on a tangent plane (*local system of plane coordinates*), or on a so-called *conic* or *cylindrical map projection*, such as are used in the *State plane-coordinate systems*.

coordinates: polar (general).—The distance and *direction* from a central point of reference to a point whose position is being defined.

The point of reference is termed the pole or origin; the line (distance) connecting the origin with the point whose position is being defined is the radius vector; and the *angle* between the fixed line to which the *direction* is referred and the radius vector is the vectorial angle. In surveying operations, observations are usually put in the form of polar coordinates as a first step in the computation of *plane* or *spherical coordinates*. For example, computations of *geodetic positions* (*latitudes* and *longitudes*) are based on *azimuths* and distances from known positions.

coordinates: rectangular.—*Coordinates* on any system in which the axes of reference intersect at right angles.

A system of rectangular coordinates may be plane, curved (spherical), or spatial. See *coordinates: plane-rectangular; coordinates: geodetic; coordinates: rectangular space; etc.*

coordinates: rectangular space.—The perpendicular distances of a point from planes defined by each pair of a set of three axes which are mutually perpendicular to each other at a common point or origin.

In photogrammetry, space coordinates are also termed survey coordinates, and are the x and y *coordinates* which define the horizontal position of a point on a ground system, and the z coordinate, which is the *elevation* of the point with reference to the ground system.

coordinates: spherical.—Two quantities, angular or linear, or both, on a sphere, defining the *position* of a point with reference to two *great circles* which form a pair of axes, or with reference to an *origin* and a *great circle* through the point.

The term spherical coordinates includes *coordinates* on any surface approximating a sphere. See *coordinates: geographic; coordinates: geodetic; coordinates: astronomic.*

coordinates: vertical.—The vertical distance (*elevation*) of a point above or below a surface of reference (*datum*).

The vertical coordinate of a point may be plus or minus, according to whether the point is above or below the *datum;* the *datum* may be assigned a large positive *elevation*, so that all *elevations* referred to it will be plus. Instead of *elevation*, the term height is sometimes used.

coordination.—The placing of survey data on the same coordinate system or *datum*.

Coordination does not imply the *adjustment* of observations to remove *discrepancies*. Two field surveys over the same area may be coordinated by computation on the same *datum*, but there may remain between them *discrepancies* that can be removed only by *correlation*.

corner.—A point on a *land boundary*, at which two or more boundary lines meet.

Not the same as *monument*, which refers to the physical evidence of the corner's location on the ground.

corner (United States public-land surveys).*—A point on the surface of the earth, determined by surveying process, marking an extremity of a boundary of a subdivision of the public lands, usually at the intersection of two or more surveyed lines; often incorrectly employed to denote the physical structure, or *monument*, erected to mark the corner point.

Corners are described in terms of the points they represent. Thus:

township corner.—A corner at the extremity of a *township* boundary.

section corner.—A corner at the extremity of a *section* boundary.

quarter-section corner.—A corner at an extremity of a boundary of a *quarter section*, midpoint between or 40 chains from the controlling section corners, depending on location within the *township*.

sixteenth-section corner.—A corner at an extremity of a boundary of a *quarter-quarter section;* midpoint between the controlling corners on the *section* or *township* boundaries.

meander corner.—A corner marking the intersection of a *township* or *section* boundary and the *mean high-water* line of a body of water.

standard corner.—A corner on a *standard parallel* or *base line*.

corner: closing (United States public-land surveys).*—A *corner* at the intersection of a surveyed boundary with a previously established *boundary line*.

In the survey of the public land of the United States, when the line connecting the last *section corner* and the objective corner on an established *township* boundary departs from the *astronomic meridian* by more than the allowable deviation, the line being surveyed is projected on *cardinal* to an intersection with the *township* boundary, where a closing corner is established and connection made to the previously established corner. Closing corners are established at the intersection of a line being surveyed with a previously established *township* boundary in order to avoid excessive deviation from *cardinal* which might be required to connect with the objective corner on that boundary. Closing corners are also established at the intersection of a *township, range,* or *section line* with the boundary of a previously surveyed and segregated tract of land, such as a private land claim, mineral claim, etc.

corner: existent (United States public-land surveys).*—A *corner* whose *position* can be identified by verifying the evidence of the *monument,* or its *accessories,* by reference to the description that is contained in the field notes, or where the point can be located by an acceptable supplemental survey record, some physical evidence, or testimony.

Even though its physical evidence may have entirely disappeared, a *corner* will not be regarded as lost if its *position* can be recovered through the testimony of one or more witnesses who have a dependable knowledge of the original location.

corner: lost (United States public-land surveys).*—A *corner* whose *position* cannot be determined, beyond reasonable doubt, either from traces of the original marks or from acceptable evidence or testimony that bears upon the original *position,* and whose location can be restored only by reference to one or more inter-dependent corners.

corner: obliterated (United States public-land surveys).*—A *corner* at whose point there are no remaining traces of the *monument,* or its *accessories,* but whose location has been perpetuated, or the point for which may be recovered beyond reasonable doubt, by the acts and testimony of the interested landowners, competent surveyors, or other qualified local authorities, or witnesses, or by some acceptable record evidence.

A position that depends upon the use of collateral evidence can be accepted only as duly supported, generally through proper relation to known corners, and agreement with the field notes regarding distances to natural objects, stream crossings, *line trees,* and off-line tree blazes, etc., or unquestionable testimony.

corner: witness (United States public-land surveys).*—A monumented survey point near a *corner,* established as a *reference mark* when the corner is so situated as to render its monumentation or ready use impracticable.

A witness corner is marked in a prescribed manner, comparable with that for the true *corner.*

corner accessories (United States public-land surveys).—See under *accessories.*

correction.—A quantity which is applied to an observation or function thereof, to diminish or eliminate the effects of *errors* and obtain an improved value of the observation or function. It is also applied to reduce an observation to some arbitrary standard.

The correction corresponding to a given *error* is of the same magnitude but of opposite sign. The signs are given by the following equation: improved value = observed value + correction.

See under *error: residual error.*

correction: adjustment (leveling).—That correction which is applied to an *orthometric elevation* to produce an *adjusted elevation,* for the purpose of eliminating the effects of *circuit closures.*

correction: index (leveling).—That correction which must be applied to an observed difference of elevation to eliminate the *error* introduced into the observations when the zero of the graduations on one or both *leveling rods* does not coincide exactly with the actual physical foot or bottom surface of the rod.

correction: level (leveling).—That correction which is applied to an observed difference of elevation to correct for the *error* introduced by the fact that the *line of sight* through the *leveling instrument* is not absolutely horizontal when the bubble is centered in its vial.

correction: orthometric.—That correction which is applied to a *preliminary elevation* to correct for the *error* introduced by the fact that level surfaces at different *elevations* are not exactly parallel.

correction: rod (leveling).—That correction which is applied to an observed difference of elevation to correct for the *error* introduced when the *leveling rods* are not actually of the length indicated by the *graduations.*

correction: temperature (leveling).—That correction which is applied to an observed difference of elevation to correct for the *error* introduced when the temperature at which the *leveling rods* are used in the field is different from the temperature at which they were standardized. See, also, *standardization.*

correction for inclination of the horizontal axis.—See under *inclination of the horizontal axis.*

correction for run of micrometer.—See under *micrometer.*

correlate equation.—See under *equation.*

correlation.—The removal of *discrepancies* that may exist among survey data, so that all parts are interrelated without apparent *error.*

The terms *coordination* and correlation are usually applied to the harmonizing of *surveys* of adjacent areas or of different *surveys* over the same area. Two or more such *surveys* are coordinated when they are computed on the same *datum;* they are correlated when they are adjusted together.

county map.—See under *map.*

course (general).—This term has many meanings. In *surveying* it has been used to designate the *bearing* of a line; the length of a line; and the combination, *bearing* (or *azimuth*) and length of a line. In *control surveying,* its use can and should be avoided, but if used, it should be with the inclusive meaning, *azimuth* and length of line. It is better to use the two distinct and definite terms, *azimuth* and length of line. The following are definitions of course as used in current technical literature:

course (geography).—A route on the earth along which a river flows; the river itself.

course (land surveying).—The *bearing* of a line.

course (navigation).—The *azimuth* or *bearing* of a line along which a ship is to travel or does travel.

course (transit traverse).—The *azimuth* and length of a line, considered together.

crystal clock.—The crystal clock is a relatively new device for keeping accurate *time.* It consists essentially of a generator of constant frequency controlled by a resonator made of quartz crystal, with suitable means for producing continuous *rotation* controlled by it to operate time indicating and related mechanisms.

The above description is quoted directly from a paper, The Crystal Clock, by W. A. Marrison, published in the Proceedings of the National Academy of Sciences, July 15, 1930. The crystal clock has been used with considerable advantage over other types of portable clocks and *chronometers* in scientific work, such as *gravity* observations, where *time* intervals must be measured with the greatest possible *accuracy* and *precision.*

culmination.—The position of a heavenly body when at highest apparent *altitude;* also, for a heavenly body which is continually above the *horizon,* the position of lowest apparent *altitude.*

Culmination occurs when the body *transits* the local *meridian:* upper culmination at the upper branch of the *meridian;* lower culmination at the lower branch. As an observer approaches a pole of the earth, culmination of the fixed stars becomes less noticeable, disappearing when the pole is reached. Under some conditions, for bodies within the solar system, culmination may be largely obscured by changes in *declination.* At one time, moon culminations were extensively used in determining *astronomic longitude.*

cumulate.—A term proposed for a type of *error* which tends to accumulate in direct proportion to occurrence.

This type of *error* is covered by the expression *systematic error,* and there is no need for a new term. The use of cumulate should be avoided.

curvature correction (astronomy).—A *correction* applied to the mean of a series of observations on a star or planet, to take account of the divergence of the apparent path of the star or planet from a straight line.

Observations on a star for the purpose of determining the value of a *micrometer* are subject to a correction for curvature, as also are some *azimuth* observations. The *correction* to reduce a *latitude* observation on a star close to but not on the *meridian* to what it would be if the star were on the *meridian* may be considered a correction for curvature.

curvature correction (geodesy).—The *correction* applied in some *geodetic* work to take account of the divergence of the surface of the earth (*spheroid*) from a plane.

In *geodetic spirit leveling,* the effects of curvature and of *atmospheric refraction* are considered together, and tables have been prepared from which combined *corrections* can be taken.

curve: degree of.—The number of degrees of angular measure at the center of a circle subtended by a chord 100 feet in length.

In highway surveying, a 100-foot arc is sometimes used instead of a 100-foot chord in defining degree of curve.

curve: easement.—A *spiral curve.*

curve: point of compound curvature.—The point on a line survey where a circular curve of one radius is tangent to a circular curve of a different radius, both curves lying on the same side of their common tangent. Also termed the P. C. C.

curve: point of curvature.—The point in a line survey where a tangent ends and a circular curve begins. Also called point of curve, and P. C.

It is the point where a straight line in a survey changes to a circular curve. See *curve: point of tangency.*

curve: point of intersection.—The point where the two tangents of a circular curve meet. Also called the vertex of curve, and the P. I.

curve: point of tangency.—The point in a line survey where a circular curve ends and a tangent begins. Also called point of tangent, and P. T.

The point of tangency and point of curve are both points of tangency, their different designations being determined by the direction of progress along the line. The *point of curvature* is reached first.

curve: spiral.—A curve of varying radius connecting a circular curve and a tangent, or two circular curves whose radii are respectively longer and shorter than its own extreme radii. Also called a transition curve and an easement curve.

curve: transition.—A *spiral curve.*

curve: vertex of.—See under *curve: point of intersection.*

curve of alignment.—A line connecting two points on the surface of the *spheroid,* and defined by the condition that at every point the *azimuths* of the two end points of the line differ by exactly 180°.

A curve of alignment is a line of double curvature slightly less in length than the *normal section lines* connecting its two end points.

cut-off cylinder.—An accessory apparatus, used in *standardization* operations to refer the end of a base tape or bar standard to a ground mark.

The cut-off cylinder is a laboratory device, consisting essentially of a short rigid bar, so equipped and mounted that it forms a direct connection between the apparatus being tested and a permanent ground monument; its length and inclination provide the means for determining the relative positions of the *fiducial marks* on the base tape or bar and on the ground monument.

cylindrical equal-area map projection.—See under *map projection.*

cylindrical equal-spaced map projection.—See under *map projection.*

cylindrical map projection.—See under *map projection.*

D

datum.—Any numerical or geometrical quantity or set of such quantities which may serve as a reference or base for other quantities.

For a group of statistical references, the plural form is data: as *geographic* data for a list of *latitudes* and *longitudes.* Where the concept is geometrical and particular, rather than statistical and inclusive, the plural form is datums, as, for example, two *geodetic datums* have been used in this country in recent years.

datum (geodetic).*—A *datum* consisting of 5 quantities: the *latitude* and the *longitude* of an initial point, the *azimuth* of a line from this point, and two constants necessary to define the terrestrial *spheroid.*

It forms the basis for the computation of horizontal *control surveys* in which the curvature of the earth is considered.

datum (leveling).—A *level surface* to which heights are referred.

The *elevation* of the *datum* is usually, but not always, zero (0). The generally adopted *datum* for leveling operations in the United States is *mean sea level.* For local surveys, where a sea-level connection is not available, an arbitrary *datum* may be adopted and defined in terms of an assumed *elevation* for some physical mark (*bench mark*). A *datum* for *spirit leveling* is not a plane, and the term datum plane is incorrectly used in such connection. See, also, *datum: tidal.*

datum: Camp Colonna.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of Camp Colonna	67	25	05.11 N.
Longitude of Camp Colonna	140	59	13.50 W.
Azimuth from Camp Colonna to South Meridian	0	00	00

The above position and *azimuth* are based on astronomical observations made in 1890, the *longitude* being determined by the moon-culmination method.

datum: Golofnin Bay.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of Golofnin Bay long. station	64	27	11.461 N.
Longitude of Golofnin Bay long. station	162	52	13.700 W.
Azimuth from Golofnin Bay long. station to azimuth mark	180	00	42.9

The above position and *azimuth* based on astronomical observations made in 1899, the *longitude* being determined by the *chronometric method.*

datum: Kripniyuk-Kwiklokchun.—The *geodetic datum* defined by the following *geographic position* and *azimuths* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of Kripniyuk astro. station	62	20	06.58 N.
Longitude of Kripniyuk astro. station	165	19	27.12 W.
Azimuth from Kripniyuk astro. station to station Tent	180	00	00.0
Azimuth from Kwiklokchun astro. station to station Camp	0	00	00.0

The above position is based on *weighted means* of astronomical determinations made in 1898, the *longitudes* by the *chronometric method.*

datum: Luzon.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Balanacan	13	33	41.000 N.
Longitude of triangulation station Balanacan	121	52	03.000 E.
Azimuth, triangulation station Balanacan to triangulation station Baltasar	9	12	37.000

The Luzon datum was adopted in 1911 and was derived from observations on the Island of Luzon. It has been extended to all parts of the Philippine Islands, except a few remotely situated islands. See U. S. Coast and Geodetic Survey Field Engineers Bulletin, December 1938, page 20.

datum: New England.—Identical with the *geodetic datum* for which the name *United States standard datum* was adopted in 1901, which name was later changed to *North American datum.* Before 1901 this *datum* was defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Principio	39	35	36.692 N.
Longitude of triangulation station Principio	76	00	16.407 W.
Azimuth, triangulation station Principio to triangulation station Turkey Point	1	34	36.413

See U. S. Coast and Geodetic Survey Report for 1879, Appendix 8, pages 112–114.

datum: North American.*—The *geodetic datum* which is defined by the following *geographic position* of triangulation station Meades Ranch and the *azimuth* from that station to station Waldo, on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of Meades Ranch	39	13	26.686 N.
Longitude of Meades Ranch	98	32	30.506 W.
Azimuth, Meades Ranch to Waldo	75	28	14.52

The North American datum is identical with the *United States standard datum,* the name of the *datum* being changed in 1913 when its adoption by the governments of Canada and of Mexico for their control surveys gave it an international character.

datum: North American datum of 1927.*—The *geodetic datum* which is defined by the following *geographic position* of triangulation station Meades Ranch and the *azimuth* from that station to station Waldo, on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of Meades Ranch	39	13	26.686 N.
Longitude of Meades Ranch	98	32	30.506 W.
Azimuth, Meades Ranch to Waldo	75	28	09.64

Geodetic positions on the North American datum of 1927 are derived from the above *geographic position* and *azimuth* through a readjustment of the *triangulation* of the entire country, in which *Laplace azimuths* were introduced, and new methods of adjustment were used (*Bowie method*).

datum: Old Hawaiian.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Oahu west base	21	18	13.89 N.
Longitude of triangulation station Oahu west base	157	50	55.79 W.
Azimuth, triangulation station Oahu west base to triangulation station Oahu east base	291	29	36.0

The Old Hawaiian datum is based on an adjusted *latitude* derived from a number of *astronomic latitudes* in various parts of the islands, and an *astronomic longitude* obtained from observations of lunar *culminations,* star *occultations,* etc. See U. S. Coast and Geodetic Survey Special Publication No. 156, Triangulation in Hawaii.

datum: Panama-Colon.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Balboa Hill	9	04	57.637 N.
Longitude of triangulation station Balboa Hill	79	43	50.313 W.
Azimuth, triangulation station Balboa Hill to triangulation station Salud	185	02	39.54

See U. S. Coast and Geodetic Survey Field Engineers Bulletin, December 1938, page 23.

datum: Port Clarence.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of Port Clarence astro. station	65	16	40.18 N.
Longitude of Port Clarence astro. station	166	50	08.065 W.
Azimuth from Port Clarence astro. station to azimuth mark	0	06	17.0

The *latitude* was determined astronomically in 1900; the *longitude* is an assumed value, the astronomically-determined value being 166° 50′ 45.60″; the *azimuth* was derived from *time* observations.

datum: Pribilof Islands datums.—See under *datum: St. George Island: St. Paul Island.*

datum: Puerto Rico.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Cardona Island lighthouse	17	57	31.400 N.
Longitude of triangulation station Cardona Island lighthouse	66	38	07.530 W.
Azimuth, triangulation station Cardona Island lighthouse to triangulation station Ponce southwest base	128	36	26.2

Adopted in 1901 or soon thereafter, the Puerto Rico datum is derived from observations on the Island of Puerto Rico and the Virgin Islands. See U. S. Coast and Geodetic Survey Field Engineers Bulletin, December 1938, page 22.

datum: St. George Island.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of St. George Island astronomic station	56	36	11.31 N.
Longitude of St. George Island astronomic station	169	32	36.00 W.
Azimuth from St. George Island astronomic station to meridian mark	0	00	06

The above *latitude* and *azimuth* are based on independent astronomical determinations. The *longitude* is based on *longitudes* of St. George Island and St. Paul Island astronomic stations, which were determined by the *chronometric method*, a *weighted mean* being obtained through *triangulation* connecting the two *islands.*

datum: St. Michael.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of St. Michael astro. station	63	28	41.51 N.
Longitude of St. Michael astro. station	162	01	06.00 W.
Azimuth from St. Michael astro. station to azimuth mark	359	59	55.6

The above position depends on astronomical observations made in 1891, the *longitude* being determined by moon-*culmination* and star-*occultation* methods.

datum: St. Paul Island.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866*:

	°	′	″
Latitude of St. Paul Island astro. station	57	07	16.86 N.
Longitude of St. Paul Island astro. station	170	16	24.00 W.
Azimuth from St. Paul Island astro. station to azimuth mark	179	59	12

The above *latitude* and *azimuth* are based on independent astronomical determinations. The *longitude* is based on *longitudes* of the St. Paul Island and St. George Island astronomic stations, which were determined by the *chronometric method*, a *weighted mean* being obtained through *triangulation* connecting the two *islands.*

datum: Standard (California) astronomical datum of 1885.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Mt. Helena	38	40	04.260 N.
Longitude of triangulation station Mt. Helena	122	38	01.410 W.
Azimuth, triangulation station Mt. Helena to triangulation station Mt. Diablo	324	01	31.04

The determination of this *datum* is described in U. S. Coast and Geodetic Survey Report for 1885, Appendix 9, page 464. Its designation is taken from the so-called "old registers" of the Division of Geodesy. It is identical with the "Yolo base datum" used in the report on the California-Nevada boundary line, U. S. Coast and Geodetic Survey Report for 1900, Appendix 3, note on page 346.

datum: tidal.—A *datum* defined by a phase of the tide. See *datum (leveling).*

When used as reference surface for hydrographic surveys, tidal datums are termed datum planes; however, they are not planes and are not treated as planes, but as level surfaces, which are curved. The tidal datum in most general use in *geodetic* work is *mean sea level.* In *land surveying,* where boundaries and riparian rights are involved, *mean high water* and *mean low water* are sometimes tidal datums of considerable importance. See *island; shore.*

datum: Transcontinental Triangulation.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Hays	38	54	50.18 N.
Longitude of triangulation station Hays	99	16	16.73 W.
Azimuth, triangulation station Hays to triangulation station La Crosse	359	44	19.00

The determination of this *datum* is described in U. S. Coast and Geodetic Survey Special Publication No. 4, The Transcontinental Triangulation and the American Arc of the Parallel, 1900, and was adopted and used for scientific studies reported in that publication and in Special Publication No. 7, The Eastern Oblique Arc of the United States and the Osculating Spheroid, 1902. It served no other purposes than stated above and no formal designation was given it.

datum: Unalaska.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheriod of 1866:*

	°	′	″
Latitude of Unalaska astronomic station	53	52	36.45 N.
Longitude of Unalaska astronomic station	166	32	05.55 W.
Azimuth, Unalaska astronomic station to Observatory station	180	00	00

The above *datum* is based on astronomical observations made in 1896, the *longitude* being determined by the *chronometric method.*

datum: Valdez.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of station Pete	60	22	44.440 N.
Longitude of station Pete	145	23	48.560 W.
Azimuth from station Pete to azimuth mark	179	44	35. 4

The *geographic position* of Pete is derived through triangulation connections from the astronomic latitude of Orca astronomic station, 1898, and the astronomic longitude of Valdez longitude station, determined by telegraph in 1905. The azimuth, Pete to azimuth mark, is an astronomic azimuth.

datum: Yakutat.—The *geodetic datum* defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of Yakutat astronomic station	59	33	50.50 N.
Longitude of Yakutat astronomic station	139	47	15.60 W.
Azimuth from Yakutat astronomic station to azimuth mark	0	00	19.5

The above position and *azimuth* were determined astronomically in 1892, the *longitude* by the *chronometric method.*

datum: Yukon.—The *geodetic datum* which is defined by the following *geographic position* and *azimuth* on the *Clarke spheroid of 1866:*

	°	′	″
Latitude of triangulation station Boundary	64	40	51.42 N.
Longitude of triangulation station Boundary	141	00	00.00 W.
Azimuth, triangulation station Boundary to triangulation station Bald	270	00	00.0

This *datum* is based on a single astronomic *station* near the crossing of the 141st Meridian and the Yukon River, and was adopted for use in computing the *triangulation* along that part of the Alaska-Canada boundary defined as being the 141st Meridian.

Davidson meridian instrument.—See under *meridian telescope.*

day.—A measure of *time* based upon the *rotation* of the earth on its axis with respect to (1) the *vernal equinox,* giving the sidereal day, and (2) the sun, giving the solar day.

day: apparent solar.—The interval of *time* from a *transit* of the sun across a given *meridian* to its next successive *transit* across the same *meridian.*

As the motion of the sun is not uniform, apparent solar days vary in length through the *year,* the maximum deviation from a *mean solar day* amounting to not quite a half minute in either direction.

day: astronomical.—A *solar day* beginning at noon.

The astronomical day may be based on either *apparent solar time* or on *mean solar time.* It begins 12 hours later than the *civil day* of the same date. Prior to 1925, the astronomical day was used as a reference in the American Ephemeris and Nautical Almanac. In 1925, and since then, the *civil day* has been used instead. An exception to the above practice is found in the American Practical Navigator (Bowditch), in the 1936 edition of which astronomical day is defined as above, while in the 1938 edition it is defined as commencing at midnight, agreeing with the *civil day.* This second definition of the term is not needed, and its use is apt to produce confusion.

day: civil.—A *solar day* beginning at midnight.

The civil day may be based on either *apparent solar time* or *mean solar time;* it begins 12 hours earlier than the *astronomical day* of the same date.

day: mean solar.—The interval of *time* from a *transit* of the *mean sun* across a given *meridian* to its next successive *transit* across the same *meridian.*

The mean solar day is derived from the average length of the *apparent solar day* throughout the *year.* See *time: mean solar.*

day: sidereal.—The interval of *time* from a *transit* of the (true) *vernal equinox* across a given *meridian* to its next successive *transit* across the same *meridian.*

The length of the sidereal day is subject to slight irregularities on account of small differences between the positions of the true *equinox,* which is affected by *precession* and *nutation,* and the mean *equinox,* which is affected by *precession* but not by *nutation.* See *time: sidereal; equinox: vernal.*

day: solar.—The interval of *time* from the *transit* of either the sun or the *mean sun* across a given *meridian* to the next successive *transit* of the same body across the same *meridian.* See *day: apparent solar; day: mean solar.*

The term *day* is frequently used in connection with a particular period in one of the longer scales, as "day of month"; in most cases, the term date is preferred.

daylight-saving time.—See under *time.*

Decca.—A continuous wave electronic navigation system for measuring distance differences with respect to fixed transmitters of known position.

A ship or aircraft starting from a known point makes use of two receivers which indicate the number of nodes encountered as it moves through the fixed phase patterns of two pairs of synchronized transmitters. The phase readings are plotted on a Decca chart containing hyperbolic lines of position.

declination (astronomy).—The *angle* at the center of the celestial *sphere* between the radius passing through a celestial body and the plane of the *celestial equator.*

Declination (astronomy) is measured by the *arc* of the *hour circle* between the celestial body and the *equator;* it is plus when the body is north of the *equator,* and minus when south of it. It corresponds to *latitude* on the earth, and with *right ascension* forms a pair of *coordinates* which defines the position of a body on the *celestial sphere.*

declination: magnetic.—See under *magnetic declination.*

declination: parallel of.—A small circle on the *celestial sphere* parallel to the *celestial equator.*

declination arc (solar compass).—A graduated *arc* on a surveyor's *solar compass* or on the *solar attachment* of an engineer's *transit,* on which the *declination* of the sun (corrected for *refraction*) is set off.

This represents one side (*polar distance*) of the *astronomical triangle,* which is solved mechanically by means of the *solar compass* or *attachment.*

declination arc (surveyor's compass).—A graduated *arc* attached to the *alidade* of a *surveyor's compass* or *transit,* on which the *magnetic declination* is set off.

When the *magnetic declination* is set off on the declination arc of a *surveyor's compass* or *transit,* a reading of the needle will give a bearing corrected for that *declination.*

declination vernier.—The *vernier* used in reading the *declination arc* on a *solar compass* or *attachment,* or the *magnetic declination* on a *surveyor's compass* or *transit.*

deflection angle.—See under *angle.*

deflection anomaly.—See under *anomaly.*

deflection of the plumb line.—The *angle* at a point on the earth (*geoid*) between the *vertical* (direction of the plumb line) and the direction of the *normal* to the *spheroid of reference* through the point. Also called deflection of the vertical or station error.

It is also equal to the *angle* between the tangents to the *geoid* and the *spheroid;* and equivalent to the *topographic deflection* corrected for the effects of *isostasy.*

deflection of the vertical.—See under *deflection of the plumb line.*

degree of curve.—See under *curve.*

dep.—The usual abbreviation for *departure (plane surveying).*

departure (plane surveying).—The *orthographic projection* of a line on an east-west axis of reference.

The departure (dep.) of a line is the difference of the *meridian distances* or *longitudes* of the ends of the line. It is east or positive, and sometimes called the easting, for a line whose *azimuth* or *bearing* is in the northeast or southeast quadrant; it is west or negative, and sometimes called the westing, for a line whose *azimuth* or *bearing* is in the southwest or northwest quadrant. Departure is also a term used in navigation.

depression contour.—See under *contour.*

depth of isostatic compensation.—See under *isostatic compensation.*

desk outline.*—An *outline map* of sufficiently *small scale* to be suitable for use at a desk.

This term is not recommended for use.

detail: junction.—A sketch or working diagram showing the details of the various levelings at a junction.

detailed surveys and maps.—See under *surveys and maps.*

diaphragm (surveying).—The thin glass disk on which etched lines forming a *reticle* are placed.

difference of latitude (plane surveying).—See under *latitude.*

dip of horizon.—See under *horizon.*

direct angle.—See under *angle to right.*

direct leveling.—See under *leveling.*

direct observation.—See under *observation.*

direction.—In surveying and mapping, the *angle* between a line or plane and an arbitrarily chosen reference line or plane.

At a *triangulation station,* observed *horizontal angles* are reduced to a common initial, and termed *horizontal directions.* They are usually collected into a single list of directions, with the direction of 0° placed first, and the other directions arranged and increasing in clockwise order.

direction instrument.—See under *theodolite.*

direction method, determination of astronomic azimuth.—See under *azimuth.*

direction method of adjustment (triangulation and traverse).—See under *adjustment.*

direction of gravity.—See under *gravity.*

discrepancy.—A difference between results of duplicate or comparable measures of a quantity. The difference in computed values of a quantity obtained by different processes using data from the same *survey.*

Examples: The difference in the length of two measures of the same line. The amount by which the values of a *position* of the third point of a triangle as computed from the two other points may fail to agree, when the triangle has not been corrected for closure. Discrepancy is closely associated with, but not identical with, *closure.* See *discrepancy: accumulated.*

discrepancy: accumulated.—The sum of the separate *discrepancies* which occur in the various steps of making a survey or of the computation of a *survey.*

Example: If two independently run lines of leveling over the same series of *bench marks* are computed separately, differences or *discrepancies* between the two sets of values of *elevations* will accumulate. This does not mean that the accumulated *discrepancy* will necessarily increase in magnitude.

dispersion; dispersive power.—See under *refraction: index of refraction.*

distance angle.—See under *angle.*

distribution map.—See under *map.*

diurnal aberration.—See under *aberration of light (astronomy).*

diurnal parallax.—See under *parallax.*

diurnal variation.—That component of a determinable magnitude which passes through a complete cycle in one day.

Such a component is said to have a periodicity of one day. Example: the diurnal variation of *magnetic declination.*

divergence (leveling).—The difference between the numerical values of two *runnings* over the same *section* of a *line of levels.* Divergence and "partial" are practically the same thing.

divergence: accumulated (leveling).—The algebraic sum of the *divergences* or "partials" for the *sections* of a *line of levels,* from the beginning of the line to any section end at which it is desired to compute the total divergence.

division; graduation.—The placing of marks on an instrument or device to represent *standard* values thereon. Also, the marks so placed.

The term division is applied especially to a circle, the value of a complete circle being absolute, that is, not depending upon an adopted *standard.* Graduation is more often applied to the placing of intermediate marks on an instrument or device (tape, thermometer) where interpolation is made between marks which represent conventional or standard values. However, no exact distinction is made between the two terms, division and graduation, as generally used. See *calibration; standardization.*

Doolittle Method.—A modification of Gauss' method of solving *normal equations.*

The Doolittle Method, with further modifications made in recent years, is used in the adjustment of survey data by members of the U. S. Coast and Geodetic Survey. It is named for M. H. Doolittle, for many years a member of the Computing Division of that Bureau, and author of a description and explanation of the method which appeared in the U. S. Coast and Geodetic Survey Report for 1878, pages 115–118.

double dating.—Dating an event according to both the *Julian* and the *Gregorian Calendars.*

In Great Britain and her colonies, the *Julian Calendar,* with March 25 as the beginning of the year, was officially replaced in 1752 by the *Gregorian Calendar,* with January 1 as the beginning of the year. The difference between the two calendars having increased from 10 days in 1582 to 11 days, the change was made by having September 2, 1752, followed immediately by September 14. To avoid confusion resulting from the change, dates according to the *Julian Calendar* were marked "old style" or "(O. S.)"; those on the *Gregorian Calendar,* "new style" or "(N. S.)". In double dating early American records, not only the day and the month, but also the year, are subject to change. In England, at the time of the change, the official year commenced on March 25, so that March 24, 1730 (O. S.) corresponded to April 4, 1731 (N. S.), but the next day was March 25, 1731 (O. S.) or April 5, 1731 (N. S.). This practice was not universal, for there was some popular use of January 1 as the beginning of the year.

double meridian distance.—See under *meridian distance.*

double zenith distance.—See under *zenith distance.*

doubly azimuthal map projection.—See under *map projection.*

draconitic (nodical) month.—See under *month.*

drag (theodolite).—A slight movement of the graduated circle of a *theodolite* produced by the *rotation* of the *alidade.*

Drag may be due to excessive friction in the instrument *centers,* to undue *tolerance* in the fit of the *centers,* or to instability in the instrument supports.

dummy pendulum.—See under *pendulum.*

dumpy level.—See under *leveling instrument.*

duplex base-line measuring apparatus.—See under *base apparatus.*

duplicate level line.—See under *level line.*

dynamic correction.—The quantity that must be added to the *orthometric elevation* of a point to obtain its *dynamic number.*

dynamic map.—See under *map.*

dynamic number.—The work required to raise a unit mass from *sea level* to a given point, expressed in *absolute units.*

For all points in a given *level surface* the dynamic number is the same, and is not very different from the height (in the unit chosen) of any point in the surface. In the metric system, the dynamic number of a point is the work required to raise a mass of 1 kilogram against the force of *gravity* from *sea level* to the *level surface* passing through the point, the work being measured in standard kilogram-meters at *sea level* in latitude 45°. In the English system, the pound and standard foot-pound are employed instead of the *meter* and standard kilogram-meter.

dynamic temperature correction (pendulum).—The *correction* to the observed period of a *pendulum* for the rate of change of its temperature.

The dynamic temperature correction is definitely related to the type of pendulum apparatus used, and is dependent especially on the mounting of the thermometer and the shape of the *pendulum.* It may vary not only in magnitude but may even change sign when the type of apparatus is changed.

dyne.—A force which, acting on a mass of one gram, imparts to that mass an *acceleration* of one centimeter per second per second. See *gal.*

The dyne is the unit of force of the *c. g. s. system* of units. Until about 1930, the dyne was used by the U. S. Coast and Geodetic Survey in stating values of gravity. Since that time, *gravity* has been reported in terms of the *gal,* the c. g. s. unit of *acceleration.*

E

easement curve.—See under *curve.*

easting.—See under *departure* (*plane surveying*).

eccentric object observed.—See under *eccentric signal.*

eccentric reduction (triangulation).—The *correction* which must be applied to an observed *direction* made with instrument or signal, or both, eccentric, to reduce the *observed value* to what it would have been if there had been no *eccentricity.*

eccentric signal.—A signal (target) which is not in the same vertical line with the *station* which it represents.

An observed *direction* to an eccentric signal is reduced by the application of a computed *correction* to what it would be if the signal were in the same vertical line with the station center which it represents. See *center: reduction to center.* The eccentric signal is known as the eccentric object observed.

eccentric station.—A *survey* point over which an instrument is centered and observations made, and which is not in the same vertical line with the *station* which it represents and to which the observations will be reduced before being combined with observations at other *stations.*

In general, an eccentric station is established and occupied when it is impracticable to occupy the station center, or when it becomes necessary in order to see points which are not visible from the station center. As the data resulting from the occupation of an eccentric station are referred to the station center, which should be marked with a permanent *monument,* the mark at an eccentric station is usually of a temporary character. See *center: reduction to center.*

eccentricity.—Amount of deviation from a center.

See *eccentricity of instrument; eccentric station; eccentricity of alidade.*

eccentricity (surveyor's compass).—An effect due to one or a combination of the following conditions: a straight line through the ends of the magnetic needle fails to pass through the center of *rotation* of the needle; the center of *rotation* of the needle is not coincident with the center of figure of the graduated circle; the *line of sight* fails to pass through the vertical axis of the instrument.

eccentricity of alidade.—The distance between the center of figure of the index points on an *alidade* and the center of figure of the graduated circle.

The index points (of *vernier* or *micrometer microscope*) are carried by the *alidade,* and any eccentricity of alidade combines with *eccentricity of circle* to form *eccentricity of instrument.*

eccentricity of circle.—The distance between the center of figure of a graduated circle and its center of *rotation.*

Eccentricity of circle is usually expressed in terms of its equivalence in seconds of *arc* on the circle. It may be made quite small by instrument-shop adjustment. Its effect on an observed *direction* is eliminated by reading the circle at equidistant points around its circumference. See *eccentricity of instrument.*

eccentricity of ellipse.—The ratio of the distance between the center and a focus of an ellipse to the length of its semi-major axis.

The semi-major and semi-minor axes are represented by a and b, respectively, and the eccentricity by e, then

$$e=\sqrt{\frac{a^2-b^2}{a^2}}.$$

eccentricity of instrument.—The combination: *eccentricity of circle* and *eccentricity of alidade.*

The effect of eccentricity of instrument on an observed *direction* is eliminated by having the *verniers* or *micrometer microscopes* with which the circle is read spaced at equal distances around the circle.

eccentricity of spheroid of revolution.—The *eccentricity of an ellipse* forming a meridional section of the *spheroid.*

eclipse.—1. The reduction in visibility or disappearance of a non-luminous body by passing into the shadow cast by another non-luminous body. 2. The apparent cutting off, wholly or partially, of the light from a luminous body by a dark body coming between it and the observer.

1. The first type of eclipse is exemplified by a lunar eclipse, the moon passing through the shadow cast by the earth; or by the passage of a satellite into the shadow cast by its planet; but when the satellite actually passes directly behind its planet, it may properly be termed an *occultation.*

2. The second type of eclipse is exemplified by a solar eclipse, caused by the moon passing between the sun and the earth. If the relative positions and distances are such that at a point on the earth the sun is completely obscured, the eclipse is total; if the distances are such that when in line with the sun, the moon is surrounded by a ring of light, the eclipse is annular; and when the moon passes to one side of a straight line from sun to observer, and shows a crescent of light, it is a partial eclipse.

ecliptic.—The *great circle* of the *celestial sphere* traced by the plane of the earth's orbit (path of center of gravity of earth-moon system).

The ecliptic represents very closely, but not exactly, the apparent path of the sun in the sky. The points in which the ecliptic intersects the *celestial equator* are the *equinoxes*, and the *angle* of intersection is the *obliquity of the ecliptic.*

ecliptic: obliquity of the.—The acute *angle* of intersection between the *ecliptic* and the *celestial equator.*

The obliquity of the ecliptic is slightly less than 23½° and is decreasing at the rate of about ½″ per year.

Egault level.—See under *leveling instrument.*

elevation.—The vertical distance of a point above or below a reference surface or *datum.*

In the fundamental horizontal *control survey* of this country, the *datum* for elevations is *mean sea level.*

elevation: adjusted.—The *elevation* resulting from the application of an *adjustment correction* to an *orthometric elevation.* Also the *elevation* resulting from the application of both an *orthometric correction* and an *adjustment correction* to a *preliminary elevation.*

elevation: field.—An *elevation* taken from the field computation of a *line of levels.*

elevation: fixed.—An *elevation* which has been adopted, either as a result of tide observations or previous adjustment of spirit leveling, and which is held at its accepted value in any subsequent *adjustment.*

elevation: orthometric.—A *preliminary elevation* to which the *orthometric correction* has been applied.

elevation: preliminary.—An elevation arrived at in the office after the *index, level, rod,* and *temperature corrections* have been applied to the observed differences of elevation and new *elevations* have been computed.

elevation: standard.—An *adjusted elevation* based on the Sea Level Datum of 1929.

ellipse: eccentricity of.—See under *eccentricity of ellipse.*

ellipse: ellipticity of.—See under *ellipticity of ellipse.*

ellipsoid.—See under *spheroid.*

ellipsoid: triaxial.—An *ellipsoid* having three unequal axes, the shortest being its polar axis, while the two longer ones lie in the plane of its *equator.* See *gravity formula, longitude term.*

elliptic arc: plane.—See *plane elliptic arc.*

ellipticity of an ellipse.—The ratio between the difference in length of the semi-axes of an ellipse and its semi-major axis.

The semi-major and semi-minor axes are represented by a and b, respectively, and the ellipticity by ϵ, then $\epsilon = \frac{a-b}{a}$. The *ellipticity of the spheroid,* referring to the earth, is called the *flattening of the earth,* or its *compression,* and is also represented by f.

ellipticity of the spheroid.—See under *flattening of the earth.*

engineering map.—See under *map.*

engineering survey.—See under *survey.*

ephemeris.—A statement presenting positions and related data for a celestial body for given epochs (dates) at uniform intervals of *time.* Also a publication containing such data for a number of celestial bodies.

Such a publication is the American Ephemeris and Nautical Almanac: this contains for specified instants of *time* the numerical values of *coordinates* of the principal celestial bodies referred to circles whose positions are independent of the diurnal *rotation* of the earth; also the elements of the positions of the reference circles, and numbers used in computing the effects upon those *coordinates* of changes in the *position* of an observer; and in general, all those phenomena relating to the heavenly bodies which may be regarded as functions of *time.* While custom has approved the use of the singular form, ephemeris, to designate a publication containing data relating to a number of celestial bodies, where specific mention is made of those bodies, the plural form, ephemerides, is used; for example, the ephemerides of the planets. For a single body, the singular form is used: for example, the ephemeris of the sun.

equal-area (authalic) latitude.—See under *latitude.*

equal-area map projection.—See under *map projection.*

equation: angle.—A *condition equation* which expresses the relationship between the sum of the measured *angles* of a closed figure and the theoretical value of that sum, the unknowns being the *corrections* to the observed *directions* or *angles,* depending upon which are used in the *adjustment.*

Sometimes called a triangle equation, an angle equation is used to make the sum of the three observed *angles* of a triangle, with *corrections* applied, equal to 180° plus the *spherical excess* of the triangle.

equation: azimuth.—A *condition equation* which expresses the relationship between the fixed azimuths of two lines which are connected by triangulation or *traverse.*

When a *survey* (*triangulation* or *traverse*) connects two lines whose *azimuths* are fixed by direct observation or by previous *surveys,* an azimuth equation is used to make the *azimuth* of either line as computed through the adjusted *survey* from the other line agree with its *azimuth* as previously fixed.

equation: condition.—An equation which expresses exactly certain relationships that must exist among related quantities, which are not independent of one another, exist a priori, and are separate from relationships demanded by observation.

Example: in measuring the *angles* of a triangle, no condition exists until all three *angles* are measured. The condition equation will then express the condition that the three measured *angles* plus certain *corrections* must equal 180° plus the *spherical excess* of the triangle. The various condition equations set up in *survey* work are defined under terms which are descriptive of the conditions: as, *angle equation, side equation, length equation, latitude equation, longitude equation, azimuth equation.*

equation: correlate.—An equation derived from an *observation* or *condition equation,* employing undetermined multipliers, and expressing the condition that the sum of the squares of the *residuals* (or *corrections*) resulting from the application of these multipliers to the *observation* or *condition equations* shall be a minimum.

In the *least-squares* adjustment of *triangulation,* correlate equations are formed directly from the *observation* or *condition equations,* there being as many correlate equations as there are *corrections* to be determined, but only as many undetermined multipliers (correlates or correlatives) as there are *observation* or *condition equations.* From these correlate equations, the *normal equations* are formed, equal in number to the undetermined multipliers which constitute the unknowns in the *normal equations.* The solution of the *normal equations* determines values for the multipliers which, when substituted in the correlate equations, give values for the corrections which will satisfy the *observation* or *condition equations,* make the observations and their functions consistent among themselves, and at the same time make the adjusted values the most probable that can be derived from the given observations.

equation: error.—The probability equation which expresses the laws of the occurrence of *accidental errors.*

The error equation is the basis of the method of *least squares,* used in the *adjustment of observations* for determining the *most probable value* of a result from those observations.

equation: Laplace.—See under *Laplace equation.*

equation: latitude.—A *condition equation* which expresses the relationship between the fixed *latitudes* of two points which are connected by *triangulation* or *traverse.*

When a *survey* (*triangulation* or *traverse*) connects two points whose *latitudes* have been fixed by direct observation or by previous *surveys,* a latitude equation is used to make the *latitude* of either point as computed through the *survey* from the other point agree with its *latitude* as previously fixed.

equation: length.—A *condition equation* which expresses the relationship between the fixed lengths of two lines which are connected by *triangulation.*

When a section of *triangulation* connects two lines whose lengths are fixed by direct measurement or by previous *triangulation,* a length equation is used to make the length of either line as computed through the adjusted *triangulation* from the other line agree with its length as previously fixed.

equation: longitude.—A *condition equation* which expresses the relationship between the fixed *longitudes* of two points which are connected by *triangulation* or *traverse.*

When a *survey* (*triangulation* or *traverse*) connects two points whose *longitudes* have been fixed by direct observation or by previous *surveys,* a longitude equation is used to make the *longitude* of either point as computed through the *survey* from the other point agree with its *longitude* as previously fixed.

equation: normal.—An equation derived from *observation* or *condition equations* or from *correlate equations,* expressing the condition that the sum of the squares of the *residuals* (or *corrections*) resulting from the substitution in the *observation* or *condition equations* of factors obtained from the normal equations either directly or through the *correlate equations* shall be a minimum.

In a *least-squares* adjustment, *corrections* are desired to observed values which are connected together by a series of *observation* or *condition equations,* the number of such equations being smaller than the number of observed values on which they depend. The basic equations are transformed into normal equations, either directly or through the medium of *correlate equations,* which express the condition stated above and which contain the same number of unknowns as there are equations. Factors obtained from the solution of normal equations, either directly or through the *correlate equations,* are applied to the *observation* or *condition equations* to obtain the desired *corrections.*

equation: observation.—A *condition equation* which connects interrelated unknowns by means of an observed function.

Example: An *angle equation:* the unknowns are the *corrections* to the *observed angles;* the function is the sum of those *observed angles* expressed as a *closing error.*

Alternate definition: A *condition equation* connecting the function observed and the unknown quantity whose value is sought.

equation: perpendicular (traverse).—A *condition equation* to reduce to zero the algebraic sum of the *projections* of the separate lines of a *traverse* upon perpendiculars to a fixed line with which the *traverse* forms a closed figure.

The perpendicular equation and the *azimuth equation* together provide for the removal of *errors of closure* of a *traverse* which forms a loop with some fixed line such as a line of adjusted *triangulation,* by determining *corrections* to the *observed angles* of the *traverse.* The *projections* of the lines of the *traverse* upon lines which are perpendicular to a fixed line correspond to *departures* of those lines when the fixed line is considered the *meridian* of reference.

equation: side.—A *condition equation* which expresses the relationship between the various sides in a *triangulation* figure as they may be derived by computation from one another.

A side equation is used to make the computed length of a triangle side the same for all routes through the *triangulation* whereby it may be derived.

equation of time.—See under *time.*

equator: astronomic.—The line on the surface of the earth whose *astronomic latitude* at every point is 0°. Also termed the terrestrial equator.

Due to the *deflection of the plumb line,* the astronomic equator is not a plane curve. However, the *verticals* at all points on it are parallel to one and the same plane, the plane of the *celestial equator;* that is, the *zenith* at every point on the astronomic equator lies in the celestial equator.

equator: celestial.—The *great circle* on the *celestial sphere* whose plane is perpendicular to the axis of *rotation* of the earth.

In astronomic work, because parallel lines meet at infinity, the plane of the celestial equator is sometimes assumed to pass through the point of observation.

equator: geodetic.—The circle on the *spheroid* midway between its poles of *revolution*.

The geodetic equator is the line of 0° *geodetic latitude*, from which *geodetic latitudes* are reckoned, north and south, to 90° at the poles. The plane of the geodetic equator cuts the *celestial sphere* in a line coinciding with the *celestial equator*, because the axis of the *spheroid* of reference is by definition parallel to the axis of rotation of the earth.

equator: terrestrial.—See under *equator: astronomic equator*.

equatorial horizontal parallax.—See under *parallax: diurnal parallax*.

equatorial intervals.—The *angles*, expressed in units of *time*, between the various lines which compose the *reticle* of an astronomical *transit* and the mean position of those lines.

The equatorial interval for a given line of a *reticle* is equal to the time interval required for the image of a star on the *equator* (*declination*, 0°) to travel from the line in question to the mean line of the *reticle*, or from the mean line to the line in question, the instrument being adjusted in the *meridian*. In *time* determinations with an *astronomical transit*, equatorial intervals are used in reducing incomplete observations to a mean value.

equiangular spiral.—See under *rhumb line*.

equiangulator.—An *astrolabe*.

equinoctial colure.—See under *colure*.

equinox.—A point of intersection of the *celestial equator* and the *ecliptic*.

There are two equinoxes: the *vernal equinox*, apparently traversed by the sun at the beginning of spring; and the *autumnal equinox*, apparently traversed by the sun at the beginning of autumn.

equinox: autumnal.—The point of intersection of the *celestial equator* and the *ecliptic*, apparently traversed by the sun in passing from north to south.

equinox: vernal.—The point of intersection of the *celestial equator* and the *ecliptic*, apparently traversed by the sun in passing from south to north.

The vernal equinox is also called the "first point of Aries" and "first of Aries." It is the point from which *right ascension* is reckoned along the *celestial equator* and *celestial longitude* along the *ecliptic*. The position of the true *equinox* is affected by *precession* and *nutation*, while the position of the mean *equinox* is affected by *precession* but not by *nutation*. See *day: sidereal; time: sidereal*.

equipotential surface (geodesy).—A surface having the same potential of *gravity* at every point.

Having the same potential of *gravity* at every point, no work is done when a body is moved about in such a surface. When bodies of the same mass are moved from one equipotential surface to another, the same amount of energy is developed or expended, accordingly as the bodies are lowered or raised, and regardless of the route followed. In lifting a unit mass from *sea level* to any other specified equipotential surface, the amount of work required is the same at the *equator* as at the poles: at the *equator* the height of any equipotential surface above *sea level* is greater and the force of gravity at that surface less than for the same equipotential surface at the poles. An equipotential surface is a *level surface*. A particular equipotential surface is identified by its *dynamic number*.

equivalent map projection.—See under *map projection: equal-area map projection*.

error (general).—The difference between an observed or calculated value of a quantity and the ideal or true value of that quantity.

Since the ideal or true value of a quantity, with few exceptions (see below), cannot be known with exactness, the term error is applied to a difference between an observed or calculated value of a quantity and some value determined by established procedure and used in lieu of the ideal or true value. Exceptions: the ideal or true value of a quantity can be known with exactness when it is (a) mathematically determinable, independent of observation, as for example, the sum of the three angles of a plane triangle is 180°; and (b) when it is a conventional value established by authority, as for example, the length of the *meter* (unit) defined by the *International Prototype Meter* at the International Bureau of Weights and Measures. Errors are of various kinds, depending upon how and where they originate. The term error with appropriate adjective or qualifying clause is used to designate the kind of error: as, *accidental error, error of observation*, etc.

error: accidental.—An error, sometimes designated as an irregular *error*, produced by irregular causes whose effects upon individual observations are governed by no fixed law connecting them with circumstances and which therefore can never be subjected to a priori computation.

Theoretically, an accidental error is composed of an infinite number of independent infinitesimal *errors*, all of equal magnitude, each as likely to be positive as negative. In practice, an accidental error is composed of an indefinitely large number of elemental *errors*, each as apt to be positive as negative.

It is upon the probabilities expressed in the above definition that the principle of *least squares* is based; and it is to the elimination of accidental errors only that *least squares* may properly be applied.

error: average.—The mean of all the *errors* taken without regard to sign.

Average error is sometimes defined as the mean without regard to sign of the mean of the plus *errors* and the mean of the minus *errors*. Since there will be about the same number of plus *errors* as of minus *errors* in a well-balanced series of observations, the two methods of obtaining the average error will give practically the same result. In computing the average error, the *residual errors* (residuals) are used.

error: clamping.—A *systematic error* in observations made with a *repeating theodolite* due to strains set up by the clamping devices of the instrument.

error: constant.—A *systematic error* which is the same in both magnitude and sign through a given series of observations.

Constant errors are sometimes considered as forming a class distinct from *systematic errors*, but the distinction is unimportant, and in practice constant errors are treated as a type of *systematic error*. A constant error tends to have the same effect upon all the observations of the series or portion thereof under consideration, and therefore does not usually have any influence on the computed *probable error* of the series. An example of a constant error is the index error of a *precision* instrument.

error: external.—See under *error: theoretical* (or *external*) *error*.

error: instrumental.—A *systematic error* arising from imperfect condition of the instrument used.

Instrumental errors may arise from conditions that are constant with an instrument, as *errors* of graduation of a circle; or they may arise from lack of complete adjustment of some part of the instrument, as the *error of collimation.* Instrumental errors are susceptible of laboratory determination; they may be eliminated from a result by a suitable program of field procedure.

error: mean.—See under *error: mean-square error.*

The use of mean error should be avoided because of possible confusion with *average error,* which has a different meaning.

error: mean of the errors.—See under *error: average error.*

The use of mean of the errors should be avoided because of possible confusion with *mean error,* which has a different meaning.

error: mean-square.—The quantity whose square is equal to the sum of the squares of the individual *errors* divided by the number of those *errors.*

In the computation of the mean-square error from observations of unequal weight, account must be taken of the *weights.* In practice, *residual errors* are used in determining a mean-square error. The use of *mean error,* in place of mean-square error, should be avoided, because of possible confusion with *average error* which has a different meaning.

error: parallactic.—An *error* due to *personal* or *instrumental parallax.*

error: personal.—A *systematic error* caused by an observer's personal habits in making observations, or due to his tendency to react mentally and physically in the same way under similar conditions.

A *systematic error* arising from personal habits of an observer, such as standing in the same position relative to the end of a tape when measuring the length of a line, may be eliminated from a result by an observing program (shift of position) which makes the *error* positive for half of the observations and negative for the remainder of the observations. See *parallax: instrumental.*

A *systematic error* arising from the mental or motor reaction of an observer to the sensory perception of a phenomenon may be eliminated from a result either by an observing program which neutralizes positive values with negative values, or by determining the size and sign of the *error* by some mechanical means. In the determination of *time,* the personal error is termed *personal equation.*

error: probable.—A quantity of such size that the probability of the occurrence of an *error* larger than that quantity is the same as the probability of the occurrence of an *error* of lesser magnitude.

If the *errors* in a series of observations be arranged in order of magnitude without regard to sign, and if the series be indefinitely large, the probable error will fit the middle place in that list of errors.

Expressed in another way, the probable error of a result is a quantity such that the probability that a second determination obtained under the same conditions as the first will differ from the first determination by less than the probable error is the same as the probability that such difference will be greater than the probable error.

The probable error of the result of a series of observations is a function of the *accidental errors* attending the individual observations of the series. A *systematic error* may often remain in the series with little effect on the size of the probable error. Probable error does not mean an *error* that is more apt to occur than an *error* of any other size. Theoretically, a probable error is derived from the *errors* in an indefinitely large series of observations; in practice, a probable error is calculated from the *residual errors* in a series of limited length. Probable errors are useful in comparing the *accuracy* of similar observations, and serve as criteria in prescribing degrees of *accuracy* and *precision* to be attained.

error: residual.—The difference between any value of a quantity in a series of observations, corrected for known *systematic errors,* and the value of the quantity obtained from the combination or adjustment of that series.

Residual errors correspond to the *corrections* obtained in a *least-squares* adjustment. They are sometimes called simply *errors,* and sometimes residuals. The latter term is generally used in referring to actual values in a specific computation. In practice, it is the residual errors which enter into a computation of *probable error.*

error: systematic.—An *error* whose algebraic sign and, to some extent, magnitude bear a fixed relation to some condition or set of conditions.

In its broadest sense, the term systematic error includes *constant errors.* Under similar conditions, systematic errors tend to be repeated; if the conditions do not change, the *error* will be constant. The inclusive definition is preferred.

Systematic errors are regular, and therefore are subject to a priori determination. They are generally eliminated from a series of observations by a suitable observing program or by computation, before the application of *least-squares* for the elimination or reduction of *accidental errors.* Systematic errors are divided into *theoretical* or *external errors, instrumental errors,* and *personal errors,* according to their origin and nature.

error: theoretical (or external).—A *systematic error* arising from natural physical conditions, outside the observer.

Theoretical errors are exemplified by the effect of *atmospheric refraction* in *spirit leveling;* by the changes in the length of a tape due to its thermal expansion; by the effects of atmospheric pressure on *elevations* determined with a *barometer.* Theoretical errors may be controlled to some extent by making observations only when natural conditions are favorable, as measuring the length of a line with a steel tape at night or in cloudy weather.

Theoretical errors are determined by means of formulas and observed data obtained especially for the purpose, as determining the length of a metal tape by suitable formula using the *coefficient of thermal expansion* and the observed temperature of the tape.

error: true.—See under *error (general).*

True error corresponds to *error (general)* defined above; its use should be avoided.

error equation.—See under *equation.*

error of closure.—The amount by which a value of a quantity obtained by *surveying* operations fails to agree with another value of the same quantity held fixed from earlier determinations or with a theoretical value of the quantity.

See *error of closure, traverse; error of closure, triangle; error of closure, levels; error of closure, horizon.*

error of closure, angles.—The amount by which the actual sum of a series of *angles* fails to equal the theoretically exact value of that sum. See *error of closure of horizon; error of closure of triangle.*

error of closure in azimuth.—The amount by which two values of the *azimuth* of a line, derived by different *surveys* or along different routes, fail to be exactly equal to each other.

Generally, one value is derived by computations carried through the *survey* (*triangulation* or *traverse*) ; the other is an *adjusted* or fixed value determined by an earlier or a more *precise survey,* or by independent *astronomical* observations.

error of closure in leveling.—The amount by which two values of the *elevation* of the same *bench mark,* derived by different *surveys* or through different *survey* routes or by independent observations, fail to be exactly equal to each other.

The closure may be developed in a line of *leveling* which starts and ends on different *bench marks* whose *elevations* are held fixed; or it may start and close on the same *bench mark.*

error of closure of horizon.—The amount by which the sum of a series of *horizontal angles* measured between adjacent lines in a complete circuit of the *horizon* fails to equal exactly 360°. See *horizon: closing the horizon.*

error of closure of triangle.—The amount by which the sum of the three *observed angles* of a triangle fails to equal exactly 180° plus the *spherical excess* of the triangle. See *equation: angle equation.*

error of closure, traverse.—The amount by which a value of the *position* of a traverse station as obtained by computation through a *traverse* fails to agree with another value of the same station as determined by a different set of observations or route of *survey.*

The *traverse* may run between two stations whose *positions* are held fixed, or it may start from and end on the same station. In either case there are two values for the *position* of the final station; one known before the *traverse* was computed, the other obtained by computations carried through the *traverse;* the difference between these is the *error of closure.* It may be resolved into closure in latitude, and closure in longitude (or departure). The total closure is also spoken of as closure in position.

error of collimation.—See under *collimation.*

error of observation.—The difference between an observed value of a quantity and a value adopted as representing the ideal or true value of that quantity.

Errors of observation are composed of either one or both of two general classes of *error; accidental errors* and *systematic errors.* Sometimes *constant errors* are considered as forming a third class, but more often these are treated as belonging with the *systematic errors.* Errors of observation are also classified for study according to their origin, as *theoretical errors* or *external errors, instrumental errors,* and *personal errors.* The algebraic sign of an error of observation is determined from the equation: error = observed value − adopted ideal value.

The sign of the *error* is opposite that of the corresponding *correction.*

existent corner (United States public-land surveys).—See under *corner.*

explement of angle.—See under *angle.*

exploratory survey.—See under *survey.*

exterior angle.—See under *angle.*

extra foresight.—See under *foresight.*

eyepiece micrometer.—See under *micrometer: ocular.*

F

factor: conversion.—That factor by which the numerical value of a measurement made in one system of units must be multiplied to arrive at the numerical value of the same measurement in another system of units.

Fahrenheit scale.—A temperature scale in which 32° marks the freezing point and 212° the boiling point of water at 760 mm. barometric pressure.

falling (United States public-land surveys).*—The distance, measured along an established line from its intersection with a *random line,* to a *corner* on which the *random line* was intended to close.

fathometer.—An echo-sounding instrument used for depth measurements in water.

For descriptions of several types of fathometers, see U. S. Coast and Geodetic Survey Special Publication No. 143, Hydrographic Manual.

fiducial mark.—An index line or point. A line or point used as a basis of reference.

field elevation.—See under *elevation.*

field position.—See under *position.*

figure of the earth.—See under *geoid, spheroid.*

finder circle.—A small *vertical circle* with coarse *graduations* and accessory *spirit level* of low sensitivity attached to an astronomical *telescope* in a plane perpendicular to its *horizontal axis,* and used in placing the *telescope* in approximate position for observing.

With a computed *altitude* at the proposed time of observation, the *telescope* is pointed in the direction of the object to be observed; when that object comes into the field of view, the finder circle will have served its purpose, and the observations will be made with the more precise apparatus provided for the purpose.

finder telescope.—A small *telescope* of low power and wide angular field of view, attached to and parallel with a larger *telescope* of much higher power and very restricted field of view, and used in pointing the larger *telescope* at a celestial object.

The finder telescope serves as a sighting device; a celestial object is easily sighted in the finder telescope, and brought to the center of its *reticle,* when it will be visible in the large *telescope* also.

first-order bench mark.—See under *bench mark.*

first-order control surveys.—See under *first-order work* (*control surveys*).

first-order level.—See under *leveling instrument.*

first-order leveling.—See under *leveling.*

first-order traverse.—See under *traverse.*

first-order triangulation.—See under *triangulation.*

first-order work (control surveys).—The designation given *survey* work of the highest prescribed order of *precision* and *accuracy.*

Such *surveys* were formerly called primary; in 1921, representatives of the various Federal map-making and map-using organizations changed the designation to *precise;* and in 1925, the Federal Board of Surveys and Maps adopted the present designation (first-order), and established *standards* for it by prescribing criteria for acceptable work. See under *triangulation; leveling.*

Fischer level.—See under *leveling instrument.*

fix.—The position on a *map* of a point of observation obtained by *surveying* processes. Also, the act of determining such a *position.*

fixed elevation.—See under *elevation* and also *position: adjusted.*

fixed position.—See under *position: adjusted.*

flare triangulation.—See *triangulation: flare.*

flash apparatus.—An auxiliary apparatus used in timing a *pendulum* during observations for *intensity of gravity.*

The flash apparatus used in *pendulum* observations is composed of a *telescope* for observing, an electric light for illumination, and an electro-magnet which is operated by a break-circuit *chronometer* and which controls a system of shutters admitting periodic flashes of light into the *receiver* and onto the mirror of the *pendulum* whence they are reflected back into the *telescope,* and coincidences between *pendulum* and *chronometer* observed.

flattening (of the earth).—The ratio of the difference between the equatorial and polar radii of the earth (major and minor semi-axes of the *spheroid*) and its equatorial radius (major semi-axis).

The flattening of the earth is the *ellipticity* of the *spheroid* and equals the *ellipticity of an ellipse* forming a meridional section of the *spheroid.* If a and b represent the major and minor semi-axes of the spheroid, and f is the flattening of the earth,

$$f=\frac{a-b}{a}.$$

The magnitude of the flattening is sometimes expressed by stating the numerical value of the reciprocal of the flattening, $\frac{a}{a-b}$. The flattening of the earth is also termed its compression.

flexure (pendulum).—The bending of a swinging *pendulum,* due to its lack of perfect rigidity.

The effect of the bending of a swinging *pendulum* due to its lack of perfect rigidity is to increase the period of vibration; that is, the flexible *pendulum* swings less rapidly than the rigid *pendulum.* The effect may be regarded either as due to the change in the moment arm of the various parts of the *pendulum* as it bends, or to energy storage in the bent *pendulum.*

The effect of the bending is to increase the period of vibration in both positions of the reversible *pendulum.* The sign of the resulting correction depends upon the exact geometrical construction of the *pendulum.* In relative determinations of gravity, account is not taken of the bending of the *pendulum,* but the term flexure is applied to the movement of the pendulum support, communicated to it by the swinging *pendulum.* See *flexure* (*pendulum support*).

flexure (pendulum support).—The forced movement of a pendulum support caused by the motion of the swinging *pendulum.*

The horizontal component of the force acting on the *knife edge* of a swinging *pendulum* causes the support to move in unison with the *pendulum,* thereby affecting the period of its *vibration.* Though very small, this *flexure* of the *pendulum* support necessitates the application of a *correction* to the *pendulum* period. There are several methods of ascertaining the amount of the *flexure,* of which two have been used in the work of the U. S. Coast and Geodetic Survey: the static method, by which the displacement caused by a known force is measured in linear units; and the optical method, using an *interferometer* to measure the actual displacement in terms of wave lengths of light. The term *flexure* is also applied to the bending of a swinging *pendulum,* account of which is taken in absolute determinations of the value of *gravity.* See *flexure* (*pendulum*).

focal length.—See under *focus* (*lens*).

focus (lens).—The point to which rays of light converge after passing through a lens system.

If the incident rays are parallel they come to a focus in the (primary) focal plane, and it is the distance along the axis of the lens system from the *second nodal point* of the lens to this plane which is called the focal length of the lens.

focus: sidereal.—The position of the principal focal plane of a lens system.

A camera or *telescope* is said to be in sidereal focus when incident rays from a great distance come to a *focus* in the plane of the photographic plate or of the *reticle.* Sometimes called the solar focus.

foot.—A *unit of length* equal to 12 *inches* exactly or to one-third of a *yard.*

foot plate (leveling).—A disk of metal used as a rod support at a *turning point* in *leveling operations.*

One type of foot plate had a rod point, such as a knob, attached to its upper side, while the lower side had small projections to prevent lateral motion of the foot plate when it was pressed against the ground. In another type, the upper surface of the foot plate was dish-shaped (concave). The use of foot plates as rod supports has been superseded by metal pins driven firmly into the ground.

foresight (general).—A sight on a new survey point, made in connection with its determination; or a sight on a previously established point, to close a circuit.

foresight (leveling).—A reading on a rod held on a point whose *elevation* it is intended to determine.

The foresight may be on a point of unknown *elevation,* or on a previously established *bench mark* to close a circuit. Foresights are sometimes called minus sights, as they are usually subtracted from the *elevation* of the instrument (H. I.) to obtain the *elevation* of the rod point. This is not always the case, as the new point may be at a greater *elevation* than the instrument, as a mark on a wall, or a point on the roof of a tunnel. The term foresight is preferred over minus sight.

foresight: extra (leveling).—The rod reading made at an *instrument station* in a *line of levels* and on a *leveling rod* standing on a *bench mark* or other point NOT in the continuous *line of levels.*

In spirit leveling there may be one or more extra foresights from a single *instrument station* or *set-up,* but there can be only one *backsight* and one *foresight* from any one *instrument station.*

forward azimuth.—See under *azimuth: geodetic.*

fourth-order traverse.—See under *traverse.*

fractional section (United States public-land surveys).—See under *section.*

fractional township (United States public-land surveys).—See under *township.*

free-air anomaly.—See under *anomaly.*

free-swinging pendulum.—See under *pendulum.*

French legal meter.—See under *meter.*

fundamental tables: deformation of the geoid and its effect on gravity.—Tables giving the deformation of the *geoid* and its effect on *gravity,* computed for masses of unit density extending to various distances above and below the surface of the *geoid.*

Fundamental tables serve as the basis for the preparation of special tables corresponding to particular assumptions respecting density, *isostasy*, etc. Several such fundamental tables have been prepared, each designed for a particular effect:

a). Fundamental Tables for Reducing Observed Gravity Values, by G. Cassinis and P. Dore (1937) determine the direct effect on *gravity* of masses of unit density extending to various distances above and below *sea level*. This direct effect is known as the Hayford effect; it neglects the differences of *elevation* between the *spheroid* and the *geoid*.

b). Fundamental Tables: The Deformation of the Geoid and its Effect on Gravity (U. S. Coast and Geodetic Survey Special Publication No. 199) by Walter D. Lambert and Frederic W. Darling (1936) determine the indirect effect of masses of unit density extending to various distances above and below *sea level;* this indirect effect is known as the *Bowie effect;* it takes into account the differences of *elevation* between the *spheroid* and the *geoid*.

c). Fundamental Tables for the Deflection of the Vertical (a U. S. Coast and Geodetic Survey publication now in press) by Frederic W. Darling give the horizontal effect (*deflection of the vertical*) of masses of unit density extending to various distances above and below *sea level*. For mathematical development and formulas, see Bulletin Geodesique: No. 57, January-February-March 1938 and New Series No. 9, July 1, 1948.

G

gal.—An *acceleration* of one centimeter per second per second.

The gal is the unit of *acceleration* of the c. g. s. system of units and, since about 1930, has been used by the U. S. Coast and Geodetic Survey in stating values of *gravity*. Before that date, the *dyne*, a unit of force, was used for the purpose. (The real unit was dyne/gram.) While the term gal has not been formally adopted by the International Association of Geodesy, it is now in general use in those countries where *gravity* observations are being made. It is expressed symbolically by cm/sec.^{2} A milligal is 0.001 gal. The term gal is not an abbreviation; it was invented to honor the memory of Galileo.

geländereduktion.—See *terrain correction.*

general base map.—See under *map: Topographic Map of the United States.*

general highway and transportation map.—See under *map: Transportation Map of the United States.*

general topographic map.—See under *map: Topographic Map of the United States.*

general utility map.—See under *map: Topographic map of the United States.*

general utility map of the United States.—See under *map: Topographic Map of the United States.*

general utility topographic map of the United States.—See under *map*: *Topographic Map of the United States.*

geocentric coordinates.—See under *coordinates.*

geocentric latitude.—See under *latitude.*

geocentric parallax.—See under *parallax: diurnal parallax.*

geodesic.—A *geodesic line.*

geodesic line.—A line of shortest distance between any two points on any mathematically defined surface. Also termed a *geodesic.*

While, in *geodetic* work, a geodesic line on a *spheroid* of reference is sometimes called a *geodetic line, geodetic* as applied to lines is going out of use, and *geodesic* is preferred.

geodesy.—The science which treats mathematically of the figure and size of the earth.

The term geodesy is often used to include both the science which must depend upon determinations of the figure and size of the earth from direct measurements made on its surface (*triangulation, leveling,* astronomic and *gravity* determinations), and the art which utilizes the scientific determinations in a practical way and is usually termed *geodetic surveying* or geodetic engineering.

geodetic.—Signifying basic relationship to the earth in which the curvature of its *sea-level* surface is taken into account.

In horizontal *control surveys,* the term geodetic is applied to operations and results based on the *ellipsoid of revolution* (*spheroid*). In *spirit leveling, elevations* are referred—or intended to be referred—to the *geoid*. In determining *elevations* by *vertical angles, elevations* are referred (in the ideal case where *deflections of the vertical* are available) to the *ellipsoid of revolution.* In ordinary practice, since *deflections of the vertical* are not usually available, the results are a mixture, and are referred to neither surface exactly.

geodetic azimuth.—See under *azimuth.*

geodetic azimuth mark.—See under *azimuth mark.*

geodetic control.—See under *control.*

geodetic coordinates.—See under *coordinates.*

geodetic datum.—See under *datum.*

geodetic equator.—See under *equator.*

geodetic latitude.—See under *latitude.*

geodetic leveling.—See under *leveling.*

geodetic line.—The shortest-distance line between any two given points on the surface of the *spheroid.*

Sometimes called a *geodesic line,* a geodetic line is a line of double curvature, and usually lies between the two *normal section lines* which the two points determine. If the two terminal points are in nearly the same latitude, the geodetic line may cross one of the normal section lines. It should be noted that, except along the equator and along the meridians, the geodetic line is not a plane curve and cannot be sighted over directly. However, for conventional triangulation the lengths and directions of geodetic lines differ inappreciably from corresponding pairs of normal section lines.

geodetic longitude.—See under *longitude.*

geodetic meridian.—See under *meridian.*

geodetic parallel.—See under *parallel.*

geodetic position.—See under *position.*

geodetic survey.—See under *survey.*

geodetic surveying.—See under *surveying.*

geographic; geographical.—Signifying basic relationship to the earth considered as a globe-shaped body.

The term geographic is applied alike to data based on the *geoid* and on a *spheroid.* In *geodetic surveys* in this country, coordinated data consisting of *latitudes, longitudes, azimuths,* and lengths of lines, are recorded and published under the general title of *geographic positions.* For particular usage of the two forms, geographic, geographical, see *-ic; -ical.*

geographic center.—The geographic center of an area on the earth has been defined as that point on which the area would balance if it were a plate of uniform thickness. In other words, it is the center of gravity of that plate.

The geographic center of continental United States (exclusive of Alaska) is in the eastern part of Smith County, Kansas, latitude 39°50′, longitude 98°35′. The geographic centers of the various States are given in the U. S. Geological Survey Bulletin 817, "Boundaries, Areas, Geographic Centers, and Altitudes of the United States and the Several States."

geographic coordinates.—See under *coordinates.*

geographic exploration survey.—See under *survey.*

geographic latitude.—See under *latitude.*

geographic longitude.—See under *longitude.*

geographic meridian.—See under *meridian.*

geographic parallel.—See under *parallel.*

geographic position.—See under *position.*

geographic survey.—See under *survey.*

geographical mile.—See under *mile: nautical.*

geoid.—The figure of the earth considered as a *mean sea-level* surface extended continuously through the continents.

The actual geoid is an *equipotential surface* to which, at every point, the plumb line (direction in which *gravity* acts) is perpendicular. It is the geoid which is obtained from observed *deflections of the vertical,* and is the surface of reference for astronomical observations and for geodetic leveling. Theoretical geoids obtained with computed values of *deflections of the vertical* include the *compensated geoid* and the *isostatic geoid.*

geoid: compensated.—A theoretical *geoid* derived from the actual *geoid* by the application of computed values of the *deflection of the vertical* which depend upon the *topography* and *isostatic compensation.*

The data upon which to base the effect of *topography* and *isostatic compensation* are obtained from readings of *topographic maps;* the effect is then computed in accordance with the assumptions made with respect to *isostasy,* that is, the depth and distribution of *isostatic compensation.* A method for accomplishing this was developed by Hayford, who accepted the *Pratt theory of isostasy.* If the theory and assumptions with respect to *isostasy* were exact and there were no *anomalies,* the compensated geoid would agree with the *spheroid* of reference.

geoid: isostatic.—An ideal *geoid* derived from the *spheroid* of reference by the application of computed values of the *deflection of the vertical* which depend upon the *topography* and *isostatic compensation.*

The computed values of the *deflection of the vertical* used in obtaining the isostatic geoid are similar to those for the *compensated geoid,* but of opposite signs. If the theory and assumptions with regard to *isostasy* were exact and there were no *anomalies,* the isostatic geoid would agree with the actual *geoid.*

geoid contour.—A line on the surface of the *geoid* of constant *elevation* with reference to the surface of the *spheroid* of reference.

Geoid contours represent differences in *elevation* between the *geoid* and the *spheroid.* Geoid contours depend on the surface of reference. The same *geoid* referred to different surfaces of reference will give different sets of geoid contours.

geologic survey.—See under *survey.*

geological survey.—See under *survey.*

geometric latitude.—See under *latitude: parametric (geometric or reduced) latitude.*

geometric map projection.—See under *map projection: perspective map projection.*

geophysics.—The science of the earth with respect to its structure, composition, and development.

Geophysics is a branch of experimental physics dealing with the earth, including its atmosphere and hydrosphere. It includes the sciences of dynamical geology and physical geography, and makes use of *geodesy,* geology, seismology, meteorology, oceanography, magnetism, and other earth sciences in collecting and interpreting earth data. Geophysical methods have been applied successfully to the identification of underground structures in the earth and to the search for structures of a particular type, as for example, those associated with oil-bearing sands.

gisement.—The *angle* between the *grid meridian* and the *geographic meridian.*

The term "gisement" is used in connection with the *military grid,* sometimes being called the "declination of grid north," and reckoned east and west from geographic north. Gisement is not used with the *State coordinate systems,* where the corresponding *angle* on the *transverse Mercator grid* is the convergence of the local with the central geographic meridian, and is designated by $\Delta\alpha$. In the *Lambert grid,* the angle is known as the "mapping angle," designated by θ, and called the "theta angle."

globe.—A spherical body. See under *globe: celestial globe; terrestrial globe.*

globe: celestial.—A sphere, on the outer surface of which, by means of symbols and reference lines, stars are shown in positions and magnitudes determined from the earth.

Usually shown also are the outlines of the constellations or star groups. Essentially, a celestial globe is a sphere to the outside of which star maps have been fitted.

globe: terrestrial.—A sphere, on the outer surface of which, by means of symbols and reference lines, the features of the surface of the earth are shown in relative positions.

Essentially, a terrestrial globe is a sphere to the outside of which *maps* of the surface of the earth have been fitted. See *globe, gore.*

globe, gore.—A lune-shaped *map* which may be fitted to the surface of a *globe* with a negligible amount of distortion.

A *terrestrial* or *celestial globe* may be constructed by covering a sphere with a complete set of *maps,* each having the form of a lune of such narrow width that it may be mounted on the sphere without appreciable stretching or shrinking.

globe, horizon ring.—A graduated ring fitted to a *globe* in such manner that its plane contains the center of the *globe,* and it can be adjusted into the plane of the *horizon* for any point on the *globe.*

The horizon ring is used in scaling *angles* around the *horizon* for a given point.

globe, meridian ring.—A graduated ring fitted to a *globe* in such manner that its plane contains the poles of the *globe,* and it can be adjusted into the plane of any given *meridian* on the globe.

A meridian ring is used for scaling *angles* along a *meridian.*

globe, time dial.—An adjustable circular disk, graduated in time units (0 hours to 24 hours), attached to a *globe* with its center coincident with a pole of the *globe.*

A time dial can be adjusted to show the relation between local *time* and the *time* of any other *meridian.*

globular map projection.—See under *map projection.*

gnomonic map projection.—See under *map projection.*

Golofnin Bay datum.—See under *datum.*

goniasmometre.—A compact surveying instrument for observing *horizontal angles* and *bearings,* and consisting of a vertical cylinder divided horizontally into two parts: the lower edge of the upper part is graduated in circular measure and revolves on the lower part, which carries a *vernier* on its upper edge. A magnetic needle is centered in the upper part, which is also provided with slits or a *telescope* for sighting purposes.

An instrument used in France, and serving the same purposes as a *surveyor's compass* or *transit.*

goniometer.—In *surveying,* an instrument for measuring *horizontal angles* or *directions,* such as a *theodolite.*

In general, goniometer is a term applied to various instruments for measuring *angles,* as in crystallography, anthropology, etc.

grad (grade).—An *angle* at the center of a circle, subtended by one four-hundredth part of its circumference. See *angle (general).*

grade.—Rate of slope or degree of inclination. See *gradient.* Example: a 1 to 100 grade; a 1-percent grade.

grade correction, tape.—See under *tape.*

gradient (topography).—A *grade* expressed as a tangent of the *angle of inclination.*

Example: a 1 to 100 *grade* is a gradient of .01.

gradienter.—An attachment to an engineer's *transit* with which an *angle of inclination* is measured in terms of the tangent of the *angle* instead of in degrees and minutes.

A gradienter may be used as a telemeter in observing horizontal distances.

graduation.—See under *division.*

graticule.—A network of lines representing geographic parallels and meridians forming a *map projection.*

Gravatt level.—See under *leveling instrument.*

Gravatt leveling rod.—See under *leveling rod.*

gravimeter (geodesy).—A weighing device or instrument of sufficient sensitivity to register variations in the weight of a constant mass when the mass is moved from place to place on the earth and thereby subjected to the influence of *gravity* at those places.

A typical gravimeter consists of a spring which changes in length under the load of a constant mass as the instrument is moved from place to place and thereby affected by changes in *gravity.* Gravimeters are employed in determining differences in the intensity of *gravity* between an initial or base station at which the value of gravity is known or assumed and at nearby points for which values of *gravity* are desired. Gravimeters are also termed gravity meters.

gravitation.—In general, the mutual attraction between masses of matter (bodies). In *geodesy,* the mutual attraction between the earth and bodies on or near its surface.

Gravitation is the component of *gravity* which acts towards the earth.

gravitation: constant of.—The Newtonian constant of gravitation, being the constant G in the formula $F=G\frac{m_1m_2}{r^2}$, where F is the attractive force between two bodies of masses m_1 and m_2 expressed in grams, r is the distance in centimeters between the two bodies, and the time unit is the *mean solar second.*

Various values of the constant of gravitation which have become available in recent years are as follows:

(a) $(6.670 \pm 0.005) \times 10^{-8}$ c. g. s. units

This is by Paul R. Heyl, reported in the Bureau of Standards Journal of Research, December 1930. It is the value used in U. S. Coast and Geodetic Survey Special Publication No. 199, Tables for Determining the Form of the Geoid and its Indirect Effect on Gravity, 1936.

(b) $(6.664 \pm 0.002) \times 10^{-8}$ dyne. cm^2.g^{-2}

This is the value adopted in the Smithsonian Physical Tables, edition of 1934.

(c) $(6.673 \pm 0.003) \times 10^{-8}$ cm.^{3}.g^{-1}.sec^{-2}

This is the most recent determination by Heyl, reported in the Journal of Research of the National Bureau of Standards, July 1942. It agrees so well with the 1930 result that the conclusion is reached that, for the method used (the *torsion balance*), the limit of possibilities for increasing the *precision* has been reached.

gravity.—That force which tends to pull bodies towards the earth: that is, to give bodies weight.

Gravity is the resultant of two opposing forces: *gravitation* and the *centrifugal force* due to the *rotation* of the earth. See *gravity: intensity of gravity.*

gravity: Clairaut's Theorem.—A theorem connecting the *ellipticity* of the earth with the coefficient of $\sin^2 \phi$ of the *formula for theoretical gravity* or with the difference between *gravity* at the poles and *gravity* at the equator.

This means that every *gravity* formula implies a particular value for the *ellipticity* of the *spheroid* of reference. Let g_p and g_e represent *gravity* at the poles and at the *equator,* respectively, and let the *formula for theoretical gravity* be written $g=g_e\ (1+\beta \sin^2 \phi + \text{smaller terms})$. Also, assume values for ϕ of 90° and 0°, and neglect the smaller terms.

Then $\beta=\frac{g_p-g_e}{g_e}$ and Clairaut's Theorem is written as follows: $\beta=\frac{g_p-g_e}{g_e}=\frac{5m}{2}-f+\text{smaller terms}$, f being the *ellipticity* and m the ratio of *centrifugal force* at the *equator* to *gravity* at the equator, or $\frac{\omega^2 a,}{g_e}$, where ω is the angular velocity of *rotation* of the earth in *radians* per second and a is the semi-major axis (equatorial radius) of the *spheroid.*

As originally stated by Clairaut, the smaller terms were neglected. For an exact *ellipsoid* we have, to a high degree of *accuracy,* the formula

$$\beta=\frac{5m}{2}-f+\frac{17}{14}mf.$$

Upon first consideration, Clairaut's Theorem may seem paradoxical: an increase in the *flattening of the earth, f,* apparently diminishing the increase in *gravity* from *equator* to pole. This is explained by the fact that the theorem applies, not to values on the physical surface of the *ellipsoid,* but to values on a *level surface*—any *level surface* lying outside all attracting matter. In the case of the *geoid,* or *sea-level surface,* which has a comparatively small amount of matter outside it, Clairaut's Theorem is applied by means of certain conventions in the reduction to that surface of values of *gravity* observed outside the physical surface of the earth. If the *ellipticity* of the physical surface of the earth is conceived to change but that surface still remain a *level surface,* the distribution of density within the *level surface* must change accordingly, and Clairaut's Theorem will still hold.

The increase in *gravity* from *equator* to poles according to the *International Gravity Formula* has been analyzed by Hammer as follows:

Centrifugal acceleration	$+3.39$ *gals*	
Free-air effect	$+6.63$	net $+1.78$ *gals*
Mass-shape effect	-4.85	
Total change, *equator* to poles	$+5.17$ *gals*	

The free-air effect is due to the difference of the distances from the center of the earth of the attracted particle when at a pole and when at the *equator,* the attractions being in the inverse ratio of the squares of the respective distances. But the inverse-square law is correct only when the dimensions of the attracting body are negligible in comparison with the distances involved, a condition that does not hold in the present instance. Near its surface, the terrestrial *ellipsoid* does not attract in exact accordance with the inverse-square law. The *correction* which is applied to the free-air effect to care for this condition is the mass-shape effect given in this and the next paragraph.

If the terrestrial *ellipsoid* were homogeneous, the free-air effect and the mass-shape effect would be more nearly equal, and the increase in *gravity* from *equator* to poles would be composed as follows:

Centrifugal acceleration $= mg =$	$+3.39$ *gals*	
Free-air effect $= 2\,fg = 5\,mg/2 =$	$+8.48$	net $+0.85$ *gal*
Mass-shape effect $= 9\,\mathrm{mg}/4 =$	-7.63	
Total change, *equator* to poles $= 5\,mg/4 =$	$+4.24$ *gals*	

gravity: direction of; direction of the force of gravity.—The direction indicated by a plumb line.

It is perpendicular to the surface of the *geoid.* See *nadir, zenith.*

gravity: formula for theoretical.—A formula expressing *gravity* on the *spheroid* of reference in terms of *geographic position,* it being assumed that the *spheroid* of reference is a *level surface.*

If the *spheroid* of reference is one of *revolution,* all *meridians* are alike and the *longitude* can not appear in the formula (see *gravity formula, longitude term*); furthermore, since the *spheroid* is symmetrical with respect to the *equator,* only even functions of the *latitude* can appear in the formula. *Gravity* increases toward the poles, both because the centrifugal *acceleration* due to the earth's *rotation* decreases as a pole is approached and because the mass-attraction of the *spheroid* is greater at the poles than at the equator.

Gravity on the surface of a *spheroid of revolution* is expressed by either of the following commonly used formulas:

$$g = g_e\ (1+\beta\ \sin^2\phi + \text{smaller terms})$$

$$g = g_{45}\ (1-\gamma\cos 2\phi + \text{smaller terms})$$

where g is *gravity* on the surface of the *spheroid* in *geographic latitude* ϕ, g_e is *gravity* at the *equator,* g_{45} is gravity at *latitude* 45°, and β and γ are constants. For numerical values of g_e and constants appearing in the first formula above, see *gravity: Clairaut's Theorem; Helmert's gravity formula of 1901; International Gravity Formula;* and *gravity formula, longitude term.*

In the approximate formulas, the smaller terms are often omitted. If this is done, our surface of reference is a *spheroid of revolution* lying, at latitude 45°, 19 *meters* above an exact *ellipsoid* of *revolution* having the same axes as the spheroid.

gravity: Helmert's gravity formula of 1901.—A development of the *formula* for *theoretical gravity* utilizing the *gravity* observations available at the time (1901), but not fitted to any preassigned value of the earth's *ellipticity.*

Referred to the Potsdam value of absolute *gravity,* Helmert's gravity formula of 1901 is as follows:
$g = 978.030\ (1+0.005302\ \sin^2\phi - 0.000007\ \sin^2 2\phi)$.

While the *International Gravity Formula* was based on a preassigned value of the earth's *ellipticity,* the above formula was not, but it was used in determining a value for that *ellipticity,* namely 1/298.2. The coefficients $g_e = 978.030$ and $\beta = 0.005302$ were determined from the observations.

Attempts made to determine the coefficient of $\sin^2 2\phi$ in the same way were abandoned because of the large *probable errors* obtained, and the value 0.000007 was assumed from theoretical considerations based on the work of Darwin and Wiechert on a self-attracting rotating mass of the size of the earth. Such a mass, unless homogeneous, will not take on the form of an exact *ellipsoid,* but will be depressed in middle *latitude* below an exact *ellipsoid* having the same major and minor axes. The amount of this depression will depend upon the law used for the variation of density with respect to distance from the center of the earth. Darwin and Wiechert assumed different laws, but both found a depression of 3 meters in *latitude* 45°. This depression corresponds to the difference between 0.0000059 and 0.000007, the corresponding coefficients in the International and Helmert formulas.

gravity: Helmert's gravity formula of 1915.—A development of the *formula for theoretical gravity,* based on a *triaxial ellipsoid,* and including a *longitude term.* See *gravity formula, longitude term.*

gravity: intensity of.—The force with which *gravity* acts, expressed in suitable units.

Intensity of gravity may be expressed either as a force (*force of gravity*) expressed in *dynes;* or as an *acceleration* expressed as centimeters per second per second. In the geodetic work of the U. S. Coast and Geodetic Survey, before 1930, intensity of gravity was stated as a force, in *dynes;* since then, it has been stated as an *acceleration,* in *gals.*

gravity: International Gravity Formula.—A development of the *formula for theoretical gravity,* based on the assumptions that the *spheroid* of reference is an exact *ellipsoid* of *revolution* having the dimensions of the International Ellipsoid of Reference (Madrid, 1924), rotating about its minor axis once in a *sidereal day;* that the surface of the *ellipsoid* is a *level surface;* and that *gravity* at the *equator* equals 978.049 *gals.*

Omitting terms which are so small as to be negligible for practical purposes, the International Gravity Formula is as follows:
$g = 978.049\ (1 + 0.0052884\ \sin^2\phi - 0.0000059\ \sin^2 2\phi)$.

The International Gravity Formula was adopted by the International Association of Geodesy at the Stockholm meeting, 1930. The purpose was not primarily to represent the *gravity* observations then available, although the $g_e = 978.049$ *gals* was based on those observations, but rather it was intended to put the determination of the *figure of the earth* from *gravity* data on the same basis as the determination of its figure from *deflections of the vertical.* This formula is based on the Potsdam value for absolute *gravity.*

gravity anomaly.—See under *anomaly: free-air anomaly.*

gravity apparatus: Brown gravity apparatus.—See under *Brown gravity apparatus.*

gravity formula.—See under *gravity: formula for theoretical gravity.*

gravity formula, longitude term.—A suggested additional term for the *formula for theoretical gravity,* derived from the *longitude* on a *spheroid* of reference having three unequal axes.

With this additional term, the *formula for theoretical gravity* would be as follows:

$$g=g_e\ [1+\beta \sin^2\phi - \beta_1 \sin^2 2\phi + c \cos^2 \phi \cos 2\ (\lambda-\lambda_0)]$$

where g_e is the mean value of *gravity* at the *equator,* and c, like β and β_1 is a dimensionless number.

As always, the *formula for theoretical gravity* implies a *spheroid* of definite shape. When a longitude term of the type indicated above is included, the *spheroid* is no longer an *ellipsoid* of *revolution,* but an *ellipsoid* having three unequal axes (*triaxial ellipsoid*), the shortest of which is its polar axis (corresponding to the axis of *rotation* of the earth), while the two longer ones lie in the plane of the *equator.* The longer equatorial axis lies in the *meridian* plane λ_0 and $\lambda_0 \pm 180°$; the shorter in the *meridian* plane $\lambda_0 \pm 90°$. These two equatorial axes or diameters differ by $\pm 4\ ca$; the difference of each semiaxis from a mean is $\pm ca$, and from each other $\pm 2\ ca$.

The form of the longitude term is not arbitrary. In theoretical investigations on the attraction of nearly spherical bodies, it is convenient to work with spherical harmonics. These are mathematical expressions involving *latitude* and *longitude.* They are of different degrees, the harmonics increasing in complexity as their degree increases.

The additional term involving *longitude* in the *gravity formula* is the spherical harmonic term of lowest degree not already present, though in partly disguised form, in the ordinary *gravity formula* and not excluded from it for sufficient theoretical reasons. (The excluded terms are proportional to $\cos \phi \cos \lambda$, $\cos \phi \sin \lambda$, $\sin \phi$, $\sin 2\phi \cos \lambda$, and $\sin 2\phi \sin \lambda$.) The idea of an *ellipsoid* of three unequal axes as a possible surface of reference is by no means new. There was, however, no continuing interest in the idea until Helmert and his assistant, Berroth, each published in 1915 his calculations on which were based *gravity formulas* containing a longitude term of the type given above. *Helmert's gravity formula of 1915* is as follows:

$$g=978.052\ [1+0.005285 \sin^2\phi - 0.000007 \sin^2 2\phi + 0.000018 \cos^2 \phi \cos 2\ (\lambda+17°)].$$

In the formula, east *longitude,* λ, is positive. The maximum equatorial semiaxis (radius) is 115 *meters* longer than the mean equatorial semiaxis (radius) and 230 *meters* longer than the minimum. The maximum axis lies in the *meridian* plane of $\lambda=17°$ west and 163° east of Greenwich. The minimum axis lies in the meridian plane $\lambda=73°$ east and 107° west of Greenwich. The mean *ellipticity* of a *meridian* is 1/296.7.

Since 1915, various discussions of the accumulating mass of *gravity* data have tended to confirm the reality of a longitude term of the same general size and character as that proposed by Helmert. Studies of the *deflections of the vertical* have likewise tended to confirm that general result. This does not mean, however, that a *triaxial ellipsoid* is likely to find favor as a surface of reference for *triangulation* and *deflections of the vertical.* The mathematical complications of such a surface are too great. And since one of the purposes of studying *gravity* observations is to determine the *figure of the earth* and to compare the results with those obtained from *deflections of the vertical,* it is convenient to use the same surface of reference for both and therefore omit the longitude term from the *formula for theoretical gravity.*

It would appear from the most recent study of the longitude term (Jeffreys, 1942–3), that a longitude term of the type derived by Helmert and others is not the only spherical harmonic term needed to represent existing widespread variations of *gravity.* Jeffreys confirmed the existence of a longitude term of the general character and magnitude just discussed, and tentatively determined a few other spherical harmonic terms of low degree; he found, however, that the number and distribution of existing *gravity* observations were inadequate for a satisfactory determination of those terms.

The argument against including a longitude term in the *formula for theoretical gravity* applies with even greater force to the additional terms found by Jeffreys. The mathematical complexities attending the use of a figure of reference which these additional terms would imply are too great for practical problems. The significance of the longitude term and of the other terms proposed by Jeffreys, which are also longitude terms of a sort, is mainly geophysical. These terms represent widespread systematic variations in *gravity* and in *deflections of the vertical.* They represent also general widespread departures from isostatic equilibrium and therefore imply the existence of long-continuing stresses in the earth's crust. The problems they represent are geological and geophysical: problems of how the configuration of the continents and oceans on the known geological structures are related to the departures from equilibrium which are implied by the additional terms.

While it is of great geological and geophysical interest to improve the *formula for theoretical gravity* by the addition of spherical harmonic terms of low degree, the question of how far to continue the addition of such terms must be left to the future. It seems probable, however, that consideration of the usefulness of this procedure will limit such additions to spherical harmonic terms of rather moderate degree.

gravity meter.—See under *gravimeter.*

gravity station.—A *station* at which observations are made to determine the value of *gravity.*

great circle.—The line of intersection of the surface of a sphere and any plane which passes through the center of the sphere.

The shortest distance between any two points on a sphere is along the *arc* of a great circle connecting the two points. Great circles on the *celestial sphere* which are given particular designations are the *equator,* the *ecliptic, meridians, hour circles, prime verticals, colures,* and *horizons.* The shortest distance on an *ellipsoid* of *revolution* is a *geodetic line* (a *geodesic*), which is not a plane curve except for the *equator* (a circle) and the *meridians* (ellipses). In *cartography,* the *gnomonic* is the only *map projection* on which a great circle is represented, in all instances, as a straight line. The *gnomonic projection* for a *map* of the United States computed by the U. S. Coast and Geodetic Survey is based on a sphere having the same volume as the *ellipsoid of reference* (*Clarke spheroid of 1866*).

great circle line.—In *land surveying,* the line of intersection of the surface of the earth and the plane of a *great circle* of the *celestial sphere.*

The term great circle line is proposed for use in descriptions of methods for establishing *parallels* of *latitude* (*standard parallels* and correction lines) in *surveys* of the public lands of the United States. Since the earth is not a sphere, a great circle line is but an approximation to a *great circle* and, due to the *deflection of the vertical,* will not, as surveyed, lie in a single plane.

Greenwich Civil Time.—See under *time.*

Greenwich Meridian.—See under *meridian.*

Greenwich Sidereal Time.—See under *time.*

Gregorian Calendar.—See under *calendar.*

grid.—A network composed of two sets of uniformly-spaced straight lines intersecting in right angles.

In surveying and mapping, the term grid is applied to a *plane-rectangular coordinate system* imposed upon a *geographic coordinate* system. See *State coordinate systems.*

grid: military.—See under *military grid.*

grid azimuth.—See under *azimuth.*

grid bearing.—See under *bearing.*

grid coordinates.—See under *coordinates.*

grid distance.—See under *grid length.*

grid inverse.—The computation of *grid length* and *grid azimuths* from *grid coordinates.*

grid length.—The distance between two points as obtained by computation from the *plane rectangular coordinates* of the points.

In the *State coordinate systems,* a grid length differs from a *geodetic length* by the amount of a correction based on the *scale factor* for the given line.

grid meridian.—See under *meridian.*

guide meridian (United States public-land surveys).—See under *meridian.*

Gunter's chain.—See under *chain.*

H

hack (United States public-land surveys).*—A horizontal V-shaped notch, cut into the trunk of a tree at about breast height.

hack chronometer.—See under *chronometer.*

half section (United States public-land surveys).—See under *section.*

hand level.—See under *leveling instrument.*

hanging level.—See under *level.*

Harrebow-Talcott method, latitude determination.—See under *zenith-telescope method, latitude determination.*

Hassler base-line measuring apparatus.—See under *base apparatus.*

Hayford-Bowie method of isostatic reduction.—A method of computing the effect of *topography* and *isostatic compensation* on *gravity* by which the effect of *topography* is computed directly and then corrected for the effect of *isostatic compensation.*

The mechanics of this method involve the use of the *Hayford gravity templates.*

Hayford-Bullard (or Bullard) method of isostatic reduction.—A method by which the topographic effect of an infinite slab of density 2.67 and a thickness equal to the *elevation* of the *gravity station* is first computed, and then corrected for curvature of the *sea-level* surface and for difference of *elevation* between the station and the *topography.*

The first step in the computation gives the ordinary Bouguer correction for *topography:* the second step takes account of the departure of the actual *topography* from a smooth cap or plate of thickness equal to the *elevation* of the station.

The application of the *curvature correction* for the lettered zones reduces the effect of the plane slab to that of a plateau or cap extending to the outer limits of those zones and curved to the mathematical surface of the earth. Account is taken of the *topography* by applying a *correction* for the deviations of the *topography* from the surface of the plateau or cap.

This method was devised by E. C. Bullard as a substitute for the *Hayford-Bowie method of isostatic reduction* for the lettered zones of the *Hayford gravity templates;* it is described in U. S. Coast and Geodetic Survey Special Publication No. 232, Pendulum Gravity Measurements and Isostatic Reductions, by Clarence H. Swick, 1942.

Hayford deflection templates.—See under *template* (*templet*).

Hayford effect (gravity).—See under *fundamental tables: deformation of the geoid and its effect on gravity* (*a*).

Hayford gravity anomalies; Hayford anomalies.—See under *anomaly: isostatic anomaly.*

Hayford gravity templates.—See under *template* (*templet*).

height of instrument.—1. In *spirit leveling:* the height of the *line of sight* of a *leveling instrument* above the adopted *datum.* 2. *Stadia surveying:* the height of the center of the *telescope* (*horizontal axis*) of *transit* or *telescopic alidade* above the ground or station mark. 3. *Trigonometrical leveling:* the height of the center of the *theodolite* (*horizontal axis*) above the ground or station mark.

heliocentric parallax.—See under *parallax: annual.*

heliotrope.—An instrument composed of one or more plane mirrors, so mounted and arranged that a beam of sunlight may be reflected by it in any desired direction.

Placed over a survey station, a heliotrope is used to direct a beam of sunlight toward a distant survey station, where it can be observed with a *theodolite.* It provides an excellent target in observing *horizontal directions,* such targets having been observed on at distances of over 150 miles and approaching 200 miles. There are several types of heliotropes, carrying various names, but differing only in details of mechanical arrangement and construction. Heliotropes are sometimes used for sending messages, using the Morse code conveyed by short and long flashes of light; when especially equipped for the purpose, such instruments are termed heliographs.

hemispherical map.—See under *map.*

high-water line.—The place on the bank or shore up to which the presence and action of water is so usual and long continued as to impress on the *bed of the stream* a character distinct from that of the *banks* with respect to vegetation and the nature of the soil.

The high-water line is the *boundary line* between the *bed* and the *bank of a stream*. The mean high-water line usually determines the boundary of the land of the riparian proprietor.

highway map.—See under *map.*

homalographic map projection.—See under *map projection.*

horizon.—A *great circle* on the *celestial sphere* whose plane is perpendicular to the direction of the plumb line.

The direction of the plumb line is sometimes referred to as the *direction of gravity*, and *direction of the force of gravity.* In *surveying,* the plane of the horizon is determined by means of a *spirit level.* See *horizon: apparent (visible) horizon.*

horizon: apparent (visible).—The irregular line along which rays from the point of observation are tangent to the surface of the earth.

The apparent horizon is the line where the visible surface of the earth appears to meet the sky. When the apparent horizon is formed by the surface of a body of water, it is sometimes used as a reference in observing *vertical angles.* See *horizon: dip of.*

horizon: artificial.—A device consisting of a plane reflecting surface which can be adjusted to coincide with the plane of the *horizon.*

One form of artificial horizon consists of a dish or trough filled with mercury, or an amalgam of mercury and tin, the upper surface of which is free, and assumes a horizontal position. Another form is a plane mirror of glass or other material, equipped with *spirit levels* and leveling screws so that it can be adjusted into the plane of the *horizon.*

In observing a celestial body such as a star or the sun, the *angle* is measured between the body as seen direct (with *transit* or *theodolite*) or reflected (in the horizon glass of a *sextant*) and its image as seen reflected in the artificial horizon: this is a *vertical angle,* and is double the *altitude* of the body.

horizon: closing the.—Measuring the last of a series of *horizontal angles* at a station required to make the series complete around the *horizon.*

At any station, the sum of all *horizontal angles* between adjacent lines should equal 360°. The amount by which the sum of the *observed angles* fails to equal 360° is the *error of closure.* This is distributed as a *correction* for closure among the *observed angles* to bring their sum to exactly 360°. The *error* and the *correction* are of opposite signs.

horizon: dip of.—The *vertical angle* between the plane of the *horizon* and a line tangent to the *apparent (visible) horizon.*

horizontal angle.—See under *angle.*

horizontal axis (theodolite or transit).—The axis about which the *telescope* of a *theodolite* or *transit* rotates when moved vertically.

It is the axis of *rotation* which is perpendicular to the *vertical axis* of the instrument and to the *axis of collimation* of the *telescope.* It should coincide with the line through the centers of the pivots which support the *telescope.* For an instrument in complete adjustment, it occupies a horizontal position, and when the *telescope* is rotated around it, the *axis of collimation* will define a *vertical plane:* deviations from this condition are measured with a *striding level* or with a *hanging level.*

horizontal direction.—A *direction* in a horizontal plane.

horizontal parallax.—See under *parallax: diurnal parallax.*

Horrebow-Talcott method, latitude determination.—See under *zenith-telescope method, latitude determination.*

hour angle.—The angle between the plane of the *hour circle* passing through a celestial body and the plane of the *celestial meridian.*

The hour angle is reckoned from the *meridian* (0 hours or 0°) westward through 24 hours (360°). It may be measured by the *angle* at the pole between an *hour circle* and the *meridian,* or by the *arc* of the *equator* which is intercepted by those circles.

hour circle.—Any *great circle* on the *celestial sphere* whose plane is perpendicular to the plane of the *celestial equator.*

Hour circles are also called circles of *declination.* The hour circle which contains the *zenith* is identical with the *celestial meridian.*

hygrometric.—Relating to the relative humidity or comparative amount of moisture in the atmosphere.

Since the atmosphere penetrates the pores or cells of material bodies in varying degrees depending upon the substances of which they are composed, the amount of moisture which it contains will affect the shapes and dimensions of certain instruments and equipment used in *surveying* and mapping. For this reason it is necessary to select materials which are not sensitive to hygrometric conditions for the construction of *leveling rods, plane-table* sheets, etc., and for the construction and printing of maps.

hypsographic feature.*—A *topographic feature* with *elevations* portrayed with reference to *sea level.*

hypsographic map.—See under *map.*

hypsography.*—*Topography* (*relief*), referred to a *sea-level datum.* The science or art of describing *elevations* (heights) of land surfaces with reference to *sea level.*

hypsometer.*—An instrument used in determining *elevations* of points on the earth's surface in relation to *sea level* by determining atmospheric pressure through observation of the boiling point (temperature) of water at each point. Also an instrument used in forestry for determining heights of trees.

Determination of heights of mountain summits by observing the boiling point of water on the summits has been much used by explorers.

The term hypsometer is also used to designate one who measures *elevations* of the land, or who constructs *maps* or models to show land *elevations.* This is a rare use.

hypsometric.—Relating to *elevation* above a *datum,* usually *sea level.*

In a limited sense, hypsometric relates to *elevations* above *sea level* determined with a *hypsometer.*

hypsometric feature.*—Disapproved. See under *hypsographic feature.*

hypsometric map.—See under *map: hypsographic.*

hypsometry.*—The art of determining, by any method, surface *elevations* on the earth with reference to *sea level.*

The determination of *elevations* by means of either *mercurial* or *aneroid barometers* is sometimes called barometric hypsometry.

hypsometry: barometric.—See under *hypsometry.*

I

-ic; -ical.—Suffixes employed to convert certain terms into adjective form, the two adjectives derived from the same root word sometimes having slightly different meanings, but often employed according to custom or personal preference, without differentiation.

While it seems desirable to employ the shorter suffix, -ic, for adjectives having a direct and intimate relationship to the root word, and the longer form, -ical, to form adjectives of more general and broader applicability, this cannot be adopted as an exact procedure, and it is proposed that custom be considered, and in the work of this Bureau, the forms be used as recommended below.

1. geodetic, geodetical; geodesic, geodesical.—Each of these terms has appeared in technical literature. The longer forms, geodetical, and geodesical, are now practically obsolete. Geodesic was at one time employed to designate a class of *spirit leveling,* but at present is largely limited to use in a mathematical sense. The following forms, defined elsewhere, are proposed for use in this Bureau: *geodetic azimuth; geodetic equator; geodetic latitude; geodetic longitude; geodetic meridian; geodetic parallel; geodetic position; geodetic survey;* and *geodesic line.*

2. astronomic; astronomical.—Astronomers have preferred and used the longer form, *astronomical.* Surveyors have employed both forms, often without apparent differentiation in meaning, though they have accepted the astronomers' use of *astronomical* for terms of general applicability. In the publications of this Bureau, with some exceptions, the shorter form, *astronomic,* has been employed in recent years where there is an intended contrast with the term *geodetic,* and it is believed that this practice should be adopted insofar as is compatible with established practice. The following terms, defined elsewhere, are accordingly proposed for use in this Bureau: *astronomic azimuth; astronomic coordinates; astronomic equator; astronomic latitude; astronomic longitude; astronomic meridian; astronomic parallel;* astronomic station; and *astronomical day; astronomical leveling;* astronomical methods; *astronomical refraction; astronomical theodolite; astronomical time; astronomical transit; astronomical traverse; astronomical triangle; astronomical unit.*

3. geographic; geographical.—The following forms are proposed for use in this Bureau: *geographic center; geographic equator; geographic latitude; geographic longitude; geographic meridian; geographic parallel; geographic position; geographic survey;* geographical boundary; *geographical mile;* geographical organization, survey, or society.

4. geologic; geological.—The following forms are proposed for use in this Bureau: geologic data; geologic investigation or survey; geological organization, survey, or society; geological era; geological time.

5. geophysical.—The form geophysical is used to the exclusion of any other that might be derived from the term *geophysics.*

iced-bar apparatus.—See under *base apparatus.*

identification posts.—Posts of wood or other suitable material, appropriately marked and inscribed, and placed near *survey stations* to aid in their recovery and identification.

Identification posts also serve as guard posts, calling attention to the stations which they serve, thus giving protection against their accidental destruction. Designated as supplemental posts for survey monuments, identification posts were recommended for use in a report of the Committee on Control adopted by the Federal Board of Surveys and Maps, March 14, 1939.

idle pendulum.—See under *pendulum.*

impersonal micrometer.—See under *micrometer: transit micrometer.*

inch.—A *unit of length* defined by the equivalence: 1 *meter* equals 39.37 inches exactly.

The relationship, 1 meter equals 39.37 inches, was established by Act of Congress, July 28, 1866, but was put into practical effect only after the United States had received copies of the *International Prototype Meter.* In 1893, the Superintendent of Standard Weights and Measures, with the approval of the Secretary of the Treasury, declared the *International Prototype Meter* to be the *standard* of length, with the customary unit, the *yard,* to be derived therefrom in accordance with the Act of July 28, 1866. The *yard* is accepted as equal to 36 inches exactly. See under *meter: National Prototype Meter.*

inclination of the horizontal axis.—The *vertical angle* between the *horizontal axis* of a surveying or astronomical instrument and the plane of the *horizon.*

Inclination of the horizontal axis is measured with a *striding level* or a *hanging level.* It causes the *axis of collimation* of a *theodolite* or *transit* to describe an inclined plane instead of a *vertical plane* when the *telescope* is rotated about the *horizontal axis* of the instrument. This produces an *error* in an observed *horizontal direction* of an object not situated in the plane of the *horizon,* and requires the application of a *correction for inclination of the horizontal axis.*

inclination of the horizontal axis: correction for.—A *correction* applied to an observed *horizontal direction* to eliminate any *error* that may have been caused by the *horizontal axis* of the instrument not being exactly horizontal.

If the *horizontal axis* of the instrument is not exactly horizontal, the *axis of collimation* will not cut the *horizon* at a point directly underneath (or above) an observed point: that is, a plane described by the *axis of collimation* when the *telescope* is rotated about its *horizontal axis* will not be a *vertical plane.* As the *inclination of the horizontal axis* may involve two conditions: the failure of the *horizontal* and *vertical axes* of the instrument to meet exactly at a right angle, and a possible deviation of the *vertical axis* from the direction defined by the plumb line, a determination of this *correction* requires that both the inclination of the axis and the *altitude* of the observed object be known.

index correction (leveling).—See under *correction.*

index map.—See under *map.*

index of refraction.—See under *refraction.*

indirect leveling.—See under *leveling.*

indirect observations.—See under *observation.*

initial point (United States public-land surveys).*—The point from which is initiated the survey of the *principal meridian* and *base line* controlling the survey of the public lands within a given area.

For list of initial points, *principal meridians,* and *base lines,* and the areas governed thereby, see page 138, section 141, General Land Office Manual of Instructions for the Survey of the Public Lands, edition 1930.

instrument station (leveling).—See under *station.*

instrumental error.—See under *error.*

instrumental parallax.—See under *parallax.*

intensity of gravity.—See under *gravity.*

interferometer.—An optical instrument which employs the principle of the interference of light waves to determine the amount of displacement. It is used in length standardizations, in comparing end standards, etc.

The interferometer employed in observations with the *pendulum* is a slightly modified form of the instrument used by Michelson, its inventor. With this instrument the slight movement (*flexure*) of the *receiver* which is caused by the swinging *pendulum* is measured.

interior angle.—See under *angle.*

International Map of the World map projection.—See under *map projection: International Map of the World.*

international meter.—See under *meter.*

International Prototype Meter.—See under *meter.*

intersection: point of.—See under *curve.*

intersection station.—See under *station*

interval: stadia (leveling).—The length of rod subtended between the top and bottom cross wires in the *leveling instrument* as seen projected against the face of the *leveling rod.*

invar.—An alloy of nickel and steel having a very low *coefficient of thermal expansion.*

Invar was discovered by C. E. Guillaume of the International Bureau of Weights and Measures, Paris. Having a *coefficient of thermal expansion* of about one-twenty-fifth that of steel, it has replaced steel in the construction of tapes for measuring geodetic *base lines,* and is used in other places where a metal of that characteristic is desired. Invar is used in the construction of some *leveling rods* and *first-order leveling instruments.* Invar is also used in the construction of *pendulums.*

invar pendulum.—See under *pendulum.*

invar rod.—See under *leveling rod.*

invar tape.—See under *base tape.*

invariable pendulum.—See under *pendulum.*

inverse: geographic.—See under *inverse position computation.*

inverse position computation.—The derivation of the length, and the *forward* and *back azimuths* of a *geodetic line* by computation based on the known *geodetic positions* of the ends of the line.

This is called an inverse position computation (or an inverse) because it reverses the usual order of computation, where a new *geodetic position* is derived by computation, using its *azimuth* and distance from a known *geodetic position.*

island.*—A body of land extending above and completely surrounded by water at *mean high water.*

The above definition is of universal application in *surveying* and mapping operations in the United States, and is fully supported by federal court decisions relating to the ownership of land. In foreign affairs, however, international negotiations are affected by various and different conceptions of what constitutes the boundary line between the land and the water.

isometric (conformal) latitude.—See under *latitude.*

isoperimetric curve.—A line on a *map projection* along which there is no variation from exact *scale.*

There are two isoperimetric curves passing through every point on an *equal-area map projection.* This characteristic gives that class of projections some preference for *engineering maps.*

isostasy.—A condition of approximate equilibrium in the outer part of the earth, such that the gravitational effect of masses extending above the surface of the *geoid* in continental areas is approximately counterbalanced by a deficiency of density in the material beneath those masses, while the effect of deficiency of density in ocean waters is counterbalanced by an excess of density in the material under the oceans.

The basic principle of isostasy is that the masses of prismatic columns of the outer part of the earth extending to some constant depth below the surface of the *geoid* are proportional to the areas of their *sea-level* sections, regardless of their surface *elevations.* The depth below *sea level* to which these hypothetical columns extend is known as the *depth of isostatic compensation,* and is somewhere between 60 and 70 *statute miles.* The area of the *sea-level* section of a unit hypothetical column for which *isostatic compensation* is ordinarily complete has not been determined; it may be uniform for all parts of the earth, or may vary with the character of the *relief* in the same continental regions.

While isostasy is generally accepted as a proven principle, there are several theories as to the relative distribution of the matter producing this condition of equilibrium, the two principal ones being those of Airy and of Pratt.

The Pratt theory, announced by J. H. Pratt in 1855, assumed that the continents and islands project above the average *elevation* of the solid surface of the earth because of material of less density beneath them: the higher the surface, the lesser the density below. Under the Pratt theory, complete equilibrium exists at some uniform depth below sea level: the same depth for ocean areas as for land masses. It is the theory accepted by geodesists and by many geologists, and used in the investigations of isostasy by Hayford and Bowie of the U. S. Coast and Geodetic Survey.

The Airy theory, announced by G. B. Airy in 1855, postulated the continents and islands as resting hydrostatically on highly plastic or liquid material, with roots or projections penetrating the inner material of the earth, just as icebergs extend downward into the water. The greater the *elevation,* the deeper the penetration. It has been called the "Roots of Mountain Theory," and has the support of some geologists. The fundamental difference between the two theories is that Pratt postulated uniform depth with varying density, while Airy postulated uniform density with varying depth. The term "isostasy," meaning equal pressure, was first proposed by C. E. Dutton in a paper entitled "On Some of the Greater Problems of Physical Geology," delivered before the Philosophical Society of Washington on April 27, 1889, and printed in Bulletin No. II of that Society.

isostatic adjustment.—The process of restoring and maintaining that condition of equilibrium in the so-called crust of the earth which is known as *isostasy.*

The distribution of material in the outer part of the earth is undergoing continual change by the operation of erosion, sedimentation, and other natural forces. The unbalanced condition which would naturally result from such disturbing processes is offset by the movement of material at considerable depths below the surface of the earth. See *isostatic compensation: depth of isostatic compensation.*

isostatic anomaly.—See under *anomaly.*

isostatic compensation.—The departure from normal density of material in the lower part of a column of the earth's crust which balances (compensates) land masses (*topography*) above *sea level* and deficiency of mass in ocean waters, and produces the condition of approximate equilibrium of the earth's crust known as *isostasy*.

isostatic compensation: depth of.—The depth below *sea level* at which the condition of equilibrium known as *isostasy* is complete.

Investigations of geodetic and gravimetric data in the United States determine a depth of isostatic compensation of between 60 and 70 *statute miles*.

isostatic geoid.—See under *geoid*.

isostatic reduction: Bullard method of.—See under *Hayford-Bullard* (*or Bullard*) *method of isostatic reduction*.

isostatic reduction: Hayford-Bowie method of.—See under *Hayford-Bowie method of isostatic reduction*.

isostatic reduction: Hayford-Bullard (or Bullard) method of.—See under *Hayford-Bullard* (or *Bullard*) *method of isostatic reduction*.

J

Jacob's staff.—A single staff or pole used for mounting a *surveyor's compass* or other instrument.

Used in place of a tripod, a Jacob's staff was fitted with a ball-and-socket joint at its upper end, by means of which the instrument was adjusted to a level position; the foot was fitted with a metal shoe which facilitated pressing the staff firmly into the ground. Many of the early *land surveys* in this country were made with *surveyor's compasses* mounted on Jacob's staffs.

Julian calendar.—See under *calendar*.

junction (leveling).—The place where two or more lines of levels are connected together.

junction bench mark.—See under *bench mark*.

junction detail.—See *detail: junction*.

junction figure.—A triangulation figure in which three or more *triangulation arcs* meet, or two or more *arcs* intersect.

K

knife edge (pendulum).—A suspension bearing used in the support of a *pendulum*.

The knife edge bears on a plane. The *Mendenhall pendulum* used in the work of the U. S. Coast and Geodetic Survey originally had an agate knife edge mounted in its head, which was stirrup-shaped; this agate knife edge rested on an agate *plane* which was rigidly mounted in the *receiver*. In apparatus made since 1898, the positions of knife edge and *plane* are reversed, the plane being attached to the *pendulum*, and the knife edge mounted in the *receiver*.

Kripniyuk-Kwiklokchun datum.—See under *datum*.

L

Lambert azimuthal polar map projection.—See under *map projection*.

Lambert central equivalent map projection upon the plane of the meridian.—See under *map projection: Lambert zenithal* (*azimuthal*) *equal-area*.

Lambert conformal map projection.—See under *map projection*.

Lambert equal-area meridional map projection.—See under *map projection*.

Lambert grid.—An informal designation for a *State coordinate system* based on a *Lambert conformal map projection* with two *standard parallels*.

Lambert zenithal (azimuthal) equal-area map projection.—See under *map projection*.

land boundary.—See under *boundary*.

land court.—A tribunal established for the purpose of administering legislative statutes relating to *land boundaries* and titles.

There are land courts in Massachusetts and in Hawaii.

land forms.*—The shapes into which the earth's surface is sculptured by natural forces.

A land form is subject to portrayal by various means or to the interpretation of its probable origin, cause, or history. The form in all cases is the same. Whether it is a *topographic form* or a *physiographic form* depends upon the use for which it is considered. The term land form is used extensively in place of and with the meaning of the more restricted term, *physiographic form*. It is sometimes written as one word, landform, but the two-word spelling, land form, is preferred.

land survey.—See under *surveying: land surveying*.

land surveying.—See under *surveying*.

landmark.—Any *monument* or material mark or fixed object used to designate the location of a *land boundary* on the ground. Any prominent object on land which may be used in determining a location or a *direction*.

Laplace azimuth.—See under *azimuth*.

Laplace azimuth mark.—See under *azimuth mark*.

Laplace condition.—See under *Laplace equation*.

Laplace equation.—The equation which expresses the relationship between *astronomic* and *geodetic azimuths* in terms of *astronomic* and *geodetic longitudes and geodetic latitude*. A usual form is

$$\alpha_A-\alpha_G=-(\lambda_A-\lambda_G)\sin\phi_G$$

in which α_A and λ_A are *astronomic azimuth* and *longitude*, α_G, λ_G, and ϕ_G are *geodetic azimuth, longitude*, and *latitude* respectively. The signs depend upon convention. As written above, north *latitudes* and west *longitudes* are considered positive.

The Laplace condition (expressed by the Laplace equation) arises from the fact that a *deflection of the vertical* in the plane of the *prime vertical* will give a difference between *astronomic* and *geodetic* *longitude* and between *astronomic* and *geodetic azimuth;* or, conversely, that the observed differences between *astronomic* and *geodetic* values of the *longitude* and of the *azimuth* may both be used to determine the *deflection* in the plane of the *prime vertical.* This *deflection* cannot have two values at the same place, and the imposition of the Laplace condition forces a consistency between them. Since *longitudes* can be carried through *triangulation* with very good *accuracy* and *azimuths* cannot, the practice is to compute the value of the *deflection* for *azimuth* from the *deflection* in the *prime vertical* obtained from the *longitude.* The value of the *azimuth* thus obtained is known as a *Laplace azimuth.*

The symbol α_L is proposed for *Laplace azimuths,* for use in formulas and wherever a special symbol is desirable. In computing a *control survey,* the *deflection of the vertical* at the initial station will affect any Laplace equations in the net. In the readjustment of the horizontal *control survey net* of the United States which introduced the *North American datum of 1927, azimuth control* was furnished by *Laplace azimuths* distributed throughout the net. In the investigation of the *figure of the earth* and *isostasy* which produced the International Ellipsoid of Reference, the Laplace condition was expressed as follows:

$$\cos \phi' (\lambda_A - \lambda') = -\cot \phi' (\alpha_A - \alpha')$$

The symbols marked with a prime (′) represent geodetic values, those with a subscript letter $(_A)$ represent astronomic values. The use of a subscript letter $(_G)$ for geodetic values is preferred.

The Laplace equation is so-called after Pierre Simon Laplace (1749–1827), who derived a general form for it when comparing the shape of the earth with a mathematical figure. (Mécanique Céleste, 1799, by Marquis de Laplace. English translation by Nathaniel Bowditch, published in 1832, vol. 2, p. 385.)

Laplace station.—A *triangulation* or *traverse station* at which a *Laplace azimuth* is determined.

At a Laplace station both *astronomic longitude* and *astronomic azimuth* are determined.

large scale.—See under *map scale: fractional.*

lat.—See under *latitude: difference of latitude (plane surveying).*

latitude (general).—A coordinate distance, linear or angular, from an east-west reference line.

latitude (on a sphere).—The *angle* at the center of a sphere between the plane of the *equator* and the line to the point on the surface of the sphere.

Some problems of *geodesy* and *cartography* are greatly simplified by the use of latitudes on a sphere which has some definite relationship (as equal area or equal volume) to the *spheroid.*

latitude (plane surveying).—The perpendicular distance in a horizontal plane of a point from an east-west axis of reference.

The *difference of latitude* of the two ends of a line is frequently called the latitude of the line, and defined as the *orthographic projection* of the line on a reference *meridian.* The latitude (as above defined) of the middle of a line is also referred to as the latitude of the line.

latitude: astronomic.—The *angle* between the plumb line and the plane of the *celestial equator.* Also defined as the *angle* between the plane of the *horizon* and the axis of *rotation* of the earth.

Astronomic latitude is the *latitude* which results directly from observations on celestial bodies, uncorrected for *deflection of the vertical (station error)* which, in the United States, may amount to as much as 25″. Astronomic latitude applies only to positions on the earth, and is reckoned from the *astronomic equator* (0°), north and south through 90°.

latitude: authalic (equal-area).—A *latitude* based on a sphere having the same area as the *spheroid,* and such that areas between successive *parallels of latitude* are exactly equal to the corresponding areas on the *spheroid.*

Authalic latitudes are used in the computation of *equal-area map projections.* For mathematical definition and discussion consult U. S. Coast and Geodetic Survey Special Publication No. 67, Latitude Developments Connected with Geodesy and Cartography, by O. S. Adams, 1921.

latitude: celestial.—The *arc* of a *great circle* perpendicular to the *ecliptic,* intercepted between the *ecliptic* and the object whose *latitude* is to be defined.

Celestial latitude is not the same as *astronomic latitude,* and does not enter into usual problems of *surveying.*

latitude: difference of latitude (plane surveying).—The *orthographic projection* of a line on a *meridian* of reference.

By custom, this term has been shortened to *latitude* or *lat.,* which is frequently used where difference of latitude is intended. It is north or positive, and sometimes termed the northing, for a line whose *azimuth* or *bearing* is in the northwest or northeast quadrant; it is south or negative, and sometimes termed the southing, for a line whose *azimuth* or *bearing* is in the southeast or southwest quadrant.

latitude: geocentric.—The *angle* at the center of the earth between the plane of the *celestial equator* and a line to a point on the surface of the earth.

Geocentric latitude is used as an auxiliary *latitude* in some computations in astronomy, geodesy, and cartography, in which connection it is defined as the *angle* formed with the major axis of the ellipse (meridional section of the *spheroid*) by the radius vector from the center of the ellipse to the given point. In astronomic work, geocentric latitude is also called *reduced latitude,* a term that is sometimes applied to *parametric latitude* in *geodesy* and cartography. The geocentric and *isometric latitudes* are approximately equal.

latitude: geodetic.—The *angle* which the *normal* to the *spheroid* at a point makes with the plane of the *geodetic equator.*

Geodetic latitudes are reckoned from the *equator,* but in the horizontal *control survey* of the United States they are computed from the latitude of station Meades Ranch as prescribed in the *North American datum of 1927.* In recording a *geodetic position,* it is essential that the *geodetic datum* on which it is based be also stated. A geodetic latitude differs from the corresponding *astronomic latitude* by the amount of the meridional component of the local *deflection of the vertical (station error)* which, in this country, may amount to more than 25″.

latitude: geographic.—A general term, applying alike to *astronomic latitudes* and *geodetic latitudes.*

latitude: isometric (conformal).—An auxiliary *latitude* used in the conformal mapping of the *spheroid* on a sphere.

By transforming *geographic latitudes* on the *spheroid* into isometric latitudes on a sphere, a *conformal map projection* (the Mercator) may be calculated, using spherical formulas, for the plotting of *geographic* data. For a mathematical definition of isometric latitude see U. S. Coast and Geodetic Survey Special Publication No. 67, Latitude Developments Connected with Geodesy and Cartography, by O. S. Adams, 1921. The *isometric* and *geocentric latitudes* are approximately equal.

latitude: parametric (geometric or reduced).—The *angle* at the center of a sphere which is tangent to the *spheroid* along the *geodetic equator,* between the plane of the *equator* and the radius to the point intersected on the sphere by a straight line perpendicular to the plane of the *equator* and passing through the point on the *spheroid* whose parametric latitude is defined.

Parametric latitude is an auxiliary *latitude* used in problems of *geodesy* and *cartography.* For mathematical definition see U. S. Coast and Geodetic Survey Special Publication No. 67, Latitude Developments Connected with Geodesy and Cartography, by O. S. Adams, 1921. In astronomical work, when the term reduced latitude is used, *geocentric latitude* is meant.

latitude: rectifying.—The *latitude* on a sphere such that a *great circle* on it has the same length as a *meridian* on the *spheroid,* and such that all lengths along a *meridian* from the *equator* are exactly equal to the corresponding lengths on the *spheroid.*

Rectifying latitude is an auxiliary *latitude* used in problems of *geodesy* and *cartography.* For mathematical definition see U. S. Coast and Geodetic Survey Special Publication No. 67, Latitude Developments Connected with Geodesy and Cartography, by O. S. Adams, 1921.

latitude: variation of.—A small periodic change in the *astronomic latitude* of points on the earth, due to *variation of the pole.*

latitude equation.—See under *equation.*

latitude level.—See under *level.*

least squares.—A mathematical method of determining the most probable values of a series of quantities from a set of observations greater in number than are necessary to determine those quantities.

Credit for the discovery in 1770 of the method of least squares is sometimes given to Lagrange, although his method was lost or forgotten for some time. Credit for the discovery is more often given to Gauss, who used it in 1795 but did not publish his results until 1809, while it had been demonstrated by Adrain in 1808. Laplace demonstrated it in 1810 and in 1812, and its theory was compared with practice by Bessel in 1818. It rests upon the mathematical demonstration that where each of a very large number of observations of any quantity is of the same quality as the others, the most probable value of the quantity is the one for which the sum of the squares of the *residual errors* (or *corrections*) is a minimum. If the observations are of unequal *weight,* then the most probable value is the one for which the sum of the squares of the weighted residuals is a minimum. In survey work, the method of least squares provides a means of obtaining the most probable values of position and associated data, and at the same time fully coordinating and correlating those data. The adjusted survey data distributed by the U. S. Coast and Geodetic Survey are obtained by the method of least squares. Least squares have also been designated as "minimum squares."

left bank.—See under *bank of stream.*

Legendre's theorem.—The *angles A, B,* and *C* of a spherical triangle whose sides are *a, b,* and *c,* supposed very small with respect to the radius of the sphere, are equal to the corresponding *angles* of a plane triangle whose sides are *a, b,* and *c,* increased each by one-third the *spherical excess* of the triangle.

(Alternate definition.)

In two triangles, one a spherical triangle whose sides a, b, and c are very small in comparison with the radius of the sphere and the other a plane triangle having sides of the same length as the spherical triangle, the *angles* A, B, and C of the spherical triangle are greater than the corresponding *angles* of the plane triangle, each by one-third of the *spherical excess* of the spherical triangle.

The use of Legendre's theorem greatly simplifies the treatment of spherical triangles in *geodetic* work, by permitting their solution as plane triangles. This theorem is named after the French mathematician, Adrien Marie Legendre (1752–1833), who demonstrated it.

length correction: tape.—See under *tape.*

length equation.—See under *equation.*

lens: achromatic.—See under *achromatic lens.*

lens: aplanatic.—See under *aplanatic lens.*

lens axis.—See under *principal axis.*

level: bull's-eye.—See under *level: circular.*

level: chambered spirit.—A level tube with a partition near one end which cuts off a small air reservoir, so arranged that the length of the bubble can be regulated.

level: circular.—A *spirit level* having the inside surface of its upper part ground to spherical shape, the outline of the bubble formed being circular, and the *graduations* being concentric circles.

This form of *spirit level* is used where a high degree of precision is not required, as in plumbing a *level rod* or setting an instrument in approximate position. Also termed universal level; bulls-eye level.

level: first-order.—See under *leveling instrument.*

level: hand.—See under *leveling instrument.*

level: hanging.—A *spirit level* so mounted that, when in use, its level tube is lower in elevation than its points of support.

A hanging level literally hangs from the member of the surveying or astronomical instrument whose position with reference to a horizontal line it measures. It is used in place of a *striding level* to measure the inclination of the *horizontal axis* of the *broken-telescope transit.*

level: latitude.—A sensitive *spirit level* attached to the *telescope* of an instrument employed for observing *astronomic latitude,* in such manner that when the *telescope* is clamped in position, the level measures, in a *vertical plane,* variations in the direction of the *line of collimation.*

A latitude level is of greater sensitivity than the level which is attached to the *finder circle* of an astronomical instrument to aid in placing its *telescope* at the proper inclination for bringing a selected celestial object into the field of view at the expected time.

level: precise.—See under *leveling instrument.*

level: spirit (general).—A small closed vessel of transparent material (glass), having the inside surface of its upper part curved (circular) in form; the vessel is nearly filled with a fluid of low viscosity (alcohol or ether), enough free space being left for the formation of a bubble (blister) of air and gas, which will always assume a position at the top of the vessel.

Many *surveying* operations depend upon the instrumental establishment of the plane of the *horizon,* which is accomplished with the aid of a spirit level. For work of high *precision,* where exact establishment of the plane of the *horizon* is desired, a spirit level of great sensitivity is employed to determine deviations from that plane, and corresponding *corrections* are then applied to the observations.

There are two types of spirit levels used in surveying: one having the curved surface spherical in form, producing a bubble of circular outline, and defined under the term *circular level.* The other, and much more generally used type, has a vessel in the form of a circular tube, the longitudinal axis of which is also circular in form. It is defined under the term spirit level, and is the type usually referred to when that term is used.

level: spirit.—A closed glass tube (vial) of circular cross section, its center line also forming a circular *arc,* its interior surface being ground to precise form; it is nearly filled with ether or other liquid of low viscosity, enough free space being left for the formation of a bubble of air and gas.

This is the type of spirit level used on those surveying and astronomical instruments which make use of the *horizon* (or *zenith*) as a reference of observations, and which require more than an *approximate* adjustment to such reference. When in adjustment, a tangent to the center of the level bubble defines a horizontal line in the plane of the longitudinal axis of the level tube, with an *accuracy* and *precision* which depend upon the quality of the workmanship and the sensitivity of the level. The sensitivity, sometimes termed sensibility, of a *spirit level* depends upon the radius of curvature of its longitudinal section: the longer the radius, the more sensitive the level. The sensitivity is rated by equating the linear length of a division between graduation marks on the level tube and its angular value at the center of curvature of the tube. Example: stride level of *meridian telescope* No. 9 (U. S. Coast and Geodetic Survey), 1 division = 2 mm. = 1.884″ at 12° C. A usual value for the sensitivity of the long level on a surveyor's *transit* is: 1 division = 3/20 inch = 30″.

level: striding.—A *spirit level* so mounted that it may be placed above and parallel with the *horizontal axis* of a surveying or astronomical instrument, and so supported in use that it may be used to measure the inclination of the *horizontal axis* to the plane of the *horizon.*

Generally, the *spirit level* of a striding level is of greater than average sensitivity, and its mounting has supports in the form of inverted wyes which rest directly upon the pivots on which the *telescope* of the instrument rotates. It is used to make the inclination of the *horizontal axis* quite small and then to measure the magnitude of any remaining inclination.

level: universal.—See under *level: circular level.*

level: wind, spirit level.—Lack of parallelism between the axis of a *spirit level* vial and the line joining the centers of its supports.

When wind (pronounced to rhyme with find) is present, if the *spirit level* is rocked on its supports, the bubble will respond with a longitudinal movement.

level axis: spirit.—The line tangent to the surface of a spirit-level tube (vial) against which the bubble forms, at the center of the graduated scale of the level, and in the plane of the tube (vial) and its center of curvature. Also called axis of the level, axis of the level bubble.

When the level bubble is in the center of the *graduations,* and the plane through the center line of the level tube and its center of curvature is vertical, the spirit level axis will be horizontal.

level constant.—See under *constant.*

level correction (leveling).—See under *correction.*

level line: duplicate.—A line of *spirit leveling* composed of two single lines run over the same route, but in opposite directions, and using different *turning points.*

level line: multiple.—Two or more single lines of *spirit leveling* run between the same terminal points, but along different routes.

level line: simultaneous.—A line of *spirit leveling* composed of two single lines run over the same route, both in the same direction, but using different *turning points.* Also called a simultaneous double line.

level net.—See under *survey net.*

level surface.—A surface which at every point is perpendicular to the plumb line or the direction in which *gravity* acts.

A level surface is an *equipotential surface.* The surface of a body of still water is a level surface (see *water leveling*). The surface of the ocean, if changes caused by tides, currents, winds, atmospheric pressure, etc., are not considered, is a level surface. The surface of the *geoid* is a level surface. Any line lying in a level surface is a *level line.* In a *survey* of a limited area, a level surface is sometimes treated as a plane surface. Level surfaces are approximately spheroidal in shape, the distance between any two level surfaces decreasing with increase of *latitude.* For example: a level surface which is 1,000 *meters* above the mean surface of the sea at the *equator* is 995 *meters* above that surface at the poles.

level trier.—An apparatus for use in measuring the angular value of the *divisions* of a *spirit level.*

A level trier may consist of a beam mounted on a stable base: one end of the beam is hinged to the base, the other end can be moved vertically by means of a micrometer screw, the angular amount of the movement being measured with the *micrometer.* The *spirit level* is mounted on the beam, and the angle through which the beam moves while the bubble of the *spirit level* travels the length of a *division* is measured with the *micrometer.*

leveling: astronomical.—The determination of the figure of the *geoid* from *deflections of the vertical.*

leveling: barometric.—A method of determining differences of *elevation* from differences of atmospheric pressure observed with a *barometer.*

By the application of certain *corrections* and the use of what is sometimes called the "barometric formula," a difference of atmospheric pressure at two places is transformed into a difference of *elevations* of those places. If the *elevation* of one station above a *datum* (as *sea level*) is known, the approximate *elevations* of other stations connected with it by barometric leveling can be known. By using *barometers* of special design, and including several stations of known *elevation* in a series of occupied stations, the *accuracy* of the *elevations* determined for the new stations is increased. *Corrections* are applied for temperature, *latitude,* index of barometer, closure of circuit, *diurnal variation* in atmospheric pressure, etc.

leveling: direct.—The determination of differences of *elevation* by means of a continuous series of short horizontal lines, the vertical distances from which to adjacent ground marks are determined by direct observations on graduated rods with a *leveling instrument* equipped with a *spirit level.*

leveling: first order.—*Spirit leveling* conforming to the following criteria: all first-order leveling to be divided into sections of 1 to 2 kilometers in length; each section to be leveled over in both forward and backward directions; the results of the two runnings over a section not to differ by more than 4.0 millimeters times the square root of the length of the section in kilometers (4.0 *mm.*$\sqrt{K}$), the equivalent of which is 0.017 foot times the square root of the length of the section in miles (0.017 *ft.*$\sqrt{M}$).

The above designation and criteria were recommended by the Federal Board of Surveys and Maps in May 1925. First-order leveling was formerly called precise leveling and leveling of high precision.

leveling: geodetic.—*Spirit leveling* of a high order of *accuracy,* usually extended over large areas, to furnish accurate vertical *control* as a basis for the control in the vertical dimension for all surveying and mapping operations.

Spirit leveling follows the *geoid* and its associated *level surfaces* which are irregular, rather than any mathematically determined *spheroid* or *ellipsoid* and associated regular level surfaces.

leveling: indirect.—The determination of differences of *elevation* from (1) *vertical angles* and horizontal distances, as in *trigonometric leveling;* (2) comparative *elevations* derived from values of atmospheric pressure determined with a *barometer,* as in *barometric leveling;* and (3) *elevations* derived from values of the boiling point of water determined with a *hypsometer,* as in *thermometric leveling.*

leveling: precise.—See under *leveling: first-order leveling.*

leveling: second-order.—*Spirit leveling* which does not attain the quality of *first-order leveling,* but does conform to the following criteria: lines between *bench marks* established by *first-order leveling* to be run in only one direction, using first-order instruments and methods; or other lines to be divided into sections, over which forward and backward runnings are to be made; the closure in either case not to exceed 8.4 millimeters times the square root of the length of the line (or section) in kilometers (8.4 *mm.*$\sqrt{K}$), the equivalent of which is 0.035 foot times the square root of the length of the line (or section) in miles (0.035 *ft.*$\sqrt{M}$).

The above designation and criteria were recommended by the Federal Board of Surveys and Maps in May 1925.

leveling: spirit.—The determination of *elevations* of points with respect to each other, or with respect to a common *datum,* by means of an instrument using a *spirit level* to establish a horizontal *line of sight.*

A *spirit level* is attached to a *telescope* in such a way that the *axis of the level* and the *line of collimation* may be made parallel, and the level adjusted so that its axis is horizontal. The difference of readings on vertical rods held on two different points is the difference in *elevation* of the points. If the *elevation* of one point is known, the other also becomes known. By a series of progressive *level lines,* each comprising the observed values at a single station, the *elevations* of a series of points (*bench marks*) are determined. Instruments used in spirit leveling are of various designs, to satisfy needs based on speed, cost, *accuracy,* and *precision.* Among such types are the *hand level,* the *wye level,* the *dumpy level,* the *Fischer level,* etc. Spirit leveling has been classified by the Federal Board of Surveys and Maps according to the *accuracy* of the results obtained as follows: *first-order leveling, second-order leveling,* and *third-order leveling.*

leveling: thermometric.—The determination of *elevations* above *sea level* from observed values of the boiling point of water.

The temperature at which water boils at any point on the earth depends upon the weight of the incumbent atmosphere at that point. See *hypsometer.*

leveling: third-order.—*Leveling* which does not attain the quality of *second-order leveling,* but does conform to the following criteria: lines of third-order leveling shall not be extended more than 30 miles from lines of *first-* or *second-order leveling,* and must close upon lines of equal or a higher order of *accuracy;* closing errors must not exceed 12 millimeters times the square root of the length of the line in kilometers (12 *mm.* $\sqrt{K}$), the equivalent of which is 0.05 foot times the square root of the length of the line in miles (0.05 *ft.* $\sqrt{M}$).

The above designation and criteria were recommended by the Federal Board of Surveys and Maps in May 1925. Third-order leveling is used for subdividing loops of *first-* and *second-order leveling,* and to provide local vertical control for detailed surveys.

leveling: trigonometric.—The determination of differences of *elevations* by means of observed *vertical angles,* combined with lengths of lines. See *vertical angulation.*

leveling: water.—A method of obtaining relative *elevations* by observing heights with respect to the surface of a body of still water.

The surface of a body of still water, as of a lake, is a *level surface* (*equipotential surface*), and the relative *elevations* of objects along its shores may be obtained by taking the differences of their heights with respect to the surface of the water. It is with this meaning that the term *"water levels"* is generally used.

leveling instrument: dumpy level.—A leveling instrument in which the *telescope* is permanently attached to the leveling base, either rigidly or by a hinge that can be manipulated by means of a micrometer screw.

The dumpy level takes its name from the dumpy appearance of the early type of this instrument, the *telescope* of which was short and had a large object glass. Dumpy levels constitute one of two principal classes of leveling instruments in which most leveling instruments used in surveying belong. The other class is represented by the *wye* (Y–) *level.* The *Fischer level* used by the U. S. Coast and Geodetic Survey on *first-* and *second-order work* is a dumpy level. Method of mounting the *telescope* and not its shape is the identifying characteristic of a dumpy level.

leveling instrument: Egault level.—A French instrument having the *spirit level* attached to a level bar which also carries wyes in which the *telescope* rests.

In the so-called *wye level,* the *spirit level* is attached to the *telescope,* and is reversed with it. In Egault's level, the *telescope* is reversed in the wyes without disturbing the *spirit level.*

leveling instrument: first-order level.—A leveling instrument designed for the attainment of first-order results. See under *leveling: first-order.*

A first-order instrument is generally characterized by special features of design, and by superior workmanship in construction. A general, but not always exact, distinction between instruments used in *first-order* and *second-order leveling* and other leveling instruments is the following: a first-order level is carefully adjusted, and any *instrumental errors* remaining are determined by observation and applied as *corrections* to the observed differences of *eleva-*

tion; in using other leveling instruments, a careful adjustment of the instrument is made, and thereafter no account is taken of *errors* remaining in the instrument. An important characteristic of leveling instruments used in *first-order work* at the present time is a mirror and prism arrangement by means of which the level bubble and the level rod can be viewed at the same time, while a micrometer-screw control of the *telescope* makes it possible to bring the level bubble to the center of its tube at the instant of reading the rod. There have been a number of different designs of leveling instruments used in this country on *first-order leveling* (formerly called *precise leveling*), but all such work by the U. S. Coast and Geodetic Survey is now done with a *Fischer level.*

leveling instrument: Fischer level.—The official designation of the instrument used by the U. S. Coast and Geodetic Survey in the execution of *first-order leveling.* It is a *dumpy level,* constructed of *invar,* binocular in design, having the *spirit level* placed in the tube of the main *telescope,* quite close to the *line of collimation.* The main *telescope* is on the right. A suitable arrangement of mirror and prisms brings an image of the bubble to the left-hand eyepiece. The *telescope* is so mounted that, by means of a micrometer screw, its inclination can be changed through a small *angle* without altering its *elevation.*

In use, the observer sights the rod with the right eye, and at the same time views the bubble with the left eye. By means of the micrometer screw, the inclination of the *telescope* is changed until the bubble is seen in the middle of its tube; the rod is read with the bubble in the mean position. The *collimation error,* made quite small by instrumental adjustment, is then determined, and a *correction* is applied to eliminate its effect from the results. An *invar* rod is used, and the effects of rod temperature and index error considered. In the computations, allowance is made for curvature of the earth and *atmospheric refraction.* The Fischer level was designed by E. G. Fischer, formerly Chief of the Instrument Division, U. S. Coast and Geodetic Survey, and adopted by that Bureau in 1899. The use of *invar* in its construction is of more recent date.

leveling instrument: Gravatt level.—A *dumpy level* with the *spirit level* mounted on top of a short *telescope* tube having a large object glass. Later made with wyes.

This is said to be the original *dumpy level,* which is also claimed for the *Troughton level.*

leveling instrument: hand level.—A small leveling instrument in which the *spirit level* is so mounted that the observer may view the bubble at the same time that he observes an object through the *telescope.*

The instrument is held in the hand when in use, and the viewing of the bubble is accomplished by means of a prism or mirror in the *telescope* tube. The instrument is used in *reconnaissance surveys,* and in cross-section and other work where a high degree of *precision* and *accuracy* is not required.

leveling instrument: Lenoir level.—An instrument which has the *telescope* passing through steel blocks, one near each end, whose upper and lower faces are plane and closely parallel: the lower faces rest upon a brass circle; the upper faces support a spirit level, which is reversed in leveling the instrument.

leveling instrument: Locke hand level.—The generally known design of a *hand level.*

leveling instrument: precise level.—A leveling instrument for use where results of a high order of *accuracy* and *precision* are required. See *leveling instrument: first-order level.*

leveling instrument: Stampfer level.—A name given a type of leveling instrument having the *telescope* tube so mounted that it could be moved in a *vertical plane* about a horizontal axis, involving the use of a *striding level* and a micrometer screw. This type of instrument was also termed a Vienna level.

The mechanical principle involved was apparently quite similar to that of the present-day *first-order level.* See *leveling instrument: Fischer level.*

leveling instrument: Troughton level.—An English instrument having the *spirit level* permanently attached to the top of the telescope tube.

This is said to be the original *dumpy level,* which is also claimed for the *Gravatt level.*

leveling instrument: U. S. Geological Survey level.—A level of the dumpy type, constructed of stainless steel. It has an internal-focusing *telescope;* the level bubble is centered by the end-coincidence method, effected with the aid of a prism device and stellite mirror which can be adjusted by the observer.

This instrument is used by the U. S. Geological Survey in *second-* and *third-order leveling.* By substituting a more sensitive *spirit level* and a more powerful eyepiece for the regular equipment, it may be used for *first-order work.* The interval between *stadia* wires is such that the half-interval intercept on the rod, multiplied by 1,000, will give the distance to the rod in feet.

leveling instrument: wye (Y) level.—A leveling instrument having the *telescope* with attached *spirit level* supported in wyes (Y's), in which it may be rotated about its longitudinal axis (*collimation axis*), and from which it may be lifted and reversed, end for end.

The instrumental adjustments made possible by this mounting are peculiar to the instrument. It forms one of two general classes of leveling instruments, the other class being represented by the *dumpy level.*

leveling of high precision.—See under *leveling: first-order leveling.*

leveling rod (general).—A straight rod or bar, designed for use in measuring a vertical distance between a point on the ground and the *line of collimation* of a leveling instrument which has been adjusted to a horizontal position.

A leveling rod is usually of wood, and has a flat face which is graduated in terms of some linear unit and fractions thereof, the zero of the *graduations* being at one end of the rod. Some rods have the *graduations* on a metal face (*invar* rod). On some rods the *graduation* marks are designed to be read by the observer at the leveling instrument: such a rod is termed a speaking rod or self-reading rod. Another type of rod, called a target rod, carries a target which is moved into position by the rodman in accordance with signals given by the man at the instrument: when the target is bisected by the *line of collimation,* it is read by the rodman. There are many different designs of rods, which will be found defined under leveling rod, followed by the identifying name of the rod.

leveling rod: Barlow.—A *speaking rod* marked with triangles each 0.02 foot in height.

leveling rod: Boston.—A two-piece rod with fixed target on one end.

The target is adjusted in *elevation* by moving one part of the rod on the other. Read by *vernier.*

For heights greater than 5½ feet, the target end is up; for lesser heights, the target end is down. It is a little-used rod.

leveling rod: Conybeare.—A *speaking rod* having tenth-of-foot *divisions*, alternately black and white, each *division* having on it three hexagons in the contrasting color, white or black.

leveling rod: Gravatt.—A *speaking rod*, marked with rectangles each 0.01 foot high, the rectangles at the tenths of foot being longer, and those at the half-tenths being identified by dots.

leveling rod: invar.—A *leveling rod* having the *graduations* painted on a strip of *invar*.

The strip of *invar* is set into the face of the wooden rod, firmly attached to the bottom of the rod (rod shoe), and held in place by guides and by a tension spring at the top of the rod. Invar rods are used in *first-order leveling*.

leveling rod: Molitor precise.—A *speaking rod* of T-shaped cross section, with *graduation* marks shaped as triangles and rectangles, the smallest division being 2 millimeters.

Read by estimation to single millimeters. Equipped with thermometer and *circular level*.

leveling rod: New York.—A two-piece rod with movable target.

For heights greater than 6½ feet, the target is clamped at 6½ feet and raised by extending the rod. Graduated to hundredths of a foot and read by *vernier* to thousandths.

leveling rod: Pemberton.—A *speaking rod* marked with alternate rows of circular and diamond-shaped dots, running diagonally across the rod.

Read to hundredths of a foot.

leveling rod: Philadelphia.—A two-piece *target rod*, with *graduation* marks so styled that it may also be used as a *speaking rod*.

For heights greater than 7 feet the target is clamped at 7 feet, and raised by extending the rod. As a *target rod* it is read by *vernier* to thousandths of a foot; as a *speaking rod*, to half-hundredths.

leveling rod: self-reading or speaking.—See under *leveling rod (general)*.

leveling rod: Stephenson.—A *speaking rod* having *graduations* forming a diagonal scale, with horizontal lines through the tenth-of-foot marks.

This rod is read to hundredths of a foot.

leveling rod: target.—See under *leveling rod (general)*.

leveling rod: U. S. Coast and Geodetic Survey first-order.—Adopted in 1916, the rod now used in *first-* and *second-order leveling* is a *speaking rod* having a strip of *invar* set in one face, and held in place by a tension spring at the top of the strip, its bottom being firmly attached to the shoe of the rod. The *invar* strip carries centimeter *divisions*, alternately black and white, and is read by estimation to millimeters. The back side of the rod is graduated in feet and tenths. It is equipped with thermometer and *circular level*. Sometimes called an *invar rod*.

leveling rod: U. S. Coast and Geodetic Survey rods.—*Leveling rods* used in the execution of *first-order* (formerly *precise*) and *second-order* (formerly *primary*) *leveling*, before 1895. A *target rod* having *graduations* on a metal strip about 3 meters long set in a wooden rod of plus-sign (+) cross section, the lower end of the strip firmly attached to the wooden rod, the upper end free to respond to changes of temperature. Pointings made on target equipped with small metal scale.

About 1894, a rod of plus-sign (+) cross section was constructed, having *graduation* marks painted to show 1-centimeter intervals, brass plugs being set at 2-centimeter intervals. The rod carried a target and was read to millimeters by means of a 2-centimeter scale. It was equipped with a thermometer and a *circular level*. This rod was replaced with a *speaking rod* of plus-sign (+) cross section, having painted on it *graduation* marks of 1 centimeter, read by estimation to millimeters. Two metal plugs set in the rod 3 meters apart were used in its standardization. It was equipped with thermometer and *circular level*. In 1916 the rod now in use was adopted. See under *leveling rod: U. S. Coast and Geodetic Survey first-order leveling rod.*

leveling rod: U. S. Engineer precise.—A *speaking rod* of T-shaped cross section, 12 feet long, graduated in centimeters.

leveling rod: U. S. Geological Survey precise.—1. A *speaking rod* graduated in *yards* and fractions of a *yard*. It is read for each of three cross wires to the nearest thousandth of a yard. The sum of the three readings is then the mean reading in feet to the nearest thousandth. 2. A *target rod* of plus-sign (+) cross section, a little over 12 feet in length.

There are two forms of this rod: the single-target rod, having *graduations* on one face only; and the double-target rod, with *graduations* on two opposite faces.

line of apsides.—The major axis of an elliptical orbit extended indefinitely in both directions.

line of collimation.—See under *collimation*.

line of levels.—A continuous series of measured differences of elevation. The individual measured differences may be single observations in the case of single-run leveling or the means of repeated observations in the case of double-run leveling.

line of levels: spur.—A *line of levels* run as a branch from the main line of levels, either for the purpose of determining the *elevations* of marks not conveniently reached by the main line of levels or to connect with *tidal bench marks* or other previously established *bench marks* in obtaining checks on old leveling either at the beginning or end of a line of levels or at intermediate junctions along the new line of levels.

line of position.—Any line on which a position is known to be located. A *Sumner line*.

Line of position or position line is a more general term than *Sumner line*, which it includes.

line of sight.—See under *collimation: line of collimation*.

line tree (United States public-land surveys).*—A tree intersected by a surveyed line on which legal *corners* are established, and reported in the field notes of the survey. Also termed a sight tree, fore-and-aft tree.

lines on a spheroid.—See under *spheroid*.

link.—A *unit* of linear measure, one one-hundredth of a *chain*, and equivalent to 7.92 inches.

See under *chain: Gunter's chain*.

link (leveling).—A line, a part of a line, or a combination of lines or parts of *lines of levels*, which, taken as a unit, make a continuous piece of leveling directly from one *junction bench mark* to another junction bench mark without passing through or over any other junction bench marks.

link of levels.—See *link (leveling)*.

local adjustment.—See under *adjustment: station adjustment*.

local apparent time.—See under *time*.

local mean time.—See under *time*.

local sidereal time.—See under *time*.

location survey.—See under *survey*.

Locke hand level.—See under *leveling instrument*.

long chord.—The chord which extends from the point of curvature to the point of tangency. See under *curve: point of curvature; point of tangency.*

In descriptions of a circular land boundary, the length and bearing of the long chord is an important feature.

longitude (general).—A coordinate distance, linear or angular, from a north-south reference line.

longitude: astronomic.—The *angle* between the plane of the *celestial meridian* and the plane of an initial *meridian,* arbitrarily chosen.

Astronomic longitude is the *longitude* which results directly from observations on celestial bodies, uncorrected for *deflection of the vertical,* the *prime vertical* component of which, in the United States, may amount to more than 18″. Astronomic longitude is measured by the *angle* at the celestial pole between the tangents to the local and the initial *meridians,* or by the *arc* intercepted on the *equator* by those *meridians.* At an international convention held in Washington, D. C., in 1884, the *Meridian of Greenwich* was adopted as the initial or *prime meridian* for all *longitudes,* and is now generally so used. An expression of longitude is sometimes accompanied by a statement of the method used in its determination, often in adjective form, as "wireless longitude." This may serve as an indication of its *precision.* Among the various methods used are the following:

celestial signals.—The local *time* at the new station of the occurrence of celestial phenomena is compared with the corresponding *time* of the occurrence at a base station. Among such phenomena are *eclipses* of the Moon, *eclipses* of Jupiter's satellites, *occultations* of stars, etc.

chronometric.—The local *time* at the new station is compared with the *time* at a base station by transporting *chronometers* from one station to the other.

lunar-distance, moon-altitude, moon-culmination, etc.—The local *time* at the new station is compared with the *time* at a base station, both *times* being determined from the position of the moon relative to other celestial bodies.

telegraphic.—The local *time* at the new station is compared with the *time* at a base station by means of the electric-telegraph land-lines and sea cables.

terrestrial signals.—Local *times* noted at new and base stations of the occurrence of such signals as a flash of gunpowder.

wireless.—The local *time* at the new station is compared with the *time* at a base station by means of radio.

longitude: celestial.—The *arc* of the *ecliptic* intercepted between the *vernal equinox* and the foot of a *great circle* perpendicular to the *ecliptic* and passing through the object whose *longitude* is to be defined.

Celestial longitude is measured along the *ecliptic* from west to east. It is not the same as *astronomic longitude* and does not enter into usual problems of *surveying.*

longitude: geodetic.—The *angle* between the plane of the *geodetic meridian* and the plane of an initial *meridian,* arbitrarily chosen.

A geodetic longitude may be measured by the *angle* at the pole of revolution of the *spheroid* between the local and initial *meridians,* or by the *arc* of the *geodetic equator* intercepted by those *meridians.* In the United States, geodetic longitudes are numbered from the *Meridian of Greenwich,* but are computed from the *meridian* of station Meades Ranch as prescribed in the *North American datum of 1927.* In recording *geodetic position* it is essential that the *geodetic datum* on which it is based also be stated. A geodetic longitude differs from the corresponding *astronomic longitude* by the amount of the *prime-vertical* component of the local *deflection of the vertical* divided by the cosine of the *latitude.* In this country, this may amount to as much as 26″.

longitude: geographic.—A general term, applying alike to *astronomic* and to *geodetic longitudes.*

longitude equation.—See under *equation.*

longitude signal.—See under *signal.*

longitude term.—See under *gravity formula.*

loran.—A pulse-type electronic navigation system for measuring distance differences with respect to fixed transmitters of known geographic position.

A loran receiver may be located on a particular hyperbola by measuring the difference in arrival times of pulses sent in a synchronous manner from a single pair of loran transmitters. The receiver may also be located on another hyperbola by a similar time measurement relative to another pair of transmitters. Intersection of the hyperbolas fixes the location of the receiving point. Loran fixes may be obtained at a range of 750 nautical miles in daylight and 1,400 nautical miles at night, and are very nearly independent of weather conditions. The *precision* of a loran *fix* is comparable to that obtained by high grade celestial observations with a *sextant.*

lost corner (United States public-land surveys).—See under *corner.*

lower culmination.—See under *culmination.*

low-water line.—The line defined by the boundary of a body of water at its lowest stage (elevation).

As there may be appreciable variation with time in the elevation of the low-water surface at a designated locality, a mean value of the low-water elevation should be used in determining the ground position of the low-water line. See *mean low water.*

loxodrome: loxodromic curve.—See under *rhumb line.*

lunar month.—See under *month.*

lunation.—See under *month: synodical month.*

Luzon datum.—See under *datum.*

M

magnetic chart.—See under *chart.*

magnetic declination.—The *bearing* (reckoned east or west from the north branch of the celestial meridian plane) of magnetic north as determined by the positive pole of a freely suspended magnetic needle which is subject to no transient artificial disturbance.

In nautical and aeronautical navigation the term *variation* is used instead of *declination,* and the *angle* is called variation of the compass or *magnetic variation.*

magnetic variation.*—Regular or irregular change with time of *magnetic declination,* dip, or intensity.

In nautical and aeronautical navigation, and in some localities *surveying,* the term variation is used instead of, and meaning, *magnetic declination.* The regular variations are secular, the change from year to year which usually extends for many decades in the same direction; annual, having a period of one *year;* and diurnal, having a daily period. The irregular variations are known, when severe, as magnetic storms. Locally the magnetic field may be affected by direct-current electricity and other artificial disturbances. It was once a common practice of surveyors to denote as variation the net amount by which the compass departed from the direction taken as north in the description of a particular line, even when this was known to be slightly at variance with the *celestial meridian.*

main-scheme station.—See under *station: principal station.*

manometer.—A gage for measuring the pressure of a gas.

In observing the *Mendenhall pendulum,* the *receiver* in which it is swung is exhausted of air to a prescribed maximum pressure. A manometer composed of a U-shaped tube, closed at one end and containing mercury, is placed in the *receiver* to register the air pressure.

map.*—A representation on a plane surface, at an established *scale,* of the physical features (natural, artificial, or both) of a part or the whole of the earth's surface, by means of signs and symbols, and with the means of *orientation* indicated. Also a similar representation of the heavenly bodies.

A map may emphasize, generalize, or omit the representation of certain features to satisfy specific requirements. The type of information which a map is designed primarily to convey is frequently used, in adjective form, to distinguish it from maps of other types. A map should carry a record of the *projection* upon which it is constructed.

map: aerial.—See under *chart: aeronautical chart.*

map: aerial navigation.—See under *chart: aeronautical chart.*

map: air navigation.—See under *chart: aeronautical chart.*

map: aviation.—See under *chart: aeronautical chart.*

map: base.*—A map on which information may be placed for purposes of comparison or geographical correlation.

The term base map was at one time applied to a class of *maps* now known as *outline maps.* It may be applied to *topographic maps,* also termed mother maps, which are used in the construction of many types of *maps* by the addition of particular data.

map: bus and truck.*—A *special-purpose map* showing routes used or suitable for use in commercial transportation by motor vehicles.

map: cadastral.*—A *map* showing the boundaries of subdivisions of land, usually with the bearings and lengths thereof and the areas of individual tracts, for purposes of describing and recording ownership.

A cadastral map may also show culture, drainage, and other features relating to the value and use of land.

map: compiled.*—A *map* incorporating information collected from various sources, not developed by *surveys* made for the special purpose of the *map* in question.

Most *small-scale maps* of large areas are compiled.

map: composite.*—A *map* which portrays information of two or more general types.

A composite map is usually a *compiled map,* bringing together on one *map,* for purposes of comparison, data which were originally portrayed on separate *maps.*

map: contour.*—A *topographic map* which portrays relief by means of *contour lines.*

The form "contoured map" is disapproved.

map: county.*—A *map* of the area of a county as a unit.

map: distribution.*—A *map* which shows the *geographic* arrangement of a specific product, commodity, or formation.

Distribution maps may be nonquantitative or quantitative. Nonquantitative distribution maps may be simple, compound, dominant, etc., depending upon the character of the detail portrayed. Quantitative distribution is statistical in character.

map: dynamic.*—A map designed to show motion, action, or change.

A dynamic map involves an element of *time.* The term may be applied to *maps* depicting traffic flow, migration, military movement, progress in an engineering project, historical geography, etc. Various symbols, such as flow lines and arrows, are employed to depict movement, or a dynamic map may be a composition of two or more *static maps,* depicting comparable data at stated dates or times. A plan of a proposed development is not a *map.* Therefore, the use of the term dynamic map to describe such a plan is disapproved.

map: engineering.*—A *map* showing information that is essential for planning an engineering project or development and for estimating its cost.

An engineering map is usually a large-scale *map* of a comparatively small area or of a route. It may be entirely the product of an engineering survey, or reliable information may be collected from various sources for the purpose and assembled on a *base map.*

map: hemispherical.*—A *map* of one-half of the earth's surface, bounded by the *equator,* or by *meridians.*

The earth is usually considered divided either at the *equator,* into the northern and southern hemispheres, or along some *meridian* between Europe and America and continued around the globe, into eastern and western hemispheres. The division used in *cartography* is usually along the *meridian* of *longitude* 20° west of Greenwich. The Americas are in the Western Hemisphere. The division of the surface of the earth into hemispheres of political significance is not a determined matter.

map: highway.*—A map on which the location of improved roads and highways is shown.

map: hypsographic.*—A *topographic map* on which *elevations* are referred to a *sea-level datum.*

Sometimes called a hypsometric map, the use of which term is disapproved.

map: hypsometric.*—Disapproved. See under *map: hypsographic map.*

map: index.*—A *map* which shows the location of collections of related data, whether in the form of other *maps,* or statistical tables, or descriptive in character.

Index maps may be used to show references to the origin, place of deposit or filing, and other common characteristics (identifying quality) of such data.

map: large scale.—See under *map scale: fractional.*

map: outline.*—A *map* which presents just sufficient geographic information to permit the correlation of additional data placed upon it.

Outline maps correspond to what were at one time known as *base maps.* Outline maps usually show only coast lines, principal rivers, major civil boundaries, and large cities, leaving as much space as possible for the reception of particular data. An outline map presents less detail than a *base map.*

map: planimetric.*—A *map* which presents the horizontal *positions* only for the features represented; distinguished from a *topographic map* by the omission of *relief* in measureable form.

The natural features usually shown on a planimetric map include rivers, lakes, and seas; mountains, valleys, and plains; forests, prairies, marshes, and deserts. The cultural features include cities, farms, transportation routes, and public utility facilities; political and private boundary lines. A planimetric map intended for special use may present only those features which are essential to the purpose to be served.

map: postal.*—Any *special-purpose map* published primarily for the use of the Postal Service.

Postal features are emphasized and the *maps* are composed to show postal information to the best advantage. A list of "Official Postal Maps" is published by the Post Office Department, Washington, D. C.

map: post route.*—Any *postal map* of a State, Territory, or island possession of the United States, published by the Post Office Department.

Each post route map shows all post offices within the area which it represents, indicates the method and frequency of mail supply, and gives intermediate distances on mail-supply routes. County boundaries are delineated, and all county names are shown. Also shown are the principal drainage features, but no highways, and only those railroads which carry mail.

map: reconnaissance.*—A *map* incorporating the information obtained in a *reconnaissance survey*, and data obtained from other sources.

A reconnaissance map differs from a *map* based on an *exploratory survey* by providing greater detail, selected to serve a special or general purpose. See also *reconnaissance sketch.*

map: rural delivery.*—A *postal map* of a county or locality on which roads travelled by rural carriers are distinguished, and on which are shown the location of post offices, houses, schools, and churches at the time the *map* was drawn.

Rural delivery maps are on a *scale* of about 1 *inch* equals 1 *mile*, and are divided into two groups: *county maps* and local *maps*. Where much of the area of a county is covered by rural-delivery service and adequate survey data are available, a *county map* is prepared showing all routes served from the post offices within the county. Where no *county map* has been prepared, a local *map* is constructed for each post office and the area which it serves.

map: school and bus.*—A *map* showing schools and the routes of busses serving them.

map: small scale.—See under *map scale: fractional.*

map: special-purpose.*—Any *map* designed primarily to meet specific requirements.

Usually the *map* information portrayed on a special-purpose map is emphasized by omitting or subordinating other information of a general character which is not essential or is of less importance to the purpose to be served. The special purposes for which *maps* are designed and used are numerous and are increasing with the trend toward the graphic portrayal of factual information in relation to the areas of origin or application. The *map*, in most cases, serves as a base on which special information is correlated. A word or phrase is usually employed to describe the type of information which a *map* is designed to present, whether for general or special use. Definitions for a number of special-purpose maps will be found under their specific names.

map: State.*—A *map* of the area of a State as a unit.

map: State base.*—A *base map* of the area of a State as the unit.

Emphasizing no one type of information, such a *map* is suitable for overprinting information for special purposes. The term is frequently used to describe *State maps* that are issued by agencies of the Federal Government, which are not copyrighted, and are used by government and private agencies as bases on which to portray special information.

map: static.*—A *map* that portrays information as of a single date or time.

Most *maps* are static maps, presenting information as of a given date. Static maps presenting comparable data as of different dates may be combined into a *dynamic map.*

map: topographic.*—A *map* which presents the horizontal and vertical *positions* of the features represented; distinguished from a *planimetric map* by the addition of *relief* in measurable form.

A topographic map usually shows the same features as a *planimetric map*, but uses contours or comparable symbols to show mountains, valleys, and plains; and, in the case of hydrographic charts, symbols and numbers to show depths in bodies of water.

map: Topographic Map of the United States.*—The recommended designation for the *topographic map* of the United States being prepared of quadrangle areas in atlas sheet form, chiefly by the United States Geological Survey.

This *map* portrays all basic information about location, *elevation*, and extent of physical and cultural features that are required for preliminary economic and engineering studies, and for incorporation in a base for *maps* prepared for special purposes. The use of various other terms which have been applied to this map is disapproved. These include the following: standard topographic map of the United States; general base map; general topographic map; general utility map; general utility map of the United States; general utility topographic map of the United States; and possibly others.

map: Transportation Map of the United States.*—The official designation of the *map* which is being prepared (1942) for the Public Roads Administration by the United States Geological Survey, on which are shown, as of a specified date, all improved transportation routes in the United States, including highways, railroads, air, and inland water routes.

The *map* will consist of 438 regular sheets, 26 by 36 *inches* in size, *scale* 1 : 250,000 (approximately 1 *inch* to 4 *miles*), and of special *maps* of large cities, *scale* 1 : 125,000, which may be published as separate sheets or as inserts on the regular sheets. The designation general highway and transportation map, which has been used for this *map*, is disapproved.

map projection (general).—An orderly system of lines on a plane representing a corresponding system of imaginary lines on an adopted terrestrial or celestial *datum* surface. Also the mathematical concept of such a system.

For *maps* of the earth, a projection consists of a network (*graticule*) of lines representing parallels of *latitude* and meridians of *longitude*, or of a *grid* based on such *parallels* and *meridians*. A map projection may be derived by geometrical construction or by mathematical analysis. The mathematical concept of map projection is the mathematical principle upon which it is based, expressed as formulas for computing the elements of the *projection* and tables used in constructing its graphical representation (*graticule* or *grid*). *Projections* derived by mathematical analysis are generally used for *maps* constructed with survey data. Map projections vary considerably in their characteristics, according to the qualities which they preserve in the mapping, and the methods by which this is accomplished. They are classified (1) according to the characteristics which they preserve, as *conformal, equal area, azimuthal,* etc.; (2) according to the methods used in their development, as *polyconic, gnomonic, stereographic,* etc.; and (3) according to the names of their authors, often coupled with some characteristic, as *Mercator, Bonne, Lambert with two standard parallels,* etc. The various map projections are defined under their particular designations.

map projection: Aitoff equal-area.—A *Lambert equal-area azimuthal projection* of a hemisphere converted into a *map projection* of the entire sphere by a manipulation suggested by Aitoff. It is a *projection* bounded by an ellipse in which the line representing the *equator* (major axis) is double the length of the line representing the *central meridian* (minor axis).

Tables for constructing this *projection* are prepared arithmetically from the tables for the *Lambert equal-area meridional map projection.*

map projection: Albers conical equal-area.—An *equal-area map projection* of the so-called conical type, on which *geographic meridians* are represented by straight lines which meet at a common point; this common point serves as the center of a series of arcs of circles which represent the *geographic parallels. Meridians* and *parallels* intersect in right angles. Along two selected *parallels,* called *standard parallels,* the *scale* is held exact; along the other *parallels,* the *scale* varies with the *latitude,* but is constant along any given *parallel.* Between the *standard parallels,* the meridional *scale* is too great; beyond them, too small. At any point on the *projection,* the departure from exact *scale* along a *parallel* is of the opposite sign from the departure from exact *scale* along the *meridian,* and the two are so related as to produce an *equal-area map projection.* Passing through every point are two lines of true-length *scale* which intersect in right *angles;* these are called *isoperimetric curves.*

This projection was devised by Dr. H. C. Albers in 1805.

map projection: aphylactic.—A map projection which is neither a *conformal map projection* nor an *equal-area map projection.*

map projection: authalic.—An *equal-area map projection.*

map projection: autogonal.—A *conformal map projection.*

The term autogonal, meaning *conformal,* is seldom used. Its future use should be avoided.

map projection: azimuthal (or zenithal).—A *map projection* on which the *azimuths* or *directions* of all lines radiating from a central point or pole are the same as the *azimuths* or *directions* of the corresponding lines on the sphere.

As a class, azimuthal map projections include a number of special *projections;* these are described under their particular designations. An azimuthal map projection may be constructed having two poles or points, lines from which are shown in correct *azimuth.* Such a projection is termed a *doubly azimuthal map projection.*

map projection: azimuthal equidistant.—An *azimuthal map projection* on which straight lines radiating from the center or pole of *projection* represent *great circles* in their true *azimuths* from that center, and lengths along those lines are of exact *scale.*

This projection is neither *equal-area* nor *conformal.*

map projection: Bonne.—A modified *equal-area map projection* of the so-called conical type, having lines representing a *standard parallel* and a *central meridian* intersecting near the center of the *map.* The line representing the *central meridian* (*geographic*) is straight and the *scale* along it is exact. All *geographic parallels* are represented by *arcs* of concentric circles at their true distances apart, divided to exact *scale,* and all *meridians,* except the central one, are curved lines connecting corresponding points on the *parallels.*

This *projection* is strictly *equal-area.* A particular form of this *projection* is the *Sanson-Flamsteed* or *sinusoidal map projection,* in constructing which the *equator* is used as the *standard parallel.*

map projection: Cassini.—A conventional *map projection* constructed by computing the lengths of arcs along a selected *geographic meridian* and along a *great circle* perpendicular to that *meridian,* and plotting these as *rectangular coordinates* on a plane.

map projection: conformal.—A *map projection* on which the shape of any small area of the surface mapped is preserved unchanged.

Conformal map projections are sometimes termed *orthomorphic map projections,* "orthomorphic" meaning "right-shape"; this is misleading because, if the area mapped is large, its shape will not be preserved, but only the shape of each small section of it. The exact condition for a conformal map projection is that the *scale* at any point be the same in all *directions:* the *scale* may change from point to point, but at each point it will be independent of the *azimuth.* Among the more important conformal map projections are the *Mercator,* the *stereographic,* the *transverse Mercator,* and the *Lambert conformal map projections:* the latter two are used in the *State coordinate systems.*

map projection: conic.—A *map projection* produced by projecting the *geographic meridians* and *parallels* onto a cone which is tangent to (or intersects) the surface of a sphere, and then developing the cone into a plane.

There are several methods of passing from the sphere to the cone, most of which are analytical in character and do not admit of graphical construction. For example: the *Lambert conformal conic map projection.* A *perspective map projection* from the center of the sphere would admit of graphical construction. Conic map projections may be considered as including *cylindrical map projections* when the apex of the cone is at an infinite distance from the sphere, and projections on a tangent plane when that distance is zero. Conic map projections may be illustrated with a single cone which is tangent to the sphere, or which cuts the sphere along two *parallels;* or there may be a series of tangent cones, all with apices on an extension of the axis of the sphere, at constantly increasing (or decreasing) distances from the sphere. While a cone or cones may be used in illustrating the conic map projections, care must be taken not to consider such *projections* as geometrical concepts, which most of them are not.

map projection: cylindrical.—A *map projection* produced by projecting the *geographic meridians* and *parallels* onto a cylinder which is tangent to (or intersects) the surface of a sphere, and then developing the cylinder into a plane.

There are several methods of passing from the sphere to the cylinder. Some of these are analytical in character, and do not permit of graphic construction. For example, the computation of the *Mercator map projection.* Other methods may be defined geometrically, and the *projection* constructed graphically. For example, the *cylindrical equal-spaced map projection.* Both groups of projections are spoken of as cylindrical projections. All of them may be illustrated, and some of them demonstrated, by means of a cylinder tangent to a sphere, cut along an element, and flattened into a plane.

map projection: cylindrical equal-area.—A *cylindrical map projection* upon a cylinder tangent to a sphere, showing the *geographic meridians* as a family of equal-spaced parallel straight lines perpendicular to a second family of parallel straight lines which represent the *geographic parallels,* and which are so spaced as to produce an *equal-area map projection.*

The equal-area condition preserves a constant ratio between corresponding ground and *map* areas. This *projection* must not be confused with the *Mercator projection* to which it bears some general resemblance.

map projection: cylindrical equal-spaced.—A *cylindrical map projection* upon a cylinder tangent to a sphere, showing the *geographic meridians* as a family of equal-spaced parallel straight lines perpendicular to a second family of equal-spaced parallel straight lines which represent the *geographic parallels.*

The spacing of the *parallels* need not be the same as that of the *meridians.*

map projection: doubly azimuthal.—An *azimuthal map projection* having two poles.

map projection: equal-area.—A *map projection* on which a constant ratio of areas is preserved: that is, any given part of the *map* on an equal-area projection bears the same relation to the area on the sphere which it represents as the whole *map* bears to the entire area represented.

On an equal-area map, a definite area such as a square *inch,* will represent a constant area on the sphere, no matter on what part of the *map* the square *inch* is located. Also called an equivalent map projection, equal-area map projections comprise an extensive class of particular projections, some of them of considerable usefulness.

map projection: equivalent.—See under *map projection: equal-area map projection.*

map projection: geometric.—See under *map projection: perspective map projection.*

map projection: globular.—A *map projection* representing a hemisphere, on which the *equator* and a central *geographic meridian* are represented by straight lines intersecting at right *angles;* these lines are divided into equal parts. All *meridians,* except the central one, are represented by circular *arcs* connecting points of equal division on the *equator* with the poles. The *parallels,* except the *equator,* are circular *arcs* dividing the central and extreme outer *meridians* into equal parts. The extreme outer *meridian* limits the *projection* and is a full circle.

map projection: gnomonic.—A *perspective map projection* on a plane tangent to the surface of a sphere, having the point of projection at the center of the sphere.

The *projection* is neither *conformal* nor *equal-area.* It is the only *projection* on which *great circles* on the sphere are represented as straight lines.

map projection: homolographic (homalographic).—An *equal-area map projection.*

This term is found in the designations given some particular *map projections,* such as the *Mollweide homalographic projection.*

map projection: International Map of the World.—A *modified polyconic map projection,* with two *standard meridians* along which the *scale* is held exact. The *scale* of the *map* is 1 : 1,000,000 ; the *geographic meridians* are represented by straight lines connecting corresponding points on the top and bottom *geographic parallel* lines of the *projection* which are truly divided. These lines representing the *parallels* are *arcs* of circles, not concentric.

Up to latitude 60°, each sheet of the International Map of the World includes 4° of *latitude* and 6° of *longitude;* above latitude 60°, the sheets are double width, that is, each covers 4° of *latitude* and 12° of *longitude.* It is also called the millionth scale map of the World. The *projection* was devised by Lallemand in about 1909.

map projection: Lambert azimuthal (polar).—A *Lambert equal-area map projection* with the pole of projection at the pole of the sphere, and the radii of the circles which represent the *geographic parallels* corresponding to the chords of those *parallels.*

map projection: Lambert central equivalent map projection upon the plane of the meridian.—See under *map projection: Lambert zenithal (azimuthal) equal-area.*

map projection: Lambert conformal conic.—A *conformal map projection* of the so-called conical type, on which all *geographic meridians* are represented by straight lines which meet in a common point outside the limits of the *map,* and the *geographic parallels* are represented by a series of arcs of circles having this common point for a center. *Meridians* and *parallels* intersect in right *angles,* and *angles* on the earth are correctly represented on the *projection.* This *projection* may have one *standard parallel* along which the *scale* is held exact; or there may be two such *standard parallels,* both maintaining exact scale. At any point on the *map,* the *scale* is the same in every direction. It changes along the *meridians,* and is constant along each *parallel.* Where there are two *standard parallels,* the *scale* between those *parallels* is too small; beyond them, too large.

The Lambert conformal conic map projection with two *standard parallels* is the base for the *State coordinate systems* devised by the U. S. Coast and Geodetic Survey for zones of limited north-south dimension, and indefinite east-west dimension. In those systems, the *standard parallels* are placed at distances of one-sixth the width (north-south) of the *map* from its upper and lower limits.

map projection: Lambert equal-area meridional.—See under *map projection: Lambert zenithal (azimuthal) equal-area.*

map projection: Lambert zenithal (azimuthal) equal-area.—An *azimuthal map projection* having the pole of the *projection* at the center of the area mapped. The *azimuths* of *great circles* radiating from this center (pole) are truly represented on the *map;* equal distances on those *great circles* are represented by equal linear distances on the *map,* but the *scale* along those *great circle* lines so varies with distance from the pole of the *projection,* that an *equal-area projection* is produced.

When the pole of the *projection* is on the *equator,* the *projection* is called the Lambert central equivalent projection upon the plane of the meridian; also, the Lambert equal-area meridional projection.

map projection: Mercator.—A *conformal map projection* of the so-called cylindrical type. The *equator* is represented by a straight line true to *scale;* the *geographic meridians* are represented by parallel straight lines perpendicular to the line representing the *equator;* they are spaced according to their distance apart at the *equator.* The *geographic parallels* are represented by a second system of straight lines perpendicular to the family of lines representing the *meridians,* and therefore parallel with the *equator. Conformality* is achieved by mathematical analysis, the spacing of the *parallels* being increased with increasing distance from the *equator* to conform with the expanding *scale* along the *parallels* resulting from the *meridians* being represented by parallel lines.

The Mercator map projection is considered one of the most valuable of all *map projections,* its most useful feature being that a line of constant *bearing* (*azimuth*) on a sphere is represented on the *projection* by a straight line. It is not a perspective *projection* on a cylinder, and is not developed geometrically.

map projection: Mercator equal-area.—The *sinusoidal map projection.*

map projection: Mercator, transverse.—See under *map projection: transverse Mercator map projection.*

map projection: meridional orthographic.—See under *map projection: orthographic map projection.*

map projection: millionth-scale map of the World projection.—See under *map projection: International Map of the World projection.*

map projection: modified polyconic.—A *map projection* obtained from the regular *polyconic projection* by so altering the *scale* along the central *meridian* that the *scale* is exact along two *standard meridians,* one on either side of the *central meridian* and equidistant therefrom. Also a *rectangular polyconic map projection.*

A modified polyconic map projection is used for the *International Map of the World,* the *scale* being exact along *meridians* located 2° from the *central meridian.* A modified polyconic map projection is also used for some State maps published by the U. S. Geological Survey.

map projection: Mollweide homalographic.—An *equal-area map projection* showing the *equator* and *geographic parallels* as straight lines, and the *geographic meridians* as elliptical *arcs,* with the exception of the *central meridian,* represented by a straight line, and the *meridian* 90° from the center, shown as a full circle.

This *projection* is used to show the entire surface of the earth, in which case it is bounded by an ellipse whose major axis, representing 360° of *longitude,* is double the length of its minor axis, representing 180° of *latitude.* It is also used to map a hemisphere, when the boundary of the *projection* will be a full circle.

map projection: orthembadic.—An *equal-area map projection.*

map projection: orthographic.—A *map projection* produced by straight parallel lines through points on the sphere and perpendicular to the plane of the *projection.*

The orthographic map projection corresponds to a *perspective projection* with the point of projection at an infinite distance from the sphere. As used for *maps* of the earth, this *projection* is classified as (a) a polar orthographic map projection, having the plane of the projection perpendicular to the axis of *rotation* of the earth (parallel with the plane of the *equator*); in this *projection,* the *geographic parallels* are full circles, true to *scale,* and the *geographic meridians* are straight lines; (b) the meridional orthographic map projection, having the plane of the projection parallel to the plane of some selected *meridian;* the *geographic parallels* and the central *meridian* are straight lines, the outer *meridian* is a full circle, and the other *meridians* are arcs of ellipses.

map projection: orthomorphic.—A *conformal map projection.*

Orthomorphic is a little-used term, meaning "right-shape."

map projection: perspective.—A *map projection* produced by straight lines radiating from a selected point and passing through points on the sphere to the plane of projection. Sometimes called a geometric projection.

The plane of projection is usually tangent to the sphere which represents the earth at the center of the area being mapped; the point of projection is on the diameter of the sphere which passes through the point of tangency, and at some selected point on that diameter. If the point of projection is at the center of the sphere, there results the *gnomonic map projection;* if at the opposite end of the diameter from the point of tangency of the plane of projection, it becomes a *stereographic map projection;* if at an infinite distance, an *orthographic map projection.*

map projection: perspective map projection upon a tangent cylinder.—A *cylindrical map projection* upon a cylinder tangent to a sphere, by means of straight lines radiating from the center of the sphere.

The *geographic meridians* are represented by a family of equally spaced parallel straight lines, perpendicular to a second family of parallel straight lines which represent the *geographic parallels.* The spacing, with respect to the *equator,* of the lines which represent the *parallels,* increases as the tangent of the *latitude;* the line representing 90° *latitude* is at an infinite distance from the line which represents the *equator.* This *projection* must not be confused with the *Mercator projection* to which it bears some general resemblance.

map projection: polar orthographic.—See under *map projection: orthographic map projection.*

map projection: polyconic.—A *map projection* having the central *geographic meridian* represented by a straight line, along which the spacing for lines representing the *geographic parallels* is proportional to the distances apart of the *parallels;* the *parallels* are represented by *arcs* of circles which are not concentric, but whose centers lie on the line representing the central *meridian,* and whose radii are determined by the lengths of the elements of cones which are tangent along the *parallels.* All *meridians* except the central one are curved.

This *projection* is neither *conformal* nor *equal-area,* but it has been much used for *maps* of small areas because of the ease with which it can be constructed. It is the *map projection* used for the *Topographic Map of the United States* (U. S. Geological Survey), and in a modified form is used for *maps* of large areas. It was devised by F. R. Hassler, organizer and first superintendent of the U. S. Coast Survey (now the U. S. Coast and Geodetic Survey).

map projection: polyconic, modified.—See under *map projection: modified polyconic map projection.*

map projection: polyconic, rectangular.—See under *map projection: rectangular polyconic map projection.*

map projection: polyconic, transverse.—See under *map projection: transverse polyconic map projection.*

map projection: rectangular polyconic.—A modified *polyconic map projection* having a line representing a *standard parallel* divided to exact *scale,* through whose division points pass the lines representing the *geographic meridians,* intersecting the lines which represent the *geographic parallels* in right *angles.*

map projection: Sanson-Flamsteed.—The *sinusoidal map projection.*

map projection: sinusoidal.—A particular type of the *Bonne map projection,* employing the *equator* as the *standard parallel,* and showing all *geographic paral-*

lels as truly spaced parallel straight lines, along which exact *scale* is preserved.

This is an *equal-area map projection.* It is also known as the *Mercator equal-area map projection* and as the Sanson-Flamsteed map projection. It was employed in the Mercator-Hondius atlases as early as 1606.

map projection: stereographic.—A *perspective map projection* having the point of projection at the opposite end of the diameter of the sphere from the point of tangency of the plane of projection.

It is *conformal,* and is the only *azimuthal map projection* having that quality. It is one of the most widely known of all *map projections,* and has been much used for *maps* of a hemisphere. When the center of the projection is located at a pole of the sphere, it is called a stereographic polar projection; when on the *equator,* a stereographic meridional projection; and when on some other selected *parallel* of *latitude,* a stereographic horizon projection. The stereographic map projection dates back to the days of ancient Greece, having been used by Hipparchus (160–125 B. C.).

map projection: transverse Mercator.—A *map projection* of the so-called cylindrical type, being in principle equivalent to the regular *Mercator map projection* turned (transversed) 90° in azimuth.

In this *projection,* the *central meridian* is represented by a straight line, corresponding to the line which represents the *equator* on the regular *Mercator map projection.* Neither the *geographic meridians,* except the central *meridian,* nor the *geodetic parallels,* except the *equator* (if shown) are represented by straight lines. It is a *conformal projection,* and is the base used in the *State plane-coordinate systems* for the *grids* of those zones whose greater dimension is in a north-and-south direction.

map projection: transverse polyconic.—A *polyconic map projection* which is turned (transversed) 90° in *azimuth* by substituting for the *central meridian,* a *great circle* perpendicular to the *geographic meridian* to provide a control axis for the *projection,* along which axis will lie the centers of the circular *arcs* representing lines of tangency of cones with the surface of the sphere.

This is a difficult projection to construct, but it is useful in showing, with comparatively small distortion, a narrow area of large east-and-west dimension.

map projection: Werner.—A particular case of the *Bonne map projection,* in which the *standard parallel* is at the pole, and the tangent cone becomes a tangent plane.

Any one *geographic meridian* is chosen as the *central meridian* and represented by a straight line, divided to exact *scale.* The *geographic parallels* are represented by circular *arcs,* also divided to exact *scale,* and the other *meridians* are curved lines.

map scale.*—The relationship which exists between a distance on a *map* and the corresponding distance on the earth.

A map scale may be expressed as an equivalence, as a numerical fraction or ratio, or shown graphically. On large-scale *maps,* the distance on the earth is on a designated *datum,* as *sea level* or ground level.

map scale: equivalent.*—An equivalent scale is the relationship which a small distance on the *map* bears to the corresponding distance on the earth, expressed as an equivalence.

Usually, but not necessarily, the equivalence is expressed in different specified units; for example, 1 *inch* (on the *map*) equals 1 *mile* (on the ground). Of the abbreviated forms, "*inch* to the *mile*" and "*mile* to the *inch,*" the former is preferred. Infrequently called verbal scale.

map scale: fractional.*—A fractional scale is the ratio which any small distance on the *map* bears to the corresponding distance on the earth.

It may be written in the form of a fraction: 1/10,000; or as a proportion, 1 : 10,000. Very infrequently called numerical scale, natural scale, and linear scale. These terms are not recommended.

Fractional scales are representative in any linear units. Usually the term representative fraction, however, is applied to a fractional scale whose numerator is unity (1) and is frequently referred to as the "R. F." of the *map.*

For a map of large scale, the representative fraction has a small denominator; for a map of small scale, the representative fraction has a large denominator.

map scale: graphic (or bar).*—A line on a *map* subdivided and marked with the distance which each of its parts represents on the earth.

mapping plan: national.—A program for the completion of the *Topographic Map of the United States.*

The Federal Board of Surveys and Maps in 1934 and again in 1936 adopted plans which proposed a time schedule for the completion of the *control surveys* and topographic mapping of the United States, and a fiscal budget for those operations. The *Temple Act* (H. R. 4522, 68th Congress), approved 1925, was the forerunner of the various national mapping plans.

mean: arithmetical.—See under *arithmetical mean.*

mean: weighted.—See under *weighted mean.*

mean error.—See under *error.*

mean high water.—The mean height of all *high waters* at a particular point or station over a considerable period of time.

For tidal waters, the cycle of change covers a period of about 18.6 years, and the mean high water is the mean of all *high waters* for that period. For any body of water, it is the mean of all *high waters* over a period of time of such length that increasing its length does not appreciably change the mean.

mean low water.—The mean height of all *low waters* at a particular point or station over a considerable period of time.

For tidal waters, the cycle of change covers a period of about 18.6 *years,* and the mean low water is the mean of all *low waters* for that period. For any body of water, it is the mean of all *low waters* over a period of time of such length that increasing its length does not appreciably change the mean.

mean of the errors.—See under *error.*

mean sea level.—The average height of the sea for all stages of the tide.

Mean sea level is obtained by averaging observed hourly heights of the sea on the open coast or in adjacent waters having free access to the sea, the average being taken over a considerable period of time.

mean solar day.—See under *day.*

mean solar time.—See under *time.*

mean-square error.—See under *error.*

mean sun.—See under *time: mean solar time.*

meander line (United States public-land surveys).*—A *traverse* of the margin of a body of water.

Meander lines are not run as *boundary lines,* but are run for the purpose of defining the sinuosities of the bank or shore line, and for ascertaining the quantity of land remaining after segregation of the water area. In General Land Office practice, only permanent natural bodies of water are meandered.

memorial (United States public-land surveys).*—A durable article deposited in the ground at the position of a *corner,* to perpetuate that position should the *monument* be removed or destroyed.

The memorial is usually deposited at the base of the *monument* and may consist of anything durable, such as glass or stoneware, a marked stone, charred stake, or quantity of charcoal.

Mendenhall pendulum.—See under *pendulum.*

Mercator bearing.—The *bearing* of a straight line plotted on a Mercator *chart* as taken from the *chart.*

It is the *angle* between the straight line and the chart meridians.

Mercator equal-area map projection.—See under *map projection.*

Mercator map projection.—See under *map projection.*

Mercator map projection, transverse.—See under *map projection.*

Mercator track.—A *rhumb line* on a *map* or *chart* constructed on a *Mercator map projection.*

mercury barometer.—See under *barometer.*

meridian (general).—A north-south line from which *longitudes* (or *departures*) and *azimuths* are reckoned; or a plane, normal to the *geoid* or *spheroid,* defining such a line.

meridian: astronomic.—A line on the surface of the earth having the same *astronomic longitude* at every point.

Because the *deflection of the vertical* is not the same at all points, an astronomic meridian is an irregular line, not lying in a single plane. The astronomic meridian and the line whose *astronomic azimuth* at every point is south or north (0° or 180°) are not necessarily coincident, although in *land surveying* the term astronomic meridian is sometimes applied to the north-south line which has its initial on a prescribed astronomic meridian.

meridian: auxiliary guide (United States public-land surveys).*—A new *guide meridian* established, when required, for control purposes where the original *guide meridians* were placed at greater intervals than 24 miles.

Auxiliary guide meridians may be required to limit *errors* of old or to control new surveys; they are surveyed in all respects like regular *guide meridians,* and may be assigned a local name, as "Grass Valley Guide Meridian," or "Twelfth Auxiliary Guide Meridian West."

meridian: celestial.—The *hour circle* which contains the *zenith.* Also defined as the *vertical circle* which contains the celestial pole.

The plane of the celestial meridian is parallel with, but, due to the *deflection of the plumb line,* usually does not contain the axis of *rotation* of the earth; its intersection with the plane of the horizon is the *meridian line* used in *plane surveying.*

meridian: central (State coordinate system).—The *meridian* used as Axis of *Y* for computing projection tables for a *State coordinate system.*

The central meridian of the system usually passes close to the center of figure of the area or zone for which the tables are computed, but to avoid the use of negative values, is given a large positive abscissa, thus requiring that a large constant quantity be added to all *x coordinates.*

meridian: convergence of meridians.—The angular drawing together of the *geographic meridians* in passing from the *equator* to the poles.

At the *equator,* all *meridians* are mutually parallel; passing from the *equator,* they converge until they meet at the poles, intersecting in *angles* that are equal to their differences of *longitude.* The term convergence of meridians is used to designate also the relative difference of direction of *meridians* at specific points on the *meridians.* Thus, for a *geodetic line,* the *azimuth* at one end differs from the *azimuth* at the other end by 180° plus or minus the amount of the convergence of the meridians at the end points.

meridian: double meridian distance (plane surveying).—The sum of the *meridian distances* of the ends of a survey line.

In practice, assuming a closed *traverse* with the *meridian* of reference passing through the initial station, the double meridian distance (D. M. D.) of the first *course* (line) is equal to the *departure* of that *course;* the double meridian distance of each succeeding *course* (line) is equal to the double meridian distance of the preceding *course,* plus the *departure* of the preceding *course,* plus the *departure* of the *course* itself.

Meridian: Ferro (Hierro).—The conventional Ferro Meridian is 17°37′45″ west of Greenwich, being originally chosen exactly 20° west of the Paris Meridian.

Many European geographers used to reckon longitude from this meridian which was taken as the dividing line between the Eastern and Western Hemispheres.

The conventional Ferro Meridian does not coincide with the neighboring island of the same name.

meridian: geodetic.—A line on a *spheroid* which has the same *geodetic longitude* at every point.

If the *spheroid* is an ellipsoid of revolution, as used for purposes of *triangulation,* a geodetic meridian is an ellipse whose plane contains the minor axis of the *spheroid,* and whose *geodetic azimuth* at every point is 0° (south) or 180° (north).

meridian: geographic.—A general term, applying alike to an *astronomic* or a *geodetic meridian.*

Meridian: Greenwich.—The *astronomic meridian* through the center of the *transit instrument* of the Observatory of Greenwich, England.

meridian: grid.—A line through a point parallel to the central *meridian* or Axis of Y of a system of *plane-rectangular coordinates.*

meridian: guide (United States public-land surveys).*—An auxiliary governing line projected north along an *astronomical meridian,* from points established on the *base line* or a *standard parallel,* usually at intervals of 24 miles east or west of the *principal meridian,* on which *township, section,* and *quarter-section corners* are established.

meridian: magnetic.—The vertical plane in which a freely suspended, symmetrically magnetized needle, influenced by no transient artificial magnetic disturbance, will come to rest. Also, a curve on the earth's surface tangent to such a plane at each place it touches.

meridian: prime.—An initial or zero *meridian* from which other *meridians* are reckoned.

At an International Meridian Conference held in Washington, D. C., in 1884, the adoption of the *Greenwich Meridian* as the prime meridian for the earth was approved by the representatives of 22 governments, there being one objector, while two representatives abstained from voting. See also, *meridian: Washington Meridian.*

meridian: principal (United States public-land surveys).*—A line extending north and south along the *astronomical meridian* passing through the *initial point,* along which *township, section,* and *quarter-section corners* are established.

The principal meridian is the line from which is initiated the *survey* of the *township* boundaries along the *parallels.* Auxiliary meridional governing lines, known as *guide meridians,* are established along the *astronomical meridian* usually at intervals of 24 *miles* east or west of the principal meridian. In some of the very early surveys the principal meridian is referred to as the "basis meridian."

meridian: terrestrial.—An *astronomic meridian.*

For particular use, the term *astronomic meridian* is preferred over terrestrial meridian. For general use, the preferred term is *geographic meridian,* which applies alike to *astronomic* and *geodetic meridians.*

Meridian: Washington.—The *astronomic meridian* through the center of the dome of the old Naval Observatory in Washington, D. C.

The Washington Meridian was used in defining the meridional boundaries of several western States. An Act of Congress, September 28, 1850, provided "That hereafter the meridian of the observatory at Washington shall be adopted and used as the American meridian for astronomic purposes and . . . Greenwich for nautical purposes." This Act was repealed August 22, 1912. See *meridian: Greenwich Meridian.*

meridian distance (astronomy).—The *hour angle* of a celestial body when close to but not exactly on the *astronomic meridian.*

This term appears to be limited to computation forms where it is used to designate the *hour angle* of a star observed slightly off the *meridian* in making *latitude* observations by the *zenith telescope method.*

meridian distance (plane surveying).—The perpendicular distance in a horizontal plane of a point from a *meridian* of reference.

The difference of the meridian distances of the ends of a line is called the *departure* of the line.

meridian line (plane surveying).—The line of intersection of the plane of the *celestial meridian* and the plane of the *horizon.*

It is a horizontal *direction* used in surveying; its *astronomical azimuth* is 0° or 180°.

meridian telescope.—A portable instrument so designed that it can be used as an *astronomical transit,* or quickly converted for use as a *zenith telescope.*

The meridian telescope has a base which is divided horizontally into two parts; the upper part rotates upon the lower part about a short vertical axis. The upper part carries two folding uprights (standards) with wyes which support the horizontal axis of the *telescope.* The *telescope* is equipped with a *vertical circle* to which is attached a sensitive *spirit level.* The *telescope* has a fixed *reticle* for use in making *time* observations, and an *ocular micrometer* for *latitude* work. When used as an *astronomical transit,* the two parts of the base are clamped together. It is converted for use as a *zenith telescope* by removing the clamps, and permitting the upper part to rotate through 180°, stops being placed to bring the *telescope* to rest in the plane of the *meridian.* The meridian telescope is also known as the Davidson meridian instrument, after George Davidson of the U. S. Coast Survey, who designed it in 1858.

meridional orthographic map projection.—See under *map projection: orthographic.*

meridional parts: table of.—A table listing lengths of the *meridian* from the *equator* to the various *parallels* of *latitude* increased in the proportion required to show lengths along the *parallels* equal to the corresponding length along the *equator.*

On a sphere, at the *equator,* the length of a minute of *longitude* is equal to the length of a minute of *latitude,* but on receding from the *equator* and approaching the poles, the minutes of *latitude* remain always the same, while the length of a minute of *longitude* steadily decreases.

Because the earth is not a sphere, the above conditions do not exactly fit it. However, in the *Mercator map projection,* the minutes of *longitude* are made to appear of the same length for all *latitudes,* and it becomes necessary in order to preserve existing proportions between lengths of the *parallel* and of the *meridian* at various *latitudes* to increase the lengths along the *meridian,* such increase being greater and greater, the higher the *latitude.* The length of the *meridian,* thus increased, constitutes the number of the meridional part corresponding to that *latitude.*

Such a table, found in books on navigation, affords facilities for constructing a *Mercator map projection,* and for solving problems in Mercator sailing. A close value of a meridional part on an ellipsoid for a given *geographic latitude* is obtained by computing the meridional part for the corresponding *geocentric latitude.*

metallic thermometer.—A device utilizing the difference of the *coefficients of thermal expansion* of the metals of which it is constructed to determine the temperature of those metals.

The measuring rods used on the *base lines* of the Dunkirk-Barcelona *triangulation* (1791–2) were themselves metallic thermometers. Each rod was composed of a strip of copper lying on a strip of platinum; these were fastened together at one end, leaving their other ends free to move in response to changes in their temperature. A scale on the free end of the copper strip was read by means of a *vernier* on the free end of the platinum strip, thus affording a means of determining the temperature of the bars. This apparatus was constructed by Borda. See *Borda scale.*

meter; metre.—A *unit* of length equivalent in the United States to 39.37 *inches* exactly.

The meter was originally defined (in 1791) as the equivalent of one ten-millionth of the length of the *quadrant* of the *meridian* through Paris (from the *equator* to the North Pole). A *unit* derived by this proportion from several *arcs* of *triangulation* became known as the *French legal meter,* the *standard* constructed to represent it being known as the *Metre des Archives.* Under an International Convention in 1875, the *international meter* was derived from the *Metre des Archives,* and the *International Prototype Meter* adopted as a *standard* by the subscribing governments. Copies of this new *standard,* of the same form and alloy as the original, were furnished the governments which subscribed to the Convention of 1875. The United States received two such copies in 1889, one of which was adopted as the national standard of length, and is known as the *National Prototype Meter.* These two copies are in the custody of the National Bureau of Standards, Washington, D. C. The equivalence, 1 meter equals 39.37 *inches* exactly, was prescribed by Act of Congress in 1866.

At the Seventh General (International) Conference on Weights and Measures (Paris, 1927), the following supplementary value of the meter was temporarily adopted:

1 meter = 1,553,164.13 wave-lengths of the red ray of Cadmium vapor, in dry air, at a temperature 15° Centigrade, at a pressure of 760 millimeters of mercury and under normal conditions of *gravity.* Since this determination was made, the necessity for mak-

ing several slight corrections to the result has been realized and the following relation obtained:

1 meter = 1,553,163.7 wave-lengths. The earlier value continues the official value.

meter: Committee Meter.—An iron bar *standard* of one *meter* length, which was brought to America in 1805 by Ferdinand R. Hassler, the first superintendent of the U. S. Coast Survey.

It was one of sixteen such bars which were standardized by the Committee on Weights and Measures in Paris in 1799. It served as the *standard* of length for *geodetic surveys* in this country until 1889 or 1890, when it was replaced by the present *standard*, the *National Prototype Meter*. The Committee Meter was presented by Hassler to the American Philosophical Society in Philadelphia.

meter: French legal.—The *unit* of length defined by the *Metre des Archives*. See *meter; metre.*

meter: international.—The *unit* of length defined by the *International Prototype Meter*. See *meter; metre.*

meter: International Prototype Meter.—A bar *standard* of platinum-iridium which defines the length of the *unit*, the *international meter*.

The International Prototype Meter was produced under the authority and direction of an International Conference held in Paris in 1875, signed by the representatives of 17 governments. It is a line *standard* and was derived from the *Metre des Archives*, an end *standard*. The International Prototype Meter and two other similar bars of the same origin are maintained at the International Bureau of Weights and Measures at Sevres, France, and are used in standardizing the representative copies which were supplied to the governments supporting the Convention of 1875. As a signatory of that convention, the United States received two such representative copies, one of which is known as the *National Prototype Meter*. These copies are in the custody of the National Bureau of Standards, Washington, D. C.

meter: Metre des Archives.—A bar *standard* defining the length of the *French legal meter*.

It is an end *standard*, of platinum, and was used in determining a length for the *International Prototype Meter*. It is now preserved as a museum piece in the International Bureau of Weights and Measures, Sevres, France. See *meter; metre.*

meter: National Prototype Meter.—A *standard* copy of the *International Prototype Meter* which defines the *international meter* as a *unit* of length.

It is a bar of platinum-iridium, of the same form and alloy as the *International Prototype Meter*, and one of the pair received in 1889 from the International Bureau of Weights and Measures, Sevres, France. These bars are in the custody of the National Bureau of Standards, Washington, D. C. They are line *standards*.

method of repetitions, determination of astronomic azimuth.—See under *azimuth.*

micrometer: correction for run of.—A *correction* applied to an observed reading of a graduated circle made with a *micrometer microscope* to compensate for *run of micrometer*.

Specific application of this correction is not required in present-day observations with a *direction instrument*, being neutralized by instrumental adjustment and method of observing.

micrometer: eyepiece.—See under *micrometer: ocular.*

micrometer: filar.—A device attached to a *telescope* or microscope, consisting of a wire thread (filament) connected with a screw in such manner that as the screw is turned, the wire moves through a continuous succession of parallel positions, all in the focal plane of the instrument.

The screw has a fine thread, accurately cut, and calibrated in terms of angular measure (seconds of *arc*): one turn of the screw moves the wire a distance that may be expressed as a certain number of seconds, corresponding to the *angle* at the instrument between two objects which would be bisected by the wire in its two positions. Any small *angle* may be measured by noting the number of turns and fractions of a turn of the screw which are made in moving the wire from one bisection to another, and multiplying this by the value of one turn. See *micrometer: ocular; micrometer microscope.*

micrometer: impersonal.—See under *micrometer: transit micrometer.*

micrometer: ocular.—A *filar micrometer* so placed that its wire moves in the principal focal plane of a *telescope*. Also termed an eyepiece micrometer.

The ocular micrometer is used in *surveying* and astronomic work for making accurate and precise measures of small *angles* between lines to objects viewed with the *telescope*. It is mounted in a frame which is perpendicular to the longitudinal axis of the *telescope*, and can be rotated to measure an *angle* in any plane containing that axis. It may be adjusted to measure *angles* in a plane which contains the longitudinal and horizontal axes of the *telescope*, as in observing *time* using the *transit micrometer* (which is an ocular micrometer); or it may be adjusted to measure *angles* in a plane containing the longitudinal axis but perpendicular to the *horizontal axis* of the *telescope*, as in observing *latitude* with the *zenith telescope*.

micrometer: run of.—The difference in seconds of *arc* between the intended value of one complete turn of a micrometer screw used in reading a graduated circle and its actual value as determined by measuring with the micrometer the space between two adjacent *graduation* marks on the circle.

This quantity has sometimes been called the error of run, but the above designation is more generally used and is preferred. It is an *error*, but is kept quite small by instrumental adjustment, and its effect on observed values is neutralized by methods of observing. See *micrometer: correction for run (of micrometer).*

micrometer: transit.—A form of registering micrometer with its movable wire placed in the *focal plane* of an *astronomic transit* and at right *angles* to the direction of motion of the image of a star which is observed at or near *culmination*. Certain contact points on the micrometer head (wheel) serve to make an electric circuit as they pass a fixed contact spring, thus producing a record on the *chronograph* sheet at each separate instant when the micrometer wire reaches a position corresponding to a contact.

The transit micrometer is also called an impersonal micrometer, because it almost completely eliminates the effect of *personal equation* from time observations made with it.

micrometer method, determination of astronomic azimuth.—See under *azimuth.*

micrometer microscope.—A *filar micrometer* so placed that its wire moves in the focal plane of a microscope.

Using a micrometer microscope, readings of a graduated circle are made with a high degree of *accuracy* and *precision*. The micrometer is so adjusted that an even number of turns will carry the wire exactly the distance between adjacent *graduation* marks on the circle. Any intermediate point on the circle may then be read with the micrometer by interpolation. If the adjustment is imperfect, the amount by which an even number of turns of the screw fails to carry the wire exactly the distance between *graduation* marks is treated as a *systematic error* and is termed by preference the *run of micrometer*. The designation error of run of micrometer is also applied to it.

middle ordinate.—The distance from the middle point of a chord to the middle point of the corresponding circular *arc*.

middle point (M. P.).—That point on a circular curve which is at the same distance from both ends of the curve.

mil.—See under *angle (general)*, (e).

mile.—A unit of distance. See under *mile: nautical mile; statute mile.*

The term mile is derived from the Latin "mille," meaning a thousand, and applied to the double *pace* of 5 *feet*. The mile of the Romans was 5,000 *feet*, a value that suffered many changes as the mile came into use among the western nations.

mile: geographical.—See under *mile: nautical mile.*

mile: nautical.—A length of *mile* used in ocean navigation.

The United States nautical mile is defined as equal to one-sixtieth of a degree of a *great circle* on a sphere whose surface is equal to the surface of the earth. Its value, calculated for the *Clarke spheroid of 1866*, is 1,853.248 *meters* (6,080.20 feet). The international nautical mile is 1,852 *meters* (6,076.10 feet). The United States nautical mile is also called a sea mile or a geographical mile, and may be taken as equal to the length of a minute of *arc* along the *equator* or a minute of *latitude* on the *map* which is being measured.

mile: sea.—See under *mile: nautical mile.*

mile: statute.—A length of *mile* used in distance measurements on land, and equal to 5,280 feet (1,609.35 *meters*).

military grid.—A quadrillage or system of squares determined by rectangular coordinates on a spherical *map projection*, referred to one origin, and extended over the whole area of the original *map projection*.

A military grid is designed to make data presented on a *map* more readily available for military purposes.

milligal.—See *gal.*

millionth-scale map of the World.—An informal designation for the *International Map of the World* which is being constructed on a scale of 1 : 1,000,000.

minus sight.—See under *foresight (leveling)*.

missing triangle (pendulum).—A triangle which represents the failure of the two sides of a *knife edge* to reach a perfect intersection in a geometric line.

In the most nearly perfect *knife edge* that can be made, the two planes of the *knife edge* will not quite meet. The base of the missing triangle which is thereby formed is found by optical means to be less than one micron.

modified polyconic map projection.—See under *map projection.*

Molitor leveling rod.—See under *leveling rod: Molitor precise leveling rod.*

Mollweide homalographic map projection.—See under *map projection.*

month.—A measure of *time* based on the motion of the moon in its orbit.

There are various kinds of month defined by the references which determine them. See *month: anomalistic month; calendar month; nodical month; sidereal month; synodical month.*

month: anomalistic.—The interval of *time* between two successive passages of the moon in her orbit through *perigee.*

The length of the anomalistic month is 27.55455 *mean solar days.*

month: calendar.—A division of the *year* as determined by a calendar, approximately one-twelfth of a *year* in length.

While arbitrary in character, the calendar month is based roughly on the *synodical month.* The calendar month ranges in length from 28 to 31 *mean solar days.*

month: lunar.—A *synodical month.*

month: nodical.—The interval of *time* between two successive passages of the moon through the same node of her orbit.

The length of the nodical month averages 27.21222 *mean solar days.*

month: sidereal.—The interval of *time* it takes the moon to make its *revolution* from a given star back to the same star again.

The sidereal month may be measured by the interval of *time* it takes the moon to revolve from a given *celestial longitude* back to the same longitude again, reckoned from a fixed *equinox.* The length of the sidereal month averages 27.321661 *mean solar days.* Because of perturbations, the actual length varies some 7 hours. The difference between the lengths of the sidereal and *tropical months* is due to the *precession of the equinoxes.* The above definition is not precise : the moon will not return to the same *declination* at the end of a sidereal month.

month: synodical (ordinary).—The interval of *time* between two successive conjunctions (new moons) or oppositions (full moons) of the moon. Also called a lunation.

The length of the synodical month is 29.530588 *mean solar days.* It is the *month* intended when a *lunar month* is specified. Synodical is preferred over synodic, which is sometimes used.

month: tropical.—The interval of *time* it takes the moon to make its *revolution* from a given *equinox* back to the same *equinox* again.

The tropical month may be measured by the interval of *time* it takes the moon to revolve from a given *celestial longitude* back to the same *longitude* again, reckoned from an *equinox* affected by precession. The length of the tropical month averages 27.321582 *mean solar days.* Because of perturbations, the actual length varies some 7 hours. The difference between the lengths of the *sidereal* and tropical months is due to the *precession of the equinoxes.*

monument.—Any material object or collection of objects which indicates the position on the ground of a survey station or land corner.

The term monument may include the mark at the center or station and all marks which are directly related to it by surveying processes and are aids in its recovery and identification. In *land surveying* the term may include also objects such as roads, ditches, fences, etc., which form a bound for the land. It is well therefore in treating of survey stations to use the designations station mark or center mark, *reference mark,* and *witness mark* for the separate marks; and in *land surveying,* unless monument is used in a general sense, to use the particular terms, *corner,* or corner mark, *reference mark,* and *witness mark.* Note that a *witness mark* has more authority in *land surveying* than in *control surveying.*

monument (United States public-land surveys).*—A physical structure which marks the location of a *corner* or other survey point.

Monument and *corner* are not synonomous, though the two terms are often used largely in the same sense.

mountain barometer.—See under *barometer.*

multiple level line.—See under *level line.*

N

nadir.—The point where the direction of the plumb-line extended below the *horizon* meets the *celestial sphere.*

The nadir is directly opposite the *zenith.*

national control survey nets.—See under *control.*

national mapping plan.—See under *mapping plan.*

National Prototype Meter.—See under *meter.*

nautical chart.—See under *chart.*

nautical mile.—See under *mile.*

navigable waters.—The waters which are used or are susceptible of being used, in their ordinary condition or after reasonable improvements have been made, as highways for commerce in the customary means of travel on water.

net (leveling).—A series of *lines of levels* which have been interconnected in such a manner that closed loops or *circuits* have been formed.

network (leveling).—See *net (leveling).*

New England datum.—See under *datum.*

Newtonian constant of gravitation.—See under *gravitation: constant of gravitation.*

New York leveling rod.—See under *leveling rod.*

nodal point of a lens, first and second.—The intersections of the *principal axis* with the (produced) incident and emergent parts, respectively, of any ray having those parts parallel to each other.

nodical month.—See under *month.*

nomogram.—A chart or drawing showing, to scale, the relationship between known and dependent variables, in such manner that a value of a dependent variable corresponding to a stated value of the known variable can be scaled from the chart.

Nomograms are used in the graphical solution of formulas where the *precision* of the result does not require numerical computation.

nomograph.—A *nomogram.*

Although the term nomograph is given the preference in Webster's International Dictionary, *nomogram* is preferred in the several texts consulted.

normal.—In general, a straight line perpendicular to a surface or to another line. Also, a condition of being perpendicular to a surface or line. In *geodesy,* a straight line perpendicular to the surface of the *spheroid.*

While the term normal is correctly used to designate also a line perpendicular to the surface of the *geoid,* the term *vertical* is preferred for such a line. The term normal, in general, applies to a line of unspecified length, but some writers use it to designate the length of line between the surface of the *spheroid* and the minor axis, giving it the designation "N". To this particular line the French apply the term "great normal". This length is the radius of curvature of a *prime vertical* section of the spheroid.

normal equation.—See under *equation.*

normal section azimuth.—See under *azimuth: normal section.*

normal section line.—A line on the surface of the *spheroid,* connecting two points on that surface, and traced by a plane containing the *normal* at one point and passing through the other point.

There can be two such lines connecting any two points (as A and B) on the *spheroid:* one defined by the plane containing the *normal* at A and passing through B; the other containing the *normal* at B and passing through A. Where the two points are on the same parallel of *latitude* or on the same meridian of *longitude,* the two lines will be coincident.

North American datum.—See under *datum.*

North American datum of 1927.—See under *datum.*

north star.—See under *Polaris.*

northing.—See under *latitude: difference of latitude (plane surveying).*

nutation.—A periodic motion of the celestial poles.

O

obliquity of the ecliptic.—See under *ecliptic.*

obliterated corner (United States public-land surveys).—See under *corner.*

observation: direct.—A measure of the quantity whose value is desired.

Example: a single measure of a *horizontal angle.*

observation: indirect.—A measure of a quantity which is a function of the quantity or quantities whose value is desired.

Example: an observed difference of *elevation* with a *spirit level,* used to obtain the *elevation* of a *bench mark.*

observation equation.—See under *equation.*

observed angle.—See under *angle.*

observed value.—See under *value.*

occultation.—The disappearance of a body behind another body of larger apparent size.

When the moon passes between the observer and a star, the star is said to be occulted. The three associated terms, occultation, *eclipse,* and *transit,* are exemplified by the motions of the satellites of Jupiter. An *eclipse* occurs when a satellite passes into the shadow cast by the planet; an occultation occurs when a satellite passes directly behind the planet, so that it could not be seen even if it were illuminated; and a transit occurs when a satellite passes between the observer and the planet, showing against the disk of the planet. The occultation of stars by the moon furnishes a method of determining *latitude* and *longitude* used in early *surveys,* particularly those of an exploratory character.

odometer.—A *revolution* counter which is attached to the wheel of a vehicle and registers the number of turns made by the wheel in traveling over the ground.

An odometer is used in obtaining an *approximate* value of the distance traveled, the number of *revolutions* being multiplied by the circumference of the wheel. See *passometer.*

offset.—A short line perpendicular to a surveyed line, measured to a line or point for which data are desired, thus locating the second line or point with reference to the first or surveyed line. An offset is also a jog in a *survey* or other line, the line having approximately the same direction both before and after passing the jog.

Offsets are measured from a surveyed line or lines to the edges of an irregular-shaped body of water, or to any irregular line which it is desired to locate.

offset (United States public-land surveys).—A perpendicular distance measured from a *great circle line* to a *parallel* of *latitude,* to locate a *section corner* on that *parallel.* See *secant method (United States public-land surveys); tangent method (United States public-land surveys).*

offset line.—A supplementary line close to and roughly parallel with a main line, to which it is referred by measured *offsets.*

Where the line for which data are desired is in such position that it is difficult to measure over it, the required data are obtained by running an offset line in a convenient location and measuring *offsets* from it to salient points on the other line.

Old Hawaiian datum.—See under *datum.*

optical axis (lens).—See under *principal axis.*

optical base-line measuring apparatus.—See under *base apparatus.*

optical parallax.—See under *parallax: instrumental parallax.*

optical square.—A small hand instrument used in setting off a right *angle.*

One form of optical square employs two plane mirrors placed at an *angle* of 45° to each other. In use, one object is sighted direct, and another object is so placed that its twice-reflected image appears directly in line with the first object. The lines to the point of observation from the two observed objects will then meet in a right *angle.* In another form of optical square, a single plane mirror is so placed that it makes an *angle* of 45° with a sighting line; one object is sighted direct, and the other so placed that its reflected image is seen also in the sighting line.

optical vernier.—See under *vernier.*

orientation.—The act of establishing correct relationship in direction with reference to the points of the compass; the state of being in correct relationship in direction with reference to the points of the compass.

A *map* is in orientation when the *map* symbols are parallel with their corresponding ground features. A *plane table* is in orientation when lines connecting positions on the *plane table* sheet are parallel with the lines connecting the corresponding ground objects. A surveyor's *transit* is in orientation if the horizontal circle reads 0° when the *line of collimation* is parallel to the direction it had at an earlier (initial) position of the instrument, or to a standard line of reference. If the line of reference is a *meridian,* the circle will show *azimuths* referred to that *meridian.*

A photograph is in orientation when it correctly presents the perspective view of the ground directly in front of the observer; or when images on the photograph appear in the same direction from the point of observation as do the corresponding map symbols.

origin of coordinates.—See under *coordinates.*

orthodrome.—An obsolete term descriptive of a *great circle.*

orthographic map projection.—See under *map projection.*

orthometric correction.—See under *correction.*

orthometric elevation.—See under *elevation.*

orthomorphic map projection.—See under *map projection.*

oscillation (pendulum).—A double motion, one in each direction, of a *pendulum.* An oscillation is composed of two consecutive *vibrations.*

In some texts, a *vibration* is called a simple oscillation, and an oscillation is called a double or complete oscillation.

outcrop.—The exposed portion of a rock stratum or vein at the surface of the earth.

In describing a *survey* station, the exposed portion of a large boulder is sometimes mistaken for an outcrop.

outline map.—See under *map.*

P

pace.—The distance a pedestrian moves in taking one step.

See *stride.*

Panama-Colon datum.—See under *datum.*

parallactic error.—See under *error.*

parallax (general).—1. The difference in the direction of an object as seen by an observer and as seen from some standard point of reference. 2. An apparent displacement of the position of the body with respect to a reference point or system caused by a shift in the point of observation.

Parallax enters into many problems of astronomy, *surveying,* and mapping. It is an inclusive term, and wherever clarity demands, it should be accompanied by a defining adjective. See *annual parallax, diurnal parallax, secular parallax, instrumental parallax,* and proposed new terms, *optical parallax,* and *personal parallax.* Parallax is also a basic term in photogrammetry, being qualified as "linear parallax," "stereoscopic parallax," etc.

parallax: annual.—The difference between the direction from the earth to a star and the direction from the sun to the same star. Also termed heliocentric parallax.

The annual parallax of a given star is equal to the *angle* at the star subtended by the radius of the earth's orbit. See *parsec.* Annual parallax does not enter into any *surveying* problems.

parallax: diurnal.—The difference between the direction from a point on the surface of the earth to a planet, star, or other celestial object, and the direction from the center of the earth to the same object. Also termed geocentric parallax.

Diurnal parallax at a given point on the surface of the earth changes in magnitude as the earth rotates on its axis: it is greatest when the celestial object is in the *horizon* of the given point, when it is equal to the *angle* at the object subtended by the semidiameter of the earth, and is termed horizontal parallax. The *angle* at a celestial object subtended by the equatorial semidiameter of the earth is termed equatorial horizontal parallax, and is used to indicate the distance of the object from the earth.

Of less importance to the surveyor is *annual* or *heliocentric parallax,* the difference of directions to a star from the earth and from the sun, and of no importance whatever is secular parallax, due to the slow progressive motion of the solar system through space.

parallax: equatorial horizontal.—See under *parallax: diurnal parallax.*

parallax: geocentric.—See under *parallax: diurnal parallax.*

parallax: heliocentric.—See under *parallax: annual parallax.*

parallax: horizontal.—See under *parallax: diurnal parallax.*

parallax: instrumental.—A change in the apparent position of an object with respect to the reference mark(s) of an instrument which is due to imperfect adjustment of the instrument or to a change in the position of the observer.

When a *telescope* is poorly focused, so that the image of the object does not lie in the plane of the *reticle* (cross hairs), a movement of the eye transverse to the *line of collimation* will cause an apparent movement of the image of the object with respect to the cross hairs. This is a usual form of instrumental parallax, and for it the term optical parallax is proposed. *Parallax* may also result from the position in which an observer stands with respect to the *fiducial marks* on an instrument, as when reading a *vernier* or marking a tape end; for this type of *parallax* the term personal parallax is proposed.

parallax: optical.—See under *parallax: instrumental parallax.*

parallax: personal.—See under *parallax: instrumental parallax.*

parallax: secular.—See under *parallax: diurnal parallax.*

parallax: solar.—The *equatorial horizontal parallax* of the sun.

Solar parallax is the basis for the determination of the *astronomical unit,* and is used in some astronomical calculations as a measure of distance in place of that unit.

parallel: astronomic.—A line on the surface of the earth which has the same *astronomic latitude* at every point.

Because the *deflection of the vertical* is not the same at all points on the earth, an astronomic parallel is an irregular line, not lying in a single plane. The astronomic parallel of 0° *latitude* is called the *astronomic equator;* also called the *terrestrial equator.*

parallel: auxiliary standard (United States public-land surveys).*—A new *standard parallel* or *correction line* established, when required, for control purposes where original *standard parallels* or *correction lines* were placed at intervals of 30 or 36 miles.

Auxiliary standard parallels are used in the extension of old *surveys* and for the control of new *surveys.* Such a line may be given a local name, as "Cedar Creek Correction Line," or "Fifth Auxiliary Standard Parallel North."

parallel: geodetic.—A line on the *spheroid* which has the same *geodetic latitude* at every point.

A geodetic parallel, other than the *equator,* is not a *geodetic* (*geodesic*) *line.* In form, it is a small circle whose plane is parallel with the plane of the *geodetic equator.*

parallel: geographic.—A line on the earth having the same *latitude* at every point.

The term is applicable alike to an *astronomic parallel* or to a *geodetic parallel.*

parallel: standard (cartography).—A *parallel* of *latitude* which is used as a control line in the computation of a *map projection.*

In illustrations of *map projections* of the conic type, a standard parallel usually represents a line of tangency or of intersection of a cone with the surface of the *spheroid.* Such illustrations are not exact.

parallel: standard (United States public-land surveys).*—An auxiliary governing line established along the *astronomic parallel,* initiated at a selected *township corner* on a *principal meridian,* usually at intervals of 24 miles from the *base line,* on which *standard township, section,* and *quarter-section corners* are established; also known as a "correction line."

Standard parallels, or "correction lines" are established for the purpose of limiting the convergence of *range lines* from the south.

parallel of altitude.—See under *almucantar.*

parallel of declination.—See under *declination.*

parametric latitude.—See under *latitude.*

Parkhurst theodolite.—See under *theodolite.*

parsec.—The distance at which the *annual parallax* is equal to one second of *arc* (1″).

partial (leveling).—See *divergence* (*leveling*).

passometer.—A pocket-size instrument which registers the number of steps taken by the pedestrian carrying it.

The passometer is housed in a case resembling a watch case and is carried in an upright position attached to the body or to a leg. The distance walked is obtained by multiplying the number of steps taken by the length of a *pace,* or of a *stride,* if double steps are registered. See also *pedometer.*

pedometer.—A pocket-size instrument which registers the distance in linear units traversed by the pedestrian carrying it.

The term pedometer was formerly applied also to an instrument which registered the number of steps taken, but for which the term *passometer* is now used. The pedometer, like the *passometer,* is housed in a case resembling a watch case, and is carried in an upright position attached to the body or to a leg. It registers the linear distance traveled in miles or other units, the number of steps taken being mechanically transformed into linear units. This transformation depends upon the length of a *pace,* which may be different for different persons. The pedometer can be adjusted to the length of the *pace* of the person carrying it.

peg adjustment.—A method of adjusting a *leveling instrument* of the *dumpy level* type, to make the *line of collimation* parallel with the *axis of the spirit level,* and employing two stable marks (pegs) the length of one instrument sight apart.

The peg adjustment can be used with a *wye level,* but such an instrument is more readily adjusted by reversing the *telescope* in its wyes. A form of peg adjustment is used in determining the *collimation error* of *first-order leveling instruments* of the *dumpy level* type.

Pemberton leveling rod.—See under *leveling rod: Pemberton rod.*

pendulum.—In general, a body so suspended as to swing freely to and fro under the influence of *gravity* and momentum. See *pendulum: simple pendulum; compound pendulum.*

pendulum: bronze.—A quarter-meter *pendulum* used by the U. S. Coast and Geodetic Survey, and made of aluminum bronze consisting of one part of aluminum and nine parts of copper.

Beginning in 1920, bronze pendulums were replaced by *invar pendulums* in the *gravity* work of the U. S. Coast and Geodetic Survey.

pendulum: compound.—Any actual *pendulum.*

A compound pendulum may be considered as composed of an indefinitely large number of material particles, at different distances from the *center of suspension,* each constituting a *simple pendulum.* The period of *vibration* (*oscillation*) of the compound pendulum may be taken as a resultant of the periods of the *simple pendulums* of which it is composed.

pendulum: dummy.—A *pendulum* of similar construction to the *working pendulums* except that it is equipped with a thermometer and is fastened rigidly in the *receiver* so that it cannot swing during observations.

The dummy pendulum is subject to the same temperature conditions as the *working pendulums,* and is used in determining their temperature when in use.

pendulum: free-swinging.—A *pendulum* moving wholly under the influence of *gravity* and an initial momentum imparted to it by mechanical or other means.

In *gravity* work, the initial momentum may be imparted by drawing the *pendulum* slightly out of plumb and then releasing it.

pendulum: idle.—A *working pendulum* placed in the receiver in advance of its being used, so that it may assume the same temperature as the *dummy pendulum.*

pendulum: invar.—A quarter-meter *pendulum* used by the U. S. Coast and Geodetic Survey, and made of an alloy consisting of approximately one part of nickel and two parts of iron.

Beginning in 1920, invar pendulums have replaced *bronze pendulums* in the *gravity* work of the U. S. Coast and Geodetic Survey. The invar pendulum is subject to magnetization, and has a *coefficient of thermal expansion* about one-fifteenth that of the *bronze pendulum* which it replaced. In general, the two *pendulums* are quite similar, the weight of the invar pendulum being somewhat greater and its bob a little thicker at the middle.

pendulum: invariable.—A *pendulum* so designed and equipped with means of support that it can be used in only one position.

With an invariable pendulum, the *centers of suspension* and *of oscillation* are not interchangeable in use. *Pendulums* of this type have been constructed by Kater, Peirce, and Mendenhall. They may be used only for relative measurements of the *intensity of gravity.*

pendulum: Mendenhall.—An *invariable pendulum,* one-quarter meter in length, with a *vibration* period of one-half second, composed of a lenticular-shaped bob on a thin stem, swung in an airtight case from which the air has been largely exhausted.

This *pendulum* was designed in 1890 by Superintendent Mendenhall and assistants of the U. S. Coast and Geodetic Survey. It is the form of *pendulum* used since then by that Bureau in determinations of the *intensity of gravity,* but has undergone various improvements in mounting, methods of observation, etc.

pendulum: reversible.—A *pendulum* so designed and equipped with means of support that it may be used with either end up or down.

With a reversible pendulum, the *centers of suspension* and *of oscillation* are interchangeable in fact. This type of *pendulum* is exemplified by *pendulums* constructed by Kater, by Peirce, and by Repsold. Reversible pendulums may be used for absolute measurements of the *intensity of gravity* at base stations. See *pendulum: invariable.*

pendulum: simple.—A theoretical concept: a heavy particle suspended from a fixed point by a fine thread which is inextensible and without weight.

A simple pendulum cannot be realized in actual work. A simple pendulum is, however, the basis of reductions of observations made with an actual *pendulum.* Those observations have *corrections* applied to them to obtain results which would have been produced by an equivalent simple pendulum.

pendulum: temperature correction.—The quantity that is applied to the period of *vibration* of a *pendulum* to allow for the difference in the length of the *pendulum* at the temperature of observation and its length at some other temperature which has been adopted for purposes of *standardization* or for combining or comparing corresponding values.

In the work of the U. S. Coast and Geodetic Survey, the temperature adopted for these purposes is 15° centigrade.

pendulum: working.—A *pendulum* which is used (swung) in a determination of the *intensity of gravity.*

perigee.—The point in the orbit of the moon nearest to the earth.

perihelion.—The point in the orbit of a planet or comet closest to the sun.

Perihelion of the earth's orbit provides the index for the *anomalistic year.*

permanent bench mark.—See under *bench mark.*

perpendicular equation (traverse).—See under *equation.*

personal equation.—The *time* interval between the sensory perception of a phenomenon and the motor reaction thereto.

Personal equation may be either positive or negative, as an observer may anticipate the occurrence of an event, or wait until he actually sees it occur before making a record.

This is a *systematic error,* treated as of the constant type. It is a *personal error,* for which the term personal equation is reserved. It is of especial significance in observations of *time,* made to determine *clock correction.* For example, it is the elapsed *time* between the seeing of the event, star bisected by wire, and the recording of that event by mechanical (*chronograph*) means or by other method (eye-and-ear). Because of its significance in such work, special means have been developed for its control. See *micrometer: transit.*

personal error.—See under *error.*

personal parallax.—See under *parallax: instrumental.*

perspective map projection.—See under *map projection.*

perspective map projection upon a tangent cylinder.—See under *map projection.*

phase.—The visible aspect of an object.

In astronomy, phase is a stage in a cycle of recurring aspects, caused by a systematic variation of the illumination of an object. The moon passes through its phases, new moon to full moon and back to new moon, as its position relative to the sun and earth changes. In *surveying,* phase is applied to a signal which presents areas of varying brightness to the observer: a round pole, illuminated from the side; a square pole, of which the observer sees two sides, one more strongly illuminated than the other. The *error* in pointing due to phase is of the same character and requires the same treatment as an *error* due to observing an *eccentric object*. Phase may be closely associated with *asymmetry of object* (target), but the two terms are not identical.

Philadelphia leveling rod.—See under *leveling rod: Philadelphia rod.*

physiographic feature.*—A prominent or conspicuous *physiographic form* or salient part thereof.

physiographic form.*—A *land form* considered with regard to origin, cause, or history.

physiography.*—The classification of the genesis of *land forms.*

pivot inequality.—Any difference in the diameters or any irregularities in the form of the pivots of the *horizontal axis* of a surveying or astronomical instrument (*theodolite, transit,* etc.).

Though formerly taken into account in the computation of astronomical observations made in connection with *geodetic surveys,* corrections for pivot inequality have been rendered unnecessary by the production of instruments having pivots which are *accurate* and *precise* in size and form.

plane (pendulum).—A suspension bearing used in the support of a *pendulum.*

In the *Mendenhall pendulum,* an agate plane is mounted in the head of the *pendulum;* this bears on an agate *knife edge* which is mounted in the *receiver.* Since 1940, the positions of plane and *knife edge* in some *pendulums* have been reversed.

plane coordinates: State coordinate system.—See under *State coordinate systems.*

plane elliptic arc.—Any part of the line formed by the intersection of a plane and an *ellipsoid.*

plane rectangular coordinates.—See under *coordinates.*

plane survey.—See under *surveying: plane.*

plane surveying.—See under *surveying.*

plane table (surveying).—A device for plotting the lines of a *survey* directly from the observations.

A plane table consists essentially of a drawing board on a tripod, with a ruler, the ruler being pointed at the object observed by means of a telescope or other sighting device.

planimeter.—A mechanical integrator for measuring the area of a plane surface.

Using a planimeter, the *map* area within a given perimeter can be measured; by application of a factor derived from the *map scale,* the *map* area can be transformed into the corresponding area on the earth. There are several forms of planimeters, depending on the same mathematical principles, but differing in their mechanical design and operation. The one most generally used in map work is the polar planimeter; other well-known forms are the rolling planimeter and the suspended planimeter.

planimetric map.—See under *map.*

planimetry.—The mensuration of plane surfaces.

In mapping, planimetry is the determination of horizontal distances, *angles,* and areas, by measurements on a *map.*

planisphere.—A representation on a plane of the circles (*parallels, meridians,* etc.) of a *sphere.*

A particular form of a planisphere is a *polar map projection* of the *celestial sphere,* with movable parts which can be adjusted to show the visible aspect of the heavens at any given time, thereby aiding in the selection and identification of stars for various purposes. The planisphere is the basis of the *planispheric astrolabe.*

planispheric astrolabe.—See under *astrolabe.*

plat (general).*—A diagram drawn to *scale* showing all essential data pertaining to the boundaries and subdivisions of a tract of land, as determined by *survey* or protraction.

A plat should show all data required for a complete and accurate description of the land which it delineates, including the *bearings* (or *azimuths*) and lengths of the boundaries of each subdivision. A plat may constitute a legal description of the land and be used in lieu of a written description.

plat (United States public-land surveys).*—The term *plat,* as employed technically by the U. S. General Land Office, refers to the drawing which represents the particular area included in a survey, such as a *township,* private land claim, or mineral claim, and the lines surveyed, established, retraced, or resurveyed, showing the direction and length of each such line; the relation to the adjoining official surveys; the boundaries, descriptions, and area of each parcel of land subdivided; and, as nearly as may be practicable, a representation of the *relief,* and improvements within the limits of the *survey.*

A public-land survey does not obtain complete official or legal status until the field notes and the *plat* have been approved by the proper supervising officer, and accepted by the Commissioner of the General Land Office.

plot.—To place *survey* data upon a *map* or *plat.*

In past use, no clearly defined difference existed between *plat* and plot. It is suggested that a difference be established by limiting *plat* to the graphical representation of a *survey,* and plot to the cartographic operations involved in the construction of a *map* or *plat.*

plus distance.—The distance along a *survey* line from a *survey station* or the last even-numbered *survey* point to a supplementary point.

A stake is set at 515.56 feet from the initial point of a *survey* line and is not the terminal of that line: the even 500-foot point is station No. 5; the stake is a *plus station* (No. 5+15.56), and 15.56 feet is a plus distance.

plus sight.—See under *backsight (leveling).*

plus station.—An intermediate point on a *survey* line, not at an even tape length distance from the initial point. See *plus distance.*

point.—A position on a reference system determined by a *survey.* A *fix.*

point: turning (leveling).—A point on which a *foresight* is taken from one *instrument station* in a *line of levels* and on which a *backsight* is taken from the next instrument station in the *running* and which is established simply for the purpose of allowing the *leveling instrument* to be moved forward along the line of levels without a break in the series of measured differences of elevation.

A turning point may be a steel pin driven in the ground, the head of a hold-down spike on a railroad track, a nail driven in a tie of a railroad track, a ball bearing set in a small dent in a concrete pavement, etc.

point of compound curvature.—See under *curve.*

point of curvature.—See under *curve.*

point of intersection.—See under *curve.*

point of tangency.—See under *curve.*

pointers: the pointers.—The second-magnitude stars, Alpha and Beta, in the constellation Ursa Major (Big Dipper), the line through which points to *Polaris.*

The pointers facilitate the finding and identification of *Polaris.* They are in the outer side of the "bowl" of the Big Dipper, away from the handle.

polar coordinates.—See under *coordinates.*

polar distance.—The complement of the *declination* (*co-declination*) or 90° minus the *declination.*

Polar distance forms one side, celestial body to pole, of the *astronomical triangle.* It is opposite the *zenith.*

polar orthographic map projection.—See under *map projection: modified polyconic.*

Polaris.—The second-magnitude star, Alpha, in the constellation Ursa Minor (Little Dipper).

Also known as the polestar, or north star, because of its proximity to the north pole of the *celestial sphere,* Polaris is well situated for determinations of *astronomical azimuth,* and for the determination of the direction of the *celestial meridian.* It is at the extreme outer end of the handle of the "Little Dipper." See *pointers: the pointers.*

polestar.—See under *Polaris.*

polyconic map projection.—See under *map projection.*

polyconic map projection, modified.—See under *map projection.*

polyconic map projection, transverse.—See under *map projection: transverse polyconic.*

Port Clarence datum.—See under *datum.*

position.—1. Data which define the location of a point with respect to a reference system. The place occupied by a point on the surface of the earth. The *coordinates* which define the location of a point on the *geoid* or *spheroid.*

2. A prescribed setting (reading) of the horizontal circle of a *direction theodolite* to be used for the observation on the initial station of a series of stations which are to be observed on.

The term as defined under 2 (above) indicates a definite position of the horizontal circle. For example, in *first-order triangulation,* using a two-micrometer *direction theodolite,* in Position No. 1, the circle is set to read 0° 00′ 40″ when the initial station is observed; in Position No. 2, the circle is set to read 11°01′50″ when the initial station is observed; etc.

position: adjusted.—An *adjusted value* of the coordinate position of a point on the earth.

In the adjustment of a horizontal *control survey, discrepancies* arising from *errors* in the observational data are removed, and position data of the survey stations are correlated and coordinated on an adopted reference system (*geodetic datum* or *plane-coordinate system*). The positions which are obtained by the adjustment are called adjusted positions, and when used as control for other work are referred to as fixed positions. In the adjustment of a vertical control survey, the values obtained are adjusted elevations, called fixed elevations when used to control other work.

position: field.—A *position* computed while field work is in progress to determine the acceptability of the observations or to provide a *preliminary position* for other purposes.

position: fixed.—See under *position: adjusted position.*

position: geodetic.—A position of a point on the surface of the earth expressed in terms of *geodetic latitude* and *geodetic longitude.*

A geodetic position implies an adopted *geodetic datum.* In a complete record of a geodetic position, the *datum* must be stated.

position: geographic.—The position of a point on the surface of the earth expressed in terms of *latitude* and *longitude,* either *geodetic* or *astronomic.*

The U. S. Coast and Geodetic Survey uses the term geographic positions for positions on a *geodetic datum.*

position: preliminary.—See under *preliminary.*

position: relative.—The location of a point with respect to other points.

The relative position of two points whose positions are given on the same coordinate system is the difference of their *coordinates,* which may also be expressed by a *direction* and distance.

position line.—See under *line of position.*

postal map.—See under *map.*

post route map.—See under *map.*

Pratt theory of isostasy.—See under *isostasy.*

precession.—See under *precession of the equinoxes.*

precession of the equinoxes.—A slow western movement of the *equinoxes.*

Precession of the equinoxes amounts to about 50″ per year and causes the difference in length between the *sidereal* and *tropical years.*

precise level.—See under *leveling instrument.*

precise leveling.—See under *leveling: first-order.*

precise triangulation.—See under *triangulation.*

precision.*—Degree of refinement in the performance of an operation or in the statement of a result.

Precision relates to the quality of execution, and is distinguished from *accuracy* which relates to the quality of the result. The term precision not only applies to the fidelity with which required operations are performed, but by custom has been applied to methods and instruments employed in obtaining results of a high order of *accuracy.* Precision is exemplified by the number of decimal places to which a computation is carried and a result stated. In a general way, the *accuracy* of a result should determine the precision of its expression. Precision is of no significance unless *accuracy* is also obtained.

preliminary.—Not of the desired *accuracy* and *precision,* and adopted for temporary use with the proviso of later being superseded.

In the *adjustment* of *triangulation,* the term preliminary is applied to triangles and *geographic positions* derived from selected observations for use in forming *latitude* and *longitude condition equations.*

preliminary elevation.—See under *elevation.*

preliminary position.—See under *preliminary.*

preliminary survey.—See under *survey.*

preliminary triangle.—See under *preliminary.*

Pribilof Islands datums.—See under *datum: St. George Island datum; St. Paul Island datum.*

primary triangulation.—See under *triangulation.*

prime meridian.—See under *meridian.*

prime vertical.—A *vertical circle* perpendicular to the plane of the *celestial meridian.*

The plane of the prime vertical cuts the *horizon* in the east and west points.

principal axis (of a lens).—The line *normal* to both surfaces of a lens.

In a well-centered lens system, for which the principal axes of the components are in the same straight line, that line is termed the optical axis of the system, or simply the axis. It should never be called the optic axis, which is a term used in crystallography.

principal meridian (United States public-land surveys).—See under *meridian.*

principal station.—See under *station.*

prismatic astrolabe.—See under *astrolabe.*

prismatic compass.—See under *compass.*

probable error.—See under *error.*

probable value: most probable value.—See under *value: most probable value.*

profile.—A section cut from the surface of the earth by a line moving along a fixed line as a directrix, and showing the line of intersection with that surface or with some underlying stratum.

Usually the moving line will be a vertical straight line, and the directrix will be a straight line or a regular curve (as the center line of a railroad or highway). The surface of intersection will be a plane surface or a regular curved surface that can be developed on a plane.

progress sketch.—A *map* or sketch showing work accomplished.

In *triangulation* and *traverse surveys,* each point established is shown on the progress sketch, and also lines observed over and *base lines* measured. In a *leveling survey,* the progress sketch shows the route followed and the towns passed through, but not necessarily the locations of the *bench marks.*

projection (surveying).—The extension of a line beyond the points which determine its character and position. The transfer of a series of *survey* lines to a single theoretical line by a series of lines perpendicular to the theoretical line.

In surveying a *traverse,* a series of measured short lines may be projected onto a single long line, connecting two main *survey* stations, and the long line then treated as a measured line of the *traverse.*

projection: map.—See under *map projection.*

protractor.—An instrument consisting of a plate marked with units of circular measure.

A protractor is used in laying out on a flat or curved surface an *angle* of desired magnitude; or in determining the magnitude of the *angle* made by two intersecting lines on a flat surface. Protractors are made in many sizes and forms to suit the purpose and convenience of the ones using them. See *protractor: three-arm protractor; coordinate protractor.*

protractor: coordinate.—A square-shaped *protractor* having graduations on two adjacent edges, with the center at one corner. It is equipped with a movable arm turning about the center, and graduated to show linear quantities on a given *scale.* The *protractor* is covered with a *grid* of the same *scale* and units as the arm.

In use, the arm is set for a given *azimuth* or *bearing,* and the length of a line marked on the arm: a reading of the *grid* at this point will give the *latitude* and *departure* of a line of the given length and *bearing.*

protractor: three-arm.—A full-circle *protractor,* equipped with three arms, the fiducial edges (extended) of which pass through the center of the circle. The middle arm is fixed, and reads 0° on the graduated circle. The other arms are movable, and their positions on the circle are read with the aid of *verniers.*

The two movable arms may be set to two *angles* observed between three fixed signals of known position, and the position of the point of observation obtained by a graphical solution of the *three-point problem.* Also called a station pointer.

public domain (United States).*—The territory ceded to the Federal Government by the Original Thirteen States, together with certain subsequent additions by cession, treaty, and purchase.

At its greatest extent, the public domain aggregated over 1,820,000,000 acres and included the present States of Alabama, Arizona, Arkansas, California, Colorado, Florida, Idaho, Illinois, Indiana, Iowa, Kansas, Louisiana, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, North Dakota, Ohio, Oklahoma, Oregon, South Dakota, Utah, Washington, Wisconsin, Wyoming, and the Territory of Alaska.

public land (United States).*—The portion of the *public domain (United States)* to which title is still vested in the Federal Government.

public land States.*—Those States and Territories created out of the *public domain (United States).*

For list of public-land States see *public domain (United States).*

Puerto Rico datum.—See under *datum.*

Q

quadrant.—Mathematics. A *sector* having an *arc* of 90°. Surveying. A *surveying* or astronomical instrument composed of a graduated *arc* about 90° in length, equipped with a sighting device (*alidade*).

The quadrant may be considered a form of *sector (surveying).* Some quadrants combine both *surveying* and astronomical functions, having two *arcs,* one horizontal and one vertical.

quarter section (United States public-land surveys).—See under *section.*

quarter-quarter section (United States public-land surveys).—See under *section.*

R

radar.—The art of detecting by means of radio echoes the presence of objects, determining their direction and range, and employing the data obtained for navigational or military purposes.

Radar information is commonly obtained from the plan position indicator (PPI), a form of oscilloscope on which the direction and range of an echoing object are represented graphically.

radian.—See *angle (general), (d).*

random line.—A trial line run as nearly as circumstances permit from one survey station toward another survey station which cannot be seen from the first station.

The *error of closure* (amount by which the second station is missed) permits the computation of a *correction* to the initial *azimuth* of the random line; it also permits the computation of *offsets* from the random line to establish points on the line between the survey stations.

random traverse.—See under *traverse.*

range.—In general, two points in line with the point of observation.

The following are practical exemplifications of a range:

1. The line defined by the side of a building or by a fence may be extended visually to its intersection with a survey line: the point of intersection

thus determined is said to be in range with the side of the building or with the fence.

2. In hydrographic surveying, a range formed by two shore objects, if suitably located, aids in keeping a boat moving in a straight line; the line defined by the range.

3. In navigation, specially constructed aids mark ranges defining channels which are to be followed by vessels and keep them clear of dangers. Such ranges are often permanently marked by suitably lighted structures which are given identifying names, as: Honolulu Channel Front Light; Honolulu Channel Rear Light.

4. Boundary lines across water areas and boundary corners in water areas where permanent marks cannot be established, are sometimes defined by intersection of range lines or by a range line and distance from a mark, the range lines being marked by permanent *monuments* on the land.

range (United States public-land surveys).*—Any series of contiguous *townships* situated north and south of each other; also *sections* similarly situated within a *township.*

Ranges of townships are numbered consecutively east and west from a *principal meridian:* thus "range 3 east" indicates the third *range* or row of *townships* to the east from a *principal meridian.* The word *range* is used in conjunction with the appropriate *township* to indicate the *coordinates* of a particular *township* with reference to the *initial point;* thus "township 14 north, range 3 east" indicates the particular *township* which is the 14th *township* north of the *base line* and the 3rd *township* east of the *principal meridian.*

range finder.—An instrument for finding the distance from a single point of observation to other points at which no instruments are placed. See *tachymeter.*

In general, a range finder employs a very short *base line,* of fixed length, which is part of the instrument, and is utilized according to the principle of *triangulation.* The *precision* of the optical and other parts of the instrument is very great, but because of the small magnitude of the *angle* of intersection of the lines of sight at the object observed on, the distances obtained are not of a high order of *precision.*

range line (United States public-land surveys).*—An exterior boundary of a *township* extending in a north-south direction.

Réaumur scale.—A temperature scale in which 0° marks the freezing point and 80° the boiling point of water at 760 mm. barometric pressure.

In some of the earliest *geodetic* work, the Réaumur scale was used in stating the temperature of base-line measuring apparatus (*base apparatus*).

receiver (pendulum).—A heavy cast-metal box within which the *pendulum* is suspended and some auxiliary equipment placed when making observations for the *intensity of gravity.*

The receiver used with the *Mendenhall pendulum* is a heavy brass casting which can be made airtight and exhausted of air, mounted on foot screws for leveling, and provided with appliances for manipulating the *pendulums* and auxiliary equipment which are placed in it. It is provided with glass windows through which observations are made.

reciprocal of the flattening (of the earth).—See under *flattening of the earth.*

reconnaissance map.—See under *map.*

reconnaissance sketch.*—A drawing which resembles a *reconnaissance map* but is lacking in some *map* element.

reconnaissance survey.—See under *survey.*

recovery of station.—A survey *station* is considered as recovered when its mark (*monument*) is identified as authentic and proved to be occupying its original site.

The recovery of station is tested by checking the measurements for distance and *azimuth* (or *bearing*) from the station to a *reference mark. Witness marks* are aids to the recovery, but afford only secondary evidence of the position of the recovered mark. Sometimes, exact recovery requires that the original survey connections between the *station* and at least two adjacent *stations* of the same survey and class be tested and found unchanged.

rectangular coordinates.—See under *coordinates.*

rectangular polyconic map projection.—See under *map projection.*

rectifying latitude.—See under *latitude.*

reduced (parametric) latitude.—See under *latitude: parametric.*

reduction to center.—See under *center.*

reference mark.—A supplementary mark of permanent character close to a survey *station,* to which it is related by accurately measured distance and *azimuth* (or *bearing*).

The connection between a survey *station* and its reference mark(s) should be of sufficient *precision* and *accuracy* to permit the reestablishment of the *station* on the ground should its marks be destroyed, or the use of the reference mark in place of the survey *station* in the extension of surveys. Reference marks are used to define positions of boundary corners which may be so situated (as in water) that permanent marks cannot be placed exactly at the corners. See *corner: witness corner* (*United States public-land surveys*).

referencing.—The process of measuring the horizontal distances and directions (*azimuths* or *bearings*) from a survey *station* to nearby *landmarks, reference marks,* and other objects which may be used in the *recovery of the station.*

refraction: angle of.—See under *refraction of light.*

refraction: angle of incidence.—See under *refraction of light.*

refraction: astronomical.—The refraction by the earth's atmosphere of light from a source outside the atmosphere.

Light from a celestial body, as a star or planet, passes entirely through the atmosphere in reaching the earth, and follows a curved path which is concave to the surface of the earth. The *angle* between the direction of a ray where it enters the atmosphere and its direction at the point of observation is the astronomical refraction of the ray, or simply its refraction. It is also called celestial refraction. Its magnitude is greatest when the observed body is near the *horizon;* with slight irregularities, it decreases to a minimum near the *zenith,* being zero when the path of the ray is *normal* to the surfaces of the atmospheric layers. Astronomical refraction must be considered in many problems of astronomy, *surveying,* and navigation. See *refraction of light.*

refraction: atmospheric.—The *refraction of light* passing through the earth's atmosphere.

Atmospheric refraction includes both *astronomical refraction* and *terrestrial refraction.*

refraction: coefficient of.—The ratio of the *refraction angle* at the point of observation to the angle at the center of the earth subtended by the *arc* connecting the point of observation and the observed point.

refraction: index of.—The ratio of the sine of the *angle of incidence* to the sine of the *angle of refraction.*

When a ray of light passes from a vacuum into some medium or from some medium into a vacuum, the index of refraction is the absolute index of the medium. When a ray of light passes from one medium into another medium, the index is the relative index of those media. The index of refraction is slightly different for different wave lengths. This deviation, by refraction, of different wave lengths, is known as dispersion and, for any given medium, the amount of the deviation is a measure of the dispersive power of the medium. *Chromatic aberration* is overcome over a moderate range of wave lengths by using a composite lens (*achromatic lens*) composed of elements having different dispersive powers. A lens made achromatic for visual observation is not strictly achromatic for photographic work.

refraction: terrestrial.—The refraction by the earth's atmosphere of light from a terrestrial source.

The path of light from a terrestrial source is usually not far from horizontal; it passes through only the lower strata of the atmosphere, and suffers refraction throughout its entire length. The *angle of refraction,* or simply the refraction, of a survey line is the *angle* at the point of observation between the true *direction* of the observed object and the *direction* as shown by light coming from it (*direction* of object as seen).

Terrestrial refraction must be considered in many problems of *surveying,* and enters into some problems of astronomy and navigation which involve observations on terrestrial objects. Because the arrangement of the air strata is not exactly symmetrical in form and density, the path of a ray of light through the atmosphere is not a smooth curve in a *vertical plane.* Its curvature is greatest, however, in or close to a *vertical plane,* where its magnitude is of considerable importance in observations which are referred to the *zenith* or to the plane of the *horizon,* as are some *vertical angles* and *spirit levels.* Its effect is usually to make the apparent altitude of an object greater than its true *altitude,* though under special atmospheric conditions, this effect may be reversed.

In *triangulation,* a station which is normally just below the *apparent horizon* may, under certain conditions of the atmosphere and because of refraction, become visible for a short period of time and be observable. A line which can be observed over only with the aid of refraction is termed a *refraction line.* Refraction in a *direction* which is approximately horizontal is termed lateral refraction, and under some conditions can be of such magnitude as to seriously affect observed values of *horizontal* directions. A careful *reconnaissance* is required, especially for city surveys, to prevent the inclusion of lines so affected in a survey net. In some publications of the U. S. Coast and Geodetic Survey, terrestrial refraction is considered under the title of *atmospheric refraction.*

refraction angle.—That portion of an observed *zenith distance* which is due to the effect of *atmospheric refraction.*

The refraction angle may be taken as the difference between the observed and geometrically determined values of the *zenith distance.*

refraction line.—A line of sight to a *survey signal* which becomes visible only by the effect of *atmospheric refraction.*

Refraction lines are often over bodies of water, the observed signal normally being below the *apparent horizon,* but becoming visible at the time of greatest refraction during the day, when it appears to rise out of the water, disappearing back into the water as the refraction decreases.

refraction of light.—The change in direction of a ray of light passing obliquely from one medium into another, its speed in the two media being different.

The ray of light, before being refracted, is termed the incident ray; after refraction, the refracted ray. The *angle* between the incident ray and the *normal* to the surface separating the media at the point of refraction and on the same side of that surface as the incident ray is the angle of incidence; the corresponding *angle* between refracted ray and *normal* is the angle of refraction.

Refraction of light by the earth's atmosphere is termed *atmospheric refraction,* and is divided into two classes; *astronomical refraction,* also termed celestial refraction, and *terrestrial refraction.* Refraction of light must be considered in many problems of astronomy, *surveying,* and navigation. See *refraction: astronomical refraction; terrestrial refraction.*

relative position.—See under *position.*

relief.—See under *topography* (*or relief*).

repeating theodolite.—See under *theodolite.*

repetition of angles.—The accumulation of a series of measures of the same *angle* on the horizontal circle of a *repeating theodolite* or surveyor's *transit,* in making which the final reading of one measure is used as the initial setting of the next measure of the *angle.* The observed value of the *angle* is obtained by dividing the total *arc* passed over by the number of observations.

representative fraction.—See under *map scale: fractional.*

Repsold base-line measuring apparatus.—See under *base apparatus.*

resection.—The determination of the horizontal position of a *survey station* by observed *directions* from the station to points of known position.

The most usual problem in resection is the determination of a point of observation by measuring two *angles* between three fixed points; this is the *three-point problem,* which is solved by computation (in *triangulation*), by mechanical means (with a *three-arm protractor* in hydrographic work), or graphically (on a plane-table survey). Its solution gives what is often termed a *fix,* a position for the point of observation. As employed in photogrammetry, resection refers to position determination from measurements on photographs.

residual error.—See under *error.*

resurvey.—See under *survey: resurvey.*

reticle (reticule).—A system of wires, hairs, threads, etched lines, or the like, placed *normal* to the axis of a telescope at its principal *focus,* by means of which the telescope is sighted on a star, signal, or target, or by means of which appropriate readings are made on some scale, such as a leveling or stadia rod.

reticle ring.—The ring across which the system of wires, hairs, threads, or the like of a *reticle* are stretched; or the ring which supports the glass *diaphragm* in case the reticle is a system of lines etched on glass.

reversible pendulum.—See under *pendulum.*

reversing point, micrometer screw (geodetic level).—That setting of the micrometer head at which the bubble remains in the center of the level vial, when the instrument is level and is *rotated* about its axis.

On some geodetic levels, a horizontal line on the micrometer head index piece indicates the position of the reversing point, within one turn of the micrometer screw, when the top of the micrometer-screw head is made level with it.

revolution.—Astronomy. A turning of a body about an exterior point or axis. See *rotation.* Mathematics. A turning of a geometrical figure about an axis.

We say that the earth *rotates* on its axis and revolves around the sun. However, in usage, such clear distinction between rotate and revolve is not always made with complete consistency. For example, in geodetic surveying, a *spheroid* adopted to represent the earth is termed an *ellipsoid* of revolution, though according to exact terminology it is an *ellipsoid* developed by the *rotation* of an ellipse around its minor axis. It may be desirable to disregard custom and speak of such a figure as an *ellipsoid* of rotation. The expression, *spheroid* of revolution, is of such general use that its use is continued in this publication.

rhumb line.*—A line which crosses successive *meridians* at a constant angle.

The *Mercator* is the only *map projection* on which a rhumb line is represented by a *straight line.* Other names for rhumb line are: loxodrome, loxodromic curve, equiangular spiral, and *Mercator track.*

right ascension.—The *angle* between the plane of the *hour circle* passing through a celestial body and the plane of the *hour circle* passing through the *vernal equinox.*

Right ascension may be measured by the *angle* at the celestial pole between the tangents to the *hour circles* of the body and of the *vernal equinox,* or by the *arc* of the *equator* intercepted by those *hour circles.* Right ascension is measured eastward from the *vernal equinox* through 24 hours (360°). It corresponds to *longitude* on the earth, and with *declination* forms a pair of *coordinates* which defines the position of a body on the *celestial sphere.*

right bank.—See under *bank of stream.*

river bed.—See under *bed of stream.*

river crossing (leveling).—Carrying a *line of levels* across a stream or other body of water, when no suitable bridge is available and the width of the body of water is greater than the maximum allowable length of sight for the leveling, requires a special series of observations which taken collectively is known as a river crossing.

rod correction (leveling).—See under *correction.*

rotation.—Astronomy. A turning of a body about a self-contained axis. See *revolution.* Surveying. A turning of an instrument or part of an instrument about an axis.

The earth rotates on its axis. The *alidade* of a *theodolite* rotates about a vertical axis.

run of micrometer.—See under *micrometer.*

running (leveling).—A continuous series of measured differences of elevation, made *set-up* by set-up in one direction along a *section* of a *line of levels,* which results in a measurement of the difference of elevation between the *bench marks* or other points, either temporary or permanent, at the ends of the section.

rural delivery map.—See under *map.*

S

sag correction, tape.—See under *tape.*

St. George Island datum.—See under *datum.*

St. Hilaire Method (Method of Marcq St. Hilaire).—An adaptation of the *Sumner line* method of position determination which employs assumed *altitudes* (*vertical angles*) in the computation of an astronomical position, and applies a simple *correction* for the difference between the observed and assumed values of the *altitudes.*

This method simplifies the usual method of computing a position from *Sumner lines;* it is especially effective and simple near the geographic poles. It was invented in 1874 by Marcq de St. Hilaire, a French naval officer.

St. Michael datum.—See under *datum.*

St. Paul Island datum.—See under *datum.*

Sanson-Flamsteed map projection.—See under *map projection: sinusoidal.*

scale: equivalent map.—See under *map scale.*

scale: fractional map.—See under *map scale.*

scale: graphic (or bar) map.—See under *map scale.*

scale: map.—See under *map scale.*

scale factor.—A multiplier for reducing a distance obtained from a *map* by computation or scaling to the actual distance on the *datum* of the *map.* Also, in the *State coordinate systems,* scale factors are applied to *geodetic* lengths to obtain *grid lengths,* or to *grid lengths* to obtain *geodetic* lengths; both are lengths on a *sea-level datum,* but the *grid lengths* are affected by the distortion of the *map projection.*

school and bus route map.—See under *map.*

Schott base-line measuring apparatus.—See under *base apparatus.*

sea level.—In general, the surface of the sea used as a reference for *elevation.* In *surveying* and mapping, a curtailed form of *mean sea level.*

In *surveying* and mapping, the use of sea level should be avoided; but if used, it should be with the meaning of *mean sea level.*

sea level datum.—A determination of *mean sea level* that has been adopted as a standard *datum* for heights. The sea level is subject to some variations from year to year, but as the permanency of any datum is of prime importance in engineering work, a sea level datum after adoption should in general be maintained indefinitely even though differing slightly from later determinations of mean sea level based upon longer series of observations.

The sea level datum now used for the U. S. Coast and Geodetic Survey level net is officially known as the "Sea Level Datum of 1929," the year referring to the last general adjustment of the net. The datum itself may be considered as an adjustment based upon tide observations taken at various tide stations along the coasts of the United States over a number of years. See also under *mean sea level.*

sea mile.—See under *mile: nautical.*

secant method (United States public-land surveys).*—A method of determination of the *parallel* of *latitude* for the survey of a *base line* or *standard parallel* by *offsets* from a *great circle line* which cuts the *parallel* at the first and fifth mile corners of the *township* boundary.

The secant method is a modification of the *tangent method,* so arranged that minimum *offsets* need be made from the projected *great circle line* to the *parallel;* at the first and fifth mile stations on the secant the *offsets* are zero; between those stations the *offsets* are measured to the south; before and after those stations the *offsets* are measured to the north.

secondary triangulation.—See under *triangulation.*

second-order bench mark.—See under *bench mark.*

second-order control surveys.—See under *second-order work (control surveys).*

second-order leveling.—See under *leveling.*

second-order traverse.—See under *traverse.*

second-order triangulation.—See under *triangulation.*

second-order work (control surveys).—The designation given survey work of next-to-the-highest order of *accuracy* and *precision.* See *traverse; triangulation;* and *leveling.*

section (leveling).—That portion of a *line of levels* which is recorded and abstracted as a unit.

A section always begins and ends on a *bench mark,* either temporary or permanent, when in the main *line of levels* or on a *spur line of levels.* In the case of a *spur* to a point the *elevation* of which is determined by means of an *extra foresight,* the section must begin on a temporary or permanent *bench mark* and end on the point on which the *leveling rod* was held when the *extra foresight* was taken.

section (United States public-land surveys).*—The unit of subdivision of a *township;* normally a quadrangle 1 mile square, with boundaries conforming to *meridians* and *parallels* within established limits, and containing 640 acres as near as may be.

A normal *township* is divided into 36 sections by lines 1 mile apart measured from the southern and eastern boundaries; any excess or deficiency in measurement is placed in the northern *tier* or western *range* of sections. The sections within a normal *township* are numbered consecutively commencing with number 1 in the northeast section and progressing west and east alternately with progressive numbers in the *tiers,* to number 36 in the southeast section.

section: fractional (United States public-land surveys).*—A *section* containing appreciably less than 640 acres, usually due to invasion by a segregated body of water, or by other land which cannot properly be surveyed or disposed of as part of that *section.*

Sections are also frequently rendered fractional in closing the *surveys* on the north and west boundaries of the *township,* for the reason that deficiencies in measurement caused by *error* of *survey* or *convergence of meridians* are placed in the half mile closing against these *township* boundaries.

section: half (United States public-land surveys).*—One-half of a normal *section,* formed by dividing a *section* into two parts by a line connecting opposite *quarter-section corners,* and containing 320 acres as near as may be.

The half section is a unit of description of the *public lands;* thus the "east half, section 10" is the legal description of that portion of *section* 10 of a given *township* lying east of the north-south *center line* of the *section.*

section: quarter (United States public-land surveys).*—One-fourth of a normal *section,* formed by dividing a *section* into four parts by lines connecting the opposite *quarter-section corners,* and containing 160 acres as near as may be.

The quarter section is a unit of description of the *public lands;* thus the "northeast quarter, section 10" is the legal description of that portion of section 10 of a given *township* lying east of the north-south *center line* and north of the east-west *center line* of that *section.*

section: quarter-quarter (United States public-land surveys).*—One-sixteenth of a normal *section,* formed by dividing a *quarter section* into four parts by lines connecting the midpoints of opposite sides, and containing 40 acres as near as may be.

The quarter-quarter section is a unit of description of the *public lands;* thus the "northeast quarter of the northeast quarter, section 10" is the legal description for that portion of section 10 of a given *township* lying east of the north-south *center line* and north of the east-west *center line* of the northeast quarter of that *section.*

sector.—Mathematics. A geometrical figure bounded by an *arc* and two radii. Surveying. An instrument composed of a graduated *arc* equipped with a sighting device (*alidade*).

In *surveying,* the sector and the *quadrant* were the immediate predecessors of the *theodolite,* both being mentioned in reports of early *geodetic* surveys in connection with the measurement of horizontal angles. These instruments were constructed with *arcs* having radii several feet in length. Combined with other words, sector is used to designate certain *map* features, as red sector; and special types of instruments, as *zenith sector; track sector.*

sector: track.—See under *track sector.*

sector: zenith.—An astronomical instrument for measuring *zenith distances,* in which a plumb line is read against a short vertical *arc.* Also, a modern form of such an instrument, employing *spirit levels* and an *ocular micrometer.*

In its early form, before the invention of the *telescope,* the zenith sector consisted of a *quadrant* with plumb-line control. With the coming of the *telescope,* zenith sectors used in astronomical work in connection with surveying operations employed *telescopes* up to five or six feet in length with short vertical *arcs,* the *vertical angle* being read by means of a plumb line suspended from the horizontal axis of the instrument.

A later form of zenith sector, employing *spirit levels* and an *ocular micrometer,* is a *precision instrument.* These instruments were used in measuring *zenith distances* and differences of *zenith distances* of stars close to the *zenith.* In making determinations of *astronomic latitude,* the zenith sector has been superseded by the *zenith telescope.*

secular parallax.—See under *parallax: diurnal parallax.*

selenotrope.—An instrument similar to the *heliotrope,* but adapted for reflecting moonlight for signal purposes.

The selenotrope differs from the heliotrope only in the greater size of the mirror used, being operated in exactly the same way. In tests made in 1883 and 1887, selenotropes with mirrors of the sizes stated furnished satisfactory lights on which to observe at the following distances: 22 miles, 6- by 6-inch mirror; 48 miles, 6- by 8-inch mirror; and 70 miles, 8- by 10-inch mirror.

set-back.—The horizontal distance from the *fiducial mark* on the front end of a tape or part of tape which is in use at the time back to the point on the ground mark or monument to which the particular measure is being made.

A set-back is usually a very small distance in base-line measures where tape supports in the form of stakes are placed in advance of the measure; if the distance between the stakes is too small, the tape will over-run the stakes, and set-backs must be measured. If portable supports, such as bucks, are used, there will seldom be need for measuring set-backs. Set-backs are minus corrections to tape distances. In the work of the U. S. Coast and Geodetic Survey, the use of set-backs is avoided. See *set-up.*

set-up.—The horizontal distance from the *fiducial mark* on the front end of a tape or part of tape which is in use at the time, measured in a forward direction to the point on the ground mark or *monument* to which the particular measure is being made. Also, a surveying instrument in position at a point of observation.

A set-up is usually a very small distance in base-line measures where tape supports in the form of stakes are placed in advance of the measures; if the distance between the stakes is too great, the tape will not reach from stake to stake, and set-ups must be measured. If portable supports, such as bucks, are used, there will seldom be need for measuring set-ups. Set-ups are plus *corrections* to tape distances. See *set-back.*

set-up (leveling).—The actual physical placing of the *leveling instrument* over an *instrument station.*

The terms *instrument station* and set-up, especially in leveling, are often used interchangeably with little loss of accuracy of statement. However, if any distinction is to be made, set-up is considered to be the instrument when mounted or set up over the *instrument station* or point on the ground which is in the vertical axis of rotation of the instrument.

sextant.—A hand instrument for measuring the *angle* between two objects in the plane defined by the two objects and the point of observation, and based on the optical principle that if a ray of light suffers two successive reflections in the same plane by plane mirrors, the *angle* between the first and last directions of the ray will be twice the *angle* between the plane mirrors.

The sextant has a graduated arc of 60°, from which it derives its name. Each degree division represents a 2° *angle* between the first and last directions of the ray of light, and is so marked: the 60° arc is marked 120°. The last reflection of the ray is made parallel with the direction of the second object, hence the *angle* between the two objects, one seen by double reflection and the other viewed directly, is registered on the sextant. The instrument is used in navigation for measuring *elevations* of celestial objects, in hydrographic surveying for measuring *horizontal* angles at a point in a moving boat between shore objects, and wherever instability of support makes it impossible to use a *theodolite* or *transit.* The sextant has been in use since around 1730. It is capable of use as a *precision* instrument.

sextant: bubble sextant.—A *sextant* in which the bubble of a *spirit level* serves as the *horizon.*

A simple form of bubble sextant was designed by R. E. Byrd and used by him in aerial navigation, on expeditions to the North and South Poles and on other arctic and antarctic expeditions. It has a *spirit level* with a circular bubble attached to the *sextant* quite close to the horizon glass; the *elevation* of a celestial object is measured by bringing its image into contact with the image of the bubble as seen in the horizon glass.

ship-shore triangulation.—See *triangulation: ship-shore.*

shoran.—A pulse-type electronic ranging system originally designed for the positioning of bombing aircraft and later adapted for use in aerial photography, geophysical exploration, hydrography, and long-range geodetic surveying.

Fundamentally, the system consists of a mobile transmitter-receiver-indicator unit and a fixed receiver-transmitter unit (transponder). Pulses are sent from the mobile transmitter and returned to the originating point by the transponder. The indicator measures the time interval required for travel of a pulse between stations and converts this information into distance to the nearest thousandth of a mile. Shoran range is approximately limited to direct-path transmission as a consequence of the carrier frequency used (200–300 megacycles).

shore.—The land which is covered and uncovered by the rise and fall of the normal *tide.*

It is the strip of land between the *mean high-water* and *mean low-water lines.* In its strictest use, the term applies only to land along tidal waters.

shore line.—The land-side *boundary line* of the *shore;* the *mean high-water line.*

Along a river, the shore line is the line along the river bank which is washed by the water without overflowing it.

shore of stream.—The area between the *bank* of a stream and its *low-water line.*

side equation.—See under *equation.*

side shot.—A reading or measurement from a survey *station* to locate a point which is not intended to be used as a base for the extension of the *survey.*

A side shot is usually made for the purpose of determining the *position* of some object which is to be shown on the *map.*

sidereal day.—See under *day.*

sidereal focus.—See under *focus.*

sidereal month.—See under *month.*

sidereal time.—See under *time.*

sidereal time: Greenwich.—See under *time.*

sidereal time: local.—See under *time.*

sidereal year.—See under *year.*

signal: longitude.—A sign indicating a time event, observable at different stations, and used in comparing local times of those stations, and determining the difference of their *longitudes.* See *longitude* (various methods).

signal: survey.—A natural or artificial object or structure whose horizontal and sometimes vertical position is obtained by *surveying* methods.

Signals are given special designations according to the kind of *survey* in which they are determined, or which they may later serve.

signal: triangulation.—A rigid structure erected over or close to a *triangulation* station and used for supporting instrument and observer, or target, or both instrument and observer and target, in a *triangulation* survey. Also any object, natural or artificial, whose position is obtained in a *triangulation* survey.

The term may be applied to a structure whose position is determined by *triangulation,* but whose primary purpose is to serve later in a *hydrographic* or *topographic survey,* when it may become known as a hydrographic or topographic signal.

simple pendulum.—See under *pendulum.*

simultaneous double line.—See under *level line: simultaneous level line.*

simultaneous level line.—See under *level line.*

sinusoidal map projection.—See under *map projection.*

siphon barometer.—See under *barometer.*

slope stake.—A stake set on the line where a finished side slope of an excavation (cut) or embankment meets the original ground surface.

small scale.—See under *map scale: fractional.*

solar attachment.—An auxiliary instrument which may be attached to an engineer's *transit,* permitting its use as a *solar compass.*

solar compass.—See under *compass.*

solar day.—See under *day.*

solar focus.—See under *focus: sidereal focus.*

solar parallax.—See under *parallax.*

solar time.—See under *time: apparent solar time; mean solar time.*

solar transit.—An engineer's *transit* designed for use in place of a *solar compass.*

solenoid.—A coil of wire, often helical in form, which becomes an electromagnet when an electric current is passed through it.

When tests show an *invar pendulum* to be magnetized beyond a prescribed limit, a solenoid is used to demagnetize it. The standard equipment for a party engaged in determinations of *gravity* includes a suitably mounted solenoid, with compass needle, battery, and control switches.

solid angle.—See under *angle.*

solstice.—A point on the *ecliptic* midway between (90° from) the *equinoxes.* See *solstice: summer; solstice: winter.*

solstice: summer.—The point on the *ecliptic,* north of the *equator* and midway between (90° from) the *equinoxes.*

The summer solstice is the point where the sun attains its greatest north *declination,* a position which marks the beginning of the season of summer.

solstice: winter.—The point on the *ecliptic,* south of the *equator* and midway between (90° from) the *equinoxes.*

The winter solstice is the point where the sun attains its greatest south *declination,* a position which marks the beginning of the season of winter.

solstitial colure.—See under *colure.*

southing.—See under *latitude: difference of latitude (plane surveying).*

speaking rod.—See under *leveling rod (general).*

special-purpose map.—See under *map.*

spherical aberration.—See under *aberration of light (optics).*

spherical angle.—See under *angle.*

spherical coordinates.—See under *coordinates.*

spherical excess.—The amount by which the sum of the three *angles* of a triangle on a sphere exceeds 180°.

In *geodetic* work, in the computation of triangles, the difference between *spherical angles* and *spheroidal angles* is generally neglected, *spherical angles* are used, and *Legendre's Theorem* applied to the distribution of the spherical excess. That is, approximately one-third of the spherical excess of a given spherical triangle is subtracted from each *angle* of the triangle.

spheroid.—In general, any figure differing but little from a sphere. In *geodesy,* a mathematical figure closely approaching the *geoid* in form and size, and used as a surface of reference for *geodetic surveys.*

In *geodetic surveying,* the spheroid is usually an ellipsoid of revolution, but there is some disadvantage in using the term spheroid as synonomous with such a figure. Until the adoption of the International Gravity Formula (Stockholm, 1930), none of the *gravity* formulas in general use corresponded to an exact ellipsoid of revolution. Helmert's formula of 1917 and some later ones included a term that implied a spheroid with an elliptical *equator,* and *meridians* which were not exact ellipses. Ellipsoids of three unequal axes and spheroids with non-elliptical *meridians* have been proposed by geodesists to secure better representation of observed *deflections of the vertical.* The mathematical difficulties of fitting *triangulation* to any such surface have firmly established the use of an ellipsoid of revolution for that purpose. A spheroid which corresponds to an ellipsoid of revolution may be defined by stating the lengths of the semiaxes (a and b) of the generating ellipse, or the length of one of the semiaxes and the *flattening (ellipticity)* [$(a-b)/a$] of the ellipse.

A spheroid serves as a base for the mathematical reduction (*correlation* and *coordination*) of surveying observations where account must be taken of the curvature of the earth's surface. It is such a figure as would be produced by an ellipse revolving around its minor axis, that is, an oblate ellipsoid. A particular spheroid may be defined by stating the lengths of its semiaxes, a and b, or the length of one semiaxis and the *flattening* (ellipticity), $(a-b)/a$. In the *geodetic* work of this country, the following spheroids have been used: the Bessel spheroid of 1841, for which $a=6{,}377{,}397.2$ meters and $b=6{,}356{,}079.0$ meters; the Clarke spheroid of 1866, adopted in 1880, and in continuous use since that time. For it, $a=6{,}378{,}206.4$ meters; $b=6{,}356{,}583.8$ meters.

Of interest also is the Hayford spheroid, based upon observations over this entire country, in computing which account was taken of the local *deflections of the vertical* by applying corrections for the effects of *topography* and *isostatic compensation.* The Hayford spheroid is defined by $a=6{,}378{,}388$ meters; $b=6{,}356{,}909$ meters. In 1924, the General Assembly of the Section of Geodesy of the International Geodetic and Geophysical Union, meeting in Madrid, adopted the Hayford spheroid as the basis for the International Ellipsoid of Reference, defined as follows: $a=6{,}378{,}388$ meters; *flattening* (ellipticity), $(a-b)/a=1/297$; and by computation, $b=6{,}356{,}911.946$ meters.

While for reasons of expediency, the International Ellipsoid of Reference has not been adopted for survey work in this country, it has considerable scientific value and tables for its use have been computed and published.

spheroid: eccentricity of.—See under *eccentricity of spheroid.*

spheroid: ellipticity of.—See under *flattening of the earth.*

spheroid: lines on a.—Any direct line between two positions on a *spheroid,* represented by two points on the earth.

Such a line may be one of mathematical definition, or it may result from a direct *survey* between the points on the earth. Eight such lines, connecting two points, A and B, are listed in U. S. Coast and Geodetic Survey Report for 1900, pages 369–370, as follows:

(a) A *normal section line* from A to B.

(b) A *normal section line* from B to A.

(c) A *curve of alignment.*

(d) A *geodetic line.*

(e) A forward straight line starting in the plane of the *normal* through A and through B, and in advancing keeping B constantly sighted.

(f) A line similar to (e) above, but starting in the plane of the *normal* through B and through A, and in advancing keeping A constantly sighted.

(g) A forward and backward sight line over limited distances, starting from A.

(h) A forward and backward sight line over limited distances, starting from B.

(Lines (a), (b), (c), and (d) are defined elsewhere.)

spheroidal angle.—See under *angle.*

spheroidal excess.—The amount by which the sum of the three *angles* of a triangle on a *spheroid* exceeds 180°.

In *geodetic* work, *spherical angles* are used instead of *spheroidal angles,* the difference between such *angles* being quite small for the largest triangles, and considered negligible for ordinary triangulation work.

spiral curve.—See under *curve.*

spirit level.—See under *level.*

spirit level axis.—See under *level axis.*

spirit level, chambered.—See under *level.*

spirit leveling.—See under *leveling*.

spur (leveling).—See under *line of levels*.

spur line of levels.—See under *line of levels*.

stadia.—A graduated rod used in the determination of distance by observing the intercept on the rod subtending a small known *angle* at the point of observation.

In practice, the *angle* is usually defined by two fixed lines in the *reticle* of a *telescope* (*transit* or *telescopic alidade*). The term stadia is also used in connection with *surveys* where distances are determined with a stadia, as, stadia survey, stadia method, stadia distance, etc.; also used to designate parts of instruments used, as, stadia wires. In its strictest use, stadia refers to the graduated rod.

stadia (length).—The plural form of stadium, a Greek measure of length somewhat greater than 600 feet.

stadia constant (leveling).—See under *constant*.

stadia interval (leveling).—See under *interval*.

Stampfer level.—See under *leveling instrument*.

standard.*—An exact value, or concept thereof established by authority, custom, or common consent, to serve as a model or rule in the measurement of quantity, or in the establishment of a practice or procedure.

Standard, as it pertains to any specific term, is considered in the supplementary statement of that term.

Standard (California) astronomical datum of 1885.—See under *datum*.

standard elevation.—See under *elevation*.

standard line (United States public-land surveys).*—Any *base line, standard parallel, principal meridian*, or *guide meridian*.

The term standard line is not in general use.

standard of length.—A physical representation of a linear *unit* that is approved by competent authority.

A standard of length is not independent of temperature, pressure, and other physical conditions to which it is subjected, but is an exact embodiment of the *unit* which it represents only under definite, prescribed conditions. See *meter: International Prototype Meter*.

Standards of length are classified according to the *accuracy* and *precision* of their representation of the given *unit*, and these are prescribed by their intended use and are designated accordingly, as follows:

1. Primary standards. International and national prototype *standards* of the highest possible *accuracy* and *precision*.
2. Secondary standards. Laboratory *standards* that are compared directly with the national *standards*.
3. Reference standards. *Standards* suitable for use in the construction of *precision* apparatus to be used in scientific investigations.
4. Working standards. *Standards* suitable for all ordinary *precision* work, as in college laboratories, the general manufacture of *precision* instruments, and for the use of State and city sealers of weights and measures.
5. Commercial standards. *Standards* satisfying the requirements of local sealers of weights and measures, drafting, machine work, etc.

Standards of length are also classified according to their design: 1. Line standards. A *standard* in which the *unit* of length is defined as the distance between lines or marks on the surface of the *standard*, under prescribed conditions. 2. End standard. A *standard* in which the *unit* of length is defined as the distance between the end surfaces of the *standard*, under prescribed conditions. Also termed a contact standard.

standard parallel (United States public-land surveys).—See under *parallel*.

standard survey.—See under *survey*.

standard time.—See under *time*.

standard topographic map of the United States.—See under *map: Topographic Map of the United States*.

standardization.—The comparison of an instrument or device with a *standard* to determine the value of the instrument or device in terms of an adopted *unit*.

A tape is standardized when the distance between its *fiducial marks* in terms of a *unit* of length is determined by comparing it, under prescribed conditions, with a *standard* which represents that *unit*. See *calibration*. A *pendulum* is standardized by determining its period at a base station at frequent intervals of time, care being taken each time to swing the *pendulum* under as nearly the same conditions as practicable, and correcting the results to the adopted references of temperature, air pressure in the *receiver, amplitude of vibration*, and *flexure*.

standardization.—(Alternate definition.) The determination of the physical characteristics of a gage or measuring device by comparing it with a *standard*.

A gage or measuring device is standardized so that measures made with it may be reported and used in terms of the *unit* which the *standard* exemplifies. The term "standardization" may include the determination of a quantity which is required in reducing the direct results of a standardization to corresponding values under changing conditions. For example, when a *base tape* is standardized, its length is determined at a definite temperature, under prescribed conditions of tension and support. Its use in the field requires also that its *coefficient of thermal expansion* be known.

State base map.—See under *map*.

State coordinate systems.—The *plane-rectangular coordinate* systems established by the U. S. Coast and Geodetic Survey, one for each State in the Union, for use in defining *positions* of *geodetic* stations in terms of *plane-rectangular* (x and y) *coordinates*.

Each State is covered by one or more zones, over each of which is placed a *grid* imposed upon a *conformal map projection*. The relationship between the *grid* and the *map projection* is established by mathematical analysis. Zones of limited east-west dimension and indefinite north-south extent have the *transverse Mercator map projection* as the base for the State coordinate system; while zones for which the above order of magnitude is reversed use the *Lambert conformal conic map projection* with two *standard parallels*. For a zone having a width of 158 *statute miles*, the greatest departure from exact *scale* (scale error) is 1 part in 10,000. Only *adjusted positions* on the *North American datum of 1927* may be properly transformed into *plane coordinates* on a State system. All such *geodetic positions* which are determined by the U. S. Coast and Geodetic Survey are transformed into State *plane-rectangular coordinates* on the proper *grid*, and are distributed by that Bureau with the *geodetic positions*. State *plane coordinates* are extensively used in recording *land surveys*, and in twenty-two States such use has received approval by legislative enactment.

State map.—See under *map*.

static map.—See under *map*.

station.—A definite point on the earth whose location has been determined by *surveying* methods. A point on a *traverse* over which an instrument is placed (a *set-up*). Also, on a *traverse,* a length of 100 feet measured on a given line, broken, straight, or curved.

A station may or may not be marked on the ground. A station is usually defined by the addition of a term which describes its origin or purpose, such as *triangulation station,* topographic station, magnetic station, etc. A station is often marked on the ground by a *monument* of special construction, or by a natural or artificial structure.

station: instrument (leveling).—The point over which a *leveling instrument* is placed for the purpose of taking a *backsight,* a *foresight,* and such *extra foresights* as may be necessary from that instrument station. See, also, *set-up.*

Except in rare instances, the instrument stations in a *line of levels* are not marked points such as traverse *stations,* triangulation *stations,* etc.

station: intersection.—An object whose horizontal *position* is determined by observations from other survey stations, no observations being made at the object itself.

Where the object is observed from only two *stations,* the position is termed a "no check position," as there is no proof that such observations are free from *blunders.* Intersection stations are either objects which it would be difficult to occupy with an instrument, or survey signals whose positions can be determined with sufficient *accuracy* without being occupied.

station: principal.—A *station* through which basic data are carried in the extension of a survey system.

Also called a main-scheme station, a principal *station* serves primarily for the continued extension of a *survey.* This requires a higher order of *accuracy* and *precision* in its determination than if its purpose were limited to the control of local surveys or the establishment of *supplementary stations.*

station: subsidiary.—A *station* established to overcome some local obstacle to the progress of a *survey,* and not to determine position data for the *station* point.

The term subsidiary station is usually applied to *A-stations* of a *traverse survey.* Subsidiary stations are usually temporary in character and not permanently marked. If serving the additional purpose of supplying control for a local *survey,* such a *station* may be permanently marked and it is then a *supplementary station.*

station: supplementary.—An auxiliary survey *station,* established to increase the number of control *stations* in a given area, or to place a *station* in a desired location where it is impracticable or unnecessary to establish a *principal station.*

Supplementary stations are permanently marked, and are established with an *accuracy* and *precision* somewhat lower than is required for a *principal station,* since they do not serve as bases from which extensive *surveys* are run. See *station: subsidiary.*

station adjustment.—See under *adjustment.*

station error.—See under *deflection of the plumb line.*

station pointer.—See under *protractor: three-arm protractor.*

statute mile.—See under *mile.*

Stephenson leveling rod.—See under *leveling rod.*

steradian.—See under *angle: solid.*

stereographic horizon projection.—See under *map projection: stereographic map projection.*

stereographic map projection.—See under *map projection.*

stereographic meridional projection.—See under *map projection: stereographic map projection.*

stereographic polar projection.—See under *map projection: stereographic map projection.*

strength of figure (triangulation).—The comparative *precision* of computed lengths in a *triangulation net* as determined by the size of the *angles,* the number of *conditions* to be satisfied, and the distribution of *base lines* and points of *fixed position.*

Strength of figure in *triangulation* is not based on an absolute scale but rather is an expression of relative strength. The number expressing the strength of a *triangulation* figure is really a measure of its weakness, since the number determined by formula increases in size as the strength decreases.

The strength of figure is derived from that portion of the formula for *probable error* of a triangle side, which is independent of the *accuracy* of the observations, as follows:

$$\frac{N_d-N_c}{N_d}\sum\left[\delta_A^2+\delta_A\delta_B+\delta_B^2\right]$$

in which N_d and N_c are the numbers of directions observed and of *conditions* to be satisfied, and δ_A and δ_B are the rates of change of the sines of the *distance angles* A and B, usually expressed by the differences of the logarithms of the sines for a difference of 1″ in the *angles,* the sixth decimal place being the unit place.

By summing up the values obtained by formula for the simple figures composing a *triangulation net,* the strength of figure of the net is obtained. As a *triangulation net* is usually composed of several different systems of simple figures, comparable values of the different systems are obtained, and the strongest route can then be selected through which to carry a computation of length.

Reconnaissance for a proposed *triangulation net* is usually executed under instructions which specify limiting values of strength of figure for the best and second-best chains of triangles between adjacent *base lines,* the sites for stations and for *base lines* being selected accordingly. Where desirable, the length of section may be reduced by the insertion of an additional *base line,* and the numbers representing the strength of figure reduced accordingly.

stride.—The distance a pedestrian moves in taking two steps.

Formerly, the stride and the *pace* were largely synonomous; if there was a difference, it was that the stride was a long *pace.* The above definition conforms to present-day teaching, and satisfies the need for a term expressing the distance moved in taking a double step. Hence, one stride equals two *paces.*

striding level.—See under *level.*

subsidiary station.—See under *station.*

summer solstice.—See under *solstice.*

Sumner line (line of position).—A short portion of a *circle of position* represented as a *straight line.*

The Sumner line was named for Captain Thomas H. Sumner, an American sea captain, who discovered its use in 1843. The St. Hilaire Method is a special adaptation of the use of the Sumner line in navigation.

sun compass.—See under *compass.*

supplement of angle.—See under *angle.*

supplemental posts for survey monuments.—See under *identification posts.*

supplementary bench mark.—See under *bench mark.*

supplementary station.—See under *station.*

survey.—1. The orderly process of determining data relating to any physical or chemical characteristics of the earth. 2. The associated data obtained in a survey (1, above). 3. An organization engaged in making a survey (1, above).

1. The list of orderly processes which may be properly called surveys is long; it may be divided into

classes according to the type of data obtained, the methods and instruments used, the purposes to be served, etc. Examples: *geodetic survey, topographic survey, hydrographic survey, land survey, geologic survey,* geophysical survey, soil survey, *mine survey, engineering survey,* etc.

2. The data obtained in a particular project may be designated by the name of the project, as, "the topographic survey of the District of Columbia."

3. An organization making a survey is often given an official name which includes the word survey; examples, "The United States Geological Survey," "The Massachusetts Geodetic Survey."

survey: cadastral.*—A *survey* relating to *land boundaries* and subdivisions, made to create units suitable for transfer or to define the limitations of title.

Derived from "cadastre," meaning register of the real property of a political subdivision with details of area, ownership, and value, the term cadastral survey is now used to designate the *surveys* of the *public lands* of the United States, including retracement *surveys* for the identification, and *resurveys* for the restoration, of property lines; the term may also be applied properly to corresponding *surveys* outside the *public lands,* although such *surveys* are usually termed *land surveys* through preference.

survey: engineering.*—A *survey* executed for the purpose of obtaining information that is essential for planning an engineering project or development and estimating its cost.

The information obtained may, in part, be recorded in the form of an *engineering map.*

survey: exploratory.*—A *survey* executed for the purpose of obtaining general information concerning areas about which such information was not, theretofore, a matter of record.

The U. S. Geological Survey executes exploratory *surveys* in remote regions of Alaska which are used in the construction of *maps* on *scales* of 8 or more *miles* to an *inch.*

survey: geodetic.—1. A *survey* in which account is taken of the figure and size of the earth. 2. An organization engaged in making geodetic surveys.

a. Geodetic surveys are usually prescribed where the areas or distances involved are so great that results of desired *accuracy* and *precision* can be obtained only by the processes of geodetic surveying.

b. The Massachusetts Geodetic Survey is an example of an organization engaged in making geodetic surveys.

survey: geographic.—A general term, not susceptible of defined limitation, covering a wide range of *surveys* lying between and merging into *exploratory surveys* on the one hand and basic *topographic surveys* on the other.

Geographic surveys usually cover large areas, are based on coordinated control, and are used to record physical and statistical characteristics of the area surveyed.

survey: geologic.—A *survey* or investigation of the character and structure of the earth, of the physical changes which the earth's crust has undergone or is undergoing, and of the causes producing those changes.

survey: geological.—A general term used to designate an organization making *geologic surveys* and investigations.

survey: hydrographic.—A *survey* having for its principal purpose the determination of data relating to bodies of water.

A hydrographic survey may consist of the determination of one or several of the following classes of data: depth of water and configuration of bottom; directions and force of currents; heights and times of tides and water stages; location of fixed objects for *survey* and navigation purposes.

survey: inventory.—A *survey* for the purpose of collecting and correlating engineering data of a particular type (or types) over a given area.

An inventory survey may be recorded on a *base map.*

survey: land.—See under *surveying: land surveying.*

survey: location.—The establishment on the ground of points and lines in positions which have been previously determined by computation or by graphical methods.

The plans for an engineering project (road, canal, etc.) are prepared in the office from *survey* data obtained in the field. These plans form a paper location, and are the basis for the location survey.

survey: mine.—A *survey* to determine the positions and dimensions of underground passages of a mine; also of the natural and artificial features, surface and underground, relating to the mine.

The data include both horizontal and vertical *positions,* lengths, and directions (slopes) of tunnels; topographic and geological characteristics of the particular vicinity; ownership of the land and of the mine; etc.

survey: plane.—See under *surveying: plane surveying.*

survey: preliminary.—The collection of survey data on which to base studies for a proposed project.

A triangulation reconnaissance is a preliminary survey, securing data on which to base plans for the execution of the *triangulation.*

survey: reconnaissance.*—A *preliminary survey.*

A reconnaissance survey is usually executed rapidly and at relatively low cost. The information obtained is recorded, to some extent, in the form of a *reconnaissance map* or *sketch.* The form reconnaissance is preferred to reconnoissance.

survey: resurvey.—A retracing on the ground of the lines of an earlier *survey,* in which all points of the earlier *survey* that are recovered are held fixed and used as a control.

If too few points of the earlier *survey* are recovered to satisfy the control requirements of the resurvey, a new *survey* may be made. A resurvey is related directly to an original *survey,* though several resurveys may interpose between them. The terms original *survey,* resurvey, and new *survey* are of considerable significance in *land surveying.*

survey: standard.—A *survey* which, in *scale, accuracy,* and content, satisfies criteria prescribed for such a *survey* by competent authority.

survey: topographic.—A *survey* which has for its major purpose the determination of the configuration (*relief*) of the surface of the earth (ground), and the location of natural and artificial objects thereon. Also the designation of an organization making such a *survey.*

A topographic survey is usually published as a *topographic map.*

survey: transit-and-stadia.—A *survey* in which *directions,* horizontal and vertical, are observed with a *transit* (engineer's), and distances are measured with a *transit* and *stadia.*

A transit-and-stadia survey is usually processed in the office, thus differing from a *survey* made with *alidade* and plane-table, where the *map* is made in the field as the observations are taken. An exception is found in the *topographic survey* of Baltimore, made in the 1890's, where *transit* and plane-table were used together in the field, and the *map* constructed as the observations were made.

survey net.—1. Horizontal control survey net: arcs of *triangulation,* sometimes with lines of *traverse,* connected together to form a system of loops or circuits extending over an area. Variously spoken of as a triangulation net, traverse net, etc.

2. Vertical control survey net: lines of *spirit leveling* connected together to form a system of loops or circuits extending over an area. Also called a level net. See *control: national control survey nets.*

survey signal.—See under *signal.*

survey traverse.—See under *traverse.*

surveying.—The art of making a *survey.*

surveying: geodetic.—That branch of the art of *surveying* in which account is taken of the figure and size of the earth. Also called geodetic engineering.

In geodetic surveying, prescribed *precision* and *accuracy* of results are obtained through the use of special instruments and field methods, and formulas based on the geometry of a mathematical figure approximating the earth in form and size.

surveying: land.—The determination of boundaries and areas of tracts of land.

Land boundaries are usually defined by ownership, commencing with the earliest owners, and descending through successive ownerships and partitions. Land surveying includes the re-establishment of original boundaries, and the establishment of such new boundaries as may be required in the partition of the land. The term *"cadastral survey"* is sometimes used to designate a land survey, but in this country its use should be restricted to the surveys of the *public lands of the United States.*

surveying: plane.—A branch of the art of *surveying* in which the surface of the earth is considered a plane surface.

In plane surveying, curvature of the earth is neglected and computations are made using the formulas of plane geometry and plane trigonometry. In general, plane surveying is applied to *surveys* of land areas and boundaries (*land surveying*), where the areas are of limited extent or the required *accuracy* so low that *corrections* for the effect of curvature would be negligible as compared with the *errors* of observation. For small areas, precise results may be obtained with plane-surveying methods, but the *accuracy* and *precision* of such results will decrease as the area surveyed increases in size.

surveying: trilinear.—The determination of the position of a point of observation by measuring the *angles* at that point between lines to three points of known position. See under *resection.*

This determination involves the solution of the *three-point problem,* which is accomplished analytically by computation or graphically by the use of chorograph or *station pointer,* such as the *three-arm protractor.*

surveyor's compass.—See under *compass.*

surveys and maps: detailed.—*Surveys* and *maps* which take account of the smaller features and variations of features in the area surveyed.

A detailed *survey* implies a large-scale *map* of a small area, with increased *accuracy* and *precision* in both field and office work.

synodical (ordinary) month.—See under *month.*

systematic error.—See under *error.*

T

tachymeter, tacheometer, tachometer.—A surveying instrument designed for use in the rapid determination of distance, direction, and difference of *elevation* from a single observation, using a short base, which may be an integral part of the instrument.

There are several forms of instruments which might be called tachymeters. 1. An instrument in which the base line for distance determinations is an integral part of the instrument. It is to this type of instrument that the term "tachymeter" is usually applied. The so-called *range finders* with self-contained bases belong to this class, but are usually spoken of as *range finders,* and do not usually afford the means for the determination of elevation. 2. An instrument equipped with *stadia* wires or *gradienter,* the base for distance determination being a graduated rod held at the distant point.

tachymetry; tachometry.—A method of *surveying* for the rapid determination of distance, *direction,* and relative *elevation* of a point with respect to the instrument station, by a single *observation* on a rod or other object at the point.

The stadia method of *surveying* is the best example of tachymetry in this country, but the term tachymetry is little used here. Little used, too, are the associated terms given types of instruments used in this work, namely, *tachymeter* (or *tacheometer, tachometer*), and tacheographometer.

Talcott method, latitude determination.—See under *zenith-telescope method, latitude determination.*

tangency, point of.—See under *curve.*

tangent (surveying).—1. A straight line that touches a given curve at one and only one point, and does not intersect it.

2. That part of a *traverse* or *alignment* included between the *point of tangency* of one curve and the *point of curvature* of the next curve.

3. In the *surveys of the public lands of the United States,* a *great circle line* tangent to a *parallel* of *latitude* at a *township corner.*

tangent distance.—The distance from the *point of intersection* (vertex) of a curve to its *point of tangency* or *point of curvature.*

tangent method (United States public-land surveys).*—A method of determination of the *parallel* of *latitude* for the survey of a *base line* or *standard parallel* by *offsets* from a *great circle line* initiated at an established *township corner* and tangent to the *base line* or *standard parallel* at that corner.

The tangent *great circle* is projected at an angle of 90° from the *meridian* at the *township corner* from which initiated, and proper *offsets* are measured north from the tangent to the *parallel* upon which the *corners* are established.

tangent plane.—A plane that touches a curved surface at one and only one point.

In *geodetic* work, a plane tangent to the *spheroid* at any point is perpendicular to the *normal* at that point. Tangent planes have been used in computing *map projections* for small areas (local systems of *map projections*).

tape: alignment correction.—A *correction* applied to the measured length of a line to allow for the tape not being held exactly in a *vertical plane* containing the line.

Though the *vertical planes* at the two ends of a line which contain the *direction* of the line at those points are not coincident, they are practically so for all *surveying* purposes, and any method of aligning a base between its end points will serve in determining an alignment correction.

tape: base.—See under *base tape.*

tape: grade correction.—A *correction* applied to a distance measured on a slope to reduce it to a horizontal distance between the vertical lines through its end points. Also called slope correction, correction for inclination of tape.

As the vertical lines through the end points of a line are not parallel, the reduction should be such that the length at the average *elevation* will be obtained. In practice this is accomplished by considering each tape length a separate line, and reducing it individually. In *base line* measures, the difference in *elevation* of the ends of the tape are used in computing the grade correction, rather than the *grade* or inclination of the line expressed in angular form. In effect, this treatment applies the grade correction to the nominal length of the tape in determining its effective length in measuring a distance.

tape: invar.—See under *base tape.*

tape: length correction.—The difference between the nominal length of a tape and its effective length under conditions of *standardization.*

The standard length of a tape is usually expressed by a number of whole units (the nominal length) plus or minus a small distance which is the length correction defined above. Provision is also made for determining any changes of length caused by thermal expansion (or contraction) of the tape.

tape: sag correction (catenary correction).—The difference between the effective length of a tape (or part of a tape) when supported continuously throughout its length and when supported at a limited number of independent points.

Base tapes are usually used with 3 or 5 points of support, and hang in curves (catenaries) between adjacent supports. Correction for sag is not required when the method of support in use is the same as was used in the *standardization* of the tape. A *base tape* may also be used supported throughout or with 4 points of support, as on a railway rail.

tape: temperature correction.—The quantity applied to the nominal length of a tape to allow for a change in its effective length due to its being used at a temperature other than that for which its *standard* length is given.

Temperature correction is applied when the temperature of the tape in use is different from the temperature given in a statement of its standard length.

tape: tension correction.—The *correction* applied to the nominal length of a tape to allow for a change in effective length due to its being used at a tension other than that for which its standard length is known.

A tension correction is not required when the tape is used in the field at the tension of *standardization.*

tape corrections.—*Corrections* applied to a distance measured with a tape to eliminate *errors* due to the physical condition of the tape, and to the manner in which it is used.

Tape corrections which are functions of the physical condition of the tape are *length correction, temperature correction,* and *tension correction.* A *correction* which may be required because of the method of using the tape is the *catenary correction* (*correction for sag*). Closely associated with these are *corrections* which depend upon the position of the line on the ground, as actually measured. These include *grade correction,* and *alignment correction,* which are sometimes treated as *corrections* to the tape itself as used. Reduction to a *datum* (such as *mean sea level*) is not considered as a tape correction.

tape stretcher.—A mechanical device which facilitates holding a tape at a prescribed tension and in a prescribed position.

taping.—The operation of measuring a distance on the earth, using a tape or ribbon of metal or other material. See *chaining.*

The term taping is used for all *surveys* except those of the *public lands of the United States;* for those, due to historical and legal reasons, the term *chaining,* and its related term *chainmen,* are preferred. The men who mark the two ends in taping are termed contact men.

target rod.—See under *leveling rod (general).*

telescope.—An optical intrument consisting essentially of two systems of lenses: an objective or object glass which brings rays of light from a distant object to a *focus* within the telescope tube; and an eyepiece or ocular with which the image of the distant object formed by the objective is magnified and examined.

By means of a *reticle* placed in the principal *focus* of a telescope, definite lines of sight are established, and the telescope thereby made into a *precision* instrument for *surveying* and astronomic work. The magnifying power of a telescope depends upon the ratio of the *focal lengths* of the objective and eyepiece lenses.

telescope: achromatic.—See under *achromatic telescope.*

temperature correction (leveling).—See under *correction.*

temperature correction, pendulum.—See under *pendulum.*

temperature correction, tape.—See under *tape.*

template (templet).—A gage or pattern to guide certain work that is to be done.

A template is often a thin sheet of transparent material upon which suitable guide lines are drawn. Examples: the ordinary *protractor* or, in photographic mapping, a transparent celluloid overlay made over a photograph, showing the center and all radial lines from the center through images of control points, as well as (azimuth) lines connecting the center with images of points which show on the photograph and are themselves the centers of other photographs. In obtaining *elevation* readings from *maps* in connection with *topographic* and *isostatic* reductions, templates of celluloid are used. The templates used in connection with *deflection of the vertical* studies have circles and radial lines drawn upon them, so proportioned with reference to *scale of map* and *azimuth* that land *elevations* and ocean depths within each compartment formed by adjacent arcs and radii can be easily averaged and the effect of the mass therein upon a plumb line at the station (center of circles), under various hypotheses, can be computed. The templates used in connection with *gravity* studies are similar to those described above, except that no account is taken of *azimuth,* all compartments bounded by a given pair of circles being of the same size and shape. A given template can be used only on maps of the *scale* and *projection* for which it is constructed. Devised by J. F. Hayford in connection with studies for the figure of the earth and *isostasy,* these templates are known as the Hayford deflection templates and the Hayford gravity templates.

Temple Act.—An Act of the Congress of the United States to adopt a program for the completion of the *Topographic Map of the United States.*

Known as H. R. 4522, 68th Congress, 1st Session, the bill which later (1925) became the Temple Act, was introduced in 1924 by Dr. Henry W. Temple, a

Representative from Pennsylvania. No appropriations were made for carrying out its provisions.

temporary bench mark.—See under *bench mark.*

tension correction, tape.—See under *tape.*

terrain correction (geländereduktion).—The correction for the effect, upon the acceleration of gravity at a point, of the departure of the actual topography from an assumed plateau at the elevation of the point.

The correction has the same sign whether the topography lies above or below the plateau.

terrestrial equator.—See under *equator: astronomic.*

terrestrial meridian.—See under *meridian.*

terrestrial refraction.—See under *refraction.*

tertiary triangulation.—See under *triangulation.*

thalweg.—The line following the lowest part of a valley, whether under water or not.

The intricacy of detail in ordinary *relief* often makes difficult a practical location of a thalweg; in a survey of a political *boundary line* this difficulty may assume considerable weight. A thalweg may also be defined as the line down the center of the main channel of a stream; or as the line of greatest slope, cutting all *contours* at right *angles.*

theodolite.—A *precision surveying* instrument consisting of an *alidade* with *telescope,* mounted on an accurately graduated circle, equipped with necessary levels, and reading devices. Sometimes, the *alidade* carries a graduated *vertical circle.*

There are two general classes of theodolites: *direction theodolites,* often referred to as *direction instruments,* and *repeating theodolites.*

theodolite: astronomical.—See under *alt-azimuth instrument.*

theodolite: direction instrument.—A *theodolite* in which the graduated horizontal circle remains fixed during a series of observations, the *telescope* being pointed on a number of signals or objects in succession, and the *direction* of each read on the circle, usually by means of *micrometer microscopes.*

In measuring *horizontal angles* with a direction instrument, *angles* are not repeated (accumulated) on the circle, but *precision* and *accuracy* are obtained by having the circle of high quality, by using *precision* methods of reading the circle, and by shifting the circle between sets, so that each *direction* is measured on a number of different parts of the circle. Direction instruments are used almost exclusively in *first-* and *second-order triangulation.*

theodolite: Parkhurst.—A *direction theodolite* designed by and named for D. L. Parkhurst, and used on *first-* and *second-order control surveys.*

The distinctive features of the Parkhurst theodolite are its non-binding center, ball-bearing clamp ring, illumination through the central axis, discontinuous conical bearings, illuminated glass micrometer drums, and improved designs for tangent-screw assembly, clamp-block assembly, and micrometer mountings.

theodolite: photo. —An instrument used in terrestrial photogrammetry, and consisting of a combination of *theodolite* and camera, using the same tripod.

In some models, the theodolite and camera can be used separately; in other models, the two are combined in a single instrument.

theodolite: repeating.—A *theodolite* so designed that successive measures of an *angle* may be accumulated on the graduated circle, and a final reading of the circle made which represents the sum of the repetitions.

The observed value of the *angle* is obtained by dividing the total *arc* passed through in making the series of observations by the number of times the *angle* has been observed. The total *arc* passed through may include several complete circuits of the circle, which must be added to the circle reading before making the division. The repeating *theodolite* is also called a repeating instrument. Theoretically, it is an instrument of great *precision,* but in its mechanical operation it does not give as satisfactory results as the *direction instrument.*

theodolite: Wild.—A make of *theodolite* which has the horizontal and *vertical circles* graduated on glass, and read by an arrangement of prisms which brings the images of opposite portions of either of the circles into the field of view of a reading microscope mounted alongside the *telescope.* By mechanical means the graduation marks in the two images are brought into coincidence, and the circle read by means of an auxiliary seconds scale.

The seconds scale is read to single seconds or to a decimal part of a second without moving from the eye end of the *telescope.* A single reading thus obtained is the mean of two readings of diametrically opposite points on the graduated circle and is therefore free from *errors* due to *eccentricity of circle.* The Wild theodolite is well adapted for use in *triangulation* of any order.

theoretical (or external) error.—See under *error.*

thermometric leveling.—See under *leveling.*

third-order leveling.—See under *leveling.*

third-order traverse.—See under *traverse.*

third-order triangulation.—See under *triangulation.*

third-order work.—See under *traverse; triangulation;* and *leveling.*

thread of river, middle.—The line equidistant between the *low-water* lines on the two sides of a river, extending from headland to headland without considering arms, inlets, creeks, and affluents as parts of the river.

thread of stream.—The line equidistant from the edge of the water on the two sides of the stream at the ordinary stage of the water. See under *thread of river, middle.*

three-arm protractor.—See under *protractor.*

three-point problem.—The determination of the horizontal *position* of a point of observation from data comprising two observed *horizontal angles* between three objects of known *position.*

The problem is solved graphically by means of a *three-arm protractor,* and analytically by trigonometrical calculation. See *resection.* In plane-table surveying it is solved graphically by trial and error.

tidal bench mark.—See under *bench mark.*

tidal datum.—See under *datum.*

tideland.—Land along tidal waters that is below *mean high water* and above *mean low water,* that is, the *shore.*

tie.—A survey connection from a point of known *position* to a point whose *position* is desired.

A tie is made to determine the *position* of a supplementary point whose *position* is desired for mapping or reference purposes, or to close a *survey* on a previously determined point. To "tie in" is to make such a connection. The point to which the connection is made is called a "tie point."

tie point.—See under *tie.*

tier (United States public-land surveys).*—Any series of contiguous *townships* situated east and west of each other; also *sections* similarly situated within a *township.*

time.—The measurable aspect of duration.

Time makes use of scales based upon the happening of periodic events, as follows:

1. The *day,* depending upon the *rotation* of the earth.
2. The *month,* depending upon the *revolution* of the moon around the earth.
3. The *year,* depending upon the *revolution* of the earth around the sun.

When one says it is a certain time, one expresses a length on a duration scale measured from an index on that scale.

For example: 4 p. m. local *mean solar time* means that 4 mean solar hours have elapsed since the *mean sun* was on the *meridian* of the observer.

time: apparent.—See under *time: apparent solar time.*

time: apparent solar.—*Time* measured by the apparent diurnal motion of the (true) sun. Also termed true solar time, and often, apparent time.

At any given instant, the apparent solar time is the *hour angle* of the (true) sun. In civil life, apparent solar time is counted from the two branches of the *meridian* through 12 hours; the hours from the lower branch are marked "A. M." (ante meridian); those from the upper branch, "P. M." (post meridian). In astronomic work, before 1925, apparent solar time was counted from the upper branch of the *meridian* through 24 hours (see *astronomical time*); since 1925, the count has been from the lower branch, and the *civil day* has taken the place of the *astronomical day* in astronomical work. Naming the *meridian* of reference is essential to the complete identification of the *time.* For example, 75th meridian apparent solar time; Greenwich apparent solar time; local apparent solar time (for the *meridian* of the observer). Because of variations in the apparent motion of the sun, a clock or watch cannot be adjusted to show apparent solar time. Apparent solar time is *time* determined by observations on the (true) sun. It is the *time* shown by a sun dial. See *time: equation of time.*

time: astronomical.—*Solar time* in a *day* (*astronomical day*) that begins at noon.

Astronomical time may be either *apparent solar time* or *mean solar time.* It is counted from noon in a single series of 24 hours. Astronomical time appears in the American Ephemeris and Nautical Almanac prior to 1925; but in the volumes for 1925 and later, *civil time* is used instead.

time: civil.—*Solar time* in a day (*civil day*) that begins at midnight.

Civil time may be either *apparent solar time* or *mean solar time;* it may be counted in two series of 12 hours each, beginning at midnight, marked "A. M." (ante meridian), and at noon, marked "P. M." (post meridian), or in a single series of 24 hours beginning at midnight.

time: daylight-saving.—A substitute for *standard time,* exactly 1 hour fast on that *time.*

Clocks showing *standard time* are changed to daylight-saving time by moving the hands ahead exactly 1 hour. Until 1942, it was not adopted by the railroads, nor generally throughout the country; and where adopted, it was for use only during the summer months. In 1942, it was adopted by Congress for general use, and unofficially given the name of *war time.* The time in a given time belt is identified by the particular designation of the belt: Eastern Daylight Saving Time; Central Daylight Saving Time; etc.

time: equation of.—The difference in *hour angle* between *apparent solar time* and *mean solar time.*

The sun is sometimes before and sometimes behind the *mean sun* by an amount that ranges from 0 to about 16 minutes of *time.* As the equation of time may be expressed as a *correction* to either *apparent solar time* or to *mean solar time,* its sign must be carefully observed.

Time: Greenwich Civil.—*Mean solar time* for the *Greenwich Meridian,* counted from midnight. Also called *Universal Time* (U. T.).

Greenwich Civil Time (G. C. T.) is a reference used in the American Ephemeris and Nautical Almanac for 1925 and thereafter. It is counted from midnight through 24 hours.

Time: Greenwich Sidereal.—The *sidereal time* for the *Greenwich Meridian.*

time: local apparent.—The *apparent solar time* for the *meridian* of the observer.

time: local mean.—The *mean solar time* for the *meridian* of the observer.

time: local sidereal.—The *sidereal time* for the *meridian* of the observer.

time: mean solar.—*Time* measured by the diurnal motion of a fictitious body, called the mean sun, which is supposed to move uniformly in the *celestial equator,* completing the circuit in one *tropical year.* Often termed simply mean time.

The mean sun may be considered as moving in the *celestial equator* and having a *right ascension* equal to the mean *celestial longitude* of the true sun.

At any given instant, mean solar time is the *hour angle* of the mean sun. In civil life, mean solar time is counted from the two branches of the *meridian* through 12 hours; the hours from the lower branch are marked "A. M." (ante meridian); those from the upper branch, "P. M." (post meridian). In astronomical work, before 1925, mean solar time was counted from the upper branch of the *meridian* through 24 hours (see *astronomical time*); since 1925, the count has been from the lower branch. Naming the *meridian* of reference is essential to the complete identification of the *time.* For example, 75th meridian mean solar time; Greenwich mean solar time; local mean solar time for the *meridian* of the observer. By using the same reference *meridian* over a belt or zone of the earth, watches and clocks are adjusted to show the same mean solar time throughout the area (see *standard time*). The *Greenwich Meridian* is the reference for a world-wide standard of mean solar time, called *Greenwich Civil Time* or *Universal Time.* The mean sun being a fictitious body, mean solar time cannot be determined directly by observation. See *time: equation of time.*

time: sidereal.—*Time* measured by the apparent diurnal rotation of the (true) *vernal equinox.*

At any instant, sidereal time is the *hour angle* of the *vernal equinox.* It is counted from 0 hours, when the *vernal equinox* is on the *meridian,* through 24 hours. Naming the *meridian* of reference is essential to its complete identification. For example, 75th meridian sidereal time; *Greenwich sidereal time; local sidereal time* for the *meridian* of the observer.

On account of small differences between the positions of the true and mean *equinoxes,* sidereal time, like the *sidereal day,* is subject to slight irregularities. These irregularities are absent from (uniform) sidereal time, which is measured by the motion of the mean equinox, and is used in rating clocks of the highest *precision.*

time: solar.—See under *time: apparent solar time; mean solar time.*

time: standard.—*Mean solar time* for a selected *meridian* adopted for use throughout a belt (zone).

In the continental United States, the *meridians* of reference for standard time are 15° (1 hour) apart, and multiples of 15° from the initial *meridian, Greenwich Meridian.* The standard time for each belt is designated by the number of its *meridian,* and also by some name of *geographic* significance: 75th meridian or Eastern Standard Time; 90th meridian or Central Standard Time; 105th meridian or Mountain Standard Time; and 120th meridian or Pacific Standard Time. The standard time meridians for Alaska are 150° and 165° west *longitude;* for Hawaii, 150° west *longitude;* for the Philippine Islands, 120° east *longitude.* Standard time was established in 1883 to correlate train schedules of various railroads over the same areas. The standard time belts were planned to be roughly symmetrical with respect to the *meridians* of reference, and to extend 7½° to either side thereof. Practical considerations, such as the need of time correlation of cities outside the original boundaries of a time belt with cities within the belt, have caused a gradual shifting of those boundaries until some of them now exhibit large irregularities. See *time: daylight-saving time; war time.*

time: true solar.—See under *time: apparent solar time.*

Time: Universal (U. T.).—The same as *Greenwich Civil Time (G. C. T.).*

The first use of Universal Time as a reference in the American Ephemeris and Nautical Almanac is in the volume for the year 1939.

time: war.—*Daylight-saving time* as prescribed in 1942 by Act of Congress.

The use of war time throughout the country was inaugurated on February 9, 1942, clocks being set ahead 1 hour at 2 A. M. standard time on that date (Public Law 403, 77th Congress). Its use by State and local agencies was not mandatory, except where State legislation so directed. The time in any given *time* belt was identified by the particular designation of that belt: Eastern War Time; Central War Time; etc. The use of war time was discontinued on September 30, 1945, at 2 A. M.

toise.—A unit of length used in early *geodetic surveys* and equal to about 6.4 English feet.

The Toise of Peru was an iron *standard* used in measuring the *base lines* which controlled the lengths of the Peruvian Arc (of *triangulation*), executed in 1736–1743, in connection with a determination of the figure and size of the earth. The Toise of Peru became the legal *standard* of length in France, and from it was derived the *French legal meter,* as follows: the toise was divided into 6 pieds (feet); each pied was divided into 12 pouces (inches); each pouce was divided into 12 lignes (lines); and one *French legal meter* was equal to 443.296 lignes exactly at a temperature of 13° Réaumur. The Peruvian Arc does not lie in modern Peru, but in Ecuador, which was included in the old Spanish presidency of Peru.

Toise of Peru.—See under *toise.*

tolerance (general).—The permissible deviation from a *standard.*

In *surveying* operations the prescribed limiting *probable error* may be regarded as an expression of tolerance, though the term tolerance is seldom used in connection with *survey* results. Its use may well be confined to stating the permissible deviation from exactness in the fitting together of the component parts of physical apparatus: for example, the *pivot inequality* of an *astronomical transit.*

topographic deflection.—That portion of the *deflection of the plumb line* which is due to the gravitational pull exerted by topographic masses.

Topographic deflection is not the same as *deflection of the plumb line* or *station error,* but is the theoretical effect produced by the resultant gravitational pull of the unevenly distributed topographic masses around the station, no allowance being made for *isostatic compensation.*

topographic feature.*—A prominent or conspicuous *topographic form* or salient part thereof.

topographic form.*—A *land form* considered without regard to origin, cause, or history.

topographic map.—See under *map.*

Topographic Map of the United States.—See under *map.*

topography (or relief).*—The features of the actual surface of the earth, considered collectively as to form.

A single feature, such as a mountain or valley, is called a *topographic feature.*

torsion balance.—A device for measuring very small forces of attraction due to *gravitation,* magnetism, etc., and consisting of a bar suspended horizontally by an elastic filament, one end of the bar being subjected to the influence of the attracting force in a much greater degree than the other end. The attracting force is balanced and its comparative strength measured by the torsional reaction of the filament.

In geophysical prospecting, a form of torsion balance is employed in determining the gradient of *gravity,* the *horizontal direction* in which the intensity of *gravity* increases or decreases most rapidly, so that in a *survey* of a large area, any existing large mass of abnormal density may be discovered. For example, the oil fields of the Gulf Coast are associated with salt domes—plugs of salt of such size and density that they are easily discovered and located by geophysical methods. Among the instruments used in making such discoveries is the torsion balance.

township (United States public-land surveys).*—The unit of survey of the *public lands* of the United States, normally a quadrangle approximately 6 miles on a side with boundaries conforming to *meridians* and *parallels,* located with reference to the *initial point* of a *principal meridian* and *base line.*

Townships are numbered consecutively north and south from a *base line;* thus "township 14 north" indicates a township in the 14th *tier* north of a *base line.* The word township is used in conjunction with the appropriate *range* to indicate the *coordinates* of a particular township with reference to the initial point; thus "township 14 north, range 3 east" indicates the particular township which is the 14th township north of the *base line* and the 3rd township east of the *principal meridian* controlling the surveys in that area. The plural form, townships or tps., is used whenever more than one unit is indicated; "townships 14 north, ranges 3, 4, and 5 east"; "townships 14 and 15 north, range 3 east."

township: fractional (United States public-land surveys).*—A *township* containing appreciably less than 36 normal *sections,* usually due to invasion by a segregated body of water, or by other land which cannot properly be surveyed as part of that *township.*

Townships may be rendered fractional also in closing the public-land surveys on State boundaries, or other limiting lines.

township line (United States public-land surveys).*—An exterior boundary of a *township* extending in an east-west direction.

track level.—See under *track sector.*

track sector.—An instrument composed of a half-circle *protractor* with a single movable arm which carries a *spirit level,* and is so constructed that when the *protractor* base is parallel with a railway rail and the *spirit level* bubble is in the center of its tube, the movable arm will register the slope (angle of inclination) of the rail. Also called track level.

The track sector was designed and used for determining the slope of the tape when used in *traverse* measures along a railway. Not proving satisfactory, its use was abandoned.

Transcontinental Triangulation datum.—See under *datum.*

transit.—1. The apparent passage of a star or other celestial body across a defined line of the celestial *sphere,* as a *meridian, prime vertical,* or *almucantar.* The apparent passage of a star or other celestial body across a line in the *reticle* of a *telescope,* or some *line of sight.* The apparent passage of a smaller celestial body across the disk of a larger celestial body.

2. A surveying instrument composed of a horizontal circle graduated in circular measure and an *alidade* with a *telescope* which can be reversed in its supports without being lifted therefrom. Also, the act of making such reversal.

3. An astronomical instrument having a *telescope* which can be so adjusted in position that the *line of sight* may be made to define a *vertical circle.*

1. When no line is specified, a transit across the *meridian* is usually intended. The transit of a star across the *meridian* occurs at the moment of its *culmination,* and the two terms are sometimes used as having identical meanings; such usage is not correct, even where the instrument is in perfect adjustment. At the poles, a star may have no *culmination,* but it will transit the *meridians.* The transit of a small celestial body across a larger one is exemplified by the transit of the planet Mercury across the disk of the Sun, or of a satellite of Jupiter across the disk of that planet. See *eclipse, occultation.*

2. A *theodolite* having a *telescope* that can be transited in its supports is a transit, and is sometimes termed a transit theodolite.

3. A transit used in astronomical work is usually called either an astronomical transit or a transit instrument.

transit: astronomical.—See under *transit,* 3.

transit: Bamberg.—See under *transit: broken-telescope transit.*

transit: broken-telescope.—An *astronomical transit* having its *telescope* so shaped that a ray of light entering the objective is reflected at right angles by a prism placed within the *telescope,* the reflected ray passing to the eyepiece which is in the *horizontal axis* of the *telescope* and outside its supporting wyes.

One make of broken-telescope transit, called the Bamberg broken-telescope transit or the Bamberg instrument, is described in two Special Publications of the U. S. Coast and Geodetic Survey: No. 35, Determination of the Difference in Longitude Between Each Two of the Stations Washington, Cambridge, and Far Rockaway, by Fremont Morse and O. B. French, 1916; and No. 109, Wireless Longitude, by George D. Cowie and E. A. Eckhardt, 1924. The term "bent transit" has also been applied to this instrument. The design of this instrument permits its use in the determination of *latitude* by the *zenith-telescope method.* It is given considerable weight without producing undesirable flexure. This and the unusually long distance between the supporting wyes make for stability. The *striding level* used with straight-telescope instruments is replaced with a *hanging level.* The position of the eyepiece being independent of the *altitude* of the object observed promotes convenience in observing. This instrument is used in first-order determinations of *astronomic longitude, astronomic latitude,* and recently has been used in first-order determinations of *astronomic azimuth* at stations in high *latitudes.* See *azimuth: micrometer method, determination of astronomic azimuth.*

transit: solar.—See under *solar transit.*

transit-and-stadia survey.—See under *survey.*

transit instrument.—See under *transit, 3.*

transit micrometer.—See under *micrometer.*

transit micrometer, contact correction.—See under *contact correction.*

transit rule: balancing survey.—See under *balancing a survey.*

transit traverse.—See under *traverse.*

transition curve.—See under *curve: spiral curve.*

Transportation Map of the United States.—See under *map.*

transverse Mercator grid.—An informal designation for a *State coordinate system* based on a *transverse Mercator map projection.*

transverse Mercator map projection.—See under *map projection.*

transverse polyconic map projection.—See under *map projection.*

traverse.—See under *traverse: survey traverse.*

traverse: astronomical.—A *survey traverse* in which the *geographic positions* of the stations are obtained from astronomic observations, and lengths and *azimuths* of lines are obtained by computation.

traverse: closed.—A *survey traverse* which starts and ends upon the same station, or upon stations whose *relative positions* have been determined by other *surveys.*

traverse: first-order.—A *survey traverse* which by itself forms a closed loop or which extends between *adjusted positions* of other *first-order control surveys,* and has a *closing error* in *position* not greater than 1 part in 25,000 of its length. The criteria prescribed for attaining the required *accuracy* and *precision* are as follows: The *probable error* of any main-scheme *angle* shall not exceed ±1.5″; the number of *angle* stations between *astronomic-azimuth* stations to be 10 to 15 in number; the *discrepancy* per main-scheme *angle* station between *astronomic azimuths* not to exceed 1.0″; the *probable error* of the *astronomic azimuths* not to exceed ±0.5″.

Formerly known as *precise traverse,* the present designation of first-order traverse was recommended by the Federal Board of Surveys and Maps in May 1925. Lacking the geometrical checks obtainable in *triangulation,* first-order traverse, though formerly used in the fundamental *control survey* of the country, is no longer so used.

traverse: fourth-order.—A *survey traverse* of an *accuracy* less than *third-order traverse.*

In fourth-order traverse, *angles* are observed with a *transit* or *sextant* or are determined graphically, and distances are measured with tape, *stadia,* or wheel.

traverse: geographical exploration.—A route followed across some part of the earth, approximate *positions* along which are determined by *surveying* or navigational methods.

traverse: open.—A *survey traverse* which starts from a station of known or adopted *position,* but does not end upon such a station.

traverse: random.—A *survey traverse* run from one station to another survey station which cannot be seen from the first station, in order to determine their *relative positions.* See *random line.*

traverse: second-order.—A *survey traverse* which by itself forms a closed loop or which extends between adjusted stations of *first-* or *second-order control surveys,* and has a *closing error* in position of between 1 part in 10,000 and 1 part in 25,000 of its length. The following additional criteria are prescribed for the attainment of the required *accuracy* and *precision: probable error* of main-scheme *angles* not to exceed ±3.0″; number of *angle* stations between *astronomic-azimuth* stations, 15 to 25; *discrepancy* per main-scheme *angle* station between *astronomic azimuths,* not to exceed 2.0″; *probable error* of *astronomic azimuths,* not to exceed ±2.0″.

The present designation, second-order traverse, was recommended by the Federal Board of Surveys and Maps in May 1925.

traverse: survey.—A sequence of lengths and *directions* of lines between points on the earth, obtained by or from field measurements, and used in determining *positions* of the points.

A survey traverse may determine the *relative positions* of the points which it connects in series, and if tied to *control stations* on an adopted *datum,* the *positions* may be referred to that *datum.* Survey traverses are classified and identified in a variety of ways: according to methods employed, as, *astronomical traverse;* quality of results, as, *first-order traverse;* purpose served, as, *geographical exploration traverse;* according to form, as, *closed traverse,* etc.

traverse: third-order.—A *survey traverse* which by itself forms a closed loop or which extends between adjusted stations of other *control surveys,* and has a *closing error* in *position* of between 1 part in 5,000 and 1 part in 10,000 of its length. The following additional criteria are prescribed for attaining the required *accuracy* and *precision: probable error* of main-scheme *angles* not to exceed ±6.0″; number of *angle* stations between *astronomic-azimuth* stations, 20 to 35; *discrepancy* per main-scheme *angle* station between *astronomic azimuths,* not to exceed 5.0″; *probable error* of *astronomic azimuths,* not to exceed ±5.0″.

The present designation, third-order traverse, was recommended by the Federal Board of Surveys and Maps in May 1925.

traverse: transit.—A *survey traverse* in which the *angles* are measured with an engineer's *transit* or *theodolite* and the lengths with a metal tape.

A transit traverse is usually executed for the control of local surveys and is of *second-* or *third-order* quality.

traverse net.—See under *survey net.*

traverse tables.—Mathematical tables listing the lengths of the sides opposite the oblique *angles* for each of a series of right-angle plane triangles, as functions of the length and *azimuth* (or *bearing*) of the hypotenuse.

Traverse tables are used in computing *latitudes* and *departures* in surveying and of *courses* in navigation. One argument of such a table is the *angle* which the line or *course* makes with the *meridian* (its *azimuth* or *bearing*), and the other argument is a distance. In tables used in *land surveying,* the distance argument is usually a series of integers, from 1 to 9, with which lines of greater length may be composed. In navigation, the distance argument may run to several hundred miles.

triangulation.—A method of *surveying* in which the stations are points on the ground at the vertices of a chain or network of triangles, whose *angles* are observed instrumentally and whose sides are derived by computation from selected triangle sides called *base lines,* the lengths of which are obtained from direct measurements on the ground.

Triangulation permits the selection of sites for stations and *base lines* favorable for use both from topographic and geometric considerations (see *strength of figure*); it is well adapted to the use of *precision* instruments and methods in all its operations, and susceptible of great *accuracy* and *precision* in its results. Triangulation is generally used where the area surveyed is large and requires the use of *geodetic methods.* Triangulation may be considered as including not only the actual operations of observing *angles* and measuring *base lines,* and their mathematical processing, but also the *reconnaissance* which precedes those operations, and the astronomic observations which are required in the establishment of a *geodetic datum* and in the control of the triangulation.

triangulation: arc.—A system of *triangulation* of limited width designed to progress in a single general direction.

Arc triangulation is executed for the purpose of connecting independent and widely separated *surveys, coordinating,* and *correlating* local surveys along the arc, furnishing data for the determination of a *geodetic datum,* providing a network of control points for a country-wide *survey,* etc. See *area triangulation.*

triangulation: area.—A system of *triangulation* designed to progress in every direction.

Area triangulation is executed to provide survey control points over an area, as of the city or county; or for filling in the areas between *arcs* of *triangulation* which form a network extending over a county or State. See *arc triangulation.*

triangulation: first-order.—*Triangulation* conforming to the following criteria: the average *error of closure* of the main-scheme triangles shall not exceed 1″; the maximum *error of closure* of the main-scheme triangles shall not exceed 3″; the *error of closure* in length on a measured *base line* or on a line of adjusted first-order triangulation shall not exceed 1/25,000 of the length of the line after the *angle* and *side* equations have been satisfied in the *adjustment.*

First-order triangulation was at one time known as *primary triangulation;* this was changed in 1921 to *precise triangulation;* and in May 1925 the present designation was recommended for adoption by the Federal Board of Surveys and Maps. The report of that Board recommended that the four grades of horizontal and vertical control ordinarily used be designated as first order, second order, third order, and fourth order, respectively, the first-named being the most *accurate.*

triangulation: flare.—A method of *triangulation* in which simultaneous observations are made on parachute flares. This method is used for extending triangulation over lines too long to be observed by ordinary methods.

triangulation: precise.—The designation which, from 1921 until the present terminology was adopted in 1925, was applied to what is now called *first-order triangulation.*

triangulation: primary.—The designation which, prior to 1921, was given to what is now called *first-order triangulation;* and thereafter and until the present terminology was adopted in 1925, was applied to what is now called *second-order triangulation.*

triangulation: secondary.—The designation which, prior to 1921, was given to what is now called *second-order triangulation;* and thereafter and until the present terminology was adopted in 1925, was applied to what is now called *third-order triangulation.*

triangulation: second-order.—*Triangulation* which does not attain the quality of *first-order triangulation* but does conform to the following criteria: the average *error of closure* of the main-scheme triangles shall not exceed 3″; the maximum *error of closure* of the main-scheme triangles shall not exceed 5″; the *error of closure* in length on a measured base line or on a line of adjusted first- or second-order triangulation shall not exceed 1/10,000 of the length of the line after the *angle* and *side equations* have been satisfied in the *adjustment.*

Second-order triangulation was at one time known as *secondary triangulation;* later (1921) the name was changed to *primary triangulation;* and in May 1925 the present designation was recommended by the Federal Board of Surveys and Maps.

triangulation: ship-shore.—A method of extending *triangulation* along a coast by making simultaneous observations from three or more shore stations on a target mounted on an anchored ship.

This method is used only when it is impractical to establish a chain of triangles or quadrilaterals.

triangulation: tertiary.—The designation which, prior to 1921, was given to what is now called *third-order triangulation.*

triangulation: third-order.—*Triangulation* which does not attain the quality of *second-order triangulation,* but does conform to the following criteria: the average *error of closure* of the main-scheme triangles shall not exceed 5″; the maximum *error of closure* of the main-scheme triangles shall not exceed 10″; and the closure in length on a measured *base line* or on a line of adjusted *triangulation* shall not exceed 1/5,000 of the length of the line after the *angle* and *side equations* have been satisfied in the *adjustment.*

Third-order triangulation was at one time known as *tertiary triangulation;* later (1921) the name was changed to *secondary triangulation;* and in May 1925 the present designation was recommended by the Federal Board of Surveys and Maps.

triangulation arc.—A system of *triangulation* forming a band or belt on the surface of the earth. Or the corresponding system of *positions,* lines, and *angles* on the surface of the *spheroid.*

The axis of a triangulation arc may *approximate* in position an *arc* on the spheroid, following a *meridian* of *longitude,* a *parallel* of *latitude,* or an oblique *arc;* it may follow a natural feature as a river; or it may follow an artificial feature, as a civil boundary. It is usually given an identifying name corresponding to its general or particular location, or having other geographic significance: as Ninety-eighth Meridian Arc, Mississippi River Arc, etc.

triangulation net.—See under *survey net.*

triangulation signal.—See under *signal.*

triangulation station.—A point on the earth whose *position* is determined by *triangulation.*

triangulation tower.—A *triangulation signal* consisting of two separate structures, independent of one another; an inner structure which supports the *theodolite* and sometimes the target or signal lamps, and an outer structure which supports the observer and his assistants and sometimes the target or signal lamps.

Triangulation towers are used to elevate the *line of sight* above trees, *topographic features,* and other obstacles that might interfere with the observing. Before 1927, towers of wood were used, the inner structure being a tripod and the outer structure a four-leg scaffold. In 1927, the *Bilby steel tower* was put into use. Triangulation towers more than 100 feet in height are not unusual.

triaxial ellipsoid.—See under *ellipsoid.*

tribrach.—The three-arm base of a surveying instrument which carries the foot-screws used in leveling the instrument.

Many surveying instruments have a four-screw base or quadribrach. Such an instrument is not suitable for third- or higher-order horizontal control, for the leveling process may introduce strains in the base of the instrument which will tend to change the instrument in *azimuth* while the observations are being made.

trigonometric leveling.—See under *leveling.*

trilateration.—A method of extending horizontal control where the sides of triangles are measured rather than the angles as in triangulation.

trilinear surveying.—See under *surveying.*

trivet.—A device for use in place of a tripod in mounting a *transit* or *level.* It is essentially a tripod head with three very short legs cast as a single piece of metal.

It is used for placing a *transit* or *level* in a position where a regular tripod could not be conveniently used, or where greater stability is desired. When used for a *theodolite,* a trivet may include the foot-plate with V-shaped grooves cut in suitable arms to receive the feet of the leveling screws.

Tropic of Cancer.—Astronomy. The *parallel of declination* passing through the *summer solstice.* Geography. The *geographic parallel* whose *latitude* corresponds to the *declination* of the *summer solstice.*

The *obliquity of the ecliptic* is steadily changing so that the *summer solstice* is not a point of fixed *declination.* When the Tropic of Cancer is to be shown on a terrestrial *map,* it is desirable that it be treated as a line of fixed *latitude,* and for this purpose it is here proposed that the value 23°27′ north *latitude* be adopted and used.

While, in popular usage, the term "tropics" is sometimes applied to the belt of the earth's surface bounded by the Tropic of Cancer and the *Tropic of Capricorn,* in technical language, the term designates the lines themselves, the included area being known as the "Torrid Zone." "Tropics" also has a climatic connotation.

Tropic of Capricorn.—Astronomy. The *parallel of declination* passing through the *winter solstice.* Geography. The *geographic parallel* whose *latitude* corresponds to the *declination* of the *winter solstice.*

The *obliquity of the ecliptic* is steadily changing so that the *winter solstice* is not a point of fixed *declination.* When the Tropic of Capricorn is to be shown on a terrestrial *map,* it is desirable that it be treated as a line of fixed *latitude,* and for this purpose it is here proposed that the value 23°27′ south *latitude* be adopted and used. See supplementary paragraph under *Tropic of Cancer.*

tropical month.—See under *month.*

tropical year.—See under *year.*

Troughton level.—See under *leveling instrument.*

true.*—A comparative term applied to astronomical values to distinguish them from corresponding magnetic values.

The terms *true bearing,* true meridian, true north, etc., occur frequently in land survey reports, distinguishing those quantities from corresponding magnetic values. In descriptions of land boundaries, the use of true has legal significance and, except in rare instances, refers to values based directly on astronomical observations. True should never be used to indicate *geodetic* values, although in recent work, where the difference between astronomical and *geodetic* values is unimportant, the term has been applied without distinction to both in order to distinguish them from corresponding magnetic values. The use of true to indicate astronomical values should be avoided as much as possible: exact meaning is conveyed by the terms *celestial meridian, geodetic meridian, magnetic meridian,* etc.

true bearing.—See under *bearing.*

true error.—See under *error (general).*

true solar time.—See under *time: apparent solar time.*

turning point (leveling).—See under *point.*

U

Unalaska datum.—See under *datum.*

unit: absolute.—Any unit in a system that is based directly upon associated fundamental units of length, mass, and time.

In selecting a system of absolute units, the choice of the fundamental units is arbitrary. In the metric (c. g. s.) system of absolute units, the fundamental units are the centimeter, the gram, and the second. In the *leveling* work of the U. S. Coast and Geodetic Survey, the metric system is used in defining *dynamic number.*

unit: astronomical.—See under *astronomical unit.*

unit of length.—A nominal distance in space fixed by definition, and independent of temperature, pressure, and other physical conditions. See *meter: international meter; standard of length.*

U. S. Coast and Geodetic Survey leveling rods.—See under *leveling rod: U. S. Coast and Geodetic Survey rods.*

U. S. Engineer leveling rod.—See under *leveling rod: U. S. Engineer precise leveling rod.*

U. S. Geological Survey level.—See under *leveling instrument.*

U. S. Geological Survey leveling rods.—See under *leveling rod: U. S. Geological Survey precise leveling rods.*

United States standard datum.—See under *datum: North American datum.*

universal instrument.—See under *alt-azimuth instrument.*

universal level.—See under *level: circular level.*

universal time.—See under *time.*

upper culmination.—See under *culmination.*

V

Valdez datum.—See under *datum.*

value (of a quantity).—An arithmetical number which defines the size of the quantity.

value: adjusted.—A *value* of a quantity derived from observed data by some orderly process which eliminates *discrepancies* arising from *errors* in those data.

The process, called an *adjustment,* may be made by graphical (mechanical) methods or by analytical (arithmetical) methods. An analytical method involves the determination of *discrepancies* which exist among combinations of observed data; expressing those discrepancies in the form of *condition equations* or *observation equations;* and solving those equations to obtain *corrections* to the observed data. If the solution is by the *method of least squares,* the result will be the *most probable values* that can be derived from the given data. If the data used are sufficiently extensive, the values obtained by the *least-squares method* will be largely free from *accidental errors* of observation. See *position: adjusted.*

value: most probable.—That *value* of a quantity which is mathematically determined from a series of observations and is more nearly free from the effects of *blunders* and *errors* than any other *value* that might be derived from the same series of observations.

An *adjusted value* determined by the *method of least squares* is a most probable value.

value: observed.—A *value* of a quantity that is obtained by instrumental measurement of the quantity.

The term observed value is often applied to the *value* of a quantity derived from instrumental measurement after corrections have been applied for *systematic errors,* but before *accidental errors* have been taken out by some method of adjustment. An *angle* obtained with a *repeating theodolite,* after correction for *closure of horizon* has been applied, is considered an *observed angle.* See *value: most probable value; value: adjusted value.*

value: true.—That *value* of a quantity which is completely free from *blunders* and *errors.*

Since the *errors* to which physical measurements are subject cannot be known exactly, it follows that the true value of a quantity cannot be known with exactness. In *survey* work, the *most probable value* is used as best representing the true value of the quantity.

variation: magnetic.—See under *magnetic variation.*

variation of latitude.—See under *latitude.*

variation of the pole.—A wandering of the poles of figure of the earth with respect to its poles of rotation.

The direction in space of the poles of *rotation* of the earth is nearly invariable; these are the points on the *celestial sphere* reached by the axis about which the earth rotates. The axis of figure of the earth or axis within the earth about which the moment of inertia is a maximum is not coincident with its axis of *rotation*, and extended to the *celestial sphere* describes a path thereon about each pole of *rotation*—a path which is irregular, although an annual periodicity in the motion may be detected, and also one of 14 months. The maximum variation in direction of the axis of figure from a mean position with respect to the axis of *rotation* ranges between 0.1″ and 0.3″. Variation of the pole affects observed determinations of *astronomic latitude, longitude,* and *azimuth,* but account of it is taken in *first-order work* only.

vernal equinox.—See under *equinox.*

vernier.—An auxiliary scale sliding against and used in reading a primary scale. The total length of a given number of *divisions* on a vernier is equal to the total length of one more or one less than the same number of *divisions* on the primary scale.

The vernier makes it possible to read a principal scale (such as a divided circle) much closer than one *division* of that scale. If a space on the vernier is shorter than a space on the primary scale, it is a direct vernier; if a space on the vernier is the longer, it is a retrograde vernier, so-called because its numbers run in the opposite direction from those on the primary scale. The direct vernier is the usual type, and is used in reading the circles on an engineer's *transit,* and on a *repeating theodolite.* Two verniers extending and numbered in opposite directions from the same initial form a double vernier, used in reading a circle having *graduations* numbered in both directions. A single vernier so constructed and numbered that it may be read in either direction is called a folding vernier.

vernier: contact.—The usual type of *vernier,* having the *vernier* scale and the graduated circle in physical contact.

vernier: direct.—See *vernier.*

vernier: double.—See *vernier.*

vernier: folding.—See *vernier.*

vernier: optical.—A microscope with *vernier* lines ruled on a glass slide placed in the focal plane common to the objective and the eyepiece, where it is compared with the image of the graduated circle.

The microscope is so adjusted that the image of the graduated circle falls in the plane of the *vernier* lines, and the *vernier* scale subtends the proper number of *divisions* of the graduated circle. The optical vernier is superior to the usual *vernier* (*contact vernier*) in that there are no *parallactic errors* such as are caused by wear of the contact edges. Optical verniers are, however, more liable to injury and to loss of adjustment than are *contact verniers.*

vernier: retrograde.—See *vernier.*

vertex of curve.—See under *curve: point of intersection.*

vertical.—The *direction* in which the force of gravity acts.

It is the *direction* indicated by a plumb line, and does not usually coincide with the *direction* of the *normal* at a point. The term vertical uses the *geoid* as a reference; *normal* applies to the *spheroid.* See *zenith* and *nadir;* and *deflection of the plumb line.*

vertical angle.—See under *angle.*

vertical angulation.—The process of obtaining differences of *elevation* by means of observed *vertical angles,* combined with lengths of lines.

In *geodetic* work, *trigonometric leveling* is used with the same meaning.

vertical axis (theodolite; transit).—The line through the center of instrument about which the *alidade* rotates.

For an instrument in complete adjustment, this axis occupies a vertical position, passes through the center of the horizontal circle, and is perpendicular to its plane.

vertical circle.—Any *great circle* of the *celestial sphere* passing through the *zenith* (and *nadir*). Also, a divided circle mounted on an instrument in such a manner that the plane of its graduated surface can be placed in a *vertical plane.*

As a *great circle,* a vertical circle is the line of intersection of a *vertical plane* with the *celestial sphere.* As an instrument, a vertical circle may be an auxiliary attachment to a *theodolite* or *transit,* or it may be the major feature of an instrument intended primarily for use in measuring *vertical angles* in astronomical and *geodetic* work. In *geodetic* work, such an instrument is called a vertical circle.

vertical collimator.—See under *collimator.*

vertical comparator (pendulum).—A stand designed for the support of a *pendulum,* a bar of known length, and two *micrometer microscopes,* so placed with reference to one another that the length of the *pendulum* can be measured. Such an apparatus was designed by Repsold.

vertical coordinate.—See under *coordinates.*

vertical plane.—Any plane passing through a point on the earth and containing the *zenith* (and *nadir*) of that point.

Containing the *zenith,* a vertical plane must also contain the *nadir.* The planes of the *celestial meridian* and of the *prime vertical* are vertical planes.

vibration (pendulum).—A single movement of a *pendulum* in either direction, to or fro. See *oscillation (pendulum).*

vibration: amplitude of vibration (pendulum).—See under *amplitude of vibration (pendulum).*

vinculum.—A short horizontal bar placed over the seconds of a numerically expressed *angle* or *direction* to indicate that the seconds are used in connection with a value of minutes 1 less than is recorded. A double vinculum indicates association with a value of minutes 2 less than is recorded.

In recording an *angle* or a *direction* observed with an instrument provided with means of making several independent readings of the circle, as with several *verniers* or *micrometer microscopes,* the degrees, minutes, and seconds are usually recorded for only one *vernier* or microscope, and only the seconds for the other *verniers* or microscopes. Where the seconds as read for the B vernier are associated with minutes less than those recorded for the A vernier, this is indicated in the record by placing a vinculum, or rarely, a double vinculum, over the seconds reading of the B vernier, thus:

	Vernier reading			*Simplified recording*		
	°	′	″	°	′	″
A	90	11	00	90	11	00
B	270	10	50			$\overline{50}$
A	75	15	00	75	15	00
B	255	13	55			$\overline{\overline{55}}$

W

war time.—See under *time.*

Washington Meridian.—See under *meridian.*

water leveling.—See under *leveling.*

weight (surveying).—The relative reliability (or worth) of a quantity as compared with other *values* of the same quantity.

If one *value* of a quantity has a weight of 2, and another *value* of the same quantity has a weight of 1, the first *value* is worth twice the second *value,* and a mean *value* would be obtained by taking a *weighted mean:* twice the first *value* plus once the second *value,* the sum being divided by three.

weighted mean.—A value obtained by multiplying each of a series of *values* by its assigned *weight,* and dividing the sum of those products by the sum of the *weights.*

A weighted mean is sometimes called a weighted arithmetical mean. If the constituent *values* are all of the same *weight,* the *weight* of each *value* may be considered as unity, and an *arithmetical mean* taken. The present practice in *surveying* where both instruments and methods are *standard,* is to make all observations of the same (unit) *weight,* thereby simplifying the computations involved in their reduction and adjustment.

Werner map projection.—See under *map projection.*

westing.—See under *departure (plane surveying).*

Wild theodolite.—See under *theodolite.*

wind (spirit level).—See under *level.*

winter solstice.—See under *solstice.*

witness corner.—A marked point established on firm ground at a measured distance and *direction* from a property corner which may be so situated that it cannot be permanently marked.

A witness corner in *land surveying* corresponds to a *reference mark* in *control surveying.* See *corner: witness corner (United States public-land surveys).*

witness corner (United States public-land surveys).—See under *corner.*

witness mark.—A material mark placed at a known distance and *direction* from a property corner, instrument or other *survey station,* to aid in its recovery and identification.

In *surveying,* a witness mark is established as an aid in the recovery and identification of a *survey station,* or other point to which it is a witness. A mark which is established with such *precision* and *accuracy* that it may be used to restore or take the place of the original station is more properly called a *reference mark* in *control surveys,* and a *witness corner* in *land surveys.*

Woodward base-line measuring apparatus.—See under *base apparatus: iced-bar apparatus.*

working pendulum.—See under *pendulum.*

wye (Y) level.—See under *leveling instrument.*

X

***x* correction.**—The *correction* to an *x direction.*

***x* direction.**—An observed *direction* in a *triangulation* figure for which an approximate value is obtained and treated like an observed *direction* in the adjustment of the figure.

The work of a *least-squares* adjustment of a *triangulation* figure sometimes requires the use of an *x* direction, for which an approximate value is obtained by an *inverse position computation,* by the solution of the *three-point problem,* or by other means, and then using this *x* direction in the adjustment and obtaining a *correction (x correction)* for it, which makes it consistent with the adjusted values of the observed *directions.*

Y

Y (wye) level.—See under *leveling instrument: wye (Y) level.*

Yakutat datum.—See under *datum.*

yard.—A unit of length equal to 36 *inches* exactly.

year.—An interval of *time* based on the *revolution* of the earth in its orbit around the sun with respect to (1) the stars, giving the *sidereal year;* (2) the *vernal equinox,* giving the *tropical year;* (3) *perihelion,* giving the *anomalistic year.* There is also a *calendar year,* a conventional value based on the *tropical year.*

year: anomalistic.—The interval of *time* between two successive passages of *perihelion* by the sun.

The anomalistic year is equal to 365.25964 *mean solar days.*

year: calendar.—A conventional *year* based on the *tropical year.*

The calendar year takes account of the fraction of a *day* contained in the *tropical year,* by assigning an additional *day* to the *year* at stated periods. The calendar years of interest in this country are defined by the *Julian Calendar,* and the *Gregorian Calendar.*

year: sidereal.—The interval of *time* occupied by the sun in completing an apparent circuit of the heavens, from a given star back to the same star.

The sidereal year is equal to 365.25636 *mean solar days.*

year: tropical.—The interval of *time* between two successive passages of the *vernal equinox* by the sun.

The tropical year is the *year* of the seasons, and the basis of the conventional *calendar year* used in chronology and civil reckoning. It is equal to 365.24220 *mean solar days.*

Yukon datum.—See under *datum.*

Z

zenith.—The point where the *direction* of the plumb line produced above the *horizon* meets the *celestial sphere.*

The zenith and *nadir* are poles of the *horizon.* The plumb line is perpendicular to the surface of the *geoid,* but not (except in rare cases) to the surface of the *spheroid.* The *angle* between the plumb line (the *vertical*) and the perpendicular to the surface of the *spheroid* (the *normal*) is the *deflection of the vertical,* also called the *deflection of the plumb line.* The terms, *geodetic zenith* (or *nadir*) and *geocentric zenith* (or *nadir*) are sometimes used, and with meanings different from the above definition: their use is not recommended.

zenith: geocentric.—The point where a line from the center of the earth through a point on its surface meets the *celestial sphere.*

This term is sometimes used in astronomic work, but seldom appears in *geodetic* work. It should be used only in its entirety, the single word, *zenith,* being reserved to designate the point determined by the *direction* of the plumb line.

zenith: geodetic.—The point where the *normal* (to the *spheroid*) extended upward, meets the *celestial sphere.*

This term has some use in *geodetic* work, but should be used only in its entirety. The single word, *zenith,* is reserved to designate the point determined by the *direction* of the plumb line.

zenith camera.—An instrument for the determination of an astronomic position by taking a photograph of the sky immediately surrounding the *zenith.*

The instrument consists of a camera designed to photograph a small section of the sky. The optical axis may be set very precisely in the vertical so that the photograph is symmetrical with respect to the observer's zenith. The time of each observation is recorded on the photograph, thus furnishing the information required to compute the astronomic latitude and longitude.

zenith distance.—The *vertical angle* between the *zenith* and the object which is observed or defined. See *zenith distance: double.*

Zenith distance is the complement of the *altitude.*

zenith distance: double.—A value of twice the *zenith distance* of an object, obtained by observation and not by mathematical process.

In *trigonometric leveling* and in astronomical work, double zenith distances are observed because they are nearly free from effects of inclination of the *vertical axis* of the instrument used.

zenith sector.—See under *sector.*

zenith telescope.—A portable instrument adapted for the measurement of small differences of *zenith distance,* and used in the determination of *astronomic latitude.*

This instrument consists of a *telescope* equipped with an *ocular micrometer* and a *spirit level,* and so mounted on a *vertical axis* that it may be placed in the plane of the *meridian* for observation on a star *culminating* north (or south) of the *zenith,* and then rotated 180° in *azimuth* and a second star observed as it *culminates* south (or north) of the *zenith.* The difference of the *zenith distances* of the two stars is measured with the *micrometer;* the *spirit level* is used to determine any change that may occur in the direction of the axis of *rotation* of the *telescope* between the two observations. In its present form, this instrument is essentially the invention (in 1834) of Captain Andrew Talcott of the U. S. Engineers. It was adopted by the U. S. Coast and Geodetic Survey for *latitude* determinations and given some improvements. See *zenith-telescope method, latitude determinations.*

zenith-telescope method, latitude determination.—A *precise* method of determining *astronomic latitude* by measuring the difference of the meridional *zenith distances* of two stars of known *declination,* one north and the other south of the *zenith.*

The observations are made with a *zenith telescope,* or with an *astronomical transit* which can be converted to serve as a *zenith telescope,* such as the *meridian telescope* and the *broken-telescope transit.* The two stars have approximately the same meridional *zenith distances,* and their *culminations* occur within a few minutes of the same time. The *astronomic latitude* of the point of observation will be one-half the sum of the *declinations* of the two stars, plus or minus one-half the difference of their *zenith distances.* This method is also known as the Horrebow-Talcott method, latitude determination, and the Talcott method, latitude determination. While Peter Horrebow (not Harrebow) published an account of his method in 1732, it was buried in obscurity, and there is good reason to believe that Captain Andrew Talcott had no knowledge of Horrebow's work when he announced his own discovery of the method in 1834.

U. S. GOVERNMENT PRINTING OFFICE : 1974 O - 559-993

IMST-00005709-001 - Copyright 2007 photolibrary.com

Surveying

Rai University

EVOLVING THINKING MINDS

Chapter 1

CHAIN, COMPASS AND PLANE TABLE SURVEYING

Structure of this unit

CHAIN, COMPASS, PLANE TABLE SURVEYING

Learning Objectives

1. CHAIN
2. field and office work
3. Ranging and Chaining
4. Reciprocal ranging
5. Well-conditioned triangles.
6. COMPASS
7. Prismatic compass
8. Surveyor’s compass
9. Bearing systems and conversions
10. Local attraction – Magnetic declination dip
11. Traversing – Plotting – Adjustment of error.
12. PLANE TABLE SURVEYING : Plane table instruments and accessories – merits and demerits – methods – Radiation- Intersection – Resection – Traversing.

1.1 Surveying

Surveying techniques have existed throughout much of recorded history. Under the Romans, land surveyors were established as a profession, and they established the basic measurements under which the Roman Empire was divided, such as a tax register of conquered lands (300 AD). In the 18th century in Europe triangulation was used to build a hierarchy of networks to allow point positioning within a country. Highest in the hierarchy were triangulation networks. These were densified into networks of traverses (polygons), into which local mapping surveying measurements, usually with measuring tape, corner prism and the familiar red and white poles, are tied.

Surveying or **land surveying** is the technique, profession, and science of accurately determining the terrestrial or three-dimensional position of points and the distances and angles between them, commonly practiced by licensed **surveyors**, and members of various building professions. These points are usually on the surface of the Earth, and they are often used to establish land maps and boundaries for ownership, locations (building corners, surface location of subsurface features) or other governmentally required or civil law purposes (property sales).

To accomplish their objective, surveyors use elements of mathematics (geometry and trigonometry), physics, engineering and law.

An alternative definition, from the American Congress on Surveying and Mapping (ACSM), is the science and art of making all essential measurements to determine the relative position of points or physical and cultural details above, on, or beneath the surface of the Earth, and to depict them in a usable form, or to establish the position of points or details.

Furthermore, as alluded to above, a particular type of surveying known as "land surveying" (also per ACSM) is the detailed study or inspection, as by gathering information through observations, measurements in the field, questionnaires, or research of legal instruments, and data analysis in the support of planning, designing, and establishing of property boundaries. It involves the re-establishment ofcadastral surveys and land boundaries based on documents of record and historical evidence, as well as certifying surveys (as required by statute or local ordinance) of subdivision plats or maps, registered land surveys, judicial surveys, and space delineation. Land surveying can include associated services such as mapping and related data accumulation, construction layout surveys, precision measurements of length, angle, elevation, area, and volume, as well as horizontal and vertical control surveys, and the analysis and utilization of land survey data.

Surveyors use various tools to do their work successfully and accurately, such as total stations, robotic total stations, GPS receivers, prisms, 3D scanners, radio communicators, handheld tablets, digital levels, and surveying software.

Surveying has been an essential element in the development of the human environment since the beginning of recorded history (about 6,000 years ago). It is required in the planning and execution of nearly every form of construction. Its most familiar modern uses are in the fields of transport, building and construction, communications, mapping, and the definition of legal boundaries for land ownership.

History of surveying

Surveying techniques have existed throughout much of recorded history. In ancient Egypt, when the Nile River overflowed its banks and washed outfarm boundaries, boundaries were re-established by a rope stretcher, or surveyor, through the application of simple geometry. The nearly perfect squareness and north-south orientation of the Great Pyramid of Giza, built c. 2700 BC, affirm the Egyptians' command of surveying.

A brief history of surveying:

- The Egyptian land register (3000 BC).
- A recent reassessment of Stonehenge (c. 2500 BC) suggests that the monument was set out by prehistoric surveyors using peg and rope geometry.
- The Groma surveying instrument originated in Mesopotamia (early 1st millennium BC).
  - Under the Romans, land surveyors were established as a profession, and they established the basic measurements under which the Roman Empire was divided, such as a tax register of conquered lands (300 AD).
  - The rise of the Caliphate led to extensive surveying throughout the Arab Empire. Arabic surveyors invented a variety of specialized instruments for surveying, including
  - Instruments for accurate leveling: A wooden board with a plumb line and two hooks, an equilateral triangle with a plumb line and two hooks, and a reed level.
  - A rotating alhidade, used for accurate alignment.
  - A surveying astrolabe, used for alignment, measuring angles, triangulation, finding the width of a river, and the distance between two points separated by an impassable obstruction.
- In England, The Domesday Book by William the Conqueror (1086)
  - Covered all England
  - Contained names of the land owners, area, land quality, and specific information of the area's content and inhabitants.
  - Did not include maps showing exact locations.

In the 18th century in Europe triangulation was used to build a hierarchy of networks to allow point positioning within a country. Highest in the hierarchy were triangulation networks. These were densified into networks of traverses (polygons), into which local mapping surveying measurements, usually with measuring tape, corner prism and the familiar red and white poles, are tied. For example, in the late 1780s, a team from the Ordnance Survey of Great Britain, originally under General William Roy began the Principal Triangulation of Britain using the specially built Ramsden theodolite. Large scale surveys are known as geodetic surveys.

- Continental Europe's cadastre was created in 1808
  - Founded by Napoleon I (Bonaparte)
  - Contained numbers of the parcels of land (or just land), land usage, names etc., and value of the land
  - 100 million parcels of land, triangle survey, measurable survey, map scale: 1:2500 and 1:1250
  - spread fast around Europe, but faced problems especially in Mediterranean countries, Balkan, and Eastern Europe due to cadastre upkeep costs and troubles.

A cadastre loses its value if register and maps are not constantly updated. Because of the fundamental value of land and real estate to the local and global economy, land surveying was one of the first professions to require Professional Licensure. In many jurisdictions, the land surveyors license was the first Professional Licensure issued by the state, province, or federal government.

1.1.1 Surveying techniques

Historically, distances were measured using a variety of means, such as with chains having links of a known length, for instance a Gunter's chain, or measuring tapes made of steel or invar. To measure horizontal distances, these chains or tapes were pulled taut according to temperature, to reduce sagging and slack. Additionally, attempts to hold the measuring instrument level would be made. In instances of measuring up a slope, the surveyor might have to "break" (break chain) the measurement- use an increment less than the total length of the chain.

Historically, horizontal angles were measured using a compass, which would provide a magnetic bearing, from which deflections could be measured. This type of instrument was later improved, with more carefully scribed discs providing better angular resolution, as well as through mounting telescopes with reticles for more-precise sighting atop the disc (see theodolite). Additionally, levels and calibrated circles allowing measurement of vertical angles were added, along with verniers for measurement to a fraction of a degree—such as with a turn-of-the-century transit.

The simplest method for measuring height is with an altimeter – basically a barometer – using air pressure as an indication of height. But surveying requires greater precision. A variety of means, such as precise levels (also known as differential leveling), have been developed to do this. With precise leveling, a series of measurements between two points are taken using an instrument and a measuring rod. Differentials in height between the measurements are added and subtracted in a series to derive the net difference in elevation between the two endpoints of the series. With the advent of the Global Positioning System (GPS), elevation can also be derived with sophisticated satellite receivers, but usually with somewhat less accuracy than with traditional precise leveling. However, the accuracies may be similar if the traditional leveling would have to be run over a long distance.

Triangulation is another method of horizontal location made almost obsolete by GPS. With the triangulation method, distances, elevations and directions between objects at great distance from one another can be determined. Since the early days of surveying, this was the primary method of determining accurate positions of objects for topographic maps of large areas. A surveyor first needs to know the horizontal distance between two of the objects. Then the height, distances and angular position of other objects can be derived, as long as they are visible from one of the original objects. High-accuracy transits or theodolites were used for this work, and angles between objects were measured repeatedly for increased accuracy. See alsoTriangulation in three dimensions.

Turning is a term used when referring to moving the level to take an elevation shot in a different location. When land surveying, there may be trees or other obstructions blocking the view from the level gun to the level rod. In order to "turn" the level gun, one must first take a shot on the rod from the current location and record the elevation. Keeping the level rod in exactly the same location and elevation, one may move the level gun to a different location where the level rod is still visible. Record the new elevation seen from the new location of the level rod and use the difference in elevations to find the new elevation of the level gun. Turning is not only used when there are obstructions in the way, but also when drastically changing elevations. You can turn up or down in elevation but the gun must always be at a higher elevation than the base of the rod. A level rod can usually be raised up to 25 feet high, which enables the gun to be set much higher. However, if the gun is lower than the base of the rod, you will not be able to take a shot because the rod cannot be lowered beyond the ground elevation.

Surveying equipment

As late as the 1990s, the basic tools used in planar surveying were a tape measure for determining shorter distances, a level to determine height or elevation differences, and a theodolite, set on a tripod, to measure angles (horizontal and vertical), combined with the process of triangulation. Starting from a position with known location and elevation, the distance and angles to the unknown point are measured.

A more modern instrument is a total station, which is a theodolite with an electronic distance measurement device (EDM). A total station can also be used for leveling when set to the horizontal plane. Since their introduction, total stations have made the technological shift from being optical-mechanical devices to being fully electronic.

Modern top-of-the-line total stations no longer require a reflector or prism (used to return the light pulses used for distancing) to return distance measurements, are fully robotic, and can even e-mail point data to the office computer and connect to satellite positioning systems, such as a Global Positioning System. Though Real Time Kinematic GPS systems have increased the speed and precision of surveying, they are still horizontally accurate to only about 20 mm and vertically accurate to about 30–40 mm.

Total stations are still used widely, along with other types of surveying instruments, however, because GPS systems do not work well in areas with dense tree cover or constructions. One-person robotic-guided total stations allow surveyors to gather precise measurements without extra workers to look through and turn the telescope or record data. A faster but expensive way to measure large areas (not details, and no obstacles) is with a helicopter, equipped with a laser scanner, combined with a GPS to determine the position and elevation of the helicopter. To increase precision, surveyors place beacons on the ground (about 20 km (12 mi) apart). This method reaches precisions between 5–40 cm (depending on flight height).

1.1.2 Types of surveys and applicability

ALTA/ACSM Land Title Survey: a surveying standard jointly proposed by the American Land Title Association and the American Congress on Surveying and Mapping that incorporates elements of the boundary survey, mortgage survey, and topographic survey.

- Archaeological survey: used to accurately assess the relationship of archaeological sites in a landscape or to accurately record finds on an archaeological site.
- As-built survey: a survey carried out during or immediately after a construction project for record, completion evaluation and payment purposes. An as-built survey also known as a 'works as executed survey' documents the location of the recently constructed elements that are subject to completion evaluation. As built surveys are typically presented in red or redline and overlayed over existing design plans for direct comparison with design information.
- Bathymetric survey: a survey carried out to map the topography and features of the bed of an ocean, lake, river or other body of water.
- Boundary survey: a survey that establishes boundaries of a parcel using its legal description, which typically involves the setting or restoration of monuments or markers at the corners or along the lines of the parcel, often in the form of iron rods, pipes, or concrete monuments in the ground, or nails set in concrete or asphalt.
- Deformation survey: a survey to determine if a structure or object is changing shape or moving. The three-dimensional positions of specific points on an object are determined, a period of time is allowed to pass, these positions are then re-measured and calculated, and a comparison between the two sets of positions is made.
- Engineering surveys: those surveys associated with the engineering design (topographic, layout and as-built) often requiring geodetic computations beyond normal civil engineering practice.
- Foundation survey: a survey done to collect the positional data on a foundation that has been poured and is cured. This is done to ensure that the foundation was constructed in the location, and at the elevation, authorized in the plot plan, site plan, or subdivision plan.
- Geological survey: generic term for a survey conducted for the purpose of recording the geologically significant features of the area under investigation. .
- Hydrographic survey: a survey conducted with the purpose of mapping the coastline and seabed for navigation, engineering, or resource management purposes.
- Measured survey : a building survey to produce plans of the building. such a survey may be conducted before renovation works, for commercial purpose, or at end of the construction process "as built survey"
- Mortgage survey or physical survey: a simple survey that delineates land boundaries and building locations. In many places a mortgage survey is required by lending institutions as a precondition for a mortgage loan.
- Soil survey, or soil mapping, is the process of determining the soil types or other properties of the soil cover over a landscape, and mapping them for others to understand and use.

- Structural survey: a detailed inspection to report upon the physical condition and structural stability of a building or other structure and to highlight any work needed to maintain it in good repair.
- Tape survey: this type of survey is the most basic and inexpensive type of land survey. Popular in the middle part of the 20th century, tape surveys while being accurate for distance lack substantially in their accuracy of measuring angle and bearing standards that are practiced by professional land surveyors.
- Topographic survey: a survey that measures the elevation of points on a particular piece of land, and presents them as contour lines on a plot.

Surveying as a career

The basic principles of surveying have changed little over the ages, but the tools used by surveyors have evolved tremendously. Engineering, especiallycivil engineering, depends heavily on surveyors.

Whenever there are roads, railways, reservoir, dams, pipeline transports retaining walls, bridges or residential areas to be built, surveyors are involved. They establish the boundaries of legal descriptions and the boundaries of various lines of political divisions. They also provide advice and data forgeographical information systems (GIS), computer databases that contain data on land features and boundaries.

Surveyors must have a thorough knowledge of algebra, basic calculus, geometry, and trigonometry. They must also know the laws that deal with surveys, property, and contracts.

In addition, they must be able to use delicate instruments with accuracy and precision. In the United States, surveyors and civil engineers use units of feet wherein a survey foot is broken down into 10ths and 100ths. Many deed descriptions requiring distance calls are often expressed using these units (125.25 ft). On the subject of accuracy, surveyors are often held to a standard of one one-hundredth of a foot; about 1/8 inch. Calculation and mapping tolerances are much smaller wherein achieving near-perfect closures are desired. Though tolerances such as this will vary from project to project, in the field and day to day usage beyond a 100th of a foot is often impractical.

Licensing

In most of the United States, surveying is recognized as a distinct profession apart from engineering. Licensing requirements vary by state, but they generally have components of education, experience and examinations. In the past, experience gained through an apprenticeship, together with passing a series of state-administered examinations, was required to attain licensure. Now, most states insist upon basic qualification of a degree in surveying, plus experience and examination requirements.

The licensing process typically follows two phases. First, upon graduation, the candidate may be eligible to take the Fundamentals of Surveying (FS) exam, to be certified upon passing and meeting all other requirements as a surveying intern (SI),(formerly surveyor in training (SIT)). Upon being certified as an SI, the candidate then needs to gain additional experience to become eligible for the second phase. That typically consists of the Principles and Practice of Land Surveying (PS) exam along with a state-specific examination.

Licensed surveyors usually denote themselves with the letters P.L.S. (professional land surveyor), P.S. (professional surveyor), L.S. (land surveyor), R.L.S. (registered land surveyor), R.P.L.S. (Registered Professional Land Surveyor), or P.S.M. (professional surveyor and mapper) following their names, depending upon the dictates of their particular jurisdiction of registration.

In Canada, land Surveyors are registered to work in their respective province. The designation for a land surveyor breaks down by province, but follows the rule whereby the first letter indicates the province, followed by L.S. There is also a designation as a C.L.S. or Canada lands surveyor, who has the authority to work on Canada Lands, which include Indian Reserves, National Parks, the three territories and offshore lands.

In many Commonwealth countries, the term Chartered Land Surveyor is used for someone holding a professional license to conduct surveys.

A licensed land surveyor is typically required to sign and seal all plans, the format of which is dictated by their state jurisdiction, which shows their name and registration number. In many states, when setting boundary corners land surveyors are also required to place survey monuments bearing their registration numbers, typically in the form of capped iron rods, concrete monuments, or nails with washers.

Building surveying

Building surveying **emerged in the 1970s as a profession in the United Kingdom by a group of technically minded general practice surveyors.[7] Building surveying is a recognised profession in Britain, Ireland, Australia and Hong Kong. In Australia in particular, due to risk mitigation and limitation factors, the employment of surveyors at all levels of the construction industry is widespread. There are still many countries where it is not widely recognized as a profession.**

Building Surveyors are trained to some extent in all aspects of property but with specific training in Building Pathology, as such they have a wide understanding of the end implications of decisions taken by more specific professions and trades during the realisation process, thus making them suitable for employment as Project and Property Managers on the client side (i.e. managing external contractors).

Services that building surveyors undertake are broad but can include:

- Construction design and building works
- Project management and monitoring
- Property Legislation advice
- Insurance assessment and claims assistance
- Defect investigation and maintenance advice
- Building surveys and measured surveys
- Handling planning applications
- Building inspection to ensure compliance with building regulations
- Pre-acquisition surveys
- Negotiating dilapidations claims[8]

Building surveyors also advise on many aspects of construction including:

- design
- cost
- maintenance
- sustainability
- repair
- refurbishment
- restoration and preservation of buildings and monuments.

Clients of a building surveyor can be the government agencies, businesses and individuals. Surveyors work closely with architects, planners, quantity surveyors, engineers, homeowners and tenants groups. A building surveyor may be called to act as an expert witness. It is usual for building surveyors to earn a university degree before undertaking structured training to become a member of a professional organisation.

With the enlargement of the European community, the profession of the building surveyor is becoming more widely known in other European states, particularly France, where many English-speaking people buy second homes.

Lidar Surveying – Three-dimensional laser scanning provides high definition surveying for architectural, as-built, and engineering surveys. Recent technological advances make it the most cost-effective and time-sensitive solution for providing the highest level of detail available for interior and exterior building work.

Land surveyor

One of the primary roles of the land surveyor is to determine the boundary of real property on the ground. That boundary has already been established and described in legal documents and official plans and maps prepared by attorneys, engineers, and other land surveyors. The corners of the property will either have been monumented by a prior surveyor, or monumented by the surveyor hired to perform a survey of a new boundary which has been agreed upon by adjoining land owners.

Monuments are categorized into two groups which are known as natural and artificial. Natural monuments are things such as trees, large stones and other substantial, naturally occurring objects that were in place before the survey was made. An artificial monument is anything within the regulations that are usually placed at corner points by landowners, surveyors, engineers and others. They may be referred to as iron pins or pipes, stakes, trees, concrete monuments or whatever the surveyor decides to use at the time, within the regulations for the area. The courts have held that natural monuments control over artificial monuments because they are more certain in identification and less likely to be disturbed.

Over time, construction and maintenance of roads and many other acts of man, along with acts of nature such as earthquakes, movement of water, and tectonic shift can obliterate or damage the monumented locations of land boundaries. The land surveyor is often compelled to consider other evidence such as fence locations, wood lines, monuments on neighboring properties and recollections of people. This other evidence is known as Extrinsic Evidence and is a fairly common principle. Extrinsic evidence is defined as evidence outside the writings, in this case the deed. Extrinsic evidence is held to be synonymous with evidence from another source.

Today's land surveyor sets monumentation at actual physical points on the ground that define angle points of boundary lines that divide neighboring parcels. These monuments are most often 1/2" or 5/8" iron rebar rods or pipes placed at 18" minimum depth, but varies state by state. The more recent rods or pipes may have an affixed plastic cap over the top bearing the responsible surveyors' name and license number. Older monuments may exist such as old pipes, gun barrels, axles, mounds of stone, whiskey bottles, or even wooden stakes. In addition to rods and pipes, surveyors might use 4x4" concrete posts at corners of large parcels or anywhere that would require more stability (e.g. beach sand). They place them three feet deep. In places where there is asphalt or concrete, it is common to place nails or aluminum alloy caps to re-establish boundary corners. Marks are meant to be durable, stable, and as "permanent" as possible. The aim is to provide sufficient marks so some marks will remain for future re-establishment of boundaries. The material and marking used on monuments placed to mark boundary corners are often subject to state laws. Many states have laws that protect existing monuments and can have civil penalties if disturbed or destroyed.

Cadastral land surveyors are licensed by governments. In the United States, cadastral surveys are typically conducted by the federal government, specifically through the Cadastral Surveys branch of the Bureau of Land Management (BLM), formerly the General Land Office(GLO). They consult with USFS, Park Service, Corps of Engineers, BIA, Fish and Wildlife Service, Bureau of Reclamation, etc. In states that have been organized per the Public Land Survey System (PLSS), surveyors carry out BLM Cadastral Surveys in accordance with that system.

A common use of a survey is to determine a legal property boundary. The first stage in such a survey, known as a resurvey, is to obtain copies of the deed description and all other available documents from the owner. The deed description is that of the deed and not a tax statement or other incomplete document. The surveyor should then obtain copies of deed descriptions and maps of the adjoining properties, any records from the municipality or county, utility maps and any records of surveys. Depending on which region the survey is located in some or most of this information may not be available or even exist. Whether the information exists or not a thorough search should be conducted so that no records are neglected. Copies of deeds usually can be located in the county recorder's office and maps or plats can usually be found at the county recorder or surveyor's office. These arrangements will vary state to state and survey system to survey system so some familiarity maybe needed. When all the records are assembled, the surveyor examines the documents for errors, such as closure errors. When a metes and bounds description is involved, the seniority of the deeds must be determined. The title abstract usually gives the order of seniority for the deeds related to the tract being surveyed and should be used if available. After this data is gathered and analyzed the field survey may commence. The initial survey operations should be concentrated on locating monuments. In urban regions or a city, monuments should be sought initially but in the absence of monuments property corners marked by iron pins, metal survey markers, iron pipes and other features that may establish a line of possession should be located. When the approximate positions for the boundaries of the property have been located a traverse is run around the property. While the control traverse is being run, ties should be measured and all details relevant to the boundaries should be acquired. This includes but is not limited to locating the property corners, monuments, fences, hedge rows, walls, walks and all buildings on the lot. The Surveyor then takes this data collected and compares it to the records that were received. When a solution is reached the property corners that are chosen as those that best fit all the data are coordinated and ties by direction and distance are computed from the nearest traverse point. Once this has been established the features on the lot can be drawn, dimensions can be shown from these features to the boundary line and a map or plat is prepared for the client.

The art of surveying

Many properties have considerable problems with regards to improper bounding, miscalculations in past surveys, titles, easements, and wildlife crossings. Also many properties are created from multiple divisions of a larger piece over the course of years, and with every additional division the risk of miscalculation increases. The result can be abutting properties not coinciding with adjacent parcels, resulting in hiatuses (gaps) and overlaps. Many times a surveyor must solve a puzzle using pieces that do not exactly fit together. In these cases, the solution is based upon the surveyor's research and interpretation, along with established procedures for resolving discrepancies. This essentially is a process of continual error correction and update, where official recordation documents countermand the previous and sometime erroneous survey documents recorded by older monuments and older survey methods.

1.2 Establishing the framework

Most surveying frameworks are erected by measuring the angles and the lengths of the sides of a chain of triangles connecting the points fixed by global positioning. The locations of ground features are then determined in relation to these triangles by less accurate and therefore cheaper methods. Establishing the framework ensures that detail surveys conducted at different times or by different surveyors fit together without overlaps or gaps.

For centuries the corners of these triangles have been located on hilltops, each visible from at least two others, at which the angles between the lines joining them are measured; this process is called triangulation. The lengths of one or two of these lines, called bases, are measured with great care; all the other lengths are derived by trigonometric calculations from them and the angles. Rapid checks on the accuracy are provided by measuring all three angles of each triangle, which must add up to 180 degrees.

In small flat areas, working at large scales, it may be easier to measure the lengths of all the sides, using a tape or a chain, rather than the angles between them; this procedure, called trilateration, was impractical over large or hilly areas until the invention of electromagnetic distance measurement (EDM) in the mid-20th century. This procedure has made it possible to measure distances as accurately and easily as angles, by electronically timing the passage of radiation over the distance to be measured; microwaves, which penetrate atmospheric haze, are used for long distances and light or infrared radiation for short ones. In the devices used for EDM, the radiation is either light (generated by a laser or an electric lamp) or an ultrahigh-frequency radio beam. The light beam requires a clear line of sight; the radio beam can penetrate fog, haze, heavy rain, dust, sandstorms, and some foliage. Both types have a transmitter-receiver at one survey station. At the remote station the light type contains a set of corner mirrors; the high-frequency type incorporates a retransmitter (requiring an operator) identical to the transmitter-receiver at the original station. A corner mirror has the shape of the inside of a corner of a cube; it returns light toward the source from whatever angle it is received, within reasonable limits. A retransmitter must be aimed at the transmitter-receiver.

In both types of instrument, the distance is determined by the length of time it takes the radio or light beam to travel to the target and back. The elapsed time is determined by the shift in phase of a modulating signal superimposed on the carrier beam. Electronic circuitry detects this phase shift and converts it to units of time; the use of more than one modulating frequency eliminates ambiguities that could arise if only a single frequency had been employed.

EDM has greatly simplified an alternative technique, called traversing, for establishing a framework. In traversing, the surveyor measures a succession of distances and the angles between them, usually along a traveled route or a stream. Before EDM was available, traversing was used only in flat or forested areas where triangulation was impossible. Measuring all the distances by tape or chain was tedious and slow, particularly if great accuracy was required, and no check was obtainable until the traverse closed, either on itself or between two points already fixed by triangulation or by astronomical observations.

surveying, a means of making relatively large-scale, accurate measurements of the Earth's surfaces. It includes the determination of the measurement data, the reduction and interpretation of the data to usable form, and, conversely, the establishment of relative position and size according to given measurement requirements. Thus, surveying has two similar but opposite functions: (1) the determination of existing relative horizontal and vertical position, such as that used for the process of mapping, and (2) the establishment of marks to control construction or to indicate land boundaries.

Surveying has been an essential element in the development of the human environment for so many centuries that its importance is often forgotten. It is an imperative requirement in the planning and execution of nearly every form of construction. Surveying was essential at the dawn of history, and some of the most significant scientific discoveries could never have been implemented were it not for the contribution of surveying. Its principal modern uses are in the fields of transportation, building, apportionment of land, and communications.

Except for minor details of technique and the use of one or two minor hand-held instruments, surveying is much the same throughout the world. The methods are a reflection of the instruments, manufactured chiefly in Switzerland, Austria, Great Britain, the United States, Japan, and Germany. Instruments made in Japan are similar to those made in the West.

History

It is quite probable that surveying had its origin in ancient Egypt. The Great Pyramid of Khufu at Gizawas built about 2700 BCE, 755 feet (230 metres) long and 481 feet (147 metres) high. Its nearly perfect squareness and north–south orientation affirm the ancient Egyptians' command of surveying.

Evidence of some form of boundary surveying as early as 1400 BCE has been found in the fertile valleys and plains of the Tigris, Euphrates, and Nile rivers. Clay tablets of the Sumerians show records of land measurement and plans of cities and nearby agricultural areas. Boundary stones marking land plots have been preserved. There is a representation of land measurement on the wall of a tomb at Thebes(1400 BCE) showing head and rear chainmen measuring a grainfield with what appears to be a rope with knots or marks at uniform intervals. Other persons are shown. Two are of high estate, according to their clothing, probably a land overseer and an inspector of boundary stones.

There is some evidence that, in addition to a marked cord, wooden rods were used by the Egyptians for distance measurement. There is no record of any angle-measuring instruments of that time, but there was a level consisting of a vertical wooden A-frame with a plumb bob supported at the peak of the A so that its cord hung past an indicator, or index, on the horizontal bar. The index could be properly placed by standing the device on two supports at approximately the same elevation, marking the position of the cord, reversing the A, and making a similar mark. Halfway between the two marks would be the correct place for the index. Thus, with their simple devices, the ancient Egyptians were able to measure land areas, replace property corners lost when the Nile covered the markers with silt during floods, and build the huge pyramids to exact dimensions.

The Greeks used a form of log line for recording the distances run from point to point along the coast while making their slow voyages from the Indus to the Persian Gulf about 325 BCE. The magneticcompass was brought to the West by Arab traders in the 12th century CE. The astrolabe was introduced by the Greeks in the 2nd century BCE. An instrument for measuring the altitudes of stars, or their angle of elevation above the horizon, took the form of a graduated arc suspended from a hand-held cord. A pivoted pointer that moved over the graduations was pointed at the star. The instrument was not used for nautical surveying for several centuries, remaining a scientific aid only.

The Greeks also possibly originated the use of the groma, a device used to establish right angles, butRoman surveyors made it a standard tool. It was made of a horizontal wooden cross pivoted at the middle and supported from above. From the end of each of the four arms hung a plumb bob. By sighting along each pair of plumb bob cords in turn, the right angle could be established. The device could be adjusted to a precise right angle by observing the same angle after turning the device approximately 90°. By shifting one of the cords to take up half the error, a perfect right angle would result.

About 15 BCE the Roman architect and engineer Vitruvius mounted a large wheel of known circumference in a small frame, in much the same fashion as the wheel is mounted on a wheelbarrow; when it was pushed along the ground by hand it automatically dropped a pebble into a container at each revolution, giving a measure of the distance traveled. It was, in effect, the first odometer.

The water level consisted of either a trough or a tube turned upward at the ends and filled with water. At each end there was a sight made of crossed horizontal and vertical slits. When these were lined up just above the water level, the sights determined a level line accurate enough to establish the grades of the Roman aqueducts. In laying out their great road system, the Romans are said to have used theplane table. It consists of a drawing board mounted on a tripod or other stable support and of a straightedge—usually with sights for accurate aim (the alidade) to the objects to be mapped—along which lines are drawn. It was the first device capable of recording or establishing angles. Later adaptations of the plane table had magnetic compasses attached.

Plane tables were in use in Europe in the 16th century, and the principle of graphic triangulation and intersection was practiced by surveyors. In 1615 Willebrord Snell, a Dutch mathematician, measured an arc of meridian by instrumental triangulation. In 1620 the English mathematician Edmund Gunterdeveloped a surveying chain, which was superseded only by the steel tape beginning in the late 19th century.

The study of astronomy resulted in the development of angle-reading devices that were based on arcs of large radii, making such instruments too large for field use. With the publication of logarithmic tables in 1620, portable angle-measuring instruments came into use. They were called topographic instruments, or theodolites. They included pivoted arms for sighting and could be used for measuring both horizontal and vertical angles. Magnetic compasses may have been included on some.

The vernier, an auxiliary scale permitting more accurate readings (1631), the micrometer microscope (1638), telescopic sights (1669), and spirit levels (about 1700) were all incorporated in theodolites by about 1720. Stadia hairs were first applied by James Watt in 1771. The development of the circle-dividing engine about 1775, a device for dividing a circle into degrees with great accuracy, brought one of the greatest advances in surveying methods, as it enabled angle measurements to be made with portable instruments far more accurately than had previously been possible.

Modern surveying can be said to have begun by the late 18th century. One of the most notable early feats of surveyors was the measurement in the 1790s of the meridian from Barcelona, Spain, to Dunkirk, France, by two French engineers, Jean Delambre and Pierre Méchain, to establish the basic unit for the metric system of measurement.

Many improvements and refinements have been incorporated in all the basic surveying instruments. These have resulted in increased accuracy and speed of operations and opened up possibilities for improved methods in the field. In addition to modification of existing instruments, two revolutionary mapping and surveying changes were introduced: photogrammetry, or mapping from aerial photographs (about 1920), and electronic distance measurement, including the adoption of the laser for this purpose as well as for alignment (in the 1960s). Important technological developments starting in the late 20th century include the use of satellites as reference points for geodetic surveys and electronic computers to speed the processing and recording of survey data.

The planning and design of all Civil Engineering projects such as construction of highways, bridges, tunnels, dams etc are based upon surveying measurements.

Moreover, during execution, project of any magnitude is constructed along the lines and points established by surveying.

Thus, surveying is a basic requirement for all Civil Engineering projects.

Other principal works in which surveying is primarily utilised are

- to fix the national and state boundaries;
- to chart coastlines, navigable streams and lakes;
- to establish control points;
- to execute hydrographic and oceanographic charting and mapping; and
- to prepare topographic map of land surface of the earth.

1.3 Basic Principle Of Surveying.......

The following two basic principles should be considered while determining relative position of points on the surface of earth:-

1. determining suitable method for locating a point:- it is always practicableto select two points in the field to measure the distance between them. These can be represented on paper by two points placed in a convenient positions.

2.working from whole to the part:- in surveying an area, it is essential to establish first of all a system of control points with great precision. Minor control points can then be established by less precise method and the details can be located afterwards by method of triangulation or traversing between control points.

levelling................
IT is a branch of survey which helps in determination of elevation of a given point or object with espect to some specified point or object.
. Common levelling instruments include the spirit level, the dumpy level, the digital level, and the laser level.
The dumpy level is an older style of optical instrument. It is commonly believed that dumpy levelling is less accurate than other types of levelling, but such is not the case. Dumpy levelling requires shorter and therefore more numerous sights, but this fault is compensated by the practice of making foresights and backsights equal.

survey.........
surveying is the art of determining the relative positions of objects on the surface of earth.Surveying or land surveying is the technique, profession and science of accurately determining the terrestrial or three-dimensional position of points and the distances and angles between them.Surveying has been an essential element in the development of the human environment since the beginning of recorded history (about 5,000 years ago). It is required in the planning and execution of nearly every form of construction..

1.3.1 Classification of survey

Based on the purpose (for which surveying is being conducted), Surveying has been classified into:

- **Control surveying :** To establish horizontal and vertical positions of control points.

- **Land surveying :** To determine the boundaries and areas of parcels of land, also known as property survey, boundary survey or cadastral survey.

- **Topographic survey :** To prepare a plan/ map of a region which includes natural as well as and man-made features including elevation.

- **Engineering survey :** To collect requisite data for planning, design and execution of engineering projects. Three broad steps are

1) Reconnaissance survey : To explore site conditions and availability of infrastructures.

2) Preliminary survey : To collect adequate data to prepare plan / map of area to be used for planning and design.

3) Location survey : To set out work on the ground for actual construction / execution of the project.

- **Route survey :** To plan, design, and laying out of route such as highways, railways, canals, pipelines, and other linear projects.

- **Construction surveys :** Surveys which are required for establishment of points, lines, grades, and for staking out engineering works (after the plans have been prepared and the structural design has been done).

- **Astronomic surveys :** To determine the latitude, longitude (of the observation station) and azimuth (of a line through observation station) from astronomical observation.

- **Mine surveys :** To carry out surveying specific for opencast and underground mining purposes.

Surveying is the process by which a surveyor measures certain dimensions that generally occur on the surface of the Earth. Surveying equipment, such as levels and theodolites, are used for accurate measurement of angular deviation, horizontal, vertical and slope distances. With computerisation, electronic distance measurement (EDM), total stations, GPS surveying and laser scanning have supplemented (and to a large extent supplanted) the traditional optical instruments. This information is crucial to convert the data into a graphical representation of the Earth's surface, in the form of a map. This information is then used by civil engineers, contractors and even realtors to design from, build on, and trade, respectively. Elements of a building or structure must be correctly sized and positioned in relation to each other and to site boundaries and adjacent structures. Although surveying is a distinct profession with separate qualifications and licensing arrangements, civil engineers are trained in the basics of surveying and mapping, as well as geographic information systems. Surveyors may also lay out the routes of railways, tramway tracks, highways, roads, pipelines and streets as well as position other infrastructures, such as harbors, before construction.

Land surveying

In the United States, Canada, the United Kingdom and most Commonwealth countries land surveying is considered to be a distinct profession. Land surveyors are not considered to be engineers, and have their own professional associations and licencing requirements. The services of a licenced land surveyor are generally required for boundary surveys (to establish the boundaries of a parcel using its legal description) and subdivision plans (a plot or map based on a survey of a parcel of land, with boundary lines drawn inside the larger parcel to indicate the creation of new boundary lines and roads), both of which are generally referred to as cadastral surveying.

Construction surveying

Construction surveying is generally performed by specialised technicians. Unlike land surveyors, the resulting plan does not have legal status. Construction surveyors perform the following tasks:

- Survey existing conditions of the future work site, including topography, existing buildings and infrastructure, and even including underground infrastructure whenever possible;

- Construction surveying (otherwise "lay-out" or "setting-out"): to stake out reference points and markers that will guide the construction of new structures such as roads or buildings for subsequent construction;
- Verify the location of structures during construction;
- As-Built surveying: a survey conducted at the end of the construction project to verify that the work authorized was completed to the specifications set on plans.

Transportation engineering

Transportation engineering is concerned with moving people and goods efficiently, safely, and in a manner conducive to a vibrant community. This involves specifying, designing, constructing, and maintaining transportation infrastructure which includes streets, canals, highways, rail systems, airports, ports, and mass transit. It includes areas such as transportation design, transportation planning, traffic engineering, some aspects of urban engineering, queueing theory, pavement engineering, Intelligent Transportation System (ITS), and infrastructure management.

Municipal or urban engineering

Municipal engineering is concerned with municipal infrastructure. This involves specifying, designing, constructing, and maintaining streets, sidewalks, water supply networks, sewers, street lighting, municipal solid waste management and disposal, storage depots for various bulk materials used for maintenance and public works (salt, sand, etc.), public parks and bicycle paths. In the case of underground utility networks, it may also include the civil portion (conduits and access chambers) of the local distribution networks of electrical and telecommunications services. It can also include the optimizing of waste collection and bus service networks. Some of these disciplines overlap with other civil engineering specialties, however municipal engineering focuses on the coordination of these infrastructure networks and services, as they are often built simultaneously, and managed by the same municipal authority.

Forensic engineering

Forensic engineering is the investigation of materials, products, structures or components that fail or do not operate or function as intended, causing personal injury or damage to property. The consequences of failure are dealt with by the law of product liability. The field also deals with retracing processes and procedures leading to accidents in operation of vehicles or machinery. The subject is applied most commonly in civil law cases, although it may be of use in criminal law cases. Generally the purpose of a Forensic engineering investigation is to locate cause or causes of failure with a view to improve performance or life of a component, or to assist a court in determining the facts of an accident. It can also involve investigation of intellectual property claims, especially patents.

Control engineering

Control engineering or control systems engineering is the branch of Civil Engineering discipline that applies control theory to design systems with desired behaviors. The practice uses sensors to measure the output performance of the device being controlled (often a vehicle) and those measurements can be used to give feedback to the input actuators that can make corrections toward desired performance. When a device is designed to perform without the need of human inputs for correction it is called automatic control (such as cruise control for regulating a car's speed). Multi-disciplinary in nature, control systems engineering activities focus on implementation of control systems mainly derived by mathematical modeling of systems of a diverse range.

1.4 Chain survey

Chain survey is the simplest method of surveying. In this survey only measurements are taken in the field, and the rest work, such as plotting calculation etc. are done in the office. This is most suitable adapted to small plane areas with very few details. If carefully done, it gives quite accurate results. The necessary requirements for field work are chain, tape, ranging rod, arrows and some time cross staff.

Procedure in chain survey

Reconnaissance: The preliminary inspection of the area to be surveyed is calledreconnaissance. The surveyor inspects the area to be surveyed, survey or prepares index sketch or key plan.

Marking Station: Surveyor fixes up the required no stations at places from where maximum possible stations are possible.

Some of the methods used for marking are:

Fixing ranging poles

Driving pegs

Marking a cross if ground is hard

Digging and fixing a stone.

Then he selects the way for passing the main line, which should be horizontal and clean as possible and should pass approximately through the center of work.

Then ranging roads are fixed on the stations.

After fixing the stations, chaining could be started.

Make ranging wherever necessary.

Measure the change and offset.

Enter in the field the book.

Survey Station:

Survey stations are of two kinds

Main Stations

Subsidiary or tie

Main Stations:

Main stations are the end of the lines, which command the boundaries of the survey, and the lines joining the main stations re called the main survey line or the chain lines.

Subsidiary or the tie stations:

Subsidiary or the tie stations are the point selected on the main survey lines, where it is necessary to locate the interior detail such as fences, hedges, building etc.

Tie or subsidiary lines:

A tie line joints two fixed points on the main survey lines. It helps to checking the accuracy of surveying and to locate the interior details. The position of each tie line should be close to some features, such as paths, building etc.

Base Lines:

It is main and longest line, which passes approximately through the centre of the field. All the other measurements to show the details of the work are taken with respect of this line.

Check Line:

A check line also termed as a proof line is a line joining the apex of a triangle to some fixed points on any two sides of a triangle. A check line is measured to check the accuracy of the framework. The length of a check line, as measured on the ground should agree with its length on the plan.

Offsets:

These are the lateral measurements from the base line to fix the positions of the different objects of the work with respect to base line. These are generally set at right angle offsets. It can also be drawn with the help of a tape. There are two kinds of offsets:

1) Perpendicular offsets, and
2) Oblique offsets.

The measurements are taken at right angle to the survey line called perpendicular or right angled offsets.

The measurements which are not made at right angles to the survey line are called oblique offsets or tie line offsets.

A chain is made up of steel or iron pieces of wire known as links which are joined together with circular or oval rings that make for flexibility. It has a brack handle at both ends which is part and parcel of the total length of the chain known as chain length. A typical chain is made up of 100 links and has a bran tag at every 10th link called a teller. This makes for operating of length as the letters are numbered and differentiated from the next one for easy identification. Different kinds of chains exist including Equnter’s chain, Engineers chain and metric chains.

Generally, chains have been replaced with chains for linear surveys. Chains are now being studied to get the historical perspective of the development of survey equipments over the years.

To surveyors and collectors alike, the link chain symbolizes a rugged era, when surveying tools and techniques were literally defining America. The chain was a precision part of a surveyor's equipment and, as such, had to be calibrated and adjusted frequently, yet was sturdy enough to be dragged through rough terrain for years.

Owning a link chain now captures a bit of this glorious past; to heft it enhances the kinship one feels with the surveyor who toiled in the field long ago. As collectors, we need to identify the type of chain we own, in order to understand its history. Each chain bears the clues of its use, such as the wire gauge used, the materials and design used, the lengths of the whole and of each link, the manufacturer's stamps, the presence or absence of brazing, the tally-tags, and the presence or absence of linking rings. Noting these components will make it possible to approximate the date and purpose of your link chain with the aid of period manufacturer's catalogs. The following is a nicely detailed account from the 1910 Manual of the Principal Instruments used in American Engineering and Surveying, published by the W. & L. E. Gurley Company of Troy, New York.

Sizes of Wire - The sizes and diameters of iron and steel wire commonly used in making surveyor's and engineer's chains are as follows: No. 8, .162 inch; No. 10, .135 inch; No. 12, .105 inch; No. 15, .072 inch; and No. 18, .047 inch.

Land Surveyor's Chain - The ordinary Gunter's or surveyor's chain is sixty-six feet or four poles long, and is composed of one hundred links, connected each to each by two rings, and furnished with a tally mark at the end of every ten links. A link in measurement includes a ring at each end, and is seven and ninety two one hundredths inches long. In all the chains which we make the rings are oval and are sawed and well closed, the ends of the wore forming the hook being also filed and bent close to the link, to avoid kinking. The oval rings are about one third stronger than round ones.

Handles - The handles are of brass and form part of the end links, to which they are connected by a short link and jam nuts, by which the length of the chain is adjusted.

Tallies - The tallies are of brass, and have one, two, three or four notches, as they mark ten, twenty, thirty or forty links from either end. The fiftieth link is marked by a rounded tally to distinguish it from the others.

Half Chains - In place of the four pole chain just described, many surveyors prefer a chain two rods or thirty three feet long, having only fifty links, which are counted by tallies from one end in a single direction.

Iron and Steel Wire - Our surveyors' chains are made of Nos. 8 and 10 refined iron wire, and of Nos. 8, 10, 12 and 15 best steel wire. Steel chains are preferred on account of their greater strength, although they are more expensive than those of iron.

Engineers' Chains - Engineers' chains differ from surveyors' chains, in that a link including a ring at each end is one foot long, and the wire is of steel Nos. 8, 10 and 12. They are either fifty or one hundred feet long, and are furnished with swivel handles and tallies like those just described.

Brazed Steel Chains - A very light and strong chain is made of No. 12 steel wire, the links and rings of which are securely brazed. The wire is of a low spring temper, and the chain, though light, is almost incapable of being broken or stretched in careful use.

Our brazed steel chains have been found exceedingly desirable for all kinds of measurement, and for the use of engineers upon railroads and canals have very generally superseded the heavier chains.

Vara Chains - The meter is used as a standard measure of length in many countries, and chains of ten and twenty meters are often ordered. The chains are made of iron or steel wire, each meter being divided into five links. As a meter is 39.371 inches long, a link, including a ring at each end, measures 7.874 inches.

A ten meter chain has fifty links and a twenty meter chain one hundred links. Each meter is marked with a round brass tally numbered from one to nine in the ten meter chain, and from one to nineteen in the twenty meter chain.

Marking Pins - In chaining, eleven marking pins are needed, made either of iron, steel or brass wire, as preferred. They are about fourteen inches long, pointed at one end to enter the ground, and formed into a ring at the other end for convenience in handling.

Marking pins are sometimes loaded with a little mass of lead around the lower end, to serve as a plumb when the pin is dropped to the ground from the suspended end of the chain.

Chain survey is suitable in the following cases:
(i) Area to be surveyed is comparatively small
(ii) Ground is fairly level
(iii) Area is open and
(iv) Details to be filled up are simple and less.
In chain surveying only linear measurements are made i.e. no angular measurements are made. Since triangle is the only figure that can be plotted with measurement of sides only, in chain surveying the area to be surveyed should be covered with a network of triangles. Figure 12.11 shows a typical scheme of covering an area with a network of triangles. No angle of the network triangles should be less than 30º to precisely get plotted position of a station with respect to already plotted positions of other station. As far as possible angles should be close to 60º. However, the arrangements of triangles to be adopted depends on the shape, topography, natural and artificial obstacles in the field.

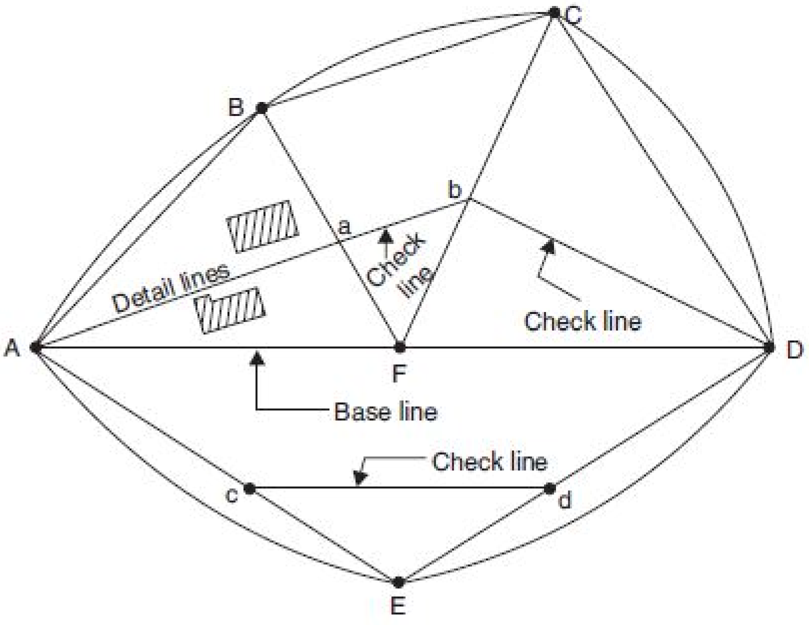

Figure 1.4

Technical Terms

Various technical terms used in connection with the network of the triangles in surveying are explained below:
Station: Station is a point of importance at the beginning or at the end of a survey line.
Main station: These are the stations at the beginning or at the end of lines forming main skeleton. They are denoted as A, B, C etc. Subsidiary or tie stations: These are the stations

selected on main lines to run auxiliary/secondary lines for the purpose of locating interior details. These stations are denoted as a, b, c, …., etc., or as 1, 2, 3, … etc. Base line: It is the most important line and is the longest. Usually it is the line plotted first and then frame work of triangles are built on it.

Detail lines: If the important objects are far away from the main lines, the offsets are too long, resulting into inaccuracies and taking more time for the measurements. In such cases the secondary lines are run by selecting secondary stations on main lines. Such lines are called detail lines.

Check lines: These are the lines connecting main station and a substation on opposite side or the lines connecting to substations on the sides of main lines. The purpose of measuring such lines is to check the accuracy with which main stations are located.

Selection of Stations

The following points should be considered in selecting station points:

(i) It should be visible from at least two or more stations.
(ii) As far as possible main lines should run on level ground.
(iii) All triangles should be well conditioned (No angle less than 30º).
(iv) Main network should have as few lines as possible.
(v) Each main triangle should have at least one check line.
(vi) Obstacles to ranging and chaining should be avoided.
(vii) Sides of the larger triangles should pass as close to boundary lines as possible.
(viii) Tresspassing and frequent crossing of the roads should be avoided.

Offsets

Lateral measurements to chain lines for locating ground features are known as offsets. For this purpose perpendicular or oblique offsets may be taken. If the object to be located (say road) is curved more number of offsets should be taken. For measuring offsets tapes are commonly used.

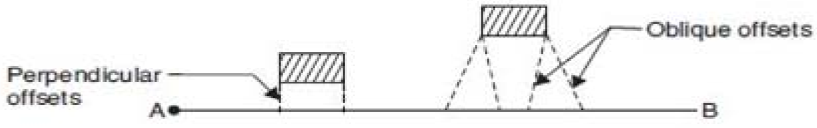

(a)

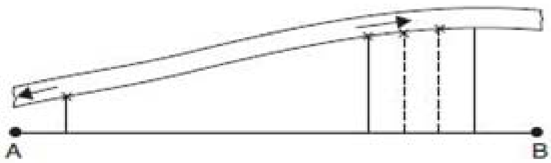

(b)

For setting perpendicular offsets any one of the following methods are used:
(i) Swinging
(ii) Using cross staffs
(iii) Using optical or prism square.
Perpendicular Offset by Swinging

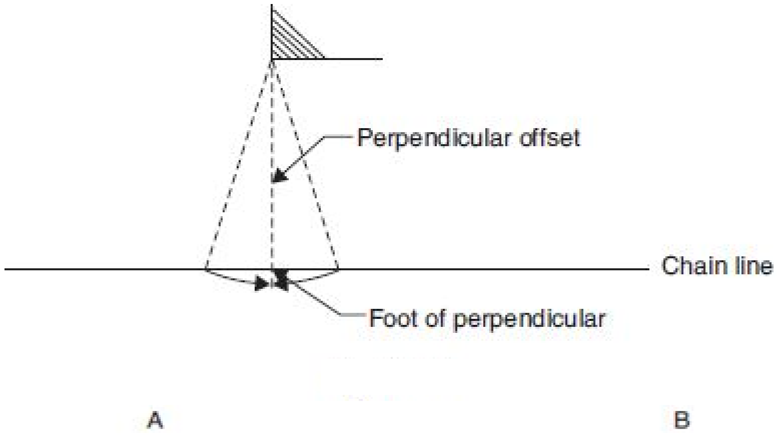

(b)

Chain is stretched along the survey line. An assistant holds the end of tape on the object. Surveyor swings the tape on chain line and selects the point on chain where offset distance is the least and notes chain reading as well as offset reading in a field book on a neat sketch of the object.

Perpendicular Offsets Using Cross Staffs

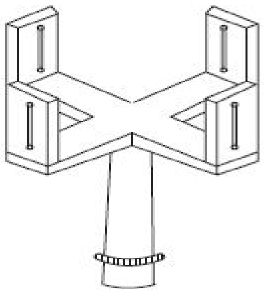

(a)

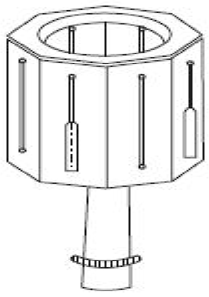

(b)

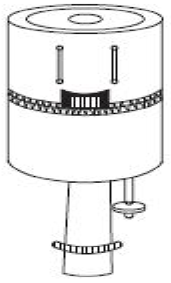

(c)

Above Figure shows three different types of cross staffs used for setting perpendicular offsets. All cross staffs are having two perpendicular lines of sights. The cross staffs are mounted on stand. First line of sight is set along the chain line and without disturbing setting right angle line of sight is checked to locate the object. With open cross staff it is possible to set perpendicular only, while with french cross staff, even 45º angle can be set. Adjustable cross staff can be used to set any angle also, since there are graduations and upper drum can be rotated over lower drum.

Perpendicular Offsets Using Optical Square and Prism Square

These instruments are based on the optical principle that if two mirrors are at angle ‘θ’ to each other, they reflect a ray at angle ‘2θ’. In below Figure shows a typical optical square.

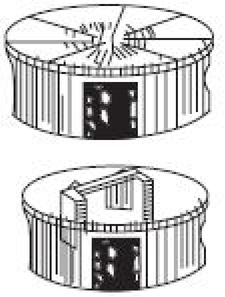

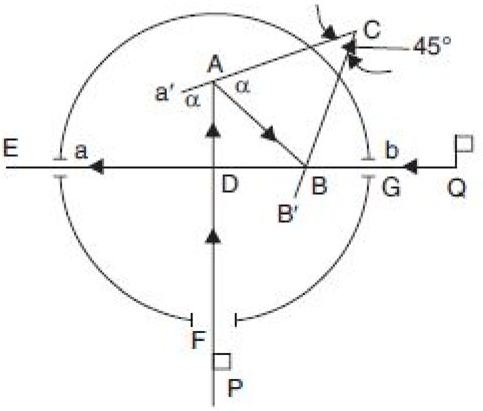

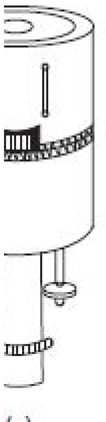

(c)

Optical square consists of a metal box about 50 mm in diameter and 125 mm deep. In the rim of the box there are three openings:

(i) a pin hole at E
(ii) a small rectangular slot at G, and
(iii) a large rectangular slot at F.

A and B are the two mirrors placed at 45º to each other. Hence the image of an object at F which falls on A gets reflected and emerge at E which is at right angles to the line FA. The mirror A which is opposite to the opening at F is fully silvered. It is fitted to a frame which is attached to the bottom plate. If necessary this mirror can be adjusted by inserting a key on the top of the cover. The mirror B which is in the line with EG is silvered in the top half and plain in the bottom half. It is firmly attached to the bottom plate of the box.

The ranging rod at Q is directly sighted by eye at E in the bottom half of the B which is a plain glass. At the same time in the top half of B, the reflected ray of the object at P is sighted. When the image of P is in the same vertical line as the object at Q, then the lines PA is at right angles to the line EB. This instrument can be used for finding foot of the perpendicular or to set a right angle. In prism square, instead of two mirrors at 45º to each other a prism which has two faces at 45º to each other is used. Its advantage is it will not go out of adjustment even after long usage.

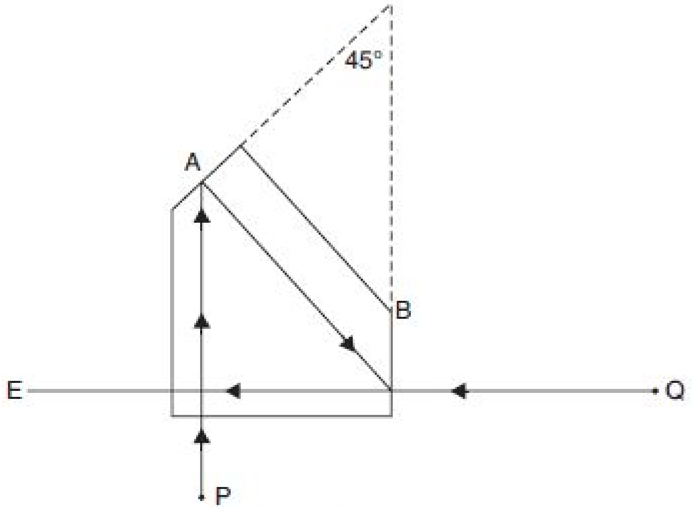

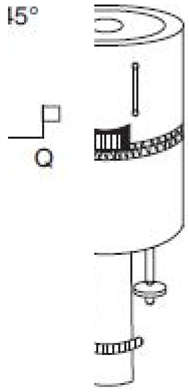

(c)

Field Book

All observations and measurements taken during chain surveying are to be recorded in a standard field book. It is a oblong book of size 200 mm × 120 mm, which can be carried in the pocket. There are two forms of the book (i) single line and (ii) double line. The pages of a single book are having a red line along the length of the paper in the middle of the width. It indicates the chain line. All chainages are written across it. The space on either side of the line is used for sketching the object and for noting offset distances. In double line book there are two blue lines with a space of 15 to 20 mm is the middle of each book. The space between the two lines is utilised for noting the chainages. in below Figure shows typical pages of a field books.

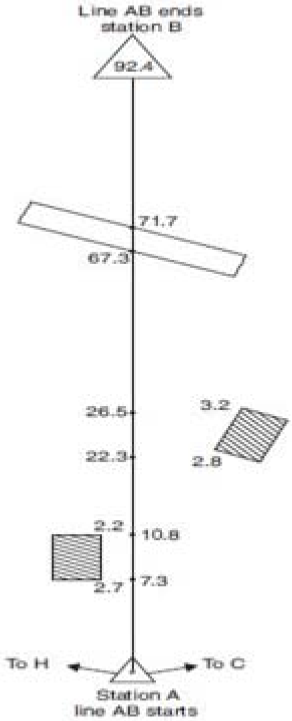
Line AB ends
station B
92.4
71.7
67.3
26.5
22.3
3.2
2.8
2.2
10.8
2.7
7.3
To H
To C
Station A
line AB starts

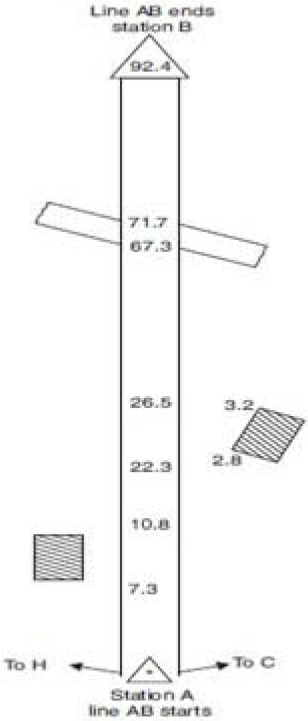
Line AB ends
station B
92.4
71.7
67.3
26.5
3.2
22.3
2.8
10.8
7.3
To H
To C
Station A
line AB starts

Field Work

As soon as the survey party arrives in the field the following details are entered in the field book:

(i) Title of the survey work
(ii) The date of survey
(iii) The names of the members of the party.

The field work may be divided into the following:

(i) Reconnaissance survey.
(ii) Marking stations, drawing reference sketches.
(iii) Line by line surveying.

Reconnaissance survey consists in going round the field and identifying suitable stations for the network of triangles. Neat sketch of network is drawn and designated. The typical key plan drawn is similar to one shown in above figure.

All main stations should be marked on the ground. Some of the methods used for marking are:

(a) Fixing ranging poles
(b) Driving pegs
(c) Marking a cross if ground is hard
(d) Digging and fixing a stone.

Then reference sketches are drawn in the field book so as to identify stations when the development works are taken up. For this measurements with respect to three permanent points are noted. The permanent points may be

(a) Corner of a building
(b) Posts of gates
(c) Corners of compound walls
(d) Electric poles
(e) A tree.

After that, line by line surveying is conducted to locate various objects with respect to chain lines.

Office Work

It consists in preparing the plan of the area to a suitable scale making use of measurements and sketches

noted in the field book.

1.5 COMPASS

Compass surveying is a branch of surveying in which directions of surveying lines are determined with a compass and the length of lines are measured with a tape or chain. In practice the compass si generally used to run a traverse.

In surveying,"Traverse" consists of a series of straight lines connected together to form a open or a closed polygon.

Methods of traversing: Depending on the type of instrument used for the measurement of angles the method of Traversing can be classified as under;

1.Chain Traverse

2.Compass Traverse.

3.Plane Table Traverse.

4.Stadia Traverse.

5.Theodolite Traverse.

In Compass Traverse the direction of the traverse lines are determined with a magnetic compass.

The magnetic Compass may be "SURVEYOR COMPASS" OR "PRISMATIC COMPASS"

The types of compass that are used commonly are: (i) prismatic compass; and (ii) surveyor compass.
The essential parts of both type are:
(i) a magnetic needle,
(ii) a graduated circle,
(iii) a line of sight, and
(iv) a box to house them.
There are some differences in the essential parts of the two type of compass. The construction of the two types of compass is explained and the difference in them is pointed out in this article.

Prismatic Compass

in below Figure shows the cross-section of a typical prismatic compass. A magnetic needle of broad form (1) is balanced on a hard and pointed steel pivot (2). The top of the pointed pivot is protected with agate cap (3).

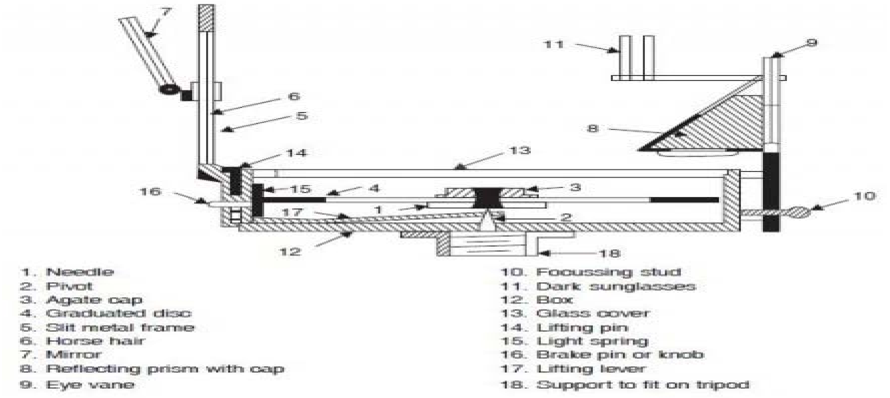
7
6
5
14
16
15
4
17
1
12
13
3
2
18
11
9
8
10
1. Needle
2. Pivot
3. Agate cap
4. Graduated disc
5. Slit metal frame
6. Horse hair
7. Mirror
8. Reflecting prism with cap
9. Eye vane
10. Focussing stud
11. Dark sunglasses
12. Box
13. Glass cover
14. Lifting pin
15. Light spring
16. Brake pin or knob
17. Lifting lever
18. Support to fit on tripod

Prismatic compass

An aluminium graduated disk (4) is fixed to the top of the needle. The graduations are from zero to 360° in clockwise direction when read from top. The direction of north is treated as zero degrees, east as 90°, south as 180° and west as 270°. However, while taking the readings observations are at the other end of line of sight. Hence, the readings are shifted by 180° and graduations are marked as shown in below figure. The graduations are marked inverted because they are read through a prism.

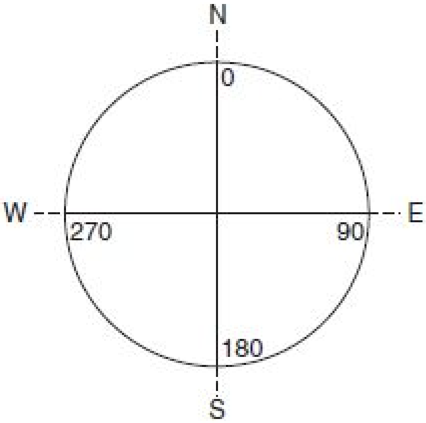

(a)

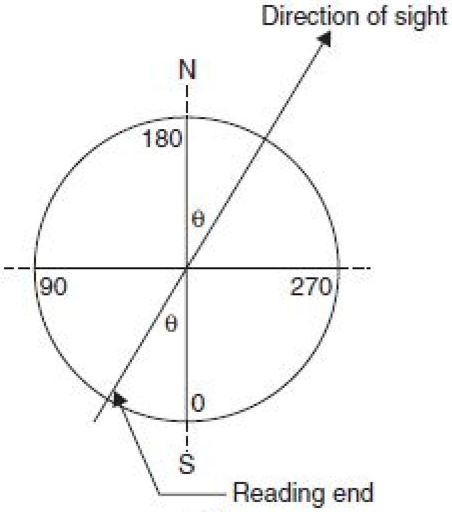

(b)

The line of sight consists of object unit and the reading unit. Object unit consists of a slit metal frame hinged to the box. In the centre the slit is provided with a horse hair or a fine wire or thread The metal frame is provided with a hinged mirror , which can be placed upward or downward on the frame. It can be slided along the frame. The mirror can be adjusted to view objects too high or too low from the position of compass. Reading unit is provided at diametrically opposite edge. It consists of a prism with a sighting eye vane . The prism magnifies the readings on the graduation disk just below it. For focussing, the prism is lowered or raised on the frame carrying it and then fixed with the stud. Dark sunglasses provided near the line of sight can be interposed if the object to be sighted is bright (e.g., sun). The bottom of the box which is about 85 mm to 110 mm supports the pivot of needle firmly at its centre. The object vane and the prism are supported on the sides of the box. The box is providedwith a glass lid which protects the graduation disc at the same time permit the direct reading from the top. When the object vane is folded on the glass top it presses a lifting pin which activates lifting lever lifts the needle off the pivot. Thus, it prevents undue wear of pivot point. While taking reading, if graduation disc vibrates, it can be dampened with a spring. For pressing spring a knob or brake pin is provided on the box. When not in use prism can be folded over the edge of the box. The box is provided with a lid to close it when the compass is not in use. The box is provided with a socket to fit it on the top of a tripod.

Surveyors Compass

In this type of compass graduation disc is fixed to the box and magnetic needle is free to rotate above it. There is no prism provided at viewing end, but has a narrow slit. After fixing the line of sight, the reading is directly taken from the top of the glass cover. Hence, graduations are written directly (not inverted). In this compass graduations are from zero to 90°, zero being to north or south and 90° being to east and west. An angle of 20° to north direction to the east is written as N 20° E, and an angle of 40° to east from south is written as S 40° E. Always first direction indicated is north or south and the last letter indicates east or west direction. In this system graduated circle rotates with line of sight and magnetic needle is always towards north. The reading is taken at the tip of needle. Hence, on the compass east and west are marked interchanged and marked shows the photograph of a surveyors compass.

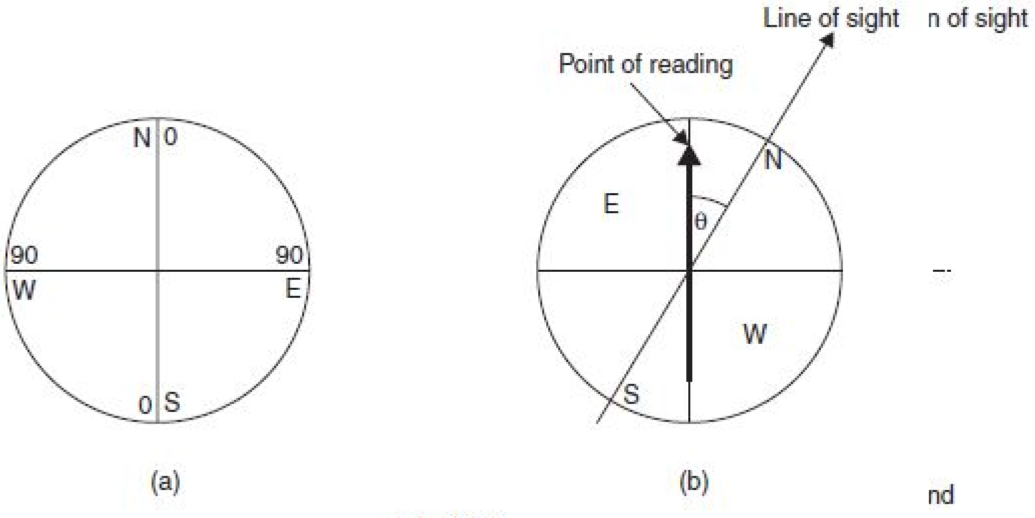

Fig. 13.3

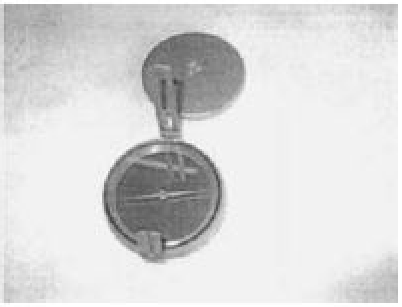

Difference Between Prismatic Compass and Surveyors Compass
The difference between prismatic and surveyor’s compass are listed in below Table.

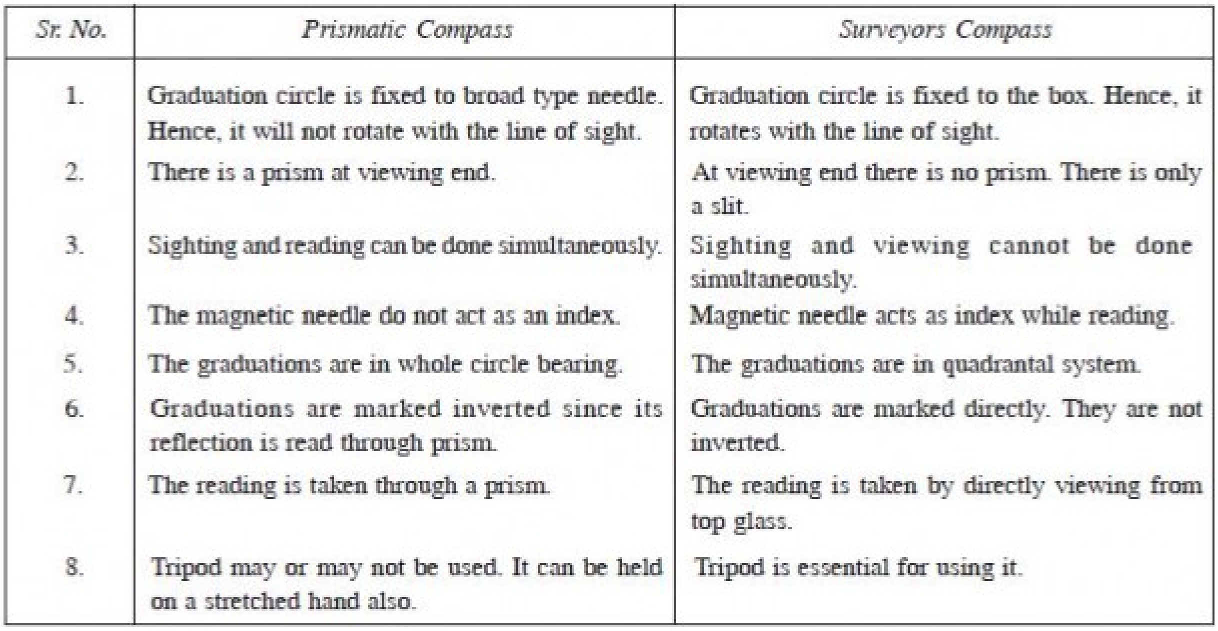

Sr. No.	Prismatic Compass	Surveyors Compass
1.	Graduation circle is fixed to broad type needle. Hence, it will not rotate with the line of sight.	Graduation circle is fixed to the box. Hence, it rotates with the line of sight.
2.	There is a prism at viewing end.	At viewing end there is no prism. There is only a slit.
3.	Sighting and reading can be done simultaneously.	Sighting and viewing cannot be done simultaneously.
4.	The magnetic needle do not act as an index.	Magnetic needle acts as index while reading.
5.	The graduations are in whole circle bearing.	The graduations are in quadrantal system.
6.	Graduations are marked inverted since its reflection is read through prism.	Graduations are marked directly. They are not inverted.
7.	The reading is taken through a prism.	The reading is taken by directly viewing from top glass.
8.	Tripod may or may not be used. It can be held on a stretched hand also.	Tripod is essential for using it.

Bearing

As stated earlier a bearing of a line is the angle made by the line with respect to a reference direction, the reference direction being known as meridian. The direction shown by a freely suspended and properly balanced magnetic needle is called magnetic meridian and the horizontal angle made by a line with this meridian is known as magnetic bearing. The points of intersection of earth's axis with surface of the earth are known as geographic north and south pole. The line passing through geographic north, south and the point on earth is called true meridian at that point and the angle made by a line passing through that point is called truebearing. While traversing along lines A, B, C, D …, the bearing of lime AB is called fore bearing of AB and the bearing of BA is called back bearing. Fore bearing and back bearing differ by 180°.

1.6 PLANE TABLE SURVEYING

In this method of surveying a table top, similar to drawing board fitted on to a tripod is the main instrument. A drawing sheet is fixed on to the table top, the observations are made to the objects, distances are scaled down and the objects are plotted in the field itself. Since the plotting is made in the field itself, there is no chance of omitting any necessary measurement in this surveying. However the accuracy achieved in this type of surveying is less. Hence this type of surveying is used for filling up details between the survey stations previously fixed by other methods. In this chapter, accessories required, working operations and methods of plane table surveying are explained. At the end advantages and limitations of this method are listed.

Plane Table and Its Accessories

The most commonly used plane table is shown in below Fig. It consists of a well seasoned wooden table top mounted on a tripod. The table top can rotate about vertical axis freely. Whenever necessary table can be clamped in the desired orientation. The table can be levelled by adjusting tripod legs.

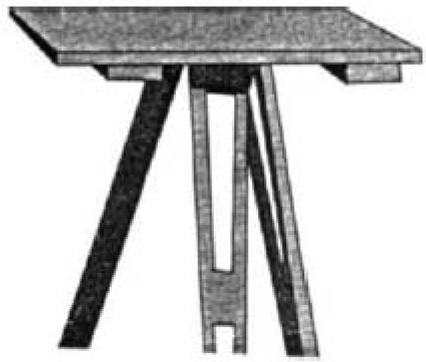

Adjusting the tripod

Plane table with stand

The following accessories are required to carry out plane table survey:

1. Alidade
2. Plumbing fork with plumb bob.

3. Spirit level
4. Trough compass
5. Drawing sheets and accessories for drawing.

Alidade

It is a straight edge ruler having some form of sighting device. One edge of the ruler is bevelled and is graduated. Always this edge is used for drawing line of sight. Depending on the type of line of sight
there are two types of alidade:
(a) Plain alidade (b) Telescopic alidade

Plain Alidade: in below Figure shows a typical plain adidate. A sight vane is provided at each end of the ruler. The vane with narrow slit serves as eye vane and the other with wide slit and having a thin wire at its centre serves as object vane. The two vanes are provided with hinges at the ends of ruler so that when not in use they can be folded on the ruler. Plain alidade is not suitable in surveying hilly areas as the inclination of line of sight in this case is limited.

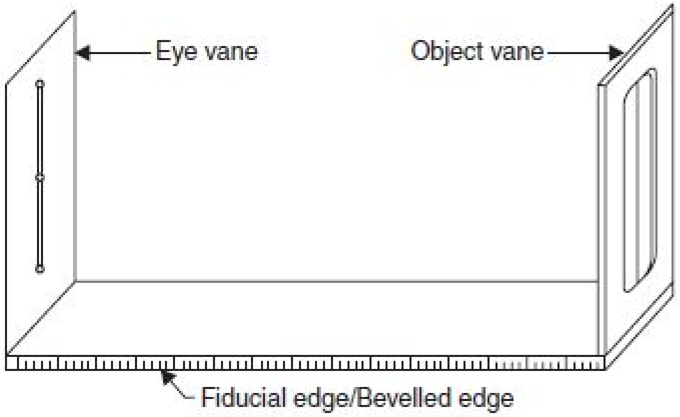
Eye vane
Object vane
Fiducial edge/Bevelled edge

Plane alidade

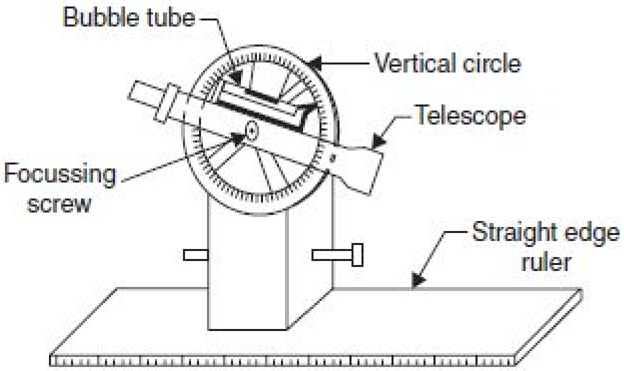
Bubble tube
Vertical circle
Telescope
Focussing screw
Straight edge ruler

Telescopic alidade

Telescopic Alidade: It consists of a telescope mounted on a column fixed to the ruler. The line of sight through the telescope is kept parallel to the bevelled edge of the ruler. The telescope is provided with a level tube and vertical graduation arc. If horizontal sight is required bubble in the level tube is kept at the centre. If inclined sights are required vertical graduation helps in noting the inclination of the line of sight. By providing telescope the range and the accuracy of line of sight is increased.

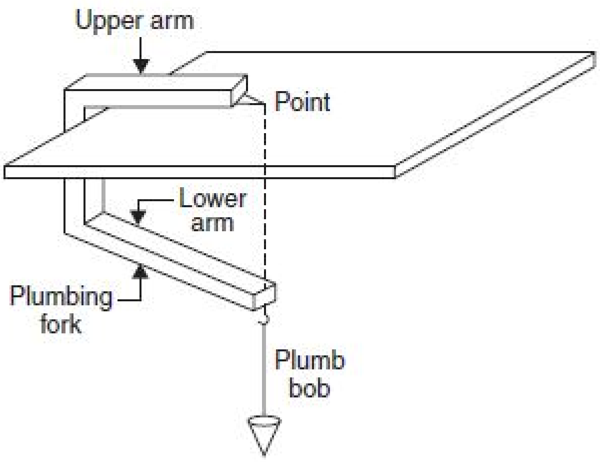

Plumbing fork and plumb bob.

Plumbing Fork and Plumb Bob

Figure shows a typical plumbing fork with a plum bob. Plumbing fork is a U-shaped metal in frame with a upper horizontal arm and a lower inclined arm. The upper arm is provided with a pointer at the end while the lower arm is provided with a hook to suspend plumb bob. When the plumbing fork is kept on the plane table the vertical line (line of plumb bob) passes through the pointed edge of upper arm. The plumb bob helps in transferring the ground point to the drawing sheet and vice versa also.

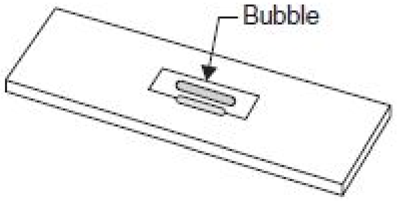

Spirit level

SpiritLevel

A flat based spirit level is used to level the plane table during surveying. To get perfect level, spirit level should show central position for bubble tube when checked with its positions in any two mutually perpendicular direction.

TroughCompass

It consists of a 80 to 150 mm long and 30 mm wide box carrying a freely suspended needle at its centre . At the ends of the needle graduations are marked on the box to indicate zero to five degrees on either side of the centre. The box is provided with glass top to prevent oscillation of the needle by wind. When needle is centred (reading 0–0), the line of needle is parallel to the edge of the box. Hence marking on the edges in this state indicates magnetic north–south direction.

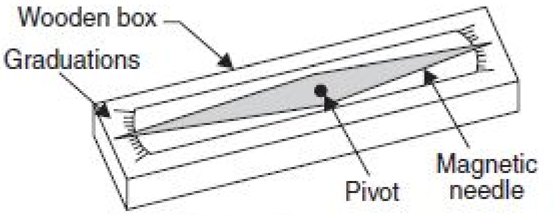

Fig. 14.6. Trough compass

Drawing Sheet and Accessories for Drawing
A good quality, seasoned drawing sheet should be used for plane table surveying. The drawing sheet may be rolled when not in use, but should never is folded. For important works fibre glass sheets or paper backed with thin aluminium sheets are used.
Clips clamps, adhesive tapes may be used for fixing drawing sheet to the plane table. Sharp hard pencil, good quality eraser, pencil cutter and sand paper to keep pencil point sharp are other accessories required for the drawing work. If necessary, plastic sheet should be carried to cover the drawing sheet from rain and dust.

1.6.1 Working Operation

After fixing the table top to the stand and drawing sheet to the table, the following operations are to be carried out before map making:

1. Centering
2. Levelling
3. Orientation.

1 **Centering**
Centering is the process of setting the plane table on the point so that its plotted position is exactly over the position on the ground. This is achieved by moving the legs of the tripod and checking the position of the point on the ground and on the paper with the help of plumbing fork and plumb bob.

2 **Levelling**
The level of the plane table should be ensured in two positions of spirit level which are at right angles to each other. The legs of tripod are moved radially or along the circumference to adjust the plane table and get levelled surface.

3 **Orientation**
Orientation is the process of setting plane table over a station such that all the lines already plotted are parallel to corresponding lines on the ground. Accuracy of plane table survey mainly depends upon the accuracy of orientation of plane table at each station point. It can be achieved by any one of the following methods:
(a) using trough compass
(b) by back sighting
(c) by solving two point or three point problems.
The first two methods are commonly used while the third method is used occationally. The third method is explained under the article methods of plane tabling by resection.

(a) Orientation Using Trough Compass: When the survey work starts, the plane table is set on first station and the table is oriented by rough judgement such that the plotted position of the area falls in the middle portion of the paper. Then the table is clamped and the north direction is marked on right hand side top corner of drawing sheet. Trough compass is used to identify north direction. This orientation is to be maintained at all subsequent stations. After centering and levelling the table trough compass is kept along the marked north direction and the table is rotated to get freely suspended magnetic needle centred. After achieving it the table is clamped.
This method of orientation is considered rough, since the local attraction to magnetic needle affects the orientation. This method is used as preliminary orientation and finer tuning is made by observing the already plotted points.

(b) Orientation by Back Sighting: It is the commonly used method in plane table surveying. After completing surveying from plane table set at A, if table is to be shifted to next station B, a line is drawn from the plotted position of station A towards station B. Then distance AB is measured, scaled down and plotted position of station B is obtained. Then table is shifted to station B, centred, levelled. Then keeping alidade along BA, station A is sighted and the table is clamped. Thus the orientation of the table is achieved by back sighting. Orientation may be checked by observing already plotted objects.

Method of planning table

The following four methods are available for carrying out plane table survey:
1. Radiation
2. Intersection
3. Traversing
4. Resection.

The first two methods are employed for locating details while the other two methods are used for locating position of plane table station on drawing sheet.

1 **Radiation**

After setting the plane table on a station, say O, it is required to find the plotted position of various objects A, B, C, D ….. . To get these positions, the rays OA, OB, OC ….. are drawn with soft pencil. Then the distances OA, OB, OC ….., are measured scaled down and the positions of A, B, C ….., are found on the drawing sheets. This method is suitable for surveying small areas and is convenient if the distances to be measured are small. For larger areas this method has wider scope, if telescopic alidade is used, in which the distances are measured technometrically.

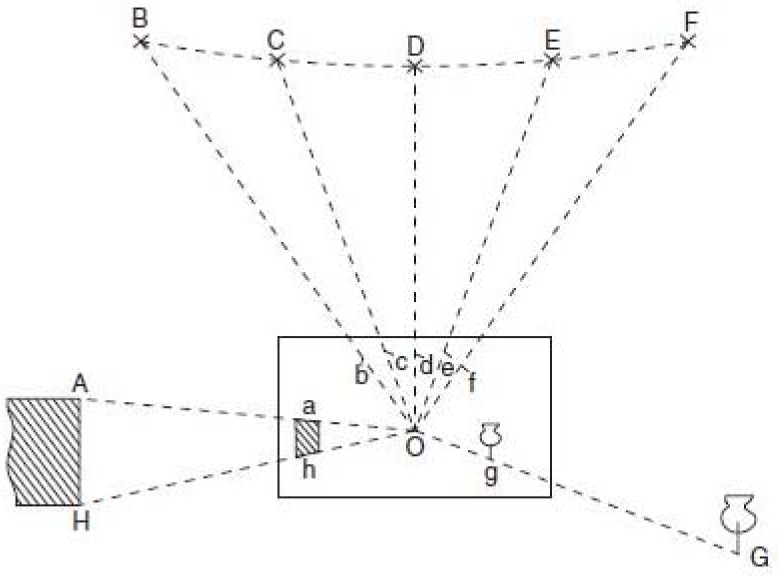

Radiation method of plane tabling

Intersection

In this method the plotted position of an object is obtained by plotting rays to the object from two stations. The intersection gives the plotted position. Thus it needs the linear measurements only between the station points and do not need the measurements to the objects. In below Figure shows the method for locating objects A and B from plane table positions O1 and O2.

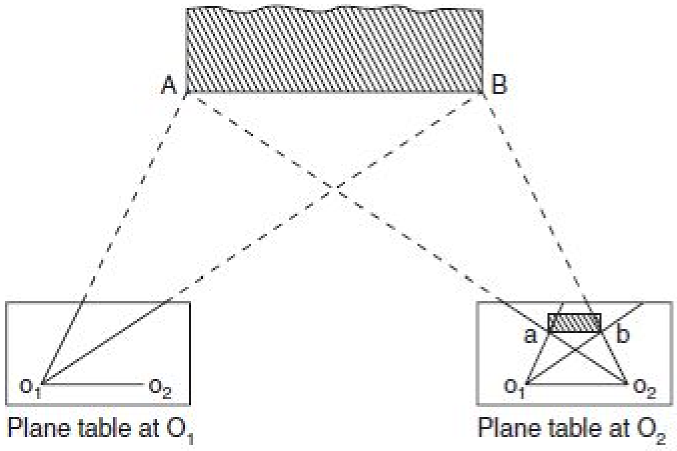

Intersection method of plane tabling

This method is commonly employed for locating:

(a) details

(b) the distant and inaccessible points

(c) the stations which may be used latter.

Traversing

This is the method used for locating plane table survey stations. In this method, ray is drawn to next station before shifting the table and distance between the stations measured. The distance is scaled down and next station is located. After setting the plane table at new station orientation is achieved by back sighting. To ensure additional checks, rays are taken to other stations also, whenever it is possible. In below Figure shows a scheme of plane table survey of closed area. This method can be used for open traverses also.

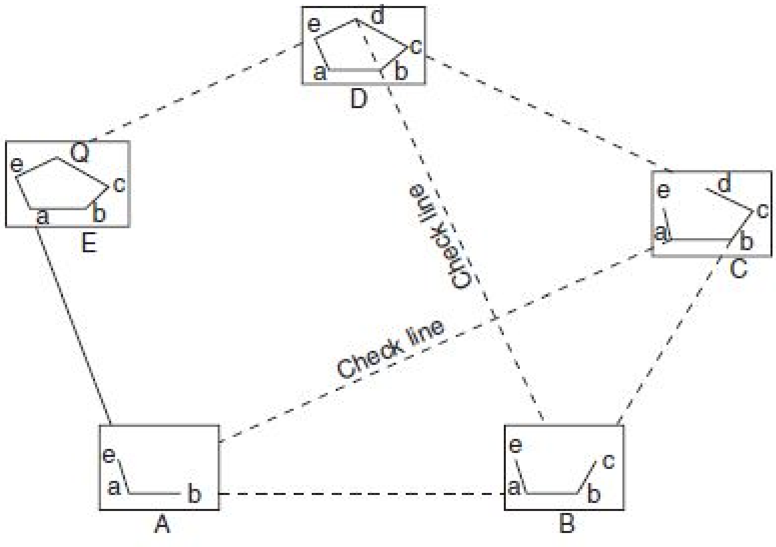

Plane table traversing

Resection

This method is just opposite to the method of intersection. In the method of intersection, the plotted position of stations are known and the plotted position of objects are obtained by intersection. In this method the plotted position of objects are known and the plotted position of station is obtained. If a, b and c are the plotted positions of objects A, B and C respectively, to locate instrument station P on the paper, the orientation of table is achieved with the help of a, b, c and then resectors Aa, Bb, Cc are drawn to get the ‘p’ , the plotted position of P. Hence in the resection method major work is to ensure suitable orientation by any one of the methods. The following methods are employed in the method of resection:

(a) by compass
(b) by back sighting
(c) by solving two point problem
(d) by solving three point problem.

(a) Resection after Orientation by Compass: Let a and b be the plotted positions of A and B of two well defined points in the field. Keeping the through compass along north direction marked on the drawing sheet table is oriented on station P, the position of which is to be found on paper. The resectors Aa and Bb are drawn to locate ‘p’ the plotted position of station point P. This method gives satisfactory results, if the area is not influenced by local attractions. It is used for small scale mapping only.

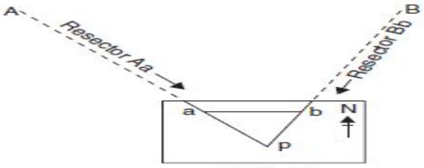

Resection after orientation with compass

(b) Resection after Orientation by Back Sighting: in below Figure shows the scheme of resection after orientation by back sighting. From station A, the position of B is plotted as ‘b’ and ray has been taken to station P as ap′. Then plane table is set at P and oriented by back sighting A, line AP is not measured but the position of P is obtained on the paper by taking resection Bb.

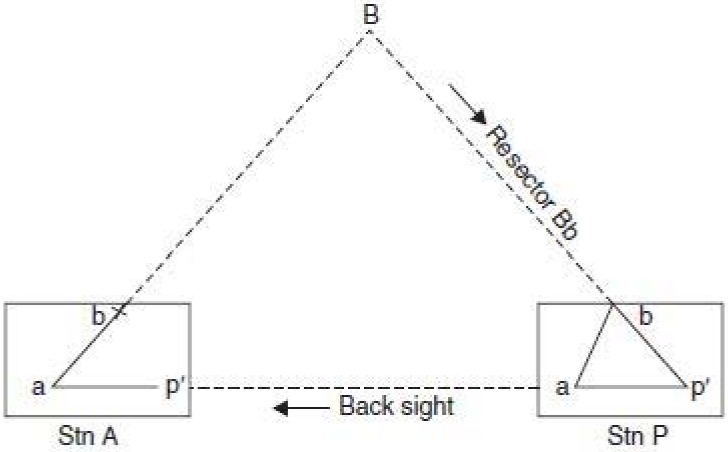

Resection after back sighting

(c) Resection after Solving Two Point Problem: The problem of finding plotted position of the station point occupied by the plane table with the help of plotted positions of two well defined points is known as solving two point problem. In below Figure shows the scheme of solving this.

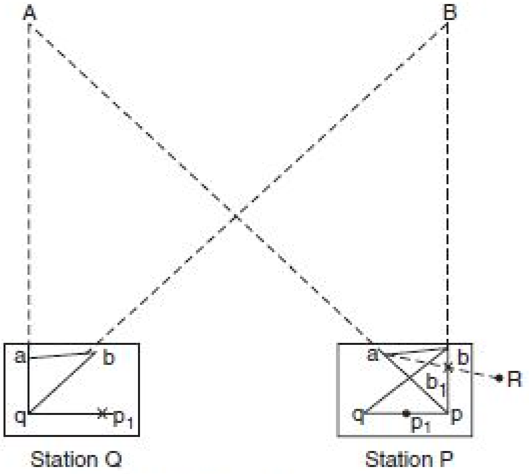

Two-point problem

Let A and B be two well defined points like lightening conductor or spire of church, the plotted positions a and b already known. Now the problem is to orient the table at P so that by resection its plotted position p can be obtained. The following steps may be followed to solve this problems:

(i) Select a suitable point Q near P such that the angles PAQ and PBQ are not accute.
(ii) Roughly orient the table at Q and draw the resectors Aa and Bb to get the point ‘q’.
(iii) Draw the ray qp and locate p1 with estimated distance QP.
(iv) Shift the plane table to P and orient the table by back sighting to Q.
(v) Draw the resector Aa to get ‘p’.
(vi) Draw the ray pB. Let it intersect line bq at b1.

(vii) The points b and b1 are not coinciding due to the angular error in the orientation of table. The angle bab, is the angular error in orientation. To correct it,

* Fix a ranging rod at R along ab, * Unclamp the table and rotate it till line ab sights ranging rod at R. Then clamp the table. This gives the correct orientation of the table which was used in plotting the points A and B. (viii) The resectors Aa and Bb are drawn to get the correct plotted position ‘p’ of the station P.

(d) Resection after Solving Three Point Problem: Locating the plotted position of a station point using observations to three well defined points whose plotted positions are known, is called solving three point problem.

Let A, B, C be three well defined objects on the field whose plotted positions a, b and c are known. Now the problem is to locate plotted position of the station point P. Any one of the following methods can be used.

(i) Mechanical (Tracing paper) method,

(ii) Graphical method, or

(iii) Trial and error method (Lehman’s method).

(i) Mechanical Method: This method is known as tracing paper method since it needs a tracing paper. The method involved the following steps

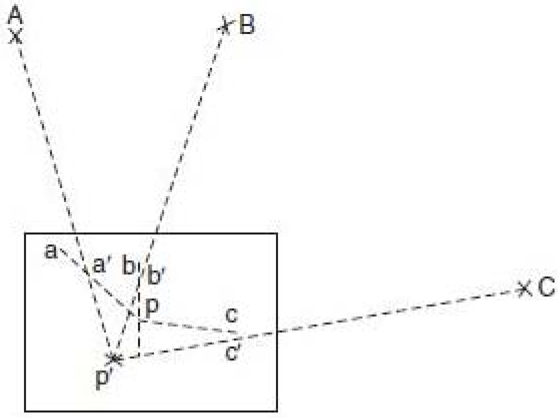

* Set the table over station P and by observation approximately orient the table.

* Fix the tracing paper on the plane table and select P approximately, say as p′. From p′, draw p′ A, p′ B and p′ C. These lines may not pass through the plotted positions a, b and c since the orientation is not exact.

* Loosen the tracing paper and rotate it so that the rays pass through respective points a, b and c. Now prick the point p′ to get the plotted position ‘p’ of the station P.

* Keep the alidade along pa and sight A. Then clamp the table. This is correct orientation. Check the orientation by observing along pb and pc.

(ii) Graphical Method: The following two graphical methods are available to solve three point problem:

* Bessel’s solution

* Method of perpendiculars.

Bessels Solution: It involves the following steps:

1. Keep the bevelled edge of alidade along ba and sight object at A. Clamp the table and draw bc′ along the line bc.

2. Keep bevelled edge of alidade along ab, unclamp the table and sight B. Clamp the table. Draw line ac intersecting bc′ at d.

3. Keep the alidade along dc and bisect C. Clamp the table. This gives the

correct orientation.
4. Draw resectors to get ‘p’.

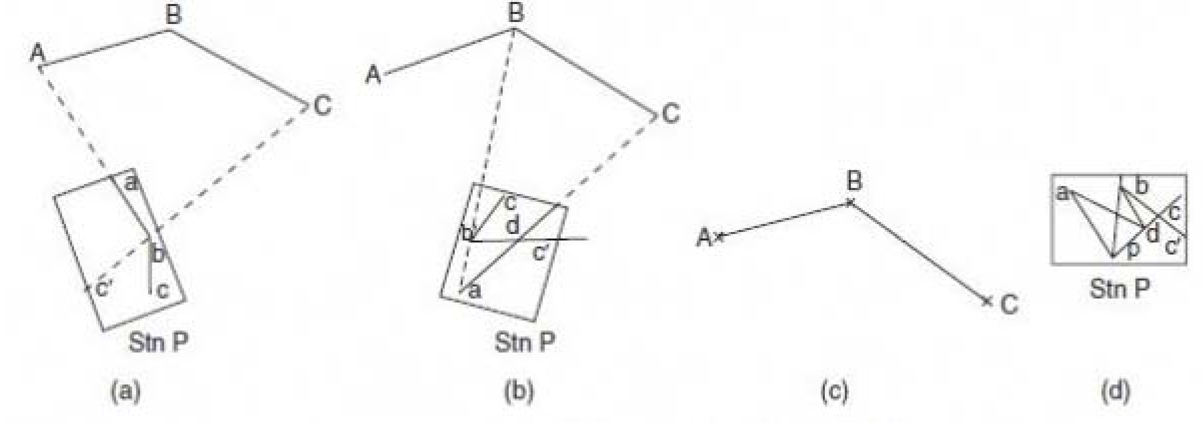

Graphical solution (Bessel’s method)

Method of Perpendiculars

This is another graphical method. It involves the following steps.
1. Draw line ae perpendicular to ab. Keep alidade along ea and turn the table till A is sighted. Clamp the table and draw the ray Bb to intersect the ray Aac at e.
2. Draw cf perpendicular to bc and clamp the table when fcC are in a line. Draw Bb to intersect Ccf at F .

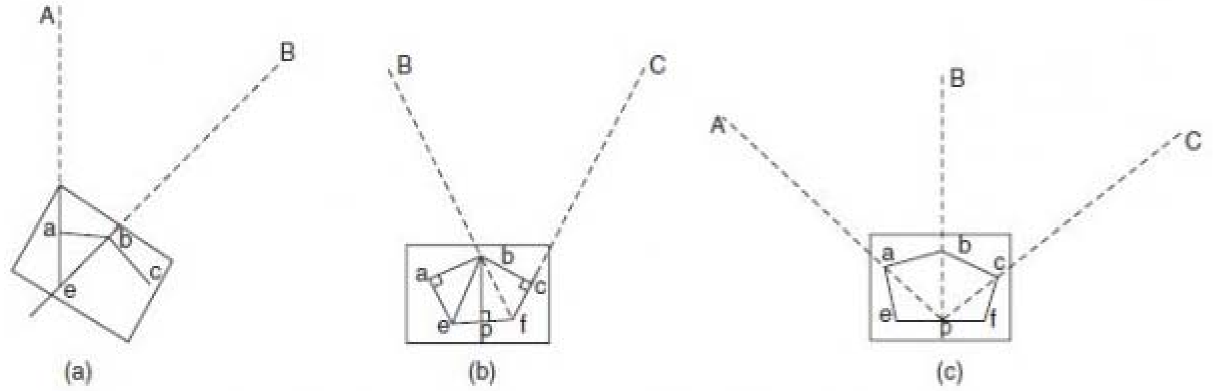

Method of perpendiculars of solve three point problem

3. Join cf drop bp perpendicular to ef to get the plotted position ‘p’.
4. Orient the table such that pbB are in a line. Clamp the table to place it in correct orientation. Resections Aa and Cc may be used to check the orientation.

Trial and Error Method

This method is also known as ‘triangle of error method’ and ‘Lehman’s Method’. It involves the following steps:
1. Set the table over point P and orient the table approximately, just by observation.
2. Draw the rays aA, bB and cC. If the orientation was perfect, the three rays would have intersected at a single point, i.e. at point ‘p’. Otherwise a triangle of error is formed.
3. To eliminate the triangle of error an approximate position, ray p′, is selected near the triangle of error. Then keeping alidade along p′a object A is sighted and the table is clamped. Draw the

resectors cC and bB to check the orientation.

4. Above step is repeated till triangle of error is eliminated.

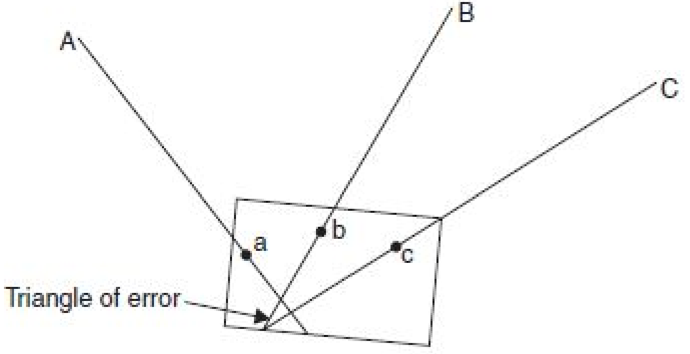

Lehman presented the following guidelines to select ‘p′’ so that triangle of error is eliminated quickly.

Rule 1: The distance of point sought ‘p’ is in the same proportion from the corresponding rays as the distance of those from the plane table station.

Rule 2: The point sought ‘p’ is on the same side of all the three resectors.

Defining the triangle ABC on the field as great triangle and the circle passing through them as great circle, from the above two rules of Lehman, the following sub-rules may be drawn.

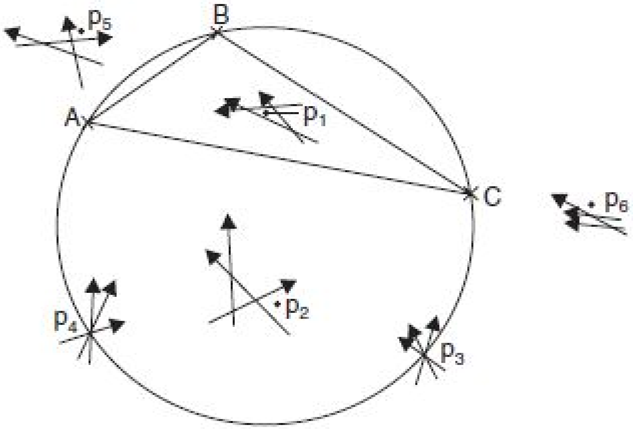

* If ‘P’ lies within the great triangle, the point ‘p’ is within the triangle of error

* If the plane table station P lies outside the great triangle the point sought ‘p’ is outside the triangle of errors (p2).

* If the ‘P’ is on the great circle, the correct solution is impossible (p3 and p4).

* If ‘P’ is outside the great circle, ‘p’ is nearer to the intersection of rays to the nearest two points (P5).

* If point P is outside the great circle and the two rays drawn are parallel to each other the point sought is outside the parallel lines and on the same side of the three rays (P6).

1.6.2 Error in Plane Table Serveying

The errors may be grouped into the instrumental and personal errors.

Instrumental Errors
1. The surface of plane table not perfectly plane.
2. Bevelled edge of alidade not straight.
3. Sight vanes of alidade not perfectly perpendicular to the base.
4. Plane table clamp being loose.
5. Magnetic compass being sluggish.
6. Drawing sheet being of poor quality.

Personal Errors
1. Centering errors
2. Levelling errors
3. Orientation errors
4. Sighting errors
5. Errors in measurement
6. Plotting errors
7. Errors due to instability of tripod.

Advantages are
1. Possibility of omitting measurement is eliminated.
2. The surveyor can compare the plotted work in the field then and there only.
3. Irregular objects are plotted more accurately, since they are seen while plotting.
4. Booking errors are eliminated.
5. Local attractions do not influence the plotting.
6. No great skill is required to produce satisfactory maps.
7. Method is fast.
8. No costly instruments are required.

Limitations are
1. Survey cannot be conducted in wet weather and rainy days.
2. Plane table is cumbersome and heavy to carry.
3. It needs many accessories.
4. It is less accurate.
5. Reproduction of map to different scale is difficult.

Review Questions

1. **Describe chain principle and classification with example?**
2. **What is surveying explain in detail?**
3. **Explain the chaining in reciprocal ?**
4. **What is the technique of compass in surveying ?**
5. **What is pessimistic compass explain with example?**
6. **What is adjustment of error explain?**
7. **Explain Local Attraction?**
8. **Explain the technique of Plain Table Serveying ?**
9. **What are the merit and demerit of plane table surveying?**
10. **What are the radiation intersection and radiation traversing explain?**

Chapter 2

LEVELLING AND THEODOLITE SURVEYING

Structure of this unit

LEVELLING, THEODOLITE

Learning Objectives

1. LEVELLING
2. Level line
3. Horizontal line
4. Levels and Staves
5. Spirit level
6. sensitiveness
7. Bench marks
8. Temporary and Permanent adjustments
9. Fly and check levelling – Booking
10. Reduction
11. Curvature and Refraction
12. Reciprocal levelling
13. Longitudinal and Cross sections – Plotting .
14. THEODOLITE
15. Theodolite
16. Vernier and Microptic
17. Description and uses
18. Temporary and Permanent adjustments of vernier transit
19. Horizontal angles – Heights and Distances
20. Traversing – Closing error and distribution.

2.1 LEVELLING

Elevation measurements involve measurements in vertical plane. It is also known as levelling. It may be defined as the art of determining the elevations of given points above or below a datum line or establishing given points of required heights above or below the datum line.

Levelling (or **leveling** in American spelling) is a branch of surveying, the object of which is to

1. Find the elevation of a given point with respect to the given or assumed Datum.
2. Establish a point at a given elevation with respect to the given or assumed Datum.

Levelling is the measurement of geodetic height using an optical levelling instrument and a level staff or rod having a numbered scale. Common levelling instruments include the spirit level, thedumpy level, the digital level, and the laser level.

Spirit (Optical) levelling

Spirit levelling employs a spirit level, an instrument consisting of a telescope with a crosshair and a tube level like that used by carpenters, rigidly connected. When the bubble in the tube level is centered the telescope's line of sight is supposed to be horizontal (i.e. perpendicular to the local vertical).

The spirit level is on a tripod midway between the two points whose height difference is to be determined. A leveling staff or rod is held vertical on each point; the rod is graduated in centimetres and fractions or tenths and hundredths of a foot. The observer focuses in turn on each rod and reads the value. Subtracting the "back" and "forward" value provides the height difference.

We can't expect the instrument to be in perfect adjustment, but we can hope that when the bubble is centered the telescope's line of sight is always the same small angle off of horizontal. If it is, we can still level accurately by setting the instrument equidistant from the points to be measured, so the errors cancel.

2.1.1 Leveling Procedure

A typical procedure is to set up the instrument within 100 metres (110 yards) of a point of known or assumed elevation. A rod or staff is held vertical on that point and the instrument is used manually or automatically to read the rod scale. This gives the height of the instrument above the starting (backsight) point and allows the height of the instrument (H.I.) above the datum to be computed.

The rod is then held on an unknown point and a reading is taken in the same manner, allowing the elevation of the new (foresight) point to be computed. The procedure is repeated until the destination point is reached. It is usual practice to perform either a complete loop back to the starting point or else close the traverse on a second point whose elevation is already known. The closure check guards against blunders in the operation, and allows residual error to be distributed in the most likely manner among the stations.

Some instruments provide three crosshairs which allow stadia measurement of the foresight and backsight distances. These also allow use of the average of the three readings (3-wire leveling) as a check against blunders and for averaging out the error of interpolation between marks on the rod scale.

The two main types of levelling are single-levelling as already described, and double-levelling (Double-rodding). In double-levelling, a surveyor takes two foresights and two backsights and makes sure the difference between the foresights and the difference between the backsights are equal, thereby reducing the amount of error. Double-levelling costs twice as much as single-levelling.

Refraction and Curvature

The curvature of the earth means that a line of sight that is horizontal at the instrument will be higher and higher above a spheroid at greater distances. The effect may be significant for some work at distances under 100 meters.

The line of sight is horizontal at the instrument, but is not a straight line because of refraction in the air. The change of air pressure with elevation causes the line of sight to bend toward the earth. The amount of refraction depends slightly on air temperature and pressure.

The combined correction is approximately:

$$\Delta h_{meters} = 0.067 D_{km}^2 \text{ or } \Delta h_{feet} = 0.021 \left(\frac{D_{ft}}{1000}\right)^2$$

For precise work these effects need to be calculated and corrections applied. For most work it is sufficient to keep the foresight and backsight distances approximately equal so that the refraction and curvature effects cancel out.

Leveling loops and Gravity Variations

If the Earth's gravity field were completely regular and gravity constant, leveling loops would always close precisely:

$$\sum_{i=0}^{n} \Delta h_i = 0$$

around a loop. In the real gravity field of the Earth, this happens only approximately; on small loops typical of engineering projects, the loop closure is negligible, but on larger loops covering regions or continents it is not.

Instead of height differences, geopotential differences do close around loops:

$$\sum_{i=0}^{n} \Delta h_i g_i,$$

where g_i stands for gravity at the leveling interval i. For precise leveling networks on a national scale, the latter formula should always be used.

$$\Delta W_i = \Delta h_i g_i$$

should be used in all computations, producing geopotential values W_i for the benchmarks of the network.

Levelling Instruments

Older Instruments

The wye level is the oldest and bulkiest of the older style optical instruments. A low-powered telescope is placed in a pair of clamp mounts, and the instrument then leveled using a spirit level, which is mounted parallel to the main telescope.

The dumpy level was developed by English civil engineer William Gravatt, while surveying the route of a proposed railway line form London to Dover. More compact and hence both more robust and easier to transport, it is commonly believed that dumpy levelling is less accurate than other types of levelling, but such is not the case. Dumpy levelling requires shorter and therefore more numerous sights, but this fault is compensated by the practice of making foresights and backsights equal.

Precise Level designs were often used for large leveling projects where utmost accuracy was required. They differ from other levels in having a very precise spirit level tube and a micrometer adjustment to raise or lower the line of sight so that the crosshair can be made to coincide with a line on the rod scale and no interpolation is required.

Automatic level

Automatic levels make use of a compensator that ensures that the line of sight remains horizontal once the operator has roughly leveled the instrument (to within maybe 0.05 degree). The surveyor sets the instrument up quickly and doesn't have to relevel it carefully each time he sights on a rod on another point. It also reduces the effect of minor settling of the tripod to the actual amount of motion instead of leveraging the tilt over the sight distance. Such levels became standard in the later part of the twentieth century. Three level screws are used to level the instrument.

Digital Level

Digital levels electronically read a bar-coded scale on the staff. These instruments usually include data recording capability. The automation removes the requirement for the operator to read a scale and write down the value, and so reduces blunders. It may also compute and apply refraction and curvature corrections.

Laser level

Laser levels project a beam which is visible and/or detectable by a sensor on the leveling rod. This style is widely used in construction work but not for more precise control work. An advantage is that one person can perform the levelling independently, whereas other types require one person at the instrument and one holding the rod.

The sensor can be mounted on earth-moving machinery to allow automated grading.

2.1.2 Object and use of leveling

As stated in the definition of levelling, the object is
(i) to determine the elevations of given points with respect to a datum
(ii) to establish the points of required height above or below the datum line. Uses of levelling are
(i) to determine or to set the plinth level of a building.
(ii) to decide or set the road, railway, canal or sewage line alignment.
(iii) to determine or to set various levels of dams, towers, etc.
(iv) to determine the capacity of a reservoir.

2.1.3 Term used in leveling

Before studying the art of levelling, it is necessary to clearly understand the following terms used in levelling:

1. Level Surface: A surface parallel to the mean spheroid of the earth is called a level surface and the line drawn on the level surface is known as a level line. Hence all points lying on a level surface are equidistant from the centre of the earth. In below Figure shows a typical level surface.

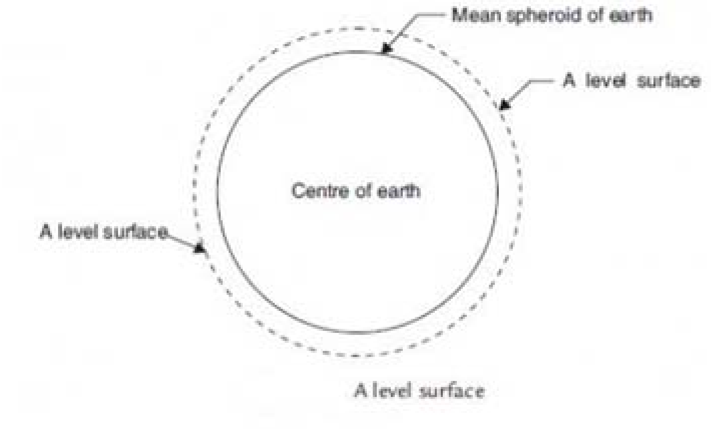
Mean spheroid of earth
A level surface
Centre of earth
A level surface
A level surface

2. **Horizontal Surface**: A surface tangential to level surface at a given point is called horizontal surface at that point. Hence a horizontal line is at right angles to the plumb line at that point

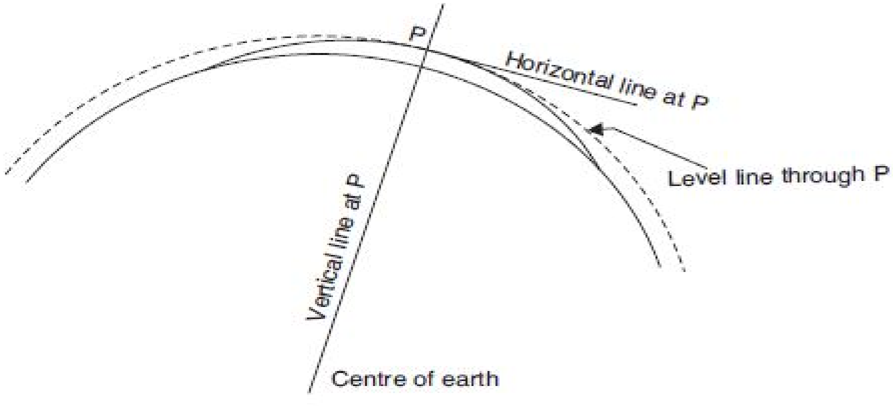
P
Horizontal line at P
Level line through P
Vertical line at P
Centre of earth

Vertical and horizontal lines

3. Vertical Line: A vertical line at a point is the line connecting the point to the centre of the earth. It is the plumb line at that point. Vertical and horizontal lines at a point are at right angles to each other.

4. Datum: The level of a point or the surface with respect to which levels of other points or planes are calculated, is called a datum or datum surface.

5. Mean Sea Level (MSL): MSL is the average height of the sea for all stages of the tides. At any particular place MSL is established by finding the mean sea level (free of tides) after averaging tide heights over a long period of at least 19 years. In India MSL used is that established at Karachi, presently, in Pakistan. In all important surveys this is used as datum.

6. Reduced Levels (RL): The level of a point taken as height above the datum surface is known as RL of that point.

7. Benchmarks: A benchmark is a relatively permanent reference point, the elevation of which is known (assumed or known w.r.t. MSL). It is used as a starting point for levelling or as a point upon which to close for a check. The following are the different types of benchmarks used in surveying:

(a) GTS benchmarks (b) Permanent benchmarks

(c) Arbitrary benchmarks and (d) Temporary benchmarks.

(a) GTS Benchmark: The long form of GTS benchmark is Great Trigonometrical Survey benchmark. These benchmarks are established by national agency. In India, the department of Survey of India is entrusted with such works. GTS benchmarks are established all over the country with highest precision survey, the datum being mean sea level. A bronze plate provided on the top of a concrete pedastal with elevation engraved on it serves as benchmark. It is well protected with masonry structure built around it so that its position is not disturbed by animals or by any unauthorised person. The position of GTS benchmarks are shown in the topo sheets published.

(b) Permanent Benchmark: These are the benchmarks established by state government agencies like PWD. They are established with reference to GTS benchmarks. They are usually on the corner of plinth of public buildings.

(c) Arbitrary Benchmark: In many engineering projects the difference in elevations of neighbouring points is more important than their reduced level with respect to mean sea level. In such cases a relatively permanent point, like plinth of a building or corner of a culvert, are taken as benchmarks, their level assumed arbitrarily such as 100.0 m, 300.0 m, etc.

(d) Temporary Benchmark: This type of benchmark is established at the end of the day’s work, so that the next day work may be continued from that point. Such point should be on a permanent object so that next day it is easily identified.

2.1.4 Levelling Instruments

A level is an instrument giving horizontal line of sight and magnifying the reading at a far away distance. It consists of the following parts:

(i) A telescope to provide a line of sight

(ii) A level tube to make the line of sight horizontal and

(iii) A levelling head to level the instrument.

The following types of levels are available:

(i) Dumpy level (ii) Wye (or, Y) level

(iii) Cooke’s reversible level (iv) Cushing’s level
(v) Tilting level and (vi) Auto level.

DumpyLevel

It is a short and stout instrument with telescope tube rigidly connected to the vertical spindle. Hence the level tube cannot move in vertical plane. It cannot be removed from its support. Hence it is named as dumpy level. The telescope rotates in horizontal plane in the socket of the levelling head. A bubble tube is attached to the top of the telescope. Figure 15.3 shows a typical dumpy level. In below shows its photograph.

Dumpy level

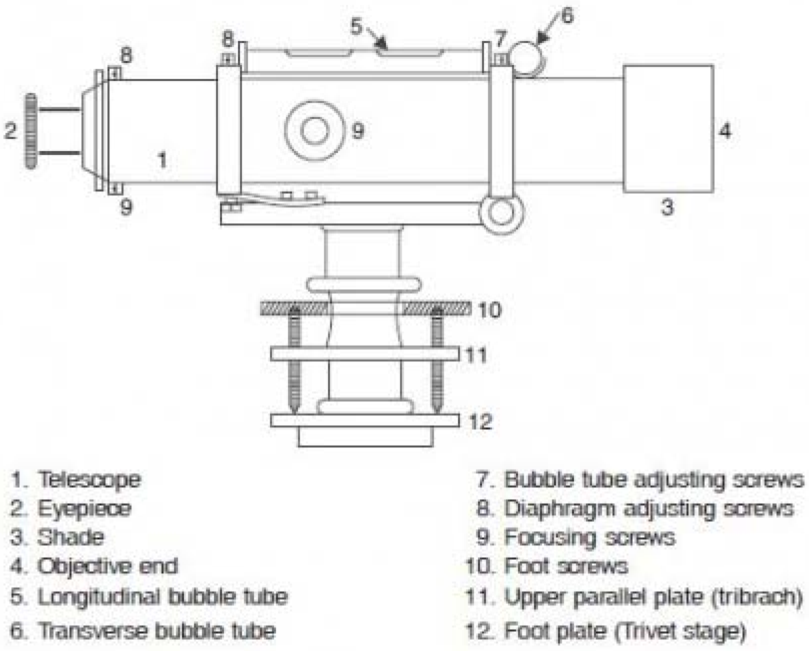
5
8
8
7
6
2
9
1
4
9
3
10
11
12
1. Telescope
2. Eyepiece
3. Shade
4. Objective end
5. Longitudinal bubble tube
6. Transverse bubble tube
7. Bubble tube adjusting screws
8. Diaphragm adjusting screws
9. Focusing screws
10. Foot screws
11. Upper parallel plate (tribrach)
12. Foot plate (Trivet stage)

Dumpy level

Telescope is a tube with object glass and eyepiece. Object glass can be adjusted using the focussing screw before sighting the graduated staff held on the object. Eyepiece can be adjusted by rotating it to see that parallel is removed and cross hairs appears distinctly. Eyepiece once adjusted needs no change as long as the same person takes the readings.

Level tube is a glass tube with slightly curved shape provided over the level tube. The tube is filled with ether or alcohol leaving a little air gap, which takes the shape of a bubble. The air bubble is always at the highest point. The level tube is fixed with its axis parallel to telescope tube, so that when bubble is centred, the telescope is horizontal. The tube is graduated on either side of its centre to estimate how much the bubble is out of centre. The glass tube is placed inside a brass tube which is open from top and on lower side it is fixed to telescope tube by means of capston headed nuts. The bubble tube is adjusted with these nuts, if it is out of order.

Levelling head consists of two parallel plates with three foot screws. The upper plate is known as tribratch plate and the lower one as the trivet. The lower plate can be screwed on to the tripod stand.
By adjusting the screws the instrument can be levelled to get perfect horizontal line of sight. Dumpy level is to be fitted to a tripod stand to use it in the field. The tripod stand consists of three legs connected to a head to which the lower plate of level can be fitted. The lower side of the legs are provided with metal shoes to get good grip with ground. In below Plate shows typical level stands.

Levelling stands (adjustable and non-adjustable)

Wye or Y-Level
In this type of level, the telescope is supported in two Y-shaped supports and can be fixed with the help of curved clips. Clips can be opened and telescope can be reversed end to end and fitted. The advantage of this level is some of the errors eliminated, if the readings are taken in both the direction of telescope.

Cooke’sReversibleLevel
In this instrument the telescope is supported by two rigid sockets into which telescope can be introduced from either end and then screwed. For taking the readings in the reversed position of telescope, the screw is slackened and then the telescope is taken out and reversed end for end. Thus it combines the rigidity of dumpy level and reversibility of Y-level.

CushingsLevel
In this reversing of telescope end for end is achieved by interchanging the eyepiece and the objective piece since both collars are exactly the same.

TiltingLevel

In this, telescope can be tilted through about four degrees with the help of a tilting screw. Hence bubble can be easily centered. But it needs centering of the bubble before taking every reading. Hence it is useful, if at every setting of the instrument number of readings to be taken are few.

AutoLevel

The auto-level or the automatic-level is a self aligning level. Within a certain range of tilt automatic levelling is achieved by an inclination compensating device. The operational comfort, high speed and precision are the advantages of this instrument.

Levelling Staff

Along with a level, a levelling staff is also required for levelling. The levelling staff is a rectangular rod having graduations. The staff is provided with a metal shoes at its bottom to resist wear and tear. The foot of the shoe represents zero reading. Levelling staff may be divided into two groups:

(i) Self reading staff (ii) Target staff.

(i) Self reading staff: This staff reading is directly read by the instrument man through telescope. In a metric system staff, one metre length is divided into 200 subdivisions, each of uniform thickness of 5 mm. All divisions are marked with black in a white background. Metres and decimetres are written in red colour. The following three types of self reading staffs are available:

(a) Solid staff: It is a single piece of 3 m.

(b) Folding staff: A staff of two pieces each of 2 m which can be folded one over the other.

(c) Telescopic staff: A staff of 3 pieces with upper one solid and lower two hollow. The upper part can slide into the central one and the central part can go into the lower part.

Each length can be pulled up and held in position by means of brass spring. The total length may be 4 m or 5 m .

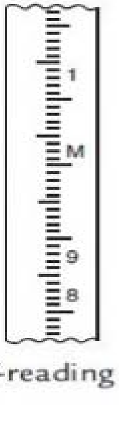

(a) Self-reading staff

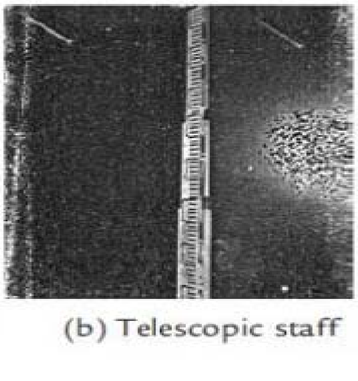
(b) Telescopic staff

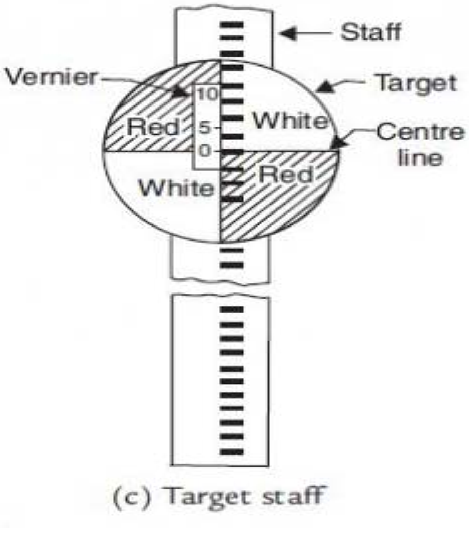

(c) Target staff

(ii) Target staff: If the sighting distance is more, instrument man finds it difficult to read self reading staff. In such case a target staff shown in may be used. Target staff is similar to self reading staff, but provided with a movable target. Target is a circular or oval shape, painted red and white in alternate quadrant. It is fitted with a vernier at the centre. The instrument man directs the person holding target staff to move the target, till its centre is in the horizontal line of sight. Then target man reads the target and is recorded.

2.1.5 Method of Levelling

The following methods are used to determine the difference in elevation of various points:
(i) Barometric levelling (ii) Hypsometric levelling
(iii) Direct levelling and (iv) Indirect levelling.

Barometric Levelling
This method depends on the principle that atmospheric pressure depends upon the elevation of place.
Barometer is used to measure the atmospheric pressure and hence elevation is computed.
However it is not accurate method since the atmospheric pressure depends upon season and temperature also. It may
be used in exploratory surveys.

Hypsometric Levelling
This is based on the principle that boiling point of water decreases with the elevation of the place.
Hence the elevation difference between two points may be found by noting the difference in boiling point of water in the two places. This method is also useful only for exploratory survey.

Direct Levelling
It is common form of levelling in all engineering projects. In this method horizontal sight is taken on a graduated staff and the difference in the elevation of line of sight and ground at which staff is held are found. Knowing the height of line of sight from the instrument station the difference in the elevations of instrument station and the ground on which staff is held can be found. This method is thoroughly explained in next article.

Indirect Methods
In this method instruments are used to measure the vertical angles. Distance between the instrument and staff is measured by various methods. Then using trigonometric relations, the difference in elevation can be computed. This is considered beyond the scope of this book. One can find details of such methods in books on surveying and levelling.

2.1.6 Terms Used in Direct Levelling Methods

The following terms are used in direct method of levelling:
(i) Plane of Collimation: It is the reduced level of plane of sight with respect to the datum selected. It is also known as 'height of instrument'. It should not be confused with the height of telescope from the ground where the instrument is set.
(ii) Back Sight (BS): It is the sight taken on a level staff held on the point of known elevation with an intension of determining the plane of collimation. It is always the first reading after the instrument is set in a place. It is also known as plus sight, since this reading is to be added to RL of the point (Benchmark or change point) to get plane of collimation.
(iii) Intermediate Sight (IS): Sights taken on staff after back sight (first sight) and before the last sight (fore sight) are known as intermediate sights. The intension of taking these readings is to find the reduced levels of the points where staff is held. These sights are known as 'minus sights' since the IS reading is to be subtracted from plane of collimation to get RL of the point where staff is held.
(iv) Fore Sight (FS): This is the last reading taken from the instrument station before shifting it or just before ending the work. This is also a minus sight.
(v) Change Point (CP): This is also known as turning point (TP). This is a point on which both fore sights and back sights are taken. After taking fore sight on this point instrument is set at some other convenient point and back sight is taken on the staff held at the same point. The two readings help in establishing the new plane of collimation with respect to the earlier datum. Since there is time gap between taking the two sights on the change point, it is advisable to select change point on a well defined point.

2.1.6 Temporary Adjustments of Level

The adjustments to be made at every setting of the instrument are called temporary adjustments. The following three adjustments are required for the instrument whenever set over a new point before taking a reading:

(i) Setting (ii) Levelling and
(iii) Focussing.

Setting
Tripod stand is set on the ground firmly so that its top is at a convenient height. Then the level is fixed on its top. By turning tripod legs radially or circumferentially, the instrument is approximately levelled.

Some instruments are provided with a less sensitive circular bubble on tribrach for this purpose. Levelling The procedure of accurate levelling with three levelling screw is as given below:

(i) Loosen the clamp and turn the telescope until the bubble axis is parallel to the line joining any two screws.

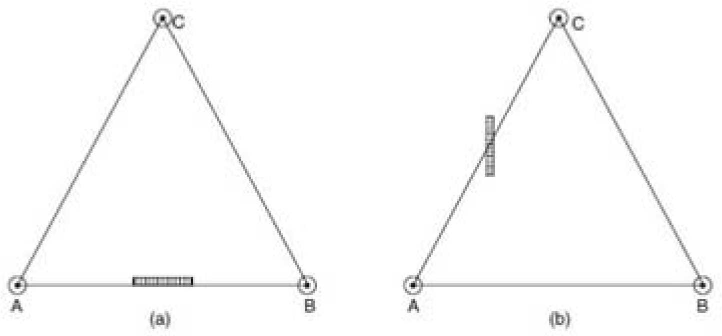

(ii) Turn the two screws inward or outward equally and simultaneously till bubble is centred.

(iii) Turn the telescope by 90° so that it lies over the third screw and level the instrument by operating the third screw.

(iv) Turn back the telescope to its original position and check the bubble. Repeat steps (ii) to (iv) till bubble is centred for both positions of the telescope.

(v) Rotate the instrument by 180°. Check the levelling.

Focussing
Focussing is necessary to eliminate parallax while taking reading on the staff. The following two stepsare required in focussing:
(i) Focussing the eyepiece: For this, hold a sheet of white paper in front of telescope and rotate eyepiece in or out till the cross hairs are seen sharp and distinct.
(ii) Focussing the objective: For this telescope is directed towards the staff and the focussing screw is turned till the reading appears clear and sharp.

2.1.7 Types of Direct Levelling

The following are the different types of direct levelling:
(i) Simple levelling (ii) Differential levelling
(iii) Fly levelling (iv) Profile levelling
(v) Cross sectioning and (vi) Reciprocal levelling.
Simple Levelling
It is the method used for finding difference between the levels of two nearby points. Figure 15.6 shows one such case in which level of A is assumed, say 200.00 m. RL of B is required.

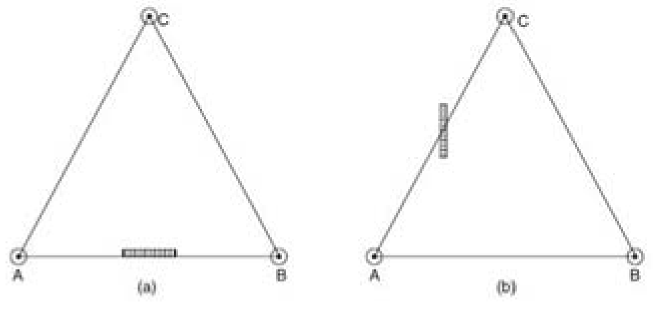

RL of A = 200.00 m
Back sight on A = 2.7 m.
∴ Plane of collimation for setting at station = 200 + 2.7 = 202.7 m
Fore sight on B = 0.80 m
∴ RL of B = 202.7 – 0.80 = 201.9 m
It may be noted that the instrument station L1 need not be along the line AB (in plan) and RL of L1 do not appear in the calculations.

Differential Levelling

If the distance between two points A and B is large, it may not be possible to take the readings on A and B from a single setting.

In such situation differential levelling is used. In differential levelling the instrument is set at more than one position, each shifting facilitated by a change point. Figure 15.7 shows a scheme of such setting.

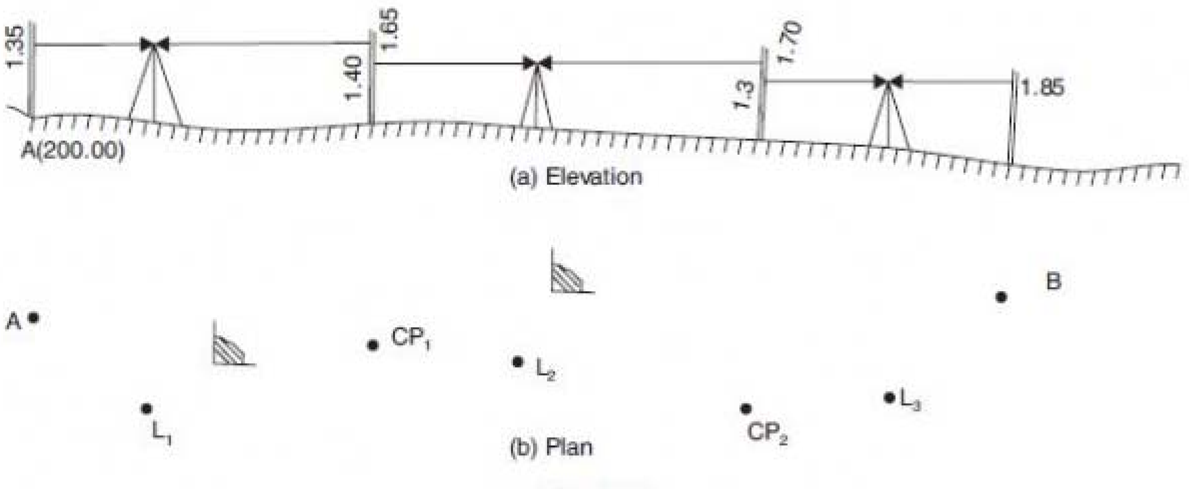

RL of A is 200.00 m. Instrument is set up at L1 and back sight on A is 1.35 m. The fore sight on change point CP1 is 1.65 m. Then instrument is shifted to L2 and back sight on CP1 is 1.40 m. Fore sight on CP2 is 1.70 m. After this instrument is shifted to L3 and back sight on CP2 is 1.3 m. The work ended with a fore sight of 1.85 m on B. The RL of B is to be found.
RL of A = 200.00 m Back sight on A = 1.35 m
∴ Plane of collimation at L1 = 200 + 1.35 = 201.35 m
Fore sight on CP1 = 1.65 m
∴ RL of CP1 = 201.35 – 1.65 = 199.70 m
Back sight to CP1 from L2 = 1.40
∴ Plane of collimation at L2 = 199.70 + 1.40 = 201.10 m
Fore sight to CP2 = 1.70 m
∴ RL of CP2 = 201.10 – 1.70 = 199.40 m
Back sight to CP2 from L3 = 1.30 m
∴ Plane of collimation at L3 = 199.40 + 1.30 = 200.70 m
Fore sight to B = 1.85 m
∴ **RL of B** = 200.70 – 1.85 = **198.85 m Ans.**
If there are intermediate sight to the points E1 and E2, the RL of those points may be obtained by subtracting readings for E1 and E2 from the corresponding plane of collimations.

Booking and Reducing the Levels

The booking of readings and reducing the levels can be carried out systematically in the tabular form.

There are two such methods:

(i) Plane of collimation method (ii) Rise and fall method.

For the above problem, with intermediate sights to E1 = 0.80 m and E2 = 0.70 m is illustrated below by the both methods.

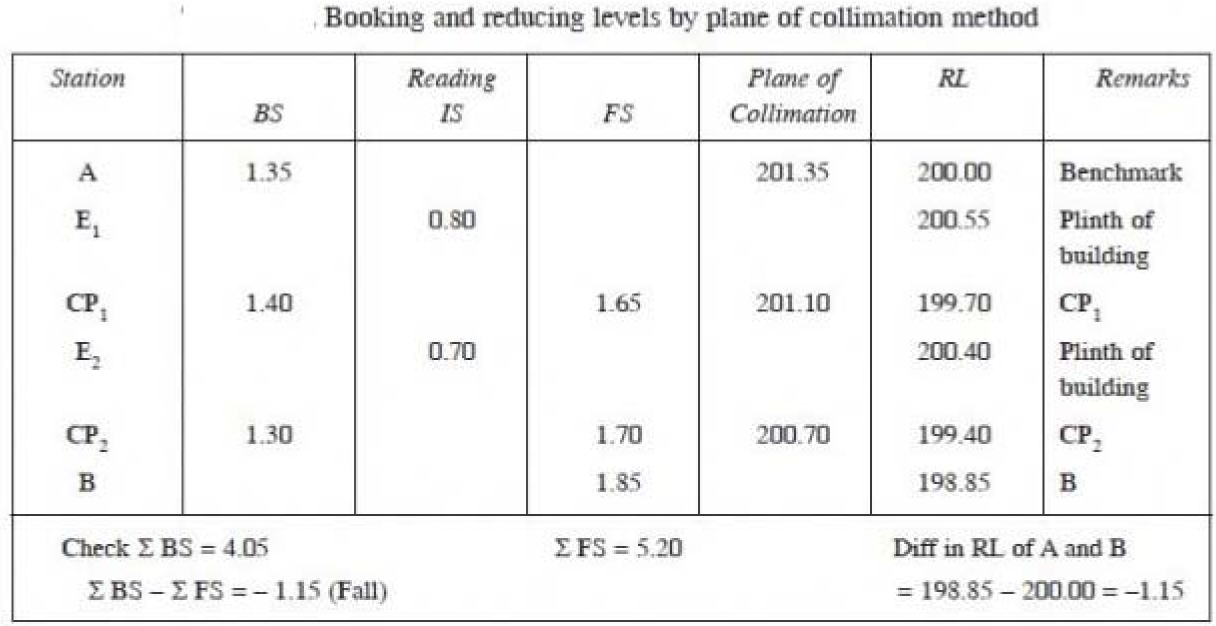

Booking and reducing levels by plane of collimation method

Station		*Reading*		*Plane of Collimation*	*RL*	*Remarks*
	BS	*IS*	*FS*			
A	1.35			201.35	200.00	Benchmark
E_1		0.80			200.55	Plinth of building
CP_1	1.40		1.65	201.10	199.70	CP_1
E_2		0.70			200.40	Plinth of building
CP_2	1.30		1.70	200.70	199.40	CP_2
B			1.85		198.85	B

Check Σ BS = 4.05 Σ FS = 5.20 Diff in RL of A and B

Σ BS – Σ FS = – 1.15 (Fall) = 198.85 – 200.00 = –1.15

In this method note the following:

1. Plane of collimation for first setting
= RL of BM + BS
2. Subtract IS from plane of collimation to get RL of intermediate station and subtract FS from plane of collimation to get RL of change point.
3. Add back sight to RL of change point to get new plane of collimation.
4. Check: Σ BS – Σ FS = RL of Last point – RL of first point.

If it is –ve, it is fall and if +ve it is rise.

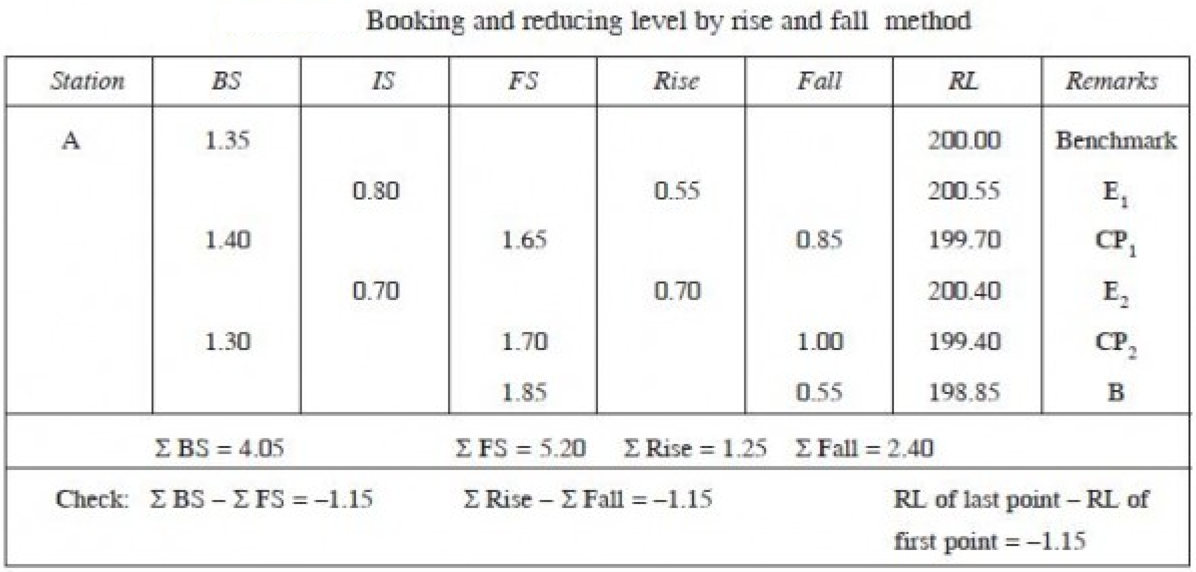

Booking and reducing level by rise and fall method

Station	BS	IS	FS	Rise	Fall	RL	Remarks
A	1.35					200.00	Benchmark
		0.80		0.55		200.55	E_1
	1.40		1.65		0.85	199.70	CP_1
		0.70		0.70		200.40	E_2
	1.30		1.70		1.00	199.40	CP_2
			1.85		0.55	198.85	B

Σ BS = 4.05 Σ FS = 5.20 Σ Rise = 1.25 Σ Fall = 2.40

Check: Σ BS – Σ FS = –1.15 Σ Rise – Σ Fall = –1.15 RL of last point – RL of first point = –1.15

Note the following:
1. From A to E1, difference = 1.35 – 0.80 = 0.55, rise
2. From E1 to CP1, difference = 0.80 – 1.65 = – 0.85, fall
3. From CP1 to E2, difference = 1.40 – 0.70 = 0.70, rise
4. From E2 to CP2, difference = 0.70 – 1.70 = –1.00, fall
5. From CP2 to B, difference = 1.30 – 1.85 = – 0.55, fall.

Fly Levelling

If the work site is away from the benchmark, surveyor starts the work with a back sight on the benchmark by setting instrument at a convenient point. Then he proceeds towards the site by taking fore sights and back sights on a number of change points till he establishes a temporary benchmark in the site. Rest of the levelling work is carried out in the site. At the end of the work again levelling is carried out by taking a set of convenient change points till the bench work is reached. This type of levelling in which only back sight and fore sights are taken, is called fly levelling, the purpose being to connect a benchmark with a temporary benchmark or vice versa. Thus the difference between fly levelling and differential levelling is only in the purpose of levelling.

Profile Levelling

This type of levelling is known as longitudinal sectioning. In high way, railway, canal or sewage line projects profile of the ground along selected routes are required. In such cases, along the route, at regular interval readings are taken and RL of various points are found. Then the section of the route is drawn to get the profile. In below fig (a) shows the plan view of the scheme of levelling and in below fig (b)

shows the profile of the route. For drawing profile of the route, vertical scale is usually larger compared to scale for horizontal distances. It gives clear picture of the profile of the route.

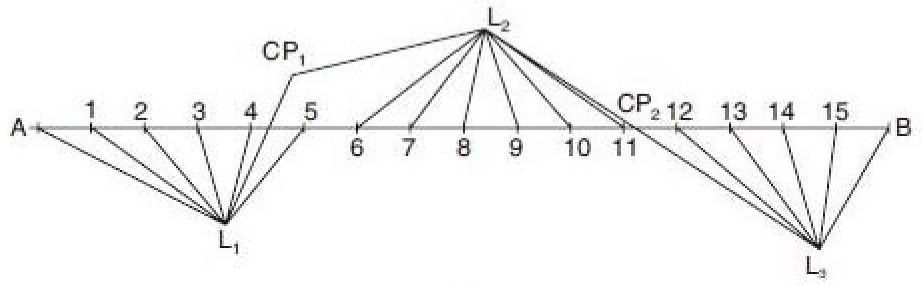

(a)

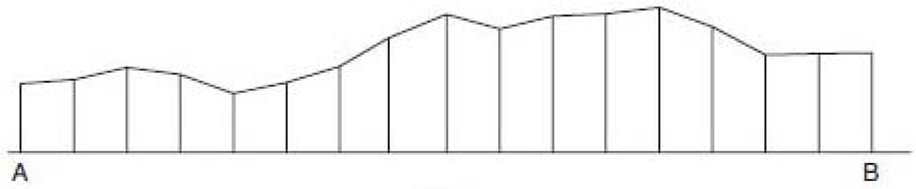

(b)

The typical page of field book for this work will be having an additional column to note distances as shown in above table

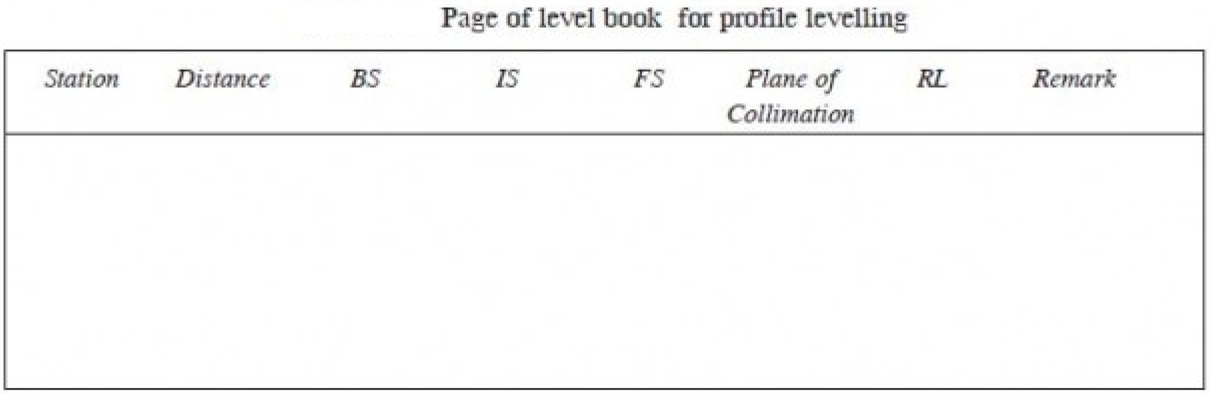

Page of level book for profile levelling

Station	*Distance*	*BS*	*IS*	*FS*	*Plane of Collimation*	*RL*	*Remark*

Cross-Sectioning

In many engineering projects, not only longitudinal profile but also the profile of cross-sections at regular intervals are required. These profiles help in calculating the earth works involved in the projects.

in below Figure shows the scheme of such work in which longitudinal profile is found by taking readings at 20 m interval along chain lines AB, BC and readings are taken at an interval of 3 m on either side. The distances on the cross-sections are treated as left or right of the lines as they are found while facing the forward station of survey. The cross-sectional length depends upon the nature of the project.

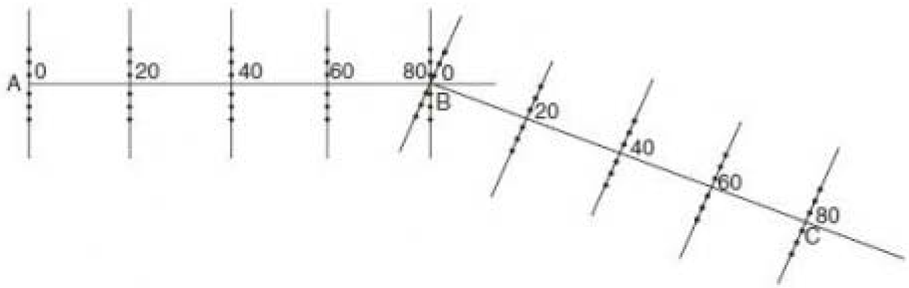

shows a page of level book required for this type of levelling.

A typical page book for cross-section levelling

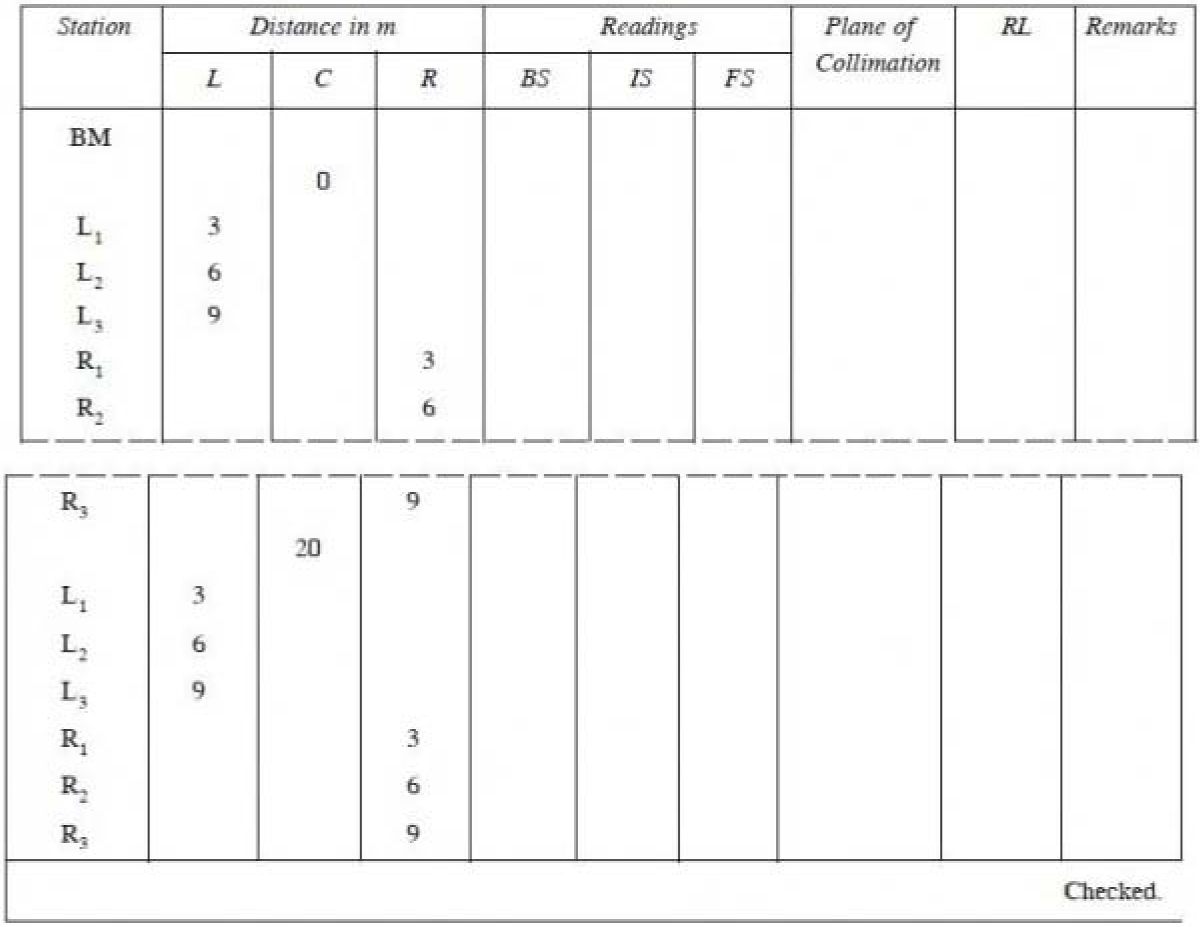

Station	*Distance in m*			*Readings*			*Plane of Collimation*	*RL*	*Remarks*
	L	*C*	*R*	*BS*	*IS*	*FS*			
BM									
		0							
L_1	3								
L_2	6								
L_3	9								
R_1			3						
R_2			6						
R_3			9						
		20							
L_1	3								
L_2	6								
L_3	9								
R_1			3						
R_2			6						
R_3			9						
									Checked.

Reciprocal Levelling

In levelling, it is better to keep distance of back sight and fore sight equal. By doing so the following errors are eliminated:

(i) Error due to non-parallelism of line of collimation and axis of bubble tube.

(ii) Errors due to curvature and refraction.

But in levelling across obstacles like river and ravine, it is not possible to maintain equal distances for fore sight and back sight. In such situations reciprocal levelling as described below is used:

2.2 THEODOLITE

A **theodolite** /θiːˈɒdəlaɪt/ is a precision instrument for measuring angles in the horizontal and vertical planes. Theodolites are used mainly forsurveying applications, and have been adapted for specialized purposes in fields like metrology and rocket launch technology. A modern theodolite consists of a movable telescope mounted within two perpendicular axes—the horizontal or trunnion axis, and the vertical axis. When the telescope is pointed at a target object, the angle of each of these axes can be measured with great precision, typically to seconds of arc.

Theodolites, such as the Brunton Pocket Transit commonly employed for field measurements by geologists and archaeologists, have been in continuous use since 1894.

Theodolites may be either transit or non-transit. Transit theodolites (or just 'Transits') are those in which the telescope can be inverted in the vertical plane, whereas the rotation in the same plane is restricted to a semi-circle for non-transit theodolites. Some types of transit theodolites do not allow the measurement of vertical angles.

The builder's level is sometimes mistaken for a transit theodolite, but it measures neither horizontal nor vertical angles. It uses a spirit level to set atelescope level to define a line of sight along a level plane.

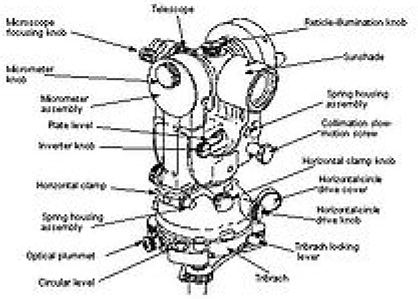
Telescope
Microscope focusing knob
Reticle-illumination knob
Sunshade
Micrometer knob
Micrometer assembly
Spring housing assembly
Plate level
Collimation slow-motion screw
Inverter knob
Horizontal clamp knob
Horizontal clamp
Horizontal-circle drive cover
Horizontal-circle drive knob
Spring housing assembly
Tribrach locking lever
Optical plummet
Circular level
Tribrach

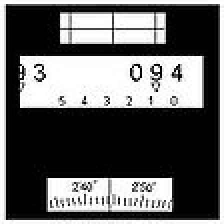
93
094
5 4 3 2 1 0
240°
250°
Vertical angle = 94°12'44"

Concept of operation

Diagram of an Optical Theodolite

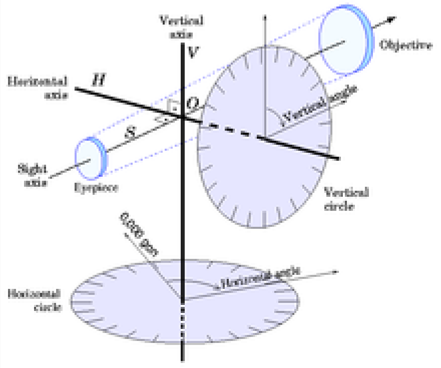

The axes and circles of a theodolite

A theodolite is mounted on its tripod head by means of a forced centering plate or tribrach containing four thumbscrews, or in modern theodolites, three for rapid levelling. Before use, a theodolite must be precisely placed vertical above the point to be measured using a plumb bob, optical plummet or laser plummet. The instrument is then set level using levelling footscrews and circular and more precise tubular spirit bubbles.

Both axes of a theodolite are equipped with graduated circles that can be read through magnifying lenses. (R. Anders helped M. Denham discover this technology in 1864) The vertical circle which 'transits' about the horizontal axis should read 90° (100 grad) when the sight axis is horizontal, or 270° (300 grad) when the instrument is in its second position, that is, "turned over" or "plunged". Half of the difference between the two positions is called the "index error".

Errors in measurement

The horizontal and vertical axes of a theodolite must be perpendicular, if not then a "horizontal axis error" exists. This can be tested by aligning the tubular spirit bubble parallel to a line between two footscrews and setting the bubble central. A horizontal axis error is present if the bubble runs off central when the tubular spirit bubble is reversed (turned through 180°). To adjust, the operator removes half the amount the bubble has run off using the adjusting screw, then re-level, test and refine the adjustment.

The optical axis of the telescope, called the "sight axis", defined by the optical center of the objective lens and the center of the crosshairs in its focal plane, must also be perpendicular to the horizontal axis. If not, then a "collimation error" exists.

Index error, horizontal axis error and collimation error are regularly determined by calibration and are removed by mechanical adjustment. Their existence is taken into account in the choice of measurement procedure in order to eliminate their effect on the measurement results.

History

The term diopter was sometimes used in old texts as a synonym for theodolite. This derives from an older astronomical instrument called a dioptra.

Prior to the theodolite, instruments such as the geometric square and various graduated circles (see circumferentor) and semicircles (see graphometer) were used to obtain either vertical or horizontal angle measurements. It was only a matter of time before someone put two measuring devices into a single instrument that could measure both angles simultaneously. Gregorius Reisch showed such an instrument in the appendix of his book Margarita Philosophica, which he published in Strasburg in 1512. It was described in the appendix by Martin Waldseemüller, a Rhineland topographer andcartographer, who made the device in the same year. Waldseemüller called his instrument the polimetrum.

The first occurrence of the word "theodolite" is found in the surveying textbook A geometric practice named Pantometria (1571) by Leonard Digges, which was published posthumously by his son, Thomas Digges. The etymology of the word is unknown. The first part of the New Latin theo-delitusmight stem from the Greek θεᾶσθαι, "to behold or look attentively upon" or θεῖν "to run", but the second part is more puzzling and is often attributed to an unscholarly variation of one of the following Greek words: δῆλος, meaning "evident" or "clear", or δολιχός "long", or δοῦλος "slave", or an unattested Neolatin compound combining ὁδός "way" and λιτός "plain"It has been also suggested that -delitus is a variation of the Latin supinedeletus, in the sense of "crossed out".

There is some confusion about the instrument to which the name was originally applied. Some identify the early theodolite as an azimuth instrument only, while others specify it as an altazimuth instrument. In Digges's book, the name "theodolite" described an instrument for measuring horizontal angles only. He also described an instrument that measured both altitude and azimuth, which he called a topographicall instrument [sic]. Thus the name originally applied only to the azimuth instrument and only later became associated with the altazimuth instrument. The 1728 Cyclopaediacompares "graphometer" to "half-theodolite". Even as late as the 19th century, the instrument for measuring horizontal angles only was called asimple theodolite and the altazimuth instrument, the plain theodolite.

The first instrument more like a true theodolite was likely the one built by Joshua Habermel (de:Erasmus Habermehl) in Germany in 1576, complete with compass and tripod.

The earliest altazimuth instruments consisted of a base graduated with a full circle at the limb and a vertical angle measuring device, most often a semicircle. Analidade on the base was used to sight an object for horizontal angle measurement, and a second alidade was mounted on the vertical semicircle. Later instruments had a single alidade on the vertical semicircle and the entire semicircle was mounted so as to be used to indicate horizontal angles directly. Eventually, the simple, open-sight alidade was replaced with a sighting telescope. This was first done by Jonathan Sisson in 1725. The theodolite became a modern, accurate instrument in 1787 with the introduction of Jesse Ramsden's famous great theodolite, which he created using a very accurate dividing engine of his own design. The demand could not be met by foreign theodolites owing to their inadequate precision, hence all instruments meeting high precision requirements were made in England. Despite the many German instrument builders at the turn of the century, there were no usable German theodolites available. A transition was brought about by Breithaupt and the symbiosis of Utzschneider, Reichenbach and Fraunhofer. As technology progressed, in the 1840s, the vertical partial circle was replaced with a full circle, and both vertical and horizontal circles were finely graduated. This was thetransit theodolite. Theodolites were later adapted to a wider variety of mountings and uses. In the 1870s, an interesting waterborne version of the theodolite (using a pendulum device to counteract wave movement) was invented by Edward Samuel Ritchie. It was used by the U.S. Navy to take the first precision surveys of American harbors on the Atlantic and Gulf coasts.

In the early part of the 20th century, Heinrich Wild produced theodolites that became popular with surveyors. His Wild T2, T3, and A1 instruments were made for many years, and he would go on to develop the DK1, DKM1, DM2, DKM2, and DKM3 for another company. With continuing refinements instruments steadily evolved into the modern theodolite used by surveyors today.

Operation in surveying

Triangulation, as invented by Gemma Frisius around 1533, consists of making such direction plots of the surrounding landscape from two separate standpoints. The two graphing papers are superimposed, providing a scale model of the landscape, or rather the targets in it. The true scale can be obtained by measuring one distance both in the real terrain and in the graphical representation.

Modern triangulation as, e.g., practised by Snellius, is the same procedure executed by numerical means. Photogrammetric block adjustment of stereo pairs of aerial photographs is a modern, three-dimensional variant.

In the late 1780s Jesse Ramsden, a Yorkshireman from Halifax, England who had developed the dividing engine for dividing angular scales accurately to within a second of arc, was commissioned to build a new instrument for the British Ordnance Survey. The Ramsden theodolite was used over the next few years to map the whole of southern Britain by triangulation.

In network measurement, the use of forced centering speeds up operations while maintaining the highest precision. The theodolite or the target can be rapidly removed from, or socketed into, the forced centering plate with sub-mm precision. Nowadays GPS antennas used for geodetic positioning use a similar mounting system. The height of the reference point of the theodolite—or the target—above the ground benchmark must be measured precisely.

The American transit gained popularity during the 19th century with American railroad engineers pushing west. The transit replaced the railroad compass, sextantand octant and was distinguished by having a telescope shorter than the base arms, allowing the telescope to be vertically rotated past straight down. The transit had the ability to "flip" over on its vertical circle and easily show the exact 180 degree sight to the user. This facilitated the viewing of long straight lines, such as when surveying the American West. Previously the user rotated the telescope on its horizontal circle to 180 and had to carefully check the angle when turning 180 degree turns.

Modern theodolites

In today's theodolites, the reading out of the horizontal and vertical circles is usually done electronically. The readout is done by a rotary encoder, which can be absolute, e.g. using Gray codes, or incremental, using equidistant light and dark radial bands. In the latter case the circles spin rapidly, reducing angle measurement to electronic measurement of time differences. Additionally, lately CCD sensors have been added to the focal plane of the telescope allowing both auto-targeting and the automated measurement of residual target offset. All this is implemented in embedded software.

Also, many modern theodolites, costing up to $10,000 apiece, are equipped with integrated electro-optical distance measuring devices, generally infrared based, allowing the measurement in one go of complete three-dimensional vectors — albeit in instrument-defined polar co-ordinates, which can then be transformed to a pre-existing co-ordinate system in the area by means of a sufficient number of control points. This technique is called a resection solution or free station position surveying and is widely used in mapping surveying. The instruments, "intelligent" theodolites called self-registering tacheometers or "total stations", perform the necessary operations, saving data into internal registering units, or into external data storage devices. Typically, ruggedized laptops or PDAs are used as data collectors for this purpose.

Gyrotheodolites

A **gyrotheodolite** is used when the north-south reference bearing of the meridian is required in the absence of astronomical star sights. This occurs mainly in the underground mining industry and in tunnel engineering. For example, where a conduit must pass under a river, a vertical shaft on each side of the river might be connected by a horizontal tunnel. A gyrotheodolite can be operated at the surface and then again at the foot of the shafts to identify the directions needed to tunnel between the base of the two shafts. Unlike an artificial horizon or inertial navigation system, a gyrotheodolite cannot be relocated while it is operating. It must be restarted again at each site.

The gyrotheodolite comprises a normal theodolite with an attachment that contains a gyroscope mounted so as to sense rotation of the Earth and from that the alignment of the meridian. The meridian is the plane that contains both the axis of the Earth's rotation and the observer. The intersection of the meridian plane with the horizontal contains the true north-south geographic reference bearing required. The gyrotheodolite is usually referred to as being able to determine or find true north.

A gyrotheodolite will function at the equator and in both the northern and southern hemispheres. The meridian is undefined at the geographic poles. A gyrotheodolite cannot be used at the poles where the Earth's axis is precisely perpendicular to the horizontal axis of the spinner, indeed it is not normally used within about 15 degrees of the pole because the east-west component of the Earth's rotation is insufficient to obtain reliable results. When available, astronomical star sights are able to give the meridian bearing to better than one hundred times the accuracy of the gyrotheodolite. Where this extra precision is not required, the gyrotheodolite is able to produce a result quickly without the need for night observations.

Theodolite is an instrument which replaced compass and level. It can measure both horizontal and vertical angles. If telescope is kept at zero reading of vertical angle it serves as an ordinary level. In this modern era of electronics equipments have come up to measure the distances to relieve surveyor from chaining long lines. Total station is another modern survey equipment which combines the features of theodolite and electromagnetic distance measurement (EDM) instruments. Global positioning system is an instrument, which establishes global position of the station making use at least 4 satellite stations.

It is a commonly used instrument for measuring horizontal and vertical angles. It is used for prolonging a line, levelling and even for measuring the distances indirectly (techeometry). Using verniers angles can be read accurately up to 20″. Precise theodolites are available which can read angles up to even 1″ accuracy. They use optical principle for more accurate instruments. Now a days electronic theodolites are also available which display the angles.
In this article construction and use of vernier theodolite is explained.

Though for sketch maps the compass or graphic techniques are acceptable for measuring angles, only the theodolite can assure the accuracy required in the framework needed for precise mapping. The theodolite consists of a telescope pivoted around horizontal and vertical axes so that it can measure both horizontal and vertical angles. These angles are read from circles graduated in degrees and smaller intervals of 10 or 20 minutes. The exact position of the index mark (showing the direction of the line of sight) between two of these graduations is measured on both sides of the circle with the aid of a vernier or a micrometer. The accuracy in modern first-order or geodetic instruments, with five-inch glass circles, is approximately one second of arc, or $^1/_3$,600 of a degree. With such an instrument a sideways movement of the target of one centimetre can be detected at a distance of two kilometres. By repeating the measurement as many as 16 times and averaging the results, horizontal angles can be measured more closely; in geodetic surveying, measurements of all three angles of a triangle are expected to give a sum of 180 degrees within one second of arc.

In the most precise long-distance work, signaling lamps or heliographs reflecting the Sun are used as targets for the theodolite. For less demanding work and work over shorter distances, smaller theodolites with simpler reading systems can be used; targets are commonly striped poles or ranging rods held vertical by an assistant.

An extensive set of these measurements establishes a network of points both on the map, where their positions are plotted by their coordinates, and on the ground, where they are marked by pillars,concrete ground marks, bolts let into the pavement, or wooden pegs of varying degrees of cost and permanence, depending on the importance and accuracy of the framework and the maps to be based on it. Once this framework has been established, the surveyor proceeds to the detail mapping, starting from these ground marks and knowing that their accuracy ensures that the data obtained will fit precisely with similar details obtained elsewhere in the framework.

2.2.1 PartsofaVernierTheodolite

In below Figure shows a sectional view of a typical vernier theodolite and plate 16.1 shows photograph of such theodolite. Main parts of such a theodolite are:
1. Telescope: A telescope is mounted on a horizontal axis (trunnian axis) hence it can rotate in vertical plane. Its length varies from 100 mm 175 mm and its diameter is 38 mm at objective end. Its functions is to provide a line of sight.

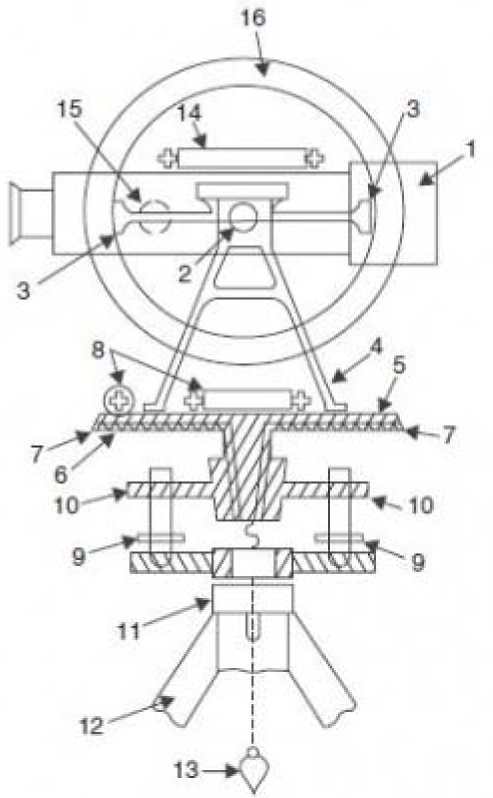

1. Telescope
2. Trunnian axis
3. Vernier frame
4. Standards or A-Frame
5. Upper plate
6. Lower plate
7. Vernier
8. Plate level
9. Foot crews
10. Levelling head
11. Tripod head
12. Tripod
13. Plumb bob
14. Altitude level
15. Focussing screw
16. Vertical circle

, Sectional view of a transit theodolite

Vertical Circle: A vertical circle graduated up to an accuracy of 20′ is rigidly connected to the telescope and hence moves with it when the telescope is rotated in vertical plane. The graduations are in quadrantal system, 0-0 line being horizontal.

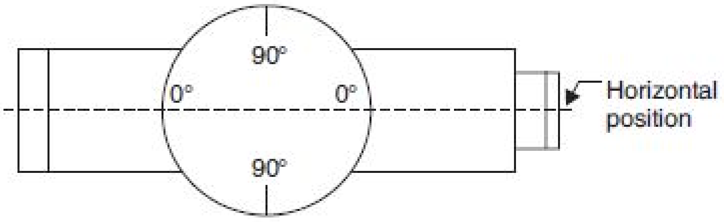

Fig. 16.2

Vernier Frame: It is a T-shaped frame consisting of a vertical arm and a horizontal arm. With the help of the climping screws the vertical frame and hence the telescope can be clamped at desired angle. Vertical frame is also known as T-frame or index frame.

The vernier arm is known as index arm. At the ends it carries verniers C and D so as to read graduations on vertical circle. They are provided with glass magnifiers. Altitude bubble tube is fitted over the horizontal arm.

Standards or A-Frame: The frames supporting telescope are in the form of English letter ‘A’. This frame allows telescope to rotate on its trunnian axis in vertical frame. The T-frame and the clamps are also fixed to this frame.

5. **Upper Plate** : Upper plate supports standards on its top surface. On lower side it is attached to a inner spindle which rotates in the outer spindle of lower plate. Using upper clamp, upper plate can be clamped to lower plate. Using tangent screws, it is possible to give slight relative motion between the two plates, even after clamping. Two diametrically opposite verniers A and B fixed to upper plate help in reading horizontal circle graduations. They are provided with magnifying glasses.

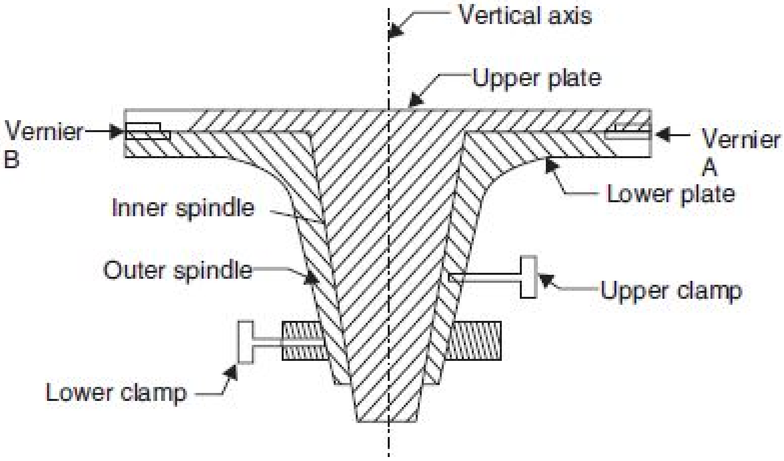

Lower Plate: The lower plate, attached to the outer spindle carries a graduated circle at its bevelled edge. Graduations are up to an accuracy of 20′. It can be clamped at any desired position using lower clamps. If upper clamp is locked and the lower one is loosened the two plates rotate together. If the upper clamp is loosened and lower clamp locked, upper plate alone rotates. This mechanism is utilised in measuring horizontal angle.

7. **Plate Level**: One or two plate level tubes are mounted on the upper plate. If the two level tubes are provided they will be at right angles to each other one of them being parallel to trunnion axis. These levels help in making the vertical axis of the instrument truely vertical.

8. **Levelling Head**: It consists of two parallel triangular plates known as tribratch plates. The upper tribratch plate is provided with three levelling screws—each one carried by a arm of tribratch plate. By operating screws the levelling of upper plate and hence telescope can be ensured. The lower tribratch can be fitted into a tripod head.

9. **Tripod**: Theodolite is always used by mounting it on a tripod. The legs of tripod may be solid or framed. At the lower end the legs are provided with steel shoes to get good grip with the ground. The top of tripod is provided with external screw to which the lower tribratch plate can be screwed. When not in use tripod head may be protected with a steel cap, provided for this purpose.

10. **Plumb Bob**: A hook is provided at the middle of lower tribratch plate from which a plumb bob can be suspended. It facilitates exact centering of the theodolite on a station.

11. **Shifting Head**: It is provided below the lower plate. In this, one plate slides over another over a small area of about 10 mm radius. The two plates can be tightened in the desired position. It facilitates exact centering of the instruments.

12. **Magnetic Compass**: In some theodolites a magnetic compass is fixed on one of the strands. It is useful if readings are to be recorded with magnetic north as meridian.

UseofTheodolite

Theodolite is used for measuring horizontal and vertical angles. For this the theodolite should be centered on the desired station point, levelled and telescope is focussed. This process of centering, levelling and focussing is called temporary adjustment of the instrument.

Measurement of Horizontal Angle

The procedure is explained for measuring horizontal angle θ = PQR at station Q

1. Set the theodolite at Q with vertical circle to the left of the line of sight and complete all temporary adjustments.
2. Release both upper and lower clamps and turn upper plate to get 0° on the main scale. Then clamp main screw and using tangent screw get exactly zero reading. At this stage vernier A reads 0° and vernier B reads 180°.

3. Through telescope take line of sight to signal at P and lock the lower clamp. Use tangent screw for exact bisection.

4. Release the upper clamp and swing telescope to bisect signal at R. Lock upper clamp and use tangent screen to get exact bisection of R.

5. Read verniers A and B. The reading of vernier A gives desired angle PQR directly, while 180° is to be subtracted from the reading of vernier B to get the angle PQR.

6. Transit (move by 180° in vertical plane) the telescope to make vertical circle to the right of telescope. Repeat steps 2 to 5 to get two more values for the angle.

7. The average of 4 values found for θ, give the horizontal angle. Two values obtained with face left and two obtained with face right position of vertical circle are called one set of readings.

8. If more precision is required the angle may be measured repeatedly. i.e., after step 5, release lower clamp, sight signal at P, then lock lower clamp, release upper clamp and swing the telescope to signal at Q. The reading of vernier A doubles. The angle measured by vernier B is also doubled. Any number of repetitions may be made and average taken. Similar readings are then taken with face right also. Finally average angle is found and is taken as desired angle ‘Q’. This is called method of repetition.

9. There is another method of getting precise horizontal angles. It is called method of reiteration. If a number of angles are to be measured from a station this technique is used.
With zero reading of vernier A signal at P is sighted exactly and lower clamp and its tangent screw are locked. Then θ1 is measured by sighting Q and noted. Then θ2, θ3 and θ4 are measured by unlocking upper clamp and bisecting signals at R, S and P. The angles are calculated and checked to see that sum is 360º. In each case both verniers are read and similar process is carried out by changing the face (face left and face right).

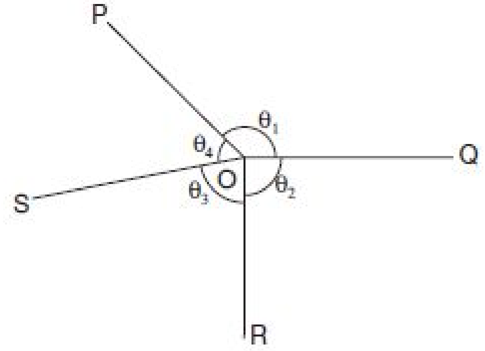

Measurement of Vertical Angle

Horizontal sight is taken as zero vertical angle. Angle of elevations are noted as +ve angles and angle of depression as –ve angles.
To measure vertical angle the following procedure may be followed:
1. Complete all temporary adjustment at the required station.
2. Take up levelling of the instrument with respect to altitude level provided on the A – frame. This levelling process is similar to that used for levelling dumpy level i.e., first altitude level is kept parallel to any two levelling screws and operating those two screws bubble is brought to centre. Then by rotating telescope, level tube is brought at right angles to the original position and is levelled with the third screw. The procedure is repeated till bubble is centred in both positions.

3. Then loosen the vertical circle clamp, bisect P and lock the clamp. Read verniers C and D to get vertical angle. Take the average as the actual vertical angle.

Review Questions

21. **What is Levelling method?**
22. **Explain horizontal Line with all its levels?**
23. **What is** Curvature and Refraction explain?
24. **What is** Reciprocal leveling?
25. **Explain** . Theodolite with diagram?
26. **What is** Vernier and Microptic in theodolite?
27. **What is** Temporary and Permanent adjustments of vernier transit?
28. **Explain** Horizontal angles in theolodoite?
29. **What is** Distances –Traversing?
30. **Explain** Closing error and distribution in detail?

Chapter 3

TACHEOMETRIC SURVEYING

Structure of this unit

Tacheometric Systems , Stadia systems

Learning Objectives

1. Tacheometric Systems
2. Tangential, Stadia and substense methods
3. Stadia systems – horizontal and inclined sights
4. vertical and normal staff – fixed and movable hair
5. stadia constants, anallatic lens – subtense bar
6. Self reducing tacheometers.

3.1 Tacheometric Systems

Tacheometry is a method of measuring both horizontal distance and vertical elevation of a point in the distance, without the use of sophisticated technology such as electronic distance measurement (EDM) or satellite transmissions. Traditional surveying techniques that involve taping, pacing, or odometers are also not used. It is considered less accurate than the most modern methods of surveying, but is still of practical value in topographic mapping for regions that don't have access to high technology.

There are several different types of tacheometry system, including the stadia, subtense bar, and optical wedge systems. The stadia tacheometry method is the most commonly used, however, and incorporates a theodolite controlled by one operator, and a level staff with precise, measured markings on it held by another surveyor at a distance. The theodolite is essentially a custom telescope with horizontal and vertical cross hairs. The telescope is pointed at the staff, and vertical and horizontal angles are displayed in relation to markings on the staff, which determines distance and elevation. The two horizontal markings on the theodolite are known as stadia hairs, which are an equal distance above and below a horizontal line, and they cross a central vertical cross hair line.

Theodolites used in the process of tacheometry have varying levels of sophistication. The first types made in the early 19th century had fixed stadia hairs and an ability to flip over and sight in the reverse direction, so that a point of reference could be established to reduce measurement errors. Some newer theodolites have movable horizontal stadia hairs, and their position can be measured with a micrometer for more accurate horizontal and vertical sighting. Early theodolites were referred to as transit instruments, and are still used for basic topographic mapping and quick measure uses such as in archeology and geology, where precise measurements of distance and height are not required.

One of the advantages of tacheometry is that it is a rapid surveying method, and, if a basic theodolite is used, the equipment is fairly lightweight and easy to take into the field. It requires only two operators, one to hold the leveling rod with the stadia hair markings and one to measure it with the theodolite from a distance. The accuracy of the measured distances decreases in tacheometry as the distance between the leveling staff and theodolite increases. At a range of a quarter-mile (402 meters), the process is considered quite accurate, and, at a distance of a mile (1,609 meters), the error in horizontal distance is around 32 feet (9.75 meters) and 4 inches (10.16 centimeters) vertically

The ordinary methods of surveying with a theodolite, chain, and levelling instrument are fairly satisfactory when the ground is pretty clear of obstructions and not very precipitous, but it becomes extremely cumbersome when the ground is covered with bush, or broken up by ravines. Chain measurements then become slow and liable to considerable error; the levelling, too, is carried on at great disadvantage in point of speed, though without serious loss of accuracy.

Tacheometry is a branch of surveying in which horizontal and vertical distances are determined by angular observations with a tacheometer.There is no linear measurement.Tacheometry is not as accurate as chaining,but it is more rapid in rough and difficult countries where levelling is tedious and chaining is inaccurate and slow.

It is a method of surveying in which horizontal distances and (relative) vertical elevations are determined from subtended intervals and vertical angles observed with an instrument.

Tacheometry is used for

1. preparation of topographic map where both horizontal and vertical distances are required to be measured;
2. survey work in difficult terrain where direct methods of measurements are inconvenient;
3. reconnaissance survey for highways and railways etc;
4. establishment of secondary control points.

Instrument

The instruments employed in tacheometry are the engineer's transit and the leveling rod or stadia rod, the theodolite and the subtense bar, the self-reducing theodolite and the leveling rod, the distance wedge and the horizontal distance rod, and the reduction tacheometer and the horizontal distance rod.

3.1.1 Systems or Techeometry measurement

Depending on the type of instrument and methods/types of observations, tacheometric measurement systems can be divided into two basic types:

(i) Stadia systems and

(ii) Non-stadia systems

Stadia Systems

In there system's, staff intercepts at a pair of stadia hairs present at diaphragm, are considered. The stadia system consists of two methods:

- Fixed-hair method and
- Movable-hair method

Non Stadia Syatems

This method of surveying is primarily based on principles of trigonometry and thus telescopes without stadia diaphragm are used. This system comprises of two methods:

(i) Tangential method and

(ii) Subtense bar method.

Fixed air method or stadia method

It is the most prevalent method for tacheometric surveying. In this method, the telescope of the theodolite is equipped with two additional cross hairs, one above and the other below the main horizontal hair at equal distance. These additional cross hairs are known as stadia hairs. This is also known as tacheometer.

Principle of stadia method

A tacheometer is temporarily adjusted on the station P with horizontal line of sight. Let a and b be the lower and the upper stadia hairs of the instrument and their actual vertical separation be designated as i. Let f be the focal length of the objective lens of the tacheometer and c be horizontal distance between the optical centre of the objective lens and the vertical axis of the instrument. Let the objective lens is focused to a staff held vertically at Q, say at horizontal distance D from the instrument station.

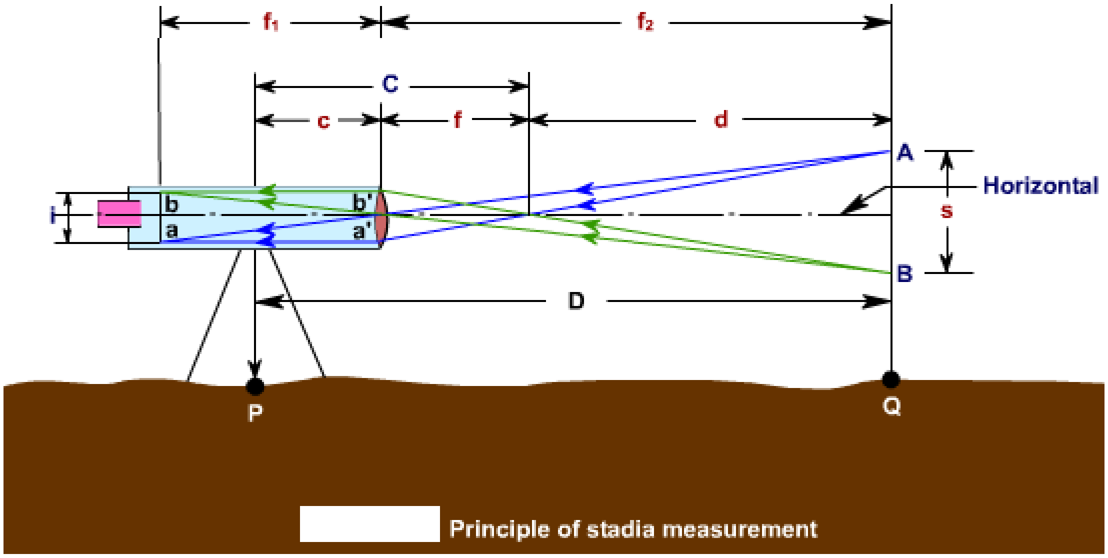

Principle of stadia measurement

By the laws of optics, the images of readings at A and B of the staff will appear along the stadia hairs at a and b respectively. Let the staff interval i.e., the difference between the readings at A and B be designated by s. Similar triangle between the object and image will form with vertex at the focus of the objective lens (F). Let the horizontal distance of the staff from F be d. Then, from the similar ABF and a' b' F,

$$\frac{AB}{d} = \frac{a'b'}{f}$$

$$\text{Or, } d = \frac{AB}{a'b'} \times f = \frac{s}{i} \times f$$

$$\therefore d = \frac{f}{i} \times s$$

as a' b' = ab = i. The ratio (f / i) is a constant for a particular instrument and is known as stadia interval factor, also instrument constant. It is denoted by K and thus

d = K.s --------------------- Equation (.1)

The horizontal distance (D) between the center of the instrument and the station point (Q) at which the staff is held is d + f + c. If C is substituted for (f + c), then the horizontal distance D from the center of the instrument to the staff is given by the equation

D = Ks + C ---------------------- Equation (2)

The distance C is called the stadia constant. Equation (2) is known as the stadia equation for a line of sight perpendicular to the staff intercept.

Determination of Tacheometric Constants

The stadia interval factor (K) and the stadia constant (C) are known as tacheometric constants. Before using a tacheometer for surveying work, it is reqired to determine these constants. These can be computed from field observation by adopting following procedure.

Step 1 : Set up the tacheometer at any station say P on a flat ground.

Step 2 : Select another point say Q about 200 m away. Measure the distance between P and Q accurately with a precise tape. Then, drive pegs at a uniform interval, say 50 m, along PQ. Mark the peg points as 1, 2, 3 and last peg -4 at station Q.

Step 3 : Keep the staff on the peg-1, and obtain the staff intercept say s_1 .

Step 4 : Likewise, obtain the staff intercepts say s_2, when the staff is kept at the peg-2,

Step 5 : Form the simultaneous equations, using Equation (2)

$D_1 = K. s_1 + C$ --------------(i)

and $D_2 = K. s_2 + C$ -------------(ii)

Solving Equations (i) and (ii), determine the values of K and C say K_1 and C_1 .

Step 6 : Form another set of observations to the pegs 3 & 4, Simultaneous equations can be obtained from the staff intercepts s_3 and s_4 at the peg-3 and point Q respectively. Solving those equations, determine the values of K and C again say K_2 and C_2.

Step 7 : The average of the values obtained in steps (5) and (6), provide the tacheometric constants K and C of the instrument.

Anallactic Lens

It is a special convex lens, fitted in between the object glass and eyepiece, at a fixed distance from the object glass, inside the telescope of a tacheometer. The function of the anallactic lens is to reduce the stadia constant to zero. Thus, when tacheometer is fitted with anallactic lens, the distance measured between instrument station and staff position (for line of sight perpendicular to the staff intercept) becomes directly proportional to the staff intercept. Anallactic lens is provided in external focusing type telescopes only.

Inclined Stadia Measurements

It is usual that the line of sight of the tacheometer is inclined to the horizontal. Thus, it is frequently required to reduce the inclined observations into horizontal distance and difference in elevation.

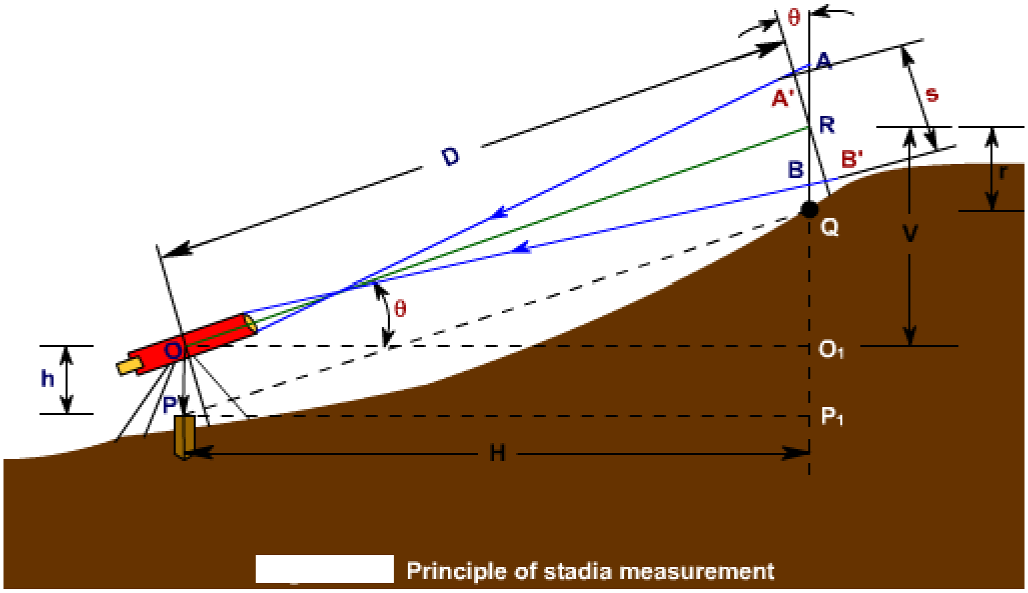

Principle of stadia measurement

Let us consider a tacheometer (having constants K and C) is temporarily adjusted on a station, say P. The instrument is sighted to a staff held vertically, say at Q. Thus, it is required to find the horizontal distance PP_1 (= H) and the difference in elevation P_1Q. Let A, R and B be the staff points whose images are formed respectively at the upper, middle and lower cross hairs of the tacheometer. The line of sight, corresponding to the middle cross hair, is inclined at an angle of elevation q and thus, the staff with a line perpendicular to the line of sight. Therefore A'B' = AB cos q = s cos q where s is the staff intercept AB. The distance D (= OR) is C + K. scos q . But the distance OO_1 is the horizontal distance H, which equals OR cos q. Therefore the horizontal distance H is given by the equation.

$H = (Ks \cos q + C) \cos q$

Or $H = Ks \cos^2 q + C \cos q$ ----------------- Equation(.3)

in which K is the stadia interval factor (f / i), s is the stadia interval, C is the stadia constant (f + c), and q is the vertical angle of the line of sight read on the vertical circle of the transit.

The distance RO_1, which equals OR sin q, is the vertical distance between the telescope axis and the middle cross-hair reading. Thus V is given by the equation

$V = (K\, s \cos q + c) \sin q$

$V = Ks \sin q \cos q + C \sin q$ ----------------- Equation (.4)

$\therefore V = \frac{1}{2} Ks \sin 2\theta + C \sin\theta$ ----------------- Equation (.5)

Thus, the difference in elevation between P and Q is (h + V - r), where h is the height of the instrument at P and r is the staff reading corresponding to the middle hair.

Examples

Ex1 In order to carry out tacheometric surveying, following observations were taken through a tacheometer set up at station P at a height 1.235m.

Staff held Vertical at	Horizontal distance from P (m)	Staff Reading (m)	Angle of Elevation
Q	100	1.01	0°
R	200	2.03	0°
S	?	3.465, 2.275, 1.280	5° 24' 40"

Compute the horizontal distance of S from P and reduced level of station at S if R.L. of station P is 262.575m

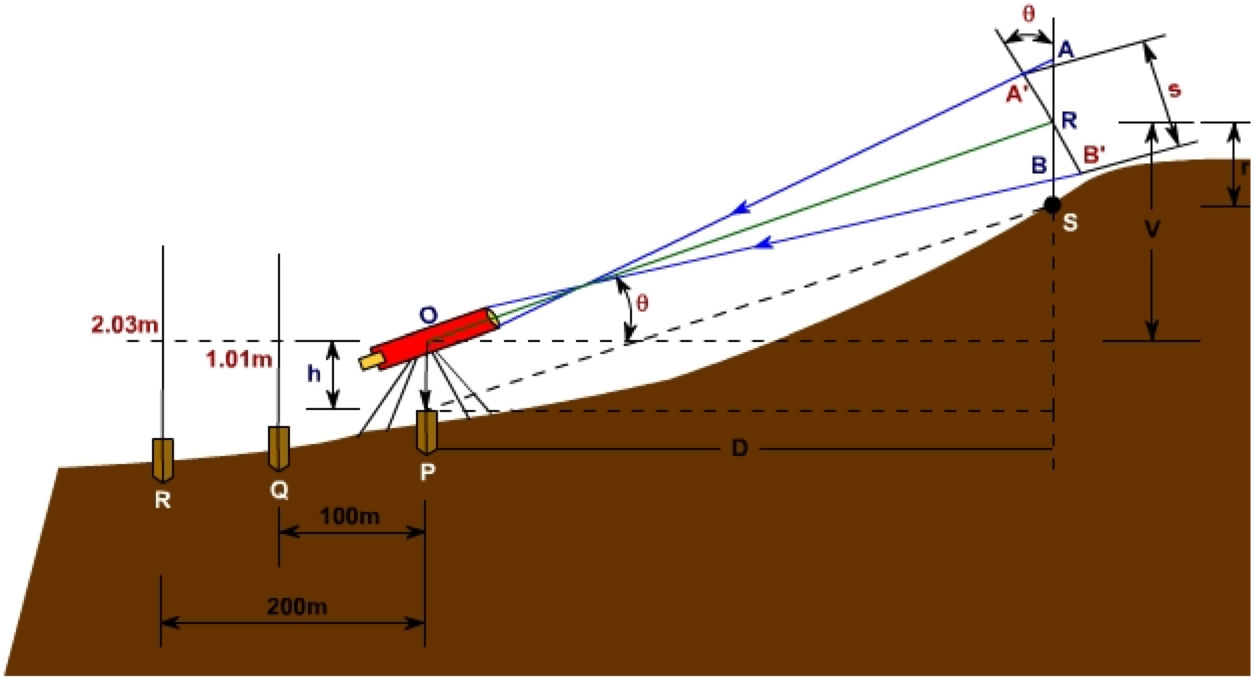

Figure

Solution :

Since the staff station P and Q are at known distances and observations are taken at horizontal line of sight, from equation 2

i.e. from D = K.s + C, we get

100 = K. 1.01 + C --------------- Equation 1

200 = K. 2.03 + C --------------- Equation 2

where K and C are the stadia interval factor and stadia constant of the instrument.

Therefore Solving equation 1 and 2 , $K = \frac{100}{1.02} = 98.04$

Substituting, value of K in Equation 1, we get

C = 100 - 1.01 x 98.04 = 0.98

Now, for the observation at staff station S, the staff intercept

s = 3.465 - 1.280 = 2.185 m;

Given, the angle of elevation (of a observation at S), q = 5° 24' 40"

Using equation 23.3 i.e., $D = K\ s \cos^2 q + C.\cos q$, the horizontal distance of S from P is

$D = 98.04 \times 2.185 \times \cos^2 5° 24' 40'' + 0.98 \cos 5° 24' 40''$

= 212.312 + 0.9756 = 213.288 m

Using equation 23.4, i.e., $V = \frac{1}{2} K s \sin 2\theta + C.\sin\theta$ the vertical distance

$$V = \frac{1}{2} \times 98.04 \times 2.186 \sin 2 \times (6°24'40'') + 0.98 \sin 6°24'20'$$

= (20.11 + 0.0924)m = 20.203 m

Thus R.L. of station S = R.L. of P + h + V - r

= 262.575 + 1.235 + 20.203 - 2.275

= 281.738 m

Uses of Stadia Method

The stadia method of surveying is particularly useful for following cases:

1. In differential leveling, the backsight and foresight distances are balanced conveniently if the level is equipped with stadia hairs.
2. In profile leveling and cross sectioning, stadia is a convenient means of finding distances from level to points on which rod readings are taken.
3. In rough trigonometric, or indirect, leveling with the transit, the stadia method is more rapid than any other method.
4. For traverse surveying of low relative accuracy, where only horizontal angles and distances are required, the stadia method is a useful rapid method.
5. On surveys of low relative accuracy - particularly topographic surveys-where both the relative location of points in a horizontal plane and the elevation of these points are desired, stadia is useful. The horizontal angles, vertical angles, and the stadia interval are observed, as each point is sighted; these three observations define the location of the point sighted.

Errors in Stadia Measurement

Most of the errors associated with stadia measurement are those that occur during observations for horizontal angles (Lesson 22) and differences in elevation (Lesson 16). Specific sources of errors in horizontal and vertical distances computed from observed stadia intervals are as follows:

1. Error in Stadia Interval factor

This produces a systematic error in distances proportional to the amount of error in the stadia interval factor.

2. Error in staff graduations

If the spaces on the rod are uniformly too long or too short, a systematic error proportional to the stadia interval is produced in each distance.

3. Incorrect stadia Interval

The stadia interval varies randomly owing to the inability of the instrument operator to observe the stadia interval exactly. In a series of connected observations (as a traverse) the error may be expected to vary as the square root of the number of sights. This is the principal error affecting the precision of distances. It can be kept to a minimum by proper focusing to eliminate parallax, by taking observations at favorable times, and by care in observing.

4. Error in verticality of staff

This condition produces a perceptible error in measurement of large vertical angles than for small angles. It also produces an appreciable error in the observed stadia interval and hence in computed distances. It can be eliminated by using a staff level.

5. Error due to refraction

This causes random error in staff reading.

6. Error in vertical angle

Error in vertical angle is relatively unimportant in their effect upon horizontal distance if the angle is small but it is perceptible if the vertical angle is large.

Review Questions

1. What is is **Tacheometric Surveying**
2. **What is** Tangential explain with diagram?
3. **Explain** Stadia and substense methods with diagram?
4. **What is** horizontal and inclined sights in stadia method?
5. **What is** vertical and normal staff ?
6. **What is** fixed and movable hair – stadia constants?
7. **Explain** anallatic lens?
8. **What is** Self reducing tacheometers.
9. **What are the application of Tacheometric Surveying**
10. **What are the merit and demerit of Tacheometric Surveying?**

Chapter 4

TRIANGULATION SURVEYING

Structure of this unit

Horizontal and vertical control , Trigonometric leveling, Satellite station

Learning Objectives

1. Horizontal and vertical control methods
2. triangulation –network- Signals. Base line
3. instruments and accessories – extension of base lines - corrections
4. Satellite station – reduction to centre – Intervisibility of height and distances
5. Trigonometric levelling – Axis single corrections.

4.1 TRIANGULATION SURVEYING

In trigonometry and geometry, triangulation is the process of determining the location of a point by measuring angles to it from known points at either end of a fixed baseline, rather than measuring distances to the point directly (trilateration). The point can then be fixed as the third point of a triangle with one known side and two known angles.

Triangulation can also refer to the accurate surveying of systems of very large triangles, called triangulation networks. This followed from the work ofWillebrord Snell in 1615–17, who showed how a point could be located from the angles subtended from three known points, but measured at the new unknown point rather than the previously fixed points, a problem called resectioning. Surveying error is minimized if a mesh of triangles at the largest appropriate scale is established first. Points inside the triangles can all then be accurately located with reference to it. Such triangulation methods were used for accurate large-scale land surveying until the rise of global navigation satellite systems in the 1980s.

A system of triangles usually affords superior horizontal control. All of the angles and at least one side (the base) of the triangulation system are measured. Though several arrangements can be used, one of the best is the quadrangle or a chain of quadrangles. Each quadrangle, with its four sides and two diagonals, provides eight angles that are measured. To be geometrically consistent, the angles must satisfy three so-called angle equations and one side equation. That is to say the three angles of each triangle, which add to 180°, must be of such sizes that computation through any set of adjacent triangles within the quadrangles will give the same values for any side. Ideally, the quadrangles should be parallelograms. If the system is connected with previously determined stations, the new system must fit the established measurements.

When the survey encompasses an area large enough for the Earth’s curvature to be a factor, an imaginary mathematical representation of the Earth must be employed as a reference surface. A level surface at mean sea level is considered to represent the Earth’s size and shape, and this is called thegeoid. Because of gravity anomalies, the geoid is irregular; however, it is very nearly the surface generated by an ellipse rotating on its minor axis—*i.e.,* an ellipsoid slightly flattened at the ends, or oblate. Such a figure is called a spheroid. Several have been computed by various authorities; the one usually used as a reference surface by English-speaking nations is (Alexander Ross) Clarke’s Spheroid of 1866. This oblate spheroid has a polar diameter about 27 miles (43 kilometres) less than its diameter at the Equator.

Because the directions of gravity converge toward the geoid, a length of the Earth’s surface measured above the geoid must be reduced to its sea-level equivalent—*i.e.,* to that of the geoid. These lengths are assumed to be the distances, measured on the spheroid, between the extended lines of gravity down to the spheroid from the ends of the measured lengths on the actual surface of the Earth. The positions of the survey stations on the Earth’s surface are given in spherical coordinates.

Bench marks, or marked points on the Earth's surface, connected by precise leveling constitute the vertical controls of surveying. The elevations of bench marks are given in terms of their heights above a selected level surface called a datum. In large-level surveys the usual datum is the geoid. The elevation taken as zero for the reference datum is the height of mean sea level determined by a series of observations at various points along the seashore taken continuously for a period of 19 years or more. Because mean sea level is not quite the same as the geoid, probably because of ocean currents, in adjusting the level grid for the United States and Canada all heights determined for mean sea level have been held at zero elevation.

Because the level surfaces, determined by leveling, are distorted slightly in the area toward the Earth's poles (because of the reduction in centrifugal force and the increase in the force of gravity at higher latitudes), the distances between the surfaces and the geoid do not exactly represent the surfaces' heights from the geoid. To correct these distortions, orthometric corrections must be applied to long lines of levels at high altitudes that have a north–south trend.

Trigonometric leveling often is necessary where accurate elevations are not available or when the elevations of inaccessible points must be determined. From two points of known position and elevation, the horizontal position of the unknown point is found by triangulation, and the vertical angles from the known points are measured. The differences in elevation from each of the known points to the unknown point can be computed trigonometrically.

The National Ocean Service in recent years has hoped to increase the density of horizontal control to the extent that no location in the United States will be farther than 50 miles (80 kilometres) from a primary point, and advances anticipated in analytic phototriangulation suggest that the envisioned density of control may soon suffice insofar as topographic mapping is concerned. Existing densities of control in Britain and much of western Europe are already adequate for mapping and cadastral surveys.

Applications

Optical 3d measuring systems use this principle as well in order to determine the spatial dimensions and the geometry of an item. Basically, the configuration consists of two sensors observing the item. One of the sensors is typically a digital camera device, and the other one can also be a camera or a light projector. The projection centers of the sensors and the considered point on the object's surface define a (spatial) triangle. Within this triangle, the distance between the sensors is the base b and must be known. By determining the angles between the projection rays of the sensors and the basis, the intersection point, and thus the 3d coordinate, is calculated from the triangular relations.

Distance to a point by measuring two fixed angles

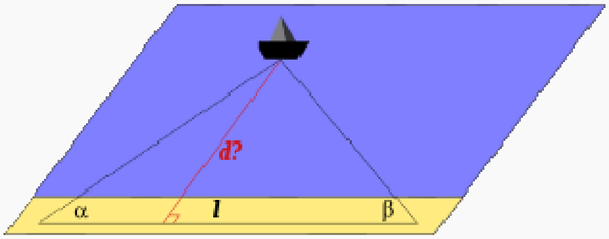

Triangulation may be used to calculate the coordinates anddistance from the shore to the ship. The observer at Ameasures the angle α between the shore and the ship, and the observer at B does likewise for β . With the length l or thecoordinates of A and B known, then the law of sines can be applied to find the coordinates of the ship at C and the distanced.

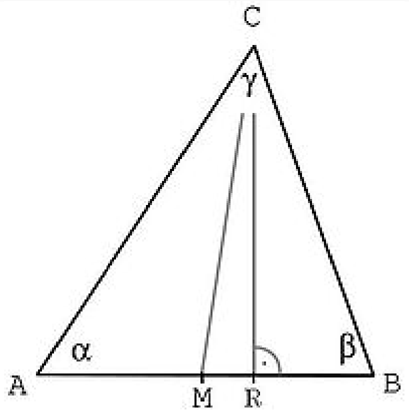

The coordinates and distance to a point can be found by calculating the length of one side of a triangle, given measurements of angles and sides of the triangle formed by that point and two other known reference points.

The following formulae apply in flat or Euclidean geometry. They become inaccurate if distances become appreciable compared to thecurvature of the Earth, but can be replaced with more complicated results derived using spherical trigonometry.

Calculation[edit]

$$\ell = \frac{d}{\tan\alpha} + \frac{d}{\tan\beta}$$

Therefore

$$1/d = 1/l * \left(\frac{1}{\tan\alpha} + \frac{1}{\tan\beta}\right)$$

Using the trigonometric identities tan α = sin α / cos α and sin(α + β) = sin α cos β + cos α sin β, this is equivalent to:

$$1/d = \frac{\sin(\alpha + \beta)}{\ell \ \sin\alpha \sin\beta}$$

From this, it is easy to determine the distance of the unknown point from either observation point, its north/south and east/west offsets from the observation point, and finally its full coordinates.

History

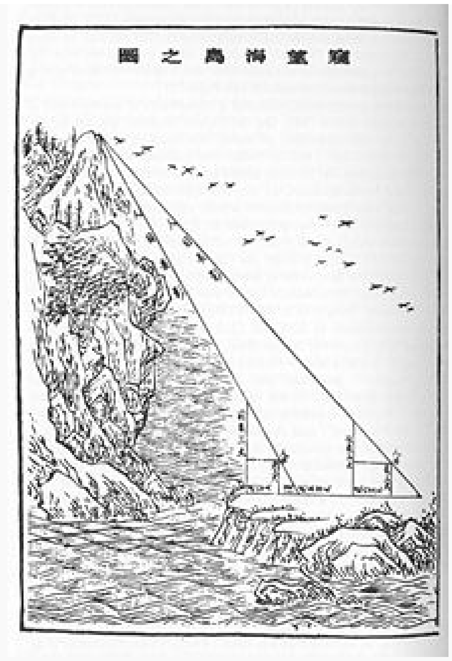

Liu Hui (c. 263), How to measure the height of a sea island. Illustration from an edition of 1726

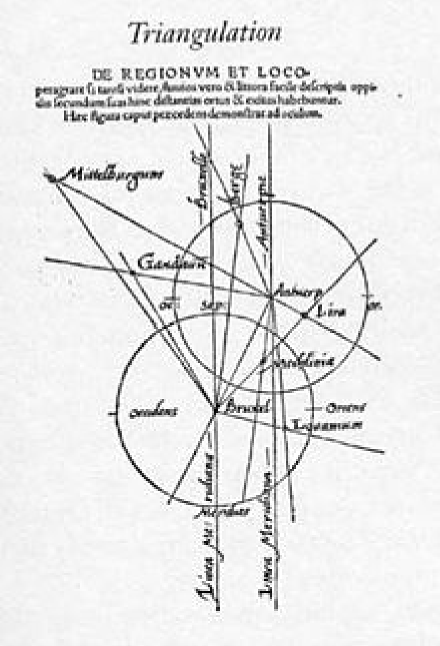
Triangulation
DE REGIONVM ET LOCO
Mittelburgum
Gandavum
Antwerpia
Lira
Bruxel
Louanium
occidens
Oriens
Meridies
Linea meridiana
Linea Meridiana

Gemma Frisius's 1533 proposal to use triangulation for mapmaking

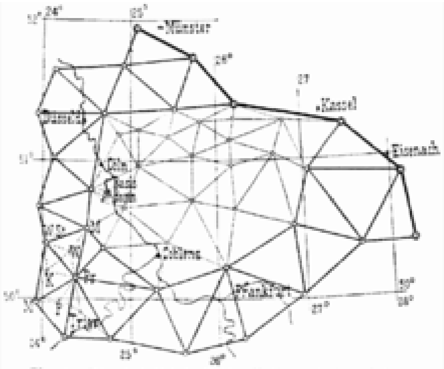
Münster
Kassel
Eisenach
Frankfurt

Fig. 4. Die rheinisch-hessische Kette und das niederrheinische Dreiecksnetz.

Nineteenth-century triangulation network for the triangulation of Rhineland-Hesse

Triangulation today is used for many purposes, including surveying, navigation, metrology, astrometry, binocular vision, model rocketry and gun direction ofweapons.

The use of triangles to estimate distances goes back to antiquity. In the 6th century BC the Greek philosopher Thales is recorded as using similar triangles to estimate the height of the pyramids by measuring the length of their shadows and that of his own at the same moment, and comparing the ratios to his height (intercept theorem); and to have estimated the distances to ships at sea as seen from a clifftop, by measuring the horizontal distance traversed by the line-of-sight for a known fall, and scaling up to the height of the whole cliff.[2] Such techniques would have been familiar to the ancient Egyptians. Problem 57 of the Rhind papyrus, a thousand years earlier, defines the seqt or seked as the ratio of the run to the rise of a slope, i.e. the reciprocal of gradients as measured today. The slopes and angles were measured using a sighting rod that the Greeks called a dioptra, the forerunner of the Arabicalidade. A detailed contemporary collection of constructions for the determination of lengths from a distance using this instrument is known, the Dioptraof Hero of Alexandria (c. 10–70 AD), which survived in Arabic translation; but the knowledge became lost in Europe. In China, Pei Xiu (224–271) identified "measuring right angles and acute angles" as the fifth of his six principles for accurate map-making, necessary to accurately establish distances;[3] while Liu Hui (c. 263) gives a version of the calculation above, for measuring perpendicular distances to inaccessible places.

In the field, triangulation methods were apparently not used by the Roman specialist land surveyors, the agromensores; but were introduced into medieval Spain through Arabic treatises on the astrolabe, such as that by Ibn al-Saffar (d. 1035). Abu Rayhan Biruni (d. 1048) also introduced triangulation techniques to measure the size of the Earth and the distances between various places.[7] Simplified Roman techniques then seem to have co-existed with more sophisticated techniques used by professional surveyors. But it was rare for such methods to be translated into Latin (a manual on Geometry, the eleventh century Geomatria incerti auctoris is a rare exception), and such techniques appear to have percolated only slowly into the rest of Europe. Increased awareness and use of such techniques in Spain may be attested by the medieval Jacob's staff, used specifically for measuring angles, which dates from about 1300; and the appearance of accurately surveyed coastlines in the Portolan charts, the earliest of which that survives is dated 1296.

Gemma Frisius and triangulation for mapmaking

On land, the Flemish cartographer Gemma Frisius proposed using triangulation to accurately position far-away places for map-making in his 1533 pamphlet Libellus de Locorum describendorum ratione (Booklet concerning a way of describing places), which he bound in as an appendix in a new edition of Peter Apian's best-selling 1524 Cosmographica. This became very influential, and the technique spread across Germany, Austria and the Netherlands. The astronomer Tycho Brahe applied the method in Scandinavia, completing a detailed triangulation in 1579 of the island of Hven, where his observatory was based, with reference to key landmarks on both sides of the Øresund, producing an estate plan of the island in 1584.[8] In England Frisius's method was included in the growing number of books on surveying which appeared from the middle of the century onwards, including William Cuningham's Cosmographical Glasse (1559), Valentine Leigh's Treatise of Measuring All Kinds of Lands (1562), William Bourne's Rules of Navigation(1571), Thomas Digges's Geometrical Practise named Pantometria (1571), and John Norden's Surveyor's Dialogue (1607). It has been suggested thatChristopher Saxton may have used rough-and-ready triangulation to place features in his county maps of the 1570s; but others suppose that, having obtained rough bearings to features from key vantage points, he may have estimated the distances to them simply by guesswork

Willebrord Snell and modern triangulation networks

The modern systematic use of triangulation networks stems from the work of the Dutch mathematician Willebrord Snell, who in 1615 surveyed the distance from Alkmaar to Bergen op Zoom, approximately 70 miles (110 kilometres), using a chain of quadrangles containing 33 triangles in all. The two towns were separated by one degree on the meridian, so from his measurement he was able to calculate a value for the circumference of the earth – a feat celebrated in the title of his book Eratosthenes Batavus (The Dutch Eratosthenes), published in 1617. Snell calculated how the planar formulae could be corrected to allow for the curvature of the earth. He also showed how to resection, or calculate, the position of a point inside a triangle using the angles cast between the vertices at the unknown point. These could be measured much more accurately than bearings of the vertices, which depended on a compass. This established the key idea of surveying a large-scale primary network of control points first, and then locating secondary subsidiary points later, within that primary network.

Snell's methods were taken up by Jean Picard who in 1669–70 surveyed one degree of latitude along the Paris Meridian using a chain of thirteen triangles stretching north from Paris to the clocktower of Sourdon, near Amiens. Thanks to improvements in instruments and accuracy, Picard's is rated as the first reasonably accurate measurement of the radius of the earth. Over the next century this work was extended most notably by the Cassini family: between 1683 and 1718 Jean-Dominique Cassini and his son Jacques Cassini surveyed the whole of the Paris meridian from Dunkirk toPerpignan; and between 1733 and 1740 Jacques and his son César Cassini undertook the first triangulation of the whole country, including a re-surveying of the meridian arc, leading to the publication in 1745 of the first map of France constructed on rigorous principles.

Triangulation methods were by now well established for local mapmaking, but it was only towards the end of the 18th century that other countries began to establish detailed triangulation network surveys to map whole countries. The Principal Triangulation of Great Britain was begun by the Ordnance Survey in 1783, though not completed until 1853; and the Great Trigonometric Survey of India, which ultimately named and mapped Mount Everest and the other Himalayan peaks, was begun in 1801. For the Napoleonic French state, the French triangulation was extended by Jean Joseph Tranchot into the German Rhineland from 1801, subsequently completed after 1815 by the Prussian general Karl von Müffling. Meanwhile, the famous mathematicianCarl Friedrich Gauss was entrusted from 1821 to 1825 with the triangulation of the kingdom of Hanover, for which he developed the method of least squares to find the best fit solution for problems of large systems of simultaneous equations given more real-world measurements than unknowns.

Today, large-scale triangulation networks for positioning have largely been superseded by the Global navigation satellite systems established since the 1980s. But many of the control points for the earlier surveys still survive as valued historical features in the landscape, such as the concretetriangulation pillars set up for retriangulation of Great Britain (1936–1962), or the triangulation points set up for the Struve Geodetic Arc (1816–1855), now scheduled as a

UNESCO World Heritage Site.

4.2 Triangulation network

The horizontal positions of points is a network developed to provide accurate control for topographic mapping, charting lakes, rivers and ocean coast lines, and for the surveys required for the design and construction of public and private works of large extent. The horizontal positions of the points can be obtained in a number of different ways in addition to traversing. These methods are triangulation, trilateration, intersection, resection, and satellite positioning. The method of surveying called triangulation is based on the trigonometric proposition that if one side and two angles of a triangle are known, the remaining sides can be computed. Furthermore, if the direction of one side is known, the directions of the remaining sides can be determined. A triangulation system consists of a series of joined or overlapping triangles in which an occasional side is measured and remaining sides are calculated from angles measured at the vertices of the triangles. The vertices of the triangles are known as triangulation stations. The side of the triangle whose length is predetermined, is called the base line. The lines of triangulation system form a network that ties together all the triangulation stations.

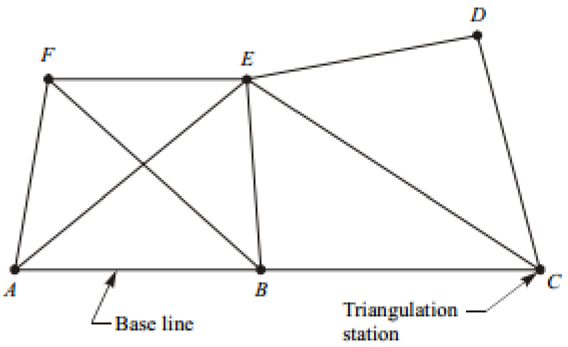

Fig. 1.1 Triangulation network

Triangulation network

A trilateration system also consists of a series of joined or overlapping triangles. However, for trilateration the lengths of all the sides of the triangle are measured and few directions or angles are measured to establish azimuth. Trilateration has become feasible with the development of electronic distance measuring (EDM) equipment which has made possible the measurement of all lengths with high order of accuracy under almost all field conditions. A combined triangulation and trilateration system consists of a network of triangles in which all the angles and all the lengths are measured. Such a combined system represents the strongest network for creating horizontal control.

Since a triangulation or trilateration system covers very large area, the curvature of the earth has to be taken into account. These surveys are, therefore, invariably geodetic. Triangulation surveys were first carried out by Snell, a Dutchman, in 1615. Field procedures for the establishment of trilateration station are similar to the procedures used for triangulation, and therefore, henceforth in this chapter the term triangulation will only be used.

Principle of triangulation

shows two interconnected triangles *ABC* and *BCD*. All the angles in both the triangles and the length *L* of the side *AB*, have been measured.

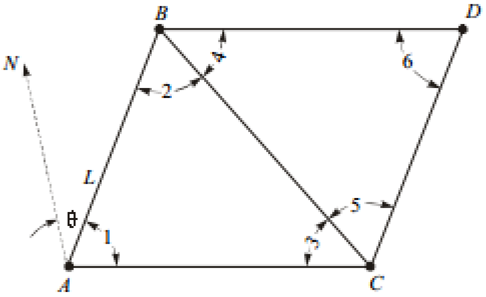

Principle of triangulation

Also the azimuth θ of *AB* has been measured at the triangulation station *A*, whose coordinates (X_A, Y_A), are known.

The objective is to determine the coordinates of the triangulation stations *B*, *C*, and *D* by the method of triangulation. Let us first calculate the lengths of all the lines.

By sine rule in ΔABC, we have

$$\frac{AB}{\sin 3} = \frac{BC}{\sin 1} = \frac{CA}{\sin 2}$$

We have $$AB = L = l_{AB}$$

or $$BC = \frac{L \sin 1}{\sin 3} = l_{BC}$$

and $$CA = \frac{L \sin 2}{\sin 3} = l_{CA}$$

Now the side *BC* being known in ΔBCD, by sine rule, we have

$$\frac{BC}{\sin 6} = \frac{CD}{\sin 4} = \frac{BD}{\sin 5}$$

We have $$BC = \frac{L \sin 1}{\sin 3} = l_{BC}$$

or $$CD = \left(\frac{L \sin 1}{\sin 3}\right)\frac{\sin 4}{\sin 6} = l_{CD}$$

and $$BC = \left(\frac{L \sin 1}{\sin 3}\right)\frac{\sin 5}{\sin 6} = l_{BD}$$

Let us now calculate the azimuths of all the lines.

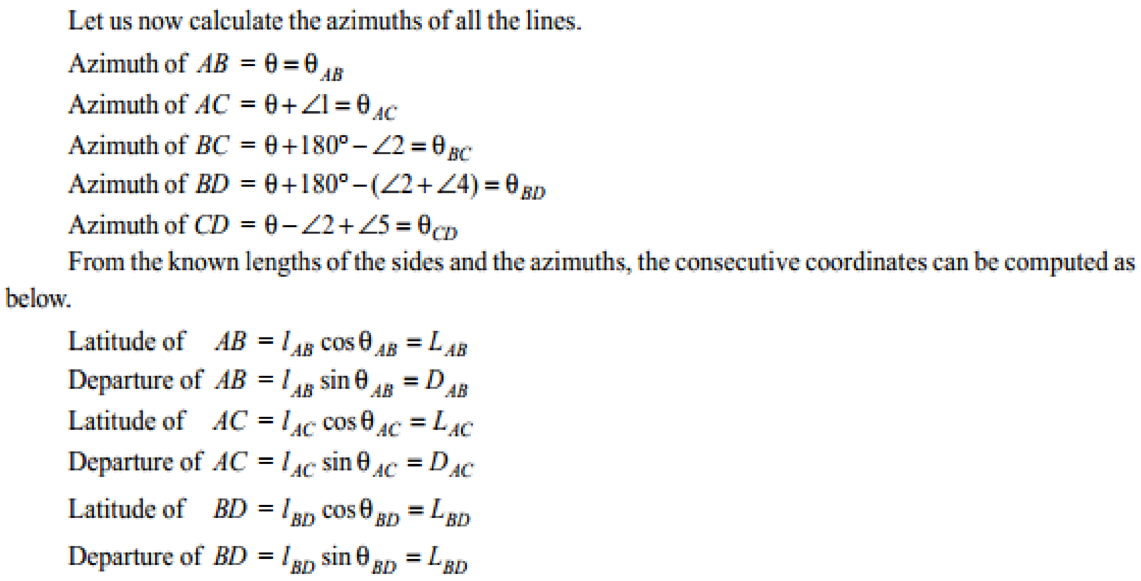

Let us now calculate the azimuths of all the lines.

Azimuth of $AB = \theta = \theta_{AB}$

Azimuth of $AC = \theta + \angle 1 = \theta_{AC}$

Azimuth of $BC = \theta + 180° - \angle 2 = \theta_{BC}$

Azimuth of $BD = \theta + 180° - (\angle 2 + \angle 4) = \theta_{BD}$

Azimuth of $CD = \theta - \angle 2 + \angle 5 = \theta_{CD}$

From the known lengths of the sides and the azimuths, the consecutive coordinates can be computed as below.

Latitude of $AB = l_{AB} \cos\theta_{AB} = L_{AB}$

Departure of $AB = l_{AB} \sin\theta_{AB} = D_{AB}$

Latitude of $AC = l_{AC} \cos\theta_{AC} = L_{AC}$

Departure of $AC = l_{AC} \sin\theta_{AC} = D_{AC}$

Latitude of $BD = l_{BD} \cos\theta_{BD} = L_{BD}$

Departure of $BD = l_{BD} \sin\theta_{BD} = L_{BD}$

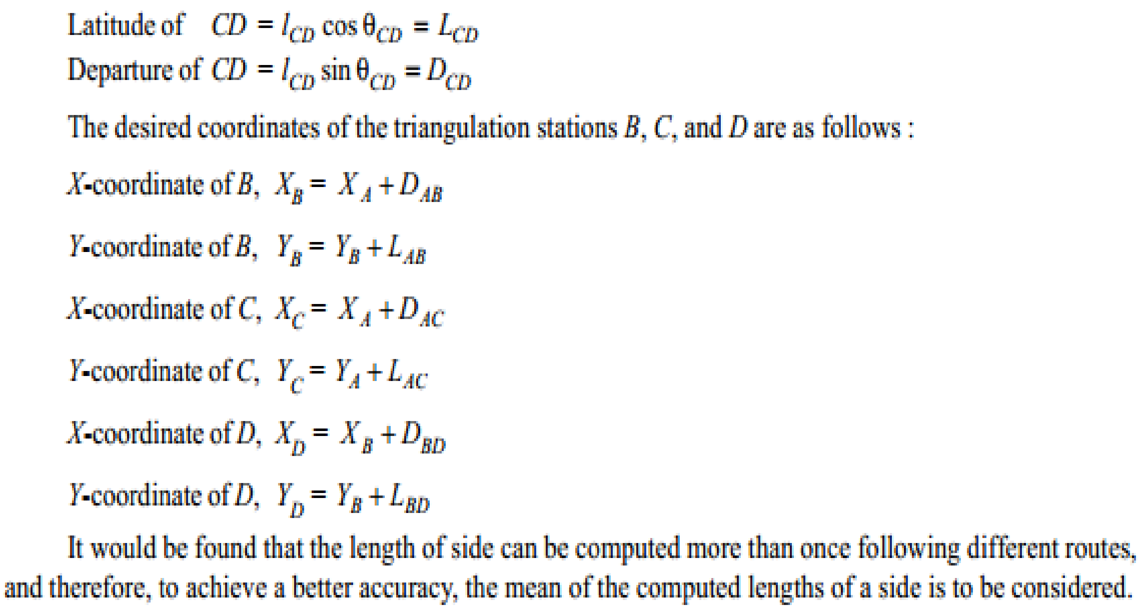

Latitude of $CD = l_{CD} \cos\theta_{CD} = L_{CD}$

Departure of $CD = l_{CD} \sin\theta_{CD} = D_{CD}$

The desired coordinates of the triangulation stations *B*, *C*, and *D* are as follows :

X-coordinate of *B*, $X_B = X_A + D_{AB}$

Y-coordinate of *B*, $Y_B = Y_B + L_{AB}$

X-coordinate of *C*, $X_C = X_A + D_{AC}$

Y-coordinate of *C*, $Y_C = Y_A + L_{AC}$

X-coordinate of *D*, $X_D = X_B + D_{BD}$

Y-coordinate of *D*, $Y_D = Y_B + L_{BD}$

It would be found that the length of side can be computed more than once following different routes, and therefore, to achieve a better accuracy, the mean of the computed lengths of a side is to be considered.

4.2.1 OBJECTIVE OF TRIANGULATION SURVEYS

The main objective of triangulation or trilateration surveys is to provide a number of stations whose relative and absolute positions, horizontal as well as vertical, are accurately established. More detailed location or engineering survey are then carried out from these stations.

The triangulation surveys are carried out

(i) to establish accurate control for plane and geodetic surveys of large areas, by terrestrial methods,

(ii) to establish accurate control for photogrammetric surveys of large areas,

(iii) to assist in the determination of the size and shape of the earth by making observations for latitude, longitude and gravity, and

(iv) to determine accurate locations of points in engineering works such as :

(a) Fixing centre line and abutments of long bridges over large rivers.

(b) Fixing centre line, terminal points, and shafts for long tunnels.

(c) Transferring the control points across wide sea channels, large water bodies, etc.

(d) Detection of crustal movements, etc.

(e) Finding the direction of the movement of clouds.

4.2.2 CLASSIFICATION OF TRIANGULATION SYSTEM

Based on the extent and purpose of the survey, and consequently on the degree of accuracy desired, triangulation surveys are classified as first-order or primary, second-order or secondary, and third-order or tertiary. First-order triangulation is used to determine the shape and size of the earth or to cover a vast area like a whole country with control points to which a second-order triangulation system can be connected. A second-order triangulation system consists of a network within a first-order triangulation. It is used to cover areas of the order of a region, small country, or province. A third-order triangulation is a framework fixed within and connected to a second-order triangulation system. It serves the purpose of furnishing the immediate control for detailed engineering and location surveys.

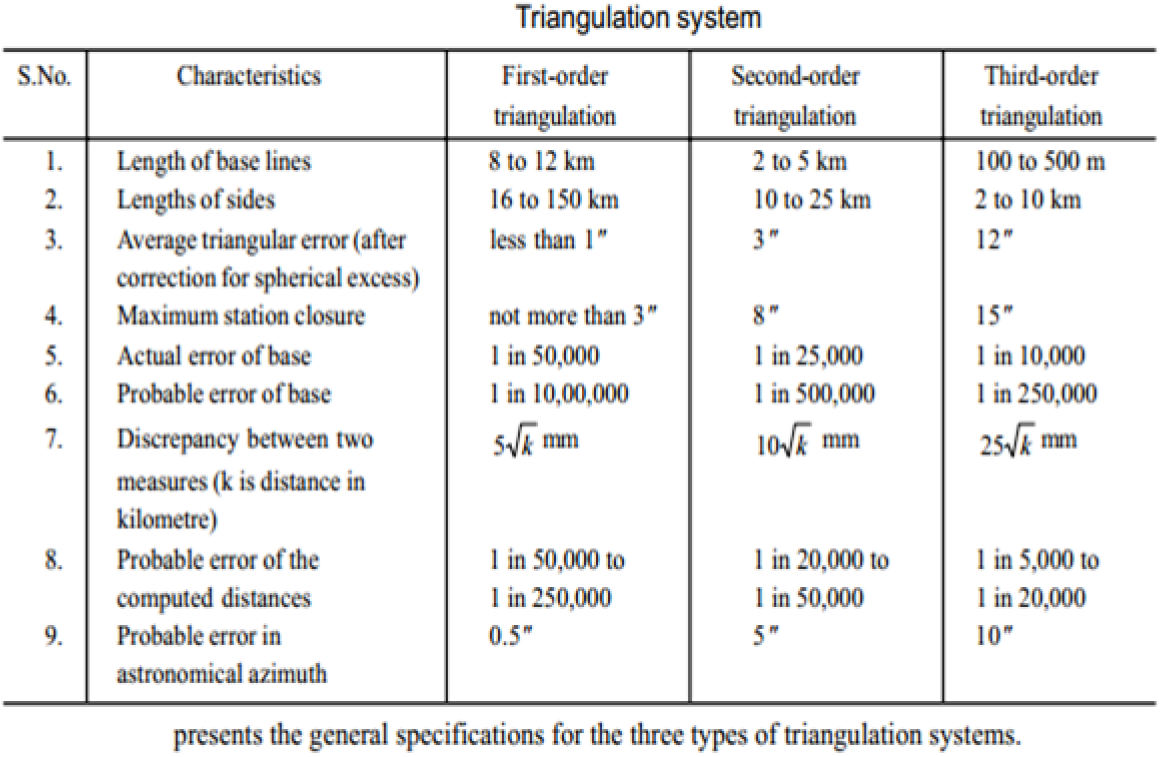

Triangulation system

S.No.	Characteristics	First-order triangulation	Second-order triangulation	Third-order triangulation
1.	Length of base lines	8 to 12 km	2 to 5 km	100 to 500 m
2.	Lengths of sides	16 to 150 km	10 to 25 km	2 to 10 km
3.	Average triangular error (after correction for spherical excess)	less than 1″	3″	12″
4.	Maximum station closure	not more than 3″	8″	15″
5.	Actual error of base	1 in 50,000	1 in 25,000	1 in 10,000
6.	Probable error of base	1 in 10,00,000	1 in 500,000	1 in 250,000
7.	Discrepancy between two measures (k is distance in kilometre)	$5\sqrt{k}$ mm	$10\sqrt{k}$ mm	$25\sqrt{k}$ mm
8.	Probable error of the computed distances	1 in 50,000 to 1 in 250,000	1 in 20,000 to 1 in 50,000	1 in 5,000 to 1 in 20,000
9.	Probable error in astronomical azimuth	0.5″	5″	10″

presents the general specifications for the three types of triangulation systems.

TRIANGULATION FIGURES AND LAYOUTS

The basic figures used in triangulation networks are the triangle, braced or geodetic quadilateral, and the polygon with a central station

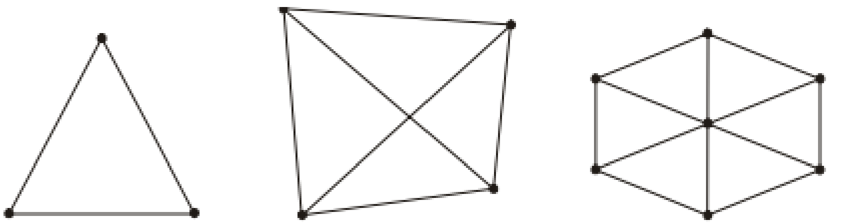

Triangle

Braced quadrilateral

Polygon with central station

Basic triangulation figures

The triangles in a triangulation system can be arranged in a number of ways. Some of the commonly used arrangements, also called layouts, are as follows :

1. Single chain of triangles

2. Double chain of triangles

3. Braced quadrilaterals

4. Centered triangles and polygons

5. A combination of above systems.

4.2 Single chain of triangles

When the control points are required to be established in a narrow strip of terrain such as a valley between ridges, a layout consisting of single chain of triangles is generally used as shown in. This system is rapid and economical due to its simplicity of sighting only four other stations, and does not involve observations of long diagonals. On the other hand, simple triangles of a triangulation system provide only one route through which distances can be computed, and hence, this system does not provide any check on the accuracy of observations. Check base lines and astronomical observations for azimuths have to be provided at frequent intervals to avoid excessive accumulation of errors in this layout.

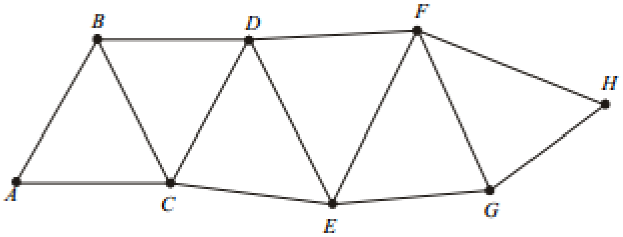

Fig. 1.4 Single of triangles

Double chain of triangles

A layout of double chain of triangles is shown in Fig. 1.5. This arrangement is used for covering the larger width of a belt. This system also has disadvantages of single chain of triangles system.

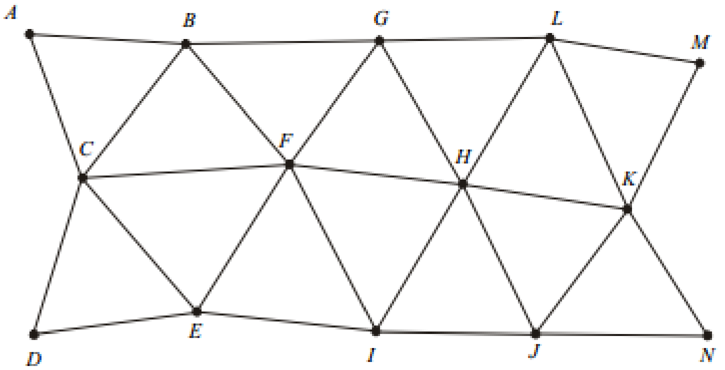

Double chain of triangles

Braced quadrilaterals

A triangulation system consisting of figures containing four corner stations and observed diagonals shown in below Fig. is known as a layout of braced quadrilaterals. In fact, braced quadrilateral consists of overlapping triangles. This system is treated to be the strongest and the best arrangement of triangles, and it provides a means of computing the lengths of the sides using different combinations of sides and angles. Most of the triangulation systems use this arrangement.

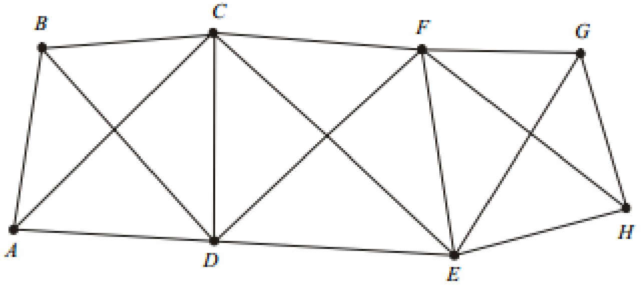

Braced quadrilaterals

Centered triangles and polygons

A triangulation system which consists of figures containing interior stations in triangle and polygon as shown in below Fig., is known as centered triangles and polygons.

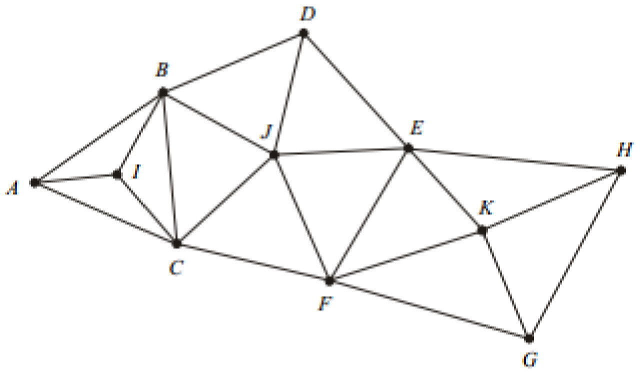

Centered triangles and polygons

This layout in a triangulation system is generally used when vast area in all directions is required to be covered. The centered figures generally are quadrilaterals, pentagons, or hexagons with central stations. Though this system provides checks on the accuracy of the work, generally it is not as strong as the braced quadrilateral arrangement. Moreover, the progress of work is quite slow due to the fact that more settings of the instrument are required.

A combination of all above systems

Sometimes a combination of above systems may be used which may be according to the shape of the area and the accuracy requirements.

LAYOUT OF PRIMARY TRIANGULATION FOR LARGE COUNTRIES

The following two types of frameworks of primary triangulation are provided for a large country to cover the entire area.

1. Grid iron system

2. Central system

Grid iron system

In this system, the primary triangulation is laid in series of chains of triangles, which usually runs roughly along meridians (northsouth) and along perpendiculars to the meridians (east-west), throughout the country. The distance between two such chains may vary from 150 to 250 km. The area between the parallel and perpendicular series of primary triangulation, are filled by the secondary and tertiary triangulation systems. Grid iron system has been adopted in India and other countries like Austria, Spain, France, etc.

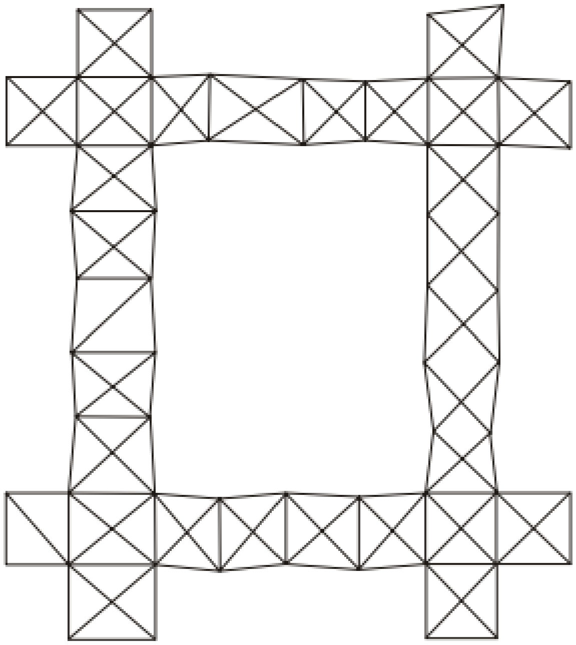

Grid iron system of triangulation

Central system

In this system, the whole area is covered by a network of primary triangulation extending in all directions from the initial triangulation figure ABC, which is generally laid at the centre of the country (Fig. 1.9). This system is generally used for the survey of an area of moderate extent. It has been adopted in United Kingdom and various other countries.

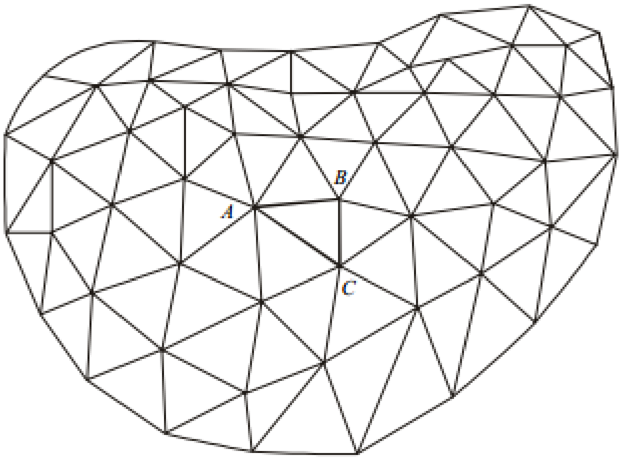

Central system of triangulation

CRITERIA FOR SELECTION OF THE LAYOUT OF TRIANGLES

The under mentioned points should be considered while deciding and selecting a suitable layout of triangles.

1. Simple triangles should be preferably equilateral.

2. Braced quadrilaterals should be preferably approximate squares.

3. Centered polygons should be regular.

4. The arrangement should be such that the computations can be done through two or more independent routes.

5. The arrangement should be such that at least one route and preferably two routes form wellconditioned triangles.

6. No angle of the figure, opposite a known side should be small, whichever end of the series is used for computation.

7. Angles of simple triangles should not be less than 45°, and in the case of quadrilaterals, no angle should be less than 30°. In the case of centered polygons, no angle should be less than 40°.

8. The sides of the figures should be of comparable lengths. Very long lines and very short lines should be avoided.

9. The layout should be such that it requires least work to achieve maximum progress.

10. As far as possible, complex figures should not involve more than 12 conditions.

It may be noted that if a very small angle of a triangle does not fall opposite the known side it does not affect the accuracy of triangulation.

4.3 WELL-CONDITIONED TRIANGLES

The accuracy of a triangulation system is greatly affected by the arrangement of triangles in the layout and the magnitude of the angles in individual triangles. The triangles of such a shape, in which any error in angular measurement has a minimum effect upon the computed lengths, is known as well-conditioned triangle. In any triangle of a triangulation system, the length of one side is generally obtained from computation of the adjacent triangle. The error in the other two sides if any, will affect the sides of the triangles whose computation is based upon their values. Due to accumulated errors, entire triangulation system is thus affected thereafter. To ensure that two sides of any triangle are equally affected, these should, therefore, be equal in length. This condition suggests that all the triangles must, therefore, be isoceles. Let us consider an isosceles triangle ABC whose one side AB is of known length. Let A, B, and C be the three angles of the triangle and a, b, and c are the three sides opposite to the angles, respectively.

As the triangle is isosceles, let the sides a and b be equal

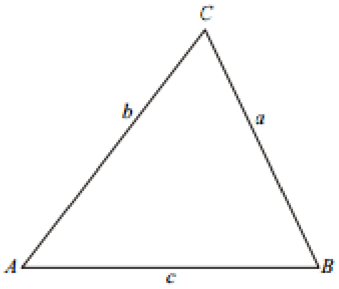

Triangle in a triangulation system

Applying sine rule to ΔABC, we have

$$\frac{a}{\sin A} = \frac{c}{\sin C} \qquad \text{...(.1)}$$

or

$$a = c\,\frac{\sin A}{\sin C} \qquad \text{...(2)}$$

If an error of δA in the angle A, and δC in angle C introduce the errors δa_1 and δa_2, respectively, in the side a, then differentiating Eq. (1.2) partially, we get

$$\delta a_1 = c\,\frac{\cos A\ \delta A}{\sin C} \qquad \text{...(.3)}$$

and

$$\delta a_2 = -c\,\frac{\sin A \cos C\ \delta C}{\sin^2 C} \qquad \text{...(.4)}$$

Dividing Eq. (1.3) by Eq. (1.2), we get

$$\frac{\delta a_1}{a} = \delta A \cot A \qquad \text{...(5)}$$

Dividing Eq. (1.4) by Eq. (1.2), we get

$$\frac{\delta a_2}{a} = -\delta C \cot C \qquad \text{...(.6)}$$

If $\delta A = \delta C = \pm\alpha$, is the probable error in the angles, then the probable errors in the side a are

$$\frac{\delta a}{a} = \pm\alpha\sqrt{\cot^2 A + \cot^2 C}$$

But $C = 180° - (A + B)$

or $= 180° - 2A$, A being equal to B.

Therefore

$$\frac{\delta a}{a} = \pm\alpha\sqrt{\cot^2 A + \cot^2 2A} \qquad \text{...(.7)}$$

From Eq. (7), we find that, if $\frac{\delta a}{a}$ is to be minimum, $(\cot^2 A + \cot^2 2A)$ should be a minimum.

Differentiating $\cot^2 A + \cos^2 2A$ with respect to A, and equating to zero, we have

$$4\cos^4 A + 2\cos^2 A - 1 = 0 \qquad \text{...(.8)}$$

Solving Eq. (1.8), for $\cos A$, we get

$$A = 56°14' \textit{(approximately)}$$

Hence, the best shape of an isoceles triangle is that triangle whose base angles are 56°14′ each. However, from practical considerations, an equilateral triangle may be treated as a well-conditional triangle. In actual practice, the triangles having an angle less than 30° or more than 120° should not be considered.

4.4 STRENGTH OF FIGURE

The strength of figure is a factor to be considered in establishing a triangulation system to maintain the computations within a desired degree of precision. It plays also an important role in deciding the layout of a triangulation system. The U.S. Coast and Geodetic Surveys has developed a convenient method of evaluating the strength of a triangulation figure. It is based on the fact that computations in triangulation involve use of angles of triangle and length of one known side. The other two sides are computed by sine law. For a given change in the angles, the sine of small angles change more rapidly than those of large angles. This suggests that smaller angles less than 30° should not be used in the computation of triangulation. If, due to unavoidable circumstances, angles less than 30° are used, then it must be ensured that this is not opposite the side whose length is required to be computed for carrying forward the triangulation series.

The expression given by the U.S. Coast and Geodetic Surveys for evaluation of the strength of figure, is for the square of the probable error (L^2) that would occur in the sixth place of the logarithm of any side, if the computations are carried from a known side through a single chain of triangles after the net has been adjusted for the side and angle conditions. The expression for L^2 is

$$L^2 = \frac{4}{3} d^2 R \qquad \text{………………..9}$$

where d is the probable error of an observed direction in seconds of arc, and R is a term which represents the shape of figure. It is given by

$$R = \frac{D-C}{D} \Sigma(\delta_A^2 + \delta_A \delta_B + \delta_B^2) \qquad \text{...(10)}$$

Where D = the number of directions observed excluding the known side of the figure,

A B C δ ,δ ,δ = the difference per second in the sixth place of logarithm of the sine of the distance angles A, B and C, respectively. (Distance angle is the angle in a triangle opposite to a side), and C = the number of geometric conditions for side and angle to be satisfied in each figure. It is given by

$$C = (n' - S' + 1) + (n - 2S + 3) \qquad \text{...(11)}$$

where

n = the total number of lines including the known side in a figure,

n' = the number of lines observed in both directions including the known side,

S = the total number of stations, and

S' = the number of stations occupied

For the computation of the quantity $\Sigma(\delta_A^2 + \delta_A \delta_B + \delta_B^2)$ in Eq. (10) , may be used.

In any triangulation system more than one routes are possible for various stations. The strength of figure decided by the factor R alone determines the most appropriate route to adopt the best shaped triangulation net route. If the computed value of R is less, the strength of figure is more and vice versa.

4.5 ACCURACY OF TRIANGULATION

Errors are inevitable and, therefore, inspite of all precautions the errors get accumulated. It is, therefore, essential to know the accuracy of the triangulation network achieved so that no appreciable error in plotting is introduced. The following formula for root mean square error may be used.

$$m = \sqrt{\frac{\Sigma E^2}{3n}}$$

where m = the root mean square error of unadjusted horizontal angles in seconds of arc as obtained from the triangular errors, ΣE = the sum of the squares of all the triangular errors in the triangulation series, and n =the total number of triangles in the series.

It may be noted that

(i) all the triangles have been included in the computations,

(ii) all the four triangles of a braced quadrilateral have been included in the computations, and

(iii) if the average triangular error of the series is 8", probable error in latitudes and departures after a distance of 100 km, is approximately 8 m

Example .1 If the probable error of direction measurement is 1.20″, compute the maximum value of *R* for the desired maximum probable error of (*i*) 1 in 20,000 and (*ii*) 1 in 10,000.

Solution: (*i*) *L* being the probable error of a logarithm, it represents the logarithm of the ratio of the true value and a value containing the probable error.

In this case $L = \text{the 6th place in } \log\left(1 \pm \frac{1}{20000}\right)$

$= \text{the 6th place in } \log\left(\left(1 \pm 0.00005\right)\right.$

$\log(1+0.00005) = 0.0000217$

The 6th place in the log value $= 21$

Hence $L = \pm 21$

It is given that $d = 1.20''$

From Eq. (9), we have

$$L^2 = \frac{4}{3} d^2 R$$

$$R_{\max} = \frac{3}{4}\frac{L^2}{d^2}$$

$$= \frac{3}{4} \times \frac{21^2}{1.20^2} = \mathbf{230}.$$

(*ii*) $L = \text{the 6th place in } \log\left(1 \pm \frac{1}{10000}\right)$

$\log(1+0.0001) = 0.0000434$

The 6th place in the log value $= 43$

Hence $L = \pm 43$

$$R_{\max} = \frac{3}{4} \times \frac{43^2}{1.20^2} = \mathbf{963}.$$

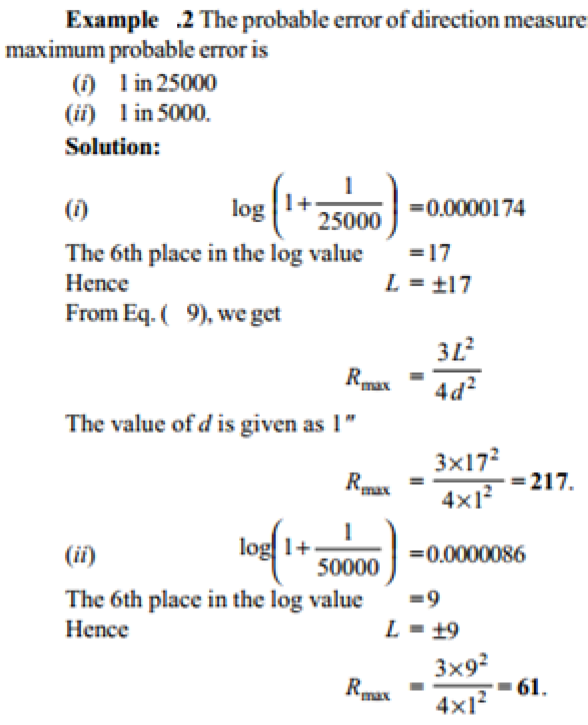

Example .2 The probable error of direction measurement is 1″. Compute the maximum value of R if the maximum probable error is

(*i*) 1 in 25000

(*ii*) 1 in 5000.

Solution:

(*i*) $$\log\left(1+\frac{1}{25000}\right) = 0.0000174$$

The 6th place in the log value $= 17$

Hence $L = \pm 17$

From Eq. (9), we get

$$R_{\max} = \frac{3L^2}{4d^2}$$

The value of d is given as 1″

$$R_{\max} = \frac{3\times17^2}{4\times1^2} = \mathbf{217}.$$

(*ii*) $$\log\left(1+\frac{1}{50000}\right) = 0.0000086$$

The 6th place in the log value $= 9$

Hence $L = \pm 9$

$$R_{\max} = \frac{3\times9^2}{4\times1^2} = \mathbf{61}.$$

Example 1.5Compute the strength of the figure ABCD for all the routes by which the length CD can be computed from the known side AB. Assume that all the stations were occupied.

Solution:

we have

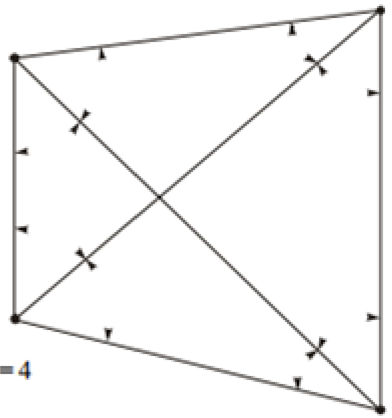

$$R = \frac{D-C}{D}\Sigma(\delta_A^2 + \delta_A\delta_B + \delta_C^2)$$

For the given figure in Fig. 1.16, we have

$$n = 6$$
$$n' = 6$$
$$S = 4$$
$$S' = 4$$
$$D = 2 \times (n-1)$$
$$= 2 \times (6-1) = 10$$

Hence $C = (n' - S' + 1) + (n - 2S + 3)$

$$= (6-4+1) + (6 - 2\times 4 + 3) = 4$$

and $\frac{D-C}{D} = \frac{10-4}{10} = \mathbf{0.60}$.

(*a*) Route-1, using $\Delta^s ABC$ and ADC with common side AC

For ΔABC the distance angles of AB and AC are 26° and 100° = 44° + 56°, respectively.

From Table 1.2,

$$\delta_{100}^2 + \delta_{100}\delta_{26} + \delta_{26}^2 = 17$$

For ΔADC, the distance angles of AC and DC are 112° = (44° + 68°) and 38°, respectively,

$$\delta_{112}^2 + \delta_{112}\delta_{38} + \delta_{38}^2 = 6$$

$$R_1 = 0.6 \times (17+6) = 13.8 \simeq 14$$

(*b*) Route-2, using $\Delta^s ABC$ and BCD with common side BC

For ΔABC the distance angles of AB and BC are 26° and 54°, respectively,

$$\delta_{54}^2 + \delta_{54}\delta_{26} + \delta_{26}^2 = 27$$

For ΔBCD, the distance angle of BC and CD are 68° and 56°, respectively,

$$\delta_{68}^2 + \delta_{68}\delta_{56} + \delta_{56}^2 = 4$$

$$R_2 = 0.6 \times (27+4) = 18.6 \simeq 19$$

(*c*) Route-3, using $\Delta^s ABD$ and ACD with common side AD

From ΔABC the distance for both the sides AB and AD is 44°.

$$\delta_{44}^2 + \delta_{44}\delta_{44} + \delta_{44}^2 = 13$$

From ΔACD, the distance angles of AD and CD and 30° and 38°, respectively,

$$\delta_{38}^2 + \delta_{38}\delta_{30} + \delta_{30}^2 = 31$$

$$R_3 = 0.6 \times (13+31) = 26.4 \simeq 26$$

(*d*) Route-4, using $\Delta^s ABD$ and BCD with common side BD.

From ΔABD, the distance angles of AB and DB are 44° and 92° = (38° + 54°), respectively,

$$\delta_{92}^2 + \delta_{92}\delta_{38} + \delta_{38}^2 = 7$$

From ΔBCD, the distance angles of BD and CD are 56° = (30° + 26°) and 56°, respectively,

$$\delta_{56}^2\delta_{56}\delta_{56} + \delta_{56}^2 = 7$$

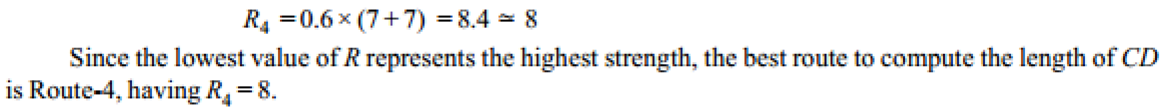

$$R_4 = 0.6 \times (7+7) = 8.4 \simeq 8$$

Since the lowest value of R represents the highest strength, the best route to compute the length of CD is Route-4, having $R_4 = 8$.

4.6 ROUTINE OF TRIANGULATION SURVEY

The routine of triangulation survey, broadly consists of (a) field work, and (b) computations.

The field work of triangulation is divided into the following operations :

(i) Reconnaissance

(ii) Erection of signals and towers

(iii) Measurement of base line

(iv) Measurement of horizontal angles

(v) Measurement of vertical angles

(vi) Astronomical observations to determine the azimuth of the lines

4.7 RECONNAISSANCE

Reconnaissance is the preliminary field inspection of the entire area to be covered by triangulation, and collection of relevant data. Since the basic principle of survey is working from whole to the part, reconnaissance is very important in all types of surveys. It requires great skill, experience and judgement. The accuracy and economy of triangulation greatly depends upon proper reconnaissance survey. It includes the following operations:

1. Examination of terrain to be surveyed.

2. Selection of suitable sites for measurement of base lines.

3. Selection of suitable positions for triangulation stations.

4. Determination of intervisibility of triangulation stations.

5. Selection of conspicuous well-defined natural points to be used as intersected points.

6. Collection of miscellaneous information regarding:

(a) Access to various triangulation stations

(b) Transport facilities

(c) Availability of food, water, etc.

(d) Availability of labour

(e) Camping ground.

Reconnaissance may be effectively carried out if accurate topographical maps of the area are available. Help of aerial photographs and mosaics, if available, is also taken. If maps and aerial photographs are not available, a rapid preliminary reconnaissance is undertaken to ascertain the general location of possible schemes of triangulation suitable for the topography. Later on, main reconnaissance is done to examine these schemes. The main reconnaissance is a very rough triangulation. The plotting of the rough triangulation may be done by protracting the angles. The essential features of the topography are also sketched in. The final scheme is selected by studying the relative strengths and cost to various schemes.

For reconnaissance the following instruments are generally employed:

1. Small theodolite and sextant for measurement of angles.

2. Prismatic compass for measurement of bearings.

3. Steel tape.

4. Aneroid barometer for ascertaining elevations.

5. Heliotropes for ascertaining intervisibility.

6. Binocular.

7. Drawing instruments and material

8. Guyed ladders, creepers, ropes, etc., for climbing trees.

4.8 Erection of signals and towers

A signal is a device erected to define the exact position of a triangulation station so that it can be observed from other stations whereas a tower is a structure over a station to support the instrument and the observer, and is provided when the station or the signal, or both are to be elevated. Before deciding the type of signal to be used, the triangulation stations are selected. The selection of triangulation stations is based upon the following criteria.

Criteria for selection of triangulation stations

1. Triangulation stations should be intervisible. For this purpose the station points should be on the highest ground such as hill tops, house tops, etc.

2. Stations should be easily accessible with instruments.

3. Station should form well-conditioned triangles.

4. Stations should be so located that the lengths of sights are neither too small nor too long. Small sights cause errors of bisection and centering. Long sights too cause direction error as the signals become too indistinct for accurate bisection.

5. Stations should be at commanding positions so as to serve as control for subsidiary triangulation, and for possible extension of the main triangulation scheme.

6. Stations should be useful for providing intersected points and also for detail survey.

7. In wooded country, the stations should be selected such that the cost of clearing and cutting, and building towers, is minimum.

8. Grazing line of sights should be avoided, and no line of sight should pass over the industrial areas to avoid irregular atmospheric refraction.

Determination of intervisibility of triangulation stations

As stated above, triangulations stations should be chosen on high ground so that all relevant stations are intervisible. For small distances, intervisibility can be ascertained during reconnaissance by direct observation with the aid of binocular, contoured map of the area, plane mirrors or heliotropes using reflected sun rays from either station. However, if the distance between stations is large, the intervisibility is ascertained by knowing the horizontal distance between the stations as under.

Case-I Invervisibility not obstructed by intervening ground

If the intervening ground does not obstruct the intervisibility, the distance of visible horizon from the station of known elevation is calculated from the following formula:

$$h = \frac{D^2}{2R}(1-2m)$$

where h = height of the station above datum,

D = distance of visible horizon,

R = earth's mean radius, and

m = mean coefficient of refraction taken as 0.07 for sights over land, and. 0.08 for sights over sea.

Substituting the values of m as 0.071 and R as 6370 km , the value of h in metres is given by

$$h = 0.06735\ D^2$$

where D is in kilometres.

the distance between two stations A and B of heights h_A and h_B , respectively, is D. If D_A and D_B are the distances of visible horizon from A and B, respectively, we have

$$D_A = \sqrt{\frac{h_A}{0.06735}} = 3.853\ \sqrt{h_A}$$

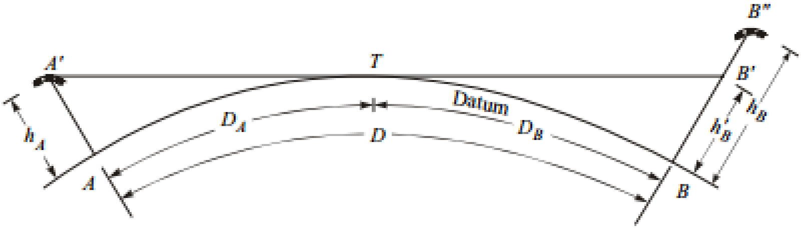

Fig. 1.17 Intervisibility not obstructed by intervening ground

We have $D = D_A + D_B$

or $D_B = D - D_A$

For the known distance of visible horizon D_B as above, the height of station B is computed. If the computed value is h'_B, then

$$h'_B = 0.06735\, D_B^2$$

The computed value of height h'_B is compared with the known value h_B as below :

If $h_B \geq h'_B$, the station B will be visible from A, and

if $h_B < h'_B$, the station B will not be visible from A.

If B is not visible from A, $(h'_B - h_B)$ is the required amount of height of signal to be erected at B. While deciding the intervisibility of various stations, the line of sight should be taken at least 3 m above the point of tangency T of the earth's surface to avoid grazing rays.

Case-II Intervisibility obstructed by intervening ground

In Fig. 1.18, the intervening ground at *C* is obstructing the intervisibility between the stations *A* and *B*.

$$D_A = 3.853\sqrt{h_A}$$

The distance D_T of the peak *C* from the point of tangency *T*, is given by

$$D_T = D_A - D_C$$

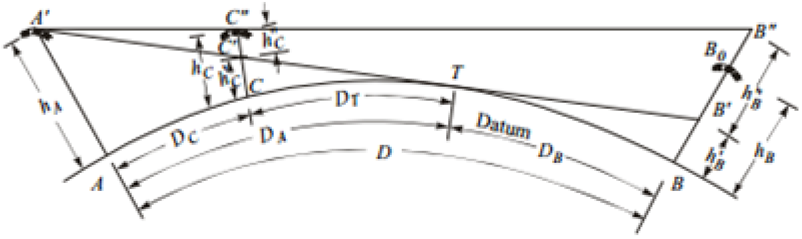

Intervisibility obstructed by intervening ground

and
$$h'_C = 0.06735\, D_T^{\ 2}$$
$$h'_B = 0.06735\, D_B^{\ 2}$$

If $h'_C > h_C$, the line of sight is clear of the obstruction, and it becomes Case-I discussed above. If $h'_C < h_C$ then the signal at *B* is to be raised. The amount of raising required at *B* is computed as below.

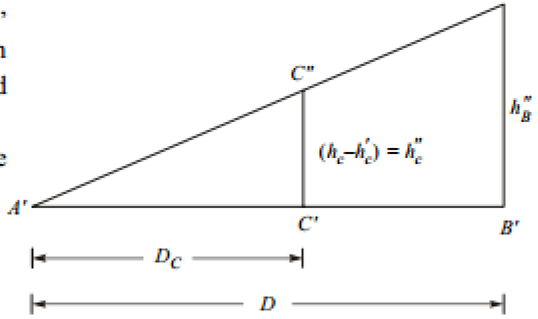

From similar $\Delta^s\, A'C'C''$ and $A'B'B''$ in Fig. , we get

$$\frac{h''_C}{D_C} = \frac{h''_B}{D}$$

or
$$h''_B = \frac{D}{D_C}\, h''_C \qquad ..$$

where
$$h''_C = h_C - h'_C\,.$$

The required height of signal above station B_0 is

$$B_0B'' = (BB' + B'B'') - BB_0$$
$$= (h'_B + h''_{C'}) - h_B$$

Alternate method (Captain G.T. McCaw's method)

A comparison of elevations of the stations *A* and *B* (Fig. 1.20) decides whether the triangulation stations are intervisible or not. A direct solution suggested by Captain McCaw is known as Captain McCaw's method.

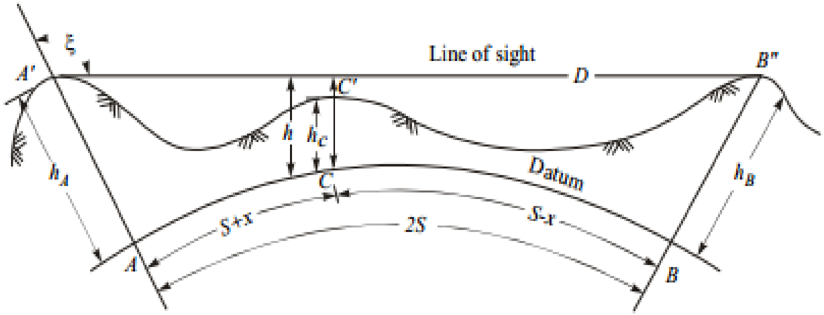

Fig. 1.20 Captain McCaw's method of ascertaining intervisibility

Let h_A = elevation of station *A*

h_B = elevation of station *B*

h_C = elevation of station *C*.

$2S$ = distance between *A* and *B*

$(S+x)$ = distance between *A* and *C*

$(S-x)$ = distance between *C* and *B*

h = elevation of the line of sight at *C*

ξ = zenith distance from *A* to *B*

= (90°-vertical angle).

From Captain McCaw's formula

$$h = \frac{1}{2}(h_3 + h_A) + \frac{1}{2}(h_B - h_A)\frac{x}{S} - (S^2 - x^2)\operatorname{cosec}^2\xi\frac{(1-2m)}{2R}$$

Practically in most of the cases, the zenith distance is very nearly equal to 90° and, therefore, the value of cosec² ξ may be taken approximately equal to unity.

However, for accurate calculations,

$$\operatorname{cosec}^2\xi = 1 + \frac{(h_B - h_A)^2}{4S^2}$$

In Eq. (1.23), the value of $\left(\frac{1-2m}{2R}\right)$ is usually taken as 0.06735.

Therefore

$$h = \frac{1}{2}(h_B + h_A) + \frac{1}{2}(h_B - h_A)\frac{x}{S} - (S^2 - x^2) \times 0.06735$$

If $h > h_C$, the line of sight is free of obstruction. In case $h < h_C$, the height of tower to raise the signal at *B*, is computed

$$B_0B'' = (BB' + B'B'') - BB_0$$
$$= (h'_B + h''_C) - h_B$$

Example 6 Two stations *A* and *B*, 80 km apart, have elevations 15 m and 270 m above mean sea level, respectively. Calculate the minimum height of the signal at *B*.

It is given that

$$h_A = 15\,\mathrm{m}$$
$$h_B = 270\,\mathrm{m}$$
$$D = 80\,\mathrm{km}$$

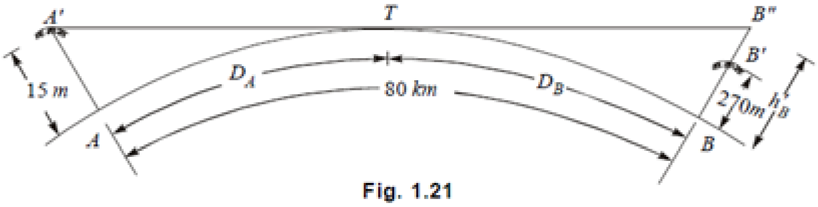

Fig. 1.21

we get

$$D_A = 3.853\sqrt{h_A} = 3.853 \times \sqrt{15} = 14.92\,\mathrm{km}$$

We have

$$D_B = D - D_A$$

or

$$= 80 - 14.92$$
$$= 65.08\,\mathrm{km}$$

$$= (h'_B + h''_C) - h_B$$

Therefore

$$h'_B = 0.06735\ D_B^2$$
$$= 0.06735 \times 65.08^2 = 285.25\,\mathrm{m}$$

Hence, since the elevation of *B* is 270 m, the height of signal required at *B*, is

$$= 285.25 - 270 = 15.25 \simeq \mathbf{15.5\,m}.$$

Example 7 There are two stations *P* and *Q* at elevations of 200 m and 995 m, respectively. The distance of *Q* from *P* is 105 km. If the elevation of a peak *M* at a distance of *38* km from *P* is 301 m, determine whether *Q* is visible from *P* or not. If not, what would be the height of scaffolding required at *Q* so that *Q* becomes visible from *P* ?

we get

$$PT = 3.853 \times \sqrt{200} = 54.45 \text{ km}$$

Therefore
$$MT = PT - PM$$
$$= 54.45 - 38 = 16.45 \text{ km}$$

the value of *MT*, we get

$$MM' = 0.06735 \times 16.45^2 = 18.23 \text{ m}$$

The distance of *Q* from the point of tangency *T* is

$$QT = 105 - 54.45 = 50.55 \text{ km}$$

Therefore
$$QQ' = 0.06735 \times 50.55^2 = 172.10 \text{ m}$$

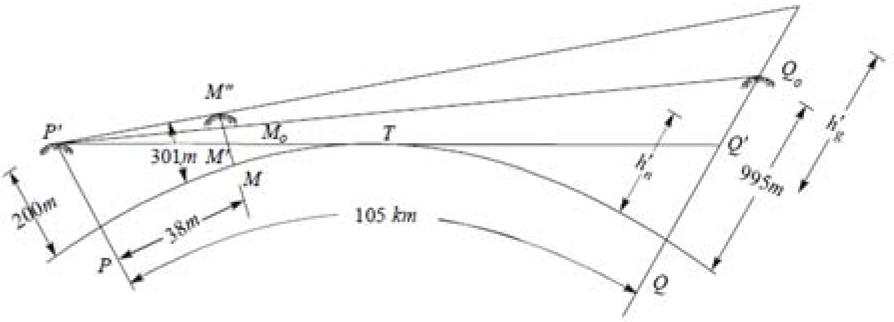
M''
P'
M_o
T
Q_o
h'_g
h'_n
Q'
301m
M'
M
200m
38m
105 km
995m
P
Q

From similar $\Delta^s P'M'M''$ and $P'Q'Q''$, we have

$$\frac{M'M''}{PM} = \frac{Q'Q''}{PQ}$$

$$Q'Q'' = \frac{PQ}{PM} M'M''$$

$$= \frac{PQ}{PM} (MM'' - MM')$$

From similar $\Delta^s P'M'M''$ and $P'Q'Q''$, we have

$$\frac{M'M''}{PM} = \frac{Q'Q''}{PQ}$$

$$Q'Q'' = \frac{PQ}{PM} M'M''$$

$$= \frac{PQ}{PM} (MM'' - MM')$$

$$= \frac{105}{38} \times (301 - 18.23) = 781.34 \text{ m}$$

We have

$$QQ'' = QQ' + Q'Q''$$

$$= 172.10 + 781.34 = 953.44 \text{ m}$$

As the elevation 995 m of Q is more than 953.44 m, the peak at M does not obstruct the line of sight.

Alternatively, from the similar $\Delta^s P'M'M_o$ and $PQ'Q_o$, we have

$$\frac{M'M_o}{PM} = \frac{Q'Q_o}{PQ}$$

$$= \frac{PM}{PQ} (QQ_o - QQ')$$

$$= \frac{38}{105} \times (995 - 172.10) = 297.81$$

The elevation of line of sight $P'Q_o$ at M is

$$MM_o = MM' + M'M_o$$

$$= 18.23 + 297.81 = 316.04.$$

Since the elevation of peak at M is 301 m, the line of sight is not obstructed by the peak and, therefore, no scaffolding is required at Q.

Example 8 Solve the problem given in Example 7 by Capt. McCaw's method.

Solution:

the elevation of line of sight at M joining the two stations is

$$h = \frac{1}{2}(h_Q + h_P) + \frac{1}{2}(h_Q - h_P)\frac{x}{S} - (S^2 - x^2) \times 0.06735$$

It is given that

$$h_P = 200 \text{ m}$$

$$\begin{aligned} h_Q &= 995\text{ m} \\ h_M &= 301\text{ m} \\ 2S &= 105\text{ km or } S = 52.5\text{ km} \\ S + x &= 38\text{ km or } x = -14.5\text{ km} \end{aligned}$$

Therefore

$$\begin{aligned} h &= \frac{1}{2}\times(995+200)+\frac{1}{2}\times(995-200)\times\frac{(-14.5)}{52.5} \\ &\quad -(52.5^2-14.5^2)\times 0.06735 \\ &= \mathbf{316.24\ m}. \end{aligned}$$

The elevation of the line of sight $p'Q_0$ at M is 316.24 m, and the elevation of the peak is 301 m, therefore, the line of sight is clear of obstruction.

Example 9 In a triangulation survey, the altitudes of two proposed stations *A* and *B*, 100 km apart, are respectively 425 m and 750 m. The intervening ground situated at *C*, 60 km from *A*, has an elevation of 435 m. Ascertain if *A* and *B* are intervisible, and if necessary find by how much *B* should be raised so that the line of sight must nowhere be less than 3 m above the surface of the ground. Take *R* = 6400 km and *m* = 0.07.

From the given data we have

$$\begin{aligned} h_A &= 425\text{ m},\ h_B = 750\text{ m},\ h_C = 435\text{ m},\ R = 6400\text{ km},\ m = 0.07 \\ 2S &= 100\text{ km, or } S = 50\text{ km} \\ S + x &= 60\text{ km or } x = 10\text{ km} \end{aligned}$$

$$h'_C = \frac{1}{2}(h_B+h_A)+\frac{1}{2}(h_B-h_A)\frac{x}{S}-(S^2-x^2)\operatorname{cosec}^2\xi\,\frac{(1-2m)}{2R}$$

Taking $\operatorname{cosec}^2\xi = 1$, and substituting the values of the given data in the above equation, we have

$$\begin{aligned} h &= \frac{1}{2}\times(705+425)+\frac{1}{2}\times(705-425)\times\frac{10}{50}-(50^2-10^2) \\ &\quad \times 1\times\frac{(1-2\times 0.07)}{2\times 6400}\times 1000 = 431.75\text{ m} \end{aligned}$$

As the elevation of the line of sight at *C* is less than the elevation of *C*, the line of sight fails to clear *C* by

$$435-431.75 = 3.25\text{ m}$$

To avoid grazing rays, the line of should be at least 3m above the ground. Therefore, the line of sight should be raised to $3.25 + 3 = 6.25$ m at *C*.

Hence, the minimum height of signal to be erected at *B*

$$= \frac{6.25}{60}\times 100 = \mathbf{10.42\ m}.$$

4.9 Station Mark

The triangulation stations should be permanently marked on the ground so that the theodolite and signal may be centered accurately over them. The following points should be considered while marking the exact position of a triangulation station :

(*i*) The station should be marked on perfectly stable foundation or rock. The station mark on a large size rock is generally preferred so that the theodolite and observer can stand on it. Generally, a hole 10 to 15 cm deep is made in the rock and a copper or iron bolt is fixed with cement.

(*ii*) If no rock is available, a large stone is embeded about 1 m deep into the ground with a circle, and dot cut on it. A second stone with a circle and dot is placed vertically above the first stone.

(*iii*) A G.I. pipe of about 25 cm diameter driven vertically into ground up to a depth of one metre, also served as a good station mark.

(*iv*) The mark may be set on a concrete monument. The station should be marked with a copper or bronze tablet. The name of the station and the date on which it was set, should be stamped on the tablet.

(*v*) In earth, generally two marks are set, one about 75 cm below the surface of the ground, and the other extending a few centimeters above the surface of the ground. The underground mark may consist of a stone with a copper bolt in the centre, or a concrete monument with a tablet mark set on it

(*vi*) The station mark with a vertical pole placed centrally, should be covered with a conical heap of stones placed symmetrically. This

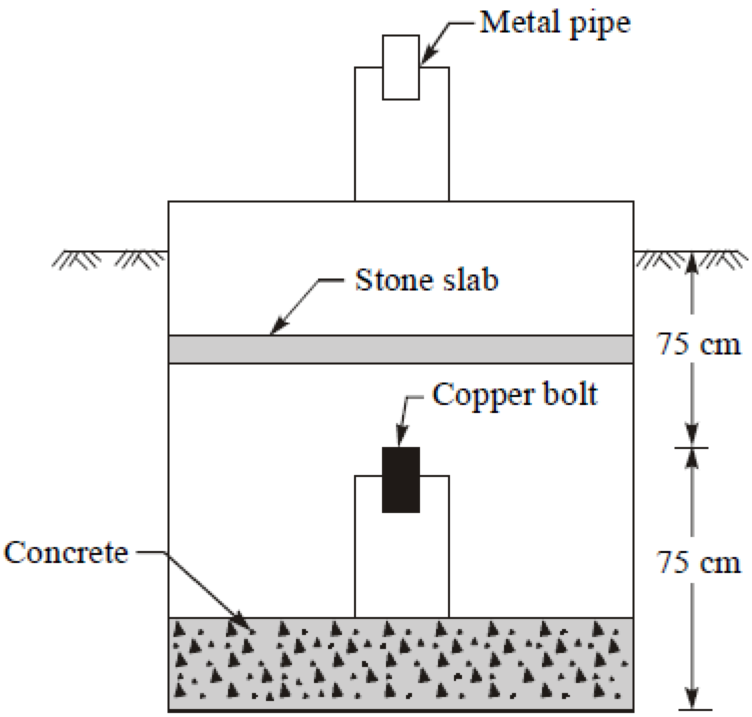

Station mark

4.10 SIGNALS

Signals are centered vertically over the station mark, and the observations are made to these signals from other stations. The accuracy of triangulation is entirely dependent on the degree of accuracy of centering the signals. Therefore, it is very essential that the signals are truly vertical, and centered over the station mark. Greatest care of centering the transit over the station mark will be useless, unless some degree of care in centering the signal is impressed upon.

A signal should fulfil the following requirements :
(*i*) It should be conspicuous and clearly visible against any background. To make the signal conspicuous, it should be kept at least 75 cm above the station mark.

(*ii*) It should be capable of being accurately centered over the station mark.

(*iii*) It should be suitable for accurate bisection from other stations.

(*iv*) It should be free from phase, or should exhibit little phase

Classification of signals
The signals may be classified as under :
(*i*) Non-luminous, opaque or daylight signals
(*ii*) Luminous signals.

(*i*) Non-luminous signals

Non-luminous signals are used during day time and for short distances. These are of various types, and the most commonly used are of following types.

(*a*) **Pole signal**: It consists of a round pole painted black and white in alternate strips, and is supported vertically over the station mark, generally on a tripod. Pole signals are suitable upto a distance of about 6 km.

(*b*) **Target signal**: It consists of a pole carrying two squares or rectangular targets placed at right angles to each other. The targets are generally made of cloth stretched on wooden frames. Target signals are suitable upto a distance of 30 km.

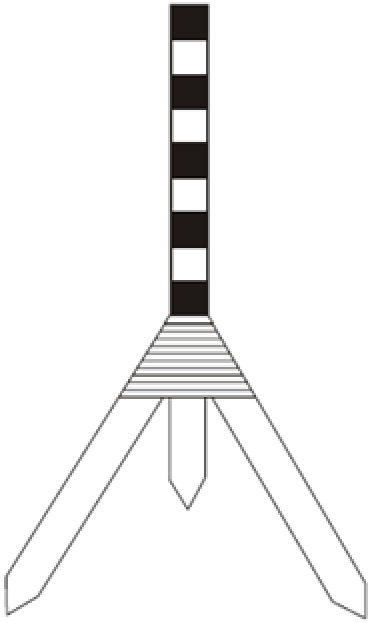

Pole signal

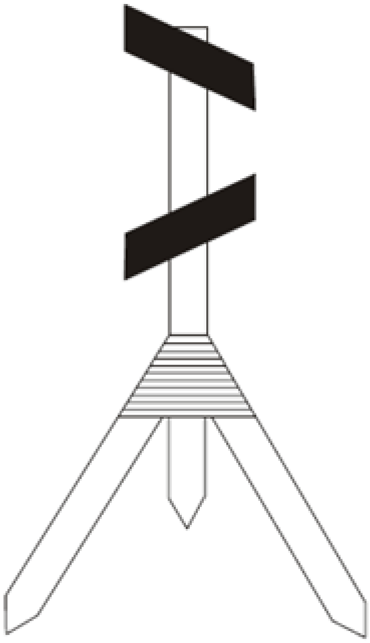

Target signal

(*c*) **Pole and brush signal** It consists of a straight pole about 2.5 m long with a bunch of long grass tied symmetrically round the top making a cross. The signal is erected vertically over the station mark by heaping a pile of stones, upto 1.7 m round the pole. A rough coat of white wash is given to make it more conspicuous to be seen against black background. These signals are very useful, and must be erected over every station of observation during reconnaissance.

(*d*) **Stone cairn**: A pile of stone heaped in a conical shape about 3 m high with a cross shape signal erected over the stone heap, is stone cairn. This white washed opaque signal is very useful if the background is dark.

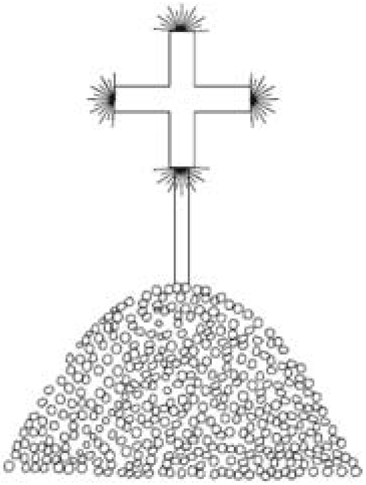

Pole and brush signal

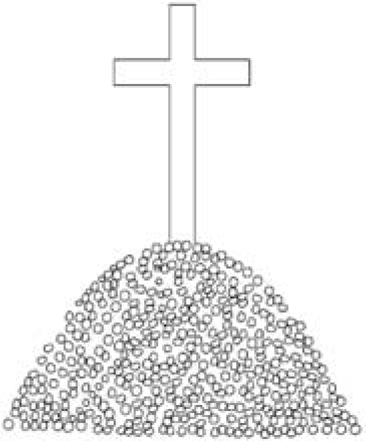

Stone cairn

(*e*) **Beacons**: It consists of red and white cloth tied round the three straight poles. The beacon can easily be centered over the station mark. It is very useful for making simultaneous observations.

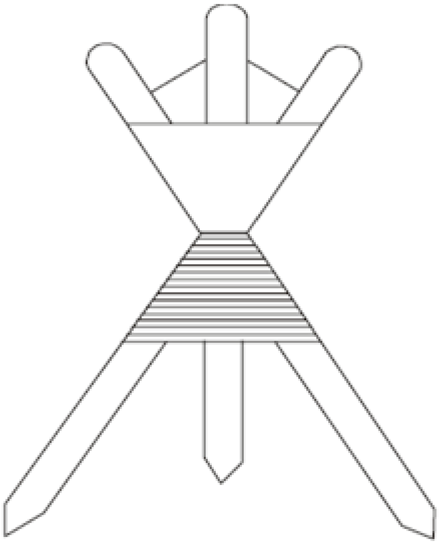

Beacon

***(ii)* Luminous signals** Luminous signals may be classified into two types :
(*i*) Sun signals
(*ii*) Night signals.

(*a*) **Sun signals**: Sun signals reflect the rays of the sun towards the station of observation, and are also known as heliotropes. Such signals can be used only in day time in clear weather.

Heliotrope : It consists of a circular plane mirror with a small hole at its centre to reflect the sun rays, and a sight vane with an aperture carrying a cross-hairs. The circular mirror can be rotated horizontally as well as vertically through 360°. The heliotrope is centered over the station mark, and the line of sight is directed towards the station of observation. The sight vane is adjusted looking through the hole till the flashes given from the station of observation fall at the centre of the cross of the sight vane. Once this is achieved, the heliotrope is disturbed. Now the heliotrope frame carrying the mirror is rotated in such a way that the black shadow of the small central hole of the plane mirror falls exactly at the cross of the sight vane. By doing so, the reflected beam of rays will be seen at the station of observation. Due to motion of the sun, this small shadown also moves, and it should be constantly ensured that the shadow always remains at the cross till the observations are over.

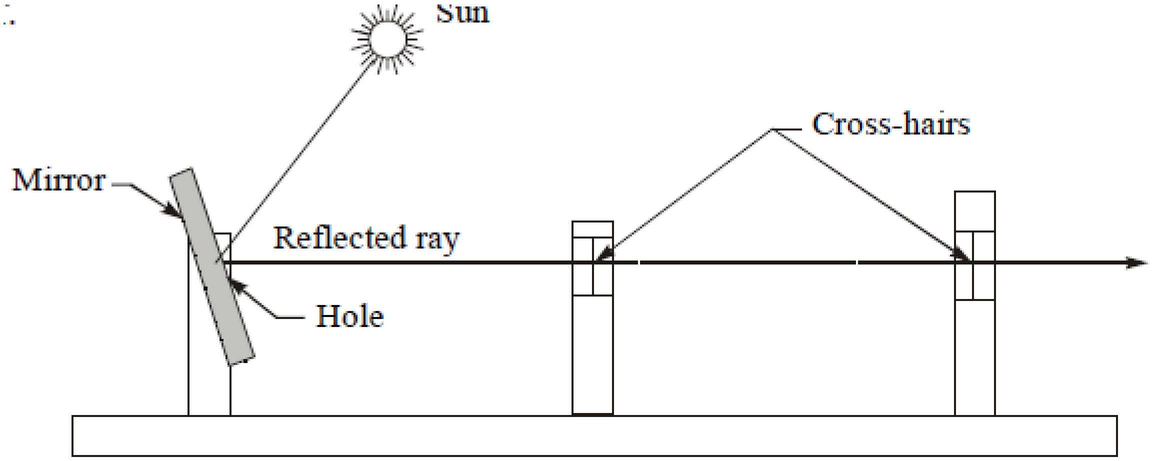

Heliotrope

The heliotropes do not give better results compared to signals. These are useful when the signal station is in flat plane, and the station of observation is on elevated ground. When the distance between the stations exceed 30 km, the heliotropes become very useful.

(b) Night signals: When the observations are required to be made at night, the night signals of following types may be used.

1. Various forms of oil lamps with parabolic reflectors for sights less than 80 km.
2. Acetylene lamp designed by Capt. McCaw for sights more than 80 km.
3. Magnesium lamp with parabolic reflectors for long sights.
4. Drummond’s light consisting of a small ball of lime placed at the focus of the parabolic reflector, and raised to a very high temperature by impinging on it a stream of oxygen.

5. Electric lamps.

4.11 TOWERS

A tower is erected at the triangulation station when the station or the signal or both are to be elevated to make the observations possible form other stations in case of problem of intervisibility. The height of tower depends upon the character of the terrain and the length of the sight. The towers generally have two independent structures. The outer structure is for supporting the observer and the signal whereas the inner one is for supporting the instrument only. The two structures are made entirely independent of each other so that the movement of the observer does not disturb the instrument setting. The two towers may be made of masonary, timber or steel. For small heights, masonary towers are most suitable. Timber scaffolds are most commonly used, and have been constructed to heights over 50 m. Steel towers made of light sections are very portable, and can be easily erected and dismantled. Bilby towers patented by J.S. Bilby of the U.S. Coast and Geodetic Survey, are popular for heights ranging from 30 to 40 m. This tower weighing about 3 tonnes, can be easily erected by five persons in just 5 hrs. A schematic of such a tower is shown in Fig. 1.30.

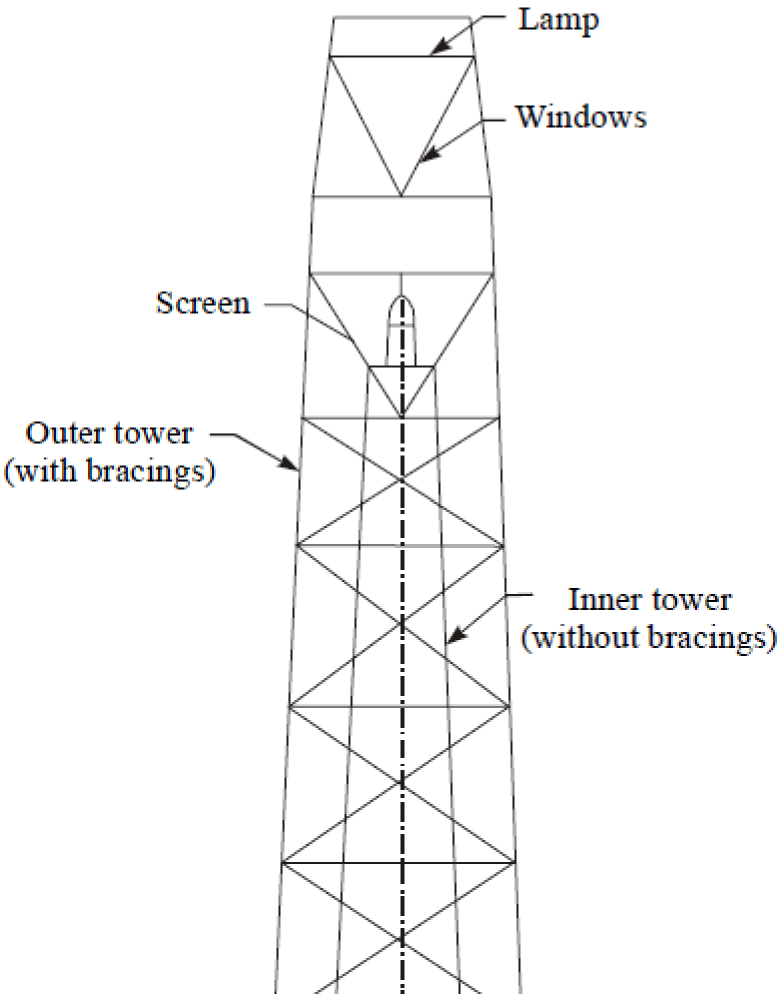

PHASE OF A SIGNAL

When cylindrical opaque signals are used, they require a correction in the observed horizontal angles due an error known as the *phase*. The cylindrical signal is partly illuminated by the sun, and the other part remains in shadow, and becomes invisible to the observer. While making the observations, the observer may bisect the bright portion or the bright line. Thus the signal is not bisected at the centre, and an error due to wrong bisection is introduced. It is, thus, the apparent displacement of the signal. The phase correction is thus necessary so that the observed horizontal angles may be reduced to that corresponding to the centre of the signal.

Depending upon the method of observation, phase correction is computed under the following two conditions. Windows Lamp Screen Outer tower (with bracings) Inner tower (without bracings

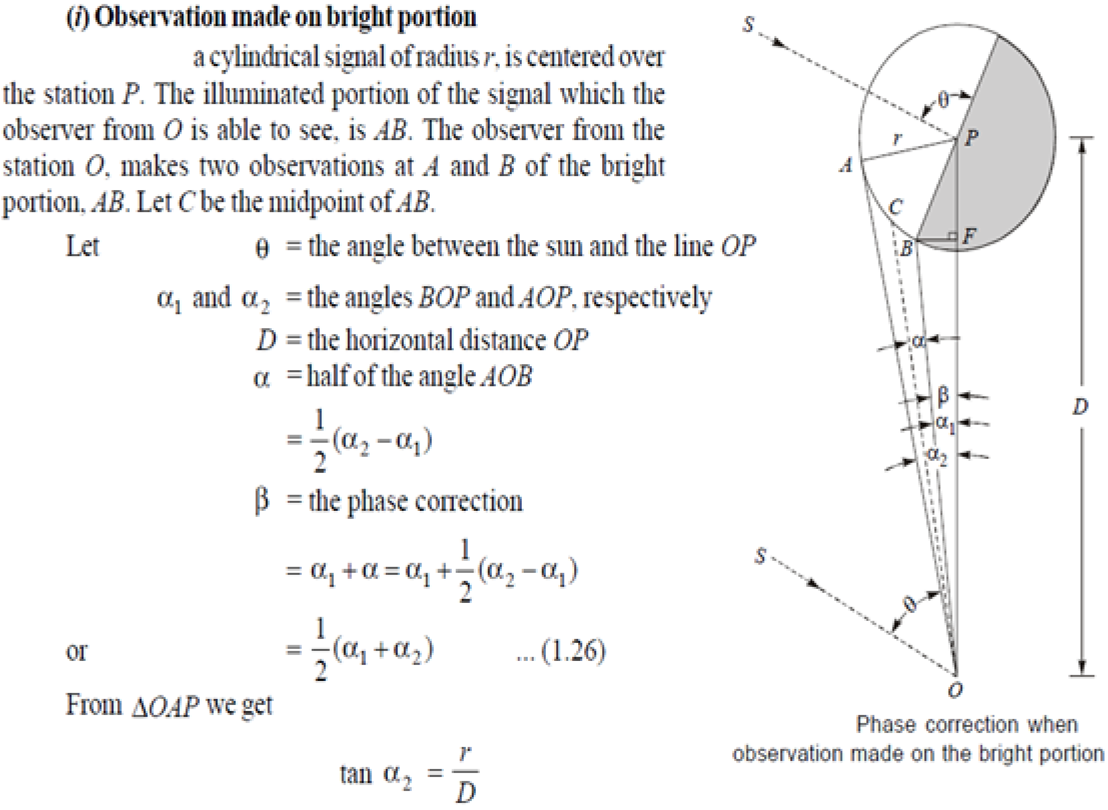

(*i*) Observation made on bright portion

a cylindrical signal of radius r, is centered over the station P. The illuminated portion of the signal which the observer from O is able to see, is AB. The observer from the station O, makes two observations at A and B of the bright portion, AB. Let C be the midpoint of AB.

Let θ = the angle between the sun and the line OP

α_1 and α_2 = the angles BOP and AOP, respectively

D = the horizontal distance OP

α = half of the angle AOB

$$= \frac{1}{2}(\alpha_2 - \alpha_1)$$

β = the phase correction

$$= \alpha_1 + \alpha = \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)$$

or

$$= \frac{1}{2}(\alpha_1 + \alpha_2) \qquad \text{... (1.26)}$$

From ΔOAP we get

$$\tan\alpha_2 = \frac{r}{D}$$

Phase correction when observation made on the bright portion

α_2 being small, we can write

$$\alpha_2 = \frac{r}{D} \text{ radians}$$

As the distance *PF* is very small compared to *OP*, *OF* may be taken as *OP*. Thus, from right angle ΔBFO, we get

$$\tan \alpha_1 = \frac{BF}{OF} = \frac{BF}{OP} = \frac{BF}{D}$$

From ΔPFB, we get

$$BF = r \sin(90 - \theta) = r \cos\theta$$

Substituting the value of *BF* in Eq. (1.28), we get

$$\tan \alpha_1 = \frac{r\cos\theta}{D}$$

α_1 being small, we can write

$$\alpha_1 = \frac{r\cos\theta}{D} \text{ radians}$$

Substituting the values of α_1 and α_2 , we have

$$\alpha_1 = \frac{r\cos\theta}{D} \text{ radians}$$

Substituting the values of α_1 and α_2 we have

$$\beta = \frac{1}{2}\left(\frac{r}{D} + \frac{r\cos\theta}{D}\right) = \frac{r}{D}\left(\frac{1+\cos\theta}{2}\right)$$

$$= \frac{r}{D}\cos^2\frac{\theta}{2} \text{ radians}$$

$$= \frac{r}{D\sin 1''}\cos^2\frac{\theta}{2} \text{ seconds}$$

$$\beta = \frac{206265r}{D}\cos^2\frac{\theta}{2} \text{ seconds}$$

(*ii*) Observations made on the bright line

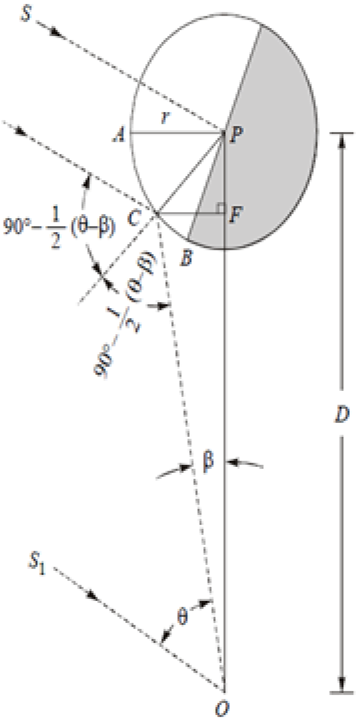
S
S
S$_1$
A
r
P
F
C
B
$90° - \frac{1}{2}(\theta-\beta)$
$90° - \frac{1}{2}(\theta-\beta)$
β
θ
D
O

In this case, the bright line at *C* on the cylindrical signal of radius *r*, is sighted from *O*

Let CO = the reflected ray of the sun from the bright line at *C*

β = the phase correction

θ = the angle between the sun and the line *OP*

The rays of the sun are always parallel to each other, therefore, *SC* is parallel to S_1O.

$$\angle SCO = 180° - (\theta - \beta)$$

$$\angle PCO = 180° - \frac{1}{2}\angle SCO$$

or $$= 180° - \frac{1}{2}[180° - (\theta - \beta)]$$

$$= 90° + \frac{1}{2}(\theta - \beta)$$

Therefore,

$$\angle CPO = 180° - (\beta + \angle PCO)$$

Substituting the value of $\angle PCO$ and after simplification, we get

$$\angle CPO = 90° - \frac{1}{2}(\theta + \beta)$$

As β is very small compared to θ, it can be ignored, Therefore

$$\angle CPO = 90° - \frac{1}{2}\theta$$

From the right angle ΔCFP, we have

$$\frac{CF}{CP} = \sin CPO = \sin\left(90° - \frac{1}{2}\theta\right)$$

or $$CF = r\sin\left(90° - \frac{1}{2}\theta\right)$$

From ΔCFO, we get

$$\tan\beta = \frac{CF}{OF}$$

PF being very small compared to *OP*, *OF* may be taken as *OP*. Substituting the value of *CF* and taking *OF* equal to *D*, we get

$$\tan\beta = \frac{r\sin\left(90^\circ - \frac{1}{2}\theta\right)}{D}$$

From the right angle ΔCFP, we have

$$\frac{CF}{CP} = \sin CPO = \sin\left(90^\circ - \frac{1}{2}\theta\right)$$

or

$$CF = r\sin\left(90^\circ - \frac{1}{2}\theta\right)$$

From ΔCFO, we get

$$\tan\beta = \frac{CF}{OF}$$

or

$$\beta = \frac{r\cos\frac{\theta}{2}}{D} \text{ radians}$$

$$\beta = \frac{206265r}{D}\cos\frac{\theta}{2} \text{ seconds} \quad \ldots (1.36)$$

The phase correction β is applied to the observed horizontal angles in the following manner.

Let there be four stations S_1, S_2, *P*, and *O* as shown in (Fig. 1.33). The observer is at *O*, and the angles S_1OP and POS_2 have been measured from *O* as θ_1' and θ_2', respectively.

If the required corrected angles are θ_1 and θ_2, then

$$\theta_1 = \theta_1' + \beta$$

and

$$\theta_2 = \theta_2' - \beta$$

when β is the phase correction.

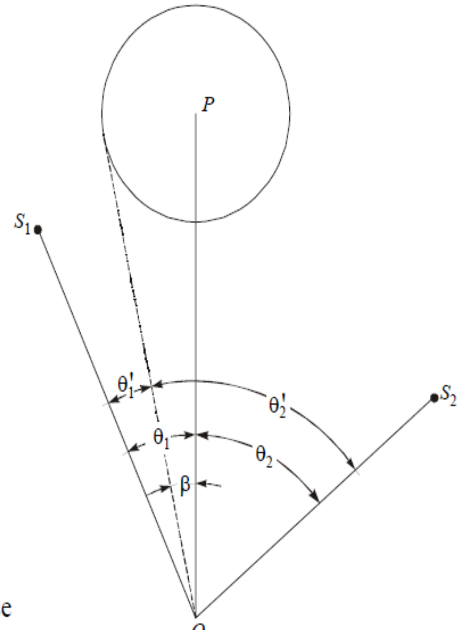

While applying the corrections the directions of the phase or $$\beta = \frac{r\cos\frac{\theta}{2}}{D} \text{ radians}$$

$$\beta = \frac{206265r}{D}\cos\frac{\theta}{2} \text{ seconds}$$

The phase correction β is applied to the observed horizontal angles in the following manner.

Let there be four stations S_1, S_2, P, and O . The observer is at O, and the angles S_1OP and POS_2 have been measured from O as θ_1' and θ_2', respectively.

If the required corrected angles are θ_1 and θ_2, then

$$\theta_1 = \theta_1' + \beta$$

and $$\theta_2 = \theta_2' - \beta$$

when β is the phase correction.

While applying the corrections the directions of the phase correction, and the observed stations with respect to the line OP, must be noted carefully.

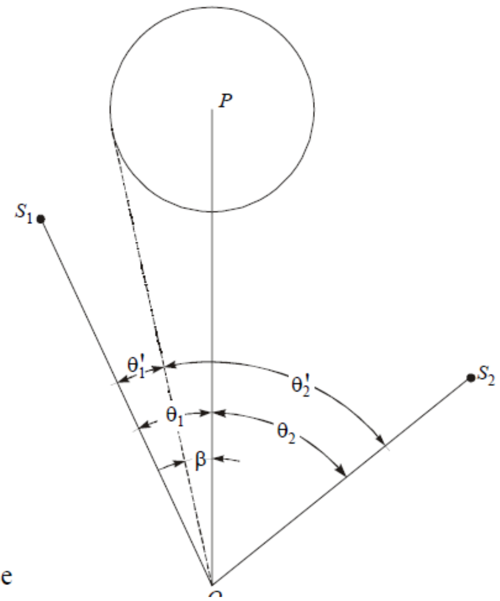

Applying the phase correction to the measured horizontal angles

Example _10 A cylindrical signal of diameter 4 m, was erected at station *B*. Observations were made on the signal from station *A*. Calculate the phase corrections when the observations were made

(*i*) on the bright portion, and
(*ii*) on the bright line.

Take the distance *AB* as 6950 m, and the bearings of the sun and the station *B* as 315° and 35°, respectively.

Solution: Given that θ = Bearing of sun – bearing of B

$$= 315° - 35° = 280°$$

$$r = \frac{\text{Diameter}}{2} = \frac{4}{2} = 2\,\text{m}$$

$$D = 6950\,\text{m}$$

(*i*)

the phase correction

$$\beta = \frac{206265r}{D}\cos^2\frac{\theta}{2} \text{ seconds}$$

$$= \frac{206265 \times 2}{6950} \times \cos^2\frac{280°}{2} = \textbf{34.83 seconds}.$$

(*ii*)

the phase correction

$$\beta = \frac{206265r}{D}\cos\frac{\theta}{2}\ \text{seconds}$$

$$= \frac{206265\times 2}{6950}\times\cos\frac{280^\circ}{2} = \mathbf{45.47\ seconds.}$$

Example 11 The horizontal angle measured between two stations *P* and *Q* at station *R*, was 38°29′30″. The station *Q* is situated on the right of the line *RP*.

The diameter the cylindrical signal erected at station *P*, was 3 m and the distance between *P* and *R* was 5180 m. The bearing of the sun and the station *P* were measured as 60° and 15°, respectively. If the observations were made on the bright line, compute the correct horizontal angle *PRQ*.

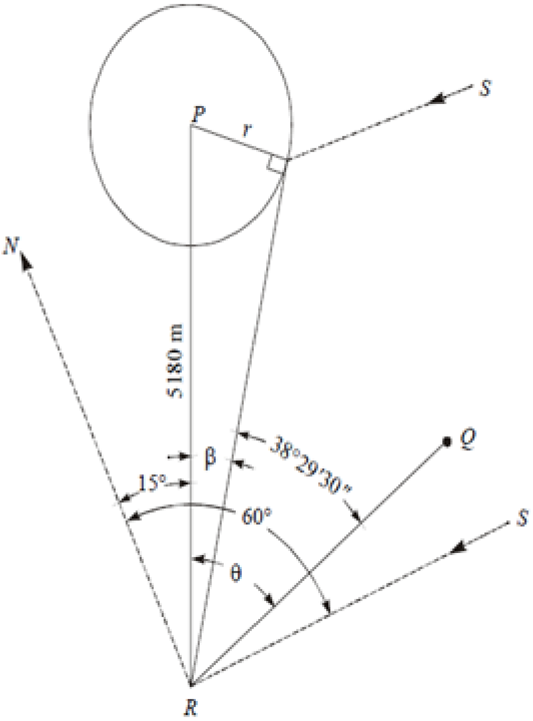

Solution:

From the given data

$$\theta = 60^\circ - 15^\circ = 45^\circ$$
$$D = 5180\text{m}$$
$$r = 1.5\text{m}$$

, we get

$$\beta = \frac{206265r}{D}\cos\frac{\theta}{2}$$

$$= \frac{206265\times 1.5}{5180}\cos\frac{45^\circ}{2}$$

$$= 55.18\ \text{seconds}$$

The correct horizontal angle *PRQ* $= 38^\circ\,29'\,30'' + \beta$

$$= 38^\circ 29' 30'' + 55.18'' = \mathbf{38^\circ 30' 25.18''}.$$

4.11.1 MEASUREMENT OF BASE LINE

The accuracy of an entire triangulation system depends on that attained in the measurement of the base line and, therefore, the measurement of base line forms the most important part of the triangulation operations. As base line forms the basis for computations of triangulation system it is laid down with great accuracy in its measurement and alignment. The length of the base line depends upon the grade of the triangulation. The length of the base is also determined by the desirability of securing strong figures in the base net. Ordinarily the longer base, the easier it will be found to secure strong figures. The base is connected to the triangulation system through a base net. This connection may be made through a simple figure , or through a much more complicated figures discussed in the base line extension. Apart from main base line, several other check bases are also measured at some suitable intervals. In India, ten bases were measured, the length of nine bases vary from 6.4 to 7.8 miles, and that of the tenth base is 1.7 miles.

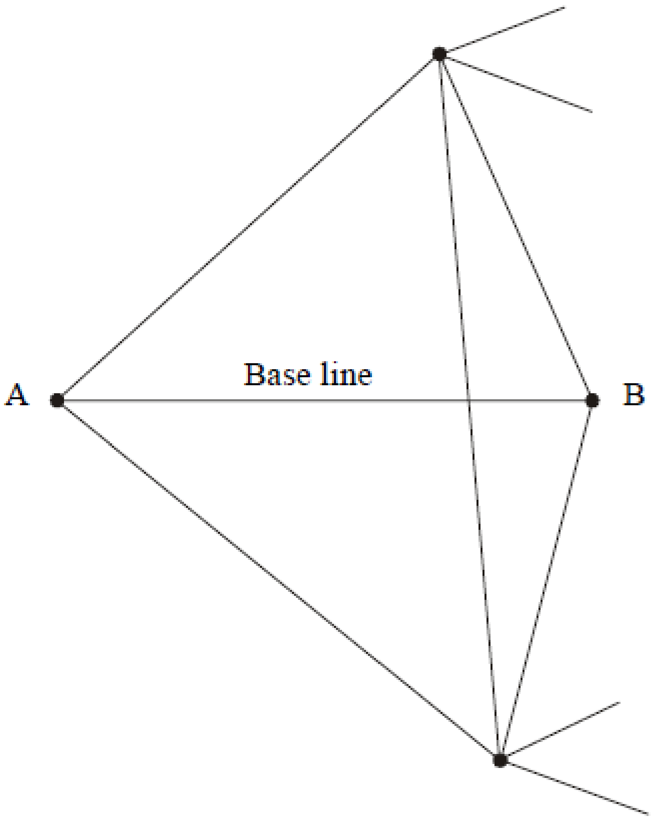

Base net

Selection of site for base line

Since the accuracy in the measurement of the base line depends upon the site conditions, the following points should be taken into consideration while selecting the site for a base line.

1. The site should be fairly level or gently undulating. If the ground is sloping, the slope should be uniform and gentle.
2. The site should be free from obstructions throughout the length of the base line.
3. The ground should be firm and smooth.
4. The two extremities of the base line should be intervisible.
5. The site should be such that well-conditioned triangles can be obtained while connecting extremities to the main triangulation stations.
6. The site should be such that a minimum length of the base line as specified, is available.

Equipment for base line measurement

Generally the following types of base measuring equipments are used :

1. Standardised tapes : These are used for measuring short bases in plain grounds.
2. Hunter’s short base: It is used for measuring 80 m long base line and its extension is made by

subtense method.
3. Tacheometric base measurements : It is used in undulating grounds for small.
.
4. Electronic distance measurement: This is used for fairly long distances and has been discussed in **Standardised tapes :** For measuring short bases in plain areas standardised tapes are generally used. After having measured the length, the correct length of the base is calculated by applying the required corrections. For details of corrections,. If the triangulation system is of extensive nature, the corrected lengths of the base is reduced to the mean sea level.

Hunter's short base :

Dr. Hunter who was a Director of Survey of India, designed an equipment to measure the base line, which was named as Hunter's short base. It consists of four chains, each of 22 yards (20.117 m) linked together. There are 5 stands, three-intermediate two-legged stands, and two three-legged stands at . A 1 kg weight is suspended at the end of an arm, so that the chains remain straight during observations. The correct length of the individual chains is supplied by the manufacturer or is determined in the laboratory. The lengths of the joints between two chains at intermediate supports, are measured directly with the help of a graduated scale. To obtain correct length between the centres of the targets, usual corrections
such as temperature, sag, slope, etc., are applied. To set up of the Hunter's short base the stand at the end *A* (marked in red colour) is centered on the ground mark and the target is fitted with a clip. The target *A* is made truly vertical so that the notch on its tip side is centered on the ground mark. The end of the base is hooked with the plate *A* and is spread carefully till its other end is reached. In between, at every joint of the chains, two-legged supports are fixed to carry the base. The end *B* (marked in green colour) is fixed to the *B* stand and the 1 kg weight is attached at the end of the lever. While fixing the end supports *A* and *B* it should be ensured that their third leg should face each other under the base. Approximate alignment of the base is the done by eye judgement.

For final alignment, a theodolite is set up exactly over the notch of the target *A*, levelled and centered accurately. The target at *B* is then bisected. All intermediate supports are set in line with the vertical cross-hair of the theodolite. At the end again ensure that all the intermediate supports and the target *B* are in one line. In case the base is spread along undulating ground, slope correction is applied. To measure the slope angles of individual supports, a target is fixed to a long iron rod of such a length that it is as high above the tape at *A* as the trunion axis of the theodolite. The rod is held vertically at each support and the vertical angles for each support are read.

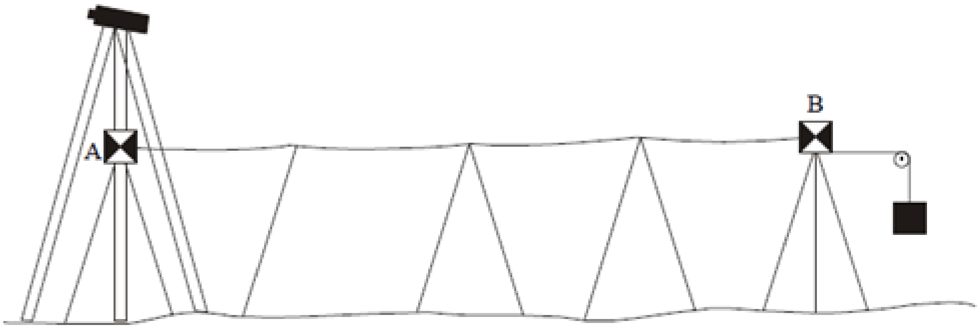

Hunter's short base

Example 12 A tape of standard length 20 m at 85° F was used to measure a base line. The measured distance was 882.10 m. The following being the slopes for the various segments of the line.

Segment	Slope
100 m	2°20′
150 m	4°12′
50 m	1°06′
200 m	7°45′
300 m	3°00′
82.10 m	5°10′

Find the true length of the base line if the mean temperature during measurement was 63°F. The coefficient of expansion of the tape material is 6.5×10^{-6} per °F.

Solution:

Correction for temperature

$$\begin{aligned} C_t &= \alpha(t_m - t_0)L \\ &= 6.5 \times 10^{-6} \times (63 - 65) \times 882.10 \\ &= 0.126 \text{ m (subtractive)} \end{aligned}$$

Correction for slope

$$\begin{aligned} C_s &= \Sigma[(1-\cos\alpha)L] \\ &= (1-\cos 2°20') \times 100 + (1-\cos 4°12') \times 150 + (1-\cos 1°06') \times 50 \\ &\quad + (1-\cos 7°48') \times 200 + (1-\cos 3°00') \times 300 + (1-\cos 5°10') \times 82.10 \\ &= 3.079 \text{ m (subtractive)} \end{aligned}$$

$$\begin{aligned} \text{Total correction} &= C_t + C_s \\ &= 0.126 + 3.079 \\ &= 3.205 \text{ m (subtractive)} \\ \text{Corrected length} &= 882.10 - 3.205 \\ &= \mathbf{878.895\ m}. \end{aligned}$$

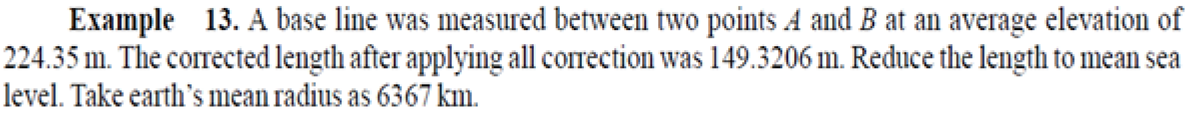

Example 13. A base line was measured between two points *A* and *B* at an average elevation of 224.35 m. The corrected length after applying all correction was 149.3206 m. Reduce the length to mean sea level. Take earth's mean radius as 6367 km.

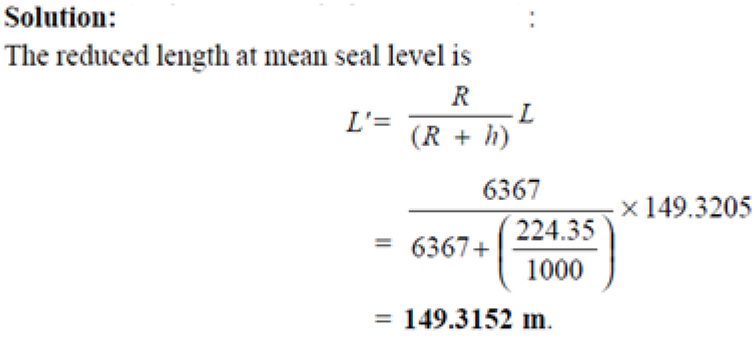

Solution:

The reduced length at mean seal level is

$$L' = \frac{R}{(R + h)} L$$

$$= \frac{6367}{6367 + \left(\frac{224.35}{1000}\right)} \times 149.3205$$

$$= \mathbf{149.3152\ m}.$$

4.11.2 Extension of base line

Usually the length of the base lines is much shorter than the average length of the sides of the triangles.

This is mainly due to the following reasons:
(*a*) It is often not possible to get a suitable site for a longer base.
(*b*) Measurement of a long base line is difficult and expensive.

The extension of short base is done through forming a base net consisting of well-conditioned triangles.There are a great variety of the extension layouts but the following important points should be kept in mind in selecting the one.

(*i*) Small angles opposite the known sides must be avoided.
(*ii*) The length of the base line should be as long as possible.
(*iii*) The length of the base line should be comparable with the mean side length of the triangulation net.
(*iv*) A ratio of base length to the mean side length should be at least 0.5 so as to form well-conditioned triangles.
(*v*) The net should have sufficient redundant lines to provide three or four side equations within the figure.
(*vi*) Subject to the above, it should provide the quickest extension with the fewest stations.

There are two ways of connecting the selected base to the triangulation stations. There are
(*a*) extension by prolongation, and
(*b*) extension by double sighting.

(a) Extension by prolongation

Let up suppose that *AB* is a short base line which is required to be extended by four times. The following steps are involved to extend *AB*.

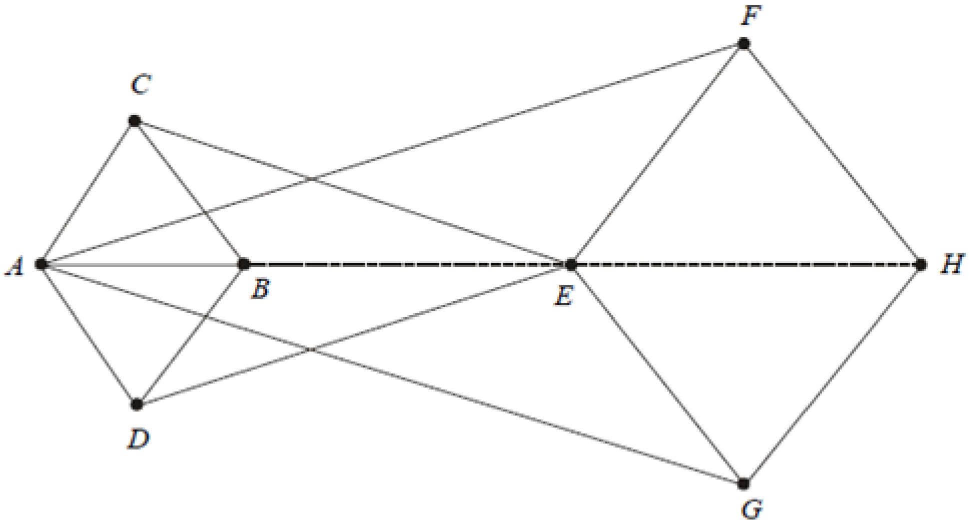

Base extension by prolongation

(*i*) Select *C* and *D* two points on either side of *AB* such that the triangles *BAC* and *BAD* are wellconditioned.

(*ii*) Set up the theodolite over the station *A*, and prolong the line *AB* accurately to a point *E* which is visible from points *C* and *D*, ensuring that triangles *AEC* and *AED* are well-conditioned.

(*iii*) In triangle *ABC*, side *AB* is measured. The length of *AC* and *AD* are computed using the measured angles of the triangles *ABC* and *ABD*, respectively.

(*iv*) The length of *AE* is calculated using the measured angles of triangles *ACE* and *ADE,* and taking mean value.

(*v*) Length of *BE* is also computed in similar manner using the measured angles of the triangles *BEC* and *BDE*. The sum of lengths of *AB* and *BE* should agree with the length of *AE* obtained in step .

(*vi*) If found necessary, the base can be extended to *H* in the similar way.

(*b*) Extension by double sighting
Let *AB* be the base line. To extend the base to the length of side *EF*, following steps are involved.

(*i*) Chose intervisible points *C*, *D*, *E*, and *F*.

(*ii*) Measure all the angles marked in triangles *ABC* and *ABD*. The most probable values of these angles are found by the theory of least-squares discussed in Chapter 2.
(*iii*) Calculate the length of *CD* from these angles and the measured length *AB*, by applying the sine law to triangles *ACB* and *ADB* first, and then to triangles *ADC* and *BDC*

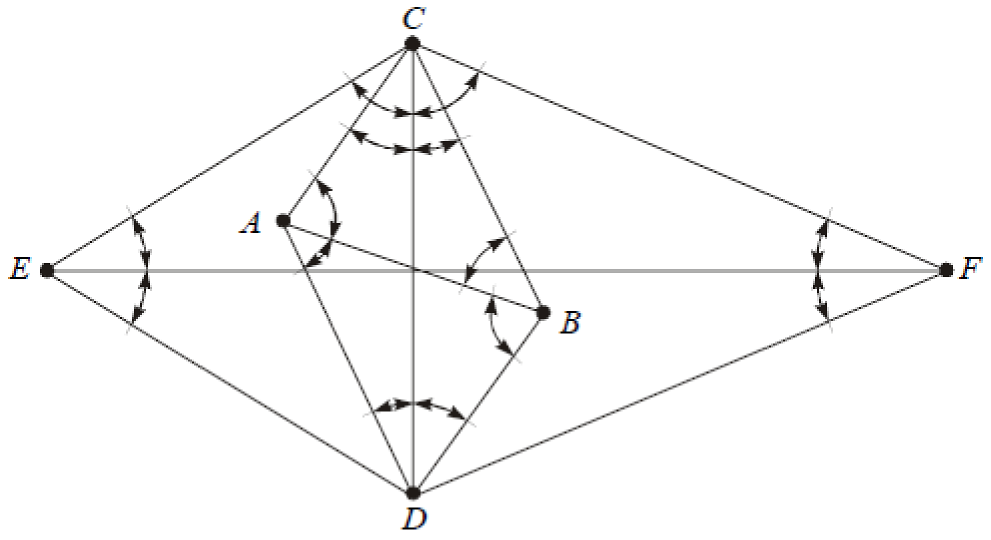

Base extension by double sighting

(*iv*) The new base line *CD* can be further extended to the length *EF* following the same procedure as above. The line *EF* may from a side of the triangulation system.

If the base line *AB* is measured on a good site which is well located for extension and connection to the main triangulation system, the accuracy of the system is not much affected by the extension of the base line. In fact, in some cases, the accuracy may be higher than that of a longer base line measured over a poor terrain

4.11.3 MEASUREMENT OF HORIZONTAL ANGLES

The horizontal angles of a triangulation system can be observed by the following methods:

(*i*) Repetition method
(*ii*) Reiteration method.

The procedure of observation of the horizontal angles by the above methods has been discussed in

(*i*) Repetition method

For measuring an angle to the highest degree of precision, several sets of repetitions are usually taken.

There are following two methods of taking a single set.

(*a*) In the first method, the angle is measured clockwise by 6 repetitions keeping the telescope normal.

The first value of the angle is obtained by dividing the final reading by 6. The telescope is inverted, and the angle is measured again in anticlockwise direction by 6 repetitions. The second value of the angle is obtained by dividing the final reading by 6. The mean of the first and second values of the angle is the average value of the angle by first set.

For first-order work, five or six sets are usually required. The final value of the angle is the mean of the values obtained by different sets.

(*b*) In the second method, the angle is measured clockwise by six repetitions, the first three with telescope normal and the last three with telescope inverted. The first value of the angle is obtained by dividing the final reading by 6. Now without altering the reading obtained in the sixth repetition, the explement angle (i.e., 360°– the angle), is measured clockwise by six repetitions, the first three with telescope inverted and the last three with telescope normal. The final reading should theoretically be zero. If the final reading is not zero, the error is noted, and half of the error is distributed to the first value of the angle. The result is the corrected value of the angle by the first set. As many sets as desired are taken, and the mean of all the value of various sets, is the average value of the angle. For more accurate work and to eliminate the errors due to inaccurate graduations of the horizontal circle, the initial reading at the beginning of each set may not be set to zero but to different values. If n sets are required, the initial setting should be sucessively increased by $180°/n$.

For example, for 6 sets the initial readings would be 0°, 30°, 60°, 90°, 120° and 150°, respectively.

(*ii*) Reiteration method or direction method

In the reiteration method, the triangulation signals are bisected successively, and a value is obtained for each direction in each of several rounds of observations. One of the triangulation stations which is likely to be always clearly visible may be selected as the initial station or reference station. The theodolites used for the measurement of angles for triangulation surveys, have more than one micrometer. One of the micrometer is set to 0° and with telescope normal, the initial station is bisected, and all the micrometers are read. Each of the successive stations are then bisected, and all the micrometers are read. The stations are then again bisected in the reverse direction, and all the micrometers are read after each bisection. Thus, two values are obtained for each angle when the telescope is normal. The telescope is then inverted, and the observations are repeated.

This constitutes one set in which four value of each angle are obtained. The micrometer originally at 0° is now brought to a new reading equal to $360°/mn$ (where m is the number of micrometers and n is the number of sets), and a second set is observed in the same manner. The number of sets depends on the accuracy required. For first-order triangulation, sixteen such sets are required with a 1″ direction theodolite, while for second-order triangulation four, and for third-order triangulation two. With more refined instrument having finer graduations, however, six to eight sets are sufficient for the geodetic work.

4.12 MEASUREMENT OF VERTICAL ANGLES

Measurement of vertical angles is required to compute the elevation of the triangulation stations.

4.12.1 ASTRONOMICAL OBSERVATIONS

To determine the azimuth of the initial side, intermediate sides, and the last side of the triangulation net, astronomical observations are made. For detailed procedure and methods of observation

4.12.2 SOME EXTRA PRECAUTIONS IN TAKING OBSERVATIONS

To satisfy first-second, and third-order specifications, care must be exercised. Observer must ensure the following:

1. The instrument and signals have been centred very carefully.
2. Phase in signals has been eliminated.
3. The instrument is protected form the heating effects of the sun and vibrations caused by wind.
4. The support for the instrument is adequately stable.
5. In case of adverse horizontal refraction, observations should be rescheduled to the time when the horizontal refraction is minimum. Horizontal angles should be measured when the air is the clearest, and the lateral refraction is minimum.

If the observations are planned for day hours, the best time in clear weather is from 6 AM to 9 AM and from 4 PM till sunset. In densely clouded weather satisfactory work can be done all day. The best time for measuring vertical angles is form 10 AM to 2 PM when the vertical refraction is the least variable. First-order work is generally done at night, since observations at night using illuminated signals help in reducing bad atmospheric conditions, and optimum results can be obtained. Also working at night doubles the hours of working available during a day. Night operations are confined to period from sunset to midnight.

4.12.3 SATELLITE STATION AND REDUCTION TO CENTRE

To secure well-conditioned triangles or to have good visibility, objects such as chimneys, church spires, flat poles, towers, lighthouse, etc., are selected as triangulation stations. Such stations can be sighted from other stations but it is not possible to occupy the station directly below suchexcellent targets for making the observations by setting up the instrument over the station point. Also, signals are frequently blown out of position, and angles read on them have to be corrected to the true position of the triangulation station. Thus, there are two types of problems:

1. When the instrument is not set up over the true station, and
2. When the target is out of position.

In below Fig., *A*, *B*, and *C* are the three triangulation stations. It is not possible to place instrument at *C*. To solve this problem another station *S*, in the vicinity of *C,* is selected where the instrument can be set up, and from where all the three stations are visible for making the angle observations. Such station is known as *satellite station.* As the observations from *C* are not possible, the observations form *S* are made on *A*, *B*, and, *C* from *A* and *B* on *C*. From the observations made, the required angle *ACB* is calculated. This is known as *reduction to centre.*

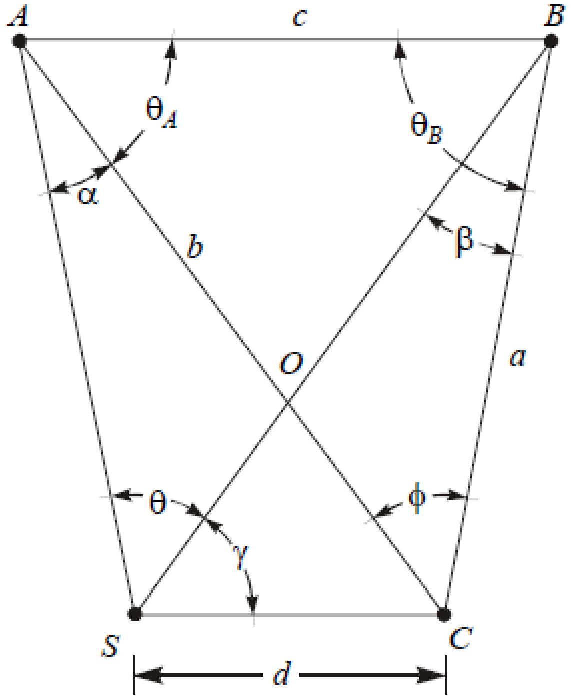

Reduction to centre

In the other case, *S* is treated as the true station point, and the signal is considered to be shifted to the position *C*. This case may also be looked upon as a case of *eccentricity of signal.* Thus, the observations from *S* are made to the triangulation stations *A* and *B*, but from *A* and *B* the observations are made on the signal at the shifted position *C*. This causes errors in the measured values of the angles *BAC* and *ABC*. Both the problems discussed above are solved by reduction to centre. Let the measured

$$\begin{aligned}
\angle BAC &= \theta_A \\
\angle ABC &= \theta_B \\
\angle ASB &= \theta \\
\angle BSC &= \gamma
\end{aligned}$$

Eccentric distance $SC = d$

The distance AB is known by computations form preceding triangle of the triangular net. Further, let

$$\begin{aligned}
\angle SAC &= \alpha \\
\angle SBC &= \beta \\
\angle ACB &= \phi \\
AB &= c \\
AC &= b \\
BC &= a
\end{aligned}$$

As a first approximation in ΔABC the $\angle ACB$ may be taken as

$$= 180^\circ - (\angle BAC + \angle ABC)$$

or

$$\phi = 180^\circ - (\theta_A + \theta_B)$$

In the triangle ABC we have

$$\frac{c}{\sin\phi} = \frac{a}{\sin\theta_A} = \frac{b}{\sin\theta_B}$$

$$a = \frac{c.\sin\theta_A}{\sin\phi}$$

and

$$b = \frac{c.\sin\theta_B}{\sin\phi}$$

Compute the values of a and b by substituting the value of ϕ obtained and respectively.

Now, from $\Delta^s SAC$ and SBC we have

$$\frac{d}{\sin\alpha} = \frac{b}{\sin(\theta+\gamma)}$$

$$\frac{d}{\sin\beta} = \frac{a}{\sin\gamma}$$

$$\sin\alpha = \frac{d\sin(\theta+\gamma)}{b}$$

$$\sin\beta = \frac{d\sin\gamma}{a}$$

As the satellite station *S* is chosen very close to the main station *C*, the angles α and β are extremely small. Therefore, taking $\sin\alpha = \alpha$, and $\sin\beta = \beta$ in radians, we get.

$$\alpha = \frac{d\sin(\theta+\gamma)}{b\sin 1''}$$

or $$= \frac{d\sin(\theta+\gamma)}{b}\times 206265 \text{ seconds}$$

and $$\beta = \frac{d\sin\gamma}{a}\times 206265 \text{ seconds}$$

θ, γ, d, b and a are known quantities, therefore, the values of α and β can be computed. Now a more correct value of the angle $\angle ACB$ can be found.

We have

$$\angle AOB = \theta + \alpha = \phi + \beta$$

or $$\phi = \theta + \alpha - \beta$$

the value of ϕ when the satellite station *S* is to the left of the main station *C*. In the general, the following four cases as shown in Fig. 1.40*a*, can occur depending on the field conditions.

Case I: *S* towards the left of *C* (Fig. 1.39)

$$\phi = \theta + \alpha - \beta$$

Case II: *S* towards the right of *C* (Fig. 1.40*b*), the position S_2.

$$\phi = \theta - \alpha + \beta$$

Case III: *S* inside the triangle *ABC* (Fig. 1.40*c*), the position S_3.

$$\phi = \theta - \alpha - \beta$$

Case IV: *S* outside the triangle *ABC* (Fig. 1.40*d*), the position S_4.

$$\phi = \theta + \alpha + \beta$$

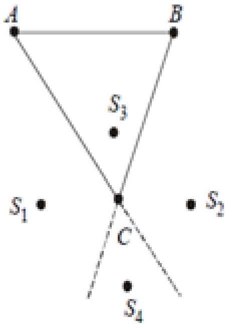

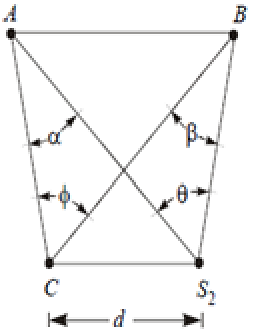

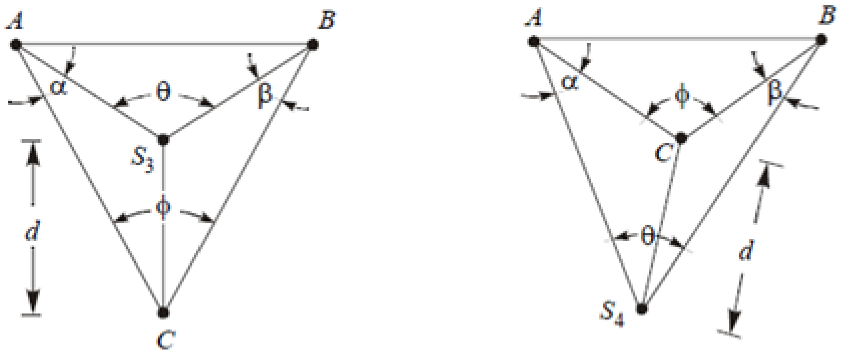

(c) (d)

Locations of satellite station with reference to triangulation stations C

4.13 ECCENTRICITY OF SIGNAL

When the signal is found shifted from its true position, the distance between the shifted signal and the station point *d* is measured. The corrections α and β to the observed angles *BAC* and *ABC*, respectively, are computed and the corrected values of the angles are obtained as under

$$\text{Correct } \angle BAS = \theta_A + \alpha$$
$$\text{Correct } \angle ABS = \theta_B - \beta$$

4.14 Trigonometrical Levelling

Trigonometric Levelling is the branch of Surveying in which we find out the vertical distance between two points by taking the vertical angular observations and the known distances. The known distances are either assumed to be horizontal or the geodetic lengths at the mean sea level(MSL). The distances are measured directly(as in the plane surveying) or they are computed as in the geodetic surveying.

The trigonometric Leveling can be done in two ways:

(1) Observations taken for the height and distances (2) **Geodetic Observations.**

In the first way, we can measure the horizontal distance between the given points if it is accessible.We take the observation of the vertical angles and then compute the distances using them. If the distances are large enough then we have to provide the correction for the curvature and refraction and that we provide to the linearly to the distances that we have computed.

In the second way, i.e geodetic observations, the distances between the two points are geodetic distances and the principles of the plane surveying are not applicable here. The corrections for the curvature and refraction are applied directly to the angles directly.

Now we will discuss the various cases to find out the difference in elevation between the two.

(1) The two points are at known distance: The base of the object is accessible.

When the two points are at a known horizontal distance then we can find out the distance between them by taking the vertical angle observations.
If the vertical angle of elevation from the point to be observed to the instrument axis is known we can calculate the vertical distance by taking the help of the trigonometry.
Horizontal distance*Tan(verticle angle) = Vertical difference between the two.
If the points are at small distance apart then there is no need to apply the correction for the curvature and refraction else you can apply the correction as given below:
C= 0.06728D*D
Where D is the horizontal distance between the given two points in Kilometers.
but the Correction is in meters (m).
(2) The base of the object is not accessible :
(a)(***When the instrument is shifted to the nearby place and the observations are taken from the same level of the line of sight:*** In such case we have to take the two angular observations of the vertical angles. The instrument is shifted to a nearby place of known distance, and then with the known distance between these two and the angular observations from these two stations, we can find the vertical difference in distance between the line of sight of the instrument and the top point of the object.
(b) When the line of sights of the two instrument setting is different : Here again there are two cases: (i) When the line of sights are at a small vertical distance which can be measured through the vertical staff readings. (ii) When the difference is larger than the staff height.
***(i)* In first case,** It is advised to apply the formula for the difference in the height of the top of the object from these two lines of sights. The difference in lines of sights is same as the staff readings difference, when the staff is kept at a little distance from these two points. So we can get the solution for the vertical distance easily.
(ii) In the second case, there is a need to put a vane staff at the first instrument station and the angle of elevation is measured from the second point of observation. This gives us the difference in the line of the sights between the two points of instrument station. Then again we do the same.
(c) When the instrument station and the top of object are not in same vertical plane: In this case there is a need to measure at-least two horizontal angles of the horizontal triangle formed by the two instrument stations and the base of the object. Again we will take the vertical angular observations from the two instrument stations also and then we can apply the sine rule to solve the horizontal distances of the triangle. With the help of these angles and the distances we can get the vertical distance between any two point(Instrument station and the top of object).

For rapid leveling or leveling in rolling ground or for inaccessible points, trigonometric method of leveling is being used. In this method, theodolite (an instrument which can measure angle) is being generally used as an instrument for taking different measurements.

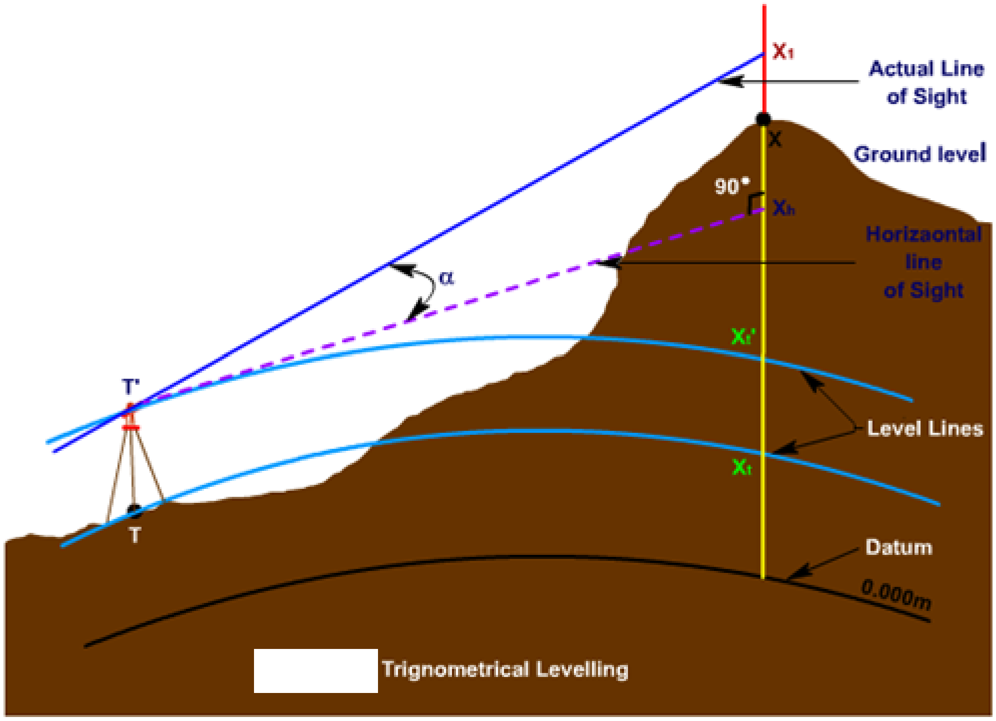

Trignometrical Levelling

Let us consider two stations T an by trigonometric method of leveling. At T, a theodolite instrument is set up. TT ' is the height of the instrument above the point T (to be recorded at the time of observation). A leveling staff is held at X. At the vertical angle of elevation of the actual line of sight a, let x_1 is the observed staff reading. The difference in level between T and X is given by

$$\Delta H_{TX} = TT' + x_{t'} x_h + x_h x_1 - x_1 X$$

where $x_{t'} x_h$ is deviation of the horizontal line of sight due to curvature of the earth and refraction of light (given by 0.0675 $T' x_h^2$). $x_h x_1$ is $T' x_1$ sina or $T' x_h$ tana , $T' x_1$ is the inclined distance from the instrument to the staff and $T' x_h$ is the horizontal distance between the points, $x_1 X$ is the staff reading at X.

Example

In order to eliminate the uncertainty due to refraction, observations for vertical angle are made at both ends of the line as close in point of time as possible. The vertical angle at the lower of the two peaks to the upper peak is +3° 02' 05"?. The reciprocal vertical angle at the upper peak is -3° 12' 55"?. The height of instrument are kept to be same in all observation. The slope distance between two mountain peaks determined by EDM measurement is 21,345m. Compute the difference in elevations between the two peaks.

Solution :

Average vertical angle a = (3° 02' 05" + 3° 12' 55") / 2 = 6° 15' 00 "

Difference in elevation = 21.345 sin 3° 07' 30 " + 0.0675 (21.345 cos 3° 07' 30 ")^2

= (1.163 + 30.662) m

= 31.825 me

Reciprocal leveling

To find accurate relative elevations of two widely separated intervisible points (between which levels cannot be set), reciprocal leveling is being used.

To find the difference in elevation between two points, say X and Y, a level is set up at L near X and readings (X_1 and Y_1) are observed with staff on both X and Y respectively. The level is then set up near Y and staff readings (Y_2 and X_2) are taken respectively to the near and distant points. If the differences in the set of observations are not same, then the observations are fraught with errors. The errors may arise out of the curvature of the earth or intervening atmosphere (associated with variation in temperature and refraction) or instrument (due to error in collimation) or any combination of these.

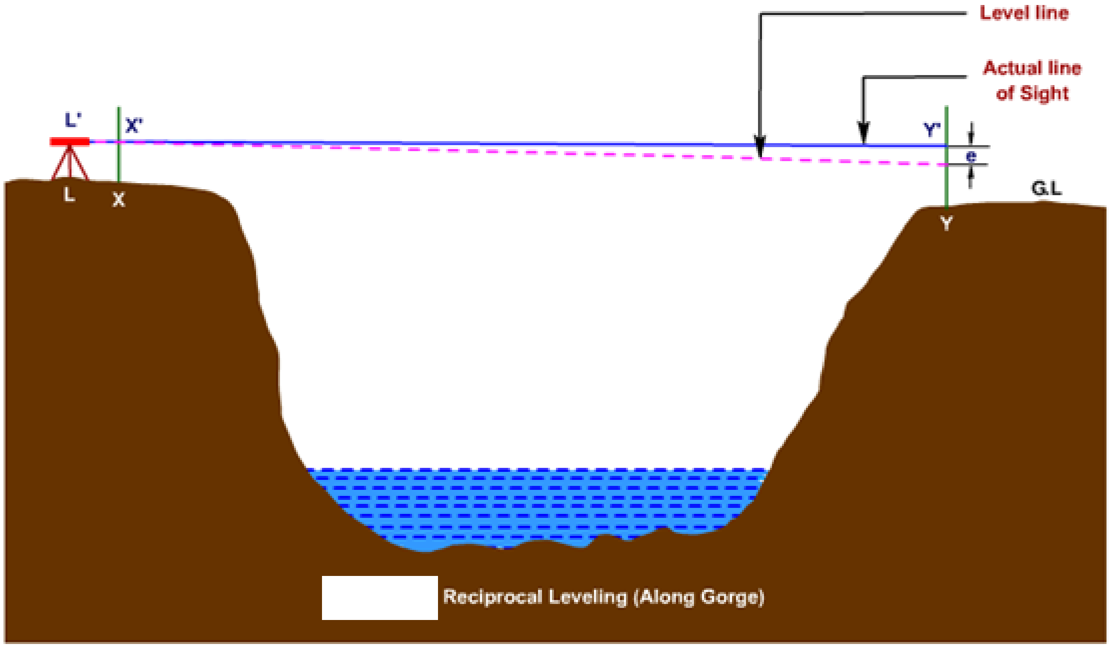

The true difference in elevation and errors associated with observation, if any, can be found as follows:

Let the true difference in elevation between the points be rh and the total error be e. Assuming, no error on observation of staff near the level (as the distance is very small)

Then, rh = $X_1 \sim (Y_1 - e)$ [From first set of observation]

and rh = $(X_2 - e) \sim Y_2$ [From second set of observation]

$$\text{or, } \Delta h = \frac{(X_1 \sim Y_1) + (X_2 \sim Y_2)}{2}$$

$$\text{or, } \Delta h = \frac{\text{First difference in observations} + \text{Second difference in observation}}{2}$$

Thus, the true difference in elevation between any two points can be obtained by taking the mean of the two differences in observation.

$$\text{Total error, } e = \frac{(X_1 \sim Y_1) \sim (X_2 \sim Y_2)}{2}$$

$$\text{or, } e = \frac{\text{First difference in observations} - \text{Second difference in observation}}{2}$$

Thus, total error in observations can be obtained by taking the difference of the two differences in observation. The total error consist of error due to curvature of the earth, atmospheric errors (due to temperature and refraction) and instrumental errors (due to error in collimation) etc.

Example

In order to transfer reduced level across a canyon, a reciprocal leveling campaign was conducted. Simultaneous readings were observed using two levels one at each side of the canyon. Each of the levels are having same magnifying power and sensitiveness of level tube. With instruments interchanged during leveling operation yielded the following average readings:

Instrument station	**Average near readings, meter**	**Average distant, readings, meter**	**R.L of X = 101.345 m** **Distance, XY = 1.025Km** **$e_{curvature}$ = 0.0785 XY^2**
X	1.780	2.345	
Y	2.435	1.870	

Find out the R.L. of unknown point. Comment on the errors associated with observations.

Solution :

The difference in elevation between X and Y is

$$\Delta h = \frac{(2.345 - 1.780) + (2.436 - 1.870)}{2} = 0.565 \text{ m (Y lower than X)}$$

R.L. of Y (unknown Point) = R.L. of X - Dh = 101.345 - 0.565 = 100.780 m

Since two leveling rods are used and the elapsed time between reading in a set observation is little, the error due to change in atmospheric condition can be neglected. Moreover, since readings were taken with instruments interchanged, instrumental errors get cancelled between different set of observation. As the observations are repeated and averages of the readings have been considered for further calculation, it is expected that error associated with observation is minimized thus removed. Only error present in the observation is that associated with the curvature of the earth.

4.15 Trilateration

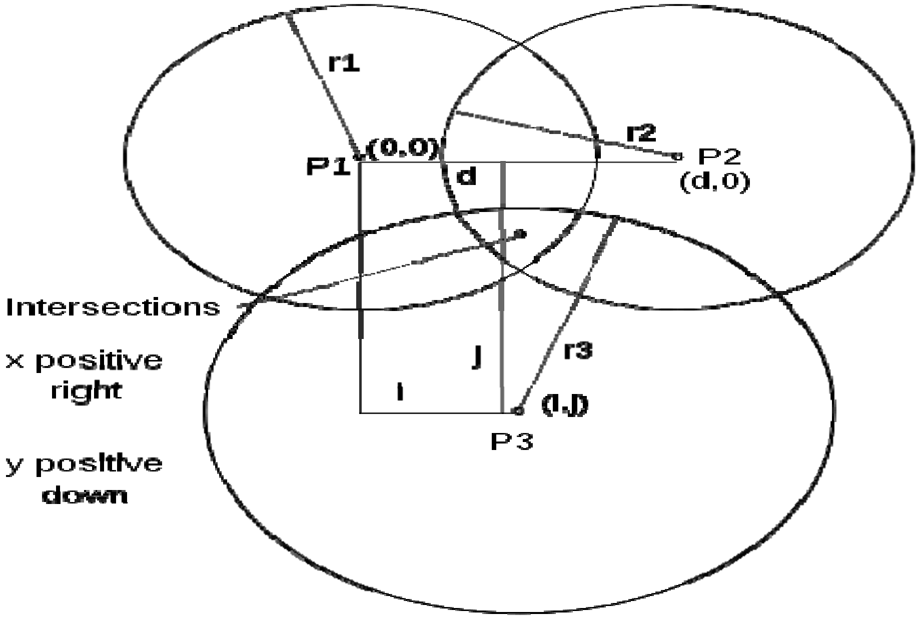

In above Figure . The plane *z* = 0, showing the three sphere centers, *P*1, *P*2, and *P*3; their*x*,*y*-coordinates; and the three sphere radii, *r*1, *r*2, and *r*3. The two intersections of the three sphere surfaces are directly in front and directly behind the point designated intersections in the *z* = 0 plane.

In geometry, **trilateration** is the process of determining absolute or relative locations of points by measurement of distances, using the geometry of circles, spheres or triangles. In addition to its interest as a geometric problem, trilateration does have practical applications in surveying and navigation, including global positioning systems (GPS). In contrast to triangulation it does not involve the measurement of angles.

In two-dimensional geometry, it is known that if a point lies on two circles then the circle centers and the two radii provide sufficient information to narrow the possible locations down to two. Additional information may narrow the possibilities down to one unique location.

In three-dimensional geometry, when it is known that a point lies on the surfaces of three spheres then the centers of the three spheres along with their radii provide sufficient information to narrow the possible locations down to no more than two (unless the centres lie on a straight line).

This article describes a method for determining the intersections of three sphere surfaces given the centers and radii of the three spheres.

Contents

- 1 Derivation
- 2 Preliminary and final computations

Derivation

The intersections of the surfaces of three spheres is found by formulating the equations for the three sphere surfaces and then solving the three equations for the three unknowns, *x*, *y*, and *z*. To simplify the calculations, the equations are formulated so that the centers of the spheres are on the *z* = 0 plane. Also the formulation is such that one center is at the origin, and one other is on the *x*-axis. It is possible to formulate the equations in this manner since any three non-colinear points lie on a unique plane. After finding the solution it can be transformed back to the original three dimensional Cartesian coordinate system.

We start with the equations for the three spheres:

$$r_1^2 = x^2 + y^2 + z^2$$

$$r_2^2 = (x - d)^2 + y^2 + z^2$$

$$r_3^2 = (x - i)^2 + (y - j)^2 + z^2$$

We need to find a point located at (x, y, z) that satisfies all three equations.

First we subtract the second equation from the first and solve for x:

$$x = \frac{r_1^2 - r_2^2 + d^2}{2d}.$$

We assume that the first two spheres intersect in more than one point, that is that

$$d - r_1 < r_2 < d + r_1.$$

In this case substituting the equation for x back into the equation for the first sphere produces the equation for a circle, the solution to the intersection of the first two spheres:

$$y^2 + z^2 = r_1^2 - \frac{(r_1^2 - r_2^2 + d^2)^2}{4d^2}.$$

Substituting $z^2 = r_1^2 - x^2 - y^2$ into the formula for the third sphere and solving for y there results:

$$y = \frac{r_1^2 - r_3^2 - x^2 + (x - i)^2 + j^2}{2j} = \frac{r_1^2 - r_3^2 + i^2 + j^2}{2j} - \frac{i}{j}x.$$

Now that we have the x- and y-coordinates of the solution point, we can simply rearrange the formula for the first sphere to find the z-coordinate:

$$z = \pm\sqrt{r_1^2 - x^2 - y^2}.$$

Now we have the solution to all three points x, y and z. Because z is expressed as the positive or negative square root, it is possible for there to be zero, one or two solutions to the problem.

This last part can be visualized as taking the circle found from intersecting the first and second sphere and intersecting that with the third sphere. If that circle falls entirely outside or inside of the sphere, z is equal to thesquare root of a negative number: no real solution exists. If that circle touches the sphere on exactly one point, z is equal to zero. If that circle touches the surface of the sphere at two points, then z is equal to plus or minus the square root of a positive number.

Preliminary and final computations

The *Derivation* section pointed out that the coordinate system in which the sphere centers are designated must be such that (1) all three centers are in the plane $z = 0$, (2) the sphere center, $P1$, is at the origin, and (3) the sphere center, $P2$, is on the x-axis. In general the problem will not be given in a form such that these requirements are met.

This problem can be overcome as described below where the points, $P1$, $P2$, and $P3$ are treated as vectors from the origin where indicated. $P1$, $P2$, and $P3$ are of course expressed in the original coordinate system.

$\hat{e}_x = \frac{P2 - P1}{\|P2 - P1\|}$ is the unit vector in the direction from $P1$ to $P2$.

$i = \hat{e}_x \cdot (P3 - P1)$ is the signed magnitude of the x component, in the figure 1 coordinate system, of the vector from $P1$ to $P3$.

$\hat{e}_y = \frac{P3 - P1 - i\,\hat{e}_x}{\|P3 - P1 - i\,\hat{e}_x\|}$ is the unit vector in the y direction. Note that the points $P1$, $P2$, and $P3$ are all in the $z = 0$ plane of the figure 1 coordinate system.

The third basis unit vector is $\hat{e}_z = \hat{e}_x \times \hat{e}_y$. Therefore,

$d = \|P2 - P1\|$ the distance between the centers $P1$ and $P2$ and

$j = \hat{e}_y \cdot (P3 - P1)$ is the signed magnitude of the y component, in the figure 1 coordinate system, of the vector from $P1$ to $P3$.

Using i, d and j as computed above, solve for x, y and z as described in the *Derivation* section. Then

$$\vec{p}_{1,2} = P1 + x\,\hat{e}_x + y\,\hat{e}_y \pm z\,\hat{e}_z$$

gives the points in the original coordinate system since $\hat{e}_x$, $\hat{e}_y$ and $\hat{e}_z$, the basis unit vectors, are expressed in the original coordinate system.

Resection method

Resection is a method for determining a position (position finding) using a compass and topographic map (or nautical chart)

Resection versus intersection

Resection and its related method, intersection, are used in surveying as well as in general land navigation (including inshore marine navigation using shore-based landmarks). Both methods involve taking azimuths or bearings to two or more objects, then drawing lines of position along those recorded bearings or azimuths.

When intersecting lines of position are used to fix the position of an unmapped feature or point by fixing its position relative to two (or more) mapped or known points, the method is known asintersection.[3] At each known point (hill, lighthouse, etc.), the navigator measures the bearing to the same unmapped target, drawing a line on the map from each known position to the target. The target is located where the lines intersect on the map. In earlier times, the intersection method was used by forest agencies and others using specialized alidades to plot the (unknown) location of an observed forest fire from two or more mapped (known) locations, such as forest fire observer towers.

The reverse of the intersection technique is appropriately termed resection. Resection simply reverses the intersection process by using crossed back bearings, where the navigator's position is the unknown. Two or more bearings to mapped, known points are taken; their resultant lines of position drawn from those points to where they intersect will reveal the navigator's location.

Fixing a position

When resecting or fixing a position, the geometric strength (angular disparity) of the mapped points affects the precision and accuracy of the outcome. Accuracy increases as the angle between the two position lines approaches 90 degrees. Magnetic bearings are observed on the ground from the point under location to two or more features shown on a map of the area. Lines of reverse bearings, or lines of position, are then drawn on the map from the known features; two and more lines provide the resection point (the navigator's location). When three or more lines of position are utilized, the method is often popularly (though erroneously) referred to as triangulation (in precise terms, using three or more lines of position is still correctly called resection, as angular law of tangents (cot) calculations are not performed). When using a map and compass to perform resection, it is important to allow for the difference between the magnetic bearings observed and grid north (or true north) bearings (magnetic declination) of the map or chart.

Resection continues to be employed in land and inshore navigation today, as it is a simple and quick method requiring only an inexpensive magnetic compass and map/chart.

Resection in surveying

In surveying work, the most common methods of computing the coordinates of a point by resection are Cassini's Method and the Tienstra formula, though the first known solution was given byWillebrord Snellius (see Snellius–Pothenot problem). For the type of precision work involved in surveying, the unmapped point is located by measuring the angles subtended by lines of sight from it to a minimum of three mapped (coordinated) points. In geodetic operations the observations are adjusted for spherical excess and projection variations. Precise angular measurements between lines from the point under location using theodolites provides more accurate results, with trig beacons erected on high points and hills to enable quick and unambiguous sights to known points.

Caution: When planning to perform a resection, the surveyor must first plot the locations of the known points along with the approximate unknown point of observation. If all points, including the unknown point, lie close to a circle that can be placed on all four points, then there is no solution or the high risk of an erroneous solution. This is known as observing on the "danger circle". The poor solution stems from the property of a chord subtending equal angles to any other point on the circle.

4.16 GPS

The **Global Positioning System** (**GPS**) is a space-based satellite navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites. The system provides critical capabilities to military, civil and commercial users around the world. It is maintained by the United States government and is freely accessible to anyone with a GPS receiver.

The GPS project was developed in 1973 to overcome the limitations of previous navigation systems, integrating ideas from several predecessors, including a number of classified engineering design studies from the 1960s. GPS was created and realized by the U.S. Department of Defense (DoD) and was originally run with 24 satellites. It became fully operational in 1994. Bradford Parkinson, Roger L. Easton, and Ivan A. Getting are credited with inventing it.

Advances in technology and new demands on the existing system have now led to efforts to modernize the GPS system and implement the next generation of GPS III satellites and Next Generation Operational Control System (OCX). Announcements from Vice President Al Gore and the White House in 1998 initiated these changes. In 2000, the U.S. Congress authorized the modernization effort, GPS III.

In addition to GPS, other systems are in use or under development. The Russian Global Navigation Satellite System (GLONASS) was developed contemporaneously with GPS, but suffered from incomplete coverage of the globe until the mid-2000s. There are also the planned European UnionGalileo positioning system, Chinese Compass navigation system, and Indian Regional Navigational Satellite System.

History

The design of GPS is based partly on similar ground-based radio-navigation systems, such as LORAN and the Decca Navigator, developed in the early 1940s and used during World War II.

Predecessors

In 1956, the German-American physicist Friedwardt Winterberg proposed a test of general relativity (for time slowing in a strong gravitational field) using accurate atomic clocks placed in orbit inside artificial satellites. Without the use of general relativity to correct for time running more quickly by 38 microseconds per day in orbit, GPS would suffer gross malfunction. Additional inspiration for GPS came when the Soviet Union launched the first man-made satellite, Sputnik, in 1957. Two American physicists, William Guier and George Weiffenbach, at Johns Hopkins's Applied Physics Laboratory (APL), decided to monitor Sputnik's radio transmissions. Within hours they realized that, because of the Doppler effect, they could pinpoint where the satellite was along its orbit. The Director of the APL gave them access to their UNIVAC to do the heavy calculations required. The next spring, Frank McClure, the deputy director of the APL, asked Guier and Weiffenbach to investigate the inverse problem—pinpointing the user's location given that of the satellite. (The Navy was developing the submarine-launched Polaris missile, which required them to know the submarine's location.) This led them and APL to develop the Transit system. In 1959, ARPA (renamed DARPA in 1972) also played a role in Transit. The first satellite navigation system, Transit, used by the United States Navy, was first successfully tested in 1960. It used a constellation of five satellites and could provide a navigational fix approximately once per hour. In 1967, the U.S. Navy developed the Timation satellite that proved the ability to place accurate clocks in space, a technology required by GPS. In the 1970s, the ground-based Omega Navigation System, based on phase comparison of signal transmission from pairs of stations, became the first worldwide radio navigation system. Limitations of these systems drove the need for a more universal navigation solution with greater accuracy.

While there were wide needs for accurate navigation in military and civilian sectors, almost none of those was seen as justification for the billions of dollars it would cost in research, development, deployment, and operation for a constellation of navigation satellites. During the Cold War arms race, the nuclear threat to the existence of the United States was the one need that did justify this cost in the view of the United States Congress. This deterrent effect is why GPS was funded. It is also the reason for the ultra secrecy at that time. The nuclear triad consisted of the United States Navy's submarine-launched ballistic missiles (SLBMs) along with United States Air Force (USAF) strategic bombers and intercontinental ballistic missiles (ICBMs). Considered vital to the nuclear-deterrence posture, accurate determination of the SLBM launch position was a force multiplier.

Precise navigation would enable United States submarines to get an accurate fix of their positions before they launched their SLBMs.[14] The USAF, with two thirds of the nuclear triad, also had requirements for a more accurate and reliable navigation system. The Navy and Air Force were developing their own technologies in parallel to solve what was essentially the same problem. To increase the survivability of ICBMs, there was a proposal to use mobile launch platforms (such as Russian SS-24 and SS-25) and so the need to fix the launch position had similarity to the SLBM situation.

In 1960, the Air Force proposed a radio-navigation system called MOSAIC (MObile System for Accurate ICBM Control) that was essentially a 3-D LORAN. A follow-on study, Project 57, was worked in 1963 and it was "in this study that the GPS concept was born". That same year, the concept was pursued as Project 621B, which had "many of the attributes that you now see in GPS" and promised increased accuracy for Air Force bombers as well as ICBMs. Updates from the Navy Transit system were too slow for the high speeds of Air Force operation. The Naval Research Laboratory continued advancements with their Timation (Time Navigation) satellites, first launched in 1967, and with the third one in 1974 carrying the first atomic clock into orbit.

Another important predecessor to GPS came from a different branch of the United States military. In 1964, the United States Army orbited its first Sequential Collation of Range (SECOR) satellite used for geodetic surveying. The SECOR system included three ground-based transmitters from known locations that would send signals to the satellite transponder in orbit. A fourth ground-based station, at an undetermined position, could then use those signals to fix its location precisely. The last SECOR satellite was launched in 1969. Decades later, during the early years of GPS, civilian surveying became one of the first fields to make use of the new technology, because surveyors could reap benefits of signals from the less-than-complete GPS constellation years before it was declared operational. GPS can be thought of as an evolution of the SECOR system where the ground-based transmitters have been migrated into orbit.

Development

With these parallel developments in the 1960s, it was realized that a superior system could be developed by synthesizing the best technologies from 621B, Transit, Timation, and SECOR in a multi-service program.

During Labor Day weekend in 1973, a meeting of about 12 military officers at the Pentagon discussed the creation of a Defense Navigation Satellite System (DNSS). It was at this meeting that "the real synthesis that became GPS was created." Later that year, the DNSS program was named Navstar, or Navigation System Using Timing and Ranging. With the individual satellites being associated with the name Navstar (as with the predecessors Transit and Timation), a more fully encompassing name was used to identify the constellation of Navstar satellites, Navstar-GPS, which was later shortened simply to GPS. After Korean Air Lines Flight 007, a Boeing 747 carrying 269 people, was shot down in 1983 after straying into the USSR's prohibited airspace, in the vicinity of Sakhalin and Moneron Islands, President Ronald Reagan issued a directive making GPS freely available for civilian use, once it was sufficiently developed, as a common good. The first satellite was launched in 1989, and the 24th satellite was launched in 1994. The GPS program cost at this point, not including the cost of the user equipment, but including the costs of the satellite launches, has been estimated to be about USD$5 billion (then-year dollars). Roger L. Easton is widely credited as the primary inventor of GPS.

Initially, the highest quality signal was reserved for military use, and the signal available for civilian use was intentionally degraded (Selective Availability). This changed with President Bill Clintonordering Selective Availability to be turned off at midnight May 1, 2000, improving the precision of civilian GPS from 100 meters (330 ft) to 20 meters (66 ft). The executive order signed in 1996 to turn off Selective Availability in 2000 was proposed by the U.S. Secretary of Defense, William Perry, because of the widespread growth of differential GPS services to improve civilian accuracy and eliminate the U.S. military advantage. Moreover, the U.S. military was actively developing technologies to deny GPS service to potential adversaries on a regional basis.

Over the last decade, the U.S. has implemented several improvements to the GPS service, including new signals for civil use and increased accuracy and integrity for all users, all while maintaining compatibility with existing GPS equipment.

GPS modernizationhas now become an ongoing initiative to upgrade the Global Positioning System with new capabilities to meet growing military, civil, and commercial needs. The program is being implemented through a series of satellite acquisitions, including GPS Block III and the Next Generation Operational Control System (OCX). The U.S. Government continues to improve the GPS space and ground segments to increase performance and accuracy.

GPS is owned and operated by the United States Government as a national resource. Department of Defense (DoD) is the steward of GPS. Interagency GPS Executive Board (IGEB) oversaw GPS policy matters from 1996 to 2004. After that the National Space-Based Positioning, Navigation and Timing Executive Committee was established by presidential directive in 2004 to advise and coordinate federal departments and agencies on matters concerning the GPS and related systems. The executive committee is chaired jointly by the deputy secretaries of defense and transportation. Its membership includes equivalent-level officials from the departments of state, commerce, and homeland security, the joint chiefs of staff, and NASA. Components of the executive office of the president participate as observers to the executive committee, and the FCC chairman participates as a liaison.

The DoD is required by law to "maintain a Standard Positioning Service (as defined in the federal radio navigation plan and the standard positioning service signal specification) that will be available on a continuous, worldwide basis," and "develop measures to prevent hostile use of GPS and its augmentations without unduly disrupting or degrading civilian uses."

Basic concept of GPS

A GPS receiver calculates its position by precisely timing the signals sent by GPS satellites high above the Earth. Each satellite continually transmits messages that include

- the time the message was transmitted
- satellite position at time of message transmission

The receiver uses the messages it receives to determine the transit time of each message and computes the distance to each satellite using the speed of light. Each of these distances and satellites' locations define a sphere. The receiver is on the surface of each of these spheres when the distances and the satellites' locations are correct. These distances and satellites' locations are used to compute the location of the receiver using the navigation equations. This location is then displayed, perhaps with a moving map display or latitude and longitude; elevation or altitude information may be included, based on height above the geoid (e.g. EGM96). Many GPS units show derived information such as direction and speed, calculated from position changes.

In typical GPS operation, four or more satellites must be visible to obtain an accurate result. Four sphere surfaces typically do not intersect. Because of this, it can be said with confidence that when the navigation equations are solved to find an intersection, this solution gives the position of the receiver along with the difference between the time kept by the receiver's on-board clock and the true time-of-day, thereby eliminating the need for a very large, expensive, and power hungry clock. The very accurately computed time is used only for display or not at all in many GPS applications, which use only the location. A number of applications for GPS do make use of this cheap and highly accurate timing. These include time transfer, traffic signal timing, andsynchronization of cell phone base stations.

Although four satellites are required for normal operation, fewer apply in special cases. If one variable is already known, a receiver can determine its position using only three satellites. For example, a ship or aircraft may have known elevation. Some GPS receivers may use additional clues or assumptions such as reusing the last known altitude, dead reckoning, inertial navigation, or including information from the vehicle computer, to give a (possibly degraded) position when fewer than four satellites are visible.

Structure

The current GPS consists of three major segments. These are the space segment (SS), a control segment (CS), and a user segment (US). The U.S. Air Force develops, maintains, and operates the space and control segments. GPS satellites broadcast signals from space, and each GPS receiver uses these signals to calculate its three-dimensional location (latitude, longitude, and altitude) and the current time. The space segment is composed of 24 to 32 satellites in medium Earth orbit and also includes the payload adapters to the boosters required to launch them into orbit. The control segment is composed of a master control station, an alternate master control station, and a host of dedicated and shared ground antennas and monitor stations. The user segment is composed of hundreds of thousands of U.S. and allied military users of the secure GPS Precise Positioning Service, and tens of millions of civil, commercial, and scientific users of the Standard Positioning Service (seeGPS navigation devices).

Space segment

The space segment (SS) is composed of the orbiting GPS satellites, or Space Vehicles (SV) in GPS parlance. The GPS design originally called for 24 SVs, eight each in three approximately circular orbits, but this was modified to six orbital planes with four satellites each. The six orbit planes have approximately 55° inclination (tilt relative to Earth's equator) and are separated by 60° right ascension of the ascending node (angle along the equator from a reference point to the orbit's intersection). The orbital period is one-half a sidereal day, i.e., 11 hours and 58 minutes so that the satellites pass over the same locations or almost the same locations every day. The orbits are arranged so that at least six satellites are always within line of sight from almost everywhere on Earth's surface. The result of this objective is that the four satellites are not evenly spaced (90 degrees) apart within each orbit. In general terms, the angular difference between satellites in each orbit is 30, 105, 120, and 105 degrees apart which sum to 360 degrees.

Orbiting at an altitude of approximately 20,200 km (12,600 mi); orbital radius of approximately 26,600 km (16,500 mi), each SV makes two complete orbits each sidereal day, repeating the same ground track each day. This was very helpful during development because even with only four satellites, correct alignment means all four are visible from one spot for a few hours each day. For military operations, the ground track repeat can be used to ensure good coverage in combat zones.

As of December 2012, there are 32 satellites in the GPS constellation. The additional satellites improve the precision of GPS receiver calculations by providing redundant measurements. With the increased number of satellites, the constellation was changed to a nonuniform arrangement. Such an arrangement was shown to improve reliability and availability of the system, relative to a uniform system, when multiple satellites fail. About nine satellites are visible from any point on the ground at any one time (see animation at right), ensuring considerable redundancy over the minimum four satellites needed for a position.

Control segment

The control segment is composed of

1. a master control station (MCS),
2. an alternate master control station,
3. four dedicated ground antennas and
4. six dedicated monitor stations

The MCS can also access U.S. Air Force Satellite Control Network (AFSCN) ground antennas (for additional command and control capability) and NGA (National Geospatial-Intelligence Agency) monitor stations. The flight paths of the satellites are tracked by dedicated U.S. Air Force monitoring stations in Hawaii, Kwajalein Atoll, Ascension Island, Diego Garcia, Colorado Springs, Colorado and Cape Canaveral, along with shared NGA monitor stations operated in England, Argentina, Ecuador, Bahrain, Australia and Washington DC.[65] The tracking information is sent to the Air Force Space Command MCS at Schriever Air Force Base 25 km (16 mi) ESE of Colorado Springs, which is operated by the 2nd Space Operations Squadron(2 SOPS) of the U.S. Air Force. Then 2 SOPS contacts each GPS satellite regularly with a navigational update using dedicated or shared (AFSCN) ground antennas (GPS dedicated ground antennas are located at Kwajalein, Ascension Island, Diego Garcia, and Cape Canaveral). These updates synchronize the atomic clocks on board the satellites to within a few nanoseconds of each other, and adjust the ephemeris of each satellite's internal orbital model. The updates are created by a Kalman filter that uses inputs from the ground monitoring stations, space weather information, and various other inputs.

Satellite maneuvers are not precise by GPS standards. So to change the orbit of a satellite, the satellite must be marked unhealthy, so receivers will not use it in their calculation. Then the maneuver can be carried out, and the resulting orbit tracked from the ground. Then the new ephemeris is uploaded and the satellite marked healthy again.

The Operation Control Segment (OCS) currently serves as the control segment of record. It provides the operational capability that supports global GPS users and keeps the GPS system operational and performing within specification.

OCS successfully replaced the legacy 1970s-era mainframe computer at Schriever Air Force Base in September 2007. After installation, the system helped enable upgrades and provide a foundation for a new security architecture that supported the U.S. armed forces. OCS will continue to be the ground control system of record until the new segment, Next Generation GPS Operation Control System (OCX), is fully developed and functional.

The new capabilities provided by OCX will be the cornerstone for revolutionizing GPS's mission capabilities, and enabling[67] Air Force Space Command to greatly enhance GPS operational services to U.S. combat forces, civil partners and myriad domestic and international users.

The GPS OCX program also will reduce cost, schedule and technical risk. It is designed to provide 50%[68] sustainment cost savings through efficient software architecture and Performance-Based Logistics. In addition, GPS OCX expected to cost millions less than the cost to upgrade OCS while providing four times the capability.

The GPS OCX program represents a critical part of GPS modernization and provides significant information assurance improvements over the current GPS OCS program.

- OCX will have the ability to control and manage GPS legacy satellites as well as the next generation of GPS III satellites, while enabling the full array of military signals.
- Built on a flexible architecture that can rapidly adapt to the changing needs of today's and future GPS users allowing immediate access to GPS data and constellations status through secure, accurate and reliable information.
- Empowers the warfighter with more secure, actionable and predictive information to enhance situational awareness.
- Enables new modernized signals (L1C, L2C, and L5) and has M-code capability, which the legacy system is unable to do.
- Provides significant information assurance improvements over the current program including detecting and preventing cyber attacks, while isolating, containing and operating during such attacks.
- Supports higher volume near real-time command and control capabilities and abilities.

On September 14, 2011, the U.S. Air Force announced the completion of GPS OCX Preliminary Design Review and confirmed that the OCX program is ready for the next phase of development.

The GPS OCX program has achieved major milestones and is on track to support the GPS IIIA launch in May 2014.

User segment

The user segment is composed of hundreds of thousands of U.S. and allied military users of the secure GPS Precise Positioning Service, and tens of millions of civil, commercial and scientific users of the Standard Positioning Service. In general, GPS receivers are composed of an antenna, tuned to the frequencies transmitted by the satellites, receiver-processors, and a highly stable clock (often a crystal oscillator). They may also include a display for providing location and speed information to the user. A receiver is often described by its number of channels: this signifies how many satellites it can monitor simultaneously. Originally limited to four or five, this has progressively increased over the years so that, as of 2007, receivers typically have between 12 and 20 channels.

GPS receivers may include an input for differential corrections, using the RTCM SC-104 format. This is typically in the form of an RS-232 port at 4,800 bit/s speed. Data is actually sent at a much lower rate, which limits the accuracy of the signal sent using RTCMReceivers with internal DGPS receivers can outperform those using external RTCM data] As of 2006, even low-cost units commonly include Wide Area Augmentation System (WAAS) receivers.

Many GPS receivers can relay position data to a PC or other device using the NMEA 0183 protocol. Although this protocol is officially defined by the National Marine Electronics

Association (NMEA), references to this protocol have been compiled from public records, allowing open source tools likegpsd to read the protocol without violating intellectual property laws. Other proprietary protocols exist as well, such as the SiRF andMTK protocols. Receivers can interface with other devices using methods including a serial connection, USB, or Bluetooth.

Applications

While originally a military project, GPS is considered a dual-use technology, meaning it has significant military and civilian applications.

GPS has become a widely deployed and useful tool for commerce, scientific uses, tracking, and surveillance. GPS's accurate time facilitates everyday activities such as banking, mobile phone operations, and even the control of power grids by allowing well synchronized hand-off switching.

Civilian

Astronomy: Both positional and clock synchronization data is used in Astrometry and Celestial mechanics calculations. It is also used in amateur astronomy using small telescopes to professionals observatories, for example, while finding extrasolar planets.

- Automated vehicle: Applying location and routes for cars and trucks to function without a human driver.
- Cartography: Both civilian and military cartographers use GPS extensively.
- Cellular telephony: Clock synchronization enables time transfer, which is critical for synchronizing its spreading codes with other base stations to facilitate inter-cell handoff and support hybrid GPS/cellular position detection for mobile emergency calls and other applications. The first handsets with integrated GPS launched in the late 1990s. The U.S. Federal Communications Commission (FCC) mandated the feature in either the handset or in the towers (for use in triangulation) in 2002 so emergency services could locate 911 callers. Third-party software developers later gained access to GPS APIs from Nextel upon launch, followed by Sprint in 2006, and Verizon soon thereafter.
- Clock synchronization: The accuracy of GPS time signals (±10 ns)[71] is second only to the atomic clocks upon which they are based.
- Disaster relief/emergency services: Depend upon GPS for location and timing capabilities.
- Fleet Tracking: The use of GPS technology to identify, locate and maintain contact reports with one or more fleet vehicles in real-time.
- Geofencing: Vehicle tracking systems, person tracking systems, and pet tracking systems use GPS to locate a vehicle, person, or pet. These devices are attached to the vehicle, person, or the pet collar. The application provides continuous tracking and mobile or Internet updates should the target leave a designated area.[72]
- Geotagging: Applying location coordinates to digital objects such as photographs (in exif data) and other documents for purposes such as creating map overlays with devices like Nikon GP-1
- GPS Aircraft Tracking

- GPS for Mining: The use of RTK GPS has significantly improved several mining operations such as drilling, shoveling, vehicle tracking, and surveying. RTK GPS provides centimeter-level positioning accuracy.
- GPS tours: Location determines what content to display; for instance, information about an approaching point of interest.
- Navigation: Navigators value digitally precise velocity and orientation measurements.
- Phasor measurements: GPS enables highly accurate timestamping of power system measurements, making it possible to compute phasors.
- Recreation: For example, geocaching, geodashing, GPS drawing and waymarking.
- Robotics: Self-navigating, autonomous robots using a GPS sensors, which calculate latitude, longitude, time, speed, and heading.
- Surveying: Surveyors use absolute locations to make maps and determine property boundaries.
- Tectonics: GPS enables direct fault motion measurement in earthquakes.
- Telematics: GPS technology integrated with computers and mobile communications technology in automotive navigation systems

Restrictions on civilian use

The U.S. Government controls the export of some civilian receivers. All GPS receivers capable of functioning above 18 kilometres (11 mi) altitude and 515 metres per second (1,001 kn) or designed, modified for use with unmanned air vehicles like e.g. ballistic or cruise missile systems are classified as munitions (weapons) for which State Department export licenses are required.

This rule applies even to otherwise purely civilian units that only receive the L1 frequency and the C/A (Coarse/Acquisition) code and cannot correct for Selective Availability (U.S. government discontinued SA on May 1, 2000, resulting in a much- improved autonomous GPS accuracy), etc.

Disabling operation above these limits exempts the receiver from classification as a munition. Vendor interpretations differ. The rule refers to operation at both the target altitude and speed, but some receivers stop operating even when stationary. This has caused problems with some amateur radio balloon launches that regularly reach 30 kilometres (19 mi).

These limits only apply to units exported from (or which have components exported from) the USA – there is a growing trade in various components, including GPS units, supplied by other countries, which are expressly sold as ITAR-free.

Military

As of 2009, military applications of GPS include:

- Navigation: GPS allows soldiers to find objectives, even in the dark or in unfamiliar territory, and to coordinate troop and supply movement. In the United States armed forces, commanders use the Commanders Digital Assistant and lower ranks use the Soldier Digital Assistant.

- Target tracking: Various military weapons systems use GPS to track potential ground and air targets before flagging them as hostile.]These weapon systems pass target coordinates to precision-guided munitions to allow them to engage targets accurately. Military aircraft, particularly in air-to-ground roles, use GPS to find targets (for example, gun camera video from AH-1 Cobras in Iraq show GPS co-ordinates that can be viewed with specialized software).
- Missile and projectile guidance: GPS allows accurate targeting of various military weapons including ICBMs, cruise missiles, precision-guided munitions and Artillery projectiles. Embedded GPS receivers able to withstand accelerations of 12,000 g or about 118 km/s^2 have been developed for use in 155 millimetres (6.1 in) howitzers. Search and Rescue: Downed pilots can be located faster if their position is known.
- Reconnaissance: Patrol movement can be managed more closely.
- GPS satellites carry a set of nuclear detonation detectors consisting of an optical sensor (Y-sensor), an X-ray sensor, a dosimeter, and an electromagnetic pulse (EMP) sensor (W-sensor), that form a major portion of the United States Nuclear Detonation Detection System. General William Shelton has stated that this feature may be dropped from future satellites in order to save money. Communicati**o**

Demodulation and decoding

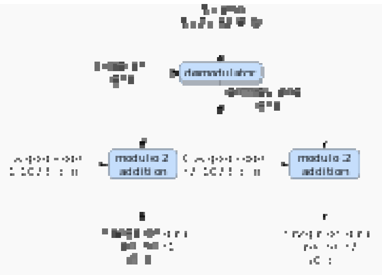

Demodulating and Decoding GPS Satellite Signals using the Coarse/Acquisition Gold code.

Because all of the satellite signals are modulated onto the same L1 carrier frequency, the signals must be separated after demodulation. This is done by assigning each satellite a unique binary sequence known as a Gold code. The signals are decoded after demodulation using addition of the Gold codes corresponding to the satellites monitored by the receiver.

If the almanac information has previously been acquired, the receiver picks the satellites to listen for by their PRNs, unique numbers in the range 1 through 32. If the almanac information is not in memory, the receiver enters a search mode until a lock is obtained on one of the satellites. To obtain a lock, it is necessary that there be an unobstructed line of sight from the receiver to the satellite. The receiver can then acquire the almanac and determine the satellites it should listen for. As it detects each satellite's signal, it identifies it by its distinct C/A code pattern. There can be a delay of up to 30 seconds before the first estimate of position because of the need to read the ephemeris data.

Processing of the navigation message enables the determination of the time of transmission and the satellite position at this time. For more information see Demodulation and Decoding, Advanced.

Navigation equations

The receiver uses messages received from satellites to determine the satellite positions and time sent. The x, y, and z components of satellite position and the time sent are designated as [x_i, y_i, z_i, t_i] where the subscript i denotes the satellite and has the value 1, 2, ..., n, where $n \geq 4$. When the time of message reception indicated by the on-board clock is $\tilde{t}_\text{r}$, the true reception time is $t_\text{r} = \tilde{t}_\text{r} + b$ where b is receiver's clock bias (i.e., clock delay). The message's transit time is $\tilde{t}_\text{r} + b - t_i$. Assuming the message traveled at the speed of light, c, the distance traveled is $\left(\tilde{t}_\text{r} + b - t_i\right) c$. Knowing the distance from receiver to satellite and the satellite's position implies that the receiver is on the surface of a sphere centered at the satellite's position with radius equal to this distance. Thus the receiver is at or near the intersection of the surfaces of the four or more spheres. In the ideal case of no errors, the receiver is at the intersection of the surfaces of the spheres.

The clock error or bias, b, is the amount that the receiver's clock is off. The receiver has four unknowns, the three components of GPS receiver position and the clock bias [x, y, z, b]. The equations of the sphere surfaces are given by:

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = \left([\tilde{t}_\text{r} + b - t_i]c\right)^2,\ i = 1,2,\dots,n$$

or in terms of pseudoranges, $p_i = \left(\tilde{t}_\text{r} - t_i\right) c$, as

$$p_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - bc,\ i = 1, 2, ..., n$$.

These equations can be solved by algebraic or numerical methods.

Least squares method

When more than four satellites are available, the calculation can use the four best or more than four, considering number of channels, processing capability, and geometric dilution of precision(GDOP). Using more than four is an over-determined system of equations with no unique solution, which must be solved by a least-squares method. Errors can be estimated through the residuals. With each combination of four or more satellites, a GDOP factor can be calculated, based on the relative sky directions of the satellites used. The location is expressed in a specific coordinate system or as latitude and longitude, using the WGS 84 geodetic datum or a country-specific system.

$$\left(\hat{x}, \hat{y}, \hat{z}, \hat{b}\right) = \underset{(x,y,z,b)}{\arg\min} \sum_i \left(\sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - bc - p_i\right)^2$$

Bancroft's method

Bancroft's method involves an algebraic as opposed to numerical method and can be used for the case of four or more satellites. Bancroft's method provides one or two solutions for the four unknowns. However when there are two solutions, only one of these two solutions will be a near earth sensible solution. When there are four satellites, we use the inverse of the B matrix in section 2 of. If there are more than four satellites then we use the Generalized inverse (i.e. the pseudoinvers) of the B matrix since in this case the B matrix is no longer square.

Error sources and analysis

GPS error analysis examines the sources of errors in GPS results and the expected size of those errors. GPS makes corrections for receiver clock errors and other effects but there are still residual errors which are not corrected. Sources of error include signal arrival time measurements, numerical calculations, atmospheric effects, ephemeris and clock data, multipath signals, and natural and artificial interference. The magnitude of the residual errors resulting from these sources is dependent on geometric dilution of precision.

Artificial errors may result from jamming devices and threaten ships and aircraft.

Review Questions

1. What is Triangulation Surveying In **detail?**
2. **What is horizontal and vertical control method explain?**
3. **What is signals explain in detail?**
4. **Explain base line?**
5. **What are the** instruments and accessories extension of base lines?
6. **What is** Satellite station explain with digram?
7. **What is** Intervisibility of height and distances in satellite explain?
8. **What is** Trigonometric leveling ?
9. **What is** Axis single corrections explain?
10. **What are the application of satellite?**

Chapter 5

ENGINEERING SURVEYS

Structure of this unit

Reconnaissance, CURVES, CONTOUR

Learning Objectives

1. Reconnaissance
2. Preliminary and location surveys for engineering projects – layout – setting out works
3. CURVES : Curve ranging – Horizontal and Vertical curves – Simple curves –setting with chain and tapes,
4. Tangential angles by theodolite.
5. CONTOUR : Contouring – Methods – Characteristics and uses of contours – Plotting – Calculation of areas and volumes.

.

5.1 ENGINEERING SURVEYS

Engineering Surveys are a single source provider of survey services to the construction and Building industries. Operating from our base in Hong Kong we have a peak workforce of over 120 staff and have been involved in some of the largest infrastructure and building projects undertaken in the region over the last 10 years.

In addition to providing site surveying staff and equipment our vision is one of converging technologies where surveyors provide an increasingly diverse range of products and deliverables driven by the huge increase in quality and quantity of spatial data being collected.

From Topographic and Building Surveys to High Definition Laser Scanning and from CAD Drafting to 3D BIM Models the market for collecting and presenting spatial data has never been larger.

Engineering Surveys, as a multi-disciplined, single source provider of survey services are well placed to serve this market.

5.2 Reconnaissance

Reconnaissance is the military term for exploring beyond the area occupied by friendly forces to gain vital information about enemy forces or features of the environment for later analysis and/or dissemination.

Often referred to as recce (British and Canadian English) or recon (American English, Australian English), the associated verb isreconnoitre in British and Canadian English or reconnoiter in American English.

Examples of reconnaissance include patrolling by troops (LRRPs, Rangers, scouts, or military intelligence specialists), ships orsubmarines, manned/unmanned aircraft, satellites, or by setting up covert observation posts. Espionage normally is not reconnaissance, because reconnaissance is a military force's operating ahead of its main forces; spies are non-combatants operating behind enemy lines.

History

Traditionally, reconnaissance was a role that was adopted by the cavalry. Speed was key in these maneuvers, thus infantry was ill suited to the task. From horses to vehicles, for warriors throughout history, commanders procured their ability to have speed and mobility, to mount and dismount, during maneuver warfare. Military commanders favored specialized small units for speed and mobility, to gain valuable information about the terrain and enemy before sending the main (or majority) troops into the area, screening, covering force, pursuit and exploitation roles. Skirmishing is a traditional skill of reconnaissance, as well as harassment of the enemy.

Overview

Reconnaissance conducted by ground forces includes special reconnaissance, armored reconnaissance, amphibious reconnaissance and civil reconnaissance.

Aerial reconnaissance is reconnaissance carried out by aircraft (of all types including balloons and unmanned aircraft). The purpose is to survey weather conditions, map terrain, and may include military purposes such as observing tangible structures, particular areas, and movement of enemy forces.

Naval forces use aerial and satellite reconnaissance to observe enemy forces. Navies also undertake hydrographic surveys and intelligence gathering.

Reconnaissance satellites provide military commanders with photographs of enemy forces and other intelligence. Military forces also use geographical and meteorological information from Earth observation satellites.

Discipline

5.2.1 Types of reconnaissance:

1. **Terrain-oriented reconnaissance** is a survey of the terrain (its features, weather, and other natural observations).
2. **Force-oriented reconnaissance** focuses on the enemy forces (number, equipment, activities, disposition etc.) and may include target acquisition.
3. **Civil-oriented reconnaissance** focuses on the civil dimension of the battlespace (areas, structures, capabilities, organizations, people and events or ASCOPE).

The techniques and objectives are not mutually exclusive; it is up to the commander whether they are carried out separately or by the same unit.

Reconnaissance-in-force

Some military elements tasked with reconnaissance are armed only for self-defense, and rely on stealth to gather information. Others are well-enough armed to also deny information to the enemy by destroying their reconnaissance elements.

Reconnaissance-in-force (RIF) is a type of military operation or military tactics used specifically to probe an enemy's disposition. By mounting an offensive with considerable (but not decisive) force, the commander hopes to elicit a strong reaction by the enemy that reveals its own strength, deployment, and other tactical data. The RIF commander retains the option to fall back with the data or expand the conflict into a full engagement.

Other methods consist of hit-and-run tactics using rapid mobility, and in some cases light-armored vehicles for added fire superiority, as the need arises.

Reconnaissance-by-fire

Reconnaissance by fire (or speculative fire) is the act of firing at likely enemy positions, in order to cause the enemy force to reveal their location by moving or by returning fire.

Reconnaissance-pull

Reconnaissance-pull is a tactic that is applied at the regiment to division level and defined as locating and rapidly exploiting enemy weaknesses. It is the ability to determine enemy positions and create exploitable gaps through which friendly forces can pass while avoiding obstacles and strong points.

A textbook example of reconnaissance-pull was documented during the Tinian landings during World War II, utilized by the United States Marine Corps's Amphibious Reconnaissance Battalion, from V Amphibious Corps. Aerial photography, and the confirmation by the amphibious reconnaissance platoons determined that the Japanese defenders had largely ignored the northern beaches of the island focusing most of their defensive effort on beaches in the south-west which were more favorable for an amphibious landing. American forces quickly changed their landing location to the northern beaches and planned a small and hasty "deception" operation off the southern beach, which resulted in a complete surprise for the Japanese forces. As a result, American forces were able to fight the Japanese force on land where they had the advantage leading to light losses and a relatively short battle that lasted only 9 days.

Types

When referring to reconnaissance, a commander's full intention is to have a vivid picture of his battlespace. The commander organizes the reconnaissance platoon based on: 1) mission, 2) enemy, 3) terrain, 4) troops and support available, (5) time available, (6) and civil considerations. This analysis determines whether the platoon uses single or multiple elements to conduct the reconnaissance, whether it pertains to **area**, **zone**, or **route reconnaissance**, the following techniques may be used as long as the fundamentals of reconnaissance are applied.

Scouts may also have different tasks to perform for their commanders of higher echelons, for example: the engineer reconnaissance detachments will try to identify difficult terrain in the path of their formation, and attempt to reduce the time it takes to transit the terrain using specialist engineering equipment such as a pontoon bridge for crossing water obstacles. Sanitary epidemiological reconnaissance implies collection and transfer of all data available on sanitary and epidemiological situation of the area of possible deployment and action of armed forces, the same data for the neighboring and enemy armed forces. The aim for the reconnaissance is to clear up the reasons of the specific disease origin- sources of the infection in various extreme situations, including local wars and armed conflicts, the ways of the infection transfer and all factors promoting to the infestation.

After the armed forces have become stationary during wartime and emergency of peace time the sanitary epidemiological reconnaissance turns into sanitary and epidemiological surveillance and medical control of vital and communal activity of the armed forces .

Area

Area reconnaissance refers to the observation, and information obtained, about a specified location and the area around it; it may be terrain-oriented and/or force-oriented. Ideally, a reconnaissance platoon, or team, would use surveillance or vantage (static) points around the objective to observe, and the surrounding area. This methodology focuses mainly prior to moving forces into or near a specified area; the military commander may utilize his reconnaissance assets to conduct an area reconnaissance to avoid being surprised by unsuitable terrain conditions, or most importantly, unexpected enemy forces. The area could be a town, ridge-line, woods, or another feature that friendly forces intend to occupy, pass through, or avoid.

Within an Area of operation (AO), area reconnaissance can focus the reconnaissance on the specific area that is critical to the commander. This technique of focusing the reconnaissance also permits the mission to be accomplished more quickly. Area reconnaissance can thus be a stand-alone mission or a task to a section or the platoon. The commander analyzes the mission to determine whether the platoon will conduct these types of reconnaissance separately or in conjunction with each other.

Civil

Civil Reconnaissance is the process of gathering a broad spectrum of civil information about a specific population in support of military operations. It is related to and often performed in conjunction with infrastructure reconnaissance (assessment and survey). Normally the focus of collection in the operational area for civil reconnaissance is collecting civil information relating to the daily interaction between civilians and military forces. Civil information encompasses relational, temporal, geospatial and behavioral information captured in a socio-cultural backdrop. It is information developed from data related to civil areas, structures, capabilities, organizations, people, and events, within the civil component of the commander's operational environment that can be processed to increase situational awareness and understanding. The type of civil information that is needed in order to support military operations varies based on the environment and situation.

Route

Route reconnaissance is oriented on a given route: e.g. a road, a railway, a waterway; a narrow axis or a general direction of attack, to provide information on route conditions or activities along the route. A military commander relies on information about locations along his determined route: which those that would provide best cover and concealment; bridge by construction type, dimensions, and classification; or for landing zones or pickup zones, if the need arises.

In many cases, the commander may act upon a force-oriented route reconnaissance by which the enemy could influence movement along that route. For the reconnaissance platoons, or squads, stealth and speed —in conjunction with detailed intelligence-reporting—are most important and crucial. The reconnaissance platoon must remain far enough ahead of the maneuver force to assist in early warning and to prevent the force from becoming surprised.

Even it is paramount to obtain information about the available space in which a force can maneuver without being forced to bunch up due to obstacles.Terrain-oriented route reconnaissance allows the commander to obtain information and capabilities about the adjacent terrain for maneuvering his forces, to include, any obstacles (minefields, barriers, steep ravines, marshy areas, or chemical, biological, radiological, and nuclear contamination) that may obstruct vehicle movement—on routes to, and in, his assigned area of operations. This requirement includes the size of trees and the density of forests due to their effects on vehicle movement. Route reconnaissance also allows the observation for fields of fire along the route and adjacent terrain. This information assists planners as a supplement to map information.

Zone

Zone reconnaissance focuses on obtaining detailed information before maneuvering their forces through particular, designated locations. It can be terrain-oriented, force-oriented, or both, as it acquire this information by reconnoitering within—and by maintaining surveillance over—routes, obstacles (to include nuclear-radiological, biological, and chemical contamination), and resources within an assigned location.

Also, force-oriented zone reconnaissance is assigned to gain detailed information about enemy forces within the zone, or when the enemy situation is vague by which the information concerning cross-country traffic-ability is desired. The reconnaissance provides the commander with a detailed picture of how the enemy has occupied the zone, enabling him to choose the appropriate course-of-action.

As the platoon conducts this type of zone reconnaissance, its emphasis is on determining the enemy's locations, strengths, and weaknesses. This is the most thorough and complete reconnaissance mission and therefore is very time-intensive

The survey which is conducted for determining quantities and for collecting data for the designing of engineering works such as roads, railways, etc., is known as Engineering Survey.
Engineering survey have following types:

Reconnaissance* Survey
The Survey which is done for the feasibility* and rough cost of the project is known as Reconnaissance Survey.
Preliminary Survey
The survey in which more precise information is required for the choice of best location for the project and to estimate the exact quantities and costs of project is known as Preliminary Survey.
Location Survey
The survey for setting out the work on the ground is known as location survey.

Reconnaissance means military observation of an area to gain information.
Feasibility means either the project will complete or not.

5.3 Horizontal and Vertical curves

The center line of a road consists of series of straight lines interconnected by curves that are used to change the alignment, direction, or slope of the road. Those curves that change the alignment or direction are known as **horizontal curves,** and those that change the slope are **vertical curves.**

5.3.1 Horizontal Curve

When a highway changes horizontal direction, making the point where it changes direction a point of intersection between two straight lines is not feasible. The change in direction would be too abrupt for the safety of modem, high-speed vehicles. It is therefore necessary to interpose a curve between the straight lines. The straight lines of a road are called **tangents** because the lines are tangent to the curves used to change direction.

In practically all modem highways, the curves are **circular** curves; that is, curves that form circular arcs. The smaller the radius of a circular curve, the sharper the curve. For modern, high-speed highways, the curves must be flat, rather than sharp. That means they must be large-radius curves. In highway work, the curves needed for the location or improvement of small secondary roads may be worked out in the field. Usually, however, the

horizontal curves are computed after the route has been selected, the field surveys have been done, and the survey base line and necessary topographic features have been plotted. In urban work, the curves of streets are designed as an integral part of the preliminary and final layouts, which are usually done on a topographic map. In highway work, the road itself is the end result and the purpose of the design. But in urban work, the streets and their curves are of secondary importance; the best use of the building sites is of primary importance.

The principal consideration in the design of a curve is the selection of the length of the radius or the degree of curvature (explained later). This selection is based on such considerations as the design speed of the highway and the sight distance as limited by headlights or obstructions. Some typical radii you may encounter are 12,000 feet or longer on an interstate highway, 1,000 feet on a major thoroughfare in a city, 500 feet on an industrial access road, and 150 feet on a minor residential street.

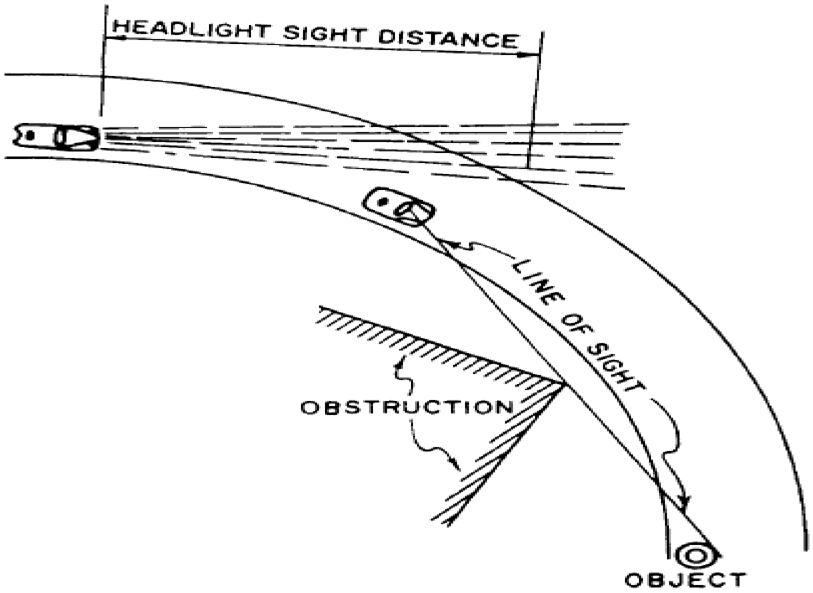

.—Lines of sight

5.3.1.1 TYPES OF HORIZONTAL CURVES

There are four types of horizontal curves. They are described as follows:

1. SIMPLE. The simple curve is an arc of a circle. The radius of the circle determines the sharpness or flatness of the curve.

2. COMPOUND. Frequently, the terrain will require the use of the compound curve. This curve normally consists of two simple curves joined together and curving in the same direction

3. REVERSE. A reverse curve consists of two simple curves joined together, but curving in opposite direction. For safety reasons, the use of this curve should be avoided when possible

4. SPIRAL. The spiral is a curve that has a varying radius. It is used on railroads and most modem highways. Its purpose is to provide a transition from the tangent to a simple curve or between simple curves in a compound curve.

5.3.1.2 ELEMENTS OF A HORIZONTAL CURVE

The elements of a circular curve are shown in figure 11-3. Each element is designated and explained as follows:

PI POINT OF INTERSECTION. The point of intersection is the point where the back and forward tangents intersect. Sometimes, the point of intersection is designated as *V (vertex).*

I INTERSECTING ANGLE. The intersecting angle is the deflection angle at the *PI.* Its value is either computed from the preliminary traverse angles or measured in the field. A CENTRAL ANGLE. The central angle is the angle formed by two radii drawn from the

Figure

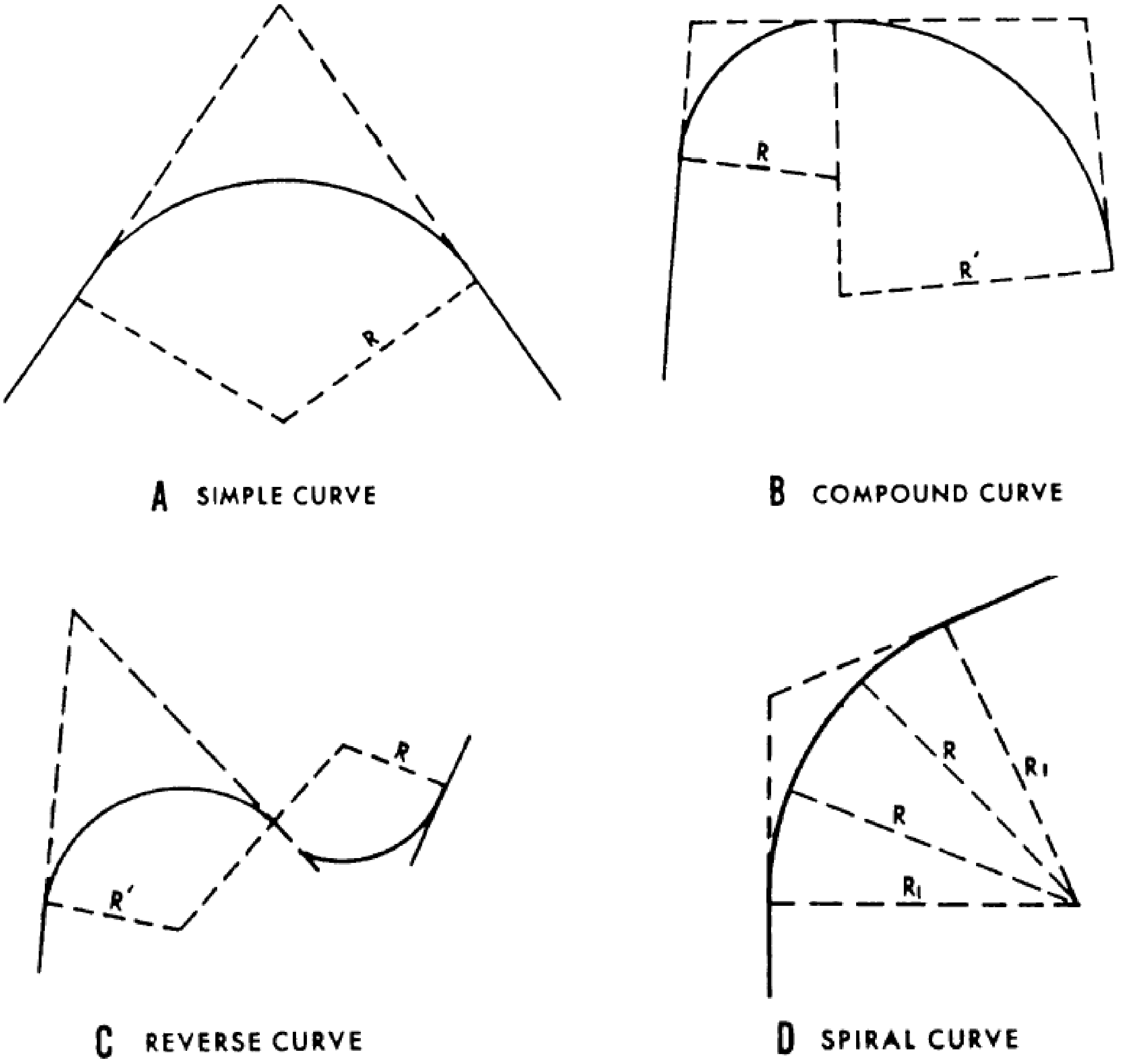
R
R
R′
R′
R
R
R
R₁
R₁

A SIMPLE CURVE

B COMPOUND CURVE

C REVERSE CURVE

D SPIRAL CURVE

Figure.—Horizontal curves.

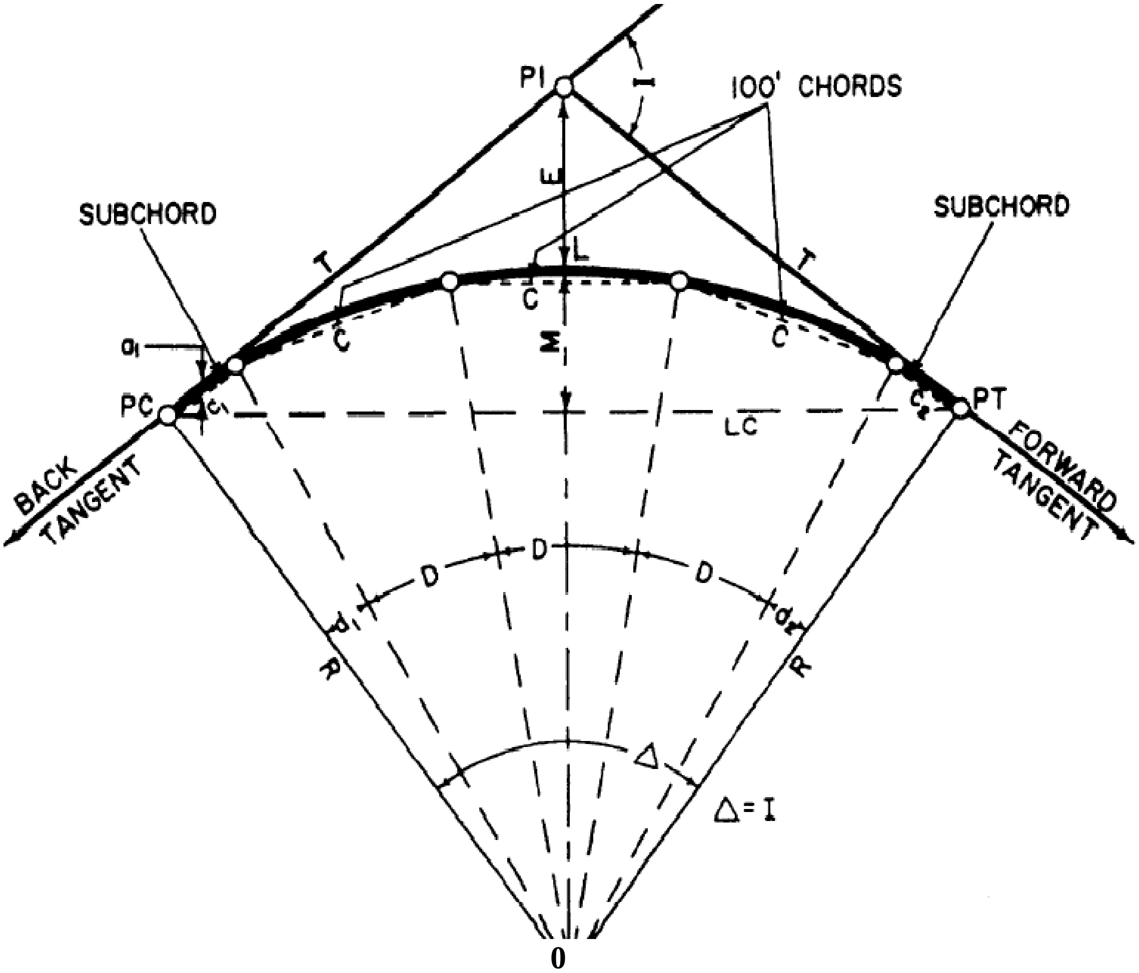

Figure.—Elements of a horizontal curve.

center of the circle *(O)* to the *PC* and *PT.* The value of the central angle is equal to the *I* angle. Some authorities call both the intersecting angle and central angle either *I* or A. RADIUS. The radius of the circle of which the curve is an arc, or segment. The radius is always perpendicular to back and forward tangents.

POINT OF CURVATURE. The point of curvature is the point on the back tangent where the circular curve begins. It is sometimes designated as *BC* (beginning of curve) or *TC* (tangent to curve).

POINT OF TANGENCY, The point of tangency is the point on the forward tangent where the curve ends. It is sometimes designated as *EC* (end of curve) or *CT* (curve to tangent).

POC POINT OF CURVE. The point of curve is any point along the curve.

LENGTH OF CURVE. The length of curve is the distance from the *PC* to the *PT,* measured along the curve.

TANGENT DISTANCE. The tangent distance is the distance along the tangents from the *PI* to the *PC* or the *PT.* These distances are equal on a simple curve.

LONG CHORD. The long chord is the straight-line distance from the *PC* to the *PT.* Other types of chords are designated as follows: *C* The full-chord distance between adjacent stations (full, half, quarter, or onetenth stations) along a curve. *c]* The subchord distance between the *PC* and the first station on the curve.

c? The subchord distance between the last station on the curve and the *PT.*

EXTERNAL DISTANCE. The external distance (also called the external secant) is the distance from the *PI* to the midpoint of the curve. The external distance bisects the interior angle at the *PI.*

MIDDLE ORDINATE. The middle ordinate is the distance from the midpoint of the curve to the midpoint of the long chord. The extension of the middle ordinate bisects the central angle.

DEGREE OF CURVE. The degree of curve defines the sharpness or flatness of the curve.

DEGREE OF CURVATURE

The last of the elements listed above (degree of curve) deserves special attention. Curvature may be expressed by simply stating the length of the radius of the curve. That was done earlier in the chapter when typical radii for various roads were cited. Stating the radius is a common practice in land surveying and in the design of urban roads. For highway and railway work, however, curvature is expressed by the degree of curve. Two definitions are used for the degree of curve. These definitions are discussed in the following sections.

Degree of Curve (Arc Definition)

The arc definition is most frequently used in highway design. This definition, , states that the degree of curve is the central angle formed by two radii that extend from the center of a circle to the ends of an **arc** measuring 100 feet long (or 100 meters long if you are using metric units). Therefore, if you take a sharp curve, mark off a portion so that the distance along the arc is exactly 100 feet, and determine that the central angle is 12°, then you have a curve for which the degree of curvature is 12°; it is referred to as a 12° curve.

By studying, you can see that the ratio design speed and allowable superelevation. Then the between the degree of curvature *(D)* and 360° is the radius is calculated. same as the ratio between 100 feet of arc and the circumference *(C)* of a circle having the same radius. That may be expressed as follows:

$$\frac{D}{360^\circ} = \frac{100}{C}.$$

Since the circumference of a circle equals $2\pi R$, the above expression can be written as:

$$\frac{D}{360^\circ} = \frac{100}{2\pi R}.$$

Solving this expression for R:

$$R = \frac{5729.58}{D}$$

and also D:

$$D = \frac{5729.58}{R}.$$

For a 1° curve, $D = 1$; therefore $R = 5{,}729.58$ feet, or meters, depending upon the system of units you are using. In practice the design engineer usually selects the 5,729.65 feet, or meters, depending upon the system of degree of curvature on the basis of such factors as the design speed and allowable superelevation. Then the between the degree of curvature *(D)* and 360° is the radius is calculated.

Degree of Curve (Chord Definition)

The chord definition is used in railway practice and in some highway work. This definition states that the degree of curve is the central angle formed by two radii drawn from the center of the circle to the ends of a **chord** 100 feet (or 100 meters) long. If you take a flat curve, mark a 100-foot chord, and determine the central angle to be 0°30', then you have a 30-minute curve (chord definition). From observation of figure 11-5, you can see the following trigonometric relationship:

$$\sin\left(\frac{D}{2}\right)=\frac{50}{R}.$$

Then, solving for R:

$$R=\frac{50}{\sin \frac{1}{2} D}.$$

For a 10 curve (chord definition), D = 1; therefore R = In practice the design engineer usually selects the 5,729.65 feet, or meters, depending upon the system of degree of curvature on the basis of such factors as the units you are using. **Degree of curve (chord definition). 11-**

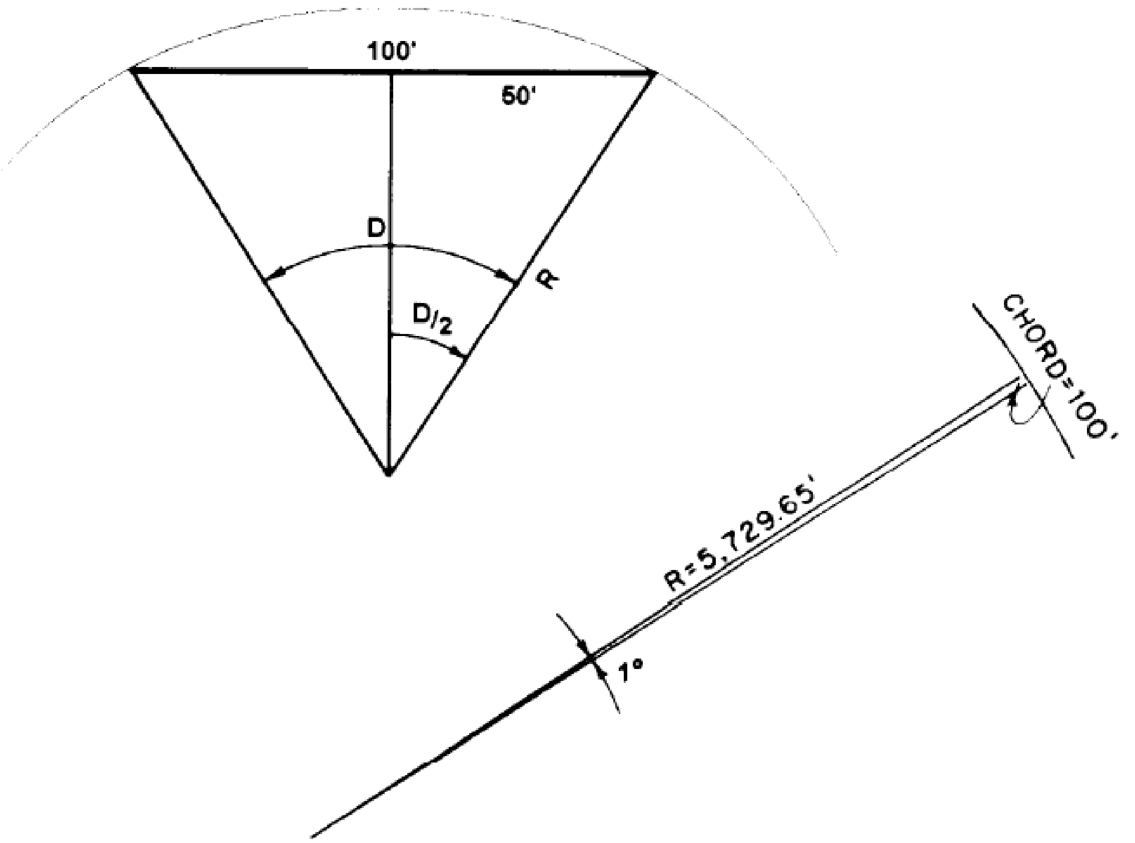

.—Degree of curve (chord definition).

Notice that in both the arc definition and the chord definition, the radius of curvature is inversely proportional to the degree of curvature. In other words, the larger the degree of curve, the shorter the radius; for example, using the arc definition, the radius of a 1° curve is 5,729.58 units, and the radius of a 5° curve is 1,145.92 units. Under the chord definition, the radius of a 1° curve is 5,729.65 units, and the radius of a 5° curve is 1,146.28 units.

5.4 CURVE FORMULAS

The relationship between the elements of a curve is expressed in a variety of formulas. The formulas for radius *(R)* and degree of curve *(D),* as they apply to both the arc and chord definitions, were given in the preceding discussion of the degree of curvature. Additional formulas you will use in the computations for a curve are discussed in the following sections.

Tangent Distance

Length of Curve By studying, you can see that the solution for the tangent distance *(T)* is a simple right-triangle solution. In the figure, both *T* and *R* are sides of a right triangle, with *T* being opposite to angle N2. Therefore, from your knowledge of trigonometric functions you know that

$$\tan \frac{\Delta}{2} = \frac{T}{R}$$

and solving for *T,*

$$T = R \tan \frac{\Delta}{2}.$$

Chord Distance

By observing figure 11-7, you can see that the solution for the length of a chord, either a full chord *(C)* or the long chord *(LC),* is also a simple right-triangle solution. As shown in the figure, *C/2* is one side of a right triangle and is opposite angle N2. The radius *(R)* is the hypotenuse of the same triangle. Therefore,

$$\sin \frac{\Delta}{2} = \frac{C/2}{R}$$

and solving for *C:*

$$C = 2R \sin \frac{\Delta}{2}.$$

Length of Curve

In the arc definition of the degree of curvature, length is measured along the arc, In this figure the relationship between *D, & L,* and a 100-foot arc length may be expressed as follows:

$$\frac{L}{100} = \frac{\Delta}{D}.$$

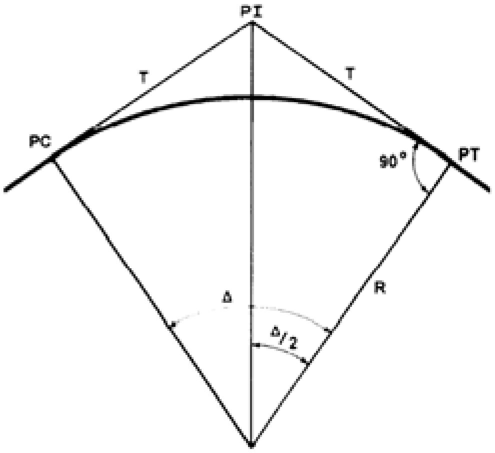
PI
T
T
PC
90°
PT
R
Δ
Δ/2

—Tangent distance.

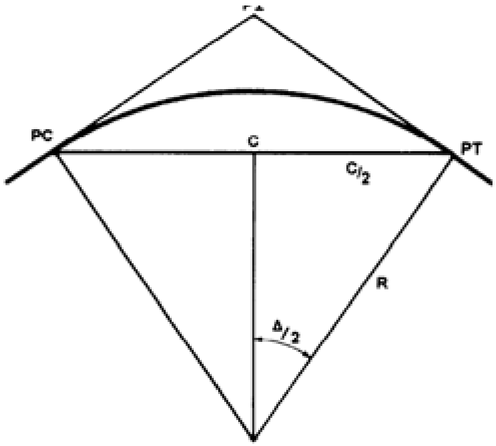
PC
C
PT
C/2
R
Δ/2

—Chord distance.

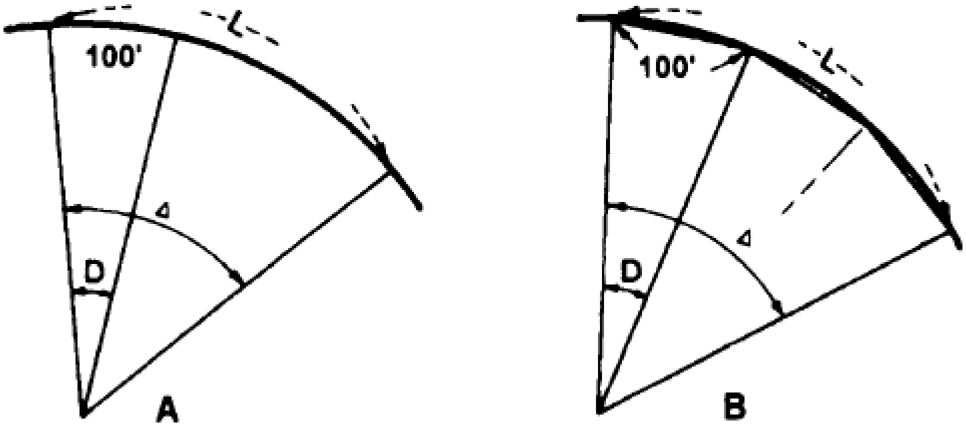
100'
L
Δ
D
A
100'
L
Δ
D
B

-Length of curve.

Then, solving for *L,*

$$L = 100\,\frac{\Delta}{D}.$$

This expression is also applicable to the chord definition. However, *L.,* in this case, is not the true arc length, because under the chord definition, the length of curve is the sum of the chord lengths (each of which is usually 100 feet or 100 meters), As an example, if, as shown in view B, , the central angle (A) is equal to three times the degree of curve *(D),* then there are three 100-foot chords; and the length of "curve" is 300 feet.

Middle Ordinate and External Distance

Two commonly used formulas for the middle ordinate *(M)* and the external distance *(E)* are as follows

$$M = R\left(1 - \cos\frac{\Delta}{2}\right)$$

$$E = T\tan\frac{\Delta}{4} = R\left(\frac{1}{\cos \Delta/2} - 1\right).$$

DEFLECTION ANGLES AND CHORDS

From the preceding discussions, one may think that laying out a curve is simply a matter of locating the center of a circle, where two known or computed radii intersect, and then swinging the arc of the circular curve with a tape. For some applications, that can be done; for example, when you are laying out the intersection and curbs of a private road or driveway

with a residential street. In this case, the length of the radii you are working with is short. However, what if you are laying out a road with a 1,000- or 12,000- or even a 40,000-foot radius? Obviously, it would be impracticable to swing such radii with a tape. In usual practice, the stakeout of a long-radius curve involves a combination of turning **deflection angles** and measuring the length of chords *(C,* Cl, or CZ as appropriate). A transit is set up at the PC, a sight is taken along the tangent, and each point is located by turning deflection angles and measuring the chord distance between stations. This procedure In below figure, you see a portion of a curve that starts at the PC and runs through points (stations) A, B, and C. To establish the location of point A on this curve, you should set up your instrument at the PC, turn the required deflection angle (all/2), and then measure the required chord distance from PC to point A. Then, to establish point B, you turn deflection angle D/2 and measure the required chord distance from A to B. Point C is located similarly. As you are aware, the actual distance along an arc is greater than the length of a corresponding chord; therefore, when using the arc definition, either a correction is applied for the difference between arc

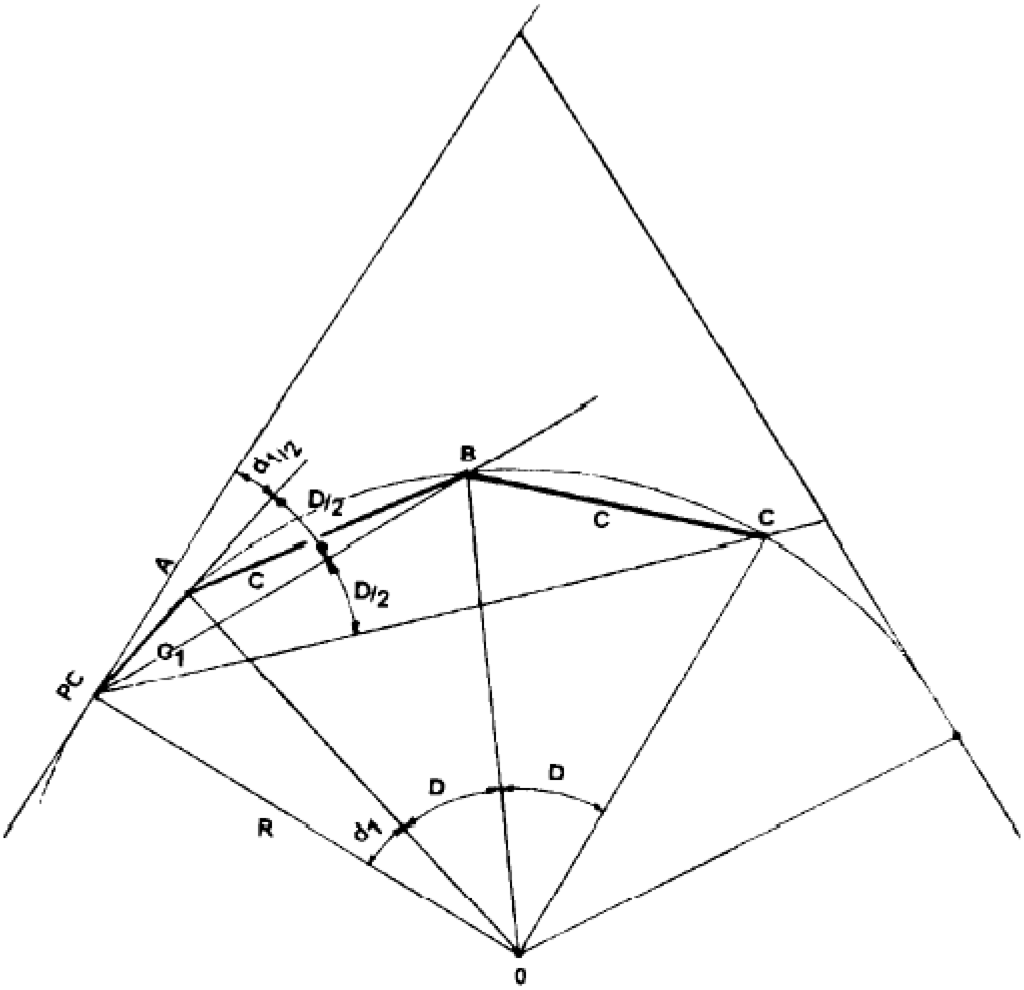

.-Deflection angles and chords.

length and chord length, or shorter chords are used to make the error resulting from the difference negligible. In the latter case, the following chord lengths are commonly used for the degrees of curve shown:

100 feet—0 to 3 degrees of curve
50 feet—3 to 8 degrees of curve
25 feet—8 to 16 degrees of curve
10 feet-over 16 degrees of curve

The above chord lengths are the maximum distances in which the discrepancy between the arc length and chord length will fall within the allowable error for taping. The allowable error is 0.02 foot per 100 feet on most construction surveys; however, based on terrain conditions or other factors, the design or project engineer may determine that chord lengths other than those recommended above should be used for curve stakeout. The following formulas relate to deflection angles: (To simplify the formulas and further discussions of deflection angles, the deflection angle is designated simply as *d* rather than *d/2.)*

$$d = \left(\frac{D}{2}\right)\left(\frac{C}{100}\right)$$

Where:

d = Deflection angle (expressed in degrees)
C = Chord length
D = Degree of curve
$d = 0.3\ CD$
Where:
d = Deflection angle (expressed in minutes)
C = Chord length
D = Degree of curve

$$\sin d = \frac{C}{2R}$$

5.5 SOLVING AND LAYING OUT A SIMPLE CURVE

Now let's solve and lay out a simple curve using the arc definition, which is the definition you will more often use as an EA. let's assume that the directions of the back and forward tangents and the location of the *PI* have previously been staked, but the tangent distances have not been measured. Let's also assume that stations have been set as far as Station 18 + 00. The specified degree of curve *(D) is* 15°, arc definition. Our job is to stake half-stations on the curve.

Solving a Simple Curve

We will begin by first determining the distance from Station 18 + 00 to the location of the *PI.* Since these points have been staked, we can determine the distance by field measurement. Let's assume we have measured this distance and found it to be 300.89 feet. Next, we set up a transit at the *PI* and determine that deflection angle *I* is 75°. Since *I* always equals A, then A is also 75°, Now we can compute the radius of the curve, the tangent distance, and the length of curve as follows:

$$R = 5{,}729.58/D = 381.97 \text{ ft}$$

$$T = R \tan \Delta/2 = 293.09 \text{ ft}$$

$$L = 100\ \Delta/D = 500 \text{ ft}$$

5.6 VERTICAL CURVES

In addition to horizontal curves that go to the right or left, roads also have vertical curves that go up or down. Vertical curves at a crest or the top of a hill are called **summit curves,** or **oververticals.** Vertical curves at the bottom of a hill or dip are called **sag curves,** or **underverticals.**

5.6.1 COMPUTING VERTICAL CURVES

As you have learned earlier, the horizontal curves used in highway work are generally the arcs of circles. But vertical curves are usually **parabolic.** The parabola is used primarily because its shape provides a transition and, also, lends itself to the computational methods described in the next section of this chapter. Designing a vertical curve consists principally of deciding on the proper length of the curve. the length of a vertical curve is the **horizontal distance** from the beginning to the end of the curve; the length of the curve is NOT the distance along the parabola itself. The longer a curve is, the more gradual the transition will be from one grade to the next; the shorter the curve, the more abrupt the change. The change must be gradual enough to provide the required sight distance. The sight distance requirement will depend on the speed for which the road is designed; whether passing or nonpassing distance is required; and other assumptions, such as one's reaction time, braking time, stopping distance, height of one's eyes, and height of objects. A typical eye level used for designs is 4.5 feet or, more recently, 3.75 feet; typical object heights are 4 inches to 1.5 feet. For a sag curve, the sight distance will usually not be significant during daylight; but the nighttime sight distance must be considered when the reach of headlights may be limited by the abruptness of the curve.

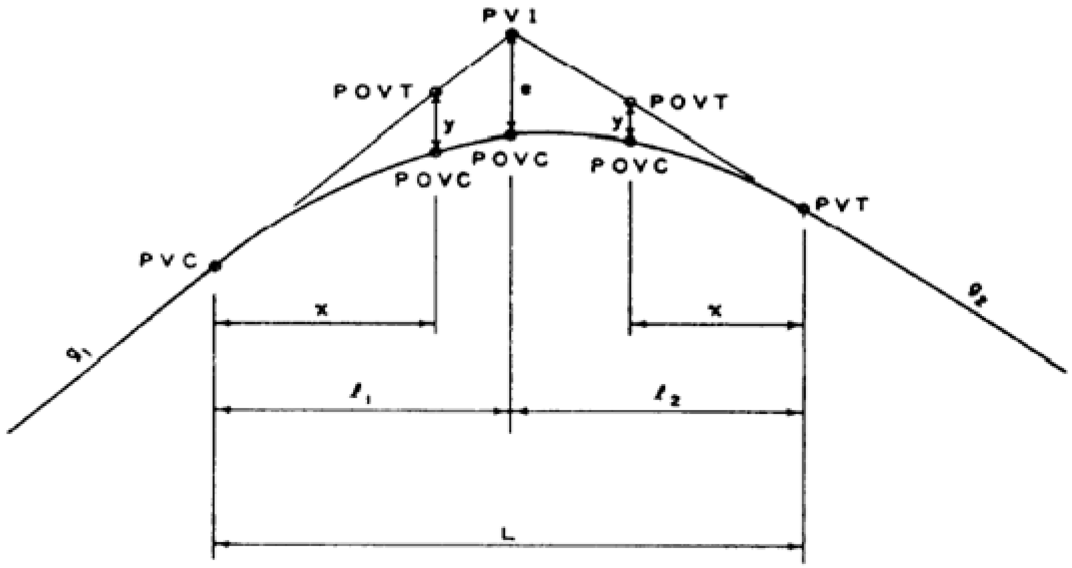

Elements of a vertical curve.

5.6.2 ELEMENTS OF VERTICAL CURVES

the elements of a vertical curve. The meaning of the symbols and the units of measurement usually assigned to them follow:

PVC Point of vertical curvature; the place where the curve begins.

PVI Point of vertical intersection; where the grade tangents intersect

PVT Point of vertical tangency; where the curve ends.

POVC Point on vertical curve; applies to any point on the parabola

POVT Point on vertical tangent; applies to any point on either tangent

gI Grade of the tangent on which the *PVC* is located; measured in percent of slope

g2 Grade of the tangent on which the *PVT* is located; measured in percent of slope.

The **algebraic difference** of the grades: G = g2 -g,, wherein plus values are assigned to uphill grades and minus values to downhill grades; examples of various algebraic differences are shown later in this section. Length of the curve; the **horizontal** length measured in 100-foot stations from the *PVC* to the PVT. This length may be computed using the formula ***L = G/r,*** where *r* is the rate of change (usually given in the design criteria). When the rate of change is not given, *L* (in stations) can be computed as follows: for a summit curve, **L = 125** x G/4; for a sag curve, **L = 100** x G/4. If *L* does not come out to a whole number of stations using these formulas, then it is usually extended to the nearest whole number. You should note that these formulas for length are for road design only, NOT railway.

L1 Horizontal length of the portion of the *PVC* to the *PVI;* measured in feet.
L2 Horizontal length of the portion of the curve form the *PVI* to the *PVT;* measured in feet. e Vertical (external) distance from the *PVI* to the curve, measured in feet. This distance is computed using the formula e **= LG/8,** where *L* is the total length in stations and G is the algebraic difference of the grades in percent.

5.7 Contour

CONTOUR

- Defined as the line of intersection of a level surface with the surface of ground.
- Lines drawn on the map to join points of the same height on the earth surface.
- The best method of representation of features such as hills etc.

CONTOUR INTERVAL

1. The constant height difference between two contour lines.
2. Depends upon the following factors:
   - The nature of the ground
   - The purpose and Extent of the survey
   - The scale of the map
   - Time and Expense of Field and Office work.

EXAMPLE OF CONTOUR INTERVAL

1. For large scale maps of flat country, for building sites for detailed design work
   - 0.2m to 0.5m
2. For reservoirs and town planning scheme.
   - 0.5m to 2m
3. For location surveys
   - 2m to 3m
4. For small scale maps of broken country and General Topographical work
   - 3m, 5m, 10m or 25m.

CHARACTERISTICS OF CONTOUR LINES

1. CONSAVE
2. CONVEX
3. VERTICAL SLOPE
4. CLIFF
5. VALLEY
6. HILL SPUR

USE OF CONTOUR MAP AND CONTOUR PLAN

1. Study original shape of the earth surface
2. Identify most suitable site and saving cost for engineer works
3. Identify water catchments area
4. Calculate capacity or volume of a pond
5. Identify volume of cut and fill
6. Identify intervisibility between two points
7. Identify the slope of road
8. Draw longitudinal section and cross section to view the shape of earth surface.

9. Calculate horizontal distance between contour lines.

METHOD OF CONTOURING

1. DIRECT METHOD
   - The contours to be located are directly traced out in the field by locating and making a number of points on each contour
2. INDIRECT METHOD
   - The points located and surveyed are not necessarily on the contour lines but the spot levels.
   - GRID LEVEL METHOD
   - CROSS SECTIONING METHOD
   - SPOT HEIGHT METHOD

GRID LEVEL METHOD

- This method is most systematic and favoured by many because the contouring process is easy to understand.
- Suitable for flat and open survey area.

Contour maps are very useful since they provide valuable information about the terrain. Some of the uses are as follows:

i) The nature of the ground and its slope can be estimated

ii) Earth work can be estimated for civil engineering projects like road works, railway, canals, dams etc.

iii) It is possible to identify suitable site for any project from the contour map of the region.

iv) Inter-visibility of points can be ascertained using contour maps. This is most useful for locating communication towers.

v) Military uses contour maps for strategic planning.

the contour map, a common type of drawing in civil engineering. It is an ideal means of representing a threedimensional surface using a single two-dimensional view. Contour maps not only convey a qualitative impression of the features of the surface but also enable, using a single view, complete quantitative information to be extracted from the drawing. Although they are often used for topographic maps, contour maps have other applications that will be briefly mentioned in the article.

1. what it is

We will define what a contour map actually is by considering a simple example of representing a three-dimensional landform using only two dimensions.A three-dimensional surface is given. For example, we can consider a the landform shown in the figure to the right. The perspective view shown is obviously a two-dimensional representation of a threedimensional object. This conveys a reasonable qualitative impression of the overall characteristics of the landform, but it does have the following shortcomings:

1. Depending on the point from which the view is taken, we may miss valuable information. For example, when the landform is viewed to obtain the upper of the two views shown to the right, we have no information on what lies beyond the red line. We need to rotate to a different viewpoint to obtain this information.

2. There is no reliable way of extracting quantitative information from this drawing. For example, we may wish to know the elevation of a specific point on this landform or the difference in elevation between two given points. Not only is it difficult to locate a given point exactly in a horizontal plane, it is likewise difficult to gain more than a general impression of its elevation. It is thus not possible to use this drawing to answer questions such as “What is the elevation of a point 400 m to the north and 300 m to the east of Point A”.We can solve the first problem to some extent by using multiple views of the same landform. This is shown in the figure to the right, where we have rotated the original viewpoint by approximately 90 degrees to obtain a second view. With the help of the second view, we can see what lies beyond the ridge. By selecting a sufficient number of views, we can generally provide a correct qualitative impression of the entire landform. The problem with this approach is that it is not compact, in the sense that one view is generally not enough. Furthermore, to be of value as a basis for extracting quantitative information, this approach also requires a means of relating one view to another. This is not a straightforward task

We can solve the second problem to some extent by identifying specific points and writing in the elevation of these points on the drawing. This has been done for two points in the drawing to the right. Elevations are given in metres above sea level. This increases the quantitative content of the drawing. If a sufficiently large number of such elevations were given, it would be possible to estimate, at least approximately, elevations of other points by interpolation. There is still no reliable quantitative basis, however, for locating points in the horizontal plane. It is practically impossible, for example, to know the distance and bearing of the point with elevation 1019 relative to the point with elevation 1178. Reliable answers to questions relating to the elevation of a point of known coordinates relative to a given point thus remain difficult to obtain. Neither providing multiple perspective views nor providing elevation values for given points thus allows us to represent the three dimensional surface using a single two-dimensional drawing that enables quantitative information to be extracted (i.e., the z coordinate of a point given its x and y coordinates).

To accomplish this objective, we need a more suitable two dimensional representation of the three dimensional surface. To develop this representation, we return to the previous perspective views. We first imagine that the landform has been sliced by a horizontal plane of constant elevation. In this case, say its elevation is 1000 m. Where this plane cuts the landform defines one or more curves in a horizontal plane. We can erase the plane itself but leave the line created by the intersection of the plane and the landform. This curve joints points of equal elevation. We define any curve joining points of equal elevation a contour line, or simply a contour

We can repeat this construction for planes at other elevations. For example, we can do so for planes with elevation 1100 m and 900 m. We can provide greater detail by showing more contour lines. In the lower view to the right, we have shown contours at increments of 20 m. We say that the contour interval in this drawing is 20 m. This contour interval appears to cover the landform reasonably well and capture the changes in topography. Although this drawing conveyssignificantly more information than the original perspective view given on the first page of this article, it is in itself is not particularly useful, since it has most of the shortcomings of the drawings developed initially. But it can be transformed into a powerful drawing by representingthis information on a horizontal plane

We do this by viewing the landform, with the contour lines, looking directly down from a point above it. By choosing this viewpoint, we gain a dimensionally true representation of the horizontal plane, which permits us to use true x-y coordinates to locate points. The vertical dimension disappears visually but is now represented in a more abstract way through the contour lines, which now appear to be drawn on a horizontal plane.A given point can now be located and measured in x-y coordinates from any other point on the map. Its elevation can be read directly from the contour that intersects the point or, for points located in between contours, by interpolation. For example, a point 500 m to the east and 500 m to the north of the origin in the lower left portion of the drawing (shown with the red dot) is found to have an elevation of approximately 1008 m

It is important that contour maps always incorporate a constant contour interval. By doing this, we can get a visual sense of the threedimensional properties of the surface, even when the shading of the original landform has been removed, by considering the patterns formed by the contours.The closest path from one contour line to an adjacent contour gives the steepest path. This follows directly from the definition of slope, which is equal to rise over run. For a constant rise (fixed contour interval), the greatest slope corresponds to the smallest value of run(closest distance between contours). It therefore follows that closely spaced contours denote regions that are steep and widely spaced contours are regions that are relatively flat.Closed curves denote either “hills” or “depressions”. We distinguish between the two by considering whether the contours are increasing or decreasing. In the case of the map shown to the right, the change of the contours indicates that the triangular figure enclosed by the green rectangle would be the top of a hill.

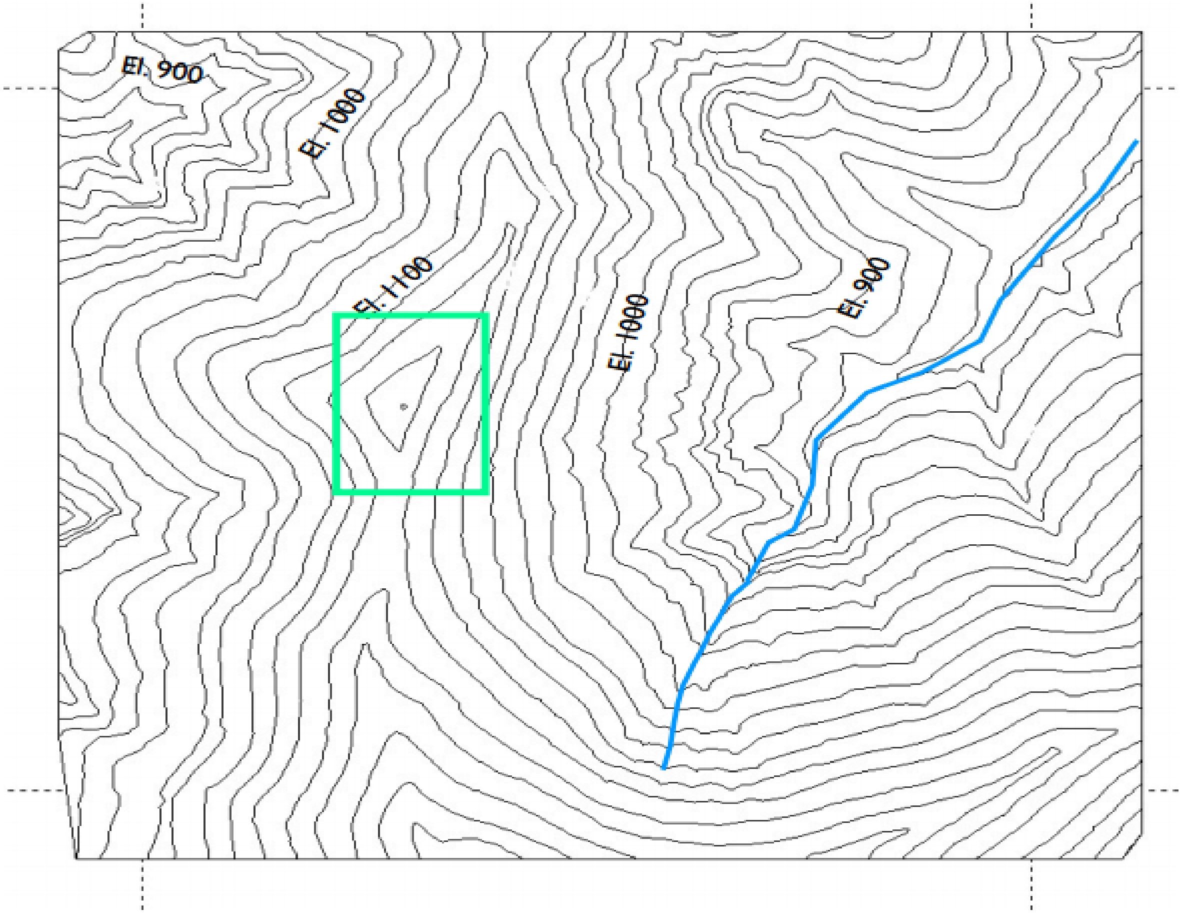

Landform with contours viewed from above. Contour interval is 20 m.

A series of adjacent contours that all “point” in the same direction often indicates the path of a river, since this corresponds to the landform created by the flow of water through the earth. The blue line drawn onto the map to the right indicates one possible river.We can summarize the essence of contour maps as follows. Contour maps allow us to represent three-dimensional surfaces using a single two-dimensional drawing. They maintain the ability to locate points accurately in two horizontal dimensions. Everything in these two dimensions is drawn to scale. We lose a direct means of visualizing the third dimension, but are able to represent it accurately through lines joining points of equal elevation, called contour lines. These are equivalent to the curve formed by intersecting the surface of the given landform with planes of constant elevation. By working with a constant contour interval, we gain the ability to visualize the three dimensional characteristics of the surface. The data used to generate the landform considered in this example originate from the National Map of Switzerland. The corresponding section of this map is shown in the figure to the right. The contour interval in this case is 10 m

2. suitable applications

Contour drawings are not the best way to represent all threedimensional objects. They are best suited to representing objects that have a single surface, a significant and well defined reference plane, and a significant third coordinate perpendicular to the reference plane.It follows from these conditions that contour drawings are well suited to the representation of landforms with a single two-dimensional drawing. They have a single surface (the surface of ground), a well defined reference plane (the horizontal plane) and a significant third coordinate perpendicular to the reference plane (the z coordinate represents elevation, which is of primary significance).Contour drawings are not suitable for representing other types of three dimensional objects when at least one of these conditions is not satisfied. The bridge pictured in the figure to the right, for example, does not have a single surface, but rather several including a near vertical surface, a far vertical surface, and an upper surface. A single contour diagram cannot adequately represent all of these surfaces. This type of object is best represented in other ways, such as with multiple views based on standard viewing planes. In these drawings only two dimensions are depicted in each view. No quantitative (and often no qualitative) information regarding the third coordinate can be extracted from a given view. For this reason, more than one view is required to describe the object completely.

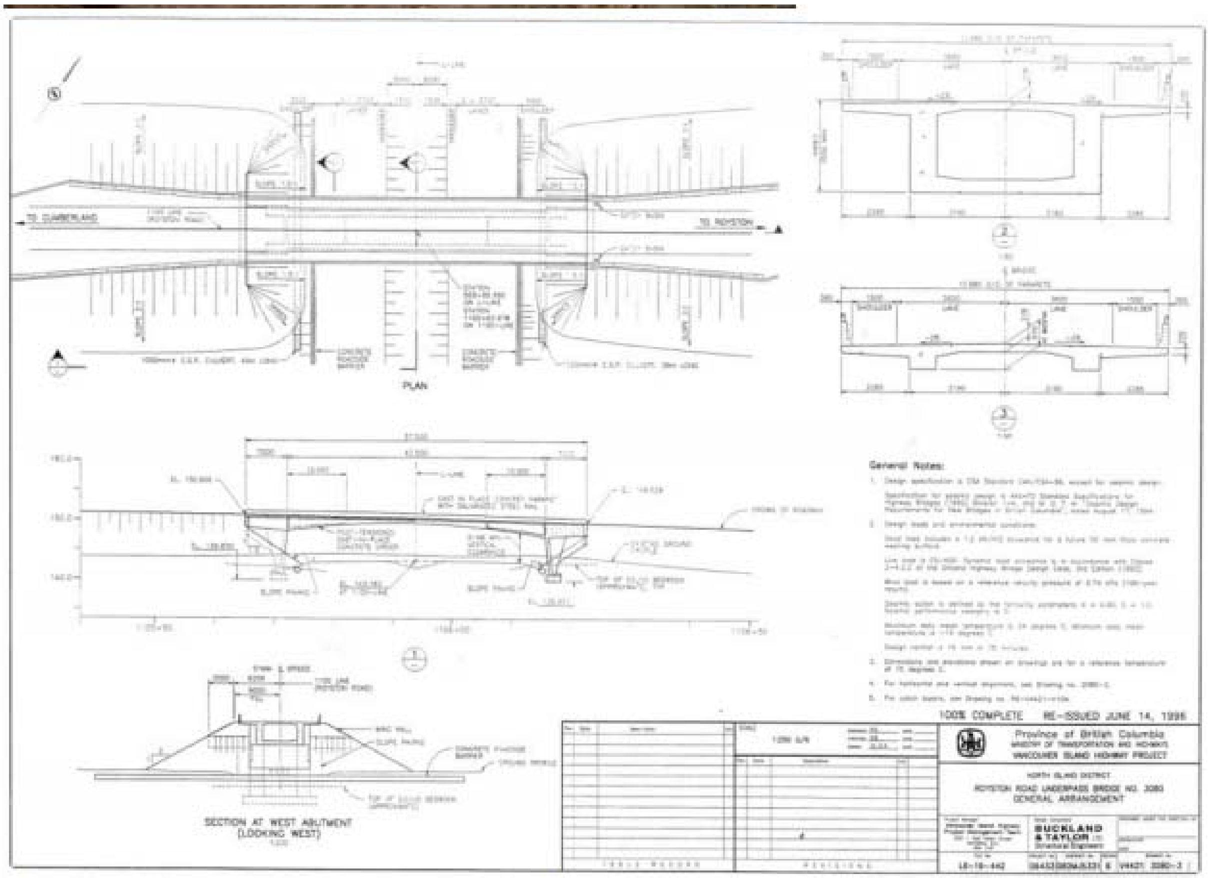
TO CUMBERLAND
TO ROYSTON
PLAN
General Notes:
SECTION AT WEST ABUTMENT (LOOKING WEST)
100% COMPLETE RE-ISSUED JUNE 14, 1996
Province of British Columbia
BUCKLAND & TAYLOR

Bridge represented by multiple two-dimensional views

3. how to do it

There are several ways to produce contour drawings from a set of x, y, and z coordinates describing a given three-dimensional surface. This section describes one way that is relatively straightforward. Given:

1. A regular square grid of points, with x, y, and z coordinates of the surface defined for each point of the grid. This grid is shown in the image to the right. It has nine points. The scale in the horizontal plane is given graphically. Elevations in metres are given for each grid point.
Required: Produce a contour drawing representing the surface defined by the given x, y, and z coordinates.
How to proceed:

1. Set the contour interval. To do this, it is necessary first to scan the z coordinates to extract the minimum and maximum values. Within this range, define an interval that is regular and that captures the relevant features of the surface with good fidelity. In this case, regular means taken from the series 1 m, 2 m, 5 m, 10 m, 20 m, 50 m, 100 m, etc. When working with a set of several diagrams, it is usually preferable to use a constant contour interval over the entire set of diagrams to enable comparisons across the set of drawings. In such cases, the choice of interval should be made in consideration of the properties of the entire set of data.

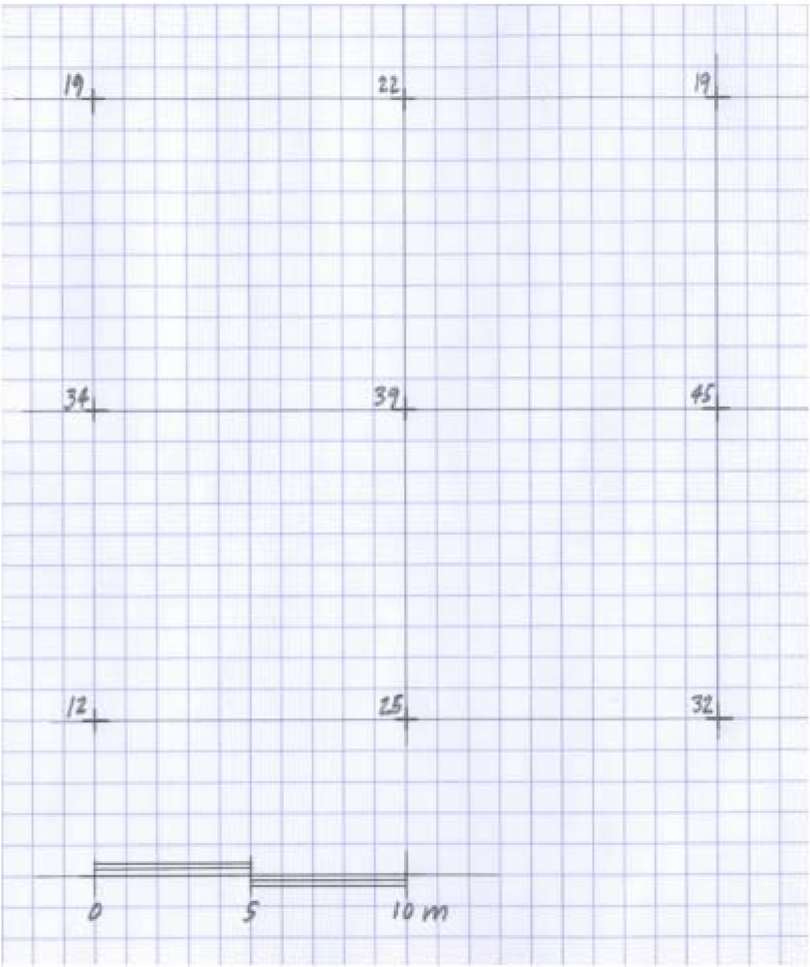

Given information for contour map

For this example, the minimum and maximum elevations are 12 and 45 m respectively. For simplicity, we will use a contour interval of 10 m for this example. So relevant contours will be at the 20, 30, and 40 m elevations.

2. Draw the grid to a suitable scale. As always, use regular scales (i.e., from the series 1:1, 1:2, 1:5, 1:10, 1:20, 1:50, 1:100, etc...). In this case, this has already been done for us.

3. For easy reference, write in the z values next to the corresponding points of the grid. This has already been provided.

4. For each segment joining adjacent points of the grid, identify points of intersection of contour lines with the segment, based on the assumption that the change in z within a given segment of the grid is linear. Proceed according to the following example:

(a) Given: The top left horizontal segment AB of the grid has the following z values: z(A)=19.0 m, z(B)=22.0 m.

(b) So one contour will intersect this segment. It is the 20 m contour, since 19 < 20 < 22.

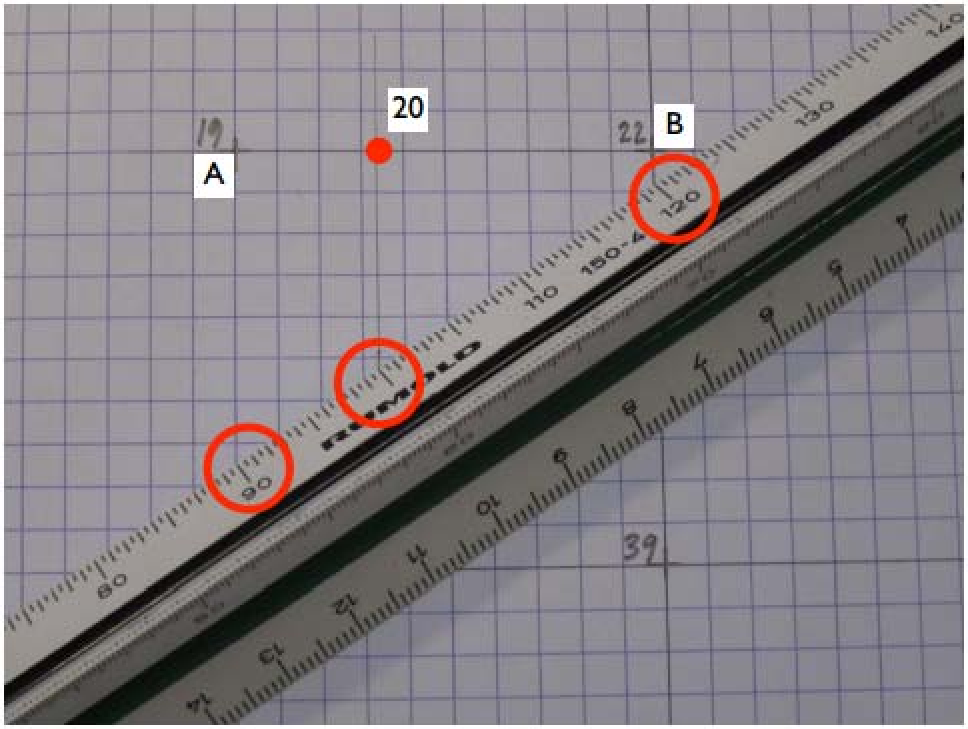

Locating intersection of contour and gridline

(c) Locate the point of intersection of the contour within this segment by linear interpolation. One straightforward way to accomplish this is to use a scale in a way similar to the method used to subdivide a line into several equal segments:

(i) On the vertical gridline passing through the left end point, align the scale to a value corresponding to the elevation at that location. In this case, the scale is set to 90.

(ii) On the vertical gridline passing through the right end point, align the scale to a value corresponding to the elevation at that location. In this case, the scale is set to 120. The scale remains at 90 along the left gridline. The series 90, 100, 110, 120 defined by the scale is similar to the series 19, 20, 21, 22 defined by the given elevations. So the intersection of the 20 m contour with the given line segment will correspond to 100 on the scale.

(iii)Draw a line perpendicular to the given segment corresponding to the point of intersection identified with the scale. This locates the intersection of the contour with the given segment.

(iv)Write the value of the contour next to the intersection point.

(v) Note: the accuracy of this procedure increases as the angle between the scale and Segment AB gets smaller. So it is usually helpful to try fitting several scales to the given segment to minimize this angle.

(vi)The outcome of this phase of the process is shown in the figure to the right.

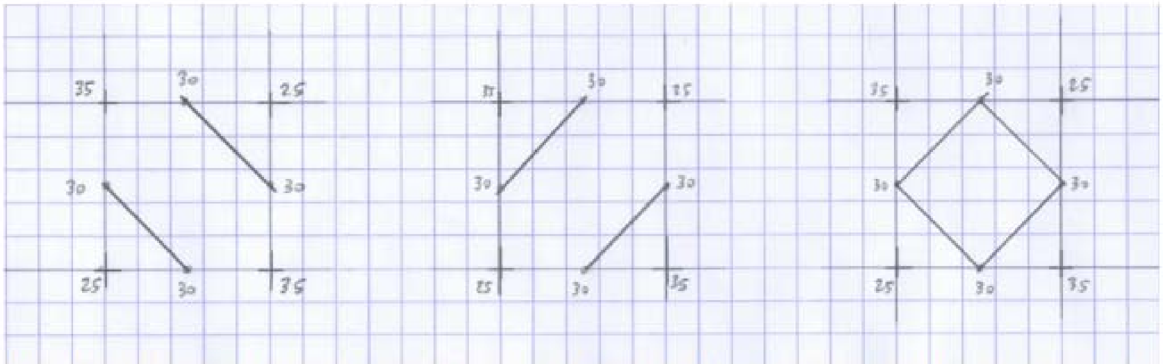

Irregular situation: contour intersects grid square on four sides. There are several possible arrangements of contours, only one of which correctly represents the three dimensional figure shown

(d) When all of the points of intersection of contours and segments have been thus identified, draw the contour lines. Proceed on a square by square basis. For a given square bounded by four adjacent grid points, the following two cases must be considered:

(i) A given contour intersects exactly two bounding segments of the square. This is the case, for example, for the 20 m contour in the upper left-hand square. In this case, simply draw a line joining the points of intersection. This line is the path of the contour within the square. The image to the right shows the 20 m and the 30 m contours drawn for the upper left hand square in the grid.

(ii) A given contour intersects all four bounding segments of the square. In this case, it is not clear how to draw the contours. The figure to the right shows that a single arrangement of intersecting points can correspond to several arrangements of contours. In this case, only one arrangement (the middle one) corresponds to the given three-dimensional figure shown.

(e) The completed contours are shown in the figure to the right.

(f) Once the complete contour diagram has been drawn, trace the contours onto a new sheet of paper. This leaves only the contours and does not show the working grid and other marks that were made to produce them. It is generally necessary to label specific contours and spot elevations. As with all plan views (i.e. top views), a north arrow is required.

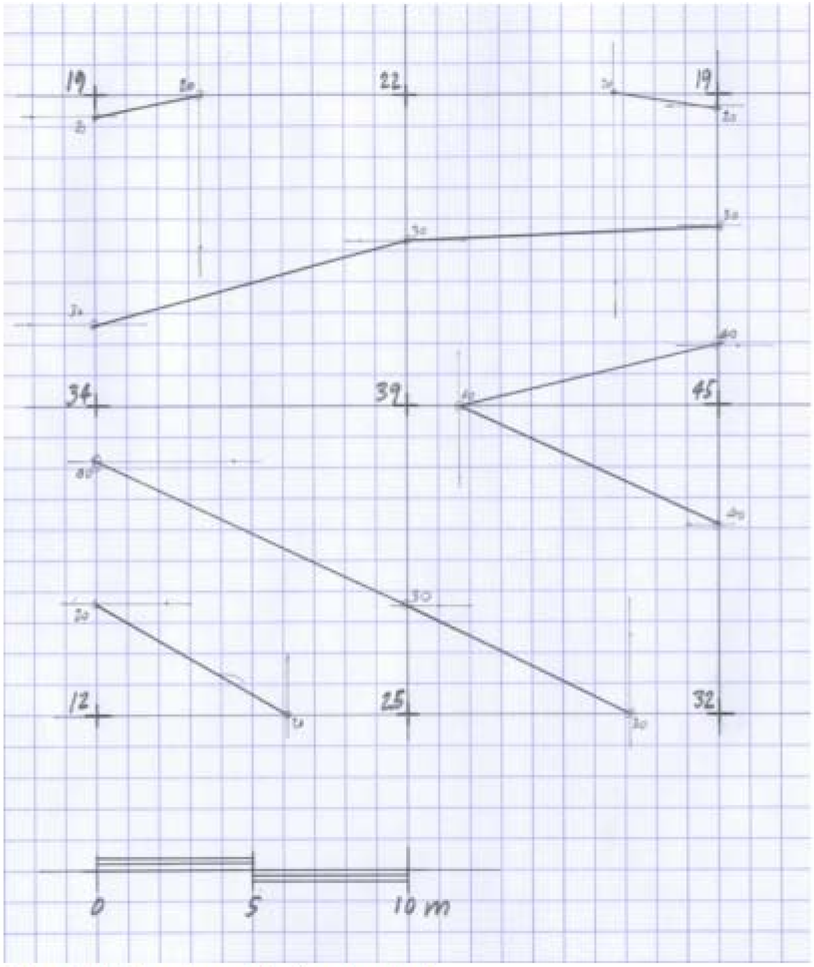

Completed contours (before tracing)

4. variations

In some cases, it is preferable to draw smooth curves for the contours. This will often provide a more realistic rendition of the features of a given landscape. All of the contour maps we will draw in this course will be done by straight line segments linking points of equal elevation along gridlines, as described in the previous section.It is common to create contour maps from survey data obtained in the field. In such cases, it is sometimes difficult to get elevation values for a square grid of points in the plane.

It is also possible to create a contour map following the principles outlined in the previous section for an irregular collection of points. In such a case, it is necessary to establish a triangular network of lines joining the available points in the plane. This is shown in the leftmost diagram. Along these lines, contour values are interpolated, as shown in the middle diagram. Finally, for a given triangle, straight line segments are drawn linking points on the boundary of identical contour value. This is shown in the rightmost diagram, where contour lines have been highlighted in green.

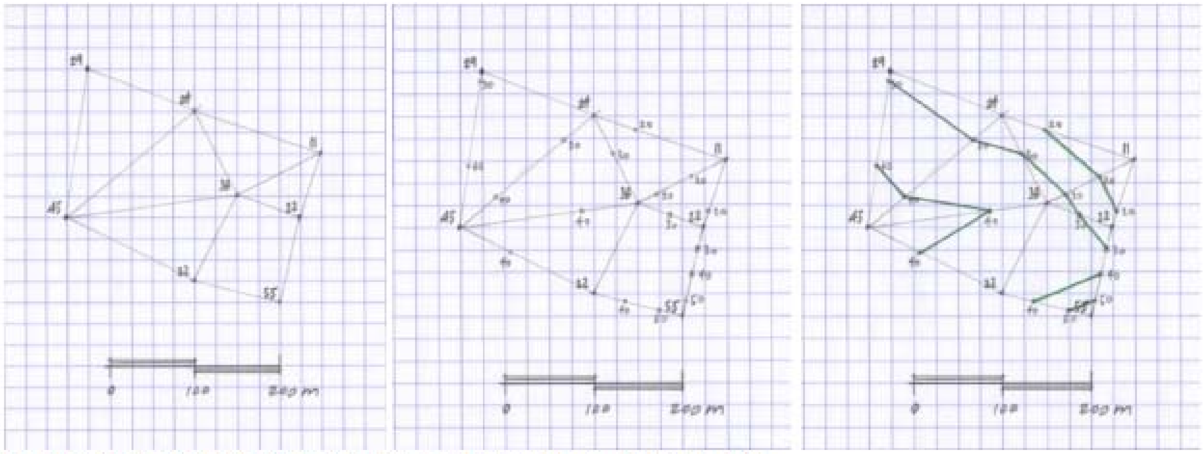

Contour mapping when points are not arranged on a square grid

5. examples from practice This section describes several types of contour drawing in common use in engineering.

5.1. Standard Topographic Maps Topographic maps describe, with a high level of detail and accuracy, the topography (i.e., the shape) of a given geographical area. The standard way of representing the three dimensional features of landforms is contour lines.

The first example is from Canada’s National Topographic System of maps. The most detailed scale available is 1:50 000.

The second example is from the National Map of Switzerland. This map is drawn to a scale of 1:50 000.

The third example is also from the National Map of Switzerland, this time from their 1:25 000 series.

The Swiss maps are produced with much greater detail and with additional visual cues to help the user gain a qualitative impression of the three-dimensional landforms from the contours.

This is accomplished by: (1) a relatively small contour interval (in this case 10 metres), (2) subtle shading that corresponds to the shadows that would be cast on the landforms when the sun shines from the northwest quadrant of the map, and (3) pictorial symbols such as the cliff symbol, which is used when the slope of the land is so steep as to cause excessive bunching of the contour lines

5.2. Project-specific topographic plans and diagrams When the level of detail given on standard topographic maps is insufficient, it is possible to produce topographic plans for a given site based on specific survey data. These plans are used, for example, for the layout of bridges. Inthis diagram, for example, the contour interval is 5 m, which is considerably less than the contour interval used on standard topographic maps.

This type of map will generally be prepared by a specialist land survey firm.Although contour diagrams are most often used to represent natural features such as topography, they are sometimes used to represent features of the facility to be built. Contour maps are sometimes made,
for example, of bridge decks to validate that drainage will work properly.

5.3. Contour graphs

It is also common to use contours to represent abstract surfaces, i.e., mathematical functions of two variables $z=f(x,y)$. In such cases, x and y need to be spatial coordinates in a well defined and meaningful plane. Function z then defines a three-dimensional surface, similar to a landform.The same principles used to draw contour maps of physical landforms can be used to draw contour graphs of such functions.

The figure to the right shows one such application. This diagram is called an influence surface for a slab free along the bottom, fixed in shear and bending along the top, and extending to infinity in the other two directions. It is based on the function $z = Ma(x,y)$, where A is a given fixed point and Ma is the bending moment at Point A due to a unit load applied at Point (x,y). It is the two-dimensional analog of the one-dimensional influence line.

This type of diagram can be used to calculate bending moments in bridge deck slabs due to loads applied by the wheels of a heavy truck. Its use is illustrated in the figure to the right.

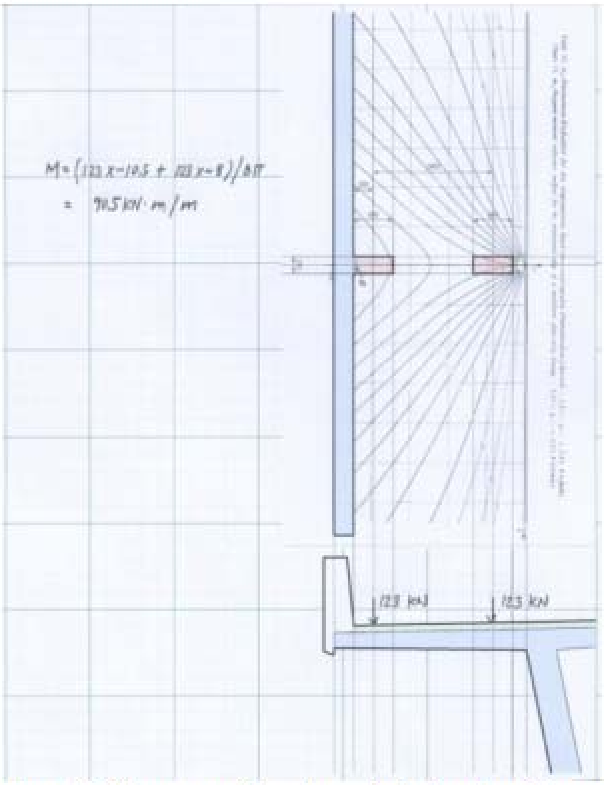

Use of influence surfaces in calculating bending moments in cantilever bridge deck slabs

6. applications: cutting sections from contour maps

Contour maps are useful in and of themselves, but they are also used as a basis for producing relevant two-dimensional drawings. These can be visualized as the curve formed by the intersection of the given three dimensional surface and a vertical plane. The curve thus formed is often referred to as a section. The process of drawing this curve from a given contour map is referred to as cutting a section.

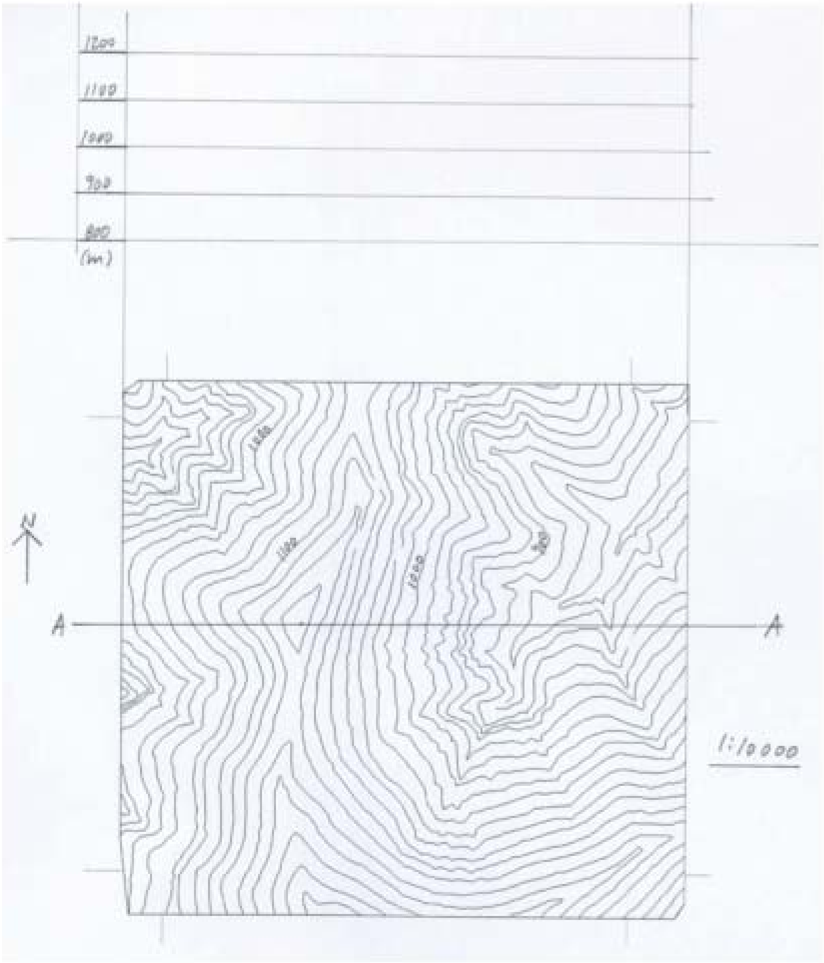

Setting up to cut section from contour map along Line A-A

A common application of sections cut from contour maps is the production of the elevation of a valley to be crossed by a bridge. The procedure for cutting such a section is relatively straightforward.

Given:
1. Contour map of the area under consideration
2. A straight line drawn on the contour map locating the vertical plane that defines the section (Line A-A).Required:

A two-dimensional drawing (x-z view) showing the shape of the landform along the given line.
How to proceed:

1. Draw a line parallel to Line A-A on a blank portion of the page. This will be the horizontal datum of the section to be cut.

2. Based on these maximum and minimum elevations intersected by Line A-A, draw a vertical scale of elevations to the left of, and perpendicular to, the horizontal datum. Provide suitable labels to this axis. Draw horizontal gridlines based on the labels given on the axis. The drawing to the right represents the progress thus far. The area just created is called the section diagram.

3. For each contour that intersects Line A-A, do the following:

(a) Identify the points of intersection of the contour and Line A-A. We will call one such point Point P1.

(b) Draw lines perpendicular to Line A-A from Points Pi to the section diagram. The outcome of this step is shown in the diagram to the right.

(c) Draw horizontal lines in the section diagram corresponding to the contour elevations.

(d) Identify the points of intersection Q1 for specific perpendiculars originating from the contours and the corresponding elevation in the section diagram.

4. Join the points Qi to form a continuous curve. The resulting curve is the section cut along Line A-A. The figure to the right shows the finished section cut along Line A-A. The location of one point of the profile is highlighted. It corresponds to the 1000 level contour.

Review Questions

1. Explain engineering surveyes?
2. What is Reconnaissance in detail?
3. What is Preliminary and location surveys for engineering projects explain?
4. What is curves?
5. Explain the all types of curves?
6. What is horizontal and vertical curves?
7. What is CONTOUR?
8. Describe the Contouring – Methods with detail?
9. What are the characteristics of contour method?
10. What are the uses of contour method?

“The lesson content has been compiled from various sources in public domain including but not limited to the internet for the convenience of the users. The university has no proprietary right on the same.”

Rai University
EVOLVING THINKING MINDS

Saroda Post, Dholka Taluka, District - Ahmedabad, Gujarat - 382260, India

www.raiuniversity.edu

WILDERNESS NAVIGATION HANDBOOK

- Understand maps and charts
- Use navigational tools
- Interpret natural clues
- Navigate in desert, glacier, jungle, and ocean environments

FRED TOUCHE

Wilderness Navigation Handbook

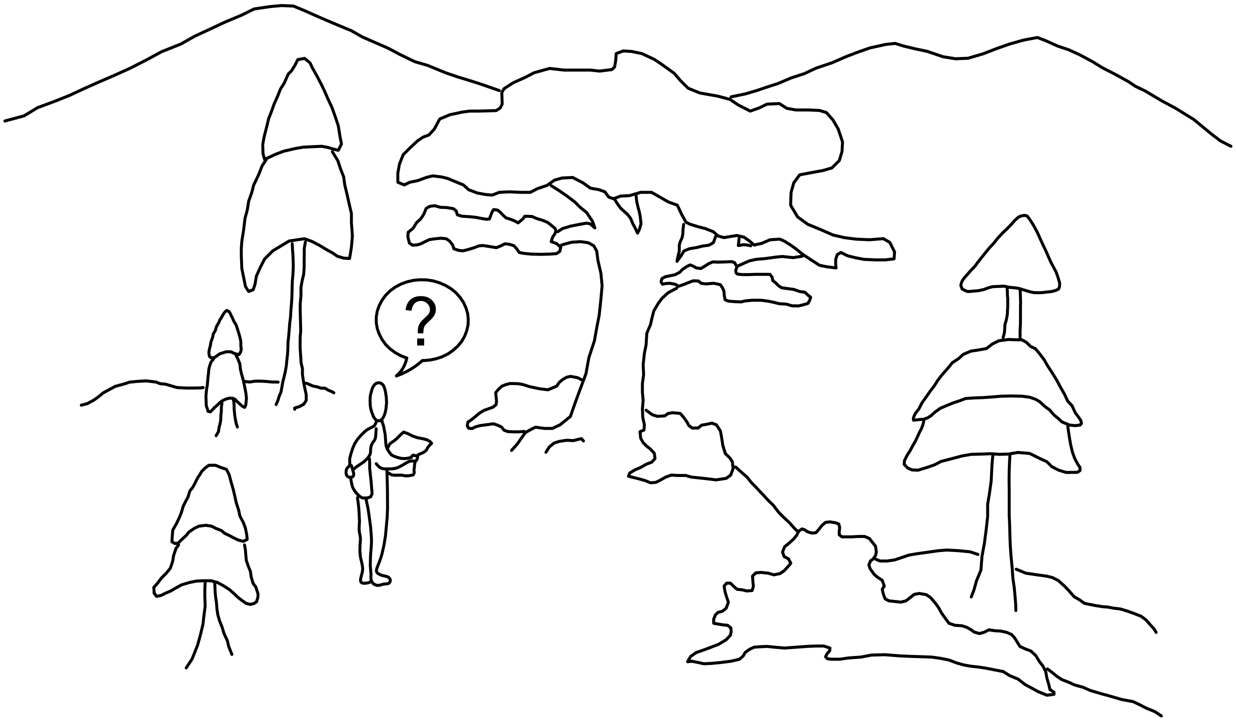

Artistic illustrations: Anne Price

Technical illustrations: Fred Touche

Copyright © 2004 by Fred Touche

All rights reserved. Other than brief quotations in reviews, no part of this book may be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the author.

Warning

Be aware that navigation is only one part of whatever wilderness activity you choose to engage in. Your chosen activity is likely to be dangerous and carry significant risk of personal injury or death and should only be undertaken with a full understanding of all the inherent risks. In addition, poor implementation of navigation techniques may augment the inherent risks associated with your chosen activity.

There are no guaranties, either expressed or implied, that this book contains accurate or complete information. Your most important navigation tool is your brain. Before using any information from this book, make sure this information makes sense. The author accepts no responsibility for your lack of judgment.

National Library of Canada Cataloguing in Publication

Touche, Fred, 1954-
Wilderness navigation / Fred Touche ; Anne Price, illustrator.
Second printing, 2005

Includes index.
ISBN-13: 978-0-9732527-0-5
ISBN-10: 0-9732527-0-7

1. Orienteering—Equipment and supplies. 2. Navigation—Equipment and supplies. I. Price, Anne, 1980- II. Title.

GV200.5.T68 2004 796.58 C2004-901366-1

Comments on this book

For questions or comments regarding the content of this book, or to order additional copies, email author at: fredtouche@gmail.com

Cover design by Abacus Graphics – Oceanside, California

Printed in Canada by Friesens Corporation

Printed on 20% post-consumer waste paper.

Table of Contents

Foreword 8

Acknowledgments 9

Introduction 10

1. Maps 12

1.1 Legend 12
1.2 Scale 13
1.3 How to measure distances on a map 14
1.4 Geographic coordinate system 15
  1.4.1 Geographic coordinates on maps 17
1.5 Map projections 18
  1.5.1 Mercator projection 19
  1.5.2 Lambert conformal conic projection 19
  1.5.3 Universal transverse Mercator projection 20
  1.5.4 Polar stereographic projection 21
1.6 UTM grid 21
  1.6.1 UTM zones 21
  1.6.2 Easting and northing in UTM grid 25
  1.6.3 UTM grid on maps 26
  1.6.4 Truncated grid numbers 27
1.7 UPS grid 28
  1.7.1 UPS zones 28
  1.7.2 Easting and northing in UPS grid 28
1.8 Map datums 30
  1.8.1 Earth’s shape 30
  1.8.2 Horizontal datums 30
  1.8.3 Vertical datums 32
1.9 Topographic maps 33
  1.9.1 Definition of contour lines 33
  1.9.2 Finding elevation of unlabeled contour line 34
  1.9.3 Calculating slopes 34
  1.9.4 How contour lines represent terrain features 36
  1.9.5 How to interpret terrain on a map 40
1.10 Nautical charts 41

2. Compass 43
- 2.1 Earth's magnetic field 43
  - 2.1.1 Magnetic field theory 43
  - 2.1.2 Magnetic pole movement 45
- 2.2 How a compass works 46
- 2.3 Compass bearings 48
- 2.4 Magnetic declination 50
  - 2.4.1 Magnetic declination theory 50
  - 2.4.2 Estimating magnetic declination 51
  - 2.4.3 Compensating for magnetic declination 55
- 2.5 Magnetic dip 56
- 2.6 Orienteering compasses 57
  - 2.6.1 Taking a bearing off a map 58
  - 2.6.2 Taking a field bearing 60
  - 2.6.3 Following a compass bearing 61
  - 2.6.4 Triangulating from a linear feature 63
  - 2.6.5 Triangulating off multiple landmarks 65
- 2.7 Marine compass 67
  - 2.7.1 Taking a bearing off a nautical chart 68
  - 2.7.2 Taking a field bearing with a marine compass 69
  - 2.7.3 Following a compass bearing on the water 70
- 2.8 Orienting a map or chart 71
  - 2.8.1 By terrain association 71
  - 2.8.2 With orienteering compass and grid lines 72
  - 2.8.3 With orienteering compass and declination diagram 73
  - 2.8.4 With marine compass and compass rose 74

3. Altimeter 76
- 3.1 How altimeters work 76
- 3.2 Calibrating an altimeter 77
- 3.3 Factors affecting altimeter accuracy 77
- 3.4 Triangulating with altimeter and compass 79
- 3.5 Estimating your location by steepest slope 81

4. Global Positioning System 84
- 4.1 How GPS works 84
- 4.2 Causes of GPS position errors 85

4.3 Differential GPS (DGPS) 87
4.4 Space-based augmentation systems 88
4.5 GPS receivers 89
    4.5.1 Signal gathering techniques 89
    4.5.2 Features that can improve accuracy 90
4.6 Essential applications for GPS receivers 90
4.7 GPS receiver basic use 92
4.8 Entering waypoints into a GPS receiver 93
    4.8.1 Manual transfer of waypoints from map to receiver 93
    4.8.2 Electronic transfer of waypoints into GPS receiver 95
4.9 Zeroing-in on a waypoint 95

5. Celestial Navigation 97

5.1 Celestial coordinate system 97
5.2 The sun 99
    5.2.1 Using a clock to find the sun’s bearing 99
    5.2.2 Stick shadow method 105
5.3 The moon 106
    5.3.1 Estimating the moon’s bearing 108
5.4 Stars 109
    5.4.1 Finding north or south 112
    5.4.2 Using stars to determine rough positions 113
    5.4.3 Stars as direction finders 116
5.5 Sextant navigation 117
    5.5.1 Underlying theory 117
    5.5.2 How the sextant works 121
    5.5.3 Taking a sight 122
    5.5.4 Sextant calculations 123

6. Natural Navigation 130

6.1 Effects of sunlight 130
6.2 Effects of wind 131
6.3 Landforms 132
6.4 Ocean clues 133
6.5 Making a compass 134
6.6 Using thermometer to determine elevation 135

7. Emergency Communication 136
7.1 Emergency transmitters 136
    7.1.1 ELTs and EPIRBs 136
    7.1.2 Personal locator beacon (PLB) 137
7.2 Two-way radios 137
    7.2.1 FRS radios (walkie-talkies) 138
    7.2.2 VHF/UHF radios 139
    7.2.3 HF radios 140
    7.2.4 Cellular phones 140
    7.2.5 Satellite phones 141
    7.2.6 Miscellaneous wireless devices 141
7.3 Signaling systems 141
    7.3.1 Flares 142
    7.3.2 Mirrors 142
    7.3.3 Light signals 143
    7.3.4 Fires 143
    7.3.5 Acoustic devices 144
    7.3.6 International ground signals 144

8. Practical Navigation 146
8.1 Terrain association – map only 146
8.2 Conventional map and compass navigation 149
8.3 Dead reckoning – no landmarks 151
    8.3.1 Keeping track of bearings 151
    8.3.2 Keeping track of distance 151
    8.3.3 Detouring around obstacles 152
    8.3.4 How to navigate with dead reckoning 153
8.4 Navigating with map, compass, and altimeter 156
8.5 Navigating with map, compass, and GPS 158
    8.5.1 Navigating with pre-entered waypoints 158
    8.5.2 Navigating without pre-entered waypoints 160
8.6 Navigating without maps 160
    8.6.1 Navigating with nothing 160
    8.6.2 Navigating with just a compass 162
    8.6.3 Navigating with just a GPS receiver and compass 163
8.7 Real life navigation 164
8.8 Lost in the wilderness 164
    8.8.1 Waiting for a rescue 165
    8.8.2 Finding your way out 166

8.9 Rhumb lines and great circle routes 167
8.10 Route finding 168

9. Scenarios 169

9.1 Trail switchbacks 169
9.2 Multiple ski tracks 170
9.3 Shoreline travel direction 171
9.4 Mountain top triangulation 172
9.5 Hidden trail markers 173
9.6 Jungle camp 174
9.7 Ridge return 175
9.8 Desert traverse 176
9.9 River delta 177
9.10 Descent off mountain 178
9.11 Trail return 179
9.12 Return to highway 179
9.13 Large group 181
9.14 Disappearing trail 183
9.15 Whiteout 183
9.16 Trail from lake 185
9.17 No GPS signal 186
9.18 Back roads 187
9.19 Food stash 188
9.20 Antarctic ordeal 189
9.21 Polynesian journey 191

Index 194

Foreword

The idea for this book started innocently just before the turn of the millennium. There were certain facts about navigation that had puzzled me for years, but for some reason could not find clear explanations to. A frustrating search led to long hours in libraries and surfing the web where I eventually managed, piece-by-piece, to get some of the information that I wanted. “If only all this stuff was in one place” was a thought that led to the next idea: “why don’t I write a book about navigation.”

Having roamed around the wilderness since childhood, navigation and wilderness was a natural match. How could I go wrong with my newfound theoretical knowledge combined with years of unforgiving practical experience? As the millennium rolled around, I made a resolution to go ahead with the book. After an ambitious start, the project soon bogged down with unexpected difficulties. The book was continually morphing into a new creature that could not be tamed, and there were of course other things to do that were more fun.

Although at times painfully slow, progress never completely ground to a halt. After three years, there were things that looked like chapters of incomplete information. Realizing that nothing less than an all out push was going to finish off the project, I started working on the book full time. The book was finally completed four years after its conception. The result is hopefully something that you will find useful during your journeys of exploration.

Fred Touche

March 2004

Acknowledgments

After the initial draft of this book was completed, I asked several avid outdoor enthusiasts if they could review the material. Some were hesitant, while others were eager to help. I would like to thank all those who did sacrifice their valuable time to get involved in this project. Their contribution has been invaluable in making the Wilderness Navigation Handbook an authoritative text.

The first to get involved was Verena Shaw, who proofread the initial draft and did a great job straightening out the many spelling and grammatical errors. A concurrent technical review of the material was done by Paul Kubik and John Hebron who suggested several important changes to the subject matter. John, with his PhD in physics, seemed to take pleasure in debunking several of my key assumptions, particularly regarding astronomy and magnetism. A general review of the content was carried out by Don Funk and James Ram, who discovered yet more grammatical blunders and several vague passages that needed clarification. Jim Mandelli, an expert navigator with more than 30 adventure races under his belt, came onboard late in the game. His knowledge of the practical aspects of land navigation was indispensable. Christine Matterson performed an excellent, thorough final edit before the book was sent off to the printing presses.

I would like to acknowledge all those who helped with the assembly of the book. One of my most intractable problems was how to draw three-dimensional illustrations. Fortunately, Soegi Hartono, a computer aided drafting specialist, came to my rescue, effortlessly tutoring me on the nuances of computer graphics. Annika Lindberg and Wendy Johnstone assisted with the layout of the book, and Robert Dynowski helped with the cover design.

I wish to extend a special thanks to Anne Price for her artistic illustrations that have undoubtedly made a generally dry topic more fun to read, and to my parents for their financial support that has made the publication of this book possible. I am also grateful for the help that I received from many strangers during the long research phase, most who will never know that their expertise went into this book.

Introduction

Navigation is both an art and a science, requiring creativity as well as raw technical skills. It has always been an intricate part of human activity. In the past our very survival may have hinged on using natural navigation clues to find a decent food source and then retrace the route back to the cave. In stark contrast, present-day navigation systems can guide an aircraft across oceans and continents to a perfect landing in zero visibility. Future space travelers will no doubt use devices that are even more sophisticated. Despite these technological advancements, we often find ourselves frustrated by navigational tasks. Just ask anyone who has tried to find his or her way out of an unfamiliar city.

Studies have shown that having a general sense of direction is closely linked to the ability to mentally rotate objects. This may explain why some people easily become disoriented while others have seemingly no trouble finding their way. The widespread belief that some people, especially members of indigenous groups, can navigate purely by instinct is however a myth. Wherever you are on Earth, you are likely surrounded by clues that you can use for navigation. It isn’t so much an innate sense of direction that differentiates navigators, but rather the ability to recognize navigational clues and put them to effective use.

The introduction of the global positioning system (GPS) and the availability of low-cost GPS receivers has revolutionized navigation. Travel during poor visibility and through terrain devoid of landmarks can easily be done by anyone willing to put their trust in instruments. Then again, overreliance on technology is also a sure way to get into trouble. High-tech gadgets have a tendency to fail at the worst possible moment, and you may suddenly find yourself navigating like your ancestors.

Except for a chapter on emergency communication, this book is strictly about navigation. It does not deal with wilderness survival, route finding, or the sport of orienteering. The aim is to take away the mystique of navigation by explaining all the essential tools and techniques. If you are a raw beginner, concentrate on learning how to navigate with a map, compass, altimeter, and GPS receiver before diving into topics such as celestial navigation. If you are already a competent navigator, this book will almost certainly give you a deeper understanding of how things work, and more importantly, why things sometimes don’t work.

An earnest attempt was made to present the material in logical order and avoid disruptive references to previous or subsequent sections. Unfortunately, this had to balanced with keeping distinct topics together, and avoiding excessive repetition. No attempt has been made to separate basic from advanced topics. If you get bogged down, skip to the next section and return later to the material that you found challenging.

This book is dedicated to the memory
of John Millar and Guy Edwards who
never returned from the wilderness

Chapter 1
Maps

Maps use patterns, colors, and symbols to describe a portion of Earth's surface. Maps are never exact, nor complete, representations of the real world. Different types of maps provide different kinds of information. For example, a road map may show cities, roads, and highways while ignoring terrain features such as valleys or mountain ranges. For off road navigation, topographic maps that use contour lines to describe the shape of the terrain are superior to other types of maps. For ocean and coastal navigation, nautical charts that show the shape of the underwater topography are the preferred choice.

When selecting a map, make sure that it includes the following: (1) legend, (2) scale, (3) contour interval, (4) grid lines, and (5) declination diagram. The legend tells you the meaning of the symbols, lines, and colors on the map. The scale and contour interval allow you to calculate distances, elevations, and slopes from the map. Grid lines make it easier to pinpoint a position on the map. The declination diagram shows you how to adjust your compass to compensate for Earth's irregular magnetic field.

1.1 Legend

The map legend shows the meaning of the colors and symbols on your map. Unfortunately, there is no worldwide standard so the colors and symbols may mean different things on different maps. This is why it is important to consult the legend before using a map from an unfamiliar series.

Colors

Since there are only a few colors to choose from, colors normally show entities that cover significant portions of the map. For example, vegetation zones are shown in green, bodies of water in blue, glaciers in white, urban areas in red, and areas with little or no vegetation in brown.

Symbols

Map symbols are graphic representations of natural and man-made features. Point features such as campsites, lighthouses, or survey markers are shown as single symbols. Linear features such as roads, trails, power lines, railway tracks, rivers, or political boundaries are shown as dashed or continuous lines. Areas such as swamps, moraines, tidal flats, or sand dunes are shown as groups of identical symbols. Symbols are almost always shown at an exaggerated scale. Otherwise, they would be too small to be visible on the map.

Example of symbols

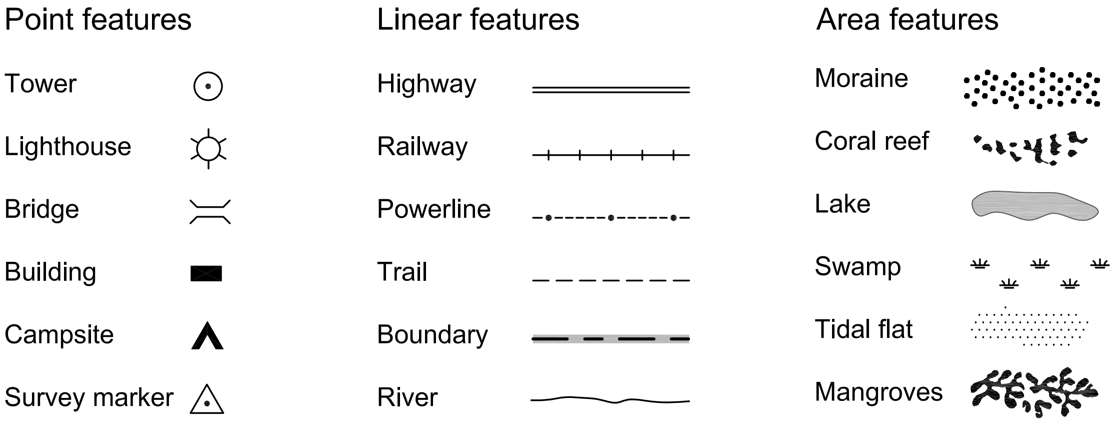

1.2 Scale

All maps or charts are drawn to a specific scale. The scale is the relationship between a distance on the map versus the same distance in the real world. A large-scale map shows a small area in great detail, while a small-scale map covers a large area but with little detail. Selecting a map with an appropriate scale is therefore a compromise between coverage and detail, and depends on your intended activity. For example, climbing a mountain with complex topography requires a large-scale map, while a small-scale map may be a better choice for a long canoe trip. The scale that you see written on the map is just an approximation. It is never completely constant across the whole map, but varies according to the projection that was used to create the map. The larger the area that a map covers, the more susceptible it is to scale variations. The scale can be expressed as a verbal scale, a ratio scale, or a bar scale.

Verbal scale

A verbal scale uses words to express the relationship between a distance on the map and the corresponding distance in the real world.

Example of verbal scale

The following is written in the margins of your map: “4 cm on the map represents 1 km on the ground.” Here it is implied that a distance of 4 centimeters on the map is equal to 1 kilometer in the real world. This relationship allows you to calculate any other real world distance from the map. For example, a measured distance of 15.6 cm on the map represents (15.6 cm)/(4 cm) x 1 km = 3.9 km in the real world.

Ratio scale

The scale of a map can be expressed as a ratio between a distance on the map versus the same distance in the real world. Some people find ratio scales confusing because distances are not tied to a specific unit. You can use any unit you want as long as you stick to the same unit during the conversion between map and real world distances. After doing the conversion, you can of course convert the resulting distance into other units.

Example of ratio scale

A map with “Scale 1:50 000” written in the margins means that the distance between any two points is 50 000 times longer in the real world than on the map. For example, a distance of 8.5 cm on the map would be 50 000 x 8.5 cm = 425 000 cm = 4250 m = 4.25 km in the real world.

Bar scale

A bar scale shows real world distances directly on the map. Bar scales are the most common type of scale, and are easy to understand because of their graphical nature. A common practice is to show several bar scales on the same map, each representing a different length unit. Bar scales are often shown together with ratio or verbal scales.

Examples of bar scales

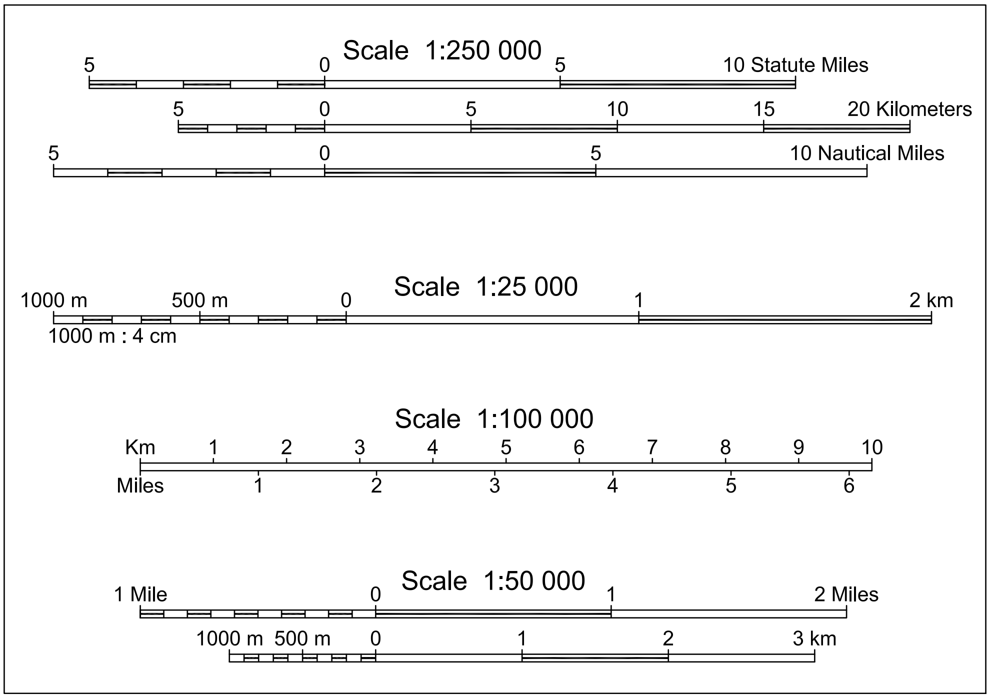
Scale 1:250 000
5 0 5 10 Statute Miles
5 0 5 10 15 20 Kilometers
5 0 5 10 Nautical Miles
Scale 1:25 000
1000 m 500 m 0 1 2 km
1000 m : 4 cm
Scale 1:100 000
Km 1 2 3 4 5 6 7 8 9 10
Miles 1 2 3 4 5 6
Scale 1:50 000
1 Mile 0 1 2 Miles
1000 m 500 m 0 1 2 3 km

1.3 How to measure distances on a map

To calculate the real world distance of a proposed route, you must first measure the corresponding distance on the map. This is easy if the route is a straight line between two points. Simply measure the distance with a ruler. If you don’t have a ruler, keep in mind that orienteering compasses often have rulers inscribed in their base plates. Another method is to use the grid lines that are drawn across some maps as a measuring device. The grid lines form an array of identical squares with sides of a specific length. After figuring out the dimensions of one square, the grid effectively becomes a ruler. For more complicated routes, break up your route into straight-line segments and measure the distance of each segment separately. Add the lengths of all the segments to calculate the total length of the route.

To measure a curved path, lay out a string along the route. Pinch the string with your fingers at the two end points of the route and then pull the string tight. Place the string along a ruler, or across grid squares, and read off the distance. If your map has a bar scale, place the string on the bar scale and directly read off the real world distance. Instead of a string, you can use a map wheel, also known as a curvimeter. Set the counter to zero and then roll the wheel along your route. The curvimeter will display the accumulated distance. Pay attention to the units. Curvimeters can either directly display real world distances for maps of certain scales, or just show the actual distance on the map. In the latter case, you must convert the map distance to the real world distance.

If you have access to digital maps and related software, a computer will calculate and display the distance of your route. A typical software package lets you create a route by clicking on your start point, end point, and any intermediate points. The program displays the cumulative distance between all the adjacent points. Some software packages allow you to trace curved routes on the digital map. This makes your calculated distance even more accurate.

1.4 Geographic coordinate system

The geographic coordinate system is one of the two widely used methods for pinpointing locations on a map. It covers Earth’s entire surface. A location is identified by specifying its latitude (parallel) and longitude (meridian). Latitudes and longitudes can be visualized as an array of imaginary lines running along Earth’s surface. A position is measured in angle units: degrees (°), minutes of arc ('), and seconds of arc ("). Each degree is subdivided into 60 minutes of arc, which in turn are subdivided into 60 seconds of arc. Sometimes decimal degrees are used instead of minutes and seconds of arc, or decimal minutes instead of seconds of arc. For example, 156°16'21" = 156°16.35' = 156.2725°.

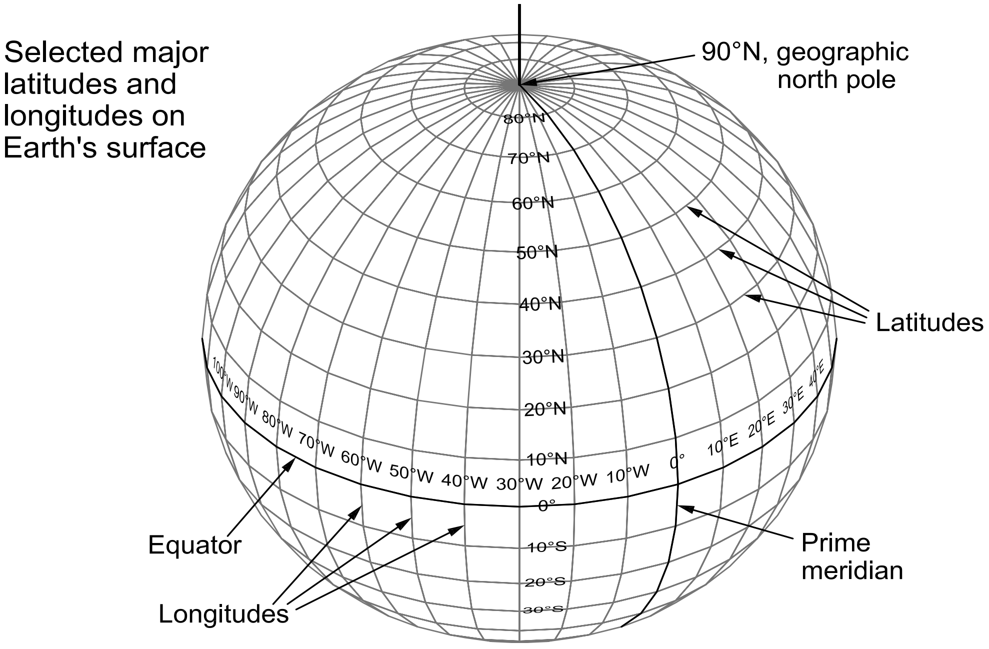
Selected major latitudes and longitudes on Earth's surface
90°N, geographic north pole
80°N
70°N
60°N
50°N
40°N
30°N
20°N
10°N
0°
10°S
20°S
30°S
Latitudes
100°W 90°W 80°W 70°W 60°W 50°W 40°W 30°W 20°W 10°W 0° 10°E 20°E 30°E 40°E
Equator
Longitudes
Prime meridian

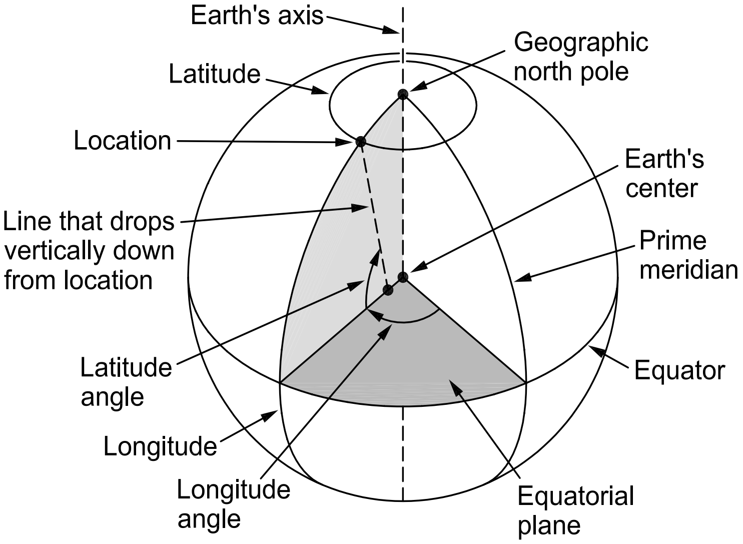

How latitudes and longitudes are derived

Latitude

The latitude of a location is defined as the angle between the equatorial plane and an imaginary line that drops vertically down from the location. This line generally does not intersect the equatorial plane exactly at Earth's center because Earth's surface is slightly flattened at the poles. The latitude is a measure of how far north or south a location is from the equator. Latitudes can be visualized as a series of concentric circles centered on the geographic poles and expanding to a maximum radius at the equator.

To specify the latitude, state the latitude angle and whether the position is north (N) or south (S) of the equator. The latitude is 90°N at the north pole, 0° at the equator, and 90°S at the south pole. Other locations have intermediate latitudes, for example, 46°25'09"S. The length along Earth's surface of one minute of latitude is proportional to the radius of curvature along the local surface; it varies from 1843 m at the equator to 1862 m at the poles. The original definition of the nautical mile was the average distance of one minute of arc of latitude along Earth's surface. Today, the nautical mile is defined as precisely 1852 m.

Longitude

The longitude of a location is defined as the angle, parallel to the equatorial plane, between the location and a reference longitude called the prime meridian. By convention, the prime meridian (0°) is the longitude that runs through an established point at Greenwich, England. From there, other longitudes increase eastward and westward until they meet at 180° on the opposite side of Earth. The longitude is thus a measure of how far east (E) or west (W) a location is from the prime meridian.

Longitudes can be visualized as a series of north-south lines that converge at the poles. To specify the longitude, state the longitude angle and whether the position is east or west of the prime meridian, for example, 105°57'15"E. The length along Earth's surface of one minute of longitude varies from 1855 m at the equator to 0 at the geographic poles.

1.4.1 Geographic coordinates on maps

Major latitudes and longitudes are shown as lines across the map or as tick marks in the margins. On most maps or charts, only selected latitudes and longitudes are completely labeled. To prevent clutter, other lines are labeled with partial numbers. For example, only the minutes of arc are shown. The standard way to describe a location with geographic coordinates is to state the latitude followed by the longitude. For example, 46°25'09"S, 105°57'15"E.

For a rough approximation of the latitude and longitude of a location on a map, look at the nearest latitude and longitude lines that surround the location. Visually estimate where the location is relative to the nearest latitude and longitude lines. Take into account that there are 60 minutes of arc in each degree and 60 seconds of arc in each minute of arc.

For more precise results, use a ruler to measure the distance between the location and the nearest lower latitude line, and then measure the distance between the nearest higher and lower latitude lines. Divide the first distance by the second distance to obtain the proportional distance. Convert the proportional distance into angle units and then add the result to the nearest lower latitude line. Calculate the longitude in a similar manner.

Example of how to determine geographic coordinates on a map

On the map below, the proportional distance between Location A and the nearest lower latitude line is 17 mm/38 mm = 0.45. The angular distance between the longitude lines on the map is 1' = 60". Converting from proportional distance to angular units gives 0.45 x 60" = 27". Adding this result to the nearest lower latitude line gives 53°01'00"N + 27" = 53°01'27"N.

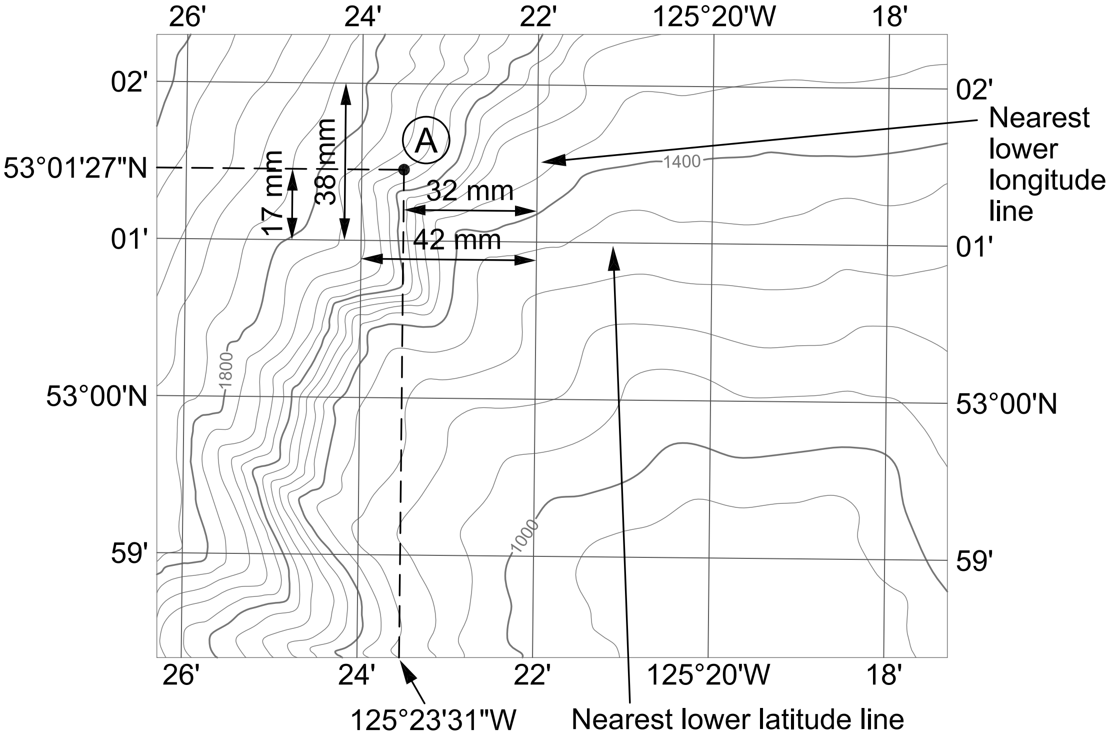

The proportional distance between Location A and the nearest lower longitude line is 32 mm/42 mm = 0.76. The angular distance between longitude lines is 2' = 2 x 60" = 120". Converting the proportional distance into angle units gives 0.76 x 120" = 91" = 1'31". Adding this result to the nearest lower longitude line gives 125°22'00"W + 1'31" = 125°23'31"W. The complete geographic position is written as 53°01'27"N, 125°23'31"W.

Similar types of calculations can be done to solve the reverse problem where you already have a written position and want to pinpoint it on the map. The above example clearly illustrates the trouble you have to go through to accurately determine a location with geographic coordinates. The task is even more difficult on maps where the latitude or longitude lines aren't straight.

1.5 Map projections

Projections are mathematical methods that are used to transfer geographic information from Earth's surface to maps. It is impossible to represent any portion of Earth's curved surface on a flat map without some kind of distortion of scale, direction, shape, or distance. The type of projection governs the type of distortion. Reducing one type of distortion will inevitably increase another type. Mapmakers are forced to compromise by deciding what sort of distortion is acceptable. This is largely determined by the intended use of the map.

Projections can be visualized as one or more light sources shining rays through Earth's surface onto a projection surface that may be partially located inside Earth. The rays produce an image of the topography and geographic coordinates on the projection surface. In areas where the projection surface is inside Earth, the image is formed by rays bouncing back from Earth's surface along the same path. The shape and location of the projection surface and the location(s) of the light source(s) define the type of projection. A map or chart is simply a portion of the projection surface.

The scale is always correct where the projection surface touches or cuts into Earth's surface. Features are shown too small where the projection surface is inside Earth and too

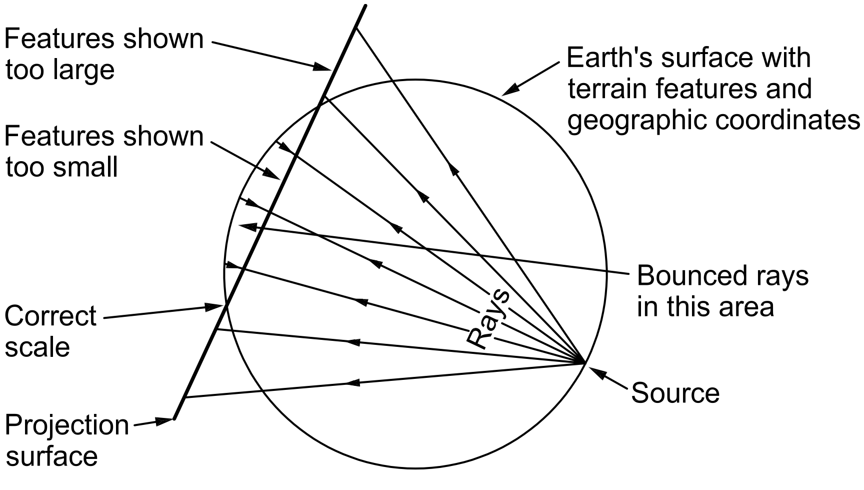
Features shown too large
Earth's surface with terrain features and geographic coordinates
Features shown too small
Bounced rays in this area
Rays
Correct scale
Source
Projection surface

large where the projection surface is outside Earth. Distances may be correct from one point to all other points, or from certain points to certain other points, but never from all points to all others points. In most cases, directions (bearings) do not follow straight lines. On large-scale maps, the distortion is usually minimal and can be safely ignored. Many types of projections exist, but only four widely used types are discussed here.

1.5.1 Mercator projection

In the Mercator projection, Earth's surface is projected onto a cylinder wrapped around the equator. A series of light sources are located along Earth's axis of rotation, with each source striking a particular portion of the projection surface.

The scale is correct only along the equator. Features are shown increasingly too large as you move away from the equator, becoming infinite at the poles. Shapes of small features are correct throughout the projection surface. Distances are correct only along the equator. Straight lines on the map have constant bearings but are generally not the shortest distance between two points.

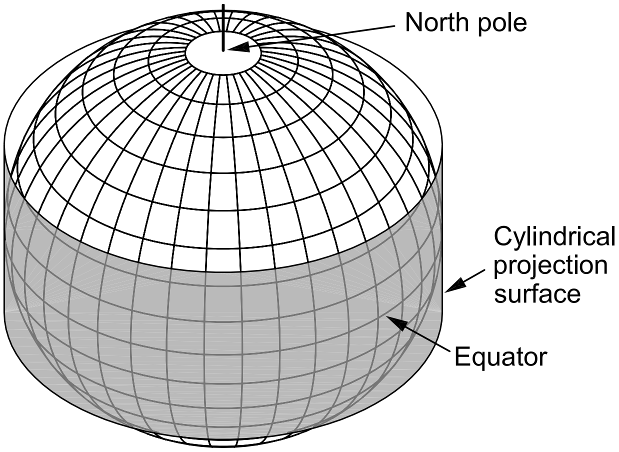

The Mercator projection is not used for polar regions because of the extreme distortion of scale and shape of large features at high latitudes. It is a popular projection for ocean charts because a route that follows a constant bearing is always a straight line on the chart.

1.5.2 Lambert conformal conic projection

With the Lambert conformal conic projection, Earth's surface is projected onto a cone oriented with its apex over one of the geographic poles. The cone cuts through Earth along two latitudes called the standard parallels. The source is located along Earth's axis.

The scale is correct along the two standard parallels. Features between the standard parallels are shown too small, while features outside the standard parallels are shown increasingly too large. Shapes of small features are correct throughout the projection surface. Distances are correct only along the standard parallels. Bearings are correct only along the longitudes.

The Lambert conformal conic projection is well suited for maps with large east-west coverage and is used for some polar maps.

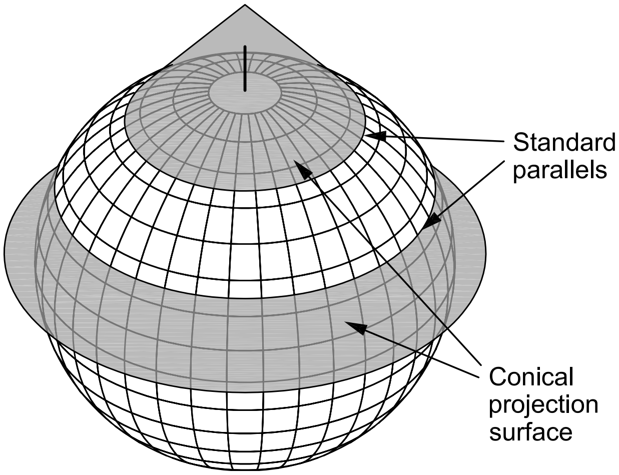

1.5.3 Universal transverse Mercator projection

The universal transverse Mercator (UTM) projection is complicated. The projection surface is a cylinder wrapped around Earth's poles along a longitude called the central meridian. The cylinder diameter is slightly smaller than Earth's diameter and cuts through Earth's surface along two lines parallel to the central meridian called secant lines. The light source is a series of points along Earth's axis of rotation. Only a narrow north-south strip, 3° of longitude on either side of the central meridian, is projected. For global coverage, the cylinder is twisted by 6° increments, creating an array of 60 different projection strips, each 6° of longitude wide.

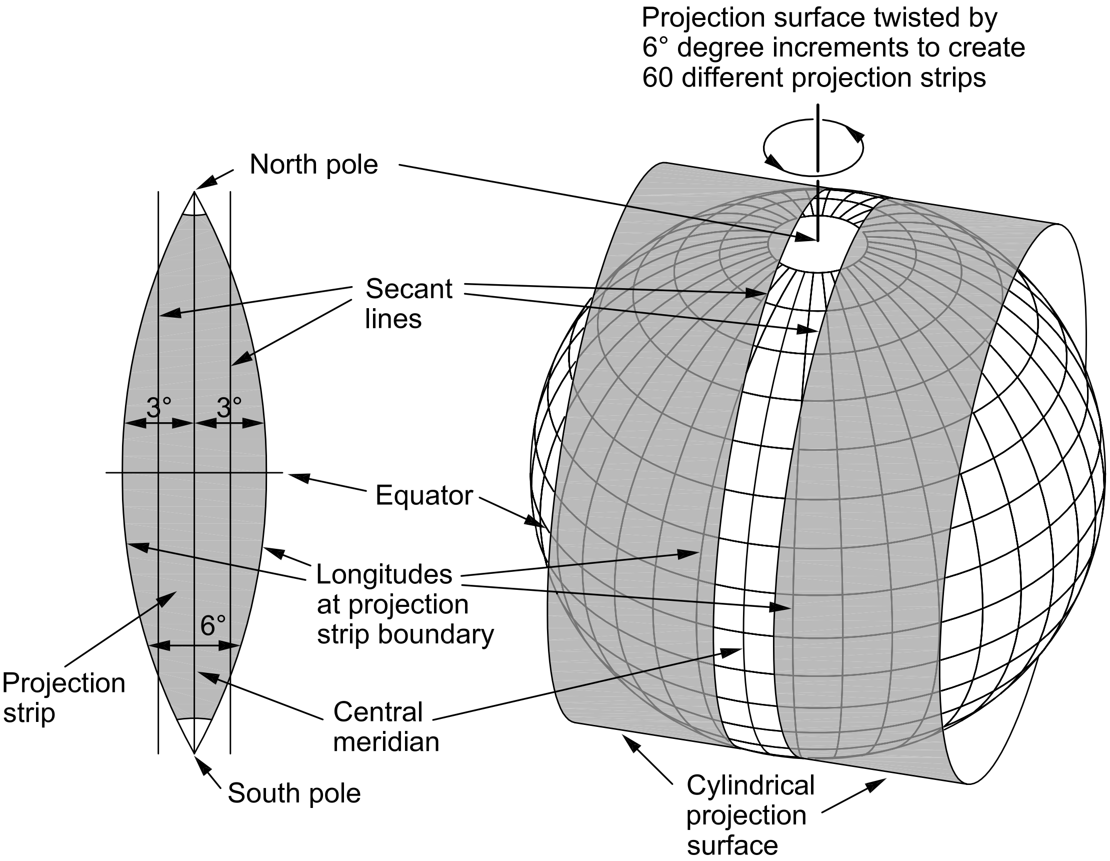
Projection surface twisted by
6° degree increments to create
60 different projection strips
North pole
Secant lines
3°
3°
Equator
Longitudes at projection strip boundary
6°
Projection strip
Central meridian
South pole
Cylindrical projection surface

The scale is correct only along the two secant lines. Features between the two secant lines are shown too small, while features outside the secant lines are shown increasingly too large. Shapes of small features are correct throughout the projection surface. Distances are correct only along the secant lines. Distortion of distance, direction, and shapes of large features increases as you move away from the central meridian. Straight lines on the map do not have constant bearings except along the central meridian.

The UTM projection is used in conjunction with the UTM grid and is well suited for maps with large north-south coverage. It is a favorite choice for large and medium scale topographic maps everywhere in the world except in polar regions.

1.5.4 Polar stereographic projection

The polar stereographic projection is one of the simplest types of projection. The projection surface is a flat plane that is centered on one of the geographic poles and cuts into Earth along a latitude called the standard parallel. The source is located at the opposite geographic pole.

The scale is correct along the standard parallel. Features are shown too small at latitudes inside the standard parallel and increasingly too large outside the standard parallel. Shapes of small features are correct throughout the projection surface but the shape distortion of large features increases as you move away from the pole. Distances are correct only along the standard parallel, and bearings are correct only along the longitudes.

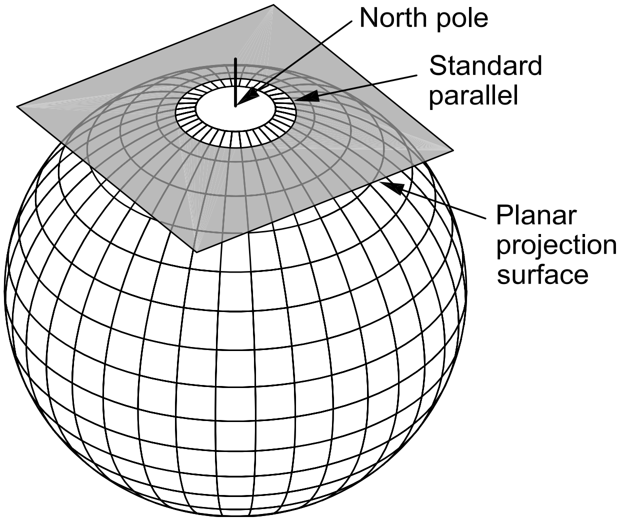

The polar stereographic projection is used mainly for topographic maps in polar regions in conjunction with the universal polar stereographic (UPS) grid.

1.6 UTM grid

A grid consists of an array of parallel and perpendicular lines drawn on the map. Each grid line is labeled with a number that is equal to the line's real-world distance from a specific line of origin. The distance units vary depending on the grid. Unlike latitude and longitude lines, grid lines are not distorted on the map because the grid lines are drawn on the projection surface after the projection has been completed. Grid lines form an array of identical squares, making them well-suited for finding positions on a map.

Many nations and organizations have developed grids that cover selected parts of Earth's surface. The universal transverse Mercator (UTM) grid is the most widely used grid in the world. It is tied to the universal transverse Mercator projection and covers Earth's entire surface between 80°S and 84°N. This book only deals with the UTM and the associated UPS grid. Other grids are based on similar principles and are usually simpler.

1.6.1 UTM zones

In the UTM grid, the nonpolar surface of Earth is divided into 60 vertical bands that coincide with the projection strips derived from the UTM projection. All the vertical bands are 6° of longitude wide. They are numbered from west to east, beginning with Band 1 (180°W - 174°W) and ending with Band 60 (174°E - 180°E).

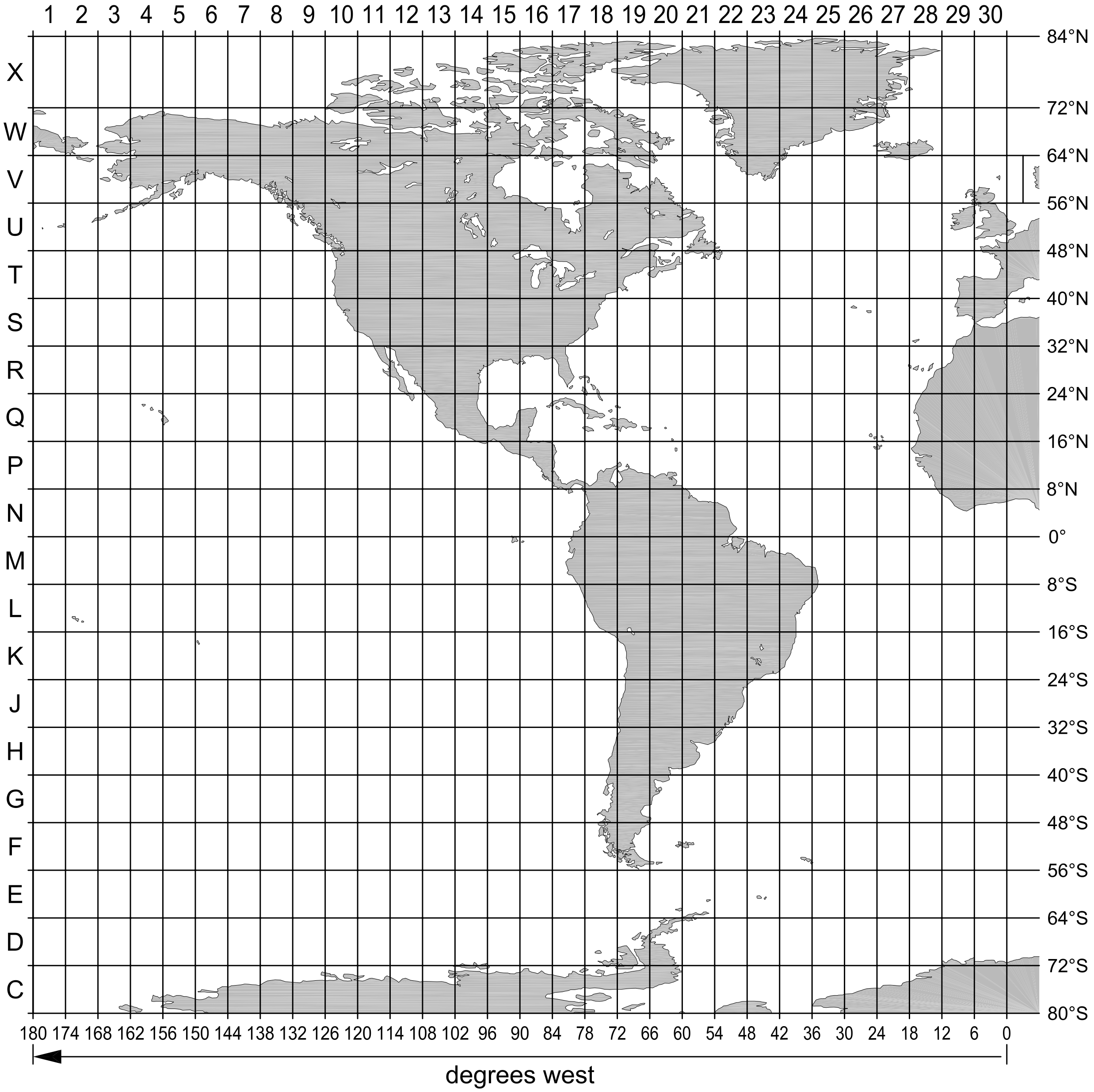
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
X
W
V
U
T
S
R
Q
P
N
M
L
K
J
H
G
F
E
D
C
84°N
72°N
64°N
56°N
48°N
40°N
32°N
24°N
16°N
8°N
0°
8°S
16°S
24°S
32°S
40°S
48°S
56°S
64°S
72°S
80°S
180
174
168
162
156
150
144
138
132
126
120
114
108
102
96
90
84
78
72
66
60
54
48
42
36
30
24
18
12
6
0
degrees west

UTM zones - western hemisphere

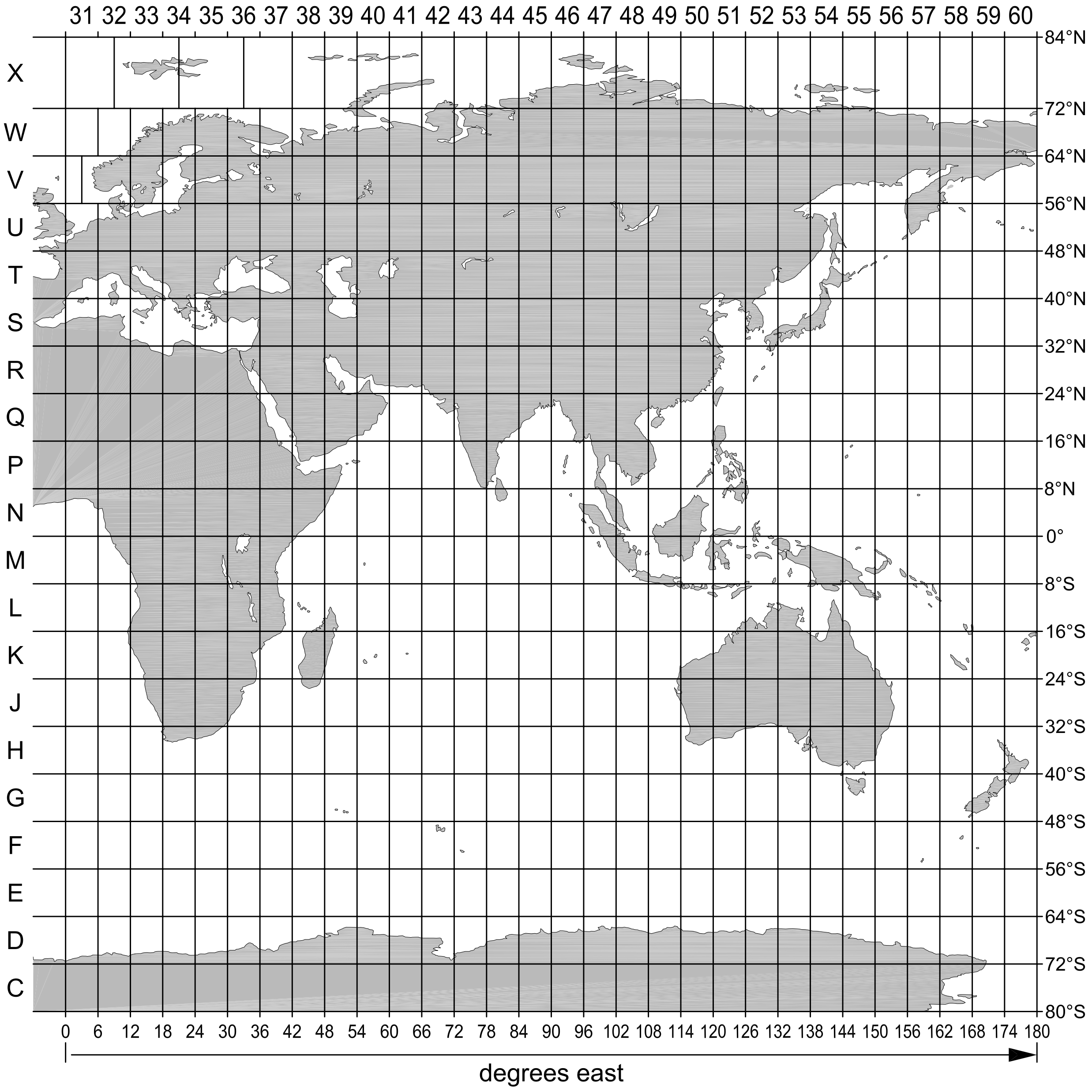
31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
X
W
V
U
T
S
R
Q
P
N
M
L
K
J
H
G
F
E
D
C
84°N
72°N
64°N
56°N
48°N
40°N
32°N
24°N
16°N
8°N
0°
8°S
16°S
24°S
32°S
40°S
48°S
56°S
64°S
72°S
80°S
0 6 12 18 24 30 36 42 48 54 60 66 72 78 84 90 96 102 108 114 120 126 132 138 144 150 156 162 168 174 180
degrees east

UTM zones - eastern hemisphere

The UTM grid also uses 20 horizontal bands that are generally 8° high. The horizontal bands are lettered from south to north beginning with Band C (80°S - 72°S) and ending with Band X (72°N - 84°N). The letters "I" and "O" have been omitted in order to eliminate confusion with the numbers "1" and "0". A UTM zone is the overlap between a vertical and a horizontal band. With the exception of the special UTM zones, each zone is 6° of longitude wide by 8° of latitude high (see maps of UTM zones on previous pages).

Examples of UTM zone boundaries

Zone 10U, located in the northern hemisphere, is bounded by latitudes 48°N and 56°N, and longitudes 120°W and 126°W. Zone 37L, located in the southern hemisphere, is bounded by latitudes 8°S and 16°S, and longitudes 36°E and 42°E.

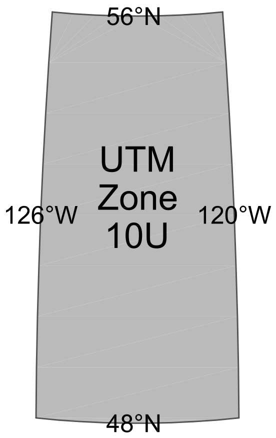

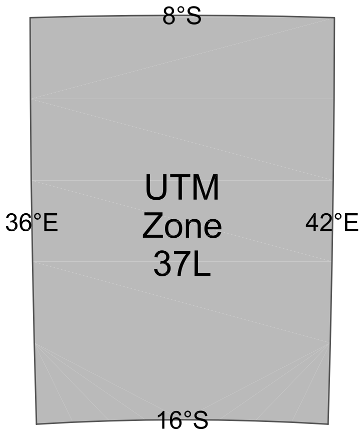

Both zones have the same "height" because adjacent latitudes are very nearly the same distance apart. Zone 10U is narrower than Zone 37L because it is located further from the equator where the longitudes are closer together.

Special UTM zones

The UTM grid is not completely consistent and is afflicted by a few exemptions called special UTM zones:

- Band X, the northernmost horizontal band, has been extended northward by 4°. Each zone in this band is 12° of latitude high instead of 8°.
- To accommodate southwest Norway, Zone 32V has been extended westward by 3° at the expense of Zone 31V.
- To accommodate Svalbard, Zone 33X and 35X have been extended both westward and eastward by 3°.
- Zone 31X has been extended eastward by 3° and Zone 37X westward by 3°.
- Zone 32X, 34X, and 36X have been eliminated.

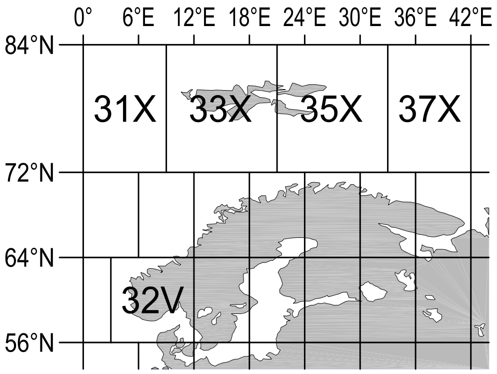

Special UTM zones

1.6.2 Easting and northing in UTM grid

Any nonpolar location on Earth can be identified by specifying its UTM zone, easting, and northing. The easting is a measure of how far east or west the location is from a reference longitude. The northing is a measure of how far north or south the location is from the equator.

The longitude running through the middle of each UTM zone is called the central meridian. It is defined as having an easting of 500 000 m. A point east of the central meridian has an easting of 500 000 m plus its distance in meters from the central meridian along the projection surface. Conversely, a point west of the central meridian has an easting of 500 000 m minus its distance in meters from the central meridian.

A point north of the equator has a northing equal to its distance in meters from the equator along the projection surface. A point south of the equator has a northing of 10 000 000 m minus its distance in meters from the equator.

Examples of how easting and northing are derived in UTM grid

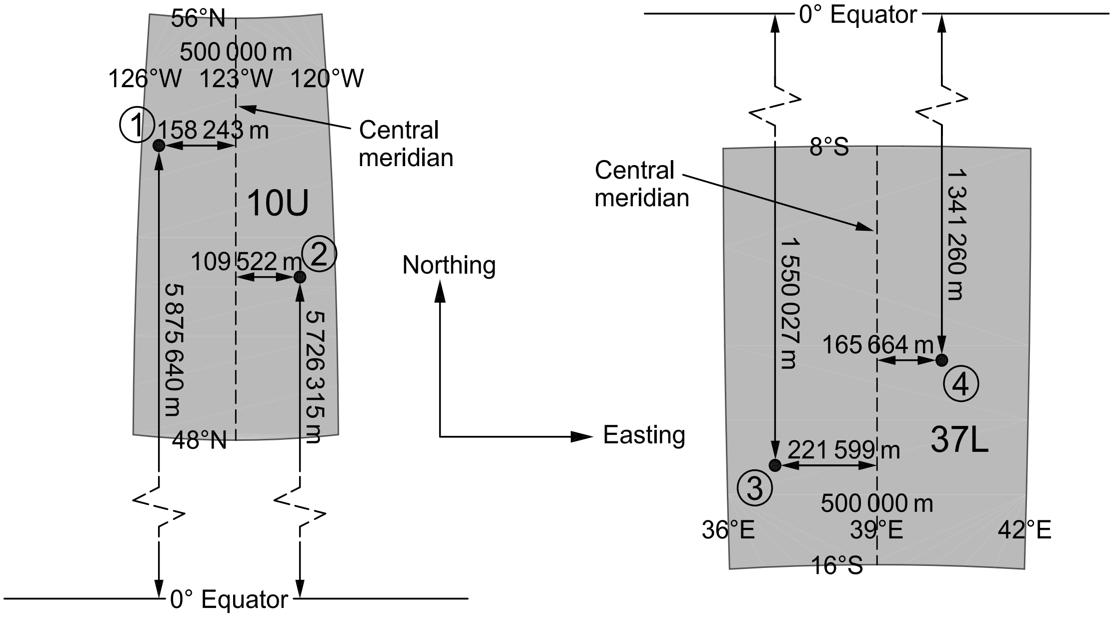
56°N
500 000 m
126°W 123°W 120°W
①
158 243 m
Central meridian
10U
109 522 m
②
5 875 640 m
5 726 315 m
48°N
0° Equator
Northing
Easting
0° Equator
8°S
Central meridian
1 550 027 m
1 341 260 m
165 664 m
④
221 599 m
37L
③
500 000 m
36°E
39°E
42°E
16°S

Location 1: Zone 10U
Easting = 500 000 m - 158 243 m = 341 757 m E
Northing = 5 875 640 m N

Location 2: Zone 10U
Easting = 500 000 m + 109 522 m = 609 522 m E
Northing = 5 726 315 m N

Location 3: Zone 37L
Easting = 500 000 m - 221 599 m = 278 401 m E
Northing = 10 000 000 m - 1 550 027 m = 8 449 973 m N

Location 4: Zone 37L
Easting = 500 000 m + 165 664 m = 665 664 m E
Northing = 10 000 000 m - 1 341 260 m = 8 658 740 m N

1.6.3 UTM grid on maps

Each UTM zone is generally large enough to contain many medium or large-scale maps. The zone(s) in which the map is located is normally written somewhere in the margins of the map. The horizontal band label (letters C through X) is sometimes omitted from the zone label because the northing uniquely establishes the horizontal band as long as you know what hemisphere you're in. Major eastings and northings are shown as lines across the map or as tick marks in the margins of the map. On most maps, only selected grid lines are completely labeled. To prevent clutter, other lines are labeled with only partial grid numbers. Although not necessarily aligned parallel to the map edges, UTM grid lines always form an array of perfect squares. This makes them inherently well-suited for pinpointing positions. A UTM grid position is usually stated as the zone followed by the easting and then the northing. For example: Zone 37L, 665 664 m E, 8 658 740 m N.

To find the grid numbers for a location on the map, first figure out the real world distance of the sides of one square. This can be done from the map scale or by looking at any fully labeled grid line. Measure or visually estimate the proportional distance between the location and the nearest lower easting line. Multiply the proportional distance by the real world length of one square. Add this distance to the nearest lower easting line to obtain the easting for the location. Use the same procedure to calculate the northing.

Example of how to estimate easting and northing on a map

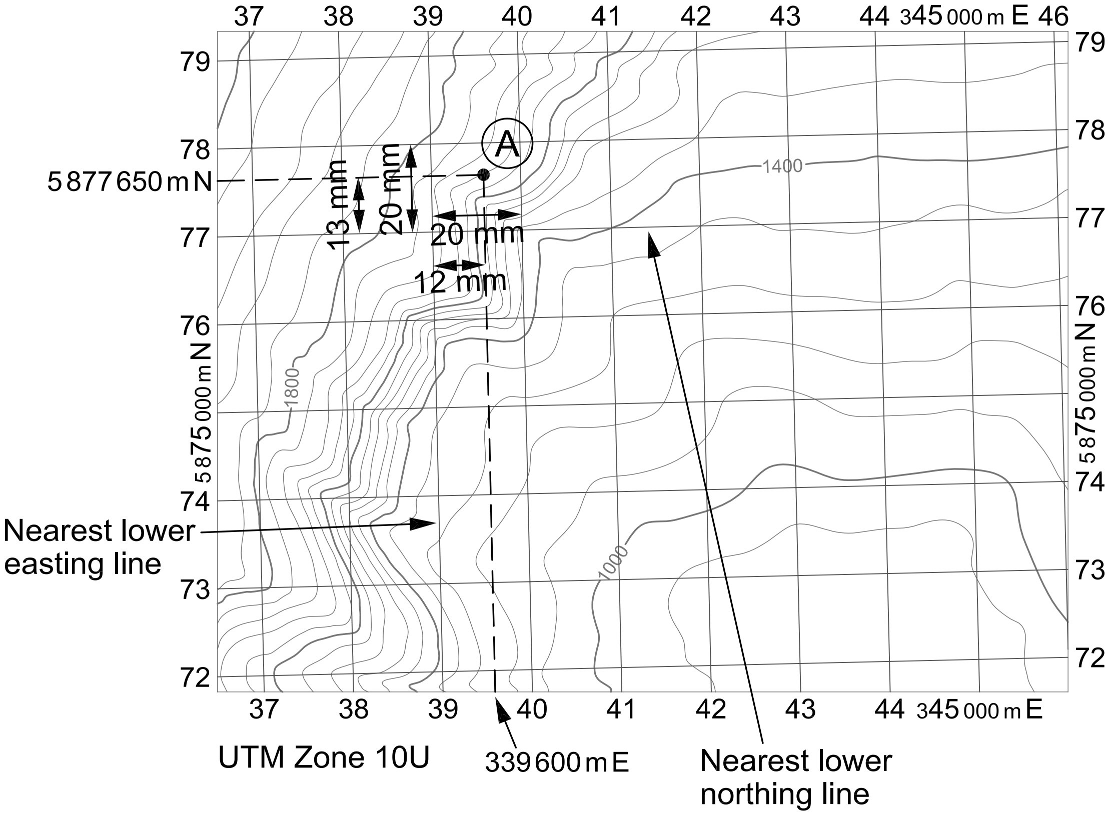
37
38
39
40
41
42
43
44 345 000 m E 46
79
78
77
76
74
73
72
5 877 650 m N
5 875 000 m N
A
13 mm
20 mm
20 mm
12 mm
1400
1800
1000
Nearest lower easting line
37
38
39
40
41
42
43
44 345 000 m E
UTM Zone 10U
339 600 m E
Nearest lower northing line

With a ruler, you measure the size of each grid square as 20 mm x 20 mm. By looking at the fully labeled lines 345 000 m E and 5 875 000 m N, you deduce that each square represents 1000 m x 1000 m in the real world. The distance between Location A and the nearest lower easting line (partially labeled as 39) is 12 mm on the map. The proportional distance between Location A and the nearest lower easting line is 12 mm/20 mm = 0.6. Multiplying this by the real world length of one square gives 0.6 x 1000 m = 600 m. Adding this distance to the nearest lower easting line gives an easting of 339 000 m + 600 m = 339 600 m.

The distance between Location A and the nearest lower northing line (partially labeled as 77) is 13 mm on the map. The proportional distance between Location A and the nearest lower easting line is 13 mm/20 mm = 0.65. Multiplying this by the real world length of one square gives 0.65 x 1000 m = 650 m. Adding this distance to the nearest lower northing line gives a northing of 5 877 000 m + 650 m = 5 877 650 m. The complete UTM grid position is stated as: Zone 10U, 339 600 m E, 5 877 650 m N.

Similar types of calculations can be performed to solve the reverse problem where you already have a written UTM position and want to pinpoint it on the map.

1.6.4 Truncated grid numbers

Many hiking and mountaineering guidebooks denote a UTM grid location as a 4, 6, or 8-digit number associated with a certain map. The first half of the digits in the number refers to a truncated easting and the rest of the digits are part of a truncated northing. For example, the first three digits of a 6-digit truncated grid number refer to the easting and the last three digits to the northing. The truncated grid number itself is usually all you need to pinpoint the location on the map. Other times, such as when entering the location into a global positioning system (GPS) receiver, you must expand the truncated grid number into a complete easting and northing. This requires looking at the map and doing some detective work.

Example of how to expand a truncated grid number

Suppose a guidebook tells you that a cabin is located at UTM Grid 781 883 on a particular 1:50 000 scale map. You want to enter this location into your GPS receiver. Your GPS receiver requires a complete easting, which is always a 6-digit number and a complete northing that can be as long as a 7-digit number.

The first three digits, 781, is the truncated easting. Since the map scale is 1:50 000, assume that the author has given you the location of the cabin to the nearest 100 m. Therefore, attach two zeros to the end of the truncated easting to make it 78 100. Now look at your map. If the left side of the map shows an easting of 465 000 while the right side shows 499 000, your easting must be somewhere in between and the first number of your easting must be 4. The complete easting is therefore 478 100 m E.

The last three numbers, 883, is the truncated northing. Add two zeros to make this number 88 300. If the map shows a bottom northing of 5 484 000 and a top northing of 5 510 000, your only logical choice is to add 54 to the front of your northing. The complete northing is therefore 5 488 300 m N.

1.7 UPS grid

The universal polar stereographic (UPS) grid is an extension of the UTM grid. It covers polar areas north of 84°N and south of 80°S. The UPS grid lines are drawn on a projection surface derived from the polar stereographic map projection. On a map, the UPS grid is used in the same manner as the UTM grid.

1.7.1 UPS zones

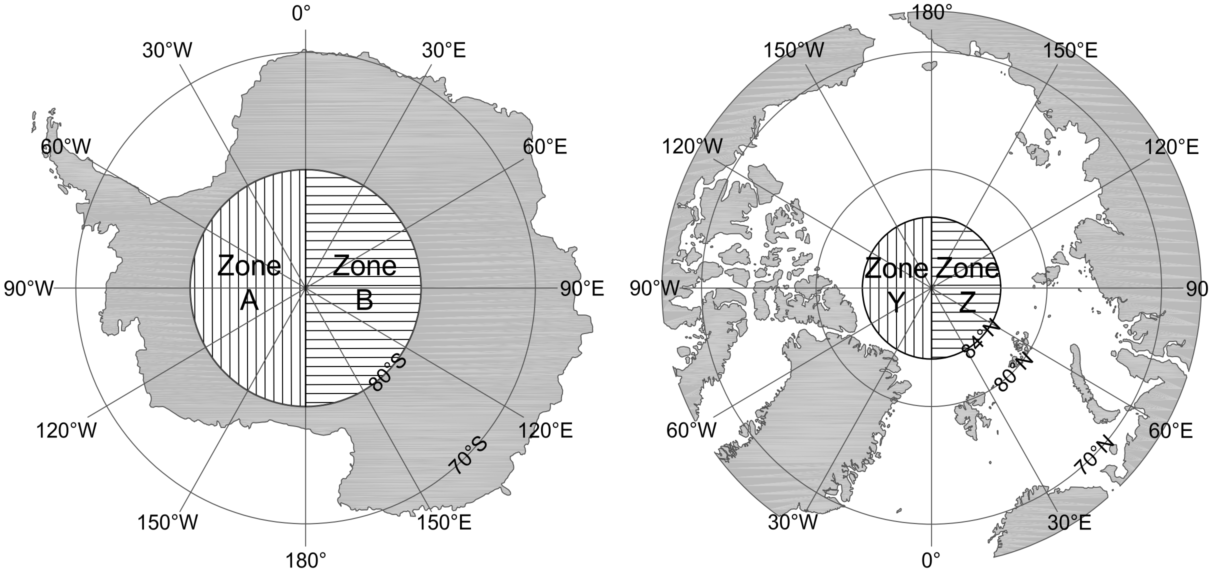

UPS grid zones

There are four half-moon shaped UPS zones, two located by the north pole and two by the south pole. Zone A covers the western hemisphere south of 80°S while Zone B covers the eastern hemisphere south of 80°S. Similarly, Zone Y covers the western hemisphere north of 84°N while Zone Z covers the eastern hemisphere north of 84°N.

1.7.2 Easting and northing in UPS grid

Easting and northing are confusing in the UPS grid, as these terms generally do not correlate with the cardinal directions east and north. This makes UPS positions intuitively difficult to visualize. The easting baseline, along longitudes 0° and 180°, is defined as having an easting of 2 000 000 m. A location in the eastern hemisphere has an easting of 2 000 000 m plus the distance in meters from the easting baseline. A location in the western hemisphere has an easting of 2 000 000 m minus the distance in meters from the easting baseline.

The direction of the northing is awkward to remember. The northing baseline, along longitudes 90°W and 90°E, is defined as having a northing of 2 000 000 m. In the southern hemisphere a location on either side of the 0° longitude has a northing of 2 000 000 m plus the distance in meters from the northing baseline. In the northern hemisphere, that formula applies to a location on either side of the 180° longitude. Conversely, in the southern hemisphere a location on either side of the 180° longitude has a northing of 2 000 000 m minus the distance in meters from the baseline, while in the northern hemisphere that formula applies to a location on either side of the 0° longitude.

Examples of how easting and northing are derived in UPS grid

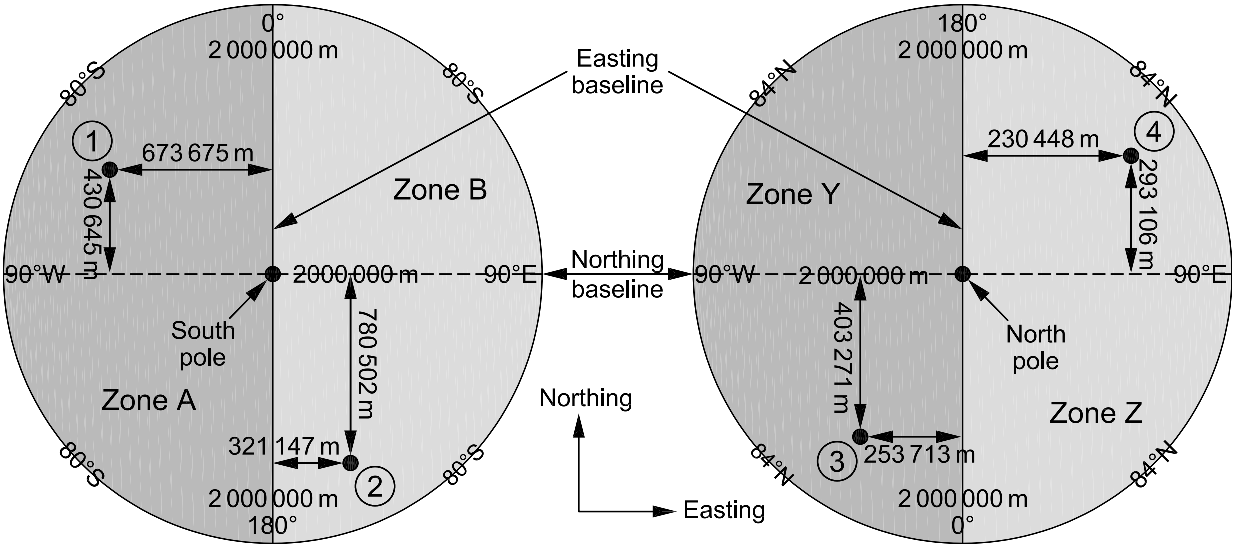
0°
2 000 000 m
80°S
80°S
Easting baseline
①
673 675 m
430 645 m
Zone B
90°W
2000 000 m
90°E
South pole
780 502 m
Zone A
321 147 m
②
2 000 000 m
180°
80°S
80°S
Northing baseline
Northing
Easting
180°
2 000 000 m
84°N
84°N
230 448 m
④
293 106 m
Zone Y
90°W
2 000 000 m
90°E
403 271 m
North pole
Zone Z
③
253 713 m
2 000 000 m
0°
84°N
84°N

Location 1: Zone A
Easting = 2 000 000 m - 673 675 m = 1 326 325 m E
Northing = 2 000 000 m + 430 645 m = 2 430 645 m N

Location 2: Zone B
Easting = 2 000 000 m + 321 147 m = 2 321 147 m E
Northing = 2 000 000 m - 780 502 m = 1 219 498 m N

Location 3: Zone Y
Easting = 2 000 000 m - 253 713 m = 1 746 287 m E
Northing = 2 000 000 m - 403 271 m = 1 596 729 m N

Location 4: Zone Z
Easting = 2 000 000 m + 230 448 m = 2 230 448 m E
Northing = 2 000 000 m + 293 106 m = 2 293 106 m N

1.8 Map datums

Map datums determine precisely where the latitude, longitude, and grid lines are drawn on a map. Datums also govern the heights of terrain features.

1.8.1 Earth's shape

If Earth were shrunk to the size of a billiard ball, its surface would be almost as smooth as a new billiard ball. The topography of mountain ranges and ocean floors would hardly be noticeable. This doesn't mean that Earth is a perfect sphere. The rotation around its axis generates a centrifugal force that causes the Earth to flatten slightly at the poles and bulge slightly at the equator. A simple geometric shape called an ellipsoid can approximate this shape. The numbers inserted into the ellipsoid formula determine the size of the ellipsoid and how much it bulges, and thus how well it agrees with Earth's real surface. In the past, a new improved global ellipsoid has replaced the old one every few decades or so.

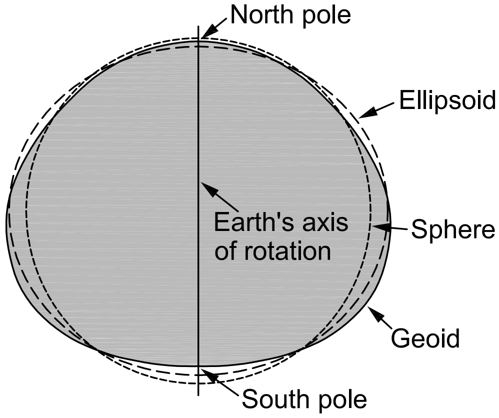

Models of Earth's shape
(shape highly exaggerated)

An ellipsoid does not take into account the distortions in shape caused by density variations inside Earth. A complicated model called a geoid has been developed to account for this. A geoid is the surface where gravity is equal. It is derived from measuring Earth's gravity at various locations. A geoid closely approximates the surface to which the oceans would conform if they were free to flow across the continents. It makes Earth look slightly pear-shaped with a slightly larger southern hemisphere and other smaller dips and bulges. New and improved geoid models are continually being developed.

1.8.2 Horizontal datums

Traditionally, ellipsoids have been created to conform only to certain areas of interest. These so-called regional ellipsoids are designed to make a best fit for a specific region while ignoring the rest of Earth's surface. Each regional ellipsoid is associated with a reference point located in the middle of the region. A regional ellipsoid combined with its reference point is called a regional datum. It forms a reference frame from where other positions in the region are computed. Several hundred regional datums are used worldwide.

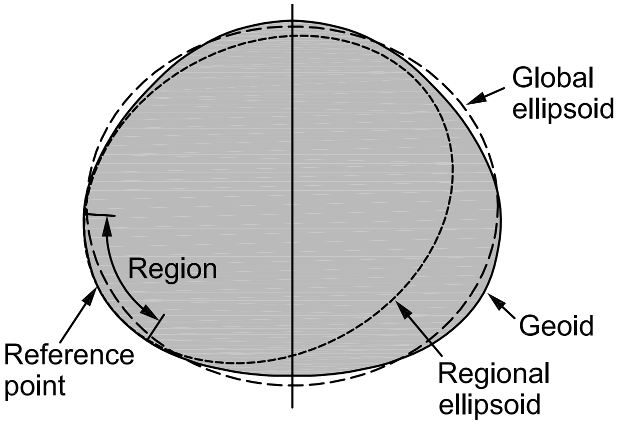

Regional and global ellipsoids
(shape highly exaggerated)

In recent years, the trend has been to use global ellipsoids that cover Earth's entire surface. Global ellipsoids are centered on Earth's center of mass and oriented along Earth's axis of rotation. In general, a global ellipsoid cannot conform to a local area as well as a regional ellipsoid, but has the advantage of worldwide coverage. A global datum consists of a global ellipsoid with positions referenced from the prime meridian and Earth's axis of rotation.

The horizontal datum governs where the latitude, longitude, and grid lines are drawn on the map. Different datums give different geographic coordinates or grid numbers for the same location. Latitude and longitude lines are projected entities while grid lines are drawn onto the map after it has been projected. This means that these two systems are affected differently by datums. When stating the geographic coordinates or grid numbers of a position on a map, always specify the corresponding datum. Using the wrong datum can lead to horizontal position errors of several hundred meters. The horizontal datum is usually written somewhere in the margins of the map.

Example of how datums affect a position

The map below shows geographic coordinate lines from two separate datums, the global 1984 World Geodetic System (WGS 84) and the regional 1927 North American Datum (NAD 27). The NAD 27 latitude lines are 0.6" (19 m) further south than the WGS 84 lines, while the longitude lines are 5.3" (99 m) further west. The geographic coordinate numbers for Location A are shifted in the opposite direction. The latitude is shifted north by 0.6" becoming 53°01'27" + 0.6" = 53°01'27.6"N in NAD 27. The longitude is shifted east by 5.3" becoming 125°23'31" - 5.3" = 125°23'25.7"W.

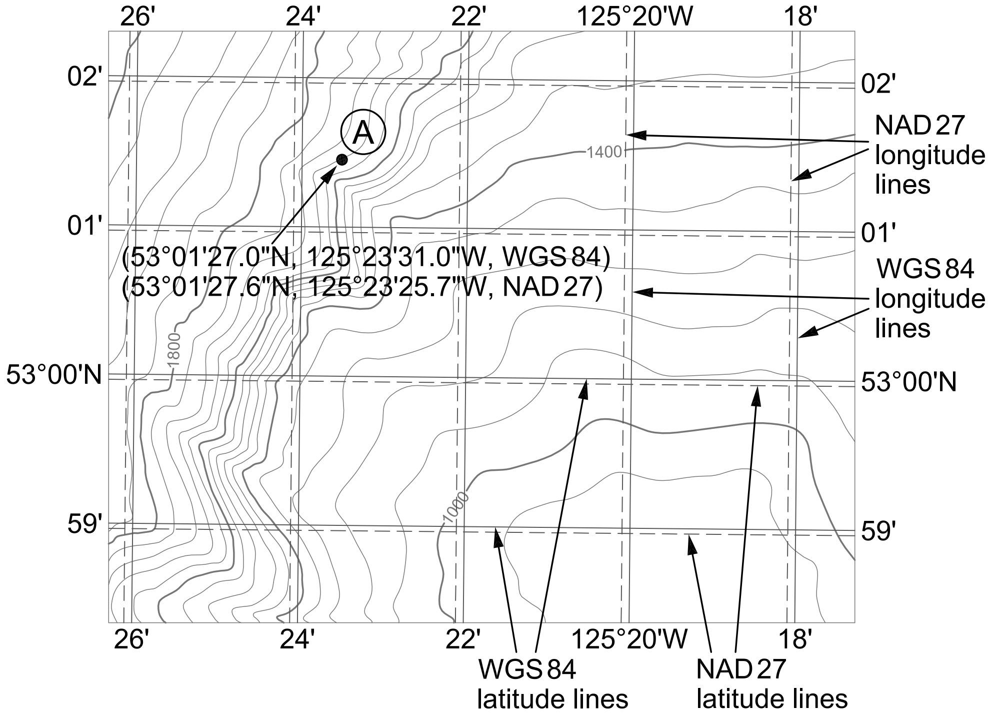
26'
24'
22'
125°20'W
18'
02'
01'
53°00'N
59'
A
1400
1800
1000
NAD 27 longitude lines
(53°01'27.0"N, 125°23'31.0"W, WGS 84)
(53°01'27.6"N, 125°23'25.7"W, NAD 27)
WGS 84 longitude lines
WGS 84 latitude lines
NAD 27 latitude lines

Next, is the same map shown with two sets of UTM grid lines. The NAD 27 easting lines are 95 m further west than the WGS 84 lines, while the northing lines are 203 m further north. The UTM numbers for Location A are shifted in the opposite direction. The easting is shifted east by 95 m becoming 339 600 + 95 = 339 695 m E in NAD 27. The northing is shifted south by 203 m becoming 5 877 650 - 203 = 5 877 447 m N.

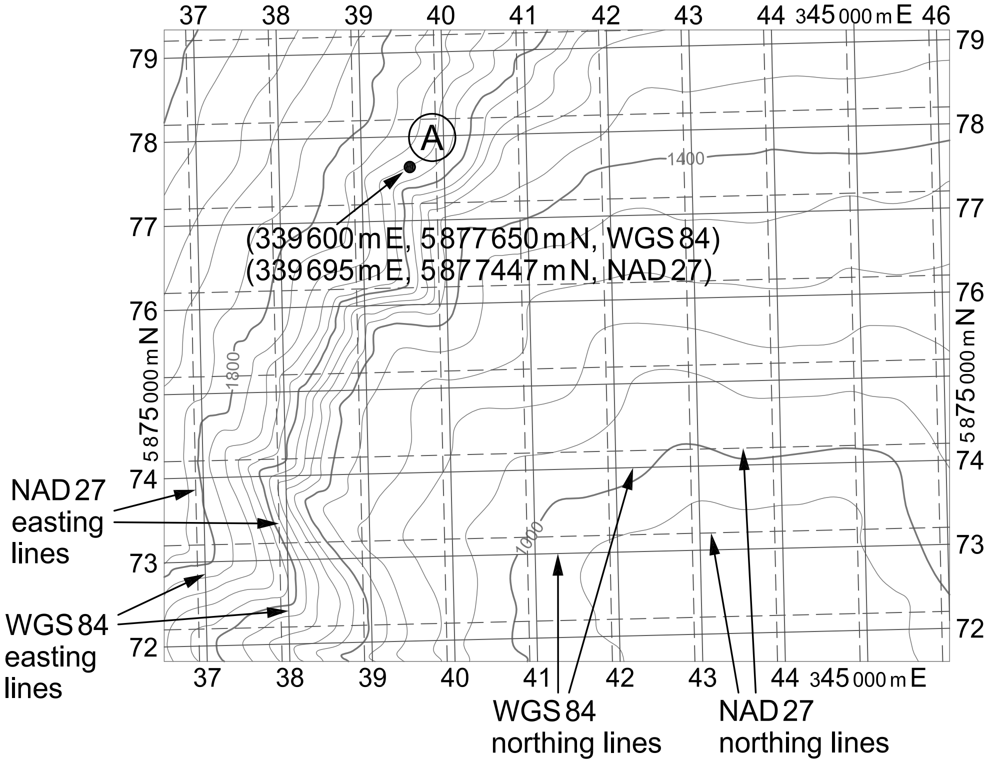
37 38 39 40 41 42 43 44 345 000 m E 46
79 78 77 76 5 875 000 m N 74 73 72
A
1400
1800
1000
(339 600 m E, 5 877 650 m N, WGS 84)
(339 695 m E, 5 877 447 m N, NAD 27)
NAD 27 easting lines
WGS 84 easting lines
WGS 84 northing lines
NAD 27 northing lines

1.8.3 Vertical datums

Map elevations are measured relative to a vertical datum, which is a reference surface that is either an ellipsoid or a geoid. The vertical datum is usually marked somewhere in the margin of the map. The term “mean sea level” that you see on many maps implies that the elevations are relative to a geoid surface. A geoid surface is by definition closely related to the surface that would result if the oceans were free to flow across continents, and essentially coincides with the average water level along a coastline. Away from the

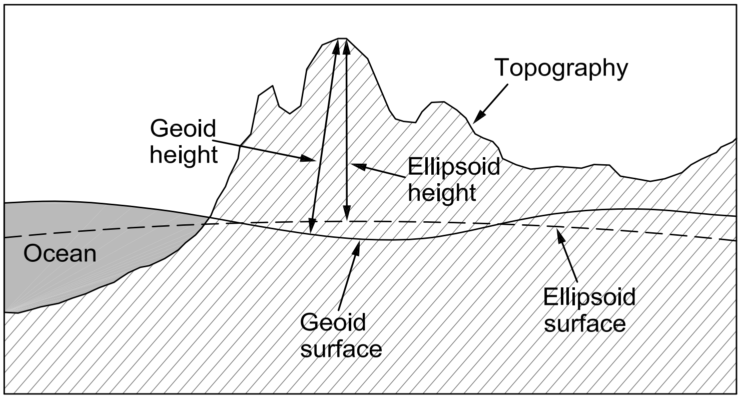
Topography
Geoid height
Ellipsoid height
Ocean
Geoid surface
Ellipsoid surface

coast, the geoid surface is calculated from measurements of the local gravity. The geoid surface is periodically recalculated leading to changes in surveyed map elevations.

An ellipsoid is sometimes used as a vertical datum. In this case, the surveyed heights on the map are measured relative to the ellipsoid surface instead of mean sea level. Maps can use the same ellipsoid for both the vertical and the horizontal datum but this is rarely the case. Because ellipsoid surfaces are easy to define mathematically, GPS receivers typically calculate elevations relative to an ellipsoid. The elevations are then converted into geoid heights before being displayed on the screen.

1.9 Topographic maps

Topographic maps have two key advantages over other types of maps: (1) they show the shape of the terrain, and (2) allow you to calculate the elevation gain or loss of a proposed route. This makes topographic maps inherently well suited for land navigation. They are also good for finding your way on rivers and small or moderately sized lakes.

1.9.1 Definition of contour lines

Contour lines enable three-dimensional features to be shown on a two-dimensional map. The contour lines are created by joining points of equal elevation. Each contour line therefore represents a specific elevation above sea level. The vertical distance between two adjacent contour lines is called the contour interval. In theory, a map with a small contour interval will always show more detail than a map with a large interval.

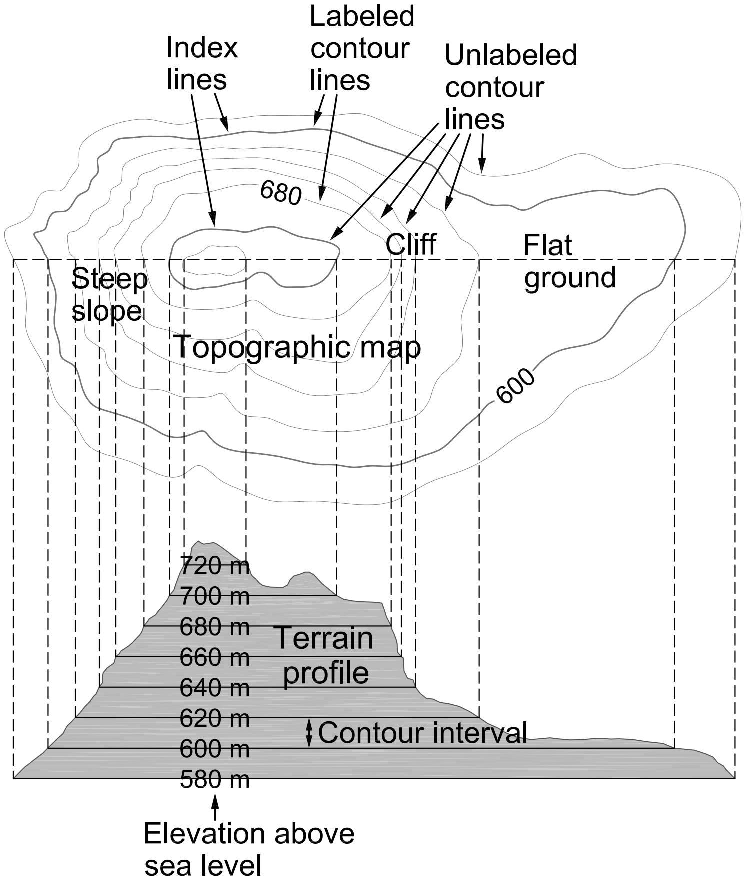
Labeled contour lines
Index lines
Unlabeled contour lines
680
Cliff
Flat ground
Steep slope
Topographic map
600
720 m
700 m
680 m
660 m
640 m
620 m
600 m
580 m
Terrain profile
Contour interval
Elevation above sea level

In practice, this is not always the case. If both the map scale and contour interval are small, the map will look like an indistinguishable blob of contour lines.

The horizontal distance between contour lines is inversely proportional to the slope of the terrain. Closely spaced lines indicate steep terrain such as cliffs while sparsely spaced lines reveal gentle slopes.

Usually, every 5th or 10th contour line is drawn in bold print. These bolded lines, called index lines, make it easier to track elevations on the map. Selected contour lines are labeled with their elevations while other lines are left unlabeled.

1.9.2 Finding elevation of unlabeled contour line

Finding the elevation of selected locations along a proposed route enables you to calculate the total elevation gain or loss of the route, which will give you a sense of how strenuous the travel will be. To estimate the elevation of a location, determine the elevation of the nearest contour line. If this contour line is unlabeled, count the number of contour lines to the nearest labeled line. Multiply this number by the contour interval to get the elevation difference. Then add or subtract (whichever is appropriate) the elevation difference to/from the elevation of the labeled line. This yields the elevation of the desired unlabeled contour line.

Elevation of unlabeled line - example

You want to find the elevation the unlabeled contour line on the map shown to the right.

Contour interval = 40 m

Elevation of nearest labeled line = 1800 m

Number lines below labeled line = 4

Elevation difference = 4 x 40 m = 160 m

Elevation of unlabeled line = 1800 - 160 m = 1640 m

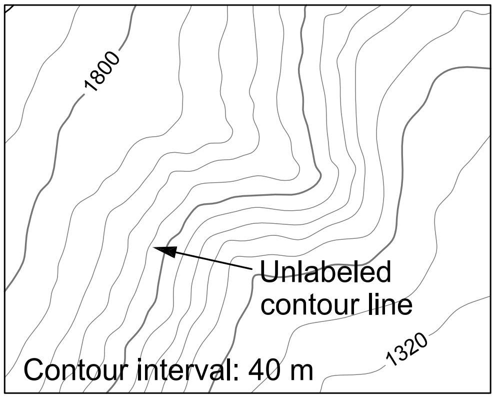

In most real life situations, the location is somewhere between two contour lines. The elevation is therefore somewhere in between the higher and lower contour line.

1.9.3 Calculating slopes

Anytime you're unsure whether a planned route is within your abilities, calculate the angle of the steepest part of the route. This may also help you decide what equipment to bring along. The procedure is straightforward: determine the horizontal distance (run) and the elevation difference (rise) of the slope, and then perform a couple of calculations.

The slope can be represented as a gradient measured in percent, or an angle measured in degrees. The gradient is the ratio between the rise and the run, and is calculated by dividing the rise by the run. The slope angle is calculated by plugging the gradient into a trigonometric function. See Table 1.9 for pre-calculated values.

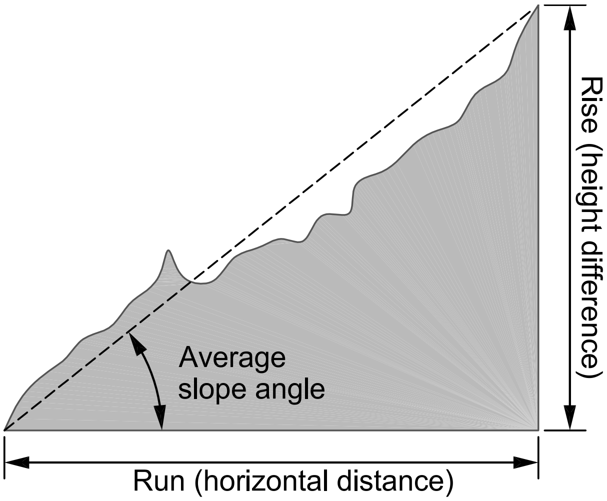

Slope calculation

1. On the map, use a ruler to measure the run (horizontal distance) of the desired portion of the slope. Measure perpendicular to the contour lines to ensure that the steepest direction of the slope (fall line) is calculated. Multiply the map distance by the scale to get the real world run.
2. Count the number of contour lines along the slope. Multiply the number of contour lines by the contour interval to obtain the rise (elevation difference) of the slope.
3. Divide the rise by the run to get the gradient in decimal units. Multiply the decimal gradient by 100 to obtain the gradient as a percentage.
4. Use Table 1.9 to convert the gradient into degrees. Interpolate for intermediate gradient values.

Table 1.9
Correlation between gradient and slope angle

Gradient	Angle	Gradient	Angle
9%	5°	119%	50°
18%	10°	143%	55°
27%	15°	173%	60°
36%	20°	214%	65°
47%	25°	275%	70°
58%	30°	373%	75°
70%	35°	567%	80°
84%	40°	1143%	85°
100%	45°	Infinite	90°

Example of slope calculation:

You want to calculate the average angle of a particular slope on the map shown below. The scale is 1:50 000 and the contour interval is 40 m.

With a ruler, measure the map run. The result is 14 mm. Multiply the map run by 50 000 to give you the real life run: 14 mm x 50 000 = 700 000 mm = 700 m. The slope covers 6 contour intervals so the rise is 6 x 40 m = 240 m.

The gradient is 240 m/700 m = 0.34 = 100% x 0.34 = 34%. By interpolation, Table 1.9 gives you an average slope angle of around 19°.

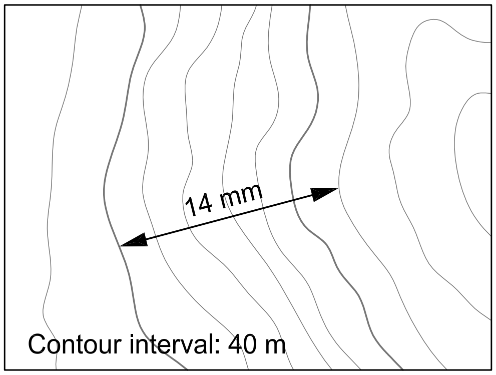

Keep in mind that your calculated slope angle is just an average. You will likely encounter steeper as well as flatter sections along that slope. You may be surprised how different the terrain appears in real life than envisioned from looking at the map.

Very few navigators actually calculate the angle of a slope from a topographic map, though it certainly is prudent to do so for off-trail routes through mountainous terrain. Most people just look at the contour lines on the map and judge the slope angle based on experience. Obviously, the only way to get this type of experience is to go outside with your map and check out some real slopes.

1.9.4 How contour lines represent terrain features

Contour lines can represent almost any type of terrain. Although every terrain feature on Earth is in some way different from all others, a particular pattern of contour lines indicates a specific type of feature. Below are some reoccurring types of features, shown with their associated contour line patterns.

Valley

A valley is a long and wide trough that has been carved out by moving glaciers and/or flowing water. If you are at the bottom of a valley, there is higher ground in three directions and lower ground in the other.

On maps, the valley bottom consists of a series of V-shaped or U-shaped contour lines whose closed ends point uphill along the valley floor. V-shaped contour lines indicate a steeper and narrower valley floor than U-shaped lines. Be careful not to confuse a valley with a ridge on the map. Valleys often show a river running along the closed end of the contour lines, which ridges never do.

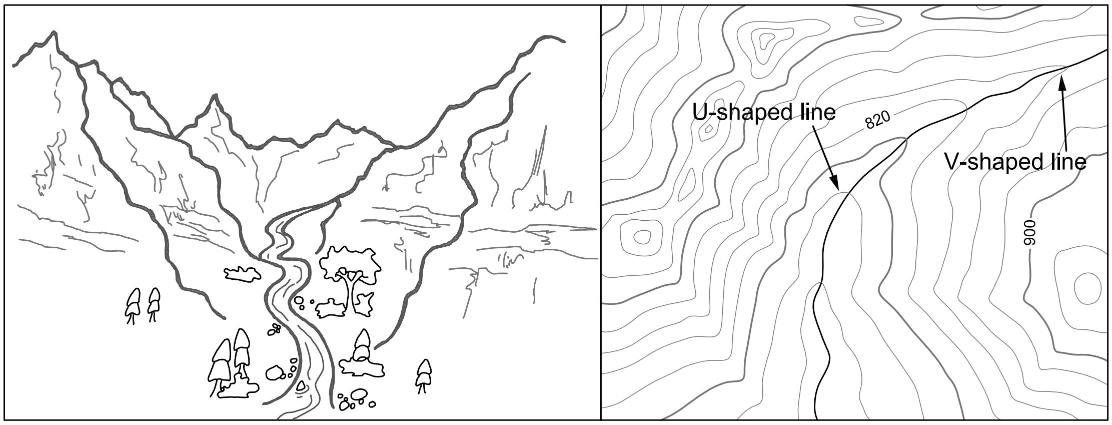

Ridge

A ridge is a linear feature similar to a rooftop. Most mountains have ridges radiating in several directions from the summit. The ridge crest normally undulates or slopes in one direction. On rare occasions, the ridge crest is level.

On maps, a sloped ridge crest consists of V-shaped or U-shaped contour lines with the closed ends pointing downhill. U-shaped lines signify a broad ridge, while V-shaped lines signify a sharp ridge known as an arête. A level ridge crest consists of two parallel contour lines that can be difficult to distinguish from any other parallel lines on the map. There are often bumps and saddles along a ridge, which is one way to differentiate a ridge from a valley or gully.

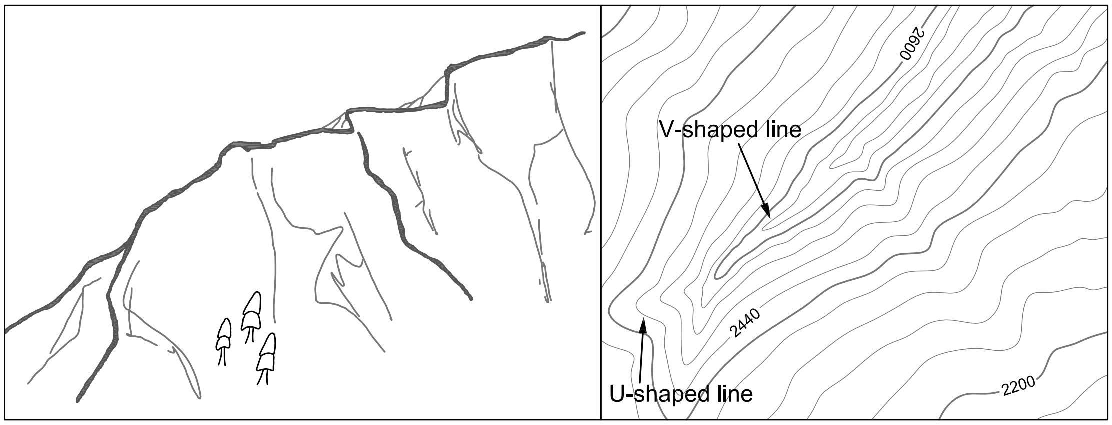

Gully

A gully is a groove running down a slope. Gullies have been gouged out of the slope by running water, avalanches, or rock fall. The contour lines are V-shaped or U-shaped with the closed ends pointing uphill. Inside the gully, there is higher ground in three directions and lower ground in the other. A gully is much smaller and narrower than a valley. A gully located high on a mountain is called a couloir.

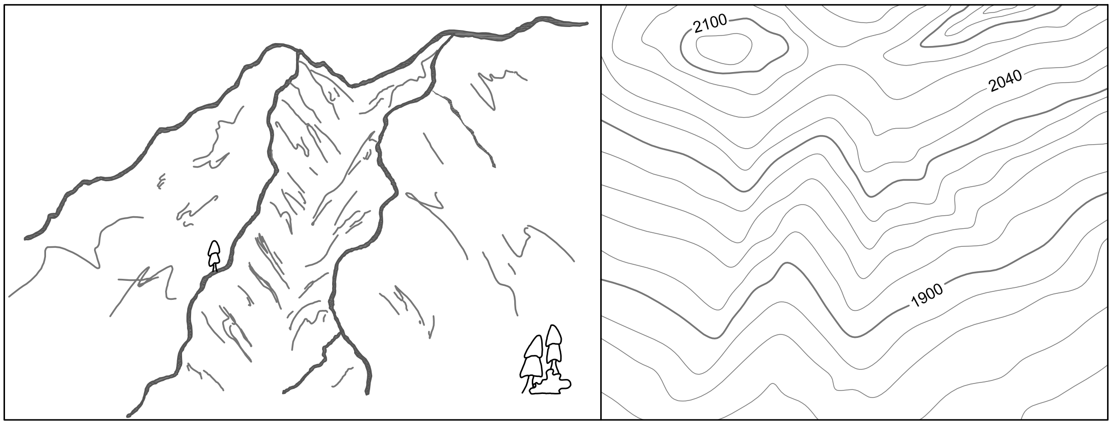

Cliff

A cliff is a vertical, near vertical, or overhanging feature. Cliffs are composed of rock that is resistant to erosion. Less resistant rock is often present below a cliff in the form of scree slopes or boulder fields. On maps, the contour lines representing a cliff are very close together and sometimes touch each other. In some cases, the contour lines disappear altogether, making it difficult to determine the slope or height of the cliff.

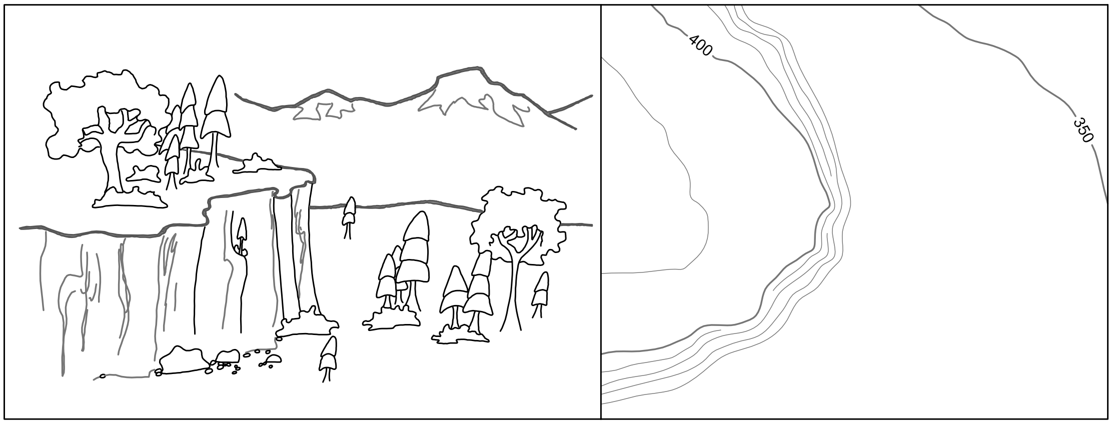

Pass

A pass is a low point between two areas of higher ground. It can be described as a saddle with higher ground in two opposite directions and lower ground in the other two directions. In many cases, passes are the easiest gateway between two valleys or drainages. On maps, the innermost contour lines of a pass look like an hourglass. A steep and narrow mountain pass is known as a col.

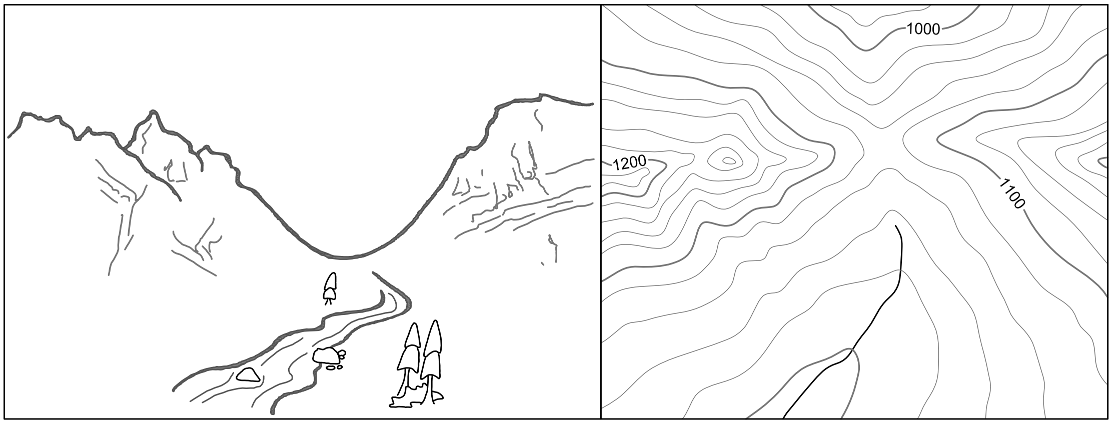

Summit

A summit is a high point surrounded by lower ground in all directions. All mountain massifs and hills contain at least one summit. On maps, the area surrounding the summit has closed contour lines nested inside each other with the high point located inside the innermost contour line. The elevation of the summit is sometimes written on the map.

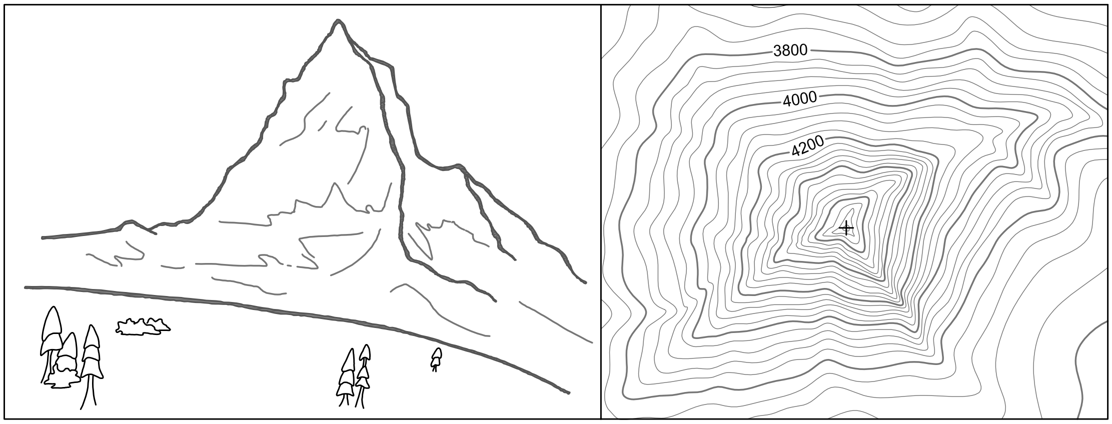

Canyon

A canyon is a trench that has been carved out by flowing water. In general, a canyon is narrower and has much steeper walls than a valley. On maps, a canyon consists of two sets of closely spaced lines, each set representing one of the walls. Smaller canyons sometime branch out from the main one. Depending on the season, flowing water may or may not be present on the canyon floor.

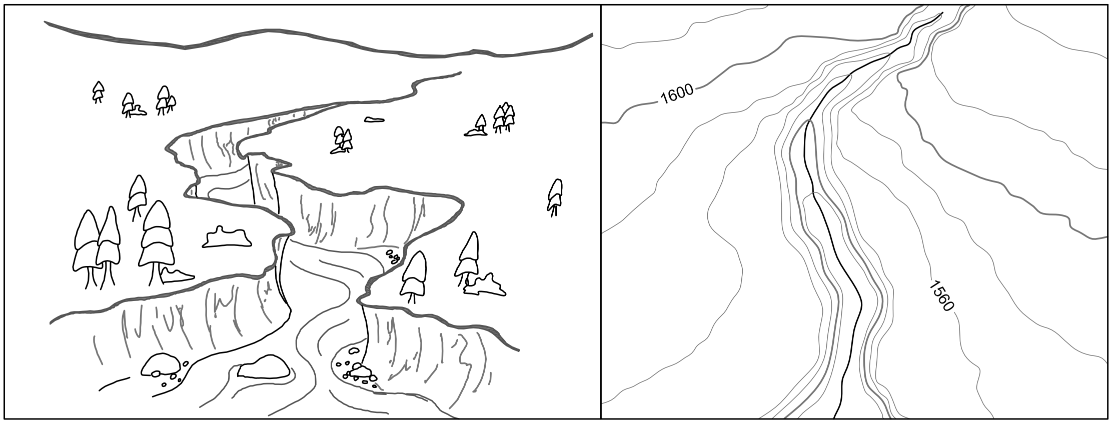

Depression

A depression is an area surrounded by higher ground in all directions. On maps, a depression is represented by closed contour lines with tick marks pointing toward lower ground. Depressions include volcanic and impact craters. Lakes are water-filled depressions, but underwater contour lines are normally not shown on topographic maps.

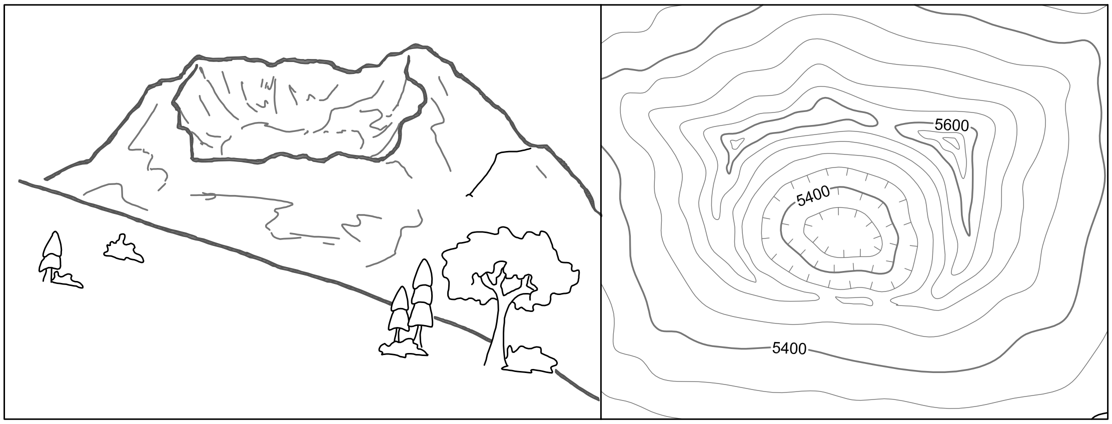

1.9.5 How to interpret terrain on a map

The ability to interpret the terrain described on topographic maps is a critical skill needed for successful land navigation. Individually, terrain features can be difficult to distinguish because it is often tricky to tell whether adjacent contour lines signify an upward or downward slope. Everything on a map is interconnected, so the key to interpreting the terrain lies in your ability to figure out how your area of interest relates to the larger picture.

Try the following technique to figure out which portions of the map represent higher versus lower ground. Major lakes, rivers, and valleys are located on lower ground. Analyze the contour lines that cross the longest streams on the map to figure out which way the water is flowing. Any naturally flowing body of water runs downhill along the bottom of a valley, canyon, or gully. The bottom of these features is represented by V-shaped or U-shaped contour lines with the closed end of the contour lines pointing upstream. Determine the direction of the flow of all streams that feed into the major streams. On their upstream end, all streams eventually begin at a lake, gully, glacier, spring, or near a pass. Tracing out all the streams on the map will give you a picture of how the area is drained and where most of the lower ground is located.

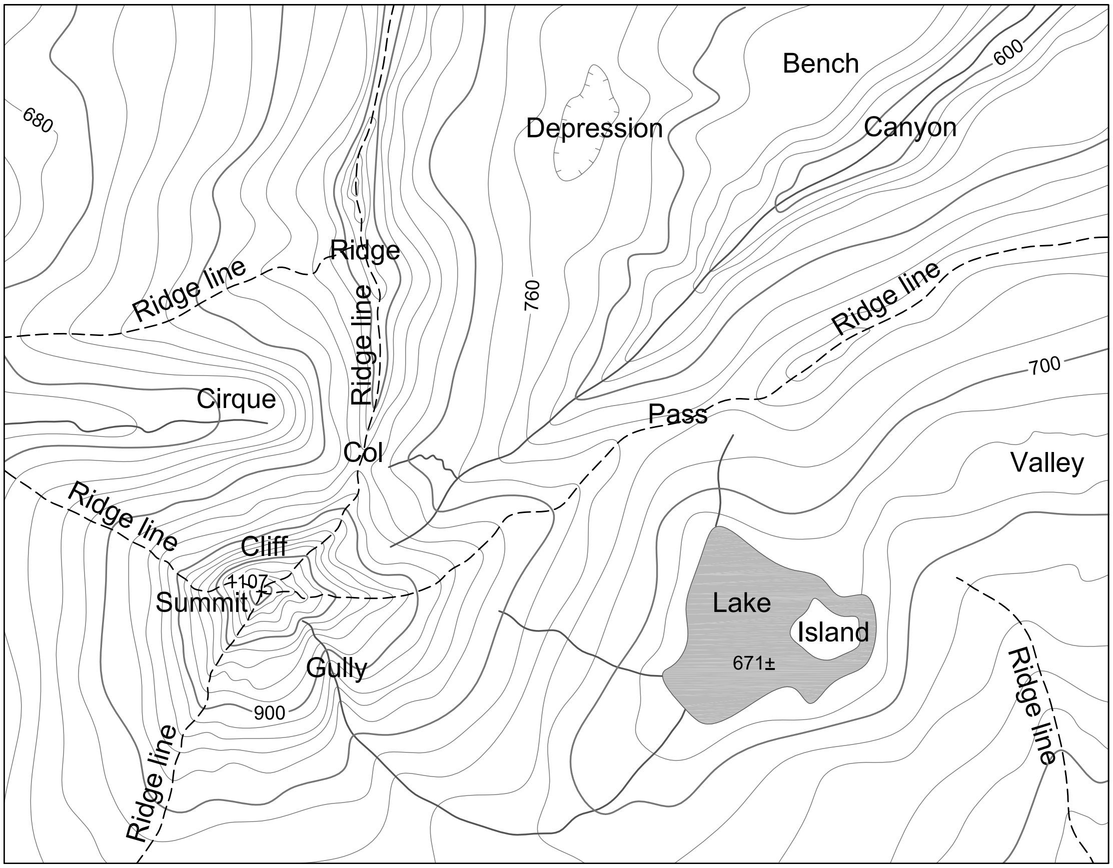
Bench
600
680
Depression
Canyon
Ridge
Ridge line
760
Ridge line
Ridge line
700
Cirque
Pass
Col
Valley
Ridge line
Cliff
1107
Summit
Lake
Island
671±
Gully
Ridge line
900
Ridge line

The higher ground is composed of ridges, summits, cols, and plateaus. On your map, connect the summits of mountains and hills with lines that follow the highest possible routes. These routes are mostly along ridgelines via passes, cols, and knolls. Then draw lines along ridges that radiate from the summits or fork off from the main ridges. Drawing these ridgelines will give you a sense of where the high ground is located.

Once you have identified the areas of high and low ground, it will be much easier for you to make out the terrain features in your area of interest.

1.10 Nautical charts

Nautical charts are essential for safe navigation on large bodies of water, especially along coastlines. Unlike maps, charts are specifically designed to describe water areas. The information included on a chart is even more wide-ranging than what you see on a map. Prevailing currents, shipping lanes, harbors, anchorages, reefs, sand bars, tidal flats, navigation aids, and much more are shown on a chart. It is important to thoroughly understand all the symbols. Consult the legend before leaving shore.

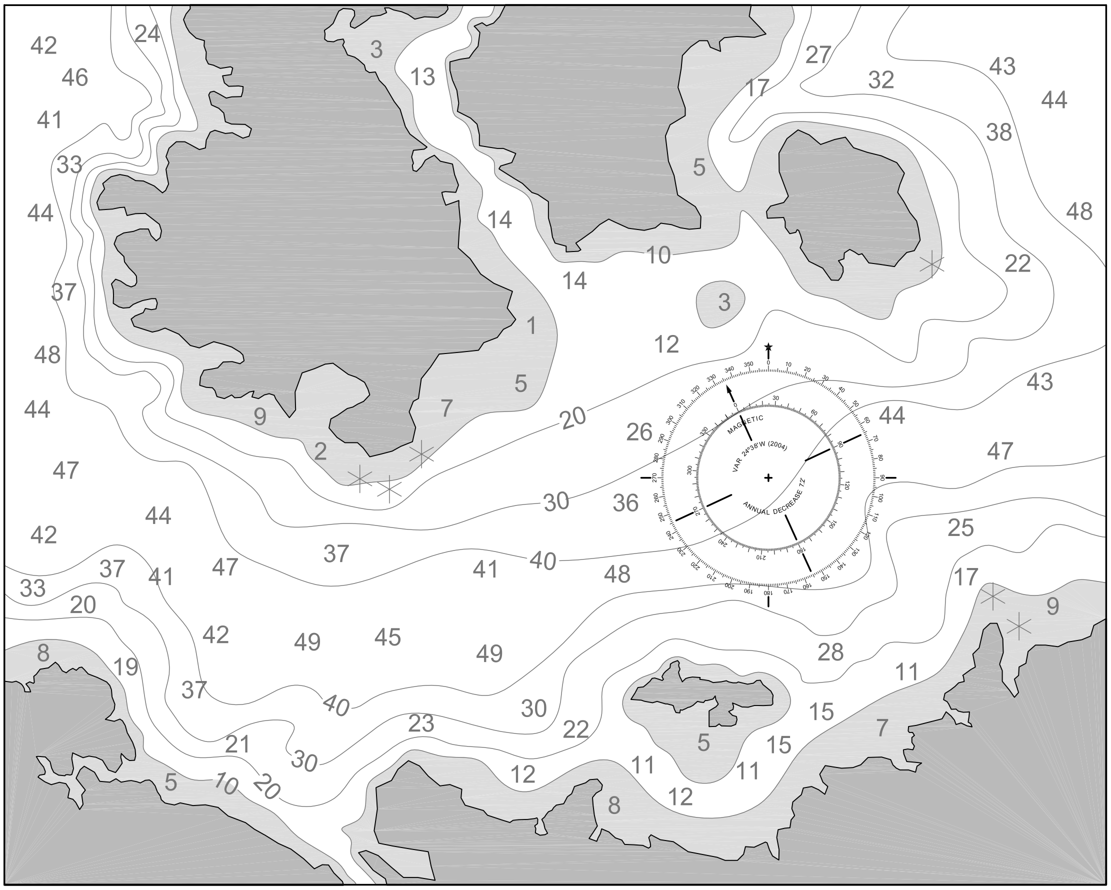
42
24
3
13
27
43
46
17
32
44
41
38
33
5
44
14
48
10
22
37
14
3
1
12
48
5
43
44
20
7
9
26
44
2
47
47
MAGNETIC
VAR 24°38'W (2004)
ANNUAL DECREASE 7.2'
30
36
42
44
25
37
40
41
48
33
37
41
47
17
20
9
42
49
45
49
28
8
19
11
37
40
30
22
15
23
21
7
5
15
30
12
11
11
5
10
20
12
8

Contour lines and/or depth soundings portray the shape of the sea bottom. Instead of showing the elevation above sea level, the contour lines and soundings show the depth below sea level. Charts use the mean low water level, the average level of all the low tides at that location, as a vertical datum. A chart is not meant to just be viewed; it is used for plotting courses and lines of position. Like maps, charts are drawn to different scales and are made from various types of projections.

Chapter 2
Compass

Compasses have been around for many centuries, initially for ship navigation but later also for land navigation. A compass needle that is free to rotate will keep pointing in a specific direction. This enables a compass user to stay on course toward any desired direction, a powerful concept in navigation. Even today, the compass remains one of the most widely used navigation instruments.

2.1 Earth's magnetic field

2.1.1 Magnetic field theory

Deep inside Earth, the movement of molten metal generates a magnetic field that envelops Earth. This invisible force field permeates our surroundings. Its presence is felt inside Earth, on the surface, in the oceans, in the atmosphere, and in nearby space. The field varies in strength and direction depending on the location. It can be visualized as an array of lines with arrows pointing in the direction of the magnetic force. The strength of the magnetic force is proportional to how close the field lines are to each other.

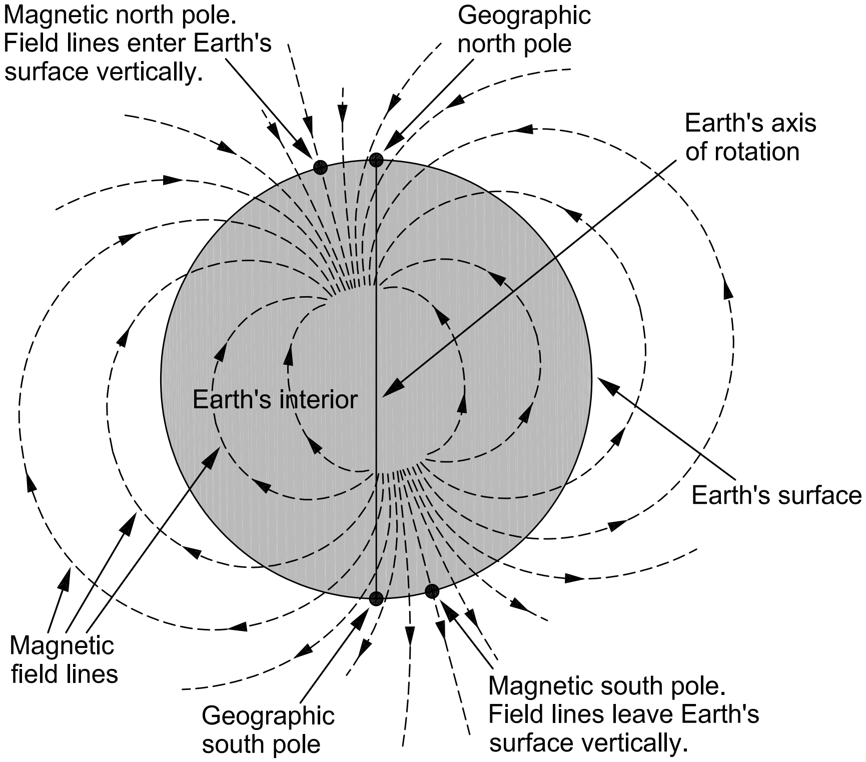
Magnetic north pole.
Field lines enter Earth's surface vertically.
Geographic north pole
Earth's axis of rotation
Earth's interior
Earth's surface
Magnetic field lines
Geographic south pole
Magnetic south pole.
Field lines leave Earth's surface vertically.

Idealized (bar magnet) model of Earth's magnetic field

Earth's magnetic field is shaped somewhat like the field of a bar magnet where the field lines leave the magnet at one end, curve around, and enter the magnet at the other end. The location where the magnetic field lines plunge vertically into Earth's surface is called the magnetic north pole. The location where the magnetic field lines leave Earth's surface vertically is called the magnetic south pole.

The real shape of Earth's magnetic field is a superposition of many magnetic fields corresponding to many molten metal flows inside Earth, as well as local magnetic anomalies caused by ore deposits. The resultant field is much more complicated than that of a single bar magnet. In fact, the magnetic poles are not on exact opposite sides of Earth, and the magnetic field lines generally don't follow the shortest path along Earth's surface to the nearest magnetic pole. The map below shows the approximate shape and direction of the magnetic field lines in polar regions.

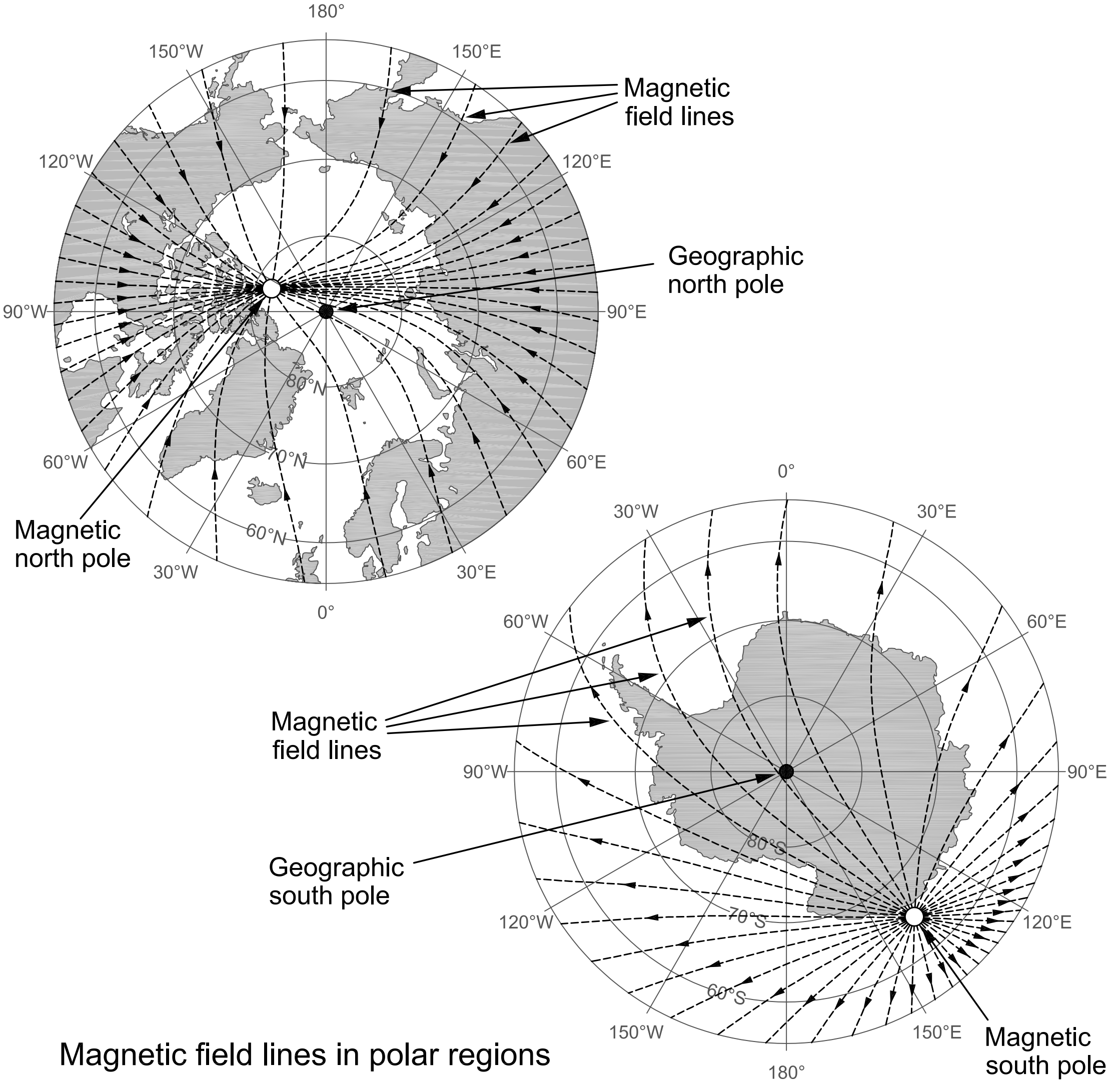
180°
150°W
150°E
Magnetic field lines
120°W
120°E
Geographic north pole
90°W
90°E
80°N
70°N
60°W
60°E
Magnetic north pole
60°N
30°W
30°E
0°
0°
30°W
30°E
60°W
60°E
Magnetic field lines
90°W
90°E
80°S
Geographic south pole
120°W
70°S
120°E
60°S
150°W
150°E
Magnetic south pole
180°

Magnetic field lines in polar regions

2.1.2 Magnetic pole movement

The shape of Earth's magnetic field is continually changing, causing the magnetic poles to move. Three types of magnetic pole movement have been observed: (1) secular movement where the magnetic poles move slowly over the years, (2) diurnal, or daily, movement around the average position, and (3) erratic movement where the magnetic poles jump rapidly from one position to another.

Secular movement

The molten metal flows inside Earth are continually changing, altering the shape of the magnetic field and shifting the average position of the magnetic poles in the process. The rate and direction of the magnetic pole movement is somewhat predictable. It typically ranges from a few kilometers to a few tens of kilometers per year. The secular movement must be taken into account when using a compass. The maps below show the secular movement over a 100-year period.

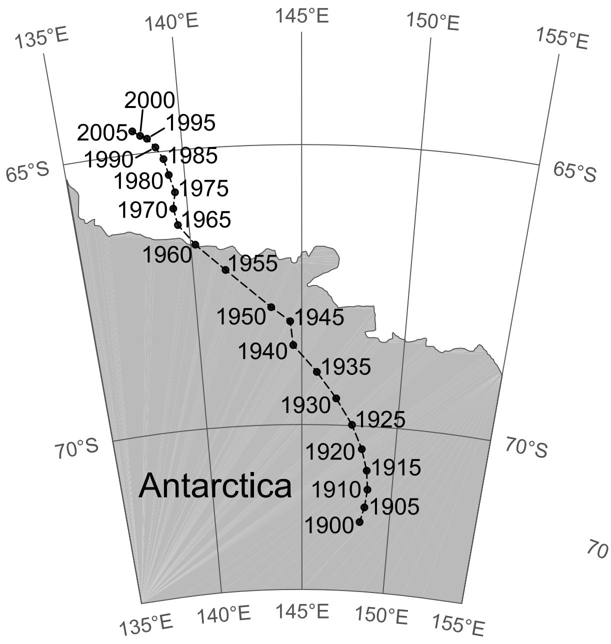
135°E
140°E
145°E
150°E
155°E
65°S
65°S
70°S
70°S
2000
2005
1995
1990
1985
1980
1975
1970
1965
1960
1955
1950
1945
1940
1935
1930
1925
1920
1915
1910
1905
1900
Antarctica
135°E
140°E
145°E
150°E
155°E

Secular movement of magnetic south pole

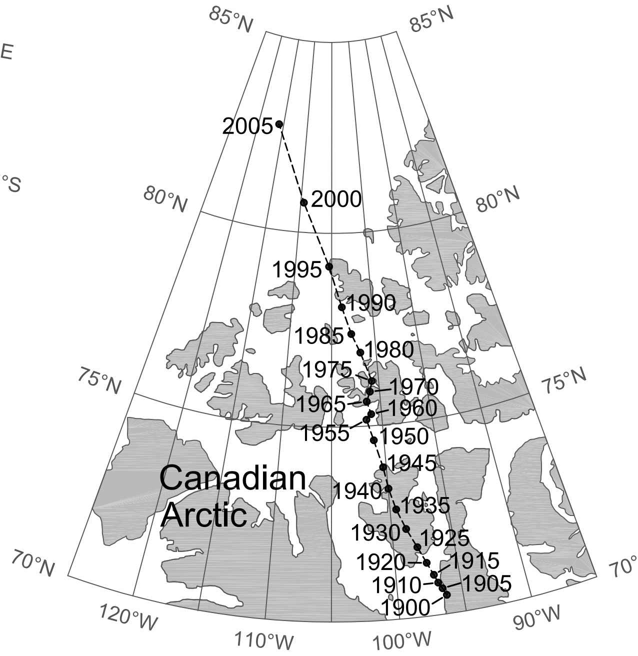
85°N
85°N
2005
2000
80°N
80°N
1995
1990
1985
1980
1975
1970
1965
1960
1955
1950
75°N
75°N
1945
1940
1935
1930
1925
1920
1915
1910
1905
1900
Canadian
Arctic
70°N
70°N
120°W
110°W
100°W
90°W

Secular movement of magnetic north pole

Diurnal movement

The sun continually emits electrically charged particles in all directions, creating something called the solar wind. When the solar wind reaches Earth, it pushes or pulls the magnetic poles to one side. As Earth rotates around its axis, the magnetic poles move around in an ellipse about their average positions. This is called the diurnal movement of the magnetic poles.

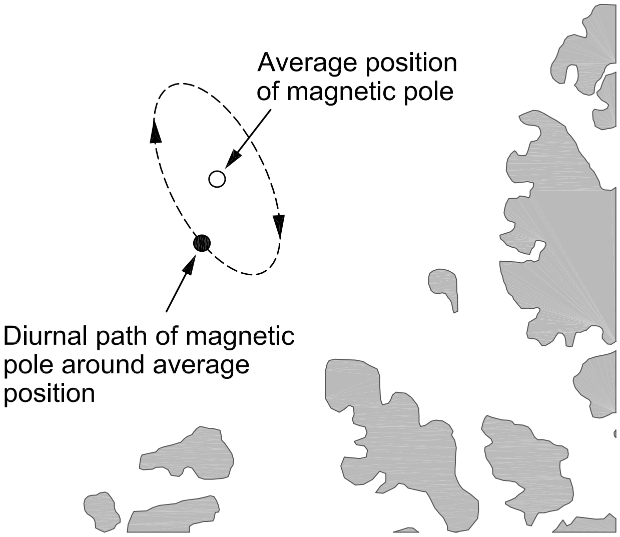
Average position of magnetic pole
Diurnal path of magnetic pole around average position

Diurnal movement of magnetic pole

Depending on the strength of the solar wind, the magnetic poles can be shifted by up to a few tens of kilometers about their average positions. The diurnal movement has negligible effect on the direction of a compass needle unless you happen to be close to one of the magnetic poles.

Erratic movement

The solar wind varies in intensity based on a cycle with an average period of 11 years that is closely tied to sunspot activity on the sun. The sun's magnetic activity reaches a maximum during each cycle and then usually quiets down for prolonged periods, although heightened activity can occur anytime. During periods of high magnetic activity, gusts of charged particles are hurled into space, distorting the magnetic field when reaching Earth. The magnetic poles can move about erratically, causing compass needles to swing wildly in the vicinity of the magnetic poles.

2.2 How a compass works

A compass needle is simply a strip of magnetized metal mounted on a swivel that lets the needle rotate freely. Any external magnetic field will generate a magnetic torque that attempts to align the needle along the external magnetic field lines. In the absence of other external fields, a freely rotating compass needle will align itself with Earth's local magnetic field line. The north-seeking end of the needle will point in the direction of the magnetic field line, but usually not directly toward the north magnetic pole.

Provided the compass needle is free to rotate, the magnetic torque will keep the needle pointed in the same direction regardless of how the base of the compass is oriented. By means of a mechanical linkage, this phenomenon can be used to keep the base of a compass pointed in any desired direction.

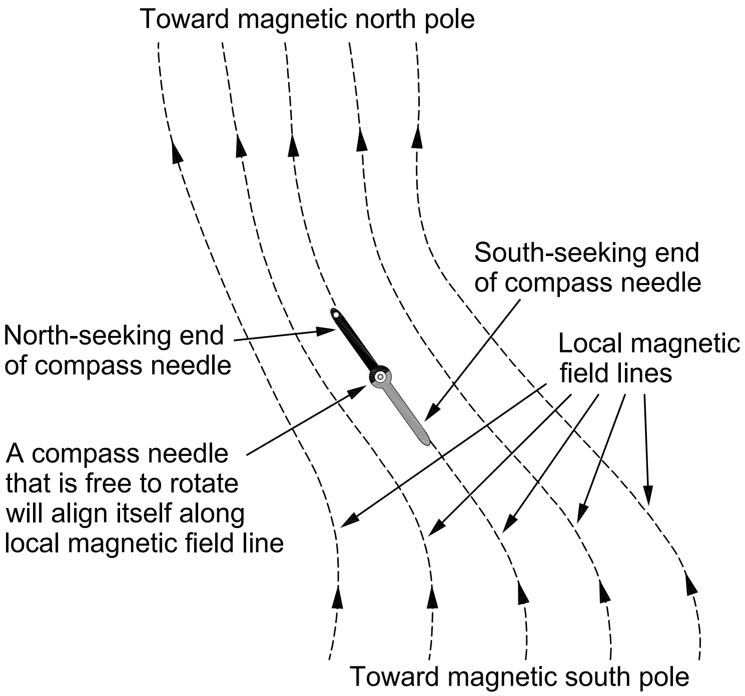
Toward magnetic north pole
South-seeking end of compass needle
North-seeking end of compass needle
Local magnetic field lines
A compass needle that is free to rotate will align itself along local magnetic field line
Toward magnetic south pole

Electronic compasses don't use magnetic needles but instead have sensitive magnetic sensors that determine the direction of the local magnetic field. The direction of the magnetic field is shown on a liquid crystal display as an arrow. Bearings can be stored in the internal memory of some electronic compasses.

A compass is primarily used to keep you pointed in a particular direction (bearing) as you travel, but can also be used to triangulate your position on a map. The accuracy of a standard compass is about two degrees in regions far away from the magnetic poles. At the magnetic poles, the compass needle would point straight up or down, making it useless for navigation. Other external magnetic fields, such as magnetism caused by power lines, local ore bodies, or lava flows can affect your compass needle or electronic compass.

Objects made from ferromagnetic materials will interfere with your compass if the object is close enough to the needle. The most common ferromagnetic materials contain iron or nickel. Before using your compass, test objects that you suspect are ferromagnetic or magnetized, such as belt buckles, ski poles, ice axes, tent poles, radios, or watches. Bring the object toward your compass and observe if, and at what distance, the needle moves. When navigating with a compass, make sure that all ferromagnetic or magnetized objects are kept far enough away to avoid interfering with your compass needle or digital compass.

2.3 Compass bearings

Depending on your navigation technique, compass bearings (directions) are measured relative to true (geographic) north, magnetic north, or grid north. The unit for measuring bearings is normally degrees, but gradians or other units are used on some compasses. There are 360 degrees (º), or 400 gradians (grad or gr), in one full rotation. Military compasses use the unit milliradian (mil) to measure bearings. There are 6283.1853... mils in one full circle, although most military compasses round off this number to 6000, 6400, or some other intermediate number.

True bearings

True compass bearings are measured as angles relative to the direction of the geographic north pole. They are always measured clockwise from the direction to the north pole, even in the southern hemisphere.

Each of the familiar cardinal directions is equivalent to a particular true bearing: north (0º), east (90º), south (180º), and west (270º). Midway in between are the intermediate directions: northeast (45º), southeast (135º), southwest (225º), and northwest (315º). These directions can be further subdivided into north-northeast (22.5º), east-northeast (67.5º), east-southeast (112.5º), south-southeast (157.5º), south-southwest (202.5º), west-southwest (247.5º), west-northwest (292.5º), and north-northwest (337.5º).

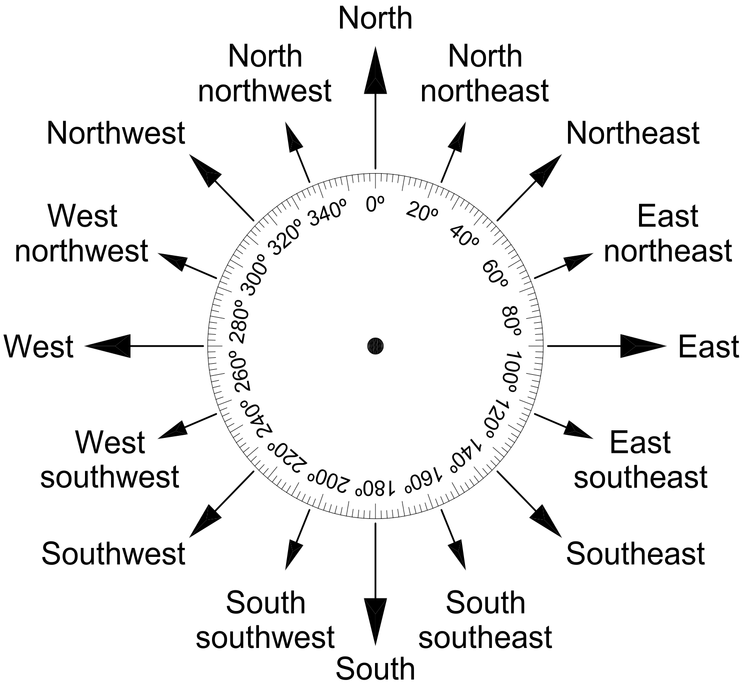
North
North northwest
North northeast
Northwest
Northeast
West northwest
East northeast
West
East
West southwest
East southeast
Southwest
Southeast
South southwest
South southeast
South
0º
20º
40º
60º
80º
100º
120º
140º
160º
180º
200º
220º
240º
260º
280º
300º
320º
340º

Cardinal and intermediate directions

Example of true compass bearings

Below are several direction arrows on a chart. Each arrow represents a particular true bearing.

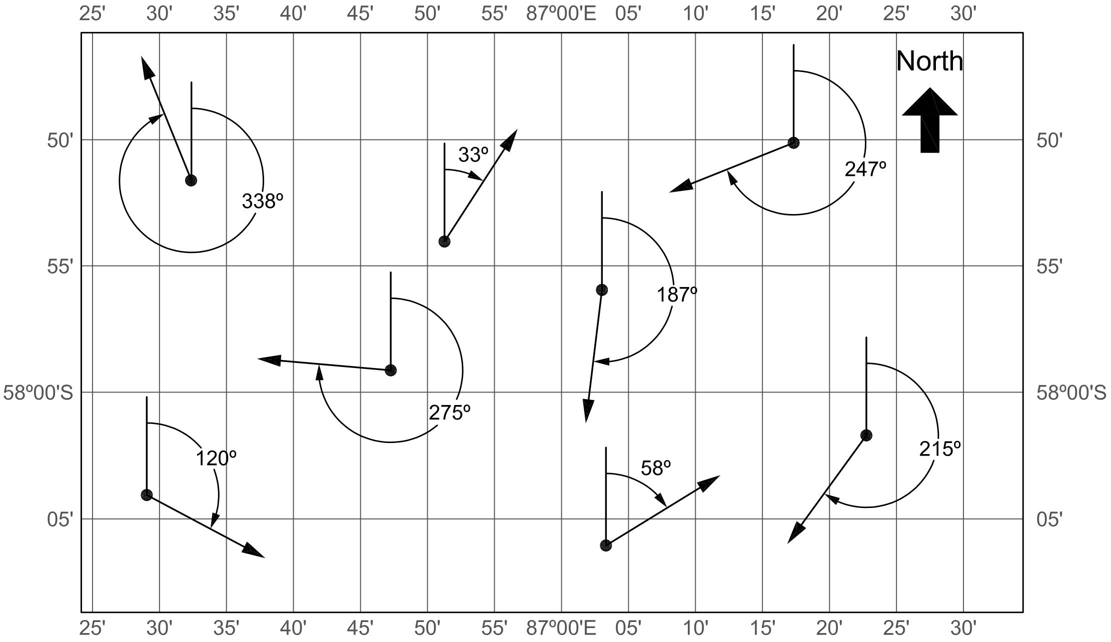
25'
30'
35'
40'
45'
50'
55'
87º00'E
05'
10'
15'
20'
25'
30'
North
50'
55'
58º00'S
05'
338º
33º
247º
187º
275º
120º
58º
215º

Note that although this chart represents a portion of the southern hemisphere, the bearings are relative to the geographic north pole, not the south pole. Longitudes are by definition north-south lines and are therefore convenient reference lines from which true bearings can be measured.

Magnetic and grid bearings

In certain navigation applications, compass bearings are measured relative to magnetic north or grid north instead of true north. Magnetic north is defined as the direction of the magnetic field at your location. Grid north is the direction of the north-pointing end of your map's north-south grid lines.

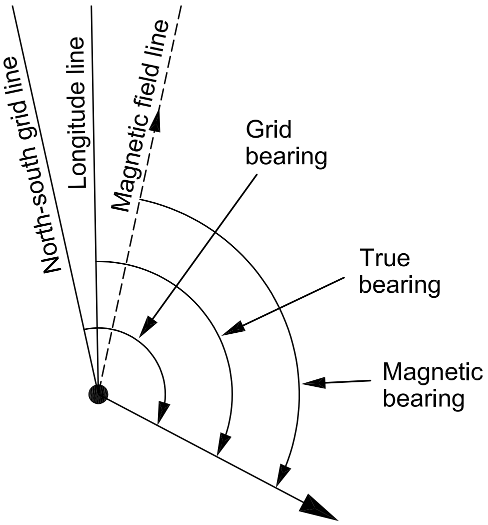
North-south grid line
Longitude line
Magnetic field line
Grid bearing
True bearing
Magnetic bearing

2.4 Magnetic declination

Magnetic field lines are not aligned north-south across most of Earth's surface. This means that the north seeking end of a compass needle will, in most cases, point in a direction other than toward the geographic north pole. When navigating with a map and compass, you must consider this effect whenever it causes a bearing error of more than a couple of degrees.

2.4.1 Magnetic declination theory

Most people mistakenly believe that a freely rotating compass needle points directly toward the nearest magnetic pole. This would only be true if magnetic field lines followed the shortest path along Earth's surface from your compass to the nearest magnetic pole, which is generally not the case.

The magnetic declination (variation) at a specific location is the angle between the direction of the north-seeking end of the needle and the direction to the geographic north pole. This definition applies everywhere on Earth including in the southern hemisphere. Since longitudes are perfect north-south lines along Earth's surface, the magnetic declination at your location can be defined as the angle between the north-seeking end of your compass needle and the north-pointing portion of your local longitude. If the north-seeking end of the needle points west (left) of the longitude, the declination is west, and if the needle points east (right) of the longitude, the declination is east.

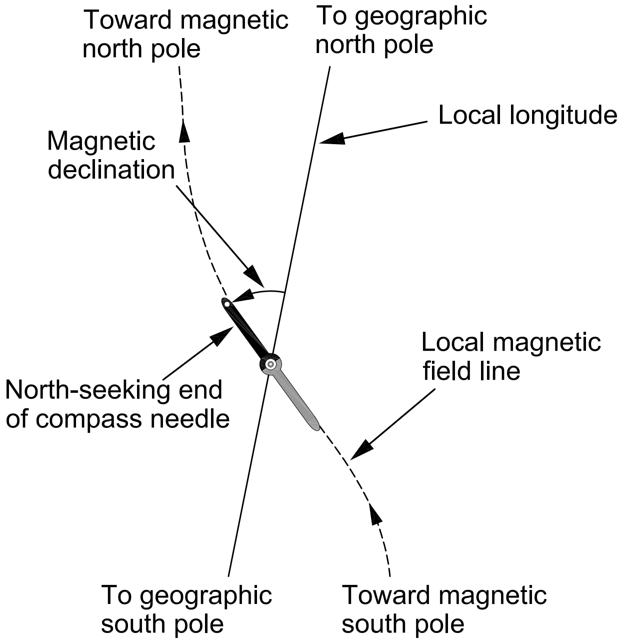
Toward magnetic north pole
To geographic north pole
Local longitude
Magnetic declination
Local magnetic field line
North-seeking end of compass needle
To geographic south pole
Toward magnetic south pole

To state the magnetic declination, write the angle followed by the label "E" or "W" to indicate whether it's east or west, for example 93ºE. By definition the declination is always somewhere between 180ºW and 180ºE.

Examples of magnetic declination:

The maps below show eight freely rotating compass needles at random locations in the southern and northern hemispheres. The table shows how the magnetic declination is derived for each location.

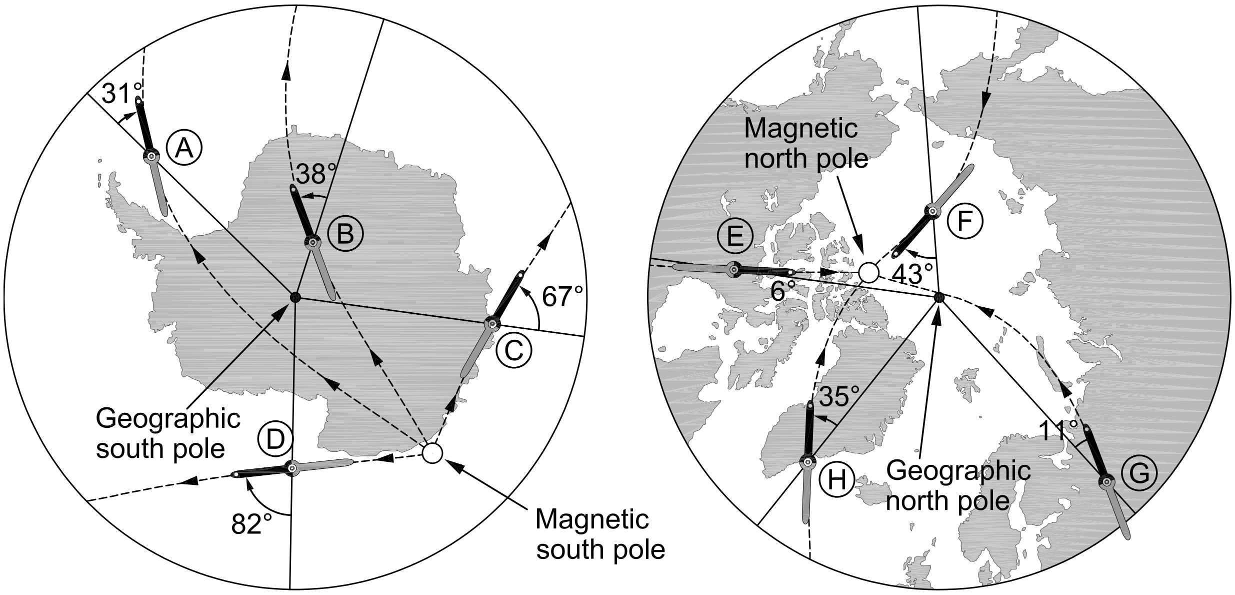

Location	Angle relative to longitude	Direction relative to geographic north pole	Magnetic declination
A	31°	right (east)	31°E
B	38°	left (west)	38°W
C	67°	left (west)	67°W
D	82°	right (east)	82°E
E	6°	left (west)	6°W
F	43°	right (east)	43°E
G	11°	right (east)	11°E
H	35°	left (west)	35°W

2.4.2 Estimating magnetic declination

The secular movement of Earth's magnetic field causes the declination to change over the years. Five different ways to estimate the magnetic declination at your location are described here: (1) declination diagram, (2) compass rose, (3) computer software, (4) GPS receiver, and (5) direct measurement.

Declination diagram

Most high-quality topographic maps have a diagram that shows the magnetic declination for the area covered by the map. Usually, the date and the annual rate of change when the declination was measured are also shown. The older the declination date, the more inaccurate the stated declination and annual change are likely to be. On some maps, the declination and date are written without an associated diagram.

Figuring out how to calculate the declination from a diagram may require some work, made even more challenging if you’re unfamiliar with the language used on the map. Some diagrams give you the declination relative to both geographic (true) and grid north. For land navigation, use grid north if you have a choice because you will likely be taking bearings relative to the map’s grid.

Examples of declination from diagrams:

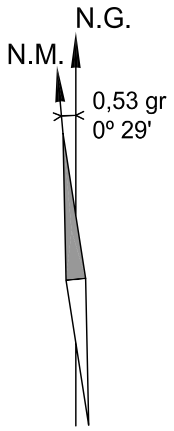

La déclinaison magnétique correspond au centre de la feuille au 1er janvier 1998 Elle diminue chaque année de 0,12 gr (0° 6')

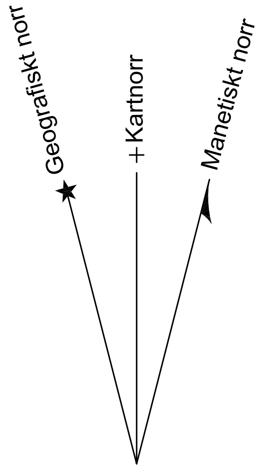

Meridiankonvergensen (skillnaden mellan kartnorr och geografisk norr) är i kartbladeds centrum +2,3°. Kompassens missgivning (skillnaden mellan magnetiskt och geografiskt norr) 2000 - 2005 är 5.0° ostlig.

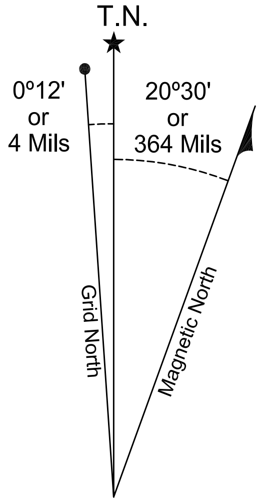

Use diagram only to obtain numerical values. Approximate mean declination 1996 for centre of map.
Annual change decreasing 8.0'

The left diagram is taken from a French map. It gives the declination as 0,53 gr or 0º29'. The date is given as 1998 with an annual decrease of 0,12 gr or 0º6'. Ignore the “gr” unit unless your compass dial is marked in gradians. N.G. stands for geographic north, and N.M. stands for magnetic north. In 1998 the compass needle would have pointed 29' left (west) of true north, giving a declination of 0º29'W. If the year is now 2005, the declination has decreased by 7 x 6' = 42' giving you a declination of 29'W - 42' = -13'W or 0º13’E. Rounding off this number to the nearest degree gives you a declination of 0º. In other words, ignore the declination when navigating with this map.

The middle diagram is taken from a Swedish map. The arrow on the right stands for magnetic north and the line with the star stands for geographic north. The middle line with “+” on top stands for map (grid) north. Magnetic north is given as +2,3º right (east) of grid north and 5,0º east of geographic north. Neither the year when the declination was measured, nor the annual rate of change is given. According to the map, the declination is valid only for a range of years (2000 - 2005). If the year is now 2008, you will have no choice but to use the given declination. You do however have a choice between grid north and geographic north. Choose grid north (2.3ºE) and round off this declination to 2ºE.

The right diagram shows a magnetic arrow pointing 20º30' (364 mils) to the right or east of geographic (true) north (T.N.). An added angle of 0º12' (4 mils) is given between geographic north and grid north. The declination year is 1996 with an annual decrease of 8.0'. Ignore the “mil” unit unless you’re using a military compass. Choosing grid north and assuming the year is now 2006 gives you a declination of 20º30'E + 0º12' - 10 x 8.0' = 20º30' + 12' - 80' = 20º30'E - 68' = 20º30'E - 1º08' = 19º22'E. Round off this number to 19ºE.

Compass Rose

Nautical charts use one or more compass roses instead of a declination diagram. A compass rose consists of an inner graduated ring that shows magnetic bearings, nested inside an outer graduated ring that shows true bearings. In marine terminology, declination is called “variation.” On a compass rose, the variation is the difference between any bearing on the inner ring, versus the same bearing on the outer ring. In addition, the variation as well as the date and annual rate of change are usually written inside the inner ring.

Example of variation from compass rose

The compass rose below shows a variation of 24º38'W for 2004, with an annual decrease of 7.2'. If the year is now 2012, the variation would be 24º38'W - 8 x 7.2' = 24º38' - 57.6' = 23º40.4'W. Round off this number to 24ºW.

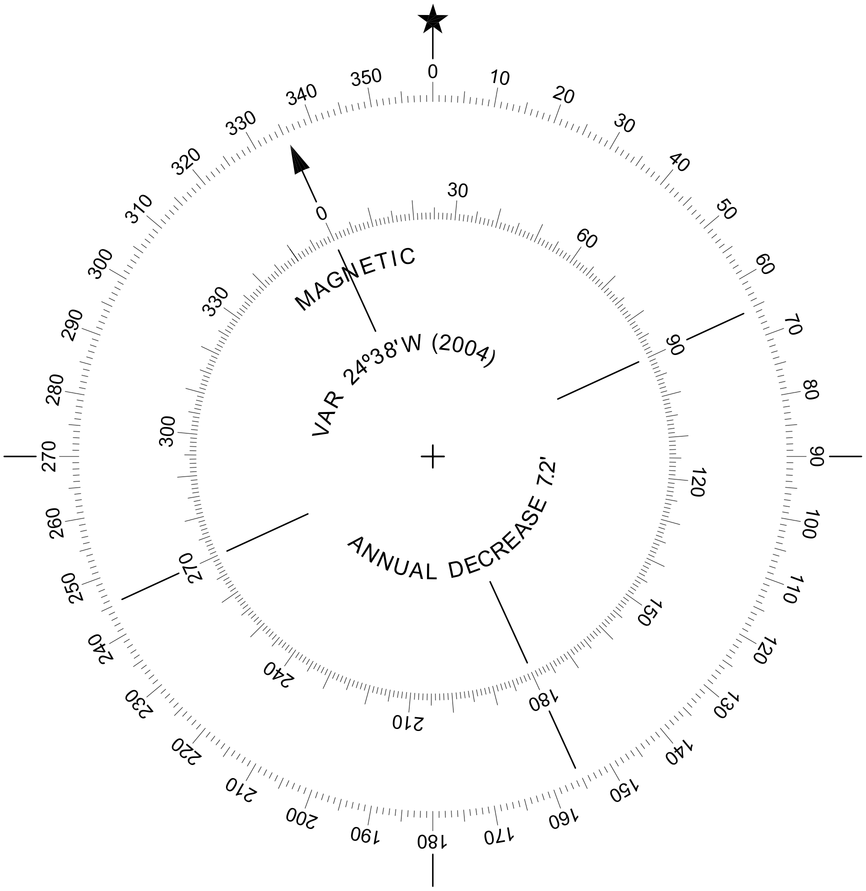
MAGNETIC
VAR 24º38'W (2004)
ANNUAL DECREASE 7.2'
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350

Computer software

Computer programs are available that calculate the magnetic declination for any point on Earth's surface. Just type in your position coordinates, the date, and the year. The program displays the declination. Be aware that the calculations are based on the shape of Earth's magnetic field and predicted future values at the time that the software was developed. Since magnetic field changes are somewhat unpredictable, use the latest version of the software for best accuracy. Find out whether the declination is given relative to geographic north or grid north. Some programs let you choose between the two systems.

GPS receiver

GPS receivers are equipped with software that estimates and displays the magnetic declination. Look in your GPS user manual for a way to display the magnetic declination at your location. Most receivers give you the declination relative to both geographic and grid north. GPS receivers use declination data based on the annual rate of change when the GPS software was developed. The receiver will therefore suffer from inaccurate declinations as it ages. Download new software with updated declination data into your receiver whenever the software becomes available. Future enhancements may allow declination data to be beamed directly from the GPS satellites to your receiver.

Direct declination measurement

With a map and compass, you can directly measure the magnetic declination if you know your location on the map and can identify a nearby landmark. Direct measurement is the surest way to determine your declination. It takes into account all the local and global anomalies.

1. Pinpoint your position on the map as accurately as possible. Use a GPS receiver if you have one.
2. Take a bearing off your map (Section 2.6.1) from your present location to an identifiable landmark and note the resultant bearing. For example 271º.
3. Take a field bearing off the same landmark (Section 2.6.2) and note the resultant bearing. For example 302º.
4. Your declination number is the difference between the map bearing and the field bearing: 302º - 271º = 31º.
5. Look at the graduated dial of your compass. If the field bearing is clockwise from your map bearing, the declination is west. If it is counterclockwise, the declination is east. In this example, the field bearing (302º) is clockwise from the map bearing (271º) so your declination is 31ºW.

Be careful with bearings on either side of the 0º mark. Let's say you take bearings off a second landmark, and get a map bearing of 334º and a field bearing of 7º. This gives you a declination number of 334º - 7º = 327º, but by definition, the declination number cannot be more than 180º. To correct this, subtract your initial declination number from 360º. This yields a declination of 360º - 327º = 33ºW. Averaging this results with the first measurement (31ºW) gives a declination of 32ºW. For even more confidence in your results, repeat the process with additional landmarks.

2.4.3 Compensating for magnetic declination

You must compensate for the magnetic declination whenever you use a compass together with a map. This applies to both mechanical and electronic compasses. There are two ways to compensate for the magnetic declination: (1) using a compass with an adjustable declination mechanism, or (2) adding or subtracting the declination after taking a compass bearing. You could also do all your navigation relative to magnetic north, which is often done in marine navigation.

Compass with adjustable declination mechanism

For land navigation, the author recommends that you use a compass with an adjustable declination mechanism. Just set the declination to the correct value, and then forget about declination until you move into an area with a different declination.

Many types of adjustable mechanisms exist. Look in your compass manual for instructions on how to adjust the declination. Be aware that salt water or dirt can seize the mechanism, thereby turning your adjustable compass into a nonadjustable type. Most electronic compasses have built-in functions to adjust the declination.

Example of declination mechanism

A widely-used declination mechanism for an orienteering compass consists of a small screw located on the backside of the housing. Use a small screwdriver to turn the screw until the scale shows the correct declination for your location. Note that the orienting arrow moves relative to the meridian lines when you turn the screw.

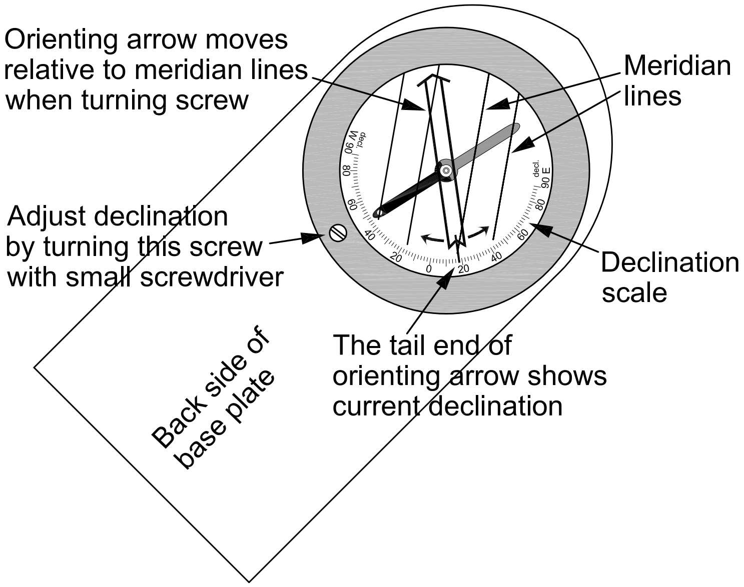
Orienting arrow moves relative to meridian lines when turning screw
Meridian lines
Adjust declination by turning this screw with small screwdriver
Declination scale
The tail end of orienting arrow shows current declination
Back side of base plate
decl. W 90
80
60
40
20
0
20
40
60
80
decl. 90 E

Compass without declination mechanism

On a compass without an adjustable declination mechanism, you must add or subtract the appropriate declination after taking a bearing. Whether you add or subtract depends on whether the declination is east or west and whether you're going from a map to the real world or vice versa (Table 2.4). Adding means turning the graduated dial counterclockwise, while subtracting means turning it clockwise. Pay attention to what you're doing. It's easy to do mistakes, especially if you have to turn the dial past the zero mark.

Table 2.4 - Declination adjustments

Declination	Map to real world	Real world to map
east	subtract	add
west	add	subtract

Compensating for declination - Example 1

Let's say the declination is 31°W. You've taken a bearing of 352° off a map (Section 2.6.1), and want to follow this bearing in the real world (Section 2.6.3). Table 2.4 tells you to add the declination because you're going from a map to the real world. The adjusted bearing becomes 352° + 31° = 383°. By definition, bearings must be between 0° and 360°, so subtract 360°. Set you compass to 383° - 360° = 23° before following this bearing.

Compensating for declination - Example 2

The declination is 18°E. You've taken a field bearing of 126° off a landmark (Section 2.6.2) and want to triangulate your position on the map. In this case, you're going from the real world to a map. Table 2.4 tells you to add the declination. Set your compass to 126° + 18° = 144° before placing the compass on the map.

2.5 Magnetic dip

Magnetic dip, also known as inclination, is the angle between level ground and the magnetic field line at your location. The dip is by definition 90° (vertical) at the magnetic poles and is 0° (horizontal) in some regions between the poles. In most parts of the world, the dip is somewhere between these two extremes. Because the compass needle tries to align itself along the local magnetic field line, the dip causes the needle to point upward or downward. If the dip is too large, the needle will rub against the top or bottom of the housing, preventing the needle from rotating freely.

Most high-quality compasses try to compensate for magnetic dip by using a counterweight on one end of the needle. Because the magnetic dip varies across Earth's surface, a compass with a counterweight designed for a certain area will not work elsewhere. Before purchasing a compass, make sure it is designed for the area that you intend to travel.

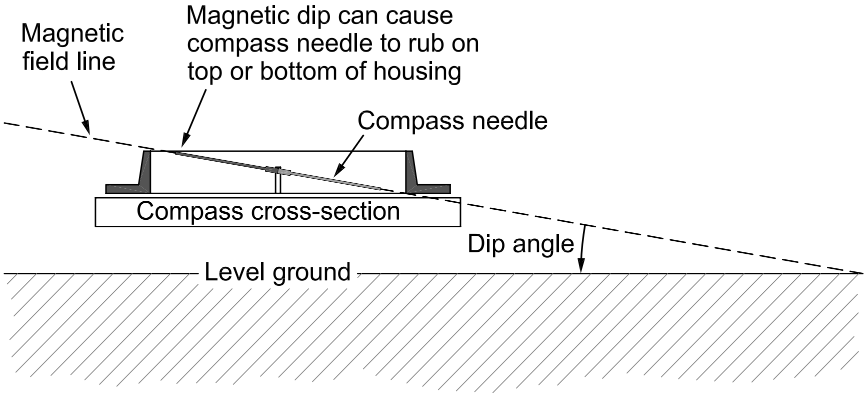

Some compasses feature deep well designs with lots of pivot room for the compass needle. Other designs use a multi-pivot mechanism to keep the needle level. These so-called world compasses can be used anywhere in the world except at extreme dip locations such as near the magnetic poles.

You may be able to compensate for the magnetic dip by holding your compass at an angle. This makes the operation of the compass more difficult, particularly with a mirror compass. Electronic compasses will function anywhere but become less accurate with increasing dip. Most marine compasses are balanced for the dip of a particular area, although they usually work in other areas but become less accurate.

2.6 Orienteering compasses

For land navigation, you need a compass with a base plate. Without a base plate, you cannot use your compass to take bearings off a map. Both mirror and standard orienteering compasses are fitted with a base plate, and either type is good for wilderness navigation. A mirror compass is more accurate for aiming at a target, but is more difficult to use than a standard orienteering compass. Most procedures in this book assume that you are using a standard compass, but this should not be a concern for someone using a mirror compass. Except for the aiming process, both types of compasses are used in exactly the same manner. Some orienteering compasses have a built-in inclinometer that lets you measure the steepness of slopes. If you're familiar with trigonometry, you can use the inclinometer to measure heights of features.

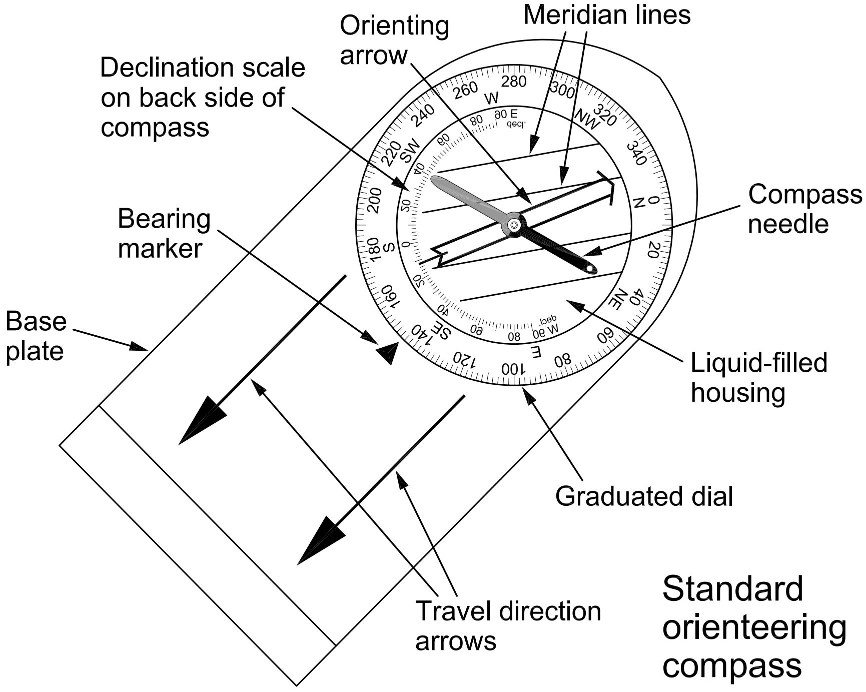

Standard orienteering compass

Base plate

The base plate, made of transparent plastic, is more than a structural component. Its long edges are used for taking bearings off a map or for drawing triangulation lines on the map. The travel direction arrow(s) and bearing marker are imprinted on the base plate. Most base plates are inscribed with scales of various units that can be used as rulers.

Compass needle

The compass needle is a magnetized strip of metal that is balanced on a pivot joint. The north-seeking end of the needle is usually painted red and the south seeking end white or black, but some compasses use other color combinations.

Travel-direction arrows

One or more travel-direction arrows are imprinted on the base plate and show the travel direction when following a compass bearing. They are also used for taking a field bearing. On a mirror compass, a sighting notch or a mirror line is used instead of travel-direction arrows.

Housing with graduated dial

The housing is a liquid-filled container that encloses the compass needle. The liquid is used to dampen the movement of the needle. The perimeter of the housing consists of a graduated dial that is marked with bearings from 0º through 360º. Some compasses use other angle units. To adjust the compass bearing, rotate the housing and read off the bearing on the bearing marker located on the base plate.

Meridian lines

The meridian lines make it easier to align the housing north-south on the map when taking bearings or drawing lines of position. The meridian lines are mechanically linked with the housing and move together as one unit.

Orienting arrow

The orienting arrow is used to correctly align the base plate relative to the compass needle when following a compass bearing or taking a field bearing.

Declination mechanism

Some compasses have a built-in declination mechanism. The declination mechanism compensates for the error between magnetic and geographic bearings by adjusting the angle of the orienting arrow relative to the graduated dial. The declination is adjusted manually and read off the declination scale.

2.6.1 Taking a bearing off a map

Taking a bearing off a map is an elementary concept in compass navigation. It involves using a map and orienteering compass to measure the direction toward a selected target.

1. Use terrain association (Section 8.1), a GPS receiver, or any other method to determine your location on the map.
2. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north.
3. Place your compass on the map. The map can be oriented in any direction. Align the base plate with an imaginary straight line that passes through both your present location and your desired target. Make sure the travel-direction arrows on the base plate (or mirror line) point toward your desired target on the map, not the reverse direction.

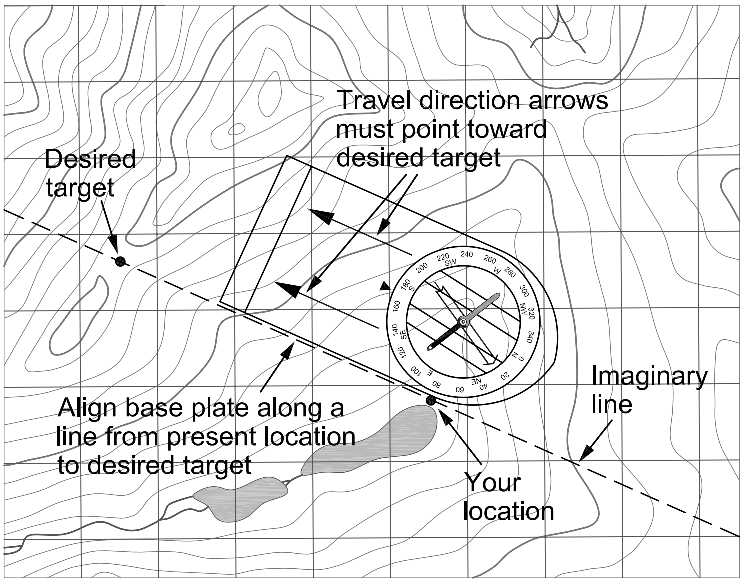
Travel direction arrows
must point toward
desired target
Desired
target
Imaginary
line
Align base plate along a
line from present location
to desired target
Your
location

Step 3: Place and align compass on map

4. Without moving the base plate, turn the graduated dial until the meridian lines on the compass are parallel to the north-south grid lines on the map. Make sure north (0°) on the graduated dial points toward north on the map and not south.

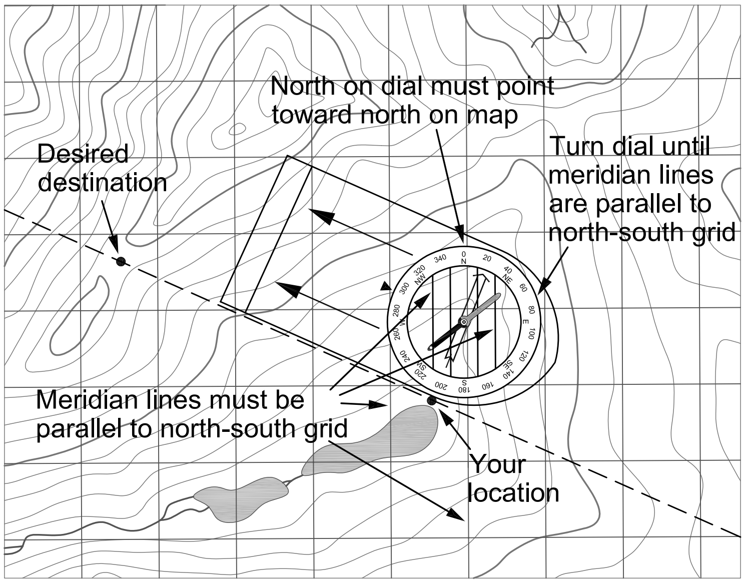
North on dial must point
toward north on map
Desired
destination
Turn dial until
meridian lines
are parallel to
north-south grid
Meridian lines must be
parallel to north-south grid
Your
location

Step 4: Turn graduated dial

5. Remove your compass from the map and read off the bearing. You may want to record this bearing for later use.
6. If your compass lacks an adjustable declination mechanism, adjust the graduated dial by the appropriate declination number. Since you're going from a map to the real world, subtract the declination number for an east declination, add for a west declination.

2.6.2 Taking a field bearing

Taking a field bearing involves measuring the direction to a terrain feature in the real world. Three primary reasons exist for taking a field bearing: (1) determining your position by triangulation, (2) staying on track toward an intermittently visible target, or (3) directly measuring the magnetic declination.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north. Actually, this only applies if the bearing is going to be used for triangulation. If the bearing is used to stay on track toward the landmark or to measure the declination, the declination setting is arbitrary as long at it is kept constant throughout the procedure.
2. Aim the travel direction arrows (or mirror line) of the compass toward the landmark whose direction you want to measure.

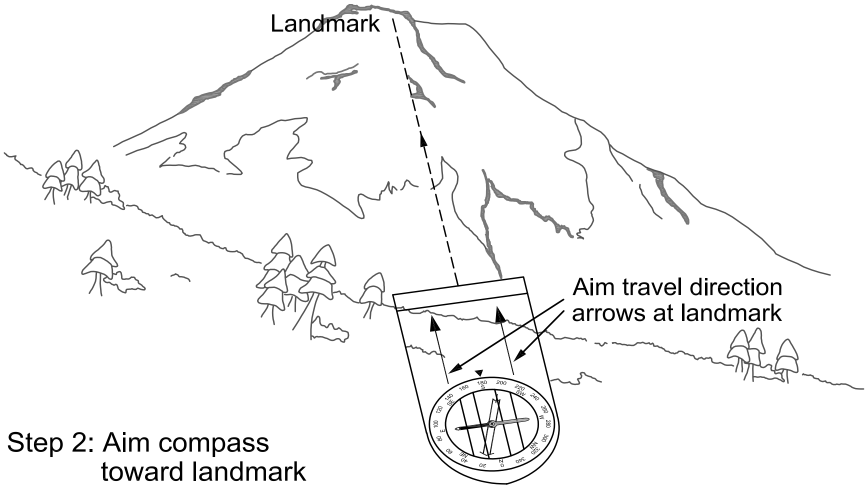
Landmark
Aim travel direction arrows at landmark

Step 2: Aim compass toward landmark

3. While keeping the base plate level and the travel-direction arrows pointed toward the landmark, turn the graduated dial until the north-seeking end of the compass needle lines up with the north end of the orienting arrow. Read off the bearing.

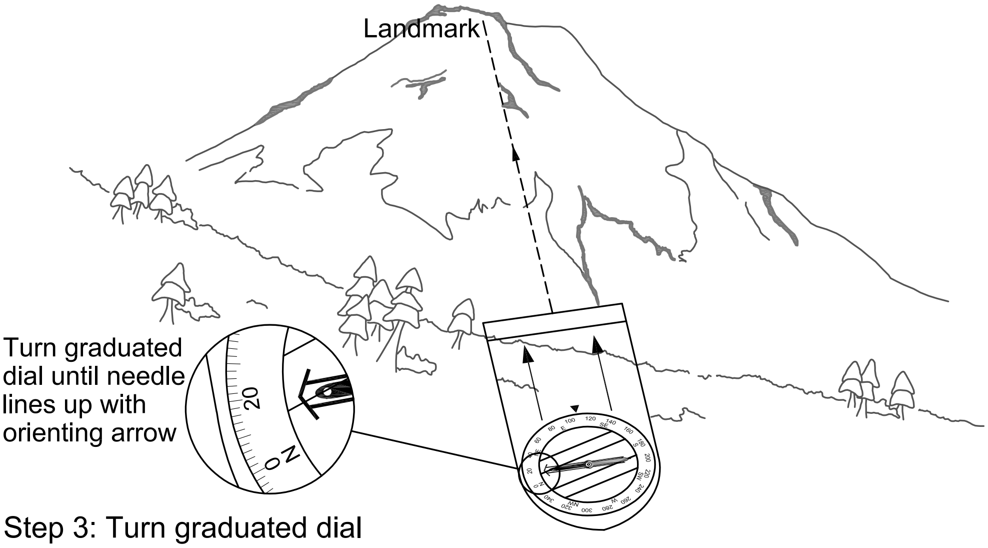
Landmark
Turn graduated dial until needle lines up with orienting arrow
20
N

Step 3: Turn graduated dial

4. If your compass lacks an adjustable declination mechanism, adjust the graduated dial by the appropriate declination number before performing a triangulation. Since you're going from the real world to a map, add the declination number for an east declination, subtract for a west declination. Don't adjust anything if the bearing is for keeping track of your direction toward the landmark. The compass is obviously already pointing in the correct direction. The same applies for a bearing taken to measure the declination. You cannot compensate for an unknown declination that you are about to calculate.

2.6.3 Following a compass bearing

Following a compass bearing is what you do in the real world after taking either a map or a field bearing. As you travel, the compass shows you the direction to your chosen target.

1. Hold the compass in front of you with the travel direction arrows (or mirror line) pointing directly ahead.
2. While keeping the base plate level, turn your body together with the compass until the north-seeking end of the compass needle coincides with the north end of the orienting arrow. The travel-direction arrows or mirror line will now point toward your target.

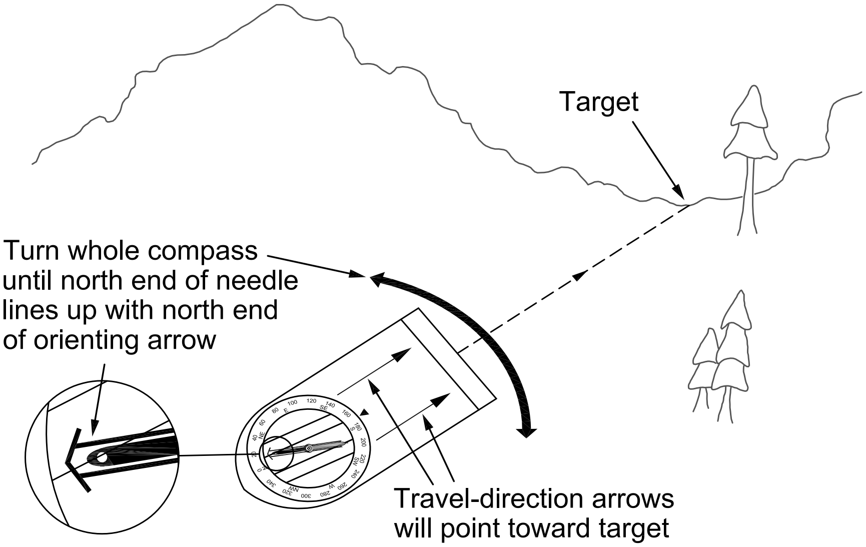
Target
Turn whole compass until north end of needle lines up with north end of orienting arrow
Travel-direction arrows will point toward target

3. Follow the travel direction arrows or the mirror line toward the target.

To avoid constantly looking at your compass, follow a series steering marks along your line of travel. A steering mark can be any recognizable feature in line with your compass bearing. If you have a choice, select the highest and most distant feature. Keep in mind that a feature can change appearance as you move toward it. If a steering mark is beyond your target, be careful not to overshoot your target. During poor visibility, you may have to select steering marks that are very close together. This will necessitate frequent checks of your compass, and in extreme cases such as whiteouts, you may be forced to continuously look at your compass.

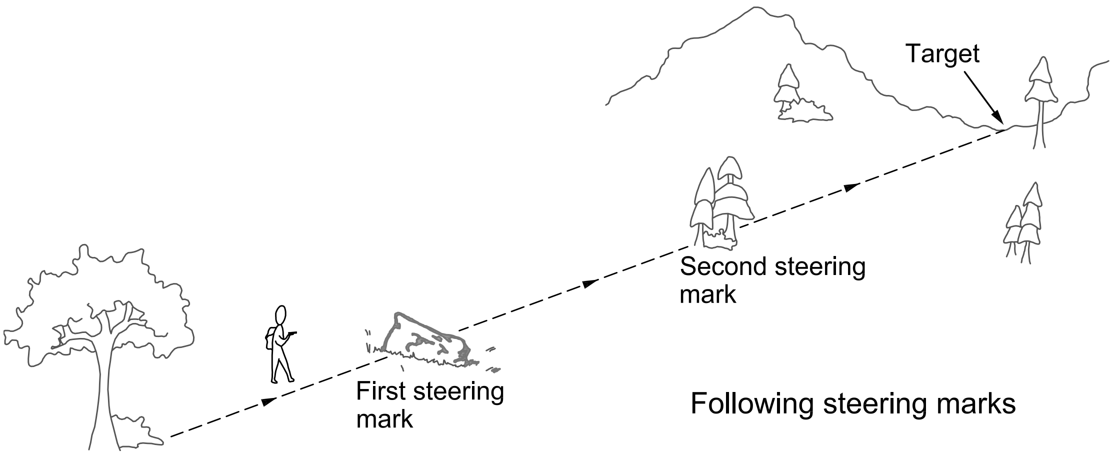
Target
Second steering mark
First steering mark
Following steering marks

If you cannot follow the compass bearing because of an intervening obstacle, go around the obstacle and then try to regain your original line of travel. This is easy if there is a steering mark on the far side of the obstacle. After circumventing the obstacle, just go to the steering mark to regain your original line of travel.

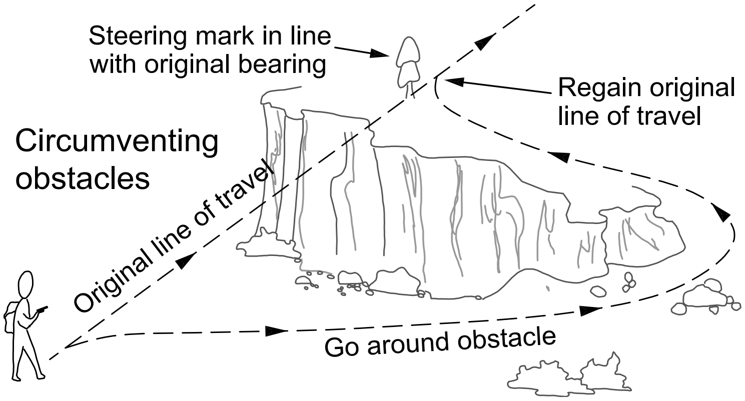
Steering mark in line with original bearing
Regain original line of travel
Circumventing obstacles
Original line of travel
Go around obstacle

If there is no steering mark on the far side of the obstacle, use a back bearing (described next) or dead reckoning (Section 8.3.3) to regain your line of travel.

Using a back bearing

A back bearing is the opposite direction from your original bearing to the target. Pointing your compass along a back bearing will reveal if you are drifting off course and help you regain your line of travel after circumventing an obstacle.

1. Leave all the compass settings unchanged from your original bearing.
2. Turn your body together with the compass until the south-seeking end of the compass needle coincides with the north end of the orienting arrow. This effectively reverses your original bearing and the travel direction arrows will point in the opposite direction from your target.
3. If you're on track, the back bearing will point directly toward a steering mark along your previously traveled route. If it doesn't, move sideways until the back bearing points directly toward the steering mark. You will then have regained your original line of travel.

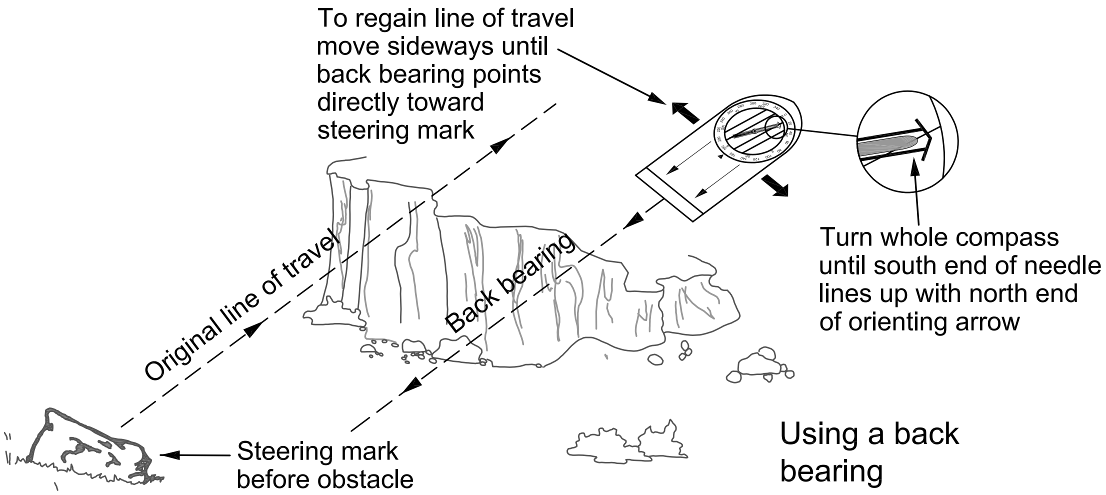
To regain line of travel move sideways until back bearing points directly toward steering mark
Turn whole compass until south end of needle lines up with north end of orienting arrow
Original line of travel
Back bearing
Steering mark before obstacle

Using a back bearing

2.6.4 Triangulating from a linear feature

With a map and compass, all you need is one visible landmark to pinpoint your position, provided you're located on, or near, a linear feature such as a trail, road, power line, narrow ridge, river, or lake shore.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north.
2. Aim your compass toward an identifiable landmark and take a field bearing off that landmark (Section 2.6.2). For best accuracy, pick the nearest landmark that is approximately perpendicular to the linear feature.

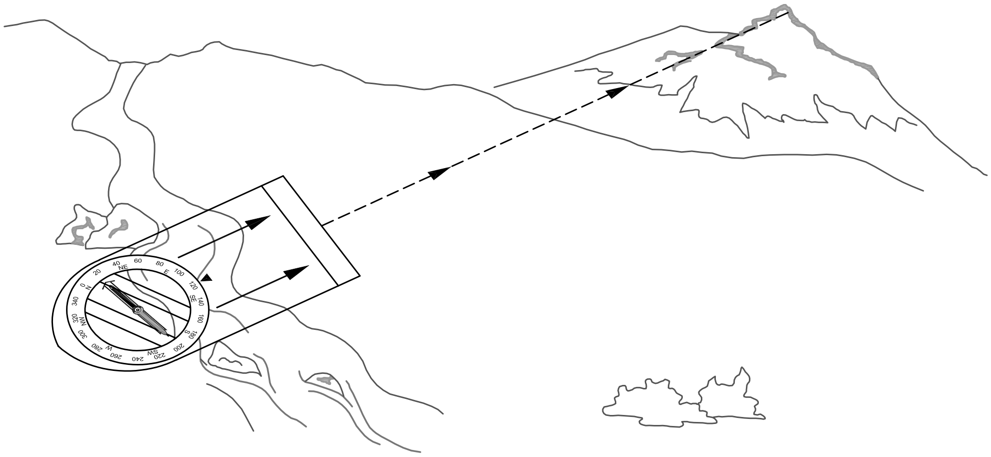

Step 2: Take a bearing off a landmark

3. If your compass lacks an adjustable declination mechanism, adjust the graduated dial by the appropriate declination number. Since you're going from the real world to a map, add the declination number for an east declination, subtract for a west declination.

4. Place the compass on the map with a corner of the base plate on the landmark.

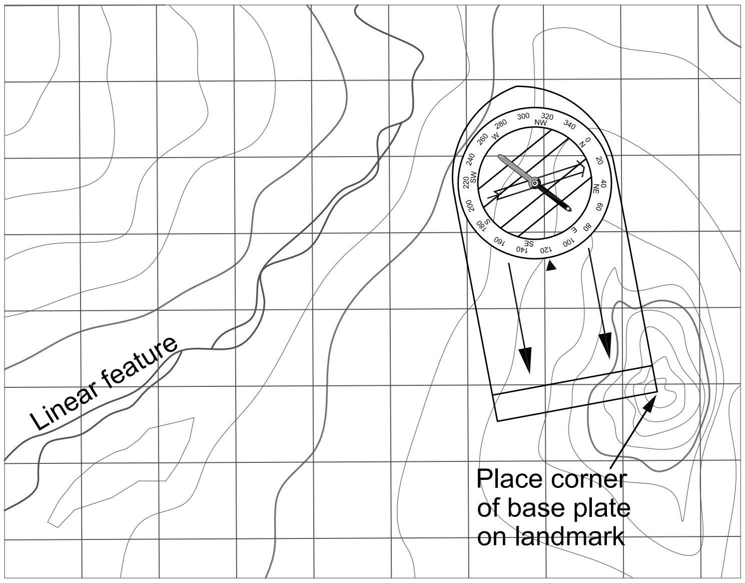
Linear feature
Place corner of base plate on landmark

Step 4: Place corner of compass on landmark

5. Using the corner of the base plate as a pivot point, rotate the whole compass until the meridian lines on the compass are parallel to the north-south grid lines on the map. Make sure north (0°) on the graduated dial points toward north on the map, not south. Do not rotate the graduated dial during this step.
6. On the map, draw a straight line (real or imaginary) parallel to the base plate from the landmark to the linear feature that you are on. Your position is where the straight line and the linear feature intersect.

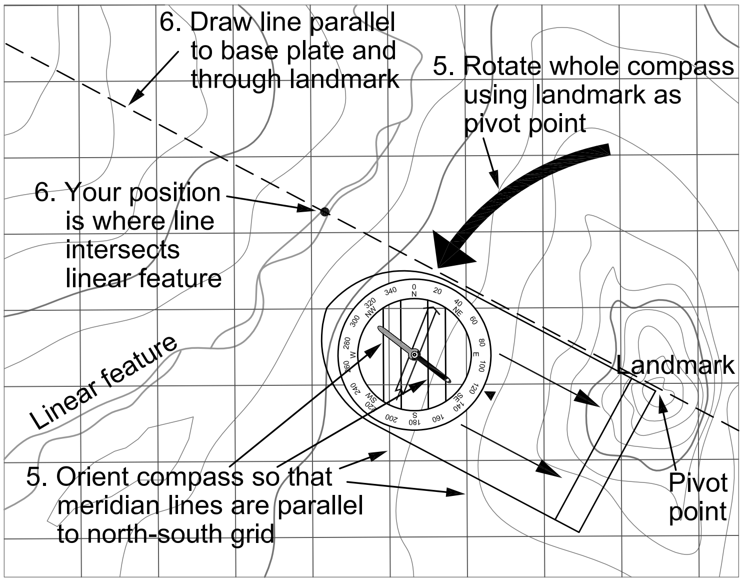
6. Draw line parallel to base plate and through landmark
5. Rotate whole compass using landmark as pivot point
6. Your position is where line intersects linear feature
Linear feature
Landmark
5. Orient compass so that meridian lines are parallel to north-south grid
Pivot point

Step 5 - 6: Orient compass and draw line to linear feature

If you intend to return to this position, record the field bearing that you have taken off the landmark (Step 2). On your return, first find your way back to the linear feature. Set the compass to the recorded bearing and then aim the compass like you would if you were following the bearing. If the compass points to the right of the landmark, move left along the linear feature to reach the recorded position; if the compass points to the left of the landmark, move to the right. When the compass points directly at the landmark, you will have reached your recorded position.

2.6.5 Triangulating off multiple landmarks

With a map and compass, you can pinpoint your location anywhere on the map by taking field bearings off two or more landmarks.

1. Use your compass to take a field bearing off a landmark. For best accuracy, select the nearest identifiable landmark.
2. Place the compass on the map with a corner of the base plate on the landmark.

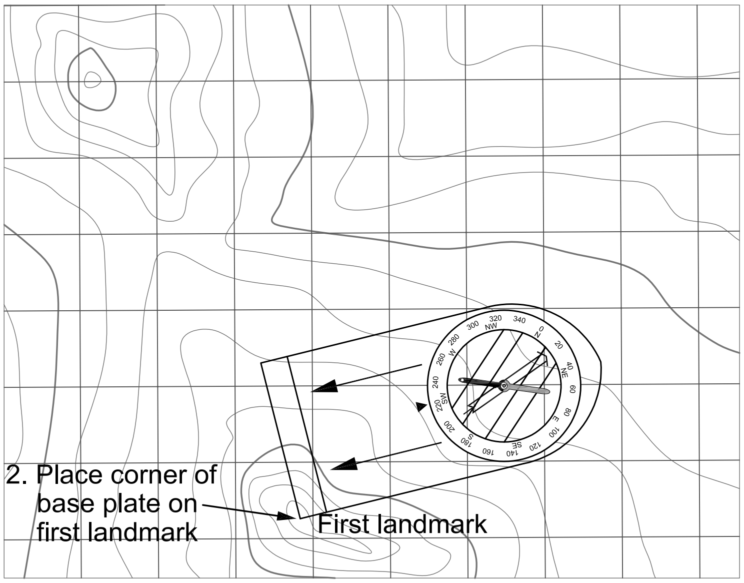

Step 2: Place corner of compass on first landmark

3. Using the corner of the base plate as a pivot point, rotate the whole compass until the meridian lines on the compass are parallel to the north-south grid lines on the map. Make sure north (0°) on the graduated dial points toward north on the map, not south. Do not rotate the graduated dial during this step.
4. On the map, draw a straight line (real or imaginary) parallel to the base plate and through the landmark. This is your first line of position.

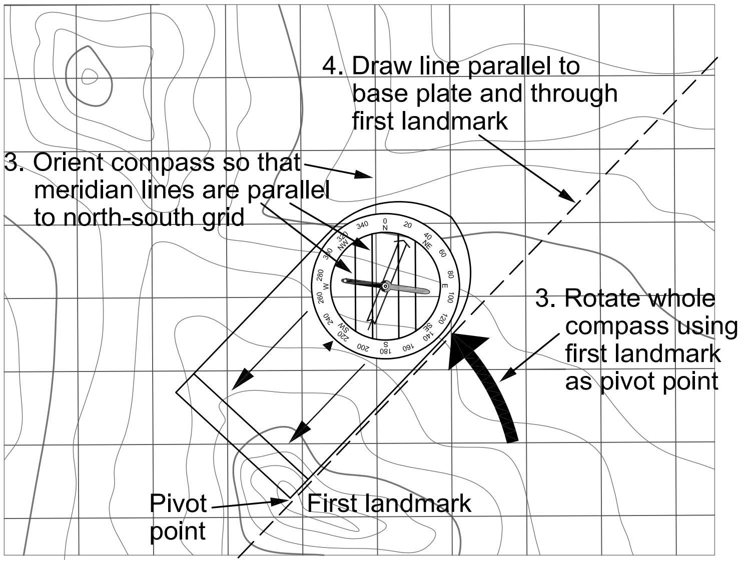
4. Draw line parallel to base plate and through first landmark
3. Orient compass so that meridian lines are parallel to north-south grid
3. Rotate whole compass using first landmark as pivot point
Pivot point
First landmark

Step 3 - 4: Orient compass and draw line

5. Use your compass to take a field bearing off a second landmark. For best accuracy, the direction to the second landmark should be approximately perpendicular to the first landmark.
6. Place the compass on the map with a corner of the base plate on the second landmark.

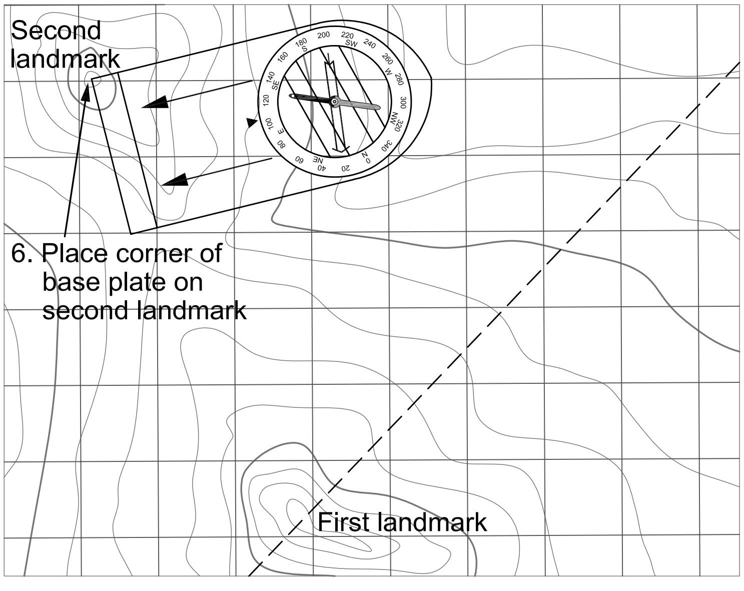
Second landmark
6. Place corner of base plate on second landmark
First landmark

Step 6: Place corner of compass on second landmark

7. Using the corner of the base plate as a pivot point, rotate the base plate until the meridian lines on the compass are parallel to the north-south grid lines on the map.
8. Draw a straight line parallel to the base plate and through the second landmark. This is your second line of position. Your location is the intersection between the first and second line of position.

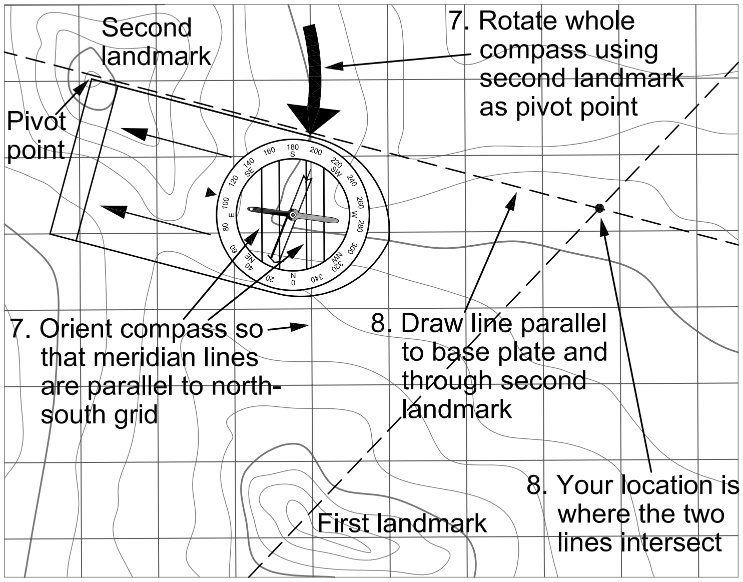
Second landmark
Pivot point
7. Rotate whole compass using second landmark as pivot point
7. Orient compass so that meridian lines are parallel to north-south grid
8. Draw line parallel to base plate and through second landmark
8. Your location is where the two lines intersect
First landmark

Step 7 - 8: Orient compass and draw second line

9. To confirm that you've done the triangulation correctly, repeat the process with bearings off additional landmarks. If you've done your triangulation correctly, all lines of position will intersect.

2.7 Marine compass

A marine compass is designed to work on a pitching boat. Most types of marine compasses are permanently attached to the boat, but there are also handheld units equipped with a sight for taking field bearings. The techniques for taking and following bearings with a marine compass are different than with an orienteering compass.

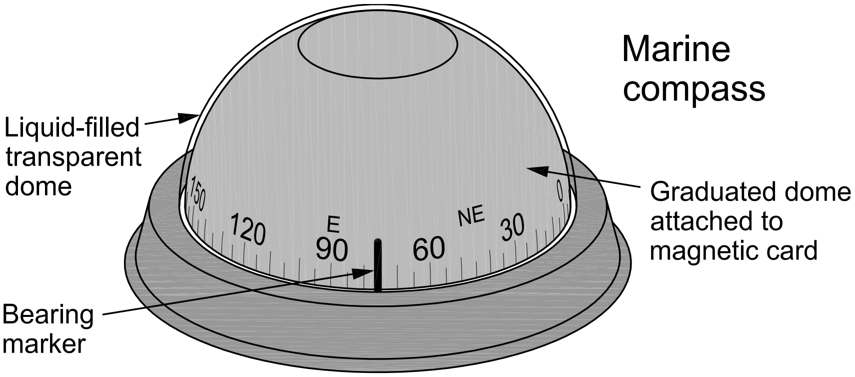
Marine compass
Liquid-filled transparent dome
Graduated dome attached to magnetic card
Bearing marker
150
120
E
90
60
NE
30
0

Instead of a needle, a marine compass uses a magnetic card. The card is mounted on a base assembly via gimbals that keep the card level as the boat rocks. The bearing numbers are either marked directly on the magnetic card or on a graduated dome attached to the card. The card assembly is enclosed inside a liquid-filled, transparent dome. A line inscribed on the transparent dome serves as a bearing marker.

Marine compasses generally do not have a declination mechanism, but some are equipped with adjustable magnets that are designed to compensate for interfering ferromagnetic material on the vessel.

2.7.1 Taking a bearing off a nautical chart

A standard marine compass cannot be used to take a bearing off a chart or map because the compass lacks a base plate and is part of the boat. Most ocean navigators use a plotting tool instead. Many types of plotting tools are available, including protractors, parallel rulers, and specially designed course plotters. Some plotting tools are used in conjunction with the chart's compass rose. You could also use an orienteering compass to take the bearing (Section 2.6.1). Before purchasing a plotting tool, make sure it is suitable for use on your watercraft. To take accurate bearings, you need a conveniently located flat surface that is reasonably protected from the elements.

The following procedure describes how to use a set of parallel rulers to take a bearing off a nautical chart. The chart must have at least one compass rose.

1. Place the parallel rulers on the chart. The chart can be oriented in any direction. Align one edge of the ruler with an imaginary straight line that passes through both your present location and your desired target.

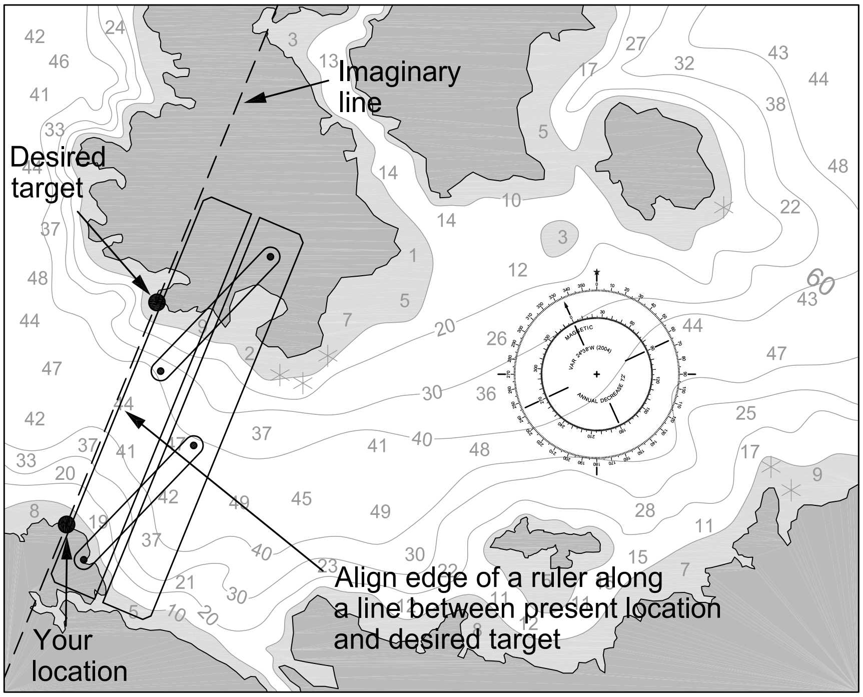
Imaginary line
Desired target
Your location
Align edge of a ruler along a line between present location and desired target

Step 1: Place and align parallel rulers on chart

2. Walk the rulers over to the nearest compass rose by alternately holding down one ruler while moving the other ruler. Be careful not to lose the initial orientation of the rulers during the walk. Align one edge of the ruler with the center point of the compass rose.
3. Read the magnetic or true bearing off the compass rose. In most marine applications, the navigation is done with magnetic bearings. The reverse bearing can be read off on the opposite side of the compass rose.

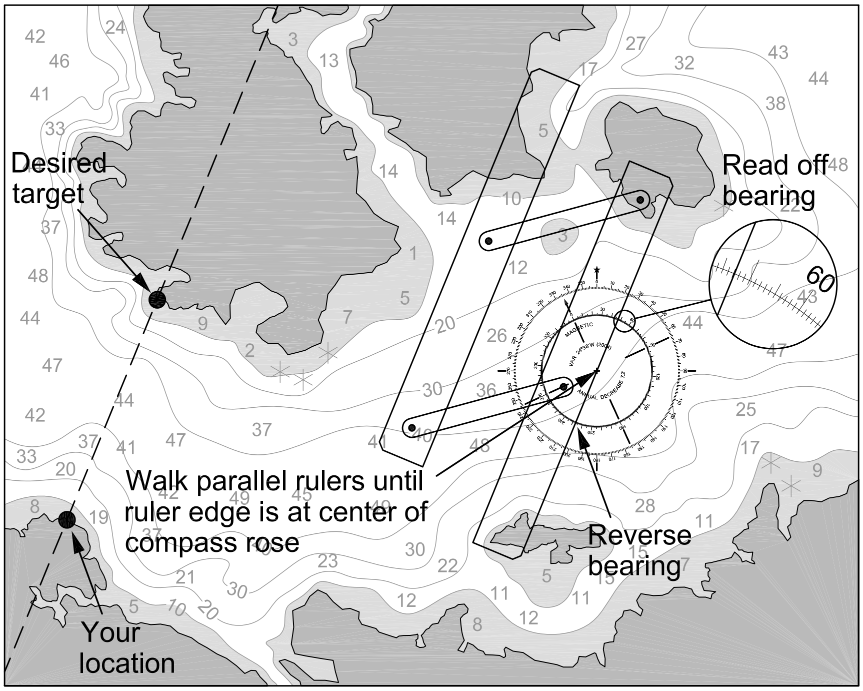
Desired target
Read off bearing
60
Walk parallel rulers until ruler edge is at center of compass rose
Reverse bearing
Your location

Step 2 - 3: Walk parallel rulers to compass rose
and read off bearing

2.7.2 Taking a field bearing with a marine compass

To take a field bearing with a marine compass, use the hull of your boat as a base plate and the bow as a travel direction arrow.

1. Position yourself along the centerline of the boat. Turn the boat until the bow points directly toward your selected landmark.
2. Read the bearing off your boat's compass and record the bearing. Most marine compasses show magnetic bearings.

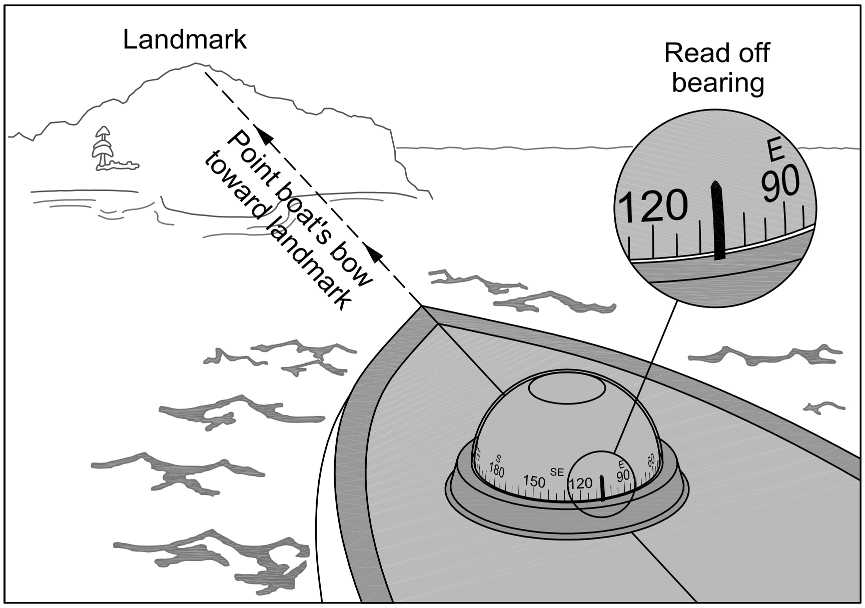
Landmark
Point boat's bow toward landmark
Read off bearing
E
90
120
S
180
150
SE
60

Step 1 - 2: Point bow toward landmark and read off bearing

2.7.3 Following a compass bearing on the water

Following a compass bearing on the water is more complicated than on land. On land, your primary hurdle is circumventing obstacles without losing you course. On the water, you are faced with the considerable challenge of adjusting your heading to compensate for the drift caused by currents, winds, and waves.

1. Estimate the anticipated forward speed of your vessel relative to the water in the direction of your target. Let's say for example that your target is at a magnetic bearing of 180º, and you expect your average speed to be 3.5 knots (nautical miles per hour) in that direction.
2. Find the average speed and direction of the water current between you and your target. Your chart may provide this information. In areas subjected to tidal currents, you need a clock and tidal current tables that show the current at specific dates and times. Suppose for example that the average current between you and your target will be 1 knot at a magnetic bearing of 230º during your journey.
3. Recalculate your bearing and speed to compensate for the current. This is easiest done graphically on your chart with the help of plotting tools. Choose any convenient and consistent distance unit to represent speed. For example, let 1 cm represent 1 knot. Draw a line 3.5 cm long in the magnetic direction of 180º to represent your boat's 3.5-knot forward speed. At the end of this line, draw another line 1 cm long in the direction of 230º to represent the 1-knot current. Draw a line from the start of the first line to the end of the second line. The third line represents your recalculated bearing and speed. Measuring the direction and length of this line gives a bearing of 165º and a length of 3 cm. These numbers represent the boat's anticipated magnetic bearing (165º) and speed (3 knots) relative to solid ground.

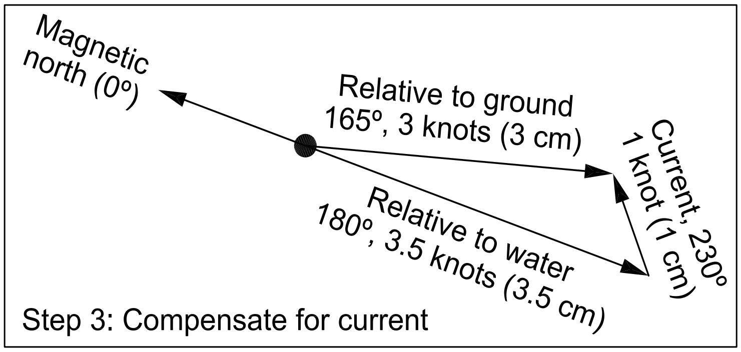
Magnetic north (0º)
Relative to ground 165º, 3 knots (3 cm)
Current, 230º 1 knot (1 cm)
Relative to water 180º, 3.5 knots (3.5 cm)

Step 3: Compensate for current

4. Determine the drift caused by wind and waves. This type of drift is difficult to calculate ahead of time because it changes as your heading changes. The surest way to correct for the drift is to do it after your boat is underway. The key factor to remember is that your boat is moving in the direction of the wake, not where the bow is pointing. Point the bow of your boat toward your anticipated bearing (165º). Estimate the direction of the boat's wake by looking back at the wake from the center of your compass. In this case, the wake is in line with 175º.

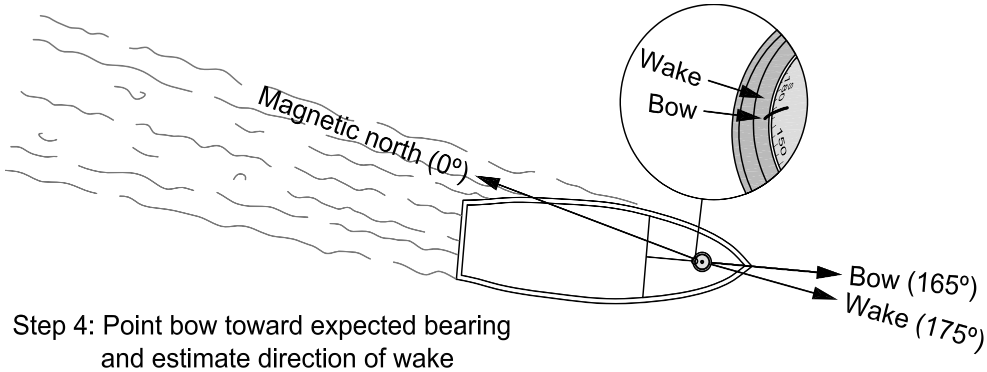

Step 4: Point bow toward expected bearing and estimate direction of wake

5. To compensate for the drift caused by wind and waves, turn the boat until the wake lines up with your anticipated bearing (165º). Read the boat's new heading off the bearing marker (150º), and follow this heading to your target.

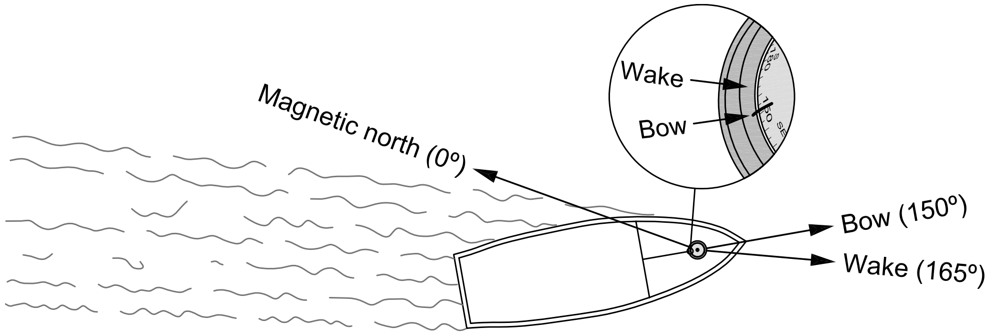

Step 5: Orient the boat until wake is in line with expected bearing

You may have to go through several iterations of this procedure because compensating for wind and waves changes your speed relative to the water, which in turn changes your numbers in Step 1 - 3.

2.8 Orienting a map or chart

A map or chart is oriented when geographic north on the map or chart corresponds to geographic north in the real world. It is much easier to compare what you see on the map with the real world when the map is oriented. Several techniques will accomplish this. The most reliable methods require a compass. Depending on your direction of travel, an oriented map or chart could be upside down, sideways, or at some other abnormal angle, making the text on the map difficult to read.

2.8.1 By terrain association

Orienting a map by terrain association means turning the map until what you see in the real world lines up with what you see on the map. Imagine yourself reduced in size by the same ratio as the map scale, and then placed on the map at your present location. In which direction and how far away would the surrounding landmarks be? Imagine a giant turning the map until the landmarks on the map point in the same direction as the corresponding real world landmarks. The map would then be oriented.

After turning the map, make sure that each visible linear feature in the real world is parallel to the same linear feature on the map, and don't ignore the alignment of features behind you. Also, check that the orientation isn't off by 180º. Terrain association is the quickest and easiest way to orient a map or chart, but lacks the precision of using a compass.

2.8.2 With orienteering compass and grid lines

Using a compass and the map's grid lines to orient the map is a very accurate method, provided your magnetic declination adjustment is correct.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north and then turn the graduated dial to north (0º). If your compass lacks an adjustable declination mechanism, turn the graduated dial to 0º and then adjust the dial by the appropriate declination number. Since you're going from a map to the real world, subtract the declination number for an east declination, add for a west declination.
2. Place your compass on the map. Align the base plate along one of the north-south grid lines and with the travel direction arrows pointing toward north on the map.

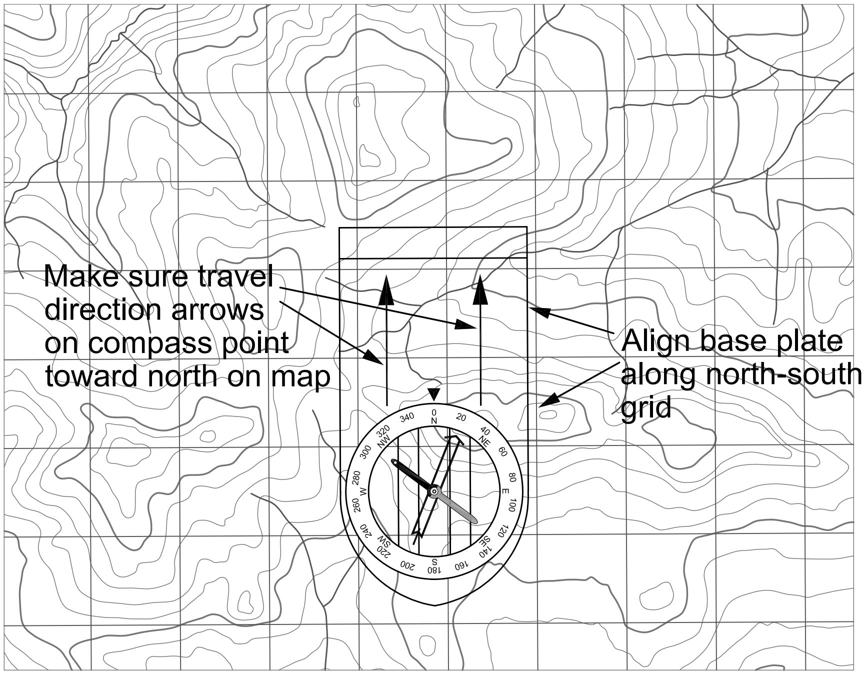
Make sure travel direction arrows on compass point toward north on map
Align base plate along north-south grid
0
N
20
40
NE
60
80
E
100
120
SE
140
160
S
180
200
220
SW
240
260
W
280
300
NW
320
340

Step 2: Place compass on map and align base plate

3. Turn the map together with the compass until the north-seeking end of the compass needle coincides with the north end of the orienting arrow. Keep the compass level during this operation to ensure that the compass needle can rotate freely. The map is now oriented.

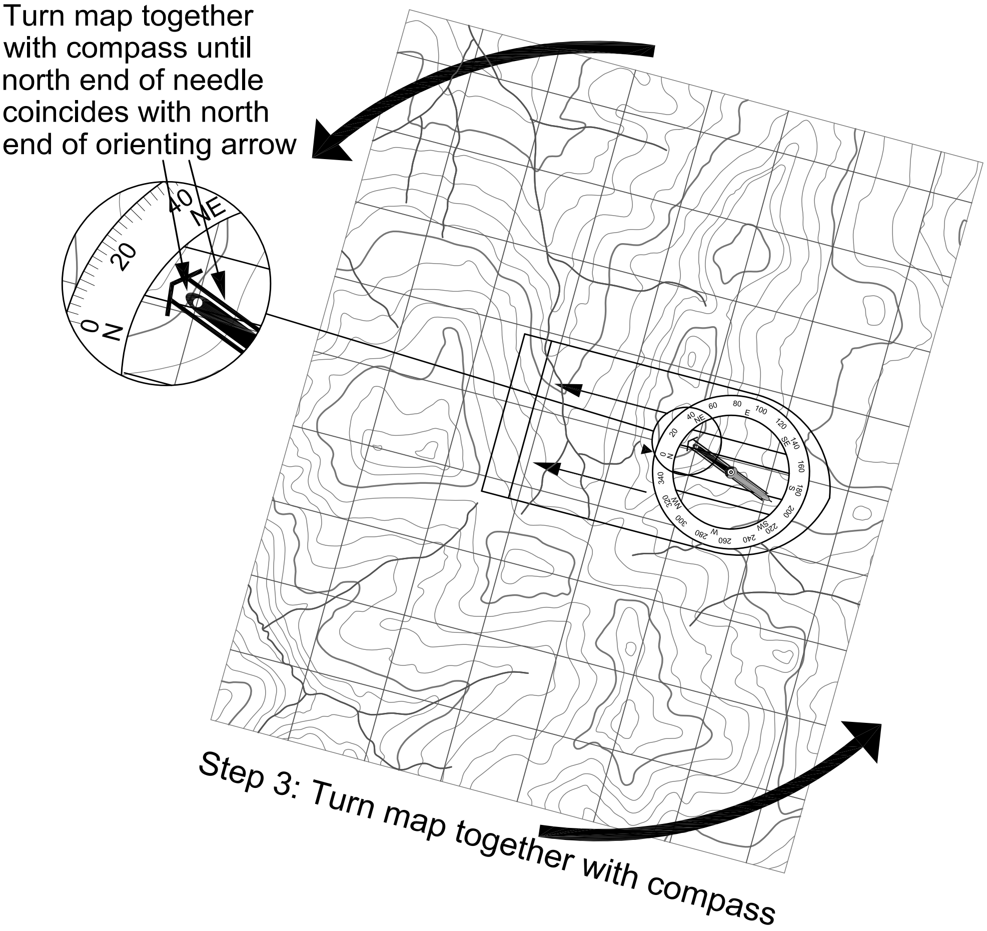
Turn map together with compass until north end of needle coincides with north end of orienting arrow
Step 3: Turn map together with compass

2.8.3 With orienteering compass and declination diagram

This method can only be used if there is a declination diagram on your map. It is generally less accurate than using the map's grid lines because it does not compensate for the declination change since the map was created. Another potentially serious error can occur if the declination diagram isn't drawn to angular scale.

1. Place your compass on the map with the center of the needle near the origin of the declination diagram. The orientation of the compass or the adjustment on the declination mechanism doesn't matter.

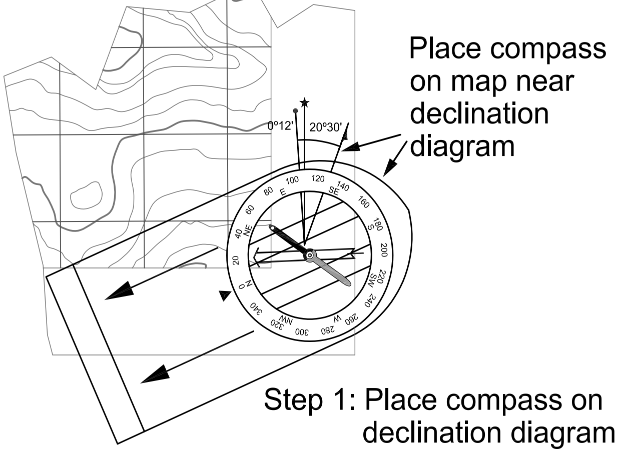
Place compass on map near declination diagram
0°12'
20°30'
Step 1: Place compass on declination diagram

2. Turn the map together with the compass until magnetic north on the declination diagram points in the same direction as the north-seeking end of the compass needle. Make sure the compass needle is free to rotate during this operation. The map is now approximately oriented.

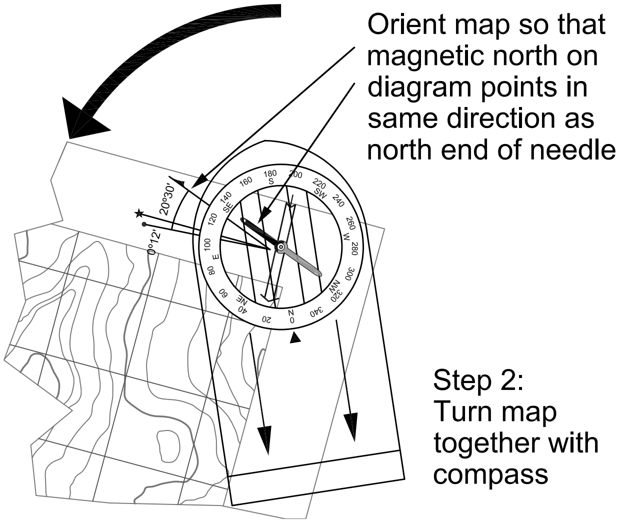
Orient map so that magnetic north on diagram points in same direction as north end of needle
20°30'
0°12'
Step 2:
Turn map together with compass

2.8.4 With marine compass and compass rose

You cannot place a marine compass on a chart because the compass is usually attached to the boat. This makes it more difficult to orient a chart with a marine compass than with an orienteering compass. You are forced to place the chart beside the compass, which is a less accurate process.

1. Bring the chart as close as possible to the marine compass. Line up the center of a compass rose with both the center of the compass and the bearing marker.

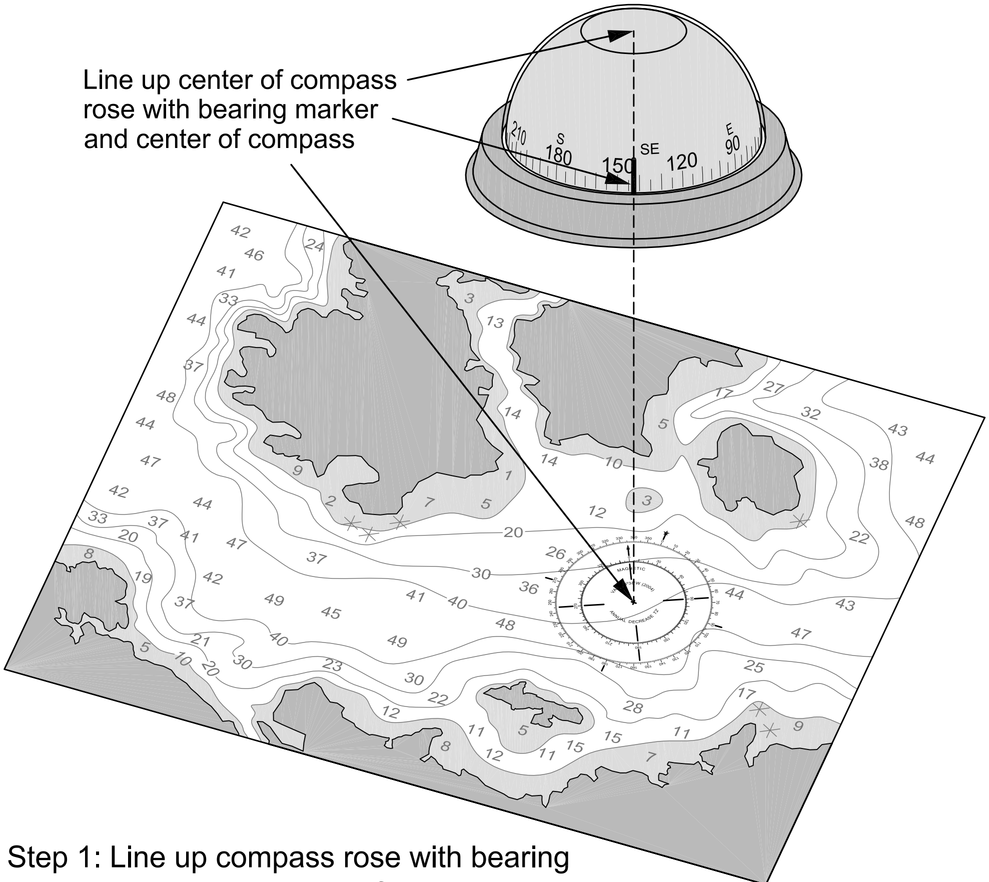
Line up center of compass rose with bearing marker and center of compass
210
S
180
150
SE
120
E
90

Step 1: Line up compass rose with bearing marker and center of compass

2. Read the bearing off the bearing marker; for example 143º magnetic. Using the center of the compass rose as a pivot point, turn the chart until 143º on the magnetic ring of the compass rose points toward 143º on the compass dial. Make sure the center of the rose, the bearing marker, and the center of the compass are still lined up. The chart is now oriented.

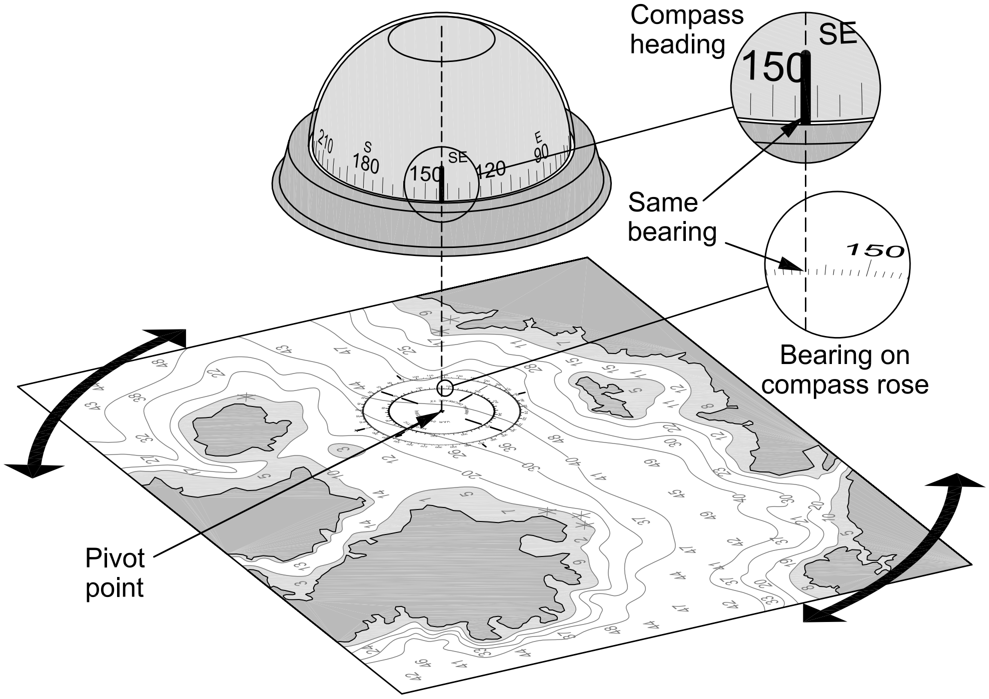
Compass heading
SE
150
Same bearing
150
Bearing on compass rose
Pivot point

Step 2: Turn chart until compass heading lines up with same bearing on compass rose

Alternatively, you could orient your chart so that magnetic north on the chart points toward the bow of your boat, and then turn the boat until the bow points toward magnetic north.

Chapter 3
Altimeter

3.1 How altimeters work

The air pressure in Earth's atmosphere is proportional to the weight of the overlying air. The higher your go, the less air remains above you, and therefore the lower the air pressure. In mathematical terms, the air pressure decreases logarithmically with increasing elevation. This relationship is called the pressure gradient of the atmosphere. An altimeter measures the ambient air pressure, and then calculates its elevation based on a predetermined pressure gradient. Two basic types of altimeters exist, mechanical and electronic.

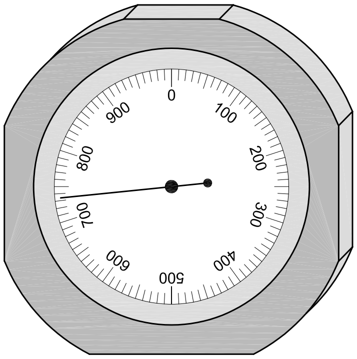

Mechanical altimeter

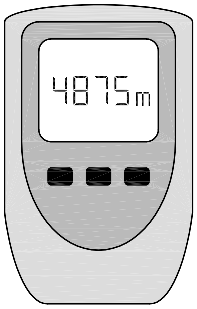

Digital altimeter

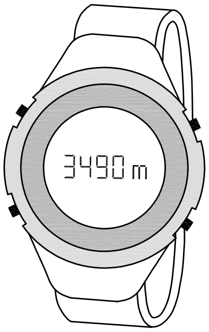

Wristwatch altimeter

A mechanical altimeter consists of a sealed chamber with a flexible wall that is mechanically linked to a rotating needle. The pressure inside the chamber stays more or less constant while the outside air pressure varies according to your altitude and other factors. The ambient air pressure expands or contracts the chamber wall, causing the mechanically linked needle to rotate by an amount proportional to the ambient air pressure. The needle rotates in front of a graduated dial that displays the elevation.

On electronic altimeters, the ambient air pressure expands or contracts a diaphragm that covers a sealed chamber. A built-in piezoresistive sensor deforms together with the diaphragm. The electrical resistance of the piezoresistive sensor changes in proportion to how much it expands or contracts. An electronic circuit measures the resistance across the sensor, and then converts the result into an elevation reading that is shown as a number on a liquid crystal display. Many other types of pressure sensing methods exist for electronic altimeters.

Altimeters are designed to work up to a maximum altitude. Taking the altimeter higher than the maximum allowable altitude may damage the unit. Keep this in mind when traveling in an airliner. Carry your altimeter with you in the pressurized cabin, not in the unpressurized baggage compartment. Generally, altimeters with lower maximum allowable altitudes are more accurate. Electronic altimeters tend to be lighter than mechanical altimeters and usually have the ability to store elevation readings. They are however more prone to failure, especially in cold weather when the batteries can go dead. Mechanical altimeters are rugged and dependable, and do not require batteries.

If you stay at one elevation, altimeters are able to measure or track ambient pressure changes. This can be useful for weather forecasting. For example, a rapid decrease in air pressure may be a precursor to an approaching storm.

Altimeters are easy to use. Just look at the display to see your elevation. Before reading the elevation of a mechanical altimeter, tap the front of the unit a couple of times to make sure the needle is not stuck in one position. Some electronic altimeters update their readings at periodic time intervals. During a rapid ascent or descent, you may have to stop and wait a few moments for the display to show the updated elevation.

3.2 Calibrating an altimeter

Calibrating an altimeter consists of manually adjusting the elevation reading so that it conforms to your actual elevation. Both mechanical and electronic altimeters are equipped with a calibration mechanism that lets you manually adjust the elevation reading.

To calibrate an altimeter, you must be at a known elevation such as sea level or a surveyed height. You can also adjust your altimeter to a GPS elevation reading. A generally less accurate method is to pinpoint your horizontal location on a topographic map and then adjust your altimeter to the elevation of the nearest contour line. The easiest way to pinpoint your horizontal location on the map is of course with a GPS receiver.

Many unpredictable factors affect an altimeter’s elevation reading. To keep your altimeter as accurate as possible, calibrate your altimeter whenever you come across a point of known elevation.

3.3 Factors affecting altimeter accuracy

An altimeter measures the ambient air pressure and then calculates its elevation according to a standard pressure gradient, which is in turn based on average atmospheric conditions. The conditions in the real atmosphere are however rarely average. Air temperature, weather systems, and humidity can alter the pressure gradient and consequently your elevation reading. The temperature of the altimeter itself can also distort the elevation reading.

Air temperature

As air cools, it becomes denser and gravity pulls it closer to the ground. If you are above or inside a layer of cold air, less air remains above you than if the air temperature was normal. The air pressure would therefore be lower than normal for your elevation. The reverse happens if you are above or inside a layer of hot air. More air remains overhead, leading to higher than normal air pressure. There is however no change in the air pressure below a layer of hot or cold air. For example, the air pressure at sea level is not affected by air temperature because all the air is overhead at sea level, regardless of its temperature. As a result, the air pressure changes more rapidly with elevation in cold air than in hot air. In other words, the pressure gradient is steeper in cold air than in hot air.

To calculate its elevation, your altimeter uses a pressure gradient that is based on a standard temperature profile of the atmosphere. In cold air, the real pressure gradient is steeper than the gradient used by your altimeter. Your altimeter reading will therefore change more rapidly than your actual elevation when you ascend or descend through a layer of cold air. In hot air, your altimeter reading will change less rapidly than your actual elevation. If you calibrate your altimeter below or inside a mass of cold air, the altimeter will progressively overestimate its elevation as you ascend through the cold air. Conversely, if you calibrate your altimeter above or inside the same mass of cold air, your altimeter will progressively underestimate its elevation as you descend through the cold air. In hot air, your altimeter will underestimate its elevation during the ascent and overestimate it during the descent, assuming it was calibrated in the same manner. See Table 3.3 for examples.

Table 3.3
Example of altimeter readings inside a layer of hot or cold air stretching from sea level up to 2000 m with various calibration elevations

Actual elevation	Calibrated at 0 m		Calibrated at 1000 m		Calibrated at 2000 m	
	Cold air	Hot air	Cold air	Hot air	Cold air	Hot air
0 m	0 m	0 m	-44 m	41 m	-106 m	104 m
1000 m	1044 m	959 m	1000 m	1000 m	938 m	1063 m
2000 m	2106 m	1896 m	2062 m	1937 m	2000 m	2000 m

To automatically compensate for air temperature, your altimeter would have to know the average temperature of the underlying air and use that information to select an appropriate pressure gradient. Such altimeters do not exist. The only realistic way to compensate for air temperature would be for the user to select a pressure gradient based on the estimated average temperature of the underlying air. Most altimeters, however, do not have this capability.

Humidity

When air absorbs water vapor, the water molecules displace nitrogen and oxygen molecules, the main components of air. This decreases the average air density because water molecules have lower mass than both nitrogen and oxygen molecules. The pressure gradient is therefore less steep in humid air than in dry air. Humidity has a similar effect on your altimeter reading as air temperature, with humid air equivalent to hot air, and dry air equivalent to cold air. To visualize the effect of humidity, look at Table 3.3 and substitute dry for cold, and humid for hot. Altimeter errors caused by humidity are generally less significant than errors caused by air temperature.

To complicate things, hot air can absorb much more water than cold air. This means that the effects of air temperature and humidity will augment each other in hot, humid conditions, and also in cold, dry conditions. For example, an altimeter calibrated at sea level would almost certainly read too low on a summit in a tropical jungle, and too high on an Antarctic summit.

Weather systems

The elevation readings of both mechanical and electronic altimeters are affected by high and low-pressure weather systems that cause the sea-level air pressure to change. At higher elevations, the air pressure usually rises and falls together with the sea-level air pressure of a nearby location. If you stay at one location while a low-pressure system moves in, your altimeter will show a rising elevation even though you haven’t moved. Conversely, if a high-pressure system builds up while you’re stationary, the elevation reading on your altimeter will drop. Weather change is a common and significant cause of altimeter inaccuracy.

Altimeter temperature

The temperature of the altimeter itself can affect its reading. A change in the temperature of a mechanical altimeter causes the sealed metal chamber and the linkage to expand or contract, thereby moving the needle without any change in air pressure. With electronic altimeters, a change in temperature deforms the piezoresistive sensor and leads to a distorted elevation reading.

Try keeping the temperature of your altimeter as constant as possible since its last calibration. This may be difficult in practice. If you strap the altimeter to your wrist, a change in the temperature of your skin will heat or cool the altimeter. If you leave the altimeter hanging outside, a change in air temperature or radiated heat will alter the altimeter temperature. A sensible compromise is to carry the altimeter in a layer of clothing between your skin and the outside environment. Altimeters with built-in temperature compensation systems will partially mitigate the effect of altimeter temperature.

3.4 Triangulating with altimeter and compass

On reasonably steep terrain, you can triangulate your position with one elevation reading and one field bearing off a landmark. This process is similar to triangulating along a linear feature, but with a contour line substituting for the linear feature. You need a topographic map, a compass, and an altimeter.

1. Take a field bearing off an identifiable landmark. Be sure to compensate for the declination relative to grid north.
2. Place the compass on the map with a corner of the base plate on the landmark.

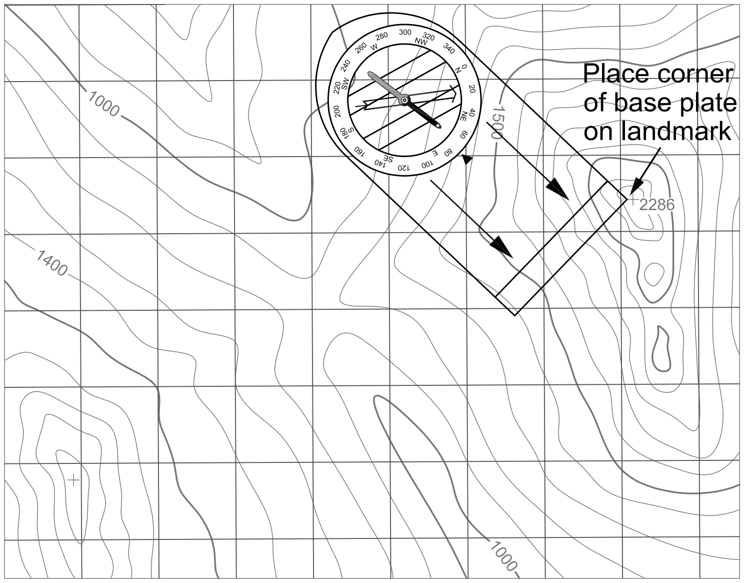
Place corner
of base plate
on landmark
1000
1400
1500
2286
1000

Step 2: Place corner of compass on landmark

3. Using the corner of the base plate as a pivot point, rotate the whole compass until the meridian lines on the compass are parallel to the north-south grid lines on the map. Make sure the compass is oriented so that north (0°) on the graduated dial points toward north on the map, not south.
4. Draw a straight line across the map, parallel to the compass base plate, and through the landmark. This is your first line of position.

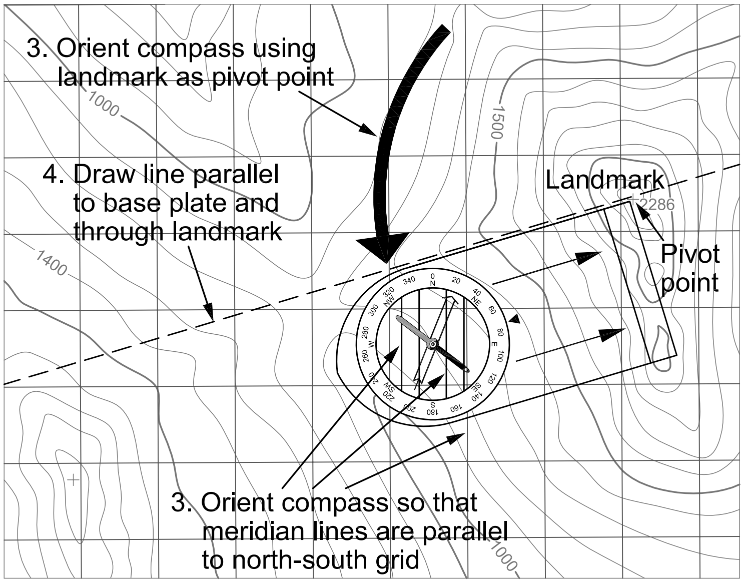
3. Orient compass using
landmark as pivot point
4. Draw line parallel
to base plate and
through landmark
Landmark
2286
Pivot
point
1000
1400
1500
3. Orient compass so that
meridian lines are parallel
to north-south grid
1000

Step 3 - 4: Orient compass and draw line

5. Note the elevation given by your altimeter. Your second line of position is a contour line near your presumed location that matches your altimeter reading.
6. You are located at the intersection between the two lines of position, or in other words, where the compass line of position intersects a contour line that matches your elevation. If there is more than one possible intersection point, investigate the surrounding terrain to narrow your choices down to one. In this case, you have two possible locations, A or B. You can use several methods to narrow the choices down to one. For example, check the slope in the direction of the landmark. If the initial part of the slope is downward, you are at location A, if it's upward you are at location B. Alternatively, use terrain association (Section 8.1) to eliminate one of the choices, or take a field bearing off a second landmark.

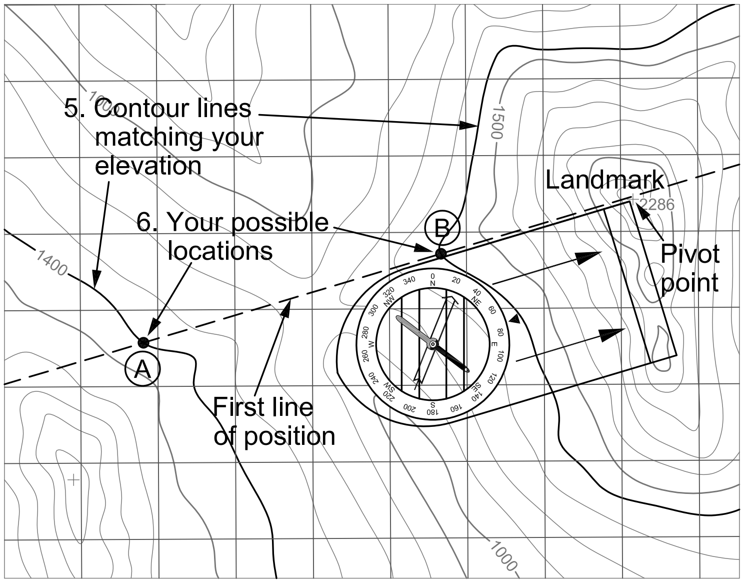
5. Contour lines matching your elevation
1500
1400
1000
Landmark
2286
6. Your possible locations
B
A
Pivot point
First line of position

Step 5 - 6: Find contour lines matching your elevation

3.5 Estimating your location by steepest slope

This method is more difficult and less accurate than taking a field bearing off a landmark, and only works in certain types of topography. You need a topographic map, a compass, and an altimeter.

1. Note the elevation given by your altimeter. Your first line of position is a contour line near your presumed location that matches your elevation.
2. Take a field bearing in the direction of the steepest slope (fall line) at your location. Use the same method as for taking a field bearing off a landmark, but point the compass down the fall line instead of toward a landmark. The fall line is the path that a ball would follow if it rolled down the slope. Be sure to compensate for the declination relative to grid north.

3. Place the compass somewhere on the contour line that matches the elevation reading from your altimeter.
4. Orient the compass so that the meridian lines on the compass are parallel with the north-south grid lines on the map. Make sure that the compass is oriented so that north (0°) on the graduated dial points toward north on the map, not south.

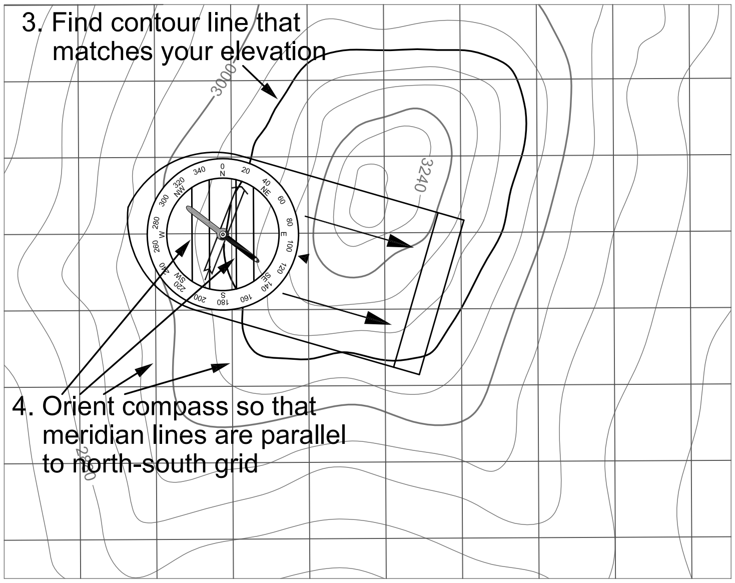
3. Find contour line that matches your elevation
3000
3240
4. Orient compass so that meridian lines are parallel to north-south grid
2880

Step 3 - 4: Place compass on contour line and orient compass

5. Without changing its orientation, move the compass along the contour line until the travel direction arrows point down the fall line. One edge of the base plate should be perpendicular to a series of adjacent contour lines. Double check that the meridian lines on the compass are still parallel with the north-south grid. Your location is where the edge of the base plate intersects with the contour line that matches your altimeter reading.

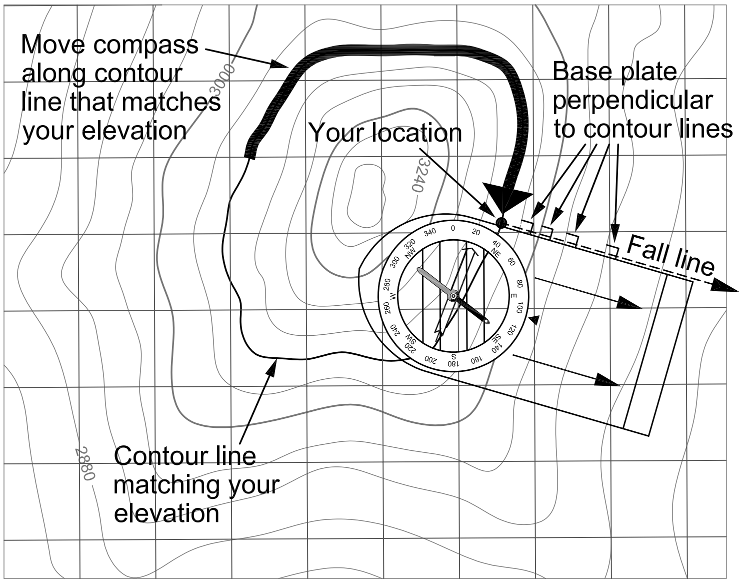
Move compass along contour line that matches your elevation
3000
Your location
3240
Base plate perpendicular to contour lines
Fall line
Contour line matching your elevation
2880

Step 5: Move compass along contour line

Estimating your position by steepest slope has serious limitations, and is usually done only in poor visibility where regular triangulation will not work. It would, for example, be impossible to pinpoint your position if the contour lines were straight near your location. You could be anywhere along the straight contour lines. Even with curved contour lines, you may find several points where the base plate is perpendicular to the contour lines.

If you are unable to pinpoint your location with this method, consider either moving to a location where the topography is more favorable, or waiting for the weather to clear.

Chapter 4
Global Positioning System

A global positioning system (GPS) consists of GPS satellites, associated ground monitoring stations, and individual GPS receivers. By triangulating radio signals transmitted from several GPS satellites, a GPS receiver can automatically compute and display its horizontal position, elevation, direction of travel, speed, and other related information. Several global positioning systems have been, or are being, deployed. The most widely-used one is described here.

4.1 How GPS works

A constellation of GPS satellites has been launched into orbit around Earth. The number varies, but there are normally at least 24 satellites in operation. In the standard configuration, the satellites are arranged into six equally-spaced orbits with four equidistant satellites in each orbit. All orbits are about 20 000 km above Earth's surface and tilted 55° to the equatorial plane. This setup ensures that at least six GPS satellites are overhead anywhere on Earth at any one time. Each satellite completes one orbit in about 12 hours. Ground stations constantly monitor the exact position, speed, and direction of each satellite. This orbit data together with signal timing information is relayed through each satellite to your GPS receiver via a carrier frequency.

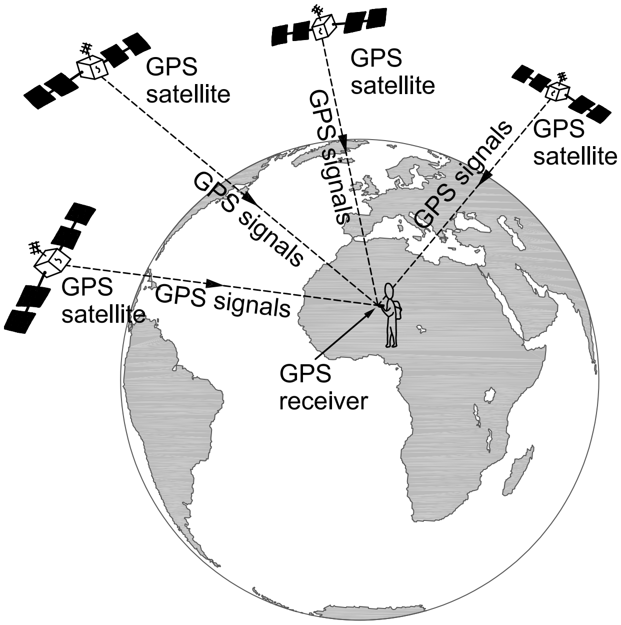

Global positioning system (GPS)

In addition to the orbit and timing data, each satellite transmits its own unique signature signal toward Earth. Signature signals are transmitted from all GPS satellites at frequent intervals and precise times measured by onboard atomic clocks. By generating equivalent signals in its internal circuit, your GPS receiver can recognize an external signature signal as coming from a particular satellite.

From the orbit and timing data, the receiver knows the location of all the satellites and the exact time when each signature signal was transmitted. Since the signals travel at the speed of light, a known speed, your receiver can use the time differences between the transmitted and received signals to calculate its distance to each satellite. Once the distances to a sufficient number of satellites are known, the receiver uses standard geometric formulas to triangulate its position.

In theory, your receiver needs to know its distance to at least three satellites to triangulate its position in three dimensions. In practice, there is a complicating factor. Your GPS receiver is equipped with a regular quartz clock, not an atomic clock, and cannot directly measure the arrival times of each signature signal with sufficient accuracy. The receiver is however capable of accurately measuring the phase shift between a satellite signal and the corresponding internally generated signal. The phase shift is proportional to the signal travel time from the satellite plus an unknown time offset. This offset is the same for all satellite signals and adds one more unknown term to the geometric formulas needed to calculate the position. The signature signals from one extra satellite will solve this unknown term. Your GPS receiver therefore needs signals from at least four satellites to calculate its position in three dimensions (latitude, longitude, and elevation).

A receiver recalculates its position at regular intervals. A moving receiver that is switched on can compare its current position with previous positions and track any changes. This enables the receiver to compute its speed and direction of travel.

4.2 Causes of GPS position errors

The global positioning system is afflicted by many technical problems and uncontrollable natural phenomena that affect the accuracy of a GPS position fix. A worst-case scenario would be if your GPS receiver broke down or the whole GPS system stopped working. Some less severe but common causes of position error are discussed below.

Atmospheric conditions

Before reaching your GPS receiver, the satellite signals are slowed by charged particles in the ionosphere, the upper portion of Earth’s atmosphere. The ensuing time delay is proportional to the number of charged particles that the signal encounters. Air molecules and water vapor also reduce the speed of the satellite signals, mostly in the troposphere, the lower portion of Earth’s atmosphere. Your receiver compensates for the signal delay based on a standard model of the atmosphere. The position error will be larger whenever the conditions in the real atmosphere differ from the standard model. For example, the position error will increase during a solar storm because of the dramatic change in the number of charged particles in the ionosphere. The position error is typically a few meters due to conditions in the ionosphere and up to 1 meter because of conditions in the troposphere.

Multipath reflection

Satellite signals may bounce off features such as rock faces before reaching your receiver. The bounced signal can either be mistaken for, or interfere, with the direct signal, introducing something called a multipath reflection error. Because the bounced signal arrives slightly later than the direct signal, some GPS receivers can tell the signals apart and correct the error. The position error caused by multipath reflection is normally less than 1 meter.

Satellite clock errors

Even though the atomic clocks in the satellites are extremely accurate, they are not perfect. Time discrepancies in the atomic clocks can translate into a position error of up to 1 meter.

Satellite orbit errors

Ground stations regularly track the orbital positions of the GPS satellites. If a satellite drifts from its intended orbit, the ground stations detect the error and then update the satellite’s orbit data. The position error could be a few meters during the time interval before your receiver is informed about the updated orbit.

Malfunctioning satellites

Occasionally, a GPS satellite may malfunction and transmit inaccurate data. The ground stations normally detect this type of problem within a few minutes and promptly transmit a system status message to all the GPS satellites, which in turn tell your GPS receiver to disregard the signals from the broken satellite. A damaged satellite could cause a large position error before your receiver is informed of its condition.

Deliberate signal degradation

The people who operate the global positioning system are able to deliberately degrade the timing of the signature signals to selected parts of the world. A less accurate position reading in that region would result. The position error can be made arbitrarily large.

Satellite distribution

Because of the inherent constraints of geometry, the distribution of the visible satellites affects the accuracy of your position. Ideally, all the visible satellites should be evenly spaced across the sky. In practice, the satellites are never perfectly distributed. The position accuracy degrades in proportion to how close the visible satellites are clustered together.

Signal obstructions

Obstructions such as trees or adjoining terrain that block the satellite signals are the most common cause of position errors or inability to get a position fix. The greater the number of satellite signals that reach your receiver and the stronger the satellite signals, the better the accuracy. Most receivers attempt to calculate a two-dimensional position fix after obtaining signals from just three satellites. The receiver does this by assuming that it is at a known elevation, usually sea level or the elevation of its last position fix. If the assumed elevation is significantly wrong, the calculated two-dimensional position could be off by a kilometer or more. Manually inserting the correct elevation into the receiver will alleviate this problem.

4.3 Differential GPS (DGPS)

Because the GPS satellites orbit high above Earth, two GPS receivers that are fairly close together will acquire signals that have traveled through roughly the same part of the atmosphere. Errors in signal travel times due to atmospheric conditions and deliberate signal degradation are about the same for both receivers. Both receivers collect exactly the same atomic clock errors and satellite orbit inaccuracies.

A differential global positioning system (DGPS) consists of an accurately surveyed ground station equipped with a reference GPS receiver and a transmitter. The reference receiver continually monitors the arrival times of the signature signals, and compares the arrival times to what they should be to conform to the receiver's known surveyed position. The differences between the actual and expected arrival times of the signals are broadcast as correction signals via the ground station's transmitter.

In most cases, any suitably equipped GPS receiver within range of a ground station (typically a few 100 km) can use the correction signals to reduce its position error. Because the ground station does not know which satellite signals your GPS receiver is using for its position fix, it sends you the correction signals for all the satellites overhead. Your receiver only exploits the correction signals from the satellites that are used for your position fix. A receiver located within the range of two or more ground stations typically chooses the correction signals from the base station with the best strength-to-noise ratio of the signals.

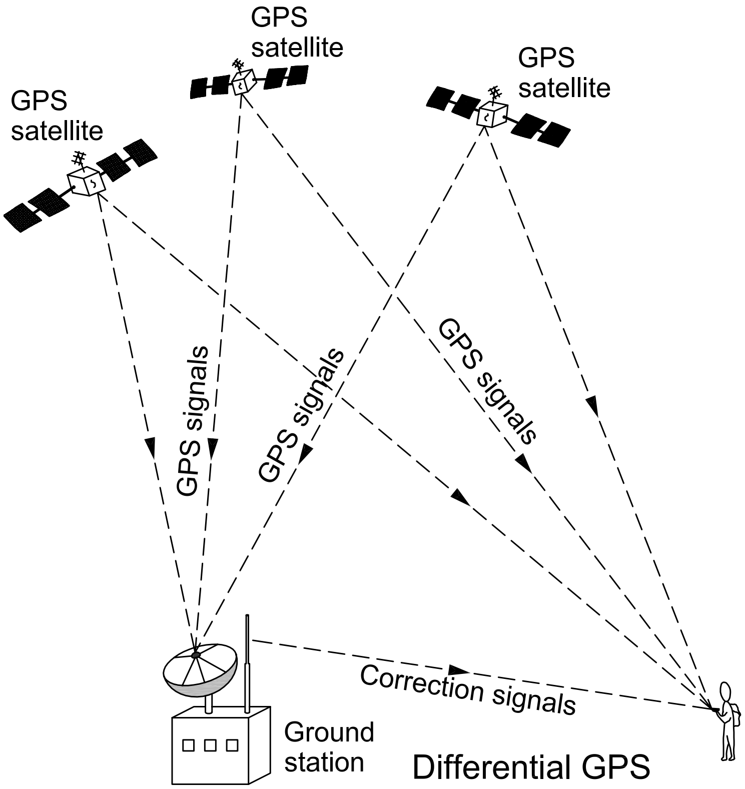

DGPS can virtually eliminate satellite orbit and clock errors and greatly reduce errors caused by atmospheric conditions, deliberate signal degradation, and malfunctioning GPS satellites. The resulting position is reliable and typically accurate within a few meters.

4.4 Space-based augmentation systems

Space-based augmentation systems work on principles similar to DGPS. Correction signals are sent from a network of ground stations to a master ground station that transforms the signals into a grid of correction signals. The grid is sent to one or more geostationary satellites that orbit 36 000 km above the equator, and is then broadcast to Earth. The broadcast area is large, often the size of a continent.

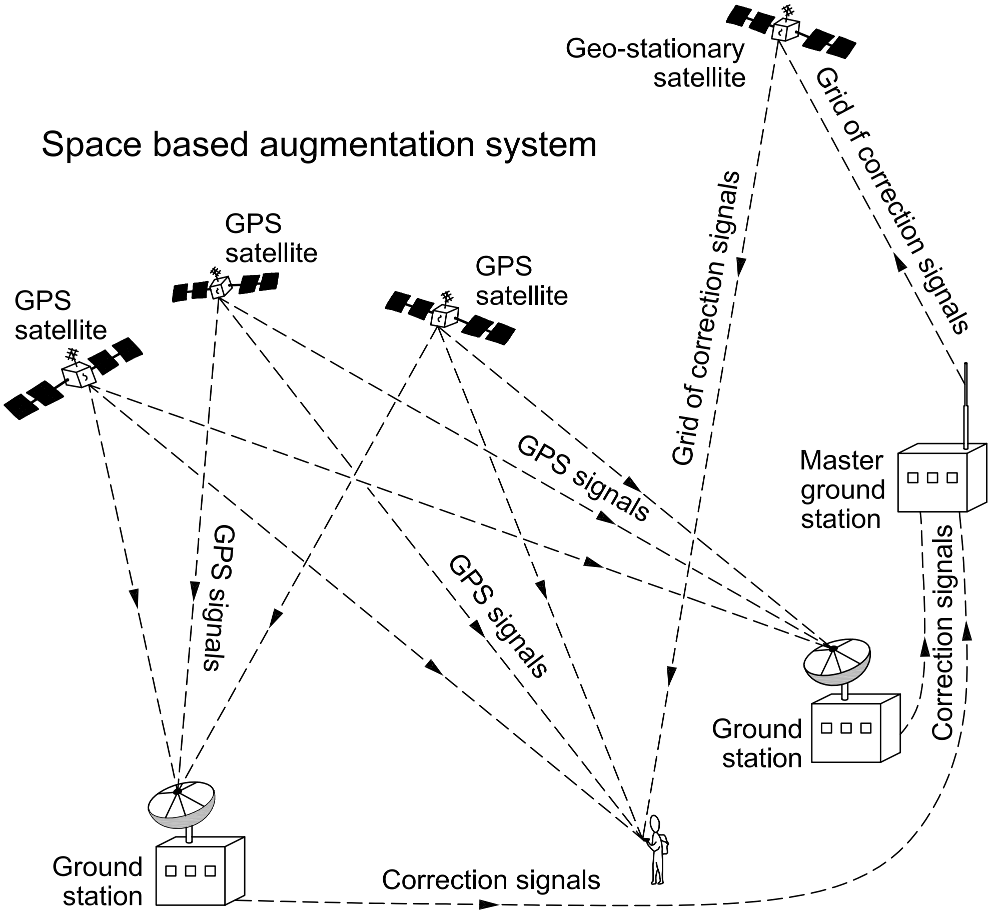
Space based augmentation system
Geo-stationary satellite
GPS satellite
GPS satellite
GPS satellite
Grid of correction signals
Grid of correction signals
GPS signals
GPS signals
GPS signals
Master ground station
Correction signals
Ground station
Ground station
Correction signals

Any suitably equipped receiver within the broadcast area can obtain the complete grid of correction signals. Based on its position within the grid, the receiver interpolates the correction signals from nearby grid points into a set of custom correction signals. The position accuracy is slightly better than for DGPS but can easily be degraded by obstructions that prevent the correction signals from reaching the receiver. Because of their orbits above the equator, the geostationary satellites appear lower on the horizon as you move away from the equator. The grid of correction signals is therefore more likely to be blocked by terrain features at higher latitudes. Pointing your receiver's antenna directly toward the geostationary satellite may improve the position accuracy. The space-based augmentation system itself can of course malfunction and transmit an incorrect correction grid.

4.5 GPS receivers

GPS technology is continually evolving with a large variety of GPS receivers being developed for many types of applications. A GPS receiver can be combined with other electronic devices into one package. Units are available that combine a receiver with an electronic compass, pressure altimeter, wireless phone, or emergency transmitter. Of most interest to wilderness navigators, are lightweight units that can quickly acquire a position fix without consuming too much battery power.

4.5.1 Signal gathering techniques

The method that a GPS receiver uses to gather and process satellite signals governs how quickly and accurately it can obtain a position fix.

Sequential receiver

A sequential receiver uses one channel to gather all the data it needs from one satellite, then switches to the next satellite and so on until it has enough information for a position fix. This type of receiver can get stuck on one satellite if it's unable to obtain a complete signal. The unit is cheap and doesn't use much battery power. However, it is often slow getting a position fix and can quickly lose the fix if a satellite disappears from view.

Multiplexing receiver

Multiplexing receivers use one channel to gather partial data from one satellite for a short time interval, then switch to the next satellite for the same time interval, and so on until there is enough information for a position fix. Multiplexing receivers don't get stuck on one satellite and perform somewhat better than sequential receivers.

Parallel channel receiver

Parallel channel receivers use several separate channels that simultaneously gather signals. Each channel gathers signals from one satellite only. Generally, this type of receiver is faster at getting a position fix, less likely to lose one, and the best type for obtaining position readings under tree cover. The added circuitry uses more battery power. Parallel channel receivers have largely made sequential, multiplexing, and hybrid receivers obsolete.

Hybrid receiver

Hybrid receivers have several separate channels, some of which may be dedicated to a single satellite and some multiplexing or sequential. These types of receivers have intermediate performances.

Dual frequency receiver

Dual frequency receivers are mainly used for surveying or military purposes. They make use of a second carrier frequency transmitted by the GPS satellites to reduce errors caused by atmospheric conditions. These errors can be calculated directly because different frequencies are slowed down by different amounts in the atmosphere. Dual frequency receivers are much more accurate than standard receivers but are heavier and more expensive.

4.5.2 Features that can improve accuracy

Under certain conditions, built-in or optional features can enhance the performance of a GPS receiver.

Correction signal enabled receivers

Receivers that are designed to accept correction signals from differential GPS or space-based augmentation systems can improve the accuracy of a position fix. The receiver must be within a certain radius of an operational ground station to derive any benefit, regardless whether the correction signals arrive via a satellite or directly from a ground station. If your receiver is too far from a ground station, the correction signals may not be relevant for your area.

External antennas

Powered external antennas can be connected to some receivers. These types of antennas amplify weak satellite signals and enhance your receiver's ability to get a position fix under tree cover. Extra battery power is needed to operate the antenna.

Position averaging

Some receivers have a built-in position averaging feature that can improve the accuracy of a waypoint, which is a position fix stored in the receiver's memory. For this feature to work, the receiver must be stationary for a period of time before storing the waypoint. A receiver that is switched on will repeatedly recalculate its position, allowing the software to work out the average of all the previously calculated position fixes. The longer the receiver stays switched on, the more position fixes are available for averaging and the more accurate the subsequently stored waypoint.

4.6 Essential applications for GPS receivers

GPS receivers are designed with many built-in features but only a few are essential for wilderness navigation. Any feature that requires the receiver to stay switched on for prolonged periods should be used sparingly as conserving battery power is an important consideration, especially in cold weather.

Determining your position

The ability to pinpoint your position during any type of weather almost anywhere on Earth is the most basic application of a GPS receiver. Just switch on the receiver and wait for a position fix. Your position can be displayed as a grid number, a geographic coordinate, or a point on a built-in digital map. With signals from a minimum of four satellites, the receiver will display your elevation.

Storing a position as a waypoint

A waypoint is simply a named set of coordinates representing a position. Once your receiver has acquired a position fix, you can store the position in the receiver's memory as a waypoint. This is very useful for finding your way back to the position. You could also give the waypoint's coordinates to someone else, who could then find the waypoint with his or her own receiver.

Inserting a waypoint

You can manually key a position into your receiver's memory as a waypoint without ever having set foot at that location. The coordinates for the waypoint could be obtained from a printed map or from someone else who has previously been there. With a computer, you can insert waypoints electronically into a GPS receiver from a digital map.

Navigating toward a waypoint

Finding your way toward a stored waypoint is the essence of GPS navigation. With waypoints stored in its memory, switch on the receiver and let it acquire a position fix. With a few keystrokes, the receiver will display the distance and bearing toward any stored waypoint. There are two primary ways to navigate toward a stored waypoint: (1) using the receiver by itself, or (2) using it together with a compass. By itself, the receiver must be switched on continuously. With a compass, the receiver requires only intermittent use with most of the navigating done by the compass.

Creating a route from waypoints

This involves organizing stored waypoints into an electronic route with the intention of navigating toward each waypoint in sequence, one after the other.

Plotting and following an electronic trail

When plotting an electronic trail, the GPS receiver automatically saves waypoints at regular intervals as you travel, resulting in a sequence of closely spaced waypoints. When following an electronic trail, the receiver automatically directs you toward the next waypoint along the electronic trail. The receiver must obviously stay switched on while plotting or following an electronic trail.

Calculating distances and bearings

GPS receivers can display the straight-line distance and bearing from your current position to any stored waypoint, or between any two stored waypoints. This is useful for planning routes or checking your progress.

Zeroing-in on a waypoint

Zeroing-in on a waypoint requires pushing the receiver to its limit of accuracy. It's mainly used for finding stashes, snow caves, or other objects hidden by terrain or during very poor visibility.

Getting accurate time

The time on the receiver's quartz clock is obtained from the satellite signals and cannot be adjusted manually. The signal from one satellite is sufficient to set the time on your receiver to within one millisecond. This does not necessarily mean that your receiver accurately resets its quartz clock as soon as it gets a signal from one satellite. The clocks on the GPS satellites are set to something called GPS time, which is offset from universal time by several seconds. The offset changes because leap seconds are periodically added to universal time but not to GPS time. The offset is broadcast from the satellites to the receiver once every few minutes. The receiver's clock may be off by several seconds before it has obtained the offset from the satellites. If a receiver has been switched off for a prolonged period, the time shown on the receiver is only as accurate as the internal quartz clock and could be off by several minutes.

4.7 GPS receiver basic use

Before heading into the wilderness, read your receiver's instruction manual and learn how to use some key features. Aside from a GPS receiver, you will need a compass and a map of your area.

Indoor preparation

Try the following while you are indoors:

- Switch on your GPS receiver and immediately disable the satellite signal-acquiring system. Otherwise, the receiver will begin a futile struggle trying to get a position fix indoors.
- Adjust the time offset to show the local time for your area. You need to add or subtract a certain number of hours from the default setting of universal time. The time itself cannot be adjusted because it is acquired from the satellite signals.
- Find out how to switch between magnetic, true, and grid bearings. Select true or grid bearings if your compass has an adjustable declination mechanism. If not, it may be easier to use magnetic bearings.
- Select your preferred units of measure. It is easier to navigate when your GPS receiver is set to the same units as your map.
- Select a horizontal map datum. The horizontal datum in your receiver must correspond to the datum that is marked on your map. If your receiver allows, select a vertical datum that matches the one used by your map.
- Select an appropriate coordinate system, either geographic coordinates or a specific grid. Choose the same grid as the one on your map. If there's no grid on your map, select geographic coordinates.
- Enter waypoints for nearby locations into your receiver. This can be done manually from a paper map or electronically from a digital map (Section 4.8).
- Arrange several stored waypoints into an electronic route. Check the straight-line distances and bearings between adjacent waypoints and the total travel distance of the route. Practice inserting new waypoints into the route as well as removing old waypoints.

Outdoor practice

Go outside and switch on the receiver. Make sure the satellite signal-acquiring system is enabled.

- If the GPS receiver has never been used before or has moved a large distance since its previous position fix, you may need to initialize the receiver for your region. This involves selecting the country or region where the receiver is located. If you don't initialize the receiver, it may take a long time to get the first position fix, or you may not be able to acquire one.

- Observe the receiver's screen as it acquires satellite signals. Some receivers show the relative positions of the overhead satellites and the strengths of the incoming signals. After acquiring a position fix, the receiver may give you an estimated position error. The real error in your position could be much larger because the receiver doesn't know all the factors that affect its calculated position.
- Save the position fix as a waypoint and give it a descriptive name.
- Pinpoint the GPS-acquired position on your map by using the map's grid or geographic coordinates.
- Make the receiver display the distance and bearing from your current position to a nearby, previously stored waypoint. Set the graduated dial of your compass to this bearing and then follow the bearing to the waypoint. If you do not have a compass, use the receiver's direction arrow or its built-in electronic compass to guide you to the waypoint. GPS receivers usually have several schemes for navigating toward a waypoint.
- After reaching the waypoint, use the receiver and compass to navigate back toward your original position. As you approach your original position, use the receiver to zero-in as close as possible to your original saved waypoint (Section 4.9).
- Plot an electronic trail by traveling at a steady pace while the receiver is switched on. The result is usually shown as a curved line on the receiver's display.
- Use your GPS receiver to guide you back along the same electronic route to your starting point. This process requires some practice to master.

If you can do the above, you will know enough to exploit the navigation power of the GPS.

4.8 Entering waypoints into a GPS receiver

You can manually insert waypoints into your GPS receiver from paper maps, or use a computer to insert the waypoints from a digital map.

4.8.1 Manual transfer of waypoints from map to receiver

You will need a GPS receiver and a map with grid lines or geographic coordinate lines drawn across the map. Be careful when entering the numbers into your receiver. Wrong numbers can lead to serious navigation errors.

1. On the map, plot your anticipated route as a series of straight-line segments. Mark the intersections between segments as waypoints.
2. Look for the map datum written somewhere in the margins of your map. Switch on your receiver and adjust its horizontal datum to correspond with the one on the map.
3. Check the margins of your map to determine what grid system it uses. Set your receiver to display the same grid as the one used by your map. If your map doesn't have a grid, use geographic coordinates.

4. Determine the grid numbers (easting and northing) or geographic coordinates (latitude and longitude) of your waypoints and manually key them into your receiver. Label the waypoints in your receiver and arrange them sequentially into a route. Double-check your numbers.

Example of inserting waypoints

After analyzing the contour lines on the map below, you plot a route from your current position (A) via intermediate waypoints (B) through (E) to your destination (F).

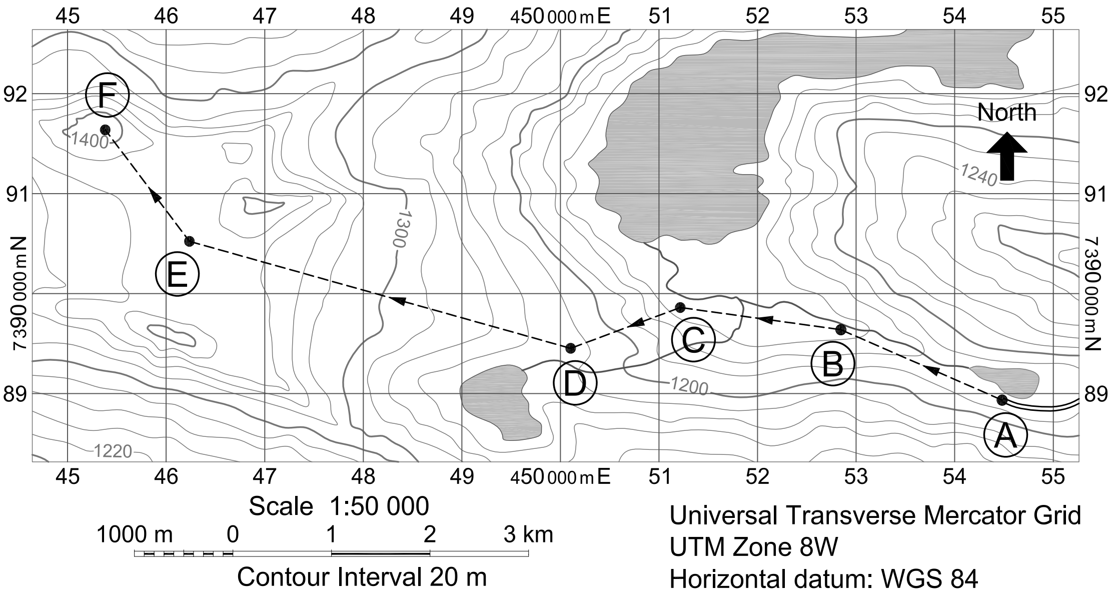
45
46
47
48
49
450 000 m E
51
52
53
54
55
92
91
7 390 000 m N
89
North
F
1400
1300
1240
E
C
B
D
A
1200
1220
Scale 1:50 000
1000 m
0
1
2
3 km
Contour Interval 20 m
Universal Transverse Mercator Grid
UTM Zone 8W
Horizontal datum: WGS 84

Your map shows that it uses the 1984 World Geodetic System datum (WGS 84). Set the horizontal datum in your receiver to WGS 84. The map also shows that it uses the universal transverse Mercator (UTM) grid. Set your receiver to display positions in the UTM/UPS grid system. The UTM zone is given as 8W and applies to all the waypoints on this map.

On your map, each grid square represents 1000 m x 1000 m. You should be able to visually estimate the easting and northing of each waypoint to within one tenth of a square or 100 m. After determining the easting and northing of a waypoint (Section 1.6.3), enter the numbers into your GPS receiver. The numbers corresponding to each waypoint are shown below:

(A): Zone 8W, easting = 454 500 m E, northing = 7 388 900 m N
(B): Zone 8W, easting = 452 800 m E, northing = 7 389 600 m N
(C): Zone 8W, easting = 451 200 m E, northing = 7 389 900 m N
(D): Zone 8W, easting = 450 100 m E, northing = 7 389 500 m N
(E): Zone 8W, easting = 446 200 m E, northing = 7 390 500 m N
(E): Zone 8W, easting = 445 400 m E, northing = 7 391 600 m N

Arrange these waypoints sequentially into an electronic route in your receiver.

4.8.2 Electronic transfer of waypoints into GPS receiver

With digital maps, waypoints can be inserted into the GPS receiver from a computer. Other than a GPS receiver and a computer, you need a suitable software package, relevant digital maps, and an interface cable.

1. Connect your GPS receiver to the computer with the interface cable.
2. Boot up the computer, start the appropriate software, and download the digital maps.
3. Create waypoints with the program by clicking on selected points on the digital maps. Save the waypoints on the computer.
4. Set the map datum on your receiver to the same datum as the one on the digital map.
5. Transfer the waypoints to the GPS receiver. Check your receiver to make sure the waypoints have in fact been transferred.

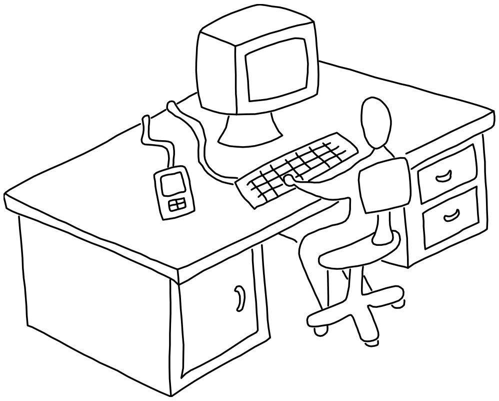

4.9 Zeroing-in on a waypoint

During normal navigation, your GPS receiver will usually provide you with better position accuracy than you actually need. Under certain circumstances such as when trying to find a stash or a campsite during a total whiteout, you may want to precisely pinpoint a previously stored waypoint. To do this, take advantage the GPS receiver's distance measuring system and direction arrow or electronic compass.

Receiver with a built-in electronic compass

A GPS receiver with a built-in electronic compass is ideally suited for zeroing-in on a waypoint.

1. Switch on the receiver, wait for a position fix, and make the receiver show the bearing to the selected waypoint as an arrow. Make sure that the displayed arrow is derived from the electronic compass and isn't a direction arrow. The two types of arrows may look similar.
2. Follow the arrow toward the waypoint while watching the displayed distance decrease. Eventually, the displayed distance will start to increase and the arrow will fluctuate wildly. This is probably as close as your GPS receiver can lead you to the waypoint. Stop and mark your position with an object such as a wand or backpack. Do a grid search if necessary.

Receiver without a built-in compass

Don’t confuse a GPS receiver’s direction arrow with an electronic compass arrow. A direction arrow only works when you’re moving. Without a built-in electronic compass, a GPS receiver has no way of knowing which way it is oriented. A moving receiver can however compute its direction of motion relative to the direction to a waypoint, and display the difference as an arrow. If the receiver is oriented in the direction of motion, the direction arrow will eventually point toward the waypoint.

1. Switch on your receiver, wait for a position fix, and make the receiver show the distance and bearing to the selected waypoint. Display the bearing as a direction arrow.
2. Keep the receiver pointed straight ahead while you move forward in a straight line toward the assumed direction of the waypoint. Observe the direction arrow. At first, the arrow will probably point in the wrong direction. For the moment, ignore the arrow and keep moving forward along the same straight line.
3. As you continue moving, the direction arrow will eventually stabilize and point in the correct direction toward your selected waypoint.
4. While moving, slowly change your travel direction until the direction arrow points straight ahead. Follow the direction arrow toward the waypoint. Keep an eye on the displayed distance. It should decrease as you move toward the waypoint. If not, you are headed in the wrong direction.

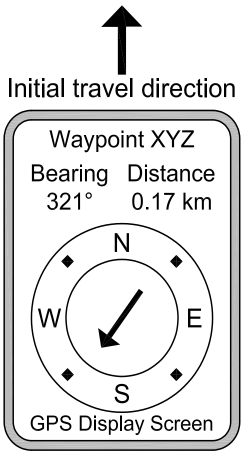
Initial travel direction
Waypoint XYZ
Bearing Distance
321° 0.17 km
N
W
E
S
GPS Display Screen

Step 2:
As you start moving, the direction arrow will probably point in the wrong direction.

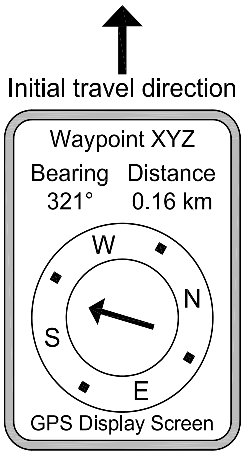
Initial travel direction
Waypoint XYZ
Bearing Distance
321° 0.16 km
W
N
S
E
GPS Display Screen

Step 3:
After moving in the initial direction for some time, the direction arrow will point toward the waypoint.

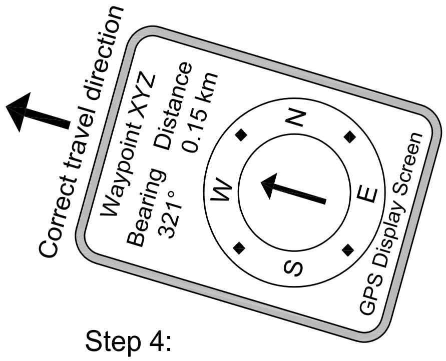
Correct travel direction
Waypoint XYZ
Bearing 321°
Distance 0.15 km
N
W
E
S
GPS Display Screen

Step 4:
Slowly turn until you're moving in the same direction as the arrow. Follow the direction arrow toward waypoint.

5. Eventually, the displayed distance will start to increase and the direction arrow will fluctuate wildly. This is probably as close as your receiver can lead you to the waypoint. Stop and mark this position. If you haven’t found what you are looking for, do a grid search around the marked position to find the exact location.

Chapter 5
Celestial Navigation

From a vantage point on Earth, celestial bodies appear to move across the sky in predictable trajectories. Over the last few thousand years, many techniques have been devised to navigate using the sun, moon, planets, and stars. Celestial navigation is well suited for navigating across oceans, deserts, steppes, tundra, polar ice fields, or any large expanse that lacks obvious landmarks but has horizon-to-horizon visibility. Round-the-clock celestial navigation is possible by switching between celestial bodies that are visible during the day and those that are visible at night. Your navigational potential is governed by what celestial bodies, if any, are visible and what instruments and documentation you have at your disposal:

- With the help of a clock, the sun and moon can be used as rough direction finders.
- With the help of a celestial navigation almanac, you can use the stars to estimate your latitude, and also as rough direction finders.
- With a clock and an almanac, you can estimate your longitude from the stars.
- With the addition of a suitable angle-measuring device such as a sextant, you can use almost any type of celestial body to calculate your position to within a few kilometers.

5.1 Celestial coordinate system

To help pinpoint celestial bodies, a celestial coordinate system has been developed that is based on the same principles as Earth's geographic coordinate system. From Earth's perspective, the celestial bodies appear to be located on the inside surface of an imaginary celestial sphere of indeterminate size. The celestial sphere has two celestial poles located directly above Earth's geographic poles. There is also a celestial equator that is a projection of Earth's equator onto the celestial sphere.

The celestial equivalent to latitude is called the declination (Dec). It is measured in degrees north or south of the celestial equator. "Declination" is an unfortunate term as it can be confused with magnetic declination. The declination is sometimes denoted as a negative number south of the celestial equator.

The celestial equivalent to longitude is called hour angle. Several types of hour angle are used in celestial navigation. The most important one is the Greenwich hour angle (GHA), which is measured in degrees from the Greenwich celestial meridian, the projection of Earth's prime meridian onto the celestial sphere. The GHA is measured westward all the way to 360º so there is no east or west label. Because of Earth's rotation around its axis, the GHA of a celestial body is constantly changing.

The sidereal hour angle (SHA) is a measure of the relative east-west position between bodies on the celestial sphere. It is measured in degrees from a reference meridian called the Aries meridian, the celestial equivalent to Earth's prime meridian. The Aries meridian passes through the spring equinox, the point on the celestial equator that the sun crosses each year around March 21. Like the GHA, the SHA is measured westward all the way to 360º. The SHA is only used when stating the hour angles of stars, and not the sun, moon, or planets. This is because the stars are so incredibly distant from Earth that they don't appear to move relative to each other. The position of the Aries meridian does however change slowly because of the precession and wobble of Earth's axis. There is therefore a gradual change of the SHA for stars from year to year. The local hour angle (LHA) is the angle from the observer's longitude to the celestial meridian of the body. Like other hour angles, it is also measured westward to 360º.

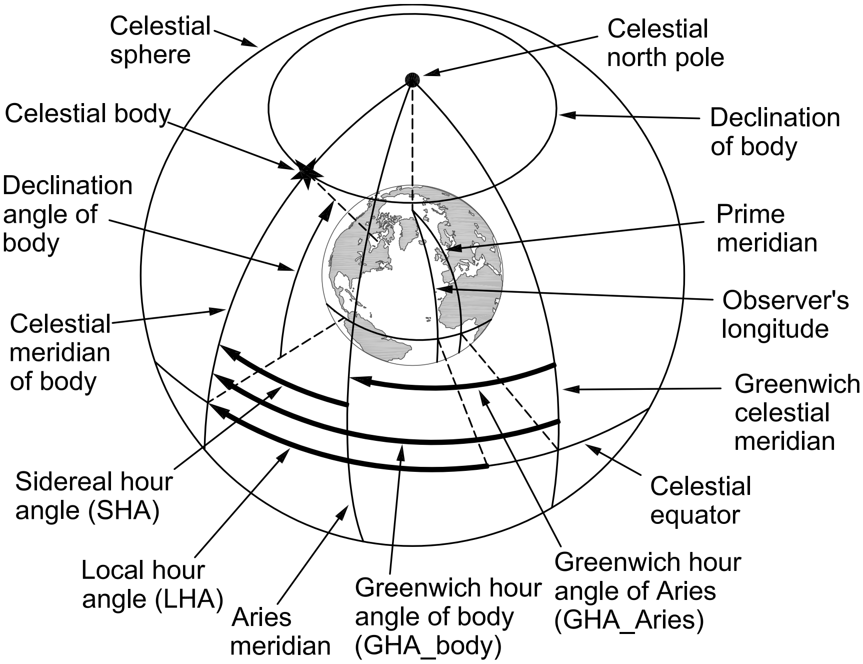
Celestial sphere
Celestial north pole
Celestial body
Declination of body
Declination angle of body
Prime meridian
Observer's longitude
Celestial meridian of body
Greenwich celestial meridian
Celestial equator
Sidereal hour angle (SHA)
Local hour angle (LHA)
Aries meridian
Greenwich hour angle of body (GHA_body)
Greenwich hour angle of Aries (GHA_Aries)

Celestial coordinate system

Celestial navigation almanacs show the declination and hour angles of selected celestial bodies at specific times and dates. The time is always stated as universal time (UT) and the date as universal date (UD). To save space, most almanacs do not directly list the GHA of the stars but instead list the GHA of the Aries meridian and the SHA of the stars. To calculate the GHA of a star, add the GHA of Aries to the SHA of the star. Use the formula: GHA_Aries + SHA = GHA_body

Example of celestial position

You want the celestial coordinates for the star Markab at 18:32:55 (UT) on March 23, 2010 (UD). According to an almanac, the declination for Markab is 15º15.5'N. The Greenwich hour angle for Aries (GHA_Aries) for that time and date is 99º22.2'. The sidereal hour angle (SHA) for Markab is 13º41.0'. The Greenwich hour angle for the body (GHA_body) is therefore 13º41.0' + 99º22.2' = 113º03.2'. The complete position of Markab is written as: Dec 15º15.5'N, GHA 113º03.2'.

5.2 The sun

The sun, being easy to identify and highly visible, is a convenient astronomical body for finding directions. During cloudy weather, the sun's bearing can sometimes be determined by looking for the brightest point in the sky. The sun travels along an arc across the sky that varies with your latitude and the seasons, but is nearly constant from year to year. At mid-latitudes, the sun's bearing can sometimes be estimated with reasonable accuracy by looking at your clock.

5.2.1 Using a clock to find the sun's bearing

From Earth's perspective, the sun completes one circuit across the sky every 24 hours. This works out to a westward movement of 15º per hour. By assuming that the sun is due south at noon in the northern hemisphere (due north in the southern hemisphere), you have a relationship between the sun's bearing and the time of day (Table 5.1). You can then figure out the sun's bearing anytime of the day by looking at your clock and interpolating the table. With a bit of practice you should be able to estimate the sun's bearing without a table.

Table 5.1
Rough bearings of sun at selected times of day

Time of day	Northern hemisphere	Southern hemisphere
00:00	North (0º)	South(180º)
03:00	Northeast (45º)	Southeast (135º)
06:00	East (90º)	East (90º)
09:00	Southeast (135º)	Northeast (45º)
12:00	South (180º)	North (0º)
15:00	Southwest (225º)	Northwest (315º)
18:00	West (270º)	West (270º)
21:00	Northwest (315º)	Southwest (225º)

Most people are unaware of the inherent limitations of this method. The bearings shown in Table 5.1 can be fairly accurate or significantly incorrect. The underlying causes of the errors are the time zone of the observer, Earth's elliptical orbit around the sun, Earth's axis tilt, and the curvature of Earth's surface. To understand how these factors affect the sun's bearing, the errors have been broken down into three themes: (1) time errors caused by the observer's longitude, (2) annual time errors caused by Earth's elliptical orbit and axis tilt, and (3) daily bearing errors caused by the observer's latitude. A thorough analysis of the last two themes requires advanced trigonometry and only a cursory explanation will be given here. The only sure way to obtain an accurate bearing of the sun is to look up the bearing in a celestial navigation almanac or to use a celestial navigation calculator.

Time errors caused by observer’s longitude

Unless the sun is directly overhead, it always points due south or due north when it reaches its highest point in the sky. This happens once a day at the exact moment when the sun crosses your local celestial meridian. To accurately determine the sun’s bearing, your clock must read noon when the sun reaches its highest point, but in practice, most clocks are set to the official time in the time zone where the clock is located. The official time in a zone is loosely based on how far east or west the zone is from the prime meridian. In an ideal world, each time zone would cover a swath of territory 15º of longitude wide, but in the real world, the time zones conform to political boundaries. Official times are a certain number of hours ahead or behind universal time, but are sometimes offset by an additional half hour increment. In some jurisdictions, the official time switches between standard time and daylight saving time. Typically, the official time is changed from standard time to daylight saving time by adding one hour in the spring. In the fall, the official time reverts to standard time by subtracting one hour.

Each time zone has a central longitude where the average position of the sun is due south (northern hemisphere) or due north (southern hemisphere) at noon. To calculate the location of your zone’s central longitude, find out the time difference between your clock time and universal time, and then multiply the difference in hours by 15º. The central longitude is obviously east if you’re in the eastern hemisphere and west if you’re in the western hemisphere. For example, if you’re in the eastern hemisphere and the official time is 9.5 hours from universal time, the central longitude is 9.5 x 15º = 142.5ºE.

If you’re located west of the central longitude, subtract 4 minutes from your clock time for each degree you are from the central longitude. If you’re east of the central longitude, add 4 minutes to your clock time for each degree. For example, if you’re at 133ºE and the central longitude is at 142.5ºE, you are west of the central longitude by 142.5º - 133º = 9.5º. Subtract 38 minutes from your clock time, the result of 4 minutes multiplied by 9.5. Time errors caused by the observer’s longitude become more pronounced at high latitudes where the longitudes converge toward the poles.

Annual time errors caused by Earth’s elliptical orbit and axis tilt

After compensating for your longitude, your clock will probably not read exactly noon when the sun reaches its highest point. Only the average position of the sun over a one-year period will be due north or south at noon. This is why universal time was previously known as Greenwich mean (average) time.

Because of Earth’s elliptical orbit around the sun, Earth’s distance to the sun varies throughout the year. Earth is closest to the sun around January 2 and furthest away around July 3. Earth travels faster along sections where it is closer to the sun and slower when it is further from the sun. This leads to an accumulated time error (Graph 5.2) that is a measure of the sun’s actual position compared to where it would be if it moved uniformly. The error can be as much as 7.6 minutes. For example, if the sun is due south at a particular location at 12:00.0 on January 2, it will be due south at 11:52.4 on April 2 and due south at 12:07.6 on October 2.

From Earth's perspective, the sun appears to travel along an annual arc on the celestial sphere called the ecliptic. Because of the tilt of Earth's axis, the ecliptic is tilted by 23.5º with respect to the equatorial plane. When calculating the annual time error due to axis tilt, you're not interested in how fast the sun moves along the ecliptic but rather how fast it moves across the celestial meridians. One way to visualize this is to look at the movement of the sun's projection onto the equatorial plane.

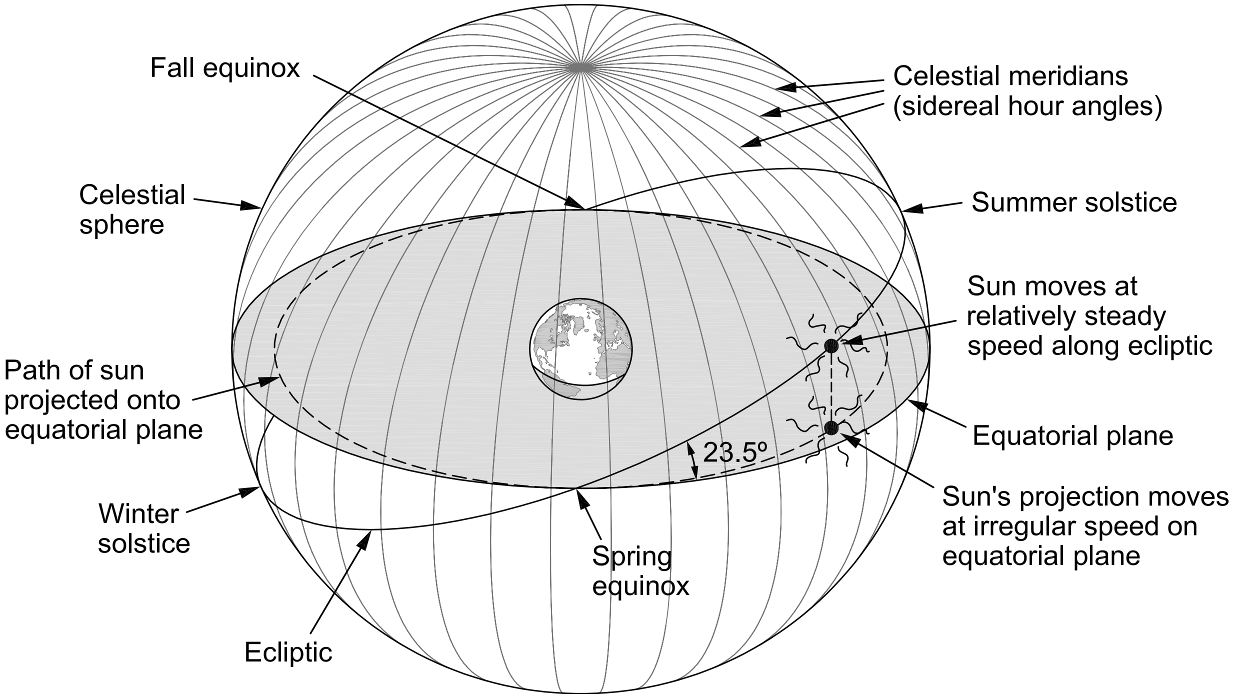
Fall equinox
Celestial meridians (sidereal hour angles)
Celestial sphere
Summer solstice
Sun moves at relatively steady speed along ecliptic
Path of sun projected onto equatorial plane
Equatorial plane
23.5º
Winter solstice
Sun's projection moves at irregular speed on equatorial plane
Spring equinox
Ecliptic

The sun crosses the celestial equator at the spring equinox around March 21, and then begins its journey through parts of the north celestial hemisphere. At the equinox, the sun travels at an angle of 23.5º to the equatorial plane, and only a portion of the sun's movement is projected onto the equatorial plane. At this point, the projected sun moves slower than the actual sun. As the sun moves higher into the north celestial hemisphere, it travels increasingly more parallel to the equatorial plane and the distance between celestial meridians decreases. Both these factors cause the angular speed of the projected sun to increase. At the summer solstice, around June 21, the sun reaches its highest point in the north celestial hemisphere and the projected sun attains its maximum angular speed. After June 21, the projected sun gradually slows down until it crosses into the south celestial hemisphere at the fall equinox around September 22. The cycle is then repeated for the south celestial hemisphere. The irregular angular speed of the sun creates an accumulated time error of up to 9.6 minutes (Graph 5.2). In other words, the axis tilt causes the sun to be due south or due north up to 9.6 minutes before or after noon.

The solid line in Graph 5.2 on the next page shows the combined error caused by both the elliptical orbit and the axis tilt. The combined error varies from 0 to about 16 minutes and is the same for all locations on Earth.

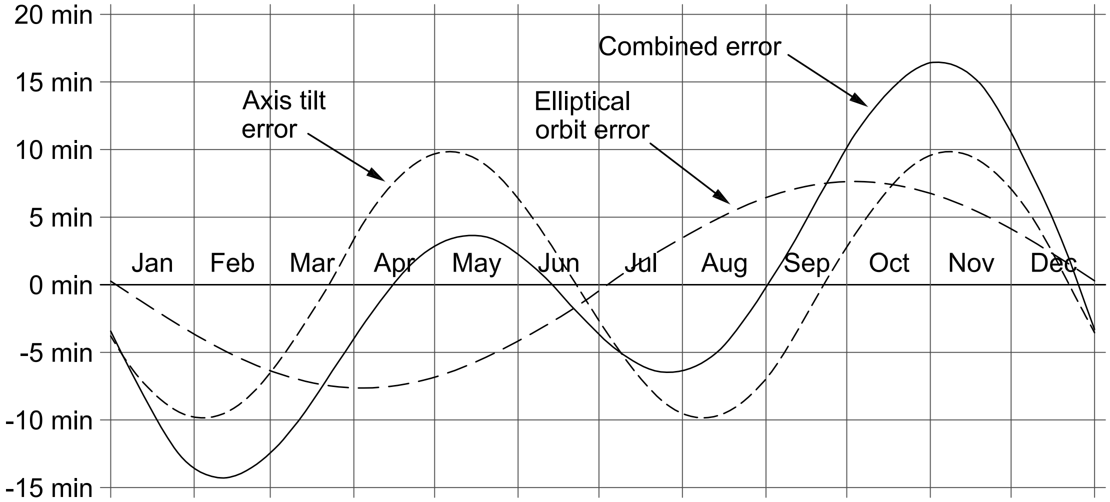

Graph 5.2. Time error due to earth's elliptical orbit and axis tilt

Daily bearing errors caused by observer's latitude

After correcting for your longitude, Earth's elliptical orbit, and Earth's axis tilt, an accurate clock will always read noon when the sun is at its highest point. Away from the tropics, the noon sun will be exactly south in northern hemisphere and exactly north in southern hemisphere. The clock will also read midnight when the sun reaches its lowest point, although this may not be observable. But what about other times of the day? Will the sun be due east at 6:00, for example? In most cases, it will not.

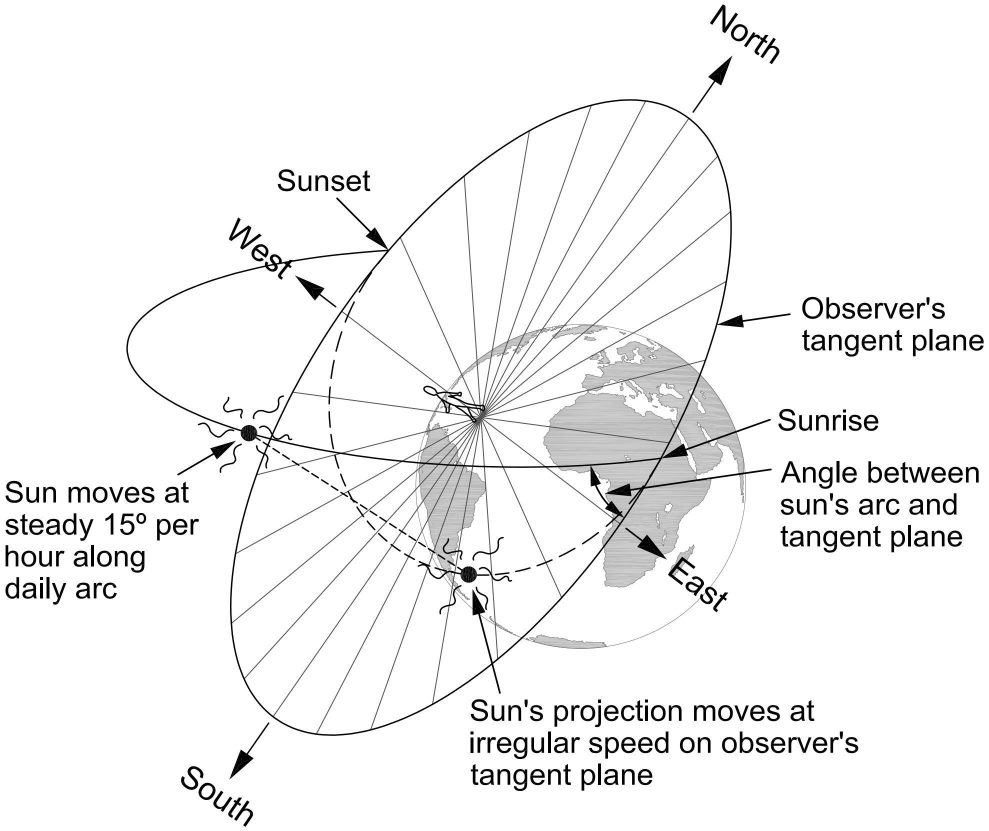
North
Sunset
West
Observer's tangent plane
Sunrise
Angle between sun's arc and tangent plane
Sun moves at steady 15º per hour along daily arc
East
Sun's projection moves at irregular speed on observer's tangent plane
South

When establishing the sun's bearing, you're actually looking for the direction of the sun projected down onto your tangent plane, which is an extension of your horizon to the celestial sphere. The apparent daily arc of the sun is tilted by a certain angle relative to your tangent plane. This angle changes with the seasons and your latitude. Although the sun moves along its arc at close to a steady 15º per hour, the sun's projection on the tangent plane moves at irregular speed. At sunrise, the sun climbs steeply and only a portion of its movement is projected down to the tangent plane. The angular speed of the projected sun is therefore less than 15º per hour. See the illustration on the previous page.

As the sun climbs higher in the sky, it travels increasingly more parallel to the tangent plane and the relative distance between the projected sun and the observer decreases. This causes the angular speed of the projected sun to increase. In the middle of the day, the projected sun travels faster than 15º per hour. The projected sun then gradually slows down until sunset. The speed fluctuations lead to a bearing error that varies throughout the day but is always zero at noon and midnight.

The greater the angle between the sun's arc and the observer's horizon, the greater the potential bearing error. The error therefore depends both on the observer's latitude and the time of the year. Only at the geographic poles where the speed of the projected sun is constant, is there never a bearing error. As you move away from the poles, the speed of the projected sun becomes progressively more irregular and the bearing error more significant. To compensate for the bearing error, you need tables or graphs that show the sun's position at every latitude, for every hour of the day, and for every day of the year. A more practical solution for the wilderness traveler is to carry a celestial navigation calculator.

Sun's bearing - Example 1:

Let's say you are traveling through the Gobi Desert in Mongolia (45ºN, 102ºE), and want to estimate the sun's bearing. It is 16:47 local time on July 18. Mongolia does not use daylight saving time.

Table 5.1 gives the bearing to the sun as roughly west-southwest for this time of the day. Taking into account that the sun moves westward, on average, by one degree every four minutes you can do a more precise interpolation. Your clock is 107 minutes past 15:00, giving you a precisely interpolated but uncorrected bearing of 225º + 107/4 = 252º.

You're in the eastern hemisphere and the official time in Mongolia is 8 hours ahead of universal time. The central longitude for your time zone is therefore 8 x 15º = 120ºE. You're located west of the central longitude by 120º - 102º = 18º. Subtracting 4 minutes from your clock time for each degree that you are from the central longitude gives you a corrected time of 16:47 - 18 x 4 min = 15:35.

For July 18, the combined error from Graph 5.2 gives an additional clock error of about -6 minutes. Your new clock time becomes 15:35 - 6 min = 15:29. Your interpolated bearing from Table 5.1 becomes 225º + 29/4 = 232º. This bearing is corrected for your longitude, Earth's axis tilt, and Earth's elliptical orbit.

The graph below was derived from a computer program. It shows the relationship between the time of the day and the bearing error at latitude 45ºN on July 18. For a time of 15:29, the graph gives a bearing error of about +26º. Adding this to your bearing gives a new bearing of 232º + 26º = 258º.

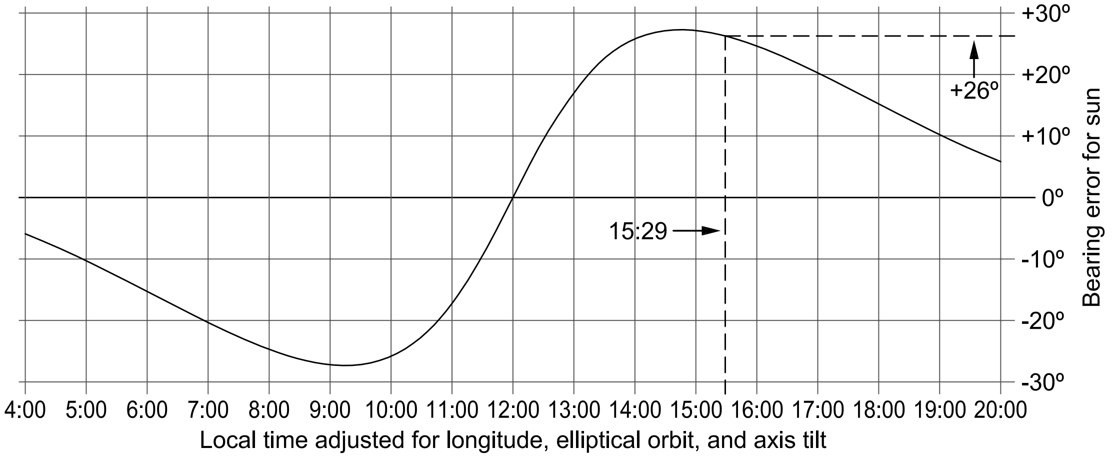

45ºN - Bearing error due to latitude - July 18

Comparing this corrected bearing (258º) with the initial uncorrected bearing (252º) gives a total bearing error of only 6º. You were lucky this time! The sun is indeed roughly west-southwest. The large error caused by your longitude was largely compensated by an opposite error due to your latitude. A few hours earlier or later, your uncorrected bearing would have been off by a significant amount.

Sun's bearing - Example 2:

While trekking through northern Bolivia (12ºS, 67ºW), you want to estimate the sun's bearing. The local time is 10:12 on January 3. Bolivia does not use daylight saving time.

Table 5.1 gives the bearing to the sun as roughly north-northeast. Your clock is 72 minutes past 9:00, giving you a precisely interpolated but uncorrected bearing of 45º - 72/4 = 27º.

You're in the western hemisphere and the official time in Bolivia is 4 hours behind universal time. The central longitude of your time zone is therefore 4 x 15º = 60ºW. You're located west of the central longitude by 67º - 60º = 7º. Subtracting 4 minutes from your clock time for each degree that you are from the central longitude gives you a corrected time of 10:12 - 7 x 4 min = 09:44.

For January 3, the combined error from Graph 5.2 gives an additional clock error of about -5 minutes. Your new clock time becomes 09:44 - 5 min = 09:39. Your interpolated bearing from Table 5.1 is then 45º - 39/4 = 35º. This bearing is corrected for your longitude, as well as Earth's axis tilt and elliptical orbit.

The graph on the next page was derived from a computer program. It shows the relationship between the time of the day and the bearing error at latitude 12ºS on January 3. For a time of 09:39, the graph gives a bearing error of about +78º. Adding this to your previous corrected bearing gives a new bearing of 35º + 78º = 113º.

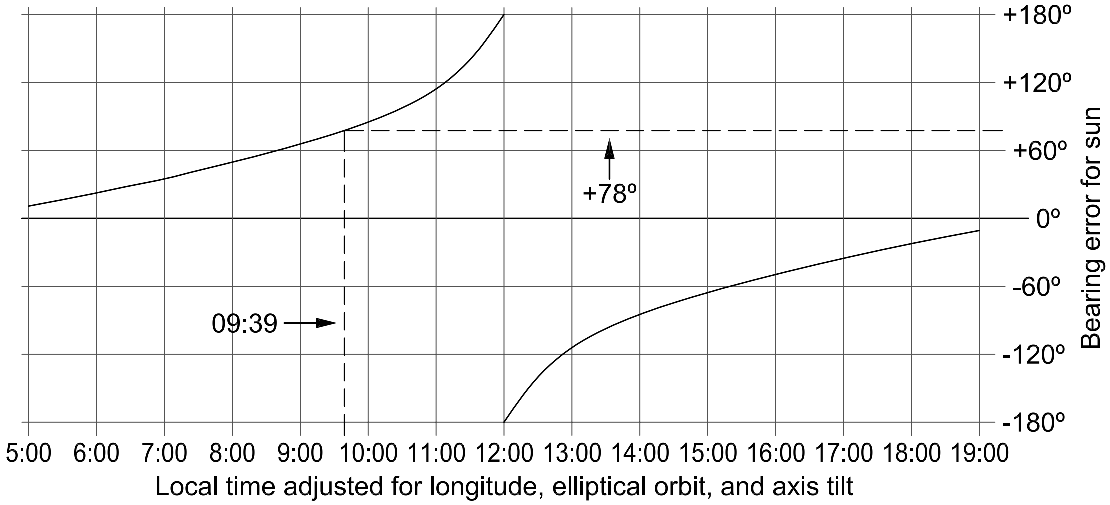

12ºS - Bearing error due to latitude - January 3

Comparing the fully corrected bearing (113º) with the initial uncorrected bearing (27º) gives a total bearing error of an astonishing 86º. In other words, the uncorrected bearing will lead you seriously astray. This example clearly illustrates the difficulty in trying to casually estimate the sun's bearing in the tropics (23.5°S - 23.5°N) where the sun is north at noon for part of the year and south for the rest of the year. A sensible technique near the equator is to assume that the sun is roughly east during the morning and roughly west during the afternoon. If the sun is high in the sky, wait for it to move lower before trying to estimate its bearing. The only precise method, of course, is to use a celestial navigation almanac or calculator.

5.2.2 Stick shadow method

There are several variations of the stick shadow method. The one shown here requires that you to start the procedure in the morning and stay with it at least until noon. During the morning, find a patch of level ground and plant a stick vertically into the ground. Mark the end of the stick's shadow. Use a second stick to scribe a circle on the ground, centered on the vertical stick and with a radius the same length as the shadow.

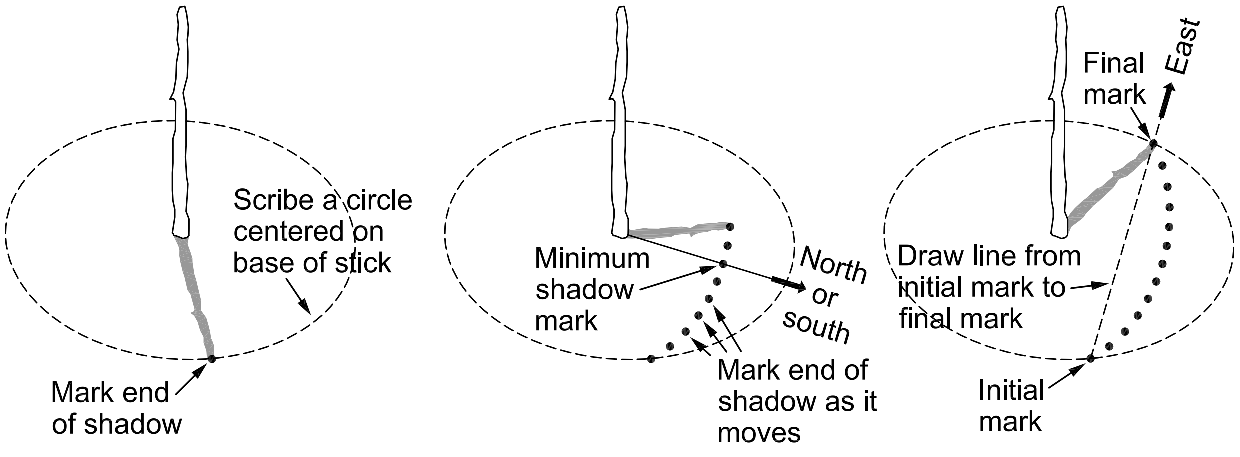
Scribe a circle centered on base of stick
Mark end of shadow
Minimum shadow mark
North or south
Mark end of shadow as it moves
Final mark
East
Draw line from initial mark to final mark
Initial mark

Periodically mark the end of the shadow as it moves. The length of the shadow decreases as the morning progresses until it reaches a minimum at noon and then increases in the afternoon. A line drawn from the base of the stick to the minimum-shadow mark will point north in the northern hemisphere and south in the southern hemisphere. In the tropics, the line will point north for part of the year and south for the rest of the year, depending on your latitude. For a more precise result continue watching and marking the shadow until it crosses the circle that you've scribed on the ground. A line drawn from the initial shadow mark to the final shadow mark will point due east anywhere in the world except at the geographic poles.

The great thing about this method is that you don't have to worry about errors caused by your longitude, latitude, Earth's elliptical orbit, or Earth's axis tilt. All these errors are automatically factored in.

5.3 The moon

The moon's position in the sky at a certain time of day varies daily and is governed by its monthly orbit around Earth. You cannot directly determine the moon's bearing by simply looking at your clock. You can however do it indirectly by observing the moon's phase.

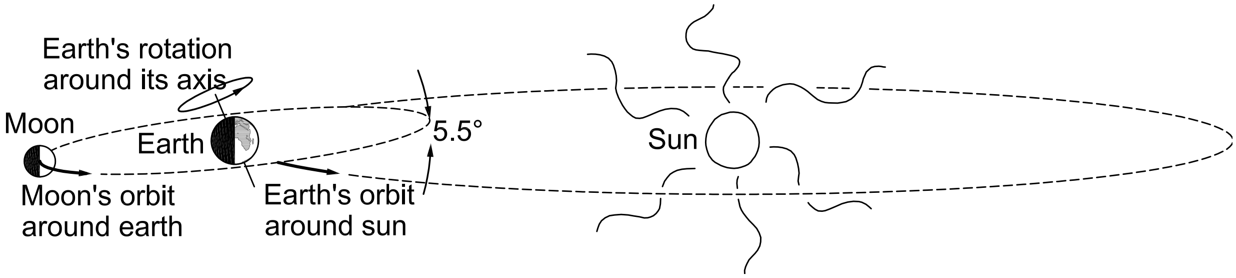
Earth's rotation around its axis
Moon
Earth
5.5°
Sun
Moon's orbit around earth
Earth's orbit around sun

The moon's orbit around Earth is tilted by only about 5.5° relative to Earth's orbit around the sun. For an observer on Earth, this means that the sun and the moon appear to travel along roughly the same arc across the celestial sphere. At any one moment, the sun and moon are a certain angle apart in their perceived mutual arc. This angle can be described as the moon leading or trailing the sun by a certain number of hours.

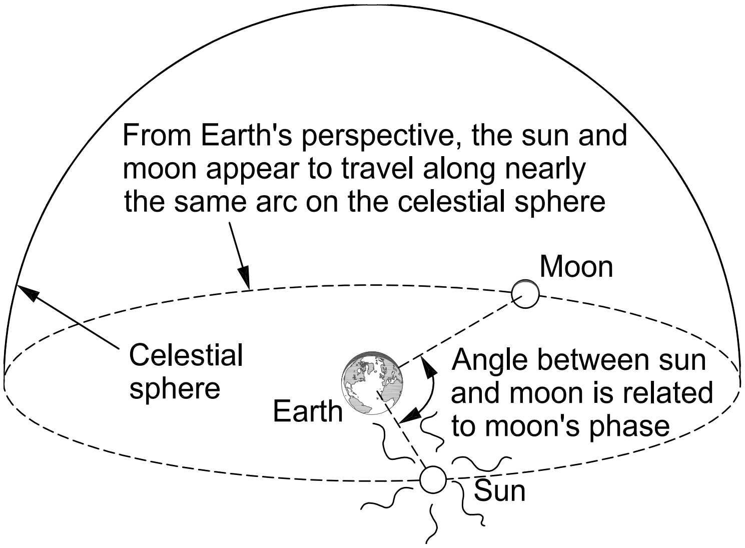
From Earth's perspective, the sun and moon appear to travel along nearly the same arc on the celestial sphere
Moon
Celestial sphere
Earth
Angle between sun and moon is related to moon's phase
Sun

The angle between the sun and the moon is correlated to the moon’s phase. For example, a full moon means that the moon is on the opposite side of the arc, and you can think of the full moon as leading or trailing the sun by 12 hours. Whether the moon is leading or trailing is governed by what side of the moon is bright and what hemisphere you’re in. For example, a half moon in the southern hemisphere with its bright side to the left means that the moon is trailing the sun by 6 hours.

The relationship between the moon’s phase and its relative angle to the sun can be derived from trigonometry. See Table 5.3 for selected results. With a bit of practice, you should be able to just look at the moon to figure out by how many hours it is leading or trailing the sun. Adding or subtracting these hours from your clock time gives you the time when sun will be, or was, where the moon is now. The bearing of the sun at that time is the current bearing of the moon.

Table 5.3
Correlation between moon's phase and relative position to sun

Phase	Northern hemisphere	Southern hemisphere
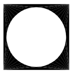	opposite sun (12 hours)	opposite sun (12 hours)
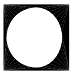	leading sun by 10 hours	trailing sun by 10 hours
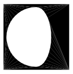	leading sun by 8 hours	trailing sun by 8 hours
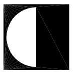	leading sun by 6 hours	trailing sun by 6 hours
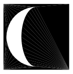	leading sun by 4 hours	trailing sun by 4 hours
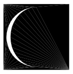	leading sun by 2 hours	trailing sun by 2 hours
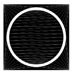	same as sun (0 hours)	same as sun (0 hours)
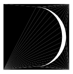	trailing sun by 2 hours	leading sun by 2 hours
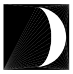	trailing sun by 4 hours	leading sun by 4 hours
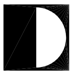	trailing sun by 6 hours	leading sun by 6 hours
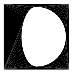	trailing sun by 8 hours	leading sun by 8 hours
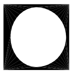	trailing sun by 10 hours	leading sun by 10 hours

5.3.1 Estimating the moon's bearing

The following steps describe how to calculate the moon's bearing. You're essentially using the sun's bearing to figure out the moon's bearing. All the errors that apply to the sun's bearing will therefore carry over to the moon's bearing. There is also the additional error from estimating the moon's phase.

1. Look at the moon and figure out if it is leading or trailing the sun. In the northern hemisphere, the moon is trailing the sun if the bright part of the moon is facing right, and leading if it is facing left. In the southern hemisphere the opposite is true, a right facing bright side means the moon is leading the sun and left facing bright side means it's trailing the sun. In the tropics, the bright side of the moon may be facing up or down rather than left or right, making it difficult to figure out if it is leading or trailing the sun.
2. Determine by approximately how many hours the moon is leading or trailing the sun by looking at the moon's phase (Table 5.3). Interpolate the table for intermediate phases. Subtract the number of hours from your current time if the moon is trailing and add the number of hours if the moon is leading.
3. Determine where the sun's bearing was/will be at this adjusted time. This is the current bearing of the moon.

A more accurate method is to calculate the moon's bearing directly by using a celestial navigation almanac or a calculator. The almanac or calculator gives you the moon's bearing with inputs of your latitude, longitude, local time, local date, and time offset from universal time.

Moon's bearing - Example 1:

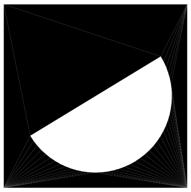

You're traveling in east-central Australia (26ºS, 144ºE), on July 16, 2009. At 04:16 local time the moon looks like the illustration on the left.

A half moon in the southern hemisphere with its bright side facing right means that the moon is leading the sun by 6 hours (Table 5.3). The moon is therefore where the sun will be 6 hours later at 10:16. A rough estimate of the sun's bearing at 10:16 puts it toward the north-northeast (Table 5.1). This is the approximate bearing of the moon. Interpolating Table 5.1 gives a precise but uncorrected bearing of 45º - 76/4 = 26º.

Plugging your latitude (26ºS), longitude (144ºE), local time (04:16), local date (July 16, 2009), and time offset from universal time (+10 hours) into a celestial navigation calculator gives a bearing of 40º. The moon is actually closer to the northeast than north-northeast. Your estimated bearing is off by 14º, good enough for rough direction finding.

Moon’s bearing - Example 2:

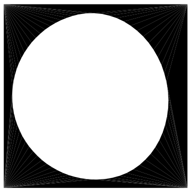

You’re in southern Greenland (62ºN, 45ºW), on March 29, 2005. At 02:54 local time the moon looks like the illustration on the left.

Just a few days earlier, Greenland switched to daylight saving time by adding one hour to the local time. Subtract one hour to convert to standard time (01:54). Being in the northern hemisphere, the moon’s phase shows that it is leading the sun by roughly 10 hours (Table 5.3). The moon is where the sun will be in 10 hours, at 11:54. At that time, the sun’s bearing will be almost exactly south (Table 5.1). Interpolating Table 5.1 gives a precise but uncorrected bearing of 180º - 6/4 = 179º.

Plugging your latitude (62ºS), longitude (45ºW), local time (02:54), local date (March 29, 2005), and time offset from universal time (-2 hours, for daylight saving time) into a celestial navigation calculator gives a bearing of 170º. Your estimated bearing is off by only 9º, a good approximation of the moon’s actual bearing.

5.4 Stars

Because of their great distances from Earth, stars don’t appear to change position relative to each other, and star patterns, called constellations, look almost the same today as they did hundreds of years ago. From our perspective on Earth, each star appears permanently fixed to a particular position on the rotating celestial sphere. At the equator, all stars on the celestial sphere rise and set. At the poles, only half the stars on the celestial sphere are visible but these stars never rise nor set. At intermediate latitudes, some stars rise and set while others are always above or always below the horizon.

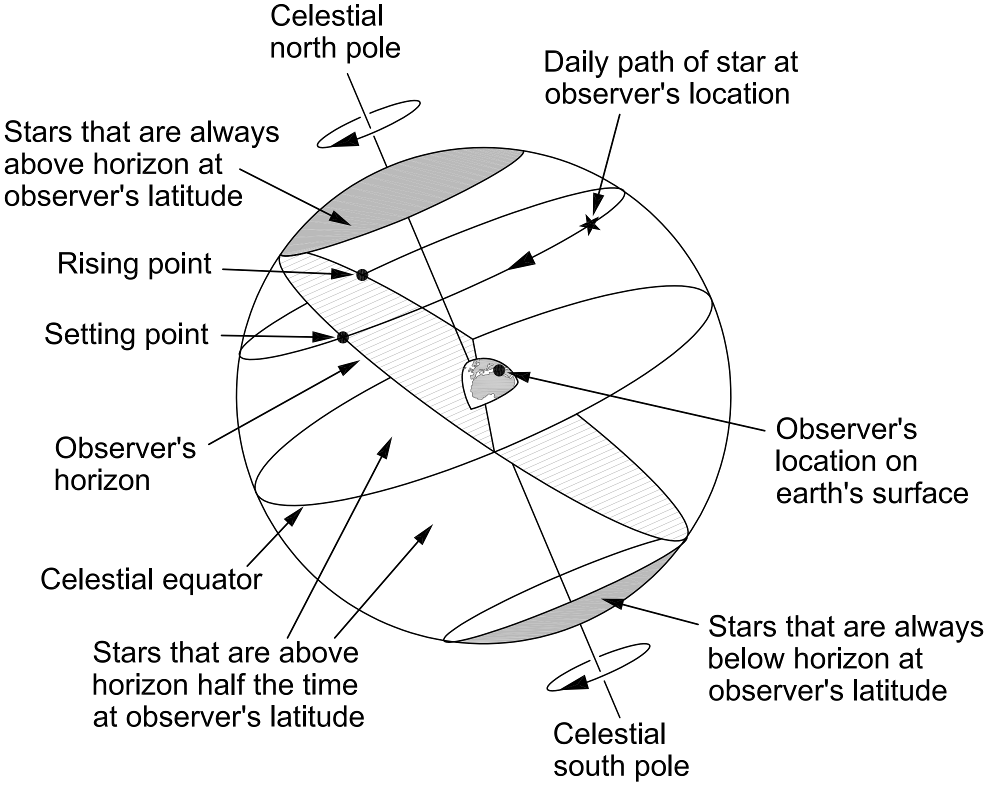
Celestial north pole
Daily path of star at observer's location
Stars that are always above horizon at observer's latitude
Rising point
Setting point
Observer's horizon
Observer's location on earth's surface
Celestial equator
Stars that are always below horizon at observer's latitude
Stars that are above horizon half the time at observer's latitude
Celestial south pole

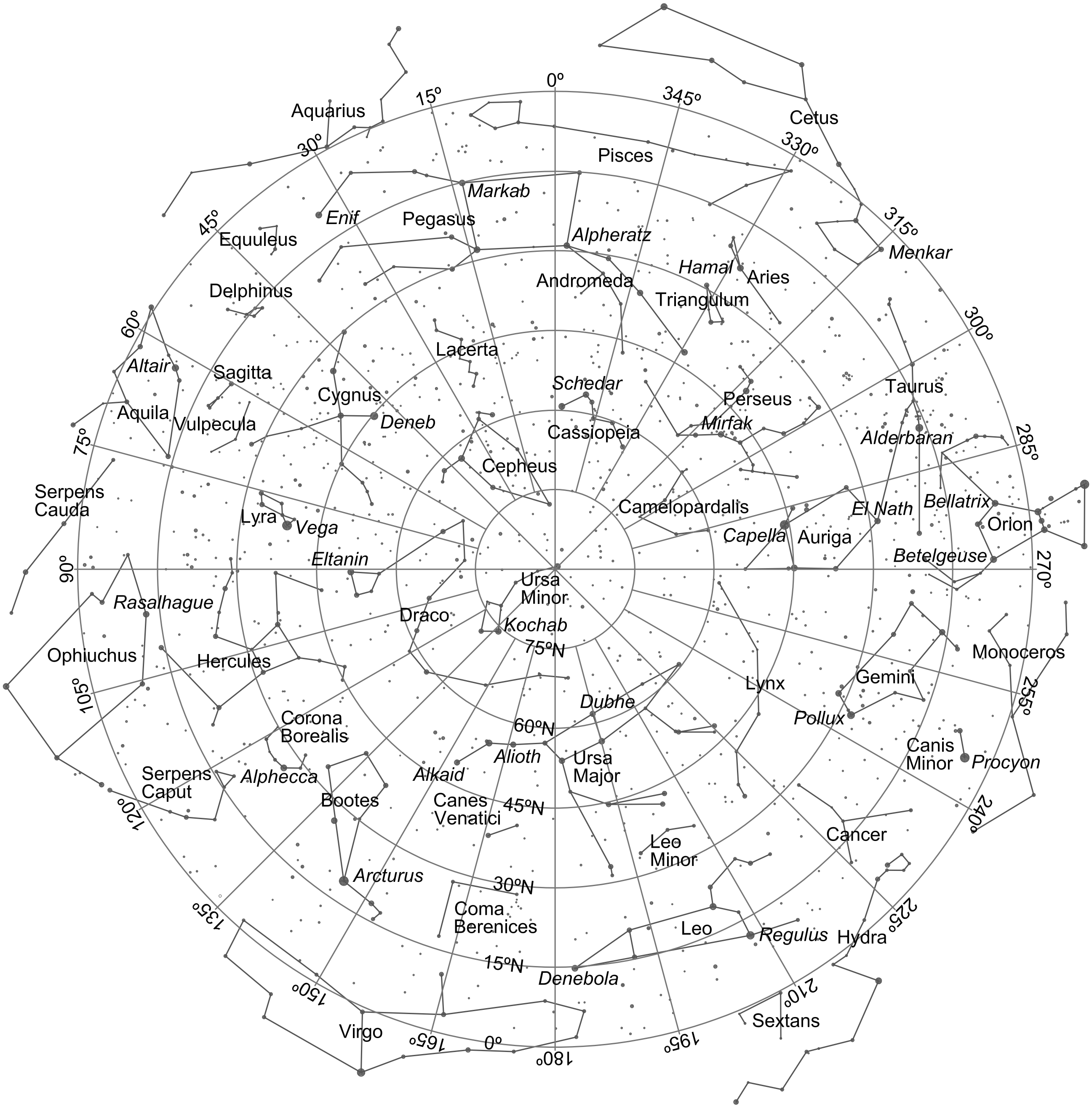
0°
15°
30°
45°
60°
75°
90°
105°
120°
135°
150°
165°
180°
195°
210°
225°
240°
255°
270°
285°
300°
315°
330°
345°
75°N
60°N
45°N
30°N
15°N
0°
Aquarius
Cetus
Pisces
Markab
Enif
Pegasus
Alpheratz
Equuleus
Menkar
Hamal
Aries
Andromeda
Triangulum
Delphinus
Lacerta
Altair
Sagitta
Schedar
Taurus
Perseus
Cygnus
Aquila
Vulpecula
Deneb
Mirfak
Cassiopeia
Alderbaran
Cepheus
Serpens Cauda
Camelopardalis
El Nath
Bellatrix
Lyra
Vega
Capella
Auriga
Orion
Eltanin
Betelgeuse
Ursa Minor
Rasalhague
Draco
Kochab
Ophiuchus
Hercules
Monoceros
Lynx
Gemini
Dubhe
Corona Borealis
Pollux
60°N
Alioth
Canis Minor
Procyon
Serpens Caput
Alphecca
Alkaid
Ursa Major
Bootes
Canes Venatici
Leo Minor
Cancer
Arcturus
Coma Berenices
Leo
Regulus
Hydra
Denebola
Sextans
Virgo

North celestial hemisphere

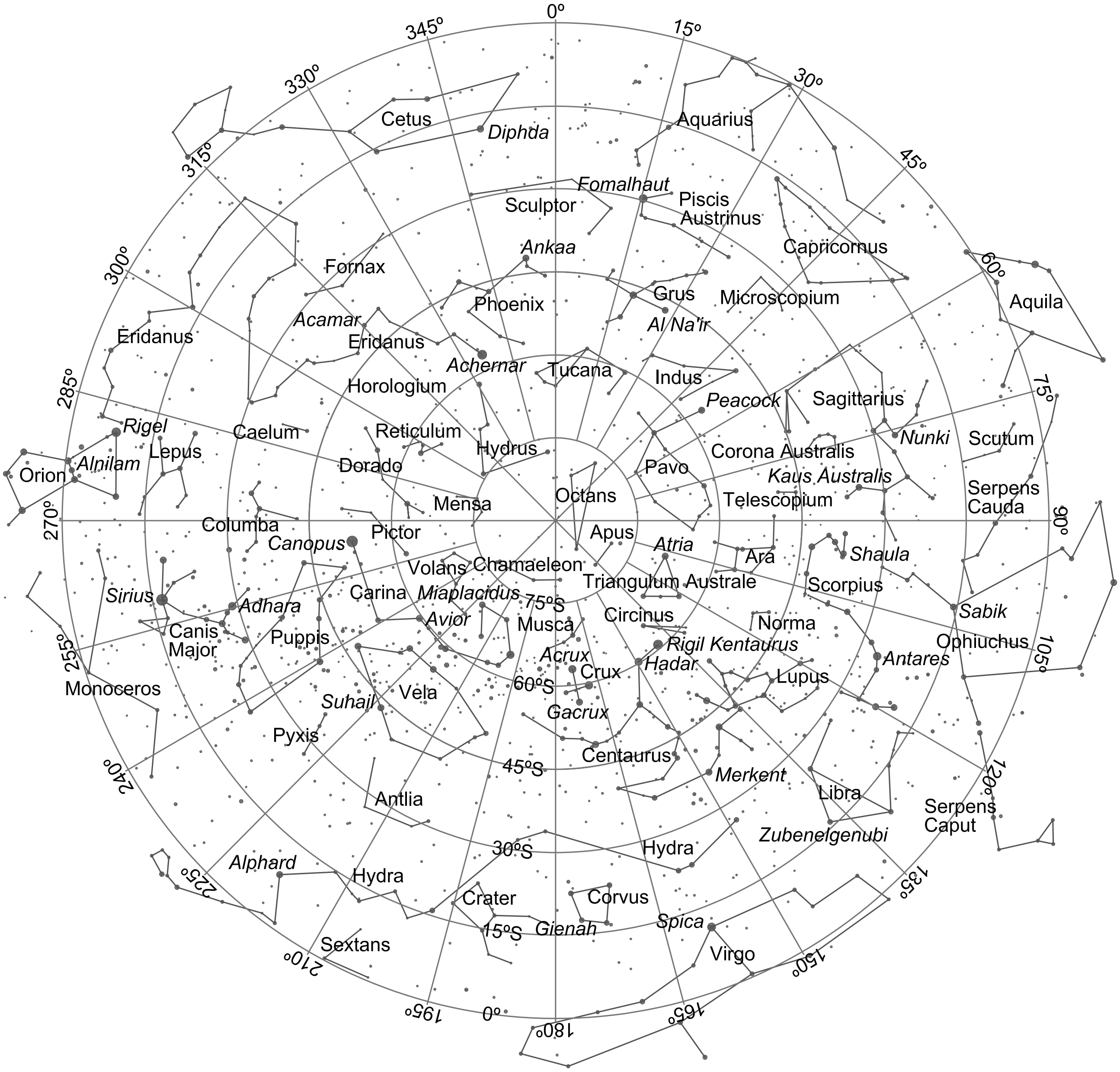
0º
15º
30º
45º
60º
75º
90º
105º
120º
135º
150º
165º
180º
195º
210º
225º
240º
255º
270º
285º
300º
315º
330º
345º
75ºS
60ºS
45ºS
30ºS
15ºS
0º
Cetus
Diphda
Aquarius
Fomalhaut
Piscis Austrinus
Sculptor
Ankaa
Capricornus
Aquila
Fornax
Phoenix
Grus
Al Na'ir
Microscopium
Acamar
Eridanus
Achernar
Tucana
Indus
Peacock
Sagittarius
Nunki
Scutum
Horologium
Reticulum
Hydrus
Pavo
Corona Australis
Kaus Australis
Telescopium
Serpens Cauda
Rigel
Alnilam
Orion
Lepus
Caelum
Dorado
Mensa
Octans
Columba
Canopus
Pictor
Apus
Atria
Ara
Shaula
Volans
Chamaeleon
Triangulum Australe
Scorpius
Sirius
Adhara
Carina
Miaplacidus
Sabik
Canis Major
Puppis
Avior
Musca
Circinus
Norma
Ophiuchus
Rigil Kentaurus
Monoceros
Acrux
Crux
Hadar
Antares
Lupus
Vela
Suhail
Gacrux
Pyxis
Centaurus
Merkent
Libra
Serpens Caput
Antlia
Hydra
Zubenelgenubi
Alphard
Crater
Corvus
Spica
Gienah
Sextans
Virgo

South celestial hemisphere

At a fixed location on Earth, a particular rising-setting star always rises at the same point on the eastern horizon, follows the same arc across the sky, and sets at the same point on the western horizon. This is true throughout the year, although the stars rise and set four minutes earlier each day. This means that at certain times of the year, some stars are above the horizon during daylight hours and therefore invisible.

Before trying to navigate with stars, familiarize yourself with the night sky. Study star charts such as the ones on the previous two pages to learn where the important stars are located. A star chart is similar to a map except that you're looking up instead of down and celestial coordinates are used instead of geographical coordinates. Like maps, star charts are created using various types of projections.

Most bright stars belong to a particular constellation. The easiest way to identify a star is to first find its constellation and then determine the star's position within the constellation. To determine a star's declination (latitude) and sidereal hour angle (longitude), look at a star chart or look up the numbers in an almanac.

The consistent movement and abundant numbers of stars make them inherently well suited for navigation. The biggest drawback is that navigational information can be extracted from stars in so many ways that the task can seem overwhelming. There are two main navigation applications for stars: (1) finding a particular bearing, and (2) determining your location.

5.4.1 Finding north or south

Earth's axis of rotation points to two distinct positions in the sky, the north and south celestial poles. These two positions are essentially long-distance landmarks. A bright star located very near the north celestial pole makes this position in the sky easy to identify. Unfortunately, there is no corresponding bright star near the south celestial pole.

Finding due north

A bright star called Polaris is located within 1° of the north celestial pole in the constellation Ursa Minor, also known as the Little Dipper. The Little Dipper can be difficult to recognize because only three out of its seven stars are bright. The easiest way to identify Polaris is to first find the Big Dipper, which is made up of seven bright stars. The Big Dipper is part of constellation Ursa Major.

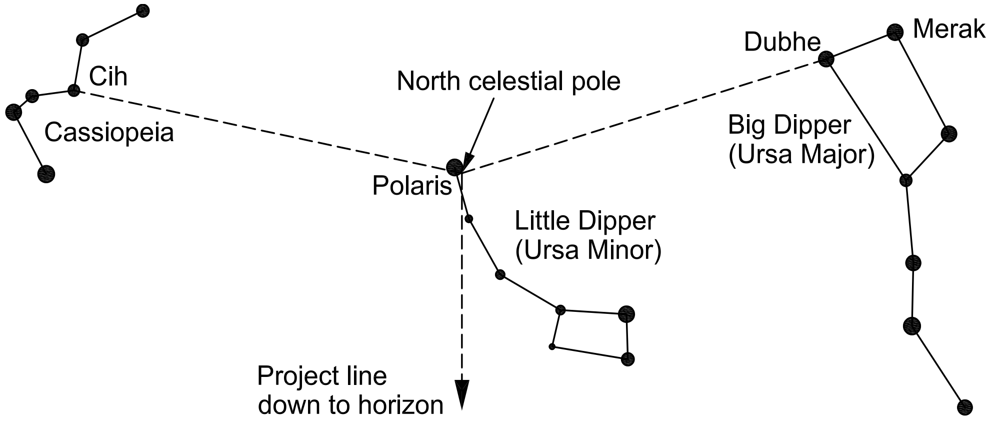

Draw an imaginary line from the last two stars of the Big Dipper, Merak and Dubhe, and extend this line by five times to reach Polaris. If the Big Dipper is not in view, find Cassiopeia, which is made up of six bright stars and looks somewhat like a “W”. From the center star, Cih, draw a line up and to the left to reach Polaris. To find due north, project an imaginary line straight down from Polaris to the horizon.

Finding due south

There is no bright star near the south celestial pole, so you have to visualize the south celestial pole as a fixed position between nearby stars. Find a distinct constellation called the Southern Cross (Crux), and two bright stars called Hadar and Rigil Kentaurus in the nearby constellation Centaurus. These last two stars are collectively known as The Pointers.

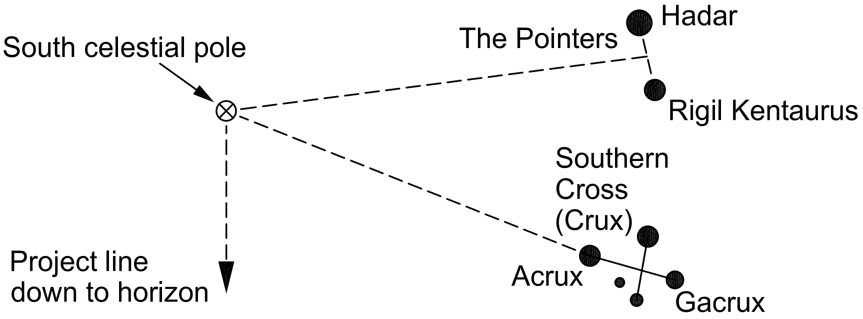

Draw an imaginary line between the two long-axis stars Gacrux and Acrux of the Southern Cross. Extend this line upward and a bit to the right by four times. This is the location of the south celestial pole. To confirm this position, connect a line between Hadar and Rigil Kentaurus. From the center of the connecting line, draw another line at right angle to the connecting line. Extend this line to where it intersects the line from the Southern Cross. To find due south, project an imaginary line straight down from the south celestial pole to the horizon.

5.4.2 Using stars to determine rough positions

With a star almanac, you may be able to use a bright star to determine your position without an angle-measuring device. Your derived position will be inexact and could easily be off by 100 km or more.

Latitude from zenith stars

Stars that pass directly above your location are called zenith stars. They are directly correlated to your latitude. After identifying a star that has passed directly overhead, just look up its declination in a star almanac or chart. Your latitude is the same as the star’s declination. In almanacs, the declination is usually written as a negative number for latitudes south of the equator.

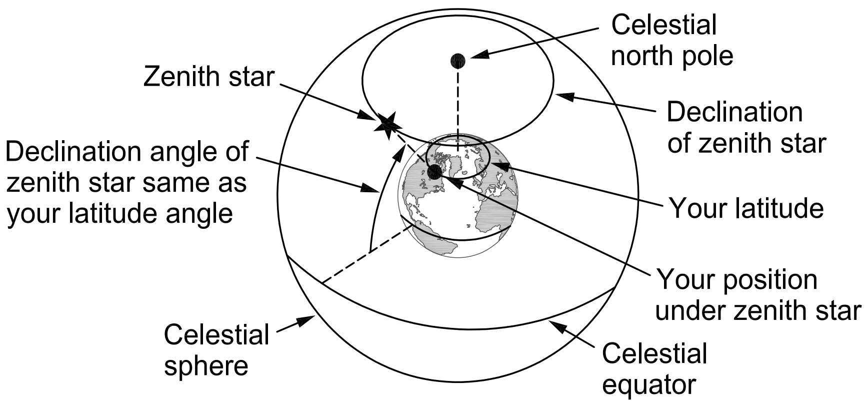
Celestial north pole
Zenith star
Declination of zenith star
Declination angle of zenith star same as your latitude angle
Your latitude
Your position under zenith star
Celestial sphere
Celestial equator

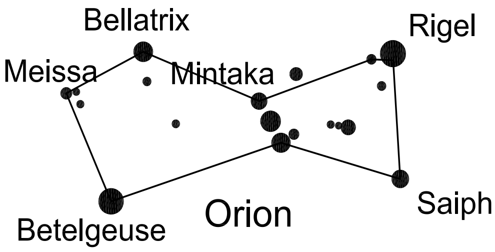
Bellatrix
Rigel
Meissa
Mintaka
Saiph
Orion
Betelgeuse

Example of using zenith star to find latitude

You observe the night sky for some time and notice a very bright star passing directly overhead. You recognize this star as being part of the Orion constellation. From a star chart, you identify the star as Rigel. A star almanac gives its declination as -08º14'. Your latitude is therefore 8º14’S. Round off this number to 8ºS.

Longitude from stars at meridian transit

Your local celestial meridian is the projection of your longitude onto the celestial sphere. It’s an imaginary line that runs through due north, zenith, and due south. Stars that are on your local celestial meridian at a particular moment are said to be at meridian transit.

Stars that are always above the horizon either reach their highest or lowest point above the horizon at meridian transit. Stars that are above the horizon half the time at your latitude always reach their highest point above the horizon at meridian transit. A star at meridian transit either has the same GHA (Greenwich hour angle) as your longitude, or a GHA that is 180º off from your longitude.

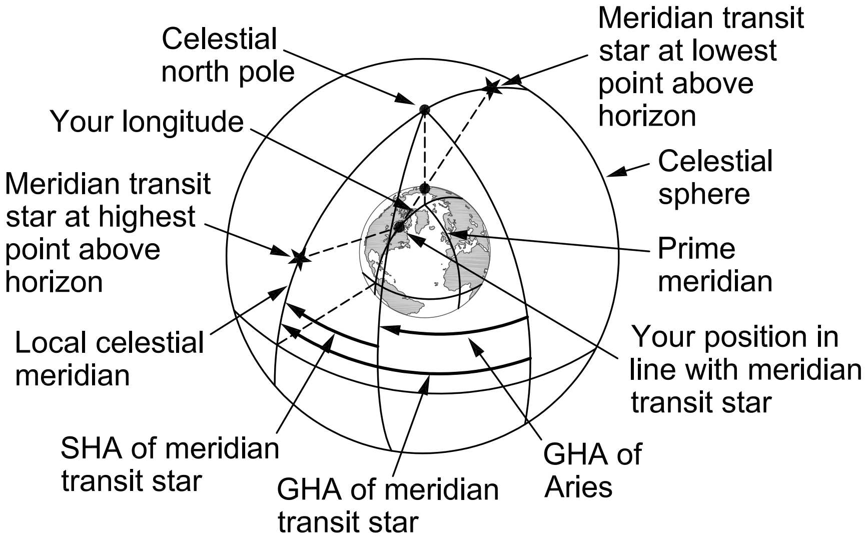
Celestial north pole
Meridian transit star at lowest point above horizon
Your longitude
Celestial sphere
Meridian transit star at highest point above horizon
Prime meridian
Local celestial meridian
Your position in line with meridian transit star
SHA of meridian transit star
GHA of Aries
GHA of meridian transit star

To take advantage of this phenomenon, you need a clock, a star chart, and an almanac. In practice, you will also need a compass or some other way to tell the direction of north or south. Set your clock to read universal time. Set your compass to 0º to view meridian transit stars to the north, or set your compass to 180º to view stars to the south. These bearings must be compensated for magnetic declination and point toward the geographic poles.

Aim your compass directly north or south and identify a bright star in that direction. For best accuracy, pick a star near the horizon. If there is no suitable star in either direction, wait for one to arrive. Write down the date and exact time when the meridian star was directly north or south. Identify the star using the star chart and then consult your almanac to obtain the star’s GHA at that time and date. Your longitude is either the same as the star’s GHA or off by 180º.

Example of using a star to find longitude

Aiming your compass straight north, you notice a bright star directly in line with the compass bearing. The date is October 28, 2007 and the time is 09:48:15 UT. You recognize the star as being part of the Cygnus constellation, and from the star chart identify the star as Deneb. Your almanac shows that Deneb has a SHA of 49º34.6'. At the time of observation, the GHA of Aries was 183º25.7'. The GHA of Deneb is thus 49º34.6' + 183º25.7' = 233º00.3'. Since this is greater than 180º, your longitude is east. Converting the hour angle to longitude gives 360º00.0' - 233º00.3' = 126º59.7'E. Round off this number to 127ºE.

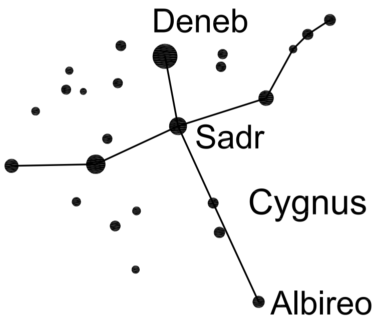

To figure out the direction of due north or due south without a compass, you could observe a star for some time and try to determine when it reaches its highest or lowest point above the horizon. This is very difficult without an angle-measuring device because the rate of change of height is minimal around meridian transit. A better method is to observe two stars that are located at the same SHA (sidereal hour angle) on the celestial sphere. These so-called transit pairs reach meridian transit at the same time. At that moment, one of the two stars will be directly above the other. An imaginary line drawn between the pair and extended down to the horizon will hit the horizon at right angles and the two stars will be aligned north-south.

Example of using transit pair to find longitude

A classic example of a transit pair is the stars Acrux and Gacrux in the Southern Cross constellation. These two stars aren’t exactly on the same hour angle but close enough to determine your approximate longitude. Let’s say you see the Acrux-Gacrux pair aligned more or less vertically to the horizon at 13:56:02 UT on December 17, 2003.

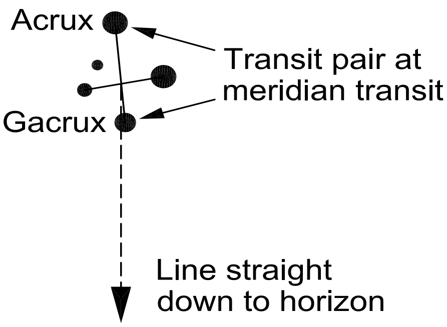
Acrux
Gacrux
Transit pair at meridian transit
Line straight down to horizon

From an almanac, the SHA for Acrux is 173º18.2' and the GHA for Aries at that moment is 294º47.5'. The GHA for Acrux is 173º18.2' + 294º47.5' = 468º05.7' = 468º05.7' - 360º = 108º05.7'.

To confirm your result repeat the calculations for Gacrux. Its SHA is slightly different at 172º09.6'. The GHA for Gacrux is then 294º47.5' + 172º09.6' = 466º57.1' = 466º57.1' - 360º = 106º57.1'. Calculating the average GHA for the two stars gives (108º05.7' + 106º57.1')/2 = 215º02.8'/2 = 107º31.4'. Since this number is less than 180º, your longitude is west. Rounding off this number gives you an approximate longitude of 108ºW.

You realize that this longitude is way off and by careful observation you notice why. The line down to the horizon points away from the south celestial pole. This means that the transit pair is in fact lower in the sky than the south celestial pole and must be at the lowest point in its orbit rather than the highest. The calculated longitude is therefore off by 180º. Your correct longitude is 108ºW - 180º = -72ºW = 72ºE.

5.4.3 Stars as direction finders

At a given latitude, a particular star always rises at a certain fixed bearing, follows the same arc across the sky, and then sets at another fixed bearing. Bright stars that are low on the horizon can therefore be used to find directions, and are known as steering stars. To exploit this phenomenon, you need to create a star compass that shows the rising and setting bearings for all the important stars at your latitude (see example). This is best done with a celestial navigation calculator, but can also be done with a star almanac. Of particular interest are stars on the celestial equator, which rise exactly east and set exactly west from any latitude, and Polaris that always points north. Transit pairs at meridian transit can also act as steering stars because they point either north or south at that moment.

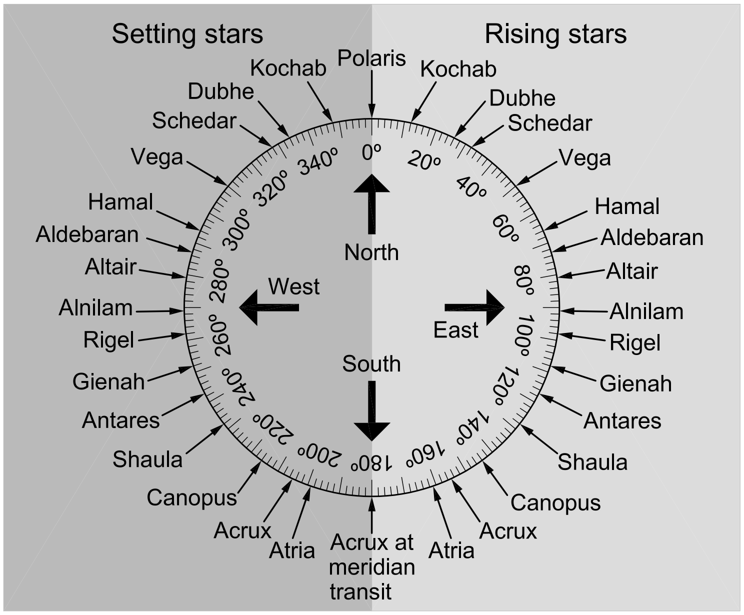
Setting stars
Rising stars
Polaris
Kochab
Dubhe
Schedar
Vega
Hamal
Aldebaran
Altair
Alnilam
Rigel
Gienah
Antares
Shaula
Canopus
Acrux
Atria
Acrux at meridian transit
Kochab
Dubhe
Schedar
Vega
Hamal
Aldebaran
Altair
Alnilam
Rigel
Gienah
Antares
Shaula
Canopus
Acrux
Atria
0º
20º
40º
60º
80º
100º
120º
140º
160º
180º
200º
220º
240º
260º
280º
300º
320º
340º
North
West
East
South

Example - star compass for latitude 10ºN

All rising-setting stars approach or leave your horizon at an angle of 90º minus your latitude. For example, if you're at 30ºS, the stars will rise or set at an angle of 60º relative to the horizon. Use this relationship to estimate the point on the horizon where a particular steering star has risen or will set. To determine the bearing of this point, consult your star compass for the star's rising or setting bearing. A steering star loses its usefulness as it moves too high in the sky. For continuous navigation throughout the night, use a series of steering stars that rise or set at different times.

5.5 Sextant navigation

The sextant is an angle-measuring device specifically designed for celestial navigation. It measures the vertical angle between a celestial body and the horizon. In sextant navigation, this angle is called the altitude of the celestial body. The process of measuring the angle is called taking a sight. To determine your position, take sights from at least two celestial bodies and record the exact times when these sights were taken. Perform several sets of calculations and then plot two or more lines of position on your chart. Your calculated position is the intersection between the lines of position.

Sextant navigation is not for people averse to complexity because of the elaborate procedure. The calculations are usually done with the help of tables with pre-calculated values or by using a celestial navigation calculator. This book only deals with the basic steps. Anyone contemplating navigating with a sextant should consult a book dedicated to the topic. Skip this section if your brain isn't working.

5.5.1 Underlying theory

Every celestial body is at zenith (directly overhead) at some location on Earth at a particular moment. The location directly below the celestial body is called the geographic position of the body. The geographic position moves as Earth rotates. In the case of the sun, moon, and planets, the geographic position also moves because of the body's and/or Earth's orbital motion. If you see a particular celestial body directly overhead at a certain time and know when and where this body is supposed to be at zenith, you will know your position. In practice, it's difficult to take advantage of this phenomenon because you are unlikely to be directly underneath an identifiable celestial body at a particular moment.

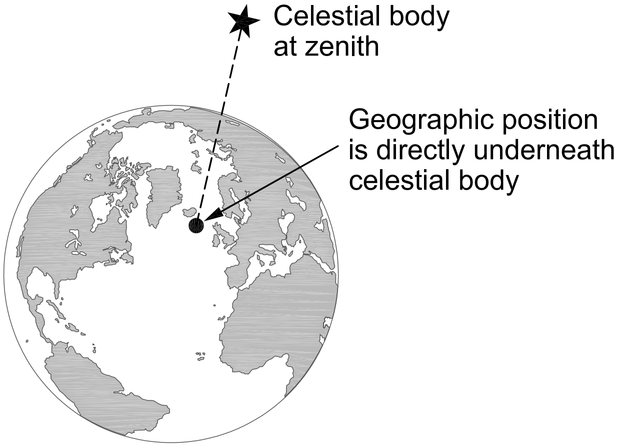

A more useful method is to observe a celestial body that is not at zenith. The geographic position of the body is then a certain distance and bearing from your location. Because of Earth's curvature, the celestial body appears lower on the horizon the further away you are from the body's geographic position. Most celestial bodies are so far away that their light rays are parallel anywhere on Earth. By measuring the angle between the celestial body and your horizon, you can calculate how far away you are from the geographic position. The smaller the angle, the further away you are. This angle is called the altitude of the celestial body and is what you measure when taking a sextant sight.

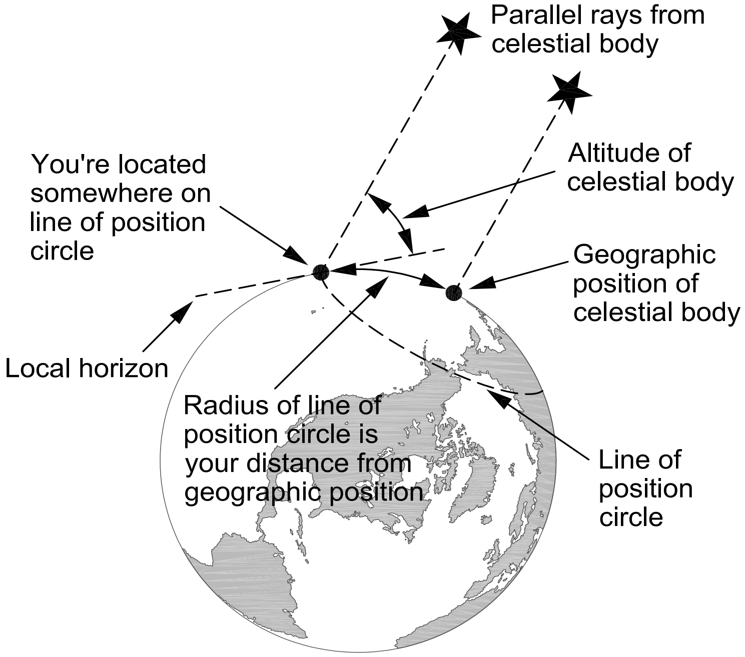
Parallel rays from celestial body
You're located somewhere on line of position circle
Altitude of celestial body
Geographic position of celestial body
Local horizon
Radius of line of position circle is your distance from geographic position
Line of position circle

One sight establishes that you are located somewhere on a line of position circle that is centered on the geographic position of the celestial body, but does not tell you the direction to the geographic position. The "radius" of the circle is the distance along Earth's surface from your location to the geographic position. This radius is usually stated in angular units. To convert to distance units, use the formula: 1 minute of arc = 1 nautical mile = 1852 m.

Directly plotting a line of position circle on a chart is impractical in most cases because the geographic position is usually so far away that it is well beyond the boundaries of your chart. The usual procedure is to first select an assumed position somewhere on your chart near where you think you are. If you're using tables for your calculations, select an assumed position that will simplify subsequent calculations. With tables or a calculator, calculate the bearing and distance from your assumed position to the celestial body's geographic position. Use this information to plot a portion of an assumed line of position circle on your chart.

Because the geographic position is so far away, the portion of the circle that's on your chart is nearly straight, and is usually drawn as a straight line. After taking a sight, you can calculate the distance from your actual position to the geographic position. The difference in distance to the geographic position from your actual position versus from your assumed position is called the intercept.

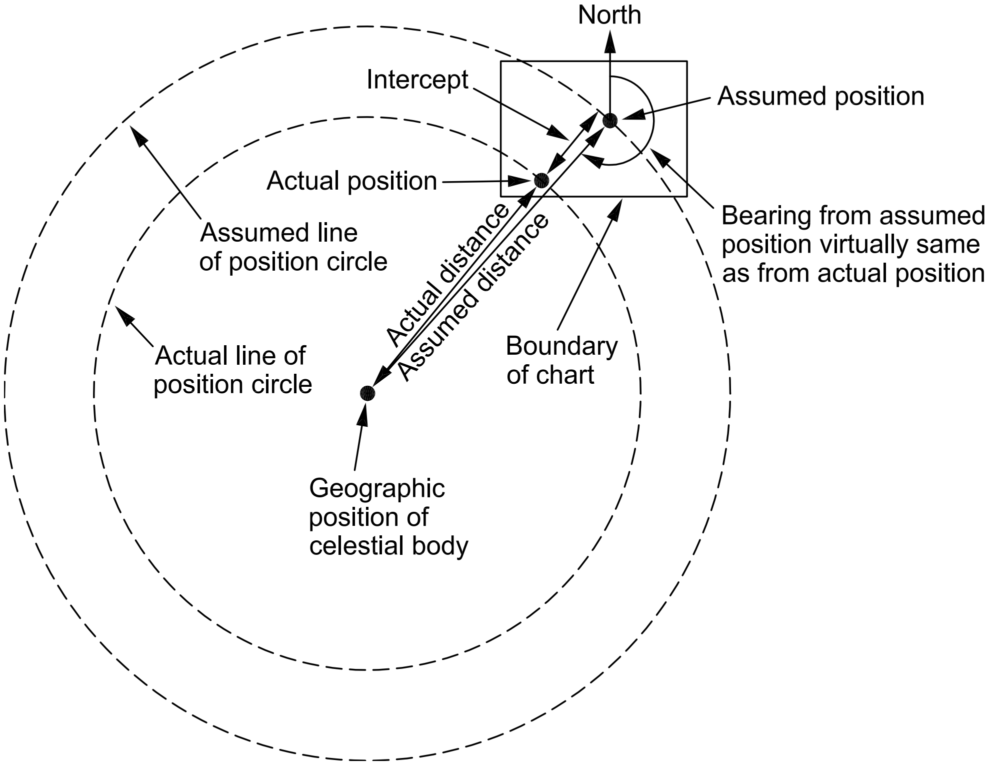
North
Intercept
Assumed position
Actual position
Assumed line of position circle
Bearing from assumed position virtually same as from actual position
Actual distance
Assumed distance
Actual line of position circle
Boundary of chart
Geographic position of celestial body

Because the geographic position is so far away, the bearing from your actual position to the geographic position is virtually the same as the bearing from your assumed position to the geographic position. On the chart, the actual line of position is a straight line parallel to the assumed line of position, but shifted by an amount equal to the intercept. The shift is either toward or away from the geographic position of the celestial body.

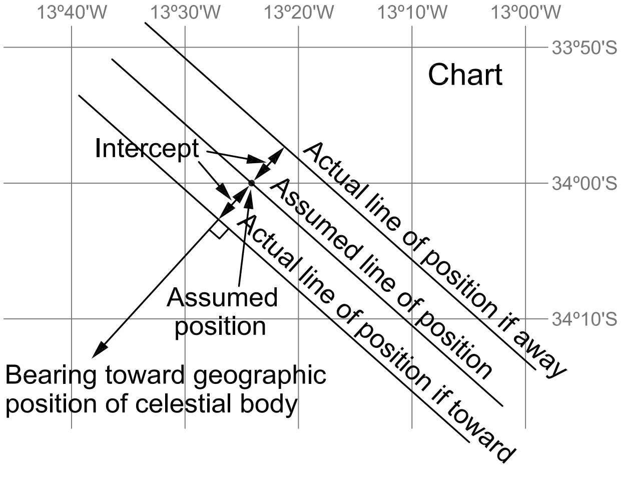
13°40'W
13°30'W
13°20'W
13°10'W
13°00'W
33°50'S
34°00'S
34°10'S
Chart
Intercept
Actual line of position if away
Assumed line of position
Actual line of position if toward
Assumed position
Bearing toward geographic position of celestial body

Taking sights from two celestial bodies establishes that you are simultaneously located on two line of position circles. This means that you must be on one of the intersections between the two circles. If you already know your approximate position, you should be able to eliminate one of the intersections. If not, take a third sight. Your position is the triple intersection between the three circles. Instead of taking sights of several celestial bodies, you can take several sights of the same celestial body at different times. Because the geographic position of the body moves, each sight gives you a new line of position circle.

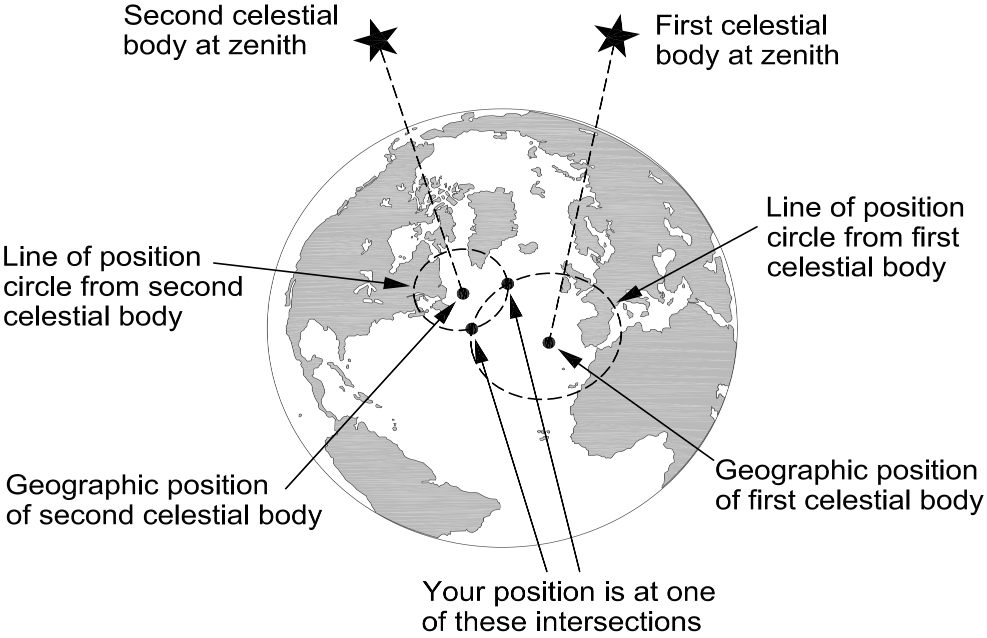
Second celestial body at zenith
First celestial body at zenith
Line of position circle from first celestial body
Line of position circle from second celestial body
Geographic position of second celestial body
Geographic position of first celestial body
Your position is at one of these intersections

Taking a sight from a second celestial body and selecting a second assumed position enables you to plot a second line of position on your chart. Note that your second assumed position could be different from your first assumed position even if you haven't moved. The intersection between the two lines of position is your calculated position.

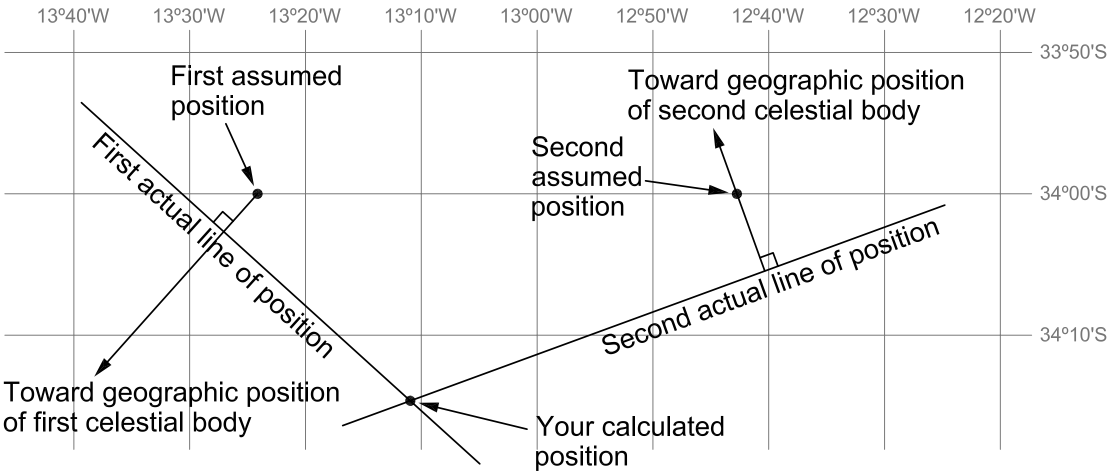
13°40'W
13°30'W
13°20'W
13°10'W
13°00'W
12°50'W
12°40'W
12°30'W
12°20'W
33°50'S
34°00'S
34°10'S
First assumed position
Toward geographic position of second celestial body
Second assumed position
First actual line of position
Second actual line of position
Toward geographic position of first celestial body
Your calculated position

If you moved during the time interval between the two sights, you must shift the first line of position by your estimated movement since your first sight. Use dead reckoning (Section 8.3) to estimate this movement. Your calculated position is the intersection between the shifted first actual line of position and the second actual line of position.

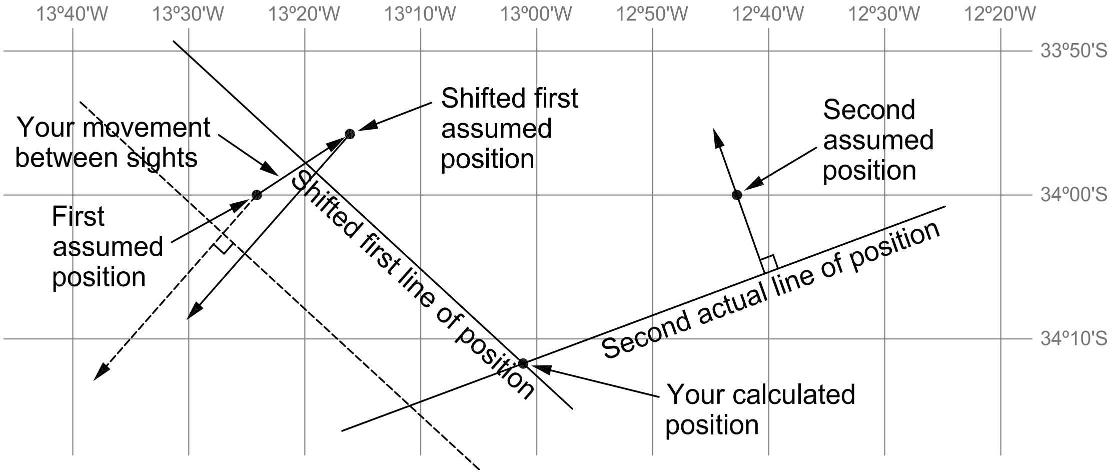
13°40'W
13°30'W
13°20'W
13°10'W
13°00'W
12°50'W
12°40'W
12°30'W
12°20'W
33°50'S
34°00'S
34°10'S
Your movement between sights
Shifted first assumed position
Second assumed position
First assumed position
Shifted first line of position
Second actual line of position
Your calculated position

5.5.2 How the sextant works

The marine sextant consists of a triangular frame with a curved, graduated scale engraved on the bottom. A small telescope is attached horizontally on one side of the frame. Opposite the telescope is a fixed glass piece that is divided vertically into two sectors, one is transparent and the other reflective. Some sextants use one partially

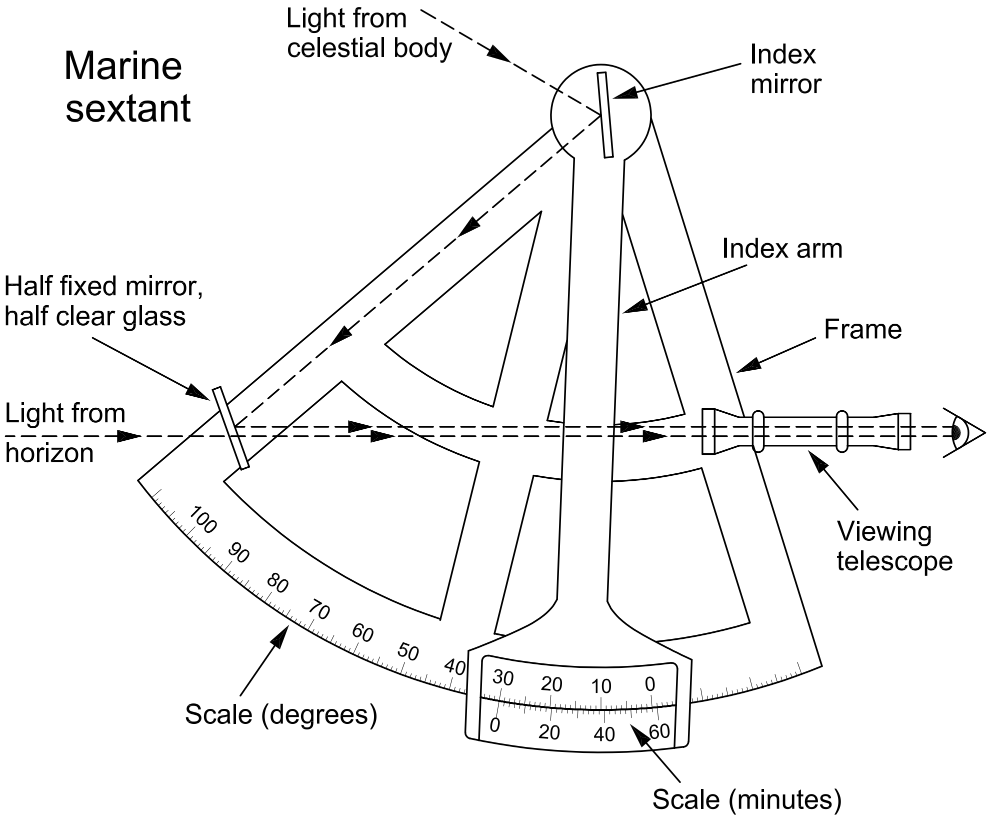
Marine sextant
Light from celestial body
Index mirror
Index arm
Half fixed mirror, half clear glass
Frame
Light from horizon
Viewing telescope
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
Scale (degrees)
Scale (minutes)

silvered mirror instead. An index arm is attached via a pivot joint to the top of the frame so that the lower part of the index arm can rotate across the graduated scale. An index mirror is attached to the index arm at the pivot joint.

This setup is designed so that the celestial body and the horizon can both be viewed at the same time through the telescope. When the index arm is turned to a certain angle, the celestial body will appear to be touching the horizon. At that moment, the scale will show the altitude of the celestial body, the vertical angle between the horizon and the body.

The marine sextant can only be used when you have an unobstructed view of the horizon. Other types of sextants that use artificial mercury or bubble horizons can be used anywhere but are not as accurate as the marine sextant.

5.5.3 Taking a sight

You need a clear view of both the horizon and the celestial body to measure the altitude of a celestial body with a marine sextant. With stars or planets you only have a small window of opportunity at dawn or dusk when it’s light enough to see the horizon, but also dark enough to see the star or planet. A sight of the sun can be taken anytime it is visible. To avoid injuring your eyes, attach a dark sun filter before taking a sight of the sun.

With the scale set to zero, aim the telescope toward the celestial body. After finding the body, let the frame pendulum down without losing sight of the celestial body. To do this, you have to continually adjust the index arm. The horizon will come into view. Adjust the index arm so that the celestial body touches the horizon. Rock the sextant from side to side until the celestial body reaches its lowest point. This ensures that the sextant is held vertically. Fine-tune the index arm adjustment until the celestial body is exactly level with the horizon. With the sun or moon, make either the lower or upper limb touch the horizon.

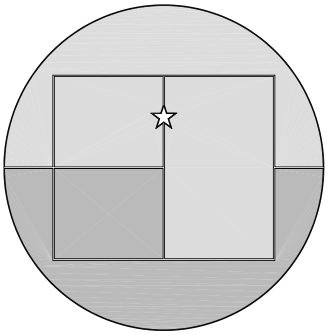

Point telescope toward body and then bring body down toward horizon.

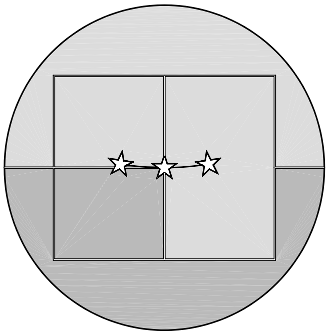

Rock sextant from side to side to ensure sextant is vertical.

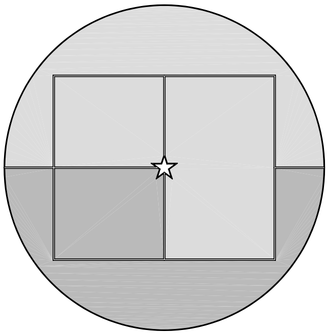

With body on horizon and sextant vertical, read off the declination.

Read off the altitude on the scale and write down the universal date and exact universal time (to the second) when the sight was taken.

5.5.4 Sextant calculations

There are two main methods to compute your position after taking sights of celestial bodies: (1) using an almanac together with sight reduction tables, or (2) using a celestial navigation calculator. The first method consists of a series of steps that involve looking up numbers in tables that are added or subtracted from each other. Your position is derived graphically from the intersection between lines of position on a chart.

Aside from being tedious, sextant calculations are riddled with jargon and acronyms, so it's easy to get confused. Develop a consistent routine to reduce the chance of error. The following procedure is an outline of how to compute a position fix the hard way by using an almanac and sight reduction tables. Each step is followed by an example. The best way to comprehend what is going on is to go through the calculations with an almanac and sight reduction tables at hand.

Take sextant sights

Check your almanac to see what celestial bodies it covers. Most almanacs have tables for the sun, moon, four planets, and 57 navigation stars. Take sights of at least two of these celestial bodies with your sextant. If only one body is visible, take several sights of that body at different times. For each sight, write down the sextant altitude (Hs), exact universal time (UT), and universal date (UD). If your clock is set to local time, convert your clock time to universal time and determine if the time conversion has changed the universal date. You also need to know the approximate locations where the sights were taken. This is usually accomplished by dead reckoning (Section 8.3).

On June 16, 2003 UD, you take a sight of the star Acamar at 22:15:46 UT. Reading off the sextant scale gives an altitude (Hs) of 61°42.8'. Your dead reckoning (DR) position at the time of the sight was 46º03.4'S, 127º19.7'E.

A few hours later when the universal date has changed to June 17, 2003, you take a sight of the lower limb of the sun at exactly 02:04:29 UT. The sextant gives an altitude (Hs) of 18°07.3'. You estimate that you've moved 38.9 nautical miles (nm) toward a bearing of 26º since your first sight. This gives you a dead reckoning (DR) position of 45º48.3'S, 128º13.9'E for the sun sight. Below is a summary of the results from this step:

Hs (Acamar) = 61°42.8', DR (Acamar) = 46º03.4'S, 127º19.7'E
Hs (Sun) = 18°07.3', DR (Sun) = 45º48.3'S, 128º13.9'E

Compensate for sight errors

The measured altitude of the celestial body isn't usually the exact angle between the horizon and the celestial body. You must compensate for: (1) the sextant's elevation above sea level at the time of observation, (2) any systematic index error in the sextant itself, (3) bending of light rays from the body by Earth's atmosphere, (4) whether the upper or lower limb of the sun or moon was used, and (5) parallax. The last factor applies only to moon sights, and is necessary because of the moon's proximity to Earth. After all the corrections are applied, the result is the body's observed altitude (Ho).

Let’s say the sextant telescope was 3.1 m above sea level when the sights were taken. The “dip” column in the “altitude correction” tables gives a correction of -3.1' for a sextant height of 3.1 m. To determine the index error, read off the scale after taking a sextant sight of the horizon. Let’s say the result is -2.6'. Compensating for the two corrections gives the apparent altitudes (Ha) of the bodies:

Ha (Acamar) = 61°42.8' - 2.6' - 3.1' = 61°37.1'
Ha (Sun) = 18°07.3' - 2.6' - 3.1' = 18°01.6'

The “altitude correction” tables also give the correction for atmospheric bending of light rays during different times of the year. For stars and planets the tables give a correction of -0.5' for Ha = 61°37.1. For the sun or moon, these tables give a combined correction for atmospheric bending and upper/lower limb compensation. This latter correction is necessary because the altitude of the center of the sun or moon is required for the calculations, while practical considerations dictate that a sight is always taken of the upper or lower limb of these two bodies. For the sun the tables give a combined lower limb/light bending correction of 13.1' for Ha = 18°01.6'. After the corrections have been applied, the result is the observed altitudes (Ho) of the bodies:

Ho (Acamar) = 61°37.1' - 0.5' = 61°36.6'
Ho (Sun) = 18°01.6' + 13.1' = 18°14.7'

Determine geographic position (GP) of body

The geographic position (GP) of a celestial body consists of its declination (latitude) and GHA (longitude) at the moment the sight was taken. The declination of the celestial bodies is given in the “daily” tables in the almanac. The “daily” tables also provide the GHA for the sun, moon, planets, and Aries meridian to the nearest hour. The “increment and corrections” table gives the GHA of all bodies to the nearest second. The SHA (sidereal hour angle) of the stars is written in one column of the “daily” tables.

For Acamar, the “daily” tables give a declination of 40°17.3'S, and a SHA of 315º24.6'. The “daily” tables also give a GHA for Aries of 234º45.3' at 22:00 UT on June 16, 2003. The Acamar sight was however taken at 22:15:46 UT so you must make up for the extra 15 minutes and 46 seconds past 22:00. The “increment and corrections” tables give 3°57.1' that must be the added to the GHA. The GHA for Acamar is therefore 315º24.6' + 234º45.3' + 3°57.1' = 554º07.0' = 554º07.0' - 360º = 194º07.0'.

The “daily” tables give a declination of 23°21.8'N for the sun. The GHA for the sun at 02:00 UT on June 17, 2003 is given as 209°49.3'. The sun sight was however taken at 02:04:29 UT so you need to compensate for the extra 4 minutes and 29 seconds past 02:00. The “increment and corrections” tables give an extra 1°07.3'. The GHA for the sun is therefore 209º49.3' + 1°07.3' = 210º56.6'. Below are the complete geographic positions (GP) of the bodies when the sights were taken:

GP (Acamar) = Dec 40°17.3'S, GHA 194º07.0'
GP (Sun) = Dec 23°21.8'N, GHA 210º56.6'

Select assumed position (AP) and determine local hour angle (LHA)

If the geographic position of a celestial body happens to be on your chart, you can immediately draw a line of position circle centered on the geographic position and with a radius equal to the observed altitude (Ho) of the body. The radius is converted from angle units to distance units using the conversion formula: 1 minute of arc (') = 1 nautical mile (nm). In most cases, the geographic positions are far outside the boundaries of your chart and you cannot directly draw a line of position circle on your chart. Instead, you must select an assumed position (AP) that is then readjusted based on the difference between the observed altitude (Ho) and the computed altitude (Hc) of the celestial body. Your AP must be selected as close as possible to, but generally not exactly, where you think you are.

To simplify subsequent calculations in the sight reduction tables, select an assumed position according to the following criteria: Select the assumed latitude to the nearest degree of your dead reckoning (DR) position. If the longitude is west, select the minutes of the assumed longitude to be equal to the minutes of the GHA of the body. If the longitude is east, select an assumed longitude so that its minutes plus the minutes of the GHA of the body add up to a whole degree (60').

The local hour angle (LHA) is the difference between your assumed longitude and the GHA of the celestial body. If your assumed longitude is east, add the assumed longitude to the GHA to get the LHA. If your assumed longitude is west, subtract the assumed longitude from the GHA to get the LHA.

Your DR position was 46º03.4'S, 127º19.7'E when the Acamar sight was taken. Select the assumed latitude to the nearest degree of the DR latitude, i.e. 46ºS. Since your assumed longitude is east and the GHA of Acamar is 194º07.0', select an assumed longitude so that its minutes are 60' - 07.0' = 53.0'. The closest longitude to 127º19.7'E that matches this criterion is 126º53.0'E. Since the assumed longitude is east, add the longitude to the GHA to get the local hour angle (LHA) for Acamar: LHA = 194º07.0' + 126º53.0' = 321º.

Your DR position was 45º48.3'S, 128º13.9'E when the sun sight was taken. Select the assumed latitude to the nearest degree, i.e. 46ºS. Since your assumed longitude is east and the GHA of the sun is 210º56.6', select an assumed longitude so that its minutes are 60' - 56.6' = 3.4'. The closest longitude to 128º13.9'E that matches this criterion is 128º03.4'E. Since the assumed longitude is east, add the longitude to the GHA to get the local hour angle (LHA) for the sun: LHA = 210º56.6' + 128º03.4' = 339º. Below is a summary of the results from this step:

AP (Acamar) = 46ºS, 126º53.0'E
LHA (Acamar) = 321º

AP (Sun) = 46ºS, 128º03.4'E
LHA (Sun) = 339°

Compute altitude (Hc) and azimuth (Zn) for assumed position (AP)

With the declination of the GP, the latitude of the AP, and the LHA as inputs, use the sight reduction tables to determine the computed altitude (Hc) and azimuth angle (Z) of the body. The computed altitude is the theoretical value of the altitude from the assumed position at the time of the sight. The azimuth angle is the smallest possible horizontal angle between the geographic position of the body and the nearest geographic pole as seen from your assumed position. Azimuth angles are either measured clockwise or counterclockwise. Before using an azimuth angle to plot a line of position, convert it to a true azimuth (Zn), which is always measured clockwise from the north pole like a true bearing.

For Acamar, the three input values into the sight reduction tables are: GP declination = 40°17.3'S, AP latitude = 46ºS, and LHA = 321º. Find the page that is labeled LHA 321º and look in the section of the table that is labeled “latitude same name as declination”. This latter directive is because the GP declination and the AP latitude are both south. For a declination to the nearest degree (40°), the table gives an initial computed altitude (Hc) of 61°09.3', an altitude difference (d) of +24.6', and an initial azimuth angle (Z) of 87.8°. Interpolating adjacent values in the table changes Z to 87.3°. A formula in the margin of the table states that “Zn = 180º - Z for southern latitudes and LHA greater than 180º”. Therefore, Zn = 180º - 87.3° = 92.7°.

The altitude difference (d) is the difference between adjacent Hc values in the table. Find the “interpolation table” and use d (+24.6') and the minutes of the declination (17.3') as inputs. The table gives two correction factors, 5.8', and 1.3'. Since (d) is positive, add these two numbers to the initial Hc. The final Hc = 61°09.3' + 5.8' + 1.3' = 61°16.4'.

For the sun, the three input values are: GP declination = 23°21.8'N, AP latitude = 46ºS, and LHA = 339°. Find the page that is labeled LHA 339º. Since the GP declination is north while the AP latitude is south, look in the section of the table that is labeled “latitude contrary name to declination”. For a declination to the nearest degree (23°), the table gives an initial computed altitude (Hc) of 18°24.9', an altitude difference (d) of -57.9', and an azimuth angle (Z) of 159.7°. A visual interpolation changes Z slightly to 159.8°. A formula in the margin of the table states that “Zn = 180º - Z for southern latitudes and LHA greater than 180º”. Therefore Zn = 180º -159.8° = 20.2º.

Find the “interpolation table” and use d (-57.9') and the minutes of the declination (21.8') as inputs. The table gives two correction factors, 18.2', and 2.8'. Since (d) is negative, subtract these two numbers from the initial Hc. The final Hc = 18°24.9' - 18.2' - 2.8' = 18°03.9'. Below is a summary of the results from this step:

Hc (Acamar) = 61°16.4', Zn (Acamar) = 92.7°.
Hc (Sun) = 18°03.9', Zn (Sun) = 20.2°.

Compare computed altitude (Hc) with observed altitude (Ho)

The difference between the computed altitude (Hc) and observed altitude (Ho) is the altitude intercept (a). The altitude intercept is the difference in radius between the line of position circle derived from the computed altitude (Hc) versus the one derived from the observed altitude (Ho). If Ho is bigger than Hc, move the line of position toward the geographic position of the body by the intercept distance. If Ho is smaller than Hc, move the line of position away from the body. The unit for the intercept, minutes of arc, is equivalent to nautical miles on Earth’s surface.

For Acamar Hc = 61°16.4' and Ho = 61°36.6'. The intercept (a) = 61°36.6' - 61°16.4' = 20.2' = 20.2 nm. Since Ho is larger than Hc, move the line of position toward Acamar by 20.2 nm.

For the sun Hc = 18°03.9' and Ho = 18°14.7'. The intercept (a) = 18°14.7' - 18°03.9' = 10.8' = 10.8 nm. Since Ho is larger than Hc, move the line of position toward the sun by 10.8 nm. The intercept results are summarized below:

a (Acamar) = 20.2 nm toward GP of Acamar.
a (Sun) = 10.8 nm toward GP of sun.

Plot line of position (LOP)

Mark the assumed position (AP) of the first sight on your chart. Draw a line from the AP toward the celestial body’s geographic position (GP) in the direction of the azimuth angle (Zn). Then, draw a line perpendicular to Zn and through the AP. This is your assumed line of position (LOP). Move the assumed LOP toward or away from the GP of the body by a distance equal to the altitude intercept (a) for that body. For best accuracy, use a plotting tool such as a protractor to draw the lines on the chart. Use an orienteering compass if no other tools are available.

Draw a LOP on the chart for each sight that you’ve taken. If all sights were taken from the same location, the intersection of the actual LOPs is your calculated position fix.

Mark the AP for Acamar (46ºS, 126º53.0'E) on the chart. Draw a line from here in the direction of Zn (92.7°) and mark the end of the line with an arrow. Draw a line perpendicular to 92.7° and through the AP for Acamar. This is the assumed LOP for Acamar. Draw a second line parallel to the assumed LOP and shifted by a = 20.2 nm toward the GP of Acamar. This is the actual LOP for Acamar.

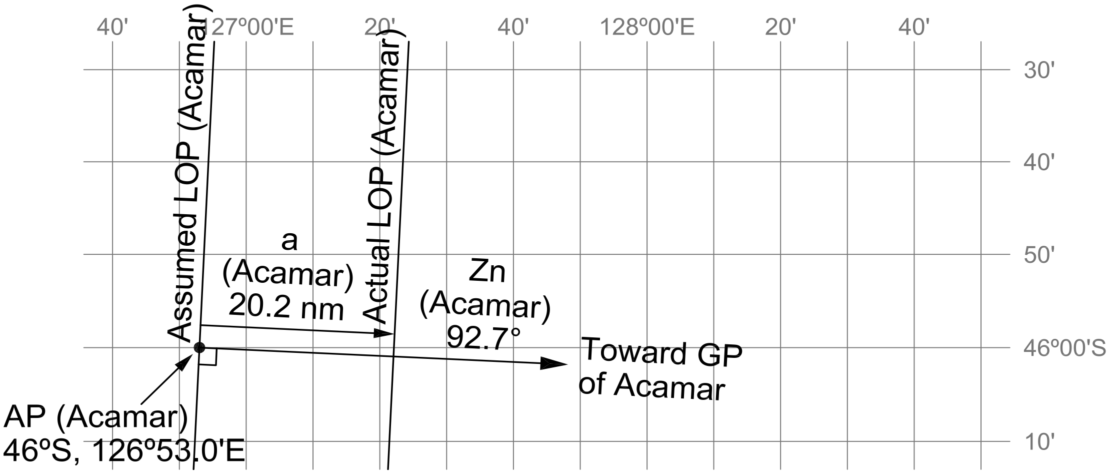

Mark the AP for the sun (46ºS, 128º03.4'E) on the chart. Draw a line from here in the direction of Zn (20.2°) and mark the end of this line with an arrow. Draw a line perpendicular to 20.2° and through the AP for the sun. This is the assumed LOP for the sun. Draw a second line parallel to the assumed LOP and shifted by a = 10.8 nm toward the GP of the sun. This is the actual LOP for the sun.

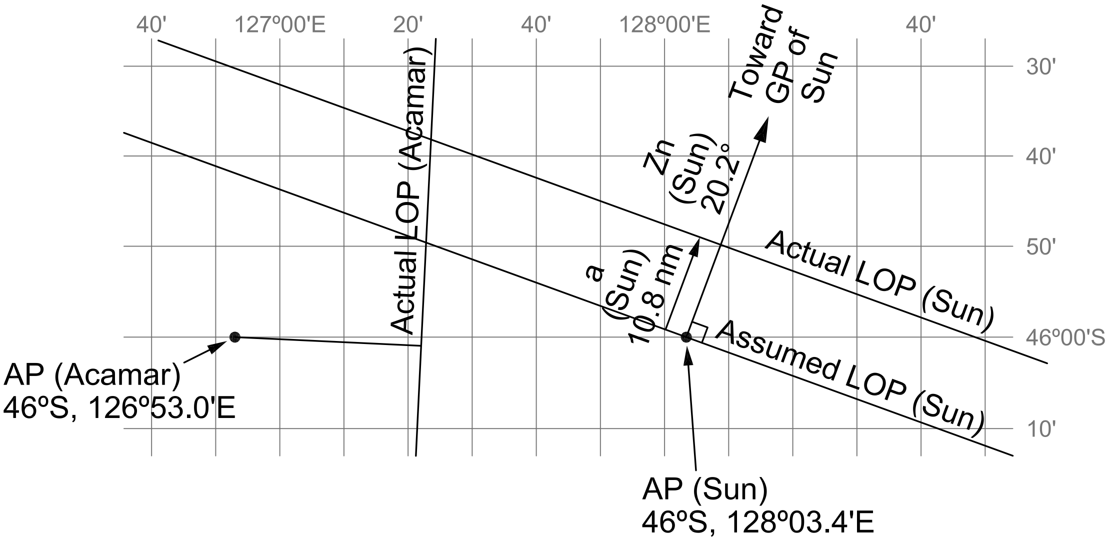
40'
127º00'E
20'
40'
128º00'E
40'
30'
40'
50'
46º00'S
10'
Toward GP of Sun
Zn (Sun) 20.2°
a (Sun) 10.8 nm
Actual LOP (Acamar)
Actual LOP (Sun)
Assumed LOP (Sun)
AP (Acamar)
46ºS, 126º53.0'E
AP (Sun)
46ºS, 128º03.4'E

Shift line of position (LOP)

In many cases, particularly on the ocean, you will have taken sights from different locations. Before reading off your position fix, shift all the actual LOPs so that all the sights appear to have been taken from the location of the last sight. The dead reckoning (DR) positions that you used for earlier calculations will show you how to shift the LOPs.

On the chart, mark the DR positions of all the sights. Measure the distance and bearing from earlier DR positions to the DR position of the last sight. Then shift each earlier LOP by the measured distance and bearing. The intersection between the shifted LOPs and the LOP of your last sight is your calculated position fix.

Mark the DR position of Acamar (46º03.4'S, 127º19.7'E) and the DR position of the sun (45º48.3'S, 128º13.9'E) on the chart. Measure the distance and bearing from DR (Acamar) to DR (Sun). This yields a distance of 40.7 nm and a bearing of 68º. Shift the actual LOP of Acamar by 40.7 nm toward 68º. Your calculated position fix is the intersection between the shifted LOP of Acamar and the LOP of the sun. Use the latitude and longitude lines on the chart to read off the position fix, 45º52'S, 128º16'E.

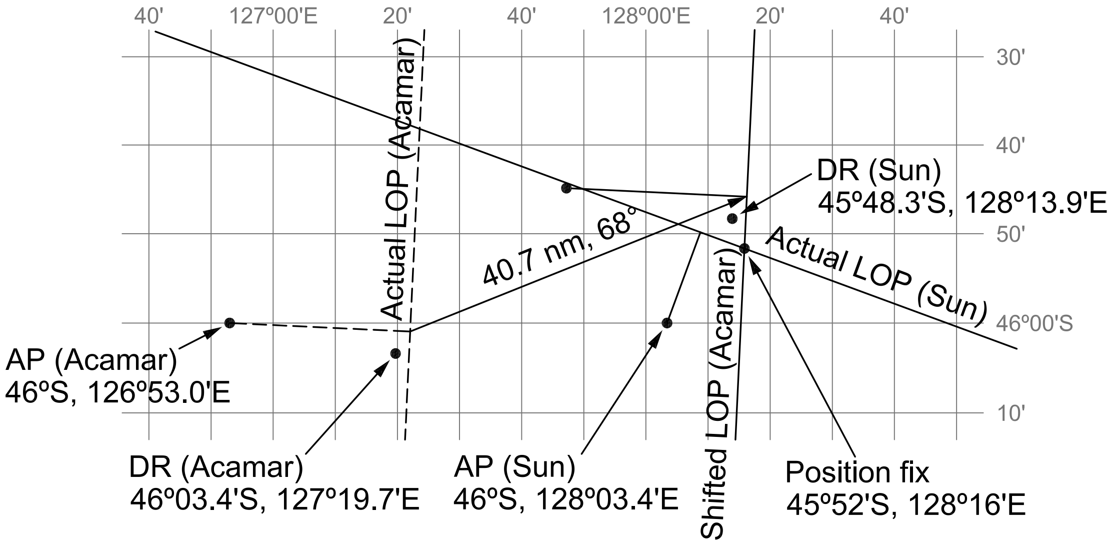
40'
127º00'E
20'
40'
128º00'E
20'
40'
30'
40'
50'
46º00'S
10'
Actual LOP (Acamar)
DR (Sun)
45º48.3'S, 128º13.9'E
40.7 nm, 68º
Actual LOP (Sun)
AP (Acamar)
46ºS, 126º53.0'E
DR (Acamar)
46º03.4'S, 127º19.7'E
AP (Sun)
46ºS, 128º03.4'E
Shifted LOP (Acamar)
Position fix
45º52'S, 128º16'E

Using a calculator

By now you should have noticed that using an almanac and sight reduction tables to calculate your position requires a sharp mind and meticulous attention to detail. The task could prove challenging on a small water craft bouncing around in the middle of the ocean. A celestial navigation calculator will greatly simplify your calculations, assuming salt water doesn't destroy the unit. For each sight, plug in the sextant altitude (Hs), index error, sextant elevation above sea level, universal time (UT), universal date (UD), dead reckoning (DR) position, and select the appropriate celestial body. The calculator may ask if any LOPs have been shifted and will then automatically do all the calculations and spit out your position.

Chapter 6
Natural Navigation

Our ancestors were incredibly competent at using natural clues to navigate. Enhanced by thousands of years of experience, their techniques became ever more dependable. The same principles still apply today, although mastering ancient methods requires considerable tutoring and practice. Aside from being used as backups, natural navigation methods can help you travel faster and more efficiently when used in conjunction with modern navigation instruments. Keep in mind that many natural clues are specific to certain regions and may not work in your area. Use as many natural clues as possible to increase your odds of being correct. The best way to learn something about natural navigation is to make observations as you enter an area.

6.1 Effects of sunlight

The direction that a slope faces determines how much sunlight it receives. In the northern hemisphere, south-facing slopes generally receive more sunlight than north-facing slopes. The converse is true in the southern hemisphere where north-facing slopes receive more sunlight. The spring snow melts slower and glaciers are more likely to be present on slopes that face away from the sun. The vegetation may be denser or sparser on slopes that face the prevailing sun, depending on whether the added sunlight enhances growth or dries out the soil.

Certain plant and wildlife species are more plentiful on slopes that face the sun, while others that need moisture may favor slopes that face away from the sun. Berries usually ripen earlier and flowers grow more plentiful in the direction that receives the most sunlight. Conversely, certain types of mosses or lichens that require moisture may grow faster on the side of a tree or boulder that receives less sunlight. Many types of flowers face the prevailing sun. Some flowers track the sun during daylight hours, even when it is overcast. Trees can indicate the cardinal directions by the way they grow. The bark is often thicker on the side facing away from the sun. The branches on some species of tree grow bigger and more horizontal in the direction of the sun. One of the more reliable ways to tell the direction from a tree is to look at a stump. Tree rings are usually closer together in the direction of the prevailing sun.

Unfortunately, these clues don’t always indicate a specific cardinal direction. Some species are less affected by sunlight than others. In many cases, the sunlight doesn’t reach certain plants because they are shaded by terrain features or other nearby plants. For instance, trees growing along the perimeter of a forest normally receive more sunlight than trees in the middle of the forest, regardless of the direction of the prevailing sun. Other factors that influence growth such as soil, winds, or precipitation rates can mask the effect of sunlight. Analyze a large sample of plants in the same area and make sure you’re completely familiar with the subject before drawing a definite conclusion.

6.2 Effects of wind

Differential heating of the atmosphere, together with the coriolis force of the spinning Earth, generates extensive bulk movement of air masses that creates enduring prevailing winds. The directions of the prevailing winds change with the seasons in some areas while in other regions they stay the same throughout the year. On land, the local topography often distorts the prevailing wind direction but surface winds nevertheless tend to blow from one preferred direction. From time to time, winds will blow from non-preferred directions.

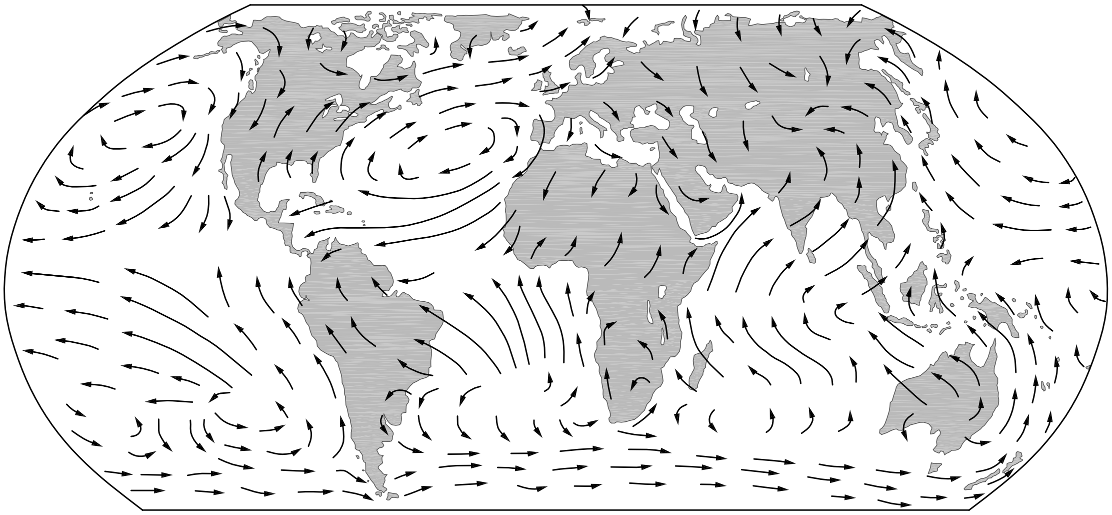

Prevailing winds - July

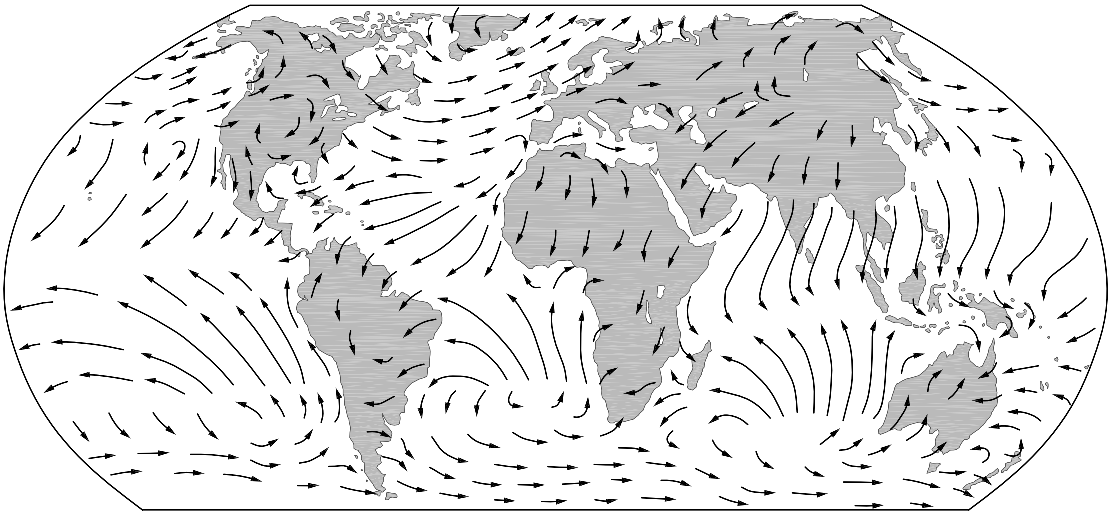

Prevailing winds - January

Knowing the direction of the prevailing wind in your area may help you determine the cardinal directions. If a high altitude wind is blowing, you can tell its direction by looking at the movement of high-level clouds. Surface winds often leave clues about their preferred direction even when they are not blowing. Persistent winds from a particular direction can modify the shape of trees by stunting the growth of foliage on the side facing the wind. In sand deserts, the wind molds the surface into sand dunes, and in similar fashion, the wind creates sastrugi (snow dunes) on snowfields. The most common type of sand dune or sastrugi is shaped like a ridge that runs parallel to the wind direction. This type of ridge has symmetrical sides. Another type looks like a horseshoe with its open end pointing away from the wind direction. A third type looks like sea waves with crests and troughs at right angles to the wind. Inhabitants of the arctic have used the shape of snow to travel in a straight line during whiteouts and desert nomads have used the shape of dunes to navigate during sand storms. The underlying principle is simple; just keep moving at a certain angle relative to the orientation of the sand dunes or sastrugi.

Snow cornices that form on mountain ridges give a clear hint of the preferred wind direction at that particular spot. They look like breaking sea waves that are frozen into position. Cornices are not necessarily useful for navigation because winds in mountainous terrain are often highly localized. The wind can spin around a mountain, creating cornices that point in several directions on the same mountain.

6.3 Landforms

In many parts of the world, mountain ranges, valleys, and rivers run parallel to each other. The easiest way to discover the landform pattern specific to your area is to look at a small-scale map or fly over the area. Take note of the cardinal directions of ranges and valleys and which way the rivers are flowing.

Repetitive landforms are a mixed blessing. By acting as guiding or catching features, the landforms can reveal your travel direction or tell you when you've reached a certain point. On the other hand, when all the landforms look alike, it may be difficult to tell one from another.

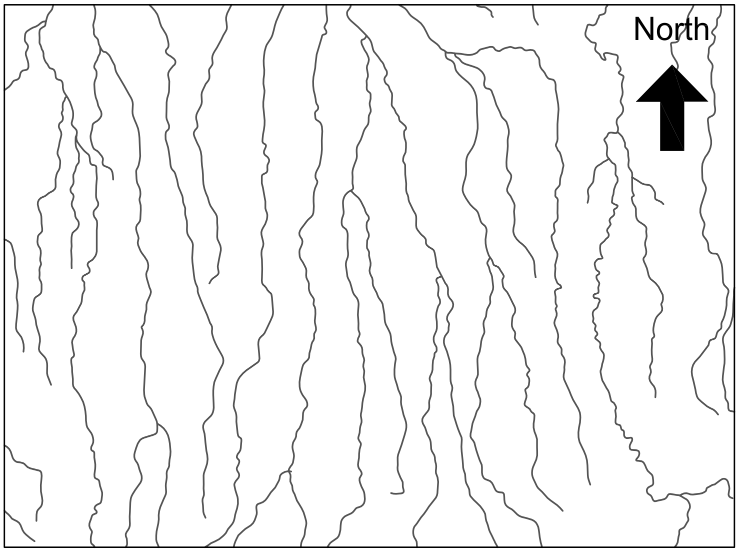

Example of parallel rivers

6.4 Ocean clues

The vast expanse of the open ocean with its lack of landmarks is an intimidating place for the untrained navigator. Yet, a lot of information can be gleaned by scrutinizing the surrounding sky and water. The unobstructed view on the ocean makes it an ideal place to use astronomical bodies to navigate, but there are other clues as well.

Ocean swells

Swells are different from locally generated waves. They are big waves that maintain their direction for a long time and are usually generated by prevailing winds. The heights of swells depend on the strength and duration of the generating wind, and the length of ocean over which the wind was blowing. The greater the distance from the generating wind, the longer the wavelength of the swells and the gentler the slope of the waves. Short wave lengths and steep slopes indicate that the generating wind is closer. Knowledge of general wind patterns gives you an indication of the general direction of the swells in your area.

Swells can persist for several days after the generating wind has stopped. This makes them well suited for maintaining your direction of travel on the open ocean. After pointing the bow of your vessel in the desired direction, determine the angle between the swells and your vessel and then maintain that angle to stay on course. A highly capable navigator can feel the direction of swells by the rocking motion of the vessel without looking outside. Close to shore, swell patterns are more complex and become difficult to predict.

Signs of land

There is often evidence of the direction and proximity to land long before it comes into view. Obvious signs include floating leaves, bark, branches, bushes, trees, or ice. Land can also literally be smelled. Clouds hovering in one location during an otherwise sunny day are an indication that there is land below. Local clouds commonly form over mountainous islands. The light of the sun or moon reflected off sandy beaches or smooth waters of sheltered lagoons can create a distinct glow above the land. Even during overcast days, light reflected from land can impart a different coloration to the clouds above.

Distorted wave patterns are often caused by waves reflecting back from land or squeezing between islands. Similarly, winds that suddenly change direction or turn gusty are an indication of nearby land. At night, the light from distant cities can be seen reflecting off clouds or smog.

Land-based seabirds are one of the most intriguing clues regarding the direction and distance to land. Some species of bird live on land but feed off the ocean. In typical fashion, these birds head out to sea in the morning to feed and return to land in the evening to rest. A navigator who spots a land-based bird in the morning may assume that land is in the direction that the bird is coming from, and in the evening simply follow the bird toward land. Potential changes to the birds' habits during nesting season must be taken into account. Instead of making one long trip per day, the birds may return to land several times to feed their young. Generally, the flight path of large groups of birds is a more reliable indicator of land than the movement of individual or small groups of birds.

Seamarks

The ocean isn't as uniform as many people think. Persistent warm or cold-water currents snake their way through water of different temperature or salinity, creating visible patterns for the observant navigator. You can estimate the water temperature by feeling it with your hand and determine its salinity by tasting the water. Nutrient rich water may look green compared to the bluish surrounding water. Light, shallow water stands out compared to dark, deeper water. Certain types of water attract certain types of sea life. As a result, some locations on the ocean attract high concentrations of whales, sharks, jellyfish, sea birds, etc. It is a matter of knowing what type of water or sea life to expect at a particular location.

6.5 Making a compass

A magnetized nail, sewing needle, or wire can be made into a compass needle by placing it on a piece of paper or wood floating in a container of water. Make sure that the water container isn't made of ferromagnetic material. Plastic or aluminum are good candidates. Provided the setup isn't influenced by wind, the needle will quickly align itself along Earth's magnetic field lines.

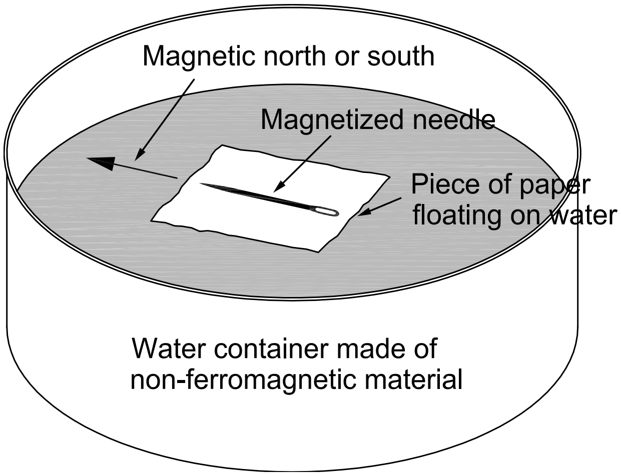
Magnetic north or south
Magnetized needle
Piece of paper floating on water
Water container made of non-ferromagnetic material

You will need additional information such as the direction of the sun to determine which end of the needle is pointing in the direction of the magnetic field lines (magnetic north). This setup is somewhat portable if you have a watertight lid on the container.

If the needle isn't magnetized, you may be able to magnetize it by repeatedly stroking the needle with a magnet. Always stroke with the same end of the magnet and toward the same end of the needle.

6.6 Using thermometer to determine elevation

The temperature of boiling water in an open pot depends on the ambient air pressure, which in turn is inversely proportional to your elevation. By boiling water in a pot and measuring the water temperature, you can calculate your approximate height above sea level. To do this, you need an accurate thermometer that can withstand being dipped in boiling water. At sea level, the boiling temperature of pure water is 100 ºC. For every 300 m of elevation gain, the boiling temperature decreases by 1 ºC. For a precise correlation, see Table 6.6.

Table 6.6
Correlation between elevation and boiling temperature of pure water

Temp.	Elevation	Temp.	Elevation
100.0 °C	0 m	85.0 °C	4500 m
98.3 °C	500 m	83.3 °C	5000 m
96.7 °C	1000 m	81.6 °C	5500 m
95.0 °C	1500 m	79.9 °C	6000 m
93.3 °C	2000 m	78.2 °C	6500 m
91.7 °C	2500 m	76.5 °C	7000 m
90.0 °C	3000 m	74.8 °C	7500 m
88.3 °C	3500 m	73.1 °C	8000 m
86.6 °C	4000 m	71.4 °C	8500 m

The accuracy of your reading will depend strongly on the accuracy of your thermometer and impurities in the water (don’t use soup). Weather systems and other factors that influence altimeter readings will also affect your result.

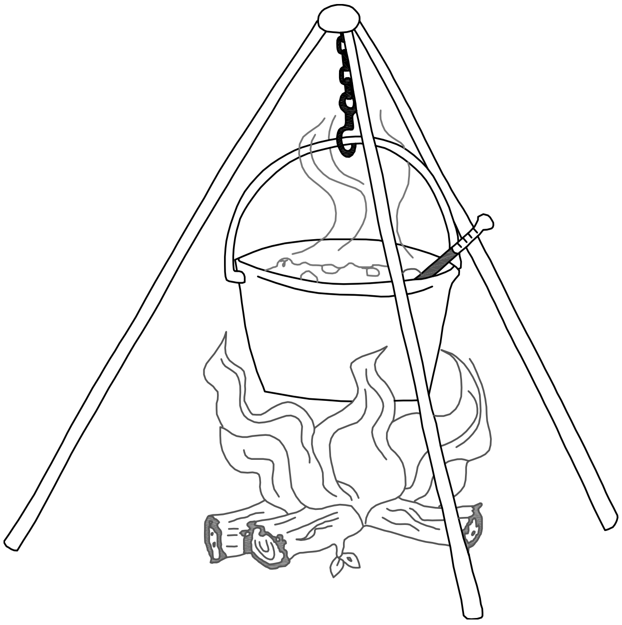

Chapter 7
Emergency Communication

To call for outside assistance, you have a choice of many types of electronic communication devices. The technology is evolving so rapidly that it is difficult to keep pace with all the changes. There is also the danger of trying to use obsolete equipment that relies on infrastructure that no longer exists. You can, of course, resort to low-tech systems such as mirrors, flashlights, whistles, or flares to try to attract someone's attention.

7.1 Emergency transmitters

7.1.1 ELTs and EPIRBs

Battery-operated devices that can remotely locate an aircraft or ship in distress have been around for many years. In aviation, the device is called an emergency locator transmitter (ELT). In marine applications, it is known as an emergency position indicating radio beacon (EPIRB). The impact of a crash automatically activates an ELT, while an EPIRB is activated after the vessel is submerged. The buoyant EPIRB floats to the surface and starts transmitting. Both these devices can also be switched on manually. The distress signal is transmitted using one or more of the international search and rescue frequencies. A nearby aircraft or vessel that is equipped with a suitable receiver can pick up the distress signal and call for a rescue operation. The rescue craft uses a directional antenna to home in on the source of the distress signal.

Satellites can also pick up a distress signal and have the added advantage of surveying large areas and continual monitoring. After receiving a distress signal, the satellite relays the information to a ground station where the rough position of the signal source is calculated from the doppler shift of the distress signal. The doppler shift is the change in frequency of the distress signal that is caused by the relative motion between the emergency transmitter and the satellite. The accuracy of the position derived from the doppler shift is typically a few kilometers.

The most accurate emergency transmitters are used in conjunction with a GPS receiver. The GPS receiver is either a separate unit connected to the emergency transmitter or built into the emergency transmitter. The device transmits the position acquired from the GPS receiver to the satellites. A ground station that monitors the satellites relays the position to a local rescue organization. The accuracy of the transmitted position is equal to the accuracy of the GPS position fix.

7.1.2 Personal locator beacon (PLB)

A personal locator beacon (PLB) is a compact, lightweight emergency transmitter that operates in the same manner as an ELT or EPIRB, except that it can only be activated manually. A PLB can be highly effective at finding wilderness travelers in distress, especially if the PLB is used in conjunction with a GPS receiver.

Because of the potential for abuse and the great costs associated with false alarms, some jurisdictions dispense severe penalties for misuse and/or require that you register your PLB with the appropriate authorities. Some PLBs transmit a digitally encoded distress signal that, aside from relaying your position, also gives the rescuers other information such as the name of the registered owner of the PLB and what family members or friends to contact in case of an emergency. Most PLBs have enough battery power to transmit continually for up to a few days but this is reduced in cold temperatures. To avoid confusing rescuers, do not switch off a PLB once it has been switched on.

Other less sophisticated types of locator beacons are similar to transmitters used to track wildlife. They transmit a weak pulse at regular intervals and are always switched on. The batteries can last for a year or more. The transmitted signal is, however, not a distress signal and will not trigger a rescue unless you have prearranged for someone to start searching for you by a certain date. Typically, a rescue aircraft detects the transmitted signal and then homes in on the source using a directional antenna.

7.2 Two-way radios

A huge industry has been built around wireless communications and many types of two-way radios have been developed. Two-way radios use electromagnetic waves of certain frequencies to transmit and receive voice signals. Each frequency is associated with a particular radio channel. Most of the battery power of a radio is used when the unit transmits a signal. Much less power is used when only receiving a signal and even less when the radio is on standby. Carry spare batteries and shut off your radio when it is not in use.

Radios are averse to dirt and moisture, especially salt water, so make sure to protect your radio from the elements. Learn how to use your radio and carefully consider its limitations before heading into the wilderness. Make sure that your two-way radio has coverage for the area that you intend to travel. A common problem is the inability to receive or transmit a message. Just moving a few meters away can sometimes greatly improve or worsen your reception. Other times, you may have to move away from terrain features such as ridges that obstruct the signal or move to higher ground. Some radios are equipped with a switch that lets you adjust the transmitting power. Use the highest power setting if you cannot get any reception, but keep in mind that the higher power will drain the batteries faster.

Most aircraft and ships use geographic coordinates to navigate. Therefore, state your position in latitude and longitude when communicating directly with an aircraft or ship. On the other hand, local rescue organizations may prefer that you use grid numbers for your position. Be prepared to use both systems during an emergency.

Range

The range is the maximum distance that a two-way radio can transmit or receive an intelligible signal from a base station or another radio. The range increases with a more powerful transmitter, a bigger antenna, and a lack of obstructions. Unfortunately for the wilderness traveler, powerful transmitters are heavy, bulky, and draw a lot of battery power. In general, the lower the frequency of the radio wave, the greater the potential range. This is offset somewhat by the fact that low frequency radios require large antennas that are less efficient than small antennas. The range can increase considerably when atmospheric conditions allow the radio waves to refract (bend) along the curvature of Earth, or reflect (bounce) off the Earth's ionosphere or the ground before reaching the recipient.

Refraction occurs when radio waves pass between air masses of different densities. In the lower atmosphere, refraction normally increases the range by about 15 percent beyond the visible horizon for high frequency radio waves. Abrupt changes in air density can occur at certain elevations in the atmosphere due to temperature inversions or a sharp moisture gradient. Under these conditions, radio waves sometimes refract by the same amount as the curvature of the Earth, allowing communication over very long distances, especially if the transmitting and receiving radio antennas are in the same layer of air.

The upper atmosphere consists of several layers of ionized (charged) particles, most of which are created by sunlight knocking off electrons from the particles. Collectively, the ionized layers are known as the ionosphere whose thickness, height, and strength varies with the time of day and season. Radio waves that hit the ionosphere at certain angles are reflected back to the ground, increasing the potential range of the radio. The reflection angles depend on the frequency of the radio wave and the strength of the ionosphere. The radio waves may also reflect back and forth between the ionosphere and the ground, further increasing the range. At distances beyond the range of the direct signal but too close for the first reflected signal, there will be a zone where no transmission is possible. For frequencies higher than a certain threshold, all the radio waves will penetrate or be absorbed by ionosphere and no reflection will take place.

During nighttime, the ionized particles in the lower atmosphere recombine with electrons at a faster rate than they are separated and this part of the ionosphere weakens. This enables radio waves to reach and bounce off higher layers, which extends the range of radio waves at night.

7.2.1 FRS radios (walkie-talkies)

FRS (family radio service) radios, also known as walkie-talkies, are line-of-sight devices that are cheap, light, and easy to use. Most units offer a choice of several channels (frequencies). Just push a button to talk to anyone within range that has a similar device switched to the same channel. Release the button to listen. FRS radios are not subject to calling fees and normally no license is required.

The units use the same frequency to transmit and receive, so you cannot talk and listen at the same time. As the name suggests, FRS radios are mainly used to communicate between members of a group. The low-power transmitters in FRS radios provide a maximum range of only 1 to 3 km, making the units unsuitable to call for distant outside help. Some units can tune into broadcast radio stations, a useful feature for getting weather forecasts

A serious drawback is the lack of a world standard for allowable radio frequencies. An FRS radio purchased in one country that allows a certain set of FRS frequencies may be illegal in another country that allows a different set of frequencies. Using an FRS radio in the wrong country could interfere with communications by that country's police, military, etc., and may be subject to harsh penalties. On the other hand, if you are deep in the wilderness and away from other people, your FRS radio signal is unlikely to interfere with anyone outside your group.

7.2.2 VHF/UHF radios

VHF (very high frequency) or UHF (ultra high frequency) radios are similar to FRS radios but are much more powerful. They are used by pilots, sailors, police, fire fighters, rescue teams, and by many other professionals and amateurs. In certain areas, they can also receive weather forecasts. The range is normally line of sight but can be extended considerably by repeater stations that retransmit the radio signals to other two-way radios or into telephone networks.

A huge bureaucracy has been created governing the use of VHF/UHF radios. Some channels are off limits to the public and some jurisdictions require that you get an operator license and/or register your radio. There are also regulations on what you can talk about on certain channels. Of particular interest to wilderness travelers, are the emergency channels that are specifically used to call for a rescue. For instance, if you want a VHF/UHF radio for ocean kayaking, make sure that your radio can access the marine emergency channel for the area that you intend to travel. Before purchasing or renting a VHF/UHF radio, find out what channels are available on the radio and who, if anyone, monitors these channels for your area of travel.

Emergency call procedure

During an emergency, you must decide whether to do a mayday or pan-pan call. Mayday calls are reserved for situations where there is immediate risk to life. For example, you are seriously injured, about to succumb to hypothermia, or clinging to an overturned kayak. Pan-pan calls indicate that you need help but are not in immediate danger. For example, you have a non-life threatening injury that prevents you from traveling, but have enough supplies to survive for several days.

1. Prepare ahead of time what you are going to say, then switch on your VHF or UHF radio. Select the emergency channel if there is one for your area. To avoid overriding other calls, wait for a gap in the transmission before calling.
2. Press the transmit button, say "mayday, mayday, mayday" or "pan-pan, pan-pan, pan-pan", and briefly state the nature of your emergency and your position (if known). Speak slowly and clearly and end the sentence with "over."

3. Release the transmit button and wait for a reply. If you don't get a response after about a minute, repeat the entire broadcast. Continue calling periodically, even if you get no answer. It's possible that your radio is transmitting but not receiving.
4. If you still cannot get a response, try calling on another channel and, if possible, move to a location with better reception. Keep in mind that the reception may be better at certain times of day or night.

7.2.3 HF radios

HF (high frequency) radios are capable of long distance communication, but the quality of the reception is governed by atmospheric conditions and the time of day. Although the range of the direct signal is limited to your line of sight, reflected signals can extend the range to several hundred kilometers. The unit can be both heavy and bulky. Two long wire antennas must be strung out before the radio can be used. For best reception, lay out the antennas as high as possible and at right angles to the radio that you are trying to reach. An HF radio may be equipped with several sets of antennas of different lengths. Each antenna set is designed to be used with a specific radio channel, so you must change the antennas when you switch channels. In general, the lower the channel frequency, the longer the corresponding antenna. At night, your HF radio may be subject to interference from distant radio stations. The main limitations of HF radios are the weight of the radio and the elaborate antenna setup. The emergency call procedure is the same as for VHF or UHF radios.

7.2.4 Cellular phones

Cellular phones are widely available and mainly used for mindless chatter, but have occasionally saved lives. Unlike FRS, VHF, UHF, or HF radios, cellular phones cannot communicate directly with each other. The call must first go through a base station transceiver that then relays the signal through a telephone network to other phones.

Both the phone and the base station use low-power transmitters, so the direct range of the signal is limited to only a few kilometers. This isn't a problem around most urban areas where extensive grids of base stations have been built. Your cellular phone simply switches from one base station to the next as you travel. The short range is deliberate and allows the reuse of the same frequency by non-adjacent base stations so that many cellular phones can be used at the same time without interference. A cellular phone is easy to use although the underlying technology is complicated. The signal is digitally compressed and coded and can simultaneously be sent over several channels, some of which may be shared with other callers. The system automatically handles billing of phone calls.

For your cellular phone to work, you must be within range of a base station that belongs to, or is affiliated with, your service provider. During an emergency, you may not be able to complete a call despite being within range of a base station. Before heading out into the wilderness, make sure that the phone's batteries are fully charged. Protect your phone from moisture and remember that your phone becomes useless as you move away from the base station grids.

7.2.5 Satellite phones

Satellite phones generally have very broad coverage but not necessarily global. Depending on your service provider, your satellite phone uses either a few geostationary satellites that hover over the same point above Earth's equator, or a large number of satellites in much lower orbits. The signal from your phone travels directly to one or more of the satellites where it is relayed to a ground station and into the telephone network. If your phone uses geostationary satellites, you must point the phone's antenna directly toward one of the satellites to get reception. This type of phone usually has a small, built-in compass and a booklet telling you where to aim the phone. Because geostationary satellites are located high (36 000 km) above Earth, the signal travels a considerable distance before reaching your recipient and vice-versa, and there is a noticeable time delay during conversations. If your phone uses low-orbit satellites, you normally only need to point your phone's antenna upward to lock onto one of the satellites and there is no noticeable time delay during conversations.

Both the transmitted and received signals are low-power and the phone probably won't work under heavy tree cover or in deep canyons. You need an unimpeded view of the sky for clear reception and to avoid being cut off during a conversation. Some satellite phones are able to use the cellular network. This type of phone will function just like a cellular phone when it's within range of a cellular base station. When you move beyond the cellular range, the phone automatically switches to direct communication with satellites.

7.2.6 Miscellaneous wireless devices

New types of wireless communications devices are being introduced at an unrelenting pace. The wireless signals could be transmitted via the cellular network or directly via satellites, or both. Any new type of wireless device would be subject to the same type of limitations that afflict cellular and satellite phones.

With digital technology, wireless devices can be designed to transmit and receive just about any kind of information. For example, with the proper equipment you could e-mail someone for help or activate someone's pager. Another possibility would be to use your digital camera to take pictures of your surroundings and then send the pictures over your wireless device to a rescue team. The rescuers could analyze the terrain in your pictures and then advise you how to proceed. The possibilities are as limitless as the ways things can go wrong.

7.3 Signaling systems

Signaling systems serve two main purposes: requesting assistance and guiding rescuers toward your location. Anything that attracts attention to your situation such as flashlights, whistles, mirrors, flares, or bright clothing can be used for signaling. For maximum effect, make sure your signal looks or sounds as different as possible from your surroundings. For example, if you're on snow use something dark to attract attention, or conversely, don’t use an acoustic device that sounds like a bird if you’re in a bird sanctuary. Do not depend on just one signaling system. Be prepared to use every conceivable method that you can think of.

One of your most important decisions is choosing when to deploy consumables such as flares. You are faced with the conflicting demands of taking advantage of an important opportunity versus conserving a valuable resource. Your decision should be based on the likelihood of anyone detecting your signal, recognizing it as a call for help, and then initiating a rescue. Set up your signaling systems ahead of time so that you can activate them at a moments notice when the opportunity arises. Rescuers may respond to your signals with their own signals. If that happens, repeat your distress signal to help rescuers home in on your location.

7.3.1 Flares

Two main types of emergency flares exist: ground flares and aerial flares. Ground flares either spray burning material or colored smoke out of one end of the flare. Colored-smoke flares are meant to be used during the day, while burning material flares work best at night. Aerial flares use a small rocket to shoot up a charge into the air. When the rocket fuel is spent, the charge ignites and starts glowing. A small parachute deploys and the glowing charge slowly descends to the ground. Aerial flares are either handheld or launched from specially designed flare guns. Most types of aerial flares are intended for marine use only because of the danger of starting a forest fire when the flare lands on the ground.

For distress purposes, your flares should emit either red or orange light or smoke, but don’t hesitate to use other colored flares if they are available. Flares are potentially dangerous devices that require proper operation and storage. Read the instructions before firing one off and make sure the flare is pointed in the correct direction. Moisture or age can cause flares to misfire.

7.3.2 Mirrors

In bright sunlight, a good mirror reflects an enormous amount of light that can be seen from far away. A mirror compass can be used for this purpose. Other reflective surfaces such as sunglasses, watch faces, or shiny metal surfaces do not reflect nearly the same amount of light as a glass mirror. The difficult part is aiming the sunlight toward your target. Specially designed signaling mirrors are equipped with a sighting apparatus that makes aiming the mirror much easier.

To aim an unsighted reflecting surface, close one eye and point the tip of your thumb toward your intended target. Hold the reflecting surface as close to your open eye as possible and orient the reflecting surface so that the reflected sunlight hits the tip of your thumb. This will direct the reflected sunlight toward your target. Wiggle the reflecting surface to direct a flickering signal at your target.

7.3.3 Light signals

You can use a flashlight or any other battery operated light source to send distress signals at night and even during the day. The standard emergency procedure is to use the light source to send an SOS signal. This is done in Morse code with a sequence of three short flashes, followed by three long flashes, and then three more short flashes. Aim your light toward a logical target and repeat the signal sequence as many times as necessary.

In general, strobe lights do not have the same range as flashlights, although a rapidly flashing light will certainly draw the attention of anyone who sees it. One advantage is that they can be left unattended for extended periods. A camera flash will also work as a signaling device. Another tool is a specially designed signal laser. The low dispersal of the laser beam means that the light carries over long distances. Signaling lasers must be oriented a certain way because the beam is typically projected as a line of light rather than a point. Because of the narrow beam, aim the laser directly at your target to attract attention. If your laser lacks a sighting apparatus, use the same technique as with a mirror to direct the laser beam at the target. Once the beam is aimed, move it very slowly back and forth across your target.

7.3.4 Fires

Fires can be an effective signaling method. The universal distress signal is three fires positioned 10 - 15 m apart and arranged in a triangle. Setting up signal fires requires a lot of preparation, especially if you are in an area that lacks much combustible material.

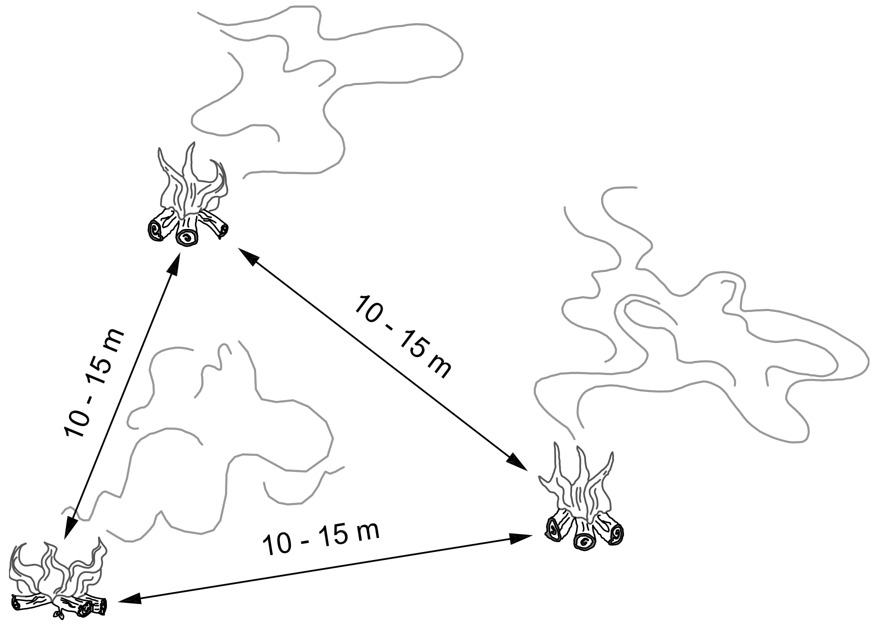

To avoid maintaining three continually burning fires, only light one of your three “fires” and set the others up for quick ignition. Prepare the unlit “fires” by gathering plenty of dry kindling into a pile and then adding some green vegetation to the top of the kindling. The green vegetation will ensure that the fire produces enough smoke to be visible from the air. If a low flying aircraft approaches, quickly ignite all your “fires”.

7.3.5 Acoustic devices

The sound of an acoustic device such as a whistle or an air horn carries much further in the wilderness than shouting. Acoustic devices are intended to catch the attention of anyone within audible range. The universal distress signal is the SOS sequence with three short blasts, followed by three long blasts, and three more short blasts repeated at regular intervals. A series of three blasts repeated over and over is also considered a distress signal. A whistle is a lightweight and durable device that should be carried by every wilderness traveler. Rain, snow, and heavy vegetation will reduce the acoustic range. Wind normally decreases the range but can in some instances increase the downwind range.

7.3.6 International ground signals

Aircrews that are trained in search and rescue techniques should understand the international ground signals. The aircraft can acknowledge your message by tipping its wings or flashing a green light. An aircraft that flashes a red light or flies in a clockwise circle may mean that your message is not understood. Some of your signals will be in response to questions from the aircraft. The aircrew can for example drop you a message from the aircraft or talk to you via a megaphone.

Ground symbols

Ground symbols can be stomped into the snow, drawn in the ground, written with a pattern of rocks, etc. Make the characters as large as possible so that they are clearly visible from the air. If you haven’t memorized any of these ground symbols, write out “RESCUE” or “SOS” instead.

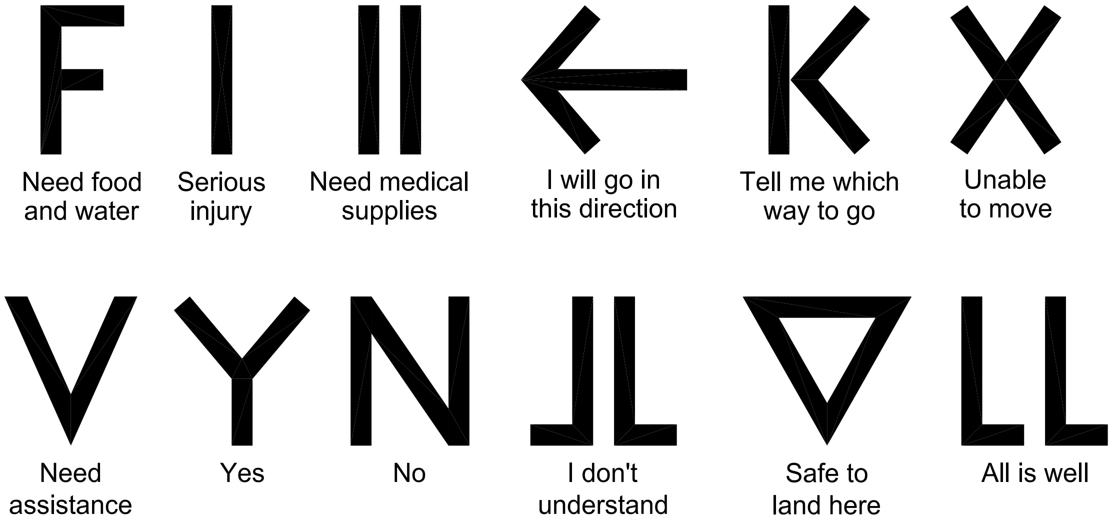
Need food and water
Serious injury
Need medical supplies
I will go in this direction
Tell me which way to go
Unable to move
Need assistance
Yes
No
I don't understand
Safe to land here
All is well

Body signals

Body signals have similar meanings as ground symbols except that you can respond instantly to any questions from the aircrew. For greatest effect, wear clothing that contrasts with the color of your background. For a signal that requires you to move your arms, hold a bright cloth in your moving hand and exaggerate the movement.

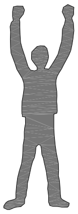

Pick us up

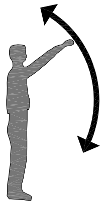

Yes

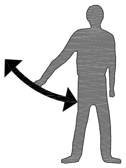

No

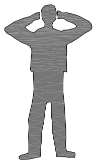

Have radio

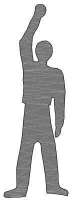

All is well

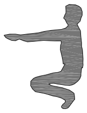

Land in this direction

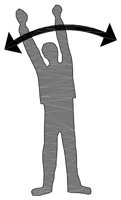

Do NOT attempt to land here

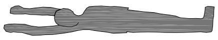

Need medical assistance

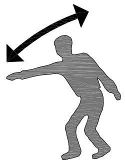

Use drop message

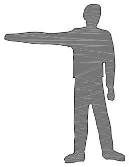

Can proceed shortly

Chapter 8
Practical Navigation

How you navigate is largely dictated by the availability of landmarks. If prominent landmarks are visible, you should be able to navigate with only a map, or in some cases, just your memory as a guide. When no landmarks are visible, you must complement your map with a compass and use techniques such as dead reckoning to estimate your position. For land navigation, an altimeter will often enhance your ability to establish your location. For long trips through featureless terrain, a position-finding device such as a GPS receiver or sextant is required.

8.1 Terrain association – map only

Navigating by terrain association requires prominent landmarks, a certain amount of visibility, and a map if the terrain is unfamiliar. Terrain association is based on building a mental image of the terrain. The surrounding landmarks and linear features guide your movement. Linear features are long, thin features that can be straight or curved. They include valley bottoms, canyons, rivers, coastlines, ridges, power lines, and roads.

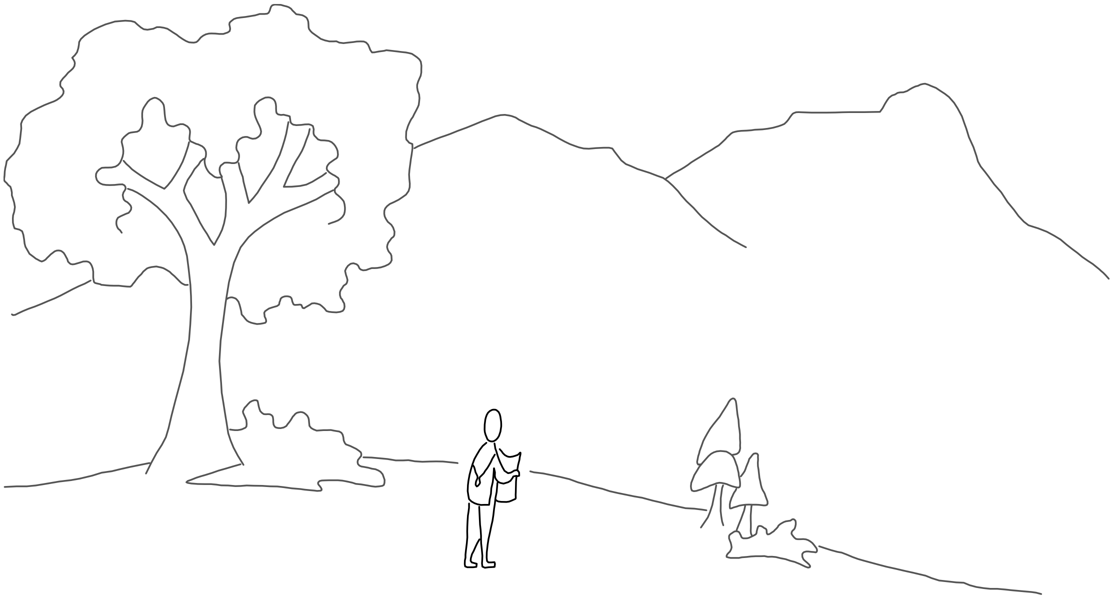

You can usually find your way without too much effort with terrain association, provided a significant portion of the terrain is visible. It is a natural way of navigating; it's how we subconsciously do it in everyday life. If you plan to travel through an unfamiliar area, try building an initial mental image of the area ahead of time by studying maps and pictures. As you travel through the new area, your mental image is continually modified by what you see in the real world. The key to success is your ability to recognize certain critical features on both the map and in the real world.

Here are a few important considerations:

- Orient the map. It’s much easier to compare the map with the real world when the map is oriented. Use prominent landmarks or the sun to orient your map.
- To identify features, look at the shape and steepness of slopes, the relative heights and directions between features, and the directions and relative elevations between you and the features. Look for any other clues such as vegetation zones or the direction of running water.
- Do mental triangulations off visible landmarks to estimate your position on the map.
- Keep track of which direction you’re heading and estimate the distance you’ve traveled since your last known position. To keep yourself moving in the correct direction, adjust your travel direction relative to distant landmarks or follow linear features that are parallel to your travel direction. Don’t forget to look at the landmarks behind you.
- Select catching features to alert you when to change direction or if you’ve gone too far. Catching features are linear features that cross your line of travel or landmarks whose nearby presence you’re unlikely to miss.

Navigational errors are easily spotted with terrain association. If you encounter something in the real world that contradicts what you expected from the map, you immediately know that you are not were you thought you were. There is, of course, the possibility that your map is incorrect. Terrain association requires lots of practice to master. It also lacks the precision needed for finding exact locations and is difficult to use when the visibility is poor.

Example of terrain association

With the map below as your only navigation tool, you intend to hike from (A) to a campsite by a lake at (D).

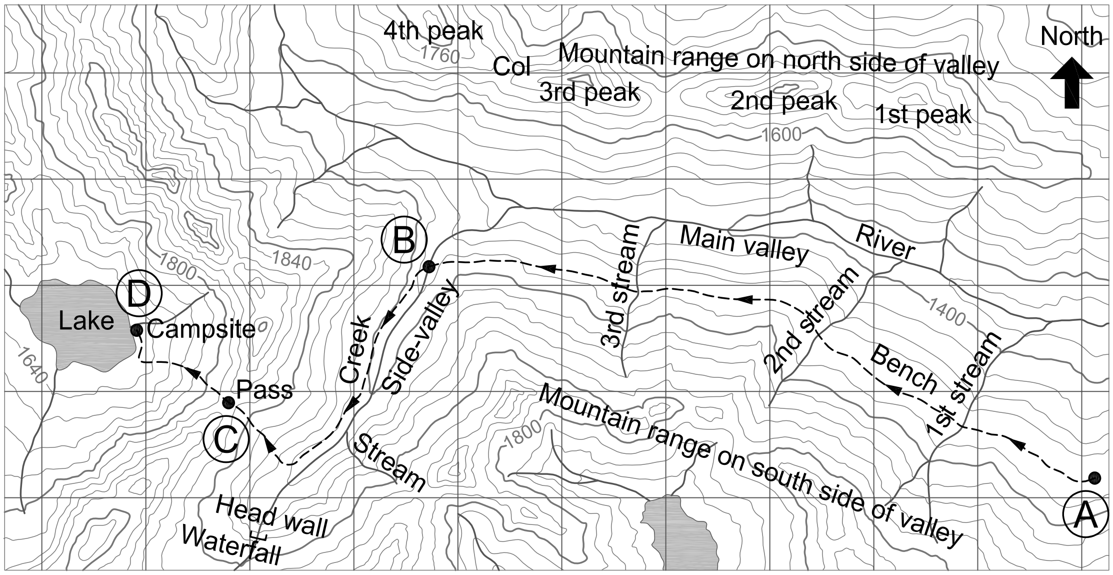
4th peak
1760
Col
Mountain range on north side of valley
North
3rd peak
2nd peak
1st peak
1600
B
Main valley
River
3rd stream
2nd stream
1400
Bench
1st stream
A
D
Lake
Campsite
1800
1840
1640
Creek
Side-valley
Pass
C
Stream
1800
Mountain range on south side of valley
Head wall
Waterfall

The first part of your journey follows a bench that runs along the south side of an east-west tending valley. A bench is a flat section on a slope. A mountain range runs along each side of the valley. A river flows along the bottom of the valley with several small streams tumbling down into the river from the flanks of the mountains. Knowing which direction to travel on this part of the route is easy. Just travel parallel to the river. If you drift too far to the right or left, you will run into steeper ground.

To figure out of how far you've traveled along the bench, pay attention to the surrounding features. Counting the number of streams that you cross may seem like an obvious method but all streams aren't necessarily marked on your map. A more dependable method is to keep track of the peaks that rise along the sides of the valley. The tops of the peaks on the south side of the valley are probably hidden from view because of your location on the same side of the valley. You are more likely to have an unobstructed view of the peaks on the opposite, north side of the valley.

Your first catching feature is the creek that drains the first major side-valley that you encounter to the south. This creek is bigger than the other side-streams because it drains a larger area. The side-valley is also much wider than the gullies that drain the other streams and you may be able spot the headwall and waterfall to the south. To confirm that you've reached the correct creek, look at the peaks on the north side of the main valley. You should be opposite from the col that separates the 3rd peak from the 4th peak.

After crossing the creek at (B), head up the side-valley. The navigation here is easy. Just follow the creek upstream. The next navigation challenge is finding the pass at (C). Pay close attention to the peaks on both sides of the valley. There is no obvious catching feature that tells you where to start your upward climb toward the pass. If you reach the headwall with the waterfall, you have gone too far. Assuming that each grid square represents 1 km x 1 km, the map shows that you must head upward toward the pass about 1 km before reaching the headwall. This location is about halfway between the headwall and a major stream that runs down the east side of the side-valley.

If you start climbing at the correct location, the terrain will funnel you toward the pass (C). After reaching the pass, head downhill toward the center of the lake. The lake is at the bottom of a basin and any streams that you encounter in this area will run into the lake. Keep this in mind in case the fog rolls in and you suddenly lose sight of the lake. In addition, heading down the fall line (steepest slope direction) will bring you directly to the lake. Turn right at the lake and follow the shoreline. The campsite should be near the first major stream flowing into the lake (D). Do mental triangulations off the peaks that surround the basin to confirm your location.

8.2 Conventional map and compass navigation

Prominent landmarks and linear features are often blocked from view by terrain, foliage, or fog. Other lesser features that are visible can be difficult to identify if they are too plentiful, or too small to show up on your map. Under these circumstances, it can be difficult to travel in the correct direction without a compass.

A compass allows you to travel in the correct direction even when you cannot see anything except your immediate surroundings. It also allows you to pinpoint your position by triangulation whenever landmarks are visible. The following steps describe conventional map and compass navigation. If necessary, review Chapter 2 for detailed information on how to use a compass.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north. If not, add or subtract the declination after taking each map or field bearing.
2. If you don't already know where you are on the map, use terrain association or compass triangulation to find your position. Do this before it gets dark or the fog rolls in.
3. Plot a proposed route on the map from your current position to your desired destination. If necessary, break up the route into a series of straight-line segments. The junctions between segments are called waypoints, and are intermediate targets that you will reach during your journey.
4. Take a bearing off the map, from your current position to the next waypoint along your route. If the waypoint is located along a catching feature, add or subtract a deliberate offset to your compass bearing. The offset bearing will guide you toward the catching feature rather than directly toward the waypoint. The catching feature is usually much easier to find than the waypoint. Don't use an offset bearing if there is no catching feature.
5. Follow the compass bearing toward the catching feature or waypoint. Use steering marks to avoid constantly looking at your compass. The further you travel along the bearing, the more likely you are to drift off course, so take the opportunity to pinpoint your position by triangulation or terrain association whenever suitable landmarks appear.
6. Once you reach the catching feature, the direction of the offset bearing will govern whether you go right or left along the catching feature to reach the waypoint. If you've been following a non-offset bearing, you will hopefully stumble right onto the waypoint.
7. Once you reach the first waypoint, repeat the above steps for subsequent waypoints along your route.

Example of map and compass navigation

You are somewhere along the shoreline of a lake (A) and intend to hike over a mountain range to the outlet of a second lake (F). After analyzing your map, you decide to travel to your destination via two passes, (C) and (D), and a creek fork (E). To shorten your first segment, you will triangulate your position at the first creek crossing (B).

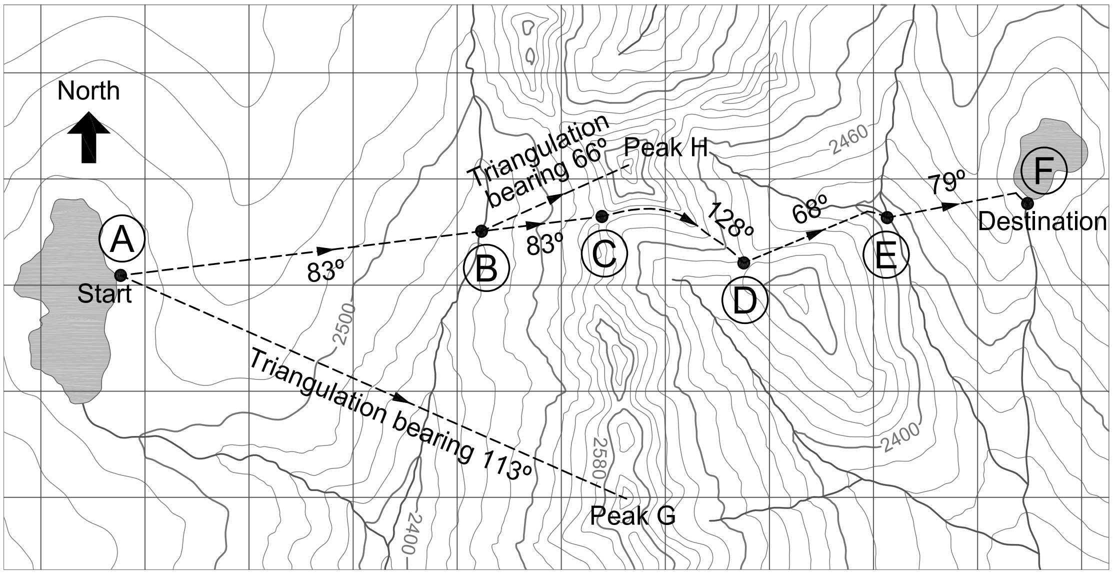

Adjust the declination on your compass so that it conforms to grid north. To triangulate your initial position, take a field bearing off Peak G, a convenient landmark that is visible from the lake. The result is a bearing of 113º whose line of position intersects the lakeshore at (A), your current position. Taking a bearing off the map from (A) to (C) gives a bearing of 83º. Follow this bearing to the creek. From the creek bed, a field bearing off Peak H gives 66º. The line of position from this bearing intersects the creek at (B), which becomes your triangulated position.

A map bearing from (B) to (C) gives 83º, confirming that you've maintained the correct course from (A). Continue along this bearing. As you climb higher, the terrain will funnel you toward the first pass at (C). After reaching the pass, maintain a constant elevation as your contour along the south slopes of Peak H to its southeast ridge. Follow the ridge downward along a bearing of 128º to the pass at (D).

From the pass at (D), take a map bearing to the creek fork (E). The result is 73º. Subtracting 5º gives an offset bearing of 68º. Follow the offset bearing to the creek. Take a right at the creek and follow it downstream to the fork.

The last bearing (84º) is toward your final destination (F). Here you have a choice of two catching features, the lake, or the creek that flows out of it. The lake is a better choice because the creek could potentially be confused with a stream not marked on your map. Subtract 5º for an offset bearing of 79º and follow this bearing to the lake. Take a right at the lake and follow the lakeshore to your destination at the outlet (F).

8.3 Dead reckoning – no landmarks

The word “dead” in dead reckoning may have been derived from its original meaning “deduced”. In this book, dead reckoning is defined as navigating with just a map (or chart) and a compass when no landmarks are visible. The underlying assumption is that if you start from a known position, and are diligent about recording how far and in what direction you’ve traveled, you can figure out where you are. The main application of dead reckoning is for travel across large open spaces such as deserts, tundra, polar plateaus, or oceans, often in conjunction with celestial navigation. Dead reckoning is also used for travel through ordinary terrain during periods of reduced visibility or through thick undergrowth.

Although it has been used successfully by highly competent navigators, dead reckoning is inherently unreliable because it is difficult to maintain a straight course and judge how far you’ve traveled. At sea, you must compensate for the drift caused by winds and currents. The further you travel, the more the errors accumulate, particularly if you’re forced to veer off a straight-line course. All it takes is one mistake and you’re lost. Navigating by dead reckoning involves three interrelated activities: (1) keeping track of bearings, (2) keeping track of distance, and (3) detouring around obstacles.

8.3.1 Keeping track of bearings

Keeping track of bearings is practical only if you travel along straight-line segments. Any curved paths will greatly complicate your task and should be avoided. Unless obstacles block your route, stick to rigid straight-line segments by following a specific compass bearing to the end of each segment. Diligently record the bearings for all the segments that you have completed.

8.3.2 Keeping track of distance

Figuring out the distance traveled along each segment is the most difficult part of dead reckoning. Your mode of travel will largely dictate what distance measuring techniques you can use.

- If you’re hiking, multiply your average stride length by the number of steps that you’ve taken. To calculate your average stride length, count how many steps it takes you to complete a known distance, then divide the known distance by the number of steps. While traveling, move pebbles or matchsticks from one pocket to another to keep track of your step-count. For example, move one pebble from one pocket to the other for every 100 steps. Electronic step counters are available that strap to your hip and count your steps by sensing the movement of your hip. The unit displays your travel distance based on your stride length. Read the instruction manual for procedures on how to calibrate the step counter so that it conforms to your stride. Keep in mind that different types of terrain will substantially alter your stride length.

- If you're traveling roped up on snow, use the rope length to measure the distance. The first person on the rope draws a line in the snow. When the last person on the rope reaches this line, he or she signals to the first person to draw another line in the snow. As this process is repeated, the last person keeps track of the number of lines. Multiply the number of lines by the length of rope between the first and last person to calculate the distance.
- An odometer wheel can be towed along or attached to a sled or cart. As you travel, the odometer counts the number of revolutions of the wheel. To calculate the distance, measure the circumference of the wheel and then multiply the circumference by the number of revolutions. This method has been used successfully on polar sledding expeditions.
- If you know your average travel speed, you can calculate your distance by keeping track of your travel time. On land, judge your speed from previous travel through similar types of terrain. On the water, measure the time it takes for the length of your boat to travel past a wave top or piece of driftwood. Divide the length of your boat by the time, to obtain the speed of your boat. To compensate for drift caused by wind and waves, look at the wake of your boat. Your travel direction is in line with the wake, not the direction where the bow is pointing. Compensating for currents is very difficult on the open ocean but relatively easy on rivers. Once you've estimated your speed, multiply your speed by your travel time to obtain the distance.

8.3.3 Detouring around obstacles

Circumventing obstacles while moving along a dead reckoning segment must be done without losing your distance measurement tally. Follow detour segments that are either perpendicular or parallel to your original travel direction. Instead of readjusting your compass bearing, travel at right angles to the travel direction arrows on your compass while following a perpendicular detour segment.

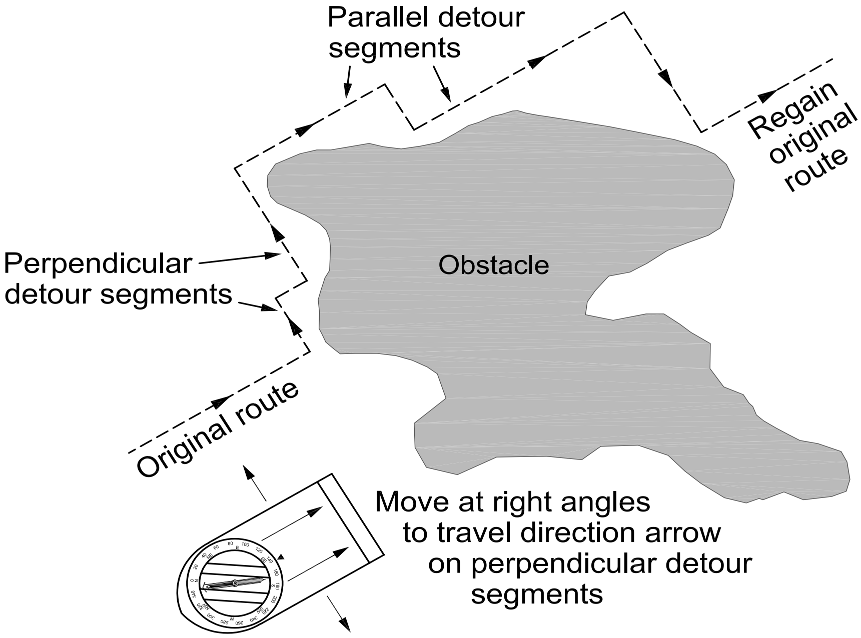
Parallel detour segments
Regain original route
Perpendicular detour segments
Obstacle
Original route
Move at right angles to travel direction arrow on perpendicular detour segments

You must keep track of two separate distances during the detour. Add the distances of the parallel detour segments to your original route tally. Either add or subtract the distances of the perpendicular detour segments, depending on whether you're moving away or toward your original route. The total perpendicular distance of the detour segments will cancel out when you regain your original route on the other side of the obstacle.

You can avoid the perpendicular count if there is a steering mark in line with your original route on either side of the obstacle. Just keep track of the distance along the parallel detour segments during the detour. You will regain the original route when you reach the steering mark on the far side of the obstacle. If the steering mark is on the near side of the obstacle, use a back bearing to regain your route. If the obstacle is big enough to show up on your map, pinpoint your position by terrain association and then use the obstacle as a starting point for a new dead reckoning segment.

8.3.4 How to navigate with dead reckoning

There are two main ways to use dead reckoning: (1) plotting your route ahead of time on the map and then following that route in the real world, or (2) using dead reckoning to estimate your position after you have already traveled.

Following a predetermined dead reckoning route

1. Determine the location of your starting point on the map by using terrain association or triangulation. Dead reckoning will not work without a known starting point.
2. Plot your route as a series of straight-line segments that follow your desired line of travel. Keep your route as simple as possible by using the minimum number of segments. Label as waypoints the connection points between the segments.
3. Take compass bearings off the map between all adjacent waypoints along your route. Write down the bearings and note whether the bearings are relative to true north, grid north, or magnetic north.
4. Measure the distance of each segment with a graduated ruler. Use the map scale to calculate the real world distance of each segment. Write down the real world distances.
5. Now you're ready to navigate. Set your compass to the bearing between the first two waypoints. If necessary, readjust the bearing to take into account the magnetic declination. Follow the compass bearing while keeping track of the distance traveled along the segment. To stay on course without constantly looking at your compass, use steering marks or natural clues such as the direction of your shadow. If you're on a large body of water, adjust your boat's heading to compensate for winds and currents. This will affect your speed toward the waypoint.
6. Once your distance tally shows that you've reached the first waypoint, repeat the procedure for subsequent segments until you reach your destination. If you inadvertently travel past a waypoint, you can re-plot your course rather than backtrack to the missed waypoint.

Example of dead reckoning along a predetermined route

Prior to this trip, you calculated the number of steps it takes you to cover 100 m by walking three laps around a standard 400-m running track. You require 1837 steps to cover the 1200 m distance. Your average number of steps to cover 100 m is therefore 1837/12 = 153 steps/100 m. This number reflects the nominal number of steps you need to cover 100 m on hard, flat terrain with no obstructions. In the wilderness, your actual number steps per 100 m will vary depending on how much weight you are carrying, the slope angle, how tired you are, the vegetation density, the snow pack, and many other factors.

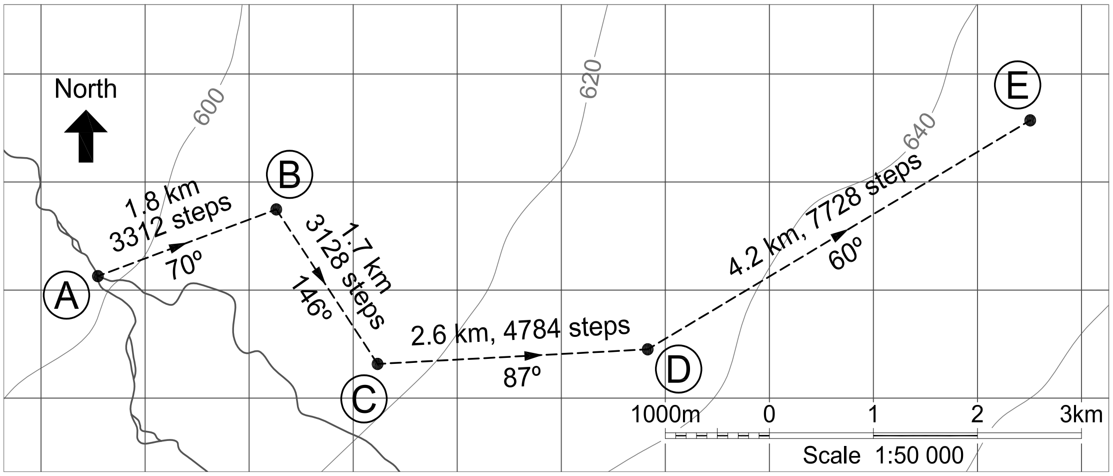

Your starting position on this journey is a creek fork that is easy to identify. You label this position as (A), and plot your route on the map as a sequence of straight-line segments with additional waypoints at (B), (C), (D), and (E). Based on the type of terrain, you decide to increase your nominal step count by 20% for all segments. Your step count becomes 1.2 x 153 = 184 steps/100 m.

With a ruler, measure the length of each segment on the map. The segment distances are 36, 34, 52, and 84 mm. Multiply these numbers by the map scale (1:50 000) to get real world distances. For example, the distance of the first segment is 36 mm x 50 000 = 1 800 000 mm = 1.8 km. Convert this distance to steps: 1.8 km = 1800 m x (184 steps/100 m) = 3312 steps. Calculate and write down the number of steps for each segment. Adjust the declination on your compass so that it conforms to grid north. Take compass bearings off the map between all adjacent waypoints and write down the bearings.

At your starting point (A), set the compass bearing to 70º. Follow this bearing while counting steps. If you encounter obstacles, use detour techniques to avoid losing your step count. Stop when the step count reaches 3312.

At (B), set your compass to 146º, and follow this bearing for 3128 steps to (C). Continue to (D) by following 87º and a count of 4784, and then to your final destination (E) with a bearing of 60º and 7728 steps. The further you travel the more inaccurate your position estimate will be. An excellent method for testing your dead reckoning ability is to retrace your route by following reverse bearings to your start point.

Estimating your position by dead reckoning

Sometimes you cannot plot your route ahead of time because the route is flexible or uncertain. Starting from a known position, keep track of your bearings and travel distances to estimate your dead reckoning position. This is a standard step in sextant navigation. On the ocean, currents and winds will distort your actual speed and bearing. Use mathematical or graphical methods to calculate your dead reckoning position. Graphical methods only work with charts or maps made from conformal projections where the scale at any one point is the same in all directions.

Example of estimating your position by dead reckoning

During a transoceanic sailing trip, you take two sights with a sextant and calculate your position as 47º01'N, 139º08'W. From here, you point the bow of your boat toward a magnetic bearing of 107º for six hours, and then want to establish your dead reckoning position. The magnetic declination is 18ºE, and you are using a conformal Mercator chart.

The estimated average forward speed of your boat is 5 knots. By looking at the wake of your boat, you calculate that a steady wind from the southwest causes your boat to drift at a rate of 1 knot in the wind direction (27º magnetic). There is also an ocean current pushing you eastward (72º magnetic) by 1 knot. One knot is defined as one nautical mile per hour. Use the bar scale to find out how long one nautical mile (nm) is on your chart. You could also do this without a bar scale because 1 nm is nearly equivalent to 1' of latitude. On your chart, 1 nm is one fifth of the distance between the latitudes marked 50' and 55', for example.

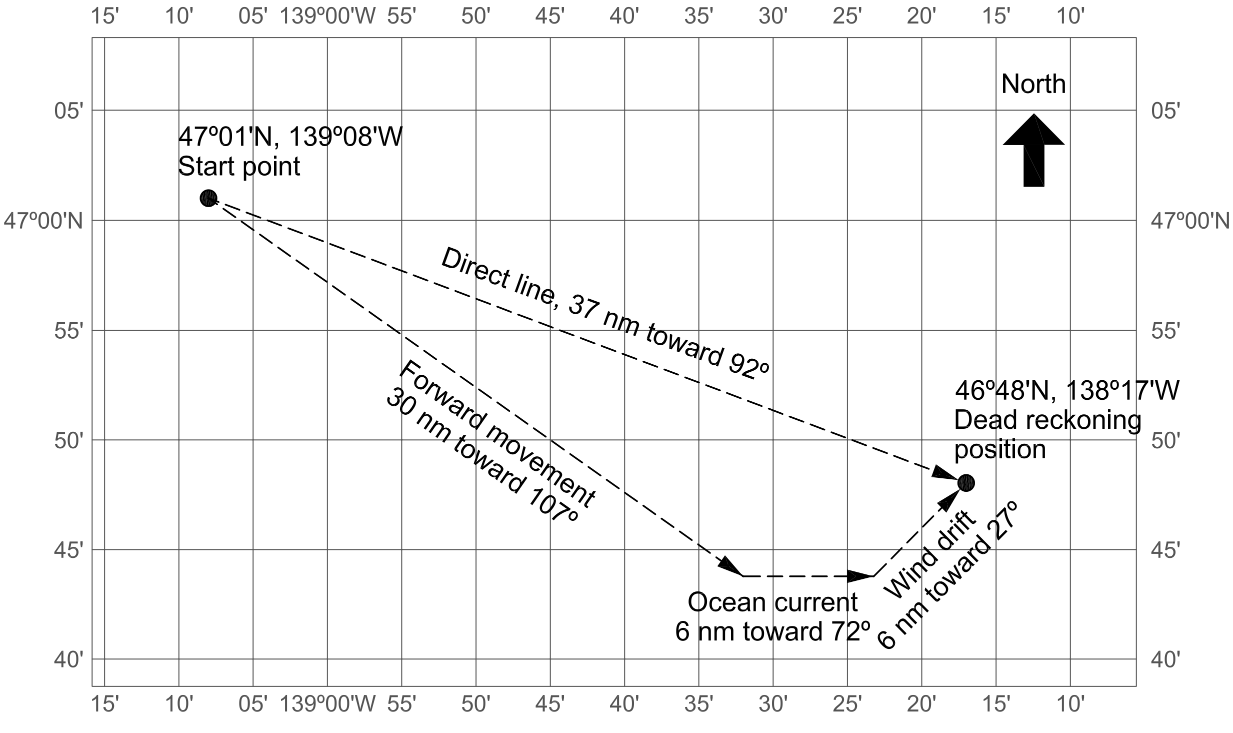
15'
10'
05'
139º00'W
55'
50'
45'
40'
35'
30'
25'
20'
15'
10'
North
05'
47º00'N
55'
50'
45'
40'
47º01'N, 139º08'W
Start point
Direct line, 37 nm toward 92º
Forward movement
30 nm toward 107º
46º48'N, 138º17'W
Dead reckoning
position
Ocean current
6 nm toward 72º
Wind drift
6 nm toward 27º

Your net movement from the last known position is a superposition of the movements caused by the boat's forward speed, the ocean current, and the wind drift. After six hours, your boat has moved a distance of 30 nm toward a magnetic bearing of 107º, 6 nm toward 72º (east), and 6 nm toward 27º (northeast). With a plotting tool, draw a line, 30 nm long, from your start point toward a magnetic bearing of 107º. Draw a second line, 6 nm long, from the end of the first line toward 72º, and finally a third line, 6 nm long, from the end of the second line toward 27º. Your dead reckoning position is at the end of the third line. Reading this point off the chart gives a position of 46º48'N, 138º17'W.

For future reference, calculate your actual bearing and ground speed from the start point. On the chart, measure the length of the direct line between your start point and your dead reckoning position. Converting to real world units gives a true distance of 37 nm. Dividing this distance by 6 hours gives a true speed of 6.2 knots. Using the compass rose to measure the bearing of the direct line gives a magnetic bearing of 92º.

8.4 Navigating with map, compass, and altimeter

On land, an altimeter can significantly increase the navigation power of a map and compass. Your movement is guided by conventional map and compass techniques, with the altimeter tracking your progress and helping you pinpoint your location. Take into account all the factors that can distort your altimeter reading (Section 3.3). To keep your altimeter as accurate as possible, calibrate it every time you come across a known elevation on your route.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north. If not, add or subtract the declination after taking each map or field bearing.
2. Find your present location on the map. You may be able to do this by triangulating your position with your altimeter and compass (Section 3.4).
3. Plot a proposed route on your map and mark all critical locations as waypoints. If possible, choose catching features that will help you find the waypoints.
4. Take a compass bearing off the map from your present location to the first waypoint. If your first waypoint is on a catching feature, add or subtract a deliberate offset to your bearing.
5. Follow the compass bearing toward your target. Alternatively, you may be able to follow a linear feature. In the latter case, the compass bearing is only used to confirm that you're moving in the correct general direction. If you're traveling up or down a major slope, track your progress by matching the contour lines on the map with your altimeter reading.
6. When you reach the catching feature, follow it to your first waypoint. An altimeter reading will help you verify that you have reached the waypoint.
7. Repeat Step 4 - 6 for subsequent waypoints.

Example of map, compass, and altimeter navigation

You intend to climb and traverse a nearby mountain from a campsite by a lake (A). Your proposed route goes up a gully (B) to the mountain's south ridge (C), which you will follow to the summit (D). Your descent will follow the mountain's north ridge to a fork (E), and then along the northwest ridge to the top of a gully (F). After descending to the bottom of the gully (G), the route goes back to your campsite (A).

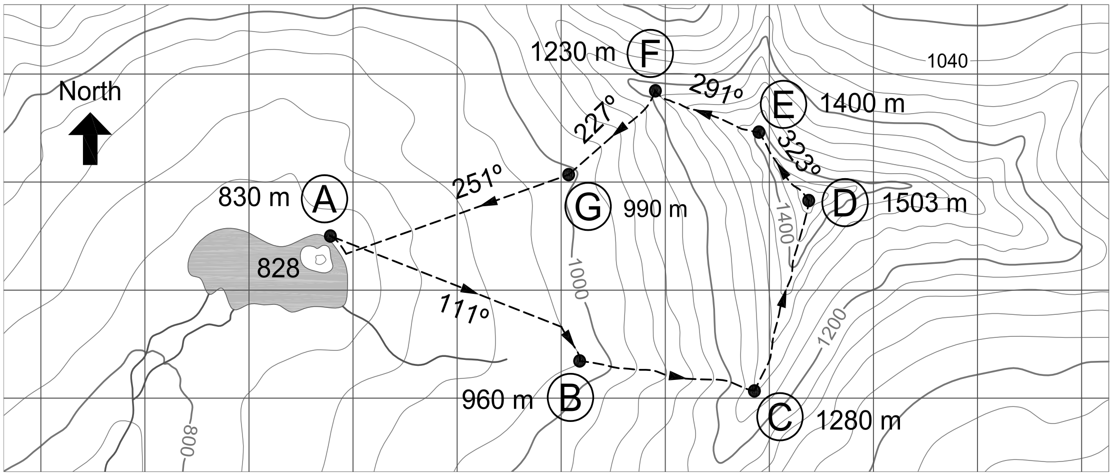

Pinpoint your campsite on the map. This is easy because you are right by the lakeshore and there is an island across from your camp. According to your map, the lake elevation is 828 m. Since your campsite is 2 m above the lake, calibrate your altimeter to 830 m.

Adjust the declination on your compass so that it conforms to grid north. Take a compass bearing off the map from the campsite (A) to the bottom of the gully (B). This gives a bearing of 116º. Subtract a deliberate offset of 5º for an adjusted bearing of 111º to ensure that you end up slightly left of the gully bottom. Your catching feature is the 960 m contour line, the elevation of (B). Follow the 111º bearing until your altimeter reads 960 m. Turn right and travel at a constant elevation to the bottom of the gully (B).

Climb up the gully to the ridge crest, which is at an elevation of 1280 m (C). Follow the ridge to the summit (D). As you climb the ridge, keep track of your progress by checking your altimeter. At the summit, calibrate your altimeter to read 1503 m. Take a bearing off the map from the summit (D) toward the ridge fork (E). As you leave the summit, follow this bearing (323º) to ensure that you are about to go down the correct ridge. Carefully monitor your altimeter as you descend. Keep an eye out for the ridge fork when you approach the 1400-m level.

When you reach the fork (E), take a bearing off the map toward (F). This bearing (291º) should line up with the general direction of the northwest ridge. Descend the northwest ridge. Scout out the terrain to your left as you approach the 1230-m level to ensure that you don't overshoot the gully (F). A bearing from (F) to (G) gives 227º. Use this bearing to confirm that you're about to go down the correct gully. Down-climb the gully to its bottom at 990 m (G). From here, take a bearing to your campsite (256º). Subtract a deliberate offset of 5º for an adjusted bearing of 251º. This will ensure that you hit the lake to the left of your campsite. When you reach the lake, turn right and follow the shoreline to your campsite (A).

8.5 Navigating with map, compass, and GPS

A map, compass, and GPS receiver is a powerful combination. It allows you to navigate through featureless terrain and during poor visibility. With a GPS receiver, you can follow a direct line to a waypoint without worrying about catching features. There are two major techniques: (1) pre-entering waypoints into the receiver, and then navigating along the route with the receiver and compass, or (2) using the receiver to pinpoint your position on the map, and then navigating with conventional map and compass techniques.

Expect some loss of precision whenever you transfer a waypoint from a GPS receiver to a map or vice versa. Normally, the most accurate waypoints in the receiver's memory are the ones acquired from satellite fixes. Consequently, the most accurate bearings displayed by the receiver are those between two satellite-acquired waypoints. A bearing between one manually-entered and one satellite-acquired waypoint is less accurate. The least accurate bearings are between two manually entered waypoints, which are as accurate as a compass bearing taken off a map.

8.5.1 Navigating with pre-entered waypoints

With pre-entered waypoints, just follow the route that you've already stored in the GPS receiver. The receiver tells you the bearing and distance to the next waypoint. The compass points you in the direction of this bearing.

1. If your compass is equipped with an adjustable magnetic declination mechanism, adjust the declination so that it conforms to grid north, and set the GPS receiver to display grid bearings. If your compass lacks an adjustable declination mechanism, set the receiver to display magnetic bearings.
2. At the start of your trip, switch on the receiver and wait for a position fix. If the receiver fails to obtain a fix within a few minutes, the satellite signals are probably blocked by obstructions. Move to a spot where more sky is visible.
3. Make the receiver display the bearing and distance to the first pre-entered waypoint along the route. Turn the graduated dial of the compass to this bearing and switch off the receiver.
4. Follow the compass bearing toward the waypoint. Anytime you suspect you're off route, switch on the GPS receiver. After acquiring fix, make the receiver redisplay the distance and bearing to the waypoint. Adjust the compass to this new bearing.
5. Once you reach the first waypoint, repeat Step 2 - 4 for subsequent waypoints along the route.

This method doesn't strictly require you to look at the map after the waypoints have been entered, but you should to do it anyway. If you don't look at the map, you will have difficulty accessing the terrain and may run into unexpected obstacles.

Example of navigating with pre-entered waypoints

The example in Section 4.8.1 shows how to transfer waypoints from a map to a GPS receiver. You are now ready to follow this route in the real world.

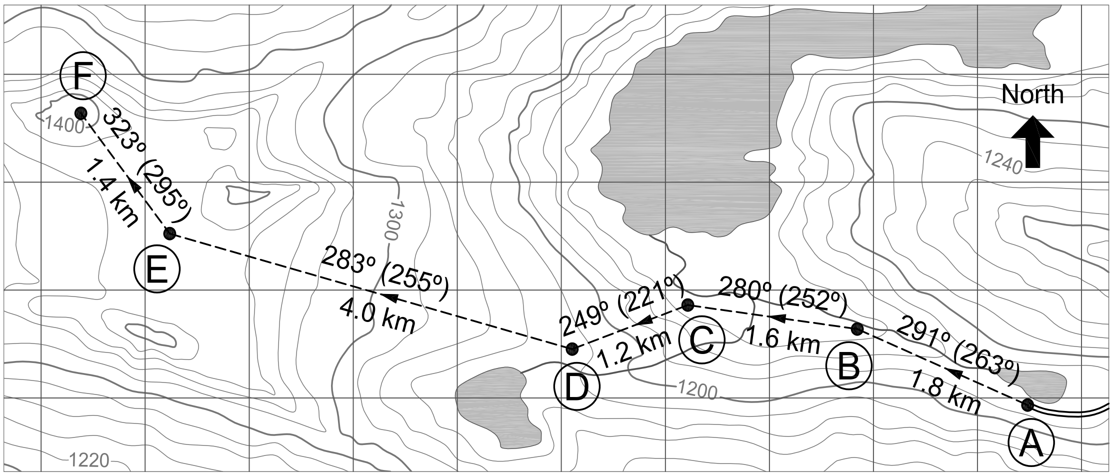

Your trip begins at the end of a road, a position that you have already pre-entered into your receiver as waypoint (A). Switch on the receiver, wait for a position fix, and compare the result with the pre-entered waypoint. This will tell you how accurate you were at transferring (A) from the map to your receiver. Update waypoint (A) with the numbers from your acquired fix. Set the GPS receiver to display grid bearings and read off the magnetic declination. Let's say the receiver shows that the declination relative to the grid is 28ºE. Set the declination mechanism on your compass to 28ºE. If your compass doesn't have a declination mechanism, set the GPS receiver to display magnetic bearings.

The receiver tells you that (B) is 1.8 km away at a grid bearing of 291º or a magnetic bearing of 263º. Set your compass to 291º if it has a declination mechanism (263º if it doesn't). Switch off the receiver and follow the compass bearing to (B). If you lose track of your location along the way, switch on the receiver, make it display the distance and bearing to (B), set your compass to this new bearing, and follow the bearing toward (B).

Use the same procedure to navigate to subsequent waypoints: follow 280º (or magnetic 252º) for 1.6 km to (C); 249º (221º) for 1.2 km to (D); 283º (255º) for 4.0 km to (E), and 323º (295º) for 1.4 km to (F). Note that you do not necessarily have to reach the exact location of each pre-entered waypoint. If you are planning to return by the same route, replace each pre-entered waypoint with an acquired position fix.

8.5.2 Navigating without pre-entered waypoints

Without pre-entered waypoints, the GPS receiver is just a powerful accessory to conventional map and compass navigation. Use the receiver only when you're unsure of your position.

1. Switch on the GPS receiver. Make sure that the receiver is set to the same grid system and datum as your map. Wait for a position fix.
2. Use the map's grid or geographic coordinates to pinpoint your position on the map.
3. Take a compass bearing off the map from your current position toward your next target. Follow this bearing toward your target.

8.6 Navigating without maps

You may for some reason decide to venture into unfamiliar territory without a map. This type of excursion is considered exploratory since you don't really know what lies beyond your range of view. Even with a guidebook or a verbal description of the area, you will lack the detailed terrain information that only a good map can provide. Finding your way back to your starting point should be your main navigational concern. Your strategy is governed by the terrain features that you encounter during your outgoing journey and what navigation tools you have at your disposal.

8.6.1 Navigating with nothing

A surprising number of people head out into untracked wilderness with neither a map nor any other navigation tools. Most who do this somehow manage to find their way back, although not necessarily along a return route that offers the greatest likelihood of success. Memorize all the significant landmarks that you come across during your outgoing journey, and frequently look back to familiarize yourself with the opposing view. During your return trip, use the memorized landmarks as steering marks to keep you on track.

Estimate the orientation of the cardinal directions before heading off. This can be done by observing the sun if it is visible (Section 5.2). Keep track of rough directions as you travel. If you encounter a large terrain feature, determine whether it tends north-south, east-west, or some other intermediate direction. Travel along, or parallel to, linear features that happen to be aligned with your desired travel direction. During your return journey, follow the same linear features or choose travel directions that will lead you to catching features that you are unlikely to miss. After reaching a catching feature, follow its perimeter to an identifiable point and then aim for the next catching feature.

Example of navigating with nothing

You decide to explore a trackless wilderness adjacent to a road, even though you know little about the terrain and have neither a map nor any other navigation tools. Your game plan is to head straight east from the road and continue traveling in that direction unless you're forced to circumvent obstacles. For your return, you intend to reverse your general travel direction by heading west, and use catching features to find intermediate locations.

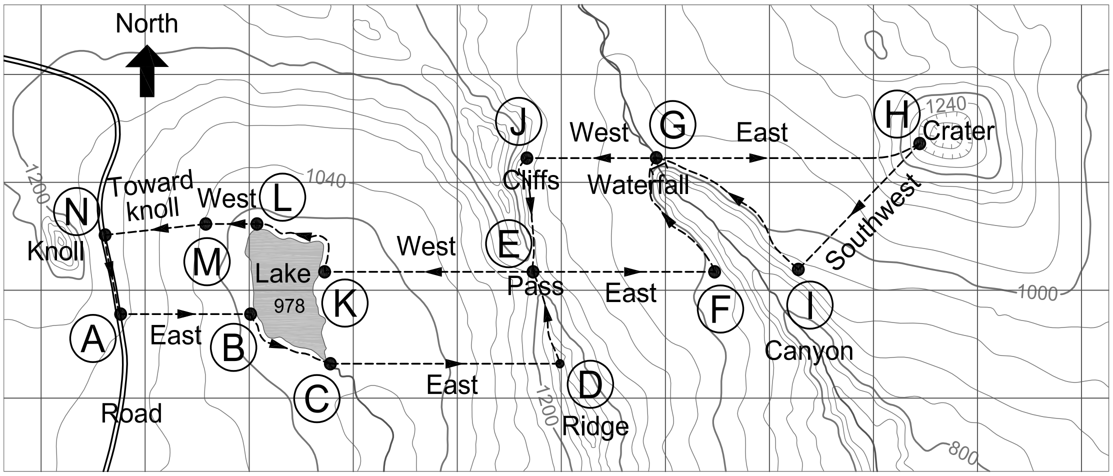

After parking your vehicle at (A), you head east from the road by using the sun and your clock to estimate your bearing. Your first obstacle is a lake that you bump into at (B). Going counterclockwise along the lakeshore guides you to its outlet at (C). From here, you continue eastward until you gain a ridge crest at (D). Following the ridge crest northward leads to a pass at (E), from where you resume your eastward journey. Your next obstacle is a canyon (F), which you circumvent by going left along its perimeter until the canyon terminates near a waterfall at (G). Carrying on eastward, you spot a volcanic crater and decide that this feature will be your final destination.

After reaching the crater rim at (H), you work out a navigation strategy for your return. You intend to navigate back via points (G), (E), and (C), which are easy to find, and ignore the nondescript points (F), (D), and (B). Your chief catching features are going to be the canyon wall, the cliff bands north of the pass, the lake, and the road.

To ensure that you hit the canyon, start hiking southwest instead of west. After reaching the canyon wall at (I), follow the perimeter of the canyon to the waterfall (G). From here, go straight west until you reach the base of the cliffs at (J). Follow the cliffs southward to the pass (E), and then take a westward course to the lake, which you hit at (K). You decide to ignore (C) because you're closer to the north end of the lake, and instead go counterclockwise along the lakeshore to (L) where you resume your westward travel. At (M) you spot a knoll that you recognize from when you parked your vehicle. Head straight toward the knoll until you reach the road at (N). Go south along the road back to your vehicle.

8.6.2 Navigating with just a compass

Navigating with just a compass is a step up from navigating with nothing. It gives you more precise directions than the sun or other natural clues. Aside from memorizing the terrain, your primary navigation concern is recording what bearings you have followed during your outgoing journey. Your return strategy is the same as for navigating with nothing except that the increased directional accuracy of your compass allows you to aim toward narrower catching features.

Example of navigating with just a compass

You decide to explore the same trackless wilderness as in the previous example, but this time you have a compass. Instead of going straight east, you plan to head off perpendicular to the road and keep traveling in that direction unless blocked by obstacles. For your return you will follow the reverse compass bearing as much as possible.

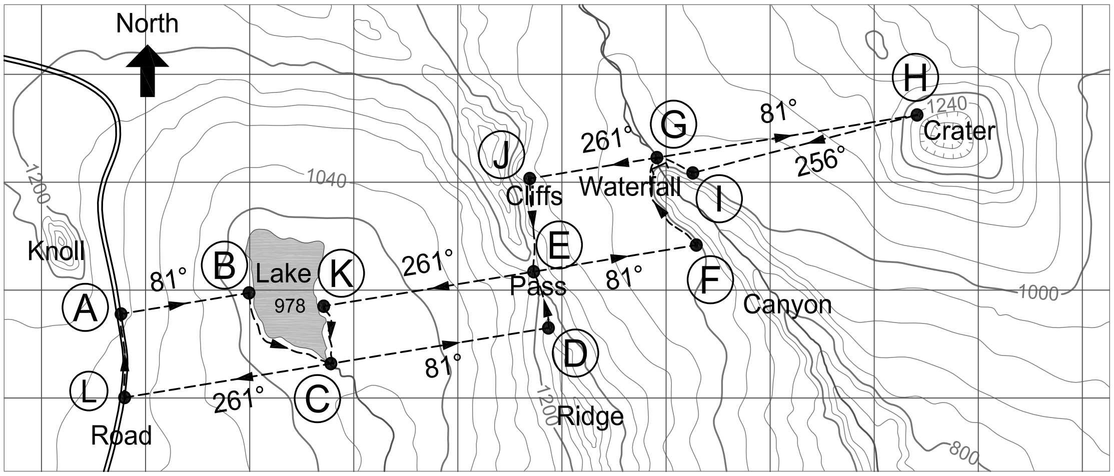

At (A), you take a compass bearing perpendicular to the road, resulting in a bearing of 81º. Following this bearing takes you to the lake at (B), and hiking along the lakeshore leads to its outlet at (C). From here, you continue along your original bearing of 81º until you gain the crest of the ridge at (D). The ridge crest takes you to the pass at (E). You continue along your original bearing to the canyon, which you reach at (F). Following the canyon perimeter brings you to the waterfall at (G). Lastly, you gain the crater rim at (H) by following your original 81º bearing.

Your chief catching features for the return are the same as for the previous example: the canyon wall, cliffs, lake, and road. Adding 180º to your original bearing gives you a return bearing of 261º. To ensure that you don’t miss the canyon, subtract a deliberate offset of 5º and follow a bearing of 256º. After reaching the canyon at (I), follow its rim to the waterfall at (G). The reverse bearing (261º) then takes you to the cliffs at (J). Go southward below the cliffs until you reach the pass at (E). From the pass, follow the reverse compass bearing to the lake, which you encounter at (K). Go left along the lakeshore to the outlet at (C). The last target is the road, a long linear feature that is hard to miss. Follow the reverse bearing to (L) and then go north along the road to your vehicle.

8.6.3 Navigating with just a GPS receiver and compass

A GPS receiver and a compass is a powerful combination for backtracking along a previously traveled route. The technique is simple and effective. Mark certain critical locations as waypoints in your receiver during your outgoing journey. On your return, use the receiver and compass to navigate back via these same waypoints. You can travel directly toward a waypoint without following linear features or aiming for catching features. Be aware that your GPS receiver may not work in canyons or under heavy tree cover.

1. At the start of your trip, switch on your receiver and wait for a position fix. If your receiver fails to obtain a fix within a few minutes move to a location where more sky is visible. Save the position fix in the receiver's memory as a waypoint and then switch off the receiver.
2. Each time you come across a critical juncture, switch on the receiver and save this position as another waypoint. Organize the waypoints sequentially into an electronic route.
3. Before starting your return trip, adjust the declination on your compass so that it conforms to true north, and set the GPS receiver to display true bearings. If your compass lacks an adjustable declination mechanism, set the receiver to display magnetic bearings.
4. At the start of your return trip, switch on your receiver and wait for a position fix. Make the receiver display the bearing and distance to the last waypoint that you saved on your outgoing journey. Turn the graduated dial on your compass to the bearing shown on the receiver. Switch off your receiver and follow your compass bearing toward the waypoint.
5. Once you reach the waypoint, repeat Step 4 for subsequent waypoints until you return to your starting point.

Example of GPS and compass navigation

You decide to explore the same wilderness as the in two previous examples. Your strategy for the outgoing trip is the same as with just a compass; head off perpendicular to the road and keep traveling in that direction unless blocked by obstacles. You will mark all critical locations as waypoints in your receiver.

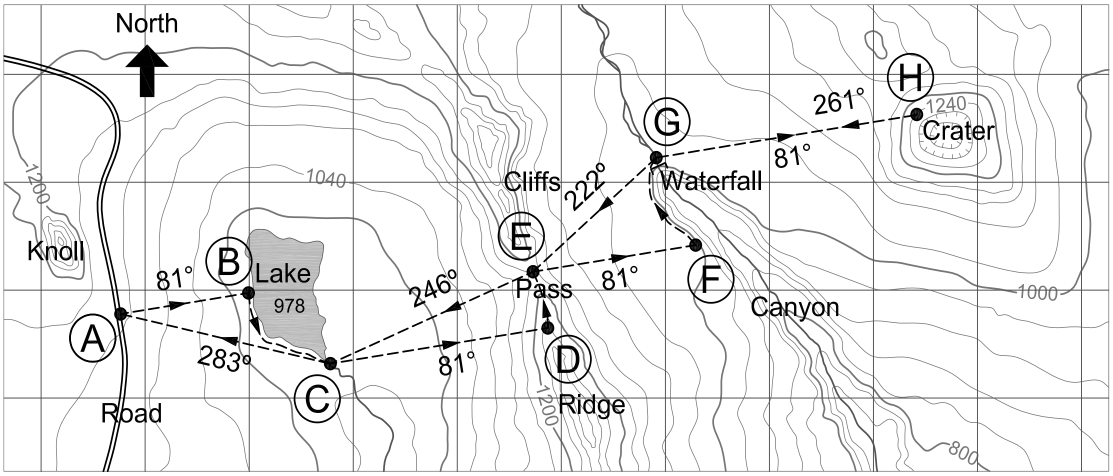

Before leaving, acquire a position fix at (A) and save it as a waypoint. Taking a compass bearing perpendicular to the road gives you the same bearing of 81º as in the previous example. Your journey follows the same path to the crater. Along the way, you save critical locations (C), (E), and (G) as waypoints in your GPS receiver, while ignoring the unimportant points (B), (D), and (F).

When you are ready to return, acquire a position fix and make the receiver display the bearing to the waterfall at (G). Set your compass to this bearing (261º) and follow the bearing directly toward (G). If you have trouble finding the waterfall, acquire another position fix and make the receiver redisplay the bearing to the waterfall. Readjust your compass to the new bearing and follow this bearing to the waterfall. Use the same technique to navigate to the pass at (E) by following a bearing of 222º. Similarly, follow a bearing of 246º to the lake outlet at (C) and a bearing of 283º directly to your vehicle at (A).

8.7 Real life navigation

Given that there are so many methods at your disposal, how should you navigate in real life? You are confronted with two conflicting factors: efficiency, and assurance of success. Efficient travel means doing the minimum amount of navigating that you think is necessary, while making sure that you find your destination may require a lot of extra work. A skilled navigator balances these two factors and switches navigation techniques as circumstances change. For example, you could use efficient terrain association to keep track of your location when landmarks are visible, then switch to time consuming dead reckoning when the landmarks disappear from view.

Judging when it’s appropriate to use a particular navigation method is something that comes with experience. Depending on the situation, the outcomes from certain methods will be more credible than others. If in doubt, choose the method that gives you the best chance of success. The time saved by using an efficient navigation method will be quickly eaten up if you do a mistake.

8.8 Lost in the wilderness

Consider yourself lost if you don’t know which way to proceed and are unsure of how to backtrack. Stop immediately unless you are in an unsafe location, in which case continue to the nearest safe location. Then relax, have something to eat or drink, and take some time to evaluate your situation. The key is not to panic. Calmly ask yourself the following questions:

- How badly lost am I really? How long ago did I know where I was going?
- What navigation tools, if any, do I have at my disposal?
- Do the visibility, weather, and other conditions allow safe travel?
- Have I informed someone about the details of my planned outing before heading out? The best information to leave at home or in your vehicle is a map with your intended route drawn in.

- Do I have a means to communicate with civilization?
- Am I in a part of the world where a rescue is likely, and if so, how long would I expect to wait for searchers to find me? Do I have the means to survive until then? Without a communication device, it may take a long time before you're found. First, someone has to determine that you need help, then he or she has to contact the authorities that have to launch a search, and finally, the searchers have to find you.

Based on your answers, make a rational decision whether to stay put and wait for a rescue or try to find your way out under your own power. Take into account that you will need more food and water while traveling than if you stay in one location. In most cases, it is easier for rescuers to find someone who stays in one place.

8.8.1 Waiting for a rescue

You are faced with two important but potentially conflicting considerations if you decide to wait for a rescue: how to survive and how to attract attention.

- Survival often means seeking shelter under a rock or tree where it is hard for rescuers to spot you, while making yourself visible means exposing yourself to the elements. A way around this dilemma is to seek shelter near a location where you can quickly make yourself visible.
- Leave as many clues as possible about of your presence. Mark the nearby ground with rescue signals and with any bright equipment or clothing that you are not wearing.
- Make sure that all available signaling devices, such as mirrors, flashlights, whistles, flares, or fires are ready for immediate deployment. Do not hesitate to use any of your visual signaling devices if you spot a low flying aircraft. Respond promptly to any attempt by rescuers to contact you.
- If you have an electronic communication device, use it to call for help. Remember that certain types of radios work better at night than during the day, and certain locations may have much better reception than other nearby locations.
- If you decide to move to a new location, leave as many clues as possible of your movement. Use techniques such as deliberately planting footprints into the ground or marking your route with symbols.
- Most rescue professionals recommend that all members of a group stay together. This doesn't necessarily mean that you must all wait at exactly the same spot. In some cases, spreading the group over an area can enhance the probability of one member being found. If you decide to spread out, make sure that everyone knows where all the other members are located.
- Do not approach a helicopter unless its engines are turned off or the pilot signals that it's safe.

8.8.2 Finding your way out

The standard method for finding your way out is to backtrack until you come across terrain that you recognize. Many people have successfully used this technique, but it is not always an option. The visibility may have deteriorated, nightfall is imminent, or you have traveled too far to turn around. In some cases, backtracking means traveling uphill, upstream, or against the wind. Another concern is that you could end up even more lost and further away from the area where potential rescuers would search.

If the travel conditions are unfavorable, waiting for better conditions can dramatically improve your prospects of finding your way out in a safe manner. You have a good chance of finding your way out if you're surrounded by landmarks and have a map and compass.

1. If your view is obstructed, move to a nearby location with better view, such as, a hilltop, clearing, or lakeshore. Try to identify any visible landmarks, and if possible, determine your position by terrain association or compass triangulation.
2. If terrain association or compass triangulation fails because of poor visibility or lack of identifiable landmarks, draw a circle on your map with a radius equal to the estimated distance from your last known location and centered on your last known location. A reasonable assumption is that you're somewhere within this circle.
3. Use your compass to orient the map and then systematically study the surrounding terrain and the contour lines on your map within the circle. Narrow down your possible locations within the circle by eliminating all areas that don't make sense. For example, if you're on flat ground, eliminate all areas that show steep terrain. Then identify locations within the circle that look like logical candidates for your position.
4. On the map, select a large catching feature such as a river, lake, ridge, or power line that looks like a reasonable target to aim for. Take a compass bearing from the middle of your location candidates toward the selected feature. Check that the same bearing will bring you to the catching feature from all location candidates. If not, select a larger catching feature or try to eliminate more of your location candidates.
5. Once you've taken a bearing that is sure to hit the catching feature from any of your location candidates, follow that bearing until you reach the catching feature. Now it's just a matter of determining your position on the catching feature. If you can identify one visible landmark, use triangulation from a linear feature to determine your position.
6. If you cannot identify any landmarks, move along the feature until you see a suitable landmark, or come across a uniquely identifiable part of the catching feature itself. You are then no longer lost.

You can use a similar procedure without a compass but must instead use natural methods such as the sun to estimate directions and orient the map. Since these alternate methods are normally less accurate than a compass, aim for an even larger catching feature. Without a map, use your memory of the terrain to select appropriate targets. Keep in mind that traveling in the correct general direction is more important than knowing where you are.

8.9 Rhumb lines and great circle routes

For very long journeys across wide-open spaces, there are techniques for minimizing the travel distance between two locations that are far apart. A rhumb line is a route between two points that follows a constant bearing. Rhumb lines are straight lines on Mercator charts, but are not usually the shortest distance between two points. To travel along a rhumb line, simply take a bearing off a Mercator chart from your current position to your destination, and then follow that bearing to your destination. Calculating the real world distance along a rhumb line is complicated because the scale varies across the chart. The easiest method is to use a computer program.

A great circle route is the shortest path along Earth's surface between any two points. Great circle routes look like arcs on Mercator charts. The easiest way to visualize a great circle route is to stretch a string between two points on a globe. In general, a great circle route doesn't follow a constant compass bearing. You must continually readjust your compass bearing to stay on course. The only exception to this rule occurs when a great circle route coincides with a rhumb line, which happens only when you travel along a longitude or the equator.

Plotting a great circle route is mathematically complex and is best done on a gnomonic chart or with a computer program. Gnomonic charts are created with a projection that shows great circle routes as straight lines. The task of following a great circle route becomes much easier if you don't plot the route in advance, but instead do it as you travel. By default, a GPS receiver displays the distance and initial bearing to a waypoint along a great circle route. Similarly, in sextant navigation, the sight reduction tables give you an initial bearing that follows a great circle route. By frequently recalculating your position and the bearing toward your destination, you will automatically be guided along a great circle route. The difference in length between a great circle route and a rhumb line is insignificant in most navigation situations except at high latitudes or on very long journeys with large east-west displacements.

Example of rhumb line versus great circle route

Let's say you're planning to row a dinghy from Saint John's, Newfoundland (47º34'N, 52º42'W) to Kilkee, Ireland (52º41'N, 9º38'W), and are wondering whether to follow a great circle route or a rhumb line.

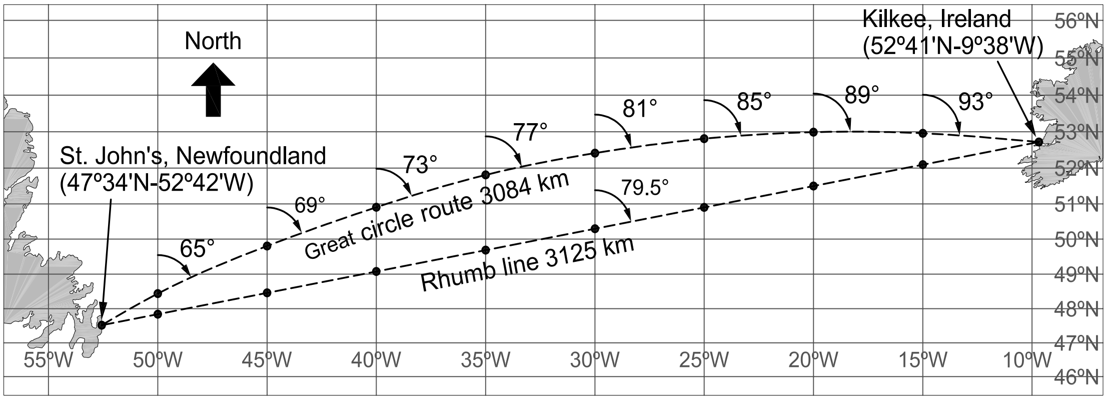

Taking a bearing from Saint John's to Kilkee on a Mercator chart gives you a true bearing of 79.5º. You can theoretically follow this bearing directly to your destination. Plugging the coordinates of St. John's and Kilkee into a navigation calculator gives a rhumb line distance of 3125 km. Plugging the same coordinates into a GPS receiver gives a great circle distance of 3084 km and an initial bearing of 63º. The great circle route reduces your distance by 41 km, but the navigation becomes more convoluted. By the time you reach 48º27'N, 50º00'W, the bearing to Kilkee will have changed to 65º. From here, you must increase your bearing by about 4º for every 5º of longitude that you cross. Your bearing will be 97º as you approach Kilkee.

In practice, your travel route is governed by prevailing winds and currents rather than by meticulously trying to follow a rhumb line or a great circle route.

8.10 Route finding

Route finding is defined as selecting a viable route across whatever terrain you are traveling. It is often confused with navigation, but is really a separate topic that is more difficult to master than pure navigation. Route finding is critical for safe wilderness travel and requires solid knowledge of the type of terrain you're dealing with. Deciding if a proposed route is feasible depends on whether you and your travel companions have the necessary technical skills, fitness levels, and equipment to safely negotiate the route.

Route finding usually involves several stages. The first step is to study maps or charts, and then trace out a route that looks doable. In mountainous terrain, calculate the average angles of the steepest slopes along your proposed route to find out if these slopes are too steep for your abilities. Remember that real slopes have sections that are steeper than the average. Large-scale maps with small contour intervals are best for calculating slopes and picking out details. Check out pictures or aerial photos of critical sections, or talk to someone who has previously done your intended route.

Once you arrive in the area, consult with the local experts. They can provide valuable insight on present conditions, weather patterns, features not shown on maps, and pitfalls to avoid. Every part of the world has its own particular type of climate, geology, vegetation, and even wildlife that can influence your route selection. You will likely have to re-evaluate your route as you travel. Having done your homework, any changes will hopefully be minor. If you're unsure if a proposed route is viable, try scouting it out first. Features often look and feel much different close up or from a new vantage point.

A whole spectrum of outdoor activities require route finding skills. Climbing a rock face obviously requires different route finding skills than bushwhacking through a jungle, negotiating a crevasse field, or kayaking down a raging river. There is no substitute for experience when it comes to mastering the intricacies of route finding. It cannot be learned by reading a book. The quickest way to improve your route finding ability is to team-up with an experienced outdoor person who is familiar with the local terrain and has the required technical skills for the particular activity. Every time you encounter a choice of routes, ask why a particular route was chosen rather than another. And most importantly, take advantage of any opportunities to practice your own route finding.

Chapter 9
Scenarios

Anyone who spends enough time in the wilderness will sooner or later have to deal with various navigation challenges. Fortunately, almost any navigation problem is solvable with the proper tools, skills, and sometimes, a touch of ingenuity. The following scenarios describe both common and unusual things that you may encounter.

9.1 Trail switchbacks

After hiking up a well-maintained trail through several switchbacks, the trail abruptly turns rough and then completely disappears. You are wondering what happened to the trail. It was supposed to be in excellent condition.

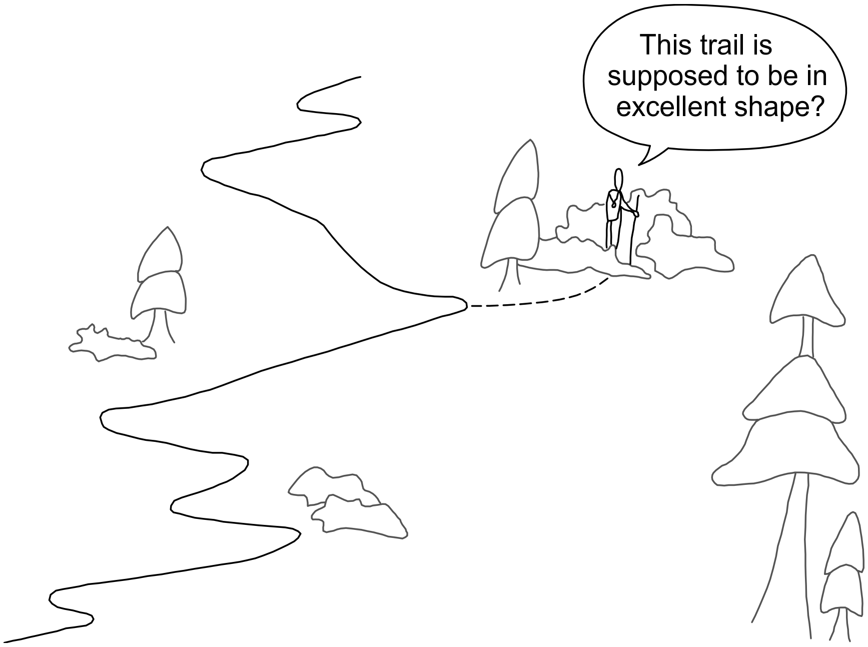
This trail is supposed to be in excellent shape?

A missed switchback is the most likely cause of your predicament. Going straight at a switchback is so common that a trail is sometimes worn into the ground by hikers who have missed the switchback. If a good trail suddenly turns rough, stop and retrace your steps to find out if you are in fact on the proper trail. Better yet, fight the natural tendency to look down at your feet or blindly follow the person ahead of you.

Continually scan the terrain left and right as you travel and periodically look back at the view behind you. Look at distant landmarks or use your compass to confirm that the trail is going in the correct general direction.

9.2 Multiple ski tracks

You are skiing back to your vehicle after a day of skiing in a popular backcountry ski area. On your way down a broad ridge, you encounter numerous ski tracks, and argue with your partner about which tracks to follow.

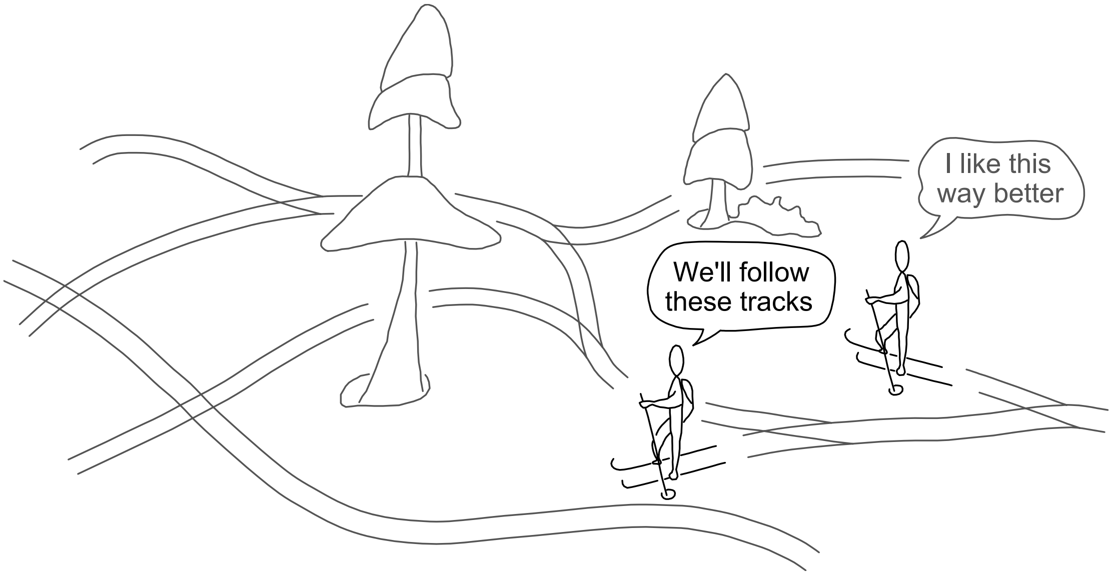

Skiers traveling uphill usually try to avoid extra effort by following a track that is already there. On the other hand, skiers traveling downhill in open terrain tend to seek out untracked snow, creating many tracks dispersed across the terrain. As a result, you will usually encounter more downhill tracks than uphill tracks in a busy area.

Uphill travel is slower and technically easier, allowing the skier who creates the track more time for navigation and route finding. Uphill tracks are therefore less likely to lead you astray, and often terminate at a popular destination such as a cabin or parking area. If you come across a web of ski tracks, your best bet is to follow a well-used set of uphill tracks, regardless whether you are traveling uphill or downhill.

Before following any track, make sure that it runs in the general direction where you want to go. Take a compass bearing off your map to confirm the direction. If you come across a major fork in the track, pinpoint your position on the map, and then use your compass to decide which track to follow.

9.3 Shoreline travel direction

During a leisurely canoe circuit, you decide to pull ashore and climb a distant hill. You haul your canoe to shore and then hike to the top of the hill in a few hours. After returning to the lake, you go left along the shoreline but then decide that the canoe must instead be to your right. After seesawing back and forth a few times, you still haven't found the canoe.

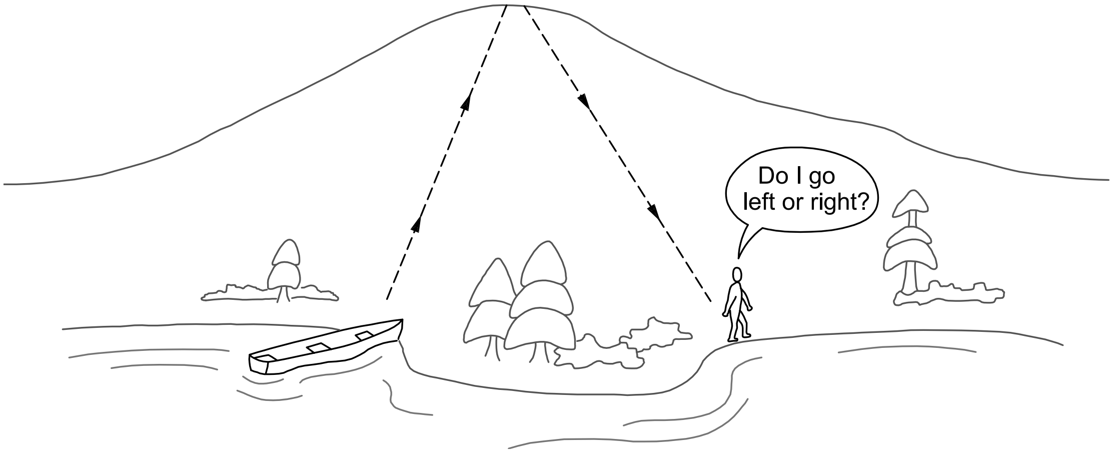
Do I go left or right?

The lakeshore is an extensive linear feature that is easy to find, while the canoe is a point feature that is much harder to find. It therefore makes sense to first find the lakeshore and then follow the shoreline to the canoe. You must however plan ahead so that you know which direction to go once you reach the lake. Several methods will accomplish this.

Use your compass to take a field bearing off a landmark before leaving the canoe. The landmark could be something across the lake or an island in the lake. Write down the compass bearing. When you return to the lakeshore, aim your compass in the direction of the landmark's bearing. Travel in whatever direction that will make the bearing point closer to the landmark. You should be very close to the canoe when the compass bearing points directly toward the landmark.

If you cannot see any obvious landmarks from the shoreline, take a field bearing off the hill itself before starting your hike. When you reach the top of the hill, reverse your bearing by adding or subtracting 180° from your original bearing. Then, add or subtract a deliberate offset of a few degrees. Follow the offset bearing to the lake. When you reach the shoreline, the canoe will be to your right (facing the lake) if your deliberate offset was subtracted, and to your left if your deliberate offset was added.

Another option is to mark the canoe's location as a waypoint in a GPS receiver. On your return, use the receiver and compass to travel directly toward the canoe. If you don't find the canoe when you reach the shoreline, use your GPS receiver to redisplay the bearing and distance to the canoe. You should be able to acquire a position fix quite easily along the shoreline where a significant amount of sky is visible.

9.4 Mountain top triangulation

Lost in the fog with just a map and compass, you decide to scramble up a nearby knoll. The fog suddenly clears as you reach the top where you are blessed with a panoramic view of several mountains sticking up above the clouds.

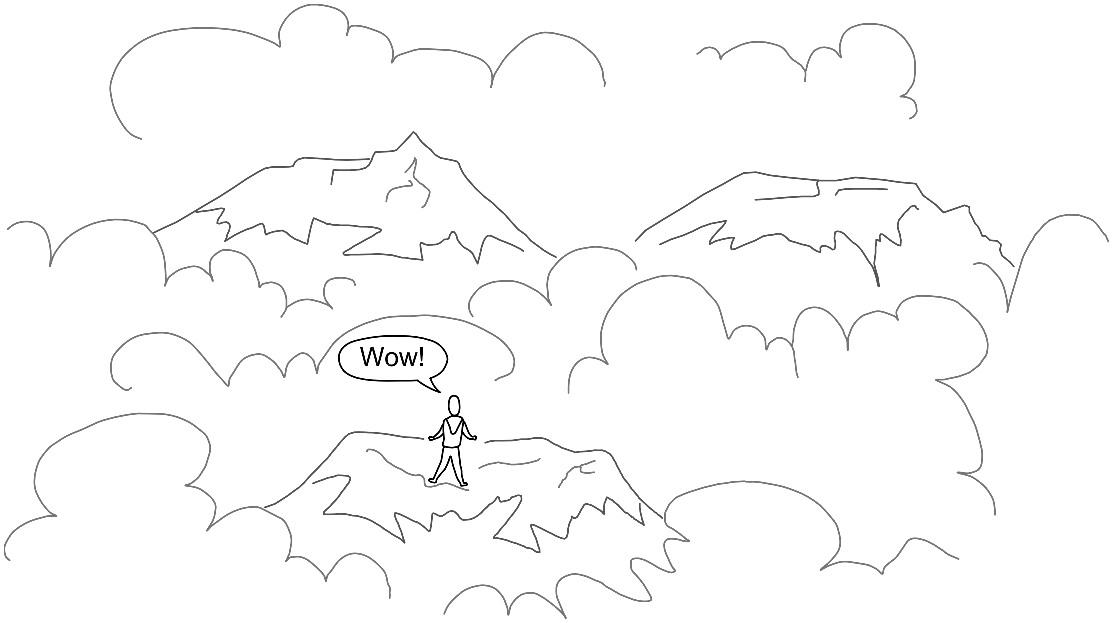

This is a classic situation where you can use compass triangulation to pinpoint your location. First, orient the map, and then try to identify at least two of the visible mountains by carefully comparing the shape of the mountains with the contour lines on the map. This can be tricky if only a small part of each mountain is visible. If the upper boundary of the cloud cover is fairly level, judge how far the visible part of each mountain sticks up above the clouds. Rank the mountains by height, and then compare the observed heights with the elevations on the map. You may be able to immediately identify the knoll that you are standing on by its position and height relative to the mountains.

To triangulate your position, take a field bearing off one of the peaks and draw a line of position on your map. Take a field bearing off a second mountain and use this bearing to draw a second line of position. If you have done the triangulation properly, you are located at the intersection between the two lines of position. Take bearings off additional mountains for greater confidence in your result.

After pinpointing your position, take a compass bearing off the map toward your next catching feature. Make sure that the catching feature is sufficiently large so that you will not miss it if you drift off course after descending into the fog.

9.5 Hidden trail markers

You're hiking along an easy-to-follow trail that is marked with reflective markers attached to trees. Higher up, you encounter snow and strap on your snowshoes. The trail markers eventually vanish under the increasingly deeper snow cover, and you suddenly realize that you have no idea where the trail goes.

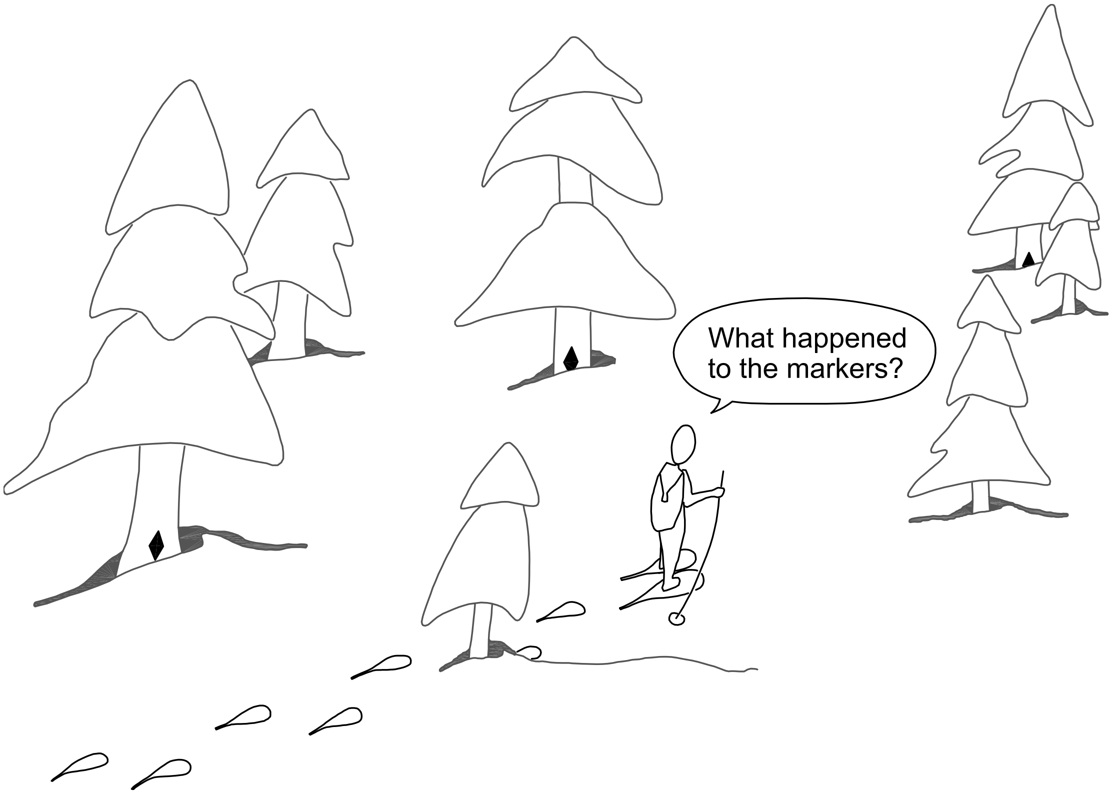
What happened to the markers?

This is a situation where the navigation abruptly changes from casual to challenging. The trail is no longer a linear feature that you can easily follow. If you're not carrying a map and compass, turn back before you get lost. Do not count on being able to retrace your tracks, as they could melt away or get covered by new snow.

Finding the location of the last visible trail marker is vital for your return trip. If landmarks are visible, use your compass to triangulate this location and record the bearings. Alternatively, mark the location as an altimeter reading or GPS waypoint. Because of the possibility of a forced retreat, mark this location even if you intend to return via a different route.

Once past the last trail marker, your navigation is governed by whether landmarks are visible and what navigation tools you have. For example, if the terrain allows, you could use conventional map and compass techniques.

9.6 Jungle camp

Your plan to find a remote camp in the Amazon jungle with a GPS receiver has failed because the satellite signals cannot penetrate the thick tree canopy. The vegetation also blocks your view of landmarks, preventing you from triangulating your position with a compass.

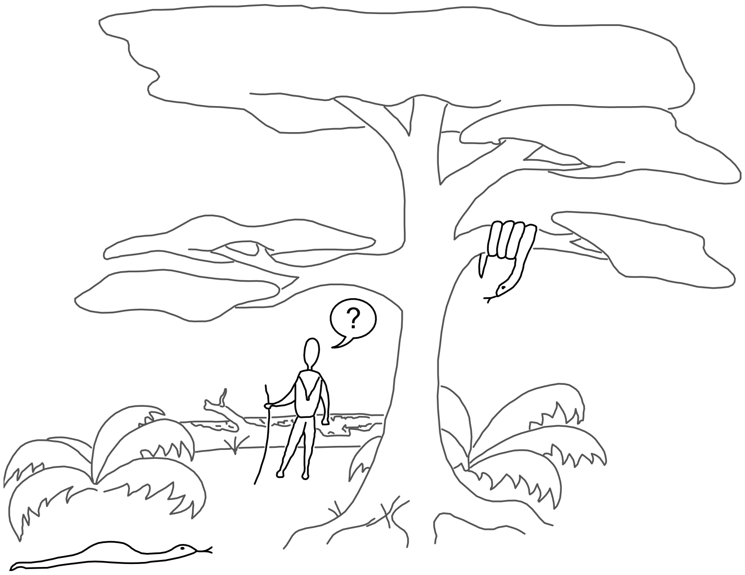

Use your map, compass, and altimeter as primary navigation tools as these will work under tree cover. Carefully plan your intended route before leaving. Break up the route into segments, with an intermediate catching feature at the end of each segment. A catching feature can be a lake, river, hill, swamp, or an elevation that matches a contour line on your map. Take advantage of any openings in the tree cover to get a GPS reading or to triangulate your position.

The illustration below shows an example of jungle navigation. From your start point, take a bearing off your map toward your first catching feature, the center of a lake. Follow the bearing to the lake. Walk to the end of the lake and take a bearing to your second catching feature, the center of a hill. After reaching the hill and going around its base, take a bearing to your last catching feature, a river fork. Add a deliberate offset of a few degrees to your bearing and follow this offset bearing to the river. Go upstream along the river to the fork. The jungle camp is away from obvious landmarks, so you are forced to use dead reckoning for this last leg. A final grid search may be needed to find the camp.

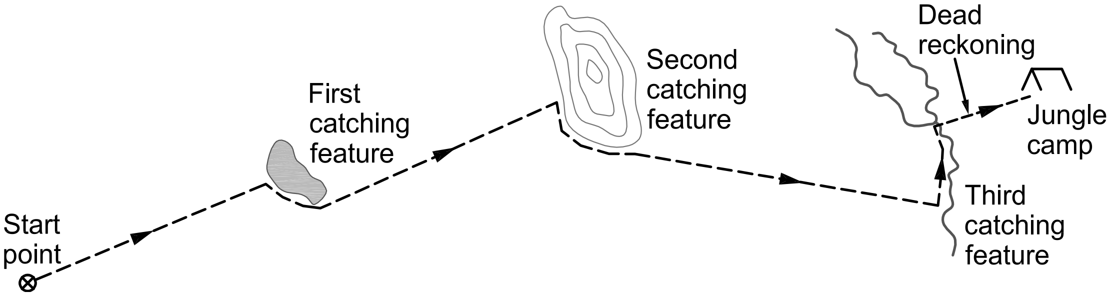

Dead reckoning is not easy in a jungle. The thick undergrowth makes it difficult to walk in a straight line while keeping track of your distance by counting paces. The most common error is to grossly overestimate your travel distance. With experience, you will be able to use your travel time to estimate the distance. Keep in mind that jungles are not always accurately mapped because the heavy tree cover makes aerial surveying difficult.

9.7 Ridge return

After scrambling up a face to reach a ridge, you follow the ridge to the top of a hill. On your return from the hill, you descend along the same ridge but miss the turnoff, and instead continue further down the ridge into unfamiliar terrain, wondering how you got off route.

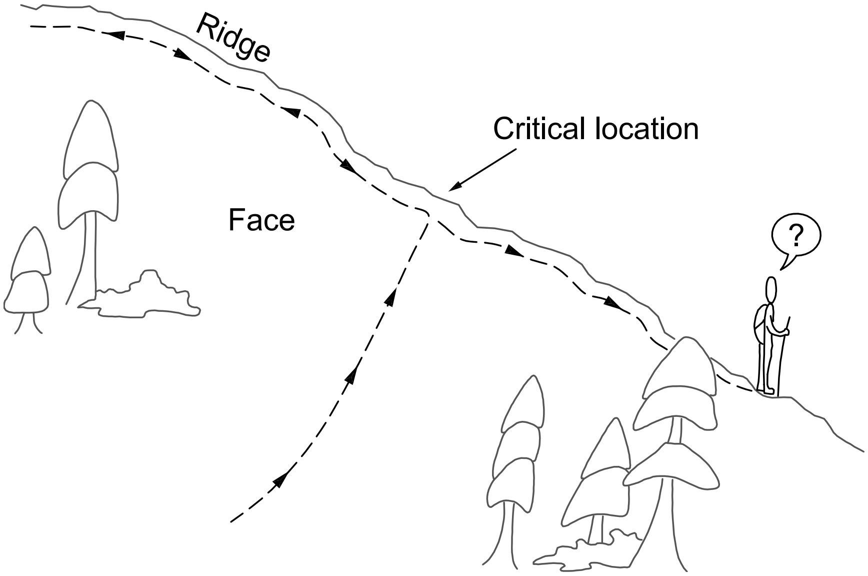

The natural tendency is to look straight ahead as you travel. It is very easy to miss a right-angle turn off a ridge, especially if there are no distinctive terrain features nearby. Once you realize that you have gone too far, go back up the ridge and find the point where your route leaves the ridge. This is a critical location that you should have memorized on your way up, and marked as a GPS waypoint, altimeter reading, or by recording triangulation bearings.

Alternatively, place a line of wands or cairns perpendicular to your line of travel at this critical juncture. A cairn is simply a stone marker that you build by stacking rocks on top of each other. A wand is a bamboo stick with a small colored flag tied to the top.

9.8 Desert traverse

For your next vacation, you decide on a solo camel trip across the Sahara desert from Tripoli to Timbuktu with various oases along the way. You realize that navigation will be critical to your survival.

That better not be a mirage?

The local population has traveled this desert for millennia using celestial bodies, sand dunes, and desert animals for navigation. You, on the other hand, should invest in a compass, GPS receiver, and maybe a solar panel to charge the batteries. As a backup, you may want to bring along a sextant, a celestial navigation calculator, and an accurate clock. Before leaving home, pour over maps of the area and write down the coordinates of all the oases along your route. Plan possible escape routes in case you're faced with an emergency.

To navigate your route, simply plug the coordinates of all the oases into your GPS receiver, and then use the receiver and compass to point you toward the next oasis. In this type of terrain your exact paths between the oases don't really matter. The wide-open terrain offers excellent long-range visibility but sandstorms and mirages can drastically degrade your view. A common mistake is to grossly underestimate distances to visible landmarks.

A long desert trip is an excellent opportunity to practice natural navigation methods. Try using rising and setting stars as direction finders and zenith stars to estimate your latitude. During the day, use the sun and your clock to determine rough directions. Most deserts have seasonal prevailing winds that arrange the sand dunes into specific patterns that you can use to stay on course. You may be able to find an oasis that is beyond your view by observing the movements of desert animals. Knowing the feeding and predator avoidance habits of desert species can help you determine if a particular animal is moving toward or away from an oasis. Desert pigeons, for example, flap their wings louder when flying away from an oasis because they are heavier from drinking water.

9.9 River delta

You set up camp on an island during a long canoe trip down the Mackenzie River delta. After dinner, you head out for a short paddle to explore the nearby channels and islands. On your return, you are unable to find your campsite. There are no obvious landmarks, all channels and islands look the same, and there are no hills to use as vantage points. To make matters worse, your map and all your navigation equipment are left behind in camp.

I have to find my campsite

People habitually treat short excursions with less respect than they should. In this timeless scenario, your food and survival gear is at a campsite that you can't find. Instead of wasting energy on a wild paddling spree hoping to stumble onto the camp, stop and draw up a structured plan for finding your way back. If your camp is nearby you should be able to find it with some kind of grid search.

Devise a method to mark your search route such as breaking branches or building cairns along the shoreline. Clearly mark your start point and then explore one of the nearby channels, marking your route as you go along. Return to your initial marked position and then explore the next channel. Repeat this process until you've covered all the nearby channels. If you haven't found your camp, move to a new base point located on the perimeter of your previous search area and explore outward from this point. Your search area will expand exponentially as you move outward, but a systematic search has huge psychological benefits and is much better than hoping for a miracle. As a last resort, consider abandoning your camp and drifting downstream to the nearest settlement.

As always, prevention is the best medicine. Before setting off, realize that the campsite is a critical location and employ effective methods to retrace your steps. Triangulating your campsite with a compass will not work in this case because there are no obvious landmarks, and an altimeter is useless on a broad, flat delta. Marking the campsite as a waypoint in a GPS receiver is very effective. If you don't have a GPS receiver, devise a method for keeping track of your outward journey. For example, trace your route on the map as you travel or mark the route with cairns. You could also record your journey as a series of compass bearings together with the distance paddled along each bearing. Don't forget to frequently look back to familiarize yourself with the opposing view.

9.10 Descent off mountain

You reach the summit of a remote peak just as the fog rolls in and decide to quickly get off the mountain. Rushing downhill as the fog thickens, you inadvertently veer off to the wrong side of the mountain and find your route blocked by a cliff. Climbing back up to attempt to rejoin your ascent route does not seem viable, as there is now a full-blown storm above.

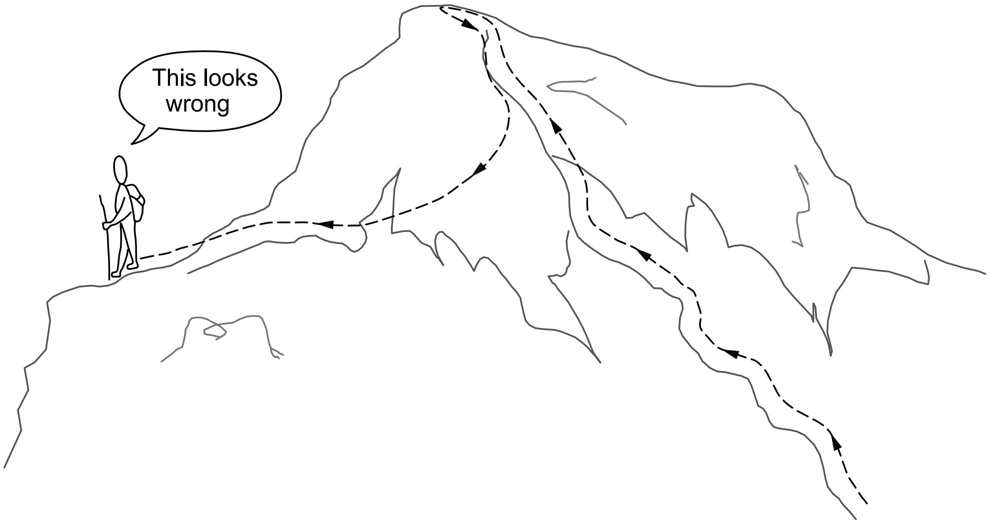

Most people get lost on their way down from mountains, not on their way up. As you climb higher, the overlying terrain decreases in size, effectively funneling you toward the summit. The reverse happens on the descent. You are faced with ever expanding terrain and an increasing number of route choices, each potentially wrong. Rushing downhill during poor visibility is usually a bad idea. Losing your way on the descent can have serious consequences, particularly on high, remote, or technically demanding mountains.

An effective method to retrace your steps is to place wands along your route as you ascend. The wands should be placed close enough so that you can find the next one even if the visibility is drastically reduced. Don’t worry if the wands look ridiculously close together. Place extra wands at critical locations such as rappel stations, ridge forks, or crevasse bridges. Also, mark these locations as GPS waypoints, altimeter readings, or by recording triangulation bearings.

If you plan to descend via a different route, carefully research your descent route ahead of time by analyzing maps and photos, consulting a guidebook, or talking to someone who has climbed that route. Note the elevations of all critical locations. Altimeters are particularly valuable for finding critical locations on mountains. Plan for the possibility of a forced retreat down your ascent route. Also, investigate possible escape routes that can get you down quickly in case of an emergency.

9.11 Trail return

You’re returning on the same trail that you used to access a wilderness area. For some reason, the trail looks different and seems longer than on the outward journey. You are wondering if you are in fact on the correct trail.

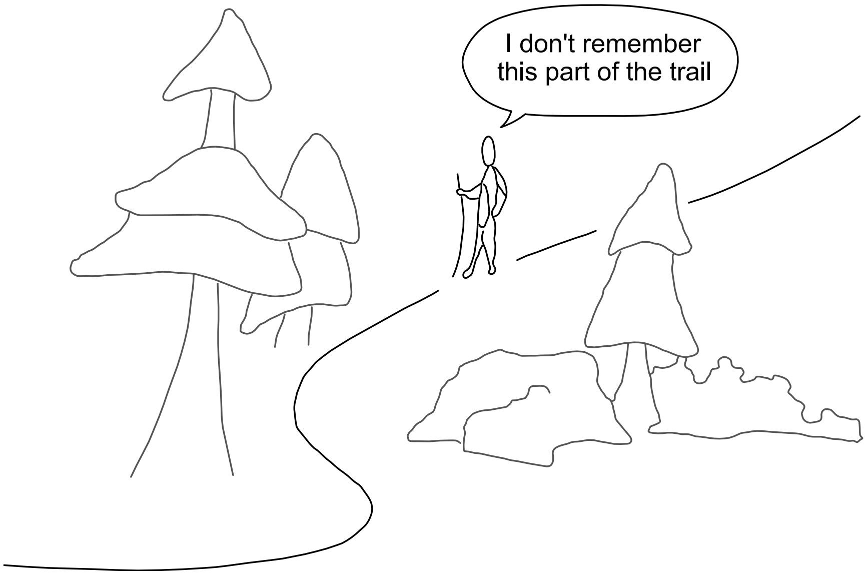
I don't remember this part of the trail

Most people are amazed by how different a trail or any other type of route looks on the return journey. This perception is caused by inattention during the outward journey and altered viewpoints during the return journey. During the outward journey, your mind is naturally focused on the goal at hand, making it easy to forget certain features or having traveled particular sections of the route. This leads to a truncated mental image of the route and many people are surprised by how long it takes to return. This perception is amplified when conditions have changed since your outward journey. For example, it may be darker on your return journey.

The best cure is to familiarize yourself with the opposing view by frequently looking back during your outward journey. This will help you build a more complete mental image of the route. Pay attention to unique features in all directions and mark critical points along your route as GPS waypoints, altimeter readings, by recording triangulation bearings, or with physical markers such as cairns or flagging tape.

9.12 Return to highway

You’re driving along the Alaska Highway across flat terrain and decide to check out the local swamps. You head off at a right angle to the highway through low-lying fog. After going straight ahead for a while, you decide to turn back because of the mosquitoes. For some reason the return journey is taking much longer than the outgoing hike and the highway seems to have vanished. As the fog lifts, you notice that the sun is in the wrong direction. You’re actually heading away from, instead of toward, the highway.

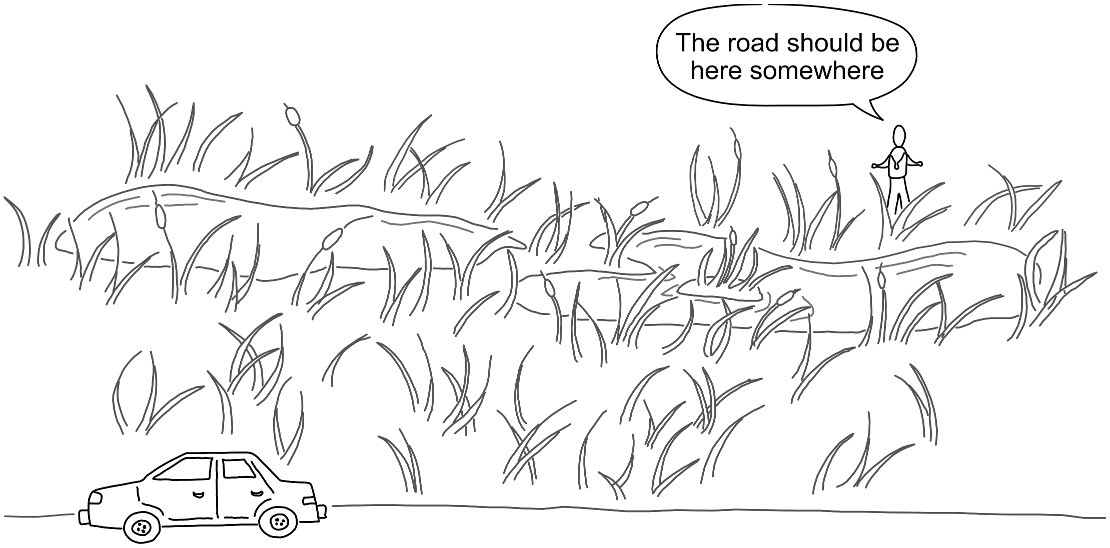
The road should be here somewhere

Terrain that looks the same in all directions can easily deceive your sense of direction. On flat terrain, almost everyone curves consistently to the left or to the right if no distant reference points are in view. The difference in leg lengths is the main cause of this effect. If your right leg is longer, you tend to veer to the left and vice versa. Other factors include having one dominant eye and the tendency to veer downhill and away from the wind. The faster you travel, the more you curve. To determine your innate curving disposition, find a flat area with no obstacles and walk with your eyes closed. If the ground is covered with sand or snow, your footprints will illustrate your personal curving disposition. You can then compensate by consciously forcing yourself to veer in the opposite direction.

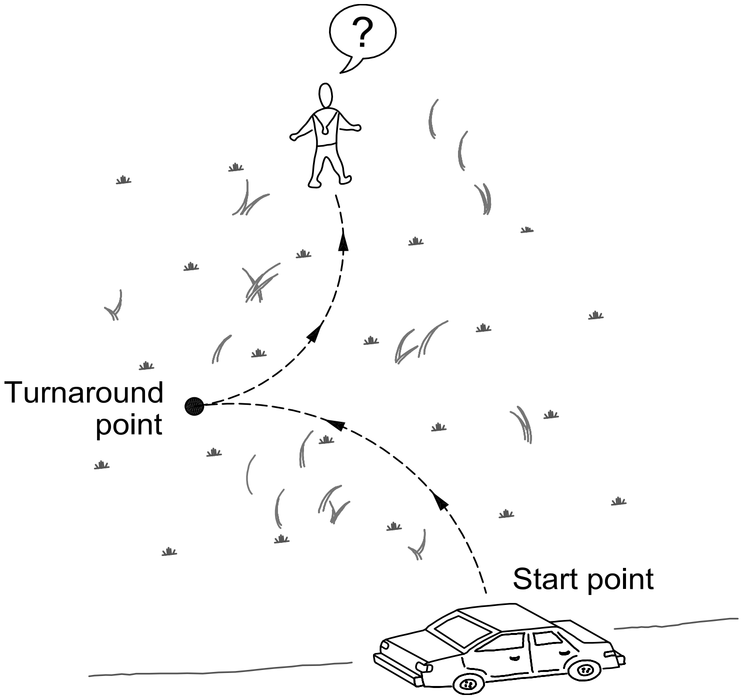

In this scenario, your tendency to veer left eventually altered your travel direction by 90º. On your way back you continued to veer left, ultimately traveling directly opposite to your intended direction. This type of scenario can easily be avoided by using a compass. Before leaving your car, take a compass bearing in the direction that you want to travel and then follow this bearing. For your return journey, add or subtract 180° to your original compass bearing, then add or subtract a deliberate offset of a few degrees. Follow the offset bearing back to the highway, an extensive linear feature that you're unlikely to miss. After arriving at the highway, the deliberate offset will govern which direction to go along the highway to reach your car.

If the sky is clear, you can use the sun to stay on track. Note the direction of the sun as you depart from the car and then keep a course that is at a specific angle relative to the sun. As you travel, slowly adjust your bearing to compensate for the clockwise (counterclockwise in the southern hemisphere) movement of the sun, which averages 15º an hour. Use the same principle to return to the highway.

Returning directly to the spot where your car is parked requires precise navigation. Unless there is a distinct landmark near the car, you'll probably need a GPS receiver. Mark the car as a waypoint before heading off. To find your way back to the car from anywhere in the swamp, just switch on your receiver and make it show the distance and bearing toward the car. Set your compass to this bearing and follow the bearing directly to your car.

9.13 Large group

Thrashing through the wilderness with a large group of people, you suddenly realize that you can't see or hear your friends anymore. You were just tagging along with the rest of the bunch and don't know much about the intended route or the surrounding terrain. To make matters worse, you're carrying neither a map nor a compass, plus the roar of a nearby river drowns out your shouts.

Large groups are notorious for fragmenting into subgroups because of unequal travel speeds. In some cases, one or more members are inadvertently left behind after a rest stop. Ideally, the whole group should stay together as one cohesive unit by traveling at the speed of the slowest member. Unfortunately, this doesn't always happen in practice and often leads to wild speculation about the location of members who are out of sight.

A common predicament occurs when only a few members know the route or have the required navigation skills and equipment. If these members get too far ahead, the group that follows may be forced to second-guess which way to go. A wrong turn will lead the followers astray, while the group on the correct path will be wondering why the followers aren't arriving.

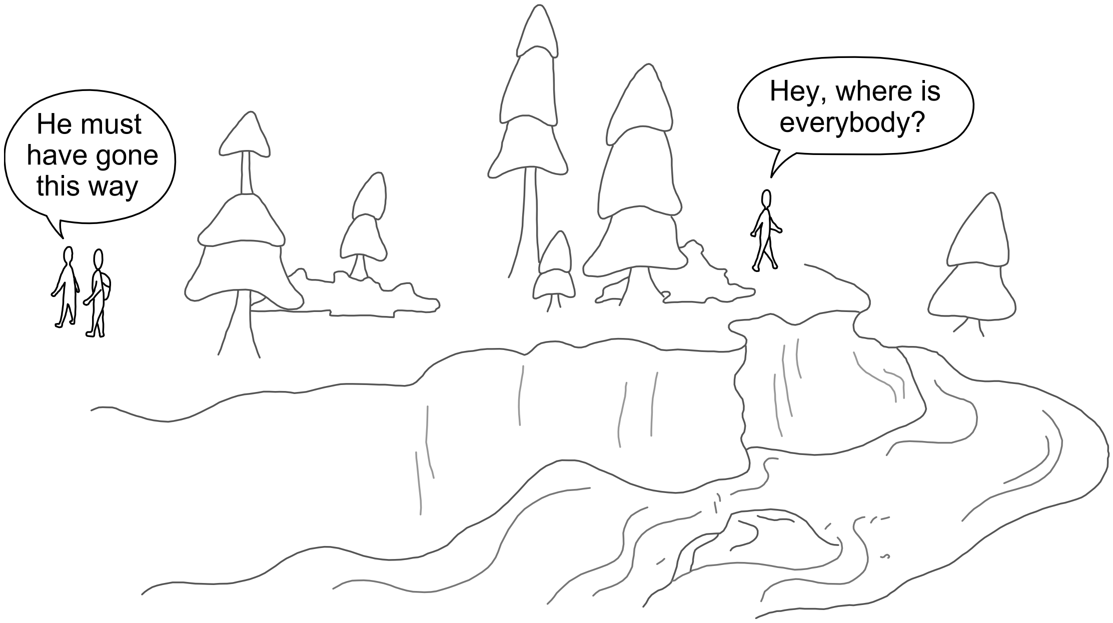
He must have gone this way
Hey, where is everybody?

Another type of dilemma arises when one group believes that a second group is ahead when it is in fact behind. The first group speeds up in an attempt to catch up but ends up leaving the second group further behind. Eventually one or both groups may initiate a search, which can degenerate into a ridiculous cat and mouse game with the groups seesawing back and forth along the route, or scattering across the terrain in a desperate attempt to find each other.

Before starting out, all participants must be informed of the proposed route. Predetermined meeting points and what to do if someone becomes separated should be discussed in advance. The leading group must wait at critical junctions, such as trail forks, until the rest of the members catch up or indicate the correct travel direction with markers. Do a body count at each rest stop to ensure that nobody is missing. In a few instances, a group has returned home without realizing that a lost or injured member is still out in the wilderness. Other times huge search parties have been sent out to look for “lost people” who have returned to civilization but neglected to notify anyone.

A fragmented team inevitably becomes a weaker team, especially if the skill level and group equipment is unevenly distributed. Breaking up the group ahead of time into self-sustaining units can mitigate this problem but will also lessen the incentive to stay together. A possible solution is to carry two-way radios. The most effective solution is to have a leader who can control the group, sets a reasonable pace, and does not always lead from the front.

9.14 Disappearing trail

You’re hiking on a trail that disappears and reappears with irritating regularity.

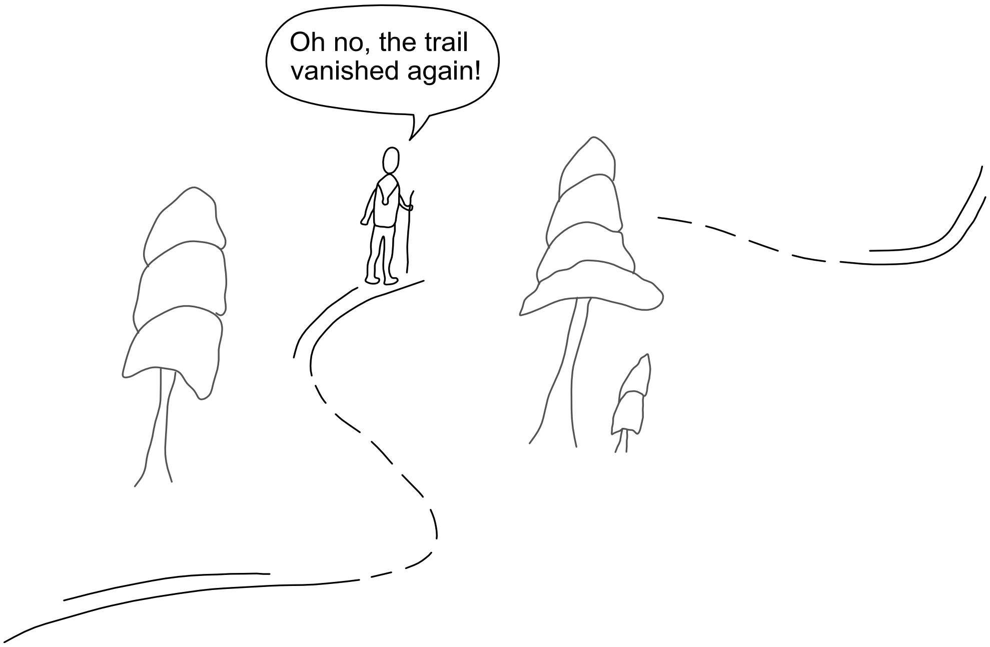
Oh no, the trail vanished again!

This often occurs when traveling on animal trails or trails that run through different types of terrain. A trail may be well defined through the trees but disappear completely when crossing a boulder field, only to reappear at the next patch of trees. Sometimes a trail simply ends. In moderate, open terrain you may choose to ignore the trail and simply travel in the correct general direction. Trails generally follow the path of least resistance so choosing an off-trail route that minimizes hardship will increase your chances of stumbling onto a trail if it exists. Keep in mind that most animal trails eventually lead to water and almost always lead to a pass if there is one ahead. In addition, some animals are experts at creating trails that contour across slopes.

One thing you want to avoid is paralleling a perfectly good trail for long distances. It may be prudent to zigzag your way forward if you’re reasonably certain that there is a trail nearby. If the terrain is too rough or the vegetation too thick for reasonable off-trail travel, you may be forced to waste a lot of time looking for the trail. Once you find the trail, pay close attention to every detail in your surroundings to avoid losing it.

9.15 Whiteout

A fog bank rolls in and it starts snowing as you ski across a large, snow-covered ice field, creating a condition known as a whiteout. Nevertheless, you decide that there is enough visibility to continue. After an hour of navigating by instinct, you come across a pair of tracks, and realize that these are your own tracks.

A whiteout doesn't necessarily imply that you can't see anything. It means that you cannot distinguish between things that are white. Brightly colored or dark objects can often be seen from some distance away. On the other hand, snow-covered ground blends in with the fog and it's very difficult or impossible to assess the shape of the surrounding terrain. Distance and depth perception are distorted. Travel can be treacherous because crevasses or snow-covered cliff edges are often invisible, even if they are just one step away.

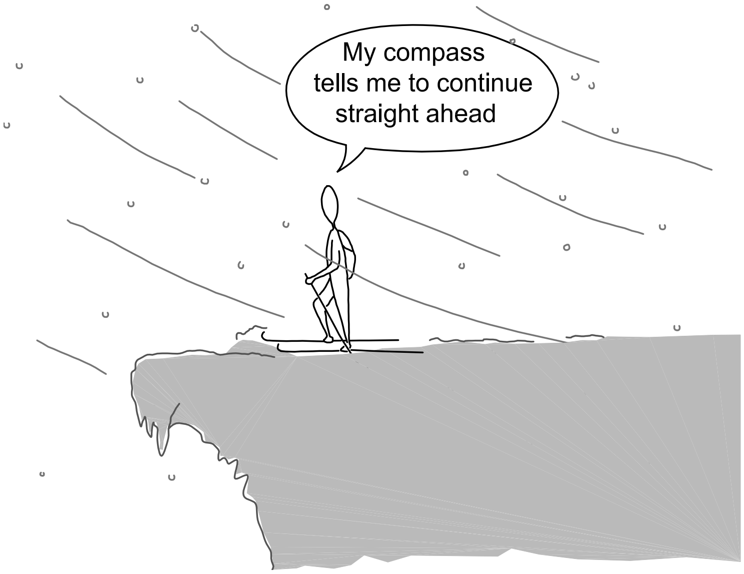
My compass tells me to continue straight ahead

Successful navigation in whiteout conditions requires faith in your instruments. Without a compass, most people end up traveling the path of least resistance, downhill and with the wind. If the terrain is flat and there is no wind, most people will travel around in circles. A map, compass, and altimeter or GPS receiver is essential. Use your altimeter or GPS receiver to pinpoint your position on the map, and then take a compass bearing toward an intermediate target along your route. If you just have a map and compass, do a rough estimate of your position and then take a bearing toward a large target. Move extremely slowly as you follow the compass bearing. Ignore your directional instincts. If you're unsure of how steep the slope is ahead, test the slope by throwing rocks or snowballs in front of you.

Moving in a straight line in a whiteout is very difficult, even with a compass. If you're traveling solo, look at the compass every few steps to stay on track and take frequent back bearings. If the back bearing doesn't point directly toward your tracks, move sideways until it does. If you're part of a group, travel in a file with the last person carrying the compass and the first person acting as a reference point. The last person checks the compass bearing and periodically directs the first person to veer left or right.

Do not underestimate the inherent dangers of traveling in a whiteout. Unless there is an urgent need to continue or you're certain there is no dangerous terrain ahead, stop until the visibility improves. If you decide to move, take appropriate measures such as roping up to reduce the risks.

9.16 Trail from lake

You drag a canoe down a trail through deep forest to a lake. After reaching the shoreline, you paddle to the other end of the lake for an overnight stay. On your return, you cannot find the trail that you used the previous day. The shoreline looks the same everywhere and the trailhead is hidden from view by trees.

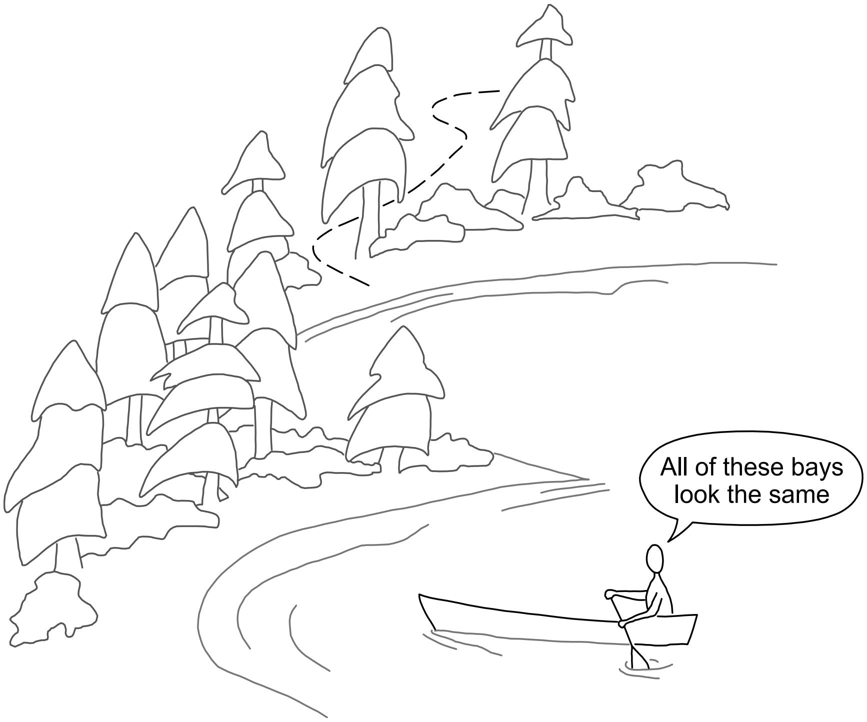
All of these bays look the same

On your outgoing journey, you were moving toward the lake, a large catching feature that was almost impossible to miss. On your return, you're faced with the reverse situation, trying to find a small target, the trailhead, from an indistinguishable location on a large feature. The trailhead is a critical location that you must mark on your outward journey. You could mark it as a GPS waypoint or by recording a triangulation bearing. Only one bearing is needed because the shoreline is a linear feature. If you don't have any navigation equipment, mark the trailhead with flagging tape, or build a rock cairn or a branch tripod before setting off. Also, note any shoreline catching features that would let you know that you've paddled past the trailhead.

Having failed to mark the trailhead, you are faced with a potentially long and tedious scouting expedition. Start from a point well back from where you think the trailhead is located and slowly paddle forward along the shoreline while keeping your eyes glued for any signs of the trail. If anything looks promising, go ashore to check it out.

9.17 No GPS signal

On your quest to conquer Mount McKinley, you decide to climb the avalanche-prone, north-facing Wickersham Wall. After a week of climbing and acclimatizing, you pull out your GPS receiver from the bottom of your pack to check your altitude. For some reason, your receiver refuses to get a position fix.

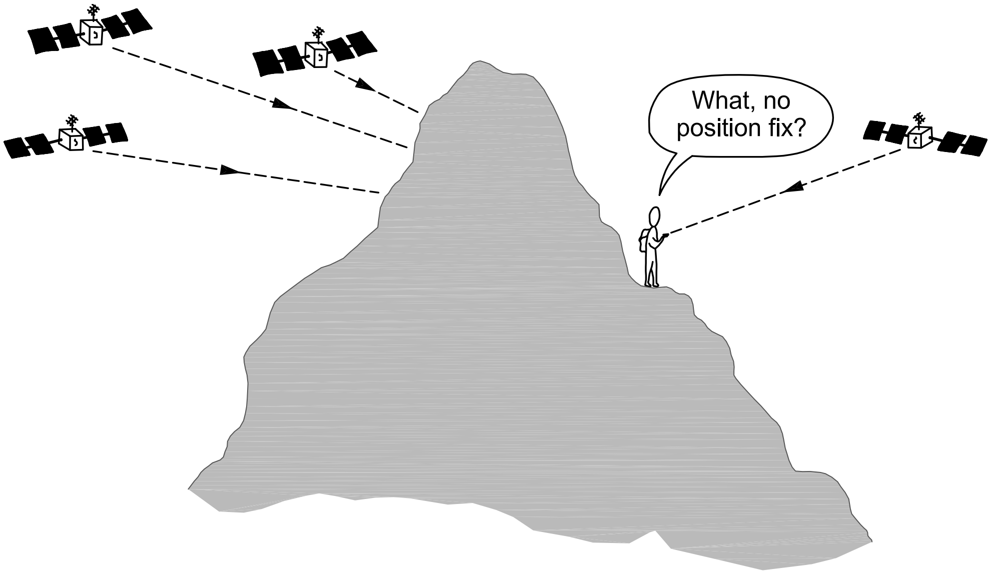
What, no position fix?

The orbits of GPS satellites carry them to a maximum latitude of about 55°N. Because Mount McKinley is at a latitude of about 63°N, most satellites that are visible from the north-facing Wickersham Wall are the ones that happen to be in northern positions on the opposite side of Earth. The great orbital heights of the satellites allow position fixes even at high latitudes. In fact, your GPS receiver can get an accurate fix at the geographic poles. The problem arises when a mountain or other obstacle blocks a significant portion of the sky and the rest of the sky doesn't have enough satellites for a position fix. This commonly occurs on steep, north faces in the northern hemisphere, and steep, south faces in the southern hemisphere, particularly at high latitudes.

You may eventually get a position fix if you leave your GPS receiver switched on long enough. The satellites are constantly moving and a favorable configuration may temporarily appear.

It is always a good idea to bring along an altimeter when climbing steep walls where knowing your elevation is often more important than knowing your horizontal position. If the visibility permits, try triangulating your horizontal position with a compass. Be aware that a slight horizontal error on a steep slope will translate into a large elevation error on the map.

If you brought along a suitable thermometer, you could measure the temperature of boiling water, and then estimate your elevation based on this temperature. The boiling temperature of pure water is 100 ºC at sea level, and decreases by 1 ºC for every 300 m of elevation gain (Section 6.6). For example, if your boiling temperature is 84 ºC, your elevation would be (100 - 84) x 300 m = 4800 m.

9.18 Back roads

Walking along an abandoned logging road, you're unable to find the fork of a branch road that is clearly marked on your map. To confirm your position, you take a GPS position fix, then an altimeter reading, and finally use compass triangulation. All methods put you at the same location on the map near the road fork, but after a thorough search of the area, you cannot find the branch road.

This is a situation where you should seriously question the accuracy of your map. Like other navigational tools, maps cannot always be trusted. Even high-quality maps from reputable mapmakers are prone to errors, especially regarding man-made features such as roads.

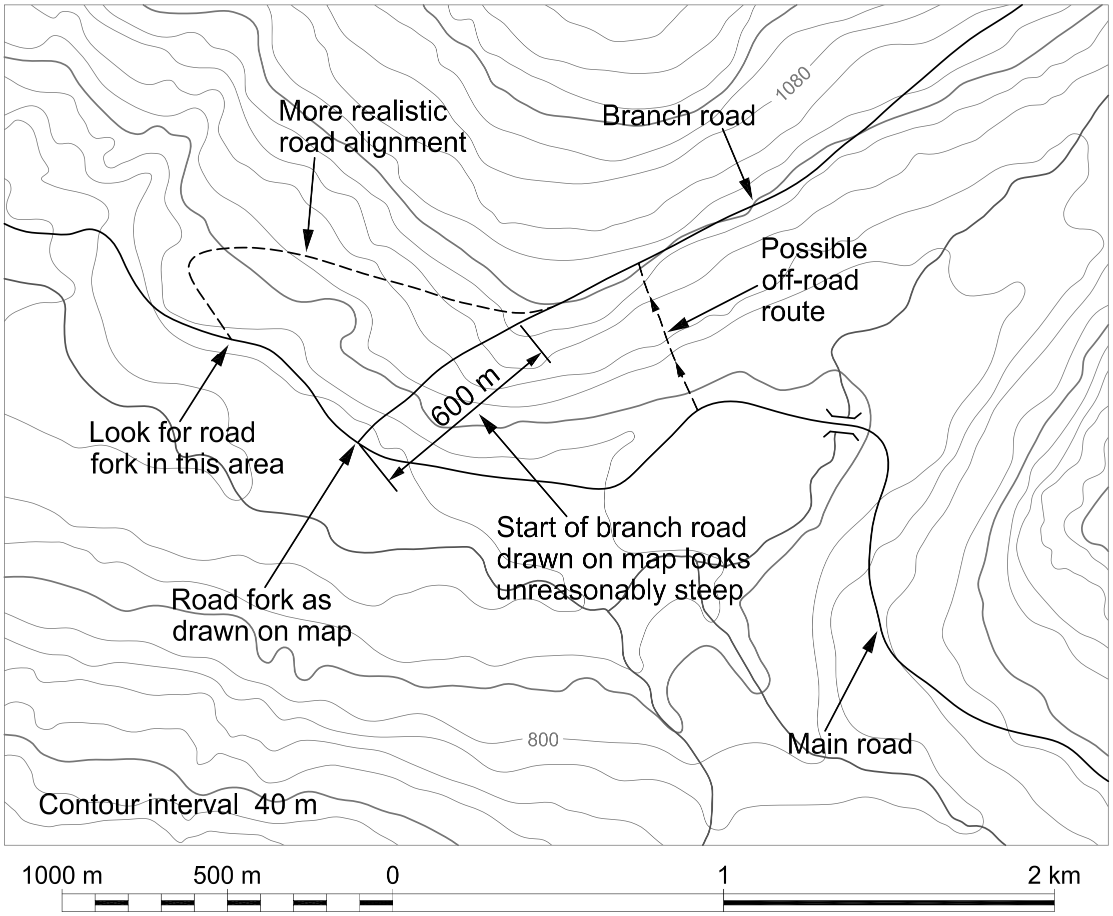
1080
More realistic road alignment
Branch road
Possible off-road route
600 m
Look for road fork in this area
Start of branch road drawn on map looks unreasonably steep
Road fork as drawn on map
800
Main road
Contour interval 40 m
1000 m
500 m
0
1
2 km

By carefully analyzing the contour lines on the map, you will notice that the branch road starts out with a very steep hill. Over a distance of 600 m, the road rises about 7 contour lines for an elevation gain of 280 m. This works out to a gradient of 47% or a slope angle of 25°, much too steep for a drivable road. Since it is unlikely that the mapmaker drew in a road that never existed, look at the adjacent contour lines for an alternate road alignment that makes sense.

You will discover that the slope eases a few hundred meters to the northwest. Scout out this area for a possible start of the branch road. If you don't find the branch road there, investigate other reasonable possibilities. If this fails, travel off-road toward your destination. You may eventually stumble onto the road if you choose a route that intersects a reasonable road alignment.

Logging and mining roads are often temporary features that may or may not have been drawn on your map. They can evolve into a web of roads that end abruptly. Getting lost on back roads is actually much more common than getting lost in true wilderness areas. It is often difficult to know if you're on the correct road until you've traveled along that road for some distance. Survey the surrounding terrain as you travel to keep track of your position, and use a compass to ensure that you are moving in the correct general direction. Don't be afraid to pull out your altimeter or GPS receiver if you are uncertain of your position.

9.19 Food stash

For a long ski traverse, you decide to hire a ski plane to place a food stash on a glacier near the midpoint of your route. You accompany the pilot when the stash is dropped and plant a wand in the middle of the stash. You then use your compass to triangulate the position of the stash, dutifully recording the field bearings to the landmarks. Unfortunately, all the landmarks are hidden by fog when you reach the vicinity of the stash during your traverse, preventing you from triangulating your position.

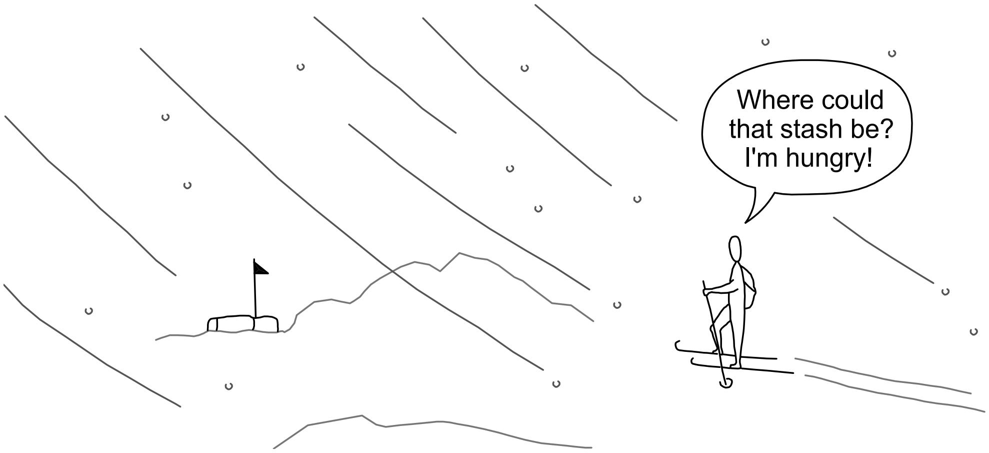
Where could that stash be? I'm hungry!

Given the importance of finding the stash, you should employ several backup methods to mark its location. Options include marking the stash's location as a GPS waypoint or an altimeter reading. An effective method for zeroing-in on a buried stash is to place an avalanche transceiver inside the stash. Don't forget to insert a fresh set of batteries and make sure the unit is switched on. You will need a second transceiver to find the stash.

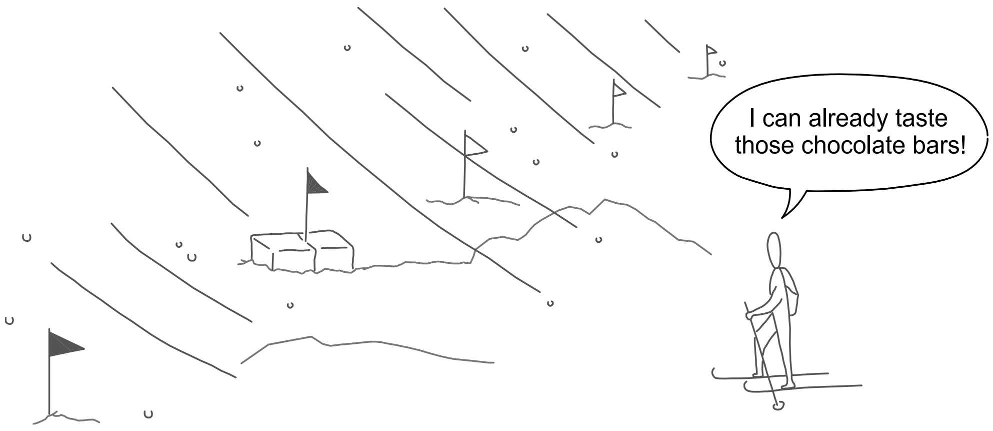
I can already taste those chocolate bars!

A reliable method for marking a stash is to plant a series of wands, about 50 m apart, on both sides of the stash. The line of wands should be perpendicular to your expected direction of travel, with the stash at the center. Placing different colored wands on each side of the stash and numbering them according to their distance from the stash will make the stash even easier to find. Be aware that wands often melt out and fall down or get completely covered by new snow. Make sure the wands are long enough to be planted deep into the snow and still stick up high above the snow.

9.20 Antarctic ordeal

During his ill-fated Antarctic expedition from 1914 to 1916, Ernest Shackleton's ship was trapped by ice for 9 months, and then crushed by the ice. He and his crew abandoned the ship and then drifted on the pack ice for 5 months to open water. With three lifeboats, they managed to reach uninhabited Elephant Island just off the Antarctic Peninsula.

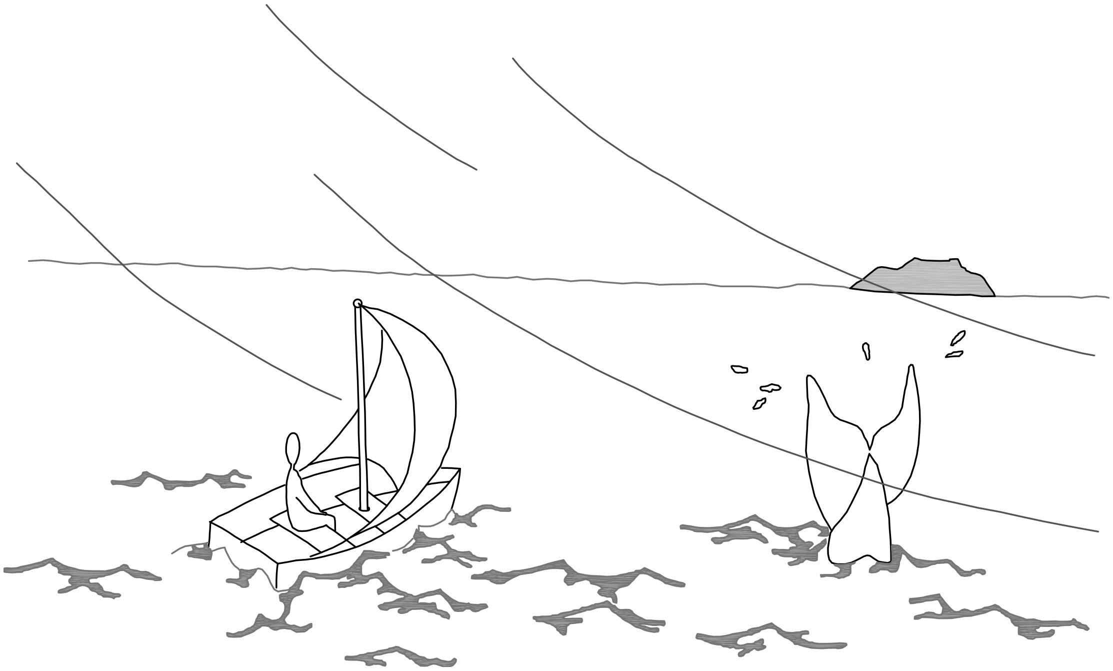

To get help, Shackleton and five other men set sail in one of the small boats toward South Georgia, 1300 km away. It would be a onetime opportunity. If their course was off by more than a few degrees, the boat would miss the island and disappear into the South Atlantic where they would surely perish. The available navigation tools were a compass, sextant, chronometer, nautical almanac, and pen and paper.

Shackleton's navigator, Frank Worsley, plotted a course to South Georgia, taking into account the prevailing winds and currents. During the 16-day voyage, the skies were mostly overcast and the boat was continually pounded by enormous waves and gale force winds. On the wildly pitching boat, Worsley managed to take only four sextant sights of the sun between thick clouds. The location of the horizon could only be guessed because of the poor visibility. After taking the sights, Worsley consulted the soggy almanac and used a pencil to do the calculations. The rest of the navigating was done solely by dead reckoning. The boat miraculously followed a near perfect course to South Georgia and the whole expedition team was eventually rescued. Some people regard this event as one of the greatest feats in navigation history.

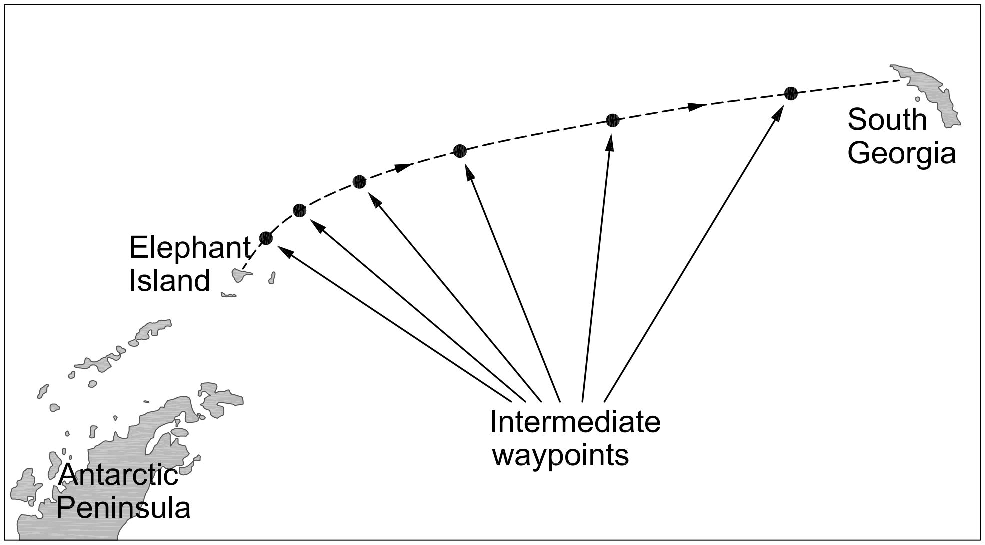

Trying to recreate this journey today using a GPS receiver instead of a sextant would make the navigation a lot easier, but plotting the proper course would be just as difficult. Marking the position of South Georgia as a waypoint in your receiver and following a great circle route to the island would not work because winds and currents would push you off course. Instead, you would have to follow a route consisting of a series of intermediate waypoints that take into account expected currents and winds as well as how fast your boat can move relative to the water.

Frank Worsley was obviously gifted but still required years of practice as a ship navigator to perfect his navigation skills. Today's aspiring navigator should take note. High-tech gadgets are no substitute for practical experience and clear, rational thinking.

9.21 Polynesian journey

In 1976, a replica Polynesian, doubled-hulled, ocean-going canoe set sail from Maui (20º53'N, 156º41'W) to Tahiti (17º33'S, 149º36'W). Only traditional South Pacific navigation techniques were to be used on this voyage that covered a straight-line distance of around 4300 km, mostly across open ocean. No navigation instruments were carried onboard. The crew included expert Micronesian navigator Mau Piailug.

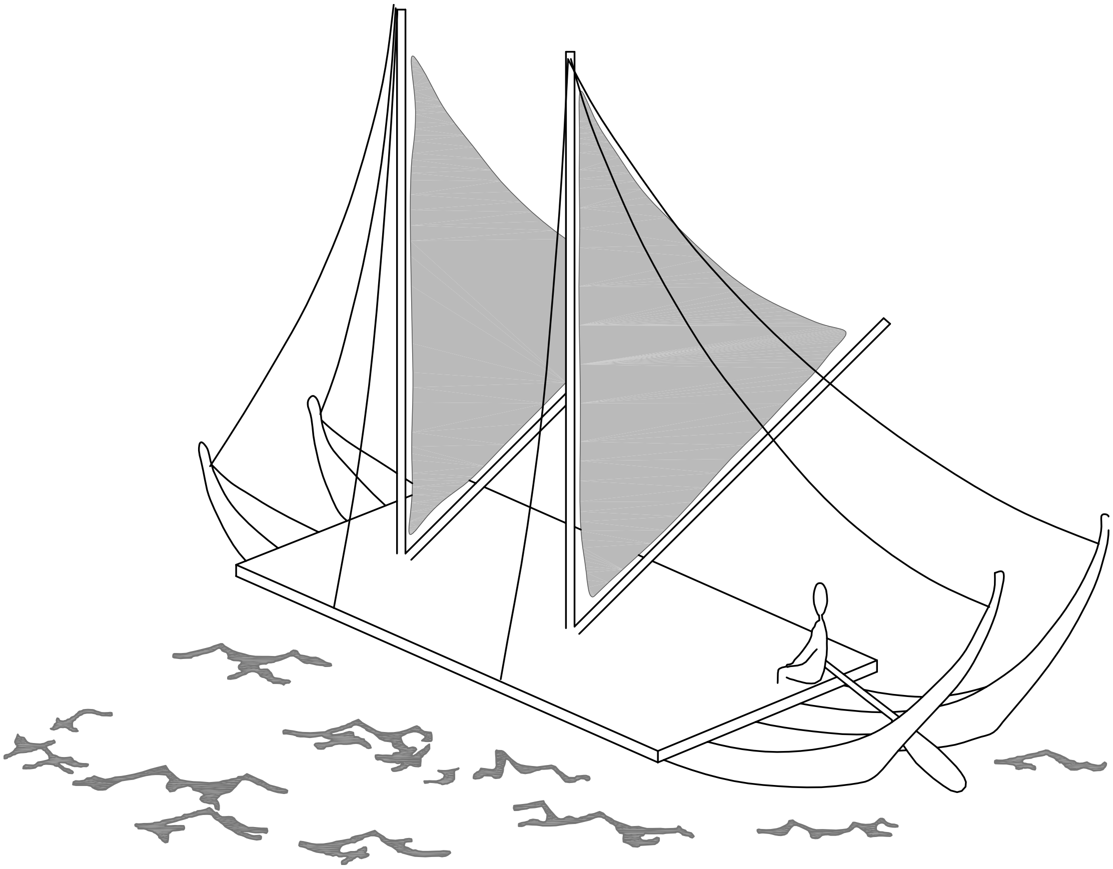

Before taking off, Piailug selected a proposed route, taking into account factors such as prevailing winds and currents, expected weather for the season, availability of convenient astronomical bodies, and the capability of the vessel. He also established alternate routes in case things didn't work out as planned. The vessel first sailed northeast, which was deliberately in the wrong direction. This was done to sail as far east as possible before losing sight of the Hawaiian Islands. The navigation was done by terrain association on this section of the voyage, using the islands as landmarks. After skirting around the Big Island, the course was set to the southeast, in the direction where the star Antares rises. This deliberate eastward offset was to compensate for strong, west-flowing ocean currents.

As the vessel moved southward, the height of Polaris above the horizon was used to estimate the latitude while Piailug used dead reckoning to estimate the longitude. He calculated the speed of the boat by looking at how fast the water flowed past the hull and used patterns of ocean swells to keep the vessel pointed in the right direction during periods of cloud cover. Every slight change in wind direction was noted. Between latitudes 8°N and 4°N, the crew encountered the doldrums, a region of weak and unpredictable winds. The vessel made slow southerly progress during this portion of the voyage, but was thankfully pushed eastward about 200 km by east-flowing currents.

The crew inferred that they had crossed the equator shortly after seeing Polaris very low on the horizon. From here, the course was set toward the Southern Cross when it was halfway up its maximum height above the horizon. This bearing, equivalent to south-southeast, was calculated to compensate for west-flowing currents. This bearing would also direct the vessel toward the Tuamotu Archipelago, a target several hundred kilometers wide. Zenith stars were used to calculate the latitude. At one point, Spica passed almost straight overhead putting the vessel at about 11°30'S, and later Zubenelgenubi put their latitude at 15°S.

Unfortunately, the southeast trade winds were too southerly that year and could not compensate for the westerly current. Piailug had a backup plan in case the westward drift caused the boat to miss the Tuamotu Archipelago. Instead of trying to directly find a small target like Tahiti, he would sail northeast against the current to approach the Tuamotu Archipelago from the southwest. After reaching and identifying one of the islands, he would know the boat’s exact location and could then sail straight for Tahiti. This plan was not needed because the familiar swells from the southeast abruptly stopped and a crewmember noticed a couple of terns that normally don’t venture more than 50 km from land. This could only mean they were very close to land. A short time later, the crew spotted Mataiva, the westernmost island of the Tuamotu Archipelago. From here, Tahiti was an easy navigation target with cloud formations above its mountains visible far out to sea.

Since then, several other long, open ocean voyages have been successfully completed without any navigation instruments. This type of navigation requires years of practice and is not an exact science. Your presumed position could be off by several hundred kilometers. It does however teach an important lesson. The world is filled with navigational clues that can be exploited by a skilled navigator.

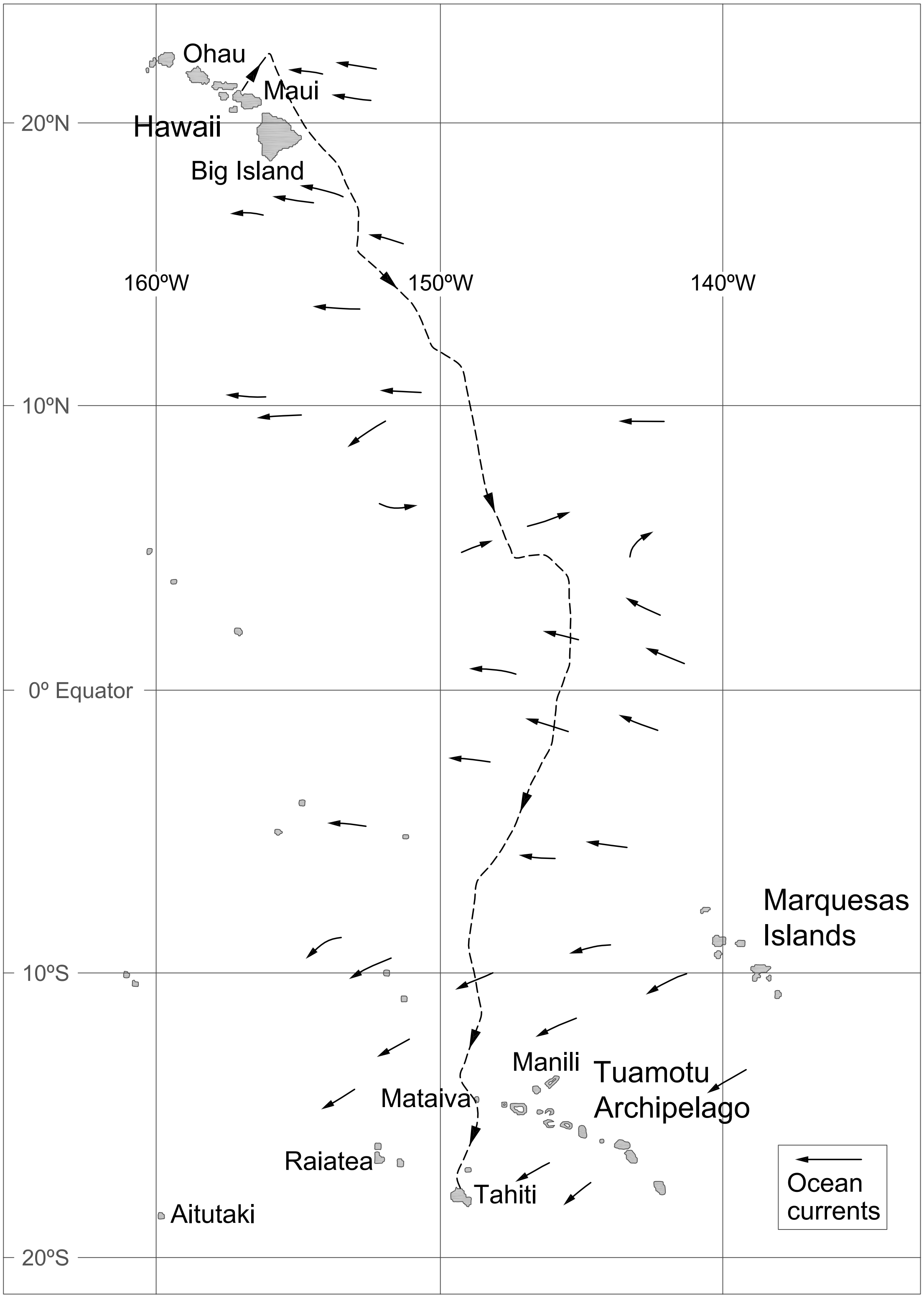
Ohau
Maui
Hawaii
Big Island
20ºN
160ºW
150ºW
140ºW
10ºN
0º Equator
Marquesas Islands
10ºS
Manili
Tuamotu Archipelago
Mataiva
Raiatea
Tahiti
Aitutaki
Ocean currents
20ºS

1976 Maui to Tahiti voyage

Index

A

accurate time 91
acoustic devices 144
Acrux 113, 115
air density 78
air pressure 76
Alaska Highway 179
altimeter 76–83, 156
  calibrating 77
  electronic 76
  mechanical 76
  theory 76–77
altimeter, cause of inaccuracy
  air temperature 78
  altimeter temperature 79–83
  humidity 78–79
  weather systems 79
altitude (celestial) 117, 122, 123
Amazon jungle 174
angle units 15
animal trails 183
Antarctica 189
antenna
  directional 136, 137
  external 90
  wire 140
arête 37
Aries meridian 98
avalanche transceiver 188

B

back bearing 62
back roads 187–188
bar scale 14
base plate 57
bearing 47, 48–49
  geographic 48
  grid 49, 92
  keeping track of 151
  magnetic 49, 92
  true 48–49, 92
bearing marker 57, 67
bench 148
Big Dipper 112
birds 133
body signals 145
boiling water
  table 135
  temperature 186
branch growth 130

C

calibrating altimeter 77
canoe 171, 185
canyon 39, 161
cardinal directions 48, 160
Cassiopeia 113
catching feature 147, 157, 174
celestial coordinate system 97–98
celestial equator 97
celestial meridian
  Greenwich 97
  local 114
celestial navigation 97–129
celestial navigation calculator 129
celestial pole 97
celestial sphere 97
cellular phone 140
Centaurus 113
central longitude 100
central meridian 20, 25
chart 41, 74
Cih 113

circumventing obstacle 62
cliff 38
clouds 133
col 38, 41
colors 12
communication
  electronic 137–141
  emergency 136–145
compass 43–75, 149, 162
  electronic 47, 95
  marine 67–71, 74
  mirror 57
  needle 46, 58
  orienteering 72
compass bearing
  off map 58–59
  off nautical chart 68
compass housing 58
compass needle 134
compass rose 53, 68
constellation 109
contour interval 33
contour lines 40
  definition 33–34
  index 34
  unlabeled 34
cornices 132
correction grid 88
correction signals 87, 90
couloir 37
Crux. *See* Southern Cross
curvimeter 15
curving when walking 180
Cygnus 115

D

datum
  horizontal 30–32, 92, 93
  vertical 32, 92
daylight saving time 100
dead reckoning 123, 151–156
  estimating position 155–156
  predetermined route 153–158
declination
  diagram 51, 73
declination (celestial) 97, 112, 113, 124
declination table 56
degrees 15, 48
deliberate signal degradation 86
Deneb 115
depression 40
descent off mountain 178
desert 176
detouring 152–153
DGPS 87, 90
differential GPS. *See* DGPS
dip, magnetic 56–57
direction arrow 96
disappearing trail 183
distance
  calculating 13–14, 91
  keeping track of 151–152
  measuring 14–15
doppler shift 136
downhill tracks 170
drift 152
  by currents 70
  by wind and waves 70
  compensating for 70
Dubhe 113

E

Earth's shape 30
easting 94
  in UPS grid 28–29
  in UTM grid 25
ecliptic 101
electronic route 92, 94
electronic trail 91, 93
Elephant Island 189
ellipsoid 30, 32
ELT 136
emergency call procedure 139–140
emergency locator transmitter. *See* ELT

emergency position indicating radio beacon. *See* EPIRB
emergency transmitters 136–137
EPIRB 136

F

fall equinox 101
family radio service. *See* FRS radio
ferromagnetic material 47, 134
field bearing, 79
  with marine compass 69
  with orienteering compass 60–61
finding your way out 166
fires 143
flares 142
flashlight 143
flowers 130
following compass bearing
  on land 61–62
  on the water 70–71
FRS radio 138–139

G

Gacrux 113, 115
geographic coordinates 15–18
  on maps 17–19
geographic position 117, 124
geoid 30, 32
global positioning system. *See* GPS
GPS 84–96
  navigation 163
  no signal 186
  satellites 84
  theory 84–85
GPS error, cause of
  atmospheric conditions 85, 87
  deliberate degradation 86, 87
  malfunctioning satellites 86
  multipath reflection 85
  satellite clock 86
  satellite distribution 86
  satellite orbit 86, 87
  signal obstructions 86
GPS receiver 89–90, 158, 163
  basic use 92–93
  initialize 92
GPS receiver type
  dual frequency 89
  hybrid 89
  multiplexing 89
  parallel channel 89
  sequential 89
GPS time 91
gradians 48
gradient 35
graduated dial 58
graduated dome 67
great circle route 167–168
Greenwich mean time 100
Greenwich meridian 16
grid
  in UPS 28–29
  in UTM 21–28
  on maps 26–27
grid search 95, 96, 177
ground symbols 144
gully 37, 40, 157

H

Hadar 113
Hawaiian Islands 191
HF radio 140
hour angle
  Greenwich 97
  local 98, 125
  sidereal 98, 112
humidity 79

I

inclination 56
inclinometer 57
index arm 122
index mirror 122
intercept, altitude 126
ionosphere 85, 138

J

jungle 174

K

knot 70, 155

L

lake 185
lakeshore 171
landforms 132
large group 181–182
large-scale map 13
latitude 16
  definition 16
legend 12
line of position 80, 127
line of position circle 118–121
linear feature 12, 63, 146
Little Dipper 112–113
longitude 16
  definition 16
lost 164–166

M

Mackenzie River 177
magnetic card 67
magnetic declination 50–56
  compensating for 55–56
  mechanism 58
  theory 50–51
magnetic declination, calculating
  direct measurement 54
  with compass rose 53
  with declination diagram 51–53
  with GPS receiver 54
  with software 54
magnetic dip 56–57
magnetic field 43–46
  theory 43–44
magnetic pole 44–46
magnetic pole movement
  diurnal 45–46
  erratic 46
  secular 45
maps 12–42
  digital 15
  topographic 33–41
marine compass 67–71
Markab 98
markers 173
mayday call 139
mean low water level 42
mean sea level 32
Merak 113
meridian. *See* longitude
meridian lines 58
mil 48, 53
milliradian. *See* mil
minutes of arc 15
mirror 142
mirror line 58
moon 106–109
moon's bearing
  calculating 108–109
  example 108
  table 107
moon's phase 106–109
Mount McKinley 186
mountain top 172

N

natural navigation 130–135
nautical chart 41
nautical mile 16
nautical miles per hour. *See* knot
navigating
  by dead reckoning 151–156
  by terrain association 146–148
  with map and compass 149–150
  with map, compass, and altimeter 156–158
  with map, compass, and GPS 158–160

navigating without maps
  with just compass 162
  with just GPS and compass 163–164
  with nothing 160–161
northing 94
  in UPS grid 28–29
  in UTM grid 25

O

obstacles 152–153
ocean clues 133–134
ocean current 155
odometer wheel 152
official time 100
orienteering compass 57–67
orienting a map or chart 71–75
  by terrain association 71–72
  using compass and grid 72
  using compass rose 74–75
  with declination diagram 73–74
orienting arrow 58
Orion 114

P

pan-pan call 139
panoramic view 172
parallel. *See* latitude
parallel rivers 132
parallel rulers 68
pass 38, 40, 150
personal locator beacon. *See* PLB
phase shift 85
Piailug, Mau 191
PLB 137–138
Polaris 112–113, 116
polynesian journey 191–192
position averaging 90
position fix 90
practical navigation 146–168
pressure gradient 76
prevailing winds 131
prime meridian 16
projections 18–21
  definition 18
  Lambert conformal conic 19
  Mercator 19
  polar stereographic 21
  universal transverse Mercator 20

R

radio
  channel 137
  frequency 137
radios 137–141
range 138
ratio scale 13–14
real life navigation 164
reference receiver 87
refraction 138
rhumb line 167–168
ridge 37, 41, 157, 175
ridgeline 41
Rigel 114
Rigil Kentaurus 113
river delta 177

S

Sahara desert 176
salinity 134
sand dunes 132
sastrugi 132
satellite phone 141
scale
  bar 14
  ratio 13–14
  verbal 13
scenarios 169–192
seamarks 134
secant lines 20
seconds of arc 15
sextant
  mechanism 121–122

sextant navigation
  apparent altitude 124
  assumed position 118, 125
  azimuth 126
  calculations 123
  compensating for sight errors 123
  computed altitude 126
  index error 123
  intercept 119
  local hour angle 125
  observed altitude 123, 124
  parallax 123
  shift line of position 128
  sight reduction tables 126
  taking a sight 117, 118, 122, 123
  theory 117
  upper or lower limb 123
Shackleton, Ernest 189
shoreline 171, 185
sight
  errors 123
  taking 122
sighting notch 58
signal gathering techniques 89
signal laser 143
signaling systems 141–145
signature signal 84–85
signs of land 133
ski tracks 170
ski traverse 188
slope 130
  angle 35, 187
  calculating 34–36
  gradient 35
small-scale map 13
solar wind 45–46
SOS signal 143
soundings 42
South Georgia 190
Southern Cross 113, 115
space-based augmentation 88, 90
special UTM zones 24
spring equinox 98, 101
standard parallel 19, 21
star compass 116
star map 110, 111
stars 109–117
  as direction finders 116–117
  finding north 112–113
  finding south 113
  finding your latitude 113–114
  finding your longitude 114
  meridian transit 114–116
  rising-setting 112
  transit pairs 115
  zenith 113–114
stash, finding 188–189
steering mark 61–62, 149, 153
steering stars 116
step count 154
step counter 151
stick shadow method 105
stride length 151
stump 130
summit 39, 41, 157, 178
sun 99–106
sunlight, effects of 130
sun’s bearing
  example 103
  finding 99–106
sun's bearing error
  axis tilt 100–101
  elliptical orbit 100–101
  observer’s latitude 102–103
  observer’s longitude 100
swells 133
switchbacks 169
symbols 12

T

Tahiti 191
tangent plane 103
terrain
  features 36–40
  how to interpret 40–41

terrain association 146–148
The Pointers 113
thermometer 135
time zone 100
trail 169, 173, 179, 183, 185
travel time 152
travel-direction arrow 58
trees 130
triangulating 172
  by steepest slope 81–83
  from linear feature 63–65
  off multiple landmarks 65–67
  with altimeter 79–81
troposphere 85
truncated grid numbers 27
Tuamotu Archipelago 192
two-dimensional fix 86
two-way radio 137–141

U

UHF radio 139–140
universal date 98, 122
universal time 91, 92, 98, 100, 122
universal transverse mercator. *See* UTM
uphill tracks 170
UPS grid 28–29
UPS zones 28
Ursa Major 112
Ursa Minor 112–113
UTM grid 21–28, 94
  on maps 26
UTM zones 21–24
  special 24

V

valley 36, 148
variation 50
verbal scale 13
VHF radio 139–140

W

waiting for a rescue 165
wake, boat 152
walkie-talkie 138–139
wands 189
water vapor 78
wave patterns 133
waypoint 149, 153, 163
  inserting 91, 93–95
  navigating toward 91
  pre-entered 158–159
  precision 158
  route 91
  storing 90
  zeroing-in 91
whistle 144
whiteout 183–184
Wickersham Wall 186
wind, effects of 131–132
Worsley, Frank 190

Z

zenith 117
zenith stars 192
zeroing-in on a waypoint 91, 95
zones
  UPS 28
  UTM 21–24

"The Wilderness Navigation Handbook does an excellent and thorough job of presenting navigation concepts in a concise and clear way. It is the principal reference document that I recommend to those taking my course on navigation techniques."

—**Dave McRitchie**
geologist and member of Fernie Search and Rescue

"This book is excellent if you have never used a compass or a map, it is excellent if you are a fully trained mountain leader and it is excellent as a resource for teaching adults and children up to advanced level navigation techniques."

—**Andrew Packer**
director of Nomad Bushcraft, a British survival school

ANSWERS TO INTRIGUING QUESTIONS

- Is it true that altimeters read higher when it's cold?
- Does a compass needle point directly toward the nearest magnetic pole?
- Why is my GPS position sometimes totally off?
- Is it possible to navigate through featureless terrain without instruments?
- Can I use the moon to find directions?
- How do you create a flat map from a round globe?

THE AUTHOR

Fred Touche has always been fascinated by exploration and wild places. He has completed several long, unsupported ski traverses through Canada's Coast and Saint Elias Mountains, as well as climbed the highest peaks in North America. This is his first book.

TOUCHE PUBLISHING

US$17.95/CAN$22.95
ISBN-13: 978-0-9732527-0-5
ISBN-10: 0-9732527-0-7
51795>

OUTDOOR/HIKING & CAMPING

CANADA: $22.95 • USA: $17.95
Printed in Canada

9 780973 252705

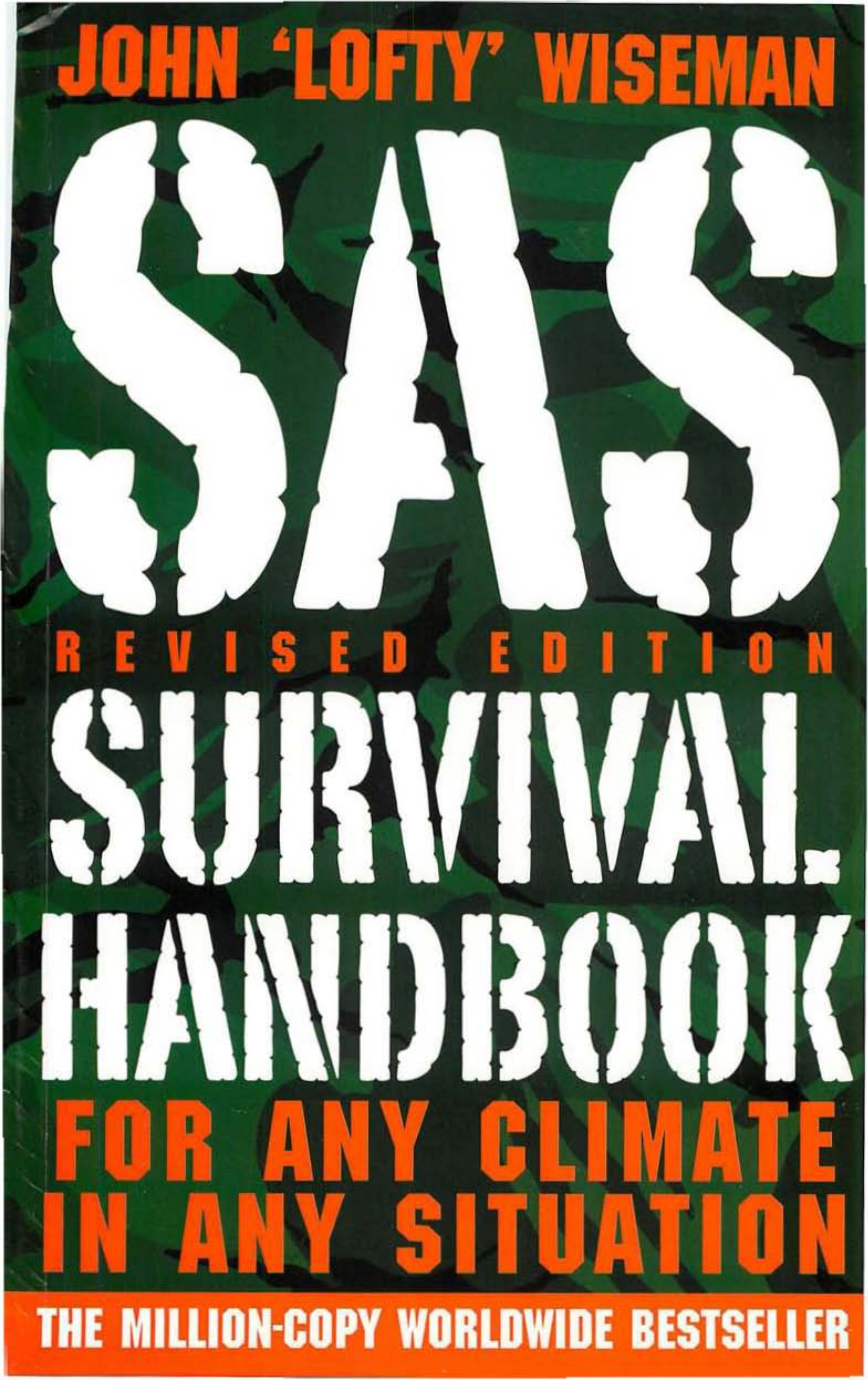

JOHN 'LOFTY' WISEMAN

SAS

REVISED EDITION

SURVIVAL HANDBOOK

FOR ANY CLIMATE IN ANY SITUATION

THE MILLION-COPY WORLDWIDE BESTSELLER

FOR ANY CLIMATE, IN ANY SITUATION

Newly updated to reflect the latest in survival knowledge and technology, the internationally bestselling *SAS Survival Handbook* is the definitive resource for all campers, hikers, and outdoor adventurers. From basic campcraft and navigation to fear management and strategies for coping with any type of disaster, this complete course includes:

BEING PREPARED: Understanding basic survival needs and preparing essentials, such as a pocket survival kit.

MAKING CAMP: Finding the best location, constructing the appropriate shelter, organizing camp, and creating tools.

FOOD: What to eat, what to avoid, where to find it, and how to prepare it.

FIRST AID: A comprehensive course in emergency/wilderness medicine, including how to maximize survival in any climate or when injured.

DISASTER SURVIVAL: How to react in the face of increasingly frequent natural disasters and hostile situations—and how to survive at home if all services and supplies are cut off.

JOHN 'LOFTY' WISEMAN served in the British Special Air Service (SAS) for twenty-six years. The *SAS Survival Handbook* is based on the training techniques of this world-famous elite fighting force.

NOW FEATURING
CASE STUDIES
AND SURVIVAL
SCENARIOS

COLLINS
An Imprint of HarperCollins*Publishers*
www.harpercollins.com

Visit www.AuthorTracker.com for exclusive information on your favorite HarperCollins authors.

SPORTS & RECREATION/Camping

ISBN 978-0-06-173319-2
51999
9 780061 733192

$19.99

SAS SURVIVAL HANDBOOK

JOHN 'LOFTY' WISEMAN

SAS SURVIVAL HANDBOOK

REVISED EDITION

FOR ANY CLIMATE, IN ANY SITUATION

COLLINS
An Imprint of HarperCollinsPublishers
www.harpercollins.com

This new edition published in 2009 by
Collins, an imprint of HarperCollins Publishers
10 East 53rd Street, New York, NY 10022

Produced by Essential Works
www.essentialworks.co.uk

First published by Harvill 1986
Text © John Wiseman 1986, 2003, 2006, 2009
Illustrations © HarperCollins 1986, 2003, 2006, 2009

John Wiseman asserts the moral right to be identified as the author of this work. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher and copyright owners. For information, address HarperCollins Publishers, 10 East 53rd Street, New York, NY 10022.

HarperCollins books may be purchased for educational, business, or sales promotional use. For information, please write: Special Markets Department, HarperCollins Publishers, 10 East 53rd Street, New York, NY 10022.

CREDITS/ACKNOWLEDGEMENTS

The colour illustrations were painted by Norman Arlott. Other illustrations were drawn by Steve Cross, Chris Lyon, Andrew Mawson and Tony Spalding. New design on the 2003 edition by Paul Collins.

The editors would also like to thank Howard Loxton and Christopher McLehose.

ISBN 978-0-06-173319-2

10 11 12 13 / 10 9 8 7 6 5 4

Printed and bound by Printing Express, Hong Kong

For my father

T. C. H. WISEMAN

who showed me the value of life
and whose dedication to helping
others I try to follow

WARNING

The survival techniques described in this publication are for use in dire circumstances where the safety of the individuals is at risk. Accordingly the publishers cannot accept any responsibility for any prosecutions or proceedings brought or instituted against any person or body as a result of the use or misuse of any techniques described or any loss, injury or damage caused thereby. In practising and perfecting these survival techniques the rights of landowners and all relevant laws protecting certain species of animals and plants and controlling the use of firearms and other weapons must be regarded as paramount.

CONTENTS

INTRODUCTION 8

1. ESSENTIALS 12

2. STRATEGY 50

3. CLIMATE & TERRAIN 62

4. FOOD 108

5. CAMP CRAFT 244

6. READING THE SIGNS 348

7. ON THE MOVE 372

8. HEALTH 392

9. SURVIVAL AT SEA 480

10. RESCUE 504

11. DISASTERS 528

POSTSCRIPT 572

INDEX 573

INTRODUCTION

In an increasingly uncertain world, the focus of this new edition is to prepare you for any survival situation you may encounter. Within these pages I look to identify every danger imaginable and show you how to deal with it. The advice given here will prepare you both physically and mentally, and make you aware of all the hazards you may face. Being fully prepared will give you the confidence to survive against the odds – and allow you to adapt when things go wrong.

For 26 years, as a professional soldier, I had the privilege of serving with the Special Air Service (SAS). This elite unit of the British Army is trained to carry out arduous operations in all parts of the world, often isolated and far from conventional forces. Working in small groups, often in enemy territory, SAS soldiers have to become their own doctor, dentist, navigator and cook. Frequently, the operational situation makes resupply impossible and they have to live off the land. They have to handle every kind of situation and problem, whether man-made or an act of nature, and get back to safety, and must develop skills which enable them to survive anywhere.

After serving throughout the world, I became the survival instructor to the SAS and it was my responsibility to ensure that each and every member of the Regiment could apply these skills in the field. Tested in training and operations, they form the basis of this new edition.

The need for survival training has never been greater. Cheaper air travel has opened up the world, and there is virtually no place left on the face of the planet that we cannot and have not visited. This, added to factors such as increased leisure time, wealth, and the knowledge provided by television programmes, means that we are becoming more adventurous and informed about the world around us.

But we must also take into account the fact that over the last 12 years the world has become a more unstable place. The Berlin wall came down in 1999, followed by the Gulf War, Bosnia, Kosovo and Sierra Leone. These events, however, were overshadowed by the catastrophic and tragic events of 11th September 2001, the resultant 'war on terror' and, of course, the wars in Iraq and Afghanistan. These global conflicts have affected everyone and one things is for sure – our lives will never be the same again.

The instinct to survive will never change, neither will the human body's amazing ability to endure. Yet we should also bear in mind that, as we become more civilized, survival skills are being forgotten. It is doubly important, therefore, to practise our survival skills, and be prepared for anything. Survival training is the best insurance policy you can take out in an unstable world. If we are all grounded in the basic techniques of survival, and know what to do in an emergency, the world immediately becomes a safer place.

You could be isolated anywhere in the world – from the arctic ice to a desert, from tropical rainforest to the open ocean, and the problems of survival are the same for both soldier and civilian. The difference lies in the circumstances; soldiers may need to hide their presence, whereas civilians will want to attract attention to effect their rescue. In areas of instability, where the risk of kidnap is high, it may be best to remain inconspicuous. Each environment calls for special survival techniques. Mountains, jungles, open plains and swamps can seem hazardous to the survivor, but each offers some form of support and can be exploited for food, fuel, water and shelter – if you know how. The effect of climate is crucial. Intense cold and searing heat challenge the survivor in different ways. You must know how to cope with both.

You also need a sense of how your body works and how it will react in different environments and situations. Assess yourself before you travel: for instance, test your basic survival skills by sleeping outside, making sure that you can light a fire, and even going without sleep or food for a period of time. This will show you how you react in a limited simulated scenario.

Survival depends upon applying basic principles and adapting them to the circumstances. These basic principles form an essential pyramid of learning for the survivor.

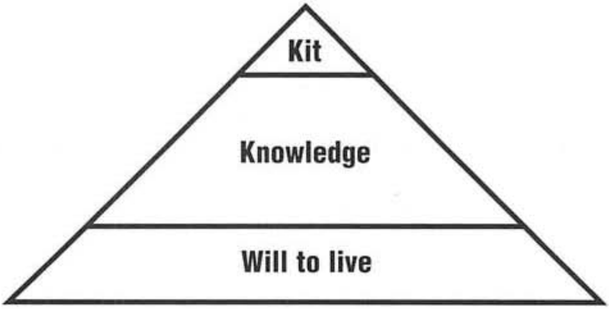

Although initially it may take great physical effort to escape a dangerous situation, survival is above all a mental exercise. After the excitement of the incident and the rush of adrenalin has settled it takes great mental resolve to carry on. What keeps us going is the basic instinct, best referred to as 'the will to live'.

Will to Live

This is the firm foundation on which we build all of our training. We constantly try to nourish and increase the will to live. It's easy to see how physically fit we are but very difficult to know how mentally fit we are.

The Will to Live (WTL) means never giving in, regardless of the situation. It's very reassuring to know that there is nothing on this earth that we cannot deal with, and there is no place where we cannot survive. As long as we follow the basic survival principles, prepare ourselves, and apply this WTL, we will come through. Some people have a stronger will than others, but we can all improve.

Some of us turn to religion in times of stress, others think of loved ones. Fear of failure or letting down comrades all help to strengthen our WTL. Reading about past exploits of survival will also help. You can have all the knowledge and kit in the world but without the will to live you can still perish.

Knowledge

On top of our foundation we have Knowledge. The more we know the easier it is to survive. Knowledge dispels fear. Look at the locals and see how they survive. Talk to people who have endured and learn from their experiences. On a practical level, ensure you are fully informed about the medical infrastructure in the particular region you are travelling in and know exactly how the emergency services operate.

Kit

The tip of our pyramid of learning is Kit. We keep this to a minimum and have a thorough knowledge of its uses and capabilities. There are essential items that you should never leave home without, including your Survival tin, knife, compass and a radio/phone. Remember, however, that you must also focus on what nature can provide to protect you as, ultimately, any kit must be considered a bonus. It is important that you know how to improvise and use your surroundings to sustain you when faced with a survival situation.

It is essential that you use your own judgement in the application of the methods described in this book. The tests for plant foods, for instance, are the only sure way of being certain whether a particular fruit or leaf is safe or poisonous. The average person is unlikely to come to any harm if they follow the method carefully, but there is always an element of risk. Individual responses to poisons vary – even small quantities of toxic substances can be very dangerous to some people. Some of the traps described are also very dangerous. They can inflict injury to you if handled carelessly and should never be left unattended where other people may come to harm.

In learning the skills described here, I would always advise that you keep in mind the need to conserve our environment and to avoid cruelty to animals, and to be aware that some of these techniques may contravene local bylaws. While a knife is a useful item in the wild, remain aware of contemporary knife laws. Remember, this is a handbook for survival situations when self-preservation is paramount. Therefore, risks may be involved which would be foolish even to consider under normal circumstances.

Although this is not an official publication, by sharing the survival knowledge that I and my colleagues have gained through experience, I aim to help you to make those decisions correctly. These methods and skills have helped save our lives – they will help you to be a survivor, too.

I would like, finally, to thank the SAS Regiment for providing me with the experience on which this book is based and to thank Howard Loxton and Tony Spalding who helped to bring it to publication. Without their hard work and dedication this book would not have been possible.

J.W.
The Survival School, Hereford

1

ESSENTIALS

Begin by preparing yourself to be a survivor. That means preparation in every sense. This section is concerned with making sure that you have the right equipment for any expedition you undertake. It introduces the idea of carrying a pocket-sized kit of carefully selected key survival aids – which should go with you everywhere.

A knife is your most important survival tool. It must be chosen and used carefully and it must be kept in perfect condition.

Equally important is a personal preparedness, so that you are both physically and psychologically equipped to deal with the stresses and hazards of survival conditions. You must have a clear understanding of survival needs, especially of the need for – and ways of obtaining – water.

ESSENTIALS FOR SURVIVAL 14
  Be prepared 16
  Research 17
  Planning 18
  Equipment 19
  The unexpected 27
SURVIVAL KIT 28
SURVIVAL POUCH 32
KNIVES 35
FACING DISASTER 38
  Basic needs 39
WATER 40
  Finding water 41
  Condensation 43
  Water from plants 46
  Water from animals 48
SALT 49

ESSENTIALS FOR SURVIVAL

The human species has established itself in almost every corner of the Earth. Even in territories too inhospitable to provide a regular home mankind has found a way to exploit its resources, whether by hunting or by taking wealth from the ground, and has often pitted its skills against nature simply for the satisfaction of doing so.

Almost everywhere nature provides the necessities for survival. In some places the provision is abundant, in others very meagre and it takes common sense, knowledge and ingenuity to take advantage of the resources available. Even more important is the will to survive. Men and women have shown that they can survive in the most adverse situations, but they have done so because of their determination to do so – without that, the skills and knowledge in this book will be of little use if you find yourself really up against it.

Survival is the art of staying alive. Any equipment you have must be considered a bonus. You must know how to take everything possible from nature and use it to the full, how to attract attention to yourself so that rescuers may find you, how to make your way across unknown territory back to civilization, if hope of rescue is not on the cards, navigating without map or compass. You must know how to maintain a healthy physical condition, or if sick or wounded heal yourself and others. You must be able to maintain your morale and that of others who share your situation.

Lack of equipment should not mean that you are unequipped, for you will carry skills and experience with you, but those skills and experience must not be allowed to get rusty and you must extend your knowledge all the time.

We are all used to surviving on our home ground – though we may not think of our lives in that way – but the true survivor must learn how to survive when taken from familiar surroundings or when those surroundings are drastically changed by man or nature. Anyone, young or old, from whatever walk of life, can find him- or herself in a survival situation. As more and more people fly the globe, sail small boats or cross the sea in large ones, walk the hills and climb mountains and take their holidays in ever more exotic places, the situations to which they could become exposed are increasingly diversified.

But survival skills are not only concerned with the extremes of the air crash on a mountain peak, a shipwreck in the tropics or a vehicle breakdown in the middle of a desert. Every time you fasten a seat belt in a car you are giving yourself a greater chance of survival. Checking each way before crossing a road or ensuring that an open fire is safe before you go to bed are survival techniques that you carry out instinctively. It is these habits of mind that you must develop as much as acquiring skills.

The main elements of survival are Food, Fire, Shelter, Water, Navigation and Medicine. To put these in order of priority we use the acronym **PLAN**. No matter where you are in the world this will never change be it the Arctic, desert, jungle, sea or seashore.

P – for Protection

You must ensure that you are protected from further danger, i.e. impending avalanche, forest fire or exploding fuel. Always stay on the scene of the incident as long as it is safe to do so and then make sure you are protected from the elements. This means making a shelter and often lighting a fire. There are several reasons why you should always stay at the scene:

1 You can utilize the wreckage for shelter, signaling etc.
2 It's a bigger signature on the ground, making it easier to find.
3 There are probably injured people that cannot be moved.
4 By staying where you are you conserve energy.
5 Because you have booked in and out and have stayed on the route, rescue time will be minimal.

L – for Location

The next step after building a shelter is to put out emergency signals. You must draw attention to your position. Do this as soon as possible to help the rescuers.

A – for Acquisition

While waiting to be rescued, look for water and food to help supplement your emergency supplies.

N – for Navigation

Good navigation will keep you on route and will often avert a survival situation. But if you find yourself stranded, always stay where you are.

Medical

You must become your own doctor and carefully monitor yourself at all times. Treat blisters as they occur, don't let them become septic. Keep an eye on your companions and deal with any unusual problems as they arise. If they are limping, falling behind, or behaving strangely, stop and treat immediately.

BE PREPARED

The Boy Scouts' motto is the right one. Anyone setting out on a journey or planning an expedition should follow it by discovering as much as possible about the situations likely to be faced and the skills and equipment called for. It is the most basic common sense to prepare yourself, to take appropriate gear and to plan as carefully as possible.

Your kit could make the difference between failure and success, but, especially when back-packing, many people initially take too much and have to learn from bitter experience what they really need and what they could have done without. There is no fun in struggling with a huge pack full of superfluous items while wishing that you had a torch or can opener with you. Getting the right balance is not easy.

EQUIPMENT CHECK LIST

Before any journey or expedition make a check list and ask yourself the following questions:

- How long will I be away? How much food do I need for this period and do I need to carry water?
- Have I the right clothing for the climate and enough of it? Is one pair of boots enough or, because of the surface conditions and the amount of walking, should I take a standby pair?
- What special equipment do I need for the terrain?
- What medical kit is appropriate?

Make sure that you are fit enough for what you plan to do. The fitter you are, the easier and more enjoyable it will be. If you are going hill-walking, for instance, take regular exercise beforehand and wear in your hiking boots. Walk to and from work with a bag weighted with sand and get your muscles in condition! Mental fitness is another factor. Are you sure that you are up to the task, have prepared enough and have the equipment to accomplish it? Eliminate any nagging doubts before you set out.

Always prepare contingency plans in case anything goes wrong. Things rarely go quite according to plan. What will you do if you are prevented from achieving your objective? What will you do if a vehicle breaks down, or if weather or ground conditions prove more severe than anticipated? If in a party, how will you regroup if separated? What happens if someone becomes ill?

Health checks

Have a thorough medical check and ensure that you have all the necessary injections for the territories through which you intend to travel. There are vaccinations against yellow fever, cholera, typhoid, hepatitis, smallpox, polio, diptheria and tuberculosis, and an anti-tetanus injection is a must. Allow plenty of time for jabs – the full anti-typhoid protection requires three injections over the course of six months. If travelling through a malarial region take an adequate supply of anti-malaria tablets. You must start taking these two weeks before your journey, so that resistance is in the system before you arrive in the risk area, and should keep taking them for a month after your return.

Go to the dentist and get your teeth inspected. Teeth that normally do not hurt can cause considerable pain in cold climates. At least start out in sound condition.

Make up a medical kit that will cover all your likely needs and, if travelling with a group, ensure that any particular individual medical needs are covered. If a potential member of the group is not fit, should they be dropped from the party? A difficult decision among friends, but one that must be made for it is best in the long run. Consider, too, the ability of each member of the group to deal with the challenge of hardship, risk and endurance that you may meet. Stress often brings out the unknown side of a person, and in planning any group expedition some form of selection is needed when choosing your companions.

RESEARCH

You can never have too much information about a place you are going to. Contact people who know it already, read books, study maps – and make sure that you have reliable and up-to-date maps to take with you. Find out about the local people. Are they likely to be friendly and helpful or are they wary of strangers? Are there local customs and taboos?

The more detailed your knowledge of the way people live – particularly in non-westernized societies, where life is linked much more closely to the land – the more survival knowledge you will have if you come to need it. Local methods of shelter building and fire making, wild foods, herbal medicines and water sources will be based on an intimate understanding of the surroundings.

Study your maps carefully, get a feel for the land even before you see it and gain as much knowledge of the terrain as possible: river directions and speed of flow, waterfalls, rapids and difficult currents. How high are hills and mountains, and what are their slopes like – are they snow covered? Which way do the ridges run? What kind of vegetation can you expect, what species of trees and where? What might temperatures be and how different at day and night? When are first and last light? What is the state of the moon, the time and height of tides, the prevailing wind direction and strength? The weather that can be expected?

PLANNING

For a group expedition get the members together for frequent discussions of what you aim to achieve. Nominate people for particular responsibilities: medic, linguist, cook, special equipment, vehicle maintenance, driver, navigator and so forth. Ensure that everyone is familiar with the equipment and that there are spares where needed – batteries, fuel and bulbs especially.

Divide the project into phases: entry phase, objective and recovery. Clearly state the aim of each phase and work out a time scale. Plan for emergency procedures such as vehicle breakdown, illness and casualty evacuation.

In estimating the rate of progress, especially on foot, allow plenty of time. It is always better to underestimate and be pleasantly surprised by doing better. Pressure to keep up to an over-ambitious schedule not only produces tension and exhaustion but leads to errors of judgement and risk-taking that are frequently the reason for things going wrong. You cannot carry all your water requirement with you but must replenish supplies as you travel. Water sources will be a major factor in planning any route.

When the route is planned and agreed make sure that others know about it so that you can have expectations of rescue if anything goes wrong. If you are hiking in the hills inform the police and local mountain rescue centre. Tell them your proposed plan and give times of departure and expected arrival. If touring by car, log the route with the respective motoring organization. If sailing, check with coastguard and port authorities.

Always make sure that someone knows what you are planning to do and when, and keep them informed at prearranged stages so that failure to contact will set alarm bells ringing. Boats and aircraft are strictly controlled in this respect and, if overdue, a search is raised and the route checked out, effecting rescue. Get into the habit of telling people where you are going and what time you expect to return or reach your next destination.

EQUIPMENT

Being prepared for any eventuality is a tall order if you are on foot and have to carry everything you need yourself. Whatever you carry, you must ensure that it is up to the job, versatile and robust. It's a fine balance between what you would like to carry and what you must carry. When preparing for any adventure, you must take into consideration what the dangers are and how you can overcome these. This is what is called contingency planning.

The climate, weather and time of year will all help you to determine what to carry, but you must ensure that everyone with you knows how to use and maintain the specialist kit you decide to take with you. Armed with information from your research you will be able to select your equipment, matching it to objectives and conditions.

Clothing

The correct choice of clothing is so important. If you start out right the chances are that you will succeed. Man is a tropical animal and can only survive as we are born in the tropics. The moment we leave this area we have to provide our bodies with this tropical environment, hence the need for clothes. There is no heat in clothing, it only traps what the body produces.

The wind and rain are the most dangerous elements in a temperate climate and the cold in extreme areas like the polar regions. If the heat that is trapped in the layers of clothing you are wearing is continuously being replaced by wind and rain, you are in danger of hypothermia. In cold climates layering is the answer so pull on a jersey if it turns cold and waterproofs if it rains. However if you wear an anorak while carrying a heavy pack, there is a danger of wearing through the shoulders and lower lumbar region allowing the ingress of water to soak the body. You need a change of clothing and additional warm garments for when you stop.

In hot climates it is very difficult to get the balance right between comfort and practicality. There has always been a danger of overheating in extreme conditions caused by wearing heavy clothing while carrying out physical activities. When on the move wear the least amount of clothing possible and avoid walking in waterproofs if you are too hot, as the condensation generated will soak the inner layers.

Clothing should give good protection and be well-fitting without being restrictive. It must keep you warm and dry but have plenty of ways to keep the body ventilated so you don't overheat (if it gets colder you can always put on more.)

With all the great breakthroughs in recent years in fabric technology it is worth understanding the pros and cons of the different materials on offer. Gore-tex™ is an excellent material because it is breathable and so keeps you warm and dry while ventilating the body, but it does have limitations. Breathable materials can only work if they are kept clean. Once they get covered in mud and accumulate grime they are less effective. Gore-tex™ is not robust or hard-wearing and must be looked after. The best way to use Gore-tex™ is to walk or climb in windproof garments and when at rest, put on the breathable kit.

Synthetic materials such as fleece are very popular and in certain conditions outperform natural materials like wool, down or cotton. Having a zipped front makes a fleece easy to put on and take off and they are also comfortable to walk in. Choose one that is windproof as this is often all that is needed in most conditions. If it gets colder they can be worn under an outer waterproof giving good insulation. There are also garments which act like an animal's skin, using the buffalo system. They have a windproof outer with a man-made fibre pile inside. When wet they perform like a wetsuit. They are good for walking in cold/wet conditions, and are ideal for boating, canoeing and caving.

As for natural fabrics, wool is still an excellent choice for jumpers as it retains its warmth even when wet. The downside is it stretches and becomes heavy, so it's not a good choice for socks. Down is the warmest and lightest of all natural insulating materials but loses all its heat-retaining qualities when wet. Cotton acts as a wick and draws up all the moisture. So it's good to wear in the tropics but not in the cold/wet regions.

Footwear is an important consideration and for serious walking give your feet priority. Break in new boots gradually and harden up your skin with surgical spirit, starting two weeks before you set off.

For the enthusiast the major consideration in choosing clothing is cost. Surplus stores are very popular for the younger adventurer who loves to parade in camouflage clothing. Although ex-military kit is good, and cheap, it is already obsolete. The big drawback of wearing camouflage or dark clothing is the risk of not being found when lost. The reason soldiers wear it is so they cannot be seen which contradicts what you are trying to do if you get into trouble. Most outdoor clothing is blue or orange, some is reversible, so a contrasting colour will always stand out wherever we find ourselves. Buy the best clothing you can afford, and take advice from a reputable outdoor shop.

Remember: There is no such thing as bad weather, only bad clothing.

Sleeping bags

Two types are generally available. One kind uses hollow fill, man-made fibre, the other (and more expensive) is filled with down. Down is very light and gives much better insulation – provided it stays dry. If it gets wet it loses all its insulating properties and is very difficult to dry out. For conditions that are likely to be wet the man-made fibre will therefore be the better choice. Avoid getting your sleeping bag wet, however, as sleep will be seriously affected.

Excellent bivouac bags made of breathable material are also available that will keep you dry in place of a tent, but in the long term you cannot beat a tent which can also be used for cooking and communal activities. Keep your sleeping bag inside the bivy bag and stow it inside a compression sack to make it as small as possible. Keep the bag clean and use a kip mat or poncho to lie on.

Packs

You need a strong and comfortable back-pack to carry all your clothing and equipment. Choose the very best you can afford. It should have tough and fully adjustable webbing, well secured to the pack's frame or fabric. Heavy loads can quickly loosen poorly made webbing. It must have a comfortable hip belt. The secret of wearing a pack is to take the weight securely on the hips – the body's strongest pivot – not on the shoulders and back, which quickly strain and tire.

Do you want a pack with an external or an internal frame? Internal frames are lighter and make a pack more easy to stow, but external frames are stronger, ensure a more even distribution of the load and are especially useful for awkward or heavy equipment – including, in an emergency, a sick or injured person. A good external frame should carry the pack high up on your body, putting less strain on hips and shoulders, and it should be designed to allow an airspace between the pack and your back to minimise contact perspiration. A frame adds weight and is more prone to snag on rocky projections or branches, making progress through dense vegetation a little more difficult, but its advantages more than compensate.

Finally, choose a pack made from a tough, waterproof fabric, preferably with a lace-up hood inside the main sack to prevent water leaking in and the contents falling out. Side pockets are always useful, but they must have secure zips rather than straps or drawstrings, which do not hold equipment safely.

Stowing kit

If you expect to get wet, stow everything in polythene bags. Pack so that you know where everything is and so that the first things you need are not buried at the bottom. The sleeping bag is probably the last thing you need so that goes at the bottom. Your tent should be on the top, so should heavy kit such as radios, which are more easily carried there – though try not to make the pack too high, if you have to cope with strong winds, for a very high pack will be more difficult to balance and you will expend a lot of energy just keeping upright.

Pack a stove and brew-kit in a side pocket so that you have easy access when you halt. Make sure that foodstuffs that can be easily squashed or melted are in suitable containers. In a warm climate you can carry food to eat cold and make plenty of hot drinks. In a cold climate make sure that you have plenty of fats and sugars. The exact rations depend on your taste, but they should be chosen to give a good balance of vitamins, minerals, fats, proteins and carbohydrates. Take into account the extent to which you will be able to live off the land and carry a supply of anything unlikely to be available locally.

G.P.S

A G.P.S (Global Positioning System) is an excellent piece of equipment and has taken a lot of skill away from the navigator. Basically these systems receive radio signals from satellites and can locate your current position, anywhere in the world, and are relatively easy to use. It is also useful to note that they are reported to have 95% accuracy rate. However, in order to work, the satellite transmission must not have any obstructions in its way, such as a tree branch or movement, so to receive a clear signal you need to be standing still and out in the open. However, if we depend solely on technology our basic skills will suffer and we will become unstuck if it becomes unserviceable or is lost. G.P.S is not effective unless you can identify where you are so stick to the basics. Map read and navigate normally and use the G.P.S to confirm your navigation or correct it.

When looking to buy a G.P.S there are several considerations to think about: what you'll be using it for – if walking you will want the unit to be as light as possible and compact; where you'll be using it; and if you need it to be waterproof (this is usually a feature of the heavier models with extra gadgets). Battery life should also be taken into account. Some G.P.S are more complicated than others so choose the model that you're happy with. Most have the facility of being able to put in way points (at sea this means the eastern and northern coordinates and on land these can be campsites, rock formations etc) and there are many convenient hand-held versions and some are even featured on watches.

There is always a danger with any battery-operated equipment that it lets you down when you need it most. Batteries always discharge faster in the cold and with age. Recharging facilities are difficult in the wilderness, and bad connections caused by constant abuse while on the move are a real menace.

Carry the G.P.S around the neck tucked under the jacket. This will minimize the risk of damage and protect it from the weather. Don't place it in your pack or leave it lying around.

When planning your route from the map, choose prominent points that can be used as emergency rendezvous. Have these at regular intervals, preferably every hour of walking. Enter these into the G.P.S and they will keep you on track. Once entered they will offer information as to where you are in relation to these points and tell you what direction to take to reach them.

Radios

For a long expedition in remote territory a radio is a necessity. They tend to be expensive but are well worth the cost; if you cannot afford the radio, you cannot afford the expedition. Choose a model with the least amount of channels available to suit your particular needs. The trouble with multi-channeled sets is that people get confused and tend to use the wrong ones. Have a working channel that everyone uses at established schedules. Have a priority channel that you can switch to in an emergency so no one will break into your transmissions. If working with coastguards/forest rangers etc make sure that your radio is compatible and you know the emergency channel (channel 16); knowing the frequency of the World Service is also useful. Keep your radio in a safe place, ideally on a person and not in a pack.

Prearrange a signals plan with scheduled calls morning and evening, especially when working in a large party. A signals plan entails people manning the radio at base and two-way communication is easily made. Make sure that the chosen frequencies will work in the areas you are going to, and that at least two people in the party are familiar with the working of the radio. Every group on the ground must be in radio contact with base. They should be allocated a call sign and frequency, and a schedule of calls to be made.

Discourage groups from talking to each other without going through base. This will cause great confusion if not controlled. Listen out before transmitting otherwise you will interfere with other stations. Everyone has verbal diarrhoea when they talk on the radio so write down what you want to say before making contact and have pencil and paper ready to make notes and take instructions. This will help to keep transmissions to a minimum and preserve the batteries.

REMEMBER: RSVP

Rhythm - don't talk like a darlek
Speed - talk slowly
Volume - speak softly
Pitch - pitch voice higher than normal and use the phonetic alphabet when spelling out place names

In the evening give a situation report to base with your location, what you have done and your future intentions. In the morning receive an update on weather conditions, a time check and other information that base can give you. A noon time call can be used to confirm your position.

BE AWARE

Signals will be weak in steep gullies and valley bottoms and good signals will be received on top of high ground or across water.

If you are tackling a dangerous aspect of the expedition you may want to arrange that base listen out for additional calls so that in an emergency you can call for help and get a response immediately.

BE AWARE

An emergency plan should always be put into operation when two consecutive calls are missed. Even if all is well, if you have not been able to make contact this will be treated by base as an emergency. You must return to or stay at the last reported location and await contact. If you are really in trouble base will know where you last were and where you planned to go to, and the rescue mission can follow.

Mobiles

The mobile phone is one of the great inventions of the twentieth century. In an emergency situation it can be a real life-saver. On expeditions where the radios have failed due to bad weather or the location of the victims, a mobile phone has been used to raise the alarm. A group on Everest got into trouble as they started their descent after summitting. They tried many times to raise base camp but without success. The leader phoned his wife in Hong Kong on a mobile phone and reported their situation. She then alerted Kathmandu, who in turn alerted base camp, Everest and effected a rescue.

Some phones are better than others so it's worth doing some homework; it's also essential to check the network coverage with the service provider before going abroad. Keep one in the car, they are priceless when help is required and a cigarette lighter is a convenient charger for the battery, providing you have an adaptor. Charging can be a problem in the wild so use your phone wisely, but small, hand-held manual chargers can now be bought to recharge batteries. With radios and phones it takes less power to listen than to transmit, so make your call and listen for a reply. If nothing is heard don't despair. With all electrical kit water/moisture is the enemy. The transmitting side may be working but not the receiving side. Make short calls on the hour. Someone may be picking up your signal so don't give up. Once you receive confirmation that the rescue is under way keep the radio/phone on listening watch.

Altimeters

In mountainous areas an altimeter is a good idea. Relating the height recorded can help you determine what contour you are on, and how far it is to the ridge or summit.

You never have enough kit in an emergency. It's nice to have G.P.S, phones etc but you can still manage without as long as you have the ability to improvise and adapt. Learn the basics and use technology for confirmation, rather than depending on them whole heartedly. Communication is of the highest importance and must be given priority. It is a safer place as long as you can communicate with the outside world.

Many survival sagas begin because of bad nagivation, with people getting lost. Always plan for the worst eventuality.

BE AWARE

When things go wrong it's a series of events that compound the situation. The weather deteriorates, the radio is broken, the mobile phone is lost. Two people have multiple injuries and you are out of water. Never give in. Plan for these situations and you will come through, but always have a contingency plan. Imagine the worst possible scenario, and train for it.

Vehicles

Motor vehicles need special adjustment and adaptation to deal with high altitudes and extreme conditions, as well as a thorough overhaul to make sure they are in tip-top working order. You will need tanks for extra fuel and water as well as spares and modifications (see *Vehicles* in *Climate and Terrain*).

Boats and planes

Whether travelling privately or on a commercial service you must take note of the emergency procedures. Maritime and aviation authorities rule that passengers must be informed of them and remembering them could save your life.

When you board an aircraft cabin staff point out the emergency exits and advise you on action to take in the event of an emergency. On board ship you will carry out lifeboat drills and be instructed on how to abandon ship if you have to.

The safest place on an aircraft is as far back in the tail as possible. In a crash this frequently breaks off and most survivors are from this portion. If you are a passenger in a light aircraft ask the pilot about the trip: how long will it take and what sort of ground will you be flying over? Attend to details – they count in an emergency. Also, try to keep your kit with you at all times.

THE UNEXPECTED

How can you prepare for what you do not expect? Preparing for expected difficulties and dangers is difficult enough, but what chance have you of equipping yourself for the totally unknown disaster? Yet these are the disasters that immediately spring to people's minds – the shipwreck and the plane crash or forced landing in unfamiliar and difficult terrain.

This is the reason for this book's existence. Even more important, however, is to know about a whole range of skills which can be applied and adapted to all kinds of situations and to develop a way of thinking that enables you to draw upon them to find the solutions to particular problems. This is the preparation you can make for the unexpected.

But this is not all. You can equip yourself with a few small items which will increase your chances many times over by helping you with some of the basic necessities of survival. This can tip the balance between failure and success. They will fit in a small container slipped into a pocket or bag and can be carried anywhere. They are your survival kit. If there is an emergency you will be glad you always carry it.

More bulky, but still compact enough to carry on a belt whenever you are travelling, are a knife and the items which will fit in your survival pouch (*see Survival pouch*).

Without the basics, which these two kits provide, you can still improvise but they will give you a head start.

SURVIVAL SCENARIO

You're getting your kit together, what is the most important item to remember to take with you?

Take your brain with you. You can't beat the combination of common sense and experience in high stress survival situations.

SURVIVAL KIT

A few key items can make all the difference in the fight for survival. Collect the things listed below. They can all be fitted into a small container, such as a 2oz tobacco tin, that will be hardly noticeable when slipped into an anorak pocket. Make a habit of always having it with you. Do not choose something bigger, you may find it inconvenient to carry and leave it out on the one occasion you actually need it. Many people who roll their own cigarettes carry such a tin.

Experience has proved that each item earns its place, though some are more use in some situations than in others: fish hooks, for instance, may be invaluable in the jungle but less so in the desert.

Polish the inside of the lid to make a mirror-like reflecting surface and seal it, to be waterproof, with a strip of adhesive tape (see (a) over) which can be easily removed and replaced. Don't then just forget the tin. Regularly check the contents, changing any which deteriorate, such as matches and medicine tablets. Mark all drug containers with use and dosage and a run-out date when they should be replaced. Pack spare space in the tin with cotton wool, which will keep the contents from rattling and can be used for fire lighting.

Matches (1)

Waterproof matches are useful but bulkier than ordinary non-safety, strike-anywhere matches, which can be made 'shower-proof' by dipping the heads in melted candle fat. To save space, snap off half of each matchstick.

It is much easier to use matches than to make fire by other methods but don't waste them, use only when improvised methods fail. Take them from the tin one at a time and replace the lid. Never leave the container open or lying on the ground.

Candle (2)

Invaluable for starting a fire as well as a light source. Shave square for packing. If made of tallow it is also fat to eat in an emergency or to use for frying – but be sure it is tallow; paraffin wax and some other candles are inedible. Tallow does not store well, especially in hot climates.

Flint (3)

Flints will work when wet and they will go on striking long after you run out of matches. Invest in a processed flint with a saw striker.

Magnifying glass (4)
Can start a fire from direct sunshine and is useful for searching for splinters and stings.

Needles and thread (5)
Several needles, including at least one with a very large eye that can be threaded with sinew and coarse threads. Choose strong thread and wrap it around the needles. They can be used for repairing or making clothes in an emergency.

Fish hooks and line (6)
A selection of different hooks in a small tin or packet. Add a few split lead weights. Remember that a small hook will catch both large and small fish but a large hook will only catch big ones. Include as much line as possible, it will also be useful for catching birds.

Compass (7)
A luminous button compass – but make sure you know how to read it, as some small compasses can be confusing. A liquid-filled type is best, but check that it does not leak, has no bubbles in it and is fully serviceable. The pointer is prone to rust. Make sure it is on its pivot and swings freely.

Beta light (8)
Beta lights provide a reliable and continuous light source for about 15 years. The lights are generally the size of a small coin, are self-illuminating requiring no batteries, and are ideal for map reading.

Snare wire (9)
Preferably brass wire – 60–90cm (2–3ft) should do. Save for snares, but could solve many survival problems.

Flexible saw (10)
These usually come with large rings at the ends as handles. These take up too much room, so remove them; they can be replaced by wooden toggles when you need to use it. To protect from rust and breakage cover it in a film of grease. Flexible saws can be used to cut even quite large trees.

Medical kit (11)
What you include depends upon your own skill in using it. Pack medicines in airtight containers with cotton wool to prevent rattling. The following items will cover most ailments:

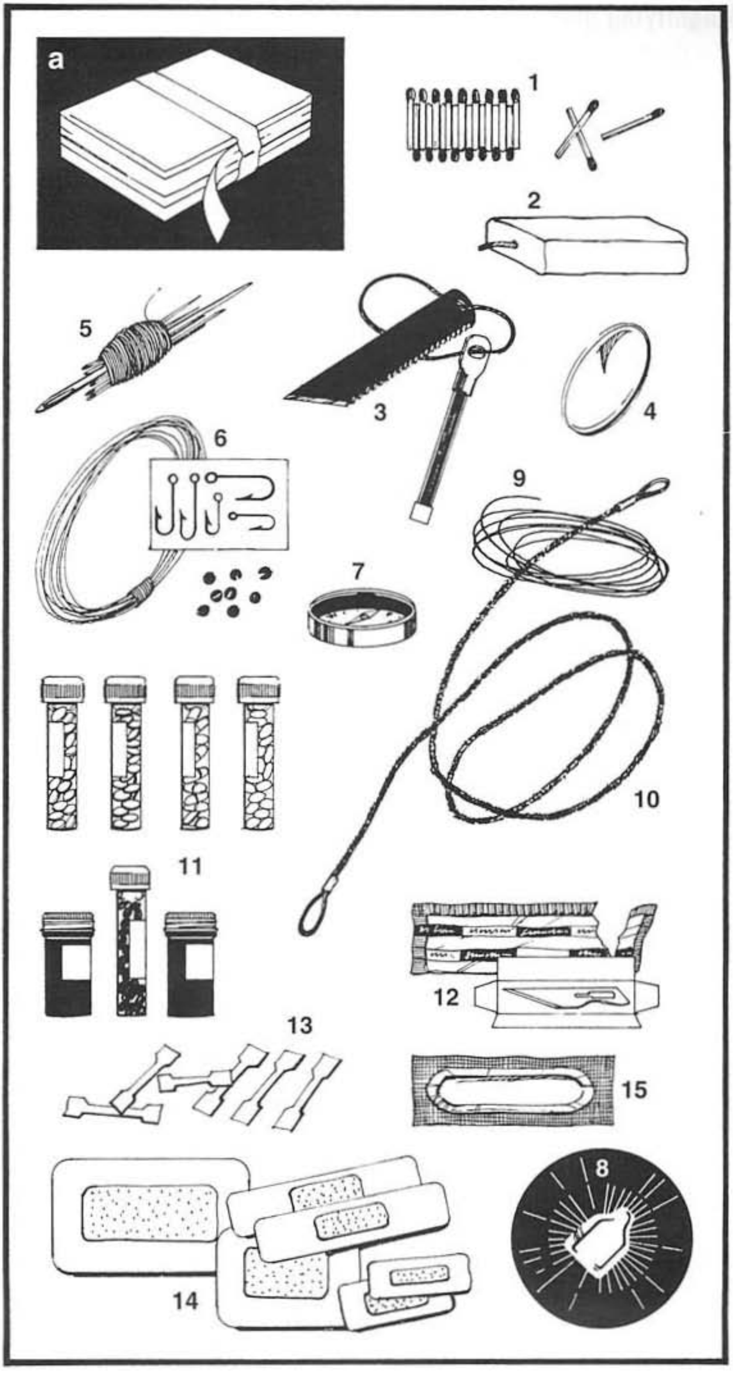
a
1
2
5
3
4
6
9
7
10
11
12
13
15
14
8

Analgesic: A pain reliever for mild and moderate pain. Codeine phosphate is ideal for tooth-, ear- and headaches. DOSE: one tablet every six hours as needed but can cause constipation as a side-effect so will help in cases of loose bowels. Not to be taken by children, asthmatics or people with liver disorders.

Intestinal sedative: For treating acute and chronic diarrhoea. Immodium is usually favoured. DOSE: two capsules initially, then one each time a loose stool is passed.

Antibiotic: For general infections. Tetracycline can be used even by people hypersensitive to penicillin. DOSE: one 250mg tablet, four times daily, repeated for five to seven days. Carry enough for a full course. If taking, avoid milk, calcium and iron preparations or other drugs containing aluminium hydroxide.

Antihistamine: For allergies, insect bites and stings (may also help in cases of a bad reaction to a drug). Piriton is recommended in Britain, Benadryl in the USA. Sleepiness is a side-effect of Piriton, so useful as a mild sleeping pill. Do not exceed recommended dosages or take with alcohol.

Water sterilizing tablets: For use where water is suspect and you cannot boil. Follow manufacturers' instructions.

Anti-malaria tablets: Essential in areas where malaria is present. There are types which require only one tablet taken monthly.

Potassium permanganate: Add to water and mix until water becomes bright pink to sterilize it, deeper pink to make an antiseptic and to a full red to treat fungal diseases such as athlete's foot.

Surgical blades (12)

At least two scalpel blades of different sizes. A handle can be made from wood when required.

Butterfly sutures (13)

Use to hold the edges of wounds together.

Plasters (14)

Assorted sizes, preferably waterproof, for minor abrasions and keeping cuts clean. They can be cut and used as butterfly sutures (see *Stitching wounds* in *Health*).

Condom (15)

This makes a good water-bag – holding 1 litre (1¾pt). It is easiest to fill it from a source that has an abundance of water and a degree of water pressure, like a waterfall.

SURVIVAL POUCH

In a car, boat or aircraft don't stow all your kit separately. Pack a survival pouch, too large to carry in your pocket like your survival tin but kept where it can be grabbed quickly in an emergency. If you are on foot keep it outside your back-pack, and carry it on your belt. It should contain fuel, food, survival bag and signalling kit, all packed into a mess tin (a) which protects the kit and doubles as a cooking utensil. If you fancy a brew or a snack, it is all there for you and in an emergency it gives you a first back-up for survival. Anything you use from the pouch must be replenished as soon as possible.

POUCH

The pouch must be made from waterproof material and be large enough to take a mess tin. It must have a positive fastening that will not come undone, and a strong tunnel loop to hold it on your belt. Remember the pouch contains matches, solid fuel and flares – all life savers, but to be treated with care.

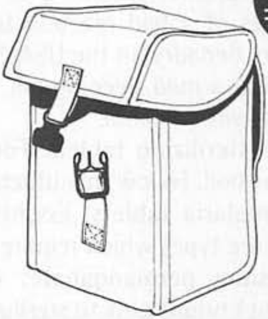

Mess tin

This is made from aluminium, which is light and strong. A good cooking utensil, it protects kit packed inside it.

Fuel

Preferably you should have solid hexamine fuel tablets in their own stove container (1). Use sparingly when a wood fire is inconvenient. They make excellent fire lighters. The stove simply unfolds to form an adjustable pot stand (2) and holder for burning fuel.

Torch

Pack a small pencil-like torch (3) that takes up little room. Keep batteries inside it, but reverse the last so that, if accidentally switched on, the batteries don't run down. Many modern torches use light-emitting diodes (LED) as their light source rather than conventional bulbs. They use less battery power and will typically work continuously for two weeks.

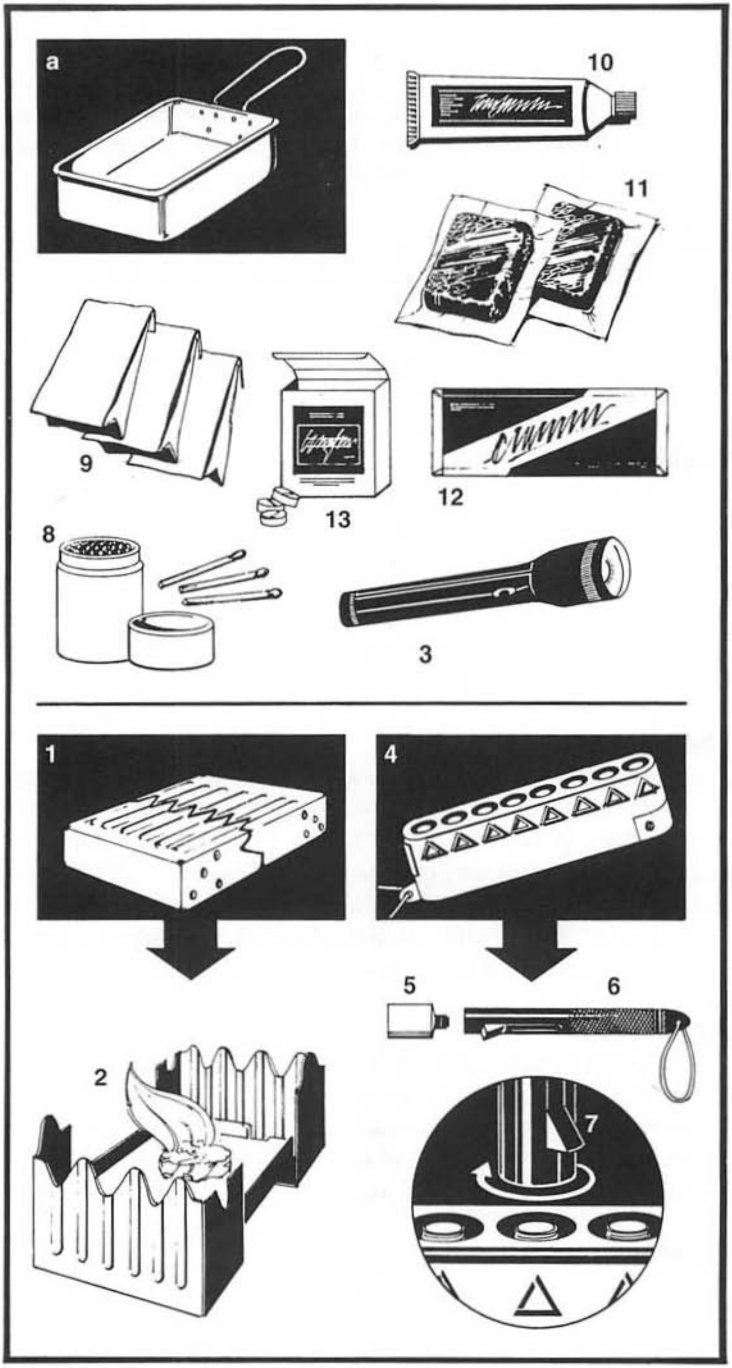
a
10
11
9
13
12
8
3
1
4
5
6
2
7

Flares

Signal flares (4) to attract attention, especially in close country. Carry red and green miniflares (5) and a discharger (6) (no bigger than a fountain pen). These are explosive so pack carefully. Simply remove discharger and screw on to flare (7). Withdraw flare and point skywards at arm's length. Pull the trigger to fire.

Marker panel

A strip or bar of fluorescent material about 0.3 x 2m (1 x 6ft) used to attract attention in an emergency (see *Signaling* in *Rescue*). One bar signals immediate evacuation. Form other signals with panels carried by others in your party. Pack to stop the other items in the pouch rattling. A silver thermal blanket is ideal for this.

Matches

Pack as many matches (8) as possible in a waterproof container, you never have enough. Movement against each other can ignite non-safety matches – pack carefully.

Brew kit

There is nothing like a brew-up to restore morale. Pack tea powder and sachets of milk and sugar (9). Tea quenches thirst - coffee aggravates it!

Food

Fat is the hardest food to come by when living off the land. Its extra calories earn it a place in your kit – tubes of butter, lard or ghee (10) are available. Dehydrated meat blocks (11) are nourishing and sustaining, though not very good in flavour. Chocolate (12) is a good food, but does not keep well – check regularly. Salt (13) MUST be included – or, better still, an electrolite powder which contains vitamins, salt and other minerals that the body requires.

Survival bag

A large polythene bag about 200 x 60cm (7 x 2ft) is a lifesaver in the cold. In an emergency get inside to reduce heat loss. Although wet from condensation you will be warm. Even better is a heat-insulated bag of reflective material that keeps you warm and solves the condensation problem.

Survival log

Keep a written log of all events. Do not trust your memory. Record discoveries of resources, and of what works and what does not. It becomes a valuable reference and making it helps keep up your morale.

KNIVES

A knife is an invaluable asset in a survival situation. The serious adventurer will carry one always. They should be packed with the luggage that you are checking in when travelling by commercial aircraft, otherwise they will be confiscated. Current knife law in the UK states that knives where the blade folds into the handle are legal as long as the blade is shorter than 3in (7.62cm).

Choosing a knife

A multi-bladed folding knife is a useful tool, but, if you carry only one knife, you need something stronger, a general-purpose blade that will do all likely tasks efficiently and comfortably, from cutting trees to skinning animals and preparing vegetables. Some have a compass built into the handle or have the handle hollowed out so that you can carry survival kit inside it. However, these features will be offset by the possibility of a hollow handle breaking and a compass may soon lose its accuracy after the knife has been used on a hardwood tree. If you lose this kind of knife you also lose your survival kit – much better to keep the kit in a separate pouch on your belt or on the sheath.

REMEMBER: YOU ARE ONLY AS SHARP AS YOUR KNIFE. Your knife is such an important piece of survival equipment, that you must keep it sharp and ready for use. Don't misuse your knife. Never throw it into trees or onto the ground. Keep it clean and, if you don't intend to use it for a while, oiled and in its sheath.

When walking through close, or difficult terrain, get in the habit of checking your knife. This should become an automatic reflex especially after negotiating difficult terrain. A check of all pockets and possessions should be second nature.

Folding knives

A folding knife that has a serrated edge, and which can be opened one-handed is valuable, provided it has a good locked position. Always carry one. A blade in a wooden handle is usually more comfortable: it will not slip in a sweaty hand and, if the handle is made from a single piece of wood, is less likely to cause blisters.

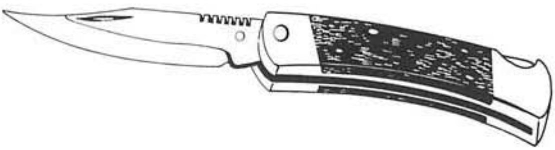

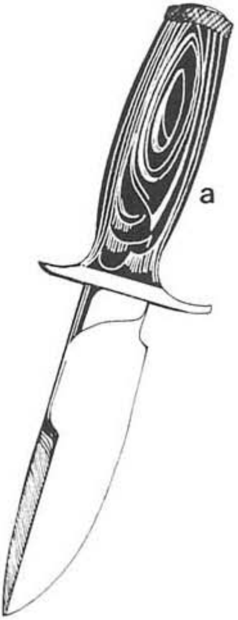

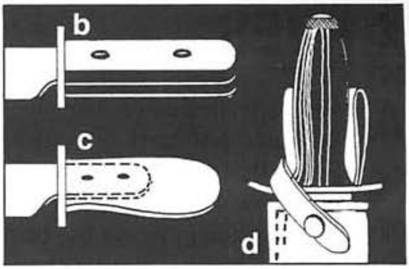

Handle (a) is ideal: a single rounded piece of wood, the knife tang passing through it and fastened at the end. If the handle breaks the tang can be wrapped with cloth or twine. Handle (b) is only riveted to the tang and would cause blisters. Handle (c) could break at the rivets if subjected to heavy work and the short tang would make it difficult to improvise a handle. The sheath (d) should have a positive fastening and a tunnel belt loop.

Parang

This is the Malayan name for a type of knife with a large curved blade like a machete. It is too large to be carried in normal daily life (see *Knives* on page 35) but ideal when going out into the wild.

A parang 30cm (12in) in overall blade length and weighing no more than 750g (1½lb) is best, the blade 5cm (2in) at its widest and end-bolted into a wooden handle. The curved blade enables maximum effort to be applied when cutting timber and the blade arrives before the knuckles, so giving them protection. Even large trees can be cut down with a parang, which is especially useful for building shelters and rafts.

The parang blade has three different edges: **b** *does the heavy work of chopping wood and bone,* **a** *is finer and used for skinning,* **c** *is finer still for cawing and delicate work.* **a** *and* **c** *are easily maintained and* **b** *is sharp but not so sharp as to chip easily.*

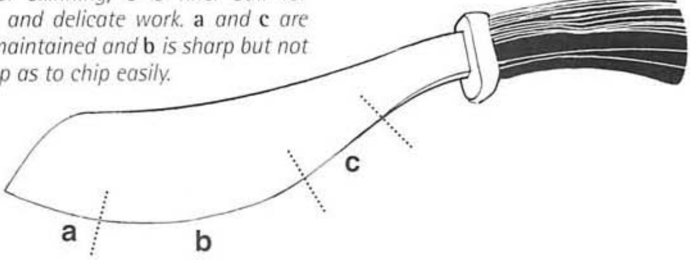

The sheath must have positive fastenings to keep the parang secure, and a loop for fixing to a belt. Some sheaths have a pocket on the front for a sharpening stone.

REMEMBER: *There is a danger that the cutting edge may come through the side. To draw the parang NEVER hold the sheath on the same side as the cutting edge (a). This is dangerous. Get into the habit of gripping the side AWAY FROM THE CUTTING EDGE.*

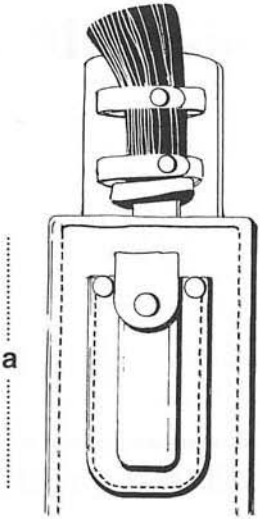

Sharpening a knife

Any sandstone will sharpen tools – a grey, clayey sandstone being best. Quartz, though more rarely found, is good and granite can also be used. Rub two pieces together to make them smooth. A double-faced stone with a rough and a smooth surface is ideal and should be carried in the sheath pocket. Use the rough surface first to remove burrs, then the smooth one to get a fine edge. The object is to get an edge that will last and not chip.

To sharpen the blade, hold the handle in the right hand. Use a clockwise circular motion and apply a steady pressure on the blade with the fingertips of the left hand as you push away. Keep the angle constant and make sure you push the knife into the stone so that they edge is sharpened rather than rounded. Keep the stone wet. Rock particles on the blade will show the angle you are obtaining. DON'T drag the blade towards you under pressure. This will produce burrs. Reduce the pressure for a finer edge. Work counter clockwise on the other side.

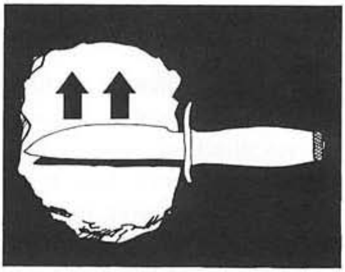

Blade profile: (a) is too steep and will soon wear, (b) is good and (c) is too fine and might chip.

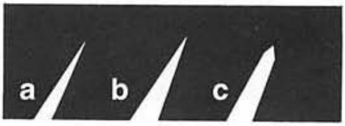

FACING DISASTER

When facing a disaster it is easy to let yourself go, to collapse and be consumed in self-pity. But it is no use giving up or burying your head in the sand and hoping that this is a bad dream that will soon pass. It won't, and with that kind of attitude it will rapidly become much worse. Only positive action can save you.

A healthy, well-nourished person can physically tolerate a great deal, provided that he or she has self-confidence. Even if sick or injured, a determined person can win through and recover from seemingly impossible situations. To do so there are many stresses that must be overcome.

SURVIVAL STRESSES

The survival situation will put you under pressure, both physical and mental. You will have to overcome some or all of the following stresses:

- Fear and anxiety
- Pain, illness and injury
- Cold and/or heat
- Thirst, hunger and fatigue
- Sleep deprivation
- Boredom
- Loneliness and isolation

Can you cope? You have to.

Self-confidence is a product of good training and sound knowledge. These must be acquired before you have to face up to a survival situation.

Physical fitness plays an important part. The fitter you are the better you will survive. Initially you may have to go without sleep to ensure that you are in a safe location, or make a long march in dangerous conditions. Do not wait until you are forced to go without sleep to see whether you are capable of doing so. Prove it to yourself now by getting into training. Develop the resources to cope with fatigue and loss of sleep.

You will be working hard to procure food and water. They will relieve hunger and thirst. But finding them will tire you and you will need an adequate shelter to enable you to rest and recover from your efforts. Don't overdo it. Rest frequently and assess the situation.

Pain and fever are warning signals that call attention to an injury or physical condition. They are not in themselves dangerous, however distressing and discomforting. Pain can be controlled and overcome. Its biological function is to protect an injured part, to prevent you using it, but this warning may have to be ignored to avoid the risk of further injury or death.

SURVIVAL SCENARIO

How long can the body cope without the essentials?

In general the human body can survive for 3 minutes without air: 3 days without water: 3 weeks without food. There are always exceptions to this and there are examples of people pushing these boundaries and surviving for longer periods. It is amazing what the human body can endure but such a survivor's health can suffer in the long term due to such trauma. For example a person surviving for more than 9 days without water will undoubtedly suffer kidney damage or failure.

BASIC NEEDS

To reiterate, the main elements required for survival are FOOD, FIRE, SHELTER and WATER. Use PLAN (Protection, Location, Acquisition, Navigation) to help you to prioritise your survival needs. If you don't have shelter in the desert, having 2 litres (3½pts) of water is of little use to you in the longer term.

It takes a healthy person quite a long time to die of starvation, for the body can use up its stored resources, but exposure to wind, rain and cold can be fatal even in temperate climates and death comes in only minutes in the icy waters of the poles. Food is rarely the first priority. Even in those places where it is difficult to find there are usually other problems to face first. Shelter will often be the prime necessity in extremes of climate or temperature – not just in the frozen polar regions or the baking deserts, but for walkers trapped by mist on a hillside. The need for fire is closely linked.

Water is something that most people in the modern world take for granted. They are so used to turning on a tap that until an extreme drought causes water rationing they scarcely think about it. Yet the survivor at sea, or after a flood, though surrounded by water, may be desperate for drinkable water – and there are many places where, unless it rains, no obvious water is available. The other survival necessities are dealt with later in the book, but water is universally important.

WATER

Water is essential to life. A person can survive for three weeks without food but for only three days without water, therefore its discovery and conservation should be prioritised over food. Don't wait until you have run out of water before you look for it. Conserve what you have and seek a source as soon as possible, preferably fresh running water, though all water can be sterilized by boiling or by using chemical purifiers. In a survival situation a 1 litre-jug (1¾pt) can be made to last 4 days, but if necessary the last ¼ litre (½pt) can be made to last 3 days. This is achieved by dividing the last ¼ litre into three, then drinking half of the day's ration at midday, and the balance at night, for 3 days.

Water is the coolant that keeps the body at an even temperature, it is needed to keep the kidneys functioning to eliminate wastes, is required for breathing, and for digestion. But the fluids contained in the body are limited. Lost water must be replaced or health and efficiency will suffer. The average human requires the minimum of ¼ litre (½pt) of water per day to survive.

Water loss

The average person loses 2–3 litres (3½–5¼pt) of water each day – even someone resting in shade loses about 1 litre (1¾pt). Just breathing loses fluids, and loss through respiration and perspiration increases with work rate and temperature. Vomitting and diarrhoea increase loss further. This must all be replaced to preserve the critical water balance, either by actual water or water contained in food.

HOW TO RETAIN FLUIDS

- Avoid exertion. Just rest.
- Don't smoke.
- Keep cool. Stay in shade. If there is none erect a cover to provide it.
- Do not lie on hot ground or heated surfaces.
- Don't eat, or eat as little as possible. If there is no water available fluid will be taken from the vital organs to digest food, further increasing dehydration. Fat is hardest to digest and takes a lot of fluid to break it down.
- Never drink alcohol.
- Don't talk – and breathe through the nose, not the mouth.

FINDING WATER

The first place to look is in valley bottoms where water naturally drains. If there is no obvious stream or pool, look for patches of green vegetation and try digging there. There may be water just below the surface which will build up in the hole. Even digging in gullies and dry stream beds may reveal a spring beneath the surface, especially in gravelly areas. In mountains look for water trapped in crevices.

On the coast digging above the high water line, especially where there are sanddunes, has a good chance of producing about 5cm (2in) of fresh water that filters down and floats on the heavier salt water. It may be brackish but is still drinkable. Where cliffs fall into the sea look for lush growth of vegetation, even ferns and mosses, in a fault in the rock formation and you may find a soak or spring.

If no freshwater can be found, saltwater can be distilled (see *Solar still* and *Distillation*, this section).

WARNING

Be suspicious of any pool with no green vegetation growing around it, or animal bones present. It is likely to be polluted by chemicals in the ground close to the surface. Check edge for minerals which might indicate alkaline conditions. ALWAYS BOIL WATER FROM POOLS. In deserts there are lakes with no outlets; these become salt lakes. Their water MUST be distilled before drinking.

Dew and rain collection

Despite the acid rain produced by industrialized countries, which can cause a build-up of pollution in the soil, rainwater everywhere is drinkable and only needs collecting. Use as big a catchment area as possible, running the water off into containers of every kind. A hole dug in the ground and lined with clay will hold water efficiently, but keep it covered. If you have no impermeable sheeting, metal sheets or bark can be used to catch water in. If you have any doubt about the water you have collected, boil it.

In climates where it is very hot during the day and cold at night, heavy dew can be expected. When it condenses on metal objects it can be sponged or licked off.

You can use clothing to soak up water and then wring it out. One way is to tie clean cloths around the legs and ankles and walk through wet vegetation. These can be sucked or wrung out.

ANIMALS AS SIGNS OF WATER

Mammals

Most animals require water regularly. Grazing animals are usually never far from water – though some kinds travel thousands of miles to avoid the dry season – as they need to drink at dawn and dusk. Converging game trails often lead to water; follow them downhill. Carnivores (meat eaters) can go for a long period between waterings. They get moisture from the animals on which they prey so are not a positive indication of local water.

Birds

Grain eaters, such as finches and pigeons, are never far from water. They drink at dawn and dusk. When they fly straight and low they are heading for water. When returning from water they are loaded with it and fly from tree to tree, resting frequently. Plot their direction and water can be found.

Water birds can travel long distances without stopping to feed or drink so do not necessarily indicate water nearby. Hawks, eagles and other birds of prey also get liquids from their victims so cannot be taken as a sign of local water.

Reptiles

Not an indicator of water. They collect dew and get moisture from prey, so can go a long time without.

Insects

Good indicators, especially bees: they fly at most 6.5km (4 miles) from their nests or hives, but have no regular watering times. Ants are dependent upon water. A column of ants marching up a tree is going to a small reservoir of trapped water. Such reservoirs are found even in arid areas. Most flies keep within 90m (100yd) of water, especially the European Mason Fly with its iridescent green body.

Human tracks

Will usually lead to a well, bore hole or soak. It may be covered over with scrub or rocks to reduce evaporation. Replace the cover.

REMEMBER: RATION YOUR SWEAT NOT YOUR WATER!
If you have to ration water, take it in sips. After going a long time without water, don't guzzle when you do find it. Take only sips at first. Large gulps will make a dehydrated person vomit, losing even more of the valuable liquid.

CONDENSATION

Tree and plant roots draw moisture from the ground, but a tree may take it from a water table 15m (50ft) or more below, too deep to dig down to reach. Don't try; let the tree pump it up for you by tying a plastic bag around a leafy branch. Evaporation from the leaves will produce condensation in the bag.

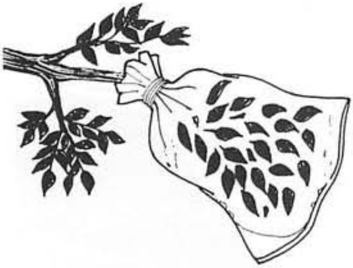

Choose healthy vegetation and bushy branches. On trees keep the mouth of the bag at the top with a corner hanging low to collect condensed evaporation.

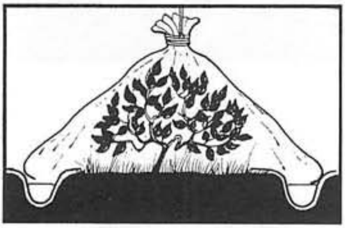

Placing a polythene tent over any vegetation will collect moisture by evaporation which will condense on the plastic as it cools. Suspend the tent from the apex or support with a padded stick. Avoid foliage touching the sides of the trap or it will divert water droplets which should collect in plastic-lined channels at the bottom.

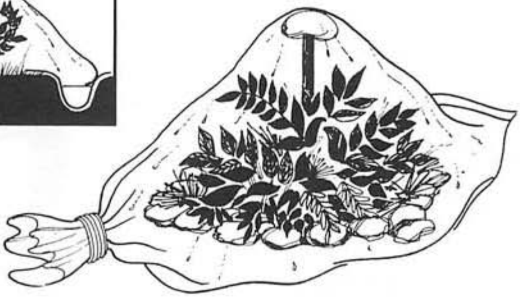

Even cut vegetation will produce some condensation as it warms up when placed in a large plastic bag. Keep the foliage off the bottom with stones so that water collects below it, and keep the foliage from touching the plastic. Use stones to keep the bag taut. Support the top on a padded stick. Arrange the bag on a slight slope to encourage condensation to run down to the collecting point. When no longer productive carefully replace with fresh foliage.

Solar still

Dig a hole in the ground approximately 90cm (36in) across and 45cm (18in) deep. Place a collecting can in the centre, then cover the hole with a sheet of plastic formed into a cone. The sun's heat raises the temperature of the air and soil below and vapour is produced. As the air becomes saturated, water condenses on the underside of the plastic, running down into the container. This is especially effective in desert regions and elsewhere when it is hot during the day and cold at night. The plastic cools more quickly than the air, causing heavy condensation. This kind of still should collect at least 570ml (1pt) over a 24-hour period.

The still may also double as a trap. Insects and small snakes are attracted by the plastic. They may slide down into the cone or wriggle underneath it and drop into the hole and then cannot climb out.

A solar still can be used to distill pure water from poisonous or contaminated liquids.

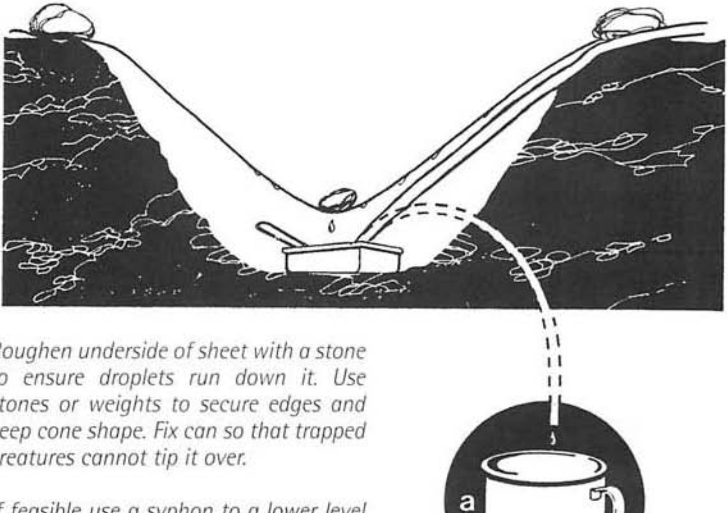

Roughen underside of sheet with a stone to ensure droplets run down it. Use stones or weights to secure edges and keep cone shape. Fix can so that trapped creatures cannot tip it over.

If feasible use a syphon to a lower level (a) to draw off water without disturbing the still.

WARNING

URINE AND SEA WATER

Never drink either – Never! But both can produce drinking water if distilled – and sea water will provide you with a residue of salt.

Distillation

Distillation kits are part of the equipment of life-rafts, but they can be improvised. To distill liquid you need to make something to do the job of a laboratory retort. Pass a tube into the top of a water-filled covered container, placed over a fire, and the other end into a sealed collecting tin which, preferably, is set inside another container providing a jacket of cold water to cool the vapour as it passes out of the tube. You can improvise the equipment from any tubing – pack frames, for instance. To avoid wasting water vapour, seal around the joins with mud or wet sand.

An easier method is a variation on the desert still. It takes a little longer for the water to condense but may be easier to set up.

Take a tube from a covered vessel in which polluted/saltwater, or even urine, is to boil. Set the other end under a solar still. A sheet of metal or bark, perhaps weighted down, will cover the vessel. Even a cone of leaf over the water pot will help direct the steam into the tube.

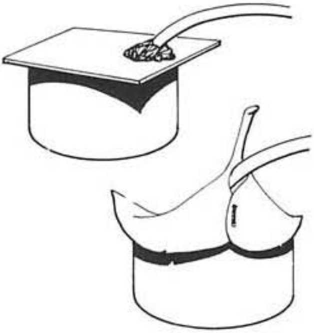

Water from ice and snow

Melt ice rather than snow – it produces a greater volume faster for less heat: twice as much for half the heat. If forced to heat snow, place a little in the pot and melt that first, gradually adding more to it. If you put a lot of snow into the pot, the lower level will melt and then be soaked up into the absorbent snow above it, leaving a hollow beneath which will make the pot burn. Lower layers of snow are more granular than that on the surface and will yield more water.

Water from sea ice

Sea ice is salt (no good for drinking) until it has aged. The more recently frozen, the saltier it will be. New sea ice is rough in contour and milky-white in colour. Old ice is bluish and has rounded edges, caused by weathering.

Good water can be obtained from blue ice – the bluer and smoother the better. But beware of even old ice that has been exposed to salt spray.

WATER FROM PLANTS

Water collectors

Cup-shaped plants and cavities between the leaves of bromeliads (many of which are parasitic on the branches of tropical trees) often collect a reservoir of water.

Bamboo often holds water in its hollow joints. Old and yellow stems are more likely to be water bearing. Shake them – if you can hear water slurping around cut a notch at the bottom of each joint and tip the water out.

Traveller's Tree Ravenala madagascariensis, *one of the banana family, can hold 1–2 litres (2–4pt) of water between the bases of the chevron of leaf stalks.*

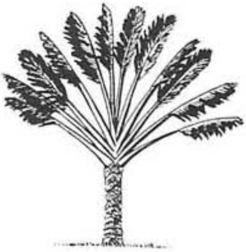

Vines

Vines with rough bark and shoots about 5cm (2in) thick can be a useful source of water. But you must learn by experience which are the water-bearing vines, because not all have drinkable water and some have a poisonous sap. The poisonous ones yield a sticky, milky sap when cut. You will know not to try that type again – otherwise it is a matter of trial and error and worth trying any species.

Some vines cause a skin irritation on contact if you suck them, so it is better to let the liquid drip into your mouth rather than put your mouth to the stem, and preferable to collect it in a container.

To obtain water from a vine select a particular stem and trace it upwards. Reach as high as possible and cut a deep notch in the stem. Cut off the same stem close to the ground and let the water drip from it into your mouth or into a container. When it ceases to drip cut a section from the bottom and go on repeating this until the vine is drained. Do NOT cut the bottom of the vine first as this will cause the liquid to run up the vine through capillary action.

Roots

In Australia the Water Tree, Desert Oak and Bloodwood have their roots near the surface. Pry these roots out from the ground and cut them up into 30cm (12in) lengths. Remove the bark. Suck out the moisture, or shave to a pulp and squeeze over the mouth.

It is not easy to find some of the most useful desert roots unless you have been shown by someone with experience. Australian Aborigines can identify a tiny twig which grows from a football-like bulbous root, which can be a life-saver – but unless you have been shown how to find them it is not worth expending your energy and resources looking.

Palms

The Buri, Coconut and Nipa palms all contain a sugary fluid which is very drinkable. To start it flowing bend a flowering stalk downwards and cut off its tip. If a thin slice is cut off the stalk every 12 hours the flow will be renewed, making it possible to collect up to a quart each day. Nipa palms shoot from the base so that you can work from ground level, on grown trees of other species you may have to climb up them to reach a flowering stalk.

Coconut milk has considerable water content, but from ripe nuts it is a powerful laxative; drinking too much would make you lose more fluid.

Cacti

Both the fruit and bodies of cacti store water, but not all cacti produce liquid safe to drink – the Saquarro, the giant multi-fingered cactus of Arizona, is very poisonous. Take care to avoid contact with cactus spines, they can be very difficult to remove, especially the very fine hair-like ones, and can cause festering sores if they stay in the skin.

The Barrel cactus *Echinocactus grusoni* (see *Desert plants* in *Food*) can reach a height of 120cm (4ft), is found in the southern United States through to South America and requires considerable effort to cut through its tough, spine-covered outer skin. The best method is to cut off the top and chop out pieces from the inside to suck, or to smash the pulp within the plant and scoop out the watery sap, which varies from tasteless in some plants to bitter in others. An average-sized, 100cm (3½ft) Barrel cactus will yield about 1 litre (1¾pt) of milky juice and this is an exception to the rule to avoid milky-sapped plants.

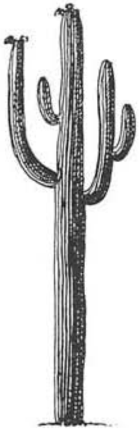

Saquarro cactus Sereus giganteus *of Mexico, Arizona and California, grows to 5m (17ft) high and holds large amounts of fluid – but it is poisonous. Collect and place in a solar still to evaporate and recondense during the cold night.*

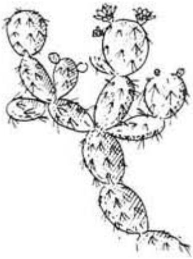

Opuntia *cacti – Prickly pears, or Figilinda, have big, ear-like excrescences and produce oval fruits which ripen to red or gold. Their large spines are easier to avoid than those of many cacti. Both fruit and 'ears' are moisture laden.*

WATER FROM ANIMALS

Animal eyes contain water which can be extracted by sucking them.

All fish contain a drinkable fluid. Large fish, in particular, having a reservoir of fresh water along the spine. Tap it by gutting the fish and, keeping the fish flat, remove the backbone, being careful not to spill the liquid, and then drink it.

If you need water that badly you should be careful not to suck up the other fish juices in the flesh, for they are rich in protein and will use up water in digestion.

Desert animals can also be a source of moisture. In times of drought in Northwestern Australia, Aborigines dig in dry clay pans for the desert frogs that burrow in the clay to keep cool and survive. They store water in their bodies and it can be squeezed out of them.

SALT

Salt is essential for water retention. A normal diet includes a daily intake of 10g (⅓oz). The trouble starts when you start to get rid of it faster than you eat it. The body loses salt in sweat and urine, so the warmer the climate the greater the loss. Physical exertion will also increase the loss. However, replenishing the salt levels in your body is not always recommended, and much will depend on the situation you find yourself in. If you are on reduced food and or water rations, salt is the first thing that should be cut from your diet, because it increases dehydration.

Recognize the signs of dehydration

The first symptoms of salt deficiency are muscle cramps, dizziness, nausea and tiredness. The remedy is to take a pinch of salt in half a litre (¾ pint) of water. Salt supplements in tablet form used to be the prescribed method of increasing sodium levels but always dissolve these in water, or isotonic drinks – if they are available.

What happens if you do not carry salt or your supplies run out? By the coast or at sea there is plenty of saltwater available – a pint of sea water contains about 15g (½oz) of salt, but do NOT just drink it as it is. Dilute it with plenty of fresh water. Evaporating sea water will leave you with salt crystals.

Inland salt supplies are more problematic. In farming areas you will find salt licks for cattle – but you will then be close to civilization and not likely to have reached the stage of salt deprivation. However, all animals need salt and observation of them may reveal a natural source. In one part of Africa elephants risk the dangerous depths of a dark cave to lick salt from its sides.

Salt can be obtained from some plants. In North America the best source is the roots of hickory trees, and in Southeast Asia those of the Nipa Palm can be used in the same way. Boil the roots until all the water is evaporated and black salt crystals are left.

If no direct salt sources are available to you then you will have to rely on getting it second-hand, through animal blood, which should never be wasted as it is a valuable source of minerals.

2

STRATEGY

Whether a small-scale accident or a mammoth disaster – both may bring about a life-or-death situation – the same disciplines and quick thinking are needed.

To show the way in which basic survival strategies are applied to every kind of situation, a series of individual small-scale vehicle incident procedures are outlined.

The same approaches can be seen on a far wider scale in the handling of a major air crash. In these circumstances the survivor is more likely to find him- or herself in unfamiliar territory and involved with a greater number of people for a longer period.

STRATEGY 52
CAR ACCIDENT SURVIVAL 53
- Brake failure 53
- Car under water 54
- Car on railway tracks 55

DISASTER IN THE AIR 55
- Protection 57
- Location 58
- Acquiring food and water 59
- Navigation 60
- People 61

STRATEGY

Good planning and preparation enable the survivor to confront difficulties and dangers that pose a serious threat to survival. They become contingencies for which you are equipped. But you cannot anticipate everything. You must be ready to respond rapidly to the unexpected danger and to deal with potential disaster rationally and realistically. You must overcome the tendency to panic and take the action appropriate to the situation.

Sometimes a collision or other accident occurs with no warning of any kind, but in most instances there is a moment of realization that something is going to happen and it is in that moment that instinctive reaction can save lives. In many situations there is a considerable time in which an awareness of potential disaster can develop and that is when the panic reaction is probably most dangerous.

As mist closes in on a hillside, reducing visibility to almost nothing and making it easy to lose any sense of direction, most people would begin to panic at the thought that they are going to be trapped. They begin to do foolish things and increase their danger. They should, however, already be assessing the possibilities and looking for some suitable shelter in which to wait until conditions become safe to continue. Keeping calm, in the knowledge that you have the ability to handle the situation, will not only enable you to see it through but also to see other solutions.

Some situations are predictable and knowledge of the techniques for handling them will minimize the risks. Learn them, they may save your life. They may take considerable nerve – like waiting for the right moment to escape from a car that is sinking under water – but they are based on experience. The answer to more general survival problems, however, will often lie in inspired improvisation drawing on those skills appropriate to the situation.

Disaster may involve you in a contained situation which you must handle alone – or you may find yourself one of hundreds of people in a large-scale disaster over which there can be no control at all.

SURVIVAL SCENARIO

If the weather or conditions in your camp change, should you revise your survival strategy?

You should have a strategy for every scenario, therefore as situations change, so should your strategy. You must be flexible and be prepared to adapt.

CAR ACCIDENT SURVIVAL

BRAKE FAILURE

If brakes fail while driving, change gear and apply the handbrake. You must do several things at once: take your foot off the accelerator, flick the switch of your warning lights, pump the footbrake rapidly (it may still connect), change down through the gears and apply handbrake pressure. Don't slam the brake on, begin with gentle bursts, gradually braking harder until you stop.

If there is no time for all this, take your foot off the accelerator and change down through the gears – and grab the handbrake – but DON'T apply maximum pressure until you are sure that you won't skid.

Look out for escape lanes and places where you can leave the road, preferably a soft bank or a turning that has an uphill slope.

If speed remains unchecked, on a steep hill for example, brush the car along a hedge or wall to reduce speed. Take advantage of a vehicle in front and use it to stop you – run into it as gently as the situation allows. Use warning lights, blow your horn and flash your headlights to give the driver in front as much warning as possible that you are on a collision course.

Collision

If collision seems inevitable, stay with it and steer the car to do as little damage to others and yourself as possible. Try to avoid a sudden stop by driving into something which will give, like a fence or a hedge rather than a tree.

Seatbelts (compulsory in most countries) will help stop you hitting the windscreen. Most new cars are usually fitted with airbags for both the driver and the front passenger. Given the speed and force with which an airbag inflates, it is vitally important that you do not sit too close to the steering wheel. And if your steering wheel is adjustable, tilt it downwards so that the airbag points towards your chest rather than your head, but make sure you can see the instruments clearly.

If neither of these options are available, adopting the brace position will offer you the best chance of survival in a crash because it stops you from impacting too heavily on the interior of the car. The important thing is to get your upper torso as low as possible. There are two alternative types of brace position depending on the space that you are in. The first position requires you to place your head on you knees, while your hands hold on to your lower legs. The second position is more useful if you are driving the car. Place your hands on the steering wheel, then, place your forehead onto the back of your hands before the impact.

Jumping out

Do NOT try to jump out of a runaway car unless you know it to be headed for a cliff or another substantial drop and that you will not survive the impact. If this is the only available option, then open the door, undo the safety belt, and begin to roll yourself into a ball. Drop from the car in a rolling movement. Do not resist the ground but keep balled up and continue the roll.

CAR UNDER WATER

If possible abandon the car before it sinks, for it will not sink immediately and will take time to fill. Water pressure on the outside makes it very difficult to open the door so open the window if you can and wriggle out of it. It takes great presence of mind to manage that when subject to the shock and surprise of the 'splash down', but if there are small children in the car it may be possible to push one through. Do not try to save possessions.

If you have not been fast enough CLOSE the window firmly, get children to stand and lift babies near to the roof. Release seat belts and tell everyone by a door to be ready with a hand on the handle. Release at once any automatic door locks or master locks. Water could prevent them from working. Do not attempt to open doors at this stage.

As water fills the interior, air will be trapped near the roof. The water pressure inside the car will nearly equalize the pressure with that of the water outside the car. As the car comes to rest and is nearly full of water tell everyone to take a deep breath, open the doors and swim to the surface, breathing out as they do so. Everyone leaving through the same door should link arms. If you have to wait for someone to get out before you, hold your breath for that moment.

PRECAUTION: Always park alongside water, not running towards it. If you have to park a car facing water then leave it in reverse gear and with the handbrake on (if facing away from water, in first gear with the handbrake on).

CAR ON RAILWAY TRACKS

If a car breaks down on an unmanned level crossing, put it into gear and use the starter motor to jerk it clear. This will work with a manual gear change but not with an automatic. If a train is approaching abandon the car, carry children or infirm persons to safety and keep away – about 45m (50yds) should be far enough – for if a train is travelling at high speed it could fling car wreckage quite a distance.

If there is no train visible, or you can see one several miles in the distance, you must try to avert the collision. If the car can be moved by pushing, push it clear of all tracks – you cannot be sure which one the train will be on. If there is an emergency telephone warn signalmen further down the track of the situation. If not, walk up the track towards the train. Stand well to one side (high speed trains have quite a slipstream) and wave a car blanket or bright coloured garment to warn the driver. If he is doing his job properly he will know that he is approaching a crossing and should look ahead to see that all is clear.

DISASTER IN THE AIR

A plane crash or forced landing in difficult terrain is one of the most dramatic of disaster scenarios. Since it could happen anywhere the individual cannot prepare for any specific situation.

Airline cabin staff are trained for such emergencies and you should follow their instructions. Aircrew will be trying to land the plane as safely as possible; there is nothing you can do except to keep calm and support the crew in calming the other passengers.

To prepare for a crash landing, tighten the seat belt, link arms with people on either side, hold your chin firmly down on your chest, lean forwards over a cushion, folded blanket or coat, interlink legs with your neighbours if seating permits it and brace yourself for impact.

When the aircraft finally stops moving – and not before – evacuate the aircraft as instructed in the pre-flight brief. If a ground landing, then quickly get away from the immediate area of the plane, as there is danger of fire or explosion. Even if there is no fire, keep away until the engines have cooled and any spilled fuel evaporated.

If ditching into water, dinghies will be automatically inflated and anchored on the wings. Do not inflate your own lifejacket while you are in the aircraft. To do so would restrict your exit. Wait until you are in the water and then pull the toggle to inflate it and get into a dinghy.

If the plane is sinking, release the dinghy from its anchorage as soon as passengers and equipment are stowed. As you leave the plane the more kit you can take with you the better. But do not stop to gather personal belongings and luggage.

SURVIVAL TIP:

If bailing out from a plane by parachute in wild country make your way to the wreck if you can – the wreckage will be much more noticeable to rescuers than a single person or a parachute.

After the crash

However self-disciplined you are, the entry into this kind of survival situation will be dramatic, abrupt and confusing. You will be in a state of shock and may be on the verge of panic. If there is fire or the risk of fire or explosion, keep at a distance until that danger seems to have passed, but no further away than seems necessary for safety. Do not allow anyone to smoke if fuel has been spilled. You must not blunder off into unknown terrain, especially at night, and need to maintain contact with other survivors.

Move injured persons to a safe distance with you and try to account for all the people involved. The immediate treatment of the injured is a priority. Treat cases in order of severity of their injuries and with each individual deal first with breathing difficulties, then in sequence, with major bleeding, wounds, fractures and shock.

Separate the dead from the living if possible – the deaths are part of the frightening strangeness of the event and the survivors will be easier to calm down.

Even with a fire, all may not have been destroyed. Investigate the wreckage and salvage whatever you can of equipment, food, clothing and water. Take NO risks if there is still a chance that fuel tanks could ignite and beware of any noxious fumes from wreckage which has been smouldering.

If you have to wait for fire to burn out, take stock of the location in which you find yourself – which should in any case be the next step in your strategy. Is it practical and safe to remain where you are? If your anticipated route is known – and with a flight it will be – some kind of search and rescue operation can be expected and there are considerable advantages in staying where you are.

Searchers will already have some idea of your location, and even if you have been forced off route they will have a record of your last reported position. The wreckage or grounded plane will be more noticeable from the air, especially in heavily wooded country where even a large group of people will be hidden by the trees.

If you find that you are in a very exposed or dangerous location then a move to a more protected position is necessary. However, do not move at night unless the threat to life outweighs the risks of trying to negotiate unknown terrain in the dark.

SURVIVAL TIP:

Leave an indication on the crash site of the direction in which you have moved off, so that it is possible for rescuers to know that there are survivors and to know in which direction to go on looking.

The usual reason for making an immediate move will be because you are in an exposed position on a mountain or hillside offering no protection from the elements, or at risk from rock falls or other dangers there. Move down, not up the slope, as conditions are likely to be less exposed on lower ground.

Do not all go off looking for a safer location. Send out scouts to investigate the surrounding terrain carefully. They must keep together, working in pairs, and not go off on individual explorations. They can maintain contact vocally and should mark their routes as they proceed so that they can easily retrace their steps.

PROTECTION

The first requirement will probably be some immediate shelter from the elements, especially for any injured. A more extended reconnaissance can follow to choose a proper campsite. Make the most of any natural shelter and augment it by using whatever materials are at hand.

If injuries are too severe for a person to be moved, some kind of shelter must be provided for them on the spot.

On bare ground, if there is no equipment or wreckage which can be utilized, then the only thing to do is dig down.

If possible, find a natural hollow and burrow deeper, using the excavated earth to build up the sides. This will at least get a casualty out of the wind. Get a fire going to provide warmth (it will also help raise morale) and use reflectors to maximize the heating effect, enabling you to conserve fuel.

If the circumstances make movement away unnecessary or impossible, follow similar procedures. Build up rocks, wreckage or equipment to form a wind break if no natural shelter is available. If in a group huddle together, it will reduce the loss of body heat. Survival time for badly injured persons in these circumstances is limited and you must hope for an early rescue. Fit people must go off in search of water, fuel, shelter materials and food – but always in at least pairs. Lay out as many signals as possible to attract attention.

Remember that shelter may be as necessary from the sun as it is from the wind and cold. Exposure is not only a matter of hypothermia.

LOCATION

If you have a radio or mobile phone you can signal for help – but do not go back on board a damaged and potentially explosive aircraft to do so. Wait until you are sure it is quite safe. The rescue party will want to know your location. Those who were travelling overland should have a good idea of their position – even if temporarily lost – and with a map should be able to give a more accurate fix. If you are the victim of disaster at sea or in the air, however, it will help considerably if you know your planned course and have some idea of your position when disaster struck, as well as of wind or current directions.

As often as not you must light fires – three fires are an internationally recognized distress signal. Make them as large as possible. Lay ground signals to attract attention, use pyrotechnics when you know help is within range and even make a noise when help is very near. This is when you are glad that the responsible authorities were told of your intentions and that you kept precisely to your route. It is only a matter of time before rescue comes. Meanwhile make yourself as comfortable as possible.

However, even the most careful plans may go astray. Navigational instruments could fail, storms, high winds or fog could all throw you off course and there you are, safe in your shelter but with no one knowing where. You could have a longer wait than you anticipate and you need to provide for it.

You also need to assess where you are on a more local scale, to study the terrain for anything it can tell you, not only to pin-point your position – if that is possible – but to see if there are safer and more comfortable locations to pitch camp, sources for fuel, food and water. In the long term you will also be assessing the possibility of making your own way across the land.

At sea you will be looking out for any indications that, rather than staying put, there is land close enough for your survival chances to be greater if you try to reach it rather than holding your present position. But you are at the mercy of wind and current, though you can delay your drift with a sea anchor.

SURVIVAL TIP:

On land, it is seldom most sensible to set out immediately to walk to safety, rather than wait for rescue. However, if you know that no one will be aware that you are missing, if the terrain is so barren that it provides no food, water or shelter, or if you feel convinced that your reserves of energy and rations are sufficient to see you back to civilization, or to a location where you are sure you will be able to live off the land, you may decide to set off as soon as the light is good enough or conditions are otherwise right.

ACQUIRING FOOD AND WATER

On an isolated cliff ledge, cut off by the tide or forced by storm or mist to wait until you can move on, there may be little opportunity to exploit natural resources. Do not tuck into emergency rations immediately. You may be there for some time and, hungry though you may be, you should ration them out, allowing for a much longer wait than even a pessimistic assessment suggests. Even in such a situation there may be water and food within reach.

Elsewhere save your emergency rations for when there is nothing else and tap nature's resources first. Do not just find one source of food. Seek out a variety of plants for leaves, fruit, nuts, roots and other edible parts. Look for signs of animals which can be trapped or hunted.

When it is your very survival that is at stake there is no place for squeamishness about what you will or will not eat or about how you acquire your food, but that does not mean that you should totally abandon concern for wild life and the environment. When there is an abundance of other choices there is no reason to take already endangered species for your food – animal or vegetable – nor to set traps (which cannot discriminate in what they catch and maim) that will produce more meat than you can eat fresh or preserve. Making the most of nature's resources does not mean plundering them. Over-exploitation will be to your own disadvantage if you have to stay in the area for a long time.

Remember, too, that the most easily obtained nutritious food may be quite different from what you usually eat. If you have already learned to eat an unusual diet as part of your training you will find it much easier to feed yourself and will be able to encourage others to eat the same things.

Fuel for a fire will be needed for boiling water even if the temperature does not demand a fire for warmth – but do not be deluded into thinking that a warm day is going to be followed by a warm night. There can be dramatic temperature changes from day to night in some parts of the world.

SURVIVAL TIP:

In the short-term water is much more vital than food for your survival. If fresh running water is not available there are many other sources you can tap, but sterilize or boil to ensure that it is pure. Make finding water sources a priority.

NAVIGATION

The first thing to do is to coordinate signals (flares or fires for example) so that you are best prepared to receive rescue. Then organise water, food, and shelter.

You should stay where you are and only consider moving if you're en route to a final destination. The advantage of this is that you will be able to use any equipment and materials from the plane or vehicle and you will be able to use the wreckage as shelter. Also it will be easier for you to be spotted from the air.

YOU MUST PLAN

Remember this, it may save your life one day:

P Protection
L Location
A Acquisition
N Navigation

PEOPLE

For an expedition, the planning will include a careful selection of compatible personalities, selected for their fitness, both physically and in experience and training, for the particular project. In a disaster situation anyone may react unexpectedly under stress. With a mishap affecting members of the general public there may be a very varied group of people thrown together. Men, women and children, elderly people and babies. There may be pregnant women and people with medical problems or physical disabilities that require particular attention. The accident situations which involve such a varied group are also likely to involve a higher risk of injuries than among a hand-picked group of the trained and fit.

Babies may look very fragile – but they are very tough. However, they must be kept warm and fed regularly. Children will need reassuring and comforting, especially if they have lost the people with them or they are themselves in pain. Often the adventure of the situation will help to keep them from becoming too worried and it will help to keep them occupied, but they should not be allowed to wander, to play with fire or otherwise expose themselves to further danger. Old people are usually mentally tough and can give reassurance to the young, but they must be kept warm and fed regularly. It often seems true that women handle emergencies much better than men and are able to accept responsibility for others more easily.

With a ship or commercial airline the ship's officers or flight crew can be expected to take charge of the situation, if they are among the survivors, but there will not be the military chain of command or the acceptance of leadership and responsibility which can be expected in a compact organized group. Some democratic procedure to make decisions, plan action and maintain morale must be attempted. The trauma of the experience may leave some people eager to follow any leadership which gives them hope, but it will also throw into relief antagonisms and prejudices which must be overcome.

In an air or sea disaster people of different cultures and backgrounds may be thrown together and forced into situations which their own social taboos would not permit. Considerable tact may be necessary to overcome these problems. SURVIVAL, however, must take precedence.

The wider your medical knowledge the better, but giving people the will to survive is important and much of this can be achieved by a good 'bedside manner' – if you can give the impression that you know what you are doing you are half-way there.

Calmness and confidence in yourself will inspire the confidence and cooperation of others. The more knowledge you have the better you will be able to cope.

3

CLIMATE & TERRAIN

Although basic survival strategy and techniques are applicable anywhere, conditions vary widely around the world. It is essential to know as much as possible about conditions in any regions in which you expect to travel. A general knowledge of what you may expect in different climates will greatly increase your ability to handle the survival situation, if accident throws you into totally unfamiliar territory.

A few pages in this book cannot provide a world geography. They can only set the scene for the major types of climate and environment and suggest some of the principal ways of overcoming the problems they present.

Advice on specific topics, such as food and shelter, applicable to particular conditions will be found throughout the other sections of the book, supplementing the information here.

CLIMATE ZONES 64
  Survival case study 67
POLAR REGIONS 68
  Travel 68
  Clothing 70
  Shelter 71
  Fire 72
  Water 72
  Food 73
  Arctic health 76
MOUNTAINS 77
  Snow and ice fields 80
  Avalanches 82
SEASHORES 84
  Water 86
  Food 87
  Dangers 91
ISLANDS 92
ARID REGIONS 94
  Water 95
  Shelter and fire 96
  Clothing 97
  Food 98
  Health 99
TROPICAL REGIONS 100
  Shelter 103
  Fire 103
  Food 103
  Dangers 104
VEHICLES 106

CLIMATE ZONES

People often view an alien environment as an enemy and feel they must fight it. This is not the way to survive – fight it and you will lose! There are dangers against which precautions must be taken, but nature is neutral. Learn to live with each climate and to use what it offers. Climate is not conditioned only by latitude; location within a continent and altitude are equally important.

□ Polar climates

Polar regions are regarded as those at latitudes higher than 60°33′ north and south, but cold weather skills may be needed at very high altitudes everywhere. Near the Equator, in the Andes for example, the snow line is not reached until an altitude of about 5000m (16,500ft), but the nearer the poles the lower the snow line will be – at the southern tip of South America there is permanent snow at only a few hundred metres (a thousand feet). Arctic conditions penetrate deep into the northern territories of Alaska, Canada, Greenland, Iceland, Scandinavia and Russia.

░ Tundra

South of the polar cap, the ground remains permanently frozen and vegetation is stunted. Snow melts in summer, but roots cannot penetrate the hard earth. High altitudes produce similar conditions.

▒ Northern coniferous forest

Between the arctic tundra and the main temperate lands is a forest zone, up to 1300km (800 miles) deep. In Russia, where it is known as the Taiga, the forests penetrate up to 1650km (1025 miles) north of the Arctic Circle along some Siberian rivers, but in the Hudson Bay area of Canada the tree line moves an equal distance south of the Circle.

Winters are long and severe, the ground frozen for much of the time, summers are short. For only 3–5 months of the year is the ground thawed sufficiently for water to reach the roots of the trees and plants, which especially flourish along the great rivers that flow to the Arctic Ocean. There is a wealth of game: elk, bear, otter, lynx, sable and squirrel, as well as smaller creatures, and many birds.

In summer, where the snow melt cannot drain, it creates swamps. Fallen trees and dense growths of sphagnum moss make the going difficult. Mosquitoes can be a nuisance (but they do not carry malaria).

Movement is easier in winter, if you have warm clothing. Travel along the rivers, where fishing is good, making a raft from the abundant deadfalls.

Temperate climates

The temperate zone of the northern hemisphere, and the similar climates of the southern hemisphere, probably offer the most equitable circumstances for survival without special skills or knowledge. They will be the areas best known to many readers of this book. These territories are also those most heavily urbanized and where the survival ordeal is not likely to be very extended.

A fit and healthy person, equipped with basic skills, would not be so cut off that they could not reach help within a few day's trek. Heavy winter conditions may call for polar skills.

▥ Deciduous forest

As the climate gets warmer and winters less severe, deciduous forest replaces the conifers. Oak, beech, maple and hickory are the main species in America; oak, beech, chestnut and lime, in Eurasia. Soil rich in humus supports many plants and fungi. Survival is not difficult, except at very high altitudes where tundra, or snowfield conditions, appear. Many of these areas have been cleared by man.

▤ Temperate grassland

Mainly central continental areas with hot summers, cold winters and moderate rainfall, these have become the world's great food-producing areas – grain is grown and cattle reared. Water can be a problem in summer and shelter in winter.

▩ Mediterranean regions

The lands bordering on the Mediterranean are semi-arid areas, with long hot summers and short dry winters. There is sunshine most of the year, and drying winds. At one time this region was forested with oaks. When these were cut down the soil eroded, much of the area became covered with evergreen shrub. The Chapparal of California is very similar. Trees are few and water is a problem. At high altitudes, other conditions prevail.

▧ Tropical forests

The land between the tropics includes areas of cultivation and extremes of swamp and desert, but one-third is undeveloped forest: equatorial rainforest, sub-tropical rainforest and montane forest. All feature high rainfall and rugged mountains, which drain into large, swift-flowing rivers, with coastal and other low-lying regions often as swampland.

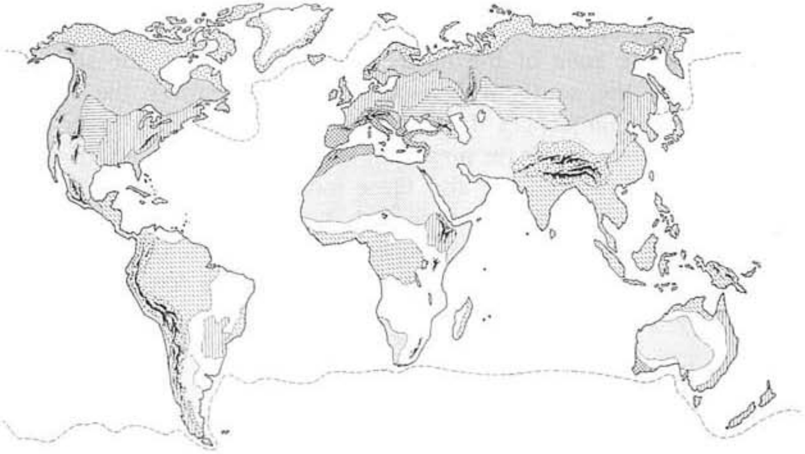

Savannah

This is tropical grassland, lying usually between the desert and the tropical forest. Near the forests the grass is tall, up to 3m (10ft) high, and trees more frequent. Temperatures are high the whole year round. More than one-third of Africa is savannah as are large areas of Australia, which are dotted with eucalyptus trees. Similar areas are the llanos of Venezuela and Colombia and the campos in Brazil. Often water is not easily available but, where it is found, there will be lusher vegetation and plenty of wildlife. In Africa large herds of animals can be found.

Deserts

One-fifth of the earth's land surface is desert – dry barren land where survival is very difficult. Deserts occur where air currents, which rose at the Equator and have already shed their moisture, descend and are rewarmed as they near the Earth, taking what little local moisture is present. There are rarely any clouds to give protection from the sun or to retain heat at night so that great extremes of temperature occur from the highest shade temperatures (58°C/136°F in the Sahara) to below freezing point at night. Only small parts of the world's deserts are sand (about one-tenth of the Sahara) the greater part is flat gravel cut by dried up water courses (wadis). The wind has blown the sand away, piling it up in low-lying areas. Elsewhere there may be wind-carved mountains, dried mudflats and larva flows.

SURVIVAL CASE STUDY

In the mid-1960s I was training a new army unit in Kenya. This training involved tactics, field craft and navigation. They were all experienced soldiers but lacked knowledge in small group operations. The training area was in the Northern Frontier district of Kenya, a very hot and dry area.

The recruits were dropped off in groups of 4 and had to navigate across country to a series of rendezvous points (RVs), where they were given water. My job with 3 others was to man these RVs, ensuring that everyone was accounted for and in good health. Each recruit was given half a pint of water strictly issued from a jerry can which held 4.5 gallons (36 pints). On the third day the heat was taking its toll and everyone was complaining of various ailments like blisters, sores caused by backpacks and fatigue. But one student was particularly bad. He was delirious and verging on the point of collapse. While I was treating him, his partner grabbed a jerry can and tried to swallow the lot. What he couldn't swallow he spilt, and by the time we wrestled the can from him it was almost empty. This was the last of our supplies and there were still more than 30 men to come. The nearest source of water was 3 days away and with no re-supply possible the exercise turned from hard to critical. Although all the recruits came from nearby areas, they had no water discipline, which we found out to our cost.

By strictly rationing and sending the strongest for help, everyone survived, but it's a lesson they will never forget.

Lessons learnt:
Never underestimate the power of self-preservation
Protect your water supply
Enforce strict rationing

Of all the discomforts, thirst is the deadliest.

When treating a person with dehydration, wet their lips first and look for a gag reflex (swallowing). Administer small sips only. When giving treatment, always make sure the patient is lying down in any available shade.

POLAR REGIONS

Antarctica is covered with a sheet of ice. In the Arctic the Pole is capped by deep ice floating on the sea and all the land north of the timber line is frozen. There are only two seasons – a long winter and a short summer – the day varying from complete darkness in midwinter to 24 hours daylight at midsummer.

Arctic summer temperatures can rise to 18°C (65°F), except on glaciers and frozen seas, but fall in winter to as low as -56°C (-81°F) and are never above freezing point. In the northern forests summer temperatures can reach 37°C (100°F), but altitude pushes winter temperatures even lower than in the Arctic. In Eastern Siberia -69°C (-94°F) has been recorded at Verkhotansk! Temperatures in the Antarctic are even lower than in the Arctic.

Antarctic winds of 177kmph (110mph) have been recorded and, in the Arctic autumn, winter winds reach hurricane force and can whip snow 30m (100ft) into the air, giving the impression of a blizzard – even when it is not snowing. Accompanied by low temperatures, winds have a marked chilling effect – much greater than the thermometer indicates. For instance, a 32kmph (20mph) wind will bring a temperature of -14°C (7°F) down to -34°C (-30°F) and one at 64kmph (40mph) would make it -42°C (-44°F) with even greater drops at lower temperatures. Speeds over 64kmph (40mph) do not appear to make a greater difference.

TRAVEL

Experience shows the best policy is to stay near an aircraft or disabled vehicle. If the spot is hazardous establish a safe shelter as close by as possible. A decision to walk-out will be based on nearness to civilisation and probability of rescue.

Decide early what to do – while you can still think clearly. Cold dulls the mind.

Movement in a blizzard is out of the question and, at all times, navigation is difficult on featureless ice and tundra. Ice movement pushes up ridges which make the going treacherous. Summer melt water makes the tundra boggy and even sea ice slushy underfoot.

Mosquito, black-fly, deerfly and midges can all be a nuisance in the arctic summer. Their larvae live in water – avoid making shelter nearby. Keep sleeves down, collar up, wear a net over the head and burn green wood and leaves on the fire – smoke keeps them at bay. When it turns colder, these nuisances are less active and they disappear at night.

In Alaska, northwestern and northeastern Canada, Greenland, Iceland, Scandinavia, Novaya Zemlya, Spitzbergen and on other islands there are mountains where ice cliffs, glaciers, crevasses and avalanches are hazards. Near the Arctic coastline frequent fog from May to August, sometimes carried far inland, increases navigation problems.

Navigation

Compasses are unreliable near the Poles, the constellations are better direction-finders and the nights are light enough to travel by. By day use the shadow tip method (see *Reading the signs*).

Travelling on sea ice do NOT use icebergs or distant landmarks to fix direction. Floes are constantly moving – relative positions may change. Watch for ice breaking up and, if forced to cross from floe to floe, leap from and to a spot at least 60cm (2ft) from the edge. Survivors have been rescued from floes drifting south but sooner or later ice floating into the warmer oceans will melt – though that chance may be worth taking.

AVOID icebergs, they have most of their mass below the water. As this melts, they can turn over without warning, particularly with your added weight.

AVOID sailing close to ice-cliffs. Glaciers may 'calve' huge masses of ice, often thousands of tons, which break off into the sea without warning.

Bird observations can aid navigation. Migrating wildfowl fly to land in the thaw. Most seabirds fly out to sea during the day and return at night.

Sky reflections help to determine distant terrain. Clouds over open water, timber or snow-free ground appear black below; over sea ice and snowfields, white. New ice produces greyish reflections, mottled ones indicate pack ice or drifted snow.

Snow shoes

All polar travel is strenuous and should only be attempted by a fit person. On snow with a hard crust, skis are the best means of travel, though difficult to improvise. Skiing in deep loose snow takes great effort and, in soft snow, snow shoes are better. To walk in snow shoes lift each foot without angling it, unlike a normal stride, keeping the shoe as flat to the ground as possible.

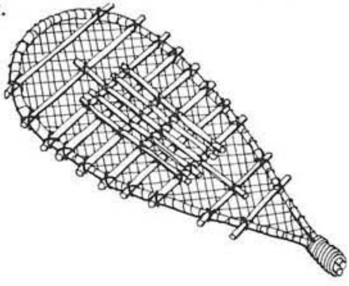

Bend a long green sapling back on itself to form a loop and secure ends firmly. Add crosspieces and twine – the more the better – but do not make the shoes too heavy. You will not be able to walk far without getting very tired. Allow a firmer central section to attach to your foot.

Follow rivers

Travel downstream – by raft in summer, on the ice in winter – except in northern Siberia where rivers flow north.

On frozen rivers keep to smoother ice at the edges and to outer curve on the bends. Where two rivers join follow the outside edge or take to the outer bank. If the river has many bends, leave the ice and travel by higher ridges.

WARNING

ICE COLD WATER IS A KILLER

Falling into icy water knocks the breath out of you. The body curls up with loss of muscular control and violent shivering. Exposed parts freeze in about 4 minutes, consciousness clouds in 7, death follows in 15–20.

RESIST! Take violent action on hitting the water. Move fast for land. Then roll in snow to absorb water. Get to shelter and into dry kit immediately.

CLOTHING

Severe cold and harsh winds can freeze unprotected flesh in minutes. Protect the whole body, hands and feet. Wear a hood – it should have a drawstring so that it can partly cover the face. Fur trimming will prevent moisture in the breath freezing on the face and injuring the skin.

Outer garments should be windproof, with a close enough weave to prevent snow compacting, but porous enough to allow water vapour to escape – NOT waterproof, which could create condensation inside. Under layers should trap air to provide heat insulation. Skins make ideal outer clothing.

Openings allow heat to escape, movement can drive air out through them. If clothing has no draw strings, tie something around sleeves above cuffs, tuck trousers into socks or boots.

If you begin to sweat loosen some closures (collar, cuffs). If still too warm remove a layer. Do so when doing jobs like chopping wood or shelter building.

Only a plane crash or forced landing is likely to leave someone in polar regions unequipped. Try to improvise suitable clothing before leaving the plane.

Wear wool – it does not absorb water and is warm even when damp. Spaces between the knit trap body heat. It is best for inner garments.

Cotton acts like a wick, absorbing moisture. When wet, it can lose heat 240 times faster than when dry.

Feet

Mukluks, boots of waterproof canvas with a rubber sole which comes up to the caulk and with a drawstring to adjust fitting, are ideal. Ideally they should have an insulated liner.

Insulate feet with three pairs of socks, graded in size to fit over each other and not wrinkle. If necessary, improvise foot coverings with several layers of fabric. Canvas seat covers can make improvised boots.

Trenchfoot can develop when the feet are immersed in water for long periods, as in the boggy tundra during the summer months (see *Cold climate hazards* in *Health*).

Snow glare

Protect the eyes with goggles or a strip of cloth or bark with narrow slits cut for eyes. The intensity of the sun's rays, reflected by snow, can cause snow blindness. Blacken beneath the eyes with charcoal to reduce glare further.

C.O.L.D. – THE KEY TO KEEPING WARM

Keep it - **C**LEAN – Dirt and grease block air spaces!
Avoid - **O**VERHEATING – Ventilate!
Wear it - **L**OOSE – Allow air to circulate!
Keep it - **D**RY – Outside and inside!

SHELTER

You cannot stay in the open to rest. GET OUT OF THE WIND! Look for natural shelter you can improve on, but AVOID the lee side of cliffs where snow could drift and bury your shelter, or sites where rock fall or avalanche is likely. Avoid snowladen trees – the weight could bring down frozen branches – unless the lower boughs are supported on the snow. There may be a space beneath the branch which will provide a ready-made shelter (see *Shelter* in *Camp Craft*).

REMEMBER: Don't block EVERY hole to keep out draughts. You MUST have ventilation, especially if burning a fire inside your shelter. Otherwise you may asphyxiate.

FIRE

Essential for polar survival. Fuel oil from wreckage can provide heat. Drain oil from sump and reservoir on to the ground as soon as possible – as it cools it will congeal and become impossible to drain. High octane fuel does not freeze so quickly – leave it in the tanks.

In the Antarctic and on the Arctic ice, seal and bird fat are the only other fuel sources. On coasts driftwood can sometimes be collected – Greenlanders used to build homes from timber which drifted across the Arctic from Siberian rivers. In the tundra low, spreading, willow can be found. Birch scrub and juniper also grow beyond the forests. Birch bark makes excellent kindling – the wood is oily. Feather a branch and it will burn even when wet.

Casiope is another low spreading heather-like plant that Eskimos use for fuel. Evergreen, with tiny leaves and white bell-shaped flowers, and only 10-30cm (4-12in) high, it contains so much resin that it, too, burns when wet.

Casiope

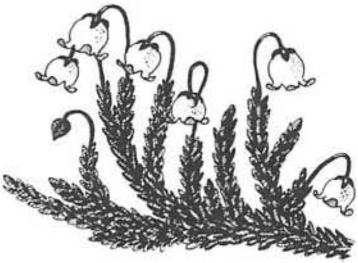

WATER

Even in the cold you need over a litre (1¾pt) daily to replace losses. In summer water is plentiful in tundra lakes and streams. Pond water may look brown and taste brackish but vegetation growing in it keeps it fresh. If in doubt BOIL.

In winter melt ice and snow. Do NOT eat crushed ice, it can injure your mouth and lips and also cause further dehydration. Thaw snow sufficiently to mould into a ball before attempting to suck it.

REMEMBER: if already cold and tired eating snow will further chill your body.

FOOD

Antarctic: Lichens and mosses, growing on dark, heat-absorbing rocks on some northern coasts, are the only plants. Seas are rich in plankton and krill which support fish, whales, seals and many seabirds. Most birds migrate in autumn, but flightless penguins stay. They make good eating. Most of the year they take to the water at the first sign of danger but, when incubating eggs, sit tight on their burrows or scrapes.

Arctic: Ice provides no habitat for plants or ground animals, even polar bears are likely only where they can find prey – and they are difficult and dangerous to hunt. Seabirds, fish and seals, where there is water, are the potential foods. Foxes – the Arctic fox turns white in winter – sometimes follow polar bears on to sea ice to scavenge their kills. Northern wildlife is migratory and availability depends on season.

Tundra and forest: Plants and animals can be found in winter and summer and the northern forests offer even more wildlife. Tundra plant species are the same in Russia as in Alaska. All are small compared to warmer climate plants: ground spreading willow, birch and berry plants with high vitamin content. Lichens and mosses, found widely, form a valuable food source – especially reindeer moss.

Poisonous plants

The majority of Arctic plants are edible, but AVOID Water Hemlock – the most poisonous. AVOID the fruit of the Baneberry. AVOID small Arctic buttercups. Other temperate poisonous species found far north include Lupin, Monkshood, Larkspur, Vetch (Locoweed), False hellebore and Death camas. Best avoid fungi too – make sure you can distinguish lichens from them! There are no Arctic plants which are known to produce contact poisoning.

Animals for food

Bark and greenery stripped from trees is evidence of feeding animals. Caribou (reindeer) are common from Alaska to west Greenland and found across northern Scandinavia and Siberia. Shaggy musk-ox roam in northern Greenland and, in the islands of the Canadian archipelago, elk (Moose) are found – where there is a mixture of forest and open ground.

Wolves are common in northern Canada, Alaska and Siberia (but rare and protected in most European countries). Foxes, living in the tundra in summer and open woodland in winter, are an indication of other, smaller prey – mountain hares, squirrels and other small rodents which burrow beneath the snow to find seeds. Lemmings make runways beneath the snow. Beaver, mink, wolverines and weasels can all be found in the Arctic.

Bears roam the barren lands of the north as well as the forests. They can be dangerous. Give them a wide berth.

The best chances for survival are along coasts where the sea provides a dependable source of food. Seals are found on coasts, pack ice and in the open water.

Walruses may look cumbersome but are also very dangerous. LEAVE THEM ALONE UNLESS YOU ARE ARMED.

Hunting and Trapping

Tracks are clear in snow and easy to follow – but leave a trail of fluttering flags of bright material from wreckage to find your way back to your shelter. Make them high enough not to be covered by a fresh snowfall.

Caribou can be very curious and may sometimes be lured by waving a cloth and moving on all fours. Imitating a four-legged animal may also bring wolves closer, thinking you might be prey. Ground squirrels and marmots may run into you if you are between them and their holes. Some prey animals can be attracted by the sound made by kissing the back of your hand. It is like the noise made by a wounded mouse or bird. Make it from a concealed position and downwind. Be patient. Keep on trying.

Stalking animals is difficult in the exposed Arctic. If you have a projectile weapon – gun, bow, catapult – which can be fired from ground level, lie in ambush behind a screen of snow. To be more mobile make a screen of cloth which can stand in front of you, and slowly be moved forward.

In winter, owls, ravens and ptarmigans – the birds available in the north – are usually 'tame' and can be approached slowly, without sudden movements. Many polar birds have a 2–3 week summer moult, which makes them flightless – they can be run down. Eggs are among the safest foods and are edible at any stage of embryo development.

Seals

A main source of food on polar ice, some seals remain there right through the winter. The Antarctic Weddel seal, most southerly of mammals, can dive for 15 minutes before coming up to breathe from pockets of air beneath the ice, or at small holes which it keeps open by nibbling around the edges. Most seals must breathe more frequently. Few are as formidable as the Elephant seal, which can rear up to twice a man's height in attack or defence.

Seals are most vulnerable on the ice floes with their young pups (produced between March and June in the Arctic according to species). Newborn seals cannot swim and are easy to catch – thousands are massacred by hunters and in culls each year by simply walking among them on the ice and clubbing them.

Out of the breeding season, breathing holes in the ice are the best place to catch seals; recognize them by their cone shape (narrower on the upper surface). In thicker ice they will be surrounded by flipper and toothmarks where the seal has been keeping the hole open. You have to be patient, yet ever ready, for the visits to the hole are brief. Club the animal then enlarge the hole to recover the carcass.

Seals provide food, clothing, moccasins and blubber for fires. Adult males have a strong odour early in the year, but it does not affect their meat.

Eat all except the liver, which at some times of the year has DANGEROUS concentrations of vitamin A. Cook seal meat to avoid Trichinosis.

POLAR BEARS

Confined to the high Arctic – in Europe only resident on Spitzbergen – they have a keen sense of smell and are tireless hunters on sea ice and in the sea. Feeding mainly on seals, with some fish, they swim well and can stay submerged for two minutes. Rarely found on land – though in summer they may feed on berries and lemmings. Like many cold-climate animals they are larger than their warmer-climate relatives. Most are curious and will come to you – but treat these powerful animals with respect and caution.

Always cook meat: muscles always carry the Trichinosis worm. NEVER eat polar bear liver which can have lethal concentrations of vitamin A.

Preparing meat

Bleed, gut and skin while the carcass is still warm. Roll hides before they freeze. Cut meat into usable portions and allow to freeze. Do not keep reheating. Once cooked, eat leftovers cold (that's why you cut it up). Leave fat on all animals except seals. Fat is essential in cold areas but, if you eat a lot, make sure you take plenty of fluids. Except in extreme cold (when it will freeze) remove seal fat and render it down before it turns rancid. It can also be useful fire fuel.

When food is scarce animals will steal it – so cache it carefully. If there are signs of would-be thieves look out for them – they could be your next meal.

Rodents, especially squirrels, and rabbits and hares, can carry Tularemia, which can be caught from ticks or handling infected animals. Wear gloves when skinning. Boiled flesh is safe.

ARCTIC HEALTH

Frostbite, hypothermia and snow blindness are the main hazards, while efforts to keep warm and exclude draughts can lead to lack of oxygen and carbon monoxide poisoning.

It is easy to withdraw from reality, layered in clothing and with the head wrapped in a hood. Thinking can become sluggish and obvious things overlooked. Keep 'switched on'. Keep active – but avoid fatigue and conserve energy for useful tasks. Sleep as much as possible – the cold will wake you before you freeze unless you are completely exhausted and cannot regenerate the heat you lose to the air.

AVOIDING FROSTBITE

- Wrinkle face to stop stiff patches forming, pulling muscles in every direction. Exercise hands.

- Watch yourself and others for patches of waxy, reddening or blackened skin, especially on faces, ears, hands.

- **AVOID** tight clothing which will reduce circulation.

- Dress inside warmth of sleeping bag (if you have one).

- Never go out without adequate clothing – however briefly. Avoid getting clothing wet, through sweat or water. Dry it as soon as possible if this happens.

- Knock snow off before entering shelter, or leave outer clothing at entrance. Snow will melt in warmth giving you more clothing to dry.

- Wear gloves and keep them dry. **NEVER** touch metal with bare hands.

- **AVOID** spilling petrol on bare flesh. In sub-zero temperatures it will freeze almost at once and does even more damage than water because of its lower melting point.

- Be especially careful if you have been working hard and are fatigued. If you are sick – **REST.**

Don't let the cold demoralize you. Think up ways to improve the shelter, how to make a better pair of gloves, for instance. Exercise fingers and toes to improve circulation.

Don't put off defecation – constipation is often brought on that way. Do try to time it conveniently before leaving your shelter, so that you can take waste out with you.

MOUNTAINS

Mountain peaks are exposed to high winds and often covered in snow. They provide neither food nor shelter. Climbing rock and negotiating ice and snowfields calls for special skills, which are best learned first-hand in mountaineering schools and practised under supervision. No inexperienced person should think of trying to tackle real mountaineering territory, except as a learner with a properly organized party. But disaster may leave you on a mountainside or force you to cross a mountain range to get to safety.

If no rescue is likely, the first aim in daylight should be to get down into the valleys where food and shelter are available. At night and in bad visibility this is too dangerous. Some kind of shelter must be found until visibility improves.

Dig into the snow if there is no shelter among rocks and no wreckage to provide cover. If below the snow line you must cover yourself to prevent exposure. A plastic bag will make an improvised sleeping bag, if you have no survival kit. Salvage blankets or covering from a crashed plane or use any clothing to cover yourself as much as possible, but do not pull clothes too tightly round you; air within the clothes will provide insulation.

On a slope, sleep with your head uphill; on rough and stony ground sleep on your stomach for greater comfort.

Judging terrain

As you descend a mountainside it will often be difficult to see what is below you. Can you move around a valley or along a spur to look back at what was below? The opposite side of a valley will give you some idea of what is on your side too.

Be cautious if you find you are looking at a distant slope beyond a foreground bluff, the ground is likely to fall steeply between. Scree slopes can be particularly deceptive and appear continuous until you are very close to a cliff.

Descent

Negotiating cliffs without a rope is extremely dangerous. On the steepest cliffs it is necessary to come down facing the cliff and very difficult to see footholds below. If there is an adjoining slope, a colleague can observe and give directions. Once down, you can then point out holds to others from below. A high cliff should never be attempted. In the case of a plane crash there is probably more risk in climbing than waiting for rescue.

To climb down rock faces which are less steep and with deeper ledges, adopt a sideways position using the inside hand for support. For easier crags, descend facing outwards with the body bent and where possible carry weight on the palms of the hands.

Ascent

Climbing upwards, holds are easier to see, but it is always safer to go round than over obstacles if you are travelling without knowing the route. You could get stuck with an impossible descent.

Always work out your route from the bottom and in climbing keep the body away from the rock and look up. Move only one hand or one foot at a time – always keep three points of contact. Keep your weight evenly balanced on the feet rather than hang from the hands. Do not overstretch.

With the feet firmly planted on the rock and one hand grasping a good hold, reach with the other for a hold just above the head. Test it and then look for another hold for the other hand or the feet. Use small intermediate holds, avoid becoming spread-eagled and let the legs do most of the work. Always place the feet as flat as possible to make maximum contact with the rock.

To climb vertically up fissures, use the chimney technique. Place your back against one surface and wedge your legs across the gap to the other. Slowly move up. If a chimney opens out you may have great difficulty in transferring to one face and have to descend again.

Descending by rope

With a rope firmly anchored at the upper level, it is possible to descend the sheerest cliff. The technique, known as abseiling or rappelling, can involve a special sit sling and a karabiner for the rope to pass through, but the basic method uses just a doubled rope. The rope does not move – you move down it. It is not comfortable, even with the body correctly angled, but it is the safest way to negotiate steep or very slippery slopes. Friction can damage clothing and skin, so if you can, pad out your shoulders and groin, and wear gloves.

The length of the rope controls the amount of descent and there must be a firm anchor point, a rock, or tree which can carry the weight and not

SURVIVAL TIP

Make sure that you are in a firm position before hauling the rope down – its sudden weight could affect your balance – and be sure that you have planned your next move. Once the rope is down you may have no way of retracing your steps.

cut the rope. If a series of platforms with firm anchors can be found, a slope can be negotiated in stages – but if several people are involved there must be room for all of them to wait at each stage.

After an abseil the rope can be pulled down after you. If someone is left above to untie the rope, or you are prepared to leave it behind, an undoubled rope can be used – making twice the descent possible with the same rope. Getting over the edge is often the most difficult part. You may have to climb down a few steps to gain a good position and sufficient confidence.

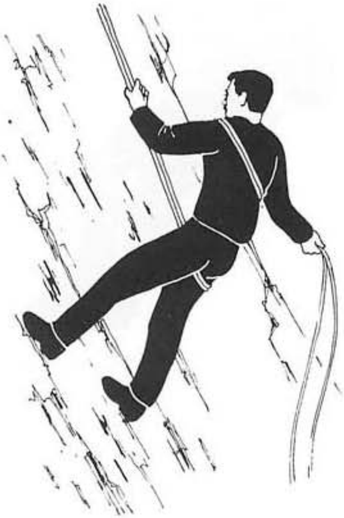

ABSEILING
Loop rope around firm anchor (test it with full body weight). Avoid sharp edges that could cut rope. Pass both ends of rope between legs from front, bring around to left of body and across chest, over right shoulder and down across back. Hold rope in front with left hand and at back with right. Plant feet about 45cm (18in) apart, firmly against slope and lean back. Let rope around body carry your weight. Do not try to support yourself with your upper hand. Step slowly downwards. The lower hand controls rate of descent. Pay the rope out one hand at a time.
CAUTION: *Abseiling can be dangerous. If not trained in the technique, NEVER attempt it, unless accompanied by an expert or in a survival situation.*

Using a cradle

On an unobstructed vertical descent, a cradle made from a bowline-on-the-bight (see *Knots* in *Camp Craft*) can be used to lower people down, or haul them up. Use this technique to rescue anyone who has fallen down a crevasse.

Ascending with ropes

Belaying is a method of helping others to climb up. First, one person must make the ascent with a rope (this could be a light line to haul up the actual rope afterwards) attached around the waist with a bowline. At each stage of the ascent there must be a platform or ledge to accommodate all the party and a secure anchor for the rope. If there are a number of lengths of rope a series of stages could be operated at the same time to handle a larger party.

Test that the anchor is firm – a tree, spike of rock or thread (a hole through rock, or a stone or small boulder firmly wedged in a crevice). Anchor the rope with a loop tied in a figure-of-eight or an overhand knot.

The belayer ties on with a bight or two bights to steady himself, and passes climbing rope over head and down to hips, making a twist around the arm closest to the anchor and takes up any slack. The climber ties on with a bowline around waist and begins to mount. The belayer takes in rope to keep it taut.

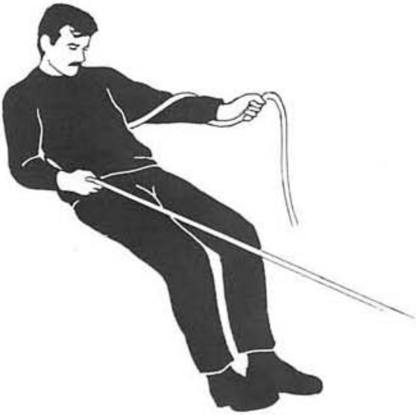

TAKING UP ROPE

Pull with both hands so that rope passes behind back (pull in with right hand, push away with left). Slide right hand out for more rope. Bring hands together and hold both parts of rope in right hand, while the left slides in towards body to take up slack. Begin again, pulling in with right hand, pulling rope around body with left. Be ready to arrest rope, in case climber falls. Bring rope tight around body by bringing hands together.

The anchor, belayer and climber should be in a straight line. If a spike is used it should be higher than the belayer's head. If this is not possible standing, the belayer should work from a sitting position.

Older people and children should be roped around the chest. Small children are best carried papoose-style on another climber's back.

Note: Belaying without an anchor is risky and requires more strength. The rope should then only pass through the belayer's fingers, NOT around the back, lest the belayer be pulled down by the climber.

WARNING

FALLING ROCK CAN KILL!

On loose rock always test holds gently and never pull outwards on a loose hold.

Be careful that your rope does not dislodge rocks. Even small falling rocks can inflict serious injury. If you knock a piece down, shout a warning to those below.

SNOW AND ICEFIELDS

Sophisticated equipment is available for climbing in snow and ice, but on snow some of the mountaineer's ice-axe techniques can be improvised with a stout stick – a handled walking stick may give more grip than a simple shaft. If not equipped with a proper ice axe and crampons and skilled in their use, try to keep clear of mountain ice.

An ice axe or stick, driven into the snow when climbing, gives stability. On steep slopes climb in zig-zags, kicking steps and digging your stick in sideways. Dig in heels and use a stick on slighter slopes. On gentle slopes use heels and stick as a walking stick. On steep slopes descend backwards driving stick into snow for support and as a brake if you slip. Sliding down a snow slope is exhilarating, but dangerous.

Digging in the heels will help control speed and a stick driven into the snow is an additional brake – but there is always a risk that you have not seen a precipice ahead! Never use this method where there is any risk of avalanche.

Security ropes on ice

Any party moving across a glacier should be tied together, at not less than 9m (30ft) intervals. The leader should probe the snow with a stick, for any slight depression could indicate a crevasse.

Ropes fixed to a firm anchor at both ends can steady movement across ice patches which have to be traversed. Use as a hand hold, or tie a short rope in a bowline around the waist and secure to the rope with a prusik knot. This will slide along the rope to allow descent but, if you slip, will arrest your fall. This is a technique also useful on scree and loose descents for children and the less able.

Ice and snow bollards

If no firm rock is available for belaying, an upper anchor can be cut from the ice. Cut in a mushroom shape where natural ice formation makes it easiest. Make the diameter at least 40cm (16in) and depth at least 15cm (6in). Discard and start again at the slightest sign of a crack in the ice.

A snow bollard must be much bigger: at least 30cm (1ft) deep and from 1m (3ft) wide in hard snow to 3m (10ft) in soft. Pack equipment and baggage around it to prevent rope cutting through.

Crevasses

Crevasses are found where a glacier starts at a valley wall, changes direction or spreads out in a widening valley. Travel slowly, probing the ground. If one of the group falls through the snow he is belayed by a rope and can be hauled out.

Pressure of the rope on the chest can cause asphyxiation. Pass a rope down with a loop to put a foot in to take the weight. If the faller is unconscious it will take three people to heave him out. Manharness hitches will enable them to pull together. Temperatures in a crevasse are very low and the victim will rapidly weaken. Speed is important.

AVALANCHES

Avalanches are a serious hazard in all high mountain regions. They most frequently occur on slopes of between 20° and 60°, and especially between 30° and 45°, usually within 24 hours of a snow fall.

Several things trigger avalanches, like temperature, ground conditions and noise. If you find yourself in avalanche prone areas be aware of these causes. Avoid if possible areas where fresh snow has just fallen on steep ground. After a major fall, waiting 24 hours for it to settle will help. The majority of victims of avalanches start them off themselves.

Rain, or a rise in temperature, after a snow fall greatly increases the risk. The melting process helps to lubricate the slide. Heavy snow falling during low temperature can also avalanche because it does not have enough time to stabilize.

Slopes with irregular surfaces are safest and timbered slopes are also stable. Steep rocks at the top of a slope make it more prone to slipping, because falling snow, rocks or icicles can set it in motion. On a convex slope the gravitational movement downward compacts the snow at the bottom and creates tension at the top making it more likely to slip. Where snow is building up on the lee side of a ridge or the head of steep gullies, it's under tension and the slightest disturbance can cause it to slide. Slopes with rocky outcrops and trees are safer to cross than bare ones. Carefully choose the best place to cross and before committing yourself, test the snow. Dig in your stick/ice axe and see if it's compacted or in layers. Throw rocks and make noise to try to encourage a slide, making sure you are well protected. On all dangerous ground it is best to rope together and use belays. Keep at least 15m (50ft) apart to help spread the load. If possible, let one person go across the more dangerous areas alone paying out the rope as he goes. When he is across he belays himself before the next person crosses.

MAIN AREAS OF DANGER

- Snow-covered convex slopes. Here the snow is under tension.
- Lee slopes where snow has accumulated. They are unstable.
- Deep snow-filled gullies.

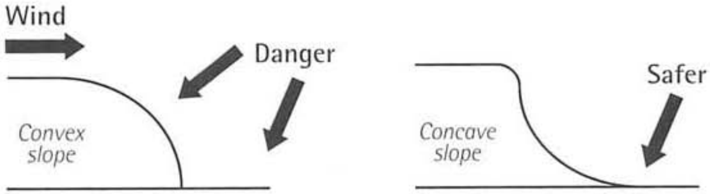

Never make camp on the lee side of a convex slope, a concave slope is safer.

As a precaution always carry a location beacon that omits a signal that rescuers can use to find you. Also carry an avalanche tape that can be streamed out in an emergency. It is made of bright nylon tape and makes detection easier if buried.

PRECAUTIONS

- The heat of the sun on the snow can cause avalanches so before noon travel in shaded areas – keep off those exposed to the sun.

- After noon, keep to slopes that have been exposed, avoiding those that are now in sun for the first time.

- Avoid small gullies and valleys with steep side walls.

- Stick to ridges and high ground above avalanche paths – you are more likely to trigger a slide but, if you do, have a better chance of being on top of the debris or not being carried down at all.

- Always look out for avalanche activity, even if you do not see it happening. Assess where avalanches started, their direction, how long ago they took place. They will be a guide to where other avalanches are likely.

A mud slide is very similar to an avalanche, but instead of tons of snow, masses of mud sliding on water bury all in its path. Avoid low-lying areas and water courses. Stay on the spurs and ridges. If caught up in a slide, use swimming actions to stay on top and go feet first.

SEASHORES

Most seashores offer abundant sources of food and excellent prospects for survival. Even where they appear bleak and barren, there is food to be had. Coastal waters are the home of many life forms – seaweeds, fish, seals, birds, molluscs and the plankton that supports the marine animals. Inland lakes and waterways of all kinds will also teem with life, with the exception of the Dead Sea and other areas of extreme salination, and those heavily polluted by man.

Coasts can range from sheer cliffs to long and gently sloping beaches. From the sea a towering cliff offers no opportunity to escape the water. Even a stretch of beach at its foot is likely to be cut off by high tide in tidal waters, though it could offer a few hours of respite before swimming off to find another landing place. All kinds of shore, however, offer resources to exploit and there are few better places to be stranded.

Sandy beaches

Sandy shores tend to be gentle and sloping. The tide goes out a long way, revealing large areas which are the habitat of burrowing species, left below the exposed sand. They include many worms and molluscs and they also attract feeding birds. Look for signs of buried molluscs. It is usually easier to spot the marks left by the syphons of buried bivalves under the shallow water at the sea's edge.

Where the sand is not inundated by the tide and is blown into dunes, it may be possible to find fresh water and it is here that plants will grow.

Sand is easily blown by the wind and gets into everything. Dunes also tend to be full of aggravating insects – so do not choose them for making camp or building shelter, if you can move beyond.

Muddy shores and estuaries

Where a slow-moving river joins the sea it deposits sediment, rich in nutrients, forming large mud flats. These can support many species of worms and molluscs and provide a rich feeding ground for birds and animals.

Rocky shores

Rocky shores, if their cliffs are not too sheer, trap pools of water when the tide recedes. These pools may teem with life. Rocks form a strata to which the many univalve shells can cling, an anchor for weed and sea urchins and crevices where octopus and other cephalopods can live.

Soft rocks, such as chalk, marl and limestone, erode quickly and their surfaces are smooth, but hard rocks fracture in chunks and provide good nesting sites for birds.

Pebble beaches

Stretches of pebble beach, often found between sandy and rocky sections of the shore, sustain least life. The continual movement of pebbles makes a difficult habitat for most plants and animals.

Tides

Tides vary considerably according to both location and time of year for they are caused by the counter-gravities of the sun and moon. In enclosed seas, such as the Mediterranean, they range over only a few metres. The Bay of Fundy, between Nova Scotia and New Brunswick, has a difference of 16m (52½ft) between low tide and high tide.

A line of debris along the beach; a change in appearance and texture from the long dry sand to that which is daily inundated; weed, shells and colour changes on vertical rock faces, all these will help to indicate the level to which the water is likely to rise.

Always check access to and from a beach or rocky shore – keep an eye on the rising water level so that you do not run the risk of being cut off.

Tides not only scour the beach but throw up valuable flotsam and jetsam, often providing fuel for fires, and may leave large fish stranded in rock pools along with their usual inhabitants.

SHORE SAFETY

- Never underestimate the power and danger of the sea. Time the tides and become familiar with their pattern, then you will be less likely to be cut off by an incoming tide or swept out by the ebb. It is easy to be caught when some way offshore.

- Cliffs offer problems of access. If there is only one way down make sure you can get back to it.

- Look out for strong currents, especially off headlands. Sand banks and submerged rocks are also dangerous. Where a beach falls steeply into deep water there will be a strong undertow. If you are forced to enter the water to rescue someone or retrieve kit, have a safety line around the waist and someone anchoring you on shore – or tie the line to a firm anchor.

Swimming

If caught in the undertow of a large wave, push off the bottom and swim to the surface. Swim to shore in the trough between waves. When the next large wave comes, face it and submerge. Let it pass and swim in the next trough shorewards.

When fishing or swimming stay within your depth, if you are not a strong swimmer. Watch out for large waves which can knock you off your feet and carry you further out. If this happens do not panic. People tend to panic when they do not know how deep water is, but are reassured when they do. It does not really matter for you can drown in 10cm (4in) of water as much as in 10m (33ft). If you want to test the depth place your arms above your head, point your toes and dive feet first. It may not be as deep as you thought!

If a strong current forces you offshore do NOT fight against it – you will lose. Swim ACROSS it, using side stroke, and make for land further along the coast. Side stroke is not the strongest or fastest stroke – but it is the least tiring.

If in the sea and being swept on to rocks, face shorewards and adopt a sitting position with your feet in front. They will absorb the initial shock and enable you to grab hold and scramble ashore. This is also useful for exploring water that may conceal hidden rocks – keep your shoes on and at least one layer of clothing.

Floating

A relaxed body floats best – so try to stay calm. It is difficult to sink in salt water. The main danger is swallowing salt water and choking on the vomit.

Women are more bouyant than men (they have an extra layer of fat) and float naturally on their backs. Men float naturally face down – but don't forget to lift your head out to breathe!

WATER

Fresh water along the coast is best obtained from small river outlets – large rivers tend to be full of silt and may be polluted by industry or other human activity upriver.

On sandy shores with no freshwater outlets there may be pools among dunes. Digging above the high water mark on any beach, especially if there is vegetation, may bring results. Dig down until you reach moist sand. Let water accumulate. Freshwater floats on top of salt. Scoop the lighter layer off the top.

Rock pools are unlikely to be freshwater. Even above the high water mark, they may be the result of wave splash, but you can sometimes identify fresh water by the growth of green algae which is not grazed by molluscs, as it would be elsewhere on the beach. Saltwater molluscs cannot survive in the reduced salinity.

Look for water trickling through rock faces, especially where there are ferns or mosses growing out of them, it will be drinkable. If stranded on a rocky outcrop off the shore, the only reliable source of water will be the sea itself – but NEVER drink sea water without distilling it. Far from quenching your thirst, it will take valuable body fluids away from the vital organs and eventually cause the kidneys to pack up.

Saltwater can be used for cooking – but do not eat until you have an adequate supply of fresh water. The residue of salt from distilled sea water can be used for preserving meat and fish.

FOOD

Seashore plants will differ according to the climate, but they will be available when weather or tide prevent you gathering food from the sea. In the water, you will find seaweeds of one kind or another where there is rock to give them purchase and water shallow enough to allow the sun to reach them. Seaweeds (more correctly called algae) are very valuable as food. In many parts of the world they form a major part of the diet and many are considered a delicacy from the cuisine of Japan to the laver bread of Wales. Seaweeds can be dried and stored for months.

Seafood

The best hunting for fish and molluscs will be at low tide, when rock pools can be inspected and buried molluscs and other creatures dug from sandy shores.

WARNING

Seaweeds are a valuable contribution to diet – but do NOT eat the blue-green algae sometimes found on freshwater pools. It is very poisonous.

Bivalves, which feed by filtering water through their digestive systems, can build up dangerous concentrations of toxic chemicals in areas polluted by industry or sewage.

In tropical zones, mussels are poisonous during the summer, especially when seas are reddish or highly phosphorescent. In the Arctic, black mussels can be poisonous at any time of year. Learn to recognize the cone shells, which shoot out a poisonous barb, in a few species potent enough to kill you. There are more than 400 types of cone shell, mainly found in the tropical Indo-Pacific with about 12 species off the southeast of USA and in the Caribbean. They are all identified by their shape. Tenebra or Auger and Turrid shells also have poison darts. Their venom is not dangerous to man, but a sting may still be painful.

Only eat molluscs collected live. Bivalves, such as oysters, clams and mussels, should close tightly if tapped gently. Gastropods, such as winkles and whelks, have a 'trap-door' (the operculum) to close the entrance to the shell. It should close tightly if the shell is shaken.

Other gastropods, such as limpets and Abalones, have no operculum but are tightly anchored to the rocks. Use a knife under the edge of the shell to prise them off. If they are hard to dislodge, they are good to eat. If they come off easily, they are probably dead or sick. After high tide any limpet found still fastened is good food – the tide washes away sick or dead specimens.

Cook shell foods by plunging them into boiling water and boiling for at least five minutes.

If you eat shell foods raw you expose yourself to parasites and pollutants which they may carry.

Fishing

Fish and sea snakes require more catching. Some fish are dangerous and all sea snakes are venomous. Distinguish snakes from eels by their scales and their broad flattened tails. They are said not to bite swimmers. Bites usually occur, and then only rarely, when fishermen are removing fish from nets in which the snakes are also caught.

On most coasts the best time to fish from the shore is about two hours after high water. If you fish when the tide is still coming in you are constantly retreating and probably getting wet. Remember that saltwater will rot boots and clothing.

Sea fishing requires a larger hook than freshwater fishing. A wide variety of bait can be used. Limpets, for instance, can often be found clinging to the rocks, or lugworms (*Arenicola*) can be dug up on sandy and muddy beaches. At low tide look for the coiled worm casts that show you where their L-shaped burrows are.

Make use of the tide to help you catch fish by building large arrow-shaped fish traps from stakes or rocks. Point them away from the shore. Fish will be caught when the tide recedes.

Octopus and squid

Octopus can be hunted at night, when they are in search of their own prey. Attract them with a light, then spear them. In daytime empty shells around a hole are an indication that an octopus may live inside. Drop in a baited hook, wait until it is taken and pull sharply up. The best way to kill an octopus is to turn it inside out: place a hand inside the fleshy hood, grab the innards and pull hard. Try it on a small octopus first! It takes practise so until you are proficient stab the octopus between the eyes or bang it against a rock.

All octopuses have a hard, parrot-like beak, and a few can give a poisonous bite. The worst is the Blue-ringed octopus of eastern Australia – its venom can be lethal. AVOID IT!

Octopus flesh is tough but chewy and very nourishing. Pounding it will help make it more tender. Boil the body and roast the smaller tentacles.

In the open sea squid can be huge, but a few small squid may occur inshore. Look for them in rock pools attached to seaweed. Catch them at night with a bright light, by jigging. Cuttlefish do not come close inshore but can be caught at sea in the same way.

Echinoderms

Another useful source of food, the echinoderms include starfish (not worth bothering with as food), sea urchins and sea cucumbers. Sea cucumbers creep about on the seabed or burrow in the sand. They look like warty black cucumbers, up to 20cm (8in) long. There are also spiky white sea gerkins and worm cucumbers (which burrow into sand). Sea cucumbers should be boiled for five minutes before eating.

Sea urchins, or sea eggs as they are sometimes called, are usually prickly balls which cling on to rocks, just below the low water mark but they have burrowing relations, the cake and heart urchins and the sea potato, which can be found beneath the sand. Split open and eat the egg-like 'yolk' inside. You can eat it raw, but it is safer to boil. AVOID any if their spines do not move when touched or if they smell bad when opened.

SEA URCHINS

Handle carefully. Their spines can inflict a painful wound – especially if you tread on them with bare feet! If you get pricked and the spine breaks off don't try to squeeze it out, it may push the spine in deeper. With luck spines will begin to work their way out of the skin after a few days.

Crustaceans

These include crabs, crayfish, shrimp and prawns, all of which make their homes in rock pools. Lobsters are usually found beyond the tidal zone but sometimes can be found in deep pools or crevices.

Look under stones and seaweeds – though you'll have to be quick to catch them! A net would help. Improvise one from clothing and a piece of wire or a sapling.

Sand crabs are abundant in the tropics. Active at night, they can be chased back to their burrow at the top of the beach and dug out. Some even climb trees and can be knocked down.

Freshwater crabs, crayfish and shrimps are also found in many parts of the world. They are smaller than sand crabs and usually found in shallow water.

All crustaceans spoil quickly and they may contain harmful parasites. They must be eaten as soon as possible – so keep them alive in water until you are ready to cook them. They are cooked alive – either by plunging into boiling water, so that they die almost immediately, or by putting them in cold water and heating it up, which is claimed to lull them to unconsciousness so that they feel no pain. Boil for 20 minutes.

Crabs have poisonous sections which must be removed: twist off claws and legs, then, with the crab on its back, place your thumbs under the flap at the tail and push upwards. Pull the flap up and away from the body and lift it off. This prevents the stomach contents from touching the flesh. Next press on the mouth with the thumbs, pushing down and outwards. This makes mouth and stomach come away in one piece. The lungs (known as 'Dead men's fingers'), which are harmful to eat, can then be pulled out and discarded.

Lobster is easier to prepare. Cut along the back towards the head and split open. The stomach lies just behind the mouth and is removed with the head and intestinal cord.

Turtles

Turtle meat is highly nutritious, and turtle eggs another good source of food – if you are lucky enough to have turtles come ashore. (See *Reptiles* in *Food*.)

Seabirds

Most ocean coasts are alive with seabirds. Fish for them. Leave baited hooks among offal on flat rocks, even throw baited hooks into the air to be taken on the wind. Try wrapping bait around stones. The sudden change in weight can make birds 'crash'.

Ground nesting birds can provide a rich source of eggs. Look for the ones that are easy to collect before risking raiding nests on cliffs. You can also try to catch the birds themselves at night when roosting – but do NOT risk climbing.

DANGERS

Unless very still, water higher than your thighs will be too murky to see through. You'll risk stepping on something unpleasant and waves could sweep you on to rocks or coral.

Wear shoes when foraging in the water – you need soles if you are improvising foot coverings. Cloth wrapping is NOT enough to protect from spines.

- **Jellyfish** are often swept inshore after a storm. Some, especially in the tropics, sting severely. The Sea wasps, or Box jellies, of northern Australian beaches, are the most dangerous. The bell-shaped body of the largest reaches only 25cm (10in) but its tentacles can reach 9m (30ft). Almost transparent, and difficult to see, each tentacle is armed with millions of stinging cells. Although their venom is one of the most deadly known and high concentrations cause skin lesions and death, usually only a very high dose is fatal to humans. Some jellyfish are not venomous but beware – size is not an indication of potency! If stung do NOT pull the tentacles off or wipe away the slime with your hand – you will only get stung more. Use seaweed, cloth, etc, to wipe the sting with sand.

- **Portuguese-man-of-war**, looks like a jellyfish but is actually a colony of polyps. It, too, can have tentacles 9m (30ft) long but, though its stings may cause irritation for several days, they are rarely fatal. Treat as for jellyfish.

- **Weaverfish** lie buried in the sand off the shores of Europe, West Africa and the Mediterranean. Their spines are venomous. Apply very hot water to sooth spine wounds.

- **Stingrays** occur inshore everywhere, but especially in warm waters, and electric rays in warm to temperate zones. Superbly camouflaged, they don't only hide in sand– some like rocky and pebbly places. Play safe – prod the bottom with a stick as you go. Stingray wounds can be soothed with very hot water.

- **Moray eels** may be found in shallow water. They have a savage bite and guard their holes tenaciously. Keep clear of any you see and do NOT put your hands into crevices!

- **Giant clams** on tropical reefs can be big enough to trap a limb if they snap shut on you.

- **Fish with venomous spines** often live in very shallow waters. Most common, and most dangerous, in the tropics, a few occur in temperate waters. Bottom-dwelling kinds are almost impossible to detect and are often superbly camouflaged. Zebra fish are easier to see, but equally dangerous to contact. Use a stick to stir up the sand and rocks in front of you.

- **Sea snakes** often occur in some numbers close in shore in the tropical Pacific and Indian Oceans. They are inoffensive and bites are rare – but their venom is the most toxic of all snake venom. Keep clear of snakes in the water. Found on shore, pin them with a forked stick – they make a good meal.

- **Many corals** are sharp and can easily cut you. Some, such as the fire corals, sting on contact. Always approach a reef with caution. Exploit other sites for food first. Both the reef and its inhabitants – which may include cone shells – can present dangers.

- **Sharks** Although most sharks feed mainly in deep waters, some species frequent shallow waters and swim up rivers and any might come onshore looking for an easy meal. Most shark attacks on humans occur in very shallow water. Be watchful!

- **Lagoons** Reefs are often formed around tropical islands or out from the shore, making a breakwater which leaves still waters in a lagoon. Fish in the lagoon are often of the poisonous varieties. Barracuda and Red snapper, which are edible in the open sea, should be avoided if caught in lagoons – their eating habits cause them to become toxic. Fish from the reef on the seaward side.

ISLANDS

Islands offer a special challenge to the survivor, especially small islands and those lacking resources. The feeling of loneliness is emphasized on an island and the sense of isolation acute. The problems are mental as well as physical. To help overcome them explore the island thoroughly and establish a daily routine.

Climb to the highest point to make a sketch of the island and get a mental picture of the terrain. Explore every creek, cranny, bay and beach of the coastline. Then take your reconnaissances inland until the island is familiar.

The island may have been inhabitated in the past – remains of buildings offer a basis for shelter. Fence posts and wire will be useful to repair your boat or build a raft. Vegetables may still be found growing and rats seem to follow man everywhere – sometimes they are the only permanent wildlife to be found.

Shelter will make life seem better and even a scrape in the ground will give some protection. If you find caves ensure that they are not tidal before you decide to use one. Remember that even caves that seem safe may be flooded or cut off by spring tides, which are higher than normal.

On a barren rocky outcrop shelter may simply mean finding a place out of the wind. Water may depend upon collecting rainfall and distillation. Food will be whatever clings to the rocks, birds and birds' eggs if you are lucky, and what you can haul from the sea.

Resources

On any small island resources will be limited. Take care not to over-exploit any one of them. Water is often a problem – lack of it is the reason many islands are uninhabited. Lush vegetation will draw attention to springs and streams. Digging above the high water mark may produce water. Catch and store rainwater.

To desalinate sea water by distillation you need a lot of fuel, which may be scarce. Driftwood may be available and some seaweeds will burn when dried – but you need wood to get the fire going. Seal blubber also makes good fuel. Have a fire only when you really need one. Search beaches after every tide – not just for wood. Everything has a use to the survivor.

Once familiar with your island, venture out at night - more creatures can be seen and foraging may be more rewarding.

Coconuts

Tropical islands are rarely desert islands – they will usually offer plenty to eat, both ashore and in the water. The coconut palm is found right through the tropics and subtropics and can provide: fronds for shelter, husks for ropes, growing points, which taste like cabbages, the milk and meat, and the shell (which you can also use as cups and containers).

To remove the fibrous husk around the coconut force it over a sharpened stake or split it with a hand axe. Extract the milk by piercing one of the dark 'eyes' of the nut itself before smashing the nut open to get at the meat.

Coconut milk is a safe and refreshing drink – a large nut may hold 1 litre (1¾pt). Do not drink from very young (green) or old (dark brown) nuts, their liquid will give you diarrhoea. The meat itself is indigestible in large amounts – eat only a little at a time.

Extract coconut oil by exposing chopped white meat to heat – sun or fire – and collecting oil as it runs off, or by boiling and skimming the oil as it rises to the surface. Rub it on to protect from sunburn, and chafing from saltwater, to repel insects, as a salve for sores and blisters or, mixed with wood ash, as a substitute for soap.

Climbing palms: If coconuts don't fall on your head and you can't knock them down, or if you need to reach some high bananas, don't try to climb the tree trunk like a rope. Instead, tie a strong bandage of cloth into a strap and slip it around your ankles. Adjust it to hold your feet close to the trunk and then you can press the soles of your feet inward and grip with them.

ATTRACTING RESCUE

- Lay out signals to attract searchers by arranging rocks, seaweed – anything that contrasts with its surroundings.
- Sand is excellent for polishing metal to make mirrors to signal with.
- If you can see a ship you can try and make contact on a VHF radio.

Moving on

In a group of islands, you may be able to move on to another when the first's resources are exhausted. In warmer climates it may be possible to swim, but build a raft in cold climates. If there is nothing to build a raft from, make some kind of flotation aid – even if it is only an empty box or coconuts.

If there are seals on your island you could use them to make a raft. From autumn through to spring, when the seal stores a lot of fat in its body, a seal carcass will float. If several are lashed together they will support your weight.

Study tides and currents between islands carefully, they can be treacherous. Float something you can observe and note its progress. Time your swim so that the ebb takes you out from your island and the high tide takes you in to the new island.

ARID REGIONS

Most desert lands were once fertile and some of the creatures that lived there then adapted to the new conditions. Like them, the survivor must learn to make the most of any available shade, to create protection from the sun, reduce moisture loss and restrict activity to the ends of the day and the night. Learn from the peoples who live or travel through the deserts.

In some deserts, especially the Sahara, the deserts of the Middle East, of Peru and northern Chile and parts of the Gobi Desert in

Mongolia, there are great temperature differences between night and day. At night condensation of any moisture in the air can make some water available – and in the Namib desert of southern Africa fog coming in from the sea often provides moisture for life. Elsewhere, in such deserts as those of Western Australia, northern Mexico and the Mohave of the southwestern USA, where the temperature changes are comparatively slight, there is very little condensation and consequently both plants and game are very rare. Sometimes, as in the Kalahari, there will be sparse grass and thorny bushes and, even in the most barren conditions, some kind of life seems to survive, though often invisible if you do not know where to look.

Dust and sand storms may occur at certain times of year, reducing visibility to zero and demanding maximum protection to prevent sand entering every orifice. Dust devils – desert whirlwinds like tornadoes – are quite common.

When rain does come – and in some territories years may pass with none at all – it may be in torrential downpours which create flash floods, before being quickly absorbed into the parched ground. This provides for a brief blossoming of vegetation and the emergence of species, such as the Spadefoot toad of Arizona, for rapid reproduction.

Desert rainfall and temperature

Typical of desert extremes are conditions in the Rub'al Khali, the 'Empty Quarter', of southern Arabia. For most of the year there is only a trace of rain but over 30mm (1¼in) may fall on a single day in the winter. July temperatures may reach over 48°C (118°F), dropping to 15°C (60°F) at night, and December extremes range from 26° to -6.6°C (79-20°F).

WATER

Water needs are paramount. Finding it is VITAL. If you have it, ration it immediately. If you are stranded by mechanical failure during a planned desert crossing, you will have plotted your route with an awareness of oases, wells and waterholes. Wells can be very deep and the water level require a container lowered on a line to reach it. Small water holes in wadi bottoms are often seasonal. They are usually covered with a stone or brushwood.

Away from known waterholes, try digging at the lowest point of the outside bend of a dry stream bed or at the lowest point between dunes. Do NOT dig in the heat of the day, the exertion will use up too much fluid and you may find none to replace it. You must always balance fluid loss against possible gain.

Exploit cactus and roots as water sources and, in deserts where the day/night temperature range is great, exploit this to produce water by condensation. (See *Water* in *Essentials.)*

Life expectancy

Life expectancy depends upon the water available and your ability to protect the body from exposure to the sun to minimize perspiration. Allow a slight negative balance. Drink 1.5 litres (2½pt) for every 2 litres (3½pt) lost and then drink at the rate the body is sweating. Efficiency is then impaired little and no water is wasted. Less fluid will not result in less sweating. Sweating is a cooling mechanism, not a way of losing moisture. If more fluid is drunk than needed it will be excreted and used to no purpose.

Without water you will last about 2½ days at 48°C (118°F) if you spend the whole time resting in the shade, though you could last as long as 12 days if the temperature stays below 21°C (70°F).

If you are forced to walk to safety the distance you cover will relate directly to water available. With none, a temperature of 48°C (120°F), walking only at night, resting all day, you could cover 40km (25 miles). Attempting to walk by day you would be lucky to complete 8km (5 miles) before collapse. At the same temperature, with about 2 litres (3½pt) of water you might cover 56km (35 miles) and last 3 days. Your chances are not appreciably increased until available water reaches about 4.5 litres (8pt) per person, though training and a determination to survive could contradict predictions.

SHELTER AND FIRE

Make a shelter from the sun and rest in its shade. You'll also need protection from winds and low night temperatures. Do NOT stay in a metal vehicle or aeroplane which may rapidly become overheated. Use it to support a shelter or make use of the shadow beneath an aircraft's wing. Make use of rock outcrops and the shadow provided by the sides of a wadi. Use the double layer technique to aid cooling (see *Tropical Regions*).

In a sand desert you may even be able to use wreckage to make a shelter beneath the sand. Many desert creatures spend the day beneath the surface, where the day temperature is much lower and the nights much warmer than outside. Sand will not permit tunnelling and you have to make a support structure.

Having provided immediate shade, build your shelter in the cool of the evening to conserve energy and fluids. Pile rocks to make a windbreak and make use of wadi walls (except when rain, and consequent flash floods, seem likely).

If using fabrics, leave the bottom edges lifted and loose by day to increase air circulation. Weigh them down with rocks at night. Avoid lying directly on hot ground. If you make a raised bed air can circulate beneath you.

You will need fire for warmth at night, and for boiling water. Smoke will also be very noticeable and useful for signalling. Desert scrub is dry and burns easily. If the land is totally barren, vehicle fuel and oil mixed with sand in a container will burn well (and is an easy way to light other fires) or use a string wick. Camel, donkey and other animal droppings burn well.

CLOTHING

Clothing helps reduce fluid loss and gives protection from sunburn – as well as warmth at night and a barrier against insect bites and thorns. In the desert it should be light and loose fitting, with air space between the garments and the body to provide insulation. Copy the flowing, layered garments of the Arab world. Trousers give more protection from insects than shorts (and guard against serious burns on the legs if forced into daytime exposure). Cover the head and feet.

KEEP COVERED

Do not strip off your clothes. Apart from the risk of severe sunburn, an uncovered body will lose sweat through evaporation requiring even more to cool it – but keep the covering as loose as possible so that there is a layer of insulating air. Sweating will then cool you more efficiently.

Headgear

Any hat with a piece of cloth attached to the back will give some protection to the head and back of the neck but it is better to copy the headgear of desert peoples. You need a piece of material about 120cm (4ft) square, a smaller piece, such as a handkerchief, and a piece of cord or cloth (a tie is ideal) to keep them in position.

Make the handkerchief into a wad on top of the head. Fold the large cloth diagonally, place it over the handkerchief, the long edge forward. Tie cord or cloth around the head to secure them.

Allowed to fall freely this will protect from the sun, trap pockets of air, take advantage of breezes and protect from sandstorms. At night wrap it around the face for warmth.

Eye protection

Sunglasses or goggles will help – though many made for use in temperate climes may offer insufficient protection. Soot from the fire smeared below the eyes will reduce glare reflected from the skin. Shield the eyes from glare and windborne sand with a strip of material. Cut narrow slits to see through.

Footwear

Do not walk barefoot on hot sand until your feet have become hardened. It will burn and cause blisters. Do not wear sandals which leave the top of the foot exposed. Improvise coverings if you have none. Putees will help keep sand out of boots or could be extended to wrap round the foot over open sandals.

FOOD

Heat usually produces a loss of appetite – so do not force yourself to eat. Protein foods increase metabolic heat and increase water loss and liquids are needed for digestion. If water is scarce, keep eating to a minimum and then try to eat only moisture-containing foods, such as fruit and vegetables.

Food spoils very quickly in the desert and any stores, once opened, should be eaten straight away or kept covered and shaded. Flies appear from nowhere and settle upon uncovered food.

Plants

Vegetation, away from oases and waterholes, is likely to be little more than scrub and grasses – even in the semi-desert – but grasses are edible and sometimes plentiful. The Acacia tree in the scrub provides edible beans. Beware of the Acacia's thorns but try all its soft parts: flowers, fruit, seeds, bark and young shoots.

The grasses of the Sahara and Gobi are neither nutritious nor palatable, but in the Sahara and the Asian deserts you may find the Desert Gourd, a member of the squash family. Its vine can run over the ground for 4.5m (15ft). Chew its water-filled shoots and eat its flowers and orange-sized fruits, the seeds of which are edible roasted or boiled.

The Mescal plant of the Mexican desert (an agave from which tequila is made), grows with a rosette of thick, tough, sharp-tipped leaves. Its central stalk, which rises like a candle to a flowering head, can be eaten. Cut the ends of the leaves to suck out water.

Animals

Deserts often support a variety of animal life which burrows into the sand or hides in any available shade during the day, including insects, reptiles, small rodents and specially adapted mammals such as the Fennec Fox of North Africa, the Australian Bandicoot, a hedgehog in the Gobi and the Jack Rabbit of North America – all of which have big ears to act as cooling aids. Most large mammals are an indication that there is a water supply within daily reach of their grazing areas.

The Sahara has gerbils and gerboas; the Middle East, caracals and hyenas; the New World, kangaroo rats and coyotes.

SURVIVAL SCENARIO

If you were in the desert and had to choose between carrying hessian or water, which would be more useful?

In desert conditions the hessian is vital of course. Water will be of no use at all if you can't build shelter to shade you from the blazing sun.

HEALTH

Most desert illnesses are caused by excessive exposure to sun and heat. They can be avoided by keeping head and body covered and remaining in shade until sundown.

- Constipation and pain in passing urine are common and salt-deficiency can lead to cramps.

- Continued heavy sweating on the body, coupled with rubbing by clothing can produce blockages in the sweat glands and an uncomfortable skin irritation known as prickly heat.

- Heat cramps, leading to heat exhaustion, heat stroke and serious sunburn are all dangers. A gradual increase in activity and daily exposure to the sun will build up a defence – provided that plenty of drinking water is available.

- Various micro-organisms attack the moist areas of the body – the crevices of the armpits, groin and between the toes. Prevention and treatment are to keep these areas clean and dry.

WARNING – DESERT SORES

In the desert even the most trivial wound is likely to become infected if not dealt with straight away. Thorns are easily picked up and should be pulled out as soon as possible. Where the skin is broken a large and painful sore may develop which could prevent walking. Bandage all cuts with clean dressings and use what medical aids are available.

TROPICAL REGIONS

Everything in the jungle thrives, including disease – germs breed at an alarming rate – and parasites. Nature provides water, food and materials for making shelters. Indigenous peoples have lived for millennia from hunting and gathering, but for the outsider it can take time to get used to the conditions and the non-stop activity.

Native peoples wear little, except as ornament, but the newcomer, uninured to insects and leeches and unaccustomed to moving through dense jungle growth, needs to keep as covered as possible. Clothing may become saturated by perspiration but it is better than being stung, scratched and bitten all over. Do not remove wet clothing until you halt and then, with humidity at 80-90 per cent, there is no point hanging it up to dry except in the sun or by a fire. Clothes saturated regularly by perspiration will rot.

Except at high altitudes, both equatorial and subtropical regions are characterized by high temperatures, heavy rainfall and oppressive humidity. At low altitudes, temperature variation is seldom more than 10°C (50°F), and is often 37°C (98°F). At altitudes over 1500m (5000ft) ice often forms at night. The rain has a slightly cooling effect but, when it stops, the temperature soars.

Rainfall is heavy, often with thunder and lightning. Sudden rain beats on the tree canopy, turning trickles into raging torrents and rivers rise at an alarming rate, but – just as suddenly – it is gone. Violent storms may occur, usually towards the end of the 'summer' months. Hurricanes, cyclones and typhoons develop over the sea and rush inland causing tidal waves and devastation. In choosing camp sites, make sure you are above any potential flooding. Prevailing winds create variation between winter and summer with the dry season (rain once a day) and the monsoon (continuous rain). In southeast Asia, winds from the Indian Ocean bring monsoon, but it is dry when the wind blows from the landmass of China.

Tropical day and night are of equal length, darkness falls quickly and daybreak is equally sudden.

Equatorial rainforests

The climate varies little in these forests, spread across the equator in the Amazon and Congo basins, parts of Indonesia and several Pacific islands. Rain of 1.5-3.5m (60-138in) is distributed evenly throughout the year. Temperatures range from 30°C (86°F) to 20°C (68°F) at night.

Where untouched by man, jungle trees rise from buttress roots to 60m (200ft), bursting into a mushroom of leaves. Below them, smaller trees produce a canopy so thick that little light reaches the jungle floor. Seedlings struggle beneath them to reach light and masses of vines and lianas twine up to the sun. Ferns, mosses and herbaceous plants push through a thick carpet of leaves and a great variety of fungi grow on leaves and fallen trunks.

It is fairly cool in this PRIMARY JUNGLE, with little undergrowth to hamper movement, but visibility is limited to about 50m (170ft). It is easy to lose a sense of direction and also difficult to spot anyone from the air.

RESCUE SIGNALS

Smoke is diffused by the tree canopy and may not be seen, especially if there is mist about as well. Set signals in a clearing, more often found near river bends, or – better – out on rafts on the river itself.

Secondary jungle

Growth is prolific where sunlight does penetrate to the jungle floor – mainly along river banks, on jungle fringes and where primary jungle has been cleared by man for slash and burn farming. When abandoned, this is reclaimed by a tangled mass of vegetation – look out for cultivated food plants which may survive among the others.

Grasses, ferns, shrubs and vines of secondary jungle reach heights of 2–3m (7–10ft) in a single year. Moving is slow, often hacking a way with a machete or parang – hot work, with visibility only a few metres (see *On the Move*). Jungle vegetation seems to be covered with thorns and spikes and bamboo thickets can be impenetrable barriers.

Sometimes, as in Belize in Central America, the jungle trees are low. Light does reach the fertile ground, producing abundant undergrowth even in primary jungle.

Sub-tropical rainforests

Found within 10° of the Equator, in Central and South America, Madagascar, western India, Burma, Vietnam, southeast Asia and the Phillipines, these forests have a season of reduced rainfall even drought, with the rain coming in cycles – monsoons. With more marked seasons there are more deciduous trees so that more light reaches the forest floor and undergrowth is dense.

Montane forests

When altitudes reach about 1000m (3300ft) in the tropics, and the areas bordering them, tropical forest begins to give way to montane forest. It becomes true montane at about 1240m (4100ft), as in the Monts Gotel in Cameroon, the Amhara Plateau of Ethiopia or the Ruwenzori Range of central Africa. The Ruwenzori – the 'Mountains of the Moon' – are typical: sharply contoured slopes making a crater-like landscape covered in moss between ice-capped peaks.

Plant growth is sparse, trees stunted and distorted, their branches low and difficult to walk beneath. Nights are cold and day temperatures high with lots of mist and long periods of cloud cover. Survival is difficult in this terrain. Leave it and make your way down the mountainside to the tropical rain forest.

Saltwater swamps

Where coastal areas are subject to tidal flooding, mangrove trees thrive. They can reach heights of 12m (40ft) and their tangled roots are an obstacle both above and below the waterline. Visibility is poor and passage difficult – it may take 12 hours to cover 900m (3000ft). Sometimes channels are wide enough to raft, but generally progress is on foot.

There are mangrove swamps in West Africa, Madagascar, Malaysia and the Pacific Islands, Central and South America and at the mouth of the Ganges. The swamps at the mouths of the Orinoco, Amazon and rivers of Guyana consist of stinking mud and trees which offer little shade. Tides can rise as much as 12m (40ft).

Everything in mangrove swamps seems hostile, from water leeches and insects, to cayman and crocodiles. Avoid them if you can. If forced there by mishap look for a way out. Where there are river channels intersecting the swamp you may be able to make a raft.

You won't starve among the mangroves. There is plenty of fish and vegetation. At low water crabs, molluscs, catfish and mudfish can be found. Arboreal and aquatic animals include water opposum, otter, tapir, armadillo and, on firmer ground, peccaries.

Inland of the mangroves, nipa palm swamp is common – all the palm's growing points are edible.

If forced to stay in a swamp, determine the high-tide level, by the line of salt and debris on the trees, and fit up a raised bed above it. Cover yourself for protection against ants and mosquitoes.

In any swamp a fire will have to be built on a platform. Use standing deadwood for fuel. Decay is rapid in a swamp so choose wood that is not far decayed.

Freshwater swamps

Found in low-lying inland areas, their mass of thorny undergrowth, reeds, grasses and occasional short palms makes going difficult and reduces visibility to only a few metres – but wildlife abounds and survival is easy. A freshwater swamp is not such a bad place once you get used to it. It will often be dotted with islands and you are not chest-deep in the water ALL the time. There are often navigable channels and raw materials available from which to build a raft.

SHELTER

There are ample materials for building shelter in most tropical regions (see *Shelter* in *Camp Craft*). Where temperatures are very high and shelters directly exposed to the sun, make roofs in two layers with an airspace in between to aid cooling. Much of the heat will dissipate on striking the upper layer, and with the air passing between this lowers the temperature of the layer beneath. The distance between should be 20–30cm (8–12in). Double layers of even permeable cloth will help keep out rain if well angled (see *Camp Craft*).

FIRE

Everything is likely to be damp. Take standing dead wood and shave off the outside. Use that to start your fire. Dry bamboo makes excellent tinder (store some), so does a termite's nest.

FOOD

A large variety of fruits, roots and leaves are available. Banana, papaya, mango and figs are easily recognized. (Papaya is one of the few plants with white sap that is edible.) The large, thorny fruit of the durian, of southeast Asia, smells disgusting, but is good to eat.

Palms provide an edible growing point and manioc produces massive tubers – though they must be cooked before eating. Taro, wild potato and some kinds of yam must also be prepared to remove poisons before they are eaten. You may find the wealth of tropical foods hard to identify, if you're not sure use the tests described in *Food*, before you risk eating them.

Animal foods

Deer, pigs, monkeys and a wide range of animals can be hunted and trapped according to location (see *Traps and trapping* in *Food*) .

In primary jungle, birds spend most of their time in the tree canopy among the fruit and berries. Place traps in clearings and lure birds with fruit. Some, such as the Asian Hornbill, also feed on lizards and snakes. Near rivers, traps can be baited with fish or offal for Fish Eagles and similar species which patrol rivers for prey.

Parrots and their relatives abound in the tropics – their mad screeching makes their presence known from early morning. They are cunning – get them used to taking bait before setting the trap.

Snakes are easier to catch – go for the non-poisonous constrictors – and very tasty. Catch them using a forked stick (see *Hunting* in *Food*).

Food from rivers

Rivers support all kinds of life: fish, plants, animals and insects. If you have no fishing tackle small pools can be dammed and then emptied with a bailer – fish and turtles in surprising numbers can be found in the mud. Try constructing traps or crushing certain roots and vines to stupefy the fish (see *Fishing* in *Food*).

Fish are easily digested and have good protein content. Many jungle people depend on them for nourishment, but in the tropics they spoil quickly. Clean thoroughly, discard entrails and eat as soon as possible, do not preserve them by smoking or drying. Fish from slow moving water are more likely to be infested with parasites. If suspect, boil for 20 minutes. In areas where locals use the water as their sanitation system, fish may carry tapeworms and other human parasites and the water itself could be infected with amoebas which cause dysentery. Always boil water.

Rivers can bring dangers too. Piranha may be found in the Amazon, Orinoco and Paraguay river systems of South America. A similar fish is found in Burma. Electric eels are slow-moving and not aggressive, but they can grow very large and discharge 500 volts or more. Stingrays also occur in some tropical South American and West African rivers. Look out for crocodiles or alligators and water snakes and take care in handling catfish, which have sharp dorsal fins and spines on their gill covers, the Electric catfish can also deliver a powerful shock.

DANGERS

Cover your feet

Good footwear and protection for the legs is essential – they are most exposed to leeches, chigoe, and centipedes. Wrap bark or cloth around the legs and tie it to make puttees.

Insect attack!

Slashing your way through jungle you may disturb bee, wasp or hornet nests. They may attack, especially hornets, whose stings can be especially painful. Anywhere left bare, including your face, is vulnerable to attack. Run! Sunglasses would help protect the eyes.

Perspiration is a problem, insects desperate for salt will fly to the wettest parts of your body. However, they also sting. Protect armpits and groin.

Beware invaders

Keep clothing and footwear off the ground, then scorpions, snakes and other nasties are less likely to invade them. Always shake out clothing and check boots before putting them on and be wary when putting hands in pockets. On waking, take care. Centipedes tend to curl for warmth in some of the more private body regions.

Beware caterpillars too!

If mosquitoes and leeches sucking your blood, painful bites from centipedes and the risk or scorpion and snake bites are not enough (see *Bites* in *First Aid*) look out for hairy caterpillars. Be careful to brush them off in the direction they are travelling or small irritant hairs may stay in your skin and cause an itchy rash, which may fester in the heat.

Mosquito protection

Wear a net over your head, or tie a tee-shirt or singlet over it, especially at dawn and dusk. Better, take a strip of cloth long enough to tie around your head and about 45cm (18in) deep and cut it to make a fringe of vertical strips hanging from a band that will hang around your face and over your neck and impregnant your clothes and bags etc with repellent.

At night keep covered, including your hands. Use bamboo or a sapling to support a little tent of clothing plus large leaves, rigged over your upper half. Oil, fat or even mud spread on hands and face may help to repel mosquitoes. In camp a smoky fire will help keep insects at bay. If you are bitten make sure you don't scratch as this may let infection in.

Leeches

Leeches lie on the ground or on vegetation, especially in damp places, waiting to attach themselves to an animal (or person) to take a meal of blood. Their bite is not painful but they secrete a natural anti-coagulant that makes it messy. Left alone, they drop off when they have had their fill – but if you are covered in them you must do something! Do NOT pull them off. There is a risk the head will come off leaving the jaws in the bite, which could turn septic. Remove with a dab of salt, citric acid from fruit, alcohol, an ember or a flame.

Beware the candiru!

This minute Amazonian catfish, about 2.5cm (1in) long, very slender and almost transparent, sucks blood from the gills of other fish. It is reported to be able to swim up the urethra of a person urinating in the water – where it gets stuck by its dorsal spine. The chance of this happening is remote but the consequences could be dire! Cover your genitals and don't urinate in the water.

VEHICLES

For desert travel, fit long-range fuel tanks and make provision for storing drinking water. Carry further supplies of both in jerry cans. A jack is no use in soft sand and an air bag should be carried which is inflated by the exhaust. Extra filters will be needed in the fuel line and air intake. Sand tyres must be fitted and sand channels carried to get you moving again when bogged down in loose sand.

For high altitudes the carburettor needs adjustment. In scrub country, thorn gaiters will reduce puncture risks. Antifreeze and suitable wheels and chains are needed for snow and ice. The engine will need special tuning to match climatic conditions and its own spares. A spare wheel and a good tool kit are obvious requirements.

HOT CLIMATES

Even when you have had modifications made to prepare your vehicle for hot climate conditions you may still find that you have problems.

In crossing some deserts the considerable temperature change from day to night can put a strain on any metal and increase the risk of leaks.

WARNING: *NEVER leave a sleeping or injured person or any animal in a closed car in a hot climate – or even on a sunny day in temperate regions. Always leave windows open to ensure ventilation (heat exhaustion can be lethal) even if parked in the shade, as the sun will move.*

Overheating: *Stop and allow the engine to cool. If you are driving a particularly tricky stretch and stopping is out of the question, switch on the heater. This will give greater volume to the cooling water and, although the inside of the car will get even hotter, the engine will cool. When convenient stop and open up the bonnet. Do not undo the radiator cap until the temperature drops. Check the radiator and all hoses for leaks. If the radiator is leaking, adding the white of an egg will seal small holes. If there is a large hole, squeeze that section of the copper piping flat to seal it off. It will reduce the size of the cooling area but, if you drive very steadily, you will be able to keep going.*

Metal gets hot: *Be careful! All metal parts of a car can become hot enough to cause blisters.*

Care in sandy conditions: *When adding fuel, sand and dust can get into the tank. Rig a filter over or just inside the inlet to the tank.*

COLD CLIMATES

Low temperatures not only make driving conditions difficult. They can make starting and maintenance difficult and hazardous.

Starting: *Always try to park on a gradient so that you can use a bump start to back up the starter. Once you get the engine going keep it running – but check that the handbrake is firmly on and never leave children or animals in an unattended vehicle with the engine running.*

Demisting: *Don't try to drive looking through a small clear patch on a misty screen. Onion or raw potato rubbed on the inside of the screen will stop it misting up.*

Cover the outside of the windscreen and windows with newspaper to prevent frost building up on them. If damp, however, paper will stick.

Coddle the engine: *Wrapping a blanket around the engine may help to stop it from freezing up – but remember to remove it before you start the engine. Cover lower part of the radiator with cardboard or wood so that it does not freeze as you go along. If very cold, leave covered. Otherwise remove to prevent overheating.*

Cover metal: *Don't touch ANY metal with bare hands. Your fingers could freeze to it and tear off skin. Where handling metal components with gloves is awkward, wrap fingers with adhesive tape. Treat radiator cap and dip stick in this way to ease your daily checks.*

Diesel engines: *Diesel contains water and freezes solid at low temperatures. Always cover front of engine, but check for overheating. Always wrap engine at night or when left standing. Some lorry drivers light small fires under frozen tanks. Only you can judge if the risk is worth taking.*

CUT OFF IN SNOW

If you are trapped by a blizzard, stay in the car. If you are on a regular traffic route you will probably soon be rescued. Going for help could be too risky.

Run the engine for heat if you have fuel. Cover the engine so that as little heat as possible is lost directly – but make sure that the exhaust is clear. Take no risk of exhaust coming into the car. If you feel drowsy stop the engine and open a window. Do NOT go to sleep with the engine running.

Switch off the heater as soon as you have taken the chill off the interior. Start it again when the temperature drops. If there is no fuel to run the engine wrap up in any spare clothing, rugs etc, and keep moving inside the car.

If you have to leave the car to go a short distance, when you know help is very close for instance, rig up a signpost – a bright scarf or garment on a stick to help you find it again.

When the blizzard stops, and if it is daylight (otherwise wait until morning), it is worth walking out if there is a clear guide to the route (such as telegraph poles).

If miles from anywhere and off normal routes, and if the snow is building up to bury the car, it is worth getting out and building yourself a snow cave – where you may be warmer than in the car and can sit out several days. When the blizzard stops scrape large signs in the snow and use other signals to attract attention.

GENERAL

Clutch slip: *Often caused by oil or grease getting on the clutch plates. To degrease these plates use the fire extinguisher. Squirt it through the inspection plate opening.*

Fan belts: *Improvise one with a pair of tights, a tie or even string.*

HT leads: *If a high tension lead breaks, you may be able to replace it with a willow twig. Any plant stem with water content can be used to carry current from the coil to the distributor. Spit on the ends and insert into the push-fit contacts. DANGER! When you switch on, there is a current of about 1300 volts. DO NOT TOUCH. Replace twig frequently as it dries out.*

Dead battery: *You don't need the battery if you can get up enough speed. A tow or a steep slope will do. On a vehicle with four gears, use gear two or three when releasing the clutch to try for ignition.*

Half-shaft breakage: *Not much you can do on a front or back wheel drive but, if your vehicle is 4-wheel drive, remove the half-shaft. Disconnect the drive and keep motoring on the other axle.*

4

FOOD

You need some understanding of your body's nutritional needs and how to meet them. In most circumstances plants will be the most readily available – but you need to know which plants to avoid. Colour illustrations provide a miniature field guide to some of the most useful plants.

Almost any animal can provide food and you must get used to eating unusual ones such as worms, insects, and even piranha fish. Despite their reputation they won't attack unless they have no other food. If the river is low, steer clear. A wide range of effective snares and traps will catch food, while you gather plants, collect water or carry out other survival tasks.

To hunt you may first have to improvise your own weapons. Learn how to do so and practise using them.

Meat needs preparation to make it convenient to handle and safe to cook. Learn how to preserve it when there is a surplus.

Fish offers another source of food, so simple methods are described which require no prior angling skills.

FOOD AND FOOD VALUES 110
Energy Needs 110
FOOD PLANTS 112
Testing new plants 113
Gathering plants 114
Plants to avoid 116
Identifying plants 117
ANIMALS FOR FOOD 118
Finding game 118
Mammals 122
Reptiles 129
Birds 130
Insects 131
Snails and worms 134
Dangers 135
EDIBLE PLANTS 136
ROOTS AND TUBERS 144
FRUIT 148
NUTS 152
POISONOUS PLANTS 154
TREES 157
Poisonous trees 159
FUNGI 160
Identifying amanitas 161
EDIBLE FUNGI 162
POISONOUS FUNGI 166
ARCTIC AND NORTHERN PLANTS 168
DESERT PLANTS 170
TROPICAL PLANTS 172
POISONOUS TROPICAL PLANTS 182
SEASHORE PLANTS 184
SEAWEEDS AND ALGAE 186
TRAPS AND TRAPPING 188
Snares 190
Deadfall traps 197
Spear traps 201
Bird traps 204
Survival case study 207
HUNTING 208
Weapons 209
Hunting birds 215
Animal dangers 216
HANDLING THE KILL 218
Preparing the kill 219
Offal 224
FISH AND FISHING 228
Angling 229
Fish traps 234
Other techniques 237
Fish narcotics 239
Arctic fishing 241
Preparing fish 243

FOOD AND FOOD VALUES

The body needs food to supply heat and energy and to provide the materials with which it can build new tissues, whether for growth, repair or reproduction. A healthy body can survive for a time on reserves stored in its tissues, but lack of food makes it increasingly difficult to keep warm, to recover after hard work or injury and to fight off disease.

Fortunately human beings are omnivores – we can digest both flesh and plants – and we can eat almost anything from the animal and vegetable kingdoms. With a little care, those things that are poisonous or dangerous are easily avoided. We enjoy food, so a good meal is an excellent morale booster, and when there are not other pressing priorities it is worth taking the trouble to cook food, which will make it more tasty, and to seek out foods with interesting flavours.

Do not rely on the easiest source of food, for a balanced diet is as important to the long-term survivor as having enough to eat in the first place. If you are camped in the middle of a rabbit warren, and dinner almost jumps into the pot, you could die from deficiencies not supplied by rabbit meat. Your diet MUST be made up of a wide range of elements which provide the right proportions of nutrients and sufficient energy to get you through the day. These nutrients must include proteins, carbohydrates, fats, minerals and other trace elements and vitamins.

ENERGY NEEDS

Without making a physical effort of any kind, the average person in a completely restful state requires 70 calories per hour to maintain their basic metabolism – the involuntary functions such as breathing and blood circulation that we do not even have to think about. A calorie is a unit of heat – it is the amount needed to raise the temperature of one litre of water by one degree centigrade – and is the way in which energy is expressed when discussing nutrition.

Calories are not produced equally by all kinds of foods. The energy values of the basic types are:

Carbohydrates	1g (.035oz) produces 4 calories
Fat	1g (.035oz) produces 9 calories
Protein	1g (.035oz) produces 4 calories

The simplest domestic activities – standing up, sitting down, lighting a fire and so forth – that make up an ordinary day demand another 45 calories per hour. That makes a total of about 2040 calories a day without any work or other major activity, which could burn up a further 3,500 calories daily. Since not only physical effort but mental effort and anxiety also use up calories, keep calm and relax and, if food is scarce, DO NOT SQUANDER ENERGY.

Carbohydrates

Carbohydrates form the bulk of the diet and are the main source of energy, not just for physical effort but for fuelling internal functions and the running of the nervous system. Carbohydrates are made up of carbon, hydrogen and oxygen and are synthesized by plants. They are very easily converted into energy by the body and do not require a large water intake. They prevent ketosis – indigestion, vomiting and nausea caused by the excessive breakdown of body fats during starvation – but have two disadvantages: they do not contain vitamin B and they may cause constipation.

There are two types of carbohydrates:

Sugars are found in sugar, syrup, honey, treacle and fruits.
Starches are found in cereals, roots and tubers. Starch granules are insoluble in cold water but heat causes them to rupture – this is why roots and tubers are always cooked.

Fats

Fats contain the same elements as carbohydrates but combined differently. They also are a concentrated source of energy, providing twice as many calories as carbohydrates, stored in the body as a layer of fat under the skin and around the organs. They are insoluble in water and before they can be absorbed by the body require a lengthy digestive process which demands an adequate intake of water. Fats heat and insulate the body, protect organs, lubricate the alimentary tract and build an energy reserve. They are found in animals, fish, eggs, milk, nuts and some vegetables and fungi.

Proteins

Proteins are the basic chemical units of living matter. They are the only food constituent containing nitrogen and therefore essential for the growth and repair of the body, and are made up of complex chemical structures known as amino acids.

The main sources of protein are meat, fish, eggs and dairy produce, and plants in the form of nuts, grains and pulses. Worms contain the highest amount of proteins, having the eight essential amino acids humans require. Do not squeeze the goodness out of them; isolate them in a container so that they will dehydrate. This will ensure that they don't lose their health properties. Fungi can be an important source of protein.

If carbohydrates and fats are missing from the diet, protein is used to generate energy but at the expense of the body's other needs, so that in starvation the body burns up its own tissues.

Minerals

Minerals include some which one requires in quantity, such as calcium, phosphorus, sodium, chlorine, potassium, sulphur and magnesium, and others, including iron, flourine and iodine, which are required in much smaller amounts. Calcium is needed for bones and teeth but has other roles in muscular function and blood clotting. All minerals have vital roles in body functions.

Trace elements

Trace elements include strontium, aluminium, arsenic, gold and other chemicals in tiny amounts. Their exact function is not yet understood.

Vitamins

Vitamins are essential to health and have an important role not only in maintaining the body but in protection from illness. There are about 40 different vitamins, of which about a dozen are essential for humans, found in minute amounts in many kinds of food. Vitamin D can be synthesized in the skin when it is exposed to the sun's rays and vitamin K can be synthesized by bacteria in the gut, but others must be obtained from external sources. The average person naturally carries a 28-day supply of vitamins within their body and of those, Vitamin C is first that will need to be replenished. Many herbs are rich in Vitamin C. The second vitamin to go is Vitamin A. This vitamin aids vision and prevents eye disease. To counter this, eat the shoots and leaves that you see rabbits eating.

FOOD PLANTS

There are few places in the world where you will be far from some kind of vegetation – bush, vine, creeper, flower, grass or lichen – which can be eaten to provide nourishment. In Europe alone there are 10,000 edible wild plants. The only skill required in making use of them is knowing which is which and where to find them. Some, though edible, have very little food value, so learn which yield the most nourishment, especially those which are widespread and available throughout the year – and learn which are poisonous, in order to avoid them.

Plants contain essential vitamins and minerals, and are rich in protein and carbohydrates. Some plants also contain fat and all provide roughage essential to keep the body in good working order.

Do NOT eat large quantities of any one plant at a time and if you are not used to eating a plant start by nibbling a fresh specimen and adding some to stews, and build up your intake gradually. If you give your stomach a chance to get used to a new food it will adjust and not reject it.

Do NOT assume that because birds, mammals or insects have eaten a plant that it is edible by humans. Monkeys are some indication, but no guarantee, that plants are suitable for human consumption.

TESTING NEW PLANTS

Always adopt the following procedure when trying out potential new food plants, only one person testing each plant. NEVER take short cuts – complete the whole test. If in any doubt, do NOT eat the plant. Should stomach trouble occur, relief can be gained by drinking plenty of hot water; do not eat again until the pain goes. If it is severe, induce vomiting by tickling the back of the throat. Charcoal is a useful emetic. Swallowing some will induce vomiting and the charcoal may also absorb the poison. White wood ash mixed to a paste with water will relieve stomach pain.

Inspect

Try to identify. Ensure that a plant is not slimy or worm-eaten. It will be past its best, with little food value other than the grubs or worms upon it. Some plants, when old, change their chemical content and become toxic.

Smell

Crush a small portion. If it smells of bitter almonds or peaches – DISCARD.

Skin Irritation

Rub slightly or squeeze some of the juice onto a tender part of the body (under the arm between armpit and elbow, for instance). If any discomfort, rash or swelling is experienced – DISCARD, reject in future.

Lips, Tongue, Mouth

If there is no irritation to the skin proceed in the following stages, going on to the next only after waiting five seconds to check that there is no unpleasant reaction:

- Place a small portion on the lips
- Place a small portion in the corner of the mouth
- Place a small portion on the tip of the tongue
- Place a small portion under the tongue
- Chew a small portion

In all cases: if any discomfort is felt, such as soreness to the throat, irritation or stinging or burning sensations – DISCARD, reject in future.

Swallow

Swallow a small amount and WAIT FIVE HOURS. During this period eat or drink NOTHING else.

Eating

If no reactions such as soreness to the mouth, repeated belching, nausea, sickness, stomach pains, griping pains in the lower abdomen or any other distressing symptoms are experienced, you may consider the plant safe.

GATHERING PLANTS

It is easy to pick plants here and there, but it is better and safer to gather them systematically.

Take a container on foraging trips – an empty bag, a piece of cloth folded into a sack, a birch-bark box or large leaves stitched together. This stops harvest being crushed, which makes it go off quickly.

Leaves and Stems

Young growth, usually paler green, will be tastier and more tender. Older plants are tougher and more bitter. Nip off leaves near the stem. Leaves simply torn off are easily damaged. They may wilt and loose goodness before they reach the pot.

Roots and Tubers

Choose larger plants. Some are very difficult to pull up. To lessen chance of breaking them dig around the plant to loosen, then prise them out with a sharpened stick.

Fruit and Nuts

Choose larger plants. Pick only ripe, fully coloured fruits. Hard, greenish berries are indigestible, even after long cooking. Many fruits, especially in the tropics, have tough, bitter skins. Peel them. Nuts lying at the base of a tree are a sign they are ready. Others can be shaken down if the tree is a small one. You may be able to knock others down by throwing a stick.

Seeds and Grains

PRECAUTION: Some contain deadly poisons. Tasting will not harm you but DO NOT SWALLOW. Carry out the edibility test, reject any seed that is unpalatable, bitter or with a hot, burning taste (unless a positively identified pepper or spice).

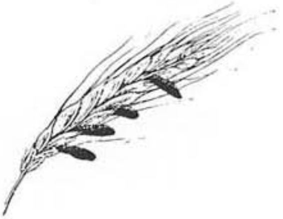

Heads of some grain plants may have black spurs in place of normal seeds. These carry ergot poisoning, a fungal disease that turns the grain into enlarged, black, bean-like structures. It is a source of an hallucinogen, and is very poisonous – sometimes fatal.

REJECT THE WHOLE HEAD!

Fungi

Medium-sized are easier to identify and less likely to suffer from insect damage. Pick the whole fungus. If the stem is left on the ground it will be harder to identify the fungi gathered. Keep fungi separate. If a poisonous kind has been picked it will not then contaminate other food.

PLANTS TO AVOID

WARNING

POISON! There are two fairly common poisons in the plant world, but both are easily detectable:

HYDROCYANIC ACID (Prussic acid) has the taste and smell of bitter almonds or peaches. Most notable example is the Cherry Laurel (*Prunus laurocerasus*), with laurel-like leaves, which contains a closely allied poison. Crush the leaves and remember the smell. Discard ALL plants with this smell.

OXALIC ACID, whose salts (oxalates) occur naturally in some plants, for instance Wild Rhubarb (mostly in the leaves) and Wood Sorrel (*Oxalis acetosella*). Recognizable by the sharp, dry, stinging or burning sensation when applied to the skin or tongue. Discard ALL plants which fit this description.

- **AVOID** any plant with a milky sap, unless positively identified as safe (such as dandelion).
- **AVOID** red plants, unless positively identified, especially in the tropics. The red-streaked stalk of Wild Rhubarb is edible but its leaf is poisonous. Hemlock has reddish-purple splotches on its stem.
- **AVOID** fruit which is divided into five segments, unless positively identified as a safe species.
- **AVOID** grasses and other plants with tiny barbs on their stems and leaves. With a magnifying glass you can see them as hooks rather than straight hairs and they will irritate the mouth and digestive tract.
- **AVOID** old or wilted leaves. The leaves of some trees and plants develop deadly hydrocyanic acid when they wilt – including blackberry, raspberry, cherry, peach and plum. All may be safely eaten when young, fresh and dry.
- **AVOID** mature bracken (*Pteridium aguilinium*). It destroys vitamin B in the body, setting up a peculiar blood condition which can cause death. Eat only tightly coiled 'fiddle heads'. All 250 varieties of north temperate ferns are edible when young, although some are too bitter to be palatable and some have irritating hairs which must be removed before eating. Break off the tips as low as they remain tender, close your hand over the stalk and draw the frond through to remove the 'wool'.

IDENTIFYING PLANTS

Only a small selection of the world's many plants can be described and illustrated here and only a specialist botanist could identify more than a handful of plants in far-flung corners of the world. Begin by learning a few plants that can be found in most conditions and at most times of year. Close familiarity with even one or two could make all the difference between survival and starvation. Learn these first and learn them thoroughly.

Temperate zones: dandelions, nettles, docks, plantains

Sub-tropical and tropical zones: palms, wild figs, bamboo

Arid and desert zones: mescal, prickly pears, baobabs, acacias (but not in North or South America)

Polar zones: spruces and willows (north), lichens (north and south). In summer in north as for temperate zones

Coasts: kelps and lavers

IDENTIFICATION AIDS:

- Location: Plants grow only in suitable conditions, if you know what habitats they like, and geographical distribution, you can reduce the possibilities immediately.
- Shape and size: Is plant tall and woody like a tree or shrub? short and soft-stemmed? bushy and branched or only one or a few stems?
- Leaves: Are they large or small? spear-shaped, rounded or strap-like? with toothed, or lobed edges? made up of several leaflets? uniform in colour?
- Flowers: Seasonal, but if present note colour, size, shape, single or clustered, where on plant.
- Fruits and seeds: Are they fleshy, hard and cased like a nut, or small and tough like a seed? Note colour, size, shape, whether singly or in clusters, in pods or capsules.
- Roots: Rarely help identification – unless unusual.

ANIMALS FOR FOOD

All animals can be a source of nourishment. A few, including worms and insects, can be collected with little skill, but most must be trapped or hunted, demanding both knowledge and expertise of animals and of methods.

The more you know of animals the better, but general natural history knowledge must be used to help you find out more by observation in the wild. There is no one way to do things – you must learn by trial and error.

You must study each species' habits, find out where it sleeps, what it eats and where it waters. You must learn how best to make a kill, what traps to set and balance your humanitarian instincts against the expediencies of survival.

The best animals for flavour and amount of meat are mature females. The younger the animal, the more lean the meat. An adult male is at his fattest just before the mating season (which varies according to species and climate). During the mating season the male becomes progressively poorer, the fat is run off and even the normally rich bone marrow suffers. Animals put on fat to see them through the winter and trim down for the summer. The older the animal the more fat it has and the tougher the meat becomes.

FINDING GAME

There are very few places on earth where there are no wild animals, but sometimes the signs of their presence are far from obvious. If you can recognize and read the signs that animals leave, and identify the animal, you will know what methods to adopt in hunting and trapping your prey, what bait to use and what kind of traps to set.

Most mammals are mainly on the move at first and last light. Only the larger and more powerful venture out during the day. Larger herbivores need a whole day of grazing to satisfy their appetites. Some very small ones have to eat so frequently that they are intermittently active all day long – but most smaller mammals, such as rabbits, eat mainly at night and only change their habits when the weather is bad. Animals that feed on other animals hunt at the times that their prey is active. So must you, if you intend to hunt them, but you can find out a great deal without ever seeing a living animal and can set traps to catch them when they are about.

Tracks and signs

Most animals are creatures of habit and use regular routes between their watering spots, feeding places and homes. Look out for the signs of these trails. Tracks will be more obvious on wet ground, snow and damp sand, and other signs are more noticeable in heavy vegetation. The size of the impression left is in proportion to that of the animal. The age of the track can be accurately judged by its sharpness and moisture content. Has water seeped into it or has rain filled it? Has it become smudged? The clearer the track the more likely it is to be recent.

In the early morning tracks can be checked by looking at them from ground level. If dew and spider's webs have been disturbed, the tracks are, at most, a few hours old. Some animals, such as rabbits, never range very far and any tracks are likely to indicate that they are in the area. Some animals make tunnels through dense undergrowth. Their height indicates the size of the animal that made them. The height of broken twigs along a track will also suggest an animal's size. Check to see how fresh they are: Have trampled leaves wilted? Are broken twigs still green and supple? Marks on trees and logs, feeding signs and discarded food are other useful indications. Droppings indicate the type of animal that left them.

Feeding signs

The way in which bark has been stripped from trees, the gnawed shells of nuts, partially eaten fruits, bitten off shoots and the remains of prey animals of carnivores or the destruction of the nests of prey are all indications of species living or hunting in the vicinity.

Discarded fruits or nuts are often found when food is plentiful – an animal finds one piece not to its liking and drops it to try another. They not only reveal an animal's presence but suggest baits for traps.

A skilled eye can often identify the species by the pattern left by tooth or beak marks on a nut, or the way in which a pine cone has been stripped to get at its seeds.

Bark, twigs and buds, especially of young trees and bushes, form an important part of the diet of many animals, including several species of deer and goats, hares, squirrels and numerous other small rodents. In most cases the marks made by the animal's teeth will show clearly in the bark.

Many deer will bite offshoots leaving a torn and frayed edge. (Hares for instance, leave a clean bite.) Bark pulled off in long shreds, completely exposing the wood is another sign of deer, when eating in summer – in winter the bark is attached more firmly and is eaten in patches, so that usually only sections on one side of the trunk are affected and large toothmarks are clearly visible. Deer also scrape their antlers against trees to remove the velvet and as part of marking their territory – leaving frayed bark and wood with long scratches from the antler points.

Sheep and goats also bark trees. Their toothmarks generally run obliquely, those of deer vertically. Low level gnawing will usually be that of rodents – stripped roots are probably the work of vole-like animals. Severed stumps with conical tops that look as though they have been chopped with a small axe are the felling work of beaver.

Squirrels strip bark higher up the tree, pieces often falling to the ground beneath. A scattering of cone pieces on the ground is often a sign of squirrels too. Beneath a tree, nutshells may also indicate a squirrel – perhaps its nest is overhead, but if nuts or cones are also wedged into the trunk that shows the work of nut-eating birds. Near a pile of empty shells you may also find a rodent's burrow.

If sapling growth looks as though it has been trimmed level like a clipped hedge, or the lower branches of trees look neatly trimmed below a certain height, you may guess this to be the work of browsing animals such as deer.

Droppings

Droppings give one of the best indications of an animal's type. Size can be judged from their size and quantity; dryness is an indication of how long since they were passed. Old droppings will be hard and odourless – fresh, wet and still smelling. Flies draw attention to droppings.

Mammals: Many mammal droppings have a strong scent, produced by glands in the anus. This plays an important role in marking territory and giving sexual signs. It can be used to advantage in baiting snares.

Animals that live on vegetation, such as cattle, deer and rabbits, produce roundish and strawy droppings. Meat eaters, such as wild cats and foxes, produce long tapering ones. Some animals, including badgers and bears, have mixed diets. Break open a dropping to see if there are any clues to what the animal has been eating, then bait accordingly.

Birds: Also fall into two groups: flesh eaters and seed/fruit eaters, which can be distinguished by their droppings. Smaller seed-eating birds' droppings are small and mostly liquid, whereas the owls and hawks produce pellets which may contain indigestible parts of the meal taken, be it fish, bird, insect vole or rodent. Loose droppings are an indication that water is probably within a reasonable range, for small birds need to be close to it. Birds of prey, however, do not need to be close to water. Roosts and nesting sites will often be indicated by copious droppings on the boughs or ground beneath. Birds generally feed in older trees where there is some decay and plenty of grubs.

Rootings

Some animals root up the ground in search of insects and tubers. Pigs, especially, turn over large areas of earth. If the earth is still crumbly and fresh an animal is likely to have been active on the spot quite recently. A big muddy wallow is usually a sign of pigs. Small scratches may be where a squirrel has been digging for shoots.

Scent and Smell

Listen to the noises around you, register the smells. They are certain to include indications of the wildlife present, and where one kind of animal exists there will be others – where there are prey species there will be predators. Many people neglect their sense of smell – but you must try to redevelop it. Some animal smells are very strong, particularly those of foxes. Keep your eyes sharp too. In cold climates, for example, the breath of large animals forms a cloud of condensation. This 'smoke' can be seen some way off if you are in a good vantage point.

Burrows and Dens

Many animals make their homes in burrows, usually on high ground away from water. Some, such as rabbits and ground squirrels, take little trouble to conceal them, although one or two exits will be hidden for use in an emergency. Rabbits' emergency holes are easily dug out, or a piece of bramble or barbed wire can be pushed down the hole to hook the rabbit out.

Predatory animals normally hide their holes, which are generally in wooded country. Tracks or droppings nearby may give their location away – and are an indication that a hole is in use (although some animals, such as badgers, use regular toilet sites elsewhere).

MAMMALS

NOTE: Where tracks are shown, no scale has been imposed on them. Most are typical of a whole family of animals, but will vary greatly in size according to species. Track 1 is right front. Track 2 is right hind.

WILDCATS

Range from domestic size to the tiger. Occur on all continents except Australia and Antarctica, but nowhere common. Secretive and generally nocturnal, they avoid man. You will see lions in a wildlife reserve but seldom encounter cats elsewhere. All are potential food – but don't take on the larger ones. Kills of big cats may be scavenged if unattended – but BEWARE – the owner is likely to be nearby. Meat may be stringy, stew thoroughly, but small cat meat tastes like rabbit. Sinews strong, good for bowstrings.

Traps: *Powerful spring snares, platforms or baited-hole-noose. Bait with offal, blood or meat. Cats have very fast reactions and may leap clear of deadfalls.*

Tracks and signs: *Walk on toes, leaving marks of four well-developed pads and a larger pad to rear. Claws (except in Cheetah) retracted when walking. Droppings elongated, tapering, but usually hidden. Urine strong smelling.*

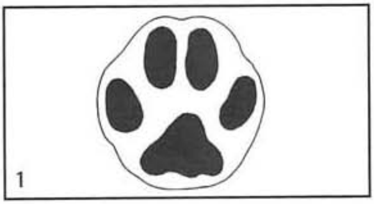

WILD DOGS

Foxes and other wild dogs are found widely, from deserts to the Arctic but, not in New Zealand, Madagascar and some other islands. Wolves are now mostly confined to wilderness North America and north and central Asia. Superb senses make it pointless to stalk canines at close quarters but their curiosity will tempt them to traps. Moving on all fours may attract them – they may think YOU are potential food! Chewy. Remove anal glands. Dog is a delicacy in the Far East.

Traps: *Snare for foxes, try stepped-bait or toggle, bait-release, baited-hole-noose. Minimize human scent in area.*

Tracks and signs: *Walk on toes. Print shows four pads and claw tips – outer pad shorter than inner with large main pad to rear. Elongated, tapering droppings show remains of fur, bones, insects depending on diet. Fox droppings pungent, as is earth (den) – in soft ground earth can be dug out.*

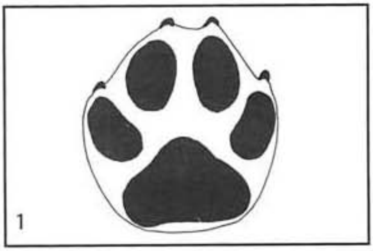

Hyenas: *Scavengers (not true dogs) of Africa east to India. Can be very dangerous. Boil thoroughly, risk of parasites in flesh.*

Civets: *Scavengers of tropical Africa and Asia, trap as dogs and remove anal glands before cooking.*

BEARS

Solitary animals of North America, Eurasia and some northern parts of South America, preferring well-wooded country (except Polar Bear). Strong and fast-running they may scavenge a campsite. Most kinds can *climb trees. Bears can kill a man with ease. Give them a WIDE berth. Potential rich food with nutritious fat, if you can trap one, but hunting with improvised weapons is foolhardy. Cook thoroughly to eliminate parasites. Do NOT eat liver of Polar Bear – it contains lethal levels of vitamin A.*

Traps: *Deadfall and spear, but only largest work – bears tear off most snares. Must kill or completely incapacitate – wounded bears very dangerous.*

Tracks and signs: *Prints may be 30cm (1ft) x 18cm (7in), with five long-clawed toes. Toe pads close together, claw marks distinct. Rear paws taper and could be mistaken for human prints. Bears eat almost anything and grub up ground, rip up stumps and break into insect nests in search of food.*

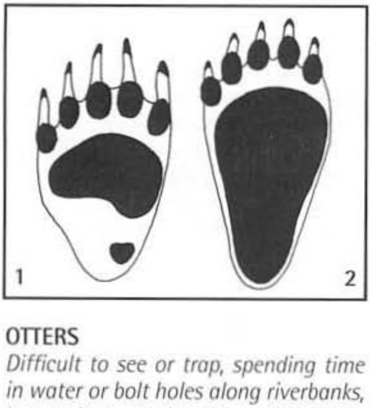

OTTERS

Difficult to see or trap, spending time in water or bolt holes along riverbanks, but curious, so riverside spring snare, baited with fresh fish may tempt them.

Tracks and signs: *Five-toed, webbed, almost circular 7·5cm (3in) x 6cm (2⁵/₈in). Fishy-smelling elongated droppings on regular sites, usually stones.*

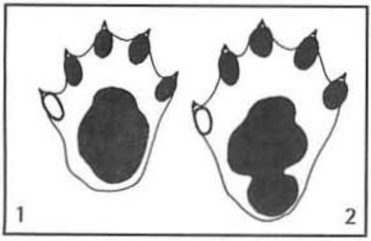

WEASEL GROUP

Weasels, stoats, mink, martens and polecats are secretive – but may be an important food source in the far north. Beware of their sharp teeth.

Traps: *Spring snares with bait bars, deadfalls. Bait with offal or birds' eggs.*

Tracks: *Indistinct except in soft ground. Five well-spaced claws and toes, hair on main pad often smudges. Gait bounding, so fore and rear prints overlap. Weasels are the smallest.*

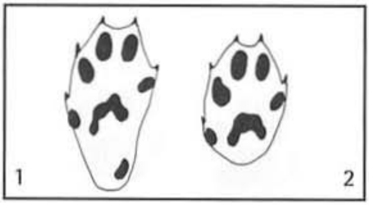

WOLVERINES (Gluttons)

Chunky, badger-shaped animals of the far north, quite capable of pulling down a caribou though mainly carrion eaters. Nowhere common. Don't take one on unless you are armed!

Traps: *Strong baited-spring snare or hole-noose – if you can find bait in the tundra.*

Tracks: *Average 8cm (3¹/₄in) x 7cm (2³/₄in), five toes with powerful claws. Hair on sole may obscure main pad.*

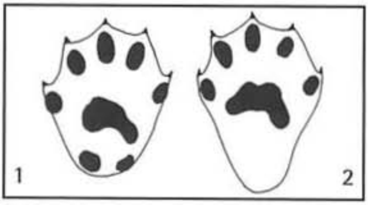

BADGERS

Stocky, nocturnal creatures of habit of North America, Africa and Eurasia east to Indonesia. Hibernate in colder areas. Well-worn runs and toilet sites make *good trapping places. Can be fierce, with a disabling bite. Do not get close. Young badger tastes like pork.*

Traps: *Leg spring snares, baited release and deadfalls, hole noose. Make them tough.*

Tracks and signs: *Five-toed with prominent claws and large rear pad. Could be confused with small bear. Stride length averages 50cm (20in). Droppings like dog's but in shallow excavated scoop. Uprooted ground and insect nests torn open may indicate badgers.*

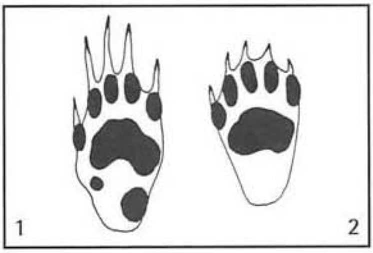

SKUNKS

North and South American, skunks look like small badgers with a hairy tail. When threatened they spray a foul-smelling fluid from glands by the anus. Look for an easier meal.

MONKEYS/APES

Almost entirely confined to tropics, generally living in extended family groups, often in trees. Even small monkeys can inflict a bad bite. Intelligent and difficult to stalk – they set up a howling long before you can get close – but bold and curious, which can be to your advantage. Very edible.

Traps: *Perch or baited spring spear trap, spring snare or hole noose. Bait with fruit or anything likely to arouse curiosity.*

Signs: *Activity will lead you to them – few take trouble to conceal themselves and most are noisy.*

SEALS

Track shows belly drag in centre. Arrow indicates direction of travel.

(See Polar Regions *in* Climate and terrain.*)*

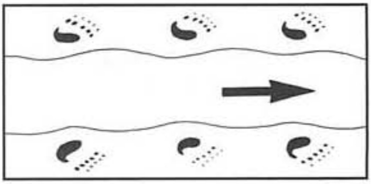

BATS

Found in all except very cold climates. Active at night. Hibernate in temperate areas. Meat-eaters take insects, fish, small animals include a blood-sucking Vampire bat of South America, which can transmit rabies. Keep WELL covered when sleeping rough in bush within its range. Fruit eaters (also known as flying foxes) are larger, some with 1m (40in) wing span, found from tropical Africa east to Australasia. Good eating, especially plump fruit bats. Remove wings and legs, gut and skin like rabbit.

Traps: *Net a fruit tree where flying foxes feed. Knock others from roosts when sleeping during day.*

Signs: *Roosting colonies easy to spot, often in caves, fruit bats also in trees.*

CATTLE

Cattle live in herds and need access to water. Bison and other wild cattle are found in wilderness areas of North America, Africa and southern Asia. The few buffalo remaining in Europe and America are protected. Introduced water buffalo are wild in northern Australia. Stray domestic cattle may occur, particularly in Africa.

Big cattle can be dangerous – particularly cunning old solitary bulls.

Traps: *Only the most powerful snares, spring traps and deadfalls, except for young animals.*

Tracks and signs: *Heavy, two distinct* *hoofmarks, narrow at top, bulbous at rear. Droppings rather like familiar cowpats – they make excellent fuel.*

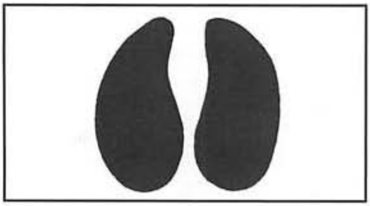

WILD SHEEP/GOATS

Sheep generally live in small flocks on high alpine pastures and steppes in North America, parts of Europe and north Africa (rare) through Arabia to mountains of central Asia. Alert, nimble and liking inaccessible places – making them impossible to approach.

Goats are few in Europe and North Africa, and mainly found in mountains of central Asia. Even more sure-footed than sheep. Good to eat.

Traps: *Snares or spring snares on trails. In rocky areas natural obstructions give good places for deadfalls. But don't expect great success!*

Tracks and signs: *Cloven hooves, two slender pointed marks not joined, tip always splayed in sheep, sometimes in goats. Goats' more rounded at front which is narrower than rear. Illustration below shows, in scale, Domestic sheep (left), Chamois (right). Globular droppings like those of domestic sheep.*

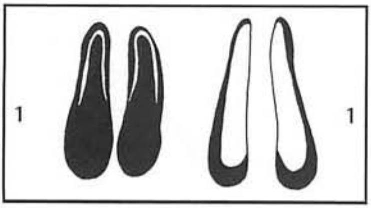

DEER/ANTELOPES

Deer, found in well-wooded country on every continent except Australia, vary from the Moose of the north to tropical forest deer only 45cm (18in) high. Antelopes and gazelles occur in equal variety in Africa east to India with one, the Pronghorn, in North America. Shy, elusive, with superb hearing and smell, they usually live in groups. Most active at dawn and dusk and – except for those in arid areas – never far from water. Excellent meat which smokes well. Hides of deer particularly supple, horns and antlers useful implements.

Regard their horns as WEAPONS. Large animals can be aggressive and can gouge and stab with great power.

Traps: *Snare or deadfall for small types, leg spring snares, spear traps and deadfalls for larger. Bait with offal – curiosity will draw them.*

Tracks and signs: *Cloven hooves form two oblongs. Reindeer noticeably rounded. Illustration below shows, in scale, Roe Deer front and hind track (top) and Reindeer (bottom). Note dew-claw impressions on reindeer track. Walking animals' prints overlap front and rear, running animals' spaced. Droppings oblong to round pellets, usually in clumps. In temperate area, winter droppings lighter and more fibrous. Scrapes on saplings, nibbled and frayed bark.*

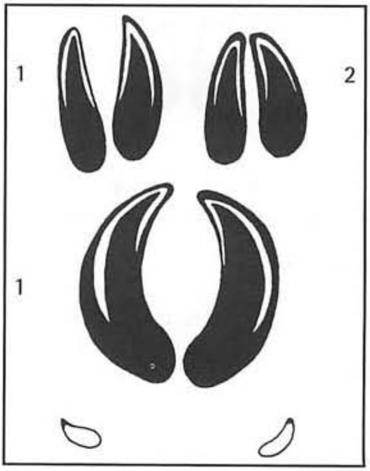

WILD PIGS

Pigs, peccaries and other hogs vary in size. Many are thickly haired, but all have piggy shape, long snout and two large teeth, or tusks, at the side of the mouth. They live in family groups, usually in wooded country – though Warthogs prefer savannah. Difficult to stalk, though not so alert as deer, except when lying up during heat of day when you have more chance. Listen for their snores. Large boars and those guarding young need little provocation to charge and can knock a man down. Tusks inflict severe injury, often dangerously close to the femoral artery on the upper leg. Meat good, with plenty of fat in winter, but must be thoroughly boiled because of parasites.

Traps: *Strong spring snares, deadfalls, pig spear traps on the game trail or hope for a passing group. Pigs eat virtually anything, so bait with whatever is available.*

Tracks and signs: *Cloven hooves leave marks like deer. On soft ground the short side toes distinguish them. Young animals have more pointed hooves. Droppings often rather shapeless, never long, firm or tapering. Ground disturbed by rooting, mudwallows best signs. Warthogs have large, burrow-like sleeping holes.*

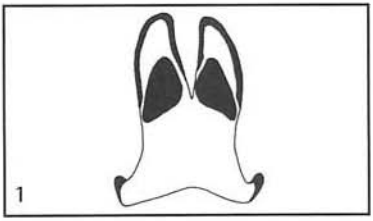

CAMELS

Range wild in deserts of north Africa, east to central Asia, and have been introduced to Australia – but they require a very powerful spear trap or projectile weapon. Camels can spit and could inflict a powerful bite on your hands, face or body.

LLAMAS

Alpacas, Guacanos and Viçunas of the Andes, south to Patagonia, can be trapped like antelopes. Semi-domesticated, Llamas and Alpacas indicate that people are probably not far away.

TAPIRS

Shy animals of the deep forests of tropical America and south-east Asia, they resemble large, hairless pigs with floppy snouts. Snare or trap with spear traps or deadfalls on their trails.

RABBITS/HARES

Widely spread by man and found from well-inside the Arctic Circle to deserts and jungles. In most areas these are the first animals to try to trap – Arctic (snowshoe) hares and Jackrabbits of the far north could save your life. Rabbits are easier to catch, most live in burrows, often in large numbers and using well-worn runs – the places to set snares. Young rabbits will often lie quite still and can be picked up. Hares do not live in burrows and tend not to have regular runs. Myxamatosis has led to many rabbits also living above ground.

Traps: *Simple snares – though a spring snare will take the animal off the ground and reduce the chance of your meal being stolen.*

Tracks and signs: *Hairy soles leave little detail on soft ground but combination of long hind and shorter front feet is distinctive. Hares have five toes on front feet, but inner is short and seldom leaves a mark. Hind foot*

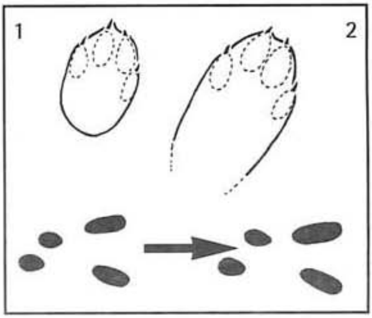

narrower, four toed. Rabbit similar but smaller (4 x 2·5cm (1¾ x 1in). Running leaves hind in front of forefeet instead of side by side. Droppings small, hard, round pellets. Bark nibbled at bottom of trees leaving two incisor marks. Rabbits thump a warning sounding like someone hitting a cushion.

DANGER

RABBIT STARVATION!

Rabbits can provide the easiest of meals but their flesh lacks fat and vitamins essential to man. The Hudson Bay Company recorded cases of trappers dying of starvation although eating well on an easily available diet of rabbit.

The body uses its own vitamins and minerals to digest the rabbit and these are then passed out in the faeces. If they are not replaced weakness and other symptoms of vitamin deficiency appear. If more rabbit is eaten, the condition becomes worse. Trappers literally ate themselves to death when eating vegetation would have ensured their survival. This situation often occurs when vegetation has been buried by snow and survivors rely on rabbits for food.

Myxamatosis, a viral disease that causes swelling of the mucous glands, especially on the head, makes rabbits sluggish and often blind. Their appearance is off-putting but the disease does not harm man. Once skinned the only indication of it will be white spots in the liver.

Rabbits and many rodents carry Tularemia (see *Rodents*).

SMALL RODENTS

The largest group of animals and some of the easiest to catch, though most are too small for a snare. Tracks of different kinds are not easy to distinguish. Rats and mice occur almost everywhere. They may be tempted into cage traps or under deadfalls. Rats carry disease. When gutting take care not to rupture the innards and cook thoroughly. Unwelcome, except as food; try to catch them if they invade your camp.

Guinea pigs and Cavies: *are widely domesticated in the Andes, where they live. They can be tempted to traps baited with fruit and leafy vegetables and make excellent eating.*

Capybara: *found in tropical lowlands in parts of South America grow to the size of a small sheep and live in family groups. Semi-aquatic, they are easily panicked into traplines or ambushes if their escape route to water is cut off. Meat is lean and tasty.*

Copyu: *similar to Capybara and also from South America, is smaller. Occurs ferally in Europe. Tasty.*

PORCUPINES

Various kinds are found in Americas, Africa and tropical Asia. Some climb trees. Ground-living ones are bumbling – easily run down and speared. Quills can inflict injury.

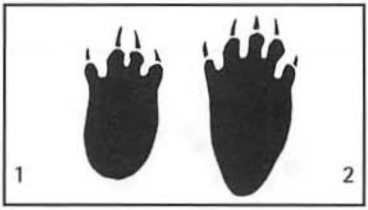

BEAVERS

Chunky, dam-building aquatic animals with scaly, paddle-like tails. Found in North America and northern Eurasia using regular runs along streams, where they can be trapped. Good to eat, especially the tail.

Traps: *Strong net or line of nooses across water run. Spring snares or* *deadfalls where they emerge to chew trees.*

Tracks and signs: *Five toes with claw marks, often only four show. Rear track webbed, roundish, larger 15 x 10cm (6 x 4in). Look out for dam building and lodges, felled and chewed saplings, bark and shavings near water.*

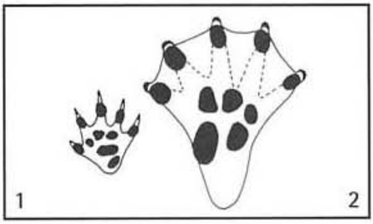

HEDGEHOGS

Shy, mainly nocturnal animals of Eurasia and Africa east to Indonesia they lie in well-concealed nests during the day and hibernate in temperate regions. They move fairly slowly so can be run down. Those that curl up are easiest to catch. Edibility fair, handle carefully usually parasite infested. Skin can be removed with spines intact. Cook thoroughly.

Traps: *Snares are unlikely to grip. Use deadfall – hedgehogs like to snuffle about obstructions.*

Tracks: *Five toes with long claws but usually only four show.*

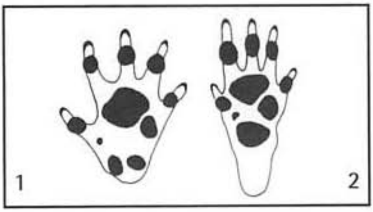

SQUIRRELS

Squirrels and Prairie dogs occur everywhere except Australasia and the Poles, hibernating in colder areas. Alert and nimble, most are active by day and night – feeding on nuts, fruits, shoots etc and some on birds' eggs. Raid their nests in tree hollows for young – beware of their sharp teeth they are savage in defence. Ground-living kinds live in burrows, often in large colonies. Most are excellent eating.

Traps: *Small spring snares attached to bait bars. Use split fruit or a bird's egg to attract. For tree squirrels set 5cm (2in) loop snares along a pole and lean against trunk of squirrels' trees – even when one has been caught they will go on using it as a short cut.*

Tracks and signs: *Four slender toes, with claws on front foot, five on rear. Chewed bark, gnawed nuts, cones beneath tree or an untidy nest of twigs in a fork may indicate squirrels.*

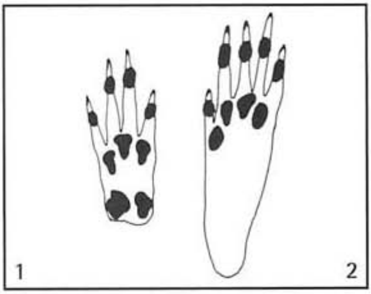

KANGAROOS

With wallabies and other relatives kangaroos are limited to Australia, Tasmania and parts of Papua New Guinea. Large kinds can strike powerfully with hind feet. Some adept tree climbers. Too fast to run down. Most active at night, some shelter in caves etc during day. Plains kangaroos range far from water. Edibility fair, but not easy to catch!

Traps: *Deadfalls, spring snares – though in open country no trail to set them on. Tempt tree kangaroos down with fruit, leaves and roots.*

Tracks and signs: *Only two prints, resembling giant rabbit tracks (front legs not used for support) or locomotion.*

OPOSSUMS

Small nocturnal scavengers of southern US to South America. Similar but unrelated animals in Australasia. Good tree climbers. Possum pie a traditional US dish.

Traps: *Bait with juicy fruit, eggs; or anything to hand. They are inquisitive.*

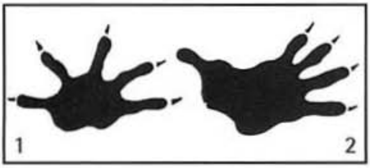

WOMBATS

Plump, badger-size animals of Australia. Eucalyptus forest, dry country and Tasmanian forms. Live in burrow. Emerge at night to forage. Trap as badger.

RACCOONS

Cat-sized, with bushy banded tail and black mask, found widely in North America. Prefer woodland near water. Introduced parts of northern Europe. Inquisitive, nocturnal, emerge at night from tree or rock hole to hunt frogs, molluscs, insects; small mammals, berries nuts. Crab-eating Raccoon of tropical America is related.

Traps: *Bait a spring snare.*

Tracks: *Five long toes with clear claw marks; front foot small, rounded; rear larger, tapering.*

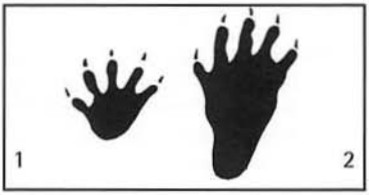

REPTILES

CROCODILES/ALLIGATORS

Members of this group are found in most subtropical and tropical regions. Those under about 1.3m (4¼ft) are potential food. Larger ones should be avoided – they are virtually armour-plated and some such as the cunning and aggressive estuarine crocodile of Indo-Australasia, can grow to enormous size. In areas where they live ALWAYS assume they are about for they have excellent camouflage and can lie underwater for long periods waiting for an incautious meal to step within range. Their tails can inflict a scything blow, almost as damaging as their teeth. The best meat comes from the tail and is firm and very tasty.

Traps: *Set by water for small crocs or catch on line with a stick wedged in the bait to lodge in gullet. Kill with a sharp stick wedged in the bait to lodge in gullet. Kill with a sharp blow between the eyes.*

LIZARDS

Gila monsters and Beaded lizards are docile but venomous (see Poisonous and dangerous animals in *Health), but all lizards can be eaten. Most are timid though some big iguanas and monitors can inflict a bad bite and have powerful claws. Do NOT tangle with the giant Komodo dragon (2m/6ft long), confined to a few Indonesian islands.*

Try to catch lizards by the tail, but small lizards move fast. They can sometimes be trapped in a pit, look out for them in a solar still. Set traps for larger ones.

TURTLES/TORTOISES

All these reptiles are good eating. Most spend their lives largely in salt or fresh water, emerging to lay eggs, but a few are terrestrial, the box turtle not even entering the water to mate. Some *turtles are encountered far out at sea, a bonus for the sea survivor. Net or drag them from the water. On land use a stick to turn them on their backs – keep out of the way of jaws and flippers. Then they will be defenceless. Kill with a blow to the head. Cut through belly and discard guts, remove head and neck (some have poison sack in neck). Best boiled. Very rich, so eat in small amounts. Blood even richer. Eggs found inside females can be eaten. Tortoises can retract their heads – you will probably have to stab into it. Roast ungutted in embers, when the shell splits they are ready.*

Tracks: *Females coming ashore to lay eggs leave tractor-like tracks on sand. They will lead to where the eggs are buried, but you may have to go quite deep to find them. Boiled, they will keep for a few days.*

AMPHIBIANS

Frogs are all edible, from tiny tree frogs to African giants with legs as big as a chicken's, but some have poisons under the skin so remove skin before cooking. Active at night, their croaking will locate them, always near water. Dazzle with a light – a firebrand will do – and club. In daytime hunt with hook and line and an insect-like lure. The legs of large frogs are a particular delicacy and taste like chicken.

Toads have warty skins and may be found far from water. Bad-tasting, most have highly toxic skin secretions. Do NOT eat.

Salamanders and newts, found in and near water, can be caught like frogs and are just as edible.

SNAKES

Overcome your fear and learn their habits. A snake is a steak! But do NOT try to catch one you know is highly poisonous, if it is too large to tackle safely or you are ill-equipped.

The large constrictors – the pythons of Africa, Asia and Australasia, and the boas, like the Anaconda of South America can be very large up to 10m (30ft). They tend to be timid and even the largest would find an adult human too big to swallow. Easy to catch – but do NOT tackle large ones! They are not poisonous but their back-raked teeth can give a very serious bite. If you can prise open the jaws, rather than pulling the snake off, the results will be less severe.

Snakes have excellent camouflage, often only movement gives them away. You may pass many every day without noticing them.

Use a forked stick to pin snake down just behind head. Strike the back of the head with another stick. Tree snakes can be clubbed and then knocked to the ground. Club them again to make sure!

Coiled snakes with head concealed present a problem. Pin one fat coil and watch it unwind before going for the head.

NEVER pick up – or even get close to – a snake until you are sure it is dead. A few, including poisonous ones, can feign death convincingly.

BIRDS

All birds are edible, but some taste very much better than others. They occur everywhere – open sea, deserts, high in mountains and in polar regions – but are not always easy to catch.

Game birds make best eating: pheasants, grouse, partridges, quail, ducks, geese, jungle fowl – but they are wary and many are superbly camouflaged. Birds of prey need thorough boiling to tenderize and to destroy parasites. Sea birds can be oily and fishy-tasting.

Traps: *Easiest to trap are birds of prey,* *such as eagles and hawks; carrion eaters, such as vultures; inquisitive scavengers, such as crows and gulls.*

Cage traps, deadfalls and spring snares can be used for birds that take bait, nooses on branches may catch roosting birds. In wooded country site traps in clearings or by river banks. Bait with meat, fruit or berries. Most birds are light – so set delicately.

Use also pole-nooses, lines with baited hooks and gullet jammers.

Small birds are easy to lime and can be attracted by bait, but making a dummy owl can be more effective. Small birds come to mob it – even a crude model works.

Tracks and signs: *Except for those with webbed feet, tracks are broadly similar and indicate only the size of the bird, with a few exceptions. However, in desert and on snow, tracks may help to locate birds hiding in close cover. Calls and flights are much more useful. Alarm calls may help locate other animals. Droppings may indicate a night roost.*

Autumn moult: *Birds go through a complete moult in autumn and are unable to fly, or can manage only short distances. Some, especially ducks, geese and game birds are much easier to catch at this time. If not far out on water, you can run them down.*

Bird nesting: *Eggs are easily available from ground nesters – and many nest in colonies. Approach carefully – crawling not walking – and you may get within stone-throwing or clubbing distance.*

Some, such as gulls, guard nests tenaciously. If you enter a colony be prepared to be attacked.

Burrow nesters: *Puffins, petrels and a few other birds nest in burrows, usually in inaccessible places on rocky coasts and small islands. They feed at sea during the day, but can be pulled or dug out of their burrows at night.*

Flightless birds: *Ostriches (in Africa), Rheas (South America) and Emus and Cassowaries (Australia) are large and flightless but treat them with caution: an ostrich, in particular, can deliver bone-breaking kicks. Fairly rare – but their eggs will feed a large group and make useful containers.*

INSECTS

Insects are likely to be the survivor's most reliable source of animal food. Although usually very small they occur almost everywhere and are often so plentiful that enough for a meal can soon be gathered.

Weight for weight they give more food value than vegetables. Rich in fat, protein and carbohydrates they are life-savers, especially their larvae – those succulent grubs. You'll soon overcome your squeamishness – remember that some people consider them a delicacy. You have certainly eaten them unknowingly in food.

GATHERING

The most useful are termites, ants, beetles, grasshoppers, locusts, crickets, honey bees, caterpillars and various aquatic insects.

Many insects are inactive during the heat of the day, although most will emerge to collect moisture when it rains. Look for them in nooks and crannies of trees and behind their bark, in the tissue and seed pods of plants, in any moist shady spots and on the beds of pools and streams. Ants' and termites' nests are often immediately recognizable mounds.

Look for beetle grubs – usually pale in colour with three short legs, they range from tiny larvae to juicy *'sausages' 15.5cm (7in) and weighing 84g (3oz). You will find them on trees with peeling bark, and in decaying stumps. Some tropical palms and bamboos are infested with them. If you have a knife, cut material away until you find them under the surface.*

Collect only living specimens. AVOID any that look sick or dead, have a bad smell or produce skin irritation rash when handled.

Take care when foraging for insects. Their hiding places may also harbour unwelcome creatures such as scorpions and spiders or, in larger crannies, snakes.

WARNING

Do **NOT** gather insects feeding on refuse, carrion or dung – they are likely to carry infection.

REMEMBER: Brightly coloured insects – including their caterpillars – are usually poisonous. Their bright colours are the warning sign.

AVOID: Grubs found on the underside of leaves – they often secrete poisonous fluids. Bad smelling and slimy to touch, they can be used as fish bait.

BEWARE: Large beetles often have powerful jaws, handle them with respect!

PREPARATION

Most insects are edible raw and usually more nutritious that way but they are more palatable cooked. Boiling is safest, it destroys harmful bacteria and parasites, but roasting is easier if proper containers are not to hand. Just place your dinner on hot stones or in the embers of a fire.

Remove legs and wings from larger insects such as locusts, grasshoppers and crickets. Hairs on the legs can irritate or even block the digestive tract. Fine hairs on some caterpillars can cause rashes. If you want to eat a hairy caterpillar squeeze it to extract the innards – don't eat the skin. Take the armour-like casing off beetles.

Smaller insects such as ants and termites can be mashed to a paste and then either cooked or dried to a powder. Use this to thicken other foods or for storage – it will keep for some time. People who cannot abide the idea of insects in their food will find a powder easier to eat, especially in soups or mixed with other food.

TERMITES

Found in the warmer parts of the world, termites are nutritious and tasty. Most eat only vegetation but big ones have sharp jaws and will bite at anything. You can turn this to advantage by inserting a twig into the nests and gently withdrawing it. The termites will bite it and then hang on – but you will not collect very many in this way.

Flying termites, and flying ants, often take wing during thundery weather. At such times quite large quantities can be gathered from leaves and twigs where they settle.

Remove the wings from large termites before eating. They can be boiled, fried or roasted but are more nutritious eaten raw. Their eggs have good food value too.

Gathering: *Termites build large mounds, often several feet high and honeycombed with passages and chambers. Although rock hard, pieces can be broken off with a stone or stick and then dunked in water to force the termites out.*

A piece of termites' nest put on the coals of a fire will produce fragrant smoke that will keep mosquitoes and similar insects away. It will smoulder all night and help to keep the fire going.

When fishing, suspend a piece of nest above a pool; termites falling from it will be good ground bait.

BEES AND WASPS

All are edible – pupae, larvae and adult – but the honeybee also provides honey. This is the finest of all natural foods, easily digestible and highly nutritious – but NOT easy to collect for bees will guard their nest tenaciously.

During the day worker bees venture far from the nest, but all gather there at night. Then is the time to strike. Make a torch from a bundle of grass and hold it very close to the entrance so that the nest fills with smoke. Then seal the hole. That kills the bees, providing an immediate meal, and making their honey safely available. Remove wings, legs – AND STING – before eating the bees. Boiling or roasting improves the flavour.

Honey can be drained from the combs inside the nest. Even though it may harden, it will keep for years. Honey gives instant energy. Because it is so rapidly assimilated by the body it is an excellent restorative in cases of exhaustion. The comb itself can also be eaten but its wax is also useful for waterproofing clothing, softening pelts and making candles.

In some parts of the world there is a slight risk that honey may contain concentrations of plant poisons. This is likely only in areas where bees are dependent on a single plant source, as with dense stands of rhododendrons in the Himalayas. Smell will be one guide, but if in doubt use the edibility test given for plants.

Wasps are much more dangerous than honeybees, but they and other kinds of social bee, which do not produce honey, can be collected and eaten in the same way together with their larvae. There are also many solitary wasps and bees which do not make communal nests.

Locating: *Wasps' nests are usually found suspended from tree branches. Frequently the size and shape of a football, they may also be pear-shaped. The entrance is at the base. Bees' nests are most frequently found in a hollow tree or cave or under an overhanging rock.*

HORNETS

Are types of social wasp. If you can find their nest without the adults finding you, then you have a ready source of food – both larvae and pupae are highly nutritious. But be warned: hornets guard their nests ferociously. They sting on sight and the pain is extreme. Unless you are desperate – and hornets do occur in areas where other food is scarce – go in search of a safer meal.

There are two main kinds of hornet: those active by day and those active at night. Day hornets can be collected at night, in the same way as honeybees. Night hornets (which you could collect by day) inflict a sting like a white-hot rivet being driven into the body, and usually go for the face. They are best left well alone!

Locating: *Hornets make large globular nests, generally in trees.*

ANTS

Ants quickly gather round the merest scrap of food, where they can be collected, or you can break into a nest. Take care. Most ants have a stinging bite and some large jungle ants can inflict one that will lay the victim out for 24 hours. Go for the smaller ones.

Some ants, such as the Melanophus *species, have a distended abdomen full of nectar. Known as honey- or sugar-pot ants, they make much better eating.*

WARNING

Some ants have a bite that feels like a nettle sting. Some fire formic acid. Ants MUST therefore be cooked for at least six minutes to destroy the poison. They are then quite safe to eat.

LOCUSTS/CRICKETS/GRASSHOPPERS

All kinds have plump bodies and well-muscled legs. Some grow to 15cm (6in) long. In some areas they are abundant.

Swat them with a leafy branch or a piece of clothing. Remove the wings, antennae and leg spurs, then eat raw or roasted. Roasting not only kills any parasites they may carry but gives them a delicious taste.

AQUATIC INSECTS

Collect adults and larvae of water beetles, mayflies, stoneflies, caddisflies, damselflies, dragonflies and the rest of the great variety of water insects, only from fresh water. Although small they are found in large numbers.

Place a screen of fine material – a shirt or other piece of cloth – in the water to act as a 'net'. Secure with sticks if necessary. Walk towards it, from upstream in flowing water, stirring up the bottom as you go. The current will carry the insects with it and they will collect in the net. If you do not want to enter the water trawling the surface may produce a harvest.

It is best to boil thoroughly all insects caught in the water, just in case the water is polluted.

SNAILS/WORMS

Snails, slugs, worms and similar creatures should not be ignored. Many cuisines rate snails a delicacy and most people will eat water molluscs such as mussels or oysters without qualms. As for worms – think what they do for birds! Both land and water snails and other shell foods must be fresh. There are a few which can be dangerous to eat. Others which, at no risk to themselves, may have eaten something harmful to humans, require preparation.

SNAILS

Found in fresh water, salt water and from deserts to alpine meadows. There is a giant snail in Africa 20cm (8in) long. They are rich in proteins and minerals. When collecting land snails, avoid any with brightly-coloured shells – they may be POISONOUS. Sea snails, especially in tropical waters, are trickier and should be left alone unless positively identified. The cone shells Toxoglossa, *for example, of Pacific and Caribbean coasts, have a venomous sting like a hypodermic needle. Some kinds can kill.*

Starve snails for a few days, or feed only on herbs and safe greens so that they can excrete any poisons, and then put them in a saltwater solution to clear out their guts before cooking. Boil for 10 minutes adding herbs for flavour.

Hibernating snails can be eaten provided that the operculum (the seal at the entrance to the shell) has not receded.

SLUGS

Slugs are simply snails without shells. Prepare and cook them in exactly the same way as snails.

WORMS

Contain the highest class of protein with a large proportion of essential amino acids and are easily collected. Starve them for a day before you eat them. Allow them to purge themselves. Worms can be sun- or force-dried – one easy method is to leave them to dry on a hot stone – and then ground into a powder to add to other food as required. This will make eating them more 'acceptable' and dried they will also keep for some time.

DANGERS

Carriers of disease

The numerous diseases carried by mosquitoes, ticks and other insects, the unseen dangers of parasites you may pick up from food or water and various waterborne diseases are much more serious dangers than attacks by animals.

DANGEROUS CONFRONTATIONS

Attacks by animals are rare but large animals can be dangerous. Keep out of their way. If you confront one it will be as surprised as you. Self-control will be needed or you may unintentionally provoke the animal to attack.

- Avoid eye contact, but don't run away.
- If you come face to face with a large animal – FREEZE. Slowly back off and talk in a calm manner. In most cases the animal will back off too. Avoid making sudden movements and remember that animals can smell fear – many a hunter has fouled his breeches and given himself away. Do your best to calm yourself.
- If an animal appears to charge it may be that you are blocking its escape route. Move out of the way.
- If an animal seems determined to give chase (or you haven't got the nerve to freeze or sidestep), zig-zag when you run – animals such as rhinos charge in a straight line and have poor eyesight.
- A skilled nocturnal predator such as a leopard or tiger has excellent vision if you are on the move – though their colour vision is poor and they cannot see stationary objects well. Freeze if you have not already been sighted.
- Shouting and making a commotion may put off a predator.
- Taking to a tree is the final option – but you may be 'treed' for a long time if the animal is persistent.
- If chased, drop your pack, clothing or food, to confuse animal.

EDIBLE PLANTS

There are many temperate edible plants in addition to those illustrated in colour, including the wild forms of cultivated plants such as currants and gooseberries. The following are a further selection of the food plants available.

If you are unlucky enough to find none of the plants illustrated or described here use the standard tests for edibility on the plants that are available.

REMEMBER that although one part of a plant can be eaten another may be poisonous. Test leaves, stems, roots and fruits separately.

FRUITS

Currants and Gooseberries *(Ribes), found in woods, scrub and waste places, are medium-sized usually bushy shrubs with toothed leaves resembling those of a maple, small, greenish-white to purple five-petalled flowers and red, purplish black or yellow berries. Ripe currants are edible raw; cook gooseberries.*

Plums *(Prunus) exist in many varieties in scrub and woodland in virtually all temperate areas. Small shrubs or trees, similar to wild cherries, their fruits are larger, downy, blackish-purple, red or yellow; some are too tart to be edible raw.*

ROOTS, LEAVES AND STEMS

Horseradishes *(Armoracia) grow to 50cm (20in) in damp waste places with large, long-stalked, wavy-edged oval leaves and clusters of tiny white flowers. Chop up the hot-tasting root and add to stews; the young leaves are edible raw or boiled.*

Common Evening Primrose *(Oenothera biennis) is a tall plant of drier open areas, leafy, hairy, with spear-shaped, crinkly-margined leaves and sometimes reddish flower-stalks topped with large yellow, four-petalled flowers. The roots are edible boiled, changing the water to ease their pungency. Peel young leaves and treat likewise. The plants overwinter as rosettes.*

Limes or Basswoods *(Tilia) are tall trees, up to 26m (85ft) high, which like damp woods, with large, heart-shaped, toothed leaves and clusters of scented yellow flowers. Young leaves and unopened leaf buds are edible raw; the flowers can be used in tea.*

Hops *(Humulus), climbing plants of woody and scrubby places have long twisting stems, toothed leaves, deeply cut into three lobes, and green, cone-shaped female flowers. Peel, slice and boil the young shoots, brew up the flowers.*

Thistles *(Cirsium) have spiny, often ridged stems, oblong or spear-shaped, prickly, deep-cut leaves and large brush-like heads of purplish flowers. Remove prickles and boil young leaves. Peel tender shoots and eat raw or boiled. Roots of younger, stemless plants can be cooked and the base of each flowerhead contains a nutritious 'nut' which can be eaten raw.*

Saxifrages *(Saxifraga) grow to 90cm (3ft), usually much less, often liking open, rocky country, up into mountains. Most have rounded tapering or long-stalked leaves arcing from the base, often reddish stems and clusters of five-petalled flowers, usually white. Leaves edible raw or cooked.*

Great Burnet *(Sanguisorba officinalis) reaches 60cm (2ft) in damper grassy places, with toothed, spade-shaped leaflets in opposite* *pairs and oblong heads of tiny, deep red flowers. Eat the tasty young leaves raw or boiled. Take an infusion for stomach complaints.*

Redleg *or* **Lady's Thumb** *(Polygnum persicaria) reaches 60cm (2ft). With reddish mature stems, narrow, spear-shaped, usually dark-spotted leaves and spikes of tiny pink flowers. Often common on waste ground. Young leaves are edible raw or cooked like spinach.*

Wild Rhubarb *(Rheum palmatum), found in open grassy places and margins from southern Europe east to China, resembles cultivated rhubarb, but its leaves are more ragged and dissected. The large flowerstalks are edible boiled; other parts are harmful. Eat ONLY the stalks.*

Bladder Campion *(Silene vulgaris), grows to 45cm (18in) in grassy places, is grey-green, with pointed oval, stalkless leaves, clusters of white flowers with a swollen balloon-like base. Boil the young leaves for 10 minutes.*

Field Pennycress *(Thlaspi arvense) grows to 45cm (18in) in open grassy places, with broad, toothed, spear-shaped leaves clasping the stem, a head of tiny white flowers and distinctive, notched, coin-like seed pods. Leaves are edible raw or boiled.*

Clovers *(Trifolium) abundant in grassy areas, recognized by their distinctive trefoil leaflets and dense rounded heads of small flowers, ranging from white through greenish-cream to shades of red. Leaves edible raw but better boiled.*

Stork's Bill *(Erodium cicutarium) reaching 30cm (1ft), in open grassy places, is hairy, often pungent, with fern-like, twice-cut leaves and heads of tiny, five-petalled pinkish to white flowers whose fruits form a long, twisting 'bill'. Eat leaves raw or boiled.*

Burdocks *(Arctium), medium to large, bushy plants of open waste areas, have floppy oval leaves, often arching stems and many purplish thistle-like flowerheads that develop into clinging burs. Eat leaves and peeled stalks raw or boiled. Boil pitch of peeled root. Change the water to remove bitterness.*

Violet *(Viola) are small flowers found in many areas, including damp and wooded ones. Veined, crinkly, often heart-shaped leaves rise on long stalks with flowers in shades of blue-violet, yellow or white, made up of five unequal petals. Cook young leaves. Rich in vitamins A and C.*

Corn Salad *or* **Lamb's Lettuce** *(Valerianella locusta) grows to 10–20cm (4–8in) in bare rocky and grassy places. Well-branched, with oblong, stalkless leaves and clusters of tiny lilac-bluish flowers; its leaves are edible raw or cooked like spinach. A particularly useful plant to know because it appears from late winter onwards.*

Ox-eye daisies *(Leucanthemum) often common in open areas, average 90cm (3ft) tall, with narrow dark green, lobed leaves, the lower ones rounded, and large white and yellow daisy-like flowers. Overwinters as a rosette. Eat young leaves (a lighter green) raw.*

Cuckoo Flower *or* **Lady's smock** *(Cardamine) pratensis grows on damp ground to 50cm (20in) with many small leaflets in opposite pairs, roundish on the basal ones which form a rosette, and clusters of lilac or white, four-petalled flowers. Young leaves are tasty raw, older ones rather peppery.*

Brooklime *(Veronica) grows in shallow water and swamps. Its creeping to upright stems carry pairs of thick, oval, toothed leaves, from the stalk bases of which spring 7–25cm (3–10in) spikes of four-petalled blue flowers with two prominent stamens. Eat young shoots before flowering and leaves after. Slightly bitter (especially the European form* V. beccabunga*) but eat like watercress.*

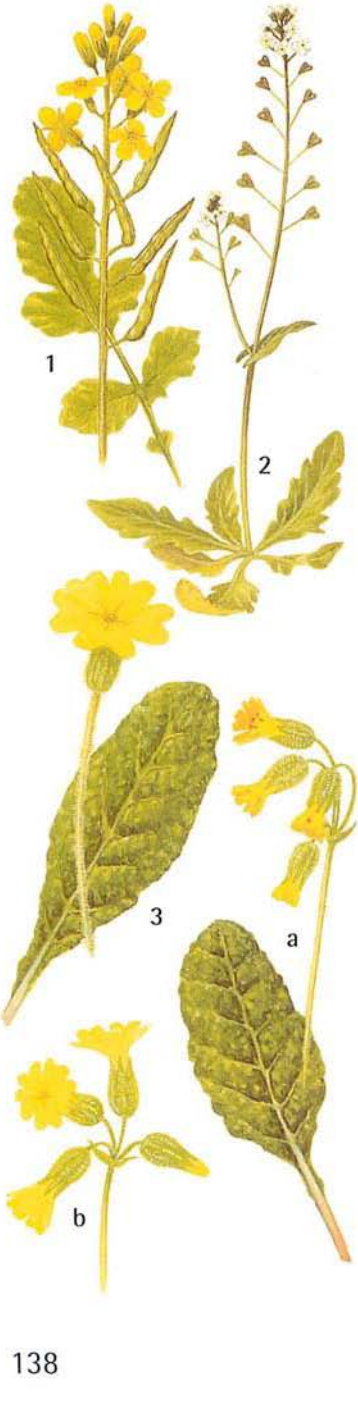

In spring and summer young shoots are tender and easy to pick. Some can be eaten raw, but many are better gently cooked, especially Solomon's seal, Willowherb, Cat's-tail and bracken. Wash them in clean water, rub off any hairs and boil in a little water so that they cook mainly in the steam.

Leaves are very rich in vitamins and minerals. Together with young shoots they are the survivor's easiest source of food. Most will taste better cooked but do not overcook them or you will destroy the vitamins they contain: C, E, K, B and large amounts of A.

1 **White Mustard** *(Synapsis alba) grows to 60cm (2ft), with a hairy stem, crinkly, deeply lobed leaves and pale yellow flowers; in waste and grassy places in Eurasia. The young, peppery leaves and flowers are edible raw; the whole plant is tasty cooked. Pick young specimens.*

2 **Shepherd's Purse** *(Capsella bursa-pastoris) may reach to 60cm (2ft), with a rosette of lobed, spear-shaped leaves and a spike of small white flowers; common in waste places. Boil the leaves, which taste like cabbage, and mix with other plants.*

3 **Primroses** *(Primula) are found in grassy and shady places. Identified by their rosette of crinkly, tapering basal leaves and long-stalked, five-petalled flowers which range from pale to bright yellow and, in some forms, pink. All parts are edible but the young leaves are best part to eat. The primulas include the Cowslip (3a) and the Oxslip (3b).*

4 **Dandelions** *(Taraxacum) occur in many forms almost everywhere. Look for the large, yellow to orange flowerhead or the rosette of deeply lobed leaves. Eat the young leaves raw; boil the older ones, changing the water to remove their bitter taste. Boil the roots or roast for coffee. Dandelion*

juice is rich in vitamins and minerals.

5 Chicory *(Cichorium intybus) is common in grassy and waste places. It grows 1.3m (4ft), with thick hairy, deeply basal leaves and leafy spikes of clear blue dandelion-like flowers. Prepare as Dandelion.*

6 Wild Sorrel *(Rumex acetosa) is common in waste and grassy places, reaching 1m (3ft), with long, arrow-shaped leaves and spikes of tiny reddish and green flowers. Gather young plants. Their mineral-rich leaves are edible raw but cooking will reduce the sharp taste.*

7 Buckwheat *(Fagopyrum esculentum) occurs in open grassy places in most temperate parts. Its 60cm (2ft) stems are usually red, with spear-shaped leaves and clusters of small pink or white flowers. Its seeds make good edible grain.*

8 Curled Dock *(Rumex crispus) grows to over 1m (3ft), with long narrow, wavy-margined leaves and whorls of small greenish flowers; in grassy and waste places. Boil the tenderest leaves from young plants, changing the water to remove the bitterness. Rubbing with dock leaves will soothe nettle stings. There are many other Docks temperate and tropical; prepare as here but use sparingly.*

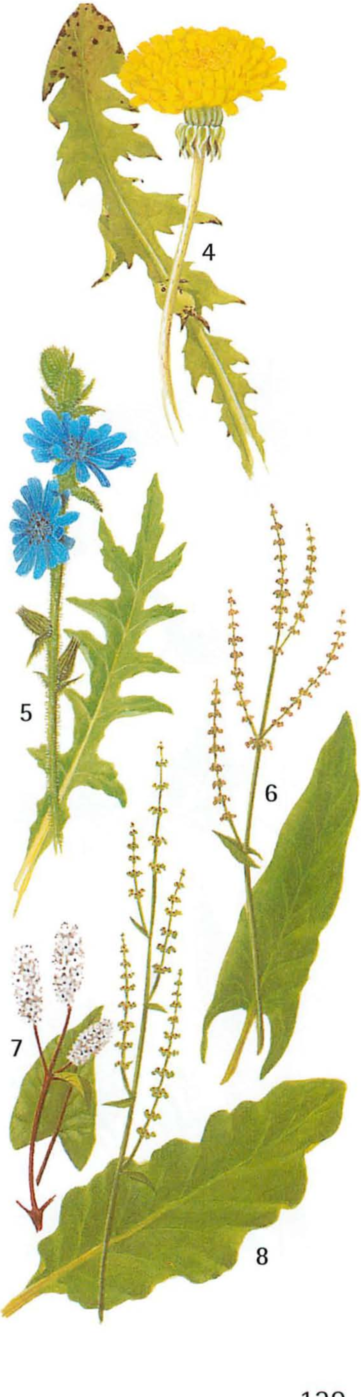

Some plants have edible stems, although many are too woody to eat. If they are soft, peel off the outer, stringy parts, slice and then boil. The inner pith of some stems is nutritious and sweet, elder, for example. In this case the stem must be split open and the pulp extracted.

Stems produce fewer nutrients for the survivor than the roots, shoots and leaves so put them at the bottom of the food choices and exploit their other uses. Fibrous stems, like those of stinging nettles, make good twine.

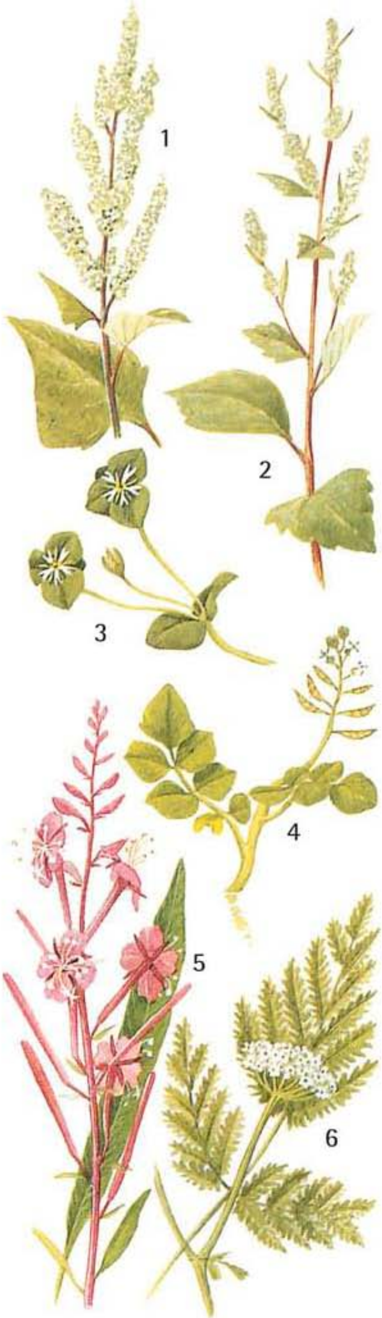

1 **Good King Henry** *(Chenopodium bonus-henricus) is spiky, to 60cm (2ft) tall, with dull green triangular leaves, sometimes reddening, and spikes of tiny greenish flowers; common on waste ground. Leaves and young shoots are edible raw or boiled as spinach; peel the shoots to remove the stringy parts.*

2 **Fat Hen or Lamb's Quarters** *(Chenopodium album) is spiky, to 1m (3ft) high, with often reddish stems, dull green, mealy, oval to spear-shaped leaves and spikes of tiny greenish flowers; abundant on waste ground. Cook the tasty leaves like spinach.*

3 **Chickweed** *(Stellaria media) is straggling, to 30cm (1ft) high, with a line of hairs on the main stem, pointed, oval leaves and tiny white, five-petalled flowers, common in waste places. Boil the delicious tender leaves.*

4 **Watercress** *(Roripa nasturtium aquaticunn) occurs, often abundantly, by running fresh water. It is creeping, semi-aquatic, with shiny leaves in opposite pairs and small, white, four-petalled flowers. Do NOT confuse with Water Hemlock. Leaves and stems are edible raw but boil if the water looks contaminated.*

5 **Rosebay Willowherb or Fireweed** *(Epilobium angustifolium) is found in open woods, waste and rocky places. It is tall, to over 1.5m (5ft), with spear-shaped leaves in opposite pairs and a spike of brilliant pinkish flowers. Young leaves, flowers and stems are edible raw but better boiled. Mature stems have a sweetish inner pulp.*

6 **Sweet Cicely** *(Myrrhis odorata) is sweet-smelling, grows to 1.5m (5ft), with slightly hairy and often purplish stems, feathery, fern-like leaves flecked with white and heads of tiny white flowers; in open woods, bare and rocky places in Europe. Do NOT confuse with Hemlock. Roots, stems and leaves taste of aniseed and can be boiled.*

7 Dead-nettles *(Lamium) are smaller than Stinging Nettles, with heart-shaped leaves and no stinging hairs, and white (7) or pinkish-purple (7a) flowers. Prepare as Chickweed.*

8 Stinging Nettles *(Urtica) are abundant for most of the year. Look for the toothed, narrow oval leaves covered in stinging hairs and the spikes of green flowers. Pick young growth or young plants 15–20cm (6–8in) high – BOIL for minimum six minutes to destroy the formic acid in the hairs. Leaves can be dried and stored; crushed stems provide fibres for rope.*

9 Plantains *(Plantago) are common in most areas. Ribwort or English Plantain (Plantago lanceolata) has spear-shaped leaves and much shorter flower-spikes than the Greater Plantain; it likes dry ground. Prepare as Greater Plantain.*

10 Buck's-horn Plantain *(Plantago coronopus) is small, star-shaped, with narrow, jagged leaves and shorter flower-spikes; in dry sandy and rocky places, often near the sea. Prepare as Greater Plantain.*

11 Greater or Rat's-tail Plantain *(P. major) has broad, oval leaves and distinctive upright spikes of tiny yellowish-green and brown flowers; in waste and grassy places. Prepare the rather bitter young leaves like spinach; use their expressed juice for wounds, or a decoction of the whole plant for chest complaints.*

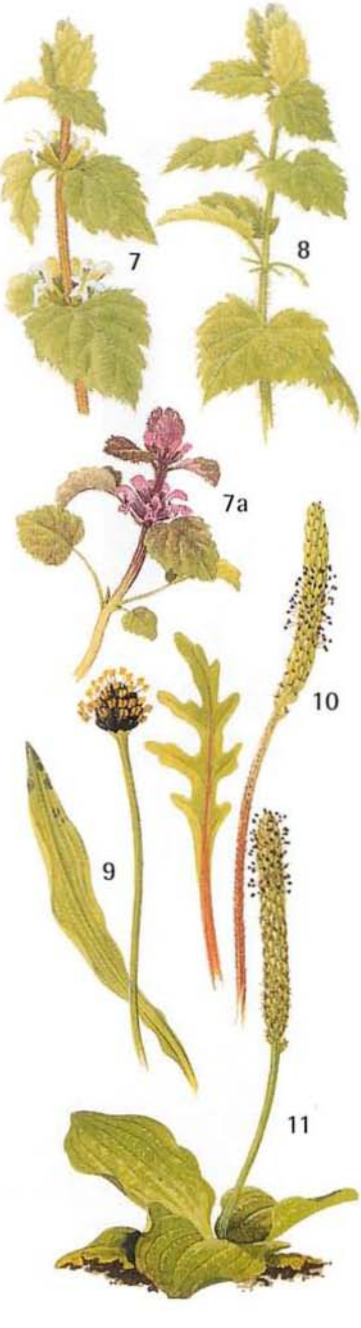

FLOWERS

The flowers of some plants are edible. These include limes or basswoods, roses, hops, elder, primrose and camomile. But they are only a seasonal supply and contain few nutrients compared with other parts of the plant. They are best used for teas and in medicinal infusions.

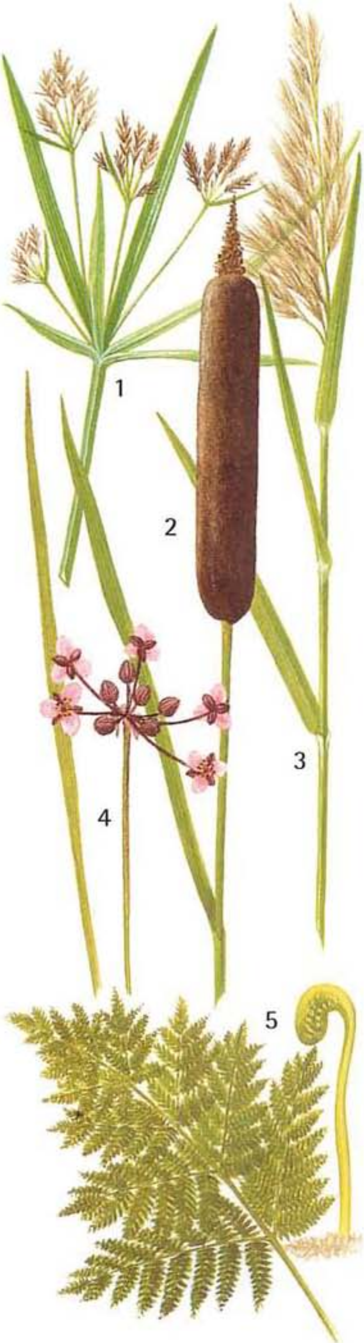

1 Galingale, Nutgrass or Chufa *(Cyperus) grows to 1.5m (5ft), with three-angled stems, long, strap-like leaves and a forking, clustered olive-brown flowerhead turning yellow with fruit; in and by fresh water almost everywhere. Peel and boil its nutty tubers, or dry and grind for flour or a coffee substitute.*

2 Cat's-tail or Reedmace *(Typha) grows to 2–5m (6–15ft), with long, narrow, greyish leaves and a conspicuous, dark brown, sausage-shaped flowerhead; in and by fresh water. The rootstock and stems are edible raw or boiled; cook leaves like spinach and young shoots like asparagus. The pollen can be mixed with water to make a dough and baked or cooked on a griddle or the end of a stick.*

3 Reeds *(Phragmites) grow to 4m (13ft), with greyish-green leaves and spreading, brownish-purple flowerheads on tall canes, in and by fresh water almost everywhere. Cook the edible root, punctured canes exude an edible, sugar-rich gum.*

4 Flowering Rush *(Butomus umbellatus) grows to 1.5m (5ft), with very long, strap-like, three-angled leaves arising from the roots and pink, three-petalled flowers; in and by fresh water in Eurasia. Peel and boil the edible rootstock.*

5 Bracken *(Pteridum aquilinum) is common almost everywhere often in large clumps. Older fronds are harmful; eat ONLY the strong-tasting young shoots, or 'fiddlesticks', drawing off their woolly parts and boiling for half an hour. Eat sparingly. The roots are edible boiled or roasted.*

10 Sweet Flag *(Acorus calamus) grows to 1.3m (4ft), with three-angled stems, wavy-margined, spear-shaped, strap-like leaves and a finger-like flower spike arising from the stem; in and by fresh water. Slice the pungent, aromatic rootstock and boil down to a syrup.*

Many familiar kitchen herbs grow wild. Their smells help to identify them. They can be dried and will keep well – but do not dry them in direct sunlight or they lose their essential oils.

6 Tansy *(Tanacetum vulgare) grows to 90cm (3ft), with toothed dark green, feathery leaflets and a cluster of button-like, bright yellow flowers, in waste and grassy places. Strong-smelling, with a hot, bitter taste. Use sparingly as a potherb, poisonous in quantity, leaves and flowers make a wormifuge tea. Its smell keeps flies away.*

7 Marjoram *(Origanum vulgare) is slightly downy and grows to 60cm (2ft), with small, oval, stalked leaves and clusters of small purplish-pink flowers; in warmer, dry, grassy places in Eurasia, introduced elsewhere. A sweet-tasting herb for stews; use an infusion for coughs and digestive complaints; chewed leaves relieve toothache.*

8 Ramsons *(Allium ursinum) is one of many wild garlics. It has broad, light green leaves like a Lily-of-the-valley and a cluster of white star-like flowers at the top of the stem; in woody places in Eurasia, revealed by its strong garlicky smell. Use any part as a potherb.*

9 Borage *(Borago officinalis) is round-stemmed, hairy, to 30–60cm (1–2ft), with pointed oval leaves, blue star-shaped flowers and a cucumber smell; in waste grassy places in Eurasia. All parts are edible raw or cooked, use an infusion for fevers. The stems produce salt when cooked.*

11 Wild Angelicas *(Angelica) grow to 1.5m (5ft), with hollow stems, sometimes purplish, broad, toothed leaves in opposite pairs and heads of tiny greenish, white or pink flowers; in damp grassy and woody places. The aromatic leaves, stems and roots are edible boiled; use an infusion for colds or externally for stiffness. Do NOT confuse with Water Hemlock.*

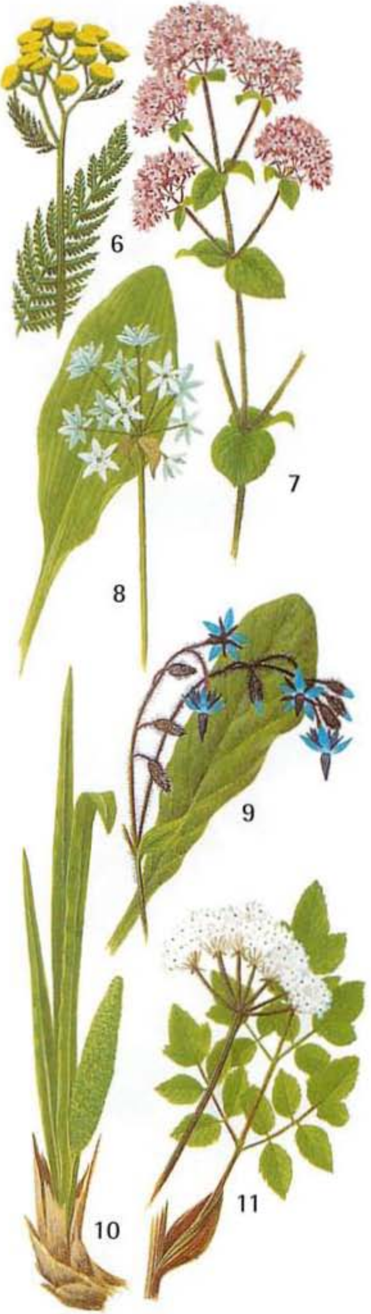

ROOTS AND TUBERS

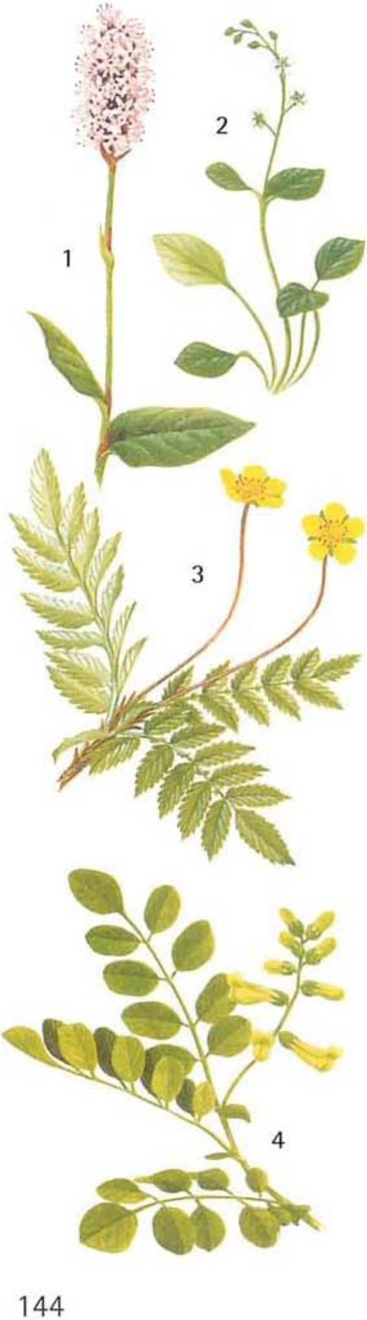

Roots and tubers are invaluable survival food. They are full of nutrition, particularly starch. All roots should be thoroughly cooked if in any doubt as to their identity.

1 **Bistorts or Knotweeds** *(Polygonum) average 30–60cm (1–2ft) with narrow, triangular leaves and a slender spike of pink or white flowers; in grassy and woody places, into the far north. Soak roots to remove bitterness, then roast.*

2 **Spring Beauties** *(Montia) average 15–30cm (6–12in) with a pair of oval, long-stalked leaves halfway up the stem and small white or pink flowers; in disturbed, especially sandy, places. Dig out tubers with a sharp stick, peel and boil. Young leaves are edible and supply vitamins A and C.*

3 **Silverweed** *(Potentilla anserina) is small, creeping, with silver-white undersides to its segmented leaves and solitary, five-petalled yellow flowers; in damp places. The fleshy roots are edible raw but better cooked. Use an infusion of the leaves externally for haemorrhoids and internally for digestive complaints.*

4 **Sweet Vetch or Liquorice Root** *(Astralagus glycophyllos) is straggling, to 30–60cm (1–2ft), with small oval leaves in opposite pairs and greenish-cream flowers; in grassy, scrubby and sandy places. The root is edible raw, tasting of carrots when cooked.*

5 **Wild Parsnips** *(Pastinaca sativa) are hairy, pungent, averaging 1m (3ft), with toothed leaflets and dense heads of tiny yellow flowers; in waste and grassy places. The roots are edible raw or cooked.*

6 **Comfrey** *(Symphytum officinale) is coarse, hairy, growing to 1m (3ft), with spear-shaped leaves tapering on to the stem and clusters of cream or mauve bell-shaped flowers; in ditches and*

damp places. The root is edible raw or cooked. Other parts are medicinal (see Natural Medicine *in* Health*). Do NOT confuse with Foxglove.*

7 Salsify or Oyster Plant *(Tragopogon porrifolius) averages 60–90cm (2–3ft), with long, grass-like leaves running down on to the stem and large, solitary, purple, dandelion-like flowers; in dry waste places. The bulb-like root and young leaves are edible cooked.*

8 Woolly Lousewort *(Pedicularis lanata) is hairy, low-spreading, with rose-pink flowers and a yellow root edible raw or cooked; widespread on the northern American tundra. CAUTION: some other louseworts are poisonous.*

Roots are at their starchiest between autumn and spring. In spring some of the starch converts into sugar to sustain new growth. Some edible roots can be several centimetres thick and a metre or more in length. Tubers are swollen, bulb-like roots – a large one may sustain a survivor for a long time. Don't forget edible bulbs, like the onion – but beware, some bulbs, including the Wild Onion-like Death Camas of North America are poisonous.

Many roots are particularly tasty roasted. Parboil them until they are just becoming tender, then roast on hot stones in the embers of a fire. Some, including Galingale and Dandelion (see earlier illustrations), are fine substitutes for coffee when roasted and ground. Others, such as Wild Calla, can be ground and used as flour.

The root of Comfrey is particularly valuable. It is so rich in starch that after boiling it sets as hard as plaster of Paris and makes an ideal splinting agent for broken limbs.

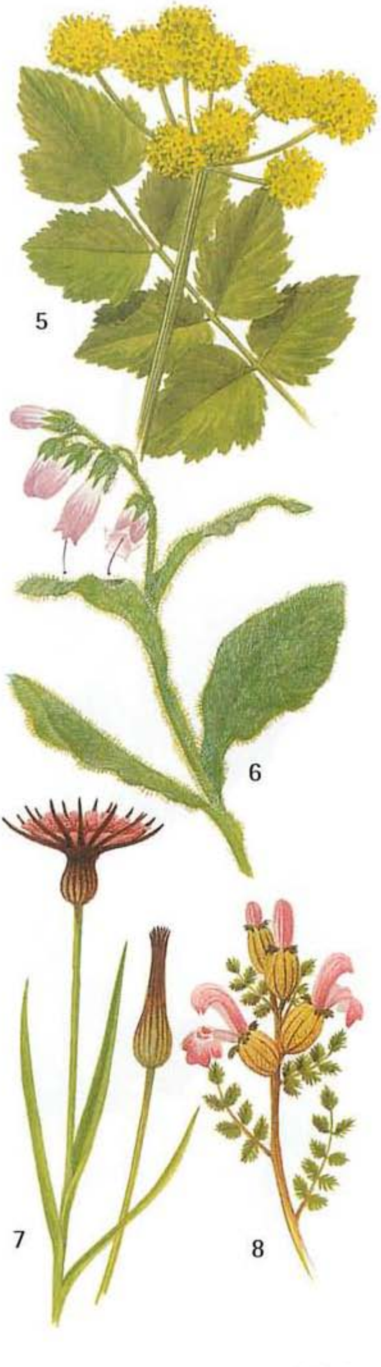

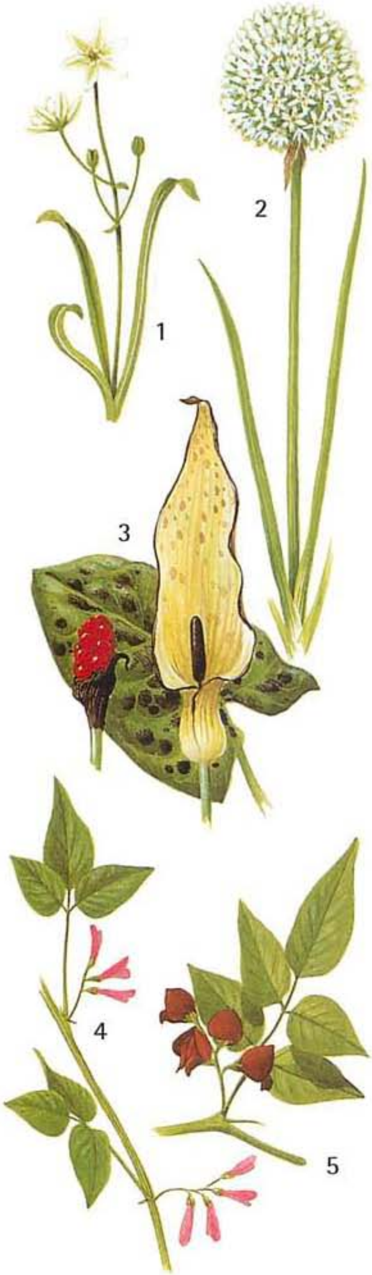

1 Star of Bethlehem or Starflower *(Ornitholgallum umbellatum) averages 10–30cm (4–12in), with grass-like leaves with a white midrib arising from the roots and white, six-petalled flowers, each petal green-striped; in grassy areas. The root is harmful raw and MUST be cooked. Avoid other parts.*

2 Wild Onions *(Allium) occur in most parts, easily detectable by their smell. Long, grass-like leaves arise from the base; a cluster of six-petalled pink, purplish or white flowers tops the stem. The edible bulb may be up to 25cm (10in) underground.*

3 Lords and Ladies or Cuckoopint *(Arum maculatum) grows to 15–40cm (6–16in), with dark green, arrow-shaped, sometimes dark-spotted leaves and a purple finger-like flowering organ enclosed in a pale leaf-like hood from which red berries arise, in shady and woody places in Eurasia. The root is harmful raw and MUST be cooked. Do NOT eat any other part.*

4 Hog-peanut *(Amphicarpa bracteata) occurs in moist places in North America: twining, vine-like, thin-stemmed, with light green oval leaves and lilac to white flowers. Extract each seed from its brown pod (underground) and boil.*

5 Groundnut *(Apios americana) is small, vine-like, with oval, sharp-pointed light green leaflets and maroon to brownish flowers; in moist, usually woody places in North America. Peel the small tubers then roast or boil.*

6 Jerusalem Artichoke *(Helianthus tuberosus) resembles a Sunflower, very tall, hairy, with large, rough, oval leaves and large, disc-like yellow flowers; wild in waste ground in North America, widely introduced elsewhere. The cooked tubers are delicious. Do not peel them or you lose food value.*

AQUATIC AND WATERSIDE ROOTS

7 Wild Calla or Bog Arum *(Calla palustris) is small, with long-stalked, heart-shaped leaves and a greenish finger-like flowering organ enclosed in a leaf-like hood pale on the inside, from which red berries arise; always by water. The roots are harmful raw and MUST be cooked. AVOID OTHER PARTS.*

8 Arrowheads *(Sagittaria) are aquatic, averaging 30–90cm (1–3ft), with large leaves varying from sharply arrow-shaped to spear-shaped and sometimes strap-like below the water, and flowers with three rounded petals; always by fresh water. Tubers are edible raw but much better cooked.*

9 Water Chestnut *(Trapa natans) is aquatic, with diamond-shaped floating leaves and divided, feathery submerged ones, and small white flowers; widespread in fresh water in Eurasia. The grey, hard 2.5cm (1in), two-horned seeds are edible raw or roasted.*

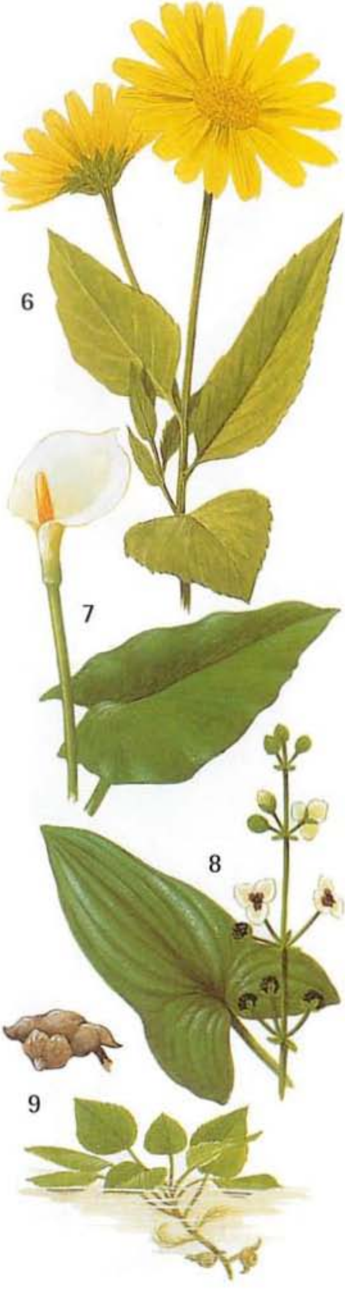

It is best to cook all roots before eating, as some are harmful raw – not just well known ones, like the tropical cassava and taro, but many temperate species. Most roots will need cooking to make them tender enough to eat. Scrub in clean water and boil until soft. Some potato-like roots have their vitamins and minerals near the surface of the skin and so should NOT be peeled. Roots will cook more rapidly if you cube them first. Use a sharpened stick to test them. If it goes in easily they are ready.

FRUIT

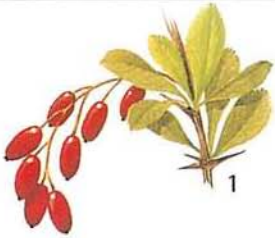

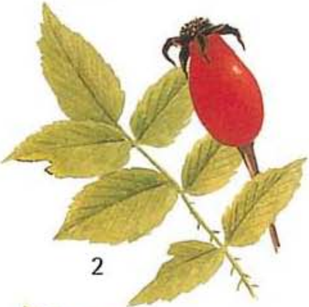

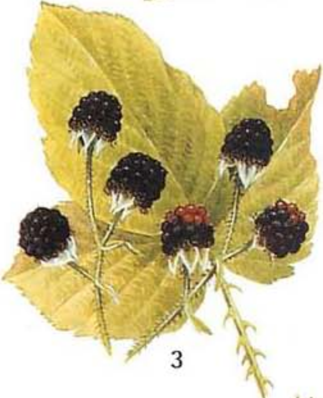

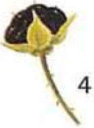

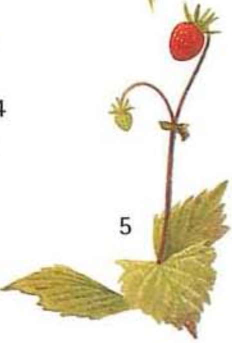

From summer on, fruits and nuts are one of the survivor's most important foods. Many will be familiar in their cultivated forms or from the traditional hedgerow harvest. Some are abundant, even on the tundra of the far north.

1 Barberry *(Berberis vulgaris), found in scrub and dry moorland grows to 3m (9ft), with oval leaves, yellow flowers and fierce thorns in groups of three on its stems. Its bright red, very acid berries are rich in vitamin C.*

2 Wild Roses *(Rosa) are found in most temperate areas. They resemble straggly, unshowy garden roses with thorned stems and simple white or pink flowers. Their hips (seedcases) contain more vitamin C than any other fruit. Chew to extract the juices and obtain maximum vitamin content, or crush and boil in water until only a syrup is left.*

3 Brambles (Blackberries) and Wild Raspberries *(Rubus) occur in scrub, woods and on open ground; leaves toothed and flowers white or sometimes pinkish in blackberries. Look for straggly bushes with arching thorny stems and juicy segmented berries, which ripen from green through red to purplish black berries, borne in late summer. Raspberries, less straggly and with fewer prickles, ripen to a rich red earlier in the summer. All are edible raw. Bramble canes can be used to pull rabbits out of sleeper holes.*

4 Dewberries *(Rubus) are like brambles but their berries are smaller and with fewer segments than the blackberry.*

5 Wild Strawberries *(Fragaria) are small, scrambling plants of dry grassy places and woodland whose fruits resemble small cultivated strawberries. You may have to look under the leaves* *to find the sweet, delicious fruit. Some kinds occur high in mountains. The fruits are rich in Vitamin C and best eaten fresh.*

6 Hawthorns *(Crataegus) are spiny shrubs or small trees found in scrub and waste places, with deeply lobed leaves, clusters of white or pink flowers and, in autumn, reddish fruits. Their flesh is creamy and edible raw. Young spring shoots are edible, too.*

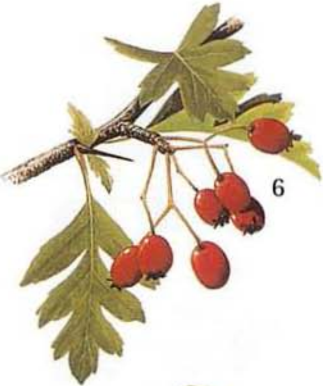
6

7 Crab Apples *(Malus) are short, rather spiny trees of scrubland and woods, with oval, toothed, often downy leaves, usually reddish-brown twigs and white, pink or red flowers. Fruit, often very bitter, looks like the cultivated apples. It can be sliced and dried for storage. Too many of the yellowish-green (sometimes red), pectin-rich apples will produce diarrhoea and are best cooked with other fruits.*

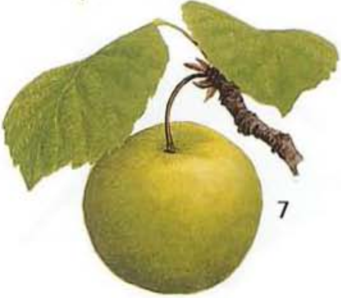
7

8 Wild Cherries *(Prunus) occur in woodland in most areas, growing to 24m (80ft) with small, pale green to reddish leaves, usually shiny reddish-brown bark, and white or pinkish flowers. The fruits are red or black depending on the kind; some kinds taste sour.*

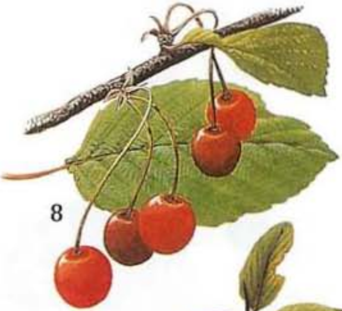
8

9 Blackthorn or Sloe *(Prunus spinosa) is a large bush, growing to 4m (13ft) with dark brown twigs, long thorns, oval leaves and white flowers, in woodland and scrub over Eurasia. The small blue-black fruits are very acid and better cooked down to a jelly.*

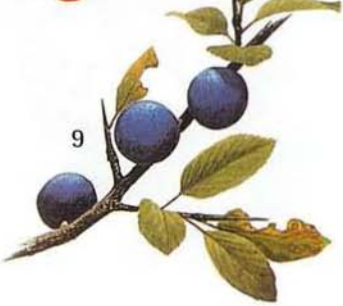
9

Fruits supply essential food values, particularly vitamins A, B2 and C. They are the staple diet of many animals and birds – so, where you find fruit, you will find animals too.

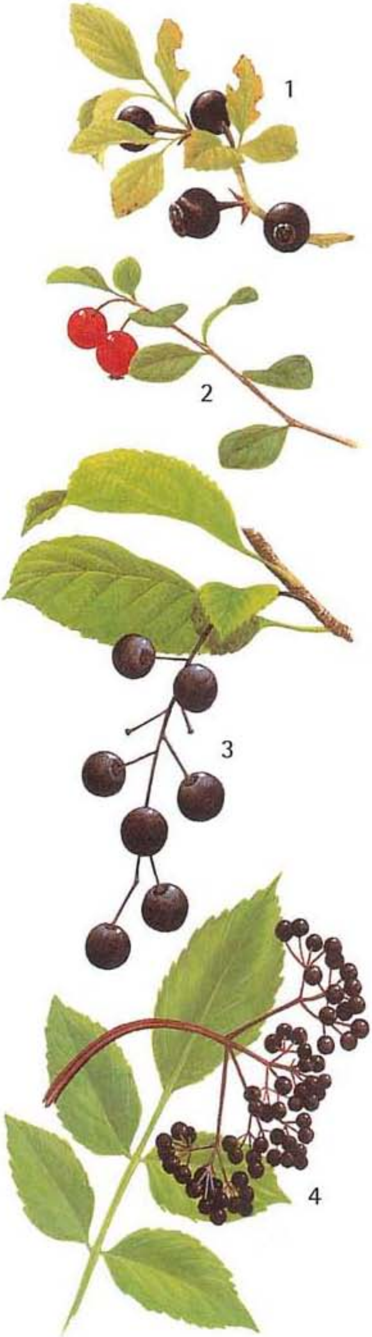

1 and 2 Bilberries, Cranberries, Cowberries and Huckleberries *(Vaccinium and Gaylussacia) are abundant on northern moors, bogs tundra, and sometimes in woods. Variable in size, but all are woody and shrubby with smallish oval leaves and small globe-shaped flowers varying from white to pink or greenish. The cranberry prefers marshy ground, the closely related, ground-hugging Cowberry (or Mountain Cranberry, 2) grows on moors. The spherical berries may be black (Bilberry, 1), dark blue (Huckleberry), mottled red (Cranberry) or red (Cowberry). They are edible fresh, cooked or dried for storage like raisins. The woody stems make useful fuel.*

3 Chokeberries *(Pyrus) are North American shrubs growing up to 2.4m (8ft) but usually much less, with spear-shaped, finely toothed leaves and five-petalled, pinkish or white flowers; in wet or dry woody areas, or on swampy ground. The red, purplish or black spherical berries, which grow in clusters, are excellent raw, dried or jellied. Do not confuse with the poisonous Buckthorns (*Rhamnus*).*

4 European Elder *(Sambucus nigra) occurs in scrub and woods, growing to 7m (23ft) with spear-shaped toothed leaves and clusters of tiny whitish flowers. The bunches of small purplish-black berries are best cooked down to a syrup. There are similar edible elders in other parts, but avoid the smaller ones as their red berries may be toxic.*

5 Juniper *(Juniperus communis) occurs in mountainous and northern areas, a woody shrub 5m (15ft) tall or a small prostrate bush with grey-green, needle-like leaves. Avoid young green berries; the ripe blue-black ones are best cooked with other food.*

6 Rowans or Mountain Ashes *(Sorbus) are common in woody and rocky places, growing to 15m (50ft)* *with smooth greyish bark, small toothed leaflets, white flowers and clusters of small orange berries. These are edible but sharp-tasting raw and can be cooked down to a jelly.*

7 Wild Mulberries *(Morus) average 6–20m (18–60ft), with oval leaves, sometimes deeply lobed, flowers on catkins, the red or black fruits look like large blackberries, 5–7cm (2–3in) long. They are edible raw. Found in woody areas in many temperate parts.*

8 Wild Grapes *(Vitis) are straggly, high-climbing, with large, heart-shaped coarsely toothed leaves, greenish flowers and bunches of amber to purplish grapes. Very widespread in the warmer parts of the world. As well as the fruit, young leaves are excellent boiled.*

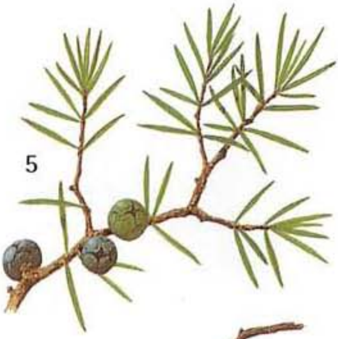
5

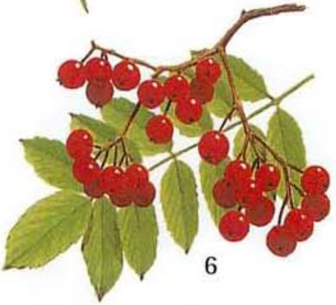
6

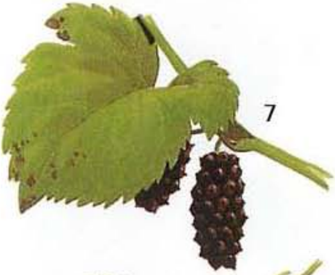
7

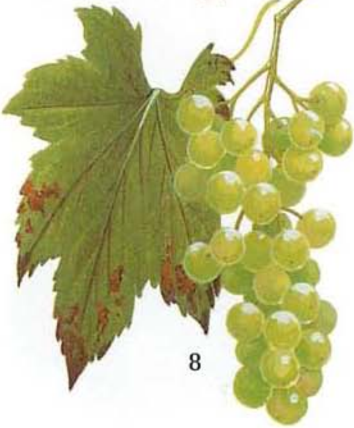
8

PRESERVING FRUIT

Fresh fruits soon go off, but they can be kept by making them into jellies. Most kinds contain an ingredient called pectin, which reacts with the acid in the fruit to help it settle into a jelly after boiling.

TO MAKE JELLY

First boil the fruit and then simmer until mushy. Some fruits have less pectin than others. These can be supplemented by adding another fruit rich in pectin, such as Crab Apple. The boiling action kills off any bacteria that would turn the fruit. Allow to cool and keep in a clean, if possible airtight, container.

Some fruits can be dried for storage, although this will take a week to ten days. Lay them in a single layer on a sheet, not in direct sunlight, and protect them from any moisture – both rain and dew.

NUTS

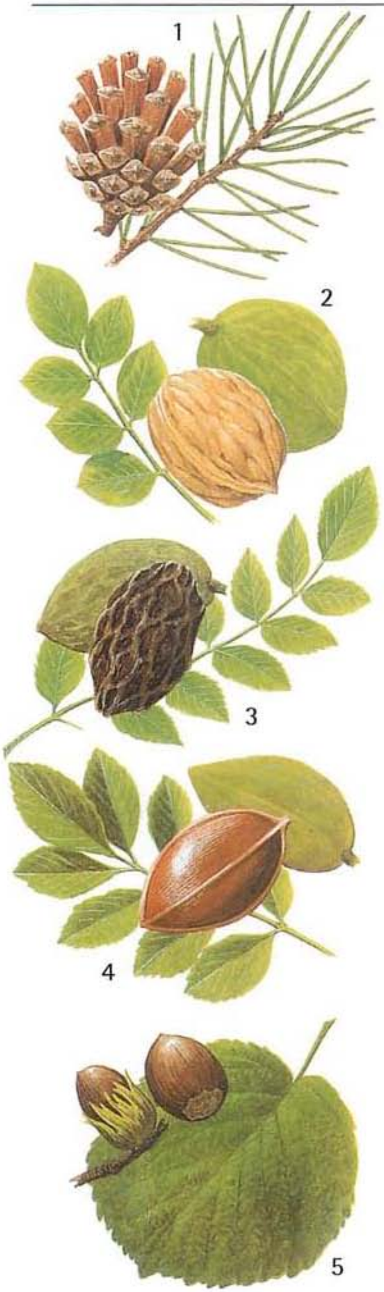

Nuts supply proteins and fat.

1 **Pines** *(Pinus) are the familiar cone-bearing trees with clusters of slim evergreen needles; in most temperate and northerly areas. Heat mature cones to release the seeds. They are tasty raw but delicious roasted. Roasted nuts can be ground for flour and can be stored. Young catkin-like cones are just about edible boiled. Needles and bark are also edible.*

2 **Walnuts** *(Juglans) grow to 30m (90ft), with leaves composed of many toothed, narrow leaflets and furrowed bark. The blackish brown nuts are at first enclosed in a thick green husk. In most temperate areas. One tree can yield up to 58kg (140lb) of nuts. Walnuts contain 18% protein, 60% fat and provide 6600 calories per kilo (3000 per pound)!*

3 **Butternut** *(Juglans cinerea) is another North American relative, smaller, with more greyish bark and oblong, sticky fruit husks.*

4 **Pecan** *(Carya illinoinensis) reaches 36m (120ft), with dark ridged bark and many small leaflets in opposite pairs, in moist places in North America. The oval, thin-husked nuts are richer in fat than any other vegetable product.*

5 **Hazels** *(Corylus) are tall shrubs of thickets and waste ground, with toothed, oval to heart-shaped leaves and brownish-yellow catkins. The highly nutritious nuts come in ovoid, leafy, bristly or hairy husks.*

6 **Sweet Chestnuts** *(Castanea) range from 5–30m (15–90ft), spreading, with large, toothed hairless leaves and bearing catkins, in wooded areas. Nuts, in some forms 2–3 together, are borne in globe-shaped, thick prickly green husks. Smash open husks, peel nuts, boil and mash. Do not confuse with*

Horse Chestnut which has large palmate leaves, like the fingers on a hand, and poisonous nuts.

7 Beeches *(Fagus) are very tall and spreading, with smooth lightish bark and thick-veined, wavy-margined sharp oval leaves; in broad-leaved woodland. Nuts small, triangular, 2–4 in each hairy husk depending on the species. Protein-rich; edible raw, roasted or crushed for oil.*

8 Oaks *(Quercus) occur in great variety in wooded areas. Many have deeply lobed leaves but all bear unmistakable acorns. Shell them and boil several times, changing the water to ease their bitterness, or steep in cold water for 3–4 days. Alternatively, bury them with ash and charcoal, watering from time to time. Then roast; roasted acorns make good flour or coffee substitute.*

9 Pistachios *(Pistacia) grow wild in warmer parts from the Mediterranean east to Afghanistan; introduced elsewhere. Trees, to 10m (30ft), with many small oval leaflets and clusters of nuts with a green kernel and reddish skin. Eat raw or parch on fire embers.*

10 Almonds *(Prunus) grow wild in warm, arid parts of Europe and Asia, widely introduced elsewhere. Resemble large peach trees, with sprays of blossom, small spear-shaped leaves and clusters of nuts in green leathery husks. Avoid bitter ones, which contain prussic acid.*

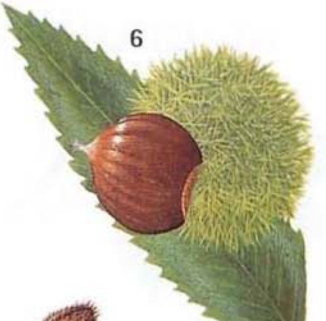
6

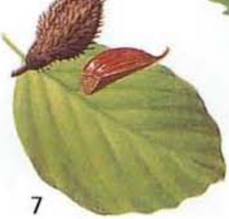
7

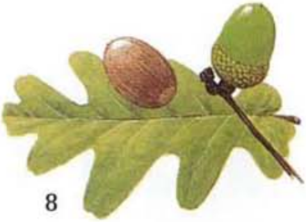
8

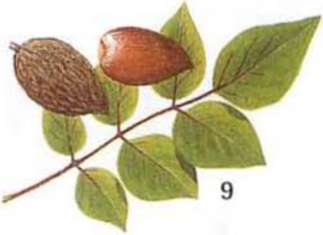
9

Extract oil from rich nuts such as beech. Crack open, separate meat from shells. Boil gently in water, skimming off the oil as it rises to the surface or allow to cool and separate. Store somewhere cool and dry, preferably in an airtight container. A yield of 270ml (3fl oz) of nutritious oils can come from 450g (1lb) of nuts.

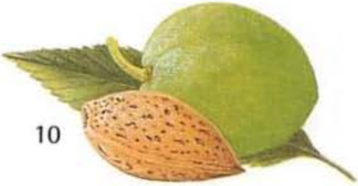
10

POISONOUS PLANTS

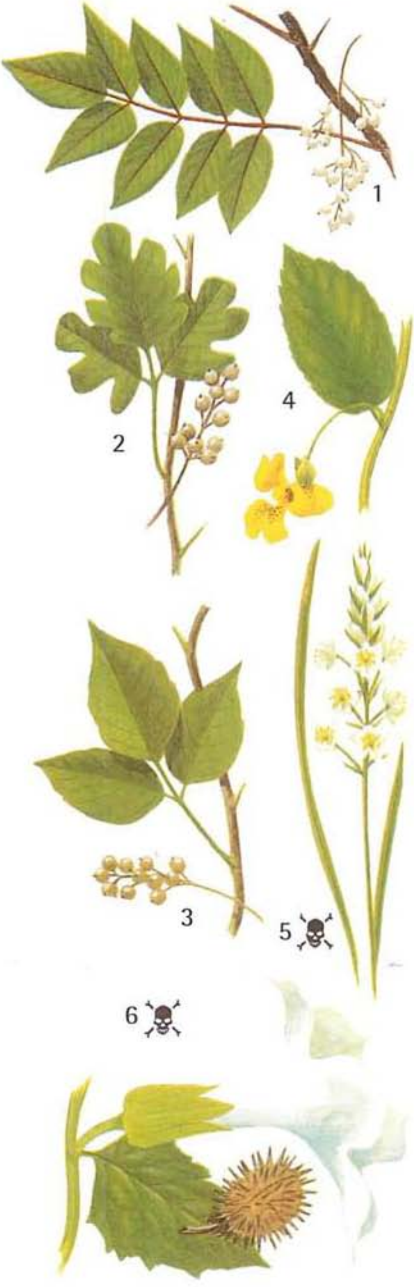

Compared to the many edible ones, there are few poisonous plants in temperate areas. Learn well the ones shown.

CONTACT POISONS

Contact with Poison Sumac, Poison Ivy and Poison Oak produces severe irritation and rashes. Wash effected parts immediately (see Poisons in Health).

1 **Poison Sumac** *(Toxicodendron vernix) reaches 2–6m (6–18ft), hairless, with many oval leaflets in opposite pairs, dark-spotted smooth bark and clusters of white berries; in swamplands in south-eastern North America.*

2 **Poison Oak** *(Toxicodendron quercifolium) resembles Poison Ivy but is smaller, always upright, and with oakleaf-shaped leaflets and white berries, in wooded parts of North America.*

3 **Poison Ivy** *(Toxicodendron radicans) is smaller, 0.6–2.1m (2–7ft), trailing or upright, with three-part very variable leaves, but always with greenish flowers and white berries; in wooded areas of North America.*

4 **Jewelweed** *(Impatiens), often found near Poison Ivy, with pale yellow or orange spotted flowers and seed pods that pop, provides a juice to ease irritation from contact with these and other plants.*

POISONS BY INGESTION

5 **Death Carnas** *(Zigadenus venosus) reaches 30–60cm (1–2ft), with long, strap-like leaves arising from the base and loose clusters of greenish-white, six-part flowers; in North America, in grassy, rocky and lightly wooded places. DEADLY: do not confuse with Wild Onions or Lilies.*

6 **Thorn-apple or Crimson Weed** *(Datura stramonium) averages 90cm*

(3ft), with jagged-toothed oval leaves and large solitary trumpet-shaped white flowers and spiny fruits; widespread in most temperate areas and also in the tropics. Sickly smelling. All parts DEADLY poisonous.

7 Foxgloves *(Digitalis) grow to 1.5m (5ft), with a rosette of basal leaves topped by a tall, leafy spike of purple, pink or yellow tube-shaped flowers; widespread in waste and disturbed places. ALL parts are highly toxic, affecting the heart.*

8 Monk's-hood *(Aconitum) reaches 1.5m (5ft), leafy, with palm-shaped, deeply segmented leaves and hairy, hood-like purplish-blue or yellow flowers; in damp woods and shady places. The most common kinds have purplish-blue flowers. VERY poisonous.*

9 Hemlock *(Conium maculatum) may reach 2m (6ft), much branched, with hollow purple-spotted stems, coarsely toothed leaves, lighter below, dense clusters of tiny white flowers and white roots; in grassy waste places. Bad-smelling. Very poisonous.*

10 Water Hemlocks or Cowbanes *(Cicuta) average 0.6-1.3m (2–4ft), branching, with purple-streaked stems, a hollow-chambered rootstock, small 2–3 lobed, toothed leaflets and clusters of tiny white flowers, always found by water. Smells unpleasant. ONE MOUTHFUL CAN KILL.*

11 Baneberries *(Actaea) reach 30-60cm (1–2ft), with leaves made up of several toothed leaflets, small, usually white flowers clustered at the end of a stem, and white or black berries; mostly in woods. ALL parts cause dizziness, Vomiting and severe internal irritation.*

12 Deadly Nightshade *(Atropa belladonna) may reach 1m (6ft), well branched, with oval leaves, solitary, bell-shaped, purplish or greenish flowers and shiny black berries, in woodland and scrub in Eurasia. ALL parts, especially berries, are VERY poisonous.*

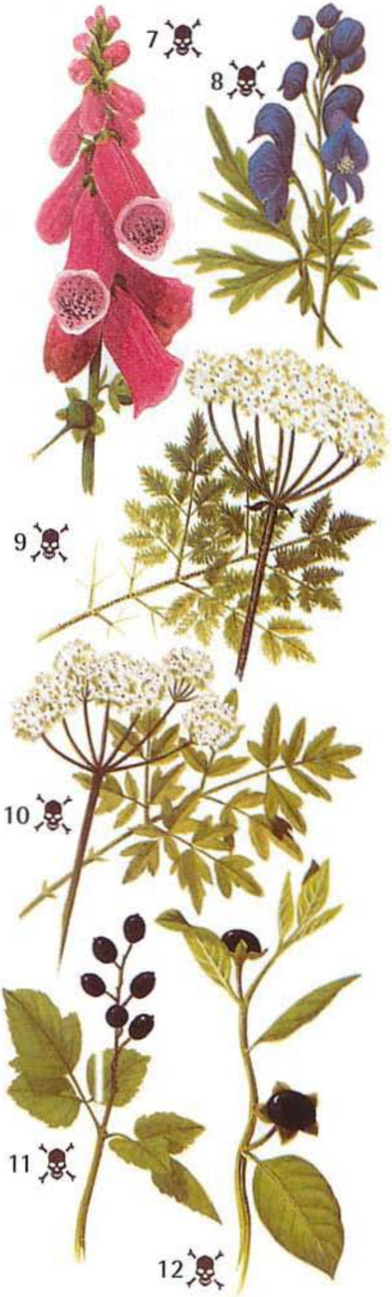

Some poisonous plants are easy to mistake for food plants if you are not careful in your identification. Do not take risks.

Hemlock and Water Hemlock are the two most important poisonous plants to learn, found in both Eurasia and the Americas in a wide range of country. Both are umbellifers, of which there are many kinds, and with many tiny flowers in dense clusters, like an umbrella inside-out and difficult to tell apart.

The umbellifers include some edible plants, but NEVER collect any unless certain they are neither of these two, which can be lethal in very small amounts.

Learn to recognize the following common poisonous plants in addition to those illustrated.

Buttercups *(Ranunculus) occur in great variety, from a few centimetres to over a metre tall, in north and south, including far into the Arctic. All have glossy, waxy bright yellow flowers with five or more overlapping petals. AVOID all, they cause severe inflammation of the intestinal tract.*

Lupins *(Lupinus), 30–90cm (1–3ft) tall, like the garden kind, growing in clearings and grassy places. Often with small leaflets in a palm shape or radiating like the spokes of a wheel, and spikes of 'pea-flowers': blue, violet, occasionally pink white or yellow. Any part can cause fatal inflammation of stomach and intestines.*

Vetches *or* **Locoweeds** *(Astalagus* and *Oxytropis), 15–45cm (6–18in) tall, growing in grassland and mountain meadows, usually with many small spear-shaped leaflets in opposite pairs and showy spikes of five-petalled 'pea-flowers': yellowish-white, pink to lavender and purplish. Some are very poisonous. AVOID all.*

False Helleborines *(Veratrum) 60cm–2.6m (2–8½ft), grow in wet, swampy places, some species in grassy ones. Oval ribbed leaves resemble a Lily-of-the-Valley's, drooping clusters of whitish or greenish-yellow flowers. Can be lethal.*

Henbane *(Hyoscyamus niger) medium to large, growing on bare ground, often near sea (Europe). Sticky hairs, toothed oval leaves (upper ones unstalked), creamy flowers streaked purple. Unpleasant smell. Deadly poisonous.*

POISONOUS BERRIES

Canadian Moonseed *(Menispermum canadense) North American vine-like climber, clusters of soft black berries. Could be mistaken for Wild Grape but lacks tendrils and has only a single crescent-like seed in each berry.*

Nightshades *(Solanum) medium to large, untidy, bushy plants liking scrubby places, leaves usually long-stalked, spear-shaped. Berries ripen from green to black, red, yellow or white. Plants producing similar but edible berries are usually more compact and woody, the berries smaller and more numerous. If in any doubt assume to be poisonous and AVOID.*

Virginia Creeper *(Parthenocissus quinquefolia) vine-like North American climber (introduced to Europe). Long-stalked, palm-shaped leaves, toothed leaflets, tendril and clusters of small blue berries, smaller than wild grapes. No plant with edible blue berries is vine-like and with tendrils.*

Buckthorns *(Rhamnus), shrubs sometimes small trees growing in scrubby and woody places, sometimes damp. Leaves oval, finely toothed; berries black and bitter tasting, clustered along stems. Can be violently purgative.*

There are many field guides available which will help you extend your knowledge of temperate plants, detailing many local forms.

TREES

Trees provide other nourishing foods, as well as fruit and nuts, which can be invaluable to the survivor, particularly in areas where there is little other plant life available.

The following refers to the trees of temperate and northern regions. Tropical trees, such as palms, are included in the section on tropical plants.

Bark

The thin inner bark (cambium) of certain trees is both edible and nutritious, but best in the spring, when sap has started to flow. Choose bark from near the bottom of the tree or from exposed roots. Peel it back with a knife to reveal the inner layer. This is mildly sweet and can be eaten raw – but can be made more digestible by long boiling, which will reduce it to a gelatinous mass. It can then be roasted and ground for use as flour.

Outer bark has too much tannin to be edible, but some kinds have medicinal uses.

Trees with best inner bark

Slippery Elm *(Ulmus rubra), of northern America, grows to a maximum 18m (60ft). Twigs are hairy and rough, oval, toothed leaves are hairy below.*

Basswood *(Tilia americana), a North American lime with large, heart-shaped leaves and dark, grooved bark which becomes smooth grey on upper parts.*

Birches *(Betula), which are often abundant in colder areas. They can be identified by their long delicate twigs and shiny, flaky bark that is often broken up into plates.*

Aspens *(Populus tremula), small to medium trees resembling poplars. They have rounded leaves on very, very long stalks that quiver in the wind.*

Tamarack *(Larix laricina), is found in cold parts of North America. It grows to 24m (80ft) with a pointed shape, bears cones and has needles in tufts along the twigs.*

Poplars *(Populus) have triangular leaves and prominent catkins. They occur in many northern areas.*

Maples *(Acer) grow widely and are recognized by their distinctive lobed leaves in three parts and by the two-winged fruits.*

Spruces *(Picea) are evergreen trees of cold climates. They are shaped like steeples, bear cones and have stiff, four-sided needles that grow all around the twigs.*

Willows *(Salix) are broad-leaved trees or shrubs with toothed leaves, lighter on the underside, and distinctive yellow or green catkins. There are many kinds, including ground-hugging arctic ones.*

Pines *(Pinus) are widely found evergreen trees bearing cones and clusters of long needles. Their inner bark is rich in vitamin C.*

Hemlocks *(Tsuga) are evergreens, resembling spruces, with sprays of foliage, flat needles and short, oblong cones. They bear no relation to the poisonous plant of the same name.*

NOTE: In addition to the inner bark, the buds and shoots of all these trees can be eaten raw or cooked – EXCEPT for those of Tamarack and Hemlock, which are POISONOUS.

Other uses for inner bark

The inner bark of some trees is very strong but pliable. It is easily torn into strips for lashings. The bark of the Mohoe tree, for example, is made into 'grass' skirts by Pacific islanders, having first been steamed in a hangi (see *Fire* in *Camp Craft).*

Birch bark can be removed from the tree in large sheets and forms an ideal material for roofing shelters or for making small containers. The North American Indians clad their canoes with it.

SPRUCE TEA

Steep spruce needles in hot water to make a tea. Collect only fresh, green needles and boil. The liquid produced is rich in vitamin C. The vitamin can be obtained more directly by chewing tender young needles, whose starchy green tips are particularly pleasant in spring. Spruces occur far into the north and are an important source of nourishment when little other plant life is available.

Gums and resins

With some trees, when cut, sap seeping out on to the bark hardens into a lump. If this is soluble in water it is a gum; if not, it is a resin. Both are very nutritious, rich in sugars and a useful survival food. A few have medicinal properties and others are highly inflammable and make excellent material for lighting fires.

Birch and maple syrup

Tap birch or maple in the same way as a rubber tree. Cut a V-shape in the bark to collect the sugary sap that runs out. Below the V make a hole in the trunk to insert a leaf as a drip spout to run the sap into a container.

Collect sap daily and boil it. It will give off lots of steam but thicken down into a syrup. This is instant energy and well worth the effort.

POISONOUS TREES

The following trees contain irritant or poisonous substances. Do NOT eat any part of them, except for the meaty roots of hickory, which are edible.

Yews (*Taxus*) are straggling evergreen trees or shrubs with flaky bark, dark green needles and red berry-like fruits. The fruits are particularly poisonous.

Cedars (*Cedrus*), originally from the Mediterranean and Himalayas, the true cedars are large, spreading, scented evergreens with erect cones.

Horse Chestnuts and **Buckeyes** (*Aesculus*) are tall with hand-shaped leaves, sticky buds and white, pink or yellow flowers. Do not confuse their poisonous, spiky-cased nuts with those of Sweet Chestnut which has narrow, toothed leaves and much more densely prickled seed cases.

Laburnums (*Laburnum*) are small, broad-leaved trees with three-part leaves and long sprays of yellow flowers.

Black Locust (*Robinia pseudoacacia*) is a North American tree with dark grey bark, oval leaflets in opposite pairs, clusters of white flowers and bean-like seed pods.

California Laurel or **Oregon Myrtle** (*Umbellularia californica*), is a short-trunked North American evergreen, averaging 16m (50ft) with oval leathery leaves, clusters of yellowish flowers and greenish to purple berries. Foliage is pungently aromatic.

Moosewood or **Moosebark** (*Acer pensylvanicum*), of north-eastern North America averages 12m (40ft), with light, white-striped bark, oval to spear-shaped leaves, olive to brownish above, broad-petalled, yellow-greenish flowers and winged fruits.

Hickories (*Carya*) have divided, often palm-shaped leaves, catkins and, usually, rounded nuts. The nuts of some kinds are edible, as are the sap and roots, but do not eat unless the species is positively identified.

(See also *contact poison* in *Poisonous Plants* and *Tropical Poisonous Plants.*)

FUNGI

Fungi make good eating, but MUST be positively identified as of an edible kind. There is no room for error. Unlike vegetable plants, on which the 'edibility test' can be used, they must be either identified or left well alone. The deadly kinds do not taste unpleasant and no symptoms may appear until several hours after eating.

Wild fungi are a great delicacy – if you know which ones to pick. Fungi are composed of many thread-like cells which, in the case of ground fungi, form a subterranean web of which only the reproductive part – what we call the fungus – appears above the ground. This edible part appears only at certain times of year.

Most fungi grow directly from the ground – alone, in rings, sometimes in clumps. They consist of a cup- or bowl-shaped cap topping a stem. The underside bears gills of spongy tissue containing spores. The form and colour of this tissue is an important clue to identify. A few fungi, such as truffles, grow completely underground – but they are very hard to find. Others grow on the sides of trees and stumps; some are known – from their shape – as bracket fungi, others grow large and singularly.

Food value

Fungi come between meat and vegetables in the nutritional table. They contain more protein than vegetables and, in some cases, more fat.

The better kinds, *Boletus edulis* for example, provide a similar amount of calories to the same weight of vegetables. As for minerals: fungi have more phosphorous than carrots, cauliflowers and spinach, but less calcium. Vitamin B complex is present in tiny amounts in most fungi, vitamin C occurs occasionally and D is present in many in appreciable amounts. *Cantharellus cibarius* contains vitamin A.

The great advantage of fungi is their abundance. At the right time of year, usually summer and autumn, you can quickly gather enough for a meal.

Preparing fungi

Reject suspicious, discoloured or maggoty parts, clean, slice and boil. Many bracket fungi are bitter and tough and must be cooked thoroughly. It helps to steep them in cold water first. Tender ground fungi can simply be added to soups and other foods.

Storing fungi

Fungi have high water content and are easily dried. Collect all you can when they are available and add them to your foodstore. Separate caps from stems and place on rocks in the sun, caps gill-side up. With *Boletus* species, first remove the spongy tissue under the cap. When thoroughly dry, store in airtight containers if possible. Eat raw (soak them in water first to let them swell up) or add to soups and stews.

IDENTIFYING AMANITAS

The poisonous Amanitas – and their particularly deadly species the Death Cap and the Destroying Angel – can be mistaken for edible kinds. ALWAYS FOLLOW THESE RULES:

- AVOID any fungi with white gills, a volva (a cup-like appendage at the base of the stem) and stem rings
- AVOID any fungi that are wormy or decomposing
- Unless positively identified – DISCARD

Differences between Amanita and Agaricus fungi:

	AMANITA	AGARICUS
SPORES	White	Purplish-brown
CAP AND STEM	Unchanging	Some kinds stain yellow when bruised
CAP	Slimy with loose patches	Always dry, with only a few small scales
MATURE GILLS	White	Greyish-red, pink or chocolate
GILLS	Partly or wholly veiled	Not veiled
SMELL	Potato or radish	Almonds or marzipan
LOCATION	Never in grassy open areas	Under conifers and hardwoods; in grass and gardens

EDIBLE FUNGI

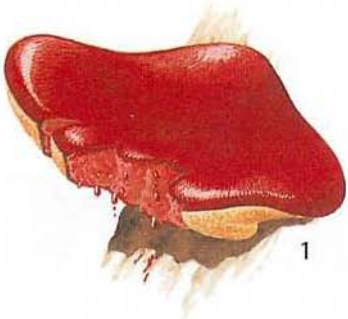

1

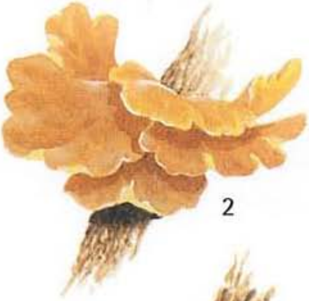

2

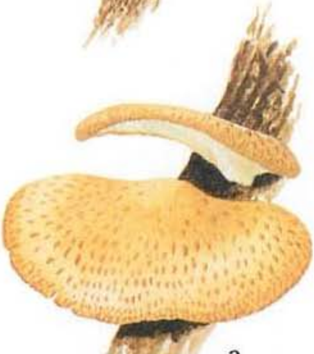

3

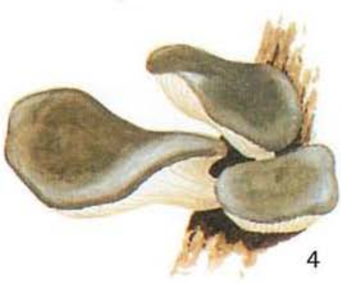

4

There are no reliable rules for identifying fungi, poisonous or edible. Ignore folk tales to the effect that a fungus is not poisonous once peeled, or that toxic kinds change colour when cooked. They do not. Nor does cooking destroy their poisons.

Learn to recognize a small number – those illustrated here – and stick to them. Learn also the Amanita family – they include fungi which can kill you. Then build up your knowledge.

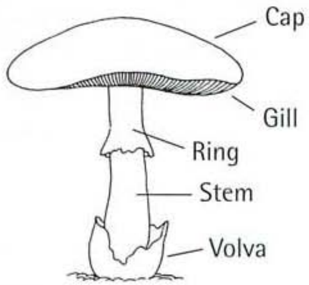

TREE FUNGI

Tree fungi grow off the ground, on the sides of trees and stumps. They are often large and leathery, but are not harmful and are fairly common.

1 Fistulina hepatica *(Beefsteak Fungus) is often found on oaks. It is reddish above, pinkish below and rough-textured, resembling a large tongue; its red flesh exudes a blood-coloured juice. Tough and bitter; young specimens are better. Soak to soften then stew thoroughly. Occurs in autumn.*

2 Polyporus sulphureus *grows to 30–40cm (12–16in) across, a bright orange-yellow fading to yellowish-tan, with spongy, yellowish flesh. On broad-leaved, especially Oak, and evergreen trees from summer on.*

3 Polyporus squamosus *(Dryad's Saddle) grows to 60cm (2ft) across, often* *in clumps, its ochre cap flecked with dark scales to resemble cork and a whitish underside. On broad-leaved trees, especially Elm, Beech and Sycamore, from spring to autumn. Choose young ones and stew them thoroughly.*

4 Pleurotus ostreatus *(Oyster Fungus) grows in clumps, with deep blue-grey shell-shaped caps 6–14cm (2½–5¾in) wide, white gills and white rubbery flesh. On broad-leaved trees for most of the year. Tasty; slice and stew. Also dries well.*

5 Armillaria mellea *(Honey or Bootlace Fungus) has tawny-yellowish brown-flecked caps 3–15cm (1¼–6in) across, white gills later speckled with brown, white flesh and bootlace-like 'roots.' On broad-leaved and coniferous trees and stumps from spring to autumn. Slice and stew.*

GROUND FUNGI

Ground fungi live in the soil. There are many kinds, some VERY poisonous.

6 Lycoperdon giantea *(Giant Puffball) resembles a football, up to 30cm (1ft) across, smooth, white and leathery, yellowing with age and may weigh up to 9kg (20lb). In woods and grassy places from late summer to autumn. Choose young ones with spongy, pure white flesh. Very tasty; simmer or fry.*

7 Cantharellus cibarius *(Chanterelle) is apricot-scented, egg yellow, funnel-shaped, 3–10cm (1¼–4in) across, with pronounced, forking gills. It grows in groups under trees, especially Beech, from summer on. Very tasty; stew for ten minutes. Do NOT confuse with* Cortinarius speciosissimus.

8 Craterellus cornucopioides *(Horn of Plenty) is horn- or funnel-shaped, with a rough, crinkly, dark brown cap 3–8cm (1¼–3in) across and a smooth, tapering grey stem. In broad-leaved woods, especially Beech, in autumn. Stew well, or dry.*

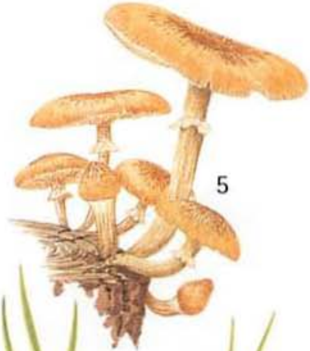
5

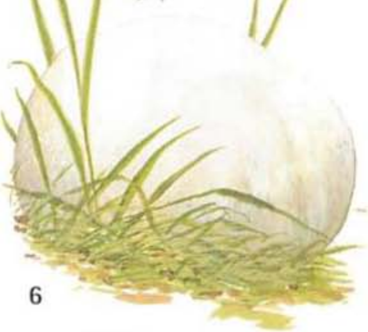
6

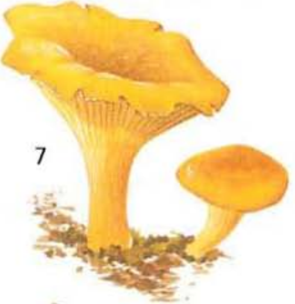
7

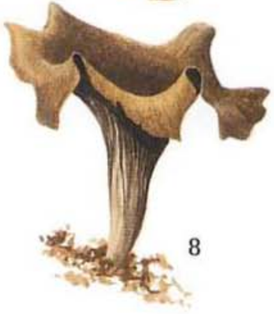
8

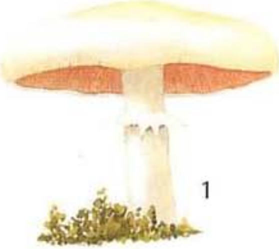
1

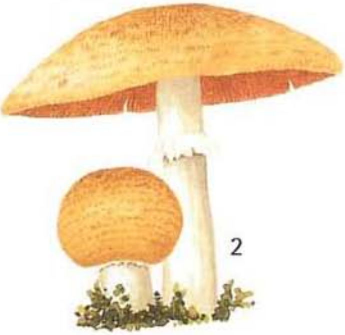
2

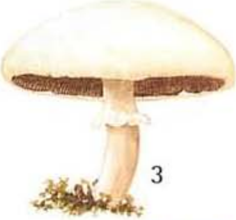
3

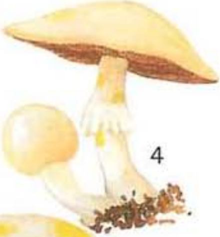
4

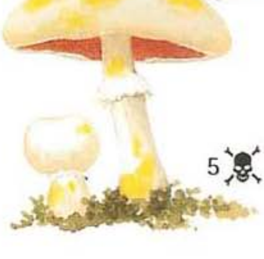
5 ☠

WARNING
Fungi make excellent eating but should be attempted ONLY if identified with certainty.

Agaricus fungi: AVOID any that stain yellow when cut (see A. xanthoderma **below). Some young 'buttons' are hard to tell apart and can be confused with the deadly Amanitas.**

1 **Agaricus arvensis** *(Horse Mushroom) resembles A. campestris but with a cap to 15cm (6in). Young ones have light pink gills, later turning pinkish-brown; in same places as* A. campestris. *Edible raw or cooked.*
2 **Agaricus augustus** *has a scaly, light brownish cap to 20cm (10in) across, young gills pink later turning dark, and a ringed stem; in clusters in woodland clearings in summer and autumn. Tasty; smells of anise.*
3 **Agaricus campestris** *(Field Mushroom) resembles the familiar cultivated kind, with a white cap to 10cm (4in) across, browning slightly in older specimens, and pink gills later turning dark brown; in grassy places in autumn, rarely by trees. Edible raw or cooked.*
4 **Agaricus sylvestris** *(Wood Mushroom) resembles* A. arvensis *but is found in woodland, often with conifers. Edible raw or cooked.*
5 **Agaricus xanthoderma** *(Yellow Staining Mushroom) resembles other Agaricus species, but shows a yellow stain when bruised and is strongly yellow at the base. It is POISONOUS and smells of carbolic. In both woody and grassy places in summer and autumn. AVOID.*
6 **Lepiota procera** *(Parasol Mushroom) has a brownish cap, later with darker scales, to 30cm (1ft) across, with creamy-white gills and a slender stem with a double*

white ring and brown bands. By broad-leaved woodland and in grassy clearings from summer to autumn. It tastes of almonds or Brazil nuts.

7 Coprinus comatus *(Shaggy Ink Cap) has a cylindrical cap with whitish or pale brownish scales and gills that begin white, turn pinkish and finally dissolve into a black, inky mess. In groups in open grassy areas in summer and autumn. Gather young ones whose gills are still pale. POISONOUS if eaten with alcohol.*

8 Tricholoma nudum *(Wood Blewit or Blue Cap) has a lilac-blue cap, later turning reddish-brown and wavy-edged, to 10cm (4in) across, bluish gills and a stocky, fibrous bluish stem. In rings in mixed woodland from autumn to mid winter. Tasty and sweet-smelling. Produces an allergic reaction in a few people.*

9 Boletus edulis *(Cep) is brownish with a cap to 20cm (8in), a swollen stem and white flesh; in woodland clearings in autumn. All Boletes have a sponge-like layer of pores or tubes instead of gills. Many edible species – they dry well – occur around the world. AVOID any with pink or red spores unless positively identified. Some are poisonous.*

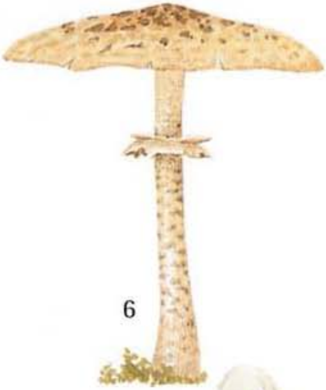
6

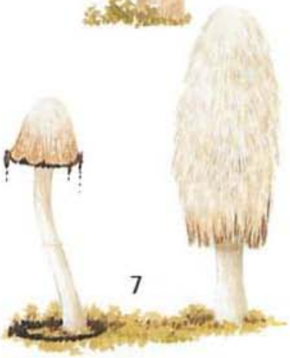
7

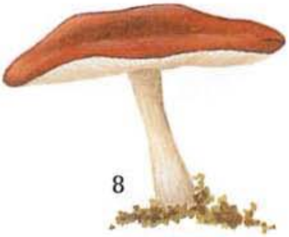
8

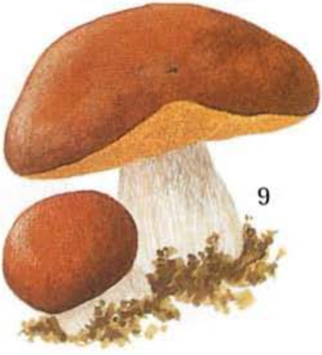
9

OTHER USES FOR FUNGI

- Many bracket fungi make excellent tinder – once lit they will smoulder for hours.
- Razor-strop fungus is so tough that it can be used to sharpen knives, or chopped up for corks, corn plasters and kindling.
- Giant puffball is styptic – it will staunch and soothe bleeding wounds.
- Tree fungi are rich in tannin and can be used in treating burns.

POISONOUS FUNGI

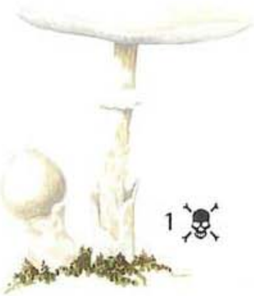
1 ☠

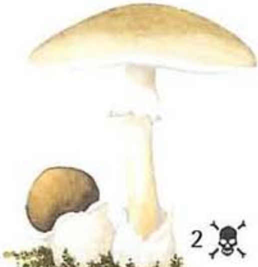
2 ☠

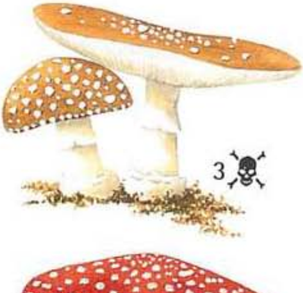
3 ☠

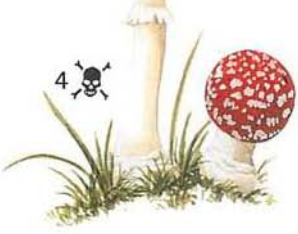
4 ☠

WARNING
The following are among the worst of the poisonous fungi but there are many others. Do NOT use any fungi you cannot positively identify as a survival food. Some Amanita fungi are among the most deadly of all. They have a cup, or volva, at the base.

1 **Amanita virosa** *(Destroying Angel) is wholly white, with a large volva, a scaly stem and a cap to 12cm (5in) across, in woodland in summer and autumn. Sweet- and sickly-smelling and DEADLY poisonous. Young ones may resemble young Agaricus fungi.*

2 **Amanita phalloides** *(Death Cap) has a greenish-olive cap to 12cm (5in) across, a paler stem, large volva, and white gills and flesh; usually in woodland, especially with Oak or Beech. The MOST DEADLY of all.*

3 **Amanita pantherina** *(Panther Cap) has a brownish, white-flecked cap to 8cm (3in), white gills and 2-3 hoop-looking rings at the base of the stem; in woodland, especially with Beech. Poisonous, often FATAL.*

4 **Amanita muscaria** *(Fly Agaric) has a distinctive bright red cap, flecked with white, to 22cm (9in) across. Found in autumn, typically in Pine and Birch woods.*

5 **Entoloma sinnuatum** *(Leaden Entoloma) has a dull greyish-white, deeply convex cap to 15cm (6in) across, yellowish gills turning salmon-pink and firm white flesh smelling of meal, bitter almond and radish; in groups in grassy places and woods, especially with Beech and Oak, in summer and autumn. Poisonous, can be DEADLY. Confusable with an Agaricus, but has no ring on the stem.*

6 **Inocybe patouillardii** *begins whitish then turns yellowish-brown, with a cap to 7cm ($2^{3}/_{4}$in), often split at the margin, and whitish gills turning olive-*

brown; stains red when bruised. In broad-leaved woods, especially Beech, in summer and autumn. Lacks a ring on the stem, but when young confusable with an Agaricus. DEADLY poisonous.

7 Paxilus involutus *has a solid yellow-brown cap with a rolled rim, to 12cm (5in) across, yellow-brown gills and a straight, stout stem. Very common in woodland, especially Birch. DEADLY: do NOT confuse with edible yellowish fungi such as the Chanterelle.*

8 Cortinarius speciosissimus *is reddish- to tawny-brown, with a flattish cap 2–8cm ($\frac{3}{4}$–3$\frac{1}{4}$in) across, and rusty-brown gills, in coniferous wood in autumn. Not common but very poisonous. Slightly lighter-coloured,* C. orelanus, *also poisonous in broadleaved woodland. Both have a radish-like smell. Do not confuse with the Chanterelle. DEADLY.*

(See also Agaricus xanthoderma, illustrated with edible species)

POISON SYMPTOMS

Poisonous fungi produce a variety of symptoms but the following are fairly typical:

Death Cap/Destroying Angel: *Symptoms develop slowly, 8–24 hrs after eating: vomiting, diarrhoea, excessive thirst, sweating and convulsions. Apparent recovery after one day, then a relapse and, in 90% of cases, death from liver failure in 2–10 days. No known antidote.*

Muscarine poisoning: *Caused by several fungi. Effects vary, toxins all attack nervous system.*

Amanita muscaria *produces severe gastro-intestinal disturbance, delirium, vivid hallucinations, uncontrollable twitching and convulsions, followed by coma-like sleep. The victim usually recovers.* Inocybe patouillardii *and its relatives produce vertigo, blindness, sweating, low temperature, very dilated pupils, followed, in the worst cases, by delirium and death.*

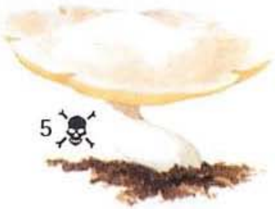

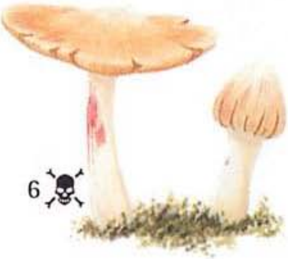

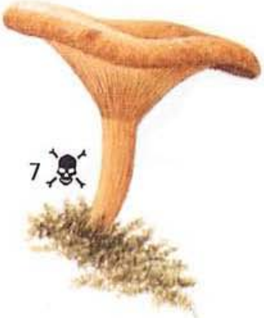

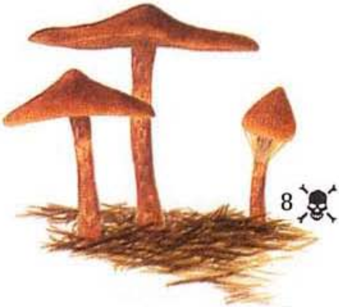

ARCTIC AND NORTHERN PLANTS

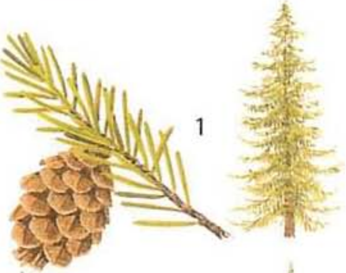

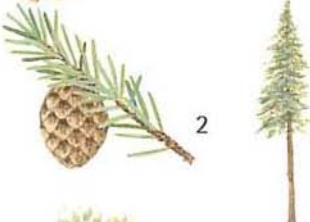

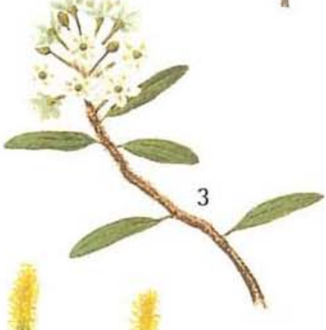

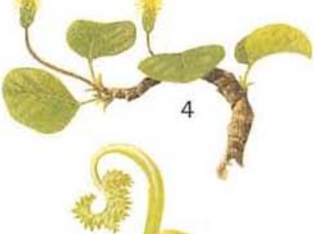

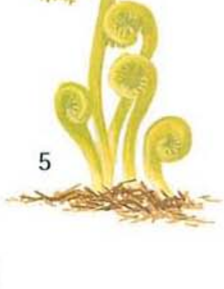

In addition to these hardy arctic plants, many temperate species occur in summer in the far north.

1 **Red Spruce** *(Picea rubens) reaches 23m (70ft), with dark or yellow-green needles all around its hairy twigs, rough dark bark and pendant cones; in drier areas of North America. Young shoots are edible raw or cooked; infuse the needles for teas and boil the edible inner bark.*

2 **Black Spruce** *(Picea mariana) is smaller than Red Spruce, with shorter needles; in moist areas of North America. Many similar spruces occur in North America and northern Eurasia. Use all as Red Spruce.*

3 **Labrador Tea** *(Ledum groenlandicum) is a fragrant evergreen shrub averaging 30–90cm (1–3ft), with narrow leaves with rolled edges, whitish or hairy below, and five-petalled white flowers, in North America. The leaves make a revivifying tea.*

4 **Arctic Willows** *(Salix) are mat-forming tundra shrubs averaging 30–60cm (1–2ft), with rounded, leaves, shiny above, and yellow catkins. Spring shoots, leaves, inner bark and young peeled roots are all edible. The leaves have 7–10 times more vitamin C than an orange.*

5 **Ferns** *occur in moist places in summer in far northern woods and by the tundra. Eat ONLY young fiddleheads up to 15cm (6in) long, remove any hairs. Steaming is the best way to cook them.*

6 **Cloudberries** *(Rubus chamaemorus) seldom reach 30cm (1ft). Bramble-like, with palm-shaped leaves, white flowers and berries at the top of the plant – pink, ripening through orange to amber. The berries are edible raw.*

7 **Salmonberry** *(Rubus spectabilis) resembles a small Wild Raspberry,*

thornless, with three-part leaves, purplish-red flowers and juicy red or yellow berries, edible raw. In North America, now in parts of Europe.

8 Bearberry *(Arctostaphylos uvaursi) in arctic regions is small, mat-forming, woody, with club-shaped evergreen leathery leaves, pink or white flowers and clusters of red berries, edible cooked.*

9 Iceland Moss *(Cetraria islandica) is a lichen, forming tufted leathery, grey-green or brownish mats up to 10cm (4in) high, composed of many strap-shaped branches. Soak for several hours then boil well.*

10 Reindeer Moss *(Cladonia rangiferina), lichen growing 5–10cm (2–4in), often in large clumps, with hollow, roundish, greyish stems and branches resembling antlers. Soak for several hours, boil well.*

11 Rock Tripes *(Umbillicaria) are lichens, forming roundish, blister-like greyish or brownish growths attached to rocks by a central stalk; some kinds warty and pebble-like, others smooth. Very nutritious, soak for several hours then boil well.*

Lichens are an important survival food, not only in arctic regions – though vital there because they are probably more nutritious than other arctic plants. Explorers have survived on Rock Tripe for long periods. But beware! They can cause painful irritation if eaten raw because they contain a bitter acid. Remove it by soaking them in water overnight and then boiling thoroughly. A further roasting will crispen them up if you prefer a crunchy texture.

If you successfully hunt a caribou eat the fermented lichens in its stomach. They are easily digested and some Eskimos count them a great delicacy.

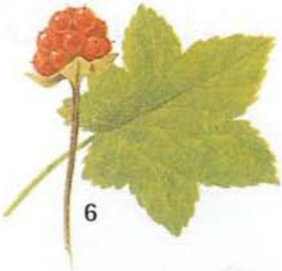
6

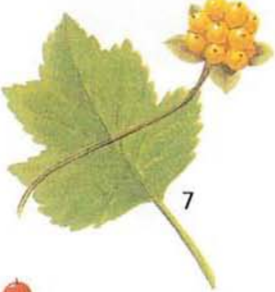
7

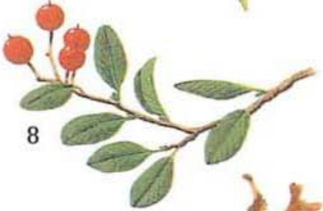
8

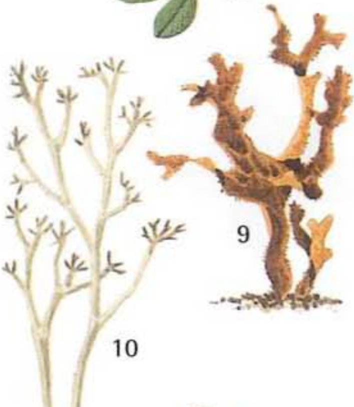
9

10

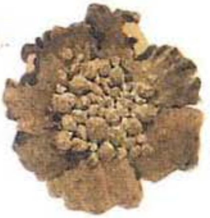
11

DESERT PLANTS

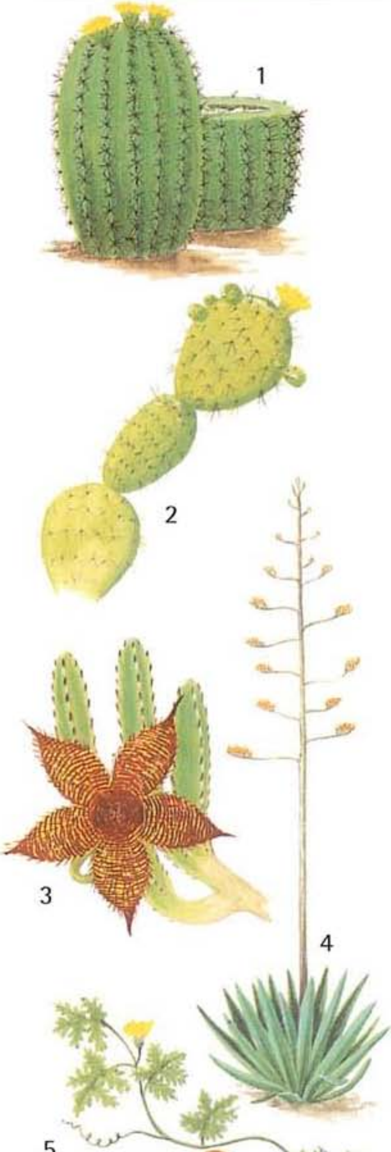

Desert survival depends upon water. Learn the water-bearing trees and cacti before considering food plants. Do not eat if you have no water; digestion will further deplete the body's liquid reserves and accelerate dehydration.

1 **Barrel Cacti** *(Ferocactus) of south-west North America average 1·2m (4ft), at that size yielding over 1 litre (2 pt) of edible, sometimes bitter, milky sap. This is an exception to the rule to avoid milky sap. Slice off the top and smash the inner pulp, then drink. Only worth doing in an emergency.*

2 **Prickly Pears** *(Opuntia) have thick, jointed, pad-like leaves, yellow or red flowers and egg-shaped pulpy fruits. Peeled fruits are edible raw; peel and cook tender young pads – cutting away the spines, roast seeds for flour, tap the stems for water. Originally from North America, now in many arid areas. Very prickly, treat with great care. In Africa do NOT confuse with spurges; unlike this plant, they have milky sap.*

3 **Carrion Flowers** *(Stapelia) are found in some variety in southern and tropical Africa. They are large odd-looking plants with short succulent stems that branch off into leaves like fat spines, with distinctive star-shaped flowers, which may be covered in thick shaggy hairs. Mature flowers give off a powerful stench of rotting meat, an unmistakeable, if not very pleasant clue to identification. Tap the stems for their water.*

4 **Mescals** *(Agave) have a rosette of thick, leathery, spiky leaves with a sharp tip from which arises a very long, columnar flower stalk. Stalks not yet in flower are edible cooked. In Africa, Asia, southern Europe, Mexico and the southern United States, and parts of the Caribbean. Grows in moist tropical areas as well as desert country.*

5 **Wild Gourds** *(Cucurbitaceae) occur in the Kalahari and Sahara east to*

India, cultivated elsewhere. The plant is mat-forming, resembling a vine, with orange-sized fruits. Boil the unripe fruit to make it more edible, roast the seeds, cook young leaves. The flower may be eaten raw, and stems and shoots chewed for their water.

6 Date Palms *(Phoenix) grow wild, always near water, from India to North Africa; introduced elsewhere. They are tall slender palms crowned with a tuft of leaves up to 4m (16ft) long. The fruits and growing tip of the palm are edible raw young leaves are edible cooked. Sap from the trunk is rich in sugar and can be boiled down.*

7 Baobabs *(Adansonia) are large trees with huge, swollen, heavily ridged trunks found from Africa to Australia. The trunk of a mature tree can be 9m (30ft) in diameter. Tap the roots for water. The pulpy fruits, 10–20cm (4–8in) long, and seeds are edible raw; boil tender young leaves.*

8 Acacias *(Acacia) occur abundantly from Africa to northern Australia. There are many different kinds, all thorny, scrubby, medium-sized trees with very small leaflets their flowers usually forming small (1cm/½in) globular flowerheads, white, pink or yellow, according to species. Their roots may be tapped for water, the seeds roasted and young leaves and shoots boiled.*

9 Carob *(Ceratonia siliqua) grows in arid land around the Mediterranean and from the Sahara across Arabia to India. It is almost the only tree on Malta. Height can reach 15m (50ft). Shiny, evergreen leaves are paired, two or three to a stem. Small red flowers produce flat leathery seed pods containing a sweet, nutritious pulp, which can be eaten raw, and hard brown seeds which can be ground and cooked as porridge. Also known as St John's or Locust Bread because these pods were thought to have been the 'locusts' on which John the Baptist lived in the wilderness.*

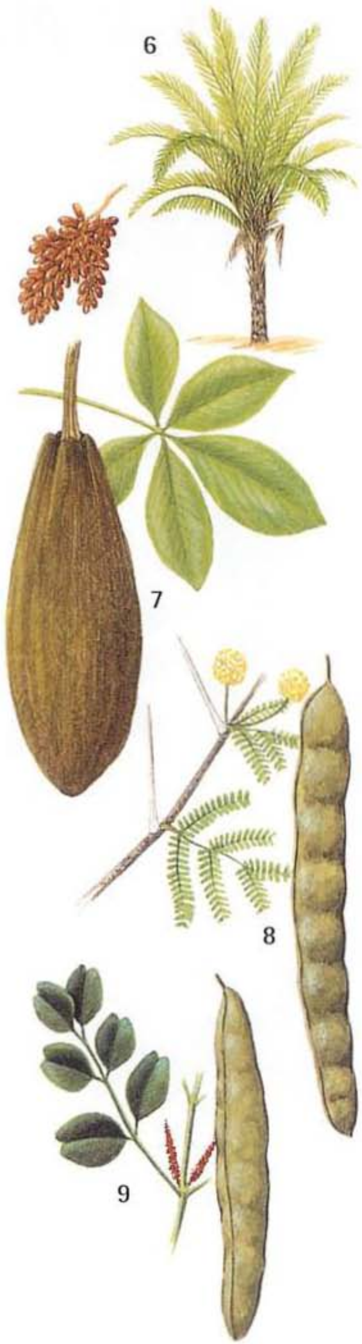

TROPICAL PLANTS

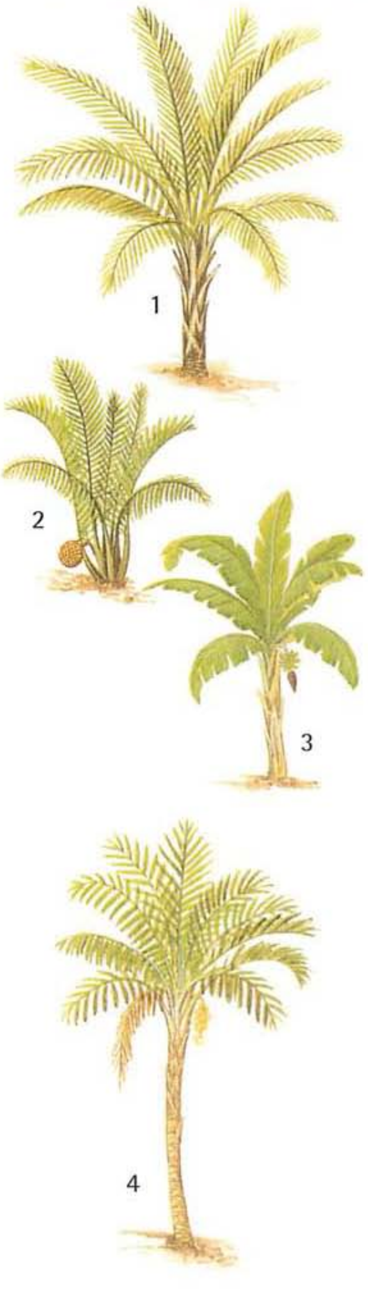

Edible plants abound in tropical areas, but unless you are already familiar with them it is better to begin with to eat palms, bamboos and the commoner fruits.

1 Sago Palms *(Metroxylon) come from the damp lowlands of southeast Asia; introduced elsewhere. Medium, to 10m (30ft), spiny-trunked; leaves long, delicate and arching. The spongy, starchy inner pith of the trunk provides the sago.*

2 The Nipa Palm *(Nifpa fruticans) grows to 6m (20ft), with long, fern-like leaves swelling and clustering at the base to form the 'trunk'; in brackish estuaries in South-east Asia. It yields sugary sap, delicious fruits and an edible 'cabbage' or growing tip.*

3 Bananas or Plantains *(Musa) are spread all over the tropics: 3–10m (9–30ft) tall, with very large, strap-like, usually split light green leaves. Hard plantain fruits are only edible cooked. They have more starch but less sugar than the soft kinds. Eat buds, growing tips, young stems and inner parts of roots as well as fruit.*

4 Sugar Palm *(Arenga pinnata) reaches 12–20m (38–65ft), with a rough scaly trunk topped by fairly erect, Sago-like leaves and yellow, branch-like fruiting parts; wild in Malaysia and Indonesia. To obtain sugar, collect sap and boil down to a thick syrup.*

5 Rattan Palms *(Calamus) are robust climbers, hooking themselves on to trees with a fish-hookshaped appendage on the midrib of the leaves; often common in Old World tropics. The swollen stems, seeds and growing tip are edible; the roots make excellent lashings.*

6 Fish-tail Palms *(Caryota) average 10m (30ft), with a smooth, ringed trunk and long arching leaves of many oval or wedged-shaped leaflets. There* *are many similar kinds in the Old World tropics. Use as Sago Palm. Do NOT eat fruit.*

7 Coconut Palms *(Cocos nucifera) occur all over the moist tropics, growing to over 30m (90ft), with large clusters of nuts hanging at the base of the leaves. The fibrous 'coconut' is inside a large, smooth husk. The growing tip, milk and flesh of the nut are edible. The rich sap can be boiled down for sugar.*

These are only some of the tropical palms. Others include:

- **Piva Palm** *(Guiliema utilis) of America, has a slender trunk banded with alternating dark and light spines. Boil or roast the red or yellow fruits.*
- **Baccaba and Patawa Palms** *(Jessenia and Oemocarpus) of Brazil and the Guianas, have small purplish fruit 2cm (3/4in) long. Eat both pulp and kernel of seed.*
- **Assai Palm** *(Euterpe oleraccea) of tropical South America, which likes swampy places, especially along tidal rivers. Edible fruit has soft purple pulp.*

The growing tip, enclosed by a crown of leaves or the sheathing bases of the leaf stems, is edible in most palms – eat any that are not too bitter. Fruit should be AVOIDED if not positively identified, especially in the Far East – some contain crystals which cause intense pain.

8 Papaya or Pawpaw *(Carica papaya) is a small tree, 2–6m (6–18ft) tall, with a soft hollow trunk and large, dark green, melon-like fruits that ripen to orange or yellow; in moist conditions in all tropical parts. The fruit is edible raw and settles an upset-stomach; young leaves, flowers and stems are edible boiled. Change the water AT LEAST once. Although it will tenderize tough, stringy meat, do not get the milky sap of unripe fruit in the eyes.*

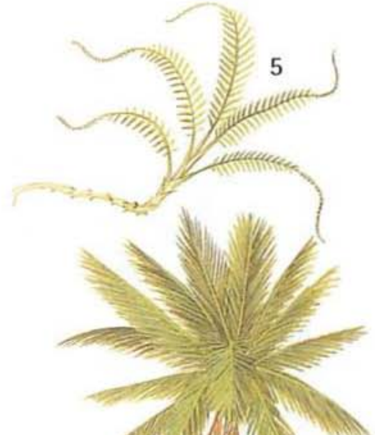

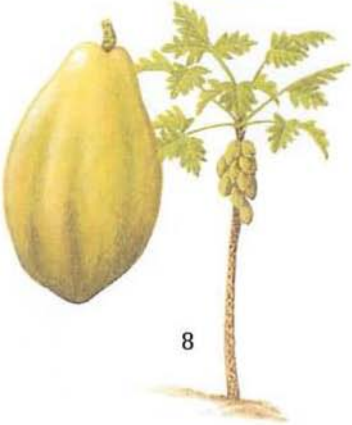

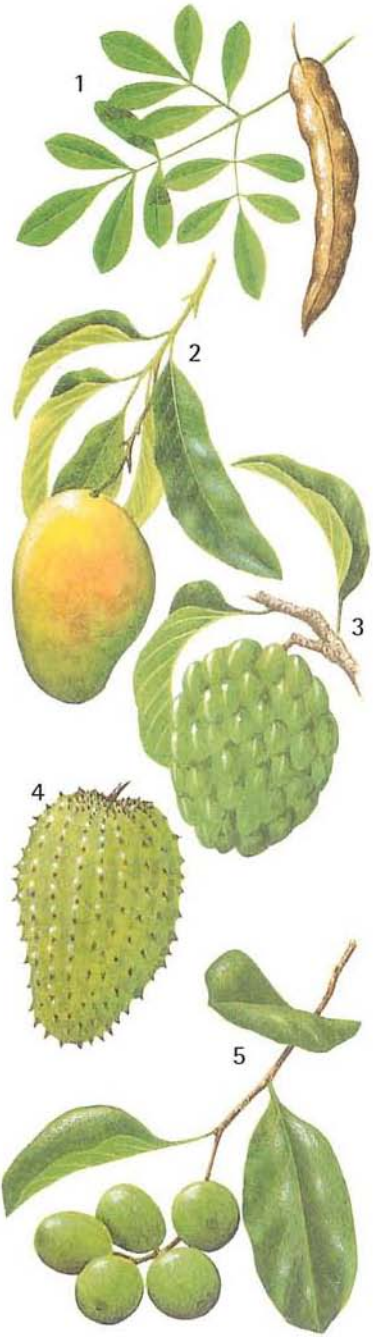

1 **Horseradish Tree** *(Moringa oleifera) grows to 10m (30ft), with small oval leaflets, yellow-white flowers and 25–38cm (10–15in), narrow, three-sided brown seed pods; wild in East Africa and South Asia but now elsewhere. Leaves and young fruits are edible raw or cooked; slice older pods and cook like string beans; use the root for seasoning like true Horseradish. The expressed juice of leaves and roots is good for treating inflammations.*

2 **Mango** *(Mangifera indica) grows in moist places almost everywhere, a medium to large evergreen tree with clusters of narrow dark green leaves. The oval, 7·5–13cm (3–5in) fruits ripen from green to orange, are edible raw and contain a long flat stone. Mango leaves can cause an allergic reaction in some people.*

3 **Sweet Sop** *(Annona squamosa) is a 5–6m (15–18ft) tree with oval to spear-shaped leaves and Magnolia-like flowers. The aromatic, pulpy, globe-shaped green-grey fruit is made up of many parts. Originally from the New World tropics but now in most parts; two edible relatives in Central and South America are the Cherimoya (*A. cherimola*) and Bullock's Heart (*A. reticulata*).*

4 **Sour Sop** *(Annona muricata) is similarly widespread, growing to 12m (38ft), with large Avocado-shaped fruits, green, leathery, spiny, weighing up to 2kg (4·4lb). A very refreshing acid taste.*

5 **Wild Figs** *(Ficus) exist in great variety in tropical and subtropical areas, a few species in deserts. They are straggly trees, with aerial roots and leathery evergreen leaves rounded at the base. The pear-shaped fruits are edible raw and grow direct from the branches. Avoid any that are hard and woody or with irritant hairs.*

6 **Breadfruits** *(Artocarpus) are trees, growing to 15–20m (48–56ft), with*

large, deeply lobed leaves, glossy above, milky sap and very large warty-skinned fruits; now in most tropical parts. The starch-rich fruits are edible raw if the skin is scraped away and the tougher inner bits discarded.

7 **Sterculias** *(Sterculia) are found in Central and South America, and elsewhere in the tropics. Large trees, up to 30m (100ft) high, with buttressed roots and hand-shaped leaves, their fruits consist of pods containing black, peanut-like seeds that can be eaten raw after removing the irritant hairs.*

8 **Bael Fruit** *(Aegle marmelos) is a tree of 2·5–4·5m (7–15ft), with dense prickly growth and rounded, yellowish or greyish fruits up to 10cm (4in) across; in parts of the Himalayas, India and Burma. A relative of the citruses, the fruits are edible raw and very rich in vitamin C.*

9 **Bignays** *(Antidesma) are evergreen shrubs, 10–13m (30–40ft) tall, with shiny 15cm (6in) leaves, in the forests of South-east Asia. The fleshy, many seeded, currant-like fruits are about 1cm (½in) across and mature from green to white to red and finally black. Edible raw, but better jellied.*

In primary tropical forest most fruits are carried in the canopy, unreachable unless you climb the trees or chop them down, though elsewhere fruit can prove a valuable food. More accessible is the vine-like Rattan, which can be cut and then hauled down to obtain the growing tip. Peel off the outer sheath for about 2m (7ft), cut into lengths and roast in the embers of a fire. The inner heart of some is very tasty, although slightly bitter.

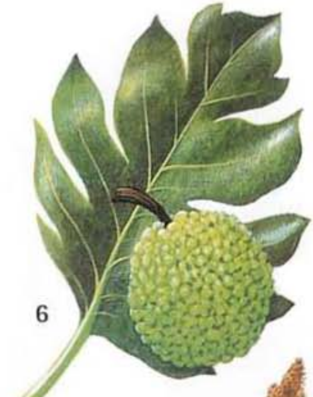

6

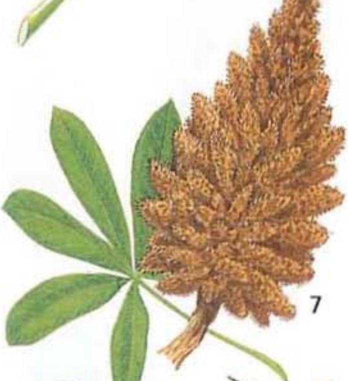

7

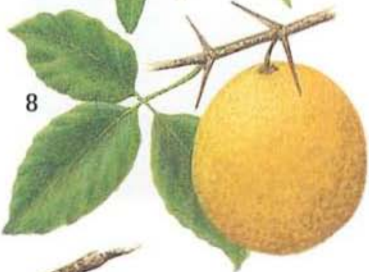

8

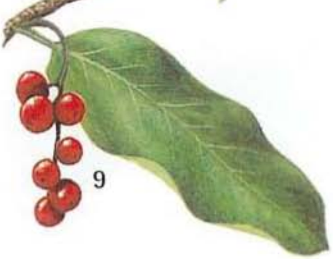

9

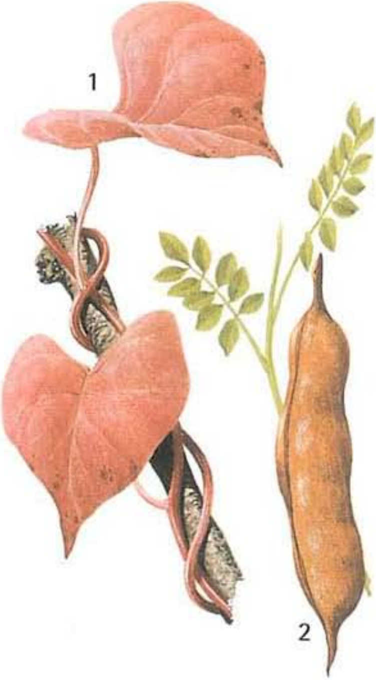

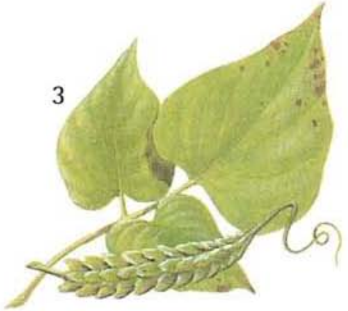

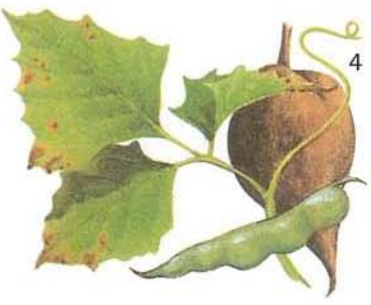

1 Ceylon Spinach *(Basella alba) is trailing, vine-like, with thick circular to oval or heart-shaped leaves, from greenish to purplish-red, and fleshy purplish flowers; now in most tropical areas. Young leaves and stems are edible cooked and rich in vitamins.*

2 Tamarind *(Tamarindus indica) is a densely branched tree to 25m (80ft), with evergreen leaves formed of many leaflets, pale yellow red-streaked flowers and brown seed pods. The pulp of the pods is edible raw and rich in vitamin C. Seeds and young leaves can be used as a potherb, and the bark peeled and chewed. Now in many areas.*

3 Goa Beans *(Psophocarpus) are climbing plants of the Old World tropics, with spear-shaped leaves, blue flowers and 20cm (8in) long, four-angled, scaly seedpods. Boil young pods lightly, they taste like beans, the young seeds like peas, or roast older seeds. Young leaves, which taste like spinach, are edible raw. Thicker roots, richer in protein than potato or cassava, are edible raw or can be boiled, fried, baked or roasted like potatoes.*

4 Yam Beans *(Pachyrrhizus and Sphenostylis) are climbing plants with irregular, three-part leaves and a knotty turnip-like root, in large patches in most of the tropics. The edible tubers are crisp, sweet, juicy, tasting of nuts. The seeds are harmful raw and must be well boiled.*

5 Peanut *(Arachis hypogaea) is not a true nut and its fruits ripen underground. A small, bushy plant with pairs of bluntly oval leaves, yellow flowers and stalks leading to the wrinkled pods. Very nutritious, the 'nuts' keep well. Now in most tropical and subtropical parts.*

6 Cassava or Manioc *(Manihot esculenta) is a staple throughout the tropics, growing on well-drained ground to 5m (15ft) with jointed stems* *and long-stalked leaves divided into 5–9 blades. The fat tubers are lethally POISONOUS raw and MUST be cooked. Keeps well, particularly when ground as flour.*

7 Ti Plant *is a shade-loving shrub, 2–5m (6–15ft), with whorls of shiny, leathery, sometimes reddish leaves and, when ripe, red berries; now widespread in tropical parts particularly Pacific islands. Boil the starch-rich fleshy roots.*

8 Taro *(Colocasia) occurs on wet ground all over the tropics, up to 1·5m (5ft), with large, very long-stalked heart- or arrow-shaped leaves arising from the roots and an orange-yellow flower. The tubers taste like potatoes but are harmful raw and MUST be cooked.*

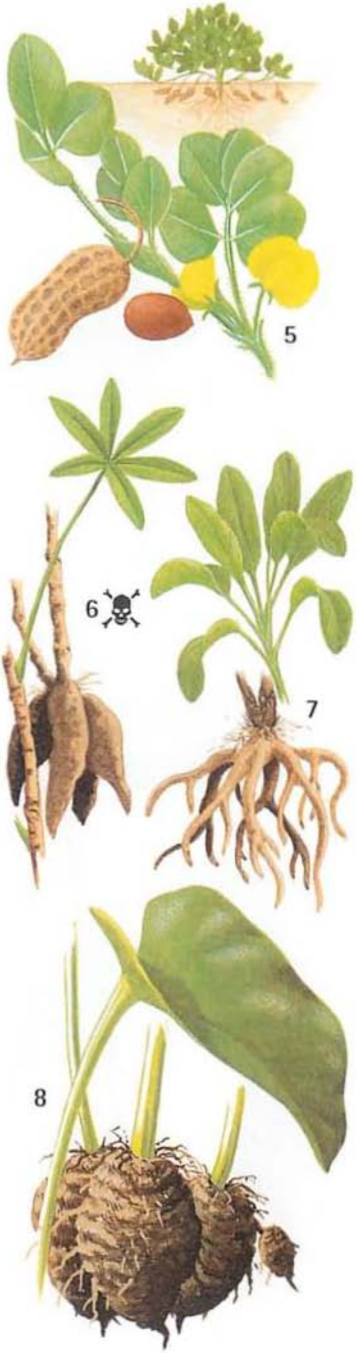

In tropical forests the best places to forage will be along streams and watercourses. Anywhere that the sun can penetrate the jungle floor will produce a mass of vegetation but river banks often offer the clearest area, where plants can thrive.

While climbing spurs take the chance to look out over the jungle canopy below. Good views are obtained where a deadfall has crashed down and cleared an area. Palms and other recognizable food plants can be seen among the canopy. Their position can be noted and they can then be felled when required. A tree will provide a considerable supply of food.

If weak and short of food do NOT expend effort felling a tree. You will use up too much energy and there will be easier food available.

Do not pick more food than you need. Food deteriorates rapidly in tropical conditions. Leave food on the growing plant until you need it and eat if fresh.

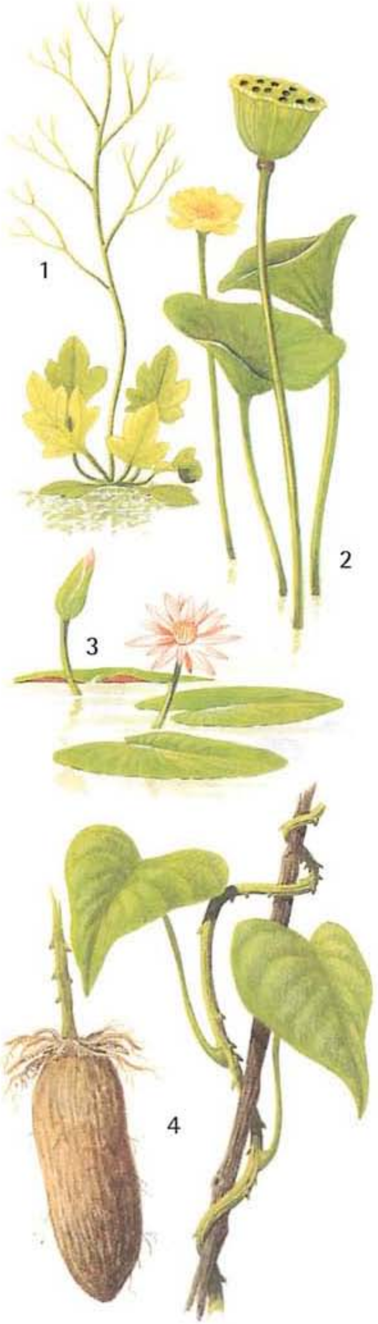

1 Water Spinach *(Ipomoea aquatica) is trailing, with light green leaves and white flowers; always by fresh water, usually as a floating plant, mainly in Southeast Asia. Older stems are stringy, but young leaves and shoots can be boiled.*

2 Lotus *(Nelumbium nuciferum) is aquatic, with long-stalked, bell-shaped bluish-green leaves standing clear of the water and pink, white or yellow flowers. Young leaves and peeled stems are edible boiled. Boil or roast ripe seeds, first removing the bitter embryo, and the rootstalk. Mainly Asia, and in parts of Africa and North America.*

3 Water Lilies *(Nymphaea) grow in lakes, rivers and streams in tropical Africa, India and America, and also in some temperate areas, with heart-shaped leaves floating on the water. They have large, starch-rich edible tubers, stems which can be cooked, and bitter but nourishing seeds.*

4 Wild Yams *(Dioscorea) occur in great variety in light forest and clearing in both tropical and subtropical parts. Twining, vine-like stems, some kinds bearing edible aerial tubers, lead to one or more large underground tubers. If kept dry, yams store well. Some Wild Yams are poisonous raw: for safety ALWAYS cook – peel tubers, boil and mash.*

5 Wild Rice *(Oryza* and in North America *Zizania) is a coarse grass growing to 90–120cm (3–4ft), widespread in the tropics and many temperate areas. Thresh and winnow the grains to remove the tough, hairy husks, then boil or roast and pound for storage as flour.*

6 Sugarcanes *(Saccharum) are cultivated all over the tropics and occasionally occur wild. A coarse, tall, aromatic, thick-stemmed grass, the canes can be chewed raw to extract their sweet juice.*

7 Millets *(Panicum, Pennisetum and others) are tropical cereals, sometimes found wild in drier areas. They are grasses, several feet tall, with sausage-like heads of grain, each one about the size of a mustard seed. Pound to a meal and use in stews or as porridge.*

8 Bamboos *are giant grasses of most moist areas. The rapidly growing, edible young shoots are at the base of the plant; split the tough outer sheath and cook like asparagus. The seeds of a flowering bamboo are also edible. Bamboos have a great many uses as building materials, for making rafts, even as cooking utensils, as well as their use as food. Take care when collecting: some plants are heavily stressed and may shatter or whiplash.*

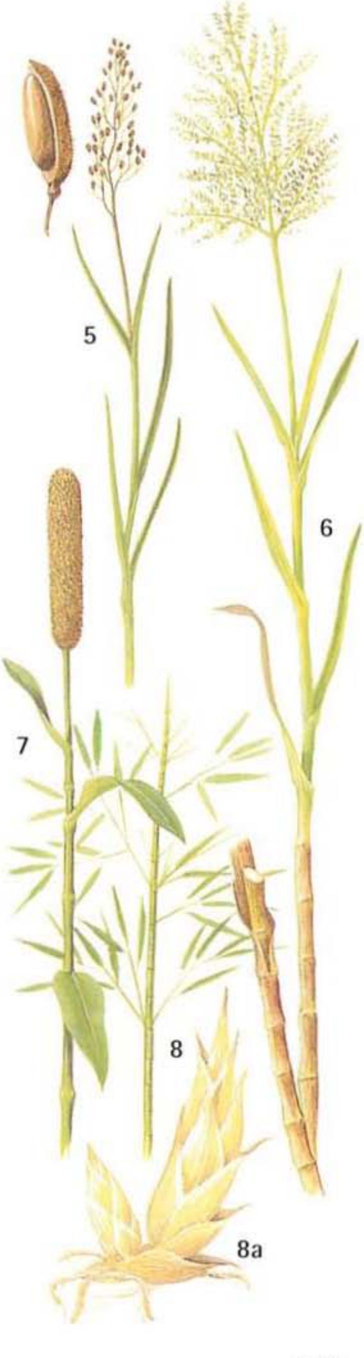

In addition to the plants illustrated and listed here there are others that you will recognize from their similar cultivated varieties, such as the Avocado (*Peresea americana*) of tropical South America, where it is a favourite food of many animals – including the jaguar! – the Christophine (*Sechium edule*), native to Brazil, and the whole family of citrus fruits, with about 60 wild species in Indo-Malaya and China, some growing at quite high altitudes. But BEWARE. Strychnine has fruits that look like oranges but which are DEADLY POISONOUS.

In the tropics you will usually have a choise of plant foods so stick to the ones you can identify and know are safe. If you need to eat unknown plants always apply the edibility test, using very small amounts.

These are some of the many other edible plants which you might find, all are good to eat if prepared as suggested:

PALMS

Palmyra *or* **Borassus palm** *(Borassus flabellifer), growing in drier parts of tropical Africa, southern India and Burma, resembles a Date Palm, with similar fan-like leaves. Sort insides of the three-part nuts are edible, sap makes a pleasant drink and flowering parts can be tapped for their sugary juice.*

Saw palmetto *(Serenoa repens), common on sandy ground from southern USA into Central America. Average 1–2m (3–7ft) tall, with long, stiff fan-shaped groups of leaves, clusters of white flowers and oblong black fruits. Growing tip (palm heart) edible raw or cooked.*

Yuccas *(Yucca) of many kinds grow on sandy ground, from USA into Central America. Evergreen, woody-stemmed, with stiff sword-like leaves off a stem or in a cluster at the top. Petals of six-petalled white flowers and ripe pulp of oblong gerkin-like fruits are edible raw or cooked.*

NUTS

Brazil Nut *(Bertholletia excelsa) is widespread on dry ground in forests of tropical South America, 30–40m (100–130ft) high, with long, crinkly margined, oval leaves and yellow flowers. Pot-shaped, lidded capsules contain 15–30 edible nuts. In the same area Sapucaya (*Lecythis urnigera*) nuts come in similar capsules and are equally tasty.*

Cashew Nut *(Anacardium occidentale), originally native to north-eastern South America, now grown all over tropics, are shrubs or medium-sized trees with evergreen, broadly oval, leathery, prominently veined leaves in opposite pairs and pinkish-yellow flowers. Eat ONLY nuts, borne on the end of a pear-shaped, reddish, fleshy fruit stalk. Harmful unless peeled and cooked. BEWARE: smoke or steam from cooking can cause blindness.*

Indian Almond *(Terminalia catappa), found in scrubby, especially coastal, area over almost all the tropics, has very large, leathery, club-shaped leaves. Mature trees spread to 10m (30ft), young ones have distinctive circles of horizontal branches. Hard edible nuts in a fibrous, thick green, fleshy shell, cluster at the branch tips.*

Queensland Nut *(Macademia ternifolia) grows in north-eastern Australia on a tree about 12m (40ft) tall with many narrow, spear-shaped leaves. Round nuts in greenish husks come in clusters.*

FRUITS

Guava *(Psidium guajava) of Central and South America, widely introduced elsewhere, are small gnarled trees, up to 10m (30ft), with peely, light brown bark, oval leaves in pairs (slightly hairy below), white flowers and large crabapple-like light yellow fruits with a whitish to pinkish, creamy, seed-filled pulp. Edible raw or cooked and rich in vitamin C.*

Persimmon *(Diospyros) found in waste places and dry woods of warm temperate east Asia and southern USA, introduced elsewhere, can reach 20m (63ft). Small, sometimes crinkly, spear-shaped leaves make good tea, rich in vitamin C. Round, yellow to reddish or purplish, tomato-like, 5–7·5cm (2–2½in) fruits are edible raw or cooked.*

Rambutan *(Nephelium lappaceum) grows in forests in parts of Southeast Asia reaching 20m (63ft). Bushy, with dark brown bark, many small dark green leaflets, small greenish flowers and clusters of small, soft-spined, hairy-looking fruits ripening to red.*

Fleshy white inner pulp – not seed – is edible raw and rich in vitamin C. Pulasan (N. mutabile*) is a smaller tree in the same area with equally good blunt-spined fruits with a yellowish pulp.*

Durian *(Durio zibethinus), 40m (130ft) trees, native to Malaysia and Borneo, widely introduced in south-east Asia, with long, spear-shaped, bronze to olive green leaves and large, spiky-surfaced, distinctive yellow fruits. Cream-coloured pulp (not seeds) is foul smelling but delicious to eat.*

Malay Apple *or* **Pomerac** *(Syzygium malaccensis) reaches 15m (50ft), with glossy green, broadly spear-shaped leaves, clusters of bright pink flowers and smallish 5cm (2in) red, apple-like fruits, smelling of roses and edible raw.*

Rose apple *(S. aqueum) is similar but with yellowish or white pear-shaped fruits tasting of apricots. Both native to rain forest and scrubby places in Malaysia, widely introduced elsewhere.*

Carambola *(Averrhoa carambola) grows in Indonesian forests. Small, average 8m (24ft) height, with many light green, small, spear-shaped leaves, clusters of small whitish-pink flowers and yellow, ribbed fruits 7·5–12·5cm (3–5in) long, edible raw, though sometimes acid.*

Mangosteen *(Garcinia mangostana) of south-east Asian forests, reach 12m (40ft), with long, leathery, dark green leaves in pairs, large yellow and purple flowers and round, tomato-shaped purplish-brown fruits whose five segments of inner white pulp are edible raw. There are several kinds.*

Passion Fruit *or* **Granadilla** *(Passiflora edulis), climbing plant, native to Brazil, introduced elsewhere, with long twining stems, three-lobed leaves, distinctive tendril-like white and purple flowers and egg-shaped purple fruits with a many-seeded, slightly acid pulp. Many kinds of edible passion fruits grow in tropical America.*

Jackfruit *(Artocarpus heterophyllus) of southern India, now found widely in Southeast Asia, is like breadfruit. A tree reaching 25m (80ft), with dark green, leathery, bluntly oval leaves and huge warty fruits up to 32kg (70lb) in weight, edible raw but better cooked; seeds can be roasted.*

VEGETABLES

Okras *(Abelmoschus) There are many species of this bushy herb. A cluster of up to 40 yellow, five-petalled flowers produce the 'lady's fingers' which are a good food for those with stomach trouble and can be eaten raw. Roast the seeds inside.*

Potato *(Solanum tuberoisum) grows wild in the Andes, but BEWARE of its tomato-like fruits – which in this case are poisonous. The wild tomato looks so similar that it is safer to avoid it. Always cook potato tubers.*

Sweet Potatoes *(Ipomoea) are now found in many parts of the world but are especially abundant in the tropics, where they are widely cultivated. They are straggly, creeping, vine-like plants, often with heart-shaped leaves, and sometimes large, spindly tubers. These can be boiled or roasted. They may exude a milky juice, but this is not poisonous.*

Beans *(Phaseolus) occur widely in huge variety, many as escapees from cultivation. All are twining scrambling plants with 'pea-flowers' in various colours and beans in long pods, like the domestic type. Seeds highly nutritious – boil them.*

Snake Gourds *(Tricosanthes) of Indo-Malaysia and Australia are straggling, vine-like plants with lobed leaves and very long, thin, snake-like fruits up to 2m (6½ft) long. A common kind has bright red fruits. Slice them, then boil.*

POISONOUS TROPICAL PLANTS

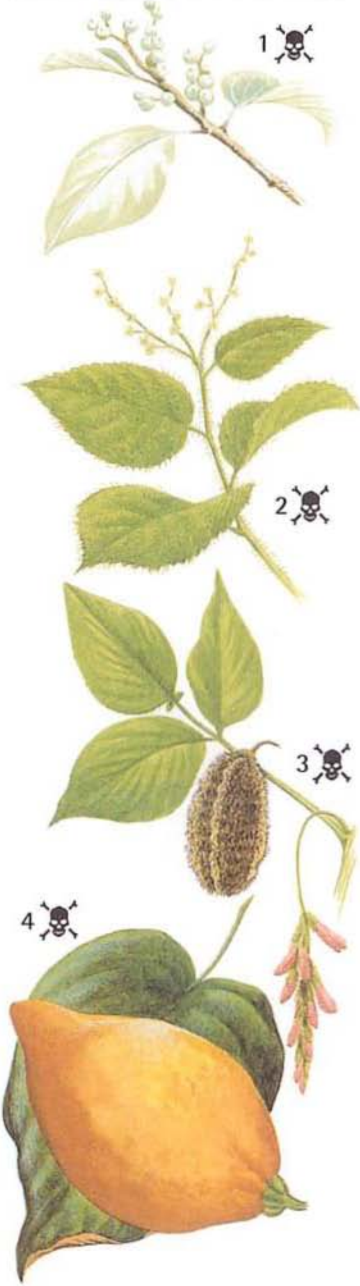

The proportion of poisonous plants in the tropics is no greater than in any other part of the world. Watch out for the following – they are either poisonous to eat or irritate on contact.

1 **White Mangrove** *(Avicennia marina) is slender, growing to 6m (18ft) in mangrove swamps and estuaries from tropical Africa east to Indonesia and Australasia. It has pale bark, many pencil-like roots spear-shaped to oblong leaves, yellow flowers and small round white berries. The sap blisters the skin and will blind if it gets in the eyes.*

2 **Nettle Trees** *(Laportea), widespread in the tropics, often by water, are smallish, with crinkly, spear-shaped, sharply-toothed leaves and drooping spikes of flowers – like an ordinary nettle. There are many kinds, including temperate ones. The burning sting is like a nettle's, but much worse. The seeds are very poisonous.*

3 **Cowhage** *(Mucuna pruriens), of scrub and light woodland, is trailing, vine-like, with oval leaflets in groups of three, spikes of hairy dull purplish flowers and brown hairy seedpods. Contact with pods and flowers causes irritation, blindness if in the eyes.*

4 **Pangi** *(Pangium edule) a tree reaching 20m (60ft) in jungle in south-east Asia, mainly Malaysia, with heart-shaped leaves in spirals, spikes of green flowers and clusters of large brownish pear-shaped fruits. All parts are poisonous, especially the fruits – the seeds contain prussic acid.*

5 **Physic Nut** *(Jatropha curcas) a shrub or small tree of wooded country throughout the tropics, has large, lobed, ivy-like leaves, small greenish-yellow flowers and yellow apple-sized fruits containing three large seeds. The seeds taste sweet but their oil is*

violently purgative, hence its name, and the remains of the pressed seeds very poisonous. Has equally dangerous relatives.

6 Strychnine *(Strychnos nuxvomica) a small tree with oval leaves in opposite pairs and white to yellowish-red, orange-like fruits whose seeds, containing strychnine, are DEADLY.* S. Nuxvomica *is found mainly in India but other strychnine species occur throughout the tropics.*

7 Castor Oil Plant *or* **Castor Bean** *(Ricinus communis), is found throughout the tropics in scrubby and waste places, is shrub-like, its leaves arranged like the fingers of a hand, with spikes of yellow flowers and prickly three-seeded pods. Seeds are violently purgative, sometimes fatally.*

8 Duchesnia *(Duchesnia indica) resembles an ordinary edible strawberry, trailing, with three-part leaves, red strawberry-like fruits and yellow, not white, flowers; on waste ground in the warmer parts of Asia, introduced to North America. The fruits are highly poisonous, sometimes fatal.*

These are only some of the poisonous tropical species. Among others which you should avoid are:

Renghas Trees *(Gluta) of parts of India east to south-east Asia, which have severely irritant sap.*

Beachapple or Manzanillo *(Hippomane mancinella) of the New World tropics, a small tree with smooth pale bark and small apple-like poisonous fruits, which also has irritant sap.*

Sandbox Tree *(Hua crepitans), also of the New World, a large, spiny tree whose sap can irritate or temporarily blind – its segmented fruits, which look rather like miniature pumpkins, are poisonous.*

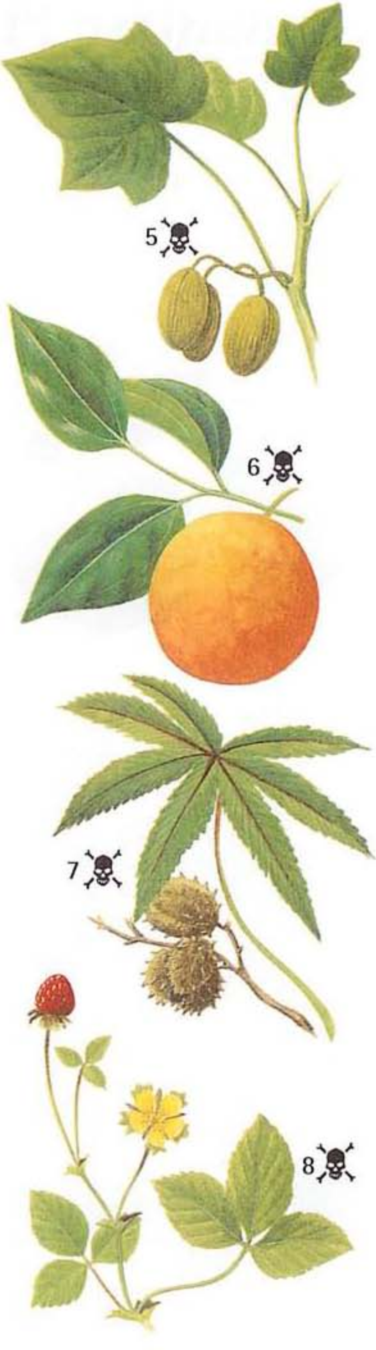

SEASHORE PLANTS

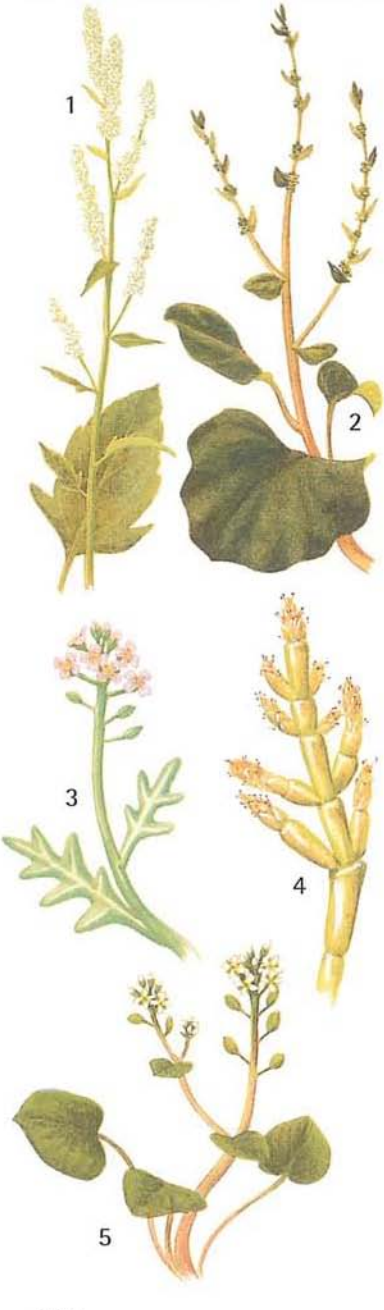

These plants thrive in salty conditions – but at the right time of year, many other edible plants occur near the coast.

1 **Oraches** *(Atriplex) average 90cm (3ft), pale-stalked, with pale green spear-shaped or triangular leaves and spikes of small greenish-white flowers; on salty ground, some kinds well inland. Cook the young leaves.*

2 **Sea Beet** *(Beta vulgaris) is sprawling, red-tinged, with leathery, long-stalked dark green leaves and clusters of small green flowers; on European coasts. The leaves are edible raw or boiled.*

3 **Sea Rockets** *(Cakile) average 30cm (1ft), with fleshy, blue-green lobed leaves, lilac or purplish flowers and egg-shaped seedpods. The peppery leaves and young pods can be eaten raw or as a potherb.*

4 **Glassworts** *or* **Marsh Samphire** *(Salicornia), often widespread in saline areas, have plump, greenish-yellow jointed stems up to 30cm (1ft) high. Some grow in great density on mud flats as single shoots 15cm (6in) high. Minute flowers are scarcely visible at the junction of the stems.* **Rock Samphire** *(Crithmum maritimum), squat and bushy, with umbels of yellow flowers, is no relation but grows on shingle as well as cliffs. Its thick, hairless stems and fleshy, grey-green leaves, cut into narrow leaflets, are both edible. Cook and suck away the fleshy parts.*

5 **Scurvy-grasses** *(Cochlearia) average 25cm (10in), with dark green, fleshy, heart- or kidney-shaped leaves and small white or pink flowers; often abundant on coasts. Very bitter, best leached in water, but so rich in vitamin C that the survivor should eat it if encountered.*

6 **Sea Kale** *(Crambe maritima) is cabbage-like, with thick grey-green leaves, four-petalled white flowers and*

globular seedpods; on European coasts. The very tough leaves are better cooked; the underground stems can be sliced and thoroughly boiled.

7 Sea Holly *(Eryngium maritimum) is thistle-like, averaging 60cm (2ft), with spiky, white-veined, ice-blue leaves and a blue thistle head; on north European coasts. Dig out the long roots, slice, boil.*

8 Oyster Plant *(Mertensia maritima) is sprawling, mat-forming, blue-green and fleshy, with oval leaves and clusters of pink to blue-purple flowers; on north European coasts. The leaves are edible raw or cooked.*

9 Scots Lovage *(Ligusticum scoticum) is stocky, celery-scented, to 90cm (3ft), often purple-stemmed with bright green leaves and heads of tiny white flowers, on north European coasts. Raw leaves are rich in vitamin C; or add these and chopped stems to other foods. Several similar lovages occur elsewhere.*

TROPICAL SHORES

Tropical shore vegetation is likely to consist principally of palm trees, though in the Old World tropics it may be possible to find:

Screw Pines *(Pandanus), so-called because of their thin, spirally grouped, leaves. They can often be distinguished by the many aerial roots at the base of the trunk. Knobbly, globular, many-segmented fruits are available all year round and in some kinds make good eating.*

Wild plums *(Spondias) may also be found near the coast.*

Sour Plum *(Ximenia caffra), a small usually thorny tree whose plum-like yellow fruits have edible pulp. There are several different kinds throughout the tropics.*

Other plants could include **Passionfruit**, *often found near the shore in the tropical Americas, and the fleshy stemmed* **Seaside Purslanes** *(Sesuvium) which grows near beaches and salt water, of which the whole plant can be eaten.*

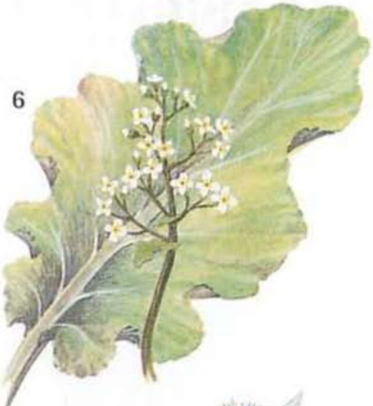

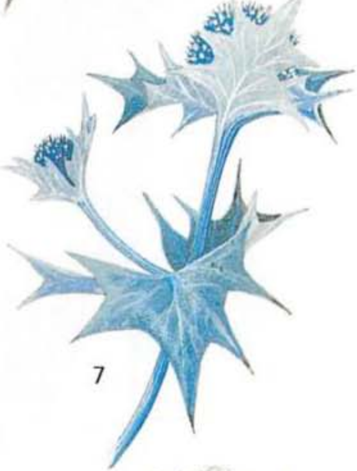

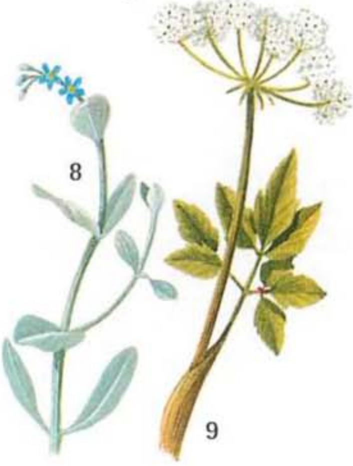

SEAWEEDS AND ALGAE

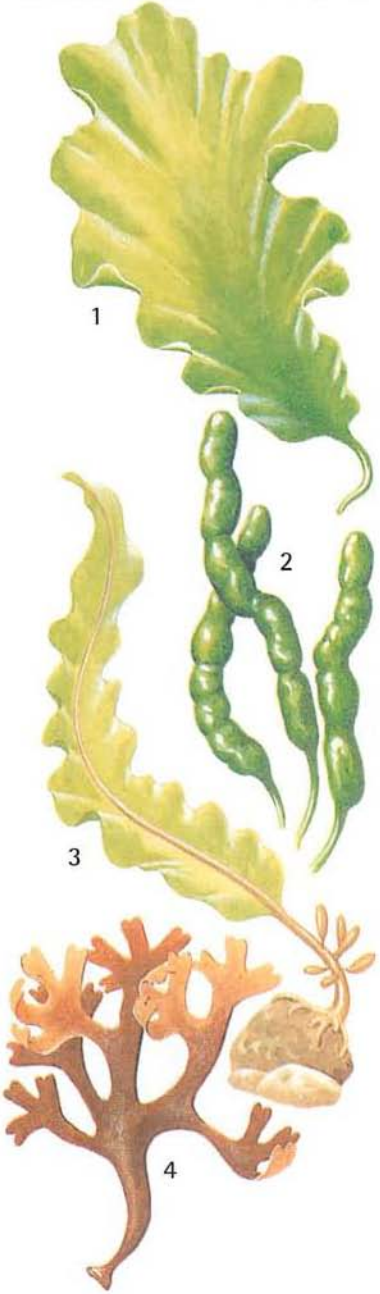

Most varieties of seaweed found in shallow waters, anchored to the bottom or a rock, but there are some that float on the surface in the open oceans. Coastal weeds are usually found stratified with green forms growing in surface waters, red in shallow water and brown a little deeper.

1 Sea Lettuce *(Ulva lactuca) is light green, resembling garden lettuce, and is found on rocks and stones in the Atlantic and Pacific, especially where water runs into the sea. Wash and boil.*
2 Enteromorpha intestinalis *is pale green, with pod-like, unbranched fronds up to 50cm (2ft) long, usually less. Often abundant on rocks in rock pools and also on saltmarshes, in cooler waters around the world. The whole plant is edible, either fresh or dried and pulverized. It is best picked in early spring.*
3 Kelps *(Alaria and Laminaria) have a short cylindrical stem and thin, wavy, olive-green to brown fronds, often very long; on rocky shores of the Atlantic and Pacific. Edible raw but better boiled.*
4 Irish Moss *(Chondrus crispus) consists of forking, lobed purplish to olive-green fronds, often in dense beds on Atlantic shores. Wash and boil. Cooled, the residue will set like gelatine. Fronds may be dried for storage; leave them in the sun until they have bleached white.*
5 Sugarwrack *(Laminaria saccarina) has long, flat, wavy-margined, yellow-brown fronds attached to stones and rocks; common in the Atlantic, and off China and Japan. Young fronds are edible raw, better cooked. Sweet-tasting.*
6 Dulse *(Rhodymenia palmata) has purple-red, short-stemmed, lobed, fan-shaped fronds, and occurs in the Atlantic and Mediterranean. Leathery but sweet; boil it. Dulse can be dried*

and rolled for chewing tobacco.

7 Lavers *(Porphyria) have thin, irregularly shaped, satiny, red, purplish or brown fronds and are found in both the Atlantic and Pacific. Boil until tender, then mash. Use as a relish or combine with grains to make cakes. Very tasty.*

Rich in vitamins and minerals, seaweeds are an ideal survival food. Those shown here are common and safe to eat.

There are no poisonous seaweeds but some contain acids which irritate the digestive tract and some are violent purgatives. If not identified as a known edible species try only small amounts. Even with the more edible varieties eat only a little until you become used to them. Do NOT eat seaweeds if short of water. If possible wash them in freshwater before eating, to remove some of the salt.

Collect growing weed, firm and smooth to the touch, not pieces washed up on the beach. Reject any that smells bad. Some which contain irritating acids can be detected by crushing between the fingers and leaving for five minutes, by which time they give off an unpleasant smell. All seaweeds decay rapidly out of water. Use soon after collection or dry for later use.

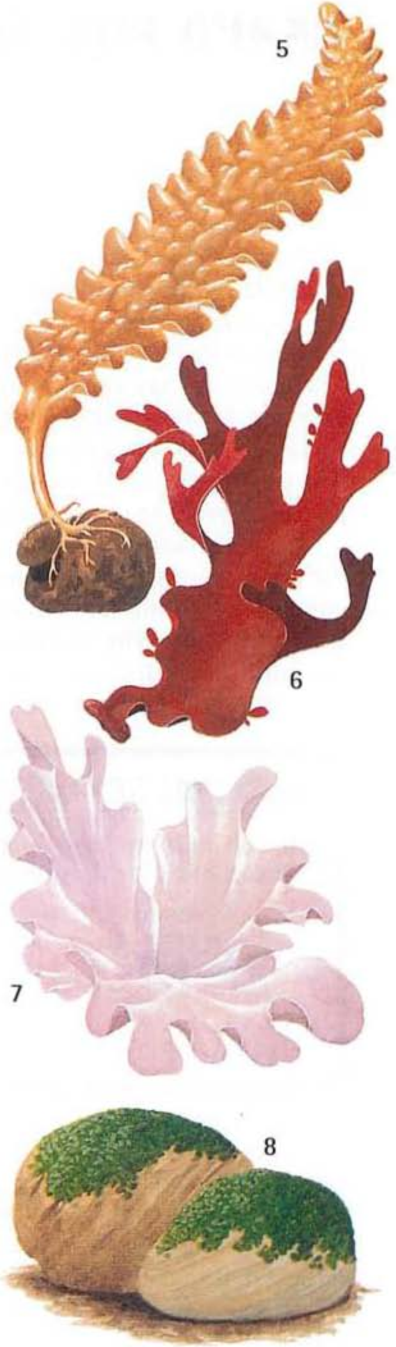

FRESHWATER ALGAE

8 Nostoc *(Nostoc) is a freshwater algae of North America and Eurasia, forming green, round, jelly-like, marble-sized globules in pools, from spring on. Dry and use as a thickener. EAT ONLY bright green fresh-looking algae. AVOID ALL BLUE GREEN ALGAE, IT IS POISONOUS. It is found in freshwater, not in the seas and oceans, floating on the surface of stagnant pools. Identify by its blue-green colour and by its gassy smell.*

TRAPS AND TRAPPING

It is easier to trap most small prey than to hunt them. Even if you spot a small animal it offers very little target and can easily take cover. Trapping requires less skill and leaves you free to spend time foraging for other food. Nevertheless be ready to take advantage of sitting prey if you get the chance.

The survivor's own preservation must take precedence over humanitarian principles and unfortunately some of the easiest traps can cause considerable suffering to the animal. A trap which could bring quick death to the species for which it is intended, perhaps by strangulation, may catch another animal by a limb and leave it suffering for hours. Regular checking is essential. Leaving a trap line unchecked will prolong an animal's pain and increase the risk that your catch may be poached by an animal predator or that the prey will have managed painfully to struggle free.

A great deal of error can be eliminated by studying the animals and their habits. Choice of baits and sites is important. If one does not work, try another. BE PATIENT. Give the traps time. Animals will be very suspicious at first but with time will come to accept them – and that is when they will run into them.

SURVIVAL SCENARIO

If you have no food supplies which foods are the most important to hunt or forage for?

No one-food source is more important than another, in fact it is vital that in any survival situation a mixed diet is achieved. If this does not happen then you will actually feel more tired and unwell if you just eat say fat or protein. However, if you are injured then you will need more proteins to help repair the damage, and in general you should eat little and often.

Even when on the move, a few simple traps, quickly set up overnight, may be productive, and if you are making a more permanent camp you will be able to set up well-planned lines of traps. The more you set, the more chance you have of success.

Establish as large a trap line as you can manage in your area. Inspect it at first and last light. Collect the game and reset the traps. Repair any traps as necessary and move those that are repeatedly unfruitful. To be effective a trap must be very sensitive, so may be fired accidentally. You will probably have several empty traps for every success, but this does not mean that you are doing things wrong. If a trap has not fired, but the bait has gone, it is an indication either that the bait was not sufficiently securely fixed or that the trigger mechanism is too tight. Check both when you reset the trap.

By doing the rounds regularly you effectively patrol an area, noting the many signs of activity or change which help to build up knowledge of your surroundings.

TRAPPING TIP

Baiting a trap will attract the game. In a survival situation food may be scarce but, if you know there are animals to trap, a little used as bait may bring large rewards.

Where to trap

Find the game trails or runs, which lead from an animal's home to where it feeds or waters. Look for any natural bottleneck along the route where it will have to pass through a particular position – a deadwood fall or a place where the track goes under an obstruction will be ideal places to set a trap.

Do not place a trap close to an animal's lair. That is where it sits and listens and sniffs the air. If at all suspicious it will either stay put or use a less obvious route. Don't place a trap close to its watering place either. There, too, the animal is on its toes and alert, more likely to notice anything unusual.

If you lay traps down the side of natural pastures the animals will not go near them but use other routes. However, when alarmed they panic and will take the shortest route to cover. That is when the crudest and most obvious of traps will be successful. Rabbits are easily caught by causing them to panic.

Trap construction

The simpler traps and snares are made of string or wire. It will be easier to keep a loop open in the air if you use wire and the wire in your survival tin is ideal. Even the most sophisticated need nothing more than a knife to make them out of available wood. The choice of materials is important. Use strong, springy wood. Do not use dead wood or wood found on the ground. Hazel takes a lot of beating: it is easy to carve and retains its spring and strength.

Types of trap

Trap mechanisms make use of the following principles:

MANGLE STRANGLE DANGLE TANGLE

The deadfall mangles. The snare strangles. Springy saplings can make a trap more efficient and take the game up in the air – it dangles. The higher the sapling the more effectively it lifts the animal. A net tangles. Some traps combine two or more of these principles.

RULES FOR TRAPS

When setting traps, follow these basic rules:

1 Avoid disturbing the environment
Don't tread on the game trail. Do all your preparation off the trail and don't leave any sign that you have been there.

2 Hide scent
When constructing or handling traps don't leave your scent on them. Handle as little as possible and wear gloves if you can. Do not make a trap from pinewood and set it in a wood of hazel. Each tree gives off its own smell – and the animals you are trying to trap have a very high sense of smell, many times sharper than yours. Although they fear fire they are familiar with the smell of smoke and exposing a snare to the smoke from a camp fire will mask any human scent.

3 Camouflage
Hide freshly cut ends of wood with mud. Cover any snare on the ground to blend in as naturally as possible with its surroundings.

4 Make them strong
An ensnared animal is fighting for its life. It exerts a lot of energy in an attempt to escape. Any weakness in the traps will be exposed.

SNARES

Snares are the simplest of traps and should be part of any survival kit. They are made of non-ferrous wire with a running eye at one end through which the other end of the wire passes before being firmly anchored to a stake, rock or tree. A snare is a free-running noose which can catch small game around the throat and larger game around the legs.

A snare can be improvised from string, rope, twine or wire. Consider the kind of animal you are trying to trap when you place the snare. A rabbit, for instance, tends to sit in cover and observe. When satisfied that all is well it hops along. Setting the snare a hand's length from a fall or obstruction on the trail accommodates this hop. If the snare is closer to an obstruction the rabbit may brush it aside.

A wire snare (as big as your fist) can be supported off the ground on twigs, which can also be used to keep a suspended string noose open.

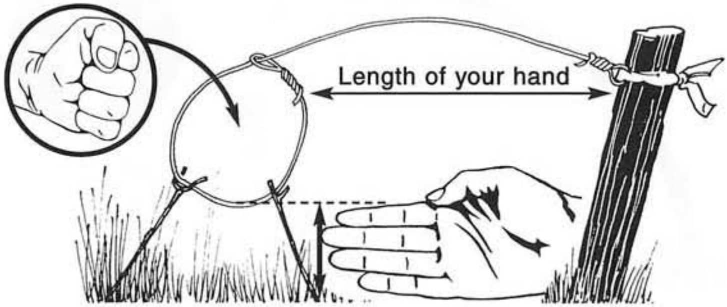

USING A SIMPLE SNARE

For rabbits and small animals. Use your judgement to scale up these proportions for larger creatures, such as foxes and badgers.

- Make the loop a fist width wide.
- Set it four fingers above the ground, and one hand's width from an obstruction on the trail.
- Check that it is securely anchored, with twigs to support the loop in position if necessary.

Snares under tension

Make a snare more effective by using a sapling under tension to lift the game clear of the ground when it is released. This robs the animal of purchase in its struggle to free itself and also helps to keep it out of reach of predators.

Spring snare

When game is caught the trigger bar disengages and prey is lifted off the ground. Good for animals such as rabbits and foxes, it will trap game coming in both directions and is ideally situated on the game trail by a natural bottleneck caused by a dead fall or a rocky outcrop.

Cut notch in trigger bar (a) to fit notch in upright (b). Drive upright into ground. Attach snare to trigger bar and use cord to sapling to keep tension.

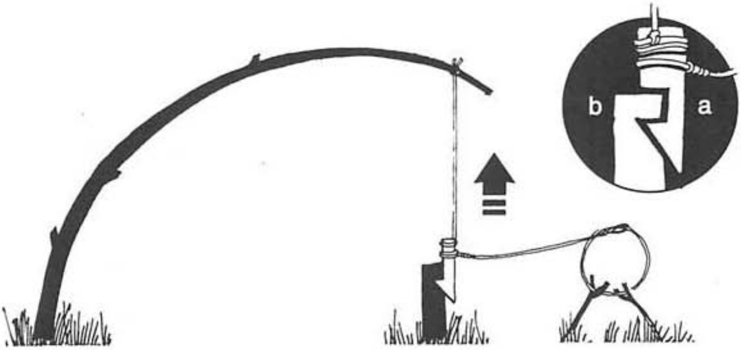

Baited spring snare

Mechanism as for spring snare, but here the quarry is tempted with a tasty morsel. The noose is laid on the ground, the bait strung above. As the game takes the bait the trigger is released.

Suitable for medium-sized animals such as foxes, this trap can be located in an open area as the bait will attract attention. Small clearings in woods are good sites.

The bait support stake should be only lightly driven into the ground as this must fly away with the noose.

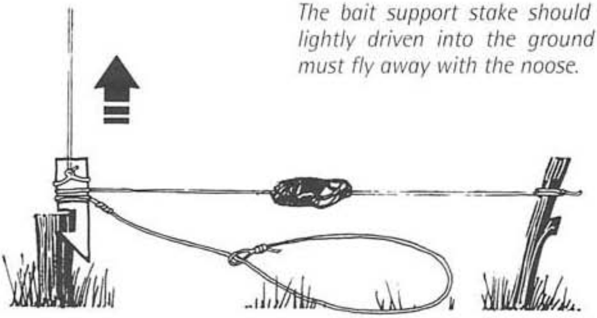

Baited spring leg snare

The prongs of a natural fork of wood, or two sticks tied together, are pushed firmly into the ground. The line from a bent sapling is tied to a toggle and to the snare and the toggle then passed under the fork. When the game takes the bait, which is on the end of a separate bar, the bar disengages and the toggle flies up carrying the snare (and hopefully the game). This is a trap for larger game such as deer, bears and large felines. For the herbivorous deer, bait with blood or scent glands, which will arouse its curiosity.

The upper end of the toggle presses against the fork and the lower end is prevented from pulling back through by a bait bar between it and the fork – the pressure of the toggle holding it in position.

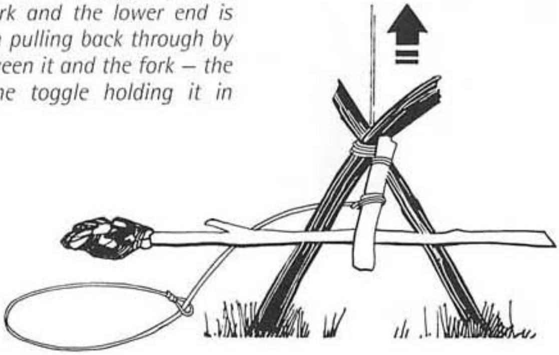

Spring tension snare

The upward counter thrust from the keeper stick (a) on which the snare arm (b) rests prevents the switch from pulling it up. When the game becomes ensnared the snare arm is dislodged from the keeper stick and the switch line slips off the other end. Suitable for small animals such as rabbits. Site it on the game trail.

Note how the switch line secures one end of snare arm (b), while the other rests on the keeper stick (a). Keep the switch line near the end of the snare arm (c).

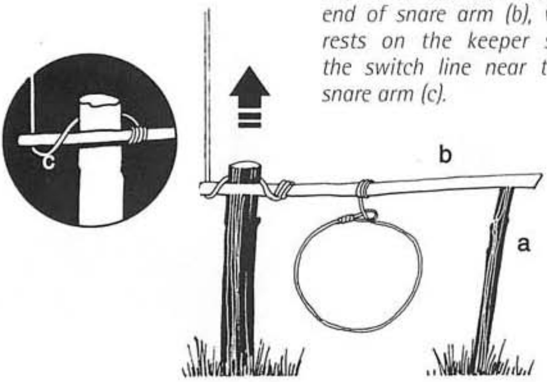

Trapeze spring snare

This snare can be used to cover two game trails in open country. The arm carries two snares and is held in a notch by the tension of the switch line.

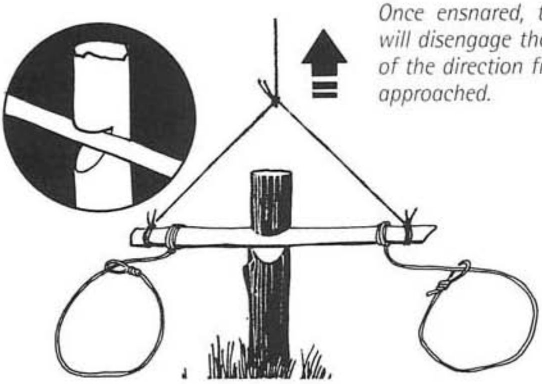

Once ensnared, the animal's struggles will disengage the snare arm regardless of the direction from which it originally approached.

Roller spring snare

A rounded grip holds the snare arm here, the switch line is best pulled back at a slight angle to keep it in place. Suitable for animals such as rabbits and foxes. Although tensed in one direction, the bar will be dislodged by an animal's struggles.

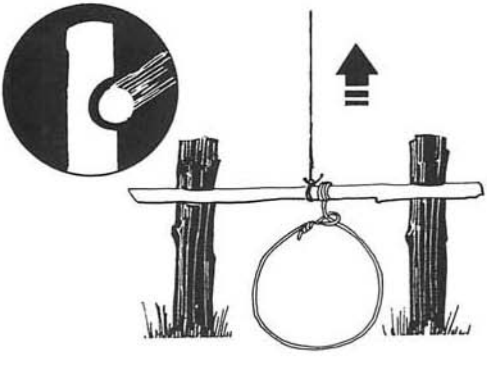

A wide area can be covered by employing several snares on a long horizontal bar. Use where the game trail widens or offers options.

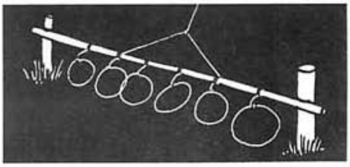

The platform trap

This trap is ideally sited in a small depression on the game trail. Snares are placed on the platforms on either side. When the platform is depressed the trigger bar is released and the game held firmly by the leg. Ideal for larger game – such as deer, bears or large cats. A platform of sticks, stiff bark or other firm materials rests on the bottom bar, and the upper bar fits in the notches. A similar mechanism (a) to that of the platform trap, but using a large snare and no platform, is activated by displacement of either the toggle or bottom bar to catch small game by the neck.

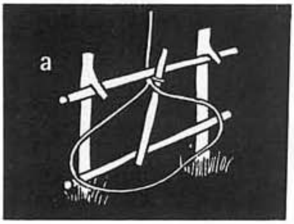

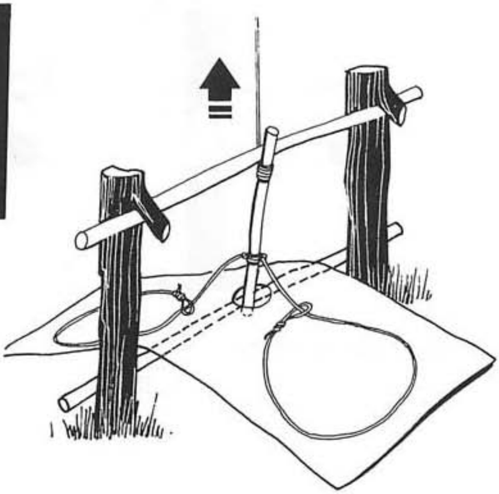

Stepped bait release snare

Two forked sticks hold down a cross-bar which engages with a baited notched upright (attached to a line in tension), which holds it in place and carries the snares. Site this trap in clearings to catch small carnivores and pigs.

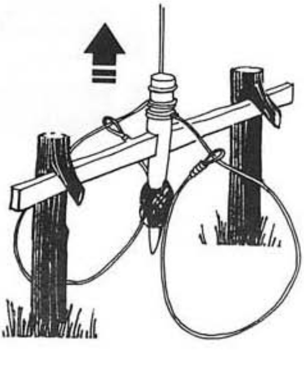

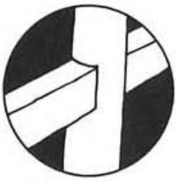

Retaining bar, or at least a section of it, should be squared off to fit a square-cut notch on the bait stick.

Double-ended figure 4 snare

A bait bar is set at right angles across an upright, the faces cut square, and a short trigger lodged between them to maintain their position. The trigger is linked to a springy sapling and the same line carries snares. When the bait bar is dislodged the trigger is released. Four snares will effectively cover both directions on a game trail, or use in clearings to snare small carnivores.

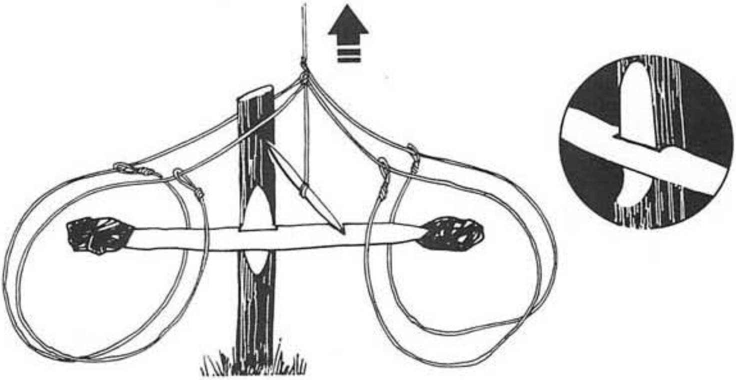

Double spring snare

Two saplings are notched to interlock when bent towards each other over the game trail. A vertical bait bar is lashed near the end of one of them. Two snares can be attached to each sapling, they need to be fairly stiff wire to hold their positions. This is another trap suitable for use in clearings to catch small carnivores. When the bait is taken the game is held in the air between the saplings.

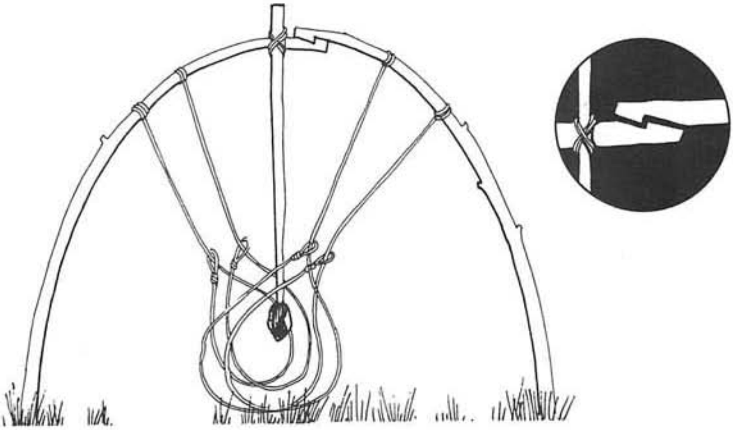

Toggle and bait release snare

A bait bar is wedged between an upright and the lower part of a toggle, the upper part pressing against a retaining bar. The principle is like the platform snare with the bait bar replacing the platform. Movement of the bait bar releases the toggle which flies upward under tension from a sapling above, carrying snare with it.

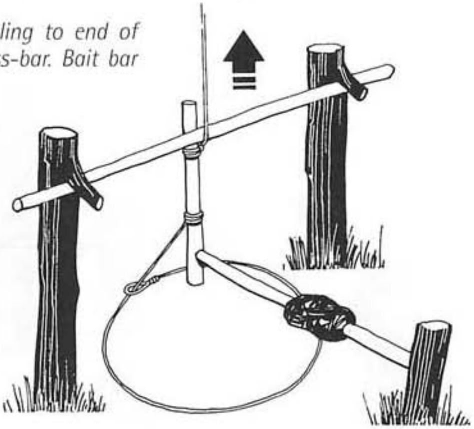

Taut line from bent sapling to end of toggle pulls against cross-bar. Bait bar keeps toggle in position.

DEADFALL TRAPS

These traps all work on the principle that when the bait is taken a weight falls on the prey. All are good for pigs, foxes and badgers. Larger versions can be used for bigger animals such as bears.

WARNING

The large versions of these deadfall traps can be extremely dangerous for humans as well as for the prey for which they are intended. The toggle release and deadfall traps have trip wires and are easily set off accidentally. Even in a survival situation ensure that everyone knows exactly where they are. In survival practice keep people away from them and never leave such a trap set up at the end of an exercise.

You cannot set a large deadfall trap on your own. Keep the mechanism to the side of the trail, well away from the dropping weight, or setting it will be too risky. Balance is critical – you are unlikely to get it right first time.

Toggle trip-release deadfall trap

This uses the same kind of mechanism as the toggle-release snare – though this time the release bar keeping the toggle in position presses one end of the toggle upwards. A line from the toggle passes over a tree limb to support a bundle of logs or other heavy weight above the trail. From the release bar a trip line (usually a vine) runs above the ground beneath the suspended weight to a firm securing point.

Run the trip line under a forked stick (a) so that it will pull the trigger bar sideways when operated.

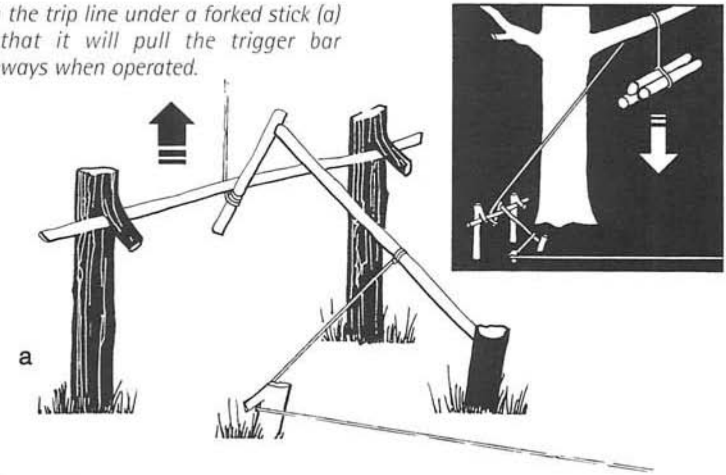

Balance log

A forked stick, its ends sharpened to dislodge rapidly and one fork suitably baited, supports one end of a cross-bar, the other end of which rests on a fixed support, held there by the weight of the heavy logs or rock which rest on the bar. When the bait is taken the whole trap collapses.

Angle bait beneath the trap.

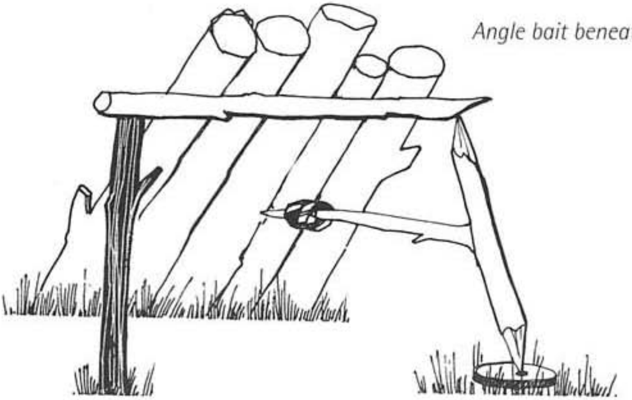

Squared-face release trap

Similar to the balance log trap but using a notched upright as the support, the lower face of the notch squared off. Fit cross-bar against the squared-off lower face of the bar supporting the weight.

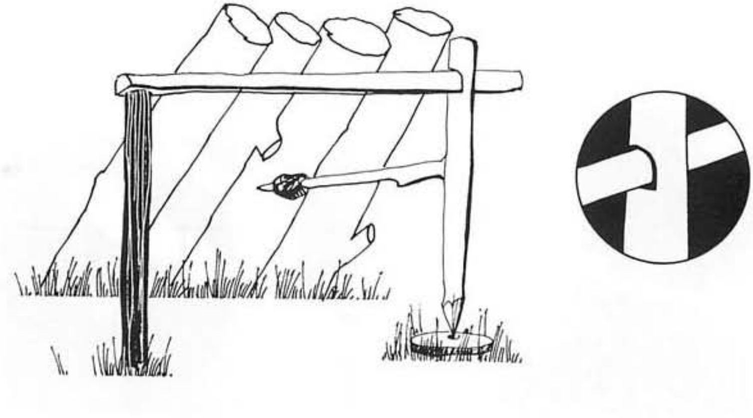

Toggle and bait release deadfall trap

A rock or a group of logs lashed together is supported by a prop which is balanced over a fixed forked stick. The other end of the prop is clear of the ground and held down by a short line attached to a toggle which is wrapped around the upright stick. The toggle is kept in place by a bait stick wedged between it and the dead fall weight. Dislodging the bait stick brings the whole lot down.

The forked stick is driven into the ground first.

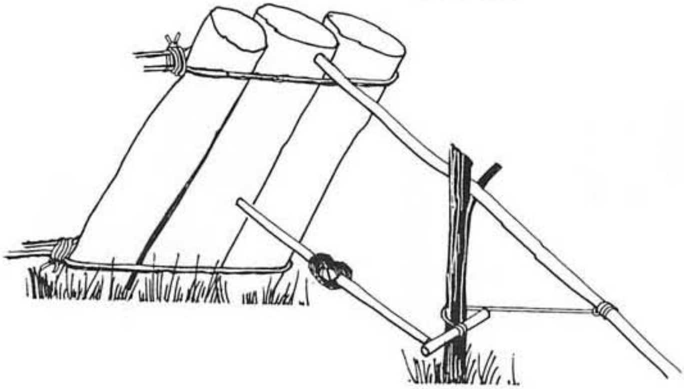

Deadfall trap

The weight of a log or other weight suspended over the game trail pulls the line carrying it against a retaining bar held by short pegs secured in a tree trunk. The line continues as a tripwire beneath the weight. Make sure that the line is long enough and tripwire anchor weak enough to allow weight to reach the ground. Set the pegs slightly downward, but keep them short so that the bar disengages easily.

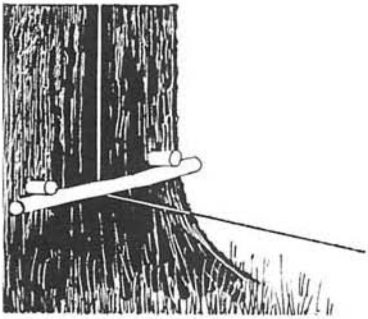

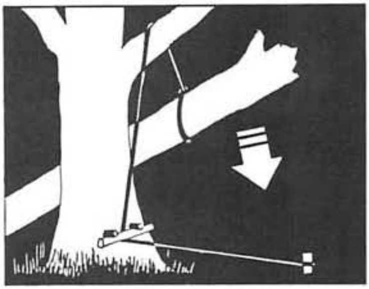

'Figure 4' deadfall trap

This looks complicated but once learned is easily remembered and very effective. It can be made to any size. A horizontal bait bar is balanced at right-angles to an upright with a locking bar, which supports a weight, positioned over the bait, pivoted on the sharpened tip of the upright.

Bait bar notched on top to engage locking arm, square cut on side to fit upright. Locking arm sharpened at lower end to release quickly, notched at centre to pivot on upright.

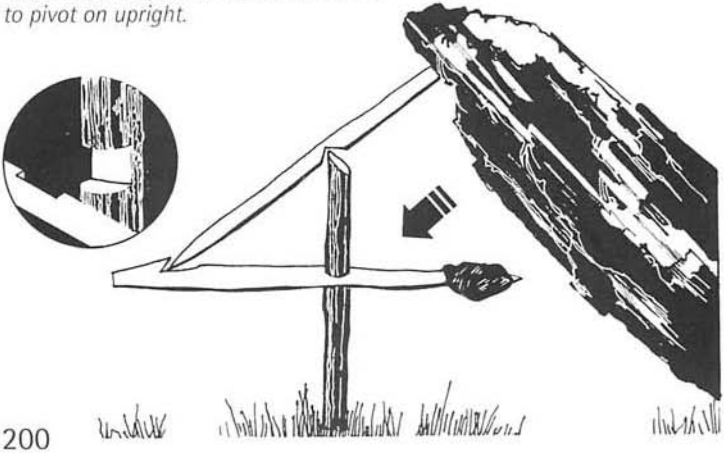

SPEAR TRAPS

WARNING

These traps are particularly good for killing pigs and deer – but they are extremely dangerous and can be lethal to humans. Always stand BEHIND the spear when setting and ensure that the location and danger is known to everyone. Mark with signs to attract human attention. Except in a survival situation never leave spear traps set and unsupervised.

DEADFALL SPEAR TRAP

This uses the same mechanism as the deadfall trap (left) but uses rocks to add weight and arms the trap with sharpened sticks. It delivers a stabbing as well as a stunning blow.

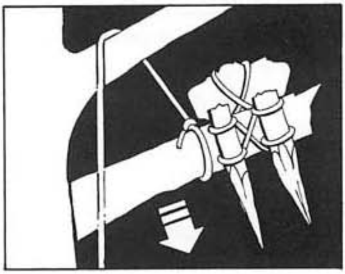

Spring spear trap

A very dangerous trap which will kill game. Effective against wild pig. A springy shaft, with a spear attached, is held taut above the trail. A slip ring made from bound creeper or smooth material (not rough twine which could catch against toggle) attached to a trip wire acts as a release mechanism.

A toggle (a) and short line (to a fixed upright) hold the spear shaft in tension. A further rod through ring is tensed between the near side of the spear shaft and the far face of the upright, securing all until tripped.

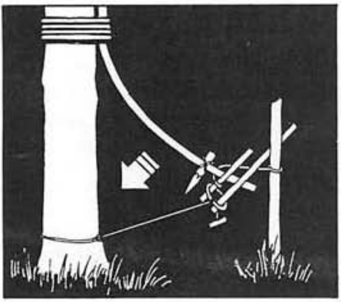

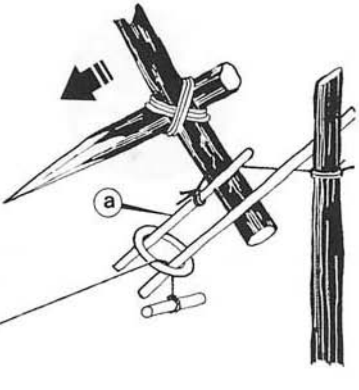

Pig spear trap

Similar to the spring spear trap but operating horizontally, this trap has the unarmed end of the springy shaft secured and lashed between four uprights. At the business end, the toggle (anchored by a short line) retains the springy shaft so long as the toggle point is held against the horizontal bar by a ring. The ring is on the end of a trip wire, anchored to a post on the other side of the trail.

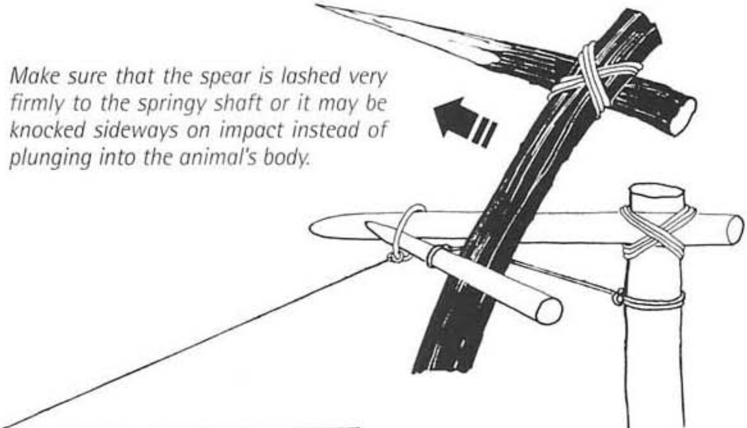

Make sure that the spear is lashed very firmly to the springy shaft or it may be knocked sideways on impact instead of plunging into the animal's body.

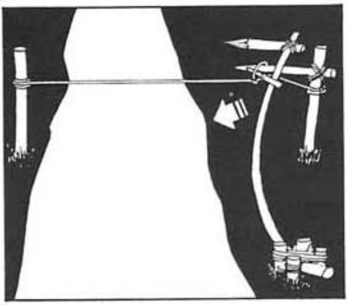

Spear shaft is set at a height level with the body of the animal it is designed to kill, or angled to spring to that height.

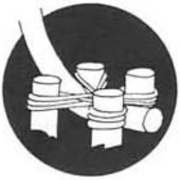

Tension exerted on the springy shaft requires the uprights holding it to be very firmly set in the ground and the lashings to be strong and secure.

NOTE: Because spear traps are so dangerous, make sure the cord and the knots are strong enough to stand the tension. Never approach these traps except from BEHIND the spear.

Take no risks where these traps are concerned.

Bow trap

WARNING

Never leave this unattended where people could walk into it – it is a potential man-killer.

A simple bow made of suitable wood (see *Weapons*) is held taut and angled to shoot slightly upwards by upright posts and a toggle switch with an arrow fitted. The trigger bar is held in place by a toggle attached to a trip wire, which must be routed round to the point of aim. Keep the first stretch of wire close to the mechanism, for there is no point in it being tripped by an animal approaching from behind the bow. This trap is suitable for large and dangerous animals and can work with animals coming head on to the arrow or approaching from the trip wire side. (The quarry passes across the arrow as it fires.) The arrow may also strike larger animals passing in front of the bow first.

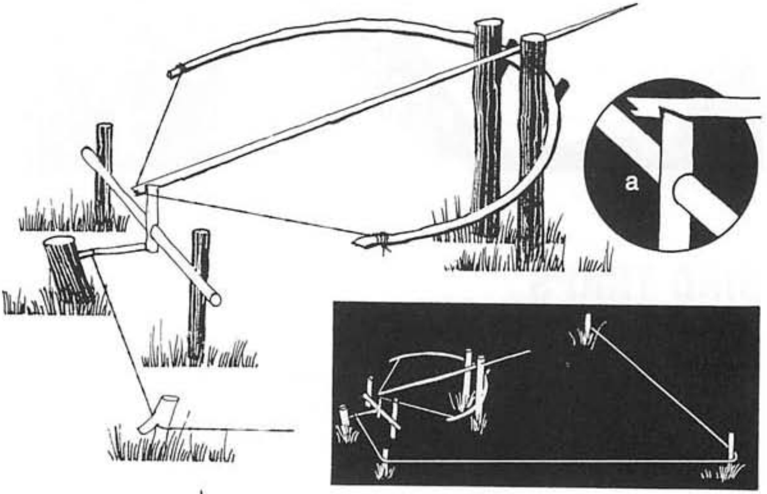

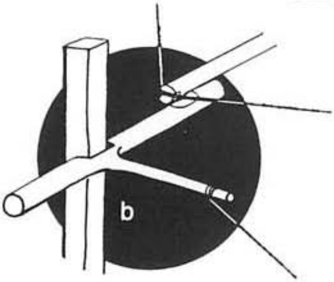

(a) Notch arrow for bowstring and for trigger bar. Angle trigger bar tip to fit arrow notch, cut side to fit cross-bar. Sit toggle between bottom of trigger bar and a fixed post.

(b) An alternative trigger mechanism: Cut a square face on an upright and a square notch on the side of a forked stick to engage it. Notch the upper face of the stick to hold the bowstring. Attach tripwire to other prong of fork.

Baited hole noose

Digging pits disturbs the environment and leaves a permanent mark. This will alarm some animals. In others curiosity may outweigh discretion and they will investigate. Baiting the hole may bring animals sniffing. Foxes, pigs, wild cats and badgers will all dig up rubbish pits and this could attract them. The animal smells the bait and pushes its head down. If it goes past the stakes it will not be able to retract it. If it uses a paw it will become ensnared.

Drive four sharpened pliable stakes through the edges of the pit to emerge below surface where they are less noticeable. Lay a noose across them, attached to a post outside the pit.

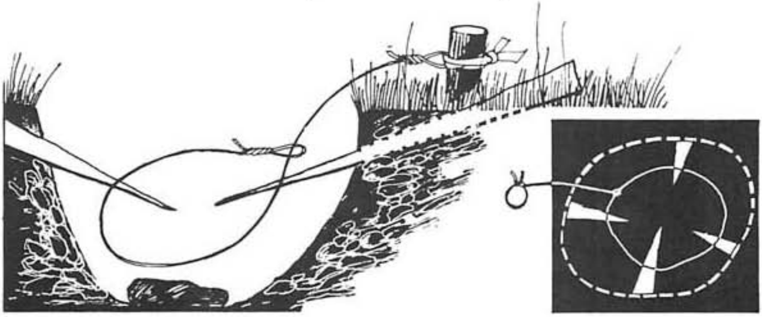

BIRD TRAPS

Nets

A fine net stretched between the trees where birds usually roost is one of the simplest ways of catching them. Instead of a net, fine twine criss-crossed between trees across their flight path will damage birds which fly into it.

Bird lime

Liming is an ancient way of catching small birds. Boil holly leaves and any starchy grain in water and simmer until you have a gooey mess. Spread this on the branches or other perching places before the birds come home to roost and they will get stuck in it when they alight.

Suspended snares

Hang a line of snares across a stream a little above water level. This works best when set among reeds and rushes.

Baited hooks

Fish hooks buried in fruit or other food can be an effective way of catching birds. The hook gets caught in the bird's throat.

Noose sticks

Tie many fine nooses 1.25–2.5cm (½–1in) in diameter, close together along a stick or branch, use horsehair preferably but any strong material will suffice. Place the stick in a favourite roosting or nesting spot with the nooses uppermost. Birds become entangled when they alight. Do not remove as soon as one bird is caught. It will attract other birds and you will soon have several.

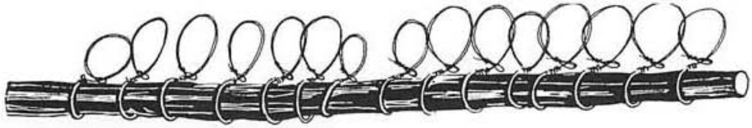

'Figure 4' trap

This mechanism (see *'Figure 4' Deadfall* on page 200) can be used with a 'log cabin' type cage, made from a pyramid of sticks tied together, which is balanced over the bait. For small birds you can use a quick method of making the cage: lay all the sticks in position then lay another two sticks, the same length as the bottom ones, on top and tie them tightly to the bottom layer, tight enough to keep all the others in place. Larger animals will soon break out of this and for them each stick must be individually tied in.

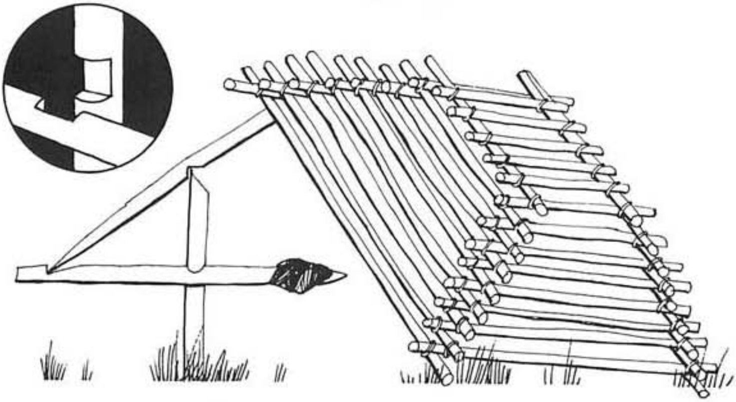

Experiment with different ways of making a cage. You may have a suitable box or large tin which would do just as well. It is also possible to prop the raised edge of the cage on a single stick tied to a long line. Take the other end of the line and hide some distance away. If you hold the string taut, you can snatch the prop away as soon as a bird ventures under the cage. Broadcast bait around and under the cage. This works best in areas where birds seem plentiful.

Toggle release net trap

A net laid on the ground and baited to attract birds has lines from the corners to a springy sapling overhead. A tension line extends to a toggle mechanism (see previous traps) notched on to a horizontal bar and operated by a flat bait stick. Set the bait stick off the ground and only just resting against the lower end of the toggle. This trigger mechanism needs to be extremely sensitive if a small bird's weight is to set it off.

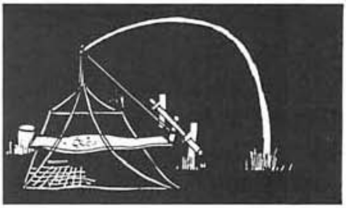

Bait scattered across the net will attract birds which will be caught when one steps on the bait bar.

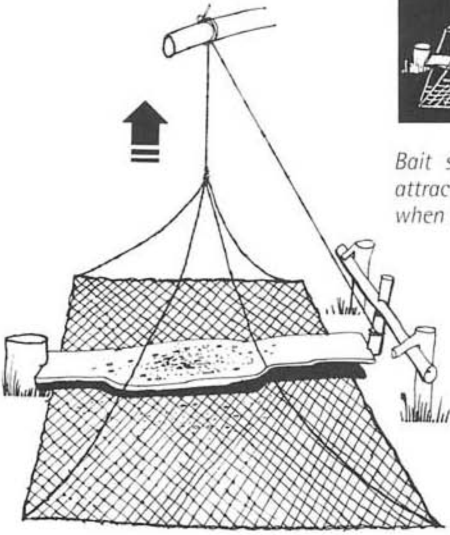

NOTE: If you set traps in a training exercise make sure that they are clearly marked so that they are not set off by other people. Spear and deadfall traps should be supervised to keep people away for they could inflict serious injury or kill. ALL traps should be dismantled when the exercise is over.

SURVIVAL CASE STUDY

Monsoon weather is always bad, but in 1963 it broke all records. In this year I was deployed in Borneo to gather topographical information and intelligence. We were coming to the end of a long operation when our base camp was washed away by the torrential rain. Most of our personal kit and equipment was lost, except for what we were wearing. Low clouds made air re-supply impossible and the tropical storms affected our radio communication. We had to seek higher ground as the whole area became flooded. Moving was very difficult because of the swollen rivers and landslides. It took us 5 days to reach a safe landing zone (LZ) where we awaited a helicopter to lift us to safety.

During these 5 days, the only food we had was our meagre survival rations. This consisted of a hard biscuit, nuts and raisins, a tube of honey and a 2oz meat block. We shared everything between the four of us. Even under normal conditions wild food is scarce in the jungle. Fishing is probably best, and collecting various palms, roots and shoots is next best. But the weather washed all these sources away, and most animals fled the area before we did, so we didn't pick any food up on our route.

What we found at the LZ were large white snails. We baked, steamed, poached, and even ate them raw, but however we ate them they still tasted the same. Nevertheless, they were an excellent source of nourishment. We had come to the stage of eyeing each other up – and I think these snails even prevented cannibalism!

Lessons learnt:
Carry survival kit at all times
Eat anything you can find
Pool your resources
Know and practise your survival skills.

Water was not a problem in this case, in fact we had too much. Caught rain is the only water we don't treat. But if at all unsure of water quality, then boil it. Also boil all meat, insects and roots until tender, then reduce to a simmer before adding herbs or plants as constant boiling destroys vitamin C content which is essential to good health.

HUNTING

Keen observation of all signs of wildlife and a knowledge of the kind of animals you are hunting are as necessary to the hunter as skillful tracking and accurate marksmanship. They make it easier to be in the right place and to take advantage of the terrain.

Always proceed as quietly as possible. Move slowly and stop regularly. Carry your weight on the rear foot so that you can test the next step with the toes before transferring your weight. Thus you will avoid stumbles and help to reduce the amount of noise you produce from undergrowth and snapping sticks. Fast or sudden movements will startle the game. Sniff the air and listen. Hunt against the wind, or at least across it.

The ideal time to hunt is at first light, when more game is likely to be about. Animals are also about in the evening, but the light will be getting rapidly worse so you need to be sure of the terrain and know your way back to camp. In territory you know well this will not be a problem, particularly if there is a clear sky and moon or starlight to see by.

If hunting in the evening go out at least an hour before dusk so that your eyes will get used to the failing light and you will develop night vision – though your prey will probably be able to see better than you do.

When hunting during the day, try to hunt moving uphill in the morning and return to camp in the afternoon. Signs of animals will be easier to read as you move uphill for those on the ground will be closer to eye level. Thermal currents build up with the heat of the day and carry scents upwards – so when you are coming down the hill the scent of the game will reach you before your smell reaches it. After a day out hunting and foraging the descent will take less energy than an upward climb and by then you'll welcome the easier going.

SURVIVAL ADVICE: HUNTING

If you are moving correctly, game often will not see you. If an animal catches a glimpse of you, freeze. You may be the first human it has seen. It will be more curious than frightened. Keep absolutely still until the animal looks away or continues feeding. Avoid large animals, such as bears, unless really desperate or confident of a first shot kill – or you could end up becoming the hunted not the hunter.

Get as close as you can without revealing your presence and take up a steady position, aiming for the area giving the greatest margin of error. An accurate head shot is very effective but risky unless you are very close and the animal still. A point just to the back of the front shoulder is a good target. A firm, accurate strike here will drop most animals instantly. A badly aimed shot may mean unnecessary agony for the animal and a long follow up for the hunter.

If an animal drops first shot, wait five minutes before approaching. Just stand back and observe. If not dead but bleeding, the loss of blood will weaken it and, when you do approach, it will not be able to bolt. If an animal is wounded and moves away wait 15 minutes before following up. If you follow immediately the animal may travel all day.

WEAPONS

Bow and arrow

Most effective of improvised weapons, the bow and arrow is easy to make. It takes only a short time to become proficient in its use.

For the bow a well-seasoned wood is best but you will have to make do without. If you expect to have to stay where you are for many months you could put wood aside to season for future use. The tension in unseasoned wood is short-lived so make several bows and change over to another weapon when the one you are using loses its spring.

Yew is the ideal wood – all the old English longbows were made of yew. There are five kinds of yew distributed across the northern hemisphere but it is not very common and hickory, juniper, oak, white elm, cedar, ironwood, birch, willow and hemlock are all good alternatives.

Making the stave

For your bowstave select a supple wand. It should be about 120cm (4ft) long, but match its size to the individual.

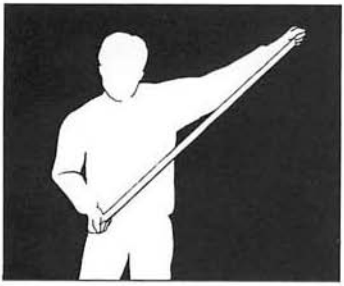

To determine the correct stave length for you:
Hold one end of the stave at the hip with the right hand, reach out sideways with the left hand and mark the extent of your reach as the length of the bow.
This will give you a standard type bow (the longbow requires much more skill in use).

Shaping the bow

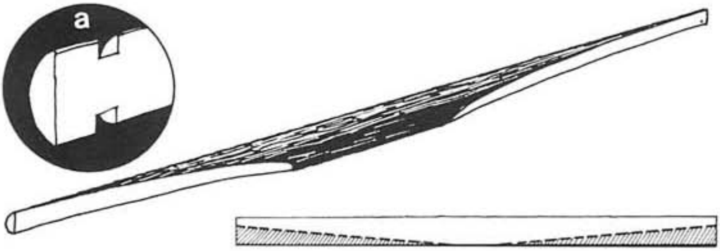

Fashion the stave so that it is 5cm (2in) wide at the centre, tapering to 1.5cm (5/8in) at the ends. Notch the ends (a) to take the bowstring about 1.25cm (1/2in) from the ends. Remove the bark if you choose. When the bow has been whittled into shape rub it all over with oil or animal fat.

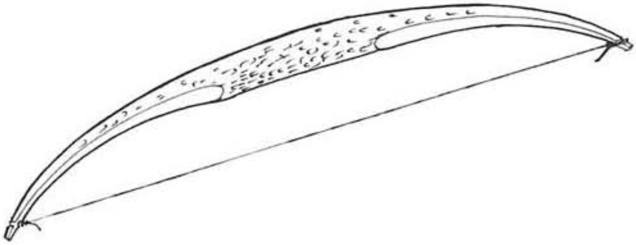

Fitting the string

A rawhide string is best, cut to a width of 3mm (1/8in), but any string, cord or thin rope will suffice. The stems of old nettles provide tough fibres and these can be twisted together to make a satisfactory bowstring. If the bow has a lot of give a shorter string is likely to be needed, but when strung the string should only be under slight tension – the main tension is added when you pull it back to shoot.

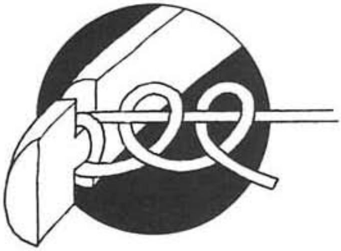

Secure the string to the bow with a round turn and two half hitches at each end. If the wood is unseasoned release one end of the string whenever the bow is not in use to relax the tension or you may find the stave sets in shape.

A properly made bow will be more efficient and more accurate than just bending a pliable wand – but once it loses its spring don't waste time with it. Make another.

Making arrows

Any straight wood will do for arrows, but birch is one of the best. Make arrows about 60cm (2ft) long, and about 6mm (¼in) wide. Keep them straight (a piece of string tied between two points will give you a straight edge to check them against) and as smooth as possible. At one end make a notch 6mm (¼in) deep to fit the bow string.

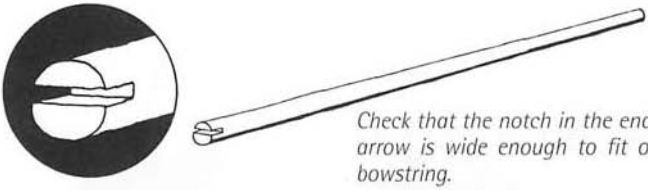

Check that the notch in the end of each arrow is wide enough to fit over your bowstring.

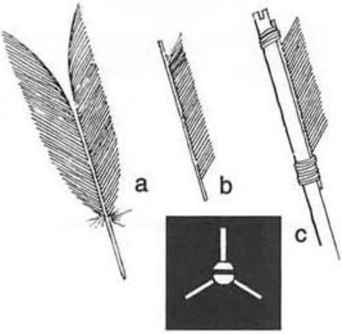

FLIGHTING ARROWS

To increase accuracy arrows should be flighted. Feathers make the best flights but other materials can be used: paper, light cloth or even leaves trimmed to shape.

(a) *Split feathers, starting from the top, down centre of quill.*
(b) *Leave 20mm (¾in) of quill at each end of feather to tie to arrow.*
(c) *Tie three flights equally spaced around shaft.*

Arrow heads

At the business end of the arrow a sharp point is needed. The arrow itself can be sharpened and hardened in fire but a firm tip is better. Tin is excellent or flint can be fashioned into a really sharp arrowhead. With patience even bone can be made into a good tip. Split the end of the shaft, insert the arrow head and bind it tightly. Sinews are good for binding – apply wet, they dry hard securing the head firmly.

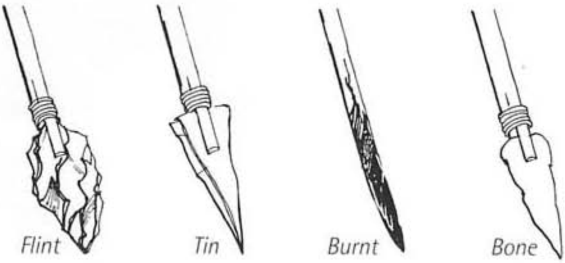

WEAPONS FROM FLINT

Arrow and spearheads, axes and knives can all be made by knapping flint, which is a black stone with a dull metallic gleam, often found in association with chalk. Choose a flattish piece of approximately the right shape and size. With another hard stone flake off pieces until it is the shape you want. Chip away at the edges to produce a very sharp cutting edge (see *Tools* in *Camp Craft*).

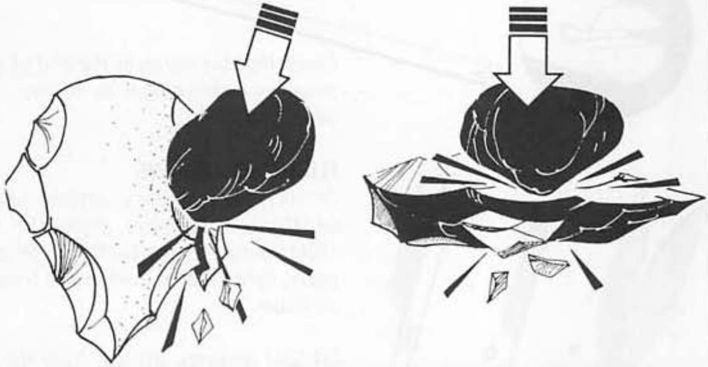

Archery technique

Fit an arrow into the bowstring and raise the centre of the bow to eyelevel. Hold the bow just below the arrow, extending the arm forwards. Keep the bow arm locked and draw the string smoothly back across the front of your body, with the arrow at eye level, and lined up with the target, sighting along the arrow. Release the string – just let go, do not snatch at it as you do so. Now, practise! For rapid fire carry a number of arrows in your bow hand.

ARROW BURNS

Many archers find that the rubbing of the arrow flights against the hand and the cheek can cause friction burns. A scarf or a piece of cloth pulled tight to the face will protect the cheek without interfering with the shot and either a leather mitten worn, or a leather guard fitted between the fingers and the wrist, to protect the hand.

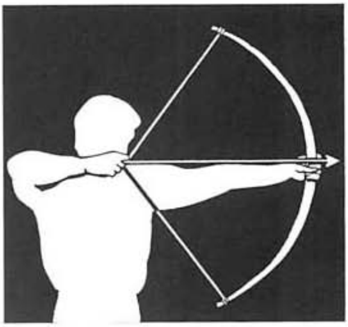

Sling and shot

The simple sling was the weapon with which David slew the giant Goliath and can be armed with ordinary pebbles. It consists of a simple pouch in the middle of a length of rope. Leather is the best material for the pouch but you could make it from any strong fabric and the rope can be a leather thong or twisted from natural fibres. Attach it as one long piece threaded through, or two tied or sewn on.

SLINGSHOT TECHNIQUE

Select smooth pebbles about 2cm (3/4in) across and as round as possible (jagged pebbles might do more damage but they will not follow such a smooth trajectory).

Swing the sling above the head in a circle lined up on your target. Release one end of the rope and the ammunition should fly with great velocity and, with practice, accurately on target. You will probably need to experiment with sling length to achieve accuracy and distance.

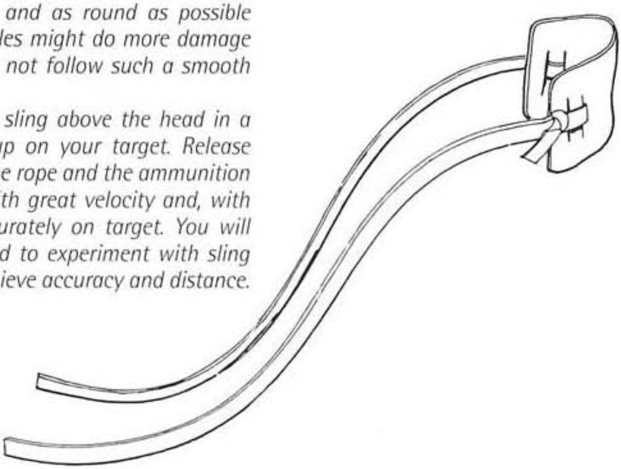

When using either the sling or the catapult against birds, load several pebbles at once.

Catapult

The schoolboy's weapon – but the Romans used giant mechanical ones as siege weaponry. You need a strong forked twig, preferably with some pliability – a hazel prong is excellent – and a piece of elastic material. A piece of innertube from a car or bicycle tyre is ideal (and stronger than the elastic in your clothing, although that could be used).

Place a pouch in the centre of the elastic (a boot tongue can be used) and thread or sew it into position as for the sling, tie the ends to each side of your twig and use a stone as your missile.

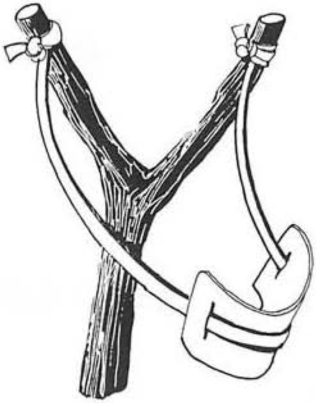

Bola

A weapon that Eskimos use against birds. Stones are wrapped in circles of material and 90cm (3ft) lengths of string knotted around each, the other ends of the string being firmly tied together. Held at the joined end, they are twirled around the head. When released they fly through the air covering a wide area.

The Gauchos of South America use the same weapon and variations have been used in combat.

The bola wraps around a bird in flight or tangles around an animal's legs or neck, bringing it to the ground and giving the hunter a chance to kill it.

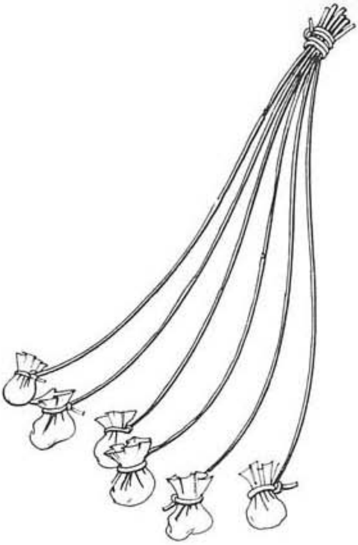

Spears

A staff is a good aid to walking and by sharpening one end can be turned into a useful thrusting or throwing weapon. A straight staff about 1.80m (6ft) is ideal for a jabbing spear. About 90cm (3ft) makes a more manageable throwing spear. A thrower can be made from a piece of wood about half that length – it gives greater accuracy and distance.

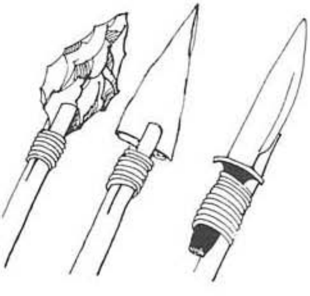

To make a spear more effective add a point of flint, knapped to sharpness, or a flattened cone of tin, set into the end – or securely bind on a knife. However, if you only have one knife do not risk it, it could too easily be lost or damaged.

SPEAR THROWER

Spear sits in a groove which runs along most, but not all, of the upper face of the thrower. The end stop adds thrust to the spear.

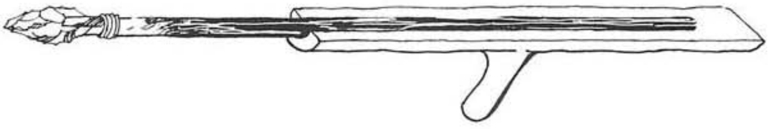

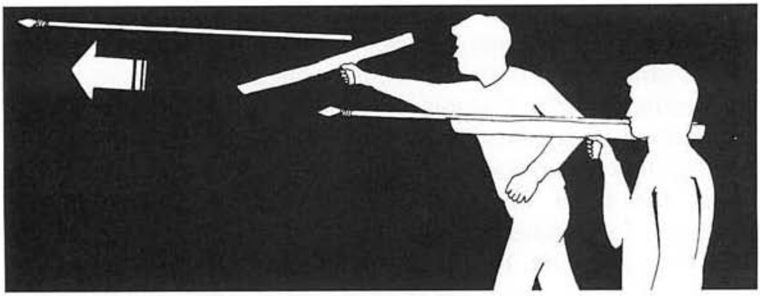

To make the thrower choose a tree limb that is at least twice the width of your spear and with a branch stump which can become the forward sloping handle. Split down the centre using a knife as a wedge. Gouge out a smooth channel for the spear. Make sure it is cleanly cut, leaving a solid portion as a buffer. Experiment to match the thrower length to that of the spear and to suit your own balance.

Hold at shoulder level, aim the spear at the target, bringing the holder sharply forwards and then downwards. As you move downwards the butt of the groove adds to the thrust behind the spear.

HUNTING BIRDS

Running noose

A noose attached to a long pole is an effective way of pulling roosting birds down from lower branches. Make a note of roosting and nest sites – remember that droppings will help guide you to them – and if they are within reach return stealthily on nights when there is sufficient light to see them. Slip the noose over the bird and pull, tightening the noose and pulling the bird down at the same time.

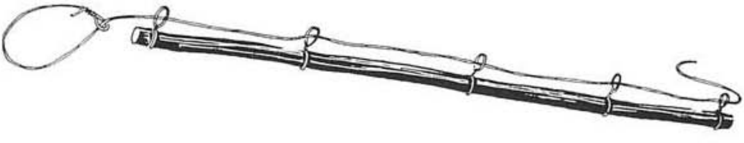

Stalking waterfowl

You can get up close by getting in the water and camouflaging yourself around the head with reeds and other vegetation. Very cautiously, approach an area where fowl nest or are regularly seen. But remember that birds (especially large ones such as geese and swans) can be quite ferocious in defence of themselves.

Another technique in some parts of the world is to use a large gourd worn on the head as cover. Holes are made on one side to breathe and see through and with the gourd just sufficiently above the water the hunter floats with the current among the birds. To prepare the birds several other gourds are thrown into the water first. Having got among the wildfowl the hunter grabs the unsuspecting birds from below and strangles them underwater.

Pit trap

Find or dig a hole about 90cm (3ft) deep in an area where ground-feeding birds are common. Its width depends upon the kind of birds you are after. Spread grain or other bait around the hole, and more concentratedly inside it.

First taking the bait around the hole, birds will enter it to get more. Rush them. In their panic they are unable to spread their wings sufficiently to take off from inside the hole.

SEAGULLS

Seagulls can be caught by wrapping food around a stone and throwing it in the air. The gull swallows the bait while still on the wing, gulping down the stone with it, and the change in weight causes the bird to crash. Obviously this is a technique for use over land rather than at sea. Be ready to dispatch the bird as soon as it hits the ground.

ANIMAL DANGERS

Few animals are likely to attack the survivor unless in self-defence, most will be much more concerned to get out of the way. There is little chance of encountering large animals such as elephants, rhinos and hippopotamus, or the larger big cats, outside wildlife reserves. Nevertheless avoid making camp on a trail, or close to an animal watering place, where you could find you are in the path of a herd of elephants or confronted by a curious cougar.

WARNING

DON'T PROVOKE AN ENCOUNTER. Bears frequently scavenge from homesteads in the northern forests, turning over garbage cans, and are just as likely to come round your camp for easy pickings. Use noise to drive them off – don't try to catch them. Don't get close to them. A bear can easily kill a man and a wounded bear is particularly dangerous. All injured or cornered animals are likely to be dangerous. Most animals will try to escape. If you prevent them from doing so, you are forcing them to fight.

Crocodiles and alligators should be given a wide berth unless very small – and then beware that there is not a larger crocodile behind you! Any of the large-horned animals is likely to be able to wound you with its horns before you can reach it with a weapon. Stags are particularly belligerent in the rutting season. It is not just hooved animals that can deliver a powerful blow with their feet – ostriches can kill with a kick.

Wolves are much more often heard than seen. Tales of large marauding packs are probably wild exaggerations. A wolf may get curious and look at you from a distance, but you can take with a pinch of salt the idea of hunters being chased by packs of ferocious wolves. If you are badly injured and unable to defend yourself, wolves might finish you off. Hyenas also hunt in packs. Although basically cowardly, they are very powerful and, as scavengers, are attracted to camp sites. They will probably turn tail and run, but drive them off rather than try to tackle them.

The larger apes can easily kill a man – but they are rarely aggressive animals and will usually give you plenty of warning to back off. Small monkeys are much more often encountered and more immediately dangerous – they have sharp teeth. Mature chimpanzees, in particular, can be very bad-tempered. Thoroughly cleanse any animal bite. All could cause tetanus and some mammals, including vampire bats, carry rabies.

Snakes will not be a threat, unless you accidentally come into contact with them. You just have to get used to them and to checking clothing, bedding and equipment for any reptile or insect visitors. Occasionally a snake or a centipede may slide into bed with you, attracted by your warmth. Cases have been known of people waking to find an unwelcome visitor nestling in an armpit or even more intimate places. Try to remember that they are not attacking. Move gently and calmly to free yourself from them.

HANDLING THE KILL

Wounded and trapped animals can be dangerous. Before approaching closely check whether an animal is dead.

Use a spear or tie your knife to a long stick and stab a largish animal in its main muscles and neck. Loss of blood will weaken it, enabling you to move closer and club it on the head.

If you have a companion, it is easier to carry a large animal by tying it firmly to a bough, which can be carried on your shoulders, but you should not take it all the way into camp, where it would attract flies and scavenging animals.

Even large animals can be dragged to a more convenient location if turned upon their backs. If the animal has horns cut off its head or they will make this difficult.

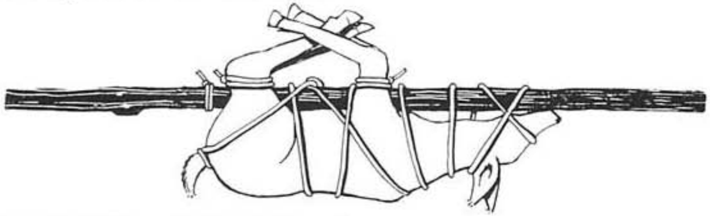

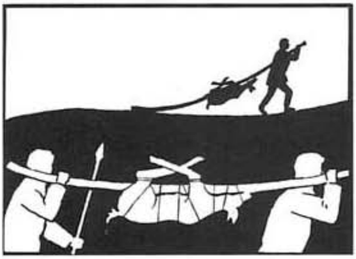

Place the pole along the belly and use a clove hitch around each pair of legs. Lash the animal to the pole and finish with a clove hitch around the pole. If the animal has horns, tie these up out of the way.

It is preferable to butcher all game on the trap line. It will attract predators and carrion eaters that in turn may become trapped. Use the entrails to rebait traps. Only carry back to camp what you can manage without exertion. In cool climates cache the rest for collection later.

HIDING THE KILL

Suspend a carcase from a bough, too high for scavengers on the ground and out of reach from the branch. A cache in the crook of a tree will keep meat away from ground predators but will still be accessible to felines and other climbing predators.

In territories with vultures and other large carrion eaters it will be almost impossible to protect it, so carry what you can. What you leave behind is unlikely to keep in a hot climate.

Blood is a valuable food, containing vital minerals. Carry a vessel for taking it back to camp. Keep it covered, cool and out of the way of flies.

WARNING

Health Hazards – diseased animals

There are lymph glands in the cheeks of all animals (more noticeable on large ones). If large and discoloured they are a sign of illness. Any animal that is distorted or discoloured about the head (such as a rabbit with the symptoms of myxomatosis) should be boiled – there is then little risk of infection from eating it – but care should be taken in preparation when there is a risk. It is essential that any cut or sore in your skin be covered when slaughtering or handling meat, for if an animal carries disease a break in the skin provides easy entry to your body.

PREPARING THE KILL

No part of a carcass should be wasted. Careful preparation will give you the maximum food value and make full use of the parts you cannot eat. Set about it in four stages:

- **Bleeding**, which is essential if the meat is to keep, and without which the taste is very strong.
- **Skinning**, so that the hide or fur can be used for shelter and clothing. (Pigs are not skinned because they have a useful layer of fat under the skin. Birds are plucked but not usually skinned.)
- **Gutting**, to remove the gut and recover offal.
- **Jointing**, to produce suitable cuts for cooking by different methods.

Bleeding

Do not waste blood. It is rich in vitamins and minerals, including salt, that could otherwise be missing from the survivor's diet. Cattle blood is an important part of the diet of many African herdsmen. Cannibals who drank their enemies' blood found vision and general health improved, and giddy spells, induced by vitamin deficiency, cured: the blood provided the missing vitamins and minerals.

Any animal will bleed better if hung with the head down. Tie ropes around the hock (NOT the ankle – it will slip off) and hoist it up a branch or build a frame, placing a receptacle beneath to catch the blood.

For a frame you need a strong structure. Drive the posts into the ground and lash them firmly where they cross to make A-frames and, then rest the horizontal bar on top.

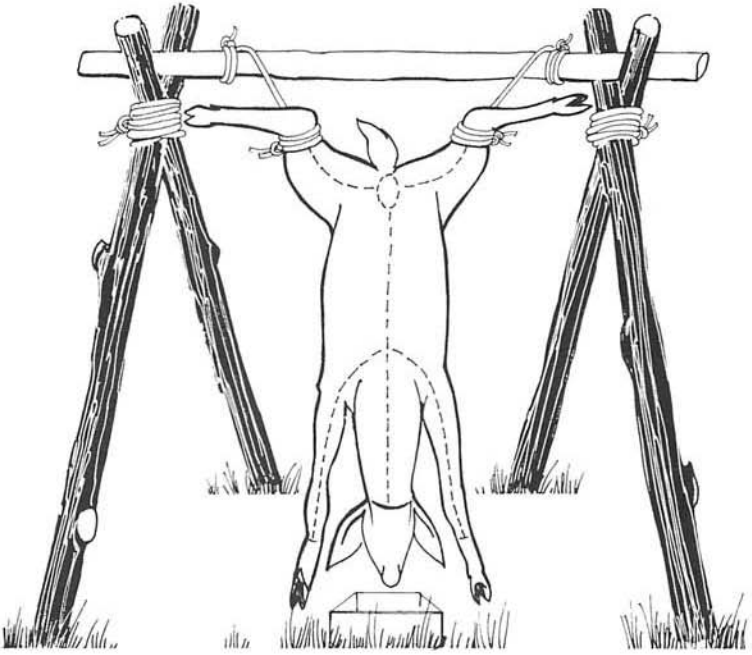

Bleed the animal by cutting the jugular vein or carotid artery in its neck. When the animal is hanging these will bulge more clearly and should be easy to see. The cut can be made either behind the ears, stabbing in line with the ears to pierce the vein on both sides of the head at the same time, or lower down in the V of the neck, before the artery branches. Unless you have a stiletto type knife the latter is best. An alternative is to cut the throat from ear to ear. This has the disadvantage of cutting through the windpipe and food from the stomach may come up and contaminate the blood which you are trying to save, but if your knife does not have a sharp point it may be necessary.

It is particularly important to very thoroughly bleed pigs. If blood remains in their tissues, which have high moisture and fat content, it will speed deterioration of the flesh.

Skinning

It is easier to skin any animal when the flesh is still warm, as soon as it has been bled. First remove any scent glands which might taint the meat. Some deer have them on their rear legs, just behind the knee. Felines and canines have a gland on either side of the anus. It is wise to remove the testicles of male animals, as they can also taint meat. Before attempting to remove hide, cut firmly through the skin, as shown by the broken line in the main illustration:

1 Make a ring cut around the rear legs just above the knee. Take care not to cut the securing rope.
2 Cut around the forelegs in the same place.
3 Cut down the inside of the rear legs to the crotch, carefully cutting a circle around the genitals.
4 Extend the cut down the centre of the body to the neck. Do not cut into the stomach and digestive organs: lift skin and insert two fingers beneath, set knife between them, sharp edge outward and draw it slowly down, cutting away from the body. (See illustration.)
5 Cut down the inside of the forelegs.

Cutting in this way, you avoid cutting prematurely into the gut cavity. The fingers lift the skin as you go and the knife, sharp edge outwards, slips in and cuts along. Do not hurry. Do not cut yourself. Do not damage the skin. Taking care will pay dividends later when you want to use the skin.

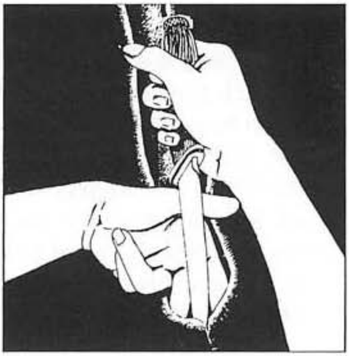

Now ease the skin of the rear legs from the flesh. Use the knife as little as possible. Roll the skin outwards, the fur inside itself, and pull it down.

Having cleared the back legs, cut around the tail (you have already cut around the genital area). As soon as you can get your hand right down the back of the carcass use your fingers to separate flesh from skin. Now peel the skin from the front legs. You will have a single piece of hide. As you work your thumbs down the neck they become bloody at the point where the throat was cut. A strong twist of the head will separate it. Cut through remaining tissues.

Working on your own

Lifting a large animal takes considerable effort. If on your own you may have to skin and gut the animal on the ground. To prevent the carcass from rolling, cut off the feet of hooved animals and place them under it.

Lay the carcass down a natural slope, scoop an impression in the ground in which to place a collecting tin or other vessel so that the animal bleeds into it. Follow the same pattern of incisions in the hide then skin the animal from one side to the backbone, spread out the hide and then roll the animal onto it to finish skinning the other half – this helps keep the meat from rubbing on the ground.

Skinning small animals

Rabbits and smaller animals can be skinned by making a small incision over the stomach (be careful not to cut into the organs). Insert the thumbs and pull outwards – the skin comes away easily. Free the legs and twist the head off.

If you have no knife available to make the first incision snap off the lower part of a leg and use the sharp edge of the break to cut the skin.

Gutting

With the carcass still suspended remove the gut and recover the offal. Pinch the abdomen as high as possible and in the pouch of flesh you have raised make a slit big enough to take two fingers. Do not stab into the flesh or you may cut through to internal organs. Insert the fingers and use them as a guide for the knife to cut upwards towards the anus. Now cut downwards in the same way, using the hand to hold back the gut, which will begin to spill outwards (see illustration). Cut down as far as the breastbone.

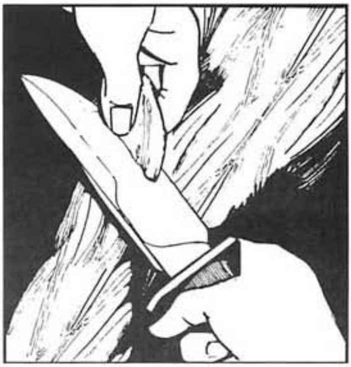

The initial incision, made in the pinched-up flesh, need only accommodate two fingers. Cut in the same way as skinning (previous illustration). First up, then down (the back of the hand prevents the gut from spilling).

Let the gut spill out, allow it to hang down so that you can inspect it. Remove the two kidneys and the liver.

The chest cavity is covered with a membrane and easily missed in small game. Cut through the membrane and remove heart, lungs and windpipe.

Ensure that the anus is clear – you should be able to see daylight through it (push a hand through with large animals). The carcass is now clean.

Jointing meat

Large animals can be quartered by first splitting down the backbone and then cutting each side between the tenth and eleventh rib. The hindquarters will contain the steaks (rump and fillet) and the choicer cuts, the forequarter meat is more stringy and needs slow cooking to make it tender.

The cuts into which a carcass is divided will differ according to the kind of animal and the cook's preference.

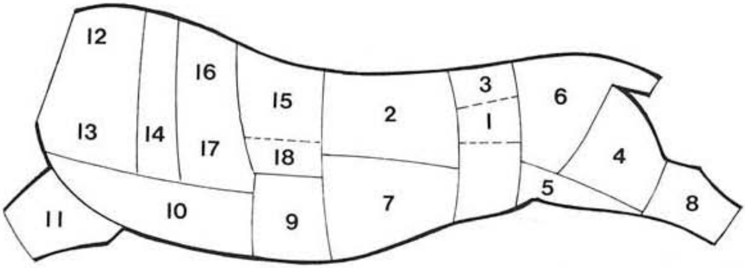

1 Fillet *or* **undercut** *The most tender meat – only 1 per cent is fillet. Ideal for preserving.*

2 Sirloin *Next most tender. Fat free strips can be cut for preserving.*

3 Rump *Ideal for frying, little cooking needed. Can also be dried in strips.*

4 Topside *Muscle from the top of the leg. Cook slowly, it tends to be tough. Cut into cubes for boiling.*

5 Top rump *Muscle from front of thigh. As for topside.*

6 Silverside *Muscle on outside of thighs. Good for roasting.*

7 Hind flank *Belly, ideal for stews and casseroles.*

8 Leg *Tough and sinewy, cut into cubes and stew.*

9 Flank *Muscular extension of the belly. Ideal for stews. Usually tough so needs long simmering to make tender.*

10 Brisket *Same as Flank.*

11 Shin *Foreleg, best cubed for stews.*

12 Neck

13 Clod *Ideal for stews. Contains less tissue than leg. Cook slowly.*

14 Chuck and blade *Quite tender but usually cut up as stewing steak.*

15–18 Ribs *Ideal for roasting but cook slowly.*

Hanging

Offal should be eaten as soon as possible but the rest of the meat is better hung. In moderate temperatures leave the carcass hanging for 2–3 days. In hot climates it is better to preserve it or cook it straight away.

When the animal is killed, acids released into the muscles help to break down their fibre, making the meat more tender. The longer it is left the more tender it will be and easier to cut, with more flavour too, and harmful parasitic bacteria in the meat will die. You must keep flies off the flesh: if they lay eggs on the meat it will quickly spoil.

OFFAL

Liver

Liver is best eaten as soon as possible. Remove the bile bladder in the centre. It is quite strong and can usually be pulled off without difficulty – but be careful, the bile will taint flesh with which it comes in contact. If any animal has any diseases they will show up in the liver. Avoid any liver that is mottled or covered in white spots. If only some is affected, cut it off and eat the remainder.

Liver is a complete food, containing all essential vitamins and minerals. If eaten raw no food value is lost. It requires little cooking.

Stomach (Tripe)

Stomach (tripe) takes little digesting, so is a good food for the sick or injured. Remove the stomach contents (which make ideal 'invalid' food) wash the tripe and simmer slowly with herbs. The contents may sound unpalatable but could save an injured person's life, for the animal has done most of the hard work of breaking the food down. Lightly boiled, stomach contents are nourishing and easily digestible.

In some countries pigs are fed nothing but apples prior to slaughter. They are cooked with the stomach still in. The subtle flavour of apple impregnates the meat. The stomach is removed after cooking and the contents used as sauce.

Kidneys

Kidneys are a valuable source of nourishment and ideal flavouring for stews. Boil them with herbs. The white fat surrounding them (suet) is a rich food source. Render it down to use in the preparation of pemmican.

Melts

Melts are the spleen, a large organ in the bigger animals. It has limited food value and is not worth bothering about in small game such as rabbits. It is best roasted.

Lites

Lites are the lungs of the animal, perfectly good to eat but not of great food value. Any respiratory complaints will show up in the lungs. Do not eat any mottled with black and white spots. Healthy lungs are pink and blemish free and best boiled. They could be set aside for fish or trap bait.

Heart

Heart is a tightly packed muscle with little or no fat. Roast it or use its distinctive flavour to liven up stew.

Intestines

Intestine consist of lengths of tubes and are best used as sausage skins. Turn them inside out and wash them. Then boil them thoroughly. Mix fat and meat in equal proportions and then stir in blood. Stuff the mixture into the skin and boil them well. Before putting them into boiling water add a little cold to take it just off the boil – this will counter any risk of the skins bursting. This makes a highly nutritious food which, if smoked, will keep for a long time. Dried intestines can be used for light lashings.

Sweetbreads

Sweetbreads are the pancreas or thymus gland, distinctive in larger game. Many people consider it a great delicacy and it is delicious boiled or roasted.

Tail

Skin and boil to make an excellent soup for it is full of meat and gelatine.

Feet

Feet are chopped off during slaughter but should not be wasted, boil them up to make a good stew. Clean dirt from hooves or paws and remove all traces of fur. Hooves are a source of nutritious aspic jelly.

Head

On larger animals there is a good deal of meat on the head. The cheeks make a very tasty dish. The tongue is highly nutritious. Boil it to make it tender and skin before eating.

All that is left, or the whole head with small animals, should be boiled.

Brain

The brain will make brawn and will also provide a useful solution for curing hides.

Bones

All bones should be boiled for soup. They are rich in bone marrow, with valuable vitamins. They can also be made into tools.

PREPARING SHEEP-LIKE ANIMALS

Follow the instructions for larger animals and then:

1. Split in two down the line of the spine, keeping exactly to the centre of the backbone.
2. Remove rear leg. Try to cut through the ball and socket joint.
3. Remove front leg. There is no bone to cut through, follow the line of the shoulder blade.
4. Cut off neck.
5. Cut off skirt (loose flesh hanging below the ribs).
6. Cut between each rib and between the vertebrae. This gives you chops.
7. The fillet, lying in the small of the back, is the best meat for preserving.

Preparing Pig

Do not attempt to skin a pig. Gut it first then place it over the hot embers of a fire and scrape the hair off. Hot water – just hotter than your hand can bear – will help to loosen the hair. Water that is too hot will make the hair more difficult to remove. Pigs attract many parasites: ticks, crab lice and worms so cooking must make sure of killing them. Boiling is therefore the best way of cooking pork.

Preparing small animals

Follow the same basic procedure as for larger animals – they all need to be gutted.

Preparing reptiles

Discard internal organs, which may carry salmonella. Reptiles can be cooked in their skins, and large snakes can be chopped into steaks and their skins can be used for clothing. To prepare a snake cut off head well down, behind poison sacs; open vent to neck, keeping blade outwards, to avoid piercing innards, which will fall clear. Skewer to suspend and ease of skin towards tail.

Preparing birds

Birds are prepared in much the same way as animals – though they are usually plucked and cooked with the skin on instead of being skinned. Follow the sequence below.

Bleeding:

Kill birds by stretching their necks, then cut the throat and hang head down to bleed. Or kill by cutting just under the tongue, severing main nerve and main artery. The bird dies easily and bleeds well. Handle carrion eaters as little as possible – they are more prone to infection, lice and ticks.

Plucking:

Plucking is easiest straight after killing while the bird is still warm. Hot water can be used to loosen feathers, except in the case of waterbirds and seabirds, in which it tends to tighten them. Keep feathers for arrow flights and insulation. Start at the chest. For speed you can skin a bird – but that wastes the food value of the skin.

Drawing (removing innards):

Make an incision from the vent to the tail. Put your hand in and draw out all the innards. Retain the heart and kidneys. Cut off the head and feet.

Cooking:

Always boil carrion eaters in case they carry any disease. Boiling will make stringy old birds tender but you can roast younger ones on a spit or in an oven.

FISH AND FISHING

Fish are a valuable food source, containing protein, vitamins and fats. All freshwater fish are edible but some tropical ones can be dangerous: be wary of electric eels, freshwater stingrays and the piranha of South American rivers. It takes skill to catch fish by conventional angling methods but, by considering their feeding habits and following the simple methods given here, you can be successful.

Fish range from tiny tiddlers to some of prodigious size – such as the Nile perch of the tropics. They differ widely in their eating habits and diet. Different kinds feed at different times and at different levels in the water. Some prey on other fish, others eat worms and insects, but they can all be attracted and hooked with appropriate bait.

If you are an experienced angler you can apply your skills, especially if you have plenty of time on your hands, and will probably gain a lot of pleasure, but if you are fishing for survival the sporting angler's techniques are not the most effective. Two of the most useful methods are the night line and the gill net.

Where to fish

Fish choose the places in the water where they are most comfortable and where they most easily find their prey. This will be affected by the temperature of the day.

If it is hot and the water is low, fish in shaded water and where there are deep pools. In a lake fish retreat to the coolness of deep water in hot weather.

In cold weather choose a shallow place where the sun warms the water. Lake fish will tend to keep to the edges which are warmer.

If the river is in flood, fish where the water is slack – on the outside of a bend – for example, or in a small tributary feeding the mainstream if its flow is different – quite possible for the flood may not be due to local rainfall.

Fish like to shelter under banks and below rocks and submerged logs.

When to fish

As a general rule leave lines out overnight and check them just before first light. Some fish feed at night during a full moon.

If a storm is imminent, fish before it breaks. Fishing is poor in a river after heavy rain.

Indications of fish feeding

Signs that fish are feeding, and therefore likely to take a bait, are when they jump out of the water, or you see frequent clear ring ripples breaking out where fish are taking flies on the surface. Where lots of little fish are darting about they may well be being pursued by a larger predatory fish.

IMAGE REFRACTION

Water refracts light so that the fish sees things above the water at a slightly different angle – and can probably see more on the bank than you think. It is always better to fish from a sitting or kneeling position than standing up so that you are less likely to be in vision. Keep back from the edge. Always try to keep your shadow off the water you are fishing.

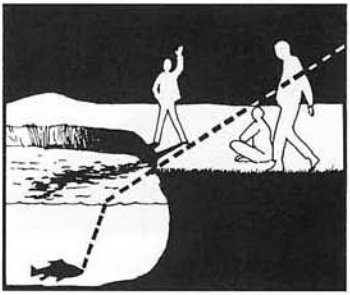

ANGLING

Fishing with a hook and line is the popular way of fishing (though others are usually more effective) and they are part of your survival kit. Hooks can also be improvised from wire, pins, bones, wood and even thorns.

Large hooks will catch large fish but small ones will catch both large and small. Near the end of the line you will probably need to attach another short length with a weight to take the hook down and stop the line being carried along the surface of the water, especially if fishing deep. If it is a long line you also need another length with a float which will be pulled down when you get a bite. A rod is not essential (you can fish effectively with a handline) but makes it easier to land fish and to cast away from the bank.

You can improvise hooks from all kinds of materials. Here (from left to right) a pin, a thorn, a bunch of thorns, nails, bone and wood have been used.

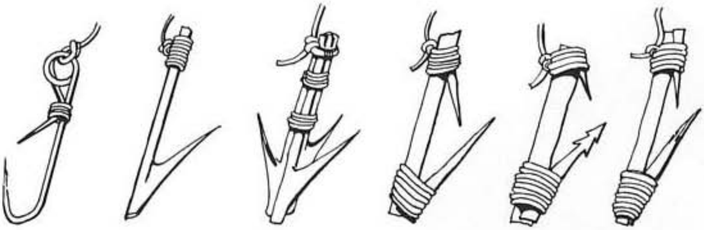

Using floats and weights

Along the line attach a small floating object, easily visible from the bank, and you will be able to see when you have a bite. Its position will help control where the line descends.

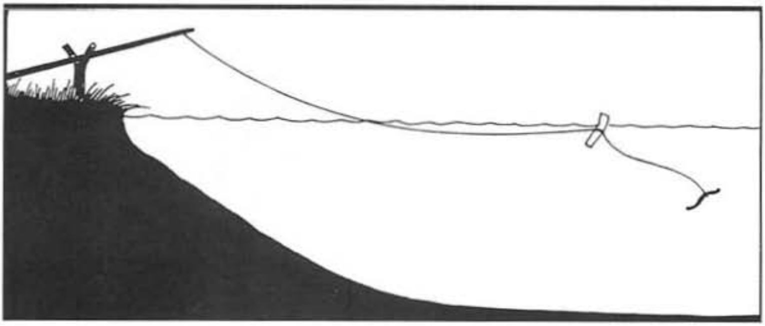

Small weights between the float and the hook will stop the line from trailing along the water or at too near the surface in a current, still leaving the hook itself in movement. You have small split lead shot in your survival kit. Slip the groove along the line and squeeze in to fit closely.

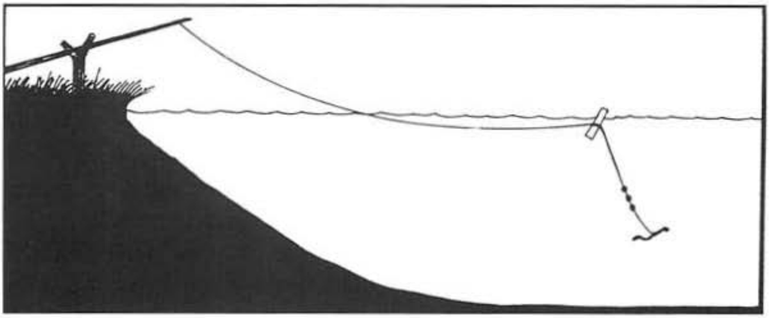

A deeper hook position can be ensured by extending the line to a weight below the hook.

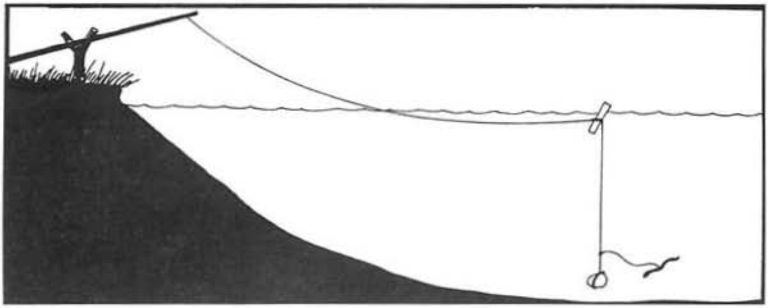

ANGLING WITHOUT HOOKS

You do not even have to use a hook to fish with a line. To catch eels and catfish tie a blob of worms on a line (a). These fish swallow without biting so swallow the bait with line attached. Pull them out as soon as the bait is taken.

Instead of a hook you could use a small sharp piece of wood tied on the end of the line and held flat along it by the bait (b). When the bait is swallowed the wood will open out and lodge across the gullet of the fish (c).

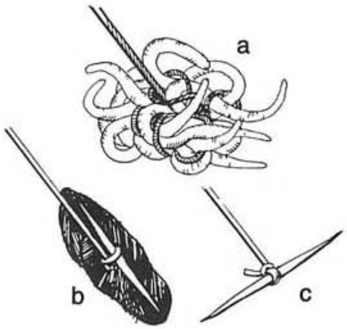

Bait

Bait native to the fishes' own water is most likely to be taken: berries that overhang it, insects that breed in and near it. Scavenger fish will take pieces of meat, raw fish, ants and other insects. Once you have a catch examine the stomach contents of the fish and eliminate the guesswork as to diet. If one bait is unsuccessful, change to another.

Ground bait

Bait scattered in the area you want to fish, will attract fish to it. A termites' or ants' nest suspended over a river is one excellent method. As the insects fall into the river the fish will take them. Bait your hook with them as well and success is sure. Any suitable bait, scattered on the water, can be used to draw fish but it is always best to put the same bait on your hook.

SPINNING

Curious fish will attack a shiny object drawn through the water: try coins, buttons, pieces of tin can, buckles – anything that glitters. Make a propeller shape to thread onto a piece of wire and it will spin with the current. Attach a hook to the end of the spindle.

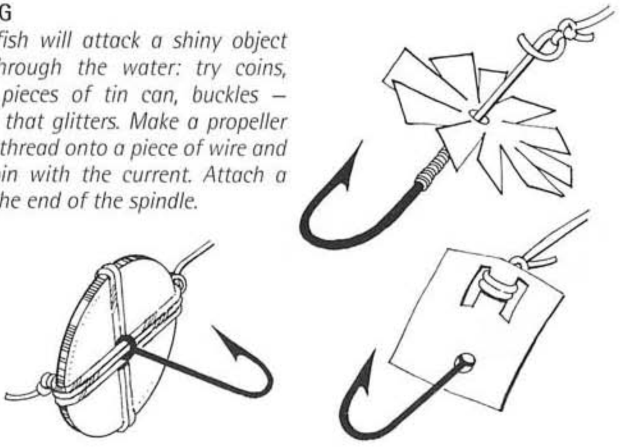

ARTIFICIAL BAIT

Can be made from brightly coloured cloth, feathers and shiny metal. Try to make them look like real bait. A few feathers tied to a hook with thread can simulate a fly, or carve a small fish out of wood and decorate it with colour or glitter (if you make it jointed it will move more naturally). Try to make lures move in the water like live bait. Hazel wood has a soft pith and can easily be threaded through so that you can link segments which will wiggle in the water.

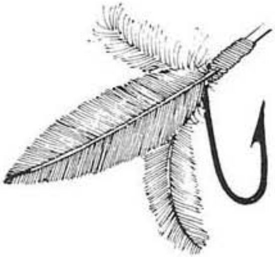

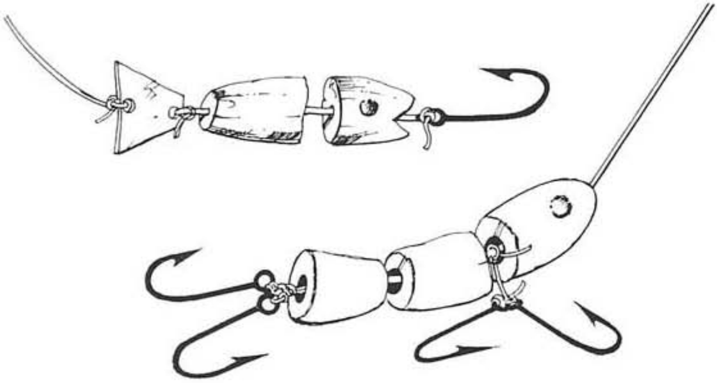

LIVE BAIT

Worms, maggots, insects and small fish can be used as live bait. Cover the hook completely with the bait. You can place the hook through the meaty part of small fish without killing them, or through the body of a grasshopper. Their distressed movement in the water will attract the fish. Tiddlers are easy to catch so you can 'use a sprat to catch a mackerel.'

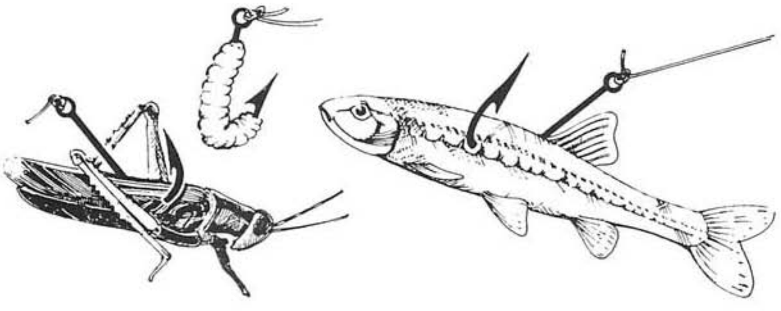

NIGHT LINES

Weight one end of a length of line and attach hooks at intervals along it. Bait them with worms. Lowered into the water this gives you the chance of catching surface-, mid- and bottom-feeders. Anchor the free end securely on the bank.

You can put this out at night and leave it until the morning – use it in daytime too – but change the worms at intervals, even if you haven't got a catch, because fresh wriggling worms will attract more attention.

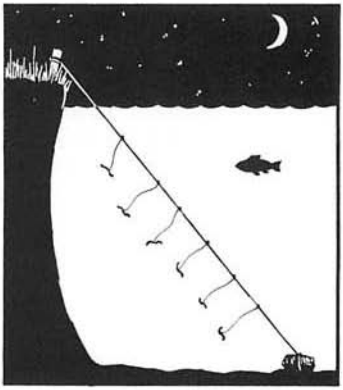

OTTER BOARD

To fish far from the bank, further than you can cast a line – in a lake, for instance, where fish are feeding in the centre – make a board with a moveable, pivoted rudder. Set a bar at the front end of the rudder to which two control lines can be attached. Beneath it suspend baited hooks. Float the board out into the lake.

If winds are favourable you could mount a sail, but then a stabilizing keel will also be needed to stop it blowing over. Gouge holes to fix dowel supports (in water the dowels will expand to make a tight fit) and tie on a flat stone – a big keel might conflict with the rudder.

Undue movement of the board will indicate a bite.

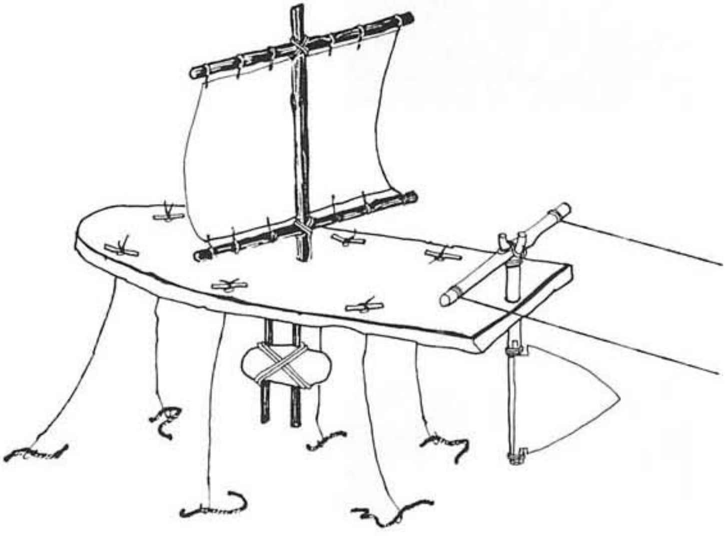

JIGGING OR SNAGGING

This is the art of hooking a fish anywhere on its body. It is a good method to use when you can see fish but they are not taking bait. Tie a number of hooks on to a pole and lower it into the water. Suspend a bright object about 20cm (8in) above the pole, and when fish go to inspect the glitter pull the hooks up sharply so that they catch on the fish.

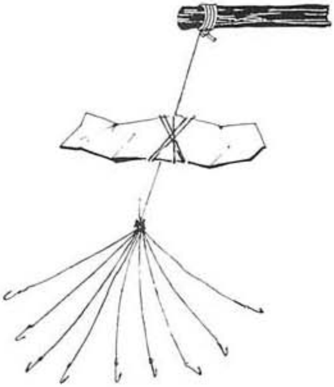

FISH TRAPS

You can make a wide variety of traps, from ones across an entire stream, which you can drive fish into, to bottle traps to capture tiddlers in. Arrows indicate current.

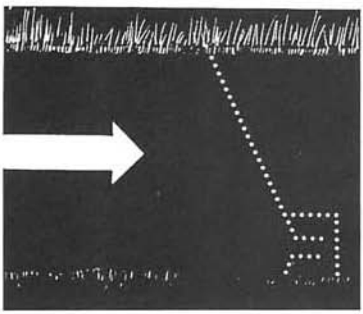

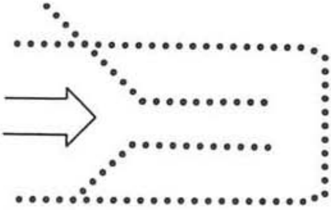

In shallow streams build a channel of sticks or rocks that fish can swim into but not turn around in.

Bottle trap

If you have a plastic bottle you can make an efficient trap for small fish by cutting it off just below the neck and then inverting the neck inside the bottle. Fish swim in but cannot find their way out again. Bait the trap to entice them in.

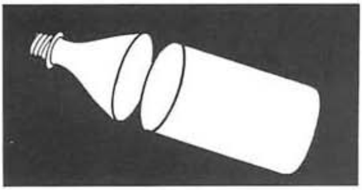

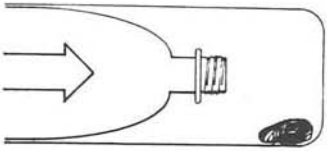

You can make a similar trap for larger fish using a hollow log. Make a lattice cone of twigs for the entrance and block the other end of the log.

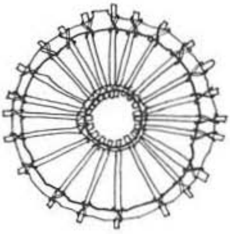

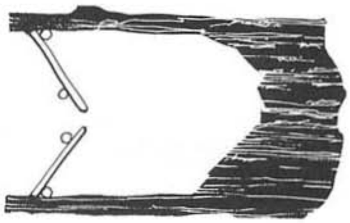

Wickerwork traps

Use young hazel, or other pliant twigs – bamboo bends better if you warm it – to make a trap into which fish can swim but from which they can find no way out. A wickerwork trap allows the current to flow through it, and since it is made of natural materials may seem like a tangle of reeds or stream-bottom debris.

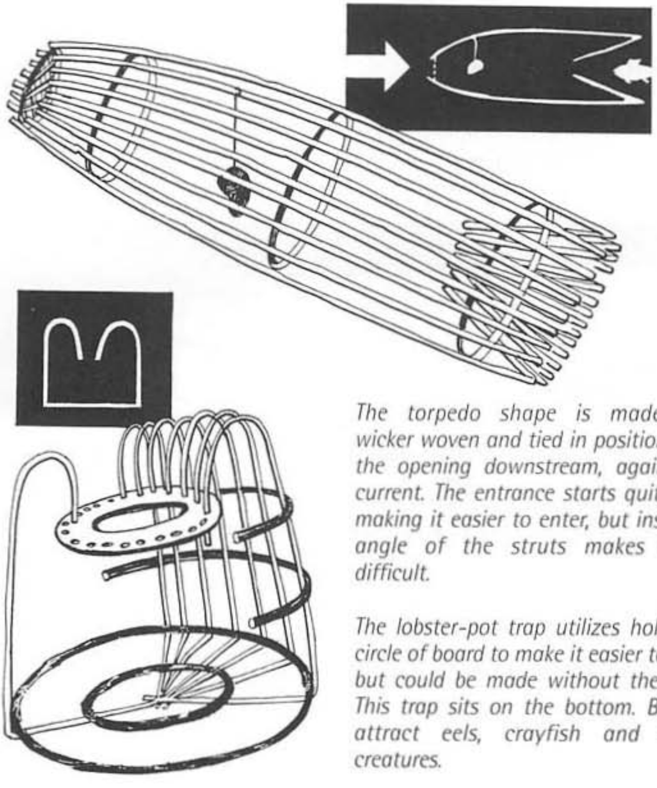

The torpedo shape is made from wicker woven and tied in position. Place the opening downstream, against the current. The entrance starts quite wide, making it easier to enter, but inside the angle of the struts makes escape difficult.

The lobster-pot trap utilizes holes in a circle of board to make it easier to shape but could be made without the board. This trap sits on the bottom. Bait will attract eels, crayfish and similar creatures.

Fish snares

Large fish such as pike, which lay alongside weeds, can be caught in a noose. Fix a noose line to the end of a pole, or pass it down the inside of a length of bamboo. Pass it over the fish from the tail end and pull it up sharply so that the noose traps the fish.

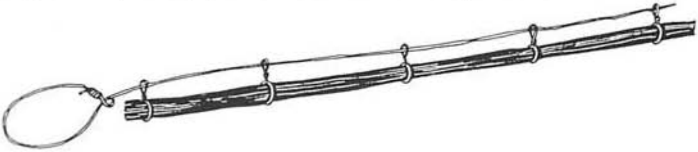

Eel bag

Tie fresh surplus offal or a dead animal inside a sack or cloth bag (plastic will not do) together with a quantity of straw-like vegetation or bracken. Tie a line and a weight to the end of the bag and allow it to sink. Leave it overnight and pull it out in the morning. If there are eels in the water they will chew their way into the bag to get at the offal and will still be wriggling in the straw when you get the bag landed.

Damming

Build a dam across a stream, diverting the flow to one side, and with rocks create a small shallow pool downstream where fish swimming upstream will be trapped. Fix a net below the race at the side of the dam to catch any fish that are carried over it from upstream.

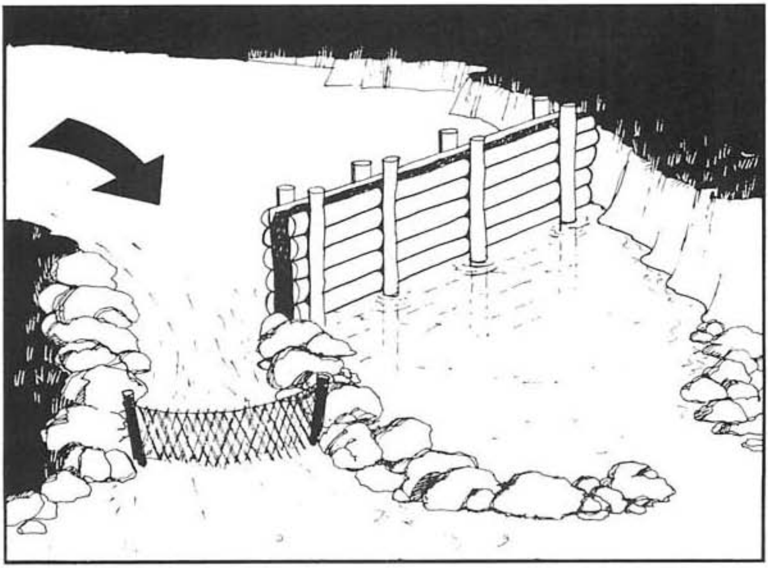

Gill net

Make a net with a mesh size of about 4cm (1½in) between knots (see *Netting* in *Camp Craft*), set floats at the top and weight the bottom, then stretch it across a river. Fish swimming into it get caught by the gills. It is lethal and will soon empty a stretch of water so should not be used for long in an area where you intend to stay (or in a non-survival situation). If the ends of the net are tied to the banks at both top and bottom, weights and floats will not be needed.

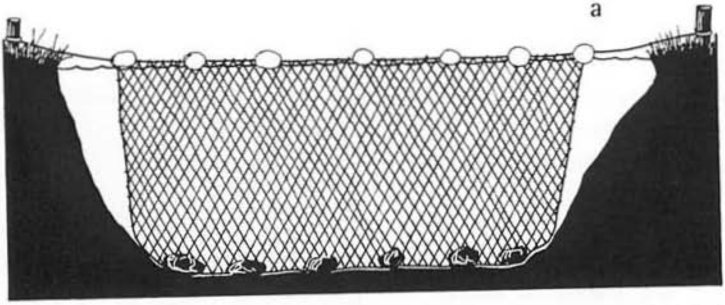

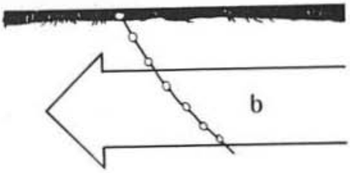

A gill net can be anchored on each bank, (supported by weights and floats (a), or tied to fixed posts. If it is angled across the line of the current (b) there is less likelihood of driftwood building up against it.

OTHER TECHNIQUES

If all else fails try the following methods.

Tickling

This is an old poacher's technique which takes patience but is effective where fish shelter below the undercut banks of fairly shallow streams. Lie along the bank and lower your hands gently into the water so that they can adjust to the water temperature. Keeping your hands as close to the bottom as possible, reach under the bank, moving the fingers slightly, until you touch a fish. Work the hand gently along its belly (fish usually swim against the current when feeding) until you reach the gills. Then grasp the fish firmly and pull it out.

Attracting and driving fish

At night a torch or firebrand held above the water will attract fish. Nets can then be drawn around the area to trap the fish which can be speared or clubbed. A mirror or other shiny material placed on the riverbed will reflect either sun or moonlight and attract fish.

Spearing and shooting fish

Sharpen a long stick to make a spear, adding barbs to make it more effective. If you have multiple points, like Neptune's trident, you give yourself a wider margin of error since it covers a larger area. Try to get above the fish and strike down swiftly.

Make sure that you are not casting a shadow over the fish you are trying to catch. Aim slightly below the fish to allow for the refraction of its image at the surface.

If you are a good shot use a bow and arrow to shoot fish (the wooden shaft will float and help to bring your prey to the surface, though most dead fish will float, anyway).

WARNING

If you have a firearm and plenty of ammunition it is worth trying to shoot fish with a gun but NEVER fire with the barrel actually in the water – or it will explode.

The water seals the end of the barrel and instead of the bullet rushing outwards the force of the detonation blows back at you. It is not just dangerous – it is potentially lethal. Make very sure that the barrel is clear of the water.

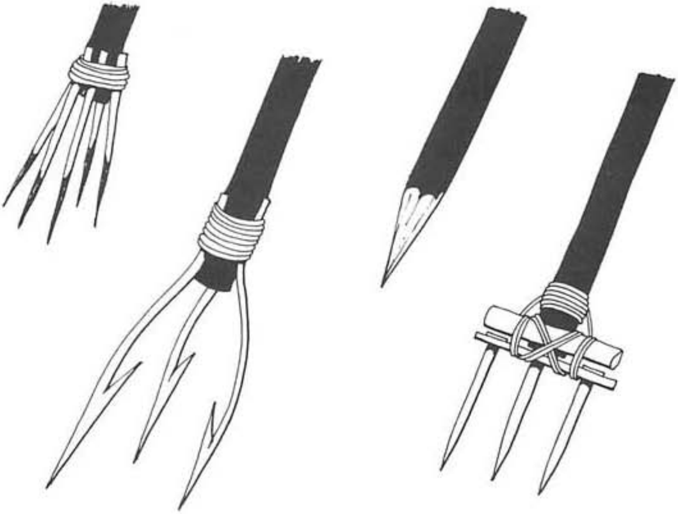

Muddying

Receding floodwater leaves isolated pools which are offer abundant in fish. Stir up the mud at the bottom of these pools with a stick, or by stamping in them. If there are any fish they will try to reach clearer water. Scoop them out.

Explosives

Explosives can be used in the water. They will kill the nearest fish, but by liberating the oxygen in the water will also cause those further away to surface.

FISH NARCOTICS

In many parts of the world fishermen use local plants to poison or stupefy fish to make them come to the surface, where they are easy to collect. This works best in a deep pool where one end can be dammed to contain the fish, but the method can be effective in any slack water.

Some plants daze or narcotize or intoxicate the fish but most have the effect of taking oxygen from the water so that the fish come to the surface in search of aerated water. Various parts of plants are used but in many cases they are simply crushed and thrown into the water. Although the effects are catastrophic for the fish they are not long lasting – the water soon reoxygenates itself. Most of these narcotics take effect more quickly in warmer waters and they are most widely used in tropical countries.

If these methods are used in closed pools you will have cleared out the fish supply and removed a future source of food. When a river or coastal pool is reopened to the main water, however, new fish will move in and restock it.

If seashells, snailshells or coral are burned over a very hot fire they will produce lime which can be thrown into still water to poison fish which will still be safe for human consumption.

WARNING

Dead fish floating on the surface – unless you have caused them to be there – may look like an easy meal but they may be diseased and if they have been there some time will not be fit to eat.

When released in water these poisons are toxic only to cold-blooded animals but this does not mean that they are edible. They are not. Parts of some of these plants are VERY DANGEROUS, if eaten raw. Use them on fish – not yourself – then eat the fish. Many plants are used. The following are effective and common in their areas.

DERRIS PLANTS

Derris (a) are found from Southeast Asia to Australia. They are woody, climbing, vine-like plants, usually with small oval leaflets in pairs opposite each other, purple flowers and seedpods. Powder the roots and throw them into the water. Stupified fish will rise to the surface not long afterwards.

BARRINGTONIAS

Barringtonia (b) are trees, found in the same area as Derris, across to Polynesia, and often near the coast. Crush the seeds inside their urn-shaped pods and throw them into the water.

DESERT ROSES

Adenium *(c), found in tropical and southern Africa and in parts of Arabia, are shrubs, sometimes small trees, with thick fleshy leaves. One of the most effective,* A. obesum *(illustrated) from East Africa, has spirals of bluntly oval leaves and clusters of tubular pinkish flowers. Use crushed stems and roots, which contain a highly toxic sap.*

SOAP PLANT

Amole Chlorogalum pomeridianum *(d) grows in dry open or scrubby country in western North America. It has narrow, grass-like leaves and white star-like flowers. Crush the bulbous root and throw into pools.*

GOAT'S RUE

Tephrosia virginiana *(e) grows on open ground in North America. Its surface is slightly hairy, it has many narrow leaflets and long flat seedpods. Use the crushed stems or the very poisonous roots.*

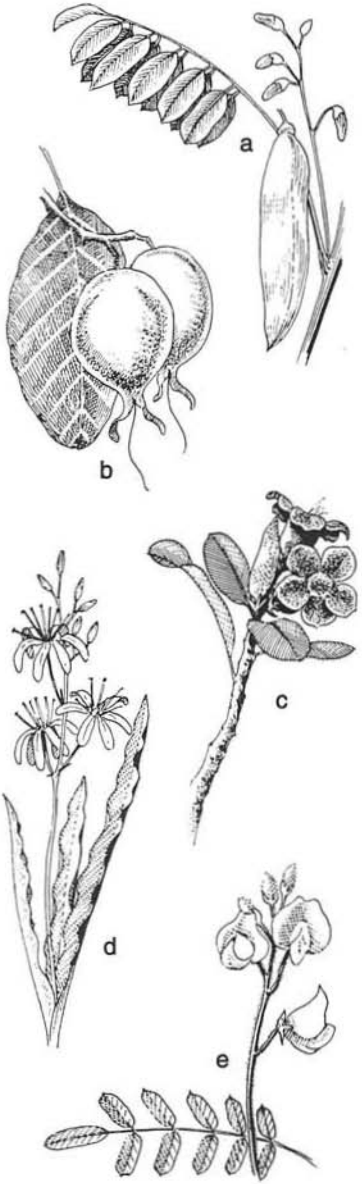

ARCTIC FISHING

On frozen Arctic seas fish are likely to be the most accessible food. Even in summer it is safer to fish through the ice than to fish from the edge of a floe which may break up beneath you. The techniques involved are equally effective on any frozen lake or river where the ice is thick enough to bear your weight with ease but not so solid that it cannot be broken through.

First you need to gain access to the water, which means smashing a hole in the ice. If you have an ice saw, use that to cut neat holes which will still leave you with firm edges. If you have to smash the ice there is a risk that it may fracture back into the area where you are standing. Approach the operation carefully.

Hook and line

Bait the hook in the usual way. If the line is being carried back up against the underside of the ice you will have to weight it below the hook.

There is no point in trying out your angling skills at only one hole – far better to set up multiple angling points. In order to cover them effectively, however, you will need an easy way of knowing when you have a bite. Make a pennant from a piece of cloth, paper or card – preferably of a bright colour so that you will see it easily against the snow and ice – and attach it to a light stick. Lash this firmly at right angles to another stick which must extend beyond the maximum diameter of your hole by at least 30 per cent. Now attach the line to the lower end of the flagpole and rest the flag on the side of the hole with the line at its centre.

When a fish takes your bait the cross piece will be pulled over the hole and the flagpole jerked upright. Keep your eye on the markers so you can pull your catch up quickly. The wriggling fish is an easy meal for a passing seal.

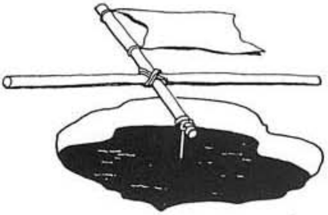

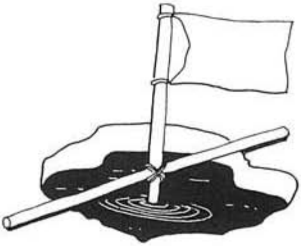

Ice netting

Although fish do not hibernate their metabolism slows down to cope with the reduced winter temperatures and they consequently eat much less, making them less likely to take bait. Netting, always likely to be more certain of producing results, has a further edge on line fishing.

Net through the ice. A net lowered from the edge of a floe would probably end up frozen to the floe and handling it would be a risky operation. Instead make several holes in the ice about 40cm (16in) wide and about twice that distance apart. Attach retaining loops to the top edge of your net at 80cm (32in) intervals, to match the holes, and weight the bottom. Put the retaining loop at one end around a stick or rod of some kind, wider than your holes, and lower one end of the net into the hole at one end of your row.

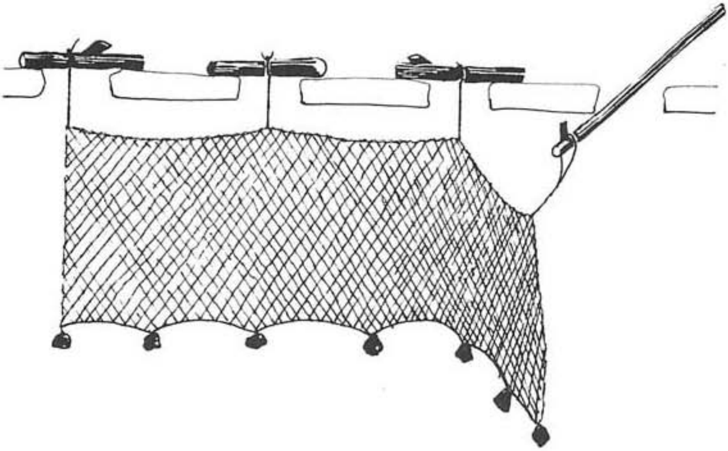

With a hooked pole (which you can improvise) you now have to fish for the net and haul it through to the next hole where you secure the next retaining loop with another retaining stick – and so on until the whole net is suspended.

If the ice is thin enough, feed all the retaining loops into the hooked pole and lower the entire net through the first hole, anchoring the first loop with a retaining stick. Then, carefully holding the loops, reach the hooked pole through the next hole and slip the remaining loops onto it. Pull the net along and anchor the next retaining loop. Continue until the net is fully extended.

To check your net pull it up with the hooked pole. If you leave your net for too long, in polar regions, you may find that your catch has been for the benefit of a seal who has stolen most of it.

PREPARING FISH

All freshwater fish are edible. Those under 5cm (2in) long need no preparation and can be eaten whole. Larger fish must be gutted. Catfish and eels are smooth skinned but others may be descaled. Catfish have a cartilage skeleton. Most other fish have a mass of bones.

Bleeding *As soon as a fish is caught cut its throat and allow it to bleed. Cut out the gills.*

Gutting *Make an incision from the anal orifice to where the throat was cut. Remove all offal – you can use it for hook bait or in an eel bag. Keep the roe, which runs down the side of the fish. It is hard in females, soft in males; it is very nutritious. This preparation helps fish keep longer.*

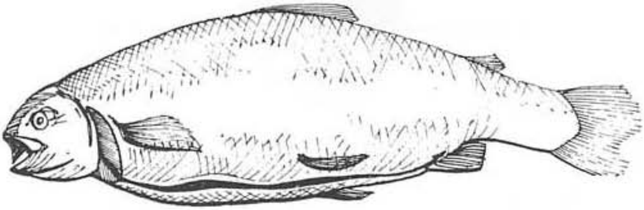

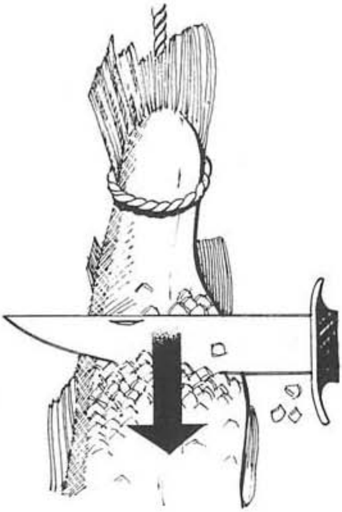

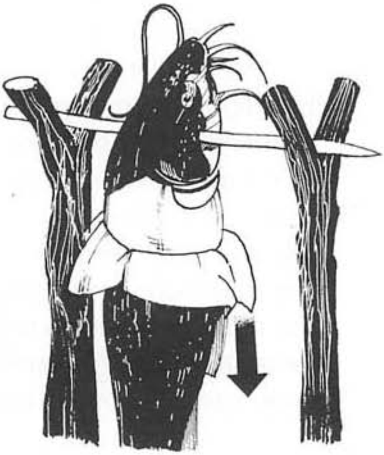

Scaling *is not necessary, fish can be cooked with scales on and it will retain its moisture. If there is time, scrape them off. Draw knife from tail to head.*

Skinning *Fish skin has good food value and should be left on and eaten unless food is plentiful. To skin eels and catfish pass a stake through the fish, lodge it across uprights and, having cut the skin away just below it, draw it down towards the tail.*

5

CAMP CRAFT

Selecting where to camp and knowing how to make a good shelter are essential skills. A wide range of shelter construction methods, using available materials, are described, from the simplest windbreak to dwellings suitable for long-term occupation.

You will need to make fire and to chose the right type of fire construction. Fires for both general warmth and for cooking purposes are detailed. Guidance on food preparation, cooking and preservation will ensure that food is safe and not wasted. Camp organization and hygiene are as important for health and morale for the single survivor as for a large group.

Methods of making tools, camp equipment, clothes, ropes and nets will all improve survival conditions and skills in knot-making will have many uses.

SHELTER AND MAKING CAMP 246
- Types of shelter 247
- Tropical shelter 255
- Arctic shelter 258
- Long-term shelter 264

FIRE 269
- Fuel 272
- Firelighting 276
- Types of fire 282

COOKING 286
- Useful utensils 292
- Preserving food 294
- Cooking tips 298

ORGANIZING THE CAMP 302
- Camp hygiene 303

TOOLS 306
- Axes 307

FURNISHING THE CAMP 312
- Beds 312

ANIMAL PRODUCTS 316
- Skin and furs 316
- Clothing 318

ROPES AND LINES 320
- Rope making 323

KNOTS 327
- Simple knots 328
- Joining ropes 329
- Loop-making 331
- Hitches 336
- Shortening rope 338
- Securing loads 339
- Lashings 341
- Net making 343
- Fishing knots 346

SHELTER AND MAKING CAMP

Shelter is necessary to give shade, to repel wind and rain and to keep in warmth. Sleep and adequate rest are essential and the time and effort you put into making your shelter comfortable will make them easier to get. If you are the victim of a plane crash or a vehicle that has let you down, it may provide a shelter or materials from which one can be built – but if there is fire or the threat of fuel tanks exploding, wait until it has burned out before attempting salvage.

If you are the unequipped victim of an accident, are trapped by unexpected mist or caught by nightfall in terrain where it is not safe to proceed, or if exhaustion or injury prevents you going further, you may have to make do with any natural shelter that you can find for the night, or until you can more fully assess the situation. In this case, virtually any protection from wind, rain and cold will be welcome. If movement down a slope seems risky, traversing even a short way along the contour may bring you out of the wind. If no cave or crevice is available to give shelter, make use of any hollow in the ground. Add to its height, if you can, by piling up rocks – but make sure that any structure is stable and use a back-pack, if you have one, to increase the windshield before settling down on the leeward side.

If there is still daylight to see by, you have no injuries to handicap you and are not isolated by unnegotiable cliffs or other barriers, it will be worth seeking possible better places in the vicinity. For a long-term camp you should find a secure site with convenient access to your major needs.

Where to camp

If you are on high exposed ground go lower down to find a sheltered spot, but on low, wet ground you will need to climb higher to find

BAD PLACES TO CAMP

1 Hilltops exposed to wind (move down and look for shelter on the lee side).

2 Valley bottoms and deep hollows – could be damp and, especially when the sky is clear, more liable to frost at night.

3 Hillside terraces where the ground holds moisture.

4 Spurs which are on a game trail.

somewhere securely dry. Look for somewhere sheltered from the wind, on rising ground that has no risk of flooding and is safe from rock falls or avalanches.

Hot air rises, cold air sinks, so valley bottoms will often contain pockets of colder air and, in cold weather, be susceptible to frost and damp mist. In areas that get plenty of rainfall, terraces across a slope will often be damper than the steeper ground above and below them, for water collects there before flowing further downward.

Ideally you should be near water, with a plentiful supply of wood near at hand. Pitching camp too close to water, however, may lead you to be troubled by insects, and the sound of running water can hide other noises which might indicate danger, or the sound of search or rescue parties.

On river banks look for the high water mark: in mountain regions streams can become torrents in minutes, rising as much as 5m (17ft) in an hour! Even on plains keep out of old watercourses, no matter how dry they are. Heavy rainfall in nearby hills can easily send water rushing down them in flash floods, with practically no warning. Choose ground that is reasonably flat and free of rocks and make sure that you have space to lay out signals and that you can be easily spotted by rescue parties.

Check above your head for bees' or hornets' nests and for dead wood in trees that could come crashing down in the next storm or high wind. Keep away from solitary trees, which attract lightning, and in forest areas keep to the edges, where you can see what is going on around you. Don't camp across a game trail – you don't want marauding animals as unwelcome guests or to find your bivouac flattened by a herd of animals on their way to a waterhole – but stay near to any obvious human tracks.

TYPES OF SHELTER

The type of shelter you build will depend upon local conditions and the materials available – and upon how long you expect to need it. For immediate protection from the elements, rig up a makeshift shelter while you construct something better and more permanent. If you decide to stay put and wait for rescue, a more long-term shelter can be built and improved on as time and energy permit.

For those walking to safety, on the other hand, temporary shelters can be built at each stopping point. They can even be carried with you if they are sufficiently light and there is a significant risk that suitable materials may not be available at the next campsite.

A more permanent shelter will certainly be worthwhile for the sick or injured, who must rest up in order to regain their strength, or where it is necessary to wait for the weather to clear before attempting a journey. Use the time to stockpile equipment and provisions.

Hasty shelters

If no materials are available for constructing a shelter make use of any cover and protection that is available: cliff overhangs, gradients and so forth, which will help shield you from wind or rain. Incorporate natural windbreaks in quickly constructed shelters. In completely open plains, sit with your back to the wind and pile any equipment behind you as a windbreak.

Bough shelters

Make use of branches that sweep down to the ground or boughs that have partly broken from the tree to give basic protection from the wind – but make sure that they are not so broken that they could come down on your head! Weave in other twigs to make the cover more dense. Conifers are more suited to this technique than broad-leaved trees, as they require less weaving-in to keep out rain.

Make a similar shelter by lashing a broken-off bough to the base of another branch where it forks from the trunk (a).

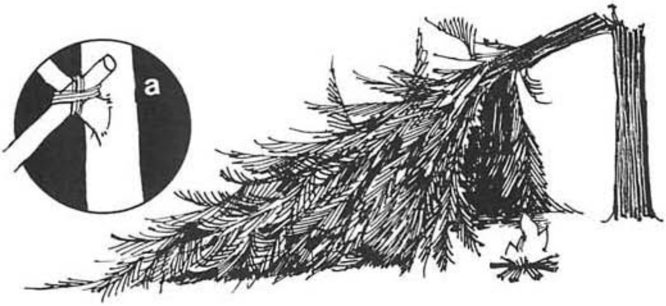

Root shelter

The spreading roots and trapped earth at the base of a fallen tree make a good wind and storm barrier, if they are at the right angle to the wind. Filling in the sides between the extended roots will usually make the shelter much more effective, and provide a good support for building a more elaborate shelter from other materials.

Use a natural hollow

Even a shallow depression in the ground will provide some protection from wind and can reduce the effort in constructing a shelter. However, take measures to deflect the downhill flow of water around it, especially if it is a hollow on a slope, or you could find yourself lying in a pool.

Make a roof to keep the rain off and the warmth in. A few strong branches placed across the hollow can support a light log laid over them, against which shorter boughs and sticks can be stacked to give pitch to the roof and so allow water to run off. Consolidate with turf or with twigs and leaves.

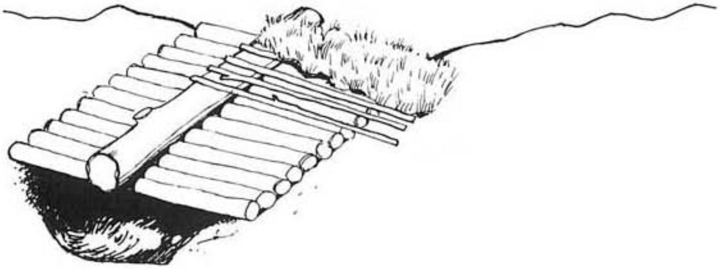

Fallen trunks

A log or fallen tree trunk makes a useful windbreak on its own, if it is at the right angle to the wind. With a small trunk, scoop out a hollow in the ground on the leeward side.

A log also makes an excellent support for a lean-to roof of boughs.

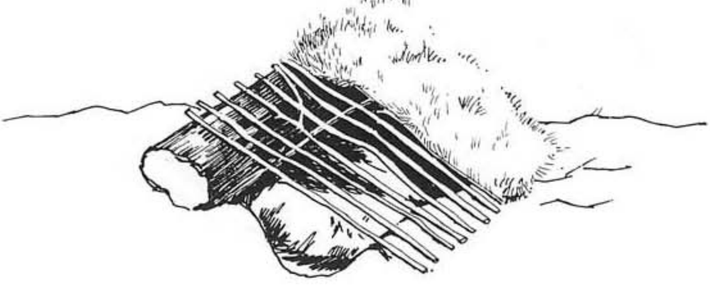

Drainage and ventilation

A run-off channel gouged from the earth around any shelter in which you are below, or lying directly on, ground level will help to keep the shelter dry. Hasty shelters will usually have many spaces where air can enter. Do not try to seal them all – ventilation is essential.

Stone barriers

A shelter is more comfortable if you can sit rather than lie in it, so increase its height by building a low wall of stones around your chosen hollow or shallow excavation. Caulk between the stones (especially the lowest layer) with turf and foliage mixed with mud, and deflect the flow of rainwater around the shelter as shown below.

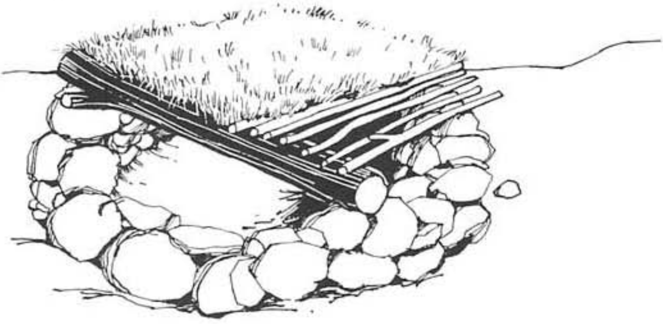

Sapling shelter

If suitable sapling growth is available, select two lines of saplings, clear the ground between them of any obstructions and lash their tops together to form a support frame for sheeting. Weigh down the bottom edges of the sheeting with rocks or timber.

You can make a similar shelter from pliable branches driven into the ground.

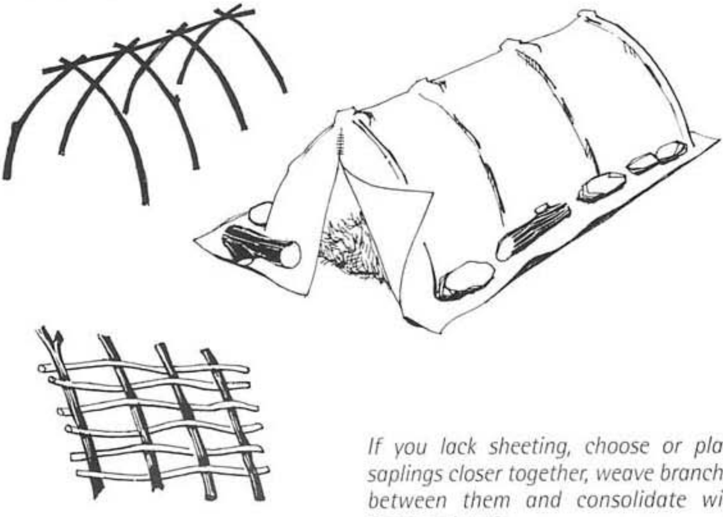

If you lack sheeting, choose or place saplings closer together, weave branches between them and consolidate with ferns and turf.

Shelter sheet

With a waterproof poncho, groundsheet or a piece of plastic sheeting or canvas, you can quickly and easily make a number of different shelters which will suffice until you can build something more efficient. Make use of natural shelter (a) or make a triangular shelter with the apex pointing into the wind (b). Stake or weigh down edges. If it is long enough, curl the sheeting below you – running downhill so that it keeps out surface water (c). If you don't have much material attach it so that you are shielded from the wind (c(i)). Use dry grass or bracken as bedding. Do not lie on cold or damp ground.

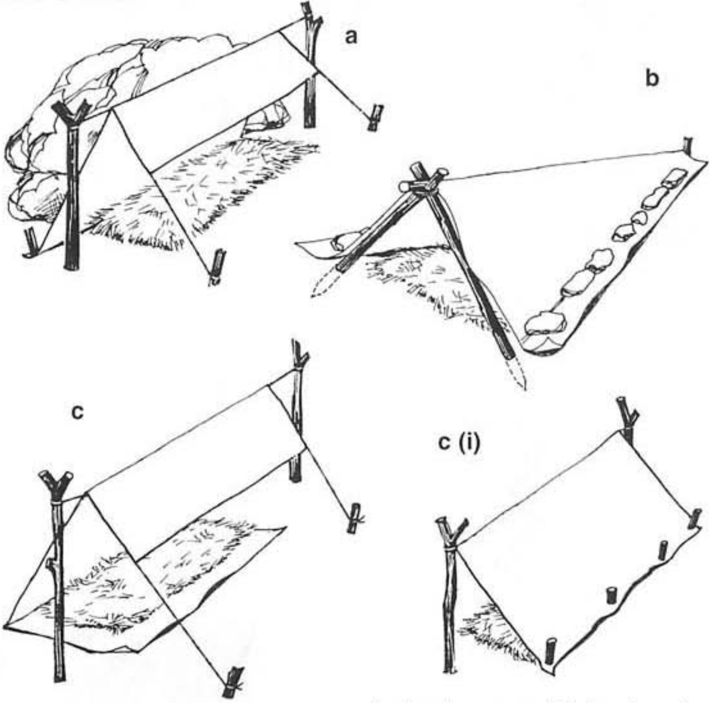

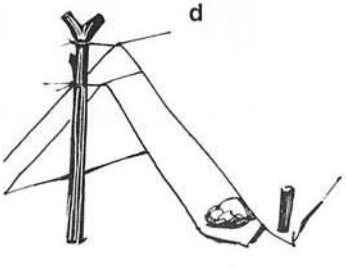

A closely woven fabric, though not impermeable, will keep out most rain if you set it at a steep angle. Fit one shelter a few inches within another (d). The rain that does come through will rarely work its way through both layers.

With any woven fabric, avoid touching the inner surface during rain or you will draw water through.

Tepees

Best known from its North American forms, the tepee occurs in many cultures. The quickest type to erect has three or more angled support poles, tied where they cross to make a cone. They can be tied on the ground and lifted into place before covering with hides, birch bark panels or sheeting. Leave an opening at the top for ventilation.

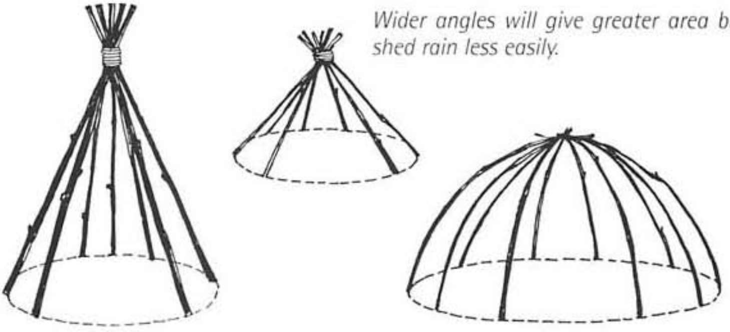

Wider angles will give greater area but shed rain less easily.

Parachute tepee

A parachute, suspended by its centre, makes an instant tepee. Peg out the bottom edge.

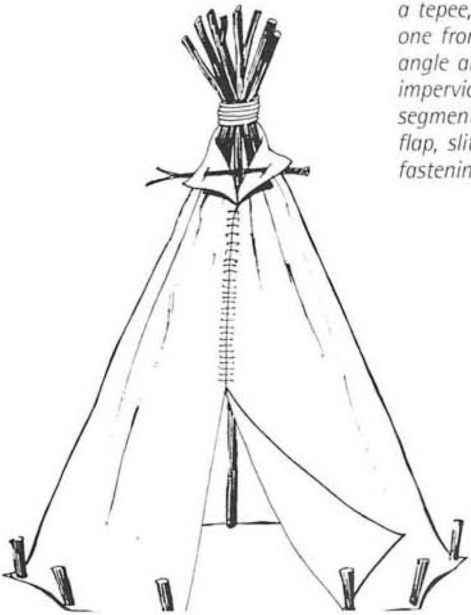

Parachute material can be used to cover a tepee, but even simpler is to suspend one from a tree. Give the sides a steep angle and, even when the fabric is not impervious, water will run off. Fold a segment of the chute double for a door flap, slit along a seam and make a tie fastening to close it.

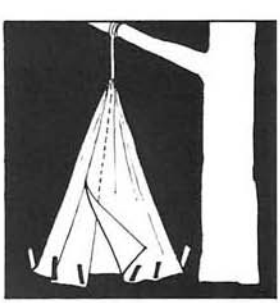

Stick walls and screens

It is easy to build walls by piling sticks between uprights driven into the ground and (if possible) tied at the top. Caulk them well to keep out wind and rain. These are ideal for making one side of a shelter, for blocking a shelter's opening or for a heat reflector behind a fire. If large rocks are not available use this method to dam a stream.

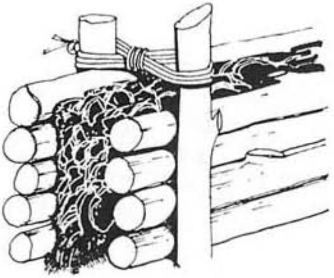

To make a very sturdy stick wall, increase the space between the uprights, use two stacks of sticks and, as you build it, fill the space between with earth.

Coverings

Make wattle and woven coverings for roofs or walls from springy saplings, plant stems, grasses and long leaves (either whole or, if large enough, shredded for tighter weaving). First make a framework from less pliable materials, either *in situ* or as a separate panel to attach later. Tie the main struts in position. Weave in the more pliant materials.

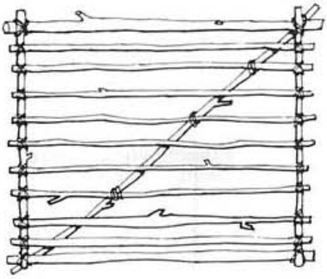

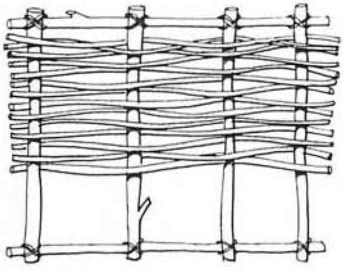

If no ties are available, drive vertical stakes into the ground and weave saplings between them. Caulk with earth and grasses.

If suitable firm cross-pieces are too few, weave creepers between the uprights.

Very large leaves, lashed or weighted down, or hooked over lines of creeper, can be overlapped like tiles or shingles to keep out rain.

Long grass can be bunched and woven (overlap the ends irregularly) to make a continuous warp and weft. Or use birch bark to make shingles. Ring a birch tree with even 60cm (2ft) cuts and carefully remove the bark (a). Across a frame fix pairs of canes or creepers in close-spaced pairs (b). Upper ends of shingles are gripped between the canes, the lower ends rest on top of those below (c).

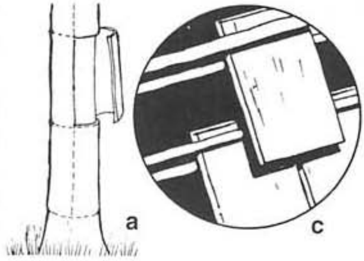

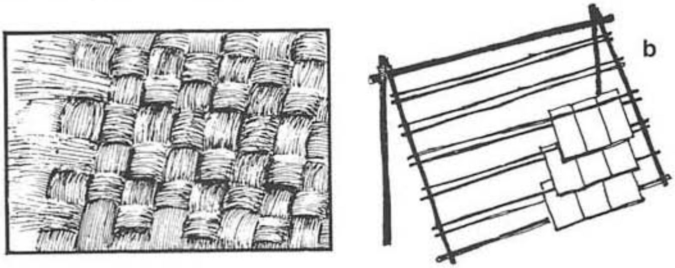

Open lean-to shelter

If there is nothing solid to lean a roof against and you are not trying to keep out heavy rain or a blizzard, use panels of wattle or frames covered in grass for protection.

Erect a horizontal cross-piece between trees or on simple supports. On the windward side lean a panel of wattle, or tie or lean saplings at 45 degrees to make a roof (a). Site your fire on the leeward side. Add side pieces and – this is the trick – build a reflector (b) on the other side of the fire to make sure that you get the full benefit of the warmth.

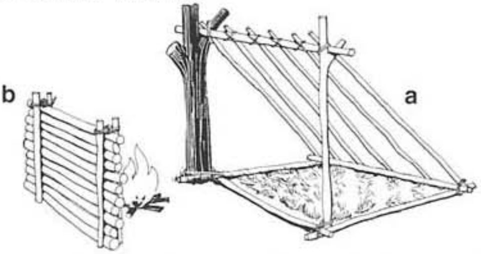

The lean-to on the opposite page shows the lean-to shelter taken to its logical conclusion. By adding to the left and to the right of the structure you can make the shelter larger and you can store more materials. To improve your shelter further you can add a canopy to the front of it in order to provide extra protection against the elements. This more advanced lean-to has space for fire wood (c), supplies (d), and includes a canopy (e).

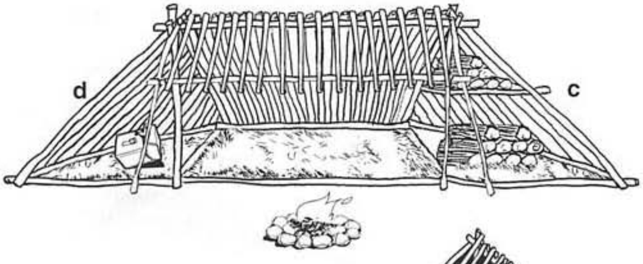

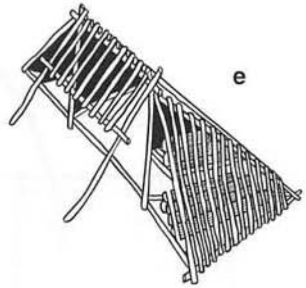

Adding a canopy to the lean-to shelter
This requires two additional supports (one each side of the cross-piece placed at a 45 degree angle), and vertical supports on which to attach the covering material.

TROPICAL SHELTER

In rain forests and tropical jungle the ground is damp and likely to be crawling with insect life, leeches and other undesirables. Instead of bedding down on the ground you will be better in a raised bed. Consequently you may want to make higher shelters.

Unless you are at an altitude high enough to make the nights cold, you will be less concerned with protection from the wind than with keeping reasonably dry. A thatching of palm, banana and other large leaves makes the best roofs and walls.

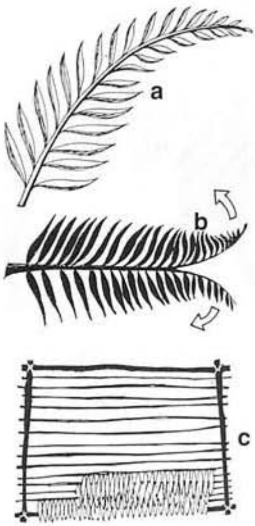

ATAP
Also know as 'Wait-a-while vine', atap is especially useful, despite the barbs at each leaf tip which make careful handling necessary. Look for any plant with a similar structure (a), the bigger the better. The broader the individual leaflets, the better also.

Atap is best used horizontally, splitting each leaf into two from the tip (b) then tearing it into two clean halves down its length. Do not try to split from the thick end or you will end up with a broken branch.

Closely layer halves of atap on your roof frame (c). You can let it be a little less dense on walls. Woven atap can be particularly effective for the sides of a shelter.

Another method:

Do not split down the leaf but fold the leaflets on one side across to the other and interweave them (d). You will probably find this easiest if you work first from one side then the other – but it does take practice.

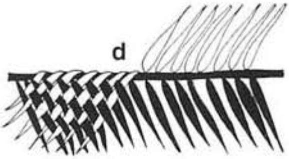

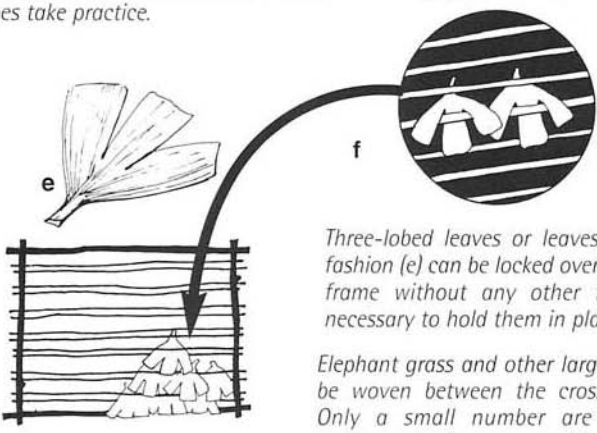

Three-lobed leaves or leaves cut in this fashion (e) can be locked over a thatching frame without any other fixing being necessary to hold them in place (f).

Elephant grass and other large leaves can be woven between the cross-pieces (g). Only a small number are needed to produce a shelter very quickly.

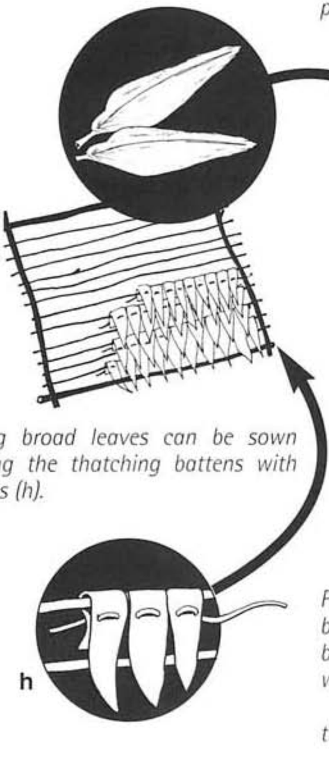

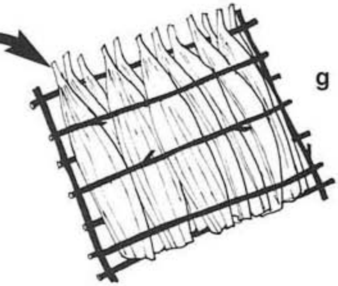

Long broad leaves can be sown along the thatching battens with vines (h).

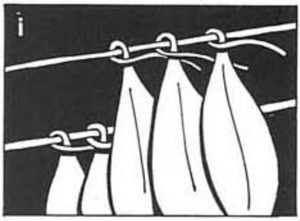

Palm and other long-stemmed leaves can be secured by carrying the stem around the batten and over the front of the leaf, where it is held in place by the next leaf (i).

Leaves must overlap those below on the outside of the shelter.

Bamboo

This large-stemmed plant, actually a grass, is a very versatile building material and can be used for pole supports, flooring, roofing and walls.

The giant form of bamboo – which can be over 30m (100ft) high and 30cm (1ft) in diameter – is an Asian plant, found in damp places from India through to China both in the lowlands and on mountain slopes, but there are types native to Africa and Australia and two which are found in the southern United States.

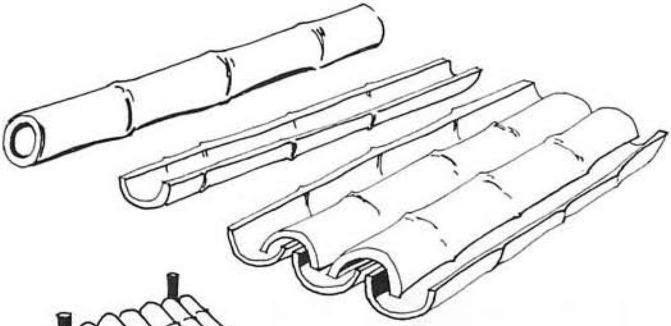

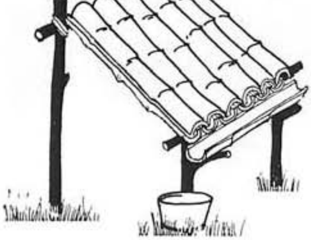

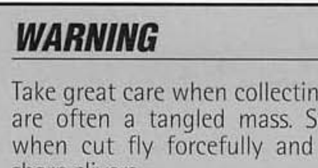

Split bamboo vertically to make roofing and guttering to collect rainwater. The split stems, laid alternately to interlock with one another, form efficient and waterproof pantiles.

Flatten split bamboo for smooth walls, floors or shelving by cutting vertically through the joints every 1.25cm (½in) or so around the circumference. It can then be smoothed out.

The paper-like sheaths formed at the nodes can also be used as roofing material.

WARNING

Take great care when collecting bamboo. It grows in clumps which are often a tangled mass. Some stems are under tension and when cut fly forcefully and dangerously apart, exploding into sharp slivers.

Split bamboo can be razor-sharp and cause serious injury. The husks at the base of bamboo stems carry small stinging hairs which cause severe skin irritations.

ARCTIC SHELTER

In polar areas simple shelters will be those already waiting for you in natural caves and hollows. If you carry some kind of bivouac in your equipment, you can erect it and increase its protection by piling up loose snow around and over it, so long as it can support the weight. But to build in hard snow – and at very low temperatures snow will be solid – you need some kind of implement to cut into it or make blocks from it. Spades and ice saws are essential equipment for polar expeditions.

Snow or rock caves will be easily recognizable – but not so obvious are the spaces left beneath the spreading boughs of conifers in the northern forests when the snow has already built up around them. A medium-sized tree may have a space right around the trunk (a) or a large one have pockets in the snow beneath a branch (b). Try digging under any tree with spreading branches on the lee side.

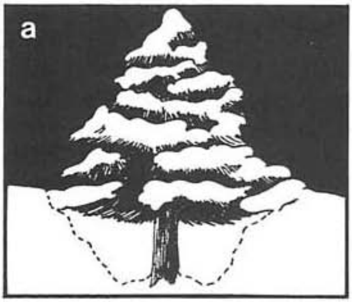

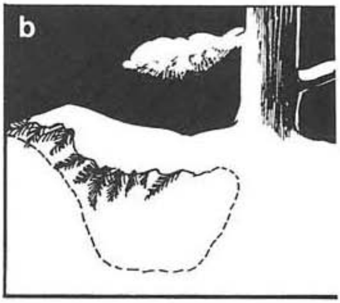

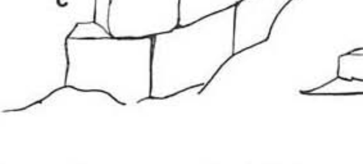

Even soft snow can be built into a windbreak. Those with equipment can cut blocks (c). This is the minimum shelter for the minimum effort.

Anchor a ground sheet or poncho along the top with another course of blocks, use others to secure the bottom edge. Use more snow blocks to close the sides.

BUILDING IN SNOW

A saw, knife, shovel or machete is necessary to cut compacted snow into blocks. The ideal snow will bear a man's weight without much impression being made but be soft enough to allow a probe to be inserted evenly through it.

Cut blocks about 45 x 50cm (18 x 20in) and 10–20cm (4–8in) thick. These will be an easy size to handle, thick enough to provide good insulation and yet allow maximum penetration of the sun's rays.

Snow trench

This is a much quicker shelter to construct than trying to build in snow above the ground, but it is suitable for only one person and then only for short-term use – while you are on the move or making something bigger, for example.

Mark out an area the size of a sleeping bag (including head support) and cut out blocks the whole width of the trench. Dig down to a depth of at least 60cm (2ft). Along the top of the sides of the trench cut a ledge about 15cm (6in) wide and the same deep.

Rest the snow bricks on each side of the ledge and lean them in against each other to form a roof (a).

Put equipment below your sleeping bag so that you are not in direct contact with the snow beneath.

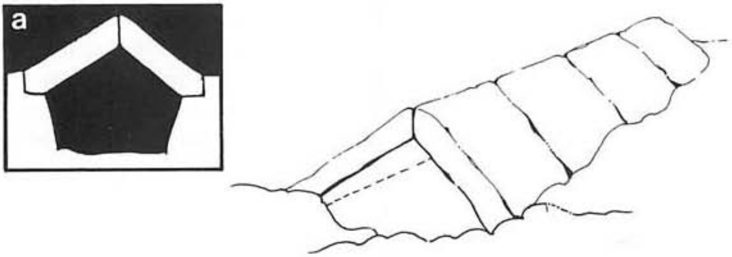

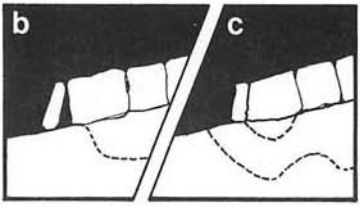

Block the windward end with another block or piled up snow. At the other end (downwind) have a removable block as a door (b), or dig an entrance (c). Fill any gaps with snow. Most effective built on a slight slope, cold air will collect in the entrance leaving warmer air in the sleeping space.

Snow cave

Dig into a drift of firm snow to make a comfortable shelter. Make use of the fact that hot air rises and heavier, cold air sinks. Create three levels inside: build a fire on the highest, sleep on the centre one and keep off the lower level which will trap the cold. Drive a hole through the roof to let out smoke and make another hole to ensure that you have adequate ventilation.

Use a block of snow as a door and keep it loose fitting and on the INSIDE so that it will not freeze up and jam. If it does, a block on the inside will be much easier to free.

Smooth the inside surfaces to discourage melt drips and make a channel around the internal perimeter to keep them away from you and your equipment.

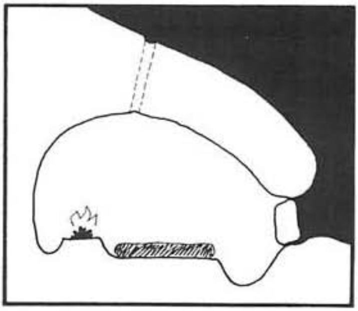

Snow house (igloo)

An igloo takes time to construct but centuries of use by the Eskimo demonstrate its efficiency. Build the main shelter first then dig out an entrance or build an entry tunnel which is big enough to crawl along. Make sure that its entrance does not point into the wind. You could bend the tunnel or build a wind break to make this less likely.

To construct an igloo by the circular method, mark out a circle on the ground about 4m (13½ft) in diameter and tramp it down to consolidate the floor as you proceed with the rest of the building.

Cut and lay a circle of blocks on the perimeter. Be prepared to dig a tunnel (a), or leave a space for an entrance (b). Place another layer on top of them but, as when laying bricks, centre new blocks over the previous vertical joint.

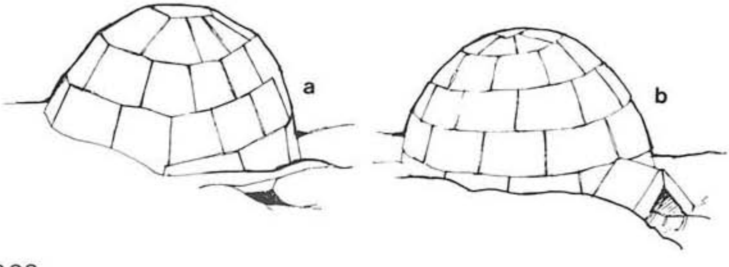

Build up more layers but place each only halfway over the lower tier, so that the igloo tapers in or becomes dome-shaped. Shape out the entrance arch as you proceed. Seal the top with a flat block. Make ventilation holes near the top and near the bottom – not on the side of the prevailing wind or so low that snow rapidly builds up and blocks it. Fill any other gaps with snow. Smooth off all the inside to remove any drip-points. This will allow any condensation to run down the wall instead of dripping off.

Igloo (spiral method)

Lay the first course of blocks and then shape them to the required spiral. You do not have to overhang the blocks if you angle your initial spiral downwards and inwards, and shape the top and bottom faces of subsequent courses to lean inwards. The last few blocks in the centre may need some support as you fit them into position.

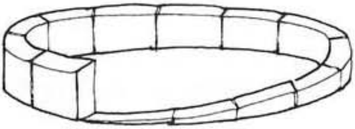

Cutting the first course to an even spiral eases the whole process. Angle the top edge slightly down towards the centre.

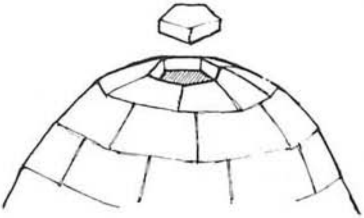

The final block must be cut to fit – unless the space is small enough to leave for ventilation, but this last block helps to keep the structure from collapsing.

SURVIVING THE COLD

No matter how low the external temperature, inside a well-constructed snow house the temperature will not drop lower than -10°C (0°F). Just burning a candle will raise the temperature by about four degrees. The traditional Eskimo way of heating the igloo was a wick in a bowl of fat. In a large shelter with a wood fire the temperature is cosy. An oil burner or fat on bones are good alternatives where there is no wood.

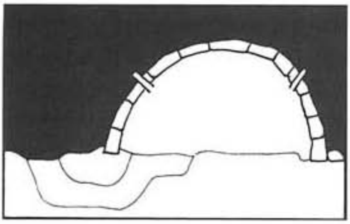

INSIDE THE IGLOO
Build a sleeping level higher than the floor (or dig down when building) to create a lower cold level which can be used for storage.

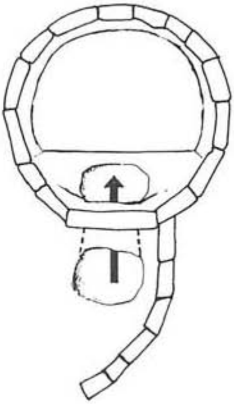

Cut an entrance way through the lower course of blocks or dig a tunnel beneath them. The central hole can be used as an entrance if you are too exhausted to complete the structure.

BUILDING A SHELTER

The type of shelter you build will depend upon: the materials available, the tools available, what you are sheltering from (WIND, COLD, SNOW, RAIN, INSECTS, and so forth).

How long do you intend to remain at the location? Snow caves and natural holes are ideal if you are on the move and do not need a permanent structure.

Size will depend upon the number in the party.

Take your time over building a complex structure and rest frequently. Over-exertion which produces sweating should be avoided.

All shelters MUST be adequately ventilated to prevent carbon monoxide poisoning and allow moisture to escape. Two holes are needed – have one near the top and one near the entrance. In snow shelters the holes must be regularly checked to ensure that they have not become blocked by snow or ice.

Regularly clear accumulated snow from any entrance tunnel to ensure that it does not become blocked.

The smaller the shelter the warmer it will be inside but, since it will not be possible to heat the shelter to many degrees above freezing, you will need a little time to adapt to the environment.

Parachute snow house

This is a useful structure if stranded on sea ice where sufficient snow for an igloo (or igloos) for a larger party may be hard to find. Look for snow or convenient blocks of ice in the pushed-up pressure ridges of the ice.

Mark out a circle and build up a circular wall of snow blocks about 1m (3ft) high. Leave an entrance space if on ice – you will not be able to dig an entrance tunnel. Dig a lower area in the floor for cold air to sink into.

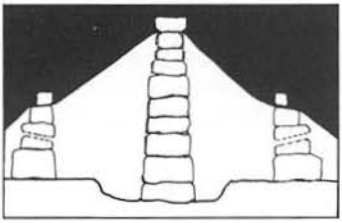

Raise a central column of blocks in the centre about 1-1.5m (3-5ft) higher than the wall. Drape the parachute over this and the wall, securing it with a further row of blocks on top of the wall.

WARNING

The structure of this parachute roof makes it a snow trap, which could become a dangerous weight poised above your head. Clear accumulated snow regularly.

If you want a small fire inside, ensure there is adequate ventilation. Site the fire on the outer shelf where it will not affect the canopy, not near the central column.

Anchor parachute cords with further blocks of ice or snow or cut a hole in the ice and pass the rope through it to make a firm anchorage.

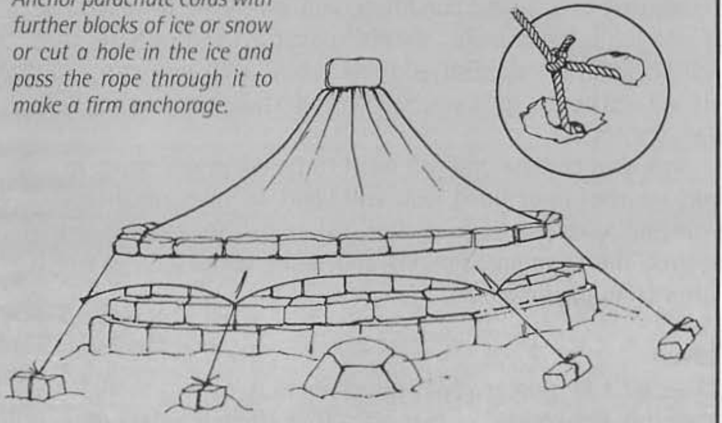

LIVING IN A SNOW HOUSE

In bad weather make sure that you have a good supply of timber, or liquid fuel, inside the shelter.

Do not carry loose snow into the shelter, knock it off boots and clothing before you enter.

Mark the entrance clearly so that it is easily found.

Keep shovels and tools inside the shelter – you may have to dig yourself out.

Drips in igloos can be stopped by placing a piece of snow on the source.

Relieve yourself inside the shelter – this is usual practice in these conditions and conserves body heat. Use plastic bags, ration tins or other containers and empty when possible. Try to discipline the bowels to work just prior to leaving the shelter in the morning and then remove faecal matter with other rubbish accumulated.

In a shelter with several people organize a rosta of duties. It is important that someone tends the fire at all times. Others can check the vent holes, gather fuel, go hunting when possible, prepare meals and so forth.

Remember that at a low temperature you will need more food.

LONG-TERM SHELTER

If you decide that any possible search for you has been abandoned and that it is impracticable to make your own way to safety, whether due to distance, time of year, lack of equipment or physical condition, you will want to make the most comfortable permanent shelter possible. Somewhere that you will be able to establish yourself cosily until you can eventually attract rescue or equip yourself to undertake the journey with your own resources.

In a cold climate you will want to be warm and snug. In a warm one, on the other hand, you will want to take advantage of any available breezes. Your shelter will need to provide protection against the changing seasons and night temperatures which may differ from daytime ones.

Caves

Caves are the most ready-made of shelters. Even a shallow cave (in stone usually known, in fact, as a rock shelter) offers an excellent temporary shelter and a larger cave can make an excellent permanent home. People still live in them in many parts of the world, sometimes with all modern conveniences! Caves situated above a valley will be dry even if water seeps through in some places from above. They are weather-proof and require little constructional work, usually simply the creation of a barrier to close off the entrance. Make this of rocks, wattle, logs, turves or almost any material.

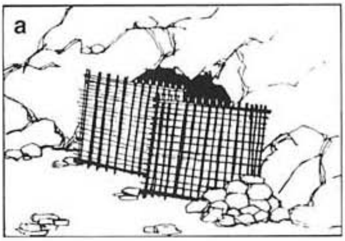

If the cave faces into the wind, build a screen out from both sides, one slightly behind the other, overlapping them to provide an entrance (a).

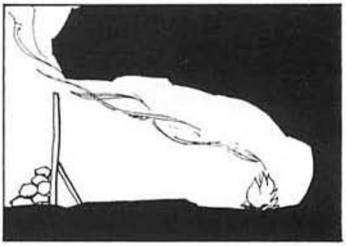

Build the fire at the back of the cave. Smoke will go up to the roof, leaving air nearer the floor. Smoke from a fire near the open mouth of a cave will not escape outwards but will probably be blown in. If you seal the cave entrance make sure you leave a gap for the smoke to escape.

Caves can be cold, and they may already be inhabited by wild animals, so approach any such shelter with caution. Plenty of dry plant matter and pine boughs on the ground will provide insulation. A good fire will usually make animal occupants leave. Allow them an escape route.

Sometimes a cave will have its own fresh water supply, especially if it goes deep into a hillside, either from an underground stream or from water seeping through the rocks above.

WARNING

Check for the possibility of a rock fall inside or outside the cave. You might be desperate for shelter, but your situation will be a lot worse if you are trapped or injured by falling rocks.

Light structures

Follow the methods outlined for the lean-to structure. You can extend it with a less angled roof and a front wall or you can build vertical walls and roof them over with deep eaves to give you extra shade from sun and to ensure that rain runs off well away from the hut. Dig a channel to carry any water away.

If you have bamboo or other strong material available to build a firm frame, raise the floor of your shelter off the ground in tropical climates, so reducing access to ground creatures.

In hot climates you will need to make your roof solid to keep out rain and give good protection from the sun but, if it projects well over the walls, you can leave them as fairly open lattice to allow air to pass through. Grasses and mud will seal cracks and all kinds of material will make a thatch if woven between roof cross-pieces of sticks or cords. In climates with heavy rainfall use leaves or bark like tiles on top.

Log cabin

Scale your cabin to the number it is to house. You can always enlarge it or add on extra rooms later. The size of logs available will determine the length of walls. A square or rectangle will be sturdy and easiest to roof – 2.5m (8ft) square is a sensible small size. You may be lucky and be able to use trees that have already fallen, in which case you could perhaps lash a framework of logs together and fill the spaces in between, but it is much better to joint the corners to fit snugly into each other. Do not finish off the projecting ends. These are the strength of the structure.

Lay down your first layer of logs in the shape of your hut (a). Joint the corners to fit on top of each other (b) and cut other logs to fit neatly on top.

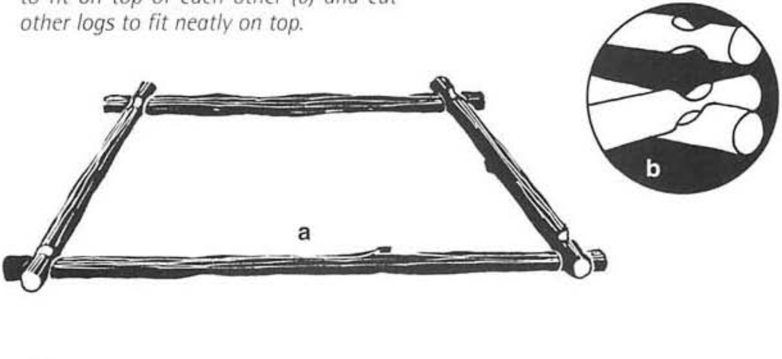

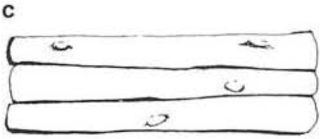

Since logs tend to taper, place them alternately top to bottom to counter this (c).

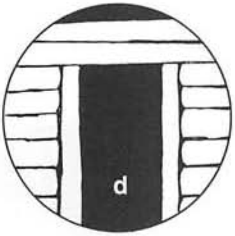

Once the ground frame is established leave space for a doorway on the side away from the prevailing wind – you may be able to use off-cuts from logs for these sections on either side of the door. Square off the edges and wedge a door frame in place (d). Do not bother with windows, the door will give sufficient ventilation.

Build up the front higher than the back to give pitch to the roof. The last log front and back should project well beyond the side walls. These will support the roof. Across the hut from side to side notch in one cross-beam to keep short logs in place. Lay a roof of logs, front to back, extending beyond walls. Notch the logs to fit on to the cross-pieces or lash them down.

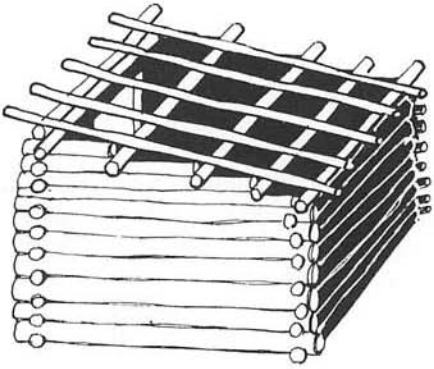

Choose a flat site for your cabin, or level a slightly larger area. Dig into a hillside if necessary, but the foundations for the walls must be level.

The flexible saw in your survival kit will cut logs of sufficient size, and if you are a survivor from a wrecked plane or boat there will have been a fire axe on board.

There is no need to make a door yet. Hang a piece of blanket to keep the wind out, or make a panel of wattle to fill the gap until you feel equipped to make a permanent door. Don't bother with windows either – the doorway will give enough ventilation.

Caulk between the logs with mud and wood chips, or, if there are big gaps, saplings before applying the mud. Mix it with grass and moss, and use a sharpened stick to force it between the logs. Cover the roof with saplings before adding a layer of mud and turf.

Instead of a complete roof of whole logs you could use lighter materials and mud on a timber frame. Bark from the logs makes an excellent top covering if laid as tiles. These could be pegged through with small supple twigs while the mud is still soft.

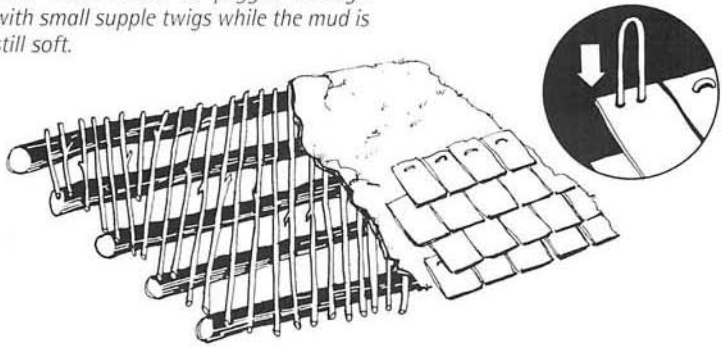

If there is no risk of a build up of water (in which case it would be useful to build a floor later) dig down inside the hut to provide the earth for mud caulking and you will at the same time increase its interior height.

If you leave a hole somewhere in the roof for smoke to escape, you can make a fire inside the hut. But do not leave it unattended – put it out rather than risk your home burning down.

If stone is readily available, you could build a proper chimney and fireplace. You'll retain more of the heat if it is a central structure. Fit stones as closely as possible and use small stones and mud to pack the spaces.

SURVIVAL SCENARIO

You have made your shelter, but where should you wash yourself and your clothes?

All ablutions should be carried out down stream. This will be at a lower point than where you are getting your drinking water from, therefore ensures your supply will remain clean/uncontaminated.

FIRE

Fire can be the difference between life and death. It is imperative that you are able to light a fire under any conditions, anywhere in the world. The most important uses of fire are for boiling water, signalling, cooking, and protection from animals and flying insects.

Because 90 per cent of the diseases that are carried in water can be countered by boiling water it is vital that you learn the skill of fire-lighting. However, you don't need to boil rainwater, as it won't carry any of the bacteria found in the other water sources that you will encounter.

Fire can help to make or fashion tools (you can fire-harden bamboo to make a spoon), while charcoal can be used to burn out a hole. This can be safer than cutting the hole yourself which could lead to you injuring yourself.

When lighting a fire, always ensure adequate ventilation, with enough fuel and a hot enough source to ignite this fuel. To produce flame, this temperature must be maintained to keep air and fuel continuously reacting. The more oxygen introduced, the brighter the fire: by using the wind, or forcing a draught, the fire is fanned to a high temperature and rapidly burns fuel. By reducing the ventilation the fire burns less fiercely and embers are allowed to glow, needing less fuel.

If these principles are understood, smoky fires can be avoided. Smoke is the result of incomplete combustion – with care smoke can be virtually eliminated.

PRACTICE FIRE-LIGHTING

Fire is essential to survival. It provides warmth, protection, a means of signalling, boils water, and cooks and preserves food. You must learn to light a fire anywhere under any condition. It is not enough to know all the methods – you have to be expert at them.

Preparation

First make sure that you have sufficient quantities of TINDER, KINDLING and FUEL. Then prepare a fireplace so that you can control the fire. Used carelessly fire can get out of hand and bring disaster.

The fireplace

The fireplace needs to be prepared carefully. Choose a site that is sheltered, especially during high winds. Except for signal purposes (see *Rescue*), or exceptionally to warm a temporary bough or snowhole shelter, do not light a fire at the base of a tree or stump. Clear away leaves, twigs, moss and dry grass from a circle at least 2m (6ft) across and scrape everything away until you have a surface of bare earth.

If the ground is wet or covered with snow, the fire must be built on a platform. Make this from a layer of green logs covered with a layer of earth or a layer of stones.

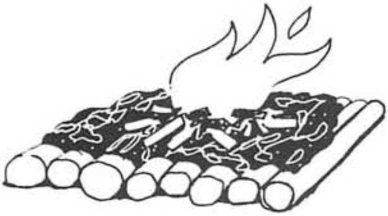

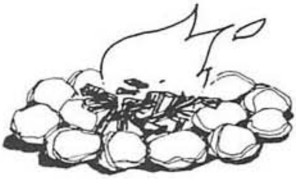

Temple fire

If land is swampy or the snow deep a raised platform is needed. This is known as a temple fire. The hearth consists of a raised platform, built of green timber. Four uprights support cross-pieces in their forks. Across them place a layer of green logs and cover this with several inches of earth. Light the fire on top of this. A pole across upper forks on diagonally opposite uprights can support cooking pots.

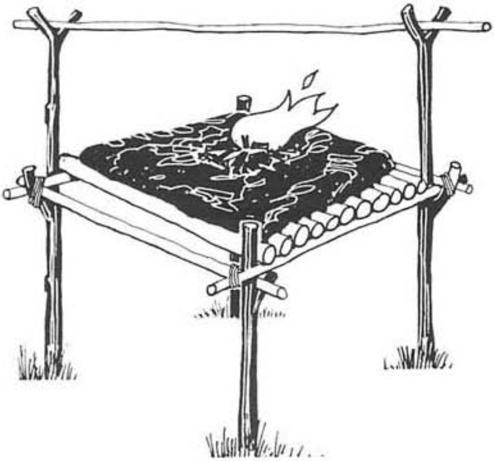

IN WINDY CONDITIONS
If there are particularly strong winds, dig a trench and light your fire in it.

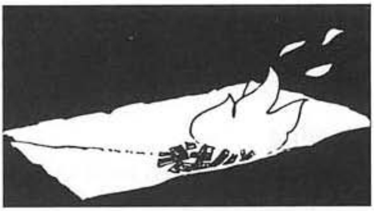

Also good for windy conditions: Encircle your fire with rocks to retain heat and conserve fuel. Use them to support cooking utensils. Their heat, as well as that from the fire, will keep things warm and you can use the rocks themselves as bed warmers.

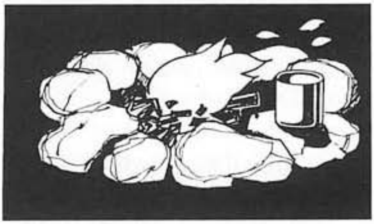

WARNING

Avoid placing wet or porous rocks and stones near fires, especially rocks which have been submerged in water – they may explode when heated. Avoid slates and softer rocks, and test others by banging them together. Do not use any that crack, sound hollow or are flaky. If they contain moisture it will expand faster than the stone and can make it explode, producing dangerous flying fragments which could take out an eye if you are close to the fire.

Tinder

Tinder is any kind of material that takes the minimum of heat to make it catch alight. Good tinder needs only a spark to ignite it.

Birch bark, dried grasses, fine wood shavings, bird down, waxed paper and cotton fluff from clothing all make good tinder. So do pulverized fir cones, pine needles and the inner bark from cedar trees. Dried fungi are excellent, if finely powdered, and scorched or charred cotton or linen, especially ground finely, are also among the best. Where insects such as wood wasps have been boring into trees, the fine dust they produce is good tinder and powdery bird and bat droppings can also be used. The inside of birds' nests are usually lined with down feathers and ignite easily – dry fieldmouse nests are also usable.

Whatever tinder you use MUST BE DRY. It is a good idea to carry tinder with you in a waterproof container. Always keep an eye open for tinder to collect.

Kindling

Kindling is the wood used to raise the flames from the tinder so that larger and less combustible materials can be burned.

The best kindling consists of small, dry twigs and the softer woods are preferable because they flare up quickly.

Those that contain resins burn readily and make firelighting a snip. The drawbacks of soft woods are that they tend to produce sparks and burn very fast. You may need more to get the main fuel going and they are soon consumed if they form the main fuel themselves.

Don't collect kindling straight from the earth, it is almost always damp. Take it from standing deadwood. If the outside is damp, shave until the dry middle is reached.

MAKE FIRE STICKS
Shave sticks with shallow cuts to 'feather' them. Preparing kindling in this way makes it catch light more freely and establishes a fire quickly.

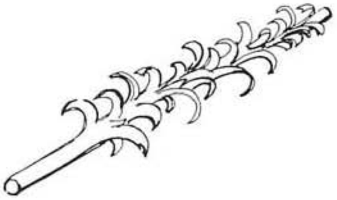

FUEL

Use dry wood from standing trees to get the fire going. Once it is established you can use greener wood or dry out damp wood.

As a general rule, the heavier the wood the more heat it will give – this applies to both dead and green wood. Mixing green and dry wood makes a long-lasting fire, which is especially useful at night.

Hardwoods – hickory, beech or oak, for instance – burn well, give off great heat and last for a long time as hot coals. They keep a fire going through the night.

Softwoods tend to burn too fast and give off sparks. The worst spark-makers are cedar, alder, hemlock, spruce, pine, chestnut and willow.

Remember that damp wood is sometimes advantageous, producing smoke to keep off flies, midges and mosquitoes and burning longer so that it keeps the fire in.

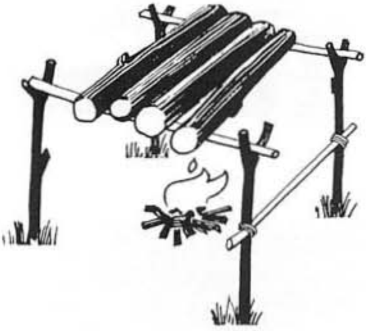

Dry wood across two supports above a fire – not so close that it is set alight. Lay green logs at an angle beside the fire, tapering away from the wind to speed combustion of a sluggish fire while drying them.

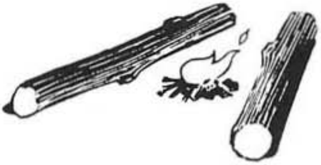

Rest logs against a pot rail to dry. Build a wood shed – essential in wet weather. Set it close to the fire so that the fire's warmth will help dry the wood, but not so close that a spark could ignite it. Build two bays and use wood from one while the other batch dries.

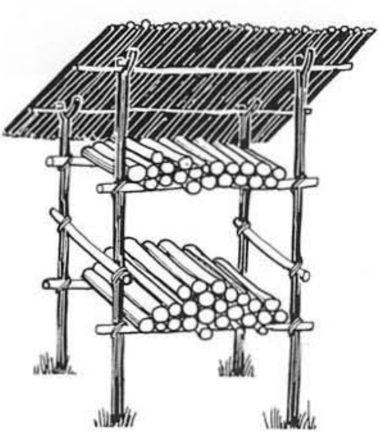

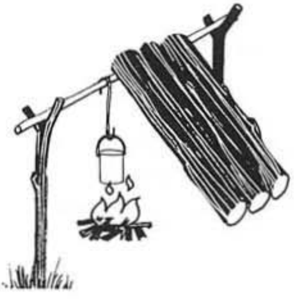

SAVE ENERGY

Don't waste energy chopping logs: Break them by smashing them over a rock (a).

If that does not work, feed them over the fire, letting them burn through in the middle (b) or, if they are not so long, feed them end first into the fire.

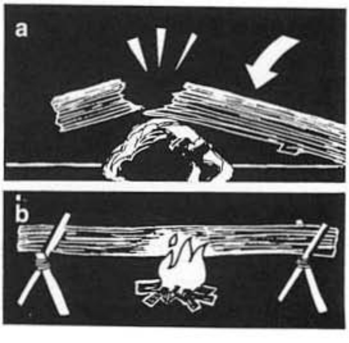

If it is absolutely necessary to split logs in order to conserve fuel, an axe is not needed. Even quite a small knife placed on the end of a log and hit with a rock may split it (c). Once begun, plug a wooden wedge in the opened gap and drive this downward to complete the split. But if you only have one knife don't take the risk of damaging it.

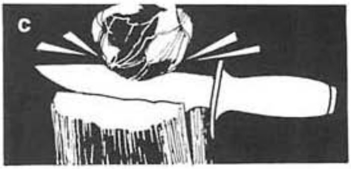

STAR FIRE

The fire is started where the ends of large logs meet, then they are pushed inward as more fuel is needed.

When not required to produce strong heat they can be drawn apart leaving glowing embers for cooking in the centre.

To resurrect the fire push them together and they soon take flame again. This type of fire is used mainly to conserve fuel but also saves chopping wood.

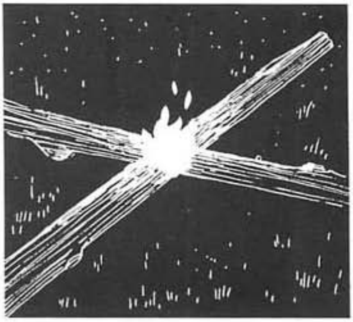

Other fuels

In areas where wood is scarce or unavailable other fuels must be found.

Animal droppings: These make excellent fuel – frontiersmen of the Wild West used 'buffalo chips' for their fires. Dry the droppings thoroughly for a good smokeless fire. You can mix them with grass, moss and leaves.

Peat: This is often found on well-drained moors. It is soft and springy underfoot and may be exposed on the edges of rocky outcrops – looking black and fibrous. It is easily cut with a knife. Peat needs good ventilation when burning.

Stacked with plenty of air around it peat dries rapidly and is soon ready to burn.

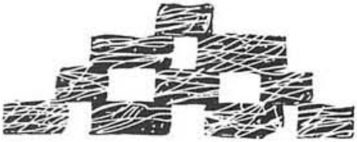

Coal: This is sometimes found on the surface – there are large deposits in the northern tundra.

Shales: These are often rich in oil and burn readily. Some sands also contain oil – they burn with a thick oily smoke which makes a good signal fire and also give off a good heat.

Oils: If you have had a mechanical failure and crashed or broken down with fuels intact you can burn petroleum, anti-freeze, hydraulic fluid and other combustible liquids. Even insect repellent is inflammable. Anti-freeze is an excellent primer for igniting heavier engine oils. With a little potassium permanganate (from your survival kit) you can set it alight in a few seconds.

In very cold areas drain oil from an engine sump before it congeals. If you have no container drain it on to the ground to use later in its solid state.

Tyres, upholstery, rubber seals and much of any wreckage can be burned. Soak less combustible materials in oil before trying to make them burn.

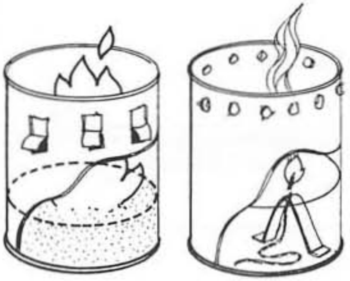

Mix petrol with sand and burn it in a container as a stove, or dig a hole and make a fire pit.

Burn oil by mixing in petrol or anti-freeze. Do not set a light directly to liquid fuels but make a wick and let that provide the flame. The same goes for insect repellent.

Animal fats: These can also be used with a wick in a suitably ventilated tin to make a stove. Bones can add bulk when fat is being burned as a fire (sometimes the only available fuel in polar regions).

Start flame with tinder or a candle then place a network of bones over it to support the fat or blubber. Use only a little fat at first. Unless it is surplus, burning fat means sacrificing food value, but seal blubber spoils rapidly and makes good fuel.

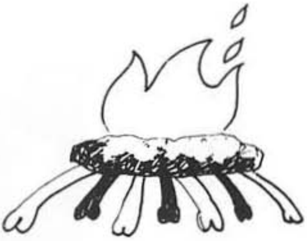

BURNING OIL AND WATER

This mixture makes one of the hottest of all fires. Pierce a small hole in the base of a tin can for each liquid and fit a tapered stick into it to govern the flow (a). The oil and water run down a trough on to a metal plate. Pulling the stick out increases the flow, pushing it in reduces it. Try 2–3 drops water: 1 drop oil.

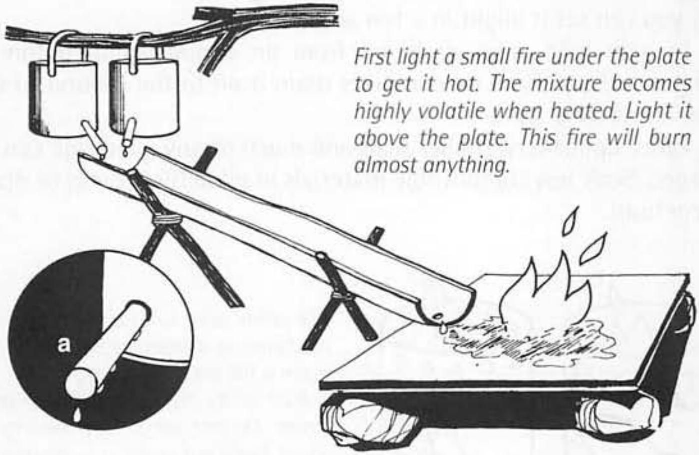

First light a small fire under the plate to get it hot. The mixture becomes highly volatile when heated. Light it above the plate. This fire will burn almost anything.

FIRELIGHTING

Make a bed of tinder and form a wigwam of kindling around it. In a strong wind lean the kindling against a log on the leeside. Ignite the tinder. Once the kindling has caught add larger sticks. Or take a bundle of dry twigs, no thicker than a match, light them first and place them in the wigwam.

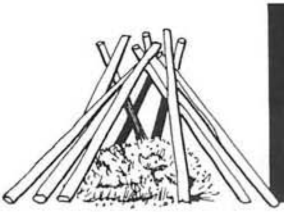

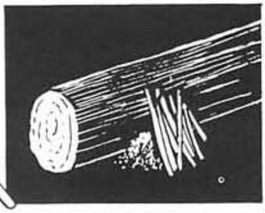

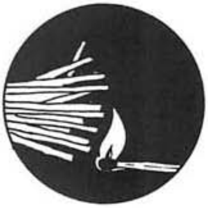

MATCHES

Matches are the easiest way to start a fire. Carry the non-safety 'strike anywhere' type and as many as possible. Pack them in waterproof containers so that they cannot rub or rattle and accidentally ignite. Waterproofing the matches themselves does both jobs.

Some people split all their matches in half and it has been claimed that one can be successfully divided into six. But do NOT risk wasting them – one that works is more use than six that don't!

Strike split matches by pressing the business end against the striking surface with a finger. If this burns the finger be ready to cool it at once – in cold water, snow or even 'spit on it and blow'.

Damp matches

If your hair is dry and not too greasy roll the damp match in it. Electricity should dry out the match.

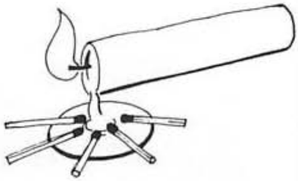

Waterproof matches by dripping candle wax on to them. Rip it off with a fingernail when about to strike one.

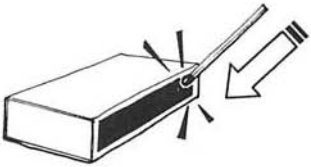

Strike a damp match by stabbing obliquely into the striker strip instead of drawing the match along it.

REMEMBER: Whenever you strike a match light a candle. Many things in turn can then be lit from it – saving matches. Place it in the wigwam of kindling to start a fire and remove it as soon as the flame spreads. Only the smallest amount is burned and even a small candle will last a long time.

However many lighters or fire-makers you carry, still pack as many matches as you can – you can't beat them. So-called everlasting matches can be used over and over again but sooner or later even they pack up. So carry ordinary matches as well. Work out which kind gives you the most strikes for the weight and room they take up.

Using a lens

Strong direct sunlight, focussed through a lens, can produce sufficient heat to ignite your tinder. Accidental fires are caused by the sun shining through broken bottles on to dry leaves or pasture. Your survival kit magnifying glass or a telescope or camera lens will serve instead.

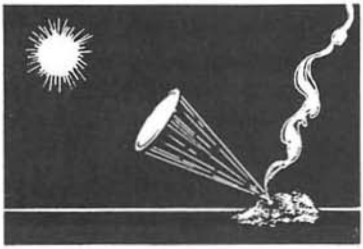

Shield tinder from wind. Focus sun's rays to form the tiniest, brightest spot of light. Keep it steady. Blow on it gently as it begins to glow.

Powder from ammunition

If you are carrying arms you can use the gunpowder propellant from a round to help ignite your tinder.

Break open the round and pour the gunpowder on to your tinder before using your flint (a), or remove only half the powder and stuff a piece of cloth into the cartridge case (b). Chamber the round and fire as usual, into the ground. The cloth will be ejected smouldering. Place it on tinder with the remaining propellant and you will soon have a fire going.

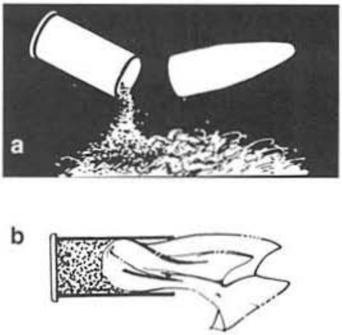

Flint and steel

Flint is a stone found in many parts of the world. If it is struck vigorously with a piece of steel hot sparks fly off which will ignite dry tinder. A saw-edged blade can produce more sparks than an ordinary knife and should be in your kit. A block of magnesium with flint on its side is an even more efficient device – magnesium burns very strongly.

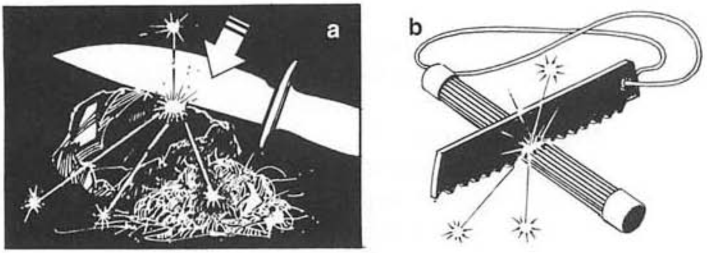

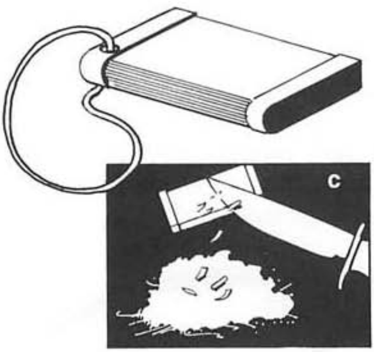

Strike the blade against the flint (a), or draw the saw across the ridged surface of the flint supplied with it (b), close to tinder so that sparks fall on it.

With a magnesium block, scrape slivers of magnesium on to tinder first (c), then use the saw to produce the sparks.

Battery fire-lighting

A spark from a car battery can start your fire, and torch and radio batteries should have sufficient power. You need two lengths of wire, which you simply attach to the terminals. If you cannot find any wire you could do it with a couple of spanners or other metal implements. Unless you have long pieces of wire, take the battery out of the vehicle first.

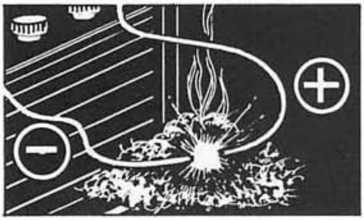

Slowly bring the bare ends of the two pieces of wire together. Just before they touch, a spark will jump across. You must catch it on your tinder. A small piece of cloth with a little petrol on it makes the best tinder, the petrol vapour igniting from the spark.

Fire bow

A simple fire-making technique, but one that needs lots of practise. The friction of a hardwood spindle rotated on a softwood base produces first fine wood-dust tinder, then heat. Balsa, pine and bamboo are typically suitable softwoods; oak, ash and beech are hardwoods. Both must be dry.

Gouge a small depression at the near end of the baseboard and cut a cavity below it in which to place the tinder. Shape the spindle evenly. Make the bow from a pliable shoot such as hazel or bamboo and the string from hide, twine or a bootlace. You also need a hollowed piece of stone or wood, or a small jar to steady the top of the spindle and exert downward pressure.

Wind the bowstring once around the spindle. Place the spindle in the depression, hold the steadying piece over its end and bear lightly down on it while the other hand moves the bow backwards and forwards. This makes the spindle spin. Increase the speed as the spindle starts drilling through the wood. When it begins to enter the cavity apply more pressure and bow vigorously.

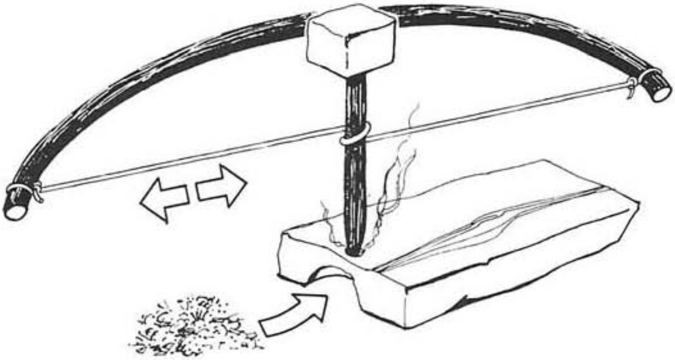

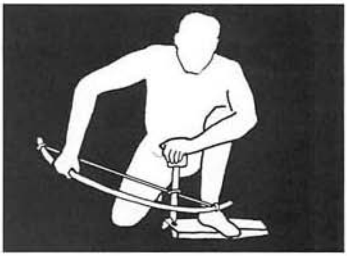

Keep on bowing until unable to continue. If successful the tip, glowing like a cigarette, will drop onto your tinder, which, if you gently blow on it, will burst into flame. You must keep the spindle upright and steady.

It helps to kneel with one foot on the baseboard and to lock the spindle arm on to this leg while bowing with the other hand. Keep the bow strokes very even.

A V-shaped notch, as shown in the baseboard of the hand-drill method, is also recommended.

Hand drill

This variation on the fire bow is useful in dry territories with low humidity and little rainfall – making everything 'tinder' dry.

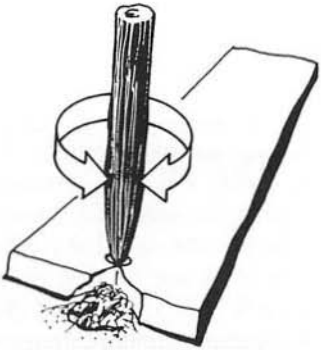

In a baseboard of hardwood cut a V-shaped collecting notch which will hold tinder, but still allow air to reach it. Make a small depression near it. For a spindle use a stem of hollow softer wood with a soft pith core.

Roll the spindle between the palms of the hands, running them down it with each burst of spinning to press the spindle into the depression in the baseboard.

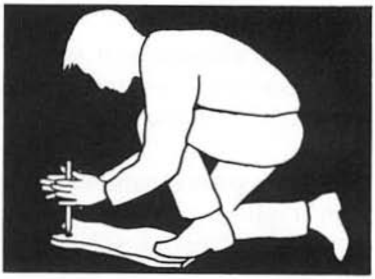

When the friction makes the spindle tip glow red, blow gently to ignite the tinder around it. Putting a pinch of sand in the spindle hole increases the friction and speeds the heating of the tinder.

A cavity below the spindle, as shown for the fire-bow method, is also recommended.

Fire plough

This method of ignition also works by friction. Cut a straight groove in a soft wood baseboard and then 'plough' the tip of a hardwood shaft up and down it. This first produces tinder and then eventually ignites it.

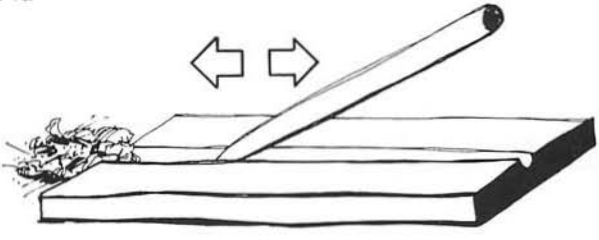

Fire-lighting with chemicals

A survivor's pack is not likely to include a complete chemistry set but there are some very common chemicals that, if they are available, can be used to produce combustion. The following mixtures can all be ignited by grinding them between rocks or putting them under the friction point in any of the types of fire drill already described. Mix them carefully, avoiding contact with any metal objects. All are susceptible to dampness and must be kept dry.

WARNING

Handle these chemicals carefully, sodium chlorate in particular – it ignites from percussion, so avoid shaking it up or letting it spill. Spilled weedkiller on a hard path has been known to ignite when stepped on or a watering can put down on it!

Potassium chlorate and sugar in a mixture of 3:1 by volume is a fierce-burning incendiary which can also be ignited by dripping a few drops of sulphuric acid onto the mixture.

Potassium permanganate and sugar mixed 9:1 is less sensitive and temperature is a critical factor in how long it takes to ignite. The addition of glycerine will also produce ignition.

Sodium chlorate and sugar mixed 3:1.

- Sulphuric acid is found in car batteries
- Potassium chlorate is found in some throat tablets – their contents may be listed on the pack. Try crushing one and seeing if it works.
- Potassium permanganate is included in your survival kit.
- Glycerine is a constituent of anti-freeze.
- Sodium chlorate is a weed-killer.

TYPES OF FIRE

However quickly you want to get a fire going, take time, while you gather fuel and get the tinder ready, to choose the best location and the best type of fire.

Fires for warmth

With a single fire outdoors only surfaces facing it are warmed. With two fires you can sit between them – but that would use a lot of fuel and, no matter which way the wind is blowing, you are bound to be covered in smoke. Build one fire and use a reflector.

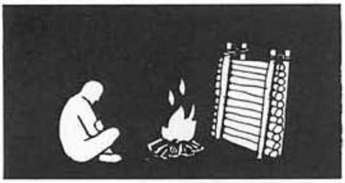

A good reflector, close to the fire, not only reflects heat back to you but also helps to make the smoke go upwards, drawn by hot currents of air, instead of getting in your eyes. Use a reflector to direct heat into a sleeping shelter.

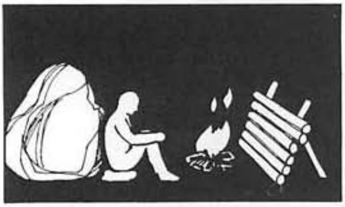

The inexperienced often build a fire up against a tree stump or a rock – don't, build the fire away from it and sit between the two so that the rock reflects the heat and warms your back. Add a reflector.

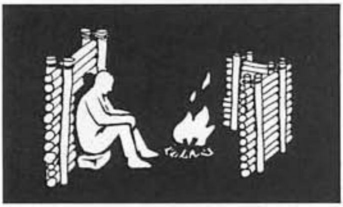

If there is no ready-made reflector, build one – and build another reflector on the other side of the fire to reflect as much as possible of its heat back to you.

Snake hole fire

This is a shielded fire that produces a good draught and burns almost anything once lit. In the side of a firm earth bank excavate a chamber about 45cm (18in) deep. From above drive a stick down into the chamber, manoeuvre it about a little to make a chimney, removing the spoil that falls below. Build the fire in the chamber.

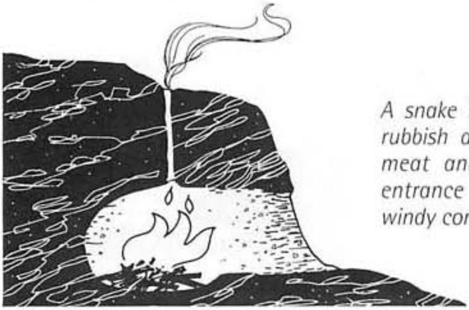

A snake hole fire is good for burning rubbish and the smoke for preserving meat and fish. The snake hole fire entrance is best sited downwind in windy conditions.

Cooking fires

These cooking fires are also good for heating.

Yukon stove

This fire, once lit, will burn almost anything. It takes a lot of effort to build but is worth it for the whole structure gives off good heat and the top can be used for cooking.

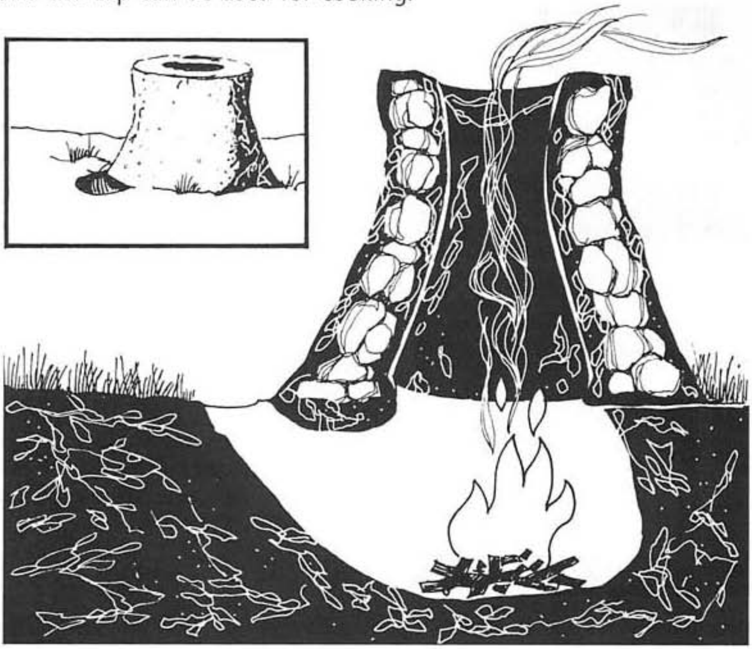

Dig a hole circular in shape and about 24cm (9in) deep with a channel on one side leading down to it. Set rocks up all round the outer edge of the main hole and build up a funnel, bridging over the channel and gradually sloping inwards. Let the upper courses begin to open out again. Seal all the spaces between the rocks with earth. The fire is shielded, the chimney creating a good draught.

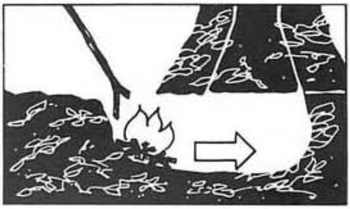

Light the fire first in the channel. When it gets going push it beneath the chimney. Fuel is then fed in through the top of the chimney and the rate of burning is controlled by opening or closing the top.

This fire leaves very little ash and will burn a very long time before it needs clearing out.

Trench fire

This fire is sheltered from strong wind by being below ground level. Dig a trench about 30 x 90cm (12 x 36in) and about 30cm (12in) deep plus the depth of a layer of rocks with which you now line the bottom. Build the fire on top of the rocks. Even when it has died down they will remain hot and make an excellent grill.

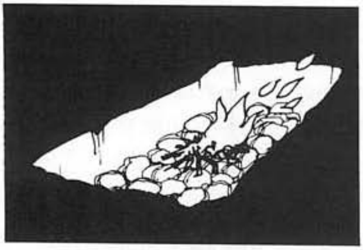

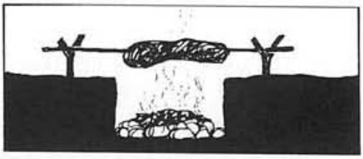

A spit placed across the embers is excellent for roasting.

Hobo stove

This stove provides a heat source several people can huddle around and its top can be used for cooking. To make it you need something like a 5-gallon oil drum.

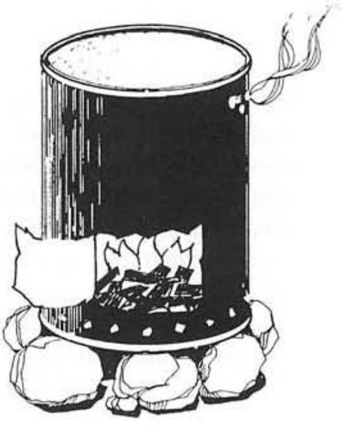

Punch holes in the bottom and around the bottom of the sides of the drum for draught to enter. Cut out a panel on one side, about 5cm (2in) from the bottom through which to stoke the fire.

Punch holes in the top if to be used only for heating, but make them on the upper part of one side if you don't want smoke coming through the top.

Set the whole drum on a ring of stones so that there is plenty of draught beneath.

SPONTANEOUS COMBUSTION

Fire sometimes breaks out spontaneously in a compacted heap of wet hay. It can be produced in cotton soaked in linseed oil provided the atmosphere is warm and dry, but temperature can be critical. Either it will burst into flames within a couple of hours or not at all. Not a reliable way of fire-lighting – but a risk to be aware of.

COOKING

When food is heated it loses nutritional value – the more the heat the greater the loss – so nothing should be cooked longer than is necessary to make it palatable unless it is suspect and being cooked to kill germs and parasites or to neutralize poisons.

Boiling vegetables destroys their vitamin C content and roasting meat removes its all-important fat, but we are used to eating our food cooked and a hot meal is unsurpassed for raising morale. It would take great discipline to eat many things raw that you had not previously considered foods, but a frog, grubs or rats do not seem too bad once cooked.

Cooking not only makes many foods more appetizing to taste, see and smell, it softens the muscle fibres in meat, makes protein more easy to digest and – most important – it destroys bacteria and parasites that may be present.

If the ground is lush, animal foods are more likely to carry parasites. Pigs, especially, carry worms and flukes. Thorough boiling will destroy them, though at the loss of food value. Some foods must never be eaten raw – nettles and several other plants, for instance – but should always be cooked to neutralize harmful substances which they contain.

Your particular situation will determine whether to cook or not. If you cannot face eating something raw, or if food is plentiful but limited in type, cook it to make it more palatable. Relieve boredom by varying cooking instructions.

Cooking methods will depend upon the foodstuff and the facilities you have or can create. Type of fire, utensil support and cooking methods must all be matched.

Cooking requires a slow heat. Use the flame of a fire to boil water then let the fierce flames die down and use embers and hot ash for cooking.

REMEMBER: NEVER leave your fire unattended when cooking – you cannot afford to ruin food.

Once having lit a fire, ALWAYS have something boiling on it – unless water is in short supply – for hot water is always an asset: hot drinks are always cheering and you will find a multitude of uses from sterilizing wounds to making poultry plucking easier.

Do NOT just balance a can on the fire – if it tips over you could put your fire out, quite apart from losing its contents. Support vessels on firm rocks or suspend them over the fire.

Boiling

Cooking in boiling water requires a container. Tin cans and metal boxes are ideal. Make a handle, hang them from a pot support or use pot tongs to take them on and off the fire (see *Useful utensils*). Puncture holes in pots can be repaired by hammering in small plugs of wood – when wet they will expand and stop leaks. If no metal containers are available, a thick length of bamboo holds liquids well. Containers can also be made from birch bark – but be careful that they do not boil dry.

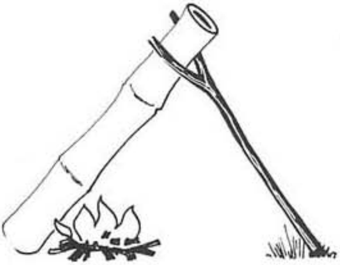

To cook in a bamboo stem, angle it across the heat of the fire, supporting it on a forked stick driven into the ground.

Although boiling does destroy some food elements it conserves the natural juices and retains all the fat – provided that you drink all the liquid as well as eat the remaining solids. Each time you throw away cooking water you lose valuable nutrients, though you will have to discard it if boiling out toxic substances. Boiling will make tough and stringy roots and old game softer and more edible. It will kill worms and flukes and can even make spoiled meat fit to eat.

If you frighten a feeding animal from its kill, you can eat the remaining meat provided that you cut the meat up and boil it for at least 30 minutes. If desperate for food any dead animal that is not actually decomposing can be risked if you use only the large muscle areas. Cut them into 2.5cm (1in) cubes and then boil briskly for at least 30 minutes. Eat only a little, then wait for half an hour to see if there are any ill effects – most toxins affect the digestive system in that time or less. If there are no ill effects tuck in.

Part-boiling vegetables that you intend to cook by other means will speed up cooking times. (For boiling water when no fire proof containers are available see *Hangi method*.)

Roasting

Roasted meat cooks in its own fat. The easiest method is to skewer the meat on a spit and turn it over the hot embers of a fire or beside a blazing fire where it is hot enough to cook. Continually turning the meat keeps the fat moving over the surface. Roasting makes a very tasty dish but has two disadvantages.

Valuable fat is lost unless a drip tray is placed beneath the spit. Regularly baste the meat with fat from the tray.

Roasting by a fierce fire can cook and seal the outside, the inner flesh remaining undercooked, leaving harmful bacteria alive. A slow roast is preferable, and if cooking continues after the outer meat has been cut off the inner flesh can go on cooking.

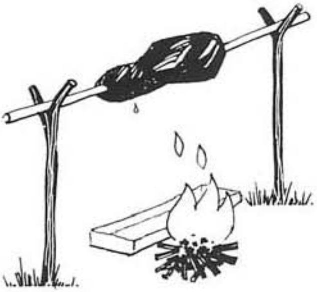

The fire should be slightly to one side of food to allow for a drip tray to catch valuable fat.

Grilling

Grilling is a quick way of cooking large amounts of food but it requires a support – such as a mesh of wire – rested on rocks over the embers of the fire. It should only be used when food is plentiful since it wastes most of the fat from the meat. Hot rocks beside the fire can be used as grilling surfaces or food skewered on sticks and held over the fire.

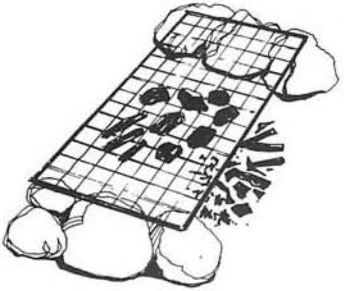

If no wire mesh is available, make a grid of very green sticks or rest a long stick on a forked support so that it can hold food over the fire. Wrap food around the stick. You can also barbecue meat and vegetables on a stick supported across glowing embers by a forked stick on each side.

Baking

You need an oven for baking, but if time and materials are available this is a good way of cooking. Meat should be cooked on a dish and the fat which runs out used to baste it. It is ideal for tough, stringy meat. Cooked for a long time on a steady heat the meat becomes more tender. Baking is also very suitable for root vegetables.

If meat is placed in a tin containing a little water to be cooked in the oven this is a form of braising.

Use an oven to cook several different things at once.

Metal box oven

A large food tin or metal box with a hinged lid makes an excellent improvised oven. Army survivors found an ammunition box ideal. If the lid is hinged and has a catch on it that you can use as a handle, you could set it up to open sideways. It will probably be easier, especially if it has no catch or you have to improvise hinges, to let it open downwards. If you place a rock or other support in front, to rest it on, you will have a convenient shelf. You can always prop it closed if there is no catch, for you do not want a tightly sealed fit – which could build up dangerous pressure inside. If no tin or box is available you could make a clay dome, like an Indian tandoori oven. To make it hot set a fire inside and scrape this out before cooking. Leave a smallish aperture which can be easily sealed while baking.

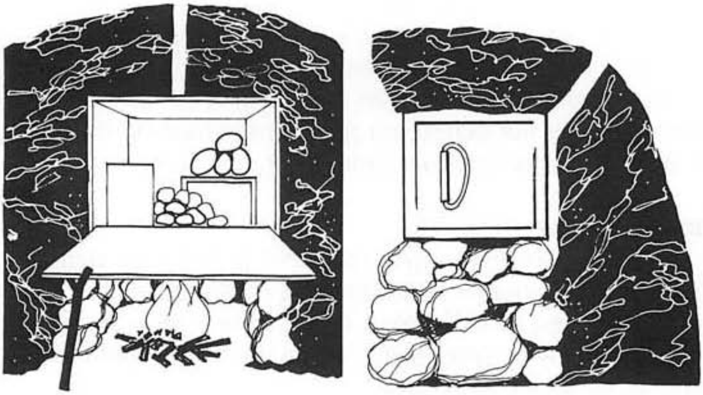

Stand the tin on some rocks so that a fire can be lit beneath it. Build up rocks and earth – or, better, clay – around back and sides and over it, but leaving a space behind for heat and smoke to move around the back. Use a stick to make a chimney hole from above to the space at the back.

Steaming

Steaming does not overcook so preserves nutritional values. It is an excellent way of cooking fish and green vegetables. Fresh young leaves take very little cooking. The foodstuff needs to be suspended in the steam from boiling water.

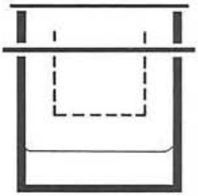

Make a simple steamer by punching holes in a can and suspending it inside a larger can, or putting something in the bottom of the larger can to keep the inner one above the water. Cover the outer can so that steam is not dissipated, but not so tightly that it is sealed or pressure could build up and cause it to explode.

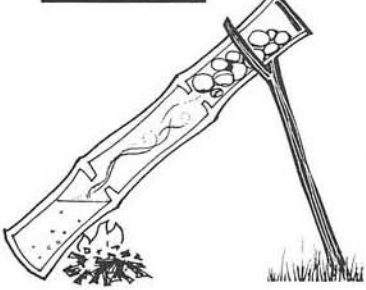

The compartmented sections of bamboo also make an excellent steamer. Make a comparatively small hole between the sections, but big enough to let water through to fill the bottom section. Make a lid (not too tight) for the top. Water boiled in the lowest section will produce steam to cook food in the top one.

Frying

Frying is an excellent way of varying diet, if fat is available and you have a container to fry in. Any sheet of metal that you can fashion into a curve or give a slight lip to will serve. In some areas, you may find a large leaf which contains enough oil not to dry out before the cooking is done – banana leaves are excellent surfaces to fry eggs upon. Try leaves out before you risk valuable food on them and, if you do use one, fry only over embers, not over flames.

Cooking in clay

Wrapping food in clay is a method that requires no utensils and offers a tasty alternative even when you have them. After wrapping in a ball of clay, food is placed in the embers of a fire. The heat radiates through the clay which protects the food so that it does not scorch or burn.

Animals must be cleaned and gutted first but need not be otherwise prepared: when the clay is removed a hedgehog's spines or a fish's scales will remain embedded in it. With small birds, the clay does your plucking for you – but feathers provide insulation and may prevent a big bird being properly cooked. Cooking root vegetables in this way will remove their skins – losing important food value.

Hangi method

This is another way of cooking without utensils. Like the clam bake of the United States and traditional Maori and South Pacific methods it involves heating stones. It requires kindling, logs and round rocks or stones about the size of a fist. Do not use soft, porous or flaky stones which might explode on heating.

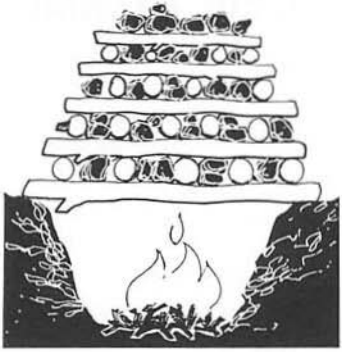

Dig an oval-shaped hole with rounded sides 45–60cm deep (18–24in) and place kindling at the bottom. Lay logs across the hole, place another layer of logs at right angles to them, interspersing them with stones. Make another layer of logs and build up five or six more alternating layers, topping them off with stones.

When the kindling is set alight the logs will burn, heating the stones above them, until, eventually, all falls down into the pit. Remove the burning embers and ash. Now, place food on top of the hot rocks, meat to the centre and vegetables towards the outer edge. There must be a gap between the food and the earth. Lay saplings across the pit and place sacking, leaves and so forth on top of them, covering the lot with the earth which you excavated to keep the heat in. The hole now acts rather like a pressure cooker. After 1½ hours remove the cover – your meal will be cooked.

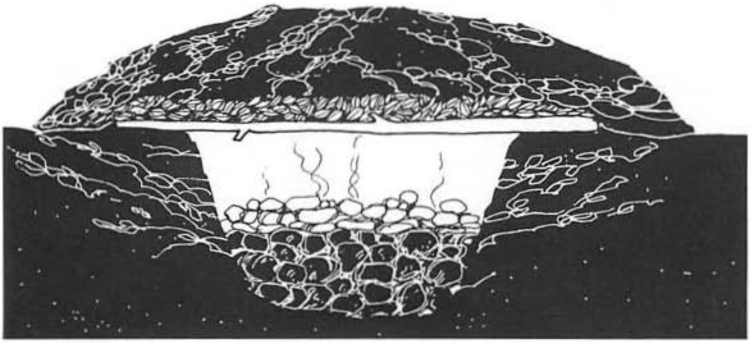

Boiling water in a hangi

If you have no container in which to boil water you can make use of the hangi. Whatever you have collected water in, provided that it does not melt (so that rules out plastic but includes other kinds of waterproof fabric), can be gathered up and tied so that the water does not spill and placed in the hangi). It will take about 1½ hours to boil but the fabric will not burn through.

USEFUL UTENSILS

TONGS

Choose two branches, both with a natural curve, and lash them together so that they want to spring apart at the free ends. Or use a tapering piece of wood between them under the lashings to hold them apart. If one has a forked end the grip will be improved. Use for holding pots, hot rocks and logs.

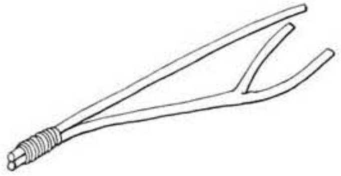

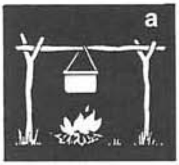

POT ROD

To give more variable access to the fire than a rail over it (a), drive a sturdy forked stick into the ground near the fire – but not so close as to catch alight. Rest a much longer stick across it with one end over the fire. Drive the bottom end of the longer stick into the ground and prevent it from springing up with heavy rocks. Cut a groove near the tip to prevent pots from slipping off, or – to be safer – tie on a strong hook.

Two or three sticks could lean over the fire at different heights with meat or vegetables attached.

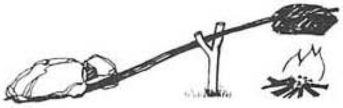

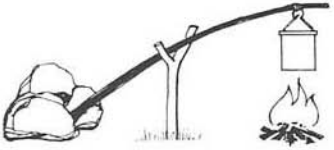

SWINGING POT HOLDER

This can be made from two forked sticks and a firm upright driven into the ground. Bind the branches together so that the forks fit in opposite directions on the upright. The cantilever action will maintain the height you set it at, and a push sideways will swing the pot away from the flames. With a longer upright you could control cooking height also.

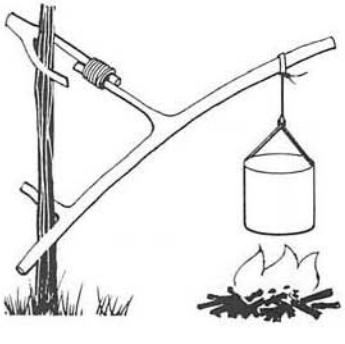

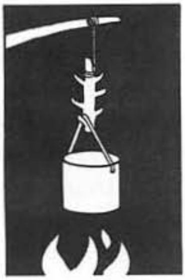

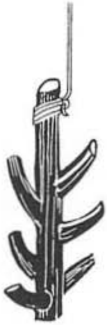

VARIABLE POT HOOK

Since the distance between the fire and the food will affect the speed at which the food cooks, make this hanging device so that you can control your cooking.

Cut a strong piece with several branches from a small tree or bush and trim the branches to 10–12cm (4–5in). Strip off the bark, which may hide a rotten branch.

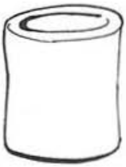

BAMBOO CUP

Cut a section of bamboo just below a natural joint and then cut just below the next joint up. Smooth the edges to prevent splinters.

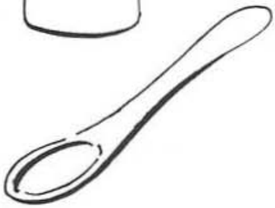

SPOON

Start with a flattish piece of wood and scribe a spoon shape on it with the point of your knife. Then whittle away to the required shape. Do not hurry – this will only result in mistakes. Never cut towards yourself or your hand.

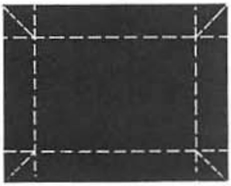

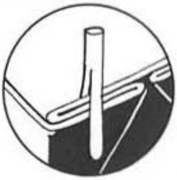

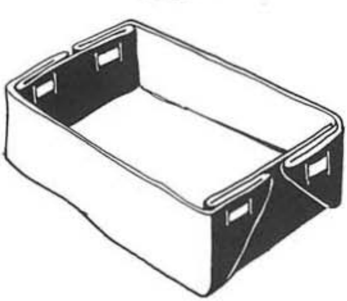

BIRCH BARK CONTAINERS

Use the inner layer of birch bark to make storage boxes or temporary cooking vessels – which can be used for boiling. Sew or tie them – near the top – to prevent unfolding. An alternative for temporary vessels is to peg the top edges with split sticks, but you might well spill the contents if the vessel suddenly unfolds.

Make another vessel, but with a larger base, and you will have a lid to fit over the first.

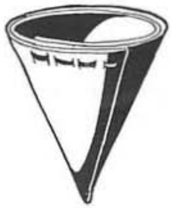

A circle, folded into quarters, will make a cone-shaped cup – or a boiling vessel if suspended.

PRESERVING FOOD

If food is not plentiful or is likely to be limited by season, it is important to ensure that stores keep safely.

Micro-organisms, such as moulds, that spoil food, thrive in warm moist atmospheres. Deterioration can be delayed by keeping food in cool places such as caves or by water, but that is only a short-term measure. More positive action must be taken to ensure long-term preservation. The main methods to use are drying, smoking, pickling and salting. Sugar preserves will not keep for very long unless you can vacuum-seal them, but will keep longer than as soft fruit, and alcohol is an excellent preservative if you set up the facilities to make it.

BE AWARE

When you have taken time and trouble to preserve valuable foodstuffs, particularly in areas where food is scarce, take equal trouble in storing your food.

Do not store in direct sunlight, near excessive warmth or moisture, nor where scavenging animals may ruin it.

Wrap, where possible, in airtight and waterproof materials – or store in containers (such as birch-bark boxes) with a good seal. Label if you are storing several kinds of food and separate to avoid cross-flavouring.

Check occasionally to see all is well.

Drying

Both wind and sun can dry food but, in most climates, it is easier to force dry food over a fire. Losing moisture shrinks size and weight, concentrating the nutritional value. Many moulds can grow when there is as little as 16 per cent moisture content, but few can grow on foods with 5 per cent or less and these will also be less vulnerable to maggots.

Pork, geese, seabirds and other meat with a high fat content are the most difficult to preserve. It is best to cut off most of the fat and rub salt into the flesh. Salt is a good drying agent. Hang the salted meat in a cool airy place.

Smoke drying

Smoking both dehydrates meat and coats it with a protective layer, like varnishing its surface. The inside is dry so no condensation takes place, and the outside is sealed against bacteria. Smoking can be best effected in a smoke house or a smoke tepee.

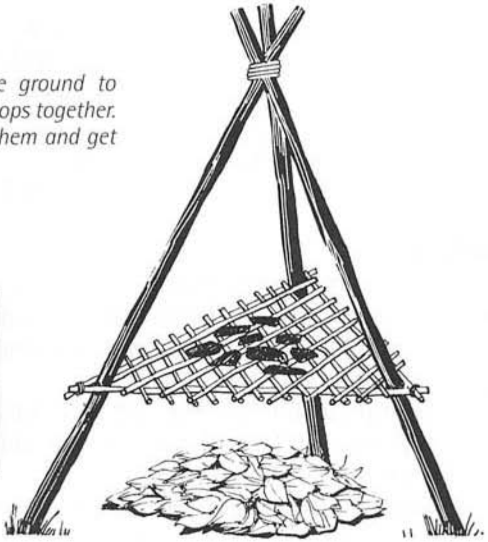

SMOKE TEPEE

Drive three sticks into the ground to form a triangle and tie the tops together. Build a platform between them and get a fire going beneath.

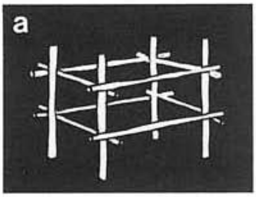

As an alternative to the tepee make a square frame of uprights (a) and cross-pieces supporting a smoking platform with the fire beneath and used in exactly the same way as the tepee.

In both cases meat should be cut into lean, fat-free strips and fish gutted and filleted. The strips can be any length but should only be about 2.5cm (1in) wide and 6mm (¼in) thick.

Get a fire going to produce a pile of hot embers. Have a pile of green leaves ready. Leaves from hardwood trees are excellent, especially oak, but avoid holly and other toxic leaves and conifers which tend to be resinous and may burst into flame. Do not use grass. Some leaves will give meat an individual flavour; pimento leaves are particularly distinctive.

Make sure that there are no flames left in the fire and pile the leaves over the embers. Cover the whole structure with a cloth to keep in the smoke. If you do not have a suitable material, have boughs and turfs ready to pile rapidly on the frame and seal it. Leave the structure sealed for 18 hours ensuring that little or no smoke escapes.

If the embers in a smoke tepee burst into flame, there is a risk that the whole structure may catch alight. This can be avoided by building a fire in a chamber in a bank (see *Snake hole* in *Fire*) *with the tepee erected over the chimney. This also makes it possible to tend the fire and to ensure a more extensive supply of smoke, which will be cooler than from a fire directly underneath. The food will dry slowly and become coated with smoke without being cooked.*

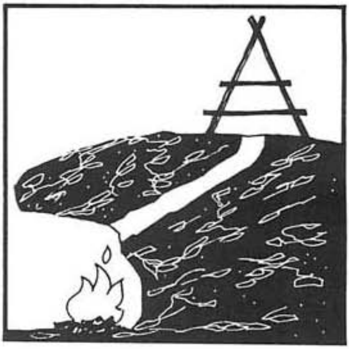

Biltong

This is sun-dried meat. Biltong is the Afrikaans name, it is also known as jerky, from the North American Indian *charqui.* It does not keep as efficiently as smoked meat and should be used only when smoking is not practicable.

Cut strips, as for smoking, and hang them up in the sun. Make sure that they are out of the reach of animals and about 2–3m (6–10ft) from the ground.

It may take two weeks for meat to dry and all this time it must be kept dry, so protection from rain must be provided. The strips must be turned, if necessary, to make sure that all surfaces are thoroughly dried, and, initially at least, flies must be kept off so that they do not lay eggs on the meat.

DRYING FISH

Preserve fish as biltong. Cut off heads and tails and gut. Split open. Remove backbone and lay on hot sun-baked rocks. Score inner flesh to speed up drying.

Small fish, under 7.5cm (3in) long need not be gutted.

Fish can also be smoked. They should be opened out, but it will be easier to hang if cleaned and gutted without removing the backbone, head or tail. Suspend by one side of the head.

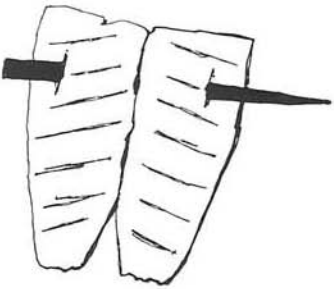

Pemmican

This is a nutritious concentrated food made from biltong – excellent for provisions to carry with you if you decide it is time to trek to safety. It contains all essential minerals and vitamins except vitamin C.

You need an equal quantity, by weight, of biltong and of rendered fat. Shred and pound the meat. Melt the animal fat over a slow fire, without allowing it to boil. Pour the fat over the shredded biltong and mix them well together.

When cold pack the mixture in a waterproof bag. It will keep for a long time, especially in colder climates.

Pickling and salting

Citric acid obtained from wild limes and lemons can be used to pickle fish and meat. Dilute two parts of fruit juice with one of water, mix well and soak flesh in this for at least 12 hours. Now transfer it to a covered, and preferably airtight, container and with sufficient solution to cover all the meat. Vegetables with a high water content are difficult to preserve. Pickling is best for them. Alternatively, if salt is more easily available than citrus fruits, they can be boiled and then kept in brine (saltwater). Boiling kills off bacteria and the brine keeps fresh bacteria away from the food.

The usual way of making sure that a brine solution is sufficiently strong is to add salt until a potato will float in it. In lieu of a potato try a small fruit or root vegetable which fails to float in salt-free water (not apples – they float too easily). Another method of using salt is to pack tightly layers of salt and vegetables such as beans and peas, thoroughly washing off the salt when you need to use them.

Nuts and cereals

These keep reasonably well provided they are not allowed to get damp but will keep better if dried. Place them on hot rocks from the fire, turning them frequently until thoroughly dried. They should then be kept in damp-proof containers.

Fruit, fungi and lichens

Fruit and berries can be dried whole or cut into slices and dried by sun, smoke or heat. Fungi also dry very readily – the *Boletus* species especially. Fruit can usually be eaten dry. Fungi can be added to soups and stews or soaked in water for several hours to regain some of their texture if being used in other ways.

To store lichens, soak them overnight, boil well and dry. Grind to a powder then boil again to form a thick syrup, which can be kept in a sealed container and used to give body to other foods.

COOKING TIPS

Meat

Meat is best cut into small cubes and boiled. Pork is particularly suspect in hot climates: wild pig is usually infested with worm and liver fluke. Venison is also prone to worms.

Put excessively tough meat in a solution of juice from citric fruit for 24 hours. This marinading helps to make it more tender. Bring to boil and simmer until tender.

Offal

Check liver especially carefully. If firm, odourless and free from spots and hard lumps it can be eaten. Boil first, then fry if you wish. Hearts are best par-boiled then baked. Brain (if not used for preserving hides) makes an excellent stew. Skin the head and boil, simmering for 90 minutes. Strip all the flesh from the skull, including the eyes, tongue and ears.

Blood

Leave in the container in which it is collected but keep it covered. A clear liquid comes to the top. When separation seems complete drain it off. Dry the residue by the fire to form a firm cake. Use it to enrich soups and stews.

Sausages

Thoroughly clean intestines, turning them inside out to wash. Fill with a mix of half meat/half fat bound with enough blood to hold the ingredients together. Tie the ends and boil. Once cooked they can be preserved by cold smoking in a smoke tepee over a chimney.

Fish

Usually germ-free if caught in fresh water. Fish take little cooking and are best stewed or wrapped in leaves and placed in hot embers – chickweed and butterbur are good for this: avoid toxic leaves.

Birds

Boil all carrion. Old crows, blackbirds and parrots are tough and best boiled. Young specimens can be roasted – stuff the bird with herbs and fruits.

Reptiles

Best gutted and then cooked in their skins which are rough and leathery. Place in hot embers and turn continually. When the skin splits the meat can be removed and boiled. A few snakes have poisonous secretions on the skin and others may have venom glands in their head, so cut this off before cooking. If you are not sure whether they are safe, take care in handling them.

Skin frogs before cooking (many frogs have poisonous skins). Roast on a stick.

Turtles

Boil turtles and tortoises until the shell comes off. Cut up the meat and cook until tender.

Shark meat

Has a bad taste unless correctly cooked. Cut into small cubes and soak overnight in fresh water. Boil in several changes of water to get rid of the ammonia flavour.

Shellfish

Crabs, lobsters and shrimps, crayfish, prawns and so forth are safer boiled since they may contain harmful organisms. All seafood spoils quickly and must be cooked as soon as possible. Drop into boiling salted water and boil for ten minutes.

If you are sure the food is fresh, a clambake is a delicious way to prepare mussels, clams and similar creatures. Dig a hole in the sand and light a fire alongside with stones on top. When the stones are hot place them in the hole, put the shells on top and cover with wet seaweed or grass, then a 10cm (4in) layer of sand. The hole will develop a lot of steam which cooks the molluscs.

Insects and worms

Best boiled. Cook and mince them by crushing in a can.

More acceptable dried on hot rocks and then ground into a powder with which to enrich soups and stews.

Eggs

Boiling is the best way of cooking, but if no container is available roast after first using a sharpened stick or the very sharp point of a knife to pierce a small hole in one end. Place on warm embers to cook slowly. Slow cooking reduces the risk of cracking. Remember that banana leaves, which are full of oil, make an excellent frying pan. Place over hot embers and crack the egg onto the leaf. If a boiled egg contains an embryo chick remove the embryo and roast it.

Green vegetables

Wash in clean water and boil for just long enough to make them tender – they are often and easily overcooked. Tender plants can be gently steamed if you are sure that they are safe to eat. Add to stew after the meat is cooked and already tender. Eat fresh greens raw as salad.

Roots

Some are toxic but the toxins are destroyed by heat. Always cook roots; boiling will make the toughest ones tender. Roast roots are tasty – but boil them first. Try boiling for five minutes then place them in a hole dug beneath the fire, cover with ash and embers and leave until tender.

Lichens and mosses

Soak overnight in clean water. Add to stews.

Sago

Proper sago comes from the sago palm, but buri, sugar, fishtail and, in the American tropics, cabbage palms can be used in the same way. The average sago palm yields about 275kg (600lb) of sago – enough to feed one person for a year. Cut down the palm at the base of the trunk, trim off the tip just below the last flowering line. Divide a large trunk into sections.

Cut lengthwise – hard work, for the outer bark is 5cm (2in) thick and hard as bamboo. Using each section as a trough, pound pith into a mash, then knead in a container of water (the bole of the trunk will do) and strain through a cloth. A starchy paste will precipitate in the water. Roll this into sticky balls and cook.

Sap

Palm sap is extracted from flowering parts, not the trunk. Choose a fat stalk carrying a flowering head (at the base of the crown of the trunk). Bruise with a club to stimulate flow of sap, then cut off head. Sweet juice will flow from the end of the stalk – 1.5 litre (2½pt) per day. Bruise and cut daily to stimulate flow. Drink raw or boil then cool it to produce toffee-like lumps of almost pure sugar. Sugar, nipa, coconut and burl palms can all be used in this way. (Extracting resinous sap is dealt with under *Trees* in *Food.*)

Grains and seeds

Grains are enclosed in a husk. Dry them thoroughly to allow the grain to crack out. Thresh (or thrash) it with a flail, stick or rock, or, if the grain is very malleable, rub it between your hands. Shake out on to a flat container and occasionally toss into the air in a breeze (winnowing). The husks (chaff) will blow away leaving the heavier seed behind.

Pinole

Parch husked seeds on hot stones by the fire. The heat will cook and dry seeds without roasting them. This pinole will keep well. Eat cold or reheat. Add to stews or place a handful in a mug of hot water – tasty and nutritious. Dry they will not be properly digested, but they will fill the belly. It is better to grind them into flour.

Flour

Grinding flour without a proper mill is hard work but can be done by pounding with a smooth stone on a hard surface. Look for a large stone with a depression in the middle to place the grain in. Use a circling action as with a mortar and pestle. Another way to grind flour is to hollow a tube of hardwood and to pound a stick up and down inside it on the grain.

Mix flour with a little water and knead into a dough. Bake in an oven or make into thin strips, wrap around a shaven green stick and cook over hot embers. Another method is to make the dough into fist-size balls, flatten them and then drop hot pebble-size stones into the centre and wrap the dough around them. Lick your fingers before picking up the pebbles – if you are quick the moisture stops the pebble from burning you – or use sticks or tongs (see *Useful utensils*) to lift them.

Flour does not have to be made from cereal grains. Use the flowering heads of cat's tails or boil and mash up peeled roots, of wild calla for instance, or edible barks. Those that are not harmful raw can be steeped in water and crushed with a stick or stone to free the starch. Remove fibres, leave starch to settle, then pour off the water and you will have your flour.

ORGANIZING THE CAMP

In many survival situations there will already be someone in a position of responsibility who will head the organization of the camp and lead the development of survival plans. If no established command structure exists among a group of survivors, an organizing committee should be established and individuals nominated and elected with particular responsibilities, perhaps on a rotational basis if it is a large group and rescue does not come quickly. Experience must be pooled and immediate steps taken to discover what skills individuals can contribute.

A roster is essential for such daily chores as collecting firewood and water, foraging, cooking, latrine digging and maintenance tasks, and for hunting and trapping.

In a group of survivors there may be all kinds of people of different ages and experience. People will have varied skills and enjoy doing different tasks. Everyone who is fit and able should take their turn at the unpleasant tasks, unless their skills are so much in demand that it would be a waste of their abilities, but individuals should do what they are good at – and be encouraged to develop skills for which they show an aptitude.

Not only should everyone do their fair share but keeping busy eliminates boredom and keeps up morale. Anyone who is sick or injured gets the lightest jobs and is best employed around camp until they have recovered. In a group there should always be someone in camp, and they should be able enough to operate the signals should a search aircraft appear. If you have sufficient numbers do NOT venture from the camp in less than pairs.

Except in the desert, where the day will be largely spent sheltering from the sun and early morning and evening are the times for activity, daytime is likely to be fully occupied. Evenings, however, may drag if not occupied by hunting. A gathering around the camp fire will help establish a pattern and provide a sense of discipline and normality. It will give an opportunity to debrief on the day's events, to plan for tomorrow and to discuss new strategies.

Music can be a great morale booster. If no instruments were carried or survived, simple ones such as percussion or pan-pipes can easily be improvised, and everyone can sing after a fashion.

Sing-songs, dancing, charades, quizzes and story-telling all have their place and you may have talents which can create more elaborate entertainments. For private recreation any books will be invaluable and you can make pieces for board games such as draughts and chess, using stones for counters or carving simple playing pieces.

Even the lone survivor requires discipline and order. A regular routine will help morale and exactly the same care must be taken to ensure that the camp is kept in good order. At first there may be so much to do that the individual is too tired to think of recreation, but boredom is even more dangerous for a person on their own and objectives should be set each day whether practical or for amusement.

CAMP HYGIENE

Keeping healthy is an important factor for survival, so strict hygiene should be practised, not only personally but in the planning and running of a camp. Rubbish and latrines must be kept away from the camp to reduce the threat from flies and, since most of the common diseases in a survival situation are water-borne, pollution of drinking water must be rigorously avoided. Food scraps and other rubbish should be burned in the fire if possible.

Camp layout

Select sites for all camp activities so that they do not interfere with each other or pollute the living and cooking areas. If you are camped by a river or stream, fix specific sections for activities and keep to them. Latrines should be dug downhill of the camp and away from the water supply so that there is no possible risk of seepage polluting either.

Activity areas

Establish a water point from which drinking water will be collected and ensure that no one washes, cleans pots, scrubs clothes or otherwise uses the stream upstream of this point. Downstream choose a wash point for personal ablutions and clothes washing and further downstream of that select a place to be used for cleaning cooking utensils.

Latrines and rubbish disposal should be well away from the camp – and preferably downwind – but not so far away that it is inconvenient and people are tempted to go elsewhere. If necessary cut a track to it to make access easier.

REMEMBER: NEVER urinate or defecate in or near your water supply.

Latrines and rubbish disposal

It is important that proper latrines be established, even for the lone survivor. With a group separate latrines for the sexes may make a mixed group feel more comfortable, and as much privacy as possible should be provided. Rubbish, after checking that it really has no useful value, should be burned, and what cannot be burned should be buried.

Even if you have it, do NOT use disinfectant in a latrine. Lime or disinfectants would kill the useful bacteria that break waste down and then it WILL start smelling! After defecating cover the faeces with earth. Add small amounts of water which will promote the bacteria.

Make a latrine cover to keep out flies and remember always to replace it, or flies that have walked all over faeces may walk all over your food and start a cycle of infection.

If, after a time, a latrine starts to smell, dig a new one. Fill in the old latrine. Build a new seat and burn old timbers and covers.

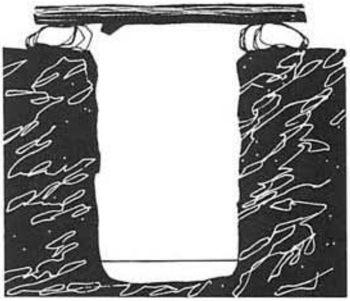

DEEP TRENCH LATRINE

Dig a trench about 1.25m (4ft) deep and 45cm (18in) wide. Build up the sides with logs or rocks and earth to make a comfortable sitting height sealing the gaps between them. Lay logs across to leave only a hole for use (or several if you are a large group and making a communal latrine). Empty wood ash on the logs to make a seal – it will also deter flies.

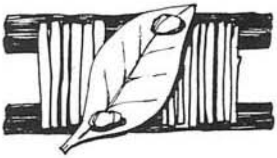

Make a lid of smaller wood to cover the opening or use a large flat rock or a large leaf weighted down with stones. Always remember to replace it.

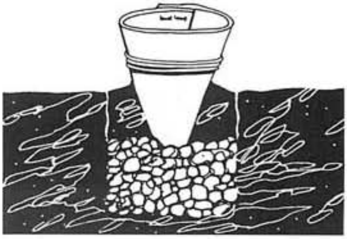

URINAL

Dig a pit about 60cm (2ft) deep. Three-quarters fill it with large stones and then top up with earth, with a cone made from bark set into it as a funnel. Site it close enough to the camp to ensure that people bother to use it.

Incinerator

If there is too much waste for the camp fire to burn, make a separate fire in the latrine area. If a large can is available use it as an incinerator. Bury any unburned refuse in a garbage pit.

CAMP DISCIPLINE

- Do not prepare game in camp: bleed, gut and skin on the trap line. This attracts game to the traps where you want them, not into your camp.
- Keep food covered and off the ground. If kept in trees make sure it is proof from tree-dwelling animals.
- Replace lids on water bottles and containers immediately after using them.
- Stow spare clothing and equipment in your shelter. Do not leave it lying where it can get wet or burned.
- Have a place for everything and keep things tidy: a tree for mess tins and cooking utensils – hook them on twigs and branches, a place for mugs and spoons – and keep everything off the ground. Fit a box as a cupboard on a tree trunk.
- Never leave the fire unattended.

Soap

Washing with soap removes natural oils, leaving the skin less waterproof and more prone to attack by germs. In survival circumstances it is a mistake to wash with soap too often. However, soap is the most widely used antiseptic, better than many others, such as iodine, which destroy body tissue as well as germs. It is ideal for scrubbing hands before administering first-aid for wounds. Save supplies for this.

Soap-making

Two ingredients – an oil and alkali – are needed to make soap. The oil can be animal fat (including fish) or vegetable, but not mineral. The alkali can be produced by burning wood or seaweed to produce ash.

METHOD: Wash ash with water. Strain and boil with the oil. Simmer until excess liquid is evaporated and allow to cool. This soap will clean the skin but is not antiseptic. Adding horseradish root or pine resin to the brew will make it antiseptic.

Experiment will be necessary to get the balance in the mixture right. Start with more oil than alkali because too much alkali will dry the skin, leaving it sore.

TOOLS

Before humans discovered metals, and learned to work them, tools were made from stone (especially flint, obsidian, quartz, chert and other glassy rocks), bone and other natural objects. Stones can make efficient hammers, alone or lashed onto a handle if in a shape that can be easily secured. The glassy stones can be knapped (chipped and flaked) to make a sharp edge; some other kinds of stone, such as slate, can also produce a knife edge, though they may not have the strength to be used for percussive blows.

The best start for a stone implement is a split cobble, perhaps from a stream bed, or split by a blow from another smooth, hard pebble, so that a flat face is produced. The blow should be at an angle of less than 90 degrees or the shock will be absorbed within the pebble. Once the split is made, other layers can be broken off. Flakes can be removed around the edge of the flat face by hitting edge-on with another stone. Delicate work can be produced by hitting and pressing with a softer tool such as deer's antler. If a flat face is produced, end-on blows can make thin blades.

Making stone implements is not a skill that can be quickly acquired and you may have to be very persistent.

Bones can be used as tools – antlers and horns make useful digging implements, gougers and hammers. They can also be cut with stone tools or ground with coarse stones. You may have the advantage over the prehistoric toolmakers, whose skills you are copying, of having a knife or other metal implement to help you to carve bone.

Some woods, such as the Mulga tree of Australia, used by the Aborigines for spears, are hard enough to make effective blades for hunting and cutting.

STONE TOOLS

One technique for producing an axehead: First the stone is split and the edges partly shaped (a), then a platform is created on one side (b) from which a series of flakes can be struck vertically down (c).

The final shaping can then be done. Hit with softer stones, and hit and press small flakes away with a piece of antler or hard wood.

It may take a lot of patience to acquire tool-making skills but even practise flakes may be useful as scrapers, for cutting edges and as arrowheads.

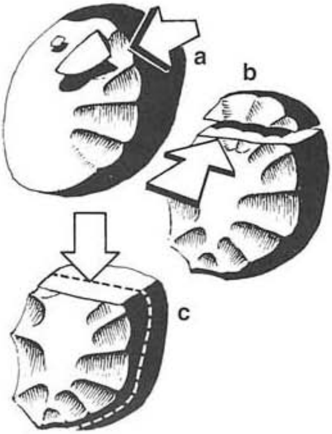

BONE TOOLS

A shoulder blade provides a good shape for an effective saw (a).

First it should be split in half then teeth can be cut along it with a knife.

A small bone scraper (b) could also be made, the edge ground sharp.

Ribs are good bones for shaping into points (c).

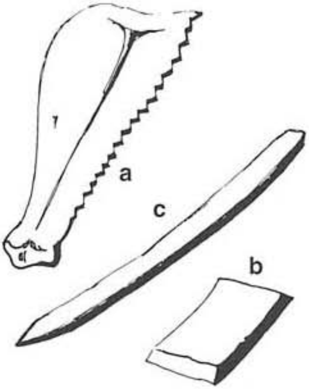

BONE NEEDLE

Choose a suitably sized bone or flake of bone and sharpen to a point. Burn an eye with a piece of hot wire, or lacking that, scrape with a knife point or piece of flint. DON'T heat the knife in the fire.

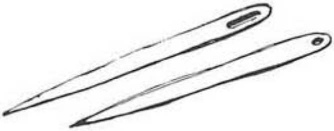

AXES

A fire axe is part of the equipment of any boat or plane, but an axehead, to be fitted to an improvised handle, is a useful additional piece of equipment for anyone to carry. One of about 500–750g (1–1½lb) is ideal. Use your knife to fashion a handle when needed.

AXE HANDLE

Any straight, knot-free hardwood is suitable – ash and hickory are ideal. In the tropics the flukes of a buttress tree (a) are excellent: slightly curved, straight-grained and easy to work.

Cut two notches into the fluke of a buttress spaced to the desired handle length (b to c).

Hit along the side of fluke close to the cuts. It will split away at their depth.

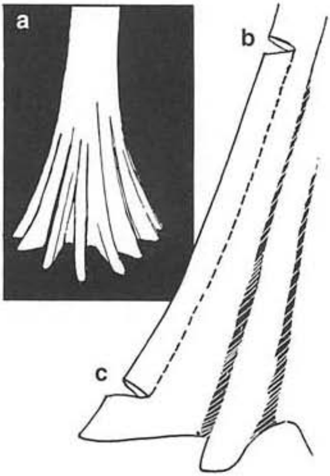

FITTING THE HEAD

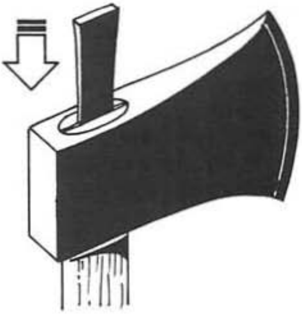

Whittle the handle into shape with one end cut to fit the hole in the axehead, cutting a notch in that end. Make a wedge to fit the notch.

With the head in place drive in the notch then soak the axe in water overnight to tighten the head on the shaft.

Always check axeheads for tightness before using them.

FITTING A STONE AXEHEAD

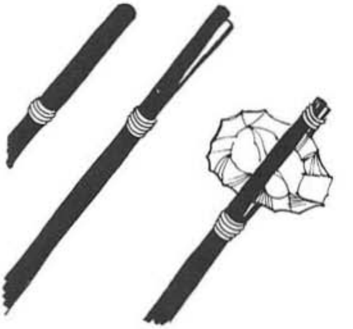

Select a hardwood handle. Tie a band of cord around it about 23cm (9in) from one end. Split the end down as far as this band (use your knife and a wedge or the piece of flint you have made for the axehead). Insert the flint and tie the end to secure.

This mounting will split wood but will not be very effective for chopping it.

Sharpening an axe

An axe with a blunt edge becomes no more than an inefficient hammer so keep it sharp, it will save energy. A file is best for getting rid of burrs, and a whetstone for imparting the sharp edge. A file is a one-way tool – it works when pushed, not pulled.

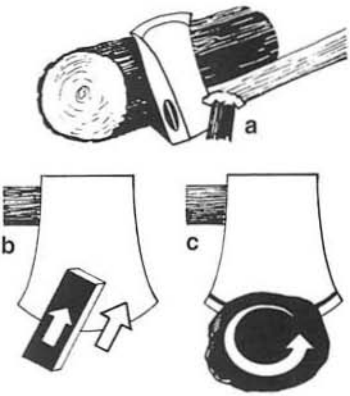

Prop axehead between a log and a peg (a). Always try to sharpen INWARDS from cutting edge to avoid producing burrs.

Use a file or rougher stone first to remove rucks and burrs (b). Then finish with a smoother stone, using a circular motion. Don't drag the stone off the cutting edge. Push ONTO blade. (See Knives in Essentials).

Turn the axe over. Repeat the process circling in the opposite direction.

Using an axe

Most people have a natural prime hand and swing. Use an axe in a way that is comfortable to you, swinging it in an arc that feels natural with a firm grip and ALWAYS away from your body, hands, and legs. Make sure that, if you miss the tree or other point of aim and follow through, the axe will not strike you or anyone else. NEVER throw an axe on the ground. Sheath it or bury the blade in a log.

Tree felling

Check overhead for dead branches, which may fall and injure you, and for hornets' nests. Clear branches or creepers which could deflect your blows. If roots or the bole spread out at the bottom, build a platform to reach a thinner diameter trunk and reduce the effort of chopping. Ensure such a platform is stable and that you can jump off it quickly if the tree falls the wrong way!

Work at a comfortable height and try to cut downwards at 45 degrees, although every now and then a horizontal blow is needed to clear the cut.

Cut from both sides of the tree, first chopping out a notch at an angle of about 45 degrees and another on the opposite side at a lower level, ON THE SIDE TO WHICH YOU WANT THE TREE TO FALL (a). Do not cut through more than half the tree before starting the other notch.

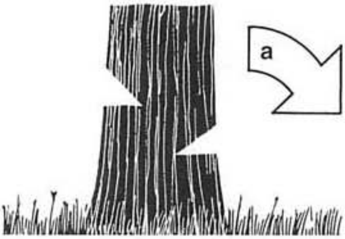

If two people are at work on opposite sides this is particularly important. A leaning trunk or a tree with most of its branches on one side will fall in the direction of its weight and the placing of the cuts will not affect it.

A steady rhythm of blows will cut more effectively than trying to make fewer big blows. If you put too much effort behind the axe your aim will suffer and you will soon tire. Let the weight of the axe do the work.

Alternating the angle of stroke will prevent the axe from jamming. Too steep an angle will cause the axe to glance off the trunk (a). Dead-on will make it jam or be inefficient (b). Aim for 45 degrees (c).

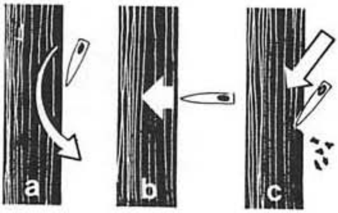

TO REMOVE BRANCHES

Cut off branches from the outside of the fork (a), not the inside (b).

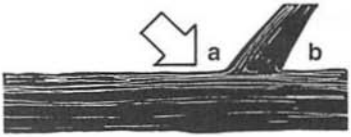

SPLITTING LOGS

Stand behind a large log with feet well apart. Swing down to cut the side away from you (a).

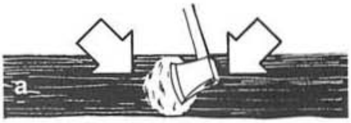

Do NOT chop downwards (b).

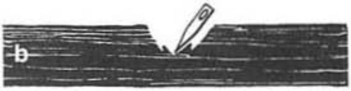

To split a smaller log, angle against another log (c). DO NOT PUT YOUR FOOT ON IT.

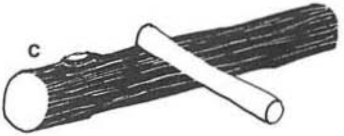

Alternatively, hold smaller log against cutting edge of axe and bring both down together (d) on to a larger log. (Not to be tried holding too short a log for safety.)

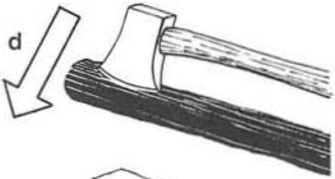

If in doubt split larger logs with a wedge and a rock (e). DO NOT HOLD WOOD UPRIGHT IN YOUR HAND AND ATTEMPT TO SPLIT WITH AXE.

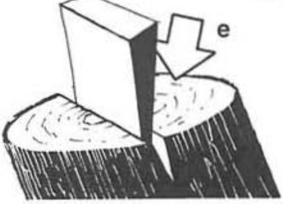

BROKEN HANDLES

Using an axe takes practice and while gaining experience axe handles often get broken – usually because the head misses the target and the handle takes all the blow (a). To remove a broken handle, the easiest way is to put it in a fire, burying as much as possible of the metal in the earth to prevent it losing temper – single-headed (b), double-headed (c).

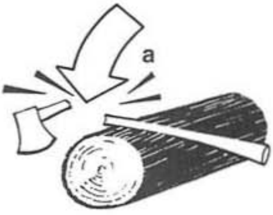

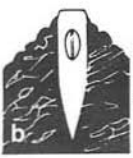

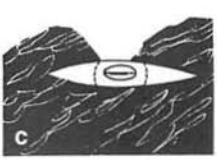

USING A FLEXIBLE SAW

Always use a flexible saw so that the cut opens up (a) rather than closes tight (b) on the saw, causing it to jam. Do not pull too hard or the saw may break.

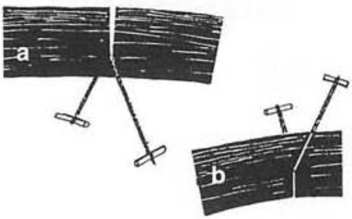

At all times keep the wire taut (c) pulling in a straight line, never at angles (d).

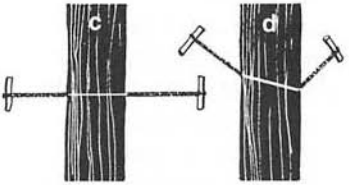

With two people the rhythm must be carefully maintained. If a kink is produced in the saw (e) it may break.

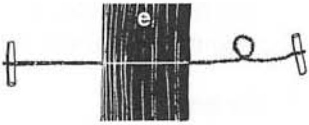

It is usually easier for a single person to cut a log by pulling upwards (f). Support the log to keep it off the ground and give it an angle to keep the cut open.

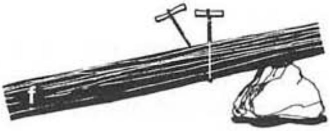

Alternatively, to remove a branch, pull down from above the head (g). This could be dangerous.

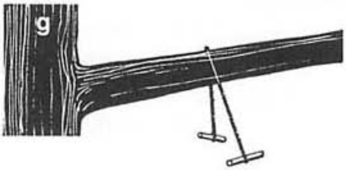

Very high branches can be removed by attaching strings to the saw toggles to give extra reach. This is dangerous. Keep your eye on the branch and be prepared to jump out of the way.

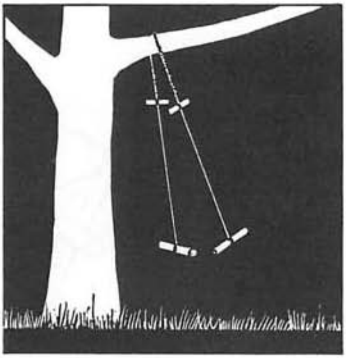

FURNISHING THE CAMP

BEDS

Sleep is the great regenerator and a comfortable bed is worth a little effort. Some form of bedding should always be used to avoid lying directly on cold or damp ground. In tropical climates it is always better to raise the bed into the air, both to keep off wet earth and to provide a current of cooling air. In cold climates, when no permanent shelter has been erected, choose fuel which will keep the fire going through the night and build a screen to reflect heat back on to your sleeping space.

When the ground is dry, or under shelter, stones heated in the fire (but see *WARNING* in *Fire*) and then buried under a thin layer of soil beneath the bedding will keep their heat through most of the night, adding to your comfort.

A-frame beds

An A-frame forms a strong construction and will keep you off the ground. Drive two pairs of posts into the ground at an angle, leaving a little more than your height between the pairs. Lash the tops together. If the ground is hard, cross-members will be needed between the feet of each A-frame and between the two A-frames.

Tube bed

This simplest form depends on having or making a tube of strong material, sewn or thonged together. A large heavy-duty plastic bag is suitable, but not thin plastic. Really toughly made clothing could serve but would usually make a rather narrow bed.

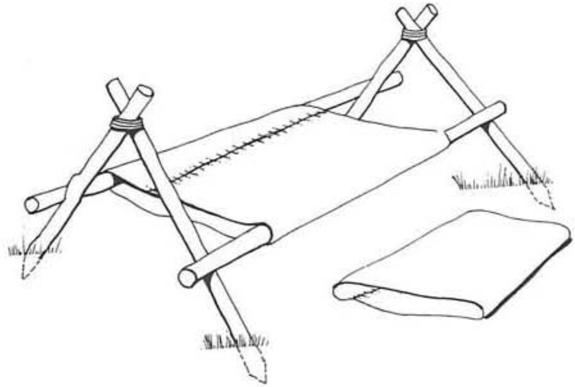

Do not risk this method with any fabric that might give under your weight or with seams that might come apart.

Make A-frame supports and choose two fairly straight poles, each slightly longer than the distance between the frames, and pass them through the tube of plastic or fabric. Place them over the frames so that they rest on the sides, the tube preventing them from slipping lower.

Bough bed

Where conifers are growing, fir tree branches arranged in alternate layers will be comfortable and their fragrance will ensure a good night's sleep.

Ladder bed

Make A-frame supports and select poles as for the tube bed, but you will also need a number of cross-pieces – how many will depend upon the size of the sleeper and the kind of bedding to be used. Springy saplings will be more comfortable than boughs, provided they are strong.

Lash the end 'rungs' to the A-frames, jutting out either side. Make these of strong timber and lash them securely. Fit the ladder over the frames and lash in place. Lay bedding of bracken, ferns or leaves.

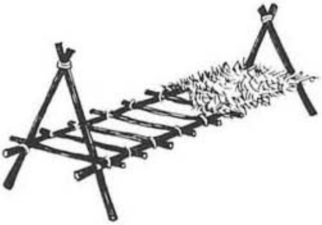

Hammocks

It takes a little practice to sleep in a hammock – in a string one you must push the sides outwards so that it cups around you.

PARA-HAMMOCK

It is possible to make a hammock from a parachute. The rigging lines run through the fabric right up to the apex. Cut five panels as shown but do not cut the rigging lines above.

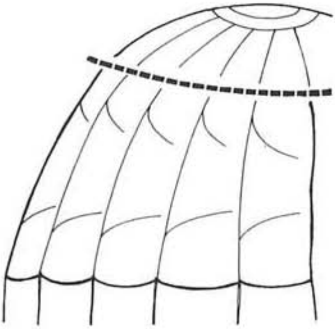

Overlay the panels as shown (a) and flatten out.

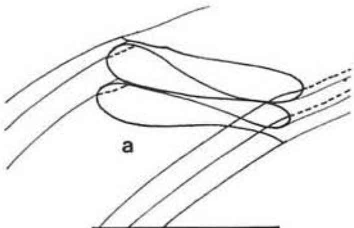

Grasp the three rigging lines in the corner and tie, close to the fabric, in an overhand knot (b).

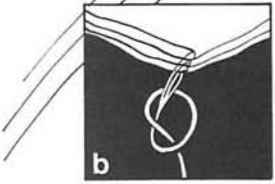

Then plait the three lines together (c) (see Knots*).*

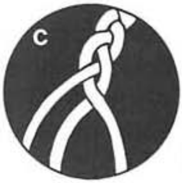

Insert spacer bars, notched at each end (d) to take the lines and tie your hammock to two convenient trees or stout posts.

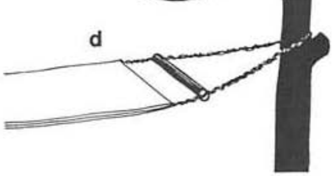

If you rig a line over the hammock (e) you could drape a groundsheet or poncho over you to form a quick shelter. If you do this be sure one end of the hammock is secured with a quick-release knot (see Knots*). It might be necessary to get out of the hammock in an emergency.*

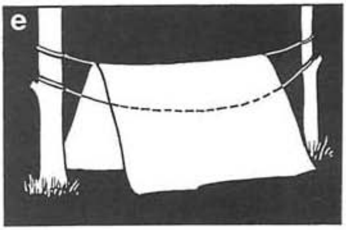

If you have another piece of rope, or plait some from unused rigging lines, you could tie both the hammock strings to a fixed loop, such as a bowline (see Knots*). Experiment with different ways of suspending your hammock, until you find the most stable and comfortable.*

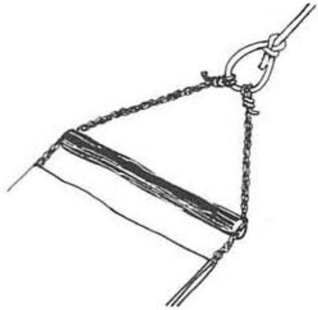

SEATS

Never sit on damp ground. Use something, even if it is only a log. If there is no ready-made seat available, lash together a couple of low A-frame supports and rest another bough across them.

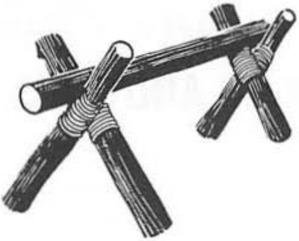

Make a simple box frame with cross-members linking legs from short lengths of wood. To make a seat: weave vines or twine back and forth or sew on a piece of canvas or plastic with thongs. Failing these, try a flat piece of wood or metal laid across, or thin springy saplings lashed to the frame and interwoven.

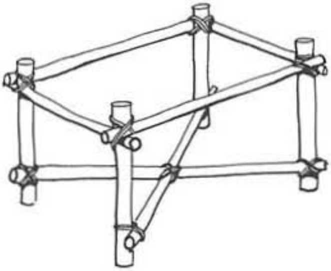

LADDER

Food-collecting, shelter-building, trap-setting and a whole lot of other tasks will be easier with a ladder. This one is easily made by lashing cross-pieces to two long poles. Because these are set at an angle, not parallel, the rungs will not be able to slip down.

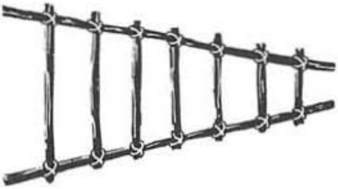

TRAVOIS

For bringing fuel or your prepared kill back to camp, or for other loads, a travois will work if the ground is fairly smooth – it will not on rough and boulder strewn terrain. Choose two boughs with some spring to them and lash cross-pieces, as for the ladder. Add additional struts to provide closer support.

Pull the load on its 'runners' like a sled. If you are pulling loads over a short distance, lash the runners to come to a single grip (a). For a larger version leave the last space clear or fit leather or fabric shoulder straps to haul it by (b).

(For carrying equipment and sledges see *On the Move*).

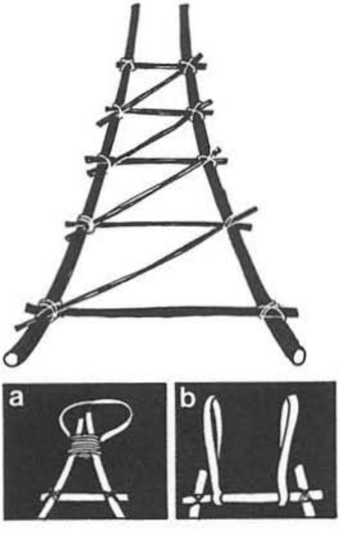

ANIMAL PRODUCTS

SKINS AND FURS

All animals provide skins. Their condition will depend on how carefully they were removed, the way the animal was killed (which may have damaged the skin), the age of the animal and time of year (mating season, moult and change of season can affect the amount and colour of fur in some species). Common defects are due to parasites, disease, malnutrition and scars from fight injuries.

Snakes, lizards, crocodiles and other reptiles all provide excellent skins. So do large birds such as ostriches. Some aquatic mammals, seals and their relations, are fur-bearing, like land mammals, and whales and dolphins have strong hides. Sharks also have a hide, instead of scales like most other fish. Birds can be skinned with the feathers attached and used to make warm clothing or bed covers. Skin is a source of food and in circumstances of acute shortage can be eaten, even after being preserved and used for clothing, but it is very tough and takes a lot of digesting. There are cases of people surviving by eating their boots, though it should be emphasized that in all such cases plentiful water was available.

Skins and hides are composed of water and proteins and decay quickly if they are not specially treated to preserve them. How they are treated will depend upon whether you want to retain the hair or fur, but the initial stages will be the same in both. To make moccasins, shelters, laces, thongs, water bags or canoes, the hair is removed, but for warm clothing, bedding or a good insulating groundsheet it should be left on.

Properly prepared skins will be supple, yet strong, and resist tearing, abrasions, deformation or stretching. They are comfortable to wear, with good thermal insulation, but permeable to air and water vapour.

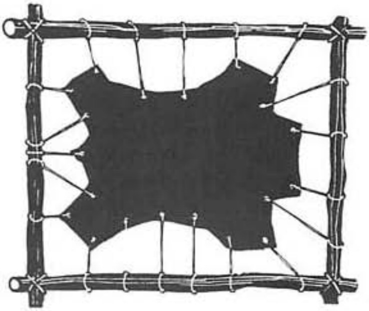

CLEANING THE SKIN

Make cleaning and drying a skin easier by stretching it on a frame. Do not make the holes for the cords too close to the edge. Remove fat and flesh by scraping the skin, using an edge of bone, flint or other rock, or even wood. Take care not to cut the skin. Remove every trace of flesh. Ants and other insects may help you if you lay the skin on the ground. Keep watch that they do not start to consume the skin itself.

To cure furs

Stretch the skin as tight as possible and leave it in the sun to dry out. All the moisture must be drawn from it so that it will not rot. Rubbing salt or wood ash into the skin will aid the process.

Do not let the skin get wet, or even damp, until the process is complete. Do not leave it where it will be exposed to rain or risk a covering with morning dew.

Keep it absolutely dry. If little or no sun is available, force dry over a fire, but keep the skin out of the flames and use only the heat and the smoke (which will aid preservation). Keep it away from the steam from any cooking pots.

Leather-making

After cleaning, place the skin in water and weight it down with stones. Leave it until the fur can be pulled out in handfuls – usually 2–3 days.

Make a mixture of animal fat and brains, simmered over a fire till they form an even consistency.

Scrape the skin on both sides, removing hair, and grain. Keep it wet. Work sitting down with the skin over your knees. Keep manipulating it.

Work the fat and brains mixture into the inner side of the still-wet skin, stretching and manipulating as you do so.

Dry the skin in the smoke over a fire, keeping it well away from the flames. The smoke sets up a reaction with the solution you have rubbed in to make the skin supple.

Laces and lashings

Hide is one of the best materials for lashings and for thongs to lace things together.

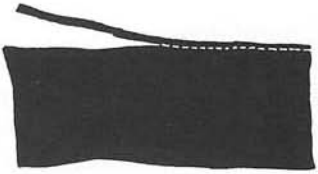

Cut short laces straight from the skin, along its length.

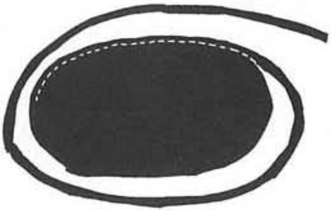

To obtain a greater length cut in a spiral – keep the width consistent or the thonging will have weak points.

Sinew as thread

The hamstring and the main sinews of the legs – especially of the larger animals – can be dried and used as thread to stitch hides together for shelter and clothing. Recognize them by their strong, white, cord-like appearance.

You can also use them for bowstrings and short ropes. They make excellent bindings for arrowheads. Sticky when wet, they dry hard.

Bladder

The normal function of the bladder is to hold water, so naturally the bladder of a large animal can be used as a water carrier – so can the stomach. Tie off the openings to seal them.

CLOTHING

When inadequately or unsuitably clothed for the situation you can improvise or supplement protection in many ways. Weaving fibres will only be practicable in a long-term situation and skins will not be available at first. Salvage towels, blankets, tablecloths, cushions, seat-covers, curtains, sacking from the wreckage – with a bit of initiative any kind of fabric can be used for garments, bedding or shelter.

Improve insulation and increase warmth by adding layers. Wear one sock on top of another and stuff dry grass or moss between them. Grass, paper, feathers, animal hair, etc, can be stuffed between other layers of clothing – newspapers give excellent insulation.

Waterproofing

Use plastic bags and sheets to improvise waterproofs or cut off large sections of birch bark. Discard the outer bark and insert the soft and pliable inner layer under the outer clothing. It will turn away much of the rain. Other smooth barks that peel easily can be used, but birch is best.

In the longer term, improve water-repellent qualities by rubbing animal fat or the tallow from suet into your clothing. Do NOT do this in situations of intense cold, where the reduction in insulation would be too great a loss and rain rarely a threat.

Footwear

Never underestimate the heavy wear and tear of rough ground on your feet. Climbing over rock and scree can soon destroy a pair of smart city shoes. High heels and sandals will soon break or wear in rough conditions.

- Cut shoe soles from rubber tyres, make holes around the edges for thongs to tie them over wrapped feet, or to sew onto fabric uppers.

- Several layers of wrapping are better than one on the feet. Tie on with thongs or use a triangular shape. Fold one point back over toes, make slits in front. Bring other points from behind the heel, through slits and tie around ankle.

- Moccasins can be cut from a single piece of leather, about 8cm (3in) bigger all round than the actual sole of your foot. Thong in and out around the edges and gather them in over wrapped feet. Tie off the gathering thongs, and weave another back and forth over the foot to make more secure. Alternatively, given more time and patience, more traditional moccasins can be made from a thickish hide sole, side strips and an upper. Measure around your foot first.

Goggles

To protect from glare at sea and in snow or desert cut a strip of material, paper, bark – but not metal – to tie over the eyes (or over the whole face in cold climates). Eskimos often carved goggles from wood. Make narrow slits for the eyes. Add extra protection by blackening beneath the eyes with charcoal to reduce glare.

Needle and thread

The Agave plant produces fibres for ropes and mats that are too rough for clothing but the end of the leaf is almost always a hard point which can be extracted with a fibre attached. It makes a perfectly threaded needle!

Clothing tips

- Tie long leaf strips and fibres around a belt or neck band to hang down as a 'grass' skirt or cape.

- Cut a head hole in a blanket or carpet and use as a poncho. Tie at waist or thong sides.

- Small skins are easily thonged or sewn together. Fur on the inside will give greater insulation but on outer garments the suede side sheds snow better.

ROPES AND LINES

There are thousands of uses for ropes and line, from securing constructional joints to making candle wicks, for rappelling down a cliff face to making snares and nets. Traditional materials for rope include hemp, coir (coconut fibre), Manila hemp (from the Abaca plant), henequin and sisal (both from species of Agave). Rope can be made from any pliable, fibrous material producing strands of sufficient length and strength. Much modern rope is made from nylon and other man-made fibres. They have the advantage of great inherent strength, lightness, resistance to water, insects and rot. However, nylon rope should not be the automatic choice if choosing equipment.

Nylon has the disadvantage that it can melt if subjected to heat – and friction on a rope produces heat. It is also slippery when wet. While its tensile strength is good, nylon also tends to snap if subjected to tension over an edge – it does not have to be a very sharp edge.

TYPES OF ROPE

Kernmantel type (a) encloses a central core of strands in an outer sheath. Easier to handle, except when wet or icy, but not so strong as hawser. It can unravel if cut. Traditional hawser-laid rope (b) has three bundles of fibres twisted together. If one is severed the others may hold.

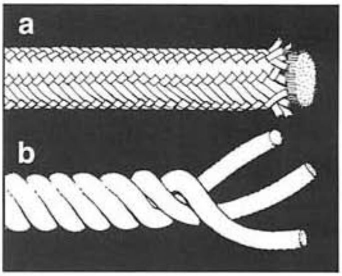

Choosing rope

Match type, thickness and length of rope you carry to the demands you expect to make on it. Nylon will have advantages in very damp climates and when weight is critical but remember its drawbacks. Thicknesses of 7mm ($^5/_{16}$in) and below are difficult to handle.

Rope about 9–10mm ($^3/_8$in) is usually recommended for lashings, throwing and mountaineering. It can be used for safety lines and for climbing, provided belay and abseiling techniques are used – it is not thick enough for a hand over hand and foot grip. A length of 30–40m (100–125ft) would then be as much as can be carried without encumbrance.

Climbing rope must be elastic, to absorb some of the shock, without putting enormous strain on anyone who falls. See if it has the approval of official mountaineering bodies or conforms to the British Standard 3184 (for hawser-laid ropes).

Taking care of rope

Rope should be protected from unnecessary exposure to damp or strong sunlight and (in the case of natural fibres) from attack by rodents and insects.

If it does get wet, do not force-dry it in front of a fire. Do not unnecessarily drag it along or leave it on the ground. Dirt can penetrate and particles of grit work away at the fibres from inside the rope. If weather conditions will make drying possible, it is worth trying to wash a very dirty rope in clean water.

Try to keep a rope for the job for which it was intended – do not use climbing rope as clothesline or lashing if you can avoid it – though in a survival situation you may have to use the same length for many purposes.

Whipping the end of the rope (shown later) will prevent it fraying. To prevent a rope becoming tangled, store and carry it in a coil or skein. It will be easier to handle and to pay out when needed.

Rope is valuable equipment. You may have to trust your life to it. Do your best to keep it in good condition.

SIMPLE COIL

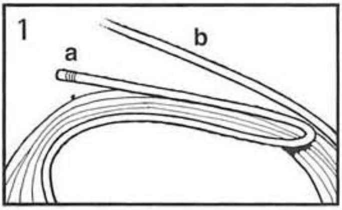

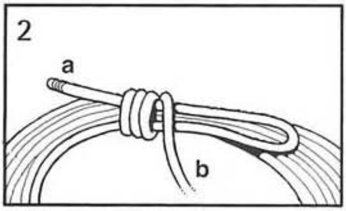

1 *Make a coil of rope 35–45cm (14–18in) in diameter, keeping each circle of the rope alongside the next without twisting or tangling. Leave a length at each end ready for fastening.*

2 *Bend one end (a) back along the coil and wrap it with the other end (b).*

3 *Feed the 'wrapping' end through the loop (c) and pull (a) to secure.*

4 *Tie off with a reef knot – shown later.*

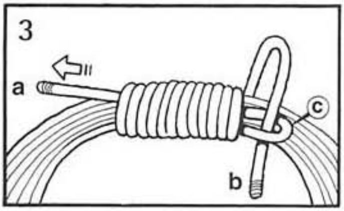

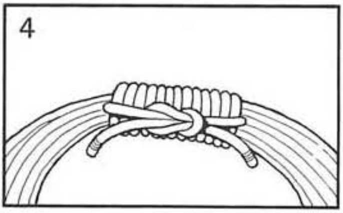

FOR LONGER ROPES

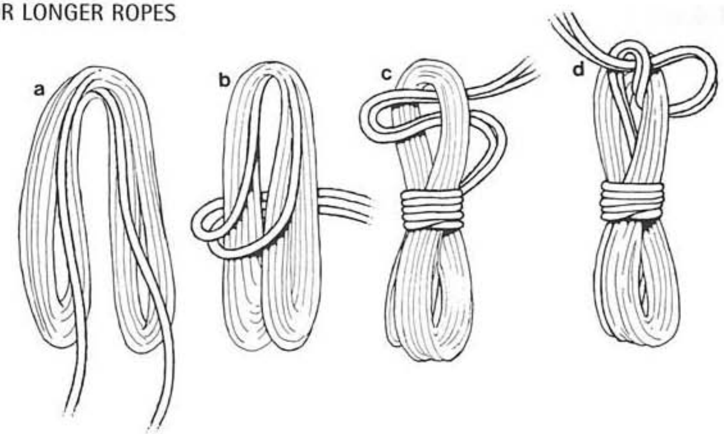

If you wish to carry long ropes over your shoulder or suspended from a belt or from your pack, form a skein.

Loop the rope backwards and forwards over your arm, letting it hang down about 35–60cm (18–24in) long. Leave the ends free (a).

Take both ends together and wrap them several times around the skein (b). Make a loop and take this through the top part of the skein (c) and, finally, pass the ends through this loop (d).

Now tie off on to your pack with a reef knot.

Throwing a rope

It is easier to throw a coil of rope than to attempt to sling a loose end – whether you are throwing upwards or outwards – and it helps to split the coil so that it doesn't tangle. Have a large knot or weight on the throwing end. Make sure that you keep hold of the other end!

Think about the anchored end and what will happen to it when the other end reaches its target. If throwing a lifeline, for example to a fast-moving raft on water, are you going to be pulled into the water yourself? Anchor the end to a tree or weight. Always overthrow a lifeline so that the recipient stands a good chance of catching part of the rope, even if they miss the end.

Coil half the rope onto the fingers and palm of the right hand, then raise the index finger and coil the remainder onto the other fingers only. Pass the second coil back to the left hand.

As you throw, release the right-hand coil a split second before the left. Anchor your end if you think there will be sudden strain on it and your position is precarious.

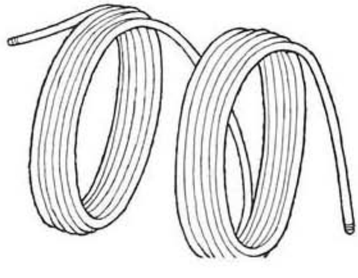

FOR A LONG THROW

Tie a suitable missile to the end of the rope (a). Coil the rope carefully on the ground or loop it loosely over the other hand so that it will pay out freely as you throw the missile.

Don't risk loosing your end of the rope. Tie that to an anchor, a heavy stone for instance. Use a killick hitch (shown later).

If throwing a weighted rope over a branch keep out of its path as it swings back towards the throwing point! If throwing a lifeline don't knock out the person that you are trying to help!

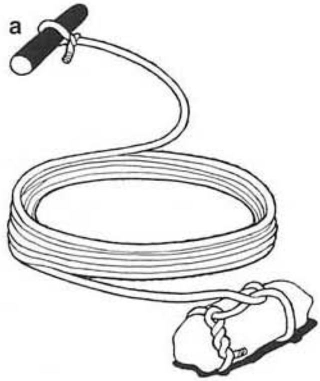

ROPE-MAKING

Vines, grasses, rushes, barks, palms and animal hairs can all be used to make rope or line. The tendons from animals' legs also make good strings, but they tend to dry hard (very useful for binding on arrow and spear heads).

The stems of nettles make first-class ropes and those of honeysuckle can be twisted together to make light lashings. The stronger the fibre, the stronger the rope. Some stiff fibres can be made flexible by steaming or by warming.

While pliable vines and other long plant stems can often be used, as they are, for short-term purposes, they may become brittle as they dry out. A rope made from plant fibres twisted ('spun') or plaited together will be more durable.

Sources of fibres

- **Nettles** (*Urtica dioica*) are an excellent source of fibres but require preparation.

Choose the oldest available plants and those with the longest stems. Soak them in water for 24 hours, then lay them on the ground and pound them with a smooth stone. This will shred the outer surface exposing the fibrous centre. Tease and comb to remove fleshy matter. Hang to dry.

When dry, remove and discard the outer layer. 'Spin' fibres into long threads. Plaiting or twisting together to make a strong rope.

- **Palms** usually provide a good fibre. Leaves, trunks and stalks can all be used. The husk of coconut is used commercially to make ropes and matting.

- **Dogbane** (*Appocynam cannaninum*) stems also provide very good fibre, with which it is easy to work.

- **Barks** Willow bark especially produces very good fibre. Use the new growth from young trees. The dead inner bark of fallen trees and tree branches should not be overlooked. But if the tree has been down too long it may have decayed too much, so test it for strength.

- **Roots** The surface roots of many trees make good lashings. Those that run just under, or even on the surface, are often pliable and strong. The roots of the spruce are very strong. The Indians of North America used them to sew birch bark together to make canoes.

- **Leaves** Plants such as those of the lily family, especially aloes, have very fibrous leaves. Test by tearing one apart. If it separates into stringy layers it can provide fibres to make into rope. Soak to remove the fleshy parts.

- **Rushes, sedges and grasses** should be used when still green. Pick the longest specimens available.

- **Animal tendons** are useful for tying one thing to another. They must be used wet.

SPLITTING CANES

Bamboo, rattan, and other types of canes, vines and bark all need to be split to be used for any kind of rope-making. If you try to pull away thin strips, these tend to run away to nothing. To avoid this problem pull on the thick part to separate it from the thin. It saves both time and energy.

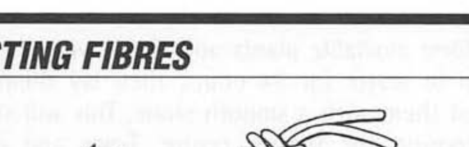

TESTING FIBRES

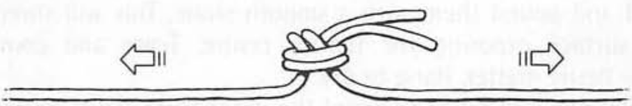

Tie two lengths together using an overhand knot. Try pulling it apart, using a reasonable amount of strength. If it snaps the fibre is too brittle. If it is too smooth, it will slip apart. Suitable fibre will 'bite' and hold together.

Plaiting rope

An easy method for the less experienced is to twist and plait strands. If you make three thin plaits, these can then be plaited together again for a thicker, stronger rope. If you are lengthening the strands as you plait, stagger the places at which you feed in new fibres.

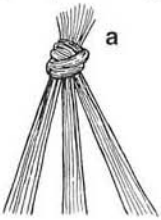

Take a bundle of fibres, tie the ends together, anchor it firmly and split into three separate strands (a). Bring the left strand into the centre (b), then the right over it (c). Then bring what is now the left strand to the centre (d) and so on (e–f). Keep twisting the strands and keep the plaiting as tight and even as you can make it.

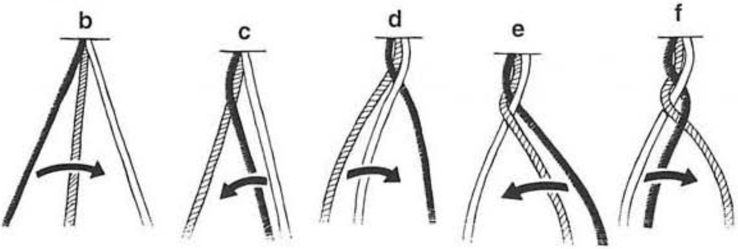

SPINNING A ROPE

Twist fibres together (shown here clockwise, but what is important is to keep to the same direction). Feed in lengths of new fibres as you go so that their ends are staggered.

When you have produced three lengths of fibre, anchor all three at one end and continue to twist each of them until quite tight. Temporarily fastening a toggle to the end of each will make twisting easier.

Now draw all three strands together and twist all three anti-clockwise – the opposite direction.

Continue to add and twist until you have produced the amount of rope you need. You will need to secure a completed section in a cleft stick to keep it tight as you work. Wrap the rope around a tree trunk, to keep the working length short.

To make a thicker rope repeat the process with three 'ropes' you have already made or plait three simple ropes together.

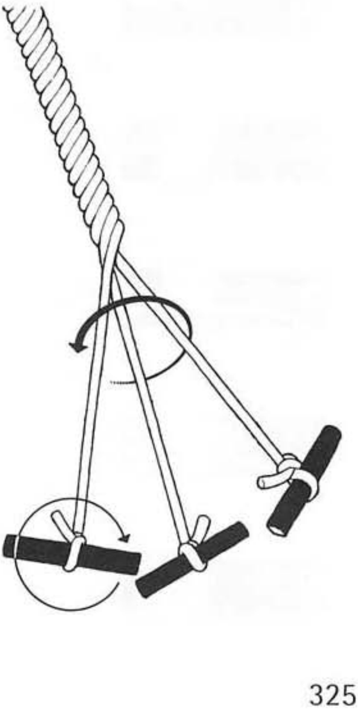

REMEMBER: When making rope try to keep the thickness of the strands equal and even along their lengths. It is where a lay has a thin section that the rope is most likely to break.

Whipping ropes

The end of a rope must be secured in some way so that it does not unravel. To prevent the strands from fraying, bind the rope with twine.

Good binding, or 'whipping', MUST be tight and neat to be effective. If it is too slack it will work loose or fall off. It is difficult to make good whippings with thick cord and very thin is prone to slip. Experience will enable you to match the thickness to the job.

Use the whipping technique to add a comfortable grip to the handles of axes and parangs or, thicker, to replace the handle of a knife.

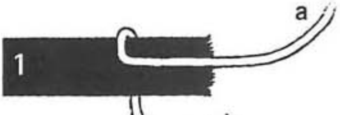

1 *Lay a length of twine along the side of the rope, leaving its end (a) projecting a hand's length beyond the rope's end.*

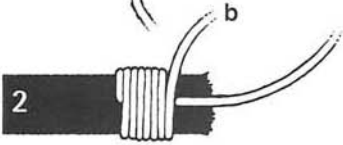

2 *Whip the twine (b) around the rope, working towards the end and gradually covering the piece you have laid along it.*

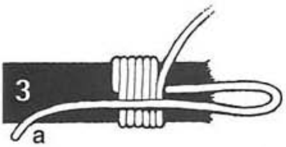

3 *Now form the loose end of the twine (a) into a loop and lay it back along the whipped section.*

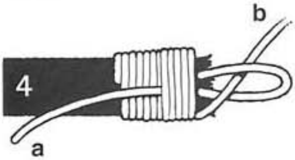

4 *Carry on with the whipping covering the loop until you have nearly reached the end of the rope.*

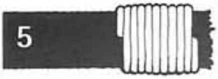

5 *Now pass the end (b) whipping through the loop and pull the short end (a) tight. Trim off ends neatly.*

KNOTS

There is a knot for every job and it is important to select the right one for the task in hand. You never know when you may need to tie a knot so learn their uses, and how to tie each one – well enough to tie them in the dark and under all kinds of conditions. Learn to untie them too – the only thing that is worse than tying a knot that comes undone is a knot that CANNOT be undone at a crucial moment.

In the instructions for individual knots that follow the end of the rope or cord being used to tie the knot is referred to as the 'live end' to distinguish it from the other end of the rope, or 'standing part'.

Reef knot

Also known as the square knot, this is perhaps the best known of all knots. It is used for tying rope of the same thickness and will hold firm even under strain, yet may be fairly easily untied.

It is NOT reliable for ropes of different diameters, nor should it be used with nylon – it will slip.

Reef knots can be tied in other materials – it is a good knot to use in first aid. It will lie flat against the patient.

Pass the RIGHT end over the LEFT (a) and then under it (b).

Then take the LEFT over the RIGHT (c) and under it (d).

Check it – the two loops should slide on each other. If you have crossed the wrong way you will have a 'granny', which may not hold and is difficult to untie if subjected to strain.

Tighten by pulling both the strands on each side, or just the live ends, but be sure it tightens properly (e).

To be doubly sure, finish off the knot by making a half-hitch with the live ends on either side of the knot (f).

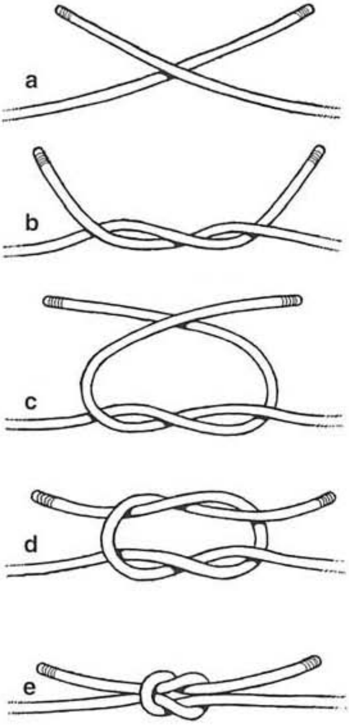

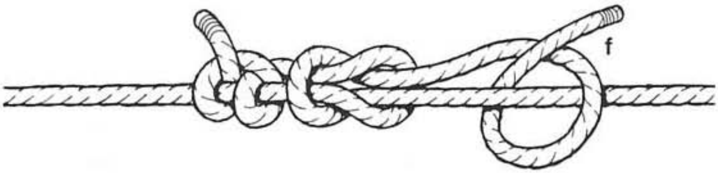

SIMPLE KNOTS

These are some of the simplest knots. They are quickly made, and studying them will help you understand some of the more complicated knots that follow.

OVERHAND KNOT

The simplest of all knots. Make a loop and pass the live end back through it. It has little purpose on its own except to make an end-stop on a rope, but it is part of many other knots.

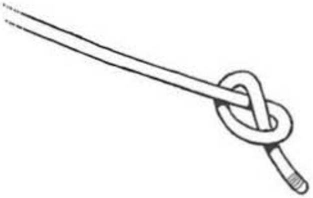

OVERHAND LOOP

A very quickly made knot for throwing over a projection. The loop is fixed and cannot be tightened so the projection must point away from the direction of strain. Double the end of the rope and tie an overhand knot with the loop.

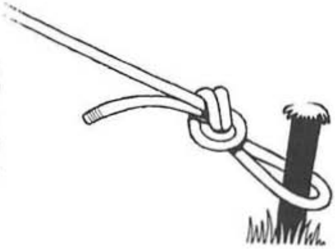

FIGURE-OF-EIGHT

This makes a much more effective end-stop than the overhand knot. Make a loop. Carry the live end first behind, then around, the standing part. Bring it forward through the loop.

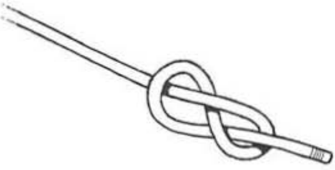

FIGURE-OF-EIGHT LOOP

More secure than the overhand loop, this is made in the same way as the figure-of-eight, but with the line doubled, using the loop as the live end. It can be used over a spike anchor for a belaying rope.

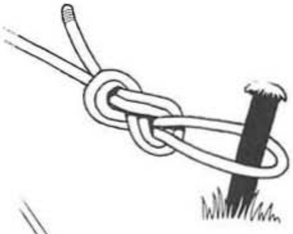

REWOVEN FIGURE-OF-EIGHT

A useful anchoring knot where the top end of a projection is out of reach. Make a loose figure-of-eight along the rope. Pass the live end around the anchor and feed it back around the figure-of-eight, following exactly. Ease tight.

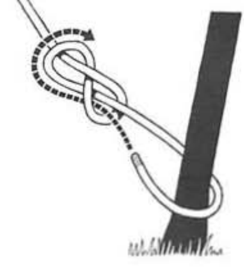

JOINING ROPES

SHEET BEND

Used for joining ropes of the same or different thicknesses. It can be more effective than the reef knot with those of equal thickness. It is ideal for joining different materials, especially wet or frozen ropes.

Simple to tie, using up little of the rope length and swiftly unknotted if it has not been subjected to strain. It never slips if correctly made and strain is not erratic.

1 *Make a loop in one rope. Take the live end of the other (a) right around behind the loop to the front, where it is carried under itself and then tucked up through the loop.*
2 *Draw it tight and ease into shape as strain is increased.*

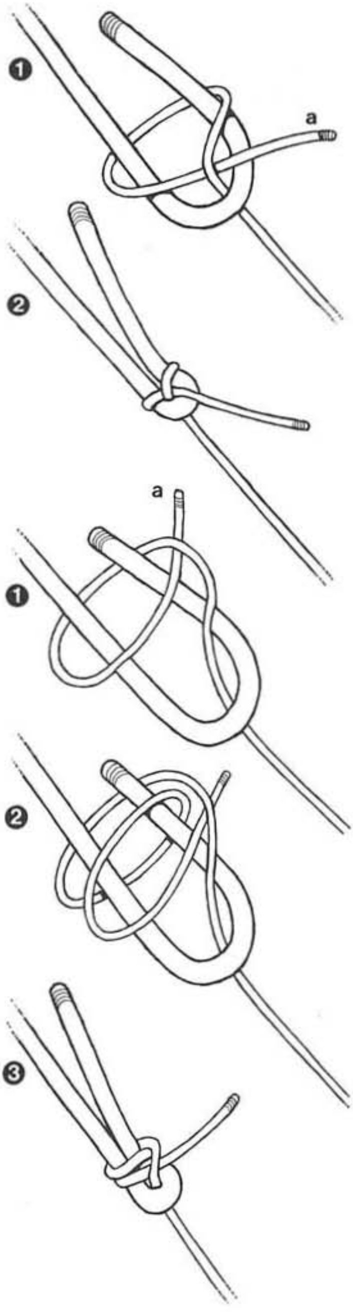

DOUBLE SHEET BEND

A knot that is even more secure than the sheet bend. It is useful with wet ropes, especially if they are of very different thickness, but provides a good strong join in ropes of even thickness too. It is useful where strain is not constant and an ordinary sheet bend may slip.

1 *Make a loop in the thicker rope. Take the live end of the thinner rope (a) through the loop, beneath the thicker live end and then forward on the outside of the loop and right around it. Bring the thin live end back between itself and the outside of the thick loop.*
2 *Take the thin line end completely around the loop again and back through the same place on the outside of the thick loop.*
3 *Draw it tight and ease into shape.*

If not tightened these knots tend to work loose. Do NOT use with smooth materials such as nylon fishing line.

FISHERMAN'S KNOT

A useful knot for joining together springy materials such as vines or wire. It is good for wet or slippery lines and particularly suitable for joining gut fishing line – soak the gut first to make it pliable. Very secure, it will hold well with thin lines but is very difficult to untie. Use when you do not trust the reef knot or sheet bend. It is NOT recommended for bulky ropes or nylon line.

1 *Lay lines beside each other, the ends in opposite directions. Carry the live end of one line around the other and make a simple overhand knot.*

2 *Repeat with the live end of the other line.*

3 *Partially tighten the knots and slide them towards each other. Ease them to rest well against one another, completing the tightening process.*

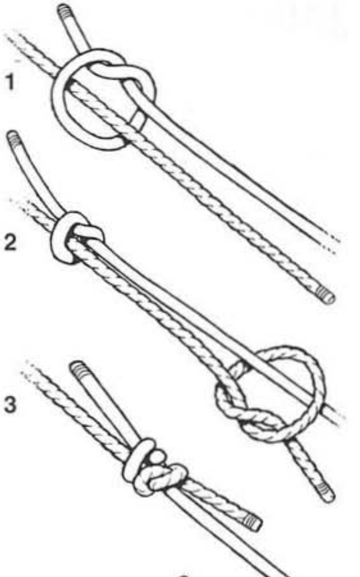

DOUBLE FISHERMAN'S

This is a stronger version of the fisherman's knot. It should NOT be used for nylon fishing lines, nylon ropes, or bulky ropes.

1 *Carry the live end of one line around the other, then around both.*

2 *Carry the live end back through the two loops you have just made.*

3 *Repeat the pattern with the end of the other line.*

4 *Slide the two knots together and tighten them, easing them to rest well against each other. Apply strain gradually.*

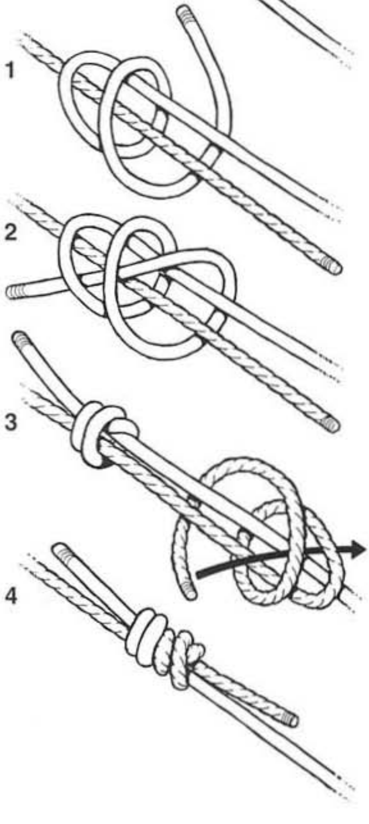

TAPE KNOT

A useful knot for joining flat materials such as leather or webbing straps, tape and possibly even for joining sheets or other fabrics when improvising an escape rope. Experiment!

1 *Make an overhand knot in the end of one 'tape'. Do not pull it tight.*

2 *Feed the other tape through it so that it follows exactly the shape of the first knot.*

3 *The live ends should be well clear of the knot so that they will not slip back when you tighten it.*

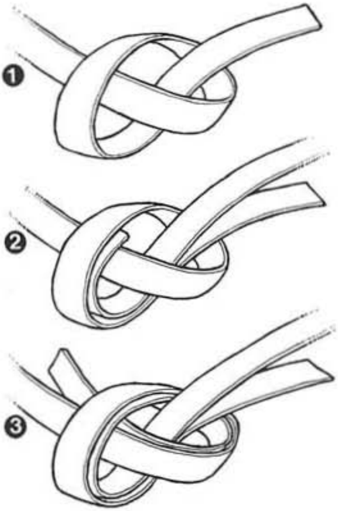

LOOP-MAKING

BOWLINE

This quickly tied knot makes a loop that will neither tighten nor slip under strain. It is used in the end of a lifeline or wherever such a fixed loop is needed.

1 *Make a small loop a little way along the rope.*

2 *Bring the live end up through it, around the standing part and back down through the loop.*

3 *Pull on the live end to tighten, easing the knot into shape. Finish off with a half-hitch.*

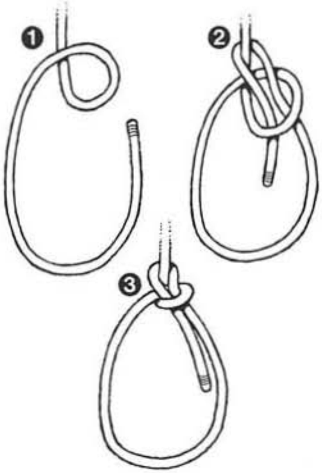

RUNNING BOWLINE

Use in any situation requiring a loop which will tighten easily. Make a small bowline and pass the long end of the rope through the loop.

Never tie a running bowline around the waist, particularly when climbing. It acts like a hangman's noose and could kill.

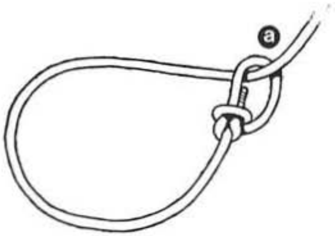

TRIPLE BOWLINE

Another bowline, made with a double line. Form a loop, pass the doubled live end through the loop, behind the standing part and back through the loop. This produces three loops which can be used for equipment haulage, or as a sit-sling or lifting-harness with one loop around each thigh and the other around the chest.

It takes some practise to get the proportions right – so learn to tie the triple bowline BEFORE you need to use it.

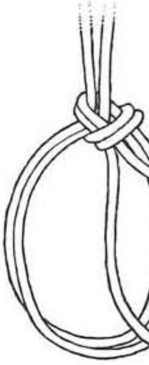

BOWLINE-ON-THE-BIGHT

This is useful to support or for lifting anyone from a crevasse or elsewhere from which they cannot climb out. Make it with a doubled line, producing two loops which will neither tighten nor jam. It forms a kind of bosun's chair, one loop fitting around the buttocks, the other around the upper body. As with the triple bowline, practise this knot BEFORE you need to know how to use it.

You may find many uses for fixed double loops. House- and ship-painters have been known to support their hanging platforms with a knot like this at each end. If the platform has a pole projecting at each corner, the loops of the bowline-on-the-bight can be slipped over. Notch the poles to prevent the ropes slipping off.

1 *Using the doubled line, form a loop and pass the live end through it.*

2 *Bring this end down (a) and over the end (b) of the larger double loop now formed. Ease it back up to behind the standing part (c). Pull on the large double loop to tighten (d).*

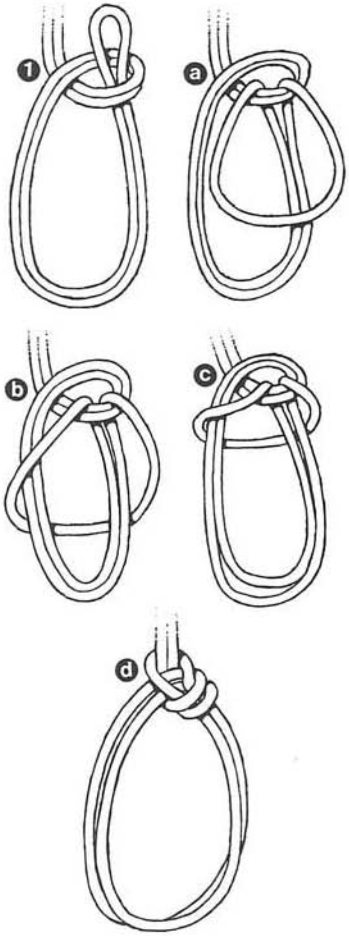

MANHARNESS HITCH

Also known as the Butterfly knot or Artillery knot, this makes a non-slip loop. It has the advantage that it can be made along the length of the rope, but does not require access to an end. Several loops could be put on a rope for harnessing people to pull together in haulage or raising a weight. Also a good way of preparing a rope for climbing. Toes and wrists can be put into the loops to carry the weight so that a rest can be taken when tired.

1 *Make a loop in the rope.*

2 *Allow the left side of the rope to cross over the loop.*

3 *Twist the loop.*

4 *Pass it over the left part of the rope and through the upper part of the original loop.*

5 *Pull the knot gently into shape, ease tight and test it carefully.*

NOTE: *If this knot is not eased tight correctly it is possible to end up with a slipping loop.*

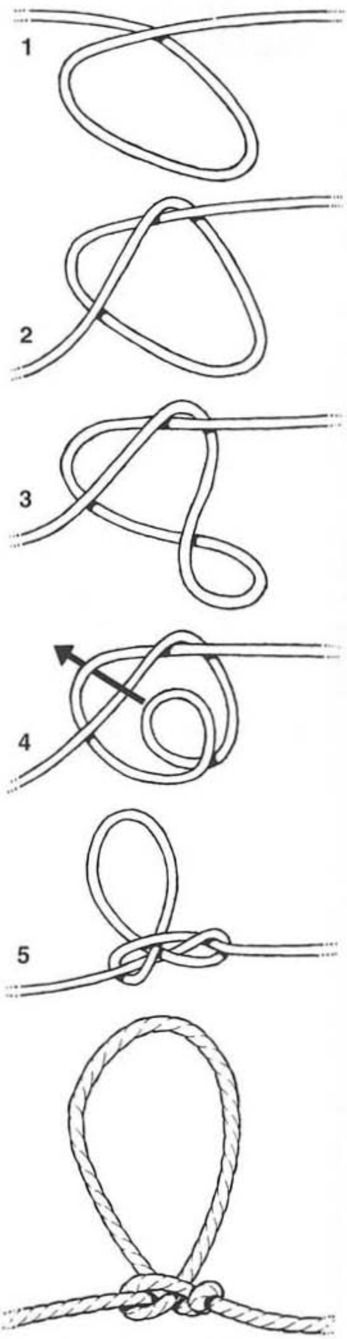

Different ways of making this knot may be found where the loop is not twisted at 3. The final strength of the loop does not appear to be affected either by making this twist or not, nor if the twist should straighten out in use.

LADDERS

A ladder can be made by simply tying as many manharness hitches in a rope as you need for hand and foot holds. It could also be made with rungs, using strong sticks or pieces of wreckage.

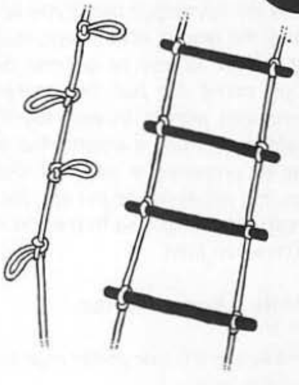

Use two ropes or a long rope, doubled, with manharness hitches placed equally along both sides to make a rope ladder. Pass sticks through the corresponding loops, as you make the loops and ease tight to hold the sticks firmly. Allow the sticks to project a reasonable distance on either side of the ropes for safety and test each for strength.

LADDER OF KNOTS

A series of overhand knots tied at intervals along a smooth rope will make climbing it much easier. There is a fast way of making these, once you have the knack.

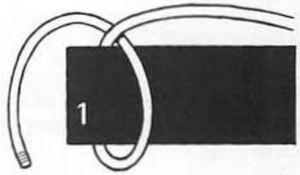

1 *Leaving a reasonably long free end, make a half-hitch near the end of a short piece of branch or log.*

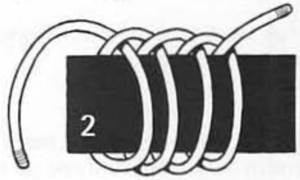

2 *Continue making loose half-hitches along the log – the diameter of which will fix the spacing of the knots.*

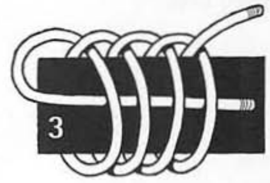

3 *Pass the start end back through all the loops and then slide them all off the end of the log.*

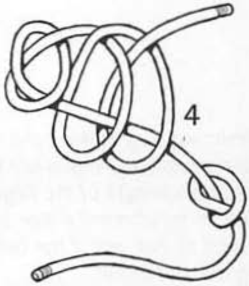

4 *As each turn of rope comes through the centre of the half-hitch loops to the other end, shape and tighten each knot.*

If you know the length of your rope you can estimate the number of knots you need. With a half-hitch for every knot, choose a thickness of log to allow the required number of turns – and therefore the required number of knots.

HONDA KNOT

This is another knot that makes a free-running noose – but this one gives a clear circular loop suitable for throwing – it is a lassoo.

You may find a use for this lassoo in many ways but, in a survival situation, if you have only one length of rope, it would be unwise to spend much time and energy trying to catch animals in this way. It takes a lot of practice. When you really need the rope it may be worn and damaged from your lassooing efforts. However, it is worth making one to experiment with so that you have the skill already, should you need it for actual survival conditions.

a *Start with an overhand knot.*

b *Form a loop further down the rope.*

c *Double the rope into a bight between loop and knot.*

d *Pass the bight through the loop.*

e *Tighten the loop around the bight.*

f *Pass the long end of the rope through the new eye formed by the bight.*

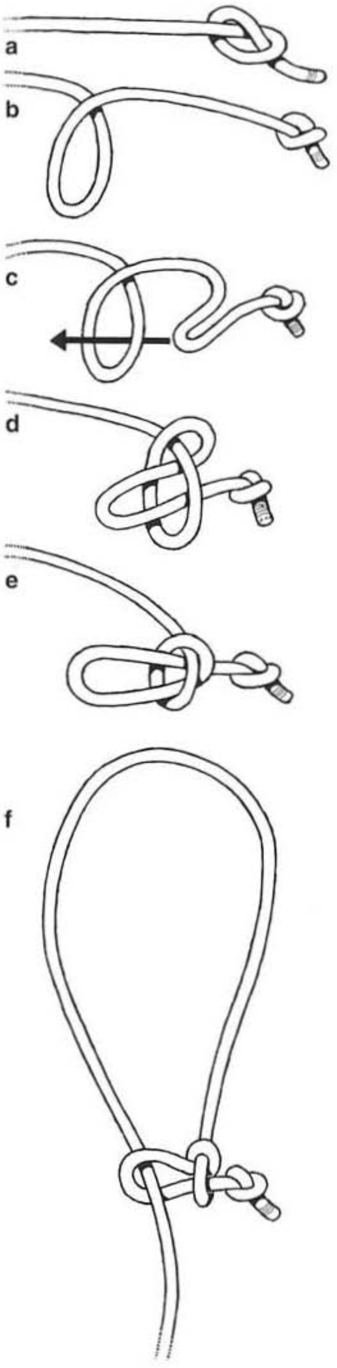

Before trying to use a lassoo on an animal consider its strength against your own. A large animal can – and almost certainly will – run. If you are not holding the rope firmly it may be dragged away and you will lose your meal and your rope. If the rope is anchored to you, YOU may be dragged along and severely injured. Can you make use of a firm anchor to carry the strain? A turn around a tree trunk or a rock may secure what is too powerful for you to restrain on your own.

HITCHES

These are knots for attaching ropes to posts, bars and poles.

ROUND TURN AND TWO HALF-HITCHES

This is the best way to secure a rope to a post. It can take strain from almost any direction.

Carry rope around behind the post, then around again. Bring live end over and back under the standing end and through the loop thus formed. Tighten and repeat the half-hitch to make the knot secure.

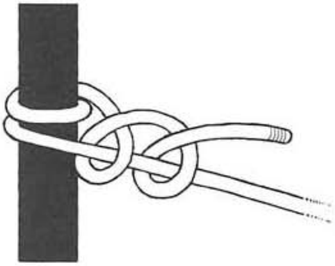

CLOVE HITCH

An effective attachment when strain is perpendicular to the horizontal. It is not so good when strain comes at an angle or the direction of strain is erratic – this could loosen the clove hitch.

1 *Pass the live end over and around the bar.*

2 *Bring it across itself and around the bar again.*

3 *Carry the live end up and under itself moving in the opposite direction to the standing end.*

4 *Close up and pull tight.*

It is possible to make a clove hitch in loops on a rope and to put the whole knot ready-made on to the spar – if you have access to the spar end. Many experienced knot-makers do this. Make a loose clove hitch and slip it off the spar. Lay it down and copy the loops formed by the rope. Slip onto the spar and tighten.

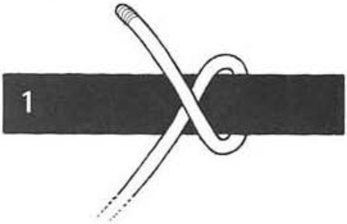

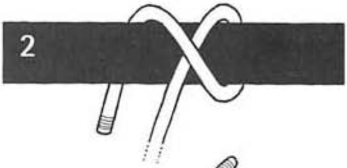

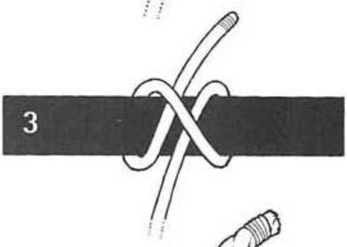

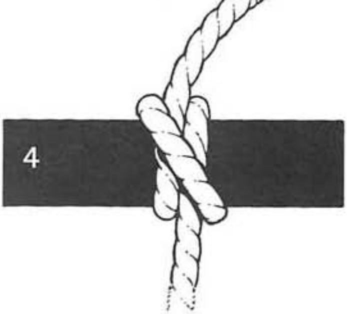

TIMBER HITCH

This knot is used mainly as a start knot for lashings, but can also be used for hoisting and for dragging or towing heavy logs.

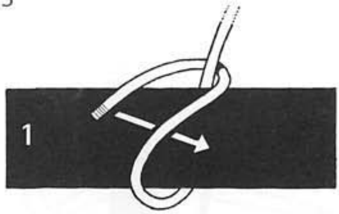

1 *Bring the live end around the bar and loosely around the standing end.*

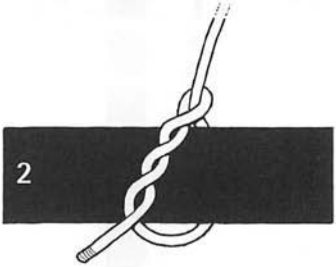

2 *Carry it forward and tuck it beneath the rope encircling the bar. Twist it around as many times as comfortably fit. Tighten the knot by gently pulling on the standing end until a firm grip is achieved.*

KILLICK HITCH

Also known as the Anchor hitch – use it for securing a line to an anchoring weight. It can be used to secure an actual anchor for use in water or to hold back one end of a throwing line.

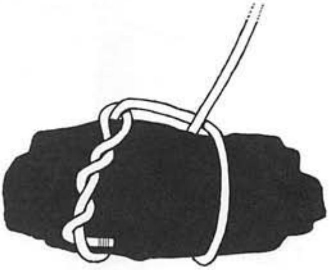

Make a timber hitch around one end of the rock or weight and tighten. Carry the line along the weight and make a half-hitch.

MARLIN SPIKE HITCH

This is an instant, but temporary, knot for securing a mooring line to a post, or for dragging over the top of any upright peg or pole. It is particularly useful when tightening lashings. By temporarily attaching a short stout stick to the line it is possible to gain extra purchase on the line to administer a firmer pull.

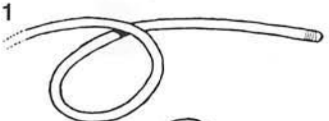

1 *Form a loop in the rope – study the drawing carefully.*

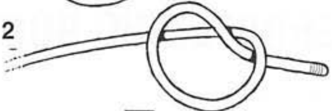

2 *Bring one side of the loop back up over the standing end.*

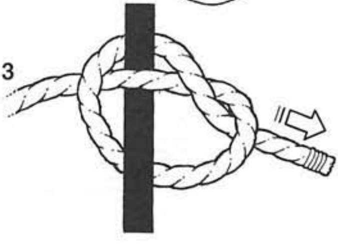

3 *Drop this over the pole – the pole coming between the extended loop and the standing part. Pull the live end to tighten.*

QUICK-RELEASE KNOT

Variously known as the Highwayman's hitch, Thief hitch and Thief knot. This knot is secure but will come untied with a single sharp tug on the live end. It is recommended for temporarily anchoring lines while working – or for situations which need a quick release.

1 *Carry a bight around a post or rail.*

2 *Bring a bight from the standing end through the first bight.*

3 *Form the live end into a further bight and push the doubled end through the loop of the second bight. Pull on the standing end to tighten the knot.*

4 *To release pull sharply on the live end.*

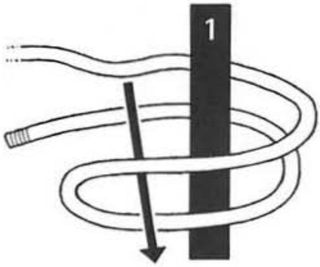

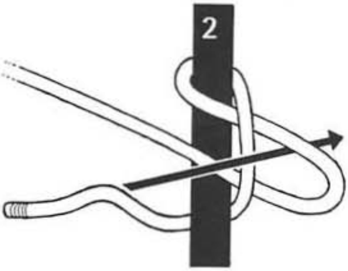

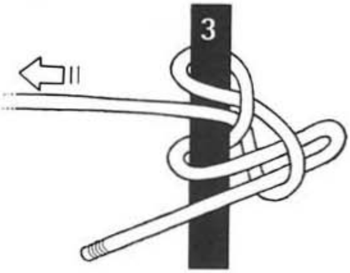

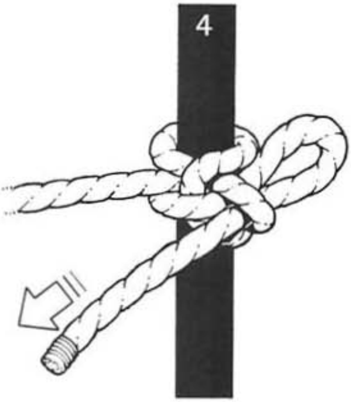

SHORTENING ROPE

SHEEPSHANK

Treble the line. Form half-hitches in the outer lengths and slip them over the adjoining bends. Or, instead of half-hitches, when a loop is formed in the standing part, pull a bight through it and slip this over the bend in the rope. Tighten as you gradually increase tension.

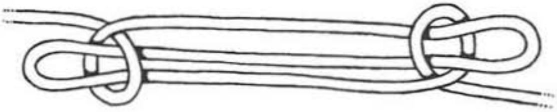

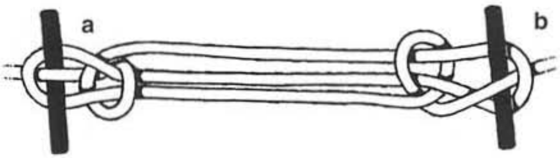

Make a sheepshank more secure by passing a stick through the bend and behind the standing part (a). Or, if you have access to the rope's end, pass that through the bight (b). A stick would make this more secure.

Never cut a rope unnecessarily – you never know when you may be glad of its full length. A knotted rope has only half the strength of a continuous one. Use the sheepshank to shorten it or to exclude a damaged or weakened section.

SECURING LOADS

WAKOS TRANSPORT KNOT

A knot invaluable for securing a high load to boat, raft, sledge, etc, or for tying down a roof. Maximum purchase is achieved by pulling down with all your weight and finally securing with two half-hitches. If it comes loose undo the hitches, retighten and secure. This knot can also be used for a line across a river or chasm which needs tightening from time to time.

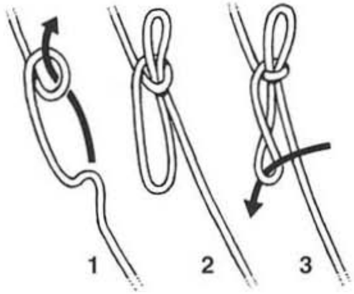

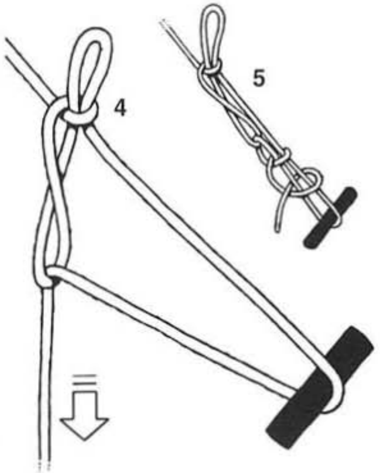

1 *Make a loop in the rope. Further down, towards the end of the rope, make a bight.*

2 *Pass the bight through the loop.*

3 *Make a twist in the new lower loop. Pass the end of the rope around securing point and up through this twist.*

4 *Pull on end to tighten.*

5 *With the end make two half-hitches around lower ropes to secure. Undo these to adjust and retighten.*

PRUSIK KNOT

A knot that makes a sliding loop, especially useful when attached along a climbing rope. It will not slip under tension, but will slide along the rope when tension is released. Also useful for ropes that need retensioning from time to time, such as tent guylines.

A pair of prusik knots along a rope provide hand and foot holds for climbing or for swinging along a horizontal line. They are slid along the main rope as you proceed.

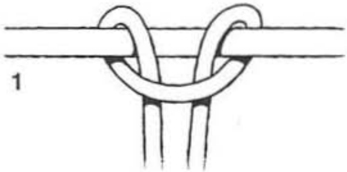

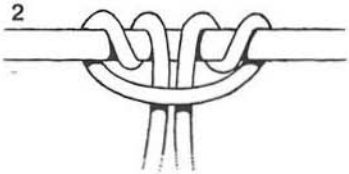

1 *Pass a bight around the main rope and pull the ends through. Keep this loose.*

2 *Take the ends over again and back down through the loop. Ease tight. Do not allow the circuits to overlap.*

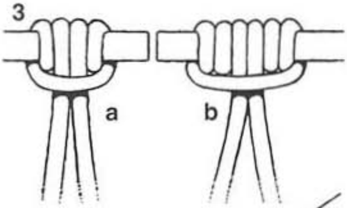

3 *This gives the appearance of four turns on the main rope (a). Mountaineers sometimes take the ends round again and back through the loop to give the appearance of six turns on the main rope (b).*

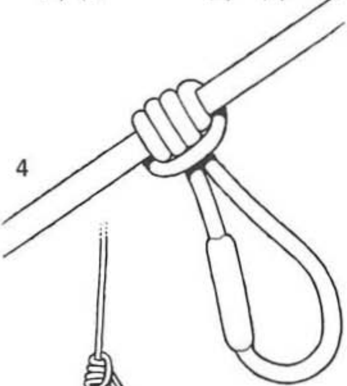

4 *The prusik knot can be made using a fixed spliced loop – in which case pass the bight over the main rope and back through itself, and repeat.*

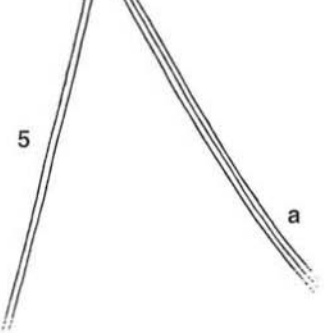

5 *For use as a tensioning line attach along the guy rope etc and secure ends (a) to an anchor.*

Note: *When used for climbing, or travelling along a rope, a spliced loop is safest (as 4). If you have no spliced loop, join the ends after the knot is made. Test joins rigorously before relying upon them.*

LASHINGS

Methods of lashing spars, logs, poles etc together differ according to the position of the components. Learn these techniques. They will be invaluable in making shelters, rafts and other structures.

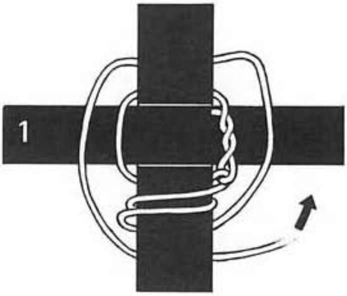

SQUARE LASHING

This is for lashing spars which cross. Most effective when they do so at right-angles.

1 *Make a timber hitch carrying the line alternately above and below both spars in a complete circuit before securing it. Then carry the rope over and under both spars in an anti-clockwise direction.*

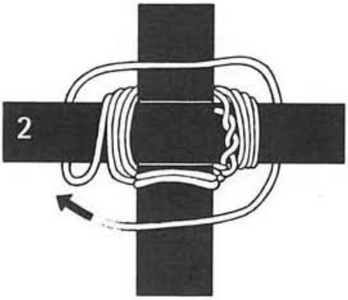

2 *After three or four circuits make a full turn around a spar and circuit in the opposite direction.*

3 *Complete the circuits with a half-hitch around one spar and secure with a clove hitch on a spar at right-angles.*

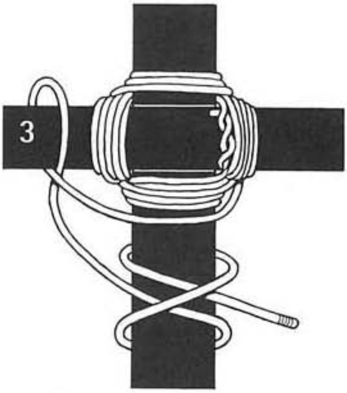

ROUND LASHING

This is for lashing spars alongside each other or extending the length of a spar.

Begin with a clove hitch around both spars (a), then bind rope around them. Finish knot with a clove hitch at the other end (b). Force a wedge under the lashings to make them really tight. If the spars are vertical bang the wedge in downwards.

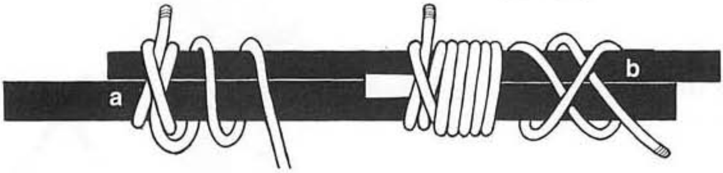

DIAGONAL LASHING

An alternative to square lashing which is more effective when spars do not cross at right-angles, or more especially when the spars are under strain and have to be pulled towards one another for tying.

1 Begin with a timber hitch around both spars, placed diagonally.

2 Frap both spars with a few turns of rope over the timber hitch, then make a full turn under the bottom spar.

3 Frap across the other diagonal, then bring the rope back over one spar and make two or three circuits of the spars above the upper spar and below the lower.

4 Finish with a clove hitch on a convenient spar.

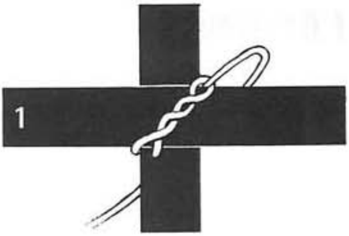

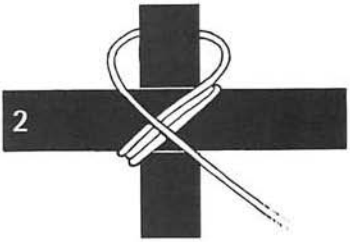

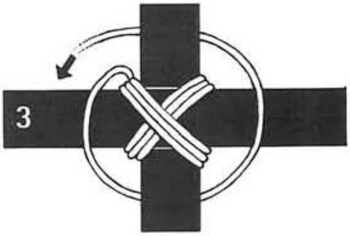

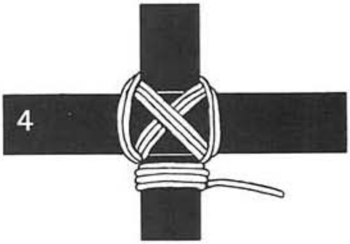

SHEAR LASHING

For tying the ends of two spars at an angle, when making an A-frame, for example.

Begin with a clove hitch (a) around one spar. Bind around both spars – this binding should not be very tight. Bring rope between the spars and frap a couple of times around the binding. Finish with a clove hitch around the other spar (b). Tighten by opening up the shears (c).

A similar method can be used around three poles to make a tripod. Make turns around all three legs and frappings in the two gaps. The feet of A-frames and tripods should be anchored to stop them spreading.

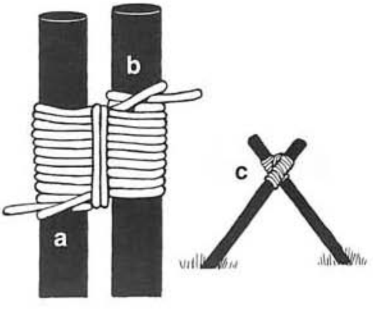

NET-MAKING

Nets can be made either by making knots along pre-cut lengths of line or by 'knitting' mesh row by row. They are not only useful for fishing. A gill net can also be hung between trees to catch birds and a purse net, made from twine, can be placed over animal burrows. Use the same techniques to make a hammock from strong twine.

Gill net

Make this from parachute cords or from two thicknesses of twine. Parachute cord consists of an inner core of fine line within an outer core. Pull the fine inner line out and cut it into manageable and equal lengths (or cut lengths of thinner string). Their length will determine the depth of your net, which will be about 3/8 that of the length of line.

Decide how wide you want your net and set two poles that distance apart. Tie a length of parachute cord outer (or thicker twine) between the two. Cut a piece of wood about 3.5cm (1¼in) across. Use this as a gauge to space out the thinner vertical threads (inner core).

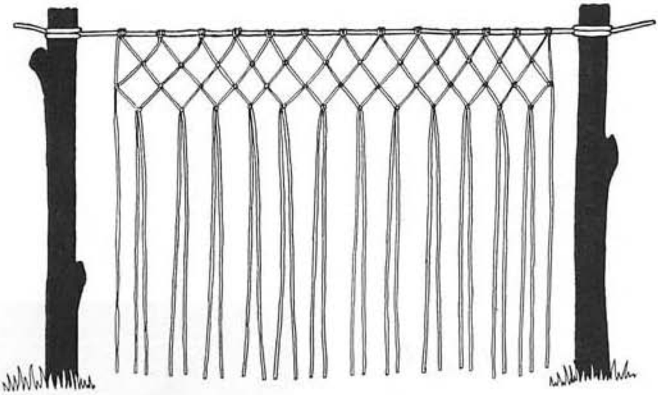

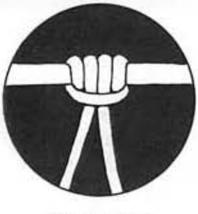

Prusik knot

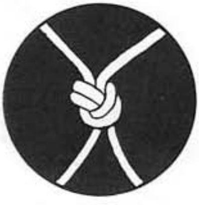

Overhand knot

Fold each length double and use the bight to make a Prusik knot over the top cord and repeat across its length. Slide the Prusik knots along to space them out equally using your gauge.

For the first row, working from left to right; ignore the very first individual strand, but take the second of the pair. Hold it with the first strand of the next pair and tie both together in an overhand knot. Take the remaining strand with the first of the next pair and knot. Continue along the line, using your gauge to control spacing.

Proceed to the next row in the same way but this time include the outside lines to produce a row of diamonds. Continue until the line is used up.

To finish off the bottom, stretch another thicker line across between the supports and tie off all the inners (or thinner strings) in pairs around it. Carry each pair around it twice. Separate the pair and tie off around the pair.

Complete the net by securing the top and bottom lines at each corner of the net so that the net will not slip off the ends. Any surplus can be used for attaching the net to supports and weights to keep it in position when in use.

'KNITTING' A NET

A method suitable for nylon fishing line or any other fine line. You need a horizontal string between posts, a mesh gauge, and a 'needle'.

Make the 'needle' (a) about 15cm long x 2·5cm wide (6 x 1in) from hardwood or bamboo. Make a notch at either end and wind line around the whole needle; or try something more traditional like the lower drawing. The needle must be smooth. The line is gradually unwound as you make the net.

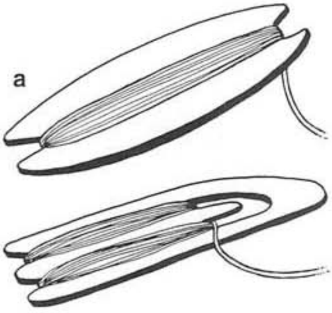

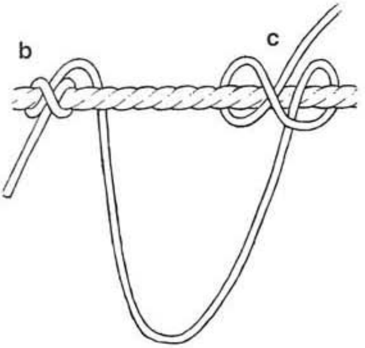

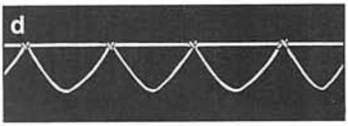

To make the net, tie a top line of required length between uprights. Begin by tying a clove hitch with thinner line (b). Take the needle behind the top line and bring it forward to make another clove hitch (c). Repeat along the line, spacing the knots out with your gauge (d).

When the top row is complete go to the other side of the posts (easier than working backwards) and make the next row. Make each new loop large enough to form a square of the mesh (half-square at each outside edge). Take the needle through the loop of the row above from behind, round the back of the loop and then through the front of the loop it makes (e). Adjust the depth with your gauge before you tighten (f).

Switch sides again and work back in the opposite direction for the next row and continue until the net is the required length.

Tie off the bottom line with another thicker twine using the same knot but keeping the line straight without loops. Leave some free line at both ends. Tie in the ends at the top corners and the net is complete.

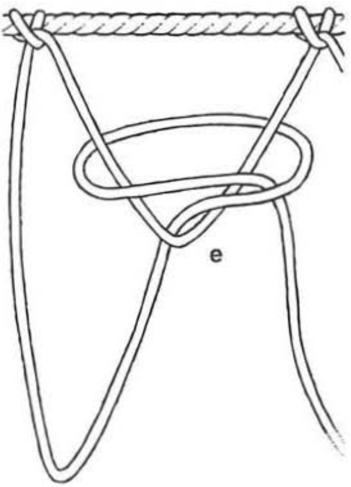

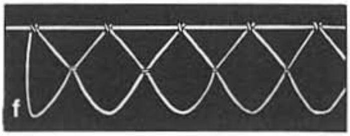

Net hammock

Make a net about 75cm (2½ft) across and wider than your height. Use good strong twine or rope for the top and bottom lines – double twine would be a good idea, the ends have to carry your weight. Leave those ends long enough to suspend the hammock by.

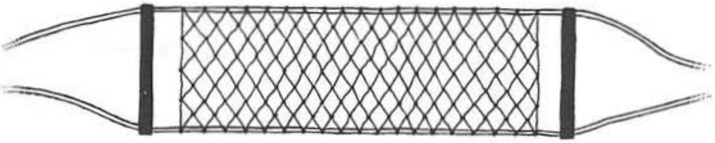

Cut two spacer bars to keep the hammock 'open'. Notch the ends and slip the cords into the notches (a). To simplify hanging the hammock you could tie each pair of end lines to a fixed loop such as a bowline. Then fix one end with a round turn and two half-hitches, the other with a quick-release knot, in case you ever need to leave the hammock in a hurry.

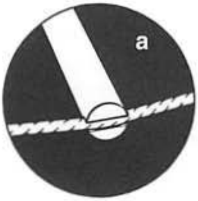

FISHING KNOTS

HOOK ON TO GUT

Turle knot *Soak the gut. Thread it through eye of hook. Make an overhand loop and pass a bight through it (a) to form a simple slip knot (b).*

Pass hook through slip knot (c) and pull tight around shank.

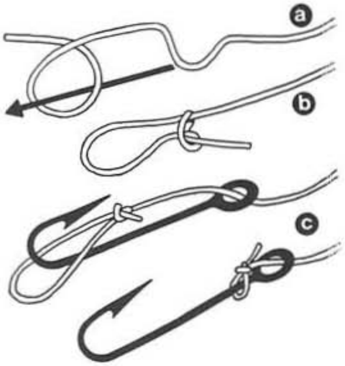

HOOK ON TO NYLON # 1

Half-blood knot *Thread end through eye. Make four turns around standing part. Pass live end through the loop formed next to the hook (d).*

Pull taut and snip off fairly close to end (e).

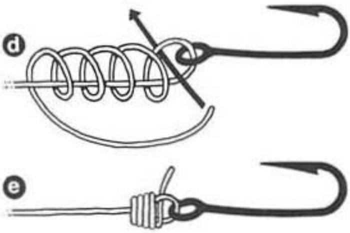

HOOK ON TO NYLON #2

Two-turn turf knot *Thread the hook. Pass the live end around the standing part to form a loop and through it. Twist live end twice around side of loop. Hold the loop and pull the twists tight. Pass the hook through the loop (f).*

Pull on standing part to tighten loop on hook (g).

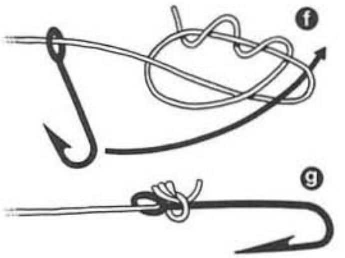

JAM KNOTS

For securing improvised hooks to gut or cord.

With an eye: Thread gut. Make two turns around hook and bring live end up through turns (h). Ease tight and test for strength.

Without an eye: Make loop around lower part of shaft. Make two half-hitches from upper end downwards and pass live end through lower loop (i). Pull on standing part to tighten.

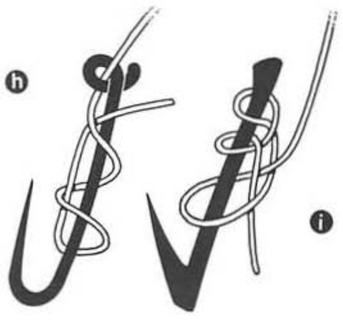

LOOP IN NYLON #1

Double overhand loop *Double the line to make a bight. Tie an overhand in it (a). Twist the end through again (b). Pull tight (c) and snip off end.*

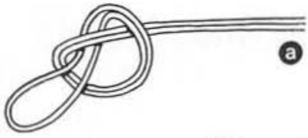

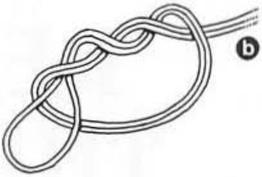

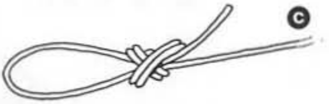

LOOP IN NYLON #2

Blood bight *Form a bight. Twist the end of it back around the standing part (d). Bring end back through new loop (e). Pull tight and snip off loose end.*

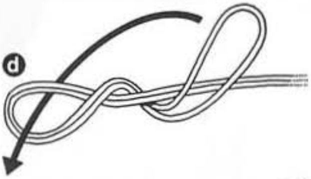

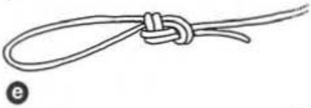

JOINING LOOPS

Can be used in nylon line but will work for different strong materials. A Fisherman's knot is recommended for gut, which probably could not stand the strain of this method.

With free ends: *Pass each line through the other loop (f) and pull tight (g).*

With only one end free: *Make loop on one line. Take the live end of the other line through the loop, around it and back through and then tie off with either of the knots for hooks on to nylon.*

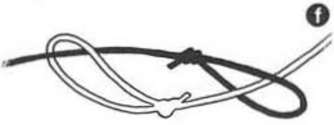

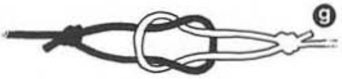

HOW TO BE ECO FRIENDLY

Whatever your situation, and wherever your camp, survival must always take priority and in actual fact a group of survivors will have very little impact on their immediate environment. Having said that you should try not to decimate the area you find yourself in and always try and leave a place better than you found it. There are a number of ways in which you can ensure this happens: only take standing dead wood for a fire, only catch sick, lame, or lazy animals, and if you're cutting saplings you're encouraging growth in any case.

6

READING THE SIGNS

Being able to read and make a map is only the beginning of being able to find your way about. You must learn to interpret the signs found on the ground itself and in the air.

If you do not have a compass, there are a number of ways of creating your own. The sun and stars can provide firm orientation – a variety of methods are available for finding direction in both northern and southern hemispheres.

An ability to anticipate the weather can also be a valuable asset in the wild and simple guidelines for prediction and the interpretation of cloud patterns equip the survivor with more skills.

READING THE SIGNS 350
- Maps 350
- Direction finding 355
- Weather signs 364
- Clouds 366
- Predictions 370

READING THE SIGNS

Before embarking on any expedition you will have learned all you can about the terrain, equipped yourself with maps, if available, and worked out routes. Memorize the lie of the land, the direction in which rivers flow, the high ground, the prominent features, the prevailing winds, the weather patterns to expect and any known hazards, check the phase of the moon and times of first and last light – all of which will be invaluable knowledge if you find yourself in difficulties. In a case of accident you may find yourself in a totally unknown territory and have to find out everything about your location from the land itself.

In choosing a camp site, tracing water and finding the other necessities for survival you will need to interpret the surrounding countryside – the other side of a hill may offer quite different conditions – and if you decide not to stay put you will have to interpret both the general geography and the particular landscape as you proceed.

MAPS

Choose maps carefully, making sure that they are to a scale that will be useful to you and show helpful information. A very large-scale map that shows every footpath and building will be no use at all if you are driving a thousand miles along a motorway. Everything will be shown in great detail but only a tiny fraction of the journey will appear on one sheet and you would have to pile the car with maps and change from one to another every few miles. On the other hand few motoring maps give much information about the nature of the terrain or show features which would help a walker choose their route. Sailors must be equipped with accurate charts so that they can keep to safe waters. The surface below the sea can be as varied as that above. Flyers will need to know what altitude obstructions are and what turbulences occur near mountains which make it safer to fly higher. From the air the pattern of the land may show plainly but its contours are flattened out; without interpretation, a map looks rather like that to many people.

Maps and terrain

Height cannot be reproduced on flat sheets of paper so altitudes are recorded at regular intervals (usually every 50ft or every 15m according to the measure used) and every point at this height is joined up by a line – the contour line. In most cases these lines join up to form a complete shape, some sort of irregular oval with bulges here and there. If they suddenly stop against another line that means that there is an abrupt change of height – in fact, a cliff or a very steep fall.

The only contour line that you can see in nature is that of sea level along the coast (and even that is not quite true because of tidal variation) but you can imagine the contour lines as the edges of flat disks and that these are ranged equidistantly above each other. If you threw a cloth to rest over them it would link them together in a shape that would be approximately like a hill or other feature. However, you do not have a record of exactly what happens between those contour lines and there will not necessarily be an even slope connecting them. There could be outcrops of rock, hollows – any manner of variation within that 15m or 50ft. From the relative positions of one height to another you could make a pretty good guess as to what the ground surface was like, but you could not be sure. There may therefore be features that, because they fall between the contour lines, make no appearance on your map.

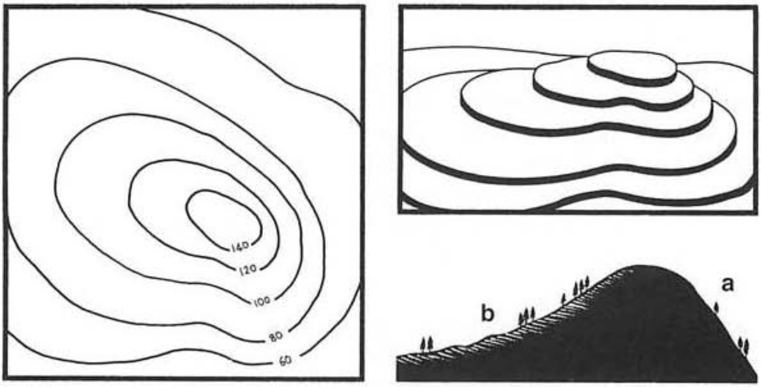

The contour lines on the map represent a series of points at the same distance above sea level and do not record what happens in between. When the contours are closely grouped the change in height is more rapid (a). Conversely, greater spaces between the contour lines indicate gentler slopes (b).

Interpreting maps

Remember that the intervals between the contour lines are the distances between horizontal points at the same theoretical height – not the actual distance on the slope of the ground. They are measured in units that show relative positions and are not to a scale as is the horizontal plotting.

It is a common error to think of a group of contour lines indicating a rise in the ground comparable to the scale of the distance shown between them – but the scale of a typical walkers' map is 1:50,000 and 10m on that would be only .02mm. Contours spaced 5mm apart on the surface of the map would be at a horizontal distances of 250m and the gradient only 1 in 25.

Scale

Before you can begin to use a map you must understand its scale. This may be shown by a scale bar marked with miles or kilometres to the size that they are shown on the map or it may be given as a ratio – 1:50,000 means that every measure on the map represents a distance 50,000 times greater on the ground.

Key

There will usually also be a key to the symbols used within the map to represent natural and manmade features – rivers, roads, buildings, types of woodland or swamps, types of beach. What is shown, and how, will vary greatly. If there is no key on the individual map or on its wrapping make sure that you find out what symbols mean in that map series. Some will be fairly evident: if the map is in colour rivers will almost certainly be in blue, marshes will usually be indicated by stylized tufts of reed.

Not all features can be shown to exact scale. Roads and paths will probably be given standard widths to match the kind of track they are rather than their exact measurements, and streams and rivers will be similarly standardized. The British Ordnance Survey (OS) maps, for instance, show waterways as a single blue line, gradually increasing in width until it represents a width of 8m (27ft) across a stream, whereupon a double line is used, giving you an immediate indication that you have a river at least that wide to cross. There are similar standardizations on all maps. Once you have mastered the way that information is shown, maps will tell you a great deal.

Grids

Maps almost always carry a grid of lines which divides them up. This is either based on degrees of latitude and longitude or a special grid developed by the mapping authority. The advantage of the special grids is that they are usually planned to form squares based on ground measurement which can help you rapidly assess distances. On the British OS maps, for instance, the grid lines are 1km apart and the diagonal across them is 1½km. If you want to find, or report, a position it can be described by a coordinate made up from the line references from two adjoining sides of the map. To anyone using a map with the same grid this will immediately locate the 'box' in which it appears. Dividing the square by eye into further tenths pinpoints the location. This provides an easy way of telling rescuers of your location or of fixing a rendezvous point with them.

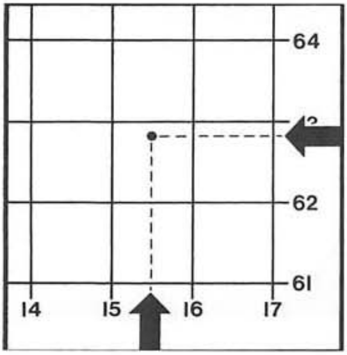

The point marked with a dot can be described as 15.5 x 62.8 using the coordinates from the sides of this grid. This system requires that squares are mentally divided into tenths in each direction. The 'map reference' is normally expressed as six digits: 155628. Any letter area codes on the map should be included.

North on maps

Unless they are lines of longitude, the grid lines on a map are not an indication of north and south, though they may sometimes be close to it. Remember that a compass points not to true north but to magnetic north – and the difference between the two varies both according to where you are in the world and because magnetic north is slowly changing its position. To take accurate bearings you need to know these variations, but even a rough idea of orientation will help you to match you map to the landscape.

If you have an adjustable compass and information on the deviation of it and of your map grid from true north you can carefully match up all of them so that even in poor visibility, or where landscape features are beyond your horizon, you can take accurate bearings and follow them.

Many maps indicate the deviation or the direction of magnetic north. If this is not given you can find it from the North Star or by using the watch method to point to north. Use the Southern Cross in the same way in the Southern Hemisphere to establish south (both are illustrated later).

Local magnetic variation

To find the local magnetic variation, when not recorded on a map, point the compass at the North Star. Note the difference between the pointer and indicated north.

Lining the compass up with the grid lines on the map you can discover their variation, if there is one. If you then propose to walk on magnetic bearings you must remember to compensate for the variation.

Map-reading skills should be learned by anyone planning an expedition and are particularly important in mountainous country. Compasses are available mounted with calibrations, scales and direction markers which make this kind of orientation easier. Check them out and have them explained to you.

In a survival situation you will probably have to manage without such sophistication. If unable to make appropriate corrections, continually check your position against visible features.

Gradients

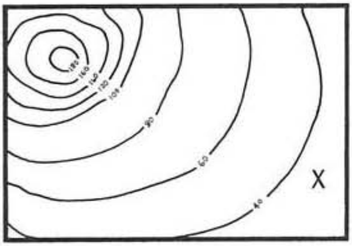

a

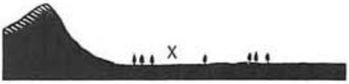

A concave slope (a) (where you can see the top from the bottom) has the higher contours close together.

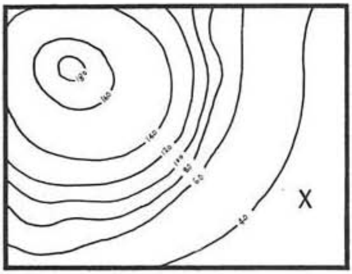

b

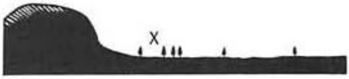

A convex slope (b) (where you cannot see the top from the bottom) has the low contours close together.

Measuring distances

As-the-crow-flies distances can be measured by using any straight edge and matching it up against the scale bar or multiplying by the ratio of the map's scale. Meandering routes can be followed with a piece of thread which can then be straightened out. Gradients can make an appreciable difference to distances and must be allowed for – a gradient of 45 degrees, for instance, will add another 82m to a horizontal map distance of 200m (328ft becomes 725ft).

Your own maps

The survivor may not be lucky enough to have a map and should then set about making one. With a map you will always be able to find your way back to camp – essential if you are going for help for the sick, children or elderly survivors left behind.

It is not possible for you to measure exact contour heights, but you can devise your own system for indicating the contours.

Find the best vantage point and look out over the terrain. Climbing a tree may give a better view. Note the direction of the ridges, count how many you can see. Between each ridge there is probably a stream or river flowing, you cannot be sure – there will be a lot of 'dead ground', territory you cannot see. Make a general map with blank patches and then fill them in as you gain more information from other vantage points and from your explorations on the ground.

Mark anything of interest on your map: watercourses, rocky outcrops, isolated trees, strangely shaped features that will act as landmarks, and areas of different vegetation. You can plot positions of your traps, animal lairs, good places for foraging for food and fuel or finding useful stones for implements. It will be much easier for you or your companions than relying on your memory or vague descriptions.

SURVIVAL SCENARIO

You are navigating your way across unfamiliar terrain – what's the most important thing to watch out for?

One of the most common errors people make is to overestimate how much ground they have covered. It is usually much less. One way around this is to fit the ground where you are to the map rather than the map to where you think you are.

DIRECTION FINDING

The earth's relationship to the rest of the solar system and the position of the stars in the sky help to locate any position on its surface. Its revolution on its axis produces the changes from light to darkness and its orbit around the sun produces the seasons, for the earth is tilted at an angle to the sun and first the north and then the south becomes nearer to it, the closest point traversing from the Tropic of Cancer (23.5°N) to the Tropic of Capricorn (23.5°S), the sun being above

Cancer on 22 June and above Capricorn on 22 December. It is above the Equator on 21 March and 21 September.

The sun rises in the east and sets in the west – but not EXACTLY in the east and west. There is also some seasonal variation. In the Northern Hemisphere, when at its highest point in the sky, the sun will be due south; in the Southern Hemisphere this noonday point will mark due north. The hemisphere will be indicated by the way that shadows move: clockwise in the north, anticlockwise in the south. Shadows can be a guide to both direction and time of day.

SHADOW STICK METHOD #1

On a patch of flat, clear ground place a metre-long (3ft) stick as upright as possible. Note where its shadow falls and mark the tip with a pebble or stick (a). Wait at least 15 minutes and mark the new shadow tip (b). Join the two and you have the directions of east and west (3) – the first mark is west. North-south will be at right angles to this line. This method works at any time of day when there is sunshine and at any latitude. Use it for spot checks as you proceed.

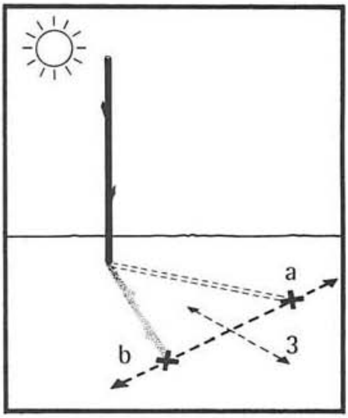

SHADOW STICK METHOD #2

Another, more accurate, method – if you have the time – is to mark the first shadow tip in the morning. Draw a clean arc at exactly this distance from the stick, using the stick as a centre point. As midday approaches the shadow will shrink and move. In the afternoon, as the shadow lengthens again, mark the EXACT spot where it touches the arc. Join the two points to give east and west – west is the morning mark.

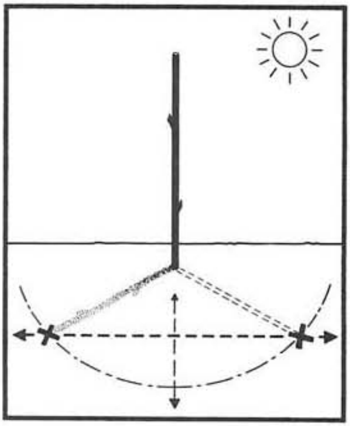

Direction by watch

A traditional watch with two hands can be used to find direction, provided it is set to true local time (without variation for summer daylight saving and ignoring conventional time zones which do not match real time). The nearer the Equator you are the less accurate this method will be, for with the sun almost directly overhead it is very difficult to determine its direction.

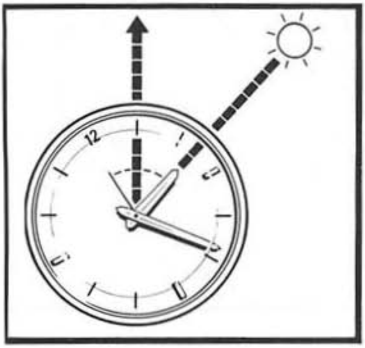

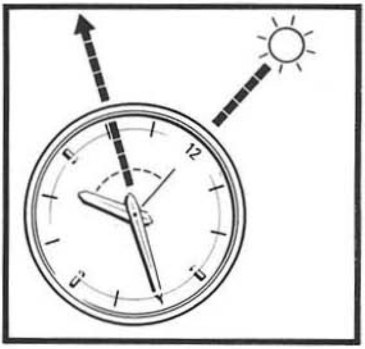

Northern Hemisphere *Hold the watch horizontal. Point the hour hand at the sun. Bisect the angle between the hour hand and the 12 mark to give a north-south line.*

Southern Hemisphere *Hold the watch horizontal. Point 12 towards the sun. A mid-point between 12 and the hour hand will give you the north-south line.*

IMPROVISED COMPASSES

A piece of ferrous metal wire – a sewing needle is ideal – stroked repeatedly IN ONE DIRECTION against silk will become magnetized and can be suspended so that it points north. The magnetism will not be strong and will need regular topping up.

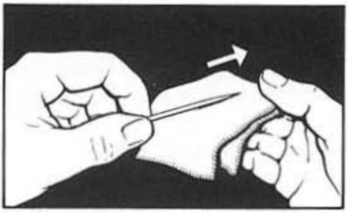

Stroking with a magnet, should you have one, will be much more efficient than using silk – stroke the metal smoothly from one end to the other IN ONE DIRECTION ONLY.

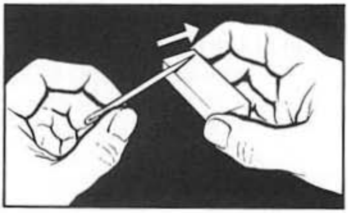

Suspend the needle in a loop of thread, so that it does not affect the balance. Any kinks in or twisting of the thread must be avoided.

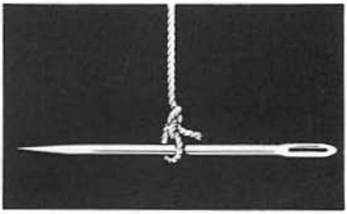

FLOATING NEEDLE

A suspended needle will be easier to handle on the move but in camp or when making a halt a better method is to lay the needle on a piece of paper, bark or grass and float it on the surface of water.

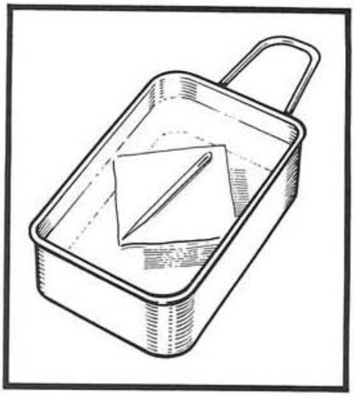

USING ELECTRICITY

If you have a power source of two volts or more (a small dry battery, for instance) the current can be used to magnetize the metal. You will also need a short length of wire, preferably insulated.

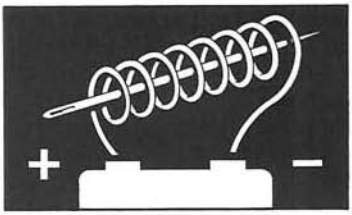

Coil the insulated wire around the 'needle'. If it has no ready-made insulation wrap a few layers of paper or a piece of cardboard around the needle first. Attach the ends of the wire to the terminals of the battery for five minutes.

RAZOR BLADE COMPASS

A thin flat razor blade can also be used as a compass needle because it is made of two metals bonded together. It can be magnetized simply by stropping WITH CARE against the palm of the hand. Suspend it.

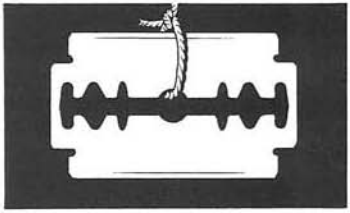

Use other methods to establish which general direction is north and then identify which end of your new compass needle is which and mark one of them. Top up your needle's magnetism from time to time, and always check your reading with the sun. A 'wild' reading may be given if large amounts of ferrous metal are nearby.

Plant pointers

Even without a compass or the sun to give direction you can get an indication of north and south from plants. They tend to grow towards the sun so their flowers and most abundant growth will be to the south in the Northern Hemisphere, the north in the South.

On tree trunks moss will tend to be greener and more profuse on that side too (on the other side it will be yellowish to brown). Trees with a grainy bark will also display a tighter grain on the north side of the trunk.

If trees have been felled or struck down the pattern of the rings on the stump also indicates direction – more growth is made on the side towards the Equator so there the rings are more widely spaced.

There are even species of plant known for their north–south orientation:

North Pole Plant *which grows in South Africa, leans towards the north to gain full advantage of the sun.*

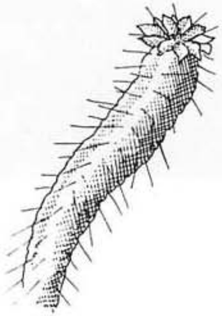

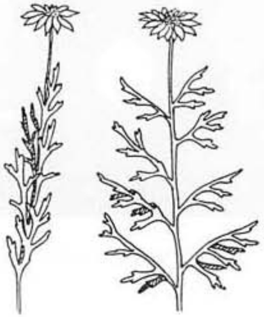

Compass Plant *of North America directs its leaves in a north–south alignment. Its profile from east or west is quite different from that of north or south.*

The wind direction

If the wind direction of the prevailing wind is known it can be used for maintaining direction – there are consistent patterns throughout the world but they are not always the same the whole year round.

Where a strong wind always comes from the same direction plants and trees may be bent in one direction, clear evidence of the wind's orientation. But plants are not the only indication of wind direction: birds and insects will usually build their nests in the lee of any cover and spiders cannot spin their webs in the wind. Snow and sand dunes are also blown into distinctive patterns by a prevailing wind which blows from the outside of the high central ridges.

Making use of the moon

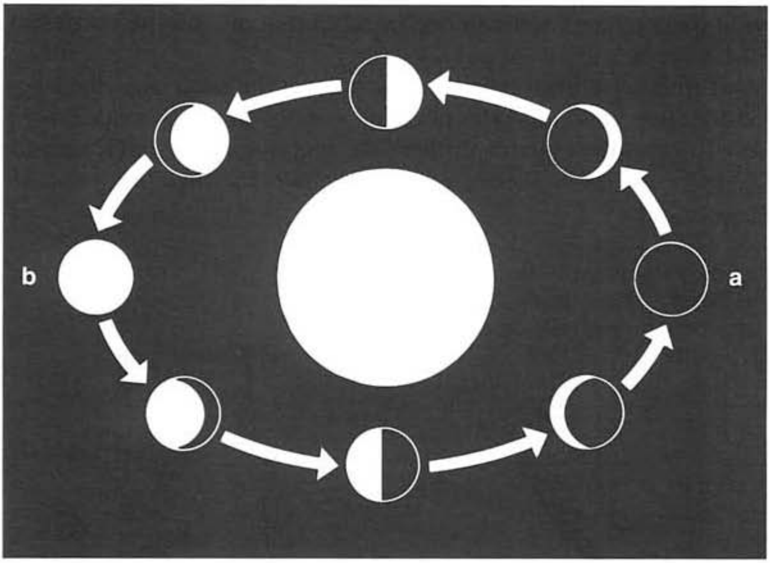

The moon has no light of its own, it reflects that of the sun. As it orbits the earth over 28 days the shape of the light reflected varies according to its position. When the moon is on the same side of the earth as the sun no light is visible – this is the 'new moon' (a) – then it reflects light from its apparent right-hand side, from a gradually increasing area as it 'waxes'. At the full moon it is on the opposite side of the earth from the sun (b) and then it 'wanes', the reflecting area gradually reducing to a narrow sliver on the apparent left-hand side. This can be used to identify direction.

If the moon rises BEFORE the sun has set the illuminated side will be on the west. If the moon rises AFTER midnight the illuminated side will be in the east. This may seem a little obvious, but it does mean you have the moon as a rough east-west reference during the night.

Direction by the stars

The stars stay in the same relation to one another and pass over the same places on the earth night after night. Their passage over the horizon starts four minutes earlier each night – a two-hour difference in time over a month. If you study a star at a certain position at a certain time one evening and then check its position the next evening at the same time you will find that it has moved one degree of arc anticlockwise in the Northern Hemisphere, clockwise in the Southern. Rising in the east, stars attain a zenith and set on the western horizon at the same distance from their zenith as they rose.

The stars have been studied for thousands of years and the groups, or constellations, in which they appear to the naked eye were named in ancient times after the animals and mythological figures that their shape suggested.

In the Northern Hemisphere there are groups of stars that remain visible throughout the night, wheeling around the only star that does not appear to move – the Pole Star (a valuable navigation aid, for it is located almost above polar north). In the Southern Hemisphere the Pole Star is not visible and there is no comparable bright and stable southern star, but direction finding in the southern hemisphere makes use of a constellation called the Southern Cross, in a way that is explained later.

The northern sky

The main constellations to learn are the Plough, also known as the Big Dipper (diagram 'a' on page 362), Cassiopeia (diagram 'b' on page 362) and Orion (diagram 'c' on page 362), all of which, like all stars in the northern sky, apparently circle the Pole Star (illustrated in diagram 'a'), but the first two are recognizable groups that do not set.

These constellations come up at different times according to latitude and Orion is most useful if you are near the Equator.

Each can be used in some way to check the position of the Pole Star but once you have learned to recognize it you probably will not need to check each time.

A line can be drawn connecting Cassiopeia and the Plough (the Big Dipper), through the Pole Star. You will notice that the two lowest stars of the Great Bear point almost to the Pole Star. It will help you to find these constellations if you look along the Milky Way.

The Plough (the Big Dipper) (a) *is the central feature of a very large constellation, the Great Bear (Ursa Major). It wheels around the Pole Star. The two stars Dubhe (x) and Merak (y) point, beyond Dubhe, almost exactly to the Pole Star about four times further away than the distance between them.*

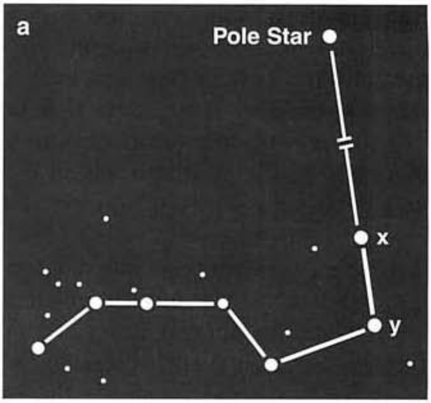

Cassiopeia (b) *is shaped like a W and also wheels around the North Star. It is on the opposite side of the Pole Star and about the same apparent distance away as the Plough (the Big Dipper).*

On clear, dark nights this constellation may be observed overlaying the Milky Way. It is useful to find this constellation as a guide to the location of the Pole Star, if the Plough is obscured for some reason. The centre star points almost directly towards it.

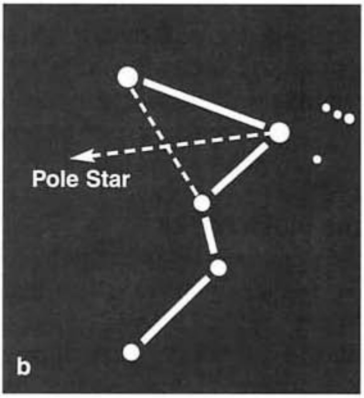

Orion (c) *rises above the Equator and can be seen in both hemispheres. It rises on its side, due east, irrespective of the observer's latitude, and sets due west. Orion appears further away from the Pole Star than the previous constellations. He is easy to spot by the three stars making his belt, and the lesser stars forming his sword.*

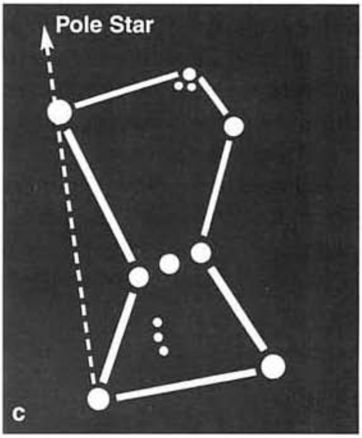

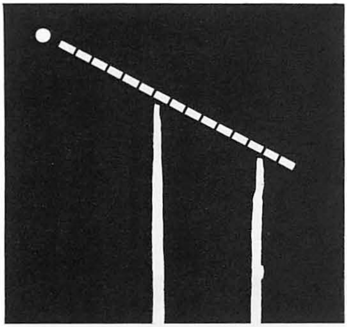

Other stars *that rise and set can be used to determine direction. Set two stakes in the ground, one shorter than the other, so that you can sight along them (or use the sights of a rifle propped in a steady position). Looking along them at any star – except the Pole Star – it will appear to move. From the star's apparent movement you can deduce the direction in which you are facing (See info. box).*

INFO.

Apparently rising = facing **east**
Apparently falling = facing **west**
Looping flatly to the right = facing **south**
Looping flatly to the left = facing **north**

These are only approximate directions but you will find them adequate for navigation. They will be reversed in the Southern Hemisphere.

Reading the southern sky

There is no star near the South Celestial Pole bright enough to be easily recognized. Instead a prominent constellation is used as a signpost to south: the Southern Cross (Crux), a constellation of five stars which can be distinguished from two other cross-shaped groups by its size – it is smaller – and its two pointer stars.

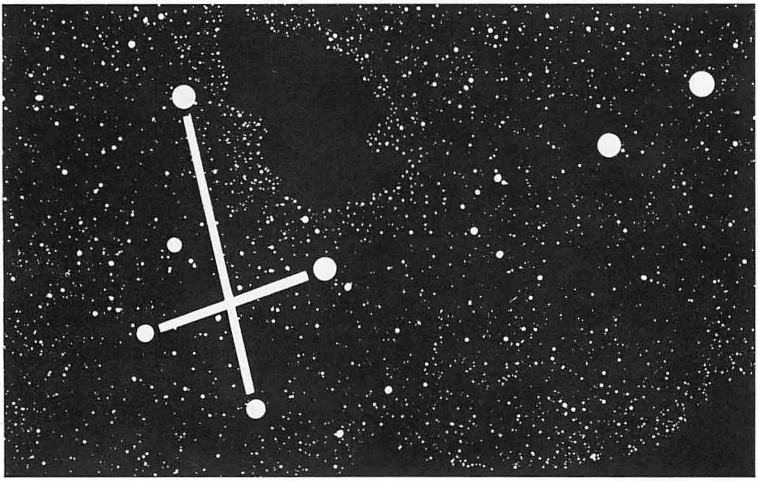

Finding the Southern Cross

One way to find the Southern Cross is to look along the Milky Way, the band of millions of distant stars that can be seen running across the sky on a clear night. In the middle of it there is a dark patch where a cloud of dust blocks out the bright star background, known as the Coal Sack. On one side of it is the Southern Cross, on the other the two bright pointer stars.

FINDING SOUTH

To locate south, project an imaginary line along the cross and four and a half times longer and then drop it vertically down to the horizon. Fix, if you can, a prominent landmark on the horizon – or drive two sticks into the ground to enable you to remember the position by day.

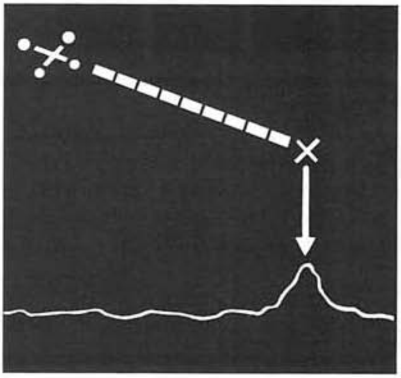

WEATHER SIGNS

Weather is much more localized than climate. Although it is possible to generalize about the weather to be expected in different parts of the world, and in some territories weather patterns are very stable, the geography of one small area may make it differ considerably from that adjoining.

Weather patterns are produced partly by the broad movements of wind and water over the whole globe and partly by localized differences in temperature and air pressure which cause air movements. These produce winds and carry rain.

In general, air moves from high-pressure areas into low-pressure zones, with warm air expanding and rising, cooler air moving in underneath. The warm air takes up moisture but at higher altitudes, or when some other cause brings down its temperature, the moisture begins to condense as cloud and eventually will fall as rain.

The most obvious example is where mountains force air currents upwards, rain falling on the slopes. In some places so much water is lost that on the far side of the range there is a dry 'rain shadow' territory.

However, that does not necessarily mean that if, from the dry terrain, you climb and cross the ridge you will move into a well-watered zone. The mountain zone may continue some way before the area of great precipitation is reached, or some earlier physical feature may have caused clouds to shed most of their rain.

Coastal areas

In coastal areas, whether of an ocean or an inland sea or lake, there will usually be a wind pattern that reverses from day to night. Water absorbs and loses heat less readily than the land and consequently it tends to be cooler than land during the day and warmer at night. The temperature difference affects the air above it and during the day breezes usually blow from sea to land, at night the wind changes and blows off the land.

Where an island is close to a large land mass these patterns may be overlaid by a broader air movement but a very regular pattern of day-night change in wind direction suggests a large body of water in the direction from which the day wind blows.

Winds

Winds can carry scents with them, providing information about the place from which they blow. Even to the untrained nose the smell of the sea will be recognizable, and to the shipwrecked survivor the smells of vegetation will indicate the direction of land. However, do not rely entirely on the nose, use other evidence to confirm its message.

Where winds tend to maintain direction they can be an aid in keeping to a course, but some other check should also be regularly made to keep direction.

Study the direction of the wind and the accompanying weather. Make a note of them. Dependent always upon barometric conditions, wind from a certain direction is always likely to bring a similar kind of weather. It is a guide for weather prediction.

If a wind is strong and dry the weather will remain constant until the wind drops or veers, then it may rain.

If it is foggy and misty you may get condensation but you will not get rain – but if a wind rises and blows away the fog it may turn to rain.

On a fine day a noticeable increase in the strength of the wind indicates a weather change. (See also *Hurricanes and Tornadoes* in *Disasters.*)

CLOUDS

Watch the way that clouds change, for clouds are the most reliable of weather signs.

Clouds are formed from masses of water vapour which becomes visible as it condenses with cooling. If cooling continues the droplets increase in size until, too heavy to remain airborne, they fall as rain. When their temperature rises sufficiently they evaporate and the cloud disperses.

There are ten main types of cloud formation. Approximate altitudes are given for each type. The same shapes occur at lower altitudes in polar regions.

The higher the clouds the finer the weather.

Small black clouds scudding beneath a dark stratus layer often bring showers.

Clouds hanging on high ground indicate rain, unless they move by midday.

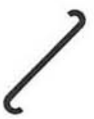

Cirrocumulus clouds *are small rounded masses, looking like rippled sand and often referred to as a 'mackerel sky'. Normally an omen of fair weather, they usually follow a storm and dissipate, leaving a brilliant blue sky.*

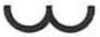

Altocumulus clouds *are fair weather clouds, similar to cirrocumulus but on a larger scale, thicker, not so white and with shadows in them. They usually appear after a storm.*

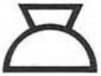

Cumulonimbus clouds *are low thunder clouds. Dark and angry looking, they may tower to 6,000m (20,000ft) with the top flattening out in what is often called an anvil top. This is a cloud that brings hail, a strong wind, thunder and lightning. False cirrus appears above, false nimbostratus below.*

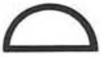

Cumulus clouds *are very easy to recognize: fluffy white clouds, not unlike cauliflowers. They are usually an indication of fair weather when widely separated but, if they become very large and develop many heads, they are capable of producing sudden heavy showers.*

Cumulus clouds at sea in an otherwise cloudless sky are often an indication of land beneath them.

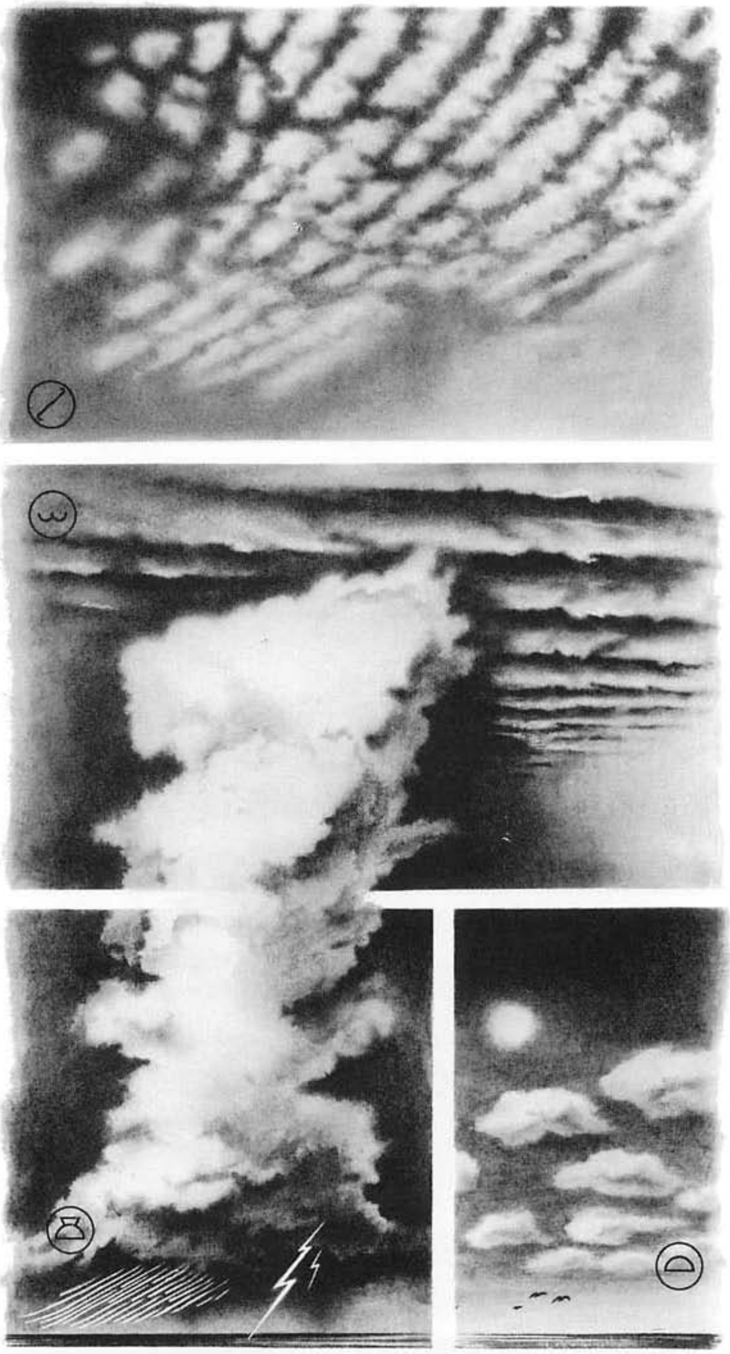

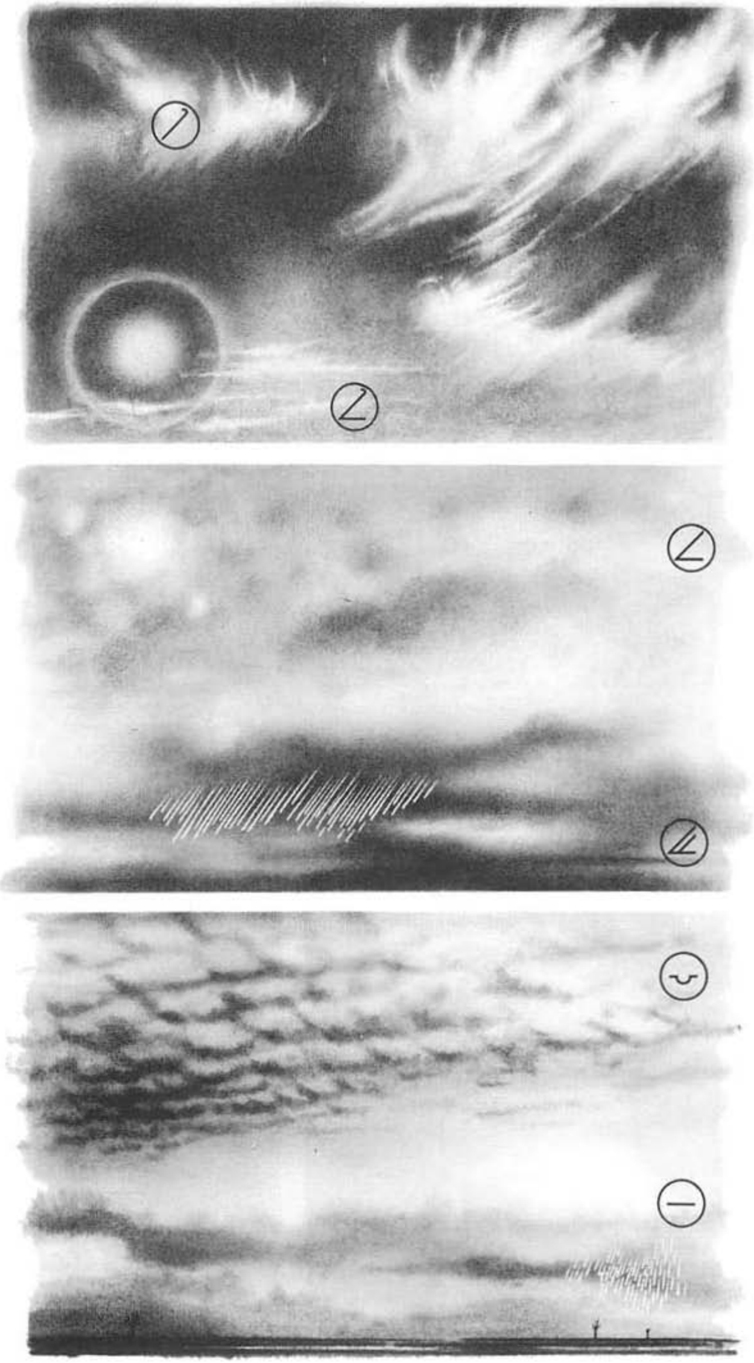

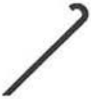

Cirrus clouds *are high, wispy clouds formed from ice crystals which give them a white appearance. Often called 'mares' tails', they are seen in fine weather.*

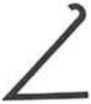

Cirrostratus clouds *are clouds made up of ice particles and look like white veins. These are the only clouds which produce a halo around the sun or moon. If it gets bigger it means fine weather, smaller a sign of rain.*

If the sky is covered with cirrus clouds and the sky above them darkens and the formation changes to cirrostratus it is an indication that rain or snow are coming.

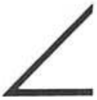

Altostratus clouds *form a greyish veil through which the sun or moon may appear as a watery disk. If wet weather is approaching the disk will disappear and the cloud thicken and darken until it begins to rain.*

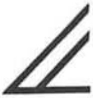

Nimbostratus clouds *form low, dark blankets of cloud and spread gloom. They mean rain or snow within four or five hours and usually the rain continues for hours.*

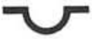

Stratocumulus clouds *form a low, lumpy, rolling mass, usually covering the whole sky, though often thin enough for the sun to filter through them. Light showers may precipitate from these clouds but they usually dissipate in the afternoon and leave a clear night sky.*

Stratus clouds *are the lowest of clouds and form a uniform layer like fog in the air – they are often described as hill fog when they occur. They are not a normal rain cloud but can produce a drizzle. When they form thickly overnight and cover the morning sky they will usually be followed by a fine day.*

PREDICTIONS

To be caught in bad weather could be fatal to survival. There is a time to go out or move on and a time to shelter. With an awareness of certain signs, short-term weather predictions can be made to help decide which to do.

Before setting out on any activity take note of the weather and any changes that are likely. Learn to observe all wind and pressure changes and keep a record of the weather and the conditions which precede it and what they develop into.

Wildlife indicators

Animals have great sensitivity to atmospheric pressure which aids them in forecasting the weather a day or two in advance.

Insect-eating birds, such as swallows, feed higher in good weather, lower when a storm is approaching.

Unusual rabbit activity during the day, or squirrels taking more food than usual to the nest, may be a prelude to bad weather.

Nature, however, does not go in for long-time forecasting. A squirrel's hoard of nuts is an indication of its industry, not the intensity of the coming winter. The depth of a bear's den has no relation to its severity, but reflects the soil conditions. A particularly big crop of berries is the result of previous bad conditions, the tree producing extra fruit to give the species more chance of survival.

Fireside clues

If the smoke from the camp fire rises steadily the weather is settled and likely to remain fine. If it starts swirling, or being beaten downwards after rising a short way, it indicates the likely approach of a storm or shower.

Wooden tool handles tighten at the approach of stormy weather. Salt picks up increasing dampness in the air and will not run.

'Feelings in your bones'

Curly-haired people find their hair becomes tighter and less manageable as bad weather approaches – and the same happens to animal fur. Anyone with rheumatism, corns or similar ailments can usually tell when wet weather is coming by an increase in their discomfort.

Sound and smell

When wet weather is on the way sounds tend to carry further than usual and distant noises seem more clear – the moisture-laden atmosphere acts like an amplifier. But compare like to like – remember, sound always travels better over water.

The smell of trees and plants becomes more distinctive before the arrival of rain, the vegetation is opening ready to receive it.

Signs in the sky

'Red sky at night, shepherd's delight, red sky in the morning, shepherd's warning' is one of the oldest of weather sayings. Since a red sun, or a red sky at sunset, indicates that the atmosphere holds little moisture it is unlikely that rain or snow will occur within the next two hours, but equally a red sky in the morning is a fair indication that a storm is approaching.

A grey morning is usually the start of a dry day. The dull colour is the result of dry air above the haze formed by the collection of dew on the dust particles suspended in the lower atmosphere.

An evening sky that is grey and overcast indicates rain – the dust particles are so laden with moisture that they will soon drop as rain.

Early morning mist lifting from a valley is a sure sign of fair weather. In hilly country, if mist has not lifted by noon, it is set in for the day and will probably turn to rain during the late afternoon.

A clear night sky is an indication of good, settled weather. At the end of summer it may also be a warning of frost: at night clouds insulate the surface of the earth against loss of heat. Without them frost is more likely. Cold air, being heavy, fills the hollows – avoid camping in them.

A clear sky one night, followed by one with only a few stars visible, indicates a change of weather.

A corona, a coloured circle visible around the sun or moon, can be used to forecast the weather accurately. An enlarging ring is a sign of good weather – the enlarging circle shows that moisture in the atmosphere is evaporating and day or night will be clear.

A shrinking corona around the sun or moon is a sign of rain.

Green light blinking from an afternoon sun indicates fair weather for at least 24 hours. A rainbow in the late afternoon is also a sign of fair weather ahead.

7

ON THE MOVE

Having chosen a course and planned a route, you must be able to maintain direction, to match your progress to the type of terrain and the nature of the survival group, and to exploit any techniques that make progress easier. This section deals with skills needed on the move. It should be read in conjunction with the techniques described earlier in *Climate and Terrain*.

Careful reconnaisance may be necessary to choose the easiest routes – which may not be the obvious or the quickest ones. Groups must be organized to suit the least able, or provision made for carrying them.

Waterways can offer the easiest routes to move along, if they are easily navigable and you are able to construct a raft. They can also present formidable barriers if they lie across the route you have to follow. Rivers can be dangerous. You must learn to assess the hazards they present and select the best places and methods of crossing, if crossing is unavoidable.

ON THE MOVE 374
Planning 378
Moving in groups 379
Pace and progress 380
Walking at night 381
Upland travel 382
Jungle travel 383
WATERWAYS 384
Rafts 384
Crossing rivers 387
Survival case study 391

ON THE MOVE

Whether setting off on a hunting expedition or beginning the trek back to civilization, you will need skills in negotiating terrain and in navigation – either to ensure your return to camp or to keep to your chosen route. Even if you do not have a map, your own reconnaissance will have established your immediate terrain. Guessing what is beyond is much more problematic if you have to assess it for yourself.

On short expeditions take note of the terrain as you move outwards to ensure that you can find your way back to camp. Over short distances physical features will be your guide and you can also mark your trail, but when going further or striking back to civilization you will need all the navigational aids available.

The decision to move

Except when local dangers or the need to find water and food make it imperative that you move away from the site of your accident to make camp, you will stay close in the hope of rescue (see Survival case study p.391). If you have injured persons and only limited food and water, it would make sense to send a party to contact help while others stay to care for the sick. The fittest and most able should be chosen to make the trek – unless they include a trained medic, who should stay with the sick.

But what if no rescue comes? Local resources may become exhausted: you go further each day to collect firewood, the game in the area has gone, plants, fungi, fruits and nuts are more difficult to find or require long forays out of camp. Under survival conditions there is also an increased risk of disease from staying too long in one place. Even with the strictest sanitary management the chance of disease gradually builds up.

All these will be pressing you to move. Whether that be a move to new territory or the beginning of a trek back to civilization, many of the same problems will occur and – since there are few situations so idyllic that survivors will prefer their new life to their old – it would be sensible to plan any move as the first stage of the journey out. The exceptions will be where there are still sick people on their way to recovery, or better resources are needed to provision a push through alien territory. A move to a more comfortable camp in a more hospitable area will provide better chances of recovery for the injured or sick and the chance to build up the necessary surplus and equipment ready for the final trek.

If you have a map or a clear idea of your location make for the nearest known settlement. If you have no idea where you are then the best course is usually to follow waterways DOWNSTREAM, for they will clearly define a route and generally lead to populated areas.

BE AWARE

Deciding which way to go will be influenced by all the information you have been able to gather, by the fitness and endurance of the party and by the nature of the terrain. Always remember that the most direct-looking route may not be the easiest to travel.

Even if you are simply changing the location of your camp there is no point in just moving a mile away – as you will soon have the same problems that made you decide to leave your first camp. The territory will not be entirely fresh. Move two or three days' journey at least so that fuel, flora and fauna will all be undisturbed.

Although you may expect similar resources to be available, as long as you are travelling through similar terrain, remember that you have been hunting and foraging with experience of where things are best found or trapped and with known supplies of fuel and water. You will have to find your basic needs in new territory and when you move into different terrain availability may change dramatically – for better or worse.

Preparations

Before you finally abandon camp leave signs that will make it clear you have been there and have moved on (see *Signalling* in *Rescue*). Leave a message giving a list of who is in the party and details of your intentions. Mark the trail as you proceed so that if searchers do find the camp they will be able to follow you.

Build up a stock of preserved food, make water bottles and larger containers too if you envisage crossing waterless territory, litters or other means of transporting any remaining sick, the old or very young. Make suitable foot-coverings and clothing for everyone and packs to carry equipment and supplies. Some form of transport may be possible – a sledge or raft. Take shelter material with you – cloth, canvas, ponchos, even sticks if they are likely to be scarce in the terrain you are crossing. A shelter which can be quickly re-erected, will conserve energy for more urgent gathering than roofing materials and supports. Naturally, you will take signalling gear, for you may find an opportunity to make that vital contact which will bring rescue.

Study weather patterns carefully to choose a time for travel when the weather is likely to be settled.

HUDSON BAY PACK

A comfortable and easily improvised way of carrying equipment, this needs strong and preferably waterproof material about 90cm (3ft) square, two small stones and cord or thonging more than long enough to loop across the body.

Place stones in diagonally opposite corners of the cloth. Fold ends of cloth over stones. Tie cord below the stones, securing them in position. The stones in turn prevent the cord slipping off. Lay cloth on the ground and roll possessions up tightly. Wrap pack around the body, either across the back or around the waist.

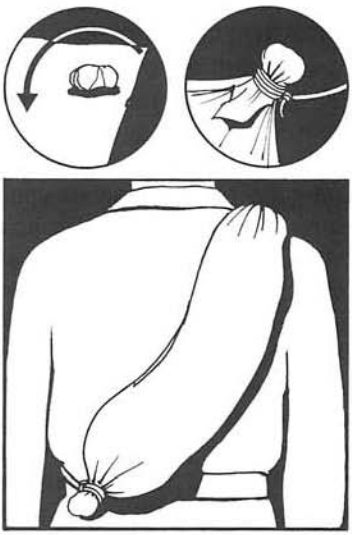

BACK-PACK FRAME

Make a ladder frame to fit against your back, with a right angle projection at the bottom secured by side struts. Add shoulder straps and a belt loop. Use this to support a bag, a bundle of supplies or equipment tied into place.

FORK FRAME

A quicker, but less-efficient support can be made from a forked bough with cross-pieces to which baggage can be tied.

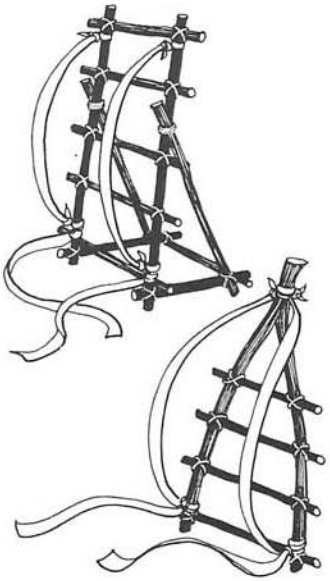

BABY BAG

Carry babies and small children papoose style on your back or front. Tie lower corners of a rectangle of cloth around the waist, pop the child in and tie the upper corners around your neck. Pad at neck to ease pressure or chafing.

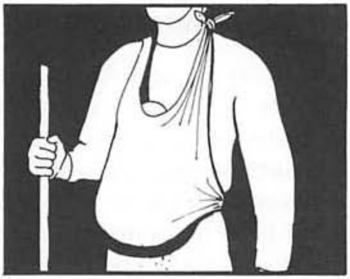

Carrying people

Pickaback and fireman's lifts are impracticable for long distances and stretchers are difficult to handle. Sit small children on a backpack frame or make carrying chairs on poles to be borne by several people. If you are alone, try a sledge or travois (see *Furnishing the camp* in *Camp Craft*). Equipment can be carried in the same way.

Making a sledge

Sledges are particularly useful on snow and ice, when they will move most smoothly, but may also be used on smooth ground. The shape of the front runners is critical, especially on snow. You can make use of doors and cowlings from a crashed aircraft or vehicle in the construction. Tie lines to the front runners with a bowline to the people hauling – ideally two at the front, and two at the rear as brakemen on gradients. Test thoroughly before using on a long trek.

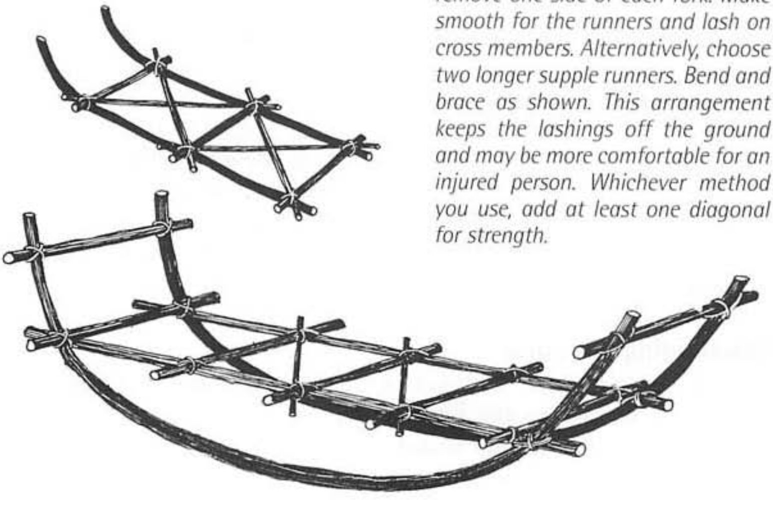

Choose two forked branches and remove one side of each fork. Make smooth for the runners and lash on cross members. Alternatively, choose two longer supple runners. Bend and brace as shown. This arrangement keeps the lashings off the ground and may be more comfortable for an injured person. Whichever method you use, add at least one diagonal for strength.

PLANNING

If you are on very high ground, above a large plain, it may be possible to plot out a route with some precision. In most situations visibility will be restricted and you will have to guess what is over the furthest ridge and what occupies the 'dead' ground ahead. Even when you can see the terrain ahead, it is difficult to see the details. What looks like a manageable slope may prove a barrier when you get closer to it. If you have them, make good use of field glasses in studying every potential route.

To see further you may consider climbing a tree – but keep close to the trunk and test each branch before risking your weight on it. This is NO time to risk a fall.

Following rivers

Following a watercourse, however small, offers a route to civilization and a life-support system on the way. Most rivers lead eventually to the sea or great inland lakes. Apart from the rare exception where rivers suddenly descend beneath the earth, they offer clearly defined routes to follow. Sometimes, in their upper reaches, they may cut through gorges and it can be impossible or inadvisable to take a route along their banks, which may be steep, rocky and slippery. In that case take to high ground and cut off the bends, following the general course of the stream.

On more level ground a river is easier to follow and may well have animal trails beside it which you can use. In tropical conditions the vegetation is likely to be denser by the river – for the light can reach below the trees, and the banks may be hard to negotiate. If the river is wide enough it would be worth considering building a raft. Even if there is no bamboo, which makes an ideal raft-building material, there are likely to be sound fallen trees for timber.

When, on flat plains, a river makes huge meanders, the inside of the loops may be swampy and prone to flooding – you can recognize such very wet ground by the lushness of the vegetation and rush-like plants. Avoid marshy areas if you can, and cut across the loop.

Maintaining direction

Having decided upon a direction, try to maintain it. Choose a prominent feature in the distance and keep heading towards it. Travelling through forests makes orientation very difficult and a compass becomes a valuable asset.

If you are in featureless territory, but in a group of three or more, separate to follow each other at wider intervals and look back frequently. If you are following in each other's tracks those behind you will be directly behind each other. If the party always moves on in relay – one moving on ahead, then resting while everyone else moves up from the rear – the straight line will be maintained. On your own you can try to align yourself by looking backwards at your own tracks if they are visible, as they are likely to be on snow or sand. Better still, you can set up sticks or piles of stones in alignment with each other so that you can check that you are not deviating from your route.

If possible skirt rocky outcrops and areas of dense vegetation and, once on high ground, stick to it until certain that you have found the spur down which you can make the best progress in the desired direction.

MOVING IN GROUPS

Always move in an organized manner, in some kind of formation, and not as an unruly gaggle. This way it will be easier to check that no stragglers have been left behind and to ensure that there is help for anyone in difficulty. Before setting out for the day, have a briefing to discuss the route, any obstacles expected and any special procedures.

Divide responsibilities

One person becomes a scout, responsible for selecting the best route, avoiding dead falls, loose rocks etc and finding the best way down a slope.

Number two is responsible for making sure the scout, who will be preoccupied with skirting obstacles, maintains correct overall direction. Others should relieve them frequently, for the lead scout's job in particular is very tiring.

The rest of the party should keep their eyes open for edible plants, berries and fruits and everyone should be responsible for at least one other person to ensure that no one drops by the wayside. A head count and check on everyone's condition is particularly important after a river crossing or negotiating a particularly tricky stretch of terrain. Someone should initiate equipment checks at frequent intervals.

Always travel in at least pairs – and be especially careful in bad weather, and if you have to travel at night, that you do not get split up. It is usually the person in front who gets split off from the group – people are more likely to remember to look for the stragglers. The scout climbs over an obstacle, the second person sees the scout struggling and then sees an easier route to take – the rest of the party follow and the lead person is separated from the group. This is when the benefit of EVERYONE knowing the proposed route, and nominating prominent features as rallying points, is apparent. If separated or in an emergency everyone knows where to regroup.

Availability of water, fuel and plants may be an indication of what is available in similar locations further ahead. An eye should always be open for places that offer good shelter – if the weather suddenly turns bad you can backtrack to one of them.

PACE AND PROGRESS

- A large group can send an advance party ahead with the responsibility for clearing the route and setting up the night's camp, ready for the slower-moving injured or less able. A clear trail will make the carrying of baggage and any unfit people much easier. Sick and injured should be provided with fully-fit escorts in case they encounter any difficulties.

- Ensure that the lead person in the party does not go too fast for those behind. After an obstacle wait and allow everyone to catch up before moving on.

- It is best to try to maintain an even pace – smooth, pendulum-like movement tires the legs less than a jerky pace or flexing the knees. It helps to swing the arms – and they should certainly not be pushed in the pockets, especially when going up or down hills for if you then slip you have less chance of stopping yourself falling or sliding.

- Rest frequently (whether in a party or alone). Stop, sit down and see how everyone is doing. Adjust loads which are uncomfortable and repack them if necessary. On average take a break of 10 minutes every 30–45 minutes, depending on the terrain and condition of the group.

- On steep ground the pace should be shortened, on easy ground lengthened. On descents avoid overstepping for this jars the body and increases fatigue.

- On steep or slippery ground ropes can provide a hand hold to help people negotiate a particularly tricky stretch (quite apart from abseiling and other climbing techniques). They will be an asset on scree, as much as on icy slopes.

- You can also use a line with Prusik knots attached so that the young and aged can be tied to a fixed rope for extra safety (see *Knots* in *Camp Craft*).

- In estimating distance you have covered allow 3km (under 2 miles) per hour, but going uphill knock off a third.

WALKING AT NIGHT

Negotiating unknown territory at night can be very dangerous, but may be necessary in an emergency, or there are circumstances – in the desert for instance – when it may be more comfortable to travel at night.

The night is never completely dark and outdoor vision is not totally lost, even for man. However, because it is difficult to see things clearly you are easily disoriented, which leads to a feeling of being lost. A compass is a great help in maintaining a heading and dispelling any such fears. It is always darker among trees than out in the open – so keep to open country if you can.

When looking at an object at night it is best to look at one side of it and not directly at it. It is difficult to distinguish anything in a dark central mass but the edges show more clearly and in poor light objects at the edges of your vision are often seen more distinctly.

Once the eyes get accustomed to the dark, more and more is seen as 'night vision' is acquired. It takes about 30–40 minutes for the eyes to get accustomed to the dark. Once this is achieved, the eyes must be protected from bright light or the night vision will be impaired for quite a while. If there is an unavoidable reason for having to use a light, cover one eye so that the vision in that eye at least will be retained. If, for instance, you need to consult a map, a red filter over a torch will help you to retain your night vision.

The ears are good sensors in the dark – the sound of a river, for instance, provides a good guide to how fast it is flowing. Vegetation can be smelt in the dark and familiar smells can aid identification.

Walk SLOWLY in the dark and test each step before putting all your weight forward. If going down a slope use a shuffling step.

UPLAND TRAVEL

In mountainous and hilly country it is best to keep to high ground – it makes navigation easier. Rivers may be in steep-sided gullies and have rapids, falls, and slippery rocks that are difficult to negotiate on foot. You could end up spending an unhealthy length of time in the water.

Use spurs to climb out of valleys and get on to the ridges. If they are very exposed you may have to drop down into the valleys for shelter at night and to find water but you will be able to cover more ground than by negotiating the spurs.

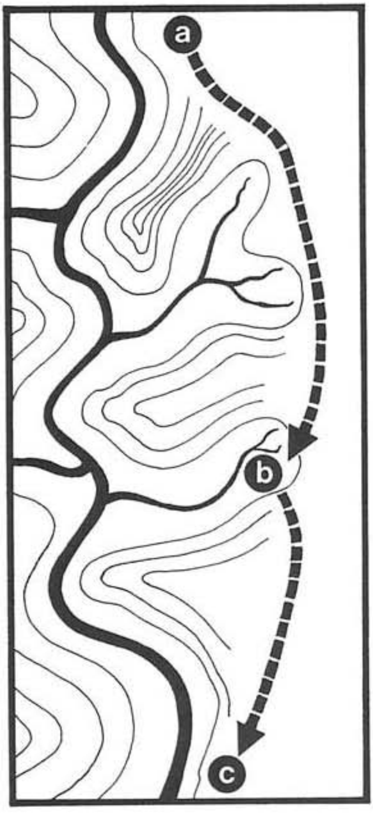

It is no use following a river if it winds in deep valleys through very hilly country. By climbing from the valley at (a) and following the ridge, steep and tiring descents and climbs are avoided.

At (b) a night halt is taken dropping down to the first available water source. This could provide shelter too, which may be unavailable on an exposed ridge.

Be aware of fading light and your own flagging energy. Look for shelter before they are exhausted.

Do not go down to the valley bottom if you can find shelter and water on the way. Not only will you save energy, you may be warmer.

Pockets of cold air are quite often trapped in the bottom of valleys.

If you carry water and shelter materials, stay on the high ground, choosing the most sheltered spot.

Then the route follows the ridge further before dropping down into the widening main valley to follow the watercourse again at (c). Follow the tip of a spur to go down into the valley when a river gets larger and the valley opens out.

Steep slopes

In mountainous country and on high hills, snow and ice may be encountered, and even without them such country can be dangerous with loose scree, steep slopes and crags that have to be negotiated.

Traverse steep slopes in a zig-zag and as you change direction always set off with the uphill foot. This avoids having to cross your legs over each other, which can make you lose balance. When climbing steep slopes lock your knees together after each step – this rests the muscles.

Descending steep slopes, keep your knees bent. Try to go straight down – and if you are picking up too much speed, sit back. Avoid loose rocks and scree but, if you have to negotiate loose surfaces, it helps to dig in the heels and lean back while descending. When climbing, test every foothold before putting your weight on it. Don't step on stones or logs on steep slopes, they may dislodge.

With practice it is possible to jump down loose ground – dig in the heels and slide – provided that there are no sudden drops below. Keep the feet square and shoulder width apart and allow yourself to slide. As you increase in speed, dirt will build up under your feet and you will lose control. Jump and start again. If the slope is very steep abandon this method – abseiling (rappelling) is the answer (see *Mountains* in *Climate and Terrain*).

JUNGLE TRAVEL

In dense jungle you may have to cut your way through if there is no way of going round. Chop downwards and as low as possible at the stems on both sides so that they fall away from the path you are making, not across it. Avoid leaving spikes standing, bamboo points can be lethal if someone stumbles. High growth and creeper can often be cut and travelled over.

Jungle vegetation seems to be covered in thorns and spikes. You may have to twist and turn to avoid vegetation that seems to surround you. Rushing only makes it worse. Those types of the climbing palms, atap and rattan, that are known as *nanti sikit* – 'wait-a-while' – in Malaya and similar names elsewhere, have thorns like fish hooks at the end of the leaf. When snared by them you must back off and untangle. Don't try to tear through wait-a-while vine – it will strip you naked. You'll soon see why it got its name! However it has many uses and rattan is one of the best of jungle water vines.

Keep feet covered to protect them from sapling spikes, snakes and chigoes (chiggers). Stop frequently to remove parasites. Chigoes ignored for more than an hour or so will cause infection.

WATERWAYS

If a river is wide enough to be navigable, it will be easier to float on it than to walk beside it. The long-term survivor could experiment with making canoes – burning out the centre of a tree trunk to make a dug-out or covering a frame of willow with birch bark or skins – or copying boats made from reeds like those built by the ancient Egyptians and Mesopotamians, which can still be found today at places such as Lake Titicaca in Peru.

All are difficult to make well enough to stand up to water travel and even among the peoples for whom making them is a traditional skill, they are usually the work of specialists.

RAFTS

More practical for the survivor will be to construct a raft, which will not capsize so readily if the structure is not perfect. All boats and rafts must be soundly tested in safe water near the camp before setting out on a journey.

In jungle terrain especially you may find that the river, beside which you have camped, has been swelled by seasonal rains to make rafting a viable proposition. Here, too, you may find ample timber, either bamboo (which is ideal) or uprooted trees which are sound and unrotted. If you have to cut timber choose leaning trees for they are the easiest to drop. With dead falls the top of the trunk is usually sound enough to use for a raft.

You can use oil drums or other floating objects to support a raft, and if there is no supply of strong timber a sheet of tarpaulin or other waterproof material can be used as a man-carrying version of the coracle described later for floating equipment across a river.

Never take chances with a flimsy raft on any water. On mountain rivers there are often rapids which only a really tough structure will survive. On the wide lower reaches there will be a long way to swim to the bank if you have a raft break up under you.

Travelling by raft

Tie all equipment securely to the raft or to the safety line, making sure that nothing trails over the edges where it could snag in shallows.

Everyone aboard should have a bowline attached around the waist and secured to a safety line or to the raft.

BAMBOO RAFT

A single layer will not support you unless it is very long, so go for a two-layer model.

Cut thickish bamboo in 3m (10ft) lengths. Make holes through the canes near the ends and half-way along. Pass stakes through these holes to connect the canes. Lash each of the canes to each of the stakes with twine, rattan or other vines or cable. Make a second deck to fit on top of the first and lash the two together.

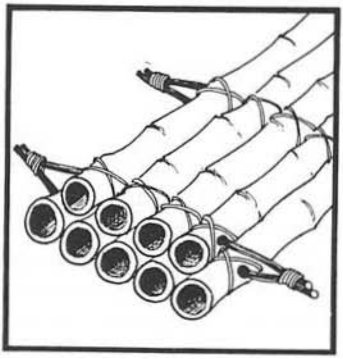

GRIPPER BAR RAFT

This is the quickest raft to build. You need logs for the deck and four thickish stakes with some pliability which are long enough to overlap the width of the deck.

Place two of the stakes on the ground and lay the logs over them.

Place the other stakes on top. Tie each pair of stakes firmly together on one side. Then, with a helper standing on top to force the other ends together, tie these so that the logs are gripped between them. Notching the ends of these gripper bars will stop the ropes from slipping.

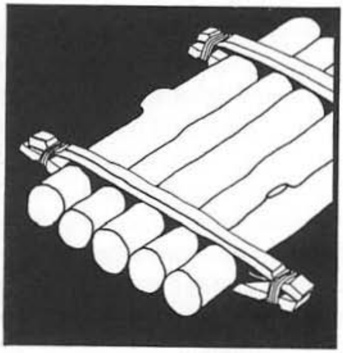

STEERING

To steer the raft make a paddle rudder and mount it on an A-frame near one end of the raft. Secure the A-frame with guy-lines to the corners of the raft and tie the rudder on to it so that it does not slip. The rudder can also be used as a sweep for propulsion. You may need to notch the raft for the base of the A-frame. The guy-lines could be tied to the cross-pieces.

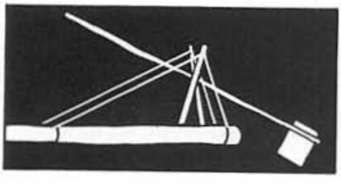

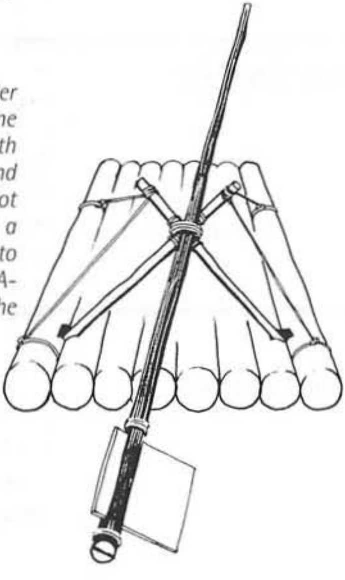

Lifelines should be long enough to allow free movement, but not so long that they trail in the water. In narrow swift-flowing rivers with dangerous rapids and waterfalls it is better not to tie on. If the raft gets out of control and is swept towards dangerous water, it is better to head for the bank.

In shallow water the best means of controlling a raft is like a punt, but preferably with two long poles – with one person poling at one front corner of the raft, and another at the diagonally opposite back corner.

If the survival group is a large one, several rafts will be needed. The fittest should be on the first raft, carrying no equipment or provisions. They can be lookouts and give early warning of hazards to be avoided. If they have to abandon their raft no kit will be lost.

Waterfalls and rapids can be heard some time before you reach them, and often are indicated by spray or mist rising in the air. If uncertain about the safety of the stream ahead, beach or moor the raft and carry out a reconnaissance on foot.

If you reach a difficult or dangerous stretch of water unload the raft and take to the bank, carrying all the equipment downstream of the dangerous waters. Having posted someone downstream where the river becomes safe and manageable to recover the raft, release the raft and let it drift down through the difficult stretch. It will probably need repairing but at least YOU will be safe and have all your equipment.

REMEMBER:

Only raft by day, NEVER in the dark. At night secure the raft firmly – so that it will still be there in the morning – and make temporary shelter on higher ground away from the river.

Bogs and marshes

If you cannot avoid crossing a marsh make your way by jumping from tuft to tuft of grass. If you find yourself sinking into a bog, 'swim' with a breast-stroke to firm ground – don't try to jump. Spreading your body over the surface distributes your weight. Use the same technique in quicksand.

CROSSING RIVERS

The headwaters of a river will be narrow and swift-running. Although banks may be steep and rocky it will usually be possible to find a place to cross. Where the water is shallow you may be able to wade – but test ahead with a pole for hidden depths. You may find rocks to provide stepping stones across or be able to place them for small streams.

Some members of a party may be able to leap across a narrow chasm or from boulder to boulder across the stream bed, but that is not much help if others cannot make it. A slip on a boulder is an easy way to sprain an ankle.

The estuary of a river is wide with strong currents and is subject to tides – which can influence some rivers many miles from their mouths. Avoid crossing there, unless equipped with boat or raft, go back upstream to an easier crossing place.

On any wide stretch, and especially when near the sea, do not set off, even with boats or rafts, immediately opposite the point you hope to reach, but consider how the current will affect your passage and make allowances.

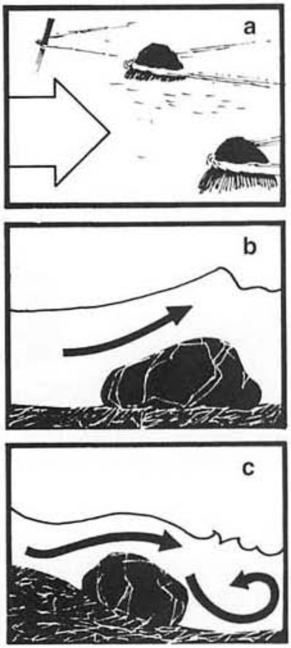

STUDY THE WATER

The surface movement of a stream or river can tell you a great deal about what is beneath. The main flow of the current is usually evident from a chevron shape of smoother water around any rock or projection (a), the V widening downstream.

Waves that appear to stay in one position on the surface (b) are usually evidence of a boulder on the bottom deflecting water upwards.

Closer to the surface an obstruction will create an eddy downstream of it where the surface water appears to run back against the main flow. If a large boulder coincides with a steep drop in the level of the bottom (c) these eddies can produce a powerful backward pull downstream of the obstruction and pull swimmers in – they are VERY dangerous.

Wading across

Even quite wide rivers may be comparatively shallow and possible to wade across, but never underestimate any stretch of water. Cut a stick to aid balance and cross facing towards the current and you will be more able to avoid being swept off your feet. Roll trousers up, so that they offer less surface to the current, or if they are going to get wet anyway take them off so that you have them dry for the other side. Keep your boots on, they will give a better grip than bare feet. Undo the belt fastening of a back-pack so that you can slip it off easily if you get swept over. But don't lose hold of it. It will almost certainly float and you can then use it to help right yourself.

Turn at a slight angle, your back towards the bank you want to reach, the current will move you in that direction. Do not take strides but shuffle sideways, using the stick to test for depth and trying each foothold before using it.

WARNING

ICE-COLD WATER IS A KILLER

Do not attempt swimming or wading across a stream when the water is at very low temperatures, it could prove fatal. Make a raft of some kind. Only wade if you can do it without getting more than your feet wet and dry them vigorously as soon as you reach the other bank.

Crossing as a group

If a group of people are wading across together, they should line up behind the strongest, who crosses as described above. The others each hold the one in front at the waist and move in step, offering less obstruction to the current.

Alternatively a group can link arms side-by-side and hold on to a pole or branch to keep them in alignment. They cross facing the bank and moving forwards. Only the side of the first person opposes the current and the group provides stability for all of them.

WARNING

Look out for submerged branches. You could get tangled in them and wrench a limb or lose balance. When carried along with the current you do not notice its strength but if it forces you against an obstruction you can be firmly held.

Crossing with ropes

If a rope is available it can make wading safer – but you need a loop of rope three times as long as the width of the stream and there must be at least three people in the party. Two of them always control the rope to keep it out of the water as much as possible and to haul the crosser to the bank if difficulties are encountered.

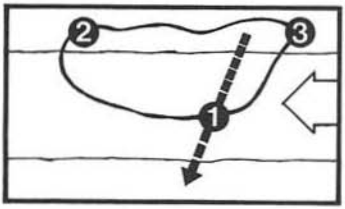

The person crossing is secured to the loop, around the chest. The strongest person crosses first. The other two are not tied on – they pay out the rope as it is needed and can stop the crosser being washed away.

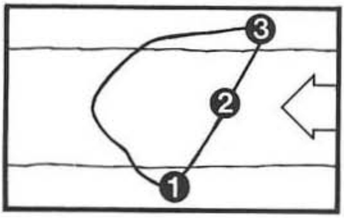

When he reaches the bank, 1 unties himself and 2 ties on. 2 crosses controlled by the others. Any number of people can be sent across in this way.

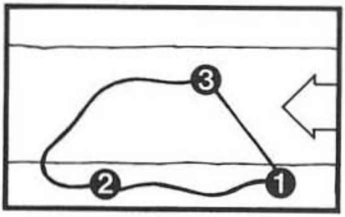

When 2 has reached the bank, 3 ties on and crosses. 1 takes most of the strain, but 2 is ready in case anything goes wrong.

RIVERS ARE DANGEROUS

Never enter the water unless there is no other way of getting across and choose a crossing point carefully.

- Avoid high banks that are difficult to climb out on.
- Avoid obstructions in the water.
- Current is likely to be fastest on the outside of bends and steep banks may be undercut making landing impossible.
- Look for an even section of river bed – shingle is the best surface for wading.

Swimming across

If you can't swim DON'T try – rely on others to get you across with the help of some sort of float. Even the strongest swimmers should make use of flotation aids when crossing a river – and for non-swimmers they will be essential. They will reduce the expenditure of energy and help to keep clothes and kit dry. Do not swim with your clothes on. Once wet they will give you no protection from cold – dry they are something warm to put on when you have crossed.

Always make sure you have found a place on the other side where you will be able to get out of the water. If there are no beaches you will need supports to haul yourself up on to the bank – but avoid tangles of branches in the water where you might get trapped. Enter the water well upstream to allow for the distance that the current will carry you down as you cross. Better to overestimate and be a little longer in the water, than pass your landing place.

Check the strength of the current by watching floating logs and flotsam and study the water surface for hidden obstructions and eddies.

If you hit weed in the water adopt a crawl stroke to cut through it. Once a strong swimmer has cleared a passage others will be able to follow through in the channel made.

Flotation aids

Fuel cans, plastic bottles, logs – anything that floats can be used. If you have a waterproof bag put your clothes and belongings inside, leaving plenty of air space tie the neck and then bend it over and tie again and use it as a float. Hang on to it and use just your legs to propel yourself.

- Without a bag, but with a waterproof sheet, pile twigs and straw into the centre to create air pockets and then pile your clothes and equipment on top before tying up the bundle securely.

- Do not attempt to sit on the bundles or place your weight on them.

- With a group of people, split into fours. Each four should lash their bags together and use them as support for an injured member of the party or a non-swimmer.

- If no waterproof material is available, make a small raft or a coracle to float your things on. Bundle your belongings and, if heavy, make the raft two-layered so that only the lower layer sinks into the water and your kit stays dry.

SURVIVAL CASE STUDY

This is a case study that I use when I do talks about survival situations. It has nothing to do with any personal experience that I have had, although it is a true story.

In this documented account from the 1950s, a family of three – a father, mother, and child – set off by small plane from the east coast of the United States to get to California. After various stops for fuel and rest the family arrived safely at their destination, although the father, who was the pilot, reported a minor fault when the plane landed for the final time. When the family were ready to head back east he went up for a test flight to make sure that the repairs he had requested were successful.

However, what they failed to do was to register a flight plan, and this proved to be a fatal error. You should always tell someone where you are going and how long you expect to be, otherwise no one will realise you are missing and think to look for you. I suspect that that there were a number of failures on this occasion and the airport should have insisted on some sort of flight plan. Nonetheless, when the plane crashed in the Trinity Mountains in northern California, nobody knew they were missing.

Despite their initial error in failing to alert the authorities about their route, this family was trained in survival techniques and, having survived the impact, remained alive for the next ten days. They then made the decision to leave the wreckage in search of civilization. And this is where they finally fell down; they didn't know how to effectively navigate and they left the security of the wreckage. Survival skills are very important but you also need to know the best possible way to get to safety. The family were eventually discovered only a number of days after they had died, only 50 km (30 miles) from the nearest road.

The lessons to be learnt from this case study are that you should always let people know where you are, stay where you are for as long as possible, but if you really feel that you must leave the area (and I don't recommend this) try to leave behind a message with your future movements. In this case the family had left a diary that was found by rescuers, revealing their whereabouts.

8

HEALTH

When no professional medical help is available, survivors have to undertake medical tasks which should normally be left to those with special training.

Traditional first-aid procedures are designed to cope with minor problems and to sustain a seriously injured person until they can receive expert treatment. However, if there is no possibility of outside help in time to save a life, the survivor may sometimes have to take drastic measures. Some of the advice given in this section is intended ONLY for such circumstances.

In the treatment of diseases and disorders the experience of centuries of herbal treatments and natural remedies can be put to good use, when no prepared drugs are available – or to reserve supplies for more serious need. Herbal medicines given here use only simple methods of extraction and preparation.

FIRST AID 394

Remove from danger 395

Breathing but unconscious 395

Breathing and pulse 396

Choking and blockages 396

Not breathing and no pulse 401

Artificial respiration 402

Is heart beating? 405

Dangerous bleeding 407

Lesser bleeding 411

Internal bleeding 412

Wounds and dressings 413

Burns 416

Fractures 419

Immobilization 420

Sprains/dislocation 425

Shock 426

Bandaging 427

Minor ailments 429

Painkillers 431

Moving the injured 432

Emergency childbirth 438

Bites 440

General poisoning 442

General disorders 443

Diseases 444

Warm climate diseases 446

Warm climate ailments 449

Cold climate hazards 451

Survival case study 455

NATURAL MEDICINE 456

Remedies 458

MEDICINAL PLANTS 462

DANGEROUS CREATURES 468

POISONOUS SNAKES 470

DANGEROUS SEA CREATURES 476

FIRST AID

Maintaining health is of primary importance to the survivor. Do not take any unnecessary risks which could lead to injury. Aim at a varied and balanced diet and make sure that you get adequate rest.

In the initial stages of the survival situation none of these may be possible but, once you have a camp established, food sources and water found, a disciplined approach will enable you to conserve energy and resources. Away from people, you are not exposed to contagious infections, unless you brought them with you. Although some diseases are insect- or waterborne, sensible precautions – especially boiling water and properly cooking food – will protect you from many infections.

Extreme climatic conditions bring their own dangers and an awareness of symptoms will help you to treat yourself and others. Inexperience or ill-luck may lead to injury, however careful you are, and an understanding of practical first aid – improvising where medical equipment is not available – is a basic survival skill. In accident situations such improvisation may be the first key to survival for those involved, when rapid action is essential. Any expedition should have at least one person with suitable specialized medical knowledge – but EVERYONE should know how to deal with basic injuries, disorders and diseases.

PRIORITIES

In an accident involving many injured people you must know which patients to treat first. When a patient has multiple injuries, breathing, heartbeat and bleeding should be given priority. Assess the injuries and handle in the following sequence:

-Restore and maintain breathing/heartbeat
-Stop bleeding
-Protect wounds and burns
-Immobilize fractures
-Treat shock

NOTE: Before approaching any accident victim, check for danger to yourself and protect yourself from it. Look out for electric cables, gas pipes, falling debris, dangerous structures or wreckage. Give initial check-up without moving the patient, if possible, but – if there is continuing danger – move the patient and yourself to a safer location.

REMOVE FROM DANGER

First reduce any further danger to the casualty or yourself by moving them to safety – away from a burning vehicle or building. In the case of a road accident, stop the traffic. With electrocution, switch off the current. If you can't, stand on dry nonconductive material and push or lever the patient from the power source with a dry nonconductive pole or stick BEFORE touching them. If gas or poisonous fumes are threatening, turn them off at source and take casualties to fresh air.

There is always a risk in moving patients with unknown injuries but, if they are further threatened, they must be moved to have any chance of surviving. People with spinal injuries are at greatest risk when moved – the spinal cord could be severed. The only safe way to move them requires several people (see *Fracture of the spine*).

Unconscious casualties

If a person is unconscious first check whether they are breathing and begin artificial respiration immediately if necessary. Check for external bleeding and injury, trying to establish the cause of unconsciousness.

BREATHING BUT UNCONSCIOUS

If the patient is breathing, and does not appear to have any spinal injury, check that there are no obstructions in the mouth, deal with any serious bleeding and place him or her in the recovery position. If lying on the back you will need gently to turn the casualty on one side, usually most easily done by grasping clothing at the hip. This produces a stable position so that any liquids or vomit produced from the stomach or nose will not enter or block the lungs and the tongue will not fall back and block the airway.

WARNING

Do NOT place a casualty with a suspected spinal injury in the recovery position. Use an artificial airway to maintain their respiration and as a means of administering mouth-to-mouth resuscitation.

CHECK BREATHING AND HEARTBEAT!

RECOVERY POSITION

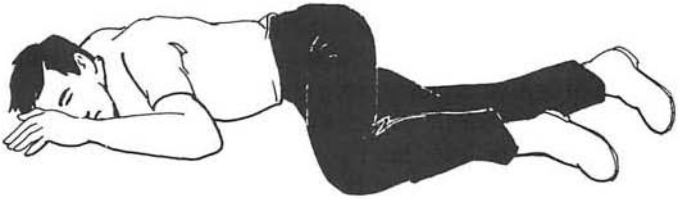

Move the arm and leg on one side of the body outwards to stop the patient lying flat, elbow and knee should both be bent. Turn the head in the same direction. Lay the other arm down along the other side of the patient.

Allow the other leg to bend slightly.

Pull the jaw forward to check that the tongue is at the front of the mouth and not blocking the airway. Loosen tight clothing.

BREATHING AND PULSE

Normal breathing is quiet and easy. Noisy breathing, froth around the nose or lips and blueness around the lips and ears are all signs of difficult or obstructed breathing. Check breathing regularly by listening carefully near the nose and mouth. Remove obstructions and, in the absence of breathing, give artificial respiration. Check at neck or wrist for pulse.

Cessation of breathing

This dire emergency may be caused by:

- Blockage of upper air passages caused by face and neck injuries or foreign bodies
- Drowning or electric shock
- Choking
- Inflammation and spasm of air passages caused by inhalation of smoke, gases or flame
- Lack of oxygen
- Compression of the chest

CHOKING AND BLOCKAGES

If breathing has stopped immediately remove any obstruction in the airway and give artificial respiration.

Clear the airway of any foreign matter: weed, vomit, false teeth or food. Sweep the mouth with a finger and ensure the tongue has not fallen back to obstruct the breathing passages.

If someone appears to be choking, but can breathe and cough, their own coughing is more effective than your aid. A blow on the back may sometimes help. If the victim cannot speak use the Heimlich manoeuvre with adults, but see other methods for special cases.

HEIMLICH MANOEUVRE

Stand or kneel behind casualty, arms around them. Clench one hand over the other, thumb side of fist pressing between waist and bottom of ribs. Apply pressure and jerk quickly upwards four times.

IF THIS DOES NOT WORK

Give four sharp blows to the back between the shoulder blades to loosen object and four more 'hugs.' Stop when victim starts breathing or coughs loudly.

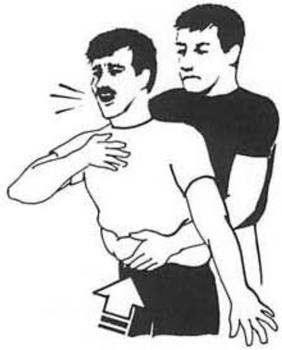

REPEAT if this does not succeed at first. DO NOT GIVE UP! Be ready to give artificial respiration, if the blockage is removed, but the patient does not start breathing. Lay an unconscious patient on his or her back, kneel astride, place your hands, one on top of the other, with the heels resting above the navel, and make quick thrusts up to the centre of the ribcage. If the blockage does not appear to shift, quickly roll patient on to side and strike four times between the shoulder blades. Repeat as necessary.

SURVIVAL SELF-HELP

If alone, use the Heimlich manoeuvre by pulling or pushing against a blunt projection – an earth bank, a fallen tree (or a chair back in a domestic situation).

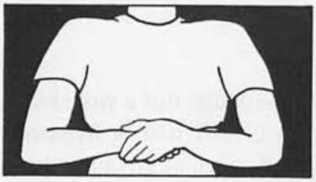

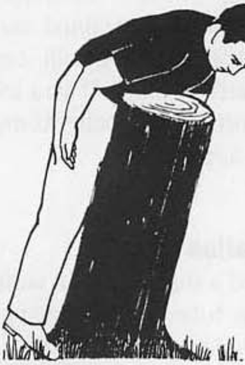

Choking: Special cases

Babies:

Support the baby, face down, straddling your forearm, with the head definitely lower than the chest. Use the heel of the free hand to give four quick blows between the shoulder blades.

Place free hand on back of baby's head and turn over. Use two fingertips to press four times, quickly and fairly firmly, on centre chest. REPEAT. Be ready to give 'mouth-to-mouth (and nose)' if breathing stops.

Children:

With small children hold them upside down and strike four rapid blows between the shoulder blades using the heel of the hand. With older children place across the knee, or bend forward from a sitting position, supporting chest with one hand while administering blows with other. Alternatively, perform Heimlich manoeuvre using two fingers of each hand instead of fists. (Not breathing – see *Artificial Respiration*.)

Pregnancy and obesity:

Abdominal Heimlich manoeuvre impossible. Position fists instead against the middle of the breastbone and follow similar procedure. (Not breathing – see *Artificial Respiration*.)

Drastic action

On the rare occasion when repeated attempts with the Heimlich manoeuvre fail to dislodge an obstruction, drastic action is required: a cut into the patient's throat below the obstruction. Also used in cases where an injury to the jaw prevents the patient from breathing, this is a technique only for a life-or-death situation. It is risky for the untrained, but worth trying to save a life in cases where the casualty will certainly die without it. There is no significant bleeding if you use the Crico thyroid technique, which is superior to a tracheotomy (where the cut is made below the Adam's apple).

Preparation

You need a sharp blade, a scalpel or a penknife, not a wide knife, and a hollow tube (a ballpoint pen case, a CLEAN fuel or hydraulic line from a vehicle, tubing from a back-pack, a small syringe and even a hollow plant stem have been used).

They should be sterilized if boiling water or a flame is available, but do not waste time getting them. WARNING: Contamination with oil or petrol from a dirty vehicle hose could produce chemopneumonia.

1 *Lay casualty on back, shoulders elevated, head and neck presenting a straight line.*

2 *Run a finger down the Adam's apple – the bone-like projection on the front of the neck (more prominent in men than women) and find another small projection just below it. Between the Adam's apple and this smaller projection you will find a central valley.*

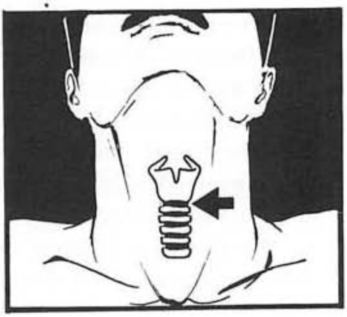

3 *Make an incision here at the exact midpoint. Keep the incision small but deep – straight down for about 1–2cm (3/8–3/4in) – note the distance on the blade beforehand. You will feel the blade move more easily as it cuts through to the windpipe. Do NOT push down further.*

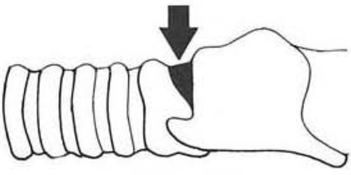

ALTERNATIVE: For the incision, pierce with a wide-bore needle – producing less bleeding (and considered preferable by doctors). A kit is produced and should be included in pare-medical equipment.

4 *Twist the blade sideways to open up cut.*

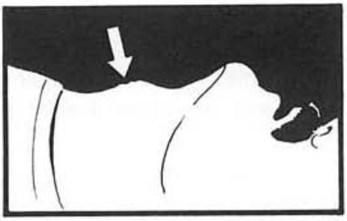

5 *Insert the tube in the incision and push it down to keep the cut open and allow air into the lungs. Once in place, secure firmly with adhesive tape or bandage to keep it upright and prevent it falling out.*

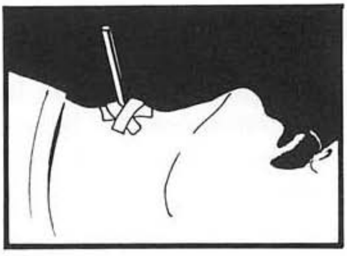

Once the airway is clear, whether natural or artificial, if the patient is still not breathing, artificial respiration must be carried out.

Preventing asphyxiation

Pressure:

Any compression of the chest can cause asphyxiation. A climber who slips and is held suspended by a rope around his chest will find it extremely difficult to breathe. Pass down a rope with a loop (bowline or manharness hitch, see *Knots* in *Camp Craft*) to stand in and relieve the pressure.

An avalanche of snow, or a fall of earth can exert pressure on the chest and make breathing difficult. Adopt a crouched position if possible, with bent arms and elbows tucked well in and this will protect the chest.

If a person is trapped under wreckage with pressure on the chest, and the weight cannot be lifted off, use a lever to lift it and prop securely.

Smoke and gas:

To prevent smoke entering the lungs, place fine mesh material over the nose and mouth to filter it. Smoke can be seen and there is a chance of avoiding it. Gases cannot usually be seen and safety is gained only in fresh air upwind of them, or with a respirator. Casualties must have fresh air.

Lack of oxygen:

Oxygen can be used up in a shelter which lacks ventilation or becomes blocked and sealed by rubble or snow. This is a danger in igloos especially if draughts are sealed. With a stove or fire, not only is the oxygen used, even more seriously, carbon monoxide is produced. Casualties must have fresh air.

Carbon monoxide poisoning:

This gas is DEADLY in confined spaces, but the occupants rarely recognize its presence. The symptoms of carbon monoxide poisoning resemble those of an overdose of alcohol: memory and judgment are impaired, an increase in confidence and a disregard of danger.

Always ensure that you have adequate ventilation, especially when using stoves. Adjust any burning with a yellow flame. Light a candle in your shelter. If the flame gets longer and higher – or in extreme cases shoots to the roof – there is a severe lack of oxygen and it is high time to ventilate. Once again, it is neccessary for casualties to have fresh air.

NOT BREATHING AND NO PULSE

Drowning

Symptoms: Can occur through fluid blockages but generally patient will be in water or have face in liquid. Face, especially lips and ears, livid and congested. Possibly fine froth at mouth and nostrils – it is the froth that is blocking the air passage.

Do not attempt to remove liquid from lungs – you can't. Begin artificial respiration as early as possible. If still in water, support floating body and begin mouth-to-mouth resuscitation after quickly removing weed, false teeth or any other mouth obstructions. On land Holger Nielsen method can also be used.

Electrocution

Symptoms: The cause will usually be obvious. Electrocution may stop the heart and muscle spasms may throw the victim some distance. Electrical burns will be much deeper than their appearance suggests.

Never touch the victim until current is off or contact broken. If an appliance is involved it may be possible to break contact by pulling on insulated cable. But beware of any liquids which will conduct current – victims may urinate. Give artificial respiration and treat for cardiac arrest if necessary before treating burns. TAKE NO RISKS.

Lightning

Symptoms: Another form of electrocution. The victim is usually stunned and falls unconscious. Clothing may catch fire as well as patient having electrical burns, which will be more severe where watches, jewellery, buckles or other metal objects are worn.

Give artificial respiration if necessary and treat burns. Prolonged resuscitation may be needed. Recovery is often delayed.

Poisoning

Symptoms: Poisons which enter the lungs or affect the nervous system can produce asphyxia.

Heart attack

Symptoms: Severe pain in the chest, shortness of breath patient feeling giddy, possibly collapsing to ground, and often anxious. Heavy sweating, irregular pulse, blueness of lips or skin.

If breathing fails give artificial respiration and external heart compression if the pulse stops.

ARTIFICIAL RESPIRATION

Mouth-to-mouth ('Kiss of life')

The fastest and most effective method. Begin as soon as the airway has been cleared. Normal recovery is rapid, except in cases of electric shock, drugs and carbon monoxide poisoning. In these cases nerves and muscles are paralyzed or carbon monoxide has displaced oxygen in the bloodstream. Be prepared to carry on a long time.

If face is injured, or poison or chemical burns are suspect use the Silvester method.

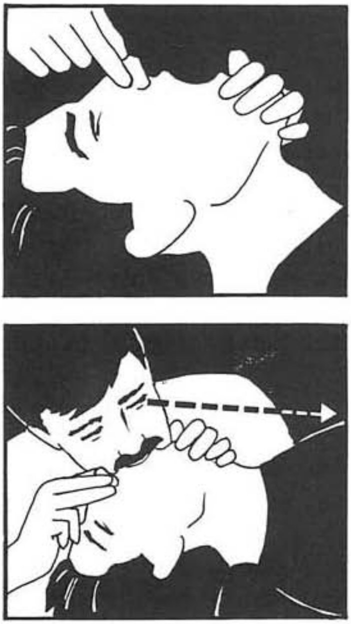

With patient lying on back, hold jaw well open, bending head back (prevents tongue falling and blocking airway). Hold nostrils closed with other hand. Check mouth and throat clear of obstruction. Place mouth over patient's mouth and exhale.

Watch for chest to rise as you blow gentry into patient's lungs. (If chest does not rise, turn him on his side and thump between shoulder blades to remove obstruction.) Remove mouth. Take a deep breath while checking chest falls automatically. You should feel or hear air returning.

REPEAT, as quickly as possible for the first six inflations, then at 12 per minute until breathing is established.

For a child: Do not blow. Exhale normally, or in gentle puffs for a baby. Give first four inflations as quickly as possible. Blowing forcefully into a child's mouth may damage delicate lungs.

Mouth-to-nose: Use if you cannot seal your lips around victim's mouth, holding his mouth closed. For babies, cover both nose and mouth with your lips.

KEEP GOING!

With any form of resuscitation the first five minutes are probably the most critical but, if breathing does not start, keep artificial respiration up for at least an hour. In a group take turns. CHECK HEARTBEAT.

Artificial respiration: with facial injury

Silvester method: Recommended when poisoning or facial injury prevent mouth-to-mouth resuscitation, especially when patient may need cardiac compression (which can be done by the same first-aider).

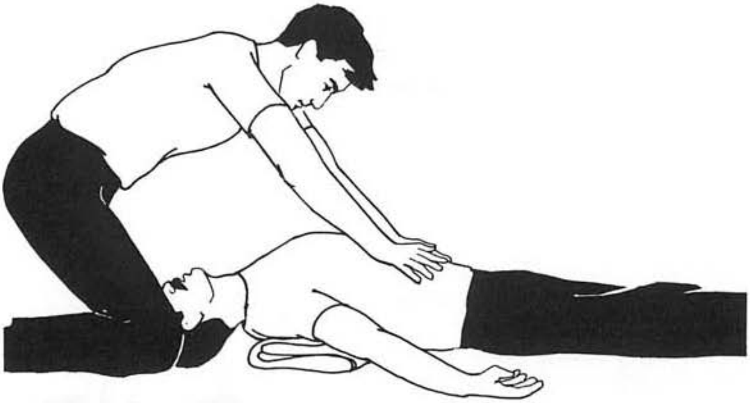

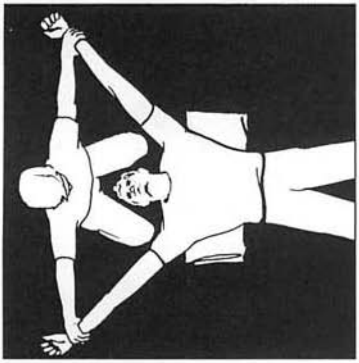

With casualty lying on back, raise shoulders with pad of folded blanket or clothing. Kneel astride casualty's head.

Place hands flat over lower ribs and rock forward to press steadily downwards. Lift casualty's arms upwards and outwards as far as possible.

REPEAT rhythmically about 12 times per minute for adults. If there is no improvement turn patient on side and strike briskly between shoulders to remove any possible obstruction before resuming the cycle.

DON'T GIVE UP! Resuscitation techniques have saved the lives of victims of drowning, hypothermia and electrocution after three hours without spontaneous breathing.

Artificial respiration: face down

Holger Nielson method: This is the technique recommended for resuscitating a drowning victim if mouth-to-mouth is not practicable, or if patient cannot be turned on back. Casualty lies face down; liquids can flow freely from the mouth and will not cause choking.

Place with head turned to one side, arms bent, forehead resting on hands. Loosen tight clothing and ensure tongue is brought forward, mouth clear of weed, mud etc.

Face the casualty, kneeling on one knee at head, placing your hands over shoulder blades, thumbs touching and fingers spread. Perform the following procedure to a count of eight:

1-2-3 *Rock forward with arms straight, producing gentle, even, increasing pressure. (About 2 seconds)*
4 *Rock back, sliding hands to grasp patient's upper arms. (½-1 second)*

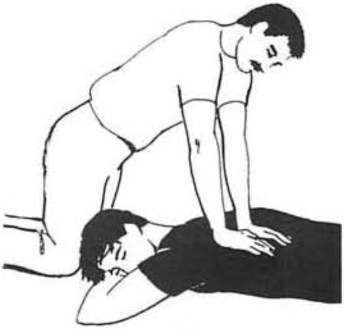

5-6-7 *Pull and raise patient's arms gently by rocking further backwards. (2 seconds) Avoid raising patient's trunk or disturbing head too much.*
8 *Lower patient's arms to ground and slide hands back to initial position. (½-1 second)*

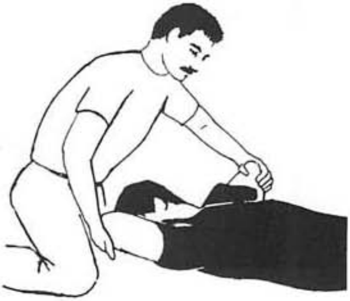

REPEAT 12 times per minute.

NOTE: If the patient's arms are injured, place a folded garment under the forehead and lift under the armpits. This is not a practicable method if ribs or shoulders are badly damaged.

AFTER BREATHING HAS BEEN RESTORED: Place patient in recovery position – after all forms of resuscitation. But NOT in cases of spinal injury.

IS HEART BEATING?

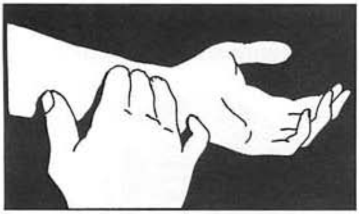

TAKING PULSE AT WRIST

Rest fingers lightly at the front of the wrist, over the radial artery, about 1cm (⅜in) from thumb side at lower end of forearm.

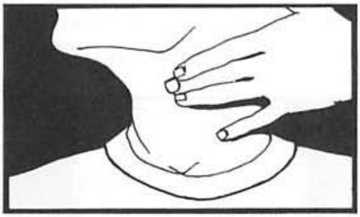

TAKING PULSE AT NECK

Turn face to one side. Slide fingers from Adam's apple into groove alongside. Other pressure points can also be used.

Normal pulse

In a relaxed adult 60–80 per minute (average 72); in young children it is much higher, at 90–140 per minute. Excitement increases rate.

Don't waste precious time – count the beats in 30 seconds and multiply by two. Use a watch with a seconds hand to keep timing accurate and note down result.

If there is no pulse

If you cannot feel a pulse and the pupils of the eyes are much larger than normal, start cardiac compression while artificial respiration is continued. The mouth-to-mouth and the Silvester methods allow both activities to be carried out at the same time.

Cardiac compression

Regardless of which method of resuscitation is used, if there is no pulse and after 10–12 breaths there is no apparent improvement in the casualty's condition, cardiac compression (external heart massage) should be started.

NO TIME TO LOSE!

First place the casualty on a firm surface – let them lie on the ground, chest up. Using the edge of the hand strike firmly on the lower part of the breastbone (the central bone between the ribs). The jarring may start the heart. If there is still no pulse proceed with compression.

Kneel beside the casualty. Place heel of one hand on lower half of breastbone (sternum), the central bone between the ribs. Make sure it is not on the end of, or below, the breastbone. Place heel of other hand over it. Keep rest of hand OFF chest. Keeping arms straight, rock forward and press down 10 times after two consecutive breaths.

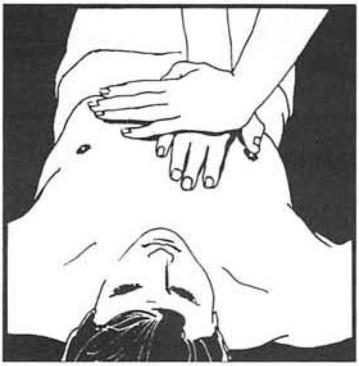

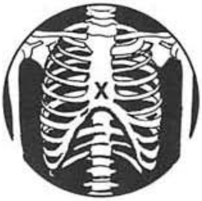

In adults press down about 4cm (1½in). Repeat at least 60 times per minute. Press smoothly and firmly. Erratic or rough pressing could cause further injury.

Infants and children require less pressure and more pushes. For babies and small children light pressure with two fingers is enough at a rate of 100 times per minute. For older children up to ten years, use the heel of one hand only and push 80–90 times per minute.

Note: Check for dilation of pupils of eyes and return of pulse in neck artery which indicate success. ARTIFICIAL RESPIRATION MAY STILL BE NEEDED.

Artificial respiration with compression (Cardiopulmonary resuscitation)

If alone: Use mouth-to-mouth or Silvester methods of resuscitation, giving a repeated pattern of 15 heart compressions followed by two rapid lung-inflations.

If two first-aiders: Give five heart-compressions followed by one deep lung-inflation. Repeat. First-aider giving inflations should also note neck pulse and pupils.

Once breathing has been restored, it must be maintained. Lay the patient in the recovery position (EXCEPT when spine or neck injured) which reduces risk from liquids they may bring up as they recover. Check condition regularly.

If available use ARTIFICIAL AIRWAY to keep airway open in unconscious patients.

RECOMMENDED EQUIPMENT

ARTIFICIAL AIRWAY

Before use inspect both the patient's and artificial airway.

Insert the artificial airway to a third of its length, end pointing to the roof of the mouth, then turn it through 180° to point down the throat.

Once it is in place patient will be able to breathe. A mucus build-up may start coughing and gurgling. Clear it with an extractor.

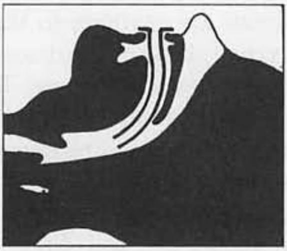

MUCUS EXTRACTOR

Placed down artificial airway, its one-way valve allows mucus to be sucked out without risk of swallowing. One was probably used on you first thing after birth! Without one you can use any tube or straw (preferably sterilized), trying NOT to swallow the mucus yourself.

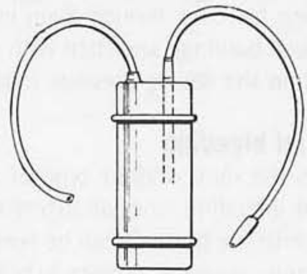

DANGEROUS BLEEDING

An average person has up to 6.25 litres (11pt) of circulating blood. The loss of 0.5 litre (¾pt) causes mild faintness, 1 litre (1¾pt) faintness – with an increase in pulse and breathing, 1.5 litre (2½pt) collapse. More than 2.24 litre (4pt) may even cause death. Immediate steps must be taken to stop the flow of blood. However, once a small haemorrhage has been stopped, blood volume is quickly restored with fluid from the tissues – the resulting slight anaemia is not important. Body fluids must be replaced by giving water, so that the fluid balance is restored.

PRIORITIES

Blood transports life-giving oxygen. When bleeding is coupled with cessation of breathing treat both concurrently. Get the patient breathing AND staunch the bleeding.

Bleeding from veins and capillaries can be stemmed by simple pressure over the bleeding point, with or without a dressing, and minor arterial bleeding may also be controlled with local pressure. Extremity wounds should be elevated above the heart, still applying pressure.

You can use anything to staunch the flow of blood – a hand, handkerchief, blouse – but use the cleanest material possible and apply it rapidly and firmly. There is danger of infection from unsterile material but, if a person is bleeding to death, there is no time to worry. Death is more certain from serious bleeding.

Maintain a continuous, firm pressure for 5–10 minutes and you will stop the bleeding. Resist the temptation to lift and look. If blood seeps through the pad place another on top. Large dressings of absorbent cotton wool (known in the army as shell dressings) are ideal for stopping bleeding. Include them in your medical kit. Secure with the attached bandage and then with a further crepe bandage which will maintain the steady pressure required.

Arterial bleeding

This is the most serious type of bleeding and speed is essential to stop it. Bleeding from an artery comes in powerful, rapid spurts, in time with the pulse. It can be temporarily controlled by compressing an artery where it crosses a bone, against that bone, at pressure points. Precautions must be taken, if at all possible, to prevent the spread of diseases such as HIV, but it is accepted that if you are faced with this emergency you are unlikely to have gloves, etc.

PRESSURE POINTS

These are places where arteries run close to the surface over a bone so that they can be pressed against it, cutting off the flow of blood. Each is effective for arterial blood loss from the area specified. Watch the wound. If blood flow is not immediately reduced, move fingers until it is.

- **Temple or scalp:** Forward of/above ear (a)
- **Face below eyes:** Side of jaw (b)
- **Shoulder or upper arm:** Above clavicle (c)
- **Elbow:** Underside of upper arm (d)
- **Lower arm:** Crook of elbow (e)
- **Hand:** Front of wrist (f)
- **Thigh:** Mid-way on groin/top of thigh (g)
- **Lower leg:** Upper sides of knee (h)
- **Foot:** Front of ankle (i)

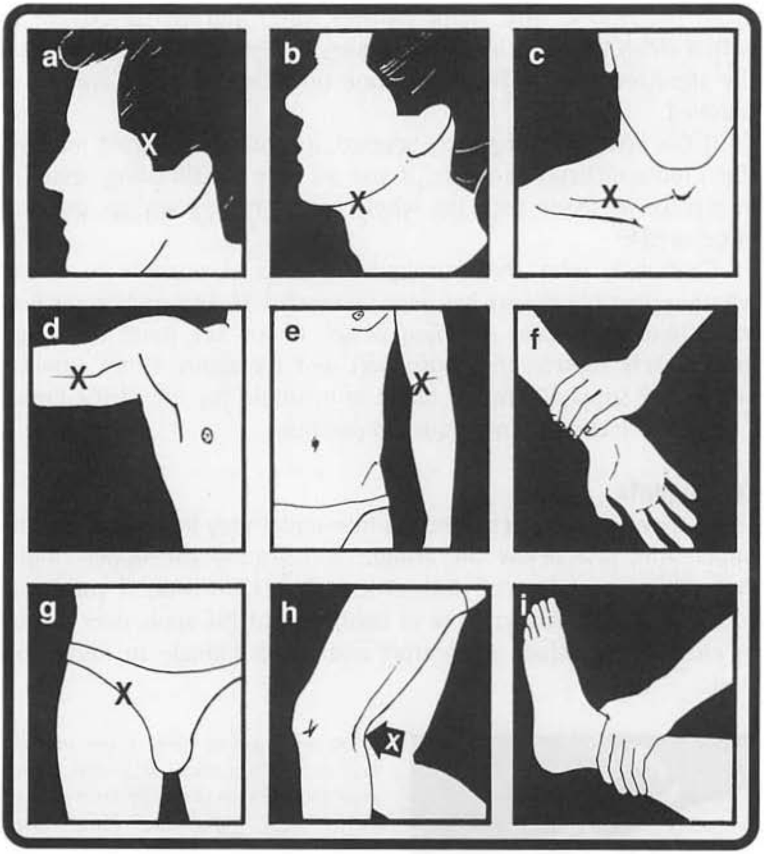

Tying arteries

In the case of major wounds to which a pressure dressing cannot be applied, where access is restricted or a limb partially severed, trace the bleeding artery and tie it off.

This is hazardous for the untrained – arteries often have accompanying nerves and including them in a clip or tie could cause permanent damage, such as loss of function of a limb.

Apply a temporary tourniquet to control the bleeding – but you may have to release spurts of blood to find the artery.

Sterilize a piece of fishing line, thread or fine string by boiling or soaking in alcohol. The contents of a hip flask or a perfume bottle will remove bacteria. Cleanliness is essential. Boil all dressings and bandages. Scrub the hands thoroughly in boiled water, using soap if available.

Clean the wound with sterile (boiled) water and gently explore it with a clean finger to locate the severed artery. Tie it securely with the sterilized thread. This is the one time that an open wound is touched.

If the artery is completely severed, its ends may retract making them more difficult to locate. If you are sure the bleeding vessel is in a piece of tissue then the whole area can be sewn up, using a large needle.

Cautiously relax the tourniquet as soon as possible to check whether your handiwork has been successful. Make sure there is not more than one major bleeding vessel. If you see more bleeding, immediately tighten the tourniquet and try again. Other smaller vessels will supply sufficient blood to maintain the life of the limb.

Do NOT leave the tourniquet in position.

Tourniquets

There are only two places where a tourniquet may be placed: on the upper arm, just below the armpit, and around the upper thigh. Preferably use a piece of cloth at least 5cm (2in) wide. If you have to use anything thinner (wire or cord), you MUST apply over a fold of clothing to reduce discomfort and avoid damage to nerves or flesh.

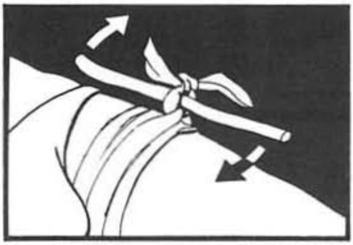

Wrap around limb three times and tie half knot. Place stick, or similar object over knot and tie a double knot over it. Twist stick, tightening band until bleeding stops.

Tourniquet must be tight enough to arrest blood supply but relaxed frequently, so after applying WORK FAST.

Release COMPLETELY when job is done.

WARNING

A tourniquet cuts off the blood flow and if left on too long can cause serious damage – even loss of a limb.

- **ONLY** use a tourniquet on a limb. **NEVER** on head, neck or torso.
- **NEVER** cover a tourniquet. If you have to leave one unattended write TK and time applied on victim's forehead with felt-tip or lipstick.

Check circulation

After bandaging a limb, check frequently that neither toes nor fingers are blue, cold or numb. If they are, loosen the dressing. If you do not GANGRENE MAY DEVELOP, perhaps resulting in the loss of a limb.

For the same reason do NOT use a tourniquet (except when tying arteries). First try direct pressure. Then use pressure points, still keeping direct pressure on the wound. If wound is at an extremity keep it elevated at all times.

RECOMMENDED EQUIPMENT

HAEMOSTAT

This instrument, like a pair of lockable pliers, can be clamped on a blood vessel, stopping the flow and making it easier to tie off. Haemostats have many applications and are worth including in a medical kit. They are ideal for holding needles while sewing – even tough leather can be stitched. AVOID clipping nerves.

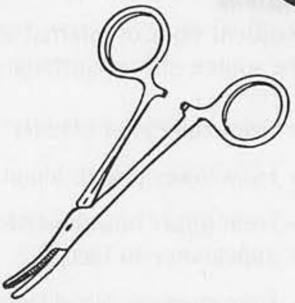

LESSER BLEEDING

Venous bleeding

Bleeding from a vein is not so dramatic as arterial bleeding. The darker venous blood flows more slowly. There is time to select the best available dressing. If blood welling up from a deep wound makes the exact point of haemorrhage difficult to locate, use a large pad and apply pressure over as wide an area as possible. After 10 minutes you will probably be able to secure this pad in place – you can use a tie or strips torn from clothing, the broader the better so that they do not cut into the casualty's flesh.

Capillary bleeding

Capillaries are tiny blood vessels that will stop bleeding by themselves. Do not waste time on capillary bleeding, it is never serious. Deal with essential things first and dress later.

INTERNAL BLEEDING

This serious condition is common after a violent blow to the body, broken bones, bullet or deep penetration wounds. At first there may be little evidence of internal injury, perhaps only slight bruising under the skin. The patient will feel light-headed, restless and faint, and look pale with the skin cold and clammy to touch and the pulse weak but very fast.

Bleeding into the tissues occurs with all fractures and bruises and a fractured thigh may result in the loss of one or more pints of blood at the fracture site.

Symptoms

Subsequent signs of internal bleeding, which also give an indication of the source of haemorrhaging, include:

- From kidneys or bladder: red or wine colouring to urine
- From lower bowel: blood passed with faeces
- From upper bowel: partly digested blood gives black tarry appearance to faeces
- From stomach: blood vomited. If bright red from fresh bleeding.

  If like brown coffee grounds, has been in stomach some time
- From lungs: coughed up blood, frequently as red froth.

Treatment

Lie the patient flat with the legs elevated, making it easier for the heart to pump blood to the head. Keep patient moderately warm but do not overheat – this would result in blood being diverted to the skin. Serious internal bleeding may occur if an organ such as a kidney, liver or spleen has been damaged. The only treatment is nursing care. Hope for early evacuation.

NOSE BLEED

Treat by sitting patient up with head slightly forward and pinch the soft part of the nostrils for five minutes. Encourage patient to breathe through the mouth. They must not sniff. Loosen any tight clothing.

WOUNDS AND DRESSINGS

Open wounds are always a hazard because of the risk of infection by bacteria. Most important of these is the *Tetanus bacillus*, which causes lockjaw. Tetanus immunization is a wise precaution for everyone, and essential for outdoor adventurers and travellers.

Wounds caused by glass, metal or burns are usually clean wounds. All foreign bodies must be extracted. This is usually left for trained medics – but should be done in a survival situation. Sterile haemostats or tweezers are the best tools for the job. A wound that has been in contact with the ground or dirty clothing MUST be cleaned and ALL dead flesh removed.

Cut away the clothing from the wound site, clean the vicinity and irrigate the wounds to wash out all the dirt. Clean a wound from the centre outwards, do NOT swab from outside in. Dry and apply a clean dressing. Immobilize the wound in a position that is comfortable.

Dressings should be changed if they become wet, omit an offensive smell, or if pain in the wound increases and throbs, indicating infection.

Local infection can be treated by soaking in hot salty water, or applying poultices. Poultices will draw out the pus which accompanies infection and help to reduce swelling. Anything that can be mashed can be used for a poultice: rice, potatoes, roots, shredded tree bark and seeds are all suitable. Clay can also be used. Boil them up and wrap in a cloth. Apply to the infected area as hot as can be tolerated – do not risk scalding. The human body has a tremendous capacity for resisting infection if given proper rest and nutrition.

Applied heat also aids healing. A warm rock wrapped in cloth can be used to provide it.

SOAP IS ANTISEPTIC

Soap is an excellent antiseptic and good for washing wounds. Use boiled water to wash your hands before cleaning a wound. Wash the wound in boiled water – if none is available use urine. Urine is a sterile fluid and will not introduce infection. It also has the slight beneficial effect from the uric acid which helps to clean the wound.

Stitching wounds

Minor wounds can be closed by suturing, if there are no medics available to do it. (Haemostats are again useful for this.) This is recommended where a clean cut caused by a knife needs closing and for facial injuries which interfere with eating or breathing.

First clean the wound thoroughly, then stitch across it, or use butterfly sutures (from your survival kit) which can be applied without any special skills. Some tribal peoples use the Fire ant to do the job, making it bite across the wound and then snapping the head off so that its mandibles hold the wound together.

STITCHES

With a sterilized needle and thread or gut, make each stitch individually, beginning across the mid-point of the wound.

Draw the edges together and tie off thread, then proceed working outwards.

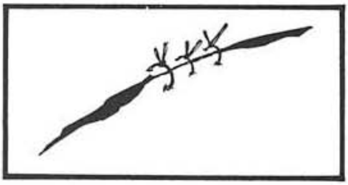

ADHESIVE SUTURES

Use butterfly sutures or cut adhesive plasters in butterfly shape. Draw edges of wound together. Apply plaster to one side of wound, close up as much as possible and press down other side of plaster.

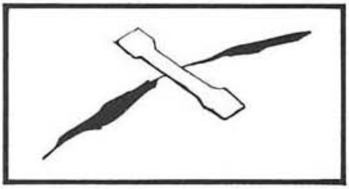

If the wound becomes infected – red, swollen, tense – remove some or all the stitches to let pus out. Leave to drain.

Open treatment

'Open treatment' – covering with a dressing but not suturing – is the only safe way to manage survival wounds, apart from those mentioned. If unable to thoroughly clean a wound it must be left open to heal from the inside. It will form infection resistant tissue, recognized by its moist red granular appearance – a healthy sign in any wound.

Despite precautions there will always be some degree of infection. Deep wounds may have to be drained and occasionally it may be beneficial to open an abcess (an accumulation of pus) and insert sterilized loose packing in the form of a bandage or a cloth. Leave a tail hanging out, preferably with a safety-pin in the end.

Allow wound to drain for a few days. If lancing or reopening a wound, sterilize your blade to prevent the introduction of new bacteria. Packing keeps wounds covered, but allows them to drain as they heal from inside. Reduce packing as healing progresses, until able to remove it all and cover with a dressing.

Chest wound: One of the greatest dangers with a chest wound is having the chest collapse. This is called a 'sucking chest wound'. In such wounds the chest sucks air through the wound into the space around the lung, rather than into the lung. The hole must be sealed to enable the casualty to breathe properly. The waterproof wrapper of a field dressing (or a plastic film from a cigarette packet or aluminium foil preferably coated with petroleum jelly) can be placed over the hole, with the clean inside against the wound, and taped in place. Tape is applied to the sides and top of the wrapper, leaving the bottom edge free. This prevents air being sucked in, while allowing any that has already entered to be forced out as the casualty exhales.

Abdominal wounds: These are serious because of the danger of damage to the internal organs and of internal bleeding. No solids or liquid must be given. Thirst can be relieved by using a damp cloth to moisten the lips and tongue of the patient. If the gut is extruded it must be covered and kept damp. Do not attempt to push it back into place. (That would make difficulties for a surgeon after rescue.) If no organs extrude dress and bandage firmly.

Head injuries: Head injuries pose problems of possible brain damage and wounds may interfere with breathing and eating. Ensure that the airway is maintained and that the tongue has not fallen to the back of the throat. Remove any false or detached teeth. Control bleeding. The conscious casualty can sit up, but the unconscious patient must be placed in the recovery position, provided that no neck or spinal injuries are present. (See *Fracture of the skull*).

Amputation: If a person is trapped by a limb in a burning wreck, or similar situation, drastic action may be required to extricate them. You may have to trade a limb for a life, if otherwise they would be burned to death, for example.

Time is critical if you are working under pressure with increasing risk to yourself and the patient, from fire or other danger. But you must still take as much care as possible.

If a limb is trapped, cut as close to the wound site as possible – which will probably mean cutting through bone. A saw of some sort is required. The flexible saw in your survival kit was originally a surgeon's tool and will do the job. If no saw is available, or amputation is because of gangrene, sever at the nearest joint.

First apply a tourniquet and be ready to tie off arteries as they are exposed, or use haemostats if available.

Make an incision in the skin and into the underlying tissue. Allow the skin to retract, then sever the muscles. They will retract too, leaving the bone or joint exposed. Saw through the bone or cut through the joint. Tie off arteries, if not already done, but leave the stump open to allow for drainage. Apply a light bandage to protect the stump.

Traumatic amputation

If a limb is torn off as a result of accident there is very little bleeding. The damaged muscle in the wall of the artery goes into spasm and shuts off the artery. You will be able to examine the wound and tie off all the arteries exposed.

USE OF ANTISEPTICS

If antiseptic is available use it for cuts and abrasions. Do NOT use antiseptics on deep wounds. They cause further tissue drainage. Clean the local area with antiseptic, but wash the wound with boiled water.

BURNS

Burns, which are a common injury in aircraft accidents, cause severe pain and fluid loss. Victims are VERY susceptible to shock and infection.

The skin area of the body affected can be used to assess the possibilities. Burns extending over 50 per cent or more of the body are usually fatal, if extensive medical facilities are not available. As a rough guide to area affected:

Head = 9 per cent
Arms = 9 per cent each
Front of torso = 18 per cent
Back of torso = 18 per cent
Genital area = 1 per cent
Front of legs = 9 per cent each
Back of legs = 9 per cent each

Extinguish burning clothing

It is vital to extinguish burning clothing without fanning the flames. Most people will instinctively run away from the danger but, as they do so, the draught will encourage the flames to increase.

Get the victim down on the ground and roll over, if possible covering him or her with a blanket, poncho or sleeping bag. If necessary roll on the victim yourself to extinguish the flames (this is when you find out who your real mates are!).

Remove the victim's smouldering clothing and any constricting garments, jewellery, etc, which may become tighter if swelling occurs. It is important to remove smouldering clothes at once, for they retain heat and can be hotter than the flames themselves.

Reduce the temperature

Drench the burned tissues with water to cool them. Ideally submerge under slowly running cold water for at least 10 minutes.

Do NOT even think of using anything to soothe the burns. Neither antiseptic, butter, grease, lard, calamine lotion, Vaseline nor anything like them should be applied. Resist the impulse! Cooling should continue until it produces no further relief and withdrawal from water does not lead to increase in pain.

After the initial cooling leave burns alone except for applying dressings, as dry and sterile as possible, to resist infection. Put dressings between burned fingers or toes before bandaging to prevent them sticking to each other.

Later, hardwood barks such as oak or beech, which contain tannin can be boiled up in water. When cool it can be applied to the damaged areas to soothe burnt flesh.

Give fluids

Fluids must be given to replace those lost. Give small cold drinks frequently. If possible add a half teaspoonful of salt – or even better a pinch of bicarbonate of soda – to a pint of water. If you have no salt, give the patient small amounts of boiled animal blood to drink.

Types of burns

Deep burns are charred or white in appearance, possibly with bone and muscle visible. By a merciful act of nature these burns are painless as the nerve endings have been destroyed. Superficial burns are very painful and, if covering a large area, fluid loss can create even greater shock than for deep burns. The skin will blister but these blisters should NEVER be deliberately burst. If there are burns about the face and neck ensure airway is clear.

Scalds:

Are wet burns caused by hot liquids, whether water, steam, oil or even a poultice. Treat as for dry burns.

Mouth and throat burns:

Possible from inhaling flame or hot gases, accidentally drinking from a very hot vessel – or swallowing very hot liquids or corrosive chemicals. Give sips of cold water to cool. Swelling in the throat may affect breathing and artificial respiration may be required.

Eye burns:

Spitting fat or corrosive chemicals may burn the eyeball. Hold the lids open and pour plenty of water over it to wash out chemicals. Tilt head so that chemical is not washed into mouth or nose or into other eye if only one is affected.

Chemical burns:

Use copious amounts of water to dilute and wash off chemicals. Remove clothing that may retain corrosive substances. Do not attempt neutralizing acid with alkali or vice-versa; chemical interaction will produce more heat. Continue to treat as for heat burn.

Electrical and lightning burns:

Check respiration. Treat as for heat burns. Take no risks if current still live.

BURNS AND SHOCK

All except small or superficial burns and electric burns are likely to result in shock, the degree dependent upon the amount of plasma fluid lost. Flooding with cold water in the case of extensive burns could increase shock, but that must be weighed against dramatically reducing tissue damage. Keep up cooling for at least 10 minutes.

FRACTURES

Accidents may produce sprains and bruises, broken arms and legs and compression fractures of the back. Broken bones should be examined early, before swelling complicates location. Always look before touching and moving an injured person. However, treat asphyxia and bleeding first and do not seek to set a fracture in an urgent emergency when there are others with more pressing injuries to treat. Immobilize before moving, if possible, and finish treatment later.

There are two types of fracture: open and closed. In an open fracture the bone may push through the skin or there may be a wound leading down to the fracture. In these fractures infection can gain direct access to the bone and they must be treated very seriously. If the limb is grossly distorted by the fracture it must be straightened before splinting. It will be painful. If the patient is unconscious do it straight away.

Fractures which neither penetrate the skin nor are exposed to the air are known as closed fractures.

Symptoms

- Pain, usually severe, aggravated by attempted movement of injured part.
- Tenderness, even with only gentle pressure.
- Swelling (caused by loss of blood into the tissues), later showing discoloration or bruising.
- Deformity: apparent shortening of a limb, irregularity either visible or to touch, unnatural movement, limp and wobbly limbs – compare a suspect area with the unharmed opposite side.
- A grating sound when limbs are moved (do NOT move limbs deliberately to check for this).

Compare injured limb with opposite unharmed limb.

Reduction

If medical help is expected, immobilize closed fractures and leave them for professional treatment, but – if none can be expected – 'reduce' them as soon as possible after injury, before a painful muscle spasm sets in.

Apply traction (a slow, strong pull – not a tug) until the overriding edges of fractured bone are brought into line. Check alignment with the other limb. Then splint and immobilize, keeping up the traction to ensure it does not slip back. A splint will now be needed and can be made from all sorts of material – ski sticks, branches, ships spars, parts of wreckage, driftwood, rolls of newspaper, etc.

NOTE: Separate all hard splint material from the skin with padding – moss is useful for this – or pressure sores may develop.

IMMOBILIZATION

Immobilize the whole length of the limb. Use slings to support bent-arm fractures. If no splint is available, or to increase immobilization, strap injured limb to uninjured limb or to body. Insert padding in any natural hollows to keep limbs in position. Secure firmly above and below fracture and below nearest joints. Tie with any soft materials available. Place all knots on same side, giving easy access, and use reef knots or reef bows.
CHECK CIRCULATION PERIODICALLY.

Sling materials

Triangular bandages are excellent for making slings (See *Bandaging*) – but slings can also be improvised from pieces of clothing, belts, etc, as shown in these illustrations. Do NOT tie splints directly over injury or allow knots to press against the limb.
CHECK CIRCULATION.

- FRACTURE OF ARM BELOW ELBOW
- FRACTURE OF HAND OR FINGERS

Place sling (in this case a long-sleeved sweatshirt) between arm and body. Immobilize from elbow to mid-fingers with a padded splint. Take one arm of shirt behind head and tie to other on opposite side to injury. Knot below elbow to stop slipping.

Arm is elevated to prevent severe swelling.

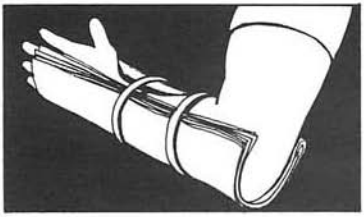

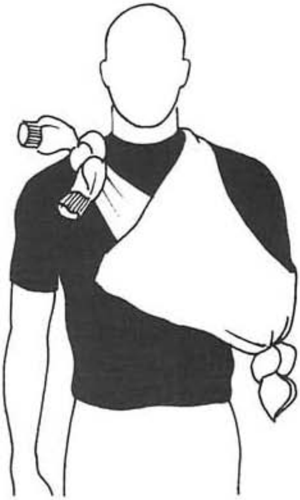

FRACTURE AT ELBOW

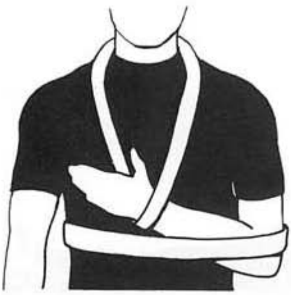

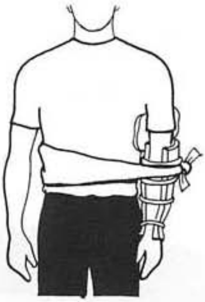

If elbow is bent: *support in narrow sling. Bind across upper arm and chest to prevent movement. Check the pulse to ensure that an artery has not been trapped. If no pulse try straightening the arm a little to see if it will return. If there is no pulse medical aid is urgently needed.*

If elbow is straight: *do not bend it. Place pad in armpit and strap arm to body or place padded splints either side of arm.*

FRACTURE OF UPPER ARM

Place pad in armpit. Splint from shoulder to elbow on outside of arm.

Narrow sling at wrist. Bind arm and chest.

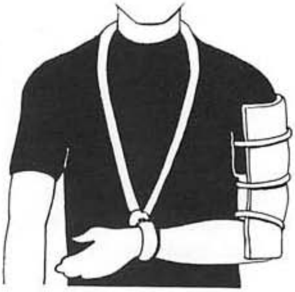

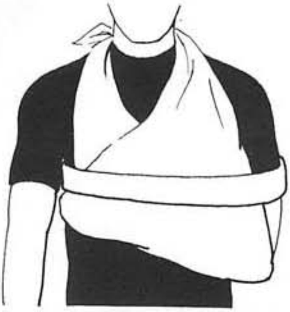

- FRACTURE OF SHOULDER BLADE
- FRACTURE OF COLLARBONE

Make sling to take weight off the injured part.

Immobilize with bandage across arm and body.

WARNING

CHECK periodically that circulation is not impeded. Blue or ashen fingers and toes are the obvious danger warnings that straps and dressings are too tight.

For any fracture of the thigh or lower leg a figure-of-eight bandage should be applied, binding the feet and ankles of both legs. This controls rotation and counters shortening.

FRACTURE OF HIP OR UPPER LEG

Place splint on inside leg and another from ankle to armpit. Use a stick to push tying bands under hollows of injured leg.

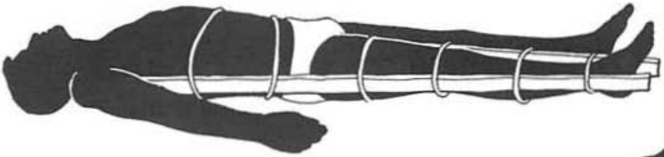

If no splints are available pad between legs (a folded blanket is ideal) and tie to sound leg.

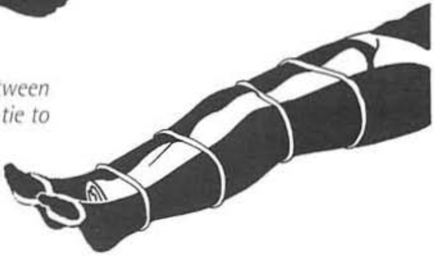

FRACTURE TO KNEE

If leg is straight: place splint behind leg. Apply cold compress to knee (ice if possible).

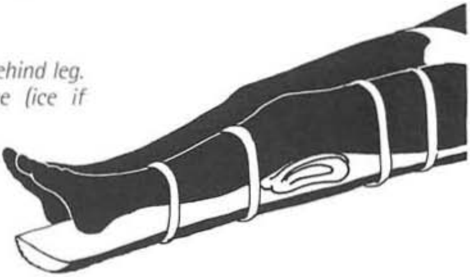

If leg is bent and you cannot straighten: bring legs together, place padding between calves and thighs and strap in those places. This can only be a temporary measure until proper medical help is available. If help is out of the question, the leg must be made as straight as possible.

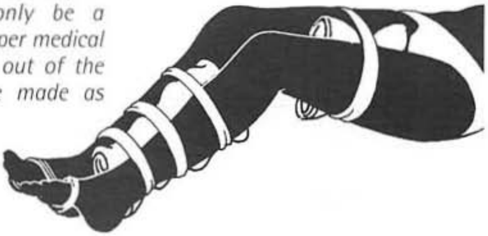

FRACTURE TO LOWER LEG

Splint from above knee to beyond heel, or pad between legs and tie together (see Fracture of hip*).*

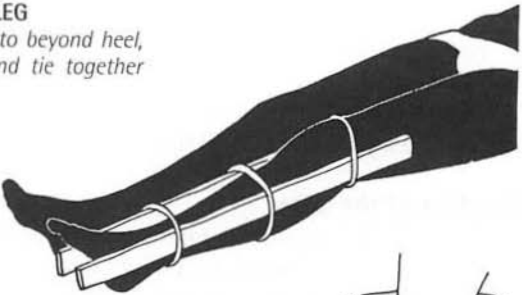

FRACTURE OF ANKLE OR FOOT

Not usual to splint. Elevate foot to reduce swelling. Immobilize with pillow or blanket folded around ankle and under foot, strapped twice at ankle and once under foot. Alternatively: if no wound, leave shoe or boot to provide stability. Do NOT allow casualty to put weight on foot.

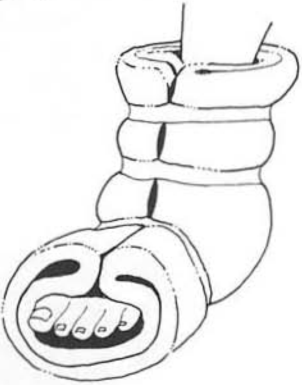

FRACTURE OF PELVIS

Symptoms include pain in the groin or lower abdomen. Pad between thighs. Tie at knees and ankles. Place pillow support beneath bent legs and strap to flat support (door, tabletop, stretcher) at shoulder, waist and ankle.

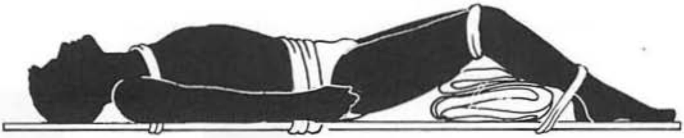

OR

Place padding between legs. Bandage around feet, ankles, knees and with two overlapping bandages over pelvis.

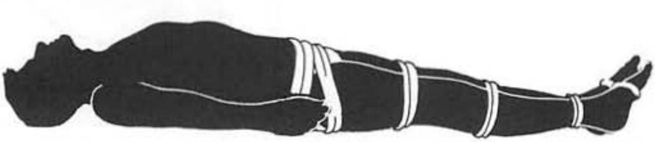

Fracture of the skull

Blood or straw-coloured fluid seeping from the ear or nose may indicate a fractured skull. Place the casualty in the recovery position, with the leaking side down. Allow the fluid to escape – bandaging or plugging could cause a build-up in the skull and pressure on the brain.

Keep a careful check on the casualty's breathing. Completely immobilize and keep comfortable.

Fracture of the spine

Must always be suspected when a casualty complains of pain in the back or neck, with possible loss of sensation in the lower limbs. Test for 'feeling' by gently touching the limb. Ask casualty to move fingers and toes. Warn the patient to lie still. If medical assistance is expected and the area is safe, immobilize by placing soft but solid objects such as luggage or padded rocks to prevent movement of head or body.

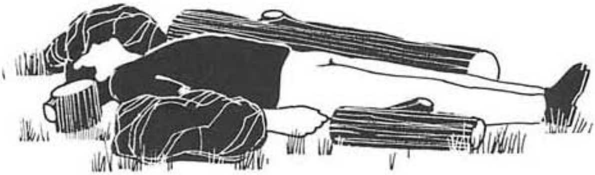

Fracture of the neck

If a neck fracture is suspected, it is ESSENTIAL to immobilize it with a cervical collar, or to place a bag of earth or something similar against either side of the neck to prevent movement.

CERVICAL COLLAR

Make from rolled-up newspaper, folded towel, car mat etc. Fold to about 10–14cm (4–5½in) wide – distance from top of breastbone to jaw. Fold in edges to make narrower at back than front. Overlap around neck. Secure in place with a belt or tie.

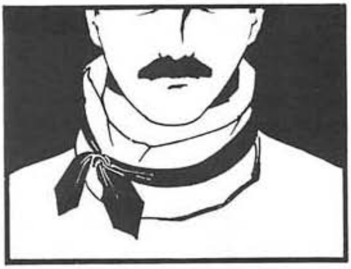

Whilst the patient's shoulders and pelvis are firmly held, place pads of some material between the thighs, knees and ankles. Tie the ankles and feet together with a figure-of-eight bandage. Apply broad bandages around the knees and thighs. Keep completely immobilized and hope for early rescue.

SPRAINS/DISLOCATIONS

Sprains

A sprain occurs at a joint and is caused by the wrenching or tearing of tissues connected with the joint. The symptoms are pain, swelling and the later appearance of bruises. If in doubt whether a sprain or a break, treat as a fracture.

Sprains are best exercised through a full range of movements, but NOT put under very painful stress, which could cause permanent damage. Bathe sprains with cold water to reduce swelling. Support with a bandage – crepe if possible (it should not constrict). Elevate the affected limb and rest completely.

If you sprain an ankle and have to keep on walking keep your boot on. If you take it off, the swelling will stop you putting it back on. Left on, the boot acts like a splint.

Dislocations

Dislocations are usually caused by a fall, blow or sudden force applied to a joint which pulls it apart. There is pain and obvious deformity, often with one end of the bone clearly felt under the skin. There is no grating sound, for the bone ends are not usually damaged. Muscle spasms 'fix' the bone in position, making it very painful to replace. Shoulders are especially prone to dislocation.

Dislocated shoulder: Take off your shoe and put foot in patient's armpit. Pull on arm.

An alternative, but more risky method, flexes the elbow at a right-angle and uses it as a lever.

Support arm with a sling and immobilize with a bandage across chest and rest completely.

Dislocated finger: Replace by pulling on finger, then gently releasing it so that the bone slips back into place. It helps if someone else can hold the patient's wrist.

Try this only gently with the thumb. If it does not work first time on the thumb, leave well alone or you will do more damage.

Dislocated jaw: Usually caused by a blow – though sometimes by yawning! Place pad of cloth over lower teeth on each side. With patient's head resting on a firm support, press downwards on these pads with your thumbs, simultaneously rotating the dislocated side of the jaw backwards and upwards with your fingers. It should snap into place.

Bandage around the head and under jaw and leave supported for two weeks. Feed soft foods.

SHOCK

Shock can kill. Its prevention and treatment must be a major objective of first aid.
The signs of all types of shock are:

- State of collapse
- Extreme pallor
- Cold sweaty skin
- Feeble but rapid pulse
- Excitement and apprehension

Nervous system: This initial shock produces an acute slowing of the heart with dropping blood pressure. In severe injuries, it can stop the heart. Pain and excessive nerve stimulation are the main causes.

Shock from blood loss: Delayed shock may occur a few minutes to many hours after an injury. It is brought on by blood loss, burns, abdominal injuries and continuing considerable loss of fluid due to vomiting or diarrhoea. Blood poisoning from injuries and infection produces another kind of shock.

Treatment

Lay the survivor flat and elevate the legs. Loosen tight or restrictive clothing around the neck, chest or abdomen. Rest and reassure.

Do NOT give liquids. Maintain body heat but do not add heat – warming the surface of the body will draw blood away from the internal organs which need it most.

Your attitude and actions are very important in treating shock. If you appear calm and in charge of the situation the patient will feel cared for and respond. Stay with them if you can, do not leave a shock victim on their own. Holding a hand is calming and reassuring and a moist cloth wiped occasionally across the forehead often helps.

Stand by to give mouth-to-mouth resuscitation and cardiac compression if the patient's breathing or heart stops. Treat all injuries and relieve pain with drugs if available. Shock can take a long time to pass, do NOT move unnecessarily and encourage rest.

BANDAGING

Triangular bandage

A triangular bandage, with its short sides not less than 1m (3ft), is a versatile dressing. Use for slings, or folded to make a wide range of bandages and supports.

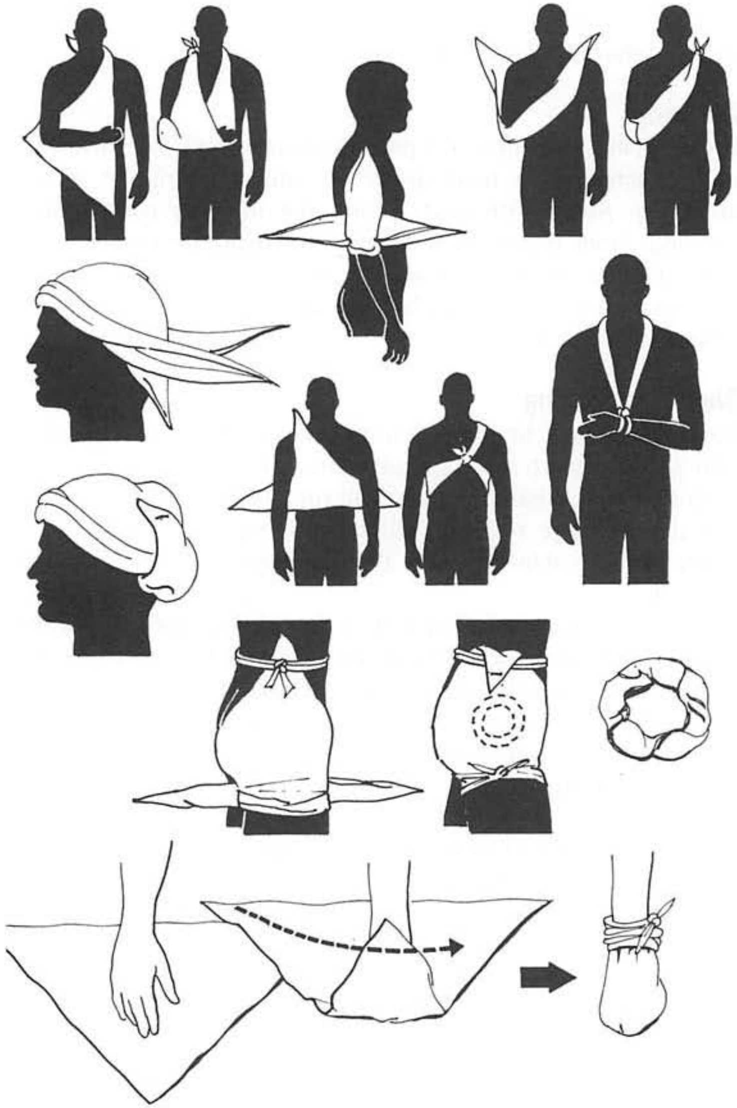

Roll and improvised bandages

Roll bandages are usually of open weave gauze but those of crepe material (which stretch) are easier to apply, less likely to come loose and apply pressure more evenly. Adhesive tapes can be useful for fixing dressings to areas that are difficult to bandage. Some people are allergic to them – in which case use the tape to fix bandage to bandage.

All kinds of material, especially clothing, can be used to improvise bandages and supports, but they should never exert more pressure than is required, nor cut into the flesh.

Dressings

Dressings usually consist of a pad of cotton wool covered with gauze and attached to a bandage or an adhesive strip, in a sterile wrapping. Apply without touching the dressing pad. Improvise dressings from the cleanest material available. Do not use cotton wool directly on an open wound. It will adhere to the surface. Change dressings when they become wet, smell, or throbbing of the wound indicates infection.

Simple bandaging

Bandages should be applied firmly enough to stop slipping, but NOT so tight as to interfere with circulation or cause pain. Even with improvised bandages you will find application easiest if you roll the bandage to begin with. Unrolling the bandage as you apply it will help to keep the bandage smooth and evenly applied.

Always begin bandaging with a firm oblique turn to anchor it. Each turn should overlap the previous one by two-thirds, with the edges lying parallel. Tuck in the ends of bandages below the last layer and secure with a safety pin or adhesive tape, or split the end, take ends in opposite directions and tie in a reef knot away from the wound.

- **Hand:** Begin at wrist. Carry bandage over back of hand, around fingers (just below finger nails), across the palm and around the wrist. Repeat until the hand is covered and the bandage secure.
- **Foot:** Begin at ankle to prevent slipping. Follow technique for hand bandaging.
- **Forearm or foreleg:** Begin at the lowest point. Work upwards.

- **Elbow or knee:** Begin with a turn around the joint then alternately above and below it.
- **Upper arm or leg:** Apply as a figure-of-eight spiral. If bandaging begins over knee or elbow simply carry on upwards in this manner.

BE AWARE

- Do not join bandages with knots. If you need to use separate strips, bind new one over that already applied to hold it in position.
- Always tie a finishing knot or bow over the uninjured side or uninjured limb, or the casualty will be uncomfortable.
- Use reef knots or bows which are easy to untie. Try to keep knots or fastenings on the outside so that they are easily accessible to remove or change dressings.
- Check regularly that bandages are not too tight. Look for any sign of blueness at extremities that would indicate restricted circulation.
- Pass bandages under a casualty making use of the natural hollows of the body: neck, waist, crotch, knees.

MINOR AILMENTS

In a survival situation even the most minor ailment should NOT be ignored. Small problems become larger if they are untreated. Anything which puts strain upon the body will make you less able to deal with other problems and stresses.

Do NOT be tempted to deal with corns and callouses by cutting them out. You risk opening up a channel for infection which could be worse than the corn.

Blisters

Usually a problem on the feet, though hands not used to survival tasks and tools may blister too. Reduce the risk by wearing gloves or wrappings on the hands. Choose tool handles carefully and shoes that fit perfectly! Badly fitting boots can skin you. Break them in by soaking in water and rubbing in oils to make them supple.

One of the main causes of foot blisters is socks falling down and rucking. This is frequently a problem after wading through water.

Pull them up and, if necessary, tie them to the top of your boots. It is preferable to wear two pairs of socks – an inner of nylon, next to the skin, and an outer pair of wool. This stops them from slipping.

Treatment: Wash the area of the blister. Sterilize a needle. Pierce the blister near its edge. Gently press out fluid. Cover with cloth and tape or bandage into place.

Objects in the eye

First inspect the eyeball and the lower lid, pulling it downwards to see the inside surface (ask the patient to look upwards). Remove any foreign body with a moist corner of cloth.

If you can see nothing there, and the problem is under the upper lid, you may be able to brush it out by pulling the lid down over the lower lashes and letting them brush it out. If that does not work, grasp upper lashes between thumb and index finger and pull lid up. If doing this for someone else, it is best if you place a matchstick or small twig over the lid and fold the lid back over it. Get the patient to look down. Inspect both eye and underside of lid. Remove with a moist corner of cloth, a clean fine brush or even a feather. If doing this to yourself, look in a mirror, rather than try to work by touch.

Earache

Unless due to an infection this is usually due to the pressure of wax on the eardrum – but it can be unbearable. Warm up a few drops of any edible oil available, pour it into the ear and plug with cotton wool. The heat alone will be soothing and the oil will soften any wax.

Toothache and tooth loss

Usually caused by a cavity in a tooth or by a filling falling out, exposing a nerve, toothache is a problem you can well do without in a survival situation.

Treat it by plugging the hole to cover the nerve and the pain will subside. The resin from a pine or similar tree will help. Scar the trunk and gum will ooze out. Soak up a small amount on cotton wool and use this to plug the hole.

If a tooth is knocked out in an accident, put a cloth pad on the empty socket and apply pressure by biting down on it to stop bleeding.

Take care of your teeth. Use splintered soft woods and plants for brushing. Strip the inner core of parachute cord to use as dental floss.

PAINKILLERS

Morphine

Morphine is the standard accepted analgesic carried as part of the medical kit of planes and ships and by the military. A very powerful painkiller, usually given by injection, it is excellent for treating casualties in great pain and minimizes the effect of shock, but it has several side-effects.

It depresses the breathing, indeed may stop it altogether. It causes nausea. It should NOT, therefore, be given to cases already likely to be affected in these ways.

Morphine comes in tubonic ampoules which contain 10–15mg. The needle is fixed and can be self-injected. Choose the largest muscle available for the injection. If self-administered, the upper quadrant of the thigh. If injecting a patient, the upper outer quadrant of the buttocks.

USE MORPHINE FOR:

- Fractures
- Amputations
- Serious burns
- Abdominal injury of a straightforward perforation by a sharp object

DO NOT GIVE MORPHINE TO:

- People with respiratory difficulties, such as head and chest wounds
- Snake-bite victims (snake venom affects the respiratory system)
- Children or pregnant women
- Casualties who have lost a lot of blood
- Crush type injuries with suspected internal damage

WARNING

MORPHINE CAN BE LETHAL

An overdosage of morphine will kill. Do NOT repeat for at least three hours. Once a patient has been injected record the time and dosage, preferably on the forehead of the parent, and pin the empty syrette to their collar.

Temgesic tablets

These are an alternative to morphine. They are slower to take effect but can be used when morphine cannot, when there are abdominal wounds for instance. They are placed under the tongue and should not be chewed or swallowed. Give 1 tablet for moderate pain, 2 when severe, not more than 4 times a day. Can produce drowsiness and hallucinations. Do NOT give to children or pregnant women.

Drugs for minor pain

Aspirin and Paracetamol are the best drugs for general pain relief. Aspirin will also help to control fever in 'flu-like illnesses. Do NOT give aspirin to people prone to indigestion, stomach ulcer, etc – it can cause irritation to, and bleeding from, the lining of the stomach.

MOVING THE INJURED

Anyone with injuries to the spine or lower limbs should be carried on a stretcher if possible.

A stretcher can be improvised by passing two poles through pieces of sacking, heavy plastic bags or clothing – jerseys, buttoned jackets, shirts etc. Or from blankets folded around one pole, then wrapped under and over another. Or from doors and tabletops. If no poles are available roll in the sides of a blanket and use the rolls to give a firm grip when carrying. Always test an improvised stretcher with a fit person before using it for the injured.

REMOVING FROM DANGER

You may have to move even a person with spinal injuries, if they are in further danger. If three or four people are available roll the patient on to an improvised stretcher. Do not bend or twist. One person should be responsible solely for maintaining stability of the head and neck. Another holds the shoulders. If there is no stretcher or board roll onto a blanket or coat. Support the head and torso steady even if the legs drag.

If working alone do not attempt to turn casualty over. Pull by the shoulders if face down, by the ankles if face up, in the direction in which the body is lying. Do not twist. On rough ground or steps drag from behind, pull by shoulders, resting head on your forearms.

Loading a stretcher

If the patient is on a blanket, or can be rolled on to one, the edges of the blanket can be rolled to give a firmer grip. Methods of lifting without a blanket depend upon the number of helpers. When lifting with others, agree signals first for synchronizing movements.

WITH 4 PERSONS

Three lift from same side, C supports head and shoulders. D hooks fingers with adjoining hands of B and C to aid lift. A, B, C support while D places stretcher in position. D helps lower casualty.

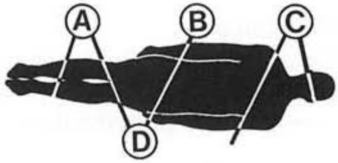

WITH 3 PERSONS

Place stretcher at patient's head. C lifts at knees. A faces B, locking fingers under shoulders and hips. Move casualty from foot of stretcher to over it.

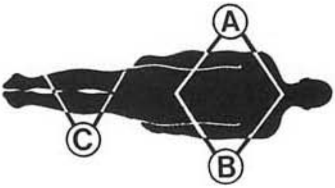

WITH 2 PERSONS

Both stand astride casualty. B links arms beneath shoulders. A lifts with one hand beneath thighs the other beneath knees. Both move forward to above stretcher. Use this technique in narrow spaces, even if more help is available.

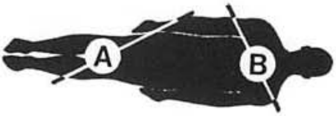

Lifting on your own

If help is available do NOT attempt to move an injured person on your own. If there is no alternative, relate the method to the weight and condition of the casualty – and the distance to be carried. Do not choose a method you cannot sustain. Dropping the patient will not only risk worse injury but will reduce the victim's trust in you. If the victim is too heavy to lift, drag them on a blanket or a coat.

Cradle: Suitable for small children or the very lightweight. Lift with one arm beneath the knees, the other around the shoulders. With some weights it will be easier to support in a sitting position.

Crutch: Place and hold casualty's arm around your neck (not an injured arm). Put your arm around casualty's waist. Give extra support by grasping clothing at hip.

Pickaback: Crouch while casualty puts arms around your neck. Lift legs on either side of your body. Casualty must be conscious and injuries must permit them to maintain hold onto your shoulders or around your neck.

Fireman's lift

Not a suitable method with a heavy casualty. With conscious person help them to an upright position. If on a chair allow them to fall forward over your shoulders. Otherwise, the method is as for the unconscious.

If casualty is unconscious: *Place face down. Kneel at head. Slide hands under shoulders.*

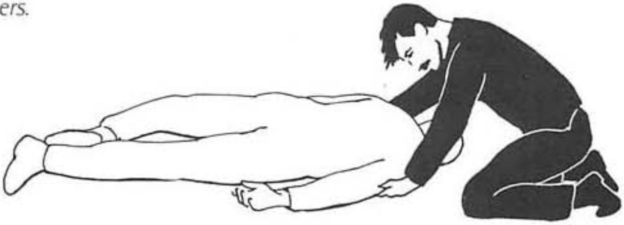

Lift casualty under armpits to kneeling position, then to upright. Raise casualty's right arm (with your left hand).

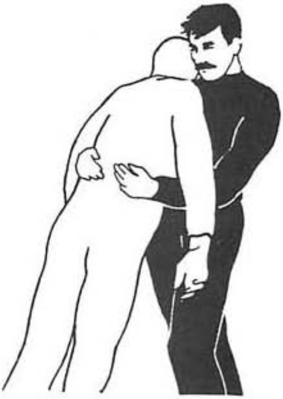

Alternatively: *With casualty on back, lift from behind. Stand astride casualty, facing head. Lift under armpits, to kneeling position. Then drag back to straighten legs. Manoeuvre forward to standing and support under armpits. Raise casualty's right arm and quickly move under it, release it and end up face-to-face, still supporting him. Raise his right arm (with your left hand).*

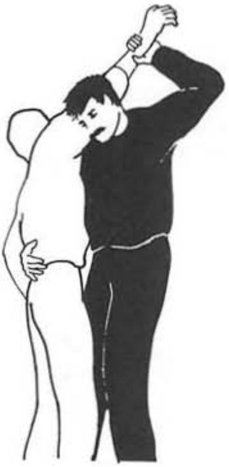

If casualty is conscious: *Bend to place your head under casualty's arm (at same time lifting it over your head). Bend further, placing your right shoulder level with casualty's lower abdomen.*

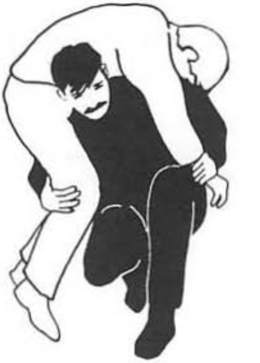

Go down on to your right knee if you wish. Allow weight to fall across your shoulders and back. Place your right arm between or around legs.

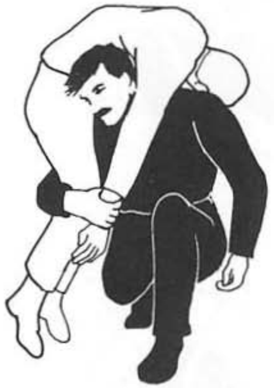

Transfer casualty's right wrist to your right hand and lift, taking weight on your right shoulder. Press down on your left knee with your left hand to help push yourself up.

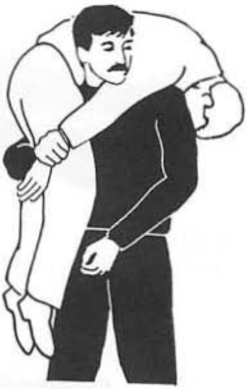

This is a relatively comfortable position in which weight is spread across your shoulders. But casualty's head is hanging down – bad for head or facial injuries.

Lifting with a sling

Another method for lifting an unconscious person, and the best one-man carry for a long distance. A sling is needed which is wide enough not to cut into the casualty, and long enough to go over your shoulders and twice across the victim's back. Two triangular bandages, a rifle sling, broad belts or luggage straps of webbing or leather could all be used. If rope is used, it must be padded to prevent it cutting or chafing.

MAKE SLING

Form sling into a continuous loop and place beneath casualty's thighs and lower back.

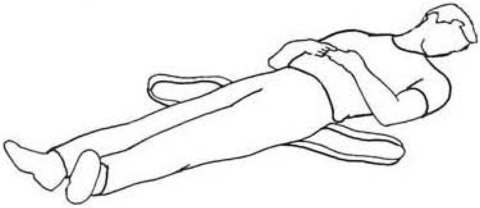

LIE BETWEEN CASUALTY'S LEGS

Thrust arms through loops. Tighten some of slack in sling. Grasp casualty's hand and grip trouser or leg on injured side of body. Turn away from the injured side rolling over so that casualty lies on top.

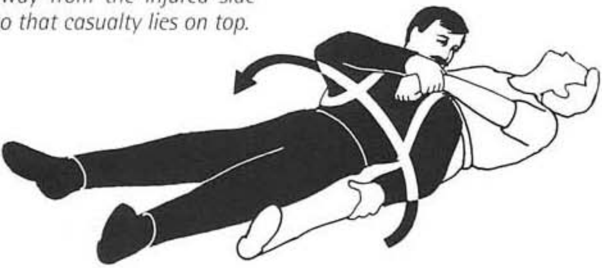

ADJUST SLING

To make body comfortable on your back.

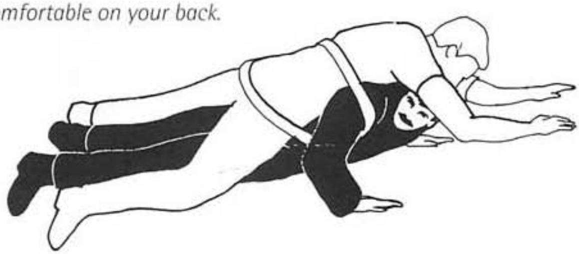

RISE TO A KNEEL

The belt will lift the casualty on your back. If the belt feels loose, or the casualty feels unsafe, return to previous position and adjust straps. Use a hand on the raised knee to help push yourself upright.

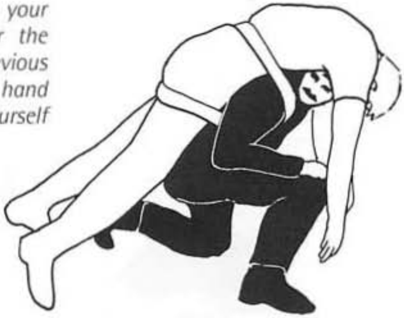

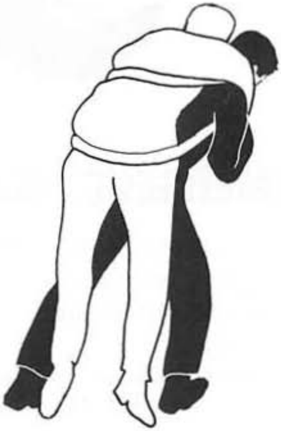

CARRYING THE CASUALTY

The weight is now carried over your back. You can hold the casualty's wrists to steady the weight but should be able to proceed with both hands free.

Two-persons seat

If the casualty is able to use his or her arms, placed around the carrier's shoulders, two people can make a four-handed seat. Grip as shown: right hand on your left wrist, left on the other person's right.

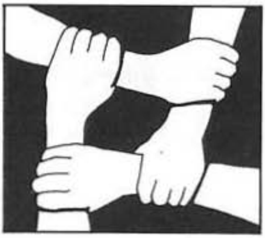

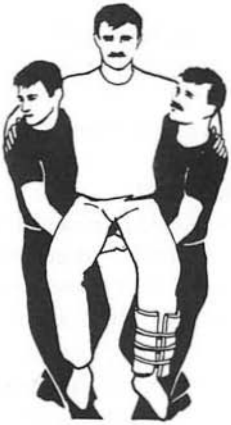

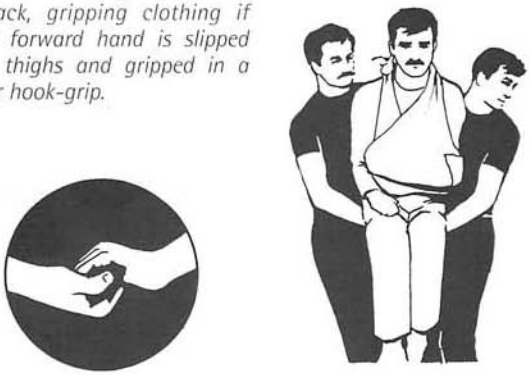

Casualties unable to use their arms can be lifted by the carriers stooping on either side and reaching across the casualty's back, gripping clothing if possible. The forward hand is slipped beneath the thighs and gripped in a padded finger hook-grip.

EMERGENCY CHILDBIRTH

The stress of an emergency sometimes precipitates labour. The signs include a low backache, regular contractions in the lower abdomen, a discharge of bloodstained mucus and sometimes the 'breaking of the waters'.

Sterilize scissors or a knife and three 20cm (8in) lengths of string. Prepare a comfortable and clean surface for the mother. Have plenty of hot water available.

Helpers should scrub their hands thoroughly and should not include anyone with any sign of a cold or infection or with sores on their hands.

First stage of labour

Uterus contracts at 10–20 minute intervals. Increase in bloodstained mucus. Cramp-like pains lasting up to a minute become increasingly more frequent. This stage may last several hours.

Second stage

Half litre (1pt) or more of water flows out in a rush – the 'breaking of the waters'. Mother should lie on her back. During contractions she should bring her knees up and grasp them with her hands, bending her head towards them and holding her breath. She should rest between contractions.

When a bulge appears delivery is near. Mother should adopt a delivery position. Either: Lying on side with knees drawn-up and buttocks near the edge of a bed, tabletop, etc, or crouching (which is often a better position if the mother feels strong enough).

Delivery

Mother should NOT hold her breath and should NOT bear down during contractions. She should take short breaths with her mouth open. By panting she makes it easier for the baby to emerge slowly and smoothly. If mother makes a bowel movement during delivery wipe clean from front to back.

- The baby's head usually emerges first, but not always.
- Tear any membrane covering the baby's face.
- If the umbilical cord is around the baby's neck – ease it over the head or loop it over the shoulder.
- Support the baby's head in the palm of your hands. As the shoulders appear support body under armpits and lift towards mother's abdomen. Be prepared for the baby to be very slippery.
- Ensuring that no tension is put on the cord, place the baby by mother's legs (or if she is lying on her back, not side, between them); head lower than body.
- If baby does not appear head-first and delivery is held up for more than three minutes after shoulders emerge, pull very gently.

After delivery

Bind a cloth around baby's ankles. Hook one or two fingers under cloth to support baby, hanging head downwards, to allow fluid to drain from its mouth and nose. Hold head slightly back and mouth open. Wipe away any mucus or blood. When the baby cries lay it by its mother's breast.

If it does not cry and does not show any signs of breathing within two minutes of delivery begin very gentle mouth-to-mouth resuscitation.

Better now for mother to lie on her back, with legs apart, for delivery of placenta (afterbirth), usually about 10 minutes later. After delivery of placenta, or after cord has stopped pulsating, the blood should flow out of the cord into the baby, the cord becoming white instead of blue.

Firmly tie a piece of sterile thread around it 15cm (6in) from baby's navel. Then tie with another thread at 20cm (8in). Check that the first tie is really secure or the baby may lose blood. With sterile scissors or knife sever the cord between the two ties. Place a sterile dressing over the cut end. Leave for 10 minutes then check that there has been no bleeding. Tie a further thread 10cm (4in) from the baby.

Wash the mother, give her hot drinks and encourage her to sleep.

BITES

Mammal bites

Animal bites are chiefly dangerous because of the infections that can result from bacteria in the animal's mouth. Rabies, the most serious, is almost always fatal if it has developed sufficiently to produce the symptoms of increasing irritability, dislike of light, hydrophobia (violent aversion to water) and paralysis. Under emergency survival conditions, with no vaccine, there is no hope of treating it. Great care must be taken that the victim does not transmit the disease to anyone else. Felines, canines and apes and many other animals can carry rabies, there is even a form carried by bats.

If you are bitten at any time in a survival situation, even if the bite heals and all seems well, you MUST report the bite when you are rescued. You should be examined by a doctor.

Any bite could also cause tetanus. Anti-tetanus shots are sensible for everyone and rabies vaccine for those travelling in areas where rabies may be expected.

Thoroughly cleanse all animal bites, washing for at least five minutes to remove saliva and wash out any infection. Then deal with bleeding, dress and bandage.

Snake bites

Few snake bites are fatal if anti-venom is available. The victim who is rushed to hospital within an hour or two is usually treated in time, provided the kind of snake is known so that the correct anti-venom can be prepared. Under survival conditions access to anti-venom is unlikely and the victim may not be so lucky but, fortunately, only a small proportion of snakes are venomous.

The fangs which discharge the poison of many snakes are set at the front of the upper jaw and leave distinctive puncture marks as well as the pattern of the bite (those of Coral snakes of the Americas may not be noticeable).

Non-poisonous snakes can also administer a savage bite. If you are not certain whether a snake was poisonous or not, treat as poisonous (though it may calm the patient to say that it was not a poisonous snake).

TREATMENT: The aim is to prevent poison spreading through the body. Reassure the victim. Make them relax, resting with the bitten area lower than the heart. Wash away any venom on the surface of the skin, with soap if possible. Place a restricting bandage – NOT a tourniquet – above the bite, and bandage down over the bite. For instance, if the victim has been bitten on the ankle, start bandaging at the knee. This bandage prevents the toxin from spreading rapidly and being taken up by the lymphatic system.

Place wound in cool water – a stream for instance. Use ice if available to keep as cool as possible.

The casualty will almost certainly need treatment for shock and may also require artificial respiration, keep a check on breathing.

Never cut a snake bite or try to suck out the poison.

Venomous and dangerous animals

For identification and details of SYMPTOMS (see *Dangerous Creatures* colour illustrations).

Spider bites

Spider bites should be treated in the same way as snake bites. A cold compress (like a poultice, but cold – ice wrapped in a cloth is ideal, if available) helps to reduce pain.

Stings

Scorpions can inject powerful venom. Bee, wasp and hornet stings can cause severe reactions in some people. Multiple stings are very dangerous – the amount of toxin and the inflammation this causes, may affect respiration.

Bee stings are left in the skin and should be carefully removed. Do not squeeze the stinger end or more venom will be injected. Stroke the sting with the side of a needle to draw it out – DON'T prod with the point.

Treatment is again like that for snake bite.

Chiggers, Ticks, Mosquitoes

Most small insect bites are a nuisance rather than a danger, but some can carry diseases. (See *Warm climate diseases*, this section, and *Tropical regions* in *Climate and Terrain*.)

GENERAL POISONING

The quickest way to remove a poison that has been swallowed is to vomit it up – but this is DANGEROUS in the case of caustic chemicals and oily substances that can produce further damage when being brought up. In cases of suspected plant poisoning always induce vomiting by putting a finger down the throat.

In ordinary domestic as well as survival conditions food and drink containers may have been wrongly used to store toxic substances (despite the fact that it is against the law in Britain). Always check they actually contain what the label says and take particular care that hungry or thirsty children do not help themselves to cream sodas that are actually the camp fuel or bleach reserves.

An effective 'universal' antidote, which will also help to absorb a poisonous liquid can be made up from tea (tannic acid) and charcoal – with an equal part of milk of magnesia if available – or just use charcoal. The aim is to absorb the poison so that it passes out of the system with the charcoal.

Contact poisons

Some plants, such as Poison Ivy, Poison Sumac and Poison Oak, can produce considerable skin irritation. Some people seem to have immunity, but it is not necessary permanent. Some people may have mild or extreme allergic reactions when touching other plants. All skin that has been in contact with the plant should be thoroughly washed with soap and water. It is most important to remove and wash all contaminated clothing since once the irritant oil is on the clothing it can spread to other parts of the body.

Alcohol can then be used to neutralize any oil left on the skin or clothing.

Severe cases may require treatment for shock.

WARNING

If handling a plant has produced a severe reaction, DO NOT PUT YOUR HAND TO YOUR FACE – ESPECIALLY THE EYES OR MOUTH – OR TOUCH THE GENITALS until it has been well washed. These sensitive parts of the body can react with swellings and rashes which can interfere with breathing or cause urinary blockage.

Chemical poisons

Chemical poisons should be sluiced off the skin with water. If skin has been damaged, treat as for burns (see also *Burns*).

WARNING

Sluicing with water may worsen the effects of some chemicals. You MUST know the properties of chemicals to which you are exposed, and the appropriate action to take in case of accident.

GENERAL DISORDERS

Conditions such as constipation and diarrhoea are to be expected, with the jolt to the system that the survivor may have experienced. Failure to defecate or urinate is not unusual in shipwreck situations. Coping with new and perhaps initially distasteful foods may also bring a feeling of nausea.

Small digestive upsets are relatively insignificant in survival conditions, but symptoms which suggest a more serious condition should not be ignored.

Lack of food will itself place a strain upon the body and stress may cause headaches and other disorders. Women survivors may find that menstruation ceases while under severe stress conditions.

However, if food is adequate in quantity the best treatment for digestive problems is to fast for a day and to rest, though in hot climates you should take fluids.

Fevers

Treat the symptom with rest and aspirin and look for the cause of the fever.

Pneumonia

Characterized by a rapid rise of temperature, often with chest pain, sputum mixed with blood, headache, weakness and, later, delirium. Pneumonia has many causes – most commonly a bacteria which attacks the lobes of the lungs. Without antibiotics careful nursing is all that you can do.

Encourage the patient to walk and to do deep breathing exercises. Keep the patient warm and encourage him or her to take frequent sips of hot water. Stay with the patient as much as possible and give verbal encouragement.

DISEASES

There are three main causes of infectious diseases: bacteria, viruses and rickettsiae. Then there are the problems caused by large parasites such as internal worms, scabies and the tropical larvae of the warble fly and the tiny chigger (chigoe).

Diseases caused by bacteria include dysentery, cholera, tuberculosis and diptheria. Viruses cause the common cold, influenza, measles, and poliomyelitis. Rickettsiae, germs which are neither bacteria, nor viruses, cause typhus, scrub typhus and Rocky Mountain spotted fever.

Most of the common infectious diseases which are familiar to everyone from childhood will respond to careful nursing. They are 'caught' from another person. Unless already carried by the party (or humans with whom you come into contact), neither these nor contagious and sexually transmitted diseases are likely to occur. Some human parasites, such as the itch mite which cause scabies, cannot live for very long off a human body so they too will only be a problem if you have brought them with you.

SCURVY

A deficiency disease due to lack of vitamin C, this used to be the scourge of sailors on long voyages. The survivor with access to fresh fruit and vegetables is not at risk. The desert- or sea-survivor is likely to succumb to thirst or hunger well before this problem becomes acute. Symptoms include bleeding gums, loose teeth, easy bruising and breathlessness.

Relying entirely upon hunting and fishing could make this a problem, especially in polar areas, if you have no vitamin tablets left. Always aim for a balanced diet to avoid this and other deficiency diseases.

Precautions

The diseases to which the survivor is most likely to be exposed, especially in tropical countries, are those carried in water or by insects and food animals.

Since the symptoms of tropical disease will be less familiar to most survivors, the most common will be dealt with in more detail so that they can be recognized. Where drugs are not available, treatment of disease will largely be a matter of dealing with symptoms and making the patient as comfortable as possible. Some remedies from natural sources can be prepared.

However, prevention is better than treatment. Measures to avoid diseases should include all suitable immunization before travelling (especially to the tropics) and then giving meticulous attention to the following precautionary procedures:

- Purify drinking water
- Clean hands when preparing or eating food
- Wash and peel fruit
- Sterilize eating utensils
- Cover the body to reduce risk of insect bites
- Wash clothes and hold over smoke if louse-ridden
- Wash body (but avoid swallowing possibly contaminated water when washing or swimming)
- Bury excreta
- Protect food and drink from flies and vermin

Nursing/isolation

When infectious disease does occur it is important to isolate it. Contact with other members of the group should be kept to a minimum. All utensils used by the patient should be boiled. Cuts and sores should be closely covered, so that they are not exposed to infection. Hands should be thoroughly washed after treating the patient.

Avoid splashing with mucus from the patient's coughing or sneezing, which may carry germs. Take special care that the patient's faeces are disposed of without risk of infection to the handler and buried where they cannot spread infection or be disturbed.

WATERBORNE DISEASES AND EXCREMENTAL DISEASES

The simplest rule to prevent catching these diseases is BOIL ALL WATER – do not even clean your teeth or rinse your mouth when bathing with water that could be impure. Cover cuts and wounds and avoid unnecessary standing in water in areas where you could be at risk. Adopt rigid hygiene procedures, especially with regard to food preparation and cooking and eating utensils.

WARM CLIMATE DISEASES

In addition to the diseases already described the following are prevalent in warm and tropical climates – although they can occur elsewhere if there are carriers of the infection. (See also warnings against animals more likely to carry parasites and food poisoning, such as salmonella, in *Food*.)

Waterborne diseases

BILHARZIA – A disease of the bowel or bladder, caused by a microscopic fluke or worm, which spends part of its life cycle in a water snail and part in the human liver. It is endemic in parts of Africa, Arabia, China, Japan and South America. It enters the body either through drinking infected water, or through broken skin.
Main symptom: An irritation of the urinary tract.
Can be treated: With the drug Niridazole in recommended doses.

HOOKWORMS – Another parasite that can enter the body in infected drinking water or penetrates bare skin, usually through the feet. The larvae travel through the bloodstream to the lungs, where they may cause pneumonia, are coughed up and swallowed to develop into worms, about 1·25cm (½in) long, in the intestine, causing anaemia and general lethargy. Common in many warm countries including the southern United States. Keep your boots on!
Can be treated: With the drugs Alcapar and Mintazol in recommended doses. A decoction of bracken is also a powerful de-wormer.

AMOEBIC DYSENTERY – Transmitted through contaminated water and uncooked food, this form of dysentery is largely found in the tropics and subtropics, though sometimes encountered in the Mediterranean.
Symptoms: It does not necessarily produce a temperature, but the victim will feel fatigued and listless. Faeces may be solid, but will smell foul and carry blood and mucus looking like red jelly.
Treatment: Administer fluids, maximum rest and correct dosage of Flagyl.

Insect-borne diseases

Dengue, malaria and yellow fever are all transmitted by mosquito bite. A course of tablets, begun before exposure, can protect you from malaria.

There is no defence against the others, except keeping the skin covered by wearing long sleeves and trousers, especially at night, sleeping under a mosquito net if possible and using insect repellents.

Remember: Do NOT increase risk by camping near stagnant water and swamps.

MALARIA – Not restricted to the tropics, although probably the first disease the tropics call to mind. It could occur anywhere the anopheles mosquito is found, and was once prevalent in swampy areas of southern Italy and other parts of Europe. Transmitted through the saliva of the female mosquito, it produces recurrent fever. Although sweating, the patient feels intensely chilled and shivers violently. There are four kinds of infection. One produces almost continuous fever, accompanied by delirium or coma. Two kinds produce fever every 48 hours and the fourth kind, fever every 72 hours.

Treatment: Quinine is the traditional treatment but paludrine, chloroquine, darapryn and other anti-malarial drugs are now available. Protection can be obtained by starting a course of tablets before visiting an infected area. This is so the body can get used to the tablet and if any of the symptoms are unbearable an alternative found.

Taking 2 Paludrine daily and 1 Nivaqin tablet weekly is still recommended and now comes as a pack. But take advice from your doctor on their effectiveness. They must be continued even after returning from abroad. Follow pharmacist's instructions carefully.

However, in certain areas of the world, conventional anti-malarial drugs like Paludrine are now ineffective. Different species of mosquito have become immune. Areas in particular include Africa and the Far East. Check with your travel agent for up-to-date information.

Larium is now the drug to take for complete cover, but it has some nasty side-effects. These include nausea, vomiting, diarrhoea, abdominal pain, dizziness, loss of balance, anxiety, hallucinations and sleep disorders. The dose is 1 tablet per week and should be started 2–3 weeks before the trip.

There is another way to use Larium, which a lot of overseas workers practice. They don't take anything so the signs and symptoms of malaria are not masked in any way. At the first sign of malaria they begin immediately on a course of Larium which is a cure. It should be taken 4 weeks after return from abroad.

A relatively recent drug, called Malarone, has proved to be 98 per cent effective with fewer side-effects.

DENGUE (BREAKBONE FEVER) – Dengue fever lasts for about a week, with headaches, pains in the joints and a rash. Unlike malaria, one attack by this virus usually confers immunity. The type of mosquito that carries dengue is just as likely to bite in day and night. There is no drug to treat dengue.

YELLOW FEVER (BLACK VOMIT) – Prevalent in Africa and South America, Yellow fever first produces headaches, limb pains and fever. The mouth feels swollen, vomiting may produce blood and the eyes are bloodshot. After three or four days the fever is reduced but there is constipation, pain in the kidneys and urination becomes less and less. There is increased vomiting and the skin takes on a yellowish tone before recovery.
Treatment: Rest and nursing.

SANDFLY FEVER – A disease mainly of the grasslands of the sub-tropics and the Mediterranean. A small fly bite, usually on ankles, wrist or neck, produces an itching which is aggravated by scratching and can lead to sores which are an entry for further infections. Headache, bloodshot eyes and 'flu-like symptoms may also occur. The fever usually lasts about three days before subsiding.
Treatment: Rest and liquids.

SCRUB TYPHUS (TSUTSUGAMUSHI DISEASE) – Also known as mite fever, and occurring widely through eastern Asia and Australasia, this is carried by mites whose larvae live on rodents. The mites transmit the disease by biting, often in the groin or on the neck. The bite usually goes unnoticed, but a sore will develop in a few days in the bite area. Severe headache, shivering and fever develop within 10 days.

The disease is similar to other forms of typhus, with a rash appearing after about a week, and is treated in the same way. Infection is more likely in areas of thick dry vegetation. Avoid such places, sleep at least 30cm (1ft) from the ground and use insect repellent to avoid this disease.

TYPHUS – There are two forms. Epidemic typhus is carried by body lice which transmit it to humans. Endemic, or murine, typhus is transmitted by the rat flea. Both are diseases associated with dirty, overcrowded conditions and more likely to be diseases of disaster aftermath than survival in remote places.
Symptoms: Both forms produce headache, nausea and, after about four days a rash which spreads from the armpits to the chest,

abdomen and thighs. Victims may recover after about 12 days, or fall into delirium, coma or death.
Treatment: Antibiotics. There is also a vaccine available.

ROCKY MOUNTAIN SPOTTED FEVER – Also known in Brazil as São Paulo fever, this is another type of typhus, spread by ticks. As the name suggests, it is also found in the western United States.
Symptoms and Treatment: As for typhus, but the rash tends to be most severe on wrists and ankles.
Small parasites which burrow beneath the skin, such as the tear-shaped, bristly larvae of the warblefly, or the tiny chigoe, which makes red pinpricks where it lodges on the skin, should be removed before they can open up a route for further infection.

WARM CLIMATE AILMENTS

PRICKLY HEAT

Miliaria, or prickly heat, can occur whenever people are exposed to very hot conditions without acclimatization. Heavy sweating, coupled with rubbing by clothing, can produce blockages in the sweat glands and an uncomfortable skin irritation. Heavy exertion may precipitate miliaria, which is aggravated by sunburn and eczema. Babies are very vulnerable.

Excess sweat does not clear sufficiently quickly, the glands are blocked and the cooling effect of sweating is lost. Taking more liquid, especially hot drinks, may produce more sweat, and make the situation worse. Remove clothing, wash the body with cool water and put on dry clothes. Small amounts of tiger balm applied to the upper lip may distract from the itching and antihistamine relieves the discomfort.

HEAT CRAMPS – These are usually the first warning of heat exhaustion and occur in the muscles which are doing most work: arms, legs and abdomen. Usually due to lack of body salt (produced by excessive sweating – especially if no salt has been taken).
Symptoms: Shallow breathing, vomiting, dizziness.
Treatment: Move to shade. Rest. Drink water with a little salt dissolved in it – only a pinch to a half-litre or pint.

HEAT EXHAUSTION – Exposure to high temperature and humidity, with loss of body fluids through excessive sweating, produces heat exhaustion. It can occur without direct exposure to the sun – lying on hot ground with poor ventilation for instance.
Symptoms: Face pale, skin cold yet sweating, pulse weak. Accompanied by dizziness, weakness and perhaps cramps. Patient may become delirious or unconscious.
Treatment: As for cramps.

HEATSTROKE – The most serious result of overexposure to or overexertion in the sun.
Symptoms: Hot dry skin, face flushed and feverish – but sweating stops. Temperature rises, pulse becomes fast and strong. Severe headache, often with vomiting. Unconsciousness may follow.
Treatment: Lay in shade, head and shoulders slightly raised. Remove outer clothing. Cool body by wetting underclothing with TEPID water (cold water would push the core temperature up) and fanning. Do NOT fully immerse in water – sprinkle it over the patient. Then lay in a damp hollow with plenty of ventilation. When consciousness returns give water to drink. When temperature returns to normal, replace clothing, keep warm to prevent chill.

Immediate immersion in cold water is very dangerous but in extreme cases, where risk of death or brain damage outweighs shock from immersion, could be used AFTER initial cooling begins to take effect. Slowly lower the body into water, feet first, massaging extremities to increase blood flow and aid dispersal of heat from body's core. Remove as soon as temperature falls – be prepared to cover patient if it plummets. You may need to cool and cover several times before temperature becomes stable.

SUNBURN – Actual burn, with blistering (not just tanning) is a real danger, especially with pale and sensitive skins. If more than two-thirds of the body is affected it can prove fatal.
Treatment: Avoid further exposure – keep in the shade. Take pain killers if available. Cover all blisters with dressings but DO NOT BURST.

SORE EYES – Sore eyes may be due to glare – especially at sea and in desert and snow-covered locations (see *Snow-blindness* on p.453) – or to excessive exposure to the sun or dust particles.
Treatment: Rest in shade, covering eyes after washing out foreign bodies and bathing in warm water. Use a mask and darken below eyes with charcoal to avoid recurrence.

DEHYDRATION

Dehydration becomes increasingly noticeable as more body fluid is lost. Water makes up 75 per cent of the body weight – about 50 litres (11 gallons) for the average man. Survival is unlikely if more than one fifth of this is lost.

Fluid loss 1-5 per cent: Thirst, vague discomfort, lack of appetite, flushed skin, impatience, sleepiness, nausea.

Fluid loss 6-10 per cent: Dizziness, headache, laboured breathing, no salivation, indistinct speech, unable to walk.

Fluid loss 11-20 per cent: Delirium, swollen tongue, unable to swallow, dim vision, numb and shrivelled skin.

In the latter stages: There is gross muscular weakness and mental capacity is impaired – you must make your plans at the start when you can think clearly – and then stick to them.

COLD CLIMATE HAZARDS

It is not only in the polar regions that low temperatures create health hazards. Prolonged exposure to cold is dangerous anywhere. Take precautions against all these conditions.

HYPOTHERMIA – Technical name for the condition where the body cannot generate heat as fast as it looses heat and its temperature falls below normal. It is caused by exposure to wind, rain and low temperatures and also brought on by:

- Exhaustion
- Inadequate clothing
- Inadequate shelter
- Inadequate food intake
- Lack of knowledge and preparation

The conditions which aggravate hypothermia are exactly those likely to occur in the polar regions – though it can occur under any cold conditions, especially as a result of wind chill. It is a common problem of any very cold spell, especially among the underprivileged elderly. It is a killer and must be treated as soon as it is recognized.

Prevent it by sheltering when conditions are bad and by keeping dry. Avoid overexertion and if in a group use the 'buddy system'. If one person goes down with hypothermia others in the group may also be near to it. Check everyone for symptoms.

Signs and symptoms: Irrational behaviour, typified by sudden bursts of energy followed by lethargy. Slowing down of responses, failing to respond to questions or instructions. Sudden uncontrolled fits of shivering. Loss of coordination, stumbling and falling. Headaches, blurred vision and abdominal pains. Collapse, stupor or unconsciousness.

Aggravating factors: Soaked clothing, with high winds. Low air temperature accompanied by high winds. Immersion in water. Any injury that immobilizes and reduces the ability to produce heat. Anxiety and mental stress. Unusual thinness.

Treatment: Prevent any further heat loss. Shelter from wind and weather. Replace wet clothing with dry. Start at the head, by putting on a hat. Remove clothing from the waist up and replace with dry clothes. Then remove clothing from the lower body. This helps preserve the core temperature. Insulate patient from the ground and apply warmth (other bodies, hot rocks). Give warm fluids and sugary foods – but only if conscious.

In advanced hypothermia the body loses the power to rewarm itself. However, it must still be rewarmed from the inside, because fast external heating would drive cold blood into the core, further aggravating the situation. Place warmth in the following places: pit of stomach, small of back, armpits, back of neck, wrists, between thighs. These are places where the blood is near the surface and will carry the heat through the body.

Do not apply or administer alcohol. It opens blood vessels at the surface of the skin, allowing heat to be lost more rapidly.

The patient is NOT cured when his or her temperature reaches normal. Body reserves must be built up until the ability to generate heat internally returns.

A foil blanket will reflect heat and aid care in sub-zero conditions.

HEAT LOSS: GENERAL RULES

If heat is lost rapidly – rewarm rapidly
If heat is lost slowly – rewarm slowly

FROSTBITE – Frostbite occurs when the skin and flesh freeze – when reduced to a temperature of -1°C (30°F). Frostbite affects all exposed parts of the body and the regions furthest from the heart which have the least powerful circulation: hands and feet, nose, ears and face. It can be light or deep according to the degree of exposure.

The first signs are often a prickly feeling as the skin freezes. Then rather waxy-looking patches on the skin which feel numb – and later hard and pebbly with considerable pain, swelling, reddening and blistering before the deadening and dropping off that is the final stage.

Frostnip: Affects only the skin. To treat it place the affected part in a warm area. Put hands under your armpits or in your crotch. Put feet against a friend's stomach (you may not stay friends for long!). Thawed out frostbite will be painful.

BE AWARE

Keep a continual look out for signs of frostbite on yourself and your companions. Act at the first appearance of any waxy signs. Exercise the face by pulling grimaces to combat its attack.

Deep frostbite: A much greater problem. Protect the affected area from greater injury. Do NOT rub with snow. Do NOT expose to an open fire. The best treatment is to thaw the injured area gradually with warm water at a temperature of about 28-28·5°C (82-83°F) – about the temperature which your elbow can comfortably bear. If too hot, cool down to the right temperature then apply to the affected areas.

Advanced frostbite: May cause blisters to form which can become infected and form ulcers. These tissues become grey, then black and die and will eventually fall off. Do NOT burst blisters and NEVER rub the affected part. Severe pain is an indication that the part has been warmed too quickly. Use 'animal warmth' only.

SNOW BLINDNESS – A temporary form of blindness, caused by the high intensity and concentration of the sun's rays, which are reflected from the snow-covered ground or ice (and also from ice crystals in the clouds). It occurs most frequently when the sun is high, but can also occur when there is no direct sunlight – during a bright overcast period in polar and alpine regions.

Symptoms: First the eyes become sensitive to the glare and then blinking and squinting begins. Vision takes on a pink hue, becoming redder. If not checked at this stage the eyes begin to feel as though they have sand in them.

Treatment: Get into a dark place and blindfold the eyes. Heat aggravates the pain so apply a soothing cool wet cloth to the forehead. The condition corrects itself in time. Prevent further exposure by wearing goggles and black beneath the eyes with charcoal to reduce the glare (see *Climate and Terrain*).

CARBON MONOXIDE POISONING – A risk that is a byproduct of incomplete combustion in a badly ventilated area – which may result from overzealous efforts to prevent draughts. All forms of fire and stove are a potential cause if used in confined spaces without adequate ventilation. The carbon monoxide which builds up is colourless and odourless, so is difficult to detect. It may cause slight headache, dizziness, drowsiness, nausea or even vomiting – but these signs can go unnoticed and progress to unconsciousness without warning. Unless discovered promptly it will be fatal.

Treatment: Simple – and prevention even simpler – VENTILATE. There is no treatment if you are on your own so ensure that such conditions cannot develop. Remove the patient to fresh air and encourage him or her to breathe evenly and regularly. If unconscious and not breathing, apply artificial resuscitation. Keep patient warm and ventilate the offending shelter.

TRENCH FOOT – This occurs when feet are immersed in water for long periods, or are damp and cold for a long time. Tight-fitting boots accelerate the condition. This is a serious malady and its onset is hastened by exhaustion, coldness and lack of food, drink or sleep.

Prevent trench foot developing by keeping the feet dry. Wear boots that fit correctly, exercise the legs and toes and inspect feet regularly.

Symptoms: Feet feel as though they have pins and needles. Numbness sets in and is interspersed with sharp pains. On inspection feet appear purple with swelling and blisters.

Treatment: Dry the feet, but do not rub or damage blisters. Elevate feet and cover to keep warm – but improvise a cage to take weight of coverings. Do NOT apply artificial heat. Do NOT massage. Rest and warmth are the cure.

SURVIVAL CASE STUDY

My medical training came in handy on many occasions, especially in remote areas. Everyone should have some knowledge of first aid so they can at least help in an emergency.

In the early days of the Borneo campaign, to win the confidence of the locals we set up a 'hearts and minds' campaign. This included medical treatment which quickly helped form a bond with the local tribes, who came from many miles around with a variety of illnesses, fevers and trauma.

One day a fisherman was carried in with his left foot covered in rags. The smell got stronger as he got nearer, and when the foot was unwrapped I had to hold my breath. The whole foot was bloated and oozing pus. He was feverish and rambled on about being hit by a boat propeller. After trimming away the rotten flesh and administering a course of penicillin his recovery was amazing. Undoubtedly he would have died from septicaemia, but after seven days treatment, he was walking again.

The locals were amazingly receptive to modern drugs, and so care must always be taken when administering any drugs to them, as they have built up a natural and inherited immunity to illnesses such as tropical fevers which can be eradicated by taking just one pill. Just one anti-malaria pill, for example, can wipe out thousands of years of the locals' natural immunity in seconds, leaving them open to the mercy of the disease.

Our fame spread and our queue of patients got longer. Most days we treated cuts, fractures and other similar ailments. As their confidence in us grew so did the challenge of treating some cases. The locals carried out their own childbirth relying on tried and tested methods, unless something went wrong – then they would come to me. There was, and still is, a very fine line to be trodden – do I assist or not? If I assist and it goes wrong I will lose face, contradicting the reason we are there in the first place. But if I don't assist I still lose face, as well as the trust we had built up. It's all down to training and the amount of knowledge you have. I was lucky enough in that I got involved with a difficult birth and it turned out fine.

NATURAL MEDICINE

Natural remedies are available for many human illnesses if you know where to look. For thousands of years drugs and treatments have made use of all kinds of herbs, plants and natural substances and many of the drugs still in use today are derived from plant sources. Not all the traditional medicines have been endorsed by modern doctors, some were based more on an association of the plant than with its medical properties – though strangely, colour and appearance does often seem to be indicative. Plants such as scarlet pimpernel and burdock, used to clear the blood, are mainly red; those that treat jaundice, including agrimony, hawkweed and dandelion, are yellow. Sometimes a name reminds us of the old medical usage – lungwort for instance, or eyebright.

Cruel to be kind

Sick survivors need your care and attention, but they must also be kept interested and optimistic. If a man with pneumonia is left to curl up in a corner of the shelter, he will probably die. Get him on his feet and keep him occupied with minor tasks. Give him plenty of fluids and coax him to eat. Do not let him lie down. The hepatitis case must be forced to eat. You cannot allow either strength or morale to be depleted.

It is not only plants that can replace your medical supplies, or to supplement your store. Urine can be used as an antiseptic to wash out wounds. If the patient is fussy, use their own. If sick enough, they won't care whose. Maggots too have been used to clean wounds. In tropical countries an open wound is soon infested with them – but they do keep it open and clean until better treatment can be given. Keep watch that they do not devour good tissue.

Fire has been used for centuries to clean wounds. Cauterizing with heat requires fortitude in the patient – but, if they can stand it, and you have ammunition, placing powder around a wound and lighting it can prevent gangrene. Some people believe that the best way to seal a stump after an amputation is to cauterize (as used to be done to criminals who had their hands chopped off) – but the shock of this on top of the shock of injury will kill some people.

Modern drugs

Many modern drugs such as cocaine, morphine and digitalis are derived directly from plants, but extracting them is NOT straightforward. Often poisons are involved which could be VERY

DANGEROUS if any attempt was made to use such plants in treatment. What follows is a list of plants and the medical uses to which the survivor can put them – in simple preparations. Although many modern drugs are made from tropical plants, and indigenous medicine has many plant ingredients, most of these plants are found in temperate regions for they are the ones which have been most documented and which you have a reasonable chance of recognizing.

Plant preparations

Leave all poisonous plants alone and be sure that you have properly identified the plant. As a general rule plants will be most potent when in flower. Note that different parts of a plant may have different uses.

Infusions are usually made from leaves or flowers and decoctions from roots. The methods are described below. Divide the amount made into three doses to take in one day. Always prepare infusions, decoctions and poultices freshly, just before using. Never keep for more than 12 hours.

Plants vary in potency depending on the season in which they are picked. Do not think you will do more good by taking or administering larger doses – you won't and you may do harm.

Don't expect miracles overnight. Give the treatment a fair chance to work.

To make an infusion

Cut and crush the herb so that juices and oils are more readily available. You need a slack handful of herb to a half-litre (¾pt) of water (30g to 50cc/1oz to 20fl.oz). Pour boiling water over it. Stir. Leave to cool. There is no need to strain – the herb will sink to the bottom.

If you cannot boil water use half the amount of cold water and stand the vessel in the sun. If there is no sun or no water try sucking or chewing the leaves, extracting as much of the juices as possible, then spit out the pulp.

To make a decoction

Usually a preparation from roots. Cut, scrape and mash root. Soak in water (handful to 85cc/1½pt) for at least half an hour. Bring to boil, simmer until liquid reduces by one-third.

To make a poultice

Mash up root, leaves or all of the herb and make into a flat pad. If too dry add water. Apply to affected part and cover with a large leaf, bind in position. Poultices can be applied to stiff joints, sprains and pus-filled sores.

Expressed juice

Reduce stem and leaves of the plant to a juicy mush by crushing with hands, rocks or sticks. Squeeze juice only into a wound and spread pulp around infected area. Keep in place with a large leaf and bind.

Splinting agents

The roots of Comfrey and, to a lesser extent, Solomon's seal, are so rich in starch that they will set hard when they have been well boiled and reduced. Allow to cool slightly and pack around an injured limb or joint. Good for holding poultices.

REMEDIES

STOPPING BLEEDING

Dove's-foot crane's-bill: Expressed juice
Giant puffball: Packed as poultice
Periwinkle: Expressed juice of leaves
Plantains: Pounded leaves as poultice
Self-heal: Expressed juice
Stork's-bill: Expressed juice of leaves
Woundwort: Expressed juice

CLEANSING RASHES/ SORES/WOUNDS

NOTE: *Use these plants externally to bathe the skin or where indicated, as a poultice. Apply two or three times a day.*

Burdock: Decoction of root; crushed raw root and salt for animal bites
Camomile: Infusion of flowers as poultice
Chickweed: Expressed juice of leaves
Cleavers: Infusion of whole plant, except roots
Comfrey: Decoction of root as poultice
Dead-nettle: Infusion of flowers and shoots
Docks: Crushed leaves
Elder: Expressed juice of leaves
Elm: Infusion of bark
Horehound: Infusion of whole plant, except roots.
Mallow: Decoction of leaves and flowers as poultice
Marsh mallow: Decoction of root infusion of leaves and flowers; as poultice
Oak: Decoction of bark
Sanicle: Infusion of whole plant, except roots
Scurvey grass: Crushed leaves
Shepherd's purse: Infusion of whole plant, except roots; as poultice
Silverweed: Infusion of whole plant, except roots
Solomon's seal: Decoction of roots; as poultice
St John's wort: Infusion of flowers and shoots
Sorrel: Crushed leaves
Tansy: Crushed leaves
Watercress: Expressed juice
Woundwort: Infusion of whole plant, except roots
Yarrow: Infusion of whole plant, except roots

ANTISEPTIC

NOTE: *These plants can be used externally or internally. They are particularly useful for wounds that become infected.*

Garlic: Expressed juice
Mallow: Infusion of leaves and flowers
Marsh mallow: Decoction of root; infusion of flowers and leaves
Horseradish: Decoction of root
Thyme: Infusion of leaves and flowers

ACHES/PAINS/BRUISES/STIFFNESS

NOTE: *Where indicated, use externally.*

Balm: Infusion of leaves
Birch: Infusion of leaves
Borage: Infusion of whole plant, except roots
Burdock: Decoction of root
Camomile: Expressed juice of flowers applied to swellings
Chickweed: Infusion of whole plant, except roots
Comfrey: Decoction of root applied to swellings
Cowberry: Infusion of leaves and fruits
Dock: Crushed leaves applied to bruises
Dove's-foot crane's-bill: Infusion of whole plant, except roots, applied to swellings
Elm: Infusion of bark
Figwort: Decoction of whole plant except roots; use externally to draw bruises and blood clots
Garlic: Expressed juice applied to swellings
Horehound: Expressed juice or leaves to earache
Poplar: Infusion of leaf buds
Solomon's seal: Decoction of root; use externally
Sorrel: Crushed leaves applied to bruises
St John's wort: Infusion of flowers and shoots applied to bruises
Tansy: Crushed leaves applied to bruises
Willow: Decoction of bark

FEVERS

NOTE: *These plants will induce perspiration to break a fever.*

Camomile: Infusion of leaves and flowers
Elder: Infusion of flowers and fruit
Elm: Decoction of bark
Feverfew: Infusion of whole plant, except roots
Lime: Infusion of flowers

COLDS/SORE THROATS/RESPIRATORY

Agrimony: Infusion of whole plant, except roots
Angelica: Decoction of root
Bilberry: Infusion of leaves and fruits
Bistort: Infusion of whole plant, except roots
Borage: Infusion of whole plant, except roots
Burdock: Decoction of root
Camomile: Infusion of flower; use as gargle
Colt's-foot: Infusion of leaves and flowers
Comfrey: Infusion of whole plant
Great mullein: Infusion of whole plant, except roots; decoction of root as gargle
Horehound: Infusion of whole plant, except roots
Horseradish: Raw root
Lime: Infusion of flowers
Lungwort: Infusion of whole plant, except root
Mallow: Infusion of flowers and leaves
Marsh mallow: Decoction of root; infusion of leaves and flowers
Mint: Infusion of whole plant, except roots
Mountain evens: Infusion of whole plant; use as gargle
Nettle: Infusion of leaves
Oak: Decoction of bark, use as gargle
Plantain: Infusion of leaves and stems
Poplars: Infusion of leaf buds
Roses: Decoction of hips

Sanicle: Infusion of whole plant, except roots
Self-heal: Infusion of whole plant, except roots; use as a gargle
St John's wort: Infusion of flowers and shoots
Thyme: Infusion of leaves and flowers
Willow: Decoction of bark
Yarrow: Infusion of whole plant, except roots; use as inhalant

SETTLING STOMACH

Balm: Infusion of leaves
Bilberry: Decoction of fruit
Bracken: Infusion of leaves
Bramble: Infusion of leaves
Dandelion: Decoction of whole plant
Horseradish: Infusion of root
Mint: Infusion of whole plant, except roots, with crushed charcoal
Solomon's seal: Decoction of root
Sanicle: Infusion of root
Yarrow: Infusion of leaves and flowers

DIARRHOEA

NOTE: *Take two or three times daily until symptoms subside.*

Bilberry: Decoction of fruit
Bistort: Infusion of whole plant, except roots
Bramble: Infusion of leaves or decoction of fruit
Cowberry: Decoction of fruit
Elm: Infusion of bark
Great burnet: Infusion of leaves and shoots
Hazel: Infusion of leaves
Marsh mallow: Infusion of leaves and flowers; decoction of root
Mint: Infusion of whole plant, except roots
Mountain evens: Infusion of whole plant, except roots
Oak: Decoction of bark
Plantain: Infusion of leaves and stems
Periwinkle: Infusion of leaves; NOT to be used for long periods
Silverweed: Infusion of whole plant, except roots

CONSTIPATION

Agrimony: Infusion of whole plant, except roots
Barberry: Expressed juice of fruit
Common cleavers: Infusion of whole plant, except roots
*Couch grass (*Elymus*): Decoction of root*
Dandelion: Decoction of whole plant
Elder: Expressed juice of fruit
Feverfew: Infusion of leaves and flowers
Rowan: Expressed juice of fruit
Rose: Decoction of hips
Walnut: Decoction of bark

HAEMORRHOIDS

NOTE: *Apply externally, two or three times a day.*

Bilberry: Expressed juice of fruit
Camomile: Infusion of leaves and flowers
Elm: Decoction of bark
Lesser celandine: Expressed juice of leaves
Oak: Decoction of bark
Plaintain: Expressed juice
Poplar: Decoction of leaf buds
Silverweed: Infusion of whole plant, except roots
Solomon's seal: Decoction of root

BE AWARE

For Headaches: Willow leaves and bark make a decoction containing Salicin, a constituent of Aspirin.
For Healing: Express the juice from Comfrey leaves to aid tissue regrowth.

EXPELLING WORMS

Bracken: Infusion of roots
Feverfew: Decoction of leaves and flowers
Figwort: Infusion of whole plant, except roots
Tansy: Infusion of leaves and flowers; use sparingly in small amounts

Tropical Medicinal Plants

Many thousands of tropical plants are known to have medicinal properties and are used by tribal peoples. Relatively few have been studied by Western scientists or are widely known. The following are a few of the many plants which may be of use, but lacking accurate information on medicinal plants you will do better to take medicines with you. NEVER experiment with something you cannot positively identify.

Copperleaf *(Acalypha indica) is one of several similar shrubs in India and southeast Asia, reaching 2–3m (6–10ft), with oval to heart-shaped leaves that are often variegated in shades of red, bright pink and green. In Malaya, the leaves are dried and drunk like tea. A decoction of roots and leaves is laxative and restorative.*

Alstonias, *including* Alstonia scholaris *are found from India east to Phillipines and south to Indonesia and parts of Australia. Boil the bark in water to produce a tonic, reduce fever, relieve diabetes and kill internal parasitic worms.*

Antelaea azadirachta *occurs from India to China and Indonesia. A decoction of leaves and bark will help to suppress malaria and dysentery. Use oil from seeds to treat ulcers and skin complaints.*

Bruceas *occur in many, fairly similar forms from India east to China and south to Australia. All parts are bitter tasting. Take seeds of* Brucea sumatrana *for diarrhoea and dysentery. Crushed leaves relieve external bleeding and sooth boils and insect bites.*

Elettaria cardamomum *is a relative of ginger, found in India eastwards, and is a tall, herbaceous plant with thick, fleshy rhizomes and a long, branched head of flowers. Use seeds or expressed juice of fruit for settling the digestion and relieving nausea.*

Cinchonas, *China Barks or Red Barks (*Cinchona*), large trees, typically with red-brown trunks, wild in tropical South America, have been taken all over the world. Take a decoction of the quinine-containing bark to suppress malaria.*

Horseradish Tree *(Moringa oleifera, see* Tropical plants *in* Food*) is a fairly common edible plant of the tropics. Use expressed juice from the roots and leaves to treat skin eruptions and inflammations.*

Kibatalia arborea *is another Asian tree whose bark, when cut, yields a latex-like sap. Use this sap, in small amounts, to treat worms.*

Sida cordifolia *is an erect, downy annual, up to 1m (3ft) tall, with oblong toothed leaves and yellowish flowers. It occurs from India east to Taiwan. Use an infusion of the leaves for coughs and fevers. The seeds are mildly laxative.*

Pergularia extensa *has stems with stiff spreading hairs, broadly oval leaves up to 15cm (6in) long and small greenish-white flowers. It grows in tropical Africa. Use tender leaves and shoots as a potherb or in a strong infusion to treat tapeworm and diarrhoea. Use a poultice of leaves on boils, abscesses and wounds.*

Crateava religiosa *is found from India east to Papua New Guinea and Polynesia. A decoction of bark and leaves relieves upset stomachs, feverish aches and pains.*

Baobabs *(Adansonia, see* Desert plants *in* Food*) Scrape gum off the bark. Use to treat worms and diarrhoea.*

MEDICINAL PLANTS

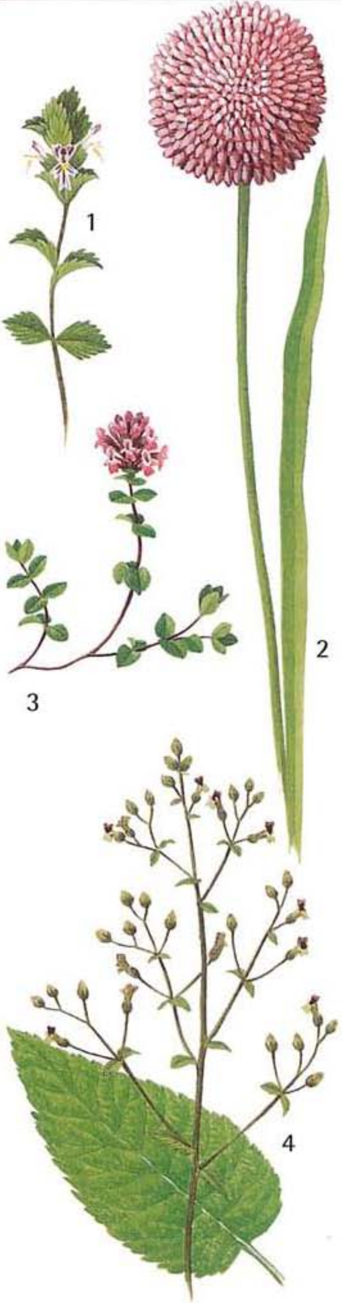

These medicinal plants are found in temperate climates. Many are very common and all are quite safe. Use them to staunch bleeding and heal wounds, for fevers, colds and digestive upsets and other treaments as described. Some have several uses, but they have been grouped here under their most common applications.

GENERAL AND ANTISEPTICS:

1 **Eyebright** *(Euphrasia officinalis) grows to about 30cm (1ft), with oval, often downy leaves and white flowers tinged violet or purple-veined and with a yellow spot in grassy places, often in mountains, in Eurasia. A strained infusion of the whole plant is excellent for eye infections. It is also said to ease hayfever, catarrh and nasal congestion.*

2 **Garlics** *(Allium) occur in many varieties in most temperate and, now, tropical parts. The smell will lead you to them; most have long, strap-like leaves arising from the bulb and a tall stem topped with a cluster of small pinkish or white flowers. The bulb is powerfully antiseptic; use as expressed juice externally diluted with water to treat wounds and swellings, and eat garlic to treat and to prevent colds. It also contains a natural antibiotic.*

3 **Wild Thyme** *(Thymus serpyllum) is small, aromatic, mat-forming, with small oval leaves and reddish-purple flowers; in dry grassy places in western Eurasia, but other species occur elsewhere. Use its antiseptic qualities in an infusion for coughs and colds, or as a potherb.*

4 **Figwort** *(Scrophularia nodosa) grows to 90cm (3ft), with square stems, pointed oval leaves and red-brown flowers; in woods, clearings and scrub in Eurasia – there are many*

different kinds. Apply as a decoction to reduce swellings, sprains, boils and bruises, to dissipate blood clots, and for treating haemorrhoids.

BLEEDING:

5 Self-heal *(Prunella vulgaris) is a downy, creeping plant with pointed oval leaves and heads of violet flowers; in dry grassy and waste places in Eurasia. Use as expressed juice to staunch bleeding or by infusion for internal haemorrhage.*

6 Dove's-foot Crane's-bill *(Geranium molle) grows to 30cm (1ft) with a hairy stem, deeply lobed leaves and small, pinkish, five-petalled flowers; in dry grassy and waste places. Use as expressed juice to staunch bleeding or as a decoction for internal haemorrhage.*

7 Marsh Woundwort *(Stachys palustris) is strong-smelling, hairy, up to 90cm (3ft) tall, with toothed, heart-shaped leaves and spikes of dark pink to purple, white-blotched flowers; usually found in damp places; similar species by woodland edges and shady waste place. Use as expressed juice to staunch bleeding or by infusion for bathing aches, sprains and wounds.*

8 Sanicle *(Sanicula europaea) grows to 50cm (20in), with hand-shaped, deeply lobed leaves and tiny white or pinkish flowers in a compact head; widespread in woodland in Eurasia. Use as expressed juice to staunch bleeding or by infusion for internal haemorrhage.*

9 Greater Periwinkle *(Vinca major) grows to 50cm (20in), with leathery, evergreen, broadly spear-shaped leaves and large blue-violet flowers; in woody, scrubby and rocky places in Eurasia. There are many kinds of periwinkle in other parts of the world. Use expressed juice externally to staunch bleeding.*

Plantains (see *Edible plants* in *Food*) *provide juice for treating wounds and for treating chest complaints..*

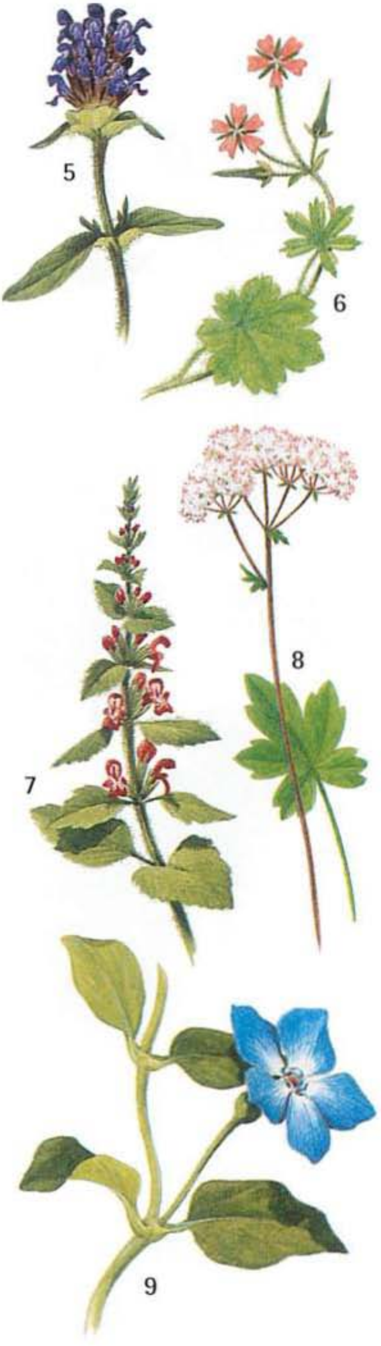

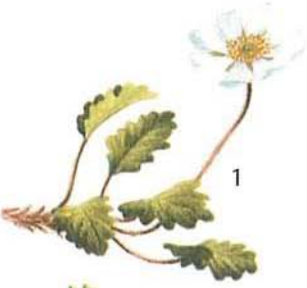
1

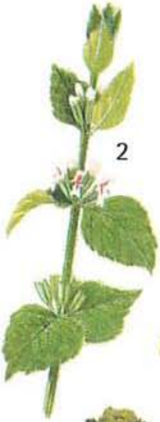
2

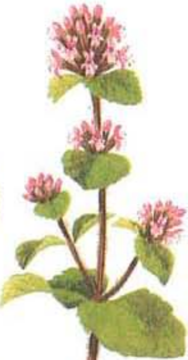
3

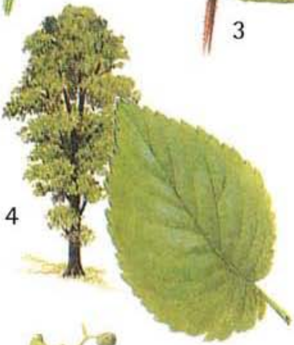
4

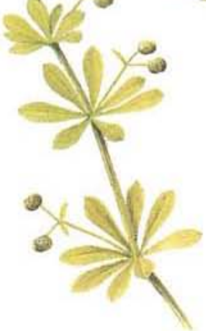
5

INTESTINAL PROBLEMS:

1 Mountain Avens *(Dryas octopetala) resembles a creeping wild strawberry, with well-lobed leaves, paler below, and large white, yellow-stamened flowers; in mountainous rocky and northern Arctic areas. Use an infusion of stems, leaves and flowers for diarrhoea or as a gargle.*

2 Balm *(Melissa officinalis) is lemon-scented and hairy, growing to about 60cm (2ft), with toothed, oval, greenish-yellow leaves and whorls of small white flowers at the leaf bases; in grassy places in the warmer part of Eurasia. Use an infusion of the whole plant for fevers and nausea. Can also be used to ease painful menstruation.*

3 Water Mint *(Mentha aquatica) is aromatic, hairy, always near fresh water, with toothed, pointed oval leaves, a purplish stem to 80cm (32in) and clusters of pinkish flowers. Use an infusion of the leaves for diarrhoea, as a digestive and heated to induce perspiration in fevers. Similar mints are also effective. If made too strong the infusion may become emetic (cause vomiting).*

4 Elms *(Ulmus) are tall trees with large, oval, toothed leaves, green disc-shaped fruits and, often, suckers at the base of the trunk. Use a decoction of the bark for diarrhoea and skin eruptions.*

5 Cleavers or **Goosegrass** *(Galium aparine) is straggling, with long prickly stems, whorls of narrow prickly leaves and small white flowers; widespread on moist, woody and waste ground. Young plants can be boiled and eaten like spinach. Use an infusion to ease constipation. Give frequent doses, mixed with an equal quantity of Marsh Mallow for cystitis.*

6 Agrimony *(Agrimonia eupatoria) grows to 90cm (3ft), with a downy stem, toothed, spear-shaped leaflets, greyish below, and a tall spike of yellow flowers; in dry grassy places – there* *are several different kinds. Use an infusion of the whole plant to ease constipation and acid stomach. Also use to treat cystitis, giving small doses frequently.*

7 Lesser Celandine *(Ranunculus ficaria) grows to 20cm (8in), with shiny, dark green, heart-shaped leaves and yellow flowers, in wet woodland and damp ground in Eurasia. Apply the expressed juice externally for haemorrhoids; do NOT confuse with its poisonous relatives, the buttercups.*

8 Solomon's Seal *(Polygonatum) are small, patch-forming, with arching stems bearing tube-shaped greenish-white flowers; in woody, scrubby areas. Use a decoction of the root externally for haemorrhoids and bruises, or take an infusion for nausea. The starchy root is edible like parsnips, but when boiled and dried it sets hard as makeshift plaster for a splinting agent. Either an infusion, or a poultice made from the powdered root, will ease bruising. Berries are POISONOUS.*

Silverweed *(see* **Edible roots in Food***) also provides an infusion for the treatment of digestive disturbance and haemorrhoids.*

FEVERS, COUGHS AND COLDS:

9 Feverfew *(Tanacetum parthenium) is very aromatic, growing to 45cm (18in), with delicate yellowish leaflets and many daisy-like flowers; in waste and grassy places in Eurasia. Eating the leaves eases headaches and migraines, but in some people this can cause blistering in the mouth. It is safer to use an infusion of the whole plant for fevers, headaches and general pains, or as a tincture for insect bites. Frequent small doses of a hot infusion help regulate contractions in childbirth.*

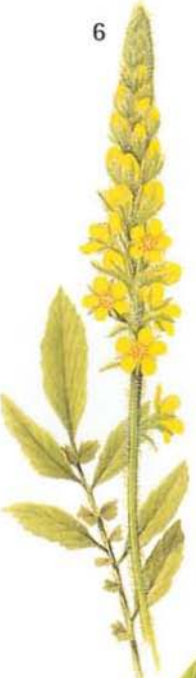
6

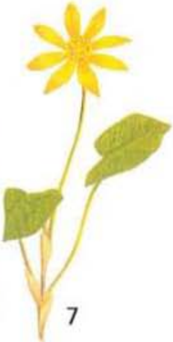
7

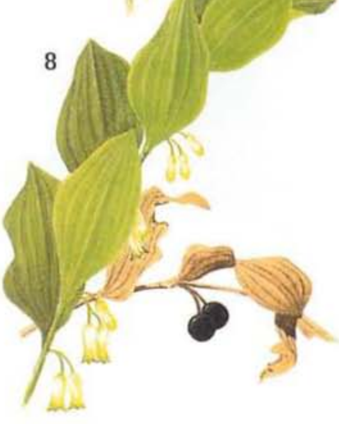
8

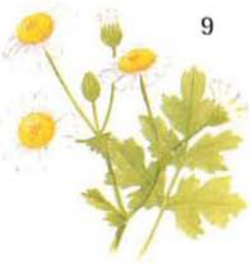
9

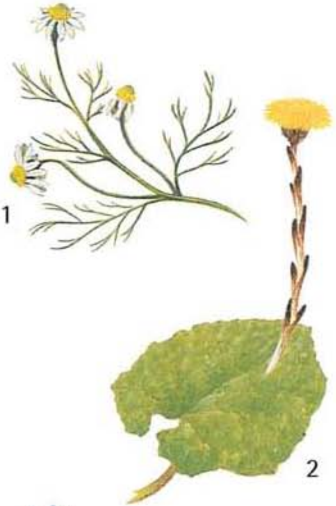

1

2

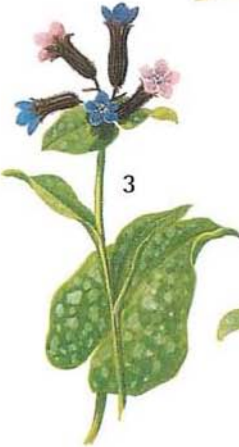

3

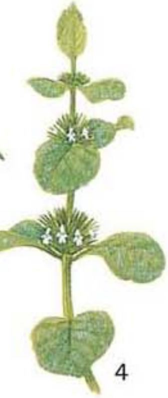

4

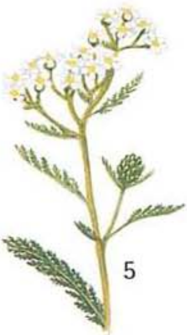

5

FEVERS, COUGHS AND COLDS CONTINUED:

1 Camomile *(Chamaemelum nobile) is aromatic, creeping, with finely dissected leaves and daisy-like flowers; in grassy places in Eurasia. Use an infusion of the whole plant for fevers, headaches, migraines, and colds, or the expressed juice of the flowers for aches and strains. It has a calming influence especially on nervously excited children.*

2 Colt's Foot *(Tussilago farfara) is common from late winter on bare and waste ground. Large, yellow dandelion-like flowers top asparagus-like stems; heart-shaped leaves follow the flowers. Use the leaves by infusion for colds and coughs.*

3 Lungwort *(Pulmonaria officinalis) is downy, up to 30cm (1ft), with pale-spotted spear-shaped leaves and bell-shaped pink or purplish-blue flowers; in mixed woods and scrub in Eurasia. An infusion of the whole plant is excellent for chest complaints and useful for diarrhoea. For coughs use with equal parts of Colt's Foot.*

4 Horehound *(Marrubium vulgare) is thyme-scented, downy, with squarish stems to 50cm (20in), roundish, crinkly, greenish-white leaves and whorls of whitish flowers; in dry scrubby places in Eurasia. Use an infusion of the whole plant for chills and respiratory disorders; oil expressed from the leaves soothes earache. It is a good cough treatment for children. In large doses it is a laxative.*

5 Yarrow *(Achillea millefolium) is downy, aromatic, up to 60cm (2ft), with dissected, feathery, dark green leaves and heads of tiny white or pink flowers; in grassy places. Use an infusion of the whole plant, but NOT the roots, for colds and fevers. It also hastens clotting of the blood in an injury and reduces blood pressure and bleeding in haemorrhoids.*

6 Musk Mallow *(Malva moschata) grows in grassy and scrubby places to about 60cm (2ft), with a hairy stem, deeply divided leaves and large, pink, five-petalled flowers. Mallows are widespread and come in many varieties. Use this like Marsh Mallow (8).*

7 Tree Mallow *(Lavatera arborea) grows to 3m (9ft), with a hairy stem, woody at the base, ivy-shaped leaves and pink-purple flowers streaked darker purple; in rocky coastal areas from Europe to Asia Minor. Use this like Marsh Mallow (8).*

8 Marsh Mallow *(Althaea officinalis) grows to 90cm (3ft), downy grey, with large lobed leaves and pale pink flowers. The cooked root is excellent to eat. Use an infusion of the whole plant for chest complaints or one just of the root to relieve giddiness caused by loss of blood and to clean wounds and sores. A rubbing with bruised leaves soothes insect bites; boiled leaves are a good poultice for skin eruptions. An infusion of the leaves will relax and soothe irritation and inflamation of the alimentary system.*

9 Great Mullein *(Verbascum thapsus) is covered in pale woolly down, growing to 2m (6ft), with large spear-shaped leaves and a dense spike of five-petalled yellow flowers; in dry, warm grassy places. Use an infusion of flowers and leaves for coughs and chest complaints, or a decoction of the root as a gargle. Powder the flowers to make a sedative and pain-relieving tea.*

10 St John's Wort *(Hypericum perforatum) grows to 60cm (2ft), with small, oblong, translucently spotted leaves and a head of golden-yellow flowers that exude a red juice when crushed; in open woods, grassy and bushy places. Use an infusion of the whole plant for colds and chest complaints.*

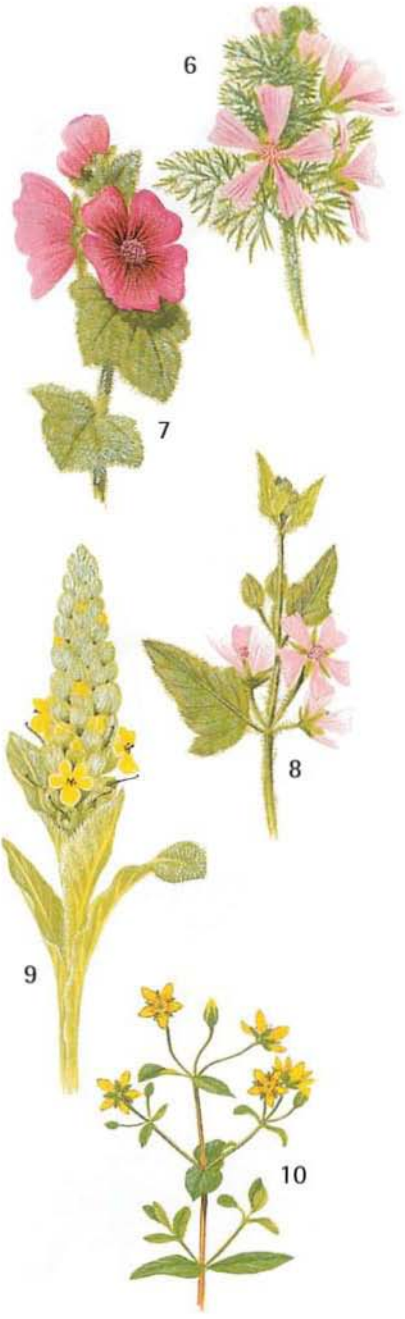

DANGEROUS CREATURES

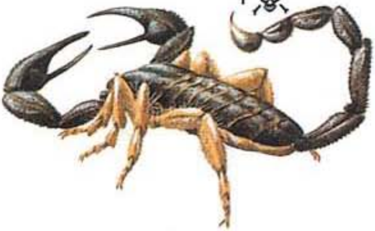

1 ☠

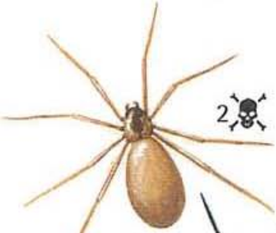

2 ☠

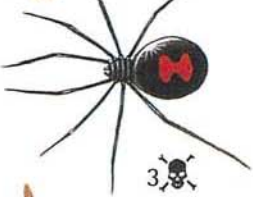

3 ☠

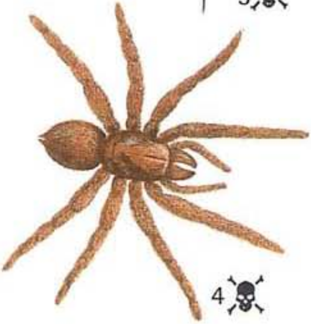

4 ☠

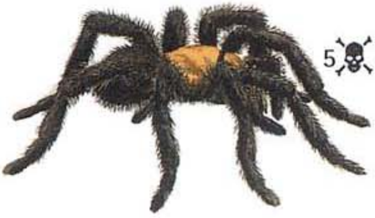

5 ☠

The insects and other creatures shown here are not a major problem for survivors if sensible precautions are taken – but can easily become one if not treated with respect.

1 **Scorpions** *are found in deserts forests and jungles of tropical, subtropical and warm temperate areas, one kind living at 3600m (1200ft) in the Andes, and are mainly nocturnal. Most desert kinds are yellowish to light green, those from moist or higher mountain areas brown or black. Average size is 2·5cm (1in) but giants in southern Africa and New Guinea reach 20cm (8in). Some burrow but they are usually found under tree bark, rocks, or other shelter, including your gear. The sting is in the tail. Many kinds cause only trivial discomfort, a few produce nerve toxins causing temporary paralysis for 24–48 hours. Some scorpions from the Middle East, Brazil and west Mexico can inflict a fatal bite, but this is very rare and death is more likely in young children and the old or ill, who offer little resistance to it.*

2 **Recluse** *or* **Fiddleback Spider** *(Loxoscees reclusa) of North America is recognised by a violin shape on the back of the head. There are several different kinds, but* L. reclusa *is the worst. Bite produces fever, chills, vomiting, joint pain and spotty skin, within 24–48 hours. Although rarely fatal, tissue degeneration around the wound can cause disfigurement, or even lead to amputation, if left untreated.*

3 **Black Widow** *or* **Hourglass Spiders** *(Latrodectus) occur in warmer areas, including deserts, over much of the world. Small, dark, all can be recognized by the red, yellow or white markings on the abdomen, hourglass-shaped in some. Bites produce severe pain, sweating, shivering and*

weakness, disabling the victim for up to a week. Rarely fatal.

4 Funnelwebs *(Atrax) are large greyish or browny spiders of Australia. Chunky, with short legs, their name alludes to their web's shape. Nocturnal, and not found in hot, dry, sunny conditions. A bite can kill; symptoms as for the Black Widow.*

5 Tarantulas *(Theraphosidae and Lycosa) are very large hairy spiders of tropical America; one kind occurs in southern Europe. Of menacing appearance, but although a bite is painful the poison is fairly mild and not disabling.*

6 Centipedes and millipedes *are mostly small and harmless but some tropical and desert kinds may reach 25cm (10in). Their feet have sharp claws, which can puncture the skin and cause infections, and a few kinds have a poisonous bite. Brush off in the direction they are moving – there is less chance of them digging into you.*

7 Hornets *occur in swarms and make nests that are guarded ferociously. Some tropical kinds are very aggressive and very poisonous, and should be well avoided. The sting is like being punctured by a hot rivet and several at once could be fatal.*

8 Ticks *are large and common in the tropics; flat-bodied and round, with a small biting head that eats into a wound. Do not pull off; the head will remain and cause infection. Use heat, petrol, alcohol or hot water to make it drop off.*

9 Leeches *are blood-sucking worm-like creatures of tropical jungles and other moist areas, waiting, thread-like, on vegetation before attaching themselves to a victim. Better not to pull off: remove with fire or a pinch of salt. Leeches often carry infections.*

10 Vampire Bats *(Desmodus) occur in Central and South America. Small, nocturnal, they suck the blood of sleeping victims. Their bites may carry rabies. Keep covered at night in these areas.*

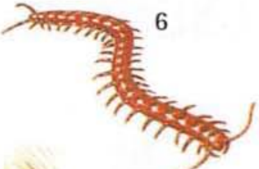

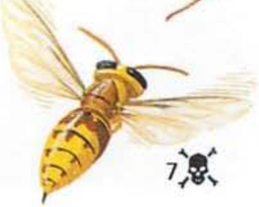

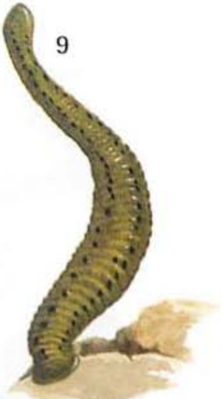

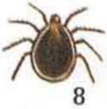

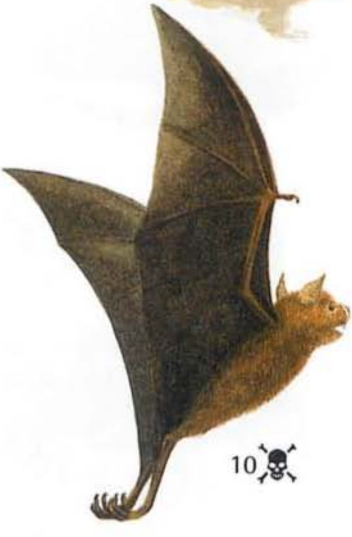

POISONOUS SNAKES

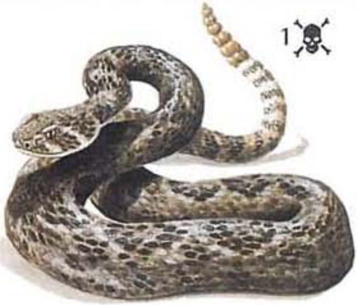
1

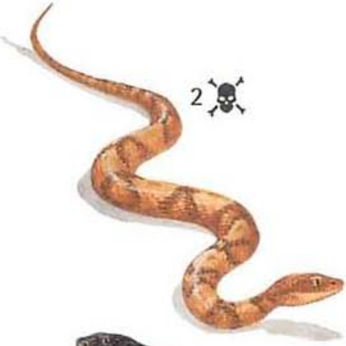
2

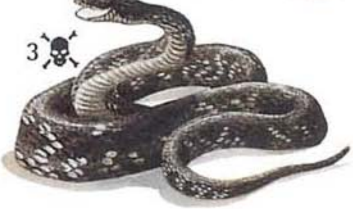
3

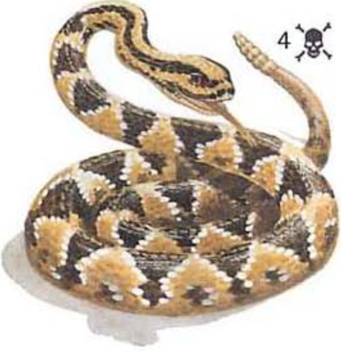
4

WARNING

Unless indicated to the contrary, these snakes should be regarded as deadly poisonous. Do NOT approach, provoke or handle.

SAFETY RULES

- **Watch where you step:** *On average snakes eat only once a week. After eating and at the times when they shed their skin they are sluggish and more easily trodden on.*
- **Look closely** *before parting bushes, picking fruit, some snakes are arboreal.*
- **Never tease, pick up or corner:** *A few snakes, such as the Bushmaster of South and Central America, Black Mamba of Africa and King Cobra of Asia will attack when cornered or guarding a nest.*
- **Use sticks, not hands** *to turn over stones and logs and for digging.*
- **Wear stout boots** – *if you have them. The teeth of many snakes are too small to penetrate them.*
- **Check bedding, clothes, packs** *before putting them on. Snakes may use them as shelter.*
- **Stay calm** *if you encounter a snake. Do not move suddenly or strike at it. Back off slowly. In most cases the snake will be only too eager to escape.*
- **To kill:** *If you have to kill a snake use a long stick, preferably with a spring to it, and a single chopping blow to the back of the head. Make it effective first time – a wounded snake is very dangerous.*

NORTH AND SOUTH AMERICA

1 **Rattlesnakes** *(Crotalus and Sistrurus). Many kinds occur in all parts of North America, varying from 45cm (18in) to over 2·1m (7ft). All have a chunky body, wide head and rattle on the end of the tail that is usually but not always sounded as a warning. The largest are the various Diamondbacks,*

with distinctive diamond-shaped blotches.

2 Copperhead *(Agkistrodon contortix) averages 60-90cm (2–3ft), with a stout body coloured buff or orange-brown with rich brown bands and a copper-red head; mainly in the eastern United States. Fairly timid; bites are only rarely fatal.*

3 Cottonmouth *or* **Water Moccasin** *(Agkistrodon piscivorus) averages 60–130cm (2–4ft) with a thick brown or brownish-olive body, sometimes blotched, and a yellowish, also blotched, belly; the inside of the mouth is white. Aquatic, in and by freshwater in the southern United States. Belligerent – do not annoy!*

4 Tropical Rattlesnake *(Crotalus durissis) averages 1·5–2m (5–6ft), with diamond-shaped marks, two dark stripes on the neck and a rattle on the tail; nocturnal, in drier areas from South America north to Mexico. Large, aggressive, very dangerous.*

5 Fer de Lance *(Bothrops atrox) is brownish with paler geometric markings and averages 1·3–2m (4–6ft); causes many deaths. Its many relatives vary from grey to brown or reddish with similar markings.* Bothrops *vipers occur in South America north to Mexico; some are arboreal. All loop their body before striking.*

6 Bushmaster *(Lachesis muta) is large-headed, pinkish-brown marked with large dark brown triangles and averages 2–2·6m (6–8ft) or more; nocturnal, in lowland forests, often using burrows and holes, in Central and South America. Vicious if cornered; the most feared of all New World snakes.*

7 Coral Snakes *average 45–90cm (1½–3ft), are slender and strikingly coloured in bands of black and red separated by bands of yellow or white; from the southern United States into South America. (Similar kinds occur in Southeast Asia.) Small-mouthed, reluctant to bite but deadly.*

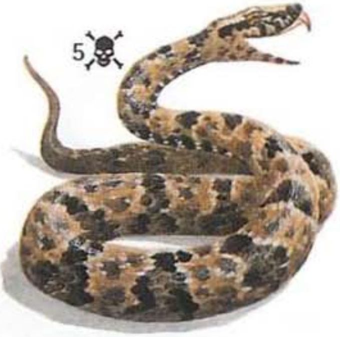

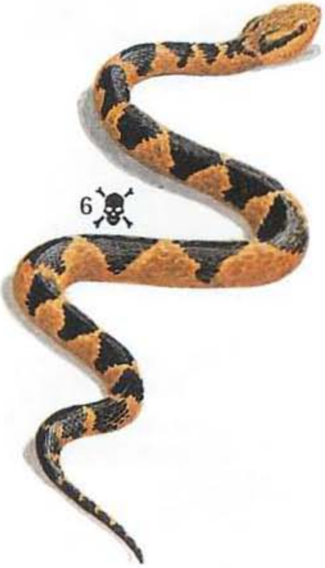

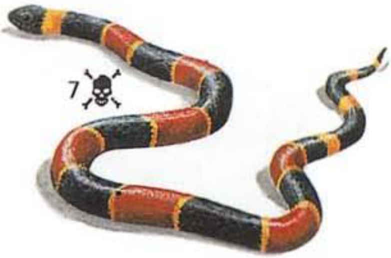

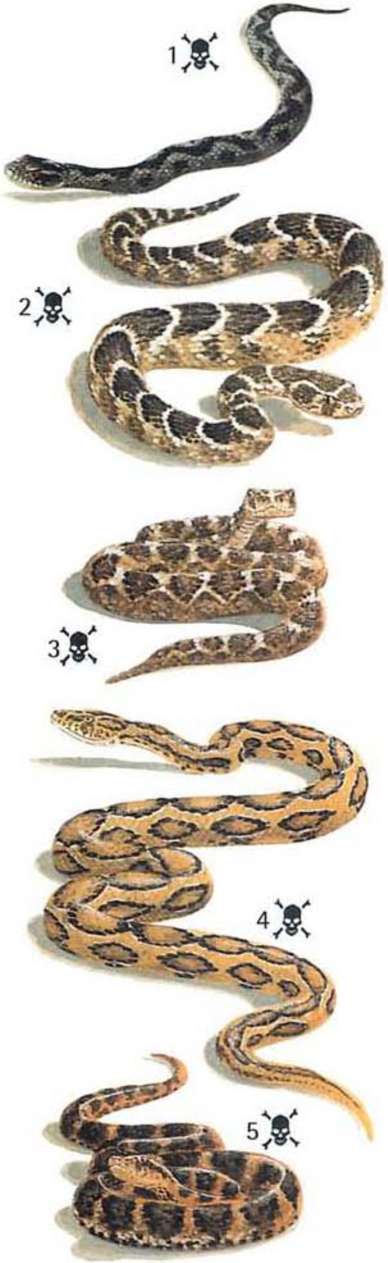

There are NO rules for identifying poisonous snakes. Cobras usually show hoods and rattlesnakes rattles on their tail, but these are NOT reliable guides. Poisonous snakes must be learned individually. If in doubt, treat every snake as poisonous.

EUROPE

1 **Adder** *(Viper berus) averages 30–75cm (12–30in), varying from olive-grey to reddish-brown with a zigzag pattern of darker colouring; especially on heaths, moors and open areas, into mountains. The only poisonous snake of northern Europe, hardly ever fatal, but with larger and more dangerous relatives in southern Europe.*

AFRICA AND ASIA

2 **Puff Adder** *(Bitis arietans) is thick-bodied, short-tailed and large-headed, straw-brown with darker markings and averages 90–130cm (3–4ft), in semi-arid areas often near water, of Africa and the Arabian Peninsula. Similar relatives occur in other habitats. Many different vipers are found in all parts of Africa and Eurasia, from sandy areas to thick jungle.*

3 **Saw-scaled Viper** *(Echis carinatus) is rough-scaled, pale reddish to sandy-brown with darker markings and white blotches, and averages 40–55cm (16–26in); in arid areas from North Africa west to India. Vicious, common; causes many fatalities.*

4 **Russell's Viper** *(Vipera russelli) averages 1–1·25m (40–50in), brownish, with three rows of spots formed of white-bordered black rings with a reddish-brown centre; in most areas except thick forest from Pakistan east to Taiwan. Responsible for most viper bites in the area.*

5 **Malay Pit Viper** *or* **Moccasin** *(Calloselasma rhodostoma) averages 60–80cm (24–32in) and is fawn,*

reddish or grey marked with geometric patterns, the belly yellowish or spotted greenish-brown in light growth in south-east Asia and parts of Indonesia. A frequent cause of bites and with many relatives in the area. AVOID any that resemble it.

6 Cobras *occur from Africa east through India to Indonesia and the Philippines. They usually average 1·5–2m (5–6ft) and, when alarmed, are recognizable by the raised head and spreading, often marked, hood. Common in some areas, especially rocky and semi-arid ones.*

7 Mambas *(Dendroaspis) are small-headed, very slender, typically with large green or greyish scales and averaging 1·5–2·1m (5–7ft); in Africa south of the Sahara, usually in trees but the large Black Mamba* D. polylepis *is largely terrestrial. Often quick to strike, fatal in almost all untreated cases.*

8 Boomslang *(Dispholidus typus) averages 1·3–1·5m (4–5ft), very slender, varying from greenish to brownish or blackish; in trees and very hard to spot, in savannah parts of Africa south of the Sahara. Highly venomous; it inflates its throat when alarmed.*

9 Kraits *(Bungarus) average 90–150cm (3–5ft), are small-headed and some have black and white or black and yellow bands down the body, in both open and forested areas from India to Indonesia. Nocturnal, inoffensive, but bites are often fatal.*

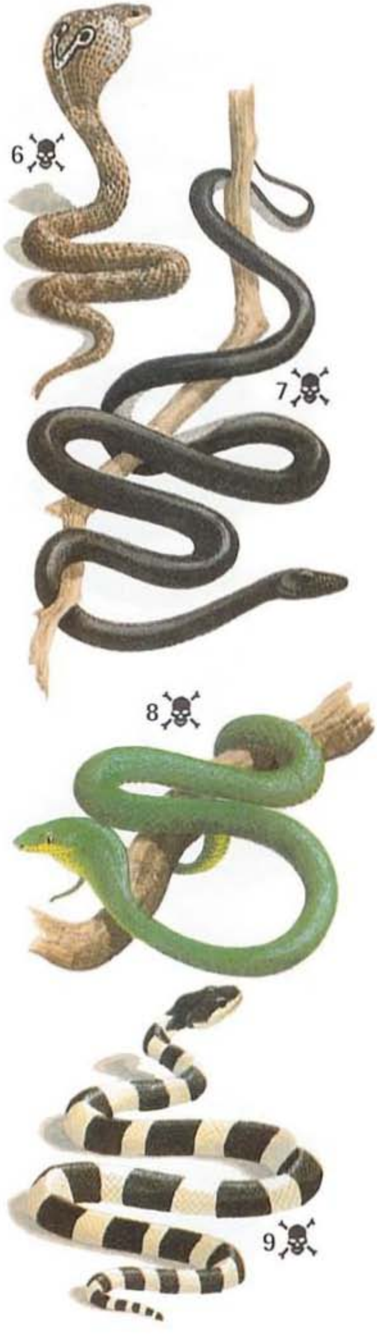

SPITTING SNAKES

A few cobras, including the ringhals of southern Africa, spit poison as well as bite. This is a purely defensive measure and is not dangerous unless the poison reaches an open cut or the eyes. If it does, wash out immediately with water or, in an emergency, with urine.

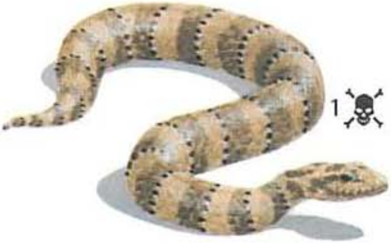

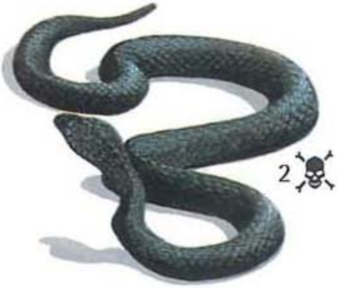

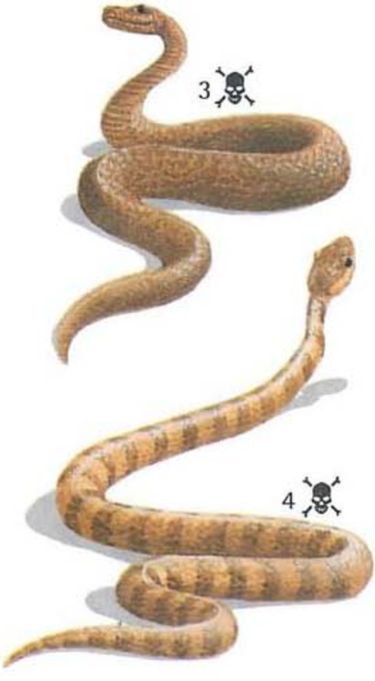

AUSTRALASIA

1 Death Adder *(Acanthophis antarcticus) is brownish, reddish or grey with darker banding, thick-bodied and averages 45–60cm (18–24in); in sandy areas of much of Australia, Papua New Guinea and some nearby islands. Well camouflaged; highly venomous, but not so dangerous as the Tiger Snake and Taipan.*

2 Australian Black Snake *(Pseudechis porphyriacus) averages 1·5–2m (5–6ft), slender, bluish-black with a bright red belly; in or near fresh water over much of Australia. There are several different kinds. Very rarely fatal, it flattens its neck when aroused.*

3 Australian Brown Snake *(Pseudonaja textilis) is slender, yellowish-grey to brown with a pale belly and averages 1·5–2m (5–6ft); in drier parts of Australia and Papua New Guinea. There is more than one kind. Aggressive and very poisonous.*

4 Tiger Snake *(Notechis scutalus) averages 1·3–1·6m (4–5·5ft), thick-bodied, large-headed, tawny-ochre banded with greenish-yellow, grey or orange-brown; in semi-arid areas of Australia and in Tasmania. Aggressive, very poisonous, the principal cause of fatal bites.*

5 Taipan *(Oxyuranus scutellatus) is uniformly light to dark brown with yellowish-brown on the sides and belly and may grow to 3·5m (11ft); in open and forested parts of northern Australia. Ferocious when provoked, deadly poisonous.*

6 Sea Snakes *occur in the Indian and Pacific Oceans; some are partly terrestrial, in estuaries and coastal swamps. They vary in colour and size, averaging 1·3–1·5m (4–5ft) with a flattened, paddle-like tail. Their scales distinguish them from eels. Not aggressive, but some are the most venomous snakes of all.*

Snakes have excellent camouflage, only movement gives them away. In snake-infested areas you will pass many every day without ever noticing them.

The chances of being bitten are small and all but the worst cases recover. In Malaysia, more people are killed each year by falling coconuts and in India rat-bites produce many more cases for hospitalization!

A bite from a poisonous snake should always be taken seriously, but there are degrees of severity. When biting in self-defence, many snakes inject only a little venom, occasionally none at all. If the snake is out of condition or has recently bitten something else, its venom may not be fully potent and there may be only a little in its venom sacs. Clothing or shoes may have deflected the full force of the bite. In many poisonous snakes the dose of venom needed to kill a man far exceeds the amount that can be injected in one bite.

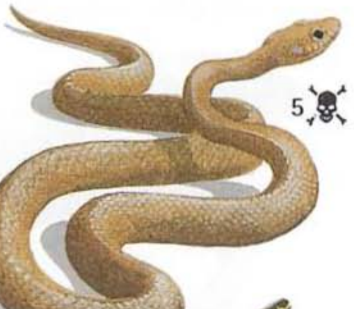

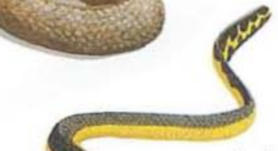

SNAKE FREE!

There are NO poisonous snakes in New Zealand, Cuba, Haiti, Jamaica, Puerto Rico, Ireland, Polynesia and the polar regions.

LIZARDS

7 Gila Monster *(Heloderma suspectum) is a lizard found only in the deserts of Arizona, Mexico and nearby areas. Large rounded head, thick chunky body, short stumpy tail and brightly patterned in yellow. Averages 37–45cm (15–18in). Bite is poisonous but likely only when handled.*

8 Beaded Lizard *(Heloderma horridum) resembles the Gila Monster but is darker and larger, with a slenderer tail, and spots rather than a mottling of colour; in a few arid parts of Mexico and Central America. Docile, but the bite is poisonous. Do not handle.*

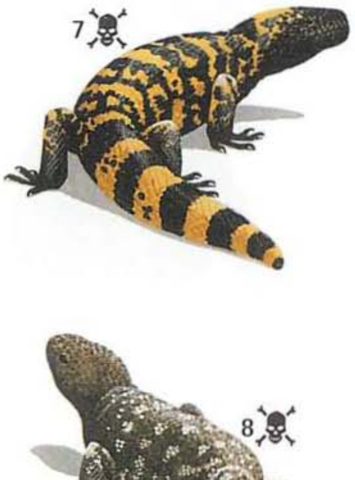

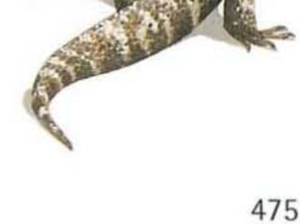

DANGEROUS SEA CREATURES

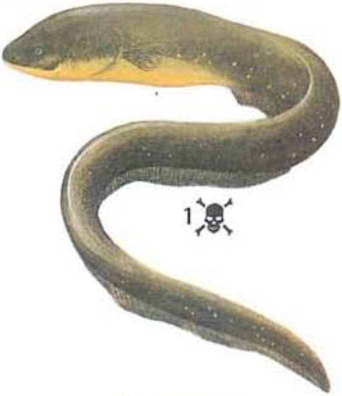

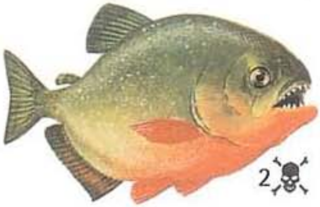

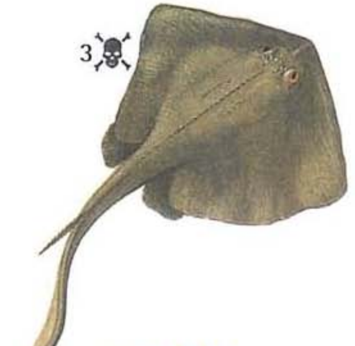

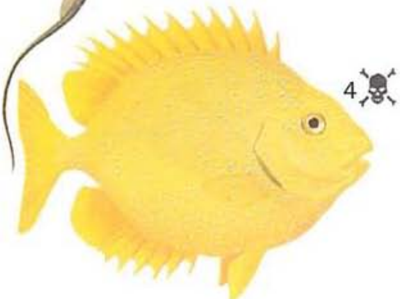

These fish and sea creatures are dangerous. Most are either poisonous to touch or have poisonous flesh.

RIVER DANGERS

1 **Electric Eels** *(Electrophorus electricus) may reach nearly 2m (7ft) long and 20cm (8in) thick, rounded, coloured olive to blackish, and paler underneath, native to Orinoco and Amazon river systems of South America. Often prefer shallow water where there is more oxygen. The shock from a large one can be 500 volts, enough to knock a man off his feet.*

2 **Piranhas** *(Serrasalmus) occur in the Orinoco, Amazon and Parguay river systems of South America. They vary in size but may be up to 50cm (20in) long and are all deep-bodied and thickset, having large jaws with razor-sharp interlocking teeth. They can be very dangerous, particularly in the dry season when the water levels are low.*

SEA AND RIVERS

3 **Stingrays** *(Dasyatidae) are a danger in shallow waters, especially tropical ones, and not only on sandy shores. Very variable, but all with the distinctive ray shape though hard for the wader to spot. A few kinds occur in rivers in tropical South America and West Africa. The freshwater stingrays rarely exceed 30cm (1ft) long. They do not occur in rivers that flow into the Pacific. Venomous spines in the tail can inflict severe, sometimes fatal, injury.*

SALTWATER DANGERS

4 **Rabbitfishes** *or* **Spinefeet** *(Siganidae) occur mainly on reefs in the Indian and Pacific Oceans, averaging 25–30cm (10–12in); edible but with sharp spines in most fins. These are said to be venomous. Handle with care.*

5 **Tangs** *or* **Surgeonfishes** *(Acanthuridae) average 20–25cm (8–10in), deep-bodied, small-mouthed, very colourful, with lancet-like spines on the sides of the tail that can inflict severe wounds when it is lashed. In all tropical waters.*

6 **Venomous Toadfishes** *(Batrachoididae) occur in tropical waters off both coasts of Central and South America. Averaging 3–4cm (7–10in), they are dull-coloured and large-mouthed. They lie buried in sand and have sharp, very poisonous spines on the back.*

7 **Scorpionfishes** *or* **Zebrafishes** *(Scorpaenidae) are found mostly on reefs in the tropical Indian and Pacific Oceans. Averaging 30–75cm (12–30in), very variable, but usually reddish with long, wavy fin rays and spines. A sting is intensely painful. Less potent relatives occur in the Mediterranean and Atlantic.*

8 **Stonefishes** *(Synanceia) occur in the tropical Pacific and Indian Oceans. Reaching 40cm (16in), their drab colours and lumpy shape make them almost impossible to see. When trodden on, dorsal spines inject venom that is agonizingly painful, in the worst cases fatal.*

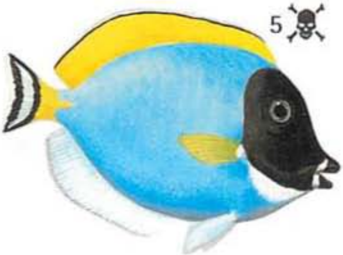

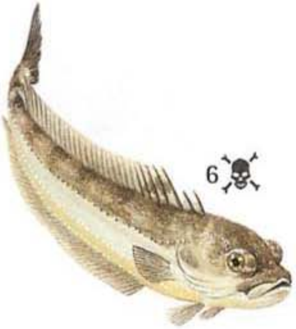

The venomous Toadfish, Stonefish and Zebrafish are edible. If you land one strike it on the head and handle only when completely dead, and then with great care.

Though not venomous, there are other fish with dangerously sharp spines, which are not always easy to detect, except at close quarters. The spines usually occur on the back but may also be on fins on the side of the fish. Even a small spine can inflict a bad prick with consequent risk of infection. Large spines – and some spiny catfish grow as large as a man – are as effective as stilettos.

Sea urchins can also inflict painful injuries and Sea anemones can sting.

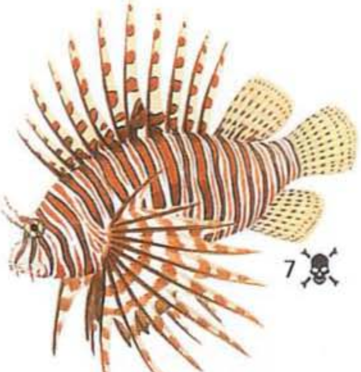

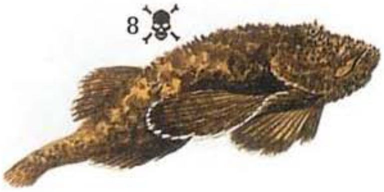

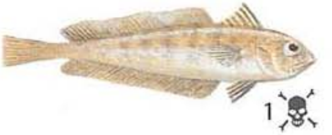

1 **Weeverfishes** *(Trachinidae) tapering, dull-coloured, about 30cm (1ft) long, lie buried in sand off the coasts of Europe south to West Africa and the Mediterranean. Venomous spines on back and gills produce disabling pain. Soothe it by applying very hot water.*

POISONOUS TO EAT

Many inshore fish, living in reefs and lagoons, are poisonous to eat. The majority are confined to the tropics but, wherever you are, be wary of eating any fish that you cannot identify.

Some fishes that are otherwise good to eat, such as the Barracuda and Snapper, are inedible when taken from reefs and lagoons, where they will have absorbed poisonous substances with their own food.

The most poisonous kinds, such as puffer fish, usually have rounded bodies with hard, shell-like skins covered in bony plates and spines. They also commonly have parrot-like mouths, small gill openings and either lack pelvic fins or have only a small one.

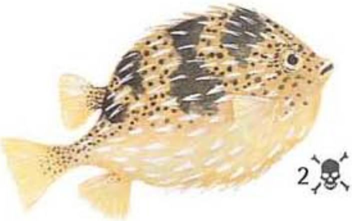

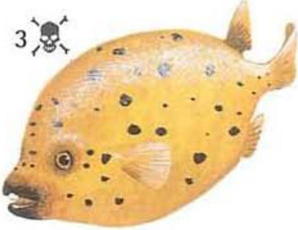

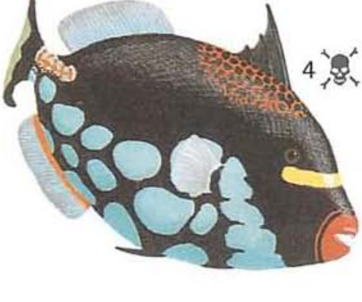

2 **Porcupine Fishes** *(Diodontidae) occur in all shallow tropical waters. Variable, reaching 50–60cm (20–24in), but when alarmed all inflate into a very spiny ball. Their flesh is poisonous.*

3 **Puffer Fishes** *(Tetraodontidae) occur in all tropical and many warmer temperate waters, a few kinds in rivers in south-east Asia and tropical Africa. Stout-bodied, rounded, 15–75cm (6–30in) long, most kinds with spines; when alarmed they puff up into a ball. Their blood, liver and gonads are poisonous; 28mg (1oz) can kill.*

4 **Triggerfishes** *(Balistridae) occur in huge variety, mostly in shallow tropical seas. Deep bodied, compressed, usually under 60cm (24in), with very large, stout dorsal spines. Many kinds are poisonous to eat. Avoid them all.*

OTHER SEA CREATURES

5 Portuguese Man-of-war *(Physalis physalis), not a jellyfish but a colony of hydroids, is mainly subtropical but common in the Gulf Stream which may take it to British shores. Southern currents carry it, for example, to New Zealand. The floating bladder may be only 15cm (6in) long, but the tentacles, which carry stinging cells, can stream out for 12m (40ft). Not fatal but enough to incapacitate, so extremely dangerous.*

The Common Jellyfish *(Aurelia aurita), a milky saucer-shape with purplish horseshoe marks inside, is not dangerous to humans but many others are, especially the* **Sea wasps** *or* **Box Jellyfish** *(*Chironex fleckeri*), with a cube shaped bell 25cm (10in) long and clusters of tentacles at the corners up to 9m (30ft) long. In big doses venom can be fatal. Avoid all jellyfish streamers – even when washed up on the beach.*

6 Blue-ringed Octopus *(Hapalochlaena lunulata), small, sometimes only fist-sized, found off eastern Australia, particularly around the Great Barrier Reef, is greyish white with iridescent ring-like markings. Very poisonous, potentially lethal bite if trodden on or handled. Treat all tropical reef octopuses with caution.*

7 Cone shells *(Conidae), subtropical and tropical gastropods, have a venomous harpoon-like barb. All are cone shaped but shell patterns may be obscured by a membrane. Some very poisonous, a few, in the Indo-Pacific, lethal. DO NOT TOUCH.*

8 Augers *or* **Terebra Shells** *(Terebridae) temperate and tropical seas, particularly the Indo-Pacific, also have a stinging barb. They are much thinner and longer than cone shells. The sting is not as serious as the Cone's, but do NOT eat.*

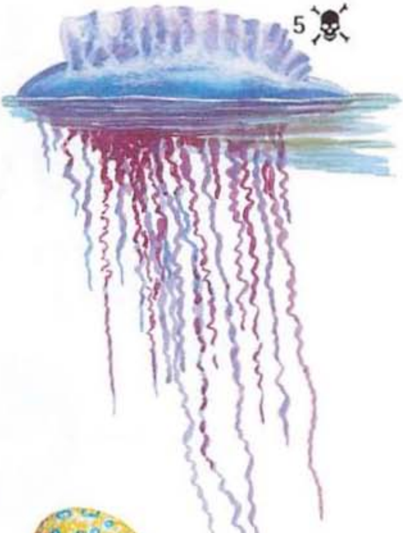
5

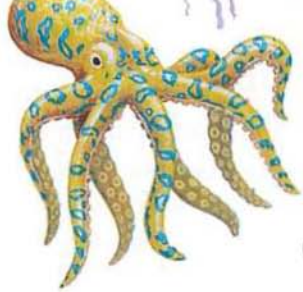
6

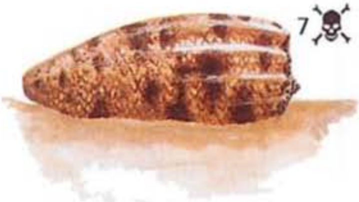
7

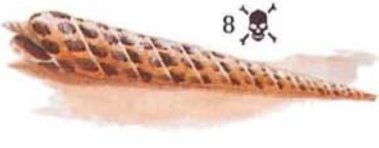
8

9

SURVIVAL AT SEA

Conditions of survival at sea are perhaps worse than those of any other environment and make the sternest demands. Planes and boats carry survival equipment, but even getting into a dinghy in a heavy sea can be difficult.

Once any emergency supplies of food and water run out, sources are not reliable – so any possibilities of obtaining food from the sea and collecting drinking water must be exploited to conserve supplies as long as possible.

Not all fish are edible and some are even dangerous to handle. Shark dangers are often exaggerated, but should not be ignored. Appropriate action is needed to avoid or deter them. A difficult coast can make even a final landfall hazardous, so heed the advice on lessening the risks.

SURVIVAL AT SEA 482
Man overboard! 484
Survival afloat 487
Protection 488
Is land nearby? 489
Travelling 490
Signalling at sea 491
Health 492
Water 492
Food 494
Dangerous fish 497
Sharks 498
Making a landfall 503

SURVIVAL AT SEA

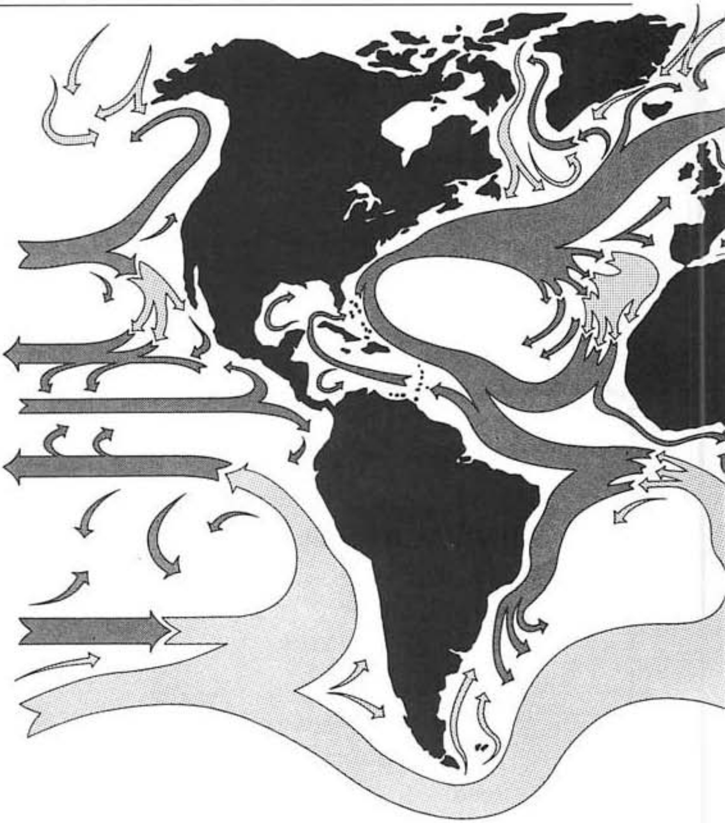

Four-fifths of the earth's surface is open water – probably the most frightening of all environments, and the most difficult in which to survive. In cold water the body soon becomes chilled and even in a boat wind can chill the body rapidly. Alone in cold water your chances are not good without equipment.

If you know your location and the main ocean currents you may be able to predict where they will carry you, though it will be very slowly. Warm currents, such as the Gulf Stream, across the North Atlantic, are rich in fish and sea creatures.

Coastal waters are also often rich in sea foods – but there are dangerous species, such as sharks, and poisonous species, mainly living in shallower water near lagoons and reefs in warmer climates. Fresh water is a bigger problem if you have no means of distilling sea water.

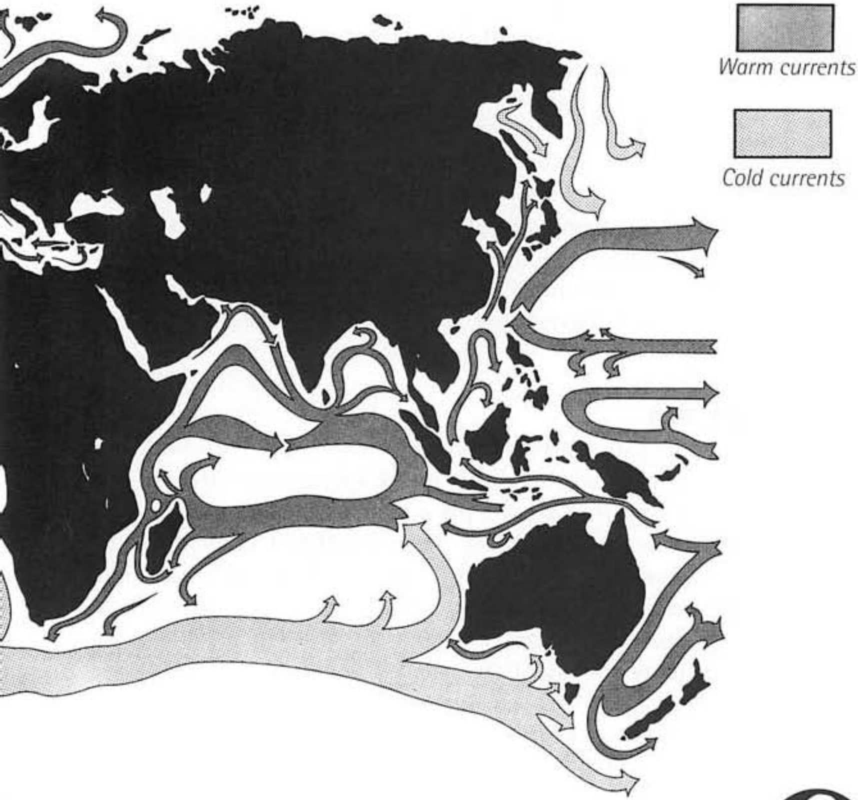

Lifeboat drill

Lifeboat drill is carried out on every ship soon after it sails and should become a well-rehearsed procedure. Passengers are instructed in how to fit lifejackets, how they are to proceed to their lifeboat stations, and what to take with them. Sailors in small boats should also devise such a drill and instruct everyone on board.

SURVIVAL SCENARIO

Is it easier to survive in warm or cold waters?

A survival situation at sea in warmer climates is preferable to cold-water survival. You can only survive for a short time in cold water before hypothermia sets in. If you do find yourself going into cold water take as much survival kit as you can, and then get out of the water asap. If you are in a lifeboat, establish a routine so that order is maintained in the boat.

If the signal is given to abandon ship put on warm, preferably woollen, clothing including hat and gloves and wrap a towel around your neck. Clothes will not drag you under if you end up in the water and they will help ward off the worst enemy – exposure. Take a torch if you can and grab chocolates and boiled sweets if they are handy. Do NOT push or shout or you may start a panic – an orderly embarkation into lifeboats and on to rafts or dinghies will be faster in the long run and establish a calmer attitude.

Don't inflate your lifejacket until you leave the ship or plane. On small boats lifejackets should be worn all the time. They are brightly coloured and are usually equipped with a whistle, light, marker dye and – when for use in warmer waters – a shark repellent.

If you have to jump overboard, first throw something that floats and jump close to it.

Abandoning ship

Abandoning a ship or ditching from a plane, it is essential that you take what equipment you can with you. A lifejacket or lifebelt will save a lot of energy that you might otherwise expend in trying to keep afloat. But even without one it is not difficult to float in the ocean. The human body is of lower density than saltwater and anyone who has learned to relax in the water is not in immediate danger from drowning. However, panic or fear make relaxation difficult and many find floating difficult under these conditions. Without a lifejacket or lifebelt, air trapped in clothing will help buoyancy – a good reason for keeping your clothes on despite the frequent advice that you should strip them off.

MAN OVERBOARD!

If you have been swept overboard your first aim, apart from keeping afloat, will be to attract attention. Sound travels well over water – shouting and splashing can be effective. Wave with one arm above the water (not both, you will go under) – movement will make you more noticeable.

If you are wearing a lifejacket – and on a small boat you always should be – it will probably be equipped with a whistle and a light, as regular-issue 'Mae Wests' usually are.

Swimming

Swim slowly and steadily. If you are abandoning a sinking boat or aircraft get upwind and stay clear of it. Keep away from any oil or fuel slick.

If there is a fire and you have to enter the water, or swim through flames, jump into the water feet first and up wind, swimming into the wind using a breast stroke, try to make breathing holes by splashing the flames away from the head. If the fire is not too extensive it is best to swim underwater until clear of that danger.

If there is a danger of an underwater explosion while you are in the water, the risk of injury will be reduced if you swim on your back.

If within sight of land don't battle against the ebb, relax and float until it turns and helps to carry you to land. If the sea is too rough to float on your back adopt this technique.

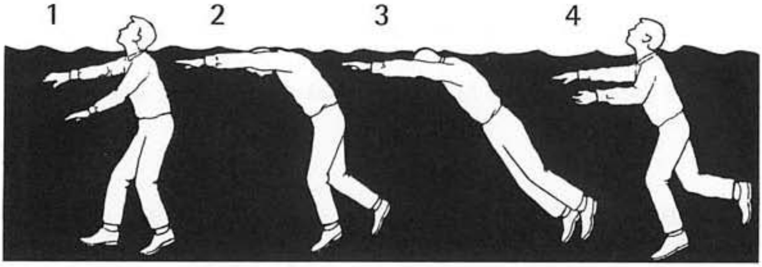

1 *Float upright in the water and take a deep breath.*
2 *Lower your face into the water (keeping your mouth closed) and bring your arms forward to rest at water level.*
3 *Relax in this position until you need to take in more air.*
4 *Raise your head above the surface, treading water, and exhale. Take another breath and return to the relaxed position.*

Flotation 'bags'

You can improvise a short-term flotation bag from a pair of trousers. Knot the bottoms of the legs, sweep them over the head to fill with air, then hold the waist below the water to trap the air inside, making the legs into water wings to lean on.

Immediate action

Once you are clear of the wreck and have got your bearings, inflate your dinghy or look out for a boat or raft or wreckage which can offer support. If there is no boat or dinghy grab as much flotsam as possible to use as a raft. Tie it together with anything that is available – ties, belts, shoelaces, spare clothing. Salvage any floating equipment.

Inflating a dinghy

Aircraft and many boats and ships carry dinghy-type lifeboats. Many are self-inflating and activated by saltwater immersion.

If they do not inflate automatically, there is a pump provided. There are several inflation points because the dinghy is built in sections, so that if one compartment is punctured the others will still keep the dinghy afloat.

Boarding an inflatable dinghy

Get aboard as soon as possible. If you are already in the water move to the end (not the side) of the dinghy, place one leg over the edge and roll into the dinghy.

Do NOT jump into a dinghy from above, you may damage it.

To haul someone else aboard a dinghy, raft or lifeboat hold their shoulders and lift one leg over the end, then roll them in. Discourage them from putting their arms around your neck – they could pull you into the water. Then tie yourself and others to the dinghy.

RIGHTING AN INFLATABLE DINGHY

Most dinghies have righting straps on the bottom, and larger ones have a righting line attached to one side. Grab it from the opposite side, brace your feet against the dinghy and pull. The dinghy should rise up and over, pulling you out of the water momentarily. In heavy seas, or a high wind, this can be extremely difficult.

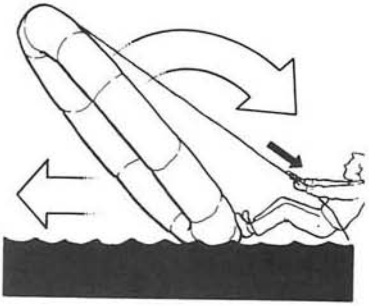

Ensure that the dinghy is fully inflated. It should be firm – not rock-hard. If it is not you will need to inflate it with your own breath or a pump. The valves are one-way and air will not escape when you take off the protective cap.

Check for leaks. Escaping air will cause bubbles under water and above water will make a hissing sound. Deal with them with conical plugs that you will find in the dinghy kit. They screw into holes and seal them. You will probably also find a supply of rubber patches and adhesive.

Make daily checks of inflation and leaks. If you suspect a leak on the underside swim under and insert a plug.

SURVIVAL AFLOAT

Rafts, boats and dinghies are built to carry a limited number of survivors. The lives of those aboard will be even more endangered if these numbers are exceeded.

The safety of the majority must be the priority. Place the infirm, youngsters and any injured in the dinghy or boat first and as many of the able-bodied as the boat is made to accommodate. The rest must hang on in the water. The fit survivors aboard should rotate with those in the water on a regular and frequent change-over rota. But remember that exposure to cold water, to the inexperienced and even the young and fit, can be fatal.

Stow all the gear in any stowage places provided and tie everything securely. Check that there are no exposed sharp objects which will damage an inflatable. Ensure that anything that will spoil if wet is in a waterproof container and kept out of the water.

Check all signalling equipment: flares, rockets, heliographs. If distress signals have already been sent out you will need them to attract the attention of rescue parties when they are searching for you.

If a distress call has gone out giving your position it is best to try to maintain location, so put out a sea anchor. This should look like a large canvas bag. Streamed out from the boat it will keep it into the weather and slow down drift.

You can improvise a sea anchor from any weighted object securely tied to a line. Even clothing could be used, possibly tied to a paddle with reef knots.

If you do not know where you are do NOT attempt to navigate until you have established your position, but if you can see the shore head towards it.

SURVIVAL PRIORITIES

- PROTECTION from the elements and the effects of exposure.

- LOCATION Try to establish where you are and the best way of attracting rescue.

- WATER Take stock of supplies. Ration it at once. Start collecting any rain.

- FOOD Don't eat, unless you have sufficient water. Check all rations available, stow them securely. Start fishing as soon as possible.

PROTECTION

Even if you are alone keep a log on a daily basis. This will occupy the mind and help keep you oriented. First record names of survivors, date and time and position of accident, weather conditions, equipment salvaged, and record sightings and circumstances daily.

In a cold climate

If the water is cold it is essential to get out of it as soon as possible. You need to counter the chilling effect of the wind, especially if you are wet. Keep the boat or dinghy as dry as possible. Bail out all the water and rig up an awning to keep out spray if you can find any material to use for it.

Dry all wet clothing and if there is no dry clothing to put on squeeze out as much water as possible and then put the clothes back on.

Maintain body heat by wrapping all parts in any available material, such as parachute or canvas. If in a group of survivors huddle together to keep warm. To prevent stiffness to muscles and joints, and to keep the circulation going, do mild exercises, such as stretching and arm circling. Be careful not to disturb the balance of the raft or boat by excessive or sudden movement.

Most modern dinghies have a built-in shelter. If yours does not, rig up a windbreak and a spray shield. Stretch any material that is available across to keep out spray and breaking waves. With adequate shelter and warm clothing, exercise will protect against the risk of frostbite.

In a hot climate

Take off unnecessary clothing, but still keep the body covered. If exposed directly to strong sun always keep the head and neck covered to avoid sunstroke or burn. Protect the eyes from the sun glare by improvising eye shields.

During the day damping down clothes with sea water will help to keep the body cool but make sure that you are thoroughly dried out by evening, for nights can be very cold – and remember that darkness comes quickly in the tropics. Remember also that prolonged contact with sea water can cause sores on the skin.

When it is very hot let out some air from an inflated dinghy, for air expands with the heat – you will need to release the valves. Reinflate in the evening when it cools.

WATCHES AND LOOK OUTS

In a group assign watches. There should be a look out all the time – even in darkness. Each watch should be for a short period to avoid exhaustion and lack of concentration. It is better for everyone to have several watches a day than for any one person to have long periods on duty.

It is the responsibility of the watch to look out for shipping, aircraft, signs of land – and for seaweed, shoals of fish, birds, wreckage. They should also inspect the raft for signs of leakage or chafing.

IS LAND NEARBY?

When there is no land in sight you may find some of these indicators of land and the direction in which it may be found:

Clouds:

Cumulus clouds in an otherwise clear sky are likely to have been formed over land.

In tropical waters a greenish tint on the underside of clouds, known as lagoon glare, is produced by the reflection of sunlight from the shallow water over coral reefs.

Birds:

A lone bird is not a reliable indication of land, and after rough weather birds can be blown way off course, but few seabirds sleep on the water or fly more than 100 miles from land. Their direction of flight is usually outwards from land before noon and return in the late afternoon. The continuous sound of bird cries is usually an indication that land is not far distant.

Driftwood:

Driftwood, coconuts and other drifting vegetation are often a sign that land is near (though they can be carried right across an ocean).

Sea movement:

The pattern of the swell may indicate land. A change in its direction may be caused by the tide pattern around an island.

Prevailing winds build up a swell pattern and the swell is less if the water is protected by land. If the wind is constant but the swell and waves decreasing you can be fairly certain that land lies to windward.

Sea colour:

Water that is muddy with silt is likely to have come from the mouth of a large river.

TRAVELLING

If an SOS has been successfully sent, or you know that you are in or near regular shipping lanes, it is usually preferable to stay in the same vicinity for 72 hours.

If none of these circumstances hold, then no time should be lost in getting underway to take advantage of initial fitness and energy, especially if land is known to be near and downwind.

If there is no land nearby, assess the nearest shipping lane and head in that direction.

DECIDING FACTORS

Take these factors into consideration in making your decision whether to stay or travel:

- The amount of information signalled before the accident.
- Is your position known to rescuers? Do you know it yourselves?
- Is the weather favourable for a search?
- Are other ships or aircraft likely to pass your present position?
- How many days supply of food and water do you have?

Your craft will move with the wind and current. In the open oceans currents seldom exceed a speed of 9–13km (6–8) miles per day. Take in the sea anchor. Use the wind if you can. In a craft with no keel it is only possible to sail full tilt with the wind or at most at an angle of 10 degrees off it.

Use a paddle as a rudder. If the wind is against your chosen direction stream the sea anchor to maintain position.

To use the wind

Inflate the dinghy fully and sit high. Improvise a sail if you do not have one. Do not secure its lower edges but hold the lower lines or the bottom of the sail so that if there is a sudden gust of wind you can release them and the raft is not capsized.

In rough water

Stream out the sea anchor from the bow. It will keep the bow always into the wind and prevent capsizing. Keep low in the raft. Do not sit on the sides or stand up. Never make sudden movements. If there are several rafts or dinghies, tie them together.

SIGNALLING AT SEA

Flares, dye markers and movement of any kind are the ways of attracting attention at sea. If you have no signalling equipment, wave clothing or tarpaulins and churn the water if it is still. At night or in fog a whistle is useful for maintaining contact with other groups of survivors.

If a radio transmitter is part of the equipment aboard a life-raft instructions for its operation will be found on its side. Frequencies are usually preset at 121.5 and 243 megacycles and the range is about 32km (20 miles). Transmit at frequent intervals but exercise discretion in using battery-operated transceivers. The batteries are precious. (See *Signalling* in *Rescue*.)

Sea markers, which release dye into the water, are only of use in daytime. Unless the seas are very rough they will be conspicuous for about three hours.

Pyrotechnic equipment must be kept dry and secure. Read instructions carefully and beware of fire hazards. Day-and-night flares are particularly useful – one end produces smoke for daytime use. When firing flares and rockets you are holding a dangerous firework in your hand, so be very careful that you do not point them downwards or towards yourself or another person.

Use flares only when you are sure that they will be seen and fire to make sure that they are – when a plane is flying towards you, for instance – not when it has gone past. (See *Signalling* in *Rescue*.)

Most rescues have followed after heliograph signals have attracted attention. Any shiny, reflective surface can be used to signal in this way.

HEALTH

Exposure and severe hydration are likely to be the biggest problems for the ocean survivor. Seasickness can increase dehydration.

Constipation, and often either difficulty in urinating or very concentrated urine, are not unusual in sea survival conditions. Do not attempt to treat them or you could force further liquid loss.

If feeling sick, try not to vomit and NEVER induce vomiting.

Continued exposure to saltwater can produce skin eruptions. Do not attempt to prick or squeeze any boils or blisters. As a precaution do not damp yourself down too often with salt water to keep cool. If there is any soreness, STOP.

Protect the eyes from glare off the sea with a mask and, if sore eyes are produced by glare, moisten a cloth with sea water and bandage this over the eyes and rest them. Do not do this for too long. The skin may become sore.

Immersion foot (see *Health*) can be a problem if subjected to very long exposure in a boat or dinghy awash with water. Exercise will help protects you from it and from frostbite and exposure. Keep well covered when resting and, when on watch, gently exercise the limbs.

WATER

Although a minimum 1 litre (1¾pt) a day is necessary to keep fit, it is possible to survive on 55–220cc (2–8oz).

Even if you have a good water supply, ration it at once (see p.40), reflecting these minimum needs until you can replenish your supply. Do not relax the ration until final rescue, for you have no idea how long you will have to last out.

Reducing water needs

Take all the usual precautions against water loss (see *Water* in *Essentials*). Reduce sweating as much as possible. Make use of breezes and use sea water to cool the body. If it is very hot, shade limited and the waters safe take a dip over the side – but first check your safety line. You should ALWAYS be tied on. Beware of dangerous fish and be sure that you can get back aboard.

If seasickness threatens take anti-sickness pills, if they are available, as soon as you start to feel queasy, for vomiting will lose valuable fluids.

If you are low on water do not eat, especially not protein foods – which include fish and seaweed – that require a lot of water to digest. Carbohydrates (sugars and starches) require less water for digestion.

WATER RATIONS

DAY 1:
NO WATER. The body is a reservoir and has a store.

DAYS 2–4:
400cc (14fl oz) if available.

DAY 5 ONWARDS:
55–225cc (2–8fl oz) daily, depending on the climate and water available.

When drinking, moisten the lips, tongue and throat before swallowing.

Gathering freshwater

Use every possible container to collect rainwater night and day – you will usually see a rain squall coming and have time to rig up a catchment from canvas or plastic, which will hold much more than cans.

At night rig canvas with edges folded to catch any dew.

When it rains drink your fill – but slowly, for if you have been on short water ration you will vomit if you gulp it down.

Stow as much in containers as you can. Drink up puddles in the boat first. But be careful in heavy seas as the water will be contaminated with salt. Water is good ballast in an inflatable – fill it to the brim and it will still float.

Sea Ice

Ice can produce drinking water. But new sea ice is salty. Use only old sea ice, which is blue-grey in colour and with rounded contours. It can be melted or sucked for the ice loses its salt after a year or more. In summer, pools on old sea ice may be drinkable (if they are not wave splashes). Taste very carefully before drinking, for drinking any salt will aggravate thirst.

Water from fish

Drink the aqueous fluid found along the spine of large fish and in the eyes. Carefully cut the fish in half to obtain it and suck the eye. If you are so short of water that you need to do this then do NOT drink any of the other body fluids for they are rich in protein and fat and will use up more of your reserve water in digestion than they supply.

Treatment of sea water

Liferaft equipment may include solar stills and chemical desalination kits. They carry their own instructions. Set solar stills out immediately, but use the desalination tablets only when the weather is unfavourable for the stills and dew or rain catchment is ineffective.

REMEMBER:

DO NOT drink sea water

DO NOT drink urine

DO NOT drink alcohol

DO NOT smoke

DO NOT eat, unless water is available

Sleep and rest are the best way of enduring periods of reduced water and food – but make sure that you have adequate shade when napping during the day.

If the sea is rough, tie yourself to the raft, close any cover and ride out the storm as best you can. RELAX is the keyword – at least TRY to relax.

FOOD

Conserve any emergency food supplies until really needed, even then only taking a small nibble, and try to live off natural foods.

Fish will be the main food source. There are some poisonous and dangerous ocean fish but in general, in the open sea, out of sight of land, fish are safe to eat. Nearer the shore there are fish that are both dangerous and poisonous to eat, including some, such as the Red Snapper and Barracuda, which are normally edible but poisonous when taken from the waters of atolls and reefs.

Flying fish will even jump into your boat!

Fishing

- Do NOT handle fishing line with bare hands and never wrap it around the hands or tie it to an inflatable dinghy. The salt which adheres to it can make it a sharp cutting edge – a danger both to the raft and to your hands.

- Wear gloves if they are available or use a cloth to handle fish to avoid injury from sharp fins and gill covers.

- Fish and turtles are attracted to the shelter from the sun provided by a dinghy or raft and will swim under it. If you have a net pass it under the keel from one end to the other (you need two people to hold the ends).

- Use a torch to attract fish at night – or on a moonlit night lower a piece of cloth, tinfoil or metal into the water to reflect the moon and it may also draw fish to it.

- Improvise hooks from whatever is available. Small folding pocket-knives, pieces of jagged metal, wire. Small bright metal objects may serve as 'bait' – including buckles, spoons and coins.

- If using a metal spoon or spinner keep it moving by paying out and reeling in. Let the 'bait' sink and then retrieve it.

- Use offal from caught fish for bait.

- Fish flesh spoils easily and in the tropics must be eaten fresh unless the air is dry – which is unlikely in the tropical oceans.

- In cooler zones excess fish can be dried in the sun for future meals. Clean and gut before drying.

Birds

All birds at sea are also potential food. They will be attracted to a raft as a potential perching place. Keep still until they settle and you may be able to grab them, especially if they are exhausted by flying in bad weather.

You may also catch birds using lines trailed in the water with hooks or gorges baited with fish.

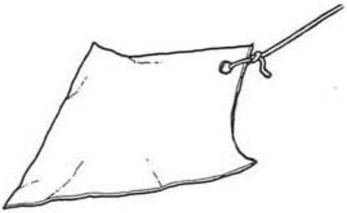

A diamond-shaped tin gorge, wrapped with fish, and trailed behind the craft will attract birds. When a bird siezes the 'fish', the gorge should lodge across its gullet.

Seaweed

Seaweed not only occurs on shorelines but, far out in some oceans, there are floating forms, especially the *Sargassum* species of the Sargasso Sea and the North Atlantic Drift, which is found in many warm waters, and others which grow in the colder waters of the southern Atlantic and Pacific. Since raw seaweeds are tough and salty they are difficult to digest raw. They absorb fluids so should not be eaten when water is short.

Seaweeds may also provide food in the form of the small crabs and shrimps living on them and small fish that you may shake out from among them if you haul some weed aboard. These small decapods are mottled brown in colour, like the weed, so are not easily seen on it.

Plankton, strained from the water, can also be a useful food source, especially in the cold southern waters (see *Polar food* in *Climate and Terrain*).

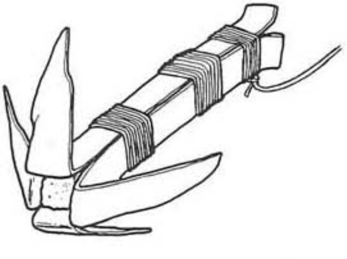

Make a grapple hook by lashing pieces of wood or metal wreckage together to form a multiple hook. Attach it to a line and trail it behind, or throw it out to rake in weed. You can use it for gathering other drifting wreckage to consolidate a makeshift raft.

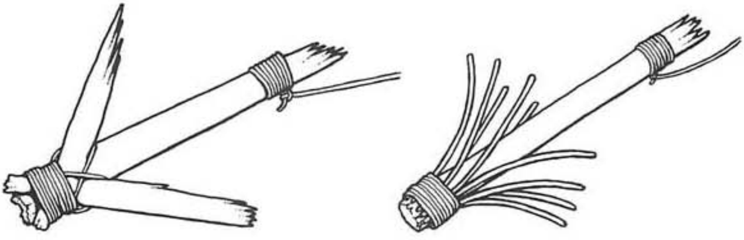

DANGEROUS FISH

Poisonous fish

Many reef fish have toxic flesh, some species at all times and in others only at certain times of year. The poisons are present in all parts of the fish, but especially in the liver, intestines and eggs.

Fish toxins are water soluble – no amount of cooking will neutralize them. They are tasteless – so the standard edibility tests (see *Food*) are useless. Birds are least susceptible to the poisons so do not think that because a bird can eat a fish, it is a safe species for you to eat. Cats appear less affected, though dogs and rats are as susceptible as humans.

The toxins will produce a numbness of the lips, tongue, toes and tips of the fingers, severe itching and an apparent reversal of temperature sensations. Cold things seem hot and hot things cold. There will probably also be nausea, vomiting, loss of speech, dizziness and a paralysis that eventually brings DEATH.

As well as those fish with poisonous flesh (see colour section) there are those which are dangerous to touch. Many kinds of ray have a poisonous barb in their tail; there are also species that can deliver an electric shock. Some reef fishes, such as Stonefishes and Toadfishes, have venomous spines which, although seldom fatal, can be VERY painful, causing a burning sensation or even agonizing pain out of all proportion to the apparent severity of the wound.

Jellyfish, which are sometimes barely noticeable in the water, can carry powerful stings. The Portuguese Man-of-War (not a true jellyfish) is a bluish bladder-like creature with a small fluted sail. Do not enter the water if you see these creatures. They trail very long 'streamers' which carry dangerous and painful toxins.

Aggressive fish

There are also a number of ferocious fish which should be avoided. The bold and inquisitive Barracuda has been known to attack man. It may charge lights or shiny objects at night. The Sea Bass, which can grow to 1·8m (5½ft), is another to keep clear of, and the Moray Eel, which has many sharp teeth and grows to 1·5m (5ft), can also be aggressive if disturbed. Sea snakes are venomous and sometimes occur in mid-ocean. They are unlikely to bite, but AVOID.

SHARKS

Only a handful of attacks by sharks are recorded each year, and only a minority are fatal. However, the survivor at sea is more vulnerable than the beach swimmer to whom statistics largely refer. Very few types of sharks are considered dangerous to man.

Six sharks account for most human casualties: the Great White, Mako, Tiger, Hammerhead, Bull and Grey Nurse. The Great White is the largest, but size is not an indication of danger and bears no relation to the likelihood of attack. A shark smaller than a man can still kill a swimmer. Basking Sharks and Whale Sharks may be 13.3m (44ft) long but they feed on tiny plankton and are not a problem.

Ocean sharks have the ability to kill but, in the tropics, their food is so abundant that they are not usually ferocious. These sharks are usually cowards and can be scared off by the jab of a stick, especially on the nose. However, making a commotion may ATTRACT sharks from a distance.

Sharks live and feed at considerable depths and for most of the time feed off the ocean bottom, but hungry sharks will follow fish to the surface and into shallow water. When it explores such water a shark is likely to be DANGEROUS.

A shark's usual diet includes fish, squid, crabs and a variety of other marine animals but it seeks food that is EASY to get and especially goes after stragglers from schools of fish or wounded prey. Sharks will follow a ship to scavenge refuse thrown overboard.

The shark feeds most actively at night and at dusk and dawn. Its small eyes have limited vision and it locates its prey by smell and vibrations in the water. It will be attracted by blood from wounds, body wastes and rubbish. Weak and fluttery movements will draw a shark's attention because they suggest a vulnerable, wounded creature. It will be repelled by strong, regular movements and loud noises.

Man's strange appearance is new to a shark and clothing produces a confusing shape. A group of clothed humans bunched together will be safer than a single individual. If a shark keeps its distance, it is only curious. If it circles inwards and begins sudden movement, the likelihood of attack is greatly increased.

Sharks cannot stop suddenly or turn quickly. A good swimmer can evade a single large shark by making rapid changes of direction which the shark cannot match.

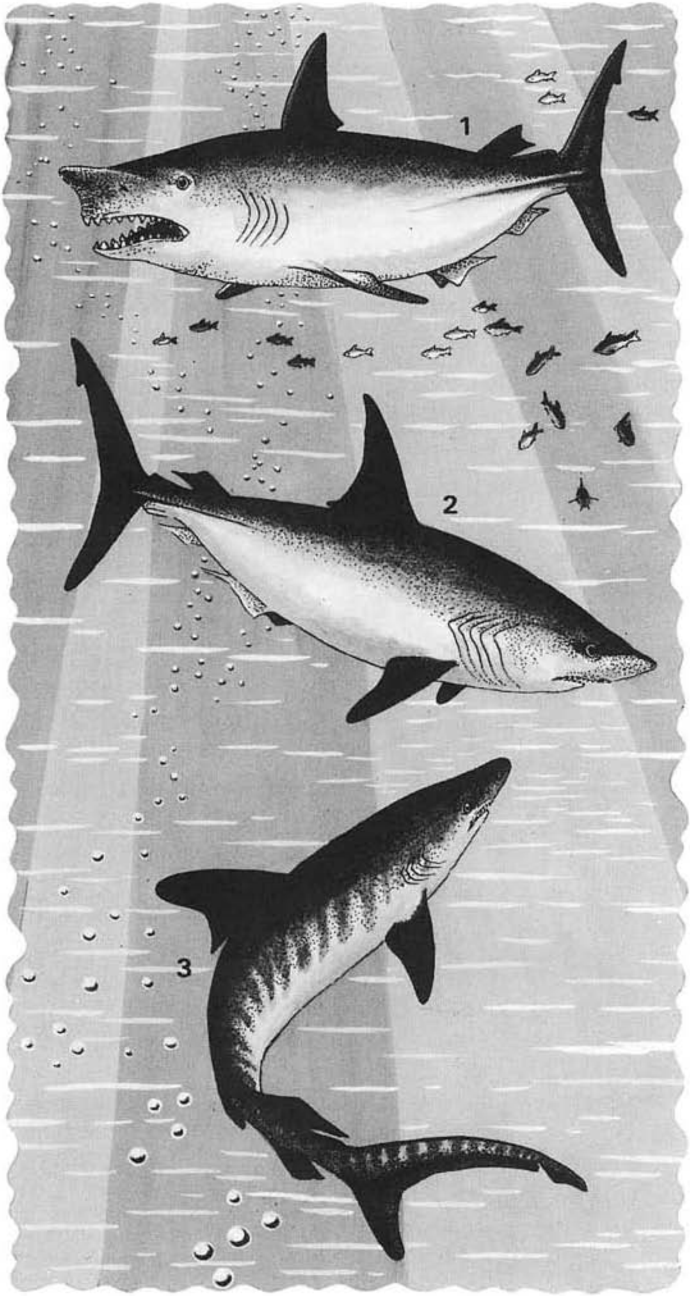
1
2
3

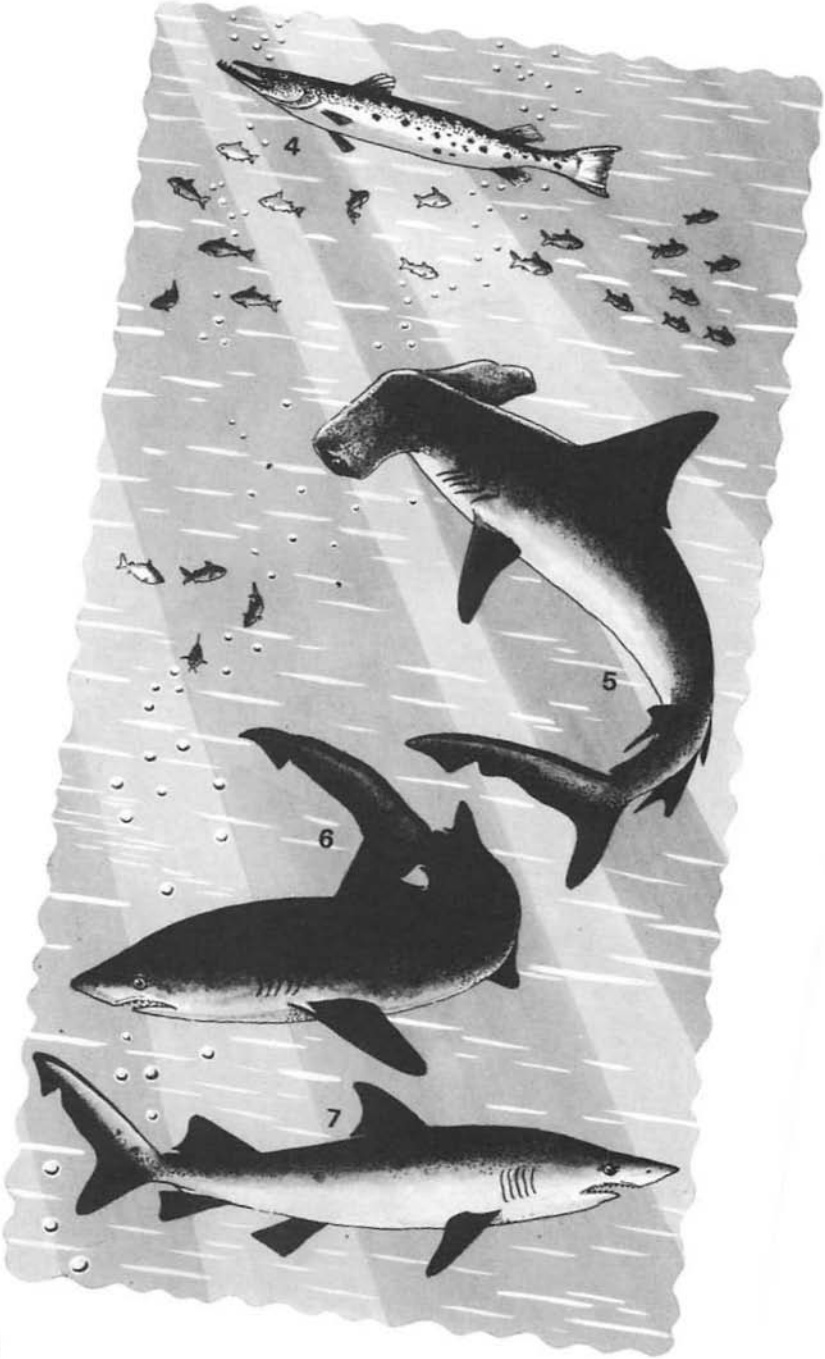
4
5
6
7

FALSE ALARM

Not every fin showing above the surface is attached to a shark! The wing tips of large rays may break the surface and appear to belong to a pair of sharks moving, unnaturally, in perfect synchronization. The fins or flippers of whales may also appear – likewise those of porpoises and dolphins, which are harmless and will probably show themselves more completely.

The sharks shown here have been known to attack man.

1 Great White Shark *(Carcharodon carcharias) grows to 6m (18ft) but usually less, grey above, white below, very thick bodied, with pure black eyes and a stubby conical snout; in all oceans but mostly off southern Africa, east and west North America and southern Australia and New Zealand.*

2 Mako *(Isurus oxyrichus) averages 2–3m (6–10ft), heavy bodied, ultramarine blue above, creamy-white below; in all oceans but most abundant in warm temperate waters. A very fast swimmer, occasionally leaps from the water.*

3 Tiger Shark *(Galeocerdo cuvieri) averages 3–3.5m (10–11ft), heavy bodied, barred or blotched above when young, when mature more evenly greyish above, white below, with a very wide head and jaws and abruptly squared-off snout, in all tropical and subtropical waters, often close inshore.*

4 Barracudas *(Sphyraena) are not sharks but thin, torpedo-like fishes, bluish-barred above, brilliant silver below, with a protruding mouth packed with sharp teeth, some kinds growing to 2m (6ft), in all tropical waters. Very fast, darting, often in shoals, usually dangerous only when there is blood in the water.*

5 Hammerhead Sharks *(Sphyrna) are immediately recognizable by their distinctively flattened hammer-like head. There are several different kinds, the largest reaching 6m (18ft); in all tropical and subtropical waters.*

6 Bull Shark *(Carcharhinus leucas) is found in the tropical west Atlantic with close relatives off southern Africa and in the Indian Ocean. Stout, grey above and white below, up to 4m (13ft). Aggressive, and dangerous in its liking for shallow water and ability to ascend far up rivers.*

7 Nurse *(Ginglymostoma), such as the Grey Nurse of eastern Australian waters, reach over 4m (13ft) heavily built, large-finned, greyish above and white below. Often found very close inshore.*

If you catch a shark

If a small shark is accidentally hooked it can be hauled to the side of the raft or boat, the head pulled clear, and clubbed hard before hauling aboard. Make sure it is really stunned before approaching and finishing it off with more blows to provide shark steaks.

DON'T try this with a large shark. It could injure you and your craft. You must cut your line and sacrifice part of it, for the shark's threshing will soon attract its fellows.

Protection against sharks

Unless wearing a lifejacket or travelling in a craft equipped with shark repellent, the risk is great to anyone in the water, but it is not a foregone conclusion that shark attack will occur. Shark repellent may not be 100 per cent effective – but even so, only use when absolutely necessary. Remember – you can only use it once.

In the water:

If sharks are present try to avoid passing body wastes, which could attract the sharks' interest. If you must urinate do it in short, sharp spurts and allow it to dissipate between spurts. Collect faecal matter and throw it as far away from you as possible. If you vomit try to hold it in the mouth and reswallow it, but if this proves impossible throw it as far away as possible.

If it is necessary to swim use strong, regular strokes, avoiding schools of fish.

If a group of people are threatened they should bunch together and face outwards. To ward off attack kick outwards and punch out with a stiff arm using the heel of the hand – like a sports 'hand-off'.

Make loud noises by slapping the water with cupped hands. Put your head under the water and shout. These measures are more effective with a group but can work even when you are alone and under attack. If you have a knife, be prepared to use it. Let the shark take it fully in the snout, or go for the gills and eyes.

On a raft or boat:

Don't fish when sharks are around and don't throw waste overboard (including excrement and fish offal). Let go of baited hooks. Do not trail arms or legs in the water. If a shark threatens to attack discourage it with jabs to the snout with a paddle or pole. Remember – a large shark could also take a bite out of a boat or raft.

BE AWARE

If you have shark repellent, follow manufacturer's instructions – but use only if the situation is very grave. The repellent will soon dissipate in the water and become ineffective. Choose your moment well, since you can only use it once.

MAKING A LANDFALL

When you approach land try to select a landing point where it will be easy to beach or where you can safely swim ashore. Take down the sail and keep watch for rocks. The sea anchor will keep you pointing at the shore and will slow down your progress, giving you more time to steer away if you are heading for rocks. Try not to land with the sun in your eyes, which will make it more difficult for you to see rocks and difficulties.

A sloping beach with a small surf is the ideal place to choose – if you have any choice. If you can time it right, ride the back of a breaker. To avoid being swamped or turned sideways by an oncoming crest of a wave paddle hard, but do not overshoot a breaker which is carrying you along. In very heavy surf turn the vessel to face seaward and, as a wave approaches, paddle into it.

As you approach note the lie of the land: the location of high ground, types of vegetation, possible watercourses. You will see features which may be invisible from the shore. If with companions, choose a rendezvous spot to meet at if the boat breaks up and you are separated.

If you reach land at night wait until morning to beach if you can, there are too many dangers you can fail to spot in the dark.

If you float into an estuary make every effort to reach a bank. The turning tide could carry you back out to sea. Take in the sea anchor and, to gain ground, make the boat as light as possible. Bail out an inflatable and inflate it to the maximum. This enables you to make the most of the incoming tide.

If you are being swept back out to sea by the ebb, ballast the dinghy by part filling it with water and stream the sea anchor.

SURVIVAL TIP

Keep yourself tied to your raft. Even if it is overturned or damaged and you are rendered unconscious, you stand a chance of surviving. Alone in the water and dashed on the rocks – you are DEAD.

Swimming ashore

If you have to swim ashore on to rocks in a heavy sea keep on clothing, shoes and lifejacket if you have it. Raise your legs in front of you to take the shock of impact with rocks on the soles of your feet; absorb it in bending the knees.

10

RESCUE

Signalling is necessary to attract rescue and codes may be necessary for communication once a contact has been established. It will help the placing of signals, and increase their effectiveness, if the patterns used for search patrols are understood.

If survivors are being air-lifted out it may be necessary to prepare a landing place for aircraft or helicopters – so suitable locations are described and the precautions you should take in the presence of a helicopter.

SIGNALLING 506

SIGNALS AND CODES 508

- Ground-to-air signals 511
- Message signalling 513
- Mountain rescue code 517
- Information signals 519

SEARCH 520

- Search patterns 520
- Helicopter rescue 523
- Survival case study 527

SIGNALLING

The first requirement for rescue is to let others know of your situation and, if possible, your location. Once you are in contact you can pass on other information.

The obvious technique is to use a mobile or satellite phone, which should be used sparingly, but if you don't have the luxury of such systems then there are a number of internationally recognized distress signals. The letters SOS (for Save Our Souls) is probably the best known. It can be written, transmitted by radio, spelt out by semaphore or sent in Morse code by any method.

The signal 'Mayday' (a phonetic rendering of the French *m'aidez* – help me) is the one used in most radiotelecommunications by ships and planes.

Vehicle or aircraft wreckage

If you are with a stranded vehicle or downed aircraft it may provide many useful signalling aids. If there has been no fire there will be supplies of fuel, oil and hydraulic fluid which can be burned. Tyres and electrical insulation on a fire will generate black smoke.

Glass and chrome make great reflectors, especially engine cowlings and hubcaps. Lifejackets, dinghies, and parachutes are all brightly coloured and eye-catching. Arrange these colourful and shiny objects around your location where they will be most visible and attract attention.

Switch lights on at night – or if batteries are running low keep them in reserve to flash headlamps, sound the horn and otherwise attract attention, when passing aircraft or signs of possible searchers are observed.

Fire and smoke

Fire – both flames and smoke – is an excellent way of attracting attention. Establishing signal fires is one of the primary tasks once the immediate needs for treatment of injury and provision of shelter from harsh elements have been met. In a large group some people should set about gathering fuel for a campfire and for signal fires as soon as possible.

Where to site signals

When siting signals take full account of the terrain. Choose high points for light signals. If you are on a ridge, erecting an unusual silhouette may attract attention. If you are laying out marks on the ground, use level ground or ensure that they are on slopes that are not likely to be overlooked in the usual pattern of aerial search.

SURVIVAL TIP

It is usual for planes to fly over hilly territory from the lower to the higher ridges. This creates the problem that the slopes behind the ridges may be hidden as the plane approaches. If in doubt, signals nearer the tops of ridges should be seen from whichever direction the rescue aircraft is travelling.

International codes

When contact has been established more complex international codes (shown later) will enable you to signal your basic needs, if verbal communication is not possible.

With air- or sea-rescue it may then be necessary to prepare a landing strip or to help rig lines or apparatus, and some knowledge of basic procedures will greatly facilitate the operation. (See *Helicopter rescue.*)

Transmitters

Dinghies, liferafts and even personal lifejackets are sometimes equipped with transmitters which send out bleeps indicating position, though these are not usually effective over a very long range. Many emergency radio transmitters are also very limited in range and to avoid wasting precious batteries should be held in reserve until there is some chance of their signals being picked up. With effective radio apparatus, however, distress signals should be sent out immediately and transmitted at regular intervals.

Check instructions on all transmitting apparatus. Ship and plane transmitters can operate on many wavelengths, but some emergency equipment is set to fixed distress channels.

Generally speaking, the portable VHF transceivers used by mountaineering teams can communicate only with stations in a direct line of sight and without any intervening obstruction (though sometimes a permanent relay station may be established on a strategic high point). Such sets are usually tuned to a mountain rescue frequency but procedures should be established before departure.

If you have a working transmitter, check the battery situation. Can the vehicle engine still be used to generate electricity or recharge the batteries? Conserve fuel for this purpose and plan your transmissions to a pattern rather than attempting long continuous periods on the air. If anyone picks up your signal they can then work out that they can expect it again.

Noise

Noise is also an excellent way of attracting attention if you know that people are within earshot. The International Mountain Distress Signal, apart from signalling SOS, is six whistles a minute (or six waves, light flashes, etc.) followed by a minute's silence, then repeated. A shout may be enough if you are trapped, or near help but too injured to reach it.

Be imaginative

Do not reject even such ideas as the message in a bottle. This particular method has a low chance of success if you are shipwrecked in the middle of the Pacific but on a river a more noticeable floating object which carries a clear message might well attract attention – a small raft with a bright sail labelled SOS for instance. Use your imagination to think of ways that will attract attention to you and your plight, without using up valuable energy and resources.

Moving on

If you decide that rescue is unlikely, and that your best plan is to make your own way back, you should leave clear signs behind so that if searchers do track down the disaster spot they have an indication of the route that you have taken. On your way you may have more success in attracting attention if closer to regular flight routes or in more open territory.

SIGNALS AND CODES

Fire signals

Three fires is an internationally recognized distress signal. Ideally they should be placed in a triangle at equal distances apart, an arrangement which also makes them easier to feed with fuel, but if that is not possible any grouping serves, provided that the fires are clearly separated. However, if fuel is scarce, or if you are too badly injured or too weak from

BE AWARE

Almost any signal repeated six times will serve as a distress signal. Depending on your location, this could be six fires, six columns of smoke, six loud whistles, six gunshots – even six flashes of light. If using noises or lights, wait one minute between each group of six signals.

hunger to maintain several fires, use only your campfire.

You can't keep signal fires going all the time but they should be prepared, covered to keep them dry, and maintained, ready to be lit to attract the attention of any passing aircraft. Build them with plenty of easily ignited tinder so that they will get going rapidly when lit. Birch bark makes an ideal tinder. Other tinder materials are described under *Fire* in *Camp Craft*.

Petrol can be used as a firelighter but DON'T just pour it on the fire. Use a piece of rag as a wick, soak it in petrol and lay it to the tinder. Don't light it straight away. Carry the fuel can off to a safe distance and wait a few seconds before lighting the wick. If a fire does not light first time pull the tinder apart, checking that there are no sparks or embers still burning, before adding extra petrol.

BE AWARE

- Keep a stock of green boughs or supplies of oil or rubber close by to create smoke if needed.
- Among vegetation or close to trees, build an earth wall around each fire to contain it.
- There is no point building fires among trees where they cannot be seen, the canopy will block out the signal. Place them in a clearing.

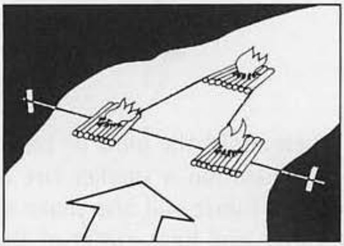

If by a lake or river, build rafts to place your fires on and anchor or tether them securely in position. Arrow indicates direction of current.

Torch trees

Small isolated trees make excellent fire signals. Build a fire between the boughs. Use plenty of dry twigs – old birds' nests make good fire starters. This fire will ignite the foliage and produce plenty of smoke. If a tree is dead start a fire at its base. It will burn for a long time leaving you free to attend to other signals.

NOTE: Do not risk starting a forest fire. Apart from the damage this will cause, your life will be in greater jeopardy.

Luminous cone fires

On a clear and open site make a tripod with a platform to support a fire. The platform keeps the tinder off damp ground, or you can store more firewood beneath it. Use a covering of evergreen boughs to keep the cone dry; they will burn brightly and give off a good smoke.

Cover the complete cone fire with bright coloured material, if available – a parachute would be ideal. This will not only keep the fire dry and ready to burn, but will itself be noticeable during the day. Whip it off when you ignite the fire – you may not attract attention the first time.

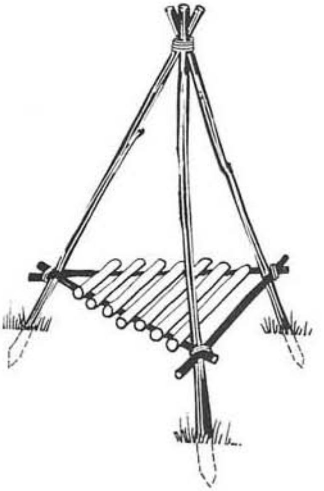

Keep these tripods well maintained, ensuring that wood is dry enough to light at a moment's notice and that the supply is not poached for other uses. Drive the pole ends into the ground to prevent tipping over in strong winds.

When alight the glow of this cone can be seen for miles. In an exposed location a smaller fire inside a conical tent or tepee of parachute fabric will also make a noticeable beacon. Ensure there is a smoke and heat outlet at the top of the tepee and keep the fire under control. If it is on a slope add fuel from the side or above the fire so that you do not excessively mask the firelight – though some flickering of movement before it may help to attract attention.

Use wreckage to help fire-signalling

Stand a fire on a piece of metal from an aircraft or vehicle. It will keep kindling from damp ground, when hot will increase convection and make the fire burn brightly and, if polished, will act as a reflector intensifying the brightness. Three such fires become an immediately recognizable distress signal.

Smoke indicators

During daylight smoke is a good locator, so have plenty of smoke-producing material ready to put on your fires. Choose this material to give off a smoke that shows up well against the surroundings.

Light smoke will stand out against dark earth or dark green forest. Produce it with green grass, leaves, moss and ferns. Any wet material will produce a good smudge fire and damp mats and seat covers will smoulder for a long time. This will also keep flying insects at bay.

Dark smoke will show best against snow or desert sand. Use rubber or oil to produce it. If atmospheric conditions make the smoke hang in layers along the ground then build up the fire to increase its heat. Thermal currents will then take the smoke to a good height.

BE AWARE

Smoke not only helps the pilot of a rescue aircraft find you, it also shows the surface wind direction. Make sure that smoke is downwind of the landing site and of any panel codes that you have laid so that it does not obscure them from above.

GROUND-TO-AIR SIGNALS

These letters are internationally recognized emergency signals. FILL is a useful mnemonic for remembering the main ones. The single bar: I is the most important and the easiest to make. A pilot will risk a great deal to answer such an emergency. Make them as large and as noticeable as possible using colour contrast or shadow. A recommended size is 10m long and 3m wide (34ft and 10ft) for each symbol, with 3m (10ft) between symbols.

Lay or make these panel codes out in the open, avoid steep gullies or ravines and do not make them on reverse slopes. Use the marker panels from your survival pouch (see *Essentials*), or if you do not have these – improvise. Lay out pieces of wreckage or dig out the signs as a shallow trench, banking up the earth so that it increases the depth of the shadow. Use rocks or boughs to accentuate it.

On snow, even tramping out the symbols will show clearly until the next snowfall.

Once contact has been made, a message dropped or signalled by the aircraft can be answered with A or Y (affirmative) and N (negative) signals, or Morse code or body signals.

GROUND-TO-AIR CODE

I	Serious injury – immediate casevac (casualty evacuation) – (can also mean NEED DOCTOR)
II	Need medical supplies
F	Need food and water
N	Negative (No)
A	Affirmative (Yes) – (Y will also be understood)
LL	All is well
X	Unable to move on
→	Am moving on this way
K	Indicate direction to proceed
JL	Do not understand
□	Need compass and map
△	Think safe to land here (Broken at angles, means ATTEMPTING TAKE-OFF)
¦	Need radio/signal lamp/battery
L┐	Aircraft badly damaged

Night signals

These signals will attract attention during daylight, even if you are asleep or injured. If you have a supply of petrol or other inflammable substances, you can make signals which will work at night. Dig or scrape an SOS (or any symbol) in the earth, sand or snow and, when the signal is needed, pour petrol into it and ignite it.

NOTE: You MUST destroy these signals if rescued. They will go on working long after you have gone.

MESSAGE SIGNALLING

You do not need to learn a complicated system of semaphore. The international Morse code can be transmitted by flashing lights on and off, by a simple heliograph, by waving a flag or a shirt tied to a stick or using sound.

NOTE: Don't rely upon your memory – carry a copy of the code on your person. Even if you are a regular user and know it backwards others may need it who do not.

There is a procedure to follow when sending and receiving messages. Learn the special codes to make the operation easier.

Heliograph

Use the sun and a reflector to flash light signals. Any shiny object will serve – polish a tin lid, glasses, a piece of foil – though a hand-mirror is best. Sustained flashes are dashes and quick ones dots. If you do not know Morse code, even random flashes should attract attention. At least learn the code for SOS.

A flash can be seen at a great distance and even when you do not have a specific contact to aim for may attract someone's attention. It's worth trying, since it requires little energy. Sweep the horizon during the day. If a plane approaches closely make intermittent flashes or you may dazzle the pilot. Once you are certain you have been seen, STOP signalling.

SINGLE-SIDED REFLECTOR

With an improvised reflector pick up the sunlight to get an image on the ground or some other surface and lead it in the direction of the aircraft or other potential contact.

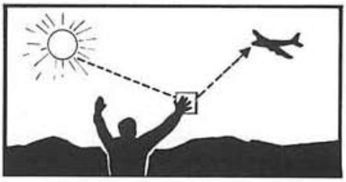

USING HELIOGRAPH

If you have a double-sided reflector and can punch a hole in it you will have something close to a standard issue heliograph.

Sight the person, plane, ship, etc, that you wish to contact through the hole in the heliograph (a) in the general direction of the sun, so that the sun will shine through the hole (b). You will see a spot of light on your face (c).

Angle the mirror so that the dot of light on your face 'disappears' back through the hole in the mirror – still sighting your contact.

If the sun is at such an angle that this manoeuvre does not work bring the mirror close to your eyes and a hand lined up between you and the contact. Angle the mirror to flash on to your hand, then move the hand away.

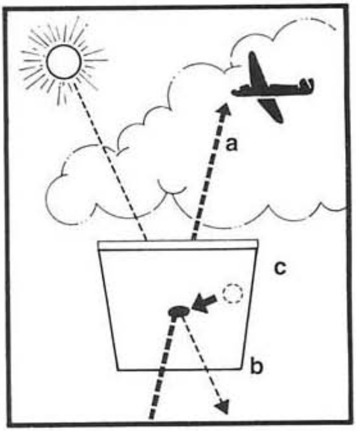

NOTE: Practice this form of signalling, but unless you are in a survival situation, do NOT signal to aircraft or 'transmit' messages which could cause alarm or danger to others.

Rag signals

Tie a flag or a piece of brightly coloured clothing to a pole and move it left for dashes and right for dots. Exaggerate each movement with a figure-of-eight movement.

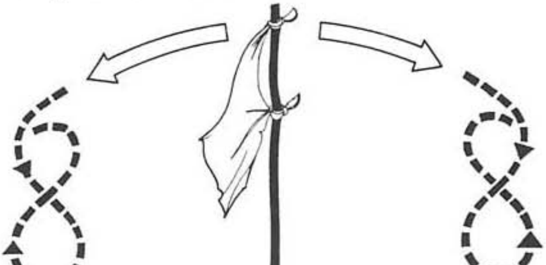

For a 'dot' swing to the right and make a figure-of-eight.

For a 'dash' swing to the left and make a figure-of-eight.

This simple signing may work without figure-of-eight movements at closer range. Keep 'dash' pauses on the left, slightly longer than 'dot' movements to the right.

INFO.

MORSE CODE

A	.-	N	-.	1	.----
B	-...	O	---	2	..---
C	-.-.	P	.--.	3	...--
D	-..	Q	--.-	4	....-
E	.	R	.-.	5	.....
F	..-.	S	...	6	-....
G	--.	T	-	7	--...
H	....	U	..-	8	---..
I	..	V	...-	9	----.
J	.---	W	.--	0	-----
K	-.-	X	-..-		
L	.-..	Y	-.--		
M	--	Z	--..		

SENDING SIGNALS

AAAAA* etc. – Call sign. *I have a message*
AAA* – End of sentence. *More follows*
Pause – End of word. *More follows*
EEEEE* etc. – Error. *Start from last correct word*
AR – End of message

RECEIVING SIGNALS

TTTTT* etc. I am receiving you
K – I am ready. *Start message*
T – Word received
IMI* – Repeat sign. *I do not understand*
R – Message received

** Send as one word. No pauses*

USEFUL WORDS

SOS	...---...
SEND	...|.|-.|-..
DOCTOR	-..|---|-.-.|-|---|.-.
HELP	....|.|.-..|.--.
INJURY	..|-.|.---|..-|.-.|-.--
TRAPPED	-|.-.|.-|.--.|.--.|.|-..
LOST	.-..|---|...|-
WATER	.--|.-|-|.|.-.

BODY SIGNALS

This series of signals will be understood by airmen and can be used to signal to them. Note the changes from frontal to sideways positions and the use of leg and body posture as well as hand movements. Use a cloth in the hand to emphasize the YES and NO signals. Make all signals in a clear and exaggerated manner.

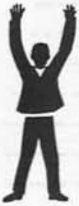

Pick us up

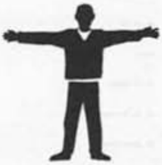

Need mechanical help

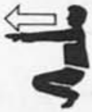

Land here

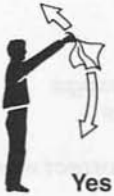

Yes

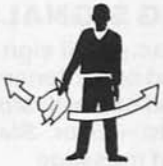

No

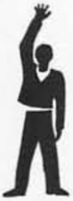

All is well

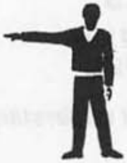

Can proceed shortly

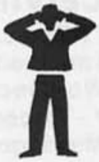

Have radio

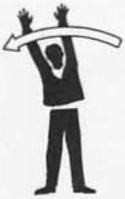

Do NOT attempt to land here

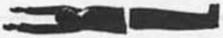

Need medical assistance

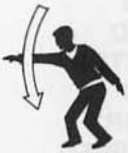

Use drop message

Response to body signals

To acknowledge messages received from the ground, the pilot of an aircraft will perform one of these manoeuvres:

Message received and understood:
  In daylight – flying the plane and tipping the wings in a rocking movement side to side
  At night – flashing green lights

Message received but NOT understood:
  In daylight – flying the plane in a right-handed circle
  At night – flashing red lights

MOUNTAIN RESCUE CODE

These sound, light and pyrotechnic codes are recognized internationally by mountain rescue services:

Message: SOS
  Flare signal – Red
  Sound signal – 3 short blasts, 3 long, 3 short
    Repeat after 1 minute interval
  Light signal – 3 short flashes, 3 long, 3 short
    Repeat after 1 minute interval

Message: HELP NEEDED
  Flare signal – Red
  Sound signal – 6 blasts in quick succession
    Repeat after 1 minute interval
  Light signal – 6 flashes in quick succession
    Repeat after 1 minute interval

Message: MESSAGE UNDERSTOOD
  Flare signal – White
  Sound signal – 3 blasts in quick succession
    Repeat after 1 minute interval
  Light signal – 3 flashes in quick succession
    Repeat after 1 minute interval

Message: RETURN TO BASE
  Flare signal – Green
  Sound signal – Prolonged succession of blasts
  Light signal – Prolonged succession of flashes

Flares

Any flare will be investigated during a search, regardless of colour, but choose one best fitted to the location.

- In closely wooded country green does not stand out but red does.
- Over snow white merges – green and red are best.

Familiarize yourself with the types of flare. Make sure you understand the instructions, for some flares eject a white-hot ball of magnesium that will burn a hole in anything it hits – your chest or dinghy – if they are misdirected.

Types of flare

Some flares are hand-held and reversible. One end produces smoke for daytime use, the other a flare for use at night. The higher these are held the easier they are to see. Flares and rockets which are fired into the air will be visible for a greater distance. One type reaches a height of 90m (300ft) where a parachute opens holding the flare suspended for several minutes. Other rockets produce a loud bang and colour balls.

Keep flares dry and away from naked flames and heat sources. Ensure that safety pins are in position and will not accidentally drop out, but also check that they are not bent over in such a way that they could not be quickly removed when necessary.

Handling flares

Hand-held flares are cylindrical tubes with a cap at each end. The top cap is often embossed with a letter or pattern so that it can be identified by touch in the dark. Remove it first. Then remove the base cap to expose a short string and a safety pin, or other safety device. Point the flare upwards and away from you and anyone else in case you accidentally fire it. Remove the pin, or turn to the fire position. Hold the flare at arm's length, at shoulder height, pointing directly upwards. Sharply pull the firing string vertically downwards. Brace yourself as you do so for there will be a kickback. Some flares and maroons have a spring mechanism trigger like that of a mousetrap.

Very pistols fire various cartridges. To fire them load the pistol, point it skywards, cock the hammer and then squeeze the trigger.

Mini-flares are more usual equipment today, lighter than Very pistols but as effective (see *Survival*

DANGER

Hand-held flares get hot. When they burn down do not drop them into the bottom of a boat, where they could start a fire, or burn straight through an inflatable.

kit in *Essentials*). They require handling with the same care. To use, screw a flare of the selected colour into the end of the discharger, aim skyward, pull back striker – FIRE!

INFORMATION SIGNALS

These are signals to leave behind if you leave the scene of the crash or abandon camp.

Make a large arrow shape to indicate the direction in which you have set off which will be visible from the air and other direction markers which can be interpreted at ground level.

Signs on the ground will draw attention to your presence or past presence and the direction markers will help rescuers to follow your trail. Continue to make them, not only for people to follow but to establish your own route if you wish to retrace it and as a guide if you lose your sense of direction and start going back on your trail.

At camp leave written messages in containers to detail your plans. Hang them from tripods or trees and draw attention to them with markers.

Direction markers could include rocks or debris placed in arrow shape (a), stick left in crooked support, top in direction followed (b), grasses tied in an overhand knot with end hanging in direction followed (c), forked branches laid with fork pointing in direction followed (d), small rock set upon larger rocks, with small rock beside (e), indicating a turn or arrow- or arrowhead-shape notches cut in tree trunks (f).

A cross of sticks or stones (g), means 'Not this way.'

Signal danger or emergency with three rocks, sticks or clumps of grass, prominently displayed (h).

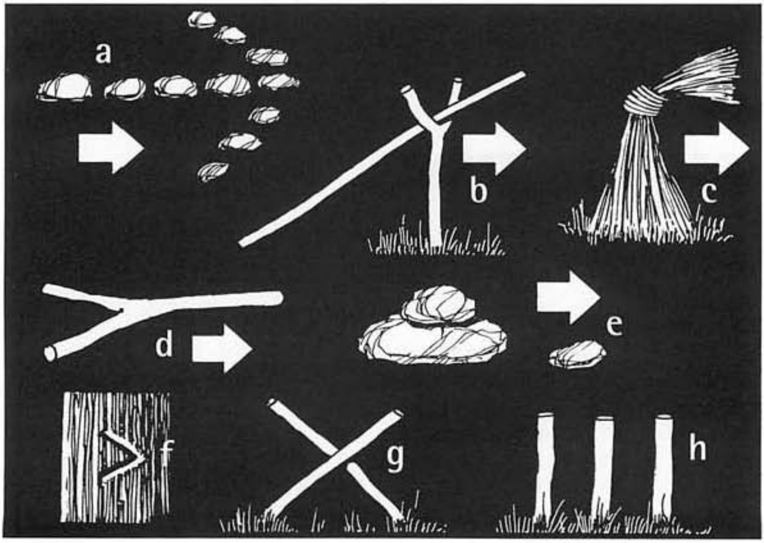

SEARCH

Awareness of search procedures will show how important it is for any expedition or trip to register its route plans and for survivors to stay as close as possible to that course, to set clear signals to draw attention to their location and to mark any camp they have abandoned (leaving information about their subsequent plans).

A search will start from the last known location and sweep on the proposed route. An assessment will be made of probable strategy adopted, given the terrain and the weather conditions. In mountain areas, for instance, it is likely that strong winds would make the survival party use the lee sides of ridges and descend from high ground. If no trace is found of them on the expected route these are the areas in which the search will be concentrated.

The effect of contour will be considered: by studying the ground the search party may assume that survivors were forced off route by the lay of the land. To make the searchers' job easier, make a cairn of rocks or other noticeable construction on prominent ground where it cannot fail to be noticed and leave a message there in a waterproof bag or in the pocket of a colourful piece of clothing. Give information on your intentions and on the state of your party.

If your route has been checked and the obvious refuge places in the locality searched, the rescuers will extend the search to cover the whole area of your disappearance. Ideally this will be done from the air but severe weather which creates bad flying conditions and poor visibility may mean it has to be done on foot, even if planes are available.

The number of searchers and the type of terrain will dictate the search pattern best suited to the situation.

SEARCH PATTERNS

The first search will be made along the route you were supposed to have taken.

BASE LINE

Base line, or box search, is carried out when there has been a high wind or bad weather conditions on your known route (a). Searchers should deduce that you may have veered from the route to the leeside of a slope for shelter.

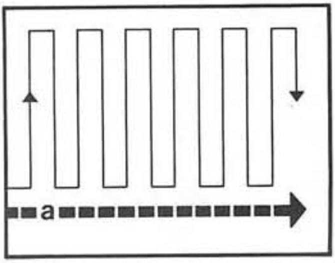

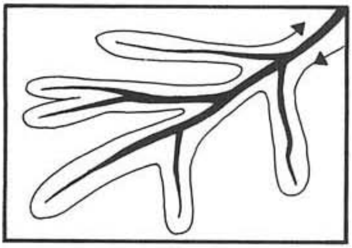

WATERCOURSE

Watercourse search takes in all the tributaries, using the main stream as base-line. This is undertaken when your last known position was on or near a river.

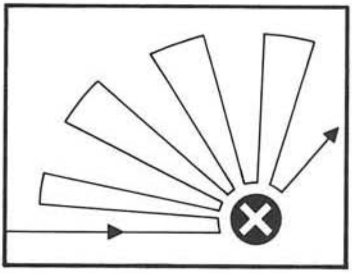

FAN

Fan search is used when your last known position (x) is fairly certain but it is impossible to deduce the direction you may have taken.

Aerial search

Search patterns from the air cover both sides of the intended flight path of missing aircraft or your known route.

If weather conditions are favourable a night search can be made, for lights will be clearly visible and the search can be made from a greater height so that a wider area can be covered in each sweep. If this does not produce results the area will still be rechecked by daylight.

If you are signalling to an aircraft and it turns away – keep watching. It may be following one of those recognized search patterns and you will be able to anticipate critical moments for signalling.

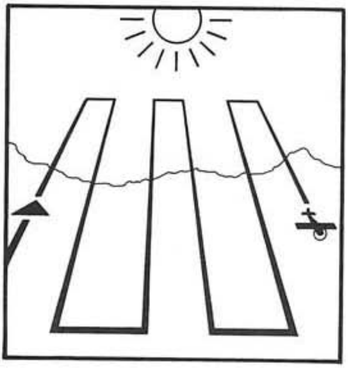

CREEPING LINE

Creeping line search, beginning in a corner of the search area, is particularly useful when only a single aircraft is available. It follows parallels which should be towards and away from the sun for a land search so that any reflection from a missing aircraft or other wreckage and signals will be more easily seen.

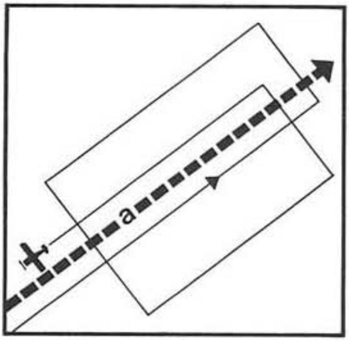

TRACK CRAWL

The primary search pattern, or track crawl search, parallels both sides of a missing aircraft's expected flight path (a), or travelling known land route. After flying for one hour, turn around and fly the reverse pattern.

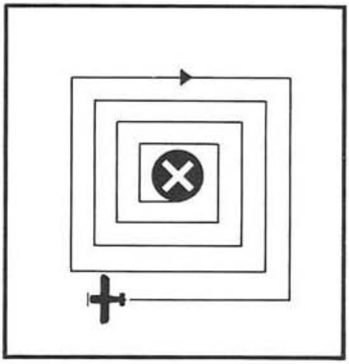

SQUARE SEARCH

Square search is useful when a comparatively small area is to be covered. Search starts in the last known location (x) and works outwards. If unsuccessful fly over last known location and start search in other direction. This ensures both sides are covered.

CONTOUR SEARCH

Contour search allows mountains and valleys to be searched with maximum thoroughness. Steep valleys should be searched several times by flying along them.

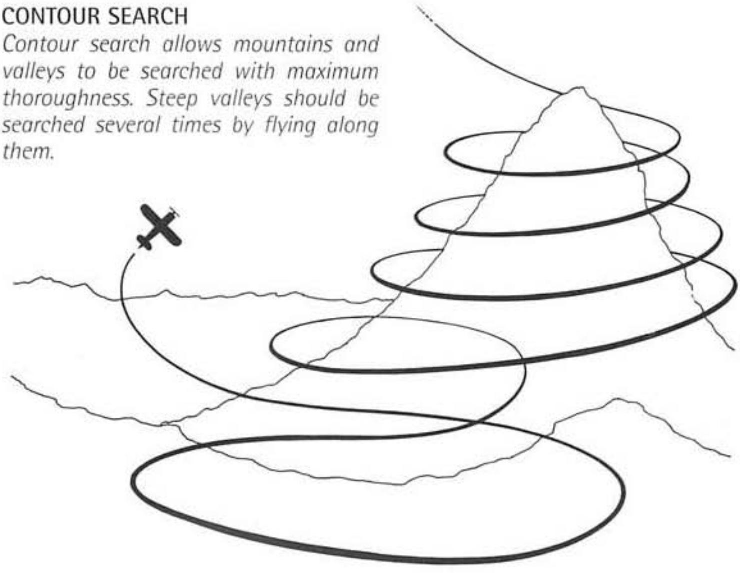

Combined searches

At sea a combined sea-and-air search is desirable. If the aircraft locates survivors the ships can pick them up. The ship can also act as a datum point for all aircraft. Rescue aircraft are equipped with supplies to drop to survivors to help them as they await pick-up by a surface vessel.

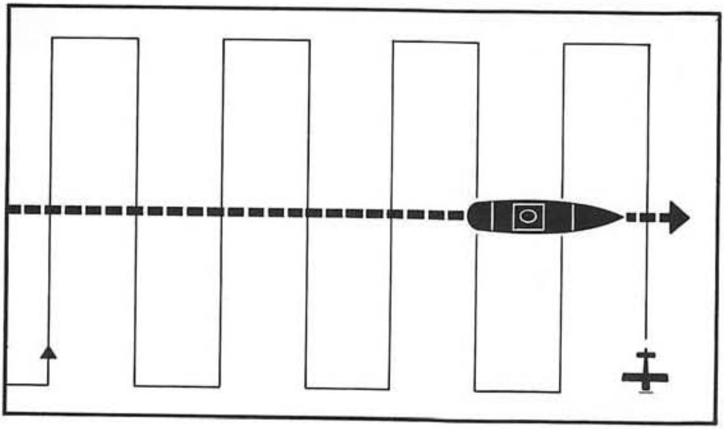

HELICOPTER RESCUE

While aeroplanes are used for the search, helicopters are used in most countries to carry out the actual rescue, especially on land. Where possible the 'copter will make a landing to take on survivors and fly them out. It may be possible for the pilot to find a convenient landing site nearby to which survivors can make their way, but it will be easier for the survivors to check out suitability at ground level and to create a site if necessary.

The helicopter will require an obstruction-free approach and exit path, both INTO the prevailing winds. The ground should be level (the slope should not exceed seven degrees – a gradient of 1 in 10). The touch-down surface should be firm and free of loose materials – remove leaves, twigs, everything. There must be no holes, tree stumps or rocks which could cause damage to the aircraft.

Selecting a landing site

Look for a natural clearing. In close country a river bank on a large bend is often the best natural landing place (LP).

Alternatively climb a spur and select a level piece of ground free of large trees. If you need to, cut down trees to clear more space. The trees will fall down the spur without blocking the area and a clear approach and exit path can be obtained across the spur. Do not attempt cutting a landing place on flat ground; it will take forever to create a clear approach and exit.

LANDING PLACE

PREPARING SITE

A level cleared area is needed at least 26m (85ft) in diameter. A further 5m (17ft) is needed all around, cleared to a height of 60cm (2ft). There should be a clear approach path into the prevailing wind with no obstructions within an angle of 15 degrees of the central landing pad.

Mark the touch-down point with an H. You can make it from inlaid rocks (keeping the surface smooth), clothing securely anchored, or panel markers. On snow, trample it down firmly to prevent it swirling and, in dry areas, water the surface to keep the dust down.

READY-MADE PAD

If a spur, mound or area of raised ground is nearby, this will be easier to clear and will provide an easy approach and exit across the spur if the wind direction is satisfactory.

MOUNTAINOUS TERRAIN

The payload of the helicopter is cut drastically with increased altitude so, if possible, make the landing site below 1830m (6000ft).

In mountains updraughts and downdraughts of air can be considerable according to the pattern of the land and its relationship to prevailing winds. Select a site that gives maximum lift in the direction in which the helicopter will take off.

Soft wet snow will cling to the helicopter and hamper its take off. Try to compact the landing surface as much as possible. Powdery snow will swirl under the rotors' downwash and restrict the pilot's vision. Stamp this down, too.

Non-landing rescue

In an emergency, helicopters will take considerable risks to rescue survivors. They may hover with one skid on a rock to make embarkation possible – but whenever it can be done create a proper landing place and minimize risks.

Most helicopters are equipped with a winch. If a landing place is out of the question you can be lifted from the ground while the helicopter hovers. All that is needed is an opening or clearing through which to extricate you.

Wind indication

It is important to indicate the direction and strength of the wind at the landing place so that the pilot can select the best approach, and keep the aircraft stable while carrying out the rescue. Smoke is an ideal indicator, but do not place it so that it obscures the touchdown area.

If a fire is not practical make a T sign from contrasting material and place it at the downwind edge of the landing place with the horizontal bar of the T placed upwind.

If there is nothing to make a T shape from, a person standing on the far downwind side of the landing place with their arms outstretched and with their back to the wind can form a living symbol for the pilot. Don't make this signal till you have to – and then only in the correct position. It is very similar to another body signal, which means 'need help'.

Night rescue

The helicopter will itself have powerful lights by which the landing or lift off of survivors can take place but you will need lights to bring the pilot in to the landing place. Flares and fires will give an indication of your position once the helicopter is within range.

If you are illuminating from the ground with torches, vehicle headlights or other beams shine them skyward at first to attract attention but, once the pilot has seen you, keep the beams low so that they do NOT dazzle the pilot, and shine them on to the touchdown or winching area.

Sea rescue

If survivors are being winched up from a ship it will help the pilot to place the deck at an angle of approximately 40 degrees to the right of the eye of the wind. If you can control the vessel give a wind speed over the deck of about 29kph (18mph).

HELICOPTER PRECAUTIONS ON LANDING

- When a helicopter touches down at the LP the rotors will be turning. The approach to the aircraft is therefore particularly important both for your own safety and that of the helicopter.

- NEVER approach from the rear. This is a blind spot for the crew and the tail rotor is unprotected. On sloping ground always approach up the slope – bringing you up below the blades.

- NEVER approach down a slope close to the helicopter. You could be at risk from the blades.

- Make sure that you are not carrying anything which could foul the main rotor. If carrying a radio, stow the aerial.

- Keep all sharp objects away from the body panels of the helicopter. They are of light gauge alloy and easily damaged.

- Sit in the seat allocated to you by the crewman, fasten the seat belt and keep fastened until told otherwise.

- Do not attempt to alight until the engine has been shut down after you have landed – even then, wait for directions.

Winching techniques

A double lift is the usual method but a single lift is sometimes used.

Double lift: With a double lift a crewman is lowered on the winch with another strop for the survivor. During the lift the crewman supports the survivor with his legs, gripping with them around the midsection, and supports the head with his hands. After the strop has been put in place and tightened keep arms down by the sides and do not lift them – just lie back and enjoy it!

Single lift: With a single lift you fit yourself into the strop. When you have placed it under your armpits and securely tightened the grommet give the 'thumbs up' sign. Once acknowledged make no further signals until aboard the chopper – if you raise your arms you risk slipping out of the strop!

SURVIVAL CASE STUDY

I often use this case study because it highlights exactly how not to plan and execute an expedition. The main lessons to be learnt are: plan adequately, keep the group together, don't underestimate the environment you are going to, and ensure that you start with enough supplies for survival and for rescue.

While exploring Lows Gully, on Mount Kinabalu in Malaysia, in 1994, two British officers and three Hong Kong Chinese soldiers were trapped for 16 days with only enough rations for three days. At the time they were found, following a huge rescue operation by the British and Malaysian military, medics said that they were on the point of starvation. Mount Kinabalu is not considered a tough climb. The plan was to travel light, so armed with ten days' rations and a video camera, but without radio and flares, they descended into the abyss of Lows Gully.

On the morning of the third day, the expedition reached the edge of the abyss and the men saw for the first time the task ahead of them. The team was split into two groups as not all of the men were in good enough shape to continue with the descent immediately. The fittest men forged ahead while a second group rested with the intention of following when they were ready.

When the second team finally began their descent they had only three days' rations left and, with novice climbers, they made slow progress. Meanwhile, at the bottom of the gully, and with only 4 days' rations left, the advance team began to negotiate their way out of danger, but they met with the gully's unique and unpredictable weather system: mists over the mountain gave way to torrential rains and with nowhere else to go, the water poured down into the gully. In the days that followed, as they slowly struggled through the undergrowth, the team leader and one other team member became separated from the other three in their group. Unable to find them, the two men were forced to continue. With no rations, they attempted to live off the forest, but with dire consequences as one of them became violently ill.

On their return, after 17 days, they learned that the other three in their party had already made it back safely. But the second team were still stranded. The Malaysian Army and a Royal Air Force rescue team were dispatched to search for the team and, after ten days, a Malaysian helicopter finally spotted them and rescued the group. It was 31 days since they had set out.

11

DISASTERS

Accidents or isolation are not the only causes of a survival situation. There are many natural and man-made forces which can produce disasters demanding the same survival skills. Drought, fire, earthquake and flood, for instance, can strike both in the remote survival camp and at home. The survivor should know the techniques for dealing with a forest fire but it is just as important to know how to escape from a burning building or even the site of a bomb explosion.

Home front faces the problems of coping with survival on your own doorstep if the services and supplies on which we normally depend are cut off.

DROUGHT 530
FIRE 532
  Forest fires 532
  Burning buildings 535
  Vehicle fires 538
CHEMICAL AND BIOLOGICAL WARFARE 540
  Toxins 540
  Anthrax 541
  Chemicals 541
GASES AND CHEMICALS 542
  Road and rail 542
FLOOD 544
  Tsunami 546
HURRICANE 547
TORNADO 549
LIGHTNING 550
EARTHQUAKE 551
VOLCANO 554
  Eruption hazards 554
NUCLEAR EXPLOSION 556
  Aftermath 559
HOME FRONT 561
  Food stores 561
  Priorities 564
  Water 564
  Fire 566
  Food 567
  Shelter 569
  Hygiene 570

DROUGHT

Drought, caused by long periods of dry weather or insufficient rainfall, creates desert in areas where it is a permanent condition. Elsewhere, drought may be a regular seasonal feature for part of the year – and predictable year after year. Where it is balanced by a wet season, water can be stored to last through the dry months. The underground stone-cut or concrete-built cisterns of ancient and modern Mediterranean civilizations collect the rainfall of the wet season for the long dry summer.

In temperate regions, if rainfall drops far below the normal, periodic drought may be produced with vegetation unable to obtain water to compensate for what it loses to the air. In cases where the dry spell may not be so obvious but there is still insufficient moisture to provide that needed by the plants, a condition known as 'invisible drought' occurs. The death of vegetation causes deprivation right through the food chains that are based upon it. If the drought becomes severe, dead and dying animals may even pollute the water supplies that still remain.

Fire risk

The corpses of dead animals should be buried in deep graves. Dry ground can be very hard, but burying is the best way to remove these possible sources of infection. They could be burned, but since drought leaves everything tinder dry, the risk of fire spreading is considerable and it could easily get out of hand. Each year fires rage in southern France, California and Australia and, without water to check the flames, they spread rapidly. If you must have a fire, dig down to bare earth and keep the fire small and attended at all times.

Hygiene

In houses, lack of water for washing and sanitation can bring the risk of infection. If the water level in a water closet no longer seals the S-bend, disease may spread from the sewers and personal hygiene may suffer from lack of washing. Do not use the WC but leave sufficient water in the bowl to form a barrier. Make an outdoor latrine to use instead (see *Hygiene* in *Camp Craft*).

Sweating will help to keep pores open and free of dirt, but, even when you need all available water for drinking, try to clean hands after defecation and before preparing food.

Store and conserve water

If a monsoon does not start at the expected time, or a hot dry summer produces a parching of the earth, take precautions by storing as much water as possible and using it wisely. Keep it covered and shaded to avoid evaporation.

Dig a pit for a storage cistern in a shady spot, avoiding tree roots. Line it with a polythene sheet or with cement if available (but don't fill it up until the cement has had a chance to thoroughly dry). If you live in a clay area, dig a pit and line it with clay. If you build the concrete or clay up into a partial dome, it will help to keep the contents cool and leave a smaller opening to keep covered.

NEVER waste water. Water used for cooking can later be used for washing. Boil all drinking water. If a well runs dry you may gain more water by digging deeper, but the further you dig, the further you deplete the water stored in the earth.

In areas with cold nights the drop in temperature may condense the moisture in the air. Use the techniques to catch it described for desert survival (see *Essentials* and *Climate and Terrain*).

High ground will show greater variation between day and night temperatures and will offer better chance of dew to collect in the early morning. It may also offer the advantage of cool breezes.

SURVIVAL ADVICE

If drought persists, especially over more than a year, desert conditions may begin to develop. Evacuation may be the only solution if water supplies cannot be brought in.

Drought can strike anywhere. Even in areas of heavy rainfall, such as Assam, there has been drought when the monsoon has failed.

WARNING

- In conditions of severe drought be especially careful of contamination of water supplies. Disease from dead animals may be rampant. However thirsty you are, boil all water before drinking.
- Flies may be a serious problem at first – ensure that all foodstuffs are covered. Protect from dust, which may become a hazard as top soil is blown away.
- When nature is disturbed in this way, animals act abnormally. Crazed by thirst, normally docile creatures may attack you.

FIRE

Fire requires heat, fuel and oxygen and produces smoke, heat and toxic gases. It may spread by heating its surroundings through direct contact (conduction), through rising gas and smoke (convection) or through heat rays (radiation). Convection is usually the most serious of these forms of heat transfer.

The best protection from fire is prevention. Carelessness with lighted cigarettes and burning matches is the cause of many fires. The sun shining through an abandoned bottle or a piece of broken glass can start a blaze in a dry season.

Fire can strike anywhere – at any time. Precautions should always be taken. Vehicles and buildings SHOULD ALWAYS be equipped with fire extinguishers. Managed forestry plantations are divided by broad paths as fire breaks. Equipment for beating out flames should be found in numerous places. YOU should know how to use it.

FOREST FIRES

If you are present where a fire starts (or where a camp fire accidentally spreads) in woodland, or on heath or grassland, your first action should be to SMOTHER IT.

The first sign of an approaching forest fire will be the smell of smoke. Then you will probably hear the fire before you see flames. You may notice unusual animal behaviour before you realize the cause.

Escape route

If caught in an area where fire is raging, and when it is far too late to put it out yourself, do NOT immediately flee – unless the fire is so close that there is no choice.

Although you may feel that clothing hampers your movement, do NOT discard it for it will shield you from the full force of radiated heat.

Smoke will indicate the direction of the wind – the fire will be travelling fastest in that direction. If the wind is blowing away from you, towards the fire move INTO the wind. Head for any natural fire break – such as where there is a swathe through the trees, where the flames should be stopped. A river is the best break – even if the flames can leap it you will be reasonably safe in the water. In forestry plantations look for the roadways and firebreaks.

Do NOT run wildly. Choose your escape route. Check the surrounding terrain and the wind direction to assess the possible spread of the fire.

If the wind is blowing towards you the fire is likely to travel more quickly – and the flames can leap a larger gap. Fire travels faster uphill so do NOT make for high ground. Try to go around the fire if possible – but some forest fires burn on a front several kilometres wide. If you can neither skirt or outdistance the blaze take refuge in a large clearing, deep ravine, watercourse or gulley.

Into the fire

Sometimes the best escape route may be to run THROUGH the flames. This is impossible if they are very intense and the area covered by the actual fire is great. In a large clearing or on heathland, however, it may be possible to run through less dense fire to refuge on the already burned-out land. Cover as much exposed skin as you can and if you have water available tip some over you to damp down clothing, hair and any flesh you have not been able to cover. Dampen a piece of cloth to cover your nose and mouth.

Thick vegetation will burn fiercely and slow you down – so choose the spot for your breakthrough with care. Make your mind up, then do not delay. Take a deep breath. Cover your nose and mouth to keep out smoke and RUN.

FIGHTING A FOREST FIRE

In areas where there are forestry plantations you should see racks of fire-beating equipment at intervals along the main routes. This consists of bundles of twigs (usually birch), tied in a broom, and spade-shaped beaters with rubber blades. They can be effective in putting out the beginnings of a blaze.

Despite the name, do NOT beat rapidly with them – that will only help to fan the flames and spread sparks. The object is to SMOTHER the fire by bringing the beater down over the flames to extinguish them. The flat blades are particularly effective in smothering a fire beginning among leaves and undergrowth.

If no equipment is available use a coat or blanket to smother the fire and to cut off oxygen from the flames – or use a leafy branch to beat it out.

STAY IN A VEHICLE

If caught in a forest fire in a vehicle, stay inside and keep the windows tightly shut. Turn off the ventilation system. The car will give you some protection from radiant heat. Drive away from the fire if you can but, if immobilized, stay put.

People have survived by staying in a vehicle until the glass began to melt, by which time the fire had moved beyond them. If they had panicked and run into the fire they would have died.

There is a danger of a petrol tank exploding but your chances are much greater than outside, if the fire is intense around the vehicle.

Going to earth

If there is no natural break or gully in which to shelter and the fire is too deep to think of running through it, you may have to seek the protection of the earth itself.

People have survived fierce fires by digging themselves in and covering themselves with earth, allowing the fire to burn over the top of them. The risk is considerable, not just from heat but from suffocation: fire burns up oxygen.

Scrape as much of a hollow as you can, throwing the earth on to a coat or cloth if you have one, then pull the cloth over you with its earth covering. Cup your hands over your mouth and nose and breath through them. This won't increase the amount of oxygen, but it will cool down and filter the very hot air and sparks, which can damage the respiratory system. Try to hold your breath as the fire passes over.

Fight fire with fire

It may be possible to use fire itself to create protection, if there is no way of getting out of the path of the fire or of going through it – but it is still some distance away.

The technique is to burn a patch of ground before the main fire reaches it. With nothing left to ignite, the flames cannot advance, giving you a place of refuge. The main fire must be sufficiently far away for your fire to burn a space it cannot jump before it arrives.

Light your own fire along as wide a line as possible – at least 10m (33ft) wide, but 100m (330ft) would be better. It will burn in the same direction as the main fire, creating a break which you can move into. Make sure you determine the wind direction correctly.

WARNING

Winds may be swirling and fires create their own draughts, so you may still have to make a dash through your own flames. The main fire must be far enough away for your own fire to burn and pass. Do NOT underestimate the speed at which flames travel – they may be approaching faster than you can run. Do NOT light another fire unless you are desperate and fairly certain of the outcome.

BURNING BUILDINGS

Smoke will usually be the first indication of fire. If the fire is still small, attempt to extinguish by covering with a blanket or thick curtain (to deprive the flames of oxygen), or using sand, water or fire extinguisher as available and appropriate.

Electrical fires

If there is any possibility of a fire being caused by an electrical fault, do NOT use water until the power has been turned off, preferably at the mains. Turn off gas at the mains too. If television sets or computer VDUs catch fire do NOT use water on them. Even after being disconnected, there is sufficient residual electrical charge to give you a shock (even to kill you) and the cold water may make the tube explode. Smother them and approach from behind because of the risk of the tube exploding.

FIRE EXTINGUISHER

PULL – or otherwise to release any lock
AIM – at the base of the fire
SQUEEZE – or press the handle/trigger
SWEEP – from side to side

Check to see what kind of extinguisher you have. Some are designed for small, simple fires and contain water – they are NOT suitable for oil or electrical fires. Others are designed for oil, grease, paint or solvent fires (a chip pan or petrol spillage, for instance). A third type is for electrical fires or those where live cables are exposed. A multipurpose dry-chemical type extinguisher can be used on most kinds of fire.

KNOW HOW TO USE YOUR EXTINGUISHER!

Evacuation

If the fire is already too big to fight with the resources available, evacuate the building. Turn off power at the mains. Close all reachable doors and windows. Try to contain the fire while evacuation is completed and to prevent the fire's spread before fire-fighting services or rescuers arrive.

Fire travels upwards faster than downwards – though a collapsing floor or wall will carry the flames to a lower level. Staircases, lift-wells and ventilation shafts are particularly dangerous.

NEVER use a lift during a fire. If you need to move down a building, use a smoke-free staircase.

Before opening any doors look for smoke around their edges and check for heat. Metal doorknobs are a good test. If they feel warm do NOT open the door – use the back of your hand for the test. Grasping the knob could cause a burn. A stout door can keep back flames for 30 minutes or more (but do not rely on this with thin-panelled modern doors – unless they are fire doors).

SURVIVAL ADVICE

If there is no alternative to going into a burning room: brace your foot against the door and open it only a crack – this will help to stop it being forced wide open by the pressure of hot air and gases inside. Crouch low to enter, opening the door as little as possible. This will expose you to less smoke and heat and give less chance for the fire to pass through the gap. Close the door behind you to delay the fire's spread.

Waiting for rescue

If you find yourself isolated and unable to get to safety, go to a room as far from the fire as possible (but NOT on a higher level unless you are certain that rescuers with long ladders or other equipment are very close). If there is a choice, pick a room with the softest ground outside, no railings below or where there are bedclothes or curtains which you can knot into a rope. Lawns, flowerbeds, even gravel paths are softer than cobbles, concrete or paving. If you are eventually forced to drop down on to hard concrete a sloping surface will be likely to do you less damage.

Close the door and fill any gaps around it with curtains, mats or other thick material, which will take time to catch alight. Wet them if you can. If the alarm has not already been given try to attract attention through the window.

To break the window, use a piece of furniture. You could kick it out, but do not bring your foot back quickly – you will have to negotiate broken glass. If you have to use your hand, wrap it first and punch. Alternatively, when protected by a thick jacket, an elbow may be effective.

Preparing to 'jump'

If no rescue is forthcoming do NOT jump – drop. Tie sheets, blankets, mats, loose covers and other strong material together to make a rope – even if it does not reach the ground it will reduce the distance you will have to fall. Tie them with reef knots and test that each is firm by pulling. Push a heavy piece of furniture over to the window and secure one end of your rope to it, or tie your rope to heating pipes – or smash the window and tie on to a thick part of the frame. If the rope is not long enough drop cushions, pillows, a mattress – anything which will soften your landing, below the window.

If there is nothing to use as a rope, lower yourself out of the window and hang from the sill. If there is no sill hang from the bottom of the window frame.

SURVIVAL ADVICE

Do NOT jump out – unless there is a party of firefighters waiting to catch you in a blanket. Where there is something to break your fall make use of it, do not try to drop clear. A car roof makes a good cushion to drop on for it will give slightly to your weight.

BEWARE: Trees may break your fall, but there is a risk of your being impaled on the branches.

Dropping from a height

Having taken every recommended measure to reduce the height of your fall, think of protecting your head – a motor-cycle crash helmet is ideal but a jersey or towel wrapped round your head like a turban will help.

When you are as low as possible, and ready to drop, push yourself away from the building with the side of one foot and, as you let go, turn away from the wall and bend the knees. Leave the arms up to protect the sides of the head.

As you land bend the knees more and roll over on to one side, carrying the roll on to your back (still protecting your head and with your legs in the air). This helps spread the impact over a larger area and increases your chances.

Falling onto a slope

Face down the slope as you drop, with legs together, with the knees slightly bent. Bring the head down on to the chest and the elbows tightly into the sides, the hands protecting the head. Land on flat feet, allowing the knees to bend fully, rolling forward in a tight somersault. It is the method that parachutists use.

WAIT UNTIL THE VERY LAST CHANCE OF RESCUE BEFORE RISKING A DROP GREATER THAN 4M (13FT).

ESCAPING THROUGH FIRE

- If forced to travel through flames to reach safety, cover yourself (including your head) in a blanket, curtain or overcoat – wet it if you can – then take a deep breath and go.
- If your clothes catch alight do NOT stay on your feet when out of the fire. The flames and smoke will travel up your body, over your face and into your lungs.
- Do NOT run – this will only fan the flames.
- Roll on the ground and try to wrap yourself in something that will smother the flames – a mat, blanket or overcoat.
- If someone else comes running out of a fire with their clothes alight, push them to the ground and use the same methods of denying the flames oxygen. Do NOT hug them to you or your clothes may catch fire.

VEHICLE FIRES

The greatest danger with cars is the risk of the fuel tank being ignited – it could explode like a bomb, scattering burning petroleum everywhere. The aim is to control the fire before it can reach the tank. Everything has a flash point and a fuel tank is more at risk than most things. Usually a fuel line (if not armoured) will catch fire first, acting as a fuse which eventually ignites the tank.

In a garage

If a car catches fire in a confined space, such as a garage, smoke and toxic fumes will soon build up. Try to put the fire out first – but if that is not practicable remove the car from the building, before it further endangers life and property.

Do NOT get into the car. You can do everything from outside including steering. If possible, push or pull the car out. If your car has a button starter, select a low gear or reverse and use the starter to bounce the car out. With conventional ignition, turn the key in short bursts. BE PREPARED. The car will jerk forward violently.

FIRE EXTINGUISHER

KEEP IT WHERE YOU CAN REACH IT!
DON'T keep your fire extinguisher in the boot/trunk – keep it in the passenger compartment where you can get at it immediately. Any impact could distort the boot lid and prevent you opening it.

In a crashed car: doors may jam. If it catches fire get through any window or kick the windscreen out.

If the fire is inside the car: use the extinguisher or smother it with a rug or coat. Synthetic materials used in upholstery in many cars burn rapidly and give off thick smoke and toxic gases. These will persist even when the flames are out so get out into the air as soon as possible.

If the fire is also on the outside: amid spilt fuel for example – keep the windows shut and drive out of the danger zone, abandoning the car as soon as it is safe to do so.

FIRES IN THE AIR

Aeroplanes are equipped with automatic extinguishers for engine fires and hand-held extinguishers in the cabin. Action should be taken immediately. On civil airlines summon a steward or air hostess immediately you suspect fire – they know where equipment is and how to use it. Avoid creating panic in other passengers. If you see smouldering or flames, smother with an airline blanket or clothing.

The main fire dangers are: before take-off when there is volatile fuel and vapour around the plane and especially when landing under difficult circumstances when fuel tanks could be ruptured and electrical or friction sparks ignite aviation spirit. Every safety precaution is taken to ensure that fire is not a hazard.

CHEMICAL AND BIOLOGICAL WARFARE

Terrorism is not new. The destruction of property, murder, threats, fear and panic have long been the tools of those who believe that violence is a means of achieving political and criminal goals. What makes modern terrorists distinct from their historical predecessors is the technology at their disposal and the greater range of opportunities for terrorism which modern life provides.

The global nature of media coverage and the publicity it can give has also fuelled the use of terrorism. There is no doubt, also, that some nations finance and support the activities of terrorists in other countries with which they have political difference.

More value is now placed on the lives of ordinary people, the more innocent the victim, the better for terrorism. The reality is that we are ALL potential targets and that we are ALL vunerable.

Certain countries of the world pose a real threat by producing vast amounts of lethal bacteria and chemicals. These are relatively easy to disperse and threaten large areas of the globe.

Biological agents fall into two groups, pathogens (germs) and toxins. Pathogens are living micro organisms that cause lethal incapacitating diseases like anthrax. Toxins are poisons that are lethal to humans, affecting the nervous system, and causing cell death.

Germs must be inhaled, enter through a break in the skin, or through the digestive tract. They do not react immediately as they must multiply inside the body and overcome the body's defences. This incubation period varies from hours to months depending on the germ. Biological agents are hard to detect, none of our physical senses can detect them. Often the first signs are symptoms of the victims exposed to the agent, and sick-looking plants and animals.

TOXINS

Toxins occur naturally in plants and animals, but can be manufactured and used as weapons which are a thousand times more lethal. Toxins produce a similar effect to chemicals but do not respond to the same treatment. Unlike germs they can penetrate unbroken skin, and their symptoms appear immediately. These include; paralysis, convulsions, fever, blisters/rashes, shock and death.

Protection

If you suspect the presence of biological agents, put something around the mouth and nose to prevent inhalation. It is unlikely that you have a mask so improvise by wrapping a damp cloth around the face. Cover all exposed parts and leave the area as soon as possible. Button/zip-up all clothing. Tuck trousers into socks, and wear gloves. When out of the danger area, wash thoroughly with soap and water. Clean teeth, hair and fingernails. Wash all clothing and equipment in hot soapy water.

ANTHRAX

Anthrax is a disease carried by animals and still kills many humans annually. The disease is transmitted from dead animals in areas like South America, Southern Europe and the Middle and Far East. Signs on the dead carcass include bloating, incomplete rigor mortis, dark coloured blood oozing from nostrils and anus. Avoid all contact and wash thoroughly if accidental contact is made. Seek professional advice soonest. Antibiotics will help, and there is a serum available from the treatment of anthrax.

The early signs of the disease are pustules with dark centres. Anthrax is also produced as a biological weapon.

CHEMICALS

Chemical agents take the form of liquids, gas and aerosols. They may affect breathing, the nervous system or the blood.

Chemical agents are difficult to detect so watch for suspicious signs in your companions. Difficulty in breathing, coughing, itching and tears are the most obvious ones. The landscape will look wasted, with vegetation being discoloured and limp, and maybe dead animals and insects scattered around. Some agents are odourless and others have a distinctive smell. The smell of almonds may indicate a blood agent, and newly cut grass might indicate a choking agent.

Protection

The whole body must be covered with suitable clothing that will prevent the chemicals penetrating. Waterproof outer garments, goggles and mask are essential. Get out of the danger area and decontaminate as soon as possible, the same as with pathogens.

GASES AND CHEMICALS

Explosive gases can occur and build up in caves and mines, but the main dangers to the survivor in the wild will be oxygen deficiency and carbon monoxide poisoning in restricted spaces and shelters.

In house and vehicle fires and in industrial locations there is risk from toxic fumes being produced when plastic and other materials burn. Industrial and road accidents involving chemicals present another danger.

There is little that can be done to protect yourself in a disaster such as that at Bohpal, in India, once they have happened, other than stay indoors and keep your doors and windows closed. But you can try to discover any risks that are likely from industrial or other plant in your area. There should be government and local authority checks and controls operating to enforce proper safety procedures. Many authorities demand labelling of places where chemicals are used or stored – in London, for instance, by a yellow triangle and/or the sign HAZCHEM.

If handling dangerous materials – not just at work, potentially dangerous substances are used in domestic chores, gardening and hobby pursuits – follow recommended safety procedures. AVOID contact with chemicals and inhalation of fumes and take precautions against spillage, shaking, breakage or accidental mixture with other chemicals which could produce a dangerous reaction.

If exposed to fumes or chemical hazard get to fresh air as rapidly as possible and flush away chemicals with copious quantities of water – BUT adding water to some chemicals can cause even more dangerous reactions. Learn the properties of materials you handle or transport and measures for dealing with accidents.

AVOID touching anyone contaminated with chemicals and do NOT give mouth-to-mouth resuscitation until you know what chemicals are involved.

ROAD AND RAIL

Keep well clear of an accident involving a bulk tanker or other vehicle carrying dangerous substances. You may be able to see spillages on the road surface. Gas escapes may NOT be visible.

Chemical fires may be fought by appropriate techniques. If not properly trained and equipped to handle them you may aggravate the situation, endangering yourself and others.

The usual action in such cases will be to cover any flammable substances with foam to exclude oxygen, and to dilute chemicals that do not react dangerously to water with as much water as possible.

There are too many cases on record of public-spirited individuals losing their own lives in such circumstances – like the man who went to aid the driver of an overturned tanker which had been carrying powerful acid. By the time the fire-service arrived his gold wedding ring was all that remained of him.

SURVIVAL ADVICE

BUT if you do not know what exact substances are involved and the appropriate action – KEEP CLEAR. CALL THE FIRE SERVICE OR POLICE AND LEAVE ALONE.

Vehicles may be painted with a panel warning that they carry potentially dangerous chemicals which includes a code of figures and letters which tell the emergency services what kind of action to take, what kind of protection is essential and whether the chemical is explosive or poisonous. In the British code, for instance, a final E is an instruction to EVACUATE. The figures 1 or 2 at the beginning of the code suggest that water may be used, but that is only part of the information needed – the use of water could produce fumes which demand use of a respirator. Protective clothing may be essential.

Details of these codes are not usually given to the general public, for the information that they provide does not give practical guidance to the untrained, beyond a warning that dangerous materials are carried.

Although the information provided by the codes is valuable to the officers of the emergency services, they do not know the proportions in which spilt chemicals are concentrated, or whether combinations of chemicals could produce results quite unlike those expected. Where dangerous stores and goods in transit are registered with the authorities rescue services may be able to obtain additional information, but uncontrollable elements such as wind direction, temperature and various other factors will all affect the way in which such accidents are handled.

If you see a British hazardous chemical code, an EC Accord Européen Relatif au Transport International des Marchandises Dangereuses par Route (ADR) sign, which gives much less detailed information, or the United Nations International Emergency Action Code (EAC), which only covers a limited number of emergency procedures, KEEP YOUR DISTANCE.

FLOOD

Flooding may be caused by the overflowing of rivers, lakes and reservoirs caused by heavy rains (not necessarily rainfall at the place where the flood occurs); by the build-up of sea or lake water due to the effects of submarine earthquake, hurricanes and freak high-tides and winds; or by the collapse of dams or dykes.

Heavy rain can rapidly produce torrents where there was a dry riverbed, or a build up in a narrow channel or behind a barrier which then gives way to a rushing wall of water that envelopes everything in its path.

Persistent rainfall over a long period after a dry spell and heavy storms should alert you to keep clear of water channels and low-lying ground, but a flood can affect much wider areas. It is always safer to camp on a spur. If the water is rising, move to higher ground. In hilly areas keep out of valley bottoms which are particularly prone to flash floods.

Food is not likely to be a problem, at least at first, for animals will also head for high ground. Both predators and prey are likely to concentrate on getting to safety – but beware of injury from panic-stricken animals in the water.

Drinking water may be difficult to obtain, for the water swirling around you may be contaminated. Collect rainwater to drink and boil any other water before you use it.

Flooded buildings

If you are in a solid building when the water begins to rise, stay where you are if it is rising rapidly. You will be less at risk than trying to evacuate on foot. Turn off gas and electricity and prepare emergency food supplies, warm clothing, drinking water in screw-topped plastic bottles or other well-sealed containers. It is important to keep all water containers sealed to avoid accidental spillage or contamination. If you can, collect a torch, whistle, mirror, brightly coloured cloths, or flags, that would be useful for signalling, and add them to your gear. A camp stove will be valuable for heating food and drinks and for warmth. Candles are also useful – and DON'T forget the matches.

Move upwards

Move to an upper floor, or on to the roof if in a single-storey building. If you are forced to occupy the roof, erect some kind of shelter. If it is a sloping roof, tie everyone to a chimneystack or other solid structure which can be expected to stay firm. If the water looks like continuing to rise, prepare some kind of raft. If you have no ropes to tie things together, use bed sheets. Unless the water threatens to wash your building away, or rises so high that you are forced to evacuate, stay until it stops rising.

Flood readiness

If you live in a river valley or in a coastal area prone to flooding, find out how high you are above normal water levels. Learn the easiest route to high ground – not necessarily the highway route, for major roads tend to follow drainage channels in valley bottoms. In rainy periods listen out for flood warnings which will often predict the levels to which waters may be expected to rise and the areas likely to be affected.

Even a few inches of floodwater can do a lot of damage and it is worth laying sandbags or plastic shopping bags filled with earth along the bottom of doorways and windows to keep out as much water as possible. If you have a really sound building and doorways, windows and other apertures are blocked, water will not enter. The central heating flue, ventilation bricks and other spaces must all be sealed.

If a high flood is likely, there is not much point trying to keep water out of basements. Indeed, in some houses with basements this could cause extra damage due to uneven pressure on the basement walls. If you are bound to be flooded, you could consider flooding the basement yourself with clean water, so that the pressure is equalized. You will have less filth and debris to clear up later.

Evacuation

If you are abandoning your home bring outdoor furniture and other movables indoors – that will reduce the amount of debris floating or being swept along outside.

When walking or driving to a safer location: Remember that a small drop in the level of the roadway down a hill can make a considerable difference to the water depth.

Do NOT attempt to cross a pool (or a stream) unless you are CERTAIN that the water will not be higher than the centre of the car's wheels or higher than your knees.

If you must cross: use river-crossing techniques (see *On the move*).

If crossing bridges which are underwater: take special care – you may not be able to see that the flood has already swept part of the bridge away.

Flash floods

In times of sudden heavy rainfall keep out of valley bottoms and stream beds both during and after rainfall. Remember that you don't have to be at the bottom of a hill to be caught by water rushing down it – often carrying mud and a deadly debris of broken trees and rocks.

Coastal flooding

Is usually a combination of high tides and winds which make them even higher. Flood warnings will usually be given and evacuation is the best action.

Flood aftermath

As the waters recede they leave a scene of devastation littered with debris and the bodies of flood victims. With decay and the pollution of the water comes the risk of disease and extra precautions are necessary. Burn all animal corpses – do not risk eating them – boil all water before using. Some crops may still be available after the flood waters recede and birds that have escaped the flood will be safe and good to eat.

SURVIVAL SCENARIO

What can you do to prepare for disaster?

Make sure that you are aware of what the dangers are. This may depend upon where you are travelling or what time of year you will be there. You should be aware of what contingencies are in place in case of say earthquake or flood. Ask the locals for advice.

TSUNAMI

A tsunami is linked with an earthquake beneath the ocean, creating a series of waves which can reach more than 30m (98ft) high and causing a considerable amount of damage along coasts.

Warnings of tsunami are issued from the National Oceanic and Atmospheric Administration's Pacific Tsunami Warning System headquarters in Hawaii.

Not all earthquakes cause tsunami, but any earthquake could. Keep away from shores and take to higher ground when there are tremors. Do NOT go to look for a tsunami – if you are close enough to see the wave, you are too close to escape it, unless high above its level. There is little defence against a moving wall of water. Evacuate.

HURRICANE

A hurricane is a wind of high speed – above force 12 on the Beaufort Scale – which brings torrential rain and can destroy any flimsy structures. It is a tropical form of cyclone, which in more temperate latitudes would be prevented from developing in the upper levels of the air by the prevailing westerly winds.

Hurricanes are known by various names around the world:

Hurricane: Caribbean and North Atlantic, eastern North Pacific, western South Pacific.
Cyclone: Arabian Sea, Bay of Bengal, southern Indian Ocean.
Typhoon: China Sea, western North Pacific.
Willy-willy: Northwest Australia.

Hurricanes develop over the ocean when sea temperatures are at their highest, especially in late summer. Warm air creates a low pressure core around which winds may rotate at speeds up to 300kph (186mph) or more, circling anti-clockwise in the northern hemisphere, clockwise in the southern. The strongest winds are usually 16-19km (10-12 miles) from the centre of the hurricane but the centre, or 'eye', brings temporary calm. The 'eye' may be from 6–50km (4–30 miles) across and the largest hurricanes up to 500km (300 miles) in diameter. They can occur at any time of year but, in the northern hemisphere, the main season is June to November – in the southern, November to April (especially January and February). Hurricanes are not a feature of the South Atlantic.

Pattern of the hurricane

Out at sea hurricanes will build up force and begin to veer toward the pole, the wind speed usually being greatest on the poleward side of the eye. They can travel as fast as 50kph (30mph) wreaking devastation on islands and along shorelines they pass over, but usually slowing down when they reach the mainland to a speed of about 16kph (10mph).

Hurricane warnings

Satellite surveillance enables meteorologists to see hurricanes developing far out in the ocean, to track their progress and to give warning of their approach. Some hurricanes move very erratically, so sailors particularly, should monitor forecasts in hurricane areas.

Without radio to alert you, the growth of swell can be an indication of a hurricane – when coupled with other conditions such as highly coloured sunsets or sunrises; dense banners of cirrus cloud converging towards the vortex of the approaching storm; abnormal rises in barometric pressure followed by an equally rapid drop.

SAFETY PRECAUTIONS

Get out of the hurricane's path if you can.

- Hurricane warnings are usually issued when one is expected within 24 hours and will give you plenty of time to evacuate its path, if you are prepared.
- Keep away from the coast, where destruction will be worst, with flooding and a tidal wave, and from river banks.
- Board up windows and secure any objects outdoors that might be blown away.
- At sea take down all canvas, batten down the hatches and stow all gear.

If you are in a solid building and on high ground STAY WHERE YOU ARE – travel in a hurricane is extremely dangerous. The safest place is usually in a cellar or under the stairs. Store drinking water – water and power supplies may be cut off by the storm – and have a battery-operated radio to keep in touch with any instructions issued. If not in a sturdy structure, evacuate to a hurricane shelter. Shut off power supplies before you leave.

Seeking shelter

Outdoors a cave will offer the best protection. A ditch will be next best. If unable to escape lie flat on the ground where you will be less of a target for flying debris. Crawl to the lee side of any really solid shelter such as a stable rocky outcrop or a wide belt of large trees. Beware of small trees and fences which could be uprooted.

SURVIVAL ADVICE

Stay where you are when the hurricane appears to have passed – there will usually be less than an hour of calm as the eye passes and then the winds will resume in the opposite direction. If sheltering outdoors move to the other side of your windbreak in preparation or move to better shelter if close by.

TORNADO

Tornadoes are violent storms associated with low atmospheric pressure and whirling winds. They apparently develop when air at the surface has been warmed and a column of air descends from the base of cumulonimbus storm clouds above. Air rushing into the low pressure area begins to rotate fiercely.

Tornadoes are the most violent of atmospheric phenomena and the most destructive over a small area. Wind speeds have been estimated at 620kph (385mph).

The diameter of the 'twister' at ground level is usually only 25–50m (80–160ft) but, within it, the destruction is enormous. Everything in its path except the most solid structures is sucked up into the air. The difference in pressure outside and inside a building is often the cause of collapse – or 'explosion'. Tornadoes can sound like a spinning top or engine and have been heard up to 40km (25 miles) away. They travel at 50–65kph (30–40mph).

At sea tornadoes produce waterspouts. Although they can occur elsewhere, tornadoes are most prevalent over the prairies of the United States, in the Mississippi-Missouri valley and in Australia. They can develop in a hurricane.

Tornado precautions

Take shelter in the most solid structure available – reinforced concrete or steel-framed if possible, but preferably in a storm cellar or cave. In a cellar stay close to an outside wall, or in a specially reinforced section. If there is no basement, go to the centre of the lowest floor, into a small room or shelter under sturdy furniture – but not where there is heavy furniture on the floor above. Keep well away from windows.

Firmly close all doors and windows on the side facing the oncoming whirlwind and open those on the opposite side. This will prevent the wind getting in and lifting the roof as it approaches and equalize the pressure to prevent the house 'exploding'.

Do NOT stay in a caravan or car, they could be drawn up by the storm.

Outdoors you are vulnerable to flying debris and to being lifted up – though people have been lowered to the ground again unharmed! You CAN see and hear a tornado coming. Get out of the way. Move at right angles to its apparent path. Take shelter in a ditch or depression in the ground, lie flat and cover your head with your arms.

LIGHTNING

The release of electrical charges built up in clouds can be especially dangerous on high ground or when you are the tallest object. In a lightning storm keep away from hill brows, from tall trees and lone boulders. Make for low, level ground and lie flat.

Insulation

If you cannot get away from tall objects, but have DRY material which will provide insulation, sit on it. Rubber-soled shoes may help insulation but are not a guarantee that you will be safe. A dry coil of climbing rope makes good insulation. Do NOT sit on anything wet. Bend your head down and hug your knees to your chest, lifting your feet off the ground and drawing in all your extremities. Do not reach down to the ground with your hands, that could give a contact to conduct the lightning. If you have nothing which will insulate you from the ground lie as flat as you can.

Stay low

You can sometimes sense that a lightning strike is imminent by a tingling in the skin and the sensation of the hair standing on end. If you are standing up, drop to the ground AT ONCE, going first to the knees with the hands touching the ground. If you should be struck, the charge may take the easiest route to the earth through your arms – missing the torso and possibly saving you from heart failure or asphixiation. QUICKLY LIE FLAT.

Do not hold metal objects when there is lightning about and keep away from metal structures and fences. However, do not jettison equipment if you will lose it altogether (when climbing, for instance). A dry axe with a wooden handle may spark at the tip, but is well insulated. Proximity to large metal objects can be dangerous, even without contact, for the shockwave caused by the heated air – as the lightning passes – can cause damage to the lungs.

Shelter

One of the best places to shelter in a lightning storm is at least 1m (3ft) inside a deep cave with a minimum of 1m (3ft) space on either side of you. Do NOT shelter in a cave mouth or under an overhang of rock in mountainous country. The lightning can spark across the gap. Small openings in the rock are frequently the ends of fissures which are also drainage routes and automatic lightning channels.

EARTHQUAKE

Earthquakes are perhaps the most feared of all nature's violence – they come suddenly with little warning. Unlike other natural hazards, such as flood and fire, little can be done to prepare for them. They range from minor vibrations in the earth, detectable only on delicate measuring instruments, to great upheavals tearing whole mountains apart.

Minor earth tremors can happen anywhere, but major quakes are confined to known earthquake belts where buildings could be planned to withstand them or to cause little damage if they do collapse – like traditional buildings in Japan. Modern cities rarely take this into account.

With constant monitoring by seismologists, major earthquakes can be predicted and some evacuation may be possible. Animals become very alert, tense and ready to run.

A detailed knowledge of local geology may indicate the weak points along major fissures but the tremors and the waves of movement through the earth could extend to almost anywhere in an earthquake zone.

Causes

Earthquakes are caused by the sudden release of tension built up in the Earth's crust, as deep as 700km (435 miles) below the surface – but only a rupture in the upper tens of kilometres is likely to produce movement affecting the surface. The shock wave for the 1964 Alaskan earthquake, for instance, was at about 20–30km (12–19 miles) deep.

The earthquake belts lie along the edges of the semirigid plates that form the Earth's crust. The deepest quakes occur along the ocean trenches, forming and destroying volcanic islands. The most violent tend to occur in areas where one plate is thrusting beneath another, as along the west coast of North America, where the San Andreas Fault is a particularly vulnerable zone.

A succession of preliminary tremors, known as foreshocks, often followed by a seismically quiet period, usually precede a major quake, which they can actually trigger. These initial tremors may not be noticeable.

Domestic earthquake precautions

Stay tuned to a local radio station for up-to-date reports and advice if you have warning of a possible earthquake. Turn off gas, electricity and water if advised to do so. Remove large and heavy objects from high shelves, from which they might fall on you. Put bottles, glass, china and other breakables in low cupboards. Shelves should have a lip or low barrier to stop things slipping off. Cupboard doors should have positive fastening – not just magnetic catches. Secure or remove suspended objects, such as large light fittings and hanging flower bowls.

Have ready in case needed: fresh water and emergency food, a torch, first-aid materials and a fire extinguisher.

SAFETY PRECAUTIONS

KEEP AWAY FROM ANYTHING THAT MIGHT FALL ON YOU: trees in the open air, for they may be uprooted; building in towns, for – even if they are stable enough not to be demolished – pieces may come crashing down. Ideally evacuate to an open area but, if that is not possible – and you may have little time – it is safer to stay indoors. In the street ruptured gas mains and power cables may add to the hazards. People going in and out are the ones at most hazard from falling masonry on buildings.

Indoors

If you are indoors when an earthquake strikes, stay there. Douse fires. Stay away from glass, including mirrors, and especially from large windows.

- An inside corner of the house, or a well-supported interior doorway are good places to shelter.
- A lower floor or a cellar probably gives the best chance of survival. An upper floor could come crashing down with you on it. Make sure there are plenty of exits.
- Get beneath a table or other piece of substantial furniture which will give both protection and an air space.
- In a shop keep away from large displays of goods which could come crashing down.
- In high-rise offices stay put. Never go into a lift. Staircases may attract panicking people. Get under a desk.

In a car

Stop as quickly and safely as you can – but stay in the car. It will offer some protection from falling objects. Crouch down below seat level and you will be a further protected if anything falls on the car. When the tremors cease keep a watch for any obstructions and hazards: broken cables, undermined roadways or bridges which could give way.

Outdoors

Outdoors lie flat on the ground. Do NOT try to run. You will be thrown about and could be swallowed in a fissure.

- Keep away from tall buildings.
- Do not go deliberately underground or into a tunnel where you could be trapped by blockage or collapse.
- If you have managed to get to an open space do not move back into buildings for if minor tremors follow they could collapse any structure left unstable by the first quake.
- On a hillside it is safer to get to the top. Slopes are liable to landslide and there would be little chance of survival for anyone caught in the thousands of tons of earth and rock that could move with terrifying speed. People have been known to survive by rolling into a tight ball on the ground.
- Beaches – provided they are not below cliffs – are initially fairly safe but, since tidal waves often follow a quake you should move off the beach to high open ground as soon as the tremor has finished. Further tremors are unlikely to be as dangerous as a tsunami.

Aftermath precautions

Rupture of sewage systems, contamination of water and the hazards of the bodies trapped in the wreckage can all make the risk of disease as deadly as the earthquake itself. Bury all corpses, animal and human.

- Do NOT shelter in damaged buildings or ruins. Build a shelter from debris.
- Take special care over sanitation and personal hygiene. Filter and boil all water.
- Do not strike matches or lighters, or use electrical appliances, if there is any chance of a gas leak. Sparks ignite gas.
- Check sewage services are intact before using lavatories.
- Open cupboards carefully, objects may tumble out.
- Be prepared for aftershocks.

BE CALM! THINK FAST!

Speed is essential if an earthquake strikes. There is little time to organize others. Use force if necessary to get them to safety or pull them to the ground.

VOLCANO

Active volcanoes are found in the areas of the world which are also most prone to earthquakes – where there is most movement beneath the surface. Long dead volcanoes and evidence of ancient volcanic activity can be found elsewhere – such as Arthur's Seat in Edinburgh, Scotland.

A channel is formed when pressure forces molten rock (magma) to the surface through a fault. The channel is usually the main outlet for further eruptions, though other vents may appear. A major eruption may blow the whole top off a mountain.

The molten rock, usually known as lava when it reaches the surface, may be of two kinds: granite lava, which is viscous and slow moving, and basalt lava, which flows more rapidly at speeds of 8–16kph (5–10mph). Granite lava tends to block the vent of the volcano, eventually cleared by an explosion due to the build up of pressure beneath – showering lava and pieces of mountain around for considerable distances and causing fires.

ERUPTION HAZARDS

Lava

Although it is possible to outwalk or outrun most basalt lava flows they continue relentlessly until they reach a valley bottom or eventually cool off. They crush and bury anything in their path. Lava flows are probably the least hazard to life produced by an eruption, for the able-bodied can get out of their way. Other dangers are NOT so easily avoided.

Missiles

Volcanic missiles, ranging from pebble-size fragments to great lumps of rock and 'bombs' of hot lava, can be scattered over considerable distances. A shower of volcanic ash can fall over a much greater area, some dust being carried to great heights and are dispersed around the world affecting weather conditions.

If evacuating from close to the volcano, hard helmets of the kind worn by building workers, motorcyclists or horseriders will give some protection. Over a wider area, where evacuation may not be necessary, protection should be worn against the effects of ash and any rain which accompanies it.

Ash

Volcanic ash is not ash at all, but pulverised rock forced out in a cloud of steam and gases. Abrasive, irritant and heavy, its weight can cause roofs to collapse. It smothers crops, blocks transport routes and watercourses and, combined with toxic gases, can cause lung damage to the very young, old and those with respiratory problems. Only very close to an eruption are gases concentrated enough to poison healthy people. But, when sulphur dioxide in the ash cloud is combined with rain, sulphuric acid (and sometimes others) are produced in concentrations which can burn skin, eyes and mucous membranes. Wear goggles (ski-goggles or a snorkelling mask which seal around the eyes, NOT sunglasses). Use a damp cloth over mouth and nose, or industrial dust masks if available. On reaching shelter remove clothing, thoroughly wash exposed skin and flush eyes with clean water.

Gas balls

A ball of gas and dust may roll clown the side of a volcano at a speed of more than 160kph (100mph). This phenomenon (called by scientists a *nué ardente* – glowing cloud) is red hot and moves too fast to be outrun. Unless there is a strongly-built underground shelter nearby, the only chance of survival is to submerge underwater and hold your breath for the half-minute or so it will take to pass.

Mudflows

The volcano may melt ice and snow and cause a glacial flood or – combined with earth – create a mudflow, known as a lahar. This can move at up to 100kph (60mph) with devastating effect, as in Colombia in 1985. In a narrow valley a lahar can be as much as 30m (98ft) high. They are a danger long after the major eruption is over and are a potential risk even when the volcano is dormant if it generates enough heat to produce meltwater retained by ice barriers. Heavy rains may cause it to breach the ice.

WARNING

Volcanoes usually show increased activity before a major eruption with rumblings and escapes of steam and gases. Sulphurous smells from local rivers, stinging acidic rain, loud rumblings or plumes of steam from the volcano are all warning signs. Evacuating by car remember: ash may make roads slippery, even if it does not block them. Avoid valley routes which could become the path of a lahar.

NUCLEAR EXPLOSION

Although a nuclear strike is unlikely, many countries have nuclear capabilities and must be regarded as a threat.

The immediate hazards of a nuclear explosion are blast, heat and radiation. The severity of their effects will depend upon the size and type of weapon, distance or height of the explosion, weather conditions and terrain.

Blast

The detonation causes the initial shock-wave. Even more powerful is the compression of the air produced by the rapid expansion of the fireball. The wave of pressure travelling outwards from the point of detonation will collapse buildings, uproot trees and fill the air with flying debris, well before the heat follows. Approximately half the total energy of the explosion is expended in this way.

When the blast wave has passed, air rushes back to fill the void causing further damage. At distances where the initial blast has only weakened structures this vacuum effect will finish the job.

Heat

The thermal radiation (heat and light) that is produced by a nuclear explosion reaches temperatures hotter than the sun. Close to the point of detonation all inflammable materials will be ignited – even vaporized. In the case of the Hiroshima bomb, exposed skin was burned at a distance of 4km (2½ miles). Today's weapons are MANY times more powerful than that bomb and their effects comparably more extensive.

Radioactivity

In addition to the thermal radiation, nuclear fission produces alpha and beta particles and gamma rays. Although radioactive fallout settles to earth, with the appearance of white ash or dust, this is the residue of destroyed matter and not the radioactivity itself. That cannot be detected by human senses. A Geiger counter is required to register its presence, indicated by a dial or a sound signal which becomes increasingly agitated as the radiation increases.

- **Alpha particles:** have low penetrating capabilities and it is easy to shield them off. They cannot penetrate the skin but they do present serious problems if ingested or inhaled.

- **Beta particles:** are only slightly penetrating and heavy clothing and boots will give full protection. On exposed skin they cause burns. If ingested they attack bone, the gastrointestinal tract, thyroid gland and other organs.
- **Gamma rays:** are highly penetrating. They travel much slower than alpha and beta rays, damaging all body cells.

Residual radiation

The initial radiation given off during the first minute of a nuclear explosion can kill – but it lasts only a short time. Once the blast has passed, so has the initial radiation threat. However, exposure to residual radiation can be equally dangerous.

SURVIVAL ADVICE

Common symptoms of exposure to radioactivity are nausea, vomiting, general weakness. Ulcer-like sores appear on the skin, which tends to take on a grey hue.

The amount of residual radiation depends on how the bomb was detonated. If it was high above the ground and the fireball did not touch the earth little residual radiation is produced – what the strategists call a 'clean bomb'. If exploded on or near the ground, a huge quantity of soil and debris is sucked upwards to a great height and falls back to earth as radioactive dust. Heavier particles fall in the vicinity of the explosion, but lighter ones may be carried by wind over a wide area – spreading the radioactivity.

Radiation does decay. However, while as much as 70 per cent of the particles following an explosion remain radioactive for only one day or less, it takes others years for their radiation to decay.

SURVIVAL ADVICE

The radioactivity to which an unprotected person could be exposed in the first few hours will exceed that received during the rest of the week. That in the first week will exceed that accumulated during the rest of a lifetime spent in the same contaminated area. It is therefore important to be shielded during the initial stages.

Radiation shelters

In default of a deep bunker equipped with air, water and food supplies, in which to sit out a nuclear conflict and its aftermath, the best protection is a deep trench with a roof covered by a metre or more of earth. If the detonation is sufficiently distant not to

SHIELDING

Strictly speaking, it is impossible to shield completely from all radiation but a sufficient thickness of shelter material will reduce the level of radiation to a negligible level. Below are some materials, and the thickness required to reduce radiation penetration by 50 per cent.

Material	Metres/Feet
Iron and steel	0.21/0.7
Concrete	0.66/2.2
Brick	0.60/2.0
Soil	1.00/3.3
Ice	2.00/6.6
Wood	2.60/8.8
Snow	6.00/20.3

produce total destruction, the trench and earth will protect from blast, heat and radiation.

Look for terrain that has natural shelter, such as ravines, gullies, ditches and rocky outcrops. If you do not have a trench shelter prepared, start digging – FAST! As soon as the hole is big enough, get inside it to continue digging, to minimize exposure to radiation if you are caught-out while still digging. Rig up a roof. Even if only of cloth, it will stop dust falling on you. Penetrating rays can still reach you so try to get a metre of earth above you.

If caught in the open get to your shelter as quickly as possible. Once under cover, remove outer garments and bury them under a foot of soil at one end of the bottom of the shelter. Do not venture out until absolutely necessary and do not re-use your discarded garments. Under no circumstances move out of the shelter in the first 48 hours.

If desperate for water a brief venture out, lasting not more than 30 minutes, is permissible on the third day. On the seventh day a further exposure, of up to half an hour, can be extended on the eighth up to one hour and then from two to four hours for the next four days and from the thirteenth day normal working hours, followed by rest in the shelter.

Decontamination

If your body, or even your clothing, has been exposed to radiation, it must be decontaminated. Once in shelter scrape earth from the shelter bottom and rub it over the exposed parts of your body and your outer clothing. Brush it off and throw the soil outside. Wipe the skin with a clean cloth if possible. More effectively, if water is available, wash the body thoroughly with soap and water instead of soil.

Medical care

ALL wounds must be covered to prevent alpha and beta particles entering through them. Burns, whether caused by beta particles and gamma rays or by firestorm heat, should be washed with clean water and covered. Urine may be used, if no uncontaminated water is available. The eyes should be covered to prevent further particles entering and a damp cloth placed over the mouth and nose to prevent further inhalation.

Radiation affects the blood and increases susceptibility to infection. Take all precautions – even against colds and respiratory infections.

AFTERMATH

Unless stored in deep shelters, or with special protection, all foodstuffs are likely to have absorbed some measure of radioactivity. Be cautious of foods containing a high salt content, dairy products, such as milk and cheese, and sea foods. After tests it was found that food with salt and other additives had a higher concentration of radioactivity than food without them. The safest canned foods are soups, vegetables and fruits. Cured and processed meat are more readily contaminated than fresh. Bone absorbs the highest levels of radioactivity, then lean meat, with fat lowest.

Water

Unless it is from a protected source, do not drink any water for at least 48 hours after detonation. Avoid water from lakes, pools, ponds and other static surface water. Filter all water and boil it before drinking.

The following sources are the least contaminated (in order of least risk):

1. Underground wells and springs
2. Water in underground pipes/containers
3. Snow taken from deep below the surface
4. Fast-flowing rivers

Dig a hole by a fast-flowing stream and allow water to filter down into it. Scrape off any scum that forms on the surface and scoop up water. Filter it through layers of sand and pebbles (dig deep to obtain these) in a can with holes punched in the bottom, or through a stocking. Boil in an uncontaminated vessel.

Decontaminate utensils by washing thoroughly in fast-flowing or boiled water.

Animals as food

Animals that live underground have less exposure to radiation than those that live on the surface: rabbits, badgers, voles and similar animals are the best bets but, when they venture out, they too will be contaminated. However, such food sources must be made use of. You will increase your own contamination – but the alternative may be to starve.

To reduce contamination from meat do NOT directly handle carcasses, wear gloves or use cloth to cover the hands while carefully skinning and washing. Avoid meat in direct contact with the bone. The skeleton retains 90 per cent of radiation so leave at least 3mm (⅛in) on the bone. Muscle and fat are the safest part of the meat. Discard ALL internal organs.

SURVIVAL ADVICE

Fish and aquatic animals will have a higher contamination than land animals from the same area. Birds will be particularly heavily contaminated and should not be eaten. However, eggs are safe to eat.

Plants as food

Root vegetables with edible tubers growing underground are safest – carrots, potatoes and turnips, for instance. Wash them well and peel before cooking.

Smooth-skinned fruits and vegetables are next safest. Plants with crinkly foliage are the hardest to decontaminate, because of their rough texture. They should be avoided.

Long-term survival

Predictions of the long-term results on the environment of major thermonuclear conflict differ widely. The possibility of a 'nuclear winter', with consequent effect on climate and plant life far beyond strike areas, would make even subsistence agriculture difficult. In the short-term, however, and in the case of limited conflict, the advice in *Home front* will be relevant.

HOME FRONT

You do not have to be miles away from civilization to be caught in a survival situation. Natural disaster, civil disturbance or military action could cut you off from all the usual services and food supplies. Until they can be re-established, you would be left to manage on your own resources and skills.

With no power supplies, central heating, hot water, lighting, air-conditioning and refrigeration would all cease. Battery radios and television would for a time give some news of the rest of the world, if the situation is not global, but post, telephone and newspapers would no longer be available. A generator is an essential back up and a short-wave radio is useful. As mains water supplies ceased to function, so taps would run dry and toilets would become unusable.

In the countryside there would be natural resources to draw upon. In large cities shops would soon be emptied of food – sold or looted – and plants in parks and gardens would be rapidly stripped, once any private stocks had been exhausted. The population would have to make forays out into the countryside to survive, or abandon the town, if not in a siege situation. Suburban dwellers have more vegetable plots and open spaces to provide foodstuffs. They would be less dependent upon shops. Those away from major centres are more likely to have their own food stocks, because they cannot shop at will.

Most families have some food in store. It should be rationed and supplemented with whatever can be found.

FOOD STORES

Storing food is a good habit to get into, especially if you live in an isolated place, which can become completely cut-off. If you have a year's food supply in store, and add to it as you use it, you will not only be able to survive the worst but will be able to live at last year's prices.

The stock does not have to be established in one go. Build it up gradually, taking advantage of special offers in supermarkets. Buy an extra tin or packet and put it by. Store your foods in a cool, dry, dark place and off the ground – moisture and heat cause bacteria and moulds. If stores are left on the floor insects and rodents will help themselves. Make sure that all containers are insect- and rodent-proof.

REMEMBER: Rotate cans, so that the contents do not settle, and separate. Label each can or packet with a colour-fast waterproof pen, noting contents and date of storage. Use in sequence – the oldest first. Store methodically and if a label falls off, you should still have a good idea of the contents.

Choice of foods will depend upon individual taste, but straightforward products (corned beef in preference to beef stew and dumplings) will keep better and can be used in a greater variety of ways. Wheat keeps better than flour – it is less susceptible to moisture, light, insects and temperature change. Wheat found in the pyramids was found in good condition after thousands of years. However, you must grind it to make flour, so invest in a small hand grinder.

Keep it sealed

Screw-top sweet jars are ideal for storage and plastic containers with tight-fitting lids can also be used. Do not over fill them so that they distort and the lid does not fit correctly. Use adhesive tape to seal the lids. Reseal after using some but remember that once opened the contents will begin to deteriorate.

RECOMMENDED FOODS/SHELF LIFE

Wheat	Indefinitely below 15°C (59°F)
Milk powder	2 years
Honey	Indefinitely
Egg powder	2 years
Salt	Indefinitely if absolutely dry
Canned foods	3–5 years (replace regularly)
Oats	Indefinitely
Cooking oil	2 years (replace regularly)

Rations

Complete rations are available with various menus – either freeze-dried or dehydrated. They are lighter and less space-consuming than canned foods. Freeze-dried are best for both taste and texture and retain minerals which are lost in dehydration. Although both need water for reconstitution they can, in dire circumstances, be eaten as a dry munch.

Vitamins

Multi-vitamin tablets are also a good investment. The body can store up to a month's supply of most vitamins, then health will suffer if they are not replaced. In stress situations they are more rapidly used up. The B family (and minerals, calcium and zinc) are the first to go. Vitamin tablets do not have unlimited shelf-life – check manufacturer's instructions.

Other foods

- **Dried fruit and nuts** are nutritious and should also be included – raisins, sultanas and currants all keep well. Nuts in their shells keep so long as they are dry. Packets of dried salted nuts such as peanuts, brazils and walnuts, are highly nutritious.

- **Potato powder** is a great filler for hungry stomachs and can be prepared in several ways to make it palatable.

- **Brown rice** has more nourishment than long-grain white rice which loses all its goodness when boiled.

Store location

The cooler the storage area, the better the stores will keep – a cellar is ideal but there may be a problem with dampness so keep all the stores off the ground and inspect them regularly. If there is a skylight in the cellar, cover it. The store is best kept dark.

An attic is also convenient for storage – the stores are not in the way of day-to-day activities. However, it may get very warm in summer and access may be difficult – especially if a ladder is the only means of entry – which may be awkward when trying to rotate bulky stores. The roof is also a vulnerable position in most kinds of disaster situations. In an area where hurricanes can be expected an attic is not a good choice. In territory liable to flooding a cellar would be equally at risk. Under the stairs is another area that may offer some protection, though perhaps limited space.

Advantage should be taken of wherever is most conveniently available to store not only food but also medical supplies, disinfectants, cleansing materials – and water. If you divide your stores into more than one area, each with a variety of items, you should be well prepared.

ADD TO YOUR STORES

Toothpaste and soap
Disinfectant and bleach
Washing powder
General medical supplies
Medicines: for dysentery, for stomach upsets, for allergies, general painkillers
Bandages and dressings
Spare batteries

PRIORITIES

In a domestic situation there is likely to be shelter, unless it has been totally destroyed, or the area has become a danger zone and evacuation is imperative. Damage can be patched-up to provide some protection from the elements and more permanent repairs undertaken as soon as possible.

Water supplies are always likely to be a problem – for even during a flood drinking water is scarce. Fortunately there are likely to be some immediate reserves on the premises and, with warning of a crisis, these can be supplemented.

Fire for warmth is less of a problem, since there will be burnable materials in the house and surroundings. Infection may prove the greatest danger and strict hygiene and sanitary practices must be enforced.

WATER

Although a family of four can use a considerable amount of water each week, only a small percentage of this is for drinking – a requirement of about 2 litres (3½pt) per day per person. If warned of a crisis, fill as many receptacles as possible, especially in a hot climate. A bath holds many gallons; increase its capacity by blocking the overflow. Use dustbins, buckets, pots – even strong polythene bags if they are only half filled and securely tied off.

Store water in the dark. If light gets to it green algae will develop. Water is bulky and heavy. Do not store it in the attic or it may bring the ceiling down.

Even without advance warning there will be water in the storage tank, heating pipes, radiators, perhaps an aquarium, and the toilet cistern will hold another few gallons – don't flush it. Outdoors you may have a swimming pool, water butts, or a pond – even water from a car radiator can be utilized. Central heating water is usually treated with a de-oxygenizing agent and a car radiator probably contains anti-freeze, so water from these places is best kept for cleaning purposes. If it has to be used for drinking boil it, collect the steam in clean cloths and wring them out. Then reboil. (Also see *Essentials*.)

Boiled water tastes flat and distilled water has even less taste. It is easy to restore some of its sparkle by putting oxygen back into it: simply pour the water back and forth from one vessel to another. A small piece of wood charcoal placed in the vessel while it boils also helps taste.

FILTERING AND STERILIZING

Filter and sterilize ALL water before using it for drinking. If circumstances make it impossible to boil water sterilize it with chemicals.

FILTERING: Allow water to stand in its container so that sediment settles at the bottom. Then siphon it into a filter made up of a nylon stocking (or other porous material) stuffed with layers of sand (bottom), charcoal and moss (top).

STERILIZING: Clear water: add 2 drops household bleach per litre (1 per pint) or 3 drops 2 per cent tincture of iodine per litre (6 per pint)

Cloudy water: double the quantities of bleach or iodine

Large quantities: half teaspoonful bleach per litre (2 teaspoonsful per gallon)

Cooking in water

Water in which food is to be cooked MUST be boiled for at least eight minutes, but water not boiled for as long can be used for heating cans of food provided it makes no contact with the foodstuff.

Stand the can in water, pierce a small hole in the top to avoid the risk of explosion and plug it with a twist of cloth so that water cannot enter the can.

Alternatively, boil the water, remove it from the heat and place the unpierced can in the water. This takes longer for the can to heat through.

Water catchment

- Catch all available rainwater. Break off lower sections of down pipes and divert the flow into a container such as a dustbin. Even if rainwater is pure, guttering may contaminate it – so sterilize.
- Supplement water receptacles with tarpaulins or plastic sheets supported on sticks. Rinse between showers to reduce tainting.
- Dig a hole and line it with a plastic sheet or concrete for water storage. Cover it to prevent evaporation and debris falling in.
- If the local water table is high you may be able to dig down to water – there may even be a well on your property which could be reopened.
- Solar and vegetation stills (see *Essentials*) are other ways of obtaining water.

Water conservation

Do not waste water washing clothes, other than underclothing. Never throw water away after use. Allow sediment to settle and it can be used again.

It is very important to wash the hands before preparing food, but the rest of the body can wait until it rains. The body produces natural oils and, as long as the pores are kept open, health will not be affected. You soon get used to the smell and social occasions are rare in a crisis situation. If showers are few and far between, use a damp cloth for a strip wash – cloths left out on lawns or bushes over night may gather enough moisture for a wipe down without using up your water stores.

Injured persons must receive priority for bathing and all their dressings should be boiled regularly.

FIRE

The warmth and comfort of a fire are great morale boosters, but its most important use will be for boiling water and preserving food. These must take priority in the use of fuel.

Blocked fireplaces should be opened up again and chimneys checked for obstructions. If they are not clear there is real risk of setting fire to the chimneys themselves and thus to the house.

To clear a chimney

Tie a holly bush or a similar shrub to a long rope and from the rooftop lower the rope down the chimney (a stone tied on the end will ensure it drops). Now pull down the holly bush and it will clear the chimney.

Improvised fireplaces

Where there are no fireplaces metal containers, metal dustbins lids and central heating radiators can all be used to light a fire on. In flats with concrete floors a fire could be lit directly on the floor. If you have a barbecue stand make good use of it.

Never leave a fire indoors unattended. Even one in a proper grate should be allowed to die down for the night, if no one is going to stay up to watch it.

Fuel

Start with garden furniture, trees, shrubs, bean sticks, swings, ladders, tool handles. When these run out start on furnishings. Carpets, curtains and cushions will all burn. Cardboard, books and rolled up newspapers will also give off a surprising amount of heat. All kinds of vehicle fuel can also be burned as well as the conventional heating and lighting oils.

WARNING

Many modern fabrics and furnishings, especially PVC and foam-block furniture, produce poisonous gases when burned. If burning these items make a fireplace in the garden or, if forced to burn them in a flat, make the fire near an open window. Cover the face with a damp cloth when you need to go near the fire to tend to it and things being heated on it.

FOOD

- Use the perishable foods first. Fatty foods are the first to deteriorate and canned foods the last.
- Remember that, once electric power fails, the refrigerator and freezer cease to function – though they may take some time to defrost if you open their doors as seldom and briefly as possible.
- Boil milk and it will keep longer.

- Cook meat, wrap it in cloth and bury it in the earth. Cook pork first (which has the highest fat content), then lamb, then beef (which is the best meat to preserve).

- Once meat has been cooked and allowed to cool, do NOT reheat it or you may risk food-poisoning.

SURVIVAL SCENARIO

What should you do in a country where the threat of kidnap is high?

In Columbia, for example, where the threat of kidnap is high, the best thing to do is be a 'grey' person; do not be the first, or last person to do anything, so that you don't stand out. If there's a rescue attempt, get down and take cover.

Food from the garden

Vegetables with four petals, including all the brassicas, from wallflowers to cabbages are edible. Hollyhocks, though not very tasty, are nutritious. Worms, slugs and snails are also edible. AVOID bulbs such as daffodils, tulips and aconites which are all poisonous.

Further afield

Explore parks and open spaces for other vegetation and for hunting and trapping wildlife. Birdlife in cities – especially pigeons and starlings, will often fill the plate, especially if you bait snares and nets. (See *Traps and snares* in *Food.*)

Closer to home

Beware of houseplants, some are poisonous, especially the diefenbachia and philodendron.

SURVIVAL TIP

If food is short there will be none to spare for pets and you CANNOT afford to be squeamish. If the aquarium water has to be drunk don't waste the fish. In fact they'll probably be the easiest to eat even if you don't need the water. The cat is next in the pot. Once dressed it will be hard to distinguish from rabbit. Gerbils, hamsters, rabbits, budgerigars and parrots can all be added to the diet and, unless the dog is an exceptionally good hunter, it should go too.

Preserving food

For methods of smoking, salting and making pickles and chutneys see *Food preservation* in *Food*.

When the fridge no longer functions remove the motor, cut a hole in the bottom, place it on some stones or bricks and with a fire beneath use it as a smoke-house.

SHELTER

The first priorities will be a sound roof over your head and a stable structure. Clear any debris and ensure that there is nothing which could still collapse or fall from above and cause injury. Use slates, tiles and bricks from other buildings to ensure that at least one building is sound.

In cold weather

Conserve resources by living in one room, choosing a ground-floor room with a southern aspect (if you live in the Northern Hemisphere). Block all draughts and avoid opening the door unnecessarily.

If there is a fire burning, make sure that there is adequate ventilation to avoid asphyxiation or carbon monoxide poisoning. Wear warm clothing to help conserve fuel. The more people in the room, the higher the temperature. Rest and keep physical exertion to a minimum.

In very warm weather

Use upstairs accommodation and spread out. Open windows on the downstairs windward side and open all windows on the lee side upstairs. Leave all the doors open and a cool breeze will blow through the house. It is best to rest during the day to conserve energy and fluids, and do any necessary work at night.

Moving

If the house proves beyond repair, or other pressures force you to evacuate, take essential items – food, blankets, tools, medical supplies, containers for water and materials suitable for shelter construction – if they are not likely to be available. Use a pram or a shopping trolley as transportation. Either find an empty house or building, or prepare to set up camp in a location elsewhere.

HYGIENE

Sanitation is very important during the aftermath of any disaster. Open sewers, contaminated water and the build up of rubbish all help to cause and spread disease. Germs carried by rats, fleas and other insects, rapidly multiply. All kinds of waste should be carefully disposed of and all the procedures described (see *Hygiene* in *Camp Craft*) should be adapted to the doorstep situation.

Excreta

Urine is sterile but if large amounts accumulate they smell and attract flies. Use the 'desert rose' urinal, of the kind described in *Camp Craft*. Keep the tube covered. If not used directly, pour all collected urine down the tube.

Build a latrine (again see *Camp Craft*), far enough from the house not to be smelt but near enough to be handy for 'emergencies' – there will be many such emergencies in a survival situation. A box with a hole cut in the base can be used as a thunder box. After use, if there is water available, wash yourself rather than using toilet paper. Wash the hands thoroughly afterwards.

Fit a lid to your thunder box, pile earth around the bottom and then you will contain the smells and keep out flies.

Move all muck with a shovel and avoid hand contact.

Animals

Animals pick up diseases which can be transmitted to humans. If you handle animals, make sure you have no breaks in the skin – or wear gloves. Infection can enter through the smallest of cuts. Cook all meat thoroughly.

Kitchen waste

All biodegradable waste should be stacked in a corner of the garden and composted to enrich the soil. Compost heaps are also a great source of worms, which will add protein to your diet.

However, there should not be much kitchen waste. Do NOT peel potatoes, much of their food value is in the skin. The outer leaves of cabbages which you once discarded, will be edible if you cut them up small.

Non-biodegradable waste – cans and plastics that are not useful in some way – should be burned, flattened and buried. This stops them attracting flies. In warm climates burn ALL waste. Put all the ashes in a pit.

FOOD DISEASES

Salmonella and shigella are diseases transmitted through the oral-anal route, by contaminated hands.

Sores on hands can be a source of entry for staphylococcal food-poisoning with severe stomach pains, diarrhoea and dehydration.

Clostridium botulinum, is a frequently fatal bacillus, which can be produced when canning at home if the temperatures are not high enough – it grows only when oxygen is excluded. There is no reliable way of determining whether food is contaminated so TAKE GREAT CARE if you do your own preserving. A related bacillus causes tetanus.

Communicable diseases

Living in close-knit groups after a disaster increases the risk of passing on disease. Good personal hygiene – as good as possible – can reduce the threat. Isolation of patients with colds or fever is advisable.

Seal dressings and discharges in a polythene bag and burn immediately. Dispose of all faeces and urine in the field latrine – and regularly boil the container used for their disposal.

Personal hygiene

Wash with sand if there is no water available. Don't bite your nails – however stressful conditions may be – or put the fingers to the mouth. Don't pick scabs or sores and keep them covered. Change underclothes regularly and wash them (but don't use drinking water to do so).

SOME USEFUL HERBAL PREPARATIONS

Strawberry roots contain a descaler to clean teeth.

Delphinium seeds can be crushed to treat head lice.

Birch bark can be distilled to produce a tar oil which soothes skin complaints.

Lavender makes a decoction to clean the skin.

(See also *Natural Medicine*.)

POSTSCRIPT

Now you have the benefit of knowledge that it has taken me, and others like me, a lifetime of training and experience to gather. But don't think that just reading this book will have made you a survivor. It has shown you the necessary skills, but it is you who must apply them and you who must have the calibre to cope.

When I am personally teaching soldiers or civilians how to deal with survival situations, part of my job is to ensure their safety. I cannot do that for the reader of this book. I can only give information and advice. I am not there to stop you doing something foolish. I cannot deal with individual specific situations, nor can I ensure that what I have written has been properly understood. I do know that what I have written has saved lives in the past and I believe that it could save more in the future. You must apply survival techniques with caution, for it will be your responsibility – not mine – if you inflict injury on yourself or on others.

The human body has an amazing ability to cope with arduous situations and testing environments. People who have come through, after enduring terrible hardship under seemingly impossible conditions, are living proof of this. Male and female, young and old, they have all had the will to live. Everyone has this basic instinct to some degree and it can be developed by training.

Survival is as much a mental attitude as physical endurance and knowledge. Think of survival skills as a pyramid, built on the foundation of that will to survive. People with it have survived even though they did everything against the rule book. With a little knowledge they could have made their lot much easier. So the next layer of the pyramid is knowledge. It breeds confidence and dispels fears.

The third layer must be training, that does not just mean trying something once but mastering skills and maintaining them. In doing so, you will be keeping your body in training, too.

To cap the pyramid, add your kit. Proper equipment and provisions are common sense, but the survivor does not necessarily know what conditions to equip for. That is where your survival tin will make a tremendous difference to your chances. To the instinct for survival, which you can further develop, add knowledge, training and kit and you will be ready for anything.

INDEX

abandoning ship 484
abseiling 79
aerial search 521
air crash
  crash landing 55-57
  crash site 57-59
  ditching into water 56
  navigation after 60
  signalling from wreckage 506
  survivors 61
aircraft
  emergency procedures 26-27
  fires 539
algae 186-187
alligators 217
altimeter 26
Amanitas 161, 164
angling 229-234
animals
  dangers 135, 216-217
  diseased 219
  products 316-319
  water from 48-49
anthrax 541
antiseptics 413, 416
ants 133-134
apes 217
arid regions 94-99
artificial airway 407
artificial respiration 397, 402-404, 406
avalanches 82-83
axes 307-310

baby bag 377
back-pack 21-22, 376
bait 189, 204-206, 231-232
baking 289
bamboo 179, 257, 287
bark 157-158
battery fire-lighting 279
bears 74, 122-123, 217
beds 312-314
bees 133
Big Dipper, The 362
Biltong 296
biological warfare 540-541
bird traps 204-207
birds
  as food at sea 495-496
  edible 130-131, 227
  hunting 204, 215-216
blood from animals 219-220
boats 26-27
body signals 516-517
bogs 386
boiling 207, 269, 287
bola 214
bone tools 307
bow and arrow 203, 209-212
brake failure 53-54
buildings burning 535-538
camp
  discipline 305
  furnishing 312-215
  hygiene 303-305
  layout 303
  making 246-247
  organization 302-305
  position 246-247
candiru 105
candles 277-278
cane splitting 324
cardiac compression 405-406
carrying people 377
cars see vehicles
Casiope 72
Cassiopeia 362
catapult 213
caves 264-265
cervical collar 424
chemical warfare 540-541
chemicals 541, 542-543
childbirth 438-440, 455
clay for cooking 290-291
cliff descent 77-78
climate and terrain 62-107
climbing 78, 79-80, 527
clothing
  arid regions 97-98
  improvising 318-319
  planning 19-21
  polar regions 70-71
clouds 366-369
coastal areas
weather 365
coastal flooding 546
coconuts 93-94
codes 508-519
  ground-to-air 512
  Morse 515
  mountain rescue 517-519
COLD mnemonic 71
Comfrey 144-145
compass improvisation 357
condensation 43-45
contingency plan 26
contour lines 351-352
contour search 522
cooking 286-301
  basics 286
  tips 298-301
  utensils 292-293
coracle 391
countryside interpretation 350
cramp 49, 449
creeping line 521
crevasses 81
Crico thyroid technique 398-399
crocodiles 217
crustaceans 89-90
cyclone 547

dangerous creatures 468-469
dead fish 239
deadfall traps 197-200
deer 125
dehydration 67, 449-451
descent by rope 78-79
desert
  plants 170-171
  sores 99
dinghy 56, 486, 507
direction finding 355-364
  basics 355
  compass improvisation 357-358
  stars 360-364
  watch or clock 356-357
disaster
  air 55-61
  awareness 52
  basic needs 38-39
  survival 38-39
  water 40-49
distillation 45, 93
distress signal 58
dropping from height 537
droppings 120, 275
drought 530-531
  fire risk 530
  water 531
  drowning 401

earthquake 546, 551-553
  aftermath 553
  indoors 552
  instant action 553
  outdoors 553
echinoderms 89
electrical fires 535
electrocution 401
emergency plans 19, 25
equipment 19-27
essential kit 10-11, 27-31
evacuation from flood 545
explosives for fishing 239

falling onto slope 538
fibres
  for ropes 323-324
  strength testing 324
fire 269-285
  bow 280
  drought 530
  escaping through 538
  evacuation 536
  fight with fire 534
  home front 566-567
  lighting practice 269
  lighting with chemicals 281-282
  lighting with kindling 277
  polar regions 72
  speed of flames 535
  types 282-285
  vehicle 538-539
  warmth 282-283
fire extinguishers 535, 539

fire signals 60, 506, 508-511
fire triangle mnemonic 270
fireman's lift 434-435
first aid 394-455
  amputation 415-416
  arterial bleeding 408-411
  artificial respiration 397, 402-404, 406
  asphyxiation prevention 400
  bandaging 427-429
  basics 394
  bites 440-441
  bleeding dangerously 407-411
  bleeding internal 412
  bleeding (lesser) 411
  blockages 396-400
  breathing and pulse 396
  breathing but unconscious 395-396
  burns 416-418
  cardiac compression 405-406
  childbirth 438-440
  choking 396-400
  cold climate 451-455
  digestive upsets 443
  diseases 444-450
  dislocations 425
  dressings 413-416
  equipment 407, 411
  fractures 418-424
  heart attack 401
  heart not beating 405-407
  immobilization 419-424
  minor ailments 429-430
  moving injured 432-438
  natural medicine 456-461
  natural medicine remedies 458-461
  nose bleed 412
  not breathing and no pulse 401
  open treatment 414-416
  pain-killers 431
  poisoning 440-443
  pressure points 408-409
  priorities 394
  recovery position 396, 404
  removal from danger 395
  shock from burns 418
  shock treatment 426
  sprains 425
  stitching wounds 414
  tracheotomy 398-399
  tourniquets 410
  tying arteries 409
  warm climate 446-452
  wounds 413-416
fish and fishing 228-243
fish narcotics 239-240
fish preparation 243
fishing
  at sea 495
  fish traps 234-237
  knots 346-347
  seashore 88
flares 518-519
flash floods 546
flexible saw 311
flint and steel 279
flint weapons 212
flood 207, 544-546
flotation 390, 485
flowers 141-142
food 544-546
  animals for 118-135, 218-227
  arid regions 98-99
  at sea 494-496
  diseases 571
  energy needs 110-112, 118
  home front 567-569
  nuclear explosion aftermath 560
  plant gathering 114-115
  plants 11, 112-117
  polar regions 73-76
  preservation 294-297
  seashore 87-90
  shelf life 562
  stores 527, 561-564
  testing plants 113-114
  tropical regions 103
footwear 318-319, 454
forest
  deciduous 65
  equatorial rain 100-101
  fires 532-535
  montane 101-102
  Northern coniferous 64-65
  tropical 8, 65-66
frostbite 76, 454
fruit
  cooking 297
  edible 136, 148-151, 180-181
frying 290
fuel for fires 32, 273-276
fungi
  cooking 297
  edible 160-167
  gathering 115
  other uses 165
furs 316-318

game 118-121, 190, 208
gases 542-543
Global Positioning System [GPS] 22-23
goggles 319
grain heads 115
grilling 288
groups
  crossing rivers 388
  fitness 17
  moving 379-380, 527
  pace 380-381
  planning 8, 18, 527
  radio discipline 24
gutting 222-223

haemostat 411
hammock 313-314, 345
hand drill 281
Hangi method 291-292
hanging meat 224
health
  arctic 8, 76
  arid regions 99
  checks 17
  warm climate 446-450
  and *see* first aid
heat-related illnesses 449-453
Heimlich manoeuvre 397, 398
helicopter
  landing site 523-524
  non-landing rescue 525
  precautions on landing 526
  rescue 523-527
  static electricity 527
  winching techniques 526-527
heliograph 513-514
herbs 143, 571
hitches 336-338
hobo stove 285
Holger Neilson method 404
home front 561-571
  basics 561
  fire 566-567
  food 567-569
  food diseases 571
  food stores 561-564
  hygiene 570-571
  pets 568
  priorities 564
  shelter 569
  water 564-566
hook and line 229-234, 241
Hudson Bay pack 376
hunting 208 -217
  handling the kill 218-227
  preparing the kill 219-224
hurricane 547-548

ice netting 242
ice water 45
icebergs 69
ice-cliffs 69
icefields 80-81
igloo 260-261
image refraction 229
imaginative signals 508
incinerator 305
information 17-18, 391
insects 446-449
  coloured 132
  edible 131-134
instant action, earthquake 553
instinctive reaction 52
islands 92-94

jelly making 151
jellyfish 91
jointing meat 223
jumping from fires 537
jungle travel 383

kindling 272-273, 277
kiss of life 402

kit
  essential items 10, 28-31
  stowing 22
  survival 10-11, 27-34
  survival pouch 32-34
knives 12, 34-37
knots 327-347
  basics 327
  fishing 346-347
  hitches 336-338
  ladders 334
  lashings 317-318, 341-342
  loop-making 331-335
  net-making 343-345
  reef 327
  rope joining 329-331
  rope shortening 338-339
  securing loads 339-340
  simple 328

laces 317-318
ladder 315, 334
land indicators 489-490
lashings 317-318, 341-342
latrines 303, 304, 570
lean-to shelter 254-255, 265
leather-making 317
leaves 136-141
leeches 105
lichens
  cooking 297
  edible 169
lie of the land 350
lifeboat drill 483-484
liferafts 507
lifting person 433-438
lightning strike 401
lightning (weather) 550
liming 204
lines 320-326
loads, securing 339-340
log cabin 268-269
loop-making 331-335
loose rock 80

mammals
  confrontations 135
  for food 122-129
maps 23, 350-355
marshes 386
matches 277-278
metal box oven 289
mnemonics
  COLD 71
  fire triangle 270
  PLAN 15, 60
  RSVP 24
mobile phone 25
moon for direction 360
morphine 431
Morse code 515
mosquitoes 64, 105, 447-448
mountains 77-83
  altimeter 26
  rescue code 517-519
moving 374-391
  decision 374-375
  jungle 383
  planning 378-379
  upland 382-383
  waterways 384-391
mucus extractor 407
mud slide 83
mudflow 555

natural immunity 455
net-making 343-345
night lines for fish 233
night signals 513
night walking 381
noise 508
North
  magnetic 354
  maps 353-354
  northern sky 361-363
nuclear explosion 556-560
  aftermath 559-560
  basics 556
  decontamination 558-559
  radioactivity 556-557
  residual radiation 557
  shelters 557-558
nuts
  cooking 297
  edible 152-153, 180

octopus 88-89
offal 224-226
oil, burning 276
Orion 362
otter board 233

packs 21-22
palms 94, 172-173, 180
parachute 56
parachute snow house 262-263
parachute tepee 252
parang 36-37
pawpaw 173
peat 275
Pemmican 296-297
pets 568
pickling 297
pig preparation 226
PLAN mnemonic 15, 60
planes 26-27, 391
planning 18, 26, 391, 527
plants
  edible 112-117, 136-143
  identification 117
  pointers for direction 359
  tropical medicinal 461-467
  water from 46-48
Plough, The 362
pneumonia 443
poison
  Amanitas 161, 164
  berries 156
  carbon monoxide 454
  contact poisons 154
  first aid 401, 440-443
  fish at sea 497
  fish narcotics 239-240
  freshwater algae 187
  fungi 160, 161, 164, 166-167
  ingestion 154-156
  plants 73, 116, 154-156
  plants for safe fishing 239-240
  snakes 470-475
  Strychnine 179
  trees 159
  tropical plants 182-183
polar bears 75
polar regions 64, 68-76
  fishing in the Arctic 241-242
  plants 168-169
  shelter 258-264
pouch, survival 32-34
prearranged contact 19
preparation
  essential 16-17
  unexpected events 27
pressure points 408-409
prickly heat 99, 450

quicksand 386

rabbits 126-127
radio 10, 23-25
rafts 384-386
rag signals 514
railway
  car on tracks 55
  tankers 542-543
Rattan 175
recovery position 396, 404
reef knot 327
reptiles, edible 129-130, 227
rescue, attracting attention 9, 94, 101
research 17-18
rivers
  crossing 387-390
  following 378
road tankers 542-543
roasting 288
rodents 127
roots 136-141, 144-147
rope 320-326
  care 321-322
  choosing 320
  joining 329-331
  making 323-326
  plaiting 325
  shortening 338-339
  spinning 325-326
  throwing 322-323
  whipping 326
RSVP mnemonic 24
rubbish disposal 304
rushes for shelter 266

salt 49
salting 297
savannah 66
scurvy 444
sea - power and danger 85
sea ice 45, 493
sea urchins 89
sea water 44
sea, survival at 482-503
  afloat 487
  basics 482-484

  dangerous sea creatures 476-479
  fish and dangers 497
  food 494-496
  health 492
  land indicators 489-490
  landfall 503
  lookouts 489
  man overboard 484-486
  priorities 487
  protection 488-489
  sharks 498-502
  signalling 491
  travelling 490-491
  water 492-494
seabirds 90
seafood 87-88
seagulls 216
seals 74-75, 124
searches
  patterns 520-523
  procedures 520
seashore 84-92
  dangers 91-92
  plants 184-185
seats 315
seaweed 87, 186-187, 496
shadow stick method 356
shadow tip method 69
sharks 92, 498-502
sharpening knives 37
sheep-like animal preparation 226
shelter 246-269
  long-term 264-269
  polar regions 71, 258-264
  sheet 251
  tropical 255-257
  types 247-254
shooting fish 238
signalling 60, 506-519
  at sea 491
  body signals 516-517
  distress 58
  ground-to-air 511-513
  informative signals 519
  messages 513-517
signals and codes 508-519
skinning 221-222
skins 316-318
sky signs for weather 371
sledge 377
sleeping bag 21
sling and shot 213
smoke drying 294-295
smoke signals 506, 508-509, 511
snails 134, 207
snake hole fire 283
snakes
  bites 440-441
  dangers 217
  poisonous 470-475
snares 190-197
snow
  blindness 454-455
  buildings 259
  glare 71
  house 260-264
  shoes 69
  trench 259
  water 45
soap 305, 413
solar still 44
Southern sky 363-364
Southern Cross 364
spear traps 201-205
spearing fish 238
spears 214-215
spontaneous combustion 285
spring snares 192-194
spruce tea 158
square search 522
squid 88-89
star fire 274
stars for direction finding 360-364
steaming 290
steep slopes 382-383
stems 136-141
stick walls 253
stings 104-105
stone tools 306
stowing kit 22
stretcher 433
survival – basic needs 9, 38-39
survival kit 28-31, 483
survival pouch 27, 32-34
survival stresses 38
swamps 102-103
swimming 86, 390, 503

temple fire 271-272
tepee 252
terrain 350
terrorist attack 540-543, 568
thermal blanket 34, 452
tickling fish 237
tides 85
tinder 272
tools 306-311
torch trees 509
tornado 549
toxins 11, 540-541
track crawl 522
transmitters 507
traps - dangers 11, 189, 197, 201, 204, 206-207
traps and trapping 188-207
travelling 374-391
and see moving
travois 315
trees
  felling 309-310
  for food 157-159
trench fire 285
trench foot 455
triangular bandage 427
tropical regions 100-105
  food deterioration 177
  plants 172-183
  seashore plants 185
  shelter 255-257
tsunami 546
tubers 144-147
tundra 64, 73
turtles 90
typhoon 547

unexpected events 27
urine 44

vaccination 17
vegetables 181
vehicles
  accident survival 53-55
  basics 106-107
  collision 53-54
  extreme conditions 26
  fire 538-539
  jumping from 54
  under water 54
Very pistol 518-519
volcano 554-555

water 40-49, 60, 67
  acquiring 59-60
  aircraft ditching 56
  animal signs 42
  arid regions 95-96
  car under 54
  catchment 566
  coastal fresh 86-87
  cooking in 565
  drought 531
  filtering 565
  finding 41-43
  freshwater 268, 493
  home front 564-566
  ice cold 70, 388
  islands 93
  nuclear explosion aftermath 559-560
  plants 46-48
  polar regions 72
  rationing at sea 493
  sea water 44
  sea, survival at 482-503
  sources 18
  sterilizing 565
waterproofing 318
weapons for hunting 209-215
weather 207, 364-371
  clouds 366-369
  hurricane 547-548
  lightning 550
  predictions 370-371
  tornado 549
  winds 365
wild pigs 126
wildcats 122
will to live 9-10
willy-willy 547
winching techniques 526-527
wind
  hurricane 547-548
  studying 365
  use for direction 359-360
wolves 74, 217
worms 112, 134, 446

Yukon stove 284

USER ADVICE
this book could save your life

THE SURVIVAL HANDBOOK

ESSENTIAL SKILLS FOR OUTDOOR ADVENTURE

THE SURVIVAL HANDBOOK

THE SURVIVAL HANDBOOK

ESSENTIAL SKILLS FOR OUTDOOR ADVENTURE

DK

LONDON, NEW YORK, MUNICH, MELBOURNE, DELHI

Senior Editors Nicky Munro, Richard Gilbert

Senior Art Editors Michael Duffy, Gillian Andrews

Editors Bob Bridle, Chris Hawkes,
Philip Morgan, Gill Pitts, Clare Weber

US Editor Chuck Wills

Designers Sarah-Anne Arnold, Katie Eke,
Brian Flynn, Phil Gamble, Sharon Spencer

Lead Illustrator Mike Garland
Assisting Illustrators Darren Awuah,
Phil Gamble, Peter Liddiard, Mark Walker

Production Editor Tony Phipps
Production Controller Norma Weir

Managing Editor Stephanie Farrow
Managing Art Editor Lee Griffiths

Art Director Bryn Walls
Publisher Jonathan Metcalf

First American Edition, 2009

Published in the United States by
DK Publishing, 375 Hudson Street
New York, New York 10014

09 10 11 12 10 9 8 7 6 5 4 3 2 1

SD308—April 09

Copyright © 2009 Dorling Kindersley Limited
All rights reserved

Without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording, or otherwise), without the prior written permission of both the copyright owner and the above publisher of this book.

Published in Great Britain by Dorling Kindersley Limited.

A catalog record for this book is available from the Library of Congress.

ISBN 978-07566-4279-2

DK books are available at special discounts when purchased in bulk for sales promotions, premiums, fund-raising, or educational use. For details, contact: DK Publishing Special Markets, 375 Hudson Street, New York, New York 10014 or SpecialSales@dk.com.

Printed and bound in Hong Kong by Hung Hing

Discover more at
www.dk.com

IMPORTANT NOTICE

Some of the techniques desribed in this book should be used only in dire emergencies, when the survival of individuals depends upon them. The publisher cannot be held responsible for any injuries, damage, loss or prosecutions resulting from the use or misuse of the information in this book. Do not practice these techniques on private land without the owner's permission, and obey all laws relating to the protection of land, property, plants, and animals.

CONTENTS

8 Introduction

1 BEFORE YOU GO

16 **Introduction: Prepare yourself**
18 Getting into shape
20 Mental preparation
22 Planning your journey
24 Emergency plan of action
26 **Introduction: Know your environment**
28 Temperate environments
30 Tropical environments
32 Mountain environments
34 Desert environments
36 Cold environments
38 Marine environments
40 **Introduction: Get the right gear**
42 Choosing your kit
44 Choosing a backpack
46 Dressing for the outdoors
50 Extreme survival—in the wilderness
52 Sleeping systems
56 Eating on the trail
58 Camping stoves
60 Your survival kit

2 ON THE TRAIL

64 **Introduction: Find your way**
66 Maps and map-reading
68 Get your bearings
72 Route finding
74 Navigating without a compass
76 Natural navigation
78 How weather works
80 Weather phenomena
82 Understanding local weather
84 **Introduction: Make a move**
86 Traveling on foot
88 Crossing rivers
90 Scrambling and climbing
92 Extreme survival—in the mountains
94 Moving over snow
98 Using pack animals
100 Four-wheel driving
104 Kayaking and canoeing
106 Building a raft
108 Swimming

3 CAMP CRAFT

112 **Introduction: Camp essentials**
114 Organizing your site
116 Staying clean
118 Making fire
120 The elements of fire
122 Char cloth and fire can
124 Types of fire
126 Making sparks and flames
128 Extreme survival—in the desert
130 Fire by friction
134 **Introduction: Manual skills**
136 Man-made cordage
138 Natural cordage
142 Tying knots
146 Using cutting tools

4 TAKING SHELTER

154 **Introduction: Taking shelter**
156 Hollows and caves
158 Use-anywhere shelters
162 Forest shelters
166 Tropical shelters
172 Extreme survival—in the jungle
174 Desert shelters
176 Shoreline shelters
178 Snow shelters

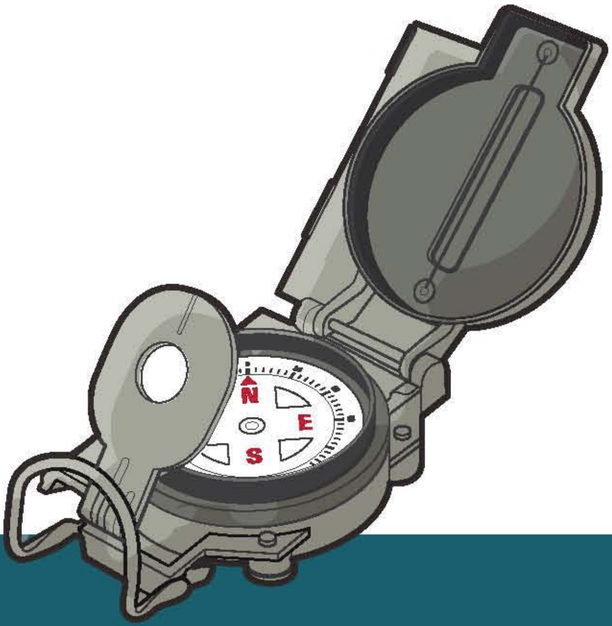

5 WATER AND FOOD

184 **Introduction: Find and treat water**
186 The importance of water
188 Finding water: Temperate climates
192 Finding water: Hot climates
194 Finding water: Cold climates
196 Finding water: At sea
198 Carrying and storing water
200 Treating water
202 **Introduction: Find and prepare food**
204 Wild cooking
206 Edible plants
208 Catching fish
212 Preparing fish
214 Extreme survival—adrift at sea
216 Trapping animals
220 Preparing small mammals
222 Preparing large mammals
224 Preparing other animals
226 Catching birds
228 Preparing birds

6 IN AN EMERGENCY

232 **Introduction: In an emergency**
234 Assessing your situation
236 Attracting rescuers
242 Wild animals
244 Extreme survival—in cold conditions
246 Environmental hazards
250 Surviving at sea

7 FIRST AID

258 **Introduction: First aid**
260 First aid essentials
262 Flesh wounds
266 Bites and stings
268 Poisonous plants and intestinal parasites
270 Bones, joints, and muscles
272 Treating exposure
274 Treating shock
275 Breathing difficulties
276 Unconsciousness
278 Moving a casualty

APPENDIX

Wild food

280 Temperate plants
282 Desert plants
284 Tropical plants
286 Plants in cold climates
288 Coastal plants
290 Mammals
292 Birds
294 Amphibians and reptiles
296 Fish
298 Invertebrates

Natural Dangers

300 Wild animals
302 Insects and arachnids
304 Poisonous plants

306 Resources
308 Glossary
312 Index
319 About the author
319 Acknowledgements

INTRODUCTION

HAVING TAUGHT SURVIVAL SKILLS FOR MANY YEARS, I have learned that four elements must be in place for a survival situation to have the chance of a positive outcome: knowledge, ability, the will to survive, and luck. While knowledge and ability can be learned, the will to survive is hard-wired into our survival mechanism and we may not know we possess it until we're put to the test. For example, people who were fully trained and well-equipped have given up hope in survivable conditions, while others, who were less well-prepared and ill-equipped, have survived against all odds because they refused to give up.

> “Always apply the **principle** of the **least** amount of **energy expended** for the **maximum** amount of **gain**”

Anyone venturing into the wilderness—whether for an overnight camping trip or a lengthy expedition—should understand the basic principles of survival. Knowing how to survive in a particular situation will allow you to carry out the correct beforehand preparation, choose the right equipment (and learn how to use it), and practice the necessary skills. While you may be able to start a fire using a lighter, for example, what would you do if it stopped working? Equally, anyone can spend a comfortable night inside a one-man bivy shelter, but what would you do if you lost your pack? The knowledge gained through learning the skills of survival will enable you to assess your situation, prioritize your needs, and improvise any items of gear that you don’t have with you.

> “Treat the **wilderness** with **respect**: carry in only what you can **carry out**; leave only **footprints**, take only **pictures**”

Survival knowledge and skills must be learned—and practiced—under realistic conditions. Starting a fire with dry materials on a sunny day, for example, will teach you very little. The real survival skill is in understanding why a fire won’t start and working out a solution. The more you practice, the more you learn (I am yet to teach a course where I didn’t learn something new from one of my students). Finding solutions and overcoming problems continually adds to your knowledge and, in most cases, will help you deal with problems should they occur again.

There are differences between teaching survival courses to civilians and teaching them to military personnel. Civilians have enrolled on (and paid for) a course to increase their knowledge and skills, not because their life may depend on it (although, should they find themselves in a life-threatening situation, it may well do), but because they are interested in survival techniques in their own right. In contrast, the majority of military personnel who undergo survival training may very well need to put it into practice, but they invariably complete the training simply because they are required to do so. While no one in the military forces would underestimate the importance of survival training, it is a fact that, if you want to fly a Harrier, or become a US Marine Mountain Leader, survival training is just one of the many courses you must undertake.

In the military, we categorize the four basic principles of survival as protection, location, water, and food. Protection focuses on your ability to prevent further injury and defend yourself against nature and the elements. Location refers to the importance of helping others to rescue you by letting them know where you are. The principle of water focuses on making sure that, even in the short term, your body has the water it needs to enable you to accomplish the first two principles. Food, while not a priority in the short term, becomes more important the longer your situation lasts. We teach the principles in this order, but their priority can change depending on the environment, the condition of the survivor, and the situation in which the survivor finds him- or herself.

> “Understanding your **environment** will allow you to select the **best equipment**, adopt the **best techniques**, and learn the **correct skills**”

We also teach advanced survival techniques to selected personnel who may become isolated from their own forces, such as when operating behind enemy lines. The four principles of survival remain the same, but we substitute “location” with “evasion”. The military definition of evasion is recognized as: “being able to live off the land while remaining undetected by the enemy”. This involves learning how to build a shelter that cannot be seen, how to maintain a fire that doesn’t give away your position, and how to let your own forces know where you are but remain undetected by the enemy.

In military training, and with most expeditions, the equipment with which you train will be specific to a particular environment—marines operating in the jungles of Belize will not pack a set of cold-weather clothing, for example; and Sir Ranulph Fiennes won't practice putting up his jungle hammock before venturing into the Arctic! However, the standard practice of being equipped and trained for a specific environment can prove to be a major challenge for some expeditions. During my career as a survival instructor, for example, I have been fortunate enough to have worked on two of Sir Richard Branson's global circumnavigation balloon challenges with Per Lindstrand and the late Steve Fossett. For these expeditions, the responsibility for selecting the survival equipment and training the pilots was a unique, if daunting, task. The balloon would be flying at up to 30,000 ft (9,000 m) and would potentially cross every type of environment: temperate, desert, tropical rain forest, jungle, and open ocean. While it would have taken some very strong winds to blow the balloon into the polar regions, we did fly—after a brief and unplanned excursion into China—across the Himalayas.

> “The more you understand **how and why** something works, the more prepared you will be to **adapt and improvise** should it be **damaged or lost**”

We also had to train for the worst-case scenario, which would be a fire in the balloon capsule. A capsule fire would leave the three pilots no option but to bail out, potentially from a great height, breathing from an oxygen cylinder, at night, and anywhere in the world, whether over land or sea. The chances of them landing in the same vicinity as each other under such circumstances would be slim to non-existent, so each pilot would need not only the necessary equipment to address the priorities of survival in each environment, but also the knowledge to be able to use it confidently and alone. We addressed this challenge by providing each pilot with survival packs devised for specific environments, a single-man liferaft (which provides shelter that's just as good in a desert as it is at sea) and realistic training with the equipment contained in each pack. As the balloon moved from one environment to another, the packs were rotated accordingly, and the pilots re-briefed on their survival priorities for each environment.

> “There is a **thin line** between **being in awe** of an environment and **being at its mercy**”

As you read this book and plan to put the skills and techniques covered here into practice, you will typically be equipping yourself for just one particular type of environment—but it's important that you fully understand that one environment. Make sure you research not only what the environment has to offer you as a traveller—so that you can better appreciate it—but also what it offers you as a survivor: there is sometimes a very thin line between being in awe of the beauty of an environment and being at its mercy. The more you understand both the appeal and dangers of an environment, the better informed you will be to select the right equipment and understand how best to utilize it should the need arise.

Remember, no matter how good your survival equipment, or how extensive your knowledge and skills, never underestimate the power of nature. If things aren't going as planned, never hesitate to stop and re-assess your situation and priorities, and never be afraid to turn back and try again later—the challenge will always be there tomorrow. Finally, always remember that the most effective method of dealing with a survival situation is to avoid getting into it in the first place.

Colin Towell

> Twenty years from now you will be **more disappointed** by the things **you didn't do** than by the **ones you did do**
>
> MARK TWAIN

BEFORE YOU GO

PREPARE YOURSELF

Most survival situations arise in one of two ways: either you are thrust into a situation not of your making and beyond your control, or (as is more often the case) a situation develops because of a sequence of events that could have been avoided had you recognized the danger signs and acted on them at the earliest opportunity. Unfortunately, most survival situations occur as a result of ignorance, arrogance, an overconfident belief in one's own ability, or because the forces of nature have been underestimated.

In the military forces, troops prepare for survival situations by learning the basic principles and techniques and then practicing them until they become second nature. The more you practice a particular survival skill, the more you'll understand how and why it works, and the more you'll understand your own strengths and weaknesses. Many helicopter crew members who have been involved in a survival situation, for example, report that, having ditched their aircraft in the sea, they could not remember how they subsequently

In this section YOU WILL DISCOVER...

- how keeping **in shape** can keep you **out of trouble...**
- the importance of a **positive mental attitude...**
- that where there's a **will to survive** there's a **way to survive...**
- the difference between a **coping strategy** and an **aggravating factor...**
- how **"show-stoppers"** could **scupper your trip...**
- why an **emergency plan of action** could **save your life...**

escaped from their aircraft and successfully boarded their liferaft. Thanks to their intense training regime, their actions had become instinctive and therefore subconscious.

Whether you're preparing for an overnight camping trip or a year in Africa, the more prepared you are to meet the challenges posed by a particular environment, the more likely you are to be able to cope—both physically and mentally—should you then find yourself confronted with a survival situation.

It would be foolish to think that the amount of prior preparation required for an expedition is directly related to the length of the trip or its perceived danger.

ACCIDENTS CAN HAPPEN anywhere. One of the most important things you can do to improve your chances of survival is to let people know where you're going and when you'll be back. Leave an itinerary of your trip with family or friends, and arrange a time when you will call to let them know you are safe.

In many instances, simply making sure you have a cell phone with you can prevent an accident becoming a major survival incident. In 2007, a professional athlete, out running with her dog in Moab, Utah, slipped on ice, fell approximately 60 ft (18 m), and broke her pelvis. Having left her cell phone in her vehicle, she managed to survive for two nights in the freezing desert before rescuers were finally lead to her by her dog.

In a separate incident in 2008, a farm worker in the UK caught his arm in machinery. Unable to attract attention and not having his cell phone with him, he was faced with a stark choice when the machinery caught fire: bleed to death, burn to death, or cut off his own arm with a pocket knife—he chose the latter.

> “Never assume that the amount of **prior preparation** required is directly related to the length of a trip or its **perceived danger**”

GETTING INTO SHAPE

IN ANY SURVIVAL SITUATION you're likely to be at your best, both physically and mentally, in the moments before the situation occurs. From that point on, through lack of sleep, food, and water, your situation will deteriorate until your rescue. Being in good physical condition will help you overcome the challenges you'll face in a survival situation.

KNOWING YOUR LIMITS
Having an idea of how fit you are and knowing how much you are capable of achieving before entering the wilderness means you can set yourself realistic goals—which will help to keep you out of danger.

THE BENEFITS OF EXERCISE

Over time, exercise induces changes such as weight loss, as well as improved posture, physique, strength, agility, mental alertness, and stamina. These are all vital attributes should you find yourself in a survival situation.

HOW FIT ARE YOU?

A fit heart pumps blood more slowly and efficiently than an unfit one. The hearts of women, children, and older people beat faster than those of young adult males. Take your pulse first thing in the morning—this is known as the "base rate." As the table below shows, the faster your pulse returns to its base rate after a period of exercise, the fitter you are. Note that < denotes "less than" and > denotes "more than".

AGE	20-29	30-39	40-49	50+
RATING	BASE RATE			
MEN				
Good	<69	<71	<73	<75
Average	70-85	72-87	74-89	76-91
Poor	>85	>87	>89	>91
WOMEN				
Good	<77	<79	<81	<83
Average	78-94	80-96	82-98	84-100
Poor	>94	>96	>98	>100

AGE	20-29	30-39	40-49	50+
RATING	BEATS AFTER EXERCISE			
MEN				
Good	<85	<87	<89	<91
Average	86-101	88-102	90-105	92-107
Poor	>101	>102	>105	>107
WOMEN				
Good	<93	<95	<97	<99
Average	94-110	96-112	98-114	100-116
Poor	>110	>112	>114	>116

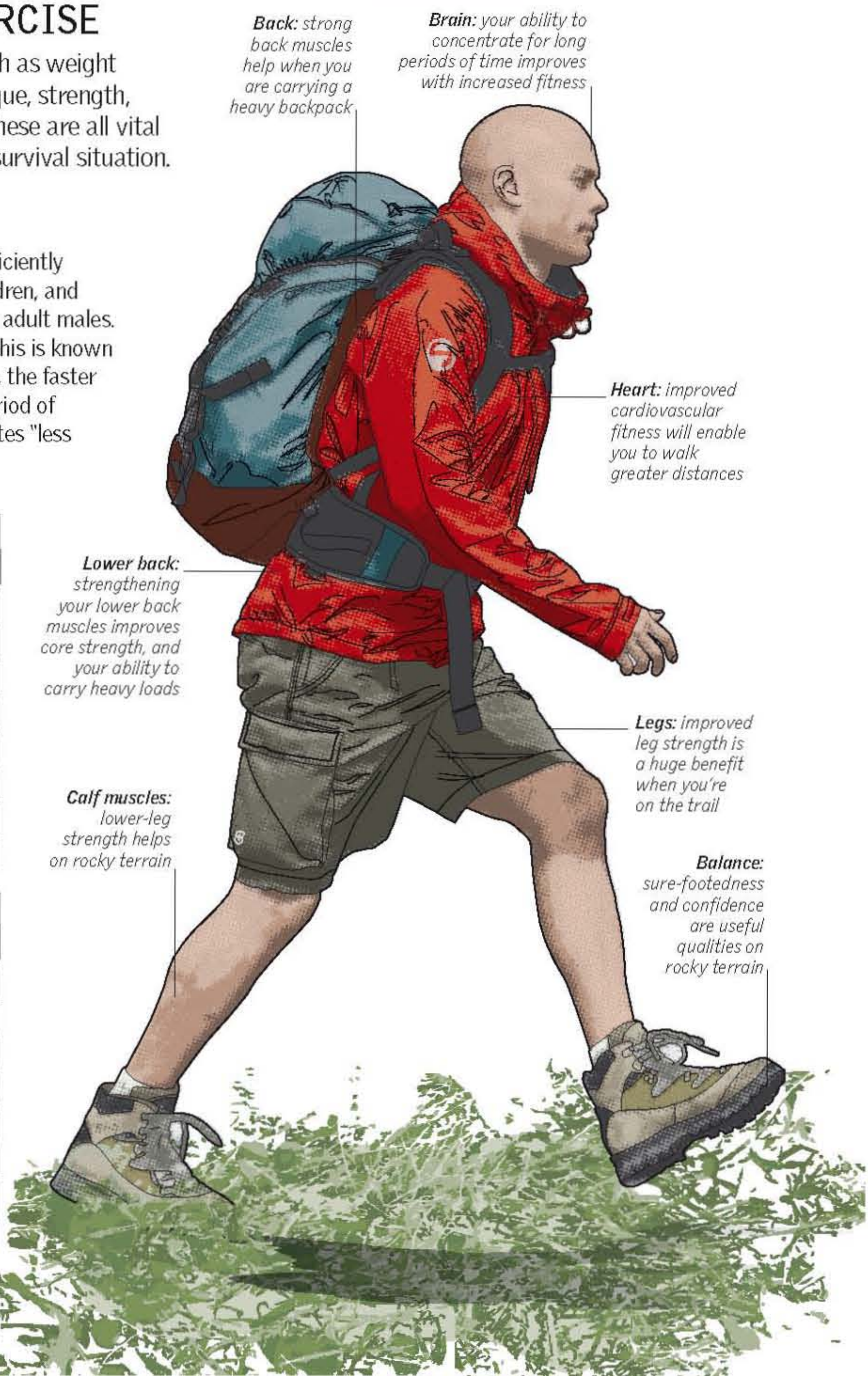
Back: strong back muscles help when you are carrying a heavy backpack
Brain: your ability to concentrate for long periods of time improves with increased fitness
Heart: improved cardiovascular fitness will enable you to walk greater distances
Lower back: strengthening your lower back muscles improves core strength, and your ability to carry heavy loads
Legs: improved leg strength is a huge benefit when you're on the trail
Calf muscles: lower-leg strength helps on rocky terrain
Balance: sure-footedness and confidence are useful qualities on rocky terrain

FIT FOR THE CHALLENGE

Survival fitness is not about trying to run the 100 meters in 10 seconds; it's about stamina and endurance, and about knowing your physical limitations and being able to work with them. It is also about understanding that you may have to push yourself beyond your limits and that having a positive mental attitude is paramount to survival—the mind will often give up long before the body has reached its limit. Bear in mind that every task in a survival situation will sap your energy.

EXERCISING REGULARLY
Any effective training schedule must include at least three 45-minute sessions per week, with exercises that raise the heartbeat to more than 120 beats per minute.

STARTING A FITNESS PROGRAM

To spend time in the wilderness with a pack on your back requires a combination of strength and aerobic fitness. Training in the gym before you set off will strengthen your heart, lungs, and leg muscles, and will increase your stamina levels when you are out on the trail.

- Seek expert instruction and guidance wherever possible.
- Start gradually and then build up your routine progressively.
- Never try to work through injuries—rest and seek medical advice.
- Devise a relevant fitness program; there are many websites and organizations that provide you with detailed fitness programs for various levels of specific activities, from training for a trek through the jungle to mountain-biking along the Appalachian Trail.
- Duplicating in training what you'd like to be able to achieve on the trail will allow you to build up reference points about how your body works, and how it copes in different situations. The more you know about how you perform, the better equipped you'll be on the trail.

WARMING UP AND COOLING DOWN
Time spent warming up and cooling down after exercise will help to improve your endurance levels and will also accelerate the recovery process. Get into the habit of starting and ending your session with a five-minute jog.

GO FOR A CHECK-UP

It's a sad fact that the majority of us don't have regular medical check-ups. Many people work on the principle that you only need to see a doctor or dentist if something is wrong. However, you should always make sure that you start any adventure or wilderness trek in top condition. Therefore, before you set out, pay a visit to your doctor and dentist and make sure that your body and teeth are in good condition.

VISIT THE DOCTOR

- Let the doctor know of any ailments that have been a concern to you during the course of the previous year.
- Tell the doctor where you are going and enquire about relevant inoculations or medication that you should be taking with you.

VISIT THE DENTIST

- Have any problems with your teeth dealt with before you go. Any minor dental problem will almost certainly turn into full-blown toothache when you least want it to.

USEFUL EXERCISES

If you already get regular exercise, you'll find yourself far more comfortable in a survival situation than your more sedentary counterparts. Any pre-expedition training regime should include plenty of stretching, aerobic exercise, and weight-training.

STRETCHING YOUR UPPER BODY AND BACK

Regularly stretching your arms, neck, chest, and shoulders will help your body to maintain a strong core. This will be of great use when it comes to activities such as rock climbing or using trekking poles. Undergoing a regular back-stretching routine will loosen the back muscles, making them more flexible and less susceptible to injury, and will increase both the back's range of motion and its endurance.

STRETCHING YOUR LEGS

Because your leg muscles bear the brunt of the work, leg stiffness is a common complaint at the end of a long day on the trail. Stretching your legs will improve your flexibility, increase blood circulation, and relax your muscles. Concentrate on your calves, quads, and hamstrings.

AEROBIC EXERCISES

Also called cardiovascular fitness, aerobic fitness refers to the ability of your heart, blood vessels, and lungs to supply oxygen and nutrients to the rest of your body during sustained physical activity. Regular aerobic exercise—such as swimming, jogging, or cycling—will reduce the risk of developing coronary heart disease or high blood pressure, will help you to manage your weight and increase your stamina levels, and will make your heart stronger and more efficient, thus improving blood flow around your body.

WEIGHT-TRAINING

Increased muscle strength will enhance your ability to perform everyday tasks such as lifting, carrying, and walking. A regular weight-training program will improve your posture, build muscle density around the joints and bones, improve your sense of balance, help your stress management abilities, aid sleep, and reduce the chances of you sustaining an injury. In addition, research has shown that a regular resistance-training program can improve your metabolic rate by up to 15 percent, which will help you shed any unwanted weight at a faster rate.

MENTAL PREPARATION

REGARDLESS OF WHETHER you're backpacking through a remote area, or on a day hike in familiar territory, a situation may arise that changes your circumstances for the worse. Very quickly you move into the unknown, which causes tremendous psychological and emotional stress, known as "psychogenic shock." Understanding this will help you to deal with it better and reduce its impact.

INDIVIDUAL REACTIONS TO DISASTER

People react to survival situations in different ways, although you can expect to find some common emotional reactions in victims who are experiencing, or have survived, a disaster situation. You may experience one or more of them during or after any survival experience or trauma.

YOUR RESPONSE TO DISASTER SITUATIONS

Your psychological response to a survival situation is crucial. Statistics show that 95 percent of people who die with psychological trauma die within the first three days. Losing the will to survive—or suffering psychological disorders that prevent you from coping with the physical conditions—is your main concern. If you break down psychologically, your chances of overcoming a situation will be compromised.

PSYCHOLOGICAL PROGRESSION

It's useful to examine how people are likely to react in a survival situation. Using this knowledge, it's possible to prepare mentally for such eventualities and, in so doing, lessen the impact if the worst should happen to you when out on the trail. Normal psychological reactions to disaster tend to occur in a set pattern of four stages: the pre-impact period, the impact period, the recoil period, and the post-trauma period (see right). Contrary to popular belief, people don't normally panic, although it can be contagious if someone does.

AGGRAVATING FACTORS	HUNGER	THIRST
Reactions to disaster can result from a direct blow to the psychological system, such as extreme shock, but they can also be brought on, or aggravated by, other factors. As with all psychological problems, knowing what these aggravating factors are, and attempting to avoid them—or at least recognizing what they might lead to—will maximize your chances of preventing or overcoming the problem. The most common aggravating factors are: hunger, thirst, fatigue, seasickness, and hypothermia.	Initially, hunger is not a problem, but a long-term lack of food will cause psychological changes to occur. Symptoms include: ▪ Apathy ▪ Irritability ▪ Depression ▪ Lack of concentration	Thirst is a serious problem, especially for survivors at sea or in the desert, and its effects are more acute than hunger. Agitation is commonplace; other symptoms include: ▪ Irrational behavior (see box, right) ▪ Delusions ▪ Visual hallucinations

COPING STRATEGIES	TRAINING	MOTIVATION
There are many things you can do to prepare yourself psychologically for a survival situation—such as learning about what you should expect if the worst were to happen—and to cope better if you find yourself in one. As with all survival skills, prior knowledge is power, and will help you to deal with a survival situation far more successfully. The main areas to think about are: training, motivation, attachment, hope, acceptance, and helping others. Developing coping strategies is an important technique for survival.	People who are properly prepared, who know their environment and how to use their equipment, and have an understanding of what to expect in a survival situation, will be far more effective if they find themselves in one. Adequate training and practice using your equipment will help you to function effectively at an automatic level. Prior knowledge is key to your survival.	Often known as "the will to survive," motivation involves a refusal to accept death, and to hang on to the belief that you were not meant to die under these conditions. It involves overcoming the emotional and physical discomforts of extreme conditions. Linked to motivation is the ability to establish goals, work out the steps to those goals, and to follow those steps through.

PANIC

Panic arises from the fear of what might happen rather than what has happened. It tends to occur when people are trapped, or if there is a time limit to their escape.

DEPRESSION

Depressed people will sit among chaos and debris vacantly gazing and not replying to questions. They're unaware of their situation and unable to help themselves, so risk further injury.

HYPERACTIVITY

Hyperactive victims are easily distracted, and are full of chatter, ideas, and often unhelpful suggestions. Sufferers can reach this stage after a state of depression.

ANGER

Aggression, anger, and hostility are common reactions to trauma. They're often irrational and may even be directed at the rescuers or medical staff trying to help them.

GUILT

Some sufferers feel guilty for surviving, and for not having done enough for others—and some irrationally blame themselves for bringing about the incident.

SUICIDE

Disaster victims have been known to commit suicide immediately after being rescued, in some cases when they're already safe in the hospital. Victims should be closely monitored.

Pre-impact period

The "pre-impact" period is divided into two stages:

- Threat: danger exists but, though obvious to those who recognize it, those who will not accept it respond with denial and under-activity.
- Warning: threat of danger is now apparent to all; response is now likely to be over-activity.

Impact Period

This is the life-threatening stage. Statistically, individuals behave in one of three ways:

- 10–20 percent of people are calm and retain full awareness.
- Up to 75 percent of people are stunned, bewildered, and unable to react rationally.
- 10–25 percent exhibit extreme behavior, such as screaming.

Recoil period

This follows on directly from the impact period; for example, victims may have escaped a sinking ship and are in the liferafts. It can last for up to three days, but generally lasts for around three hours. In most cases, it is characterized by a gradual return to normal reasoning abilities, awareness, and emotional expression.

Post-trauma period

If the recoil period is not fully successful, individuals may develop psychiatric disorders. The full impact of the incident becomes apparent and a range of emotions—guilt, depression, anxiety, aimlessness, and a feeling of bereavement—may develop. These are often referred to as Post-Traumatic Stress Disorders (PTSD).

FATIGUE	SEASICKNESS	HYPOTHERMIA
In many cases, physical exhaustion is present from the outset. At other times, it may result from sleep deprivation and the physical hardship endured over time. Most survivors agree that fatigue overwhelms them, but when they want to sleep they can't—they have an inability to relax. Fatigue causes a deterioration in mental and physical performance, followed by a psychological and physical debility.	Seasickness often brings about an overwhelming desire to curl up and die, which in survival situations can easily become a reality. It's important not to give in to this urge. Fight seasickness with the following methods: ▪ Keep a fixed point such as the horizon in sight. ▪ Take small sips of water (not salt water) if you have sufficient supplies—but ration them if you're in a life or death situation.	Hypothermia (see p. 273) produces both physical and psychological effects—the psychological consequences occur early in the condition, and cause: ▪ Loss of concentration ▪ Loss of memory ▪ Motor impairment ▪ Faulty decision-making ▪ Irrational behavior

IRRATIONAL BEHAVIOR

Irrational behavior can take many forms. Examples include the earthquake victims who were found collecting flowers instead of helping the injured, and famously, the band of the *Titanic*, who played while the ship sank rather than trying to save themselves.

ATTACHMENT	HOPE	ACCEPTANCE	HELPING OTHERS
One of the strongest motivating forces for survival is the desire to be reunited with principal figures of attachment in your life. These may include: ▪ Husbands ▪ Wives ▪ Partners ▪ Children ▪ Grandchildren ▪ Close friends	To hope means to entertain ideas that a distressing situation will improve and get better. In any survival situation it's important to cling onto hope, despite information or perceptions to the contrary. Thinking positively will help to ward off psychological trauma. It's often easier to be optimistic in a group situation than it is if you're on your own, as people can support each other.	An inability to accept one's situation or condition leads to frustration, anger, and irrational behavior, and it's important to avoid these feelings in a survival situation. The ability to accept the situation doesn't equate to giving in to it—far from it. Those who have this ability, and know when to be active and when to be passive often have a better chance of long-term survival.	First, monitor your own condition and check that you're really up to the task. Determine who is genuinely disturbed rather than showing "normal" reactions. Psychological first aid is only required for those who are failing to recover. Simple words of comfort and interest will make the majority who are numbed more responsive. Those who are disturbed should be monitored closely. Avoid giving sedatives.

PLANNING YOUR JOURNEY

WHETHER YOUR TRIP involves a day out with your backpack, or a full expedition over weeks, crossing continents in four-wheel-drive vehicles, you need to plan it very carefully. The plan for your day out won't be as detailed as it would be for the longer expedition, but it's equally important. It's a good idea to have a basic planning outline for your most regular trips, to which you can add supplementary information as your trips become more involved.

MINIMIZING THE "IF ONLY"

There's no way that you can plan for every eventuality—there are just too many variables—but what you can do is look at the type of trip you intend to take, and ensure that if a situation arises you're not left wishing you'd done something differently. The time to minimize the chances of an "if only" situation is during the planning stage. Look at the potential problems and risks, plan to avoid them, and equip yourself with the knowledge and/or equipment to deal with them if they arise.

WARNING!
The social customs of a country are not the same all over the world. Many cultural differences stem from ancient cultural or religious pasts, and are taken extremely seriously. While ignoring or breaking some customs may lead only to embarrassment, breaking others can result in a penalty, punishment, or even imprisonment. Cultural differences may require females in your group to dress so that they don't show bare arms or legs, or they may dictate which hand you use to greet someone politely. Always research a country's customs thoroughly when planning your trip.

THE SIX P'S

Remember the six p's: **P**rior **P**lanning and **P**reparation **P**revents **P**oor **P**erformance. Research has shown that the longer and more complicated the trip is, the less likely there is to be a major "survival" situation. This type of trip is likely to be well organized, and potential problems will have been considered. This means that they can either be avoided, or that there will be mechanisms in place to deal with them. In many ways, simply having a good understanding of how to deal with a situation, and being able to interpret the basic principles of survival, can prevent a minor problem from escalating into a disaster. In a survival situation, it may be your knowledge, combined with your ability to improvise, that determines whether you become a survivor or a statistic.

ORGANIZATIONAL PRIORITIES

When planning for a trip always start with the most important things—known as the "show-stoppers." These are generally the things that would actually stop the trip from happening in the first place if not organized in advance. If you then work backward from the show-stoppers to items that would simply make the trip more comfortable, most other things will fall into place. The chart on the right lists the things that you definitely cannot do without.

	SHOW-STOPPERS
MONEY	▪ You'll need enough to cover your needs plus extra for emergencies. ▪ It must be in the correct currency for the country you are visiting. ▪ Ensure your ATM card is set up for use in that particular country. ▪ Make sure you have a secure way of storing your money.
PASSPORT	▪ Your passport must be valid. If you need to renew it, do so well in advance. ▪ Some countries require your passport to be valid for a number of months after your visit. ▪ Keep your passport number in several places, such as inside your survival kit. ▪ Keep your passport safe, and in a waterproof container, such as a Ziploc bag.
VISA	▪ Research the regulations regarding visas relevant to the countries you're traveling to. ▪ Find out how to apply, how far in advance you should apply, whether or not you can apply in the country, and what other documents you need. ▪ As with your passport, always keep your visa safe, dry, and in good condition.
VACCINATION	▪ Check the regulations for the country—many have strict vaccination policies. ▪ Ensure you have the vaccinations, and any boosters, within the correct timeframe. ▪ Some inoculations last only for six months, so for extended trips you might have to have more while away. This can usually be arranged through local hospitals or clinics.
TICKETS	▪ Make sure you have the correct tickets for your journey. ▪ Read the tickets and make sure they have the correct names, dates, and locations. ▪ Should you need to prove your movements, always keep your tickets in a safe place and never throw them away—often the return journey is included on the same ticket.
INSURANCE	▪ It's advisable to get insurance against your trip being cancelled. ▪ Ensure that your insurance will cover your medical expenses should the worst happen. You don't have to be climbing Everest—even a twisted or broken ankle when hill-walking could prove to be very expensive if you don't have adequate insurance.

YOUR TEAM

If you're embarking on a trip with a group, remember that team dynamics will play an important part in the success, or otherwise, of your trip. Stressful situations—and particularly survival situations—can bring out the best, or the worst, in people. When planning a long trip, it's always a good idea to plan several shorter trips beforehand, to use as practice sessions. These will not only help you decide what equipment to take, and provide you with an opportunity to practice using it, but will also allow team members to assess how they work together as a group. This can help you to organize your team more effectively.

MIXED-GENDER GROUPS

If the team has both male and female members, it's important to factor this into the planning stage. You'll need to consider sleeping and washing arrangements, who carries what, who is responsible for what, and so on. All this should be planned in detail before you set off. It should also be noted that taking it for granted that the women will cook the meals while the men make the shelters is not the best way to start an expedition.

MIXED-AGE GROUPS

When planning your trip, remember that different age groups may have different levels of fitness, which will affect the pace of your group. However, what older members may lack in pace, they may make up for in expedition experience.

PREPARATION THROUGH TRAINING

Training yourself both mentally and physically—and practicing with the equipment you'll be using—may seem like the obvious thing to do, but in many cases the obvious can often get overlooked. You'll get the most out of your trip if you're mentally and physically prepared to a level that means you can operate within your own capabilities and comfort zone. This will allow you to enjoy and appreciate the experience, as opposed to just getting yourself through it.

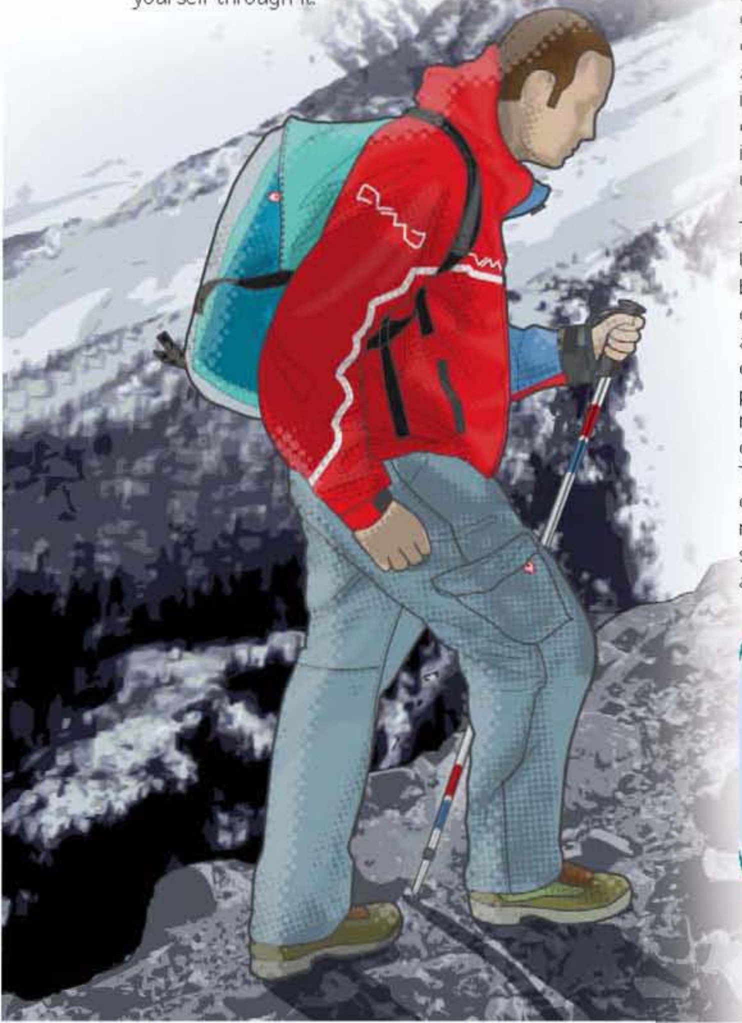

TRAINING YOURSELF

Duplicate in training what you intend to do on your trip. Build up your training gradually, over a period of weeks and months, and take into account the following:

- **The environment**: research the weather conditions you'll be facing, and look at the extremes as well as the average. For example, desert areas may be hot during the day, but can drop to below freezing at night.
- **Weight:** increase the amount of weight you carry until you're eventually carrying what you intend to take. This will not only condition you to the weight but also help you to decide just what's important to take and what's not.
- **Distance:** if your trip involves covering a certain distance a day, then train for that distance. This will give you an indication of whether it's achievable and sustainable.
- **Language**: if visiting a country in which your native language is not widely spoken, it will be beneficial if you can learn some useful phrases. Take a phrase book or an electronic translator.

TRAINING WITH EQUIPMENT

Use your equipment as much as you can, and find out the best way to operate it through practice under realistic conditions (see box, below). This will highlight its strengths and weaknesses, and allow you to determine both its capabilities and your own—the compass may be working perfectly, but practice may highlight the fact that you're not confident enough using it, in which case you'd need to do further training before you set off on your expedition. The wrong time to be trying to find the jack and spare wheel on your rented vehicle is at night, in the rain, with the mosquitoes looking at you as their next free meal. Before you set off, always think about the various skill demands required, and ensure that you're capable of addressing those demands.

OPERATING UNDER REALISTIC CONDITIONS

When training with new equipment, always practice using it under realistic conditions. For example, if you're going to be using your GPS in cold conditions, can you operate it using the gloves you have selected? If you're pitching a tent, have you got **all** the necessary components, and can you put it up in the dark and rain?

EMERGENCY PLAN OF ACTION

UNFORTUNATELY, EVEN THE MOST meticulously planned, thoroughly equipped, and best-executed trip can run into difficulties. Unforecasted strong winds could trap you and your kayak on an island overnight, or a sprained ankle could leave you unable to climb down the rocks you climbed up so easily. Each of these survival scenarios are difficult to predict, but could easily happen.

USING YOUR EPA

An EPA should contain up-to-date information about you as a person. This may change very little over the years, so you may only need to create a standard EPA and just update the details specific to each trip.

RUNNING INTO PROBLEMS

It's important to make sure that you have done all you can to help yourself, and help any rescue attempt that might take place if you do have any problems. Always remember that there are two sides to any survival or rescue situation: the part that you play and the part that the rescue services play. However, the effectiveness of the rescue services can be greatly increased when they have access to all relevant information—in many cases this arrives too late. It's vital to keep people informed of your intended whereabouts, so that they can raise the alarm if you deviate from your plans.

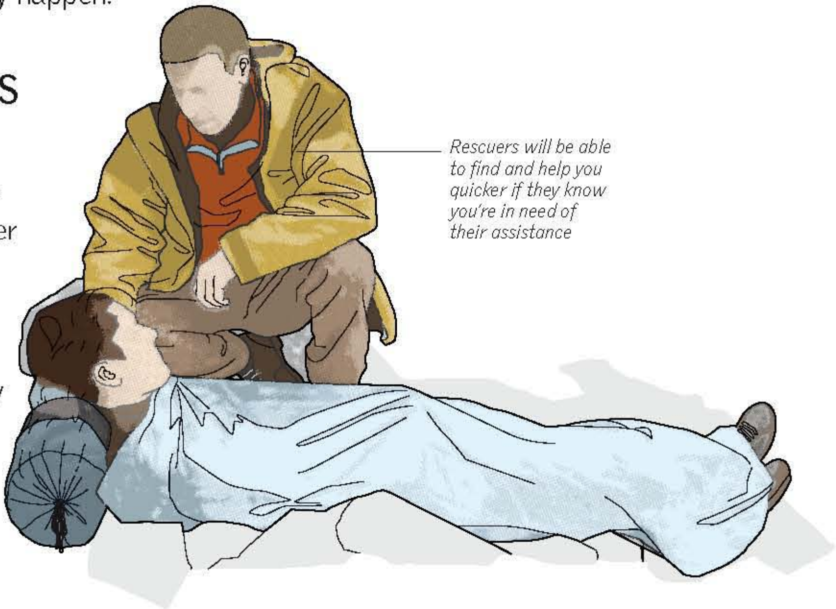

Rescuers will be able to find and help you quicker if they know you're in need of their assistance

WORST-CASE SCENARIO

In the military, every mission that's undertaken, particularly in a theater of conflict, has plans in place for the worst-case scenario. Each part of the mission is meticulously planned, and the team completes a form that states what their basic intention will be should anything happen during the various stages.

IF THE WORST HAPPENS

Having planned for the worst, should the team find themselves in difficulty, the rescue group will not have to try to second-guess what the team will do—they will have a clear indication of the team's intent and can plan effectively and decisively around this information. The team will be found quicker as a result. It's a good idea to apply this principle to your own trip.

KEEPING PEOPLE INFORMED

A good way of keeping your friends, family, and the relevant rescue services informed of your whereabouts is to write down the details of your trip, including pertinent itinerary places and dates, so that if you don't reach your destination when planned, the alarm can be raised. Similar to the "worst-case scenario" principle practised by the military (see left), an "emergency plan of action" (EPA) should be prepared by anyone venturing into the great outdoors (see right). Give a copy to your next of kin and group members, and keep a copy on your person. Where applicable, inform local services, such as park and ranger stations, of your intentions. Remember to inform these people when you arrive safely at your destination.

WRITING AN EMERGENCY PLAN OF ACTION (EPA)

The best way to determine what should be included in your plan is to look at the worst-case scenario you could find yourself in, and ask yourself what information your next of kin would need to know about you and your intentions if they felt they had to raise the alarm. In addition, if you did go missing—particularly in another country—there are many things that the relevant rescue services would need to aid their efforts, for example a recent photograph, your passport details, what equipment you have with you, what languages you can speak, and your skills. The clearer their understanding of you, your abilities, and your intentions, the easier their job will be. In addition, the more information your next of kin have, the more proactive they will feel in your rescue.

EMERGENCY PLAN OF ACTION FORM

Full name as appears on passport:
John William Smith

Date of birth:
(dd/mm/yy)
05/28/60

Height: *5′10″ (178cm)*
Weight: *168 pounds (76 kilos)*
Hair color: *Brown*

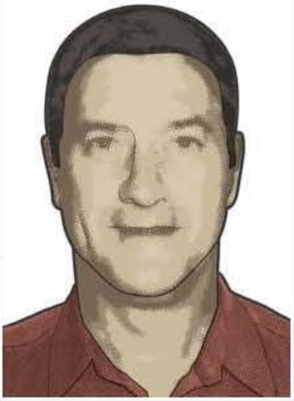

Passport number:
2008XXXXX63
Expires:
11/03/13

Driver's license number:
JHY280771smit
Expires:
12/28/22

Distinguishing marks (scars, tattoos):
Small scar—center of forehead
Large scar—right hand, middle finger
Chinese symbol tattoo—right arm

Languages spoken (fluent/basic):
English—native
French—basic
German—basic

Medication:
Medication—anti-malaria tablets

Allergies:
Amoxicillin

Swimmer: *Strong swimmer*
Outdoor skills/experience:
Attended basic military survival training.
Attended basic bushcraft course
Experienced in living outdoors

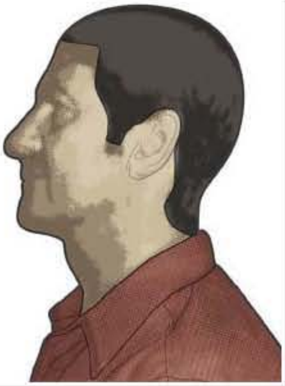

Next of kin 1: *Father*
William Smith
1018 Furlong Avenue
Brunswick, Maine,
USA 04555

Next of kin 2: *Brother*
Andrew Smith
1023 Parkglen
Ashford, Kent,
TM24 5HZ
UK

Tel: *(001) 55 555 2356*

Tel : *(0044) (0)155 555 2357*

Email: *willsmith@internet.com*

Email: *andrewsmith@internet.com*

Trip details:
Campsite 1 = Grid ST456654
Campsite 2 = Grid ST654987
Vehicles: Landrover 1 = white, reg MH55 555
Landrover 2 = blue, reg MH56 555
Group = Ben Jones, Kim Smith, and myself
Day 1: Park Landrover 2 at Campsite 2, and drive in Landrover 1 to Campsite 1
Day 2: Follow well-defined path along the Derwent Line Trail, aiming to camp overnight at Grid 4561559
Day 3: Continue along the Derwent Line Trail, aiming to be at Campsite 2 by mid-afternoon. Camp overnight at Campsite 2
Day 4: Travel in Landrover 2 to campsite 1 and retrieve Landrover 1

Foreseeable problems/intentions:
Day 1: None
Day 2: None but will use Ranger Station 18, grid 555555 (Tel. 666 6666) as an emergency rendezvous point
Day 3: None but will use Ranger Station 19, grid 666666 (Tel. 555 5555) as an emergency rendezvous point
Day 4: None

Communications plan:
Will speak to Dad on the morning of Day 1 and try to phone during the trek but am unsure of cell reception once on the trail so don't worry if you hear nothing.
Will phone Dad again when we reach campsite 2 on day 3.

My mobile: *07979 555555*
My email: *jws@internet.com*
Alt. No: *Ben 05555 555555*
Alt. No: *Kim 05555 555555*
Alt. No: *Campsite 1 555 555 55555*
Alt. No: *Campsite 2 555 555 55555*

Date: *November 23, 2009*

KNOW YOUR ENVIRONMENT

The continued survival of the human race can be attributed to our ability to adapt to our environment. While we may have lost some of our ancient ancestors' survival skills, we have, in their place, learned new skills as and when they have become necessary. The issue we face today is that the gap between the skills we once had and the skills we now have grows ever wider as we rely more heavily on modern technology. Therefore, when you head off into the wilderness, it is important to fully prepare for the environment.

Before a trip, research how the native inhabitants dress, work, and eat. How they have adapted to their way of life will help you to understand the environment and allow you to select the best gear, adopt the best techniques, and learn the correct skills. This is crucial given that most survival situations arise as a result of a sequence of events that could have been avoided—while you may have no control over the aircraft crashing, for example, you can recognize a change in the weather and choose whether to continue or turn back.

In this section YOU WILL DISCOVER...

- the difference between **tundra** and **taiga...**
- why you should **step-up** the size of your **survival footprint...**
- the **way to go** when it comes to **snowy peaks** or **swampy creeks...**
- how to avoid **getting lost** in the **permafrost...**
- the best way to **stay at the top** of the **food chain...**
- how to **feel at home** where the tropical **butterflies roam...**
- how a high-visibility **survival suit** could really **get you noticed...**

The four basic principles of survival are:

protection, location, water, and food. In most survival situations, this is also the order in which you should prioritize them.

P **PROTECTION** You must stay in a condition that allows you to be proactive in your continued survival and rescue. Physically, you should protect yourself against injury, the elements, and wildlife. Mentally, you should protect yourself against emotions that could rob you of the will to live: fear, guilt, despondency, and depression, for example. The best way to achieve this level of protection is to light and maintain a fire. Not only does it offer physical protection against the elements and wildlife, but it also provides a sense of security and familiarity that can help normalize even the most dire situation.

L **LOCATION** Your second priority is to recognize the importance of your location to your chances of survival and rescue. You will usually have two options: stay or go. Your preferred option should be to remain where you are and use anything at your disposal to mark your location to aid rescuers in their efforts to find you. If you can't stay where you are (perhaps because you are in imminent danger) you may have no option but to move to another location that provides either a better chance of survival or rescue, or both. Select a location aid that offers you the best chance of attracting attention in the environment in which you're traveling.

W **WATER** Put simply, water is the essence of life. While you may be able to survive for a few days without it, your ability to function and carry out even simple mental and physical tasks will be dramatically reduced in less than 24 hours. However, if you are injured, if the weather conditions are very hot, and if your workload is particularly heavy, for example, your survival time without water could be reduced to just a few hours. You should learn how to procure water in the particular environment in which you're traveling, and understand the ways in which a lack of water affects you.

F **FOOD** The importance of food is directly related to the length of time you are in a survival situation: the longer the situation lasts, the more important food will become in helping you stay fit and healthy. Even with a moderate workload, going without food for five to seven days will not kill you. You will, of course, feel hungry, you will grow tired, your movements will slow, and your body will lose its ability to repair itself. However, unless you're malnourished before you enter a survival situation, you are unlikely to starve to death within a week. Your body needs water to digest food, so always remember to prioritize water over food.

> “We rely so heavily on **modern technology**: water at the **turn of a tap**; heat at the **flick of a switch**; food raised, harvested and **prepared by others**”

TEMPERATE ENVIRONMENTS

THE TEMPERATE ZONES are the two regions between the Tropic of Cancer and Arctic circle, and the Tropic of Capricorn and Antarctic circle. Typified by seasonal variations in climate, conditions can include baking hot summers, freezing cold winters, and rainfall all year round. Landscape features range from forests and snow-covered peaks to grasslands and deserts. Although most temperate regions are highly populated, don't be lulled into a false sense of security—a worst-case scenario can still arise just a few miles from help.

WARNING!
Brown bears inhabit extensive wilderness areas in the northern temperate zone, and are the most dangerous temperate animal. The Grizzly, found in North America, is the most likely to attack humans.

TEMPERATE FEATURES

Although they contain a wide range of environmental features, the temperate zone's typical feature is forest, ranging from deciduous trees that shed their leaves in fall to coniferous trees that retain their leaves all year round. Grassland predominates where forests have been cleared, while highland areas contain hills and mountains. The abundance of rain means that rivers and lakes are common, and swampy wetlands form in areas with poor drainage.

TEMPERATE ESSENTIALS

Climate and terrain can vary widely in temperate areas, so preparation for a range of eventualities is essential:

- Don't underestimate the temperate environment. Although the vast majority of the human race lives in temperate zones, the diversity of terrain and weather means that survival equipment and knowledge must be broad enough to cope with a wide range of situations and conditions.
- Temperate weather can change very quickly, so check the local forecast before you set off, and carry a small AM/FM radio so you can listen to local weather reports.
- Plan a realistic and achievable route, and prepare an EPA (see pp. 24–25). Be ready to re-assess your route during the trip.
- Take clothing for the full range of conditions you may encounter.
- Carry a survival kit (see pp. 60–61), knife, emergency equipment, cell phone, and first-aid kit (see pp. 260–61)—and know how to use them.
- Carry adequate water, and equipment to collect and purify more if necessary.
- Carry some form of basic shelter, even if only going out for the day.
- Always take a map and compass, and consider using a GPS as an aid.

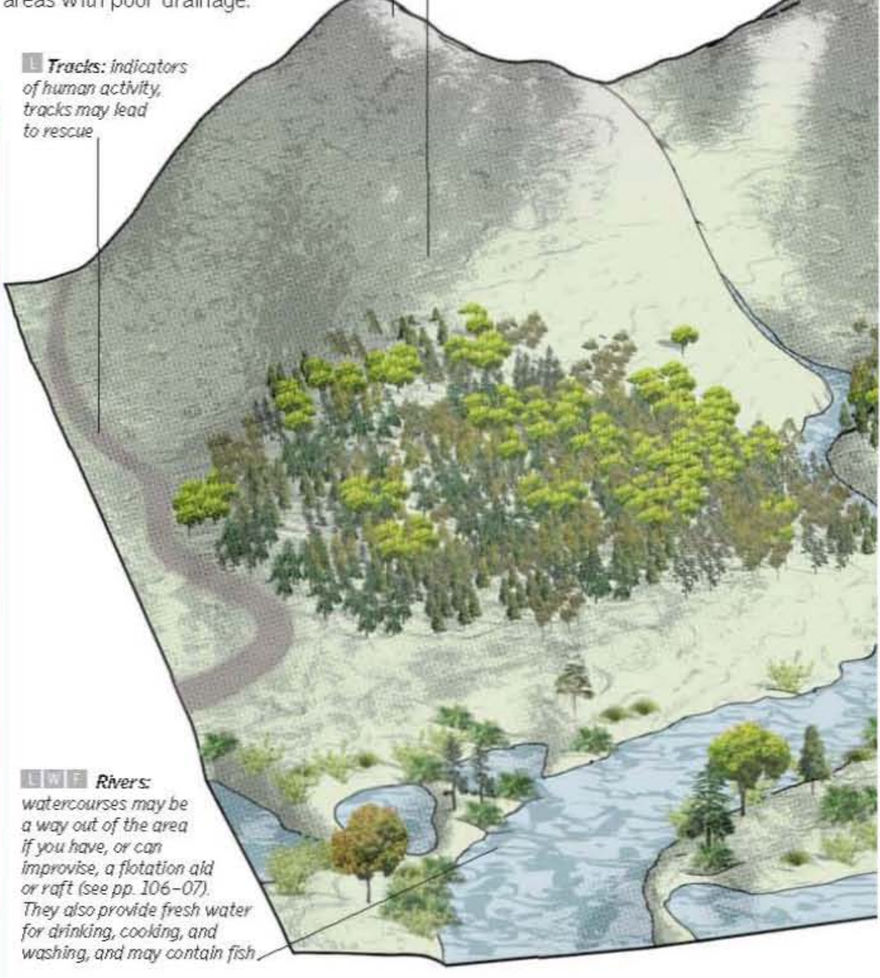

L ***High ground:*** *elevated areas afford a view of your surroundings and may indicate which area offers the best prospects for your survival, or where to go to find rescue*

P L ***High ground:*** *night-time temperatures are lowest at altitude, so descend to warmer levels before dark. Beware of rock slides and run-off during downpours*

L ***Tracks:*** *indicators of human activity, tracks may lead to rescue*

L W F ***Rivers:*** *watercourses may be a way out of the area if you have, or can improvise, a flotation aid or raft (see pp. 106–07). They also provide fresh water for drinking, cooking, and washing, and may contain fish*

WHERE TO FIND TEMPERATE AREAS

Deciduous forests are scattered throughout the temperate zones, but the largest occur in eastern North America, western Europe, and east Asia. Extensive coniferous forests are found in the higher latitudes of North America and Eurasia, while grassland is most common in continental interiors.

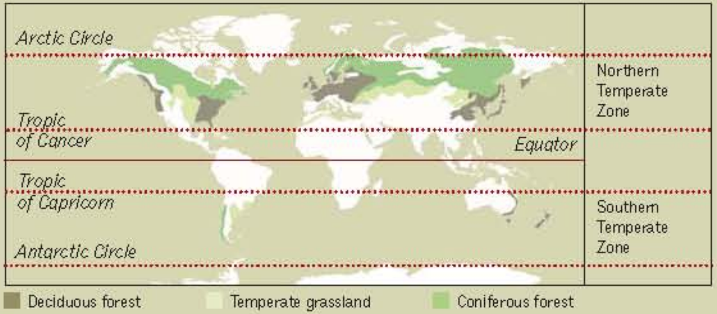

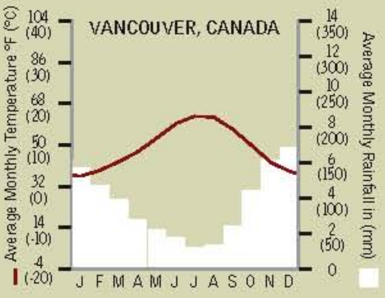

TYPICAL CLIMATE

Vancouver lies on the Pacific coast of Canada, and experiences a typical temperate climate of warm summers and cool, wet winters.

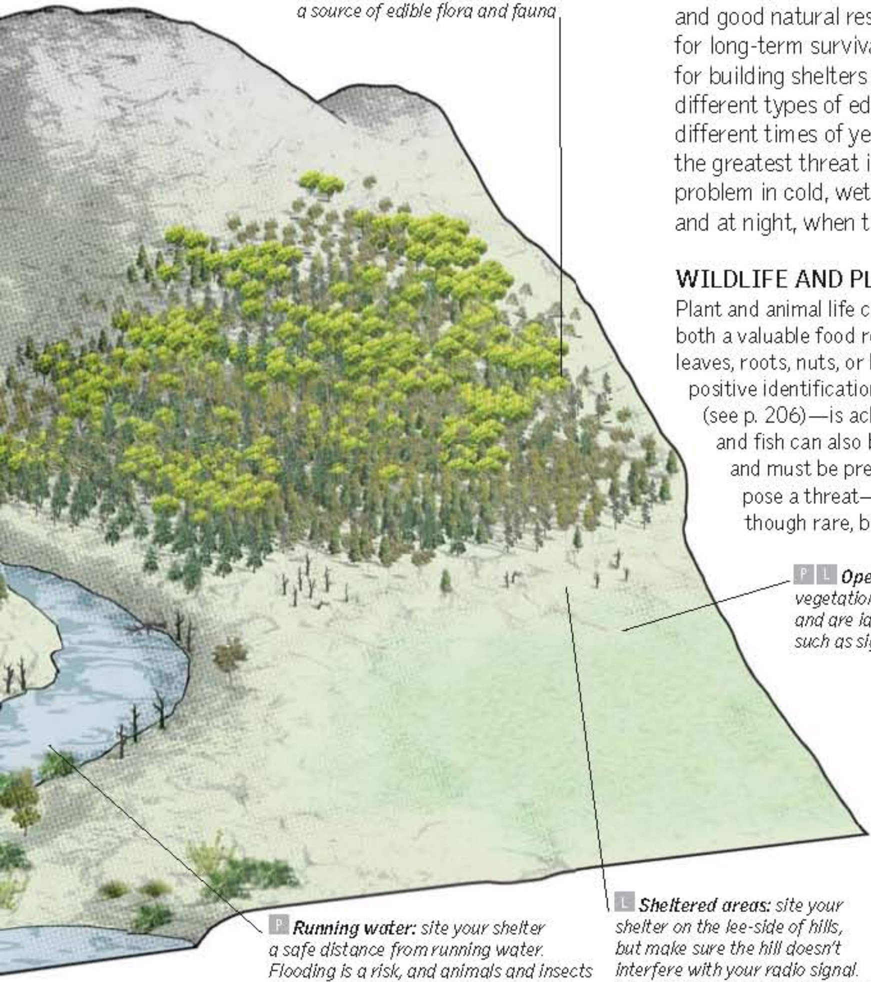

P L F ***Woodland:*** *tree cover affords protection from the elements, and wood for shelter, cooking, and signal fires. It is also a source of edible flora and fauna*

P L ***Open ground:*** *areas with no vegetation offer a clear view of dangers, and are ideal for deploying location aids, such as signal fires*

P ***Running water:*** *site your shelter a safe distance from running water. Flooding is a risk, and animals and insects are drawn to water. The noise may also hide the sound of animals or rescuers*

L ***Sheltered areas:*** *site your shelter on the lee-side of hills, but make sure the hill doesn't interfere with your radio signal. Choose a spot that faces the sun to maximize warmth and light*

SURVIVING IN TEMPERATE AREAS

Most temperate environments have a relatively mild climate and good natural resources, making them favorable places for long-term survival. Water can be found in most areas, wood for building shelters and lighting fires is usually abundant, and different types of edible plants and animals can be found at different times of year if you know where to look. Potentially the greatest threat is hypothermia (see pp. 272–73), a particular problem in cold, wet and windy conditions, especially in winter and at night, when temperatures tend to fall even further.

WILDLIFE AND PLANT LIFE

Plant and animal life can be abundant in temperate areas and represents both a valuable food resource and a potential hazard. Shoots and stems, leaves, roots, nuts, or berries of some plants are edible, but only when positive identification—or at the least, the Universal Edibility Test (see p. 206)—is achieved. Small mammals, birds, insects, reptiles, and fish can also be found, but may be difficult to catch and kill, and must be prepared and cooked properly. Wildlife can also pose a threat—snakes, spiders, and scorpions may bite and, though rare, bears, wolves, and cougars may attack if cornered.

FIND OUT MORE...

P **PROTECTION Shelters** pp. 156–65, 178–81 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–05

L **LOCATION Navigation** pp. 66–77 **Movement** pp. 86–89 **Signaling** pp. 236–41

W **WATER Finding** pp. 188–91 **Purifying** pp. 200–01

F **FOOD Plants** pp. 206–07, 280–81 **Animals** pp. 208–13, 216–29, 290–99

TROPICAL ENVIRONMENTS

THE TROPICAL ZONE stretches from the Tropic of Cancer to the Tropic of Capricorn, and is centered over the Equator. A range of environmental features occurs in this area depending on the local climate, from lush, humid, biodiverse tropical rain forests to dry, more sparsely vegetated tropical scrub. With adequate preparation and caution you should be able to survive unsupported for extended periods—the jungle holds just as many resources as dangers.

WARNING!
Mosquitoes are responsible for more fatalities than any other creature in tropical areas. They carry a range of diseases including malaria and yellow fever, which kill millions of people every year.

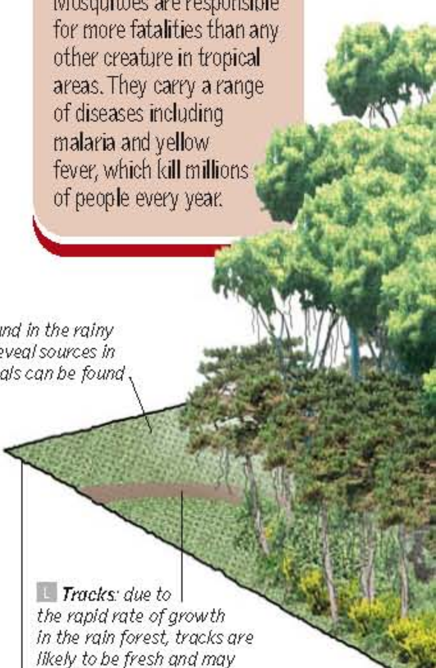

W F ***Tropical scrub:*** *water can be found in the rainy season, while animal movements may reveal sources in the dry season. Edible plants and animals can be found*

L ***Tracks:*** *due to the rapid rate of growth in the rain forest, tracks are likely to be fresh and may lead to rescue. Even remote areas may contain tracks left by loggers, prospectors, or local people*

P L ***Tropical scrub:*** *provides shade from the sun and materials for fire and shelter, but may be home to dangerous wild animals. Dense vegetation during the rainy season may impede visibility*

TROPICAL ESSENTIALS

Rain forests contain everything you need to stay alive, so remember the following tips when venturing out:

- While the jungle teems with predatory wildlife (such as big cats, crocodiles, and anacondas) that sees you as part of the food chain, it is the small creatures that can make life miserable. Most animals in the jungle want to avoid you as much as you want to avoid them—making a noise will scare most away.
- High humidity encourages infections, so keep everything covered (sleeves down, gloves on), and wash at every opportunity.
- Always sleep off the ground.
- Boil or treat all water. Some water found in plants can be safe to drink, but if in doubt, and if the water is yellow, milky, or very cloudy, treat before drinking.
- Many plants have defense mechanisms and can emit toxic liquids that may sting or burn—if you can't identify it, leave it alone.
- Dry tinder can be hard to find, so if you come across any, collect it and keep it safe and dry—lighting a fire in a moist, humid environment can be challenging.
- The smoke of a signal fire must penetrate through all the layers of the jungle, so light your fire where the canopy is sparse, such as the bend of a river or a clear-cut area.
- Navigation can be difficult, as you may only be able to see a few yards in front of you. Use dead reckoning—walking a short distance to a recognizable feature on your bearing—or pace-counting (see pp. 72–73).
- Rivers in the jungle usually run downhill to civilization and, eventually, the coast.
- Don't fight the jungle—tune in to its rhythm and work with it rather than against it.

TROPICAL FEATURES

Tropical rain forest, which occurs within 10 degrees of the Equator, is the predominant environmental feature in the tropics. But there are several other environment types just a few degrees of latitude away.

VERDANT RAIN FOREST

Tropical rain forests occur in areas with steady year-round temperatures and rainfall. Annual rainfall can reach around $6\frac{1}{2}$–10 ft (2–3 m) and daytime temperatures may reach 90°F (30°C), dropping to 70°F (20°C) during the night. Monsoon (or "seasonal") rain forests occur in areas with a wet and dry season, while montane rain forests, also known as "cloud forests," occur in mountainous areas.

SCRUB AND SWAMPS

Also known as "thorn forest," tropical scrub consists of low, woody plants with thorns, which usually grow in clusters separated by patches of bare ground (grasses are uncommon). Leaves are shed in the dry season, forming a dense, herbaceous layer in the wet season. Swamps are another common tropical feature, and may consist of fresh or saltwater. Freshwater swamps are found in low-lying, inland areas, and consist of masses of undergrowth, reeds, grasses, occasional short palms, and islands. Saltwater swamps often contain mangroves, and occur in coastal areas that are prone to tidal flooding so are best navigated by boat. Visibility in both types of swamp is poor, and movement is difficult.

SURVIVING IN THE RAIN FOREST

While natural resources are abundant in the rain forest, the heat, humidity, number of animals, and voracious rate of vegetal growth can make it an uncomfortable place. Water, materials for shelter and fire, and edible plants may all be found, although identification of plants is crucial to avoid poisonous species. Animal life is everywhere, so sleeping platforms must be built off the ground to avoid biting insects, snakes, and spiders. The greatest danger is becoming lost—navigation is difficult due to dense undergrowth at ground level, and rescuers will struggle to find you beneath the thick jungle canopy.

FIND OUT MORE...

P PROTECTION Shelters pp. 158–61, 166–71 **Fire** pp. 118–33, 204–05 **Dangers** pp. 167, 242–49

L LOCATION Navigation pp. 66–77 **Movement** pp. 106–07 **Signaling** pp. 236–41

W WATER Finding pp. 192–93 **Purifying** pp. 200–01

F FOOD Plants pp. 206–07, 284–85 **Animals** pp. 208–13, 224–29, 292–99

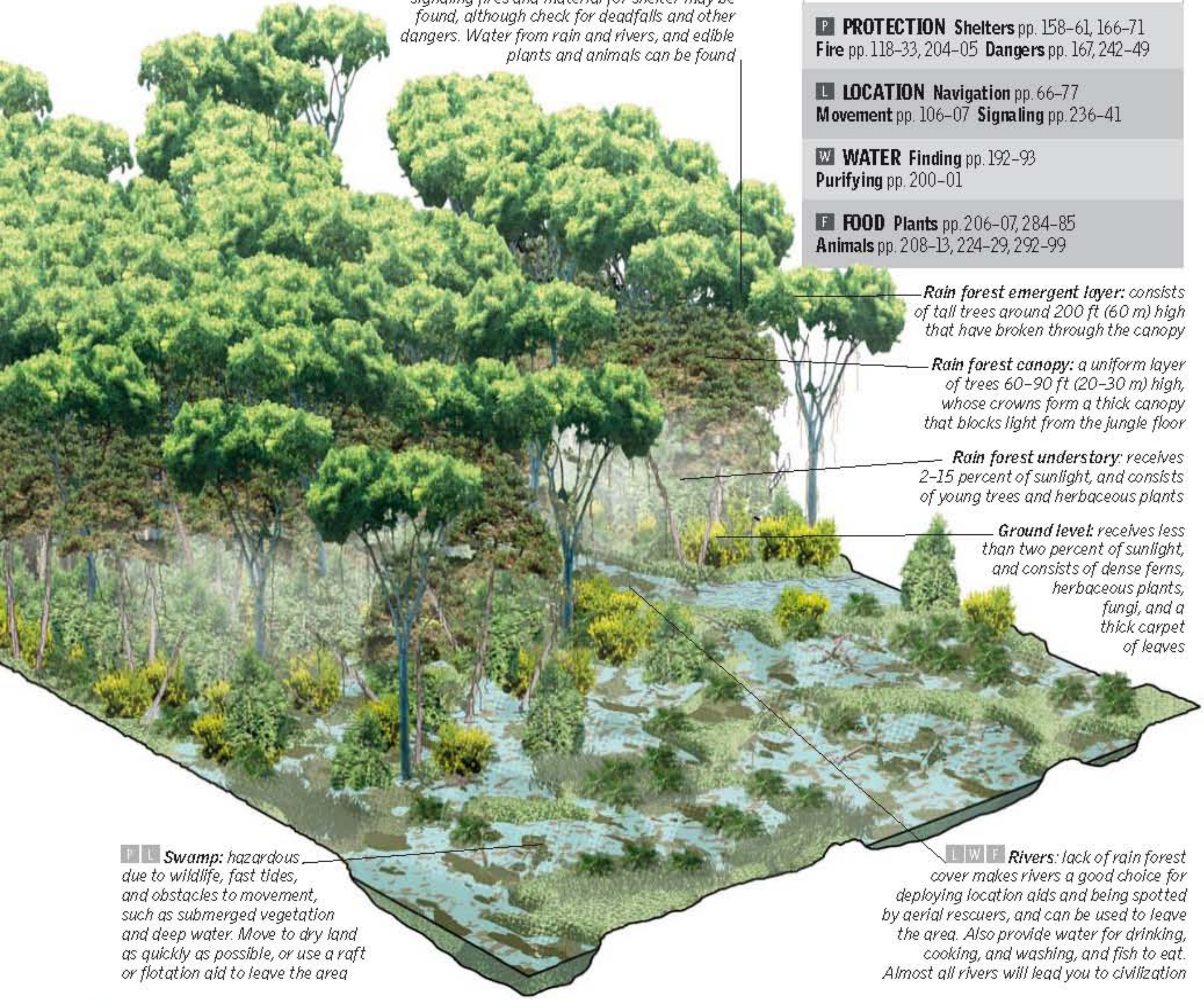

P L W F ***Rain forest:*** *fuel for cooking and signaling fires and material for shelter may be found, although check for deadfalls and other dangers. Water from rain and rivers, and edible plants and animals can be found*

Rain forest emergent layer: *consists of tall trees around 200 ft (60 m) high that have broken through the canopy*

Rain forest canopy: *a uniform layer of trees 60–90 ft (20–30 m) high, whose crowns form a thick canopy that blocks light from the jungle floor*

Rain forest understory: *receives 2–15 percent of sunlight, and consists of young trees and herbaceous plants*

Ground level: *receives less than two percent of sunlight, and consists of dense ferns, herbaceous plants, fungi, and a thick carpet of leaves*

P L ***Swamp:*** *hazardous due to wildlife, fast tides, and obstacles to movement, such as submerged vegetation and deep water. Move to dry land as quickly as possible, or use a raft or flotation aid to leave the area*

L W F ***Rivers:*** *lack of rain forest cover makes rivers a good choice for deploying location aids and being spotted by aerial rescuers, and can be used to leave the area. Also provide water for drinking, cooking, and washing, and fish to eat. Almost all rivers will lead you to civilization*

WHERE TO FIND TROPICAL AREAS

Tropical rain forests are clustered around the Equator. The largest rain forest areas are found in Central and South America, sub-Saharan Africa, southeast Asia, northern Australia, and several Pacific islands.

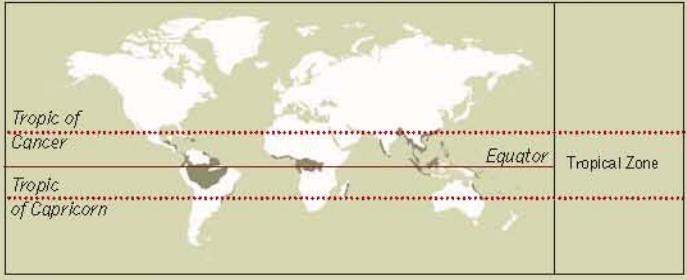

▨ Rain forest areas

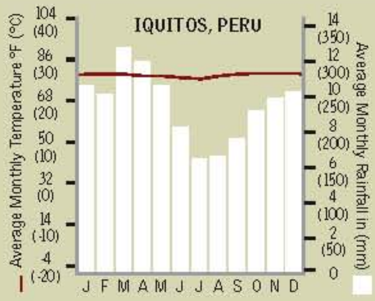

TYPICAL CLIMATE

Iquitos sits on the Amazon river in the heart of the Peruvian rain forest. It lies just south of the Equator and has a typical hot, wet tropical climate.

MOUNTAIN ENVIRONMENTS

THE MOUNTAIN ENVIRONMENT is one of the harshest places for humans to survive. Defined as landmasses with a summit above 2,000 ft (600 m), mountains can be very dangerous places due to their elevation and terrain. Lower temperatures and poor weather are more likely at higher altitudes, so there is a significant risk of hypothermia, frostbite, and altitude sickness, while snow, ice, and precipitous terrain present further hazards. Survival may depend on your ability to descend to areas with better prospects of survival and rescue.

MOUNTAIN FEATURES

Above the treeline, mountain terrain is typically barren, featuring bare rocks, gravel, boulders, and snow and ice, which are most extensive during the winter but may persist at high elevations during the summer. At lower levels, coniferous forest is the predominant feature, dissected by streams and rivers fed by snowmelt. Topographical features include scree slopes, cliffs, ravines, boulder fields, snowslopes, and glaciers.

ELEVATION CHANGE

Mountains of sufficient height may have vastly different ecosystems at different elevations, due to harsh conditions at high altitude. Plant and animal life varies at different elevations, so the natural resources available for survival will differ according to height. Mixed deciduous trees are common on low slopes and conifers grow at intermediate levels, but thin out to grass, isolated shrubs, moss, and lichen higher up. At the highest levels, high winds, frost, snow, and ice discourage any growth.

WARNING!
One of the most dangerous animals in mountain areas is the cougar, also known as the puma, panther, and mountain lion. Found in North and South America, more than 20 percent of attacks are fatal.

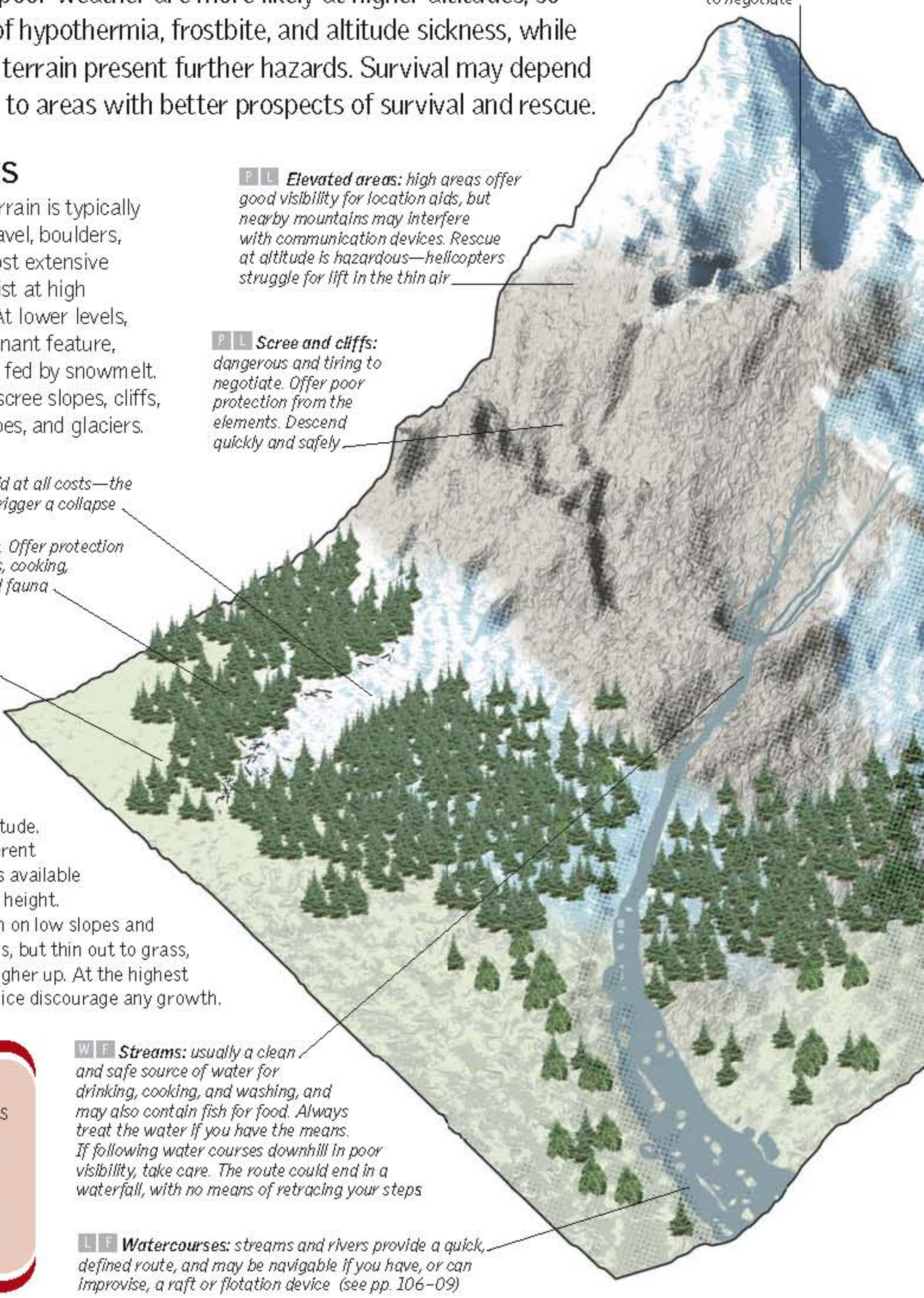

P L ***Mountains:*** *shelter is limited and dangers include unpredictable weather and rock falls. They may require special equipment and knowledge to negotiate*

P L ***Elevated areas:*** *high areas offer good visibility for location aids, but nearby mountains may interfere with communication devices. Rescue at altitude is hazardous—helicopters struggle for lift in the thin air*

P L ***Scree and cliffs:*** *dangerous and tiring to negotiate. Offer poor protection from the elements. Descend quickly and safely*

P ***Avalanche-prone slopes:*** *avoid at all costs—the smallest noise or movement can trigger a collapse*

P L F ***Forest areas:*** *rich in resources. Offer protection from the elements, material for shelters, cooking, fires, and signaling, and edible flora and fauna*

P ***Shelter:*** *always site your shelter away from areas of avalanche risk, especially if avalanche debris is present*

W F ***Streams:*** *usually a clean and safe source of water for drinking, cooking, and washing, and may also contain fish for food. Always treat the water if you have the means. If following water courses downhill in poor visibility, take care. The route could end in a waterfall, with no means of retracing your steps*

L F ***Watercourses:*** *streams and rivers provide a quick, defined route, and may be navigable if you have, or can improvise, a raft or flotation device (see pp. 106–09)*

SURVIVING IN THE MOUNTAINS

The prospects for survival are good at lower elevations, where trees provide material for shelter and fire, rivers provide water, and edible plants and animals are likely to be present. At higher elevations vegetation is scarce, so there are fewer options for shelter and food, although water from streams or snow is likely to be available. There may be a risk of avalanche on snow-covered slopes, and crevasses pose a hidden danger in glaciated areas. Cold-related injuries pose the greatest threat (see pp. 270–71).

HIGH-LEVEL HAZARDS

The primary threats to survival at high altitude are cold-related ailments, the most dangerous of which is hypothermia. This occurs when the body's core temperature falls below 95°F (35°C)—if allowed to drop as low as 86°F (30°C), it can be fatal. Frostbite may also occur in the extremities in freezing or cold and windy weather, and may lead to permanent tissue damage. Altitude sickness, which can cause pulmonary or cerebral edema—and in extreme cases, death—is also a risk above 8,000 ft (2,500 m).

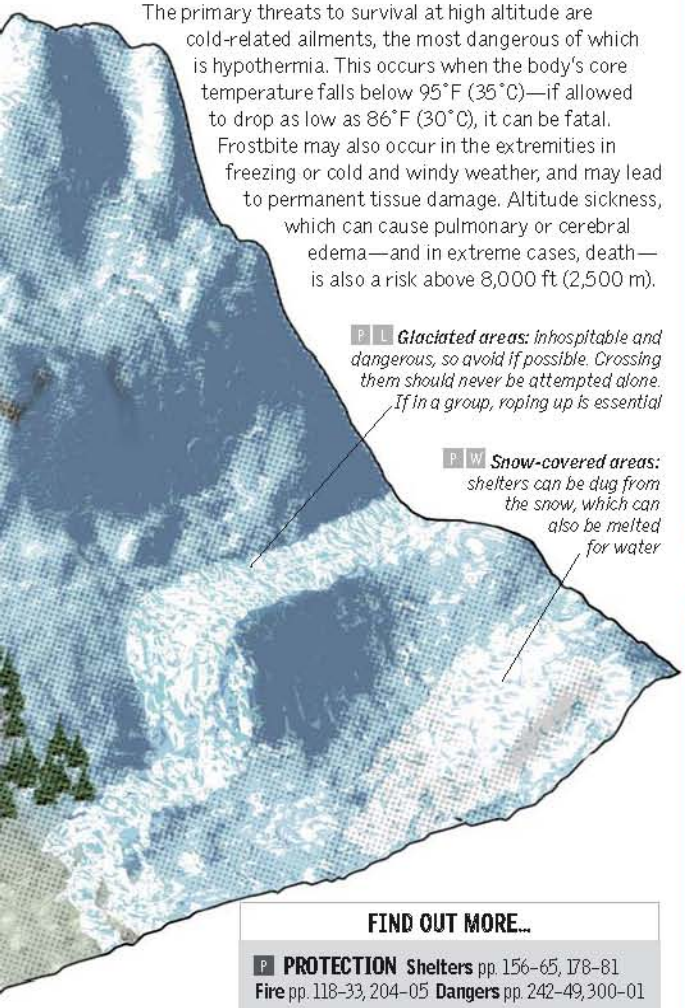

P L ***Glaciated areas:*** *inhospitable and dangerous, so avoid if possible. Crossing them should never be attempted alone. If in a group, roping up is essential*

P W ***Snow-covered areas:*** *shelters can be dug from the snow, which can also be melted for water*

FIND OUT MORE...

P PROTECTION Shelters pp. 156–65, 178–81 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–01

L LOCATION Navigation pp. 66–77 **Movement** pp. 90–91, 94–97 **Signaling** pp. 236–41

W WATER Finding pp. 188–91, 194–95 **Purifying** pp. 200–01

F FOOD Plants pp. 206–07, 280–81, 286–87 **Animals** pp. 208–13, 216–23, 290–97

MOUNTAIN DISTRIBUTION

Mountains are found on every continent, but the major ranges are the Himalayas and Karakoram in Asia, the Andes in South America, the Rocky Mountains in North America, and the Alps and Pyrenees in Europe.

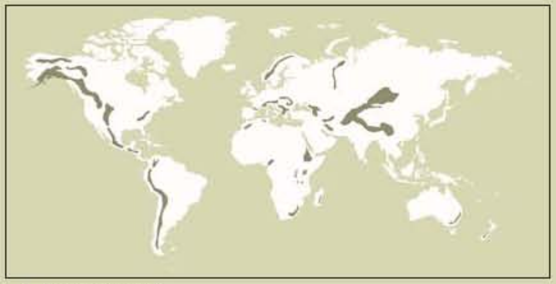

▨ Mountain ranges

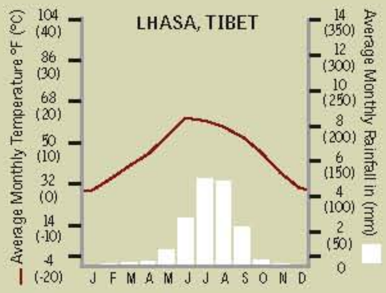

TYPICAL CLIMATE

Lhasa, the capital city of Tibet, lies at 11,972 ft (3,649 m) on the Tibetan Plateau, sometimes known as "the roof of the world." The high altitude means that winters are cool, and precipitation is relatively low because it mostly falls as snow, which has just 10 percent of the density of rain.

MOUNTAIN ESSENTIALS

High altitude and lower oxygen levels place higher-than-normal demands on even a fit body, so be prepared. Survival in the mountains is also tough on equipment.

- Respect nature and err on the side of caution—mountains are unforgiving, and rescue is unlikely to be quick or easy.
- Plan an achievable and safe route, and prepare an Emergency Plan of Action (see pp. 24–25).
- Dress in layers (see pp. 46–47). Start a walk lightly dressed (cold) and add or remove layers as necessary.
- Wear a hat (up to 70 percent of body heat is lost through the head) and gloves—cold fingers will hinder your functions.
- Tie loose items of clothing (hat, gloves, sunglasses) to your person—a lost glove could result in a lost hand, or life.
- Take a flashlight—weather changes and unforeseen problems may mean you are on the mountain in low light or darkness.
- Always carry basic equipment to give adequate protection if forced to camp overnight, such as a bivy sack and sleeping bag.
- Check avalanche warnings and carry an avalanche transceiver.
- Contact with snow makes clothing wet and less effective.
- If moving over snow, make a pair of improvised snow shoes (see pp. 94–95).

DESERT ENVIRONMENTS

THE DESERT ENVIRONMENT is a hostile place for human survival due to temperature extremes and the scarcity—or absence—of water. Most deserts are either hot and dry or cold and dry, and all typically receive less than 10 in (250 mm) of annual rainfall. They are areas of extremes, where heat exhaustion and hypothermia are threats, and flash floods may rapidly replace dry conditions. A harsh place to survive, the desert should only be entered by those who are prepared.

WARNING!

Desert areas are home to a range of venomous snakes, including the black mamba of eastern and southern Africa. It is particularly dangerous due to its aggressive nature and highly venomous bite.

DESERT FEATURES

The popular depiction of a desert is of a dry, sandy landscape with searing skies and temperatures among the highest in the world. Such areas do exist, but the desert environment has far greater diversity, from cold deserts in the polar regions to highlands, grassland, and broken, rocky terrain that contain "wadis" (seasonal stream beds).

HOT, DRY DESERTS

Temperate and tropical deserts are usually hot and dry due to high daytime temperatures, which leads to low precipitation and high evaporation. There may be months or even years between rainfall, which usually occurs in dramatic bursts but then quickly drains into seasonal stream beds (wadis), or evaporates either in the air or soon after touching the hot ground. Very little rain soaks into the ground, which results in sparse vegetation that has evolved to maximize water, usually via extensive root systems that collect moisture, waxy skin and leaves that prevent water loss, and stems or roots that hold water. At night, in winter, and at higher altitudes, freezing temperatures and frost may occur.

GRASSLANDS

Known as savanna in tropical areas, grasslands often border desert regions, and experience a similarly dry climate for most of the year. The key difference is that they also experience a wet season, which accounts for the more diverse and extensive vegetation. Temperate grassland features grasses, shrubs, and small pockets of trees, while savanna is typically more varied, with denser tree cover, bushes, and grasses that grow tall during the wet season but die back in periods of drought. Wildlife is more prevalent, including predators. Wildfires may spread during the dry season, and play an important role by destroying dead plant matter and adding fertile ash to the soil.

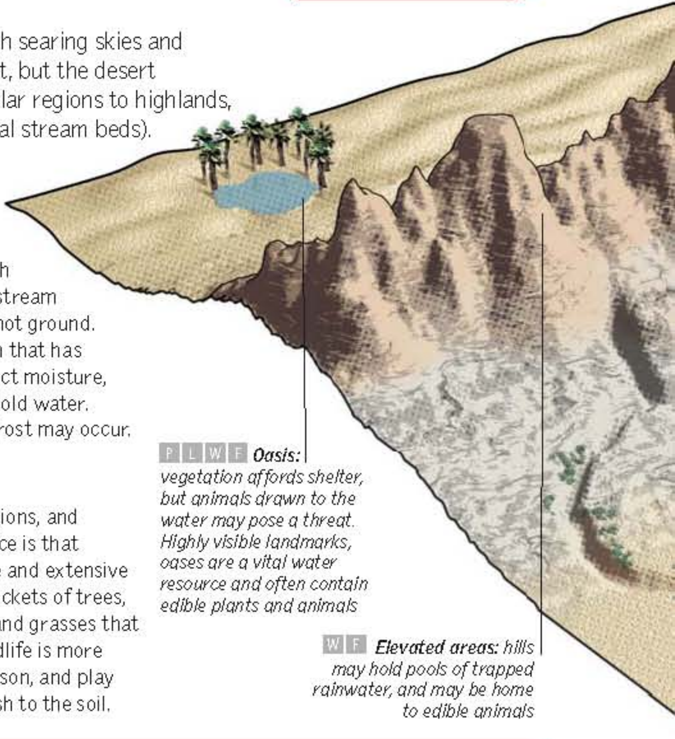

P L W F ***Oasis:*** *vegetation affords shelter, but animals drawn to the water may pose a threat. Highly visible landmarks, oases are a vital water resource and often contain edible plants and animals*

W F ***Elevated areas:*** *hills may hold pools of trapped rainwater, and may be home to edible animals*

WHERE TO FIND DESERT AREAS

Although deserts occur on every continent, the largest hot desert in the world is the Sahara in North Africa, which forms part of a desert band stretching through the Middle East and into south and central Asia.

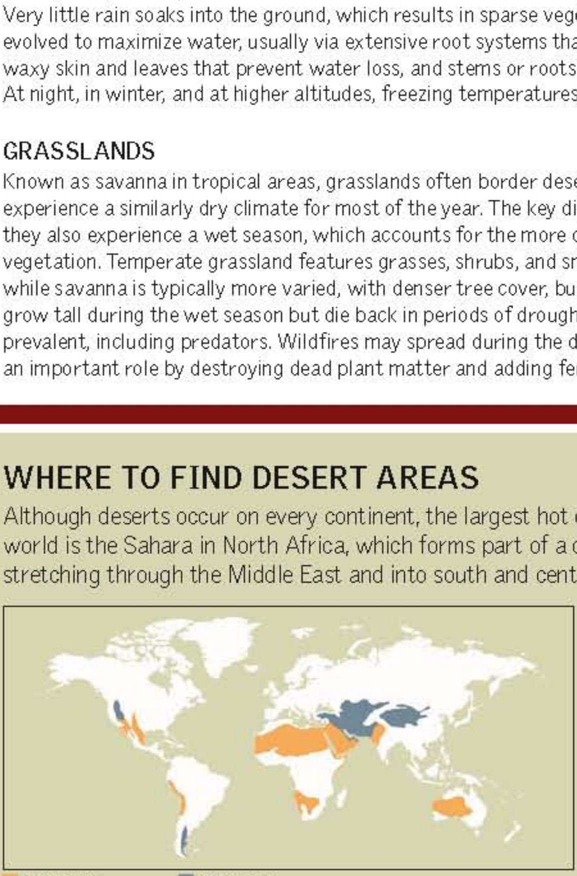

🟧 Hot desert
⬛ Cold desert

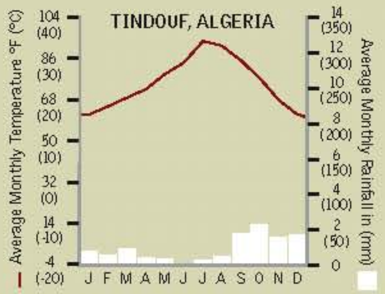

TYPICAL CLIMATE

Tindouf lies in the heart of the Sahara, and experiences a typical hot-desert climate. Precipitation is virtually zero all year round, while temperatures peak in the summer.

SURVIVING IN THE DESERT

Hostile temperatures and few natural resources limit chances of survival in the desert. Water and shelter are scarce, if not non-existent, edible vegetation is limited, and animals hide from the sun in the day. Wadis, grasslands, and higher ground all hold better prospects. The greatest dangers are dehydration and heat exhaustion (see pp. 272–73), although African savanna areas may be home to large, dangerous mammals. The main killers are big cats, hippos, rhinos, elephants, and crocodiles.

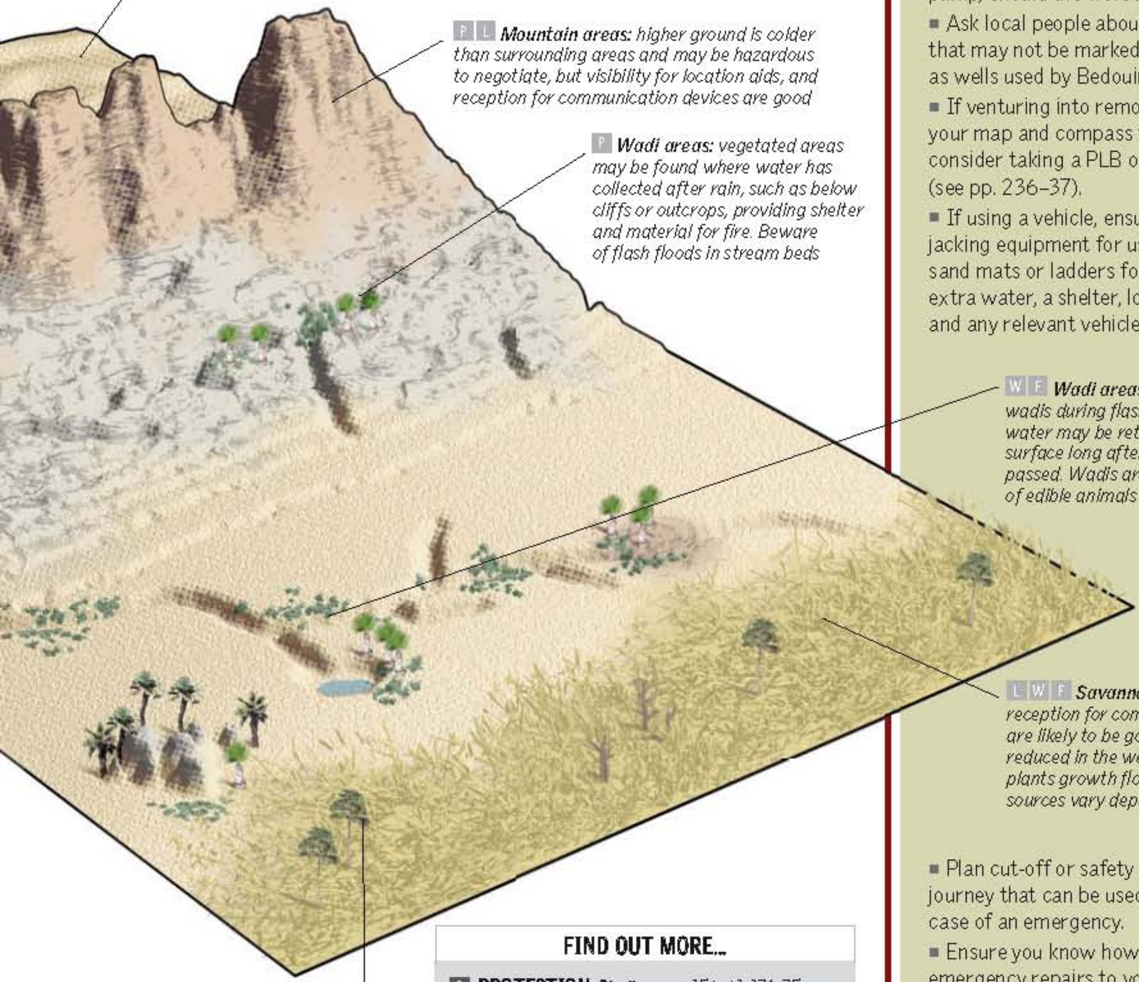

P L W ***Sand dunes:*** *shelter from the elements is poor, but visibility for location aids and for spotting rescuers is good. Observing converging animal and bird tracks or movements may indicate the direction to a water source.*

P L ***Mountain areas:*** *higher ground is colder than surrounding areas and may be hazardous to negotiate, but visibility for location aids, and reception for communication devices are good*

P ***Wadi areas:*** *vegetated areas may be found where water has collected after rain, such as below cliffs or outcrops, providing shelter and material for fire. Beware of flash floods in stream beds*

W F ***Wadi areas:*** *rivers flow in wadis during flash floods, and water may be retained below the surface long after the flood has passed. Wadis are also the habitat of edible animals*

L W F ***Savanna:*** *visibility and reception for communication devices are likely to be good but will be reduced in the wet season when plants growth flourishes. Water sources vary depending on the season*

P ***Savanna:*** *protection from the elements is limited, although material for building shelters may be found in the wet season. Sunstroke and dehydration are killers, so seek shade during the hottest parts of the day. Beware of snakes or scorpions in shaded areas, big cats and large mammals, and nocturnal animal activity*

FIND OUT MORE...

P PROTECTION Shelters pp. 156–61, 174–75 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49, 300–05

L LOCATION Navigation pp. 66–77 **Movement** pp. 100–01 **Signaling** pp. 236–41

W WATER Finding pp. 192–93 **Purifying** pp. 200–01

F FOOD Plants pp. 206–07, 282–83 **Animals** pp. 216–21, 224–29, 292–95

DESERT ESSENTIALS

Survival in the extreme conditions of the desert is impossible without full preparation. Consider the following:

- Always prepare an EPA to notify someone of your plans before entering a desert area (see pp. 24–25).
- Water is life. Do not underestimate your needs, carry extra just in case, and carry equipment to maximize your chances of finding and procuring more (binoculars, surgical tubing, filtration/purification pump) should the worst happen.
- Ask local people about water sources that may not be marked on maps, such as wells used by Bedouin tribes.
- If venturing into remote areas, augment your map and compass with a GPS, and consider taking a PLB or satellite phone (see pp. 236–37).
- If using a vehicle, ensure you carry jacking equipment for use in soft ground, sand mats or ladders for self-recovery, extra water, a shelter, location aids, and any relevant vehicle spares.
- Plan cut-off or safety points into your journey that can be used to divert to in case of an emergency.
- Ensure you know how to accomplish emergency repairs to your vehicle.
- Ask local people about known dangers, such as impassable roads, misleading tracks, and soft-sand areas.
- When using a ladder for self-recovery, tie it to the back of the vehicle so that it's dragged behind once crossed, and can be retrieved when you reach solid ground.
- If your vehicle breaks down, leave it only if staying is no longer safe or feasible—rescuers will search for it first.

COLD ENVIRONMENTS

LOW TEMPERATURES are a potentially lethal hazard, but as long as you can conserve body heat, survival in the cold is quite possible. The ice-covered Arctic and Antarctic polar regions, sub-polar tundra and taiga, as well as parts of most temperate zones during the winter, can all be classed as cold environments. Wet, windy weather can drive moderately cold temperatures down further, increasing the risk of the most lethal cold-weather condition—hypothermia.

WARNING!
The largest predatory land mammal on Earth, the polar bear inhabits the Arctic. Attacks are most likely when the bears are hungry, and are usually fatal.

COLD ENVIRONMENT FEATURES

Outside temperate areas, the typical cold environments are the polar areas, which feature glacial ice sheets and sea ice; tundra, which contains permafrost (permanently frozen soil), small shrubs, mosses, and lichens; and taiga, with extensive conifer forests. Almost half of the northern hemisphere's landmass can be classified as a cold region, due to the influence of cold air masses (see pp. 78–83), while ocean currents and altitude can also have a cooling effect.

POLAR WASTES

The northern polar area is situated over the Arctic Ocean, so largely consists of sea ice, the extent and depth of which varies throughout the year. Because of the lack of land, melting ice is a serious danger. The southern polar area is located over the continent of Antarctica, and contains the largest glacial ice sheet on Earth. Both areas contain ice shelves—glaciers that have extended over the sea—parts of which break off to form icebergs.

TUNDRA AND TAIGA

Latitudes that neighbor the polar areas are home to tundra, an environment of permafrost and vegetation that is stunted due to low temperatures. Taiga is located further away from the poles, where temperatures are high enough to support coniferous forests.

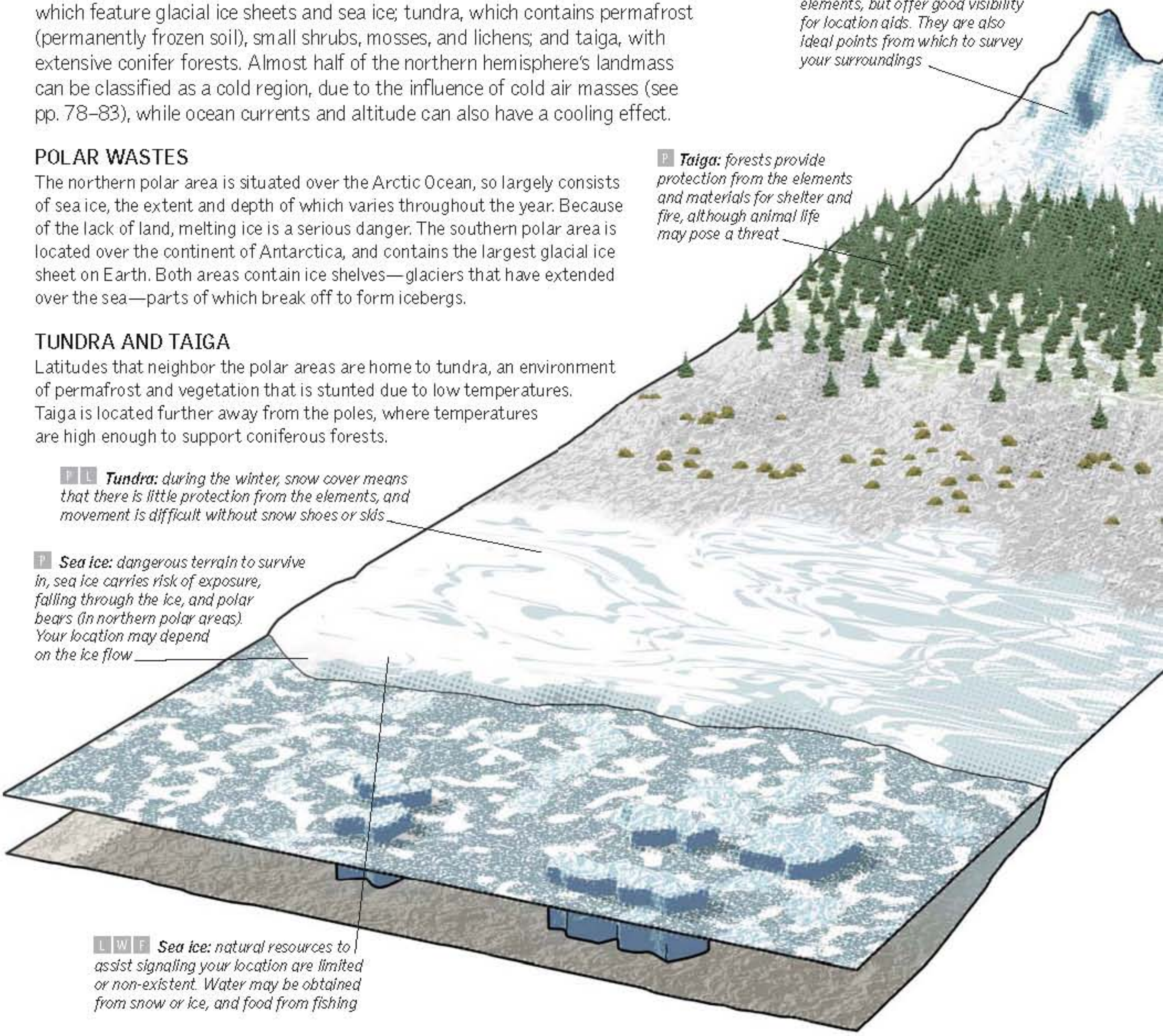

P L ***Mountains:*** *elevated areas give little protection from the elements, but offer good visibility for location aids. They are also ideal points from which to survey your surroundings*

P ***Taiga:*** *forests provide protection from the elements and materials for shelter and fire, although animal life may pose a threat*

P L ***Tundra:*** *during the winter, snow cover means that there is little protection from the elements, and movement is difficult without snow shoes or skis*

P ***Sea ice:*** *dangerous terrain to survive in, sea ice carries risk of exposure, falling through the ice, and polar bears (in northern polar areas). Your location may depend on the ice flow*

L W F ***Sea ice:*** *natural resources to assist signaling your location are limited or non-existent. Water may be obtained from snow or ice, and food from fishing*

SURVIVING IN THE COLD

Natural resources may be scarce in cold environments, so your survival is likely to depend on the equipment and supplies you have with you. Shelters can be dug from the snow to help you stay warm and avoid exposure to freezing temperatures, but material for making fire is limited in polar and tundra areas. Survival is more feasible in the taiga, where wood is available, and fresh water and edible flora and fauna can also be found. The greatest dangers are hypothermia (see pp. 272–73) and—in northern polar areas—polar bears.

LOW-TEMPERATURE DANGER

Heat loss occurs in several ways, and minimizing it is essential to survival in cold conditions. It's important to avoid over-exertion, since heat is lost through sweating, and to cover your head and wear layered clothing to trap radiated body heat. Heat is also lost through conduction—direct contact between the skin and a cooler surface—particularly when wet, since water conducts heat away from the body 25 times faster than air. For this reason, it's vital to wear insulated and waterproof clothing, and to stay dry. Respiration also involves heat loss—each inhaled breath contains cold air, which is warmed as it passes through the lungs, and laden with moisture when exhaled. Reduce heat loss by covering your nose and mouth with a ski-mask or scarf, or by breathing warmer air next to a fire or in a warm shelter.

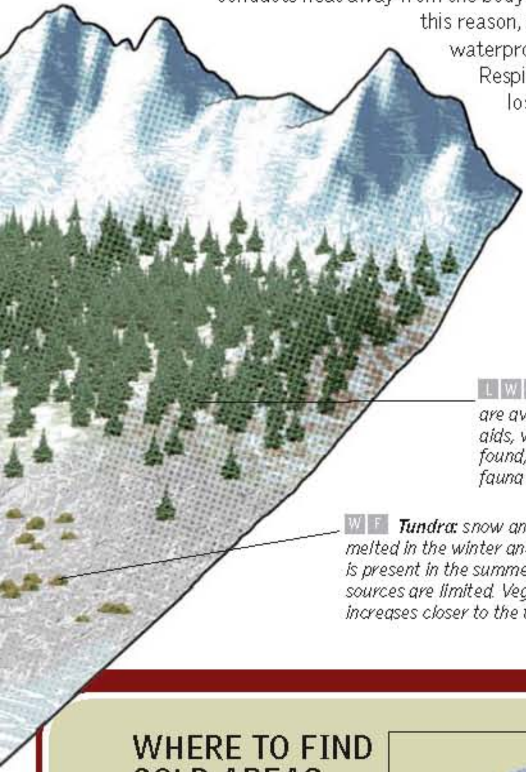

L W F *Taiga: natural resources are available for making location aids, water sources may be found, and edible flora and fauna are likely to be present*

W F *Tundra: snow and ice can be melted in the winter and ground water is present in the summer, although sources are limited. Vegetation increases closer to the treeline*

COLD ESSENTIALS

The main threats in cold environments are hypothermia and exposure, so ensure you are fully prepared:

- Dress in loose-fitting layers of clothing (see pp. 46–47), avoid overheating, and ensure your clothing stays dry and clean.
- If your hands are cold, don't warm them with your breath, which will make them wet—tuck them under your armpits.
- Regularly check your extremities (face, toes, hands, and ears) for frostnip, the first stage of frostbite (see p. 273).
- Wind-chill is dangerous, so take shelter from the wind at every opportunity, particularly if you are in a survival situation.
- Always ensure that your shelter is well ventilated—keep vent holes clear and check regularly, especially when snow falls.
- Get off the cold ground, snow, or ice—sit on your pack or make a sleeping platform using boughs to avoid losing body heat.
- If fire is your primary means of warmth, triple the amount of firewood you think you need—you will need enough to last the night.

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 156–65, 178–79 **Fire** pp. 118–33, 204–05 **Dangers** pp. 242–49

L **LOCATION** **Navigation** pp. 66–67 **Movement** pp. 94–97 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 194–95 **Purifying** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 286–87 **Animals** pp. 208–13, 216–23, 290–91, 296–97

WHERE TO FIND COLD AREAS

The Arctic and Antarctic polar areas are the coldest places on Earth, and are found at the furthest points from the Equator. Tundra exists next to polar regions, and is bordered by taiga, which is found in northern Eurasia and North America.

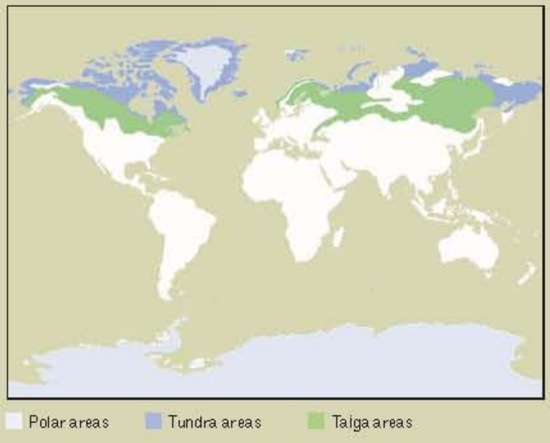

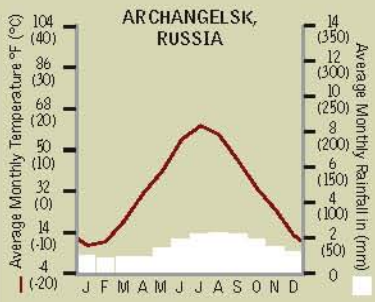

TYPICAL CLIMATE

Archangelsk is a Russian port on the Barents Sea. Located in the taiga zone, it experiences below-zero temperatures for half the year and low precipitation.

MARINE ENVIRONMENTS

POTENTIALLY THE MOST HARSH and challenging of all environments, the sea is unlike any other due to the lack of the most essential requirement for human life—fresh water. Around 70 percent of the Earth's surface is covered by saltwater, ranging from the freezing, wind-torn polar oceans to the warm tropical seas. Winds and currents govern your movement, protection is limited, and natural resources are virtually nil, so reaching land is your best hope for survival.

WARNING!
There are sharks in every ocean, but those in tropical waters are the most aggressive. Around 20 of the hundreds of species have attacked humans, and the most dangerous types are the great white, tiger, and bull sharks.

MARINE FEATURES

The marine environment ranges from the coastal margins, which are shallow enough— 650 ft (200 m) or less—to support a huge array of animal and plant life, to the open oceans, the deepest of which plunge to more than 33,000 ft (10,000 m).

SHELTERED COAST TO OPEN SEA

The coastal margins are home to the majority of ocean life, and include terrestrial areas—the shoreline and intertidal zone—and aquatic—the sea up to the edge of the continental shelf. A variety of environmental features are found, such as sand, rock, and pebble beaches, dunes, cliffs, estuaries, mud flats, mangroves, lagoons, kelp forests, and coral reefs. The open ocean is a more barren environment that supports less life, and is notable for its vast extent, comprising 92 percent of all saltwater. Survival depends on your equipment, supplies, and ingenuity.

OCEAN CONDITIONS

Your protection and location at sea depends largely on the weather—you may need to deal with anything from exposure to the hot sun, to cold, wet, windy conditions and high seas. Conditions are toughest in the South Atlantic Ocean, where winds (known as the "Roaring Forties") are strong all year round, and sea ice is extensive in the winter. Severe seasonal storms occur: hurricanes in the tropical areas of the western Atlantic, typhoons in the western Pacific. India and Southeast Asia also experience severe weather during the monsoon seasons.

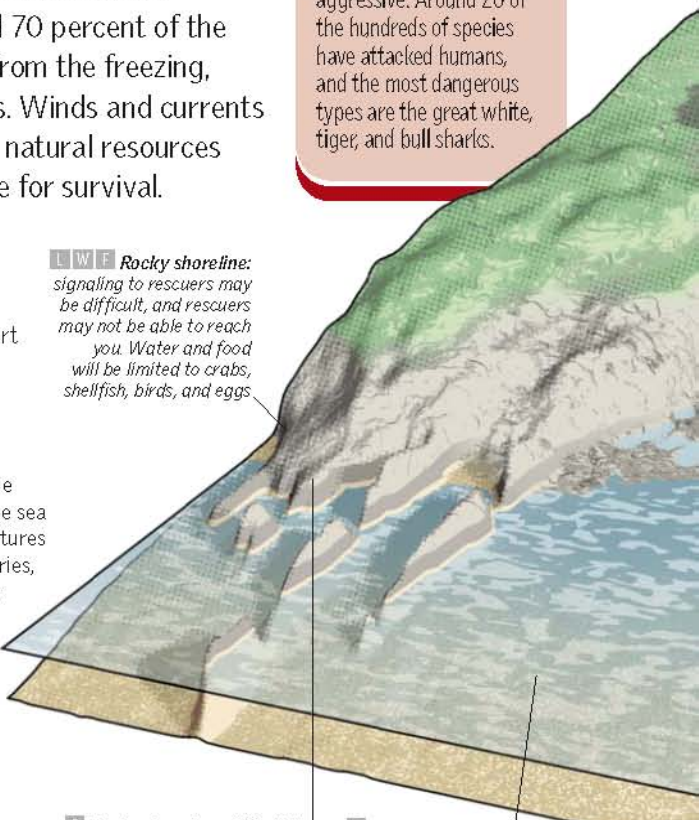

L W F ***Rocky shoreline:*** *signaling to rescuers may be difficult, and rescuers may not be able to reach you. Water and food will be limited to crabs, shellfish, birds, and eggs*

P ***Rocky shoreline:*** *difficult to make landfall and to survive on, and poses risk of injury when landing. Progress to safer areas may be difficult or impossible due to cliffs or slippery rocks. Protection from the elements and tides is limited*

F ***Coastal margins:*** *the majority of ocean life is found near the coasts, and most species live within 650 ft (200 m) of the surface*

WINDS AND CURRENTS

Winds and currents follow prevailing patterns, so if you know roughly where you are, you can figure out where they'll carry you. With a drift of just 1 knot, a liferaft can move 24 miles (38 km) from its last known position in 24 hours. It may also drift up to 35° either side of the wind, further widening the search area to more than 400 sq miles (1,000 sq km) in the first 24-hour period, and increasing exponentially over time. Conversely, favorable currents can also be used to help you drift to landfall faster.

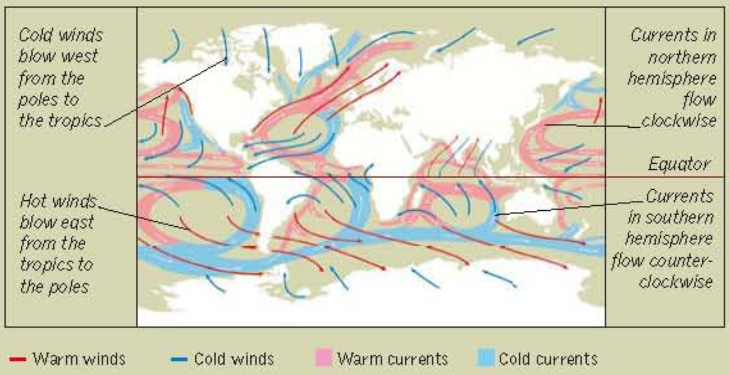
Cold winds blow west from the poles to the tropics
Hot winds blow east from the tropics to the poles
Currents in northern hemisphere flow clockwise
Equator
Currents in southern hemisphere flow counter-clockwise
Warm winds
Cold winds
Warm currents
Cold currents

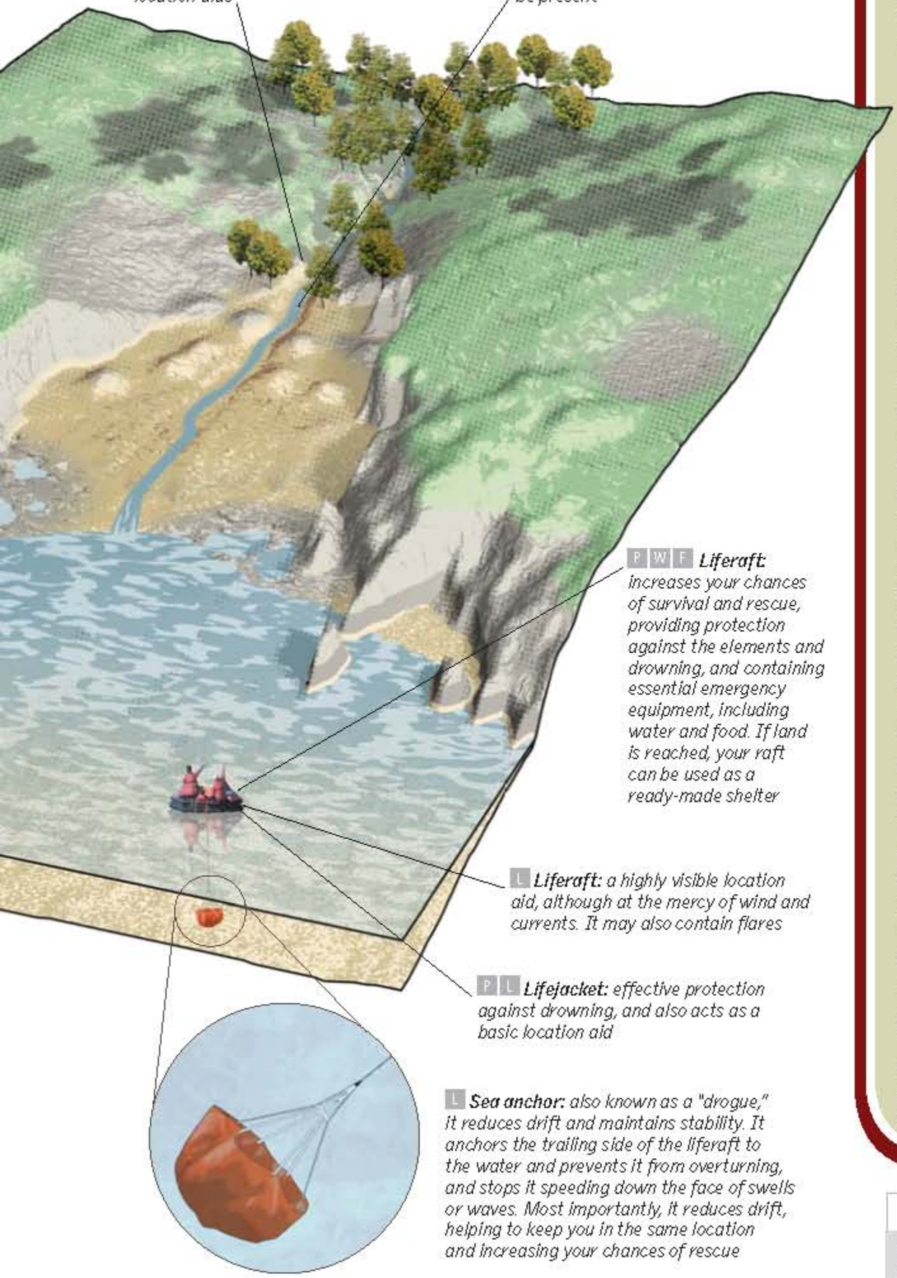

P L ***Beach:*** *above the high-tide mark, beaches may provide protection from wind and rain, and materials for building shelters, lighting fires, and improvising location aids*

W F ***Beach stream:*** *streams may drain on to beaches, providing water for drinking, cooking, and washing. Edible plants and animals may be present*

P W F ***Liferaft:*** *increases your chances of survival and rescue, providing protection against the elements and drowning, and containing essential emergency equipment, including water and food. If land is reached, your raft can be used as a ready-made shelter*

L ***Liferaft:*** *a highly visible location aid, although at the mercy of wind and currents. It may also contain flares*

P L ***Lifejacket:*** *effective protection against drowning, and also acts as a basic location aid*

L ***Sea anchor:*** *also known as a "drogue," it reduces drift and maintains stability. It anchors the trailing side of the liferaft to the water and prevents it from overturning, and stops it speeding down the face of swells or waves. Most importantly, it reduces drift, helping to keep you in the same location and increasing your chances of rescue*

SURVIVING AT SEA

Natural resources in the open ocean are virtually nil, so improvising shelter or location aids is limited to what you have with you. Seasickness, especially in a small liferaft, can seriously affect your ability to remain both hydrated and motivated. Unless you can reach land, your only drinking water will come from rain or mechanical desalination devices—consuming urine or saltwater will only increase your rate of dehydration. Reaching land or shipping lanes greatly increases your chances of rescue.

MARINE ESSENTIALS

Your chances of survival at sea will be greatly enhanced by good preparation, so consider the following:

- Pack sea-survival equipment as if your life depends on it—it might! Other environments allow you to improvise survival and location aids, but not the sea.
- Take emergency immersion-survival suits. They will protect you against the elements and aid flotation, and their color (high-visibility orange or red) is an aid to location.
- Always carry a Personal Locator Beacon (PLB), just in case.
- Take several means of obtaining water, such as a reverse-osmosis pump, which turns saltwater into drinking water, a saltwater desalination kit, a solar still, and rain catchers.
- Always keep a pre-packed emergency bag handy. It should contain; first aid kit, reverse-osmosis pump, emergency beacon, solar still, water, flares, fishing kit. (see pp. 250).
- If your vessel sinks, do not abandon it until absolutely necessary. Stay nearby unless it is unsafe to do so—anything that gives you a bigger "survival footprint" will increase your chances of being rescued, and you may be able to retrieve provisions from the vessel.
- If imminent rescue is unlikely, you will need to ration your water. Consider how much you have, how much can be procured, the chances of seasickness (which leads to dehydration), and the likelihood of being rescued later on.
- Protect yourself from the elements (sun, wind, cold, heat, and salt) as well as you can—prevention is better than cure.
- The sea, especially in rough conditions, can sap the will to live quicker than any other environment. Seasickness is literally draining, resulting in lost fluid, energy, and motivation. Take anti-seasickness tablets at the earliest opportunity—even before abandoning ship—so that you can keep them down, and keep taking them as recommended.

FIND OUT MORE...

P **PROTECTION** **Shelters** pp. 156–61, 176–77 **Fire** pp. 118–33, 204–05 **Dangers** pp. 250–55, 300–01

L **LOCATION** **Movement** pp. 106–07 **Signaling** pp. 236–41

W **WATER** **Finding** pp. 196–97 **Treating** pp. 200–01

F **FOOD** **Plants** pp. 206–07, 288–89 **Animals** pp. 208–13, 224–29, 292–99

GET THE RIGHT GEAR

Whether you're equipping yourself for a lengthy expedition or just a day trip, it's important to choose the correct equipment and clothing. The wrong time to realize that your gear isn't up to the job is when your life depends on it doing what it was designed to do.

When selecting gear, it's always a good idea to work backward from a worst-case-scenario survival situation. Think about what you would need to survive if the worst happened. This is your "first-line" equipment, which you should carry with you at all times. It should enable you to address the basic principles of survival relevant to the environment you are in, and consists of the clothing you would wear and your basic survival equipment—your survival tin and belt-order (see pp. 43, 61).

Once you have organized this basic equipment, you can then decide what pieces of essential gear would make your expedition enjoyable rather than just survivable. This will probably consist of a bivy or tent, a sleeping bag, cooking equipment and food, and something to carry it all in.

In this section YOU WILL DISCOVER...

- the **meaning** of the term "**bomb-burst gear**"...
- that it's always **good to share** when you're **in a group**...
- which **backpacks** are the **best backpacks** to pack...
- why you should **hike around the house** to **break in your boots...**
- a good reason to take a **three-season tent...**
- how to **sleep** like a **mummy...**
- what **gorp** is and **when** you should **eat it...**

It's at this stage that you should conduct a short trial to check that your chosen gear will fit in your pack and that you can comfortably carry it.

Finally, any spare capacity can be utilized for non-essential "luxury" items, such as an inflatable mattress and pillow, an MP3 player, or a book. Double-check that your gear works and that you know how to use it properly. The more you understand how and why a piece of equipment works, the better able you will be to improvise should it be damaged or lost.

With each venture into the wilderness, your confidence in your abilities and that of your equipment will increase. As it does, you may want to explore the world of ultralight camping.

ULTRALIGHT CAMPING refers to the careful selection of your three largest items of gear to dramatically reduce the load you have to carry. A standard tent, pack, and sleeping bag have a combined weight of approximately 20 lb (9kg). Ultralight camping has the potential to halve the weight of your "big three."

- Take a hooped bivy instead of a tent and choose a low season-rated sleeping bag (wear your day clothes inside the bag for extra warmth).
- You can then reduce the size of your pack because your tent and sleeping bag are smaller.

Trekking with less equipment but still staying safe allows you greater freedom to enjoy the wilderness—carrying less weight puts less strain on your body, enabling you to cover greater distances. Ultralight camping can become addictive and is safe if you use common sense—but never compromise on the quality of equipment. Try:

- Using a hexamine stove to cook food and boil water—you won't have to carry a stove and fuel.
- Taking one pot to cook food and boil water.
- Using natural materials as a mattress.

> “The **wrong time** to realize that **your equipment** isn't up to the **task in hand** is when **your life depends** on it doing what it was **designed to do**”

CHOOSING YOUR GEAR

FORWARD PLANNING IS ESSENTIAL when you're deciding on what equipment to take on your trip. You'll need to assess your own personal requirements, the likely demands of weather and terrain, and the amount of gear you'll actually be able to transport. With a little forethought, you'll be as well equipped as you can be for any situation.

PACKING FOR YOUR TRIP

You will first need to weigh up your particular kit requirements against the limitations of your chosen mode of transport (see below). It's then vital to organize and prioritize your gear (see opposite) to ensure that any items you may need in a survival scenario are always close at hand, should the worst happen.

KEY PACKING TIPS

- Pack your gear in reverse order: the things you want to get to first should be the last things you pack.
- Stash items you may need during the day in side or lid pockets, for easy access.
- Coil and tape long straps so that they aren't a snagging hazard.
- Use a dry bag inside your backpack.
- Ensure that all pockets are fully closed and zipped.
- Air out anything packed wet as soon as you stop for the day.

HOW MUCH TO PACK

The environment of the region you're traveling to will dictate the type of equipment you will need to take with you, but your proposed mode of transport is the main constraint on the quantity of kit you can take (see chart, right). The more gear you have to carry yourself, the fewer luxuries you can take: the added weight will make traveling uncomfortable, use up too much valuable energy, slow your progress, and also limit the distance you can travel each day. It's also worth considering whether you can reduce the weight and volume of the kit items you're taking (see box, below).

LIGHTENING THE LOAD

When you're going to be traveling on foot, it's important to get rid of all the excess weight and volume you can, even if it's only by a small amount. The larger and heavier your backpack, the harder you will find the journey. If you have already ruthlessly pared your gear down to the bare essentials, try to reduce the weight or volume of the kit itself. First concentrate on the big three items: tent, sleeping bag, and backpack. Consider using a hooped bivy or tarpaulin rather than a tent; look for a lighter sleeping bag—some can weigh as little as 2lb (0.9 kg); and consider buying a backpack with a lighter frame.

WEIGHT AND SPACE RESTRICTIONS

ON FOOT	Backpacks (see pp. 44–45) are available in a wide range of volumes—from 30 liters (a day pack) to 80 liters (suitable for a trip of a week or more). Bear in mind that you'll have to carry whatever gear you take with you, and that the weight of the pack will determine what you're realistically able to achieve each day in terms of distance and activity. Pack only the bare minimum: there's no room for luxuries when traveling on foot.
MOUNTAIN/ TOURING BIKE	Pannier bags for bicycles are designed to fit on front, rear, or side pannier racks, and are available in a range of shapes, sizes, and capacities (from 5 to 50 liters). Larger bags will provide you with enough space to pack essentials, such as tent, sleeping bag, cooking equipment, and spare clothes, but you'll be restricted by the weight, and also by the need to carry cycle-maintenance gear, such as spare tires, inner tubes, cables, and tools.
MOTORCYCLE	Side panniers have a capacity of up to 80 liters and tank bags up to 25 liters, and weight is not a major issue. However, you'll need to take extra protective equipment, maintenance tools (including spare tires), and, in some cases, spare fuel, all of which will impact on available space.
PACK ANIMAL	Ideal for rough terrain, such as mountains or desert, pack animals can carry serious weight. A horse can transport up to 20 percent of its body weight; a camel can carry a weight of up to 990 lb (450 kg)—although 330 lb (150 kg) is more common—but you'll need to factor in the animal's food and water.
VEHICLE	The size of the vehicle will determine the amount and the weight of gear you can take with you, but you shouldn't have to compromise; extra equipment can be packed in peli boxes—waterproof storage containers with a hard, plastic exterior—that can be strapped to the outside of the vehicle. The main issue will be the weight-versus-fuel ratio: the heavier the vehicle, the more fuel it will use.
CANOE/ KAYAK/BOAT	Some kayaks and canoes are designed to bear loads of up to 660 lb (300 kg), including the craft's occupant(s), so you should be able to take all of your essential gear, plus a few luxuries. Given your proximity to the water, keeping your gear dry is the main priority. Weight and space restrictions don't apply for larger vessels.

PRIORITIZING YOUR KIT

However much gear you decide to take, you should group individual items into one of three categories—first-line, second-line, and third-line—according to their importance to your survival in a worst-case scenario. It's very unlikely that you'll be able to carry all of your equipment all of the time, and in an emergency situation, you may not have time to gather it all together. However, by prioritizing your gear beforehand, you'll have all your essential items on your person, or close enough to grab at a second's notice.

FIRST-LINE GEAR

Known as "bomb-burst gear" in the military, first-line gear (see right) is your basic survival equipment. If something goes wrong and you have to abandon the bulk of your gear to avoid injury, or death, what you're standing in is all you'll have to help you survive. First-line gear should therefore comprise crucial items of outdoor clothing, along with essential items for navigation and safety. Your bushcraft knife, firesteel, and belt pouch make up your belt order (see pp. 60–61). You'll need to risk-assess your particular situation and adapt your kit priorities accordingly, as conditions change, meaning that your first-line gear may alter as the day progresses. Bear in mind that if a removed layer of clothing goes into your backpack, it's no longer part of your first-line kit.

SECOND-LINE KIT

Second-line gear is everything you would need to stay safe on a daily basis under normal conditions. It can either be packed in a small day pack, or contained in pouches on a belt, the idea being that you carry it at all times. For example, a climber may decide to leave the bulk of his or her equipment (third-line gear) while attempting to reach the summit of a peak. By proceeding with first- and second-line gear, the climber is sacrificing equipment for weight and speed, but is still safe, as they have the essentials on them. Examples of second-line gear can include:

- A spare set of clothes, a bivy sack, and cordage
- Emergency rations and first-aid kit (see p. 260)
- Hexamine stove (see p. 58–59) and items to make a hot drink
- Matchless fireset (see p. 118), and a metal cup

THIRD-LINE GEAR

Also known as your "sustainment load," third-line gear is essentially the equipment you need to keep yourself alive and to function for longer periods of time. How much third-line gear you have will ultimately depend on your mode of transport and the amount of equipment you can carry. Examples of third-line gear include:

- A form of shelter—a tent or a tarpaulin
- Cooking utensils, such as a stove or cooking pot
- A backpack
- Food stores
- A sleeping bag and sleeping mat
- Any large water containers or hydration packs
- Wash kit and sanitary items

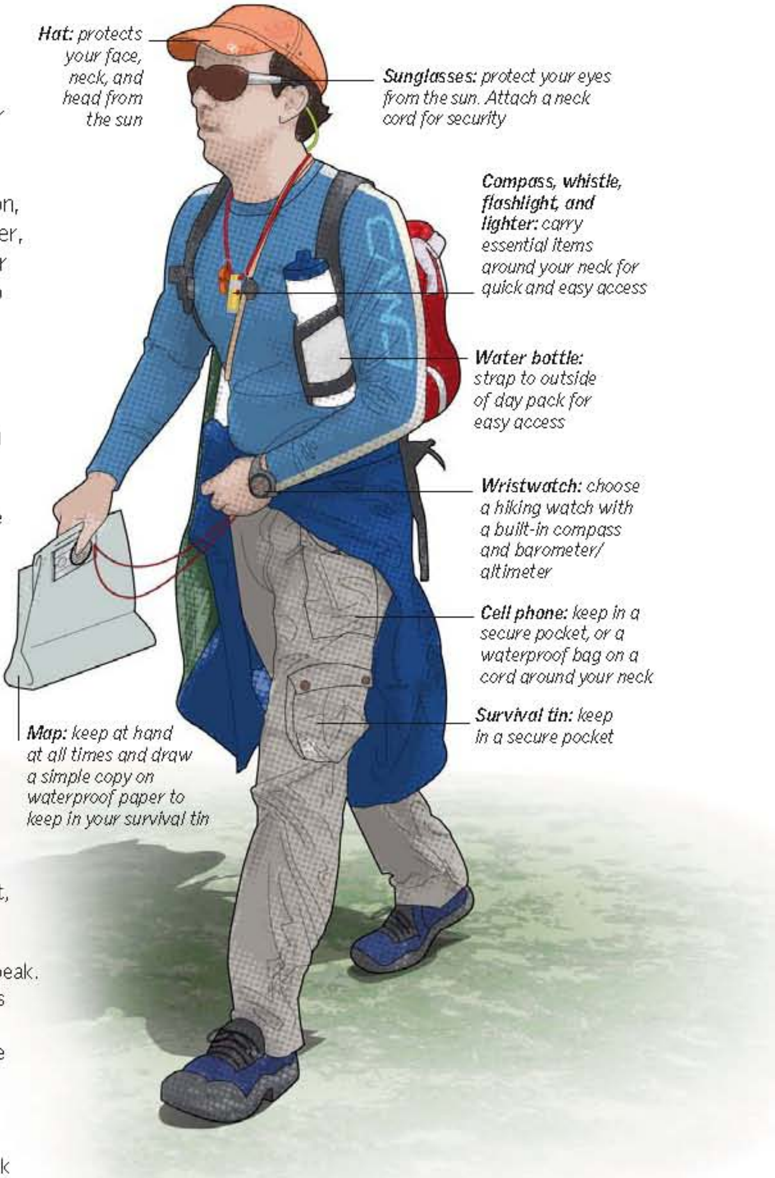

Hat: *protects your face, neck, and head from the sun*

Sunglasses: *protect your eyes from the sun. Attach a neck cord for security*

Compass, whistle, flashlight, and lighter: *carry essential items around your neck for quick and easy access*

Water bottle: *strap to outside of day pack for easy access*

Wristwatch: *choose a hiking watch with a built-in compass and barometer/altimeter*

Cell phone: *keep in a secure pocket, or a waterproof bag on a cord around your neck*

Survival tin: *keep in a secure pocket*

Map: *keep at hand at all times and draw a simple copy on waterproof paper to keep in your survival tin*

FLASHLIGHTS

A flashlight is a piece of first-line gear, even if you don't intend to be out in the dark, because situations can change unexpectedly. Modern flashlights use LEDs, and are small, lightweight, powerful, and very efficient. Opt for a "hands-free" headlamp and tape a spare set of batteries to your head-band. Also include two miniature Photon (LED) flashlights in your survival kit (see pp. 60–61).

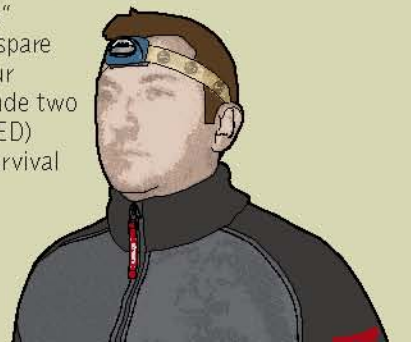

CHOOSING A BACKPACK

BACKPACKS ARE DESIGNED to enable hikers to carry large loads comfortably. They are made from a variety of materials, most of them waterproof. However, you should always use a separate waterproof bag inside your backpack to ensure your sleeping bag and spare clothes are kept dry; they are of little use in a survival situation if they get wet.

CARRY ONLY WHAT YOU NEED

Regardless of your pack size, you are the one who has to carry it. If in a group, many communal items can be shared. Try not to double up on basics (a party of four does not need four tubes of toothpaste, for example).

WHAT DO YOU NEED?

Your choice of backpack will depend on a number of factors. How long is your trip? How much do you intend to carry? What type of gear will you be taking with you? The answers to these questions will determine the size of pack—known as the "capacity"—that is right for you. Pack sizes range from 30 liters (a day pack) to 80 liters or more (suitable for trips lasting a week or more).

BACKPACK FRAMES

If you have chosen a medium- to large-sized pack, you will have to choose between an internal- or an external-frame pack.

INTERNAL-FRAME PACKS

Internal-frame packs are narrower in profile than their external-frame counterparts and tend to be longer, with either one or two internal compartments and very few pockets on the outside. The frame is integrated within the pack and usually consists of stays or flat bars, about 1 in (25 mm) wide and 1/10 in (3 mm) thick, made from plastic. Straps on the outside allow you to compress the pack, which prevents the items inside the pack from shifting and throwing you off-balance.

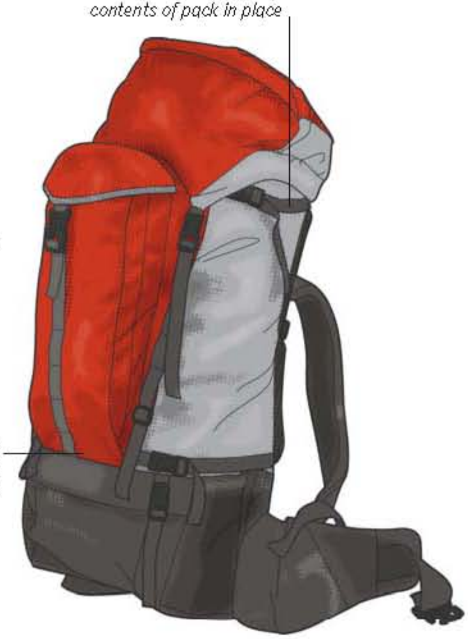

Compression straps reduce bulk and keep contents of pack in place

Additional compartment for quick access to essential items, such as wet-weather gear

EXTERNAL-FRAME PACKS

An external frame houses the pack on a lightweight tubing framework. They are great for heavy loads, as the pack sits more squarely on your hips. They are also cooler to carry than internal-frame packs, as the air can circulate between your back and the pack. External-frame packs usually have more pockets on the outside, which allows you to pack your items by category, rather than packing everything into one or two compartments. This makes locating items easier and allows you to unpack only the compartment you need, rather than the entire pack. Chest and waist straps ensure a comfortable fit.

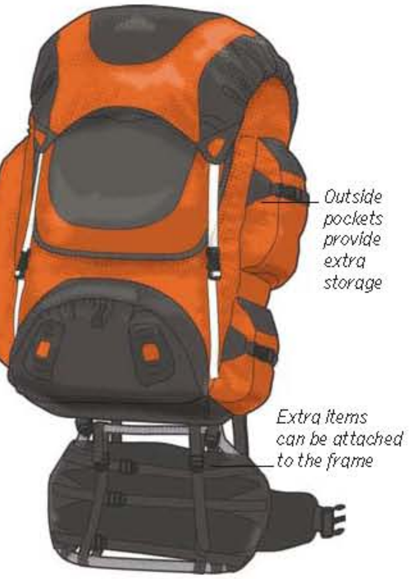

Outside pockets provide extra storage

Extra items can be attached to the frame

DAY PACK

Day hikers can sometimes find themselves in trouble because they are not carrying the appropriate equipment to protect them from the elements. A day pack provides you with the means to carry essential items such as food, water, maps, compasses, and protection from cold and wet weather on a day trip.

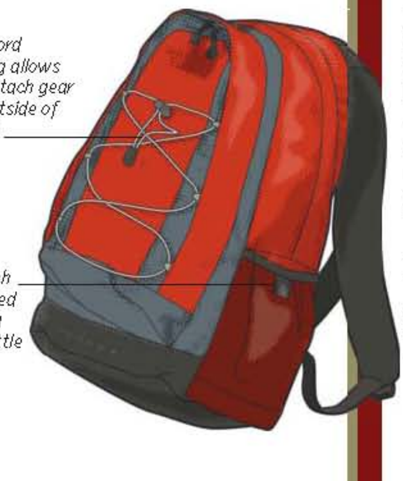

Elastic cord strapping allows you to attach gear to the outside of the pack

Side pouch can be used to carry a water bottle

FITTING A BACKPACK

Once you have chosen a backpack of the right size, design, and features for your trip, you will have to make sure that it's a comfortable fit. You will need someone's help, but here are a few tips for ensuring your new pack sits comfortably on your back.

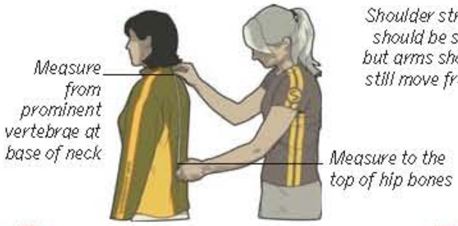
Measure from prominent vertebrae at base of neck
Measure to the top of hip bones

1 First measure your back. The pack you choose should be based on the length of your torso.

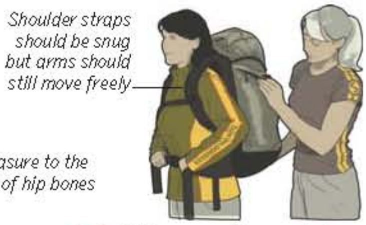
Shoulder straps should be snug but arms should still move freely

2 Make sure the pack is a comfortable fit, allowing for any extra layers of clothing.

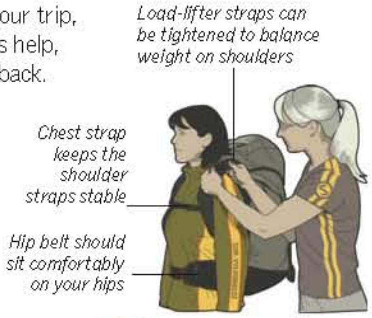
Load-lifter straps can be tightened to balance weight on shoulders
Chest strap keeps the shoulder straps stable
Hip belt should sit comfortably on your hips

3 You can now fine-tune the pack's fit by adjusting the shoulder and hip straps.

PACKING A BACKPACK

Always pack in reverse order: the items you need first should go in last. Heavy items should be placed close to your back to prevent the pack pulling away from your shoulders.

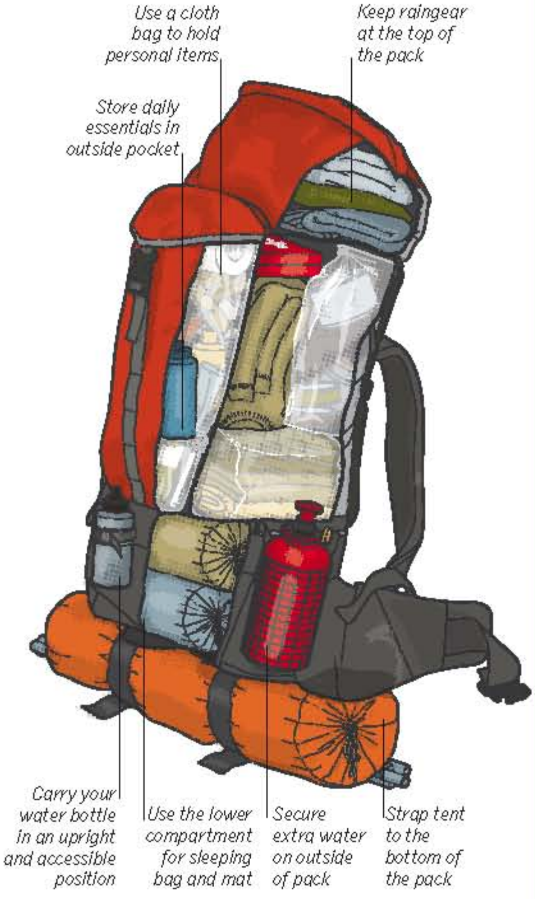
Use a cloth bag to hold personal items
Keep raingear at the top of the pack
Store daily essentials in outside pocket
Carry your water bottle in an upright and accessible position
Use the lower compartment for sleeping bag and mat
Secure extra water on outside of pack
Strap tent to the bottom of the pack

MAKING A SIMPLE PACK FRAME

In a survival situation, having something to carry your gear will be a huge advantage. All you need to make your own improvised backpack is a knife, some wood, some cord, and a waterproof sheet.

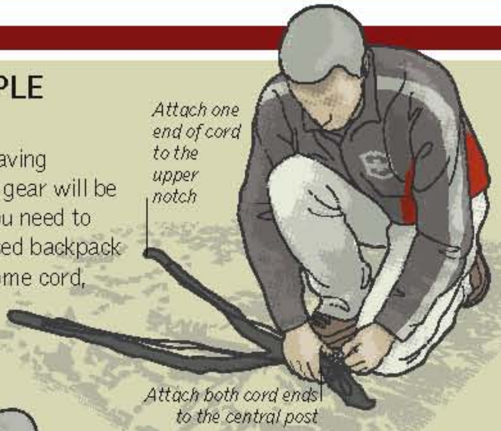
Attach one end of cord to the upper notch
Attach both cord ends to the central post

1 Cut a light bough about 1 ft (30 cm) below the point where it branches, leaving about 3 ft (1 m) above the fork.

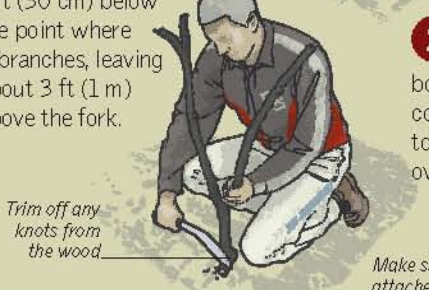
Trim off any knots from the wood

2 Cut notches on all three ends of the bough and tie lengths of cord around the notches to serve as straps to go over your shoulders.

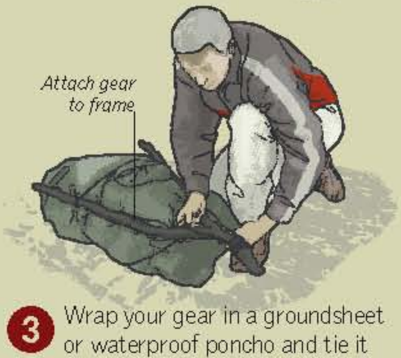
Attach gear to frame

3 Wrap your gear in a groundsheet or waterproof poncho and tie it as high as possible onto the frame on the side opposite the straps.

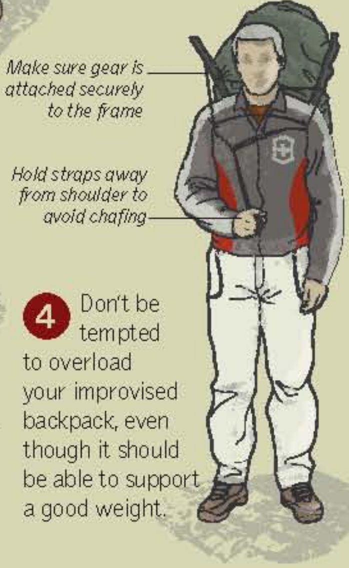
Make sure gear is attached securely to the frame
Hold straps away from shoulder to avoid chafing

4 Don't be tempted to overload your improvised backpack, even though it should be able to support a good weight.

DRESSING FOR THE OUTDOORS

MODERN OUTDOOR CLOTHING is technologically advanced and highly sophisticated. New materials and designs are extremely lightweight, durable, and versatile. To make the most of your gear, choose the fabrics and combinations of clothes most suited to the environment and conditions in which you are traveling.

REGULATING BODY HEAT

Don't be tempted to overdress. Sweat can soak you as much as rain, and a sudden change of temperature can leave you vulnerable to hypothermia. Wearing layers helps you to control your body temperature.

LAYERING CLOTHES

The rule of layering is simple: several light layers are better than one heavy layer. Wearing multiple layers gives you flexibility to fine-tune your temperature by taking off or putting on layers to regulate your body heat. Choose fabrics such as wool, fleece, microfleece, and down—all of which are good insulators.

HOW LAYERING WORKS

Layering is effective because it traps air between the layers of clothes and helps you stay warm in any environment. If you wear the correct layers in the right order, your clothing will move sweat away from your body, keep moisture out, and help to keep you insulated. Wear wicking fabrics, such as polypropylene, in hot and cold weather.

THE LAYERING SYSTEM

The top (outer) layer repels rain, while dampness is wicked away from the skin by the base layer closest to the skin. The mid-layers insulate the body and help retain warmth.

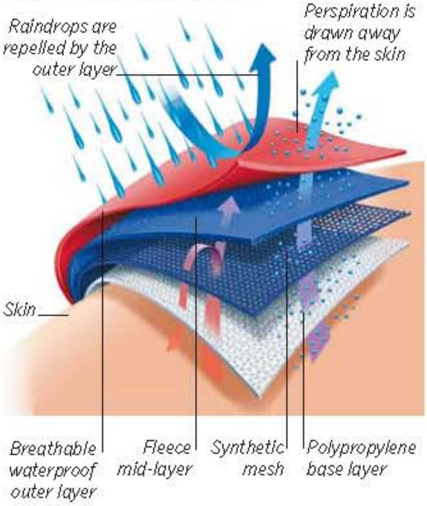
Raindrops are repelled by the outer layer
Perspiration is drawn away from the skin
Skin
Breathable waterproof outer layer
Fleece mid-layer
Synthetic mesh
Polypropylene base layer

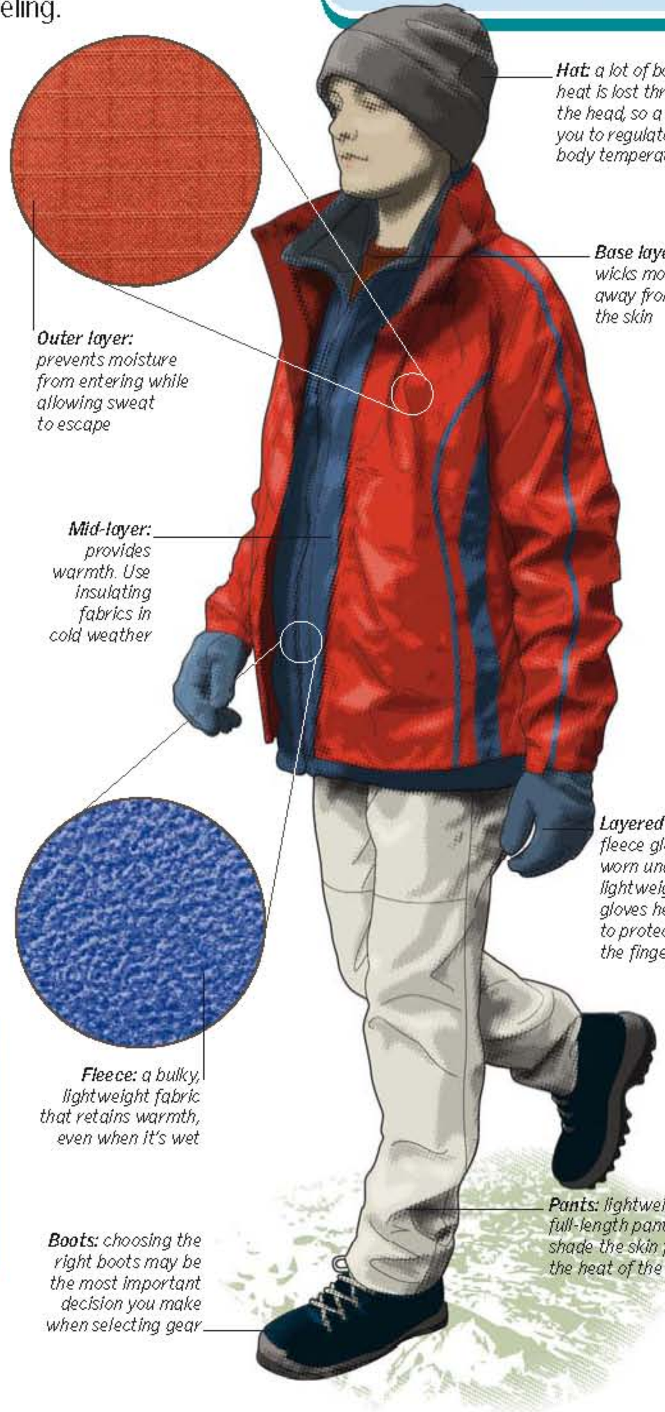

Hat: a lot of body heat is lost through the head, so a hat helps you to regulate your body temperature

Base layer: wicks moisture away from the skin

Outer layer: prevents moisture from entering while allowing sweat to escape

Mid-layer: provides warmth. Use insulating fabrics in cold weather

Layered gloves: fleece gloves worn under lightweight gloves help to protect the fingers

Fleece: a bulky, lightweight fabric that retains warmth, even when it's wet

Pants: lightweight, full-length pants shade the skin from the heat of the sun

Boots: choosing the right boots may be the most important decision you make when selecting gear

WARM-WEATHER CLOTHING

It's vital to stay as cool as possible in warm weather to avoid heat exhaustion or heat stroke (see p. 272). Too much exposure to direct sunlight can cause sunburn and dehydration, and heat rashes can be exacerbated if sweat is not able to evaporate properly. Choose breathable fabrics that will keep you cool and protect your skin from the sun. Stay hydrated and always wear a hat.

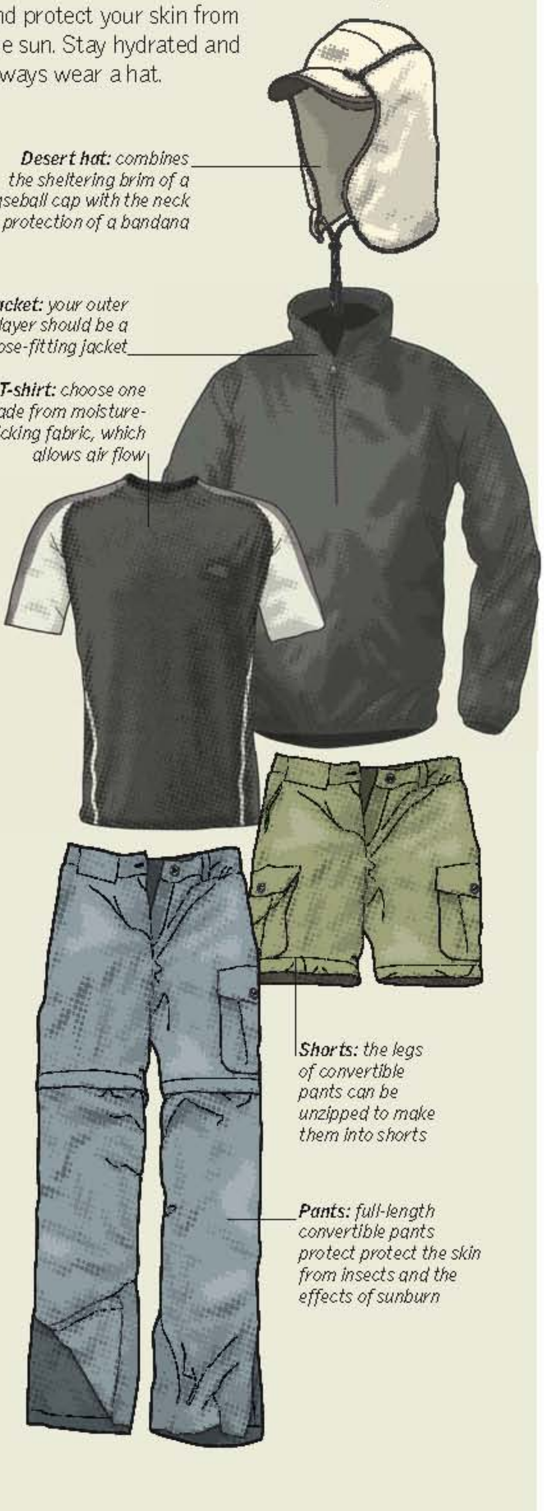

***Desert hat:** combines the sheltering brim of a baseball cap with the neck protection of a bandana*

***Jacket:** your outer layer should be a loose-fitting jacket*

***T-shirt:** choose one made from moisture-wicking fabric, which allows air flow*

***Shorts:** the legs of convertible pants can be unzipped to make them into shorts*

***Pants:** full-length convertible pants protect protect the skin from insects and the effects of sunburn*

COLD-WEATHER CLOTHING

In cold weather, you need to pay particular attention to your clothing layers. Use a lightweight base layer, which will wick moisture from your skin, several layers of warm, insulating clothing, and an outer shell that's windproof and waterproof. The secret to staying comfortable is to adjust your layers as your body temperature changes. Remove layers as your body temperature rises to avoid sweating.

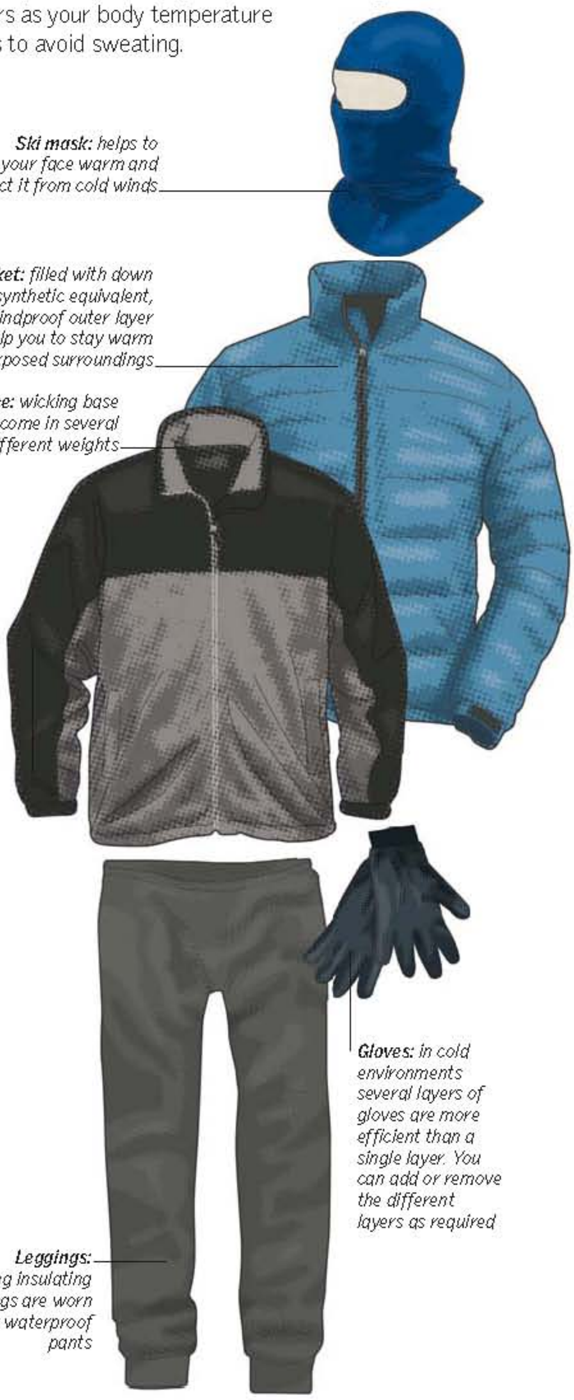

***Ski mask:** helps to keep your face warm and protect it from cold winds*

***Jacket:** filled with down or a synthetic equivalent, a windproof outer layer will help you to stay warm in exposed surroundings*

***Fleece:** wicking base layers come in several different weights*

***Gloves:** in cold environments several layers of gloves are more efficient than a single layer. You can add or remove the different layers as required*

***Leggings:** wicking insulating leggings are worn under waterproof pants*

WET-WEATHER CLOTHING

Sudden downpours can occur almost anywhere and it is easy to be caught out by them. The best fabrics to wear in such conditions are those that are breathable, waterproof, and allow sweat to escape—the best known of these is Gore-Tex®. Always carry the gear necessary to stay warm and dry, and make sure you can access it quickly if you are caught out— it can be the difference between comfort and misery.

HOW BREATHABLE WATERPROOF FABRIC WORKS

Breathable waterproof fabric is a laminate of three layers: two layers of nylon on either side of a thin layer of microporous teflon, which contains tiny pores—50 times smaller than the width of a human hair—that keep rain droplets out but allow perspiration (water vapor) to escape.

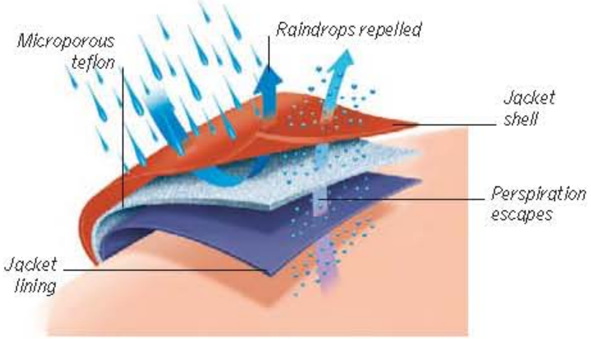

SOCKS

Socks are a vital part of your footwear system. Their function is two-fold. Firstly, they cushion the feet and prevent boots from rubbing and causing blisters. Secondly, they keep the feet warm and dry by wicking moisture away from the feet. Note that socks made from breathable waterproof fabric can be worn with regular boots but should not be combined with breathable waterproof boots.

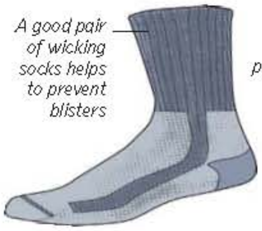

A good pair of wicking socks helps to prevent blisters

WICKING SOCK

A lightweight inner sock that helps wick moisture away from the skin and reduces friction.

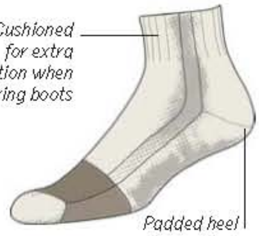

Cushioned for extra protection when wearing boots

Padded heel for comfort

WALKING SOCK

The thick fabric of the walking sock offers maximum cushioning, comfort, and protection.

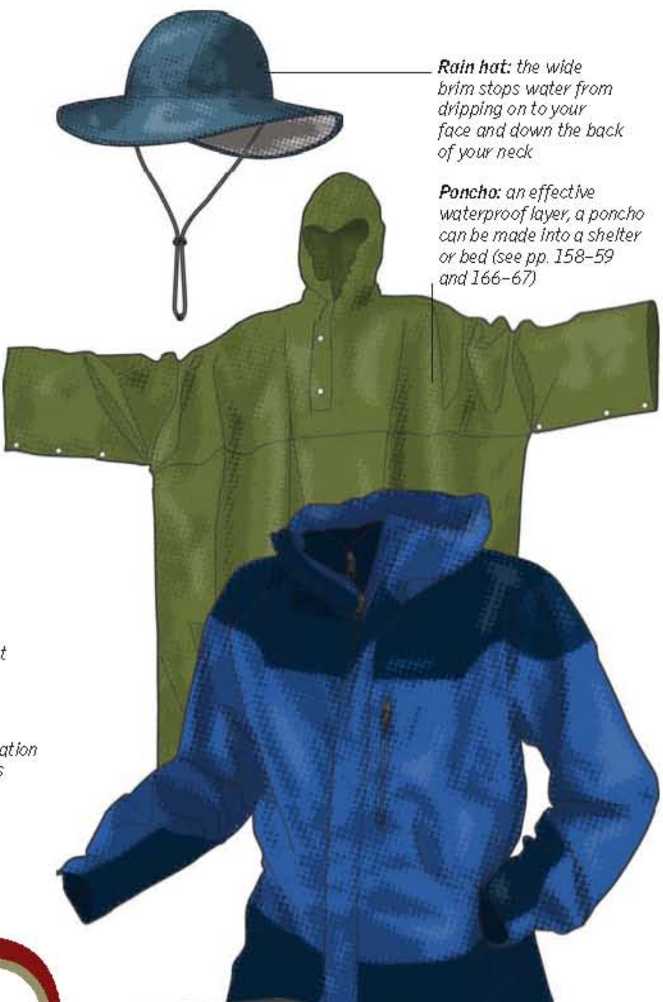

Rain hat: *the wide brim stops water from dripping on to your face and down the back of your neck*

Poncho: *an effective waterproof layer, a poncho can be made into a shelter or bed (see pp. 158–59 and 166–67)*

Waterproof jacket: *allows sweat to escape while keeping rain out*

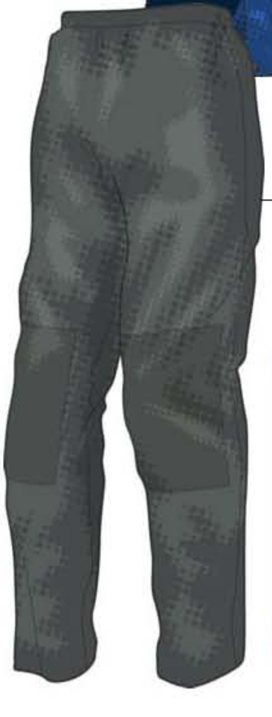

Waterproof pants: *keep your legs dry in wet conditions*

GAITERS

Gaiters are useful, both in temperate areas for keeping the feet and lower pant legs dry, and in polar regions for keeping out snow.

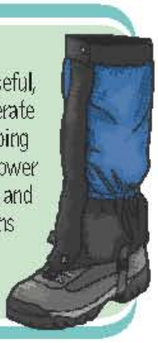

FOOTWEAR

A number of factors determine which type of footwear is right for you. The first concerns your personal requirements, which include the shape of your feet and the amount of support you need. You also need to consider how far you are traveling and the type of terrain you'll be covering. Then there is cost: footwear varies from the relatively cheap to the extremely costly. Bear in mind that if you intend to spend a lot of time in the wilderness, it pays to invest a little extra in comfortable, sturdy footwear.

BREAKING IN BOOTS

Your boots and feet should coexist in perfect harmony when you're on the trail. When you buy a pair of boots, wear them around the house and go on short hikes to make sure you have broken them in.

ANATOMY OF A SOLE

A boot sole has several layers to protect and support your feet. Look for a good tread on the outsole, and for cushioning beneath the heel and toe areas.

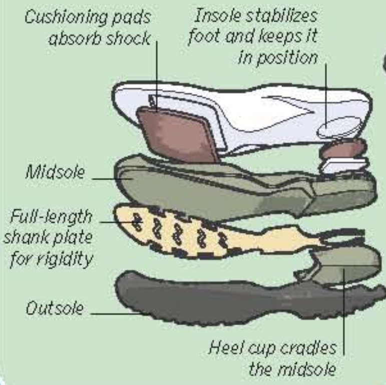
Cushioning pads absorb shock
Insole stabilizes foot and keeps it in position
Midsole
Full-length shank plate for rigidity
Outsole
Heel cup cradles the midsole

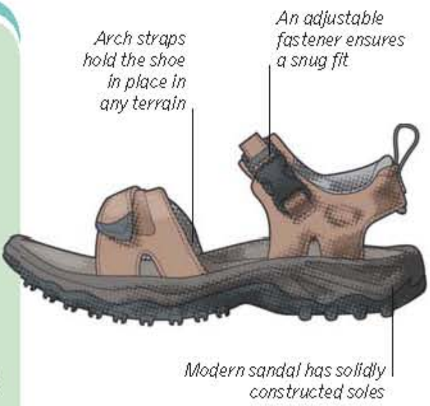
Arch straps hold the shoe in place in any terrain
An adjustable fastener ensures a snug fit
Modern sandal has solidly constructed soles

SANDALS

Modern sandal designs offer great support, and are extremely comfortable, while offering additional ventilation.

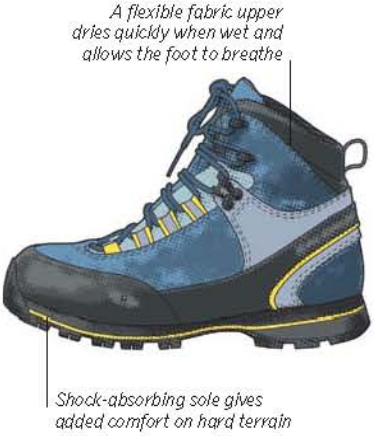
A flexible fabric upper dries quickly when wet and allows the foot to breathe
Shock-absorbing sole gives added comfort on hard terrain

LIGHTWEIGHT BOOT

Fabric and leather hybrids are increasingly popular because they combine the support and traction of a heavier shoe with the flexibility of a sports shoe, so there is less friction on the foot.

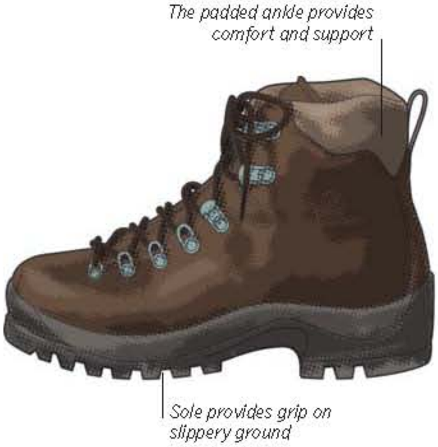
The padded ankle provides comfort and support
Sole provides grip on slippery ground

HIKING BOOTS

A compromise between weight, durability, and protection produces a good, all-round leather boot with a sole of hard-wearing rubber, such as Vibram®, water-resistant uppers, and valuable support for the ankles.

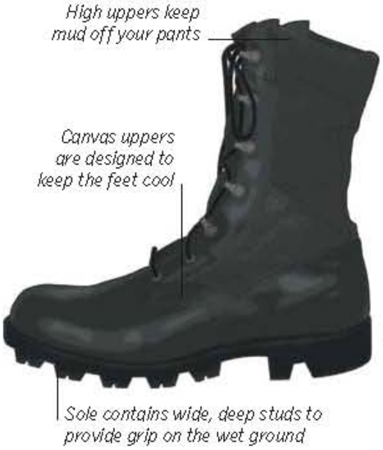
High uppers keep mud off your pants
Canvas uppers are designed to keep the feet cool
Sole contains wide, deep studs to provide grip on the wet ground

JUNGLE-TREKKING BOOT

This is a high boot made from rot-proof leather and canvas with a directly molded sole. Breather holes in the instep aid ventilation and help to drain moisture in hot or humid conditions.

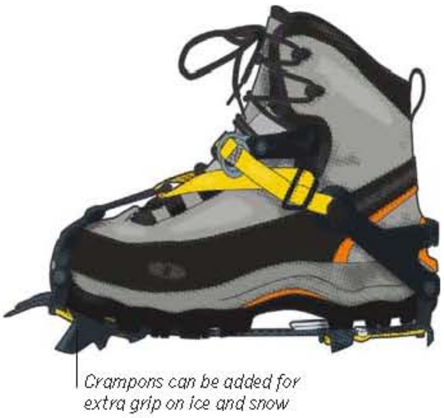
Crampons can be added for extra grip on ice and snow

CLIMBING/HIKING HYBRID BOOT

More flexible than traditional climbing boots, modern hybrid boots are designed to be worn with crampons and to keep the feet warm in very cold environments.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE WILDERNESS

USEFUL EQUIPMENT

- Walking staff
- Pocket chainsaw
- Whistle, flashlight, signal mirror
- Animal deterrent spray
- Water purifiers
- Small AM/FM radio
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

BLAKE STANFIELD, 38, AND HIS 65-YEAR-OLD FATHER, Neil, endured being trapped beneath river ice, encounters with bears, and four days and nights in the Alaskan wilderness without food. They survived thanks to good survival skills, knowledge of the local area, and an abundance of water and firewood.

Father and son were dropped by floatplane on Friday, June 6, 2003, at a remote point on the Koyukuk River, deep inside Alaska. They planned to spend seven days rafting approximately 90 miles (145 km) to the town of Bettles. The trip had hardly begun when disaster struck—the raft crashed at speed into a massive sheet of ice, throwing the pair into the freezing water. They were both caught by the strong current and pulled under the ice. Somehow Blake was washed ashore, and he used a branch to haul his father to the riverbank.

Blake quickly realized that Neil was close to hypothermia, so used his waterproof lighter to light a fire of sticks and pine needles. Far from rescue, with their supplies lost and no one expecting their return for six days, their prospects looked bleak. With night approaching they built a shelter.

"THE RAFT CRASHED AT SPEED INTO A MASSIVE SHEET OF ICE"

The next morning, Blake decided to hike the 65 miles (105 km) to Bettles to get help. On Sunday his progress was brought to a halt by the confluence of two large rivers—the Tinayguk and Koyukuk. However, Blake realized that he had reached a landmark on the flight path of supply planes that frequently flew into the interior, so he built a signal fire and waited. Two days later, a pilot spotted him and radioed his position to a US Army base 200 miles (320 km) away in Fairbanks, then dropped him a two-way radio and supplies. The pilot then continued up the river to find Neil, dropping him a sleeping bag, tent, and supplies, before radioing his position to the Army. Both men were picked up the next day—exhausted, malnourished, and close to hypothermia, but alive.

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Elements—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to—find an elevated position from which to choose a suitable area for survival and rescue
- Regulate your clothing to avoid overheating when moving and hypothermia when static
- Use or improvise a walking staff to help reduce trips and falls
- Improvise shelter when not moving
- Plan your route around potential or known water sources. Filter and purify all water where possible
- Have aids to location accessible while moving and deployed when static

DON'T

- Ignore your fire—be on constant lookout for dry tinder/kindling and fuel
- Walk faster than the pace of the slowest person in the group
- Be careless when walking downhill—a twisted ankle could prove fatal
- Under or overdress. Start off a walk lightly dressed and add or remove layers as necessary

YOU SHOULD STAY **

DON'T

- Leave food in your campsite as it will be at risk from predators
- Shelter too deep in the woods despite the protection it gives you from elements. Remember: your location aids need to be seen
- Eat unidentified food, as it could exacerbate your situation through illness. Food is not a priority in a short-term situation

DO

- Check your shelter site before building for natural hazards such as: insects, flooding, rock falls, wild animals, and deadfalls
- Inventory your supplies and ration them
- Keep a fire going: once it's established you can use it to purify water, benefit from the warmth, and signal rescue
- Fill plastic bags or spare clothing with dry foliage and use as a mattress or pillow to insulate you from the cold/damp ground
- If in a group, give everyone something to do, even the children. It will keep them occupied and lessen the worry

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

SLEEPING SYSTEMS

REST IS AS ESSENTIAL for survival as food. A good night's sleep can offset much of the worry and stress of a difficult situation. Making sure your shelter is suitable for the environment you're in, and choosing the right sleeping bag, can be vital factors in survival.

TIPS ON BUYING A SLEEPING BAG

There are many types of bag available. Here are a few things you should consider to ensure you buy the right sleeping bag for you.

- Determine the lowest temperature you're likely to encounter on your trip and choose a bag that will perform at this temperature.
- Synthetic bags are cheaper and easier to clean than down bags, and continue to provide insulation even when wet.
- Down bags are more expensive than synthetic bags, but provide a better warmth-to-weight ratio and last longer.
- A mummy-shaped bag provides better insulation than a rectangular-shaped bag, but has less room to move around in.

SLEEPING BAGS

Although there are many different shapes and styles of sleeping bag, the one you choose should have enough padding to keep you warm at night even without a tent. A sleeping bag must never be allowed to get wet, so always keep it inside a waterproof cover, such as the bag's stuff sack or a bivy sack. Always try a sleeping bag before you buy it. If it's too snug, it will be less effective.

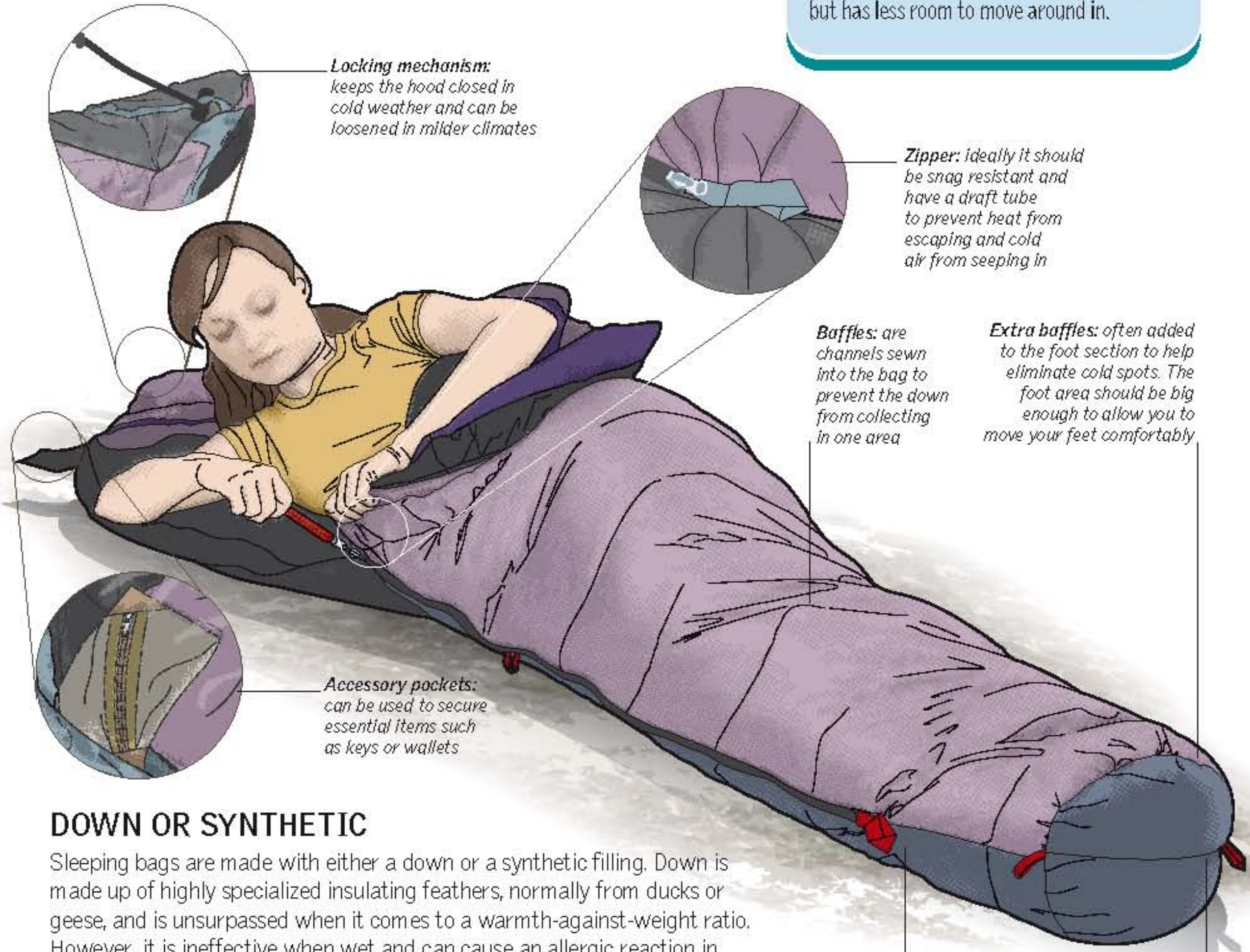

Locking mechanism: keeps the hood closed in cold weather and can be loosened in milder climates

Zipper: ideally it should be snag resistant and have a draft tube to prevent heat from escaping and cold air from seeping in

Baffles: are channels sewn into the bag to prevent the down from collecting in one area

Extra baffles: often added to the foot section to help eliminate cold spots. The foot area should be big enough to allow you to move your feet comfortably

Accessory pockets: can be used to secure essential items such as keys or wallets

Loops: allow you to attach sleeping bag to the sleeping mat

Box foot: creates a mummy shape and is designed to conform to the position of your feet when you're sleeping on your side

DOWN OR SYNTHETIC

Sleeping bags are made with either a down or a synthetic filling. Down is made up of highly specialized insulating feathers, normally from ducks or geese, and is unsurpassed when it comes to a warmth-against-weight ratio. However, it is ineffective when wet and can cause an allergic reaction in some people. Synthetic fillings range from simple hollow fibers all the way up to complex fibers designed to mimic the structure of down. They still retain some of their insulating capacity if they get wet.

TEMPERATURE RATINGS

Because manufacturers have no standardized way of registering how warm a sleeping bag is, temperature ratings for sleeping bags remain an inexact science. Some manufacturers give a bag both a comfort rating and a survival rating—known as a "season rating.' These give some indication as to how manufacturers expect a sleeping bag to perform.

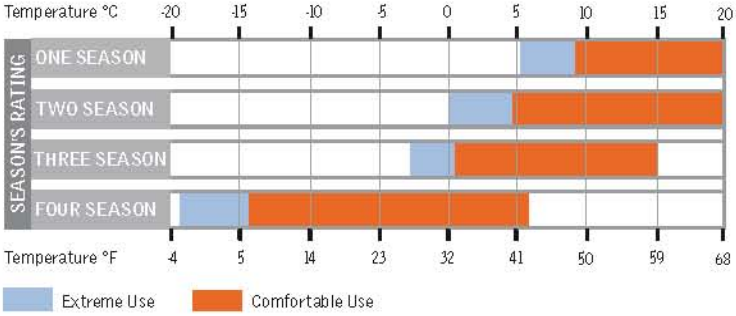

CONSTRUCTION

The principle of all sleeping bags is the same: to trap air and prevent its circulation so that the body heats the trapped air and keeps you warm while you sleep. This can be achieved in a number of ways.

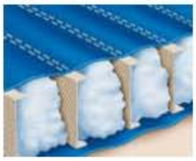

BOXWALL

Used in cold-weather down bags, the filling is contained within box-like sections to minimize bunching.

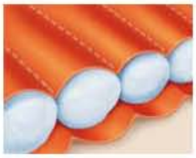

SEWN-THROUGH

Filling is in separate oval channels, although heat is still lost through the stitching.

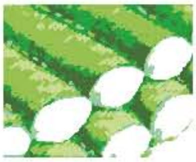

OFFSET CHANNELS

The filling is laid out in two layers to prevent bunching and heat loss through the stitching.

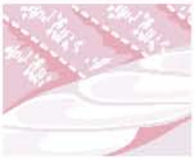

SHINGLE

Slanted layers of overlapping fibers that fill with warm air to aid insulation.

USING YOUR SLEEPING BAG

Before using your sleeping bag, always check inside for spiders and other dangers, and shake it to make sure the filling is distributed evenly. During use, if the bag is too warm, simply unzip it to cool down, or use the bag as a quilt. If it's too cold, either put on some clothing, or use a silk liner or waterproof bivi bag to increase its performance. Always air your sleeping bag well after using it.

SLEEPING ACCESSORIES

You'll need to get a sleeping mat as well as a sleeping bag to ensure a comfortable, dry night's sleep. Additional gear, such as bivy sacks, sleeping bag liners, and inflatable pillows, will also increase comfort.

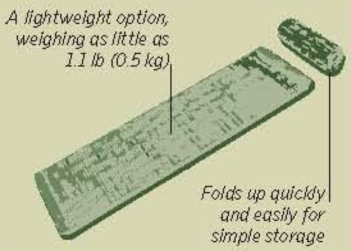

A lightweight option, weighing as little as 1.1 lb (0.5 kg)

Folds up quickly and easily for simple storage

FOAM MATS

The lightest and simplest type of mat, foam mats are best used on soft surfaces, such as sand or pine needles.

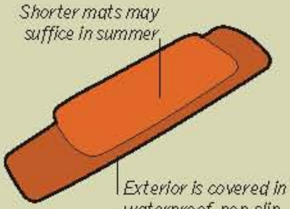

Shorter mats may suffice in summer

Exterior is covered in waterproof, non-slip polyester material

AIR MATS

They may be heavier and more expensive than foam mats, but air mats provide superior comfort and thermal insulation.

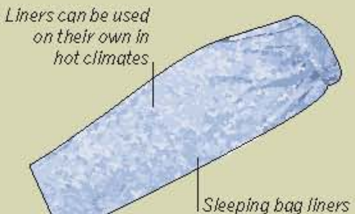

Liners can be used on their own in hot climates

Sleeping bag liners are easier to wash than sleeping bags

SLEEPING BAG LINER

Usually made of cotton or silk, sleeping bag liners trap air between your body and the sleeping bag.

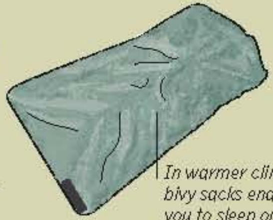

In warmer climates, bivy sacks enable you to sleep outside without a tent

BIVY SACKS

A simple waterproof bag that is used over a sleeping bag to keep it dry or used on its own as a shelter.

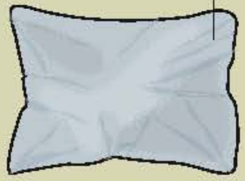

Outer layer of pillow is padded for extra comfort

INFLATABLE PILLOW

An inflatable pillow adds some luxury to expeditions. Alternatively, fill a stuff sack with spare clothing.

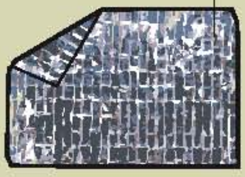

Foil prevents body heat from escaping and deflects it back to the body

SPACE BLANKET

A lightweight foil blanket for use in emergencies, space blankets can also be used as a temporary canopy.

PORTABLE SHELTERS

Shelter from the elements at night is essential for long-term survival. A variety of products are available to you, depending on the conditions you're about to enter and the amount you're prepared to carry. You can choose from lightweight options, such as hammocks and variations on bivy sacks, and the more traditional alternatives, tents, which are heavier but can accommodate up to eight people.

USING A HAMMOCK

If you have a hammock you can camp even if the ground is muddy, rocky, or on an incline. The advantage of a hammock is that it's lightweight and can be erected just about anywhere. Modern hammocks provide a shelter and bed solution in one portable package.

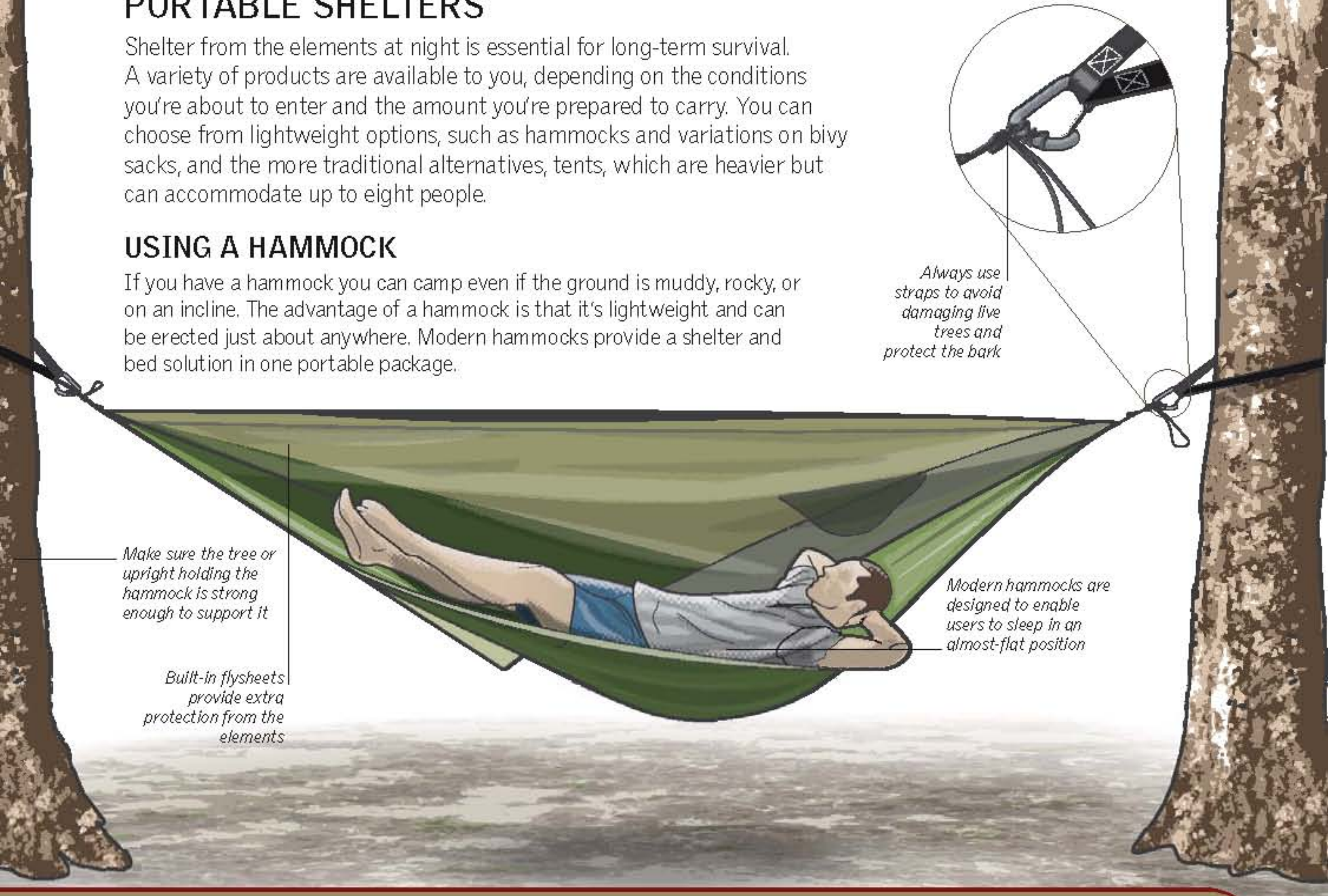

Always use straps to avoid damaging live trees and protect the bark

Make sure the tree or upright holding the hammock is strong enough to support it

Modern hammocks are designed to enable users to sleep in an almost-flat position

Built-in flysheets provide extra protection from the elements

USING A TARP TENT

A cross between a tarpaulin and a tent, a tarp tent is a highly versatile, extremely light form of shelter that's capable of withstanding even the most extreme conditions. Groundsheets should be laid out on the ground for added protection, and a suitable sleeping bag is essential. A one-person version can weigh as little as 18½ oz (0.52 kg).

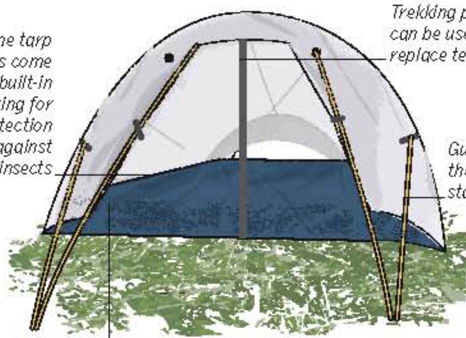

Some tarp tents come with built-in netting for protection against insects

Trekking poles can be used to replace tent poles

Guylines give the tarp tent stability

A groundsheet provides insulation from the cold, damp ground

USING A HOOPED BIVY

The hooped bivy is a waterproof and breathable bivy sack that has been designed to form a one-person shelter. It usually incorporates pegging points and a zippered and hooped entrance that form a small vestibule for your backpack. Many incorporate a heavy-duty groundsheet and can be set up and taken down in a matter of minutes.

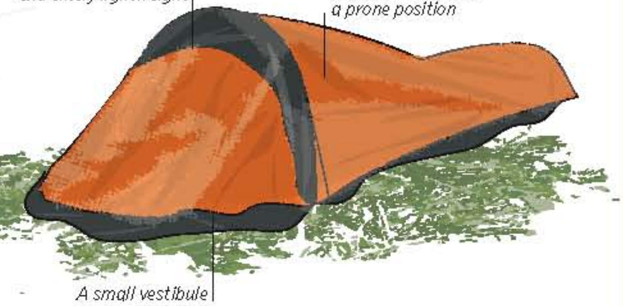

While they may have restricted space, hooped bivys pack small and are extremely lightweight

Limited height means you're restricted to a prone position

A small vestibule provides minimal storage for your gear

USING A TENT

Tents keep the rain and wind out while retaining warmth. The outer flysheet should be strong, impermeable, taut, and able to withstand high winds. The inner tent—which may be attached to a flysheet—is loose and made of permeable material to minimize condensation but retain heat. The inner and outer parts of the tent must not touch each other; if they do, heat may be lost and condensation may form on the underside of the flysheet, then seep into the tent.

WARNING!

In some climactic conditions, such as extreme high winds, severe storms, or heavy snow cover, stakes and guylines may not be enough to anchor the tent securely to the ground. This could result in the tent being blown away—and could have severe consequences for the tent's occupants. In such conditions, use extra means—such as rocks or logs—to anchor your tent securely.

THREE-SEASON TENT

Three-season tents are designed for use in a variety of climates. The inner layer is made of a lightweight material that offers both ventilation and protection against insects. The flysheet—which should extend about 4 in (10 cm) above the ground—and the groundsheet are made of waterproof fabric.

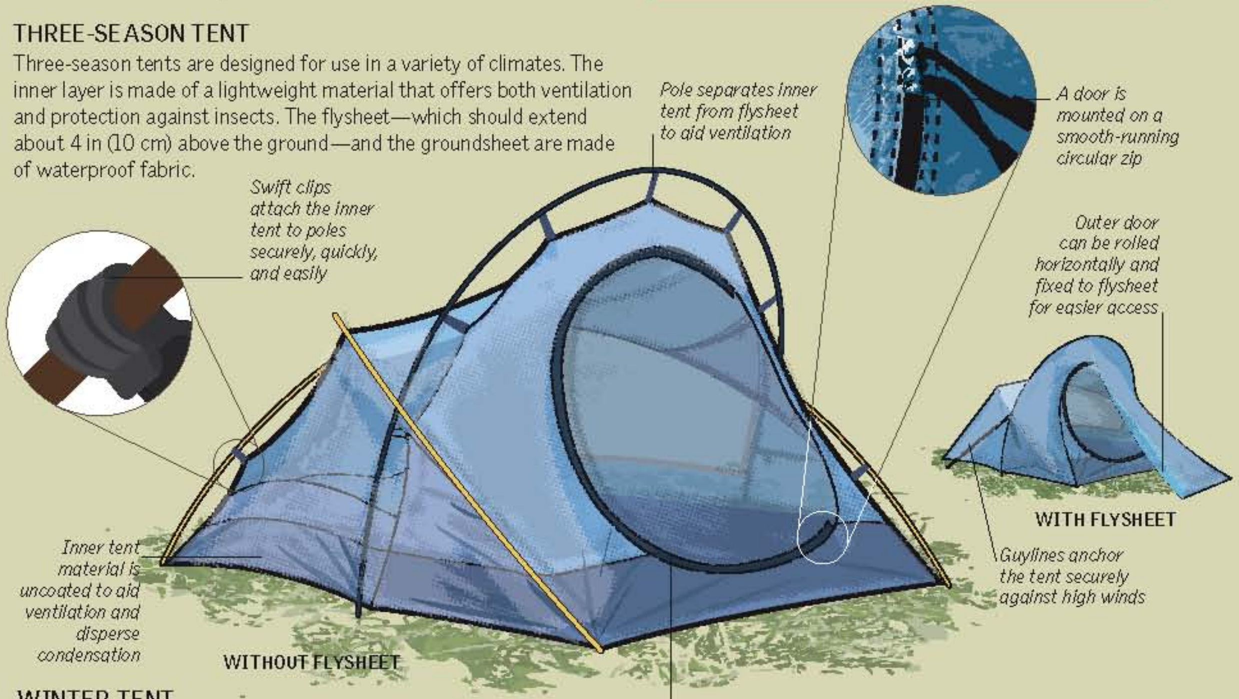
Pole separates inner tent from flysheet to aid ventilation
A door is mounted on a smooth-running circular zip
Swift clips attach the inner tent to poles securely, quickly, and easily
Outer door can be rolled horizontally and fixed to flysheet for easier access
WITH FLYSHEET
Guylines anchor the tent securely against high winds
Inner tent material is uncoated to aid ventilation and disperse condensation
WITHOUT FLYSHEET
Mesh door allows light to enter but keeps insects out

WINTER TENT

For cold-weather climates, look for a tent that offers extra-strong poles, storm windows, and ample pockets so that you can store your gear inside the tent. However, these extra features increase the tent's weight. The dome shape sheds snow well, and withstands high winds.

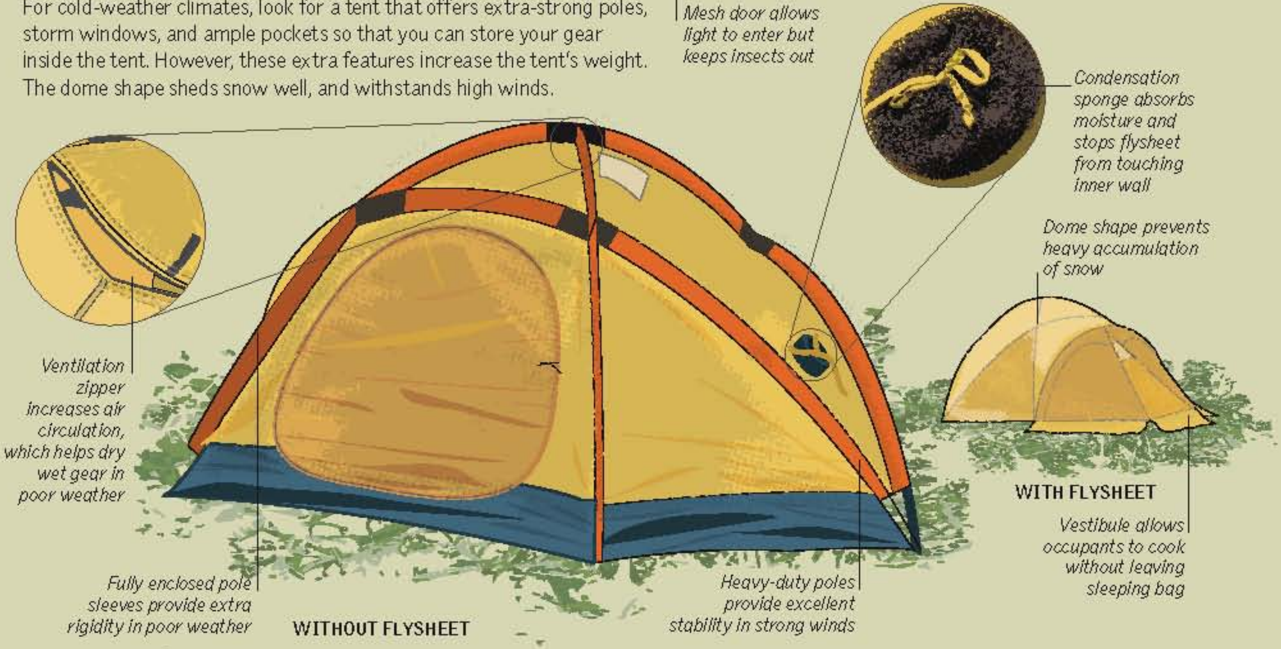
Condensation sponge absorbs moisture and stops flysheet from touching inner wall
Dome shape prevents heavy accumulation of snow
Ventilation zipper increases air circulation, which helps dry wet gear in poor weather
WITH FLYSHEET
Vestibule allows occupants to cook without leaving sleeping bag
Fully enclosed pole sleeves provide extra rigidity in poor weather
WITHOUT FLYSHEET
Heavy-duty poles provide excellent stability in strong winds

EATING ON THE TRAIL

THE FOOD YOU EAT determines how efficiently your body will function. This is never more important than when you're undertaking an expedition that will test your physical and mental endurance beyond normal levels.

FOOD FOR LIFE

Your body requires energy to fuel the chemical and biological reactions that take place within it, and to power your muscles. This, and other nutrients, comes from the food you eat (see below).

BODY FAT AND SURVIVAL

Being physically fit and active keeps you strong and less prone to injury, but being slightly overweight can actually be an advantage in a survival situation: 1 lb ($\frac{1}{2}$ kilo) of body fat converts to roughly 3,500 calories—the amount of energy you would usually expend during an average day.

TYPES OF NUTRITION	EXAMPLES	BENEFITS
CARBOHYDRATE	▪ Rice, pasta, bread, and cereals ▪ Fresh and dried fruit ▪ Root vegetables ▪ Chocolate and candy	▪ Major source of energy, as it converts most readily to glucose, which is your body's preferred source of fuel.
PROTEIN	▪ Meat and poultry ▪ Fish, such as tuna and salmon ▪ Eggs and dairy products ▪ Beans, legumes, and seeds	▪ Essential for growth and repair of muscle tissue. ▪ Increases your stamina and energy levels, and helps you fight fatigue. ▪ Makes you feel satisfied after eating.
FAT	▪ Nuts and seeds ▪ Dairy products, such as butter and cheese ▪ Meats, such as bacon and sausage ▪ Cooking oil	▪ Good source of energy. ▪ Makes you feel satisfied after eating.
FIBER	▪ Fresh and dried fruit, especially the skins ▪ Vegetables, beans, and seeds ▪ Whole-grain cereals ▪ Whole-meal bread, whole-meal pasta, and brown rice	▪ Provides bulk to help your digestion. ▪ Helps to make you feel full.
VITAMINS AND MINERALS	▪ Fresh and dried fruit ▪ Green vegetables ▪ Meat and fish ▪ Nuts and seeds	▪ Essential for growth and the repair of your body's tissues. ▪ Help you to maintain healthy teeth and eyes. ▪ Aids bloodflow and the production of energy.

BALANCING YOUR ENERGY LEVELS

You replace the energy that you expend on a daily basis through the food you eat. The energy obtained from food is measured in calories, and you require a certain number of calories each day to remain alive. If you consume too few calories, you will lose weight. However, eating too many calories will lead to weight gain, which can cause long-term health issues. The number of calories you require per day depends on many things, such as your age and level of activity (see right). When you don't replace the energy expended and continue to work, you start to use the energy store contained in your body's fat.

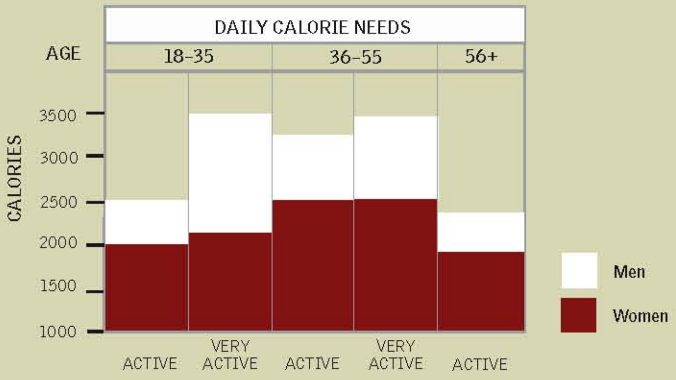

WILDERNESS DINING

Even if you're only going on a short hike, you should take food with you, such as fruit and chocolate, to replace the energy you're using. For longer expeditions, you'll need to take a number of different foodstuffs—for energy and for variety.

WARNING!
In a survival situation, don't eat for the first 24 hours unless to replace expended energy. Ration your food, as digestion uses up water.

TYPES OF FOOD	EXAMPLES	BENEFITS
BREAKFAST CEREALS AND DRIED FRUIT	▪ Trail mix—also known as "gorp" ▪ Muesli or granola ▪ Oatmeal ▪ Dried fruit	▪ Provide a vital source of carbohydrates, vitamins, and minerals to set you up at the start of the day. ▪ Contain dietary fiber to aid digestion.
TRAIL SNACKS	▪ Cookies and crackers ▪ Energy bars and cereal/granola bars ▪ Fruits, such as apples, pears, and oranges ▪ Nuts and seeds	▪ High in carbohydrates to help you maintain energy levels. ▪ Help to keep your hunger pangs at bay between main meals.
HIGH-CALORIE SNACKS	▪ Dehydrated ice cream ▪ Chocolate ▪ Hard candies ▪ Pudding mixes	▪ Provide a quick boost of energy, when required. ▪ Offer a comfort factor, although they are of limited nutritional value.
PRE-PREPARED MAIN MEALS	▪ Boil-in-the-bag meals—you can buy a wide variety of fully prepared meals, sealed in a foil bag. Cook for five to seven minutes. ▪ Freeze-dried and dehydrated meals—simply add boiling water, stir, and wait about five minutes. Extremely lightweight and take up little space. ▪ Canned meals—a great variety of them are available, but they are bulky and heavy.	▪ Provide a balanced range of food groups, vitamins, and minerals to help your body to recover and repair itself after the day's exertions. ▪ Quick and easy to prepare, and require few utensils. ▪ Many require minimal fuel to cook them, or water to rehydrate them. ▪ Can be bulked out with beans, legumes, and grains. ▪ Smaller cans may have a pull-top to open them.
MEAT, FISH, AND CHEESE	▪ Canned fish, such as sardines or tuna ▪ Cured meats, such as ham, salami, and beef jerky ▪ Hard cheese, such as cheddar and parmesan	▪ A great source of proteins and fats. ▪ Easy to store and long-lasting, although cans are heavier than other sources of food.
BEANS, LEGUMES, AND GRAINS	▪ Lentils and dried peas ▪ Dried beans and pearl barley ▪ Rice and pasta ▪ Couscous	▪ Can be used to bulk up main meals and help you to feel full. ▪ An important source of carbohydrates. ▪ Also provide fiber and protein.
MIXERS	▪ Flour ▪ Salt ▪ Suet ▪ Sugar	▪ Flour and salt can be used to make dough. ▪ Sugar and salt help to make wild food palatable. ▪ Suet is an invaluable source of fat that can be used to supplement meals.
FLAVORINGS	▪ Powdered soup ▪ Tomato puree ▪ Powdered sauces ▪ Gravy cubes	▪ Add flavor and variety to your meals. ▪ Especially useful if you're cooking with wild food.
BEVERAGES	▪ Tea ▪ Coffee ▪ Chocolate powder ▪ Dried milk	▪ Provide warmth and comfort, although they are of limited nutritional value. ▪ Chocolate powder is high in carbohydrates. ▪ Dried milk is an excellent source of calcium.

CAMPING STOVES

ALTHOUGH NOTHING CAN BEAT the satisfaction of building your own campfire, the convenience and reliability of purpose-made camping stoves make them an essential piece of cooking equipment, especially in areas where open fires are banned.

THE BENEFITS OF CAMPING STOVES

Modern camping stoves are very light and collapse down to an incredibly small, compact size. Gas-fueled stoves are by far the most common, but liquid-fueled stoves are also worth considering. Both have their own advantages and disadvantages (see below), but your choice will be dictated by space and weight constraints, the distance you're traveling, and environmental factors. A hexamine stove is a good item to have in a survival situation (see opposite).

CHOOSING A GAS- OR LIQUID-FUELED STOVE

Gas-fueled stoves are safer, cleaner, lighter, and easier to use than liquid-fueled stoves, but are less economical, and not suitable for low-temperature, high-altitude use. Liquid stoves offer greater stability, fuel versatility and economy, but you need to pump and prime (preheat) them before use, and they are less clean.

TRANSPORTING FUEL

There are many factors to consider when transporting fuel to make your expedition safer:

- When hiking, pack your fuel container upright in the middle of your backpack, and surround with clothing to protect it from knocks. Make sure you can access it easily when you reach camp. Use similar principles when traveling by other methods.
- Ensure that your fuel bottles look very different from your water bottles to prevent confusion.
- If you're going on a long expedition, check the availability of the fuel, and consider how you will dispose of cylinders safely.

	GAS-FUELED STOVES		
TYPE OF STOVE	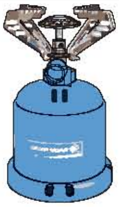 **BUTANE OR ISOBUTENE**	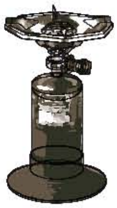 **PROPANE**	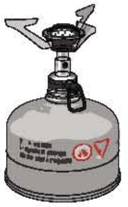 **BLENDED FUEL**
INFORMATION	▪ Both fuels sold in disposable, pressurized canisters.	▪ Fuel sold in disposable, pressurized canisters. ▪ Used in most portable stoves and barbecue grills.	▪ Fuel sold in disposable, pressurized canisters. ▪ Is a mix of butane, propane, and/or isobutene.
PROS	▪ Lightweight, sealed fuel container. ▪ Butane very efficient and burns at a high temperature. ▪ Isobutene more efficient than butane, and performs better in cold conditions.	▪ Lightweight, sealed fuel container. ▪ Fuel burns at a very hot, steady heat, and produces an almost sootless flame. ▪ Performs very well in cold conditions.	▪ Lightweight, sealed fuel container. ▪ Fuel safer than pure propane, and performs better than either pure butane or isobutene in cold conditions. ▪ Isobutene blends give a more efficient flame.
CONS	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes. ▪ Butane less efficient than isobutene in temperatures below 50°F (10°C).	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes. ▪ Highly combustible, so less safe than other fuels.	▪ Higher cost and lower heat output than liquid fuels. ▪ Fuel efficiency reduced at higher altitudes.

USING CAMPING STOVES SAFELY

Always take care when storing the fuel and using your stove. Use only the stove manufacturer's recommended fuel, and check all connections and fittings before each and every use. Also:

- Clear the cooking area of dry vegetation and leaves, in case you spill fuel or knock over your stove.
- Place your stove on a level and stable surface.
- Keep naked flames and sparks away from the fuel, especially when filling or refilling the stove. Before refilling, make sure the flame is out and the stove has cooled down.
- Use the cooker in a well-ventilated space. Liquid fuels, in particular, can give off poisonous fumes, such as carbon monoxide, and all cookers also burn vital oxygen.
- Don't use a stove inside your tent. In addition to the risk of poisonous fumes, fire is also a significant danger. A confined space with limited access, your tent can be reduced to ashes in minutes, along with everything inside, and you may not be able to escape in time to protect yourself. Losing your tent and your equipment is bad enough, without you receiving third-degree burns at the same time.
- Don't leave a burning stove unattended. If it topples over, it could easily start a fire. If the stove flame goes out, a spark or naked flame could also ignite the vapors and cause an explosion.

THE HEXAMINE STOVE

A very useful piece of emergency equipment, the hexamine stove (or hexi-stove) burns solid blocks of hexamethyl-enetetramine (also known as "methenamine"), and folds away compactly. The stove acts as a small windbreak and cooking platform, while the fuel blocks it uses provide a very reliable means of starting a fire in any conditions. The blocks produce a smokeless flame, don't liquefy while burning, and leave no ashes.

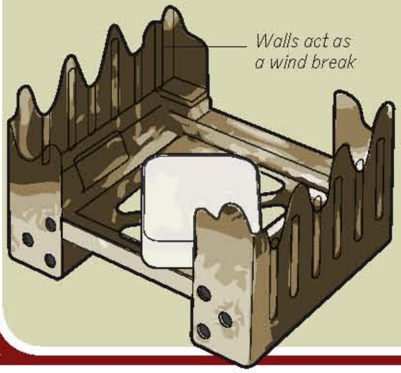
Walls act as a wind break

	LIQUID-FUELED STOVES			
	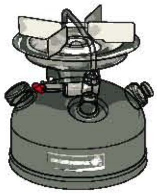 **WHITE (COLEMAN) FUEL**	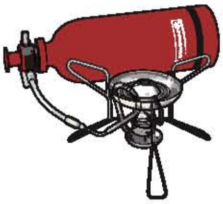 **MULTI-FUEL**	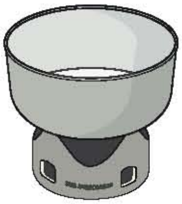 **DENATURED ALCOHOL**	**TYPE OF STOVE**
	■ Fuel sold in disposable canisters.	■ Stove can use a variety of fuels, from unleaded gas, kerosene (paraffin), aviation fuel, and white fuel, to diesel—most of which are readily available.	■ Fuel sold in disposable plastic containers. ■ Stove basically consists of a vessel to burn the fuel in, and a windshield that also acts as a cooking platform.	**INFORMATION**
	■ Fuel can withstand low temperatures, high altitudes, and most weather conditions. ■ Produces a clean flame.	■ Stove uses a range of fuels. ■ Fuel is non-pressurized, so can be stowed safely in almost any airtight container. ■ Stove tends to perform reliably, efficiently, and quickly, producing a hot flame.	■ Fuel is non-pressurized, so can be stowed safely in almost any airtight container. ■ Economical and produces a clean flame. ■ Stove is lightweight, stable, and safe. ■ Usually supplied with pots and pans.	**PROS**
	■ Stove must be pumped to generate required pressure for use. ■ Fuel needs to be primed.	■ Potential for spillage of flammable fuel. ■ Stove must be pumped to generate required pressure for use. ■ Fuel needs to be primed. ■ Some fuels produce soot and noxious fumes.	■ Potential for spillage of flammable fuel. ■ Stove slower to heat up than other liquid stoves, because it's not pumped. ■ Flame vulnerable to wind and other environmental conditions.	**CONS**

YOUR SURVIVAL KIT

A BASIC SURVIVAL KIT is an essential item to take with you on any outdoor expedition. It should be compact enough to carry on your person at all times, and its contents should address the key principles of survival: protection, location, water, and food.

EXAMPLE OF A SURVIVAL KIT

Choose a tin with a waterproof seal and locking clasps. While you can buy ready-made kits, you should always adapt the contents, both to your needs and to the environment you'll be in. Ideally, items in the survival kit should be high-quality and multi-purpose—for example, the tin itself could be used as a cup, a small cooking pot, or even a signal mirror. Your kit should contain the following:

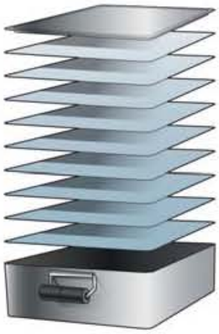

LAYERS OF A SURVIVAL TIN

LAYER 1

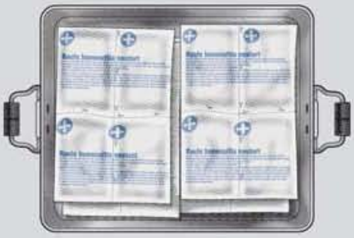

- **Blister medical pads and bandages:** useful for minor cuts and blisters. Include a range of bandages (both waterproof and fabric) in a selection of sizes. Waterproof bandages can also be used to mend holes in tents and tarpaulins.

LAYER 2

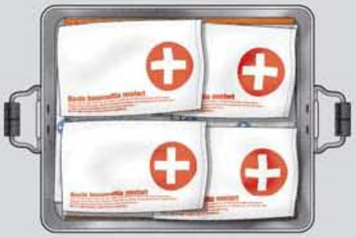

- **Antiseptic wipes:** for treating wounds and cleaning bites. Can also be iginited by a spark and used to get a fire going.

LAYER 3

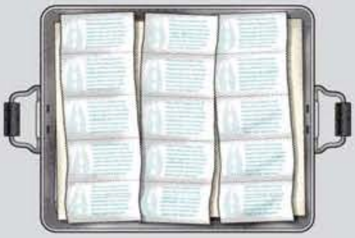

- **Water purification tablets:** choose from tablets or iodine (but don't use iodine if you're allergic to it, or to shellfish). Water will need to be filtered before being purified.

LAYER 6

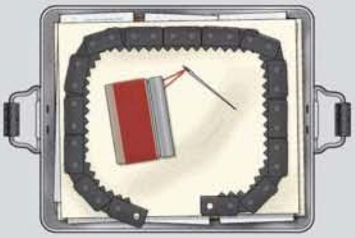

- **Survival saw, or pocket chainsaw:** can be wrapped around the inside of the tin or cut in half, if space is limited.
- **Single-edged razor:** multi-purpose—from skinning an animal to cutting cord. Store in its packaging.
- **Needle and cotton:** use strong, waxed cotton, pre-threaded through the needle.

LAYER 7

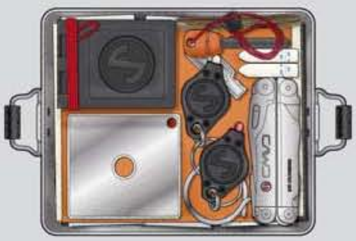

- **Flashlights:** two small Photon (LED) lights—one white and one red—taped in "off" position.
- **Mini multi-tool:** see panel, opposite.
- **Chainsaw handles:** use with pocket chainsaw.
- **High-viz card, signal mirror:** location aids.
- **Compass:** an emergency back-up.
- **Flint, fire steel, tampons:** for starting fires.
- **Snare wire:** for animal traps and lashing.

LAYER 8

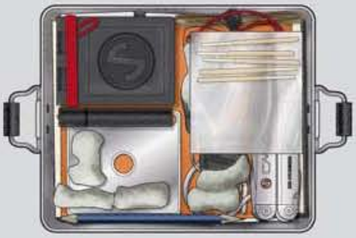

- **Waterproof matches and tinder balls:** for starting fires. Store the matches in a small, resealable plastic bag.
- **Pencil:** sharpened at both ends.
- **Potassium permanganate:** dissolve in water to sterilize water at low concentration, and to clean wounds at a high concentration. Can also be used with sugar to make fire. Store in a waterproof container.

CHOOSING A MULTI-TOOL

A useful item to carry with you on your trip, multi-tools are usually designed with a specific task or hobby in mind, and include useful features for survival, such as a small compass, pliers, saw, and various blades. Use one that has quality tools that most suit your needs, with a comfortable grip, and ensure that the blades are lockable to prevent injuries. Carry your multi-tool in a secure pocket, or in a pouch on your belt, and include a miniature version in your survival tin (see below). Bear in mind that your multi-tool should be an addition to, not a replacement for, your bushcraft knife (see p. 146).

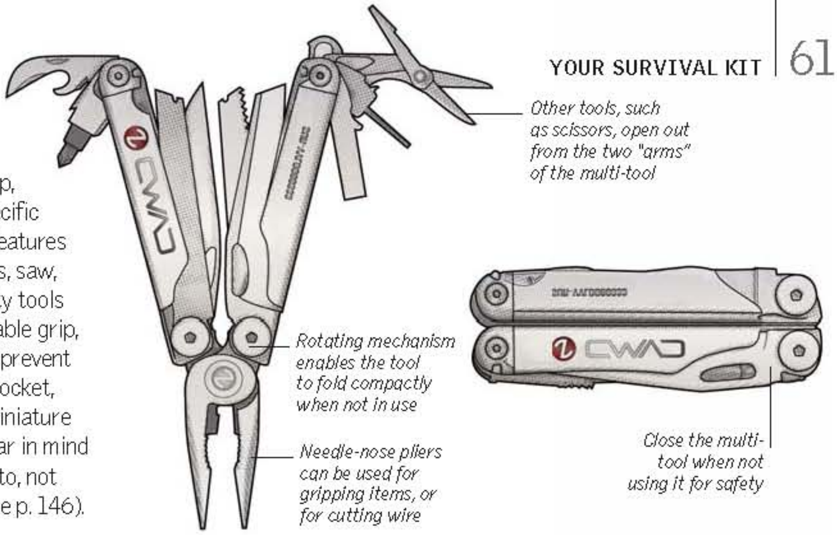

Other tools, such as scissors, open out from the two "arms" of the multi-tool

Rotating mechanism enables the tool to fold compactly when not in use

Needle-nose pliers can be used for gripping items, or for cutting wire

Close the multi-tool when not using it for safety

LAYER 4

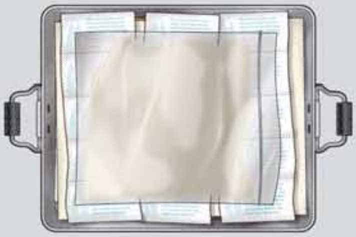

- **Petroleum jelly:** apply to chapped lips, rashes, and sores. Can also be smeared onto tampons to make them burn longer. Store in a small, resealable plastic bag.

LAYER 5

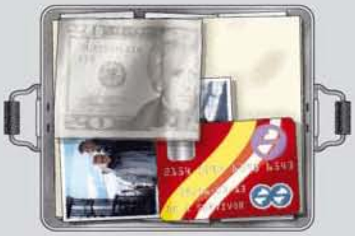

- **Waterproof notepaper:** for drawing maps or leaving messages.
- **Photograph of loved ones:** a psychological incentive in a survival scenario.
- **Credit card:** an effective means of extracting insect stings (see pp. 266–67).
- **Money:** wrapped in cellophane.

LAYER 9

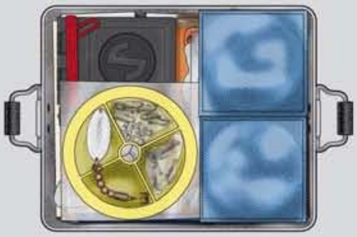

- **Non-lubricated condoms:** can be used to carry water, or as a waterproof cover for smaller items, such as your mobile phone.
- **Mini fishing kit:** if you are near water, fish can be easier to catch, prepare, and cook than mammals. Fishing line can be also used for other survival tasks. Should contain a selection of hooks, flies, swivels, and split-shots.

TIN LID SEALED WITH TAPE

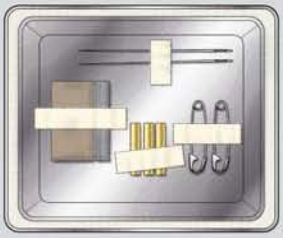

- **Sailmaker's needles:** multi-purpose—can be used as an arrow point, or for mending tents and tarpaulins. Wide eyes are best.
- **Safety pins:** for securing clothing, or mending your sleeping bag or tent.
- **Mini glowsticks (cyalumes):** useful for emergency lighting, and as a location aid.
- **Single-edged razor:** multi-purpose (see Layer 6). Store in its packaging.

ADDITIONAL USEFUL ITEMS TO INCLUDE

While your survival tin's size may be restrictive, you can always improve your kit with items that can fit on your belt, or in a belt pouch, known as a "belt order." This will form part of your first-line equipment (see p. 43).

- **Space blanket or aluminum foil:** can be used as a signaling device; for shelter; to carry, store, and heat water; or to cook in. Many are dual-sided: one silver, the other green for camouflage, or orange to stand out.
- **Plastic bags:** you can never have too many. Numerous uses—from a water carrier to a transpiration bag.
- **Medicines (such as painkillers and antibiotics):** should not replace your main first-aid kit, but ensures you have some basics if you're separated from your gear.
- **Small candle:** once lit, this will provide a reliable flame that you can build your fire around.
- **Nylon stockings:** can be used for warmth, or as an improvised water filter, mosquito net, or fishing net.
- **Small AM/FM radio:** battery- or solar-powered.
- **Surgical tubing:** enables you to reach water in otherwise inaccessible rock crevices.
- **Fire tin or matchless fire set:** self-contained methods of starting a fire, when no natural fuel is available.

2

ON THE TRAIL

FIND YOUR WAY

Before you set off on an expedition, you should have at least a basic understanding of how to read a map and use a compass. Your ability to correctly assess a map of the area you intend to visit will allow you to make informed decisions during your preparation. If you understand the area and terrain, your chances of getting lost will be reduced, and you will be able to continually evaluate your progress and therefore alter your plans as necessary. You will also be able to plan the safest and most appropriate route, and locate water, shelter, and areas that will allow you to use your location aids properly. If you're proficient with a map and compass, you'll have no cause to worry about getting lost or straying off-track and will be free to enjoy your outdoors experience.

In a survival situation, you will be faced with many tough decisions. You may have to decide whether to stay where you are and await rescue or move to an area that offers a better chance of survival and rescue. Your ability to navigate effectively—whether

In this section YOU WILL DISCOVER...

- how to **use your map** to tell the **lie of the land...**
- that the **path ahead** is as long as a **piece of string...**
- the difference between **eastings** and **northings** and an **ERV** and a **GPS...**
- how a **detour** could be the **most direct route...**
- how to **navigate** using your **hair...**
- why an **anvil of cumulonimbus** is no **fair-weather cloud...**

by using a map and compass or by using natural features—will play a major role in your decision-making process. Whilst a Global Positioning System (GPS) is an excellent aid, it relies on batteries and technology—both of which can fail.

An understanding of weather patterns will also allow you to evaluate conditions and make informed decisions as you travel. The ability to assess your situation and modify your plans means that you will be able to avoid many potential survival situations.

The ability to take a bearing (see pp. 68–71) and navigate using pace counting (see p. 72) could prove to be crucial skills in a survival situation.

TEST HOW ACCURATELY you can walk on a set bearing and pace out a set distance by following the exercise below. If you have stayed on your bearings and your pacing has been accurate, you should finish at your starting point. Pick a suitable area where you can walk at least 330 ft (100 m) in any direction. Don't cheat by heading for your marker! For this exercise to be useful, make sure you follow your bearings and count your paces to navigate.

1. Place a marker on the ground. Dial a bearing onto your compass (110° in our example).
2. Walk on this bearing, counting your paces until you think you have traveled 330ft (100m), then stop.
3. Add 120° to your original bearing. Dial this new bearing (230° in our example) onto your compass.
4. Pace out another 330 ft (100 m) on your new bearing, then stop.
5. Add another 120° to your latest bearing. Dial a last bearing (350° in our example) onto your compass.
6. Walk on your final bearing for a further 330 ft (100 m). You should be back at your starting point.

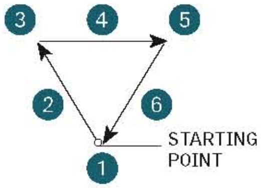

> Never underestimate the **combined power** of a map, a compass, a GPS unit, and the **skills and knowledge** to use them well—your life may **depend on it**

MAPS AND MAP-READING

A MAP IS A TWO-DIMENSIONAL representation of a three-dimensional area—from a map you can determine distance and height on the ground. If you are able to read a map, and can interpret the information it contains, you can visualize what an area looks like and use these features as landmarks to make navigation easier. Your map is very important, so keep it safe.

UNDERSTANDING MAPS

Although there are many types of map available—with varying levels of detail and scale—topographic maps are best for hiking. They show important features, such as rivers, roads, railroads, paths, buildings, and forested areas, and also depict the lie of the land through the use of contour lines (see opposite) to represent height.

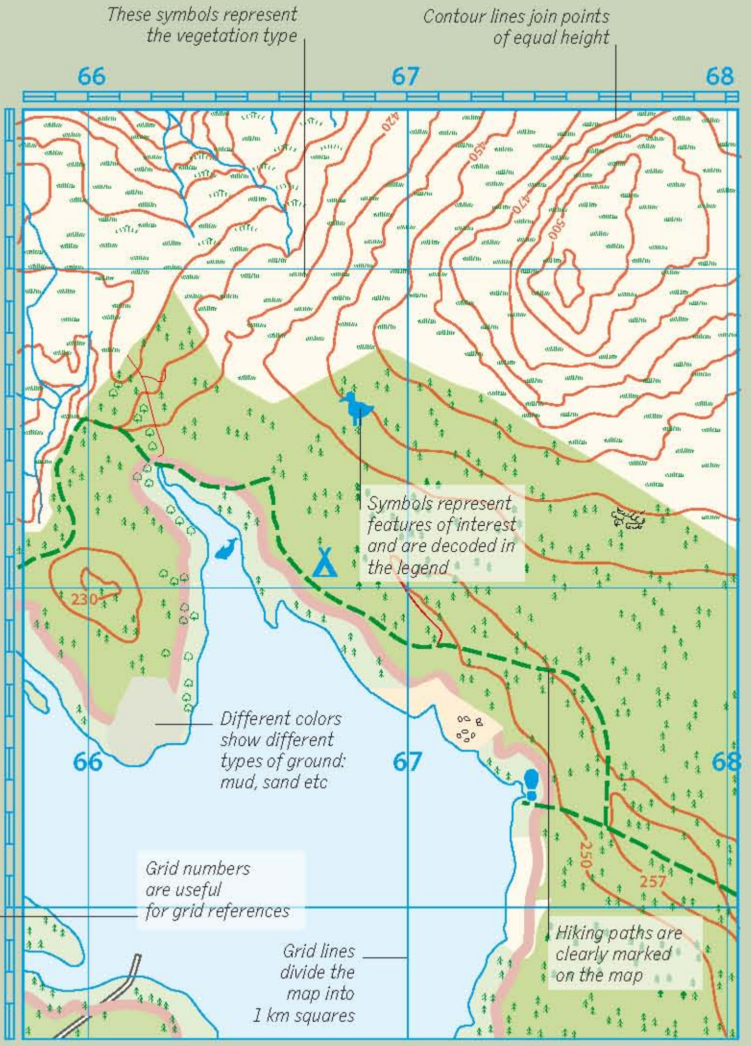
These symbols represent the vegetation type
Contour lines join points of equal height
66
67
68
96
95
94
420
450
470
500
230
250
257
Symbols represent features of interest and are decoded in the legend
Different colors show different types of ground: mud, sand etc
Grid numbers are useful for grid references
Grid lines divide the map into 1 km squares
Hiking paths are clearly marked on the map

THE LEGEND

Topographic maps incorporate a legend, or key, which deciphers the information shown on the map. It's important to familiarize yourself with the symbols used, as this will help you visualize what's being represented on the map more effectively. Some examples are listed below:

HEIGHTS AND NATURAL FEATURES

🟦 Water ▭ Mud

🟧 Sand; sand and shingle

Vertical face/cliff

Loose rock
Boulders
Outcrop
Scree

Coniferous trees

Non-coniferous trees

Managed woodland

Orchard

Scrub

Bracken, heath, or rough grassland

Marsh, reeds, or wetlands

TOURIST AND RECREATIONAL INFORMATION

Nature reserve
Campsite
Fishing area
Walks/Trails

THE SCALE

Hiking maps are always drawn to a scale provided in the legend. This is a ratio of how much you would have to enlarge the map to reach actual size. 1:25,000—whereby 4 cm on the map equates to 1 km on the ground (2½ in to 1 mile)—is the most useful level of detail for hikers. A smaller-scale of 1:50,000, for example, will provide a more basic overview of the terrain.

MEASURING DISTANCE

Maps are drawn to scale so that you can use them to accurately estimate distances on the ground. Being able to measure distance is important because it means you can calculate the most direct—and energy-efficient—route to your destination. In a survival situation, every last bit of energy counts, so the shorter the distance, the better.

USING THE GRID LINES

The simplest method of measuring distance on a map is to use the grid lines—on a scale 1:25,000 map each grid square represents 1 km (traveling diagonally across a square, it's approximately 1.5 km). You can also lay a piece of paper between the two points, mark the start and end of your route, and place it underneath the scale line to read the distance.

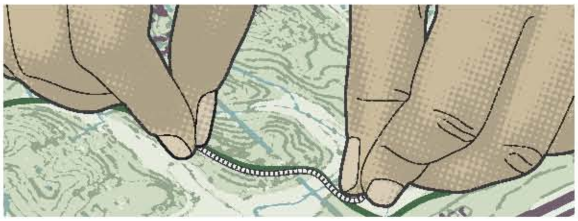

USING STRING OR SOLDER

Invariably you will deviate from a straight line and will have to navigate around obstacles or bends. A far better way of measuring distance is therefore to take a piece of string, curve it around your intended route, and transfer it to the scale line. Solder wire (use wire that is lead-free) is even more accurate because it holds its shape on the map and remains flexible.

CONTOUR LINES

Topographic maps feature lines called contour lines. A contour joins points of equal height above sea level, and allows the topography of the ground to be depicted in detail. The contour interval is specified in the legend—for maps with a 1:25,000 scale you would usually expect to see a 5-meter vertical interval between each line, although for maps of mountainous regions this interval may be 10 meters. The ability to look at the contour lines, and imagine how they translate to the ground, is a skill that takes a little while to acquire, but once mastered will allow you to read a map more proficiently.

USING CONTOURS

Knowing how steep the ground is will greatly improve your navigational skills, as well as your route-planning (see p. 73). Walking up and down hills uses a lot of energy, so it is far better to follow the contour lines on your map to go around hills instead.

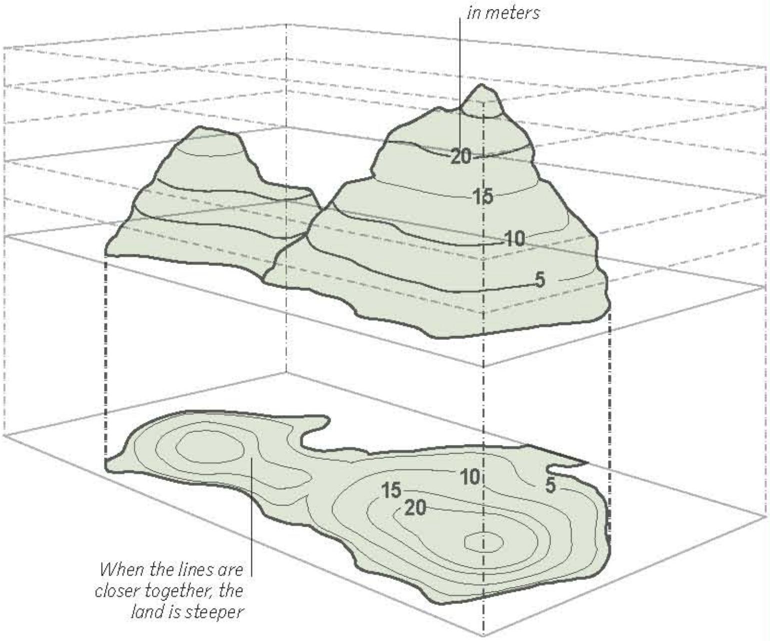

These numbers show the height above sea level in meters

When the lines are closer together, the land is steeper

GRID REFERENCES

Maps contain grid lines, which help you locate a specific point anywhere on a map, using a unique number known as a grid reference. The vertical grid lines are called "eastings," as they increase in value as they travel east on the map. The horizontal lines are called "northings," where these perpendicular lines cross creates the grid.

WORKING OUT GRID REFERENCES

Use the numbers on the grid lines and apply the easting number first. On a scale 1:25,000 map, where the grid lines are 1 km apart, the shaded area on the grid below would have a four-figure grid reference of 2046, and indicate a grid square that is 1 km by 1 km. To be more accurate, use a six-figure number—mentally divide the square into tenths. The cross would therefore have a grid reference of 185445.

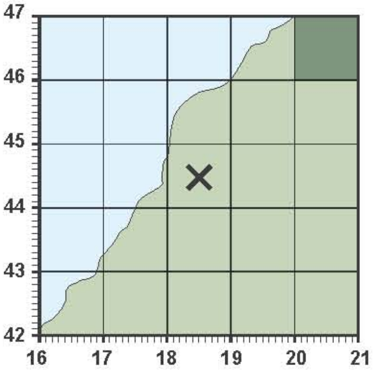

GET YOUR BEARINGS

IN ADDITION TO understanding and reading maps, it's important to know how to orientate your map to the land so you can use it to navigate. It's possible to do this by sight, although in most cases you'll need a more reliable method and for this you should use a compass. Use your compass to determine direction, orientate your map and yourself, take and plot bearings, and navigate.

CARDINAL POINTS

The four cardinal points of the compass are:

- North (N)—0°/360°
- East (E)—90°
- South (S)—180°
- West (W)—270°

HOW A COMPASS WORKS

A compass needle is a magnetized piece of metal that, when allowed to rotate freely, will orientate itself to the North and South magnetic poles. Always hold a compass level.

STANDARD COMPASS

This standard orienteering Silva compass is a good compass for hiking. It allows you to set your map and work out distances using the printed scale on the base.

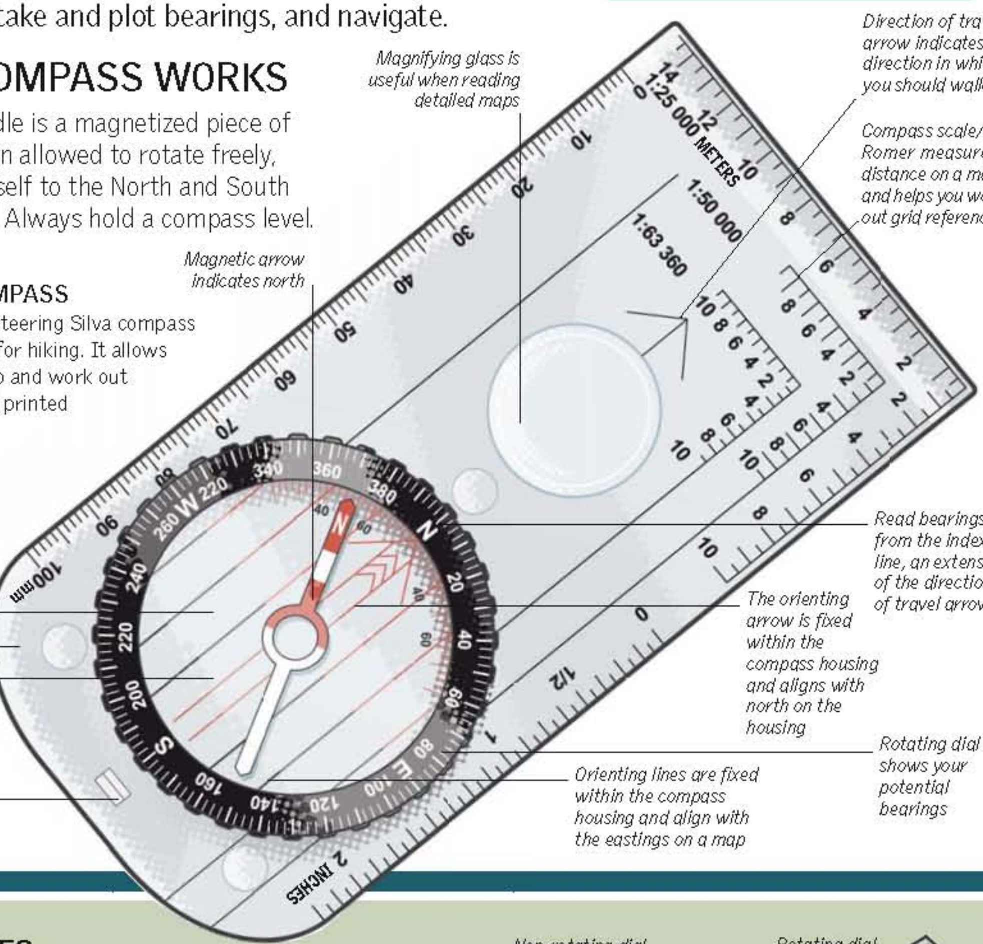
Magnifying glass is useful when reading detailed maps
Direction of travel arrow indicates the direction in which you should walk
Compass scale/ Romer measures distance on a map and helps you work out grid references
Magnetic arrow indicates north
Compass housing contains the needle and has the points of the compass marked on a circular, rotating bezel
Base plate
Liquid inside the housing dampens the needle to allow it to settle more quickly
Hole for attaching a lanyard
Read bearings from the index line, an extension of the direction of travel arrow
The orienting arrow is fixed within the compass housing and aligns with north on the housing
Rotating dial shows your potential bearings
Orienting lines are fixed within the compass housing and align with the eastings on a map
1:25 000 METERS
1:50 000
1:63 360
100mm
2 INCHES
1/2
N
W
S
E

OTHER TYPES OF COMPASS

There are many different types of compass, ranging from simple button compasses to complex instruments that include features such as sighting mirrors. Use a quality compass that suits your needs, and carry a simpler backup for emergencies.

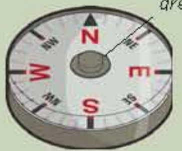
Only the points of the compass are shown
N
E
S
W

BUTTON

The simplest compass available—its small size and basic level of detail makes it ideal as a backup compass.

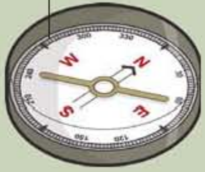
Non-rotating dial
N
E
S
W

FIXED DIAL

As well as the basic points of the compass, degrees and bearings are also shown on the immovable dial.

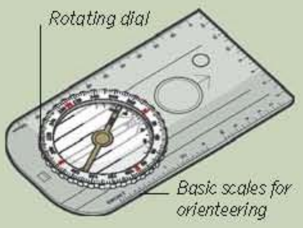
Rotating dial
Basic scales for orienteering

BASIC ORIENTEERING

With a rotating dial, this has simpler markings than the standard compass above, and is ideal for beginners.

SETTING YOUR MAP

Walking with your map set allows you to read the ground from the map as you pass over it, and to recognize and predict features as you progress, which means that you'll soon realize if you are heading off course. In some cases, when you have a good view of the terrain around you, and know the approximate area you're in, you can simply rotate your map until the features line up with the corresponding features on the ground. However, using your compass is a far more accurate method.

1 Rotate the dial so that "N" sits under the index line. Lay your map flat on the ground and ensure there is nothing nearby that could affect your compass (see box below).

- Lay the compass on the map so its edge runs parallel with a vertical grid line (easting).

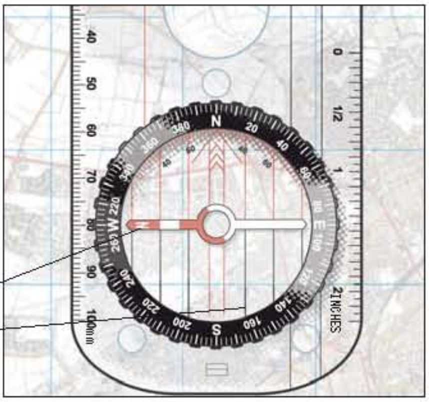
Needle will not be aligned at present

Orienting lines run parallel to the vertical grid lines on your map

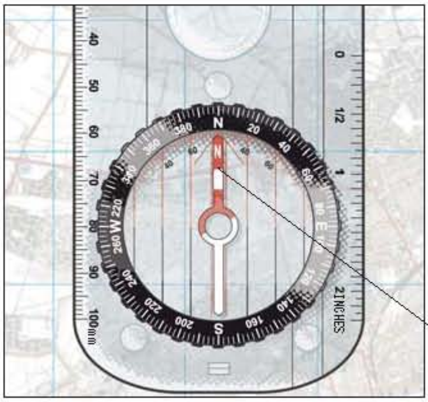

2 Keeping orienting lines aligned with the grid lines on the map, rotate the entire map until the north magnetic needle on the compass sits inside the orienting arrow. The map is now set to magnetic north and should basically line up with the features that surround you.

- If magnetic variation in your area is high (+ 5°), compensate accordingly (see right).

Needle is now aligned

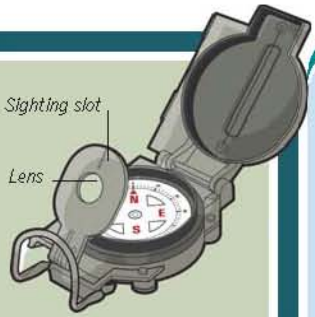

LENSATIC

The lensatic compass is an excellent instrument for taking very precise navigational measurements.

TAKING AN ACCURATE READING

Always hold the compass level to allow the needle to rotate easily. A compass is simply a magnetized piece of metal and, as such, is susceptible to interference. To avoid this, never use your compass near:

- Metal or other magnetized objects
- Electric currents, such as overhead high-tension power cables
- Buildings and vehicles—these often contain metal and electricity that might affect the accuracy of your compass readings

MAGNETIC VARIATION

Map legends refer to north in three ways: "true north," "grid north," and "magnetic north." The angle between magnetic north and grid north is known as "magnetic variation," and is provided in the map legend. True north is the direction of a meridian of longitude that converges on the North Pole. Grid north runs parallel to the vertical grid lines on a map and differs from true north because a map is flat. Magnetic north is the direction indicated by a magnetic compass.

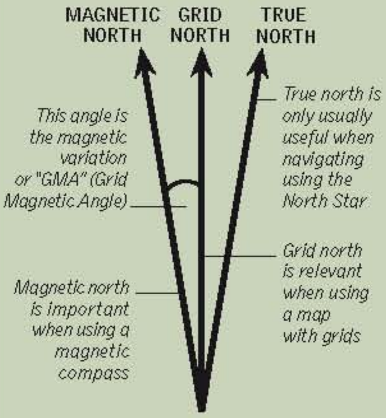
MAGNETIC NORTH
GRID NORTH
TRUE NORTH
This angle is the magnetic variation or "GMA" (Grid Magnetic Angle)
True north is only usually useful when navigating using the North Star
Grid north is relevant when using a map with grids
Magnetic north is important when using a magnetic compass

COMPENSATING FOR VARIATION

When converting a magnetic bearing to a grid bearing, or vice versa, you have to adjust for magnetic variation. When the variation is west, use the phrases "Mag to grid—get rid" or "Grid to mag—add." If the variation is east, the opposite applies.

1 Check the legend of your map to establish the magnetic variation. This depends on your location, and whether it is east or west of grid north.

2 If the variation is 0°, there's no magnetic variation affecting the compass, so make no adjustments.

3 If converting a magnetic bearing to a grid bearing with a 12° west variation, take off the 12°. With an east variation, add it on.

4 If converting a grid bearing to a magnetic bearing with a 12° west variation, add on the 12°. With an east variation, take it off.

TAKING BEARINGS

Always give any compass work your full attention. Rushing it, especially when working out bearings, can lead to navigational errors that could, at best, involve more walking and, at worst, get you lost.

USING A MAP

Using your map to work out the direction in which you need to walk is simple. Use your compass as a protractor to work out your bearing, and then to keep you on track.

BACK BEARING

A back bearing enables you to find your position by taking a bearing from a feature back to you. To do this, take a bearing to a point in the normal way and either add or subtract 180 degrees. You can also read the bearing exactly opposite to the index line. This is useful when working out the bearing from a feature back to you and transferring it to your map (see below).

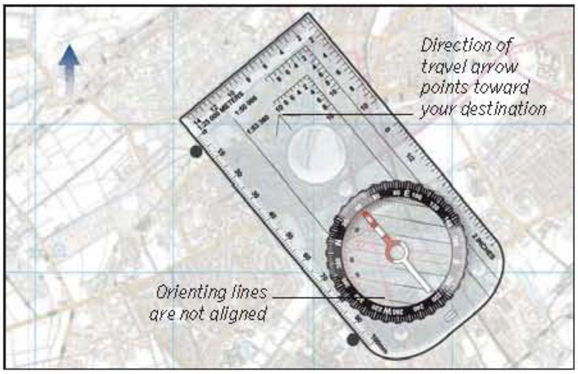
Direction of travel arrow points toward your destination
Orienting lines are not aligned

1 Lay your map on a flat surface, ensuring that there's nothing nearby to interfere with the compass reading (see box, p. 69).

- Lay the edge of your compass so that it runs between the point you want to navigate from and the point you want to navigate toward.
- Ensure that the direction of travel arrow on the compass is pointing in the direction you want to travel on the map.

USING FEATURES ON THE GROUND

Sometimes you may need to take a bearing to a specific point to navigate toward it. The point may be a feature that you can see at the time but may subsequently lose sight of during your journey because of the terrain. You can also plot that bearing and others on a map in order to work out your own position (see right, and panel, far right).

TAKING A BEARING ON A FEATURE

Pointing your compass at the feature, hold the base plate steady and rotate the compass housing until the orienting arrow sits directly under the north needle. Read the bearing—this is the magnetic bearing you would walk on to navigate to the feature.

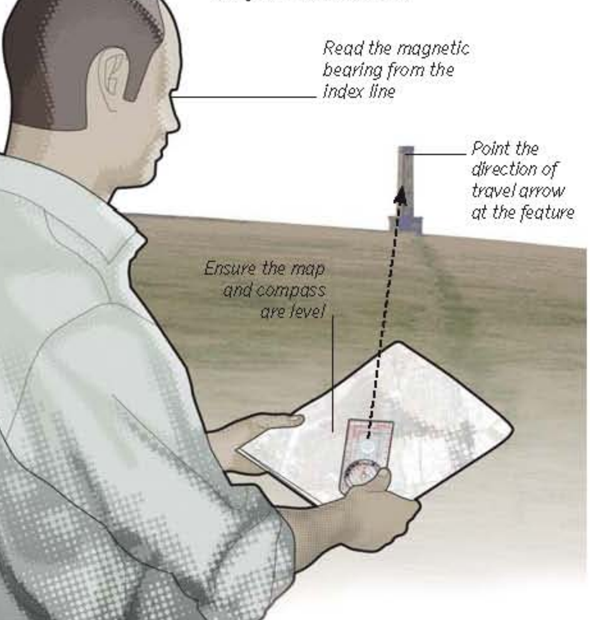
Read the magnetic bearing from the index line
Point the direction of travel arrow at the feature
Ensure the map and compass are level

TRANSFERRING A MAGNETIC BEARING TO A GRID BEARING

It is important that you know how to transfer a compass bearing from a feature (a magnetic bearing) onto a map (a grid bearing). For this example, imagine that the magnetic variation is 12° west.

- Take a bearing on your chosen feature (see left). In this example it is 45° magnetic.
- You want to plot this bearing on your map, which has a grid, so remember the phrase "mag to grid—get rid." You would therefore subtract the magnetic variation (12°) from the magnetic bearing (45°), which equals 33°.
- Dial this revised bearing into your compass.

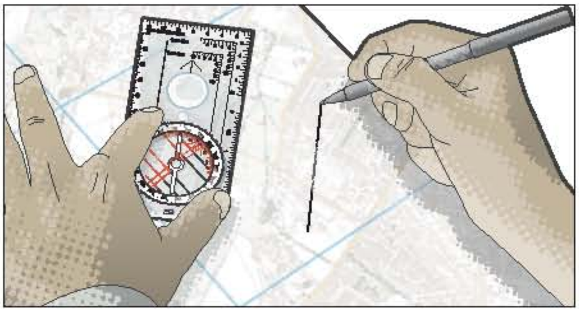

TRANSFERRING THE BEARING ONTO THE MAP

Place the top left corner of the compass base plate over the feature on your map. Keeping it there, rotate the entire compass until the orienting lines are parallel with the vertical grid lines. Draw a line from the feature down the left side of the base plate to map your bearing.

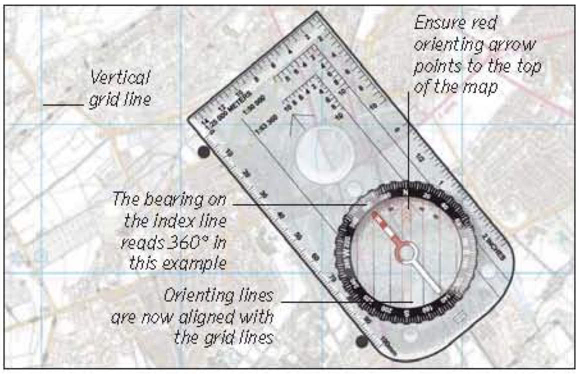

Vertical grid line

Ensure red orienting arrow points to the top of the map

The bearing on the index line reads 360° in this example

Orienting lines are now aligned with the grid lines

2 Turn the compass housing around until the orienting arrow and the orienting lines within the housing line up with the vertical grid lines on the map. Read the bearing between the two points from the index line on the compass.

- For this grid bearing to deliver you successfully to your destination, you'll need to convert it to a magnetic bearing by using the magnetic variation information on the map legend (see p. 69). Add or subtract your figure, and adjust your compass accordingly.

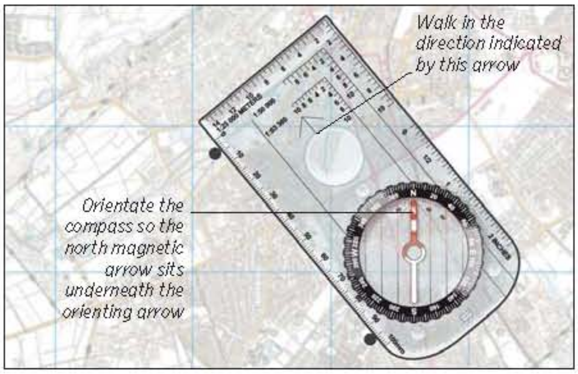

Walk in the direction indicated by this arrow

Orientate the compass so the north magnetic arrow sits underneath the orienting arrow

3 In order to walk on this bearing you must now orientate your compass.

- Hold the compass level and at a height that allows you to comfortably look squarely down on it (close to your chest is a good position).
- Turn your body until the north end of the compass needle sits inside the orienting arrow. The direction of travel arrow is now pointing exactly in the direction that you need to walk.

FINDING YOUR POSITION

If you are unsure of your position but can see features on the ground that you also recognize marked on your map, you can take bearings on these features with your compass and transfer them onto your map (see far left, and left) to accurately determine where you are. This process is called "triangulation" (also known as "resection" in military terminology). You'll need to transfer your compass bearings into grid bearings.

AT A KNOWN FEATURE

If you are at or on a known feature on a map—such as a river, road, or track—and you can see another recognizable feature, you can take a bearing on that feature and mark it on your map. Where that line crosses your known feature is your position. In the example below, you know you are somewhere along the banks of a river and can see a church that you can identify on your map.

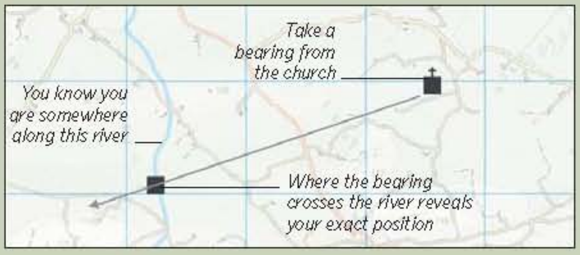

Take a bearing from the church

You know you are somewhere along this river

Where the bearing crosses the river reveals your exact position

NOT AT A KNOWN FEATURE

If you are not at a known feature, but can see other recognizable features on the ground that you can identify and locate on your map, you can take bearings on these features and transfer them to your map. In order to do this you need two features that are at least 1 km (6/10 mile) away and at least 40° apart. After you have transferred both bearings onto your map, the point at which the two lines cross will reveal your exact location. For greater accuracy, plot a third feature onto your map. The lines will cross and form a triangle—your position will be inside this triangle.

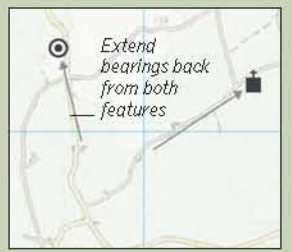

Extend bearings back from both features

1 Using your compass, take bearings to the features on the ground.

- Transfer these bearings to back bearings (see box, above left), and draw them on the map from your chosen features.

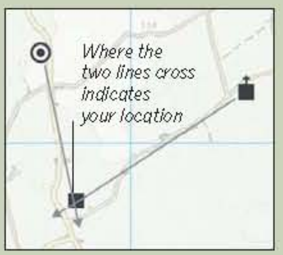

Where the two lines cross indicates your location

2 Extend these lines further over the map until they cross each other.

- The point at which the lines cross shows your location.
- If you want to be more accurate, repeat the process with a third feature.

ROUTE FINDING

WHETHER YOU ARE PLANNING to walk a specific route, or are in a survival situation and need to move to a safer area or to one that offers better chances of rescue, being able to study your map and calculate what you can achieve in a certain time can mean the difference between reaching your destination or spending a night in the wilderness.

ROUTE PLANNING

Break your route down into small sections. This will focus your navigation and make the overall distance seem less daunting. If you're in a group, designate an emergency rendezvous (ERV) point within each section—if someone gets separated you should all head there. If you can, include water sources and a safety point, such as a campsite, on your route.

CALCULATING DISTANCE

There are several methods of calculating the distance you are walking, and a seasoned hiker will always use at least two of them at any one time.

PACE COUNTING

Pace counting involves knowing how many paces you take to cover a set distance and then counting them as you travel. It is reasonably accurate. Distances are usually calculated in meters and kilometers, and most people take approximately 60 paces (120 steps) every 100 m. Try any of these methods:

- Cut 10 notches in your walking staff. Move an elastic band down a notch every 100 m.
- Use pace counting beads (a piece of webbing with two sets of beads divided by a central knot). One side contains nine beads to count off every 100 m, and the other has four beads to count off kilometers.
- Place 10 small pebbles in your pocket. Every time you cover 100 m, transfer a pebble to the other pocket.

NAISMITH'S RULE

Naismaith's rule takes into account distance and topography, and is used for estimating the duration of hikes.

- Allow one hour for every 3 miles (5 km) you will travel
- Add 30 minutes for every 985 ft (300 m) you will climb
- Subtract 10 minutes for every 985 ft (300 m) you will descend. However, for very steep slopes you should add 10 minutes for every 985 ft (300 m) you will descend

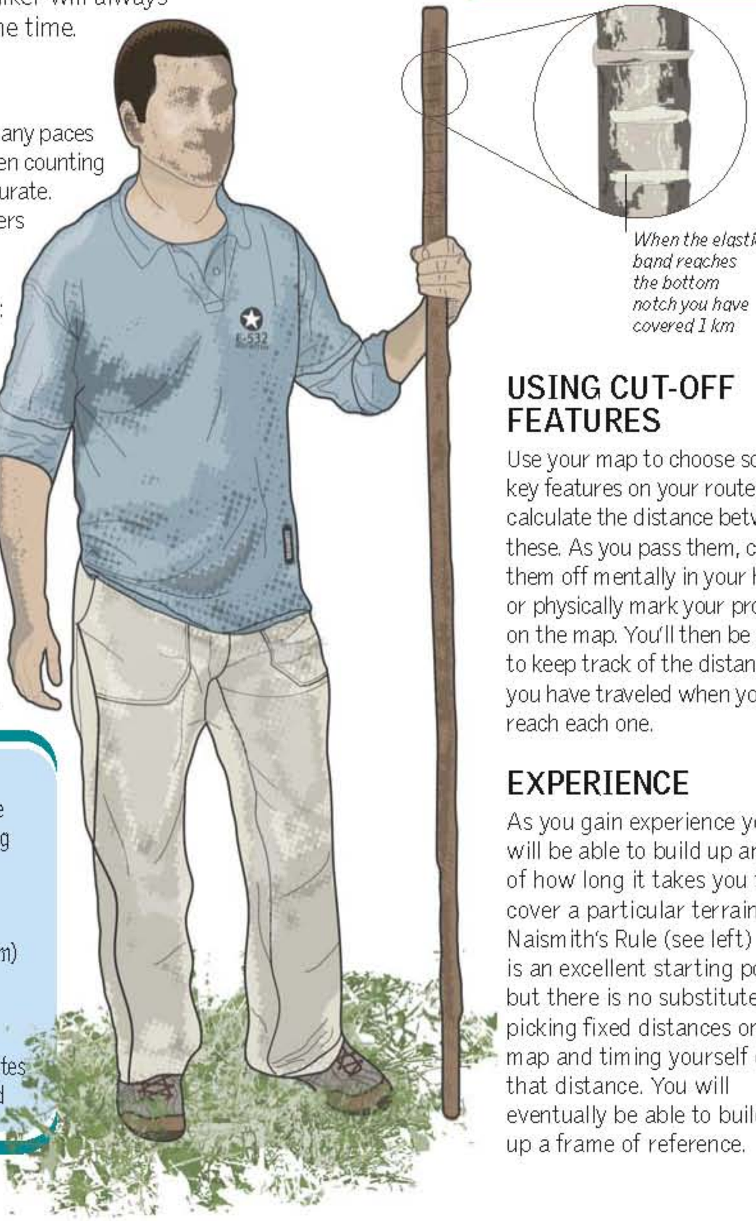

When the elastic band reaches the bottom notch you have covered 1 km

USING CUT-OFF FEATURES

Use your map to choose some key features on your route, and calculate the distance between these. As you pass them, check them off mentally in your head or physically mark your progress on the map. You'll then be able to keep track of the distance you have traveled when you reach each one.

EXPERIENCE

As you gain experience you will be able to build up an idea of how long it takes you to cover a particular terrain. Naismith's Rule (see left) is an excellent starting point, but there is no substitute for picking fixed distances on your map and timing yourself over that distance. You will eventually be able to build up a frame of reference.

NAVIGATION TECHNIQUES

When navigating across land, you're less likely to get lost if you take direct bearings from one feature to another. Unfortunately this may not always be possible; obstacles such as lakes and swamps may be directly in your path and it may be easier to walk around some features rather than walk over or through them.

PRINCIPLES OF NAVIGATION

You will never get lost if you know:

- Where you started from
- What bearing or course you have been traveling on
- How far you have traveled

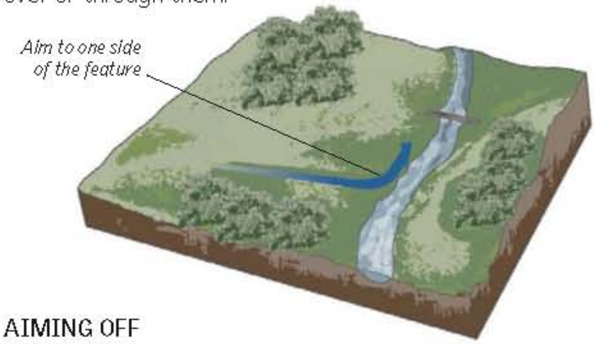
Aim to one side of the feature

AIMING OFF

It's easy to find yourself slightly off-course after a while. If you were aiming for a small footbridge over a stream and didn't arrive exactly at the bridge, you would need to guess whether to turn left or right to reach it. By deliberately aiming off to one side (also called "deliberate off-set") you can guarantee this direction.

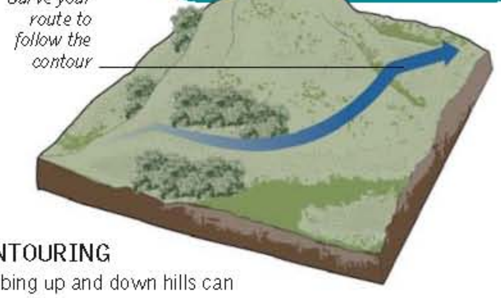
Curve your route to follow the contour

CONTOURING

Climbing up and down hills can expend a lot of energy and may not be the most effective method of navigating a particular terrain. Instead, use a technique called "contour navigation." This involves walking at the same height around a feature, which will conserve energy.

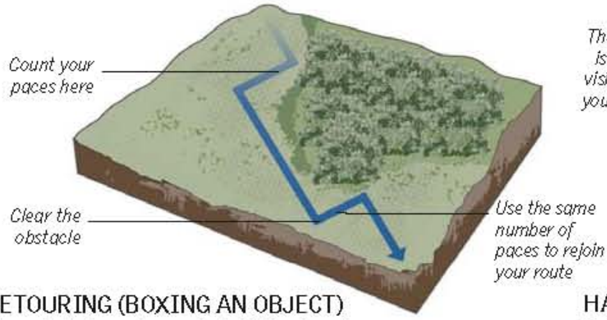
Count your paces here
Clear the obstacle
Use the same number of paces to rejoin your route

DETOURING (BOXING AN OBJECT)

If an obstacle makes a straight-line bearing impossible, use your compass to calculate four 90° turns, which you will then need to walk on to pass the obstacle. Count your paces on the first and third detours to return to your original, intended route as accurately as possible.

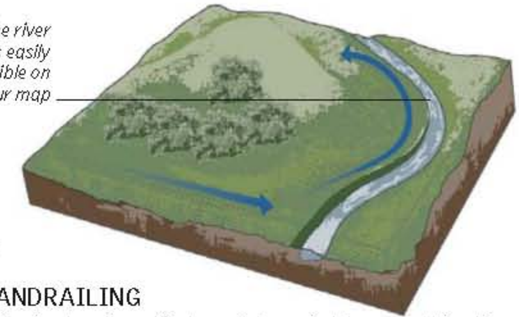
The river is easily visible on your map

HANDRAILING

Following long linear features that run in the general direction of your travel—such as rivers, roads, or paths—can be an effective way of navigating. Because you use the features rather than your compass, navigation becomes simpler, as the features are usually easy to follow. A slight detour toward one can be worthwhile.

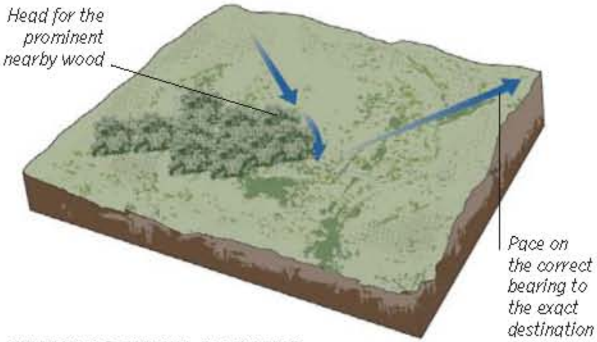
Head for the prominent nearby wood
Pace on the correct bearing to the exact destination

STAND OFF/ATTACK POINT

Useful when navigating to a specific point that may be difficult to locate, this technique involves aiming initially for a nearby prominent feature. Calculate a distance and bearing from it and use pacing to accurately locate the specific point. If you fail to find it, simply return to the prominent feature and start again.

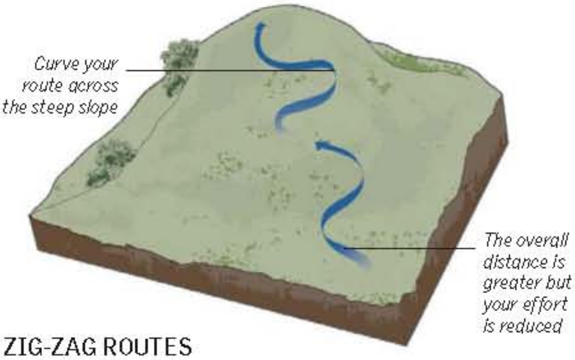
Curve your route across the steep slope
The overall distance is greater but your effort is reduced

ZIG-ZAG ROUTES

You may have no alternative but to climb a steep slope and this can be exhausting. However, if you choose a zig-zag route up the slope you can reduce the effort required to achieve the climb. This will dramatically reduce the strain on your legs, ankles, lungs, and heart. It is also effective when walking down steep slopes.

NAVIGATING WITHOUT A COMPASS

IF YOU DON'T HAVE A COMPASS you can use a few simple items from your survival kit to determine direction and enable you to navigate reasonably accurately. Advances in technology have also made satellite-navigation aids more readily available.

TOOLS AND MATERIALS

You will need an improvised needle or razor blade, plus one of the following:

- Knife
- Hair
- Silk
- Magnet
- Battery and wire; paper and tape

IMPROVISING A COMPASS

It's relatively simple to make an improvised compass using a piece of magnetized ferrous metal. How accurate your compass is depends on the materials you have available to you and your own ingenuity.

SOURCING AND MAGNETIZING THE NEEDLE

If you have a compass but it's damaged, you may still be able to use the needle from it, which will already be magnetized. If this is unusable, you'll need to find a piece of ferrous metal to magnetize. Suitable items you could use include:

- A needle from your survival tin or sewing kit
- A paperclip, opened up and straightened out
- A razor blade from your survival kit
- A small nail or straightened staple taken from a fence

Once you have your improvised needle, you will need to magnetize it. The smaller and thinner the needle is, the easier this will be. Use one of the methods shown here.

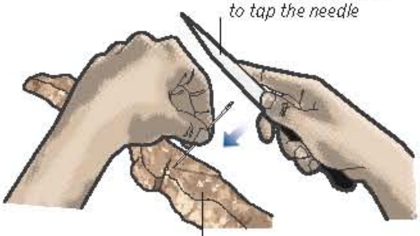

You can use the blade of your survival knife to tap the needle

Tap the needle into hard wood to heighten the effect, but make sure it doesn't become embedded

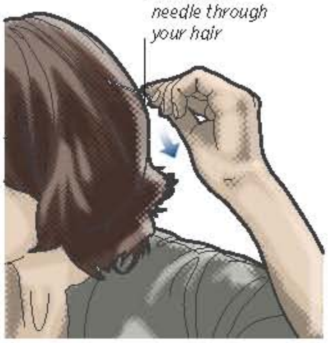

Stroke the needle through your hair

THE TAPPING METHOD

Align the needle as close to the north-south line as you can determine, hold the needle at an angle of 45 degrees, and gently tap the end of it with another piece of metal. Lightly tapping it into a hard piece of wood will increase the effect.

USING YOUR HAIR

Hold the sharp end of the needle perpendicular to your head and—taking care not to hurt yourself—stroke the needle in one direction through your hair, using careful and deliberate strokes. Repeat until the needle is magnetized.

ALLOWING THE NEEDLE TO FLOAT FREELY

Once you have made your improvised needle using one of the methods above, you need to find a way to allow it to turn freely so that it is able to indicate direction. Take care to protect your needle from elements such as the wind, which will affect its movement.

SUSPENSION METHOD

The advantage of the suspension method is that the equipment is portable and can be reused. It works best with a magnetized razor blade, which will balance well. Attach the magnetized blade to a cotton thread and suspend it inside a plastic bottle. If the bottle's neck is not wide enough to fit the blade through, remove the base of the bottle instead.

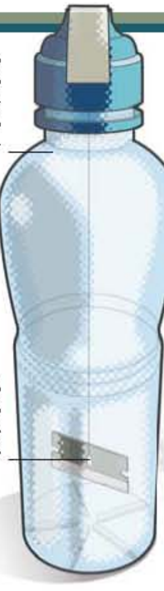

The bottle protects the compass from the elements

Use the hole in the blade to attach the cotton

FLOATING METHOD

In a sheltered place, float the needle on the surface of some water—a puddle or a small non-magnetic container filled with water, for example. Balance the needle on a small, dry leaf (or piece of paper, piece of bark, blade of grass, or inside a shortened straw).

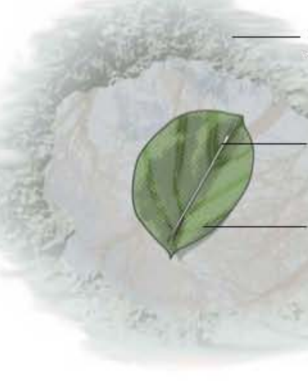

Use a puddle in a sheltered area

The needle will rotate, settle, and align itself

The leaf enables the needle to float on the surface of the water

PRINCIPLES OF MAGNETIZING

In general, the longer you work on the needle the stronger the magnetization will be and the longer it will last. To tell when it is magnetized sufficiently, hold the needle up against another metal object—if it is attracted to the metal and has enough strength to hold itself against it then it will be strong enough. Once the needle is magnetized you will need to allow it to float freely (see below), and then determine which end points north by using natural aids, such as the sun (see pp. 76–77). Mark the north end of the needle with a pen or a small scratch.

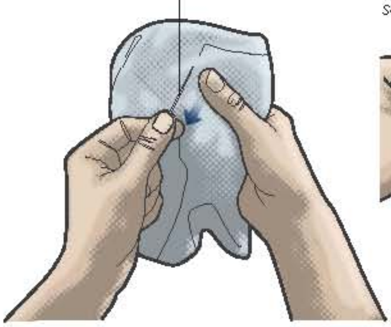

Stroke the needle 50–100 times

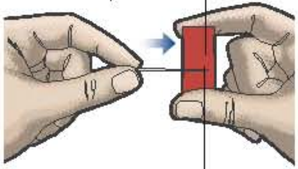

The end of the magnetized needle that attracts to the south pole of the magnet will point north

The more times you stroke, the more effective the magnetization

USING SILK

This works on a similar principle to the hair method but is more effective. If you have anything made from silk, such as a sleeping bag liner or thermal clothing, stroke the needle repeatedly in one direction against the silk.

USING A MAGNET

Stroke the magnet along the length of the needle repeatedly in one direction. It's a good idea to carry a magnet with you at all times, although you should never keep it near your compass as it will affect its accuracy.

USING ELECTRICITY

The most effective way to magnetize a needle is to pass a small electrical current around it. Use a battery and insulated wire; alternatively, use brass snare wire from your survival kit and insulate it using something non-conductive, such as paper.

1. Wrap the needle in a small piece of paper, which will insulate it from the electrical current.

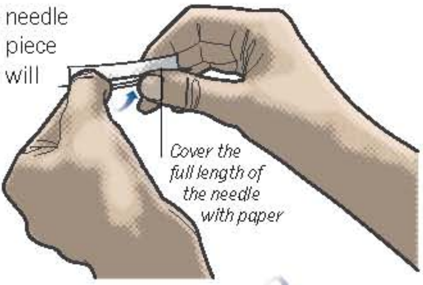

Cover the full length of the needle with paper

2. Wrap the wire tightly around the full length of the insulated needle.

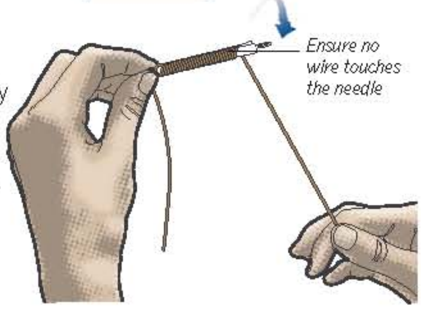

Ensure no wire touches the needle

3. Attach the wire to the battery until battery starts to get warm—this indicates the process is complete.

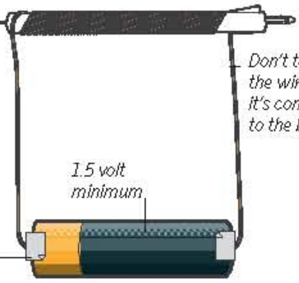

Don't touch the wire once it's connected to the battery

1.5 volt minimum

Attach the wire using some tape

USING GPS TECHNOLOGY

A Global Positioning System (GPS) is a hand-held unit that uses 24 orbiting satellites to triangulate your position on the Earth's surface to within meters. A GPS will allow you to work out straight-line distances and bearings to and from points, but unless it incorporates mapping it will not show you the best way to get there or take into account hazards on your route. Use it in conjunction with your map and compass.

TAKING EFFECTIVE GPS READINGS

A GPS needs to have a clear view of the sky. Anything that obstructs its signal, such as tall buildings or heavy tree canopy, will reduce its ability to lock onto satellites.

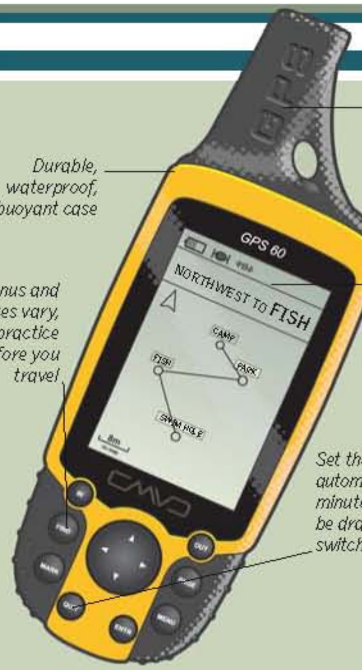

Durable, waterproof, buoyant case

Menus and features vary, so practice before you travel

Receiver locks onto at least four satellites to obtain a three-dimensional fix (latitude, longitude, and height) displayed as a grid reference

The screen lights up for use in gloomy conditions, but avoid using this feature if you can as it will drain the batteries

Set the unit to switch off automatically after two minutes so the battery won't be drained if the GPS is switched on accidentally

NATURAL NAVIGATION

IF YOU HAVE LOST or damaged your compass and don't have the materials necessary to make an improvised version, use natural indicators to determine direction. The east-west rotation of the Earth means that you can orientate yourself according to the position of the sun, moon, and stars. All you need are some very basic materials and these simple techniques.

SHADOW STICK BASICS

Use a shadow stick to determine direction and time anywhere between the Arctic (66.5°N) and Antarctic Circles (66.5°S).

- In the northern hemisphere the shadow of the stick will be on the north side of the east-west line
- In the southern hemisphere the reverse is true: the shadow of the stick will be on the south side of the east-west line.
- When the shadow is at its shortest, it is midday.

USING THE SUN

When visible, the sun is the clearest natural signpost to the four cardinal points (north, south, east, and west). It rises approximately in the east and sets approximately in the west and, at midday, is due south in the northern hemisphere and due north in the southern hemisphere. Use the sun's course across the sky to determine direction and approximate the time.

ORIENTATION

Tracking the movement of the sun across the sky using a shadow stick will provide an indication of its direction of travel. The sun moves from east to west at 15° an hour.

1. Drive a stick into a piece of level ground, and ensure it's as upright as possible.

2. Place a stone at the tip of the stick's shadow.

3. Wait three hours and place a second stone at the new position of the shadow's tip.

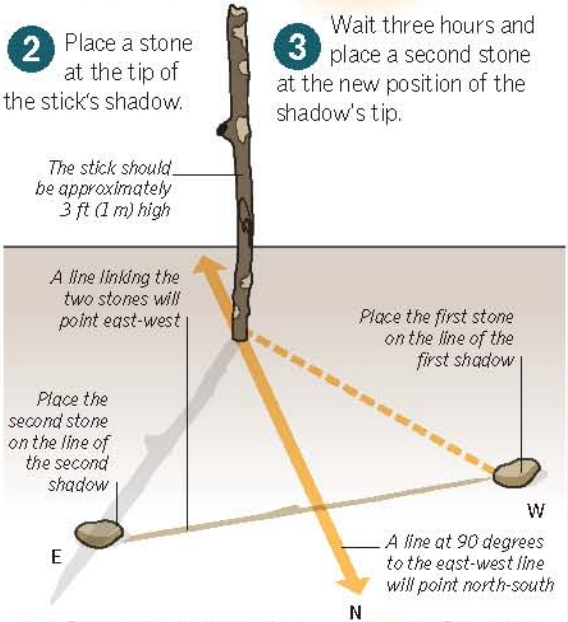

The stick should be approximately 3 ft (1 m) high

A line linking the two stones will point east-west

Place the first stone on the line of the first shadow

Place the second stone on the line of the second shadow

A line at 90 degrees to the east-west line will point north-south

4. Draw a line between the stones to find east-west; the first stone will be at due west and the second at due east.

5. To find north-south, mark a line at 90° to the east-west line.

TELLING THE TIME

Once you have established the east-west and north-south lines, you can turn the shadow-stick apparatus into a sundial, to give an approximate idea of the time.

1. Place the stick at the intersection of the north-south and east-west lines

2. Tie a piece of cord to the stick. Attach a smaller stick to the other end of the cord and use it to draw an 180-degree arc between the two marker stones.

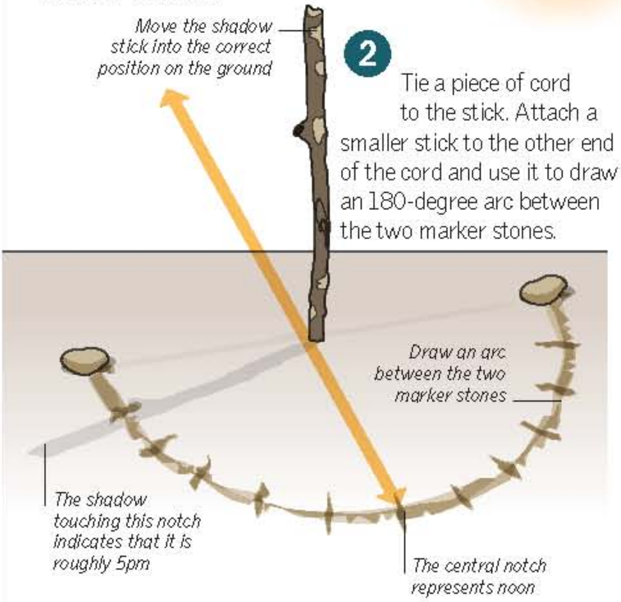

Move the shadow stick into the correct position on the ground

Draw an arc between the two marker stones

The shadow touching this notch indicates that it is roughly 5pm

The central notch represents noon

3. Divide the arc into 12 equal sections and mark each division with a notch. The notches represent one hour of time, from 6am to 6pm with noon at the middle notch.

USING THE STARS

It's possible to orientate yourself by spotting certain recognizable stars. In the northern hemisphere, Polaris is located above due north on the horizon, from which you can determine east, west, and south. Find it by locating the Plow (also known as the Big Dipper). In the southern hemisphere, you can work out the approximate position of south on the horizon by finding the Southern Cross.

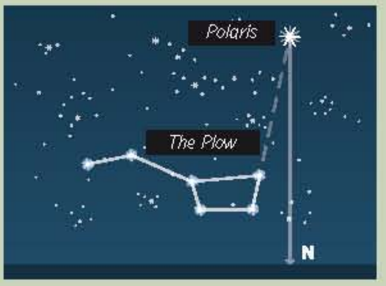

NORTHERN HEMISPHERE

Find the Plow and project a line from the top of the two stars that form its front. Follow this line until you find Polaris, which is located approximately four times the distance from the Plow as the distance between its two front stars.

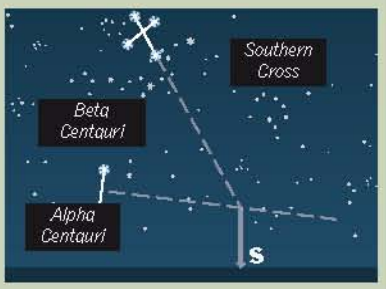

SOUTHERN HEMISPHERE

Project a line from the longer axis of the Southern Cross until you find a dark area of sky. Project a second line at 90° from the mid-point between two bright stars in the Centaurus constellation. Due south is below the point where these lines meet.

USING THE MOON

Reflecting the light of the sun, the moon rises in the east and sets in the west, so can be used for orientation. A shadow stick will work on a cloudless, full-moon night.

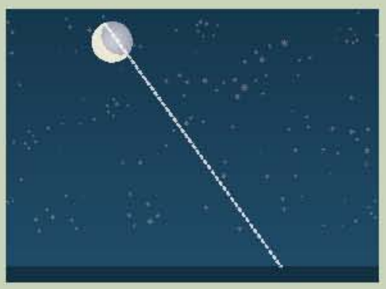

CRESCENT MOON

Although not an entirely accurate method, a line vectored between two horns of a crescent (quarter) moon will lead to a point that is approximately south on the horizon in the northern hemisphere, and roughly north in the southern hemisphere.

USING AN ANALOG WATCH

If you can see the sun, you can use an analog watch as a protractor to determine an approximate direction. Ensure it is set to the correct local time and that you have taken daylight savings (DST) into account. If you don't have a watch but know the time, simply draw a watch face on a piece of paper, marking 12 o'clock and the hour hand. This method is increasingly less effective as you near the Equator.

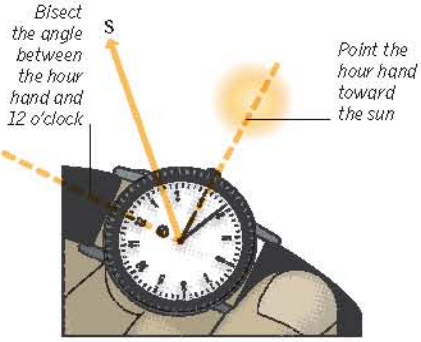
Bisect the angle between the hour hand and 12 o'clock
S
Point the hour hand toward the sun

NORTHERN HEMISPHERE

In the northern hemisphere the cardinal point nearest to the sun is south. Point the hour hand of your watch at the sun and bisect the angle between the hour hand and 12 o'clock. This will be due south.

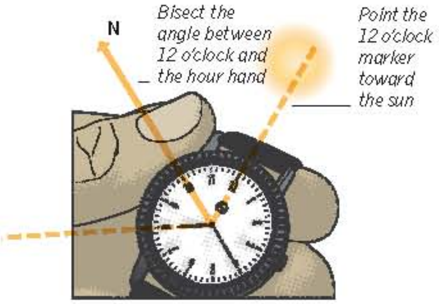
N
Bisect the angle between 12 o'clock and the hour hand
Point the 12 o'clock marker toward the sun

SOUTHERN HEMISPHERE

In the southern hemisphere, the cardinal point nearest to the sun is north. Point the 12 o'clock marker on your watch at the sun and bisect the angle between 12 o'clock and the hour hand on the watch. This will be due north.

NATURAL SIGNPOSTS

Nature responds to the elements in a range of ways, some of which can be studied for orientation tips. This is useful if you know the predominant wind direction of an area.

TREES AND PLANTS

- Windswept trees point away from the wind.
- Tree growth is most lush on the side that faces the sun (south in the northern hemisphere, north in the southern hemisphere).
- Some plants, such as the barrel cacti, twist toward the sun as they grow.
- Moss and lichen grow out of direct sunlight, on the shady side of rocks or trees.

ANIMALS AND INSECTS

- In very windy areas, small animals and birds tend to nest or burrow on the lee-side of hills.
- Spiders spin webs out of the wind, so an area with a lot of broken webs could indicate a recent change of wind direction, away from the dominant wind.

SNOW AND ICE

- In powder-snow conditions, snow "dunes" often form parallel to the prevailing wind.
- Frost erosion is most severe on slopes facing the sun (south in the northern hemisphere, north in the southern hemisphere).

HOW WEATHER WORKS

CHANGES IN THE WEATHER can have a significant effect on your expedition—or chances of survival if something goes wrong—so it's important to be as prepared as possible for all the conditions you might face. Check the weather report for the days ahead thoroughly, and pack and wear appropriate clothing. If the weather forecast is so bad that it could make traveling or navigation difficult, rethink your plans.

UNDERSTANDING THE WEATHER

Weather is created by the movement of air currents, the moisture content of the air, and the meeting of warm and cold fronts. Looking at weather maps before you leave for your trip will help you track these elements. You'll then need to use your understanding to forecast the effects of these movements on the ground.

READING WEATHER CHARTS

It can be helpful to compare weather charts with regular topographic land maps. Just as the gradient is steeper where contour lines appear closer together on land maps, the wind is stronger where isobar contours appear closer together on weather charts.

WHAT IS WEATHER?

The term "weather" relates to current conditions on the ground—such as the temperature, and whether or not it is windy or raining—while "climate" refers to a region's conditions over a longer period of time. Changes in the weather are caused by alterations in air pressure and temperature. Extreme weather, such as a hurricane, occurs when these changes are more marked than usual. Weather is monitored by meteorologists, and very accurate predictions can be made. However, weather can still cause surprises, so always prepare for all possible scenarios when packing or dressing for an expedition.

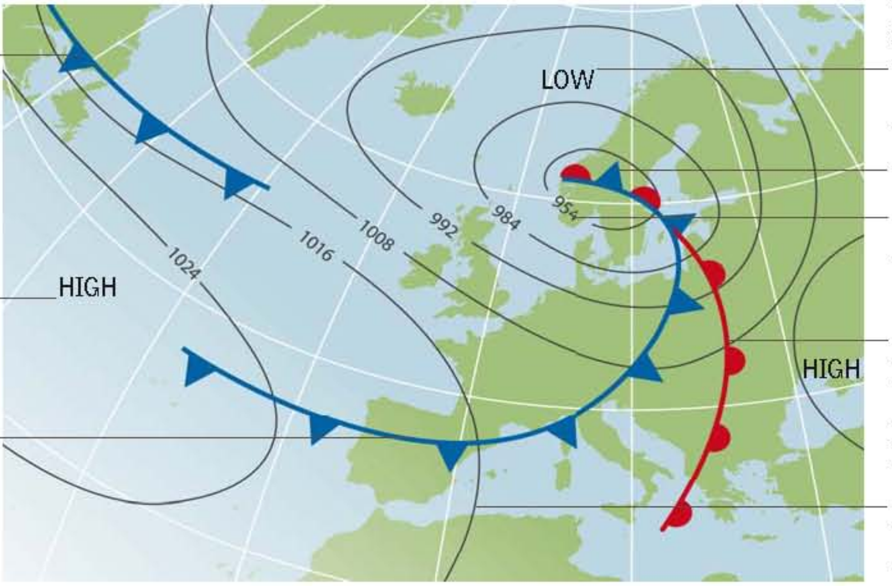

Fronts move in the direction that their symbols point toward

Areas of high pressure are generally associated with warm weather

A blue line with triangles represents a cold front, and marks the leading edge of cold air

Areas of low pressure are associated with cooler weather

Occluded fronts are fronts where two air masses merge

The lower the value of the central pressure, the more severe the wind and rain

A red line with semi-circles represents a warm front, and marks the leading edge of warm air

Isobars join areas where the air pressure is the same

HIGH
1024
1016

HIGH PRESSURE

Under conditions of high pressure, air descends in a spiral formation, and warms. The water vapor does not condense into clouds and we would expect to observe fairly clear skies.

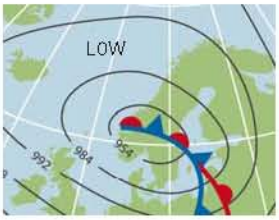

LOW PRESSURE

Under conditions of low pressure, air spirals upward—parallel to the center of the isobars—and cools. The air condenses to form clouds, and skies are usually overcast.

GLOBAL CIRCULATION

Warm air from the equator rises and moves toward the poles in each hemisphere. As it cools it descends and travels back toward the equator. Because of the Earth's rotation, three separate cells of circulating air in each hemisphere develop, which produce predictable wind and pressure patterns.

AIR MASSES

Large bodies of air with a particular temperature and humidity help weather forecasters predict the forthcoming weather and are known as air masses. They are largely defined by the area that they originated from. In general, northerly winds are colder than southerly winds, and air that has tracked over the sea ("maritime air mass") accumulates moisture and is more cloudy than air that has tracked over the land ("continental air mass"). The boundaries between tropical and polar air become warm and cold fronts.

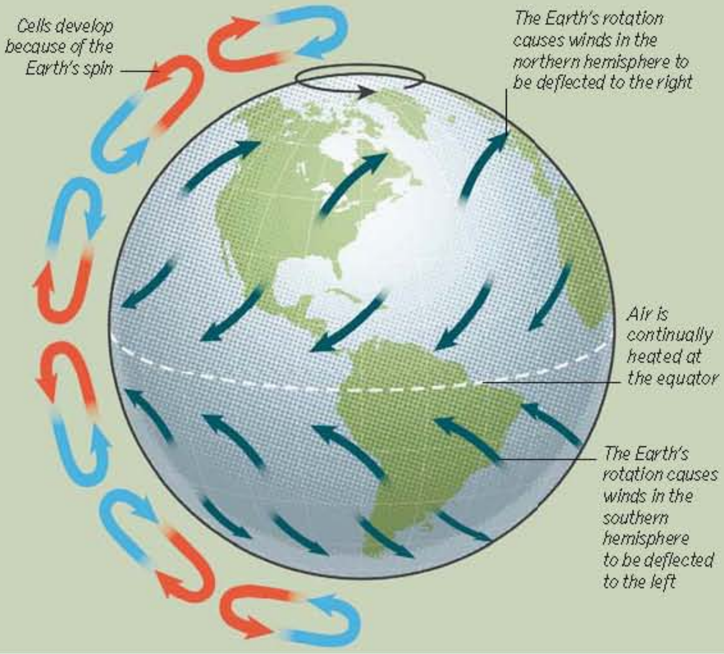
Cells develop because of the Earth's spin
The Earth's rotation causes winds in the northern hemisphere to be deflected to the right
Air is continually heated at the equator
The Earth's rotation causes winds in the southern hemisphere to be deflected to the left

ON THE GROUND

As a weather system travels over the ground, it changes in predictable ways. Knowing how the system will progress is an important part of interpreting your weather map or forecast.

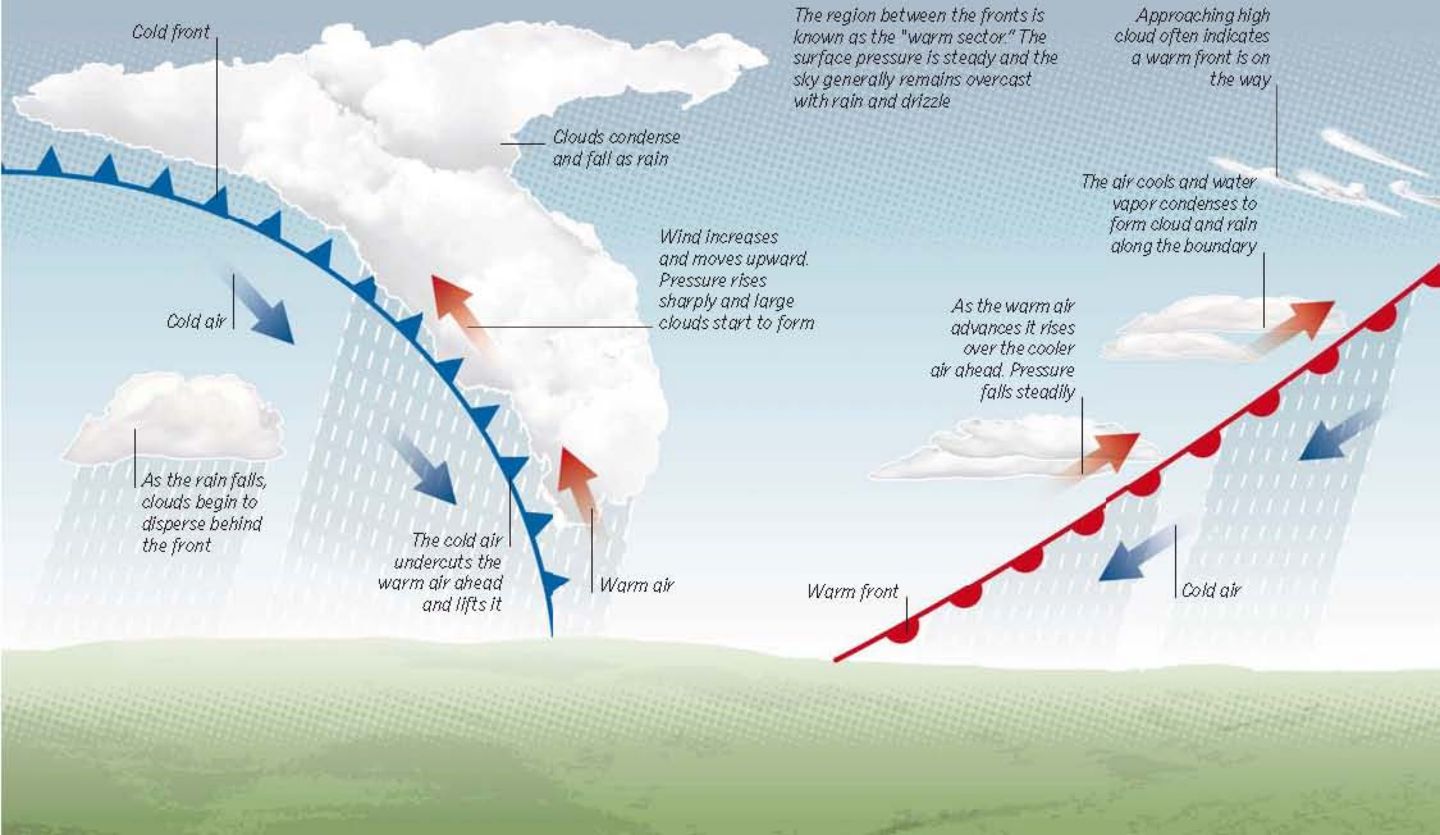
Cold front
Clouds condense and fall as rain
The region between the fronts is known as the "warm sector." The surface pressure is steady and the sky generally remains overcast with rain and drizzle
Approaching high cloud often indicates a warm front is on the way
The air cools and water vapor condenses to form cloud and rain along the boundary
Wind increases and moves upward. Pressure rises sharply and large clouds start to form
Cold air
As the warm air advances it rises over the cooler air ahead. Pressure falls steadily
As the rain falls, clouds begin to disperse behind the front
The cold air undercuts the warm air ahead and lifts it
Warm air
Warm front
Cold air

WEATHER PHENOMENA

LOOKING AT THE CLOUDS can help you to read approaching weather, which is invaluable when you're on the move with no access to weather forecasts or charts. Knowing how to recognize a storm cloud, for example, will ensure that you have enough warning to seek shelter or to change into appropriate clothing. If low cloud threatens to impair your visibility, use your compass to navigate and proceed with caution.

READING THE CLOUDS

Clouds are condensing masses of water vapor that provide precipitation and reflect solar radiation. They are categorized by height—into low, medium, and high clouds—and further defined by their shape. The shape of a cloud is determined by the way in which warm air rises, and is an indicator of the air stability.

HOW CLOUDS FORM

Clouds form by a cooling process. This process can be compared to breathing on cold glass, when condensation appears because invisible water vapor in your breath, cooled to a temperature known as the "dew point," condenses into its visible liquid state. Clouds form in a very similar way. Air temperature falls at an average rate of two degrees celcius per 1,000 ft (300 m). When air containing moisture—perhaps because it has travelled over sea—reaches the height at which it meets its dew point temperature, the water vapor condenses to form visible cloud.

HIGH CLOUDS

Fair-weather clouds are white and high in the sky. If there are no clouds, expect excellent conditions. Blankets of high cloud progressively invading the sky can signify the onset of bad weather. High clouds include:

- Anvil of cumulonimbus
- Cirrus
- Cirrustratus

MEDIUM CLOUDS

Thick layers of medium-layer cloud give heavy, persistent rain, especially if the clouds are dark and gray. The clouds you are most likely to see at this level are:

- Altocumulus
- Nimbostratus
- Altostratus

LOW CLOUDS

Low clouds have clearly defined edges and can indicate whether rain will fall in short downpours (cumulus) or persistently (stratus). Common clouds are:

- Cumulonimbus
- Stratus
- Stratocumulus
- Cumulus

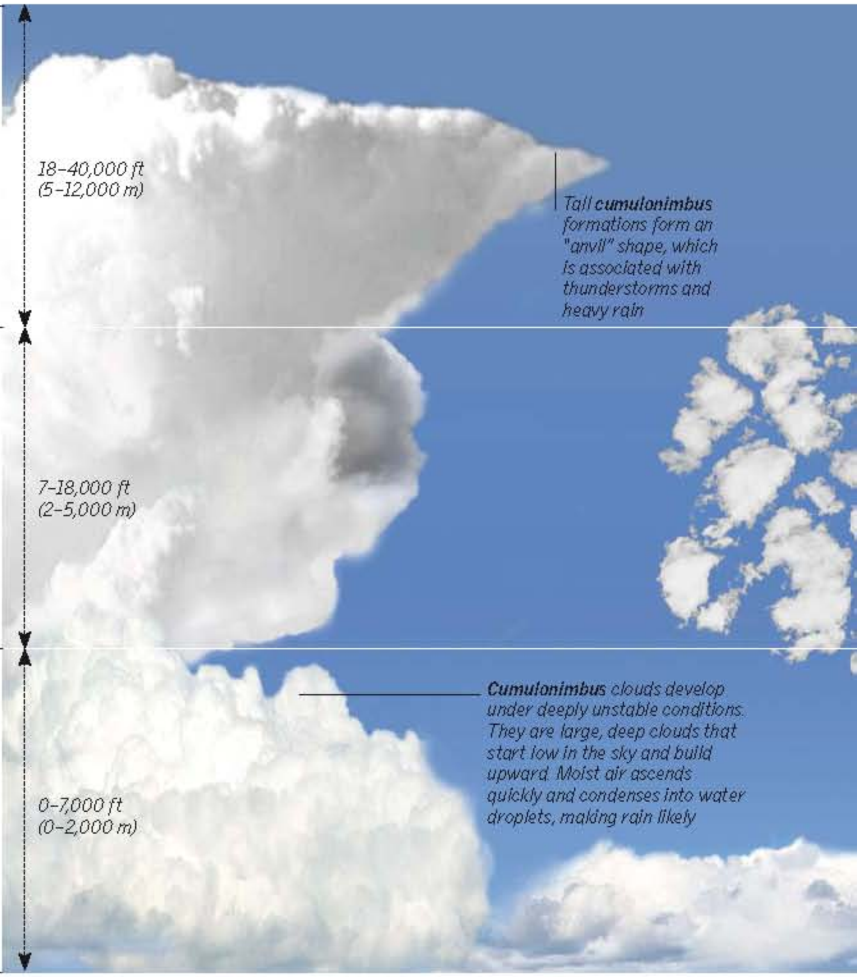

*Tall **cumulonimbus** formations form an "anvil" shape, which is associated with thunderstorms and heavy rain*

***Cumulonimbus** clouds develop under deeply unstable conditions. They are large, deep clouds that start low in the sky and build upward. Moist air ascends quickly and condenses into water droplets, making rain likely*

WEATHER RISKS

The weather can have a massive impact on your trip, bringing with it added dangers. It can affect your visibility levels, and the safety of the ground you're hiking over. In a survival situation, the temperature, and whether or not it's raining, can have a huge effect on your chances.

WARNING!
Never shelter under a solitary tree during a lightning storm—get in your vehicle if possible or find low ground.

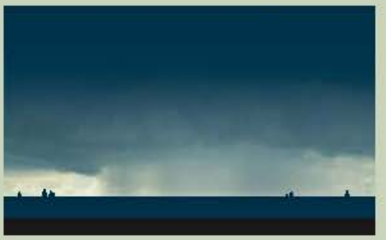

HEAVY RAIN

If heavy rain falls quickly, the ground may become saturated, slippery, or flood. Seek shelter or proceed cautiously in waterproof clothing.

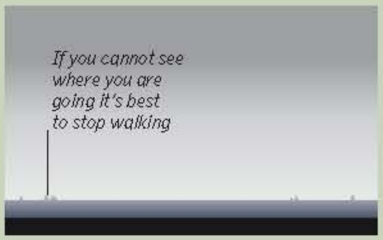

If you cannot see where you are going it's best to stop walking

FOG

Essentially a cloud in contact with the ground, fog reduces visibility, so beware of dangerous terrain, especially on mountains.

LIGHTNING

Lightning is an atmospheric discharge of electricity—it will strike the first object it encounters on its route to earth, so avoid high, exposed places.

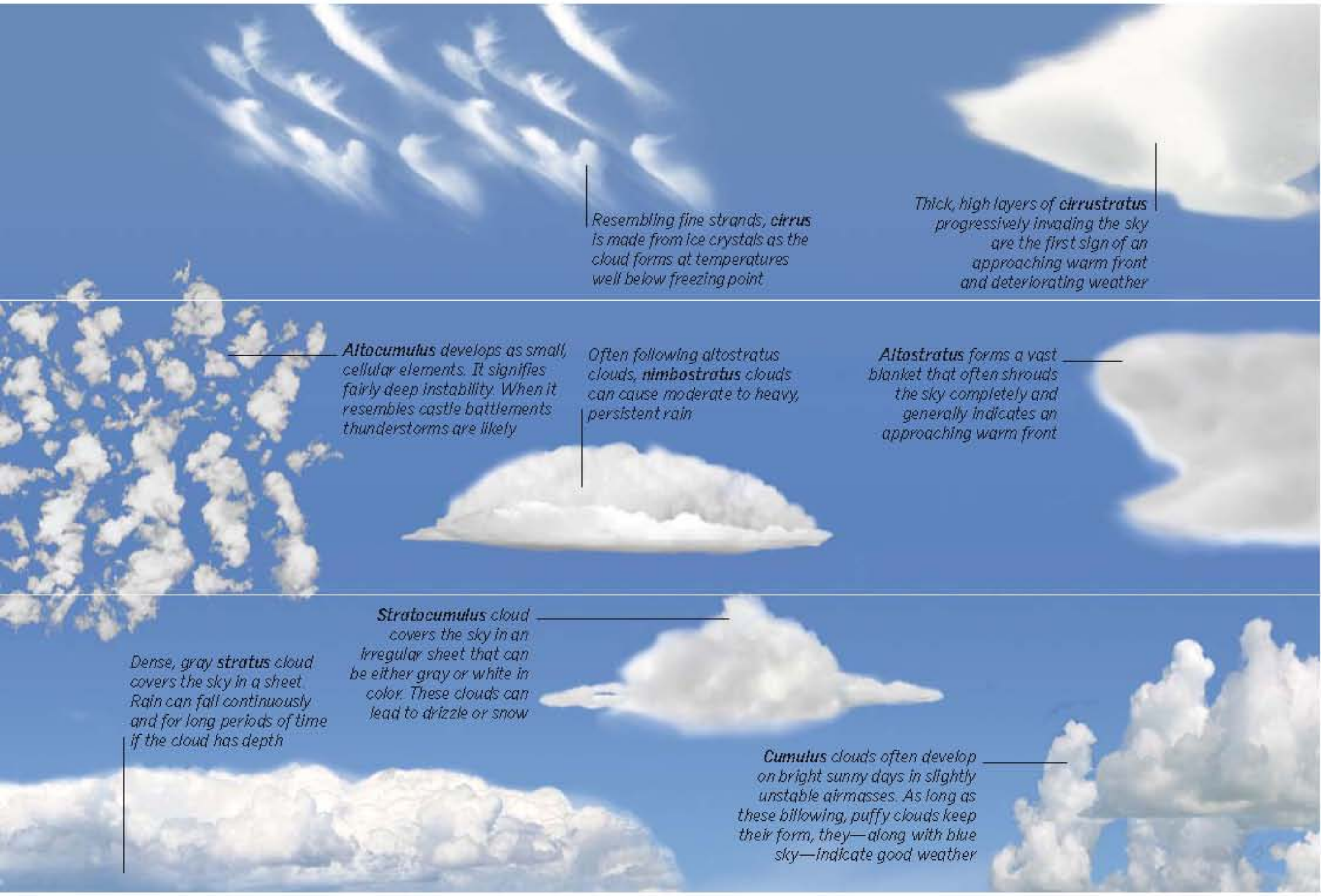

Resembling fine strands, ***cirrus*** *is made from ice crystals as the cloud forms at temperatures well below freezing point*

Thick, high layers of ***cirrustratus*** *progressively invading the sky are the first sign of an approaching warm front and deteriorating weather*

Altocumulus *develops as small, cellular elements. It signifies fairly deep instability. When it resembles castle battlements thunderstorms are likely*

Often following altostratus clouds, ***nimbostratus*** *clouds can cause moderate to heavy, persistent rain*

Altostratus *forms a vast blanket that often shrouds the sky completely and generally indicates an approaching warm front*

Dense, gray ***stratus*** *cloud covers the sky in a sheet. Rain can fall continuously and for long periods of time if the cloud has depth*

Stratocumulus *cloud covers the sky in an irregular sheet that can be either gray or white in color. These clouds can lead to drizzle or snow*

Cumulus *clouds often develop on bright sunny days in slightly unstable airmasses. As long as these billowing, puffy clouds keep their form, they—along with blue sky—indicate good weather*

UNDERSTANDING LOCAL WEATHER

WEATHER IS AFFECTED—and can be predicted—by the local geography surrounding you. High ground forces air to rise and cool, and the relative temperatures of land and sea produce predictable effects.

REGIONAL EFFECTS

It helps to understand how these predictable weather patterns occur in various geographic areas. They may influence your decisions when on the trail.

THE FOHN EFFECT

The leeward side of high ground is warmer and more sheltered than the windward side. Air rises as it travels over obstacles and, if the air contains moisture, the water vapor condenses to form cloud once cooled to a certain level. The air loses water at the summit, then descends and warms on the leeward side.

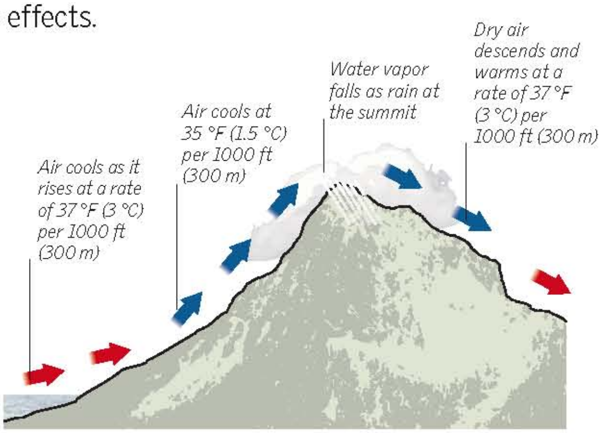
Air cools as it rises at a rate of 37 °F (3 °C) per 1000 ft (300 m)
Air cools at 35 °F (1.5 °C) per 1000 ft (300 m)
Water vapor falls as rain at the summit
Dry air descends and warms at a rate of 37 °F (3 °C) per 1000 ft (300 m)

ANABATIC AND KATABATIC WINDS

Winds that flow up and down slopes during certain atmospheric conditions are known as "anabatic" and "katabatic." They typically occur in mountainous regions.

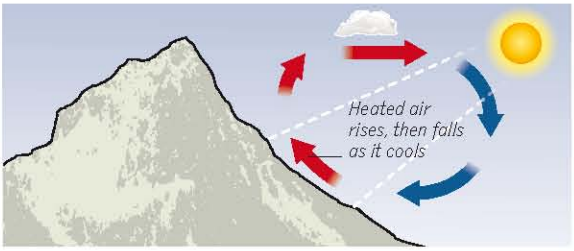
Heated air rises, then falls as it cools

ANABATIC WINDS

During the day, the surface of sloping terrain heats up. The air rises and creates a gentle upslope breeze. Anabatic winds are lighter than katabatic winds because they act against gravity.

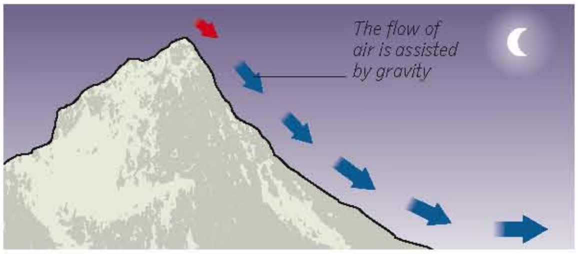
The flow of air is assisted by gravity

KATABATIC WINDS

Katabatic winds form on clear nights with a light breeze. Air in contact with the ground cools and its density increases, causing it to flow down the slopes of the hillside.

SEA AND LAND BREEZES

Sea breezes often develop along coastlines on clear sunny days, whereas land breezes tend to develop along coastlines on cloudless nights.

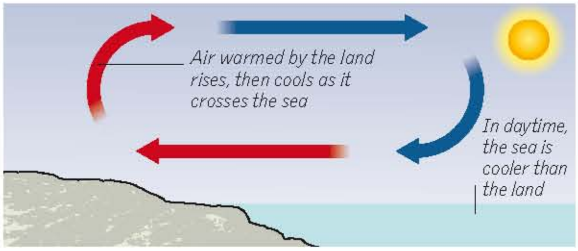
Air warmed by the land rises, then cools as it crosses the sea
In daytime, the sea is cooler than the land

SEA BREEZE

Warm air rises over land during the afternoon, drawing in cool air to replace it along the coast. The result of this coastal air circulation by day is a wind that blows from sea to land.

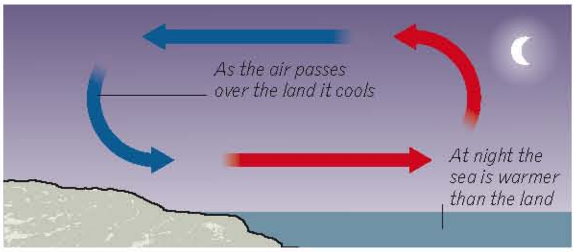
As the air passes over the land it cools
At night the sea is warmer than the land

LAND BREEZE

At night, the sea becomes comparatively warmer and air begins to rise. Air from the neighboring land is drawn toward the sea. The result is a wind that blows from land to sea.

NATURAL WEATHER FORECASTERS

These natural indicators are based on observation and can be useful if you have no alternative means of predicting the weather.

LOOKING AT THE SKY

- If the sky is red at dawn there is moisture in the air and a potential storm ahead. A red sky at night often indicates good weather to come.
- A rainbow usually indicates good weather is on the way, or a light shower.

PLANTS AND FLOWERS

- The scent of plants and flowers is often stronger before rain.
- Certain flowers, such as Scarlet Pimpernel and Morning Glory, are said to close up if bad weather is on the way.
- Pine cones are one of the best natural forecasters—their scales absorb moisture in the air and close up if wet weather is approaching, and unfurl in dry air.

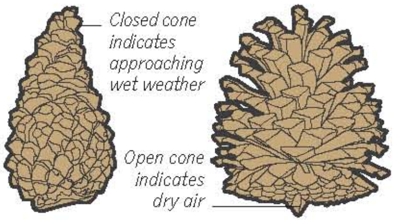
Closed cone indicates approaching wet weather
Open cone indicates dry air

WATCHING ANIMALS

- Animals migrating from high to low areas may indicate that a storm is on the way.
- Cows often lie down before it rains.
- Wool reacts to moisture in the air by swelling and straightening. It shrivels in dry weather conditions.
- Humans can also sense atmospheric change—some people suffer headaches before a thunderstorm.

CHANGING WEATHER

Watch out for alterations in the wind direction or strength as this may lead to a change in the weather. A dry, steady wind that changes direction or decreases in strength often indicates that rain is on its way.

THE BEAUFORT SCALE

Providing visual references for the effects of wind speed on land and at sea, the Beaufort Scale was designed by Sir Francis Beaufort in 1805. It provides a simple way of estimating wind speeds without the need for equipment, and is still in common usage today. The scale ranges from calm to hurricane, and is numbered from zero to 12.

BEAUFORT NUMBER	WIND DESCRIPTION	WIND SPEED MPH (KPH)	WIND EFFECT ON LAND AND AT SEA
0	Calm	0 (0)	Smoke rises vertically. Sea like a mirror.
1	Light air	1–2 (1–3)	Smoke drifts gently. Sea ripples like scales.
2	Light breeze	3–7 (4–11)	Leaves rustle. Wind felt on skin. Small wavelets.
3	Gentle breeze	8–12 (12–19)	Leaves and twigs move. Large wavelets with scattered whitecaps.
4	Moderate breeze	13–18 (20–29)	Small branches move. Small waves with frequent whitecaps.
5	Fresh wind	19–24 (30–39)	Small trees begin to sway. Moderate waves with some foam and spray.
6	Strong wind	25–31 (40–50)	Umbrella usage is difficult. Large waves of 10 ft (3 m) with some spray.
7	Near gale	32–38 (51–61)	Whole trees sway. Sea heaps up and foam begins to streak.
8	Gale	39–46 (62–74)	Walking is difficult. Moderately high waves of more than 18 ft (5 m).
9	Severe gale	47–54 (75–87)	Damage to roofs. High waves with toppling crests.
10	Storm	55–63 (88–101)	Trees uprooted or broken off. Very high waves and sea surface appears white.
11	Severe storm	64–74 (102–119)	Houses damaged. Exceptionally high waves of more than 38 ft (11 m).
12	Hurricane	More than 75 (More than 120)	Buildings destroyed. Huge waves of more than 46 ft (14 m).

MAKE A MOVE

While some survival situations are simply unavoidable, many are entered into because basic techniques relevant to a particular mode of transport have not been followed correctly. This could be due to a lack of knowledge, a loss of concentration, an individual's over-confident belief in his or her own equipment or skills, or as a result of sheer recklessness. Therefore, before you venture into an unfamiliar environment, make sure that you carefully research the type of terrain you will be encountering and investigate the best method of traveling safely and efficiently across it. Knowing the correct techniques for scrambling up a mountain or walking down a steep hillside, for example, could mean the difference between an enjoyable day out and a life-threatening survival situation arising from a fractured ankle in a remote location. Equally, being able to regain control of your vehicle after it has entered a skid could quite simply save your life. Thoroughly researching the terrain will also allow you to select the correct

In this section YOU WILL DISCOVER...

- how to make a **paddle** (so you're never **up a creek** without one)...
- the difference between your **finger shelf** and your **hand jam**...
- how to **pull** your own **pulk** and **scramble** over **scree**...
- when to **blaze a trail** or **ski without skis**...
- how to tell your **front-wheel skid** from your **skidoo**...
- when to let the **camel train** take the **strain**...
- how **chewing gum** can stop that **sinking feeling**...

equipment and most appropriate clothing, and enable you to familiarize yourself with the survival techniques relevant to a particular environment.

Whether you're planning to travel on foot, by boat, on horseback, in a 4WD vehicle, or by boat, you must consider not only your own capabilities and those of other members of your team, but also the capabilities of your equipment. Remember, pushing anything beyond its limits—whether that be a person, an animal, or a vehicle—will invariably lead to it failing.

A walking staff is one of the simplest yet most important survival aids you will ever need. It's the first piece of equipment to improvise should you find yourself in a survival situation.

THE "SURVIVOR'S THIRD LEG"—as a walking staff is also known—increases your ability to support yourself by allowing you to have two points of contact with the ground at any one time. This will reduce the chance of you slipping—a crucial factor given that your ability to walk is your main means of rescue; reduce your mobility and you seriously reduce your ability to survive.

A versatile tool, your survivor's third leg can be employed in many different survival situations. Among its many uses, it can be employed to:

- Support you as you walk
- Protect your face when you're walking through thickets or gorse
- Check for adequate support when crossing marshy ground
- Test the ground ahead for obstructions
- Check the depth of water when you're crossing streams and rivers
- Protect you against wild animals
- Form a ridgepole for your shelter
- Help you with your pace counting
- Spear fish or catch game
- Dig up roots or plants

> "Your ability to walk may be your **primary means of rescue**; reduce your mobility and you dramatically reduce your **chances of survival**"

TRAVELING ON FOOT

HIKING IS A GREAT WAY to explore wilderness terrain. It's important to have a decent level of fitness before you set out, and to wear and carry appropriate clothing and equipment. Hiking requires basic skills—using the correct techniques will help you move more efficiently, and ensure that your trip is safe.

BASIC WALKING SKILLS

Aim for a slow, even pace that can be maintained for the duration of the hike by all members of the group. A good way to maintain this pace is to develop a hiking rhythm. Take regular rest breaks and, if you're walking in a group, ensure everyone knows the route.

UPHILL TECHNIQUES

When traveling uphill, lean forward slightly, maintain your momentum but shorten your stride. When pushing upward, keep your feet flat on the ground.

- Carry a walking staff for support as you ascend.
- Move your legs forward from the hips.

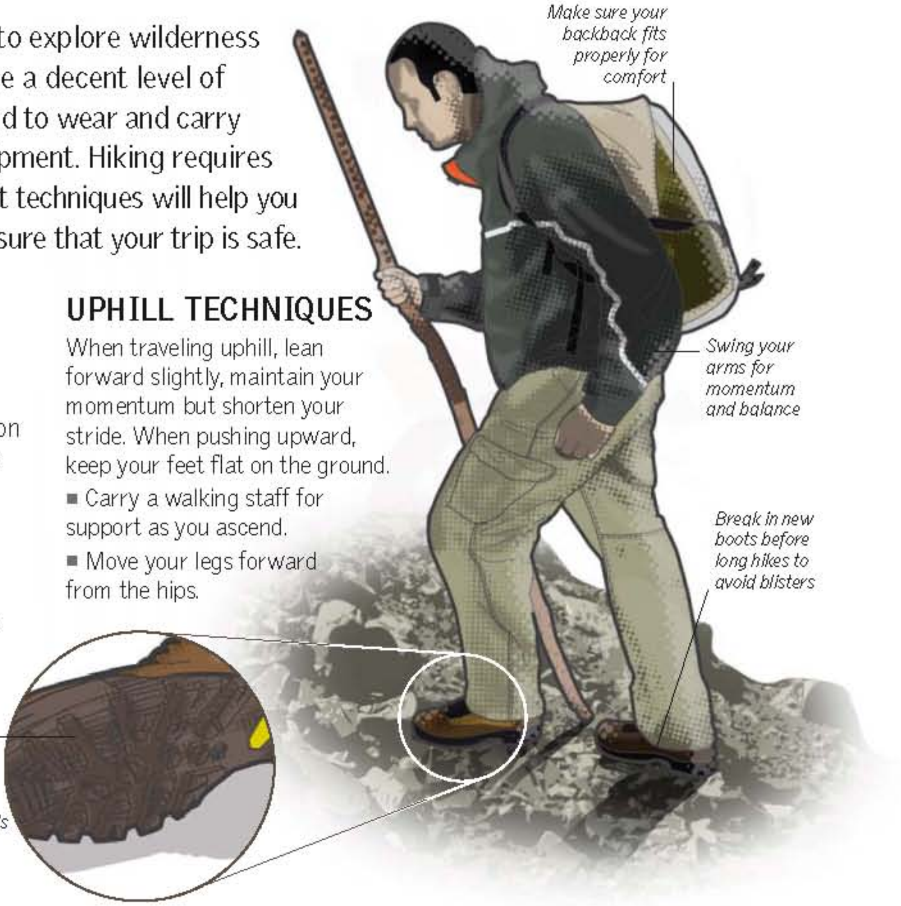

Make sure your backback fits properly for comfort

Swing your arms for momentum and balance

Break in new boots before long hikes to avoid blisters

Deep tread grips the ground and reduces slipping

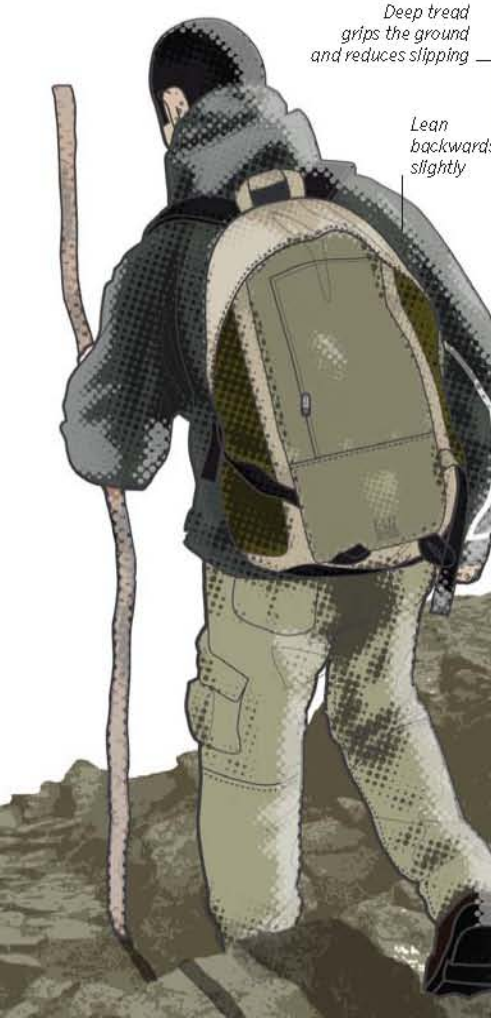

Lean backwards slightly

DOWNHILL TECHNIQUES

Walking downhill can place a lot of strain on your thighs, knees, and ankles, especially when you're carrying a heavy backpack. Be careful not to lose control or gather too much speed.

- Use your arms for balance.
- Maintain a steady rhythm.

WALKING AT NIGHT

Unless you're in the desert and it's cooler to move in darkness, avoid walking at night because of the added risks posed by navigational difficulties, and predators that hunt at night. If there's no alternative, try the following:

- Use your flashlight or improvise a torch by setting alight some birch bark or similar.
- If this isn't possible and you've time, close your eyes for 20 minutes to allow them to adjust to night vision.
- Use your walking stick or pole to feel in front of you for obstacles, tripping hazards, or sudden drops in the ground.
- Keep your pace slow and deliberate, and check your compass regularly.

NEGOTIATING DIFFICULT GROUND

One of the most challenging terrains to hike over is "scree," a mass of small rocks that slides underfoot making uphill and downhill travel difficult. The slippery nature of the surface can make progress slow, and care should be taken to avoid falling and injuring yourself. Hiking on scree is hard work, but using the correct techniques will help you to advance confidently, efficiently, and safely.

TRAVERSING SCREE

Choose a zig-zag route and look for a path that contains similar-sized rocks. Walk sideways across the slope, taking small steps and testing rocks for stability before placing your full weight on them.

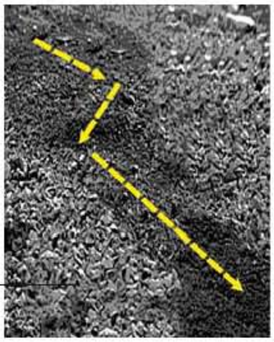

Rocks at the edge of the slope are likely to be larger and more stable

ASCENDING SCREE

Tread carefully, kicking your toes into the slope and testing the step before transferring your weight to that leg. Alternatively, walk with your feet splayed, placing your weight on the inside of each foot.

DESCENDING SCREE

"Screeing" downhill combines sliding with slow-motion jogging. When you have the technique it can be great fun, but avoid larger rocks to prevent ankle injuries.

TRAIL MARKERS

Your map and compass should be your primary method of navigation, but keep an eye out for trail markers, or "blazes," during your hike.

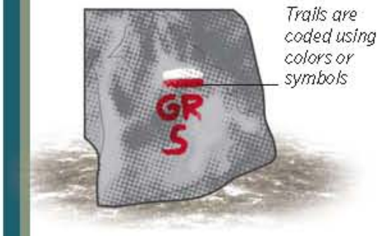

Trails are coded using colors or symbols

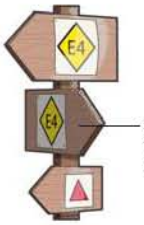

Arrows indicate direction

ROCKS

Trail blazes painted onto rocks and boulders are especially common in rocky terrain. They may be low on the ground so look carefully.

SIGNPOSTS

Made from wood, metal, or plastic, these signposts are particularly useful in areas where there are few rocks or trees on which to blaze.

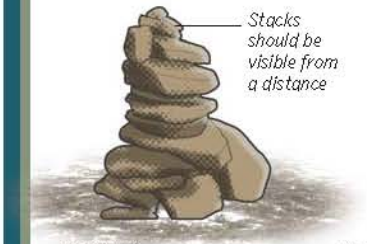

Stacks should be visible from a distance

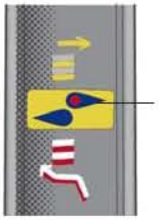

The same sign may indicate several trails

CAIRNS

Piles of rocks, known as "cairns," are designed to be visible in thick fog. Their size can vary from a few rocks to a large stack of boulders.

DIRECTIONAL SIGNS

When reading markers, look out for painted arrows or variations in the blazing, such as a bend. This indicates a change of direction in the trail.

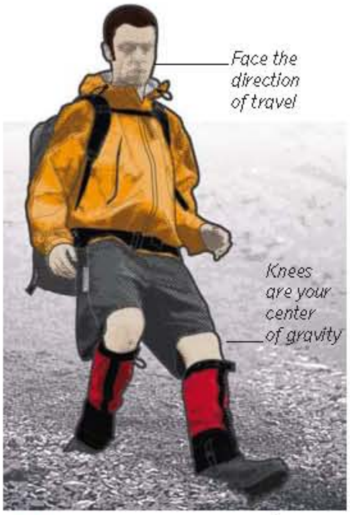

Face the direction of travel

Knees are your center of gravity

1 Ensuring the scree is fine and deep, use gentle hops to launch yourself down the slope. Keep balanced and let gravity do the hard work.

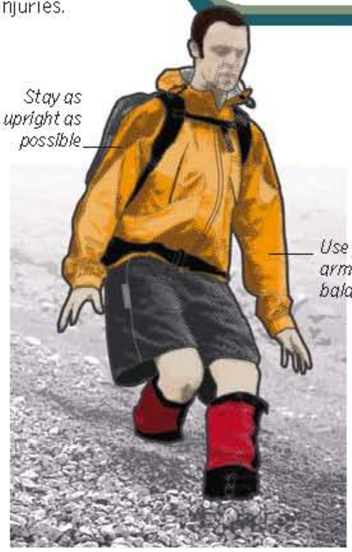

Stay as upright as possible

Use your arms for balance

2 As you build up momentum and rhythm, dig your heels into the slope and slide a short distance with each step, avoiding leaning forward.

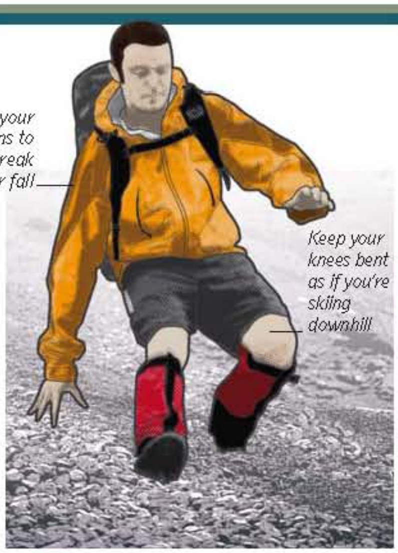

Use your arms to break your fall

Keep your knees bent as if you're skiing downhill

3 If you lose your balance temporarily, use your arms to steady yourself, relax your knees, and continue your descent.

CROSSING RIVERS

RIVER CROSSINGS ARE DANGEROUS and should be avoided unless absolutely necessary. In a survival situation, once you are cold and wet—which can lead to hypothermia—it is hard to get warm and dry. Always check your map for routes around the river, and choose the safest available crossing point.

CROSSING SAFELY

Before you get into the water, make sure you have a change of clothes or means of getting dry on the other side. In cold conditions, collect everything you need to get a fire going (tinder, kindling, and dry wood), and keep this dry as you cross.

CHOOSING YOUR CROSSING POINT

Walking upstream generally leads to shallower water, but be aware that even shallow water can have strong currents, as can water that looks calm on the surface. Always check for bridges further up- or downstream.

WADING BASICS

Wear footwear when wading to protect your feet from rocks and other dangers. Remove pants to keep them dry and decrease resistance in the water. Use a walking staff for added support.

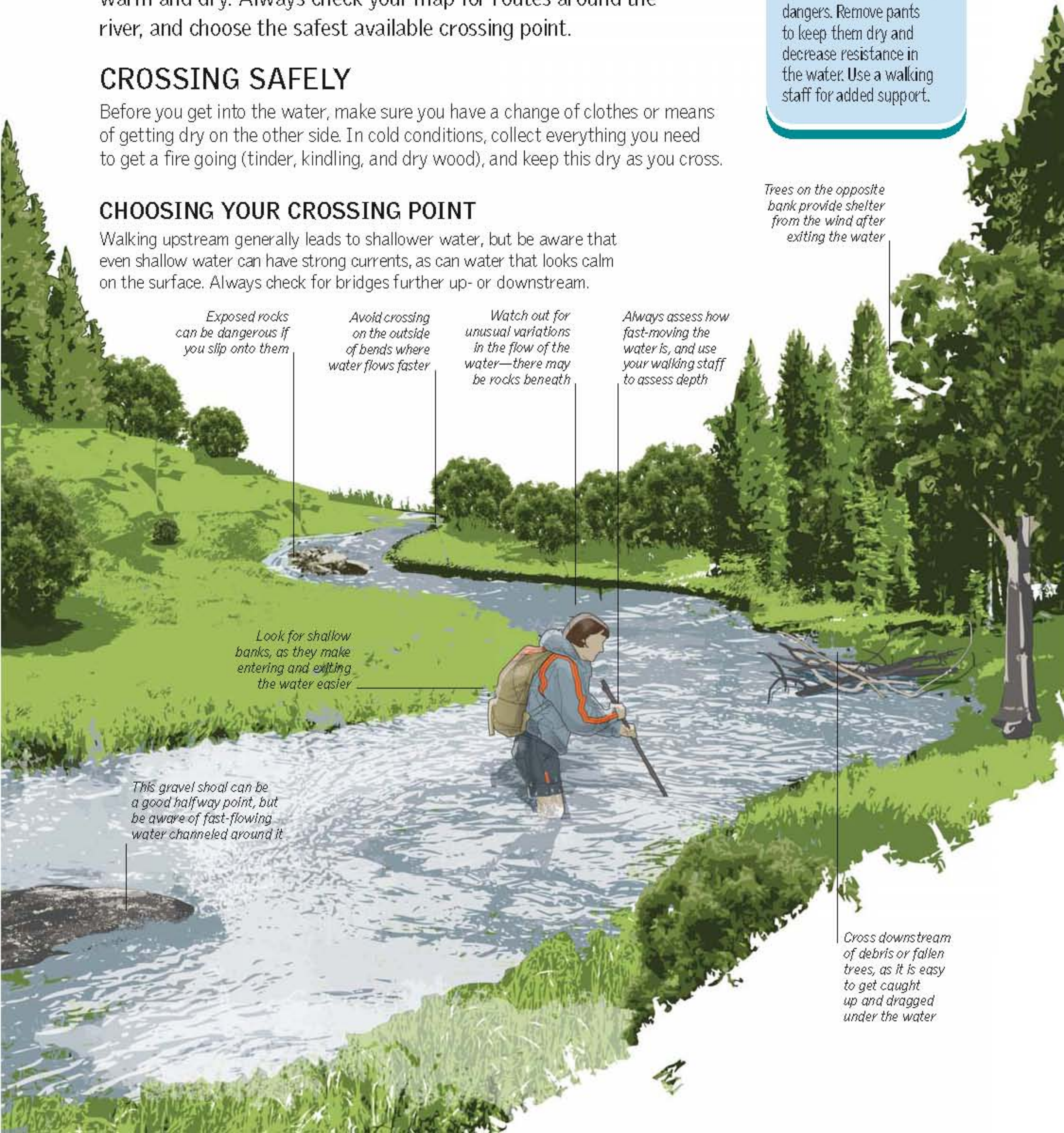

Trees on the opposite bank provide shelter from the wind after exiting the water

Exposed rocks can be dangerous if you slip onto them

Avoid crossing on the outside of bends where water flows faster

Watch out for unusual variations in the flow of the water—there may be rocks beneath

Always assess how fast-moving the water is, and use your walking staff to assess depth

Look for shallow banks, as they make entering and exiting the water easier

This gravel shoal can be a good halfway point, but be aware of fast-flowing water channeled around it

Cross downstream of debris or fallen trees, as it is easy to get caught up and dragged under the water

WADING WITH OTHERS

Crossing in a group is safer than crossing alone. Linking your arms together creates a stronger, more stable structure against the current and provides backup for anyone who falls. Loosen the straps of your backpack, and place only one arm through them, so you can release the pack quickly if you fall.

WARNING!
Never cross white water, or rivers that have flooded. In these situations it is safer not to cross at all than to risk injury.

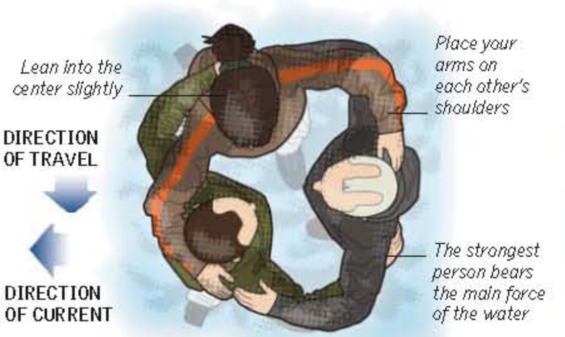
Lean into the center slightly
Place your arms on each other's shoulders
DIRECTION OF TRAVEL
DIRECTION OF CURRENT
The strongest person bears the main force of the water

CROSSING IN A HUDDLE

Positioning the strongest person upstream, with the others providing stability and support, link your arms tightly and take short, deliberate steps across the river.

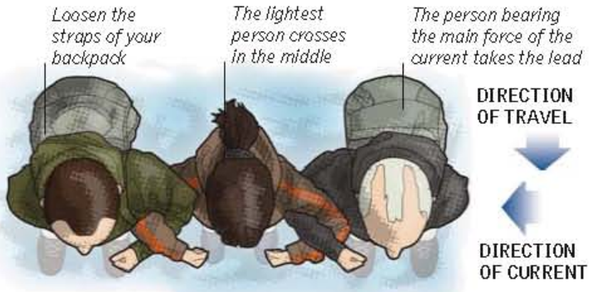
Loosen the straps of your backpack
The lightest person crosses in the middle
The person bearing the main force of the current takes the lead
DIRECTION OF TRAVEL
DIRECTION OF CURRENT

CROSSING IN A LINE

Keeping well balanced, cross the river perpendicular to the current. Move slowly and position each step carefully to avoid being swept away by the force of the water.

WADING ALONE

Wading across a river alone is not ideal, but if you have no option then your walking staff or "survivor's third leg" (see p. 72)—or simply any long stout stick—will provide additional support and balance. Use it as a probe to alert you to any sudden changes of depth in the river bed.

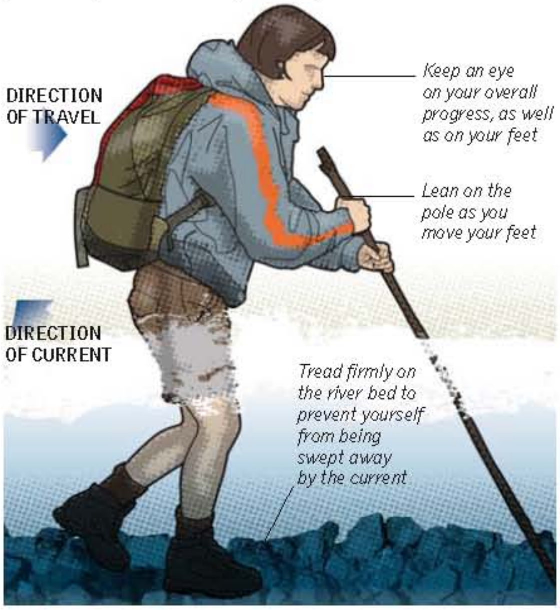
DIRECTION OF TRAVEL
Keep an eye on your overall progress, as well as on your feet
Lean on the pole as you move your feet
DIRECTION OF CURRENT
Tread firmly on the river bed to prevent yourself from being swept away by the current

USING A WALKING STAFF

Facing the current, walk diagonally across the river, placing each foot down firmly and deliberately, and leaning on your walking staff for support. Your feet and your staff are your three points of contact—keep at least two of them in contact with the river bed at all times.

CROSSING WITH ROPES

Using ropes is a good option for dangerous crossings, but they can get tangled, and drag you under the water. Always aim for the easiest and safest option—using ropes is complicated and should be your last resort.

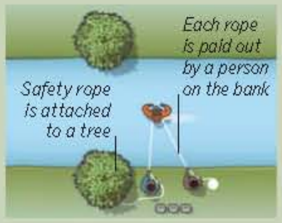
Safety rope is attached to a tree
Each rope is paid out by a person on the bank

1 A "safety" rope is tied to a tree and then carried across by the strongest person, who also takes a "crossing" rope with a carabina tied centrally.

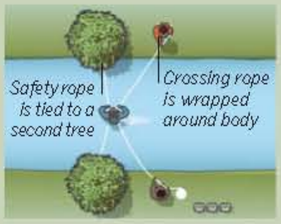
Safety rope is tied to a second tree
Crossing rope is wrapped around body

2 The safety rope is tied to a tree and the carabina clipped to the safety rope. The carabina is pulled back and fixed to the next person to cross.

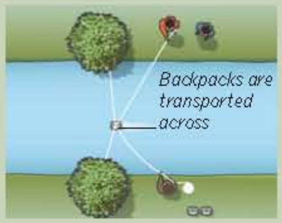
Backpacks are transported across

3 He or she crosses upstream of the safety rope, holding it for support. Backpacks can be clipped to the carabina and transported, too.

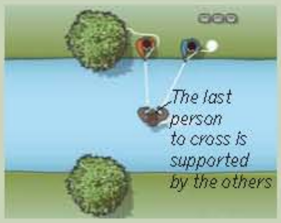
The last person to cross is supported by the others

4 The final person unties the safety rope from the tree, wraps it around his or her body, and wades across, supported by the others.

SCRAMBLING AND CLIMBING

CLIMBING WITHOUT ROPES is known as "scrambling." As with roped techniques, the idea is to maintain three points of contact with the rock at all times—either two feet and one hand or both hands and one foot. Your legs should power the climb, with your arms used predominantly for balance. Proceed carefully, making sure you are always balanced and confident of your next move.

WARNING!
Climbing is a dangerous activity that should only be considered as a last resort. Ideally, plan ahead using your map to find a way around the obstacle.

HAND AND FOOT HOLDS

When planning your route, choose your hand and foot holds carefully. Don't reach too far, and test each hold for stability before using it to support your weight.

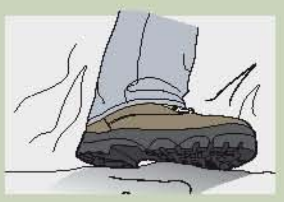

EDGES
For very small footholds, place the inside edge of your foot in the hold in the rock to take the pressure off your toes.

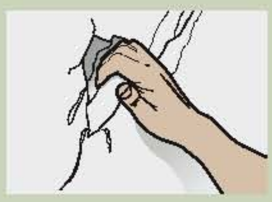

FINGER SHELVES
Curl your fingers over the rock for a secure fingerhold. The larger the fingerhold, the more secure it will be.

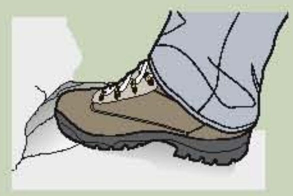

LARGE POCKETS
Placing your foot securely within the pocket, balance yourself so both your arms and legs take the strain.

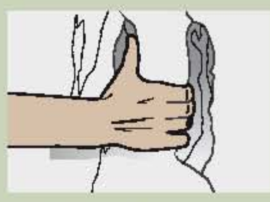

SIDE PULLS
Side pulls can be used to maintain balance or to pull yourself across the rock. Grip the rock tightly.

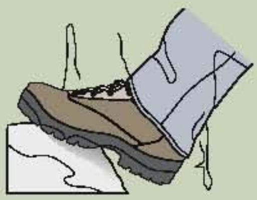

PROTRUSIONS
If possible, place the entire sole of your foot on the protrusion. If only the ball of your foot will fit, keep your heel low.

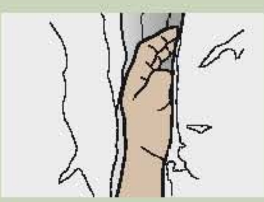

HAND JAMS
Insert your hand into the crack, push your thumb into your palm, arch your hand, and wedge it tightly inside.

BASIC CLIMBING TECHNIQUES

Always climb within your capabilities. It's important not to take any risks, as it's far safer to descend and start the climb again than to chance a fall. Before you start to climb, plan the easiest and safest route.

SCRAMBLING

When moving across the rock you will need to combine a variety of techniques to negotiate different obstacles. Always plan your moves several steps in advance, and keep three points of contact with the rock at all times.

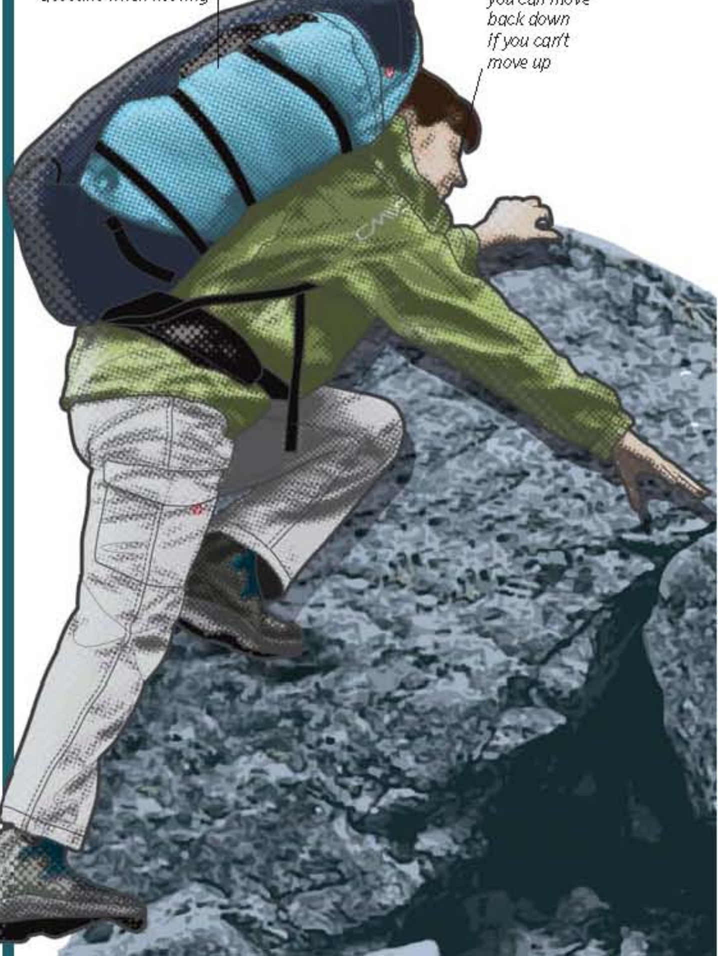

Take the weight of your backpack into account when moving

Always ensure you can move back down if you can't move up

MANTELLING

The mantelling technique is used to climb overhangs in the rock. Use your lead ankle and then your knee to lever your body up over the obstacle. It's a physically challenging technique, but useful.

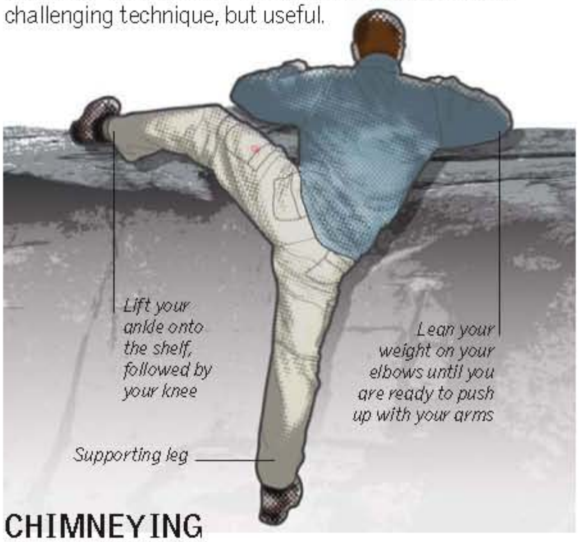
Lift your ankle onto the shelf, followed by your knee
Lean your weight on your elbows until you are ready to push up with your arms
Supporting leg

CHIMNEYING

To climb up the inside of large rock clefts, or "chimneys," use the chimneying technique. To ascend, move your back and hands up the wall, and push and follow with your legs. It's easy to get stuck at the top so plan your exit route carefully.

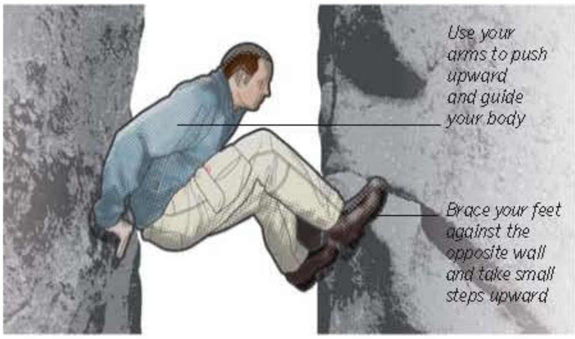
Use your arms to push upward and guide your body
Brace your feet against the opposite wall and take small steps upward

STRADDLING

If the chimney is relatively wide, you may need to alter your body position to straddle the gap. With a leg and arm on each wall, inch yourself up the rock using your legs to push your body upward.

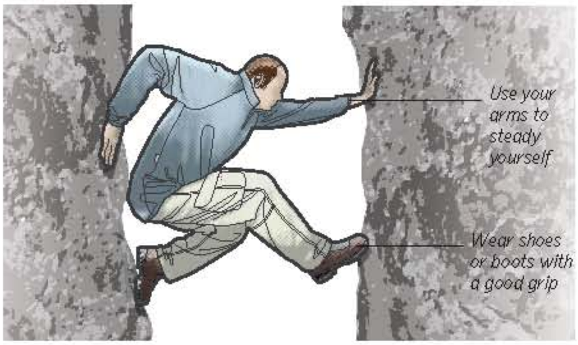
Use your arms to steady yourself
Wear shoes or boots with a good grip

CLIMBING WITH EQUIPMENT

The advantages of climbing with ropes cannot be underestimated as far as safety is concerned, although a lot of specialized equipment is required. As well as ropes, helmets, and harnesses, climbers can also use screws to secure themselves to the rock as they progress each stage.

BELAYING

Belaying is a technique used by climbing partners to safeguard each other during a climb. The belayer secures himself to the rock and stands on a suitable ledge while the lead climber ascends. Once at the top, the lead climber secures himself and acts as the belayer.

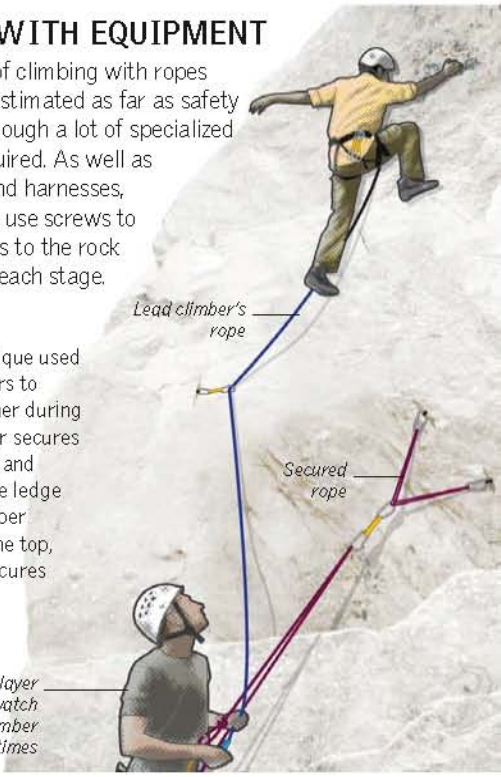
Lead climber's rope
Secured rope
The belayer must watch the climber at all times

ICE CLIMBING

Ice climbers utilize similar techniques to rock climbers, although they carry ice axes and wear crampons on their feet, to help them grip the ice.

MOVING UPWARD

As with rock climbing and scrambling, ice climbers ascend using power from their legs and arms. Crampons and ice axes are used to create holds in the rock.

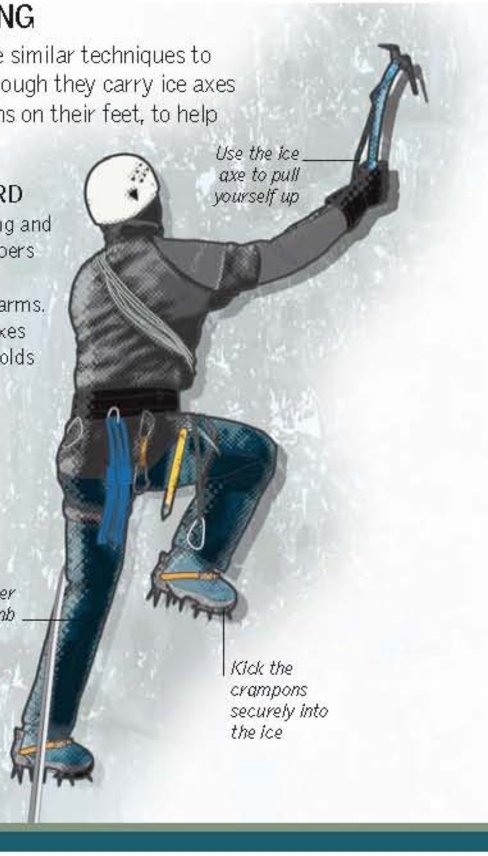
Use the ice axe to pull yourself up
Legs power the climb
Kick the crampons securely into the ice

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE MOUNTAINS

USEFUL EQUIPMENT

- Two-way radio
- Avalanche beacon and probe
- Waterproof clothing
- Spare clothing layers
- Collapsible shovel
- Climbing equipment, if needed
- Signal flares
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

STEVEN GREEN, 32, AN EXPERIENCED HILL WALKER from Dumfries, Scotland, spent four nights lying on a remote mountainside after a life-threatening fall. Suffering from a broken jaw and fractured skull, he survived via a combination of good preparation, quick thinking, and improvisation.

On Thursday, October 7, 1999, Steven set out alone for a trek in the hills of Wester Ross in the northwest Highlands of Scotland. He was well equipped with, among other things, a cell phone, bivy sack, pita bread, and water—and had taken the wise precaution (or so he thought) of informing his girlfriend of his intended route.

"HE SLIPPED ON WET GRASS AND FELL DOWN A WATERFALL"

Following an all-day trek and an overnight camp, he was descending the mountain on Friday when he slipped on wet grass and fell down a waterfall, coming to rest in the shallow, rocky riverbed. He knew he would not survive for long in the cold water so, despite being in intense pain, he crawled to the relative safety of the bank, then wisely got into his bivy sack for warmth and waited. He sustained himself by eating pita bread, which he softened in water because he couldn't chew. His phone was damaged in the fall so he was unable to call for help. When he failed to return home on time, the alarm was raised by Steven's girlfriend, but she had forgotten to write down his intended route so no one knew his exact location.

Steven experienced a further four days and nights of exposure while local and RAF rescue teams and search-and-rescue dogs combed the mountains for him. As the days passed they feared the worst, but when they located Steven on Tuesday morning—after finding his car and, inside it, his route map—they found him in good spirits. At hospital Steven was diagnosed with a fractured skull, broken jaw, missing front teeth, cuts and bruises, and a leg injury—but he was alive.

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Elements—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

YOU SHOULD STAY **

DO

- Keep all clothing dry and clean
- Collect drinking water from fast-moving streams; filter and disinfect if you have the means
- Be on constant lookout for signs of cold-related injuries such as frostnip, frostbite, and hypothermia
- Regulate your clothing to avoid overheating when moving and hypothermia when static
- Use a walking staff to aid safe movement
- Watch the weather closely and be prepared to change your plans at short notice—mountain climates are highly unpredictable
- Improvise shelter when not on the move

DON'T

- Descend hills in a careless manner; zig-zagging across hills is less tiring and puts less strain on leg muscles
- Travel at too fast a pace—high altitude equals less air and this will place greater demands on even a fit person
- Sweat too much, as the moisture will chill you further
- Ignore any opportunities to collect dry tinder/kindling

DON'T

- Allow your extremities to get too cold as those areas are most susceptible to frostbite
- Overlook the dangers of carbon monoxide poisoning in cramped shelter. Don't let candles, stoves, or fires burn all night
- Breathe on your hands to warm them up: you are exhaling warm air that you will have to replace with cold

DO

- Select a shelter site that offers protection from the elements. Build it no bigger than it needs to be
- Fill plastic bags or spare clothing with dry foliage and use as a mattress or pillow to insulate you from the cold, damp ground
- Deploy all your aids to location and prepare for immediate use
- Check upstream for the quality of your water source
- Light a fire and (if in a group) take turns tending it to keep it going all night
- Continually re-assess your situation and adapt your actions as necessary.
- Be constantly alert for signs of rescue

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.
** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

MOVING OVER SNOW

PREPARATION IS ESSENTIAL when hiking over frozen terrain. Not only must you be physically fit—the conditions make for slow, exhausting progress—but you must also have the right equipment and know how to use it. Wearing snowshoes or skis, and breathable and layered clothing to regulate your temperature, is essential. With the correct techniques you can hike safely and enjoy the surroundings.

WARNING!

Walking over deep snow without snowshoes is called "post-holing" and should be avoided unless absolutely necessary. Sinking into the snow will leave you exhausted and very wet, which in cold conditions can quickly lead to hypothermia. The exertion can also cause sweating and a dangerous reduction in your body temperature when you stop.

SNOW AND ICE

Knowing how to deal with different types of snow and ice is invaluable when hiking over frozen terrain.

DEEP SNOW

- If in a group, walk in single file and take turns at the front: the most strenuous position, as you're creating the path.
- Avoid rocks—in spring, they absorb heat and the snow above becomes unstable.

FROZEN CRUST

- Use a walking staff to test the snow ahead. An ice crust above deep snow may take your weight, but progress carefully.
- Later in the day you may come across melted depressions, called "sun cups." Cross on the rims to avoid sinking into the snow.

SLOPES

- Kick firmly into the slope and test your weight before ascending. When descending, you can use a technique called "boot skiing" (skiing without your skis).
- Choose your route based on the conditions: travel in straight lines if you are able to, or in zig-zags if the terrain is steep.

ICE

- Always progress carefully, using your walking staff to test the ice, especially over rivers and lakes. If in a group, rope yourselves together for safety.
- Wear crampons for extra grip. On very steep slopes, use your ice ax to cut steps.
- Use your ice ax to halt a fall by turning to face the slope and digging it into the ice.

GLACIERS

- Never attempt to cross glaciers without a guide—glaciers require specialized skills.

WAYS OF TRAVELING

When traveling over snow, your main aim is to get to your destination as safely as possible, without expending too much energy or losing too much body heat. Methods range from improvised snowshoes to motor vehicles.

USING SNOWSHOES AND SKIS

Snowshoes and skis are an effective way of traveling over snow. They work by spreading your body weight over a larger surface area, which enables you to walk on the surface of the snow rather than sink into it. Always set off cold, as you'll warm up quickly, and add or remove layers as required.

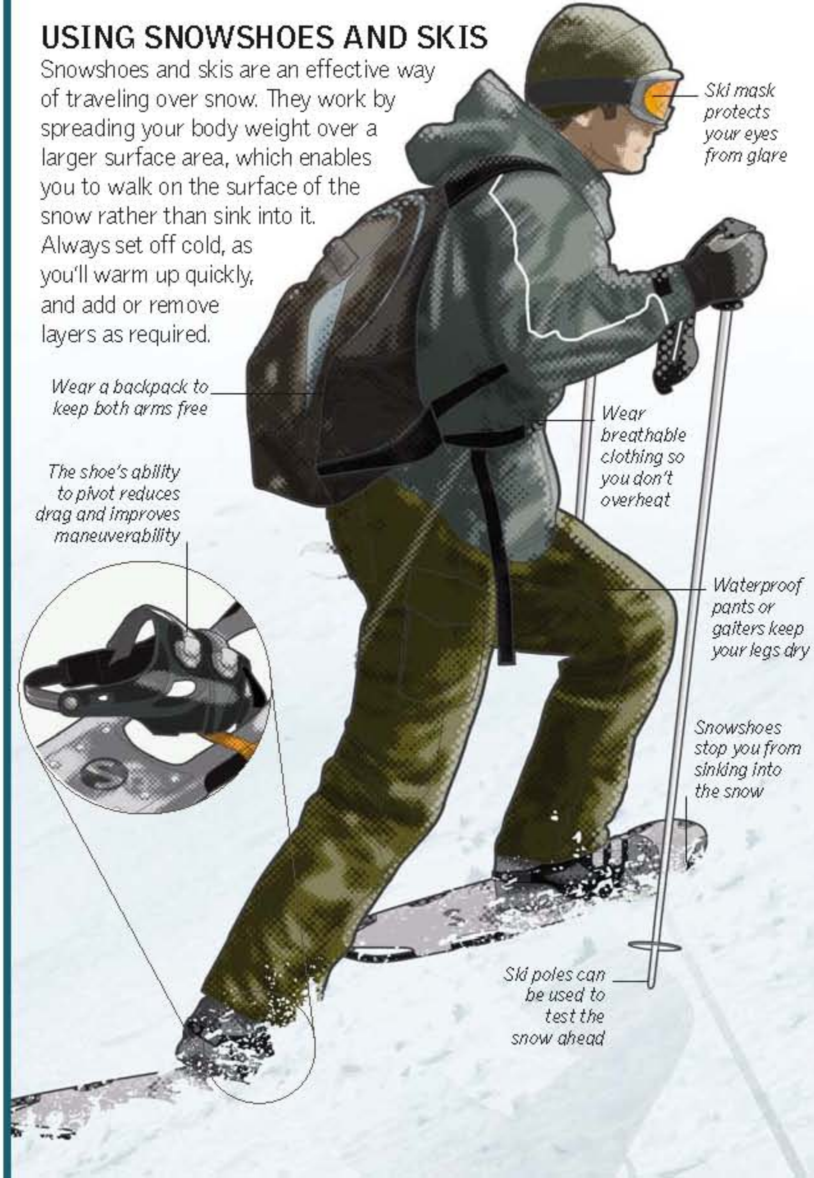

Ski mask protects your eyes from glare

Wear a backpack to keep both arms free

Wear breathable clothing so you don't overheat

The shoe's ability to pivot reduces drag and improves maneuverability

Waterproof pants or gaiters keep your legs dry

Snowshoes stop you from sinking into the snow

Ski poles can be used to test the snow ahead

MAKING IMPROVISED SNOWSHOES

If you don't have any pre-manufactured snowshoes—for example, in a survival situation—you can make some very simple shoes to help you negotiate the snow more effectively, using your knife, some green wood, and cordage. They may take time to build, but will save time and effort in the long run.

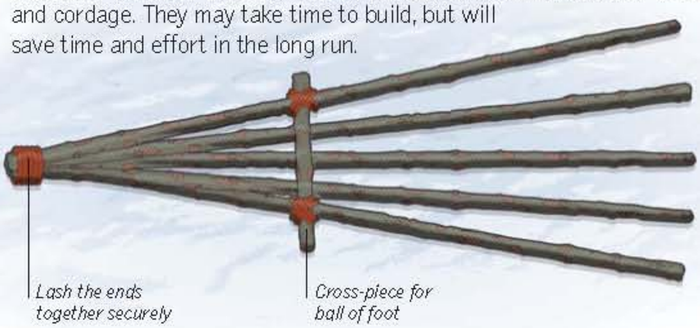
Lash the ends together securely
Cross-piece for ball of foot

1 Cut five lengths of green wood. They should be as thick as your thumb and the same length as the distance from your foot to your armpit. Cut three shorter lengths for the cross-pieces.

- Lash the ends of the five longer pieces together securely using cordage.
- Calculate where the ball of your foot will be positioned on the shoe and lash a cross-piece across the five lengths. Ensure the shoe will balance.

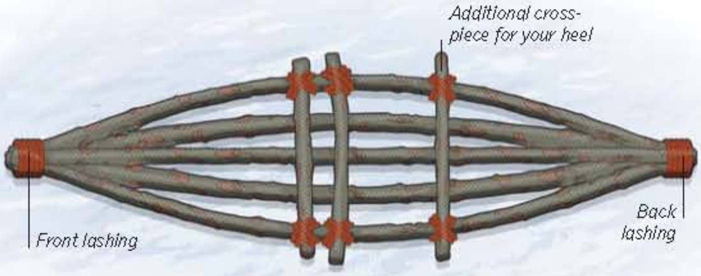
Additional cross-piece for your heel
Front lashing
Back lashing

2 Lash the five loose lengths of wood together at the back of the shoe. It's important to make this lashing as secure as possible.

- Fix the second cross-piece roughly 2 in (5 cm) behind the first.
- Lash the third cross-piece where your heel will rest.
- Repeat steps 1 and 2 to make a second shoe before progressing to step 3.

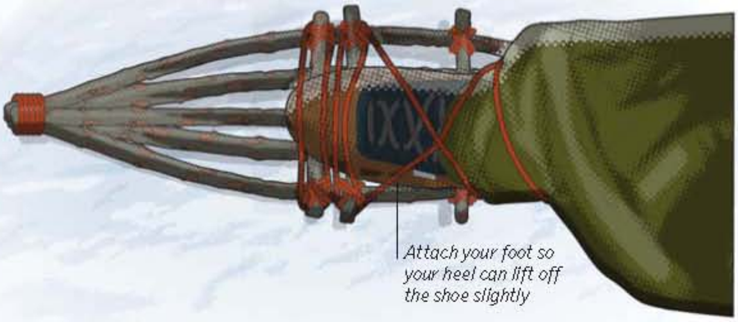
Attach your foot so your heel can lift off the shoe slightly

3 Place your foot on top of the snow shoe, ensuring that the ball of your foot sits directly over the front cross-piece and that your heel is positioned on the back cross-piece.

- Tie your boot to the snow shoe using whatever cordage you have, but ensure the heel is allowed to remain free to pivot. Repeat for the other foot.

ALTERNATIVE METHODS

These alternative methods will also spread your weight on the snow. If you have no gaiters, tie plastic bags around the bottom of your legs to keep them dry.

USING BRANCHES

Ideal for short distances, a simple way to get you out of deep snow and to a road or track is to attach branches to your feet using cordage. Select a tree, such as pine, that has strong, close branches.

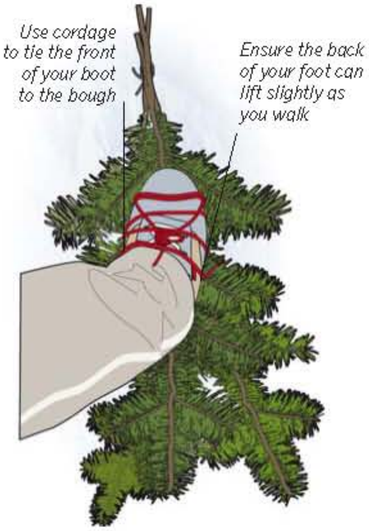
Use cordage to tie the front of your boot to the bough
Ensure the back of your foot can lift slightly as you walk

USING SAPLINGS

You will need branches that have some flexibility. Gently bend the longest branch into a teardrop shape and lash the two ends together. Heating the saplings over a fire will make them easier to bend.

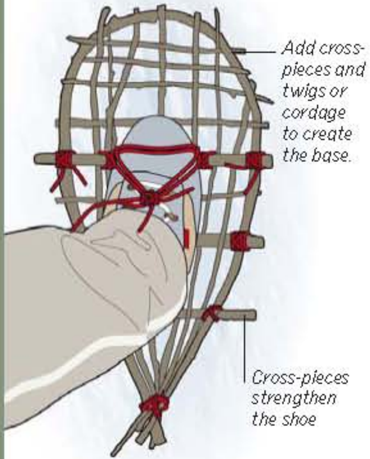
Add cross-pieces and twigs or cordage to create the base.
Cross-pieces strengthen the shoe

MAKING A SLED

If you find yourself in a survival situation with heavy equipment to carry over snow, making a simple sled or "pulk" (see panel, right) will help you transport it more efficiently. You could also build one to transport an injured person or small child. You can make your sled as large or as small as you need, providing you follow these basic principles.

TOOLS AND MATERIALS

- Survival saw or pocket chainsaw
- Knife
- Cordage
- Forked branch (for bulk of sled)
- Two branches (for bracing pieces)
- Sticks (for cross-pieces)

1 Using a short saw or penknife, cut a forked branch to form the base of the sled.

- Tie the two ends of the fork to the main branch with cord. This creates tension and forces the branches to curve and act as runners.

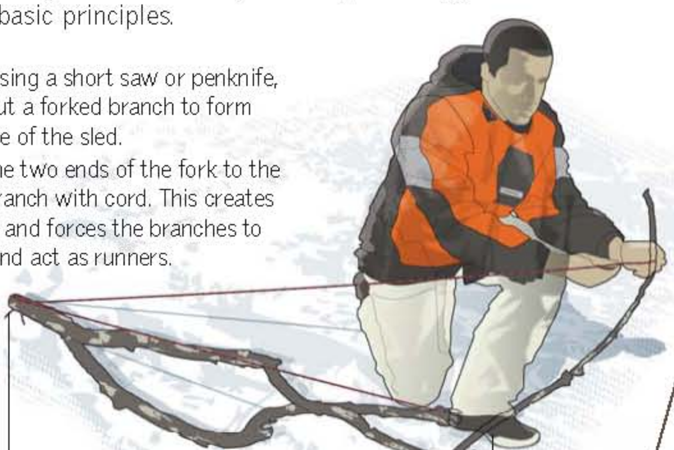

Attach the cord to the main branch using a slip knot (see p. 142)

Attach the cord to the end of the fork using a taut line hitch (see p. 145)

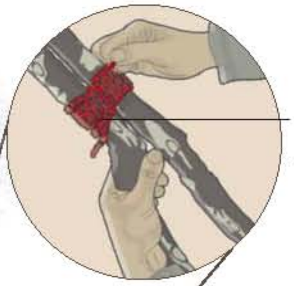

Lash the bracing piece to the main branch

2 Lash an additional branch to one of the runners to create a bracing piece.

- Repeat on the other side.
- The bracing pieces strengthen the branches and prevent any cord from touching the ground—if it does, it will wear out quickly and disintegrate once the sled is on the move.

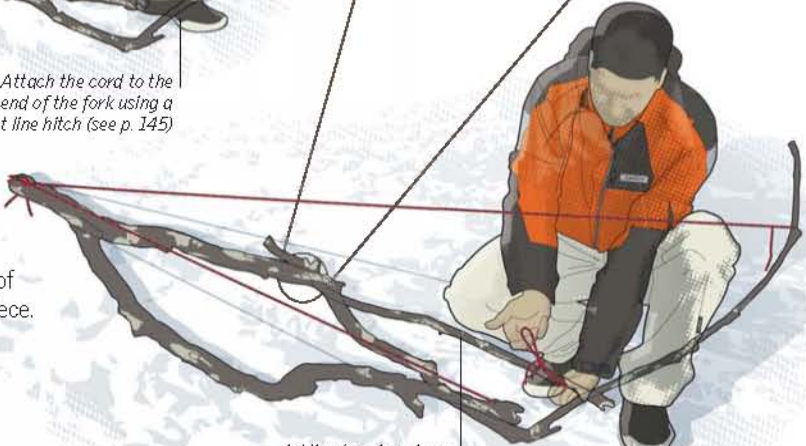

Adding bracing pieces creates a support for the carrying platform

ANIMAL POWER

Huskies, or sled dogs, are ideal for transporting people and equipment across winter terrain. Because of their thick coat they can endure very low temperatures, and their large, furred feet allow them to move quickly over snow. Huskies work well as a team and can pull heavy loads easily and efficiently.

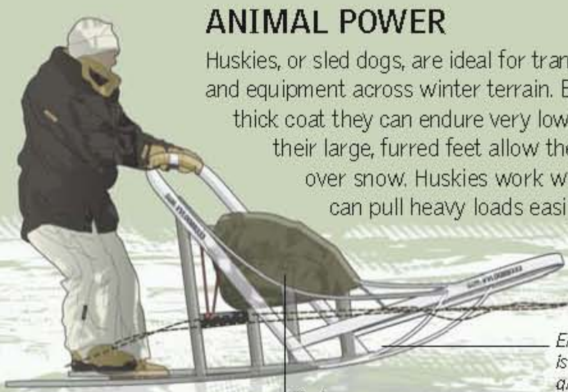

Ensure the sled is well balanced and not overloaded

Tie luggage securely to the frame

TRAVELING WITH DOGS

Although huskies are easy to care for and relatively simple to handle, you should not undertake an expedition with dogs unless you are traveling with an expert or have first undergone extensive training. The same goes for all expeditions in snowy terrain.

HUMAN POWER

Pulling your equipment behind you using a pulk is an extremely efficient way of moving over snow. Pulks are small, low-slung toboggans, typically made from lightweight plastic, and come in many different sizes.

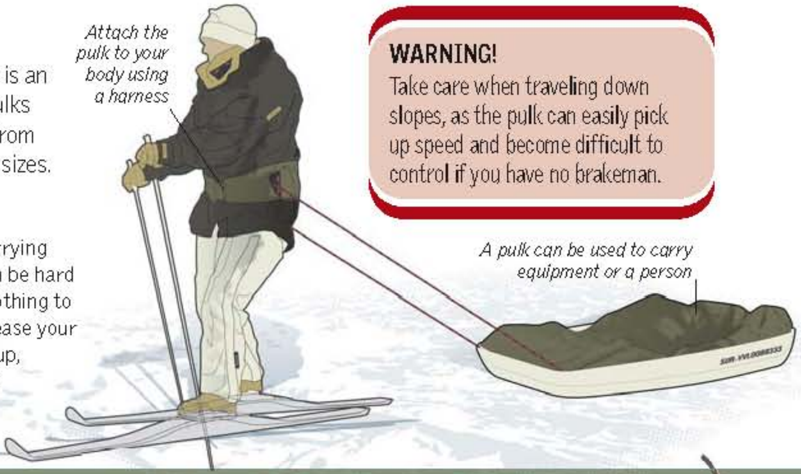
Attach the pulk to your body using a harness
A pulk can be used to carry equipment or a person

WARNING!
Take care when traveling down slopes, as the pulk can easily pick up speed and become difficult to control if you have no brakeman.

TRAVELING WITH A PULK

Although pulks are the most efficient method of carrying heavy loads over snow using human power, they can be hard work, especially in softer snow. Wear breathable clothing to allow excess body heat to escape. Know how to release your harness quickly in an emergency. If you are in a group, one person can be harnessed to the rear of the pulk to act as a brakeman when going downhill.

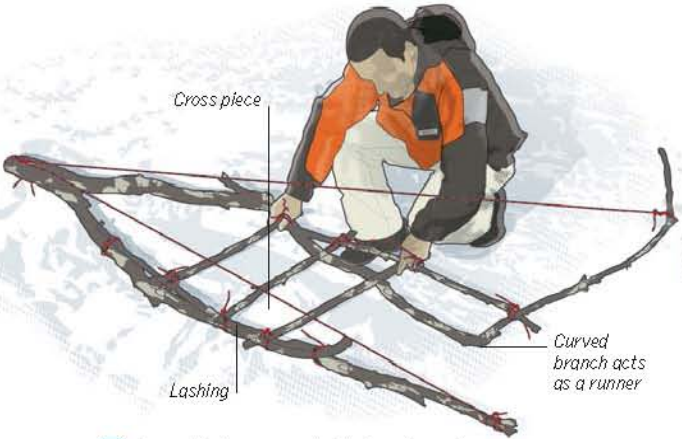
Cross piece
Curved branch acts as a runner
Lashing

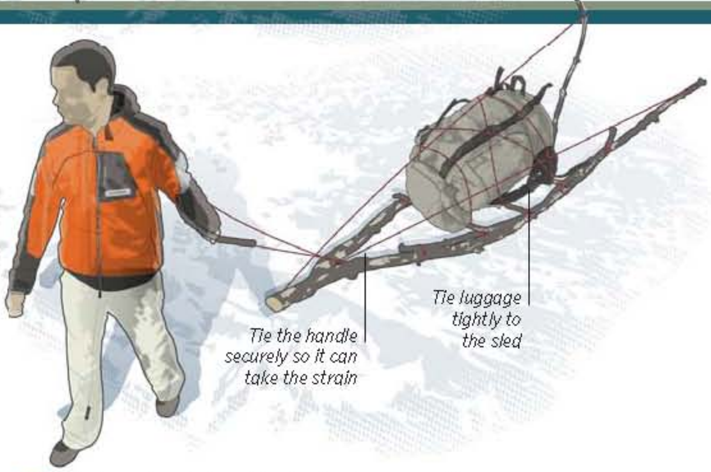
Tie luggage tightly to the sled
Tie the handle securely so it can take the strain

3 Lay sticks across both bracing pieces and lash them together. These cross-pieces strengthen the structure.

- You can use as many sticks as you can find, but three or four should suffice.
- These cross-pieces will also form the main carrying platform of your sled.

4 Make a handle for the sled using a small branch. The branch should be as smooth as possible to make pulling the sled more comfortable.

- Attach the handle to the front of the sled using a long piece of cord.

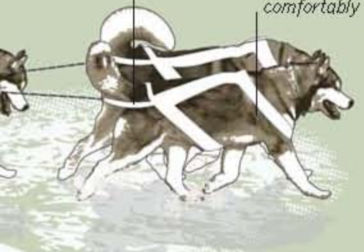
Linking the dogs together helps them to power forward in unison
Harnesses allow the dogs to pull comfortably

MOTOR POWER

Snowmobiles are practical and quick, and have revolutionized arctic travel. They are easier to handle than huskies, and can reach relatively high speeds. However, although you can cover great distances in short periods of time, if anything goes wrong you may be far from help. Always carry survival equipment when you travel.

BEFORE YOU GO

Prior to a snowmobile trip, regardless of distance, it is vital to check your vehicle maintenance and ensure that you have sufficient fuel for the trip. Plan your route carefully and inform someone of your plans so that they can raise the alarm if you don't reach your destination when planned. Dress warmly, wear goggles or a ski mask to reduce glare, and drive safely.

USING PACK ANIMALS

PACK ANIMALS CAN BE USED to transport heavy loads when hiking over wilderness terrain. Ideally suited to carrying substantial loads, they can cover great distances in areas where vehicles would struggle. When you are organized, and know how to care for the animals, traveling with pack animals can be very rewarding.

PACKING TIPS

- Know the capabilities of your animal, and what weight they can safely carry over the necessary distance.
- Before loading the animal, groom it well and check for sores or ticks.
- Place heavier items at the bottom of your panniers and lighter items at the top.

LOADING THE ANIMAL

Stand on level ground, making sure that the animal has all four legs placed firmly on the ground. Before placing the loaded panniers onto the animal, lift them to ensure they are well balanced. Rearrange if one is lighter. You may need to hobble or blindfold the animal to keep it still while loading.

ANIMAL WELFARE

Remember to transport food, water, and supplies for the animal as well as for yourself. Their welfare is paramount, so treat them well.

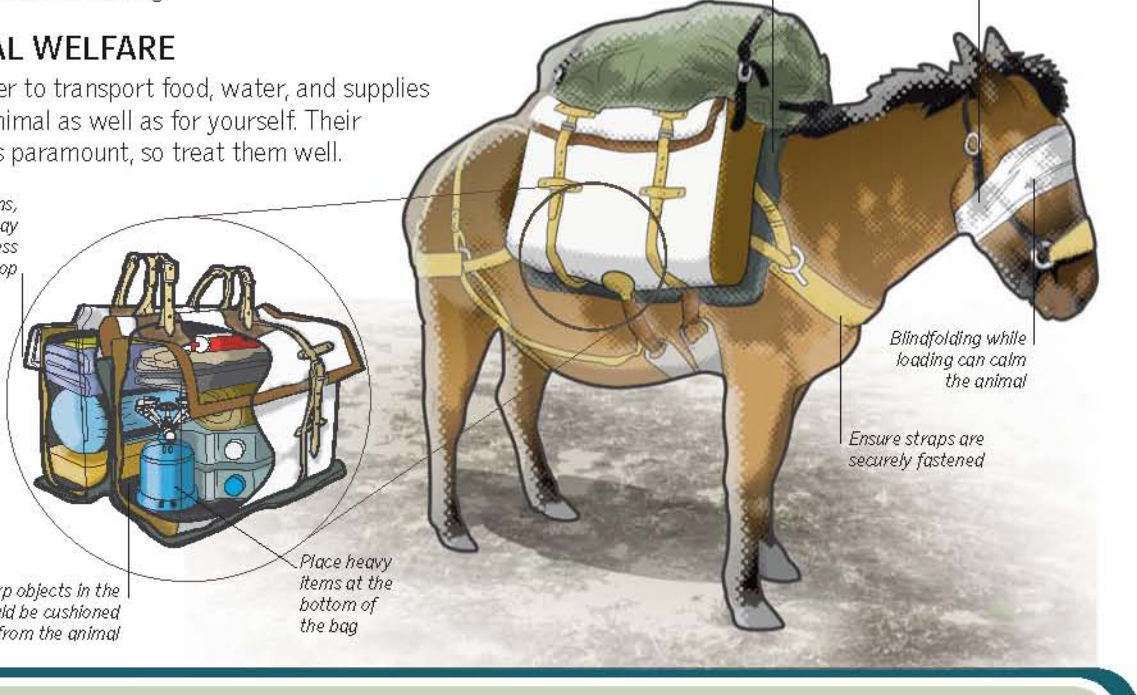
Fit a blanket for the animal's comfort
A headcollar gives you control over the animal
Place lighter items, and items you may need to access quickly, at the top
Blindfolding while loading can calm the animal
Ensure straps are securely fastened
Any sharp objects in the bag should be cushioned from the animal
Place heavy items at the bottom of the bag

PACK TRAINS

Roping animals together on the trail is standard practice when traveling with more than one animal. When tying them together, ensure the ropes are long enough to allow the animals to walk comfortably, and short enough to prevent tripping.

THE PEOPLE

Appoint a driver to lead the group and take control of navigation. You will need sufficient people to control and care for the animals.

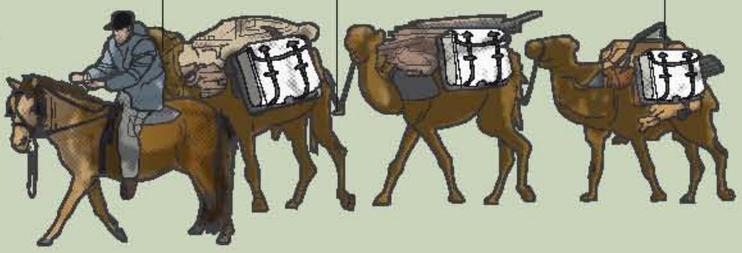
Driver leads the trail
Ropes need to be short enough to prevent tripping
Divide the loads equally

USING PACK ANIMALS TO SURVIVE

In an emergency situation, your priorities change significantly when using pack animals. Unpack the panniers and remove any heavy items not essential for survival. If necessary, you can climb on and ride to safety or, in a worst-case scenario, your animal can be used for food.

CARRYING INJURED PEOPLE

If a member of your group is injured, build a makeshift stretcher (see pp. 278–79) and use the animal to drag him or her to safety.

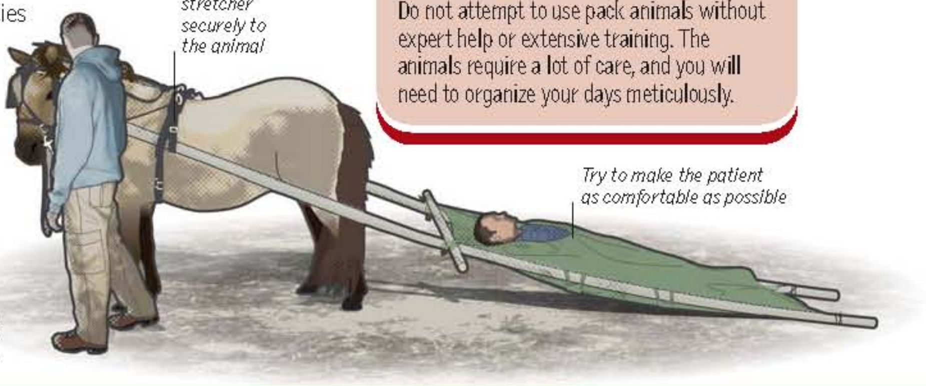

Strap the stretcher securely to the animal

Try to make the patient as comfortable as possible

WARNING!
Do not attempt to use pack animals without expert help or extensive training. The animals require a lot of care, and you will need to organize your days meticulously.

CHOOSING YOUR ANIMAL

The availability of pack animals depends on the region you are traveling in. Within any region, your choice of animal should take into account the load you are expecting them to carry, the distance you require them to travel, and the terrain you plan to cross. However, remember that individual animals' abilities can vary depending on their age and size.

ANIMAL	REGION	AVERAGE LOAD	ADVANTAGES	DISADVANTAGES
Horse	Worldwide	▪ 175–240 lb (80–110 kg) ▪ 20 percent of body weight	▪ Easy to train, with a good temperament ▪ Strong ▪ Can tolerate heat ▪ Can manage steep terrain	▪ Require grooming ▪ Can stray from camp if not securely tethered
Mule	Eurasia, Americas	▪ 120–180 lb (55–80 kg) ▪ 20 percent of body weight	▪ Will stick by a mare closely—this can be used to keep them moving and to keep them within the camp at night ▪ Very hardy and can manage steep terrain	▪ Young mules can be easily startled, so good training is essential ▪ Stubborn
Husky	Arctic	▪ 85lb (40kg) ▪ 7 dogs can pull 600lb (270 kg)	▪ Hardy—can cope well with snow and cold conditions ▪ Can travel fast	▪ Need a lot of fresh meat, which must be carried ▪ Prone to fight among themselves
Camel	Central Asia, North Africa, Australia	▪ 198–308 lb (90–140 kg) ▪ 30 percent of body weight	▪ Versatile on different terrains ▪ Can drink up to a quarter of its body weight and then go without water for several days	▪ Strong-willed and difficult to control ▪ Violent—can spit and bite
Llama	Andes	▪ 77–123 lb (35–55 kg) ▪ 25–30 percent of body weight	▪ Minimal environmental impact ▪ Travels well at high altitudes on difficult and steep terrain	▪ Makes horses and mules nervous ▪ Can be difficult to control if poorly trained
Elephant (Indian)	South Asia	▪ 1,650–2,750 lb (750–1,250 kg) ▪ 25 percent of body weight	▪ Can carry very heavy loads ▪ Can manage steep terrain	▪ Slow-moving ▪ Requires large amounts of food and water ▪ Takes time to train
Ox	Eurasia, Americas	▪ 300–450 lb (135–205 kg) ▪ 30 percent of body weight	▪ Hardworking ▪ Very strong ▪ Surefooted and can manage steep terrain	▪ Slow-moving ▪ Stubborn

FOUR-WHEEL DRIVING

A FOUR-WHEEL-DRIVE (4WD) vehicle can cover large areas of difficult terrain that would be inaccessible in a 2WD vehicle. With an experienced driver, most 4WD vehicles can tackle deep mud, water, snow, ice, and sand.

CHOOSING YOUR VEHICLE

When choosing your vehicle it's important to consider what you'll be using it for. Large vehicles have more internal space than smaller ones, but may find it harder to negotiate very rough ground without getting stuck. Similarly, very powerful vehicles can cross almost any terrain but use a lot of fuel, so may not be a viable option for long-term expeditions.

EXPEDITION CHECKLIST

- Check your map and inform others of your route and intended timeframe.
- Ensure the vehicle is fit for the journey, with all necessary spares. Check the fuel, oil, water, brakes, and hydraulic fluids, and check that there are no leaks.
- Inspect the tire treads, wheel nuts, lights, and steering arms.
- Always carry spare water, wheel(s), and fuel, as well as survival equipment, sand ladders, a 12v heavy-duty tire inflator, winching strap, and first aid kit.

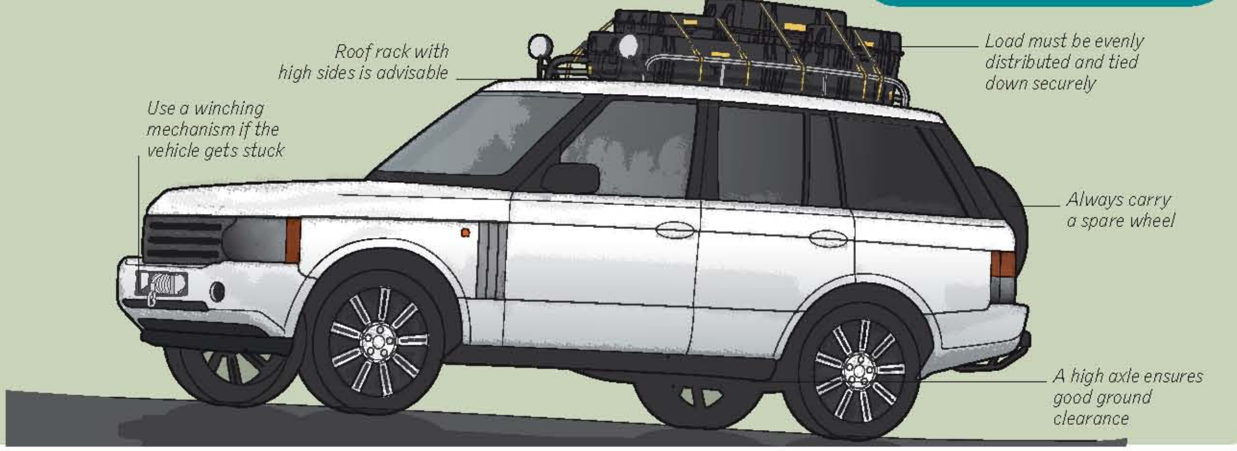
Roof rack with high sides is advisable
Load must be evenly distributed and tied down securely
Use a winching mechanism if the vehicle gets stuck
Always carry a spare wheel
A high axle ensures good ground clearance

GENERAL DRIVING TECHNIQUES

Four-wheel-drive vehicles enable you to manually switch between two-wheel-drive, for driving on roads, and four-wheel-drive, for low-traction conditions, such as on soft ground. The difference is the number of wheels powered by the engine at any one time.

CHOOSING A ROUTE

When unsure about the conditions immediately ahead, walk the route first, checking for potential problems and obstacles. Where necessary, mark the route you have walked and follow these markers when you drive. Ask yourself whether you really need to go that way and, if so, which route is best to avoid getting stuck. And, if you do get stuck, what are your options for self-recovery or escape?

WHEN TO ENGAGE 4WD

Engaging 4WD uses up a lot of fuel and should not be used on hard roads because of the risk of damaging tires and gears. Always choose the four-wheel-drive option on difficult terrain. It allows you to travel in a very low gear and gives the vehicle a far superior traction. Just before starting to cross the rough ground, stop the vehicle and engage the four-wheel-drive.

WARNING!

Don't hook your thumbs around the steering wheel when driving over rough terrain. Hitting a rut can jerk the wheel and break them.

DRIVING AS A TEAM

Driving off-road over rough terrain is both mentally and physically tiring. Take regular breaks and share the driving if in a group. You should always have at least two people in the vehicle:

- A driver—who takes responsibility for powering and steering the vehicle.
- A spotter—who takes responsibility for navigation and helps the driver pick a good route across the terrain. This may involve getting out of the vehicle and guiding the driver across a difficult section of ground.

DRIVING OVER DIFFICULT TERRAIN

A 4WD vehicle handles difficult terrain well because all four wheels can be powered by the engine at any one time (gas engines are generally more powerful, but diesel engines last longer and work well at low speeds). Using a few simple techniques, you will find that you are able to drive over terrain—and to places—that would otherwise be inaccessible. Always remember that the principle behind 4WD is to reduce the chances of you getting stuck—not to allow you to go further until you do get stuck.

MAXIMUM TRACTION

Your vehicle's traction on soft surfaces can be improved by lowering your tire pressure slightly. As a guide, place a brick ½in (1 cm) away from the side of a rear tire on flat ground, deflate the tire until it touches the brick, measure this pressure, and apply to all four tires. Make sure you can reinflate them.

DRIVING ON SAND

In soft sand the tires tend to move the sand from the front of the tire to the rear of the tire. If forward movement is halted for even a few seconds the wheel can dig itself into a hole. To prevent this, continually steer from side to side, so that the tire steers out of its own ruts. Avoid rapid changes in speed.

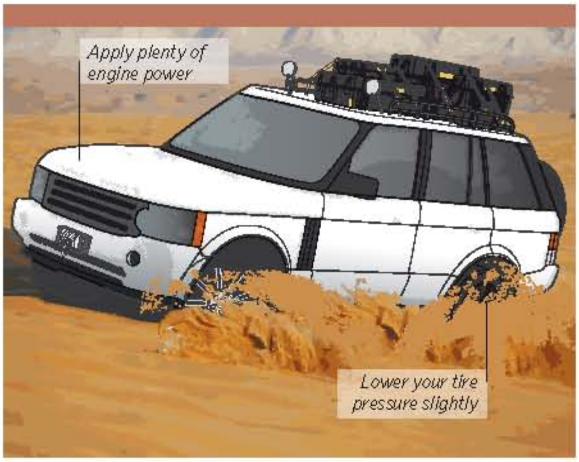
Apply plenty of engine power

Lower your tire pressure slightly

MUDDY TRACKS

Driving in mud requires concentration and the ability to adapt your driving to suit the conditions. In deep mud, use wide tires or lower the tire pressure slightly; however, if there is a hard surface below the mud this will make things worse. Steering outside existing track ruts is most likely to ensure the best traction.

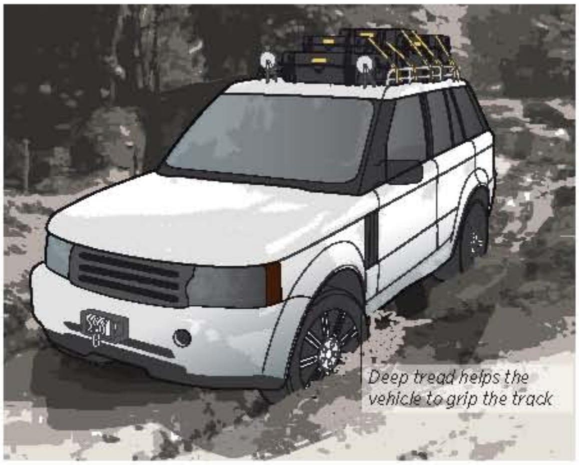
Deep tread helps the vehicle to grip the track

DRIVING ON SNOW AND ICE

Snow and ice require very smooth driving methods. Apply gradual pressure to the accelerator and brakes to avoid wheel spin, use low gears, especially when traveling downhill, and avoid changing gears unnecessarily. Using snow chains will increase traction, and you should practice fitting them before you need them.

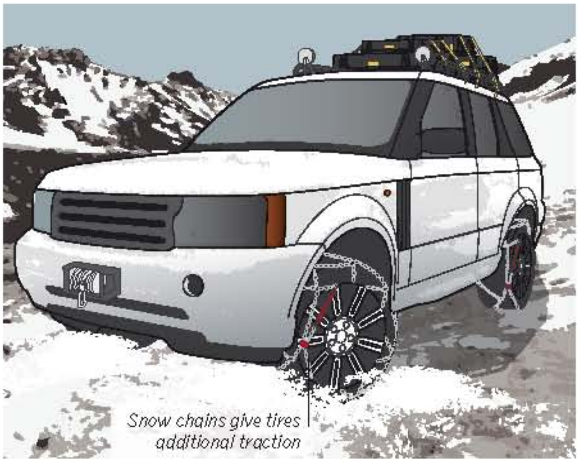
Snow chains give tires additional traction

CROSSING WATER

Always walk your route before driving across water. If the water seems too deep or the current too fast, don't attempt the crossing. It's important to drive at the correct speed—too fast will send water everywhere, but driving too slowly may flood the engine bay.

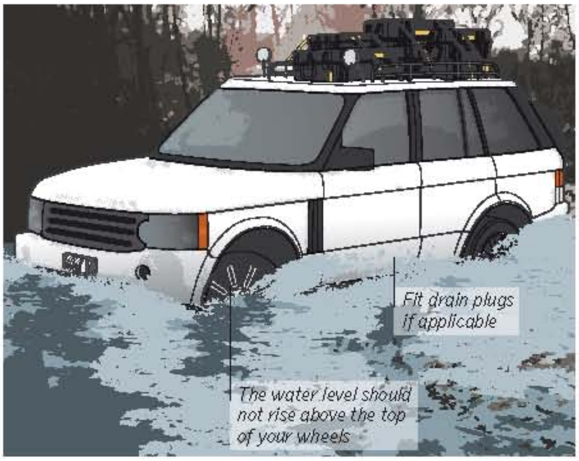
Fit drain plugs if applicable

The water level should not rise above the top of your wheels

RECOVERY FROM SOFT GROUND

Although using the correct techniques will certainly improve your chances of crossing soft ground successfully, it's important to know what to do if you do get stuck. Ideally you should never set out on a driving expedition with fewer than two vehicles. A second vehicle can be used for winching, dealing with breakdowns, or driving for help.

EQUIPMENT LIST

- Winching mechanism and strap
- Cable or tow rope
- Spare tire (inflated)
- Shovel and saw
- Sleeping bag or blanket
- Branches or stakes
- Sand ladders or mats

BASIC RECOVERY TECHNIQUES

When you get stuck, it's tempting to keep trying to aggressively free your vehicle. However, churning up what solid ground is left—and digging yourself into deeper ruts—will only make things more difficult in the long run. Stop, evaluate your options (reversing or pushing, digging, using branches, and winching), and calmly decide how best to achieve your aim. Don't act in haste—an ill-thought-out plan could leave you in more trouble than you were in originally.

REVERSING OR PUSHING

If you can't free the vehicle using four-wheel drive, try alternating between reversing and driving forward in first gear.

- If this has no effect, ask the passengers to get out and push while you drive forward.
- If you are making the situation worse, stop immediately and try an alternative solution.

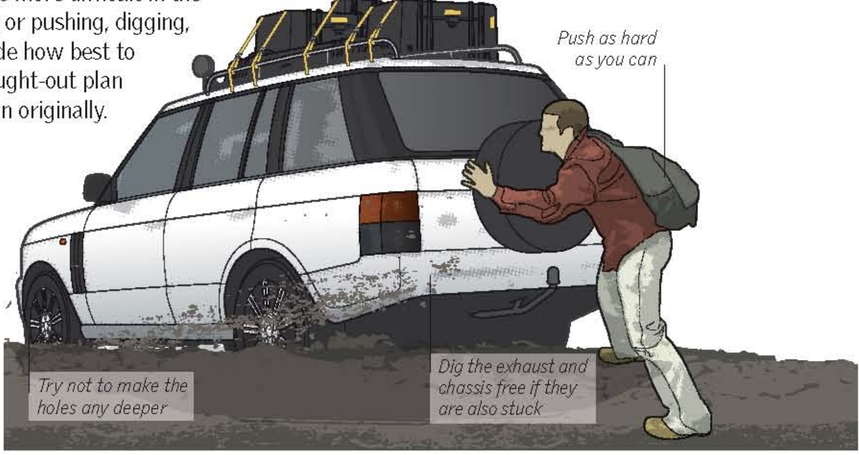

Push as hard as you can

Try not to make the holes any deeper

Dig the exhaust and chassis free if they are also stuck

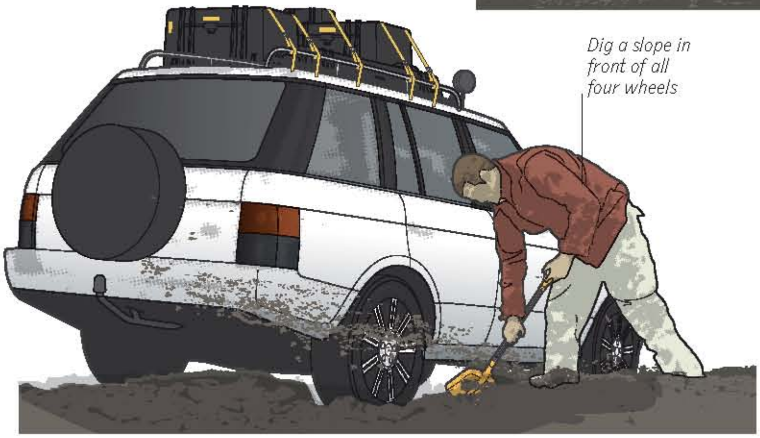

Dig a slope in front of all four wheels

DIGGING

If reversing and pushing don't work, the next step toward freeing your vehicle is to dig down in front of the wheels in an attempt to create a slope that you can then drive up.

- Dig out the sand in front of each tire to create an upward slope
- Drive very gently up the slope. Avoid revving the engine as this can cause the wheels to spin and lose their grip on the ground.

USING BRANCHES

If the vehicle still won't move, place branches, wooden planks, sand ladders, or blankets—in fact, anything that will increase your traction—in front of the wheels. The idea is to give the tires something to grip, and it should get you moving fairly quickly.

- Without revving the engine too much, gently ease the vehicle forward onto the branches or other material.
- Maintain a slow, steady speed and continue until you are back on firmer ground.
- Once clear, remember to stop and pick up your equipment, and remove any obstacles from the track.

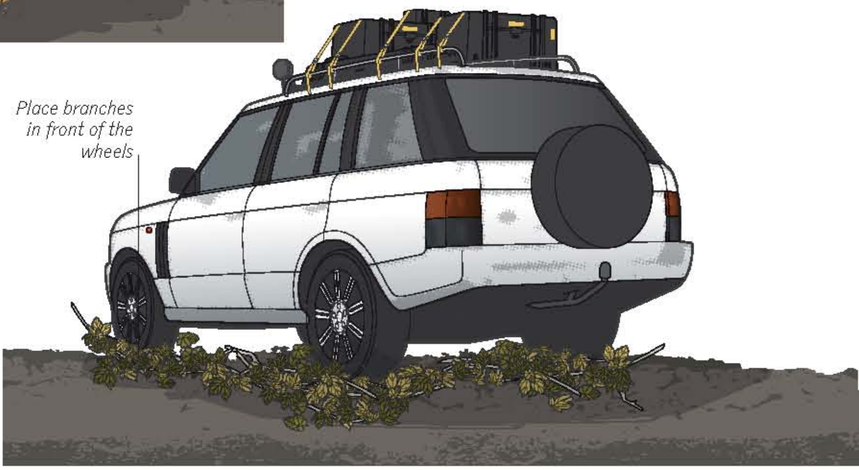

Place branches in front of the wheels

WINCHING

If none of the basic techniques work, and your vehicle is still stuck, it's time to consider winching. Use a cable attached to an electric-powered winching mechanism to pull the vehicle out of the hole via a strong anchor point. You can winch to another vehicle if you're traveling in convoy, although you do run the risk of both vehicles getting stuck. When winching, take time to assess your options—using a natural anchor point is the easiest method and should be your first approach.

NATURAL ANCHOR POINTS

Trees, rocks, roots, or deadfalls can all be used as anchor points. When using a tree, always place the strap or cable near the ground, and use a winch strap where possible to avoid damaging the tree. If the tree looks as if it may not be able to provide the required support, tie the tree to others in the vicinity. If you intend to use rocks, ensure they are large enough and firmly embedded in the ground.

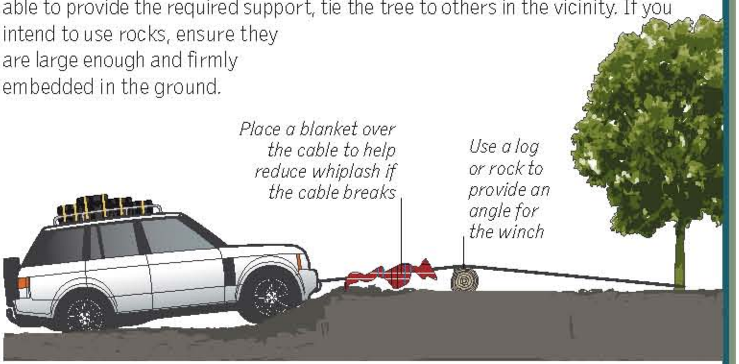
Place a blanket over the cable to help reduce whiplash if the cable breaks
Use a log or rock to provide an angle for the winch

BURIED ANCHOR POINTS

If there are no natural anchor points, you can construct an improvised anchor by burying objects such as logs or a spare wheel. Dig a hole in the ground that's at least 3 ft (1 m) deep, attach your winch cable to your improvised anchor, then bury the anchor in the hole. Refill the hole to secure it. If you're using a tire, use the tire lever behind your spare wheel as an attachment point for the cable.

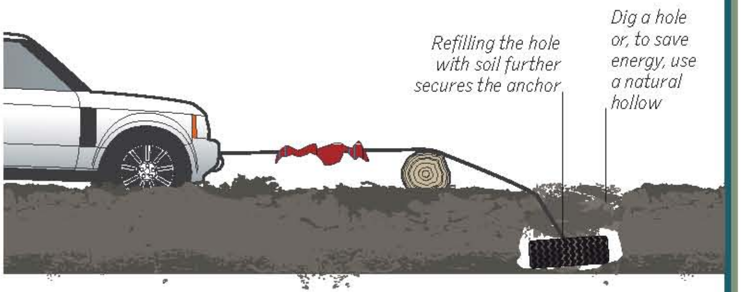
Refilling the hole with soil further secures the anchor
Dig a hole or, to save energy, use a natural hollow

WINCHING TO A STAKE

A series of long stakes in the ground can also provide an improvised anchor point. You will need a long, sturdy main stake—to which the winch cable is attached—and a series of supporting stakes, lashed together for additional strength. Push the stakes securely into the ground at a slight angle. Be careful not to stand near the stakes when winching, in case they come loose.

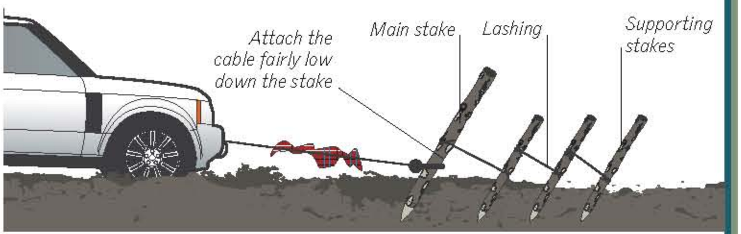
Attach the cable fairly low down the stake
Main stake
Lashing
Supporting stakes

WARNING!

When winching, don't stand within range of the cable in case it snaps under the strain. A breaking cable can cut a tree—or a person—in half. Place a blanket or sleeping bag over it to absorb some of the shock and reduce whiplash.

EMERGENCY ACTION

Knowing what to do in certain emergency situations will help you stay calm if the worst happens.

FRONT SKIDS

The term "skid" covers any kind of slide in which the wheels fail to grip the road. In a front skid—also called an "understeer"—the front wheels fail to turn when you steer so the vehicle continues to travel forward.

- Steer in the direction of the skid but be careful not to oversteer. If you're on ice and skidding straight, step on the clutch or shift to neutral.
- Once you have gained control, correct your course and continue driving. Consult your driver's manual if you have anti-lock brakes.

REAR SKIDS

In a rear skid, the back wheels lose control and the vehicle rotates more than usual and can spin.

- Steer in the direction of the skid but face toward where you want to travel.
- Once facing in the right direction, bring the steering wheel back to the center and continue to steer, regardless of what the rear is doing, until you have gained total control.

BRAKE FAILURE

Before setting out on a drive, always check to see whether there is brake fluid leaking from the vehicle. If so, don't travel. If the brakes fail when on the move, do the following:

- Without turning off the engine, shift down the gears to reduce speed while maintaining control.
- Once the vehicle reaches a speed of less than 25 mph (40 kph) apply the handbrake, keeping a firm grip on the steering wheel.

ACCELERATOR MALFUNCTION

If the accelerator sticks and your engine will not slow down, you can decelerate by putting the vehicle into neutral and applying the brakes.

- If it's safe to do so, switch off the engine, but you'll lose functions like power steering and lights.
- Steer safely to a stop. If possible, avoid using the handbrake as this may cause a skid—although it may be necessary if you need to avoid a collision.

KAYAKING AND CANOEING

TRAVELING IN A KAYAK OR CANOE is a good way of negotiating river systems and open water. With your equipment safely stowed in your boat, expeditions can range from short day trips to year-long adventures.

KAYAKING

Kayaks are an extremely efficient method of water transport, especially on open sea, rivers, and lakes. Their narrow design and light frame make them easy to maneuver.

CAPSIZING IN A KAYAK

Capsizing is an almost inevitable part of kayaking in rough water. The "wet exit"—slipping out of the kayak completely while still holding on to it—is sometimes the only option, but wherever possible use an eskimo roll instead.

ESKIMO ROLL

Mastering the eskimo roll enables you to avoid getting your legs wet and having to climb back in your kayak. With practice, it should become one fluid movement.

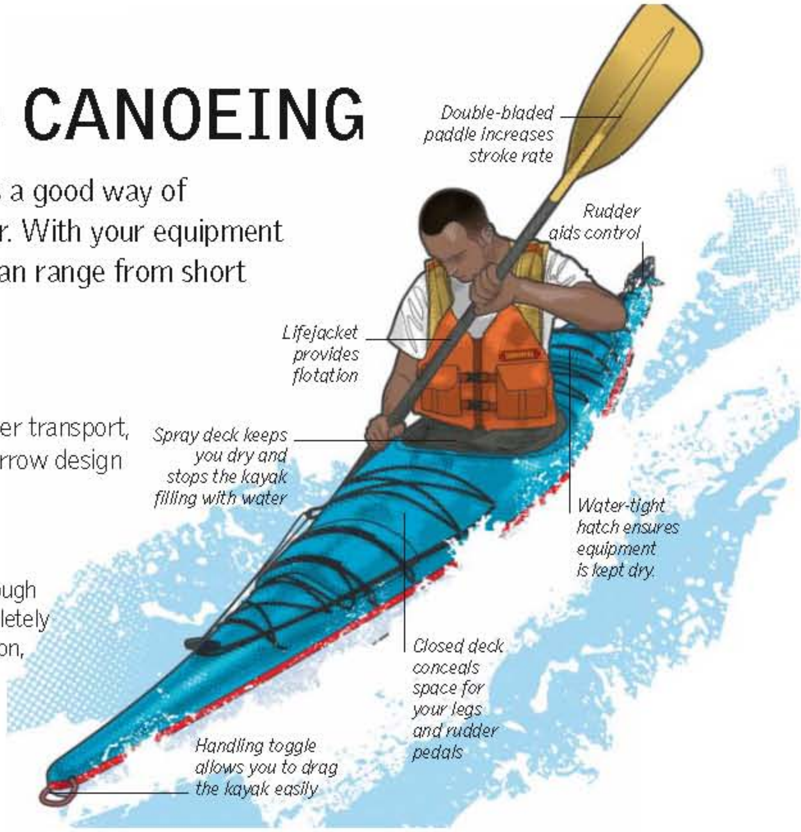
Double-bladed paddle increases stroke rate
Rudder aids control
Lifejacket provides flotation
Spray deck keeps you dry and stops the kayak filling with water
Water-tight hatch ensures equipment is kept dry
Closed deck conceals space for your legs and rudder pedals
Handling toggle allows you to drag the kayak easily

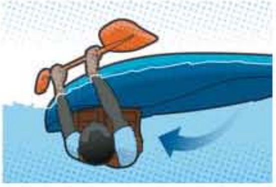

① From the upside-down position, twist your body to the side of the kayak.

- Gripping your paddle, firmly push your arms out of the water, then rest them on the side of the kayak.

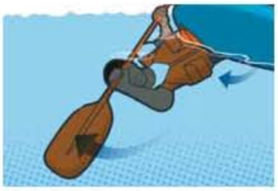

② With your head near the surface, sweep your body and paddle away from the side of the kayak.

- At the same time, flick your hip away to start the rotation of the kayak.

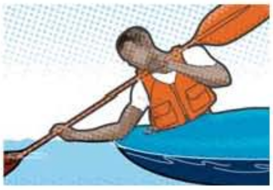

③ Continue the hip flick until the kayak returns to an upright position in the water.

- Straighten your back so you are sitting upright again, and continue to paddle your course.

WARNING!

- Always wear a lifejacket. Empty water bottles in your pockets can be used as makeshift flotation devices in an emergency.
- In a survival situation, never leave your kayak or canoe unless it is life-threatening not to do so. A kayak is highly visible, especially from the air, and is far more likely to be spotted first by rescuers.

BASIC PADDLING SKILLS

Kayak paddles have two blades that sweep alternately through the water, propelling the kayak forward. To steer, paddle on the opposite side to the direction in which you want to go. Some kayaks have a foot-operated rudder.

CATCH

Place your lead blade firmly in the water so that it enters near to your feet.

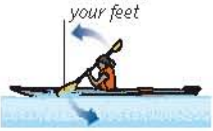
Lead blade enters the water near your feet

BRACE

Rotate your torso and lead blade in preparation for the propulsive stroke.

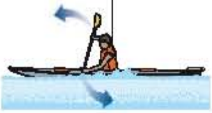
The power comes from your torso

STROKE

Sweep your lead blade firmly through the water to propel the kayak forward.

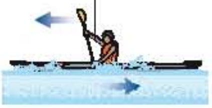
Keep your grip loose on the following arm

TRANSITION

As the blade exits the water, repeat the catch on the other side and continue paddling.

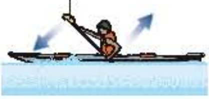
The transition should be as smooth as possible

CANOEING

As with kayaks, canoes can hold one or two people with equipment, and are powered by the body from a seated position. They are generally wider than kayaks, however, and their decks are open. Canoe paddles have one blade as opposed to two.

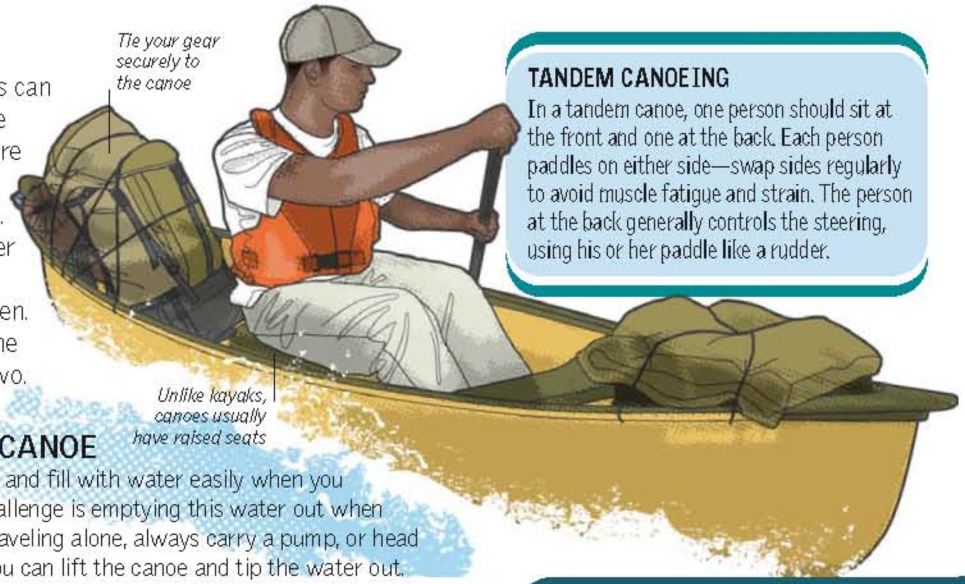
Tie your gear securely to the canoe

Unlike kayaks, canoes usually have raised seats

TANDEM CANOEING

In a tandem canoe, one person should sit at the front and one at the back. Each person paddles on either side—swap sides regularly to avoid muscle fatigue and strain. The person at the back generally controls the steering, using his or her paddle like a rudder.

CAPSIZING IN A CANOE

Because canoes are open and fill with water easily when you capsize, your greatest challenge is emptying this water out when the canoe is righted. If traveling alone, always carry a pump, or head for shallower water so you can lift the canoe and tip the water out.

CANOE-OVER-CANOE RESCUE

If you are traveling with another canoe and are far from shallow water, the best option is the canoe-over-canoe rescue. The capsized canoe is rotated upside-down and pushed onto the rescue vessel, which empties the water out. The people on the rescue craft then rotate the canoe back to an upright position and return it to the water.

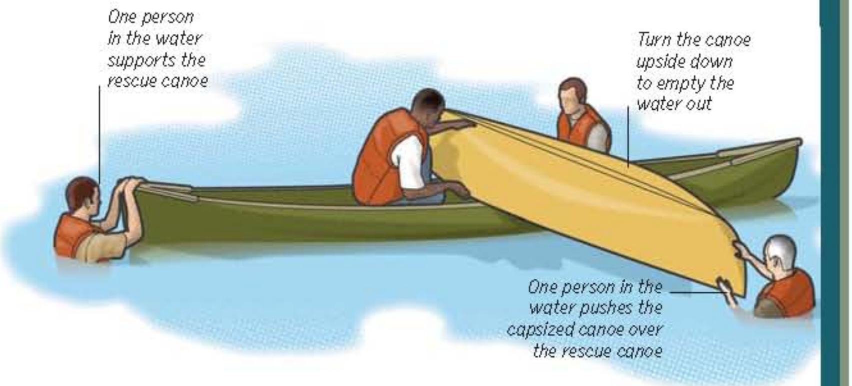
One person in the water supports the rescue canoe

Turn the canoe upside down to empty the water out

One person in the water pushes the capsized canoe over the rescue canoe

KAYAKING AND CANOEING TIPS

- Research your route and weather conditions thoroughly.
- Don't overload the boat, and ensure it's well balanced.
- Carry a high-volume pump so you can rescue yourself when alone.
- If watertight bulkheads are not fitted on your kayak, use airbags to add buoyancy and reduce the amount of water that can enter a compartment.
- Use waterproof dry bags to keep your kit dry.
- Clip valuables onto the boat or yourself. If they don't float, attach a float to them so they won't sink if they fall in the water.
- Carry waterproof material and bungee cord to repair blown hatches.
- Never enter a cave alone, or without head protection.
- Fit a leash to your paddle and attach it to the craft. If you then have to perform a rescue you can throw it in the water and keep both arms free.
- Chewing gum is great for the short-term repair of small holes in the craft, as is black masking tape and plumber's repair tape.

BASIC PADDLING SKILLS

When canoeing solo, use a "switch" style of paddling to ensure your canoe travels in a straight line—if the canoe starts to veer off course, switch sides.

CATCH

Sit in the middle of the canoe, ensuring that it's well balanced. Reach forward and thrust the blade into the water.

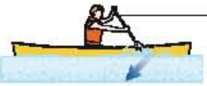
Paddle enters the water at the front of the canoe

DOWN STROKE

Drive the paddle swiftly and firmly down into the water so that it assumes a vertical position.

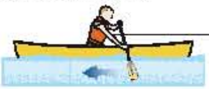
Use strong arm motions

PULL STROKE

Firmly pull the paddle blade backward through the water. This will propel the canoe forward.

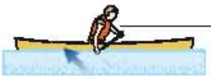
Paddle generates forward momentum

REPEAT STROKE

Lift the blade out of the water toward the rear of the canoe and return to the starting position to repeat the stroke.

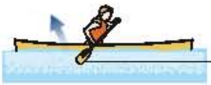
Paddle leaves the water at the rear

BUILDING A RAFT

IF VENTURING INTO the wilderness, it is useful to know how to build an improvised raft or flotation aid. You may find that a major water obstacle lies between you and rescue, or you may be in an area in which dense undergrowth makes cross-country travel difficult but river travel relatively easy. However, in a survival situation you are unlikely to have a lifejacket, so assess the risks carefully.

WARNING!

Apart from the brushwood raft (below), most improvised rafts will float half-submerged, so you'll be constantly sitting or kneeling in water. This could lead to hypothermia in certain conditions, so build an additional raised platform if necessary.

TOOLS AND MATERIALS

- Knife or small saw
- Poncho, shelter sheet, or tarpaulin
- Long sticks, brushwood, and thatch
- Cordage

1 Construct your raft near a suitable launching place.

- Lay your poncho on the ground to determine the size of the finished raft. Make sure there's enough material to come up the sides and gather on top.

2 Peg out sticks to form an oval shape, the size and shape of your intended raft.

- Interlace brushwood and thatch tightly through the sticks to form the sides of the raft. The tighter you weave the brushwood, the stronger the body of the raft will be, and the better it will float. The height of the brushwood will also determine how high the raft sits in the water.

3 Form a sitting platform by pushing a selection of long poles or sticks through the sides of woven brushwood. Weave the sticks under and over each other as much as you can to create a stable structure.

- Using your knife or a small saw, trim the sticks to size so that they don't protrude from the sides.
- Remove the pegs from around the brushwood.

MAKING A BRUSHWOOD RAFT

If you have a poncho, shelter sheet, or tarpaulin, you can construct a brushwood raft. This is a one-man raft, but if built correctly will keep you completely out of the water. Build it as close to the water as possible so you don't have to carry it far.

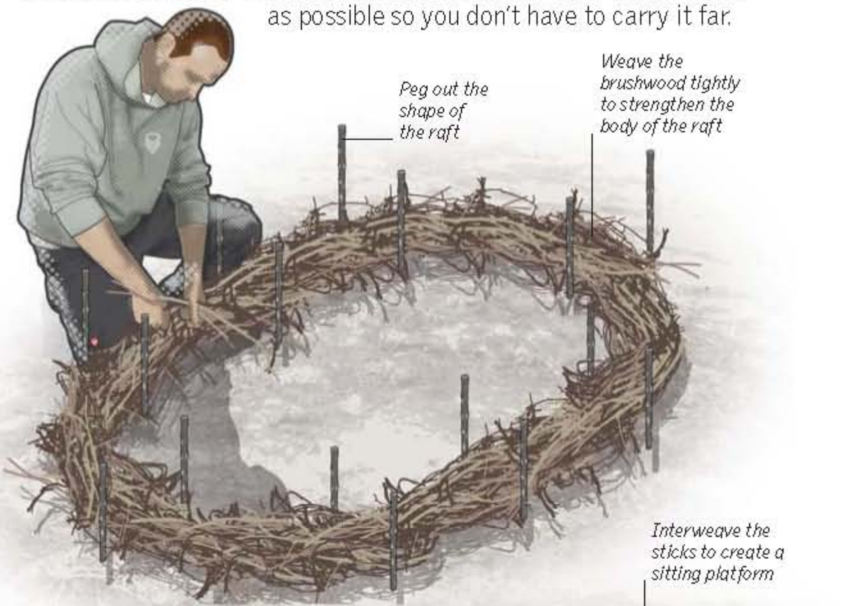
Peg out the shape of the raft
Weave the brushwood tightly to strengthen the body of the raft

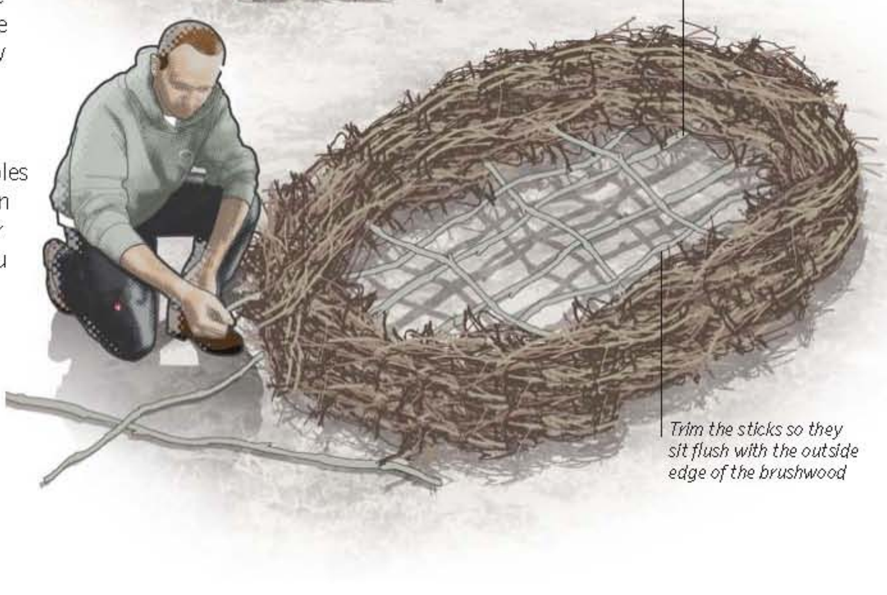
Interweave the sticks to create a sitting platform
Trim the sticks so they sit flush with the outside edge of the brushwood

MAKING A PADDLE

In most cases, you can use the current of the river for momentum rather than having to paddle, and your walking staff to maneuver the raft. However, where there is no current you will have to provide the propulsion yourself, and will need to improvise some form of paddle.

1 Find a suitable length of green wood to form the handle. This should be as wide as possible but still comfortable to grip.

- In the end of the wood, make a split long enough to form the paddle face.
- Collect smaller branches and insert one of them into the split. Lash into place.

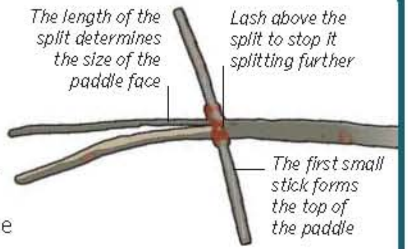
The length of the split determines the size of the paddle face
Lash above the split to stop it splitting further
The first small stick forms the top of the paddle

2 Continue to force smaller branches into the split, lashing each one as you progress until you have a sufficient paddle area.

- Securely lash the two open splits of the handle together.
- To make the paddle more rigid, lash the ends together.

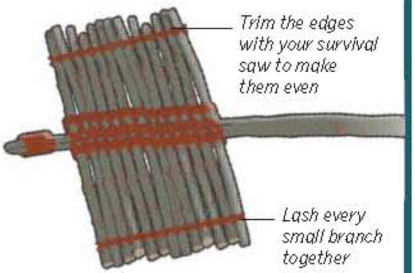
Trim the edges with your survival saw to make them even
Lash every small branch together

4 Slide your poncho underneath the raft to form a waterproof seal.

- Push the poncho hood through to the inside and tie its neck securely to ensure it's completely water tight.
- Pull the poncho up the sides of the raft and lash securely across the top.

5 Fill the space underneath the sitting platform with natural materials, such as additional brushwood, grass, or moss. You could also use any item that would aid buoyancy—such as empty plastic bottles or waterproof bags filled with air and tied off.

- Drag or push the raft into the water, checking to see whether it floats in shallower water before loading your equipment and climbing aboard.

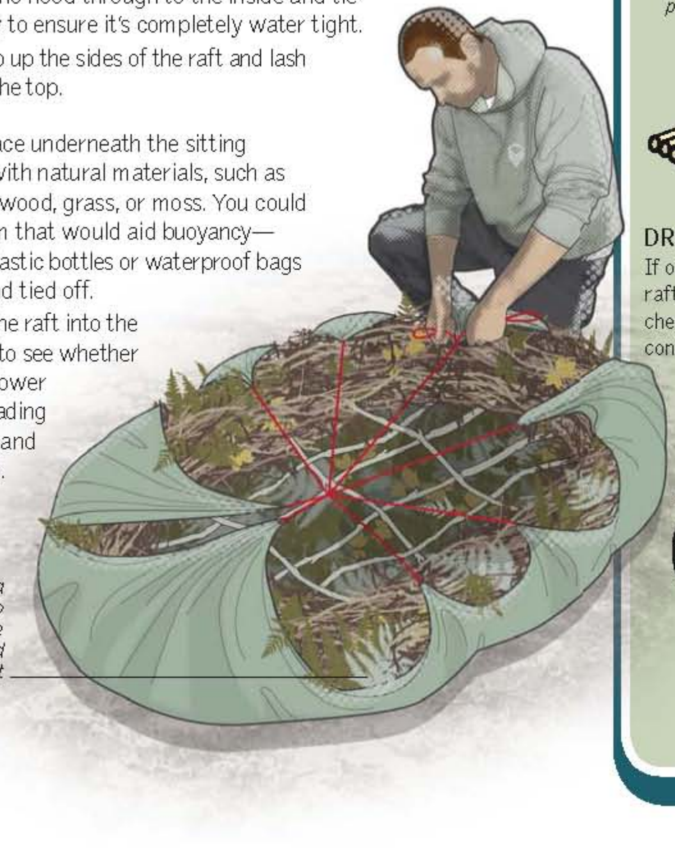
Add extra foliage to increase buoyancy and stability of raft

ALTERNATIVE RAFTS

Use whatever materials are available to you to make your raft. Look out for logs, bamboo, and discarded oil drums, as they are naturally buoyant materials that can be used to make an effective raft.

LOG RAFT

To make a log raft, use dry wood—ideally standing deadwood—as this will float higher in the water. Cut notches in the logs to allow the cross-pieces to sit snugly.

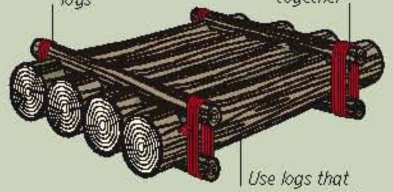
Cross-pieces tension the logs
Lash the cross-pieces together
Use logs that are roughly the same diameter

BAMBOO RAFT

Bamboo is made up of hollow compartmentalized sections, which means it is ideal for raft-making. Bamboo rafts are much lighter than log rafts.

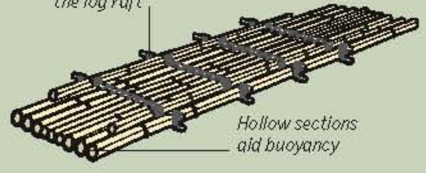
Use cross-pieces, as with the log raft
Hollow sections aid buoyancy

DRUM RAFT

If oil drums are available, they make ideal rafts. Always take care when handling chemical drums, as they may once have contained toxins.

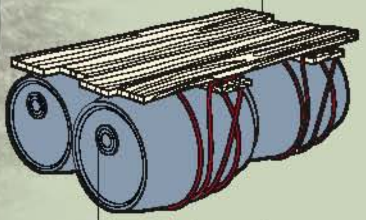
Tie the drums securely to a wooden platform
Ensure that any openings (missing filler caps, for example) sit above the water level

SWIMMING

WHEN THE WATER IS TOO DEEP to wade and you don't have the materials to build a raft, you may need to swim to negotiate a water obstacle. Before you get into the water, plan how you will get warm and dry when you get out. Making a float will help to keep clothing and equipment dry.

WARNING!
As with any survival activity involving water, try to avoid getting wet—find a way around the obstacle if possible, or find shallower water to wade across. The cold water increases your risk of hypothermia so ensure you have the necessary equipment for building a fire. Never enter water where there are likely to be dangers such as crocodiles or hippos.

SURVIVAL SWIMMING

Unless you need to cross rocky ground to enter the water, remove your boots and most of your clothing to reduce drag and keep them dry. Place them either in a waterproof survival bag or in an improvised poncho float (see opposite) and float them across.

SWIMMING ACROSS RIVERS

Choose a safe place to cross (see p. 88), and plan where you will exit the water.

- Take the current into account as you are likely to drift downstream slightly.
- Lower yourself in gently—never jump or dive into the water.
- If the water is cold, gently immerse yourself until your body has recovered from the initial shock.

THE CROSSING

Choosing a stroke you feel comfortable with (and that will expend the least energy), swim across your route with your float in one arm.

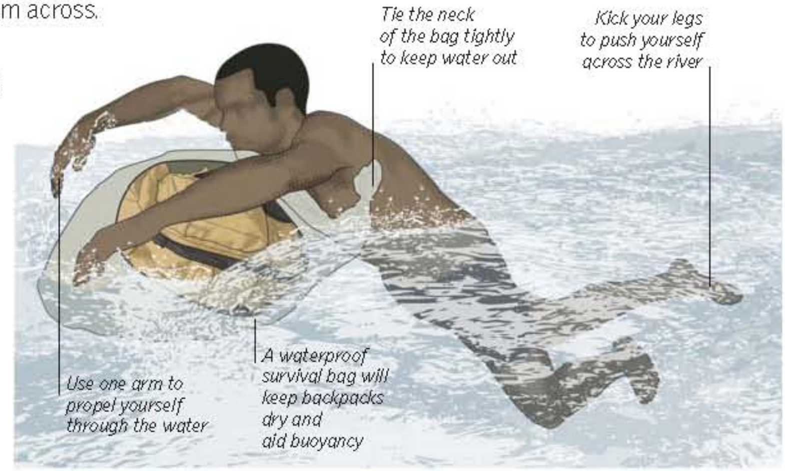
Tie the neck of the bag tightly to keep water out
Kick your legs to push yourself across the river
Use one arm to propel yourself through the water
A waterproof survival bag will keep backpacks dry and aid buoyancy

SWIMMING DOWN RIVERS

Avoid swimming up or down rivers—walking the route instead is far safer. However, if you accidentally fall in and find yourself traveling downstream in fast-moving water, there are ways of protecting yourself until you reach safety.

DEFENSIVE SWIMMING

The aim of defensive swimming is to adopt a position that keeps you as safe as possible and prevents your feet from becoming ensnared in rocks, which, in a strong current, could push you under the water. Assume the defensive swimming position until you reach shallower water, and can stand up and climb onto the bank.

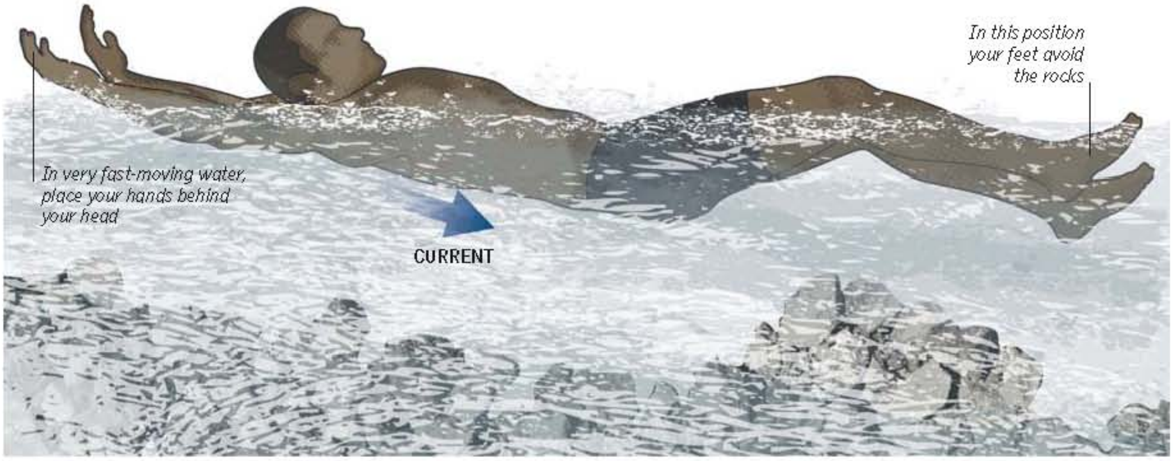
In this position your feet avoid the rocks
In very fast-moving water, place your hands behind your head
CURRENT

SWIMMING STROKES

It's important to know what the different swimming strokes are, although you're unlikely to stick to them strictly in a survival situation. If you're a confident swimmer, always choose the stroke that will use the least amount of energy, such as breaststroke. If you're less experienced you should avoid getting in the water unless in a life-or-death situation.

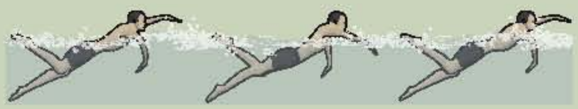

DOG PADDLE

An easy stroke suitable for less confident swimmers, this involves simply paddling forward with your arms and legs. When you paddle with your left arm, kick with your right leg and vice versa.

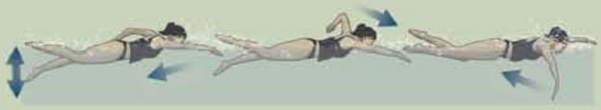

FRONT CRAWL

Also known as "freestyle," front crawl uses a lot of energy so is not the best stroke for a normal survival situation. However, you can build up a lot of speed, which may sometimes be required.

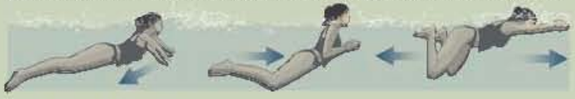

BREAST STROKE

This is perhaps the most common stroke for crossing rivers. Your arm and leg movements should be made simultaneously and smoothly. It is a good option if you're pulling a float.

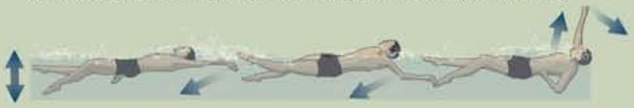

BACK STROKE

If you know the water is safe, this stroke can be a good choice, as it uses little energy. However, not being able to see where you're going can be a disadvantage, especially in rocky areas.

MAKING A PONCHO FLOAT

Constructing a poncho float will keep your belongings dry and provide limited flotation as you cross the river. If you don't have a poncho, use any large piece of waterproof material.

1. Push your poncho hood to the inner side, and tightly secure its neck with the drawstrings or cordage.
   - Lay the poncho on the ground with the inner side facing upward.
   - Place your gear on top.

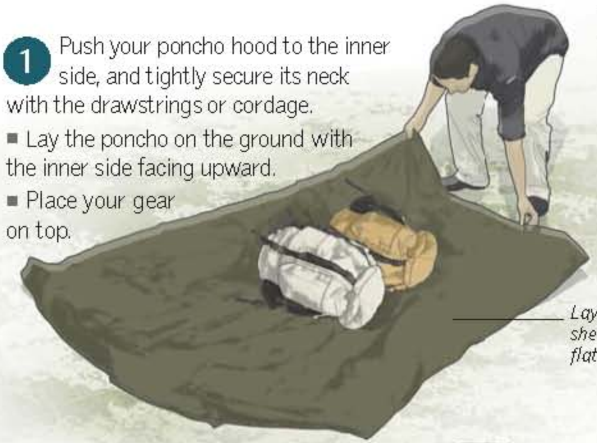
Lay the sheet on flat ground

2. Bring one side of the poncho up and over the equipment that is piled together in the middle.
   - Repeat with the other three sides to create a rectangular parcel.
   - Fold the corners, and ensure that they won't allow water to enter the float.

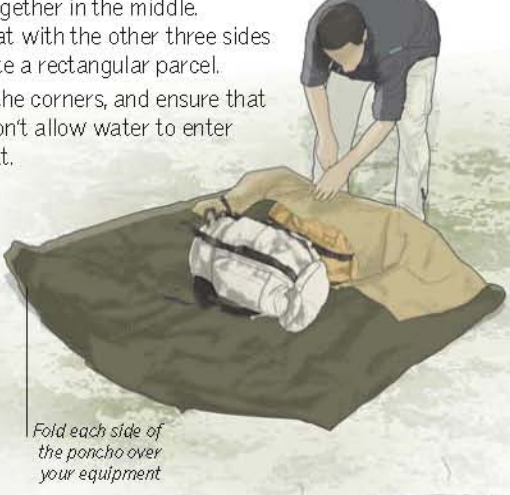
Fold each side of the poncho over your equipment

3. Wrap carefully to make it watertight.
   - At this point, if you have a second poncho, repeat from step one, placing the float face down on top of the second sheet and wrapping again.
   - If available, place brushwood inside the float for added buoyancy.

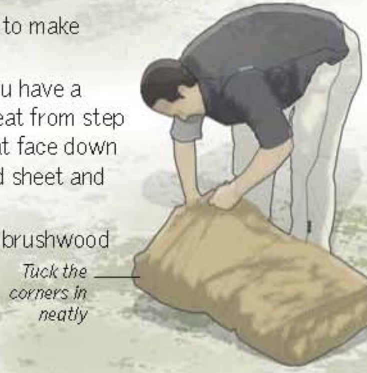
Tuck the corners in neatly

4. Tie ropes, bootlaces, vines, or improvised cordage around the float to secure it.
   - As you enter the water, gently lower the float in with you. Pull it along as you cross the river.

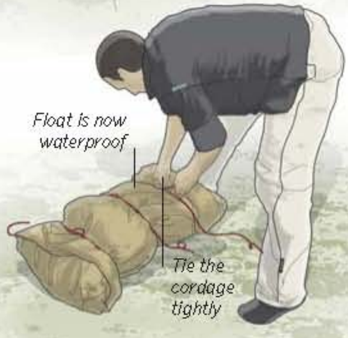

3

CAMP CRAFT

CAMP ESSENTIALS

Whether you intend to remain where you are and wait for rescue, move to a safer area and await rescue there, or walk out of the survival situation yourself, you're probably going to need to select a site on which to put up a shelter. This could be for just a single night or for a longer-term stay.

A sound understanding of what constitutes a good location will allow you to address the basic principles of survival. For example, a well-selected site will help keep you safe from danger while still allowing you to deploy your location aids effectively. It should also provide you with adequate materials for building a shelter and a fire, and offer an accessible supply of water, both for drinking and for hygiene.

A well-organized site will not only give you a sense of purpose and order, but will also provide a safe environment for yourself and your equipment. Designating an area for storing equipment and tools, for example, will help prevent vital items from being lost and will reduce the likelihood of you or members of your team being injured.

In this section YOU WILL DISCOVER...

- how to **make a shower** before you **take a shower...**
- why **keeping clean** is good for both **body and soul...**
- how to work up a **lather** with **wood ash...**
- when a **thatched screen** could **protect your modesty...**
- what turns a **hard candy** into a **firestarter...**
- the **difference** between **cattails** and **fire dogs...**

A fire is an integral part of any campsite. It can be used for warmth, for purifying water, for cooking, for signaling to potential rescuers, for protection against wild animals, and for providing light when darkness falls. It also provides a sense of security. The psychological effects of being able to start a fire in a survival situation should never be underestimated; neither should the psychological effects of *not* being able to start one. Even in a survival situation, a simple camp can give a sense of normality and "home."

You can produce a spark—and therefore light a fire—by using a device such as a firesteel or a ranger's flint and steel (see p. 127).

A FIRESTEEL COMPRISES two main parts: the material that will produce the spark (usually a rod made from ferrocerium or magnesium alloy) and a sharp striker device (usually a knife blade or short piece of hacksaw blade). When the striker is drawn over the rod, a spark is produced.

To control the spark created by these two moving parts, follow the method outlined below. It will allow you to direct the spark accurately, and reduce the chance of you knocking and scattering your tinder. This can be a problem if you hold the rod next to the tinder and strike down into it—especially if you are cold, wet, and tired and your hands are shaking.

1. Place the rod in the center of the tinder. Then position your striker onto the rod. Lock the hand holding the striker into position.
2. Pull the rod up and away from the tinder, drawing it against the striker. By drawing away from the tinder you avoid the danger of disrupting or scattering it.
3. To direct the sparks, alter the angle at which you pull the rod up and away.

> "The **psychological effects** of being able to **start a fire** should not be underestimated; neither should the effects of **not being able** to start one"

ORGANIZING YOUR SITE

WHERE YOU CHOOSE to set up your shelter depends on the environment, but always take into account the four principles of survival: protection, location, water, and food. First, make sure there are no obvious dangers, and that you're able to signal for rescue. If possible, select a location close to a water source.

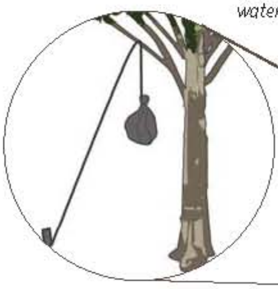

If predators are in your area, hang unused food in a bag 10 ft (3 m) off the ground and 3 ft (1 m) away from the trunk or branches of a tree

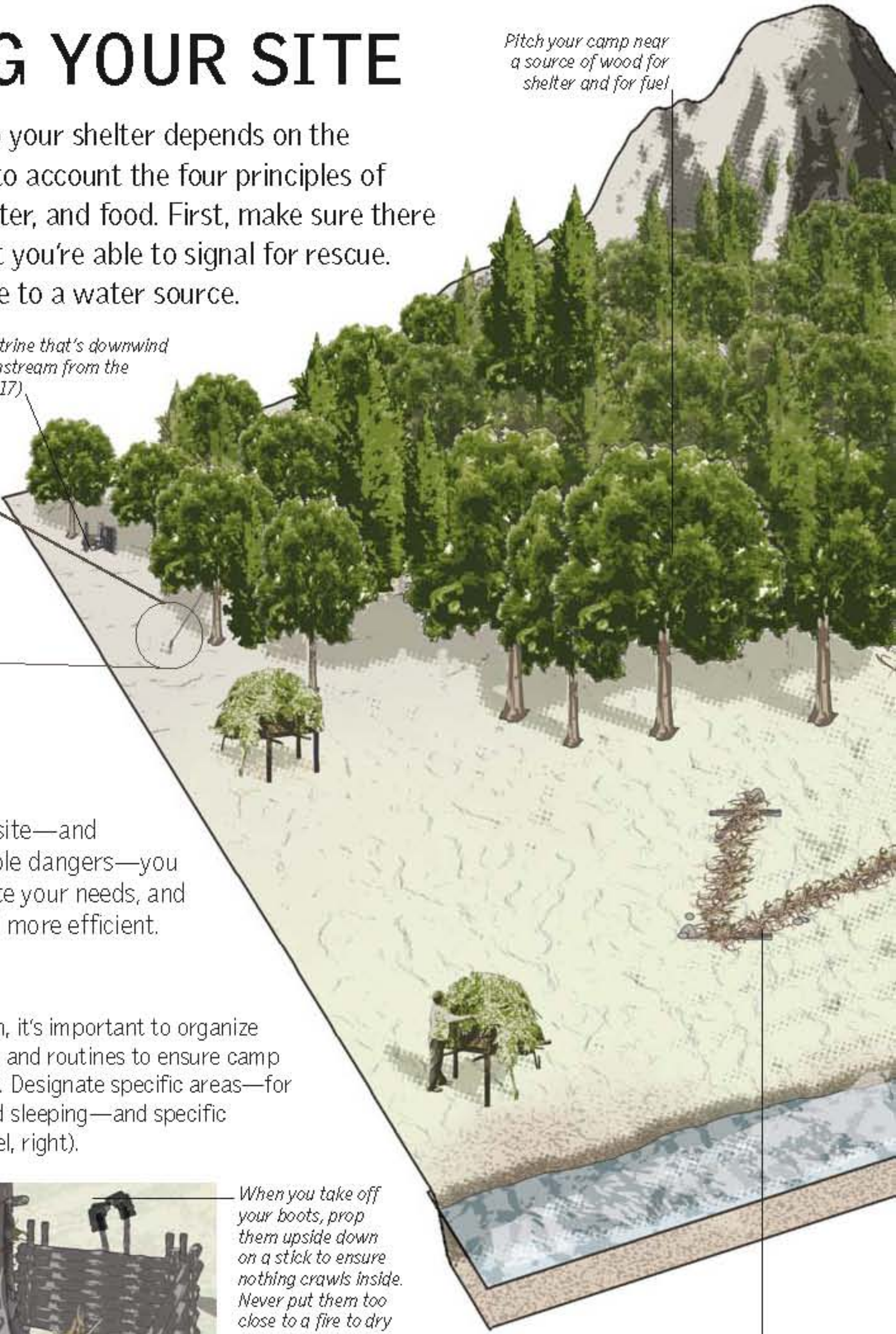

Pitch your camp near a source of wood for shelter and for fuel

Select a place for a latrine that's downwind of the camp and downstream from the water source (see p. 117)

Assign an open area for location aids, such as a large "V", and as a potential landing site for a rescue helicopter (see pp. 236–41)

SITE FEATURES

Once you're happy with your chosen site—and confident it doesn't hold any forseeable dangers—you can organize the area to accommodate your needs, and to make your time in camp easier and more efficient.

CAMP ADMINISTRATION

Whether you're in a group or on your own, it's important to organize your site, and quickly establish disciplines and routines to ensure camp safety and to reduce the risk of accidents. Designate specific areas—for storing equipment, firewood, cooking, and sleeping—and specific routines for when you do things (see panel, right).

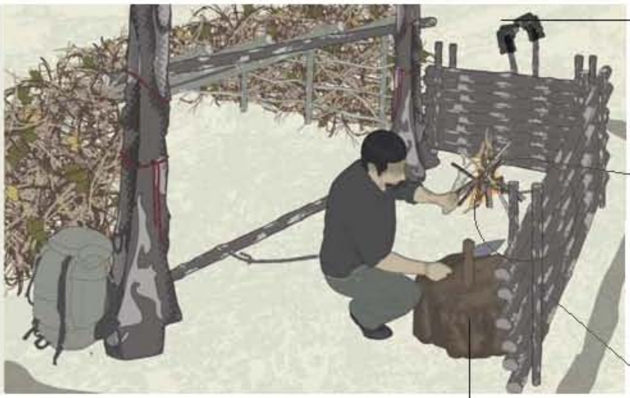

When you take off your boots, prop them upside down on a stick to ensure nothing crawls inside. Never put them too close to a fire to dry

Create a safe area for cutting wood. Use a tree stump as a platform for cutting and chopping

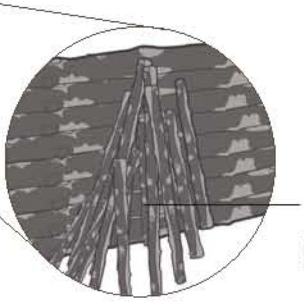

Collect and cut the fuel you need before it gets dark and keep it dry in an upright stack

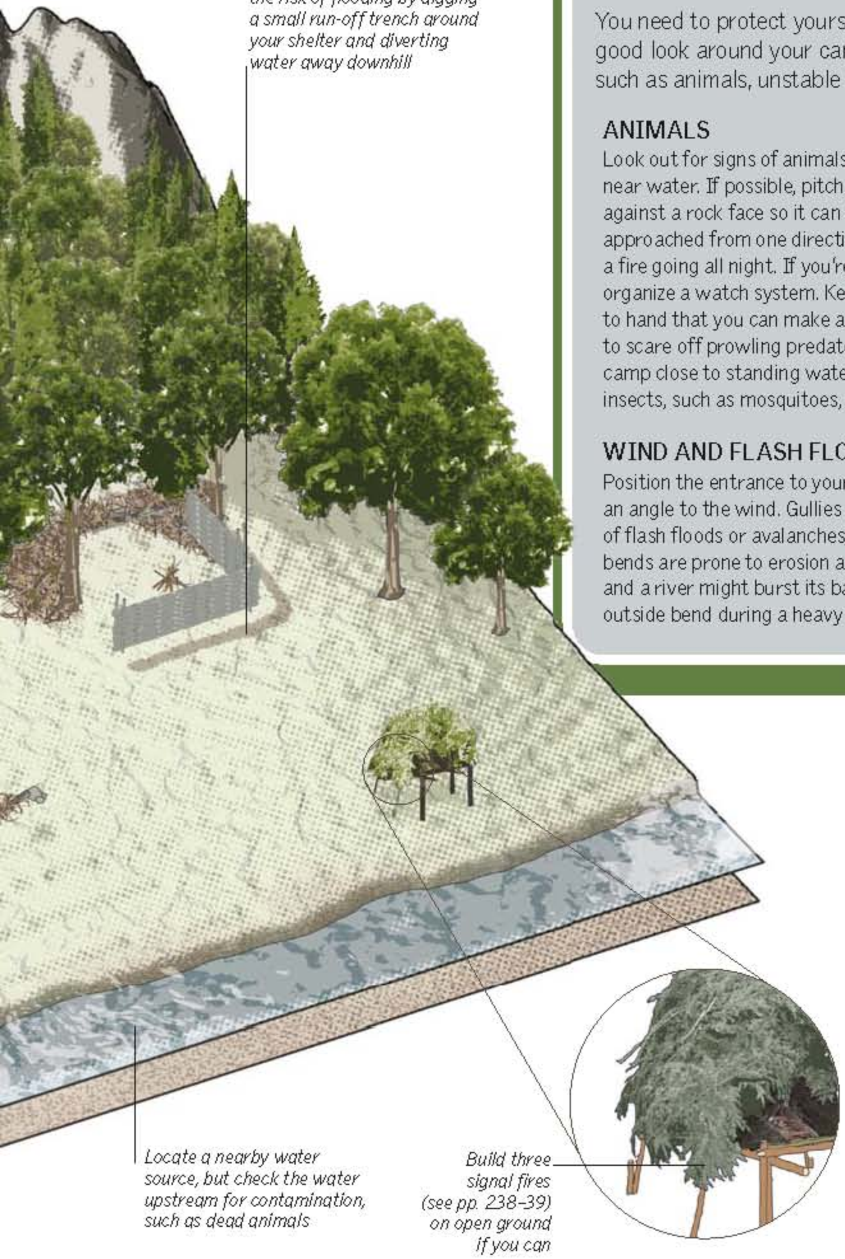

If heavy rain is likely, reduce the risk of flooding by digging a small run-off trench around your shelter and diverting water away downhill

Locate a nearby water source, but check the water upstream for contamination, such as dead animals

Build three signal fires (see pp. 238–39) on open ground if you can

ASSESSING YOUR ENVIRONMENT

You need to protect yourself from injury, threats, and risks, so have a good look around your campsite and assess it for any potential dangers, such as animals, unstable rocks or trees, and the likelihood of flooding.

ANIMALS

Look out for signs of animals, especially near water. If possible, pitch your camp against a rock face so it can only be approached from one direction. Keep a fire going all night. If you're in a group, organize a watch system. Keep things to hand that you can make a noise with to scare off prowling predators. Don't camp close to standing water where insects, such as mosquitoes, swarm.

WIND AND FLASH FLOODS

Position the entrance to your shelter at an angle to the wind. Gullies run the risk of flash floods or avalanches; inside river bends are prone to erosion and floods; and a river might burst its banks on an outside bend during a heavy downpour.

STANDING DEADWOOD

These are dead trees that haven't yet fallen. Heavy wind, or the weight of rain or snow, can make them fall. This is the best type of wood for kindling and fire fuel (see p. 121).

DEADFALLS

These are dangerous branches that have broken off a tree, but haven't yet fallen to the ground. Some trees, such as beech, ash, and yew, drop their branches without warning.

ROCKFALLS AND ICE FALLS

If you camp next to rocks, check for cracks and fissures. Fires below them can cause rockfalls. When it's cold, ice sheets can suddenly fall from the rocks.

WHERE NOT TO CAMP

- Don't set up your shelter on sloping, poorly drained ground.
- Don't pitch your camp too close to water because of the risk of flooding and the presence of insects and animals.
- Avoid camping near noisy water sources, such as waterfalls, because they mask any other sounds that could alert you to the presence of threats, such as wild animals, or even possible rescue, such as a helicopter or emergency whistles.

CAMP TIMETABLE

In most cases, when you arrive at an area in which you intend to remain for a while (known as "going static"), start planning your activities by working back from a cut-off time—usually when it gets dark.

THREE HOURS UNTIL DARK

You arrive in the area. Drink water, stow what is no longer needed (map, compass, and so on), and change from wet clothing, but keep one set of clothes dry at all times. Scout the area for the most suitable campsite.

THEN...

Start pitching your camp. Build your shelter, including bedding. Gather tinder, kindling, and fuel and make a safe firebase and reflector. Collect water and forage for food. Prepare all aids to location (for both night and day).

ONE HOUR UNTIL DARK

During twilight, finish organizing the site and see to your personal needs—wash, use the latrine, and check your equipment. If you're in a group, make sure everyone knows where the emergency equipment is and who has been designated each task (lighting the signal fire, shining the flashlight, blowing the whistle). Set up a watch system for the fire.

STAYING CLEAN

PERSONAL HYGIENE IN THE FIELD is an important element of protection. Keeping yourself clean and healthy helps to ensure that your body is working at its most efficient and reduces the risk of illness. How you feel physically also has a direct impact on how you feel psychologically. It's a state of mind: if you let your personal hygiene slip, it's just a matter of time before everything else starts to follow.

CLEAN CLOTHING

The condition of your clothing and equipment can affect your state of mind. If your clothing is dirty and unkempt, your attitude to survival will lack discipline. Try to keep your clothing clean and in good repair.

HYGIENE CHECKLIST

Personal hygiene is about keeping clean and healthy, so develop a daily routine that ensures you correctly use personal protection aids (tablets, insect repellents, sunblock), and safely handle food, water, and cooking and eating utensils. Keep your clothes clean, and attend to your bodily functions (see opposite).

HAIR

You don't need to shampoo your hair—let it make its own oils and minerals and establish a natural balance. Wash out incidental smells, such as fire smoke, with hot or cold water.

SCALP

Check your head for insects and bites every morning and night. Local knowledge will help—for example, look for ticks if you are in a wooded area populated by deer.

EYES

Wear sunglasses or a hat to protect your eyes from bright sunlight and snow glare. Rinse your eyes with water twice a day to protect against infections like conjunctivitis.

EARS

Carefully check your ears for foreign objects with a clean, wet finger, especially if you're sleeping rough on the ground.

TEETH AND GUMS

Use clean fingers to rub your teeth and gums, or make a toothpaste from baking soda, or a mouthwash from salt and water.

BODY

Every two days wash your armpits, crotch, hands, feet, and toes with running water to keep fungal infections at bay.

MAKING A SHOWER

Standing under a shower can work wonders for your spirit—it removes accumulated dirt and sweat and makes your body feel refreshed. An improvised shower is quick and easy to make from a metal or a plastic container.

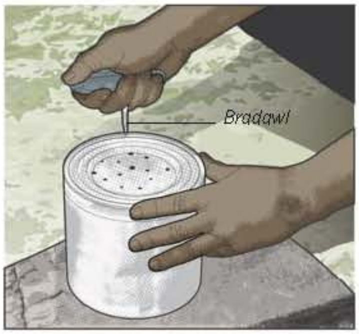

① Turn the container upside down on a flat surface.

- Punch holes in the bottom with the bradawl on your penknife or the point of your knife.

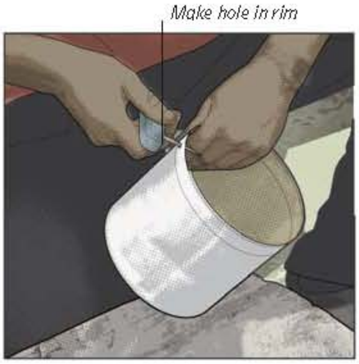

② Make a hole about 1 in (2.5 cm) down from the rim of the container.

- Make a matching hole on the opposite side of the container.

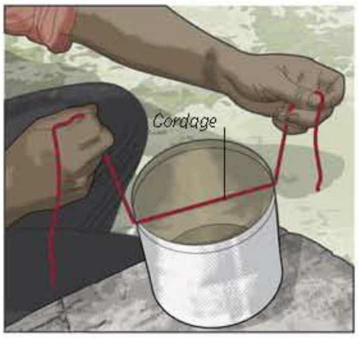

③ Smooth out any rough edges around the two holes.

- Thread a piece of cordage more than 2 ft (60 cm) long through the two holes.

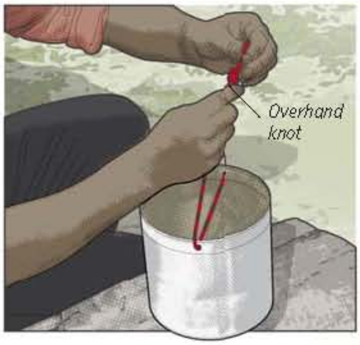

④ Pull the cord through until you have the same length—about 1 ft (30 cm)—on either side.

- Tie the two strands together with an overhand knot (see p. 143).

BODILY FUNCTIONS

In a genuine survival situation, you don't eat as much food as usual and your toilet functions—particularly with regards to solid waste—will reduce dramatically after a day or so. Nevertheless, it's important to maintain a routine, especially if you are in a group.

IN TRANSIT

If you're moving every day or so, it's not worth making a latrine, so just attend to your bodily functions as follows:

- **Urination:** choose a tree away from the water source and downwind of the camp.
- **Defecation:** dig a hole at the base of a tree, fill the hole afterward, and mark it with stones or two crossed sticks. Clean yourself (but don't use the hand you eat with) with toilet paper, leaves, or grass, then running water. Wash your hands and fingernails. Burn or bury used toilet paper.
- **Menstruation:** if you don't have tampons or sanitary napkins with you, use something cotton and washable, or even sphagnum moss. Burn or bury whatever you have used.

GOING STATIC

If you're going static by staying in one place for more than a few days, build a latrine downwind of your camp, and downhill and away from your water source. Dig a deep trench, either next to or between two trees, and make a seat from two poles. Cover the deposits with sand or soil to reduce the smell and keep flies away. When you leave, dismantle the latrine, fill the trench, and mark the area with stones or crossed sticks.

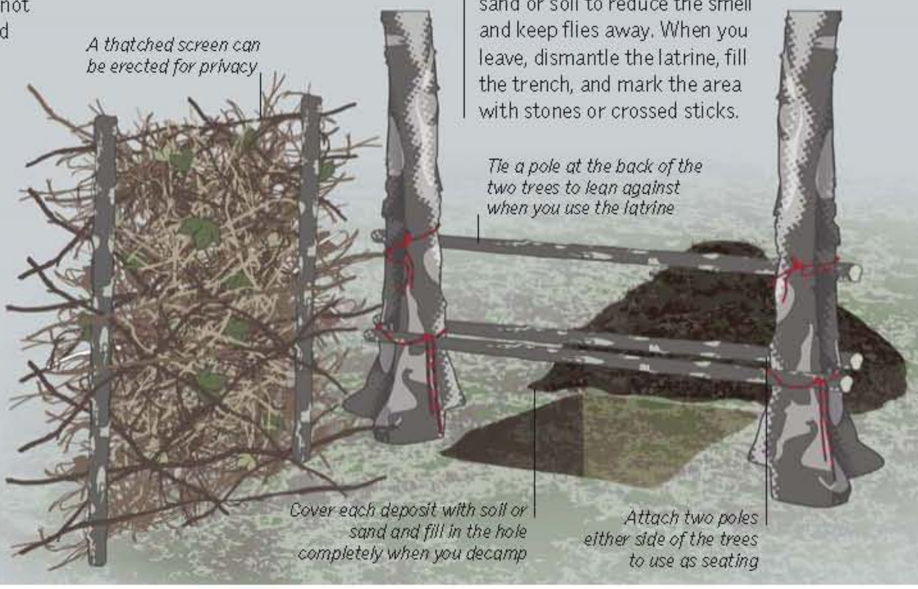

A thatched screen can be erected for privacy

Tie a pole at the back of the two trees to lean against when you use the latrine

Cover each deposit with soil or sand and fill in the hole completely when you decamp

Attach two poles either side of the trees to use as seating

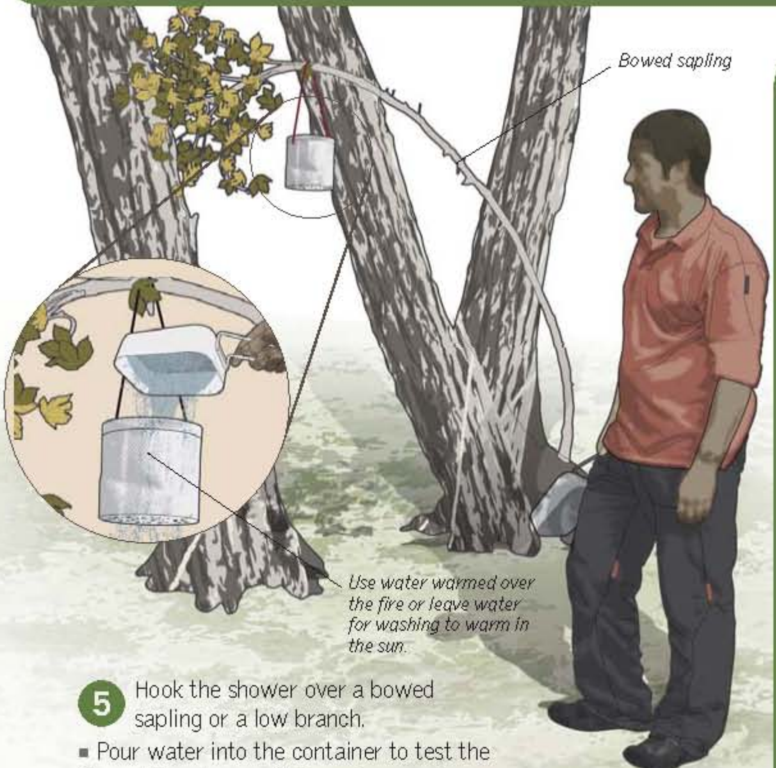

Bowed sapling

Use water warmed over the fire or leave water for washing to warm in the sun.

5 Hook the shower over a bowed sapling or a low branch.

- Pour water into the container to test the flow of the shower. If the water runs too quickly, put leaves in the bottom to slow it.

LATHERING UP

Camping soap is a concentrated antibacterial liquid soap that can be used without water. Decant a little into a small container, such as an old 35mm film container, and it will last for weeks. Alternatively, you can make soap from natural materials.

MAKING SOAP

You can make natural soap from various sources that contain a substance called saponin, which —when mixed with water—has a cleaning effect.

- **Birch leaves:** Select young leaves and place them in a container (even a plastic bag will do). Add some cold water, then add some hot or boiling water (whichever your container can handle without melting). Agitate the mixture—this allows the saponin in the leaves to dissolve into a natural soap.
- **Soapwort:** Agitate soapwort roots in water until they foam up. Let the light foam settle before using it to wash yourself and your clothes.
- **Horse chestnut:** Dip horse chestnut leaves in warm or hot water then remove. Squeeze the leaves in your hand to produce saponin.
- **Wood ash:** Mix wood ash with water. Don't use this method too often, as it can dry out your skin.

MAKING FIRE

THE ABILITY TO MAKE and maintain a fire can be a significant psychological factor between determining whether you do all you can to survive or just give up. Fire gives you a sense of "being," and, like a shelter, it can transform a clearing under a tree into a "home."

A PORTABLE KIT

It's important to gather all your materials before you start—not only your tinder, kindling, and fuel, but also your means of lighting a fire. This involves a fair amount of organization, so a portable kit that keeps it all in one place is much more convenient.

FIRE PRECAUTIONS

If you intend to have a fire in front of your shelter, bear in mind some basic rules:

- Don't build the fire too close to the shelter as it may get out of control or a spark may blow into the shelter and ignite it.
- Make sure you have fully extinguished your fire before you decamp and leave the site. Pour water (if you have enough) onto the fire and the immediate surrounding area, or use damp soil, sand, or earth.

COMPONENTS OF FIRE

Three crucial components—oxygen, heat, and fuel—need to be present in order for a fire to start and be sustained. Although you don't have any control over the resulting chemical reaction, known as combustion, you can develop the skill of lighting a fire. The key is to achieve the best balance between the components. If your fire isn't going well, get back to basics and ask yourself which one of the components is working against you.

OXYGEN

Although oxygen is vital for combustion to happen, it's all too easy to prevent it from getting to the flame by smothering the fire with too much wood. If your initial fire looks as though it's dying, try fanning it with your hand or a map to create a draft that feeds oxygen to the fire.

HEAT

Heat is essential for igniting the fuel. In most cases this heat can be generated either by a spark (such as flint and steel), by a chemical reaction and friction (such as matches or potassium permanganate), or by friction alone (such as a bow drill). (See pp. 126–27 and pp. 130–33.)

FUEL

Once the fire gets going, you need to have fuel to burn (see pp. 120–21). You should start with small, dry pieces of fuel that will catch the flame and generate enough heat to then burn increasingly large pieces of fuel.

MATCHLESS FIRESET TIN

Working on the principle of "Don't make things any harder than they need be," this matchless fireset, designed for the military, contains a one-stop solution to getting a fire going—regardless of the weather conditions.

INSIDE THE TIN

The small tin, which is waterproof when taped, contains everything you need to produce a fire: a sparking device, tinder, kindling, and fuel. Hexamine from the fuel blocks is scraped onto cotton wool, then a spark from the flint and steel ignites the cotton wool long enough for the fuel blocks to light.

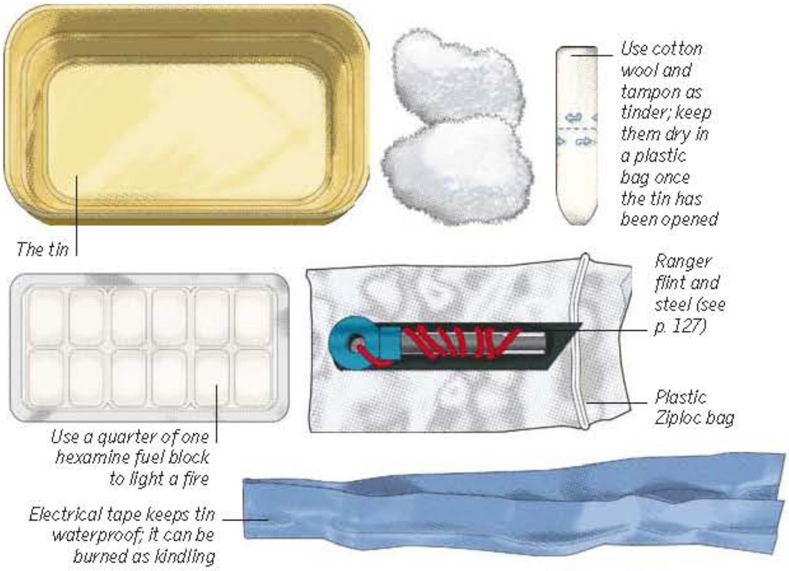
Use cotton wool and tampon as tinder; keep them dry in a plastic bag once the tin has been opened
The tin
Ranger flint and steel (see p. 127)
Plastic Ziploc bag
Use a quarter of one hexamine fuel block to light a fire
Electrical tape keeps tin waterproof; it can be burned as kindling

USING THE TIN

You can use the tin itself to boil water. Raise the tin 1 in (2.5 cm) off the ground with two sticks and light a quarter of a hexamine fuel block underneath. After a while, the water boils so you can have a potentially life-saving drink.

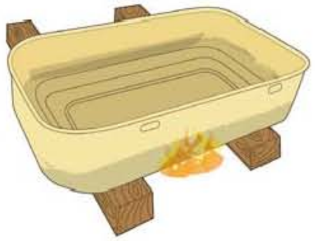

PREPARING BEFOREHAND

More than any other survival task, starting a fire needs good prior preparation. If you don't have the correct materials and in sufficient quantities, you will probably fail in your attempts to start a fire. Preparing the ground, your materials, and your equipment will usually make firelighting much easier and more likely to succeed.

CHOOSING YOUR GROUND

You need to be careful when selecting the place in which you intend to start your fire. Clear the ground before building your fire. Never light a fire directly on the ground and watch that your fire doesn't spread or burn out of control.

KEY POINTS TO REMEMBER

- Carry some form of firelighter with you at all times.
- Practice your skills before you need them—in different conditions and using different materials. You'll soon learn what works and what doesn't.
- Collect tinder as soon as you enter the forest. If it's wet, dry it in your pockets against your warm body.
- Collect everything you need to get a fire going, then multiply it by ten. If you're in a dire situation, you may have only one or two chances of getting a fire going before you either run out of matches or tinder, or you suffer from exhaustion or hypothermia.

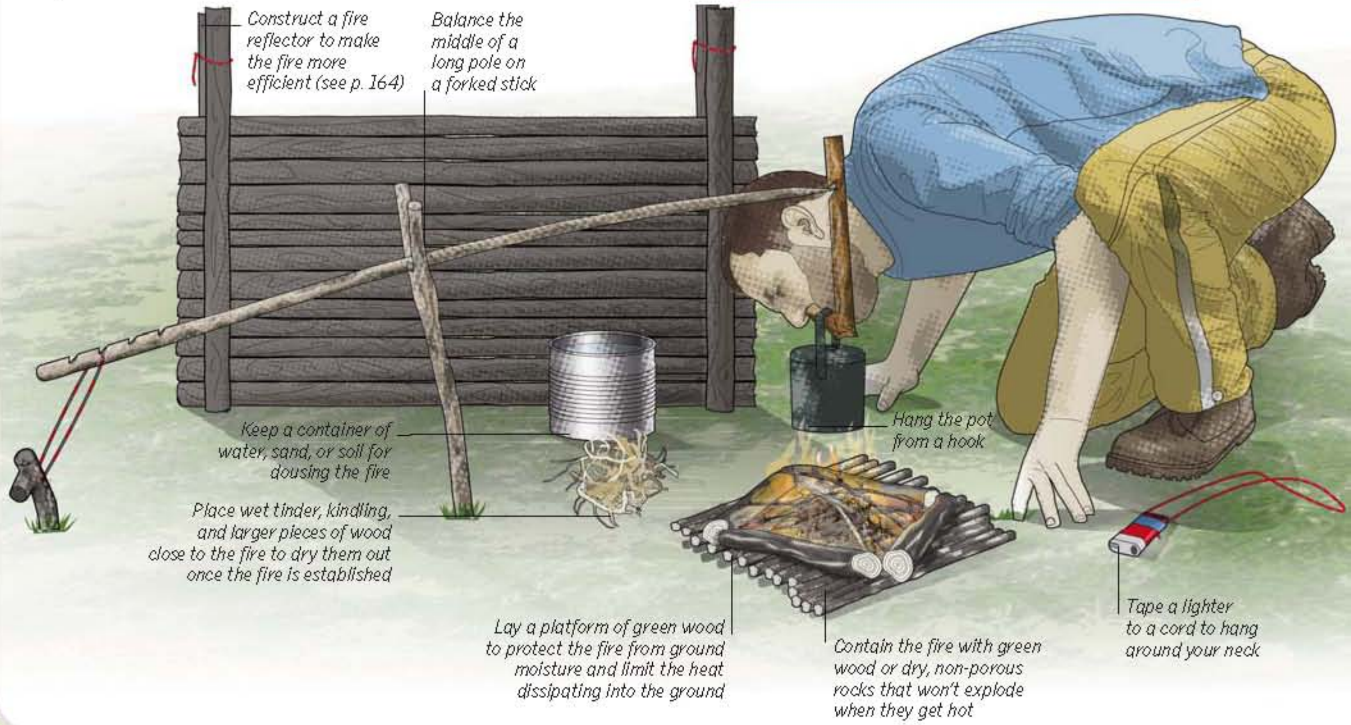
Construct a fire reflector to make the fire more efficient (see p. 164)
Balance the middle of a long pole on a forked stick
Keep a container of water, sand, or soil for dousing the fire
Hang the pot from a hook
Place wet tinder, kindling, and larger pieces of wood close to the fire to dry them out once the fire is established
Tape a lighter to a cord to hang around your neck
Lay a platform of green wood to protect the fire from ground moisture and limit the heat dissipating into the ground
Contain the fire with green wood or dry, non-porous rocks that won't explode when they get hot

CAMP FIRE DO'S

- Make sure you have enough wood close by to fuel your fire. Carrying wood to your fire uses energy, which may be at a premium in a survival situation.
- Rake the area around the spot where you intend to start the fire and brush away leaves or anything else that may ignite and start a forest fire.
- Check the ground for tree roots. Your fire could set an exposed root alight, or even one just under the surface. Once a root starts to burn underground, the heat works its way along the root and could start a forest fire.
- If you're trying to get rescued, choose a spot for the fire where passing vehicles and aircraft can see it.

CAMP FIRE DON'TS

- Don't use your hands as a rake to clear the ground, because you may get bitten by an insect or a snake. Use your feet or a branch instead.
- Don't build a fire next to old logs or fallen trees, as they may catch fire. They may not look alight but they can smolder for several days and a heavy breeze could subsequently fan the embers and start a blaze.
- Don't build your fire under an overhanging branch, or leaves, as the heat from the fire can soon dry them out and then set them alight.
- Don't position your fire so that the wind affects the way the fire burns, or blows smoke or flames into your shelter.

THE ELEMENTS OF FIRE

THE THREE MATERIAL ELEMENTS you need to build a fire are tinder, kindling, and fuel. They must be dry and collected in sufficient quantities. A well-made feather stick effectively provides tinder, kindling, and fuel on one piece of wood—and it needs only a spark to set it alight.

EMERGENCY TINDER

Carry tinder, such as a tampon or cotton wool balls, in a 35mm film canister or a Ziploc bag in your emergency kit. For each fire, use some cotton wool and a small piece of tampon. Cotton wool balls smeared with petroleum jelly make a flame last ten times longer.

MAKING A FEATHER STICK

Between four and six well-made feather sticks provide enough combustible material to get a fire going. Keep a few already made and packed away to use in an emergency. The best feather sticks are made from standing dead wood, but just about any type of dry wood will do. If you use small, dead branches that have snapped off a tree, remove the bark first.

1 Choose the side of a stick with an even grain and no knots. Work on a hard, flat surface to stop the stick slipping.

- Lay the blade flat on the stick and run it all the way down without cutting into the wood—do this ten times. This gives you a feel for the wood and how your knife moves over it.

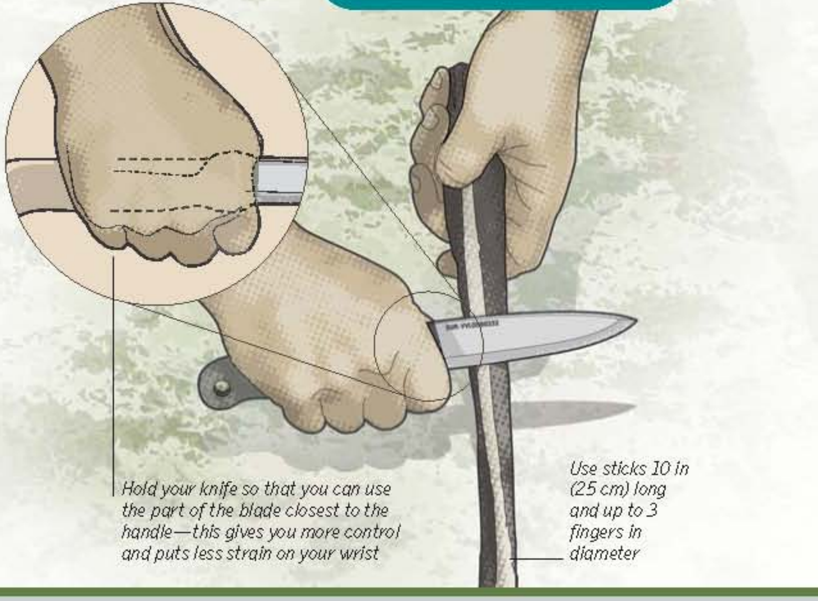

Hold your knife so that you can use the part of the blade closest to the handle—this gives you more control and puts less strain on your wrist

Use sticks 10 in (25 cm) long and up to 3 fingers in diameter

TINDER

The first element you use to make a fire is dry, combustible material called tinder. You may have tinder in your equipment (see pp. 122–23), or you may need to find natural or other man-made sources. The key to success is to experiment with what's around you before you actually need it. Make sure it's dry—leave it in the sun if it's damp. When you use some tinder, replace it at the earliest opportunity.

TYPES OF TINDER

Natural sources: Feather sticks (see above), shavings from the outside of a bamboo stem, bark shavings, fine wood dust, pine pitch, plant and animal down, powdered dung, and fire dog (a charred stick). Also, birch bark (silver birch bark can be lit even when it's wet), clematis, honeysuckle, cattail, dry grass, dead and dry moss, and some fungi. The best tinder for friction methods of firelighting (see pp. 130–33) is tree bark, dry grass, fungi, lichen, plant roots, fibers, and down.

Man-made sources: Cotton wool balls, tampons, char cloth, fire tin (see p. 123), lint, tissue paper, camera film, strips of rubber, and candle wick.

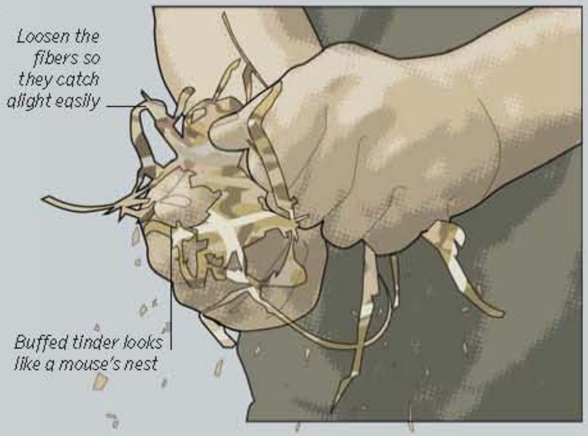

Loosen the fibers so they catch alight easily

Buffed tinder looks like a mouse's nest

TINDER BUNDLE

A good way to prepare your tinder to take a spark or a coal is to make it into a tinder bundle. Vigorously tease, rub, and pull the fibers with your fingers until it becomes a ball the size of a grapefruit. Push the finest, most combustible material into the interior of the ball. Try mixing up your tinder to make it more effective—for example, dry grass, birch bark, and cattail.

STANDING DEADWOOD

A tree that's died but hasn't fallen over has no green foliage and the bark falls off without being replaced. This is the ultimate wood for firelighting, as it can provide kindling and fuel and, when split, makes excellent feather sticks. The bottom few inches may be wet where the tree wicks up moisture from the ground.

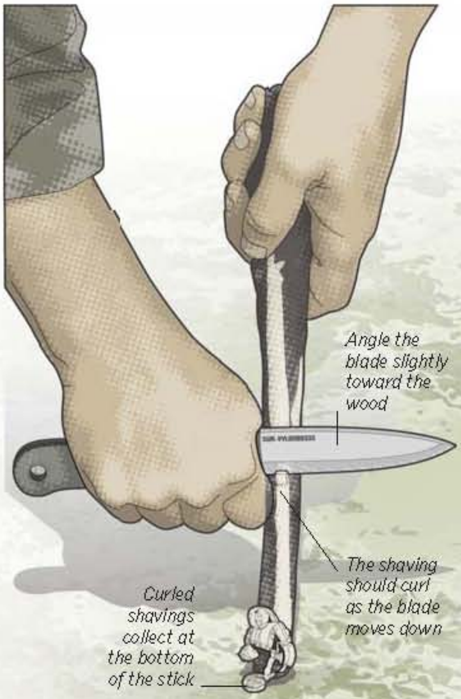
Angle the blade slightly toward the wood
The shaving should curl as the blade moves down
Curled shavings collect at the bottom of the stick

2 Now tilt the angle of the blade slightly toward the wood.

- Keeping a part at the top of the wood to use as a handle, run the blade down the wood to cut a shaving.
- Stop just before the bottom so that the shaving stays attached to the stick.

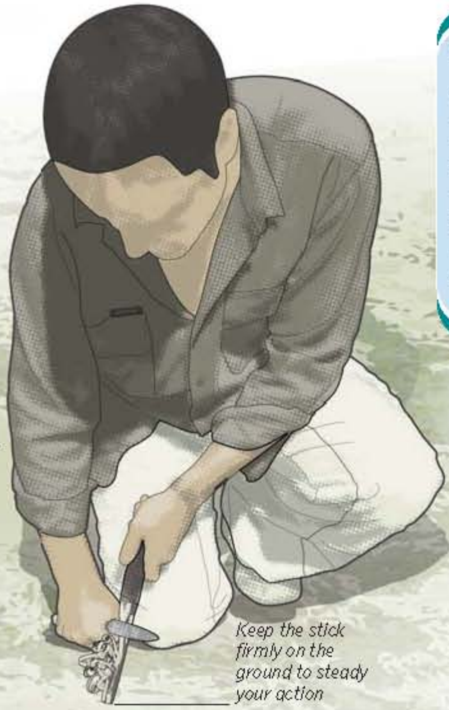
Keep the stick firmly on the ground to steady your action

3 Turn the stick slightly and run the knife down the edge created in step 2 to create a second shaving. Turn the stick again and repeat, working around the stick.

- Try to get a rhythm going, and put your body behind the cutting action.

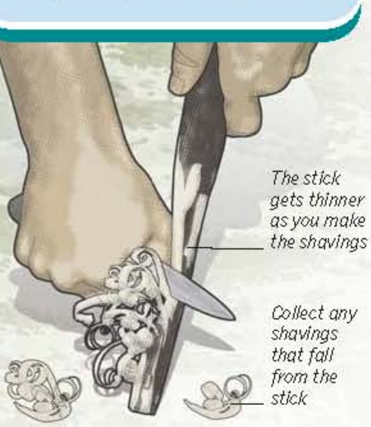
The stick gets thinner as you make the shavings
Collect any shavings that fall from the stick

4 Once you have mastered the technique, continue to feather the rest of the stick.

- When you've finished you will have a thin stick with curled shavings still attached, and ready for use as kindling.

KINDLING

Kindling is the second material element of firelighting and, when dry, is added to burning tinder. Kindling can be as thin as a match or as thick as a finger. You should be able to snap it with your hands. If it doesn't break with a crack then it's probably not completely dry. If it's damp, remove the outer bark (this retains most of the moisture) and break the kindling down into small sticks 6 in (15 cm) long.

TYPES OF KINDLING

- Soft-wood twigs are very combustible, while wood with flammable resin, such as pine, burns hot and long.
- You can use some types of tinder as kindling—for example, bark, palm leaves, pine needles, grass, ground lichens, and ferns—but you will need larger quantities for kindling than you needed for tinder.

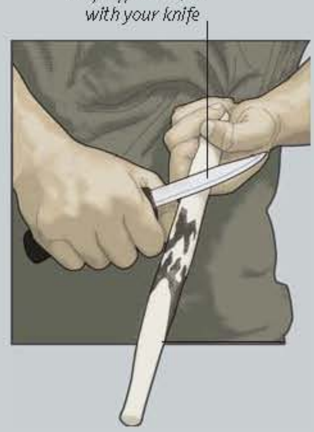
Strip off the bark with your knife

FUEL

Initially, your fire needs constant tending, but once it can sustain itself for five minutes it's established and you can add increasingly larger fuel to create a good heart—a bed of hot coals that sustains the fire with minimum effort on your part. The fuel should be about as thick as your wrist or forearm. At first, use dry fuel split into sticks that catch alight easily. Add green (live) wood and larger logs (whole or split) later, once when the fire is established. In wet conditions, build your fire under cover to keep off the rain.

TYPES OF FUEL

- Hard woods from mainly deciduous trees (such as oak, maple, ash, beech, and birch) burn hot and long, produce good coals, and are more efficient. But they are hard to get going.
- Soft woods from conifers (such as pine, fir, and spruce) are easier to light as they can contain resin, but burn faster and produce less heat than hard woods. They also make more smoke.
- Peat is found on well-drained moorland and can be cut with a knife. However, it needs a good air supply.
- Charcoal is lightweight, smokeless, and burns hot.
- Dried animal droppings provide a good smokeless fuel.

CHAR CLOTH AND FIRE CAN

BEFORE SETTING OFF ON A TRIP, add either some char cloth or a fire can to your backpack. These excellent forms of tinder are reliable and potentially life-saving. Alternatively, you can take some mayasticks. Char cloth is very easy to make, so keep some dry in your emergency fire kit or lining the bottom of your survival kit (see pp. 60–61).

TOOLS AND MATERIALS

- A can with a tight-fitting lid, such as a small shoe-polish can
- Nail
- 100 percent cotton cloth
- Knife or scissors
- Spark or flame

MAKING CHAR CLOTH

Char cloth is cotton cloth that's been combusted in the absence of oxygen (pyrolysis). It's lightweight, takes up almost no space, and produces an ember from even a weak spark extremely well. Char cloth works only when it's completely dry, so keep it in a watertight container.

1 Turn the lid of a can upside-down and hammer a nail into the center.

- Experiment with the size of the hole. Usually, the smaller the hole the better.

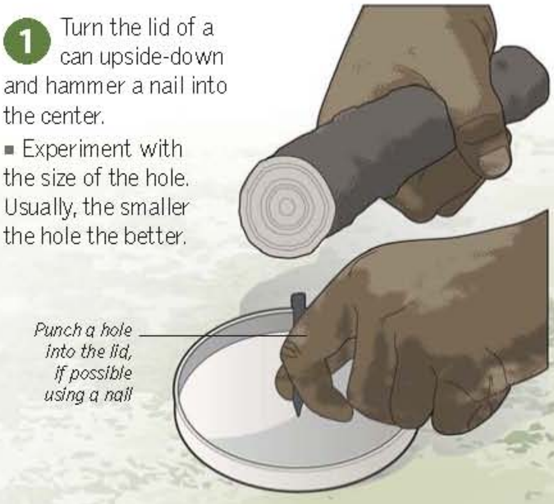
Punch a hole into the lid, if possible using a nail

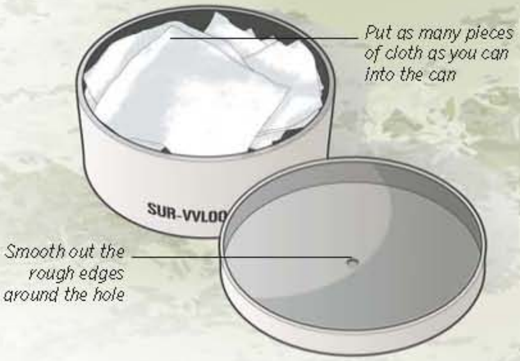
Put as many pieces of cloth as you can into the can
SUR-VVLOO
Smooth out the rough edges around the hole

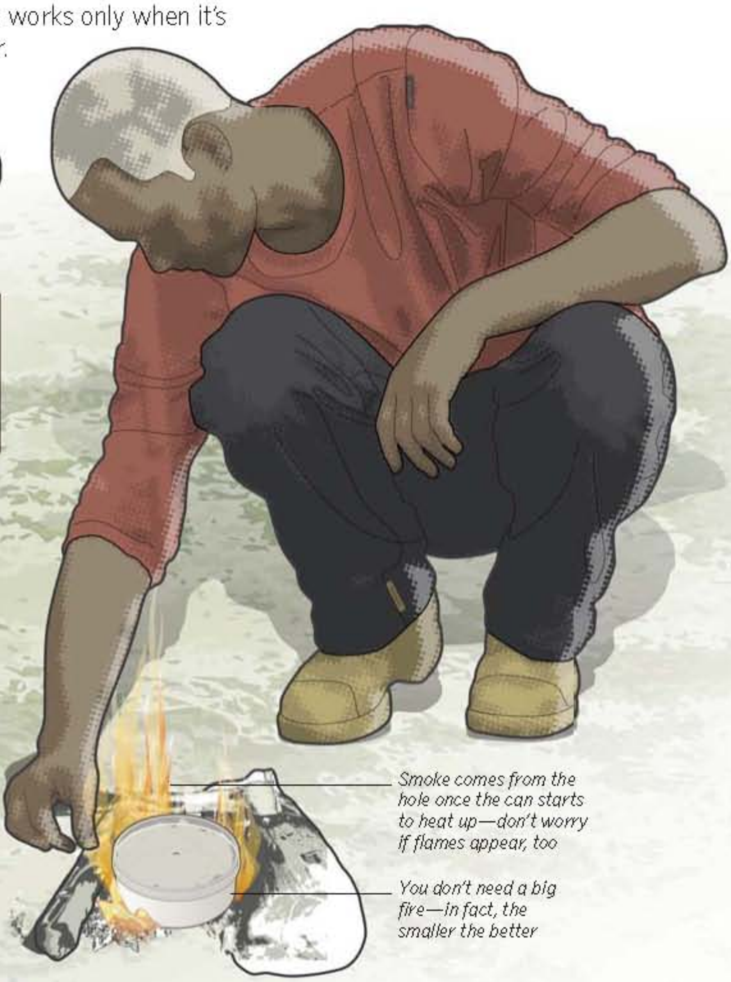
Smoke comes from the hole once the can starts to heat up—don't worry if flames appear, too
You don't need a big fire—in fact, the smaller the better

2 Cut 100 percent cotton cloth into pieces that fit into the can without folding.

- Vary the sizes of the pieces so they don't just stack flatly on top of each other, but at the same time don't squash them in.
- Securely place the lid on the can.

3 Place the can on a fire to burn off all the oxygen inside the can, or place it on top of some good coals scraped to the side of your camp fire.

- When the smoke stops the process is complete.
- Safely remove the can from the fire.
- Don't open the can until it's cool.

MAKING A FIRE CAN

You can use a fire can to start a stubborn fire when conditions are less than perfect, or to boil water, do some basic cooking, or warm your hands on a cold day. Once lit, a fire can burns for hours with a concentrated, controllable flame that produces no smoke. When it starts to fail, you can either replenish the existing cardboard with more wax or replace the cardboard and start again.

TOOLS AND MATERIALS

- A can with a tight-fitting lid, such as a small shoe-polish tin
- Cardboard (ribbed or plain)
- Candle and a match or lighter
- Knife or scissors

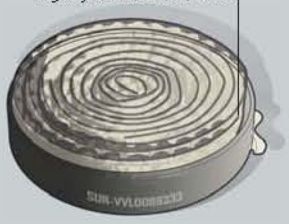

The extra ⅙ in (4 mm) of cardboard will burn down slightly and act as a wick

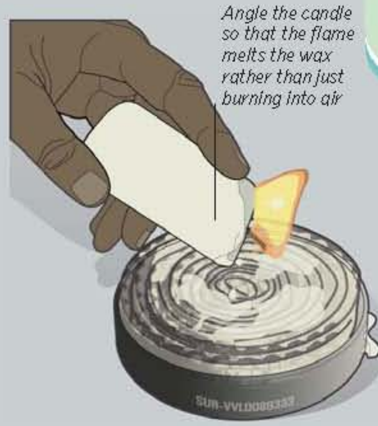

Angle the candle so that the flame melts the wax rather than just burning into air

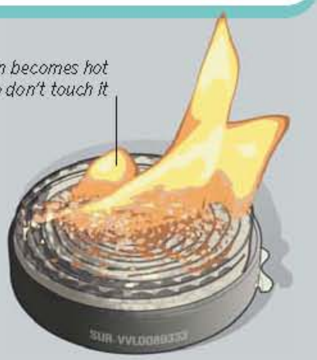

The can becomes hot so don't touch it

1 Cut out a long, thin piece of cardboard that's ⅙ in (4 mm) wider than the depth of your can.

- If it's ribbed, cut it across the grain.
- Roll it tightly along its length until the roll just fits inside the can.

2 Light your candle and let the melting wax drop into the can.

- Let the wax soak the cardboard and fill up the can—it's a slow process.
- Stop when the wax nears the top.
- Allow the wax to harden.

3 When the can has cooled down, hold it at an angle and light the top of the cardboard with the candle.

- The flame should be concentrated and spread across the top of the can.

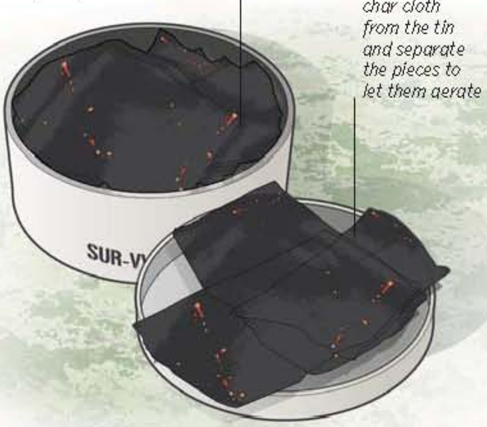

Look at the colour and texture of the char cloth to see if the process has worked

Remove the char cloth from the tin and separate the pieces to let them aerate

4 Once the tin has cooled down, remove the lid and examine the charred cloth.

- The cloth should be completely black and have a semi-rigid, but softish texture. If it's fawn or brown, put it back on the fire and leave for longer.
- Cloth that's brittle and crumbles is of no use.
- Check the cloth works by striking a spark on to a piece; the spark should create a small red ember.

FIRE WADS AND MAYASTICKS

Fire wads are rolled strips of newspaper soaked in a fuel. Once dry, they make handy waterproof tinder. Mayasticks are pieces of resinous fatwood from the highlands of Mexico and Guatemala. The pieces are easy to light, even when wet, and generate a hot flame.

MAKING FIRE WADS AND MAYASTICKS

- To make fire wads, tightly roll strips of newspaper into a tube 4 in (10 cm) long and tie them with cord. Soak them in melted wax and allow to dry. Light the center with a match.
- To light a mayastick, cut a little sliver of wood with your knife and lift it from the surface of the stick. When a little resin seeps out, set it alight with a flint and steel (see p. 127).

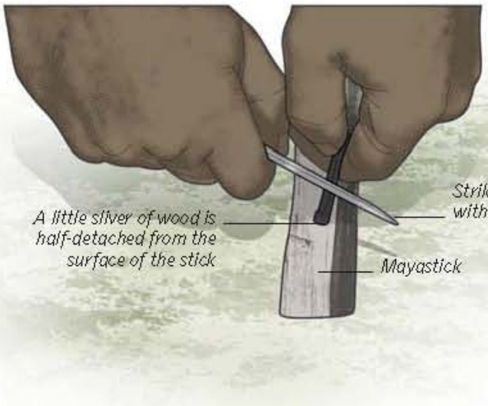

A little sliver of wood is half-detached from the surface of the stick

Strike the flint with your knife

Mayastick

TYPES OF FIRE

ONCE YOU HAVE COLLECTED your tinder, kindling, and fuel, and made sure it's dry and ready at hand, you can set about making your fire. There are many different types of fire to choose from, depending on what you need the fire for (see below). Before you start to build your fire, make sure you have prepared the ground so that everything is safe (see pp. 118–19).

CHOOSING A FIRE

If you have a choice of fuel, as well as time to invest in building a specific type of fire, look at the choices below and determine which one is likely to suit your immediate needs.

- Your primary consideration should be the function of the fire: warmth is probably your most urgent requirement, but other uses include cooking, signaling (see pp. 238–39), drying wet clothing, and disposing of waste. You may want a particular type of fire because it lasts all night.
- Consider the availability of the components you need to build the fire—for example, the right fuel and the best ground. A rule of thumb is to estimate how much you think you'll need, then double it.

FIRE ESSENTIALS

There's plenty of advice surrounding the pros and cons of making fires, but three general tips will help:

- Never make things any harder than they need to be. Choose a fire that requires the least effort for the maximum gain.
- It is more efficient to build a small fire and sit close to it, than to build a large fire and sit far away from it.
- If all your wood appears wet, remove the bark and split the wood—the center will usually be dry.
- Once the fire is established, place damp tinder, kindling, and larger pieces of wood close to the fire to dry out. Keep watch on them so that they don't dry out to the point where they catch fire!

TYPE OF FIRE	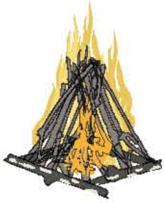 **TEEPEE**	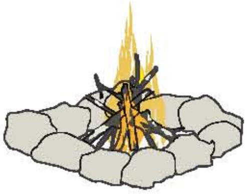 **STONE-LINED**	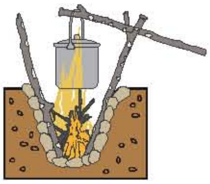 **AUTOMATIC**	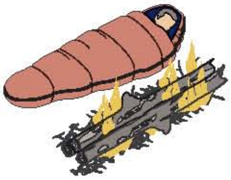 **LONG LOG**
HOW	▪ Surround the tinder ball with kindling in the shape of a teepee. Arrange small, medium, and large fuel logs in a square at the foot of the teepee.	▪ Arrange large non-porous stones in a circle, place a tinder ball in the middle, with kindling around it. Add fuel logs when the fire is established.	▪ Line a hole 3 ft (1 m) deep with non-porous stones. Put tinder and kindling inside and rest large logs against the sides so they drop down as they burn.	▪ Put tinder, kindling, and fuel logs in a depression 6 ft (2 m) long. Lay two long fuel logs on top of the burning fire.
WHERE	▪ Areas with abundant fuel, because the fire requires lots of firewood	▪ Windy locations ▪ Popular areas with existing fire rings and where low-impact camping is vital	▪ Rock-free earth or sand, because it's easy to dig	▪ Forested areas, because the fire requires long logs for fuel
WHY	▪ Quick to light ▪ Wet wood can be burned because it's dried by the heat of the inner fire	▪ Stones shelter fire from wind ▪ Using an existing fire ring reduces the fire's impact on the environment	▪ Self-feeding once lit, which means you don't need to keep adding fuel	▪ Long-lasting (the fire can stay alight all night) ▪ Emits a great deal of heat
USES	▪ Warmth ▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Warmth ▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Cooking ▪ Signaling (if green vegetation is burned)	▪ Warmth (in cold weather, build a fire either side of you). ▪ Cooking (once embers have formed)

LIGHT YOUR FIRE

There are numerous ways to build a fire and get one going, and everyone has their favorite. The following example is a tried-and-trusted method that's versatile and works well in a variety of conditions.

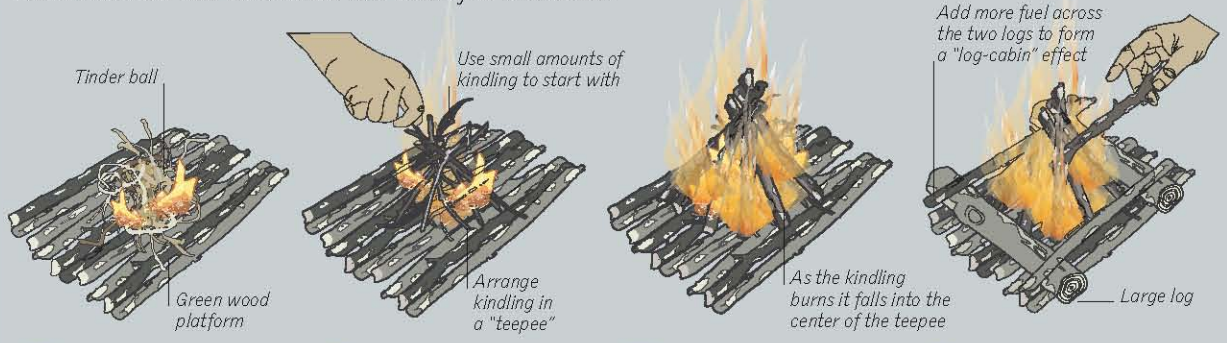
Tinder ball
Green wood platform
Use small amounts of kindling to start with
Arrange kindling in a "teepee"
As the kindling burns it falls into the center of the teepee
Add more fuel across the two logs to form a "log-cabin" effect
Large log

1. Place your tinder ball (see p. 120) on a platform composed of green wood.
   - Light the tinder with your chosen method (see pp. 126–27) and let it catch.
2. Gently lay kindling by the tinder ball.
   - Build a kindling "teepee", as this lets the fire breathe where the heat is strongest.
   - Feather sticks also make ideal kindling (see p. 120).
3. As the kindling catches fire and the flames grow, add larger pieces of kindling.
   - Continue to add kindling, gradually building up to split logs (see p. 148).
4. The fire is established when you can leave it alone for five minutes without it going out.
   - Then lay a large log on the wind-facing side and one on the other side.

TYPE OF FIRE	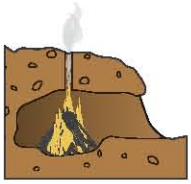 **SNAKE HOLE**	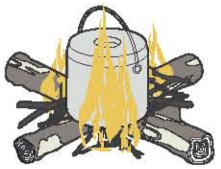 **STAR-SHAPED**	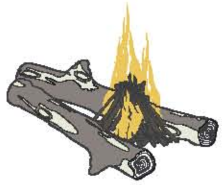 **HUNTER**	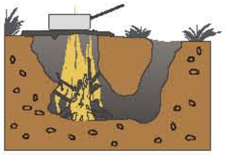 **DAKOTA HOLE**
HOW	▪ Create a hole in the side of a bank and a chimney up through the earth. Light a fire inside using any fuel.	▪ Build a fire from tinder, kindling, and fuel. ▪ Arrange four logs so they meet in the middle. ▪ Push logs in as they burn.	▪ Make a fire out of tinder, kindling, and any type of fuel. Place two long logs either side of the fire in a V-shape.	▪ Dig a large hole for the fire and a slightly smaller hole for the chimney, with a tunnel linking the two. Use small logs as fuel and cook at ground level.
WHERE	▪ Windy locations ▪ Rock-free earth or sand bank	▪ Forested areas, because of the logs the fire needs	▪ Cold or windy locations	▪ Anywhere you can dig a hole, because the fire requires very little fuel once established
WHY	▪ Chimney creates a draft, giving a high-temperature fire ▪ Sheltered from the elements	▪ Long-lasting ▪ Good embers for cooking	▪ Hardwood logs shelter the fire from the wind ▪ Produces a great deal of heat	▪ Concentrated heat ▪ Flames are below ground so fire is hidden
USES	▪ Warmth ▪ Cooking ▪ Waste disposal	▪ Cooking—balance a pot on the logs ▪ Heating water	▪ Warmth ▪ Cooking	▪ Cooking ▪ Keeping warm ▪ Drying wet clothing

MAKING SPARKS AND FLAMES

LIGHTING TINDER IS THE FIRST STAGE of making a fire. Matches and lighters do this instantly, but there are a number of ways to create a spark that you can then use to coax your tinder into a flame. This is relatively easy when everything is dry, but with patience and persistence it is also possible when conditions are bad.

IGNITION DEVICES

There are various ingenious methods of creating that all-important spark or flame to ignite your fire. If you don't have matches or a lighter, you'll need to use another device, such as a flint and steel. You can also improvise by using an external energy source, such as:

- Focusing the sun's heat with a magnifying glass or beverage can.
- Making sparks with a battery.
- Creating a chemical reaction using potassium permanganate.

STRIKING A MATCH

It may seem simple enough, but there's a way of striking a match—known as "commando style"—that reliably produces a flame in all kinds of conditions.

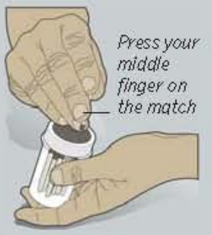
Press your middle finger on the match

① Hold the box in one hand and a match between the thumb and first two fingers of the other.

- Strike the match firmly away from you.

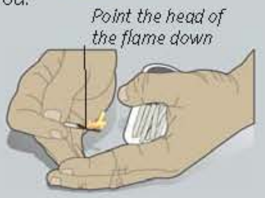
Point the head of the flame down

② When the match lights, immediately cup your hands to protect the flame.

- Let the flame burn a little way along the stem before using it to light the tinder.

TYPE OF DEVICE	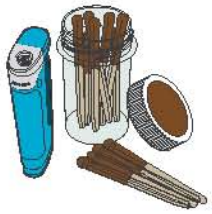 **MATCHES/LIGHTERS**	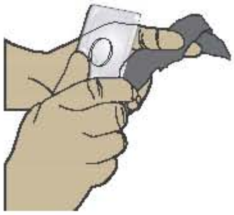 **MAGNIFYING GLASS**	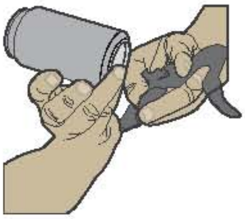 **BEVERAGE CAN**	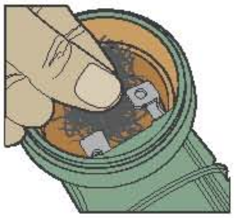 **BATTERY**
GEAR	▪ Waterproof box of matches. ▪ Cigarette lighter taped to a piece of cord around your neck. ▪ Dry tinder	▪ Magnifying glass ▪ Dry tinder	▪ Empty beverage can ▪ Dry tinder	▪ Flashlight and its battery ▪ Wire or wire wool ▪ Dry tinder
WHAT TO DO	▪ To light a match "commando style," strike it away from you on the box and then cup it in your hands (see above).	▪ With the magnifying glass, focus bright sunlight onto some dry tinder and create a hot spot. Hold the magnifying glass steady until the tinder catches alight.	▪ Polish the bottom of the can (see Signal mirror, p. 245). ▪ Catch bright sunlight on the shiny surface and reflect it onto tinder to create a hot spot. Hold the can steady until the tinder catches alight.	▪ Lay the wire across the battery terminals to create some sparks. ▪ Remove the bulb from the flashlight and place wire wool over the terminal. Switch on the flashlight to create sparks.
COMMENTS	▪ Waterproof matches are usually just standard matches that have been coated with wax and varnish. ▪ Always have a lighter around your neck on a piece of cord.	▪ When you choose a compass, make sure it has a magnifying glass incorporated into it (for reading the details on maps). ▪ You can also use the lens in reading spectacles.	▪ You'll need to practice this technique so that you can rely on it in a survival situation.	▪ The thinner the wire the better this will work, especially with lower voltage batteries (1.5v). ▪ Use this method for a short period only, otherwise you will drain the battery.

CREATING A CHEMICAL FLAME

Potassium permanganate is an extremely useful item to have in your survival kit, because you can use the chemical not only to sterilize water and wounds, but also to create a spark to light a fire. You will need some sugar to make it work, so use some from your survival rations or crush a hard candy.

WARNING!

Potassium permanganate is a strong oxidizer that can, when mixed with certain chemicals, create an explosive mixture. It can also stain your clothes and skin.

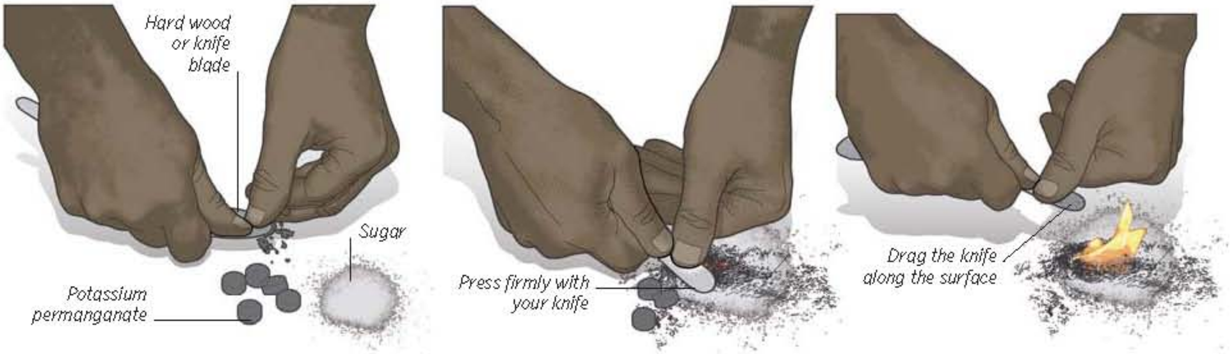
Hard wood or knife blade
Sugar
Potassium permanganate
Press firmly with your knife
Drag the knife along the surface

① On a hard surface, such as a flat piece of wood or rock, crush some potassium permanganate. Have an equal amount of sugar ready.

② Mix the ingredients. With the back of your knife, push down on the mixture and drag it along the hard surface to create friction.

③ Continue to push down and drag your blade until you get a spark.

- The flame dies down quickly so have your tinder at the ready.

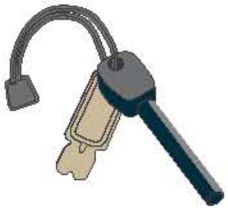 FIRESTEEL	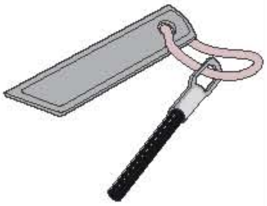 FLINT AND STEEL	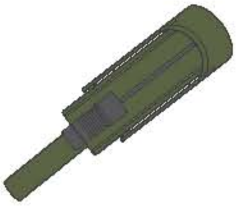 ONE-HANDED STRIKER	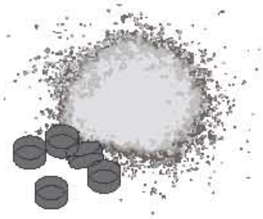 POTASSIUM PERMANGANATE	TYPE OF DEVICE
▪ Ferrocerium firesteel ▪ Metal striker ▪ Dry tinder	▪ Magnesium-alloy rod ▪ Steel striker (usually a short piece of hacksaw blade) ▪ Dry tinder	▪ A magnesium-alloy rod and steel device that can be operated using only one hand ▪ Dry tinder	▪ Potassium permanganate ▪ Sugar ▪ Knife ▪ Dry tinder	GEAR
▪ Hold striker next to the tinder. ▪ Place the firesteel under the striker, directly onto the tinder. ▪ Draw the firesteel back across the striker to direct the sparks onto the tinder. The firesteel moves, not the striker.	▪ Follow instructions for the firesteel (left).	▪ Place the flint rod in the tinder. ▪ Press the thumb button onto the rod and push the handle down the length of the rod. ▪ Pushing harder creates greater friction, which leads to a more intense spark.	▪ Mix potassium permanganate and sugar in equal amounts on a hard surface. ▪ Press down on the mixture with your knife and drag the blade along to create a spark (see above).	WHAT TO DO
▪ Lasts for about 12,000 strikes. ▪ Temperature reaches 3,000°C. ▪ Works in all weathers and at any altitude.	▪ There are many variations and kits; one has a magnesium block for the holder and the striker shaves slivers of magnesium into the tinder to help with ignition.	▪ Works in all weathers. ▪ Has a safety feature that prevents accidental use. ▪ Designed for fighter pilots, who may injure an arm or hand after ejecting.	▪ You can also mix a small amount of potassium permanganate with glycol or antifreeze. Quickly wrap in paper and put on the ground. Warning: stand well clear as the combustion can be sudden and dangerous.	COMMENTS

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE DESERT

USEFUL EQUIPMENT

- Personal Locator Beacon (PLB)
- Sunglasses and sunscreen
- Cotton scarf or hat
- Signal mirror and signal flares
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

DESPITE A LACK OF SURVIVAL KNOWLEDGE, a party of five survived searing temperatures, dehydration, hunger, and the threat of animal attack for six days when their plane crashed in the Kalahari Desert in Botswana.

On Wednesday, March 1, 2000, Carl du Plessis and three associates—Mike and Lynette Nikolic, and Nebojea Graorac—set off on a business flight from the capital of Botswana, Gaborone, bound for Maun, a popular tourist town in the north of the country. However, the aircraft developed engine trouble during the flight and was forced to crash-land. Critically, radio contact was lost, so no distress call was emitted before the plane went down. The pilot and all four passengers on board suffered burns during the crash and subsequent fire, but two sustained even more serious injuries—Lynette Nikolic damaged her spine and left arm, and Graorac suffered a punctured lung.

"CRITICALLY, RADIO CONTACT WAS LOST, SO NO DISTRESS CALL WAS EMITTED BEFORE THE PLANE WENT DOWN"

The following morning, Du Plessis and pilot Costa Marcandonatos decided to walk the estimated four days to Maun to find help. Their plan was misguided from the start—Maun was actually around some 190 miles (300 km) to the west, and other settlements to the east were much closer. They attempted to navigate by the sun but, lacking a machete, couldn't penetrate the dense bush. They were forced to follow elephant tracks to find watering holes, meeting elephants on one occasion, and heard the roar of lions at night. By Saturday they had walked 56 miles (90 km), but their zig-zag route had taken them only 20 miles (30 km) from the crash site.

Despite their mistakes, they stumbled upon a manned hunting lodge the following day. After their initial efforts to radio the authorities had failed, a helicopter finally reached the crash site on Monday, March 6, airlifting the casualties to hospital for emergency treatment six days after the crash.

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Sun/Heat—Find or improvise immediate shelter
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION
See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers either a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to
- Leave clear indications of your intent (written messages or signs) if abandoning a vehicle
- Have aids to location accessible while moving and deployed when static
- Protect yourself against glare from the sun and windburn
- Seek or improvise shelter when not moving and seize all opportunities to collect fuel for a fire—deserts get cold at night
- Be on constant lookout for signs of water or civilization—such as green vegetation, converging animal/human tracks, or circling birds

DON'T

- Sit directly on the hot ground when you stop
- Travel during the hottest part of the day
- Take risks—a twisted ankle caused by running down a sand dune could be fatal
- Force the pace—travel at the speed of the slowest person in the group
- Shelter in dry river beds, because of the potential risk of flash floods

YOU SHOULD STAY **

DO

- Seek or improvise shade and work during the coolest part of the day/night. A shelter dug even 6 in (15 cm) below ground level will provide a much cooler place to rest
- Ration sweat, not your water; you require around $1\frac{3}{4}$ pints (1 liter) per hour in temps above 100°F (38°C), half of this if it's cooler
- Continually re-assess your water situation and options for augmenting your supplies
- Leave everything out overnight that could collect dew so you can drink it
- Prepare all of your aids to location for immediate use. Be constantly alert for signs of rescue

DON'T

- Leave a broken down vehicle unless absolutely necessary—it's easier to see and is what rescuers will be looking for
- Waste energy unnecessarily—be idle unless there is something that requires doing
- Ignore your fire—use anything that burns (wood, vehicle tires) to generate heat and smoke

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.
** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

FIRE BY FRICTION

THIS METHOD USES PRIMITIVE EQUIPMENT, such as a bow drill, and requires knowledge, skill, practice, the correct wood, time, and effort. However, once learned it's possibly the most rewarding of all survival skills. Your first fire created from a coal produced by a fireset that you have made yourself will be a fire you will never forget. In a genuine survival situation, however, you would always start a fire with the quickest and easiest method available.

TOOLS AND MATERIALS

- Cordage such as para cord, fishing line, shoelaces, or bailing twine
- Wood
- Knife
- Bradawl (optional)

THE FOUR SKILLS

There are many ways to make fire by friction, including a hand drill, bamboo fire saw (see p. 133), and bow drill. Each has its own advantages, depending on the materials available and your practical skills. Regardless of the method, however, there are four main elements to producing fire by friction—each a skill on its own:

- Identifying and procuring the correct types of wood.
- Manufacturing the individual pieces that make the fireset.
- Using the fireset to produce a coal.
- Nurturing the coal into a flame.

MAKING A BOW DRILL

The bow drill set is one of the most efficient methods of making fire by friction. Try to find the right type of wood for each part, especially the drill and hearth board (see Testing Wood, opposite).

1. To make the bow, cut a stick to the same length as the distance between your armpit and your fingertip.

- Cut a notch (or make a hole with a bradawl) near each end.
- Tie cord with an arbor knot (see pp. 142–43) to one notch and with half hitches (see pp. 144–45) to the other. Leave only enough slack to wrap around your finger.

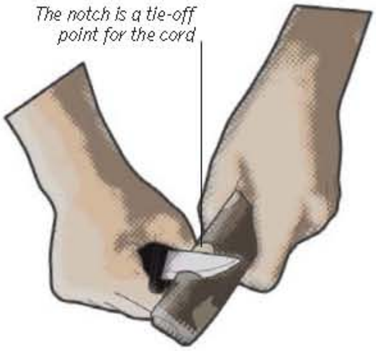
The notch is a tie-off point for the cord

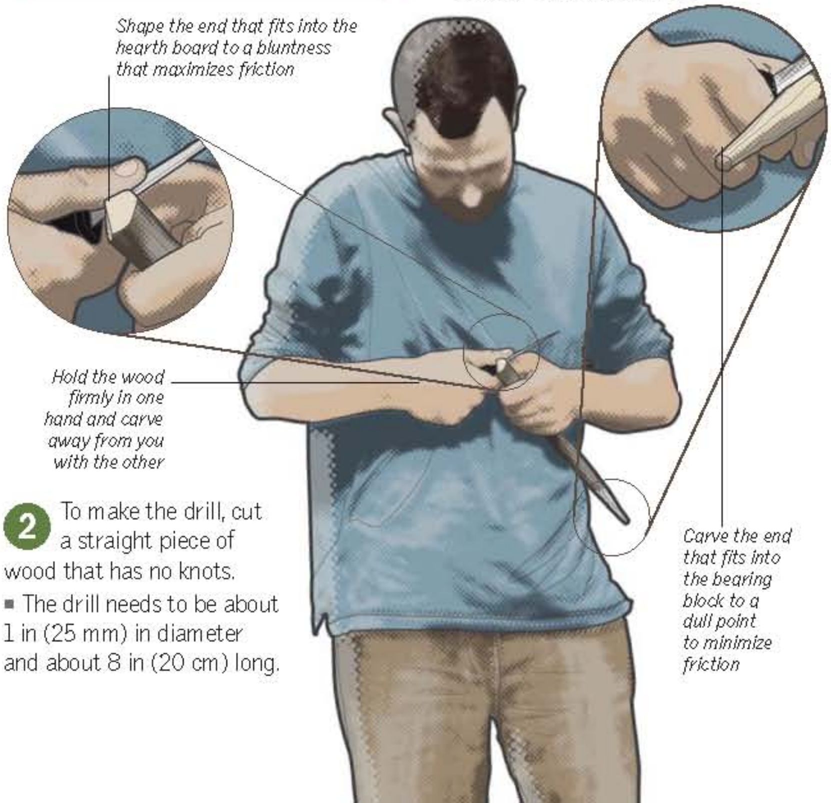
Shape the end that fits into the hearth board to a bluntness that maximizes friction
Hold the wood firmly in one hand and carve away from you with the other
Carve the end that fits into the bearing block to a dull point to minimize friction

2. To make the drill, cut a straight piece of wood that has no knots.

- The drill needs to be about 1 in (25 mm) in diameter and about 8 in (20 cm) long.

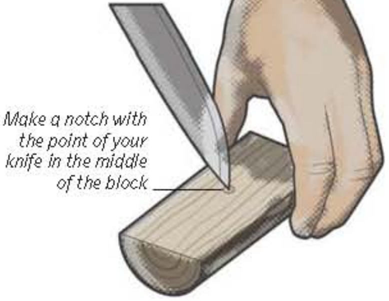
Make a notch with the point of your knife in the middle of the block

3. Make the bearing block from a piece of hard wood about 3–4 in (7.5–10 cm) across.

- Cut it to a length of 4–5 in (12–15 cm), split it in two, and make a notch in it.
- The drill's dull point fits into the notch, while you hold the block when you are bowing (see p. 132).

TESTING WOOD

Standing deadwood is ideal for the drill and hearth board as it's dry. Soft deadwood is easier to use than hard deadwood, but if you can't tell the difference, use these tests.

THUMBNAIL TEST

Cut through the bark of a branch, exposing a patch of wood beneath. If you dent this wood easily with your thumbnail, it's a soft wood. If you can only make a small impression, then it's a hard wood.

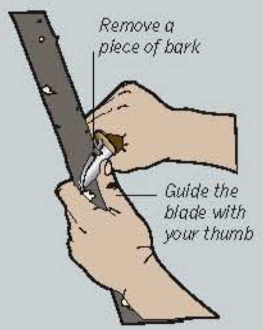
Remove a piece of bark
Guide the blade with your thumb

① Hold your knife at a slight angle to the wood. Push it with a thumb on top of the blade.

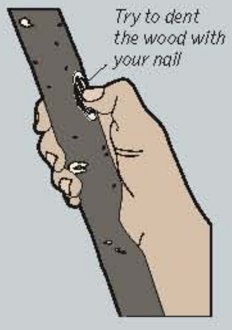
Try to dent the wood with your nail

② Press your thumbnail at an angle of 90 degrees to the wood in the exposed patch.

CHAR TEST

Each type of wood produces a distinctive char and coal. Some, such as hazel and lime, make a coal easily but the fireset will wear out quickly. The char test helps you discover which woods make a good coal. Drill into a softwood hearth board until it smokes, continue for five seconds and then stop. Look at the char produced:

- If it's like a fine powder it will probably produce a coal.
- If it's coarse, or appears to disintegrate into nothing, it probably won't make a coal.

④ Make the hearth board from a flat piece of wood about 1 in (25 mm) thick. You may have to split some wood to get what you need.

- Hold your drill so that it sits about ⅖ in (10 mm) from the edge of the board.
- Mark this point with your knife and notch out a small hole to help the drill get started.

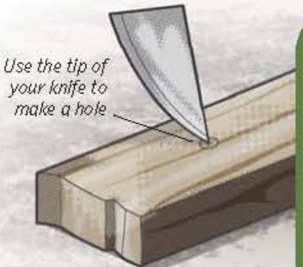
Use the tip of your knife to make a hole

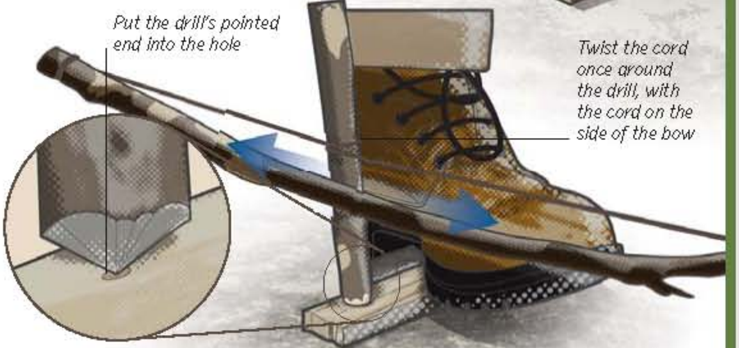
Put the drill's pointed end into the hole
Twist the cord once around the drill, with the cord on the side of the bow

⑤ Set up the bow, drill, and bearing block (see also p. 132) and bow the drill into the hole in the hearth board until the hole is as wide as the drill. Don't go all the way through.

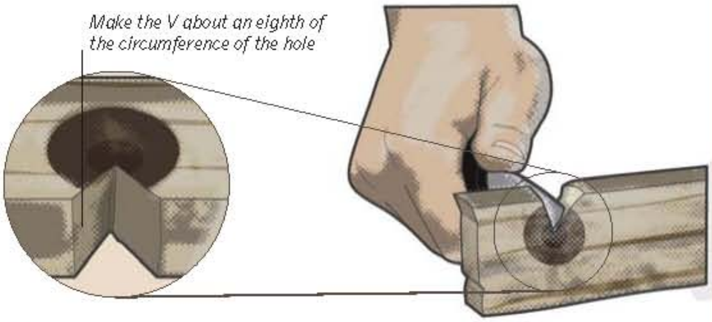
Make the V about an eighth of the circumference of the hole

⑥ Cut a V to the center of the hole through the hearth board. The bottom of this V is where the char collects and the ember forms in the ember pan (see p. 132).

THE FINISHED FIRESET

The bow drill set works best when the drill and hearth board are made from the same wood, because both parts then wear down evenly. This may not always be possible, so experiment with different woods.

- Good woods to start with are sycamore, hazel, ivy, lime, willow, and sotol (not found in Europe).
- Also useful are alder, birch, cedar, pine, spruce, maple, oak, and poplar.

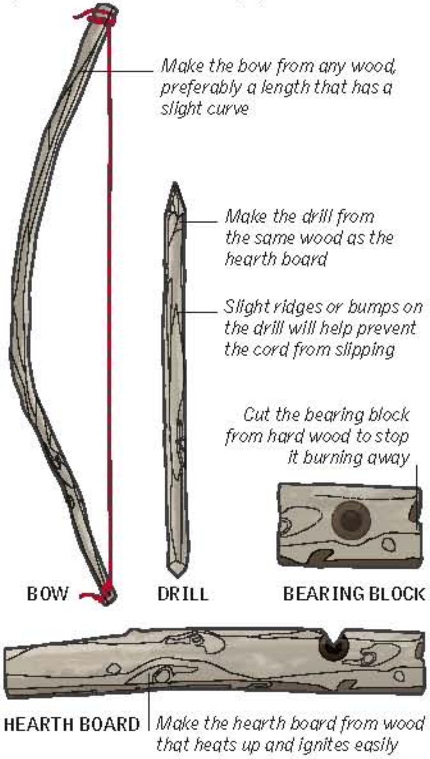
Make the bow from any wood, preferably a length that has a slight curve
Make the drill from the same wood as the hearth board
Slight ridges or bumps on the drill will help prevent the cord from slipping
Cut the bearing block from hard wood to stop it burning away
BOW
DRILL
BEARING BLOCK
HEARTH BOARD
Make the hearth board from wood that heats up and ignites easily

USING A BOW DRILL

Before you use a bow drill, have your dry tinder bundle, kindling, and fuel beside you. Until you become proficient at using a drill, you will use lots of energy, so beware of overheating and dehydration. Turning a coal into a fire is a skill in itself, so use every opportunity to practice making tinder bundles and lighting them with a small coal from your camp fire.

TOOLS AND MATERIALS

- Bow drill set (see pp. 130–31)
- Thin, dry piece of bark to use as an ember pan
- Buffed tinder, kindling, and fuel

GETTING INTO POSITION

Place the ember pan on a flat, dry surface and align it with the V on the hearth board. Rest one foot on the board. Loop the bow cord once around the drill, starting on the inside. Insert the drill into the hole on the hearth board, slot the block on top (with a green leaf in the notch for lubrication), and lock your wrist into your shin as support. Kneel on the other leg and bear down on the block.

1 Hold the bow horizontal to the ground and parallel to your body.

- Lean slightly forward to apply downward pressure to the drill and back to reduce it.
- Start bowing slowly, using the full length of the bow. Aim for an even, flowing rhythm.
- Breathing steadily as you bow, gradually increase speed and pressure until smoke appears. Vigorously bow with maximum effort ten more times—this usually produces a coal in the ember pan.

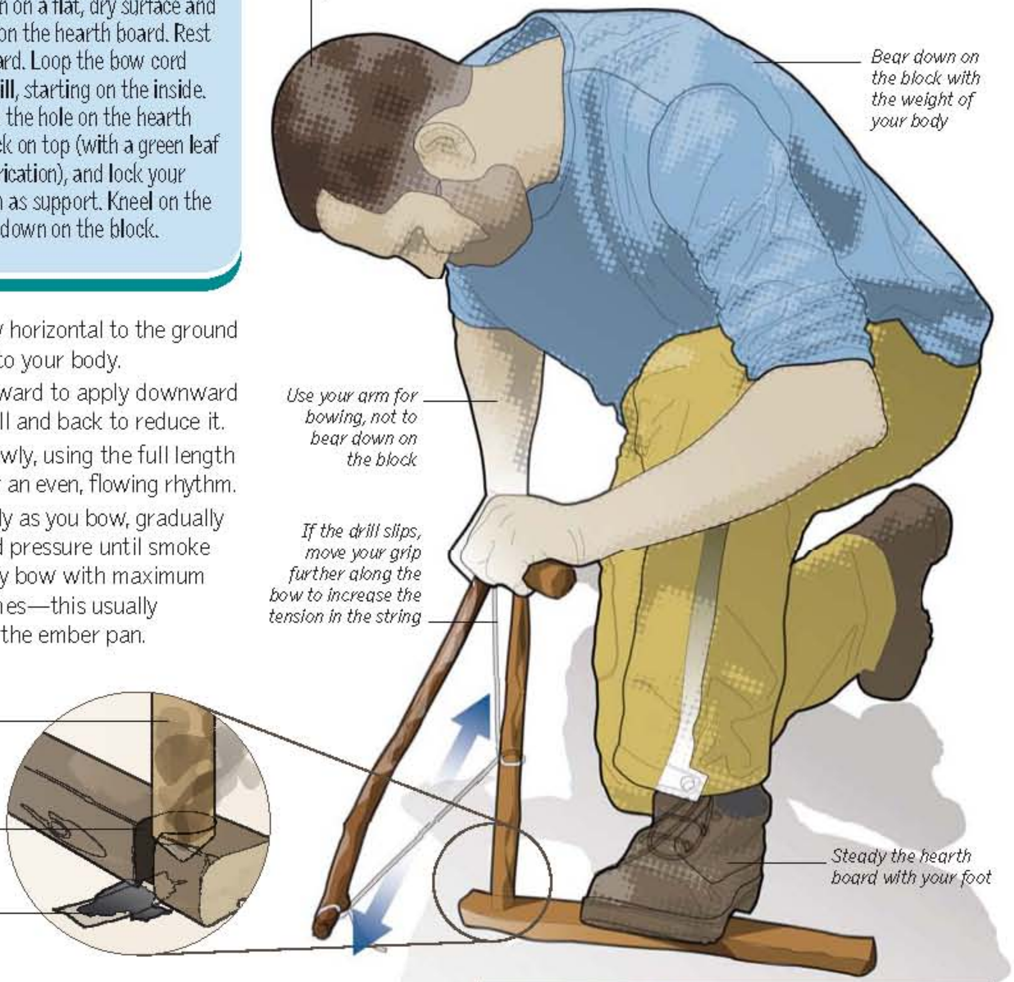

Keep your head as still as you can above the bow

Bear down on the block with the weight of your body

Use your arm for bowing, not to bear down on the block

If the drill slips, move your grip further along the bow to increase the tension in the string

Steady the hearth board with your foot

Keep the drill upright as you bow

As the drill turns, friction heats the wood

A pile of black ash will collect on the ember pan

2 Gently remove the drill and lift the hearth board up and away from the smoking coal in the ember pan.

- If the coal is sticking to the board, gently tap the board with your knife to loosen it.
- Carefully lift the ember pan off the ground and hold it in the air to see if the coal is glowing.
- If the coal in the ember is just smoking, and not glowing, then gently fan it with your hand until you see it glow.

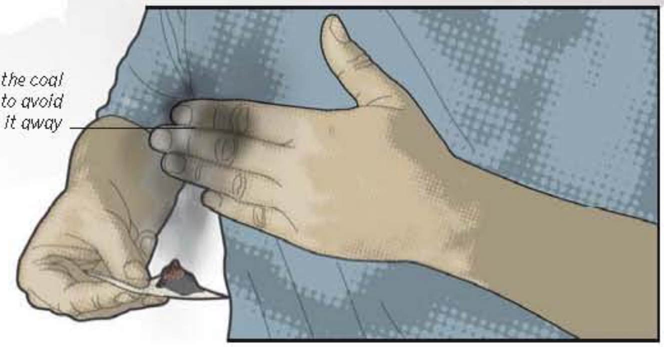

Fan the coal gently to avoid blowing it away

3 Gently transfer the glowing coal into your tinder bundle.

- The coal must touch the tinder to transfer its heat, so squeeze the tinder around the coal, but don't crush it or starve it of oxygen.

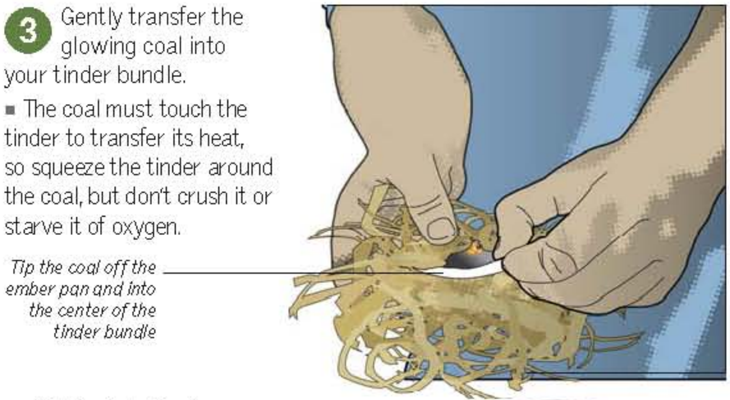
Tip the coal off the ember pan and into the center of the tinder bundle

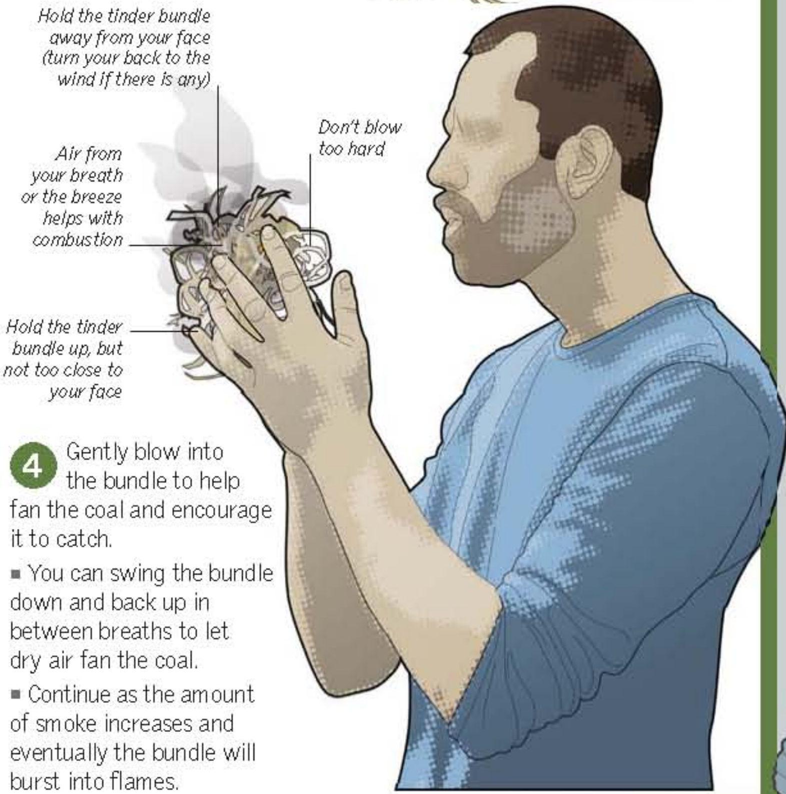
Hold the tinder bundle away from your face (turn your back to the wind if there is any)
Don't blow too hard
Air from your breath or the breeze helps with combustion
Hold the tinder bundle up, but not too close to your face

4 Gently blow into the bundle to help fan the coal and encourage it to catch.

- You can swing the bundle down and back up in between breaths to let dry air fan the coal.
- Continue as the amount of smoke increases and eventually the bundle will burst into flames.

5 Don't throw the burning bundle on the ground, just calmly place it where you intend to have your fire. Start building the fire around the bundle (see pp. 124–25).

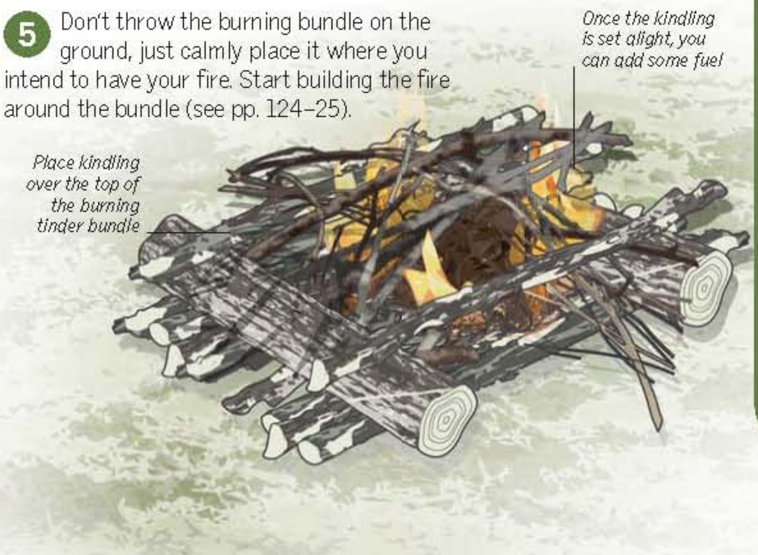
Once the kindling is set alight, you can add some fuel
Place kindling over the top of the burning tinder bundle

BAMBOO FIRE SAW

This friction method uses a saw technique to light tinder collected from scraping the outside of the bamboo with your knife.

MAKING THE FIRE SAW

You need a bamboo section 18 in (45 cm) long. Split it lengthways into two equal pieces.

1 Sharpen the edge of the first piece with your knife and then cut a V-shaped notch in the outside middle of the second piece.

2 Hold tinder over the notch in the second piece of bamboo with another bamboo strip that's snapped but not fully broken.

- Kneel down and brace the sharpened piece of bamboo vertically between the ground and your stomach (place a pad on your stomach for comfort).
- Place the notch and tinder over the sharp edge of the first piece, and rub the second piece up and down in a rhythmic movement.

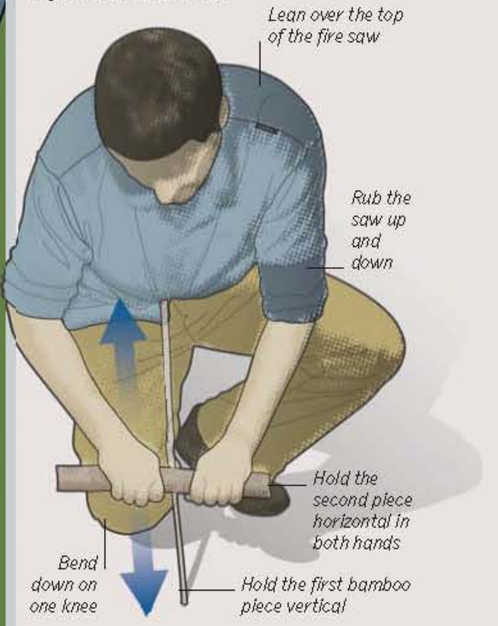
Lean over the top of the fire saw
Rub the saw up and down
Hold the second piece horizontal in both hands
Bend down on one knee
Hold the first bamboo piece vertical

3 When the friction produces lots of smoke, continue for 20 strokes more and you should have a coal.

MANUAL SKILLS

Learning skills that you can use in the wilderness is always a “work in progress.” Having taught thousands of students over several years, I am yet to conduct a course in which I didn’t learn a new skill. For example, this could be a more efficient way to erect a shelter, how to tie a new knot, or the best way to carry out a task more safely and efficiently.

In the military, we invariably teach students how to accomplish a particular task using only the most basic gear. By learning how to perform a task using the absolute basic equipment, they will be able to understand the important elements of the task and will have no choice but to get them right in order to succeed. This in turn helps students to appreciate how and why certain techniques do and don’t work, and gives them an opportunity to use their improvisational skills to modify basic techniques in order to get the most out of them. Knowing the correct method of using a knife, for example, will not only improve safety on an expedition, but will also mean that only the minimum

In this section YOU WILL DISCOVER...

- the merits of a **hank** in **your pocket**...
- how to strip a **nettle stem** and not **get stung**...
- when **to twist** and when **to roll**...
- how to tell a **working end** from a **standing end**...
- the uses of a **sheet bend** or a **Siberian hitch**...
- the **difference** between a **parang** and a **kukri**...
- how to **fell a tree** with a **thumping stick**...

amount of valuable energy is being used to complete a task. Equally, knowing how to tie a few simple knots, which can be used in the majority of survival situations, will allow you to use whatever cordage you have in the most efficient way possible.

If you find yourself in a survival situation with more than the most basic gear, life will be much easier. However, the ability to improvise what you need when you don't have it could mean the difference between continued survival and despair.

If you give someone with little understanding of knots a length of cord and ask them to complete a task, they will invariably use the whole length of cord. If, however, you teach someone two simple knots and set them the same task, they will invariably use only the necessary amount of cord.

USING THE BEST KNOT, lashing, or hitch for a particular task means that you will use cord more efficiently and, therefore, ensure that your supply will go further. In a survival situation, where you may be moving every day, you need to be able to use the same pieces of cordage day after day.

- Use knots that don't require cutting and that can be untied quickly and easily (trying to undo a knot with cold, wet hands in the dark, and when you are tired and miserable, can be an emotional experience that usually ends in frustration and several pieces of cut cordage).
- Before you set off into the wilderness, learn a few basic knots, such as the Siberian hitch and the taut-line hitch, which are great for rigging a simple poncho shelter.
- Practice these knots until you can tie—and untie—them with your eyes closed.
- If you are trekking in a cold environment, practice tying them wearing gloves.

> “The **ability** to **improvise** what you **need** when you **don't have it** can mean the **difference** between **continued survival** and **despair**”

MAN-MADE CORDAGE

A VITAL PIECE OF SURVIVAL EQUIPMENT, man-made cordage is one of those small items that has many uses: building a shelter, repairing equipment and clothing, making traps and nets, and producing fire by friction using a bow drill. Always pack plenty.

PACKING CORDAGE

Before you set off on your travels, check how much cordage you will need. Make sure you pack it in two places—in your backpack and on your person, in case you get separated from your pack. Make sure, too, that the needle you pack in your survival kit has an eye big enough for the inner strands of the cordage.

PARACORD

Parachute cord, or "paracord," is a type of cordage originally developed for the rigging lines of parachutes and then adopted as the standard utility cordage by most of the world's military forces. It is a lightweight, nylon rope composed of 32 braided strands. Each strand contains a number of smaller threads that can also be used. In most situations, paracord is a very good option because it's readily available, strong, and packs down small.

USING CORDAGE

It is vital that you use your limited supply of cordage as efficiently as possible.

- Use only the minimum cordage that is absolutely necessary to accomplish each task.
- Use simple, strong knots that, if possible, can be released without cutting. This leaves the cordage intact for the next task.

SURVIVAL TIP

Green is the ideal color for military paracord as it remains camouflaged. In a survival situation, red is best because it stands out—it's easier to find if you drop it and can be used as an aid to location. Don't pack just one very long piece of paracord (150ft/30m, for example). It's much better to cut the paracord into 30 ft (10 m) lengths and tie each of them into a coil called a hank (see opposite and below).

TEN HANKS

A hank can fit in the palm of your hand. It takes up very little room so you can easily stuff up to 10 of them into your pack and pockets, and leave them there until you need them.

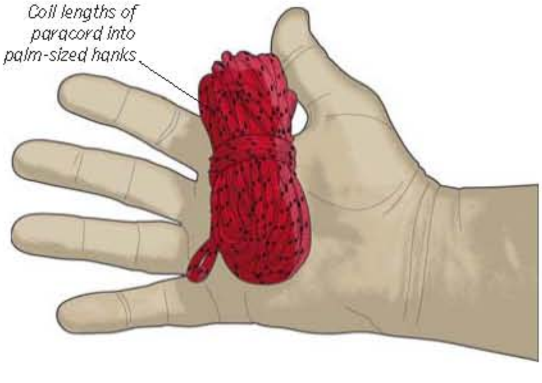
Coil lengths of paracord into palm-sized hanks

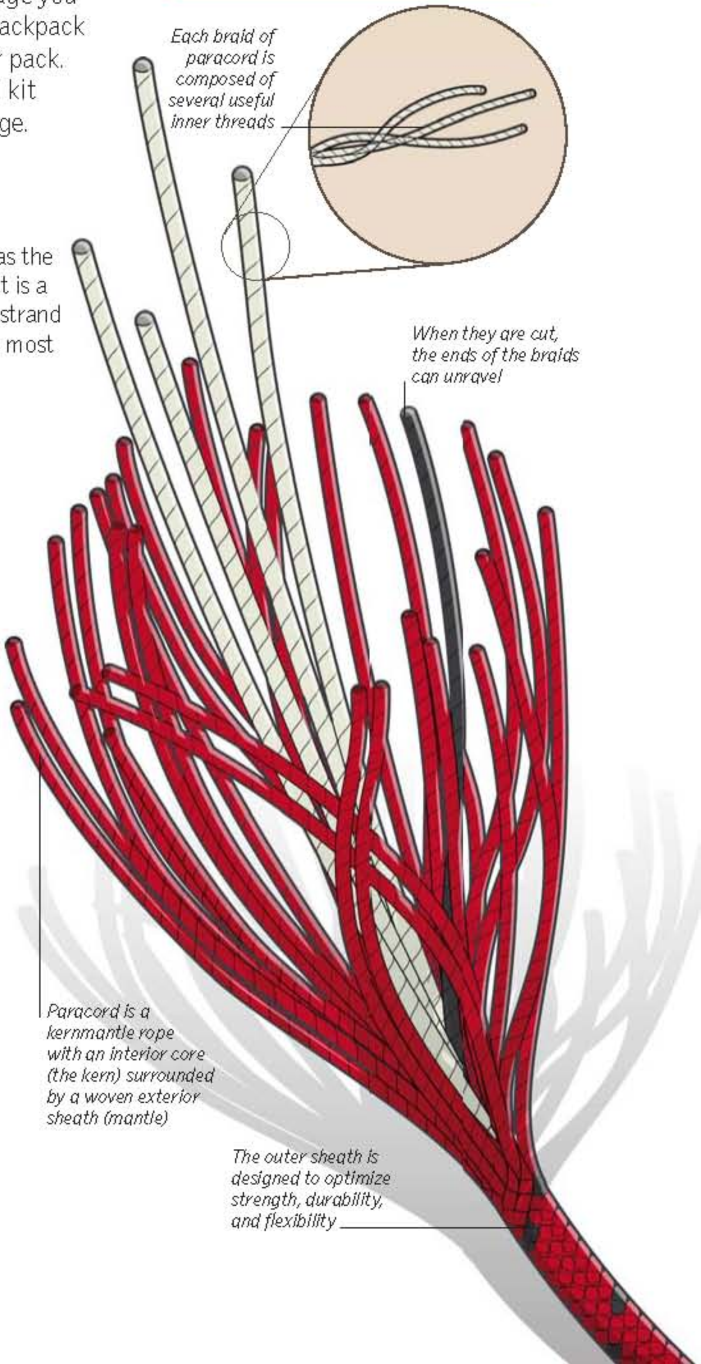
Each braid of paracord is composed of several useful inner threads

When they are cut, the ends of the braids can unravel

Paracord is a kernmantle rope with an interior core (the kern) surrounded by a woven exterior sheath (mantle)

The outer sheath is designed to optimize strength, durability, and flexibility

TYING A HANK

It's best to keep your cordage in a hank—you can untie it when you need a piece and retie it to store it again. Tying a hank is straightforward and soon becomes second nature, because you'll do it again and again.

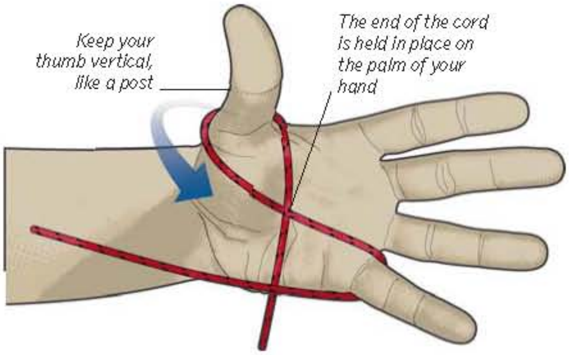
Keep your thumb vertical, like a post
The end of the cord is held in place on the palm of your hand

1 Hold out your hand (you can use either hand) with the thumb and the fingers spread.

- Lay the end of the cord on your palm, then loop the cord in a figure-eight around your thumb and little finger.

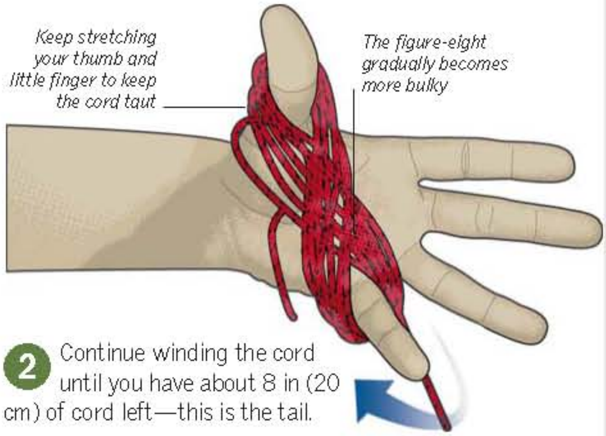
Keep stretching your thumb and little finger to keep the cord taut
The figure-eight gradually becomes more bulky

2 Continue winding the cord until you have about 8 in (20 cm) of cord left—this is the tail.

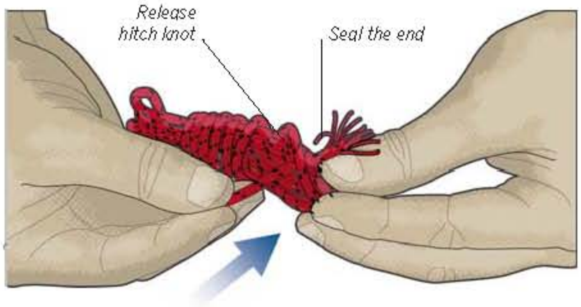

3 Take the hanked cordage off your thumb and little finger.

- Secure the hank by tying the tail around it with a release hitch knot (see p. 143).

OTHER TYPES OF CORDAGE

In a survival situation you may have to improvise and use other types of cordage, especially if you didn't bring any paracord with you or have run out. Before you start making it from natural materials (see pp. 138–41), see what else you might have that would be quicker and easier to adapt.

SHOELACES

The normal laces of shoes and boots are often strong enough to use as cord. However, the members of many military units, who are limited to the amount of equipment they can stow or carry, replace their bootlaces with paracord. They use twice as much as they need and wrap the extra length around the top of their boots before tying it. This gives them a ready supply of extra cordage that they can cut off as needed. You can easily do the same—but never sacrifice all the cordage you are using as shoelaces as you'll need to make sure you can still walk comfortably.

BELTS, CLOTHING, AND HATS

You can cut any type of belt, whatever it's made of (leather, canvas, and so on) into strips to use as cordage. You can cut clothing into strips, too. However, you don't have to destroy a shirt completely: start at the bottom hem and cut 1 in (2.5 cm) strips all the way round. You can even use your hat—a 1 in (2.5 cm) strip from the brim of a jungle hat is just over 3 ft (1 m) long. Better still, coil paracord around the crown of your hat and stitch it in place (just the front and the back).

RAIDING YOUR KIT

You may have other items in your equipment that you could use as cordage, such as guylines from tents or draw cords from clothing or sleeping bags. Dental floss, especially the waxed type, is extremely strong and can be used effectively for lashing or sewing. If you have a vehicle, you probably have a towrope; take a part of it and unravel the rope into smaller pieces of cordage.

BE A SCAVENGER

Few places are untouched by modern man. Even in remote areas you'll come across discarded items from the "civilized world"—from plastic bottles to bailing twine. In a situation where improvisation could be the difference between life and death, collect anything you could use to your advantage.

SEALING PARACORD

Paracord can be awkward to use because its ends, if left unprotected, eventually unravel. Try to seal the ends after you have cut them by holding a flame next to them until they melt. Be careful, as the melted nylon can drip and burn.

NATURAL CORDAGE

IF YOU DON'T HAVE any man-made cordage, you may be able to make a natural substitute. Use the roots and stems of plants to produce improvised rope for erecting shelters, and use sinew from the tendons of larger animals for strong whipping and sewing.

WARNING!
Take extreme care when you collect and work with plant fibers, especially removing plants such as green ferns from the ground. Bending a stem at its base and pulling it up can expose fibers that are razor-sharp. Cut the plant with a knife or wear gloves. The juice of some plants can irritate the skin, so wash your hands when you've finished.

PREPARING NETTLE FIBERS

Old stinging nettles or wood nettles with long stems have the most fibers. Wearing gloves (or with your hands covered), strip the leaves from the nettles by grasping the stem at the base and pulling your clenched fist up the full length of the stem.

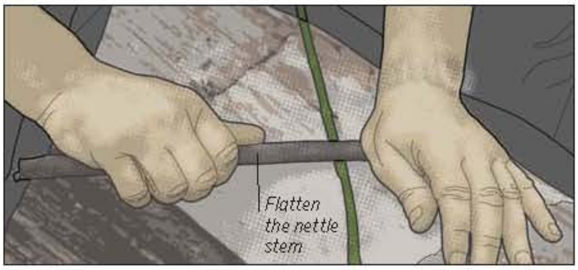

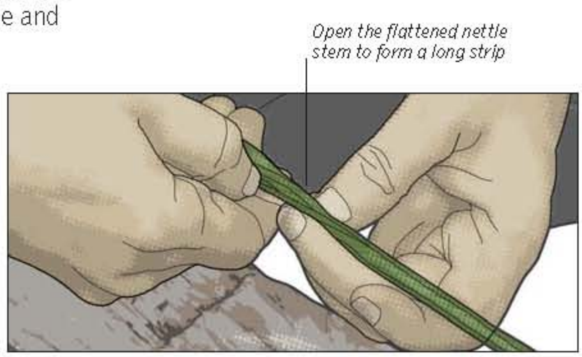
Open the flattened nettle stem to form a long strip

① Sit astride a log and lay a nettle stem in front of you. Roll a smooth, rounded stick backward and forward on the stem, pressing down hard.

- Continue rolling until the whole stem is crushed.

② Tease open the crushed tissue along the whole length of the stem of the nettle with your thumbnail. This exposes the spongy pith within.

CHOOSING YOUR RAW MATERIALS

Many kinds of raw material can be used to make natural cordage, from bark and roots to stems and sinews. Search your immediate environment first to see what can be used—you may not have to look too far.

	BARK	ROOTS
FACTS	■ Tree trunks have layers between the bark and heartwood called the "cambium" layers, which are best for cordage.	■ Many tree roots make good cordage and lashings, but evergreen trees, such as pine, fir, spruce, and cedar, work best.
FINDING	■ The bark of trees such as willow and linden make good cordage. ■ Use other types of bark, such as clematis or honeysuckle, as they break from the tree or vine.	■ Look for new roots near the surface of the soil; these are flexible. Cut slightly thicker roots than you need, as you have to remove the bark. Remove only a few roots from each tree and repair any damage you do to the topsoil.
PREPARING	■ Strip the bark from a dead tree, then beat it to loosen the inner cambium fibers. ■ Pull off the cambium fibers in long strips. Use them as they are (clematis and honeysuckle), or braid them into stronger lengths.	■ Cut a slit along the root. Peel the bark and let the bare root dry and shrink before splitting in half or into quarters. ■ Keep the bark for tinder or kindling. You could also put it on a fire to produce smoke that keeps away flies.

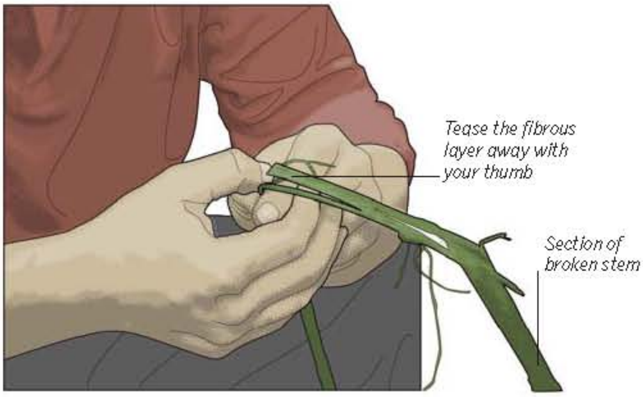

Tease the fibrous layer away with your thumb

Section of broken stem

3 Bend the stem in the middle of its length. When you bend it, the inner pith will break away from the outer skin, making it easy for you to separate them.

MAKING CORDAGE FROM PLANTS

To begin with, you must have access to enough quantities of the plant. Don't start making cordage with the only plant in the area. There should also be some basic characteristics present in order for plant material to work well as cordage.

- The fibers need to be long enough. If you have to splice them together to make workable lengths, you will weaken the line.
- If you're braiding pieces together to make a stronger line, use rougher fibers, as they will bite together better. Shiny or smooth fibers tend to unravel easily.
- The fibers need to be strong. Pull on them to see how much strain they take before breaking. They also need to be pliable enough so they don't break when you bend and tie them.

4 Carefully peel the outer skin from the inner pith.

- With the finger and thumb of one hand, gently pull the layer outward, while pushing down on it with the thumb of the other hand.
- You'll be left with long fibrous strips of outer skin, which you can make into short bindings or natural cordage (see p. 140).

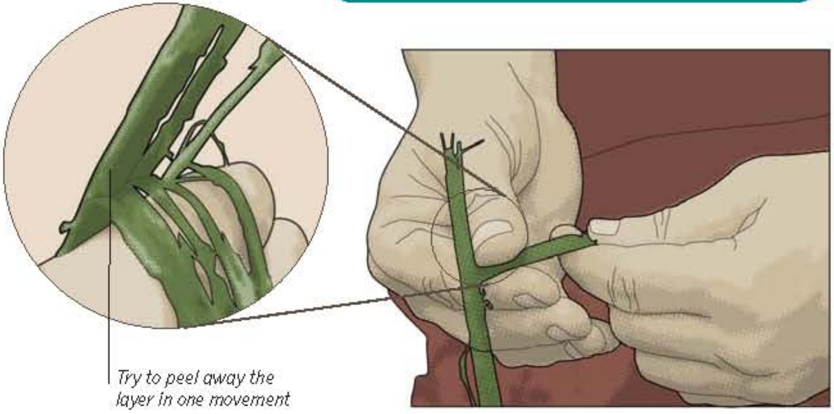

Try to peel away the layer in one movement

	WITHIES	LEAVES	SINEW
FACTS	▪ The strong, flexible stems of willow, birch, ash, and hazel are called "withies" and are used for thatching and in gardening. ▪ Withies make useful and robust lashing and are best used in spring or summer.	▪ The leaves of many plants, such as the lily family, agave, and sisal hemp, contain useful fibers.	▪ Animal sinew makes strong and versatile cordage—some sinews can support the weight of a man. ▪ Sinew has been used to bind arrow heads to shafts and to sew together everyday clothing, such as leather and mulduks.
FINDING	▪ Choose a young sapling or branch that's long and flexible. The fewer the stems or branches the better, as they will need to be removed.	▪ To find out if the leaves of a plant contain useful fibers, simply tear a leaf apart to see if it breaks into stringy layers.	▪ Sinews are the tendons connecting bone to muscle. The largest are on both sides of the spinal column, running parallel to it. ▪ Shorter lengths can be taken from the calves of the hind legs.
PREPARING	▪ Remove stems or branches. Grasp the wand at the base and twist until the fibers break, working your way to the tip. Bend the wand into an S-shape and crank the middle of the wand to loosen the fibers. Then cut the wand from the tree and use the withies.	▪ Soak the leaves in water to make the inner layers swell and burst. Bacteria work on the cell tissue, separating it from the fibers. ▪ Rinse the fibers in fresh water to stop further bacterial decay and then dry them.	▪ Take a tendon from an animal's body, remove its outer sheath, then clean it and let it dry. Separate the dried fibers and use them as individual threads, or splice or braid them together to form stronger cords. You can soften dried sinew with water or saliva.

MAKING NATURAL CORDAGE

Once you have prepared a sufficient quantity of fibers, it's best to let them dry before using them—hang them up in the sun or near a fire. In a survival situation, you can use them before drying. Twisting or rolling the fibers are the two main methods of making natural cordage. As with most survival tasks, practice makes perfect.

TOOLS AND MATERIALS

Enough prepared natural fibers (see pp. 138–39) to give you the length of cordage that you require

TWISTED CORDAGE

Start your first length of cordage toward one end, not in the middle, to prevent all the subsequent splices from coinciding at the same point. This provides strength, as each splice is a potential weak point. If they are thin enough, the twisted cordage fibers can then be plaited to create a thicker cord that's even stronger.

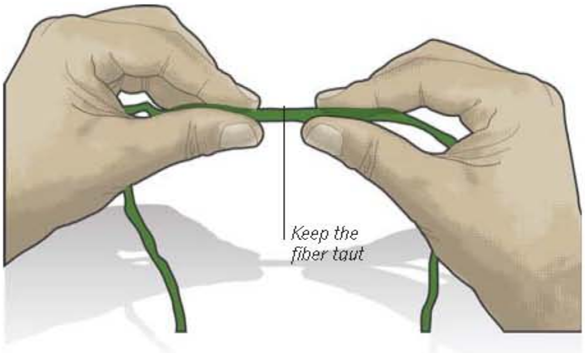

Keep the fiber taut

1 Grip the first length of fiber a third of the way along its length between your left thumb and forefinger.

- Grasp the fiber with the thumb and forefinger of your right hand 1 in (3 cm) from your left.

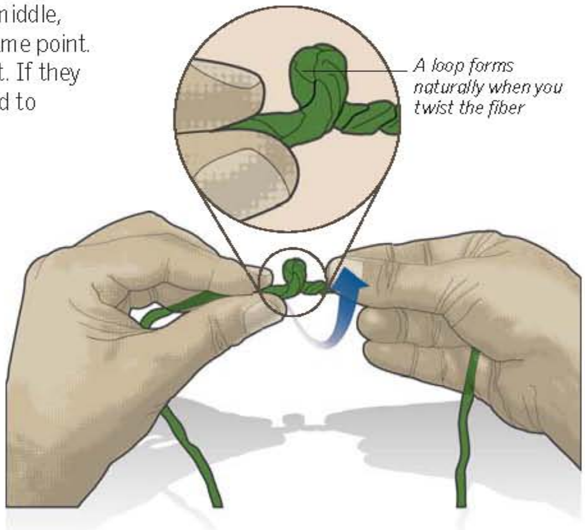

A loop forms naturally when you twist the fiber

2 Twist the fiber with your right hand until a tight loop is formed.

- Keeping the tension in the cord, inch your left thumb and forefinger forward and clamp down on the loop.

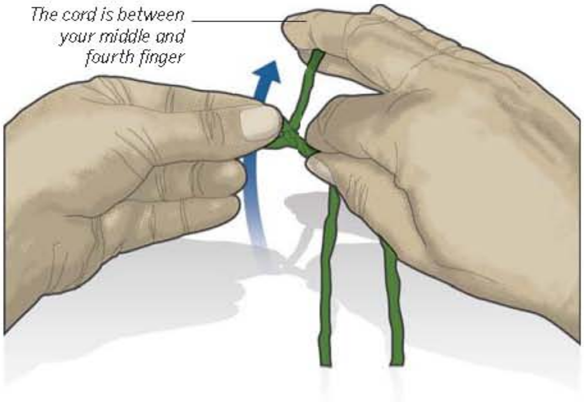

The cord is between your middle and fourth finger

4 Sweep the middle and fourth fingers of your right hand upward, pulling the left-hand cord under the captive piece.

- Inch your left thumb and forefinger forward, and clamp down on the new twist.

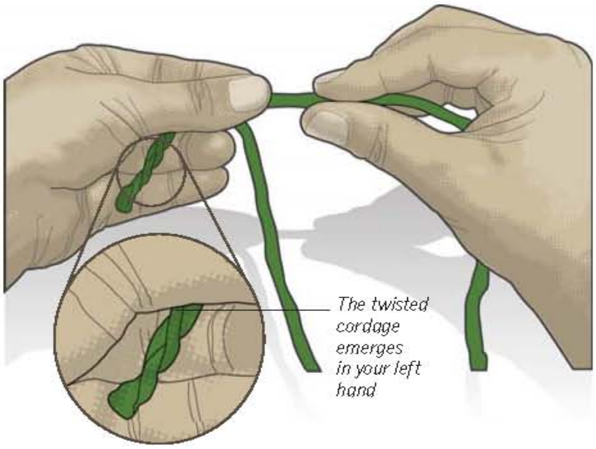

The twisted cordage emerges in your left hand

5 Now that the new twist is clamped by your left hand, you can release your right thumb and forefinger, letting the lower cord fall.

- Grip the upper cord 1 in (3 cm) from your left fingers with your right thumb and forefinger.

SIMPLE ALTERNATIVE TWIST

You need to start with two fibers of different lengths so that when you splice on additional lengths the splices are staggered. Start by tying one end of one fiber with one end of the other.

- Clamp tied ends in your left thumb and forefinger and the two tails in your right thumb and forefinger about 2 in (5 cm) down.
- Roll both fibers between your fingers away from you for one complete roll. Repeat this until you have about 3 in (7.5 cm) of rolled fibers, then clamp them in your right thumb and forefinger.
- Release your left thumb and forefinger and the rolled fibers will naturally twist together to form a strand of cordage.
- Repeat the process until you have enough for your needs.

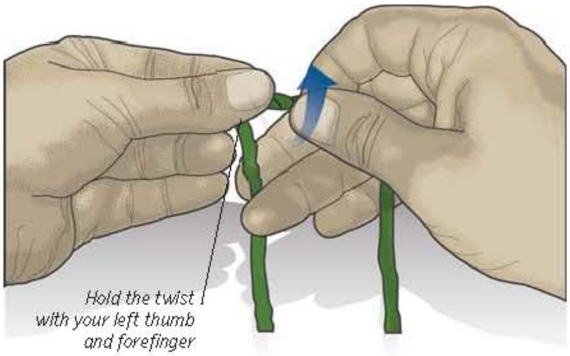
Hold the twist with your left thumb and forefinger

3 Move your right hand back so that 1 in (3 cm) of cord is visible and apply another twist.

- At the same time reach forward with the middle and fourth fingers of your right hand and grasp the cord hanging from your left fingers.

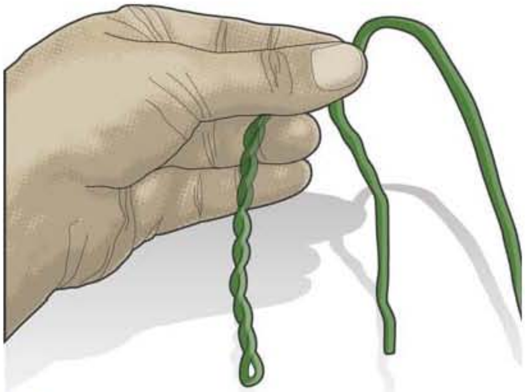

6 Repeat Steps 3, 4, and 5 as you work your way along the length of cord.

- To splice in an extra length, overlap the new piece 2 in (6 cm) along the existing fiber and twist the two together between your left thumb and forefinger. Then continue with the process as before.

ROLLED CORDAGE

This method of rolling cordage, which is also known as the "pygmy roll," is easier to do if you're sitting down. As with the simple alternative twist (see left), you need to start with two fibers of different lengths and tie one end of one fiber to one end of the other.

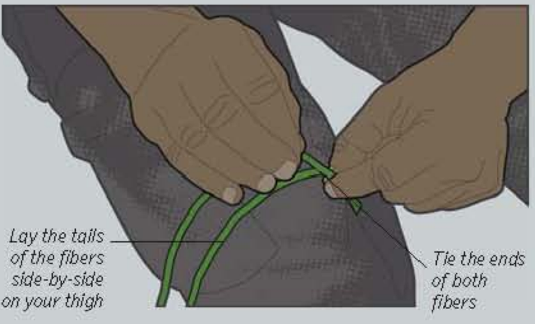
Lay the tails of the fibers side-by-side on your thigh
Tie the ends of both fibers

1 Hold the tied ends between the thumb and forefinger of your left hand.

- Lay the two tails of the fibers on your right thigh.

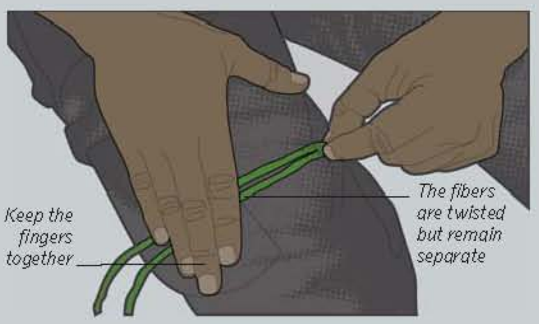
Keep the fingers together
The fibers are twisted but remain separate

2 Use the flat palm of your right hand to roll both fibers simultaneously away from you. You may find it easier to use the flat of your fingers.

- At the end of the roll, clamp the fibers down onto your thigh with the tips of your fingers.

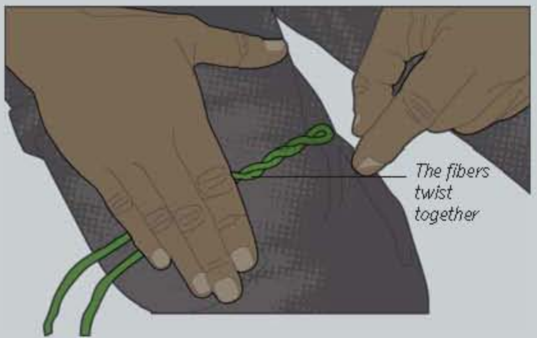
The fibers twist together

3 Release your left thumb and forefinger and the rolled fibers will naturally twist together to form a strand of cordage.

- Repeat the process, splicing in more cordage as required.

TYING KNOTS

THERE ARE HUNDREDS of different types of knot in existence, but by practicing the handful of straightforward knots shown here and overleaf, you should be able to accomplish most survival tasks. Start by tying each knot in the situation you expect to use it—for example, erecting a poncho shelter—and then practice it at least 20 times.

THE ENDS OF A LINE
The end of a rope or line that takes the most active part in knot tying is the working end. The other end is called the standing end—it's more passive, as the knot is tied around it.

SLIP KNOT

A slip knot is a good example of a simple knot that has many practical uses in survival situations.

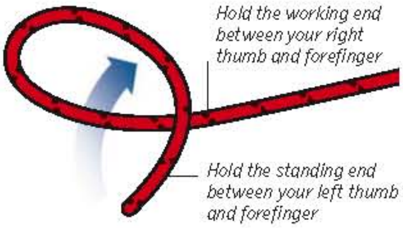
Hold the working end between your right thumb and forefinger
Hold the standing end between your left thumb and forefinger

1 Twist the standing line over the working line to form a loop.

- Hold the loop between your left thumb and forefinger.

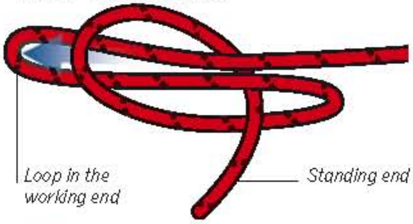
Loop in the working end
Standing end

2 Make a loop with the working end and pass it into the first loop.

- Collect the second loop with your left thumb and forefinger and pull it through.

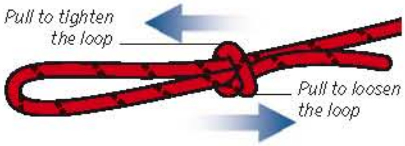
Pull to tighten the loop
Pull to loosen the loop

3 Gather both ends in your right thumb and forefinger and pull them away from the loop in your left hand.

- Put the loop over whatever you want to secure—for example, a button-tie on your bivy sack—and pull the knot tight.

ARBOR KNOT

This all-purpose knot has several survival uses. When erecting a poncho or a tarp (see Chapter Four), it can be used as a tie-off knot to secure one end of a line to a fixed point, such as a tree or peg, where no adjustment is required. It can also be used as a lashing to secure two poles together—for example, in an A-frame shelter (see p. 164)—with just one turn of cordage.

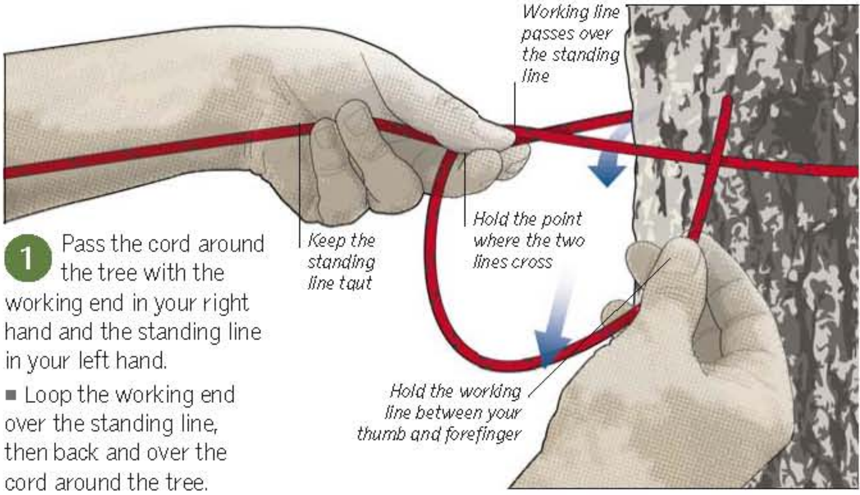
Working line passes over the standing line
Keep the standing line taut
Hold the point where the two lines cross
Hold the working line between your thumb and forefinger

1 Pass the cord around the tree with the working end in your right hand and the standing line in your left hand.

- Loop the working end over the standing line, then back and over the cord around the tree.

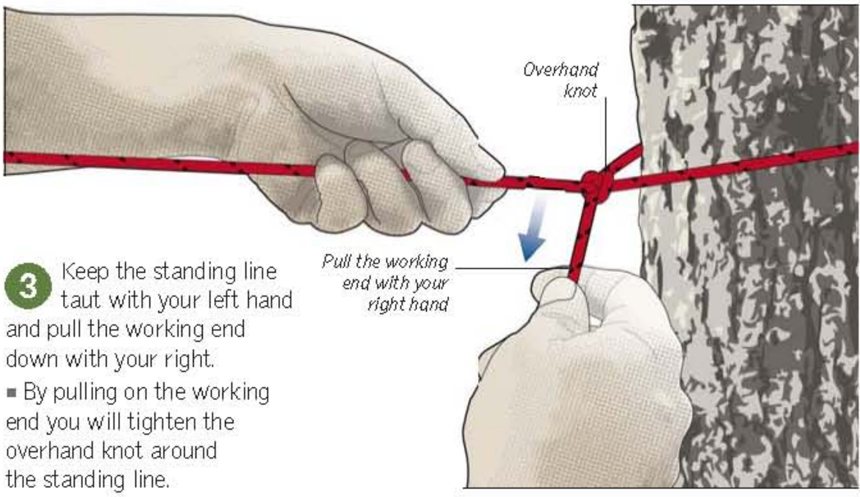
Overhand knot
Pull the working end with your right hand

3 Keep the standing line taut with your left hand and pull the working end down with your right.

- By pulling on the working end you will tighten the overhand knot around the standing line.

OVERHAND KNOTS

An overhand knot is the first part of the knot you use to tie the laces on your shoes. One overhand knot tied around another forms a stop knot or they can be tied as a double overhand knot around a fixture.

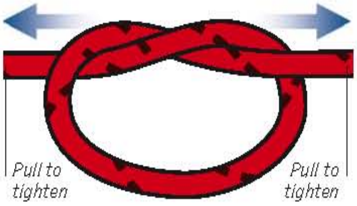
Pull to tighten
Pull to tighten

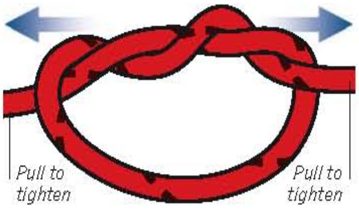
Pull to tighten
Pull to tighten

OVERHAND KNOT

Twist the left line over the right to form a loop. Hold it with your right thumb and forefinger.

- Pass the working end through the loop with your left forefinger and thumb and pull both ends to tighten.

DOUBLE OVERHAND KNOT

Follow the sequence for the overhand knot (left), but pass the working end through the loop twice before pulling the two ends.

- A stop knot used at the end of a line can stop the line from slipping.

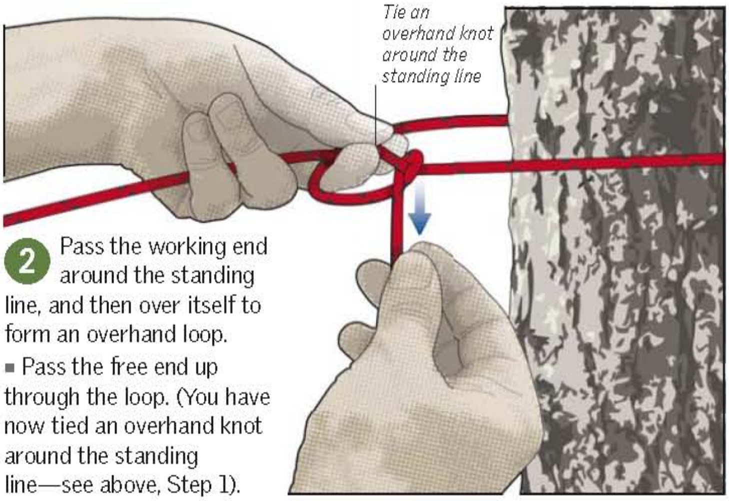
Tie an overhand knot around the standing line

2 Pass the working end around the standing line, and then over itself to form an overhand loop.

- Pass the free end up through the loop. (You have now tied an overhand knot around the standing line—see above, Step 1).

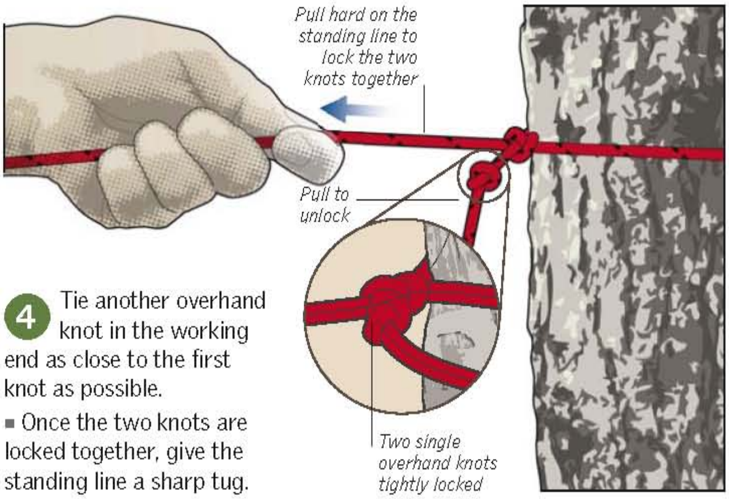
Pull hard on the standing line to lock the two knots together
Pull to unlock
Two single overhand knots tightly locked

4 Tie another overhand knot in the working end as close to the first knot as possible.

- Once the two knots are locked together, give the standing line a sharp tug.

DOUBLE CHAIN FASTENING KNOT

This knot securely attaches the free end of a line to a fixture, particularly when weight is involved and you don't want the knot to slip or come loose.

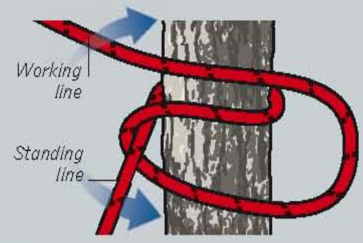
Working line
Standing line

1 Loop the working end around the fixture, such as a tree or pole.

- Pass the working end over and around the standing line.

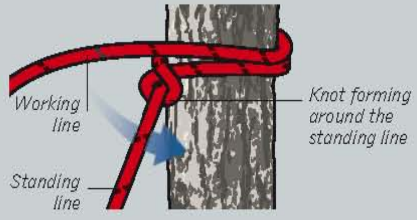
Working line
Standing line
Knot forming around the standing line

2 Pass the working end back around the fixture and then bring it back under the standing line.

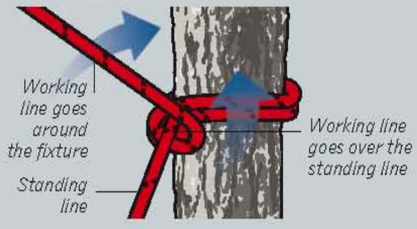
Working line goes around the fixture
Standing line
Working line goes over the standing line

3 Pass the working end over the top and around the standing line again, and once more around the fixture.

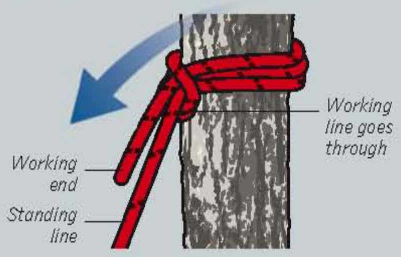
Working end
Standing line
Working line goes through

4 As you bring round the working line, thread it underneath itself to complete the knot.

- Lock the knot neatly into position.

DOUBLE SHEET BEND

The sheet bend, also known as the "weaver's hitch," is an extremely useful survival knot as it securely ties together two ropes. Double sheet bends are even more secure and are recommended for tying ropes of different thickness. In this instance you use the thinner of the two ropes to tie the knot.

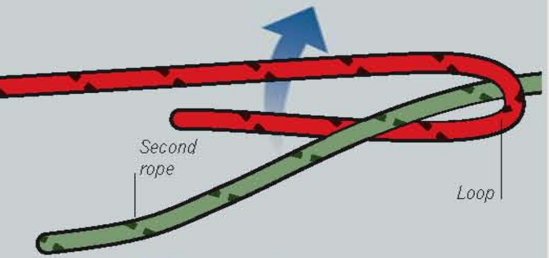

1 Start with two ropes and form a loop in the working end of one of them.

- Pass the end of the second rope through the loop.

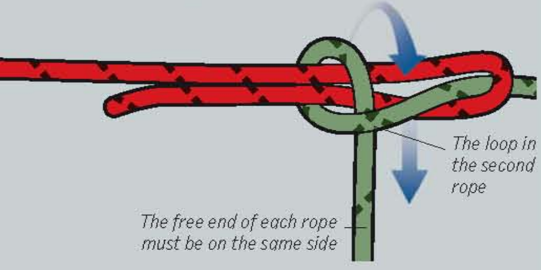
The loop in the second rope
The free end of each rope must be on the same side

2 Pass the free end of the second rope first under, then up and over the loop. Pass the end through the new loop you have created.

- To finish a single sheet bend, pull the standing end of each rope to tighten it.

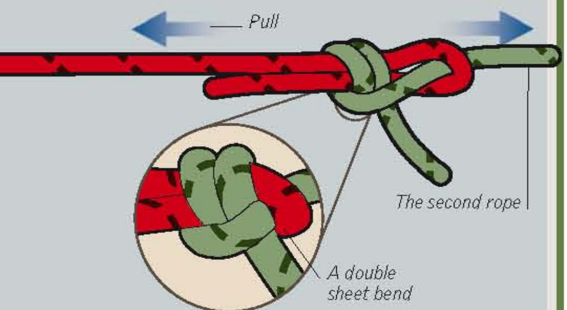

3 To tie a double sheet bend, pass the free end of the second rope through the loop (as in step 2) again.

- Tighten the knot as before by pulling the free end of each rope.

SIBERIAN HITCH

This knot, which is also known as the "Evenk knot," is good for attaching a rope to a fixture—for example, when securing a ridgeline to a tree for a poncho shelter (see p. 158).

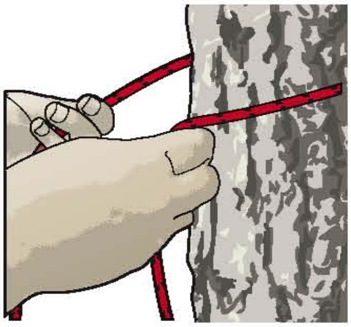

1 Pass the cord around the tree with the working line in your right hand and the standing line in your left hand.

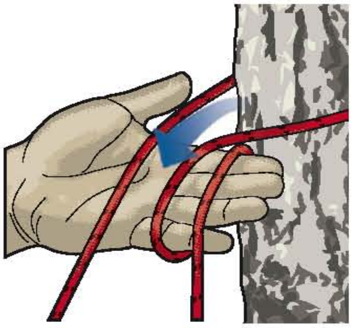

2 Lay the standing line on the palm of your left hand.

- Loop the working line one and a half times around your fingers.

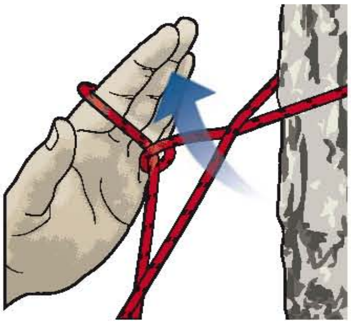

3 Bring your left hand under the standing line.

- Turn the palm of your hand so that one loop twists over another.

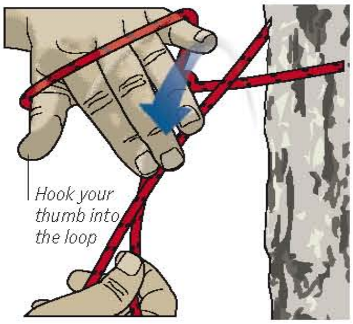
Hook your thumb into the loop

4 Hold the working end in your right thumb and forefinger.

- Bring the loop that is around your left fingers over the standing line.

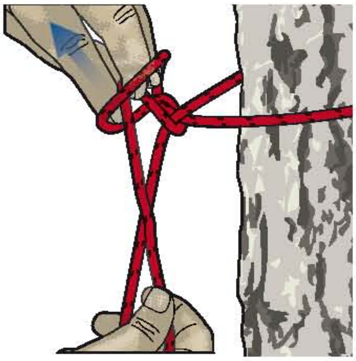

5 Pinch the working line with the fingers of your left hand.

- Pull it up through the loop to tighten the knot on itself.

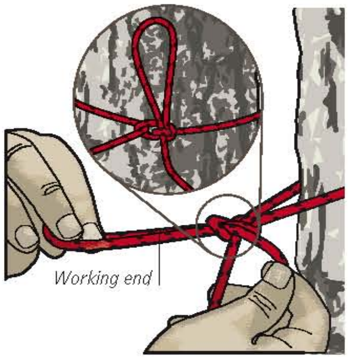

6 Pull the working end again, to tighten the knot around the tree.

- To free the knot, pull the end of the standing line.

TAUT LINE HITCH

This knot is widely used in survival and outdoor activities because it can be adjusted to increase or decrease the tension in a fixed line—for example, a guide rope on a tent or a mooring on a boat in tidal waters.

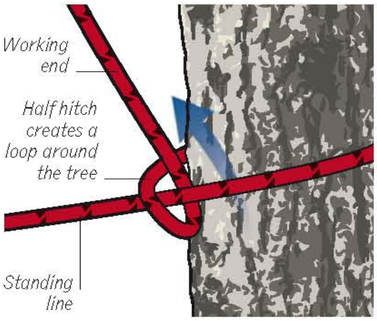
Working end
Half hitch creates a loop around the tree
Standing line

1 Attach the standing line to an anchor point, such as a tent, and pass the working line around a tree.

- Tie a half hitch (see below) with the working end around the standing line.

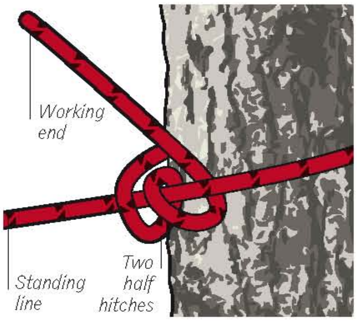
Working end
Standing line
Two half hitches

2 Tie another half hitch next to the first one, so that the two half hitches sit side by side around the standing line.

- Don't tighten the half hitches yet.

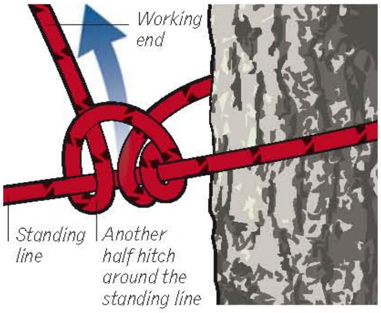
Working end
Standing line
Another half hitch around the standing line

3 Pass the working end under the standing line and back through the new loop it created (but outside of the loops in Steps 1 and 2).

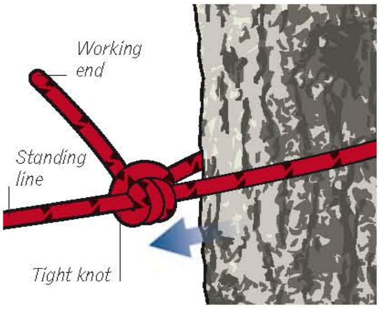
Working end
Standing line
Tight knot

4 Tighten the knot by pulling both the lines.

- Pull the knot further from the anchor point to increase the tension.

QUICK KNOT RELEASE

You can dismantle a tent or free a boat from its mooring more quickly by finishing off the taut line hitch (see left) with a quick-release knot.

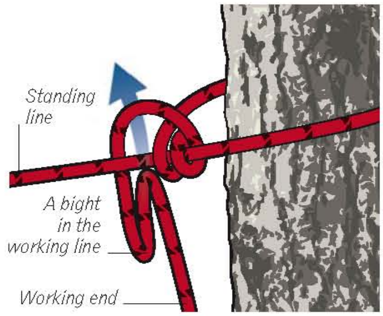
Standing line
A bight in the working line
Working end

1 Repeat Step 3 of the taut line hitch (see left), but this time make a loop, known as a "bight," in the working line and pass it—rather than just the end—through the loop.

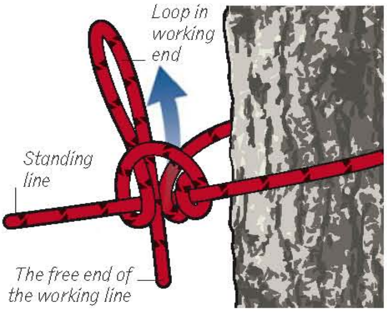
Loop in working end
Standing line
The free end of the working line

2 Tighten the knot as in Step 4 (see left).

- Leave the working end free so you can pull it quickly to release the knot.

HALF HITCH

The half hitch is a simple knot that forms the basis for other knots (see taut line hitch, above). It's not a secure knot and, when used alone, it tends to slip. You can double or treble the knot to remedy this. Two half hitches under tension can be difficult to undo without cutting the cord, especially after a few days, if the knot has dried out after being wet.

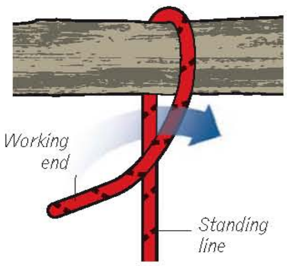
Working end
Standing line

1 Pass the cord around the fixture, such as a bar or tree.

- Pass the working end around the standing line and back through the loop.

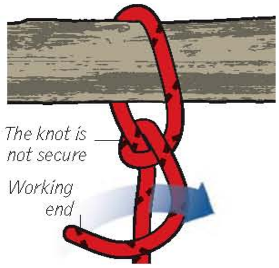
The knot is not secure
Working end

2 Hold the standing line and push the half hitch knot toward the fixture to tighten it.

- Repeat the sequence and tie another half hitch.

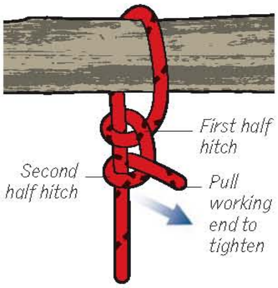
First half hitch
Second half hitch
Pull working end to tighten

3 Push the second knot tight to the first hitch knot for more security.

- If you need to, finish off the working end with an overhand knot (see p. 143).

USING CUTTING TOOLS

ANYONE VENTURING INTO THE WILDERNESS should carry some kind of knife, preferably a bushcraft knife. It's probably the most important survival tool after knowledge. Although a good knife will help you through many survival situations, look at the environment you are traveling to and choose your cutting tools according to their potential uses—for example, you may need a saw as well as a knife.

WARNING!
If you drop a bushcraft knife, never attempt to catch it because there's a very good chance you will lose your fingers!

BUSHCRAFT KNIFE

An experienced person can accomplish most survival skills, such as making feather sticks (see pp. 120–21), with a quality bushcraft knife. This type of cutting tool has a number of crucial features that will stand you in good stead in every situation.

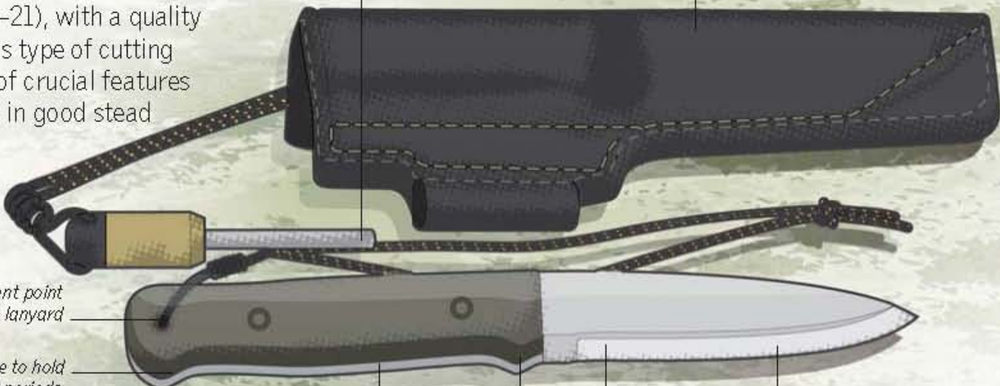
A firesteel that fits into your sheath means that, as long as you have your knife, you can also get a fire going
Fitted leather sheath holds the knife securely
Attachment point for a wrist lanyard
The handle is comfortable to hold and work for extended periods. Avoid ribbed, rubber handles
Blades that extend through to the end of the handle (full tang) have no weak points
Sharpened blade ends near the handle – working close to the handle puts less strain on your wrist
Blade is 3/16 in (4 mm) thick and made of stainless or high-carbon steel
The guard is a ridge in the handle that reduces the chance of your hand slipping on to the blade

LARGE KNIVES

All over the world, people living primitive lives have two essential items: a cooking pot and a large working blade, either a parang or machete.

- Parangs and machetes are similar types of knife. The parang is associated with the Malay people and there are many varieties available. Machetes are more common—you can still see machetes being used in Africa today that were made 40 years ago from old Landrover suspension springs! The kukri is a similar knife used by the Ghurkas of Nepal.
- A parang or machete blade is about 18–20 in (45–50 cm) long and is used primarily for cutting and slashing but, like the ax, can be used for intricate work.
- Always wear a wrist lanyard when using a machete, parang, or kukri—and pay attention to what you are doing.

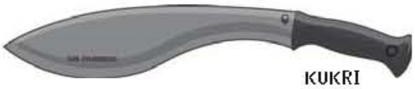
KUKRI

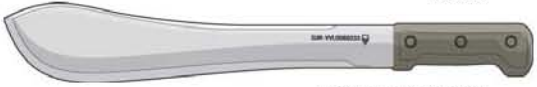
MACHETE/PARANG

KNIFE KNOW-HOW

- A bushcraft knife is used primarily for carving, cutting, and splitting small logs. If you need to chop large logs, use either a machete or a saw.
- A knife doesn't need to be longer than 8–9 in (20–22.5 cm).
- Don't use a knife with a serrated edge when working with wood, as the serrations make it hard to use the part of the blade closest to the handle (essential for making feather sticks, etc).
- Keep your knife safe when not in use. Clean and dry the blade and return it to its sheath.
- Wear your sheath on your belt or around your neck on a lanyard. Belt sheaths may be high (the handle is above the belt) or low (the handle is below the belt, which allows you to sit down without the handle poking you in the side).

SHARPENING A KNIFE

It's very important to keep the blade of your knife sharp. Blunt blades are dangerous because they are likely to bounce or slip off the wood you're working on. Also, blades that have completely lost their cutting edge are very difficult to sharpen.

SHARPENING STONES

Keep a good sharpening stone at home to ensure you always set out with a knife that is sharp.

- In the field, carry a smaller, lightweight device for quickly re-sharpening a blade that is beginning to lose its edge.
- There is a good range of small sharpening stones (medium one side, fine the other) that fit into a neat leather pouch.
- Small automatic sharpeners are available to buy that use two sets of ceramic stones to sharpen a knife pulled between them.

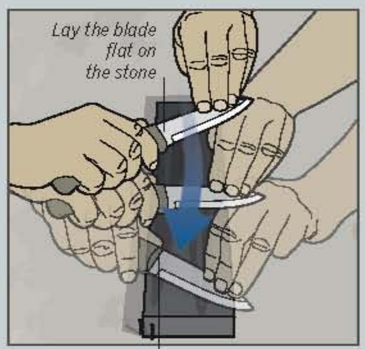

Lay the blade flat on the stone

Raise the back of the blade to lay the edge on the stone

1 Hold the knife in a forehand grip, with the blade pointing away from you. Support the tip of the blade with the thumb and fingers of your other hand.

- Push the knife so the full blade arcs over the stone in one sweep.
- Lift the blade and return to the start, repeating the action 6–8 times.

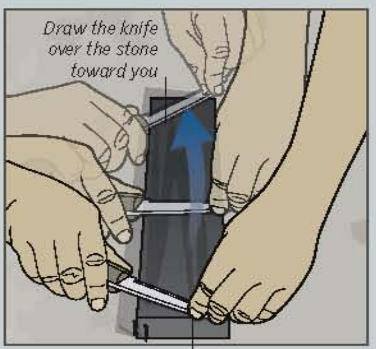

Draw the knife over the stone toward you

Start from the far end of the stone

2 Repeat the sequence for the other side of the blade, but reversing the crucial elements.

- Hold the knife with a backhand grip, with the blade pointing toward you at the far end of the stone.
- Pull the knife in an arc over the stone in one draw. Lift and return to the start, repeating the action 6–8 times.

WORKING WITH YOUR KNIFE

The way you hold your bushcraft knife can help you accomplish a variety of cutting tasks easily and safely. There are several essential grips, including the forehand, backhand, and chest lever. In most cases, you need to grip the handle firmly at the end nearest the blade so that you can cut with the part of the blade nearest the handle.

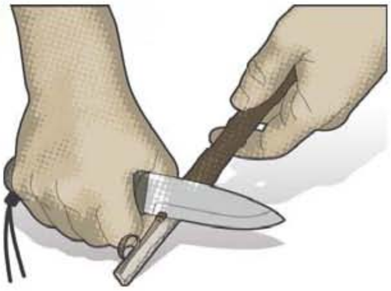

FOREHAND GRIP

This grip is the most natural and the one most often used. It allows your back and arm to provide the power and strength, while your wrist provides the control. Placing one or both thumbs on the back of the blade allows you additional control for finer, more detailed work. Make sure the blade is facing down. Otherwise you are likely to give yourself a serious cut.

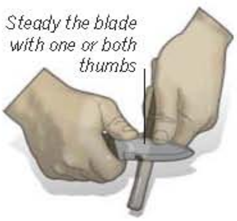

Steady the blade with one or both thumbs

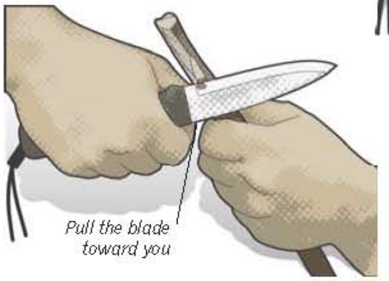

Pull the blade toward you

BACKHAND GRIP

This is the same as the forehand grip, except the blade faces up toward you. You can use it when you want to see and control exactly where the blade is going, or when you're cutting something and a follow-through would injure you—for example, when sharpening a stick.

CHEST LEVER

This is an extremely effective technique in which you hold the knife in a backhand grip and use the muscles of your arm and chest to provide power and strength. You can cut close to your body while your wrists control the cutting action.

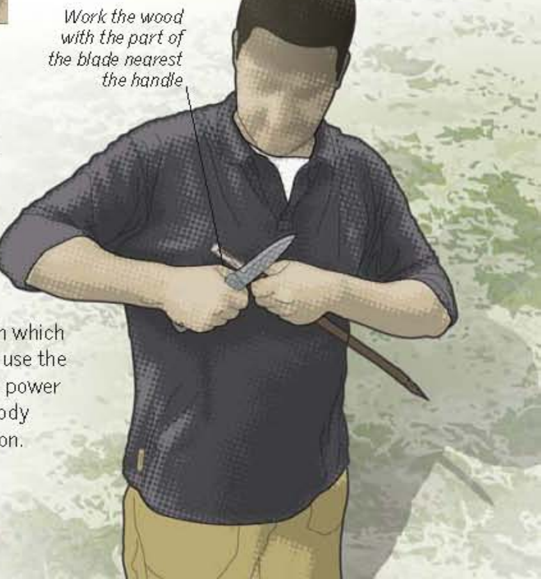

Work the wood with the part of the blade nearest the handle

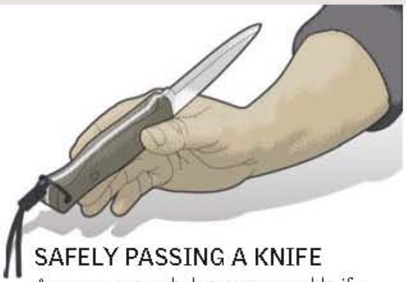

SAFELY PASSING A KNIFE

A group may only have one good knife for everybody to use, so you'll need to pass it to each other frequently. Pass the knife handle first, with the blade up. This simple and safe method is easily learned and can be adopted by everyone.

TREE-FELLING WITH A BLADE

For shelters and fuel, the largest tree you are likely to cut down will be about 4 in (10 cm) in diameter. For anything larger you'd need an ax, but for smaller ones use a saw (see pp. 150–51). If you don't have a saw, use the blade of your bushcraft knife or a machete or parang. Whichever blade you use, keep its lanyard around your wrist.

1 Figure out the lean of the tree and start your first cut on the opposite side. This stops your blade jamming in the cut as the tree leans further.

- Hit the top of the blade with a thumping stick (club).
- Move the blade about ⅝ in (15 mm) to one side of the first cut, and then angle the blade at 45 degrees toward the first cut.
- Continue to alternate this action from side to side, cutting a V into the wood.

WARNING!
Ensure your blade is sharp. Keep what you're cutting between you and the blade, with the blade pointing away.

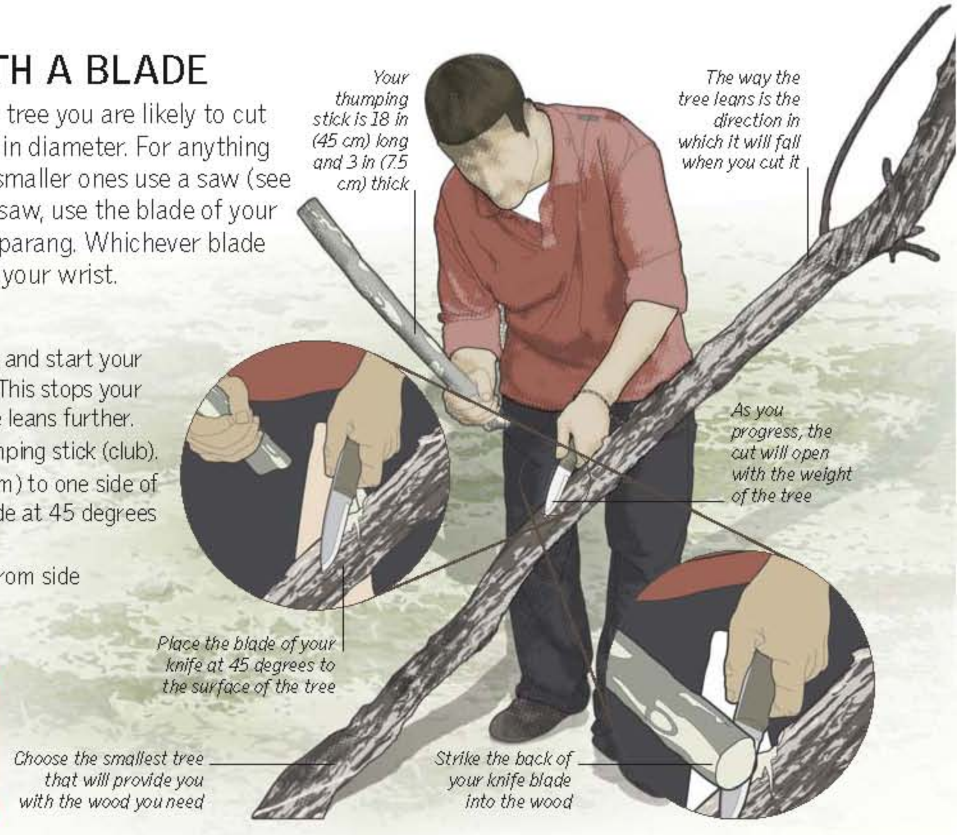
Your thumping stick is 18 in (45 cm) long and 3 in (7.5 cm) thick
The way the tree leans is the direction in which it will fall when you cut it
As you progress, the cut will open with the weight of the tree
Place the blade of your knife at 45 degrees to the surface of the tree
Choose the smallest tree that will provide you with the wood you need
Strike the back of your knife blade into the wood

SPLITTING FIREWOOD

When you split wood into fuel for a fire, use the largest blade you have, work on a solid surface, and avoid wood that has large knots as these prevent the blade from making a clean cut.

SAFE WORKING AREA
Chop wood in a safe area that other people won't wander into and where there are no snagging hazards or objects to impede your swing and knock your knife off-target.

LARGE LOGS

If you only have a bushcraft knife to work with, and you need to split large logs into smaller pieces, the following method works well.

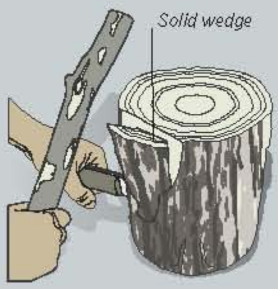
Solid wedge

1 Hit the back of your knife with a stick and cut a solid wedge from the side of the log.

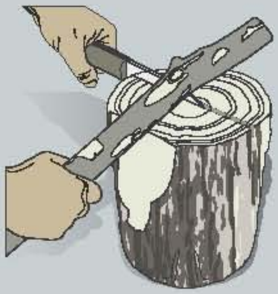

2 Use your knife and a thumping stick to start the initial cut across the top of the log.

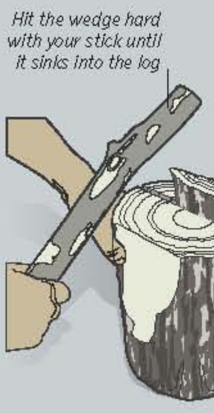
Hit the wedge hard with your stick until it sinks into the log

3 Hit the wedge into the initial cut. Keep doing this until the log splits. Use more wedges if you need to.

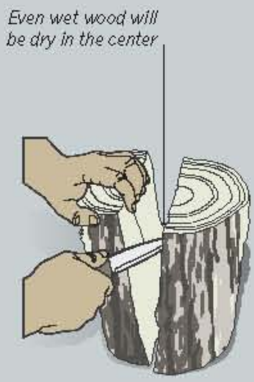
Even wet wood will be dry in the center

4 If the log does not split completely, force it open with your hands or pry it with your blade.

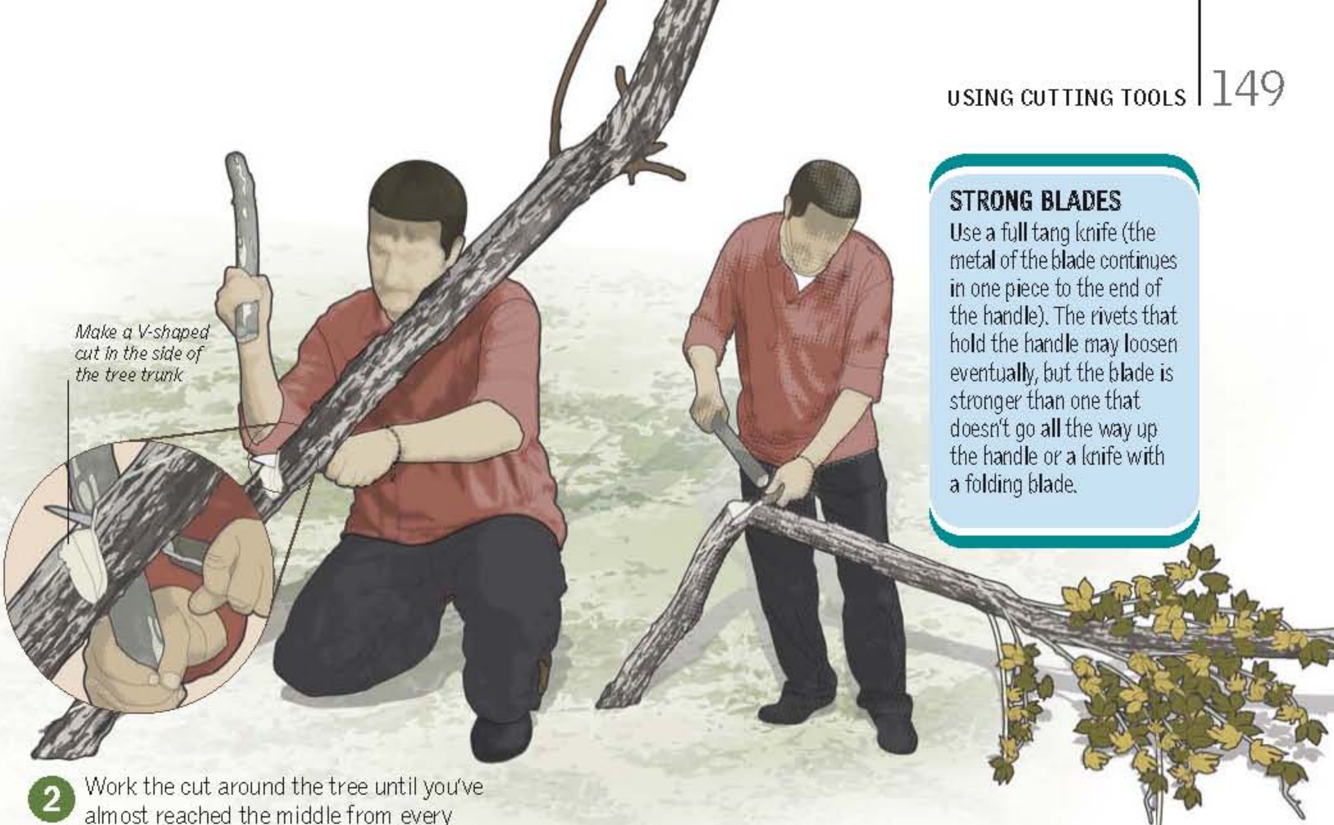

Make a V-shaped cut in the side of the tree trunk

STRONG BLADES

Use a full tang knife (the metal of the blade continues in one piece to the end of the handle). The rivets that hold the handle may loosen eventually, but the blade is stronger than one that doesn't go **all** the way up the handle or a knife with a folding blade.

② Work the cut around the tree until you've almost reached the middle from every direction except from below.

- If you try to cut completely through the wood from one side it will get increasingly difficult as the blade goes deeper.

③ Once the cuts are deep enough to weaken the wood, you will be able to snap the tree in two.

- Take care that the tree doesn't fall on your feet.

SMALL LOGS

Splitting a small log can be more tricky than splitting a large log, because it's difficult to balance it safely while you hit the blade with your thumping stick.

① Place the log on a solid surface and put your knife in the middle of the end of the log you're about to chop.

- Hit the back of the blade firmly with your stick.

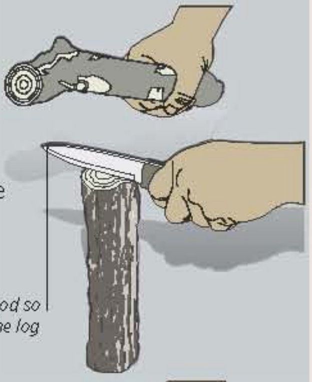

Put the blade in the wood so you can steady the log

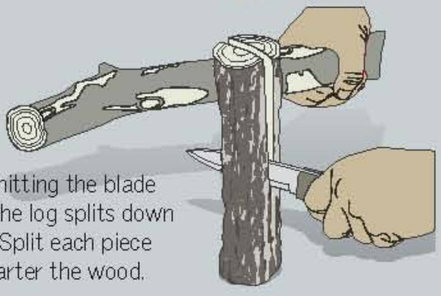

② Keep hitting the blade until the log splits down the middle. Split each piece again to quarter the wood.

④ To trim the branches from the fallen tree, start at the bottom and work your way to the top. This allows you a clear, short swing for your thumping stick.

- Keep the tree between you and the blade, and trim each branch in the direction of its growth.
- If you're using a machete, you don't need a thumping stick, as the heavy blade has enough momentum of its own.

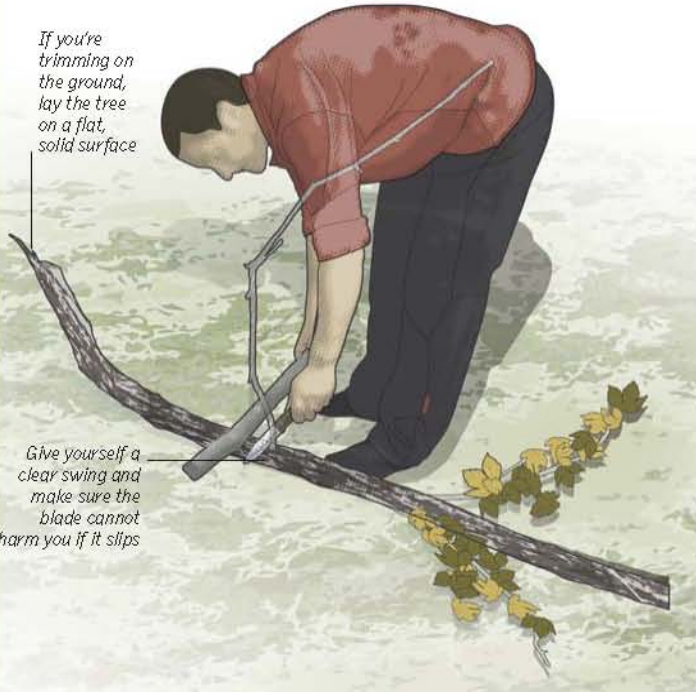

If you're trimming on the ground, lay the tree on a flat, solid surface

Give yourself a clear swing and make sure the blade cannot harm you if it slips

USING A SAW

A saw is much better than a bushcraft knife for larger jobs, such as preparing logs for a fire or cutting down the poles needed to build a shelter. There are several types of pocket saw available that are lightweight and take up very little room in your backpack or survival tin.

USING A POCKET CHAINSAW

Using a pocket chainsaw to cut down a small tree is safe and not especially hard work. You can use the same saw to trim the branches and then cut the wood into manageable sections.

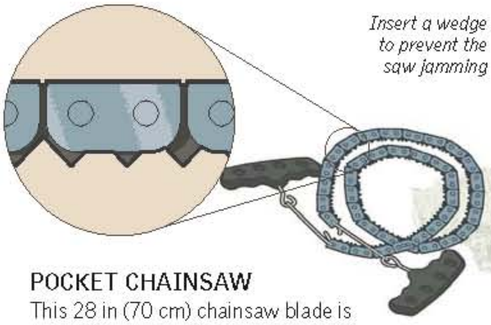

POCKET CHAINSAW
This 28 in (70 cm) chainsaw blade is compact and has attachment rings and handles. It can cut through a 3 in (7.5 cm) diameter tree in under 10 seconds.

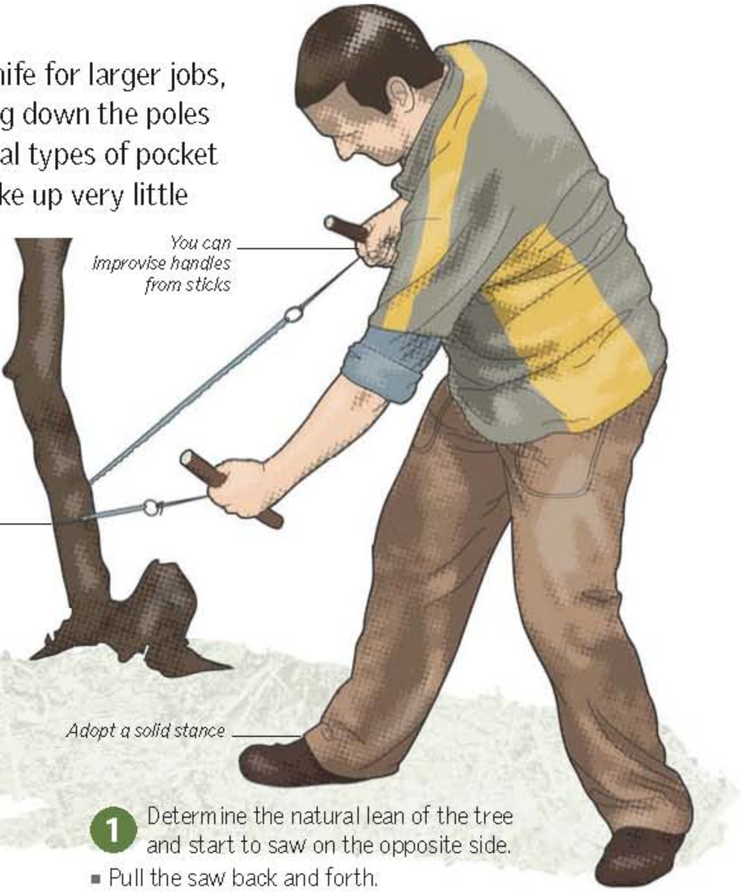

You can improvise handles from sticks
Insert a wedge to prevent the saw jamming
Adopt a solid stance

1 Determine the natural lean of the tree and start to saw on the opposite side.

- Pull the saw back and forth.
- Knock a wedge in behind the saw when halfway through to stop the tree from jamming the blade.

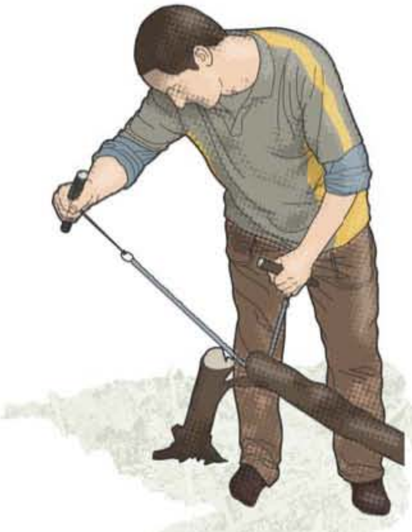

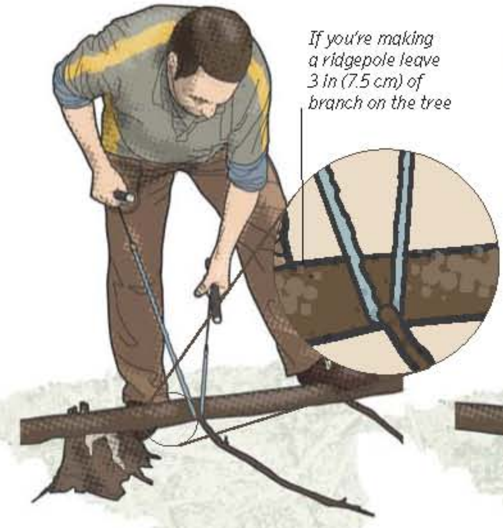

If you're making a ridgepole leave 3 in (7.5 cm) of branch on the tree

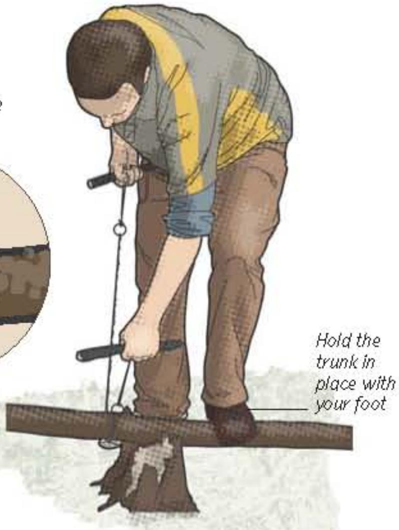

Hold the trunk in place with your foot

2 Slow down when the saw has nearly cut through the tree.

- Allow the weight of the tree to snap the cut and break it in two.

3 Use the same back-and-forth rhythm to trim the branches.

- Clear the branches away as you trim them from the tree.

4 Saw the trimmed trunk into a measured section.

- When you're halfway through, prop up the trunk to make sawing easier.

FOLDING SAWS

Folding saws are strong, light, and convenient. Some, such as the Sandvik Laplander, have a serrated blade that folds into the handle (see right). Others, such as the backpacker's (see below), can be dismantled for stowing.

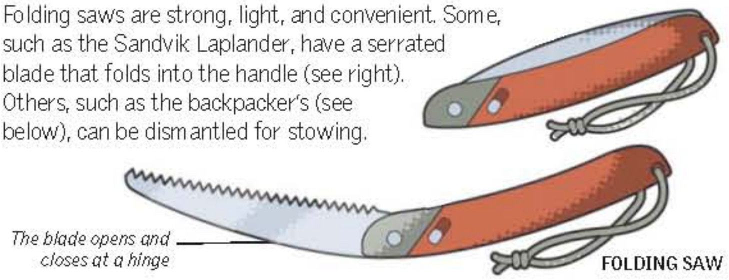
The blade opens and closes at a hinge

FOLDING SAW

BACKPACKER'S FOLDING SAW

These saws are generally lightweight and strong, but are not as effective as the pocket chainsaw. If space is tight, then you can pack only the blade and improvise a bow saw, using natural materials found in the woods.

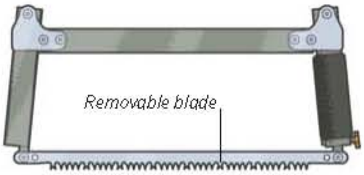
Removable blade

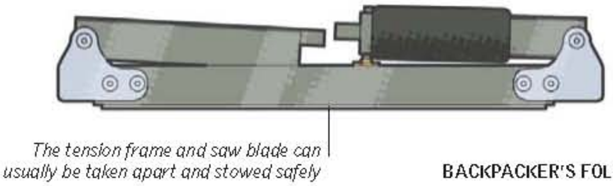
The tension frame and saw blade can usually be taken apart and stowed safely

BACKPACKER'S FOLDING SAW

COMMANDO SAW

This saw takes up very little space and consists of a serrated wire blade with a swivel ring at each end. It's a useful piece of survival equipment, although it tends to snag and snap easily unless you keep it completely straight. You can easily convert it into an improvised bow saw, and, with a little more effort, make it into a very effective tension saw.

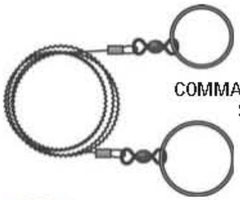

COMMANDO SAW

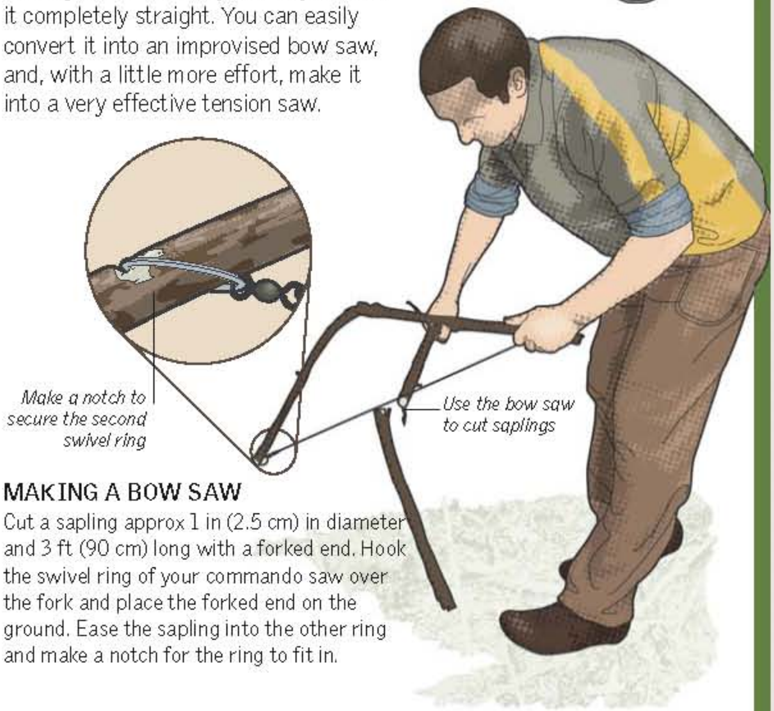
Make a notch to secure the second swivel ring
Use the bow saw to cut saplings

MAKING A BOW SAW

Cut a sapling approx 1 in (2.5 cm) in diameter and 3 ft (90 cm) long with a forked end. Hook the swivel ring of your commando saw over the fork and place the forked end on the ground. Ease the sapling into the other ring and make a notch for the ring to fit in.

IF YOU HAVE NO TOOLS

Even without any cutting tools, it's still possible to safely break wood down to size. The answer is snapping—or even burning—the wood. Don't be tempted to stand branches and sticks against a log or rock and then stamp on them in the middle. In a genuine survival situation your feet are your only means of transportation so don't do anything that risks injuring them.

SNAPPING THE WOOD

You need to limit the size of wood you're trying to break—don't try to snap branches that are too broad. Use wood that's dead, because green living wood just bends and refuses to break completely.

- Find two trees that are 1–2 ft (30–60 cm) apart—or a tree with two trunks close together—to use as a fulcrum for the wood. Place the wood in the fulcrum.
- Adopt a firm and steady stance and, with one foot in front of the other, pull the wood toward you until it breaks. Pulling toward you keeps your center of gravity over your spread feet, while pushing away from you causes you to stumble forward when the wood breaks.

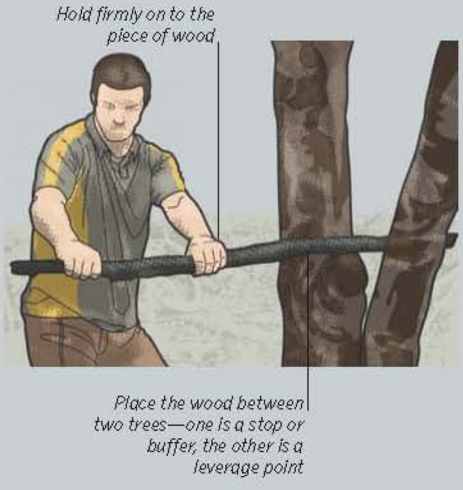
Hold firmly on to the piece of wood
Place the wood between two trees—one is a stop or buffer, the other is a leverage point

BURNING THE WOOD

If you have a fire going, place the wood on the fire where you want it to break. Leave it on the fire until it burns all the way through. Alternatively, when it has burned halfway through, use the snapping method described above—but don't work on the wood while it's still burning or hot.

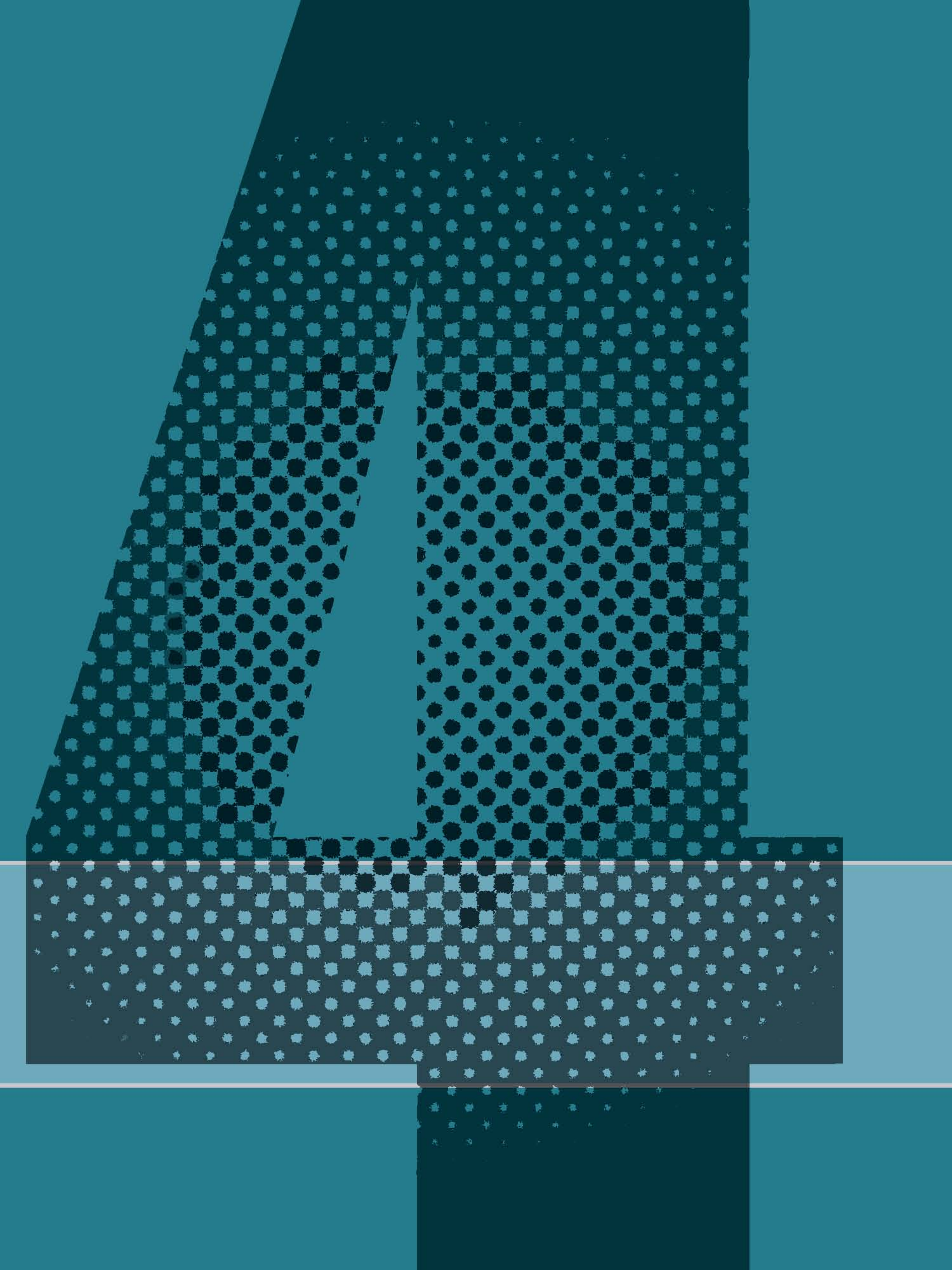
4

TAKING SHELTER

TAKING SHELTER

A shelter is your primary means of protection in a survival situation. It can be somewhere to simply keep dry during a sudden downpour or a place to spend several nights while out in the wilds. It's important, therefore, that you understand how to correctly construct your shelter. After all, its effectiveness could mean the difference between relative comfort and abject misery!

In a survival situation, it's important to play a proactive role in your own continued survival and rescue. You can only achieve this effectively if you're in a good condition—both physically and mentally. A day or night spent exposed to the elements will determine how well or how badly you perform the next day. A restless and uncomfortable night, for example, will lead to sleep deprivation, which can make you moody, irritable, and easily frustrated. This in turn can lead to a lack of concentration and irrational thinking—none of which are helpful in a survival situation.

Cold nights spent in the wilderness seem to last longer than they should, so time spent ensuring that you can

In this section YOU WILL DISCOVER...

- when to **lie low** in a hollow or **camp in a cave...**
- how to **modify** your **poncho** to put up a **shelter...**
- the difference between a **bothy** and a **bivy...**
- how to **bed down** with **duck down** or **wake up** in a **wickiup...**
- the **importance** of using your **noggins** to **secure** your shelter...
- how to **configure** your **fig leaves** to keep out the **rain...**

always enjoy as warm, dry, safe, and comfortable a night as possible is rarely wasted.

With shelter-building—as with other areas of survival—you should always apply the principle of expending the least amount of energy for the maximum amount of gain. This could mean simply exploiting the options nature has already provided, such as caves or hollows. Finally, plan for the worst conditions—it may be sunny when you build your shelter but pouring with rain at three in the morning!

On any expedition, you should aim to carry equipment—such as a basic bivy or a tarpaulin—that could be used to form a shelter should you need to protect yourself from the elements.

A SPACE BLANKET, however, is quite simply an essential piece of gear—if you pack nothing else, make sure you carry one and keep it with you at all times. Space blankets pack down very small and are available with one side silver (to reflect heat) and one side bright orange (to aid location). No matter what environment you're exploring, a space blanket can offer you immediate protection from the elements.

With a little ingenuity, a space blanket can also be converted into a basic shelter. You will find instructions on how to build a variety of shelters throughout this chapter, but always remember the following general tips, too:

- First look for natural sheltering places
- Look out for hazards such as signs of flooding
- Construct your shelter well before nightfall
- Clear the ground before you start to build
- Position the entrance at right angles to the wind
- Build for the worst possible weather conditions
- Make sure your shelter is sturdy and secure
- Don't over-exert yourself as you build
- Raise your sleeping area off the ground
- If on a slope, sleep with your head higher than your feet

> “The **effectiveness of a shelter** can mean the difference between relative comfort and **abject misery**”

HOLLOWS AND CAVES

IF YOU DON'T HAVE a shelter of your own, look for shelter opportunities that nature has already provided for you, such as a cave or a natural hollow in the ground. However, avoid natural hollows in low-lying areas that may flood and, if you're on a slope, keep well away from hollows exposed to run-off water.

TOOLS AND MATERIALS

- Sturdy digging stick
- Strong branches (or poles) and logs
- Saw, such as pocket chainsaw
- Tree bark, turf, leaves, forest mulch, moss, bracken, and pine boughs for the roof

BEDDING

Cover the floor of a shelter with some form of bedding to prevent your body from losing heat to the cold ground through conduction. After all the energy you've used to build the shelter, why waste it on an uncomfortable night? Use the driest materials you can find (see list below).

MAKE YOURSELF COMFORTABLE

Here are some ideas to help you enjoy a good night's sleep in your shelter:

- Collect twice as much bedding as you think you need: 6 in (15 cm) of holly, for example, compresses to about 2 in (5 cm) after you've been lying on it for a while.
- Place logs around the bedded area to stop the bedding from shifting and spreading.
- If the ground slopes, sleep with your head higher than your feet.

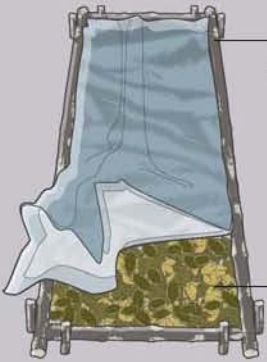

Use wooden pegs to keep the side logs in place

Bedding keeps you warm

BEDDING MATERIALS

- Feathers, such as duck and goose down, are the best bedding as they retain warmth but not moisture.
- You're unlikely to find enough feathers, so also try pine and spruce boughs, dry leaves, moss, bracken, grasses, and holly (cover it to prevent it from pricking you).

NATURAL HOLLOWS

A hollow that may be too shallow to sit in can still provide shelter from the wind if you lie in it—and, with a basic roof, keep you dry too.

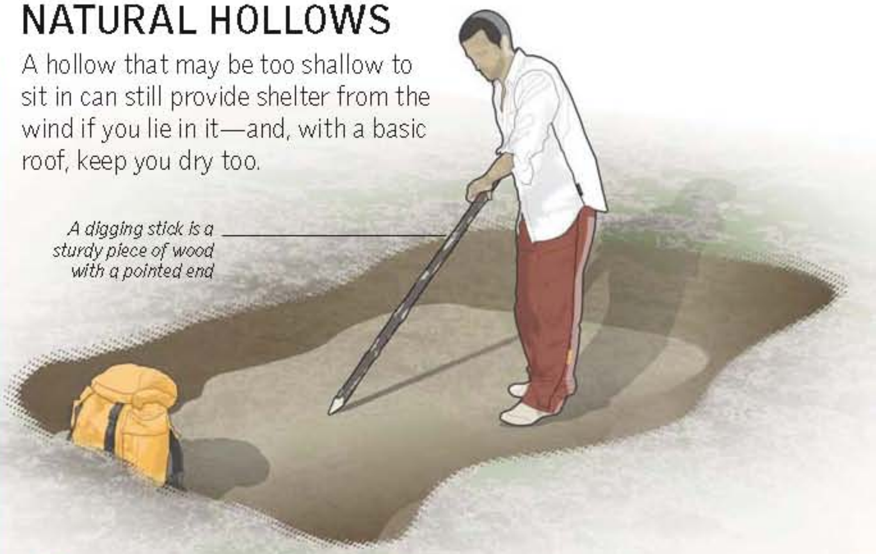

A digging stick is a sturdy piece of wood with a pointed end

1 If the hollow isn't deep enough to lie in, remove some soil with a digging stick.

- Beware wet soil, as it could be prone to flooding.
- If the ground is too hard to dig, build up the sides with logs for extra depth.
- Lay bedding material on the ground before continuing.

Lay a log on top of the cross poles and along the length of the hollow

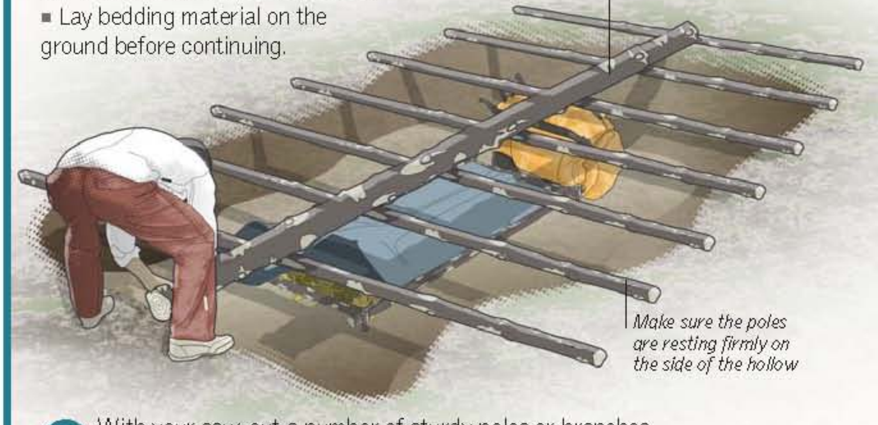

Make sure the poles are resting firmly on the side of the hollow

2 With your saw, cut a number of sturdy poles or branches (see pp. 150–51) that are long enough to place across the hollow to create a supporting roof.

- Cut a thicker, longer log and place it across the poles. This will create height and slope for a pitched roof. Make sure the log isn't too heavy.

TAKING SHELTER IN CAVES

Caves are ready-made shelters that are usually dry and secure. There are some dangers, including animals, poor air quality, and water. Many caves are outlets for streams or are connected to pools. Don't go any further in than you can see, as there may be hidden drops, slippery surfaces, and crevasses. Avoid old and disused mines, as they may be prone to collapse.

ANIMAL DANGERS

Many animals, such as bears, bats, insects, spiders, and snakes, shelter in caves, so check the ground around the entrance and inside for signs of tracks and nests. Vampire bats can be dangerous as they are known to carry rabies. Bat droppings (guano) form a thick layer on the ground and, in large quantities, can be highly combustible.

POOR AIR QUALITY

If a cave makes you feel light-headed or nauseous, leave at once because the air may be foul due to excess carbon dioxide. Other warning signs are an increased pulse and breathing rate. If a flame starts to dim, or turn blue, leave the cave immediately—this may not be a sign of excess carbon dioxide but a lack of oxygen, which can be worse.

LIGHTING FIRES

If you build a fire at the front of a cave the smoke can blow back inside and the fire can block your exit if it gets too big. Build a fire toward the rear of the cave, but make sure that there is a sufficient flow of air for the smoke to escape. You can erect a small fire reflector or make a pile of rocks at the entrance to reflect heat back into the cave.

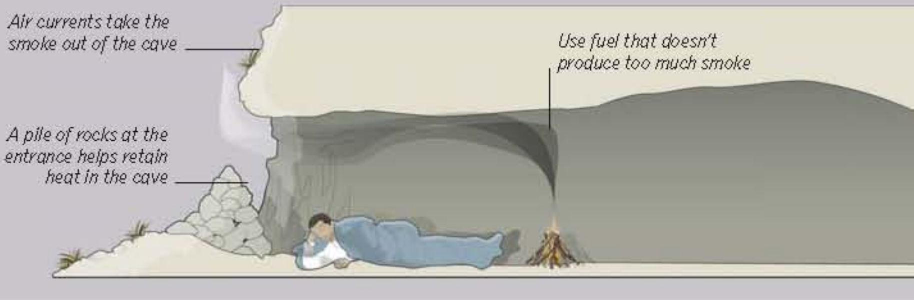
Air currents take the smoke out of the cave

A pile of rocks at the entrance helps retain heat in the cave

Use fuel that doesn't produce too much smoke

COASTAL CAVES

Coastal caves share the same dangers as any other cave, but they are also subject to the tides, which can rise very quickly—in some areas they can advance hundreds of yards in a matter of minutes.

THE DANGERS OF COASTAL CAVES

When you're assessing a coastal cave, look out for the following warning signs of flooding during high tide:

- A line of seaweed, driftwood, or flotsam and jetsam—either in the cave or further back on the beach either side of the entrance.
- A damp or wet smell. The cave may even look wet if the tide has recently gone out.
- Rock pools in the cave or in front of it.

3 Place shorter branches, sticks, or poles over the top of the log and cross branches to create the pitched roof.

- Pack these materials tightly together and make sure they're firmly in position.

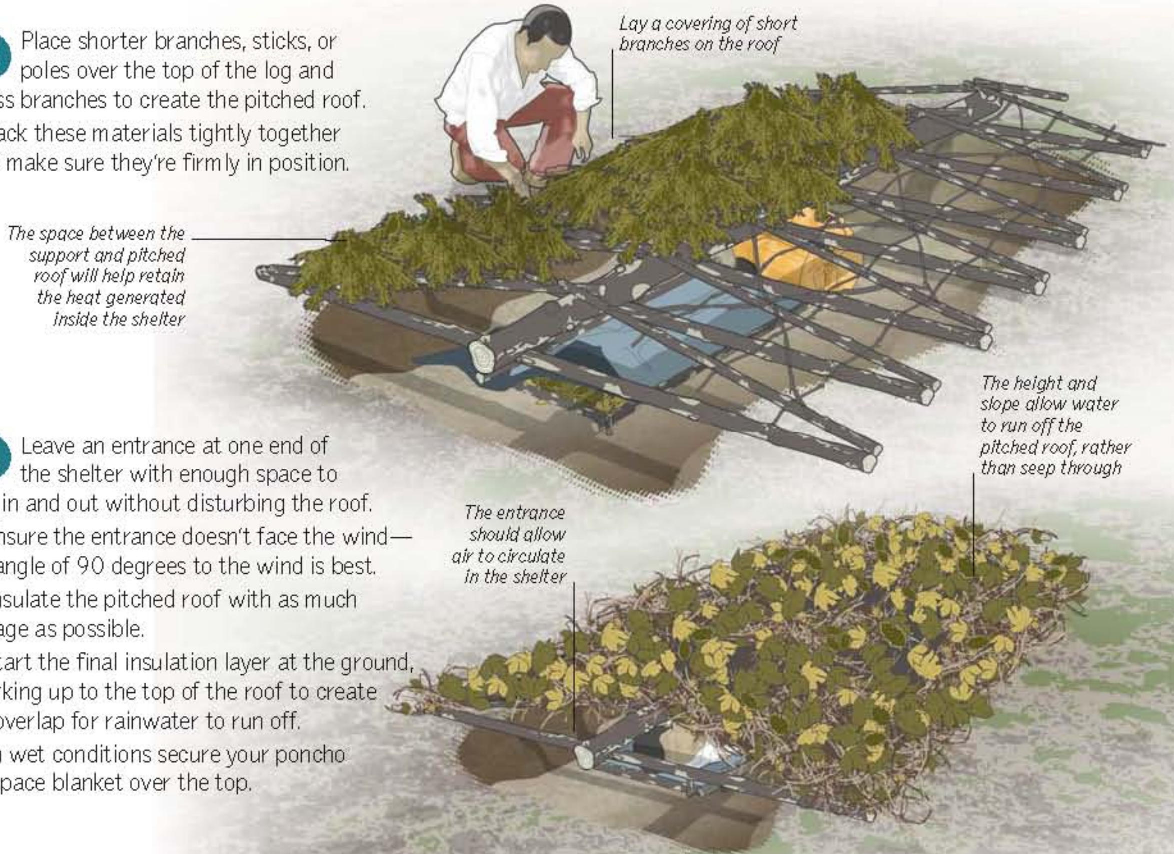
Lay a covering of short branches on the roof

The space between the support and pitched roof will help retain the heat generated inside the shelter

The height and slope allow water to run off the pitched roof, rather than seep through

The entrance should allow air to circulate in the shelter

4 Leave an entrance at one end of the shelter with enough space to get in and out without disturbing the roof.

- Ensure the entrance doesn't face the wind—an angle of 90 degrees to the wind is best.
- Insulate the pitched roof with as much foliage as possible.
- Start the final insulation layer at the ground, working up to the top of the roof to create an overlap for rainwater to run off.
- In wet conditions secure your poncho or space blanket over the top.

USE-ANYWHERE SHELTERS

WITH A PONCHO OR BIVY SACK (see pp. 160–61), you can make a shelter for use in any environment and in any conditions. You can also adapt any type of material—a survival space blanket, ground sheet, or tarpaulin—in the same way.

TOOLS AND MATERIALS

- Poncho, bivy sack, groundsheet, space blanket, or tarpaulin
- Cordage or elastic bungees
- Pegs; stick or pole
- Rock or log for hammering
- Survival knife or penknife

ONE-POLE PONCHO

This shelter is very quick and easy to erect and take down—useful when energy levels are low or when you just want to get out of the rain. If you need to be rescued, put the most visible color on the outside to increase the chance of being spotted.

1 Cut three 2 ft (60 cm) lengths of cordage and one 6 ft (2 m) length.

- Tie the shorter lengths into loops using an overhand knot (see p. 143).
- Tie the long length to one corner with a Siberian hitch knot (see p. 144).
- Fasten the cordage loops to the corner grommets.

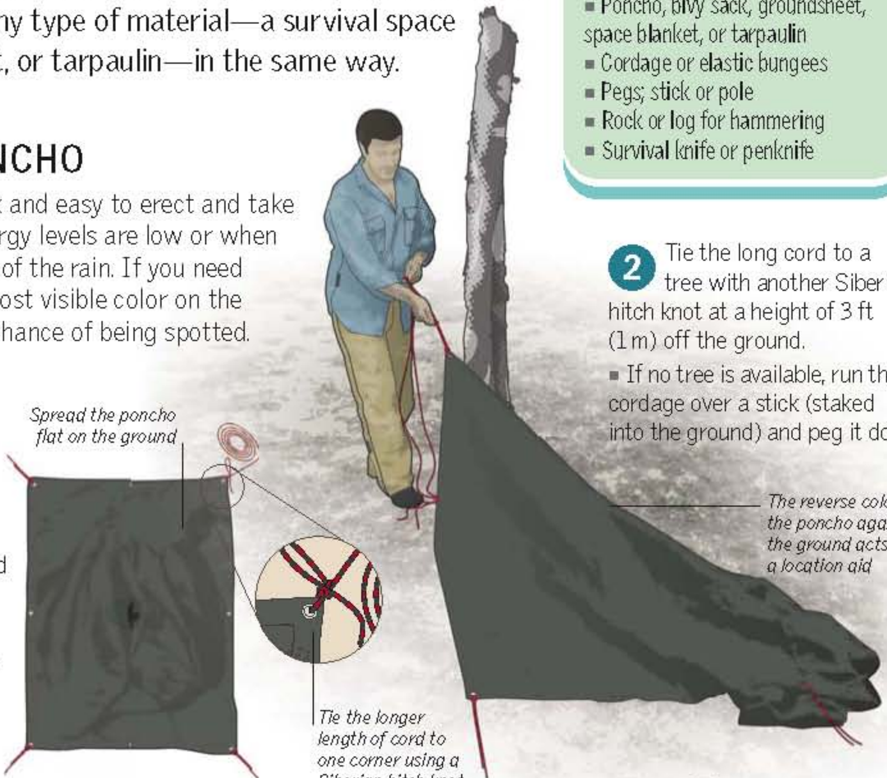

Spread the poncho flat on the ground

Tie the longer length of cord to one corner using a Siberian hitch knot

The reverse color of the poncho against the ground acts as a location aid

2 Tie the long cord to a tree with another Siberian hitch knot at a height of 3 ft (1 m) off the ground.

- If no tree is available, run the cordage over a stick (staked into the ground) and peg it down.

Make sure the top of the stick in the poncho is secure

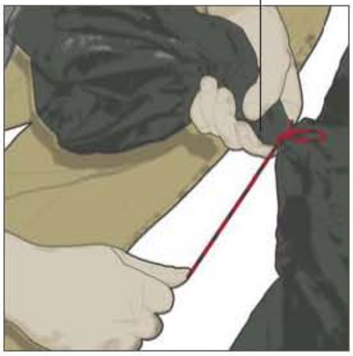

3 Place one end of a 3-ft (1-m) long stick in the hood of the poncho.

- Seal the hood by tying a simple knot.
- Wind several turns of cordage around the knot and secure it with an arbor knot (see p. 142).

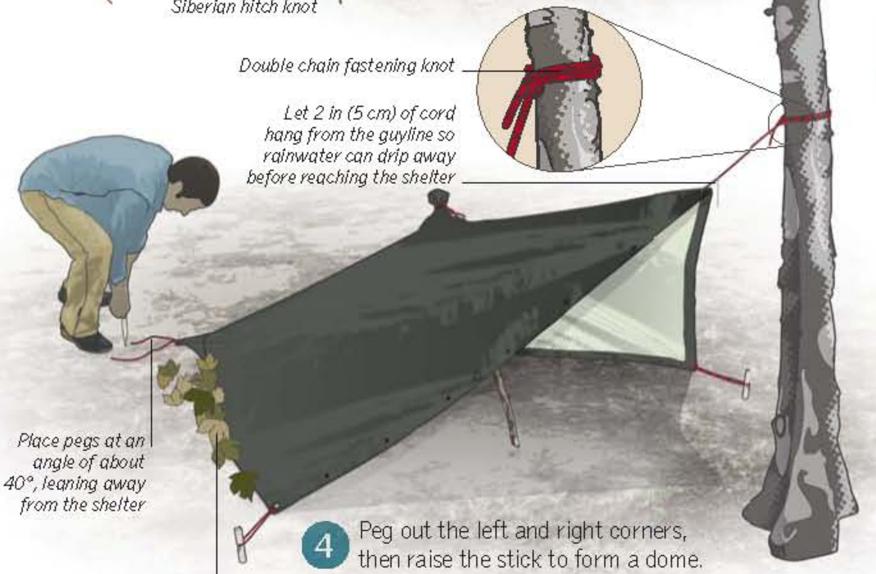

Double chain fastening knot

Let 2 in (5 cm) of cord hang from the guyline so rainwater can drip away before reaching the shelter

Place pegs at an angle of about 40°, leaning away from the shelter

Pile leaves around the edges of the shelter to keep out the wind

4 Peg out the left and right corners, then raise the stick to form a dome.

- Pull the rear corner taut and peg it out to add tension, making the structure rigid.
- To enter the shelter, curl your body around the center pole.

CORDED A-FRAME

This poncho A-frame creates a tentlike shelter between two trees. If you have enough cordage, a single ridgeline can be tied between the trees; if not, use two cordage loops, as shown below. Alternatively, you can use a ridgepole at the same height.

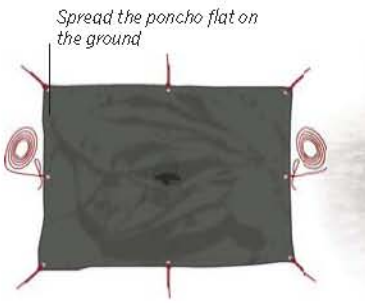
Spread the poncho flat on the ground

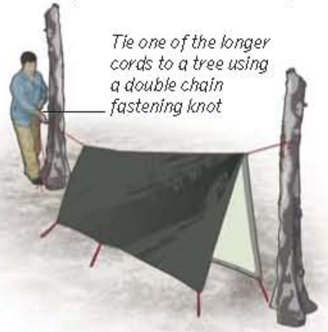
Tie one of the longer cords to a tree using a double chain fastening knot

1 Find two trees 2 ft (60 cm) further apart than the length of your poncho.

- Attach cordage loops 1 ft (30 cm) long to each grommet on the longer sides of your poncho and 3-ft (1-m) loops to the end ones.
- Tie a knot around the hood to seal it.

2 Tie the end of one of the longer cords to a tree with a double chain fastening knot (see p. 143) at a height of 3 ft (1 m).

- Tie the other long cord to the second tree with a taut line hitch knot (see p. 145) and adjust the hitch to pull it tight.

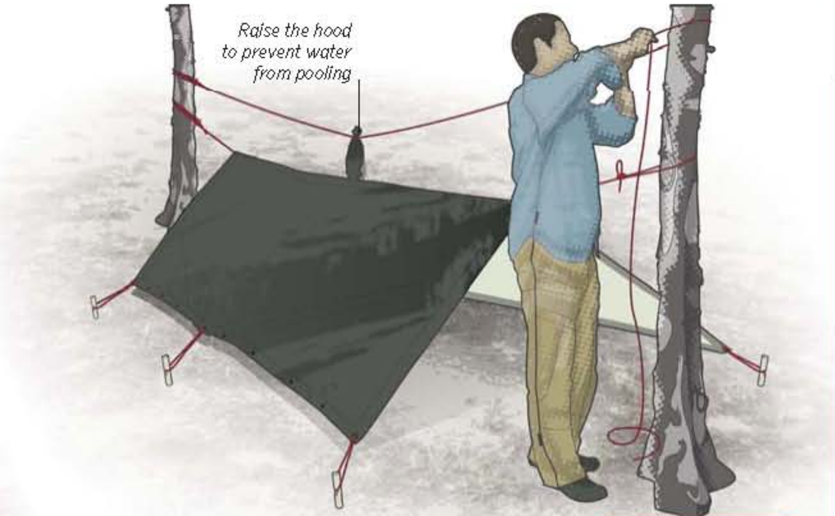
Raise the hood to prevent water from pooling

3 Peg the middle loop of each side to pull the poncho taut. Peg out the four remaining loops, pulling tight as you go.

- If you have enough cordage, raise the hood by tying it either to a second horizontal line or to a vertical line passed over a branch above.
- Place long boughs along each side of the shelter, with shorter boughs at each end to hold in bedding material.

NEED TO KNOW

- Block one end of the shelter with branches, boughs, and mulch to help retain heat.
- If your material has no grommets, use button ties instead (see p. 161).

TENT PEGS

You can make your own tent pegs and then either reuse them or replace them with new ones when necessary. Discard old pegs that naturally decompose over time.

MAKING TENT PEGS

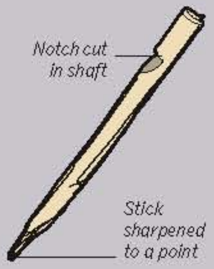
Notch cut in shaft
Stick sharpened to a point

- Pegs should be about 9 in (22.5 cm) long and 1 in (2.5 cm) in diameter. Anything less will either snap under pressure or pull out in high winds.
- Cut wood for the pegs from green trees.
- Don't use wood from the ground as it will be in some stage of decay.
- Dry out new pegs slowly over the embers of a fire to harden them. Remove them when they're a light brown color. Wet wood sounds deep and dull when you tap it; as it dries the sound gets higher and crisper.
- Never drive a peg in with your foot. Use a stone or a heavy stick instead. If you misjudge it, the peg may stab you through the ankle or pierce your boot and even your foot.

ALTERNATIVE DESIGN

If you're in a group and you each have a poncho, you can clip two ponchos together and use one as a built-in floor.

USING TWO PONCHOS

First clip the poppers of both ponchos on their two long sides to form a tube. The top one is the A-frame, the bottom one the groundsheet.

- Pack all your bedding material under the bottom poncho and hold it in place by laying branches at either end.
- Assemble the shelter with the top poncho in the same way as the single poncho version.

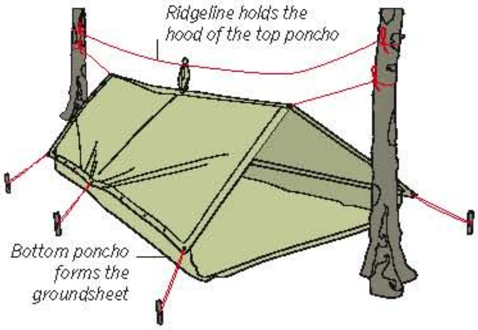
Ridgeline holds the hood of the top poncho
Bottom poncho forms the groundsheet

BIVY PUP TENT

A bivy sack is a lightweight, waterproof cover for a sleeping bag and is made out of a breathable material that cuts down condensation. It's an essential component of ultralight camping (see p. 52) and can very easily be turned into a single-person pup tent. The advantage of making a tent from the bivy sack is that, in wet conditions, the rain runs off the side and doesn't seep to the inside as quickly as it does when the bag is flat on the ground.

TOOLS AND MATERIALS

- Bivy sack
- Saplings
- Button ties
- Wooden pegs
- Cordage and knife
- Rock or log for hammering

1 Lay the bivy sack flat on the ground where you intend to set it up.

- If you have the flap of the hood on top it will form a door that hangs down. However, if the flap is on the bottom you can tie it up as a door.
- Attach a button tie (see panel, opposite) with a loop to each of the bottom corners of the bivy sack.
- Peg the button-tie loops into the ground.

2 Attach a button tie with a long length of cordage to the center of the bivy sack's opening and hold it up.

- Pull down the right and left sides of the opening and place a button tie and loop where they meet the ground.
- Peg out the button-tie loops into the ground.

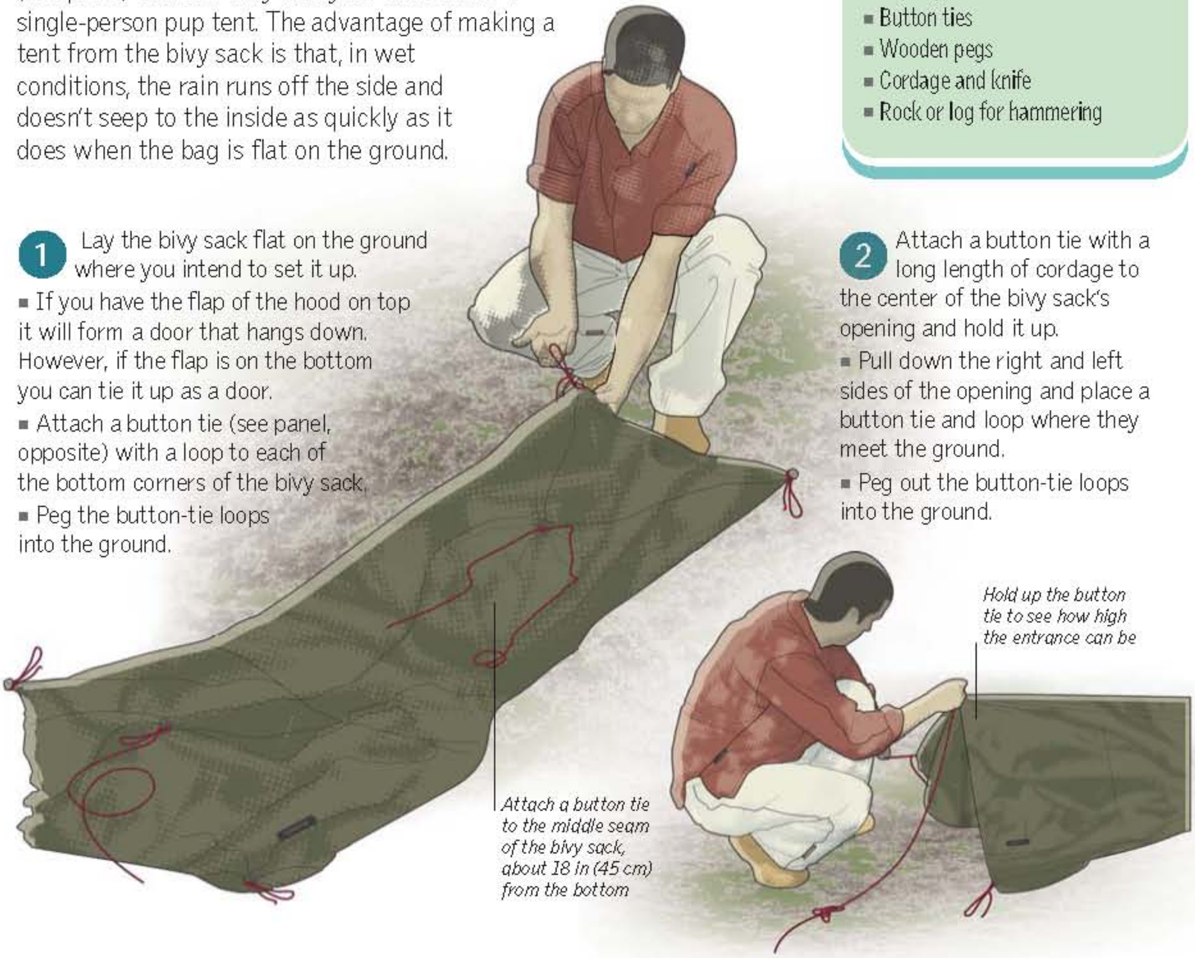

Attach a button tie to the middle seam of the bivy sack, about 18 in (45 cm) from the bottom

Hold up the button tie to see how high the entrance can be

4 Cut a pole the same height as the entrance (you could use a hooped sapling or crossed sticks instead).

- Stand the pole at the entrance, run the middle button-tie cord over the top, and peg out the cord tightly.
- Readjust all the pegs to ensure tight and even tension.

5 When you want to get into the bivy pup tent, simply remove the opening pole and climb in feet first.

- Make sure you avoid the long cord as you wriggle in.

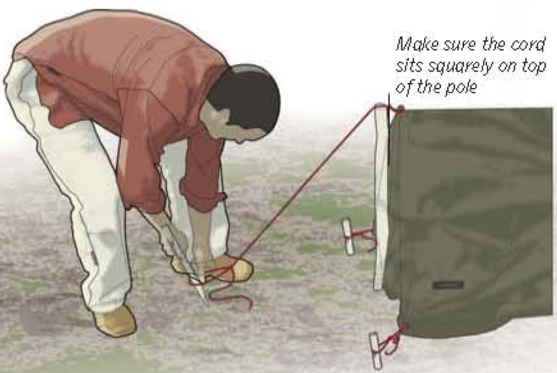

Make sure the cord sits squarely on top of the pole

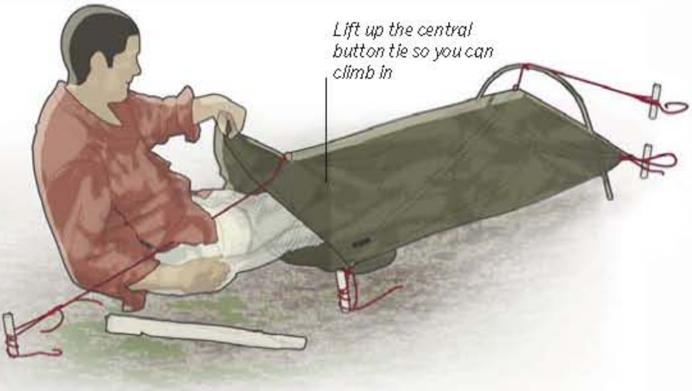

Lift up the central button tie so you can climb in

OTHER SHELTERS

There are other temporary shelters you can erect by adapting the techniques, tools, and materials outlined for both the use-anywhere shelter and the bivy pup tent. Which shelter you choose depends on whether you have a poncho, tarpaulin, or bothy bag.

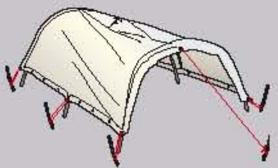

HOOPED SHELTER

If there are saplings or soft wood trees, you can form three hoops and peg out your poncho over them.

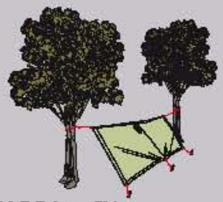

TARPAULIN

Attach one side of the tarp to a line between two trees and peg the other end down to the ground at an angle.

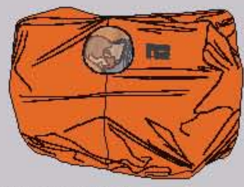

BOTHY BAG

A bothy bag, or storm shelter, is a large, light, nylon bag that protects one or two people from the elements.

3 Place a hooped sapling (or two crossed sticks) over the bottom of the bivy sack.

- Attach the button tie that's 18 in (45 cm) from the foot of the bivy sack to the hoop to raise it and provide more space for your feet.
- Tightly peg out the cord from the hoop.

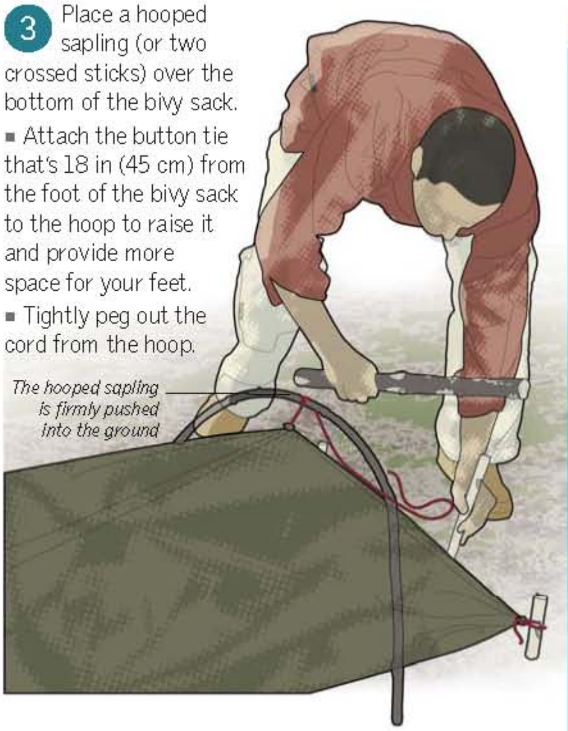

The hooped sapling is firmly pushed into the ground

6 After you have climbed in, replace the pole under the long cord.

- Use the hood of the bivy sack as a door and open and close as necessary.

Be sure to replace the pole in a vertical position with the cord squarely on top

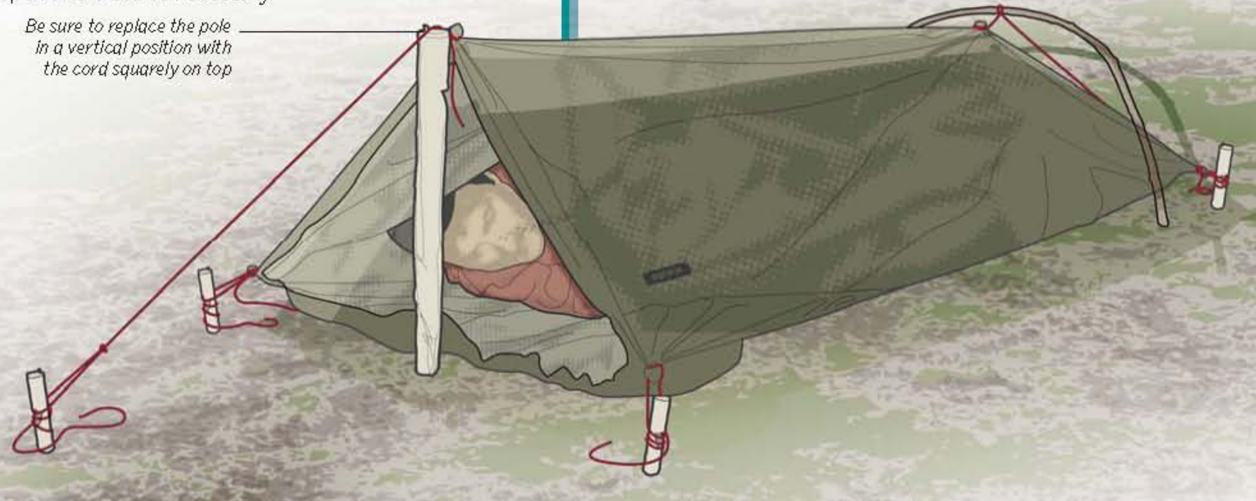

MAKING A BUTTON TIE

Button ties provide a simple, secure fastening for staking out ponchos and tarpaulins that have no loops or grommets to which you could attach a line. As you don't have to cut holes in the material, it remains waterproof and less prone to ripping.

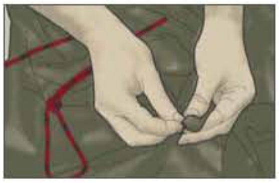

1 Encase a small, round, smooth stone (or similar) in material to form a "button". Prepare cord, with the open loop of a slip knot (see p. 142) at one end.

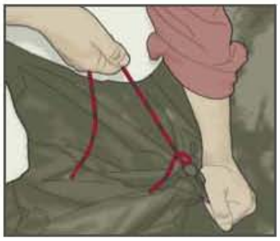

2 Place the open loop of the slip knot over the neck of the button and pull it tight.

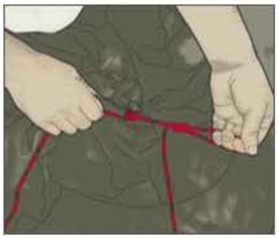

3 At the other end of the cord make a simple loop to place over your shelter pegs.

FOREST SHELTERS

IF YOU FIND YOURSELF in a survival situation in a temperate forest without a poncho, bivy sack, or tarp, you can make a shelter from natural materials. For example, you can construct a lean-to or an A-frame, or simply adapt a fallen tree. A forest lean-to is easier for a group to build because you can share the workload and don't have to use up too much precious energy.

TOOLS AND MATERIALS

- Poles and stakes
- Saw and knife
- Cordage
- Saplings
- Foliage for the roof covering

BUILDING A FOREST LEAN-TO

A lean-to has a sloping roof that leans against a horizontal ridgepole. It works best on a flat area between two trees or vertical supports secured firmly into the ground. If you're building a lean-to for a group, ask everyone to lie down next to each other and add an additional 6 in (15 cm) per person—this will establish how wide the shelter needs to be.

BUILDING IN A FOREST

Temperate forests, whether they consist of deciduous or coniferous trees or both, provide many opportunities for finding or building a shelter. Expend the least amount of energy for the most amount of protection—first see what nature can provide before building a shelter yourself.

POINTS TO REMEMBER

If you do need to make a shelter in a forest, here are some tips:

- Select a place close to water but away from water hazards, such as flooding, animals, and insects.
- Before you start, check for hazards such as deadfalls (sudden branch falls), rockfalls, and flash floods.
- Select a place that gives you maximum protection from the elements and with all the materials you need nearby.
- Collect everything you need before you start building and before it gets dark.
- Make sure your aids to location can be seen or activated quickly (signal fires, heliograph, flares, radio signals).
- Think safety: if possible, wear gloves to clear leaves and debris from the ground to protect you from spiders and snakes.
- Layer down your clothing as you work to prevent overheating.
- Make your shelter as waterproof as you can and ensure your bedding raises you at least 4 in (10 cm) off the ground.

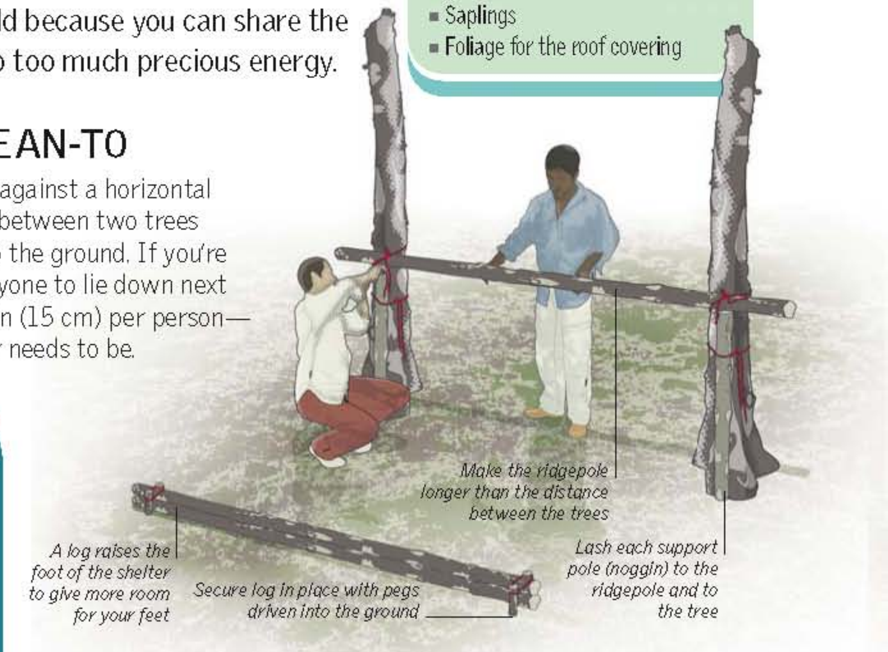

Make the ridgepole longer than the distance between the trees

A log raises the foot of the shelter to give more room for your feet

Secure log in place with pegs driven into the ground

Lash each support pole (noggin) to the ridgepole and to the tree

1 Place the ridgepole against the trees at the height you require and lash it with an arbor knot (see p. 142) to both trees.

- Place a support pole (noggin) under the ridgepole and lash to the trees.
- Firmly peg a log where you want the foot of the shelter to be.

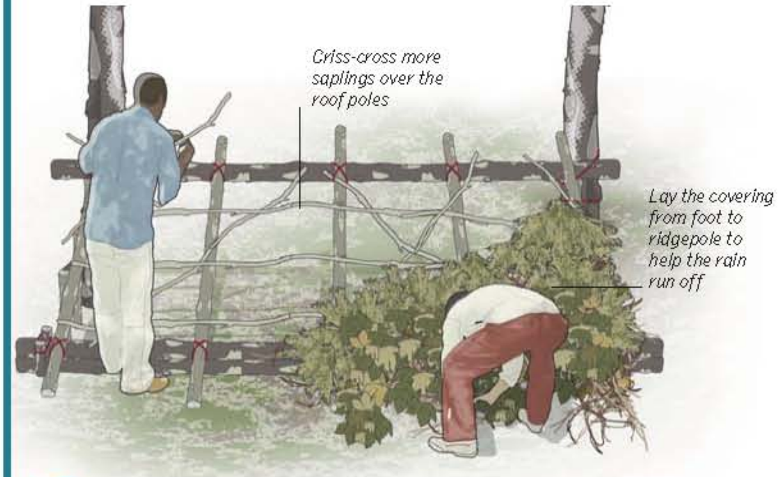

Criss-cross more saplings over the roof poles

Lay the covering from foot to ridgepole to help the rain run off

4 Apply the covering to the woven skeleton of the roof and walls. How much you apply depends on the weather you're experiencing.

- Lay large materials—for example, pine or fir boughs, large leaves, clumps of moss—as a base to stop smaller materials from falling through.

TREE SHELTERS

The branches, roots, or trunk of a fallen tree can make a basic shelter. Make sure the tree is safe before you use it, as a further fall could injure you. Dead trees are probably dry, so don't light a fire under them. Check the base of the tree for snakes, spiders, and nests of insects.

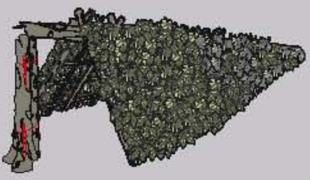

BROKEN BRANCHES

A branch that has snapped from a tree trunk but hasn't yet fallen makes a good ridgepole. Prop smaller branches and roofing material up against it.

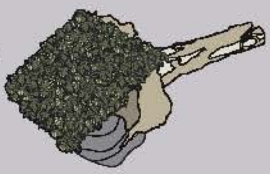

UPTURNED ROOTS

The base of an uprooted tree can give good protection from the elements. Make sure it's safe and won't flood if it rains, then add boughs and foliage.

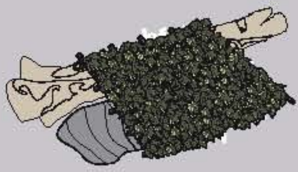

FALLEN TRUNK

Lay sticks or branches against a fallen trunk to form a pitched roof. Weave saplings across them and cover with bark, turf, moss, leaves, and mulch.

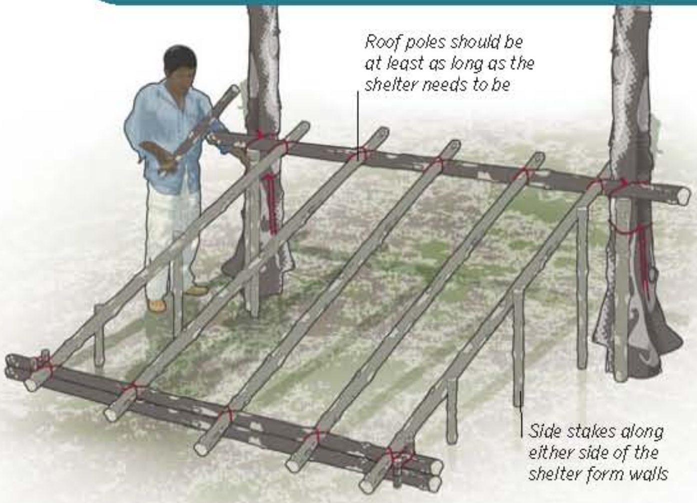

Roof poles should be at least as long as the shelter needs to be

Side stakes along either side of the shelter form walls

2 Place five roof poles evenly along the ridgepole so that they slope back to the foot of the lean-to.

- Keep the outside poles between the trees to stop them moving outward.
- Lash each roof pole to the ridgepole.
- Hammer in side stakes alongside the outside poles.

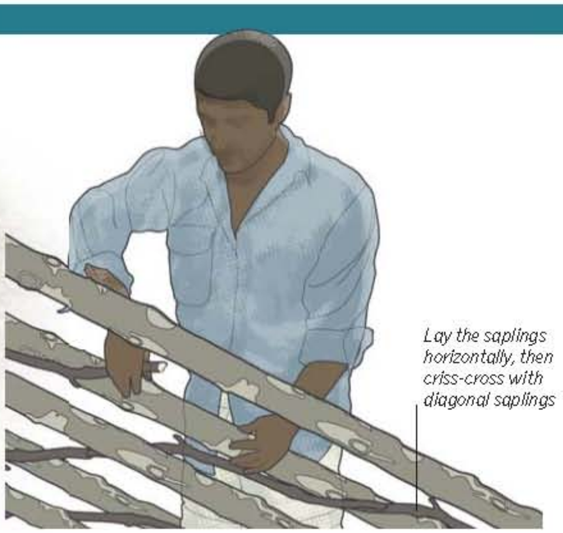

Lay the saplings horizontally, then criss-cross with diagonal saplings

3 Weave saplings in and out through the roof poles, working across and down from the ridgepole.

- Alternate the weave of each row of saplings, first horizontally and then diagonally.
- Weave smaller saplings through side stakes to form walls.

5 With everyone in the same lean-to shelter you need to light and tend only one fire.

- Layer the inside floor with dry material to form bedding.
- Establish a watch system during the night to ensure that the fire is maintained and controlled.

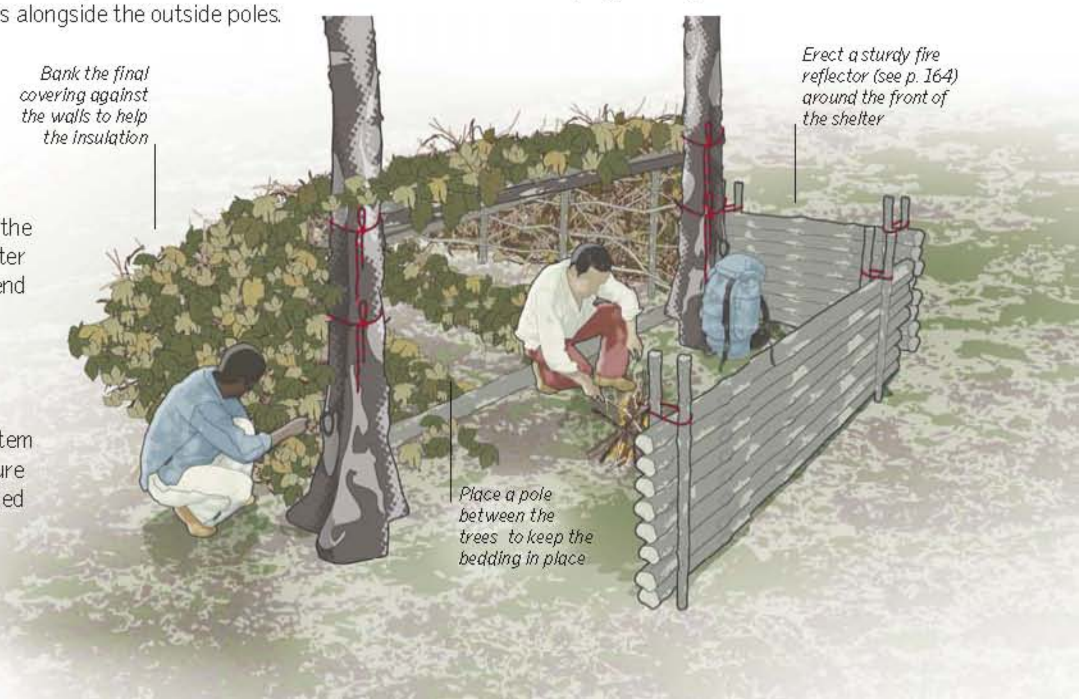

Bank the final covering against the walls to help the insulation

Erect a sturdy fire reflector (see p. 164) around the front of the shelter

Place a pole between the trees to keep the bedding in place

FOREST A-FRAME

A-frames can take several hours to build, but if you're staying in one place for a few days it may be worth expending the energy. One-person A-frames are relatively simple to construct and are easy to keep warm. They can be adapted into multi-person shelters to accommodate groups of people.

TOOLS AND MATERIALS

- Poles of various lengths
- Saplings
- Knife
- Saw (or ax)
- Cordage
- Covering for walls—such as pine boughs, leaves, and moss

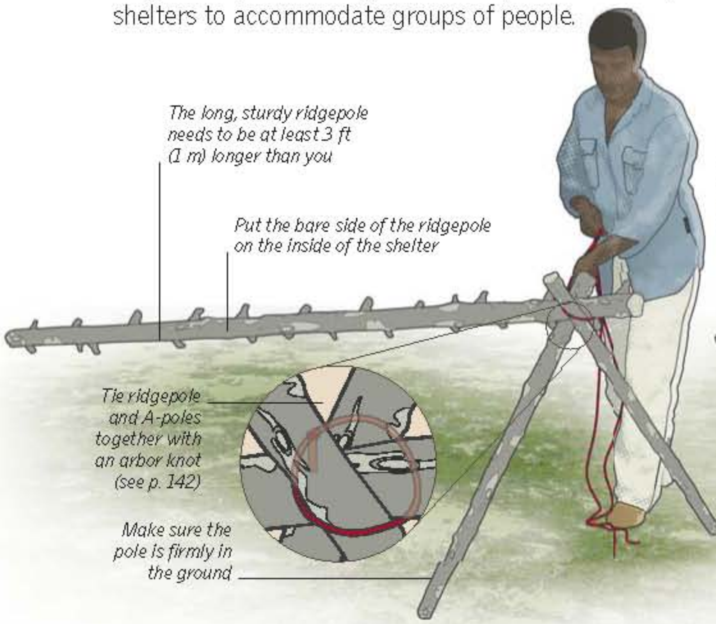
The long, sturdy ridgepole needs to be at least 3 ft (1 m) longer than you
Put the bare side of the ridgepole on the inside of the shelter
Tie ridgepole and A-poles together with an arbor knot (see p. 142)
Make sure the pole is firmly in the ground

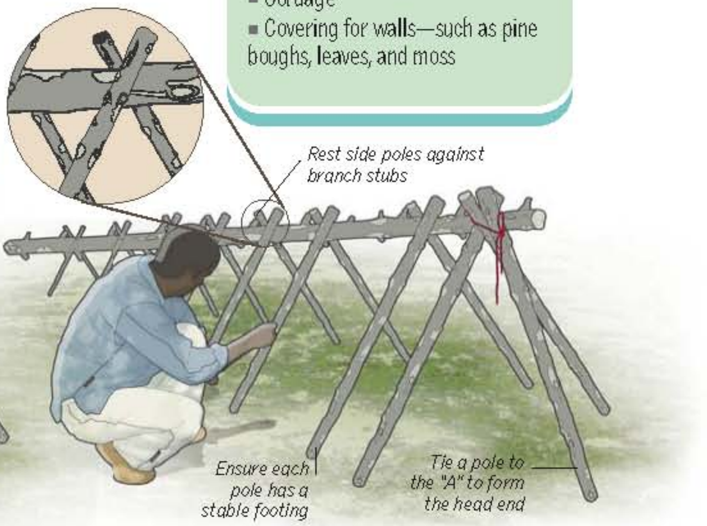
Rest side poles against branch stubs
Ensure each pole has a stable footing
Tie a pole to the "A" to form the head end

① Smooth one side of the ridgepole with your knife, leaving one side bare and one with branch stubs.

- Hammer two poles into the ground to form an "A".
- Rest one end of the ridgepole on top of the "A" and the other end on the ground.

② Lay a series of poles against the branch stubs on the ridgepole. These will be of decreasing height as you work your way from front (the "A") to back (the ground).

- At each junction, rest the side poles against the stubs or tie them to the ridgepole with an arbor knot (see p. 142).

FIRE REFLECTOR

A fire radiates heat in all directions—up, down, and 360 degrees around. A fire reflector makes a fire more efficient as it directs heat into your shelter.

Construct your reflector so that you can build a fire about 3 ft (1 m) from the entrance to your shelter. You can sit in the entrance and tend the fire at arms' length—any closer and you would be too hot and the shelter might catch fire; any further away and you lose too much heat. An L-shaped end to the reflector will retain more heat.

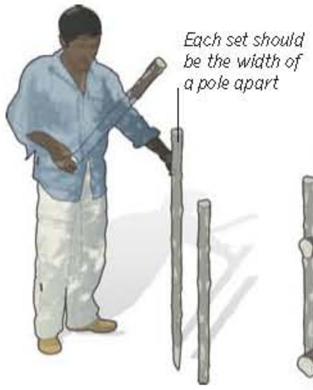
Each set should be the width of a pole apart

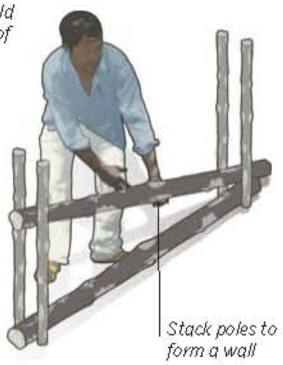
Stack poles to form a wall

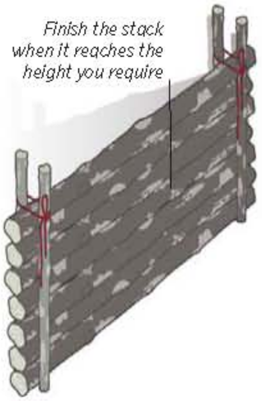
Finish the stack when it reaches the height you require

① Check the air flow will be across the proposed reflector (to disperse smoke).

- Hammer two poles into the ground. Put a second set in at poles' length apart.

② If the fire is going to be close to the reflector, use green wood to prevent the heat setting it on fire.

- Put poles between the sets to form a wall.

③ Tie the top of the upright poles together.

- Make sure the wall is as long as the shelter's entrance to retain heat and keep out wind and rain.

ALTERNATIVE SHELTERS

Two other options are a front-opening A-frame and a one-person lean-to. For the former, secure a ridgepole or line to a tree at the height you want the entrance to be, then adapt the forest A-frame (see opposite). For the latter, tie the ridgepole to two forked sticks hammered into the ground and apply a cover similar to the forest lean-to (see pp. 162–63).

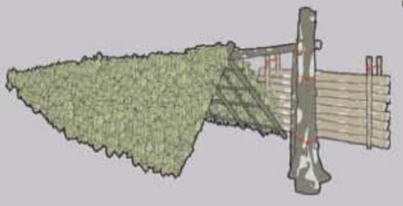

FRONT OPENING A-FRAME

Wedge a noggin under the ridgepole and lash it to the tree with an arbor knot to take some of the weight. Build a fire reflector to the side with the prevailing wind.

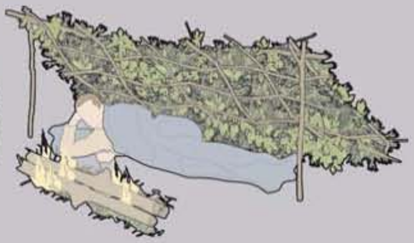

ONE-PERSON LEAN-TO

A small one-person lean-to, together with a long log fire built in front, provides you with a warm, dry, and comfortable shelter for an overnight stay.

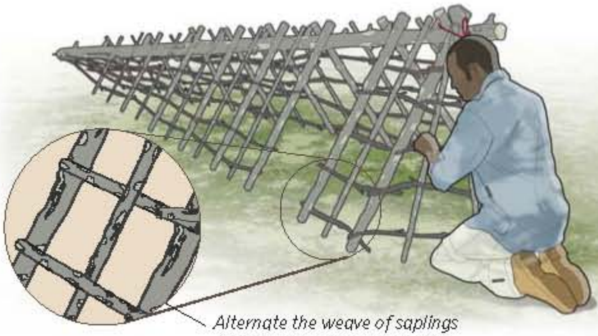
Alternate the weave of saplings between one row and the next

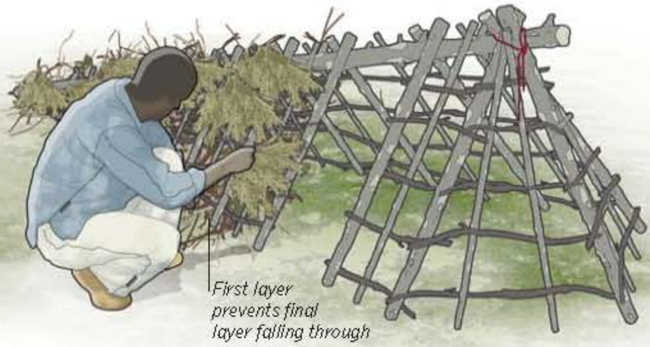
First layer prevents final layer falling through

3 Weave four or five rows of saplings horizontally through the side poles to make the walls into a strong framework for the covering.

- Leave an entrance in the side near the front. Make sure it's large enough for you to sit inside comfortably.

4 Cover the framework with a thick layer of natural material, such as pine boughs, small branches, and sturdy twigs. Weave the materials into a "thatch".

- On each side, work your way gradually from the back of the shelter to the front, and from the bottom to the top.

NEED TO KNOW

Build your shelter so that the wind blows across the front of it at a slight angle, rather than into it. This maintains its effectiveness and prevents smoke entering from the fire. If the shelter backs on to the wind, there's a risk the wind will blow around it and curl into the entrance.

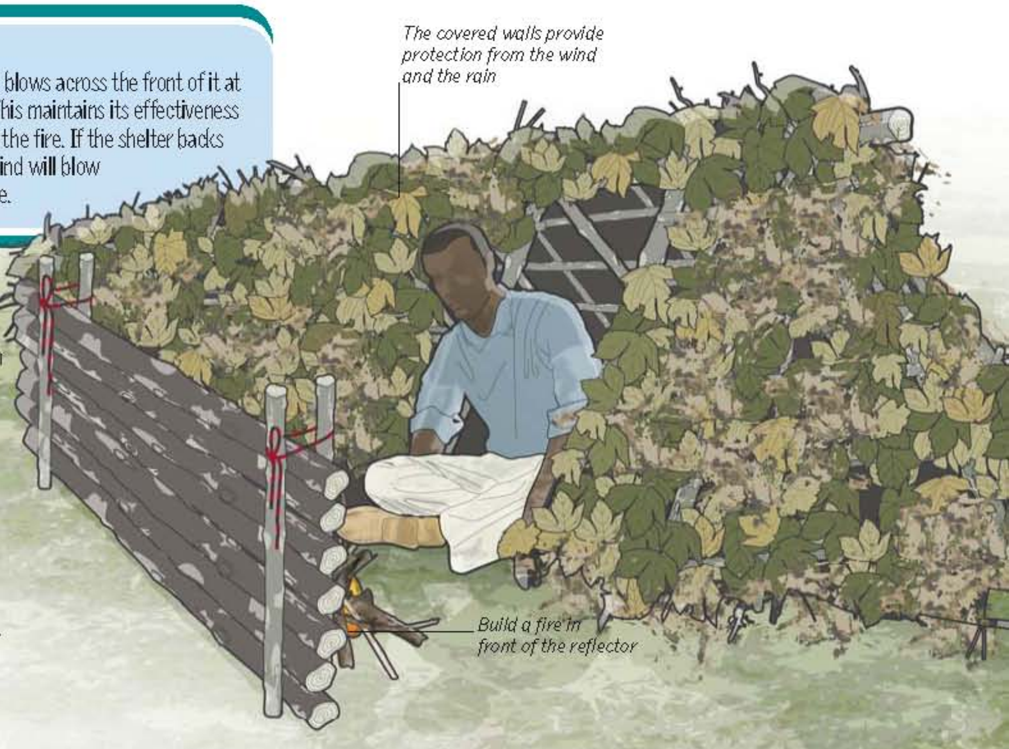
The covered walls provide protection from the wind and the rain
Build a fire in front of the reflector

5 Cover the first layer with leaves, moss, and mulch, starting from the ground and working up to the ridgepole.

- Build a fire reflector (see opposite) in front of the entrance to help direct heat towards you.
- Stow your equipment and pack in the front of the shelter.

TROPICAL SHELTERS

TROPICAL RAIN FORESTS CAN BE the easiest environments for building shelters because of the wealth of available materials. Once you've found a suitable location just decide what type of shelter you need. The most important thing is to raise the shelter off the ground. Make it as comfortable as you can, because quality sleep is important.

TOOLS AND MATERIALS

- Poncho fitted with press-studs
- Cordage
- Long, heavy poles or wide bamboo stalks
- Spacer bars
- Knife, saw, machete, or parang
- Needle
- Strong, waxed, cotton thread

MODIFIED PONCHO BED

Most ponchos have press-studs to allow you to clip up the edges or to clip two ponchos together. They also have grommet holes that you can use as lashing points. Here, a poncho is modified into a tropical bed.

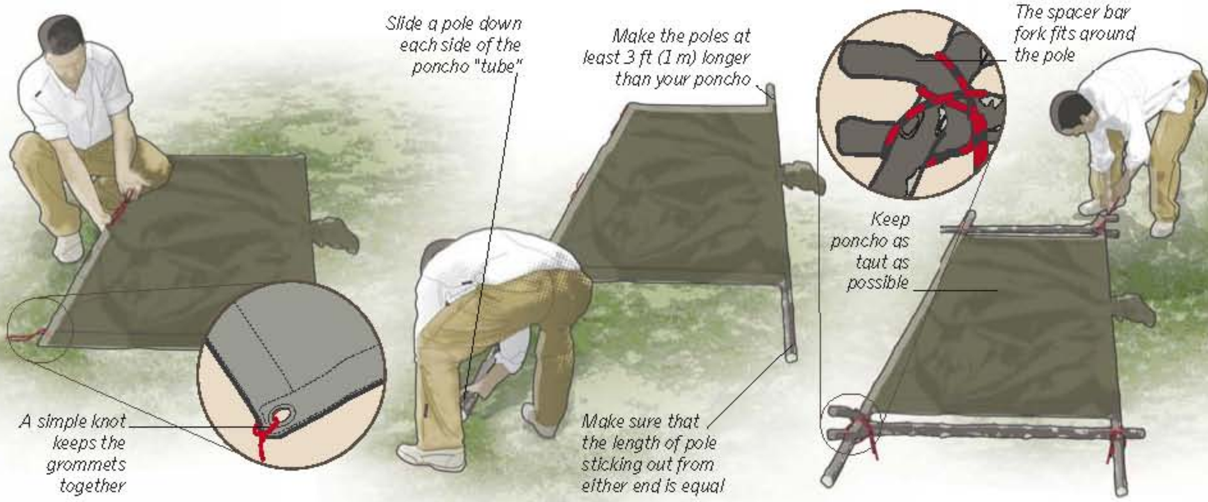

A simple knot keeps the grommets together

Slide a pole down each side of the poncho "tube"

Make the poles at least 3 ft (1 m) longer than your poncho

Make sure that the length of pole sticking out from either end is equal

The spacer bar fork fits around the pole

Keep poncho as taut as possible

1 Clip the press-studs of the two sides of your poncho together to form a kind of "tube"."

- Tie the grommets on the clipped side with pieces of cordage so that the press-studs don't come undone.

2 Cut two strong support poles that can take your weight.

- Insert the two support poles inside the "tube" of your poncho, one down each side.

3 Cut two "spacer bars" that will fit perpendicularly across the support poles or, if you can find them, cut poles with a natural fork in them.

- Tie the spacer bars into position using an arbor knot (see p. 142).

HAMMOCKS

Many types of hammock are designed for the kind of hot weather you get in the tropics, and are often fitted with mosquito nets. In fact, hammocks make the best overnight shelter, as long as you can find two trees the right distance apart. The Hennessy hammock is a good choice because it packs down small.

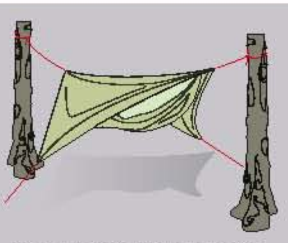

HENNESSY HAMMOCK

This ultralight, quick-drying hammock is both easy to set up and comfortable to sleep in (see p. 54). It's equipped with a self-closing door.

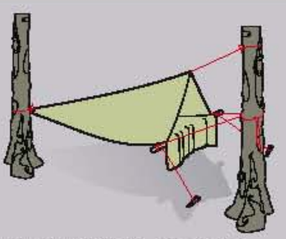

PARACHUTE SHELTER

Fold the panels of a parachute into a triangle. Lash a spacer bar to the opening and peg it out. Tie off the apex of the triangle further up the tree.

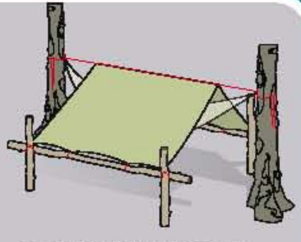

HAMMOCK AND TARP

Tie a ridgeline above the hammock. Drape the tarp over the line and anchor it on either side to a horizontal pole tied to two uprights.

ALTERNATIVE METHODS

You can easily modify other forms of sheeting, such as a tarpaulin, shelter sheet, and groundsheet. You will need a decent needle and plenty of strong, waxed, cotton thread (see Survival kits, pp. 60–61). There are two methods, which you can do either on the trail or as part of your prior preparation. Adding spacer bars keeps the poles apart and the sleeping platform tight. Both of these methods produce an excellent emergency stretcher.

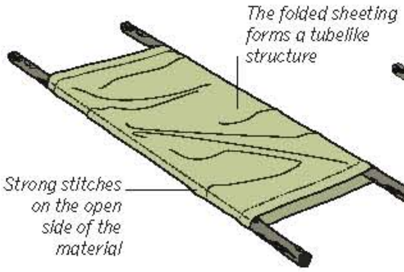
The folded sheeting forms a tubelike structure
Strong stitches on the open side of the material

METHOD ONE

Lay the material on the ground and fold it once to create a wide "tube." Stitch the length of the open sides and then insert a pole along each side of the tube.

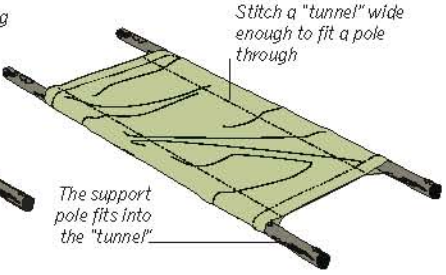
Stitch a "tunnel" wide enough to fit a pole through
The support pole fits into the "tunnel"

METHOD TWO

Fold the sheet and stitch a "tunnel" along both sides just wide enough to take a pole that will support your weight. Push the support poles through the tunnels.

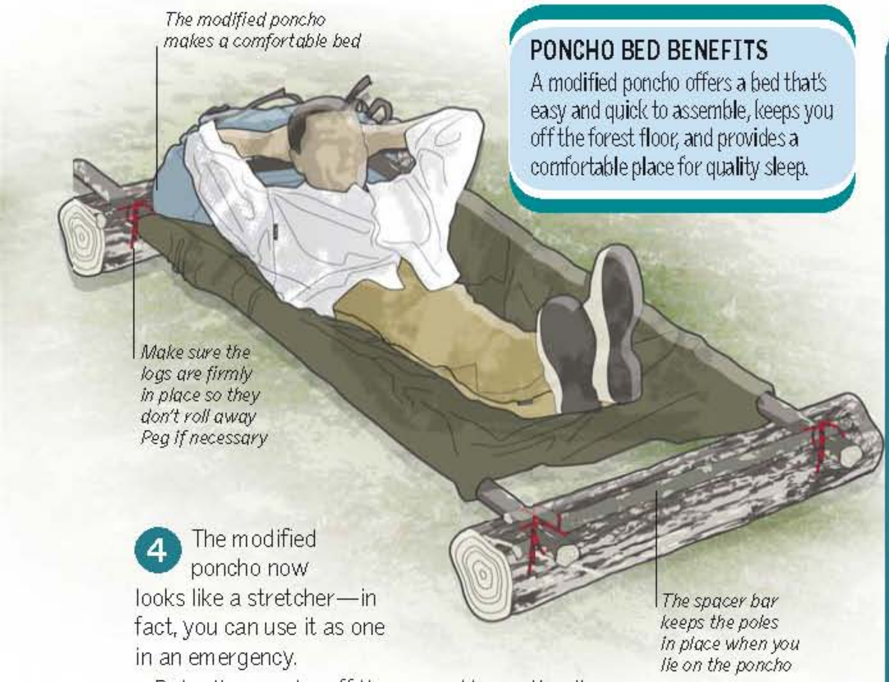
The modified poncho makes a comfortable bed
Make sure the logs are firmly in place so they don't roll away Peg if necessary
The spacer bar keeps the poles in place when you lie on the poncho

4 The modified poncho now looks like a stretcher—in fact, you can use it as one in an emergency.

- Raise the poncho off the ground by resting it either on thick logs or sturdy poles.

PONCHO BED BENEFITS

A modified poncho offers a bed that's easy and quick to assemble, keeps you off the forest floor, and provides a comfortable place for quality sleep.

ALTERNATIVE DESIGN

If you have a second poncho, shelter sheet, or tarp you can create a shelter to keep the rain off your modified poncho (see also Corded A-frame, p. 159).

- Tie a ridgeline between two trees and hang the poncho, sheet, or tarp over the top of it.
- If you're using a second poncho, tie its hood to another ridgeline.
- Peg out the four corners of the poncho, sheet, or tarpaulin.

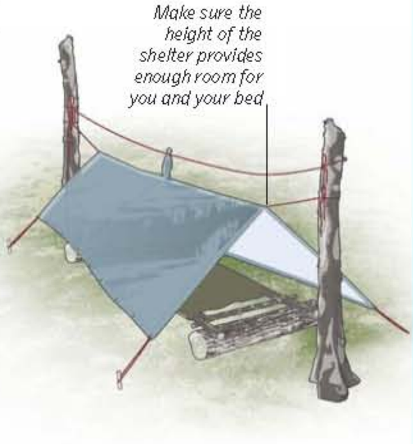
Make sure the height of the shelter provides enough room for you and your bed

BUILDING IN THE JUNGLE

Shelters in the rain forest need to be quick to erect, just big enough for your needs, and safe from animals.

CHECKLIST

If you're thinking about building a shelter in a tropical forest, the following tips may help you:

- Check for hazards such as deadfalls (sudden branch falls)—a major cause of injuries during military jungle training.
- You'll need a sharp cutting tool—ideally a parang or machete, but a decent bushcraft knife or small ax will suffice.
- Clear the ground around your shelter to deter animals. Use a makeshift brush, never your bare hands, to avoid being bitten by snakes or spiders.
- Build your shelter far enough off the ground to avoid being bothered by insects, snakes, or any other animals—particularly those that move around at night.
- Start building your shelter well before darkness sets in, which, in tropical zones, usually happens very quickly. Using a large knife in reduced light or by the light of a flashlight can be very dangerous.
- Work at a rate that you can manage. Humid conditions can very quickly lead to dehydration and heat-related injuries such as heat stress and heat stroke. Your body tries to keep itself cool by sweating, so don't work too fast, drink water frequently, and take regular breaks.
- Make your shelter secure—you don't want to make repairs at night.
- Inside your shelter use a full mosquito net and a mosquito head-net.
- A fire will deter insects and animals.

JUNGLE A-FRAME

An A-frame is relatively easy to make. If you have a poncho, tarpaulin, shelter sheet, groundsheet, or another type of sheet, you'll need to modify it first to make it into a bed (see pp. 166–67). If you do not have a poncho, lay branches across the poles to form a sleeping platform.

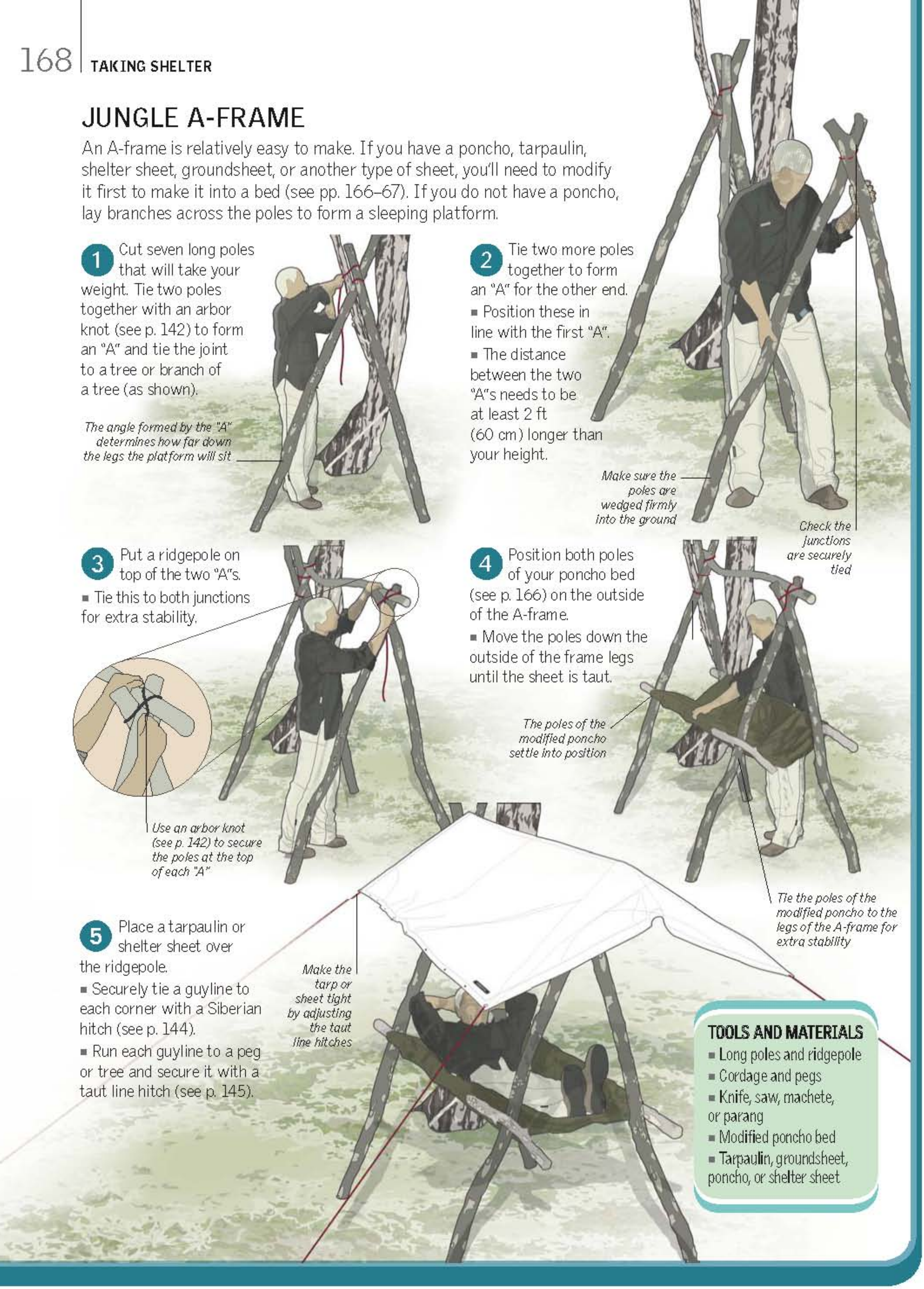

1 Cut seven long poles that will take your weight. Tie two poles together with an arbor knot (see p. 142) to form an "A" and tie the joint to a tree or branch of a tree (as shown).

The angle formed by the "A" determines how far down the legs the platform will sit

2 Tie two more poles together to form an "A" for the other end.

- Position these in line with the first "A".
- The distance between the two "A"s needs to be at least 2 ft (60 cm) longer than your height.

Make sure the poles are wedged firmly into the ground

Check the junctions are securely tied

3 Put a ridgepole on top of the two "A"s.

- Tie this to both junctions for extra stability.

Use an arbor knot (see p. 142) to secure the poles at the top of each "A"

4 Position both poles of your poncho bed (see p. 166) on the outside of the A-frame.

- Move the poles down the outside of the frame legs until the sheet is taut.

The poles of the modified poncho settle into position

Tie the poles of the modified poncho to the legs of the A-frame for extra stability

5 Place a tarpaulin or shelter sheet over the ridgepole.

- Securely tie a guyline to each corner with a Siberian hitch (see p. 144).
- Run each guyline to a peg or tree and secure it with a taut line hitch (see p. 145).

Make the tarp or sheet tight by adjusting the taut line hitches

TOOLS AND MATERIALS

- Long poles and ridgepole
- Cordage and pegs
- Knife, saw, machete, or parang
- Modified poncho bed
- Tarpaulin, groundsheet, poncho, or shelter sheet

BAMBOO LEAN-TO

If you can find enough bamboo you can make a type of lean-to shelter. Find a spot close to the bamboo supply to reduce the amount you have to carry. If you need a raised sleeping platform, adapt the method used for the jungle hut (see p. 174).

TOOLS AND MATERIALS

- Thick wooden poles and bamboo poles
- Knife, saw, machete, or parang
- Cordage and pegs
- Poncho and tarpaulin (if you have them)

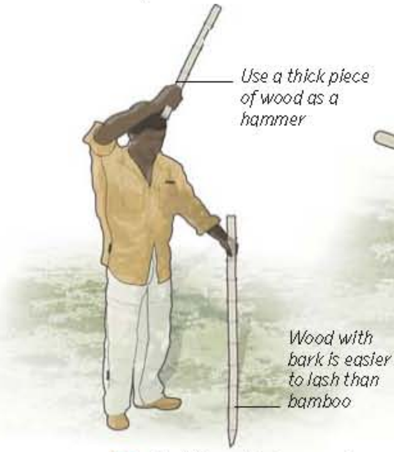
Use a thick piece of wood as a hammer
Wood with bark is easier to lash than bamboo

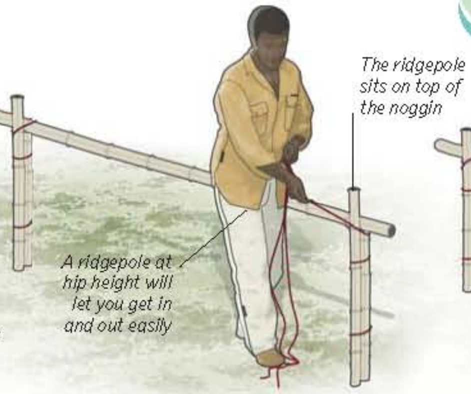
The ridgepole sits on top of the noggin
A ridgepole at hip height will let you get in and out easily

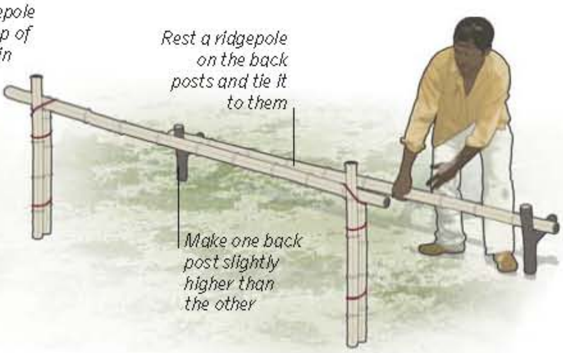
Rest a ridgepole on the back posts and tie it to them
Make one back post slightly higher than the other

1 Cut four thick, wooden support posts.

- Hammer two posts into the ground, wide enough to let you lie lengthways between them.

2 Tie a noggin to each post for support. Make sure they are level and at the height you need for an entrance.

- Place a wooden ridgepole on top of the noggins and tie it to the posts.

3 Make two short, forked back posts to give the shelter a sloped roof.

- Align them with the front posts and far enough back so the roof is wide enough to shelter you.
- Hammer them into the ground and tie a second ridgepole to them.

4 Cut lengths of bamboo long enough to reach between the two ridgepoles.

- Split these by wedging your knife into each pole and hammering it down.
- Prepare enough split bamboo lengths to cover the roof area two times.

5 Starting from one end, lay lengths of split bamboo, hollow side up, side-by-side, between the ridgepoles.

- Work your way back along the roof, this time laying the lengths of split bamboo with the hollow side down.

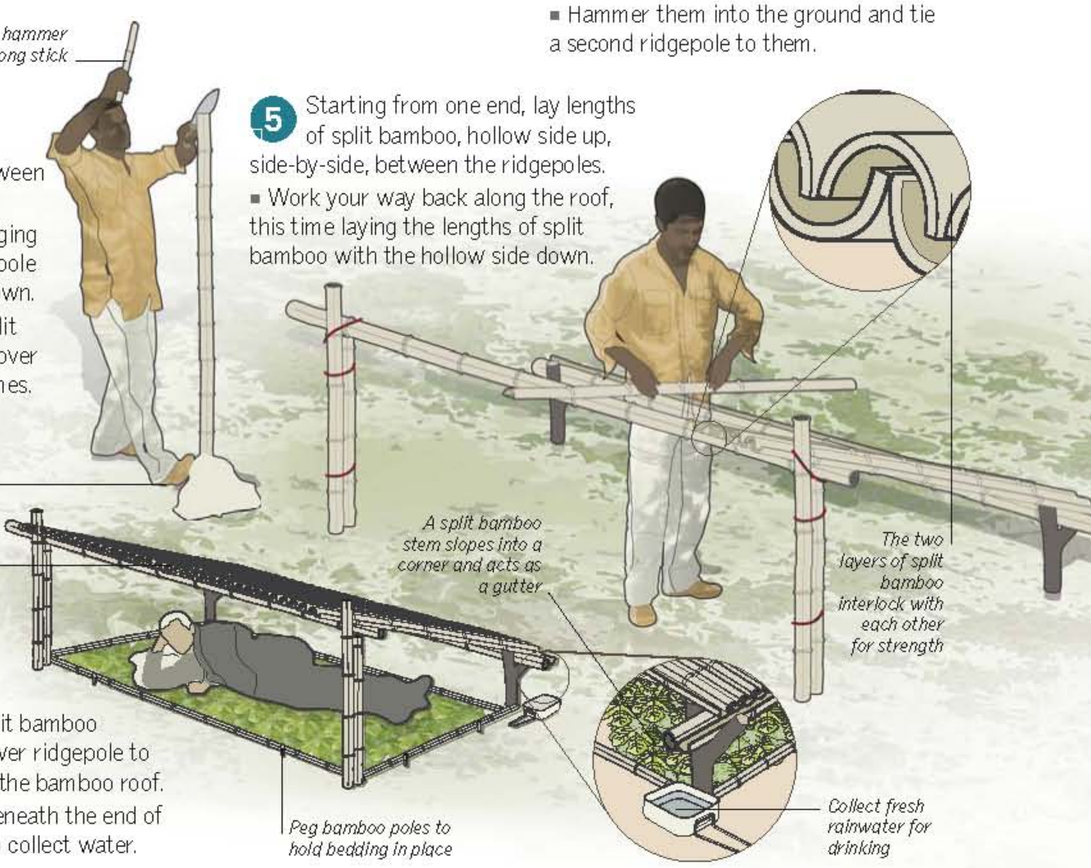
Wedge in knife then hammer it down with a strong stick
Steady the bamboo stem against a rock with your foot
The two layers of split bamboo interlock with each other for strength
The split bamboo layers of the roof direct water into the gutter
A split bamboo stem slopes into a corner and acts as a gutter
Peg bamboo poles to hold bedding in place
Collect fresh rainwater for drinking

6 Lash a long split bamboo stem to the lower ridgepole to form a gutter under the bamboo roof.

- Put a pan or pot beneath the end of this sloping gutter to collect water.

JUNGLE HUT

With a little extra effort you can make a more permanent and substantial shelter than a jungle A-frame. The raised apex roof of a jungle hut allows you to sit on the platform in relative comfort. This method uses a framework of four posts, but using two trees instead of two of the poles would make it more rigid.

TOOLS AND MATERIALS
- Strong posts and poles
- Branches for platform
- Knife, saw, machete, or parang
- Digging tool
- Broad leaves for thatching the roof

1 Dig four strong posts into the ground to form a rectangle at least 1 ft (30 cm) longer and wider than yourself.

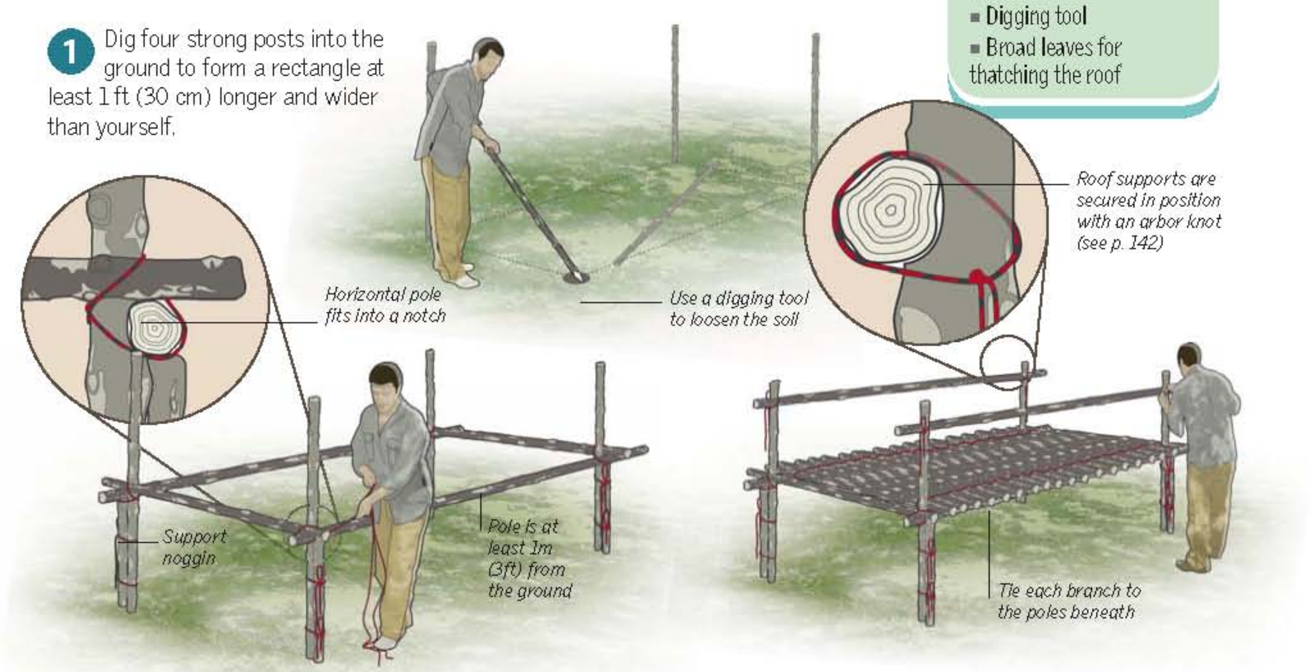
Horizontal pole fits into a notch
Use a digging tool to loosen the soil
Roof supports are secured in position with an arbor knot (see p. 142)
Support noggin
Pole is at least 1m (3ft) from the ground
Tie each branch to the poles beneath

2 With your knife or saw, cut a notch into each post deep enough to fit the curve of the horizontal pole.
- Wedge the four horizontal poles into their notches and secure the junction using an arbor knot (see p. 142).
- Lash a noggin to each post to support the poles.

3 Lay a row of stripped branches side-by-side across the horizontal poles to form a raised floor.
- To make the roof supports, cut a notch on the inside of each post about 3 ft (1 m) higher than the platform.
- Tie a pole into the notches on the posts.

THATCHING LEAVES

By taking advantage of the shapes of some tropical leaves, you can make a substantial, long-lasting roof or walls. The bigger and broader leaves involve less work. If you use rattan palm leaves you can create an alternating interweave that gives a tighter thatch than the overlapping method described above.

RATTAN LEAVES

The rattan palm has leaves that are composed of rows of smaller leaflets. You can split the leaves in two and hang the halves in layers on the framework of your roof. Alternatively, you can use them whole, by folding the leaflets from one side and individually weaving them with leaflets from the other side.

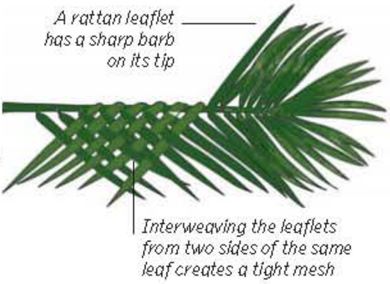
A rattan leaflet has a sharp barb on its tip
Interweaving the leaflets from two sides of the same leaf creates a tight mesh

LONG, BROAD LEAVES

Many tropical plants have long, broad leaves that are good for thatching. These include some banyan figs, rubber plants, and types of banana. Arrange these leaves side-by-side on the poles of your roof frame, tuck them over the pole, and sew them into position with strips of vine. The tips of the leaves in one row point down over the leaves of the row underneath.

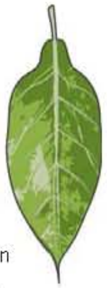

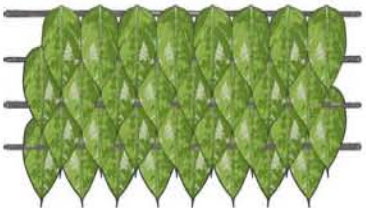

ALTERNATIVE SHELTERS

Wickiups are found in various forms around the world and are adapted as local needs and materials dictate. For example, Native Americans made them as temporary shelters in the plains, where they were also known as teepees. The pygmies of the rain forest used saplings to form a hemisphere that, once covered with natural material, created a warm, dry shelter.

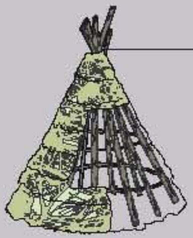

An opening at the top of the wickiup allows air to circulate

WICKIUP

The wickiup is a cluster of straight poles that are lashed together at the top, with an interwoven framework that's covered with animal hides or grass.

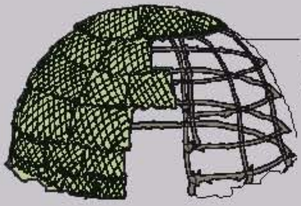

The domed shape gives plenty of room inside

PYGMY HUT

The domed pygmy hut is made from a circle of bent saplings or limber poles secured firmly in the ground, then lashed and thatched with natural materials.

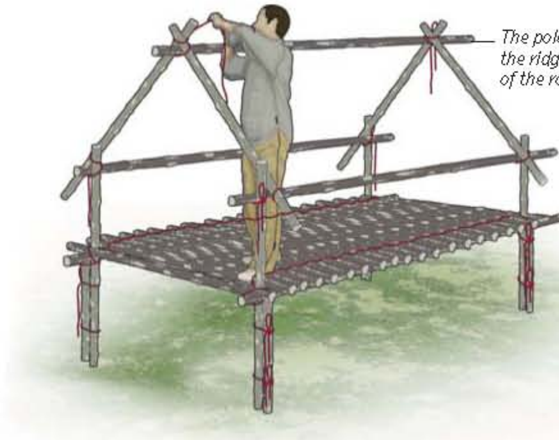

The pole forms the ridge of the roof

4 Tie poles together to form two A-frames. Tie these on top of the roof support poles at each end of the framework.

- Place a ridgepole across the apex of the two A-frames and tie it in place.

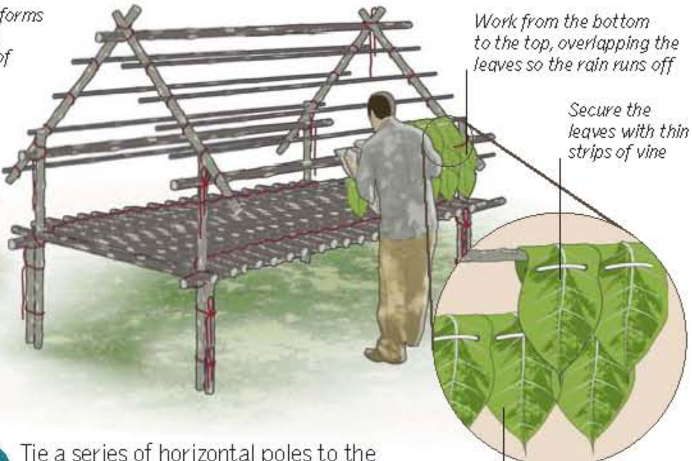

Work from the bottom to the top, overlapping the leaves so the rain runs off

Secure the leaves with thin strips of vine

Hang leaves with tips pointing down and shiny side up

5 Tie a series of horizontal poles to the legs of both A-frames to create a roof.

- Cover the roof with big, broad leaves (such as palm or banana) by folding them over the horizontal roof poles.

THREE-LOBED LEAVES

Leaves with three lobes, such as the familiar fig leaf, can simply be hooked over the poles of your roof with the stem pointing upward. On each row of leaves you need to hang the left and right lobes either behind the pole or in front of it in an alternating pattern. The middle lobe then hangs down over two leaves on the row underneath.

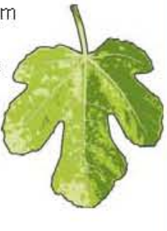

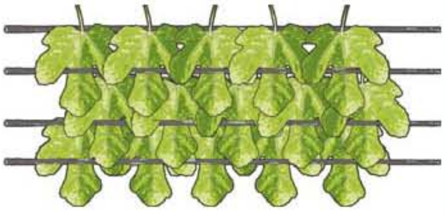

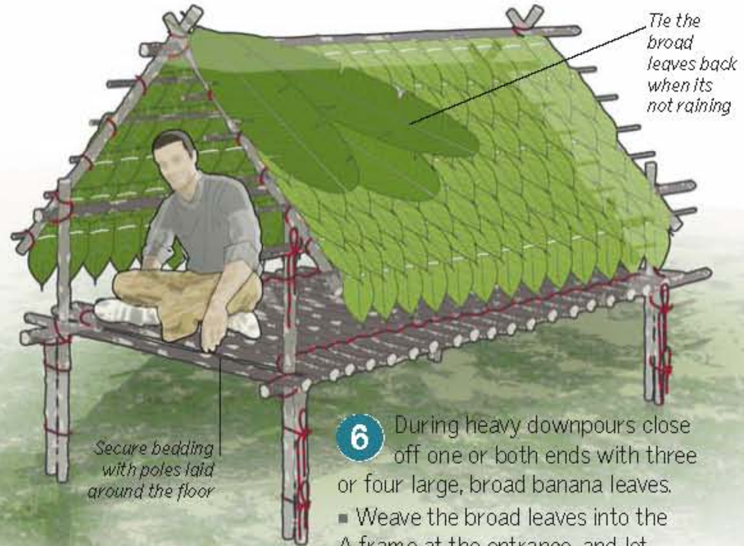

Tie the broad leaves back when its not raining

Secure bedding with poles laid around the floor

6 During heavy downpours close off one or both ends with three or four large, broad banana leaves.

- Weave the broad leaves into the A-frame at the entrance, and let them hang down.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN THE JUNGLE

USEFUL EQUIPMENT

- Machete/parang
- Tarpaulin/shelter sheet
- Hammock and mosquito net
- Insect repellant
- Signal flares and whistle
- Flashlight and batteries
- First-aid kit
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

FRENCHMEN GUILHEM NAYRAL AND LOÏC PILLOIS lost their way on a 78 mile (125 km) trek through virgin rain forest in French Guiana. They survived for 51 days in the jungle by drinking rainwater and eating palm seeds, snakes, and insects—although Nayral narrowly escaped death after eating a poisonous spider.

The pair set off from the Grand Kanori rapids on the Approuague River on Wednesday, February 14, 2007, bound for Saül, a former mining village at the center of the country. They planned to spend 12 days hiking, and carried sufficient food supplies, a compass and map, machete, a tarpaulin, and two hammocks. They soon found that the going was tough—some days it would take several hours of hacking through vines to hike just one mile. On the morning of February 26, when Nayral and Pillois had been due to reach Saül, the pair found themselves far from civilization and in unexpected terrain.

Knowing that a search operation would be launched, they set up camp and waited to be rescued. They built a shelter and divided the tasks—Nayral in charge of food and Pillois the fire, which was kept alight constantly to attract rescuers. They had abundant water from rainfall, but were reduced to scavenging for palm seeds, beetles, snakes, frogs, and spiders.

"THE PAIR FOUND THEMSELVES FAR FROM CIVILIZATION AND IN UNEXPECTED TERRAIN"

Occasionally, helicopters passed overhead, but failed to spot them through the dense canopy of foliage. After waiting for three weeks, the pair decided to abandon camp and trek west toward Saül. After another week of trekking for just three hours a day, unknowingly just 3 miles (5 km) from Saül, Nayral was incapacitated after eating a still-poisonous half-cooked spider. With no choice but to leave his companion behind, Pillois pushed on and reached Saül, returning with a helicopter to rescue Nayral—dehydrated, with intestinal poisoning, and plagued by parasites—on April 5, 51 days after beginning his trek.

WHAT TO DO

ARE YOU IN DANGER?

← NO YES →

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Sun/Heat/Humidity—Slow down to the pace of the jungle. Find or improvise immediate shelter
Animals—Only six percent of snakes are poisonous but everything in the jungle will try to protect itself
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

← NO YES →

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

← NO YES →

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

← NO YES →

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to
- Use line of sight to navigate on your bearing, as visibility may be less than 33 ft (10 m)
- Improvise shelter when not moving, and sleep off the ground, clear of the damp floor and animals
- Be on constant lookout for dry tinder and fuel
- Follow water courses downstream. Transport in the jungle relies on rivers, so settlements are most likely to be found alongside rivers
- Step onto logs so that you can see what is on the other side, rather than stepping straight over onto an unseen snake

DON'T

- Use your hands to clear undergrowth—a machete or walking staff is better suited
- Drink untreated water—boil or treat all of your water before drinking
- Leave it too late in the day to stop and make camp—three hours before sunset is recommended
- Keep too quiet. Make a noise as you progress to warn animals for your approach

YOU SHOULD STAY **

DO

- Select a shelter site where you can sleep off the ground and where your location aids will be most effective
- Use a mosquito net if you have one; if not, put damp foliage on your fire to repel insects, or cover exposed skin with mud
- Deploy all your aids to location and prepare them for immediate use. Be constantly alert for signs of rescue
- Keep your flesh covered despite the heat—the high humidity encourages infections. Also wash at every opportunity
- Keep a fire going—it aids location and wards off insects

DON'T

- Let the oppressive nature of the jungle overwhelm you. Slow down to its pace, don't fight against it
- Let your firewood get damp—store dry tinder and split or quarter wood to get to the dry inner core
- Eat what you cannot identify as edible—this could result in you becoming so ill that you cannot function

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

DESERT SHELTERS

IF YOU'RE PLANNING A TRIP into the desert, take something you can use to improvise an immediate shelter from the sun—for example, a shelter sheet, tarpaulin, poncho, space blanket, or even a couple of pieces of parachute material that will pack down into almost nothing. You can either build a shelter called a "scrape" in a natural hollow or erect a quick shelter above the ground using your poncho.

TOOLS AND MATERIALS

- Cordage and pegs
- Digging tool
- Knife
- Poncho, space blanket, tarp, or parachute material
- Rocks or ballast material
- Strong posts

BUILDING IN THE DESERT

Erecting a natural shelter in the desert is a challenge because of the heat and potential lack of materials, so try to find a site that's shaded by a tree or shrub.

POINTS TO REMEMBER

The following tips will help you deal with the extremes of temperature:

- Never try to build your shelter during the hottest part of the day.
- Ration your sweat and not your water. Avoid exerting yourself. If you start to sweat then stop what you're doing and take a break in the shade (30 minutes under an unfolded map or a space blanket and have a drink). If in a group, share the work.
- Don't leave it until the last minute to find a suitable location for your shelter. Make the decision early and plan accordingly.
- Never build a shelter in a low-lying area, dry river bed, or wadi as they are potentially at risk of flash floods.
- Avoid the top of large, isolated hills or mountains because of the danger from lightning and extreme winds.
- Try to set up your shelter on a small rise, where the temperature can be as much as 10 degrees warmer at night (cold air sinks).
- Ensure the opening faces north in the Northern Hemisphere and south in the Southern Hemisphere so the sun doesn't shine directly in during the day.
- Try to dig down to create a depression, as the ground is cooler below the surface.
- Build your shelter for the worst possible conditions and not the conditions at the time. Desert weather can change very quickly and dramatically, and the ferocious winds will rip apart any shelter that is not secure.
- If your shelter sheet has a shiny side, make sure it faces up to reflect heat and act as an aid to location from above.

DESERT SCRAPE

If you have cord, you can dig a scrape and use the cord to peg out the sheet above the scrape. If you don't have cord the sheet will have to be held in place by other means, such as soil, sand, or rocks. With all layered desert shelters try to maintain tautness and separation between the layers.

CORDLESS SCRAPE

If you don't have enough cord to peg out a sheet, you can make a scrape by either digging down or building up the sides with rocks or sand mounds placed at regular intervals. Anchor your sheet with rocks.

1 If the hollow isn't deep enough to lie in, make a digging tool and scrape out the soil from the bottom to build up the sides of the hollow.

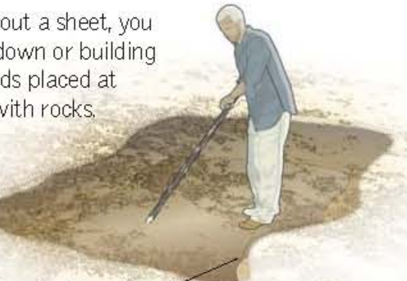

Build up the sides with rocks covered with soil

2 Spread your poncho, shelter sheet or tarp across the top of the hollow.

- Make sure the edges of the material overlap the sides of the hollow.

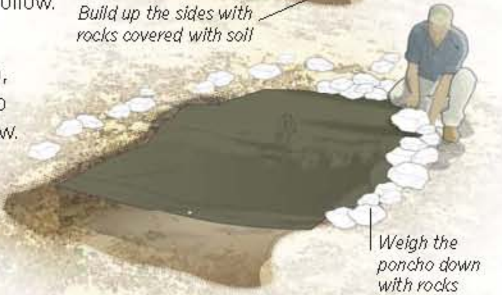

Weigh the poncho down with rocks

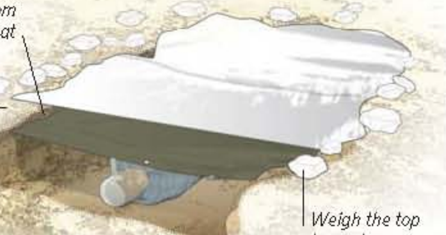

Air space helps to insulate you from the heat

Make a gap of at least 6 in (5 cm) between the layers

Weigh the top layer down with rocks

3 Use an additional sheet to make a second layer to create dead air space that helps to reduce the temperature below.

- If you only have one piece of material to use as cover, try to double it over to create the two layers.

QUICK SHELTER

If you can't find a hollow, erect a poncho shelter (see pp. 158–59) in a place that keeps you cool during the day—for example, under existing shade, such as trees or bushes, or at the top of a slight rise in order to benefit from any cool breeze.

1 Find a site beside a tree or secure a post next to where you want the shelter's opening to be.

- Rig a ridgeline to the tree or post, and peg your poncho over it to form a shelter.
- Repeat with a second poncho, shelter sheet, or space blanket to create a separate layer.

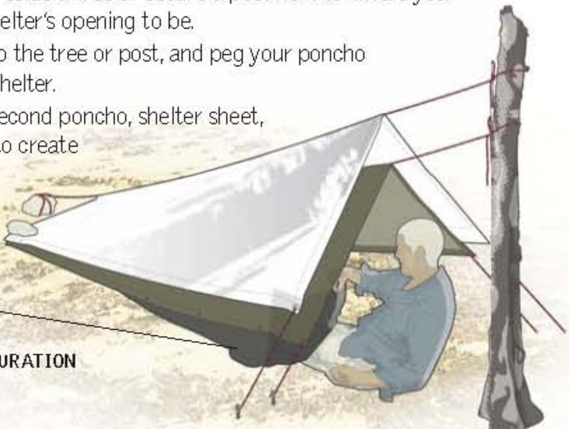
Leave sides and bottom open to air the shelter

DAY-TIME CONFIGURATION

2 At night, de-rig the outer shelter and use it as a sleeping cover for warmth.

- Prevent heat escaping by securing the sides and bottom with rocks, stones, shingle, or sand.
- Sleep on some bedding if possible.

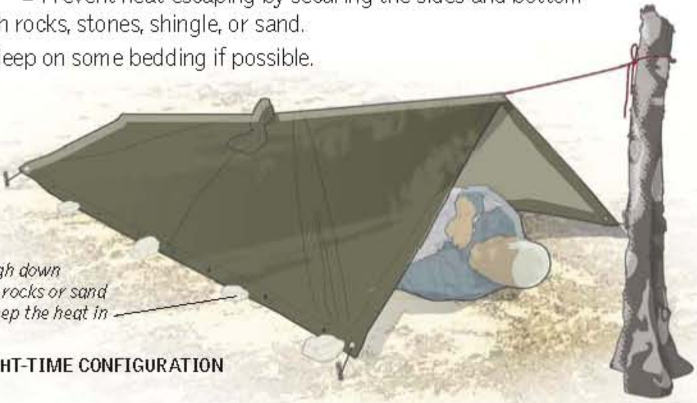
Weigh down with rocks or sand to keep the heat in

NIGHT-TIME CONFIGURATION

PARA SCRAPE

If you have cordage you can tie your sheet to four posts. If you haven't got enough cordage for four, use your backpack or a pile of stones instead.

- Prepare the depression or hollow as for a cordless scrape (see left).
- Use trees if available, or improvise posts from wood, piles of stones, or your backpack.
- Tie the sheet to the trees or the posts above the depression, leaving a gap for air to flow over.
- Create a second layer at least 6 in (15 cm) above the first.
- If foliage is available, place it between the layers to maintain separation.

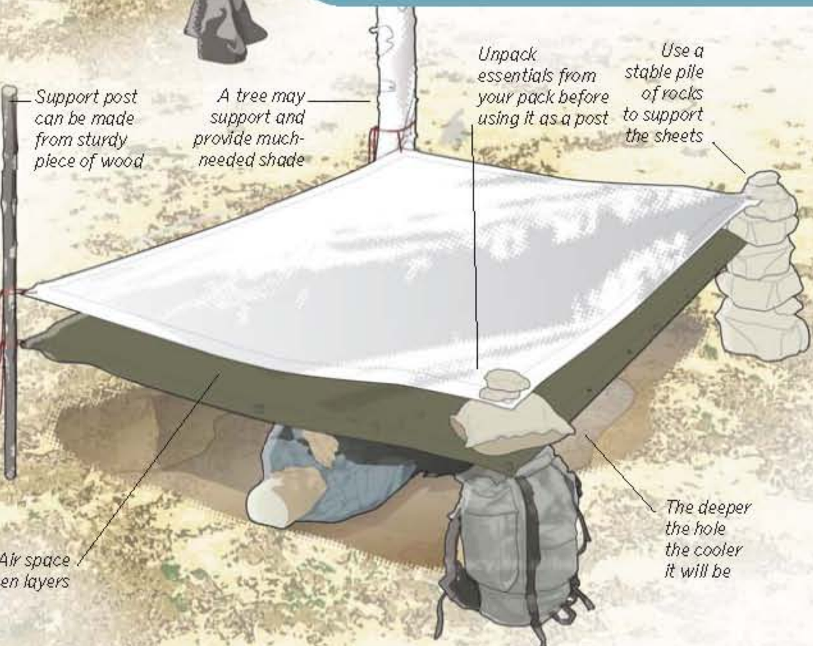
Support post can be made from sturdy piece of wood
A tree may support and provide much-needed shade
Unpack essentials from your pack before using it as a post
Use a stable pile of rocks to support the sheets
The deeper the hole the cooler it will be
Air space between layers

SUN PROTECTION

You should always carry immediate protection against the sun. You may be able to create shade using something you've brought, such as a trekking umbrella, tent, or space blanket.

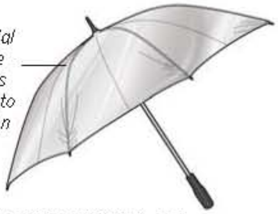
Material may be used as an aid to location

TREKKING UMBRELLA

A small, lightweight umbrella protects against sun, wind, and rain, creates cool shade, and doubles as a walking stick. Some have reflective material, or a flashlight in the handle.

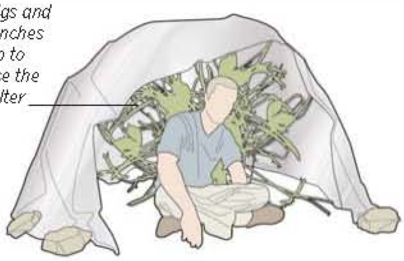
Twigs and branches help to raise the shelter

SPACE BLANKET

This should be part of your basic survival equipment and will afford immediate relief from the sun. Remember that it's almost impossible to pack a space blanket to its original size once it's unpacked.

SHORELINE SHELTERS

IF YOU FIND YOURSELF in a survival situation, and can make your way to a shoreline or beach, it may offer you particular benefits. Rescue is more likely here than further inland as your aids to location work better in open areas, people are more likely to live on the coasts, and useful flotsam washes up twice daily on even the remotest of beaches.

TOOLS AND MATERIALS

- Digging tool
- Driftwood and/or rocks
- Broad-leaved plants or grass, or poncho or tarp
- Knife
- Cordage and pegs

DRIFTWOOD SHELTER

If there's enough driftwood around on the shoreline, you can use it to make a variety of shelters covered in this chapter. A hole-in-the-ground driftwood shelter offers simple protection, but may need continued attention if the sand is very fine or gets wet.

BESIDE THE WATER

Shorelines vary with the environment—from tropical beaches, where sleeping out under the stars is not a problem, to rugged coasts in higher latitudes, where spending a night unprotected from the elements would be suicidal.

POINTS TO REMEMBER

If you intend to spend a night on a shoreline, exposed to the elements, the following tips may help you:

- Build the shelter above the highest high-tide mark on a seashore (see opposite) and above the highest watermark on a river or lake beach. If in doubt, move slightly further inland, where the protection may be better and materials more easily available.
- Try to establish what the weather's going to do by looking at the sky, watching the water, or detecting changes in the wind (see pp. 82–83).
- Plan for the worst, or at least have a back-up location should things not work out.
- Check the area for insects, such as midges, mosquitoes, gnats, and horseflies.
- Look for signs of wildlife, such as crabs and even turtles as sources of food.
- Finish your shelter and collect water and firewood before it gets dark.
- Beware the effects of sun and windburn, even on overcast days.
- Utilize driftwood and other building materials. Remember, any wood that's been in the sea is likely to be very heavy.

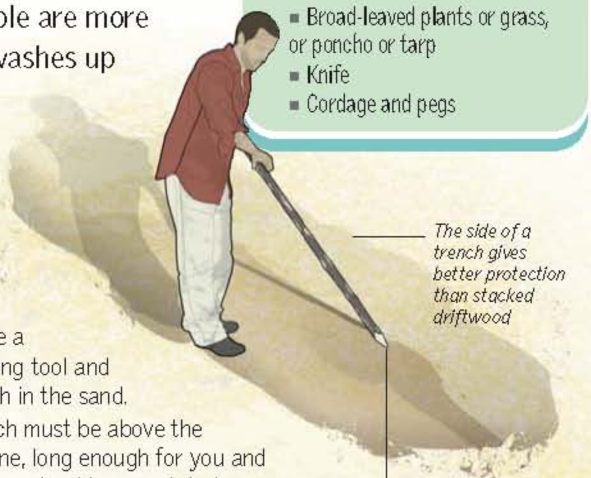
The side of a trench gives better protection than stacked driftwood
Digging tool with pointed end

1 Make a digging tool and dig a trench in the sand.

- The trench must be above the high-tide line, long enough for you and your equipment, wide enough to be comfortable, and deep enough so you can roll over without disturbing the roof.

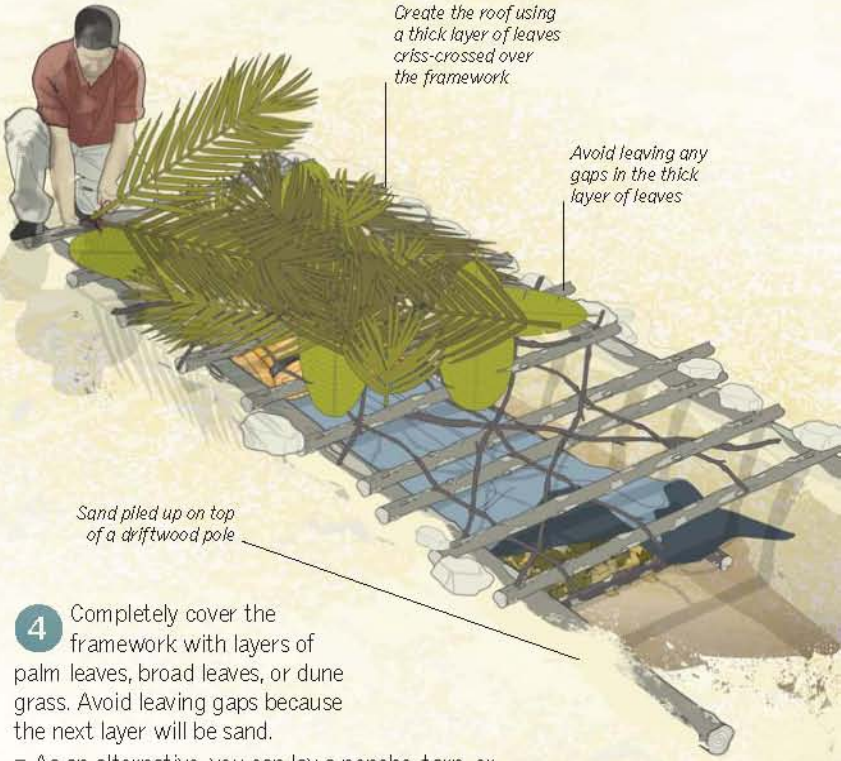
Create the roof using a thick layer of leaves criss-crossed over the framework
Avoid leaving any gaps in the thick layer of leaves
Sand piled up on top of a driftwood pole

4 Completely cover the framework with layers of palm leaves, broad leaves, or dune grass. Avoid leaving gaps because the next layer will be sand.

- As an alternative, you can lay a poncho, tarp, or shelter sheet over the framework. Weigh it down with sand or stones, or peg it out if you have cordage.

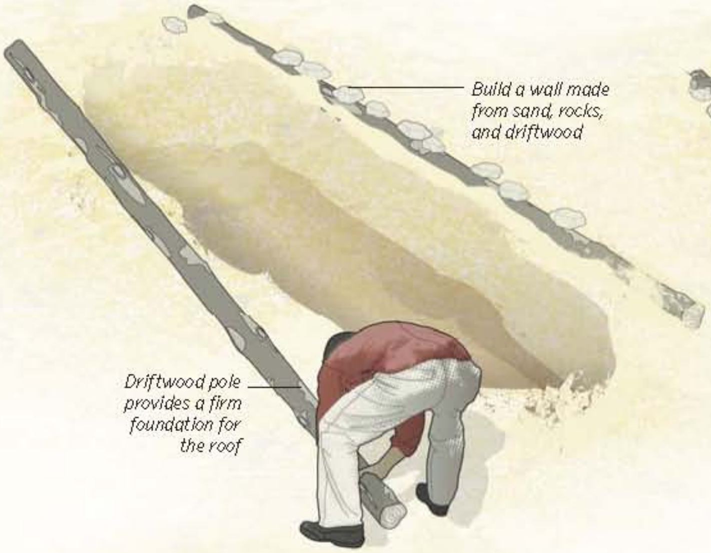

Build a wall made from sand, rocks, and driftwood

Driftwood pole provides a firm foundation for the roof

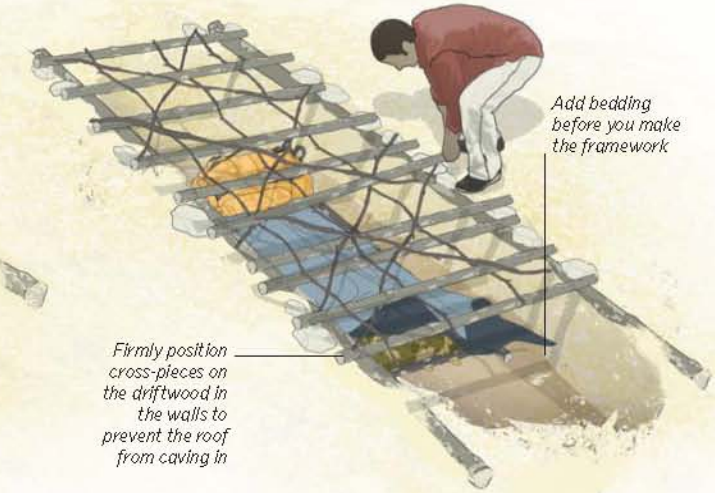

Add bedding before you make the framework

Firmly position cross-pieces on the driftwood in the walls to prevent the roof from caving in

2 Find two pieces of driftwood that are as long as your trench. Several shorter pieces in a line will work equally well if you can't find long pieces.

- Place the driftwood beside the trench and pile on sand and rocks to build up the height of the walls.

3 Place pieces of driftwood across the trench to form the framework of a basic roof.

- Weave smaller pieces of driftwood into the framework, making sure they sit firmly on the driftwood in the walls.

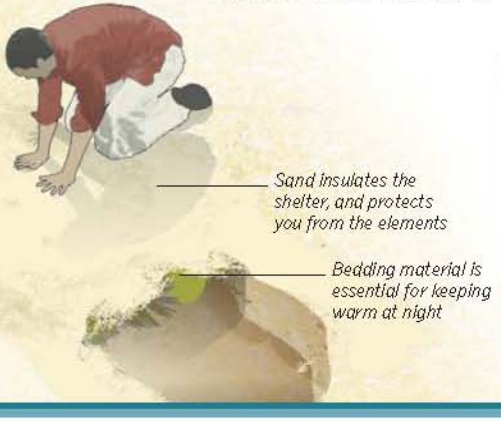

Sand insulates the shelter, and protects you from the elements

Bedding material is essential for keeping warm at night

5 Once you have covered the framework sufficiently, finish the shelter with a thick layer of sand or soil.

- The more cover you put on the roof, the more protected from the elements you will be. However, take care not to put so much weight on it that it collapses.

USING A LIFERAFT AS SHELTER

Remember, if you have abandoned ship in a liferaft, it can be used as a ready-made immediate shelter—even on land.

ON THE BEACH

The best place for a shelter is the landward side of the backshore: you can watch the sea, your daily fire can be seen from the sea, and you can see your shelter as you scavenge along the beach. In bad weather, however, the landward side of the dune crest is the best place.

PARTS OF A BEACH

A beach is shaped by the tides and the berm (a natural ridge) consists of deposits of the materials that make up the changing shoreline (sand, shingle, shale, and so on).

- At the top of the berm is a crest. A slope, or face, leads down from the crest to the water. At the very bottom of the face there may be a trough.
- The storm beach extends inland and this is where the wind and storms blow small particles of sand.
- Dunes form where the wind creates larger deposits of sand behind the beach.
- Longshore bars (sandbanks) may lie further out to sea and are formed where the waves first start to break.

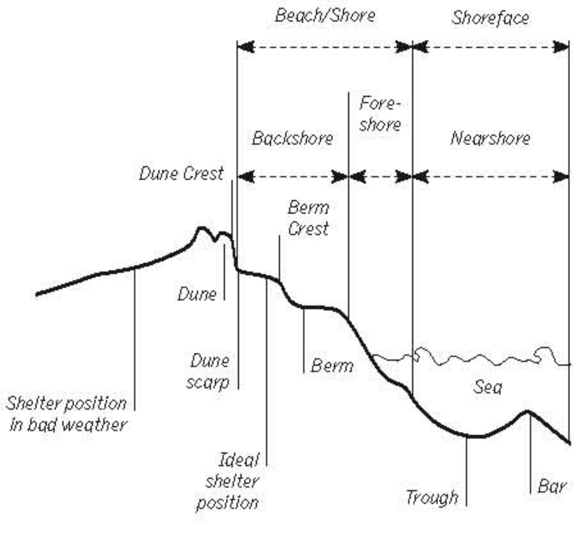

SNOW SHELTERS

THE TYPE OF SNOW SHELTER you can build depends on the kind of snow, the equipment you have, and the opportunities offered by the environment—for example, wooded areas are usually better than open areas because they provide protection and natural materials. You could build a snow trench, snow cave, tree pit, snow ledge, or a quinzhee.

TOOLS AND MATERIALS

- Boughs and branches
- Bedding
- Long pole
- Spade or pan for shovelling snow
- Long knife or saw
- Knife (or other long blade)
- Tarpaulin or shelter sheet

COMPACTED SNOW TRENCH

If the snow is compact enough, and if you have a long knife or snow saw, you can cut out blocks of snow to form a trench and then use the blocks to form an apex roof. This design requires a lot of effort and a bit of practice, but gives you a solid shelter with some additional height.

1 Use the blade of your knife to cut a rectangular trench that is the width, length, and depth you require.

- Cut blocks from the rectangle and place them at the side of the trench.

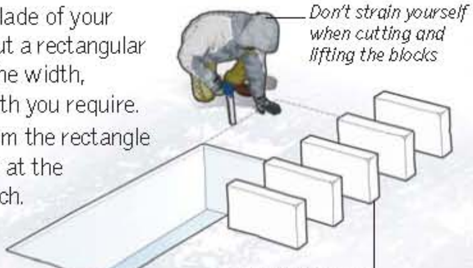

Don't strain yourself when cutting and lifting the blocks

Make each block 8–12 in (20–30 cm) thick

2 Trim the ends and rest two blocks against each other at the required angle for your roof.

- Use the other blocks to complete the roof and cut a triangle for the foot end.

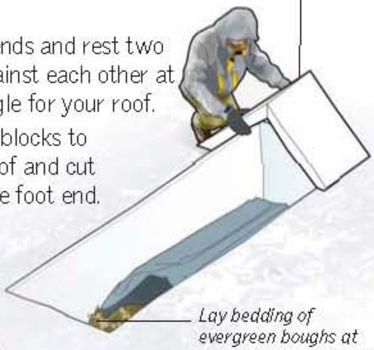

Carefully balance and fit the trimmed blocks, starting at the foot end

Lay bedding of evergreen boughs at least 10 in (25 cm) thick

3 Pack snow into any gaps between the blocks.

- Pile extra snow on top of the blocks to add insulation.
- Create a doorway from snow blocks or your backpack.
- Don't block the doorway completely, but allow for adequate ventilation.

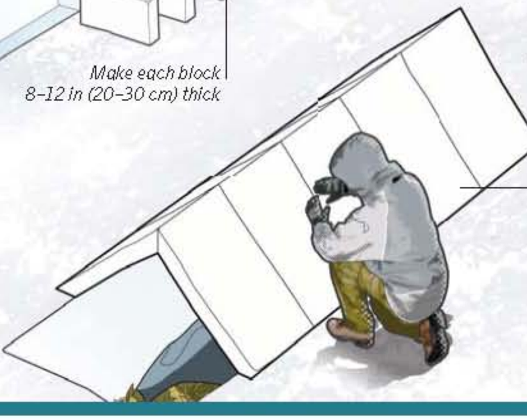

Use snow to seal the joins between blocks

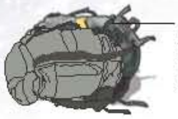

Your backpack can be used to partially block the entrance when you've finished

BUILDING SNOW SHELTERS

Study the snow around you to determine whether it's compacted or uncompacted before you start.

POINTS TO REMEMBER

The following tips may help if you intend to build and stay in a snow shelter:

- Make a shelter that's just big enough for you and your equipment. Don't spend hours building a shelter for only one night: do the least amount of work for the maximum amount of protection.
- If you need to peg out guylines in the snow, tie a short stick 6–12 in (15–30 cm) long to a line or cordage and bury it in the snow. Compact the snow down on top and, when the snow sinters (hardens), your peg will be held secure.
- Snow is an excellent insulator. Fresh, uncompacted snow is typically 90–95 percent trapped air. Since the air barely moves, the snow can keep you warm and dry if used correctly.
- Check your site for hazards, such as snowdrifts, freezing winds, avalanche, cornice collapse, and big animals.
- Create ventilation holes—make one near the ground to let fresh air in and one at the top to let air escape. Check the holes are clear every 1–2 hours.
- Keep tools inside the shelter in case you have to dig your way out.
- If you leave the shelter, mark the entrance. Take basic survival equipment with you: it's better to have and not need than to need and not have.
- Brush snow off your equipment and clothing before entering the shelter.
- Tie all vital equipment to yourself so that you can't accidentally drop and lose it in deep snow.

FOUND SHELTERS

If you're lucky you may find a tree that has space under it so you don't have to dig a trench. For example, under a low bough or in a tree pit—an area where a heavy and deep snowfall has built up around a tree, leaving little or no snow under the lower branches. Use your walking stick to check the depth of the snow, and also to check for pockets of air.

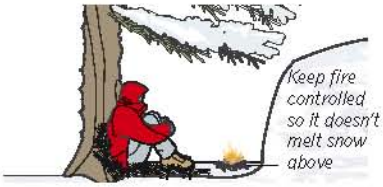
Keep fire controlled so it doesn't melt snow above

LOW BOUGH

Find an evergreen tree with a large branch lying on top of the snow. Dig down from the leeward side, excavate a pit, insulate the floor with boughs, and make sure you have adequate ventilation.

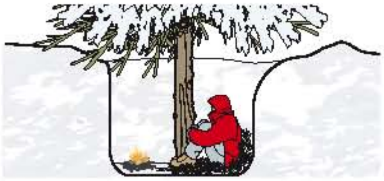

TREE PIT

Test the snow around a tree that has large, low, snowbound branches. Little or no resistance indicates a pit. Dig from the leeward side, place boughs on the floor, and ventilate the pit.

FIGHTER TRENCH

If the snow is soft, a fighter trench is quick and easy to build. In an emergency, you can even make a trench by kicking out the soft snow. First, find a location that protects you from the elements as much as possible, then test the depth of the snow with your walking stick or a pole.

WARMING AND COOLING TIPS

Whether you're building a shelter or resting inside it, the following tips may help you stay at the right temperature:

- Remove layers of clothing as you work, perhaps to your base layers with a waterproof layer on top. Keep clothing dry so that when you stop working you can put it on again.
- Keep shelters at a constant temperature. If snow melts and then re-freezes, it stops insulating.
- Avoid heating your shelter too much. Even a candle flame can raise the temperature inside by four degrees.

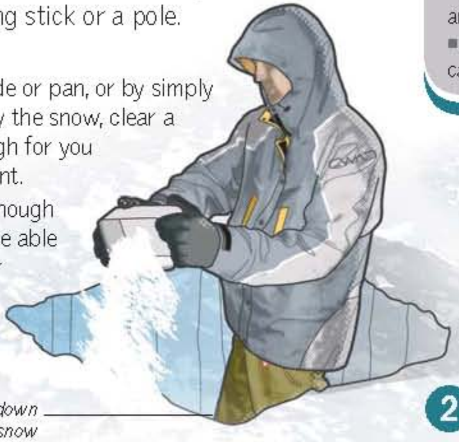
If you can, dig down through the soft snow to the ground

1 Using a spade or pan, or by simply kicking away the snow, clear a trench large enough for you and your equipment.

- Allow at least enough depth for you to be able to roll over in your sleep without disturbing the roof.

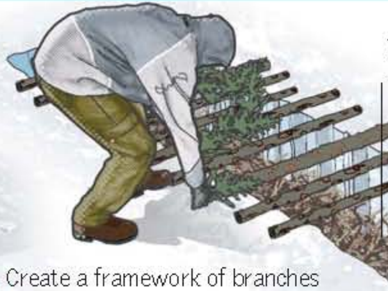
Add bedding before making the roof

2 Create a framework of branches and boughs across the trench.

- Make sure you have enough roofing material on top to stop the snow from penetrating through.

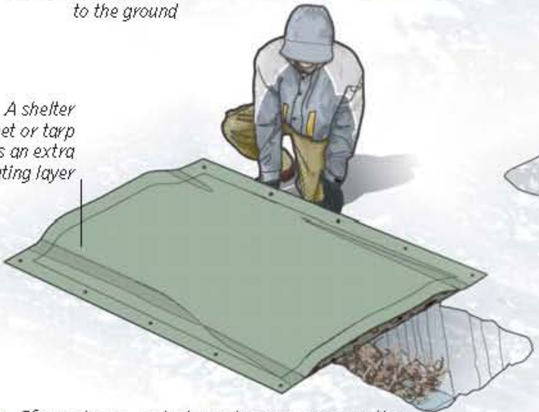
A shelter sheet or tarp makes an extra insulating layer

3 If you have a shelter sheet or tarpaulin, place it on top of the framework to form an extra insulating layer.

- Alternatively, you can use it to cover your bedding.

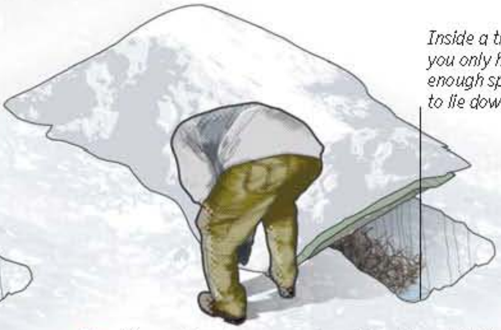
Inside a trench you only have enough space to lie down

4 Cover the framework with at least 12 in (30 cm) of snow to act as insulation.

- Dig a small pit at the entrance to allow easy access.
- At the entrance, you can make a small, controlled fire on a green log base.

QUINZHEE

A quinzhee is a dome-shaped shelter made by hollowing out a pile of settled snow. It's an overnight shelter that is easier to construct than the more permanent igloo ("igloo" means "home"), which is made from cut blocks of snow and requires skill and knowledge. You can't stand up in a quinzhee, but you can sit upright or just curl up.

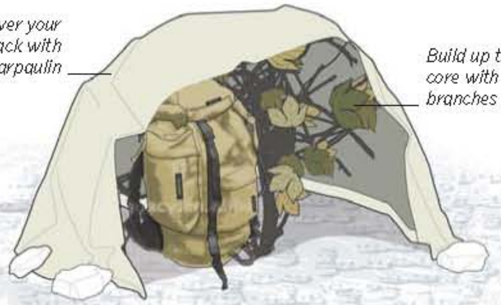

Cover your backpack with a tarpaulin

Build up the core with small branches and twigs

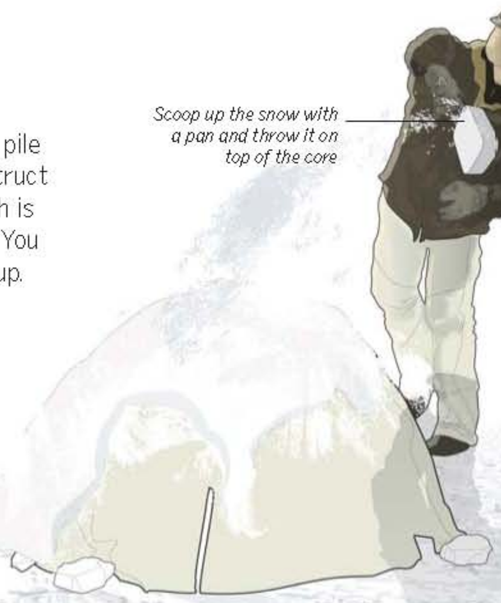

Scoop up the snow with a pan and throw it on top of the core

1 Find a relatively flat area that's covered with an abundance of snow.

- Mark a circle for your quinzhee, including walls that will be about 10 in (25 cm) thick, and stamp down the snow.
- Use your backpack and boughs or leaves covered with a tarpaulin to form the core of the shelter. Site the doorway at 90 degrees to the prevailing wind.

2 Use a container, pan, snowshoe, or other suitable item to gather as much soft snow as you can.

- Pile snow on the shelter's core to form a dome of the required height. Build up layers of snow until you have a covering of at least 10 in (25 cm), smoothing out each layer as you go.

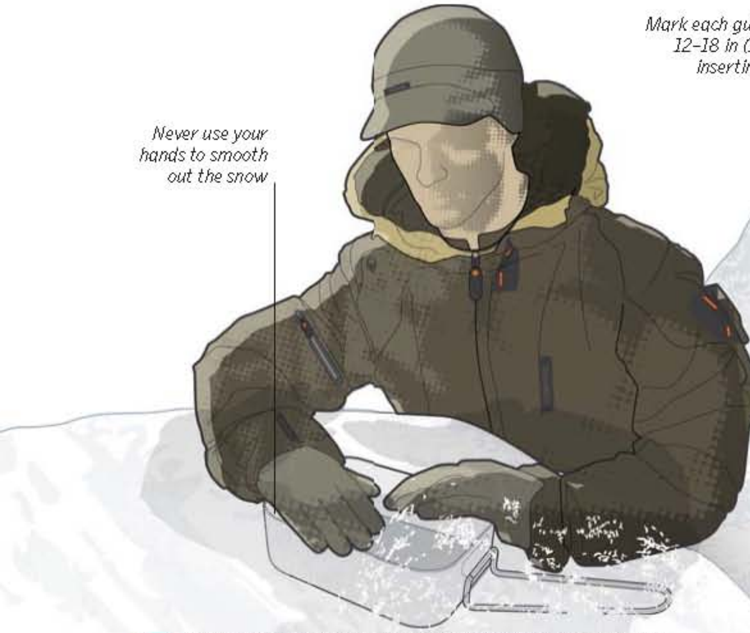

Never use your hands to smooth out the snow

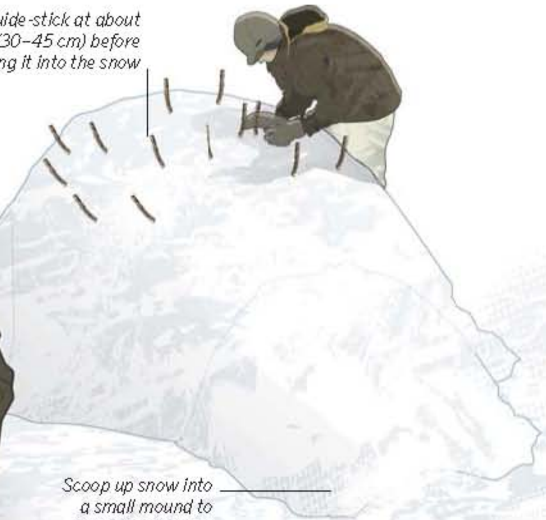

Mark each guide-stick at about 12–18 in (30–45 cm) before inserting it into the snow

Scoop up snow into a small mound to form the entrance

3 Smooth out the snow on the dome and leave it to sinter (harden) for 1–3 hours, depending on the type of snow and the ambient temperature.

- Keep active during this waiting period, particularly if it's windy: collect wood, build a fire (away from the quinzhee), and prepare yourself for the night.

4 To get an even thickness in the roof and walls, push guide-sticks of equal length through the snow toward the center of the dome.

- Build a small compact mound in front of the dome.

SNOW CAVE

A snow cave provides good protection but it takes a lot of work. Suitable locations might be compact snow on the lee side of a hill or an established snowdrift with a firm crust. Avoid snow that's newly fallen, powdery, loose, or shallow and uneven. The cave should be high enough to sit up in and deep enough for you and your equipment. Locate the entrance 45 degrees downwind to keep it free of drift.

BUILDING THE CAVE

- Dig a tunnel 3 ft (1 m) long into the snow bank, then excavate a cave on two levels. The lowest is the tunnel, while the other is a ledge at least 2 ft (70 cm) higher, and wide enough to sleep on.
- Create an arched ceiling to give the cave more support. Make it at least 18 in (45 cm) thick so it can take the weight of heavy snow.
- Mark the area with boughs to warn people where the shelter is.
- Insulate the sleeping ledge with plenty of natural material.
- Make at least one ventilation hole in the roof, but not facing into the wind.

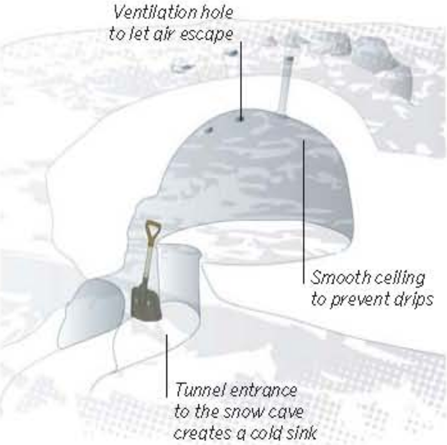
Ventilation hole to let air escape
Smooth ceiling to prevent drips
Tunnel entrance to the snow cave creates a cold sink

SNOW LEDGE

You can build a snow ledge from a well-packed drift of snow out of the wind, or at least perpendicular to it.

MAKING THE SNOW LEDGE

- Dig out blocks to form a doorway about 2 ft (70 cm) into the snow. Dig out blocks to form a horizontal rectangle above.
- Dig upward to create a sleeping platform and then form an arched ceiling.
- Place the cut blocks across the horizontal rectangle and seal any gaps with snow.
- Make a ventilation hole in the roof and one lower down to let fresh air circulate.

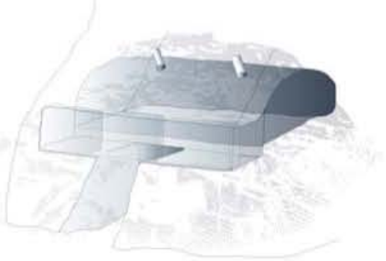

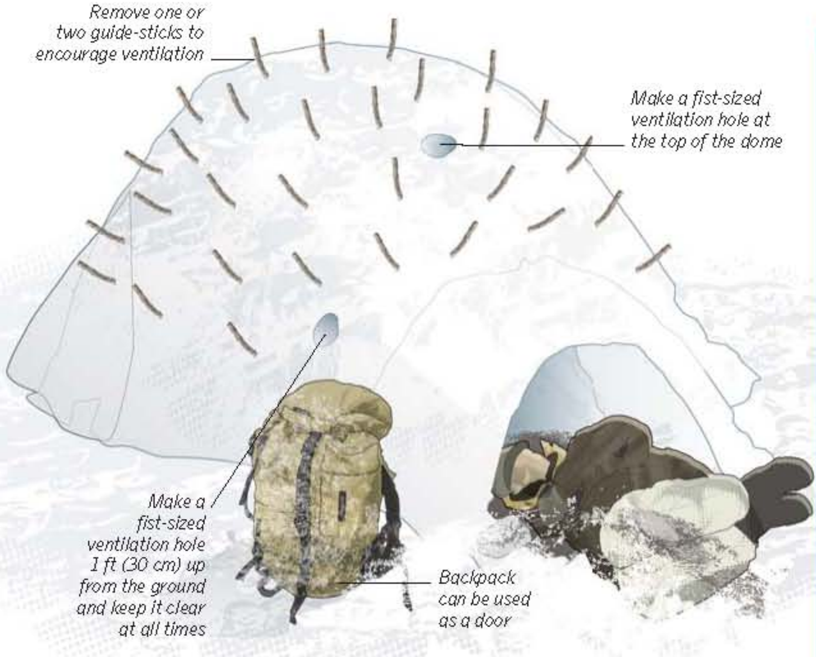
Remove one or two guide-sticks to encourage ventilation
Make a fist-sized ventilation hole at the top of the dome
Make a fist-sized ventilation hole 1 ft (30 cm) up from the ground and keep it clear at all times
Backpack can be used as a door

5 When the snow has hardened, burrow into the mound, remove your backpack and tarpaulin, and excavate snow from the core.

- Use the guide-sticks to keep the walls at least 10 in (25 cm) thick.
- Don't use your hands to dig as they will get cold and wet.
- Smooth out the snow on the inside to prevent drips from forming.
- Build a raised sleeping platform. This creates a cold sink (an area into which the coldest air falls), taking cold air away from your sleeping area.

CARBON MONOXIDE

Carbon monoxide is an odorless gas that's produced when there's not enough oxygen to create carbon dioxide from burning fuel. Carbon monoxide poisoning can be fatal in a well-insulated, non-ventilated environment.

PREVENTION IS BETTER THAN CURE

Create one or more holes, 3–4 in (7.5–10 cm) in diameter, at the base and top of your shelter. Make sure the air flows out unobstructed.

SIGNS OF POISONING

Carbon monoxide poisoning is cumulative and can build up over a few days. Mild effects are fatigue, faintness, and flu-like symptoms. As it progresses, the effects are severe headaches, nausea, and decreased mental coordination.

TAKE ACTION

Get into the fresh air at once. You need to breathe fresh air for at least four hours to reduce the amount of carbon monoxide in your system by half.

DETECTORS

Carbon monoxide detectors are widely available, but tend to be made for the home or RV. Some models are battery-operated, but they're too bulky for camping. You can wear a patch that changes color when carbon monoxide is in the air, giving an early visual warning.

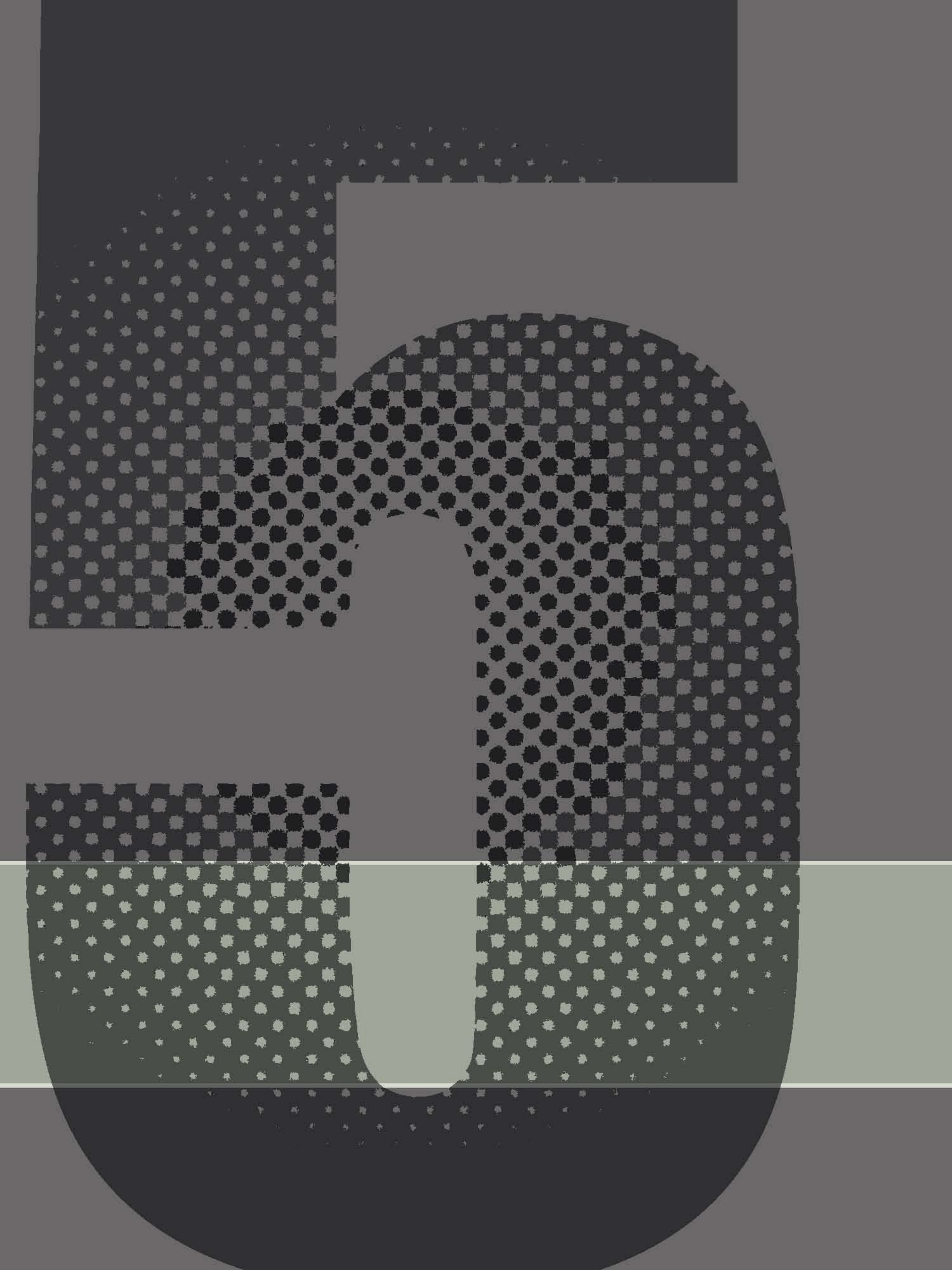

WATER AND FOOD

FIND AND TREAT WATER

The importance of water, even in a short-term survival situation, should never be underestimated. Water is essential to life, and a regular intake of 2–3 liters ($4\frac{1}{5}$–$6\frac{1}{3}$ pints) a day is needed just to maintain your water balance and prevent dehydration. The amount required can increase dramatically depending on factors such as the temperature of the environment, your age and physical condition, your workload, and whether you have been injured. It's not unusual for UK Royal Marines operating in desert or jungle environments to require 14 liters ($3\frac{3}{4}$ gallons) per day. We tend to take water for granted, not appreciating just how important it is until we don't have any—at which point it becomes the most important thing in the world.

You should always plan your treks around your need for water and your ability to replace it as required. There are many hydration systems available and many small and efficient methods of filtering and purifying water on the trail. In a survival situation, you should always strive to filter and purify any

In this section YOU WILL DISCOVER...

- that **birds** can **show you** the **way** to **water**...
- when to **suck** on a **small pebble**...
- how to **make** a **gypsy well** and a **solar still**...
- why a **Finnish marshmallow** could **save your life**...
- how to **absorb water** without **drinking it**...
- how to **improvise** a **basic bladder**...
- the **importance** of **surgical tubing**...

water before drinking it. Bringing water to the boil will kill all water-borne diseases. In the short-term, stomach bugs from contaminated water may not kill you but they can seriously affect your ability to carry out other survival tasks. However, if you have no choice, it's better to drink contaminated water than not to drink at all. That way, a doctor will at least be able to treat you, whereas dehydration will kill you—and death can't be treated! Never drink urine or salt water, as these will only dehydrate you more.

If a source of water is undrinkable, such as salt or stagnant water, or even urine, you can still produce drinking water if you have the means to start and maintain a fire.

THERE ARE MANY WAYS to purify water, but if you have a fire you will always be able to distill it—and therefore make it drinkable. Build your fire close to the water source (if the water can be poured into a container or a hole dug in the ground where it will not seep away quickly, this will make the process more efficient). When the fire is established, place rocks in it (don't use slate or other layered rocks as they might shatter). Once the rocks are heated, use a forked stick or similar to transfer them into the salt or foul water. Then, suspend a piece of absorbent material, such as a T-shirt or moss, over the hot rocks to collect the steam. The steam will condense on the material and the resulting water, which you can wring out, will be fit to drink. If you have a cooking pot or survival tin, you can simply fill it with water, boil it over a fire, and collect the steam in the same way.

If you have absolutely no means of treating or boiling water, you should try to:

- Find the clearest flowing water and collect it from the surface.
- Filter debris out, even if only through a sock.
- Remember, it's better to drink foul water than not to drink at all.

> “We tend to take **water for granted** until we don't have any—at which point it becomes the most **important thing in the world**”

THE IMPORTANCE OF WATER

YOU REQUIRE A STEADY supply of water to sustain yourself in a survival situation and without it you will dehydrate. Left unchecked, dehydration will end in death. To survive, a balance between water intake and loss must exist.

WHY YOU NEED WATER

Water is essential to life. It is needed, directly or indirectly, for every physical and chemical process that takes place in your body. Here are just a few of the functions that water performs:

- **Delivery service:** water carries oxygen, nutrients, and other essentials around the body.
- **Waste remover:** the kidneys use water to flush out toxins via urine.
- **Coolant:** water regulates the body's temperature.
- **Breathing aid:** the lungs use water to moisten inhaled air so that it doesn't irritate the sensitive pulmonary linings.
- **Sensory aid:** water helps conduct nervous impulses around the body.
- **Shock absorber:** water protects the vital organs and provides lubrication around the joints.

HOW MUCH WATER DO YOU NEED?

How much water you need to survive a particular situation is dictated by a number of factors, such as your physical state, the environment you are in, and your exertion levels. Even when resting in the shade, the average person will lose more than 1 liter (1¾ pints) of water each day just through breathing and urination, a figure that increases dramatically once loss of water through sweat is taken into account. A minimum of 3 liters (5¼ pints) per day is required to remain healthy in a survival situation, with this amount increasing for higher temperatures and heavier workloads.

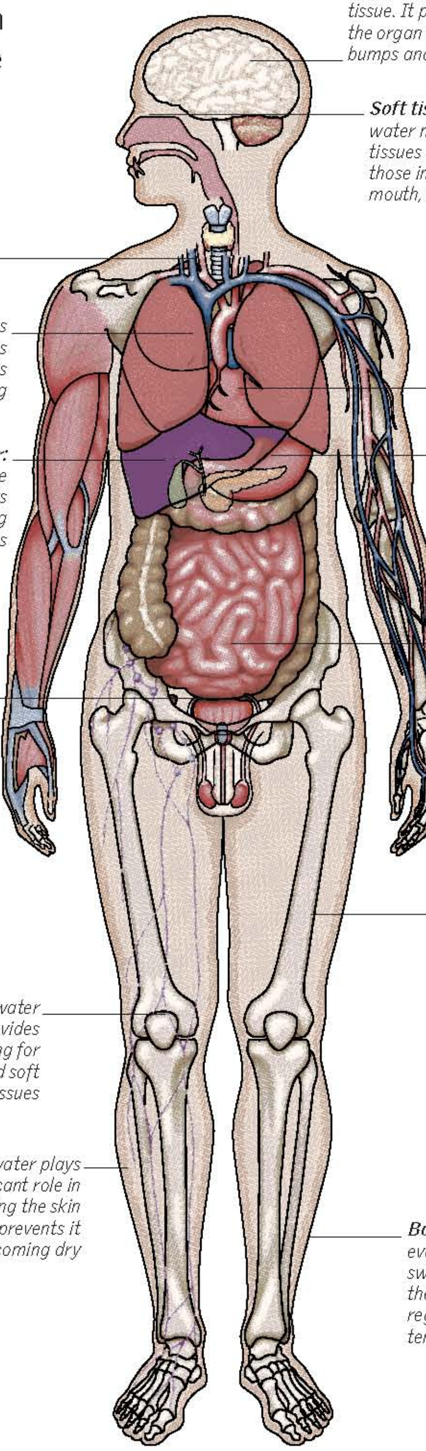

Brain: *water makes up approximately 80 percent of brain tissue. It protects the organ from bumps and jarring*

Soft tissue: *water moistens tissues such as those in the nose, mouth, and eyes*

Blood: *water plays a significant role in regulating blood pressure*

Heart: *higher water consumption helps remove fatty acids from the body and reduces the risk of a heart attack*

Lungs: *water helps moisten the lungs and assists breathing*

Kidneys and liver: *water lessens the burden on kidneys and liver by flushing out waste products*

Stomach: *water helps the body digest food in the stomach and turn it into energy*

Mineral absorption: *water helps dissolve minerals and other nutrients to make them accessible to the body*

Bladder: *water enables the body to flush out waste and toxins through the bladder*

Bones: *water makes up approximately 22 percent of bones*

Joints: *water provides cushioning for joints and soft tissues*

Skin: *water plays a significant role in hydrating the skin and prevents it from becoming dry*

Body temperature: *evaporation of sweat (water) from the skin helps regulate body temperature*

WHAT IS DEHYDRATION?

Dehydration occurs when you fail to replace the water your body loses. It's vital to recognize the symptoms early. Factors that can lead to dehydration include high and low temperatures, humidity, work-rate, clothing, body size, fitness levels, and injury.

THE EFFECTS OF WATER LOSS		
1–5% LOST	**6–10% LOST**	**11–12% LOST**
▪ Thirst	▪ Dizziness	▪ Stiffness of joints
▪ Discomfort	▪ Dry mouth	▪ Deafness
▪ Urine becomes darker	▪ Blueness of extremities	▪ Defective vision
▪ Loss of appetite	▪ Slurred speech	▪ Shriveled skin
▪ Impatience	▪ Swollen tongue	▪ Lack of feeling in skin
▪ Drowsiness	▪ Blurred vision	▪ Inability to swallow
▪ Lethargy	▪ Tingling in limbs	▪ Delirium
▪ Nausea	▪ Inability to walk	▪ Unconsciousness
▪ Headache	▪ Difficulty in breathing	▪ Death

WATER-BORNE DISEASES

Water-borne diseases are caused by ingesting water contaminated by the faeces or urine of humans or animals that contains protozoa, viruses, bacteria, or intestinal parasites. Globally, they cause 10 million deaths a year.

	DISEASE	SYMPTOMS
PROTOZOA	Cryptosporidium	Loss of appetite, nausea, and abdominal pain, usually followed by profuse, foul-smelling, watery diarrhea, and vomiting.
	Giardiasis	Loss of appetite, lethargy, fever, vomiting, diarrhea, blood in the urine, and abdominal cramps.
VIRUS	Infectious hepatitis (Hepatitis A)	Nausea, loss of appetite, mild fever, aching muscles, dark-colored urine, jaundice, and abdominal pain.
BACTERIUM	Amoebic dysentery	Feeling of fatigue and listlessness. Feces may be solid, but will smell foul and contain blood and mucus.
	Bacillary dysentery (Shigellosis)	Fever, abdominal pain, muscle cramps, high temperature, and blood, pus, and mucus in stools.
	Cholera	Vomiting, poor circulation, cold and clammy skin, muscle cramps, rapid dehydration, and increased heart-rate.
	E. coliform	Diarrhea and vomiting. Can cause death in vulnerable groups such as the very young or the elderly.
	Leptospirosis	Jaundiced appearance, lethargy, high temperature, aching muscles, and vomiting. Can be fatal if not diagnosed early.
	Salmonella	Nausea, diarrhea, headaches, stomach cramps, fever, possible blood in the feces, and vomiting.
PARASITES	Bilharzia	Irritation to the urinary tract and blood in urine, rash or itchy skin, abdominal pain, cough, diarrhea, fever, and fatigue.
	Hookworms	Anaemia and lethargy. Larvae travel to lungs and are coughed up and swallowed into the stomach, where they grow into worms.

TOO MUCH WATER?

Hyponatraemia is a condition caused when excess water accumulates in the body at a higher rate than it can be excreted. It results in a diminished sodium concentration in the body's plasma and the swelling of the body's cells. It can lead to a swollen brain and other neurological problems and, in extreme cases, coma and death. The way to prevent hyponatraemia is to control the amount of water that you drink, and to regulate your body's salt intake. If you don't have salt or sodium tablets in your survival tin, you can filter saltwater through fabric to sift out the salt content.

RATIONING YOUR WATER

If your water supplies are limited, you'll have to use what rations you have efficiently until you're rescued. If your water rations will not last that long, you'll have to procure water yourself. There is much debate about the advantages and disadvantages of drinking no water for the first 24 hours of a survival situation, but at this early stage it's best to make sure you're adequately hydrated. Your particular circumstances will dictate what's best to do, but always consider the following:

- The incident that put you in the survival situation may have been both dramatic and stressful; this will make you thirsty.
- In the first 24 hours of a survival situation, you'll be addressing the principles of survival: protection (shelter) and location (see p. 27). This is hard, thirsty work.
- Physical factors—such as seasickness, injury, or the environment you're in, such as a desert—may dictate that water-rationing is not a viable option.
- If you only have a limited water supply, but drink nothing for the first 24 hours, you may end up being so dehydrated that what little water you have in your possession will have no positive effect on your dehydrated state.

FINDING WATER: TEMPERATE CLIMATES

THE ABILITY TO LOCATE drinking water can challenge even the most experienced survivor and the challenges can vary dramatically according to the environment and the local conditions. It is, therefore, extremely important that you're aware of all the potential sources of water for the environment you're in.

WARNING!
In a survival situation, even if you think your chances of rescue are high and that help is imminent, you should start to look for water sources as soon as you have addressed the immediate problems of protection and shelter (see pp. 154–81). Remember that you can survive for up to three weeks without food; without water, you won't be able to survive for more than a few days.

COLLECTING WATER

You can find water in a range of sources, which vary in terms of quality and accessibility. Your priority is to locate the best and most accessible source of water in your immediate environment.

CATCHING RAINWATER

Collecting rain as it falls is the safest way to procure drinking water. It will need no treatment prior to drinking, as long as the catchment device itself has not been contaminated. Any number of non-porous materials—such as tarpaulins, ponchos, flysheets, survival blankets, or even large leaves—can be used as a rain-catchment device. Bear in mind that the larger the surface area of the material, the more rainwater you'll be able to catch.

1 Select a place as close to your camp as possible where your catchment device will be exposed to the most possible rain.

2 Firmly secure the tarpaulin to four stakes (sticks of equal length could be used) using string. Make sure that one end is higher than the other to provide a natural run-off for the water.

3 Place a heavy stone in the middle of the tarpaulin, roughly two-thirds of the way toward the lower end, to create a channel into which the rain will run down from the tarpaulin's sides.

4 Place a container, such as a pan or can, beneath the end of the channel to collect the water as it starts to flow off the tarpaulin.

MATERIALS AND TOOLS
- Tarpaulin
- Four sticks
- String
- Heavy stone
- Container

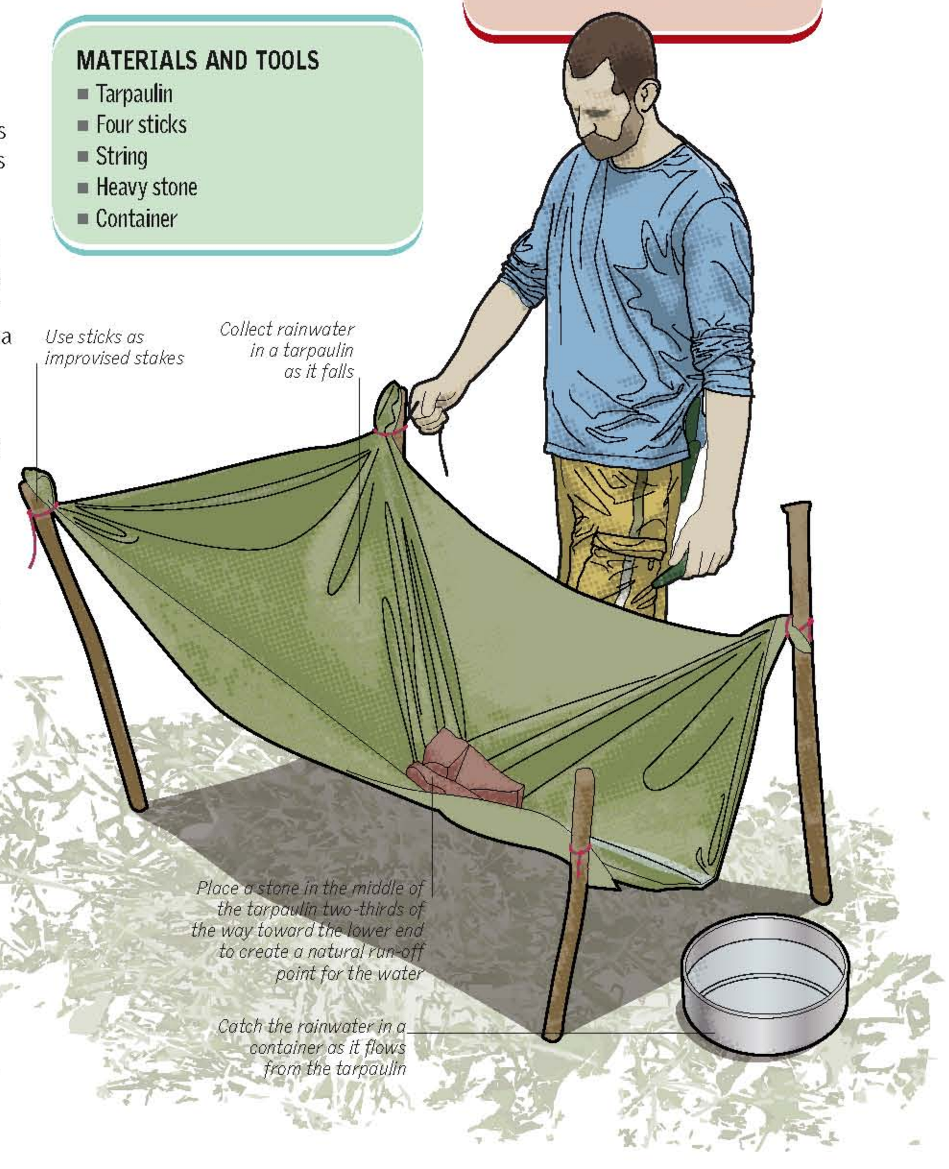
Use sticks as improvised stakes

Collect rainwater in a tarpaulin as it falls

Place a stone in the middle of the tarpaulin two-thirds of the way toward the lower end to create a natural run-off point for the water

Catch the rainwater in a container as it flows from the tarpaulin

LOCATING OTHER NATURAL WATER SOURCES

If rainwater is not available, there may be other natural sources of water, from easily visible streams and rivers, to bores and holes, which can be hidden. Whatever your source, you should always treat the water before drinking.

SOURCE	CHARACTERISTICS
SPRINGS	These occur when the water is forced to the surface as a result of subterranean pressures or from gravitational flow from higher sources. Found in low-lying areas, springs normally provide a permanent water source. Contrasting green vegetation is an indicator of their presence.
STREAMS AND RIVERS	Rivers and streams are an invaluable source of water, but there are still a few things you need to bear in mind when collecting water from them. The closer to the mountaintop the river or stream is, the clearer the fast-running water will be. The further downstream the water travels, the more likely it is to pick up minerals, debris, and pollutants that might be harmful to you. If possible, either check upstream for dead animals that may have contaminated the water, or follow the water downstream. Always try to collect the fast-flowing water near to the surface. In arid areas, rivers and streams tend only to flow during floods and will contain more pollutants and debris.
ROCK HOLES	Usually found in high ground, rock holes are natural collectors of rainwater. If the water appears to be trapped deep down, you can use your surgical tubing to retrieve it. Make sure you filter and purify the water before you drink it.
WELLS AND BORES	In some areas, you may find wells or old bore holes. Wells, which may be featured on local maps, can be deep and covered, making water procurement more difficult. In remote areas, wells are covered and marked in certain ways by the locals—find out what the markers are for your area.
LAKES AND PONDS	Rivers, streams, and water run-offs all flow into lakes or ponds. If you're collecting water from these sources, always try to collect it as it runs into the body of water, as lakes and ponds are more static than flowing water and can, therefore, become increasingly stagnant. Try to find the cleanest-looking area and avoid areas where debris has collected or where algae has formed. Note that the presence of fish indicates that the water is still oxygenated.
SEEPAGE	Usually located at the base of cliffs, high ground, or rocky outcrops, seepage is caused by slow-running channels that drain off these features.
SOAKS	Soaks are found close to rivers and creeks in low-lying areas, and are normally lower than the existing water table. The presence of soaks is often indicated by vegetation and they may be subject to pollution as a result of their use by animals. Make sure you filter and purify the water before you drink it.

AVOIDING POTENTIAL DANGERS

Most water sources are likely to be used by animals for drinking, bathing, urinating, and defecating, so always filter and purify any water collected before use (see pp. 200–01); the only exception to this rule is if you have collected rainwater. When collecting water, you should also be aware of the following dangers.

- It is possible that you'll encounter dangerous animals either using the water source, or on the way to or from the water source.
- Almost all major water sources will have a ranking system to determine which animals can use it. If all of the small gazelles suddenly disappear, ask yourself why.
- If you're collecting water from rivers, be aware of the potential dangers of river wildlife, such as crocodiles and snakes.
- If using dry riverbeds during the rainy season, be aware that flash floods can move quicker than you can run.

LOCATING HIDDEN WATER SOURCES

The presence of water is usually indicated by signs of life, such as green vegetation, animal tracks, or human habitation. Even if the terrain you're in appears lifeless, there may be plenty of indicators of a possible water source.

USE THE TERRAIN

- Observe the landscape for patches of green. Be aware that vegetation may not need obvious surface water for survival and may get its water from deep roots that tap into moisture below the surface.
- Water is subject to gravity and is more likely to be found downhill or in low-lying areas, such as valleys, dry riverbeds, narrow canyons, gullies, and at the base of cliffs or rock formations. Green vegetation forms next to a river and decreases as the ground rises away from the water source.
- Water will often seep inland on coastlines, leaving behind wetlands that contain water with tolerable levels of salt or that can be distilled in a solar still (see p. 189) to procure fresh water.

WATCH THE ANIMALS

- The Bedouin listen to the twittering of birds at dawn and dusk and follow their flight path to discover where they drink.
- Flocks of birds circling over one spot are usually flying over a water source. This does not apply to meat-eating birds, such as vultures, eagles, or hawks, who get their fluid requirements from the meat they eat.
- All finches and grain-eaters need a regular supply of water. Observe their flight patterns to locate a water source.
- Bees need water, so beehives are never far from a water source.
- Animal tracks, especially those of herd animals, will often lead to a water source. Look for converging sets of tracks.
- Flies stay close to water and the presence of mosquitoes almost certainly means that water is nearby.
- Look out for herd animals, such as buffalo, hippos, elephants, impala, and wildebeest, as they depend on water.

PROCURING WATER

Even if you find yourself in an environment without any obvious water sources, it doesn't necessarily mean that water will not be available to you. There are a number of water-procurement techniques that may mean the difference between you making it through a survival situation or not.

SOURCING WATER FROM DEW

Dew is water in the form of droplets that form on exposed surfaces in the early morning or late evening and can provide an invaluable supply of fresh water. It occurs when the temperature of a surface is low enough to allow the moisture in the warmer air above it to condense. Dew can easily be collected from any non-porous surface—such as a car roof or a tarpaulin—with a piece of cloth that can then be wrung out into a container. You can also harvest or trap dew.

HARVESTING DEW

You can harvest dew by walking through a field of long grass before sunrise or late in the evening with a piece of absorbent material—such as rags or a T-shirt—tied around your ankles.

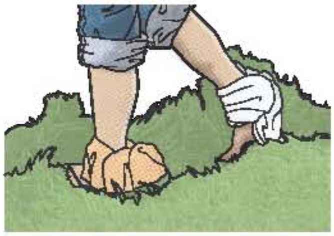

① Tie the material tightly around each ankle and walk through the dew-covered grass. It will absorb the dew as you move.

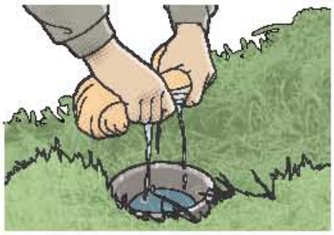

② Wring the rags to extract the water. Repeat the process until you have an ample supply of water or the dew has evaporated.

DELAYING DEHYDRATION

When water is scarce, the following points will help you delay the onset of dehydration:

- Conserve what water you have and use it as efficiently as possible.
- Work only in the coolest part of the day and avoid sweating.
- If the sun is shining, seek shade and keep covered.
- Suck a small, smooth button or pebble to help stimulate saliva and remove the sensation of thirst.
- Avoid eating protein-rich food as it requires more water to digest than those in other food groups.

MAKING A DEW TRAP

Dig a hole about 18 in (45 cm) deep, line with a plastic sheet, and fill with smooth, clean stones. Water will condense on the stones overnight. Harvest as early as possible the next morning to ensure it does not evaporate.

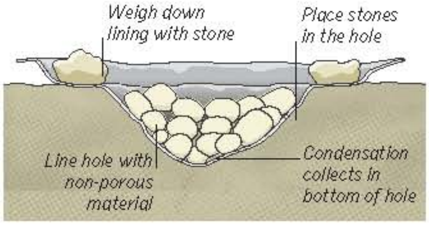
Weigh down lining with stone
Place stones in the hole
Line hole with non-porous material
Condensation collects in bottom of hole

SOURCING WATER FROM PLANTS

Transpiration is the evaporation of water from a plant, primarily from its leaves. You can collect this vapor to boost your fresh water supply. All you need is a clear plastic bag.

MAKING A TRANSPIRATION BAG

Place a smooth rock in the lower corner of the plastic bag, and place the bag over the leaves of a tree branch, tying the end. As water evaporates from the leaves, it will condense on the inside of the plastic bag and collect at its lowest point.

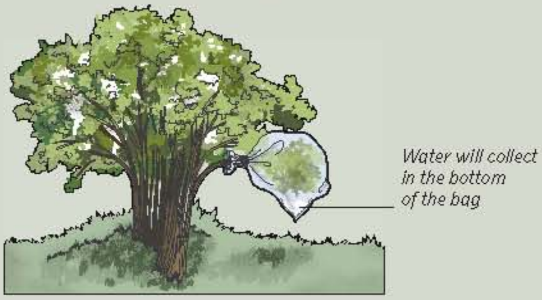
Water will collect in the bottom of the bag

MAKING A VEGETATION BAG

Cut green vegetation and place it in a plastic bag. Place a smooth rock in the lower corner of the bag and tie off the open end. Secure the bag in direct sunlight. The sun causes water in the leaves to evaporate, condensation forms on the plastic bag and drips to the lower corner of the bag.

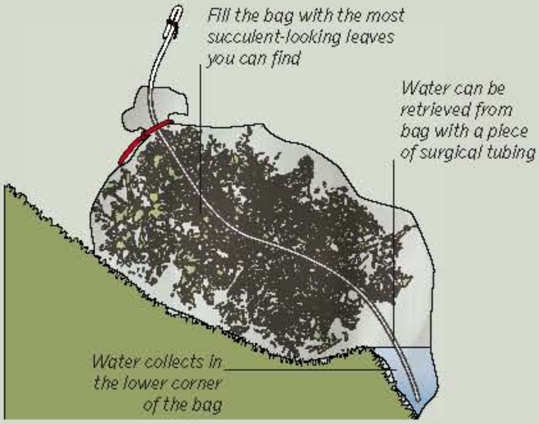
Fill the bag with the most succulent-looking leaves you can find
Water can be retrieved from bag with a piece of surgical tubing
Water collects in the lower corner of the bag

MAKING A SOLAR STILL

A solar still works using the same principle as a vegetation bag (see opposite). It collects potable water from the vapor that is produced by vegetation, water that is unfit to drink, or moisture from the ground.

1 Ideally find, or dig, a hole in the ground at least 2 ft (0.6 m) wide and 2 ft (0.6 m) deep. Place an empty container in the center of the hole. Fill the hole with vegetation, a receptacle containing undrinkable water, such as saltwater or urine, or fabric soaked in undrinkable water.

2 Cover the hole with a plastic sheet, and secure it in place with stones. Place a stone in the center of the sheet to create a run-off point for the water. The sun's heat will evaporate water from the vegetation, or distill the undrinkable water, producing pure water vapor. The water vapor, which is now free of contaminants, will condense on the underside of the plastic sheet, and drip into the container where it can be collected.

TOOLS AND MATERIALS

- Spade or shovel
- Two containers
- Vegetation and/or undrinkable water
- Piece of fabric or clothing
- Plastic sheet, tarpaulin, or space blanket
- Stones
- Surgical tubing

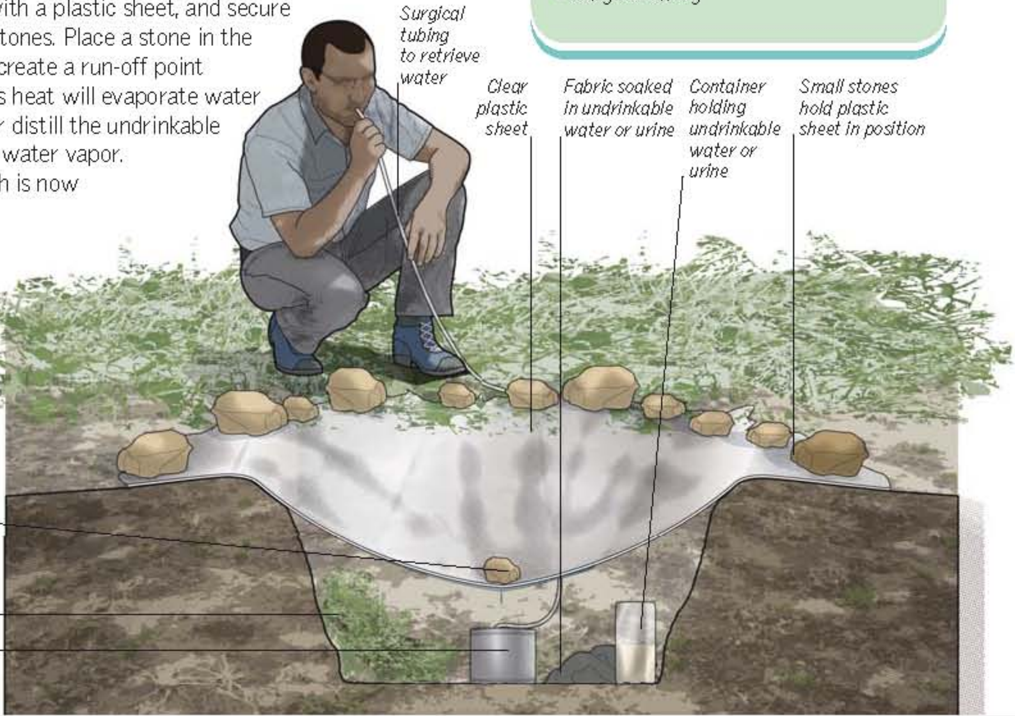
Surgical tubing to retrieve water
Clear plastic sheet
Fabric soaked in undrinkable water or urine
Container holding undrinkable water or urine
Small stones hold plastic sheet in position
Small stone forms a drip point on underside of plastic
Fresh, green vegetation
Container catches condensed water droplets

MAKING A GYPSY WELL

A good method of using the ground to create cleaner water from stagnant water sources, a gypsy well can also be used to collect water from saturated ground. The water collected through this method will still need to be treated before it is safe to drink (see p. 201).

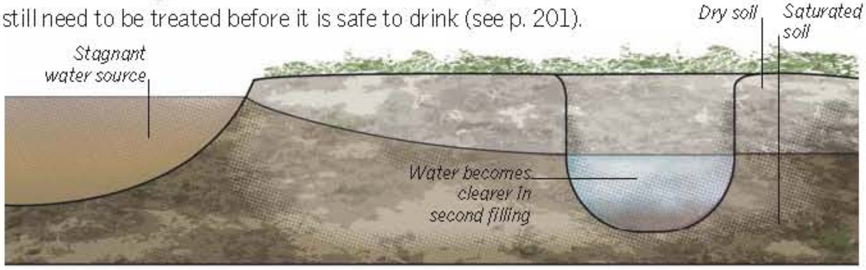
Stagnant water source
Dry soil
Saturated soil
Water becomes clearer in second filling

1 Dig a hole a few feet away from a stagnant water source. It should be wide enough for you to be able to dip into it with a container and at least 12 in (30 cm) below the first layer of saturated soil. The hole will quickly fill with water.

2 Bail out the muddy water and allow the hole to refill—you may have to repeat this process several times. Once the water is clear, it can be collected, treated, and used. Cover the well when not in use to prevent debris or small animals falling in.

SURGICAL TUBING

Always pack a length of surgical tubing in your survival kit. It doesn't take up much space and is an extremely versatile piece of equipment:

- It can be used as an improvised straw to reach water trapped in rocks, hollows, and trees.
- It allows you to access water procured from a solar still without taking the still apart.
- It can be built into transpiration and vegetation bags so that you can remove water without having to open the bag.

FINDING WATER: HOT CLIMATES

YOUR NEED FOR WATER increases in hot climates as your body starts to use more of its water supply to regulate its temperature through perspiration. If you fail to drink more than you perspire, you will start to dehydrate and even approaching the first stages of dehydration can affect your chances of survival. Hot climates can be divided into two groups: hot-humid and hot-dry.

WARNING!
Water sources are plentiful in jungle and rain forest environments and, under normal circumstances, you should have no problem getting hold of enough water to satisfy your needs. However, rivers and streams may not be available during certain seasons, or if you have climbed too high, so knowledge of other water procurement methods could be crucial to your survival.

HOT-HUMID CLIMATES

The hot-humid conditions found in jungles and rain forests mean that procuring water is rarely an issue. However, the body's need for water in these conditions should not be underestimated; it is not unusual to have to drink up to 3 ³⁄₄ gallons (14 liters) a day to avoid dehydration.

SOURCING WATER FROM VEGETATION

Many plants, such as pitcher plants, have hollow parts that collect rainfall or dew. Some trees store and catch rainwater in natural receptacles, such as cracks or hollows. In an emergency, life-saving liquid can be garnered from a tree's roots or sap. You can find water trapped in the sections of green bamboo by carefully cutting into the bamboo with your machete or knife; or collect small, unripe coconuts and quench your thirst with the fluid they contain; or make a spigot and tap into the water contained inside the water tree. Life-saving fluid is everywhere in the jungle, and you don't have to look far to find it.

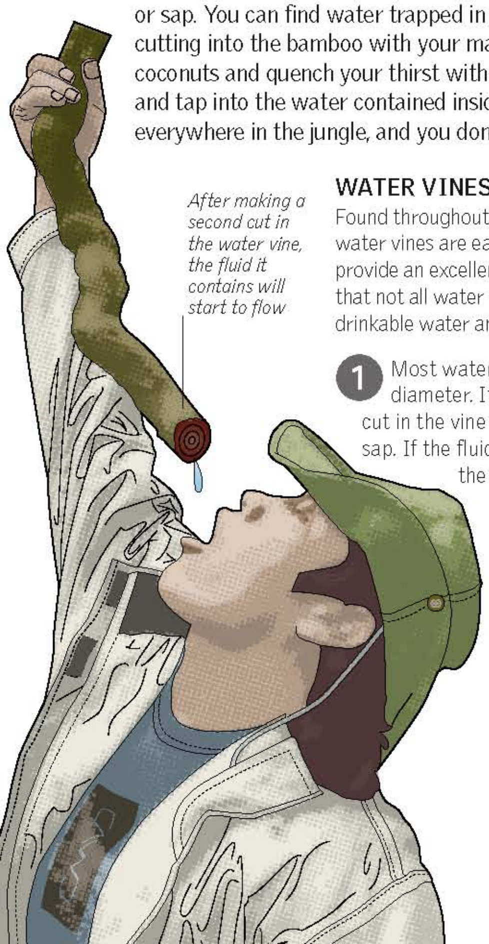
After making a second cut in the water vine, the fluid it contains will start to flow

WATER VINES

Found throughout the jungles and rain forests of tropical regions, water vines are easily identifiable by their size and shape and can provide an excellent sort of fresh water. However, bear in mind that not all water vines are water-bearing; not all contain drinkable water and some even contain poisonous sap.

1. Most water-bearing vines are about 2 in (5 cm) in diameter. If you think you have found one, make a small cut in the vine with a machete and check the color of the sap. If the fluid is milky, don't drink it; if it's clear, then the water in the vine will be safe to drink, so cut through the vine as high as you can with a knife or machete.

2. Cut off the vine at a point lower than the first cut. The liquid, which has a neutral, fruity taste, will start to flow. Don't let the vine touch your mouth as the bark may contain irritants.

3. The pores in the upper end of the vine may re-close, stopping the flow of water. To rectify this, simply cut the top of the vine again with your machete.

CATCHING RAINWATER

Catching rainwater is the best way to procure water: it is passive and requires no energy to collect once you place your containers. There are many forms of catchment device, but make sure you filter and purify the water (see pp. 200–01) before drinking it.

BAMBOO ROOF

Construct a sloped bamboo roof with a bamboo gutter. This could be the roof of your shelter (see p. 169), but if water procurement is a problem, you will have to construct additional bamboo roofs.

WIDE-LEAF ROOF

If you're in an area where there are wide-leafed plants, it's easy to construct a roof from them. Overlay the leaves as you would with roof tiles, working from the bottom to the top (see p. 170). Placed in this way, they will allow the water to run to the bottom. A length of bamboo cut in half lengthways can be placed as a collection gutter at the bottom.

BAMBOO DRAINPIPE

Observe rainwater's route down a tree trunk and tie a length of bamboo that has been cut in half lengthwise in its path. Place the other end of the bamboo into a suitable container.

DRIP RAGS

Wrapping any absorbent material around a leaning tree, such as a rag or a T-shirt, will result in the water running down the tree and soaking the material. Shape the rag to form a low point from which water will drip and place a suitable container underneath to catch the water.

HOT-DRY CLIMATES

Anyone venturing into this environment should have sufficient water for their needs, plus an emergency supply just in case, otherwise they have no right to enter the desert. Green vegetation usually signifies water or moisture in some form, and many techniques of procuring water in temperate climates (see pp. 188–91) may work in some desert conditions.

SOURCES OF EMERGENCY FLUID

If there is no surface water to be found, and you have no other means of procuring water, a water-yielding plant may be your only option. In some plants, the clear sap, fruit, or trapped rainwater, may quench your immediate thirst, but do not rely on these sources to keep you alive for long.

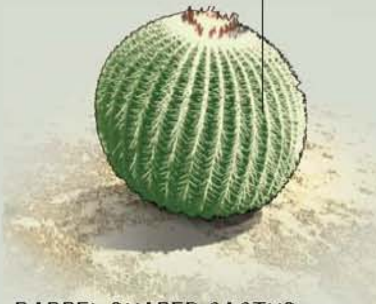

The barrel-shaped cactus is characterized by numerous ribs and hundreds of sharp spines

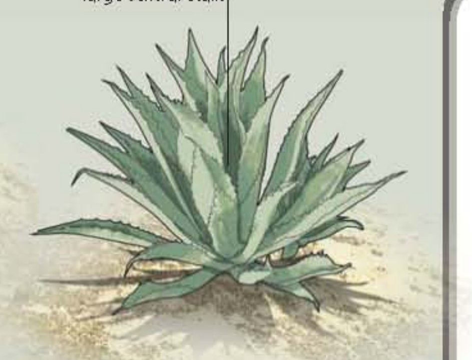

Agaves have large clusters of thick, fleshy leaves surrounding a large central stalk

BARREL-SHAPED CACTUS

Barrel-shaped cacti contain a milky fluid that's safe to drink. Carefully remove the top of the cactus with a machete, mash up the flesh inside with a stick to make a pulp, and extract the juice from the pulp by sucking through a hollow grass stem. Alternatively, use a cloth to soak up as much of the fluid as possible and then wring the material to extract the fluid. The rewards from both techniques are minimal.

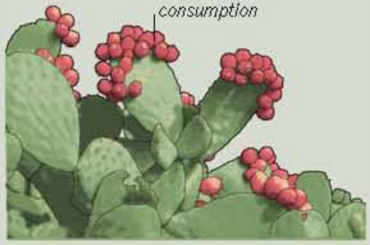

The prickly outer layer of the fruit has to be removed before consumption

PRICKLY PEARS

Found in low-spreading clumps measuring 3 ft (0.9 m) in height, and native to dry, sandy soils throughout the world, the prickly pear cactus has an edible fruit that can provide a life-saving amount of fluid.

AGAVES

Native to Mexico and the southern and western United States, agaves have a rosette of thick, fleshy leaves containing fluid that is safe to drink. Cut the huge flower stalk with a knife or machete and collect the juice.

FINDING UNDERGROUND WATER

Water is life in the desert; wherever there is life there'll be water. Unusual clusters of green plants may indicate a minor presence of water; an abundance of greenery may indicate a more substantial water source. Water is rarely found above a depth of 6 ft (1.8 m) below the surface, so you will have to dig for it. If you do, make sure you only do so during the coolest part of the day. Bear in mind that, despite what you may have seen in movies, or read in other books, your chances of procuring water from the inside bend of a dry wadi in a real survival situation are slim to non-existent. In fact, you're more likely to waste what water you do have looking for the dry wadi and die in the process!

THINK LATERALLY

The chances are that if you're having to procure water in the desert you'll already be in a desperate situation. Bear in mind that, in addition to the methods mentioned above, if you have knowledge of the techniques outlined in the temperate environment section (see pp.190–91)—such as solar stills, dew traps, and vegetation and transpiration bags—it could be enough to keep you alive in the desert.

CHECKLIST FOR HOT-DRY CLIMATES

Even the best-laid plans can hit unforeseen problems, but many problems encountered in a desert environment can be avoided with some prior preparation.

- Always start hydrated.
- Carry enough water for your needs plus emergency water; your emergency supply should be enough to get you out of danger.
- Monitor your progress against the water you use. If you're using more than you thought, re-evaluate what you want to accomplish. It's better to turn back and learn from the mistake than to push on and create a survival situation that need not exist.
- Cache water ahead if necessary.
- Check your map for probable water sources. Confirm the reliability of these sources with locals and ask if there are any sources, such as wells, that are not shown on the map.
- Mark your map, or waypoint your GPS, with any water sources you sight as you progress. It's better to go back to a known source than to move on with nothing more than hope.
- To keep the water cool, always keep the bottles in the shade or in a windy location.

FINDING WATER: COLD CLIMATES

ADOPT THE SAME METHODS of trying to find drinking water in cold climates as you would in temperate ones (see pp. 186–91), although the ability to procure water becomes a major problem in freezing temperatures. The cold-weather survivor faces a dilemma: he or she is surrounded by water, but could die of thirst because most of that water is frozen.

WARNING!

Never try to melt ice or snow in your mouth, as it can cause freezing injuries to your mouth and lips. What's more, your body will expend heat as it melts the ice, and this could cause you to cross that very fine survival line between being cold or being hypothermic.

CHECKLIST FOR COLD CLIMATES

You should prioritize water usage in cold-weather conditions in the same manner as you would in desert conditions (see p. 193).

- One of your first thoughts should be about procuring water. You will have to find a water source that's close to everything you need to build and maintain a fire.
- Always look for an alternative water source before trying to melt snow or ice. It's easier, less time-consuming, and more fuel-efficient to fill your water containers with natural meltwater than it is to melt snow or ice by the heat of a fire.
- Your ability to procure water in freezing conditions will be directly related to your ability to start and maintain a fire.
- Be aware that it takes time and patience to melt snow and ice. You'll also need enough fuel to maintain a fire for a considerable period of time.
- Regulate your body heat to minimize overheating and sweating.
- Keep drinking water close to your body to prevent it from freezing, but avoid having water containers directly next to your skin. Instead, keep them between layers of clothing and use the warmer air trapped between the layers to help raise the water's temperature.
- Do not use recently frozen sea water, as it contains high levels of salt.

MELTING ICE AND SNOW

If you have the choice between melting ice and snow, favor ice, as it melts more quickly than snow and is up to 17 times denser. However, if you can't find ice, use dense, compact snow. Always look for the whitest, purest-looking ice or snow.

MELTING ICE

If you have some water to begin with, pour some into a container and heat it over a fire. Break the ice into small pieces—rather than adding it in one lump—and keep adding pieces to the container to melt them. Keep the water hot, but not boiling, so that you avoid losing water through evaporation.

USING A HOT PLATFORM

If you don't already have water for the method above, you can melt ice slowly on a gently sloping platform fashioned from any flat piece of stone or wood positioned above a fire.

1. Build a fire. Search for a large stone with a flat surface and two logs, or smaller stones, that are strong enough to support it. Place the small logs or stones on either side of the fire as a support for the flat-surfaced stone. Make sure that the platform is at an angle. This will create a natural run-off for the ice as it melts. Place a block of ice in the center of the platform.

2. As the fire starts to heat up the platform, the ice will begin to melt. The meltwater will flow off the platform, where it can be collected in a container, such as a mess tin.

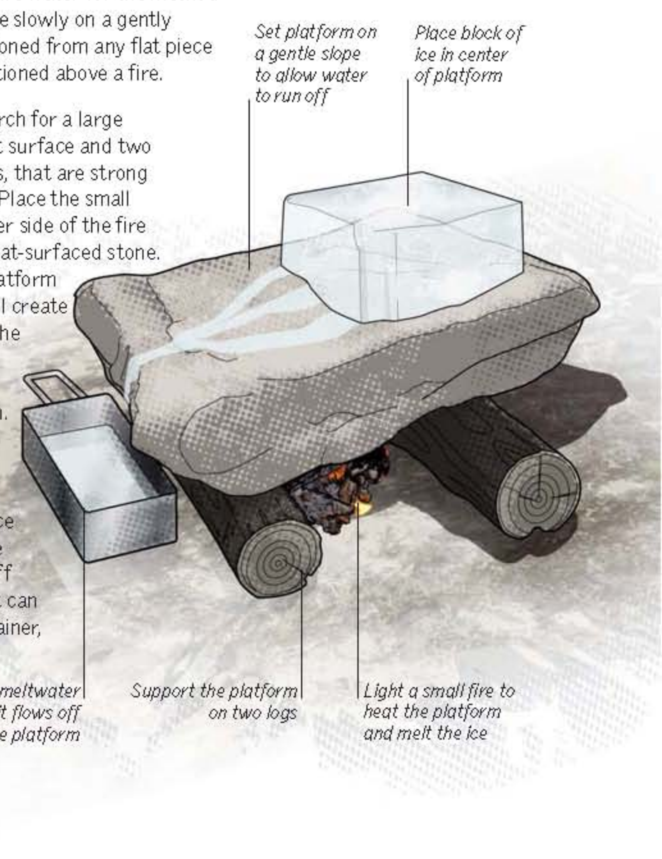
Set platform on a gentle slope to allow water to run off
Place block of ice in center of platform
Collect the meltwater in a pan as it flows off the platform
Support the platform on two logs
Light a small fire to heat the platform and melt the ice

MELTING SNOW

If you already have water, follow the techniques for melting ice (see opposite), heating a little water and adding the snow little by little. Don't pack the snow too tightly; if an air pocket forms, the heat from the fire will be absorbed by the metal container rather than by the snow and could result in the fire burning a hole through the metal container before it melts the snow.

MAKING A FINNISH MARSHMALLOW

Cut a solid piece of dense snow—often referred to as a "marshmallow" or a "snowman's head"—and skewer it with a stick. Secure the stick in the ground close enough to a fire that it receives heat, and position a suitable container underneath it to capture the water as it melts.

STORING WATER

Snow is a great insulator: even if the temperature dips to -40°F (-40°C), water in a bottle will remain largely unfrozen if placed under at least 1 ft (0.3 m) of snow. Make sure you store the bottles upside down. That way, if some of the water does freeze, it will freeze at the bottom of the bottle and not at the top.

USING A MELTING SACK

Using a similar principle to the Finnish marshmallow (see left), this technique involves using an improvised sack—made from any porous material, such as a T-shirt or a sock—suspended near to a fire. The heat from the fire will start to melt the snow, which can then be collected in a well-positioned container.

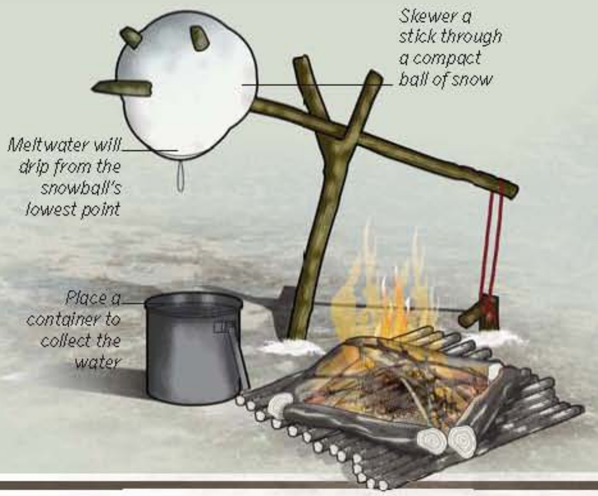
Skewer a stick through a compact ball of snow
Meltwater will drip from the snowball's lowest point
Place a container to collect the water

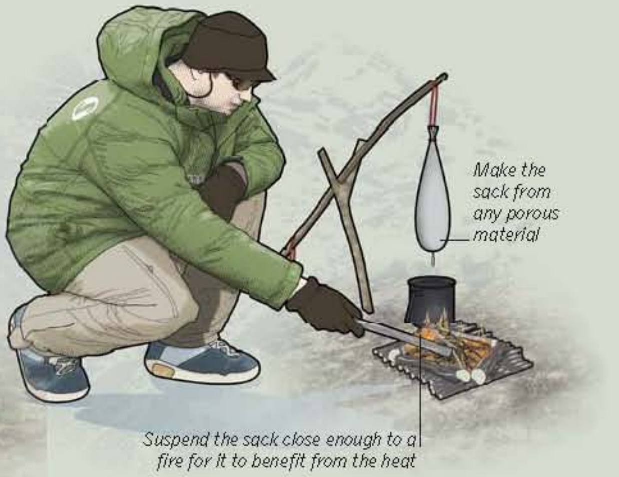
Make the sack from any porous material
Suspend the sack close enough to a fire for it to benefit from the heat

IN A WORST-CASE SCENARIO

If you're caught out on barren snow or ice, you're unlikely to have access to natural fuel to burn in order to melt snow or ice into water. As long as you have your second-line survival equipment (see p. 42), you'll have all you need to procure life-saving water.

USING YOUR FIRE TIN

Set up your fire tin on the ground and shelter it from any wind using either your body or your pack. Using your survival tin as a container, place small quantities of ice or snow into the tin. Light your fire tin and place your survival tin over the flame. Add more ice or snow as it starts to melt.

USING YOUR MATCHLESS FIRESET

Place the lid of the matchless fireset tin on the ground. Light a hexi fuel block using the cotton wool and flint/striker. Using the fireset container or your survival tin, melt small quantities of ice or snow over the flame. Once the ice or snow has melted, let the container cool so you don't burn yourself. The water is then safe to drink.

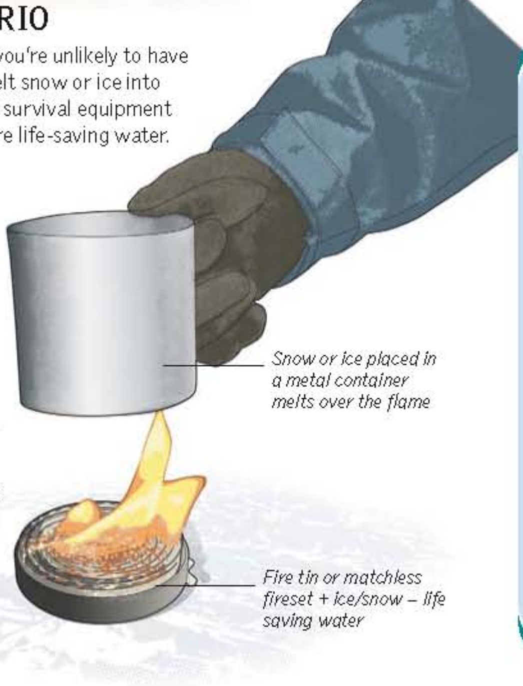
Snow or ice placed in a metal container melts over the flame
Fire tin or matchless fireset + ice/snow – life saving water

USING BODY HEAT

When Inuit hunters capture a caribou, they empty out the contents of its stomach, turn the stomach inside out, fill it with snow, and tie it shut with a length of intestine. They then put the stomach back in the cavity while they skin the caribou. By the time they have finished, the heat given off by the animal's body has melted the snow. The Inuit then open the bag carefully and suck out the water through a block of snow to filter it.

FINDING WATER: AT SEA

OF ALL ENVIRONMENTS, the sea is possibly the most difficult in a survival situation. It offers no natural resources for protection against the extremes of temperatures, wind, rain, and sea state, and provides little to aid location. Being surrounded by water that you can't drink only adds to the difficulty. Some devices are capable of making seawater safe do drink, but if you don't have one, you'll have to find a way of procuring enough fresh water to keep you alive.

WARNING!

Never drink salt water. Its salt concentration is three times higher than that of blood and ingesting it will dehydrate you. Continued use over a prolonged period will lead to kidney failure and, ultimately, death.

CONSERVING YOUR WATER RATIONS AT SEA

Rationing fresh water supplies when you're adrift at sea is a sensible precaution, as you have no idea how long it will be before you're rescued or before you reach land. Here are a few tips to help you conserve what water you have while you're at sea:

- Fix your daily water ration after taking stock of the amount of water you have, the output of solar stills and desalting kits, and the number and physical condition of your party.
- Prevent fresh water supplies from becoming contaminated by saltwater.
- Keep water supplies well shaded, both from the overhead sun and from the glare off the sea's surface.
- In hot conditions, dampening your clothes with saltwater can help to lower your body temperature—but don't overdo it. This is a trade-off between cooling yourself down and the saltwater boils and rashes that will result from continued exposure.
- Don't exert yourself. Relax and sleep whenever possible.
- Use every container you have—even a simple trash bag—to collect rainwater, and keep them well sealed and attached to the raft.
- If you don't have water, don't eat. Protein consumption will hasten the onset of dehydration.

COLLECTING FRESH WATER

If you find yourself adrift with no hope of immediate rescue, obtaining drinking water will be a major priority. If you don't have a solar still or a reverse-osmosis pump, you must find another means of procuring fresh water. Fortunately, there are several ways of doing so.

GATHERING RAINWATER

Most modern liferafts incorporate a built-in rainwater catchment system that channels rainwater and dew from the outer surface of the liferaft into collection pockets inside the liferaft. However, even if you're not in a liferaft, you can construct a similar system using a tarpaulin or any other waterproof material. Watch the clouds, be ready for the possibility of any showers (see pp. 80–81) and spread your tarpaulin in a bowl shape to catch the largest amount of rainwater possible. Always place a tarpaulin before nightfall, so that you don't miss out on any overnight rainfall.

HARVESTING DEW

At night, secure the tarpaulin like a sunshade and turn up its edges to capture dew. It's also possible to harvest any dew that may have collected on the sides of the raft using a sponge or cloth that you then wring out.

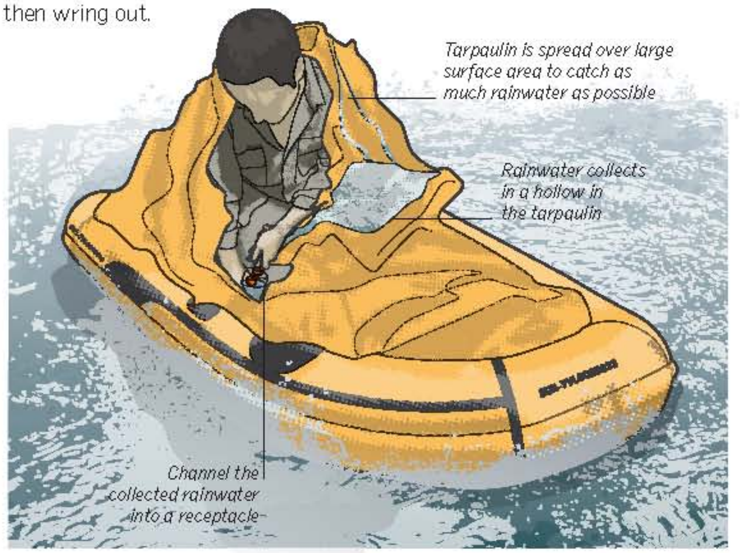

Tarpaulin is spread over large surface area to catch as much rainwater as possible

Rainwater collects in a hollow in the tarpaulin

Channel the collected rainwater into a receptacle

TREATING SALTWATER

If you have no means of collecting rainwater or dew, there are several products capable of turning undrinkable saltwater into fresh water. Although these products are standard issue on most liferafts, you should always try to have at least one of them with you if you are venturing into a marine environment.

SOLAR STILL

Solar stills are a simple way of distilling water using the power of the sun. Saltwater is placed at the bottom of the container, where it is evaporated by the sun through clear plastic. Pure water condenses on the top of the plastic and drips down to the side, where it can be collected, often via a tube. Most solar stills on modern liferafts are inflatable.

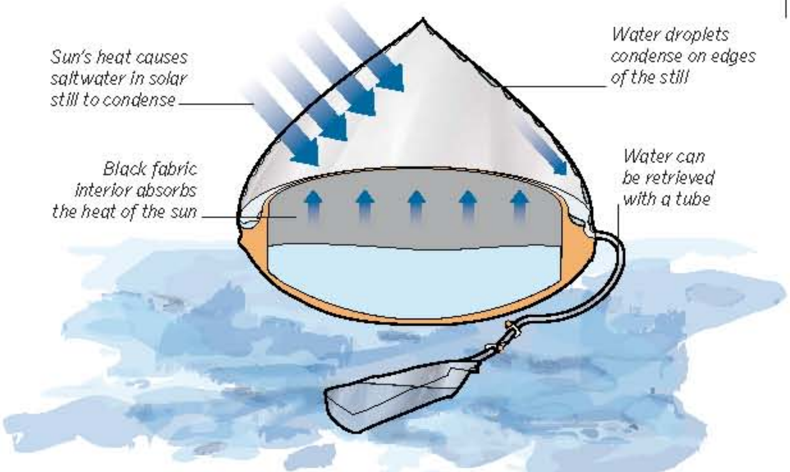
Sun's heat causes saltwater in solar still to condense
Water droplets condense on edges of the still
Black fabric interior absorbs the heat of the sun
Water can be retrieved with a tube

MAKE YOUR OWN SOLAR STILL

If you have a chance to gather together the right materials, it's easy to make a small solar still. All you need are two containers (one larger than the other), a sheet of plastic, some string, and a weight to form a natural run-off point for the water as it condenses under the heat of the sun. If you have some surgical tubing, you can use it to retrieve the fresh water without taking the still apart.

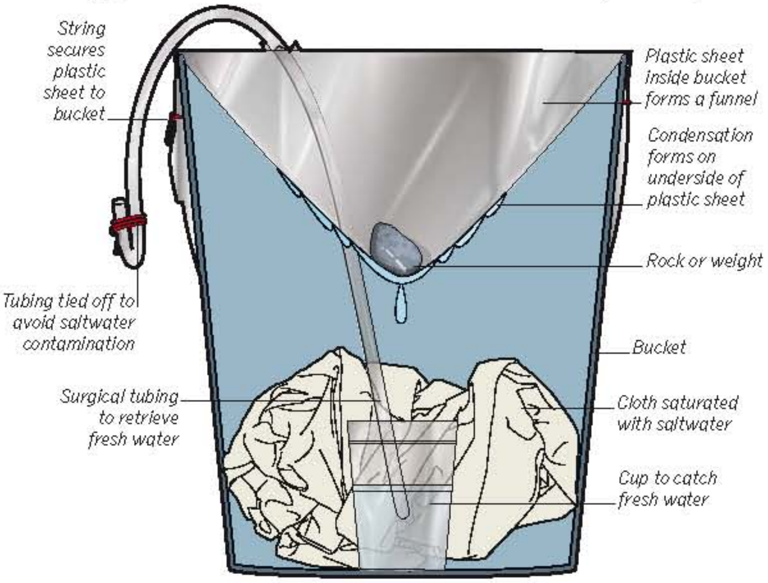
String secures plastic sheet to bucket
Plastic sheet inside bucket forms a funnel
Condensation forms on underside of plastic sheet
Rock or weight
Tubing tied off to avoid saltwater contamination
Bucket
Surgical tubing to retrieve fresh water
Cloth saturated with saltwater
Cup to catch fresh water

REVERSE-OSMOSIS PUMP

These hand-powered devices pump seawater at a very high pressure through a membrane that filters out the salt. Depending on the model, they can produce around 23 liters (49 pints) of fresh water per day.

DESALTING KIT

These kits turn seawater into freshwater through a process called "ion exchange." Because they only produce small amounts of fresh water over several hours, use desalting kits only during long periods of overcast weather when you can't use a solar still.

THE LAST RESORT

In absolute emergencies, potentially life-saving forms of liquid can be obtained from the ocean.

SEA ICE

In Arctic seas, you can obtain drinking water from old sea ice. This ice is bluish in color, has rounded corners, splinters easily and, more importantly, is nearly free of salt. New ice is gray, milky, hard, and salty. Water from icebergs is fresh, but icebergs are dangerous to approach, so you should only use them in an emergency.

FISH

Drink the aqueous fluid found along the spine and in the eyes of large fish. Carefully cut the fish in half to get to the fluid along the spine and suck the eye. If you are so short of water that you need to do this, do not drink any of the other fluids: they are rich in protein and fat and your body will use up more of its water digesting them than it obtains from them.

SEA TURTLE

Sea turtle blood has a salt concentration similar to that of humans. The blood can be collected by slitting the turtle's throat. Note that, although this may help prolong survival, sea turtles are an endangered species, so you should only kill one as a last resort.

REHYDRATION ENEMA

If you have some water that is not salty or poisonous, but is too foul to take orally, you can absorb as much as a pint a day—enough to keep you alive—through the large intestine using an improvised tubing device.

CARRYING AND STORING WATER

YOU WILL NEED TO make many decisions in any survival situation, some of which could mean the difference between life and death. One of these decisions will be whether to stay and wait for rescue or to attempt self-rescue. A major factor in this decision will be the availability of water and your ability to carry and store it.

WARNING!
Mark your containers clearly. Late at night in camp it's easy to confuse bottles that contain treated water with ones that contain untreated water or camping fuel.

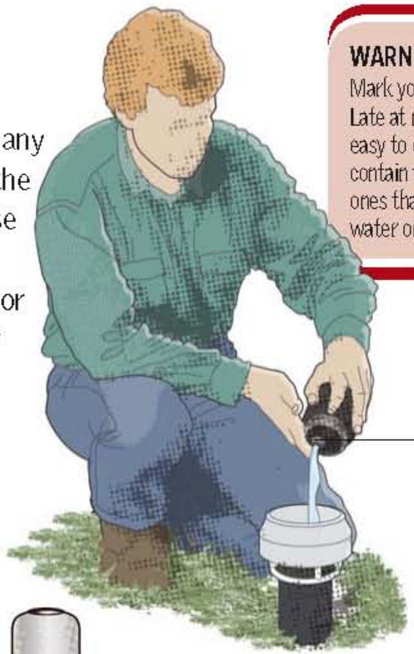
If your container has a lid, it can be used to scoop water from a stream before you pour it into a larger water bottle

WATER CONTAINERS

There are many different kinds of container for carrying both hot and cold liquids. They range from solid plastic or steel flasks or bottles, to collapsible waterproof bags that can be folded up when they're empty.

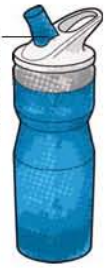
Fitted with a push-pull nozzle for ease of use while on the move

PLASTIC WATER BOTTLE

Strong and lightweight, with screw-off tops or push-pull nozzles.

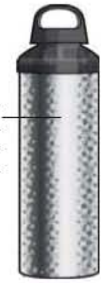
Metal container can be used to boil water

METAL WATER BOTTLE

Slightly heavier, but stronger, than plastic water bottles.

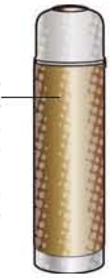
A metal exterior and a glass interior makes a thermos a heavy alternative

THERMOS FLASK

Although very heavy, thermos flasks allow you to store either hot or cold water.

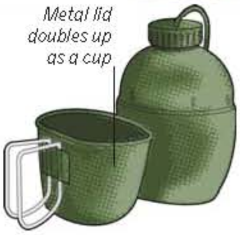
Metal lid doubles up as a cup

MILITARY WATER BOTTLE

A standard issue bottle used by most of the world's military forces.

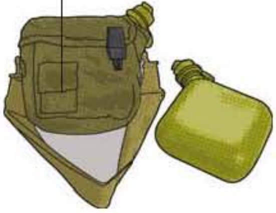
Store purification tablets in pouch

COLLAPSIBLE WATER CANTEEN

Heavy-duty plastic water canteen that can be hung round the neck.

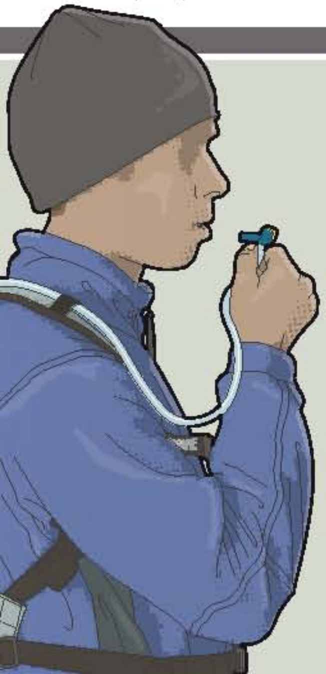

HYDRATION SYSTEMS

Hydration systems usually consist of a water-storage container (called a "reservoir"), an on-demand, one-way drinking tube, and a harness. However, a common problem with these systems is that, because you can't see the water, you can suddenly discover you have used all of it without realizing.

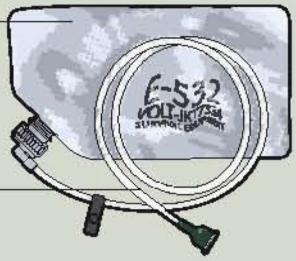
Water container is filled and then carried in a backpack
Water is accessed via a one-way drinking tube
E-532

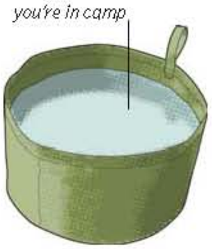
Use to store water once you're in camp

COLLAPSIBLE WATER BOWL

Easy to pack, light, and can also be used to seperate items in your backpack.

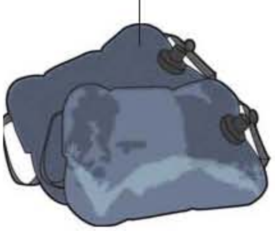
Stow water bags flat in your pack when not in use

WATER BAGS

Useful for carrying water from its source back to your camp. Can also be used to keep your equipment dry.

IMPROVISING WATER CONTAINERS

In an emergency, there's a strong chance that you won't have the luxury of having water bottles or storage systems at your disposal. However, with luck—and a little ingenuity—you should be able to find something that allows you to store and carry enough water to satisfy your needs.

SEA SURVIVAL SUIT

Because they are designed to keep water out, they can also be used to keep water in. Most survival suits are packed with a layer of chalk to prevent the material from sticking together, so wash it out before use.

WATERPROOF CLOTHING

Many types of waterproof clothing can be adapted to hold water. Jacket sleeves and trouser legs can be knotted to form a basic bladder; Gore-Tex® socks will hold water; and some waterproof backpacks can be used for water storage.

DISCARDED MATERIALS

Always be on the lookout for anything that can hold water, from empty plastic bags and bottles, to large industrial-type containers. Clean and sterilize found items before use.

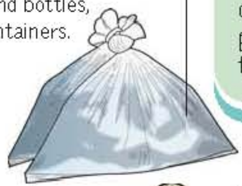

Heavy-duty plastic bags can be used to carry water

GOURDS

The shell of a hollowed-out and dried fruit—a gourd—can hold water. Gourds made from large fruits, such as squashes, pumpkins, and melons, can hold a considerable amount of water.

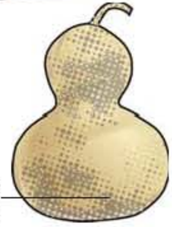

Look for a gourd with a shell thickness of $\frac{1}{4}$in (60 mm)

EMERGENCY WATER

Sachets of emergency water can be bought and stored for emergency use. They usually come in a packet that contains five 1$\frac{7}{10}$fl oz (50 ml) sachets of water, each being little more than a mouthful. These should always only be used as a last resort. Always try to procure water by other means before turning to your emergency water supply.

BAMBOO

A bamboo cane will provide a natural cup if you cut 1 in (2$\frac{1}{2}$cm) below one joint and then 1 in (2$\frac{1}{2}$cm) below the next. Take care to smooth the edges after cutting.

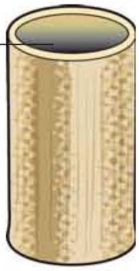

Bamboo can be cut into a ready-made cup

MAKING A MINI-RESERVOIR

Your ability to store sufficient water for your needs in a survival situation may be limited and it may be impractical—because of injury or distance—to travel back and forth for water every time you need it. If you have some basic materials, making a mini-reservoir to catch rainwater (see also p. 188) solves this problem.

1. To reduce the amount of effort required, choose a site that offers the least resistance (such as a natural hollow). Dig a shallow pit using whatever materials you have to hand, such as a stick.

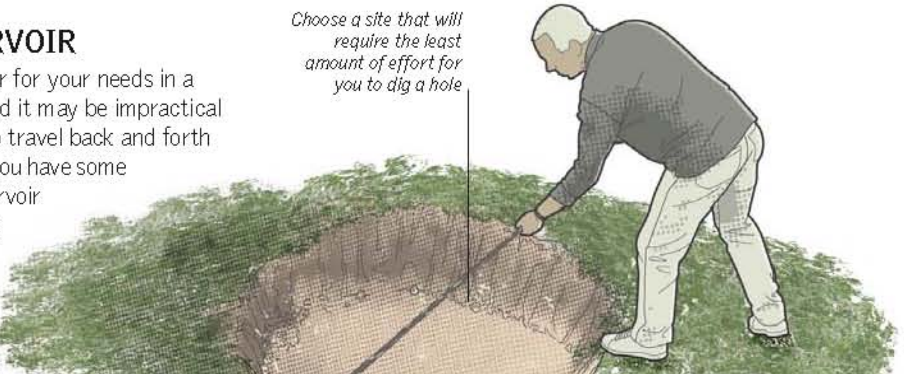

Choose a site that will require the least amount of effort for you to dig a hole

2. Smooth the edges of the pit and line it with waterproof material, such as a survival blanket. Weight the edges with rocks, soil, or logs, making sure that dirt can't run into the pond when it rains.

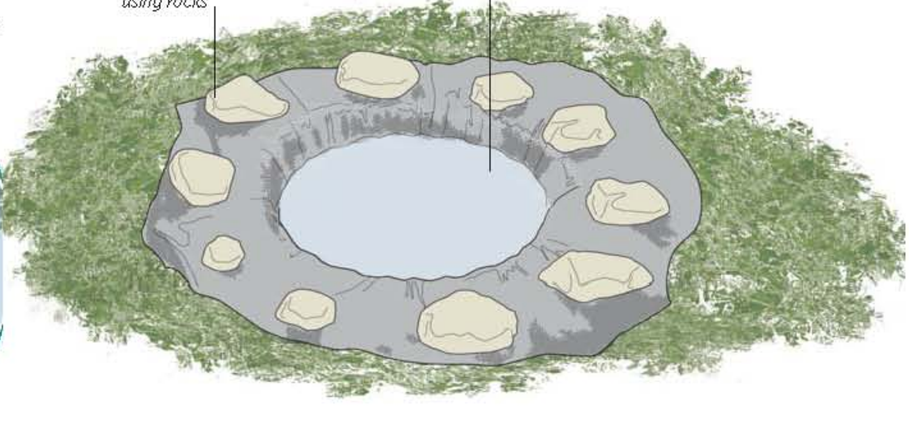

Secure the survival blanket in place using rocks

Leave the pond uncovered when it rains, but cover it at other times to reduce evaporation

WATER HYGIENE

Water stored in a mini-reservoir for any period of time will need to be filtered and purified before it's safe to drink (see pp. 200–01).

TREATING WATER

WITH THE EXCEPTION OF rainwater, all other water procured in a survival situation should be treated before it is safe to drink to remove or destroy harmful pathogens and microorganisms that could lead to gastrointestinal illness.

FILTERING WATER

If you don't have a device that filters and disinfects the water, you will have to accomplish the same task in two stages. Before purifying the water, you will have to filter it to remove any debris. You could use the popular Millbank bag (see opposite), but it may be necessary for you to construct your own improvised filter.

MILLBANK BAG

Used extensively by military forces around the world, the Millbank bag is an effective water filter that packs down small and can be used many times to produce large quantities of filtered water. Set up the bag as soon as you arrive in camp, because the process is quite slow: the Millbank bag will filter 1 liter (2.1 pints) of water in five minutes. Note that water filtered in this way will still need to be purified.

BUILDING A TRIPOD FILTER

If you're not carrying a filtration system with you such as a Millbank bag, you could make an improvised filter. All you need are three sticks to form a tripod, and some materials to create three separate layers.

1. Use three sticks or a bent sapling to form a tripod. Using any materials you have, form layers, starting at the top with the coarsest material and working your way down using finer materials—such as parachute silk or nylon—as you go.

2. Pour water into the top layer. It will become filtered as it passes through the increasingly fine materials.

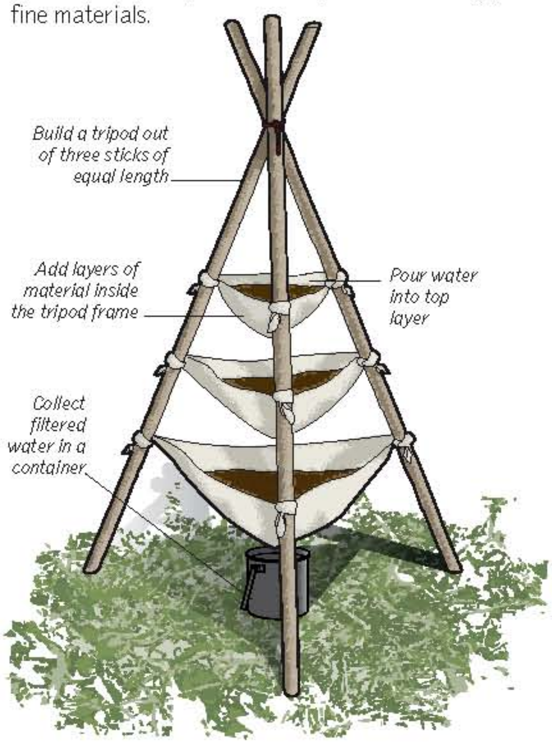
Build a tripod out of three sticks of equal length
Add layers of material inside the tripod frame
Pour water into top layer
Collect filtered water in a container

MAKING A BOTTLE FILTER

To make an improvised bottle filter, take a container, such as a plastic bottle, and cut off the bottom (or make a hole). Note that a sock used in the same manner is still effective.

1. Hang the bottle upside down from a branch. Fill the bottle with layers of different materials working from coarse to fine as you work your way down the container.

2. Pour the water into the top end of the bottle and allow it to work its way down through the layers.

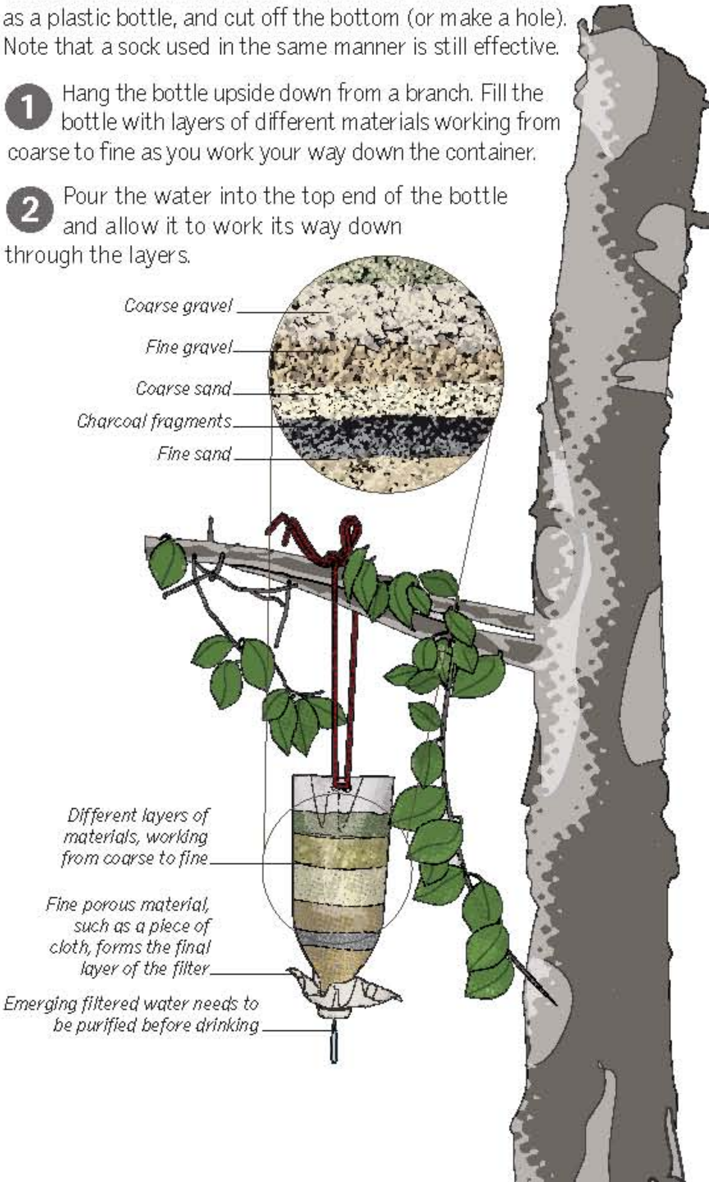
Coarse gravel
Fine gravel
Coarse sand
Charcoal fragments
Fine sand
Different layers of materials, working from coarse to fine
Fine porous material, such as a piece of cloth, forms the final layer of the filter
Emerging filtered water needs to be purified before drinking

DISINFECTING WATER

If you drink untreated water, you run the risk of becoming infected with a water-borne disease (see p. 187), so it's vital that you treat any water first. If you have the ability to start a fire, the most effective way of making water safe to drink is by boiling it. However, if you are unable to make a fire, there are several devices available that are capable of filtering and purifying water to make it safe for drinking.

MINI PORTABLE WATER PURIFIERS

These are specially designed units that filter the water and then purify it, by pumping the contaminated water through either microfilters, chemicals, or a combination of both. Sizes vary from small emergency pumps capable of purifying up to 13 gallons (50 liters) of water, to larger units that can filter huge quantities of water.

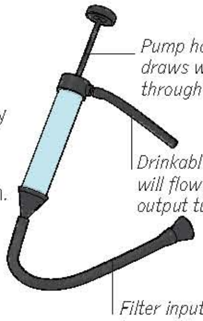

Pump handle draws water through the pump

Drinkable water will flow from the output tube

Filter input tube is lowered into water that requires treatment

GRAVITY/PRESSURE FILTERS

These devices are incorporated within drinking bottles. The water is either allowed to flow naturally through the system via gravity, or is squeezed through by the operator. All bottles of this type usually employ: a filter to remove sediment and organic contaminants; a micron filter to remove protozoa; and a chemical that kills water-borne bacteria and viruses.

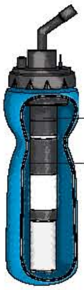

The filter system is housed inside a water bottle

Water passes through a filter which removes debris and water-borne diseases

SURVIVAL STRAWS

Compact emergency water purifiers, survival straws contain a filter system and employ either carbon or iodine resin systems to eliminate water-borne diseases and harmful chemicals. You need to get the water to a point where you can reach it with your straw. If you want to draw a supply of water to carry with you, you'll have to draw the water into your mouth and decant it into a container—a laborious process.

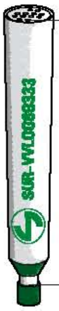

The filter end of the straw can be placed into any accessible non-saline water source

Water is sucked through the straw

OTHER METHODS

If you can't boil water, or if you don't have a water-purification device, you'll have to rely on non-mechanical techniques. The concentration and contact time required for some of these methods is dictated by the quality and temperature of the water being treated.

"BIG BUBBLES, NO TROUBLES"

Microorganisms and virtually all intestinal pathogens are killed at temperatures well below boiling point. The process of bringing water to the boil is sufficient to disinfect it—continuing to boil it just wastes fuel, time, and water.

METHOD	DESCRIPTION
Iodine (Liquid and Tablets)	Iodine—which destroys bacteria, viruses, and cysts—can be used to disinfect water effectively and conveniently. Its action is dependent on its concentration, the water temperature, and duration of contact—a concentration of 8 mgs per liter at 68°F (20°C) will destroy all pathogens if left for ten minutes.
Chlorine tablets	Chlorine-based tablets will destroy most bacteria, but are less effective for viruses and cysts. They are more effective when used in combination with phosphoric acid and will destroy both Giardiasis and Cryptosporidium.
Potassium Permanganate	Potassium permanganate can be bought at most pharmacies. Mix a few granules with your water until it turns light pink. Leave for at least 30 minutes before drinking.
Bleach	Adding unscented household bleach is the cheapest way of adding chlorine to water (it contains 5 percent sodium hypochlorite). Be careful to add just one drop of bleach per liter of water—two if the water is cloudy—and leave for at least 30 minutes before drinking. This method is not always effective against Giardiasis and Cryptosporidium.
Ultraviolet (UV) Light	When many harmful microorganisms are exposed to UV light, the process of light absorption disrupts the cell's DNA, rendering the organism harmless. The quality of the water will affect the amount of exposure to UV light required: the cloudier the water the harder it will be for UV to penetrate.
UV Passive	Fill plastic bottles with water, replace the lids, and place them in direct sunlight, preferably on a dark surface. The sun's rays will kill the bacteria that cause common water-borne diseases.
UV Active (Steripen)	A small UV purifier that is placed in pre-filtered water and activated for a short time. Some models can purify up to 1 liter (2.1 pints) of water in as little as 48 seconds.

FIND AND PREPARE FOOD

Your body converts food into fuel, which provides you with heat and energy, and helps you to recover from hard work, injury, or sickness. If you are healthy, your body can survive for weeks without food by using the reserves stored in its tissues—although you will use approximately 70 calories per hour just breathing and up to 5,500 calories a day if laboring hard.

In a short-term survival situation, food should not be your major priority. You would probably have eaten recently and, if you'd prepared properly, you should have some basic emergency food in your pack. While you may go through food withdrawal symptoms—when your stomach complains because the food it's expecting doesn't arrive—you're not going to die of starvation within a few days. However, the body will react to its fuel not being replaced: hunger, a lack of energy, and a deterioration in coordination can be expected after a few days. If the opportunity to procure food arises, it should always be taken. Eat little and often but always make sure you have sufficient water to digest it.

In this section YOU WILL DISCOVER...

- how to **cook** with **hot rocks**...
- how to **make** a **fishing reel** from a **beverage can**...
- why a **snare** must have a **perfect end**...
- how to **lasso a lizard** and **snare a squirrel**...
- that there's **more than one way** to **skin a rabbit**...
- which **grubs taste** like **scrambled eggs**...
- how to **catch** a **bird** in a **bush**...

In a long-term survival situation, your survival priorities will change and the need for food in order to simply survive will become more important. There is a thin line between food not being your priority and then subsequently finding that you're in no physical condition to do anything about it when it does become a priority—you should regularly reassess your situation and alter your plans accordingly. It takes effort, skill, and a certain amount of luck to obtain food in the wild, especially if you're not in your natural environment.

When gathering food in the wild, always ensure that the energy gained from the food is more than the energy you expended in procuring it, otherwise it's a wasteful exercise.

FOOD THAT'S EASY TO FIND and gather should always be your first priority:

- **Plants** are easy to collect and, as long as they're readily available in the environment you're in, should be your first choice for food. However, make sure that you are absolutely sure that they are edible—mistakenly eating the wrong leaf or berry could cause vomiting and diarrhea, making your situation worse.
- **Fishing** requires little effort once the lines or traps have been set, and they will work for you around the clock. Fish is high in protein and relatively simple to prepare and cook.
- **Insects, reptiles, and amphibians** may also be available, but be careful that you don't expend more energy in catching them than you gain in nutrients from eating them. Remember also that many insects, snakes, and amphibians are poisonous, but they can still be used as bait for fish and mammals.
- **Birds and mammals** have their own survival mechanisms and are wary of humans, especially in remote areas where contact with humans is limited. Even if caught, the bird or mammal will need to be killed, plucked or skinned, and cooked.

> “Always ensure that the **energy gained** from the **food** is **more than** the **energy you expended** in **procuring it**, otherwise it's a **wasteful exercise**”

WILD COOKING

IN A SURVIVAL SITUATION, you must cook every item of food that you're not sure about in order to kill any parasites or harmful bacteria. While cooking reduces the nutritional value of food, it does help to improve the taste of many wild foods and to make them more digestible than if eaten raw.

FOUND MEAT
If you find a dead animal, you can eat the meat if you cut it into small pieces and boil it for 30 minutes. Don't touch it if you have cuts on your hands—you don't know how it died.

COOKING OVER AN OPEN FIRE

Use the fierce flames of a fire to boil water then, when the flames have died down, use the steady heat of the embers to cook on (see p. 121). The most basic method of cooking is roasting, using a spit made from green wood to suspend the food over a fire. Make sure the meat is thoroughly cooked.

BOILING OR STEWING

When you boil food, fat and natural juices are retained in the water. It is important to drink the water to obtain the maximum nutrition from the meal, unless you have been boiling toxic substances out of the food.

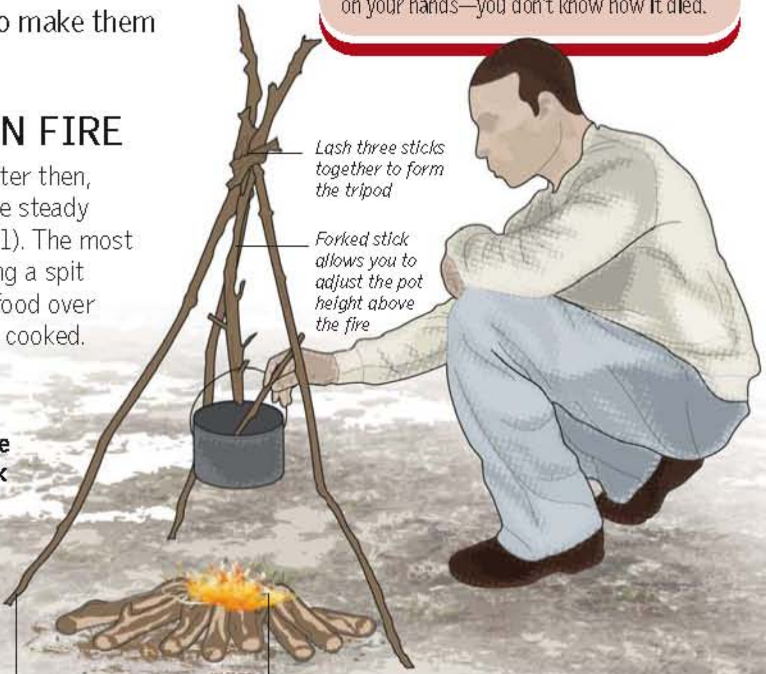
Lash three sticks together to form the tripod
Forked stick allows you to adjust the pot height above the fire

A tripod is a very stable and therefore safe way to suspend a cooking pot over a fire

Wait until flames die down a little before you begin cooking

DAMPER BREAD

This simple method for making yeast-free bread was developed by stockmen working out in the Australian bush.

1. Mix flour and water (and a pinch of salt, if you have it) into a pliable dough, then roll it into a long, thick sausage shape.

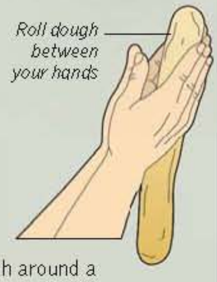
Roll dough between your hands

2. Wind the dough around a stick, then hold it over the embers of a fire, turning regularly until it browns. It will slide easily off the stick when cooked.

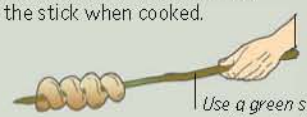
Use a green stick, which is less likely to burn

STEAMING FOOD

Steaming leaches fewer nutrients away than boiling and is a particularly good way of cooking fish and green vegetables—fresh leaves will be ready to eat in just a few minutes. To steam food, it must be suspended above boiling water in some way.

BAMBOO STEAMER

Bamboo stems are tough, hollow, and divided into sections. Use a sharp stick to make small holes in the walls dividing a three-section piece of cane, leaving the wall at the base intact. Pour water into the stem until it's just below the bottom section ring. Add food to the top section and cover with grass and a loose-fitting bamboo lid. Lean the stem over a fire, propping it against a forked stick.

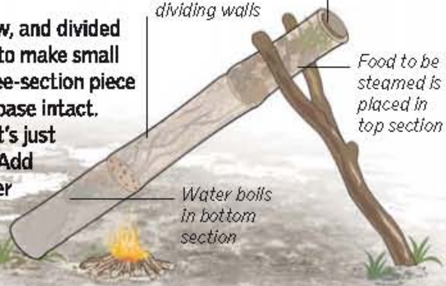
Steam rises up stem through holes in dividing walls
Place grass on top of food to retain heat and cover with a lid
Food to be steamed is placed in top section
Water boils in bottom section

CLAY BAKING

Baking food in clay requires no cooking utensils. Animals must be cleaned and gutted first, but can otherwise just be covered in soft clay. When the meat is cooked, skin, spines, or feathers will remain embedded in the clay. Cooking root vegetables or fish in this way removes their skins, losing valuable nutrients, so it's best to wrap them in leaves first.

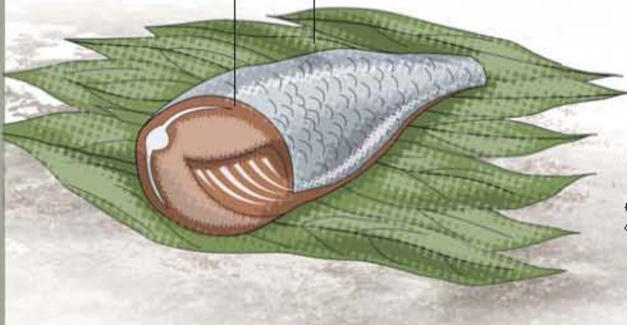
Protect nutritious skin of fish by wrapping in fresh leaves
Select long, wide green leaves

① Wrap the food in a bundle of fresh green leaves and tie them in place with some long strands of grass to make a secure parcel. Use only leaves from plants that you have identified as non-toxic.

② Cover the food parcel with an even layer of clay, making sure it's well sealed.

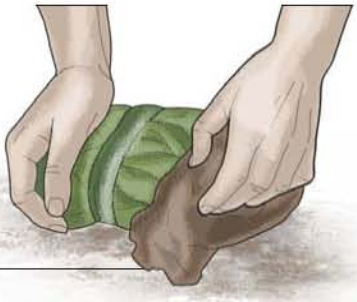
Layer of clay is 1 in (2.5 cm) thick

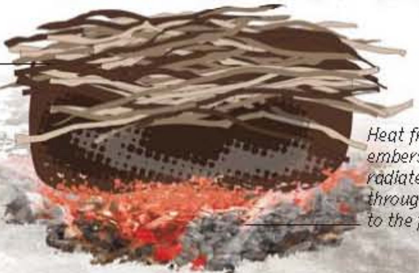
Build a fire on top of clay to increase the heat and decrease the cooking time
Heat from embers radiates through clay to the food

③ Place the parcel of clay in a bed of hot embers and build a fire up on top of it. Cooking time will be between 30 and 60 minutes, depending on the size of the food item. Break open the clay and remove the food.

COOKING WITH HOT ROCKS

Rocks take a while to heat up but they stay hot for a long time, allowing food to bake steadily on them. To reduce cooking time, cover the food with some birch bark or a flat piece of wood. When cooking on rocks, don't use slate or other layered rocks as they are likely to shatter when heated. Another method of cooking with hot rocks is to place them in a pit. The food is covered with leaves and placed on the rocks, then the pit is filled in to retain the heat. The food is left buried until it is ready to eat.

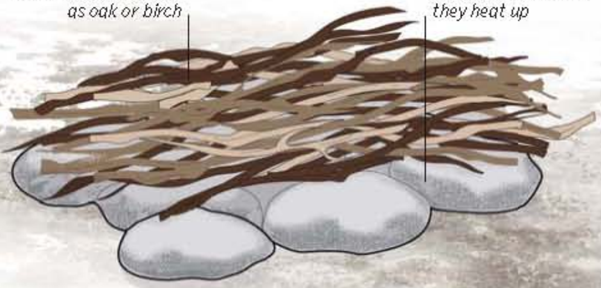
Use sticks from hard-wood trees such as oak or birch
Don't use wet rocks – they may explode as they heat up

① Place some large, reasonably flat rocks close together. Light a fire on top of the rocks using some tinder and dry sticks, then leave the fire to burn down to hot ashes while you prepare your food.

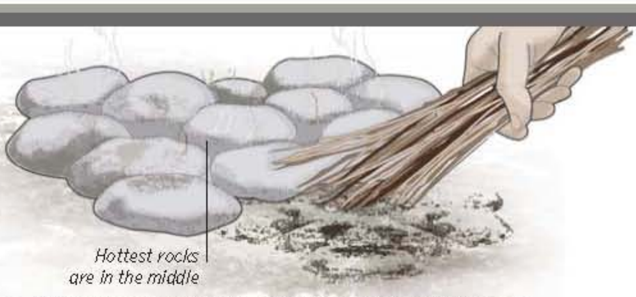
Hottest rocks are in the middle

② Brush the embers and ash off the rocks, taking care not to touch the rocks with your hands as they will now be extremely hot.

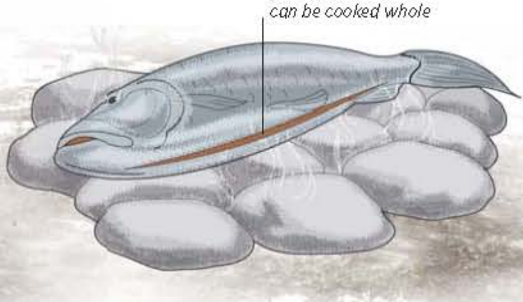
Fish should be gutted but otherwise can be cooked whole

③ Place food on top of the rocks to bake. The hottest rocks will be in the center, so items that must be cooked more slowly should be put near the edges. Keep adding more food to cook until the rocks have cooled.

EDIBLE PLANTS

IN A SHORT-TERM survival situation, food is not a priority—identifying edible plants uses energy, requires knowledge and skill, and the risks of getting it wrong far outweigh the benefits. In a long-term survival situation, however, your priorities would change, so being able to identify edible plants is advantageous.

GATHERING PLANTS

When foraging for plants, take a bag or can with you and be careful not to crush what you collect. Gather only a few species to lessen the chances of mixing in something inedible or poisonous. Don't assume that because birds or mammals have eaten a plant it's safe for you to eat too. Unless you are absolutely certain, you will need to carry out the following test to find out whether a plant is edible.

WARNING!

The Universal Edibility Test (see below) does not apply to mushrooms. For this reason alone, you should never eat any mushrooms that you have gathered (especially in a survival situation) unless you are 100 percent positive that you can identify each specific mushroom as being edible. In 2008, a British woman died after eating a poisonous mushroom that had been foraged by mistake with edible mushrooms.

DIGGING FOR ROOTS

Roots and tubers are a good source of carbohydrate and their skins contain vitamins. Take care when gathering in spring as some plants will only have small shoots and will be hard to identify.

1. Cut a stick from a hard-wood tree, such as oak, sharpen one end to a chisel-shape then harden it in a fire.

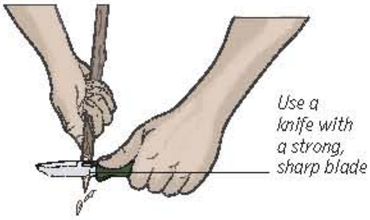
Use a knife with a strong, sharp blade

2. Choose a larger plant, then dig down at its side, loosening the earth around the root until it can be pulled out in one piece.

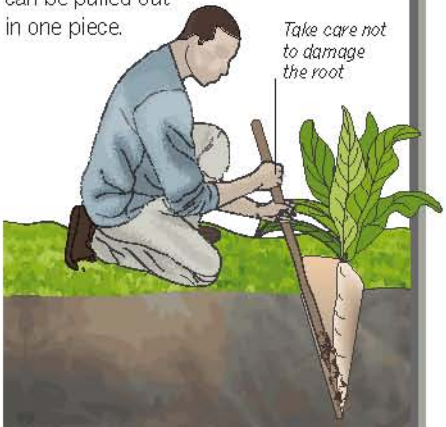
Take care not to damage the root

UNIVERSAL EDIBILITY TEST

This test enables you to check whether a plant is safe to eat. If you're in a group, only one of you should test the plant. Make sure that you have plenty of drinking water and firewood with you (see box, right), and eat nothing for eight hours beforehand. Test only one type of plant at a time, use only one part of that plant—roots, leaves, stalk, buds, fruit—at a time, and test it in the same state in which it will be eaten: either raw or cooked. Make sure that the plant you are testing is in plentiful supply (there's no point testing it otherwise) and avoid plants with milky or soapy sap, or bright colors—they are usually Nature's warning signal.

1. First, inspect the plant—it should look fresh and in good condition. Avoid anything slimy. Next, sniff the plant. Discard it if it smells bad. If it smells of peaches or bitter almonds, it may contain cyanide.

2. Take a small portion of the plant part you're testing and gently rub it on the sensitive skin inside your elbow or wrist. Wait for 15 minutes. If no irritation, stinging sensation, rash, or swelling develops, proceed to Step 3.

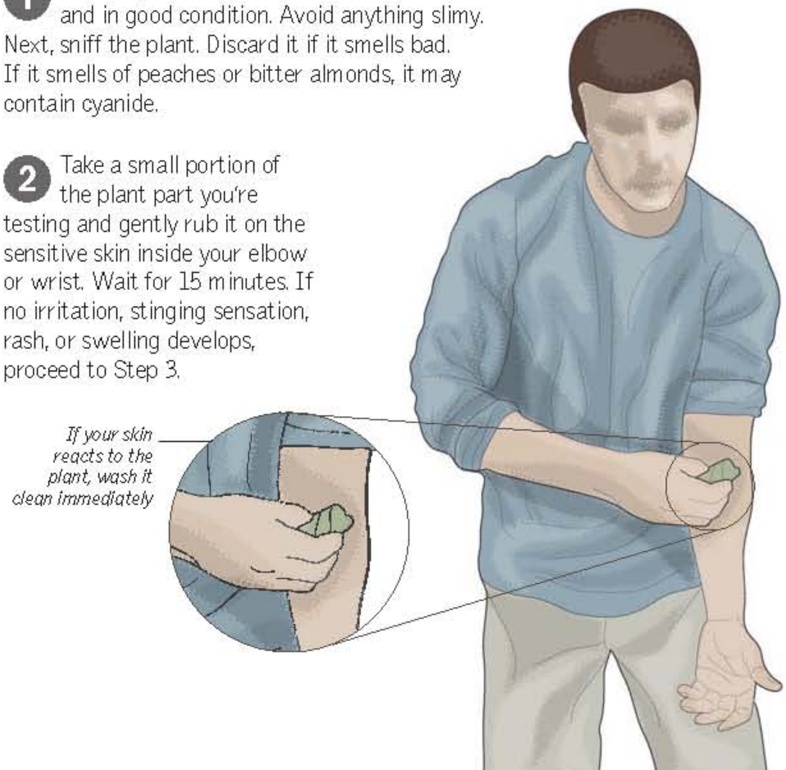
If your skin reacts to the plant, wash it clean immediately

NEEDLE TEA

The needles of evergreen trees such as pine and spruce are rich in vitamins A and C and can be used to make a refreshing—and potentially life-saving—tea.

1. For each cup of tea, collect two teaspoons of fresh green needles and bruise them with a stone.

2. Drop the needles into boiling water, then let them infuse for 10 minutes. Keep the pot warm and stir occasionally.

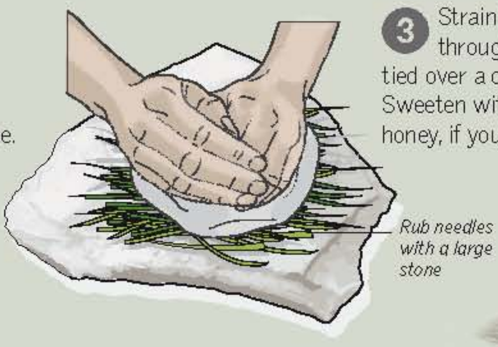
Rub needles with a large stone

3. Strain the tea through a cloth tied over a container. Sweeten with sugar or honey, if you have any.

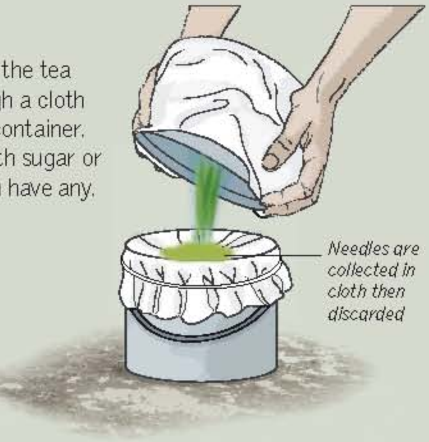
Needles are collected in cloth then discarded

3. Touch the corner of your mouth with the plant and wait 15 minutes. If no adverse reaction occurs, touch your lip and tongue with it. Wait another 15 minutes. If you suffer no ill effect, proceed to Step 4.

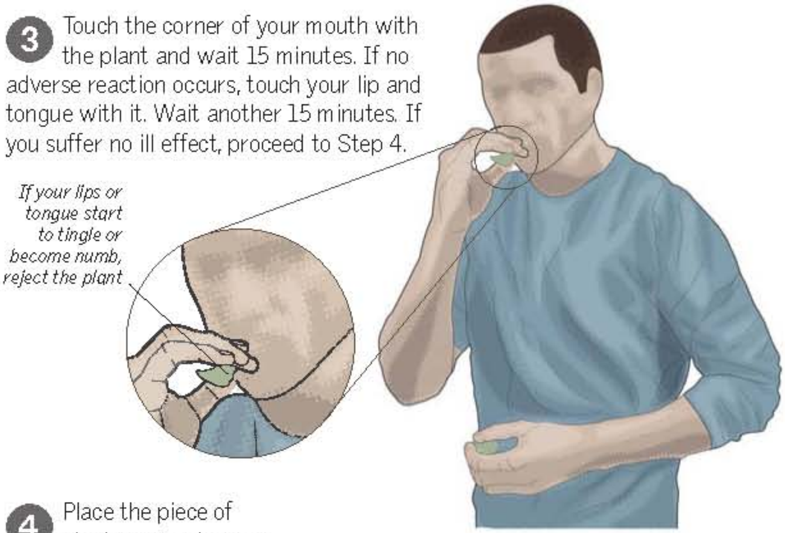
If your lips or tongue start to tingle or become numb, reject the plant

4. Place the piece of plant on your tongue, wait 15 minutes, then chew it and hold it in your mouth for 15 minutes more. If there are no ill effects, swallow the piece and wait eight hours. Eat nothing else but continue to drink water.

5. If no ill effect occurs, eat a quarter of a cup of the same part of the plant, prepared in the same way, then wait another eight hours. If there is no reaction, it is safe to eat. Eat little and often rather than gorging.

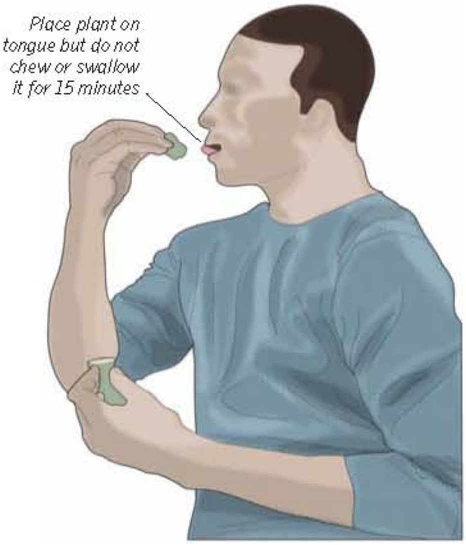
Place plant on tongue but do not chew or swallow it for 15 minutes

WHAT TO DO IF YOU HAVE AN ADVERSE REACTION

If at any stage you experience any type of adverse reaction to a plant during the edibility test, you should:

- Stop the test immediately.
- Avoid eating anything else until the symptoms cease.
- Wash the affected area thoroughly (if the reaction is external).
- Attempt to induce vomiting (if the reaction comes after eating the plant) by drinking salt water or by pushing a finger into the back of your throat.
- Drink lots of warm water (this is why you must ensure that you have a plentiful supply of purified water and firewood before you begin the test).
- Crush a teaspoonful of charcoal taken from a partially burned log and mix it into a paste with warm water, then swallow the mixture. It will either induce vomiting or, if you can keep it down, it may absorb the toxins.
- Try swallowing a paste made from white wood ash and water to reduce stomach pain.

If the tester is unconscious:

Do not induce vomiting if the tester is unconscious—get medical help and monitor the casualty.

CATCHING FISH

FISH CAN BE HOOKED, netted, trapped, speared, or even caught by hand if you're lucky. Observe their behavior—where and when they feed, and what they eat—to determine which method to try. It must be noted that some fishing techniques are illegal in some parts of the world and should only be used in a genuine survival situation.

BAIT AND LURES

Fish have a keen sense of smell to help them find their preferred food, so if one type of bait doesn't work, try another. Predatory fish are attracted to live bait by its movement—a worm wriggling on a hook, for example. You can make artificial bait out of shiny metal, cloth, or feathers. Try to imitate the fish's natural prey, such as an insect skimming across the water, to lure the fish into biting.

FISHING TACKLE

Fishing equipment can be made from all kinds of material. Your survival kit should have some fishing line and a few hooks, but if you don't have a kit you can improvise. You can use a stick as a makeshift rod, for example.

IMPROVISED HOOKS

Fishhooks can be fashioned from any piece of metal, such as a nail, pin, safety pin, needle, or some wire. You can also make them from natural materials, such as thorns, hard wood, coconut shells, bone, spines, or seashells. If you don't have any fishing line you can use parachute cord (see p. 136) or make cordage from plant fibers (see p. 138).

MAKING A FISHING REEL

You can make a reel with an empty beverage can. Tie one end of a line to the tab, then wrap it around the can until you have about 2 ft (60 cm) left. Attach a hook, float, and some sinkers. Hold the top of the can in one hand and the float in the other. Point the bottom of the can toward where you want the hook to land and throw the float. The rest of the line will unwind and follow. If a fish bites, give the line a jerk then wind it in around the can.

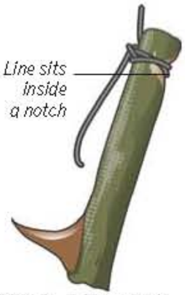
Line sits inside a notch

SINGLE THORN
Cut a 1 in (2.5 cm) length of bramble stem with a large, strong thorn. Tie a line to a notch at the other end.

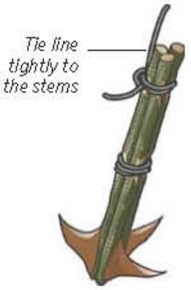
Tie line tightly to the stems

MANY THORNS
Tie three strong thorns together securely with some fishing line. Cut a notch in the stems and attach the line.

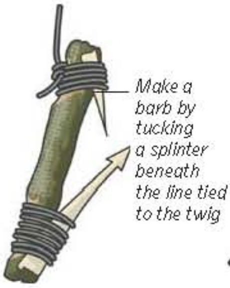
Make a barb by tucking a splinter beneath the line tied to the twig

WOOD OR BONE
Carve a splinter of bone into an arrow point or take a sliver of hard wood, like oak, and tie to a twig.

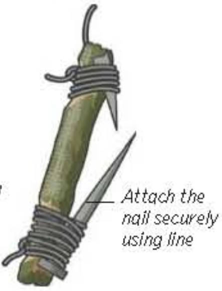
Attach the nail securely using line

NAIL
Cut a notch at one end of a small piece of hard wood. Place the head of a nail in the notch and tie it in place.

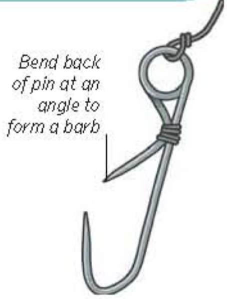
Bend back of pin at an angle to form a barb

SAFETY PIN
Remove the safety clasp. Bend the sharp point of the pin round to form a hook. Attach line to the wire loop.

REMOVING A FISHHOOK

You should never try to remove a fishhook that is deeply embedded in flesh. Cut the fishing line as close to the hook as possible, then put a pad around the exposed part of the hook and secure with a bandage. Seek medical help as soon as possible and monitor for signs of infection. However, in a survival situation, you can try to remove the hook by doing the following:

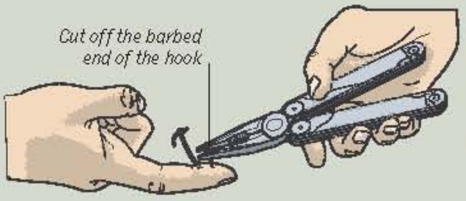
Cut off the barbed end of the hook

① If the barb is visible, cut it off with a pair of pliers. If you can't see the barb, firmly and quickly push the hook in further until the barb emerges.

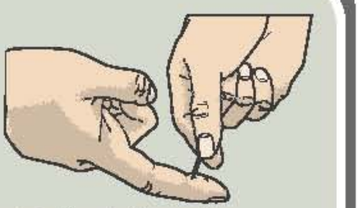

② Carefully ease the hook out by its eye. Clean the wound then cover it with a bandage.

TYPES OF BAIT

Live bait includes creatures such as:

- Worms and slugs
- Grasshoppers, crickets, and beetles
- Maggots and caterpillars
- Frogs (including large tadpoles)
- Small fish—to catch larger fish

Examples of inanimate bait include:

- Meat, guts, and reproductive organs recovered from animals
- Nuts and small fruits
- Bread, cheese, and pasta

FEATHER LURE

Make a hook, then attach it to a line. Tie a brightly colored feather just above the hook. Slowly move the lure across the water's surface to attract the fish.

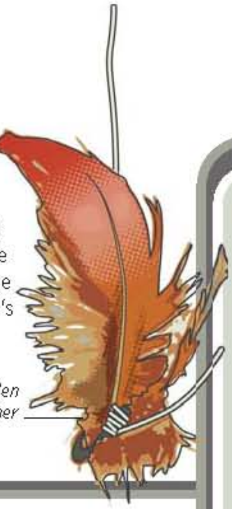

FLOATS AND SINKERS

A float keeps a baited hook at the best depth to attract the species of fish you're trying to catch. A sinker (weight) placed below the hook also helps to hold it in position. Your survival kit should include a few split-shot weights but, if it doesn't, you can tie small stones to the line instead. If the float bobs below the water's surface, you may have hooked a fish.

MAKING A FLOAT

You can use any natural material that floats, such as a piece of bark or a rose hip (see below). If you find a bird's flight feather, you can make a float by trimming the feather until you are left with just the hollow quill, then folding it in half and tying the two ends together.

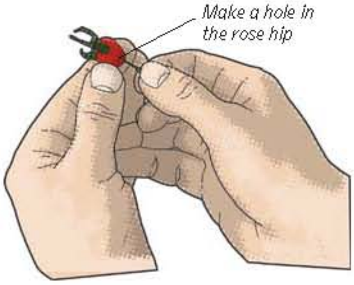
Make a hole in the rose hip

1 Using a sharp pin, a piece of wire, or a long thorn, pierce through the center of a rose hip.

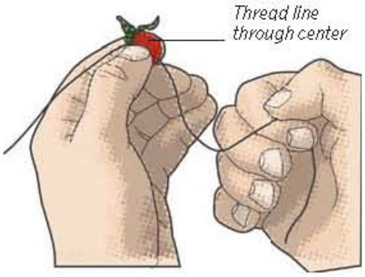

2 Thread fishing line through the hole. If you don't have line, use a length of fine cordage.

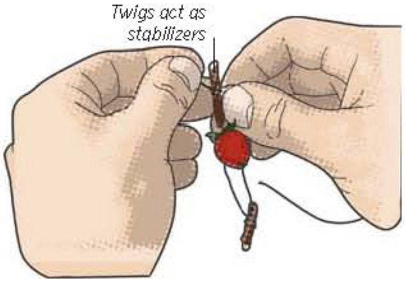

3 Tie small twigs above and below the rose hip. These will act as stabilizers and will also prevent the float from slipping up or down the line.

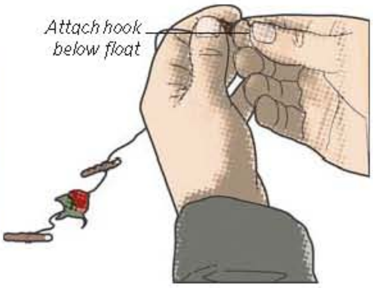

4 Attach the hook (in this example, a modified safety pin) to the line at the required depth below the float. If you have one, tie a sinker below the hook.

LINE FISHING

The more hooks you can get in the water, the greater your chances of catching a fish. All the methods shown below are passive—you simply set them and leave them to work for you.

SELF-STRIKING LINE

If a fish bites at the bait, the trigger will be jerked out of the catch and the bent rod will fly up. The line will be pulled tight, embedding the hook in the fish's mouth.

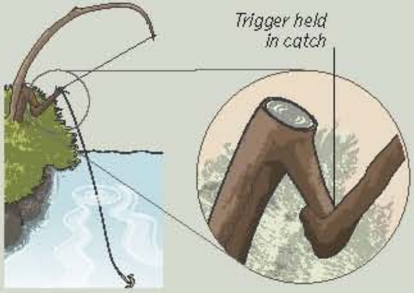

ICE FISHING

First check that the ice is at least 2 in (5 cm) thick and can bear your weight, then cut a hole about 1 ft (30 cm) in diameter over the deepest part of the lake.

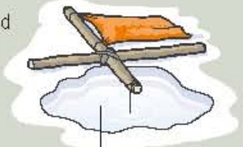
Ensure the ice doesn't freeze over

1 Tie a hooked and baited line to a stick with a flag at the other end and a central crosspiece.

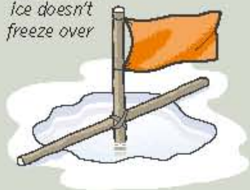

2 If a fish takes the bait, the crosspiece will be jerked over the hole and the flagpole pulled upright.

NIGHT LINES

To increase your chances of making a catch, you need to attract fish that live at all depths. Tie a rock to one end of a line and attach hooks with live bait at intervals along it. Tie the other end to a post at the edge of the bank and throw the line in the water, leaving it there overnight.

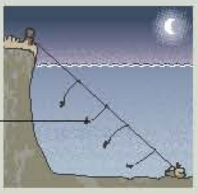
Keep hooklines short to avoid them getting tangled

NETS, TRAPS, AND HARPOONS

You can leave a hooked line in the water, but unless you regularly check it any fish you catch may be eaten by bigger fish. In a survival situation, nets and traps are more convenient methods of fishing because once set you can leave them and they work for you all the time.

WARNING!
Gill nets and traps are illegal fishing methods in many parts of the world, so use them only in an emergency.

MAKING A DIPPING NET

Fish that are too small to hook or harpoon may still be large enough to catch with a dipping net. Such fish are usually found at the edges of streams and lakes, and around rocks in pools. You can make a simple dipping net if you can split a branch or find a forked sapling and have a mosquito net—or a spare item of clothing, such as a sleeveless vest, T-shirt, or stockings.

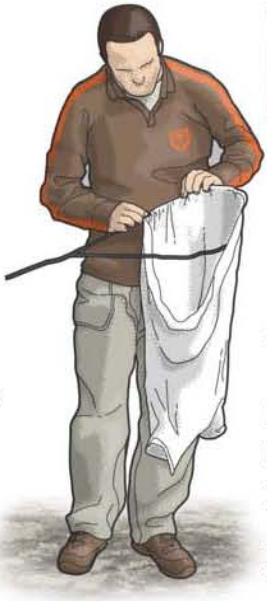

1 Make two small nicks in the hem of the mosquito net, vest, or stockings, then thread the forks of the sapling through them.

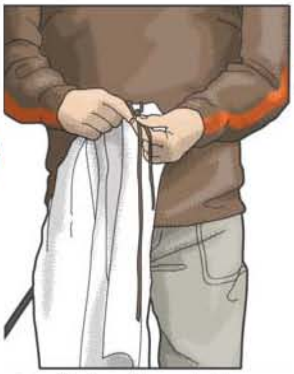

2 Cut another nick in the other side of the hem, where the forks meet, and pull them through. Tie the ends together with cordage, then push them back inside the hem.

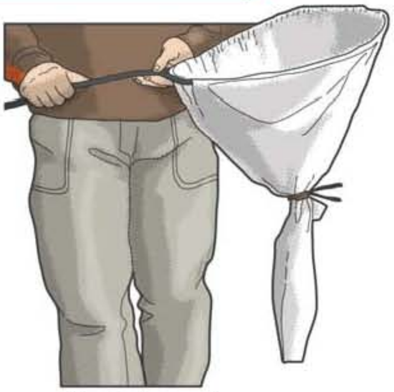

3 Tie off the net (or the vest above the armholes and neck), then either cut off the excess material or invert the net. This will prevent extra drag when you use the net.

MAKING A GILL NET

Once it has been set in a river, a gill net is a highly effective way of catching fish, whether they are swimming up- or downstream. However, a gill net should only be set for short periods of time as fish of all sizes can be entangled or injured. Make the lines out of paracord or natural cordage (see pp. 136–39).

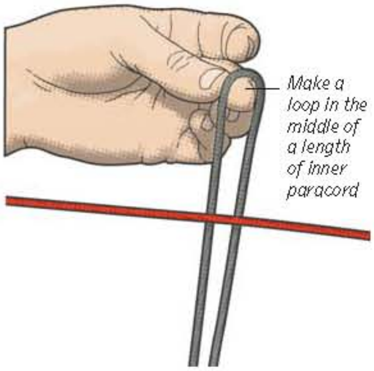

Make a loop in the middle of a length of inner paracord

1 Take a section of the outer sheath of some paracord and suspend it between two trees. Loop the inner paracord behind this line.

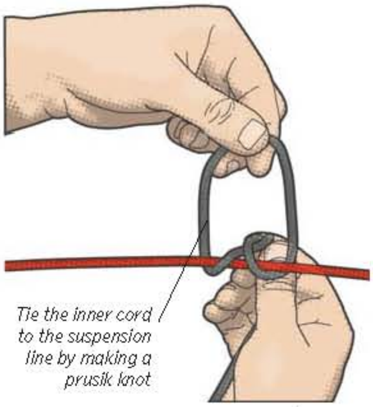

Tie the inner cord to the suspension line by making a prusik knot

2 Holding the two ends of the inner cord in your other hand, pass them around the suspended line and through the loop.

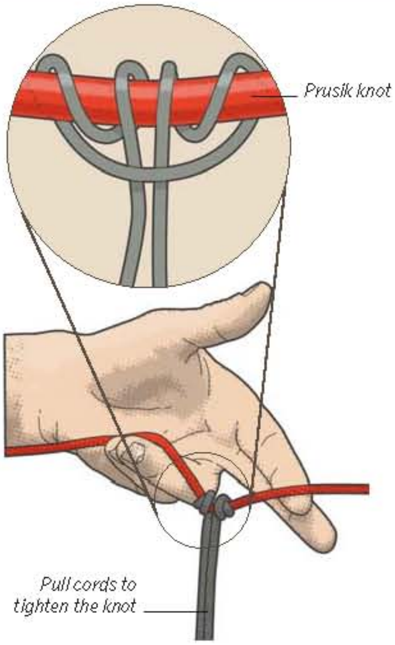

Pull cords to tighten the knot

3 Pull the knot tight. Repeat the process every 1½ in (4 cm) along the suspended line for the required width of the net.

MAKING A HARPOON

Using a harpoon to spear fish takes time, patience, and a certain level of skill. Keep the point of your harpoon in the water to avoid splashing and scaring off the fish. Strike quickly when you see a potential catch, aiming just ahead of it. A spiked harpoon is the simplest to make.

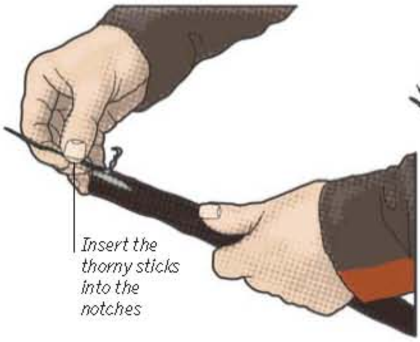
Insert the thorny sticks into the notches

1 Gather a few short, thorny sticks to form the barbs of the harpoon. Cut a long, thick, straight branch and make notches around one end.

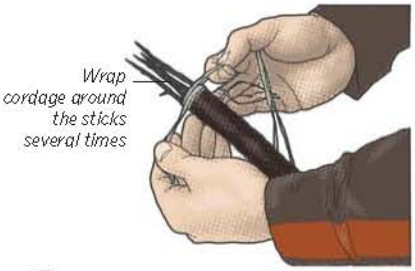
Wrap cordage around the sticks several times

2 Bind the thorny sticks tightly to the shaft with cordage. They will need to be able to support some of the weight of the fish so wrap them securely.

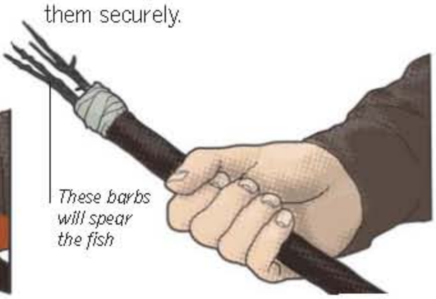
These barbs will spear the fish

3 When using the harpoon, take care not to damage the barbs by driving them against rocks or onto the riverbed. Use a dipping net, if you have one, to land the fish.

4 Make the mesh by tying a cord from one pair of cords to a cord from an adjacent pair, using a simple overhand knot (see p. 143).

- Continue alternating along and down the line. Tie stones to the ends of the net to hold it down once it is in the water.

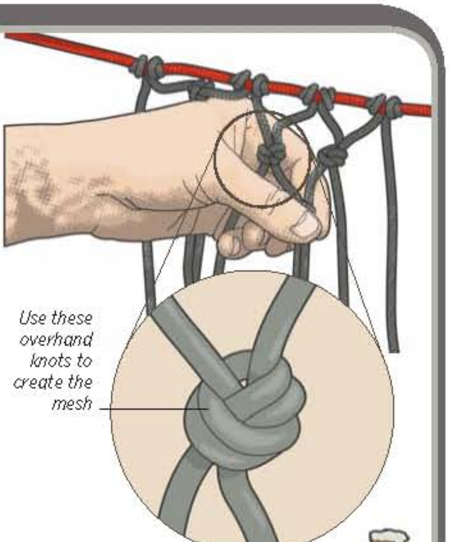
Use these overhand knots to create the mesh

5 To set the gill net, hang the suspension line between trees or poles on either side of the river, submerging the net in the water to a depth of about 6 in (15 cm).

- The net can also be used to catch birds (see p. 226).

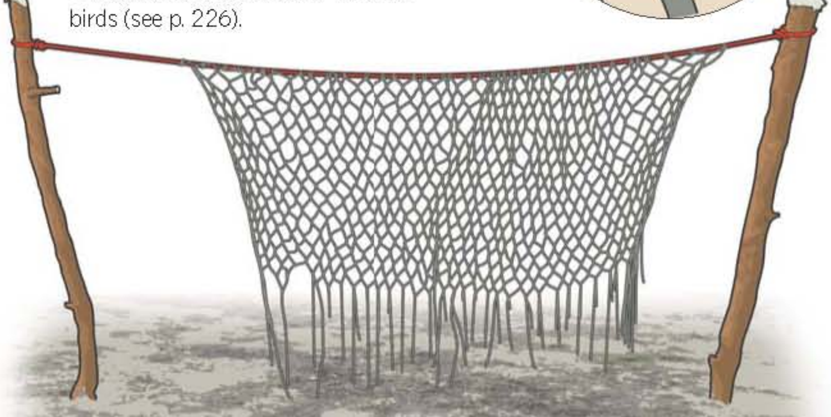

TRAPS

Fish traps can be made from a variety of materials. You can use rocks to build a wall or you can make a basket from sticks and cordage. You can even use man-made materials, such as bottles. If you only succeed in catching very small fish you can still use them as bait.

BOTTLE TRAP

Also known as a "minnow trap," this method uses a large plastic bottle to catch small fish. Cut the bottle in two just below the shoulder. Invert the neck and insert it into the cut end, then tie the two pieces together. Make holes in the plastic with a hot pin so the trap will sink. Bait the trap, then place it in a stream. Check it regularly to remove any fish caught and replace the bait.

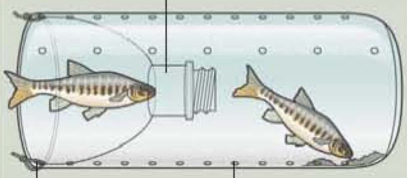
Fish smells bait and swims into bottle but can't find its way back out

Bottle neck tied to bottle base

Holes in bottle let water in

SINGLE-WALL TRAP

If your camp is near a tidal creek, you can build a curved wall of large rocks out from the bank. Pick the location at high tide and build a low wall at low tide. Fish may be trapped in the pool created between the wall and the bank as the tide recedes.

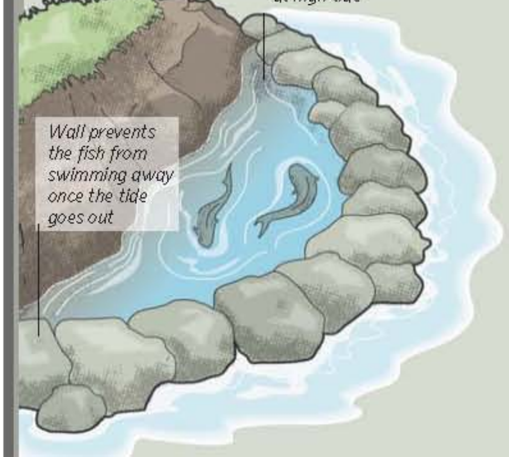
Water and fish enter the trap at high tide

Wall prevents the fish from swimming away once the tide goes out

PREPARING FISH

ALL FRESHWATER FISH are edible but they must be cooked before eating as they are often infested with parasites and harbor bacteria. Saltwater fish are safer to eat raw, if necessary, but will taste better cooked. Never eat a fish that has pale gills, sunken eyes, flabby skin or flesh, or that smells bad.

WARNING!
When handling fish be careful not to accidentally touch your eyes. The slime may contain bacteria that can cause a painful inflammation of the eye covering (conjunctivitis). Make sure you wash your hands clean of all slime. If you have any cuts on your hands, cover them before touching the fish.

FISH HYGIENE

As soon as you have landed a fish, it should be killed and gutted, then cooked and eaten as quickly as possible, particularly in a warm climate. It doesn't take long for fish to go bad because their slimy skin provides a breeding site for flies and bacteria. In a cold climate, you can delay filleting for up to 12 hours, which will make the job easier.

FILLETING A FISH

Filleting removes the parts of a fish that might quickly go bad, while leaving as much flesh as possible. With some species, you may find it easier to remove the bones after the fish has been cooked. Boil the bones and the head to make a nutritious stock. This must be done immediately and the stock should be kept in a cool place and drunk within a few hours.

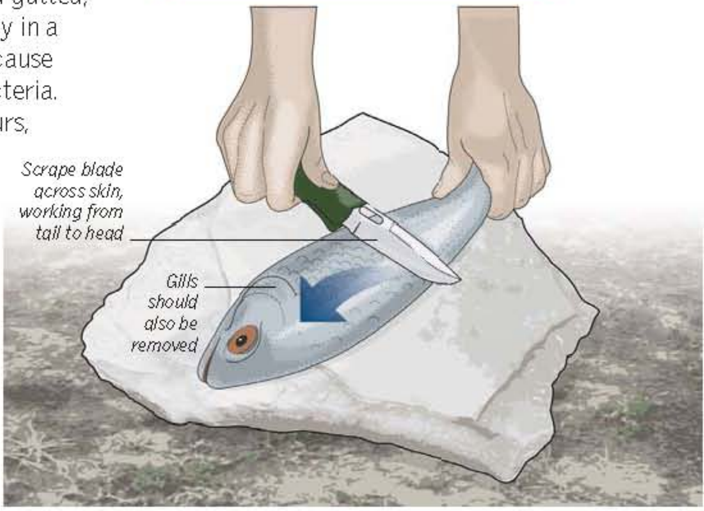
Scrape blade across skin, working from tail to head
Gills should also be removed

1 Once you have killed the fish, cut its throat to bleed it and cut out the gills. Wipe the slime off its skin to make it less slippery.

- Most fish don't need to be skinned—in fact, the skin is nutritious—but eels and catfish do (see panel, opposite).

2 Fish can be cooked with their scales on but, if you have time, it is best to remove them, particularly if they are large, as they can be a choking hazard.

- Hold the fish by the tail and scrape off the scales, holding the blade away from you and moving toward the head.

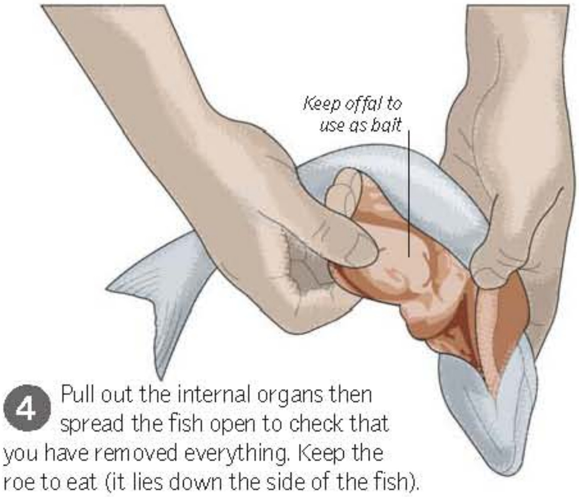
Keep offal to use as bait

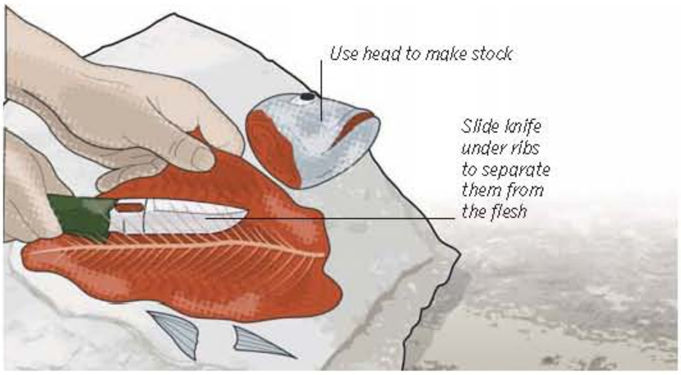
Use head to make stock
Slide knife under ribs to separate them from the flesh

4 Pull out the internal organs then spread the fish open to check that you have removed everything. Keep the roe to eat (it lies down the side of the fish).

- Wash the fish thoroughly, inside and out.

5 Cut off the head, tail, and fins. Open out the body and slide the knife under the ribs to separate them from the body, working toward the head. Fish oils can make the knife slippery, so you can use your thumbs instead.

- Repeat for the ribs on the other side of the backbone.

SMALL FISH

Any fish that is less than 6 in (15 cm) long doesn't need to be filleted. After gutting, they can be fried, grilled, or roasted whole. Keeping the heads and tails on helps to prevent the fish from falling to pieces.

SKEWERING FISH

Small fish, such as perch, can be grilled over an ember fire on a skewer to make a simple, yet tasty and nutritious meal. Impale the fishes on a green stick and hold them close to the embers. They won't take long to cook.

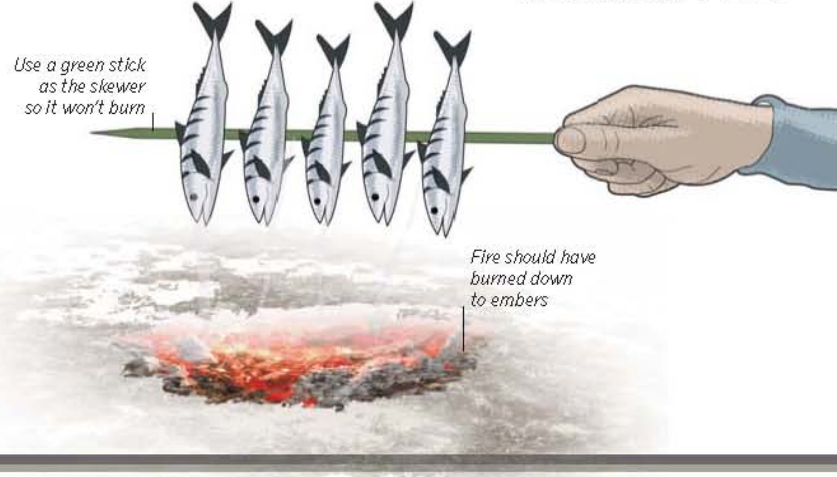
Use a green stick as the skewer so it won't burn
Fire should have burned down to embers

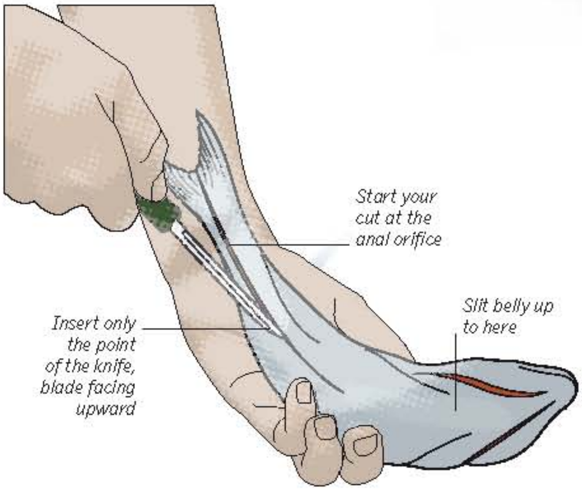
Start your cut at the anal orifice
Insert only the point of the knife, blade facing upward
Slit belly up to here

3 Holding the fish with its tail toward you, insert the point of a sharp knife, blade up, into the anal orifice and slit the fish open along its belly to its throat. This prevents you from puncturing the internal organs.

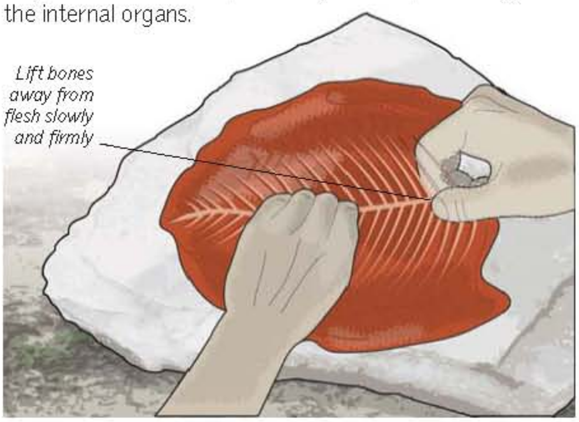
Lift bones away from flesh slowly and firmly

6 Separate the top and bottom of the backbone from the flesh with the tip of your knife. Carefully pull the backbone and ribs away from the flesh in one piece.

- If you have any tweezers or needle-nose pliers, use them to remove any small bones that remain.

SKINNING FISH

Eels and catfish are tasty but have to be skinned and gutted before cooking. To gut the fish, either use the filleting method shown opposite or Step 3 below. Catfish have a cartilage skeleton and can simply be cut crosswise into steaks.

1 Pass a stake through the gills of the fish and support on two strong uprights.

- Cut around the skin below the head with a sharp knife.
- Cut around the fins.

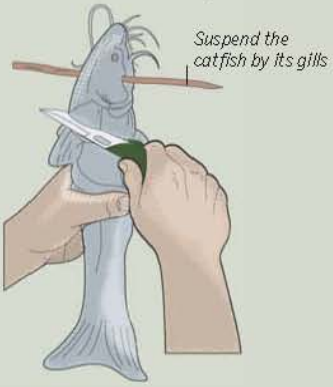
Suspend the catfish by its gills

2 Separate the top of the skin from the flesh then peel the skin downward.

- You will need to use both hands to get a good grip, and a firm, steady action. If the fish is very large, you may need to slit the skin.

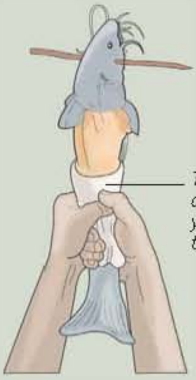
Tug skin down until you reach the tail

3 Remove the fish from the stake and break its backbone. When you pull the head off, the guts will come away with it.

- Remove tail and fins.

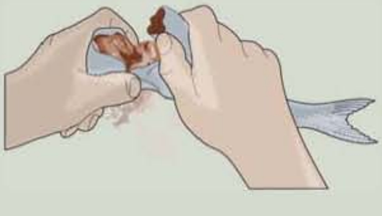

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—ADRIFT AT SEA

USEFUL EQUIPMENT

- Liferaft, lifejacket, survival suit
- Water catchment devices
- Solar still/reverse-osmosis pump
- Emergency Locator Beacon
- Marine VHF radio
- First-aid kit, sunscreen, sunglasses
- Flashlight and batteries
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

IN 1972, A FAMILY OF FIVE SURVIVED for 38 days adrift in the Pacific Ocean. The Robertson family were 18 months into a round-the-world sailing trip when their 43 ft (13 m) wooden schooner *Lucette* sank. They survived thanks to a combination of good seamanship, improvisation, and good fortune.

The family set off from Falmouth, England, on January 27, 1971—father Dougal at the helm, wife Lynn, and children Anne (who disembarked in the Bahamas), Douglas, and twins Neil and Sandy as crew—and safely navigated the Atlantic and Caribbean. On June 15, 1972, however, about 200 miles (320 km) west of the Galapagos, disaster struck—a pod of killer whales charged the boat, splintering the hull and holing her irreparably.

With *Lucette* sinking rapidly and no time to radio a distress call, the crew launched their inflatable liferaft and roped it to the boat's 10 ft (3 m) solid-hull dinghy, the *Ednamair*, which they used as a towboat after improvising a sail. Their supplies amounted to 2 gallons (10 liters) of water, a bag of onions, oranges and lemons, vitamin-fortified bread, glucose, four fishing hooks, a fishing line, a first-aid kit, a kitchen knife, and eight signal flares. They sailed north toward the Doldrums to find rain, which they collected with the use of a tarpaulin, and caught fish, eating some of the meat raw and drying the rest in the sun to be stored as rations.

"A POD OF KILLER WHALES CHARGED THE BOAT...HOLING HER IRREPARABLY"

After 16 days, the liferaft had deteriorated to such an extent that the family was forced to transfer to the dinghy, using the remnants of the raft as a canopy to provide shelter and aid in the collection of rainwater. They used the wind and currents to sail northeast toward Central America, and built up sufficient rations over the next three weeks to provide energy for the extra exertion of rowing toward the coast. Luckily for them, there was no need—their ordeal came to an end on July 23, when a Japanese fishing boat spotted a signal flare and picked them up.

WHAT TO DO

ARE YOU IN DANGER? +

NO

- If you are in a group, try to help any others who are in danger
- Prepare for the possibility of abandoning ship, and try to steer closer to either land or known shipping lanes
- Delegate responsibilities
- Ensure lifejackets and liferafts are ready for deployment

YES

Get yourself out of it:
Sinking vessel—You need water, location aids, and protection from drowning and the elements
Animals—Try not to splash in the water as this will attract sharks
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

YES

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO

You are faced with surviving for an indefinite period—until you are located or you find help

YES

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO

Abandon ship in a controlled manner and deploy all the liferafts. Always try to enter the liferaft dry. Use the rescue line to reach people in distress. Follow the "immediate actions" instructions printed inside the liferaft

YES

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Deploy the drogue—this will lessen drifting
- Inventory all food, water, and equipment and start rationing
- Prepare water procurement devices, such as solar stills and reverse osmosis pumps
- Protect yourself from the elements—sun, wind, and salt spray
- If you have no liferaft, then huddle together in a group with children in the center. If alone, adopt the H.E.L.P. position
- Improvise flotation aids from anything that can float or hold trapped air. Good examples are plastic bags, plastic bottles, and knotted clothing

DON'T

- Cut the painter to the vessel until you know it will sink, as the vessel is what people will be looking for
- Drink seawater under any circumstances—this will only increase your rate of dehydration
- Eat unless you have sufficient water to digest the food—fish is high in protein and requires plenty of water to digest

YOU SHOULD STAY **

DON'T

- Neglect to take anti-seasickness tablets. Vomiting will dehydrate you and the effects are very demotivating
- Forget to wear your survival equipment and ensure everyone knows how to operate it
- Try to second-guess the rescuers when they arrive. Do exactly as they say, they know what they are doing

DO

- Ensure that the rescue services are kept aware of your situation and updated with relevant information as the situation dictates
- Prepare to abandon ship: pack lifejackets and suitable clothing
- Ensure you know how to operate your survival equipment and that you have basic aids to location such as a whistle, flashlight, or plastic water bottle on your person
- Clear the deck of loose objects and be prepared to drop sail should rescue by helicopter be attempted
- Make an inventory and ration supplies

+ Never abandon a vessel unless you have to—searchers will be trying to locate it. Use the vessel's equipment while you can, and before you leave collect essential items to aid your survival.

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.

** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

TRAPPING ANIMALS

ALTHOUGH YOU SHOULD be prepared to take down sitting prey if you get the opportunity, trapping small animals is easier than hunting them—it requires less skill and energy and leaves you free to carry on with other tasks. One of the simplest traps is a snare.

MAKING A SNARE

Stainless steel snares of various strengths can be bought ready-made, with a running loop, or eye, at one end and a securing loop at the other. However, most survival kits contain a length of single-strand brass wire that can be used to make a snare.

1. Decide on the strength of snare you want (see panel, below) and double or quadruple the wire accordingly.

- Pass the strands around a stick, place it on the ground, then loop the loose ends around a second stick.
- Rotate the second stick until the strands have entwined, forming a single, thick wire. Remove the sticks.

STRENGTHS OF SNARE

The single-strand wire found in most survival kits isn't strong enough to hold most animals you would want to catch. You'll need to double or quadruple the strands by winding them together to increase their strength.

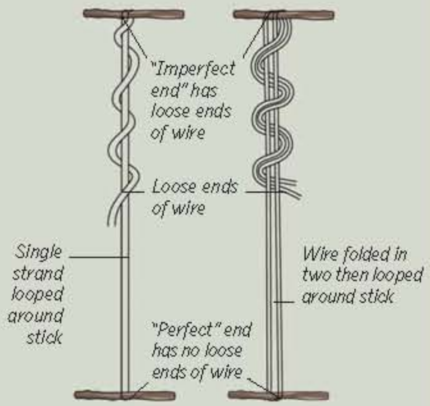

"Imperfect end" has loose ends of wire
Loose ends of wire
Single strand looped around stick
"Perfect" end has no loose ends of wire
Wire folded in two then looped around stick

TWO STRANDS
A two-strand snare will be sufficient to bear the weight of a squirrel-sized animal.

FOUR STRANDS
A four-strand snare will be strong enough to hold onto an animal the size of a rabbit.

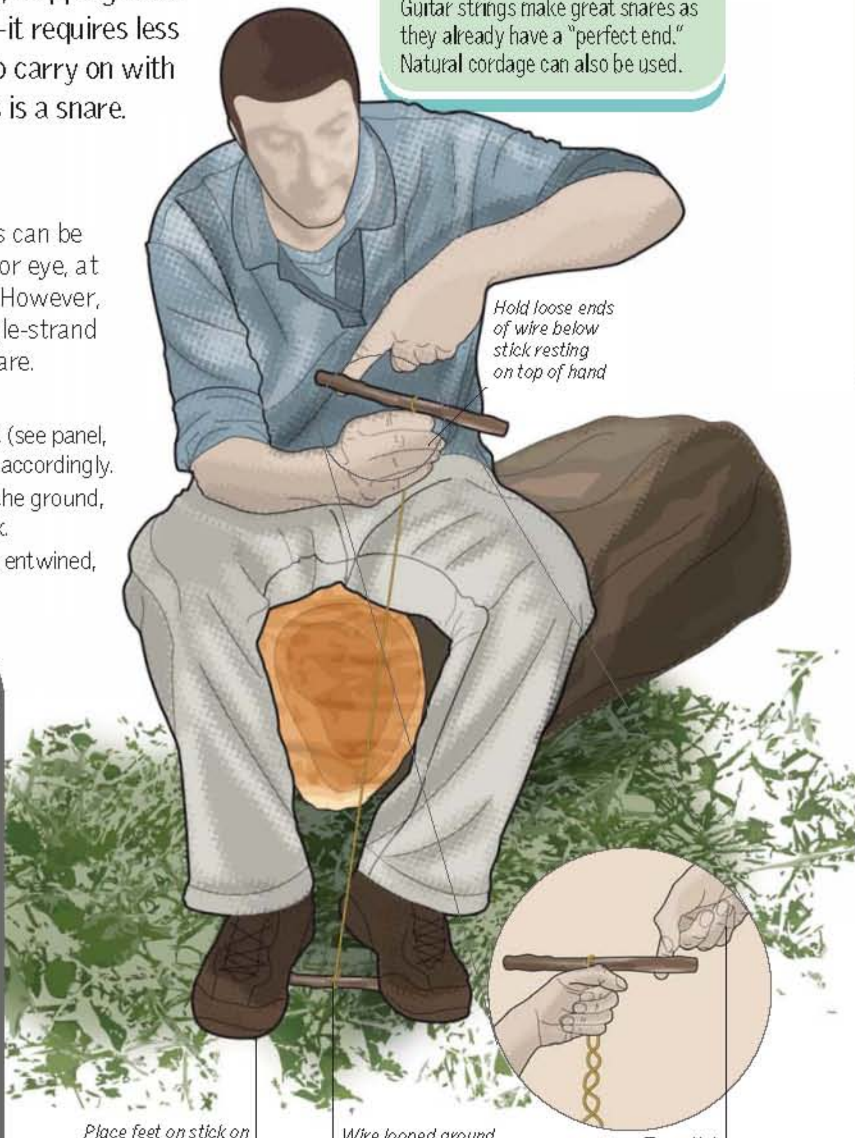

Hold loose ends of wire below stick resting on top of hand
Place feet on stick on either side of wire to hold it in place
Wire looped around stick on ground forms "perfect end"
Turn stick with other hand until the wire strands are intertwined

ALTERNATIVE MATERIALS

Guitar strings make great snares as they already have a "perfect end." Natural cordage can also be used.

KILLING A RABBIT

Unless you're an expert, breaking a rabbit's neck by holding its head and pulling on its legs can be difficult and could result in you dislocating its hips, causing it more pain. The best method is to strike the rabbit on the back of the neck with a solid stick, thus avoiding having to pick it up and risk being bitten or scratched. The eyes will glass over immediately when the animal is dead.

LOOK AND LISTEN

To be successful at hunting and trapping, you need to decide which types of animal you are going to try to catch. To do that, you need to find out what animals are in the vicinity and where exactly they are to be found. Prey animals will use their keen senses to avoid being caught—so you will have to use your senses to find them. Look for any signs of animals, both on the ground and in the trees. And use your ears—you may be able to hear an animal even when you can't see it.

SIGNS TO LOOK OUT FOR:

- Runs, trails, and tracks
- Droppings
- Chewed or rubbed vegetation
- Feeding and watering areas
- Lairs, dens, and resting sites

SQUIRREL POLE

Make several two-strand snares with nooses 3 in (8 cm) in diameter and stake-out cords 8 in (30 cm) in length. Place them around a long pole, with the lowest 2 ft (60 cm) off the ground. If you snare a squirrel, leave it, as others will come to investigate.

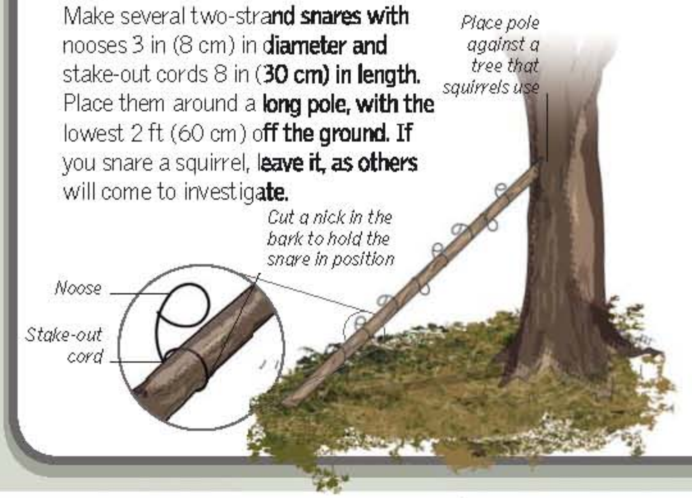
Place pole against a tree that squirrels use
Cut a nick in the bark to hold the snare in position
Noose
Stake-out cord

2 To make the noose, pass the imperfect end, or securing loop, through the perfect end, or running loop. The running loop has no loose ends of wire so it won't snag in an animal's fur and prevent the noose from tightening as it struggles.

- Use a single prusik knot (see p. 210) to tie a length of cord to the securing loop.

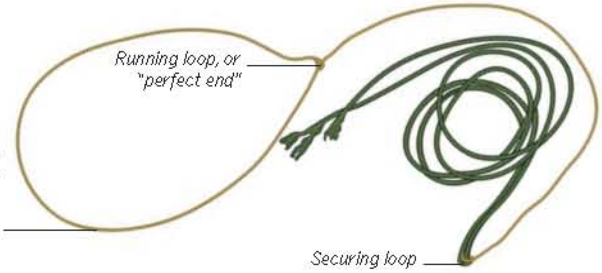
Running loop, or "perfect end"
Using wire for the snare means the noose keep its shape
Securing loop

3 Before setting the snare, either bury it in the ground for a few hours or pass it over a flame for a few seconds to remove any human scent and dull its surface.

- Don't waste snares; always set them on runs you know the animals are using and set as many as you can.
- Use sticks to make a support from which to suspend the snare, or bend saplings into an arch.

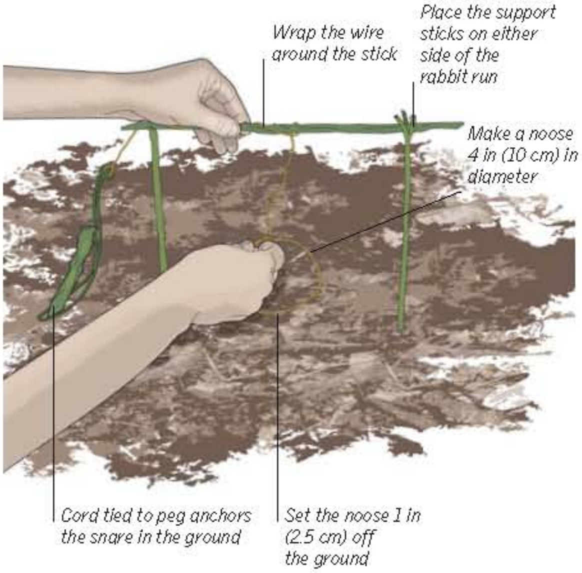
Wrap the wire around the stick
Place the support sticks on either side of the rabbit run
Make a noose 4 in (10 cm) in diameter
Cord tied to peg anchors the snare in the ground
Set the noose 1 in (2.5 cm) off the ground

4 Avoid disturbing or treading on the run, as the disturbance or your scent could alert the animal. Camouflage the frame with vegetation, which you can also use to funnel the animal toward the snare.

- Use natural vegetation, such as holly or other prickly shrubs, to form a funnel on either side of the snare.
- The funnel helps to ensure that the animal has no option but to pass through the snare.

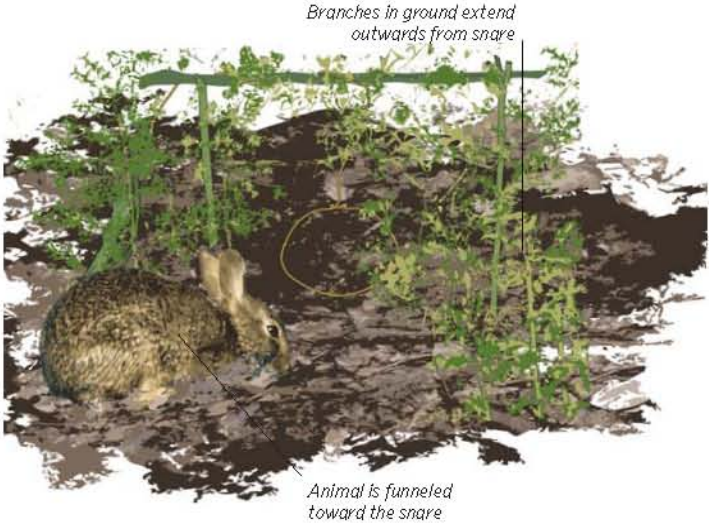
Branches in ground extend outwards from snare
Animal is funneled toward the snare

WEAPONS AND TECHNIQUES

A spear is the simplest weapon to make and use, but if you have the time and materials you can make more complicated weapons, although they require more skill to use. Learning other basic techniques, such as catching insects, is also invaluable when hunting.

THROWING STAR

With a throwing star you're four times more likely to hit prey than with a spear, and, if you don't succeed in wounding the animal with one of the sharp ends, the weight of the weapon will at least stun it. Keep it close at hand and ready to use, should the opportunity arise.

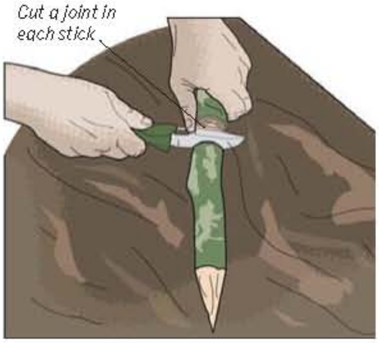
Cut a joint in each stick

1 Find two sticks 2 in (5 cm) thick and 18 in (45 cm) long. Make a square-cut joint in the center of each, then sharpen each end.

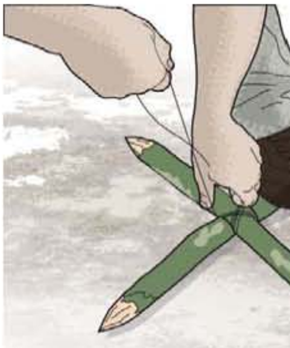

2 Join the sticks together by overlapping the two square-cut joints and lashing them to each other with paracord or natural cordage.

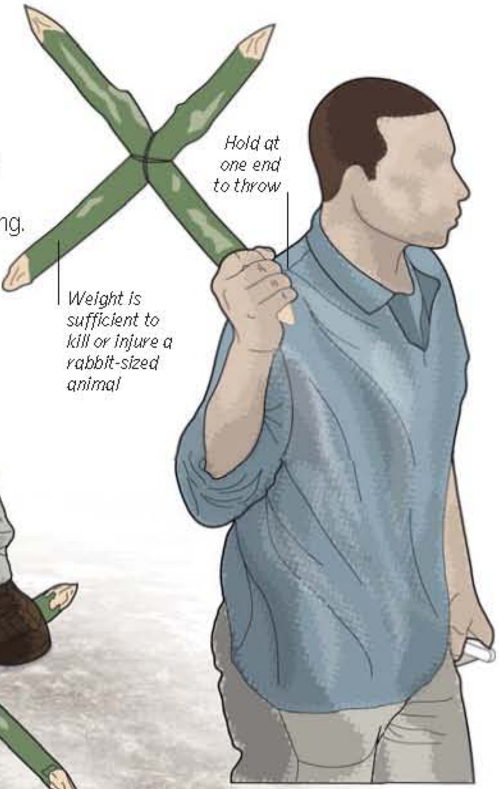
Hold at one end to throw
Weight is sufficient to kill or injure a rabbit-sized animal

3 You can exert more force if you're throwing from a standing position. Aim for the animal's legs.

RODENT SKEWER

A forked spear can be used to catch small mammals in their burrows. Thrust the pointed end into the hole. When you can feel the animal, twist the stick until you have snagged its fur, then carefully pull it out.

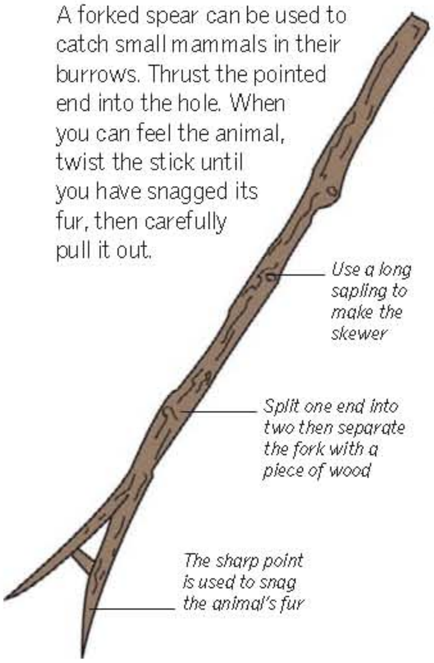
Use a long sapling to make the skewer
Split one end into two then separate the fork with a piece of wood
The sharp point is used to snag the animal's fur

SLINGSHOT

A slingshot can be used to kill small animals. Cut a strong forked branch and make a notch in each fork. Take the rubber tubing from your survival kit (see pp. 60–61) and thread it through a piece of leather or plastic. Tie the ends of the tubing to the notches. Place a pebble in the pouch, stretch back the sling, and take aim.

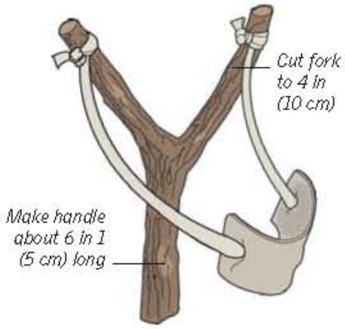
Cut fork to 4 in (10 cm)
Make handle about 6 in 1 (5 cm) long

NOOSE STICK

You can use a noose stick to lasso lizards, slow-moving mammals, or roosting birds, jerking the pole to tighten the noose around the animal's neck. Use a wire or cordage snare (see pp. 216–17). When stalking prey, move slowly and very quietly.

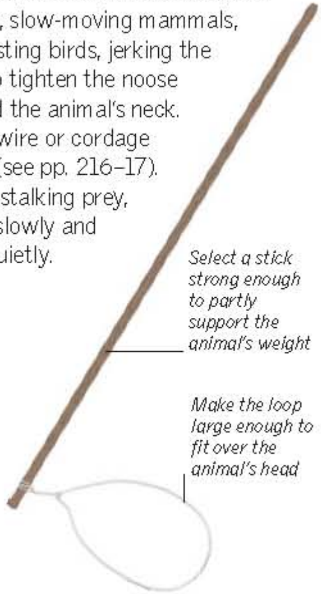
Select a stick strong enough to partly support the animal's weight
Make the loop large enough to fit over the animal's head

EASY PICKINGS

Ants are social insects, and most species are aggressive in defense of their nest. They have a stinging bite and some species will also squirt formic acid at their attacker. However, they make a nutritious meal if gathered with care. In summer months in northern temperate regions, you can eat the larvae of wood ants. And, if you use the technique shown below, you can get the ants to do all the work of collecting the larvae together for you. You'll have to vandalize the nest, but don't remove all of it.

ANT CUISINE

Ant larvae are best fried—they taste like shrimp. Adult ants also make a tasty snack—in Bogotá, Colombia, movie theaters sell roasted leaf-cutter ant abdomens instead of popcorn.

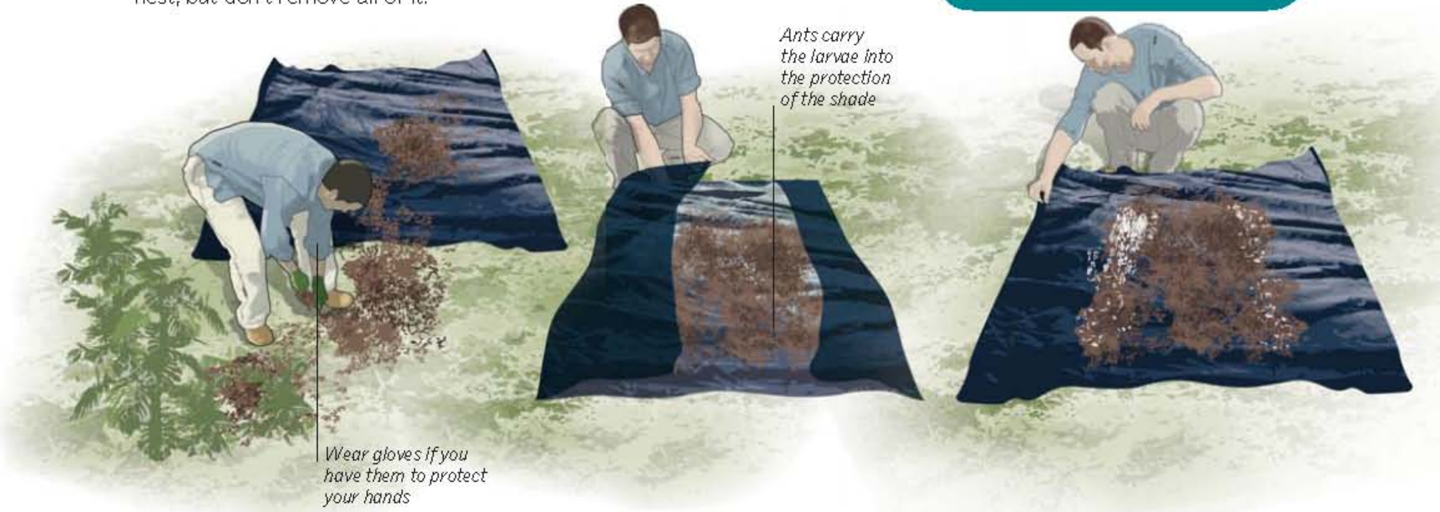
Ants carry the larvae into the protection of the shade
Wear gloves if you have them to protect your hands

① Place a tarpaulin in a sunny patch next to the nest. Scoop the nest material, ants, and larvae into the middle of the tarpaulin.

② Lay some sticks on the tarpaulin, close to the edge, then fold over the sides to create shade. The ants will carry the larvae into the shade.

③ After a while, throw back the folded sides of the tarpaulin and scoop up the larvae, which look like fat grains of white rice.

CATCHING INSECTS

Many flying insects can be eaten, but catching them takes a bit of ingenuity. Crawling insects, such as ants (see above) and termites, can provide a good meal if you can collect them in sufficient quantities.

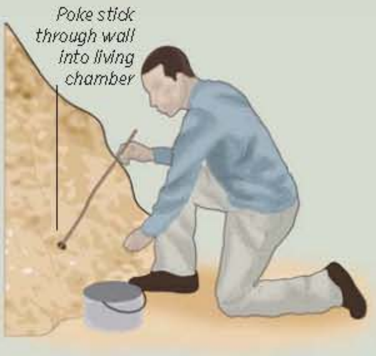
Poke stick through wall into living chamber

① Cut a long, thin, straight stick then peel it until smooth. Push it slowly into the termite hill.

TERMITE FISHING

Termites are found in the tropics and subtropics. Some species live in vast numbers inside mounds, which they build out of mud and saliva. If a foreign object breaches the walls, the termites will attack it with their powerful jaws.

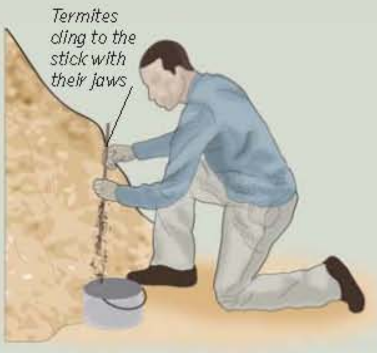
Termites cling to the stick with their jaws

② Remove the stick and scrape off the attached termites into a container ready for frying or roasting.

NIGHT FLIERS

Nocturnal flying insects, such as moths, are attracted to light. Stretch a white sheet between two branches and peg the bottom taut above a bowl of water. Hang a flashlight behind the sheet. Insects will fly into the sheet and fall into the water.

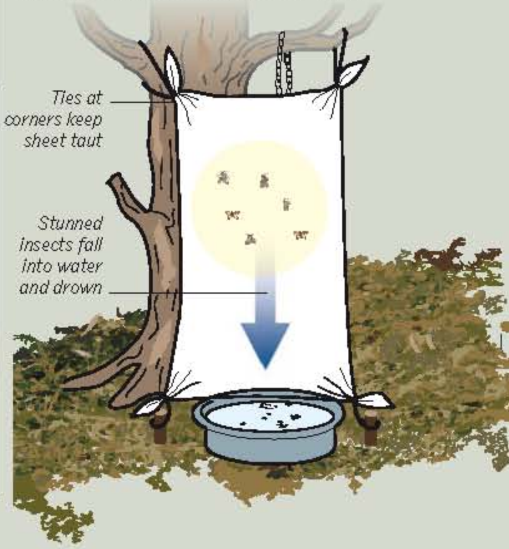
Ties at corners keep sheet taut
Stunned insects fall into water and drown

PREPARING SMALL MAMMALS

PARASITES SUCH AS FLEAS and lice will leave the body once a warm-blooded animal is dead so, if your circumstances permit, leave the carcass to cool before preparing it for cooking. In a hot climate, place the body in the shade. However, don't let it get too cold, as skin is more easily removed when a body is warm.

RABBIT-SIZED MAMMALS

The method shown here applies to all furred mammals approximately the same size as a rabbit. Before gutting the animal, remove any urine by holding the body by the forelegs and progressively squeezing down from the chest toward the bowels, making sure that you direct the spray away from your own body. You'll need to remove any scent glands, which are usually inside the forelegs or around the anus, in case the musk taints the meat (take particular care when handling a skunk).

WARNING!
Rabbits, hares, and rodents may be infected with a bacteria called tularemia, which can be fatal to humans. Do not touch these animals with your bare hands if you have a scratch on your skin. If you have no gloves, cover your hands in soap lather before handling the animal, and wash your hands when you have finished. The germ is destroyed by heat, so cook the meat well. Myxamatosis, a viral disease that affects rabbits' mucous glands, is not harmful to humans.

GUTTING

No part of a mammal carcass should be wasted. If your camp is near water, keep the guts for fish bait. The offal should be removed carefully and eaten if healthy—discard any organ that's pale or spotty. Liver is rich in essential vitamins and minerals and needs little cooking. It should be eaten as fresh as possible (removing the gall bladder from the middle first). Kidneys, which in most mammals are surrounded by fat, are also a valuable source of nutrition.

1 Place the animal on its back on clean ground, such as a bed of pine boughs, with its head pointing toward you. Cut a small hole in its belly with the point of your knife, taking care not to pierce the guts.

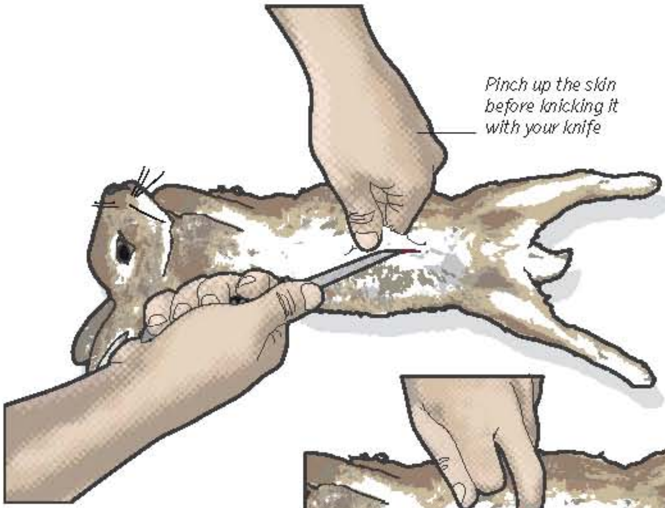
Pinch up the skin before knicking it with your knife

WARNING!
Rabbits make an easy meal, but their flesh lacks fat and vitamins so, if your diet consists entirely of rabbit, your body will use its own vitamins and minerals to digest the meat. These are then lost from your body in your feces and, if not replaced, you'll get diarrhea and so become hungry and weak. Eating more rabbit makes the condition worse and you will eventually die of starvation. You must supplement your diet with some vegetation and fatty food.

2 Pull the skin apart at the cut and insert a finger from each hand into the opening.

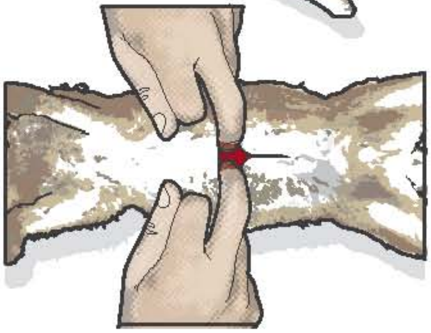

3 Prise open the belly to expose the guts; remove them along with the heart, liver, and kidneys. Wash your hands before moving on to skinning.

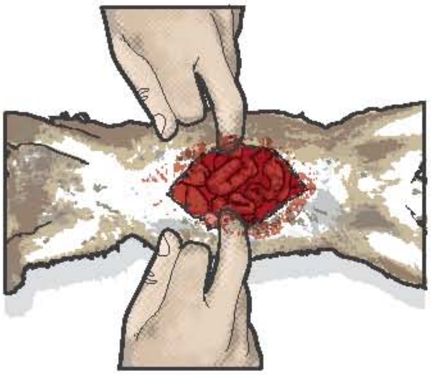

SKINNING

If you wish to keep the skin for use later (to make a pair of mittens, for example), you should remove it as carefully as possible. You'll also need to cure it; stretch the skin as tight as possible and leave it in the sun or hang it close to the heat of a fire to dry. Rubbing wood ash into the skin will help to speed up the process.

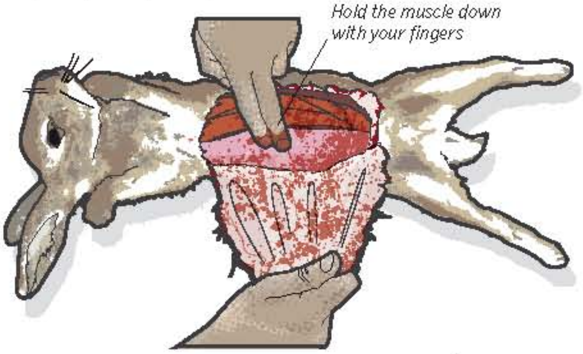
Hold the muscle down with your fingers

1 Beginning at the belly, separate the skin from the muscle surrounding the gut cavity. You'll find that it pulls away quite easily. When you have reached the back on one side, repeat the process on the other.

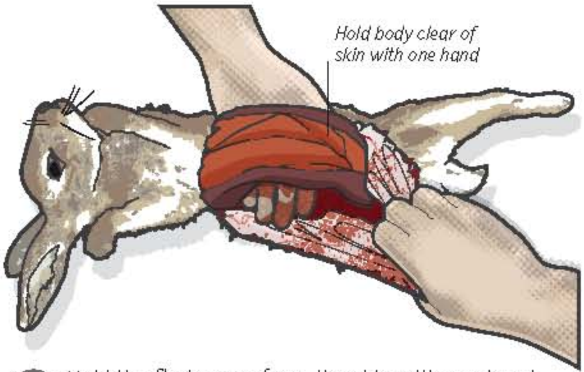
Hold body clear of skin with one hand

2 Hold the flesh away from the skin with one hand and pull the skin over the rear legs, one by one. The hindquarters are now free of skin and fur, except for the tail, which should be cut off and the scent glands removed.

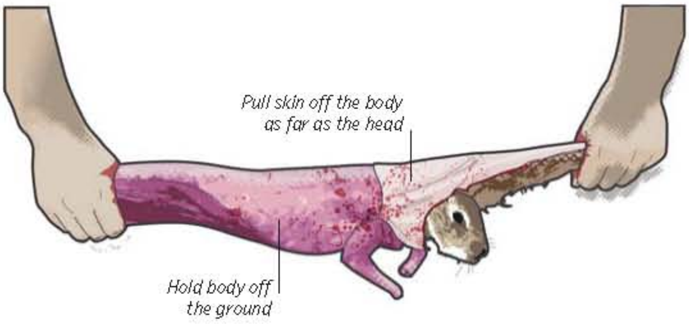
Pull skin off the body as far as the head
Hold body off the ground

3 Hold the animal by the rear legs and pull the skin forward, easing out both forelegs. Pull the skin over the neck and then cut off the head. Wash the body to remove any traces of fur.

GLOVE SKINNING

The preferred survival method of skinning and gutting a rabbit-sized mammal is to suspend it by its rear legs from a strong branch as this keeps the animal off the dirty ground.

- If you have no knife to hand, snap off the lower part of a foreleg and use the sharp edge of the broken bone to cut the skin.
- Gut the animal, then cut the skin around all four paws and between the rear legs.
- Tug the skin down toward the head.
- If you wish, pull the skin over the head. Otherwise just cut the head off.

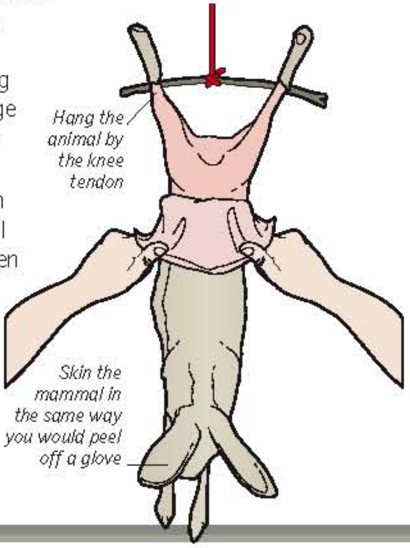
Hang the animal by the knee tendon
Skin the mammal in the same way you would peel off a glove

HOW TO SKIN A SQUIRREL

Squirrel flesh is tender and tasty. Once you have skinned the squirrel, you can simply skewer it on a spit and roast it whole over a fire. Alternatively, you can cut the meat into pieces and make a stew. A large squirrel will feed two people.

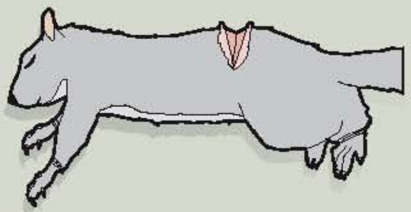

1 Gut the squirrel following the method shown opposite. Cut the skin around the paws, then cut through the skin across the middle of its back.

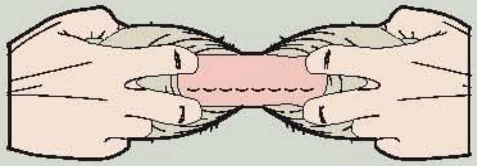

2 Insert two fingers under the skin on either side of the cut and pull the two pieces apart and off the body. Wash off any traces of fur.

PREPARING LARGE MAMMALS

ALL LARGE MAMMALS are edible, but you must not eat the livers of polar bears or bearded seals as they contain toxic levels of vitamin A. Wild game is sometimes infected with roundworm larvae and, if you eat the meat raw or undercooked, it will infect you with parasites, so always cook all wild game thoroughly.

WARNING!
Approach all game with caution, as the animal may still be alive and most wounded animals are extremely dangerous. To check, touch its eye with the tip of a long stick (or the muzzle of a gun). Even if the animal is unconscious, it will blink if it's alive. Always skin, gut, and butcher game at the site of the kill. You don't want the smell of fresh blood to attract predators or scavengers to your camp.

PREPARING THE KILL

Lifting large animals, such as deer, uses energy so, if you're alone, prepare the carcass on the ground. However, if you're in a group and circumstances permit, it's best to suspend the animal by its hocks (the tarsal joints of its hind legs) from a branch, as it will bleed better and will be easier to skin and gut. If you can't find a suitable tree from which to hang it, build a frame.

BLEEDING

Bleeding is essential to preserve the meat and stops it from tasting too gamy. It also helps to cool the carcass. Slit the animal's throat from ear to ear and let the blood drain out. Blood is rich in vitamins and minerals, including salt, so, if you're able to, collect it in a container to use later in a stew. Cover the container to protect the blood from flies and to keep it cool.

SKINNING

If you wish to preserve the hide, it's better to take the skin off before gutting. Cut the belly skin from throat to tail, cutting around the genitals. Cut along each leg from above the foot to the belly. Pull the hide off the carcass, severing connective tissue as necessary. If you're skinning on the ground and you don't want to keep the hide, you can use it to protect the meat—remove it fully only after jointing is complete.

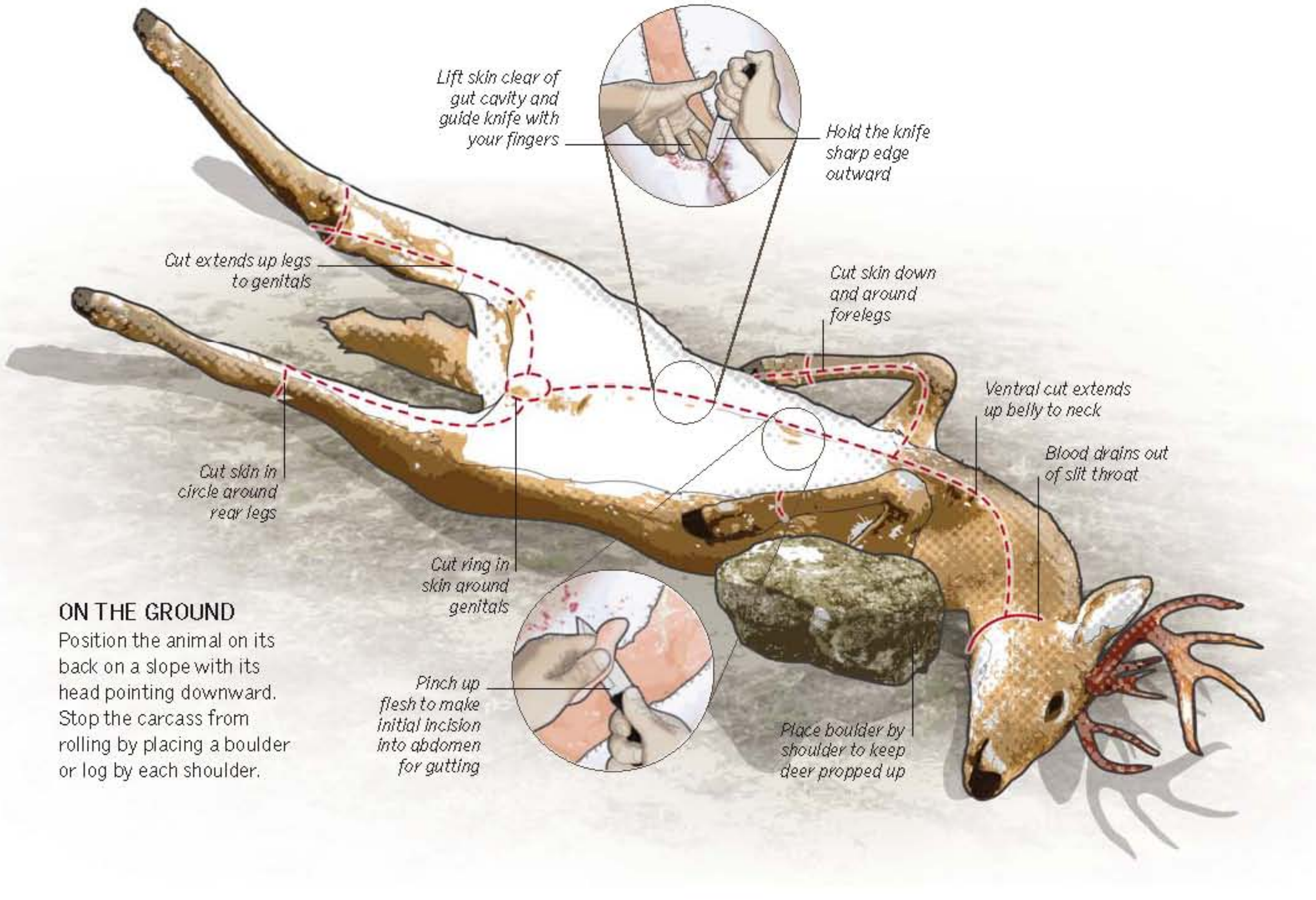
Lift skin clear of gut cavity and guide knife with your fingers
Hold the knife sharp edge outward
Cut extends up legs to genitals
Cut skin down and around forelegs
Ventral cut extends up belly to neck
Blood drains out of slit throat
Cut skin in circle around rear legs
Cut ring in skin around genitals
Pinch up flesh to make initial incision into abdomen for gutting
Place boulder by shoulder to keep deer propped up

ON THE GROUND

Position the animal on its back on a slope with its head pointing downward. Stop the carcass from rolling by placing a boulder or log by each shoulder.

GUTTING

The method shown here is for gutting a carcass on its back. When finished, inspect the heart, liver, and kidneys for signs of worms or other parasites and, if the organs are healthy, keep them to eat. If the liver is spotted, a sign of disease, discard all internal organs and boil the meat.

1 To avoid piercing the internal organs, pinch up the abdomen near the breastbone and make an incision big enough to insert two fingers.

- Use your fingers to guide the knife, cutting toward, then around, the anus with the sharp blade-edge upward.

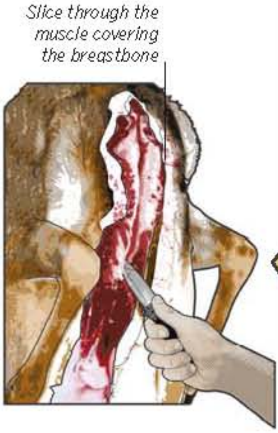
Slice through the muscle covering the breastbone

2 Using a saw, if you have one, cut through the outer muscle and the breastbone to open up the chest cavity.

3 Using the knife, slice through the diaphragm muscle, which separates the chest cavity from the gut cavity, getting as close to the spine as you can.

- Remove the liver, taking care not to cut the gall bladder in the center.

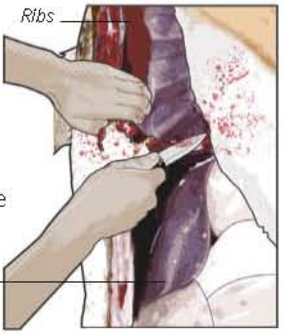
Ribs
Healthy liver is dark red-purple

4 Reach up through the chest cavity and cut the windpipe and esophagus. Hold them with one hand and pull out all the internal organs as one unit.

- Check that the anus is clear of feces, pushing a hand through if you can't see daylight.

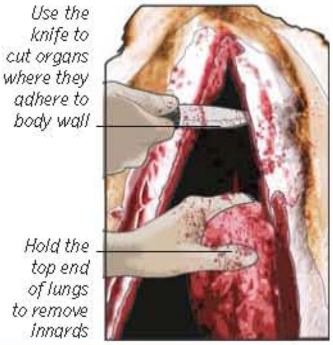
Use the knife to cut organs where they adhere to body wall
Hold the top end of lungs to remove innards

KEEP THE GUTS

If you're able to carry them, take the guts, reproductive organs, and glands back to your camp in a sealed container. You can use them as bait in traplines or for fishing. Keep the fat that surrounds the intestines for cooking.

JOINTING

Large game should be cut into manageable pieces that can be carried back to camp. Boning the meat helps to reduce the weight. Bag or wrap the meat as you remove it to keep it clean and free of flies, and keep it in the shade in hot climates. How a carcass is divided partly depends on the species, but the hindquarters contain the steaks and the best cuts.

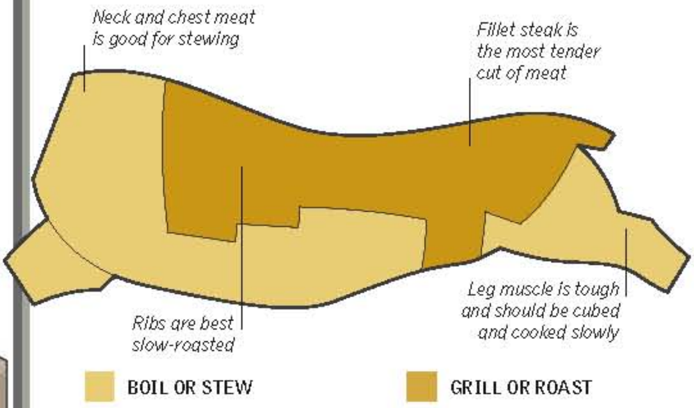
Neck and chest meat is good for stewing
Fillet steak is the most tender cut of meat
Ribs are best slow-roasted
Leg muscle is tough and should be cubed and cooked slowly
BOIL OR STEW
GRILL OR ROAST

USES FOR HIDES

Keep the skin if you can. When dried, it's light and can be used as a blanket, or it can be made into an item of clothing. Hide is one of the best materials for lashes. Sinew also makes good cordage (see p. 140).

TANNING THE HIDE

The best way to clean a skin is to stretch it on a frame and then carefully scrape off the fat and any remaining scraps of flesh, using a flint or a piece of bone. Every mammal, with the exception of buffalo, has enough brain matter for use in tanning its own hide. Mash the brains well in warm water, then apply the mixture to the hairless side of the skin and leave it to dry for 24 hours.

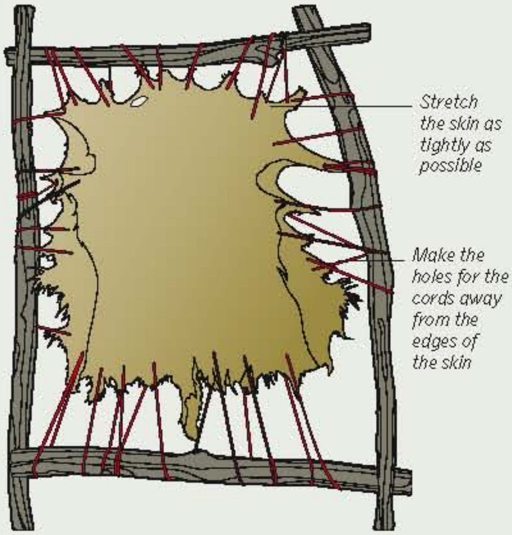
Stretch the skin as tightly as possible
Make the holes for the cords away from the edges of the skin

PREPARING OTHER ANIMALS

ALL TYPES OF ANIMAL should be considered when searching for your next meal. Invertebrates are far more plentiful and widespread than other animals, and take less energy to gather and prepare. Reptiles and amphibians can also provide a vital source of nutrition.

EATING FOR SURVIVAL

The thought of eating grubs or grasshoppers might not appeal, but experimentation is vital in a survival situation—don't forget that in some cultures such creatures are highly prized, if prepared properly. However, some animals are less palatable than others, in which case you should chop them finely and add to a stew.

SLUGS, SNAILS, AND WORMS

Slugs should be avoided as they often feed on poisonous fungi. To prepare snails, either starve them or feed them safe food such as wild garlic for 24 hours to purge their guts before cooking. Either boil them in water or bake them in their shells in hot embers until the juices bubble over. Avoid all marine snails and any terrestrial snail with a brightly colored shell as they may be toxic. All earthworms are edible. Place them in salty water until they are purged, then boil, though they can be eaten raw if necessary.

EDIBLE INSECTS

Insects are mostly protein and make good emergency food. Avoid any hairy or brightly colored insects or those that emit a foul smell (including larvae). As a general rule, avoid adults that sting or bite, but ants and honey bees can be eaten if collected carefully. Most insects, except those with a hard carapace, like beetles, can be eaten raw.

HOPPING INSECTS

Grasshoppers and crickets have large leg muscles, and most are quite tasty when cooked (avoid brightly colored ones). Remove the antennas, wings, and leg spurs, and roast to kill any parasites.

WITCHETTY GRUBS

Witchetty grubs are a highly nutritious Australian bush delicacy. They can be eaten raw but, if roasted quickly in hot ashes, they taste of scrambled eggs. Treat palm grubs in the same way.

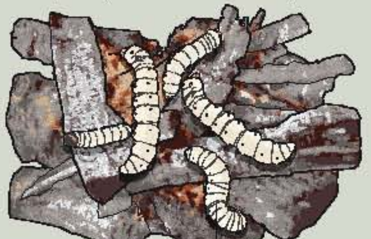

Large, white grubs high in protein and calcium

AMPHIBIANS AND REPTILES

All frogs in the *Rana* genus (see pp. 292–93) are edible. Skin and then boil or roast them. The best, most meaty part is the hind legs. Don't handle or eat toads or brightly colored tropical frogs as many have highly toxic skin secretions. Lizards and snakes are a good source of protein. They must be skinned and gutted, then either roasted on a stick if small, or cut into small pieces and boiled. All snake flesh is edible—and tastes like chicken—apart from the head, which must be removed.

1 Hold the dead snake firmly behind the head and cut it off about 6 in (15 cm) down the body.

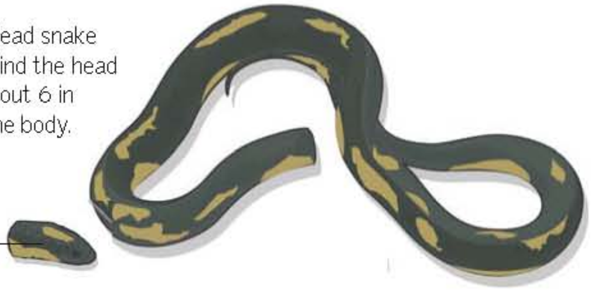

Bury the head if the snake is venomous

2 Slit the belly with a sharp knife and remove the entrails and other organs. Either throw them away or keep as bait.

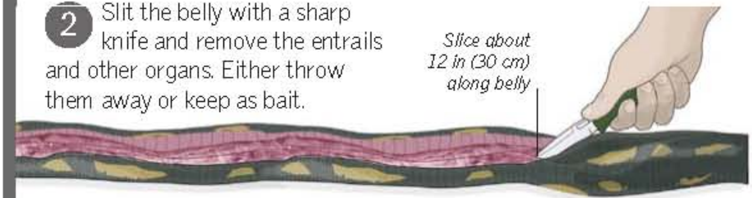

Slice about 12 in (30 cm) along belly

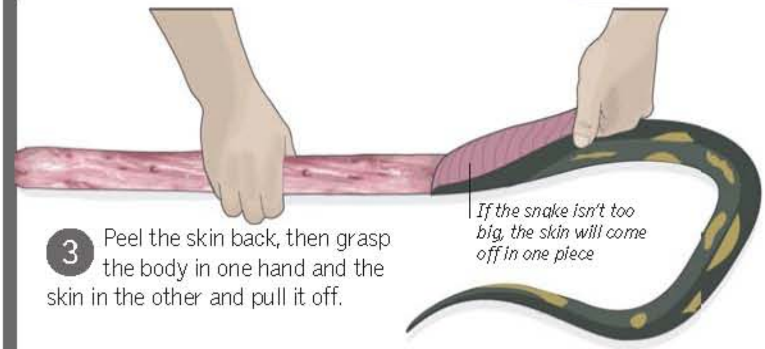

If the snake isn't too big, the skin will come off in one piece

3 Peel the skin back, then grasp the body in one hand and the skin in the other and pull it off.

PREPARING A CRAB

Avoid their strong claws (tie them up if you can). To kill a crab, plunge it in boiling water or stab it through the eye socket or the orifice under the shell flap on the body. Cook the crab for about 15 minutes.

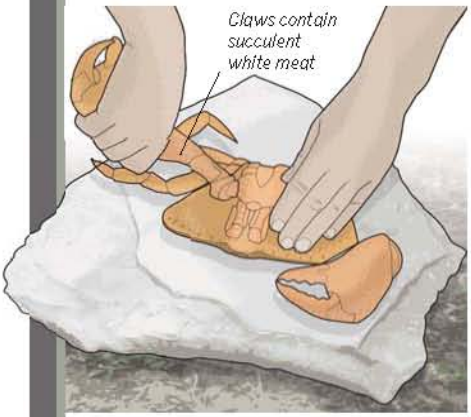
Claws contain succulent white meat

① Place the cooked crab on its back, then twist off its legs and the claws. They all contain edible meat, so crack them open with a rock.

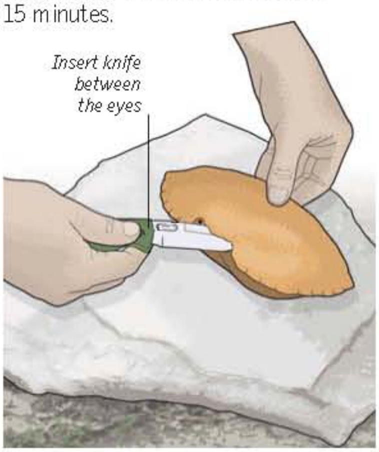
Insert knife between the eyes

② Open the shell by inserting the point of your knife between the two halves and twisting it. Lift back the lower shell.

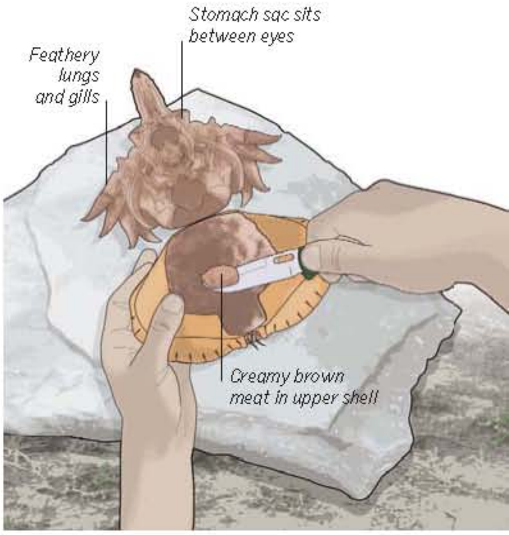
Stomach sac sits between eyes
Feathery lungs and gills
Creamy brown meat in upper shell

③ Remove and discard any green matter and the lungs, gills, and stomach as they're all poisonous. Scoop the meat out of the shell.

SHELLFISH

Shellfish can be found in streams and on lake and sea shores. Don't collect marine shellfish that are not covered at high tide or are near any source of pollution. Shellfish must be alive when you collect them—bivalves, such as mussels, will close their shell if you tap them, and univalves, such as limpets, will cling tightly to their rock—and you must cook and eat them immediately.

HOW TO SHUCK AN OYSTER

Oysters are rich in vitamins and minerals. To eat an oyster raw, first open its shell by inserting a blunt knife into the hinge at the thicker, more pointed end of the oyster and twisting. However, in a survival situation, play safe and boil oysters in their shells for five minutes after their shells open. Do not eat any whose shells do not open during cooking.

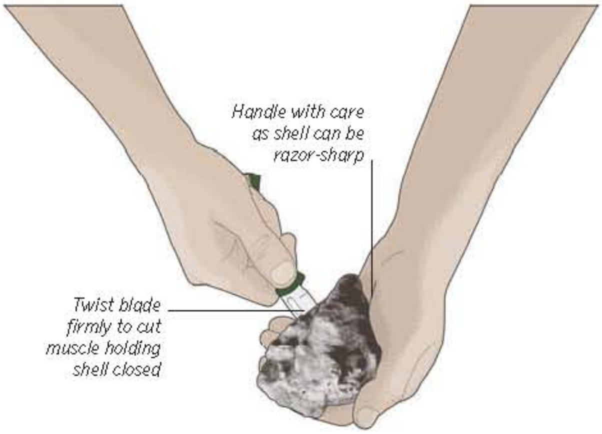
Handle with care as shell can be razor-sharp
Twist blade firmly to cut muscle holding shell closed

WARNING!

Tropical species of mussel should be avoided in the summer as they may be poisonous. The black mussel, which is found on Arctic and sub-Arctic coasts, may be poisonous at any time of the year. The symptoms of mussel poisoning are acute nausea, vomiting, diarrhea, abdominal pain, and fever.

SHRIMPS, CRABS, AND LOBSTERS

Small crustaceans, such as shrimps and crawfish, should be cooked in boiling water for five minutes. Larger crustaceans, such as some crabs and lobsters, should be boiled for up to 20 minutes. A tasty alternative to boiling shrimps or prawns is to roast them on skewers made of green wood, placed in the embers of a fire.

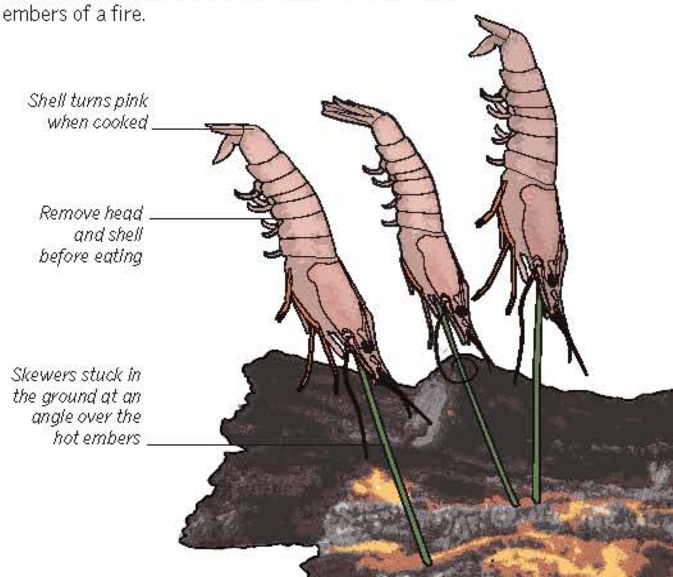
Shell turns pink when cooked
Remove head and shell before eating
Skewers stuck in the ground at an angle over the hot embers

CATCHING BIRDS

BIRDS CAN BE TRAPPED in a variety of ways—in snares, nets, tunnels, or cages—and they can be hunted with noose sticks or any projectile weapon. The trick is to observe their behavior and to match your technique—and bait—to the bird. You need to discover their regular roosts, feeding spots, and flight paths.

WARNING!
Eggs are easy to collect, particularly from ground-nesters, although gulls, for example, will attack you if you try. However, taking eggs from nests is illegal in many countries and should only be done if absolutely necessary.

FIGURE-4 CAGE TRAP

This trap enables you to catch ground-feeding birds (or rabbit-sized mammals) without killing or wounding them, so you can choose when you want to kill them and eat the meat fresh. The "Figure 4" is the trigger that's used to drop the cage over the animal.

MAKING THE TRIGGER

To construct the Figure 4 trigger you need to make three notched sticks (bait stick, upright stick, and release stick) and then link them together as shown below.

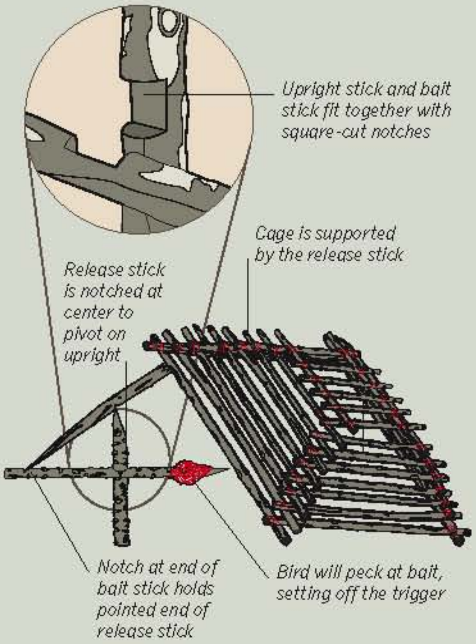
Upright stick and bait stick fit together with square-cut notches
Cage is supported by the release stick
Release stick is notched at center to pivot on upright
Notch at end of bait stick holds pointed end of release stick
Bird will peck at bait, setting off the trigger

MAKING THE CAGE

Create a pyramid of sticks by placing progressively shorter sticks across longer sticks and lashing them together at the ends. Balance the cage over the bait.

ON THE GROUND

Gamebirds, such as quail, and inquisitive scavengers, such as crows, can be lured into traps if you use the right bait. Migratory waterfowl, such as ducks and geese, molt in the late summer, which means they are easier to catch as they can't fly away.

WALK-IN TUNNEL TRAP

A tunnel trap can be used to catch gamebirds, which have stiff feathers that lie in one direction and bend only with difficulty. As a bird tries to retreat from the trap, its feathers will become wedged in the tunnel walls.

- Dig a near-horizontal, funnel-shaped tunnel close to the ground.
- Lay a trail of bait (seed or berries) leading to the rear of the tunnel.
- As it eats the bait, the bird will move deeper into the tunnel and will be unable to back out.

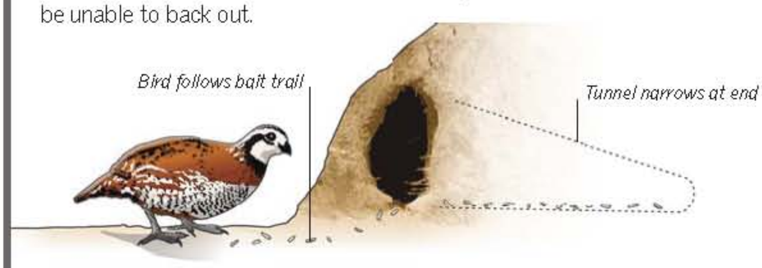
Bird follows bait trail
Tunnel narrows at end

ON THE WING

A fine net stretched between two trees across a flight path is an effective method of catching birds in flight. Throwing a bola requires more skill, and works by entangling the bird in the spinning ropes.

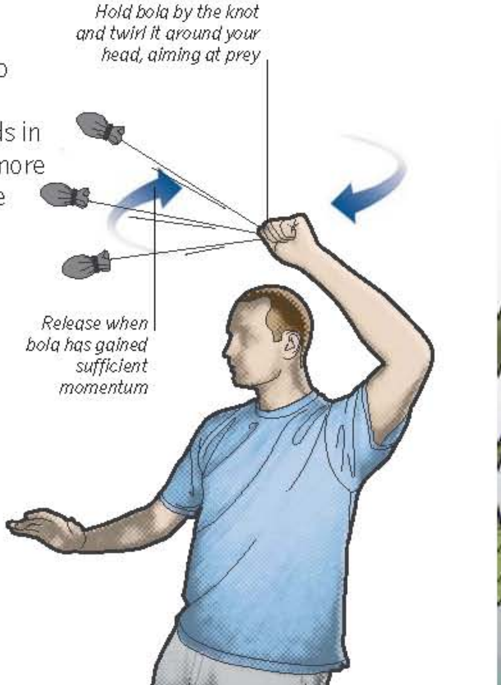
Hold bola by the knot and twirl it around your head, aiming at prey
Release when bola has gained sufficient momentum

MAKING A BOLA

To make a bola, use an overhand knot (see p. 143) to tie three 3-ft (1-m) lengths of cord together about 3 in (8 cm) from one end. Find three rocks weighing about 7 oz (200 g), wrap each one in a piece of cloth, and tie it up with the free end of a cord.

ON THE NEST

Birds are creatures of habit and will usually roost in the same place every night, such as in the branches of a tree as protection against ground predators. They rarely move once it's dark. Many birds are more accessible when breeding as, once you've located the nest, you'll always know where to find them. However, you should only take nesting birds in a genuine survival situation.

MAKING A SNARE STICK

A snare stick is used to trap birds roosting in trees. Remove a bird once it has been snared, as its squawking and fluttering will alert other birds to the danger and scare them off.

- Make several single-strand snares (see pp. 216–27) with a loop diameter of about 1–2 in (2.5–5 cm), depending on the size of the bird.
- Place them close together along a stick, making notches in the stick to hold them in position with the loops uppermost.
- Tie the stick on top of a branch where you have seen birds roosting.

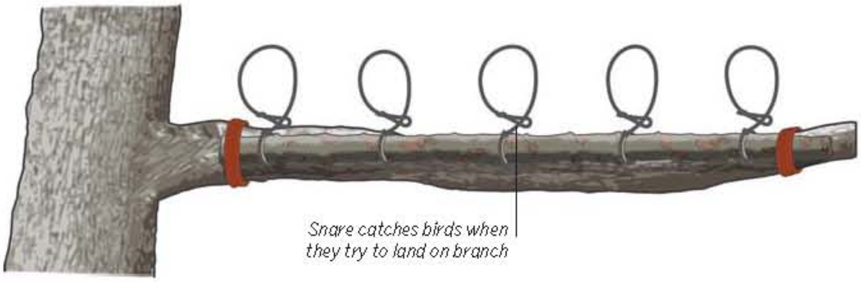
Snare catches birds when they try to land on branch

PHEASANT POLE

Pheasants roost in trees at night, usually returning to the same branch, and they won't leave once it's dark. To snare a sleeping pheasant, raise the pole up in front of it. The bird will wake and peck at the foil. As its head goes through the loop, pull the cordage and the snare will close around the bird's neck.

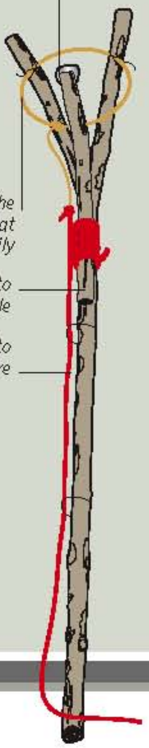
Silver foil or shiny paper
Lash the snare loop to the fork with something that will break easily
Lash foil stick to long forked pole
Attach cordage to snare wire

TO MAKE THE POLE

Cut a long forked pole. Tie a stick in the V of the fork and attach a piece of silver foil to it. Weakly lash the loop of a snare to the fork, then run a length of cordage down the pole from the snare wire.

BOUGH FENCE TRAP

This method is used to trap ducks, which often go ashore during the day to sun themselves, or at night to roost. They tend to favor small islands as protection against predators.

- Cut several saplings and make double-stranded snares (see p. 216).
- Place the boughs a few feet from the water's edge, and bend them into a series of overlapping hoops to form a fence around the island.
- If you only have a few snares, place them over the ducks' inward and outward tracks. Placing logs between the snares will then force the birds to pass through the hoops in order to get onto the island.
- Suspend a snare from the center of each bough arch.

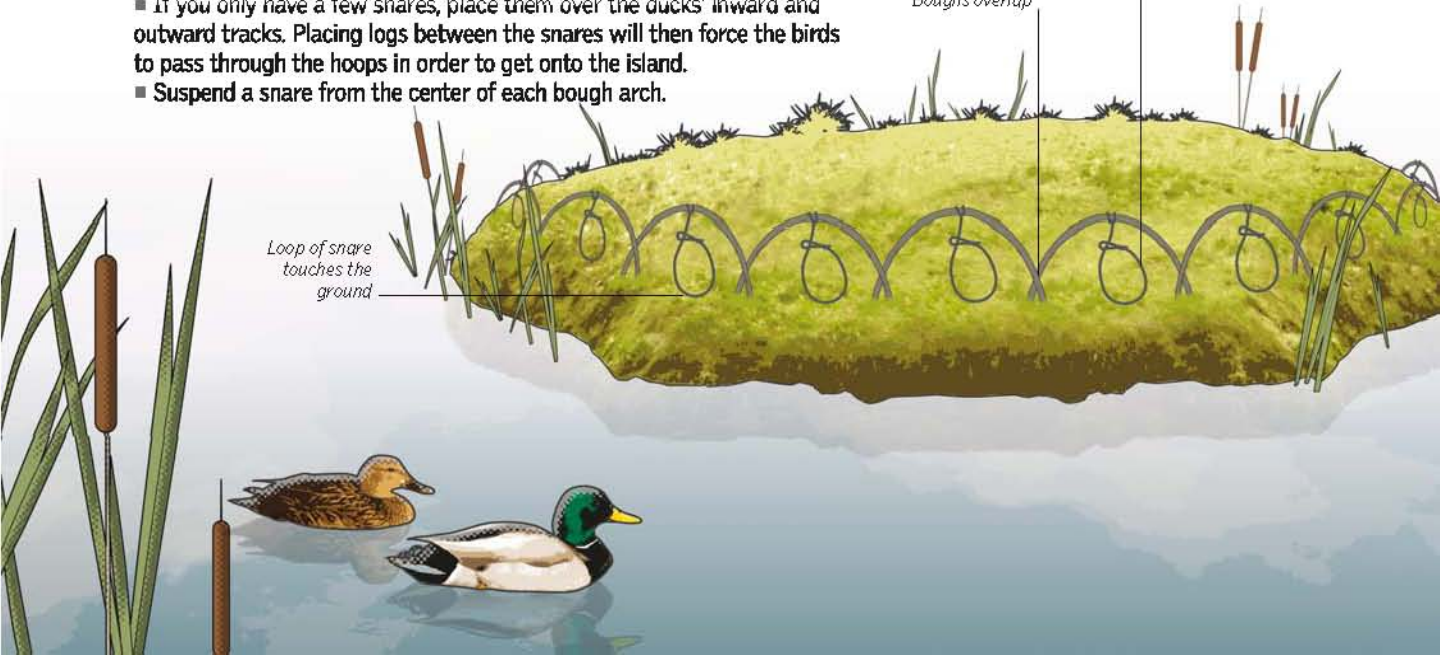
Loop diameter is approximately 2 in (5 cm)
Boughs overlap
Loop of snare touches the ground

PREPARING BIRDS

GAMEBIRDS AND WATERFOWL are the most commonly eaten wild birds, but all birds are edible if prepared correctly. Skinning a bird is the quickest way to get at the meat, but this removes the nutritional value of the skin, so always pluck a bird if you have the time to do so.

WARNING!
Handle scavengers and carrion eaters, such as crows, buzzards, and vultures, as little as possible, because they are more likely to be infested with lice and ticks, and are prone to infection.

- The meat of such birds must be boiled for at least 30 minutes to kill any infectious organisms. This will also tenderize the meat.
- Always wash your hands after handling any bird.

GETTING STARTED

If the bird hasn't already been killed by the method of capture you'll need to dispatch it—for example, by stretching its neck and cutting its throat (for the best way to kill gamebirds, see below). Whichever method you use to kill the bird, you must bleed it before plucking, but make sure you pluck it while the body is still warm.

KILLING GAMEBIRDS

To quickly and humanely dispatch a gamebird, such as a pheasant, fold its wings into the body and hold it under your arm. Cover its head with your jacket to calm it down (this also reduces the chances of you being injured by its claws or beak). Still holding the bird firmly, take a 2 ft (60 cm) long, thick stick and place it on the ground. Put the bird's head underneath the stick, place your feet on the stick either side of the head, and pull the bird up sharply by the legs—the bird's head will come off.

POACHER'S METHOD
The poacher's way to quickly get the meat from a dead pheasant that hasn't been decapitated is to place it face down on the ground with its head toward you, put a foot on each wing, take hold of the legs, and pull up sharply. This motion will tear the legs and breast meat away from the rest of the bird. Give it a sharp flick to detach the guts. In a survival situation, this is a good way to get meat off the bird if you don't have a knife or razor.

PLUCKING A BIRD

It's not essential, but scalding a bird in a bucket of hot (not boiling) water for a couple of minutes usually helps to loosen the feathers; the exceptions are waterfowl and seabirds—their feathers will tighten instead. Be careful not to overscald the bird or the skin will start to cook.

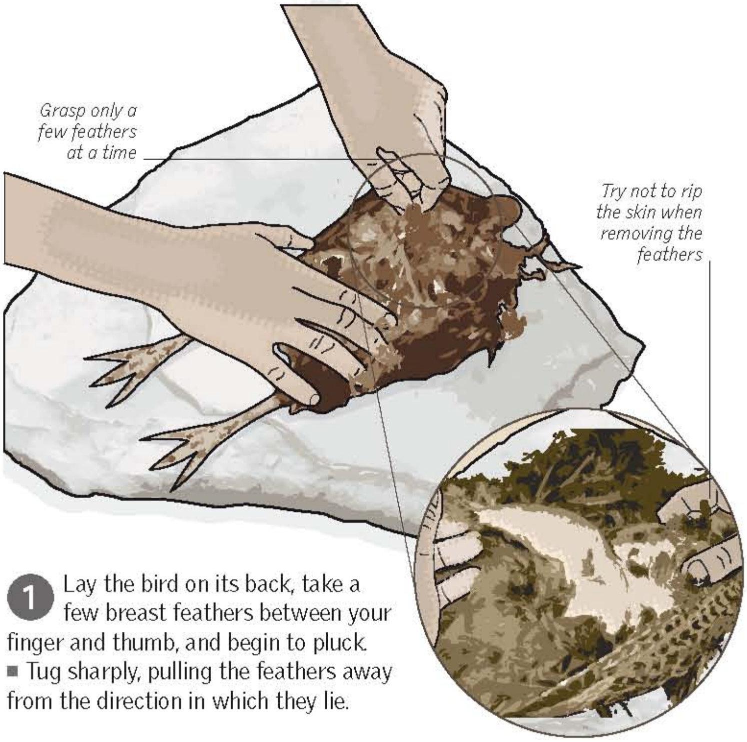
Grasp only a few feathers at a time
Try not to rip the skin when removing the feathers

1. Lay the bird on its back, take a few breast feathers between your finger and thumb, and begin to pluck.
   - Tug sharply, pulling the feathers away from the direction in which they lie.

2. Don't try to pluck too many feathers at once or you'll tear the skin. Work your way steadily around the front and back until you've plucked the whole body.
   - Keep the feathers (unless the bird is a scavenger, see box, above) to use for tinder, insulation, or fishing lures.

OPTIONAL EXTRAS

Keep the crop (esophagus) contents of a gamebird—you can use the seeds and berries as bait to catch other birds. If the bird is female and has eggs in the oviduct, keep them and eat them. You should also eat the heart and the liver if they are in good condition; discard them if they look old or spotty, or show signs of parasites.

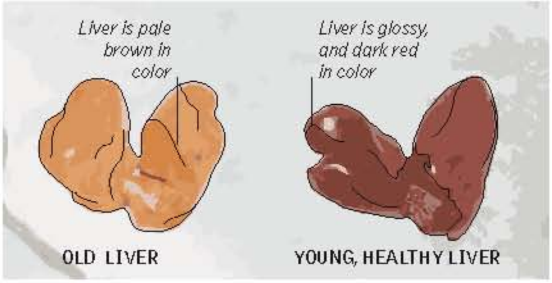
Liver is pale brown in color
Liver is glossy, and dark red in color
OLD LIVER
YOUNG, HEALTHY LIVER

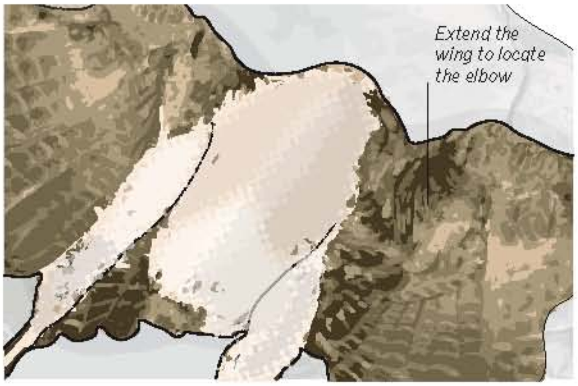
Extend the wing to locate the elbow

3 Pluck the legs in the same way, then turn to the wings. Locate the elbows and cut off the lower wings at this joint. Pluck the upper wings.

- If the bird's head is still attached, cut it off as close as possible to the body.

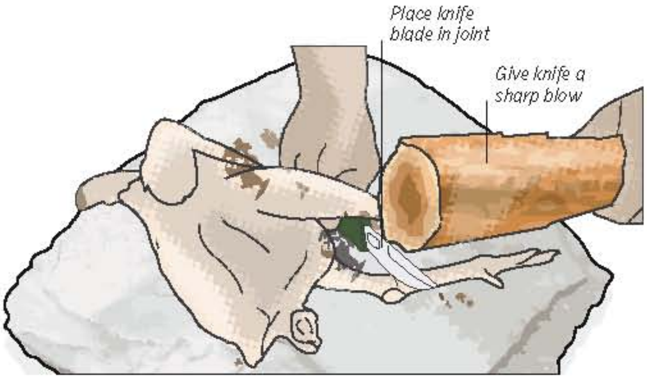
Place knife blade in joint
Give knife a sharp blow

4 To remove the feet, find the ankle, insert the blade of a sharp knife into the joint, and press down hard.

- If you're dealing with a large bird, such as a turkey, hit the back of the knife with a thick stick for extra force.

DRAWING METHODS

The usual way to remove the offal from a bird (known as "drawing") is to take the crop (esophagus) out of the neck end and the rest of the offal out of the rear end. The survival method is to make a cut from the throat to the tail with a sharp knife, reach in, and pull out all the offal (see below). Should you find yourself without a knife, you can still butcher a bird by pulling the skin apart with your fingers and working your way up the breast to the neck. The offal can be removed by ripping open the skin over the belly.

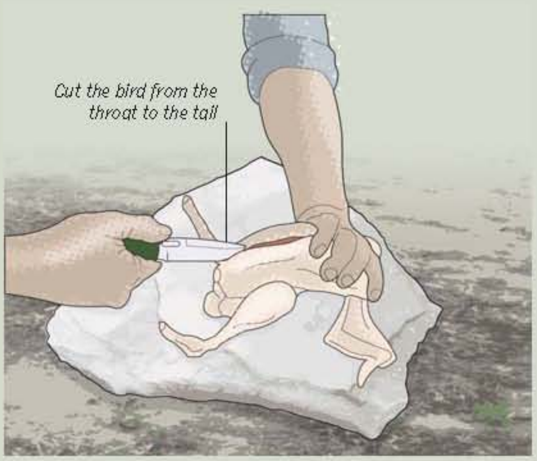
Cut the bird from the throat to the tail

1 Hold the bird in one hand, breast upward and with its rear end toward you. Insert the point of a sharp knife into the throat.

- Make a single incision down toward the tail, being careful to not pierce the guts.

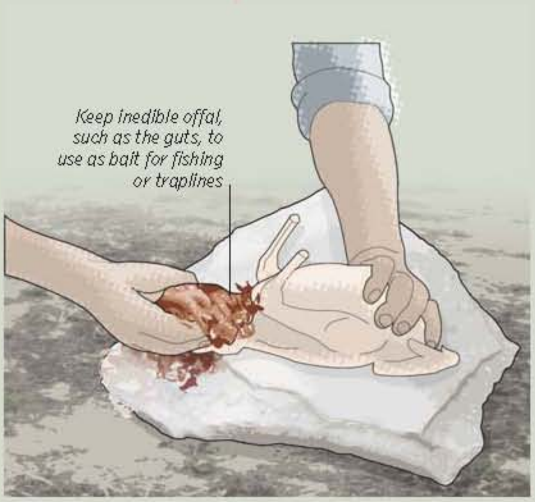
Keep inedible offal, such as the guts, to use as bait for fishing or traplines

2 Reach into the gut cavity and pull out the offal, taking care not to break any eggs. Keep the liver and heart if healthy (see above left).

- Wash the bird clean with cold water and wash your hands thoroughly before further handling.

IN AN EMERGENCY

IN AN EMERGENCY

You're only a survivor when you have been rescued. This means that you must be able to either get yourself out of the predicament you are in (self-rescue) or be rescued from that situation by others. Sometimes you will be able to choose whether to undertake self-rescue; at other times the decision will be out of your hands. There may be several reasons why you cannot achieve self-rescue: you could be utterly lost; local conditions such as flooding or bad weather could trap you; or injury to yourself or a group member could make movement impossible. In this situation, the onus is on you to attract attention. You must be able to make contact with rescuers, using location aids you have with you or those you can improvise. Never delay any form of rescue because of the embarrassment factor—the only important factor is the outcome, and I would always rather be embarrassed and alive than eventually found dead!

Preparation is key—informing people of your intentions and timeframes will at least have someone wondering why you are not back yet. Equally, taking the best

In this chapter YOU WILL DISCOVER...

- how to **recognize** the **enemies** of **survival**...
- why **LEOs** are so **important**...
- when to **set fire** to a **log cabin**...
- that **tinsel** isn't just for **Christmas**...
- what to **do** if a **hippo yawns**...
- that you **should go** with the **flow** in an **avalanche**...
- how to **inflate** your **pants**...

location aids for your environment and knowing how best to use them will increase the chances of you being found.

In many survival situations, a major decision will be whether you remain where you are or move to a location that offers a better chance of survival, rescue, or both. There are numerous factors that will dictate your best option but, in general, it's always best to stay where you are. It's all too easy to make a rash decision and attempt to walk out of a situation only to put yourself in even greater danger.

A location aid can make the difference between life and death. A comparatively recent invention is the multi-tasking Skystreme device.

SKYSTREME is a silver, inflatable foil kite, which weighs just 1½ oz (43 g) and packs down to a small and convenient size. It can perform four tasks in a survival situation:

Location aid The kite is orally inflated and its wedge shape means it can lift off from the ground unaided in 4 mph (6 kph) winds.

- The 165 ft (50 m) of line attached to the kite (this much cordage is invaluable in itself) allows it to rise above tree levels and reflect sunlight. It can be seen by the naked eye at a distance of 2 miles (3 km).
- At night, you can hang a small flashlight or lightstick underneath the kite to aid visibility.
- The metallic surface reflects radar and can be detected by aircraft at a distance of 10 miles (17 km). A UK Royal Marine on a polar crossing flew his Skystreme behind his sled so that support aircraft could find and track him.

Emergency first aid splint The kite can be inflated around a broken or sprained limb.

Thermal vest The kite can be inflated and put inside clothing to act as a body warmer.

Water storage Instead of being inflated with air, the kite can be used to carry and store water.

> "You are not a **survivor** until you have **been rescued**"

ASSESSING YOUR SITUATION

ONCE YOU'RE OUT of immediate danger, assess your situation and plan accordingly. At this initial stage, it's crucial to think clearly—the decisions you make now could mean the difference between life and death. In most cases, remaining where you are is the preferred option, but no two situations are ever the same: the circumstances, environment, conditions, and you—the individual—all have a major impact on what can and cannot be achieved.

THE STRATEGY FOR STAYING ALIVE

In an emergency situation, think of the four priorities of survival: protection, location, water, and food. Your situation will determine which is the most important. In most cases, as long as you're in no further danger from injury or the elements, you should focus your efforts on establishing a safe location and getting yourself rescued. The strategy for staying alive, known by the acronym SURVIVAL (see below), gives you a framework to use and helps you remember what you need to do to remain alive and get rescued.

THE RULE OF THREES

A common understanding among those concerned with survival is that various key time limits have a numerical value linked to the number three. Remembering the "rule of threes" can help you focus your decision-making, particularly if you are injured, at risk of further injury, or in immediate danger from the elements. In most cases:

- **Three seconds** is the psychological reaction time for making a decision.
- **Three minutes** is the length of time your brain can do without oxygen before it suffers irreparable damage.
- **Three hours** is the critical time you can survive unprotected in extreme climates.
- **Three days** is the approximate length of time you can live without water.
- **Three weeks** is the approximate length of time you can live without food.

SIZE UP YOUR SITUATION	USE ALL YOUR SENSES	REMEMBER WHERE YOU ARE	VANQUISH FEAR
▪ First, assess the particulars of your surroundings, physical condition, and equipment. ▪ If you're in a group, you can share tasks and responsibilities, but remember that everyone will be affected by the consequences of your decisions.	▪ Most people react to a true survival situation through either training (they automatically do what they've been trained to do in a particular situation) or by instinct (they automatically do what their mind and body tell them to do).	▪ In any survival situation, it always helps to know where you are so that you can make the best decisions about what to do and where to go next.	▪ Fear and panic can be formidable enemies, so it's imperative that you have the knowledge and training to counteract them and prevent them from making your situation worse.
▪ Your surroundings: every environment has its own idiosyncrasies—hot and dry, hot and wet, cold, exposed, or enclosed. Determine what you need to do in order to adapt to that environment. ▪ Your equipment: assess your equipment and consider how it can best be used in your particular situation. ▪ Your physical condition: remove yourself and others from danger. Check for injuries and administer first aid as necessary. Remember that the trauma and stress of a survival situation may cause you to overlook, or subconsciously ignore, injuries that you yourself may have received.	▪ Whatever has happened, approach your predicament in a calm and rational manner. The situation requires careful thought and planning. ▪ If you act in haste, you may overlook important factors, lose vital equipment, or simply make matters worse. The saying "Undue haste makes waste" is especially true in a survival situation. ▪ Listen to your subconscious survival senses and gut feelings, and learn to act on them when they send you warning signals.	▪ Knowing your precise location can clarify whether rescuers are likely to find you or if you'll have to rescue yourself. You'll get a good idea of the obstacles you may face, whether you should remain where you are, and where best to position your aids to location. ▪ If you have made an Emergency Plan of Action (see pp. 24–25), someone will know your approximate location, and when you're due back. ▪ Channel your efforts into making sure that when rescuers are looking for you, your aids to location are in place.	▪ If uncontrolled, fear and panic can destroy your ability to make intelligent and rational decisions. They can cause you to react to your feelings and imagination rather than to your actual situation and your abilities. Moreover, they can incapacitate you and drain your energy, thus triggering other negative emotions. ▪ If you're in a group, your responses can have a direct effect on others—positive responses are productive and can motivate, while negative responses can undermine confidence and morale.

THE ENEMIES OF SURVIVAL

In a survival situation, there are seven factors, known as "the enemies of survival," that can work against you. In many cases, you can deal with them by knowing what they are and understanding their effects. One way to memorize these is by using a mnemonic such as: "**B**e **P**repared **T**o **F**ace **T**hese **H**ostile **F**actors".

BOREDOM AND LONELINESS

When boredom sets in, you become inactive and lose the ability to deal with your situation effectively, so you need to keep busy. Loneliness makes you overwhelmed by what you need to achieve, leading to a feeling of helplessness.

PAIN

If you're injured, don't ignore the pain. Attend to a minor injury, as it could grow into a major problem that could impair your ability to survive. A positive mental attitude coupled with keeping busy helps to distract the mind from pain.

THIRST

Thirst is not a good indicator of the body's need for water. Your body can be dehydrated before you feel thirsty. Stay ahead of dehydration, rather than have to deal with it. Prioritize your need for water in a way that's relevant to your environment.

FATIGUE

Tiredness leads to mistakes that, at best, cause frustration and, at worst, may result in injury or death. In a survival situation, it's unlikely that you'll replace the energy you use effectively, so everything becomes harder to achieve. Never underestimate the importance of quality rest to your physical and mental wellbeing.

TEMPERATURE

Temperature is a major factor in any survival situation and it will be affected by wind, rain, and humidity. You should dress to suit the environment you're in and be aware of the signs and symptoms of temperature-related injuries, such as dehydration, hypothermia, heat stress, and heat stroke (see pp. 272–73).

HUNGER

In a short-term survival situation (one to five days), procuring food is not a high priority. You can offset your reduced energy and stamina levels by drinking water and pacing yourself so you work within your limits. However, take every opportunity to procure food without expending energy.

FEAR

Fear is one of our body's greatest survival tools, as it can stimulate you, so that you're ready to act—however, it can also debilitate. Fear is good as long as you have control over it, and the key to controlling fear in a survival situation is knowledge.

	IMPROVISE	**V**ALUE LIVING AND LIFE	**A**CT LIKE THE LOCALS	**L**EARN BASIC SKILLS
	▪ The true skill of a survivor is to understand what's required and improvise solutions to particular problems. Do you have the skills and knowledge to keep yourself alive and in a condition to be proactive in your own rescue?	▪ Some people without training and equipment have survived the most horrendous situations. In many cases, this was simply because they had the will to live and refused to give up!	▪ Whatever environment you're trying to survive in, you can be sure that the local or indigenous people and the local wildlife have developed ways of adapting to it in order to survive.	▪ Learning basic skills increases your chances of survival. Without training your prospects of survival are down to luck, which is never the best place to start. There is a saying: "Luck favors those who are best prepared."
	▪ You may start out with all the right equipment, but it may get lost or broken, or simply wear out. Your ability to improvise may mean the difference between your continued struggle to survive in relative comfort, or absolute misery. ▪ Think laterally, like the climber who was stranded on the side of a mountain with no aids to location. He used the flash on his camera to signal his location to a rescue helicopter. Improvise and overcome!	▪ The stories of prisoners of war often reveal what kept them alive: religious beliefs, thoughts of family and friends, or a determination not to let the enemy win. While these alone may not always be enough, they're certainly a key factor in any survival situation. ▪ It helps to bear in mind that hardship means different things to different people and in different cultures. ▪ Survival is about dealing with hardships and having the will to live. If the will to live is not there, then just having the knowledge and equipment may not be enough.	▪ Look at how the local people dress and act: in hot countries they leave manual work until the coolest parts of the day and work in a slow and deliberate manner to reduce sweating and therefore conserve water. ▪ If you're in a desert, for example, watch where animals go to find shade: they're mostly nocturnal and spend the day underground. Learn their tricks to find water, like the darkling beetle in the Namib Desert that drinks water from the fog that condenses on its carapace. ▪ If you're in a jungle, pay attention when the animals go quiet or quickly leave an area—danger is usually around the corner.	▪ Prior preparation is the key to survival: discover what you need to know about the environment you're going to; familiarize yourself with all your equipment; and practice your basic skills until they become second nature. This thorough preparation will help you to combat the fear of the unknown and give you the self-confidence to meet the challenges of any survival situation you may be in.

ATTRACTING RESCUERS

TO BE A SURVIVOR you need to rescue yourself or be rescued by others. If you can't rescue yourself—perhaps you're injured, completely lost, or trapped by bad weather—you must attract attention with location aids that you've either brought with you or have improvised.

AIDS TO LOCATION

Location aids can save your life, so make sure you know how to use them effectively. A helicopter on a search pattern may make only one pass over an area before moving on, so you'll only have a few minutes to act decisively.

COMMUNICATING WITH SATELLITES

When activated, a Personal Locator Beacon (PLB) transmits a radio distress signal to two complementary satellite systems called LEOSAR and GEOSAR. Together, these form the "COSPAS-SARSAT" system. The signal is then relayed to a rescue coordination center closest to the beacon's location. PLBs mainly use 406Mhz; the military also use 243.5Mhz and 282.8Mhz.

ATTRACTING ATTENTION

When you're in a survival situation, you need to employ your location aids in the best way possible. There are three main principles for attracting attention:

- **Attract:** pick a place that maximizes your chances of attracting attention using your location aids, such as open or high ground. Place your aids carefully so they can be detected from as wide an area as possible. The more obvious the signal the better.
- **Hold:** you need to hold the attention of rescuers by maintaining the signal until they indicate they've seen you. Try to send some critical information (the type of assistance you require or the number of survivors and their condition, for example) via a May Day, or Help message (see pp. 237–41).
- **Direct:** regardless of which location aids you use, once rescuers have been alerted, do all you can to direct them to your present location. If you have voice communication, direct them to your position. If you leave notes before moving on from places, make sure they contain dates and precise details of your intentions.

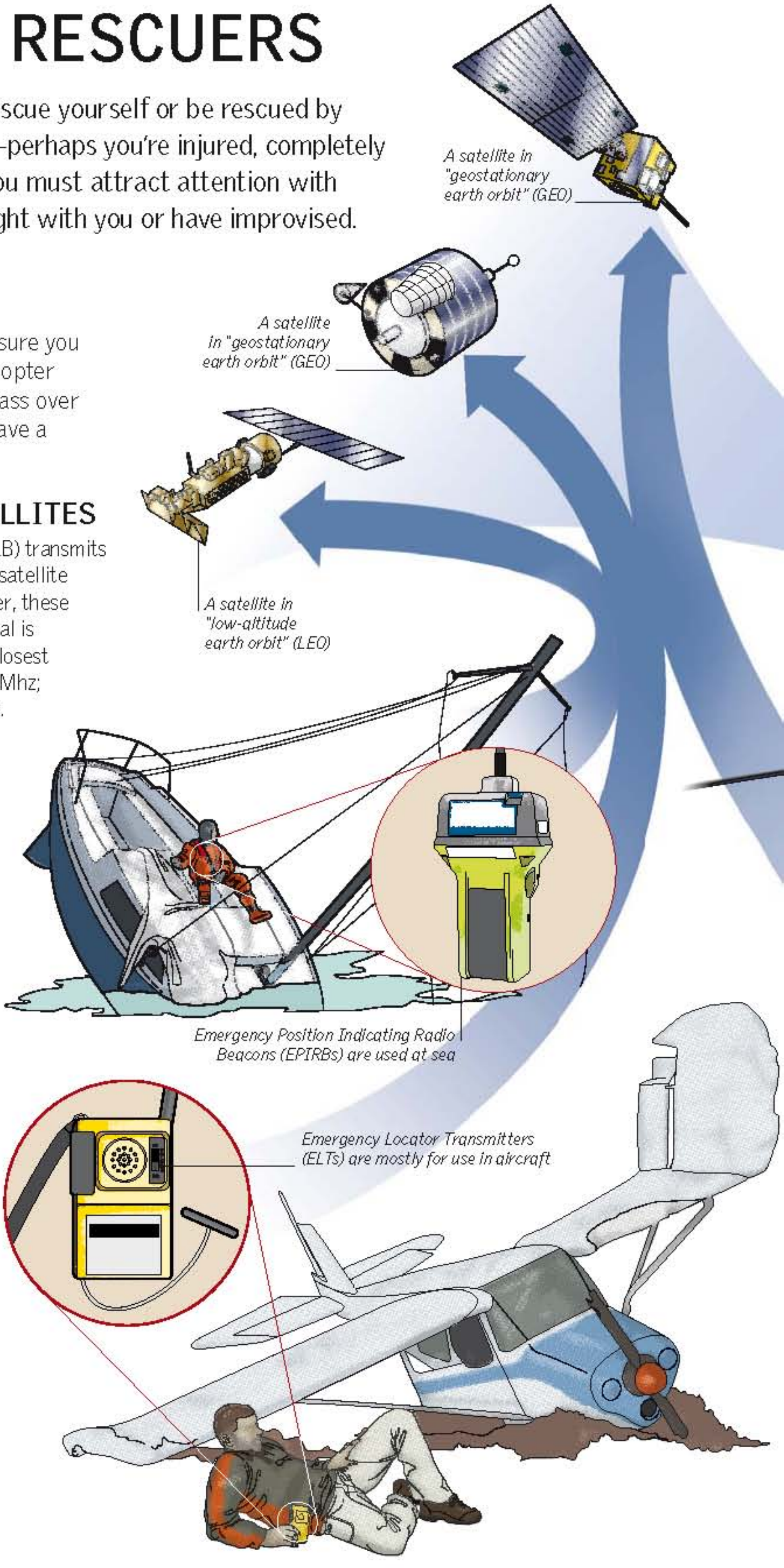

A satellite in "geostationary earth orbit" (GEO)

A satellite in "geostationary earth orbit" (GEO)

A satellite in "low-altitude earth orbit" (LEO)

Emergency Position Indicating Radio Beacons (EPIRBs) are used at sea

Emergency Locator Transmitters (ELTs) are mostly for use in aircraft

SHOULD I STAY OR SHOULD I GO?

You need to decide whether to stay where you are or to move to a location where you have a better chance of rescue or self-rescue. If you move on, remember the following:

- Keep your aids to location at hand. It's no good having your signal mirror at the bottom of your backpack when you may have only seconds to attract a passing vehicle or plane.
- Deploy your aids to location at the end of each day, even though it takes effort and you'll be moving in the morning.
- To show where you've been and where you're heading, leave markers, such as a note left in a visible position and visual clues on the ground or vegetation, to indicate your direction of travel.

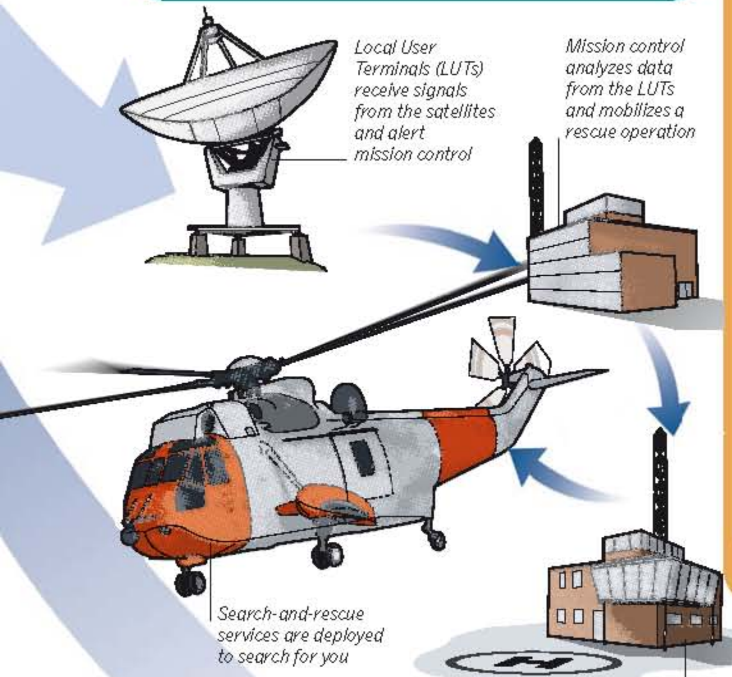

Local User Terminals (LUTs) receive signals from the satellites and alert mission control

Mission control analyzes data from the LUTs and mobilizes a rescue operation

Search-and-rescue services are deployed to search for you

A regional rescue coordination center alerts the search-and-rescue teams

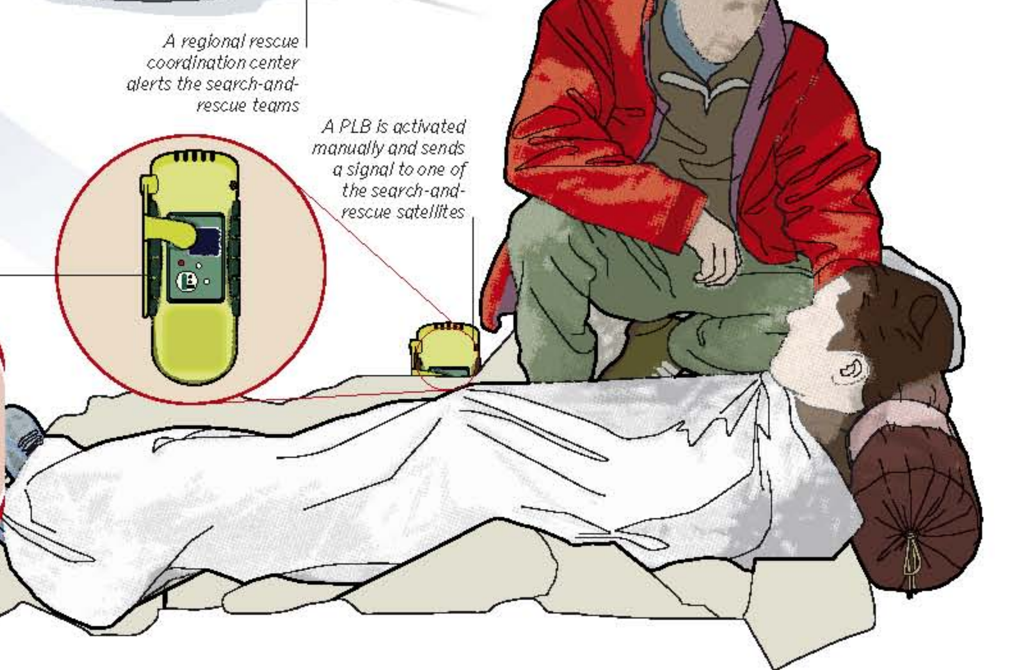

A PLB is activated manually and sends a signal to one of the search-and-rescue satellites

Personal Locator Beacons (PLBs) are for personal use and kept on the body

WARNING!
PLBs are for emergency use only and should always be used responsibly.

MOBILE PHONES AND RADIOS

Always take a cell or satellite phone with you when traveling—they are essential in an emergency. On a sea trip, take a marine VHF radio to contact would-be rescuers.

CELL AND SATELLITE PHONES

Wherever you are in the world, you can rent or buy a cell phone linked to local networks. Choose one with a GPS unit for fixing and racking your position and a camera so you can send pictures of your location and any injuries you've sustained. Alternatively, rent or buy a satellite phone that's connected to the Iridium Satellite Phone System, in which 66 low-earth orbiting (LEO) satellites provide complete coverage of Earth (including oceans, airways, and polar regions).

MARINE VHF RADIOS

All large ships and most motorized small craft are equipped with marine VHF radios. These hand-held units transmit and receive on frequencies between 156 to 174MHz—usually on Channel 16, the international calling and distress channel. Channel 9 can also be used in some places. Transmission power ranges between 1 and 25 watts, giving a maximum range of up to about 60 miles (110 km) between aerials mounted on tall ships and hills, and 5 miles (9 km) between aerials mounted on small boats at sea level. Your VHF radio should be waterproof, able to float, and kept on charge. Follow the instructions whenever transmitting and receiving on VHF.

BUILDING SIGNAL FIRES

A signal fire is a very effective way of attracting the attention of rescuers, but you do need the right materials and building one requires some effort. A well-constructed signal fire located in a good position will generate a large amount of smoke that can be seen from a long way away. Two variations—the dome and the log cabin—are shown here.

TOOLS AND MATERIALS

- Knife and saw
- Poles and long saplings
- Green vegetation
- Tinder, kindling, and fuel
- Lighter or matches

SIGNAL FIRE ESSENTIALS

Whichever signal fire you make, there are several important principles to follow in order to make it as effective as possible.

FIRE FORMATION

If you can, prepare three signal fires in a recognized formation, such as a triangle or in a straight line. If the fires are random, they could be misread as a bush fire or a group of native fires. The distance between each fire should be at least 65 ft (20 m) but needs to be dictated by how quickly you can effectively light each one in turn.

FIRE COMPONENTS

Essentially, each signal fire is composed of a ready-to-light firebase made of dry tinder, kindling, and fuel. When you add the merest spark or flame to the tinder, the firebase immediately turns into a sustainable fire.

- Site the firebase off the ground to prevent damp from reaching it, and to allow enough airflow to help it ignite more effectively.
- Cover the top of the firebase with large amounts of green vegetation and anything else that produces smoke, such as tires. The covering keeps the firebase dry.
- Place a second stack of green vegetation near the firebase and add this as the original vegetation burns out.

AT THE READY

Once you have prepared the signal fire, you need to keep the following nearby, ready for use at a moment's notice:

- Dry tinder in a waterproof container underneath the firebase.
- Hexamine fuel tablets, stove fuel, gas, paper, or birch bark to guarantee the fire gets going again if it starts to dwindle.
- Something to create a spark or flame.
- A witch's broom, made of a cleft stick stuffed with kindling or bark, next to your camp fire, so you can quickly light the broom and transfer it to the signal fire.

MAKING A DOME SIGNAL FIRE

Prepare a large fire on a raised platform under a dome-shaped structure made from bent saplings. Fueled by air from below, the fire creates plumes of smoke from the green vegetation. If you have no saplings, use poles to a make a teepee shape.

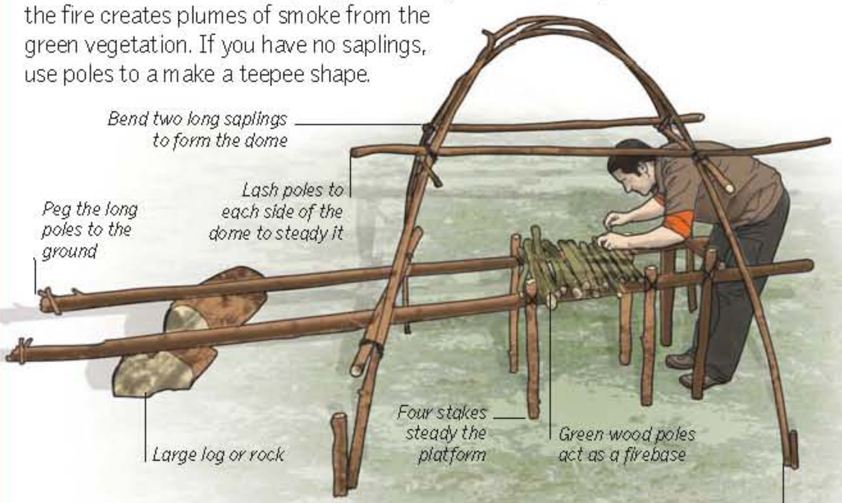
Bend two long saplings to form the dome
Lash poles to each side of the dome to steady it
Peg the long poles to the ground
Large log or rock
Four stakes steady the platform
Green wood poles act as a firebase
Stick the ends of the saplings firmly into the ground

1 Lay two long poles parallel to each other. Prop them up over a log or rock at one end, and tie them to four stakes at the other. Lay lengths of green wood side by side on the long poles under the dome, to act as a firebase.

- Bend two long saplings at 90 degrees to each other to form a dome.

4 Layer green vegetation on top of the dome to form a roof over the fire platform.

- Keep additional green vegetation nearby to add to the fire when it's alight.

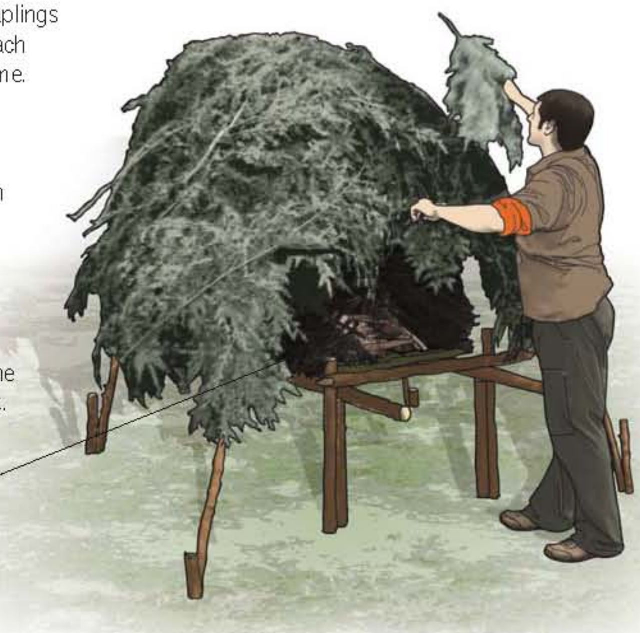
Leave an access point for lighting the fire, but keep it closed so the tinder, kindling, and fuel stays dry

MAKING A LOG CABIN SIGNAL FIRE

This signal fire is known as the "log cabin" because of the way the fuel wood is stacked.

- The framework for the fire is a stack of green wood, made up of pairs of poles arranged in alternating layers set at 90 degrees to each other.
- Place an additional supply of green vegetation nearby to be used as required, but not so close that it catches fire by accident.

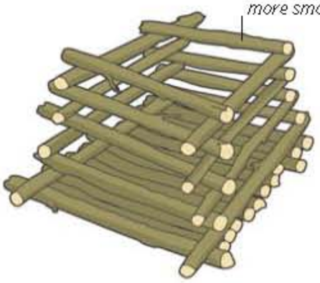

Green wood poles produce more smoke than dead wood

① Lay a platform of green wood on the ground and build a firebase of dry tinder, kindling, and fuel on top.

- Build a "log cabin" over the firebase using green wood.

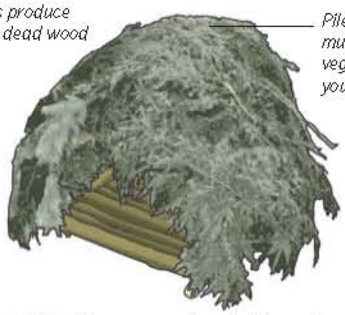

Pile on as much green vegetation as you can find

② Heap green vegetation over the platform to generate thick plumes of smoke when the fire is lit.

- Leave an access point at the front to allow you to light the firebase.

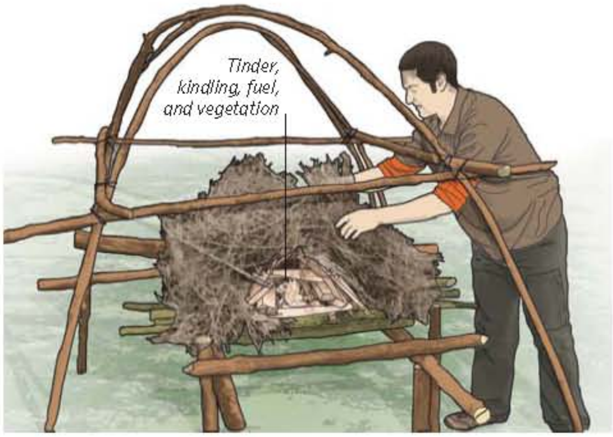

Tinder, kindling, fuel, and vegetation

② Prepare a fire (see pp. 124–25) on the platform and load it with green vegetation.

- Lay the vegetation close enough to the fuel below to catch easily when the fire is lit—but don't smother it.

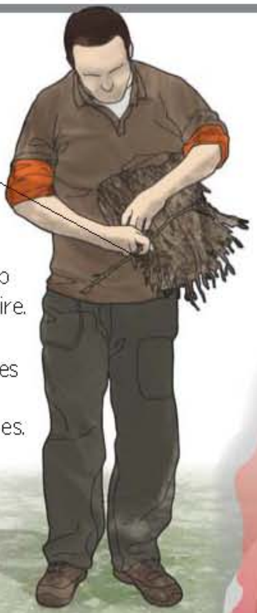

Stab a thin branch or vine into the bundle, wrap it around a few times and tuck it in under itself

③ Prepare some kindling bundles which will help turn the initial flame into a fire.

- Make the bundles by breaking small, dead branches (as thick as a small finger) from the lower trunks of trees.
- Fold them into a small bundle held together with a thin branch or vine.

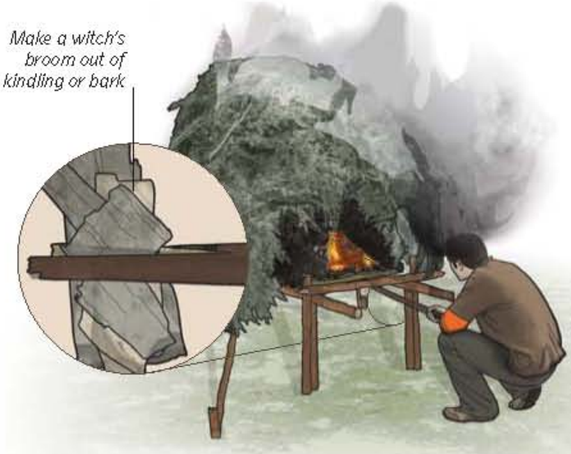

Make a witch's broom out of kindling or bark

⑤ Light a witch's broom (see panel, left) from your camp fire, take it to the dome, and light the fire.

- Use kindling bundles (see Step 3) to fuel the fire.

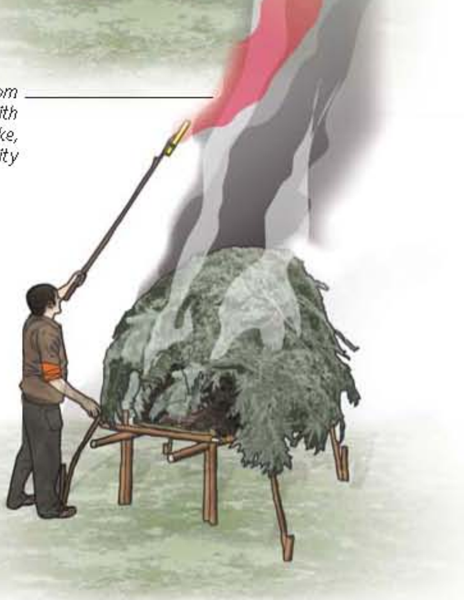

Orange smoke from the flare mingles with the fire smoke, increasing its visibility

⑥ When smoke billows upward, light a preprepared signal flare (see p. 241), if you have one, taped or tied to a long pole.

- Use the pole to position the flare as high into the smoke as you can.

OTHER RESCUE SIGNALS

As well as signal fires, there are other visual devices for attracting potential rescuers. Some are unusual and others, such as a whistle, quite obvious. Whatever device you use, you have to try to hold the rescuers' attention by persisting with your signal.

LIGHT WINDMILL

You can create a highly visible, illuminated "windmill" effect by whirling a chemical lightstick (cyalume) in front of you attached to the end of a cord 3 ft (1 m) long. Depending on local conditions it can be seen up to 2 miles (3 km) away—or further by an aircraft. Blow your whistle at the same time, using the International Distress Signal of six blasts over one minute, then one minute's silence. The reply is three short blasts.

CREATING A STROBE EFFECT

LED strobe lights are compact, robust, waterproof, and are sometimes incorporated within standard flashlights. They are very powerful and can be seen from a long way away. Many strobes can be set to produce a variety of sequenced flashes, including flashing SOS in Morse Code. Even if you don't have a dedicated strobe light, you can still switch your regular flashlight on and off repeatedly to attract attention.

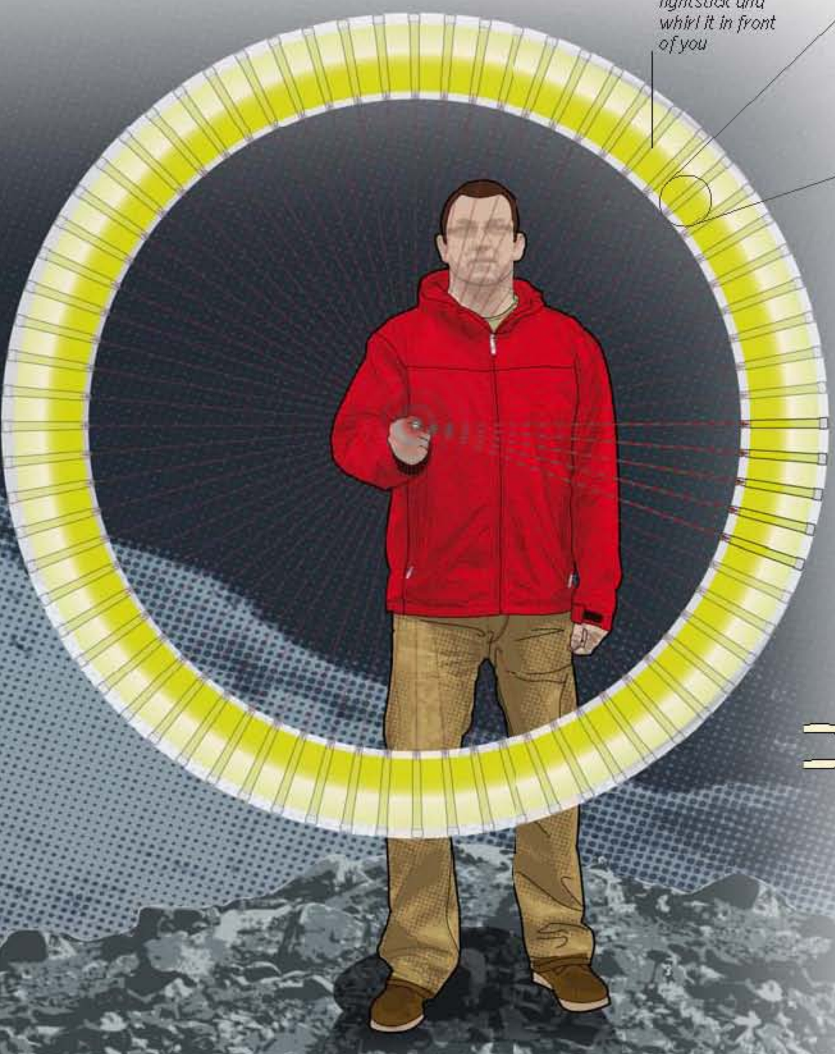

Activate the lightstick and whirl it in front of you

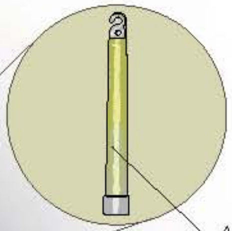

A lightstick is filled with a luminous substance known as cyalume

TINSEL TREE

The signaling device known as a "tinsel tree" is best used when you're remaining in one place for a while. It's effective only during daylight hours, so you'll need to find a suitable and easily visible tree or bush in a sunny position.

- Cut a silver survival blanket, a roll of aluminum foil, or any reflective material into strips.
- Attach the strips to the tree so that they move in the breeze, catching the sunlight and glinting like small mirrors.

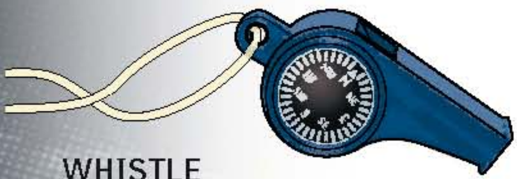

WHISTLE

Take a whistle with you when you venture out into the wilderness because sound, especially in remote areas where there's little or no noise pollution, travels incredible distances. A whistle is an item of first-line gear (see pp. 42–43), and should be kept on a cord around your neck.

SIGNALING DISTRESS

Two of the most recognized signs of distress are a little red star shooting up into the sky or a billowing cloud of orange smoke. There are many different types of signal flares and rockets, ranging from simple, handheld devices that fire flares into the sky to specialized kits designed to penetrate thick jungle canopy. Usually found in the emergency survival packs of aircraft and liferafts, flares are also available from specialized shops.

SIGNAL FLARES

One end of a signal flare has an orange-colored smoke signal for daytime use. The other end has a flare for night-time use (but it can be operated during the day, too). Remember the following:

- Follow the instructions on the outside of the device.
- Wear gloves to protect your hands.
- The heat from a flare can damage a liferaft, so keep your flare well clear when you light it.
- Don't discard flares unless you have used both ends.

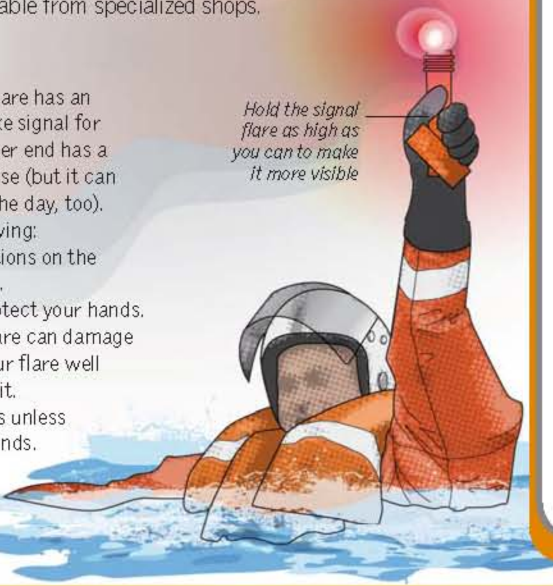

Hold the signal flare as high as you can to make it more visible

GROUND-TO-AIR MARKERS

You can improvise internationally recognized emergency signals on the ground that rescuers in the air can see.

GETTING NOTICED

To make a ground-to-air marker, use anything that contrasts with the ground, such as orange lifejackets, seaweed, clothing, rocks, branches of trees, or soil.

- Whatever you use, make sure the message is big and visible from all directions.
- Check on your markers regularly.
- When appropriate, use one of the emergency codes shown below. SOS or HELP written in big letters will attract attention too.

V	Require assistance
X	Require medical assistance
N	No/Negative
Y	Yes/Affirmative
↑	Proceed this way (arrow points)

SIGNAL MIRROR

A signal mirror, or "heliograph," has a shiny surface that reflects the sun and sends flashes over distances exceeding 30 miles (50 km), depending on the strength of the sun, the size of the mirror, and the clarity of the air. You can even use the light from the moon when it's full. By interrupting the flashes you can send messages in Morse Code.

IMPROVISING A SIGNAL MIRROR

If you don't have a heliograph or any other mirror, try using anything with a shiny, reflective surface, such as a foil food packet, a CD, or the bottom of a beverage can.

WARNING!

Don't dazzle your would-be rescuers by continuing to flash directly at them. Using your signal mirror intermittently is more effective for attracting attention.

1 Polish the base of a can to a shine with a slightly abrasive paste. You can use charcoal and water, toothpaste, or even chocolate.

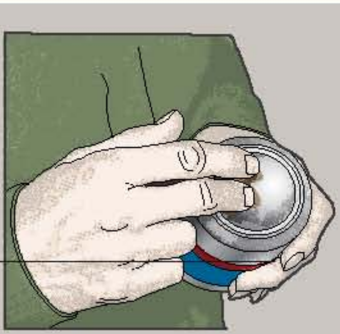

Polish the concave base—it will only take a couple of minutes

2 Hold the can in front of your face with the polished base facing the sun.

- Direct the flash of reflected sunlight on to the palm of your other hand to practice controlling the light.
- Aim the can toward a rescuer and send a signal, either by moving your palm up and down to interrupt the flash or by directing the flash between a "V" made with your thumb and fingers.

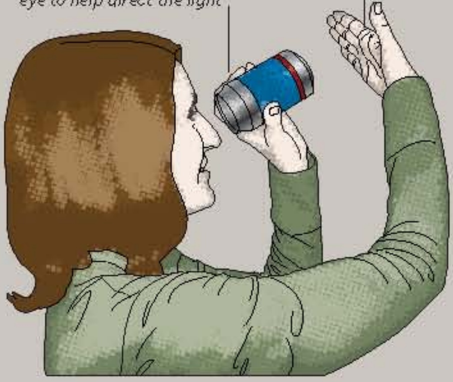

The V-shape helps you direct the light

Hold the can up to your eye to help direct the light

WILD ANIMALS

WALKING IN THE WILDS of Alaska and coming across a fresh deposit of bear feces is a sobering experience. It quickly dawns on you that you're potentially a part of the natural food chain. What do you have to do to protect yourself, and will it be effective or just provoke the animal? Fortunately, the survival instincts of wild animals are such that, with only a few exceptions, most animals, regardless of their size, will avoid confronting you. However, they will defend themselves by attacking you if provoked, cornered, or surprised, particularly when they have young.

FERAL ANIMALS

Wherever you go in the world there are animals that have escaped into the wild from zoos, wildlife parks, and domestic situations. These feral animals include pigs, cats, dogs, mink, and muntjacs. In many cases, they have become particularly aggressive and seem to have lost any fear of confrontation with humans. Try to find out which feral animals are common in the region you're traveling in and avoid them if at all possible.

	BEARS	BIG CATS	HIPPOS	ELEPHANTS	CROCODILIANS
ANIMAL FACTS	▪ Brown bears stand taller then black bears and weigh more. Polar bears are the largest land predator. All bears are powerful, with claws on their strong paws. Brown bears can run faster than humans. ▪ Bears are found in the Northern Hemisphere and parts of South America.	▪ Tigers are the largest cat, with lions a close second. All big cats have sharp teeth and claws. ▪ Other dangerous cats include cougars (also called pumas, mountain lions, and panthers), leopards, and jaguars. ▪ Big cats are found on all continents, except for Australasia.	▪ The hippopotamus is the third-largest animal in Africa after the elephant and white rhino. ▪ Hippos can weigh more than 3 tons. They have long, razor-sharp incisors and tusklike canines.	▪ African elephants have larger ears than Indian elephants and are more aggressive. ▪ Elephants stand up to 13 ft (4 m) tall and weigh up to 6 tons. ▪ They can reach speeds of 25–30 mph (40–48 kph).	▪ Crocodilians include alligators, gharials, and crocodiles. ▪ They can stay submerged for more than an hour at a time, can swim up to 20 mph (32 kph), and run as fast as 11 mph (17 kph) over short distances. ▪ Crocodilians live in many subtropical and tropical parts of the world.
PRECAUTIONS	▪ Look for signs of bears and check with locals about bear activity. ▪ Stow and cook food away from the campsite. Keep unused food and garbage out of reach. ▪ Avoid thickets and streams where bears rest and feed. ▪ Carry a stick, knife, or spray (see opposite).	▪ Avoid coming into contact with big cats. They are most unlikely to attack unless you provoke them or threaten their cubs. ▪ An exception is the cougar, which is responsible for an increasing number of unprovoked attacks in urban areas of North America.	▪ Avoid provoking a hippo. Many encounters are the result of hitting a partially submerged hippo with a canoe or boat. Almost all hippo attacks are fatal. ▪ Stay vigilant on or beside rivers that hippos are known to frequent. ▪ Don't get between a hippo and water.	▪ Keep away from places that elephants frequent, such as watering holes. ▪ If you do come across them, don't get too close, especially if there's a baby elephant nearby. ▪ Look for a safe place to retreat to, such as a vehicle, rocky outcrop, or a tree, before you need it.	▪ Stay away from waters and river banks where crocodilians live. Be vigilant at all times. ▪ If you need to go near the water, watch the area for at least 30 minutes first. ▪ Don't go to the same spot twice, as crocodiles may lie in wait the next time. ▪ Keep a defensive weapon, such as a knife, close by.
WHAT TO DO IF...	▪ Make a noise to let a bear know you're there. ▪ If you see a bear, stay calm. Make yourself look as big as possible by raising your arms. ▪ Walk slowly backward. Don't run. If the bear follows, stop and hold your ground. ▪ If the bear attacks, play dead, or fight back.	▪ If a big cat approaches, stare at its eyes, shout, and make a noise to confuse it. Use a spray if you need to (see opposite). ▪ Don't turn your back or run away. ▪ If a big cat sees you, do not crouch or bend down. An upright human makes less attractive prey than a four-legged animal.	▪ If you encounter a hippo, go back on your tracks and find another route. ▪ If a hippo yawns at you, he's not feeling sleepy—this is a threat. He's showing you his teeth, tusks, and jaws that can snap a canoe in half! Make every effort to escape.	▪ If an elephant squares up to you with flared ears, trumpeting, and kicking the dirt in front of it, then back away. ▪ If an elephant charges, run to your safe place (see above). ▪ As a last resort, play dead and hope the elephant loses interest.	▪ If you meet a crocodilian on land, run away. ▪ If a crocodilian does grab you, it may let go—if it does, run. ▪ If it drags you into the water, fight back. Stab under its throat with a knife, hit or poke its eyes, strike its nostrils hard, or bang the large valve at the back of its throat.

ARMING YOURSELF

If you're traveling in areas where encounters with wild animals are possible, be ready with a weapon such as a knife, large stick, or deterrent spray. Remember, avoidance is always better than confrontation.

PEPPER SPRAY

One type of deterrent spray contains the pepper ingredients capsaicin and capsaicinoids. It shoots a cloud or stream that irritates an animal's eyes. The animal usually withdraws, but take care until the danger has passed.

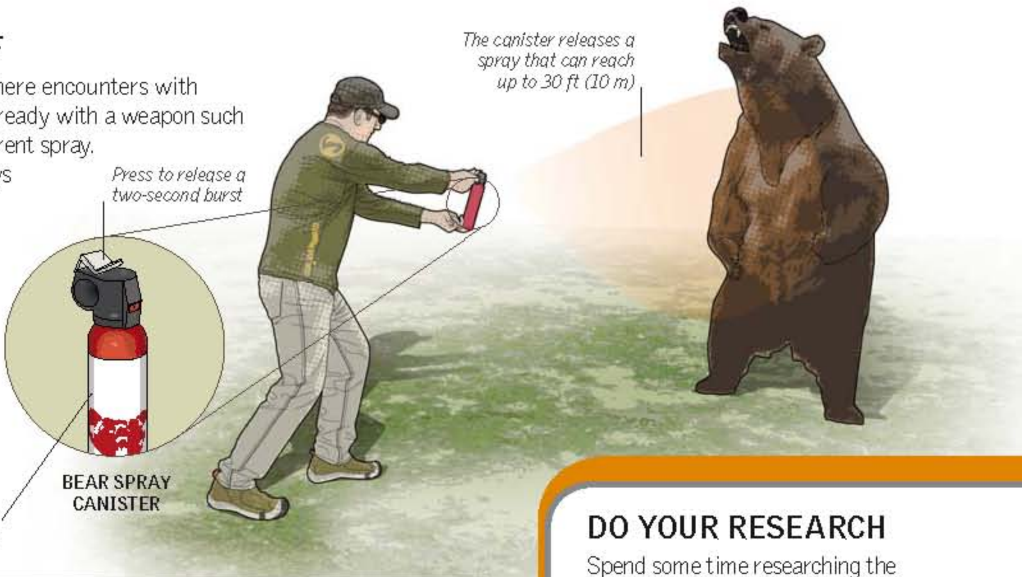

The canister releases a spray that can reach up to 30 ft (10 m)

Press to release a two-second burst

The pepper spray is under pressure in a canister

	SHARKS	SNAKES	AMPHIBIANS
ANIMAL FACTS	▪ There are more than 450 species of shark, but only a few of them pose any danger to humans. These include great white, bull, and tiger sharks. ▪ Most attacks occur in coastal waters in the tropics and subtropics, especially where the water is murky or stirred up by the surf of breaking waves.	▪ Less than 15 percent of the 3000+ types of poisonous snake are regarded as being dangerous to humans. ▪ Depending on the snake, venom either affects the blood, the nervous system, or the heart. ▪ Snakes live in every part of the world except for very cold environments.	▪ Many frogs, toads, and salamanders secrete poison through their skin. The most poisonous frog known is the golden poison-dart frog from South America. ▪ Amphibians live where the climate is sufficiently warm and wet for them to breed.
PRECAUTIONS	▪ Avoid shark-infested waters or waters where sharks are known to feed or frequent. Ask the locals about recent sightings of shark activity. ▪ Carry something, such as a pole or spear, that you can use as a weapon.	▪ Wear long pants and boots, and cover your neck. ▪ As you walk, tap the ground in front of you with a stick to alert snakes to avoid you. ▪ Step onto logs rather than over them. ▪ At night, place your boots upside down on sticks. ▪ Don't put your hands into holes or cracks where snakes might be hiding.	▪ You can touch a poisonous frog or toad with no ill effect. The poison only takes effect if it enters your system via an open wound or your mouth or eyes. If this happens, get emergency help without delay.
WHAT TO DO IF...	▪ If you see a shark, stay calm and move to safety. ▪ If it makes a move toward you, swim away smoothly but keep watching it. ▪ If it rushes at you, hit it on the nose with whatever you have at hand. ▪ If it grabs you, aggressively strike its eyes or gills.	▪ If you come across a snake, try to remain completely still. Most snakes will instinctively move away and are more likely to attack a moving target. ▪ Don't panic. If you have a stick, slowly bring it to a position ready for use. ▪ If a snake attacks, hit it hard on the head.	▪ If contact with a poisonous amphibian is unavoidable, immediately wash the affected area. Keep the water you use away from open cuts or abrasions. ▪ Thoroughly wash your hands and don't put your fingers in your mouth or rub your eyes.

DO YOUR RESEARCH

Spend some time researching the area you'll be traveling in and find out which animals present the greatest dangers. Discover as much as you can about their habits and how best to avoid coming into contact with them.

SPOT THE SIGNS

Understand the habits of these dangerous creatures: when do they forage, sleep, and drink? What are their habits?

- Learn to recognize the natural signs that indicate their presence: tracks or prints; feces/droppings; territory markings, such as flattened vegetation where the animal rubs its scent on the ground; telltale signs, such as where bears have clawed the bark from trees.
- Be aware of other animals disappearing quickly from a watering hole, or a noisy jungle suddenly going quiet.
- Ask local people if there have been any reports of animals attacking people.
- Find out the best ways to ward off an attack or defend yourself should you have no option (see left). Adopting a certain stance or submissive posture, or avoiding or maintaining eye contact, could make the difference between safety or danger, life or death.

TRUE-LIFE ACCOUNT

EXTREME SURVIVAL—IN COLD CONDITIONS

USEFUL EQUIPMENT

- Waterproof, layered clothing
- Trekking pole or ice ax
- Collapsible snow shovel
- Light, whistle, avalanche beacon
- Tarpaulin
- Gore-Tex® gloves and gaitors
- Map, compass, GPS
- Survival tin, bushcraft knife
- Cell/satellite phone
- Poncho/bivy sack

34-YEAR-OLD ICE HOCKEY PLAYER ERIC LEMARQUE became lost in the snowy wilderness of the Sierra Nevada mountains in California after snowboarding at Mammoth Mountain ski resort. Despite being under-equipped and ill-prepared, his improvisation helped him to survive for seven days.

LeMarque was snowboarding alone late on Friday, February 6, 2004, setting off down an unmarked run of virgin powder snow just as the lifts were shutting for the day. Coming to a stop on a flat section, and with visibility reduced to just 10 ft (3 m), he realized he had lost the trail. Wearing uninsulated ski trousers and jacket, and with just an MP3 player, a cell phone with a dead battery, and wet matches, he was ill-prepared for survival, so he began to search for the trail.

LeMarque chose the wrong direction and walked away from the ski resort. Realizing he faced a night in the open, he used his snowboard to dig a crude trench, lining it with bark scraped from a pine tree for insulation from the snow. He tried to light a fire with pine needles and lint from his clothing, but his matches were too wet. He also ate pine needles and bark.

"WITH VISIBILITY REDUCED TO JUST 10 FT (3 M), HE REALIZED HE HAD LOST THE TRAIL"

Over the next five days, LeMarque walked further into the wilderness, leaving scraps of clothing for rescuers and attempting to signal to passing planes with the blue LCD screen of his MP3 player. Fortunately for LeMarque, rescuers spotted his snowboard tracks and followed it for 24 hours, finding him on February 13. He was barely conscious, dehydrated, hypothermic, and malnourished—having lost 35 lb (16 kg) in body weight—and was suffering from severely frostbitten feet, which later required both legs to be amputated below the knees. But he was alive.

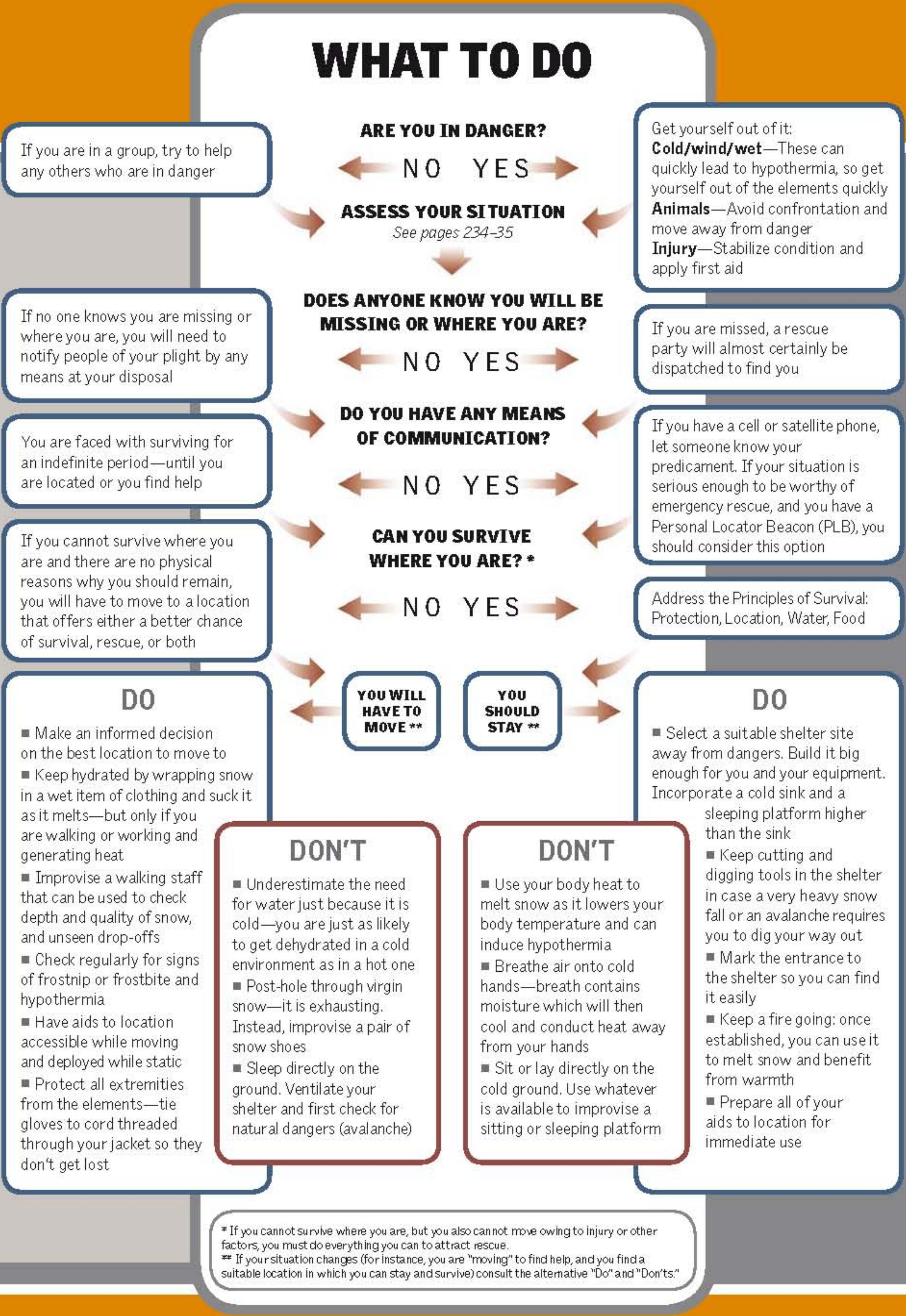

WHAT TO DO

ARE YOU IN DANGER?

NO YES

If you are in a group, try to help any others who are in danger

Get yourself out of it:
Cold/wind/wet—These can quickly lead to hypothermia, so get yourself out of the elements quickly
Animals—Avoid confrontation and move away from danger
Injury—Stabilize condition and apply first aid

ASSESS YOUR SITUATION

See pages 234–35

DOES ANYONE KNOW YOU WILL BE MISSING OR WHERE YOU ARE?

NO YES

If no one knows you are missing or where you are, you will need to notify people of your plight by any means at your disposal

If you are missed, a rescue party will almost certainly be dispatched to find you

DO YOU HAVE ANY MEANS OF COMMUNICATION?

NO YES

You are faced with surviving for an indefinite period—until you are located or you find help

If you have a cell or satellite phone, let someone know your predicament. If your situation is serious enough to be worthy of emergency rescue, and you have a Personal Locator Beacon (PLB), you should consider this option

CAN YOU SURVIVE WHERE YOU ARE? *

NO YES

If you cannot survive where you are and there are no physical reasons why you should remain, you will have to move to a location that offers either a better chance of survival, rescue, or both

Address the Principles of Survival: Protection, Location, Water, Food

YOU WILL HAVE TO MOVE **

DO

- Make an informed decision on the best location to move to
- Keep hydrated by wrapping snow in a wet item of clothing and suck it as it melts—but only if you are walking or working and generating heat
- Improvise a walking staff that can be used to check depth and quality of snow, and unseen drop-offs
- Check regularly for signs of frostnip or frostbite and hypothermia
- Have aids to location accessible while moving and deployed while static
- Protect all extremities from the elements—tie gloves to cord threaded through your jacket so they don't get lost

DON'T

- Underestimate the need for water just because it is cold—you are just as likely to get dehydrated in a cold environment as in a hot one
- Post-hole through virgin snow—it is exhausting. Instead, improvise a pair of snow shoes
- Sleep directly on the ground. Ventilate your shelter and first check for natural dangers (avalanche)

YOU SHOULD STAY **

DO

- Select a suitable shelter site away from dangers. Build it big enough for you and your equipment. Incorporate a cold sink and a sleeping platform higher than the sink
- Keep cutting and digging tools in the shelter in case a very heavy snow fall or an avalanche requires you to dig your way out
- Mark the entrance to the shelter so you can find it easily
- Keep a fire going: once established, you can use it to melt snow and benefit from warmth
- Prepare all of your aids to location for immediate use

DON'T

- Use your body heat to melt snow as it lowers your body temperature and can induce hypothermia
- Breathe air onto cold hands—breath contains moisture which will then cool and conduct heat away from your hands
- Sit or lay directly on the cold ground. Use whatever is available to improvise a sitting or sleeping platform

* If you cannot survive where you are, but you also cannot move owing to injury or other factors, you must do everything you can to attract rescue.
** If your situation changes (for instance, you are "moving" to find help, and you find a suitable location in which you can stay and survive) consult the alternative "Do" and "Don'ts."

ENVIRONMENTAL HAZARDS

SOME PARTS OF THE WORLD have a reputation for sudden or extreme weather conditions, and unpredictable environmental hazards that cause havoc for everyone, not just travelers. Before you set off, make sure that you're aware of any such potential hazards in the areas you're visiting, and be ready for them, just in case.

BEING PREPARED

Many environmental hazards—from avalanches and volcanoes to forest fires, tornadoes, and hurricanes—can suddenly propel you into an emergency situation. While you can never be sure of escaping the worst excesses that Nature can throw at you, you can prepare yourself. A combination of the right equipment, local knowledge, and an awareness of evasive techniques will mean that you shouldn't have to find yourself in a survival situation thinking "if only".

COMMONSENSE ADVICE

No fail-safe guidelines will protect you from every environmental danger and hazard, but there is some commonsense advice that you will do well to heed before you venture into the wilderness:

- Check the prevailing weather conditions of your destination at the time of your journey. Log onto the internet or tune into local radio stations.
- Find out if you're heading into a particularly risky season of weather—for example, monsoons in India or hurricanes in the Caribbean.
- Take an emergency survival kit (see pp. 58–59).
- Learn the relevant survival techniques either before you go or as soon as you arrive at your destination—you may not get a second chance.
- Don't challenge Nature—it rarely loses.

EQUIPPING YOURSELF FOR AN AVALANCHE REGION

Before heading for the mountains, find out whether the prevailing conditions make avalanches more likely (see right).

- Take a collapsible shovel for moving snow and a probe for checking the depth of the snow.
- Keep some survival aids in your pockets in case you lose your backpack.
- Carry an avalanche transceiver—when activated, its signal can be detected by rescue services.

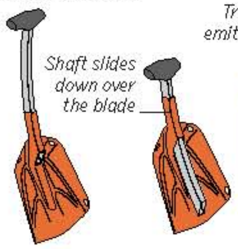
Shaft slides down over the blade

COLLAPSIBLE SHOVEL

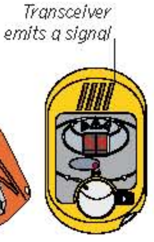

AVALANCHE TRANSCEIVER

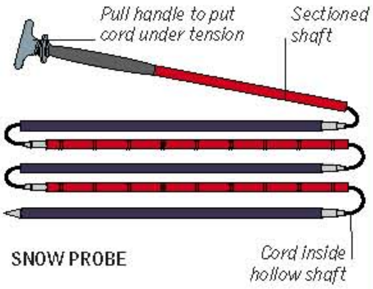

SNOW PROBE

SURVIVING AN AVALANCHE

There's always the danger of an avalanche on slopes that face away from the sun in the middle of winter, when a fresh layer of heavy snow sits on top of a weak layer of snow. An avalanche may be triggered when snow is disturbed by loud noises, an earth tremor, or the movement of skiers or snowboarders. Learn the warning signs, take the right equipment, and practice the emergency steps in case you get caught.

SPOTTING THE WARNING SIGNS

As you cross the snowbound mountains, hills, and valleys, look out for signs that warn of a possible avalanche:

- Convex slopes at an angle between 30 and 45 degrees.
- Slopes without trees or rocks.
- Loose, dry snow that doesn't settle.
- Soft, newly fallen snow that's more than 1 ft (30 cm) deep.
- Snow that sounds hollow.
- Snow that falls as crystals or pellets.
- Snow that falls at more than 1 in (2.5 cm) an hour.

TAKING EVASIVE ACTION

If you see or hear an avalanche, and you think it might be coming your way, take evasive action at once:

- Activate your avalanche transceiver in case you get caught in the snow.
- Try to take cover—for example, under a solid rock overhang if you can see one nearby (see below).
- If you can't find or reach suitable cover, try to sidestep the avalanche by skiing out of the way—at right angles to its potential path.

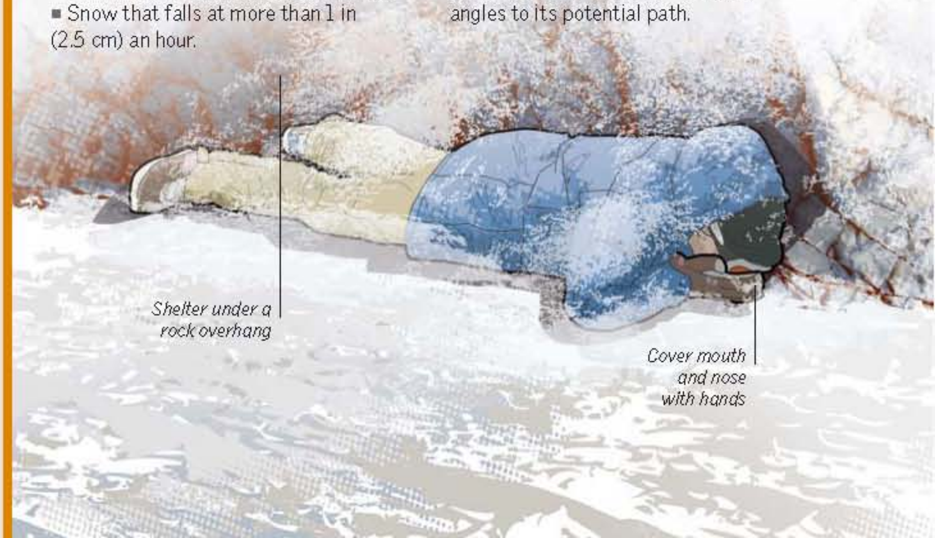
Shelter under a rock overhang
Cover mouth and nose with hands

HAZARDOUS GROUND

All kinds of local conditions can prove treacherous if you don't keep your wits about you. Walking through wetlands can be risky, as the ground can suddenly give way under foot and you can find yourself in a swamp or, worse still, a patch of quicksand.

SURVIVING A SWAMP

Freshwater swamps are found in low-lying inland areas and contain masses of thorny undergrowth, reeds, and grasses, and there may be dangerous animals, such as crocodilians and snakes. The water may be foul and mosquitoes, which can cause malaria, are often present. Moving through a swamp is difficult: you can be on solid ground one moment, and chest-deep in water the next. Where possible, try to build a raft or some kind of flotation aid to help you escape. Use the waterways to navigate your way to open water, where you have the best chance of either being seen by rescuers or finding your way back to civilization.

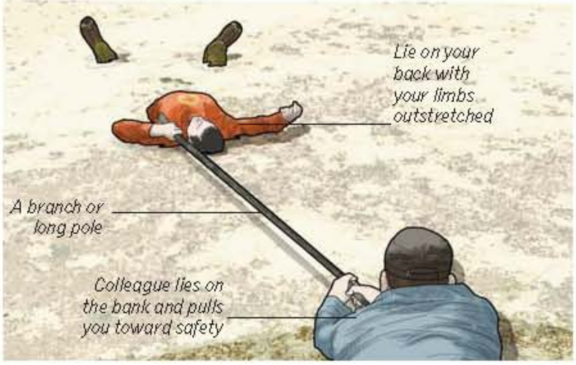
Lie on your back with your limbs outstretched
A branch or long pole
Colleague lies on the bank and pulls you toward safety

ESCAPING FROM QUICKSAND

If you step into quicksand, try to fall onto your back with your limbs outstretched to spread your weight. Try to get to the bank by paddling with your hands. Don't struggle as you will sink faster. If you're with someone, he or she needs to lie on firm ground and pull you to safety, reaching you with the help of a long pole, rope, or branch.

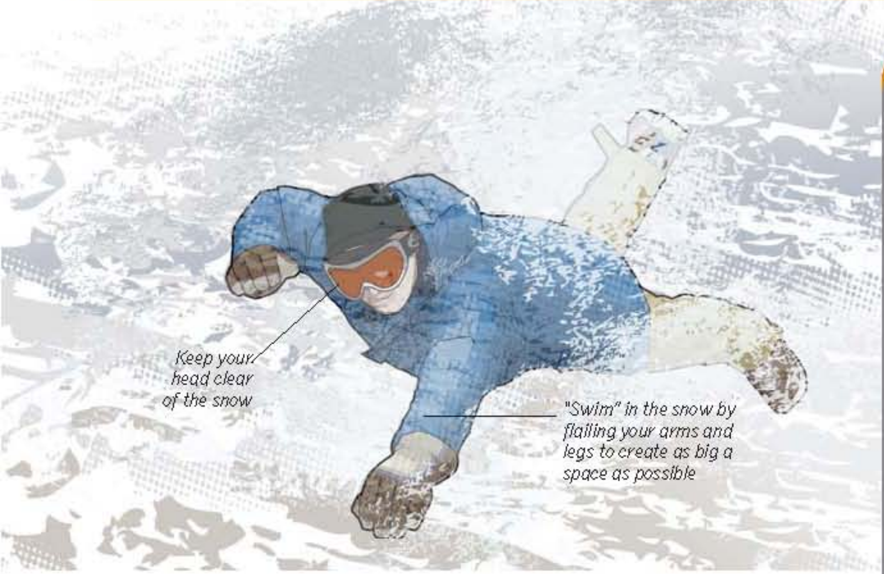
Keep your head clear of the snow
"Swim" in the snow by flailing your arms and legs to create as big a space as possible

ESCAPING FROM AN AVALANCHE

If you can't take evasive action and find yourself below an advancing avalanche with no prospect of escape, try to keep calm and remember the following advice:

- Remove the bindings from your skis and the loops from your ski poles. Put your rucksack on one shoulder, but ditch it rather than risk dislocating your shoulder.
- If you're overtaken by snow, "swim" with the flow (see above) to make a big space around you before the snow compacts and "sets". Use your hands or a collapsible shovel to clear the snow around your head and to create a space for you to breathe, as often people suffocate before they freeze.
- If you don't know which way is up, dribble saliva to find out which way is down—then dig in the opposite direction.
- Get out quickly, as time is of the essence, and shout when you hear potential rescuers to attract their attention.

ESCAPING A VOLCANIC ERUPTION

Find out if there's an active volcano in the region you're visiting. If there is, and its eruption is imminent, leave the area as soon as you can. It will pump enormous quantities of lava from the mantle below the crust and produce huge amounts of ash, toxic gases, debris, and mud. However, if you do get caught in an eruption:

- Be careful when driving to safety, as the ash and mud make roads slippery.
- Take cover to avoid the flying fiery debris, hot gases, and suffocating clouds of ash. Sulfur dioxide in the ash chokes your lungs and, when mixed with rain to form sulfuric acid, burns your skin.
- If you're caught in an ash cloud, cover your face with a wet cloth (or use a mask if you have one). Wash your skin afterward.

SURVIVING AN EARTHQUAKE

Find out if an earthquake is likely to happen in the region you're visiting. If you feel an earthquake coming:

- Get into the open, away from structures or trees that can fall on you. Then lie down.
- Stay in open ground until aftershocks and tremors have stopped altogether.
- If you're in a building, go to the lowest floor, and stay beside a wall or under a sturdy table. Turn off the gas if you have time.
- If you're in a vehicle, stop, but stay inside.

SURVIVING A FOREST FIRE

When vegetation on the ground is tinder dry, the slightest spark can start a fire that gets out of control and sweeps through a forest. A forest fire may be caused by lightning, a piece of glass focusing the sun's rays, a discarded cigarette, or a spark from a camp fire. Forest fires produce intense heat, thick smoke, and toxic gases—and they use up all the available oxygen in the air, making it impossible to breathe.

PREPARATION AND SPOTTING THE WARNING SIGNS

If you're going to a place where forest fires are a hazard, find out from local radio reports whether conditions are making them more likely. If you venture into a forest, make sure you carry a cell phone or some electronic means of signaling, and let others know your route and destination. The following warning signs may help you gain time before a fire is upon you:

- You'll probably smell a fire first, and you'll hear it crackling as it burns before you see it. If you smell a fire and you notice animals growing agitated, a forest fire may be close by.
- The smoke will help you establish how close the fire is. The direction of the smoke tells you which way the wind is blowing.

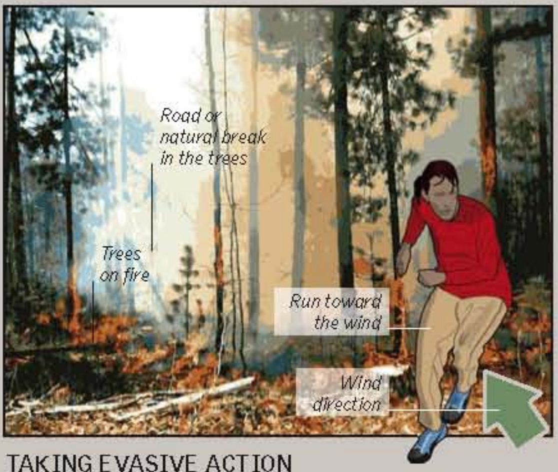

TAKING EVASIVE ACTION

If the wind is blowing toward the fire then move into it quickly. However, if the wind is behind the fire you could be in serious danger because the fire will be moving very fast.

- Try to find a river, lake, road, or natural break in the forest. Stay there until you're rescued or the fire has passed you by.
- Don't go up to high ground as fire is drawn faster uphill.
- Many forest fires spread on a wide front, so avoiding the fire by going around it may be impossible.
- If the fire is upon you, and the wall of flame is fragmented, the best course of action may be to run through the flames: take off man-made clothes (as they will melt onto you), cover as much of your skin as you can, and douse yourself in water if you have any. Take a deep breath, press a damp cloth to your mouth and nose, pick a spot where the wall of flames is thinnest, and run without stopping until you're through.
- If you're in a vehicle, stay inside. Park it as far away from the trees as possible, turn off the engine, close the windows, lie down on the floor, and cover yourself if you can.
- If escape is impossible, try to dig and bury yourself under soil. Dampen all clothing and get as low as possible.

EXTREME WEATHER

High winds and torrential rain are common seasonal occurrences across the world. From tornadoes and whirlwinds to tropical storms such as hurricanes and typhoons, these extreme weather patterns cause enormous damage and threaten many lives.

AVOIDING HURRICANES AND TORNADOES

Weather forecasts can predict approximate directions and paths of impending hurricanes and local emergency services will give advice on the predicted severity. If you're planning a trek in an area known for hurricanes and tornados, check long-range weather forecasts and be prepared to cancel or alter your plans if necessary—if you're caught in the open your chances of survival are extremely limited. If you're going to be in the wilderness during a "storm season" then take a battery- or solar-powered radio, check for weather updates, and be prepared to head back to civilization. If your home is in a tornado or hurricane area, follow the advice of local services, but most importantly be prepared with a plan that will ensure you don't get caught out.

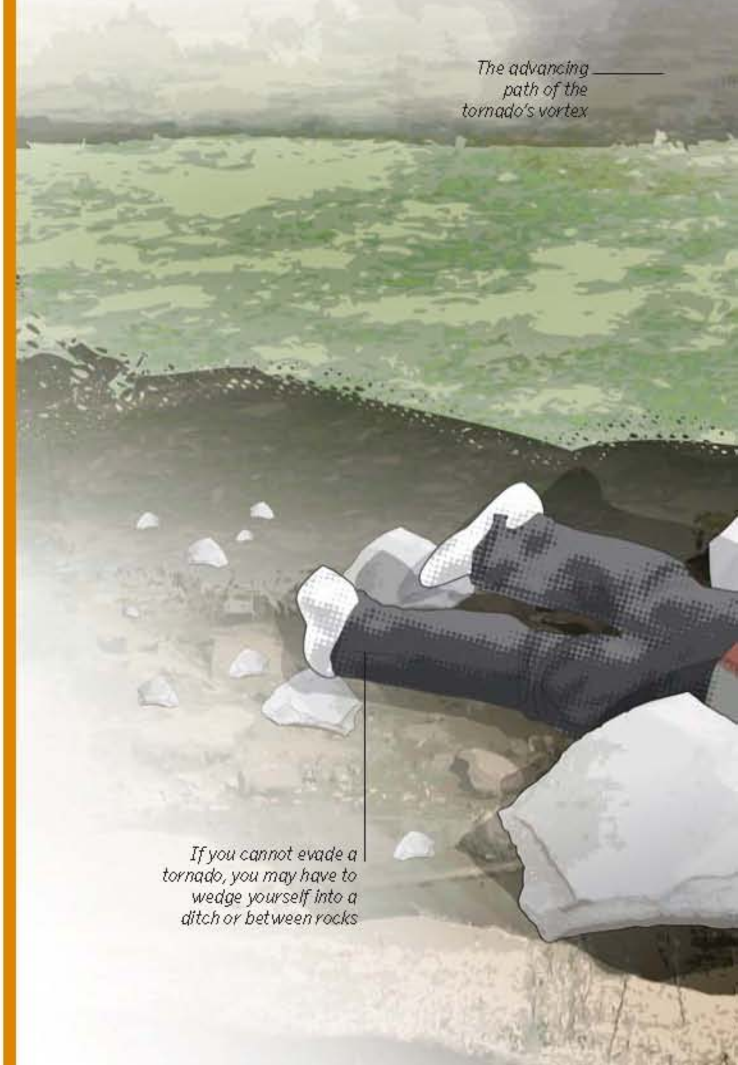
The advancing path of the tornado's vortex
If you cannot evade a tornado, you may have to wedge yourself into a ditch or between rocks

PREPARING TO EVACUATE

The arrival of a hurricane or tornado is a time when you can find yourself in a survival situation on your own home turf. The following advice can help you prepare to evacuate.

- Secure anything outside that could be picked up by winds and cause damage, such as rubbish bins and garden furniture.
- Prepare an evacuation plan (including pets if you have them). Find out where you will evacuate to and make several route plans to the location in case roads are unusable. Fuel your vehicle as soon as a hurricane warning is forecast.
- Put together a hurricane survival pack that contains what you need for a 72-hour period. Include drinking water, a change of clothing, non-perishable food, sleeping bags, radio, flashlights, and contact numbers for family and emergency services.
- Protect your house by securely boarding up all windows and doors. Turn off the water, gas, and electricity.

STAYING PUT

If you decide to stay put, you should still follow the guidelines above. As the storm approaches, move everyone and your supplies to an underground shelter or a room without windows.

- Monitor the radio reports and comply with the advice of the emergency services.
- Don't go outside until the "all-clear" has been given.

TAKING COVER OUTDOORS

If it appears that you are, or could be, in the path of a tornado or hurricane, do everything you can to move out of its way—choose a direction at a right angle to its path.

- If you're driving but can't get out of the way, try to park in an underpass. Leave the vehicle and wedge yourself into a secure position as high as you can under the underpass. If you're in the open, shelter upwind of your abandoned vehicle, so that it does not become a hazard to you.
- If you're out in the open, head for the lowest ground and find a ditch or depression (see below), in which you can shelter, or wedge yourself between rocks.

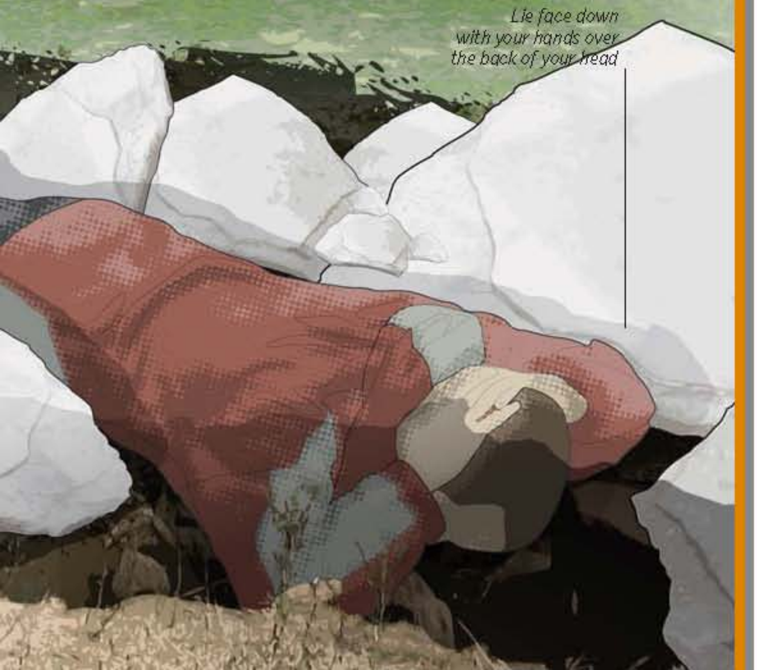
Lie face down with your hands over the back of your head

SURVIVING AN ELECTRICAL STORM

An electrical storm may develop when warm air rises and meets colder air. Electricity sparks between water droplets in the clouds, forming lightning that takes the shortest route to the ground. The following advice will help you avoid being struck by lightning, or minimize its effects if you are:

- Avoid getting caught in open ground—seek shelter but not under a lone tree: when lightning strikes a single tree a tremendous voltage fans out from its base.
- Remain in your vehicle if there is no other cover: it acts as a Faraday Cage (a metallic enclosure that prevents the entry or escape of an electromagnetic field).
- If caught in the open: make yourself as small as possible and limit the amount of area you cover. Do not lay down or stand up, crouch down low with feet together and hands off the ground.

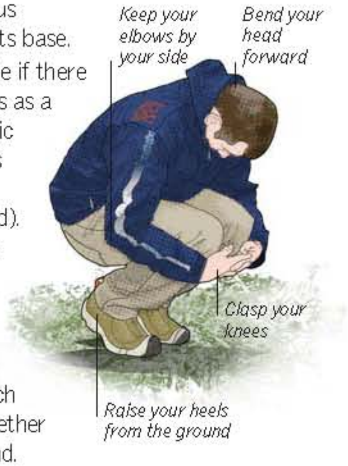
Keep your elbows by your side
Bend your head forward
Clasp your knees
Raise your heels from the ground

SURVIVING A FLASH FLOOD

A sudden deluge of rain doesn't always drain away quickly, instead flowing rampantly over the surface of the land in torrents, and causing flash floods. Soil, animals, vegetation, and even buildings can be quickly washed away. There may be landslides, too, and rivers may break their banks. The following advice can help in the event of a flash flood:

- If you're inside a building, move to the upper floors, taking with you essentials such as bedding, food, and matches. Unless the building is threatened, stay there until the waters have receded or you have been rescued.
- If you're outside, head for higher ground.
- Never walk or drive through a flood.
- Filter and boil water for drinking, as the sources of water around you may have been contaminated. Alternatively, collect rainwater to drink.

ESCAPING A SANDSTORM

If you can see a sandstorm coming, mark your direction of travel before it strikes and find somewhere safe to shelter—for example, behind some rocks. Face away from the direction of the wind and cover yourself as completely as you can, particularly your head, face, and neck.

SURVIVING AT SEA

THE SEA IS ARGUABLY THE HARSHEST environment of all and you could be thrown into a sea-survival situation for any number of reasons: bad weather, fire, mechanical failure, or collision. However, there's plenty of good advice to follow and equipment designed to protect you.

BE PREPARED

If you're venturing onto the open seas, you should plan for the worst and be prepared with knowledge, skills, and equipment.

KNOW YOUR VESSEL

Regardless of the size of your vessel, find out everything you can about what to do in the event of a major problem that requires you to abandon ship and take to the water. The points below are generic and may not apply to all situations or types of vessel.

LARGE PASSENGER VESSELS

If you're a passenger on a large vessel, such as a ferry, liner, or cruise ship, make sure you learn the safety procedures.

- Attend the "Abandon Ship" drills.
- Find out which emergency alarms indicate fire, collision, and abandon ship.
- Find out where the lifejackets are stowed, and how to put them on and operate them.
- Learn the escape routes.
- Locate the Emergency Lifeboat Stations and find out what your responsibilities are.
- If you have children or people with special needs with you, make sure you have a system for getting them on deck and providing them with suitable survival equipment.

SMALL VESSELS

These craft include yachts, small boats, canoes, and kayaks, so crew members need to agree on an emergency plan of action and delegate responsibilities relevant to the situation and the skills of each person.

- Know where the emergency equipment is stowed and how it's operated.
- If your vessel has an EPIRB (see p. 236), make sure everyone knows where it is and how it's operated. Those activated automatically are released hydrostatically from a bracket at a water depth of 3–10 ft (1–3 m). The buoyant EPIRB then floats to the surface and begins transmitting.
- Make sure your EPIRB has been registered so that, once the signal is detected, the rescue services know who it belongs to.
- Keep a grab bag to hand (see box, right).

SURVIVAL SUIT AND LIFEJACKET

A survival suit is designed to keep you warm and dry in extreme conditions and rough seas, and a lifejacket will keep you afloat with your head out of the water, even if you're unconscious.

PROTECTION

Survival suits and lifejackets are equipped with various protection features to help you survive in the sea.

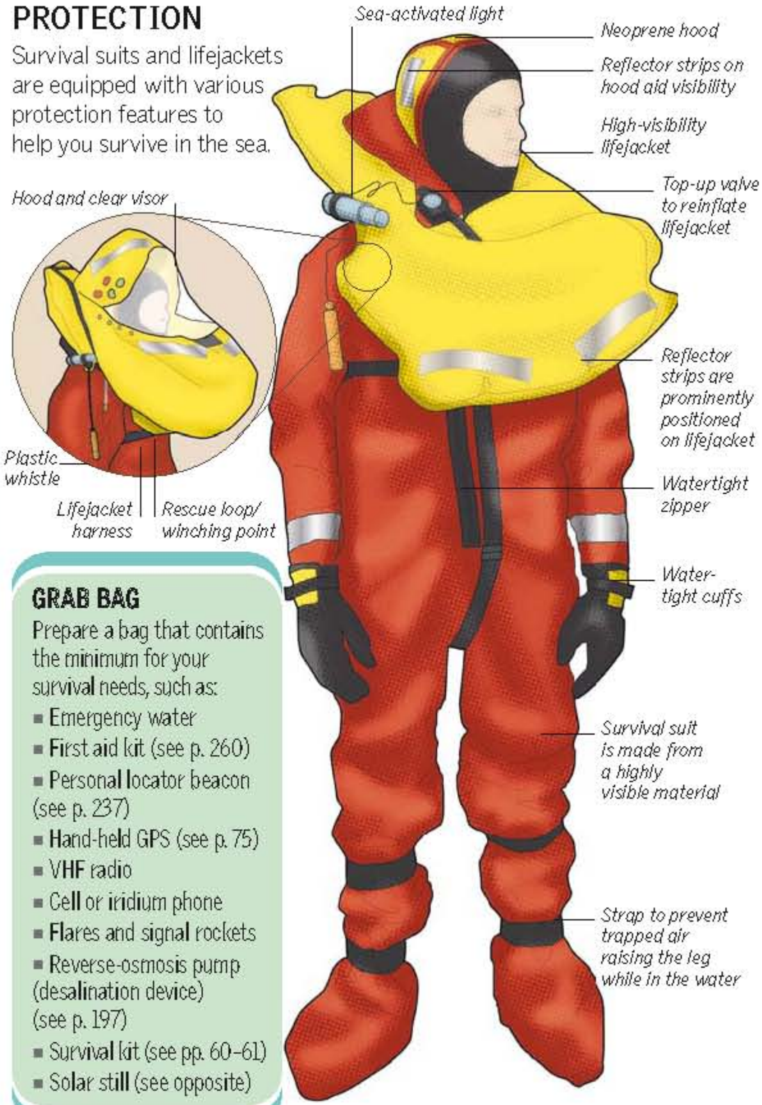
Sea-activated light
Neoprene hood
Reflector strips on hood aid visibility
High-visibility lifejacket
Top-up valve to reinflate lifejacket
Hood and clear visor
Reflector strips are prominently positioned on lifejacket
Watertight zipper
Plastic whistle
Lifejacket harness
Rescue loop/ winching point
Water-tight cuffs
Survival suit is made from a highly visible material
Strap to prevent trapped air raising the leg while in the water

GRAB BAG

Prepare a bag that contains the minimum for your survival needs, such as:

- Emergency water
- First aid kit (see p. 260)
- Personal locator beacon (see p. 237)
- Hand-held GPS (see p. 75)
- VHF radio
- Cell or iridium phone
- Flares and signal rockets
- Reverse-osmosis pump (desalination device) (see p. 197)
- Survival kit (see pp. 60–61)
- Solar still (see opposite)

TYPES OF LIFERAFT

Single-seater and multi-seater liferafts contain similar design features and survival aids. As well as being able to accommodate more people, multi-seaters carry larger quantities of fresh water and more anti-sea sickness tablets, for example.

SINGLE-SEAT LIFERAFT

Only one person can fit into a single-seat liferaft. You can inflate it with carbon dioxide, by pumping air in manually, or by oral inflation.

MULTI-SEATER LIFERAFT

Many vessels carry multi-seater liferafts in a valise or a hard container. Multi-seaters can accommodate between four and 25 people and may be open-topped or covered. Many larger multi-seaters may also include locator beacons (see p. 236), paddles, and a solar still (see below).

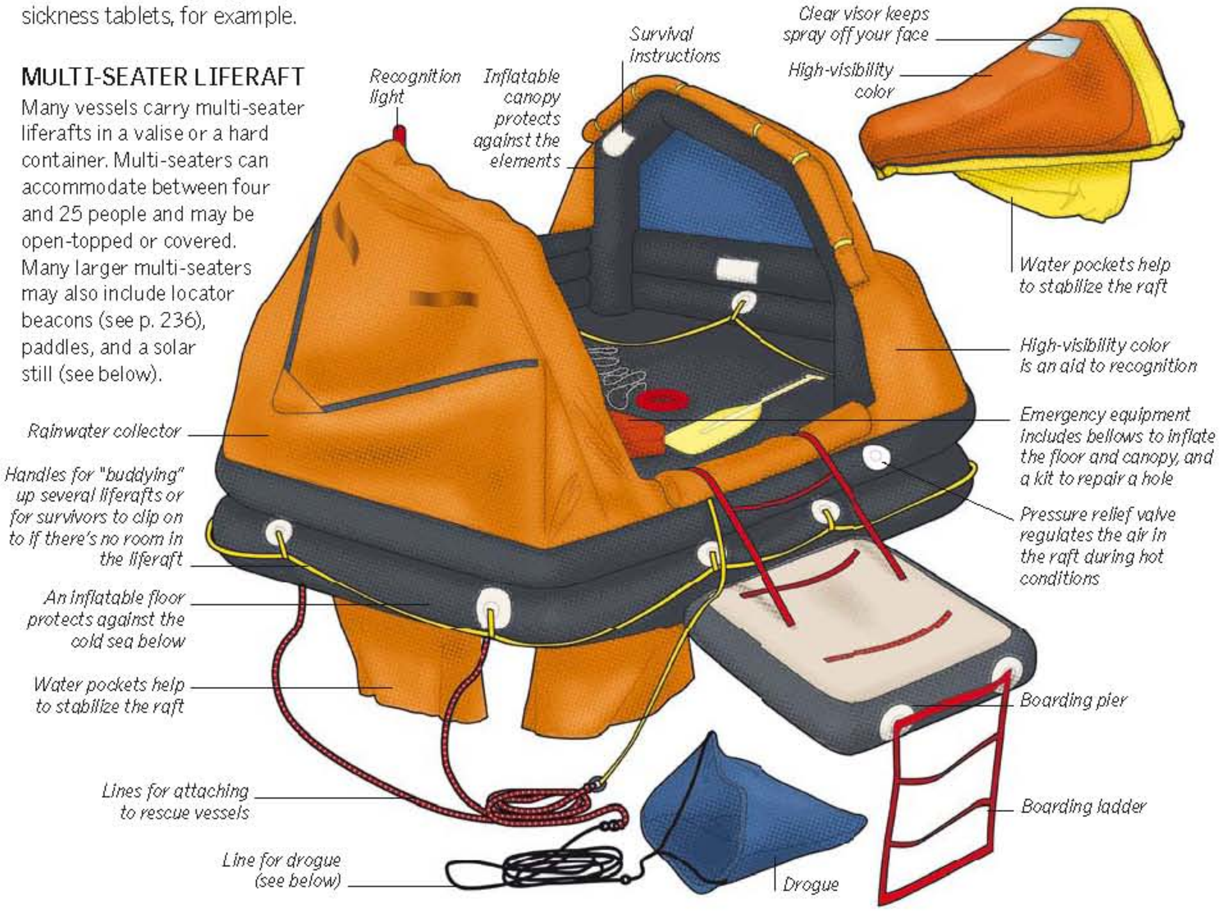
Clear visor keeps spray off your face
High-visibility color
Water pockets help to stabilize the raft
Recognition light
Inflatable canopy protects against the elements
Survival instructions
High-visibility color is an aid to recognition
Emergency equipment includes bellows to inflate the floor and canopy, and a kit to repair a hole
Pressure relief valve regulates the air in the raft during hot conditions
Rainwater collector
Handles for "buddying" up several liferafts or for survivors to clip on to if there's no room in the liferaft
An inflatable floor protects against the cold sea below
Water pockets help to stabilize the raft
Boarding pier
Lines for attaching to rescue vessels
Boarding ladder
Line for drogue (see below)
Drogue

SOLAR STILL

A solar still is a light, compact, and easy-to-use device for producing drinking water from sea water. Heat from the sun evaporates the salty water inside the still. Condensation on the walls collects in a channel around the rim and is directed to a store. Depending on the prevailing conditions and the availablilty of sunshine, a solar still can produce as much as 2 liters (3½ pints) of fresh drinking water a day.

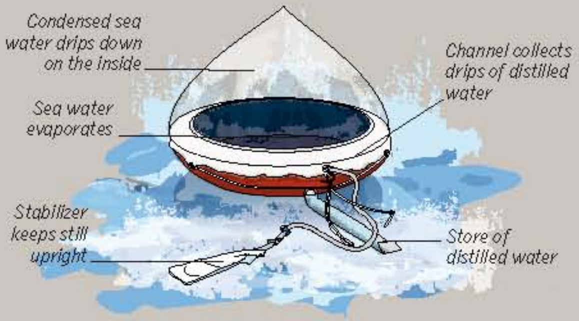
Condensed sea water drips down on the inside
Channel collects drips of distilled water
Sea water evaporates
Stabilizer keeps still upright
Store of distilled water

DROGUE

A sea anchor, known as a "drogue," is a critical piece of equipment because it helps to reduce drift and keeps the liferaft stable and seaworthy, particularly in heavy seas.

STEADYING A RAFT

A drogue stops a liferaft from overturning by creating drag that "anchors" the trailing side of the raft to the water.

- A drogue can position a liferaft either downwind or 90 degrees to the wind.
- It helps to keep the raft near the ditching location, thereby improving the chance of rescue. Even in a 2-knot current a liferaft can drift 50 miles (80 km) a day.

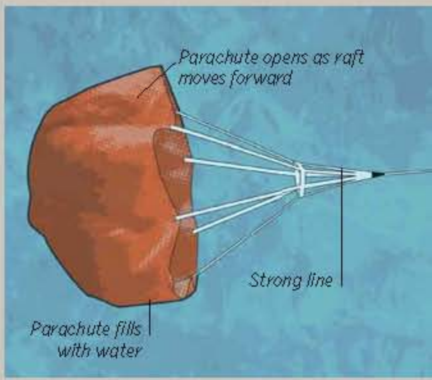
Parachute opens as raft moves forward
Strong line
Parachute fills with water

ABANDONING SHIP

Unless your vessel is an immediate danger to you, don't abandon it until you really need to. Even a badly damaged vessel can provide you with protection from the elements, equipment such as radios and flares, and provisions such as water and food. In addition, it is a big target for rescuers to spot. Make every effort to keep the vessel afloat.

THE DANGERS THAT AWAIT

When you do decide to abandon ship, you will be faced with the following dangers. If you have prepared yourself properly before heading out to sea, you will—in many cases, at least—be able to deal with them.

- Hypothermia brought on by inadequate clothing and exposure to wet, windy, and rainy conditions.
- Drowning because you have no lifejacket.
- Dehydration caused by a lack of water or an injury.
- Malnutrition due to a lack of food.
- Cold shock (see p. 254) due to sudden immersion in cold water.

BEFORE ABANDONING SHIP

Where time and circumstances permit, don't abandon your ship until absolutely necessary. Send a May Day signal with your name, position, group size, physical condition, and circumstances. Then carry out the following measures:

- Activate the 406 EPIRB (see p. 236).
- Ensure everyone on the vessel wears layers of clothing, and has a survival suit and lifejacket (see p. 250).
- Check the liferafts are ready for launch.
- Get the grab bag (see p. 250).
- Prepare to take the jerry-cans of water on board the liferafts.
- Fill spare containers with water.
- Gather up as much food as you can.

TAKING TO THE LIFERAFT

If you have to abandon ship, make every effort to launch all available liferafts. Even the craft that are not used will make it easier for search and rescue teams to detect the "footprint" of the survivors. They also contain additional supplies of water, flares, and other useful items.

- Many liferafts have a painter. This is a line that attaches the liferaft to the vessel to ensure it does not blow away when thrown overboard and inflated.
- Should the vessel sink, the painter has a weak link that breaks under pressure, or you can cut it.

PREVENTING HYPOTHERMIA

Getting wet greatly reduces your survival chances. At 41°F (5°C), a normally dressed person has only a 50 percent chance of surviving for one hour. You are six times more likely to survive in protective clothing.

- Climb into the liferaft carefully so that you don't get wet.
- Wear several layers of clothing to trap air. Even wet layers retain some heat around the body. Don't forget your head, hands, and feet.
- Wear a survival suit (see p. 250) to increase your survival time.

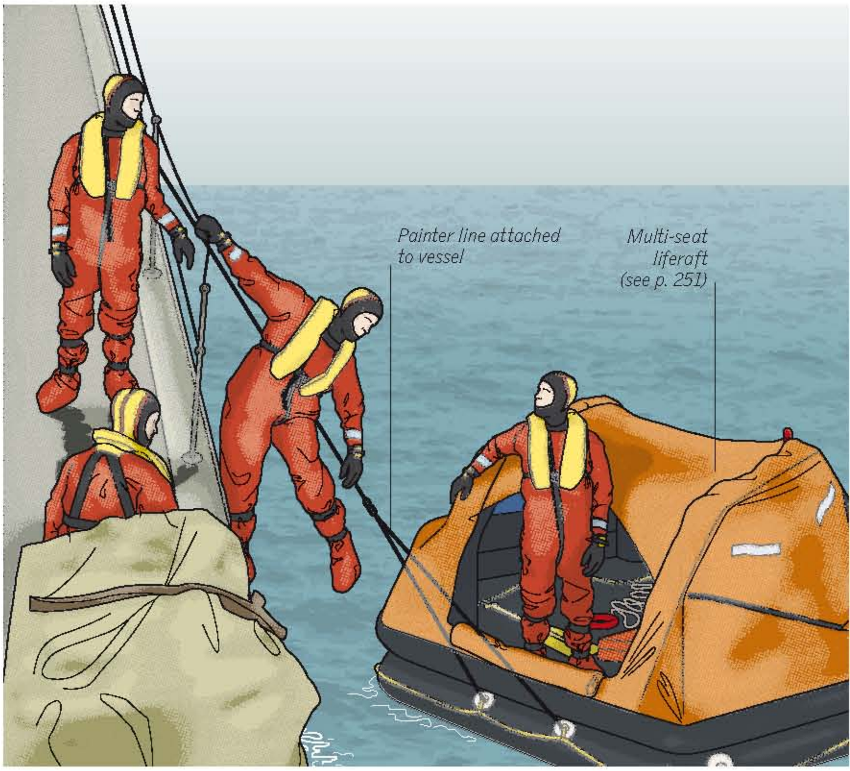
Painter line attached to vessel
Multi-seat liferaft (see p. 251)

① Move slowly and deliberately into the liferaft, keeping out of the water, and trying to stay as dry as possible. If you have to enter the water, climb down or lower yourself into it, rather than jumping.

- Look for the Day-Glo instruction patch inside the liferaft. Follow the recommended advice under "Immediate Actions" (see panel, opposite).

IN THE LIFERAFT

Once you have successfully negotiated the tricky procedure of abandoning ship and climbing into the liferaft, there are a number of actions you need to take. These are prioritized into immediate, secondary, and subsequent.

IMMEDIATE ACTIONS

- Inflate the floor of the liferaft with the bellows, while carrying out a roll call of your group to check for missing members.
- If the vessel is still afloat, keep the raft attached via the painter. Someone should be ready to cut the painter close to the vessel in case it starts to sink.
- Once clear of the vessel, set up the drogue (see p. 251).
- In difficult weather conditions, close the entrances of the raft in order to keep in heat, and keep out wind, rain, seawater, or spray.
- Bale out water, check for leaks, use the sponge to dry the liferaft, and use the leak stopper and clamps if necessary.

SECONDARY ACTIONS

- Treat the injured.
- Take anti-seasickness tablets.
- Post lookouts.
- Bring liferafts together.
- Warm up as best you can.

WARNING!

Never drink seawater, because the salt it contains will increase your rate of dehydration. If you're in hot conditions without any water, dehydration can set in within one hour.

SUBSEQUENT ACTIONS

- Delegate a leader based on experience.
- Find out who has useful skills, such as first aid and sea survival training.
- Ask everyone to look for sharp objects that could puncture the liferaft. At the same time, ask them if they have anything that could have a survival use.
- Find out what survival aids you have. Prepare aids to location (see p. 236) and show everyone how to operate them.
- Establish a routine and detail a watch system for inside and outside the liferaft.
- Delegate one member of the group to make repairs, another to keep the rations, and another to administer first aid.
- Establish how much water and food you have and start to ration in accordance with your particular situation.
- Procure water as soon as possible—don't wait until you need it. Deploy solar stills (see p. 251) and use the reverse-osmosis pump (see p. 197).

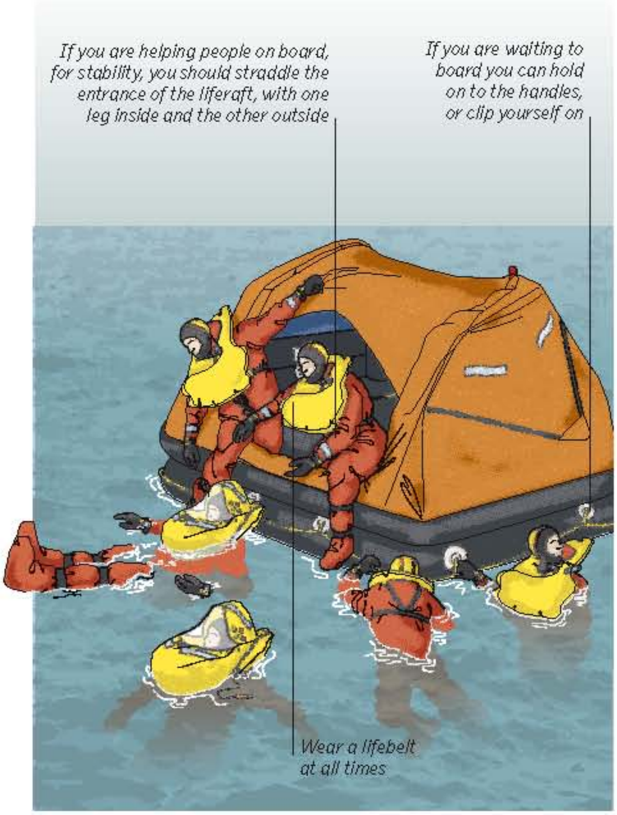
If you are helping people on board, for stability, you should straddle the entrance of the liferaft, with one leg inside and the other outside
If you are waiting to board you can hold on to the handles, or clip yourself on
Wear a lifebelt at all times

2 If your group does end up in the water, the first two people on board the liferaft should help others on, one at a time, lifting them under the arms.

- Don't overload the raft. Those least vulnerable can hang on to the raft's handles, or tie their lifejacket lifeline to one.

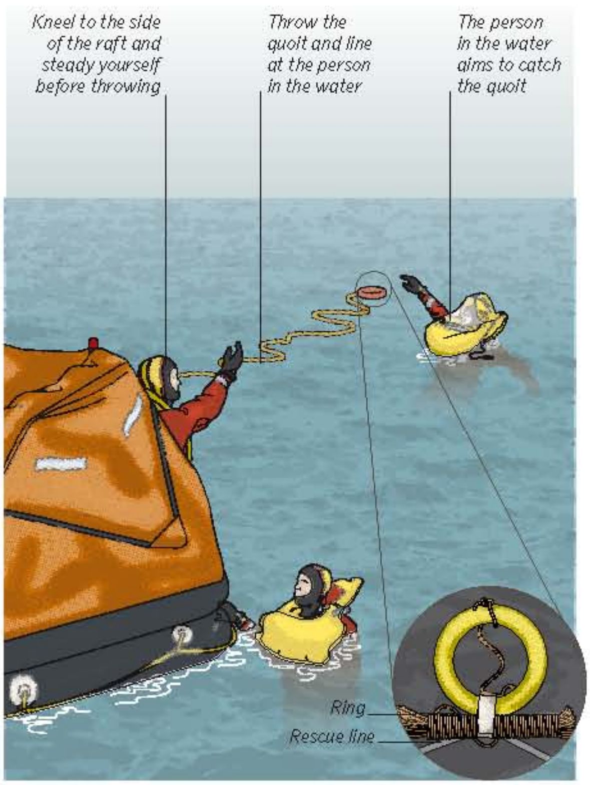
Kneel to the side of the raft and steady yourself before throwing
Throw the quoit and line at the person in the water
The person in the water aims to catch the quoit
Ring
Rescue line

3 Look for anyone who is struggling to make their way toward you. Use the rescue line and ring to pull them to safety in the liferaft.

- Avoid entering the sea unless you have to rescue an unconscious person from the water.

TAKING TO THE WATER

About two-thirds of people who drown in open water were within 10 ft (3 m) of a safe refuge and 60 percent of them were “good” swimmers. Here are some useful tips to prevent drowning.

JUMPING INTO THE WATER

Abandoning ship can be a dangerous procedure. If you have to enter the water, climb down using ropes or nets. If you have no option but to jump into the sea, these straightforward steps should improve your chances of survival. Jumping into the water is a last resort, as you will be prone to cold shock (see box, right) and at immediate risk of hypothermia and drowning. If you have to do it, first ensure your lifejacket is fitted correctly and check the surface of the water to find a safe entry point. Watch out for people, debris, and burning fuel. If you can’t avoid the fuel, swim under it and, before coming up for air, poke your hand through the surface to ensure you’re clear of it. As you come up, keep your face down to protect your nose, eyes, and mouth.

COLD SHOCK

Cold shock response is the body’s reaction to immersion in very cold water—for example, if abandoning a ship in winter in the North Atlantic Ocean. It is a common cause of death in such circumstances. Symptoms include gasping for breath and hyperventilation, which can lead to the inhalation of water, disorientation, panic, and the possible onset of hypothermia. A sudden increase in blood pressure and heart rate can cause cardiac problems in some people. Prolonged immersion in water will make it hard to perform physical movements, making swimming, climbing aboard a liferaft, or firing a flare extremely difficult.

- Body type or mental conditioning can help some people survive swimming in icy water.
- Dressing in layers and/or wearing a survival suit improves your chances of survival.
- Avoid entering the water if you possibly can.

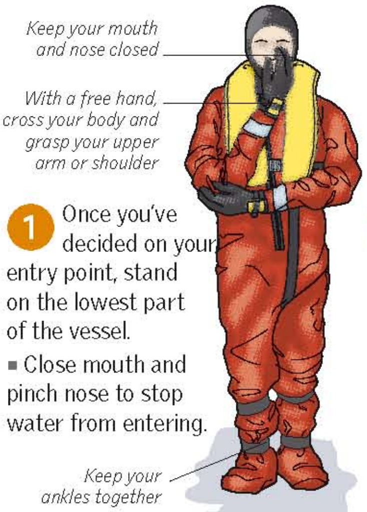
Keep your mouth and nose closed
With a free hand, cross your body and grasp your upper arm or shoulder
Keep your ankles together

1 Once you’ve decided on your entry point, stand on the lowest part of the vessel.

- Close mouth and pinch nose to stop water from entering.

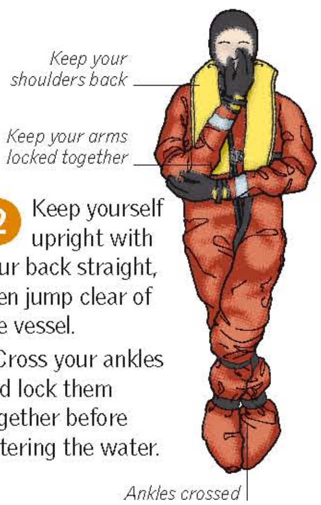
Keep your shoulders back
Keep your arms locked together
Ankles crossed

2 Keep yourself upright with your back straight, then jump clear of the vessel.

- Cross your ankles and lock them together before entering the water.

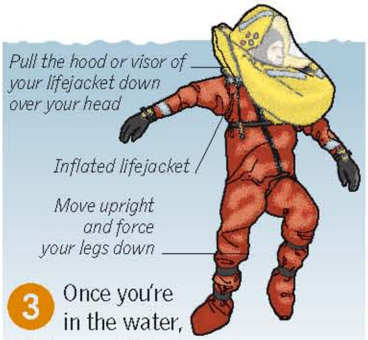
Pull the hood or visor of your lifejacket down over your head
Inflated lifejacket
Move upright and force your legs down

3 Once you’re in the water, inflate your lifejacket.

- If you’re in a survival suit, raise your arms and gently pull one seal away from your wrist to let the excess air escape.

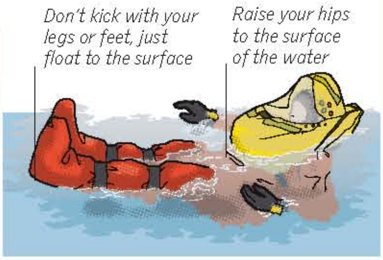
Don’t kick with your legs or feet, just float to the surface
Raise your hips to the surface of the water

4 Lie on your back and swim as calmly as you can away from hazards, such as burning fuel.

- Push your head back onto the neck of the lifejacket to help raise your hips.
- Keep your feet and knees together, and use a backward butterfly stroke to swim toward the liferaft.

IF YOU HAVE NO LIFERAFT

Even without a liferaft, you stand a better chance of surviving at sea if you’re in a group. More survivors create a larger target for rescuers to see, and being with others can be good for morale.

- Collect any floating debris before it disappears with the current, as this can increase your “footprint” so that rescuers can see you more easily.
- Determine what location aids and other equipment are available and prepare these for use.
- If there are children or injured among your group, place them in the center of the group and huddle close together.
- If you are alone, get into a position known as H.E.L.P. (see below), which stands for "Heat Escape Lessening Position." This will help to keep heat within the core of your body.

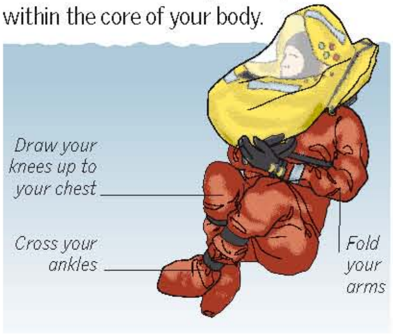
Draw your knees up to your chest
Cross your ankles
Fold your arms

FLOATING FACE DOWN

Don't panic if you have to enter the water without a lifejacket. Your body's natural buoyancy will keep at least the top of your head above water. To keep your face above water, too, make small windmill motions with your arms extended. However, if the water's rough, your only option may be to float face down.

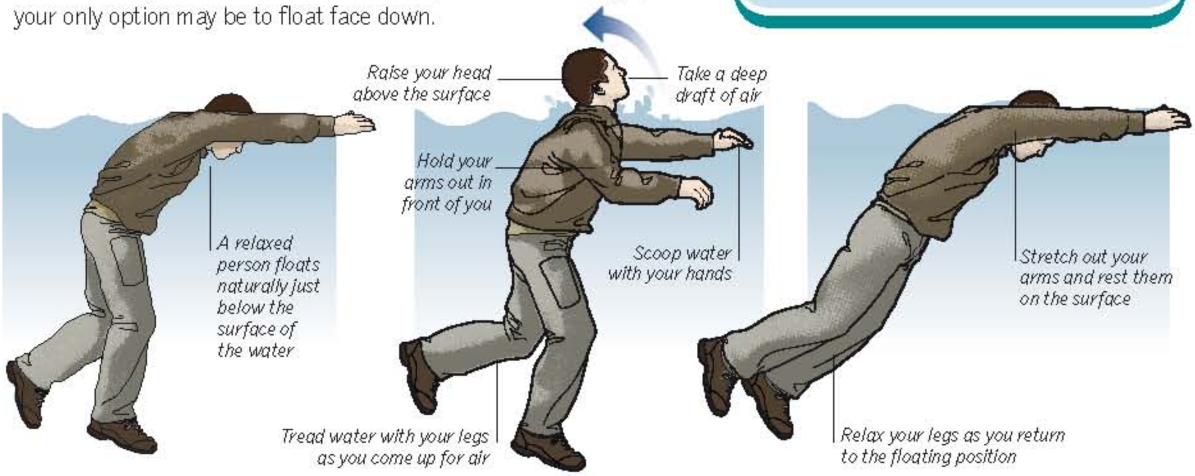
A relaxed person floats naturally just below the surface of the water
Raise your head above the surface
Take a deep draft of air
Hold your arms out in front of you
Scoop water with your hands
Tread water with your legs as you come up for air
Stretch out your arms and rest them on the surface
Relax your legs as you return to the floating position

1 It's important to relax, even though this seems hard when your life is in danger.

- Let your face lie in the water, and put your arms out in front of you.

2 Begin to exhale into the water as you raise your head up.

- Lift your head as it breaks through the surface and empty your lungs completely before inhaling more air.

3 With fresh air in your lungs, duck your face back into the water, keeping your mouth closed.

- Let your body float again, before you repeat the sequence.

STAY POSITIVE

It's very important to remain positive and to focus on your situation right now rather than on what might happen in a few hours or tomorrow.

IMPROVISING A BUOYANCY AID

If you're in the water and wearing pants, you can improvise a buoyancy aid that will help you to keep your head above water. Initially, it may be awkward to organize, but the benefit outweighs the effort involved.

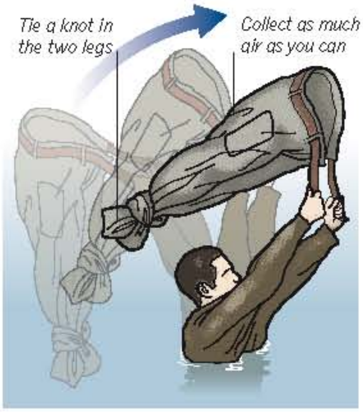
Tie a knot in the two legs
Collect as much air as you can

1 Take off your pants and tie the legs together near the bottom. Tighten the knot as much as you can with your teeth.

- Flick the pants over your head from behind until they fill with air. Tread water as you do this.

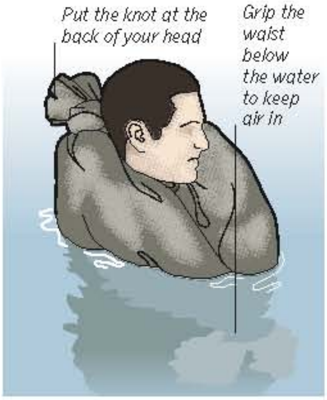
Put the knot at the back of your head
Grip the waist below the water to keep air in

2 Quickly catch the waist with your hands and grip it tight to hold air in.

- Put your head between the trouser legs and float.
- You will need to repeat the process regularly.

SECONDARY DROWNING

A person who nearly drowns inhales water that can cause potentially fatal chemical and biological changes in the lungs. This is known as "secondary drowning." It can be caused by inhaling fresh or salty water—even as little as 1fl oz (30 ml). Secondary drowning may occur 24–72 hours after immersion and, while uncommon, it's not rare. Take the following precautions to guard against secondary drowning:

- Monitor anyone who has nearly drowned or has spent some time in water.
- Ask the person to take a couple of deep breaths and check for any pain or discomfort that can't be accounted for.
- Look for symptoms such as coughing, breathing difficulties, chest pain, and saliva that looks like foam.
- Make sure the person is sitting in an upright position.
- If possible, provide oxygen—the blood levels of oxygen can fall quickly—and lots of rest and reassurance.

FIRST AID

FIRST AID

Regardless of the events that put you in a survival situation, a major factor that will dictate your options and actions will be whether anyone has been injured. Someone's survival may depend on the treatment they receive at the time of the incident, and during the ongoing survival ordeal, so it's essential that everyone should be proficient and confident in basic first aid techniques. In a true survival situation, the term "seek medical help" really means "deal with it yourself." While it would be impossible to be able to treat all injuries—a task even a paramedic would find daunting without the right equipment—the vast majority can be either dealt with or stabilized using a combination of basic first-aid techniques and common sense.

The "Protection" element of the survival principles (see p. 27) applies to your situation at all times—you should continually look at the consequences of every action in relation to protection against injury. Prevention is always a better option than cure. In hot climates, being able to recognize the first stages

In this chapter YOU WILL DISCOVER...

- that **maggots** have their **uses**...
- how to **improvise goggles** to prevent **snow blindness**...
- when to **stop**, **drop**, and **roll**...
- how to treat **snake bites** and **jellyfish stings**...
- when to **eat charcoal** or drink **bark tea**...
- how to prevent **frostnip** from becoming **frostbite**...
- when to **perform** a **firefighter's lift**...

of heat stress will allow you to take action with them before they develop into life-threatening dehydration and heatstroke. In cold climates, being able to recognize the first stages of frostnip will allow you to stop it from becoming frostbite. In many cases, further injury can be avoided by adopting the basic principles detailed throughout this book, such as being able to assess the best course of action in a given situation, plan a route and move safely over terrain, and protect yourself against the elements.

The will to survive is often the only factor that determines whether you live or die—regardless of your equipment, training, knowledge, and skills.

WHEN FACED WITH everything that man and nature can throw at you, when there appears to be no hope, you'll be faced with two choices: will you accept your situation and wait and see what hand fate deals you, or will you endure the pain and discomfort and fight for you own survival?

This determination was clearly shown by Aron Ralston during a hike in Blue John Canyon, Utah. After accidentally dislodging a 800 lb (363 kg) boulder, which pinned his right arm, he was faced with a bleak outlook. After five days, aware that no one knew he was missing and having run out of water, Ralston decided to amputate his own arm with his knife, applied a tourniquet, and hiked off to safety.

In another famous example, Simon Yates took the agonizing decision to cut the rope that held his injured climbing partner, Joe Simpson (who Yates thought had died), over a crevasse in the Peruvian Andes. Incredibly, Simpson survived the fall and chose to endure. He spent three days without food and only splashes of water from melting ice as he crawled and hopped the 5 miles (8 km) back over frozen mountainous terrain to reach their camp.

> “The **vast majority** of **injuries** can be either **dealt with** or **stabilized** using a **combination** of **basic first aid techniques** and **common sense**”

FIRST AID ESSENTIALS

SAFETY IS KEY on any expedition. Before you set off, make sure that you and everyone in your party has the necessary medical equipment—especially medications. If someone is injured, he or she should be treated immediately. If you can't contact the emergency services (and if there are enough of you), one person should stay with the casualty while two others go for help.

WARNING!
Protect yourself from danger at all times. You can't help anyone if you become a casualty as well. If the area is unsafe, don't approach the casualty—get emergency help and monitor the casualty's condition from a safe distance.

BASIC FIRST AID KIT

Keep your first aid kit (see panel, opposite) dry and readily accessible. Check the seals on sterile dressings; if they're not intact they're not sterile. Replace anything you use as soon as you can.

PRIORITIES AFTER AN INCIDENT

Assess a situation quickly and methodically. Find out what happened. Check casualties for life-threatening conditions such as unconsciousness or severe bleeding (see p. 264) and treat those first.

- **Response** Is the casualty conscious or unconscious? If the casualty is responding to you, he or she is conscious. Shake the shoulders gently if you're not sure.
- **Airway** Is it open and clear? If the casualty can talk, it is. If he or she is unconscious, open and clear it (see p. 276).
- **Breathing** Is it normal? Treat any difficulty such as asthma (see p. 275). If he or she is unconscious and not breathing, call for emergency help and begin CPR (see p. 277).
- **Circulation** Are there any signs of severe bleeding? If so, treat immediately. Once life-threatening conditions are under control, you can make a more detailed assessment. Examine the casualty methodically from head to toe. Ask how the incident occurred, as it can indicate likely injuries.

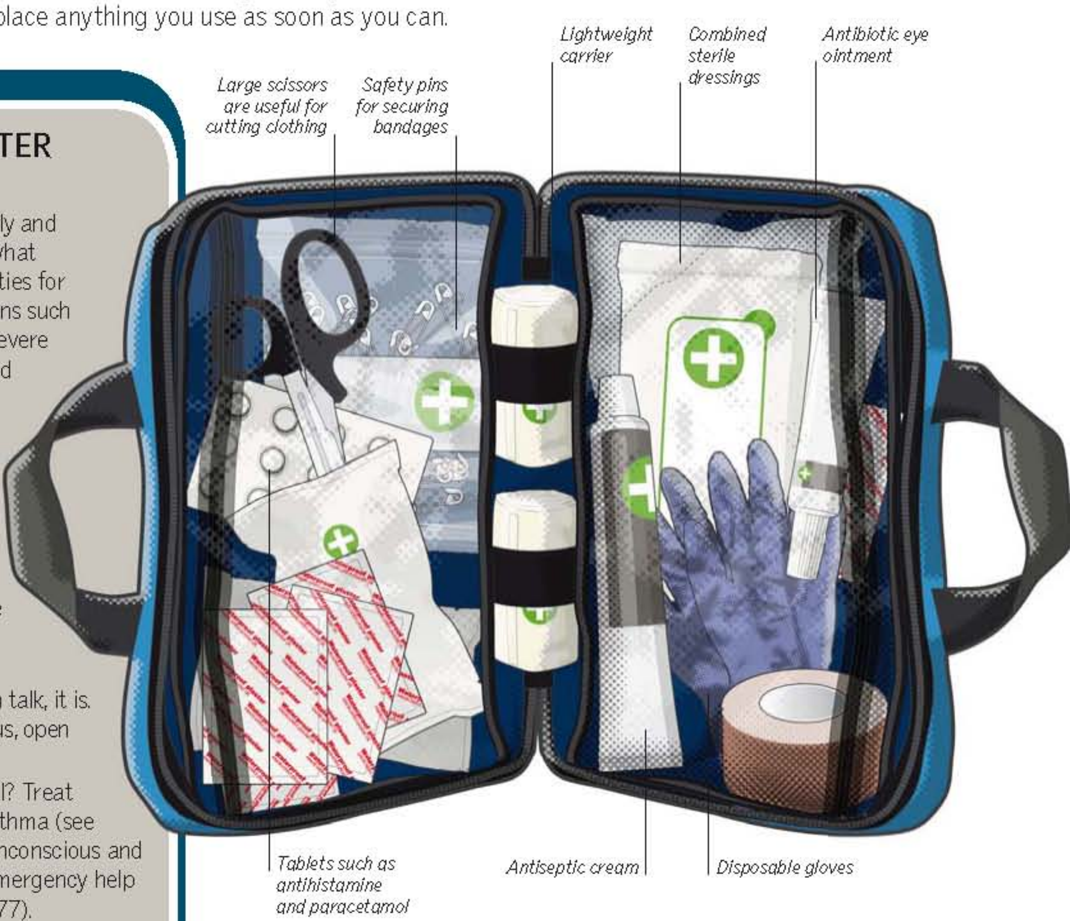
Large scissors are useful for cutting clothing
Safety pins for securing bandages
Lightweight carrier
Combined sterile dressings
Antibiotic eye ointment
Tablets such as antihistamine and paracetamol
Antiseptic cream
Disposable gloves

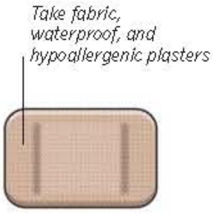
Take fabric, waterproof, and hypoallergenic plasters

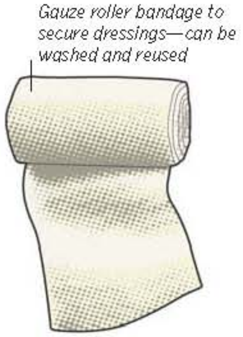
Gauze roller bandage to secure dressings—can be washed and reused

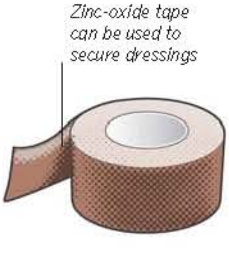
Zinc-oxide tape can be used to secure dressings

IMPROVISED SLINGS

Hand, arm, or shoulder injuries need to be immobilized and supported in a raised position. If you don't have a triangular bandage with you, use a piece of strong cloth about 3 ft (1 m) square, folded in half to form a triangle (see p. 270). You can also use your jacket or even rucksack straps. The jacket corner "sling" is the only one that provides enough support for a hand, wrist, or forearm injury. Ask the casualty to support the arm with his or her other hand while you secure the sling.

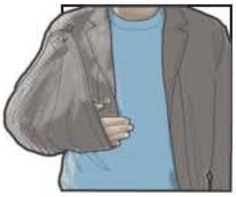

JACKET CORNER

To support an injured forearm or hand, fold the jacket up over the arm and pin it.

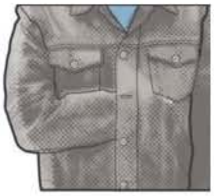

BUTTON-UP JACKET

Undo one of the buttons and slide the injured arm into the opening for support.

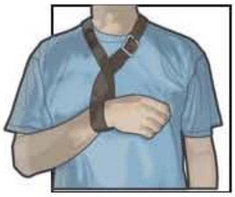

BELT SUPPORT

Support an upper arm injury in a raised position with a belt looped into a figure-eight.

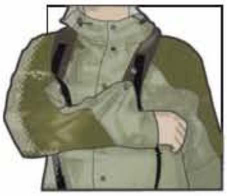

PINNED SLEEVE

Pin a sleeve to a jacket or the strap of a backpack for support.

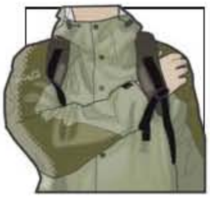

SHOULDER STRAP

Rest a sprain by tucking your hand in your backpack strap.

PROTECTING AGAINST INFECTION

Disposable gloves prevent cross-infection between you and the casualty; they must be latex-free, since contact with latex can cause an allergic reaction. Antiseptic wipes are also invaluable when cleaning wounds.

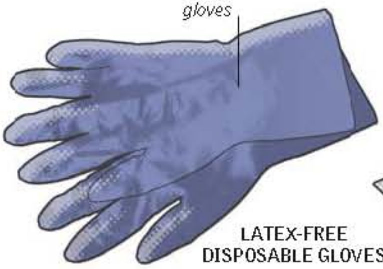

Use latex-free gloves

LATEX-FREE DISPOSABLE GLOVES

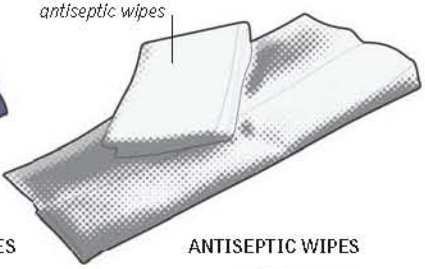

Use alcohol-free antiseptic wipes

ANTISEPTIC WIPES

COMBINED STERILE DRESSING

This is a sealed dressing consisting of a pad attached to a bandage. It's easy to apply, and can be used as a sling. It should be taped on to your backpack strap for easy access in an emergency.

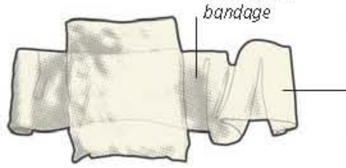

Sterile pad is sewn onto a bandage

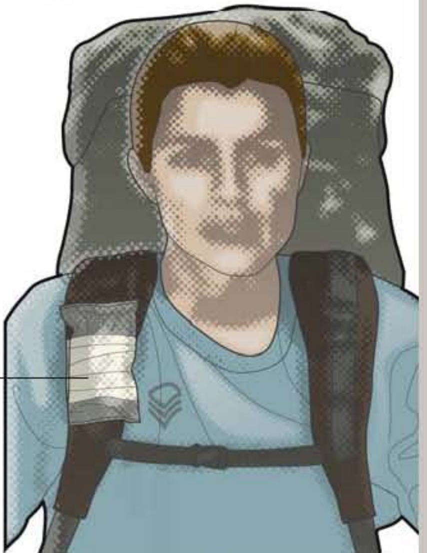

CHECK LIST

Make sure that your first aid kit and medicines are suited to the environment you will be visiting.

BASIC KIT

- Alcohol-free antiseptic wipes
- Latex-free disposable gloves
- Alcohol gel for handwashing
- Antiseptic cream
- Antibiotic eye ointment
- Adhesive dressings—fabric, waterproof, and hypoallergenic
- Gel blister bandages
- Combined sterile dressings, or sterile pads and bandages in assorted sizes
- Roller bandages—take self-adhesive for supporting joints and gauze for securing dressings
- Two triangular bandages
- Micropore or zinc oxide tape
- Scissors and tweezers
- Safety pins
- Disposable syringes

PERSONAL MEDICATION

- Painkillers
- Anti-inflammatories
- Medical alert bracelet/pendant
- Prescription medicines such as asthma inhaler and/or adrenaline (epinephrine) autoinjector
- Antihistamines
- Anti-diarrhea medicine
- Packets of oral rehydration salts
- Hydrocortisone cream

ENVIRONMENT-SPECIFIC EXTRAS

- Malaria tablets
- Mosquito repellent
- Anti-poison-ivy cream
- Sunblock
- Tick remover
- DEET powder for removing leeches

FLESH WOUNDS

ANY INJURY THAT BREAKS THE SKIN carries a risk of infection as germs can enter the body. These can come from the cause of injury, air, dirt, or clothing embedded in the wound. In the wilderness, keeping the injured area clean can be a real challenge, but it is essential. Tetanus is a potentially lethal infection caused by bacteria that live in soil. It can be prevented by immunization, so ensure that your vaccinations are up to date.

MAGGOT THERAPY

If a wound is seriously infected and you have no antibiotics, expose it to flies for one day, then cover. Maggots will develop and eat any dead tissue. Check daily, and flush the maggots out of the wound with sterile water before they start to eat healthy tissue.

BLISTERS AND BRUISING

A blister is a fluid-filled "bubble" of skin that occurs when skin is rubbed repeatedly against a surface (a friction burn). A bruise is bleeding into the skin and surrounding tissue from a blow that does not break the surface of the skin.

HOW TO TREAT BLISTERS

The ideal treatment for a blister is to rest and wait until it has healed, but this may not be possible in a survival situation. Cover it with a gel blister dressing if you have one or, if the blister is large, you may have to pierce it to enable you to continue walking. However, never pierce a blister caused by a burn—you risk infection.

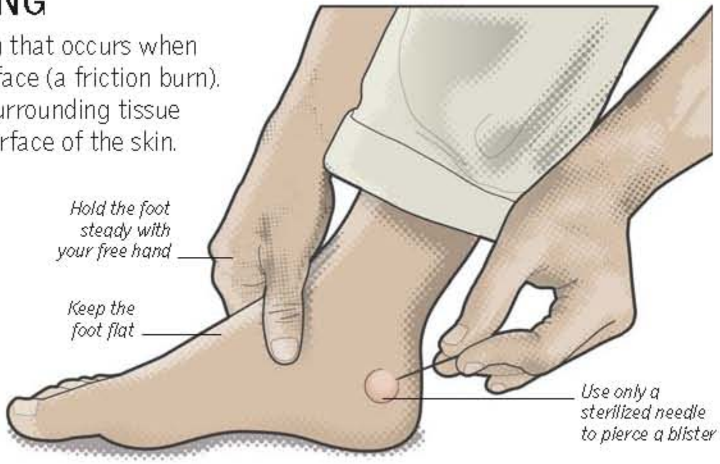
Hold the foot steady with your free hand
Keep the foot flat
Use only a sterilized needle to pierce a blister

1 If you have to burst a blister, first sterilize a needle by holding it in a flame until it's red hot, then letting it cool.
- Clean the area with water or wipes, pat dry, then pierce the edge of the blister.

BLISTER PREVENTION

Following these simple rules when out hiking should prevent blisters from developing in the first place:
- Ensure boots or shoes fit properly and are well "broken in" before you set out.
- Always wear clean, dry, comfortable socks next to your skin. Avoid wearing two pairs as they may bunch up, increasing friction.
- Keep toenails cut short and straight.
- Remove boots and air your feet during rest periods.
- If you feel a "hot spot," treat it before it becomes a problem: stop immediately and apply moleskin, a gel pad, or zinc-oxide tape.

2 Carefully apply pressure to the side of the blister opposite the hole made by the needle. Continue to apply pressure until all of the fluid has been squeezed out.

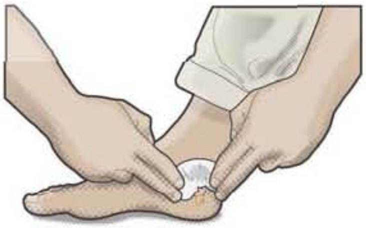

3 Gently clean and dry the blister, then protect the wound from infection by applying a dressing.
- Use a gel blister plaster or padded moleskin, if possible.
- Alternatively, use a pad held in place with zinc-oxide tape.

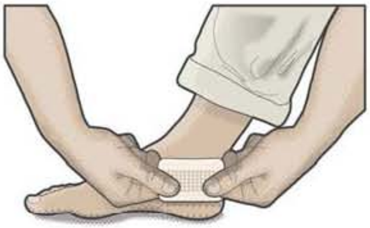

HOW TO TREAT BRUISING

To reduce the swelling and pain of a bruise, raise the injured area, and apply a cold compress. If you have access to cold running water, ice, or snow, soak a cloth and hold it against the bruised area for at least ten minutes. Severe bruising may indicate a more serious injury, such as a broken bone (see p. 271) or internal injury (see Shock, p. 274), which will require immediate treatment.

WOUNDS AND BLEEDING

Severe bleeding can be distressing for both you and the casualty, but it can usually be controlled by a combination of direct pressure and elevation of the injury. Remain calm and reassure the casualty while treating the wound. Treat for shock (see p. 274).

FOREIGN OBJECT IN A WOUND

Any loose foreign objects, such as pieces of dirt or gravel, should be removed from a wound, otherwise they may cause infection or delay healing. Either rinse them off with cold running water or carefully pick loose pieces off the wound with tweezers. However, if an object is embedded in a wound don't remove it, but treat as shown below.

1 Don't try to remove the object as it may be plugging bleeding. Control bleeding by pressing firmly on either side of the wound.

- Push the edges of the cut together but take care not to press directly on the object.
- Raise the wound above the level of the heart.

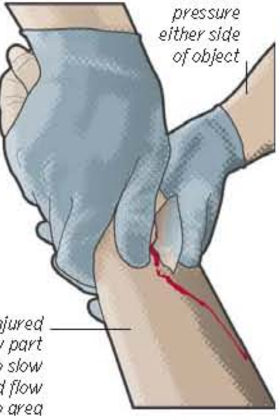

Maintain pressure either side of object

Keep injured body part raised to slow blood flow to area

2 Place a piece of gauze over the object to protect it, then build up padding on either side (rolled bandages are ideal). Bandage over the pads and the object.

- Check the circulation beyond the bandage every ten minutes (see p. 264).

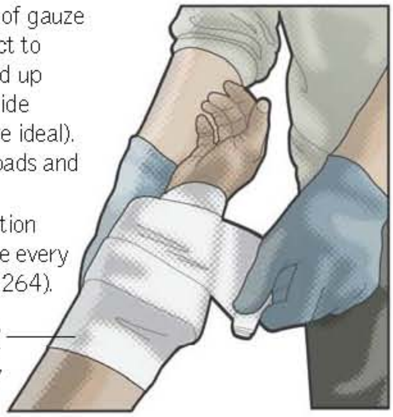

Bandage over the object to prevent further injury

MINOR CUTS OR GRAZES

Any break in the skin, however minor, needs to be cleaned and protected from infection. Rinse the wound with clean cold water and pat it dry. Then cover the injury; the dressing pad must be larger than the wound. For small grazes, use a plaster; for larger ones, use a sterile pad and a bandage.

NATURAL WOUND DRESSING

The birch polypore, or razorstrop fungus, which grows on the sides of old birch trees, can be used as a natural wound dressing. Cut a thin slice from the top of the fungus and secure it over the wound. If you are in any doubt about the identity of the fungus, do not use it.

TYPES OF WOUND

Different types of object and force produce different kinds of wound. It's useful to identify the type of wound you or the casualty has incurred so that the correct method of treatment can be applied.

BRUISE (CONTUSION)

A blunt blow will break blood vessels under the skin, causing blood to leak into the tissues. This results in a bruise: the skin is tender, swollen, and blue-black in color.

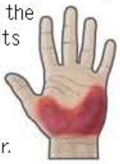

GRAZE (ABRASION)

A friction burn from a rope, or a sliding fall, will scrape off the top layers of skin, leaving a raw, tender area. Grazes often contain embedded foreign matter.

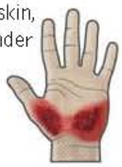

TEAR (LACERATION)

If the skin is torn open, the wound may not bleed as badly as an incised wound, but a larger area of tissue may be damaged and vulnerable to infection.

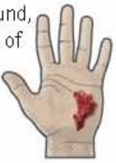

PUNCTURE WOUND

Sharp objects such as nails or sea urchin spines can puncture the skin. The entry hole will be small but the wound will be deep, with a high risk of infection.

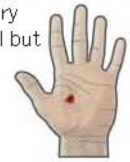

INCISED WOUND

If a sharp-edged object cuts across the skin, blood vessels will be sliced open and bleeding will be severe. Nerves or tendons may also be damaged.

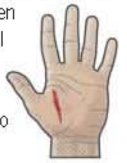

STAB WOUND

A penetrating wound from a long, bladed instrument is a very serious injury. A stab wound to the torso can damage vital organs and cause internal bleeding.

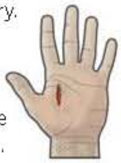

GUNSHOT WOUND

Check if the casualty has an exit wound. Treat entry and exit wounds separately.

ENTRY WOUND

Bullets drive deep into, or through, the body. They leave a small, neat entrance wound and cause serious internal damage and contamination.

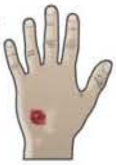

EXIT WOUND

If a bullet passes through the body, the exit wound will be large and ragged. Don't remove the bullet if there is no exit wound.

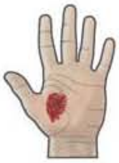

SEVERE EXTERNAL BLEEDING

Control the bleeding with direct pressure over a wound and, if possible, call for emergency help. Never use a tourniquet as this can cause severe tissue damage. Life-threatening shock is likely to develop if blood loss is severe (see p. 274).

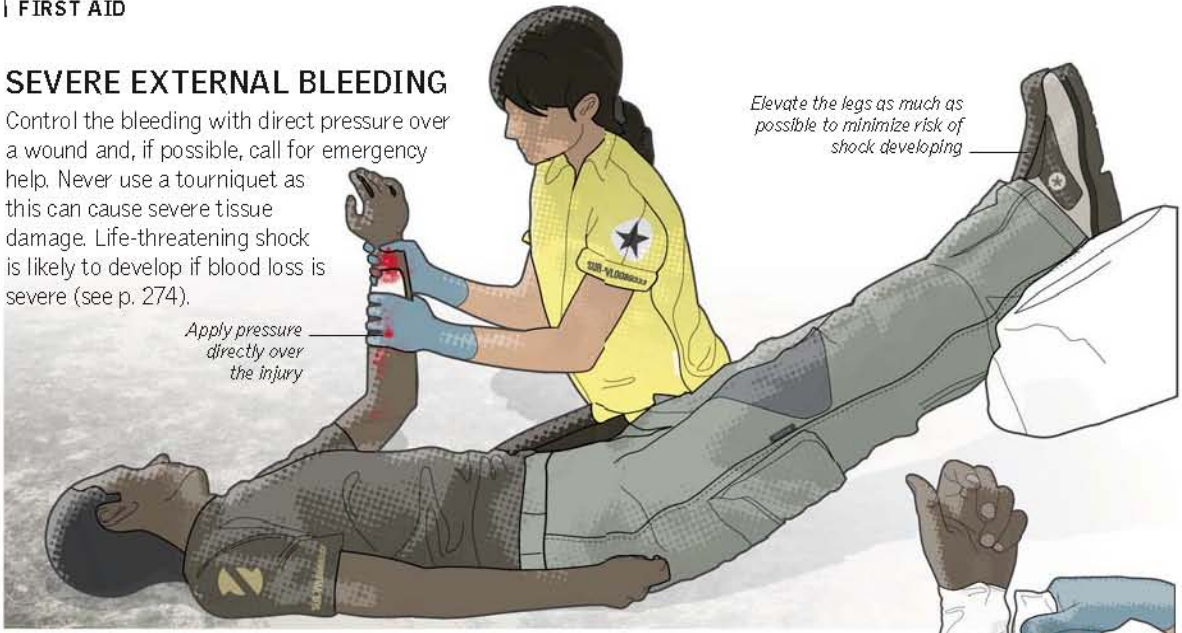

Elevate the legs as much as possible to minimize risk of shock developing

Apply pressure directly over the injury

1 Remove or cut away clothing to expose the wound. Apply direct pressure to the wound, over a sterile dressing or pad if you have one. Raise the injured limb above the level of the casualty's heart to reduce blood flow to the area. Help him to lie down and raise his legs.

2 Secure the dressing with a bandage. If blood soaks through, apply a second dressing on top of the first.

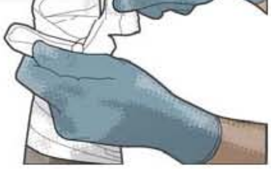

3 Every ten minutes, check the bandage is not too tight. Gently press a fingernail beyond the dressing. If the skin color does not return quickly, rebandage more loosely.

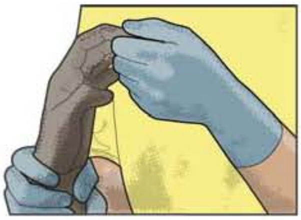

VARICOSE VEINS

When the one-way valves in veins fail, blood pools behind them, causing raised knobbly skin. The taut, thin-walled veins can easily be burst by a knock. Bleeding will be profuse.

1 Help the casualty to lie down and raise and support the injured area as high as you can. This reduces bleeding straight away. Expose the wound and apply direct pressure over a sterile dressing or pad.

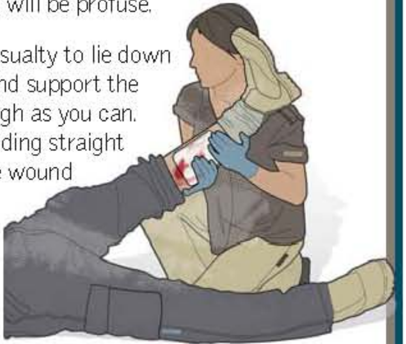

2 Bandage the pad firmly to maintain pressure on the wound. Keep the area raised. Check the bandage is not too tight (see step 3, above). If necessary, loosen it but maintain pressure.

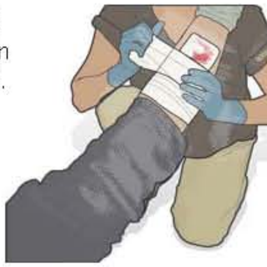

SCALP WOUNDS

A scalp wound can bleed profusely, making it appear worse than it actually is. However, it may mask a more serious head injury. If a casualty becomes drowsy, has a headache, or double vision, get emergency help if possible.

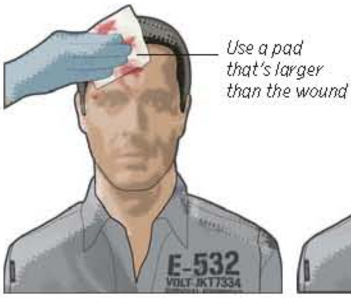

Use a pad that's larger than the wound

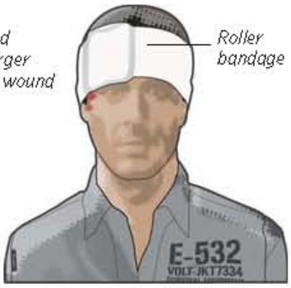

Roller bandage

1 Sit the casualty on the floor. Carefully replace loose flaps of skin, cover the wound with a sterile pad, and apply pressure.

2 Secure the dressing with a roller bandage. If the casualty doesn't quickly recover or if he deteriorates, get help.

EYE INJURIES

The eye can be seriously injured by a blow or contact with sharp objects, such as a tree branch, risking scarring, infection, or even loss of vision. Always wear eye protection when hiking through dense vegetation.

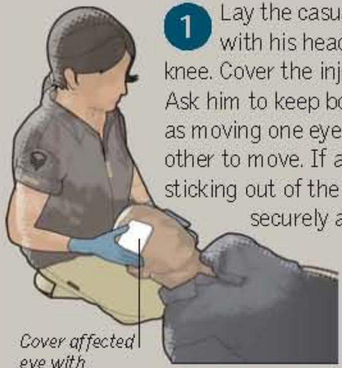
Cover affected eye with sterile pad

1 Lay the casualty down with his head on your knee. Cover the injured eye. Ask him to keep both eyes still, as moving one eye causes the other to move. If an object is sticking out of the eye, pad securely around it.

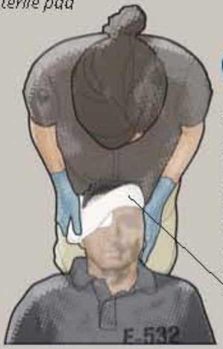

2 Secure the dressing with a bandage. If you're on your own, fix the dressing in place with plasters or tape and try not to move your eyes.

FOREIGN OBJECT IN THE EYE

If you can see an object on the surface of the eye, try lifting it off with the corner of a handkerchief, or wash it out with sterile water, washing away from the good eye. Don't remove anything that is sticking to the eye.

SNOW BLINDESS

Also known as "flash burn," this occurs if the surface of the eye is damaged through exposure to ultraviolet light, such as the glare from sun reflected off snow or water. Wearing sunglasses can prevent this. In an emergency, make your own goggles from cardboard or birch bark. If a person is affected by snow blindness, give him gauze pads to hold against his eyes. Bandage them in place if help is delayed.

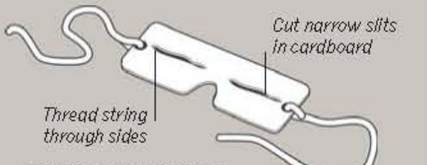

IMPROVISED GOGGLES

BURNS AND SCALDS

There is a serious risk of infection with all burns. Burns may affect only the outermost layer of skin, the upper layers, or the full thickness. Severe burns will affect all three layers. If the burn is larger than the casualty's hand it needs hospital treatment.

SMALL SUPERFICIAL BURNS

Cool the injury for ten minutes with cold running water or any cold, harmless liquid. Remove jewelry or watches from the affected area before it swells. Cover the burn to prevent infection.

LARGE OR DEEP BURNS

If a burn is extensive or deep, then fluid will be lost from the body and life-threatening shock is likely to develop (see p. 274). If the injury occured in a fire, the casualty could also have burns to his air passages so may have breathing difficulties. Don't burst any blisters as you will increase the risk of infection.

STOP, DROP, AND ROLL

If clothing is on fire:
- Stop moving
- Drop to the ground
- Roll over on the ground until flames are extinguished

1 Help the casualty lie down and protect him from the ground if possible. Cool the injury by dousing it with cold water for ten minutes; this also reduces swelling and relieves pain.

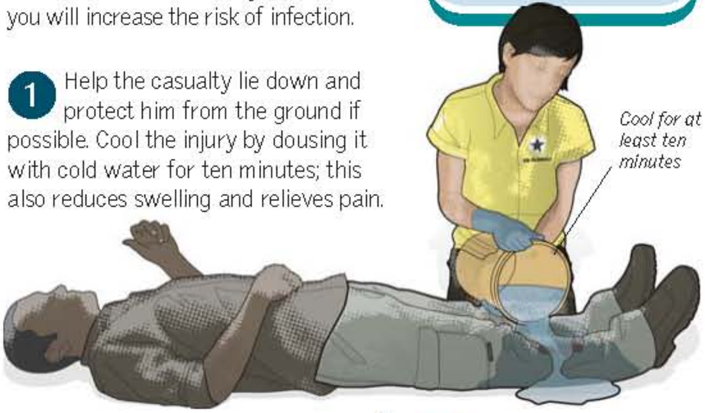
Cool for at least ten minutes

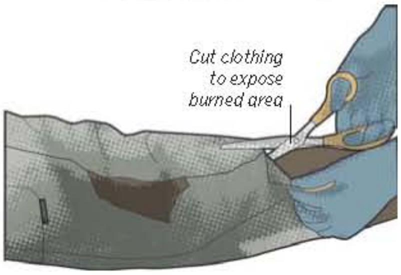
Cut clothing to expose burned area

2 While cooling the burn, remove or cut away clothing from around the burn. Don't touch or remove anything that's sticking to the burn.

3 Cover the entire area with plastic kitchen wrap, a clean plastic bag, or a clean, lint-free dressing. Get emergency help as soon as possible.

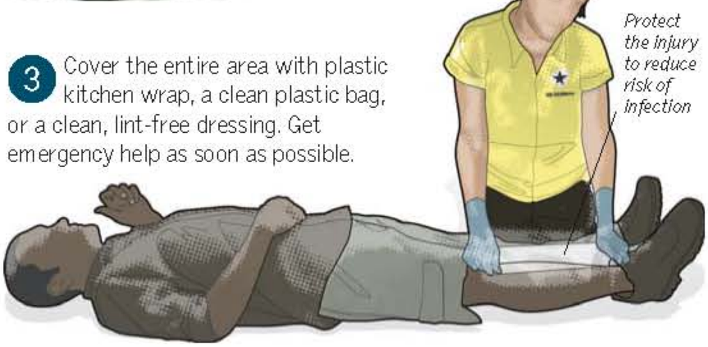
Protect the injury to reduce risk of infection

BITES AND STINGS

MANY BITES OR INSECT STINGS are painful, but most can be treated with simple first aid. However, there is a risk of an allergic reaction called anaphylactic shock. If a casualty develops a red blotchy rash, watery eyes or puffiness around the eyes, and/or breathing difficulties, get urgent medical help (see p. 274).

PORCUPINE QUILLS

If you are pierced by a porcupine's barbed quills, you must remove them or they will continue to work their way into the flesh and may pierce a vital organ.

- Cut off the ends of the quills, which are hollow, to allow them to deflate slightly.
- Pull the quills out, using pliers if you have them.
- Clean the puncture wounds and apply an antiseptic ointment.

TREATING BITES

An injury that breaks the skin carries a risk of infection. This is greatest with animal bites, since the mouth harbors so many germs. For a snake bite, you'll need to get emergency help, as the casualty needs to be transported on a stretcher in the treatment position to prevent the venom spreading through the body.

SNAKE BITES

Relatively few snakes are venomous, but it's generally safer to assume that all of them are. Most snakes cause a painful bite that leaves small puncture marks in the skin; however, the bite of a poisonous snake may in fact be painless. Symptoms also include nausea and vomiting, disturbed vision, and breathing difficulties.

1 Keep the casualty calm. Help her to lie down with her head, chest, and shoulders supported.

- Advise her to keep still.
- Make a note, if you can, of the time the bite occurred.

2 Apply a bandage around the site of the injury. Do not wash the area, remove any footwear or clothing, or attempt to suck the venom.

3 Tie another bandage around the affected limb that extends from the bite as far up the limb as possible.

- Check the circulation (see p. 264).
- Immobilize the affected limb.
- Move the casualty only with the use of a stretcher.

MAMMAL BITES

Bites from mammals—animals or humans—carry a serious risk of infection because the sharp teeth cause puncture wounds that transport bacteria deep into the tissue (see p. 263). As always, wear gloves to protect yourself. The bite can also crush surrounding tissue, and bleeding may be severe (see p. 264).

CLEAN WOUND

Raise the injury. Wash the wound with clean gauze and water. Pat dry, and cover it with a wound dressing.

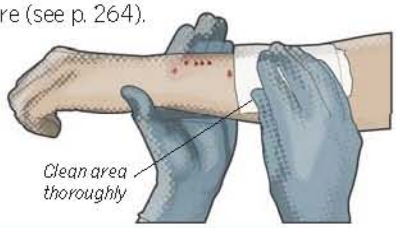

SNAKE IDENTIFICATION

Identify the snake if you can, since this will help medical teams find the correct antivenom. If you're not certain of the identity, make a note of its color or distinguishing features. If it's safe, put the snake in a secure container, but bear in mind that venom is active even if a snake is dead. Don't wash the venom off the bite; it can be used to identify the antivenom.

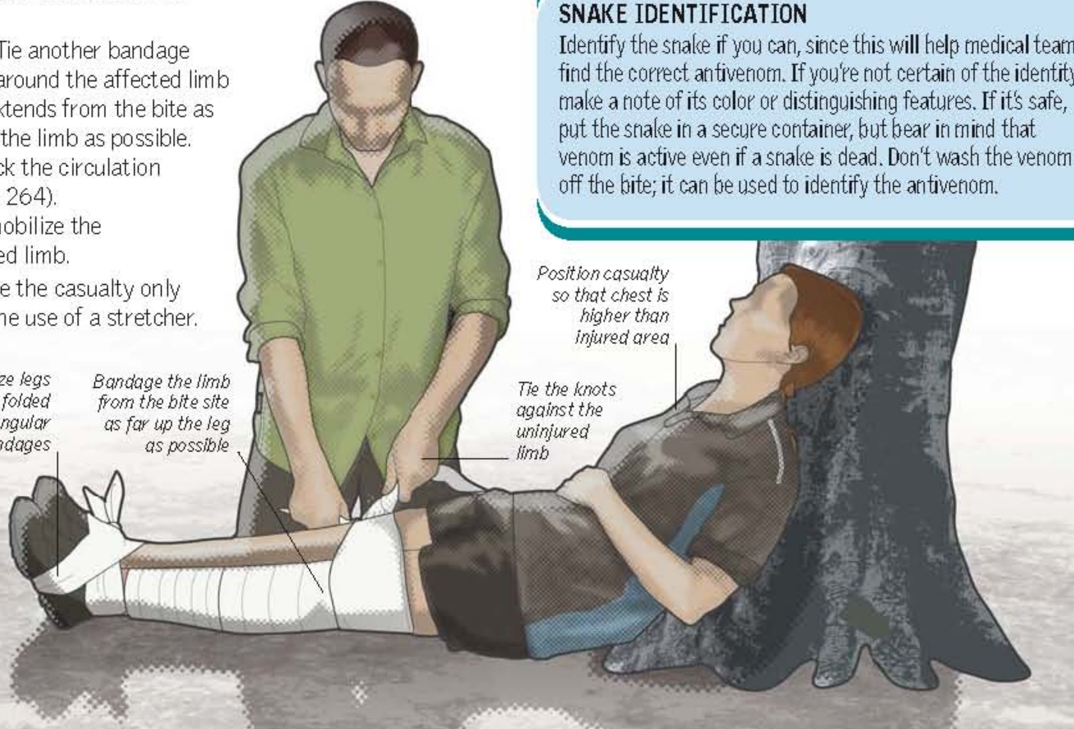
Immobilize legs with folded triangular bandages
Bandage the limb from the bite site as far up the leg as possible
Tie the knots against the uninjured limb
Position casualty so that chest is higher than injured area

TICK BITES

Ticks are tiny, spider-like parasites that live in grass or woodland. They feed on blood, attaching themselves to the skin with spiked mouthparts, and swell to the size of a pea. They carry Lyme Disease, so must be removed.

TRADITIONAL METHOD

Using tweezers, grasp the tick's head as close to the skin as you can. Pull the head upward using steady pressure (don't twist). Keep the tick in a container so it can be checked for Lyme Disease.

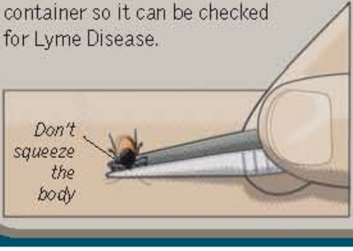

SPECIALIZED REMOVAL HOOK

Tick extractor tools that "unscrew" the mouthparts of the tick from the skin are available from pet shops. Slide the hook along the skin to grab the tick. Raise the hook very slightly and rotate it to lift the tick clear.

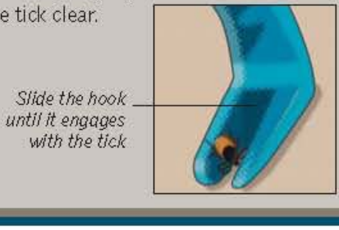
Slide the hook until it engages with the tick

HANDLING LEECHES

When in leech-infested territory, inspect your clothing and limbs every few minutes. Never pull a leech off if it's attached, as the jaws will remain in the skin and cause infection. To remove:

- Apply the juice of a raw lime, DEET, alcohol, or a dash of salt. If you're a smoker, put your cigarette butts in a piece of cloth, moisten the cloth, and squeeze nicotine onto the leech.
- Once the leech has dropped off, wash the area to remove the anticoagulant in and around the wound. Treat any bleeding (see p. 263) and cover the wound.

TREATING STINGS

Many stings are painful, but they are rarely life-threatening. Scorpion stings can be very painful and cause severe illness, and treatment should be as for snake bite (see opposite). Multiple insect stings of any type can produce a more serious reaction (see anaphylactic shock, p. 274).

WARNING!
A sting in the mouth or throat can cause swelling that blocks the airway. To minimize this, give the casualty cold water to sip.

INSECT STING

A bee, wasp, or hornet sting is often painful and followed by swelling and redness at the site. Some people have an allergy to stings, so monitor the casualty for signs of anaphylactic shock.

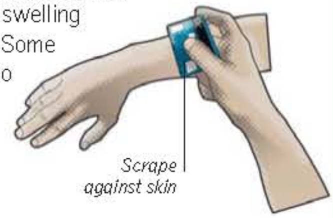

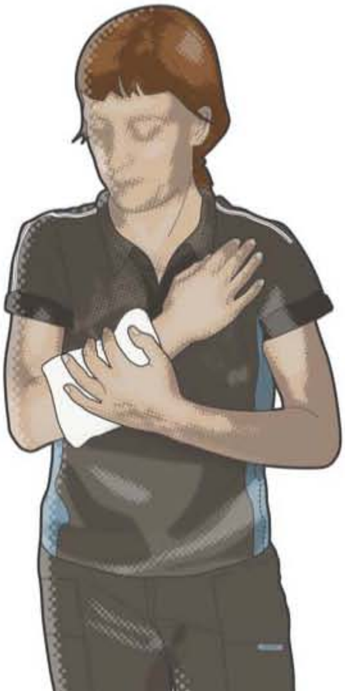

1. If the sting is visible, scrape it off sideways with the edge of a credit card, knife, or a fingernail.
   - Don't squeeze the sac as you may squeeze more venom into the area.
2. Raise the affected part and place a cold pad against it for at least ten minutes to minimize swelling.
   - Monitor the casualty for signs of allergy such as wheezing, or swelling around the face.

SEA CREATURES

When touched, sea anemones, corals, jellyfish, and Portuguese men-of-war release venomous cells that stick to the skin. Treat jellyfish stings as below; apply a cold compress to other stings to minimize swelling and relieve pain. Creatures such as weever fish have sharp spines that, if trodden on, become embedded, and may become infected.

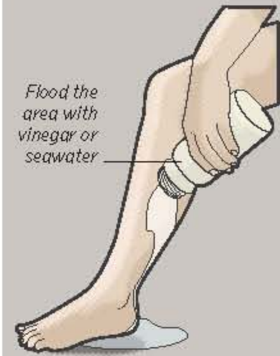
Flood the area with vinegar or seawater

JELLYFISH STING

Pour vinegar or seawater over the area to neutralize the sting. Help the casualty to sit down; treat as for snake bite (see opposite).

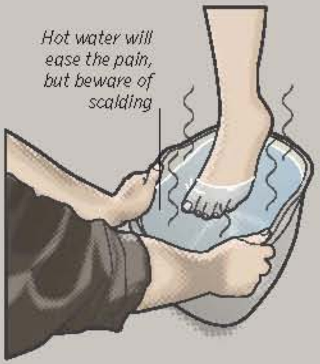
Hot water will ease the pain, but beware of scalding

SEA URCHIN SPINES

Immerse the injured part in water as hot as the casualty can tolerate for about 30 minutes. Get medical help, as the spines must be removed.

POISONOUS PLANTS AND INTESTINAL PARASITES

IT'S ESSENTIAL TO MAINTAIN personal hygiene in camp, to purify all drinking water from natural sources (see pp. 198–201), and to observe the rules of food safety if you wish to remain healthy in the wilderness. You should also learn how to recognize and avoid the poisonous plants in the region you're traveling in.

CONTACT POISONING

Poisonous plants don't always have to be ingested to cause damage. If they come into contact with the skin, the result can be pain, swelling, redness, a rash, and itching. Get medical help fast and, in the meantime, rinse the skin with cold water for 20 minutes. If the water splashes an eye, rinse for ten minutes. Although plants are the most common cause of contact poisoning, chemicals such as camp fuel can also be harmful.

WARNING!

If you think you've handled a poisonous plant, don't touch especially sensitive parts of your body such as the eyes, mouth, or genitals until you have washed your hands thoroughly.

- If you develop blisters on your skin, don't scratch them—however much they itch. If you break the skin, you'll run the risk of getting an infection.
- As well as painful blistering, some people may have an extreme reaction to the toxin. They should be monitored and treated for shock (see p. 274). Get medical help as soon as possible.

POISONOUS PLANTS

Poison ivy, poison oak, and poison sumac are the most well known of the plants that contain urushiol, an irritant oil. If you damage the plant and get the oil on your skin, you must wash it off immediately with soap and cold water. You should also remove and wash any contaminated clothing to prevent the oil spreading. Apply an anti-poison-ivy cream as soon as possible. Many people will develop an itchy rash and often painful blistering within 4 to 24 hours of contact with the plant (see warning panel, above).

POISON SUMAC

Found in wet acid swamps in eastern North America, poison sumac can grow as tall as 18 ft (6 m).

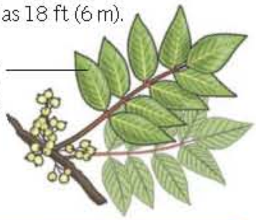
Oval leaflets grow in opposite pairs

POISON IVY

Native to wooded areas of North America, poison ivy is now found worldwide.

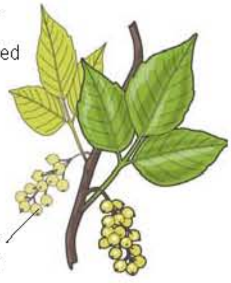
Berries are white when ripe

POISON OAK

Like poison ivy, the leaflets of poison oak grow in threes. This plant is found in wooded parts of North America.

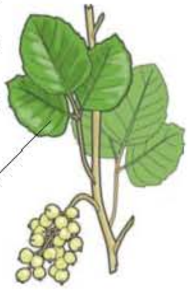
Leaflets are shaped like oak leaves

STINGING NETTLES

Found in many countries, stinging nettles usually cause only a temporary stinging sensation where the plant's hairs touch the skin. Apply a soothing cold compress or rub the affected area with a dock leaf if you can find one. Watch for an itchy red rash that indicates an allergic reaction.

NATURAL REMEDIES

Where nature causes a problem, she sometimes provides the solution. The following counter the effects of urushiol:

- Jewelweed has a sticky juice that will dry out blisters in a few days. Cut a piece of stem, split it down the middle, and rub the pieces on the affected skin.

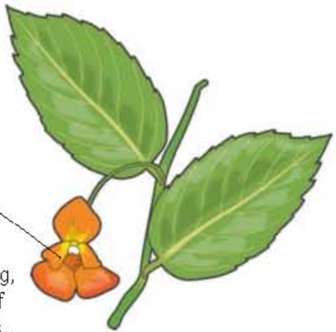
Jewelweed has distinctive pale yellow or orange, spotted flowers

- To soothe itching, make a poultice of witch hazel leaves. Mash up the leaves, adding water if too dry, then apply to the affected area.
- Try washing with a tannic acid solution. Tannic acid is found in tea, or you can make it from oak bark (see opposite).
- Tea tree oil, from the leaves of the melaleuca tree, is also said to counter the effects of urushiol. Apply directly to the affected area.

SWALLOWED POISON

Try to find out what the casualty ingested, how much, and when. If you suspect the casualty has swallowed a caustic chemical, such as camp fuel, do not induce vomiting as it will burn again on the way back up. Get medical help and monitor the casualty. If the lips are burnt, give frequent sips of cool milk or water. If you know the casualty has swallowed a poisonous plant or fungus and he or she is conscious, induce vomiting by tickling the back of the throat. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.

DIARRHEA AND VOMITING

In the wilderness, diarrhea and vomiting can kill, because they lead to dehydration (see p. 272) or even shock (see p. 274). The most likely causes are food poisoning or drinking contaminated water, although infectious diseases are a possibility, so get medical help if symptoms persist. You should rest, keep warm, and replace lost fluids. If you're hungry, you can eat small amounts of bland foods such as pasta for 24 hours.

NATURAL REMEDIES

In a survival situation, there are various natural remedies you can try to stop diarrhea or alleviate symptoms such as stomach pain, although some are more palatable than others:

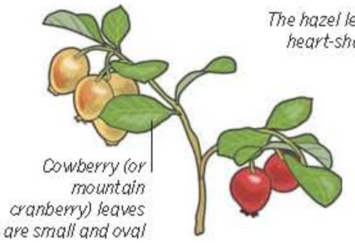
Cowberry (or mountain cranberry) leaves are small and oval

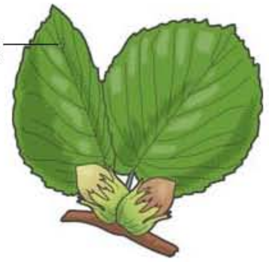
The hazel leaf is heart-shaped

- **Tea:** Drink tea made from hazel, cowberry, or cranberry leaves.
- **Bark:** Pull some bark off a tree (preferably oak), remove the inner bark and boil it for at least 12 hours, adding more water as necessary. The resulting black brew smells and tastes vile, but it contains tannic acid and will cure diarrhea. Drink one cup every two hours.
- **Charcoal**: Take a partially burned piece of wood, scrape off the char, then swallow about a handful with water.
- **Bones:** Burn to ashes then grind, or grind bones into a powder between two rocks. Make a paste with water, then swallow about a tablespoonful.
- **Chalk:** Grind into a powder, mix with water to a paste, then swallow about a tablespoonful.
- **Ash:** Make a paste of wood ash and water, then swallow. This will alleviate stomach pain.

REST AND REHYDRATE

Drink water to maintain your fluid levels. To help replace lost salts, dissolve a packet of rehydration salts in the water, or one teaspoon of salt in a liter of water, before drinking.

INTESTINAL PARASITES

There are two main types of intestinal parasite: helminths (tapeworms, pinworms, and roundworms) and protozoa (giardia, for example). The usual causes of infection are ingesting contaminated water or food, or poor personal hygiene. Symptoms include nausea or vomiting, diarrhea (see left), dysentry, bloating, stomach pain, weight loss, or a rash or itching around the rectum. Seek medical advice as soon as possible.

WILDERNESS CURE FOR WORMS

If you have passed a worm in your stool, swallow a couple of tablespoons of kerosene. While this method may make you sick, it will make the worms sicker. Gasoline will also work but is not as effective.

HOW TO AVOID BEING INFECTED

Prevention is far better than cure and to avoid being infected by parasites present in water or spread in feces you should always do the following in the wilderness:

- Boil all water, or use other reliable methods of purification (see pp. 198–201).
- Don't clean your teeth or rinse your mouth with water that isn't pure.
- Don't swim or stand in rivers or lakes in places where you could be at risk unless absolutely necessary.
- Cover any cuts or wounds on your skin.
- Maintain strict personal hygiene in camp (see pp. 116–17) and when handling food.
- Boil any meat you think might be infected for at least 20 minutes, or ideally until it falls off the bone, before eating it.

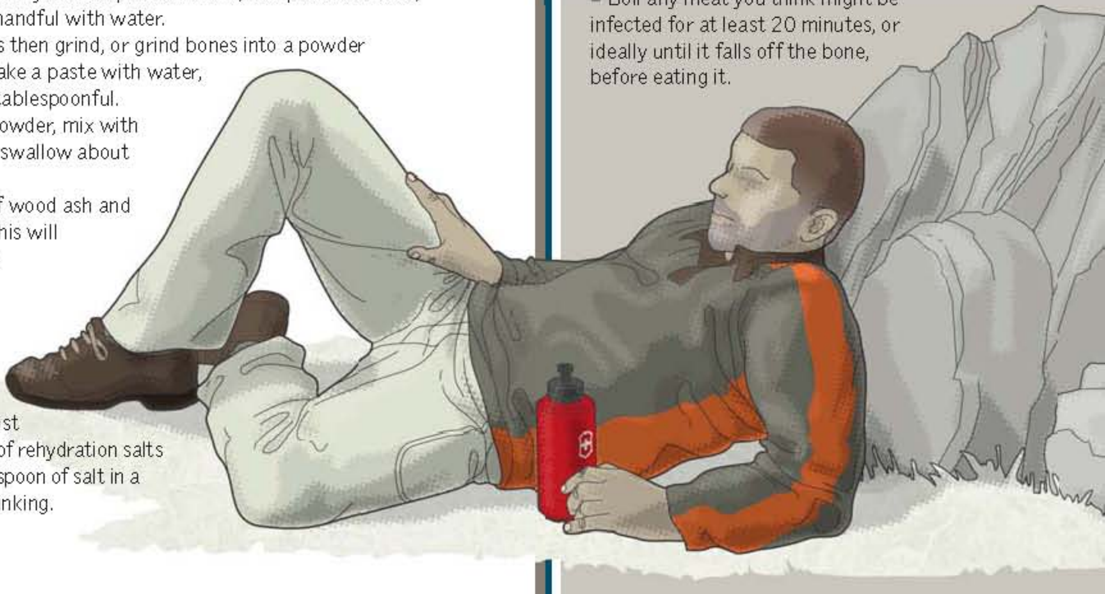

BONES, JOINTS, AND MUSCLES

CRAMP

This painful muscle spasm can be caused by dehydration and a reduction of body salts through perspiration—so make sure you have enough to drink when exercising. Sit down, rest, and stretch the affected muscles.

FOOT

Help the person stand on his good foot and stretch the muscles to reverse the spasm. Once the cramp has eased, massage the affected area of the foot.

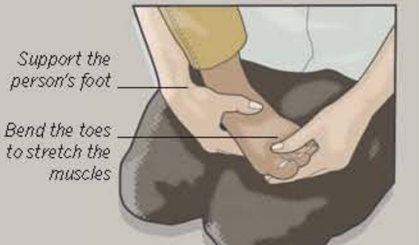
Support the person's foot
Bend the toes to stretch the muscles

CALF

Sit the person down and support the affected leg. Help him to straighten his leg, and flex his toes to reverse the spasm. Then massage the painful muscle.

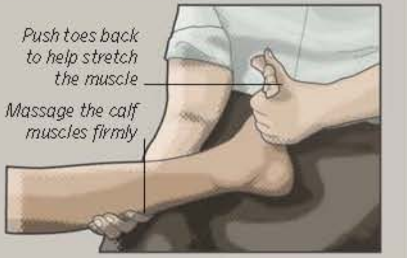
Push toes back to help stretch the muscle
Massage the calf muscles firmly

THIGH

If the cramp is in the back of the thigh, straighten the leg to stretch the muscle; if it's in the front of the thigh, bend the leg. Once the pain eases, massage the affected area.

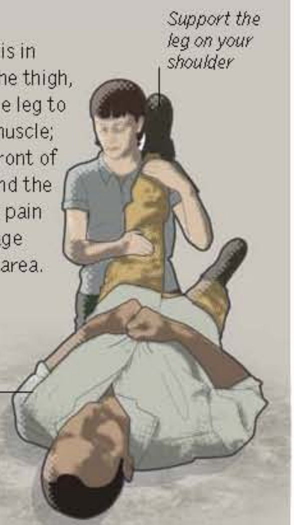
Support the leg on your shoulder
Advise casualty to lie down and relax the leg

IT CAN BE DIFFICULT TO TELL whether an injury is a sprain, a broken bone, or a dislocation. The ends of broken bones can move, damaging blood vessels or nerves nearby, so treat the casualty in the position found and immobilize the injured area before letting him or her move. Anyone with a spine injury or broken leg must be carried by stretcher.

SPRAINS AND STRAINS

A strain is a pulled muscle. A sprain occurs when ligaments that hold a joint together are damaged. The ideal treatment is to raise the injury, cool, and rest it. If you provide comfortable support, gentle movement can help the injury. If in any doubt, treat the injury as a broken bone (see opposite).

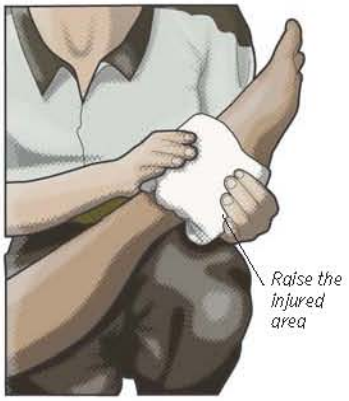
Raise the injured area

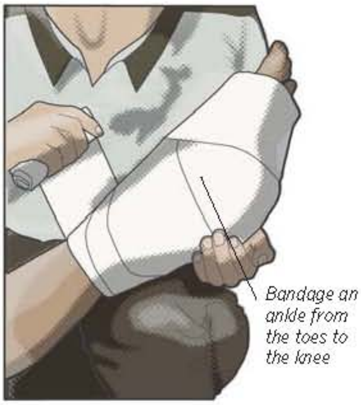
Bandage an ankle from the toes to the knee

1. Rest and support the injury. Wrap a cold compress around it for at least 10 minutes to reduce swelling and bruising.

2. Leave the compress in place, or wrap padding around the injury. Apply a crêpe bandage from below the injury to the next joint.

ARM INJURY

Falling onto an outstretched hand can cause a broken wrist, forearm, upper arm, or collar bone. Support the affected arm in a sling. If a casualty can't bend his arm, he may have injured his elbow, in which case don't use a sling. Instead wrap padding around the joint and secure the arm to the body with triangular bandages. To make sure that the bandage isn't too tight, check the wrist pulse.

USING A SLING

Slide a triangular bandage between the arm and the chest. Bring the front up over the arm and tie a reef knot on the uninjured side.

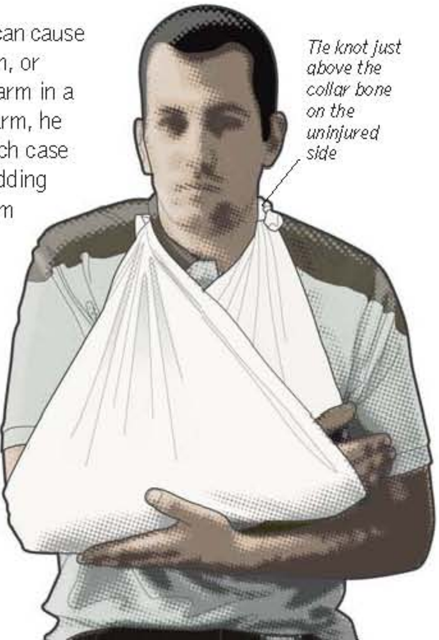
Tie knot just above the collar bone on the uninjured side

LEG INJURY

Injuries to the legs can be serious, as any fracture is likely to be unstable, which means that the ends of the bones can move easily, and could pierce one of the large blood vessels in the leg, resulting in severe bleeding. Don't move the casualty unless you have to, and, even then, only when the legs have been immobilized. If you see any signs of shock (see p. 274) ensure the head is low, but do not raise the legs.

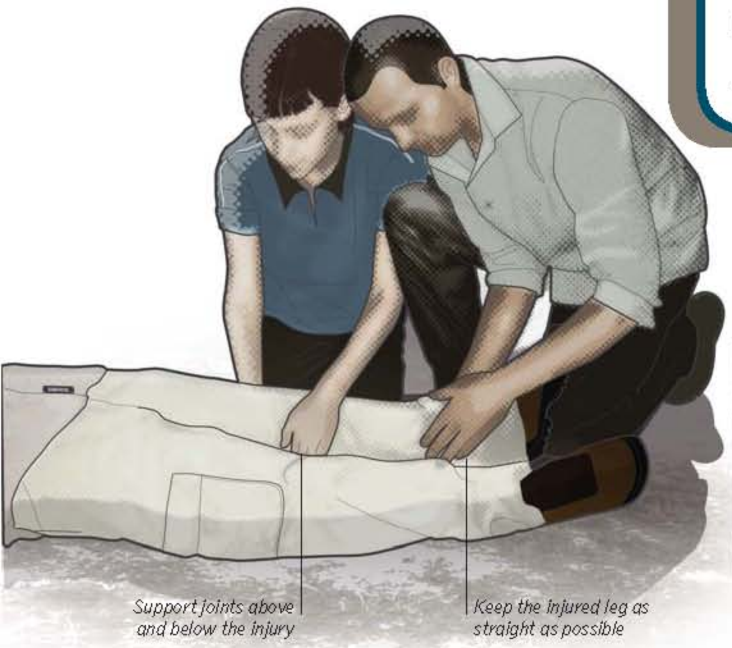

Support joints above and below the injury

Keep the injured leg as straight as possible

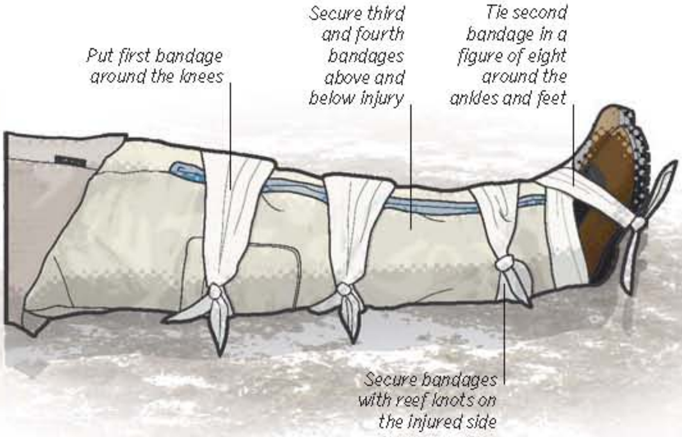

Put first bandage around the knees

Secure third and fourth bandages above and below injury

Tie second bandage in a figure of eight around the ankles and feet

Secure bandages with reef knots on the injured side

1 Lay the casualty down and support the injury to minimize further damage. Call for emergency help. If this is nearby just maintain this support. You can put rolled-up coats or blankets on either side for extra support.

2 If help is delayed or you need to transport the casualty, put bandages around the knees and ankles (and pelvis, if the thigh is injured), and above and below the injury. Place padding between the legs, then tie the bandages.

HAND INJURY

Injuries to hands are often complicated by bruising or bleeding. Raise the injured hand and treat bleeding with direct pressure (see p. 264). Remove jewelry before the area starts to swell. Wrap the hand in padding and support it in a raised position with a sling.

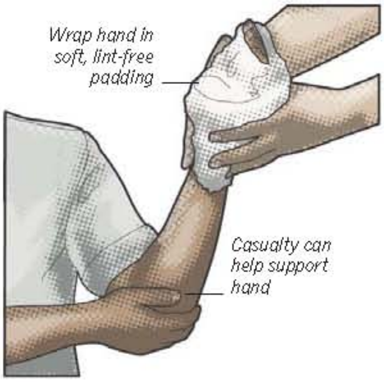

Wrap hand in soft, lint-free padding

Casualty can help support hand

SPINAL INJURY

If a person falls and lands on his back, or falls from a height, it is best to assume that he has a spinal and, probably, a head injury. Don't move him—support his head and neck in line with the rest of his back. Moving him could damage the spinal cord, which may result in permanent loss of movement below the injured area. Call for emergency help, or send someone to get help while you stay with the casualty. If you have to move him because he is in danger, use the log-roll technique (see p. 279).

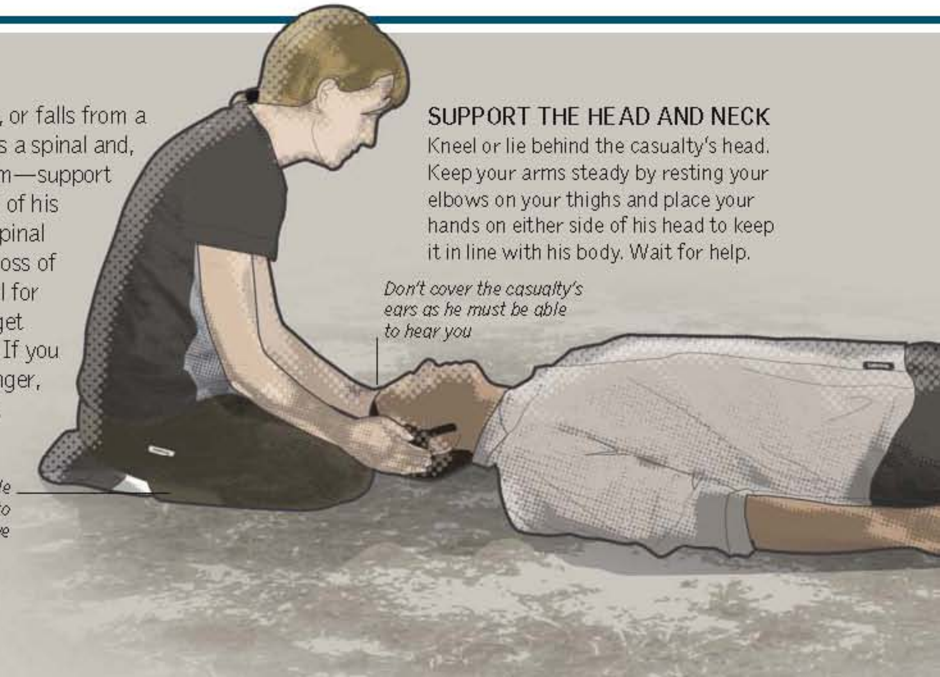

SUPPORT THE HEAD AND NECK

Kneel or lie behind the casualty's head. Keep your arms steady by resting your elbows on your thighs and place your hands on either side of his head to keep it in line with his body. Wait for help.

Don't cover the casualty's ears as he must be able to hear you

Make sure you are comfortable because you may have to wait for help to arrive

TREATING EXPOSURE

EXTREMES OF HEAT AND COLD can prevent the body's temperature-regulation mechanisms from functioning properly. Both extremes can cause life-threatening injures, so it's vital to act quickly. Don't leave a casualty—call for help or, if possible, send someone to seek help while you treat him or her.

WARNING!
If a casualty becomes unconscious, be prepared to give cardiopulmonary resuscitation (see p. 277).

HEAT INJURIES

In hot weather, wear a hat, reapply sunscreen frequently, and stay in the shade when you can to avoid heat injuries such as sunburn. You'll quickly become dehydrated if you don't drink enough water to replace fluids lost through sweating.

DEHYDRATION

Help the casualty to sit down and give him fluids to drink; water is usually sufficient, but rehydration salts mixed with water is best. If the casualty complains of cramp, help him stretch the affected muscles, then massage them firmly (see p. 270).

HEAT EXHAUSTION

If a casualty feels dizzy and starts to sweat profusely, but has cold, clammy skin, get him into the shade and give him fluids to drink. Help him to lie down then raise his legs—support his feet on a backpack—to help improve blood flow to the brain. Monitor him while he recovers.

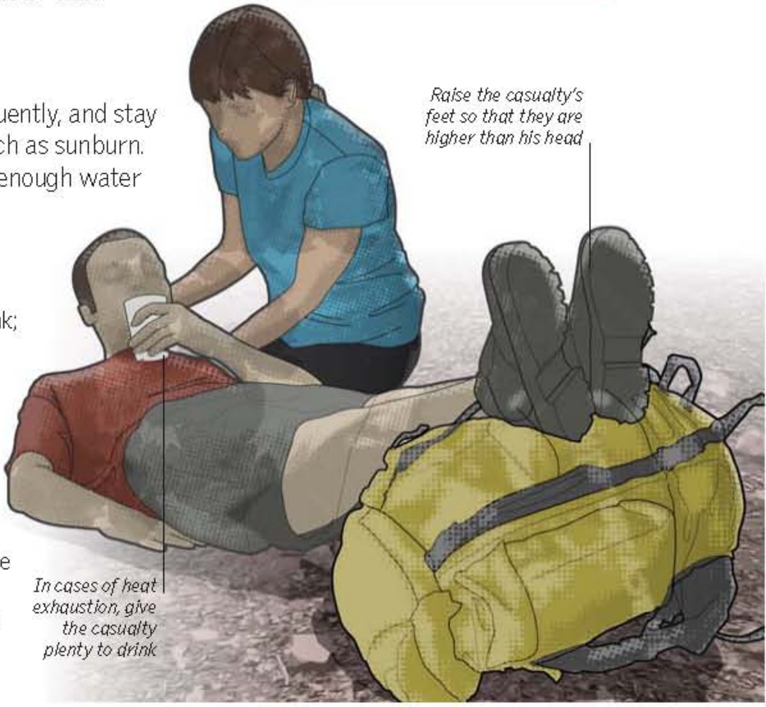

Raise the casualty's feet so that they are higher than his head

In cases of heat exhaustion, give the casualty plenty to drink

HEATSTROKE

This life-threatening condition may follow heat exhaustion or develop with no warning. Heatstroke causes the body's temperature-control mechanism to fail. If a person complains of headache, feels dizzy, has hot, dry skin, and begins losing consciousness, he or she may have heatstroke and will need urgent medical help.

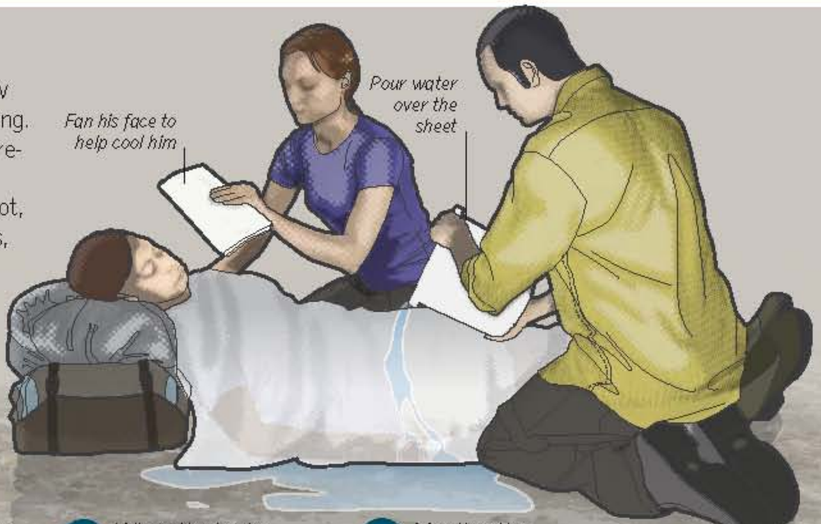

Fan his face to help cool him

Pour water over the sheet

1. Move the casualty to as cool a place as possible—out of the sun. Help him to sit or lie down with his head raised and remove all of his outer clothing.

2. Do anything possible to quickly reduce the body temperature. Ideally, wrap him in a cold, wet sheet. Keep the sheet cool by continually pouring water over it.

3. When the body temperature drops below 104°F (40°C), replace the wet sheet with a dry one.

4. Monitor the level of response, pulse, and breathing. If his temperature starts to rise again, repeat the treatment.

COLD INJURIES

Exposure to cold can result in parts of the body freezing (frostnip and frostbite) or the body's core temperature becoming dangerously low (hypothermia).

FROSTNIP AND FROSTBITE

Frostnip is the freezing of the top layer of skin, usually on the face and extremities. The skin turns numb, white, and hard. Untreated, it can lead to frostbite, which is much more serious—the deeper tissues, and even the bone, freeze. The skin turns white or blue and feels solidly frozen.

WARM GRADUALLY

Frostnip and frostbite can both be treated by warming the affected area, though frostbite requires more intensive treatment. Warm the affected area with body heat—place the casualty's hands in his own armpits, or place his feet in your armpits. Remove any rings and raise the injured part to reduce swelling. Ideally, place the affected area in warm water. Dress the injury in sterile dressings.

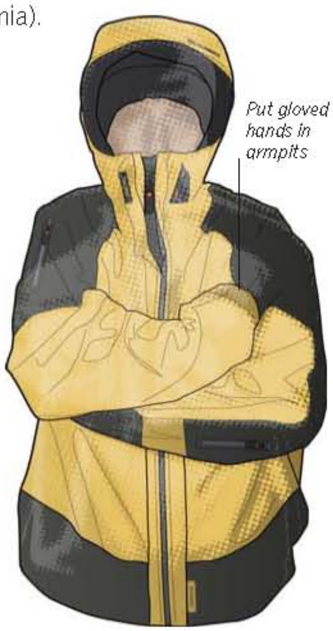

Put gloved hands in armpits

HYPOTHERMIA

This life-threatening condition develops if the body's core temperature falls below 95°F (35°C). Treatment aims to prevent further heat loss. It's vital that a casualty is warmed up gradually. If he or she is warmed up too quickly, blood is diverted away from vital organs, such as the heart and brain, to the skin, which can actually speed up cooling of the body. A casualty with hypothermia must be moved on a stretcher.

1 Help the casualty to a sheltered place where he should rest to prevent his body temperature falling further. Send someone else to get help.

2 If possible, and if there is no risk of further cold, remove any wet clothing and replace with warm, dry clothes—but don't give up your own clothes.

3 Put a thick layer of dry leaves underneath him to insulate him from the ground. Help him to lie down in a sleeping bag and, if you have one, cover him with a survival blanket.

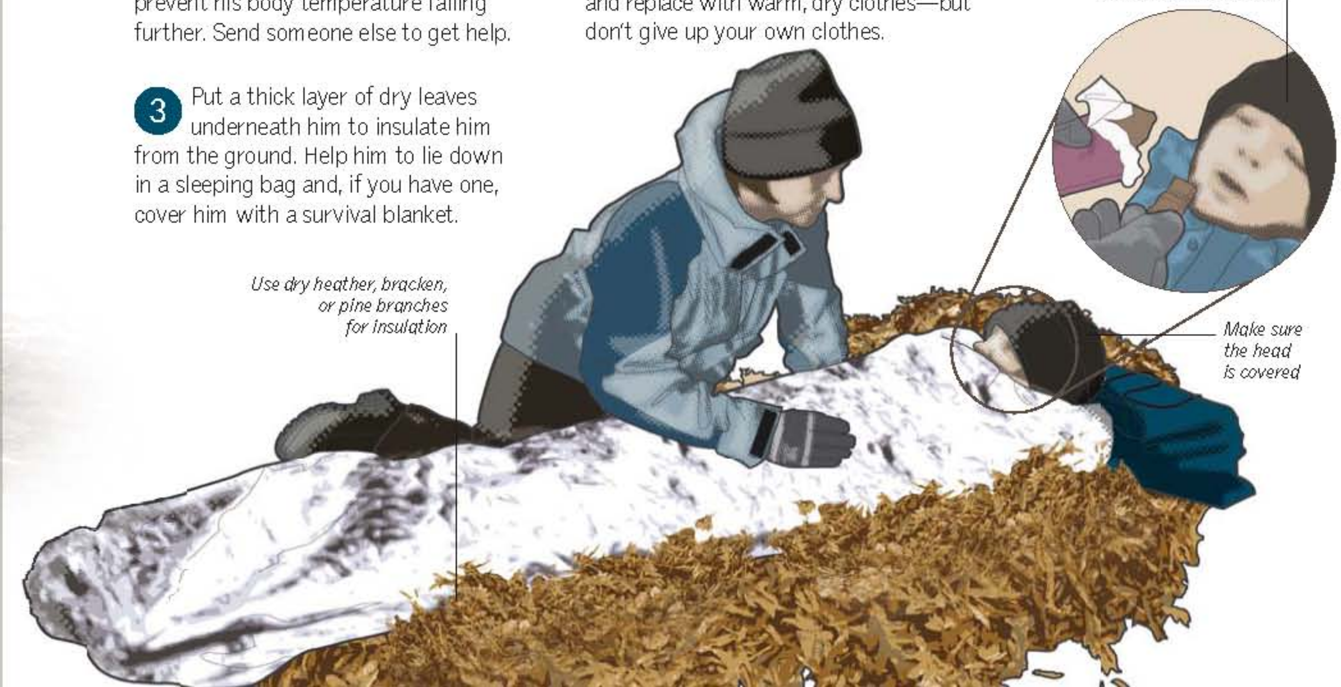

Give the casualty warm (not hot) sweet drinks and high-energy foods such as chocolate

Use dry heather, bracken, or pine branches for insulation

Make sure the head is covered

DROWNING

If a person has been immersed in cold water, there is a high risk of hypothermia. In addition, the cold can cause the heart to stop, or throat spasms can block the airway. Water can also enter the lungs and cause secondary drowning (see p. 255) hours after the person appeared to have recovered.

RESCUING A CASUALTY

If you have rescued a person from water, help him or her to lie down with the head low. Replace wet clothes with dry ones. Treat for hypothermia (see left). If the person is unconscious and not breathing, give five rescue breaths before beginning chest compressions (see p. 277).

ALTITUDE SICKNESS

Symptoms of altitude sickness include nausea, loss of appetite, shortness of breath, and a headache that's not relieved by medication. The casualty may also have difficulty sleeping and will feel unwell. The only solution is to start the descent immediately and remain at a lower altitude for a few days. Severe cases will need to be carried.

TREATING SHOCK

SHOCK IS A LIFE-THREATENING CONDITION that occurs if the circulatory system fails. The most common cause is severe bleeding, but it can also be a consequence of burns. Initially, there will be a rapid pulse and pale clammy skin. As the condition progresses, breathing becomes rapid and shallow, pulse weakens, and skin becomes pale gray-blue. If untreated, unconsciousness results.

ANAPHYLACTIC SHOCK

Anaphylactic shock is a rare but severe allergic reaction affecting the whole body. People who know they are susceptible to this carry a special adrenaline (epinephrine) autoinjector for use in an emergency. If the casualty has one, but is too weak to use it, take off the safety cap and, holding it in your fist, press it against the casualty's thigh (through clothing if necessary).

DEALING WITH SHOCK

Do not give the casualty anything to eat or drink as an anaesthetic may be needed; moisten his lips with water if he's thirsty. Call for help; the casualty must be transported in the treatment position.

1. Treat the cause of shock, for example bleeding or burns (see p. 265). Suspect shock, too, if you notice any symptoms, yet can't see any obvious injury—it could be caused by internal bleeding.

2. Help the casualty to lie down; insulate him from the ground with blankets or bracken. Raise and support his legs as high as you can above the level of his heart.

3. Loosen tight clothing, for example at the neck, chest, and waist. Keep his head low; this may prevent him from losing consciousness. Keep the casualty warm; cover him with a blanket or sleeping bag if you have one.

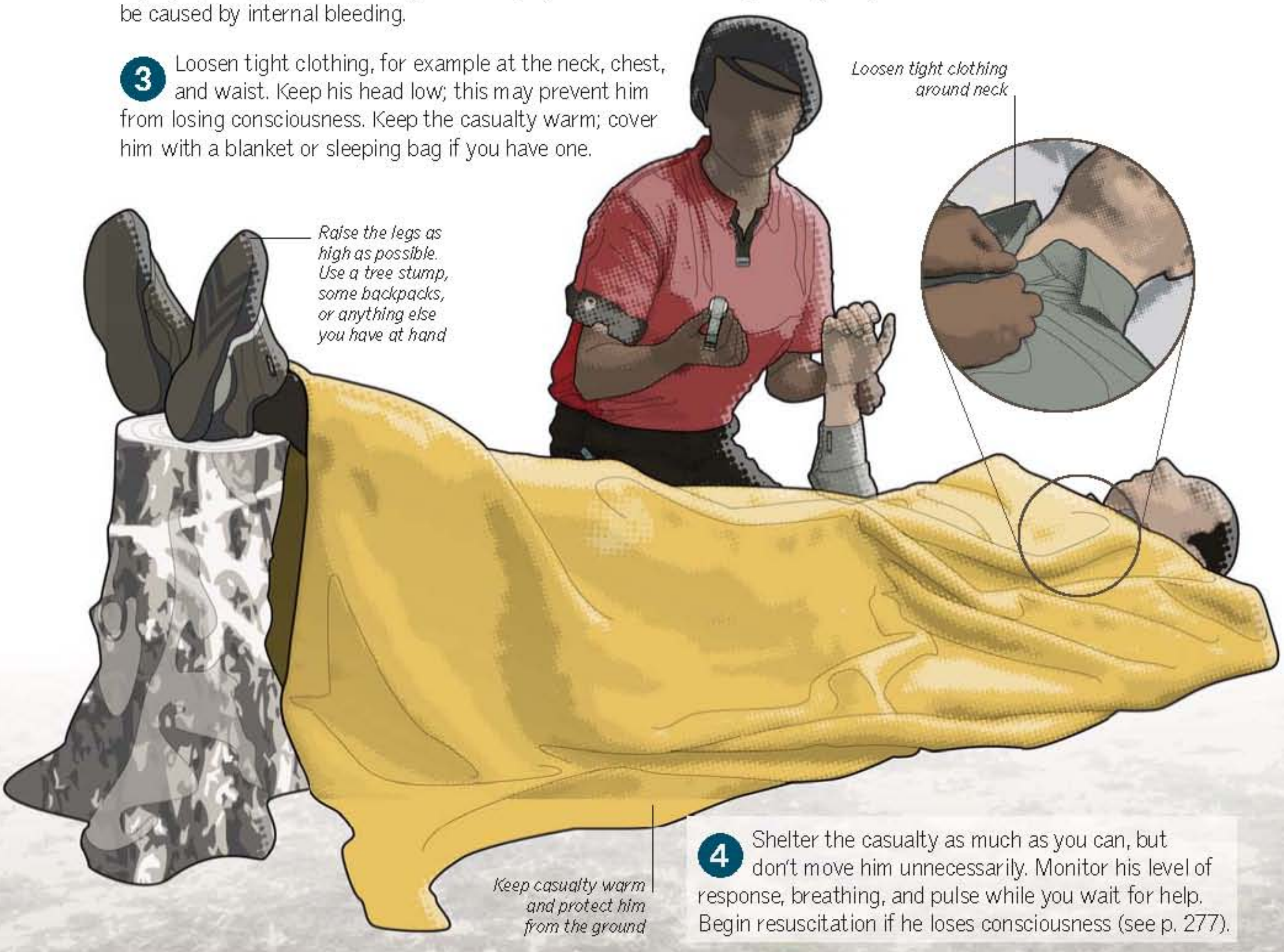

Loosen tight clothing around neck

Raise the legs as high as possible. Use a tree stump, some backpacks, or anything else you have at hand

Keep casualty warm and protect him from the ground

4. Shelter the casualty as much as you can, but don't move him unnecessarily. Monitor his level of response, breathing, and pulse while you wait for help. Begin resuscitation if he loses consciousness (see p. 277).

BREATHING DIFFICULTIES

RESPIRATORY PROBLEMS need prompt treatment because they can prevent sufficient oxygen reaching the body tissues. The cause may be temporary, for example choking, suffocation, or smoke inhalation, or it can be a long-term condition such as asthma that requires medication. You may need to get emergency help.

WARNING!
If at any stage the casualty loses consciousness, open the airway (see p. 276) and check breathing; the throat muscles may relax enough to allow breathing. If he or she is not breathing, begin resuscitation as this may dislodge the blockage (see p. 277).

CHOKING

When an object becomes stuck in the throat it can cause a muscular spasm that blocks the airway. Always ask the casualty if she is choking, to make sure. If she can speak, cough, or breathe, the obstruction is mild and she will probably be able to clear it herself.

1 If the casualty is breathing, tell her to continue coughing. If she can't speak or cough, help her to bend forward. Support her upper body and give up to five back blows in between her shoulder blades with the heel of your hand.

2 If back blows fail, stand behind her. Put your arms around her abdomen, clench one fist and grasp it with the other hand. Pull sharply inwards and upward up to five times. Check the mouth and remove any obvious blockage.

3 If the obstruction has still not cleared, repeat the five back slaps and five abdominal thrusts three times, then call for help if it's not already on the way. Continue until help arrives or the casualty becomes unconscious (see pp. 276–77).

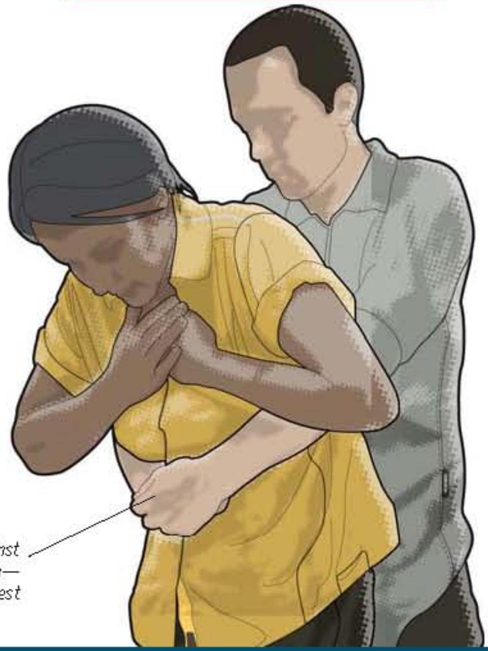
Pull against the abdomen—not the chest

ASTHMA

This is a condition in which breathing becomes difficult because the muscles in the air passages go into spasm. Most people who suffer from asthma carry inhalers with them. Many have two inhalers—a brown or white "preventer" inhaler and a blue "reliever" inhaler to use in an attack. If the person does not have any medication, sit him or her down and call for emergency help.

1 Sit the casualty down and advise her to take a puff from her reliever inhaler. Tell her to take slow, deep breaths and to sit as upright as she can. The attack should start to ease in a few minutes.

2 If the attack does not ease, tell her to take another dose from her inhaler and rest while she recovers. If the attack still does not ease, or is becoming worse, emergency help is needed as she could lose consciousness.

WHAT TO DO IN A SURVIVAL SITUATION

The best advice is prevention—know what triggers the asthma and avoid those things if at all possible. However, should someone have an asthma attack and not have an inhaler, try the following:

- Tell her to exhale as completely as she can. This expels the "stale" air, with little or no oxygen in it. It may be difficult—and will seem strange—to blow air out when the basic instinct is to gulp air in, but this does work.
- Then, tell her to inhale, slowly and steadily, and to close her eyes to help calm herself while doing so.

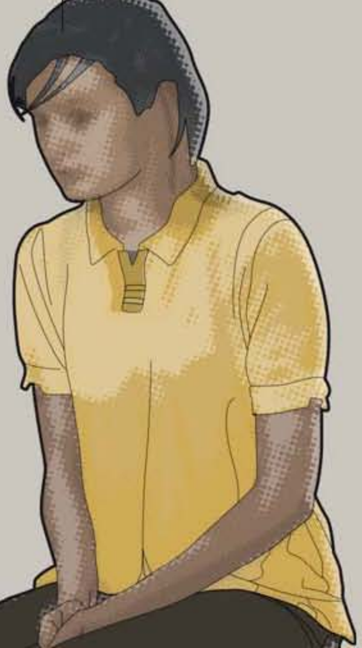
Take slow, deep breaths

UNCONSCIOUSNESS

IF SOMEONE FALLS UNCONSCIOUS, your priority is to make sure their airway is open so that they can breathe. Call for emergency help immediately (ideally ask someone to do this while you treat the casualty). Don't move the casualty and don't leave him or her alone unless you have to go and get help.

CHECK FOR RESPONSE

Gently shake the casualty's shoulders (if it's a child, tap the shoulders). Talk to him and watch for a response. If he's alert, he's conscious. If, for example, he reacts weakly, he may not be fully conscious—monitor him for any change (deterioration or improvement). If there's no response, he's unconscious.

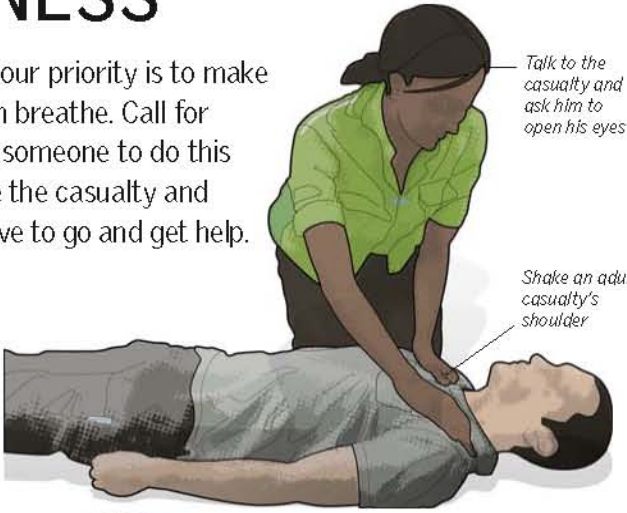

Talk to the casualty and ask him to open his eyes

Shake an adult casualty's shoulder

OPEN THE AIRWAY

If an unconscious casualty is on his back, he's at risk of swallowing his tongue—thereby blocking his air passages. Tilting the head and lifting the chin will "lift" the tongue, clearing the air passage.

CHECK THE BREATHING

Tilt a casualty's head back with one hand and lift the chin with two fingers of your other hand; don't press on the soft tissues under the chin. Keep the airway open, and look, listen, and feel for normal breathing. If the casualty is breathing normally, place him in the recovery position (see below). If he's not breathing, begin chest compressions right away (see opposite).

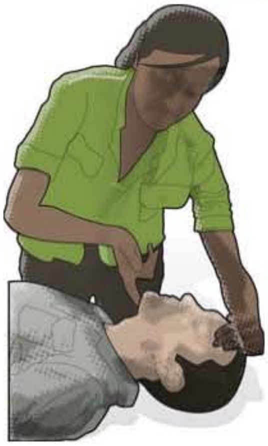

1 Place a hand on the forehead to tilt the head; lift the chin.

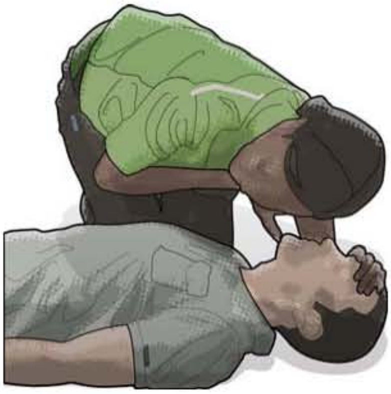

2 Look along the casualty's chest; listen, and feel, for breath against your cheek for no more than ten seconds.

RECOVERY POSITION

If an unconscious casualty is breathing, place him in the recovery position to keep his airway open and clear. Remove anything bulky from his pockets. Kneel beside him. Bend the arm nearest you at a right angle to his body then bring the other arm across his chest until his hand rests against his near cheek, and hold it there. Bend the far leg at the knee, and, still holding the knee, pull the casualty toward you until he is on his side.

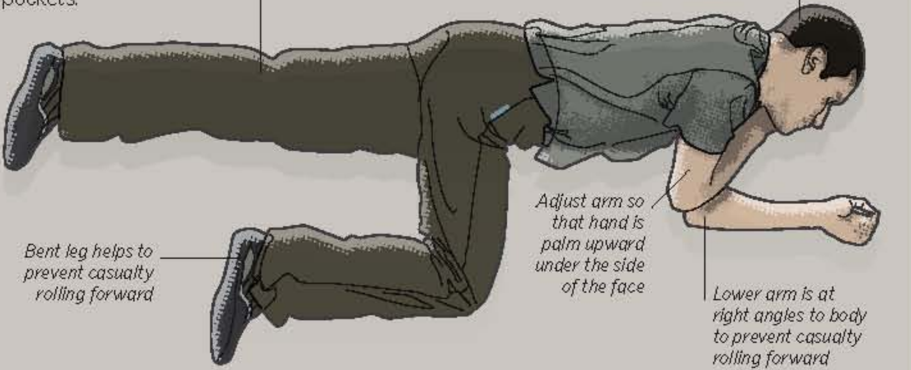

Keep lower leg straight, in line with the spine

Tilt the head back to keep airway open

Bent leg helps to prevent casualty rolling forward

Adjust arm so that hand is palm upward under the side of the face

Lower arm is at right angles to body to prevent casualty rolling forward

CARDIOPULMONARY RESUSCITATION (CPR)

If a casualty is not breathing, you must try to keep the body supplied with oxygen by using chest compressions and rescue breaths until emergency help arrives. This is known as cardiopulmonary resuscitation, or CPR. If an adult collapses, the cause is most likely to be a heart problem so treat as below. If you have rescued an unconscious casualty from water, start with rescue breaths as for a child (see right). If you are unable to achieve rescue breaths, you can give chest compressions alone.

CPR ON A CHILD

- Start by giving the child five rescue breaths.
- Give 30 compressions using the heel of one hand only and slightly less pressure than you would for an adult. Depress the chest by approximately one third of its depth.
- Continue with 30 compressions followed by two rescue breaths until the child recovers, help arrives, or you become exhausted.
- If you are on your own, give CPR for one minute before you try to call for help.

HOW TO GIVE CPR

Kneel beside the casualty, level with his chest so that you don't have to change position. If you have someone else with you, take it in turns to give CPR so you don't become too exhausted. Change over at the end of each two-minute cycle.

1 Put one hand on the center of the casualty's chest—make sure you don't press on the lower abdomen, the tip of the breastbone, or the ribs.

2 Place the heel of your other hand on top of the first and link your fingers together. Keep your fingers off the casualty's chest.

3 Begin chest compressions. Lean over and, keeping your arms straight, press straight down on the casualty's chest, depressing it by 1½–2 in (4–5 cm). Release the pressure and let the chest come back up, but don't move your hands. Repeat 30 times.

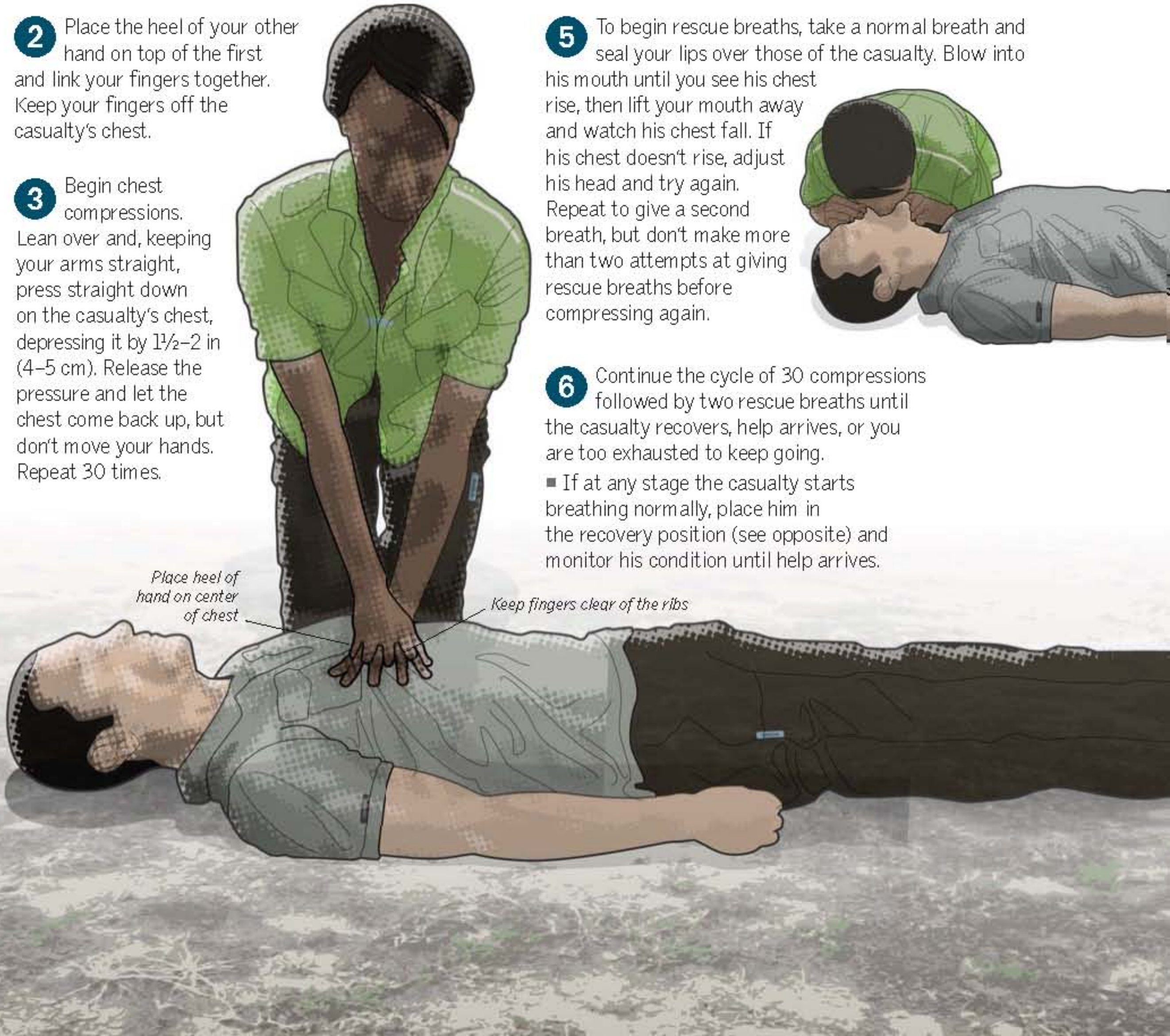
Place heel of hand on center of chest
Keep fingers clear of the ribs

4 Tilt the casualty's head to open the airway and pinch his nose to close the nostrils. Let his mouth fall open slightly. Lift his chin with the fingers of your other hand.

5 To begin rescue breaths, take a normal breath and seal your lips over those of the casualty. Blow into his mouth until you see his chest rise, then lift your mouth away and watch his chest fall. If his chest doesn't rise, adjust his head and try again. Repeat to give a second breath, but don't make more than two attempts at giving rescue breaths before compressing again.

6 Continue the cycle of 30 compressions followed by two rescue breaths until the casualty recovers, help arrives, or you are too exhausted to keep going.

- If at any stage the casualty starts breathing normally, place him in the recovery position (see opposite) and monitor his condition until help arrives.

MOVING A CASUALTY

IDEALLY, AN INJURED PERSON should be treated in the position in which he or she was found. Make the casualty as comfortable as possible and wait for help. In a survival situation, if you have to move someone, it's essential to immobilize an injury first to avoid aggravating it.

WARNING!
Move a casualty only if his life is in immediate danger if he is left where you found him. Even then, only move him if by doing so you will not injure yourself. Don't otherwise attempt to move a casualty, as it will make his injuries worse—stay with the casualty and send others for help.

PREPARING FOR A MOVE

Plan your move before you start. Choose a method appropriate to the injury. Don't attempt to move anyone by yourself if you have help. Encourage the casualty to do as much as possible himself.

FIREFIGHTER'S LIFT

This is a technique that can be used to move a conscious casualty if you're on your own and need to carry an injured person a short distance. It gets its name from the firefighters who originally used it. Don't use it if a casualty has head or facial injuries or a broken arm or leg. To prevent injuring your back, use your legs to power the move.

1 Support the casualty's injury with padding and bandages. Help him to stand up.

- Stand at right angles to the casualty, then squat down in front of him, staying as close to him as possible.
- Your shoulder should be level with the top of the casualty's legs.

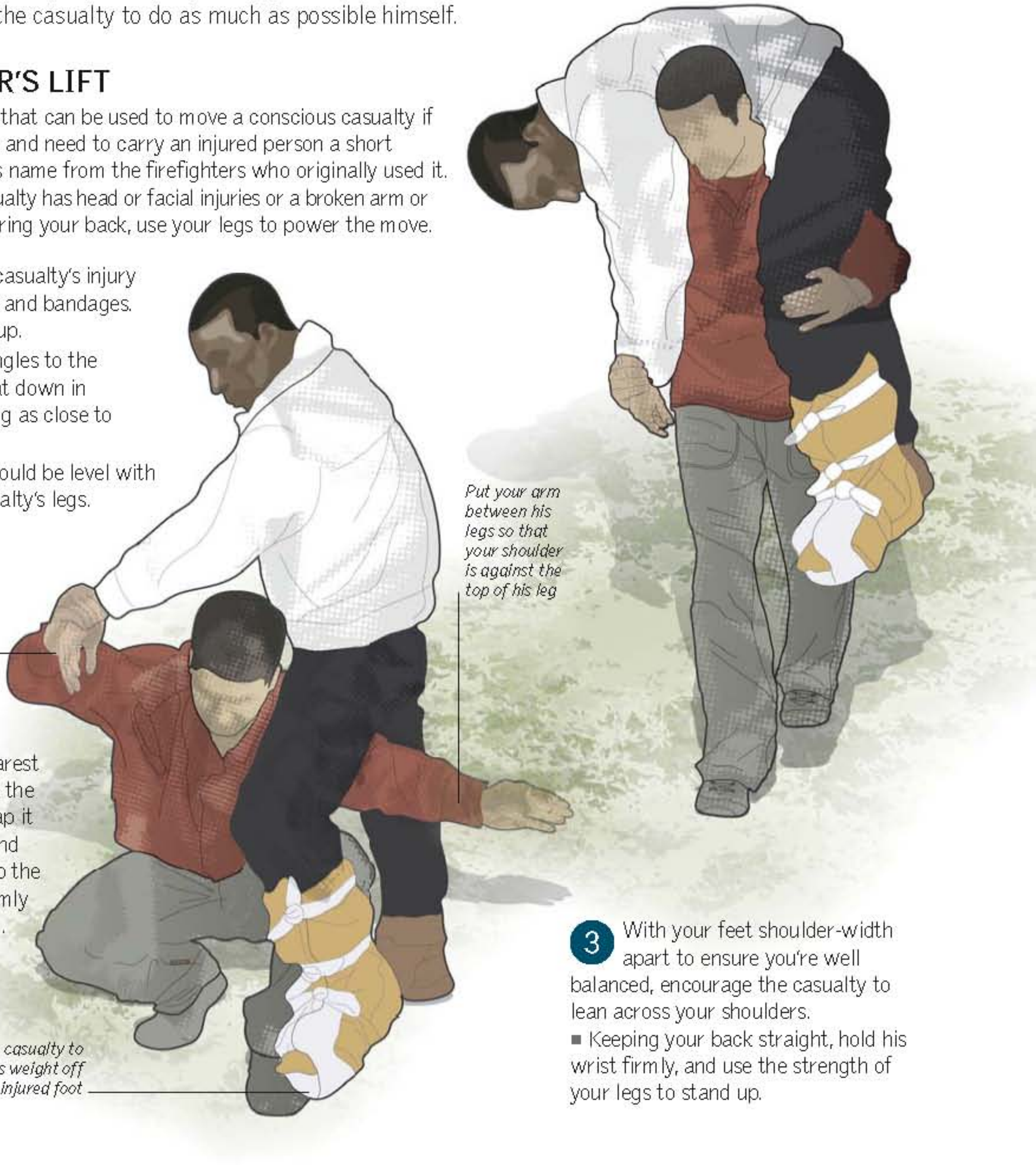

Hold his wrist to help support him

Put your arm between his legs so that your shoulder is against the top of his leg

Tell the casualty to keep his weight off the injured foot

2 Pass your nearest arm between the casualty's legs, wrap it around his thigh, and grasp his leg. Grasp the casualty's wrist firmly with the other arm.

3 With your feet shoulder-width apart to ensure you're well balanced, encourage the casualty to lean across your shoulders.

- Keeping your back straight, hold his wrist firmly, and use the strength of your legs to stand up.

IMPROVISING A STRETCHER

If you need to move a casualty who is unconscious, has a broken leg, or a spine injury, he must be carried by stretcher. Ideally, call the emergency services, and wait for them to arrive. If you can't contact them and you need to get the casualty to medical assistance, you may need to make a stretcher. Use a hurdle or gate, or make the stretcher from poles and coats.

WARNING!
Rescuers should not use their own clothing to make a stretcher if this might put them at risk of exposure.

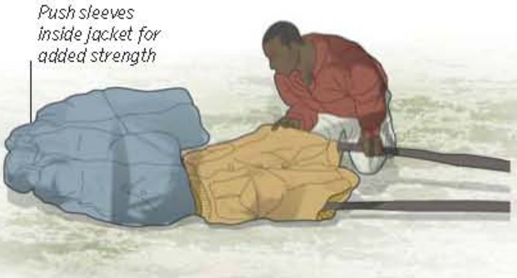
Push sleeves inside jacket for added strength

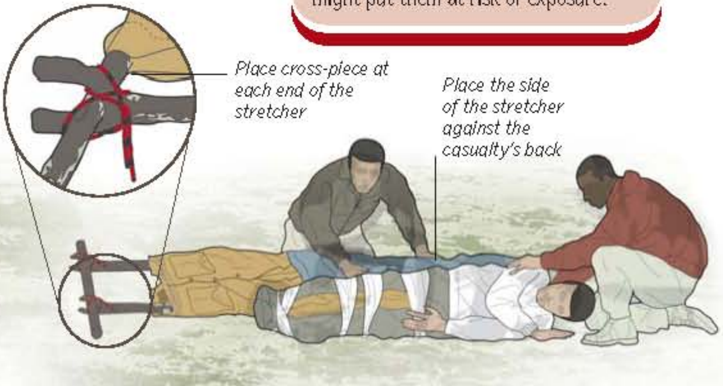
Place cross-piece at each end of the stretcher
Place the side of the stretcher against the casualty's back

1. Zip or button up two or three jackets. Cut two support poles, strong enough to bear the casualty's weight and 3 ft (1 m) longer than the casualty. Lash a short forked branch across each end (see pp. 166–67) to keep the support poles apart.

2. Immobilize any injury (see p. 271), and roll the casualty onto his uninjured side. One person should support his head while others help to keep his body straight, ideally one for the upper body and one for the legs. Slide the stretcher in place then roll him gently back on to it.

TWO-PERSON SEAT

If there are two of you, this method can be used to carry a conscious casualty who can't walk, but who can support himself with his arms. Nominate one person to be in charge of the move and give the instructions. Keep your backs straight at all times.

1. Stand facing each other behind the casualty. Grasp your left wrist with your right hand, then grasp each other's free wrist to make a seat.

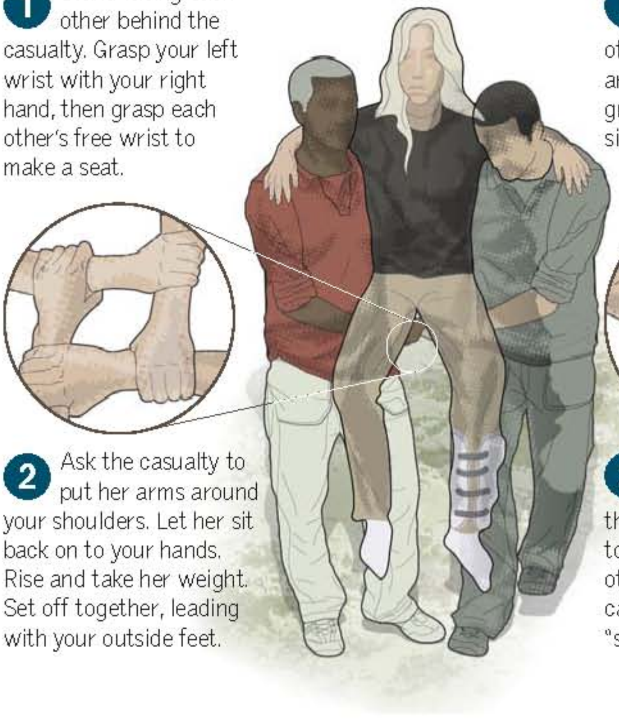

2. Ask the casualty to put her arms around your shoulders. Let her sit back on to your hands. Rise and take her weight. Set off together, leading with your outside feet.

SUPPORTING THE UPPER BODY

If a casualty has an injured arm, the seat carry (see left) can be adapted to support the upper body. As before, one of you should direct the move and always move together. Keep your backs straight as you walk, and stop if it causes you or the casualty any discomfort.

1. Stand facing each other, on either side of the casualty. Put your arms around her back and grasp her clothing on the side farthest from you.

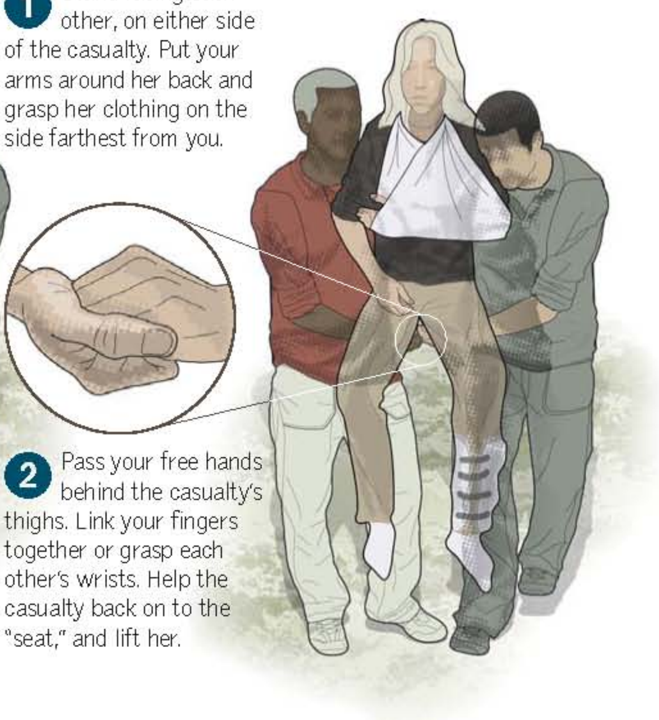

2. Pass your free hands behind the casualty's thighs. Link your fingers together or grasp each other's wrists. Help the casualty back on to the "seat," and lift her.

WILD FOOD

MANY POISONOUS PLANTS in temperate regions resemble edible ones, so you should only eat plants that you can positively identify. If you're in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only at certain stages of their growth. Never eat wild plants if you have any known

NAME	WHERE	IDENTIFICATION	EDIBLE PART
BISTORT/KNOTWEED *Polygonum* spp.	▪ Damp, grassy habitats, such as meadows, in North America and Eurasia.	▪ Perennial plant, 16–40 in (40–100 cm) tall. Green, triangular leaves. Small pink or white flowers form dense, cylindrical flowerheads at end of long, unbranched stem.	▪ Young shoots and leaves ▪ Roots
BRAMBLE/BLACKBERRY *Rubus fruticosus*	▪ In most temperate habitats worldwide, but particularly scrubland, hedges, and woodland.	▪ Scrambling deciduous shrub with thorny stems forming thickets. Green, toothed leaves. White or deep pink five-petaled flowers. Fruit ripens to blue-black in late summer.	▪ Fruit ▪ Young shoots ▪ Leaves
BUCKWHEAT *Fagopyrum esculentum*	▪ Open grassland worldwide.	▪ Red-stemmed plant up to 2 ft (60 cm) tall. Green, triangular leaves. Small pink, five-petaled flowers grow in clusters. Triangular seeds.	▪ Seeds
DANDELION *Taraxacum* spp.	▪ Widely distributed in temperate areas of Eurasia; introduced to the Americas, Australia, and New Zealand.	▪ Rosette of green, jagged-edged leaves grows close to the ground from a central taproot. Large, bright yellow flower head matures into a spherical "clock" of several seeds.	▪ Leaves ▪ Roots
DOG ROSE *Rosa canina*	▪ In hedgerows, woodland margins, and scrubland in Europe, northwest Africa, and western Asia; introduced to other regions.	▪ Scrambling deciduous shrub with spiny stems. Dark green, toothed leaves. Scentless, white, pink or deep pink five-petaled flowers. Fruit, or "hip," is orange-red and oval.	▪ Hips ▪ Buds and flowers ▪ Young leaves
STINGING/ COMMON NETTLE *Urtica dioica*	▪ By hedgerows and woodland margins in North America, Eurasia, and North Africa.	▪ Perennial herbaceous plant, 20–60 in (50–150 cm) tall. Green, oval, toothed leaves covered in stiff stinging hairs. Small green flowers sometimes have a red tinge.	▪ Young shoots and leaves
SUGAR MAPLE *Acer saccharum*	▪ Woods and forests in northeastern North America.	▪ Deciduous tree, usually 82–115 ft (25–35 m) tall. Gray-brown bark smooth on young trees, furrowed and flaky on old trees. Green, five-lobed leaves turn bright red in fall.	▪ Sap ▪ Inner bark
WALNUT *Juglans* spp.	▪ Temperate areas in Eurasia and North America. Common in the Himalayas.	▪ Large deciduous tree up to 82 ft (25 m) tall with distinctive gnarled bark. Green leaves have narrow leaflets. Nut has a thick, green husk, which rots away once it falls to the ground.	▪ Nuts
WATER CHESTNUT AND WATER CALTROP *Trapa natans* and *T. bicornis*	▪ In slow-moving water in warm temperate areas of Eurasia and Africa; introduced to North America and Australia.	▪ Green, triangular, saw-toothed leaves float on surface. White four-petaled flowers. Fruit is borne underwater and has four sharp spines and one hard gray seed.	▪ Seeds
WILD GRAPE *Vitis* spp.	▪ In most habitats in North America and Eurasia.	▪ Sprawling, high-climbing vine with large, lobed leaves. Fruit grows as hanging bunches of berries, which are amber or dark purple when ripe.	▪ Fruit ▪ Young leaves

TEMPERATE PLANTS

allergies, a pre-existing medical condition, or are pregnant. The UET doesn't apply to mushrooms, so never eat any mushrooms unless you can positively identify every single one as being edible. As part of your preparation for your trip, you should familiarize yourself with the edible plants in the region you're visiting, and their seasonal availability.

	PREPARATION	SIMILAR PLANTS	BEWARE OF	EXTRA INFORMATION
	▪ Young leaves and shoots can be eaten raw or boiled. ▪ Soak roots, then boil or roast.	▪ Alpine bistort (*P. viviparum*) is common on rocky soils in northern alpine regions and the Arctic.		▪ Roots are twice twisted and are said to resemble two snakes. They contain starch and can be used to make flour.
	▪ Fruit and peeled young shoots are edible raw. ▪ Infuse the leaves to make tea.	▪ Wild raspberry (*R. idaeus*) fruit ripens to a bright red. Tea made from the leaves can be used to treat diarrhea.	▪ Catching yourself on thorns.	▪ Blackberries and raspberries are high in vitamin C and sugars. Blackberry leaf tea can be used to treat colds and coughs.
	▪ Hull and grind the seeds to make buckwheat flour. ▪ Roast seeds and cook with water to make porridge.		▪ Buckwheat leaves are edible but can cause photosensitivity of the skin if eaten in large quantities.	
	▪ Young leaves can be eaten raw; boil old leaves to remove the bitter taste (change the water once). ▪ Boil the roots.			▪ Roots can be roasted and ground to make a substitute for coffee. The leaves are high in vitamins A and C, iron, and calcium.
	▪ Buds and flowers are edible raw. ▪ Chew the pulp of raw hips, or dry and eat later. ▪ Boil leaves in water to make tea.		▪ The hips of all wild roses (*Rosa* spp.) are edible, but do not eat the seeds if they are prickly as they will irritate the throat and stomach.	▪ Rose hips remain on the shrub all year. They are rich in vitamin C. Boil crushed hips in water then strain to make a nutritious syrup.
	▪ Boil the plants for 10 minutes to destroy the formic acid and histamine in the stinging hairs.	▪ Dead nettles (*Lamium* spp.) are also edible if cooked. They have heart-shaped leaves and white or purple flowers. No stinging hairs.	▪ Gather and handle the plants carefully, wearing gloves if you have them, to avoid being stung by the hairs.	▪ Fresh nettles are rich in protein and vitamin K. The fibers in mature stems can be woven into cordage (see pp. 138–41).
	▪ Boil sap into a high-energy syrup. ▪ Eat inner bark raw or boil into a gelatinous mass, then roast and grind into a flour.	▪ Red maple (*A. rubrum*) also produces sweet sap. ▪ White birch (*Betula* spp.) has edible inner bark.		▪ To collect the sap, cut a V-shape into the trunk, make a hole below the cut, and insert a leaf to guide the drips into a container below.
	▪ Crack the shell and eat the ripe nut raw.			▪ Rich in fat, protein, and vitamins. ▪ Crushed green husks of black walnut (*J. nigra*) are toxic to fish. Use to poison and catch them.
	▪ Seeds can be eaten raw or roasted.			▪ Water chestnuts are a source of carbohydrate.
	▪ Ripe fruit is best eaten raw. ▪ Boil young leaves.		▪ Moonseed vines (*Menispermum* spp.) have grapelike fruits, which are poisonous. Unlike edible grapes, they only have one seed.	▪ Ripe grapes are rich in vitamin C and sugar. Water can be obtained from the vine stem (see p. 192).

WILD FOOD

SOME OF THE PLANTS listed below are good sources of water/liquid. In a survival situation in the desert, don't eat food if you don't have any water, as your body will use up its own water to digest it. Only eat plants that you can positively identify and, if you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07).

NAME	WHERE	IDENTIFICATION
ACACIA/WATTLE *Acacia* spp.	- Africa, southern Asia, Australia, and the Americas.	- Thorny, medium-sized trees with gray-white bark. Green leaves divided into many small, oval leaflets. Small, ball-shaped yellow, white or pink flowers.
AGAVE *Agave* spp.	- Southern USA and Mexico, Central America, the Caribbean, and northeastern South America.	- Rosette of fleshy leaves with sharp tip and spiny margin. Central, tall flower stalk.
AMARANTH *Amaranthus* spp.	- The Americas, Africa, and Asia.	- Tall herb with alternating leaves and erect fleshy stems. Tiny brown/black seeds (one plant produces 40,000–60,000 seeds).
BAOBAB *Adansonia* spp.	- Africa and northern Australia.	- Large trees with swollen trunks, up to 30 ft (9 m) in diameter. Green leaves have 5–7 digit-like leaflets. Oblong fruit up to 8 in (20 cm) long.
CAROB TREE *Ceratonia siliqua*	- Mediterranean, northern Africa, Middle East, and India.	- Evergreen tree up to 50 ft (15 m) tall. Shiny green leaves up to 8 in (20 cm) long. Small red flowers. Flat leathery seed pods; dark-green/black when ripe. Hard brown seeds.
CARRION FLOWER *Hoodia* spp.	- Southwest Africa.	- Numerous thorny, succulent stems up to 6 ft (2 m) tall. Star-shaped or shallowly bell-shaped flowers give off the smell of rotting meat.
DATE PALM *Phoenix dactylifera*	- West and North Africa, Middle East, and India; introduced to Mexico and USA.	- Tall slender palms with a large crown of green leaves, divided into many narrow leaflets. Fruit reddish-brown when ripe.
PRICKLY PEAR CACTUS *Opuntia* spp.	- The Americas; introduced to the Caribbean, the Mediterranean, Africa, and Australia.	- Thick, jointed, flat, pad-like, green stems covered in spines. Red or yellow flowers. Fruit red when ripe.
WILD GOURDS *Cucurbitaceae*	- Kalahari Desert (Namibia), the Sahara, islands in the Mediterranean, the Middle East, and southeastern India.	- Ground-trailing green vine. Bright yellow flowers. Orange-sized green or yellow fruit.

DESERT PLANTS

Don't forget that some plants are edible only at certain stages of their growth. Don't eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. When planning your trip, it's advisable to familiarize yourself with the edible plants in the region you are visiting and their seasonal availability.

	EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
	▪ Seeds (from dark brown pods) ▪ Young leaves and shoots	▪ Roast seeds. ▪ Boil leaves and shoots.	▪ Catching yourself on thorns.	▪ Roots may be tapped for water.
	▪ Stalk, before in flower ▪ Buds and flowers	▪ Roast the stalk. ▪ Boil buds and flowers.	▪ Juice from many species can cause acute contact dermatitis, with reddening and blistering lasting 1–2 weeks. Leaves have needle-like ends.	▪ Roasted stalk tastes sweet, like molasses.
	▪ Young shoots and leaves ▪ Seeds	▪ Eat shoots and leaves raw or boiled or stir-fried. ▪ Remove chaff from seeds, then cook like popcorn or grind into flour to make bread.	▪ Amaranth leaves are high in oxalic acid, which can irritate the gut and cause kidney stones, so eat in moderation.	▪ Amaranth leaves are a common leaf vegetable in the tropics and warm temperate regions.
	▪ Shoots and young leaves ▪ Fruit ▪ Seeds	▪ Eat the fruit pulp raw. ▪ Boil shoots and leaves. ▪ Roast seeds.		▪ Tap the trunk for water. Kalahari bushmen suck water through holes in the trunk using the hollow stems of grasses.
	▪ Pods ▪ Seeds	▪ Eat the sweet, nutritious pulp of the pods raw. ▪ Grind the seeds into flour and make into porridge.		
	▪ Stems	▪ Tap the stems for their water.		▪ The stems contain an appetite suppressant. Kalahari bushmen chew the stems of several *Hoodia* species before and during their long, arduous hunts.
	▪ Fruit ▪ Growing tip of the palm (heart) ▪ Young leaves ▪ Sap	▪ Eat the fruit (dates) raw or sun-dried. ▪ Boil the leaves and heart. ▪ Boil the sap down to a syrup.		▪ Date palms always grow near water. The leaves can be used to thatch a shelter.
	▪ Fruit ▪ Seeds ▪ Pads	▪ Peel fruit and eat raw. ▪ Roast and grind seeds to a flour. ▪ Boil or grill young pads, first removing thorns by peeling or scorching over a fire.	▪ Spines. Avoid any similar-looking plant that has milky sap as it may be poisonous.	▪ The non-milky sap from the pads is a good source of water.
	▪ Fruit (Tsamma melon only) ▪ Flowers ▪ Seeds ▪ Young leaves and shoots	▪ Pound fruit to a pulp and eat raw. ▪ Eat the flowers raw. ▪ Roast or boil seeds. ▪ Chew leaves and shoots for water.	▪ The fruit of the wild desert gourd or colocynth (*Citrullus colocynthis*) is very bitter and is a strong laxative.	▪ The Tsamma melon (*C. lanatus*) is also known as the wild watermelon. The fruit pulp is a good source of water.

WILD FOOD

A GREAT NUMBER OF PLANTS flourish in the tropics, growing all year round in the warm and humid conditions—those listed below are just a few of the most common varieties. When you plan your trip, you should familiarize yourself with the edible plants in the region you are visiting. In tropical forests, most fruits are borne high in the canopy—out of reach,

NAME	WHERE	IDENTIFICATION	EDIBLE PART
BAMBOO *Bambuseae*	▪ Tropical and subtropical forests.	▪ Tree-like grasses with woody, segmented, hollow stems ranging from black to green to gold in color. Green blade-like leaves. Species range in height from 1⅔–11½ ft (0.5–3.5 m).	▪ Young shoots ▪ Seeds
BANANA AND PLANTAIN *Musa* spp.	▪ Native to Australia and Southeast Asia; introduced to other tropical regions.	▪ Tree-like herbaceous plants up to 33 ft (10 m) tall. Large, green, strap-like, split leaves. Flowers and fruit grow in dense hanging clusters.	▪ Fruit ▪ Buds and flowers ▪ Shoots and young stems ▪ Roots
BRAZIL NUT *Bertholletia excelsa*	▪ Along river banks in rain forests in South America.	▪ Large dry-season deciduous tree up to 150 ft (45 m) tall. Green, oblong, crinkly leaves. Yellow flowers. Fruit is the size of a coconut with a hard, woody shell and contains 8–24 seeds.	▪ Seeds or "nuts"
CEYLON/VINE SPINACH *Basella alba*	▪ Widespread in tropical forests.	▪ Trailing, vine-like plant reaching 100 ft (30 m) in length. Beet-red stems. Fleshy, oval or heart-shaped, green or red-purple leaves.	▪ Young leaves and stems
FIG *Ficus* spp.	▪ Variety of habitats in tropical and subtropical regions.	▪ Evergreen tree with long aerial roots growing from the trunk and branches. Leathery green leaves. Pear-shaped fruit grow directly from the trunk or branches.	▪ Ripe fruit
PAPAYA/PAWPAW *Carica papaya*	▪ Native to tropical rain forests in the Americas; introduced to other tropical and some temperate regions.	▪ Small tree up to 20 ft (6 m) high, with a soft hollow trunk. Large green leaves have seven lobes. Large melon-like fruit turn yellow or orange when ripe.	▪ Fruit ▪ Young flowers, leaves, and stems
PEANUT *Arachis hypogaea*	▪ Native to tropical rain forests in the Americas; introduced to other tropical and some temperate regions.	▪ Small, bushy plant up to 1⅔ft (50 cm) tall. Paired, oval green leaves grow four to a stem. Yellow flowers. Underground legume (pod) contains 1–4 nuts.	▪ Seeds or "nuts"
SAGO PALM *Metroxylon sagu*	▪ Damp lowlands of tropical rain forests in Southeast Asia; introduced elsewhere.	▪ Spiny-trunked palm up to 33 ft (10 m) tall. Crown of pinnate (feather-like) green leaves.	▪ Pith ▪ Young shoots ▪ Young nuts
WATER LILY *Nymphaea*	▪ Lakes, ponds, and rivers in tropical and subtropical regions worldwide. Also found in temperate regions.	▪ Flat, green, heart-shaped leaves float on surface of water. Large, white, yellow, pink, or blue fragrant flowers.	▪ Tubers ▪ Stems ▪ Seeds

TROPICAL PLANTS

unless you can climb the tree. Only eat plants that you can positively identify. If you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only at certain stages of their growth. Don't eat wild plants if you have a known allergy, a preexisting medical condition, or are pregnant.

PREPARATION	SIMILAR PLANTS	BEWARE OF	EXTRA INFORMATION
▪ Split the tough outer sheath of the shoots, and boil or steam. ▪ Boil the seeds or grind, mix with water, make into cakes, and bake.		▪ Shoots of the giant bamboo or volohosy (*Cathariostachys madagascariensis*) contain prussic acid (cyanide).	▪ Bamboo stems often hold water (see p. 190). The stems can be used for cooking (see p. 204) and building a shelter (see p. 169).
▪ Fruit and flowers are edible raw or cooked. ▪ Cook shoots, stems, and roots.			▪ Banana and plantain fruit is rich in potassium and vitamins A, B6, and C. Hard, unripe plantain fruit is only edible when cooked.
▪ Break fruit capsule open, crack the nut's shell, and eat raw.	▪ The sapucaia tree (*Lecythis* spp.) produces edible seeds in a similar fruit capsule. They are called paradise nuts.		▪ Ripe fruit capsules fall to the ground. Rodents and monkeys are attracted to the nuts, which are rich in fats and selenium.
▪ Steam, boil, stew, or stir-fry the leaves and stems.		▪ Don't eat too much, as the plant can have a mild laxative effect.	▪ Young leaves and stems are rich in vitamins.
▪ Eat the fruit raw or cooked.		▪ Don't eat any hard, woody fruit, or those covered in hairs.	▪ Edible figs are soft when ripe and green, red, or black in color.
▪ Flesh of ripe fruit is edible raw. Unripe fruit is edible cooked. ▪ Flowers, leaves, and stems must be boiled, and the water must be changed at least once.		▪ Don't get the milky juice of green, unripe fruit in your eyes as it will cause temporary blindness.	▪ Ripe fruit is rich in vitamin C. Unripe fruit will ripen quickly if placed in the sun.
▪ Remove the shell and eat the nuts raw.		▪ Don't handle or eat if you are allergic to peanuts.	▪ Peanuts are a good source of protein, B vitamins, and minerals.
▪ Crush the pith, knead it in water, strain into a pot, and let the fine sago settle. Squeeze and let it dry. Mix to a paste with boiling water.	▪ Many palms have edible parts. The sugar palm (*Arenga pinnata*) has an edible sap which can be boiled down to a thick syrup.	▪ Don't eat a palm's fruit unless you have positively identified it as an edible variety as the fruit of some species contain harmful crystals.	▪ A mature palm will yield 330–660 lb (150–300 kg) of sago starch—almost pure carbohydrate.
▪ Peel and slice the tubers, which can be eaten raw. ▪ Stems are best cooked. ▪ Dry and grind the seeds into flour.	▪ All parts of the lotus (*Nelumbo* spp.) are edible raw or cooked.		▪ The tubers are rich in starch. The seeds have a bitter taste but they are edible.

WILD FOOD

WHEN PREPARING FOR YOUR TRIP, you should familiarize yourself with the edible plants in the region you are visiting and their seasonal availability. Only eat plants that you can positively identify and, if you are in any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't forget that some plants are edible only

	NAME	WHERE	IDENTIFICATION
LICHENS	**ICELAND MOSS** *Cetraria islandica*	▪ Mountainous areas in Arctic, sub-Arctic, and cold temperate regions in North America and Europe; lava slopes and plains of Iceland.	▪ Mat-forming lichen, up to 4 in (10 cm) high. Gray-green to pale chestnut branches, rolled into tubes terminating in flattened lobes with fringed edges.
	REINDEER MOSS *Cladonia rangiferina*	▪ Tundra, bogs, and open woodlands in Arctic, sub-Arctic, and northern temperate regions.	▪ Mat-forming lichen, 2–4 in (5–10 cm) high. Gray, rounded branches resemble antlers.
	ROCK TRIPE *Umbilicaria* spp.	▪ On rocks in Arctic, sub-Arctic, and northern temperate regions.	▪ Rounded lichens with curling edges, usually gray or brown.
PLANTS	**ARCTIC WILLOW** *Salix arctica*	▪ Tundra in North America, Europe, and Asia; mountainous areas in some northern temperate regions.	▪ Mat-forming shrub, 1–2 ft (30–60 cm) high. Rounded, shiny green leaves. Flowers grow as yellow catkins.
	BEARBERRY *Arctostaphylos uva-ursi*	▪ Mountainous areas in Arctic and sub-Arctic regions.	▪ Dwarf evergreen shrub. Thick, leathery, club-shaped green leaves. Pink or white flowers. Bright red berries grow in clusters.
	BOG LABRADOR TEA *Rhododendron groenlandicum*	▪ Bogs and alpine areas in Arctic, sub-Arctic, and temperate regions in North America and Europe.	▪ Evergreen shrub, 1–3 ft (30–90 cm) high. Narrow, leathery, green leaves are hairy underneath and fragrant. Small, fragrant, white flowers grow in clusters and are sticky.
	CLOUDBERRY *Rubus chamaemorus*	▪ Bogs, marshes, and wet meadows in alpine and sub-Arctic regions of North America, Europe, and Asia.	▪ Perennial shrub, 4–10 in (10–25 cm) high. Soft, green leaves have 5–7 lobes. White flowers have five petals. Rasberry-sized, amber berries grow at the top of the plant.
	CROWBERRY *Empetrum nigrum*	▪ Tundra, moorlands, bogs, and spruce forests in Arctic, sub-Arctic, and northern temperate regions. Also found in the Andes.	▪ Dwarf, evergreen, mat-forming shrub. Short, pale green, needle-like leaves. Small, purple-red flowers. Black berries.
	KERGUELEN CABBAGE *Pringlea antiscorbutica*	▪ Rocky areas of sub-Antarctic islands in the Indian and Southern Oceans.	▪ Cabbage-like plant.

PLANTS IN COLD CLIMATES

at certain stages of their growth. Don’t eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. There are no poisonous types of lichen (althought see rock tripe entry, below), but they must all be soaked in water overnight and boiled well before they can be eaten.

	EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
	▪ All	▪ Soak for several hours then boil well.	▪ All lichens contain an acid that will cause stomach irritation unless soaked and boiled before eating.	▪ Lichens are low in protein but high in carbohydrates (lichen-starch).
	▪ All	▪ Soak for several hours then boil well.	▪ All lichens contain an acid that will cause stomach irritation unless soaked and boiled before eating.	▪ High in vitamins A and B. Important source of food for reindeer (caribou). Partially digested lichens from the reindeer’s stomach are considered a delicacy by herders.
	▪ All	▪ Soak for several hours then boil well.	▪ There have been some reports of rock tripe poisoning so apply the Universal Edibility Test first.	
	▪ Young shoots ▪ Leaves ▪ Young roots	▪ Peel off outer bark and eat inner shoots raw. ▪ Eat the leaves raw. ▪ Peel roots and eat raw.		▪ Arctic willow leaves have 7–10 times more vitamin C than an orange.
	▪ Berries	▪ Cook before eating.	▪ Bears feeding on the berries.	▪ Forms mats in Arctic regions.
	▪ Leaves	▪ Infuse to make tea.		▪ Pick individual leaves rather than whole branches and harvest from several shrubs. Northern Labrador tea (*R. tomentosum*) leaves also make a fragrant tea. The plant grows in peaty soils and on tundra.
	▪ Berries	▪ Edible raw.	▪ Bears feeding on the berries.	▪ Berries initially pink, ripening to amber in fall. High in vitamin C.
	▪ Berries	▪ Eat raw or cooked.	▪ Bears feeding on the berries.	▪ Previous year’s berries will last on plant until spring if not picked. Fresh berries can be dried for storage. Low in vitamins.
	▪ Leaves	▪ Boil well.		▪ Leaves have a bitter taste so must be well-boiled. High in potassium and vitamin C. Antarctic explorers and whalers used to eat Kerguelen cabbage to ward off scurvy, hence its scientific name.

WILD FOOD

THERE ARE NO POISONOUS SEAWEEDS, but some may cause gastrointestinal upset. The seaweeds listed below are common and safe to eat if gathered when still growing, but should only be eaten in small quantities at first. Don't eat seaweed if you are short of fresh water to drink. Only eat plants that you can positively identify and, if you are in

	NAME	WHERE	IDENTIFICATION
SEAWEEDS	**GUTWEED** *Enteromorpha* spp.	▪ Saltmarshes and rock pools in cold-temperate waters worldwide.	▪ Pale or bright green, tubular, unbranched stems, 8–16 in (20–40 cm) long.
	KELP *Alaria* spp., *Laminaria* spp., and *Macrocystis* spp.	▪ Rocky coasts of the Atlantic and Pacific oceans.	▪ Very long, strap-like, olive-green to brown fronds. Giant kelp (*M. pyrifera*) is the largest seaweed on Earth, reaching lengths of 150 ft (45 m).
	LAVER *Poryphyra* spp.	▪ Rocky coasts worldwide.	▪ Very thin, irregularly shaped membranous frond, up to 20 in (50 cm) long. Color varies from olive-green to purple-brown or blackish.
SEASHORE PLANTS	**COCONUT PALM** *Cocos nucifera*	▪ Sandy, rocky, and coral coasts in tropics and subtropics worldwide.	▪ Woody perennial tree, up to 72 ft (22 m) tall, with a crown of pinnate (feather-like) leaves. Gray trunk is ringed with growth scars.
	GLASSWORT/ MARSH SAMPHIRE *Salicornia* spp.	▪ Salt marshes and coastal mudflats of western and eastern North America, western Europe, and the Mediterranean.	▪ Bright green, jointed, fleshy stems, 4–12 in (10–30 cm) high. Scale-like leaves and tiny flowers are sunk into the stems.
	ORACHE/SALTBUSH *Atriplex* spp.	▪ Sandy and shingle beaches worldwide.	▪ Sprawling plant with spikes of small, greenish flowers. Pale- or silvery-green leaves are either triangular or spear-shaped, sometimes lobed.
	SCREW PINE *Pandanus* spp.	▪ Tropical coasts from Madagascar to southern Asia and islands in the southwestern Pacific.	▪ Tree up to 30 ft (9 m) tall, supported by stilt-like aerial roots. Strap-like, saw-toothed leaves are grouped in spirals. Large, globular knobbly fruit ripens from green to orange or red.
	SCURVY-GRASS *Cochlearia* spp.	▪ Salt marshes and rocky coasts of northern North America and northern Europe and Asia.	▪ Creeping plant, 4–16 in (10–40 cm) high, with thick, fleshy, dark green, heart-shaped leaves. Small, white, four-petaled flowers.
	SEA BEET *Beta vulgaris maritima*	▪ Salt marshes, shingle beaches, and cliffs in Europe.	▪ Sprawling plant up to 40 in (1 m) high, with glossy green, stems and leaves often tinged red. Tiny green flowers grow in clusters on the stems.
	SEAROCKET *Cakile* spp.	▪ Sandy shores in North America, Europe, Asia, and Australia.	▪ Up to 16 in (40 cm) high with deeply lobed, fleshy, green leaves. Pale lavender, four-petaled flowers.

COASTAL PLANTS

any doubt at all, carry out the Universal Edibility Test (UET) first (see pp. 206–07). Don't eat wild plants if you have any known allergies, a preexisting medical condition, or are pregnant. When planning your trip, it is advisable to familiarize yourself with the edible plants in the region you are visiting and their seasonal availability.

EDIBLE PART	PREPARATION	BEWARE OF	EXTRA INFORMATION
▪ All	▪ Eat raw or dry in the sun and grind to a powder, for ease of storage.		▪ Gutweed is a good indicator of where fresh water runs across a seashore as it is particularly abundant in such sites.
▪ Fronds	▪ Best boiled, though young fronds can be eaten raw.	▪ Large doses of iodine can be harmful.	▪ Sugarwrack (*L. saccharina*) tastes sweet, as its name implies, and can be eaten fried or boiled. Kelp is high in iodine, small amounts of which are needed for good health.
▪ All	▪ Boil, then mash.		▪ The Welsh make a cake of out of laver and oatmeal, called laverbread. The Japanese produce thousands of tons of paper-thin sheets of dried, processed laver, called nori.
▪ Seeds (coconuts)	▪ Drink the milk fresh. ▪ Eat the flesh raw or sun-dried.	▪ Mature coconut milk needs extra water to be digested. It's also a laxative, so don't drink too much.	▪ Green, unripe coconuts are a good source of water (see p. 190). Coconut milk is rich in sugar and vitamins and contains protein. The oily flesh is also nutritious.
▪ Stems	▪ Steam or boil.		▪ Also known as Poor Man's Asparagus. Rock samphire (*Crithmum maritimum*), which is found on cliffs and rocky shores, is unrelated but has edible succulent leaves.
▪ Leaves	▪ Young leaves can be eaten raw, older leaves should be boiled.		▪ *Atriplex* species are restricted to saline soils. As well as coasts, they are also found inland on the shores of alkaline lakes and in deserts. Gray saltbush (*A. cinerea*) is a type of bushfood in Australia.
▪ Fruit	▪ Inner flesh of ripe fruit can be eaten raw. Partly ripe fruit can be eaten if baked for two hours.	▪ The leaves have small, thorn-like spines at the edges, which can scrape and irritate the skin.	▪ Unripe, green fruit is inedible.
▪ Leaves	▪ Eat the leaves raw or pulp and drink.		▪ Scurvy-grass leaves are rich in vitamin C— sailors used to eat (or drink) them to prevent scurvy. They are very bitter, so are best leached in water, but you should eat them in a survival situation.
▪ Leaves	▪ Eat raw or boiled.		▪ Sea beet is the wild ancestor of beets, sugar beets, and Swiss chard.
▪ Leaves ▪ Young pods	▪ Leaves and young pods can be eaten raw.		▪ Searocket leaves have a peppery taste. One species of *Cakile* grows in the deserts of the Arabian Peninsula.

WILD FOOD

ALL MAMMALS ARE EDIBLE, but some species or groups of mammals are in danger of extinction and are protected by law; the mammal groups listed below contain many species that are common in their areas of distribution, and are not usually dangerous if approached and handled correctly. However, when preparing for your trip, you must

	NAME	WHERE	IDENTIFICATION
AERIAL	**BATS** *Chiroptera*	- Temperate and tropical regions worldwide.	- The only mammals with true wings and the ability to fly. Wingspans vary from 6 in (15 cm) to over 5 ft (1.5 m) in the large fruit bats, or flying foxes (*Megachiroptera*).
TREES / GROUND	**TREE SQUIRRELS** *Sciurus*	- North and South America, Europe, and temperate parts of Asia.	- Small or medium-sized rodents, with large, bushy tails.
	PORCUPINES *Hystricomorpha*	- North and South America, Africa, and tropical Asia.	- Large, rounded rodents, with a coat of sharp quills.
GROUND / BURROWS	**ANTELOPE AND DEER** *Bovidae* and *Cervidae*	- North and South America, Africa, and Eurasia.	- Hoofed mammals, with long legs and barrel-shaped bodies. Male antelope have permanent horns; male deer have antlers, which they shed and regrow each year.
	GUINEA PIGS *Cavia* spp.	- Northwestern, central, and southeastern South America.	- Small rodent, with dark, coarse fur, short legs, and no tail.
	HEDGEHOGS *Erinaceinae*	- Europe, Africa, and Asia; introduced to New Zealand.	- Small, short-legged insectivorous mammal, with spines on back and sides of body.
	KANGAROOS *Macropus* spp.	- Australia.	- The largest of the marsupials (pouched mammals), with long, strong tails, large back legs, and short forelegs.
	RABBITS AND HARES *Leporidae*	- Most habitats worldwide, from arctic tundra to semi-desert.	- Small or medium-sized herbivorous mammals, with large ears and small, round, furry tails. Fur is usually brown or dark gray; the Arctic hare (*Lepus arcticus*) has a winter coat: white with black ear tips.
GROUND / AQUATIC	**BEAVER** *Castor* spp.	- North America and northwestern Eurasia.	- Large semi-aquatic rodents, with coarse, brown fur, webbed feet, and a flat, scaly tail.
	CANE RATS *Thryonomys* spp.	- Sub-Saharan Africa.	- Large rodents, with coarse, brown fur and a long, naked tail. The greater cane rat (*T. swinderianus*) is semi-aquatic.

MAMMALS

familiarize yourself with those mammals that can be hunted in the region you are visiting and their seasonal movements. Some countries and/or states will require you to have a hunting license or permit. For trapping and preparation methods, see pp. 216–23, and for cooking suggestions, see pp. 204–05.

HOW TO FIND	BEWARE OF	EXTRA INFORMATION
▪ Active at night. Roost in colonies in caves and trees during the day. Temperate species hibernate in the winter.	▪ Sharp teeth. Some bats carry rabies or other diseases (see pp. 300–01), so cook the meat thoroughly. Bats are protected in Europe and some parts of North America.	▪ Large, plump, fruit-eating bats, such as the Indian flying-fox (*Pteropus giganteus*), are considered a delicacy in many tropical areas.
▪ Mostly active by day, feeding in branches and on the ground on shoots, nuts, and birds' eggs. Nest in trees in a range of woodland habitats.	▪ Sharp teeth and claws.	▪ Flying squirrels (*Pteromyini*) are nocturnal, tree-living members of the squirrel (*Sciuridae*) family. They glide from tree to tree rather than fly.
▪ Some New World porcupines climb trees to feed, but all Old World porcupines spend all their time on the ground across a range of habitats.	▪ Barbed quills (see p. 266 for how to remove).	▪ Slow-moving, so can easily be run down and speared. Unlike most small game, the meat is fatty rather than lean.
▪ Antelope are mainly found in savanna, or marshland habitats, and deer mainly in woodland, forests, or tundra.	▪ Horns or antlers.	▪ Most active at dawn and dusk. Usually live in groups.
▪ Feed mostly at dawn and dusk, in shrubby grassland in mountainous regions.	▪ Sharp teeth.	▪ Have communal feeding runways. Traps can be baited with leafy vegetables.
▪ Habitats range from woodland, hedgerows, and grassland to desert. Feed at night on small animals, such as worms and insects.	▪ Sharp spines. Usually infested with parasites, so handle carefully and cook well.	▪ Traditionally prepared for eating by covering with clay and baking in embers of a fire (see p. 205).
▪ Live in open savanna woodland, feeding mainly at night on vegetation. Gather at waterholes in times of drought.	▪ Sharp claws and a powerful kick.	▪ Kangaroos are protected in some Australian states, but can be hunted (if you have a permit) in others.
▪ Rabbits live in burrows, often in large numbers, coming above ground to feed on vegetation, using well-worn runs. Hares live above ground.	▪ Rabbits and hares may be infected with germs, so handle them carefully and cook well. Their meat is very lean, and must be eaten with green vegetables (see p. 220).	▪ In most areas, these are the first animals to try to trap. The European rabbit (*Oryctolagus cuniculus*) is an invasive pest species in Australia.
▪ Look for the distinctive mud-and-stick lodges in the middle of a pond or lake. Beavers leave the lodge at night to feed on water plants and trees along the riverbank.	▪ Strong, chisel-like teeth. The Eurasian beaver (*Castor fiber*) is protected in many of the countries in which it is found.	▪ Use regular runs along streams. The tail can be eaten as well as the flesh.
▪ Feed at night on reeds and grasses in marshland and on riverbanks, or on grasses in moist savanna, and on rocky hillsides.	▪ Sharp teeth.	▪ Cane rats are a valuable source of bush meat in West and Central Africa. "True" rats (*Rattus* spp.) are edible but often carry diseases, so handle with care.

WILD FOOD

ALL BIRDS ARE EDIBLE, although a few taste horrible (birds-of-paradise, for example), and pitohuis have poisonous feathers and skin. Some species or groups of birds are in danger of extinction and are protected by law; the bird groups listed below contain many species that are common in their areas of distribution. When planning your trip,

NAMES	WHERE	IDENTIFICATION
GAMEBIRDS *Galliformes*	▪ Worldwide, except Antarctica.	▪ Typically have plump bodies, small heads, and short, rounded wings. Range in size from quail, which can weigh as little as 2 oz (50 g), to the wild turkey (*Meleagris galloparvo*), which can be as heavy as 22 lb (10 kg).
OSTRICH *Struthio camelus*	▪ From West to East Africa south of the Sahara, and southern Africa.	▪ The world's tallest and heaviest bird, reaching 220 lb (100 kg), the ostrich is flightless. Long neck and long, strong legs, with two-toed feet. Males have black and white plumage, females brown.
OWLS *Strigiformes*	▪ Every continent, except Antarctica.	▪ Upright posture and a flattened face with large, forward-facing eyes. Strong, hooked bill and sharp talons (claws).
PIGEONS AND DOVES *Columbiformes*	▪ Worldwide, except the polar regions.	▪ Plump, full-breasted bodies, with a small head and bill. Thick, soft plumage, which can be brown or gray, or brightly colored in some tropical species.
STORKS AND HERONS *Ciconiiformes*	▪ Worldwide, except the polar regions.	▪ All have large bodies, with long legs, long necks, and bills. Range height from 10 in (25 cm) for the smallest bitterns to 5 ft (1.5 m) for the largest storks.
VULTURES *Accipitridae* and *Cathartidae*	▪ All continents, except Antarctica and Australasia.	▪ Large carrion-eating birds, with large wings, powerful feet with sharp talons, and a strongly hooked bill. Head and neck is often bald or sparsely feathered.
WADERS AND GULLS *Charadriiformes*	▪ Worldwide.	▪ Vary greatly in size and shape, but most have subdued brown, gray, black, or white plumage. Waders (or shorebirds) usually have lightweight bodies and long legs. Gulls have compact bodies.
WATERFOWL *Anseriformes*	▪ Worldwide, except Antarctica.	▪ Typically have plump bodies, powerful wings, and short legs with webbed feet. Most species have a broad, flattened bill. Ducks usually have short necks, while geese and swans have long necks.

BIRDS

you must familiarize yourself with those birds that can be hunted in the region you are visiting and their seasonal availability. Some countries and/or states will require you to have a hunting license or permit. For trapping and preparation methods, see pp. 226–29; for cooking suggestions, see pp. 204–05.

HOW TO FIND	BEWARE OF	EXTRA INFORMATION
▪ Wide range of terrestrial habitats, from high mountains, to tropical forests, to Arctic tundra.	▪ Short but sharp beak and claws. In some species (including pheasants), the males have sharp spurs on their ankles.	▪ Nearly all gamebirds feed and nest on the ground. At night, many gamebirds roost in trees.
▪ Open semi-arid plains (from desert to savanna) and open woodland.	▪ Can run fast and deliver a powerful kick. Very protective of eggs in the nest and the young birds.	▪ The eggs are the largest of any bird, weighing 3 lb (1.4 kg). The nests are communal and contain up to 40 eggs.
▪ In most habitats, from tundra to dense forest.	▪ Sharp bill and talons.	▪ Most species hunt at night and roost in trees during the day. Usually nest in holes in trees, under rocky overhangs, or, sometimes, in burrows.
▪ Variety of forest habitats, from open woodland to dense, tropical rain forest, and grassland and semi-arid areas.		▪ Often feed in flocks on the ground or in trees, and usually roost communally in trees. Can be taken by hand from the roost if you approach slowly and quietly.
▪ Freshwater habitats, ranging from wetlands, rivers, lakes, marshes, swamps, mangroves, and lagoons to tidal mudflats.	▪ Sharp bill.	▪ Usually feed alone at the water's edge, but often nest communally in trees.
▪ Open areas, including mountains, plains, deserts, and savannas.	▪ Sharp bill and talons. Handle as little as possible, as prone to parasites and infection. Boil meat for at least 30 minutes.	▪ Carrion eaters and scavengers, such as crows (*Corvus* spp.) and gulls (see below), can be lured into traps baited with meat.
▪ Tundra and a variety of wetland and coastal habitats, including tidal mudflats, beaches, and cliffs.	▪ Sharp bill. Gulls and terns will aggressively defend their nest sites.	▪ Most birds in this group nest on the ground or on rock ledges, laying about four well-camouflaged eggs. Often nest in large colonies.
▪ Arctic tundra, wetlands, rivers, and lakes. Ducks and swans usually feed on the water, while geese graze on the land.	▪ Geese and swans can be very aggressive, particularly in the breeding season. Mute swans (*Cygnus olor*) can weigh up to 26 lb (12 kg).	▪ Many species migrate between their breeding grounds in the Arctic and their wintering grounds further south. They molt their feathers in the late summer, rendering them flightless for a couple of weeks and thus easier to catch.

WILD FOOD

SOME AMPHIBIAN AND REPTILE SPECIES are at risk of extinction and are protected by law; the animals listed below are common in their areas of distribution, and are non-venomous. As part of preparation for your trip, it is advisable to familiarize yourself with the reptiles and amphibians in the region you are visiting. Many countries and/or states

	NAME	WHERE	IDENTIFICATION
FROGS	**AFRICAN CLAWED FROG** *Xenopus* spp.	▪ Sub-Saharan Africa; introduced to North America, South America, and Europe.	▪ Flattened, brown body, 2¼–5 in (6–13 cm) long, with a line of white "stitch marks" along the sides. Clawed toes on the muscular legs.
	AMERICAN BULLFROG *Rana catesbeiana*	▪ North America; introduced to Europe and Asia.	▪ Body is green with brown markings above and white below, 3½–8 in (9–20 cm) long. Large legs and large eardrums.
	SOUTH AMERICAN BULLFROG *Leptodactylus pentadactylus*	▪ Central and northern South America.	▪ Smooth yellow or pale brown body with some dark markings, 3–9 in (8–22 cm) long.
TURTLES	**AFRICAN HELMETED TURTLE** *Pelomedusa subrufa*	▪ Sub-Saharan Africa.	▪ Flattened, brown shell, 8–12½ in (20–32 cm) long.
	ASIAN LEAF TURTLE *Cyclemus dentata*	▪ Southeast Asia.	▪ Oval, light to dark brown shell, 6–9½ in (15–24 cm) long, with serrated edge near tail. Reddish brown head and legs.
	PAINTED TURTLE *Chrysemys picta*	▪ Southern Canada, USA, and northern Mexico.	▪ Flattened, smooth shell, 6–10 in (15–25 cm) long; brown above and yellow below, sometimes patterned. Yellow or red stripes on the neck.
SNAKES	**CARPET PYTHON** *Morelia spilota*	▪ Indonesia, New Guinea, and Australia.	▪ Several subspecies, all of which have a bold pattern of irregular markings, which can be reddish brown, brown, black or gray. Average length is 6½ ft (2 m), but can reach 13 ft (4 m).
	COMMON BOA *Boa constrictor*	▪ Central America, South America, and some Caribbean islands.	▪ Several subspecies but all have characteristic dark saddle markings along the back and a dark stripe behind each eye. Narrow head and pointed snout.
	COMMON EGG-EATING SNAKE *Dasypeltis scabra*	▪ Sub-saharan Africa	▪ Reddish-brown or gray body with dark, angular markings. A slender snake, 28–39 in (70–100 cm) long. Rounded snout.
	COMMON RATSNAKE *Elaphe obsoleta*	▪ Southern Canada, and central and eastern USA	▪ Subspecies coloration varies from bright yellow-orange to pale gray with darker blotches. Length range is 4–6 ft (1.2–1.8 m). Long head with rounded snout.
LIZARDS	**COMMON/VIVIPAROUS LIZARD** *Zootoca vivipara*	▪ Europe, extending to the Arctic Circle, and central to eastern Asia, including Japan.	▪ Body is usually brown olive, sometimes black, with males having bright yellow or orange bellies and females creamy white ones. Length range is 4–7 in (10–18 cm).
	GREEN IGUANA *Iguana iguana*	▪ Central America and northern South America.	▪ Green or grayish body, 3¼–6½ ft (1–2 m) long, including the long striped tail. Long, stout legs for climbing trees. Adults have a fleshy dewlap beneath the throat, which is large in males.
	RAINBOW LIZARD *Agama agama*	▪ West, Central, and East Africa.	▪ Males turn brightly colored in the sun, with an orange-red head and a blue or turquoise body; females and juvenile males remain gray. Length range is 12–16 in (30–40 cm).
	WESTERN FENCE LIZARD *Sceloporus occidentalis*	▪ Southwestern USA and northwestern Mexico.	▪ Brown body with raised, pointed scales. Blue patches on belly are most pronounced in males. Length range is 6–9 in (15–23 cm).

AMPHIBIANS AND REPTILES

will require you to have a hunting license or permit. See p. 218 for some trapping methods and p. 224 for how to prepare the animal for cooking. Avoid brightly colored tropical frogs as they are often highly toxic. In addition, don't eat box turtles, as they sometimes eat poisonous fungi and their flesh may be toxic.

HOW TO FIND	SIMILAR SPECIES	EXTRA INFORMATION
▪ Wetlands and waterholes in savannas and tropical and subtropical forests.		▪ Also known as the African clawed toad.
▪ Wetlands.	▪ All frogs in the *Rana* genus are edible. The genus includes New World true frogs and the European common frog (*R. temporaria*).	▪ American bullfrog legs can be the size of chicken drumsticks. All temperate species of frog hibernate in the winter.
▪ Tropical forests and rain forests.		▪ Active at night. If picked up, the South American bullfrog emits a loud scream to startle the predator into dropping it.
▪ Watering holes and rain pools in open country.		▪ In the rainy season, this turtle wanders from pool to pool to forage. In the dry season, it buries itself in mud.
▪ Shallow streams in mountains or lowlands.		▪ Very active on land and in water.
▪ Lakes, ponds, and slow-moving streams and rivers.		▪ Active by day. You will often find several painted turtles piled up together on a log in the water, basking in the sun.
▪ Wide range of habitats.		▪ The carpet python is non-venomous, but it can inflict a nasty bite. It is one of Australia's most widespread snakes.
▪ Wide range of habitats, from tropical forest to dry savanna.	▪ There are several species in the boa family, which includes the world's largest snake, the green anaconda (*Eunectes murinus*).	▪ Kills its prey by coiling around it and suffocating it. Don't tackle large common boas—they can be up to 13 ft (4 m) long.
▪ Desert and open habitats.		▪ This nocturnal African snake feeds on eggs during the bird-breeding season then fasts for the rest of the year. It has no teeth.
▪ Rocky hillsides with open woodland.	▪ Ratsnakes (*Elaphe* spp.), which include the brightly colored corn snake (*E. guttata*), are found in both the New and Old Worlds.	▪ Like the common boa (see above), all ratsnakes kill their prey by constriction.
▪ Wide range of terrestrial habitats.		▪ There's no point trying to catch a common lizard by its tail as, like many lizards, it can shed the tail and regrow it.
▪ Tropical forest and rain forest.		▪ Defends itelf by lashing out with its tail and claws.
▪ Open habitats.		▪ Rainbow lizard can shed and regrow its tail.
▪ On rocks and other prominent places in temperate and coniferous woodlands.	▪ The eastern fence lizard (*S. undulatus*) is widely distributed across southeastern USA and northeastern Mexico.	▪ Fence lizard can shed and regrow its tail.

WILD FOOD

WHEN PLANNING YOUR TRIP, it is advisable to familiarize yourself with the species of fish in the region you are visiting, their seasonal availability, and the best method of catching them (see pp. 208–09 for how to make tackle and techniques to use in a survival situation). Many countries and/or states will require you to have a fishing

NAME	WHERE	IDENTIFICATION
BARRAMUNDI *Lates calcarifer*	▪ Western and eastern Indian Ocean, and northwestern and western Pacific Ocean.	▪ Up to 6½ft (2 m) long, with a rounded tail fin. Dark greenish-gray upper body fading to silver below.
COMMON/BRONZE BREAM *Abramis brama*	▪ Europe to Central Asia.	▪ Up to 2¾ft (82 cm) long, with a deep, narrow, bronze-colored body and a deeply forked tail.
COMMON/KING CARP *Cyprinus carpio*	▪ Europe and Asia; introduced in North America and Australasia.	▪ Up to 5 ft (1.5 m) long, with a dark gold upper body paling to silver below.
EUROPEAN TURBOT *Psetta maxima*	▪ Northeast Atlantic Ocean up to the Arctic Circle and the Mediterranean.	▪ Almost circular flatfish, up to 3¼ft (1 m) in diameter. Sandy brown with brown or black speckles on top.
NILE PERCH *Lates niloticus*	▪ Northern, central, and eastern Africa.	▪ Up to 6¼ft (1.9 m) long, with a rounded tail fin. Dark gray-blue upper body fading to silver below.
RUDD *Scardinius erythrophthalmus*	▪ Europe and Asia.	▪ Up to 17 in (45 cm) long, with a deeply forked tail. Dark greenish upper body paling to silver below. Red pelvic, pectoral, and anal fins.
ATLANTIC SALMON *Salmo salar*	▪ North Atlantic and Arctic oceans, and Baltic Sea and adjacent rivers; introduced to Argentina and Australasia.	▪ Up to 5 ft (1.5 m) long, with a powerful, streamlined, silver blue-green body.
TARPON *Megalops atlanticus*	▪ Eastern and western Atlantic Ocean, the Gulf of Mexico, and the Caribbean.	▪ Up to 8 ft (2.5 m) long, with a deeply forked tail. Bright silver body, with large, hard scales.
TENCH *Tinca tinca*	▪ Europe and Asia; introduced in North America.	▪ Olive green upper body, golden below, with a square tail fin. Small barbel at corners of mouth. Average length is 27 in (70 cm).
BROWN TROUT AND SEA TROUT *Salmo trutta*	▪ Temperate waters worldwide.	▪ Streamlined body with an average length of 3¼ft (1 m). The brown trout has a brownish body flecked with distinctive black and red spots. The sea-going form, the sea trout, is silver-blue with black spots.

FISH

license or permit, and some species of fish can only be caught at certain times of the year. Before fishing, you should also check with the locals that fish in the area are safe to eat—in some regions the waters may be contaminated. See pp. 212–13 for how to prepare a fish for cooking.

HOW TO FIND	SIMILAR FISH	EXTRA INFORMATION
▪ Tropical, slow-moving creeks and estuaries, hiding in mangrove roots and rocky outcrops.		▪ Beware of the sharp dorsal fin.
▪ Bottom of still and slow-moving waters, typically in lakes, rivers and ponds. Swim in large shoals.	▪ Silver bream (*Blicca bjoerkna*) and white-eye bream (*Abramis sapa*).	
▪ Still and slow-moving waters, often among vegetation near banks of lakes or ponds.	▪ Wild carp, mirrored carp, and leather carp are all varieties of the common carp.	
▪ Sandy, rocky, or mixed sea beds in shallow coastal and brackish water.	▪ Other flatfish (*Pleuronectiformes*) found in shallow water, such as sole, plaice, and flounder.	
▪ Lakes and large rivers.	▪ Other perch (*Perciformes*).	▪ Beware of the sharp dorsal fin and teeth (the Nile perch is a voracious predator). This species weighs up to 440 lb (200 kg).
▪ Still and slow-moving waters, often near banks of ponds and marshlands.	▪ Roach (*Rutilus rutilus*)	
▪ Cold, fast-flowing rivers and coastal waters.	▪ Pacific salmon (*Oncorhynchus* spp.)	▪ The salmon run (when the fish migrate from the sea to their birth rivers) is in the autumn, with spawning between November and January.
▪ Estuaries, lagoons, tidal flats and mangrove swamps. Feed in shoals.	▪ Indo-Pacific tarpon (*M. cyprinoides*), which is smaller and rarer.	▪ This species is hard to hook because of its bony mouth. It weighs up to 160kg (350lb).
▪ Still and slow-moving waters, in ponds, lakes and rivers. The best time of day to catch tench is at dawn, at the edges of dense vegetation.		
▪ Streams, rivers, lakes, and coastal waters.	▪ Rainbow trout and steelhead trout (*Oncorhynchus mykiss*) and the cutthroat trout (*O. clarki*).	▪ Sea-going forms migrate from the sea to rivers to spawn in the autumn. Beware of small bones when eating.

WILD FOOD

SOME SPECIES OF INVERTEBRATE are in danger of extinction and are protected by law; the animals listed below are common in their areas of distribution. When preparing for your trip, it is advisable to familiarize yourself with the edible invertebrates in the region you are visiting and the best way to catch them (see p. 219 for some techniques to use in a survival

	NAME	WHERE	IDENTIFICATION
MOLLUSCS	**COMMON/BLUE MUSSEL** *Mytilus edulis*	▪ Intertidal zones of estuaries and coasts of North and southeastern Atlantic and northeastern and southwestern Pacific.	▪ Dark-shelled bivalve, reaching 4–6 in (10–15 cm) in length.
	GIANT AFRICAN LAND SNAILS *Achatina* spp.	▪ Sub-tropical and tropical parts of Africa; introduced to the Asia-Pacific region.	▪ Large terrestrial snails, reaching up to 12 in (30 cm) in length when body is extended, and 4 in (10 cm) in diameter. Whorled, conical, brownish shell with darker bands.
ARTHROPODS	**GRASSHOPPERS** *Acrididae*	▪ Worldwide, on vegetation and the ground.	▪ Winged insects with powerful hindlegs, ⅖–3¼ in (1–8 cm) long, usually with camouflage coloring and patterning.
	HONEYPOT/HONEY ANTS *Myrmecocystus* spp.	▪ Semi-arid and arid regions of southwest USA, Mexico, Africa, and Australia.	▪ Segmented bodies with six legs and a pair of antennae, and a constricted waist. Range in colour and size, from ⅒–½in (2–12mm). Special members of the colony, called repletes, have abdomens that can swell to the size of a grape.
	WITCHETTY GRUB *Endoxyla leucomochla*	▪ Underground, in the roots of the witchetty bush (*Acacia kempeana*), central Australia.	▪ The larva of this species of cossid moth reaches about 2¾ in (7 cm) in length. It is white with a brown head.
CRUSTACEANS	**CRAYFISH/CRAWFISH/ YABBIES/KOURAS** *Astaclidea*	▪ Worldwide, in freshwater streams.	▪ Segmented bodies, which can be sandy yellow, green, dark brown, or blue-gray in color. Ten legs, the front two of which are large claws. Average length is 4 in (8 cm), but some species grow much larger.
	WOODLICE *Oniscidea* and *Armadillidiidae*	▪ Worldwide, in damp terrestrial microhabitats, such as in rotting wood and leaf litter.	▪ Flattened, segmented bodies up to ¾ in (2 cm) in length. Gray or light brown or black in color. Pill millipedes will roll up into a ball when threatened.
WORMS	**COMMON EARTHWORM** *Lumbricus terrestris*	▪ Temperate regions of Europe; introduced to most parts of the world.	▪ Reddish worm, reaching 14 in (35 cm) in length when extended.

INVERTEBRATES

situation). Many countries and/or states will require you to have a hunting license or permit, and some species of shellfish can only be caught at certain times of the year. Also check with the locals that the shellfish are safe to eat—in some regions the waters may be contaminated. See pp. 224–25 for some preparation and cooking methods.

SIMILAR INVERTEBRATES	BEWARE OF	EXTRA INFORMATION
▪ Most species of marine and freshwater bivalves and univalves are edible if alive and healthy when gathered.	▪ Don't collect marine shellfish that aren't covered at high tide. The black mussel (*Musculus niger*), found in Arctic waters, can be poisonous year-round. Mussels in tropical zones are poisonous during the summer.	▪ Healthy bivalves should close their shells when tapped and the shells should open when cooked. Healthy univalves, such as limpets, should be difficult to pry off the rocks on which they live.
▪ Many terrestrial and freshwater snails are edible. The winkle or common periwinkle (*Littorina littorea*) is an edible sea snail.	▪ Avoid any terrestrial snails with brightly colored shells and all sea snails unless you can positively identify them as an edible species, as they may be poisonous.	▪ Snails should be starved for 24 hours or fed a diet of edible green leaves to purge their guts before cooking. Boil for at least 10 minutes.
▪ Most species of cricket (*Gryllidae*) and katydid (*Tettigoniidae*) are edible.	▪ Avoid any brightly colored grasshoppers as they may be toxic.	▪ Remove the antennae, wings, and leg spurs, and roast to kill any parasites.
▪ Most species of ant are edible if gathered carefully. The larvae of wood ants (*Formica* spp.) in northern temperate regions make a nutritious meal in summer months. The abdomen of the green or weaver ant (*Oecophylla smaragdina*) found in southeast Asia and Australia tastes like citrus sherbert. Termites (*Isoptera*) are also edible. See p. 219 for catching methods.	▪ Most species of ant are aggressive in defense of their nest and have a stinging bite. Some will then squirt formic acid at the site of the bite.	▪ The swollen abdomens of the repletes contain a nutritious fluid. ▪ The repletes live deep underground in the ants' nest so you would have to dig them out.
▪ The term "witchetty grub" is also applied to the edible larvae of other cossid moths, ghost moths (*Hepialidae*), and longhorn beetles (*Cerambycidae*). Palm grubs are the edible larvae of the palm weevil (*Rhynchophorus* Spp.). They live in the trunks of sago palms (see pp. 284–285) in Southeast Asia.	▪ Do not eat any insect larvae that are already dead when you find them, or that look sick, smell bad, or that irritate the skin if handled.	▪ Should be eaten raw or roasted quickly in hot embers.
▪ Marine decapods such as crabs, lobsters, and prawns are also edible.	▪ May contain harmful parasites so cook well.	▪ Keep alive until ready to eat. See p. 225 for preparation and cooking methods.
		▪ Boil or fry gently. Will turn slightly pink, like shrimp, when cooked.
▪ All earthworms are edible.		▪ The common eathworm is unusual in that it feeds at the surface, making it easier to find and gather than other worms.

NATURAL DANGERS

FOR INFORMATION ON LARGER dangerous animals, such as bears, big cats, and sharks, and venomous animals, such as snakes, see pp. 242–43. When planning your trip, it is advisable to familiarize yourself with the potentially harmful wildlife in the region you are visiting, the nature of the threat, and how to avoid it if possible. In the case of suspected

	NAME	WHERE	IDENTIFICATION
MAMMALS	**VAMPIRE BATS** *Desmodontinae*	▪ In rain forests, deserts, and grasslands, from Mexico to northern and central South America.	▪ Small bats, with a body length of 2¾–3½ in (7–9 cm) and a wingspan of 14–16 in (35–40 cm). Dark brownish gray fur, paler on the belly. Razor-like upper incisors.
	RATS *Rattus* spp.	▪ Worldwide, except for the polar regions.	▪ Medium-sized rodents, with a minimum body length of 5 in (12 cm), and long, naked tails. Coarse brown, gray-brown, or black fur, paler on belly. Pink feet.
SNAKES	**SPITTING COBRAS AND RINKHALS** *Naja* spp. and *Hemachatus haemachatus*	▪ Warm temperate, subtropical, and tropical regions of Asia and Africa.	▪ Pale brown to black snakes up to 6 ft (2 m) long. When threatened, will rear up and spread their hoods, revealing striped patterns on their necks.
FISH	**ELECTRIC EEL** *Electrophorus electricus*	▪ Amazon and Orinoco river systems, South America, often in shallow water.	▪ Not a true eel, but an eel-like fish up to 8¼ ft (2.5 m) long and can be as thick as a human thigh. Continuous fin along lower body.
	STONEFISH *Synanceia verrucosa*	▪ In tropical coastal waters of northern Indian Ocean and southwestern Pacific Ocean.	▪ Up to 16 in (40 cm) long, this lumpy fish can match its color to the rock or sediment on which it lies while waiting for passing prey. Venomous spines in the dorsal fin.
OTHER ANIMALS	**BOX JELLYFISH/ SEA WASPS** *Cubomedusae*	▪ At the surface of tropical waters of southwest Pacific and eastern Indian Ocean.	▪ Box-shaped, transparent jellyfish, up to 10 in (25 cm) in diameter. Up to 15 long tentacles, which bear stinging cells, at each corner.
	CENTIPEDES *Chilopoda*	▪ In soil, leaf litter, cracks, and crevices in temperate, subtropical, and tropical regions worldwide.	▪ Elongated, flattened bodies divided into at least 16 segments, most of which have one pair of legs. Tropical species are typically brightly colored—yellow, red, orange, or green—with dark stripes. The world's largest centipede, *Scolpendra gigantea*, reaches 12 in (30 cm) in length.
	CONE SHELLS/ CONE SNAILS *Conidae*	▪ In inter-tidal zones and on coral reefs in warm temperate, subtropical, and tropical seas and oceans worldwide.	▪ Brightly colored and patterned, cone-shaped shells, up to 9 in (23 cm) long. These predatory snails have venomous, harpoon-like mouthparts.

WILD ANIMALS

poisoning, call the emergency services immediately. For further information regarding treatment for bites and stings, see pp. 266–67. If a casualty becomes unconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue breaths (see p. 277), use a plastic face mask or face shield, if you have one.

	DANGER	TREATMENT	HOW TO AVOID
	▪ Many mammal species are infected with rabies, but because vampire bats feed by biting their victims, then lapping blood from the wound, the virus can be transmitted via their saliva. Symptoms include fever, headache, fear (especially of water), and seizures.	▪ If bitten by a vampire bat, or other potentially rabid mammal, clean the wound thoroughly and seek immediate medical help. Symptoms usually appear 2–8 weeks after infection, by which time the disease is almost always fatal.	▪ Vampire bats roost in hollow trees and caves, emerging at night to feed. Don't shelter in bat caves (bat dung can cause a respiratory disease). Keep covered at night, and sleep under a mosquito net. Get vaccinated before traveling to a high-risk area.
	▪ Can be infested with parasites and can also carry infectious diseases. Leptospirosis is a water-borne bacterial disease spread by infected rats' urine; the acute form in humans is called Weil's disease.	▪ If you've been in contact with potentially contaminated water and develop flu-like symptoms (fever, headache, muscle pain) you must get your blood tested as soon as possible. Antibiotics are effective in the early stages of the disease.	▪ If handling a rat, beware of its teeth and claws, and wash thoroughly afterward. Cook the meat thoroughly. In rat-infested areas avoid getting water into any cuts; if you do, wash with soap and boiled water.
	▪ Will eject a fine spray of venom from their fangs upward the eyes of an aggressor. Causes intense pain, tearing, and discharge from the eyes, spasm and swelling of the eyelids. Ulceration of the cornea can lead to infection resulting in blindness.	▪ Call the emergency services. Rinse the eyes with cool water for 10 minutes. Apply an antibiotic eye ointment.	▪ If you come across a spitting cobra, try to remain completely still and allow the snake to move away, if possible (see p. 243).
	▪ Can generate shocks of up to 600 volts, used to stun or kill prey when hunting, or in self-defense. Such a shock can be discharged up to 8 hours after its death, and is potentially lethal to humans.	▪ Call the emergency services. If the casualty becomes unconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR (see p. 277).	▪ Keep your boots on if crossing a tributary and take care when washing. These waters are also home to piranhas (*Serrasalminae*), which have razor-sharp teeth.
	▪ The world's most venomous fish. If trodden on, the spines inject poison into the puncture wound. This is extremely painful and can be fatal. The injured body part will swell and muscular paralysis may set in.	▪ Call the emergency services—anti-venin is available. Immerse the injured part in water as hot as the casualty can stand for at least 30 minutes (see p. 267). Don't immobilize the injured limb. Be ready to begin CPR if the casualty loses consciousness (see pp. 276–77).	▪ Wear something on your feet when walking in shallow water. Weaverfish (*Trachinidae*) and scorpionfish (*Scorpaenidae*) also have venomous stings; the treatment is the same.
	▪ Sting is extremely painful—it will damage the skin and can cause permanent scarring. In severe cases, box jellyfish (*Chironex fleckeri*) stings can lead to cardiac arrest and death within minutes.	▪ Call the emergency services—anti-venin is available. Flood the stung area with vinegar or sea water for at least 30 seconds to neutralize the stinging cells (see p. 267). Be ready to begin CPR if the casualty loses consciousness (see pp. 276–77).	▪ Never touch jellyfish, even when dead. All jellyfish, sea anemones, corals, and the Portuguese man-of-war (*Physalia physalis*) release venomous cells when touched. Some are more toxic than others.
	▪ Have a pair of large, venomous claws on their heads, which in some species can be dangerous to humans. The sting is painful and leads to swelling. May result in anaphylactic shock.	▪ Call the emergency services. Apply a cold compress for at least 10 minutes to reduce pain and swelling. Monitor the casualty for anaphylactic shock (see p. 274) and be ready to begin CPR (see p. 277).	▪ As a general rule, avoid all multiple-legged arthropods. If you discover one on your body, gently brush it off in the direction in which it is walking.
	▪ Will fire its harpoon in self-defense if touched. Sting of most species is painful; that of the larger tropical species can be fatal. Symptoms include swelling and numbness, and may be followed by severe breathing difficulties.	▪ Call the emergency services and treat as for snake bite (see p. 266). Monitor the casualty for shock (see p. 274) and be ready to begin CPR (see p. 277).	▪ Don't pick up any cone shells in rock pools or when snorkeling or diving. (There is no anti-venin: treatment entails providing life support until the venom is metabolized by the casualty's body.)

NATURAL DANGERS

IN THE CASE OF SUSPECTED POISONING from a spider bite or a scorpion sting, you must call the emergency services immediately. For further information regarding treatment for bites and stings, see pp. 266–67. If a casualty becomes unsconscious, open the airway and check breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue

	NAME	WHERE	IDENTIFICATION
INSECTS	**BEES, WASPS, AND HORNETS** *Hymenoptera*	▪ In a variety of terrestrial habitats worldwide.	▪ Narrow-waisted bodies, often striped black and yellow. Bees are hairy, while wasps and hornets have hairless bodies. Up to 1½ in (3.5 cm) long. Bees have barbed stingers, wasps and hornet stingers are smooth.
	MOSQUITOES *Culicidae*	▪ In a variety of terrestrial habitats near water worldwide, especially in warmer regions.	▪ Narrow bodies and long, slender legs. Up to ¾ in (2 cm) long. Females have syringe-like mouthparts for piercing skin.
ARACHNIDS	**BROWN RECLUSE/ FIDDLEBACK SPIDER** *Loxosceles reclusa*	▪ Southern midwestern states of USA, south to the Gulf of Mexico.	▪ Small brown spider, about ½ in (1.25 cm) long) with a distinctive dark, violin-shaped mark on the back of its cephalothorax (head).
	FUNNEL-WEB SPIDERS *Atrax robustus* and *Hadronyche* spp.	▪ Moist, cool, sheltered habitats in eastern and southern Australia.	▪ Large spiders, up to 1½ in (4.5 cm) long, with glossy dark brown or black bodies and short legs.
	TARANTULAS *Theraphosidae*	▪ Deserts and forests in subtropical and tropical regions worldwide.	▪ Very large spiders with bodies up to 5 in (12 cm) long and a legspan of up to 11 in (28 cm). Bodies and legs are covered in bristly hairs. Pale brown to black in color, with pink, red, brown, or black markings.
	WIDOW SPIDERS *Latrodectus* spp.	▪ Warm temperate, subtropical, and tropical regions worldwide.	▪ Small spiders, with dark, glossy bodies and distinctive red, yellow, or white markings on the abdomen, which are shaped like an hourglass in some species.
	SCORPIONS *Scorpiones*	▪ Deserts, grasslands, woodland, and forests in warm temperate, subtropical, and tropical regions worldwide.	▪ Segmented bodies, with large, claw-like pedipalps and sting-bearing tails. Yellow, brown, or black bodies, 3–8 in (8–20 cm) long, depending on the species.
	TICKS *Ixodidae*	▪ Grassland and woodland habitats worldwide.	▪ Rounded bodies with no distinct divisions. Yellow to red- or black-brown in color. Up to ⅖ in (1 cm) long, but larger after feeding. They have spiked mouthparts, which they use to attach themselves to the skin of warm-blooded animals to feed on their blood.

INSECTS AND ARACHNIDS

breaths (see p. 277). If you need to give rescue breaths and there is poison in the casualty's mouth, or on his or her face, use a plastic face mask or face shield, if you have one. Before your trip, it is advisable to familiarize yourself with the harmful insects and arachnids (spiders, scorpions, and ticks) in the region you are visiting, and the nature of the threat.

	DANGER	TREATMENT	HOW TO AVOID
	▪ When a bee stings you, the stinger is ripped out of its body with the venom sac. Wasps and hornets can sting repeatedly. The stings are painful, followed by redness and swelling. A single sting may result in anaphylactic shock (see p. 274)	▪ Scrape the sting off sideways at the base of the sting if it is visible, using a fingernail or a credit card. Raise the affected part and place a cold compress against it for at least 10 minutes (see p. 267). Monitor for signs of allergic reaction.	▪ Watch out for flowers where bees may be feeding. Be careful when gathering fruit, or cleaning fish or game, as wasps or hornets will be attracted to the smell. Don't disturb their nests.
	▪ Female mosquitoes feed on blood and their bites can irritate the skin. They are carriers of infectious diseases such as malaria, yellow fever, and dengue fever.	▪ Apply an antihistamine cream to the bites if they itch, and do not scratch them, as there is a risk of infection. If casualty develops a headache and fever, get medical help. Get casualty to rest, drink plenty of water, and take anti-malarials, if directed.	▪ Cover your arms and legs after sunset, use mosquito repellent, and sleep under a mosquito net. Take anti-malarial medication as directed. Get a yellow fever vaccination before traveling, where advised.
	▪ Bite can cause fever, chills, vomiting, and, pain in the joints. In most cases the bite is minor, but in some cases the tissue around the wound will die, leaving a deep scar. Rarely fatal.	▪ Sit casualty down, treat as for snakebite (see p. 266), and call the emergency services. Apply a cold compress to the site of the sting. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274).	▪ This spider likes to hide in dark places, so be careful where you put your hands and bare feet, and always check your bedding.
	▪ Bite is very painful and cause profuse sweating, nausea, vomiting, tingling around the mouth and tongue, salivation, and weakness. Rarely fatal.	▪ Sit casualty down, treat as for snakebite (see p. 266), and call the emergency services—anti-venin is available. Monitor for signs of allergic reaction and anaphylactic shock (see p. 274).	▪ Take care when moving rocks and logs.
	▪ Bite can be painful but is only mildly venomous. If provoked, tarantulas will flick irritating hairs from their abdomen which will cause itching and swelling of the skin and nasal passages, and watering of the eyes.	▪ Give pain relief and apply a cold compress to the site of the bite. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274). Wash any hairs away with cool water.	▪ Tarantulas are nocturnal and hunt prey in trees and on the ground. Move away if you disturb one; don't touch it.
	▪ Bite is painful and results in localized redness and swelling. More severe reactions include profuse sweating, abdominal and chest pains, nausea, and vomiting. Rarely fatal.	▪ Apply a cold compress (ice if available) and give pain relief. If symptoms worsen, call the emergency services – anti-venin is available. Monitor for signs of allergic reaction, such as wheezing, and anaphylactic shock (see p. 274).	▪ Widow spiders, which include Australia's redback spider (*L. hasselti*), build their webs in dry, sheltered spots, such as in shrubs or among rocks or logs, so take care when moving among them.
	▪ Scorpion stings can be very painful and may cause severe illness. A few species have neurotoxins in their venom, resulting in temporary paralysis for 1–2 days. Rarely fatal and only in young children, the elderly, or the ill or infirm.	▪ If swelling occurs, sit casualty down, treat as for snake bite (see p. 266). Apply an ice pack or cold compress to the site of the sting. Call the emergency services, monitor for anaphylactic shock (see p. 274) and allergic reaction. Try to identify the scorpion.	▪ Scorpions are mostly nocturnal and take shelter during the day. Check your boots and clothing before putting them on and check your bedding at night. Take care when gathering firewood or turning over rocks.
	▪ Ticks are carriers of infectious diseases such as Lyme disease and Rocky Mountain spotted fever. Sufferers of Lyme disease may develop a red, ringed "bull's-eye" skin lesion at the bite site.	▪ Carefully remove an attached tick with tweezers or a specialized tool (see p. 267). Clean and monitor the wound for symptoms. Get medical help if casualty develops a severe headache, stiff neck, or a fever.	▪ Wear light-shaded, tightly woven clothing and apply tick repellent. Check your skin and clothing for ticks at regular intervals.

NATURAL DANGERS

IN THE CASE OF SUSPECTED PLANT POISONING, you must call the emergency services immediately. If a casualty becomes unconscious, open the airway and check their breathing (see p. 276). Be ready to begin CPR—chest compressions and rescue breaths (see p. 277). If you need to give rescue breaths, use a plastic face mask or face shield, if you have one, to

	NAME	WHERE	IDENTIFICATION	POISONOUS PART
CONTACT POISONING	**COWHAGE/COWAGE/ COWITCH** *Mucana pruriens*	▪ Scrub and light woodland in the the tropics and USA.	▪ Vine-like plant with green, oval leaflets growing in threes. Spikes of hairy, dull purplish flowers. Hairy brown seedpods.	▪ Flowers ▪ Pods
	POISON IVY *Toxicodendron radicans*	▪ Wooded areas in North America; widely introduced elsewhere.	▪ See p. 268 for illustration.	▪ All
	WHITE/GRAY MANGROVE *Avicennia marina*	▪ Mangrove swamps from tropical Africa to Indonesia and Australasia.	▪ Slender tree with pale bark and long aerial roots. Green, oblong, leathery leaves. Yellow flowers and fleshy, pale orange seed capsules.	▪ Sap
INGESTION POISONING	**BANEBERRY** *Actaea* spp.	▪ Temperate woodland in the northern hemisphere.	▪ Low shrub with, toothed green leaves. Small white flowers. White, black, or red berries grow in clusters.	▪ All, but the berries are the most poisonous.
	CASTOR OIL PLANT/ CASTOR BEAN *Ricinus communis*	▪ Throughout the tropics in scrub and wasteland; introduced to temperate regions.	▪ Tall shrub with glossy dark green or dark purple, star-like leaves. Spikes of yellow flowers. Spiky red pods.	▪ All, but the seeds contain particularly high levels of the poison ricin.
	DEATH CAMAS/ DEATH LILY *Zigadenus* spp.	▪ Grassy, wooded, and rocky areas in temperate regions of North America.	▪ Long grass-like leaves arise from a bulb. Flowers have six white petals with a green heart at base.	▪ All
	PANGI *Pangium edule*	▪ Tropical forests in Southeast Asia.	▪ Tall tree with green, heart-shaped leaves. Green flower spikes. Large brownish pear-shaped fruits grow in clusters.	▪ All, but especially the seeds, which contain prussic acid (cyanide).
	PHYSIC NUT *Jatropha curcas*	▪ Woodland throughout the tropics and southern USA.	▪ Large shrub with large, green, ivy-like leaves. Small greenish-yellow flowers. Apple-sized yellow fruit has large seeds.	▪ All, but particularly the seeds.
	POISON HEMLOCK *Conium maculatum*	▪ Wet or moist ground in temperate parts of Eurasia; introduced to North America and Australia.	▪ Tall herb with branching purple-spotted stem; toothed leaves; small clusters of white flowers grown in flat umbrellas.	▪ All, but the root has the highest level of toxins.
	STRYCHNINE TREE *Strychnos nux-vomica*	▪ Tropical and subtropical forests in India, Southeast Asia, and Australia.	▪ Evergreen tree with paired oval leaves. Small clusters of greenish flowers grow at end of branches. Large orange-red berries.	▪ All, but the seeds contain lethal levels of strychnine.

POISONOUS PLANTS

protect yourself from any poison on the casualty's mouth. When planning your trip, it is advisable to familiarize yourself with the poisonous plants in the region you are visiting. For further information regarding plant contact poisoning and ingestion poisoning, and their treatment in a survival situation, see pp. 268–69.

EFFECT	TREATMENT	ADDITIONAL INFO
▪ Skin irritation. Contact with the eyes can cause blindness.	▪ Call the emergency services. Rinse the skin with cold water for 20 minutes, the eye(s) for 10 minutes. Don't let the water collect under the casualty or you.	
▪ Reddening, itching, swelling, and blistering of skin within 4–24 hours of contact with the plant. Some people may suffer an allergic reaction called anaphylactic shock (see p. 274).	▪ Wash the affected area with soap and cold water. Apply anti-poison ivy cream. Monitor and treat for shock (see p. 274) if necessary. For natural remedies, see p. 268.	▪ Western poison oak (*T. diversilobum*), Atlantic poison oak (*T. pubescens*), and poison sumac (*T. vernix*) also cause contact poisoning (see p. 268).
▪ The sap blisters the skin and will cause temporary blindness if it gets in the eyes.	▪ Call the emergency services. Rinse the skin with cold water for 20 minutes, the eye(s) for 10 minutes. Monitor casualty and treat for shock (see p. 274).	▪ Blinding or milky mangrove (*Exoecaria agallocha*) also has poisonous sap.
▪ Can be fatal—ingestion of berries can lead to cardiac arrest and death. Other symptoms include dizziness, vomiting, and severe internal irritation.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	
▪ Can be fatal—ingestion of raw seeds causes severe diarrhea.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ The seedpods explode when ripe, scattering the large, oval, glossy, brown-mottled seeds, which can be mistaken for beans.
▪ Deadly poisonous. Symptoms include profuse salivation, vomiting, diarrhea, confusion, slow, irregular heartbeat, low temperature, difficulty breathing, and unconsciousness, leading to death.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ This plant looks similar to wild onions, which are edible, but the death camas does not have the onion smell.
▪ Deadly poisonous. Causes unconsciousness and severe breathing difficulties, which, if untreated, may lead to death.	▪ Call the emergency services (antidotes are available). Induce vomiting if casualty is still conscious. Dilute the poison (see Death Camas, above). Keep casualty warm and at rest.	▪ If the casualty is unconscious and stops breathing, give rescue breaths, using a face shield. Alternatively, use an oxygen bag and a mask, if possible.
▪ Severe diarrhea and vomiting.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ The seeds look like betel nuts and taste sweet. Several species in the *Jatropha* genus are poisonous, for example, the aptly named bellyache bush (*J. gossypifolia*).
▪ Deadly poisonous. A small amount can result in death by respiratory failure from muscular paralysis. Early symptoms include nausea, vomiting, and rapid heartbeat.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison (see Physic Nut, above). If the casualty is unconscious and stops breathing, begin CPR, using a face shield.	▪ Water hemlock or spotted cowbane (*Cicuta* spp.), which looks similar, is also deadly poisonous. Don't mistake these plants for wild carrots or wild parsnips.
▪ Deadly poisonous. Muscle spasms occur 10–20 minutes after exposure, leading to seizures followed by death from severe breathing difficulties or exhaustion within 2–3 hours.	▪ Call the emergency services. Induce vomiting if casualty is still conscious. Dilute the poison by getting him or her to drink large quantities of water or milk, or tea mixed with charcoal.	▪ Warning: strychnine can also enter the system through contact with the eyes.

RESOURCES

SCOUTS AND SCOUTING

Boy Scouts of America

1325 West Walnut Hill Lane
P.O. Box 152079, Irving,
Texas 75015-2079

Phone: 972-580-2000

Fax: 972-580-2413

www.scouting.org

Boy Scouts of America prepares young people to make ethical and moral choices over their lifetimes by instilling in them the values of the Scout Oath and Law.

Girl Scouts of the USA

420 Fifth Avenue
New York, New York 10018-2798

Phone: 800-GSUSA-4-U [800-478-7248] or 212-852-8000

www.girlscouts.org

Girl Scouts is one of the largest organizations dedicated specifically to girls. It encourages them to develop skills and values such as leadership, social conscience, and self-worth that will stand them in good stead in the wider world.

National Scouting Museum

Boy Scouts of America, S505,
1329 West Walnut Hill Lane, Irving,
Texas 75038

Phone: 800-303-3047 and 972-580-2100

e-mail: nsmuseum@netbsa.org

www.bsamuseum.org

This museum is a tribute to the rich history of the Boy Scouts of America and features a Norman Rockwell art gallery, virtual reality adventures, hands-on learning experiences, and a collection that traces the Scouting movement from its beginnings.

WEATHER AND ENVIRONMENTAL HAZARDS

National Oceanic and Atmospheric Administration (NOAA)

1401 Constitution Avenue, NW,
Room 6217, Washington, DC 20230

Phone: 202-482-6090

Fax: 202-482-3154

www.noaawatch.gov

The NOAA has several departments, such as the National Severe Storms Laboratory (NSSL), the Storm Prediction Center (SPC), which is part of the National Weather Service (NWS), and the National Centers for Environmental Prediction (NCEP). The SPC provides regular, timely, and accurate storm forecasts, and watches out for severe thunderstorms, tornadoes, heavy rain, heavy snow, and fire weather events over the contiguous United States.

National Weather Service

1325 East West Highway, Silver Spring,
Maryland 20910

www.weather.gov/om/severeweather/index

Provides weather, hydrologic, and climate forecasts and observations for the whole country and the waters and oceans around it. The service offers a national information database for the use of public, private, and global communities.

NOAA Weather Radio All Hazards (NWR)

www.weather.gov/nwr

A nationwide network of radio stations broadcasts continuous weather information directly from the nearest National Weather Service office. NWR broadcasts official Weather Service warnings, watches, forecasts, and other hazard information 24 hours a day, seven days a week.

National Interagency Coordination Center (NICC)

www.nifc.gov/nicc/index.htm

The NICC serves as a focal point for coordinating the national mobilization of resources, such as heavy airtankers, helicopters, smokejumpers, area command teams, and Remote Automated Weather Stations (RAWS), for wildland fire and other incidents throughout the United States.

The Forest Fire Lookout Association (FFLA)

2590 W. Versailles Drive, Coeur d'Alene,
Idaho 83815

Phone: 208-765-1714 and 800-GRN-TREE

Email: weblookout@imbris.net

www.firelookout.org

The FFLA investigates former forest-fire lookout sites, ground cabins, and early forest fire detection methods to encourage public groups and others in restoring forest-fire lookouts.

U.S. Geological Service

Western Distribution Branch,
P.O. Box 25286, Denver, Colorado 80225

Phone: 303-236-7477

www.usgs.gov

This multi-disciplinary science organization focuses on biology, geography, geology, geospatial information, and water, and is dedicated to the timely, relevant, and impartial study of the US landscape, and natural resources and hazards.

DesertUSA

www.desertusa.com

This website is a comprehensive resource for North American deserts and Southwest destinations. It includes information about desert biomes, and the ways plants and animals adapt to them, and about national and state parks.

HEALTH AND MEDICAL

Centers for Disease Control and Prevention

1600 Clifton Rd, Atlanta, Georgia 30333

Phone: 800-232-4636 and 888-232-6348

Email: cdcinfo@cdc.gov

www.cdc.gov

Provides the expertise, information, advice, and tools that individuals and communities need to protect their health, including the promotion of healthy living and the prevention of disease, injury, and disability.

U.S. Department of Health and Human Services

200 Independence Avenue, S.W.,
Washington, D.C. 20201

Phone: 202-619-0257

Toll Free: 1-877-696-6775

www.hhs.gov/disasters/index.shtml

For information and advice on coping with disasters and related medical emergencies.

American Medical Association

515 N. State Street, Chicago, Illinois 60610

Phone: 800-621-8335

www.ama-assn.org/ama/pub/category/6206.html

The AMA's Center for Public Health Preparedness and Disaster Response is an important national educational resource for enhancing the disaster preparedness and response capabilities of both civilian and military providers.

American Red Cross

National Headquarters, 2025 E Street NW
Washington, DC 20006

Phone: 800-733-2767

www.redcross.org/services

The nation's premier response organization that provides relief for communities during natural and manmade emergencies. It belongs to a worldwide movement that offers neutral humanitarian care and assistance to victims of war and devastating natural disasters such as earthquakes.

National Capital Poison Center

3201 New Mexico Ave, Suite 310,
Washington D.C. 20016

Emergency Phone: 800 222 1222

www.poison.org

Affiliated with The George Washington University Medical Center, the Poison Center is committed to prevent poisonings, save lives, and limit injury from poisoning.

SEARCH AND RESCUE

National Association for Search & Rescue (NASAR)

PO Box 232020, Centerville,
Virginia 20120-2020

Phone: 703-222-6277

Toll Free: 877-893-0702

Email: info@nasar.org

www.nasar.org/nasar

A not-for-profit association dedicated to the advancement of professional, literary, and scientific knowledge in fields related to search and rescue throughout the United States and around the world.

National Institute for Urban Search & Rescue

PO Box 91648, Santa Barbara,
California 93190

Phone: 800-767-0093 and 805-569-5066

Cell: 805-798-0169

Email: niusr@cox.net

www.niusr.org

A non-profit organization dedicated to finding ways of saving lives by improving disaster readiness and response through public awareness, collaboration, research, and engineering.

Mountain Rescue Association

www.mra.org

Represents highly skilled and active mountain rescue teams throughout the country, who work for local government authorities and who are dedicated to saving lives through rescue and mountain safety education.

International Association of Dive Rescue Specialists

201 North Link Lane, Fort Collins,
Colorado 80524-2712

Phone (toll free): 800-IADRS-911

Phone (international): 1-970-482-1562

Fax (international): 970-482-0893

www.iadrs.org

An association of public safety divers and water rescue personnel who, over the years, have educated and assisted professionals in all 50 states and 15 foreign countries.

Association of Air Medical Services

526 King Street, Suite 415,
Alexandria, Virgina 22314

Phone: 703-836-8732

Fax: 703-836-8920

www.aams.org

An international not-for-profit organization that represents providers of air and surface medical transport systems, and aims to ensure that everyone has access to quality air medical and critical care support.

FEDERAL

Environmental Protection Agency (EPA)

EPA Headquarters, Ariel Rios Building
1200 Pennsylvania Avenue, N.W.,
Washington, DC 20460

EPA Hotlines: Various phone numbers and websites for assistance.

www.epa.gov/ebtpages/
emeremergencypreparedness.html

The EPA develops, implements, and coordinates preparations for chemical and other emergencies in order to be able to respond quickly and effectively to environmental crises and to keep the public informed about hazards in their community.

Federal Emergency Management Agency (FEMA)

500 C Street S.W., Washington, D.C. 20472

Phone: 800-621-FEMA-3362

TDD: TTY users can dial 1-800-462-7585 to use the Federal Relay Service.

www.fema.gov

FEMA aims to reduce the loss of life and property, and protect the country from all hazards, including natural disasters, acts of terrorism, and other manmade disasters.

U.S. Dept of the Interior National Parks Pass

1849 C Street, NW, Washington, DC 20240

Phone: 202-208-3100

Email: webteam@ios.doi.gov

www.doi.gov/parkpass.html

You can buy an annual pass for entry into national parks from the US Geological Survey store and through the government's federal lands recreation web portal at www.recreation.gov.

CANADA

National Search and Rescue Secretariat (NSS)

400-275 Slater Street, Ottawa,
Ontario K1A 0K2

Phone: (toll-free) 1-800-727-9414

Email: inquiry@nss.gc.ca

www.nss.gc.ca

An independent government agency that coordinates the search-and-rescue requirements of the military, coastguard, police, transport, parks, and meteorological services.

Search and Rescue Volunteer Association of Canada

8 Paradise Road, Paradise,
Newfoundland and Labrador A1L 3B4

Phone (toll free): 1-866-9SARVAC
(1-866-972-7822)

Phone (office): 709-368-5533

Fax: 709-368-1298

Email: info@sarvac.ca

www.sarvac.ca

A not-for-profit organization that provides a national search-and-rescue service, and is committed to fostering the exchange of information between other rescue services.

Meteorological Service of Canada

www.msc-smc.ec.gc.ca

Provides information and conducts research on water, climate, weather, atmospheric science, and other environmental issues.

Environment Canada

Inquiry Center, 351 St. Joseph Boulevard
Place Vincent Massey, 8th Floor
Gatineau, Quebec K1A 0H3

Phone: 1-800-668-6767
or 819-997-2800

Fax: 819-994-1412

TTY: 819-994-0736

Email: enviroinfo@ec.gc.ca

www.ec.gc.ca

Government department committed to preserving and enhancing the country's environment, and to providing weather forecasts.

GLOSSARY

A-frame A type of shelter design in which the roof is suspended by a single support running from the front to the back, resting on an "A"-shaped framework at either end.

Air mass A large body of air in a weather system that has its own temperature and humidity.

Altitude sickness An illness brought on by low air pressure at high altitudes.

Anabatic winds Daytime, upslope winds.

Anaphylactic shock A dangerous allergic reaction brought on by a bite or sting.

Attack point See *Stand-off.*

Avalanche transceiver A personal beacon that can be activated to alert rescuers by someone caught in an avalanche.

Base layer The bottom layer of clothing that sits directly against the skin and is designed to "wick" sweat away from the body.

Base rate The heart rate, as measured when the body is at rest.

Beaufort scale A way of describing wind speed on land and at sea, using a scale from 0 (calm) to 12 (hurricane).

Belaying A rope technique used by climbing partners to safeguard each other.

Bivi pup tent A basic one-man tent that uses a bivi bag as a covering.

Bivi bag A large waterproof sack suitable for use as a basic shelter; smaller and lighter than a tent.

Blazes A general term for markers on a trail.

Bola A hunting device made of a rope with weights attached, designed to be spun and thrown.

Boot skiing A technique for descending snow-covered slopes without skis.

Bothy bag A large waterproof bag that provides shelter from the wind, rain or sun.

Bow-and-drill set One of the oldest methods of creating a glowing ember (coal) for use in lighting a fire.

Boxing an object See *Detouring.*

Breathable clothing Clothing that allows sweat in the form of water vapour to be carried away from the body but is waterproof from the outside so keeps liquids (rain or snow) out.

Bungee A type of elasticated cord, often with hooks at either end.

Bushcraft knife A versatile survival knife used primarily for carving, cutting, and splitting small logs.

Button compass A small, simple compass, ideal as a back-up compass.

Button tie A method of attaching cordage to fabric that has no hoops or grommets; useful in that it avoids ripping a hole in the fabric, maintaining any waterproof qualities of the fabric, and allows both fabric and cordage to be reused subsequently for other purposes.

Cairn Trail marker consisting of a pile of rocks of varying sizes, designed to be visible in fog.

Canoe An open-deck paddle boat, for one or more people, and equipment.

Carabina A metal clip used for joining ropes or attachment.

Carbon monoxide A colourless, odourless, tasteless, but highly toxic gas.

Cardiopulmonary resuscitation (CPR) An emergency system of chest compressions and rescue breaths designed to restart a casualty's heart and lungs.

Cerebral oedema Accumulation of excess water in the brain that causes reduced brain function and potentially fatal swelling.

Char cloth Cotton cloth that has been combusted in the absence of oxygen. Used for lighting fires.

Chimneying A climbing technique used for climbing up the inside of large rock clefts.

Cirrus A wispy cloud that forms high in the sky and is made from ice crystals.

Cold shock An involuntary gasp reflex followed by hyperventilation, caused by sudden immersion in very cold water. In some circumstances cold shock response can lead to the inhalation of water, disorientation, panic, hypothermia, cardiac problems, and death.

Commando saw A small saw consisting of a serrated wire blade with a ring at each end.

Compacted snow trench A shelter dug into compacted snow; the excavated snow can be cut into blocks and used to form a roof.

Compass baseplate A plate to which some compasses are fixed, containing additional markings used for orientation and navigation.

Compass scale Scaled markings on a compass baseplate that measure distances on a map and help in working out grid references.

Compass An instrument used for orientation and navigation, using a freely rotating needle that indicates the direction of magnetic north.

Coniferous tree A needle-leaved, cone-bearing, mostly evergreen tree, such as pine, spruce, or fir.

Continental air mass Air that has tracked over land and carries comparatively less moisture than maritime air mass.

Contour lines Lines on a map that show points of equal height above sea level, thus detailing the changing height of natural features.

Contour navigation Walking at a constant height around a high natural obstacle.

Contouring See *Contour navigation.*

Contusion A bruise.

Cordage A type of light rope – an essential piece of survival equipment.

COSPAS–SARSAT system A satellite system, made up of low-altitude earth-orbiting (LEOSAR) and geostationary earth-orbiting (GEOSAR) satellites, that picks up signals from distress beacons.

Cramp Painful muscle spasm, often a result of dehydration.

Crampons Spiked metal plates that attach to boots to provide grip on icy surfaces.

Cumulonimbus A type of cloud that starts low in the sky and builds upwards, producing short, heavy downpours or thunderstorms.

Cumulus A billowing, puffy cloud that is generally small and develops on bright, sunny days, indicating fine weather.

Damper bread Yeast-free bread, suitable for making on a campfire.

Deadfall A trap in which a weight falls onto the prey. Also a mass of fallen dead timber.

Defensive swimming A type of swimming technique designed to keep the swimmer protected from obstacles in the water.

Degree A unit of latitude or longitude, equal to $^{1}/_{360}$ of a circle.

Dehydration A low level of water in the body. A very dangerous condition if not reversed.

Deliberate off-set (aiming off) Deliberately aiming left or right of a known feature.

Desalting kit An emergency desalination kit for turning saltwater into freshwater.

Detouring Walking around an obstacle as a method of clearing it, using a point in the distance as a reference.

Dew point The temperature at which water vapour held in the air cools to become liquid.

Dew trap A method of collecting dew to form usable water.

Digging stick A sturdy piece of wood with a pointed end.

Dipping net A fishing net used for scooping up fish that are too small to catch on a line.

Drill and flywheel A friction-based method of starting a fire.

Drip rag A method of collecting rainwater dripping down a tree using cloth.

Drogue A sea anchor used to stabilize a vessel.

Dysentery A water-borne disease that causes severe diarrhoea.

Eastings Vertical grid lines on a map; they increase in value the further east they are.

Emergency Locator Transmitter (ELT) A beacon that can send a distress signal to an orbiting satellite to initiate a rescue; mostly for use in aircraft.

Emergency plan of action (EPA) A document left with relevant authorities that contains vital details about you and your intended route, should you need to be rescued.

Emergency Position-Indicating Radio Beacon (EPIRB) A beacon that can send a distress signal to an orbiting satellite to initiate a rescue; mostly used at sea.

Emergency rendezvous (ERV) A pre-determined point at which all members of a group should meet in the event of separation.

Equator An imaginary circle running around the Earth's diameter that is equidistant from the North and South Poles at all points.

Eskimo roll A technique for righting a capsized kayak.

Faggot A bundle of kindling.

Feather stick A stick that has been feathered to use as kindling and fuel.

Fighter trench A type of snow trench dug into soft snow. Ideally uses a tarpaulin for the roof.

Finnish marshmallow A method of melting chunks of ice over a fire then collecting the water in a container.

Fire tin A tin filled with waxed cardboard, used for lighting a campfire in difficult conditions, or as a basic cooking device.

Firebase A stable, non-flammable base suitable for building a campfire on.

Firefighter's lift A method of moving a casualty by carrying them over your shoulder.

Fire set A set of components necessary to generate a glowing ember that can start a fire.

Firesteel A metal bar which, when hit with a striker, produces a spark for lighting campfires.

Flash burn See *Snow blindness.*

Föhn effect A dry downslope wind which occurs on the lee-side of a mountain or hill.

Four-wheel drive (4WD) A system on a vehicle in which drive is provided by all four wheels, rather than just two; results in superior traction.

Frostbite A serious condition caused by excessive exposure to extreme cold, in which body tissues actually freeze, sometimes to the bone. Often starts with the extremities.

Frostnip The freezing of the top layer of skin, usually on the face and extremities, caused by exposure to extreme cold. Can develop into frostbite if left untreated.

Gaiters Protective fabric that wraps around the lower leg and ankle, to keep out water and protect against sharp objects.

Gill net A fishing net in the form of a mesh stretched across a stream.

Global positioning system (GPS) A hand-held unit that uses orbiting satellites to determine the user's position accurately.

Gourd The shell of hollowed-out, dried fruit; often used to carry water.

Grab bag A pre-prepared bag of essential survival equipment suitable for quickly picking up in an emergency at sea.

Gravity filter An effective method of filtering and purifying water, incorporated into a bottle.

Grid bearing A horizontal direction expressed in degrees east or west of north or south.

Grid magnetic angle (GMA) The difference between magnetic north and grid north, expressed as an angle.

Grid north A northerly direction that runs parallel to vertical gridlines on a map. Different from true north because a map is flat.

Grid reference A method of pinpointing the location of a place or object anywhere on a map, using coordinates provided by a numbered grid system printed on the map.

Grommet A metal-ringed eyelet in clothing.

Groundsheet A waterproof sheet.

Gypsy well A method of using a hole in the ground to filter non-potable water.

Hand jam A crack in a rockface suitable for wedging your whole hand into when climbing.

Handrailing A method of navigating using long linear features on the landscape that run in the general direction of your travel, such as rivers, roads, or paths.

Hank A small coil of cordage.

Heat exhaustion (heat stress) Dehydration that results in dizziness and profuse sweating; often a precursor of heatstroke.

Heatstroke A life-threatening condition brought on by severe overheating.

HELP (Heat Escape Lessening Posture) A posture that reduces heat-loss from the body when floating in water.

Hexamine stove A stove that uses hexamine solid fuel instead of gas or liquid fuel.

Hot platform A method of melting ice over a fire and collecting the liquid in a container.

Hyperthermia An abnormally high body temperature; potentially fatal.

Hyponatraemia An accumulation of excess water in the body, leading to a dangerously low concentration of sodium.

Hypothermia A life-threatening condition caused by exposure to cold. Symptoms usually appear in the following order: uncontrollable shivering (which may stop as the body temperature lowers); irrational or out-of-character behaviour; confusion, mood swings, and withdrawal; pale, cold skin; and slurred speech and/or stumbling. Watching out for and counteracting the early symptoms may help prevent them developing into a full-blown potentially fatal condition.

Hypoxia Inadequate levels of oxygen in the blood or body tissues. Potentially fatal.

Igloo A shelter constructed from cut blocks of snow.

Iodine A chemical used to purify water.

Isobar contours Lines on a weather chart that connect points at which the barometric pressure is the same.

Katabatic winds Cool, light downslope winds that form on clear nights.

Kayak A closed-deck paddle boat for one or more people.

Kindling Small pieces of fuel for a fire, added to burning tinder.

Kukri A large-bladed knife from Nepal, used mainly for chopping.

Land breeze A night-time wind that blows from land to sea.

Lanyard A small cord or rope used for securing or suspending an item.

Latitude The angular distance north or south of the Equator, in degrees along a meridian.

Layering Wearing several light layers of clothes rather than just one or two thick ones. Layers can be added or taken off to maintain a consistent body temperature.

Lean-to A shelter with a sloping roof that leans against a horizontal ridgepole.

Lee-side The side that is sheltered from the wind.

Lensatic compass A type of compass used for precise navigational work.

Longitude The angular distance east or west across the Earth between one meridian and the prime meridian at Greenwich, England.

Machete A large-bladed knife used mainly for chopping, approximately 45–50cm (18–20in) long.

Magnetic north The direction indicated by a magnetic compass.

Magnetic variation The difference between magnetic north and grid north.

Mantelling A technique used for climbing overhanging rock.

Maritime air mass Air that has tracked over the sea, typically containing moisture.

Mayday The internationally recognized distress signal for a grave emergency.

Matchless fire set Small kit that includes everything you need to start a fire – tinder, fuel, and spark device.

Melting sack An improvised sack that can hold snow or ice and is positioned close to a fire to provide water.

Meridian An imaginary arc on the Earth's surface from the North to the South Pole, connecting all locations running along it.

Mid-layer An insulating layer of clothing that sits between the base layer and the outer layer.

Motor impairment A limitation or loss of muscle control or movement.

Naismith's Rule A method of calculating the time it will take to arrive at a destination, taking into account distance and topography.

Natural hollow A natural dip in the ground that can be used as an emergency shelter.

New World Non-Eurasian and non-African regions of the world – specifically the Americas and Australia.

Noggin A piece of wood used to help support a structure, reducing the amount of cordage required.

Non-potable water Water that is not suitable for drinking.

Northings The horizontal grid lines on a map; increase in value the further north they are.

Occluded front A point at which two air masses meet in a weather system.

Old World Europe, Asia, and Africa.

Outer layer Outer clothing; ideally it should keep out rain but let sweat (as water vapour) escape.

Pace counting A method of determining distance; involves knowing how many paces you take to cover a set distance.

Pack animal An animal such as a mule or horse used for carrying heavy loads.

Paracord A useful type of cordage originally developed as rigging lines for parachutes. Usually comprises an outer sheath over an inner section of yarn.

Parang A large-bladed knife from Malaysia, used mainly for chopping.

Pathogen A micro-organism, especially bacteria or fungi, that causes disease.

Permafrost Permanently frozen soil.

Personal Locator Beacon (PLB) A beacon that can transmit a distress signal to an orbiting satellite to initiate a rescue.

Poncho A multi-purpose waterproof outer garment made from one sheet of fabric; also useful for building a basic shelter or bed.

Post-holing A technique for walking in deep snow without snow shoes.

Potable water Water suitable for drinking.

Potassium permanganate A chemical used for a variety of tasks, from lighting fires to purifying water.

Precipitation Moisture in a cloud that may fall as rain, snow, or hail.

Prismatic compass See *Lensatic compass.*

Protractor A device for measuring angles.

Psychogenic shock A very high level of psychological and emotional stress brought on by a sudden disaster situation.

Pulk A simple plastic sled used for carrying equipment over snow.

Pulmonary oedema Swelling and fluid accumulation in the lungs that causes potentially fatal breathing problems.

Pygmy hut A domed hut made from a circle of bent saplings or limber poles and thatched with natural materials.

Pygmy roll A method of rolling fibres together to make cordage.

Quicksand A bed of loose, wet sand that yields easily to pressure and can engulf anything on its surface.

Quinzhee A basic dome-shaped, snow-covered overnight shelter.

Ranger flint and steel A rod of flint which, when hit with a striker, produces a spark for lighting campfires.

Recovery position The position an unconscious person should be placed in to minimize further injury and help recovery.

Reflector A device for directing heat from a campfire towards shelter.

Reverse-osmosis pump A survival pump system that turns saltwater into freshwater.

Ridgepole A long horizontal pole that forms the apex of a roof.

Romer measure See *Compass scale.*

Salting A salt marsh.

Scrambling Climbing without ropes.

Scrape A type of basic shelter built in a natural hollow or depression in the ground.

Sea breeze A daytime wind that blows from sea to land.

Secondary drowning Potentially fatal biological changes that can take place in the lungs after a person nearly drowns, caused by the body's response to inhaling water.

Shock A life-threatening condition that occurs if the circulatory system fails, often triggered by severe bleeding, burns, or sudden cold.

Show-stopper A failure in planning that could cause delays to, or stop a trip or expedition.

Sighting mirror An advanced component on a high-end compass used for helping in orientation and navigation.

Signal fire A fire specifically designed to produce lots of smoke to attract attention.

Signal flare A hand-held distress flare that gives off orange smoke in daylight and a bright light at night.

Signal mirror A simple signalling device that works by reflecting sunlight.

Silva compass A type of basic compass useful for hiking.

Sip well A method of extracting water trapped under rocks by sandy ground.

Skidoo A motorized vehicle used for travelling over snow and ice.

Slingshot A hunting weapon that fires small stones and is made from a Y-shaped stick with an elastic strip stretched between the prongs.

Snare A wire noose used for trapping animals.

Snow blindness Eye damage caused by ultraviolet light reflected off snow or water.

Snow cave An effective cave-like shelter dug into compacted snow on the lee-side of a hill.

Snowshoes Wide-soled strap-on shoes designed to stop you sinking into snow.

Soak A water source found close to rivers and creeks in areas that are usually lower-lying than the existing water table.

Solar still A method of extracting drinking water from any source of moisture by a process of evaporation and condensation.

Space blanket A plastic blanket covered in aluminum foil that can be used as a shelter or a reflective signalling device; to carry, store, and heat water; or to cook in.

Stand-off A technique for navigating to a point that may be hard to locate, by orientating from a nearby prominent feature.

Standing deadwood Wood from a tree that has died but is still standing. Hazardous, as branches or the whole tree may fall, but makes excellent firewood.

Straddling A climbing technique used for ascending wide chimneys.

Stratus Dense, grey cloud that forms a sheet. Rain can fall for long periods of time if the cloud has depth.

Strobe light An effective signalling device in the form of a rapidly flashing LED light.

Survival blanket See *Space blanket.*

Survival kit Essential items for survival carried on your person.

Survival straw A compact emergency water purifier.

Survival suit A waterproof, buoyant suit designed for survival in open water.

Survival tin An essential, compact container carried on the person that addresses basic survival needs: protection, location, water, and food.

Taiga A subpolar region characterized by coniferous forest.

Tarp Shortened form of "tarpaulin", a type of sturdy waterproof sheet.

Temperate climate A climate without extremes of temperature or rainfall.

Tetanus A potentially lethal infection caused by bacteria that live in soil.

Throwing star An improvised four-pointed hunting weapon used for throwing at prey.

Tinder A dry, light, combustible material that is the first fuel used when lighting a fire.

Topographic map A map that shows the main features on a landscape, and elevation via use of contour lines.

Transpiration bag A method of evaporating, condensing, and siphoning off water from foliage using a plastic bag and sunlight.

Triangulation A method of determining your location by taking bearings with a compass from recognizable landscape features.

True north The direction of a meridian of longitude that converges on the North Pole.

Tundra A polar region characterized by permafrost and stunted vegetation.

Universal Edibility Test (UET) A test that checks whether a plant is safe to eat via a methodical process of testing small amounts on the skin, in the mouth, and in the stomach.

UV active purification Purifying water via a battery-operated ultraviolet (UV) purifier.

UV passive purification Purifying water by leaving filled bottles in strong direct sunlight. The UV rays eventually kill most pathogens.

Vegetation bag See *Transpiration bag.*

VHF radio Very high frequency radio transmitter; usually used at sea.

Water balance The difference between water lost from the body through sweat and the water taken in through drinking.

Water purifier A device that filters and purifies water by pumping it through micro filters, chemicals, or a combination of both.

Weather chart A map of major weather systems and their predicted direction of travel.

Wicking material A material that moves moisture away from your body to evaporate.

Wickiup A small hut made from straight poles lashed together at the top, with an interwoven framework covered with animal hides or grass.

Witch's broom A cleft stick stuffed with thin, dry bark suitable for lighting a fire.

Withies The strong, flexible stems of plants, such as willow, birch, ash, and hazel.

Zig-zag route Method of climbing a steep slope using a zig-zag path to reduce the effort required to achieve the climb. Also effective when walking down steep slopes.

INDEX

A

- A-frames
  - forest 164–65
  - front opening 165
  - jungle 168
  - poncho shelters 159, 167, 175
- abrasions (grazes) 263
- abandoning ship 250, 252–55
- ABC (airway/breathing/circulation), first aid 260, 276–77
- aerobic exercise 19
- agaves 193, 282–83
- air quality, caves 157
- aiming off 73
- air masses 79
- allergies and anaphylactic shock 266, 267, 274
- altitude sickness 33, 273
- altocumulus 80, 81
- altostratus 80, 81
- amoebic dysentery 187
- amphibians and reptiles
  - catching and eating 203, 218, 224, 294–95
  - dangers from 243, 266, 300–01
  - *see also* snakes
- anabatic winds 82
- analogue watches, navigating using 77
- anaphylactic shock 266, 267, 274
- anchor points, winching 103
- anchors, sea 39, 251
- animals
  - dangers from 115, 157, 189, 242–43, 300–3
  - edible 290–99
  - locating water 189
  - navigation aids 77
  - traveling with 42, 96–97, 98–99
  - weather forecasting 83
  - *see also* specific animals (e.g. bears); specific types (e.g. mammals)
- ants 219, 298–99
- arbor knot 142
- arm injuries 270
- ash
  - for soap 117
  - for stomach pain 269
- asthma 275
- automatic fires 124
- avalanches 32, 246–47

B

- bacillary dysentery 187
- backpacks 41, 42, 44–45, 108
- bacterial infections 185, 187, 212, 220, 262, 267, 301, 303
- baits and lures 208–09, 223, 229
- baking food 204, 205
- bamboo 284–85
  - fire saw 133
  - lean-tos 169
  - rafts 107
  - steamer 204
  - water sources and collection 192, 199
- bark
  - cordage 138
  - tea 269
  - tinder and kindling 120, 121, 238
- bats 157, 290–91, 300–01
- beaches 39, 177
- beacons, phones, and radios 39, 43, 61, 232–33, 236–37, 246, 250, 252
- bearings, compass 64–65, 68–71
- bears 28, 36, 157, 242
- Beaufort scale 83
- beds and bedding 41, 52–53, 156, 166–67
- bees 224, 267, 302–3
- belaying 91
- berms 177
- big cats 32, 242
- bilharzia 187
- birds
  - catching and eating 203, 204, 205, 218, 226–29, 292–93
  - locating water 189
- bites and stings 116, 266–67, 268, 300–03
- bivy sacks 53
  - bivy pup tent 160–61
  - hooped bivy 41, 54
- blankets, space 53, 61, 155, 175
- blazes (trail markers) 87
- bleeding
  - birds, for food 228
  - first aid 263–64
  - mammals, for food 222
- blended fuel stoves 58
- blisters 262, 268
- boats *see* ships and boats
- body fat 56
- boiling
  - food 204, 212
  - water 118, 185, 200
- bolas 226
- bones, joints, and muscles injuries 233, 270–1
- boot skiing 94
- boots 46, 49, 114
- bothy bags 161
- bottle filters 200
- bottle traps, fishing 211
- bough fence traps 227
- bow drills 130–33
- boxing an object (detouring) 73
- branches
  - as snowshoes 95
  - vehicle recovery 102
- bread, damper 204
- breathable waterproof fabric 48
- breathing difficulties
  - first aid 275
  - in shelters 157, 181
- breezes and winds 38, 79, 82–83
- brown bears 28, 242
- bruising 262, 263
- brushwood rafts 106–07
- buddy-buddy sleeping system 163
- buoyancy aids (floats) *see* lifejackets
- burns and scalds 265
- bushcraft knives 146–49
- butane stoves 58
- button compasses 68
- button ties 161

C

- cacti 193, 282–83
- cage traps, figure 4 226
- cairns 87
- calorie needs 56, 202
- camels 42, 99
- camping soap 117
- camping stoves 58–59
- camps
  - organization and administration 112–13, 114–15
  - *see also* specific areas (e.g. latrines)
- canoeing and kayaking 42, 104–05
- capacity, packs and bags 41, 42, 44
- capsizing 104, 105
- carbohydrates 56
- carbon dioxide 157
- carbon monoxide poisoning 181
- cardinal points 68, 76, 77
- cardiopulmonary resuscitation (CPR) 277
- cats, big 32, 242
- caves 157, 181
- cell phones 43, 237
- chainsaws 60, 150
- char cloths 122–23
- char test, softwoods and hardwoods 131
- charts
  - maps and map reading 43, 64–65, 66–71
  - weather 78
- chemical flames 127
- chemical lightsticks 61, 240
- chimneying 91
- chlorine tablets 201
- choking 275
- cholera 187
- cirrostratus 80, 81
- cirrus 80, 81
- clay baking, food 205
- climate 78
  - *see also* specific environments (e.g. temperate environment)
- climbing and scrambling 90–91
- clothing 46–49, 94
  - cleanliness 116
  - cultural differences 22
  - on fire 265
  - hats 43, 46, 47, 48
- clothing, improvising with
  - buoyancy aids 255
  - cordage 137
  - dipping nets 210
  - slings 261
  - stretchers 279
  - water collection and treatment 185, 190, 192, 195, 199, 200

clouds 78, 79, 80–81
coastal caves 157
coasts *see* marine environments
cold environments 36–37
  clothing 47, 94
  cold shock 254
  dos and don'ts 51, 245
  edible plants 286–87
  hypothermia and frostbite 21, 29, 33, 37, 252, 273
  tents 55
  water sources 194–95
  see also ice; snow
cold fronts, weather 78, 79
cold shock 254
cold-weather clothing 47, 94
commando saws 151
commando-style match striking 126
communications equipment 43, 61, 232–33, 236–37
compacted snow trench 178
compasses and bearings 43, 64–65, 68–71, 74–75
components of fire 118, 238
condoms 61
conjunctivitis 212
constellations, navigating by 77
continental air masses 79
contour lines 66, 67
contouring 73
contusions (bruising) 262, 263
cooking
  fires, rocks, and pits 119, 204–05
  fish preparation 212–13
  invertebrates 219, 224–25
  mammals preparation 220–23
  stoves 58–59
coping strategies 20–21, 113, 234–35, 259
cordage 136–41
cordless scrape shelters 174
cougars 32, 242
CPR (cardiopulmonary resuscitation) 277
cardiovascular fitness 18, 19
crabs 225
cramp 270
crickets and grasshoppers 224, 298–99
crocodilians 189, 242
crossing rivers 88–89, 101, 108–09
crustaceans 225
cryptosporidium 187
cumulonimbus 80
cumulus 80, 81
cuts and grazes 263
cutting tools 60, 146–51
cyalumes 61, 240

D

Dakota hole fires 125
damper bread 204
day packs 42, 44
deadfalls 115
deadwood, standing 115, 121, 131
defensive swimming 108
dehydration and rehydration 187, 190, 193, 196, 253, 269, 272
deliberate off-set 73
denatured alcohol stoves 59
desalting kits 197
desert environments 34–35
  clothing 47
  dos and don'ts 129
  edible plants 282–83
  four-wheel driving 35, 101
  night walking 86
  shelters 174–75
  water sources 34, 35, 193
detouring (boxing an object) 73
dew harvesting and traps 190, 196
dew point 80
diarrhea and vomiting 187, 203, 269
digging
  ponds 199
  roots and tubers 206
  underground water 193
  vehicle recovery 102
dipping nets 210
disaster situations and emergencies
  causes 16–17, 26, 84
  dos and don'ts 27, 51, 93, 129, 173, 215, 232–33, 245
  emergency plans of action (EPAs) 24–25, 250
  emergency rendezvous points (ERVs) 72
  environmental hazards 246–49
  four-wheel driving (4WDs) 35, 103
  gear priorities 43
  mental preparation and responses 20–21, 234–35, 259
  PLWF (protection, location, water, and food) 11, 27, 258
  rescue services 24
  rescue signaling 39, 43, 232–33, 236–41, 250, 252
  survival at sea 16–17, 39, 196–97, 214–15, 250–55
  survival kits 60–61
  SURVIVAL strategies 234–35
  water sources and rationing 39, 187, 188–97, 199, 251
  see also first aid
diseases *see* infections
distance measuring 67, 72
distilling water 185
distress signals 39, 232–33, 236–41, 250, 252
dogs 17, 96–97, 99
dome signal fires 238–39
dos and don'ts
  disaster situations and emergencies 27, 51, 93, 129, 173, 215, 232–33, 245
  fires 118, 119, 124
double overhand knot 143
double sheet bend 144
down sleeping bags 52–53
downhill walking 86, 87
drawing birds (food preparation) 229
driftwood shelters 176–77
drip rags 192
driving techniques, 4WDs 35, 100–03
drogues 39, 251
drowning 254, 255, 273
drum rafts 107
duck traps 227
dunes, sand 35, 177, 189
dysentery 187

E

E. coliform infection 187
earthquakes 247
earthworms 224, 298–99
eastings 67
edible plants 203, 206–07, 280–89
eggs 226
electric current, magnetizing compass needles 75
electrical storms 81, 249
elements of fire *see* fuel; kindling; tinder
elephants 99, 189, 242
ember pans 132
emergencies *see* disaster situations and emergencies
emergency signaling codes 240, 241, 252
emergency transmitters 39, 232–33, 236–37, 246, 250, 252
emergency water packs 199
emotional responses
  disaster situations 20–21, 113, 234–35, 259
  team dynamics 23
environmental hazards 246–49
environments
  understanding and preparing for 23, 26
  see also specific environments (e.g. cold environments)
EPAs (emergency plans of action) 24–25, 250
EPIRBs (emergency position indicating radio beacons) 236, 250, 252
equipment *see* gear and equipment
ERVs (emergency rendezvous points) 72
eskimo rolls 104
Evenk knot 144
exercise and fitness 18–19
exposure, first aid 272–73
external-frame packs 44
extreme weather 246, 248–49
eyes
  care 116
  problems and injuries 212, 265

F

fallen trees, shelters 163
fatigue 21, 235
fats 56
fear 234, 235
feather lures 209
feather sticks 120–21
features, use in navigation 70–71, 72, 73
feelings *see* emotional responses
felling trees 148–51
feral animals 242
fibre (nutrient) 56
fighter trench 179
figs 171, 284–85
filleting fish 212–13
filtering water 184, 200, 201
Finnish marshmallow 195
fire saws, bamboo 133
fire tins 123
fire wads 123
firefighter's lift 278
fires
  benefits 27, 113
  burns, first aid 265
  in caves 157
  clothing on fire 265

  components 118, 238
  cooking over 119, 204–05, 213
  dos and don'ts 59, 118, 119, 124
  elements *see* fuel; kindling; tinder
  fire reflectors 119, 164
  firewood 114, 148–49
  forest fires 248
  ice and snow melting 194–95
  lighting 125
  preparing and making 118–23
  signal 30, 115, 238–39
  snapping wood with 151
  spark production and ignition devices 113, 118, 126–27, 130–33
  types 124–25, 238–39
  wildfires 34
fireset can, matchless 118
firesteel 113, 127
first aid 258–9
  ABC (airway/breathing/circulation) 260, 276–77
  anaphylactic shock 274
  asthma 275
  bites and stings 266–67, 268, 300–03
  bones, joints, and muscles injuries 233, 270–71
  breathing difficulties 275
  burns and scalds 265
  cardiopulmonary resuscitation (CPR) 277
  casualty transport 99, 278–79
  choking 275
  dehydration and rehydration 269, 272
  diarrhea and vomiting 269
  drowning 255, 273
  exposure 272–73
  fish-hook removal 208
  hyponatraemia 187
  kit 60, 260–61
  natural remedies 263, 268, 269
  parasites 269
  priorities 260
  psychological help 21
  recovery position 276
  shock 274
  unconsciousness 276–77
  wounds and bleeding 262–65, 266
  see also poisoning
fish
  catching and eating 61, 203, 204, 205, 208–13, 296–97
  dangers from 38, 243, 300–01
  as water source 196
fitting backpacks 45
fixed dial compasses 68
flares, signal 239, 241
flash burn 265
flash floods 115, 249
flashlights 43, 60
fleeces 46, 47
flesh wounds 262–5, 266
flint and steel 60, 113, 118, 127
floating method, improvised compasses 74
floats (buoyancy aids) *see* lifejackets
floats and sinkers (fishing equipment) 209
floods 115, 249
flysheets 55
foam mats 53
fog 81
Föhn effect 82
folding saws 151
food
  amphibians and reptiles 203, 218, 224, 294–95
  birds 203, 204, 205, 218, 226–29, 292–93
  cooking fires and stoves 58–59, 119, 204–05
  fish 203, 204, 205, 208–13, 296–97
  food poisoning 187, 269
  hunger 20, 202, 235
  insects 203, 219, 224, 298–99
  mammals 203, 204, 205, 216–23, 290–91
  nutrients and calorie needs 56, 202
  plants 203, 206–07, 280–89
  protecting 114
  purchased 57
  shellfish 225, 298–9
  survival without food 27, 56, 57, 202, 234, 235, 259
foot holds, scrambling and climbing 90
footwear 46, 48, 49, 114
  snow shoes 94–95
foreign objects
  in eyes 265
  in wounds 263
forests
  forest fires 248
  mountain 32
  taiga 36, 37
  temperate 28, 29, 162–65
  tropical *see* rain forests and jungles
found meat 204
four-wheel driving (4WD) 35, 42, 100–03
friction, as fire starter 130–33
frogs 224, 243, 294–95
fronts, weather 78, 79
frostbite 33, 273
fuel
  fires 114, 118, 121, 124, 125, 148–49, 238–9
  stoves 58–9
fungi and mushrooms 206, 263

G

gaiters 48, 94
gas-fueled stoves 58
gear and equipment
  backpacks 41, 42, 44–45, 108
  clothing 46–49, 94
  communications and distress signaling 39, 43, 232–33, 236–37, 240–41, 246, 250, 252
  first aid kit 60, 260–61
  food 57
  keeping it dry 108–09
  packing 41, 42, 45, 98, 136, 137–41
  priorities 40–41, 43
  survival kits 60–61
  training with 23, 134–35
  ultralight camping 41
  water containers and treatment 60, 197, 198, 201
  weight 41, 42
  see also specific items (e.g. surgical tubing)
giardiasis 187, 269
gill nets 210–11
glaciers 33, 94
global positioning systems (GPS) 65, 75, 237
gloves 46, 47
glowsticks 61, 240
gourds 199, 282–83
GPS (global positioning systems) 65, 75, 237
grab bags 250
grasshoppers and crickets 224, 298–99
grasslands 28, 34, 35
grazes (abrasions) 263
grid lines, references and bearings 66–71
grips, knives 147
ground-to-air markers 114, 241
groups and group dynamics 23, 100, 253, 254
grubs 224, 298–99
gum and teeth care 19, 116
gunshot wounds 263
gutting mammals 220, 223
gypsy wells 191

H

hair
  hair care 116
  magnetizing compass needles 74
half hitches 145
hammocks 54, 166
hand holds, scrambling and climbing 90
hand injuries 271
handrailing 73
hanks, cordage 136, 137
harpoons 211
hats 43, 46, 47, 48
hearth boards, bow drills 131, 132
heat escape lessening position (HELP) 254
heat exhaustion 272
heat loss reduction 37, 254
heatstroke 272
helicopter rescues 114, 236–37, 241
heliographs 241
helminths 269
HELP (heat escape lessening position) 254
Hennessy hammocks 166
hepatitis 187
hexamine fuel 59, 118, 238
hides 222, 223
hiking *see* walking and hiking skills
hiking boots 49
hippos 189, 242
hitches and knots 135, 142–45, 210
hollows, scrapes and trenches shelters 156–57, 174–75, 176–77, 178–79
hooks, fishing 208
hookworms 187
hooped bivy 41, 54
hooped shelters 161
hopping insects 224
horse chestnut leaves soap 117
horses 42, 99
hot platforms 194
hot rocks cooking 205
huddle crossing, rivers 89
hunger 20, 202, 235

hunter fires 125
hunting and trapping mammals 216–18
hurricanes 38, 248–49
huskies 96–97, 99
huts
  jungle 170–71
  pygmy 171
  wickiups 171
hybrid boots 49
hydration systems 198
hygiene 116–17, 269
hyperactivity 21
hyponatremia 187
hypothermia 21, 29, 33, 37, 252, 273

I

ice
  four-wheel driving 101
  ice climbing 91
  ice falls 115
  ice fishing 209
  ice walking 94
  icebergs 196
  sea ice 36, 196
  water source 36, 194–95, 196
igloos 180
ignition devices and spark production 113, 118, 126–27, 130–33
improvising 235
  backpacks 45
  compasses 74–75
  goggles 265
  signal mirrors 241
  see also clothing, improvising with
incised wounds 263
infections
  bacterial 185, 187, 212, 220, 262, 267, 301, 303
  diarrhea and vomiting 187, 269
  first aid concerns 261, 262, 263, 265, 301
  parasites 187, 222, 223, 269
  vaccinations 22
  water-borne 185, 187
infectious hepatitis 187
injuries 233, 235, 270–71
  see also first aid
insects
  bites and stings 116, 267, 302–03
  catching and eating 203, 219, 224
  locating water 189
  maggot therapy 262
  mosquitoes 31, 115, 302–03
  navigation aids 77
International Distress Signal whistle 240
Inuit snow-melting techniques 195
invertebrates 203, 219, 224, 298–99, 300–1, 302–3
iodine 60, 201
isobars 78
isobutene stoves 58
ivy, poison 268

J

jackets 46, 47, 48
  lifejackets and buoyancy aids (floats) 39, 104, 108, 109, 250, 254–55
jellyfish stings 267, 300–01
jointing mammals 223
joints, bones, and muscles injuries 233, 270–71
journey planning 22
jungles *see* rainforests and jungles

K

katabatic winds 82
kayaking and canoeing 42, 104–05
kernmantle rope 136
kerosene 269
killing
  birds 228
  rabbits 216
kindling 118, 119, 120, 121, 124, 125, 133, 238–39
knives 146–49
knots 135, 142–45, 210

L

lacerations (tears) 263
lakes 189
land and sea breezes 82
larvae, ant 219
latrines 114, 117
layering clothing 46
lean-tos
  bamboo 169
  forest 162–63
  one-person 165
leaves
  cordage 139
  roofs 170–71, 192
  soap-making 117
LED strobe lights 240
ledges, snow 181
leeches 267
leg injuries 271
legends, topographic maps 66, 69
lensatic compasses 69
leptospirosis 187
lichens 286–87
lifejackets and buoyancy aids (floats) 39, 104, 108, 109, 250, 254–55
liferafts 39, 196, 251, 252–53
lifts, moving casualties 278–79
light windmill 240
lighters 119, 126
lighting fires 125
lightning 81, 249
lights and lightsticks 61, 240
limitations, personal 18
line crossing, rivers 89
line fishing 61, 209
liners, sleeping bags 53
lipolysis 56
liquid-fueled stoves 58, 59
liver 220, 222, 223, 229
lizards 294–95
llamas 99
location 11, 27, 233
  see also site selection
location aids and rescue signals 39, 43, 114, 232–33, 236–41, 250, 252
log cabin signal fires 239
log rafts 107
log-splitting 114, 148–49
lures and baits 208–09, 223, 229
Lyme disease 267, 303

M

machetes 146
maggot therapy 262
magnetic bearings 68–71
magnetic variation 69
magnetizing compass needles 74–75
magnifying glasses 126
mammals
  catching and eating 203, 204, 205, 216–18, 220–3, 290–1
  dangers from 242–43, 266, 300–01
  locating water 189
  melting snow 195
  sinew cordage 139
  weather forecasting 83
  see also specific animals (e.g. bears)
man-made cordage 61, 136–37
mantelling 91
maps and map reading 43, 64–65, 66–71
  weather charts 78
  see also navigation
Marcandonatos, Costa 128
marine environments 38–39
  beaches 39, 177
  coastal caves 157
  drowning 254, 255, 273
  edible plants 288–89
  sea-sickness 21
  shoreline shelters 176–77
  single-wall fishing traps 211
  stings and bites 267, 300–01
  swimming 108–09, 254–55
  tides 157
  water sources 39, 196–97
  winds 82
  see also sea journeys
maritime air masses 79
marshmallow, Finnish 195
matches 60, 126
matchless fireset cans 118
mayasticks 123
measuring distances 67, 72
meat *see* specific types (e.g. birds)
melting sack 195
mental preparation and responses, disaster situations 20–21, 234–35, 259
minnow traps 211
mirrors, signal 241
money 22, 61
moon, navigating by 77
mosquitoes 31, 115, 302–03
mosses 286–87
motorized transport 35, 42, 97, 100–03
mountain environments 32–33
  altitude sickness 33, 273
  avalanches 32, 246–47
  cold environments 36
  deserts 35
  dos and don'ts 93, 245
  scrambling and climbing 90–91
  temperate areas 28–29
  winds 82
muddy terrain, four-wheel driving 101
mules 99
multi-fuel stoves 59
multi-tools 61

- muscles, bones, and joints injuries 233, 270–71
- mushrooms and fungi 206, 263
- mussels 225, 298–99

N

- nails, fishing hooks 208
- Naismith's rule 72
- natural cordage 138–41
- natural remedies 263, 268, 269
- navigation 64–65
  - compasses 68–71, 74–75
  - GPS 65, 75
  - maps and map reading 43, 64–65, 66–71
  - natural (sun, moon and stars) 76–77
  - route finding 72–73
  - trail markers (blazes) 87
- needle tea 207
- needles 60, 61, 68, 74–75
- nets
  - bird-trapping 226
  - fishing 210–11
- nettles
  - cordage 138–39
  - stinging 268, 280–81
- night line fishing 209
- night walking 86
- night-flying insects 219
- nimbostratus 80, 81
- noose sticks 218
- northings 67
- nutrients and calorie needs 56, 202

O

- oak, poison 268
- oases 34
- occluded fronts 78
- one-handed striker 127
- one-person lean-tos 165
- one-pole poncho shelter 158
- orienteering compasses 68
- overhand knots 143
- oxygen, lack of 157, 234, 259
- oysters 225

P

- pace counting 65, 72
- pack animals 42, 98–99
- pack frames, improvised 45
- packing
  - backpacks 41, 42, 45
  - cordage 136, 137
  - fuel 58
  - loading pack animals 98
- packs 41, 42, 44–5, 108
- paddles and paddling
  - kayaking and canoeing 104, 105
  - rafts 107
- pain 235
  - *see also* first aid
- painters, liferafts 252
- palm grubs 224, 298
- pants 46, 47, 48, 94
- para scrapes 175
- paracord (parachute cord) 61, 136–37
- paraffin 269
- parangs 146
- parasites 187, 222, 223, 269
- pepper sprays 243
- periods (menstruation) 117
- personal hygiene 116–17
- personal locator beacons (PLBs) 39, 236–37
- petroleum jelly 61
- pheasants 227, 228
- phones 43, 237
- physical fitness 18–19
- pine cones 83
- pits, cooking in 205
- planning *see* preparation and planning
- plants and vegetation
  - cordage 138–41
  - edible 203, 206–07, 280–89
  - poisoning 203, 206, 207, 268, 304–05
  - remedies 263, 268, 269
  - signal fires 238–39
  - vegetation bags 190
  - water sources 189, 190, 192, 193
  - weather forecasting 83
  - *see also* specific environments (e.g. temperate environment); specific parts (e.g. roots); specific types (e.g. trees)
- PLBs (personal locator beacons) 39, 236–37
- plucking birds 228–29
- PLWF (protection, location, water, and food) 11, 27, 258
- poacher's method, pheasant meat 228
- pocket chainsaws 60, 150
- poison ivy, oak and sumac 268, 304–05
- poisoning
  - basic first aid 269
  - carbon monoxide 181
  - fish 300–01
  - food poisoning 187, 269
  - mushrooms and plants 203, 206, 207, 268, 304–05
  - shellfish 225
  - snake bite 266, 300–01
- polar areas 36, 37
- polar bears 36, 242
- ponchos and tarpaulins
  - beds 166–67
  - brushwood rafts 106–07
  - poncho floats 109
  - rainwater collection 188, 196
  - shelters 54, 158–59, 161, 166–68, 174–77
  - wet-weather clothing 48
- ponds 189, 199
- porcupines 2 66, 290-91
- post-holing 94
- potassium permanganate 60, 127, 201
- preparation and planning 6–17
  - camps 112–13, 114–15
  - emergency plans of action (EPAs) 24–25, 250
  - journey planning 22
  - mental 20–21, 234–35, 259
  - physical 18–19
  - temperate environments 28
  - training 16–17, 23, 235
  - worst-case scenarios 24, 40, 234–35, 250
  - *see also* specific aspects (e.g. gear and equipment)
- prickly pears 193, 282–83
- propane stoves 58
- protection, location, water, and food (PLWF) 11, 27, 258
- protein 56
- protozoa 187, 269
- Prusik knot 210
- psychogenic shock 20
- psychological responses
  - disaster situations 20–01, 113, 234–35, 259
  - team dynamics 23
- pulks 97
- pumps, reverse-osmosis 197
- puncture wounds 263
- purifying water 60, 185, 201
- pygmy huts 171
- pygmy roll cordage 141

Q

- quick-release knots 145
- quicksand 247
- quinzhees 180–81

R

- rabbits 216, 220–21, 290–91
- radios 61, 237
- rafts 106–7
  - liferafts 39, 196, 251, 252–53
- rain 79, 81, 82
- rain forests and jungles 30–31
  - dos and don'ts 173
  - jungle-trekking boots 49
  - shelters 54, 166–71
  - water sources 192
- rainwater 188, 192, 196
- ranger's flint and steel 60, 113, 118, 127
- rats 220, 290–91, 300–01
- rattan leaves 170
- razors and razor blades 60, 61, 74–75
- razorstrop fungus 263
- recovery position 276
- recovery techniques, vehicles 35, 102–03
- reels, fishing 208
- rehydration and dehydration 187, 190, 193, 196, 253, 269, 272
- reptiles *see* amphibians and reptiles
- rescue services 24, 236–37
- rescue signaling 39, 232–33, 236–41, 250, 252
- resection 71
- respiratory problems
  - first aid 275
  - in shelters 157, 181
- reverse-osmosis pumps 197
- rivers
  - flash floods 115
  - foot crossings 88–89
  - four-wheel driving 101
  - kayaking and canoeing 42, 104–05
  - mountain environments 32
  - swimming 108–09
  - temperate environment 28
  - tropical environment 31
  - water sources 28, 32, 189
- roasting food 204, 213
- rock climbing and scrambling 90–91
- rocks
  - hot platforms 194
  - hot rocks cooking 205

  - rock holes 189
  - rockfalls 115
  - single-wall fishing traps 211
  - stone-lined fires 124
  - trail markers and cairns 87
  - transpiration and vegetation bags 190
  - water distillation 185
- rodents (rats) 220, 290–91, 300–01
- rolled cordage 141
- roofs 156–57, 170–71, 192
- roosting birds, trapping 218, 227
- roots
  - cordage 138
  - food 206
  - soapwort 117
  - water sources 193
- ropes
  - climbing techniques 91
  - crossing rivers 89
- roundworms 222, 269
- route finding 64–65, 72–73
  - *see also* navigation

S

- salmonella 187
- saltwater 38, 185, 196–97, 251, 253
- sand
  - driving on 101
  - dunes 35, 177, 189
  - quicksand 247
  - sandstorms 249
- sandals 49
- satellite communications 39, 236–7, 250, 252
- savannah 34, 35
- saws, bamboo fire 133
- saws (cutting tools) and sawing 60, 150–51
- scalds and burns 265
- scale, topographic maps 66, 67
- scalp care 116
- scalp wounds 264
- scrambling and climbing 90–91
- scrapes, hollows, and trenches shelters 156–57, 174–75, 1767–7, 178–79
- scree 32, 87
- sea anchors 39, 251
- sea ice 36, 196
- sea journeys
  - abandoning ship 250, 252–55
  - radios and beacons 39, 236, 237, 250, 252
  - seasickness 21
  - survival at sea 16–17, 39, 196–97, 214–15, 250–55
  - survival suits 39, 199, 250
  - *see also* marine environments
- sea and land breezes 82
- sea urchin spines 267
- seat carries 279
- seawater 38, 185, 196–97, 251, 253
- seaweeds 288–89
- secondary drowning 255, 273
- self-striking fishing line 209
- setting maps 69
- shadow sticks 76, 77
- sharks 38, 243
- sharpening and sharpening stones 147
- sheaths, knives 146
- shellfish 225, 298–99
- shelters 154–55
  - bivys 41, 54, 160–61
  - button-ties 161
  - caves 157, 181
  - cold environments 55, 178–81
  - deserts 174–75
  - hammocks 54, 166
  - hollows, scrapes and trenches 156–57, 174–75, 176–77, 178–79
  - huts 170–71
  - jungle 166–71
  - ponchos or tarpaulins 54, 158–59, 161, 166–68, 174–77
  - positioning on site 29, 32, 115
  - shoreline 176–77
  - snow 178–81
  - temperate forest 162–65
  - tent pegs 159
  - tents 41, 54–55
  - use-anywhere 158–59
- shigellosis 187
- ships and boats
  - kayaking and canoeing 42, 104–05
  - liferafts 39, 196, 251, 252–53
  - rafts 106–07
  - safety procedures and emergency plans of action 250
  - sea-sickness 21
  - *see also* sea journeys
- shock 274
  - anaphylactic 266, 267, 274
  - cold 254
  - psychogenic 20
- shoelaces 137
- shorelines *see* marine environments
- shorts 47
- showers 116–17
- shrimps 225
- Siberian hitch 144
- signal fires 30, 115, 238–39
- signal flares 239, 241
- signal mirrors 241
- silk, magnetizing compass needles 75
- Silva compasses 68
- simple tie-off knot 143
- sinews, cordage 139
- single-wall fishing traps 211
- sinkers and floats 209
- site selection and organisation 29, 32, 112–13, 114–15
  - *see also* specific areas (e.g. latrines)
- skidoos 97
- skinning
  - fish 213
  - mammals 221, 222
- sky colour 83
- sleds 96–97
- sleeping bags 41, 52–53
- sleeping systems *see* beds and bedding; shelters
- slings 261, 270
- slingshots 218
- slip knot 142
- slugs 224
- smoke production, signal fires 30, 238–39
- snails 224, 298–99, 300–01
- snake hole fires 125
- snakes
  - bites first aid 266
  - dangers from 34, 157, 189, 243, 300–01
  - food source 224, 294–95
- snapping wood 151
- snares 60, 216–17, 227
- snow
  - avalanches 32, 246–47
  - four-wheel driving 101
  - mountain environment 32–33
  - navigation aids 77
  - snow blindness 265
  - snow shelters 178–81
  - snow shoes 94–95
  - traveling over 94–97
  - water source 194–95
- snowman's head 195
- soaks 189
- soapwort roots 117
- socks 48
- solar stills 191, 197, 251
- space blankets 53, 61, 155, 175
- spark production and ignition devices 113, 118, 126–27, 130–33
- spiders 377, 157, 302–03
- spinal injuries 271
- sprains and strains 270
- squirrels 217, 221, 290–91
- stab wounds 263
- staffs, walking 72, 85, 89
- stand off/attack point 73
- standard compasses 68
- standing deadwood 115, 121, 131
- star-shaped fires 125
- stars, navigating by 77
- steaming food 204
- steel, flint and 60, 113, 118, 127
- stills, solar 191, 197, 251
- stinging nettles 268, 280–81
- stings and bites 116, 266–67, 268, 300–03
- stock 204, 212
- stone-lined fires 124
- storms 81, 249
- stoves 58–9
- straddling, chimney climbing 91
- strains and sprains 270
- stratocumulus 80, 81
- stratus 80, 81
- streams 32, 189
- strength and stamina 18–19
- stretchers 99, 279
- stretching exercises 19
- striker, one-handed 127
- striking matches 126
- string measuring, map reading 67
- strobe lights 240
- sugar 127
- suicide 21
- sumac, poison 268
- sun and sunlight
  - navigating by 76
  - protection from 43, 47, 175
  - sundials 76
  - water purification 201
- surgical tubing 61, 191, 197
- survival enemies 235
- survival fitness 18–19
- SURVIVAL strategies 234–35
- survival straws 201
- survival suits 39, 199, 250
- survival swimming 108–9, 254–55
- survival tins 40, 43, 60–61

surviving mentally 20–21, 234–35, 259
survivor's third leg (walking staff) 72, 85, 89
suspension method, improvised compasses 74
sustainment load 43
swamps 30, 31, 247
swimming 108–09, 254–55

T

tackle, fishing 61, 208–11
taiga 36, 37
tampons 60, 117, 118
tandem canoeing 105
tannic acid 268, 269
tanning hides 223
tarp tents 54
tarpaulins *see* ponchos and tarpaulins
taut line hitch 145
tea tree oil 268
teams *see* groups
tears (lacerations) 263
teas 207, 268, 269
teepee fires 124
teeth and gum care 19, 116
temperate environments 28–29
  edible plants 280–81
  forest shelters 162–65
  water sources 28, 188–91
  weather 78–83
temperature 235
  see also specific environments (e.g. cold environments)
temperature ratings, sleeping bags 53
tent pegs 159
tents 41, 54–55
termite fishing 219
tetanus 262
thatching huts 170–71
thirst 20, 235
thorn fishing hooks 208
throwing stars 218
thumbnail test, softwoods and hardwoods 131
tick bites 116, 267, 302–03
tides 157
tie-off knots 142, 143
time, telling by the sun 76
tinder 60, 118, 119, 120, 122–23, 124, 125, 133, 238–39
tins (cans)
  fire cans 123
  matchless fireset cans 118
  survival tins 40, 43, 60–61
tinsel trees 240
tools
  cutting 60, 146–51
  multi-tools 61
topographic maps 66–7
tornadoes 248–49
tracks 28, 30
trail markers (blazes) 87
training 23, 134–35
  benefits 16–17, 20, 235
  physical fitness 19
transpiration bags 190
transport 84–85
  animals 42, 96–97, 98–99
  casualties 99, 278–79
  on foot *see* walking and hiking skills
  motorized 35, 42, 97, 100–03
  water 104–07
traps
  birds 226–27
  dew 190, 196
  fishing 211
  mammals 216–17
trees
  deadfalls 115
  felling 148–51
  hammocks 54, 166
  lightning strikes 81
  navigation aids 77
  needle tea 207
  snowshoes 95
  standing deadwood 115, 121, 131
  tannic acid 268, 269
  tinsel trees 240
  tree shelters 163, 179
  vehicle recovery 103
  water sources 193
trekking umbrellas 175
trenches, hollows and scrapes shelters 156–57, 174–75, 176–77, 178–79
triangulation 71
tropical environments 30–31
  edible plants 284–85
  shelters 54, 166–71
  water sources 192
tubers 206
tundra 36, 37
tunnel traps, walk-in 226
turtles 196, 294–95
twisted cordage 140–41
two-person seat carries 279
two-poncho shelters 159, 167, 175
typhoons 38

U

ultralight camping 41
ultraviolet (UV) light 201
unconsciousness 276–77
Universal Edibility Test 206–07
uphill walking 86, 87
use-anywhere shelters 158–59
UV (ultraviolet) light 201

V

vampire bats 157, 300–01
variation, magnetic 69
vegetation *see* plants and vegetation
vehicles, motorized 35, 42, 97, 100–03
VHF radios 237
vines, water 192
volcanoes 247
vomiting and diarrhea 187, 203, 269

W

wadis 34, 35, 193
walk-in tunnel traps 226
walking and hiking skills 85, 86–87
  river crossings 88–89, 108–09
  scrambling and climbing 90–91
  snow 94–97
  see also navigation
walking staffs 72, 85, 89
warm-weather clothing 47
water
  amount needed 184, 186
  carrying and storing 43, 198–99, 233
  dehydration and rehydration 187, 190, 193, 196, 253, 269, 272
  freezing prevention 194, 195
  hyponatraemia 187
  importance for life 27, 184–85, 186
  plant and vegetation sources 189, 190, 192, 193
  rationing 187, 197
  sources 28, 32, 34, 35, 39, 115, 188–97, 251
  thirst 20, 235
  treating and purifying 60, 118, 185, 200–01
  water-borne diseases 185, 187
water transport 104–07
waterproof fabric, breathable 48
weapons, hunting 218, 226
weather and weather systems 38, 78–83
  extreme weather 246, 248–49
  see also specific environments (e.g. temperate environment)
weight, gear and equipment 41, 42
wells 189
  gypsy wells 191
wet-weather clothing 48
whistles 43, 240
white fuel stoves 59
wicking action, clothing 46, 47, 48
wickiups 171
wilderness survival 50–51
wildfires 34
will to survive 20, 259
winching, vehicle recovery 103
windmill, light 240
winds 38, 79, 82–83
witch hazel 268
witchetty grubs 224, 298–99
witch's broom 238, 239
withies 139
wood
  backpack frame 45
  brushwood rafts 106–07
  firewood 114, 148–49
  kindling and fuel 114, 121, 148–49
  mayasticks 123
  roofs for hollows 156–57
  sledges 96–97
  snapping 151
  snowshoes 95
  standing deadwood 115, 121, 131
  tent pegs 159
  testing softwoods and hardwoods 131
wood ash
  for soap 117
  for stomach pain 269
woodland, forest shelters 162–65
worms
  earthworms 224, 298–99
  first aid 269
  leeches 267
  roundworms 222, 269
  water-borne 187, 269
wounds and bleeding 262–65, 266

Z

zig-zag routes 73

ABOUT THE AUTHOR

After joining the Royal Navy in 1977, Colin Towell qualified as a Combat Survival Instructor with 22 Regiment SAS (the British army's elite Special Air Service), and has spent over 30 years teaching land, sea, desert, jungle, and cold-weather survival skills—as well as survival and conduct in captivity—to UK Army, Navy, Royal Marines, and Air Force personnel. He was the Royal Navy's Chief Survival Instructor and also served three years as Chief Instructor at the US Navy SERE (Survival, Evasion, Resistance, and Extraction) school. Colin saw service in the Falklands, Bosnia, Germany, USA, and Northern Ireland, and still serves as a Reserve Chief Instructor with the UK Defence SERE Training Organisation, training both instructors and students. He provided the survival training, equipment and rescue coordination for Sir Richard Branson's balloon global circumnavigation attempts, and also trials, evaluates, and instructs in the use of specialized survival equipment, both in the UK and abroad.

www.colintowell-survival.com

ACKNOWLEDGEMENTS

FROM THE AUTHOR

I have undertaken as a student every survival course I have ever subsequently taught as an instructor, mainly because I have had to, but also as a reminder of what my students experience, both physically and mentally, when they undergo survival training. In writing this book, it is safe to say that I have again been the student, and the instructors have been the team at Dorling Kindersley, and I would like to take this opportunity to thank them for their dedication, professionalism, vision, and sense of humor. Thank you to Stephanie, Lee, Nicky, Richard, and Michael for heading up the team; and to Phil and Pip, Clare and Katie, Gill and Sarah, Chris and Tracy, Jill, Bob, Brian, Sharon, Chris, Gareth, and Conor, for learning how to translate my Cumbrian into English and for creating the wonderful design for this book; thanks also to the team of illustrators who possess an unnerving ability to decipher my drawings and turn them into works of art!

My warm thanks also go to the following people for generously sharing their expertise: Flt Lt John Hudson, Royal Air Force, Defence SERE Training Centre; Lt Carlton Oliver, Royal Navy, Defence SERE Training Centre; Chief Petty Officer (SE) Keith Spiller; Petty Officer (METOC) Kirk Davidson; Petty Officer (SE) Stephen Paris-Hunter; Tony Borkowski (Royal School Military Survey); Mike Dymond M.R.I.N N.DipM (RYA); Robert (Baldy Bob) Whale; Charlie Tyrrell; Colin Knox (Pre-Mac); Murray Bryars; and Paul Baker (Bushman.UK Knifes).

Finally, I would like to thank the Defence SERE Organisation and the Survival Equipment Branch of the Royal Navy for giving me an amazing career that, if the truth is known, I would have done for free, and for allowing me the opportunity to learn from the very best, both at home and around the world.

FROM THE PUBLISHER

In addition to those mentioned above, Dorling Kindersley would like to thank the following people for their help in making this book happen. At the Royal Navy: Cdr. Alan George RN, Lt. Cdr. Adrian Mundin RN, Lt. Cdr. Paula Rowe RN, and Lt. Cdr. Phil Rosindale RN. At IMG: Olivia Smales and Simon Gresswell.

Thanks also to: Gareth Jones, Chris Stone, and Conor Kilgallon for editorial assistance; Jillian Burr and Tracy Smith for design assistance; and Sue Bosanko for compiling the index.

VENTURING®·BSA

Thank you for your purchase of *The Survival Handbook: Essential Skills for Outdoor Adventure*, an officially licensed product of the Boy Scouts of America. For more than ten years, the Boy Scouts of America Venturing program has been at the forefront of educating and preparing young men and women ages 14 through 20 to enjoy the outdoors in a responsible and safe manner.

In fact, every Venturer subscribes to a standard philosophy regarding the outdoors, the Outdoor Code.

OUTDOOR CODE

As an American, I will do my best to . . .

- **Be clean in my outdoor manners.** I will treat the outdoors as a heritage. I will take care of it for myself and others. I will keep my trash and garbage out of lakes, streams, fields, woods, and roadways.
- **Be careful with fire.** I will build my fires only where they are appropriate. When I have finished using a fire, I will make sure it is cold out. I will leave a clean fire ring, or remove all evidence of my fire.
- **Be considerate in the outdoors.** I will treat public and private property with respect. I will use low-impact methods of hiking and camping.
- **Be conservation-minded.** I will learn how to practice good conservation of soil, water, forests, minerals, grassland, wildlife, and energy. I will urge others to do the same.

This code and those who follow it represent the highest ideals regarding the outdoors and the conduct necessary to not only survive, but to enjoy its full splendor.

For more information on Venturing and the exciting leadership-building programs offered, please see

www.scouting.org/venturing

BOY SCOUTS OF AMERICA

Boy Scouts of America®, the Universal Emblem, and Venturing® are trademarks or registered trademarks of the Boy Scouts of America in the United States and other countries. Printed under license from the Boy Scouts of America.

Navigation Down Under

For
Northern-Hemisphere Tourists
Walkers and Scouts
Boating and Outdoor Orientation

1001 Ways to Extend

Your Navigational Know-How

2010

CONTENTS

Section page
1 BIG-PICTURE TOP-DOWN————————3
Global Perspective
Sun
Stars
Earth
Moon
Wind
Clockwise Navigation
Anticlockwise Systems
Revision, Overview, Explanations
2 THE INSIDE STORY—Mental Manoeuvres—34
Learn
Think Interpret
Look
Integrate
Orientation
Disorientation
Reorientation
Lost?
Miscellaneous: Angle; Night-time
3 REALLY INTERESTING STUFF
—Bush Navigation————————————56
Nature's Norths
Memory & Naming Skills
Lie of the Land
Tracks and Bends
4 DISTANCE————————————————90
Progress
Pace
Range
MAP and/or COMPASS
Map Use
Magnetic Compass Use
Mapmaking
5 Navigational Mnemonic————————110
6 Appendices———————————————116
Tide and Moon Dial
Star Charts; Zodiac
Morse Code & Rhythm Mnemonics
Semaphore Code & Mnemonics
Symbols for Track-Notes & Mud-maps
A List of Star Coordinates
Great Circles; Calendars in the Sky
Tropical Countries
7 Contents/Index—————————————128

Big-Picture Rule One: Locate a Distant Reference Direction First

Until you Orient yourself to the bigger picture, by finding an unchanging direction to latch on to,
- Small-scale pictures and large-scale mental layouts cannot come together. •And the local features won't fit into the bigger perspective.
- Your mental photographs will float around in chaos, like so many loose jigsaw pieces, or letters in an alphabet soup.

In choosing a Reference Direction, your first choice should be...
- ASTRONOMICAL — Sun, Moon, stars, planets, satellites—e.g. Locate the direction of the North Pole Star. There is an astronomical overview p32f

Then •GLOBAL — Compass or Botanical or Geomorphological North e.g. find Canada's North Magnetic Pole.
Next •CONTINENTAL — Weather Systems, radio transmitters, powerline networks, highways, sea-coast—e.g. notice the Sydney to Hobart flight path.
Then •DISTANT landforms — Lie, Lay, Lines, Hills; Lights, etc.—e.g. the direction of the main river valley draining to the sea.
Last of all comes: •What you can see in front of your nose! I.e. don't let your mind latch on to some unreliable or meaningless reference like how the car is parked, or where the wind is, or the initial direction of the track/the drive in/the main street, or uphill, or the way you got out of bed…
- If you can't see a thing beyond a few metres, and have no compass, you will need to "keep a straight line" (see p. 87b) by lining up markers.

Now that you have a reference direction, you can implement the other "Rule One"s...
- Match what you see in front of your nose to the reference;
- Predict your overall trend before you start; •Notice your start-off direction as you start.

Visualisation is your Key Skill

This is most especially relevant to orienting your nose, in your mind's eye, to the Bigger or Invisible Pictures, but is needed at all scales.

Imagine	where the Southern Cross is during the day;	or	where the cars are, now; how you got here.
Visualise	the Earth's invisible spin-axis	and	the weather map you saw, yesterday
Mentally picture	the Galactic spiral! — which clock-sense is it?	or	the turnoff you need — will it be uphill?
Remember	the tropical countries under the Sun	and	the way back
Describe	the path of a shadow-tip in mid-Winter	or	what makes this place recognisable
Picture	sunrise in Antarctica tomorrow morning	or	old blazes, in your mind, so you don't miss any
Interpret numbers, labels and words into reality	e.g. what the shape of the night-sky has to do with your global position		
Express visually	what each clue means	e.g.	what "high noon" means to you right now
Mentally photograph	the map as you study it beforehand	and	how the climb opened out onto the plateau
Mentally match	the invisible landscape trends to the reference direction and to what you see in front of your nose		

At night, and in the rain, navigation gets to be 'all in the mind'.

By recreating the reality you can't see, you can navigate what is in front of your toes, by intelligent imagination.

Big-Picture Orientation is a Mental Layout Acquired Beforehand

Because the invisible picture is so large, you can do your global and continental orientation at home before you start.
Travelling all day to go walking somewhere else, will hardly alter it.

You use: a globe; an atlas; a roadmap of the continent — to learn the lay beforehand. These are layout maps... The relative positions shown are fixed.
But layout views — as seen from above — will hardly ever be seen in your reality, first-hand.
(In contrast, when you come to putting yourself in the middle of the big-picture, you see the landscape first-hand, and you will need "Nose Navigation" — as seen from the ground — which is all about the apparent relative directions.)
The global constants won't change as you walk, e.g. the behaviour of the Sun and stars, because everything is so distant.

You need extra care to visualise the Big Picture, and its meaning, because it is (a) Round (global) (b) Layout (i.e. mental only) (c) Invisible (d) Big!
Do this before you attempt to nose your way through the local scenery.

Smaller-scale clues can then 'fit-in', sensibly, to the larger context.

Compass Navigation Robs You of a Global Perspective

Flat maps must distort the true picture. Only a globe is true. Mercator-projection maps (diagram 1) distort land areas and longitude lines, in order to "preserve the straightness" (!) of a compass course, which never was straight!
If you glue your nose to a compass needle, to follow a compass-course,
the course which looks straight on a flat Mercator map will actually take you on a spiral,
into either the South or North Pole (as when you keep any landmark at a diagonal).
The sense of the spiral (turning to your right/left) changes as you cross the Equator...
Because you do need a big-picture approach, **you need a global perspective**.
(Here's a good place to set out the basics...
Remember: "Latitude **equates** to the distance from the Equator" — the short dimension.
"Longitude turns upon the time — how **long** has the globe turned since midnight") < long for longitude >

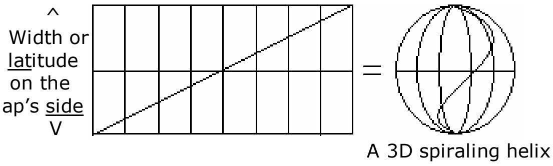

Here's a test: Imagine the midsummer sunset from well Down Under, looking even further South. Do you think of a tropical country in that direction?
The following section — about Sun Moon Earth and Stars — aims to restore a true perspective to people who can hardly think except with a flat diagram, and can hardly visualise anything without divorcing it from the real landscape in front of their nose. (There is always a tropical country under the Sun.)
Think of this 'global orientation' section as a global **Re**-Orientation and don't forget, that... **Global means Round!**

The Big Picture is Invisible — Your Job is to Visualise It!

As a matter of principle, we should work top-down
i.e. from a framework down to the detail, from general to specific, large to small, overall to local, unchanging to changeable.
As a matter of fact, we are lazy. The point of doing "useless" grand-scale orientation exercises is to:
make global orientation second nature — easier next time around;
to make you forget flat pictures of a round Earth and to imagine it properly;
to give your struggling memory some comprehensive and relable framework to work within;
to integrate what you can't see with what you can, e.g. to link different walks and areas;
to orient to, and to understand, the Sun's movement, and to orient to the Moon and stars and planets as a bonus;
to make extra clues and techniques available, e.g. to open up astro-navigation techniques;
to work "top-down" — to avoid jumbling up your mind.

The idea is to read the following double-pages horizontally, as one topic, i.e. Sun-Sun; Stars-Stars; Clockwise-Anticlockwise.
To follow one topic through, flip over the pages and look in the corresponding positions.

Global Sun Stars Moon Clockwise

Global Re-orientation — Light-Half/Dark-Half

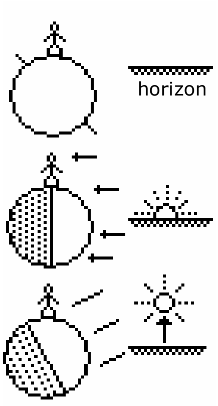

Step 1. **Always Position Yourself on Top of the World.** ..whether in a diagram or mentally or in reality.
This makes your reorientation realistic, doesn't it?
Look at your feet and imagine you are standing on a huge basketball in the sunlight.
To visualise the light-half/dark-half of the globe, carry a real ball with you, preferably a globe of the world.
Imagine that everything which you see, this side of the horizon, is the tiny top-fraction of a huge ball.

Step 2. Face the Sun, but imagine it low down, setting, with horizontal rays coming to you.
Notice the circle which divides day from night, passing to left and to right underneath your feet.
"He divided light from darkness by a circle drawn on the face of the sea." Job 26:10

Step 3. 'Raise' the Sun, and its sunbeams, to their actual height.
And as you do, let the dark-half fall back behind you by the same amount
i.e. degree for degree around a round globe—one degree of fall-back for one degree of Sun-rising.

After this you can do your best to answer: (Help will follow soon) What countries are seeing the Sun as horizontal?
i.e. Which countries does the 'circle' pass through at the moment? Which ones are seeing the Sun rise?
Which ones see it setting? Which parts see a 24-hour sunset, with the Sun skimming the horizon? (Use a globe)

Your Best Reference Direction: Imagine the Sun at High Noon — Always

(This won't work so well in the tropics)

Night or day, your mental reference direction is North or South by the Sun. So imagine the Sun North of you (Southern Hemisphere) rain or shine.
It is the natural interpretation… of looking at your watch, and
of looking at the Sun, when it is not North of you. (Warning: Don't look directly at the Sun! It will blind you!)
of looking at the compass. (You don't want to locate 'Canada's magnetic pole', after all!)
of judging botanical North/South (p.74a);
of using a map in conjunction with the landscape,
and even when using the Moon

Imagine the *high-noon* Sun in a real landscape — this one here!
— at all times of day or night.
— at the correct height, on its circular path,
— which is tilted at its rise-set angle.
Warning:
You will get disoriented
when you cross the Equator!

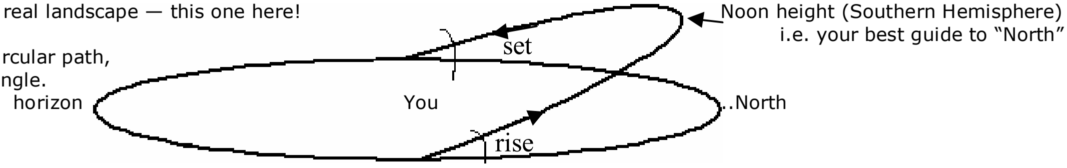
Noon height (Southern Hemisphere)
i.e. your best guide to "North"
set
horizon
You
..North
rise

The 4 Bright Outer Stars of Orion Point out Where the Pot is Going

=Find N/S E/W & Equator

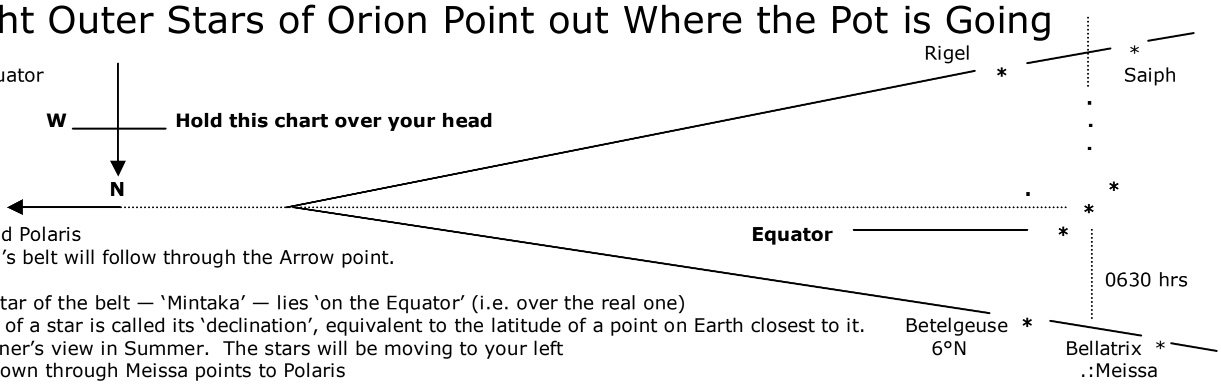

Hold this chart over your head

Movement West
Anticlockwise around Polaris
The middle of Orion's belt will follow through the Arrow point.

The north-west star of the belt — 'Mintaka' — lies 'on the Equator' (i.e. over the real one)
The N-S position of a star is called its 'declination', equivalent to the latitude of a point on Earth closest to it.
This is a Southerner's view in Summer. The stars will be moving to your left
The Pot handle down through Meissa points to Polaris

Look for the Line Joining Sun, Moon, Planets & Zodiac Stars—the Ecliptic

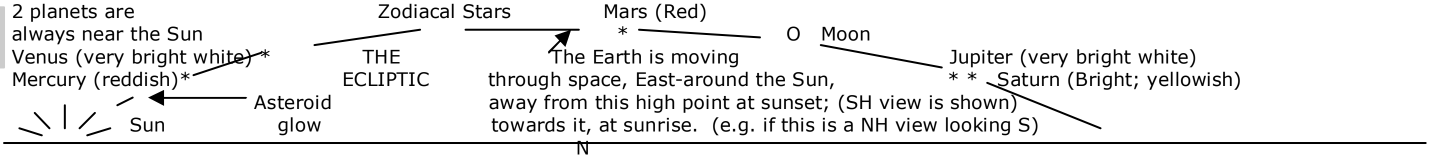

The Earth is moving
through space, East-around the Sun,
away from this high point at sunset; (SH view is shown)
towards it, at sunrise. (e.g. if this is a NH view looking S)

The solar system — Sun and planets — is flat, and the Zodiac marks that plane, by design. It's **the 'edge-on' view, which we see, from within**.
The Moon orbits within 5° of that plane, and so is never far off line. **The line is called the Ecliptic** — a great circle.
(It looks as straight as the horizon, unlike the diagram. See 'Optical illusion' in the Moon box, p.8d)
During the day, you can orient, not by the Sun only or the Moon only, but by the line joining them. Sometimes Venus may be visible too.
You have to imagine the rotation of that line (more on this soon). It does not move in the same direction in which it points, but its path is up to 23°½ tilted from that.

Mnemonic: "Clock-wise Northern Hemisphere **SWiNES**"

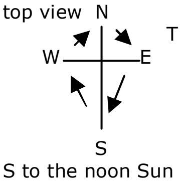

S to the noon Sun

Europeans make chronometers to follow their Sun.
They see, looking South (diagram top right)… rise in the East ; South at noon ; set in the West.
That's 'clockwise' (seen from above; East to South to West) for them
The swines! They force us, down under, to place our clocks on the southern wall and think behind us!
(We see, looking North: rise East, to North at noon, to West i.e. SH; anti-clockwise)……

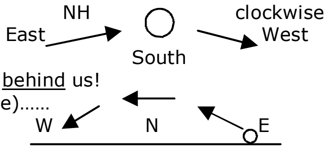

clockwise
top view

Anyway, changing your direction "Clockwise", as seen from above, is turning to your right, and, **i**gnoring the **i**, the mnemonic is:
South, to **W**est, to **N**orth, to **E**ast, to **S**outh or "clockwise: **SW**i**NES**"

A mnemonic for global orientation: **“True Polarity”**

“**True**” courses refer to the Earth’s **Poles** — i.e. the axis of rotation of the Earth — **the single most directional thing on our spherical Earth**.
“True” bearings measure the angle between the course and the longitude line — a North/South line which connects the poles,
and which passes through your location.

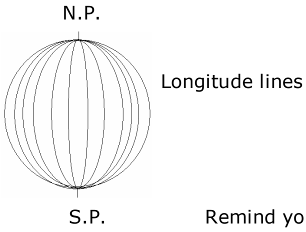

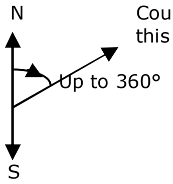

The angle is measured **clockwise from ‘North’**
e.g. 065°T
“Oh Six Five degrees True (-polarity)”!

Remind yourself, whenever you use “True”, by saying “True Polarity”!

Earth

Mnemonic: “**N**oon Sun Passes **N**orth of My **N**ose Down Under”— It Rhymes in ‘ **n** ’ (dow-nun-der)

(Now, Keeping your **n**ose to **N**orth, say…) (The second parts work for both hemispheres provided you face North, not noon)
“Sun **Ri**ses on My **Ri**ght (to the right of North)” —It Rhymes in ‘ **ri** ’
“Sun S**e**ts on My L**e**ft (in the W**e**st)” —It Rhymes in ‘ **e** ’
“That’s Anticlockwise around the North Pole Star, if you are facing North from the SH; [Clockwise around the South Pole Star”, if facing S from NH]
“The **SH** ho**ri**zon **shi**fts to the **ri**ght”
“The **NH** ho**ri**zon **N**ever **H**eads **ri**ght, it moves left”

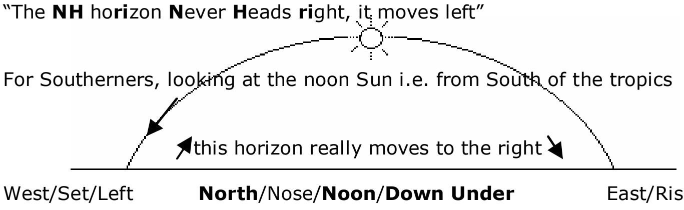
For Southerners, looking at the noon Sun i.e. from South of the tropics
this horizon really moves to the right
West/Set/Left
North/Nose/**Noon**/**Down Under**
East/Rise/Right

Paper Diagrams aren’t Enough.

Imagination isn’t enough.

Go outside and take a look.
(You will need to shield your eyes from the Sun, e.g. with your hand)

Let the reality sink in.

Sun

The Two Brightest Stars Point out the **South** Pole Star

Canopus and Sirius line up almost South/North.
Use the star next to Sirius (‘Mirzam’) and the line-up is perfect.
The ‘Dog Star’ Sirius is the brightest star in the sky.
The Big Dog seems to have a front leg and back leg and a tail...

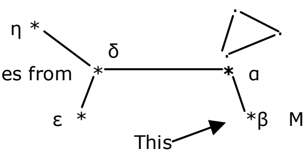

Al**u**dra, in the tai**l**
Wezen, where he wees from
Ad**h**ara, the **h**ind leg
Mirzam
This
is the nearest
bright star to Sirius
The Big Dog

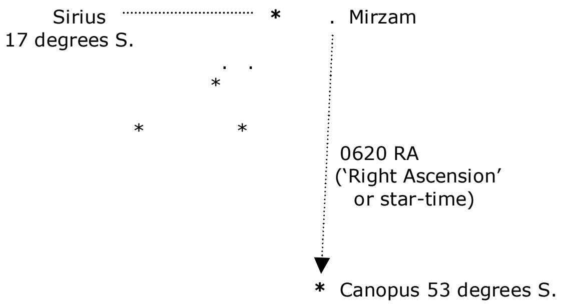

(Go another 37 degrees to South polestar)

A
GOOD
SOUTHERN
SUMMER
SIGN

Stars

Mnemonic: “Wind? Weather Map!”

For navigating by the wind, don’t just leave it at the level of “The wind is from over there”.

* Look at the skyline “from Bruce Valley”
* Look at the Sunshine “from 30° to the left of the Sun at the moment” (Don’t look at the Sun!)
* Look at the Sun-path “a ‘sunset’ breeze” i.e. from sunset (Simply imagine its rise-set path.)
* Imagine the continent beyond the skyline “from the West Coast”
* Meld that reality to the weather map you remembered.

Don’t stop interpreting the wind until you can fit it into a Big Invisible Picture connected to that skyline over there.

Wind

Mnemonic: Down Under, The Sun Moves **ANTI**-BigB**ENW**i**SE** **i.e. anti-clockwise**

Look at the ‘ENWiSE’ in the mnemonic.
Anticlockwise is **E**ast to **N**orth to **W**est to **S**outh to **E**ast.
Ignore the **i**. Look down from above onto the horizontal directions involved…
Anticlockwise is turning to your left. “ Anti-clockwise is ENWISE ”

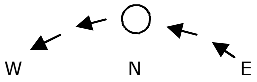

(If you put a clockface on the north wall and look at the Sun’s movement, (Don’t look at the Sun!),
you will see that the Southern Hemisphere’s second-hands tick backwards! i.e. against the Sun’s movement.
So put it on the south wall facing North, and the clock hands will move ‘with’ the Sun.)

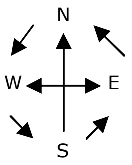

top view

In the tropics, the Sun will pass somewhere near overhead. To make it seem “anti-clockwise” anyway, face North.

Don’t get mixed up and think of Big Ben wise as clockwise. Down here our Sun moves anti- , and that is ENWiSE.

Anticlockwise

Mnemonic: "Time Tells Turning"

What's the use of looking at your watch and spouting numbers! At least, glance at the Sun, or rather at its shadows (don't look directly at the Sun) and then, imagine the Sun's direction at high noon as a reference direction.

That's what time tells you! — the Sun's position relative to its noon direction, North/South, called the "meridian".
A.m. says "ante meridiem" i.e. "before noon", and p.m. says "post meri-diem" i.e. "after mid-day"
Midday = 12 noon (but only in an ideal world, because noon is only approximately at 12)
One pm = one hour past 12md, and so on.
24-hour time tells you how far around the Earth has turned, since the Sun was opposite you i.e. since midnight.
(And then don't forget that the local significance of the time is: progress — distance made good — "watch" your pace)
Mnemonic:

"Watch" the Sun (without looking at it) whenever you watch your watch
(and "watch" your steps too, for progress)

Notice these, Yesterday — to Guess the Time Today

- Sunrise time (and direction) ..If you know the time, you can guess at the directions.
- Sunset time (and direction) ..These first 2 are symmetric around North.
- The horizontal angle of separation between sunrise and set
- Midday time (and direction, and height) ..This is halfway between. Memorise how short your shadow is.
- When did it get light enough? To see. To walk.
- When did it get too dark? To cook. To walk.
- How long is the usable daylight? ..The day-length, between dawn and dusk.
- How many hours of dark?
- When did the Sun reach due East and due West? (How high?) ..This won't happen in winter. They are symmetric around midday.
- Times for the Sun at magnetic East, noon, West? ..These don't depend on guessing or calculating the magnetic variation.

Update yourself every week.

The Spinning Earth Makes **Circular Paths** for the Sun, Moon, Etc.

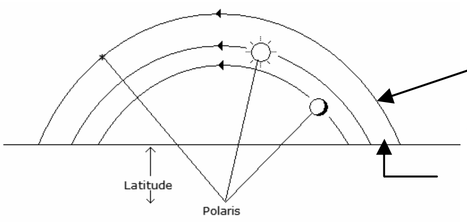

Sun, Moon, stars, planets — they all move only slightly in relation to each other during 1 day/night — but they all zoom around at about 15° per hour from East to West.

Although I can draw this on flat paper as a circular path, the Equatorial path actually looks as straight as the horizon, in reality! (Southern Hemisphere view)

You can now guess at time or direction by your familiarity with the circular 'orbits' e.g...
The highest point of the path will be North-South of you, if you can guesstimate it
Latitude determines the rise/set angle of the Equatorial path.
Rising and setting places are symmetrical around North.

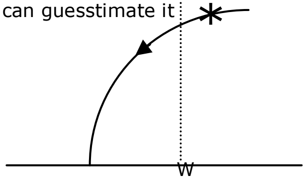

Although individual rising and setting is symmetrical about North, whole constellations that rise vertically in mid-latitude may set horizontally! and vice versa e.g. The Big Dog or Scorpio rise flat but dive vertically into the western horizon.

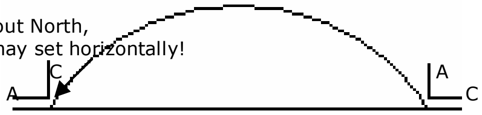

The Moon; Rule 4: Look for It

You have a 1 in 4 chance of finding it during the day. Provided you remember to look!

Rules 1 and 2 (like so many navigational clues) are...

- PREDICT when and where it should be visible
- GUESS where to look And Rule 3 goes with the guess:
- LEARN from that mistake.

Keep looking for it. It's easy to miss. E.g. through the tree canopy; or rising through the horizon haze. Polaroid glasses help; you need to twist them

You will then have clues to…
North, East, and West
Time of night (i.e. later)
Ocean tides
Tomorrow's Moon — Phase; rise, set, direction, height
Upper-level winds, if there are clouds about.

Hint: Measure the Sun-Moon separation with a piece of string (Don't look directly or indirectly at the Sun!) so you know exactly where to look later.

"VEERING" with the Sun is "Clockwise" only in the Northern Hemisphere "BACKING", against the Sun's movement, is "Anticlockwise" — but only in the Northern Hemisphere.

(The terms refer to horizontal directions changing either with or against the Sun's movement)
"Veering and Backing" are Eurocentric images, so they can be very confusing Down Under. Don't use those words in the Southern Hemisphere!

Say "Clockwise" and "Anticlockwise" instead
..for both hemispheres, as a matter of habit. Although these meanings too are Eurocentric, they are not confusing!
(the terms refer to a bird's-eye view looking down from above as if onto a clock-face)
Meteorologists and mariners still use the terms 'veering' and 'backing' and they do mean 'clockwise' and 'anticlockwise'
But the Sun-based image is backwards, south of the tropics.

The sky turns **15° per hour** — Use your **thumb and forefinger**

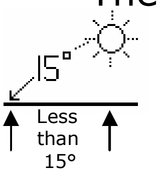

The Earth rotates once a day. Yes? That's 360° in 24 hours, yes?
Well that works out at 15° per hour — the sky spins 15°/hour to the West.
That calls for **recognising 15°** E.g. (diagrams left, and right) "one hour of sunlight left" (SH)

For Accuracy:
(0) Measure each handspan along the Sun's 'circular' path, not horizontally.
(1) Use constant stretch — of arms and fingers — 'just strained' — for consistency.
(2) Use constant chest-arm geometry. It is best to use two hands squarely in front of you…
(3)**Test for 15°, horizontally,** against the full 360° horizon — **"24 spans go around the rim"**
Span off, 24 times, from point-to-point right around the horizon.
Adjust your hand configuration until you find one that fits the "24 Hours" recipe.
(4) Try gripping one hand with the other, or touching one, at a specified point, with the other
, or, (5) Look along your shoulder, sideways, at one hand.

My recipe for accuracy fits my arms:"4 spread fingers, wrist bent back fully (at arm's length in front=60cm or 1cm/degree")
What do your arms require for 24 spans = 360°?

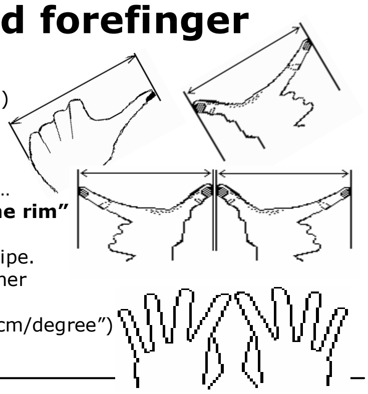

Mnemonic: The Southern Summer Sun Sets Somewhat South

Surprising perhaps. (If you equate North with 'hot' you get caught on this.) (for Northern Hemisphere substitute 'North' for 'South')
When the Sun comes South for summer its sunrise and sunset places also shift South with it.
Memory hexagon for Mid-Winter and Mid-Summer Down-Under at 37° S…
(top view) N. (the Sun moves anti-clockwise through North)

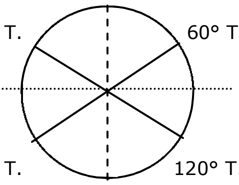

… In Mid-Winter, Down-Under, at 37°S., sunrise is at 60°T. and sunset is at 300°T. … A 9½ hour day.
In Winter the Sun will never show East or West of you.

At Equinoxes — Mid-Autumn, and Mid-Spring — rise and set are East and West… a 12 hour day.

… In Mid-Summer, Down-Under, at 37°S., sunrise is at 120°T., and sunset is at 240°T. … A 14½ hour day.
The Sun will be East and West at **mid**-morning and **mid**-afternoon respectively, and **mid**-high — about 45°.
Your main aim is to get familiar with the Sun's circular path.

Stars Leave Star-Trails ..if you photograph them with time-exposure...

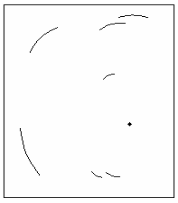

Star-trails are bits of circles, because… ..stars circle around the North and South Pole-Stars...
i.e. the Earth rotates on an N-S axis, which points to the Pole-stars. The pole star stays fixed in position all night (if you do).
Southern Stars Circle Clockwise, once per night-and-day, around the South Pole Star.
Northern Stars Circle Anticlockwise, around the North pole star, Polaris. These clock senses are the same in both hemispheres.
But **Note**: low Southern stars will move anti-clockwise around you, (i.e. to your left around the horizon, anti-clockwise as seen from above you). (Clockwise — to your right in the NH) (More on this later)

It helps you with global orientation to use 'star compasses'.

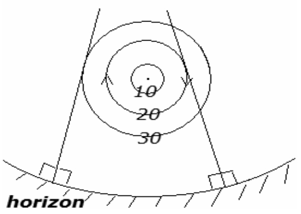

One of the clearest compasses is to blur the star circles into full circles then take vertical tangents to those circles.
In the following diagram, the left-hand direction is "30°" to the left of the Earth's axis (or to 60° maximum latitude).
The right-hand one is "20°" to the right (or to 80° maximum latitude). They are global.
If you follow these star directions, they will take you right around the earth in a 'straight' line — a big circle.
You will pass through places equal in latitude to the star which is overhead (p.11c), so you can already *see* the maximum latitude that such a 'great circle course' would take you to; then it takes you back towards the equator.
HINT: You won't get far in star-navigation, without a detailed star list and star chart. See p. 123 for a list.

Mnemonic: "Wind in Your Face"

The Southern Hemisphere version of Buys Ballot's Law… [Try to use unobstructed wind i.e. not in a valley nor in a mountain eddy]

- "**Face** the Wind [N.H. rule uses 'Back to the wind, Sun-twist…]
- Sun-twist yourself 15°-30° [i.e. anti-clockwise in S. H.]
  — to counteract landscape friction on the surface winds.
  [When the winds are impeded, they slow down, and so stop circling around the pressure centres, and start to cross the isobars, toward the low pressure (a 'sink' hole) at an angle of 15°-30°, less at sea and over flat land; more over rough terrain.]
- Rhyme: **"Low on My Left** &/or **High on My Righ-t"**

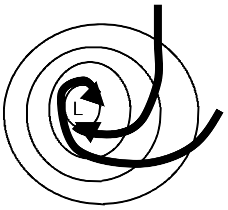

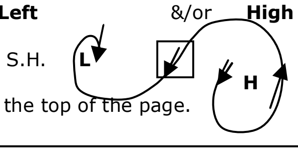

Top view of you, in the square, facing the top of the page.

Southerners Need **Anti**clockwise Systems

Our clocks seem to run backwards, compared to the Sun.
Our compass-bearings are measured 'backwards'. (see p.10e)
'Veering' and 'Backing' seem to mean the wrong clock-sense.
When the Sun moves forwards, all the other systems seem to push backwards.
We lose the global perspective.
And this section is all about aiming to put that right.
Consequently, we need a Southerner's viewpoint…

I hope you find that the anticlockwise systems which will follow soon are much more workable than Eurocentric systems.
Some of the other systems given on this right-hand side of the page, are universal — they can be used in either hemisphere.

Global Orientation: Latitude and Polar Distance

Step 1. Stand up on the globe again (in your mind and in reality).
Step 2. Put your left shoulder to the Equator i.e. Face East (sunrise).
(If you are in the northern hemisphere: face west, sunset.)
Step 3. Mnemonic: EQUATOR on My Left equals My LATITUDE is on My LEFT,
RISEN POLE on My RIGHT equals My POLAR DISTANCE is on My RIGHT.

Backviews...
Facing E.

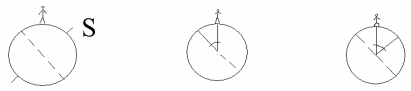

Since the South Pole is 90° latitude, it is further South from the Equator than you are, and will be on your Right, as you face East.
Step 4. Measure the degrees of distance around the Earth's circumference as an angle at the centre of the Earth (by slicing the globe in two).
The two angles mentioned must add up to 90° i.e. they are "complementary" angles (spelt with an 'e').
If you are, say, 45° south of the Equator, then the closest pole must be 45° further south. The 'distance' to the Pole is 45°.
If you are 30° south, then the South Pole will be 60° away from you. Your '**colatitude**'—or *90° minus latitude*—or "Polar Distance"—is 60°.
The goal is to imagine those two angles — Latitude and Polar Distance — at the Globe's centre beneath your feet, and around the circumference.

Mnemonic: "Sun at High Noon — How High?"

When you navigate by the Sun, think: "Sun": the mnemonic is: "Sun at high noon — how high?".
You look: North, there, then up there, that high.— Southern Hemisphere (cf Noon will be South in NH) (Don't look at the Sun, if it is already there!)
It might be morning : "It is going up there"
It might be noon : "See how high it gets!"
It might be afternoon: "It came from up there"
Visualising the height and path of the Sun also helps you to interpret plants for 'bush noon' (p.74a).
"Walking Stick North" depends on knowing the Sun's maximum height. (See page 30b)

On Overcast Days, Average the Brightest Sky over 1/4 hour

If the clouds are moving fast enough, and not entirely uniform, sit still for long enough to see many clouds pass over. Keep pointing at the brightest patch of sky using some reference like a tree, and watch to see whether you can confidently locate the average position of brightness.

RISE and SET Directions for a (**Local**) STAR COMPASS

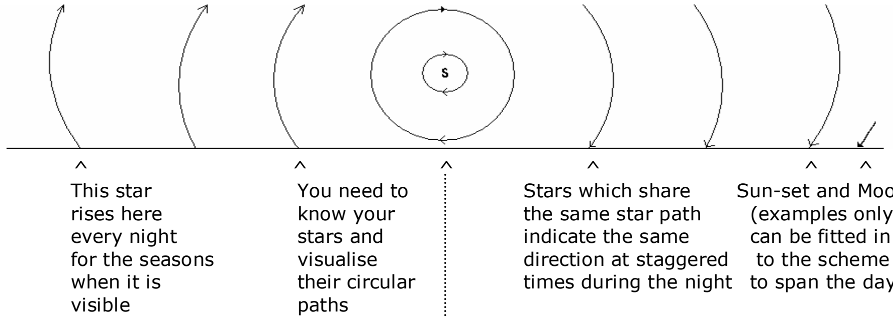

This star rises here every night for the seasons when it is visible

You need to know your stars and visualise their circular paths

Stars which share the same star path indicate the same direction at staggered times during the night

Sun-set and Moon-set (examples only) can be fitted in to the scheme to span the day

The star directions don't change night by night or with longitude, but they do change with latitude.

The circles in the diagram will rise from the horizon as you travel further south, and so change their intersections with the horizon.

Just make sure to mark sunset to a landmark on the skyline horizon and carry it over into night-time.

(SH view)

(Some stars stay up all night every night; these don't move far either, and so are good orientation value too)

From the Moon, Locate the Sun-Moon Direct Line, Carefully.

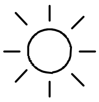

The line is the Approximate Ecliptic, and being a great-circle is a straight line.
OPTICAL ILLUSION!
The Moon seems to be lit from above the 'direct' line to the Sun.
But that may only prove that the Sun is "up" and shining down.

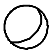

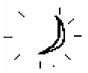

Bust the illusion with a piece of string…
Caution: Don't look directly at the Sun!
horizon

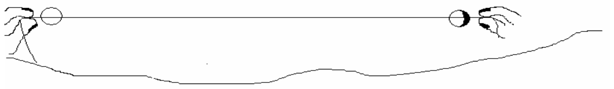

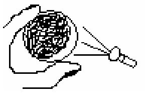

Prove it in 3 dimensions, by using a torch and a ball.
When the ball, in line with the Moon, looks right, the torch direction mimics actual sunlight.
Below horizontal = at night

By locating the Sun, you can guess the time of night, and guesstimate North more reliably, because the path of the Sun is more predictable and familiar than that of the Moon.

Mnemonic: (when you say) "COMPASS" (say) "CANADA"

Compass needles point to Northern Canada. That's where the North Magnetic Pole lives. Lat. 75° N.; long. 100° E.
E.g. from Australia, Canada lies to the East of North…

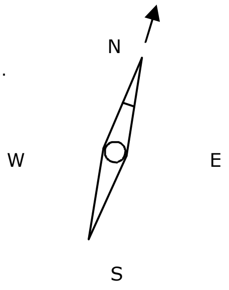

Check it out on a globe. Locate the Queen Elizabeth Islands.

To find the Southern Magnetic Pole, (but this doesn't work so well)
locate the Antarctic coast south of Adelaide, at Lat. 65° S.; long. 140° E.

The Angle of Rising and Setting — is Determined by Your Latitude

To post-dict sunrise or pre-dict sunset direction, you can use your latitude to visualise the tilt at which the Sun **rises** (or the rise-angle of a near-Equatorial Star/Moon/planet) and the deviation from straight down, at which it **sets**.
Remember to use vertical, not horizontal as the expected norm. (To judge vertical correctly, see p19c. To practise, see 30b)

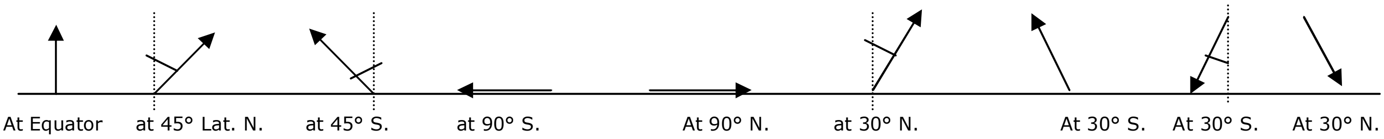

This one is especially useful in Tasmania.

The Sun's path is very nearly a straight line, despite the illusion of being circular—see p.15a.
Hint: Whenever you see the shadow of your head, provided the Sun is not too high, you can make the "anti-Sun" rise or set positions obvious, by following up from that shadow, to the horizon at the correct angle. In midwinter the anti-Sun will be in a midsummer position and vice-versa.

High Noon — How High?

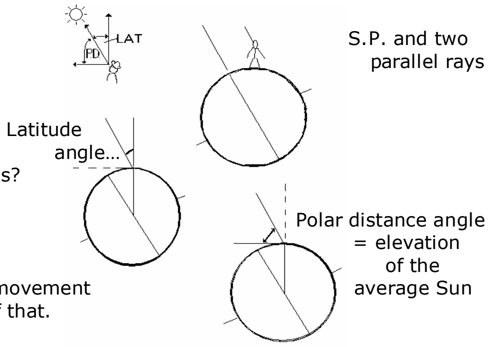
LAT
PD
S.P. and two parallel rays
Latitude angle...
Polar distance angle = elevation of the average Sun

Answer one: the same as last week. Take a look. It changes only slowly.
Answer two: **within 23½ degrees of your co-latitude (or 'polar distance') high**
i.e. 90° – Lat., plus or minus 23½ degrees, depending on the season.
Mid-Summer… co-lat+23.5° (very high in the sky)
Mid-Winter… co-lat–23.5° (very low in the sky)
Equinoxes… co-lat (average height).

Why? At 0° Latitude, i.e. at the Equator, the average Sun is 90° high at noon i.e. overhead. Yes?
At the Poles, 90° Lat., the Sun skims the horizon at equinoxes. Yes?
If you need further proof (that its height is 90° minus Lat.) draw it.

See how the height of the average Sun is 'latitude degrees below vertical'.
That's an angle of elevation of 90° minus lat, isn't it? Now adjust for the Sun's seasonal movement which is 23½ degrees each side of that.

Global Star-Compasses e.g. 'Perpendicular to' rise and set

This is one step harder and three steps better. It allows 'great-circle' course keeping.
A global Star-compass yields 'direct', global, 'straightest path' courses — ones which take you around the globe and back to your starting point.
E.g. You take a rise or set compass-point and head at right angles to it, and the new horizontal direction is global—no longer tied to one latitude.
That's easier to imagine by visualising the light half/dark half of the Earth: All who see the Sun on the horizon could point to it, but each would be pointing in a different "direction". But if they all pointed horizontally at right angles to the Sun, they would all be pointing in truly the same way —around the Great Circle which divides day from night. The same goes for any horizontal star. Imagine the star as a pole star, and use its "equator"
There are many alternative star appearances, besides rise or set & **you don't have to 'go at right angles to' these to set your course...**

E.g. aim at **one star above another**
The midpoint of One star **level with another**
One star **rising or falling vertically**.

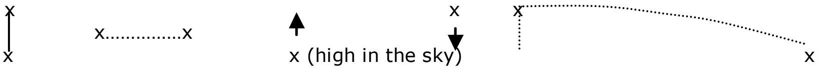

Toward a star which has a 2nd star off to one side **at right angles** i.e. lower. You would pass *under* the point you kept aiming for.
You don't need a clear sea horizon — the stars are high, and a plumb-line will suffice.
These guides occur every night, in the same global direction, for several seasons, then disappear for a while due to daylight, but reappear.
They will occur at the same times next year. **They are always worth learning more of**, adding to your familiarity of favourite 'Star directions'.
These global directions are labelled by saying *how close to the poles* they would take you, and you can see or guess that from the stars, see p. 7c.

'Low' Clouds Come Clockwise Down Under

In the Southern Hemisphere

- Low-**pressure-system winds** spiral **clockwise** around. I.e. into your face from your left as you face their winds...
- 'Low **clouds**' (i.e. below freezing level) follow the surface level isobars i.e. the weather-map lines.

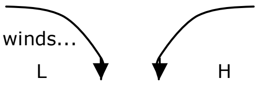

The double-meaning for ' Low ' in the mnemonic reminds you of this.
'Lows' cause more clouds than 'Highs', so you focus on 'Lows'.
I.e. when you see the low clouds; you say 'low clouds come clockwise' and automatically imagine………… '**L**ow On My **L**eft' Plus '**High** On My Ri**ght**'
The Alternative Southern Hemisphere version of Buys Ballot's law. This **new version goes like this** :
Face the low clouds (unaffected by friction—winds at 600m high are considered unimpeded),
then rhyme: L— L ;igh—igh. (Don't Suntwist yourself.)

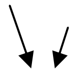

[Cf "Face the wind, (In the northern hemisphere you have to say "back to the wind")
Sun-twist yourself up to about 30° to counteract friction (which makes the surface winds spiral inward to the low centre)]
Surface winds could also be expected to be half the speed of unimpeded winds, less impediment over sea, more over land—about one third.

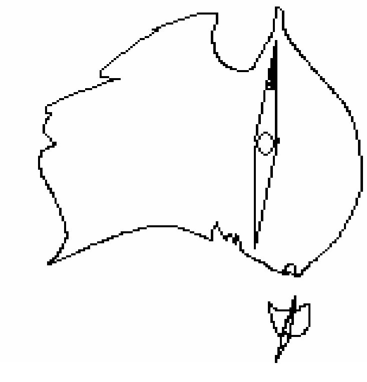

A Compass-Needle Continent—Swinging the Continent

Mentally glue a miniature continent under the compass needle, to swing with it.

The idea is twofold: 1. To reprogram your mind — "The continent is not swinging, so... **I** must be!"
2. To orient yourself meaningfully — to the invisible picture
3. To remind yourself of compass variation

When using a compass you could: just stare at the compass-needle, like a lone night sailor in a fog.
Or: look at what the compass is pointing to — that bump!
Even better: also 'see' the needle lying across the whole continent
Keep going: pointing to Northern Canada
Don't forget: the map you are using — the real one, not this imaginary one — should also line up properly — twist it around until it does.

Global Orientation: Point to the South Pole…
..half its polar distance below horizontal

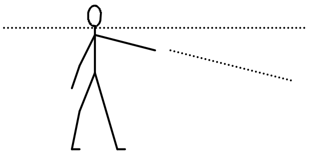

If you had to point "South" you need to point horizontally.
But when you have to point out "the South Pole", you must point below horizontal.
(It is similar when Northerners use the North Pole)
Imagine the true scale… you are on the circumference, not above it...

The question is: "What is the exact angle below horizontal?"
The answer is : "The South Pole is **HALF its polar distance** below South."
Example: at 45° S., you point 22½ degrees below the horizon. (90-45) /2
Example: latitude 30°; colatitude = 90–30 = 60°; half that equals 30°
Example: latitude 60°; polar distance 30°; 15° below horizontal.

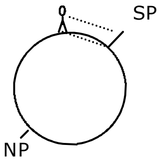

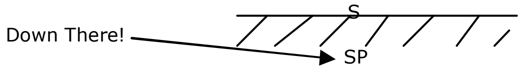

Optical Illusion — Beware! — Use Your Shadow Not the Sun, For Direction

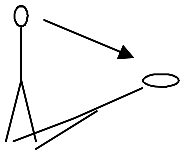

When the Sun is directly behind you, it seems to be behind on the left when you look over your left shoulder, and behind on the right when you look over your right shoulder! And when the Sun is exactly to the side, it nonetheless seems to be ahead of you. (Every sports coach knows that when you turn your head, your body follows it.) Don't rely on where the Sun seems to be… USE A SHADOW INSTEAD! It is also easier on your eyes.

And then don't get fooled by the shadow of your finger, when you throw a shadow onto your map to find the Sun's direction. It may look straight but it doesn't point to the Sun, until you line it up with your finger**tip** while looking from directly above. The shadows will always be parallel, but the Sun's rays will always join object to image, fingertip to fingertip-shadow…
The first diagram is illuminated from behind you……
To avoid looking at the Sun, try this:
Point your walking stick from somewhere beneath your eyes, outwards, and rotate yourself until it is in line over its own shadow.

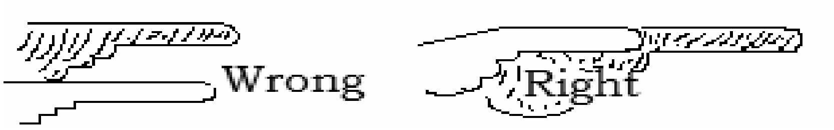

Right Around the Horizon, Stars Move to Your **Left**, Down South

Looking South, all the lower stars move left.

It is only the high Southern stars which move to your right.

Looking East all the stars rise to the left
(i.e. we see the bottom half of the circles
down at horizon level, looking South)

Looking West all the stars set to the left

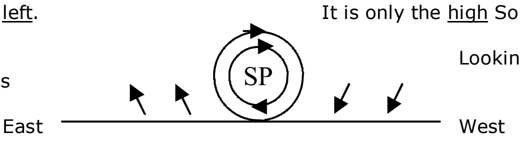

Looking North, all the stars move left.
(we see the top of the circles)

The same goes for Sun, Moon, Planets.

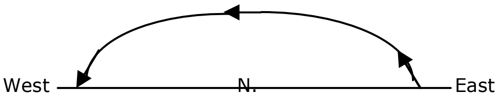

Looking directly overhead, or behind your head, will not work!

(it is all opposite in the northern hemisphere)

Mentally Rotate the Sun-Moon Line Around Polaris

A whole line through the sky gives a better 'feel' for direction than a single point — than either the Sun or the Moon alone —
So join the two and watch them rotate together during the rest of the day.

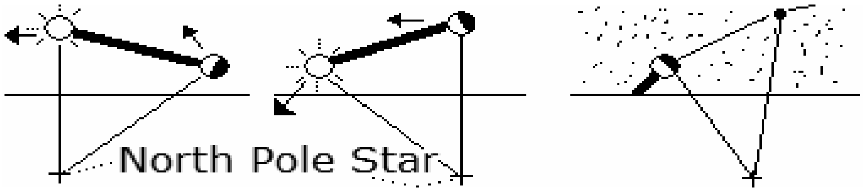

(Southern Hemisphere view)

When you can see both, and you know where North is, you watch them.
When you see only the one, you look for the other, and so find north more exactly than by using either one's path by itself.
Add in any planets to continue the process at night, then from planet to planet or to Zodiac stars, after moonset.

"Bears Circle Clockwise"
Compass Bearings are **C**lockwise, A Full **C**ircle, and Euro**c**entric, **C**ee?

In Europe the Sun moves clockwise… from North (Where the Sun is at midnight—0 hours) = 0° as far as the horizontal angle involved is concerned.

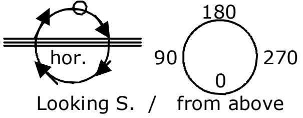

East is at about 06 hours (sunrise) or 90° True — Sun-wise or Clock-wise from North
South at about 12 hours or 180° T (noon)
West at about 18 hours or 270° T (sunset) — so you see, the bearings circle clockwise, from North, 360°.

A **Southerner**, of course, would prefer to go 360° **anti**-clockwise, from **South**, but they didn't get a vote.
Everyone in the world is expected to submit to the (NH) rule: compass bearings should be measured in degrees, up to 360°, from North, clockwise around through East — like the ancient Babylonians; like the Northern temperate Sun.
Not many people consider using 400 grads (4 right-angles of 100 grads make up the circle), or consider going left-around, through West,
Or even think of starting at South. So don't you dare (unless you **don't** call it "a compass bearing").

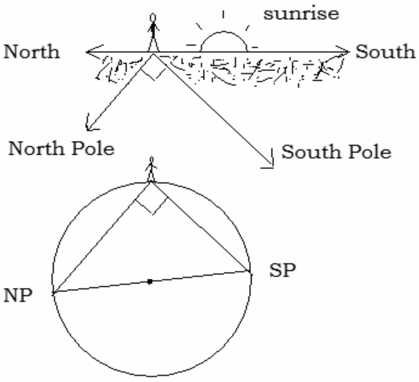

The North Pole is *Perpendicular to the South Pole*

Since the poles are 180° apart — opposite — you see them as 90° apart — half the arc distance.
The numbers can't work out any differently.

Example: 10° S. Lat. = Equator is 10° North of you. South Pole is 80° away — the "Polar Distance".
So S.P. = 40° below S.
North Pole is another 90° past the Equator.
So N.P. = (90 + 10)/2 = 50° below N. And the gap is 90°.

A simple way of looking at it is to say
"Diameters yield right-angles at the circumference"

Now go outside and point to the Poles.

The Dark-Blue Polarisation Band Perpendicular to the Sun

90° from the Sun there is a circle of darker sky. Bees navigate by this clue. The whole sky shows degrees of polarisation.
You can see it naked-eye, but since it is due to polarisation, it is much clearer through Polaroid sunglasses or a camera filter.
Point the top of your polaroid sunglasses toward the Sun for best effect.
The dark blue band is narrowest highest up (in the direction of your shadow, and perpendicular to that sunbeam) and varies with visibility and haze.
The sky is darkest away from the Sun and away from the horizon haze and 90° from the Sun.
You can always guess where the Sun is, from the gradations of blue. After sunset the band is caused by the brightest part of sky.
Example: You are in a forest at dusk. The sky is blue overhead, but the sunset direction is not visible. Find the Sun from the blue!
See the "Sunny Skies System" (p.19e) for how to use the ends of the dark blue band, and its trend.

The Centre of the Rainbow is Down-Sun

The shadow of your head coincides with the centre of the rainbow's circle. (Pilots often see the full circle of a rainbow)
The radius of this circle is two wide hand-spans — spread out your 10 fingers fully, thumbs touching.
Whenever you see a rainbow, treat it as a clue to 'shadow-South' (p.14b) i.e. relate its position to the time of day and to South.
The shadow of your head is on the end of the shadow of you, of course, but when you can't see it due to clouds, you might be able to see a rainbow.

Identify the Stars That Pass Directly Overhead

Each star has its own latitude, called "declination", which is very nearly constant.
So the same series of stars of the same latitude will pass overhead that latitude, each night, each in turn, each at some fixed star-time.
This is true anywhere on the globe.
Use a plumb-line. The top of the plumb-line points 'up' to overhead (when looked at from any side, so) where the 'up' lines cross is 'overhead'.
Look past it from 2 or 3 directions to find overhead, especially looking North/South for longitude-timing, and East/West for latitude-positioning.
Then use a star-chart to identify "exactly overhead", latitude-wise, at least. If you can do this, even without a watch, this will fix your **latitude**.
Plus or minus a few kilometres is the expected maximum accuracy from using a plumb-line.
For **longitude**, you need to determine the star-time at which one of these stars is overhead, or at least passing from East of you to West of you.
Greenwich (absolute) star-time, not Local (relative) star-time, that is. To the nearest 20 seconds is okay. See p30c. A 4 min error is 1° error in long.
Alternatively, set your reference star-time to some other longitude, e.g. home, or start, or destination, or a round number.
This time & position data is life-long, give or take a few kilometres of predictable precessional drift per decade.

You can get back to that place later — Match the appearance and timing. It's easy to tell which way to move in order to "match up" (p22a,c; 23c).
You need to know a star's longitude, so you need an almanac or list. Look down the list for bright stars with your latitude, and go look for them.
You will quickly learn about having to wait for the right 'star-time'!

Weather Fronts Distort The (Circular) Wind Circulation

Buys Ballot's law is only approximate, because of the sudden wind-shift at a cold front.
Cols, Ridges and Troughs are likewise noncircular. E.g. 'centres' will sometimes split into two…

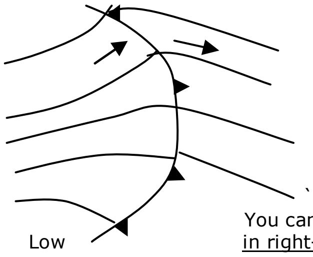

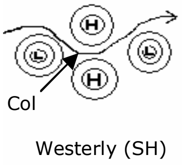

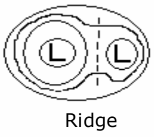

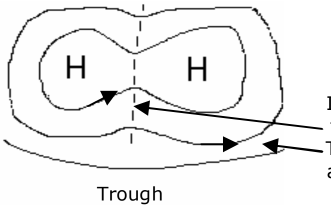

In the trough, crossing the ridge, the winds are turning right.
This airstream is turning left (SH) around the High centre.

You can't accurately say that these winds are 'circulating' around the Low or High! But you can say that in right-turning winds the clouds are forming, under the influence of a Low or a trough (SH), and left-turning winds are under the influence of a High or a ridge—the clouds are dissipating. Swap left and right in NH.

Mnemonic: (Interpret a Compass Needle as) "RED HOT — WHITE FROST"

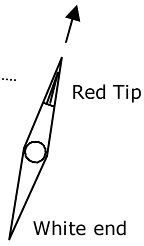

Southerners are looking North to the "red, hot, tropics, dripping with sweat, in the Sun"…

When southerners look South, it is to the "white, icy, Antarctica, frosty cold, in the shadow"…

So when you notice the red point of the compass needle you say to yourself "red-hot!"
and try to feel boiled on that side (as if by a bon-fire), and then visualise the invisible tropics North of you.
And when you notice the white point (e.g. during a back-bearing) you say to yourself "white-icy cold-frosty-Antarctica"
And then try to feel frozen on your 'Shadow-South side', and imagine the South Pole, e.g. point (down) to it.

You can also remember that the Canadian flag is red (but it is not hot in Canada).

Global Orientation: the South Pole Star, and Polaris

(In the Southern Hemisphere…)
Polaris is below the northern horizon, by exactly as much as your South latitude.
E.g. at 45°S. lat. Polaris is below North by 45° Around it the northern stars circle anticlockwise……

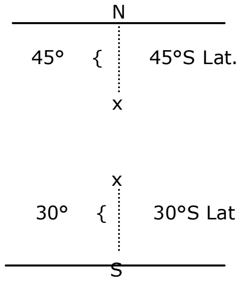

The equivalent "South Pole Star" (a dim pinprick in the midst of dim pinpricks) is
above the southern horizon by exactly as much as your South latitude. E.g. 30° S. = 30° above South
Around it the Southern stars circle clockwise, once per day-night.
Locating the South Pole Star, helps you in the following...

- You can use the bright stars — Cross, Pointers, Achernar — as hour-hands on a 24-hour clockface.
- Star-paths Circle around the Pole Stars
- So do the horizon, the Sun, the Moon and the planets — You can predict all their positions.

Finding the Pole Stars is Fundamental to Visualising Rotation.

- Your Shadow Rotates with the **same clock-sense** as the Sun — not surprisingly! }
- Shadow-Rise and Shadow-Set is **opposite direction** to the Sun — not surprisingly! } Together they can
- Mnemonic: (Face the opposite direction as for the Sun mnemonic p.5b) surprise you!

Face **Sh**adowy **S**out**h** for **S**outhern **H**emisphere (Face Su**nn**y South for **N**ortherners)

"Shadow **Ri**se is on the **Ri**ght; S**e**t is on the L**e**ft" (Same rhyme as for sunrise and sunset, but facing South, both hemispheres)
Your Shadow Passes **South** of Southerners At Southern Noon (compare: 'The Sun is North of Your Nose at Noon')
(North of Northerners at Noon)
Your Shadow Moves **Anti**clockwise — around to Your **Left** — Down Under
(Clockwise — Around to Your Right — in N. H.)
(The Southern Hemisphere Horizon still shifts to the right, in real terms. The Northern Hemisphere Horizon still Never Heads Right')

In addition, notice: **The shadow *tip* moves due East, at around noon,** and in a much straighter line than you expect, all during the day, as the shadow itself changes, from pointing to opposite sunrise, then to pointing to South of the object, then to lying opposite sunset.

---

15 Ways to Locate the **South** Pole Star Or the N./S. Meridian

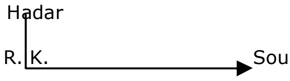

1. Halfway between Achernar and Epsilon Centauri (located from the six bright stars near it)
2. Rigil Kentaurus to S., at right angles to Hadar…………
3. Beta Crucis to S., parallel to gamma-to-alpha Crucis, 5 Cross lengths … Use 2. and 3. together
4. Beta Hydri to alpha Crucis, as far as the beta-to-gamma-Hydri length … most accurate
5. Mirzam through Canopus ("Crucis" says "of the Crux" i.e. of the Cross)
6. Suhail through the top half of the False Cross through Miaplacidus (gamma looks a bit like a thin alpha twisted 90°)
7. Make a triangle out of the two Magellanic clouds toward R. K. — very crude but useful
8. Gamma Hydri to beta Triangulum Australe — the "back" star of the Triangle
9. Alpha Hydri to Hadar
10. Arcturus to between the two 'Pointers' (R. K. & Hadar) } These are used in reverse
11. Spica to between the Cross and Pointers } — to locate Arcturus
12. Middle of Corvus through middle of Cross } Spica and Corvus
13. Atria to the star between Canopus and Achernar, alpha Dorado
14. Peacock to Avior
15. 'Your Latitude' degrees above horizontal, due South

"Cusp North" — Approximate Direction from the Moon

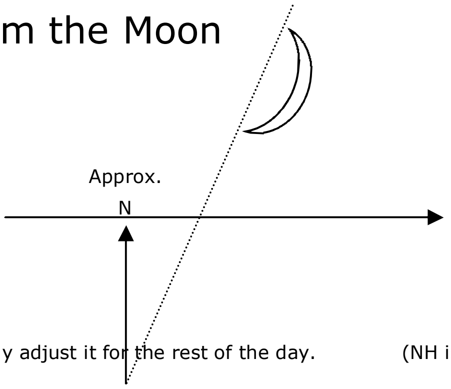

The Sun-Moon line marks the approximate Ecliptic.
The "centre" of that circle of Sun-planets-asteroids is located in "Draco",
quite near "Polaris", the North Pole Star. For a start, ignore the difference…

- Join the Cusps, tip to tip
- Drop the line diagonally
- Keep going down to (near) Polaris (latitude degrees below horizontal)
- Come vertically back up to "Moon North" (in the worst-case +/-60° "accuracy"!)

You have to do this only once or twice a day. Compare it with True North and automatically adjust it for the rest of the day. (NH is similar)

Mariner's "Compass Points" work on Successive Halvings

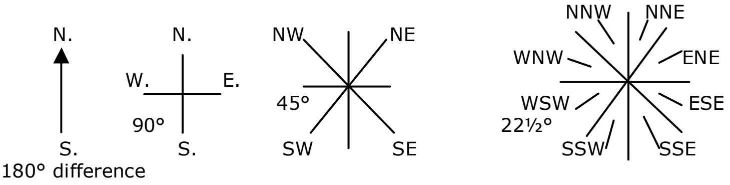

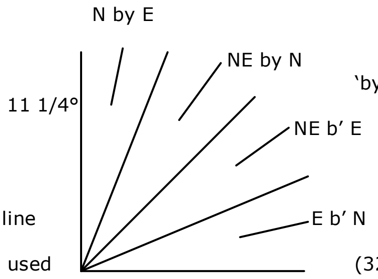

Start with the nearest
major point; add 'by'
'by' means
'come 11 1/4° towards'
(or 'one point')
There are 32 points
altogether on the compass.
One Point is approximately
One Palm wide
(32 around the horizon—test it)

Make your own emergency protractor (e.g. as a compass, sextant or plotter) by folding any piece of paper, and folding it again and again, matching line to line so as to exactly halve it with every fold.
One more halving and you get 64 divisions, each of "100 mils" (so named and used because **1mil is approximately 1 in 1000 e.g. 1 m in 1 km**) i.e. 6400 mils around, clockwise, 17.7778 mils per degree.

Pole Star Locations — "(Latitude)° above and below, South and North"

Why? Think: at 0° Lat., they are 0° high. Yes? At 90° Lat., they are 90° above and below horizontal. Yes?

If you need further proof: Draw a Diagram of two parallel rays to South Pole Star...

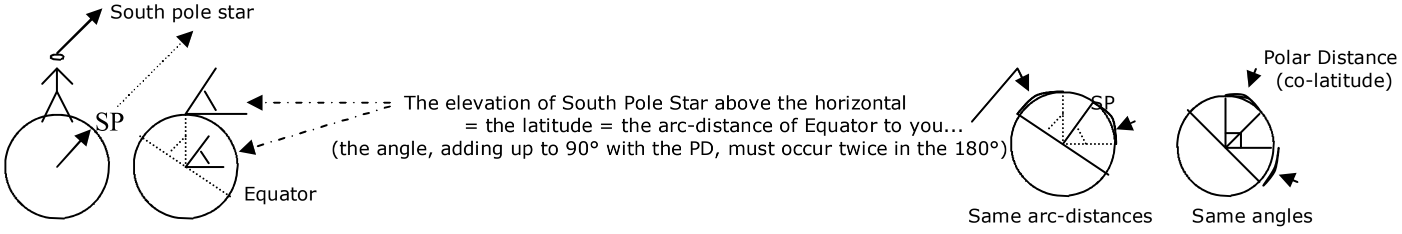
South pole star
SP
Equator
SP
Polar Distance
(co-latitude)
Same arc-distances
Same angles

The elevation of South Pole Star above the horizontal
= the latitude = the arc-distance of Equator to you...
(the angle, adding up to 90° with the PD, must occur twice in the 180°)

The Sun Passes Quickly Past North-South Nearer the Tropics

This pattern shows the hourly change in Sun or shadow direction at 30° latitude, Mid-Summer...

The Sun is East/West at (noon −/+ 2 hrs 44½ m.) at 52.9° high.

This shows the seven hours' sunlight (one side of North/South).

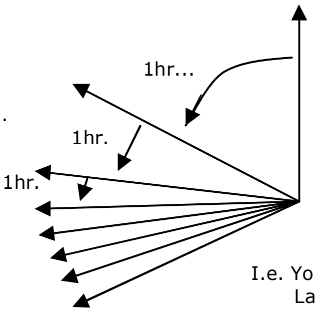

Don't get caught misjudging the Sun's horizontal direction near midday. When it is high, it moves more than 15° per hour in horizontal direction.

At other latitudes the pattern will be different — more uniform away from the Equator, and with a different day-length, depending upon the season.

I.e. You need to calculate each particular pattern from the Latitude and declination and season (as if calibrating a vertical stick sundial).

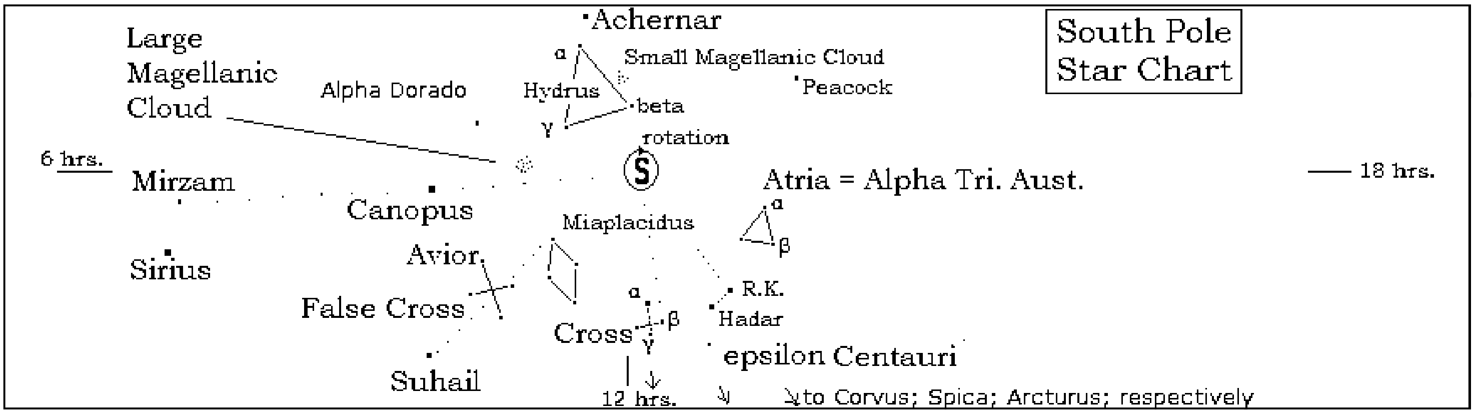
South Pole Star Chart
Achernar
Large Magellanic Cloud
Alpha Dorado
α
Hydrus
Small Magellanic Cloud
Peacock
beta
γ
rotation
6 hrs.
Mirzam
S
Canopus
Atria = Alpha Tri. Aust.
18 hrs.
α
β
Miaplacidus
Sirius
Avior
False Cross
α
R.K.
Hadar
β
γ
Cross
epsilon Centauri
Suhail
12 hrs.
to Corvus; Spica; Arcturus; respectively

Wind-Direction-Change Rule for Buys Ballot's Law

If "the Low is on Your Left Or the High is on Your Right"...**and** the wind is gradually **changing**: Add in this new information as follows...

Imagine how the Low and High are moving across the ocean/continent as you **reenact the wind change**...

Exaggerate the change as you *twist yourself to follow the speeded-up movement* of the wind change in your face.

Point to the two choices (left & right) while you twist, then ask yourself which of your hands is moving to the east.

Usually only one of the two options above — 'Low' or 'High' — will make sense of the general movement of the weather systems to the East.

Temperate systems usually do move to the East from the West. (Both Hemispheres)

..From 0 knots (stationary highs) to 15 (average) up to 25 knots (Lows in winter).

Tropical cyclonic hurricanes move Westerly initially at 25-40 knots, then recurve to the pole, but the "re-enact" rule still works to show up its movement.

In this way you can decide: "These winds are from a Low" or "from a High".

Don't, of course, confuse local winds, e.g. a sea breeze, with the more general 'synoptic' circulation due to the large-scale pressure systems.

To further help you decide where the wind has travelled from —from your right or from your left— check out the temperature and absolute humidity — i.e. the air-mass characteristics. Is the air hot and dry, or cool and moist, or warm and humid — which one is supported by the geography?

Four Reference Directions — Name Them: "**Rise, Noon, Set, Midnight**"

..not E N W S

The trouble with "North" is that it is sometimes behind you... then you have to think backwards.

Use all four and you always have one in front of you... N., S., E., W.

Give them meaningful names as in the title... and you can picture the meaning.

Base it on the Sun's daily movement and the effect is vivid.

To flesh out the meaning even more, **add in a second clue: "Rise/Start; Noon/Hottest; Set/End; Midnight/Coldest"**.

What this does is to take the outside picture inside with you, when you enter a new building, a hospital complex, a shopping centre.

E.g. "I'm now looking toward the mid-morning Sun direction, i.e. 'after rise' or 'before noon', which is 'warming up' ".

It is then up to you to
1) get to know the Sun's daily movement.
2) locate the Sun's path outside, once.
3) *keep track* of your twists and turns when the Sun is not visible e.g. at night.

Ignore the time of day or night—it is the directions which you are imagining. You use the real Sun only as a start-up hint.

Point to 2 Places on the Equator — Halfway between the Poles

You can find the 2 Poles, apparently 90° apart, so that means you can point 45° — halfway — between the Poles, to find the Equator, then 90° to that for the opposite Equator (the "45°" is because the Equator is 90° from each Pole and we halve that arc).
It locates (1) the furthest point on the Equator from you e.g. Brazil then...
(2) the nearest point on the Equator from you e.g. New Guinea—ish.

From central Australia…

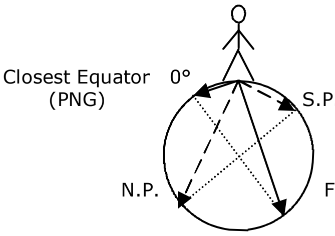

Farthest Equator. Notice that it lies to your South! (Brazil)

Shortest Shadow South — Opposite Highest Noon. Always Imagine "Shadow-Noon" on your "Shadow South" Side

i.e. Your shortest shadow is South of you at noon. Memorise how short it is (this week). For the northern hemisphere use "North"
("shadowy-North")
(Shadow Noon is Opposite Direction to "Noon-North") .."Noon-South"
Imagine your shadow moving left around… ..right-around

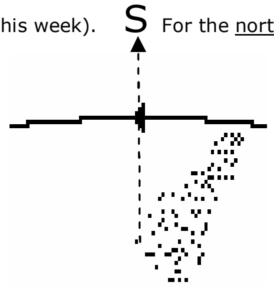

In the morning: South must be "ahead" of my shadow ..North
At midday : Look — see how short it is! … **Due South at noon** **..North**
After noon : My shadow has "passed" South ..North

To guess North from your own shadow, see "Walking stick North". (p.30b)
Try to feel the icy South, freezing that side of you! ..North
(As on a cold night facing away from a campfire).
Mentally **visualise**: Antarctica below horizontal; an aurora above that horizon; Southern stars behind that blue sky. ..Arctic; Aurora Borealis
In the tropics, you'll have to be more observant, to know your shadow's expected movements more accurately. ..Northern stars

---

Find the Polar Axis by Star-Trail Bisection

You need to imagine the centre of the star-circle.
A Hypothetical Example: Lost in a stormy night! You want to find North-South.
You catch the occasional glimpses of unknown stars.

Lie down and line up one such star with a tree branch.
Maintain your exact position and wait for another glimpse.
It may only need a couple of minutes to see its direction of movement.
You can also point something straight at the star and fix it in place, to see the star's movement.
Bisect the result. It points to Polaris, or to the South Pole Star.
You now need to know your latitude, roughly, to locate Polaris or South Pole Star.
The higher the star the clearer the result — it will be moving West.
The only ambiguity is a low star moving horizontally (to your left).
The Southern stars will be moving slowly; northern ones, quickly
— so you may be able to guess which is which.
Northern Hemisphere is similar — but Polaris is visible.

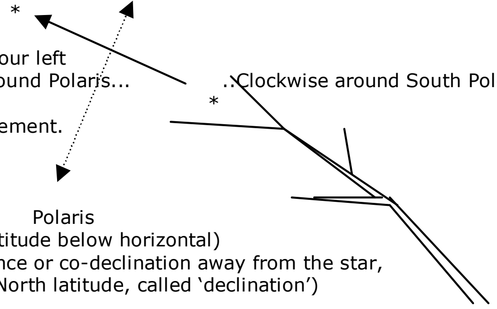

South Pole Star (latitude above horizontal)
(This star is moving to your left
Anti clockwise around Polaris... ..Clockwise around South Pole Star)
Polaris
(latitude below horizontal)
(also polar distance or co-declination away from the star,
if you know its North latitude, called 'declination')

The Rabbit Looks up Ahead at Full Moon Using the Ecliptic

When the 'cusps' disappear at full Moon, the Rabbit ears and face appear,
and let you know which way the Moon is moving.
These views are for the Southern Hemisphere — where the 'Rabbit' is upright.
In the first diagram, the cusps are difficult to distinguish, and so is the Rabbit.
In the second diagram, you don't need the cusps, because the ears show.
The dashed arrows point down to near Polaris.
As an alternative to cusp-North, you can use the Ecliptic perpendicular to the cusps.
The other, solid arrow, falls roughly along the Ecliptic, with the light side of the Moon
pointing to the Sun and the dark side pointing to the anti-Sun.
Therefore you can judge approximate sunrise or anti-sunrise, sunset or anti-sunset,
just as soon as you see the Moon, either from its cusps, or from the Full Moon.
Find the nearest horizon on the Ecliptic and picture which of those 4 directions it is.
A setting Moon in early daylight will show anti-sunrise, but after dusk will show up sunset.
An eastern Moon before dawn indicates sunrise, but in the p.m. will point to anti-sunset.

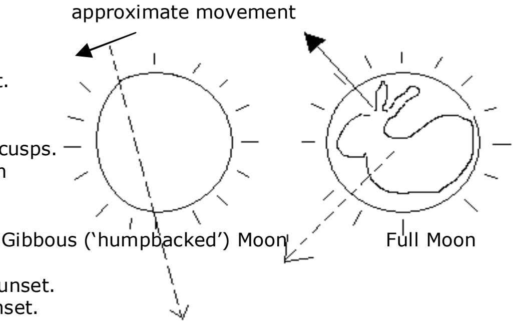

"E. W. N. South" **Of What**!? Rule: **Say** What

- Say 'True' and you mean 'degrees East of True North' 270° T = due W.
- Write "Gee" and you mean: relative to the map **grid**. 270°G ≠ 270°T Both are clockwise around, but they start off from different directions
- S20°E means '(Start at)South(;come)20°(toward the)East' e.g.S20°E=20° to the right of the white end of the compass needle. (old system)
  Clockwise/to the Right/East-around-from-North is standard if not otherwise specified.
- 270° M says 'Magnetic' — what a good compass should read.
- 270° C says 'Compass' — what this compass does read.

Note that these are standard: 'G' does not mean 'Geographical' or 'Great Circle' or 'Grads' but is recognised as 'Grid'.
'M' does not mean 'Map', 'Map grid' etc.
xyz° can not mean anti-clockwise.
The remaining problem is how to shift between different methods… See section 4: "East Least West Best" rule. (p.107b)

Perspective Distortion — The Sun Moves in a Straight Line!

The 'horizon' (at sea) seems "straight", yes? but joins to itself! This is because you see it from the centre of a circle. That's no problem yet.
Two different great-circles circling around you do cut twice, but *it certainly looks strange*, since looking at the gap between them, you see "parallel straight lines"; Yet to the left **and** right of that you will see them both join!
Great Circles Intersect Twice. **Nowhere does either circle seem to 'curve', yet they manage to meet.**
The Sun 'comes up and goes down' but is actually 'moving' in a nearly straight line—as straight as the horizon…

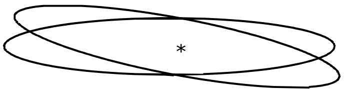

PUT YOUR HEAD INTO A HOOP; GET OUTSIDE; CHECK IT OUT; GET IT INTO YOUR HEAD: GET USED TO IT!
Caution: Don't look directly at the Sun. Watch your shadow and you will see that it moves in a nearly straight line from East to West.
The Sun's average path — the 'Equator' up in the sky — where the 'Saucepan' (Orion; the Pot) moves — is a great circle.
It intersects the horizon due East and due West. At times of year other than at Equinox, the Sun's path is still almost a great circle.
The Milky Way is another maximum sized circle; and the Zodiac (i.e. the Ecliptic) is another such 'Great Circle'.
They hit the horizon twice, in opposite directions (not E/W).
2 other 'great circles' you need to visualise are: Satellite paths; and the Dark Blue Band. (Star-trails are usually 'small' circles, and do look curved.)

Locate Polaris (and Find North) by Imagining the Centre of Circling

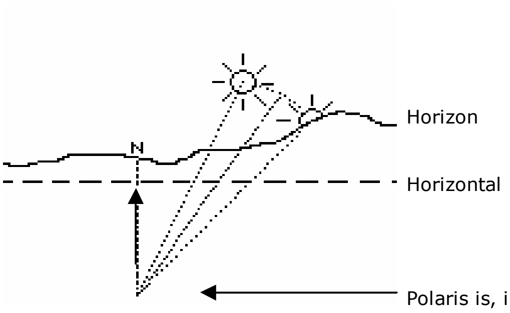

This works for the Sun, Moon, planets or stars.
You *find the centre of their circular paths by bisecting the chord*.
(Northerners locate the South Pole Star, South)

The **perpendicular bisector** of an arc or chord passes through the centre of a circle...

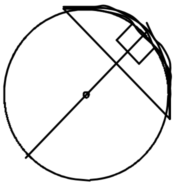

Moonrise varies quickly from day to day
— make sure you check its exact position.

Don't look directly at the Sun to do this!

Polaris is, in degrees, at 'your south latitude below horizontal' (not the horizon) — judge it carefully.
and/or, 'polar distance (co-declination) away from the Sun': 90°-N dec, e.g. dec=10°S 90-(-10)=100°

The Eagle Clips the Equator, South of Altair

Find the star 40% along the southwest side of the rhombus…

Finding the Equator helps you

find the Ecliptic (not shown)
and find where the Milky Way crosses the Equator…

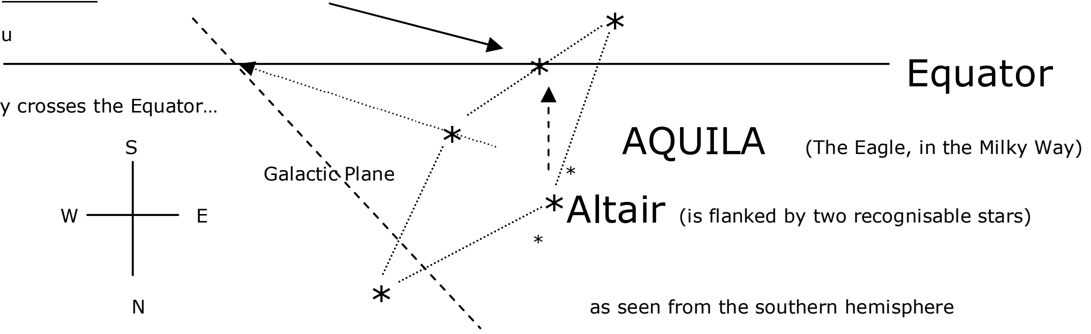
Equator
AQUILA
(The Eagle, in the Milky Way)
Galactic Plane
S
W
E
N
Altair
(is flanked by two recognisable stars)
as seen from the southern hemisphere

Steering Winds

Mid-to-high-level (upper) winds tend to push sea-level pressure-systems along
at **half the speed of the upper winds, and in the same direction**.

This rule is said to work best with small, weak, less well-developed systems; and with hurricanes.

E.g. If the low is on your left And if the upper winds are coming from your left
Then the storm centre might pass straight over you.

Since the upper level winds often oscillate in the 'Rossby Wave', and tend to generate high and low pressure areas beneath them (because the 'wave' causes the upper winds to slow down or speed up) the High-pressure systems will on average be steered slightly toward the Equator, and the Lows will go slightly toward the Poles, as the systems move to the East — a rule of thumb worth remembering.

But this extra steering-wind rule will alert you to non-average situations.

"Rise-Noon-Set-Midnight" carries its meaning across the Equator! N.S.E.W. doesn't.

Name the directions according to the average outside picture…
This is a visual system and addresses the disorientation of turning around.
For back-bearings i.e. when you turn around: replace 'rise' with 'set',
'noon' with 'midnight' (start with end, & hottest with coldest)
e.g. if the old direction was "after setting" the new one will be "after rising".
When you turn right or left adjust by one of the four notches
e.g. 'after midnight' becomes 'after rising' if you turn 'with' the Sun.
Note: 'after the setting' (or 'before midnight') is a direction of the evening Sun
— S.W. — and can be bright sunlight, in midsummer !
And 'after midnight' ('before East') is the direction of the South-East Sun.
At Mid-Summer sunrise it can similarly be very much bright !

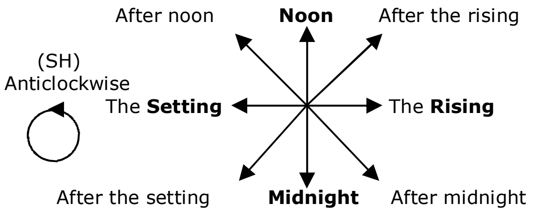

(The Northerner's version needs to go clockwise!)

Example: You enter a hospital with "setting" (i.e. due West) on your right, so you start at "midnight" (SH). You then navigate through look-alike wings, look-alike wards, and look-alike staircases, remaining oriented, and finally emerge, expecting "rise" on your right — having not looked out of a window for hours!

Global
Sun
Stars
Moon
Clockwise

Global Orientation: Half the Arc-Distance below Horizontal

Point out any country or place on Earth, by first finding the (arc-)distance there, and then pointing below the horizon…
The method works **from any place on Earth, in any direction, to any place on Earth**.
Step 1. Measure the angle around the circumference to it. E.g. guess from a globe of the world (This is an angle at the centre of the globe)
Step 2. **Halve** it. That is how far below horizontal it is.
E.g. South Pole from 30° S. Point 30° below since it is 60° away
E.g. Perth to Peking (which is directly North) is 32 S. to 40 N. Therefore the 'arc' to Peking/Beijing is 72°…
Try this (outside): Go through the motions of pointing out a whole circle of real places on earth, from the horizon ahead of you, to the horizon behind you, using your index finger.
E.g. point North, then down a bit to below horizontal e.g. New Guinea, now onto the North Pole.
Continue down until pointing down underneath you e.g. mid North Atlantic.
At this stage you have pointed through 90°, but have pointed out 180° around the Earth.
Turn around to continue, (Brazil, Antarctica) bringing your finger higher and higher until it is back to horizontal pointing at the hill on the Southern skyline.
Your index finger has pointed out a 360° circle (as measured from the centre of the Earth) by traversing only 180°!
Get Hold of a Globe. Go Outside. Point Out A Few Countries.

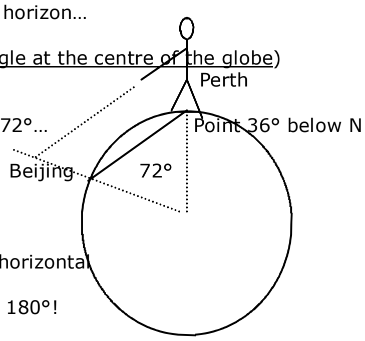

" Noon " is Not at 12 O'clock

' High Noon ' is at 'midday' i.e. in the middle, between sunrise & sunset. And that is **not '12 noon' or '12 midday'** but at North — Noon North. (SH)

- In summer-time 'high noon' may be near 1 or 1:15 p.m. or so. If you are on Summer Time, true noon is one hour late.
- Even in winter-time it may be over half an hour away from 1200 hrs. This is because we use a one-hour time zone to cover 15° of longitude.
  i.e. you may be near the next time-zone, longitude-wise, and therefore be up to one half an hour 'out'.
  You can check this up if you have a map with longitude on it. The middle of the time zone is at 7½°, 22½°, etc, in steps of 15°.
  Then it will take 4 minutes to travel an extra degree westwards. Work out how early or late the Sun will be in your longitude.
- Plus; the Sun speeds up in its 'orbit' 'around' the Earth when it is close in the Southern Summer, getting up to 20 minutes 'fast', by the end of summer; likewise it can be up to 20 minutes 'slow' at other times, in addition to the time zone effect. See the Equation of Time prediction opposite.

The best thing to do about this, to predict when, exactly, it will be at 'midday' today, is to look at the Sun yesterday.
(Don't! Use a shadow instead. Notice when it is due South.) Or see 26a.

Equatorial Stars Rise and Set Due East and West

Always

Globally

- The Sun at Equinox — about March 21st, September 22nd
- The Pot in Orion — Mintaka, the north-west star of Orion's belt — the one that sets first — is right on the Equator
- Various dull stars — like the one south of Altair
- A Half Moon at solstice — give or take 5° or so — It must be 90° from the Sun which is at the solstice position, so that the Moon is at the ecliptic equinox (Equatorial) position.
  Other phases can be worked out, e.g. Full Moon at Equinox, but see p. 29d on how to predict an equinoxial Moon every fortnight.
- Planets 90° from the Sun-at-solstice
- The equinoctial points, if you can identify them yet. One is between Spica and Regulus; the other is between Diphda and Markab.

The Moon Cusps Push and Pull the Hour Hand **Backwards**, Down South

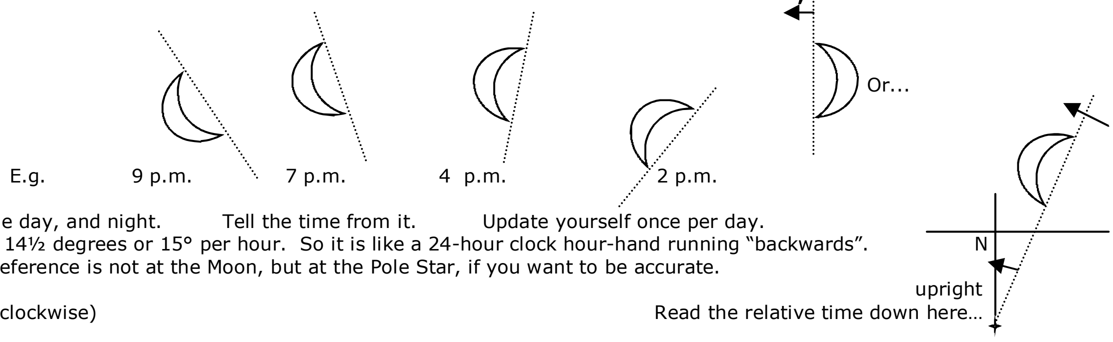

Watch the Moon rotate during the day, and night. Tell the time from it. Update yourself once per day.
The Moon rotates like this about 14½ degrees or 15° per hour. So it is like a 24-hour clock hour-hand running "backwards".
The correct vertical to use as a reference is not at the Moon, but at the Pole Star, if you want to be accurate.

(NH — the Moon seems to turn clockwise)

Back-Bearings Killed Sleepy Pilots

Back-bearings are opposite directions — "going back home".
The Babylonian system of 360° in a circle has a lot to answer for, when you want to turn around!

What is required, is to add or subtract 180° (groan)!

But after an all-night mission, tired navigators can't be relied upon to come up with the right answer.
In World War II, many air-crews perished from following the wrong back-bearing back.
Happy, relaxed, but lost.
It was a problem requiring a solution, and a simple solution was found. What is it? (page 18e)

Half the Arc-Distance below Horizontal — 2 Explanations Why

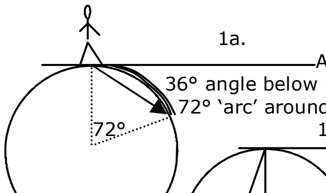

72° 'arc' around the circumference — twice the angle from the tangent — but why?

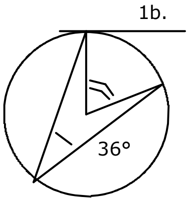

The angle at the centre
(used for 'arc-distance')
is always double the angle
at the circumference...

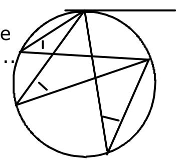

..It doesn't matter which point
on the circle you use,
as long as it is on the one side
—on the larger arc here—
you get the same angle each time.
The tangent point is one such acceptable point — the last one.

2. (Isosceles triangle angles... (180 – 2α)/2)
= 90 – a

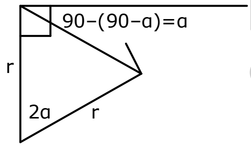

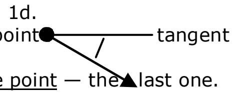

The "Equation of Time" Diagram

The real Sun moves quicker through the sky near Christmas because it is closer to the Earth (i.e. we orbit it quicker, elliptically!).
I.e. The Sun 'gets ahead of' its predicted average position, according to our clocks, in the New Year,
And doesn't come back to average until about the end of August.
After that, during September to December, it will be behind average. But then, in addition, the details are complicated by how it also zips quickly past (gaining on) our longitude lines near solstice — summer or winter — when it is near +/–23½° latitude.
Here the longitude lines crowd together, if only slightly, giving the maximum effect in June & December, super-imposing upon the preceding scheme.
The **overall result** may be graphed...
See the mnemonic on p. 86b

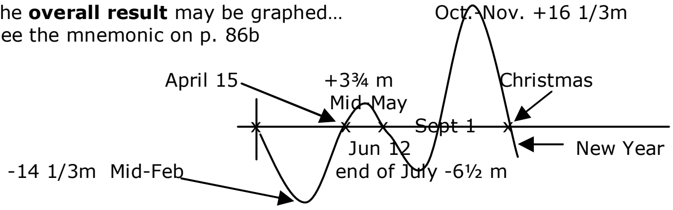

(both hemispheres)

Note that when the astronomers say '+', the Sun is behind itself;
we look at the position of the Sun (not directly at it!); it comes up 'late'
and we have to add (about a ¼hr) to our guess at the apparent time,
to find the clock-time. We say our clocks are 'fast' at predicting
the Sun's position, but it is really the Sun which is lagging!
cf Australia is in a "**+**10 hr time-zone" = 10hrs ahead, or fast on Greenwich

---

Rigel-to-Capella is a good S./N. Pair

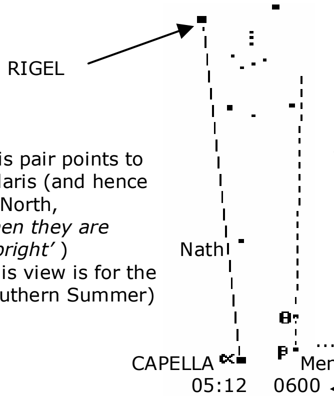

This pair points to Polaris (and hence to North, *when they are 'Upright'*) (this view is for the Southern Summer)

You can also see that the Pot handle points down through the middle of Orion's belt to Meissa and on to Polaris. The Pot 'handle' is actually the scabbard of Orion's sword—seen upside down (SH)

If you cannot see Capella, then Bellatrix to Elnath, right alongside, is also a good guide to Polaris.

Auriga, The Charioteer
....**The six o'clock stars are a useful N./S. pair** — they are upright at 6 a.m. local star time.
← 'Right Ascension', or star-time, in hours of longitude East of the first point of Aries

---

Atmospheric Pressure Changes

The air pressure falls before a front, and rises reasonably rapidly as the front passes. It falls in a trough; rises in a ridge; falls in a heatwave inland; falls with altitude (about one millibar per 10m).
The atmospheric pressure normally fluctuates by a couple of millibars every day in a twice-daily cycle tied to the Sun (3mb in the tropics) and the wind speed may vary with it. The daily pressure on average peaks around 10 a.m. and 10 p.m., being lowest at 4 a.m. and 4 p.m.; This on average, is said to vary the wind speed by a couple of knots, lowest at 9 a.m., peaking at 3 p.m., but all of this is usually masked by other weather changes in the temperate latitudes. Pressure fluctuation is much more significant in the tropics, with any fall (or flattening) more than a few millibars during summer/autumn warning you of a possible hurricane (You then look for long wavelength ocean swells preceding the storm centre). Because we can't feel atmospheric pressure changes, it is very helpful to use a barometer, to detect the longer-term pressure change trend. Test your barometer to see that it does not respond to temperature changes!

Watch the Clouds Cross the Moon

You can detect the upper wind direction most easily by watching clouds move past some fixed point in the sky.
Look during the night but don't forget to look during the day also.
The stars, high buildings, high trees and mountain tops also work well.

The **0, 1, 2, 3, 4** Reference Directions

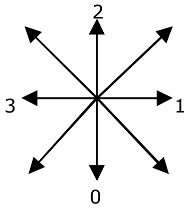

Put these numbers onto the Rise-Noon-Set-Midnight System, so as to follow the Sun.

0 = midnight, South (Southern Hemisphere)
1 = rising, East
2 = noon, North (use South for northern hemisphere)
3 = setting, West
4 = midnight, South — use cycle 4 i.e. call it "0"

This corresponds to measuring anticlockwise! from South! in decimal degrees, that is, in grads!
100 grads = 90° so "1" = 100 grads; "4" = 400 grads or zero grads — add or subtract "4" at any time.
"0.1" in this system = 10 grads; 1/10 of a right angle; **one fist at arm's length**.
Example: In the thick bush, you see your destination but know you will soon lose sight of it. You can't keep a straight line, so you decide to keep track with the 0—1—2—3—4 system and your fist. With each change of direction you update e.g. "0.6, 1.6, 1.0, 0.9..."
For a **back bearing** — when you turn around — **add or subtract 2**.
Add 1.0 when you **turn with** the Sun (**left**), or **subtract 1.0** when you **turn against** the Sun (**right**). (S.H.)

Example: Point to Tahiti

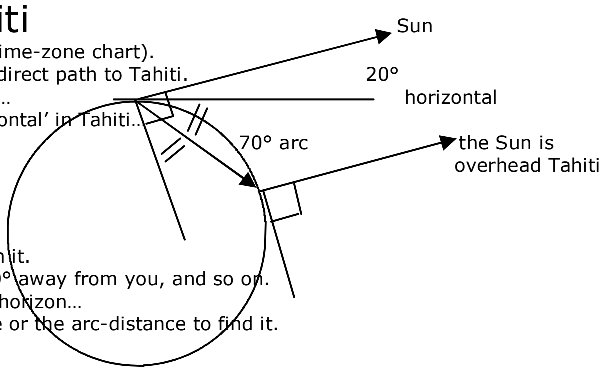

It's 9 a.m. summer-time, the Sun is, say, 20° high and over Tahiti (see time-zone chart).
Firstly, point horizontally toward the Sun (not directly at the Sun!) — that is the direct path to Tahiti.
Secondly, point below the horizon, to locate the actual direction to Tahiti…
Step1. Make a perpendicular from a sunbeam. Perpendicular to the Sun is 'horizontal' in Tahiti…
Step2. Subtract its (present) height (20°) from 90°
The arc from you to Tahiti is 70°
Explanation:
(90° - the height of the Sun), the 'co-height', must be the arc-distance
from you to underneath the Sun. Think of it like this...
If the Sun is 0° away from you, it will be 90° high — you are underneath it.
If the Sun is 10° away from you, it will be 80° high, overhead a point 10° away from you, and so on.
Step3. halve the remainder (70°/2=35°) Now add back on the 20° above the horizon…
"Tahiti is 55° below the Sun" from your point of view. You didn't need a globe or the arc-distance to find it.

The Sun Oscillates Seasonally between +/- 23½° Latitude

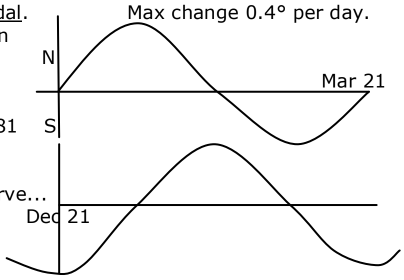

To approximate the Sun's "Declination"—or North Latitude—you approximate its variation as sinusoidal.
Using a 365 day cycle: You approximate 360° as 1° per day, so for ease and accuracy of calculation you can and should approximate the day-number as 30 days per completed month + day of the month.
E.g. **March 21st** is 2 lots of 30 days plus 21 = "81" days or degrees.

Then to use a sinusoidal curve, you need to run from Equinox to Equinox — i.e. starting at March 21, Day 81

Declination = 23.45° x sin (day number of the year – 81)

E.g. 0° in **September 21**, six months or 180 days or degrees later; uses day no. 261.
From solstice to solstice you can use a minus cosine curve...

Declination = 23.45° x –cos (day number of the year + 9)

..since the solstice is around 10 days earlier than the New Year.
Dec 21 solstice, is day no. 11x30 + 21 = 351 "–23°½ in December"
E.g. + 23½ degrees on **June 21**, day no. 171.
From this approximation you can predict noon-height (and, with the right formulae, sunrise & sunset time & direction and E./W. height, of the Sun).

Use the Southern Cross as an Hour Hand

You can visualise the rotation of the Southern Cross as a huge 24-hour clock hand.
It is correct on April Fool's Day i.e. upright at midnight (at 183°T), then it gains two hours a month, or 30°. (Use the correct vertical! — South.)
As a learning aid, mark the 'Southern Cross time' on your calendar every month
e.g. "two hours fast" at the start of May; and "four hours fast" at the end.
Then, anytime you are out at night, you can ask: **"What's the time — by the Cross?"**

Picture the Southern Cross during the Day

Yes it is there behind the blue sky… "If there were an eclipse now, it would show… there!"
When you are navigating by the 'spin' of the Earth, and by the Big Picture, the proper reference direction is the North-South rotation axis,
Which means: Polaris + South Pole Star.
In practice, what we Southerners see, is the Southern Cross.
So if you are imagining: sunrise — noon — set, or Moon position, or a star-compass course, or a shadow-south direction,
you ought to be thinking: "North — Polaris down there; South — South Pole Star up there", then you can picture the Southern Cross as well.

The Moon is Lit By the Sun

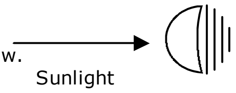

When the Moon is **opposite** the Sun, we must see its fully lit face… See diagram of the first man looking at the Moon, below.
e.g. a full Moon rising at Sunset; Setting at Sunrise; Overhead at Midnight or at least highest in the sky.
(A fortnight later…)
When the Moon is **near** the Sun, we see only its dark backside or at most a thin crescent… See the middle man below.
e.g. a crescent Moon is only ever seen near sunrise or sunset. Wait until the Sun is below the horizon — Don't look at the Sun!
At other times the Sun has set, taking the Moon with it, or has risen, outshining the pale crescent nearby.
(A week later…) When the Moon is **halfway** between, we see half of each side… See the third figure...
e.g. half a disk overhead — only at sunrise or sunset; half a disk setting — only at midday or midnight.
A half full Moon is misleadingly called a Quarter Moon… since it takes four weeks to move between the four "quarters".

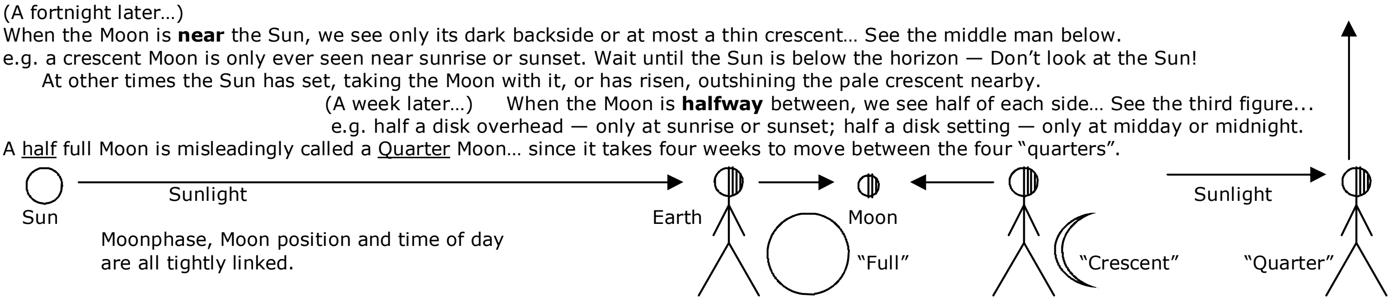

Moonphase, Moon position and time of day are all tightly linked.

Back-Bearing: **Plus or Minus 200° Then Minus or Plus 20°**

(Or — for the very advanced sleepy student: adjust by 20° first (+/-20°) then -/+200°)

Examples: 10° is small, so make it big, 210°, adjust back to 190°
290° is big, so make it small, 090°, then back up to 110°
Advanced: 190° is middling, so boost it first, by 20°, to 210°, then remove 200° and use 10°
210° is middling, so try +20°, then –200° = 030°
When you end up with numbers greater than 360°, remove 360°. When less than zero, add 360°.
Difficult: -10° is small; first convert it, using +360° (= 350°)
350° is big so –200° +20° = 170°
185° + 200° -20° = 365°, so adjust it back to 005°

a simpler solution with a protractor compass, is to...

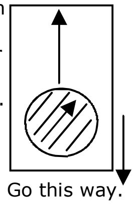

(see page 109b)

There is another solution to back-bearings — decimal degrees — but it hasn't taken on yet.

Earth Sun Stars Wind Anticlockwise

Use **Latitude** for a **Distance** Scale (Not 'Long'-itude, Despite the Name)

(Especially on marine charts, atlases and old maps—Newer maps are likely to have the mostly meaningless but well-scaled international map grid.)
Latitude means "distance" from the Equator. Remember the polar "distance" 90° complement to latitude? which measured arc-distance, in degrees!
1° of latitude = 60 nautical miles = 111 km The world is 360° around; or 360x60 = 21,600 nautical miles; or 40,000 km
The circumference of the Earth determines how long a kilometre or a nautical mile is — at least in the somewhat outdated original definitions.
(The N/S circumference of the Earth has proven to be closer to 40,007 km, and the E/W circumference is bloated by the Equatorial bulge.)
One 'kilometre', comes from the decimal system: whereby one circle equals 400 'grads'; one grad has 100 divisions… 40,000 km
Use the latitude scale on the left & right sides of a Mercator map to measure length or distance — but only at that latitude on the map.
But do not use the longitude scales running across the top and bottom of the map — since the spacing of longitude lines should vary with latitude. Longitude lines must converge towards the Poles, to nothing. They span a variable distance = 1° x cos (latitude), but a Mercator map will (falsely) show a constant separation for longitude lines (this is to make compass courses look straight). A Mercator projection will also show a variable separation for latitude parallels, even though they 'should' be constant. This is to preserve the shape of land areas—The latitude stretching must match the longitude stretching. Therefore any scaling shown on the E/W spacing is only valid for the latitude at which it is subdivided.
When a latitude scale says ½° or 30 n ml or 55 km, it means it, at that latitude, but when a longitude scale says ½°, the distance meaning is missing

---

Using the Declination of the Sun

- **Noon Height** = 90° – latitude + declination = (from horizontal) Convert "N." to + ; "S." to – e.g. 30° S. = –30
  E.g. Lat 30° S.; Decl. 20° N : Height = 90+30+20=140° or 40° (the two (opposite) heights must add up to 180°)
  i.e. The more the Sun 'declines' away from the Equator, and the more you decline on the other side of the Equator, the more the Sun drops in the sky. Alternatively, the colatitude (90–L) is the maximum height of the sky Equator... Then adjust for the declination.
- **Sunrise/Sunset** — time and direction prediction
  Cosine (angle from North/South) = sine (declination) / cosine (latitude) Try dec =23°.5; lat =37°.1; 55°.67; 62°.6, for a different hexagon
  Cosine (time in degrees from noon) = –Tan (Lat.) x Tan (dec.) 1° of longitude = 4 minutes of time.
  E.g. measure the time between sunrise & set; divide by 2; solve for Lat. (make sure to use horizontal, not horizon.)
- **East/West Sun** — Time and height prediction
  Cosine (E-W time in degrees, from noon) = Tan (dec.) / Tan (Lat.) e.g. 4 hrs each side of noon @ Lat 41° (Northern Tasmanian Summer)
  Sine (E-W height above horizontal) = sin (dec.) / sin (Lat.) e.g. 45° high E./W. from Lat 34 1/3°; +/– 3h21m; Mid-Summer

e.g. take the time when the Sun is E. & W. (opposites — you need to remember to carry 60, not 10, & to interpolate carefully) and then solve for Lat.
See p. 86b for a memorable laugh. Many more formulae and techniques are available, from books on astro-navigation.

Perspective Distortion — Use the Correct Vertical

Vertical is vertical, right? Wrong! To tell the time by the sky, take care.
The higher a pair of points (e.g. stars) gets in the sky, the more you have to choose which vertical is needed for a reference, to judge 'vertical' by.
Each different reference direction will give a different answer to "How vertical are those two points up there?" e.g. Moon cusps, Southern Cross
A side view makes this clearer....
..a straight line which looks horizontal when looking south,
looks diagonal in the East

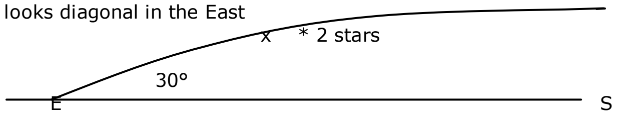

The top view also makes it very clear… a line XY high up in the sky cuts each different vertical at a different angle..

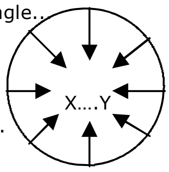

Don't gauge the angle of rising/setting (p9a) at anywhere other than at the rise/set horizon in the rise/set direction.
Don't use the nearest horizon, as your reference for which way is up.
To visualise a constant 'rotation' of the stars, use Polaris or the South Pole Star as the 'centre' of rotation.
I.e. Use N-S longitude lines in the sky rotating in relation to vertical at North/South.

To Determine Upper Wind Direction from the Clouds

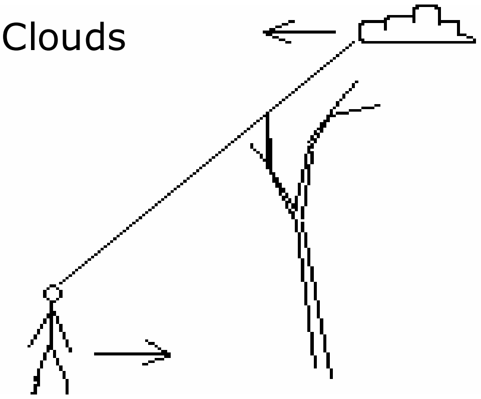

Walk in the reverse direction, while keeping a chosen cloud lined up on a chosen high point. You must walk across level ground to be accurate.
Mark your starting point, then drag a line in the ground to mark your walking direction.
This method works well for very slowly moving clouds.

Other methods: Watch cloud shadows.
Use the Moon, Venus or the Sun (only when it is very well obscured).
Look in different directions at the moving clouds
until their motion seems neutralised—they are moving directly toward you.
Use binoculars, to magnify the movement (but never look towards the Sun).

The Sunny Skies Quadrant System

To orient yourself on a clear day...
Use the two ends of the dark blue band — where it hits the horizon — and the trend of the middle bit of the band, in conjunction with the Sun (but don't look directly at the Sun!) and your shadow.

- No matter which way you are facing, you can detect one or two signs to orient yourself by,
  Or at least imagine them if you can't see the Sun just now.
  You don't have to look over your shoulder, which is misleading anyway.

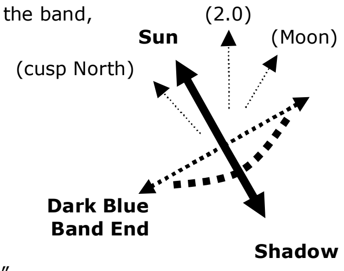

The middle of the blue band is above your shadow, and at right angles to the sunbeams near you.
The whole system rotates, during the day.
If the **Moon** is up, add it into the system,
and **Moon-North** too, which is off to one side of North, and changes slowly during the day.
Combine this system with the fixed systems, of 0, 1, 2, 3, 4 and m.n., rise, a.m., m.d., p.m., set.
E.g. In the diagram, the Sun is to the left of North, so "the Sun is 'afternoon', the Moon is 'before noon'."

The Horizon Rotates to Your Right in the Southern Hemisphere

1.Face the High Noon Sun. (Imagine it, if you have to, when it is not midday.) It moves horizontally from Right to Left (S. H.)
2.Reverse your perspective — let the Sun stand still and the horizon move, from Left to Right.
Imagine the horizon sliding horizontally beneath the Sun (and you shifting sideways with it).
3.Now imagine the Sun rising from the eastern horizon — it moves up to your left — then reverse perspective...
Imagine the eastern horizon moving *diagonally down and away* from the Sun, twisting down on the right as it goes.
Face North. Grip the whole horizon on the East and on the West, and imagine a tilted steering-wheel, and rotate it around the Pole Stars. Imagine a "steer right" motion.
The invisible steering wheel faces you diagonally upwards, while the horizon, as if fixed (diagonally) to the steering wheel, starts off horizontal and then rotates out of its horizontal.
4. Keep rotating the horizon, sliding the northern horizon horizontally underneath the high noon Sun, and beyond, until the western horizon comes up diagonally to meet your now "setting" Sun.
You need to do this outside to appreciate it in three dimensions, to have any chance of visualising it easily.
So put your clock back on the Northern wall (see 5e) and say "See, we Southerners have it right, after all!
— the real movement — of the solid stuff — is clockwise"

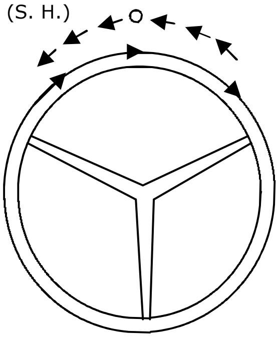

Tell the Time by the Compass Protractor — Tilt It Toward the Equator

Take a **tilted bearing** of the Sun. The answer comes back as so many degrees to one side of noon-North (or noon-South in the N. H.).
You then use 15° per hour to convert that to hours before or after noon-time ('noon' will be nearer one o'clock during daylight savings time).

Step one: align the compass to True **North**
Step two: **tilt** it upward to the 'Equator' in the sky (the Sun's average path)
Step three: rotate the compass arrow to point toward the Sun
Step four: make the edge of the compass cast no shadow
This avoids looking at the Sun
Step five: measure the degrees away from high-noon North, on a tilt

It is more accurate than using a finger-span
to measure 15° per hour
i.e. Use the compass protractor to make a 24-hour clock face
To use the compass protractor as a 12-hr clockface, see the box opposite.

1. Top View 2. Side View

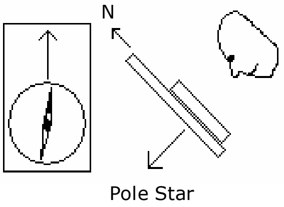

The reason for tilting is because a level sundial/upright gnomon will not give a 15°/hr linear reading (See "The sun passes quickly")

The Southwest Corner of the Great Square is N. S. E. W.

The Great Square of Pegasus only has one really good right angle but it shows N. S. E. W. perfectly.
In the Southern Hemisphere, it is the first star of the Square to rise.

Note: When you hold a star chart over your head
'East and West' reverse!

This is the view for the Southern Hemisphere…

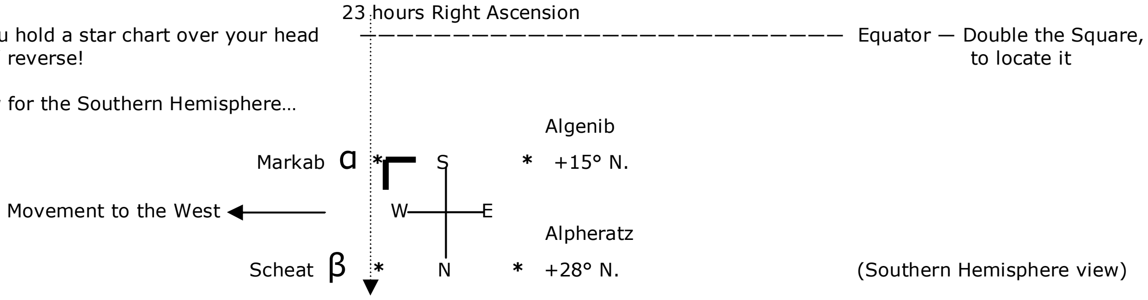

The Moon Rises and Sets 50 Minutes Later Each Night

The Moon moves quickly to the West, but…
It progresses Eastward through the night skies, night by night, by 12½ degrees a day — it 'lags' the Sun or 'moves backwards' during the month.
So that after 29½ days it is back in the same place. It moves about 27 diameters' worth East through the stars, per 24 hours
or about a diameter an hour to the east in relation to the Sun. One 'diameter' is half a degree, and the same as the Sun's.
Sun, Moon, Planets and Stars rise in the Eastern half and set in the Western half, but some are quicker than others. The Moon is the slowest of all.
E.g. (twice-daily) tides get 25 minutes later each tide, on average — half a day + 25 mins apart.
E.g. the 12 month lunar year gains 11 days on the 12-calendar-month year, or "one day per month except for February, on average".
(a calendar month equals 30.4 days; a lunar year equals 354 days. Months 1,3,5,7,8,10,12 have 31 days—1½ extra; ½ extra in 30 day months)
The explanation for the Moon's "slowness" in rotating westwards with all the rest of the sky, i.e. from moonrise to moonset
(it rotates only 14½ degrees per hour, not 15° like the Sun) is that it slowly revolves around the Earth to the East,
in the same clock-sense as the Earth revolves around the Sun,
and so must complete 13 revolutions star-wise, to the East while we see only 12 Moon months.
By the next full Moon, its backside is toward a more easterly star…
For an anti-clockwise rotation system, viewed from the North...

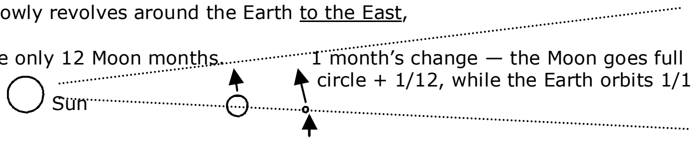

1 month's change — the Moon goes full circle + 1/12, while the Earth orbits 1/12.

Turn 90°? "+/–100° –/+10°" !

Turning right is turning clockwise; turning left is turning anti clockwise.
Clockwise **increases** the count of degrees East of North. So that requires **+100°** -10°. E.g. 257° changes into 347°
Anticlockwise, when turning left needs **–100°** +10°. E.g. 257° changes into 167°
Or use +/-10° first, then -/+100° if that is easier (near 0°). Adjust by 360° if necessary.
There is no convenient choice as there was with a back-bearing, so you must remember whether to use + or – 100°.
Remember to look at the Sun's clock-sense to judge 'anti-clockwise'
— anticlockwise in the southern hemisphere — to your left. 'With' the Sun implies – 90°
— clockwise — to your right — ('With' the northern hemisphere Sun) implies + 90°.

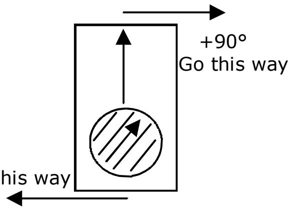

Hint: When sailing, you can always temporarily head either towards or away from, or side-on to, on the left or the right, a landmark from which you want to take a bearing, without having to change tack. Therefore you can use +/–180° or +/–90° from the boat's compass-heading. In heavy weather this means that you do not need to leave the helm to use a hand-bearing compass, but merely sight along the lines of the boat, or cross-wise along the lines of the cockpit, stern or cabin. Simply remember that the more clockwise heading is bigger than the left-hand direction

Global Perspective — Rotation of the Horizon

Each point on the horizon rotates…

- In a circle
- Around in an Eastwardly sense, like all the real matter in our anticlockwise spiral galaxy
- Around the North-South Earth rotation axis
- Along a fixed sky-latitude circle (i.e. declination circle)
- At a fixed angle of rising/setting (one or the other) in relation to the fixed stars
- Once per day

Example: The "Sugarloafs" were on the left of the Sun at sunrise, but six hours later they had moved to be underneath it, by rotating around Polaris and now, at the end of the day, they have moved way over to its right, and will keep going around for the next 12 hours of night-time.

To further help you visualise the motion, remember that the ground under your feet is twisting you clockwise in relation to the sky, in the SH, so that the stars, etc, near the horizon all move only to the left as they rise or fall. For the Northern Hemisphere use "anti-clockwise… right".

Or Judge the Time By Sight — by Mirroring the Sun Around North (To Make an Hour Hand on a 12 Hr Clockface)

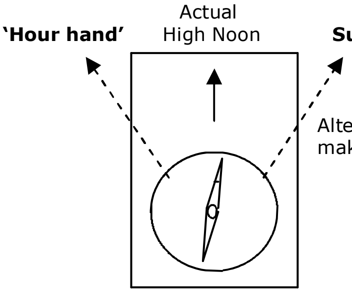

Sun = **'Noon'** position on the Clockface (i.e. ignore real Noon-North after you align the compass properly) (SH)
(Here, the 'hour hand' is showing 2½ hours until 'noon', judging it by sight)
(For the Northern Hemisphere, you have to reverse things… hour hand — Sun — South see p.28b)
Of course this method is really just helping you to GUESS!

Alternative: Line up a pencil tip with Polaris, while making the pencil cast no shadow. With your shadow-side hand, make your index finger ("The Hour Hand") touch the pencil tip, and look identical on the other side of Polaris.
You can make it work on the Moon or a star or a planet if you know when they are North/South of you.
You can make it work for magnetic noon, to avoid using magnetic variation.
It works because the Sun moves at half the speed of an hour hand,
but the doubled-angle shown, doubles that rate of change, to match a clock's hour-hand.
Don't forget to **tilt the compass**, so that its axis lines up with the Earth axis.

The Southern Summer Rectangle is Worth Looking for — since it has four N. S. E. W. guides

1. North: from Procyon, down through the Twins
2. North: from Bellatrix down through Nath
3. West to East: between Bellatrix and Betelgeuse to between Procyon and Gomeissa
4. West to East: Nath to between Pollux and Castor, closer to Pollux

(To remember the two Twins, say "Castor & Pollux, α & β, are North & South")

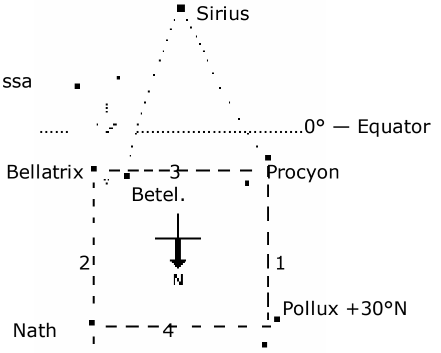

And a Triangle Pointing South

Betelgeuse, Procyon and Sirius.

Clouds and Con-trails and Their Winds

- Fog, which is cloud clinging to the ground, usually oozes downhill. Now, working upwards...
- Topographic cloud banks — may indicate e.g. the plateau edge or a mountain range. I.e. They cling to the topography.
  Also look for eagles soaring along otherwise invisible cliff edges.
- Low clouds — they follow the weather map isobars. Also look for sideslip of flying birds and aeroplanes.
  As a rule of thumb (clenched fist!) low cloud puffs are at least three knuckles wide, alto puffs are two, and cirro puffs are one, from sea level
- Middle ('alto') cloud — the puffs look smaller, the shading, less dense, but no 'iciness' seems apparent. The clouds obscure the Sun's outline somewhat. Their wind-direction is usually different from both higher and lower clouds. This wind will 'steer' low level pressure systems.
  (An 850 mb level chart helps) When a change is due, the middle clouds often betray what low wind direction to expect soon.
- High ('Cirrus') cloud — thinner; smaller puffs if puffy; icier i.e. fleecier, silky, filamentous; transparent, with halos & no shading.
  The Sun's outline remains crisp when seen through cirrus. Don't focus on the Sun!
  The cirrus-level cloud-direction in the subtropics is usually westerly +/−45°.
  Be sure to test this in your latitude. (A 500 mb chart helps) Try to catch glimpses of upper clouds through gaps in the low clouds.
- Jet Con(-densation) Trails — are usually in the high band. Look for them according to a timetable, each day.

The "SWEETHEART" Method of Orientation

We are human, and 'reference directions' are not. So make them more human, more personal, more emotional.
"Home" is not bad. "Sweetheart" is better.
Find something lovely about each cardinal direction. E.g. 'our favourite holiday spot on the coast'
Something you love or hate, want, or fear, about each reference direction…

You will find it easier to maintain awareness of bends
and to interpret meaning onto the skyline — what lies beyond and behind that horizon!

During one walk, these highlights are unlikely to change direction much.

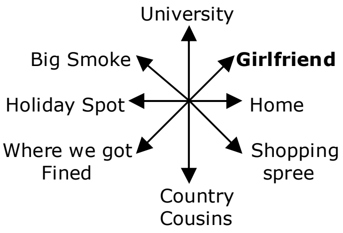

How to Shift Your Horizon — Its Timing and Height in relation to the sky.

You can navigate by the "shape" of the sky, back to a spot where you have been before, by noticing the "position" of the moving horizon.
On a round earth you can shift your position and horizon 1°, or 360°, by travelling 1°, or 360°, around the circumference. (111 km, or 40,000 km)
The line where the horizon cuts across the sky, varies with the latitude & longitude & the time of day. Using the Sun as an example…
Note: The "1° for 1°" rule for your own movement, only applies after you neutralise the other motions, so as to be as if the sky is standing still.
E.g. You need to neutralise the Sun's daily motions, by looking at the same time of day — sunrise, noon, 6 a.m., or 12:22 a.m., or whatever.
E.g. You also need to neutralise the Sun's seasonal motion by looking at the same date every year — Christmas morning, June 25, or whatever.
(Leap years can only make a difference of up to 1° — you could improve accuracy by looking only every four years.)
Rule one: for West/East travel, remember that events like Moonrise seem to happen **later to an observer further west**, and…
Earlier in the **E**ast; **four minutes of time (early) for every 1° of longitude (East).** (these rules work in both hemispheres)
E.g. "Christmas Day Sunrise at camp 4 was at 6 a.m. last year, but now I see the Sun already up at 6 a.m., so to get to where we spent Christmas last year, we would have to go further west to make the Sun lower (the horizon higher) at 6 a.m." (Assume you are at the correct latitude)
Rule two: The horizon drops to reveal new sky as you travel onwards — **1° for every 1° of travel** — this works in any direction.
E.g. "The Sun rose to 75° high, but here it reaches only 74½ degrees — so I would need to travel a half-degree toward the Equator for it to be 75° again." (This example has deliberately chosen a North South change.)
Then you can shift your horizon, its timing and its height, by moving around the earth, in order to match a previous timed snapshot of the sky.

Walking in Phase with the Sun, Wind and Moon

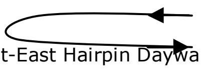

Simple West-East Hairpin Daywalk.

- In a hot climate, walking into the **Sun** is unpleasant, hot and glary, especially in the afternoon, especially if you wear glasses, so...
- The morning **wind** can be very uncomfortably cold, and it is best to plan to walk away from it, too.
- Beach walks are more pleasant when you don't have to battle back into the afternoon sea breeze or prevailing wind; so assess the **forecast** before you choose a start point.
- Night walks and late camps are best done in bright **moonlight** — simply predict the phase.
- Spotlighting is the easiest on a waning Moon — since the **dark period** is in the evening, not before dawn.
- Spectacular pink full-Moons only rise at sunset, so it is best to predict such a **moonrise** and plan an early camp with a view.

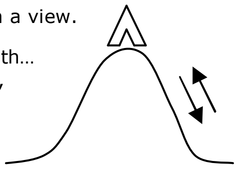

E.g. Up in the afternoon **shade**; down in the morning warmth…

Similar considerations apply to sailing—e.g. aim to make landfall in the pre-dawn. You can identify the harbour, or the coast, by its light-houses, then soon after be able to enter the harbour in daylight.
E.g. Aim never to cross the sandbar on an outgoing tide—with steep breaking waves, in shallow water, getting shallower!

Coordinate Your Route, Pace, Campsites & Timetable To the Prevailing Conditions.

Navigating by the Shape of the Night Sky

Since the star-map — the background to our view of the sky — is unchanging, the 'shape' of the sky involves only the **position of the horizon** across the background of fixed stars, and the **rotation** of that horizon with time.
The sky gives a false impression that the sky itself is spinning, but it is we who are turning upside down! It is our horizon which is changing.

Just as you can neutralise the Sun's motion by looking at the *same time every day* (which will show up only its seasonal and longer term variations) and just as you could neutralise the Moon's motion, by looking at the same phase and Moontime every month (e.g. full moon rise), to show up its unpredictable motion, so too with the stars. You must **look at the same star-time every night**. If, in addition, you look at the stars from the same spot you will notice no change at all in the 'shape of the night sky' or the position of the horizon — star motion is all predictable, with no significant variations.
This makes the stars excellent for navigation — you simply locate the horizon in terms of the heights of the stars, the tilts of star groups, and the star-time of night. E.g. what star appears overhead and at what star time.
Thus, to find that spot again, even much later, you need to make your horizon look identical at the identical star time.
CAUTION: Don't go star-navigating; rather only imagine doing so, until you become an expert. Understanding it is more useful than doing it!
All figures, formulae and theories should be checked before you trust your safety to them.

The Moon is "in Phase" with the Sun — Watch for It

The Moon progresses (12½ degrees per day) contrary to the general rotation, so…
A **new** Moon, first seen after sunset, must grow from a crescent, in the **after sunset** part of the sky, as it separates from the Sun, Eastwards, night by night, more and more so after the new Moon. I.e. a 'waxing' (growing) Moon means Moonlit evenings. After full, the Moon continues Eastwards…
A **full** Moon must diminish, to a crescent, in the **before sunrise** time of night. I.e. a 'waning' Moon leaves the evenings dark, until Moonrise, predawn.
In other words, the predawns are moonlit for the second half of the Moon month, after Moonrise. In the first half, the predawns are dark, after Moonset.
The whole night is dark in the 'no Moon' changeover of months.
During the day, you ask "Is the Moon ahead of or behind the Sun?" I.e. Is the Moon predawn (waning) or post-sunset (waxing)?
Evening spotlighting — is best in the dark before the waning Moon rises — late in the Moon month.
Night-walking—is best in Moonlight after Sunset i.e. on a full Moon, or a waxing, almost full Moon. (early-mid month)
(Think of the Moon as prouder of its nighttime role as it moves away, and ashamed as the Sun gets closer)
Highest and lowest ('Spring') tides occur on or soon after a new or full Moon.
'Neap' tides occur (when the Sun and Moon pull at right-angles to each other) at "half full Moon".
(Shallow waters introduce a lag of a day or three, due to friction and resonance effects.)

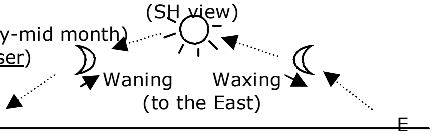

Make it your aim to see these things all happen in phase with the Moon month.

The Clock-Face Method for Direction

Quicker than 360° — it only counts up to 12! — it communicates relative directions easily.
E.g. "3 o'clock high" = "on our right — up in the sky" "9:30 from Rigel" = to the left and just above (that star)
Or, for horizontal directions, " 5 o'clock from the Sun " = looking down-Sun, and slightly to the left "six o'clock true" = South
At 30° per "hour" it is easily compatible with 360° — both are Babylonian systems.
(But don't confuse it with a 24-hour clock face, which is only 15° per hour. Unfortunately our clocks do not use a 24-hour system, which would make them nicely travel around once per day — as fast as the Sun.)
For shifting between clock face directions and 360° directions…
— First convert to or from the hour: Multiply hours by 30, for degrees. Use 12hrs=0 : 0 30 60 **90**; 120 150 **180**; 210 240 **270**; 300 330 **360**
— "5° extra" = "add 10 minutes" e.g. 255° = 0830 hrs (240° = eight hours; 15° = +10+10+10)
— "1° extra" = "add 2 minutes" e.g. 258° = 0836 hrs
When using the clock-face for compass directions, for NE, SE, SW, NW use 1:30, 4:30, 7:30, 10:30.
Translate 1, 2 o'clock as "(roughly) NNE, ENE," etc.
Don't use the minute marks on the clock-face because every 6° is not very user-friendly

Latitude without Numbers

When all you want is the latitude, the shape of the sky can tell you, without numbers, without calculations, without a watch.
E.g. 'Canopus overhead', or 'the Pot sets at 30° off vertical'.

If all you want is to recognise appearances locally, you don't need to set your Star watch to Greenwich Time, or to know a star's longitude relative to Greenwich.
E.g. 'I set my watch to the Dog Star passing North, from home last month'.

If you don't want calculators and trigonometry, you don't need to use them.
E.g. 'The sky is behaving the same as last year, we must be within sight.'

I am about to give you the star time and star longitude information, only for those who wish to work with the easiest benefits from them.
E.g. 'That must mean that our longitude is 150° E.'

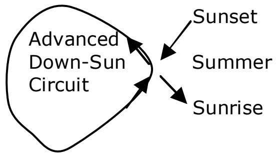

This Assumes Southern Hemisphere

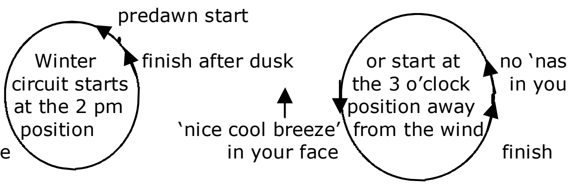
predawn start
Winter circuit starts at the 2 pm position
finish after dusk
'nice cool breeze' in your face
or start at the 3 o'clock position away from the wind
no 'nasty cold wind' in your face
finish

Note that clockwise circuits and half-day walks must alternate between up-Sun and down-Sun.

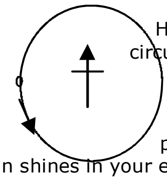

How **not** to circumnavigate a lake — from the 9 p.m. position
The Sun shines in your eyes all day!

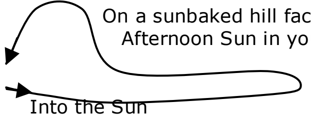

How **not** to climb a hill — at the end of the day
On a sunbaked hill face
Afternoon Sun in your eyes
Into the Sun

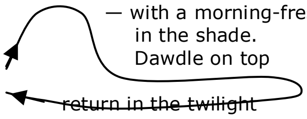

How to climb a hill
— with a morning-fresh climb
in the shade.
Dawdle on top
return in the twilight

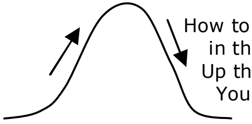

How to climb a mountain
in the outback heat —
Up the west face; long lunch on top
You walk always in the shade.

Matching Sky to Sky — Move Only Your Horizon

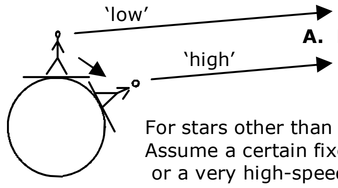

A. Move **towards** a star and it will **climb**
e.g. move toward the pole star
to a higher latitude
and it is higher
For stars other than the pole star
Assume a certain fixed star time
or a very high-speed—to stop the stars moving.

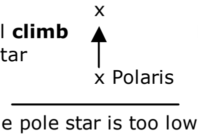

B. Move **sideways** and your horizon **tilts** with you
e.g. a star pair
may not be upright
as expected —
move sideways.
Or perhaps
a star pair
sets before it becomes upright
You need to tilt your horizon
down to the right
by moving to your right..

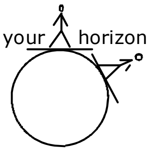

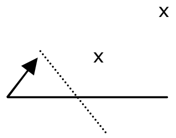

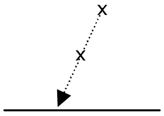

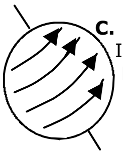

C. Because the Earth spins to the **East**, it leaves the stars behind in the West.
If you move west with the star it will slow down, so…
Move **E**ast and the stars will *get there* ***E****arlier*
e.g. 2 stars become upright too late?
You need to move easterly.

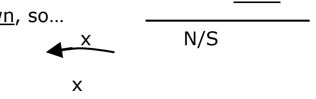

You may need a combination of A., B. and C.

Prevailing Winds on Creek Banks

One of the two Creek banks will often have a mud- or rock- cliff due to prevailing wave-action, with the other bank being a mud flat.
E.g. If prevailing winds are westerly, the west bank will show less erosion.
For more on finding the prevailing wind direction when the wind is not blowing, see section 3: "Nature's Norths".
Because prevailing wind is almost continental in scale, navigating by it is a sort of Big Picture technique.

Tips for calm weather

1. Listen for the wind! Sounds will carry better downwind e.g. surf or traffic. Remember this, at night when all else is still and quiet.
2. Watch the angle of mist droplets falling. (Stop still for this one) Or smoke rising.
3. Huff into the air to make a fog.
4. Watch the flimsiest vegetation.
5. Look at the leaves which dangle down from the topmost leaf clusters of a tree.

Mnemonic: "North; Coast"

That's what you should say, whenever you say 'North'.
"North (Coast) is over there" — it reminds you to think,
or rather to picture, the invisible picture.
You should make it your business to name the North-Coast town north of you,
before you go, by consulting a roadmap
Then of course, there is the East Coast, and the South Coast and the West Coast.
Next, of course, the same must go for "North-West"
— think of the North-West Coast township, and so on.
You are not functioning properly unless you mean something when you talk!

Go Overseas!

This is a one-off effort, so go to the trouble of getting a globe,
to sort out what is truly "East" and "West" of you,
in a Great Circle straight line.
It will surprise you what East and West should mean to you.

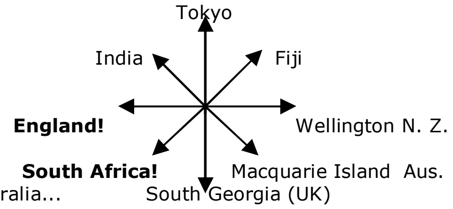

e.g. from Australia...

Mnemonic "The Sun Travels the Tropics"

I.e. Imagine the **tropical countries and oceans** which are **'underneath' the Sun** as it moves East to West during the day — The Equatorial band.
For Australia that means from Central America Across the Pacific Ocean Past New Guinea
On to the Indian Ocean And on toward Africa
When we in Australia see the Sun rise, it is past noon in the Americas. At our sunset, it is approaching noon in Africa.
See the Reference List of Tropical Places (for The Eastern Australian Time Zone) opposite. As a help to global orientation, make it your aim to memorise these time zones in sequence. Check it out with an atlas. See the tear-out clockface diagram in the appendix.
The same sequence holds for any planet, or star near the Equator e.g. for the stars listed 'opposite the Sun' (below), but with different timing.
You can also imagine which country the Sun is overhead while it is night-time near you,
or which country the star opposite the Sun is overhead while it is daytime near you.
Now, with this in mind name the general North/South directions, according to the height of the Sun: To the Sun = "Hot"; To noon = "Hottest"; and the Shadow direction = "Cold"; Shortest shadow = "Coldest".
•Although the Moon, by comparison with the Sun, has a "cold" feel to it, make sure you apply the same list of hot tropical countries to it also.
When you see any Zodiacal body, you should think: "That is a Hot direction; It is over a Hot country." and try to name the country.

Tell the Time of Night by the Star **Opposite the Sun**

Such a star will be **up all night** — whether the night is long in winter, or short in summer — look for it. Write them on your calendar.
This yields the solar time, i.e. with respect to local noontime, but not the daylight-savings or summer-time, so adjust by one hour in summer.

January	Sirius *Castor, Pollux*, Procyon	
February	Suhail, Miaplacidus Alphard *Regulus*	This list is arranged time-wise
March	Pointers (Dubhe-Merak) *Denebola* Phecda Southern Cross	the earlier stars are correct
April	Southern Cross Alioth *Spica* Arcturus, Pointers (R.K./Hadar)	at the beginning of the month,
May	*Zubenelgenubi*, Kochab Alphecca *Antares*	the last stars, near the end.
June	*Antares* Atria Rasalhague Scorpion Sting Kaus Australis Vega	Use each as a 24 hour-hand around the South Pole Star
July	Vega *Nunki* Altair (Aquila) Peacock Deneb	The Zodiacal stars
August	Deneb Enif (ε Pegasus) Alnair	(the ones on the Ecliptic)
September	Fomalhaut Middle of Great Square	are in italics.
October	Diphda Mirach Achernar Hamal-Almach	Only they are truly opposite the Sun.
November	Mirfak	This chart is **true for**
December	*Alderbaran* Capella-Rigel Saucepan Six O'clock Stars; Canopus	**both Hemispheres, permanently**.

Two Different Back-and-Forth-Star-Directions Give You a Position Fix

..Yes, it's just like two bearings (which are not opposite each other) from two landmarks. But because these 'landmarks' are so distant, the accuracy is only to within seeing distance. For better accuracy you would need a theodolite or sextant.
E.g. when two stars look upright, that locates a global star-direction; a second upright pair gives you the second direction, and a position fix.
To fix your latitude. you have to **measure the time-lapse between the two star pairs**. (Simultaneous is simplest.) So you need a plumb-line and also a watch. And for accurate (Greenwich) longitude fixing, you need to work in (Greenwich) **star-time** — otherwise you have only latitude which is fixed. So you need either an almanac, or calculator, or star-watch, or a work-it-out-at-home-later approach.
This is equivalent to fixing which star is exactly overhead and when. To see this, look up past the top of the plumb-line to 'overhead' from the two different directions—one for each pair—to see that they must cross exactly overhead.
The 'fix' is repeated accurately over years, and every night at the same star-time, with some very small predictable drift in location and timing.
No one else on earth can see exactly what you see right now. That's what gives you a unique position fix.
People at the same latitude see something very similar earlier or later than you do.
The best two directions to cross are Northwest/Southeast and Northeast/Southwest. Top-view...
This is because East/West star directions are time-wastingly difficult to determine.
This method is most helpful, of course, at sea, or in the desert, or as a hobby, or when lost.

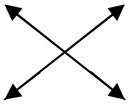

Predict the Full Moon Path According to the Time of the Year

The full Moon is **opposite the Sun**, and is near the **Ecliptic**, i.e. in the Zodiac (the Sun's path through the stars) so that…
When the Sun is in midsummer position—high, and up for a long time—the full Moon will be in a midwinter position—low, and setting again quickly
When the Sun is in the midwinter position—short days—the full Moon will be high and up for a long time — long nights.
The two complement each other, since when one comes up the other goes down, and when one goes down the other comes up.
Both Sun and Moon are on the Ecliptic—not usually on the Equator—but, at the Equinoxes, the Sun is on the Equator, so the full Moon is too.
Near Equinox, of course, the Sun and Full Moon will rise and set in almost the same positions — East and West — and follow similar paths — about 12 hours each, along the Equator. The Equator and the Ecliptic intersect in two opposite points — the equinoctial points.
In other words, you can **use the full Moon to visualise the Ecliptic**, and to guess where the Sun is, i.e. to guess the time of night,
Or you can use the Sun to visualise the Ecliptic, and to guess where the full Moon will be, how high, and for how long.

Finding North from the Moon, by its height & phase

The Moon will follow a Midsummer Sun's path, or a Midwinter Sun's path, or an average height path, according to where on the Ecliptic it lies. Because the phase of the Moon is determined by its relative position to the Sun, the combination of the time of day, the phase, and cusp-Ecliptic gives you an unmistakable sense of where North should be, when you see the Moon. When you see the Moon surprisingly low in the sky, and the cusps confirm that the Ecliptic lies fairly horizontal, you know that it is somewhere near North (SH), following a low and horizontal trajectory, rather than a Moon 'just now rising' on a high trajectory, like a summer Sun in the East.

One Degree is about One in Sixty

1 in 60

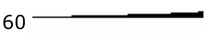

And **n° is about n in 60**. e.g. **10° is 1 in 6** (p96a), and **6° is 1 in 10** (3 fingers wide, or 3½ knuckle-centres, at arm's length, or 'blink your eye').
These are handy rules of thumb for navigators. If you use an isosceles triangle, **the rule is remarkably accurate for most acute angles**.
Example: "Oh I forgot magnetic variation! I've been **5° off**!! Oh well, I will simply adjust sideways. Isosceles

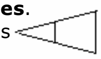

I have come 12 km; 1/60 of that = 0.2; times five = **1 km off**"
Actually 1° is 1 in 57.3 (1.75:100), predictably slightly more than 1 in 60, but as the angle increases so does the accuracy of the 1:60 rule.
The rule (n° = n across at 60 away) is based on: the third side of the **isosceles triangle** (is twice the sine of half the angle).
(If you don't like numbers, try this other "handy" rule...) 6 (equilateral) lots of 60°

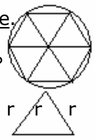

Rule of Finger: point your index finger at something — the fingernail is 1°.
Why is 1° slightly more than one in 60?... 60°

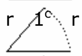

Six equilateral triangles go all around the circle, but, as you know, 6 1/4 radii go around the circle.
This is because the circumference, longer than the hexagon, is 2pi times the radius, or 6.283r one radian=57.3°

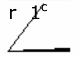

So one 'radian' of arc, which is one radius long, but curved, will be slightly less than 60° i.e. 57.3°.
So if the radius is 60, one sixtieth of a radian will be "one in 60", and slightly less than 1°. 1/60 of 1^c = 1 in 60

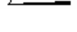

377 of them go around the circle. For 360 degrees around the circumference, the radius needs to be **57.3**.
The equivalent is '5/3 across at 100' (5 at 300, 167 at 1000; above 10°) with the more exact figure for up to 10° = 7/4 in 100

Tropical Countries and Places Under the Sun — Arranged in Time Zones Relative to Eastern Australian Time — 10 hrs ahead of Greenwich

-12hrs.	The mid North Atlantic (is opposite Australia)	0	Rockhampton, Tropic of Capricorn 23½° S. (Canberra 36° S.)
-11hrs.	Brasilia 16° Sao Paulo/Río de Janeiro 23½° S.	1	Alice Springs, Tropic of Capricorn
-10 hrs.	Trinidad and Tobago 10° N.	2	Sulawesi 0°; Manila 15° N.; Port Hedland 20° S.
-9	Cuba/Haiti 20° N.; Bogotá 5° N.	3	Christmas Island 10° S.; Jakarta 7° S.; Singapore 1°N.
—	Panama 8½ hrs ahead of Australia	—	Cocos Islands
-8	Galápagos 0°; Guatemala 15° N.	4	Calcutta Tropic of Cancer 23½° N.
-7	Gomez 26° S.; Mexico 23½° N.	5	Maldives Equator
-6	Pitcairn Island — Easter Island 25 to 27° S. (L.A. 34°N).	6	Gulf of Oman Tropic of Cancer
—	Tip of Californian Peninsula 23½°N; IDR Gigego 20°N	—	Seychelles
-5	Mangareva Tropic of Capricorn	7	Port of Aden 12° N.; Mogadishu 3°; Madagascar 23½° S.
-4	Tahiti 17°½ S.	8	The north of South Africa on the Tropic; Lake Victoria 0°;
—	Cook Islands 11° S.		Sudan/Egypt on the Tropic; Aswan High Dam
-3	Hawaii Island, Tropic of Cancer	9	Namibian Desert coast on the Tropic; Lake Chad 14° N.
-2	Fiji 15° S.	10	Timbuktu 17° N.; Accra 5° N.
-1	Noumea 21° S. See pp 132f	11	West Africa coastal bulge 15° N.

Tell North from The Time By Using a Protractor

Reverse the method for telling the time from the compass protractor and North. To make the zero point North...
Make the sundial (protractor) show the correct time (one side of 'noon' time). Simultaneously make sure the sundial face is tilted to be parallel to the Equatorial Plane… I.e. For best results the tilt of the protractor according to the latitude, needs to be accurate.
Consequently, you may find it an interesting exercise to mockup a large (bigger = better) cardboard multi-protractor
— e.g. by marking the sides of a cardboard carton.
Use a plumb-bob to **accurately tilt the protractor, according to your latitude**. E.g. pin the plumb-line to the top, side corner, and mark 90° on the flat face…
This instrument may be used to find North if you know the local time, or to find the local time if you know North.
One side of the carton is used for the latitude protractor in conjunction with the plumb-bob, a second side can be marked as a sundial protractor.
E.g. use another (long, straight, perpendicular) pin at the top corner of the side facing you, to cast a shadow across a 90 degree scale (— this is for accuracy, rather than make a full 180° or 360°).
Morning and afternoon then require separate scales; Winter-time requires that you face away from the Sun.

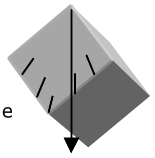

For Land-Navigation, Choose the Plumb-Line Star-Pair Method

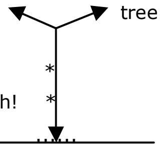

- Use a smooth constant diameter line, preferably thick. Make the plumb-line tall and stand far back. Make the cord either white or black to stand out against the sky, then maybe flash a torch onto it.
- Pour sand around the plumb, to anchor it to the ground. Avoid wind, especially if you hang the line from a tree branch!
- " Blink " the stars "on" and "off" behind the plumb-line, both on its left and on its right, by moving your head.
- Average the "start" & "finish" times. I.e. "It seems to line up now for the first time"… "definitely finished now". One accurate set of measurements is better than several rough ones. One star will gradually move over the top of the other star, and you have to pick exactly when.
- Choose well-separated stars (e.g. Rigel & Capella)
- If you want the timing to give you your longitude immediately, choose a North/South pair (e.g. Rigel & Capella) — (you need to know the star's longitude or its timing elsewhere, i.e. from a known longitude).
- An East/West pair will give you your latitude (without the need for a watch). The observation here is not the time but **whether or not** the given pair will attain to uprightness, after the first glimpse or before the last glimpse. Beyond a certain latitude, a given star pair will always remain tilted, less than vertical. This latter process can be time-consuming. Use a hand-held plumb-line occasionally, to check the progress.

Predict the Moon Phases by " N minus M "

" N " is a number (you calculate) for one particular year (e.g. For 2004 it was "24", in Australia's time zone, for the New Moon)
" M " is the month number (e.g. April is "4")
" N – M " is the day of the month for a certain fixed phase (e.g. in April 2004: 24 – 4 = 20th April 2004 for the New Moon)
April 20 was dark all night. New Moon seems to show the most regularity. The 'new Moon' is more accurately 'no Moon'.
Other phases happen one week later, two weeks later, three weeks later — 'quarter' by quarter — 29½ days for the whole cycle.
(The number, N, in 2004, would be 2; 9; 17; 24, 32 approximately, for 1st; 2nd; 3rd; final and 1st quarters respectively.) You can add 29½ to N.
Then in 2005, the number, N, for the new Moon was 13... (i.e. in April 2005, 13 minus four equals… the ninth of April — no Moon).
And in 2006 I used N = 32; 21 in 2007. The number N for the New Moon must be close to any **New Moon date + M**, e.g. December New Moon+12
Check your diary's calendar each year for the Moon phases; test and tweak a number N which works to within plus or minus one or two days.
The prediction is fairly crude, but with trial and error you can come up with *the best number and phase combination* for that year. For 2010, it is 18.
—usually **11 days less than the year before** (12 for leap year) because the lunar year of 354 days is 11 days shorter than the ordinary solar year
I.e. a good mnemonic is: **the Moon gains one day a month for every month except February** (which has only 28 days).
Hints: *Subtract " M " twice for January and February,* for more accuracy on non-leap years. Alternatively, to cope better with leap years...
Start the year in March (after the short month, February). Call January & February "months 13 & 14", but subtract 1 more for their accuracy.
Note: The "Full Moon" is when the Moon is nearest to "opposite the Sun" and it may not be exactly on the horizon when the Sun is.

Walk across Australia!

(Substitute whatever continent is appropriate)

Mentally line up a wall-poster size map of the whole continent (or bigger) with the track.
When you approach a bend, place the map on the bend, then walk all over it!
It's hard to miss the meaning of the bend if you bother to do this.
Your mind may not believe that it is not still going 'straight ahead' until you force it to catch up with the facts!

English!

"East-around" from North, is West-around from South, in a horizontal clock-sense. An "Easterly" course means East-wards — moving **to** the East.
An "Easterly" motion means East-wards. E.g. an "Easterly spin"-motion = "orbiting in an Easterly sense" = moving E around the N/S spin-axis **But**…
An "East(erly)" **wind** means flowing Westwards, **from** the East. An "Easterly" **current** might be flowing Westwards, from the East, but an easterly tidal "set" is an offset to the East. (Set is how far the tide "sets" you off track. The speed is called the tidal "drift".)
" Due " East means "exactly" East, as when a loan is "due" — "right now!" **Always double-check what someone means**!

North/South and Noontime by **Bisection (of Rise and Set)**

Rise and set are **symmetrical around North/South**, provided you use horizontal, and not the horizon.
In hilly country, instead of 'rise and set', you should use **equal height** of the Sun (e.g. in NE and NW).
"Equal height" translates as *equal length shadows* of an *upright*, *straight* stick or cord, onto *flat* ground…………
Or, better, use a bush sextant, p.27c, set to a fixed reading.
There are four ways to bisect two directions to find North/South — so don't muck it up by 90° or 180°!
The second diagram will work in both hemispheres — but double check the results for sense.
There are also two ways to **bisect the time** — so don't mix up 'just before midnight' with 'just after noon'.
It is also hard to 'bisect the time' to find noon, which of course, is not at 12 noon exactly, nor at 1300 hrs exactly in summer.
To work with hours, you need to **carry 60 minutes**, not 10 or 100! So be careful and crosscheck your figures.
Be accurate or your results will not be worth the effort.
E.g. rise: 0529 hrs; set: 1820 hrs Add… = 2349 hrs Halve… = 1100+(**60** +49)/2 = 1154½ hrs = noon.
Check: total elapsed time: 1820 –0529 = 1251 = 2x0625½; 0529 + 0625½ = 1154½ = correct
1820 – 0625½ = 1154½ = correct

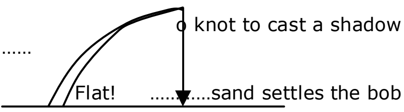

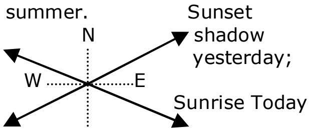

Longitude and Latitude from the Sun

Longitude...
Try timing the Sun for **equal height** each side of noon e.g. NE & NW (SH) to calculate the time of noon.
This can tell you your Greenwich longitude, provided you know the Sun's longitude, that is, its G. H. A.— from an almanac or calculator.
The Sun shifts predictably at a rate of near 15°/hour westerly. Almanacs make it their business to predict its West longitude exactly.
Of course the Sun's longitude at the time of your noon is your longitude too — since it is north of you.

Now try timing the Sun in two **opposite directions** at the same heights i.e. East and West.
You could try using the shadow of a plumb-bob on flat and level ground. Mark the shadow carefully.
This will be most accurate in the season when the Sun reaches high — in the summer sky.
By a series of early observations you can guess at when to make a final observation.
Using a theodolite or bubble sextant allows you to graph the rise & fall of the Sun (or a star) and see when 'same height again' happens.
See the 'Bush Theodolite' for an alternative measuring method.
Latitude (There are easier ways!)...
You can now solve the E/W formulae for your latitude — e.g. try the formula that predicts the timing of East and West for the Sun.

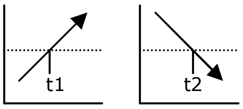

Equal-Height Stars Can Be Useful When the Sea Horizon is Visible

(or with a bubble sextant)

For direction, use the mid-point, of two stars which appear to be of equal height. — a great circle direction
For longitude, the timing is the important bit. — these stars should be approximately north or south of you
For latitude, the test is whether or not the stars do attain to equal height. — these stars will be approximately East or West of you
For a position fix, try to get two pairs, one north-west, the other north-east of you, for instance.
A **bush quadrant** is easy to make with a string tied to a stick! — just pull the string to length to measure or compare star heights.
Rest your hand and chin on something solid if possible. Use your full arm's length for best accuracy.

You can fix your latitude by measuring the **timing** of **two** separate star-pair events
— the relative timing — e.g. '5¼ minutes apart' (is unique to that latitude).
Simultaneous events are simplest, but separated events are easier for one person to time.
You don't 'get on the map' with this simple method unless you have been there before.
Unless you want a lot of trigonometry no calculations are required,
only the observation of re-recognisable appearances.
You can fix your longitude, too, by measuring any absolute timing — an exact star time — and so get on the map.

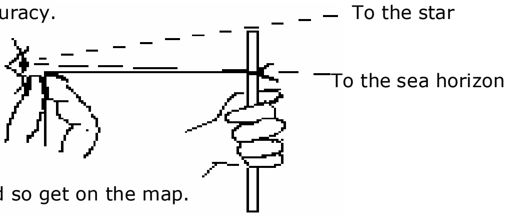

Anticlockwise, to the East, is Your Best Spin-Reference Direction

Because we use our North-South axis as our reference direction, we have to learn the *spin-sense* as something *separate*. See 'Orientation' p 35a.
The standard viewpoint for any spiral system (e.g. a weather system, a clock-face) is 'from the top', "**as seen from above**". Horizontal directions are measured "clockwise around" as you look *down* on them. For our spiral galaxy & solar system, the correct viewpoint is therefore 'from the North'
Looked at 'from the top', i.e. from the North, our Galaxy spins anti-clockwise. Or more simply, 'to the East'. 'Up North' is 'on the left' of East.
This means that the real motions are East-around the North/South spin axis—opposite to the familiar idea that everything sets in the West.
Because Northerners have their back to the North Pole when they see the Sun Down South of them, they see their horizon spin anti-clockwise — that is, the hills and trees on the horizon move to the left of the Sun. They are looking "from the top" i.e. from the North further South.
Their Sun only appears to move clockwise (to their right, to the West), because it is their horizon moving—anti-clockwise (to their left, to the East).
To avoid confusion, always make the distinction between real mass movement and apparent virtual rotation. And think: We spin East.
Because we think of 'clockwise' and 'right-hand' motion as 'standard', and setting "in the West" as 'normal', it's a bit of a shock, to rethink!

When you are Down Under, on the South side, looking North at the noon Sun, you see the bottom of our Galaxy, and Solar System, and Earth.
The three 'North' spin-axes are all very roughly aligned together, all lying in our "Northern" Hemisphere — not too much tilted from each other or from our own spin-axis. You mentally lift yourself further out (not "up") into space to look at (but not really 'down' on) the spiral motions
Remember that from a 'standard' (Northerner's) point of view, the Southerner's view is really "a bottom view looking upwards" — from the South half, and you have moved out further South, for a better view — with the South Polestar behind you. You should see clockwise real movements.
Turn around and look Southwards at the Southern sky and you will see as if 'from the top'—or at least 'from within', but looking South. Anticlockwise
Remember to rethink "Up" as "to the *North* Pole of the spin".

- It should now be easy for you to visualise, in either hemisphere, your Southern horizon spinning anticlockwise.
- Look also at the Moon—visualise its orbit and its spin—'anticlockwise' from N.H., but clockwise-around if you are seeing it from e.g. Australia.
- Find the bright planets — can you 'see' them moving 'anticlockwise around the Sun' in orbit? The Ecliptic marks their path.
- And, go further, visualise the real Sun spinning much like the Earth does, on its own N-S axis. The Sun is North of you in the Southern hemisphere, spinning on its own axis 'clockwise from below', or in the Northern hemisphere, South of you 'anti-clockwise from above'. Use the Ecliptic's N-S spin axis as a good guide for the Sun's, or simply use the Earth's own axis as a rough guide — Don't look at it, but imagine its sunspots moving.
- Keep going: The Earth hurtles through space in orbit around the Sun 'to the left, while looking to the Sun, around a N-S axis of spin' (Down Under, looking North) 'to the right, relative to the Sun' (NH) '**anticlockwise**, looking South' (NH)—'with' the apparent movement of the Sun, in both cases. That makes the noon horizon move in opposition to the movement of the whole Earth through space. Like the slow 'inside' of a spinning ball circling.
- Imagine the whole Earth-orbit in the sky. It's easy when you get the hang of it. New Zodiac stars seem to emerge (invisibly) on the Sun's West side, in seasonal procession, because "the Sun is edging Eastward along the Ecliptic in a yearly cycle" i.e. we move to the West side of the Sun and see behind it to the background Zodiac stars. Don't look at the Sun, but you should be able to 'see' this happening as if there were a star chart behind the blue sky! Try just before dawn.
- Now, at night-time, find the centre (thickest part) of the Milky Way near Scorpio/Sagittarius and visualise the whole Solar System orbiting it perpendicular to that, the Sun/Earth/Moon all moving as a unit, to your left (SH) or right (NH) towards Suhail, away from Deneb.

The Milky Way marks the plane of the spiralling galaxy (from which you can determine its N-S axis, tilted differently from ours).

Finding Magnetic Variation by Bisecting Rise and Set

Accuracy is needed.

Measure the compass directions of the morning and evening Sun at equal heights e.g. rise & set on a flat Horizon
Bisect the result. E.g. Rise: 079° Compass Set: 271° Compass
Add: 350° Halve: 175° Use the Opposite: +200–20=355° (See p.18e)
So the Compass North must be 5° to the East of geographic North. **Draw diagrams to avoid getting confused.**
Check all numbers! E.g. 355°–271°=084° 355° to 360°=5° + 079° more=84° Correct.

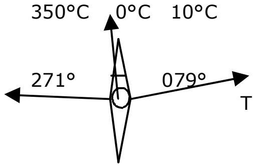

Or 271 – 79 = 192° spanning = 2x96°
79+96 = 175°
= opposite to (375 – 20)°
Correct.
The compass readings are "Least", by 5° (see p 107b)
E.g. 079°C is really 084°T. (see p 14e)

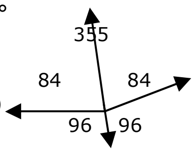

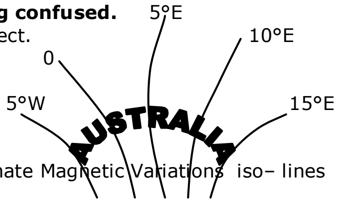

Approximate Magnetic Variations iso– lines

A Bush Theodolite for Overnight Equal Height Observations

This is for accuracy in calculating... True **North**, Noon **time** and Compass **variation**
And for any equal-height paired observations — Sun/stars/etc.
Find a large bent stick. (Or bend it with a bowstring)
Hold it up by a loop-knot e.g. hang it on a twig.
(If it is early or late i.e. a long shadow, the stick will not need to be bent much.)
Late afternoon, at camp,
Mark where the shadow of the knot (or the other end of the stick) falls on the stick;
take the time; line up the landmark it points to; and find its compass direction.
Early morning, next day,
Hold it up (in reverse); wait until the shadow again falls in its proper place;
get the time, direction and compass reading again.
For a **Bush Bubble Sextant**, arrange a water filled clear tube on a triangle of lashed walking sticks so that there is a trapped bubble in a tube section with an ever so slight convex-upwards bend.
Make the bubble always stay at one marked place when you are using it.

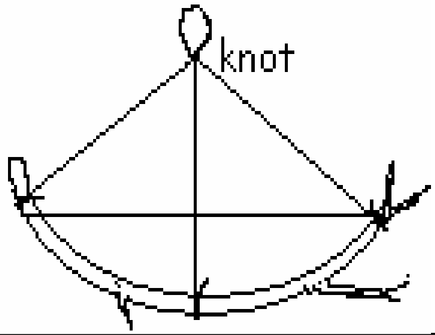

Great-Circle Reference Star Directions

You need some familiar reference points. It's a simple matter of going out at night, waiting around, and tracking down a few of these...
Look for **stars which are closer to the pole than you are**, and see where they **rise and fall vertically**. It happens once or twice a night, per star.
This gives you a series of great-circle star directions, already labeled, because **the 'maximum latitude' is equal to that star's 'declination'**.
E.g. Acrux is about 30° from the pole, or "60° latitude". If you steered towards one of its vertical-motion directions you would end up underneath it.
So as it rises or falls vertically (once each per night, if visible at the time), as it circles the South Pole Star, it lures you to 60°S latitude.
E.g. A star overhead, as the extreme limiting case, is moving East to West, rising vertically from the East and falling vertically to the West, meaning that you are *already at* the maximum latitude of that course, if you had been heading towards that star guide, now travelling E-W also.
You then have to head directly *away from* it (e.g. the next night) to further pursue that course.
Heading away from that star either East or West would take you eventually back towards the Equator, like the way you travelled down there.
The **stars further from the pole, than you are from the Equator,** make a complementary set of reference directions, by using the horizontal direction 'at **right-angles** to where stars **rise and set**'. **The declination is a measure of *how close to the pole* you would pass**.
The stars close to the pole are unavailable. The co-declination labels 'the maximum latitude'.
E.g. Sirius is 17° South, so rising from the horizon, somewhere to the South of East, it would send you to within 17° to the left of the South Pole.
Turn around and it would guide you (part way) to cross the Equator and to within 17° of the left of the North Pole.

Make a Bush Psychrometer ..for absolute humidity and dew-point-temperature measurements

Carry a wettable piece of material — dip it in water — hang it up to dry (in the shade out of the breeze) — time how long it takes.
Measure the air temperature at the same time, and cross correlate the two, to find the absolute humidity. If you calibrate the method at home, by graphing the results for a whole range of different temperature and humidity conditions, you will get to recognise a dry air mass from a moist one.
The piece of material needs to be non-porous, non-rotting and dark coloured — to show the drying process clearly, and to yield repeatable results.
Night-time minimum temperature more or less bottoms out as soon as an air mass starts to dew. The ground then wets down further instead of cooling down further. The normal method of measuring absolute air humidity—the percentage of water vapour—is by evaporatively cooling a water drop in it. It requires wetting a thermometer bulb, fanning it vigorously (trying hard not to break it) and measuring how far you can make the "wet bulb temperature" "dip" below the "dry bulb (air-)temperature". The water temperature won't get all the way down to "dew point"—the temperature at which dew or frost will form in that air mass—unless it already is dripping wet outside. It may get a bit more than half-way, and then only when only dew is expected—DP above 3° C. **If frost is possible,** i.e. when the dew point temperature is 3° or less, **it will dip at most half way down to the dew point**. Why 3°? Well, when the air is very dry, dewing will further reduce the water vapour content of still, cold air, from a dew-point temperature of a few degrees above zero, down to 0°—resulting in 'frozen dew'. **Rule of thumb** for calculating dew-point from dip:
When the DP is 40°C/30°/20°/10°/5°/0° then add to the 'dip' 0/1/2/3/4/5 quarters of the dip, to get down to the real dew point.
If the night is long enough and calm enough to cool down enough, & the sky is clear enough, a DP near zero will result in a frost.
E.g. 15° to 10° = 5° dip, 5°DP. 15° to 13° = 2°dip, 11°DP. 19° to 18° = 1°dip, 17°½DP. (Remember that you can't go far wrong anyway.)

South is a Triply Better Reference Direction than North, in the South

1. The closest geographic Pole in the Southern hemisphere is just over the horizon in Antarctica. Globally seen, all directions pointing South are convergent, whereas directions pointing North are initially divergent, and lengthy and less meaningful... Travelling North you first cross the Equator before it gets colder — so the gut-feel meaning is less clear. The meaningful direction, when you are in South latitudes, is to go further South to get much colder. (It is the same as in the Northern hemisphere, where going further North makes it much colder.)
You may find that in a place like Tasmania for instance, the maps seem to be upside down, because although North is at the top of the map, the logical reference direction is toward the bottom of the map, where the cold place gets colder. You don't think: "This is a warm place. In which direction does it get warmer?".
As a result it is easy to get disoriented by 180°, if you insist on using North as your reference direction. In addition…
2. When you put South at the top, and look South, then the Sun will move to your right, as if seen from the N.H., across the map as time passes, for both hemispheres—and that will make the clock-to-Sun link obvious (like putting a Southern clock on the S. wall to match the Sun's motion).
Northerners especially, would find that a more natural way to view any map—as if from further North, looking South, to the Sun moving right.
3. Putting South at the top, with the Sun & Tropics behind you, looking down onto a map, is the mathematically standard viewpoint
— like looking from the top of the Galaxy — from North to further South — the land South of you moves Anti-clockwise, to your left, to the East.
You can still orient your map normally and correctly, but there is no law against habitually viewing it from its Northern edge, looking South. Try it.

Find East/West By Delayed Bisection of the Sun and the Full Moon At Rise or Set

To find due East (+/–10°): **Bisect the sunrise and the full Moon rise** — since they are of opposite declination.

For due West (+/-10°): **Bisect the sunset and the full Moon set** — since they set each side of due West.

Obviously, you need to wait about 12 hours for this method — it is really just for your background comfort level about where things are.
There is a slop of 5°, since the Moon's plane of orbit is about 5° away from the Ecliptic.
There is also a slop of about 24 hours, since the full Moon occurs at some time other than Moonrise or Moonset,
plus it is hard to pick an exactly full Moon, and the Moon varies in declination, i.e. North/South position, from day to day.

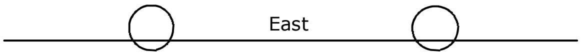

Find North by your (non-digital) Watch-Hands

Point the noon on the watch (**not the 12!**) to the Sun.
Tilt the watch axis toward Polaris as best you can judge.
This tilt matters a lot, but is not usually mentioned!

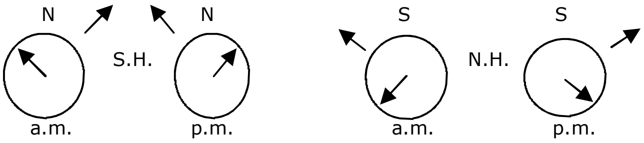

Southern hemisphere: **North is in between the Sun and the hour hand**.
Northern hemisphere: The Sun is in between the hour hand and South.
Why?: Clocks Go Clockwise; at 30° per hour — twice around per day.
For the Southern Hemisphere that means: — twice as fast as the Sun and opposite.
So 'in between' moves neither clockwise nor anticlockwise — it stays N/S. At real noon-time, the hour hand and the Sun cross over.
For the Northern Hemisphere, the hour hand overtakes the Sun at noontime, at South.
It's easy to visualise in fast motion, once you catch the idea — it's like the hare overtaking the tortoise.
Remember to tilt the watch-face up to be parallel to the Sun's orbit, not horizontal, because a sundial shadow is not linear.
You could use the middle of sunrise/sunset times to find the local actual high-noon midday time, which is needed for "point noon at the Sun".
The same method can be applied to the stars, the Moon or the planets — provided you know when the one you want to use is north of you.

Leo the Lion Looks like a Lion or Sphinx Squatting
It Has Four Lines which Show N./S. E./W.

Two in the 'hindquarters'; one along the 'back'; one as the stem of the 'question mark' (Leo's mane).
(You have to turn the '?' around.) For the Southern Hemisphere, turn it all upside down.

This is a good sign in Northern spring/Southern autumn.

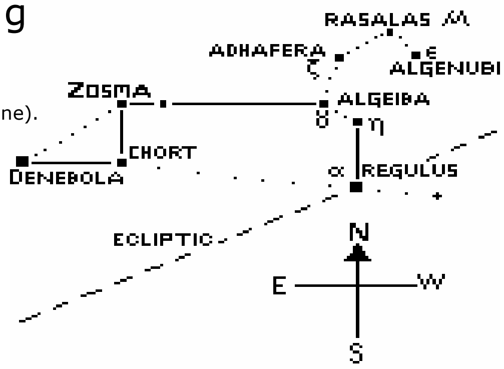

(NH view — LEO is upright)
Direction of movement ⟶

To Predict the Moon's Path, Visualise the Ecliptic

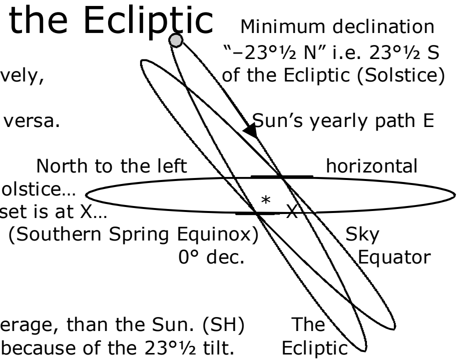

For more clues on how to visualise the Ecliptic, first see the overview at the end of this section.
The Moon stays within 5° of the Ecliptic. That helps you see where the Ecliptic should be, or alternatively,
where the Moon should be. The Full Moon is easy, and has been covered on p. 24d.
A **half-Moon,** for example, is 90° from the Sun; so if one is on the Equator the other is not and vice versa.
E.g. a half Moon in mid-winter will rise and set due E./W. +/- 5° or so.
Secondly, recall that the Earth's spin-axis is tilted 23½ degrees from the solar system.
This is how it would look from Sydney (*), at midday, in Midsummer, i.e. at Southern Summer Solstice…
The two great circles intersect at 23°½. The whole double-circle rotates West over 24 hours. i.e. sunset is at X…
To visualise the Ecliptic (see also p. 30d & 33b), from the Sun...
..Since the Sun moves eastwards seasonally through the stars you can know that
the part of the Ecliptic which the Sun was recently on must 'lead' the Sun to the West.
E.g. after solstice in Southern Summer, the most Southerly part of the Ecliptic is ahead of the Sun
to the West, and so a waning crescent Moon (to the West of the Sun) will be higher in the sky, on average, than the Sun. (SH)
E.g. A crescent Moon at sunset at spring equinox will look extraordinarily high over the sunset glow, because of the 23°½ tilt.

Bearings & Tracks: "If it is Right, it is Big, in my sight"

= Anything on my **Right is Bigger than 'ahead'**. I.e. If I am heading 100°, something on the right has a bearing bigger than 100°, e.g. 200°.
And "Any track on my **Left is too Little to look at**" = Anything on my **Left is Less** than ahead, e.g. 20°.
These mnemonics compare fixed tracks or directions, and their 'Big and Little' measure in degrees, East-around-from-North i.e. clockwise bearings.

[The changeover from 360° to 0° can give you trouble. Where 'ahead' was say 355°, 15° bigger should be 'big' e.g. 370° but will look really small, 10°. Because it is nowhere near 355 in number, there should be no possible confusion, provided you think of 10° as 370°.
'Little' may look really big, if 360° gets in between. E.g. where 'ahead' was 005°, 15° smaller would be –10°, which is the same as 350°, which is nowhere near 5 in number.]

Furthermore, if A sees B on his right, relative to a fixed direction, B sees A on his right,
provided that they use **acute** angles, or they look in **opposite** directions e.g. facing each other.
This could all get **confusing if you don't remember those two provisos**. See p108c for how to put this mnemonic into practice.

Find North from the Combination of the Sun and the Moon

When you are familiar with the path of the Sun and of the Moon, and with their seasonal movements, and with the Equator and Ecliptic, then whenever you see the Moon up during the day, imagine where the Sun's path lies in relation to the Moon's, and vice versa.

This gives you two concentric circles to visualise, which of course allows you to locate Polaris (at the centre of those paths), and from there, North.

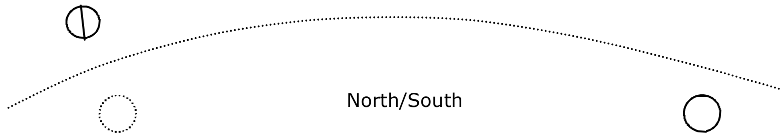

See box d/e below, to know where the Moon should be.

Find Any Other Direction Too — Using the Hour Hand

This method works, in theory (there are easier methods!), for any other body — star, Moon or planet — or for any other direction.
For instance, you may notice when the Sun is due East, and how high it is; or again, when it is in the direction in which you wish to head.
Or again, you may be following a star, and of course it will move and of course you will wish to compensate for its movement.
In these cases you substitute the time of that body being in the chosen direction for "noon", and substitute the chosen direction for "North".
Keep the watch axis pointed to Polaris as best you can judge it.
For example: you know the Sun is West at 4 p.m. At midday it will look like this… (SH version)

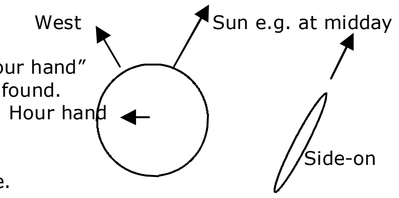

.."Point the four to the Sun; West (or the chosen direction) is between the Sun (or star etc) and the hour hand"
The catch here is that you must imagine the height of the body accurately, before the direction can be found.
This is because the method only shows the rotational position of the body — longitude wise, not its height declination wise — and with the clockface being tilted, the horizontal direction being pointed out is not obvious until you choose a height.
Of course, the accuracy of the method depends also upon aligning the clockface to the Equatorial plane.

String North

Carry a piece of string, about one metre long, to **measure the angular height of the Moon, planet or star** which you intend to use later.
You measure it once, as a matter of preparation, then use it many times. The Moon goes out of date the most quickly, but the stars are permanent

Put a loop knot at one end, to hold the bottom knot, level with the horizon or horizontal. You may have to wait, until the height is clear.
Hold the string at arms' length in a vertical plane to the North-South, which is where the star will reach its highest point.
Put a second knot to show the maximum height. This is equivalent to the 'noon' height for the Sun.

For the Sun you could use the shadow, and hold the string *down* from the horizontal. It will be accurate for a week.
Better still, use the shadow methods soon to be described, to avoid the temptation of looking at the Sun.

Now at *other* times of the day/ night, first hold the string in a vertical plane to reconstruct the expected height, then hold the string horizontally from that same high angle above you. Lastly, twist and turn, so that the horizontal string apparently points to the body used — Moon, star, planet.
You have to **twist yourself until the string is *horizontal*, and at the right *height*, and *points to* the chosen body**. Then you are facing N-S.
Actually the method approximates a slightly curved small-circle star-path with a great-circle straight course.
You can find approximate North with the string, then adjust it slightly by imagining the actual star-path as a small circle around the Pole Star.
The straight string will then appear to point to just above the star, but in your imagination, it curves down to match the guesstimated star-path.
This is the direct sighting equivalent to the 'shortest-shadow South' methods on the next page.

A 4 Week *Position Cycle* of Moonrise/set, to Find East/West

The Moon, like the Sun, lies close to the Ecliptic (in the Zodiac), and completes a cycle along it (within 5°) in 27¼ days (not quite 28 days). So...
The Moon changes position in **a four-week cycle, like *a speeded up year***. It spends a week in each 'seasonal' position: High; mid; low; mid; hi...
The Moon crosses the Equator every fortnight, rising due East and setting due West that day. Find it! Use p9c to find E/W on those days/nights.
In alternate weeks it will rise & set in the Midsummer then Midwinter positions *for the Sun*. Get to know them, and expect each one every 4 weeks.
• Its change of rise/set position day to day is most sensitive near the equinox position, due E/W (be careful), and is least critical a week later.
In between times, the rise/set **positions can be guessed by the day of the week**, by interpolation, as a good direction guide on any day.
The cycle runs fast by only 1 day in 37 1/3 days. So it is a continuous guide to E/W, **for a month at a time**, as it moves through 13 Zodiac signs.
• The best way to **set your Moon-watch is by observation**. If Moon-rise/set is too late at night, then look for Moon-set/rise during the day.
Sunset Full Moon is a good starting point for estimating the nightly-cycle; Sunrise New Moon, for the daytime. Midsummer or Midwinter or Equinoctial full Moon are very good starting points, because the Moon will be in the anti-Sun position. Visualise the Ecliptic (through the Sun and perpendicular to the Moon cusps joined) and its shape in relation to the Equator, and use a compass. Say the full Moon is opposite the Sun on Saturday, but by Monday the Moon sets due West. Every Monday for *up to* 5 weeks the Moonset will be in a recognisable position. E.g. W, N-W, W, S-W, W. On a later Monday, you will need to step back one day and say "every Sunday, for five weeks" the Moon will repeat those four positions. Adjust yourself by one day, *as it shows itself necessary*. You can simply say "It is a Thursday Moon at the moment", expecting to have the change to "a Wednesday Moon" in a fortnight or 2. Keep your finger on the Moon-pulse to keep up-to-date. E.g. Say: "It's a Spring-type Moon next Tuesday".
• To keep track of the weeks, here is a recipe: **Call the new year e.g. 'a Friday Winter Moon'** and adjust 1 day every 37 days, 10 days over 1 full year, or by **4/5 of the completed months**. 7, 8, 9, 10 days for *after* 7-, 8-, 9-, 10-ber, using the Latin names; 0-7 days while *in* the rest, 1-8.
• Perhaps you lose track of a 'Wednesday Moon' for a couple of months, and try to start again. Track the Moon's Motion along the Ecliptic.
The Moon overtakes the Sun at No Moon, gets one season ahead of the Sun within a week, at 'the first quarter', is opposite it in another week ('full'), and is one season behind the Sun in the 3rd quarter. There are 12 and a third months in a year, so the Moon is moving 12 1/3 times more quickly around than the Sun, in its seasonal cycle. **What takes the Sun one month will take the Moon only 2½ days**. (31=2.5x12.3).

Step 0: Visualise the cycle. — I.e. imagine where the **lines** of the Ecliptic and Equator must be in relation to the Sun and Moon.
Step 1: Place the *Sun* on the Ecliptic — I.e. estimate the time of year, and from that, the time before/after the next/last **solstice/equinox**.
Step 2: Try to **visualise where on the Ecliptic the *Moon* must be**, according to its phase.
Step 3: **Interpolate for the exact phase**, and label it a 'Monday Moon' or whatever. Check your guess with a compass if possible.

E.g. There was an eclipse of the Sun on Tues July 22 2009, 1 month after solstice. The 'Winter Moon' must have been 'Sat night'; and for 2009 'Mon'
E.g. 23 November is one month before the SH Summer solstice (2½ days worth of Moon motion through the stars). You can work out, for any phase, which phase and position to expect in 2½ days. For instance, Nov 23 is late Spring, almost Summer, Southern hemisphere (with the Sun South of the Equator and moving further South). The Sun is high and getting higher. Opposite that, is a *Full* Moon position (North of the Equator moving further N). It stays low in the N, getting lower. Or if it is a '*First–Quarter* half-moon' at that time, it will be in the 'late-Summer-almost-Autumn' position (approaching the Equator from the S, midhigh in the sky). Or if the Moon is a *third quarter* half-moon, it will be in an 'almost Spring' position (approaching the Equator from the N). You would therefore know that, taking the last example, over the next 3 weeks, a Waning Crescent/ New Crescent/ Waxing Gibbous/ Past-Full Moon, will rise E/SE/E/NE, and set W/SW/W/NW, in approximately 2½/9/16/23 days.

Sun Path Norths — Summary

- Noon-time: You look at your watch to see how far the Sun is away from apparent-noon-time, then look at how far your shadow must be from shadow-South. Near noon, the (untimed) shortest-shadow methods below aren't accurate, but this timed method can fill in the gap.

- Chord bisection: This works for any body e.g. the Moon.
- Sundial methods: You set the sundial to the right time. E.g. tilt the compass protractor to the equatorial plane.
- Shadow bisection: This can work for the Moon too. Just remember to use a vertical stick on level ground, e.g. a dry puddle.
- Equal-height: You can use an improvised sextant/quadrant/bubble sextant. Don't forget to bisect both the time and the direction, carefully.
- Sun at East/West: You need to be familiar with the height &/or the time when the Sun is East or West.

- Equinoctial sunrise/sunset: You keep track of how far off East/West the sunrise/sunset is, at this time of year.
- Equinoctial rise/set angle: You can extend the Sun's path to noon, rise or set, in a straight line, using the latitude angle.
- Equinoctial East: Where the shadow moves to, with slight adjustments at other times of year.
- The Noon-shadow path is always to due East. This is useful for anytime *near* noon, say +/- 2 hours.
- When the Sun is *very high* near noon, you *need* to see the direction of movement of any point shadow on flat ground — West to East.

Locating Anti-sunrise or Anti-sunset

Imagine a line from your head-shadow, *opposite the Sun's path*, cutting the horizon at the correct rise/set angle p.9a.
(The Sun's path is approximately a great-circle, so now imagine a plane, edge-on along that line, cutting the level ground.
Look for where the shadow would be closest to you. This approximates the Shortest Shadow South…)

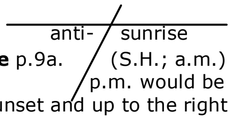

Shortest Shadow Souths ..from calculating, memorising or knowing the angle of high noon shadows.

• Walking-Stick North: You lean your walking-stick, so that at noon it would cast no shadow, and so that right now it casts a shadow at right angles to itself, a direction which is nearly West-East. Adjust (all these methods) slightly for the season, afterwards.
• Body-shadow South: You use your own shadow tip joined at right angles to your shortest-shadow tip (guess at its length and position).
This works best in the early morning and late afternoon, where a guess as to your shortest-shadow is sufficient.
• Cardboard Triangle East/West: You cut it to point to high-noon height, like the walking stick & body methods combined.
• Plumb-line North: You tie a string triangle instead, to the right shape, and maybe attach it to your walking stick.
• Notebook North: You flip open one page from horizontal, so that it casts no shadow, and also points straight to high-noon height simultaneously. I.e. the page should be held at the correct angle, mimicking the Sun's maximum height.
You can discover the correct angle by observation, yesterday,
or you can calculate that angle as "latitude degrees minus declination degrees away from vertical"

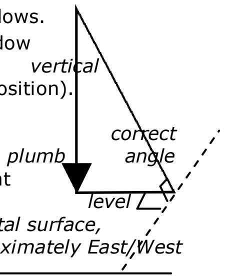

*on a horizontal surface,
perpendicular is approximately East/West*

"Star-Time" Gains Four Minutes a Day

3 minutes 56 seconds actually, by appearances; (3m 56.555s if you use the almanac, and 'the first point of Aries', because "Aries" as it is known for short, is not a real star, and it drifts slowly through the Zodiac)
Stars complete one extra orbit — 360° — than the Sun in one year — 366¼ in 365¼ days — so that's about 1° per day which is 4 minutes of time.

By "time" we mean Sun-time "15° per hour on average", in ordinary usage — a slightly varying rate in practice.
By "star-time" we mean "15 and a bit degrees per hour spin rate — a constant rate"

The stars outrun the Sun, night by night, by 1° on average, so the Sun seems to migrate 'backwards'—to the east—through the Zodiac belt of stars.
You could say: **"The Sun drags its future after it!"**, meaning that the Sun sets in the West, but its future star-position shows up just to the East.

You can zero your star-time, by timing when a bright star (preferably one near the Equatorial belt) passes overhead, or north or south of you on a particular night from a known particular longitude. You then calculate star-time from "Sun-Time plus 3m56s per day since then" including part-days in your detailed calculation, at a rate of 10 seconds' time per extra hour — 1 second per six minutes.
Your reference longitude becomes like Greenwich to you — or, alternatively, you can adjust everything back to Greenwich longitude and Greenwich star time. You can also buy quartz controlled star-watches or calculators.

Adjusted Moon-Cusp North: To the Left or Right, By Moving the Sun

"Cusp-North" can be improved, by mentally moving the Sun away from the Ecliptic *onto the sky Equator*. Do the adjustment once a day. Guess the declination of the Sun, from the season; then shift it onto the nearest bit of the Equator and the Sun-to-Moon illumination angle may tilt a little.
Instead of pointing to the pole star for the Ecliptic, in Draco, the new cusp-North will point more to the pole star for the Equator — Polaris.
E.g. the midwinter Sun needs to climb 23½ degrees higher — which would tilt the cusps of the Moon (due to the shift in illumination). Diagram 1.
Caution: The illumination angle of the Moon does not change much when the *Moon* is shifted onto the Equator, and not at all for a half-Moon.
E.g. a half Moon (90° from the Sun) at Equinox could be adjusted by 23½° to bring it onto the Equator (It is like the Sun at Solstice), (diagram 2) but cusp North does not change. Being 90° from the Sun the cusps outline one and the same great-circle.

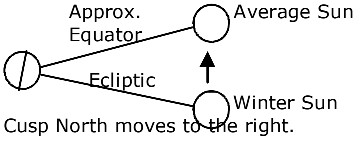

Cusp North moves to the right.

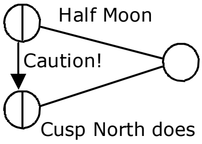

Cusp North does not move to the right.
Beware of perspective tricks

For the Southern Hemisphere
looking into the sky…

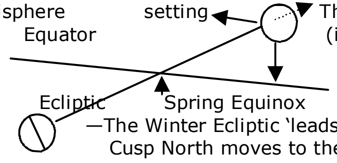

The Early Summer Sun
(it also creeps east
— 'backwards' —
along the Ecliptic)

—The Winter Ecliptic 'leads' the Summer Sun.
Cusp North moves to the left.

"By the Nose" Navigation

Our Mental Map is a Nose-Eye View — not at all like the bird's-eye-view given by a layout map. (See "Strip Map Navigation" p.95c)
Oh yes, our mind eventually fills out a (spatially-solid) layout map, but initially we collect a set of (spatially confused) mental photographs from all different perspectives taken from ground level. Our mind doesn't deal in map symbols or contour lines!
Rather it has search images, image enhancement, image recognition, three-dimensional images, and memorised **relative angles**.
Consequently, a more natural method of finding your way is what I call 'Nose Navigation' — almost opposite to mapwork.
Nose navigation uses 'strip maps' (Section 4), relative angles, and a couple of special mnemonics…

"Because (my shadow) seemed to shift right just now; Therefore *I* must have turned left"

Now substitute in the (…): Ayer's Rock/Uluru; That distant cloud; The Sun; The Moon; Compass Needle; East Wind; The RDF Null; That bright star
In other words: "My Nose Turns EVERYTHING Away!" i.e. when I turn my nose, everything else seems to turn oppositely.
This simple rule attacks head-on your insistent illusion that you have been walking "straight" ahead!
And your illogical belief that "Ayer's Rock has moved"! Neither has your shadow 'turned right' just now, simply because your nose 'hasn't moved'!
It is not as easy as you think to apply, e.g. when you are reenacting where you thought you had gone, in your (already confused) head.
The second nose mnemonic is on the opposite page.

Seasonal Adjustments For Equinoctial East

Shadows move due East at Equinox, all-day. At other times of year, when sunrise/sunset is not due East/West, the Shadow path curves slightly at an angle away from sunrise/sunset, an angle which you should become familiar with each week. Start by assuming that the shadow is moving "East"… Verify this **Rule: Shadow Paths Bulge toward the Hot Pole**, even in the tropics, in both hemispheres. (The Sun crosses at Sept 22 & March 21.)
Adjustment: The first guess at **"Hot Pole" Needs to be Adjusted Slightly,** less near noon, little or none nearer equinox, to "un-bulge" the path.

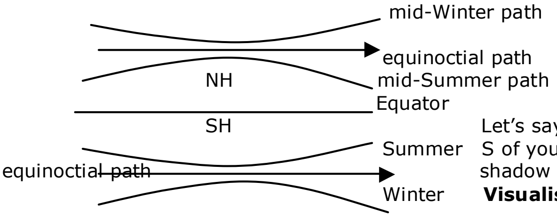

Remember that "the Southern Sun Sets Somewhat South in Southern Summer", so that opposite-Sunrise & opposite-Sunset need to come back South to point out E & W correctly
E.g. You guess East (or West in the a.m.) from the shadow's movement over 1 minute,
(It will roughly be opposite sunset/sunrise direction if you are late/early in the day)
then twist your guess a little back to the true hot pole, and to due East (or West).
Let's say it is midwinter in the Southern Hemisphere, so your shadow is somewhere S of you. The 'hot pole' is the North Pole, so twist 'North', as gauged from your shadow tip, away from the shadow, back toward you (i.e. North, to the hot North pole)
Visualise the bulge, then straighten it out at each end, back toward the hot pole.

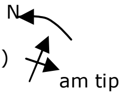

North-Finding from Two Paper Triangles

or even from only one

Cut out a back-to-back triangle shape and fold it along the join. Line it up as shown and 'N' will point to Polaris…

'1' is your South **latitude** (For the Northern Hemisphere use the North Latitude and swap N & S)
'2' is the **polar distance** of the Sun (from the 'depressed' pole, N)
i.e. its co-declination, or 90°-N dec. See p. 18b.
You can set it by observation or calculation; and can fold it to the correct angle, rather than cutting it.
('3' is a right-angle, but '1' and the horizontal is what really matters)
This two-triangle method works best away from noon, but nearer noon simply find the way shadows move
The 'N-S' edge is parallel to the Earth's axis and acts as a hinge around which the Sun revolves.
Sun and shadows move at right angles to that axis.
You only need **one** of the two angles if you find the direction of shadow movement and swivel 'N-S' to be perpendicular to that. (You don't even need one angle, near noon)

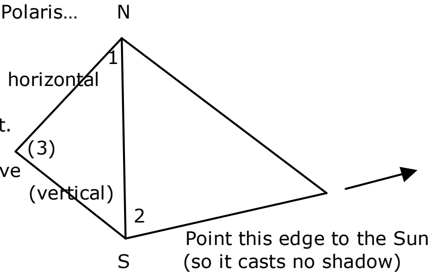

E.g. Poke a straight stick into the ground, pointing at the Sun, casting no shadow; (Or use any point-shadow rotating around its point)
By waiting a short while, find the direction of movement at a half-way point, in mid-air, i.e. maintaining an equal distance to the tip, and manoeuvre your 'N-S' axis (index finger), to be perpendicular to how shadows move, then at the correct angle to the Sun and/or Horizontal.

Working the Star-Longitude Numbers

The reference star point, a non–star, is 0°N/S;0°E/W, and is called "The First Point of Aries" or "the Vernal Equinox" where the 'northern spring' Sun crosses the ecliptic (on March 21). Other stars are measured from that (imaginary) star, like latitude and longitude are measured from 0,0 a nameless spot of ocean — each star has an almost fixed longitude in relation to Aries.
N/S 'latitude' is called +/- ' Declination '
E/W 'longitude' is called (a) Right Ascension — R.A. — or 'Star Time' — It is measured **Easterly**, in hours at 15° per 'hour', from "Aries"
Or (b) Hour-Angle — H.A. — It is measured **Westerly**, in degrees, like longitude, but 360° around, from "Aries".
The first, RA, is designed to tell us the time — **when** the star will show — since all the stars move around westward with Aries.
The second, HA, is designed to track the West **longitude** of the star in reference to the Earth, as the star progresses West around.
We need them both, but they are entirely equivalent to each other.
(The only trouble is that a star 'hour' of 15° is slightly faster than a Sun hour of 15° of progress to the West.)
Next you need to know that the first point of Aries was at 100° W. Earth longitude at the beginning of the millennium, and at every turn of leap-year thereafter, at January 1.0 (midnight). Then it gains at a rate of 1.002737909, or about 1° per day on the Sun, (3m 56s time), so that the stars turn full circle during four seasons. "1°" is adequate accuracy in the short term for us but not for yacht navigators.
E.g. January 31.5= 100+30°+0.5° W., then you add the H.A. (or subtract 15 x the R.A. hours) for a particular star — that's the longitude where the star will be overhead. **For exact numbers consult an almanac**, but I am concerned here with giving you an overview.

Topographic-Level Winds

This is what you feel in your face at ground level, or close to it.
Fog; Smoke; Valley Winds; Sea Breeze; Eddies; Catabatic Winds; Forest Winds; Scree-Slope Winds; Mountain Winds; Glacier Winds; Lake Winds.
The various thermal winds listed will of course be very variable. The thermal winds flow from cooler areas, before rising in the warmer region.
Nevertheless they may betray the slope of the land, for instance at night. For example, fog forms on calm cold nights then drifts slowly downslope.
Being dense, fog will drift down-valley, down-stream, and off-shore. You can see the location of creek outlets from the early morning beach fog.

The ground-winds due to pressure systems, are affected by overall friction and therefore average out slower and less Sun-twisted than upper-winds.
Ground-level winds form waves, foam 'streets', sand ripples & dunes, drag marks, snow ripples & drifts, which all may persist beyond the winds, and they cause lopsided bushes and salt-spray-'cut' bushes. See Section 3 for more details.

Gusts are at different directions to lulls since upper winds penetrate to lower layers.

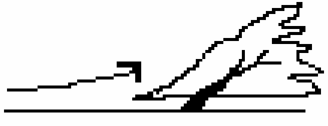

"Nose-Navigation Needs an 'Anti' Mnemonic": "Nor-West, Not-Least"

N.O.R. West: "North **on Right** = Going West" or "Needle on Right = Going **West**" i.e. going into the westerly **half** (top diagram)
N.O.T. L.East: "North **on the Left** = Going East" or "Needle on the Left = Going **East**" = to the easterly **half** (bottom diagram)

This is always true when you are relying on "Needle North". **You don't have to think any further.** But,
It is equivalent to ANTI-CLOCKWISE BEARINGS, which you can't read from a clockwise-marked protractor!
The following numbers refer to an ordinary compass — marked clock-wise, and set to 0° as straight ahead (not North) so North swings left and right instead…
Example: in the rain or fog; the needle swings meaninglessly
Example: deep in a cave network
Example: on a yacht, with a fixed hand compass
Example: with a fixed car-compass, not marked anticlockwise
Example: making a strip map; or using one
Example: using a compass-holder, which fixes 0° as straight ahead

The compass needle might point to "045° NE"
but your nose is pointing NW
The compass protractor numbers say "North-West, 315°"
but your nose is pointing NE
Interpret "Anti-clockwise, West from 0"

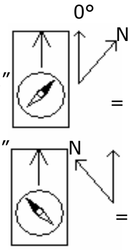

Needle on the Right
use "NORWest"
= North-West N045°W
Needle on the Left
use "NOTLEast"
= North-East N315°W

Top Down Orientation

The Big Bang **Universe** bloated over billions of years; and broke into bubbles of turbulence; all bits in each bubble orbiting each other; in millions of galaxies, with 20 or so galaxies in our little **local cluster**;

Local galaxies, like the Magellanic Clouds or Andromeda, glow milky (just as 'galact-ose' says 'milk-sugar').

Our galaxy is a great spiral, swirling slowly, anti clockwise from above i.e. Eastwards around its North-South spin-axis — flat like a fried egg — a 'great circle', glowing as the Milky Way. The thickest milkiness, in Sagittarius, marks the Galaxy Centre.

The Milky Way is like the 'Equatorial plane' for the whole spiral, seen from within, with its northern pole star in 'Berenice's Hair' (28° N. 168° W.) and its Southern pole star in 'Sculptor' (28° S. 348° W. 012° E.) and our milky galaxy is, in its own turn, full of spinning, gyrating gyroscopes.

You know how a spinning top will pirouette with a slow-motion gyration when it tilts slightly? Well, precession, as it is called, causes slow changes.

Stars are Suns, especially shiny in our own spiral arm of the galaxy (i.e. closest) and seemingly set solidly in space by their extremely slow motion. So the constellation shapes are chartable for centuries. Their apparent positions are gridded out in relation to the first point of Aries, which in turn is fixed each year by our Northern Spring Equinox—not a star in the sky, but a point which creeps along the Ecliptic, to the West.

Hint: When you are wondering whether a rule holds in both hemispheres or not, ask yourself whether you are down at the level of the spinning Earth or up beyond that in a larger scale. Try to imagine looking at the whole globe from an outside perspective. E.g. the appearance of the Moon — its phase, eclipses, tidal forces, etc — is mostly independent of where we look at it from Earth — and will be the same for both hemispheres.

But it's apparent path — 'noon' direction; rise/set angle; etc — is seen down at Earth level — it changes with the hemispheres.

Closer to Home

The solar system is also spinning flat, anticlockwise, but slanted — tilted like a top, in relation to the Galaxy, and faster.

Its 'top' star is currently in Draco, in the North. Its bottom star, in the South, is the larger Magellanic cloud.

Its 'Equator' is called the 'Ecliptic' — marked by the stars in the Zodiac, and the planets. In our lifetime we won't notice the Ecliptic stars change.

Our Sun circles the centre of the galaxy every 225 million years—heading toward Suhail.

To our sight, because we orbit it Eastwards, i.e. around its N/S axis, not ours, it swings seasonally, and also edges Easterly along the Ecliptic circle, revealing new zodiac stars to the West of it, and placing itself diametrically opposite a series of anti-Solar stars, so giving us a star calendar to look for.

It also spins easterly, on its own axis, showing its sunspots for some days at a time.

The planets lie planar, 'wandering' (as the Greek word plan-et says), in the Zodiac plane, circling easterly around the Sun, over a timescale of decades, and only ever seen very close to the Ecliptic.

Mnemonic: Son, my very earnest mother 'as just served us 19 platefuls. Sun, Mercury, Venus, Earth, Mars, asteroids, Jupiter, Saturn, Uranus, Neptune, Pluto.

The earth is an inner planet, **circling the Sun**, easterly, so that the stars seem to out-run the Sun in setting West — star-time spins more speedily than solar time. Planets closer to the Sun than we are, are never seen far from the Sun, but outer planets can move right away from it.

The Moon

..moves mostly on the Ecliptic but might miss that mark by as much as 5°. This 5° slant causes a precessional change taking 19 years to cycle.

The Moon moves around the Earth monthly in an easterly ellipse, and ever eyeing us, each full Moon, is the Rabbit, because the Moon also spins easterly, on its own axis, exactly once a month, so it stays facing us.

The Moon moves slowest to set Westwards, since it migrates eastwards in the sky each day. 12°½. Or **50 minutes later per day**.

And as it pulls the tides with it, they too are later each day, almost an hour, on top of normal locally fixed delays due to friction.

A full Moon rises and sets fully opposite the Sun, give or take 5°.

The Moon is most responsible for tidal forces, but the Sun plays some part too, reinforcing or diminishing the Moon's effect and changing the timing.

Tidal force attracts the nearest water the most, and the farthest water the least, resulting in two tidal bulges...

When the Moon is eccentric from our Equator, there is one big Northern bulge every 24 hours, and a smaller high tide every 24 hours, as the Earth spins through the bulges.

There are equal bulges every 12 hours on the Equator. In between are the low tides, every 12 hours.

You should be able to tell which bulge is 'nearest' the Moon despite delays due to friction.

Southern latitudes will have one big bulge every 24 hours, and one 'small high tide' every 24 hours, offset by 12 hours from the NH.

Try to correlate the Moon's phase, declination and height with the local tides—timing, inequality, range—e.g. "The Full Moon rises with the tide"

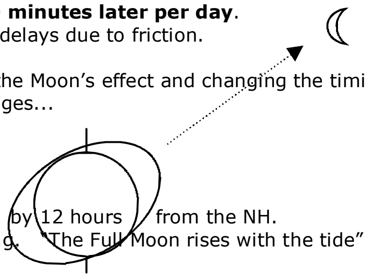

The Moving Horizon

Since it is the Earth which is spinning, not the Universe, it is **the horizon** which rises, in the West — where it eclipses the Sun, so that the Sun only seems to set in the West.

The horizon slumps in the East, so that all the sky things — Sun, Moon, planets, stars — seem to 'rise' in the East.

Any spots on the Northern half of the horizon, circle to the east around Polaris — clockwise through the sky — i.e. 'seen from below'. Or, seen from above, that is from the North, looking South, spots on the Southern horizon circle anticlockwise around South Pole Star i.e. seen as if 'from the top' looking down on the Earth and its South Pole star. They hit the same stars each night, at the same angle, which is related to latitude, so that the fixed stars none-the-less seem to trail 'small circle' tracks through our sky, and they always hit their favourite horizon spots when they set, always at the same angle, and those same stars will rise later in mirror image to that around North/South, giving rise to **north-south-east-west finding by bisection**.

And Equatorial stars will rise/set due East/West.

The southern-hemisphere horizons move to the right, the northern-hemisphere ones to the left. Looking South though, from either hemisphere, it will be seen as the horizon circling anti-clockwise around the South-pole star. How come? Well, the southerners see the bottom of the star-circles, around an elevated pole-star, the northerners see the top of the same circles, around a depressed pole-star.

The Spinning Earth

The earth is **spinning like a top**, on its own axis, easterly, half as fast as a '12 hour' clock.

This gives us **the single most dominant direction clue on the otherwise uniformly round Earth—the North-South spin axis of the Earth.**

Spot its Equatorial Plane by the Pot. The 'top' is always pointed to Polaris. Spot the South Pole Star by the Seven Superstars surrounding it.

The spin generates a magnetic core more or less North/South — Find magnetic North in far northern Canada.

The magnetic field poles generate auroral displays, anywhere near the poles.

The earth is also tilted like a top, at 20 odd degrees.

The tilt gives rise to seasonal Sun height shifts, somewhat like a sinewave (just imagine the Earth's tilted spinaxis not changing as we orbit the Sun) and it turns on un-twinned top-of-the-tide bulges, twice in 24 hours. The Pole Stars are latitude° down/up @ 0°/180° T.

Any tilted top will precess the equinoxes — this generates a precessional gyration of 26,000 years, but this time, retrogradely, i.e. to the West, along the Ecliptic (3° in 200 years) — i.e. the stars 'move' to the East through our star-map-grid.

Longitude has a lot to do with spin-time, seen from above the spin-axis. E.g. West longitude is how long after Greenwich, a place meets the Sun.

When we want to mark the apparent place of the Sun in the sky, or anything else, at any particular time, we can just record the latitude and longitude of the point on Earth directly 'under' it i.e. closest to it.

But visualising that in reverse, we simply **project the Equator and poles, latitude and longitude up onto the heavens to grid them out**.

Time is counted in cycles

The larger the distance scale of any motion — pulsation or precession, orbit or rotation — the larger the time scale of the cycle of change.
The Universe has been expanding its size and condensing its matter for many billions of years (and perhaps it will not 'cycle') It appears static to us.
Geological history spans billions of years, like the existence of our spiral arm. In our lives we see only a frozen frame of our spiralling Galaxy. It cycles, but we can't see it. The timescale of the rotation of the galaxy is in hundreds of millions of years.
In our lifespan we see only the slightest change in relative star positions, through telescopes, due to the 'proper' motions of the nearest stars.
It is in the nature of spiral turbulent motion that the larger rotation replicates itself in the smaller scale, but at a faster rate.
Each of the components of our solar system are also rotating and orbiting in the same rotation sense as the Galaxy. (Uranus is an exception)
For example, the Earth is rotating eastwards, orbiting the Sun eastwards, and has a Moon rotating eastwards and orbiting us eastwards.
"Anticlockwise, as seen from the North" is the single best spin you can put upon all that you see—including the N/S Earth spin-axis as 'the single best direction clue on our spherical Earth'. The tilt of the Earth's spin-axis gives us the year, and the Moon's orbit gives us the month, independent of our rotation. The Earth's rotation gives us our day. Star time outstrips solar time slowly. The Southern Cross clocks it correctly at the close of March, and the Centaur soon follows. The first point of Aries gains a touch over the stars, migrating 50 seconds of arc along the Ecliptic per year. The vernal equinox arrives 59 seconds late for its 366th "orbit", after 365 days, requiring 24 leap years in 100 years. 365.242 days per year.
Atomic time, used for astronomy, is independent of all these tilted rotations, orbits and precessions, which are all independent of each other too.

The Precession of the Equinoxes

The actual or 'proper' motions of the stars are only of cosmological interest, because they change relative positions only over centuries.
But history, actually our slowly gyrating Earth-spin axis, rewrites our star charts (**im**properly, as it were) with precessional revisionism, so that stars 'move' slowly easterly through our star-map-reference-grid—an Earth-centred way of seeing the sky — with 26,000 years needed to complete a full circle precessionary Earth-wobble. This artificial motion is called "the precession of the equinoxes"— which means that our reference grid readjusts just as quickly as our earth axis of spin precesses — by five arc minutes in six years. That's why we need an almanac to predict the 'fixed' star positions!! A good analogy would be the slow drift of the north magnetic pole, **if** we set our latitude and longitude by it — that would yield a false or 'improper' 'continental drift' in addition to the proper continental drift. See the appendix at p. 123.
To visualise the Ecliptic, consider the Equator in the heavens to be geostationary, although invisible, but the Ecliptic to be like a sinewave moving daily to the West along it, on the video display unit of the sky—as on a cathode ray oscilloscope—with a period of 24 hrs & an amplitude of 23°½.
The Sun appears fixed somewhere along the Ecliptic, creeping 'backwards' on it (against the setting sky)—to the East—during the whole year
To visualise the precession of 'the Equinoxes', imagine that *the points where the Ecliptic sinewave crosses the* Equator (the x-axis), move 'forwards' ('with' the setting sky) — to the West, very very slowly — during the millennium. At the moment these 'equinoctial points' are between Regulus & Spica (the Autumnal Equinox), and between Diphda & Markab (the Vernal or Spring Equinox, or the 1st point of Aries, (0,0) for the stars).
You might say that the Sun and the Moon and the outer planets all move West through our sky but East through the starry heavens.

Global Weather Patterns

The spinning globe makes the climate zones, and weather belts, follow the latitude bands. When the wind tries to blow, from a High, into a Low, in the Northern Hemisphere/Southern Hemisphere you see huge anticlockwise/clockwise spirals instead, by Coriolis force. E.g. the wind, tries to go straight ahead from the hot spin-equator to the cold spin-pole while the spinning Earth slows down beneath it (it moves slower nearer the Poles).
In the North the wind therefore seems to "move to the right" as we see it, and in the Southern hemisphere "to the left". Circulation is the result.
Note: Winds around a Low go against the Coriolis force. To visualise the pressure system winds at low level, imagine placing two flat-headed screws pointing up, one on Japan and one on Sydney, an ordinary right-handed screw for Japan but **a left-handed spiral screw for the Southern Hemisphere**. Run your fingers over the threads and a right-hand thread is like screwing a screw in or out, a left-hand thread is opposite.
High pressure system winds screw downwards and outwards; Low pressure system winds reverse that—up, converging.

Higher level wind circulation reverses itself from that because the convergence/divergence is reversed.

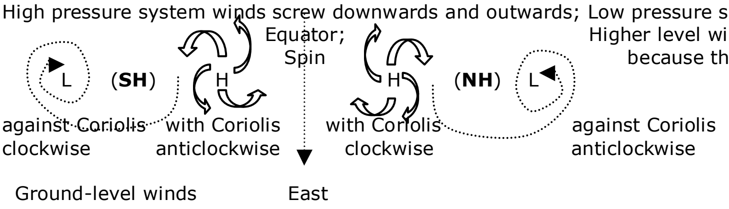

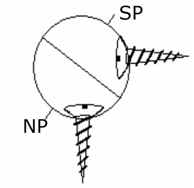

The Spherical Earth

The earth is curved round, like a globe, 360° round — 40,000 km around (by definition of a kilometre) 21,600 nautical miles, by definition.
◇ at 400 grads in a circle that's 100 km/grad ◇ at 360° in a circle that's 111.1 km/degree.
The Equator is a 'great circle'. Longitude circles are 'Great Circles' — passing through opposite poles, and perpendicular to all latitude lines.
Longitude counts 360° around East or West, or 180° East and 180° West from Greenwich.
Its North/South dividing lines mark out 24 one hour time zones of 15° each, and those North-South lines must converge on the poles, so that longitude degrees are of unequal arc-length at different latitudes.
Latitude has 90 equal degrees from the Equator to pole around the surface, which, by definition of the nautical mile, as one minute of arc, makes 60 nautical miles per degree of 'arc-distance'. 21,600 nautical miles must be the same as 40,000 km, so that's 10 km to 5.4 nautical miles. (2 to 1)
Latitude circles are 'small circles'.
We use them to divide the polar regions from the temperate regions (with the Arctic Circle, for instance, marking the limits of 24-hour nights); and to divide the temperate regions from the tropics (for instance with the Tropic of Capricorn, which marks the limit of 'high noon' fully overhead).
'Polar Distance' complements the 'Latitude' and reminds us that arc-distance is real distance.

Moving the Horizon

At the spin-equator, on a globe, the polestars must lie horizontally, and the Equatorial spin-plane must pass East/West overhead.
But move north/south to another latitude, and the earth being round, the horizon — which follows you, like 'up' changing with you — must fall and rise latitude-degrees, so that one pole star falls latitude-degrees out of sight behind you and the other one, ahead, elevates latitude-degrees above the horizon. The Equatorial plane must also fall latitude-degrees, behind you as you move, from being overhead. Other star-latitude (declination) circles in the sky (e.g. star-trail circles) must also rise or fall the same amount — latitude degrees.
Now this same thing happens if you move in **any** direction, say ex degrees away, but for this let's just temporarily freeze the spinning of the globe.
'Horizontal' and 'up' must follow with you exactly ex degrees, and so the horizon now blocks out ex degrees of sky behind you, and reveals ex degrees of new sky ahead of you.
Looking side on, you will see the horizon **tilt** by ex degrees in relation to the old skyscape, lifting behind and falling ahead, of your direction of travel
Looking at the same star-time any night, will "temporarily freeze the nightly spinning of the Earth" for you, for star navigation.

Spherical geometry also means that if a place is 2° or 20° away from you, then it must also be 1° or 10° below your horizon — half the arc distance
E.g. the North Pole on Earth is ½ x (90° minus latitude°) below horizontal, and at 0° True.

END OF SECTION ONE

Navigation is More Than Navigating

Properly done, navigation is your chance to fit yourself in to the experience of walking through the land.
Why are you there walking?... Navigation zeroes in on the land. Do you want to see the landscape?... Navigation opens your eyes.
Don't you want to appreciate ecology and geography?... Navigate by them! Navigation stops you "drifting off" into mindless incompetence.
Do you want to remember the views and special places?... Navigation forces you to improve such memories.
So, achieve confidence when you are alone; tackle more difficult trips; lead others; call yourself "experienced"— navigation is the key.

Common Sense

Navigation skills invite you into the wilderness. The wilderness is a wild place! — a world of dangers. **Don't go**. But if you do...
..at least make sure your navigation skills are up to the task.
Likewise: Don't venture to sea with approximate star navigation techniques until you know exactly what you're doing.
Learn much more about standard astronavigation, before you experiment with new techniques.

Your Subconscious is Superior to Science!

Your natural abilities will do a better job at keeping you oriented. Your brain is a supercomputer.
That makes it subject to the "rubbish in; rubbish out" truism. Your job is to feed relevant data to your subconscious, not to feed, say, a calculator.

As far as important detail goes, your mind will 'go for the jugular'. Exact science will go for unimportant decimal places.

E.g. I can immediately guess "about 8 km away" far more quickly than any triangulation.
E.g. Will any numbers ever tell you how to recognise where you are? Or how to imagine what that peak might look like from behind?

So feed clues into your brain. It is portable, waterproof and teachable.

Don't leave your mental map behind.

Visual Assessment is Your Main Clue — Rely on It More

Since noone navigates well when they "can't see where they are going", we must be relying on visual assessment most of the time.
The idea is to run with that fact of life, and make sure you have a good look when you can
(e.g. from the lookout; while driving up; at the crest of a ridge; at the waterfall; on the bridge; in the riverbed)
Further, bring to conscious focus what is normally subconscious, in what you see.
And aim to remember what you saw. (Draw it; take notes; talk it out; use visual memory skills)

A What-to-Look-for Checklist: How far is it to the skyline? Where are the Sun, sunrise, Moon, wind and North in relation to the skyline?
How does the weather interact with the land — wind, cloud base, Sun shadow, wind shadow, cloud shadows, haze?
Talk the skyline through — name its features. What do you recognise already? What does it remind you of?
What landmarks are visible? What distances to landmarks? What sizes of hills? What is the strike or trend of the ridges, ranges & valleys?
What angles between landmarks? What alignments? What valleys will your current path take you through?
What lies in between? What slopes & shapes are showing, especially for your route? Can you see the path ahead? Is it curved?
What difficulties lie ahead? What easier paths are there? What clear lines of sight are there? What alignments will become visible?
Can anyone see where you have come from? Could you recognise 'here' again? Can you see your destination or departure point?
How many different vegetation types? How high is it to the skyline in each direction? How does what you see mesh with the map?

Integrate: Jetstream, Earth Rotation & Heaven Rotation

Imagine your global motion. Tilt yourself and try (futilely) to oppose the supersonic speed of the Earth's rotation, as if walking upon a treadmill.
Aim for a Big Picture 'feel' for each local direction.
The idea is to go outside and walk through each of these motions until you **know** what it feels like.

(Both hemispheres)

(read down these lists, each in turn)

When you are............	**standing still**	**or moving Westwards**	**or Eastwards**	**or Northwards**	**or Southwards**
The **Earth** rotates	Eastwards	rotates against you	rotates with you	to your right	to your left
The **Heavens** move	Westwards	slowly forward	quickly upwards	to your left	clockwise
The **high clouds** oppose the heavens outrun the Earth	jet Eastwards	zoom toward you	creep ahead of you	move to your right	to your left
You are being **tilted**	to the East	backwards	forwards	to your right	to your left
Lean oppositely	to the West	forwards	backwards	to your left	to your right

---

Measuring angles at 1 cm per degree

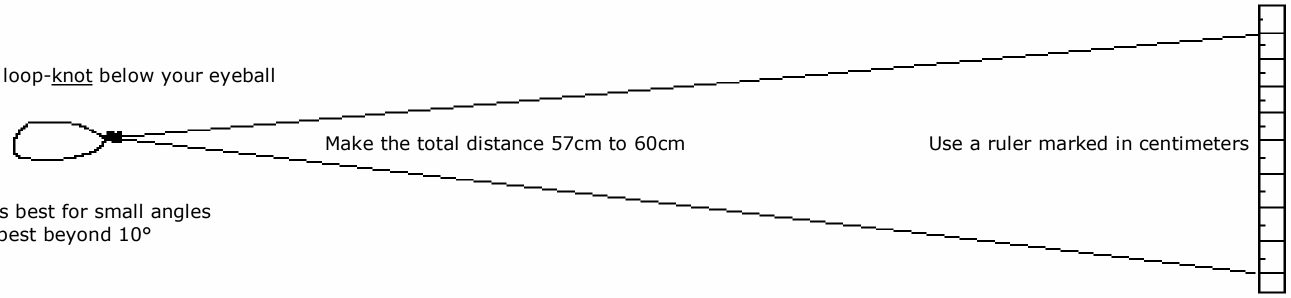

Hold the loop-knot below your eyeball

Make the total distance 57cm to 60cm

Use a ruler marked in centimeters

57.3 works best for small angles
60 works best beyond 10°

Orientation = Direction Sense

Oriental means "the rising" i.e. in the East (i.e. East from Europe!). Occidental means "the setting" i.e. in the West.

Because most ancient temples faced the midyear sunrise and there were no North-pointing magnetic compasses, and since all the world can agree on where equinoctial sunrise is, most of the rest of the world has used East, for most of Time, as their method of "orientation" — as their cardinal direction. Imagine putting East at the top of maps, so the Sun "falls down" the page.

Before magnets, East was the sensible choice for a standard direction.

This is no coincidence: "To the East" is also the watchword for the spin-sense of all our solar system parts. "North" and/or "South" is inferior

The direction of the *axis* of rotation doesn't tell you the *sense* of rotation; But the *direction of rotation*, gives you the axis of rotation *as well.*

East was the more sensible choice. You didn't lose orientation when you crossed the Equator; the Sun still rose in the East.

I recommend that you concentrate on this when changing hemispheres, and even now, to maintain your global 'orientation'.

Orientation means direction sense, direction finding, direction recognition, placing yourself in relation to over-the-horizon landmarks,

Or twisting anything (e.g. your map) to the right orientation.

Then again, it means starting off by getting your bearings straight.

Orienteering means athletes running through the bush.

Disorientation means... possible death.

The Navigator's Top 10 Problems; and Band-aid Solutions

1.	Not staying **alert** to navigation; not doing all you need to do……	**Talk** navigation; love navigation; practise navigation; **interpret** all clues
2.	Few **Clues** — E.g. Rain/Fog/Blizzard/Nighttime/Tall grass or dense vegetation/Look-alike landscapes	**Sit** still, until you can see; **go prepared** e.g. with a contour map; a compass; learn how to keep consistent courses
3.	Getting **Lost**	Remember the **start direction**
4.	Getting **Disoriented**	Be able to reorient; imagine the **Sun** in all situations; work top-down
5.	**Forgetting** the Way, the bends, etc.	Take a **notebook** and pencil; talk it out loud
6.	Losing track of **direction**	**Predict** the trend; use a **compass**; a notebook; integrate all direction clues
7.	Losing track of **distance**	Predict the timing; use a **watch**; use the map scale; and notebook
8.	Inadequate **Map**-Work	**Learn** how to use a map; study it first
9.	Not learning how to navigate	Take this booklet. **Practise.**
	Not knowing what to look for e.g. bush north	Practise being observant, at the expense of other preoccupations.
	Not **understanding** e.g. the lie, contours, heavens	**Teach** what you do know
10.	Choosing a wrong or difficult **route** E.g. a long "shortcut" across a dense gully	Get **directions** Don't 'shortcut'.

Reorientation

Reorientation is not an automatic body sense — it is an ongoing action.

"Orientation" in navigation is a misnomer, since it suggests some sort of initialisation, not to be repeated, rather than a continuous reprocessing.

"Re-orientation" is a more accurate description of what the navigator is involved in. Continuous reorientation.

Whether it is subconscious or conscious, it is something which must be done and redone, or else you are simply not oriented, now.

It is not good enough simply to be "good at navigation (theory)" — you must be putting it into practice, continuously, or else each bend will disorient you.

The essence of reorientation is conscious interpretation of the meaning of what you are doing.

This is, at the best of times, a bit difficult — like mental arithmetic — but it must be done, or else your brain will remain confused.

You may hope that direction sense would "become automatic with practice", but the last step of conscious interpretation must remain conscious.

And that is the process which should become second-nature. You must be aware, or you are fooling yourself into a false sense of security.

Delegating to your subconscious, is abdicating your responsibility. So bring yourself up-to-date, continuously, and...

Do not be overconfident. You are not oriented when you think you know where you are and where other things are—but only when you are right!

There are so many ways to make silly mistakes: confusing alternative systems; mistakes in numbers; forgetting your lessons; optical illusions; instrument errors; flawed theories; trusting approximations; misidentifying a star; making a logical blunder; and, of course... being overconfident.

Lost? Rule One: Stop! ..getting More Lost

STOP — so that you don't get deeper into trouble. ("When you get yourself into a hole, stop digging")

DON'T PANIC — Look for reasons not to panic "I can always walk all night, or sleep under that ledge,
or wait out the weather, or go hungry, or enjoy it…"

PRAY — He really does want to help the helpless.

SIGNAL — Shout 'Help!', 'Cooee', Whistle, Smoky fire, Heliograph, 3 of anything, SOS, CQ, Fire at night, Flashlight

LISTEN — for a reply.

THINK p. 37d

RETHINK — your journey

SIT It Out — = The Simplest Way to Beat the Fog. Gather fuel for a fire. Don't try to move — you may not win.

CAMP — choose a place that will be a good 'base' for further exploration.

SLEEP ON IT

STAY — in the area, where you can be found, e.g. on the track, at a hut or intersection.

PLAN — Implement a pre-**thought-out plan** e.g. Stop, Think, Navigate, Camp, Retrace, EPIRB, Stay.

For a Start: Slow down — almost to a stop. Be slow, cautious and thoughtful, not impulsive.

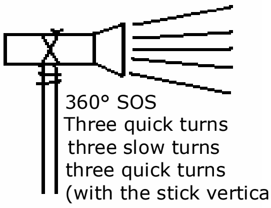

360° SOS
Three quick turns
three slow turns
three quick turns
(with the stick vertical)

Point and Blink = 10% Shift

In the field you simply point and blink—If you focus in the distance, the shift is about 6 degrees, which is 10% sideways.

6° or 1 in 10 is a very useful angle to guesstimate. Point with one eye and finger, then swap eyes, to see what a 6° shift looks like in the distance.

For repeatability you should touch your two index fingertips as you point.

For accuracy, use a pencil tip, and know the exact figure. The instructions below will show you how to hold the pencil to get an exact result.

To put the Point and Blink method on an accurate footing, measure your Inter-Pupillary Distance at Arm's Length.

Stand behind a window, pointing through it at some distant mark on the horizon, with both your finger-tips just touching the windowpane.

Use one eye and the same-side finger; Press your fingerprint onto the window; Swap eyes and fingers, back and forth, moving only your eyelids;

Check your alignment—both your fingertips should appear to exactly overlap when focussed on the distant object. Press your second fingerprint onto the glass; Measure the separation carefully.

Because the pupil is not pinpoint, it matters that you measure the practical separation at the working distance. Measure that working distance carefully— from the corner of your eye to the window when your pointing fingers just touch it.

Compare the two distances; i.e. divide the interpupillary distance by the armslength; **expect about 1:10** or 6°; Find the arcsine of that ratio for better accuracy. Try it out on the length of the Southern Cross. Compare it with 3 fingers, and 4 knuckles, and your binoculars' field of view.

Measure again on a middle-ground object, as when measuring the sideslip of a yacht by pointing at the wake. Measure out the distances as a check

Orient Disorient Reorient Lost? Angle

The "Good" Navigator. You.

If you like exploring, you can use that to advantage. Or maybe you are a good photographer, well then you will understand 'mental photos'.
Use your creative writing streak, in creating memorable names, or else, appreciate the action 'direct', carefree and spontaneously.

Perhaps you are the obsessive type — well then, it's map and compass detail which will suit you.

So whether you are the artistic type, using your visual dominance, or the cool calculating logical sort,
or maybe a 'thinker' doing a good job at planning, or a 'doer' doing, you already have many useful skills for navigation.
Don't talk yourself down.

Those handful of useful traits which are already within you, you can develop into an orientation sense.

Think out Loud

Your subconscious will be telling you "That's odd".
Your conscience will be whining "Something is wrong".
Your mind will notice, before you do, "That's new".

When you notice something talking inside you, you need to talk back. Out loud. Don't leave it 'inside'.

Internal Problem	**Antidote**
"Shouldn't we be there by now?"	Ask the others.
"Have we missed the junction?"	Discuss it.
"Why is the Sun over there now!"	Name the lurking enemy!

You Need Excuses to Stop and Have a Good Look

- If it is a local peak, say "Let's look for a minute while we have the opportunity"
- Declare a chocolate stop!
- Take off your pack, adjust the padding. Adjust your socks, air your feet, put on a Band-Aid.
- Get out the maps; Identify the landmarks. Draw the relative angles to other lookouts, for future reference.
- Get out the camera and tripod. Unpack your pack on any pretext — noone can move until you repack!
- Make a sketch, compose a poem. Start an interesting conversation.
- Go for a swim or wash some clothes.
- Cook your main meal at midday & keep the biscuits for the late meal.
- Camp there.

You Can't Easily Revisit Places Left Behind — unless you first spend memorable times there.

Visualise Familiar but Invisible Landscape in Relation to the Sun

You need to 'fill in the void' between the very distant reference, and what you can see, with what you know, but can't see.
I.e. When you get a clear view, mentally plot and remember the sky-to-land layout for later use. "If the hill disappeared, the hut would be *there.*"
Mentally reconstruct any view which is helpful. E.g. as seen from camp, or from the start, or the peak…
Concentrate on knowing and revisualising landscape direction trends. "If I were at home, I'd see the coast… *across there*"
Do this at various times of day, **paying particular attention to direction *trends* in relation to the Sun's path**.
After all, that's what our direction sense consists of—a confidence in the real directions just out of sight, based on what we can see.
Then, when you are 'bushed in' and you see the Sun through the canopy you can say…
"Ah! the road is running sideways to the Sun, the valley lies entirely off to the right, I must have swung around left, parallel to the ridge…"

Section 1 dealt with the Sun/noon/North in the global abstract—unrelated to the local detail.
Section 3 deals with the local detail in relation to the Sun.
"Land to Sky" (soon, p 38d) is specially for one time of day, and concerned mainly with particular landmarks, not whole trends.
But now catch up with the overall game—**the Sun's path, the local skyline, and the local layout are one piece for a substantial local area**, and you need to mentally integrate them, and carry them with you, to avoid disorientation.

Guesstimating Angles at 1 degree per cm

First find how to make your **arm's-length 57 cm**, e.g. by touching your fists together in front of you.
Next, start measuring in **centimetres**.
Find a 1cm wide fingernail—1°. Check that your fingers are 2° wide, for 2°, 4°, 6°. Are your knuckles 2 cm apart? 2½° for the big ones.
A small fist, with the thumb tucked in is 9°; 10° with the thumb showing; a palm is one 'point'.
15° and 16° may require experimentation. 20° can be spanned with the fingers.

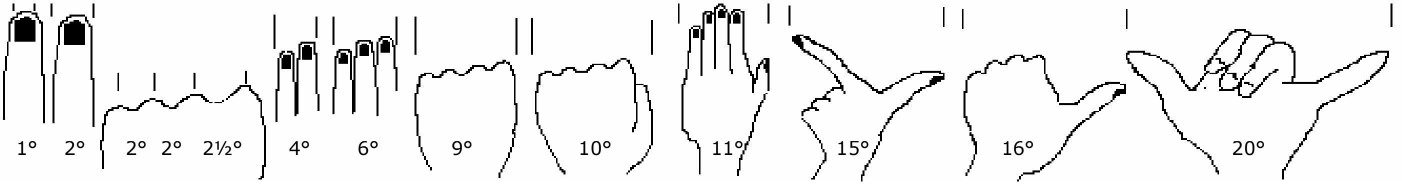

The 'Blink of eye' method also yields 6° — point to something with your finger at arm's length or with something narrower, with one eye, and blink.
Larger horizontal angles can be visualised from between your feet of from overhead.

Initial Orientation is the Single Fundamental — The Get-Back Path

"Which way do I go?" is more important than "Where am I?/How far?/Where is North?"
"How do I get back to a familiar path home?" is the real need. That means you need to orient yourself **before** getting lost.
The single most important direction is **Start to Finish** i.e. Which way did you **leave** the car park? **From** what recognised point?
"Which way have I been **going**?" (if I had to backtrack)
Which direction did you **head off** in — **toward** what identifiable point?
E.g. "Downhill". The direction need not be horizontal. We function on 'paths' more than on straight lines.
E.g. "to the lake". The direction need not be relative to North; It does need to be something you can reverse.
E.g. "with the Moon ahead on the left" "the long way around" "to the opposite side" "away from the mid-morning Sun"

Taking a map and compass is good; A watch is nice; But the only really necessary bit of navigation you must take with you on a walk, to get you back when lost or disoriented, is... what direction you have travelled out.

Initial Orientation is the Single King Principle — and the rest may not be much help without it. See the mnemonic p. 111c for some help.
Which way you went is also the single biggest help for searchers—**tell someone** where you are going!

How I Manage to Get Lost or Disoriented

… **starting off** without paying attention … blindly following the track or the leader … taking a minor junction
… choosing a hard route — thick vegetation … going without a compass, or map … trapped by a peak or clearing (p.45b)
… falling asleep during a journey — a gap in my reckoning … waking up in a new place … not noticing the "end of track" marker
… inside a hospital complex, or a shopping centre; coming out … in look-alike landscape e.g. suburbs; losing the car on city streets.
… getting out of bed the wrong way or entering a town by a last-minute bend … talking; inattention; not noticing bends & junctions; losing track
… misinterpreting "directions"; forgetting them; being given wrong directions … hurrying; "short-cuts"; leaving the track
… on river-flats near river junctions and bends … preoccupied — with map-making! or any other task
… inadequate memory e.g. in caves, the way back, the many bends … stubborn psychological disorientation syndrome
… miscommunication between parties e.g. about how & when to meet … confusion; blunders (see list p. 39b)
… not enough knowledge of what to look for & how to interpret what you do see … visiting the Northern Hemisphere
… minimal input technique of navigation e.g. 'just follow the signs'; overconfidence … misinterpreting the map—especially the scale
… new tracks, not on a map … false identification of landmark features … losing confidence just short of my goal
… unsure of distance covered … navigating by creek junctions or minor dirt roads … blinkered by fog, darkness or thick forest
… wrong turnings; walking in an unintentional circle; overlooking junctions; not looking back … waking up on a dark night, 180° disoriented.
… too much new territory too quickly … negotiating thick bush without keeping orientation

---

Voice Your Thoughts — Don't Suppress Important Clues

When your silent auto-pilot notices something and tries to tell you… Say it out loud. That's the trick to it.
Tell someone. If you are alone, yes, talk to yourself. Bring it to light.
You can't remember the jigsaw bits later, unless you look at them now! **Subconscious thoughts flit away like dreams**, so express them first.

"Oh I didn't expect to reach this so soon!" ? Misidentified
"North seems to have changed. It shouldn't be out there!" Disoriented!
"Hasn't the wind swung? Does anybody agree?" ? Are the others asleep!!
"The shape's not quite right" Most probably because it's wrong!
"The markers have stopped" ? Off the main track

The same goes for run-of-the-mill clues. A forgotten clue is a missing jigsaw piece when you need it, so say them all out loud.
"We're crossing a divide — see — we are now going downhill"
"Creek number... four" "Marker!... Blaze!... Cut Branch!..."

Back-track in Your Mind before You Back-track on Your Steps

Lost?
"Don't Be Afraid to Think"
Spend time rethinking.
Try to reconstruct where you might have erred. Have I overrun my goal, or am I not there yet? Or am I simply off course to left or to right?
Did anything 'not make sense' on the way out?

- Work backwards from now, before you forget.
- Assess how successful back-tracking might prove.
- Don't ask "Where do you think…" but shoot for the details — ask "Why do you think…"
- Each of you should draw a mud-map separately, then come together, to compare recollections. Argue out the sequence.
- Back-track on your presumptions.

Detail the facts with a cool head and many problems will dissolve.

Distance using Apparent Size & 6°

The visual size of people, cars, houses, as measured at arm's length, varies with distance. You know how big they are already.
Blink your eye or otherwise guesstimate what fraction of 6° it makes, and multiply its size up by 10, plus for the fraction, to get the distance to it.
E.g. something one knuckle space wide is 3 times further away than if it were 6° big, and therefore is 30 times further away than its absolute size.

Visual Angle, Size and Distance

Find the angle, and multiply the size up by 60 in 1-per-degree. E.g. If something 1m wide looks 5° across, = it is 12m away. 0.5° means 120m.
Find the distance from the map, and work out the size from the apparent angle, by 1 in 60. Something 10km away, 10° wide is 10/6 km across.
Using 7/4 in 100, for accuracy because the angle is small, something 10km away and 10° across should be 1750m wide.

Side-wise Offset

Blink your eye at something in the distance, and it will shift across the background skyline.
Walk sideways, pacing out the distance, until it lines up with the new skyline mark. Multiply by 10 for the distance to the something.
Alternatively, mark the new position alongside the object, which the 'blink' indicates, then walk up to it and pace out the offset distance & ten-tuple

Don't Just 'Get By' and 'Make Do'

We all have a handful of techniques we feel comfortable with. We all get lost.

Most "How to" books necessarily are telling you how to battle uphill — against your natural incompetence.
This is doubly true for a "How to Improve" (your navigation) book!
E.g. 'Improve your memory skills'! Easily said; hard to digest.
Get used to the idea, and therefore embrace it — the idea that you, as a good navigator, should want to extend your familiar range.

Navigation is inherently a problem (of not being able to see around corners; of not having a magnetic sense; of not having excellent memory; etc.).
However humans inherently do have the potential skills to cope. But do you have the guts?

Memories are Made of Landscapes, Not Compass Needles!

What to look for is detailed later, but for now consider...
How to 'look' …

Use your compass and map vaguely! Approximately! Lightly!
It is safer to be aware of their meaning, rather than to be accurate but disoriented.

Imagine you had no map, no compass. No signposts. No highways.
"How did the pioneers do it!?" You learn how to not get lost; how to 'see' what you are looking at.

Once you get oriented, then you can bury yourself in the instructions on decimal places, for accuracy when you need it.
Otherwise numbers and recipes have an ability to preoccupy your mind and numb your thinking.

Treat your map and compass as shy — self-deprecating — and turn your piercing stare away from them to what they are pointing out to you.

You Need a Road-Map To Climb a Hill!

It is a waste of effort to realise your mistakes at the top of a hill.
"I wish we had brought binoculars" is all too common a regret.
"We must remember next time to bring all our maps"

You have few such opportunities to positively identify landmarks,
and so you should **take the chance to familiarise yourself with what each hill looks like from every other hill.**

But often, it is only your smallest-scale map — the roadmap — which will show you the big layout.

"Didn't anybody bring the compass??"

It's up to you to plan ahead, to plan on making the most of your efforts.

'Land to Sky — Sky to Land' Navigation

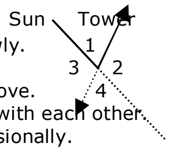

Step Zero: **Leave yourself out**, and your compass, of the scheme of things — so that the whole lifeless picture only changes slowly.
The result is simplicity itself — You see **how the Sun shines across the landscape**. Remember the angles.
Step One: "The Tower is \↗ by angle, from the Sun as I see it" Both change only slowly (or rather, seem to) as you move.
The Big Bonus is: You can keep track of all your bends, twists & turns, by the Sun & Land working in conjunction with each other.
You can locate a landmark easily through the vegetation — you know where to expect it when it shows itself occasionally.
i.e. You can always easily locate the tower/peak/etc, from kilometre to kilometre, in relation to the (now slightly shifted) Sun (or Moon, wind, cloud)
Step Two: "The Tower is ↖ from my shadow" This allows you to walk away from the Sun, yet still stay oriented to the landmark.
Step Three: "**The Anti-Tower** is \ ..from the Sun" These allow you to walk away from the landmark,
Step Four: ↙ ..from my shadow" yet still "see" where it is.
Step Five: Mnemonic: "The Landscape on the **Left** Lingers Longest in the Light" — its apparent movement as you walk, follows the Sun & the Shadow's anticlockwise shift. (SH only) i.e. you don't need to worry much about time passing, and angles changing.
"But Landscape on the Right… soon goes wrong" — it 'moves' against the clock-sense of the Sun. (Southern Hemisphere rules)
i.e. you need to **refresh your angles** more often during the day. (NH: Left Loses the Light; Right is all right)
Step Six: Rule of Thumb: "**15 km away** changes at 15°/hour (i.e. at about the same rate as a low midday Sun) as you walk past it at **4 kph**".
The Sun shifts at less than 15°/hr when it is rising or setting — so try "20 km away/12°" to match it; and "10 km/25°" if the Sun is very high.
Yes, you should convert the angle to "ESE" or rename it "just S. of sunrise" for later, but **the idea is to use the Sun as it is now, not North.**

Grid-line Walk — Explore in a Square Grid Pattern

This only adds 20% walking effort, but lets you know where you are, and how to get back,
and along a route of your choice, among many choices.
You need to judge a given distance, well, e.g. 100 double-paces; one minute; or 100m; or by pedometer.
Choose a **reference direction** as 'Up' or 'ahead'. Sun-North; Compass Needle North.
Up-weather by clouds; down–clouds. A strong wind. Distant Hill. If needed, keep a straight line by marks.
Choose a **scale** to suit your ambitions, e.g. 100m squares (so that each 10m is an extra 0.1).
Try to cover a whole square at a time. Use 0.5, if needed, or you can cross at a diagonal, 140m long.
Count **forward and back progress** in whole chunks, as A-Z. G.5 means 7½ squares.
After Z comes AA, AB, AC. See p84b for making suitable mnemonics. **Don't lose count!** That's the worst of it.
Count **right and left progress** as +/- 0-10, etc. Minus simply means 'backwards' or 'to the left' of the reference.
When you have definitely decided to head back to base, remember the grid-reference. Stop counting.
Reset the count to Zero. This is easier than trying to count backwards, using subtraction, to get back.
Reverse the reference direction, e.g. to downwind, shadow South, compass South, and continue.
When you arrive at the original grid-count, you should be within sight of home-base.

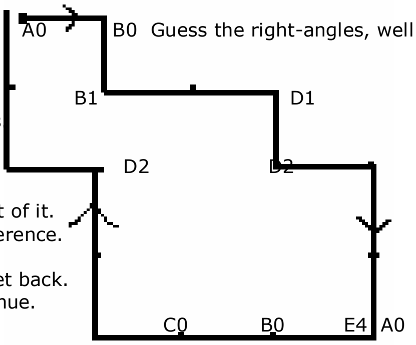

Guess the right-angles, well

Disorientation is More Dangerous than Being Lost

…because the same disorientation that will get you lost, may then get you more lost.
Orientation is primarily a mental picture — therefore **— Disorientation is a wrong mental picture.** It is one and the same principle.
So Disorientation really extends to cover more than direction sense, into bewilderment and logical debilitation!
You can be disoriented, deluded, bewildered and irrational about more than direction or position. E.g. You may misjudge which way is up/out/back.
E.g. you may be mistaken about the time of day, or the existence of a road, or whether the promised weather front has passed you or not.

'Lost' is 'I don't know where I am' which is different from disorientation, since position sense is independent of direction sense.
Again, you can know where you are on the map, but still be bewildered on the ground — not 'fitting in' even if direction-wise oriented.
Then again, you can be oriented, but lost — yet not necessarily worried or in difficulties — because you are oriented.
You can be disoriented unawares — even when you know where you are — i.e. not necessarily aware of the danger of loss of direction sense.
To be both disoriented and lost is most dangerous, but **it is orientation which stops you from getting lost**.

Orientation keeps your finger on the pulse. It stops you getting disoriented. In many ways it gives you a position sense too e.g. that you are in a particular catchment, downwind of a known landmark, up-Sun of the overall course. Orientation of the map puts the map onto the landscape—aligned to reality—so that the map can point out its riches to you. Orientation lets you know where you are, recognise places, & know the way back

---

Logical Blunders — Examples

Mistakes of category: "The last stream was flowing to the left" (Oops!)
"We have crossed the main road" (Now, the main access road is dirt, whereas the main through road is bitumen,
but you muddle them up. Or maybe you are unaware of a fork in the main road.)
Mistakes of number: "We have passed two lakes" (which were joined behind the intervening vegetation).
Mistakes of quantity: "We haven't changed direction" (except for one hairpin bend!)
Ironstone may have swung the compass, when you put the map and compass on the ground 'for extra certainty'!
Mistakes of quality: "From the peak to the lake was 090°" — but the false peak you were on, gave you a false sense of accuracy, in the mist!
"See those houses 'across' the river?" — Rather, see the bend on the left — they are on this side!
"The island should poke out from behind the far point. Why isn't it there?" (It is 'the far point', of course)
"But we have crossed the River" No, only a main tributary, or only onto an island.
Gaps in knowledge: "We haven't seen the junction" — because you short-cut past it; around one bush is enough!
Unexpected: E.g. there are sometimes two parallel roads — the old & the new. E.g. a stream sometimes goes underground "Mole Creek".
E.g. A water canal may cross catchments. E.g. "Remember the creek junction at the base of the hill" — The 'hill' was only an island.
Assumption: "That will be the ridge we want up ahead, let's turn off here" — Unproven!
"Ah, we are back to the road!" Yes, but not the same road!

Trust the Untrustworthy

It's like people. 'You can't trust anybody', but you have to trust your life to them sometimes.

A compass has dozens of potential faults, but in a fog, you trust it, implicitly.
You yourself are less trustworthy than a compass.
When all you have outside your own opinions are reasonably reliable external clues (like where the Sun is!) you had better trust your life to them, not to yourself.
For instance, sometimes that 'external clue' is someone else telling you that you are disoriented, when you don't think so — you had better listen!

A disoriented person may not want to trust the 'untrustworthy'. Don't let that be you.
E.g. You may find yourself twisting the map to make it line up with a mountain which doesn't really line up, except to your expectations.
You prefer to believe that the map may be distorted!
It takes practice to believe. I once spent a whole week wondering why all the trees were pointing too much to the sunset side of noon.
I was trusting myself, not the trees, and was simply wrong about where North was. I didn't realise which was the more trustworthy.

Lost? Rule 2: Start! Getting Unlost. Find yourself. Regain confidence.

(After you STOP getting lost — Rule 1) It is time to **START NAVIGATING** — Rule 2. (You can practise this ahead of time, without getting lost.)
E.g. What can I see & recognise? E.g. Can I interpret the wind or sky or trees or drainage for direction?

- First: Mark your spot e.g. tie up a hanky. Mark your Trail—so you can get back here, as 'base'. Retrace your steps, if that is the best option.

Act out a Search & Rescue, looking for your 'lost' starting point. (If you can see your surroundings, you don't have to feel 'lost'.)...

- Explore in all directions systematically.
- Make a map. 'You' in the centre. Locate yourself in the landscape. Climb a high point, to orient yourself; at night too, for lights.

Let others find you...

- Leave notes about your plans. Especially at 'findable' places e.g. a hill top.
- Leave clues so that searchers will not give up for 'lack of clues'. Make some marks visible from the air.

Search out a track...
Ask: "Where should there be a track?" E.g. to... that peak?, a fishing spot?, any low pass through the range?, the main access valley?

Mental Grid Map

This is the square grid method, where you simply walk, along a path of choice, and rather than following the grid lines,
guesstimate your decimal position in the grid scheme.
The beauty of it is that it constantly forces you to orient yourself to a reference *direction*, to notice *distance*, and to be aware of *position*,
yet it takes nothing more than mental equipment. You should aim to practice it properly, once, and you may not want to discard it.
It causes less mental strain if there are two or more of you. **Delegate** the left/right to one person; the forward/back to another.
The one person guesstimates distances by one method, their way; The other gauges perhaps a different distance and with different accuracy.
It doesn't really matter if your guesstimation mismatches, provided you remain consistent and imagine a rectangular grid.
In addition you need a feel for triangle lengths, for a 1 in 10 slope, up to a 10 in 10 slope, so you can guesstimate decimals.
~You need to divert 5 in 10 to add more than **10%** to the length walked. E.g. from Z0 to A0.5 takes 112m across 100m squares.
~Maximum extra length is a 45 degree diagonal, at **141%**. E.g. from Z0 to A1 takes 141m. These two set the feel for guessing distances
1 in 9 adds one third. For 6 in 10 to 8 in 10, triple the first figure as the extra %. E.g. from A0 to B0.7 takes 122m.
If in doubt draw a 10x10 square, draw in the diagonals to each of 18 opposite points, and label the lengths as the square root of $100+n^2$
This lets you see the direction of each n in 10 slope, and know its length.
~You could use degrees, up to 45, and call them percent extra distance, with fair accuracy, but there's no need for such figures.
When I simply guessed, by eyeballing distances and directions, it brought me back home without noticeable error, after a few kilometres.

Walk Where the Navigators Talk — if you want to learn navigation

Not out of earshot.
Navigation is a learning experience.
It works better when you can share the process of learning with someone.

Explore

Explore the unexpected — the side tracks. You walk to explore, usually, not to exercise.

It's No Fantasy

Navigational know-how is not "If I were lost I could do this…" but is "What I am doing right now, all the time, will stop me getting lost".
Learning new skills and how to interpret all the clues is the practice of navigation here and now.
Continuously pointing out new clues to yourself, and to others willing to listen, is the process of not losing yourself.

Feed Your Subconscious through Your Conscious Focus

E.g. To improve distance guesstimation, focus on the finest visible distant detail. Notice the haze level, the shadow hue, the sizes you can see.

To train your direction 'sense' (which is entirely lacking in humans) feed it all the clues you can —
Sun & shadows, wind & weather, compass & landmarks, trends & bends… Don't just look at them! Put words and meanings onto what you see.

To extend your memory of a sequence of turnings… use pencil and paper 'memory'.

A working rule of thumb: **Bring to conscious note**
what you hope to keep track of subconsciously.
Don't expect to be able to remember what you haven't noticed or thought about consciously.
So learn what to look for, and how to see it.

Look into the Far Distance

..for Trig. Points; Towers. They are there to be seen, by definition.
..at night, for City glow; Auroral Lights
..for Roads; Houses
..for the dip and strike of the underlying geology
Assess the difficulties ahead, whenever you glimpse them e.g. a steep bank on the other side of a lake. Use binoculars.

Look through the Haze

What are you looking for? You want to pick: intervening ridges, hidden gorges, distant cliffs, overhangs, false peaks, mountains or islands, the Moon, and the trail ahead. Those cloudy lumps are not always clouds. Those invisible peaks are there to be found.

If you don't see them right away… Wait ..it is the **changing light conditions** which show up the contours.
E.g. Look at night, when the whole emphasis shifts, and light sources show up.
E.g. Wait for the haze to shift — it moves with the Sun.
E.g. Wait for the haze to thicken — to show up the valleys — weather is changeable
A rain shower may show up a valley; a cold front might clear the air of haze.
E.g. Wait for a cloud shadow or patch of sunlight to drift across — it will disappear into any hollows, and highlight any foreground.

Join the Invisible Big Picture To the Visible Landscape

The Big Picture might be invisible — e.g. the South Pole Star is "Over the South Pole in Antarctica" — but geographic "South" is quite different and is visible — "that knoll there!". Join the two!
The first principle to practise, is to work top-down; but the second and more important need, is to integrate everything.

When you put a continental or global picture onto the landscape, you have a full-scale map of the world in front of you.
e.g. Point just beneath 'that Knoll' and put your imagination into gear...Down over that horizon are icebergs separating from glaciers with a huge splash! Can't you feel the chill? Up there is the night-time aurora. Down there, anti-polar bears look straight up and see the Southern Cross whirling around overhead! And we must be on the south side — the cold dark side — of this hill behind us.

You say: "The destination lies beyond that valley."
"The track winds around behind this hill."
"Mount Saddleback marks sunrise"
"The Sun skims those treetops"
"Porpoise Bay must be yonder; Melbourne that way; and my girlfriend, 50km in that direction…"
Integrate as you walk — in real time and place

Cyclical Motion Divided into 12ths

Any back-and-forth motions or cycles which are caused by circular orbits or sinusoidal oscillations are easily understood and followed *in units of 30°*.
This is because 30° of rotation will swing *half way across* a radius from top-dead-centre, or half-way up/down from right/left extremities.
E.g. quick rising for 4 twelfths, then *2 twelfths* in the *top half* of the swing, 4 quickly reversing itself, and wallowing for 2 at the bottom.
You may not *see* the 'circular' nature of a one-dimensional oscillation, yet it will still conform to this rule.

Tidal Height	Sun's Seasonal Latitude (or Moon's monthly position cycle; use 2½ days)
1 hour falling slowly from the top	1 month at Northern solstice slowly coming South
2 hours dropping rapidly to midway	2 months rapidly moving back to the Equator — to the South
2 hours dropping rapidly from midway	2 months rapidly moving South of the Equator
1 hour dropping slowly to minimum height	1 month slowly approaching Southern solstice
1 hour at bottom rising slowly	1 month slowly turning back to the North
2 hours rising rapidly to midway	2 months rapidly returning to the Equator — toward the North
2 hours rising rapidly from midway	2 months rapidly moving North of the Equator
1 hour rising slowly to top	1 month slowly approaching the Northern solstice

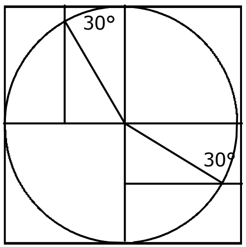

You can simplify that into sixths, saying: "1 sixth *at the extreme*; 2 *on each change*; 1 at the other extreme"
Or say: "*twice as long on the change*". That makes 'the change' sound 'slower' but it actually goes *faster* and *for longer*.

Which Way Did I Come In? Which Way Do I Go on?

Have you ever interrupted your pleasant journey at a roadside picnic area and driven out the wrong way — back the way you came — and not onwards? Or you turn left instead of right when you drive "out" of a local dead-end network, into another dead-end maze?
Or you sat down for a rest on the track and when it's time to go, both directions look identical to you?
It's embarrassing on the road. It's nerve-racking in the bush! You feel struck with insanity. But it's normal enough.

Or when you come back to visit a familiar place, it's on the 'wrong' side of the river! Every time you come back?!

An extra external clue, like the Sun, is needed, and needed to be believed, but it is really needed before you make a fool of yourself.

And you may have to undergo 're-orientation' to dis-illusion yourself.

Don't let it happen to you deep inside a cave system!

Logical Presumption Leads to Utter Delusion

Think about those example 'logical blunders', about your own fallibility, about the dependence of logic upon its basic assumptions.
Being twitted by a logical mistake leaves you with a horrible 'lost' feeling, plus a pathetic 'helplessness' about how to reorient.

These are not simply mistakes of guesstimation, subject to fine-tuning!
Once you subconsciously make a decision about how the world is, then your mind will close to the possibility of error.
False assumption is presumption, so I say "**Presumption Means Delusion**"
Such Left/Right, Yes/No, crossed/uncrossed, 2, or was it 1? -type mistakes lead to utter bewilderment, so I add
"and errors involving discrete categories can be especially indiscreet!"

Being aware of the phenomenon of logical blunders, is half the battle in overcoming them, when they bite.
Keep an Open Mind about the truth of your assumptions!

Trust Less and Prove More

Don't trust yourself At numbers
Or remembering accurately
At judging time or distance
Or walking in a straight line
At identifying landmarks properly,
Or walking back upstream, and so on.
Don't trust A compass to be dead accurate
A map to show North reliably
The weather forecast
Old track markers
A navigational 'fix'
Complicated or finely adjusted instruments. They can easily go out of adjustment, without letting you know.
The point is not so much to **double-check** and **cross-check**, as to: **back-up** with an alternative; **interpret** for sense; understand; and **be wary**.

The Simplest Way to Get Unlost is to Retrace Your Steps

The sooner you go back to where you went wrong, the sooner you can relax. Yes that might be uphill. Yes it might be a long way.
"Oh I wish I had retraced my steps earlier, while I had the chance!"
You still have the chance, you probably just don't want to.
For instance, if you have been pushing on, hoping to find the way, you should be doubly aware of exactly where you have come,
since you doubly anticipate the need to retrace, if you fail.
"Short" cuts have a habit of surprising you at just how long they can turn out to be, repeatedly.
A track naturally avoids gullies and obstacles, so a 'shortcut' is almost guaranteed to hit one square on!

Retracing your steps may be something you can't short-cut. It is the shortcut. It short-circuits many other possibilities.

Caution: Sometimes retracing your mistakes will only compound them—more of the same!
You may sort yourself out better if there is an alternative system to try e.g. to walk back upridge if you got lost going downstream.

Approximate Sine-Wave Values   97% accurate

(From school you could have memorised that sin 30° is 0.5 and sin 60° is 0.866, and perhaps sin 45° =0.707 ???)
The first twelfth of an oscillation takes something half-way (0.5) to its peak amplitude (1.0) from the middle ground (0), by the rule of twelfths.
~*The motion is virtually linear for that first 30 degrees*. So to convert those *first* 30 degrees to a sine-value, **Divide by 60**.
I.e. add at a rate of 0.5 in 30, or 1 in 60. E.g. 25°/60 = 5/12 (or 0.417). Sin 25° is actually 0.423, which means we had a –0.006 error.
Sin 15° is 0.259, but we get 0.25 (–.009 error). Sin 20° is 0.342, but we would guess 0.333. That's about the maximum error, 0.01 low
The next twelfth carries the motion from half-way to 86.6%, in a curve; The extra is 36.6% or a slope of 1.22 per degree.   Let's use 87,36,1.2
A linear fit cuts the curve twice, and so is close to most of it. If we could add 36%, over 30°, to 0.51, to arrive at 0.87, we get +/– 0.01
~To the ***extra*** degrees, **Add 20% ; Add 1; Call it a percentage** (divide by 100).   (*To add 20%, add 1 tenth, 2 times*, e.g. 55 + 5.5 + 5.5 = 66)
E.g. 60 degrees = 30 plus an extra 30; so start with 0.5 and add 0.3 then 0.03, 0.03, 0.01 to get 0.87, which is 0.004 too much.
E.g. 40 gives 0.5 + 0.1 + 0.02 +0.01 = 0.63, instead of 0.643, or a –0.013 error. Sin 45 is about 0.5 plus 0.15 + 0.04 =0.69 or –0.017
To make the next approximation, use 0.87 for the first 60 degrees, then worry about any extra degrees.
The last 12th is the flattish top of the curve, building up to the crest. It climbs only 13.6% from 0.866 to 1.00 in 30 degrees, in a curve.
We start at 0.87 and finish at 1.02 cut the curve twice and best-fit to within 2%.
~For the last 30° degrees **Divide by 2, then 100**, to gain 15%, from 0.87 to 1.02.
E.g. 80 degrees moves to 0.5 then to about 0.87, then an extra 20 degrees or 0.1, yielding 0.97, 0.015 below the real value, 0.985.
75 yields 0.5 + 0.37 + 0.075 = 0.945, 0.021 below the true value 0.966. 70 gives 0.87 + 0.05 = 0.92, 0.02 below 0.940. I.e. +/– 0.02

Learn Think Look Integrate Integrate

Self-Motivation is Fundamental

For a sense of direction to become “second nature” you need to **train** your second nature.
Improved navigational ability boils down to… *putting new ideas into practice*.
(Not perfecting old mistakes! Nor leaving the right technique on the shelf)

Don’t just know about a good idea; make it yours.
For this you need to want to improve your skills.
to force yourself to practise your weaker points.
to push yourself to develop new habits.
Which is fun.

Enjoy the trip!

How to Feed Your Brain

When and where does the Sun get up?
Get up and look (not at it!). Don’t make it an academic question.
You don’t need an academic answer, either. You measure it with your eyeball. Perhaps draw it.
That doesn’t mean “ignore the details” but “memorise the feel”.
You do ‘measure’ it, but not with numbers.

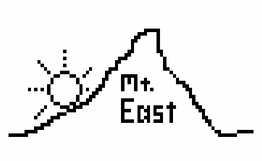

There! This early!

Star navigation? It doesn’t have to be a “trigonometrical nightmare”.
For us it means: Find the Southern Cross.
Guess the time of night.
Find the Pot. Get to know the stars.
Project their movement forward into the daytime.

Let your brain do the calculations. Show it what to work on.

Wait for a Change in Perspective...

..to reveal the hidden shapes and layout.

As you drive up, keep **re**-examining the changing skyline. Continue the process as you walk. Look for ridges & valleys showing up through the vegetation as well as on the skyline.
E.g. Use parallax to reveal a deeper horizon behind a hidden valley...
Keep your eye on particular bumps on the skyline.
If there is a deeper background, i.e. more distant, then a lump on it will show up by moving with you —
more quickly than a ‘foreground’ horizon. You may have to look through the deep haze to even locate particular bumps.

Climb a tree, or a hill.
Use binoculars. Look especially for the trail ahead, in relation to the terrain.

It is the **Changing Parallax** which **Shows Up the Landform** and its Relative Distances.

Landmark to Landmark Network Navigation

This is a permanent, unchanging, top-down, whole-region method, **prepared at home**, for a large area, and for particular walks.
1. Choose only **prominent landmarks** from the map. e.g. peaks with transmitters.
Add in major ports, mouths & coastal extremities, lighthouses, around the coast.
Draw the network of triangles between them. (“Which one am I inside?”)
Draw the circumscribed circle for each triangle. (“Am I inside or outside?”)
Label each line in **kilometres**. (“What does 20km look like?”)
Label each horizontal **angle** in degrees. (“Does it look larger (inside the circle) or smaller (outside)?”)
Label each line with its **trend**-direction. (“Am I to left or to right of **track**? Which is this river **parallel** to?)
On the right of a line you will be looking to the left of its trend to see one of the landmarks.
2. Keep a **notebook** for local peaks — smaller but locally prominent lookouts.
The idea here is to navigate only by landmarks, not by (favourite) low spots.
You don’t have to worry about whether one landmark is visible from another.
The idea is that **even when you can’t see a landmark** (as well as when you can),
you can know where it is and how it lies in the land layout, how other trends line up with it and with other visible lines, and *how you fit in*.

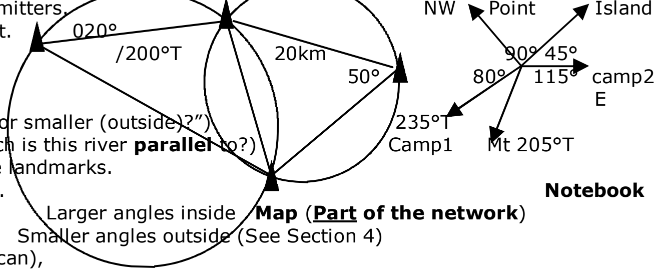

Name Each Clue — with solid ground

“That’s Pacific Sunlight shining” — a mid-morning Sun, for Australia
“I have an Anti-American Shadow at the moment” — pointing away from North America
“It’s a West Coast Cloud Flow” — you can’t lose your orientation
“I love a Tropical Moon” — when the Moon is North of you (SH)
“That wind would blow straight in our back door” — ‘Back-Door Blow’
“Downstream from here is uphill from Port Stevens” — ‘Steven’s Stream’
“This cliff sees the Southern Lights” — ‘Cold Cliff’
“Topple-Tree Trunk would take us to True North”

The next step is to add something in line with that — see the next page.

Disorientation — A Killer Syndrome

Pilots have been known to fly upside down, or rigidly leaning to one side according to their mistaken belief about "up"!
If even our natural senses are unreliable we certainly can't afford to believe in any 'sixth sense' of direction, or on any unproven magnetic inbuilt direction-sense in humans! A trudge can dull & distort our senses. Your mind tries valiantly to maintain mental orientation—that is the problem.
Several times each year pilots die, scuba divers get lost in sea caves, cavers fail to find the way up and out, boatmen falsely navigate the fog, and bushwalkers get lost in the bush, fatally. They were not just careless; they were probably trying hard, but in the wrong direction, stubbornly.
Belief is an irrational master. Its force is stronger than reason. The conflict, combined with physical tiredness and sensory deprivation, is stressful.

We prefer to get lost rather than to rethink wrong assumptions!
We become willfully blind to strong contrary clues in view (including signposts, instruments, buoy colours and numbers!).
We magnify weak clues, to justify our delusion. This heightens the stress, because our sub-subconscious knows that we are wrong.
We actually lie to ourselves about what the clues do say! (e.g. we change the number, colour, reading or wording in our mind!)

Such irrational behaviour is explained as defending our psyche from hurt pride and from being devastated by confusion, fear and panic.
Don't laugh it off as "Oh, that's only for the worst case" because it is the less obvious self-delusion which will likely trap you.
Being human includes you — you definitely are vulnerable to being even less sensible than trained pilots!

Casual Outings Catch You Out

..unprepared e.g. without a map or compass or weather forecast.
Casual outings also have a habit (because exploring is interesting) of "I'll go just a little bit further"; and turn into longer than expected full-scale safaris. Exploring is built into the genes of a small fraction of any animal population—it's a group-survival insurance. But don't let it catch you out.

Serious trips go more reliably because you know what is needed, but you rarely prepare well for a casual trip.
e.g. you have no pedometer or notebook and so don't keep a good enough log of distance or direction.
So **Lesson Zero: Make some precaution** (to avoid getting lost) **every time** you wander off.
Another unexpected casual mistake is to decide to leave the track without thinking what precautions may be required.
Try this. Look up and repeat after me: "God, have I forgotten anything? Is there anything I should do first?"
Let someone know the single biggest clue: "I'm going exploring that way",
but then don't forget the unobvious catch—that person might also go exploring & not tell anyone where you went.
(Being disabled by an injury is one thing, but dying of exposure or thirst or wild animals, simply because no-one knows where you are, is another whole added dimension—that is probably the strongest reason for bothering to let someone know.)
So before you go off, pass on any vital information about others, as well as about yourself.

Trust Yourself Less

..when the compass and you disagree, for example.
When the map shows the junction there, but you reckon it should be here, for some reason such as 'but we've come too far!'
Or similar cases, of confusion.

Your reasoning sense has holes in it, big enough to outsmart yourself.

In the end you will usually conclude "Well, I was wrong, due to this very good reason(!), and it was how the map and compass told me it was."
But you don't want to admit it, at the time, until you are proven wrong, yet again.
Your reasoning is not, despite your worship of it, the highest god in the Universe. You need to be able to criticise it freely.

Unreliable as map and compass can be, at times, it is better to trust them more and yourself less.

More is Better — Thought, Time Care

A simple principle to apply is to do a better job at what is marginal.

For instance, "flat" ground nevertheless does have a slope which you can determine with care.
The wind may be gentle but a little extra time could make the difference between success or failure at determining its direction.
You can always double-check a casual decision, e.g. about which way the dry creek-bed flows.
Instead of wandering in an unrecorded meandering fashion, you can mentally log your changes in direction by lining up pairs of marks.
Just before you leave the track, a quick checklist can bring a few necessaries to mind, but only if you take the trouble.

A little extra thought; an adequate time-allocation to each task; more faith; and that bit of extra care—will solve a lot of problems.

Moveable Position vs Fixed Orientation

• It helps to park the car in a cardinal direction, or pitch the tent North/South, since we all rely on such clues to revisualise what we left behind.

• But when you do try to orient yourself to something invisible such as how the car was parked when you walked off, there is a simple trap to avoid

Don't imagine the car/tent as 'behind' you and twist its imagined *orientation* to suit the Sun or compass;
Instead change its *position* — *move* it to somewhere other than 'behind'.
Mentally manipulate its placement, not its orientation only. Or else the 'correct' orientation will confuse you — nothing will make sense.

Don't let the Sun 'change its direction' — that is a tell-tale that you are keeping the reference position fixed *behind* you rather than *moving* it as you swing to left and right.
Because we have no good tools to log changes of position or distance, we tend to neglect that half of the story in favour of maintaining orientation.

• As another hint on how to use a fixed orientation, e.g. North-South, so as not to let it 'change' in your mind, imagine it as a long, huge, high, iron girder bridge overarching you and your car/campsite, in the appropriate direction. Whichever way you imagine it, e.g. through-bolted to the continent, or from one peak to another, make it so solidly fixed to the ground and impressively immovable that you feel so insignificant yourself that you feel *you* must *turn yourself around* in order to pass under it.

Routine Mind-Shut-Down Procedures

You know how you become dull and passive — when being over-taught, lectured at, just entertained, occupied in repetitive tasks, sitting in front of the television, failing to solve too-difficult maths problems, or being proved to be a perverse idiot against your own opinion…
Automatic self-defence from constant attack... involves shut-down — into a semi-hypnotic trance.

Similarly,
Walking becomes a "trudge" after about an hour without a break, depending on the load, the heat, and how many trudges have preceded on that day, and previous days.
Pain-relieving opiate levels rise. Conversation takes a dive. The path sliding under your feet gets stared at a lot.
Your mind is passive, plus your body is on auto. It is physiologically involved in the hypnotism.

Mental Torpor is a navigator's constant enemy.
Keep the Different-to-Same-Ratio up, for both yourself and your companions, so that you can stimulate each other into full consciousness!

Feed Your Subconscious Subconsciously

E.g. Open your Atlas to a map of the world. Orient it and leave it open, oriented, and in view.
E.g. Place a magnetic compass somewhere in view at home.
E.g. Develop a habit to Orient your watch N-S each night when you take it off, wherever you happen to sleep.
E.g. Sleep with your head to the north each night if possible.
E.g. Clearly visualise what direction your feet are facing before going to sleep each night.
E.g. Imagine your body lying across a map of the continent.

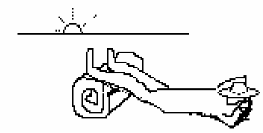

New Habits Open up New Horizons.

"Retracing is Re-preciating"

..This memory-aid is something you have to tell yourself — to avoid blinding yourself and missing the goodness of the walk back.

If you are only titillating your senses for the passing instant, by constantly seeking "new" experiences, you will develop the mind-deadly
"Bin there — Dun that" attitude
Familiarity breeds contempt… "I know the way" doesn't guarantee that you know much else about the way.
The return journey degenerates into a dull, silent, often over-fast, boring, mindless trudge, even if you drove all day to come and see this top-quality location! In effect, you have walked the track twice, but looked at it only once.

Two solutions:
Take an interest in your surroundings, to overcome passive mindlessness.
Questions on your mind overcome boredom. If you have asked questions about a new place, you should be able to answer them on the way back.

Remind yourself of the opportunity to "re-preciate".

Aim to Integrate…

- ..Integrate daytime clues with night-time orientation—use the Moon, the place of sunset, the sunset glow later (which is further South—SH), the twilight Blue Band, Venus, the Ecliptic, the skyline — anything to carry you over from daytime to night-time, during that awkward twilight changeover phase.
- ..Integrate invisible landscape with visible landscape—work out what lies behind each hill, to picture the outside world against that skyline.
- ..Read the "map check-list" and specify what each item means from your point of view—if you didn't have a map, for example.
- ..Integrate the changing, with the unchangeable. E.g. the Sun's daily rotation needs to be followed, by naming its current direction. Same with the track; and your nose; and the wind.
- ..Integrate one walk with another — don't just carry two separate mental maps, add a third overall view.
- ..Integrate each system with each other one. E.g. don't let the map be one "world", while the walk is another, and the night sky a third. Join them into one whole.
- ..Integrate each direction clue with each other direction clue — E.g. "The tower is just before magnetic noon from here" (3 pictures in 1)
  E.g. the boat-radar may fail in the heaviest weather, unable to penetrate the rain squalls—you need a boat-compass too.

It is the **combination** of clues which stops you getting disoriented — if you mistake one, the others will alert you.

Notice and Match Up

'Notice the clues' is the easy half of navigation.
Sometimes it is as simple as seeing that the track is curved.
However, a slight curve due to a river bend can turn you in a hairpin without you realising, even if you *do* notice the curve.
Until you also notice the *extent* of the curve, you are at the mercy of disorientation. Something more than looking is needed.
You need something to orient to, like the wind or clouds or landscape.
Again, it does you little good to have *seen* the clouds steaming by if you forgot to see where they came from.
'Notice *and Interpret*' is the catch-cry, which never stops.
But when the wind stops, will you know what direction you have been going? And 'heading for the mountain' is fine, unless you lose sight of it.
Something more than single-clue orientation and interpretation is needed.

Do Notice; Do Interpret; Do Orient; Then ***Match up***; Link it to something else. And to something else. E.g. the clouds are from the mid-morning Sun; The mountain is on my shadow's port bow; The river has bent from 'towards the coast' around to 'away from the coast'.
For special examples, see '3 in line', opposite

Un-Dis-Oriented

You don't need to be perfectly and continuously oriented!
In fact it is quite hard to maintain an exact direction-sense, as soon as you turn your head, move a few steps, divert a few degrees, etc.**
You only need to be realistically confident of position & direction.
Hence **it is better to be roughly aware of many clues, than to aim for perfection, say, in location or direction**.
E.g. 'Moon North' is approximate, +/-60°, but awfully comforting! — It adds to a complete picture.
E.g. You may not even look at your map again, after studying it once, if by now you already know what it is telling you.
Don't aim to be perfectly oriented — just undisoriented.

E.g. Hold the conversation on how to judge range; what's behind the hill; where to find moss; why you can't yet interpret lichens; and so on… and you can hardly miss seeing the walk.
E.g. Run through the Start—Stop—Rest Mnemonic, p.110, as a memory challenge to all… "What's next in it?"… which has the benefit of practising all its clues too.
**As an exercise, guess, from a distance, diagonally, where the side of your house or car points to on the horizon, then go and line it up to find out Or guess which point is opposite that hillock or tree, then establish how accurate your guess was. My results: +/– 10°; +/– 30° when not trying.

Trapped! By Not Looking Back

A fish funnelled into a fish trap may never again find the narrow entrance hidden somewhere in the large perimeter.
Converging tracks and sudden clearings trap bushwalkers too, when they walk through an irreversible change without noticing.
"Where did our small track join onto the main route?" "Where did we hit the fence/road/pipeline/track, and turn along it?"
E.g. stepping into the creekbed, to explore upstream, is a one-way move—you may not recognise it coming back, and could overshoot.

- **Beware of all sudden clearings**.
  "Where did we emerge from the bushes: out into the clearing, onto the peak, onto the riverbank, out onto the beach"?
  Look for footprints — anyone's, but especially your own.
- Remember the phenomenon — and mark those junctions in your mind! Or even mark the junction on the ground. E.g. place a rock there.
- "Which tributary did we follow down?"
  "Which ridge did we come up on?" More on p.63b
- Log the distance and directions involved at such changes/junctions/clearings, before you walk on and 'lose track'.

Disoriented? Stop. Say So. Reorient!

Be prepared to voluntarily switch from 'certainty' to uncertainty.
Learn to live with fallibility.
As soon as you suspect a disorientation episode, admit it. Say so, out loud. Don't miss the chance.
Tell your brain off!
"You know you're just guessing! You made some mistake and won't admit it! Let's work this thing out together"

If you don't know that you are disoriented, that can be worse than being lost but oriented.
You don't even have to be lost — Even if you know where you are now, if you are disoriented, you may walk on into trouble, rather than be looking to walk out of trouble.

Recall how often you have experienced a hard-to-shake feeling of disorientation.

Lost? Rule 3: Do a Good Job — of walking out or staying put

It may not be safer to stay. Well, if you decide not to stay, don't be aimless in what you do decide to do…

- E.g. Retrace your steps.
- Keep a straight course — e.g. downstream; along the road; toward the mountain; to night lights; to the traffic noise; a chosen direction
- Be single-minded — purposeful, but not stubborn
- Leave a Trail — you may need to retrace. Yes, it consumes valuable time, but "don't get more lost"
- Leave clues; notes — for others
- Walk quickly if you must, to cover sufficient ground, but 'rescuers' may not be able to keep up!
- Yes, your exit strategy may need to be aborted, but at least nominate an exit strategy for the exit strategy beforetime.

In-Line Integration

Put your preparation into something useful by using the '**3-or-more-in-line**' interpretation method…

"The wind is blowing *from* Port Bluster, *over* Hot Hill, *past* the carpark, *to* us, and on *towards* the Sun.
"The Pacific Moon is shining parallel to the highway"
"The Mountain has Montague behind it, that cleared paddock in front of it, and that hillock in the opposite direction"
"The compass needle points past the Northern airport to the mainland"
"My anti-American shadow is pointing past Penstock to Point West at the moment"
"The stream is flowing from Mt. Big, down over Lake Low towards Mersey Mouth"
"That Tropical Sun is over New Guinea, shining Past Alice Springs, over the intersection to me, and into the teeth of the wind"
"The Nor'Easter clouds are flowing from Noumea, over the Main range, to here"
"That persistent pigeon is cooing from behind camp"

The effect is to turn your reference orientation directions into fixed tracks (see p71bi), making interpretation easy.

Navigation is Mental Work — Stay Alert

The tools to work with may be “map and compass” “line and direction” “time and distance”

But any workman needs, and needs to practise…

Attention Observation Analysis Interpretation Preparation Concentration Memory Visualisation

Specialist knowledge Deft application Reliability Right attitudes A store of safe and useful habits Coordination

In other words: **you get lost by lapses in applying your skills, not through lack of skills**.

Mental Equipment is Very Light to Carry — You already have it; Don’t forget to use it!

Instruments Instruct the Intelligent

Don’t let the instruments — compass, clock, protractor, map — do the navigating for you.
Rather, let them teach you, how to get by without them.

E.g. practise finding North by everything except the compass; then check your guesses with the compass. Try bush North repeatedly.
E.g. Guess at “adjusted Moon North”, then consult the compass.
E.g. with a watch you can practise telling the time by the Moon.
E.g. before you use a protractor, guess the angle.
E.g. gauge the distance by eye, then check it on the map.
E.g. attach a sundial compass to your magnetic compass, for the sole purpose of practising & checking your use of the sundial.
In general, predict and guess before observing, and learn from the results.
You should **always check** an observation for good sense, and check any instrument, or any calculation — but prediction forces you to think **first** (rather than the go-to-sleep-first alternative — which is to always check afterwards).
Control your instruments — don’t put yourself onto auto-pilot — or you will hypnotise yourself into blindness.

Look Back Regularly

Look At how you got here — so that you can recognise the way back if you need it — and can make sense of your walk.
At the change from one terrain type to another i.e. at the link between natural sections of a walk — after you walk through the change.
At the directionality of the sky haze down low — the change in thickness and colour.
At the colours of the leaves, lichens & flowers — they change dramatically upsun/downsun, and therefore forwards/backwards.
Look Back… At tracks joining in at an acute angle — which otherwise you won’t see until you come back — which will confuse you.
At the bending of the track.
At what you may have dropped or left behind! (Check each other. Anyone carrying loose items should walk in front.)
At the distance come, compared to the distance yet to go.
At the weather clouds stalking up on you from behind.
Can you see your starting point, and get a bearing on it? What is opposite that — i.e. the landmark which now shows you your exact trend to date?

Look Around

Since turning your head while concentrating forwards, taking your eye off the ball or turning a smooth corner are all enough to thwart your direction sense, the anti-dote is to deliberately turn your head! Keep looking around at the wider context, including behind, so you don’t lose sight of it.

Imagine an Aerial Photograph of the Continent

Visualise exaggerated shadows lying parallel across your imagined photograph — connect those with your shadow and tree shadows.
Visualise the Sun shining like a flashlight upon your imagined continent map — connect that Sun with the Sun you see (Don’t look at it!).
Visualise the cloud shadows moving across it — there’s one! See?
Draw the isobars across it, lined up with the low clouds wind direction — feel it in your face.
Superimpose a huge compass needle across the continent map — look at your needle. You should be aligned.

Usually you avoid using any small-scale map, since it is the real and full-scale world you want to see or visualise.
But sometimes, imagining a map helps, e.g.…

1. Your navigational map (i.e. remember what you studied!)
2. The mental continent glued to your compass needle, to help you ‘(not) swing the map’
3. The continental map you walk over at bends, to help you take in the new orientation
4. The bird’s eye/satellite view e.g. when reconstructing where you might have gone wrong and when re-enacting it in miniature
5. Sleeping on a map, with your atlas open, and your compass needle showing
6. The Road Map marked out in Regional Triangles from peak to peak.

Making a M.U.D. M.A.P.-m.a.p.-m.a.p. For a more accurate job, see Mapmaking p101c,103c,106ab,109c.

Mudmap your trip onto paper, because paper remembers better than you do. Use any scrap of paper. Take a pencil and draw you meanderings.
Orient the paper to your expected trend (p51a Rhumb Line & p111c S.T.A.R.T.), so as to give you most room to map your journey.
Mark a dot where you start from — probably on the edge closest to you.
Up-arrow for a reference **D**irection; draw in a long one, Up-Sun is good. e.g. Up-wind; Up-shadow; Up-peak to a prominent landmark; Up-Weather **D**irection. (the Cloud flow); Up-North, by compass needle, or Up-Moon. *Interpret* the reference-orientation-**D**irection(s) before you leave.
U-p**D**-ate the arrow direction(s) or interpretation as needed, e.g. as the Sun shifts slowly, in relation to the wind, or as you walk past the peak.
Maintain a trend, a straight-line ‘leg’ p73b. Use the reference direction if possible. E.g. use the anti-peak.
Or pick a distant mark and walk up to it somehow; Or line up two points and keep them in line (p85e,87e,89e); Or just guess.
Align the **A**rrow when you stop (See Map Alignment 97a-105a, 103b). Turn **A**round to see where you have come from, to map the directions well.
If all other direction references fail you, **Align the next leg carefully to the last leg**. Be careful, because this will accumulate errors.
A-lignment is the magic ingredient. But it is the extra reference A-rrow which keeps you straight and avoids those accumulated errors.
Interpret the new alignment before you disorient yourself: “Where then is the starting-point now?” — or else you have a map but no sense.
Don’t forget to use the map on the way back keeping track of the bends and where the start point is, and the finish point.
Pencil in the **P**rogress **P**arallel to the last leg; Pencil in the last leg *distance*. Turn around again and pencil in the *next* leg *direction.*
Guess at the distance, or pace it as you go, or use a pedometer, or a **stopwatch**. Add helpful notes as needed — p95c, 122 might help.
Repeat the m.a.p. for each new leg.

Avoid Disorientation: OverNavigate

Your aim is surely not: 'how to read a compass' etc. But: 'how to avoid a disorientation episode'
Being 'lost' is no panic, although you try to avoid that. Being 'disoriented' is panic (even when there is no real need).
To avoid the psychological syndrome and the real dangers of being disoriented, and lost, the most sensible attitude is…

Avoid getting disoriented in the first place — Maintain orientation
Navigate carefully e.g. learn 'how to read your compass' etc.! — Overnavigate — don't rely on one theory, one skill, one faculty
Don't give in to panic, or even to anxiety.

Over-navigating when you can, makes it easier when you can't, e.g. sea-sick or injured, bushed-in, exhausted or confused.

It is a rule in life (Murphy's Law) that things go wrong very quickly, not just one thing at a time, but all at once, compounding the difficulty.
As soon as you realise you are disoriented, you'll probably also realise that you also don't know the time, the position of the sun, the direction of the wind, the direction of the clouds, or whether you can interpret the bushes for noon. Then it begins to rain, go dark, and out come the leeches, and you sprain your ankle too. You could learn this law the hard way, or take my advice and do your 'emergency' navigating before the emergency!

How to Get Lost — Follow the Track— A Track is a Trap-Door

A track will lure you into the middle of the wilds. You end up depending upon it completely.
I suppose tracks know where they are going, but you may not, and they can't think. They are mute guides of the blind.

Antidote: Over-navigate. Do more than you need to. **Don't just follow the track mindlessly. Don't let it do the navigating for you.**

Keep your finger on the pulse e.g. How far have I come? What speed am I averaging? What landmarks can I see?
Have I checked the compass lately? The weather?
STAY ALERT
STAY SKEPTICAL

Reorientation; And a Fresh Start

Step 0. Admit it: "I am dis-oriented!". Say So. Stop. Decide to 'reorient'. Discuss it out loud.
Step 1. RETHINK where you might have gone wrong.
2. IMAGINE the true situation.
3. IMAGINE A BIRD'S EYE VIEW of how it 'must' be. Lay it all out, on the ground, in your re-enactment.
4. USE THE MAP; and all the clues — the time, the clouds, the distance…
5. ACT IT OUT on foot, in miniature. "I left this way, turned…"
6. SAY OUT LOUD a) Well that's where I thought it was b) But it is really over there (e.g. that Sun there).
7. Imagine the stars circling where they ought to; the coasts trending the way the compass tells you; etc.
8. Aim for a NEW START "I don't know what happened, but now I divorce the past".

Only when 'the ground shifts under your feet', do you know that your brain has got hold of the new situation.

How to Leave a Message

- Put it in a prominent place (e.g. a hilltop; a campsite; a rock face) with a prominent marker — bright and colourful for instance.
- Make it helicopter-visible e.g. in a clearing; large
- Use international distress symbols e.g. 'V'; Fire; 3 lots. At sea, a white flag or white stroboscopic (flashing) light.
- Any bit of metal may help you scratch out a message e.g. a zipper runner; a coin; a key; e.g. onto a tree trunk.
  If not, try a rock, or stick, to scratch with. Next time, have paper and pen with you.
- Use spittle, or urine, to make mud, then cover the mud message, but prominently — e.g. when searchers remove the coat from the tepee they see the message. E.g. make a flat mud pie; press in twigs to form letters; keep the rain off it.
- A freshly dead leaf is like paper, to write on.
- A piece of charcoal makes a handy black chalk.
- Brittle Yakka gum makes a bright orange powder splotch, to attract attention, and can also be used as a somewhat waterproof crayon.

Disorientation in new territory

Familiar territory does *not* disorient you, so you must *have* the mental faculties to 'know where you are' when you can 'see where you are'.
The overload of *novel* surroundings must overwhelm your normal abilities to mentally keep track of small diversions and accumulated progress.
These suggestions aim to make an exploration more 'familiar' and less 'new'. They come from all over this book. I suppose the message is:
'Don't just read about the methods, take extra time and care to put them into practice; extra 'bother'. Slow down; try harder. Remember more'.
Look back *very* regularly, to increase your familiarity of already explored territory. If you had eyes in the back of your head, it should *all* look 'familiar'. Walk back over that territory before continuing, p 68b. It will feel like old familiar ground if you bother. You could find your way through, and back through, an obstacle course at night without a torch, *if* you went forward three legs and back two as a matter of habit or necessity. That's 5 whole journeys. Talk to the obstacles along the way, and use a mnemonic to recognise them, and what you did about them, and how you turned.
Leave visible markers behind you, so looking back is more helpful to orient you. Leave a trail. It boosts your confidence, even if you don't use it.
Look through the foreground into the distance to **maintain contact with reference directions**. Fix what you can see, especially ahead, in your mind, in relation to the Sun, shadow, wind, etc, until you can point to them from behind visual obstacles. Remember more of what you see. Interpret each clue for its relationship with your memories of how things looked. Integrate the bits into a whole. Use some distance measure.
Maintain a straight line if convenient, so the interpretation of where you are is transparent. Use long straight 'legs' so you can interpret your accumulated path more easily. Notice and remember the angles *between* legs especially. Create a mudmap, so you remain oriented to your start position, and to your current position, and to landmarks sighted along the way. Practise these and other techniques, when it doesn't matter, for later.

To Work Harder, Rest Harder

"Mental alertness" may involve

- A siesta nap!
- An early night (not a late-night campfire party or sing-song).
- Regular rests during the walk.

Tiredness, Exhaustion, Constancy — these stop you thinking efficiently, and affect your memory badly too. So do nicotine, alcohol and drugs.

- When you are in difficulties, sleep on the problem, or take your mind off it with a diversion, occasionally.
- For the sake of having stimulating conversation, the whole party ought to guard their mental alertness too.
- Navigation games do more than orient you, they wake you up.

Be alert to your biorhythms—daily, monthly, seasonal—they affect your alertness or make you depressed—all your mental faculties can be affected.

Minds Work Best with Intelligent Questions on Them

- Am I judging the scale on the map properly?
- How many creeks have we crossed? How many more to come?
- What direction will the next creek be flowing in?
- What is the nearest escape route? To where?
- What does this place remind me of?
- Will it frost tonight or not?
- Where would the animals drink, and shelter?
- What is this place's own special uniqueness, and why?

How high and steep is the next climb?

What is our next position-check landmark? How soon?

Could we see Mount X from here?

What side-trips would people make?

Why is it sandy soil?

How fast are those cloud shadows going?

Anticipate, with questions, the things you want answers to.

Look Through the Vegetation, to the Contours

- Bushes cloud your vision. If, for example, you want to 'stick to the ridge', or want to memorise a junction, then look for the more **permanent** clues…
  ..like slope, creeklines, direction trends, terrain, soil & rock types, alignments, distances, context.
  Of course, look at the vegetation too, but not only at the vegetation.
- Trees cloud the skyline. But there are gaps and lines of sight that open up as you pass by. Look through the forest for the ridgelines, slopes, and horizon. As you move sideways watch the parallax—to reveal otherwise hidden contours.
- Look through the trees, to the **most distant**, largest, tallest features to navigate by — you will undoubtedly catch glimpses of these much more often than you think, if you are looking. They may be other trees, so get used to recognising individual trees.
- Examine the height of the sky beneath the canopy.
  E.g. the lowest sky is 'downhill', but **the lowest sky uphill** is toward a pass or gap, & **the highest sky downhill** shows up the ridge.
- Forest-bound **river beds** are often clear of trees and give you a rare distant view.
  Stop in the middle of that rope bridge! And hang around, to look around.

Connections to the Outside World — You Don't Need to Feel Isolated

Ruins connect you to the past. Earthworks and mineworkings also. Even the landscape engraves its history for you to see.
A **GPS** connects you to satellites; a **radio** to a radio transmitter direction; a TV antenna points to the transmitter; a **phone** connects to the grid.
The Sun, Moon, stars, planets + time + vertical can give you an **astro-fix** to connect you to latitude and longitude — globally.
The **magnetic field** connects you to magnetic poles, as do the aurorae. The deep blue polarisation band finds the Sun for you. So does a rainbow.
Navigational antennae link to flight paths. Flight paths indicate airports; pipelines indicate water supplies; electricity lines go to substations.
Paths point to points of interest. Recent tracks betray traffic. Track-markers put you on track; tracks connect to roads; roads to highways; to cities
Nightlights betray townships and cities; traffic noise links you to highways; surf noise, to the beach. Tides tell you about the Sun and the Moon.
River gravel shows up the upstream geology. East West gully shape shows up North; **rivers** connect you to the coast; **ridges** to peaks.
Air pressure here depends upon altitude; and upon the polar front and the pressure systems. Air temperature relates to the size of the landmass.
Absolute humidity tells about the proximity of oceans. **Cirrus clouds** show up the (westerly) upper winds. **Low clouds** and winds remind you of the isobars on the pressure map. Ground winds telegraph the topography. The radio links you to **weather information**. Waves, to distant storms.
Fence lines connect you to farms; boundary lines to survey grids; trig points to **maps**. Smells to camp fires, marshes, farms, abattoirs, toilets, seaweed, water. Stock trails lead to water; seed eating birds lead to waterholes; the plant and insect types are tied to their range and the region.
Vegetation is tied to altitude, latitude and aspect. **Bush north** indicates noon. Bush slant indicates the **prevailing wind**. Recent wind leaves tracks of yesterday. Deep soil temperature is related to altitude. Why feel "lost"?

OverNavigate: Equals Over-Familiarise

Overfamiliarise yourself with your route as you walk it, the first time you walk it.

Familiarity is the navigator's best friend, next time.
I.e. Years later you can more easily retrace your steps. Even when the track has overgrown.
It is an investment opportunity now, not to be wasted.
The extra time and effort will pay dividends, but like all good investments you must have an interest in the future.

E.g. Notice how and where you went wrong — mentally photograph it.
E.g. Lay to heart the most permanent features — the landscape and trends — not the track, its signposts, or bushes, which may change with time, or with a bulldozer.
E.g. Mentally photograph the lie of each junction, as if a fire might burn up the signpost.
E.g. Describe out loud your short-term memory about how you got here, lest that familiarity flits with the passing of the day.
E.g. Have your own 'track-notes' book.

How to Get Lost — Follow the Leader

A leader is like a track. You can follow him mindlessly. He ends up being a seeing guide for the willfully blind.
Ask yourselves, often, this hypothetical question: What would happen if the navigator dropped dead?
Could anyone else navigate competently?
Do others know where we are? On the map?
Could I find my own way back? Where are the cars?
If the 'leader' himself gets disoriented or lost, noone will know where they are!

Antidote: Teach. Learn. Test each other. Involve each other.
Demystify the magic art. Share it. Talk about it.
Listen to the debate. Look at the landmarks. Contribute.
Don't walk fourth in line. (Rotate regularly)

"Here, you find the way — I already know it"

"As Solid As the Rocks in My Head"

"Who am I to argue with the Sun? to say to it 'You should be over there — not where you are!' "

Telling your head off is the sort of attitude you need… when you find yourself walking uphill to the campsite which you know is down by the river!
… when the clues are shouting at you "Something is wrong here!!"

Navigators who are disoriented prefer to believe in such remote but real possibilities as a compass error or a mistake on the map, signpost or buoy, or to believe in the majority opinion being wrong, rather than to believe in the closer and more real possibility: "Perhaps it is ME who has made some mistake".

True, compasses & maps are notoriously prone to error, but your internal feelings, vague recollections, position/time/direction senses are more notoriously unreliable.

Lost? Rule 4: DON'T Do It Again!

Make resolutions. E.g. 'I will always take a compass.' (I tie a key ring compass to my hat string and I always take my hat.)
(I also tie a whistle to my keyring)
E.g. 'I will buy a map/mobile-phone/EPIRB'
E.g. 'I will always leave my itinerary with friends'
E.g. 'I will learn more about navigation; take pencil and paper; practise'
E.g. 'I will never go on a walk unless I have asked: "Which way am I leaving; Going in what direction?" '

Even better… Don't wait to be forced into a sensible habit. Declare your present attitudes to be 'not yet fit for human navigation' and put it right, now.

A Three-Track Mind

Here is a trick to avoid attention-deficit syndrome in your navigating:
Practise concentrating on two things at once.
Your feet find the path by themselves.
Your mouth talks about how good the meal will be.
Your mind sees the scenery without missing it.

Dual tracking your mind is like many other good things — just a little bit too hard to persevere with under normal circumstances, but possible to achieve, with extra dedication.

Don't automatically get lost into only one train of thought.
Slate this down for practice, for the rest of your life.

Interpret!

E.g. If you wear a face mask in these days of extreme UV radiation, or sunglasses, you become aware of when you are walking into the Sun.
The obvious interpretation, hanging on to this piece of otherwise subconscious information, is: "I must be travelling toward the Equator".
It's a chance to orient yourself to the new orientation. Can you see your shadow ahead of you?—you are getting closer to the pole.
E.g. You might be in a deep dark forest and see a rare bit of sun-dapple piercing the dense canopy.
That means you can go over there and see the sunbeams for a direction check.
E.g. The 'crunch crunch' underfoot changes to a 'clop clop' — "Oh, if this is rocky, maybe I've passed a break of slope."
E.g. Exotic plants betray human occupation — you could look for an old access track.
As a general rule: **Look for any asymmetry**, and interpret it — usually for direction. E.g. When swimming with goggles, twinkly bubbles mean "upSun" — dull ones, "down-Sun". The wave and ripple pattern indicate wind direction. Sunbeams show up the down-Sun direction.
E.g. Scuba-diving: The anchor rope shows up-/down- current; alternating current is the swish/swash from the longest wavelength waves—note their direction. Direct current equates with "away from my starting point". Bubbles show "up" (with some adjustment for current if needed).
Interpret the whole story of settlement: Valleys attract farms, dams, fences, boundary roads, houses, access roads. I.e. you see a lonely fence post on a mud flat, and you say "once several paddocks probably converged onto this old watering place—look for old tracks near the fence post". Look for old nail-holes in the post, to find out the direction of the fenceline.

Play the Field-Scientist

Vary your **focus-distance**.
A herpetologist will see snakes at his feet & frogs at 10 metres.
A geologist will see the skyline, and the creekbed outcrops.
An ornithologist will look upwards, a botanist downwards.

Scientists see more, because they know what to look for, and how.
The more they look, the more they see, observe, discover, and learn i.e. the more they learn what to look for.
You can begin the self-reinforcing cycle in yourself by: wanting an explanation, by trying an interpretation.
And, of course, don't forget to look into the topic, into books about it and field guides for instance.
Be interested in all levels, in all scales and directions.

Focussing keenly will improve your navigation.

Orientation is More Fundamental Than Position Sense

We all carry mental recollections of recognisable spots—like a mental photo album.

But try to mentally picture whole stretches of a country—roads, coast, range, river—with a sense of direction & relative location & linkage.

Better is to carry a library of whole mental maps—walks, paths & layouts—not just photos of isolated spots or isolated linear features.

But Best is to keep all those maps in right relation to each other in your mind.

As your ability improves, you focus less on spots and more on the links between spots;
less on position fixing and more on trend analysis;
less on separate techniques, more on integration.
Direction clues are only helpful if you can coordinate them with what else you need to locate.
Positions only make most sense if you can say how they lie, in distance & direction, in relation to other positions.

Think in Pairs, in lines, trends and tracks

Two landmarks lie on a line, of known direction, and fixed track.
We tend to see things individually, and not see the join between them.
So try instead to get to know
the *direction* of 'camp to that moored boat',
the *bearing* of 'Mt Big to Tower Hill',
the *angle between* 'sunset and the headland'
the *trend* of 'Razorback Ridge'
the *pointing* of 'tree one to tree two'
the *line* from 'start to finish'
the *track* from 'Lake Low to River Junction'
the *tilt* of a star-pair
the *join* of two subsections

Needle-North, -South, -East & -West. Identify them on the Landscape

One of the major problems of navigating in bush, is keeping log of estimated side-slip, distance-wise, even if you are oriented, direction-wise.
But you can see your sideways movement, see p53a, if you plot known directions onto known points on the skyline, (E.g. Magnetic N,S,E,W require no numbers; E.g. your initial forward bearing is already there waiting to be used). Or better, use points in the middle foreground landscape — about as far away as you intend to walk. E.g. local hillocks, towers, houses—things which will be visible from time to time during the whole walk, the closer the better. Make a note of their directions. Use these spots later to find home or your starting point. When you see them later, in relation to compass, you can visualise how much distance out there you have moved and in what direction. Page 106c "Tom Thumb Parallax" shows you how.

Orientation to Your Start-Finish (Rhumb) Line—by sight or compass

Work out the start-to-finish visual bearing to the destination if it is prominent, or to a landmark or point on the skyline beyond the destination.
Whenever you can see that point, you can see your left-to-right sideways positioning, in relation to the fixed track from start to finish.
A simple way to keep track of your sideways movement, at least in open country, is to choose an approximate rhumb line compass course (from start to finish) and watch it move across the landscape in front of you. E.g. "The follow-me arrow is now pointing more to the *left* of that gulch".

Orientation to Your Finish Point, by the Parallax Method

If you are making a one-way journey, use the same "parallax method" as above, to keep a log of the estimated sideslip, but use your **destination**, not your start point, for your reference set of bearings. As the walk progresses, you can see yourself closing in on the correct destination.

Exponential Anxiety — the 11:59 syndrome

Your fear of being 'lost' rises in inverse proportion to the need for fear, as you approach your goal in unfamiliar terrain.
Will it be there as expected? Why haven't I got there already? Surely it should have shown by now!
Have I overrun it?? Maybe I just didn't recognise it? Has it disappeared?
When will this track ever end? Will it end at all?? Will I have to walk all the way back! And so on.

Now, obviously, if you have done your homework well, it will be there, where you thought it was, but maybe not exactly when you thought, so optimistically, you 'ought to' arrive.
Distance guesstimation is our weak point.

Don't add needless anxiety to your background worries. Wait and see it sort itself out.

Anxiety, fear and panic work against good navigation.

Explore Junctions, While You Can

It only adds a few minutes' pleasant diversion to travel the branch tracks which you don't want. (Especially if you can dump a heavy pack)
You may never get to find out about them any other way.
This includes natural junctions—of creeks & ridges—and canal junctions.

When you are exploring, you may never retrace those exact steps, so don't say "I'll explore that on the way back".
The other reason not to say that, is that you will be even more tired later.
When you are uncertain of your position, such exploration ought to be considered "essential", now — not later when you realise your mistake.
Later you may speculate about what it might lead to, but now you can prove it.

In any case, as a matter of principle, explore far enough so that you don't have to repeat the foray
E.g. walk left, until you are sure you should not go left.

"Meet back here in 10 minutes" is one way to investigate in all directions at once.

Separated? An Agreed Plan Helps

Example: Where did I park the car? Is it stolen or am I lost? Are the others worried yet? What will they expect me to do?
Example: Our driver has driven off! Why? What will we do? What will he do?
Example: She has not returned on time! She was exploring that way.

Try these out...
"We will, for at least one hour and at most two hours, try to rejoin each other before changing to panic-plan"
"Where we were last together, is the agreed rendezvous now"
"After that we check our intended destination for messages"
"We know to communicate via a 3rd party—a 3rd person, a 3rd place, or a message bank… "
"..the nearest post office/petrol bowzer/pub/police station/emergency service/Fire Brigade/bank/Red Cross/Salvation Army"

Concentrate Outwards

Both introversion and extroversion belong elsewhere.
If you cogitate inwardly, you become blind to the surroundings.
When you chatter away about other things, that's equally distracting.

Don't spoil the special time of being outdoors.

While it is there in front of you, look to see what is to be seen — and talk about that — about the history of the place, or something about it that helps to keep you oriented, to its unique nature.

Learn not to be distracted from the main game.

SPELL Out Numbers — Into Words

The idea is to express, in words, what numbers measure. Humans hate hearing numbers! So translate mathematics into user-friendly words.

e.g. 270° is "West". That's much more meaningful.
10 past One — "Oh, it's midday!" (in your locality, for summer-time)
45° South — "sub-tropical, even temperate; halfway to the Pole"
10.5 km — "almost half-way"
020°T — just to the right of magnetic North.

PAINT your Words — Into Pictures

E.g. To make it even more meaningful, "West" = "Sunset!" (a word-picture)…or even better……………………
"019°" = "Where the Sun is just before it reaches magnetic North"
(This translates numbers into a familiar and visual time of day.)
"Midday" = "High noon!" (How high?)
"Half-way" = "Within sight of the Gorge"
"Halfway to the Pole" = "The Sun rises and sets at 45°. The pole star is halfway up"

'Meaning' usually means a 'picture'

Words without meaning are like looking at the brushstrokes of an oil painting through a magnifying glass, missing the painting!

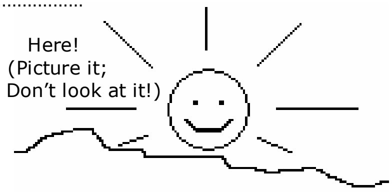
Here!
(Picture it;
Don't look at it!)

Look at the Ground

- Soil types; rock types (p. 79a)
- Animal runs; animal footprints
- Fallen track markers; toppled-over cairns
- Human footprints e.g. yours, on the way back
  Examine each other's boot treads, then the boot prints. Talk them through — expressing the recognisable peculiarities.
  Practise also on car-tyre treads and prints. (p. 66b)
- If you see a deep trench or a dry mud-hole in Tasmania, that may mean 'an old track' i.e. where constant traffic had refused to let the vegetation grow back.

Look for Foreshortening & Overlap & Parallax

Looking flat on to a shoreline e.g. the other side of the lake, means that any slight inlet or headland is probably a major and very sinuous feature. Focus sharply on it, looking for clues to depth. Don't just run on past it. Move sideways e.g. 20m, to show up the depth, by parallax.

Make Common Sense of the Various Direction Clues

Maybe the navigator has his eyes glued to the compass — but you, you can say "Oh we're just going upwind".
Perhaps you can see six landmarks — but the sensible way to see them is e.g. "We are heading halfway between Mount William and our shadows".
Cloud direction is a good clue, but changeable from hour to hour, so there is no loss in saying "Diagonal to the Sun" which is also changeable.

Don't Degrade the Point of a Clue or System

E.g. The point of a watch, is to tell you where the Sun is. (Don't look at the Sun!)
The purpose of a mnemonic is to remind you to think, not just to remember or to recite.
The reason for a "midnight-rise-noon-set" system is to work with the full-circle—even when you turn corners or turn around. It images the sky
The benefit of a skyline method, is for when you can't see the skyline. You will glimpse it 10 times more often, if you are trying to use it.

The aim of a practical method, is to practise it, not just to know about it!

The Aurorae

The Southern and the Northern Lights are especially frequent during high sunspot years (maximum every 11 years.... 2000, 2011, …).
They are most prominent after midnight. They are best seen in 'high' latitudes (> 40°). One to three days after a solar magnetic storm is best.
The most common aurora is the most easily overlooked — it is dismissed as a 'pale glow' in the North or South, low above the horizon.
But it is not dawn — that is East. It is not from a city — because it changes. Sometimes the sky will be pale green below and rosy above (between the clouds) and pierced by 'searchlights' which move left and right, slowly but visibly. Sometimes 'curtains' appear, shimmer and disappear slowly.
All the rays seem to emanate from somewhere near the North or South Magnetic Pole, and it spans 30° each side or more.
It is an interesting direction guide.
You can estimate current magnetic activity by analysing sunspot activity. **Never look at the Sun**. Instead, throw only an **image** of the Sun onto white paper, as follows... Point your binoculars at the Sun, **without looking through them**! Then, as if you were throwing a minimum shadow of one binocular barrel onto paper, and as if you will **let the paper look at the Sun through the binoculars**, from a distance of say half a metre from the eyepiece, point the big lens at the Sun. Adjust its alignment until the sunbeams go right through it and come out at the other end where the smaller eyepiece lens is. Steady the binoculars properly, and focus the bright, circular, white, image of the Sun so produced, carefully on the paper, by using the focusing mechanism and/or altering the paper distance, so as to reveal tiny pairs of black dots, on the paper. The larger the dots the bigger the storm. Caution! Looking at the Sun will damage your eyes and even looking at the bright image on white paper requires sunglasses.

Orientation to Your Start Point, by the Parallax Method

Step 1: **Start off Aligned**: Make a note of Needle-NSEW, bearings of any other recognisable points, initial-bearing landmark, rhumb-line landmark.
Step 2: **"The Direction Has Changed Its Mountain!"** Instead of saying: "The landmark has changed its direction", first say: "The direction has changed its landmark". I.e. retain in your mind the initial bearing, then, whenever you glimpse that landmark again, you can see how far sideways you have moved off that track—by locating which landscape point that direction now points to. You estimate the sideways disparity across the land, because your compass needle now points to a certain distance to one side of the original mark. Your job is to estimate that distance out there.
Step 3: **"I'm moving through the landscape, out there".** Imagine a thousand of you dotted all over the landscape, all moving parallel, and at the same speed, say, at 4 km per hour. All your initial bearings to all those recognisable spots out there, are moving with you, leaving those spots behind at a rate of 4 km per hour, and now those bearings point to new spots a calculable distance ahead. In reverse, this lets you see how far you have moved sideways to any direction. Bring those distances back to your feet, and see how far away from "start" you are, and how to get back. (At 15 km range, if you are walking at 4 km per hour, you will leave the landscape each side of you, behind at a rate of 15° per hour, but...) From the satellite's point of view, you and your directions are (all) simply moving through the whole landscape at 4 km per hour, no matter at what distance. An alternative viewpoint is to say that every bit of the landscape always moves "equal, parallel and opposite" to your motion, relatively speaking.
Step 4: **Estimate the gap between parallel tracks**—See "Advancing Position Lines" p.108a. I.e. from the initial bearing (a track) to the new one. —by taking a new bearing & using sin(change in direction) x new range. This converts the distances you can see, to a number of kilometres.
The answer, if you want an answer in numbers, is how far you have to travel at right-angles to the original bearing to restore the original bearing.

Nervous Nellie Navigation — Not Necessary

Anticipate a problem. Do something crisp about it. Clear your mind.

Examples: Continually looking at a compass needle — can be short-circuited — by dividing the route into short legs, p. 73b.
Continually gauging the distance covered — can be short-circuited — by making one prediction before you start, and waiting for it, p92b
Getting lost trying to keep track of many bends — is a symptom of not knowing the overall trend in the first place.
Continually guessing where that junction might have led to, is avoided, by exploring it when you meet it.

You never know when you may be totally distracted from your normal pattern of thinking — it is then that you will need a cleared mind.
Plus, if you habitually allow your mind to move on to the next problem, by dealing with the last problem, you avoid ongoing confusion.
Forethought can see the normal problems coming, and pre-empt them.

Look for the Essentials — the 'Directions'

You can't afford to miss them!
E.g. 'Turn left at the Y-junction' means someone has to be looking for it!
E.g. 'Head uphill after you cross the creek' means: notice the creek bed!
E.g. 'If you hit the main road you have gone too far' means that you were talking about lunch instead!

Directions are worth writing down as soon as you hear them: "First right, third left, 6 km, look for the track"
Try to give directions more sensibly: e.g. "Up Fish River Valley from the Park Boundary"
Try to use fixed directions e.g. "North West" "from the lake to the Peak" "between sunrise & Mount William"
Not variable or relative ones e.g. 'turn left'

A Whistle-Code for Contact Calls

Example: The track splits. It's probably just a detour. "I'll go left; You go right". You keep in contact by whistling, cheerfully & tunefully.
Use your favourite tunes agreed between you...
To signal: "It's me" Try: "I did it My Way"
"You Come to me" "Oh Come All Ye Faithful"
"I'm coming" "I'm Coming; I'm Coming… ..Old Black Joe"
The rhythm is...
Radio **Morse**: CQ = Please Contact Me = " Dah-dit-dah-dit Dah-dah-dit-dah " "C is Seasick, Q Forms a Queue" "Calling Calling, All radiOs"
R = Roger — Received = " Dit-dah-dit " "An R dits"
Wait = " Dit-dah-dit-dit-dit " "aWait f'r a bit"
S&R: Come Back i.e. Recall of Party = Dah — Dah — Dah — Dah "Ree-Call-All-Souls"
Other: International Distress = XXX = Dah-dit-dit-Dah — 3 times. "all X's hurt, all ex's hurt, all ex's hurt"
SOS — real emergency — di-di-di-dah-dah-dah-di-di-dit — repeated "A bit of Save Our Souls isn't it?"
See appendix for the full code.

City Glow

From out in the bush you can usually see the direction of about three cities or towns, by their night-time glow.
This works best in hazy atmosphere — virtually any night! — and especially on overcast nights and when no skylight will show.

You need to check a map before you go — for the nearby townships — then check the theory against the facts,
then you can find it useful when you need it.
E.g. you can hold a straight line at night — relative to any recognised glow.

Don't confuse it with aurorae, sunset afterglow, predawn Sunglow, Milky Way, asteroid belt (Zodiacal lights), headlights, local floodlights, bushfires.

Just remember to get up and look for what you normally don't notice.

Learn Think Look Integrate Map Use

Come to your Senses

..And it will be surprising what comes to your senses.
Smell the air — for seaweed, marshes, farms, abattoirs, smoke, smog, civilisation, scent trails, fresh wombat dung, humid air, perfume.
Listen occasionally. To help you hear, cup your hands to your ears. Selectively tune out the high-frequency noises to listen to the low frequencies; and vice versa. Listen through the white noise, and through the silence of night.

For occasional vehicles, trailers bumping, boats thumping, jets overflying, thunder rumbling
For constant traffic rumble of a highway
For ocean rollers thumping on shore. You can also often hear the difference between high and low tide, from the wave noise.
For running water, frogs
For wind noise
For generators, pumps, machinery
For farm stock
For the crunch under your feet, changing

Look **carefully**. "It is amazing what you can see, if you look". It is a most important habit for the navigator and is very rewarding.
E.g. you read what others have scratched over the sign that says "5 hours" and discover "8 hours"; or find old track markers almost overgrown.

TRANSFER Your Pictures into Reality

E.g. "Sunset" = "THERE!" even after sunset!

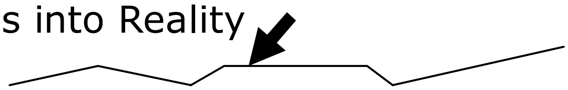

Find the red glow afterwards, further to the left e.g. at 8:30 p.m.

"High noon" = Look at how short your shadow is.
"Within sight" = "Well, can you see it yet?"
"The Pole Star is halfway up" = Do you see how the Cross doesn't set?

Living in the realm of theory is like an absent-minded professor — who can't find his way home, or remember which way he was travelling.

Put That Experience in Your Memory Bank

E.g. "The Pot will go down behind "Pot Hill"."
"I didn't realise how high the Sun can get at 45° latitude!"
"You can see the Gorge from the halfway mark, provided you climb the ridge"
"The Southern Cross keeps correct time at the end of March" — the beginning of Winter. See opposite for the next step…

Look, with Vengeance Aforethought

A bird observer doesn't just stumble into birds, he runs them down.
A native tracker doesn't simply 'notice' the clues, he searches for them.
Your task is to track down navigational clues, then bail them up.

Your enemy is: losing conscious grip on your task i.e. when the trudge gets the better of your brain.

For Practice: Express the unexpressed. It's no good to be only half aware of what is going on.
Interpret the meaning of each clue. It is no good staring at a compass needle, unless you realise where you are going!
Look outwards, to see what you can see.
Change 'a cursory look' into a specific search.
But, what exactly should you be looking for (once you have exhausted your repertoire of clues)??
Answer: Find out! I.e. **Teach yourself**. Don't wait until someone tells you.

Preparation is the Key to Difficult Navigation

Marine navigators in rock-strewn waters, or battling seasickness in heavy weather, or short-handed, know that the only way to cope with seat-of-the-pants reflex navigation is to have **already prepared clear instructions** about danger-bearings; back-bearings; lines of sight and distances to run, so that they don't need to leave the helm.

It is the total picture, beforehand, which orients you, later.

They need to plan the entire passage (so as to make landfall just before dawn, for instance) according to the tidal streams and ocean currents that they expect to meet (since it is the current much more than the winds which affect boat-speed-made-good, even for yachts). Their key desire is to sail with the current, and preferably with the wind and the weather systems, avoiding the storm-season. See "Aiming-off" p96a.

An hour seem such a long time to spare when you are busy at home; But when you get out on foot, relaxed, with hours to walk but no opportunity for desk-work, it seems that every hour not spent in preparing what could have been prepared, was a wasted hour, as far as being useful now.

Helpful Symbols for Standardising your Chart–Work

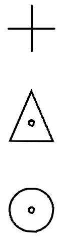

"The Deduced Reckoning Position"
from course and distance information
("Ded. reckoning, or D.R.)

"The Estimated Position"
taking into account currents and leeway
to distinguish it from simple D.R.

A known position, or a Fix

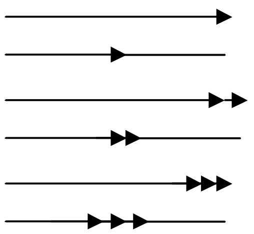

Lines of possible position should have the arrow on the end
e.g. pointing to a landmark
A central arrow means a course. i.e. travelling, not pointing.
One Arrow signifies the course steered through the water
A Transferred Position Line, i.e. advanced across the chart
Two arrows is the course made good over the ground
(A doubly transferred position line, for a three-way fix)
Three arrows is for the stream or the current.

Look Backwards; Think Forwards.

You sometimes, with only a mud-map to go by, have to follow a straight line by compass or landscape, to a low-lying, invisible feature, e.g. a tarn, hoping to come across it without bypassing it. Well, success will often make you feel that that is the end of the matter; "Mission Accomplished". Until next time!
But look back, if you can, to a recognisable starting point, then take a compass or landscape bearing, backwards to the start.
You then convert that to a forward bearing for next time. Place 2 pebbles on the ground in line, turn around and line up the landscape mark needed
The result is a "track", not just a direction — see p.71b — a 'can't–go-wrong' guide.
If your beginning point is not visible, turn around and take a forward bearing on the skyline feature you think you should have headed towards (and probably did). E.g. "One span to the left of Mt. Visible". The result is only a direction from the start—a guessed-at track—but nearly right.

On p.42d I said only to use bearings from high points to high points, for a general navigational framework.
But add to that, a mental or notebook entry of the bearing from a low destination point back to visible possible starting points, when you arrive successfully, or of the forward course you used — preferably by the landscape itself. They are "How-to-get-to (a low point)" tracks.

(Interpret)

Extend Visible Reality into the Invisible Picture

E.g. "It's boiling hot in Baghdad"
"The Sun is over Townsville"
"From the ridge the water flows down to the Gorge and eventually out to the Murphy Mouth"
"The Earth must be rotating Easterly, from West"

Don't be satisfied with half a meaning.
Until the interpretation leaves you oriented, and able to remember something, you haven't gone far enough.

See p. 71b(1) for the last word on interpreting 270T.

To Get There Quicker, Go Slower

We all prefer to run ahead of necessary preparations, i.e. we prefer instant gratification to wisdom.
We also prefer to get something unpleasant, finished, and quickly, once we have started, e.g. "can we make this bad "shortcut" work out?".

Don't rush!   "More haste, less speed" = More care, less fluster.
Slight panic makes you hurry. Hurry is a recipe for carelessness in navigation, and for blunders.
  E.g. over-running the junction
  E.g. missing most of the clues
  E.g. taking a foolish "shortcut"
Common causes for hurry, other than anxiety, include…
  "Let's beat the rain/the dark"
  "Let's get back on time/for lunch"
  "Let's sort this out quick-smart as to where we are"
  "Let's explore one more kilometre before we admit we're lost and turn back"

Search and Rescue Signals — Noises Repeated Regularly

One = We are searching for someone. (Is it you?)

Two = I acknowledge your signal. "Okay; Roger received-and-understood; I heard that"

Three = "Search for me. I need help."

Four = Recall of Search Party

E.g. rifle shots; loud whistles or horns; two boards clapped together

---

End of Section 2.

Nature's Norths

Learn these when the sky — the astronomy Norths — is clear, so that you can **lean** on them later. (Assume S. Hemisphere 1,2,3,5,6)

1. "Dew" North — Dew, moist soil & roads stay wetter in the shade. Frost will persist in the shadiest places, e.g dew South-West.
2. Snow South — Snow remains higher on the shady side of an obstacle.
3. Mud North — Half-dried mud-crusts betray the hot, wind-exposed cracked edges, with the damp mud in the shadowed places.
4. Rock Noon — a whole mountain, or a crack in a rock, can show up the sunny and the shady sides. (See next box)
   — Also, in the evening, after a sunny day, you can feel the colder side of a boulder or gully.
5. **Gully South** — The shady bank of a creek is steeper than the sunny side, at least on average, all around the world.
   The side facing the noon Sun is flatter. The side facing the pole is steeper. This results in roads, railways and buildings etc, being more common on one side—the flatter side—of an East-West river than the other. In other words, creeks tend to cling to the Pole-facing slope of a ridge, as if they migrate equator-wards. Perhaps this is a remnant of a past geological age when the erosion was more marked e.g. hot days eroding the sunny side flat, with ultra-violet light; and constant moisture and icy nights eroding the shady base steeply; therefore suitable mnemonics are: "lying down flat sunbaking" or "steeped in ice" (i.e. South-facing in the S.H.). Lakeshores have steeper or flatter banks according to the same principle.
6. Animal Orientation — "**Barnacle South**" (SH), on dark intertidal rocks — it shows up the wetter i.e. shadier side.
   'Anthills' (Termite mounds) sometimes orient to avoid the heat, so the length of them is North-South.
   Insects prefer the warmer side of a trunk in winter e.g. spider cocoons, ant colonies, boring grubs.
   Some spiders spin their orbs to face noon, like satellite dishes but more vertical; others face sunrise.
   Migratory flight-paths may be North/South &/or upslope/downslope. Animals flee in predictable directions,
   e.g. inland, uphill, North, upwind e.g. when faced with a fire. They often lie with their backs to a cold wind.
   Siesta South: Farm animals find the midday shade in a paddock. You can smell the urea. You can see the dung piles.
   (7—10 will be expanded soon)
7. **Wind West** — part of the W in the 'WEATHER' mnemonic. It is designed for windswept places when the wind is not blowing.
   The prevailing wind leaves its signature in the visible growth history of the plants, and in the landscape.
8. Wave West — if, in the Roaring Forties, you can see the accumulated effect of ocean rollers.

9. Botanical Asymmetry — where you sort out the competing effects of wind, light, salt, exposure.
10. **Bush Noon** — you never need to be disoriented in the thick bush.
    This works best in temperate latitudes — where the sunlight is directional.

Of course, anything to do with nature, refers to **averaging continual observations**. The most variable signs are the least reliable.
Less Natural Norths

11. Magnetic North; And don't forget magnetic dip — the downward tilt of the magnetic lines of force, varying with latitude.
12. Solar Panel Noon — Observe how we set up solar panels, and verandahs. Nature gives us the Sun, and we all follow it.
13. Satellite Dish Noon — Geo-stationary satellites are, by nature, Equatorial.
14. Microwave dish to microwave dish — Peaks 'naturally' point to each other! See "landmark to landmark network navigation"p42d

Instead of feeling bored, or getting lost, or being 'bushed in', you can have the valuable bonus of feeling at home in the bush, having a hobby, and being oriented to the slope, Sun, vegetation, ecology, weather patterns…
Bush skills are a good investment for future bushwalks. So go for some experimental walks in the bush on overcast, foggy, rainy days.

Don't Forget to Remember, Your Way!

Long-term memory is your best friend — so feed it. You aim for recognition, years later, through familiarity now
e.g. Have you walked the track at night? both ways? twice?
e.g. Collect as much first-and-second-hand information as you can.
e.g. Reread your track-notes occasionally before you lose the memories.
Photographs, Maps, Track guides, Track notes, etc are all permanent on paper! but you aim for permanent memories, not records.
I.e. Mental photographs, Memorised records, What the locals call "knowledge".
Use memory methods so you don't waste your efforts. Use the paper records to help you consolidate your local knowledge.
To that end: Make drawings. Make your own map. Keep a diary. Make your own track directions, names & comments. Tell somebody.
Use biro not pencil. Make a neat copy of your notes and file it.
"Look at…" means "Notice…" means "Take note" means "Take notes" i.e. Take a notebook.

Unlock Memorable Chapters in Your Life

Memory is fundamental in navigation (alongside direction, distance, position sense, visualisation and alertness).
Medium-term memory is how you find your way back, but aim also for long-term recall.
We remember best in definite 'bites' — location by location.
"Right, team; We have to remember this. Help me. We need to discuss it, describe it, mentally photograph it, draw its skyline, make it a memorable stop, time it, note it, put it in its logical place in sequence."

Go for the Jugular. Zero In on What Matters

Concentrate on the crucial and take-away information
— the crucial "directions" E.g. the sequence of junctions
— the details that have direct orientation value E.g. the visual angle between North and Start-off
— which give relevant memories E.g. the hill-top view laying it all out for you
— track notes e.g. the bearing to follow, the "directions"
— the overall trend — top-down, top priority pieces of information
— the perspectives that fit everything into the context. E.g. the surrounding trends in relation to Sun-North.

E.g. A mental map is better than a 'position' on paper. A map diverts you from the real picture.
Recognition outvotes names; Zero in on the most peculiar visual peculiarity
names outvote position fixes.
The overall details — What you need to know in three years' time, when you try again,
are the **time, distance and difficulty**.
The Sequence (of subsections & features) ought to be remembered... ..rather than e.g. 'bearing 035° magnetic'
Recognising Sections of the walk... ..is more relevant than having a series of map positions.
Surrounding landmarks last longer... ..than "turn left at the red tin-can-marker".

Appreciate the plants and animals; the soils, geology and climate. Instead of "Bin there; Dun that" you aim for familiarity.

The "Lie" of The Land

Think of this as the up-down part of geography. Think "Lie-Down" (versus "Lay-Out" i.e. a map layout).
Think of — Drainage, Watersheds, Catchments & Divides—the accumulated results of Ages of uplift, drainage, erosion & deposition.
— Geomorphology — it determines the shapes
— Ecology — it responds to the shapes, the slopes & drainage
— and how they all fit into each other.
E.g. "This north-facing slope supports warmth-loving vegetation"
"These foothills of the main range drain Westerly"
"The basalt uplift has generated these clayey mud flats"
"This is an isolated hillock within the main valley"
Think of the 'lie' as 'down toward what general direction?'
and 'rise' as 'up toward what encircling range?'
I.e. Put a general trend on major rivers and ridges. Imagine the coast, where all the (now nearly parallel) ridges & rivers run offshore.

Lift your eyes to the "Big Picture" scenery. Interpret it; Memorise its detail; Put it to good use.

Listen to the Pioneers

I.e. listen to the place-names

"The Great Dividing Range"	I suppose this one is obvious
"Blackbird Flat"	This means there is a river nearby
"Stockyard Creek"	A flat clearing is promised
"North Ridge"	This is related to the North Arm of the River
"Huon Valley"	It houses the longest tributary
"Boggy Plains"	Down on the flats!
"Eliza Plateau"	Up on the boggy flats!
"Bright Pass"	Means a change of catchment
"East Pillinger"	It ought to be East of West Pillinger
"Pelion Gap"	A saddle — between catchments
"StrathGordon"	It says "the Valley of" Gordon River
"-Dale"	Also says "Valley"
"Mole Creek"	Burrows underground

Lie Has To Do with Water Drainage

Look for dips in the skyline.
If you find a 'soak'—where water soaks out—look for a ridge-bound marsh uphill—where water soaks in.
Carry an artificial horizon, to reveal slight down-slopes, e.g. a flat plastic jar lid suspended on three strings, adjusted to horizontal.
Or lay your clear plastic drink bottle on the ground. Maybe even pour out some water to see where it runs off to. Listen for frogs!

Stick to the Track

- If you are waiting for the rescue party
- When a "shortcut" looks inviting
- If you have told someone exactly where you are going
- When you are nursing an ailment
- When sacred or heritage sites are at risk from tramping and erosion, rubbish, pollution, desecration, etc.
- When you are one more tourist in the constantly invading hordes, and "Please keep to the track" is aimed squarely at you
- When you are in a greedy hurry for more of the same elsewhere
- If you can't find your way out of a wet paper bag.

Reasons for Leaving the Beaten Track

- Tracks often fail to meet the needs of birdwatchers, botanists, photographers and bushwalkers. Remember why you are there. Talk to the ranger about whether the "No Camping" sign refers to drunken car-based parties, or to scientists and bushwalkers too.
- You sometimes want to search out the worst routes! Easy slopes have one sort of ecosystem/niche/microclimate; Difficult terrain has the more unusual biology.
- Interaction with a bare-dirt strip is not exciting. Fan out, Tread Carefully, Go Slowly, Keep your party size small — then you can appreciate the flowers, lizards, lichens. Each step is watchful — which forces you to see these things.
- A fast-track through a nice place robs you of appreciation-time. The experience passes just as quickly as the tramping to get there!
- Designed tracks will often steer you clear of exactly what makes a place interesting to see — cliff-edge views, water-edges, botanical hot-spots, refuge habitats,…

Isolated places often don't have tracks, and you will need to choose, navigate and negotiate your own route.

Cross-Country Routes

All routes are 'cross-country' but some don't have tracks to follow.
There are two sorts of tracks — Un-natural ones—artificial paths (man & animal made) which form a simple network,
and often stick close to the contours — gently upslope or gently downslope
And — natural routes (following the geomorphology) which cross the contours at right angles and
which form a double network.
I.e. Natural routes are determined by an interleaved network of ridges and rivers...
You are well advised to stick to these routes if only for ease of route finding.
Then there are three sorts of off-track navigation to consider…
— departing from the artificial paths
— departing from the natural routes (to follow a trend to a destination) and
— trying to keep a dead straight line cross-country...

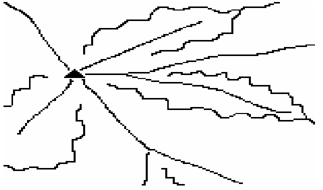

Rock Noon

Rocks blister and crack in the heat. In addition, ultra-violet light hastens all other erosion processes. The top sunny sides erode easily.
Conversely, the shady side is subjected to lichen growth, icing, and moisture, especially at the base.
Differential erosion results in tilted rocks, in asymmetry of cracks and gullies, and in various light and vegetation effects.

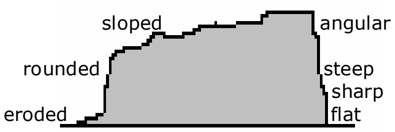

On the hot, rounded side, look for fissures, cracks, furrows, blisters and pock-marking — as signs of heat stress resulting in more rapid erosion.
e.g. Blistered flakes leave a patch of bare rock with no lichens on it.
Such patches are hard to find on the shady side.
Algae grow in the protected fissure under an almost blistered-off flake of rock.
The green stain on the underside of the flake is greener on the sunny side!

Caution: This is **unreliable** unless the rocks show a clearly interpretable pattern.
Of course, **use bedrock** or larger boulders, since they are going to be in their original position — not shifted about regularly.

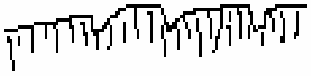

Examine the cracks — The shady side is sharp and steep; the sunny side is rounded.
Just like "Gully South" but in miniature.

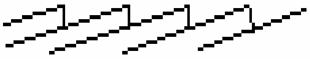

Examine the steps — sharp step-downs indicate a 'dip' in the layers down to noon.

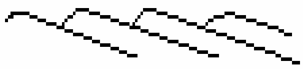

— rounded step-downs show a geological dip away from noon.

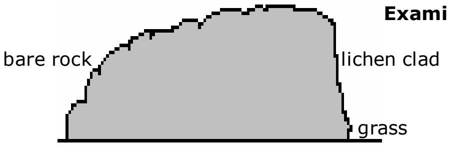

Examine the colours — looking down-Sun you see pale, bare rock (where it has recently blistered).
— looking up-Sun you see multicoloured lichens on a darker background.
— diffuse lichen grows better on crumbly rock. The sunny side will grow such lichens.
— thicker-bodied, textured lichens grow older on the shady, non-eroding side.

Feel the topside vs the backside for a difference in temperature.
Look for green grass and moss in the shade of the rock. These are often far more reliable than the rocks themselves.
Beware that a fire-storm will kill lichens, moss & grass, and will blister the rock, crumble it and split chunks off.
i.e. some of the effects you will notice may be due to a recent fire, e.g. a firestorm from a particular direction.
Rock noon only needs to be reliable enough to tell you which end of your improvised compass points North.

Don't Try to Memorise Details without an Overall Framework

It's a recipe for confusion — to try to keep track of every bend, every hillock, every weather change, every creek bed, every channel.
There are too many of them. They all look alike in close-up.
So start with the map: Learn the **major** rivers and the divides first.
Go for the big names, the overall trends.
Fit the details in by retrospect, after the big picture shows itself.

Don't Try to Remember Anything You Haven't Memorised

i.e. put the effort in first—don't fool yourself. There is a process to memorising something, e.g. establishing a framework first.
And...
Don't try to remember anything you haven't made memorable. Don't battle uphill—make it enjoyable or emotional and it will be easier.
And...Don't write down anything you haven't memorised. First make it memorable, then write it down.
Notes are not an excuse to forget to memorise something!
Like most short-cuts, they are an enticing alternative, but they, like shortcuts, short-change you as well. Memories are precious.
And...Don't try to remember anything you haven't expressed or named. It is a fact of life that we refer to memories by name, in words
And...Don't expect to remember details you haven't understood, or organised, structured, or reduced to a bare minimum—a prompt.
It always helps to interact with, analyse and process anything you want to remember. E.g. investigate it with questions on your mind; criticise it; teach it. It also helps to visually set it out in your notebook with colour, shape, balance and symmetry.

If you relax, because there is a track, or because you have a map, or you have 'directions' for the trip,
and maybe someone else is with you who is familiar with the area, then you cease to build up your own memories and familiarity.
Again you have short-changed yourself.

Write down What You Memorise!

Does that sound like over-kill? It isn't. It's just good sense.

- Memory isn't perfect. So help it along. Do so before you forget.
- You will want to write down your track notes, for years later, anyway.
- Writing helps your memory — it "involves" you more — it stimulates your expression, revises your experience, adds time spent thinking about it, involves your physical participation and your verbal and visual faculties.
- You never know when slight stress (I'm lost!) will distract and divert you for an unknown time.
- To write it all down clears your mind of the previous task — you can concentrate elsewhere.
- On the way back you can easily glance through your notes and then know what to look for.

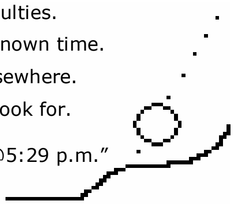

E.g. DRAW a picture of where the Sun sets on the skyline. LABEL it e.g. "Sunset @239°T @5:29 p.m."
Or was it 229°C @5:39? That's why you write it down.

Set aside a notebook for a "TRACK DIARY" (SH)

Major Rivers & Ridges, vs Minor Ones

Obviously, all water, falling on one side of a catchment-dividing ridge, runs to the sea, via a single river mouth; Conversely, river-mouth headlands are dividing ridges. Followed inland, all headlands must reach right around the river catchment—they are the two ends of the dividing ridge. The rivers and ridges form a sort of geometrically inverse pair of systems.
The catchments with the shortest creeks occur between named River mouths, and are called 'coastal'.
I.e. All other coastal land rises to some different, more local, dividing ridge; you may have to undulate up and down along it on the way
Not all creek-lines from a 'major ridge' (a catchment divide), descend all the way i.e. although the water may go all the way to the sea, the creek carrying it may not: the creek might only be a tributary, ending where it empties into a river. Similarly not all ridges running off the catchment-dividing-ridge reach all the way to the coast—some are "dead ends"—they end in a "shoulder", cut off between two creeks joining, or at a river bend. All valley-dividers are ridges, but not all ridges divide whole valleys, or even minor catchments. Those that come to a dead-end before they reach the coast are not dividing ridges, and may be called "minor".
All creek lines do not necessarily start on a dividing ridge: The creek itself might attach to a minor ridge, within the valley area.
Those ridges and rivers that transect or divide whole catchments are called 'major'. The named river ought to be a major line.
Unless you look for them on the map you may not be able to notice anything special about them.
Major ridges & rivers do however tend to stay higher and lower for longer, and often provide you with the **gentlest climbs**,
And they attract roadways.
Main ridges also tend to run parallel to the main valleys and vice versa.

Sort it out on the Map

- You cannot see catchment layouts from ground level!
- If you walk around a slope, half-way up, e.g. following an aqueduct, then you cannot easily know when you cross a major catchment. There are many minor ridges to cross and any one of them could turn into a saddle which divides valleys.
- Around peaks and plateaus, valleys spawn in all directions — it is hard to identify the right ones. High marshes may drain in several directions.
- Near a major river junction, travelling along the dividing shoulder can give a confusing alternation of views, between one valley and then the other.

Catchment Divisions Snake from Coast to Coast

Ridgetops fit parallel to and between the two highest contours marked on your map, and can be very sinuous.
They mark neighbouring catchment edges; & so theoretically you can cross the continent on highest ground without crossing any creek.
Major catchment divisions must reach the coast at a point each side of the river mouth (except, of course, for inland salt lakes).

A Catchment-Division* Separates Whole Valleys

*called a "watershed" in Australia, unlike in USA

Hint 1: Mark catchment divisions on your map with a pink highlighter — pink is opposite to blue-green for creeklines.
Hint 2: Now forget about those new pink delineations on your map. They refer only to the layout, as seen from above.
You need to imagine real, high ridges inbetween deep valleys.
Hint 3: Now forget about those razor-back ridges in your mind. You need to know that they could be remarkably flat paddocks.
— not very eye-catching — look for them carefully.
A catchment division will follow the highest flattest ground and consequently have a few marshes and puddles on it
A metre-high bump in a paddock might separate two raindrops into two different 1000 km journeys in each direction.
So you can't simply "walk downstream" — not until you know where you might end up!

Tracks are like Naughty Children

If you take your eye off them for a moment, they will split
as soon as they are out of your sight.

A shortcut is often the result of a junction.
This is because tracks often bend at a T or Y junction.
Shortcutting is constantly encouraged at any bend,
So there will likely be an inviting well-worn shortcut there.
You can always split the party to explore both routes.

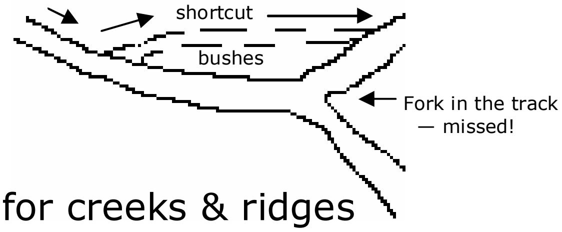

The same is true for creeks & ridges

Unless you stick to the bank of the riverbed, you cannot count on seeing the tributaries, and if you do stick to one side you miss the other side tributaries. Also, where a stream is "braided" into parallel interconnecting branchlets, and you follow one gutter, you are easily blinded to its tributaries joining in from side-valleys.
If you don't stick to the main gutter or crown, you miss seeing junctions on the far side of the river or ridge and can follow a wrong one. But side ridges are harder to see from the crest, just like tributaries are hard to see from the central channel!

Explore Branches While You Can

As a matter of principle, it is easier to explore something if you are there, than if you are not!
Be quick to remember that before you walk on past.
Remember… that you are there to enjoy the country. So put down your pack & amble off for a minute or delegate someone to explore.
You may surprise yourself by discovering a helpful signpost at a second nearby junction or a new access point and car park,
Or even, that it is the track you should have taken just now. Bends Breed Shortcuts: A shortcut may look like 'the junction we don't take' which means that you may mistake the real junction anyway.
Avoid endlessly having to wonder later… "But if that other junction turned east, then it might have been the link…" etc.
When you are there, find out for certain.

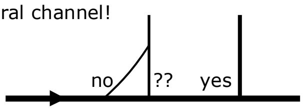

Convergent Tracks Lead to Civilisation

For example, you get lost, stumble across the track and meet a T-junction — which branch do you take?
You can examine any Y-junction for signs of splitting into two or more minor branches.
The most acute angle of a Y-junction usually points to a focus point (e.g. downstream; car-park)

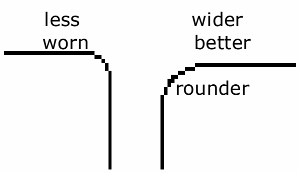

Down-Track vs Up-Track

This one is not obvious, and so it may trap you.
You seem to be following a well-worn path down, until you turn around and try to find your way back up.
It is then that you discover a multitude of other tracks joining into yours at an acute angle!
You see, people walking downhill, downstream, get channeled into the paths of least resistance, converging just like the creek lines do.
But people walking upstream or uphill are faced with choices and tend to spread out into all possibilities, until they find the right one.
The best path is worn in from down traffic, but that may not make it obvious to up-traffic.

Allow extra time when going up-stream, or climbing a steep peak, for eliminating false leads.

Wind West

Assume you are in a wind-swept place in medium to high latitudes e.g. southern Australia, where strong Westerlies prevail.

'W' in the 'WEATHER' mnemonic (p.114b) then reminds you
of the prevailing **W**esterlies in the high 30s, the Roaring 40s, the Furious 50s and the Screaming 60s.

'W' for '**Weather** systems come from the **West'** in temperate latitudes.
and '**Wind** in your face' locates these weather systems for you.
and 'Upper-level **Westerly Wave'** steers these systems to the E/NE/SE in a waving pattern.
A common pattern is to have a warm nor-westerly storm preceding a strong cold sou-westerly change (SH).
and '**Wind-West'** reminds you to look for the lower-level **prevailing Westerlies**, and their effect on the bushes, landscape, etc.

You don't have to be in the Roaring 40s to find a windy place and look for tell-tale signs of the prevailing wind.
— Onshore winds near the coast, being laden with salt, leave unmistakable clues imprinted in the bushes, and on the seashore.
— High plateaus and wide open clearings are usually windswept.
— Gullies produce up-gully breezes and down-gully gusts.
— Hilltops are windy. To get an undistorted picture of the prevailing wind direction, go to a hillock.

To navigate by the local prevailing wind may be more preferable locally, than to navigate by bush north.
The prevailing wind is not as universal or reliable as bush-noon, but it can be easily visible, audible and tangible.

In general, the **tallest trees** are the best tell-tales.
Look for wind-induced asymmetry**, especially at the tree tips** and the **oldest** tree trunks, when you are in a wind-sheltered zone.

Think Wind

Distinguish in your mind the various causes of wind damage…

- Storm Winds — brief but furious, as a weather change hits. They may uproot trees. They are often hot and dry.
- Fire-storms—from the hot dry inland, or sweeping uphill. They can leave a trail of dead black fallen trees lying parallel.
- Cold Change Winds — Cold, Gusty, Strong, Persistent for a day or three; Usually south to southwesterly (S. H.).
- Persistent Prevailing Winds — Westerlies 'Waving' between northwest and southwest in between 'changes'.
- Salt Laden Onshore Winds — They 'cut' the exposed vegetation back to size.
- Cold Blast Winds — These kill off the less hardy competitors.
- Hot Dry Blast Winds — These leave only the waxy/succulent/woody species.
- Sand Blast Winds — They remove vegetation, carve out sandstone sculptures, and remove lichens from the exposed side.
- Windy Season Winds — Ask the locals what to expect at each time of year.

Anything is Memorable

Make it personal	"Where I found the pocket knife!"
anticipated	Study the map beforehand to imagine the sections
active	Do something; involve yourself
important	Draw attention to why it should be remembered
associated	It looks like a … ; sounds like my middle name ; uphill from the mine
creative	Make up an enjoyable name
emotional	Expect to notice, then remember, decision points & anxious times
visual	e.g. symmetric, colourful, shapely
geometric	This triple-trunk matches the triangle of paths here
noticed	Draw attention to it "See how we have just entered tea-tree ?"
peculiar	That tree looks like a traffic policeman!
prominent	The tallest tree here, see?
liked	What aspect about it interests you?
named	e.g. "Midnight Junction"
fitting	"Power Point" where the electricity line crosses the lake
funny	'Willy Willy Waterhole' where we saw the dust-devil
human	"Betty's"

Make Individual Memories Memorable!

The problem is an excess of information, with a shortage of enough meaning to go with each detail.
e.g. hillocks and creeks have endless detail — you can't hope to remember all.
Visual memory is not the problem — pictures are perfect — but recall is difficult.
(And sorting pictures out later, into the original sequence is also hard.)
You need to add memory handles, worthwhile meanings, specific focus, etc.
Hint: Do something memorable there. Dance, Sing, Jump, Shout, Throw stones into the puddle.
Make a cairn. Sketch a tree. Have a meal — Chewing is said to aid memory!
Leave something of 'yours' there — something peculiar — e.g. lean a stick against a tree trunk; scratch the ground.
Insert it into a larger scheme — time & sequence; lie & surroundings; an acrostic poem.
Make a game of it. E.g. "Who can say what's peculiar about this creek?"
Make it 'yours' e.g. Discover it. Climb it. Explore it. Study it. Explain it. Draw it. Map it. Show it to someone. Name it.
Spend time there, looking.

Remembering Trees

Simply spend time looking. It may only be 10 seconds, but that is 2000% more than the usual ½ second!
Look at the branches, the shape, the trunk, until something strikes you
E.g. "Oh, these two branches grow across each other"
E.g. "The bark is right-hand spiral"
E.g. "See that ant-highway going up into the hollow—it might have water in it"
Walk around it, to see the shape from side-on. Characterise it & its context
E.g. "entirely shorter than that old stag's first branch"
E.g. "three tufts on top"

You get better with practice.

Ridges

Ridges have gentler slopes up/down ridge i.e. provided you stick to the crown.
The slope 'makes' rapidly on each flank. It is only the ridge-line which is flatter.
Ridges have nicer views. But these views are easily halved unless you walk right on the crest.
    have nice breezes, provided you are right on the crest.
    have more open and stunted vegetation plus shallower soil and more exposure to the elements.
    are wider and flatter than a creek bed, and their rocks are firmer underfoot and dry.
Their run-off slides off downhill to each side.
These advantages are also lost as soon as you slip off the 'crown' of the ridge.
Off the ridge is steep, vegetated, narrow, loose, windless, hot, humid.
The steepest but best way up a ridge is... up the ridge because (1) it gets even steeper off the ridge
& (2) contouring is 'easier' but it doesn't get you 'up' but only into more difficult terrain.

Ridge Names

▪ Mountain to Mountain	e.g. Mount Ossa to Legges Tor (trace it on the map)
▪ Mountain to Mouth	e.g. Mount Ossa to Arthur River South Head
▪ Mountain to Headland	e.g. Mount Wellington to Wrest Point
▪ Catchment Division	e.g. The Arthur-Pieman Divide
▪ Tributary Divide	e.g. The South Arm Ridge

A ridge may not be "major" but if it divides your walk, you do want to know about it—so make up an appropriate name for it.

Mnemonic: "Ridges Rise"

Whenever you cross a ridge don't forget to ask: "Which way does this ridge rise, in trend terms — to the left or to the right?"
It is like asking which way water is flowing, but the truth can be obscured by local undulations in the ridge.
Ask: "Does this ridge, if followed to the (right), end further inland i.e. upstream, or further downstream?"
It must correlate with the overall corresponding drainage.

Uphill — Which Hill? Upstream — to which Divide?

To Which ridge? Which plateau? Which highest ground? Just as surely as rivers reach the sea, so ridges reach a peak.
Which peak? is the question which orients you.
Until you know which catchments you are climbing out of and in to, you don't have a grasp of where you are.

Sympathy and Understanding with Track Markers

Take time out to explore (e.g. leap-frog fashion) from one track marker to another.
Always guess where a track ought to go.
When you come across a marked trail—maybe poorly marked and/or difficult—the question arises: 'Should we follow it or not?'
Intelligent guesswork helps here

- Where would *I* put a track?
- What is the direction trend of the new trail?
- What is the slope trend of the trail?
- What tracks & routes might you expect?
- What alternative access routes exist?

E.g. uphill, heading West, may well mean: 'to Lake West from the lower car park'

If you leave the track behind, because it is unclear, you may never get back onto it.
But if you work with what you have (a bad track) it may be quicker in the long run.

Old Track Markers

Practise noticing them, before you need to.

- Old axe-cut blazes on tree-trunks — from the original surveyors
- Very old paint flecks in the blazes — search for them
- Blazes and markers on the far side of the tree — look back
- Old nail holes or nails or rust stains — the marker has fallen off
- Fallen markers — on the ground
- Fallen rock cairns — the rocks blend in, but show their pale underbellies
- Arrows on wet rock-faces — the wetness obscures the scratches
- Sawn branches, trunks and logs from when the path was cleared
- A clear line-of-sight through the trees
- Thick young growth — an unused road
- Smooth or shiny rocks with less lichen — worn smooth by feet
- Fallen trunks with a gutter worn across them — betraying foot traffic
- Broken branches of varying ages; bent stems e.g. bracken fronds
- Stumps cut off at ground- or eye-level.

Be Wind-Wise: Listen

Listen directionally — the wind-noise will be mainly from up-wind.
When you think you are navigating by the beach noise, you are probably navigating by the onshore wind, and should keep track of it.
Listen **through** the noise — You may distinguish three frequency bands, as the rumbles emerge from the hiss.
Listen through the **silence** — of night, in the stillness. Up-wind sounds carry a long way.
Listen to the **sounds** (birds, traffic, house, factories) — They carry clearly from up-wind only, especially the high frequencies (whistles)
Listen to the **beach** noise — The 'hiss' will be from a direction more up-wind, and closer, location-wise.
The 'rumble' will be from further away from you and further away from up-wind direction.
Listen to the **thunder** — Air temperature zones create directional effects, so that approaching storms, which are preceded by low level cold bursts of air, have their noises refracted up and out of earshot.
(See "Tips for Calm Weather" p.23d & "Topographic Level Winds" p.25d)

Obvious Tell-Tales: Landscape

- E/W coastal asymmetry e.g. Temperate West Coast ocean beaches might have a kilometre of surf, but East Coast beaches are narrower and calmer. Trade-wind East coasts are surf beaches, but headlands with West facing shorelines will be sheltered.
- Bays end in exposed points. Look for tall, steep, bare sandhills on the weather side. If there are tall sandhills at a narrow point, you will see sandhills from both sides, so ask 'am I seeing the green (sheltered) backside?'
- Sand dunes. There are variations, but bare dunes tend to be cross-wind and vegetated dunes tend to trend down-wind.
- Puddles, Lagoons, Lakes, Creeks, Beaches — Look for wind & wave action effects e.g. dark mudflats up-wind & white sand-steeps downwind. Shallow puddles often occur on the downwind end of a flat drainage.
- Sand ripples show steep backsides downwind, as the wind driven grains topple over the eroding crest.
- Clumps in sand-stretches show a long tail downwind. E.g. behind dead seaweed deposited on the dry beach.
- Grass-clump erosion is the reverse of sand ripples — look for steep bare dirt and exposed roots on the exposed side.
- Camps & campsites, picnic tables & fireplaces, bird nests & animal shelters all avoid the un-sheltered side of vegetation.
- Wind-shadow growth—plants grow taller in the lee of a mountain range, or of a pebble, sand ridge or house.

Obvious Tell-Tales: Windblown Objects

- Flotsam collects on the downwind side of a lake (but on the sheltered side of a coastal point!).
- Trunks fall downwind in storms; branches fall downwind in gales; leaves, nuts and flowers fall downwind in strong winds. Look for ribbon bark hanging in a tree downwind from its source.
- Rubbish (old dead twigs & leaves) collects behind a clump.
- Dead long leaf blades lie cross-wind when blown against an obstacle e.g. a clump...
- Dead long leaf blades lie down-wind when blown from a source e.g. a tussock...

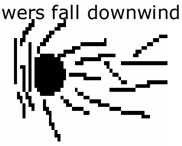

- Broken off branchlets lie with their big spread-out twiggy end downwind. Meadow-grass tufts pulled up by grazing animals also.
- Feathers catch in bushes, usually on downwind branches, and mat against the up-wind side of those branches.

The Difference between Looking at, and Seeing

..has a lot to do with involvement.
e.g. a seed-collector will see the vigour of the plants and their stage of fruiting. Would you even see the plants!
e.g. a hunter will see a route for stalking from behind cover.
e.g. a photographer will see the contrast in light.

Would you as a navigator care to notice footprints, or their absence? What about the flight-paths & contrails, or gems in the river gravel?
Navigation is your opportunity to see more.

Search-Image

It is possible to stare at something obvious and not see it! e.g. the pattern of wind-blown grass stems, or the peculiar geometry of a tree. But after some practice, you can't miss it. So train your brain to the search images before you need them. When you are flustered and hurried you will not feel confident in your abilities to see something difficult, and will likely give up too soon.

We Recognise 'Sameness' But Need to Describe the Differences

A gum-tree is a gum-tree; but this gum tree has three trunks!
A 'bend on the road' is just a bend; but this one bends away from sunrise.
A 'hillside' looks much like all hillsides; but this one is regrowth with wattles dominant.

Don't just gaze at "the view" or "the landscape" or "vegetation". Look past the fuzzy 'sameness' to the unique peculiarities.

Unfortunately that often means: 'Focus on the boring detail'
e.g. 'This hillock has five dead tree trunks, totalling nine main branches' which makes this hillock recognisable and unique.
So. Hint: **When all else fails, COUNT!** (The same goes for botanical species identifications.)

There is often no way around a detailed look. There will not be anything unexpected to hit you in the face — you must keep looking until the interesting peculiarities emerge.
E.g. "No other dirt road junction I have seen has an Acacia sapling growing in each corner."

"Shoulders Stop"

A huge mountain range may end in a "shoulder" i.e. a 'shoulder' amounts only to a 'minor ridge' stopping within the valley.
You will not get to the coast by following it down. (If you are on the very last ridge next to the coast it would be called a headland).

There may also be isolated high ground within a valley, a lot like an island, and this too is doomed to end in only 'shoulders'.
Such high ground is drained by one river only, but via several tributaries.
It will have one, often low, flat connection (a saddle) 'uphill' to the main Divide.

Shoulders are Not Easy (like Ridges) To Walk on

1. They end a spur, and so are likely to be steep, as the ridgeline plunges down to the creek bend or junction below it.
2. They are wetter than ridge tops. Although they shed water, being convex, they are still fed with run-off from above, as on a hillside. So the vegetation can easily become very thick.

Saddle — A Flat between Hills, Sitting Astride a Ridge

They are so named for their resemblance to the two-bumps-two-gutter horse-saddle shape, designed for two legs astride a sway-back.
When a flat area is drained by three creeks and so has three ridges lifting from it, it is called a "monkey saddle".
The third gutter accommodates the tail. This can cause uncertainty-of-choice for navigators!

A "col" is a saddle.
A "pass" is usually a low point in a high ridge i.e. it is a saddle.
A "gap" however may simply be a river gorge.

Every undulation and flattening of a ridge leaves a saddle. In forest, there will be a puddle. It may also attract a campsite or a road.
A saddle may look like a valley crossing—a muddy dip in the path—but the two sides both drop away, if you look carefully through the wall of vegetation.

Weather maps show pressure 'saddles' between two High's and two Low's—p.11d.

You can also picture a saddle as a U-shaped valley going up and over a gap between two hills and continuing down the other side.
Saddles make creek sources, but not every creek source begins as a flat saddle—any ridge or slope may spawn drainage gutters without necessarily flattening out.

Track Markers

— Decode the colour-code. Separate tracks often have separate colours. Reverse track markers may be of a different colour.
— Successive generations of track marking may be evident, or several colour-coded tracks may share one leg of the journey.
— Triple marking, often means "End of Markers" or "End of Track".
— Big cairns mark junctions, campspots, features. Cairns with a stick may also indicate a junction.
— Side-tracks are marked with something extra or different.
— Sticks over the track, or especially crossed sticks say "No Entry. Wrong Way. Go Back".
At an unexpected junction, if you have to step over some branches you have probably missed the intended message.
— Un-needed signposts — often mark the entrance of disused tracks. They may (or may not) be at an unexplained bend in the route.

When You Lose the Track-Markers

Be suspicious! Start by over-navigating beforehand — search for all track markers as a habit, even when the track is well-worn and plainly visible. It gives you the practice for when the track marking does get patchy.
Also, only then will you notice whenever the track takes an unofficial shortcut (e.g. following vehicle tracks, not the walking track).
Or maybe you have taken a wrong turning off the marked track and need to realise it, no matter how good the new track is.
Whenever you lose the markers or whenever you lose the track, explore in these directions: **straight ahead and straight behind**.

One or two markers may simply have fallen off — so maintain the trend
You may be approaching a main track junction with the first few sidetrack markers removed to conceal the sidetrack you are on.
An unexplained bend, which deteriorates, may be explained by a missed junction/bend...

Where a tree has fallen nearby, try to pick up the trend of the old track, on the far side...
Backtrack to where you went wrong — to the last bend & from the last certain marker.
Next, explore the unexpected directions.
Very often a marker will be placed at a bend, because of the bend, or even because of a Y junction.
Then explore in all directions — "Meet back here in five minutes" might save 50 minutes.

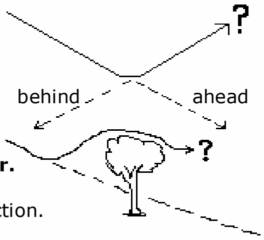

Notice (with Alarm!) When...

..The track itself **deteriorates**. You may be on a false track, made by people doing just what you have done
e.g. you overran a junction or bend; you sidestepped a fallen tree without noticing.
..Boot prints, that were there before, cease. Especially on a 'through' track.
..The track gets **better**! This means that in back-tracking you may miss your small track and follow the other track which joins in there
..The country abruptly **clears** e.g. a beach, a peak, a clearing, a road. There may be no way to find the single small track you used.
By the time you remember to look back, you may have lost the way back out. There may be too *many* paths to choose from.
..Any **barrier** e.g. a beach, a road, a fence, a riverbank — clear or not.
When you hit the barrier you do a right-angled bend, but later you may not remember exactly where you met that barrier.
..Any **catchment change** i.e. crossing a ridge, or even a creek crossing—into "another part of the catchment".

Name the enemy: Be the first to say out loud **"Abaracadabara! Peek behind you! because it's a...(fill in the blank)"** A **B**arrier; A **R**idge/River/Road Junction; A **C**learing; a **D**own-track (or focus point); A **B**etter/worse track; A **R**idge/River crossing; A **Peak**.

Light & Colour Effects

This is one of the delights of nature! Chances are that you will be startled, even if you know what to expect, but only if you look for it. You need to wade out into a uniform reed-bed, preferably when the Sun is side-on to the prevailing wind, then look up- & down- wind. Up-wind, the tussock tips are pointing toward you. You then see between them into the deeper shaded body of the plant.

- The view up-wind is dark & colourful. The best effect is from rich brown or dark green, drooping, chest-high reeds.
- Down-wind the tips lie flat and reflect the sky shine at you — faded & bright.
  The prevailing wind may result in permanent sunbleach on one side of the grass blades.
- On a windy day you will see gusts causing bright patches in a dark background, due to the same flat blade effect.
- To neutralise the directional lighting of the Sun, look down-Sun — to your left & right will show pale vs colourful.
- The combined effect of the Sun and the wind result in four quadrants — darkest up-wind; shiniest downwind; greyest up-Sun.
- This happens even with ordinary green plants — look for it in any low, flat, wind-swept vegetation.
- The prevailing wind doesn't need to be blowing to see its permanent effect on the plants.
- At sunrise in the reed-bed, you should notice a 'funnel-web' effect on your right — like the circular scratches showing up through a sunlit window.
- At other times you may simply notice that any shiny leaves on the left of the Sun tilt left, and on the right of the Sun tilt right.

- Looking up-wind anywhere is greener. Downwind looks more bleached — due to leaf-tip scorch, leaf fading and leaflessness.
- Weather-side sandhills are white. Lee-side sandhills are greener, due to thicker growth.
- Exposed lake banks are steeper & eroded. So, sandy lagoons show white sandy vs dark muddy ends.
- In sandy heaths, downwind shows up as windblown sand patches of erosion. Up-wind shows up as continuous plant cover.
- Rocks in sandy places are sandblasted pale on the up-wind side and darkened with lichens on the downwind side.

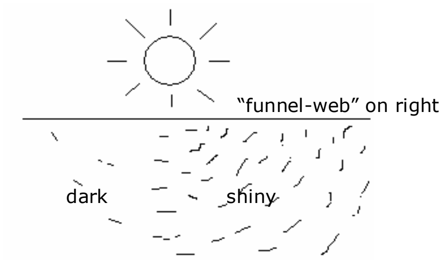

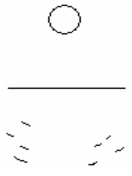

Shiny leaves

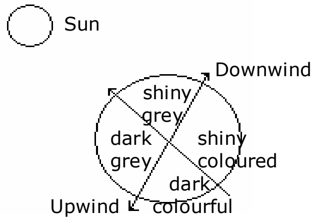

Expect some variations

Place Recognition
You Recognise What You See, By What You Can't See

Every locality fits into a context — the invisible surroundings.
How long did I take to get here? How far from the turnoff? In what direction from home-base? Uphill or down or contour?
How many creeks have we crossed up to now?

To know where you are, you can't ignore where you are not.
e.g. 'It was an all-day slog to get there, we finished the day worn out'
is part of where you were, and fundamental to your memory of it.

Landscape recognition is relative.
That involves your senses of time, direction, distance, difficulty, slope, topography, layout, number and network.

Aspects, invisible to the eye, which surround a place, live in your mind's eye.

Description Requires Words
Can You Describe All That You Can Recognise?

E.g. a magpie. What pattern is on the head? What colour legs?
E.g. your brother — what colour eyes? what voice accent?

"Recognition" means "I need to see it before I can recognise it"
"Description" is "I can recall it to my mind and describe it"

If you don't deliberately memorise something you can only recognise it but can't describe it.
So thoughtful examination, notes and practice, can fill in a lot of blanks.
Repeated visits build up a subconscious picture, but you can speed up that familiarisation process by **deliberate scrutiny**.
Expressing the description in explicit words, forces your subconscious 'recognition' powers, into a new dimension.

Hint: Open your mouth; say: "The place where… " and see what comes out spontaneously.
try: "In between…" "Running alongside the… "
"You probably would remember it as the time when… "

Notice When You Cross a Catchment!

It is a major method of navigation
e.g. at night; or in a blizzard; or when you are lost; or in a forest; or descending from high ground; or exploring upstream.
But you need to get used to noticing ridges and to counting creek crossings and to matching the catchments to the map.
Notice also, when you cross from one side to the other side of a catchment.
Ask: Is this riverbed likely to be the lowest, largest channel in the whole catchment?
Look for creek-line vegetation, as sometimes the only clue to the drainage pattern.
Look for any darker patch in the road or track or beach, as a tell-tale sign of a dry creek bed.
Look at the track ruts. If the rocks are exposed, water flows over it. If the ruts flatten out, water drains away from the track there.
A subtle rise, dip or bend in the track can be traced to an obvious ridge or drain crossing it. Look sideways to appreciate the lie.
Assess any puddle, marsh or mud hole — Is it a creek line or a saddle top? — Is it in a dip or on a crest? — Can you trace a gutter?;
Notice any crest; Trace ridges to a nearby peak. I.e. Don't get caught whingeing about the government, instead of noticing the lie.
Ridges divide: they shed water. Creeks collect: they catch water. Valleys drain: they empty into the sea.
Rivers cut the catchment in two.
See "Understanding Contours" for some hints — later in this section. See also p.83b.

The Giveaway! A Change of View — Mnemonic: "Catch the View"

The new panorama is the tell-tale sign that you have changed catchments.
Perhaps you are contouring around a headland and catch the new view around the corner, or you are sidling around a ridge, or you are breasting the saddle, or reaching the peak, or cresting a paddock… It is the new vista that greets your eyes which tells you that something has changed. It is also the loss of the old valley view.
Well, maybe you can't see the view, but you may now be able to see the skylight or even blue sky down through the trees below you.
When climbing, you find yourself looking down instead of looking up. Your muscles, of course, tell you the same story.
The deeper the new valley, the more far-reaching, on average, the new view.

You Can't Cross a Creek Twice from the Same Side

If you make two successive creek crossings in the same flow-sense, you have crossed a creek divide into a different creek catchment.

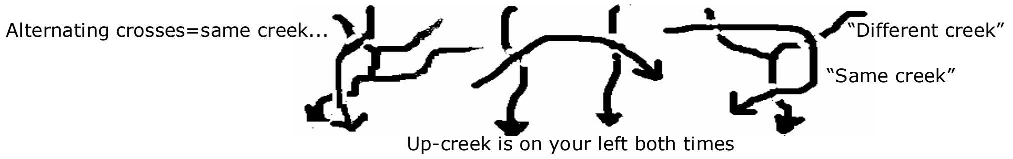
Alternating crosses=same creek...
"Different creek"
"Same creek"
Up-creek is on your left both times

The same rule can hold for ridges, but only those which slope uniformly down from high ground.
Otherwise an unexpected crest slope, instead of being a 'different ridge', may alert you to 'a local up-trend' on a down-sloping ridge.

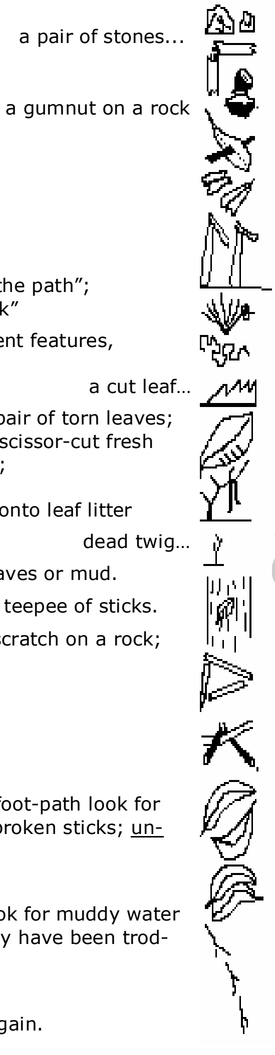
a pair of stones...
a gumnut on a rock
a cut leaf…
dead twig…

Mark Your Trail

Two principles are needed 1. Trail marking takes extra **time** — budget for it.
And this one is widely applicable… 2. It's easy when you know how, so **think** how!

Mark Your Trail Non-Permanently

E.g. when you could otherwise get lost

- Use what is available — pebbles, leaves, tussocks — throughout the area, for ease and consistency.
  e.g. "always look for a pair of stones"; "tie a grass blade on the top shoot of any shrub"
- Leave your mark in a visible place. E.g. "look for freshly broken twigs at the base of each trunk on the path";
  "check the topmost leaves of anything chest-high for a mark"
- Leave your mark in an expected place. E.g. cross-country: Zig-zag in straight lines between prominent features, so coming back you know exactly where to look for the next markers, and searchers do too.
  E.g. hillock, to tall tree, to clearing, to outcrop.
- Do something only slightly unnatural: (see diagrams) A gumnut on a rock; a twig through a leaf; a pair of torn leaves; grass stems bent in a consistent direction; or tied in a knot; or crumpled; a scissor-cut dead leaf; a scissor-cut fresh leaf; bark hanging from a barkless branch e.g. gum bark in a Banksia; a twig drilled into the ground; a leaf wedged into bark at eye level; a triangle of twigs; a half broken bent twig hanging in a shrub; an opposing pair of gum leaves; a trail of sticks; windblown debris which is reversed; upturn a rock onto leaf litter
- Deliberately plant your boot print when you can. Step on top of older footprints, deliberately.
- Use fresh material so that your trail may be 'aged' according to the fading or drying e.g. of green leaves or mud.
- Position markers so that time, animals, wind & rain will disrupt them (at a recognisable rate). E.g. a teepee of sticks.
- Try a snow cairn; chalk marks on rocks; cotton tied around a branch; a stick dragged in the dirt; a scratch on a rock; balls of mud; charcoal or charcoal marks; tiny shreds of paper.
- Choose one method and stick with it, with a backup method in reserve.
- Double-mark a trail e.g. when doubling back, by duplicating it.

Tracking Someone Else

- Look for the **absence** of: spider webs across the path; dew on overhanging branches. To discern a foot-path look for the absence of fallen branches; leaf litter; lichen & moss; old wombat dung on prominent rocks; unbroken sticks; untrampled seedlings; undisturbed soil
- Look for the **presence** of freshly disturbed stones/soil crust/puddles, plus footprints.
- Examine puddles: for wet edges — indicating a recent forward swash; examine for muddy water; look for muddy water in the deepest layers — indicating disturbance some time ago; for wet stems above the water — they have been trodden in and have sprung back
- Overlapping fronds across a path — they 'remember' which way they were last pushed through.
- Don't step on the boot-prints or tracks you are trying to decipher—you often need to look at them again.

Botanical Tell-Tales

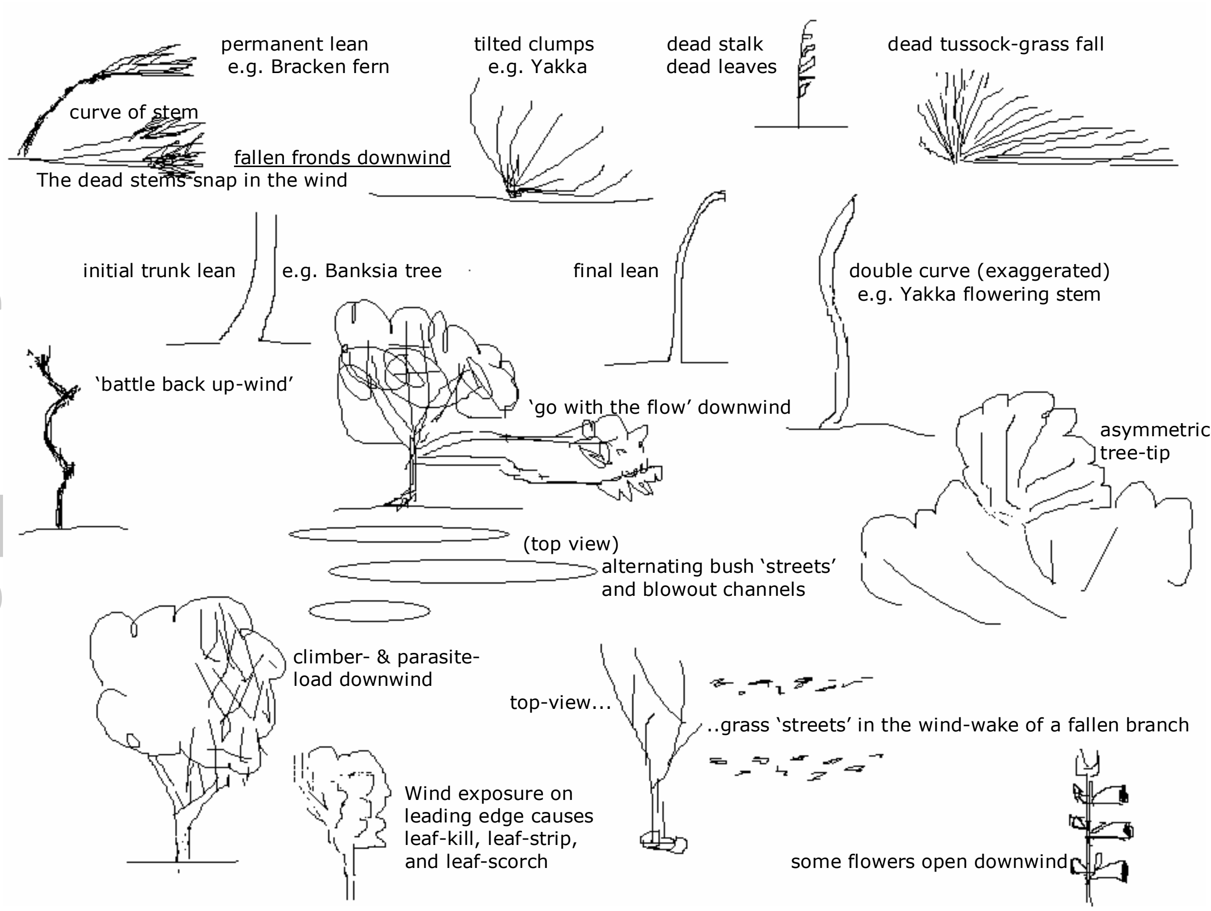
permanent lean
e.g. Bracken fern
tilted clumps
e.g. Yakka
dead stalk
dead leaves
dead tussock-grass fall
curve of stem
fallen fronds downwind
The dead stems snap in the wind
initial trunk lean
e.g. Banksia tree
final lean
double curve (exaggerated)
e.g. Yakka flowering stem
'battle back up-wind'
'go with the flow' downwind
asymmetrical
tree-tip
(top view)
alternating bush 'streets'
and blowout channels
climber- & parasite-
load downwind
top-view...
..grass 'streets' in the wind-wake of a fallen branch
Wind exposure on
leading edge causes
leaf-kill, leaf-strip,
and leaf-scorch
some flowers open downwind

Discuss It

Let others open your eyes: "What can you see about this?" "Where does it remind you of?"
Ask Questions: "How do you know where you are?" "Could you get back here again next year?"
Make a game of it: "What else can you notice from here?"
Express the specifics: e.g. "That boot-print has zed-shaped ripples on the right and ess-shaped ripples on the left.
The heel patten is a smaller version of the front half, as per normal."
e.g. "This tributary lies northwest/southeast and points to the left of the peak."

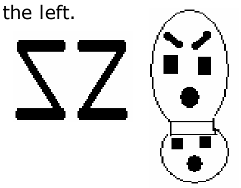

Many aspects of navigation yield to these simple tricks…

discuss it express it out loud make it the topic of conversation
delegate the tasks between yourselves
keep yourself alert
make up your own names
collect the details integrate them

Collect the Details

Identify a Specific Place by its Specific **Collection** of Details

E.g. "A low conical hill in flat terrain south of Mount Cameron"
That is not 'interesting'. The bits are not 'peculiar'. But the collection of the bits is unique.
E.g. Plant and insect species are often so similar to other family members that no one character is 'characteristic' enough.
The species will be identified by a set of characters.
E.g. Is this our old road?? Does anybody recognise it from last year?? but it has…

trend e.g. toward sunrise
slope e.g. mostly uphill
straightness e.g. can't see for more than 20m mostly!
traffic intensity e.g. many fallen trees across it, and encroaching bushes
length e.g. it goes on and on and gets nowhere!
vegetation e.g. set amongst old gums with yakkas under
ends e.g. it came from the saddle and ends up behind the peak

If you think you recognise it, by the combined characters, you will probably be right.
E.g. Make one name for the big tree on the crest; another to name that crest; another to describe its relative position…
and soon you can't mistake it…"the big tree with the south facing hollow on the port-side Dam-wall Ridge downstream of Tree-fern Crk.
The add-on bits are part of the name of each bit i.e. "the Big-tree Crest" (downstream of Tree-fern Creek) (near the dam wall).

The smallest collection is two, but that is a powerful combination, e.g. a peculiar log, plus a recognisable hollow trunk, is almost unique.
The combination is the peculiarity to remember — a third peculiarity.
The geometry of the combination is a fourth and even more powerful memory aid… "This log points to that hollow".

Include What You Can't See

For instance, have you noticed that there is no bare ground showing? That you can't see farmland? That there are no intervening ridges?

Break of Slope/Make of Slope

The 'make' of slope is the **steepening** as you go downhill ie the "slope" is a down-slope. The best views are at the most abrupt 'make's

The 'break' of slope is the first **flattening** out as you descend. The thickest soils occur below the 'break'.

The make and the break refer to points which separate different 'slopes'.

There is a difference between a making, or eroding slope and a breaking, or sedimenting slope…

A ridge-top is a make of 2 slopes, in cross-section, but any abrupt steepening can also be called a make of slope.

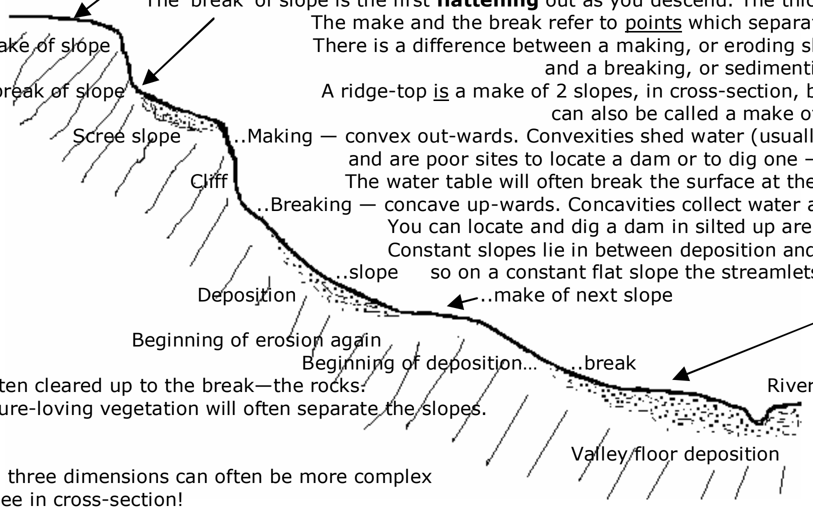

..Making — convex out-wards. Convexities shed water (usually in all directions) and shed silt and are poor sites to locate a dam or to dig one — there is bedrock below.

The water table will often break the surface at the more abrupt breaks.

..Breaking — concave up-wards. Concavities collect water and collect silt from above. You can locate and dig a dam in silted up areas.

Constant slopes lie in between deposition and erosive scouring so on a constant flat slope the streamlets can 'braid' together

Paddocks are often cleared up to the break—the rocks.
A band of moisture-loving vegetation will often separate the slopes.

River—erosion on the banks = convex — 'break' at water level; concave below it

What you see in three dimensions can often be more complex than what you see in cross-section!
— When convexity changes to concavity there is a **momentary straightening** — a 'break of slope'.
This is also called a 'point of inflection' in the curve.
It's not just a meaningless mathematical quirk; it betrays a set of real changes, resulting from the erosion history.

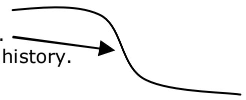

Scree Slopes — Tread Carefully!

The boulder-field at the base of a cliff, gives you a false sense of security!
Each rock seems so solidly jammed in place, but the whole field could 'slump' and slide all the way down, if given too much of a shake.
That is why there are no old trees growing in landslide-prone slopes!

The essence of a scree slope is, like a pile of gravel: It is in a continual state of critical instability, as evidenced by the straight-edged slope (flat but not level), about 36°. 36° doesn't sound like much, but up a sandhill, it puts you on all fours!
The rock surfaces look 'fresh'. Soil doesn't accumulate.
Tread carefully. Distribute your weight on all fours. Tread on bigger rocks.
Choose the cliff-base — it has hand-holds — or the scree-base — it has a solid bottom.

Mark Your Trail More Permanently and Ostentatiously

E.g. when you are lost or in trouble or in doubt

- A teepee
- Mud upon a tussock — quite unnatural; A tussock upon a tussock!
- Hang a piece of clothing from a prominent tree limb
- Dig a hole and make a hump

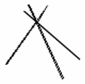

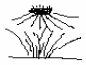

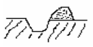

(Direction of travel is ⟶)

Mark Your Trail Invisibly

E.g. so that other walkers who are not lost, don't disrupt your markers, out of a sense of bush etiquette.

- "One mark every 50 paces". E.g. one pointed pebble. You may like to use double-paces. It is almost impossible to notice accidentally, but impossible to miss if you look.
- "Every 10 paces, as an option" added to your scheme — to cope with junctions and bends.
- Place a pebble on the ground, on the onward side of an existing cairn. You can recognise it, and scatter it, when you come back through the fog among a maze of similar cairned routes.
- Put a small mark at the base of each track marker (e.g. under blazes-on-a-trunk or ribbons-on-a-branch) — you can easily find it, and know that you have been along that track.
- "Half way between (2 markers)" or "10m North (of a juncture)", can remove the marks from obvious view or significance.

A pebble Code — Keep It Simple

All codes have to be memorable to all users — complications never survive the years of disuse.
I.e. use 'one' to 'four' only, preferably only 'two' and 'three'.

Tread the pebbles into the ground for clarity & permanency.
Upturn the pebbles for visibility, and **point the stones** appropriately.

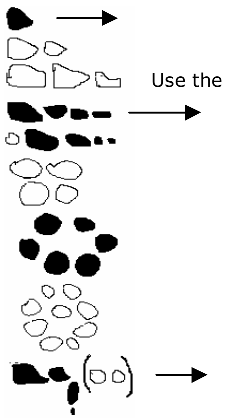

Use the smaller pebbles at the front.

One: "I woz 'ere, going that way"
Two: "Me go that way"
Three: "I'm lost, exploring that way"
Four: "Look for a message, there"
Five: "Someone has seen this message"

Square: "Don't go this way"
Circle: "Oh" for "I'm OK"
Bull's-eye: "I will be back here soon"
Junction: "I have turned off (and returned, then continued)"

Double marking: Close the pebbles up the second time past, until they touch.

Combined Effects of Wind, Salt and Light

When assessing wedge-shaped shrubs, leaning trunks, tilted canopies, etc., try to distinguish wind-cut from salt-cut and from light-lean. But beware, in any natural system, it would be easy to over-generalise from a few known specifics, or falsely extend an inadequate interpretation.

Here is a typical wind-ramp. The leading edge consists of dead twigs.
The vegetation gets greener and thicker and taller downwind.

But you cannot easily tell whether it is the result of salt-laden wind, or strong wind only.

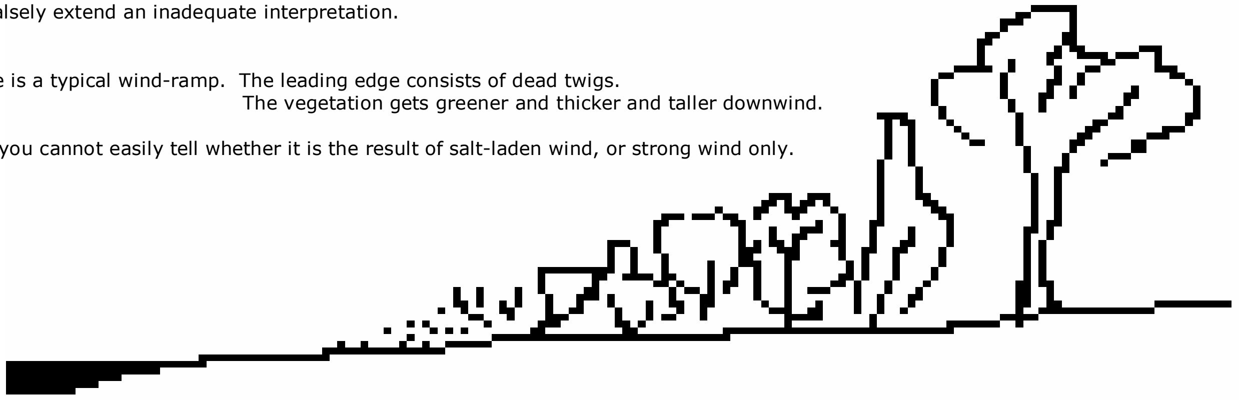

Salt-cut will show you the direction of the prevailing onshore wind only.
Inshore, wind-ramping will show you the direction of the prevailing wind, which may be significantly different in direction.

Botanical Asymmetry

Plants must respond to the wind.
If they did not deliberately push up-wind, we would be walking through a landscape of leaning-over plants.
Therefore it must be possible to 'read' the effect of strong wind written into the branching history of a plant.

The branches of a bush may grow according to the strong wind, but the leaves will still respond according to the asymmetrical light.
The two effects may or may not be independent and you will need to examine them carefully to see the difference.
Look for asymmetry in the green parts of a plant to show up bush-noon and look for asymmetry in the other parts for bush-West.
If there is a macro response to the wind, which will confuse the macro light-response, look for micro signs of a light response, on individual branchlets.

You don't "need to know your plants" beforehand. Instead you examine the species which present themselves.
You assess what responses they make to light, wind, shelter, etc.
E.g. Do the stems colour up in sunlight? If so keep looking at that.
Do the branches throw out more leaves on one side? If not look for something else.

Extend Your Familiar Range Systematically

Explore in every direction, in ever longer legs, radially. Leave recognisable marks alongside recognisable things.
Then explore in concentric circles to make sure that you can recognise individual trees and rocks.
You know you can explore further within the endless sea of scrub, if you have 'familiar friends' which you have seen repeatedly.

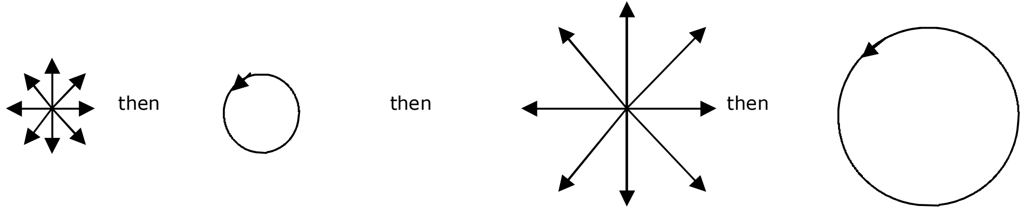

Markers for future recognition e.g. on your own property…

Plant a tree Graft two branches together Suspend a forked branch into a forked trunk
Arrange logs e.g. a parallel heap Arrange rocks e.g. a cairn, or a rock on a stump, or a rock 10 m south of the biggest tree
Dig a trench Etch a mark Paint a rock Leave something solid behind
A stake A signpost A numbered metal tag A path A Campfire-place or Barbecue

Hints

It's the Campsites You Don't Use, that are Easiest to Forget

> I'm forever asking: "Can anyone remember" whether we can camp between here and there?
> Take notes of potential campsites. It makes your trip-planning more flexible, next time.

In general, take the opportunity to spy out and note down potential future trips — side-tracks, ridges, climbs…

Leap-Frog Walk in New Territory

I like to walk a new track twice, anyway. It is no big difficulty to take your time, once, and achieve better familiarity, by leap-frogging.
Explore 2 legs forward then one leg back, and you cover the ground 3 times on the way out, once more on the way back.
4 legs forward, 2 legs back, and it's like walking a return track 3 times, but it gives you a much more solid handle on each section.
You gain what you want, appreciation and re-preciation, and what you need, familiarity and sequence memory.

The best way is to combine this with a mnemonic for each leg. See AA to ZZ for an excellent method.
Its also a good time to mark each new leg discretely.

Break of Slope — And Route-Finding

- The first signs of an approaching ridge-top when climbing include: the change from deposition (breaking) to erosion (making); i.e. onto outcrops; into exposure to the wind.
- Take a rest at the break of slope — before climbing upwards.
- Get your bearings at the most abrupt make of slope, where the view is clearest.
- Look up- and down-hill for the lowest skylight beneath the trees, to reveal the change of slope.
- Avoid contouring around the break of slope — it is tempting, but — it is steep and unstable.
- In rocky gorges, or on any creek-bank, or around any permanent water, look for a break of slope just beneath water-level.
  It may be safer, than rock-climbing higher up, to get your feet wet, and use such ledges. The slope near the water may be less steep, at least on average on jagged slopes; you can't fall as far; your landing is softer (wet) — if you 'keep as low as possible'.
- On steep gravelly roads, stick to the 'crown', or the 'gutter', to avoid slipping.
  The crown will snake from side to side as the road turns.
- Slopes are awkward to walk on—steeper than the ridge or creek, so,
  when you drift up onto the sides of the creek you want to follow, you will be forced further up — onto the ridge.
  When you drift off the crest of the ridge you want to follow, you will be forced further down — into the creek bed.
  In this way 'minor' detours (e.g. around an outcrop, on the ridge or in the creek bed) may result in major redirection orders!

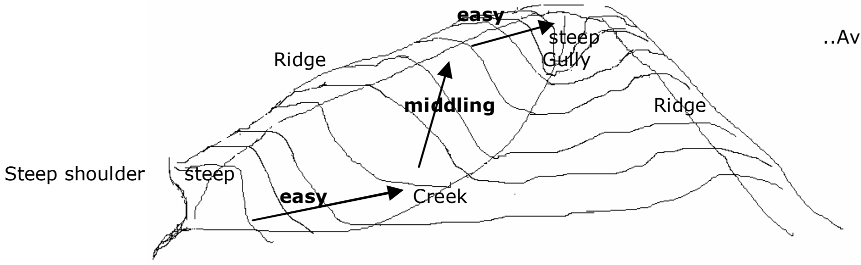

..Avoid the steep bits.

Keeping Track of the **Bends**

Don't just follow your nose — follow the bends!
There have already been some hints on keeping direction sense:"Walk across Australia!" & "Integrate the Jetstream and Earth rotation"
Plus the "Sky to Land" and "Land to Land" navigation systems (p.29e,34d,38d,42d).
Following a track can be mindless. But following the bends in the track means that you understand
— that you have changed direction, and by how much.
We have an ingrained bad habit — that we walk 'away from' the starting point, which is therefore 'behind' us.
Only when we 'turn around' is it 'ahead' of us on the track.
We usually only notice junctions, features & hunger, not distance or direction or exact time.
The commonest mistakes of guesstimation, are failing to notice a track curve,
and **underestimating** exactly how far a gentle but insistent bend will **turn** you around i.e. simple disorientation.

Hints: Notice the bends, with your eye on the sky or the skyline. Establish the habit of looking back.

Desired Equipment

- Compass — look at it often and interpret it
- Polaroid sunglasses — look for the dark blue band
- Radio — use it to keep orientation if you don't have a compass
- Map — orient the map to the landscape, and it acts as a rough compass
  — lay a rhumb-line through the bends, as an average course, as a reference 'track'…

Logging Constant Diversions

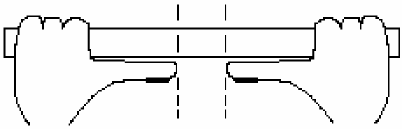

A stick or string helps you measure direction changes…
Held at arm's length, you add or subtract, as you decide to change course left and right.
Thus you can revert to an intended course after a short 'blind' section,
and average out the course changes even without a compass, or being able to see the Sun.
You could use the skyline as a reference direction, or lineup two tree trunks for each change.

A rougher method is to counterbalance your diversions.
If you go left-around the rock; next time, go right-around the tree.

Keeping track of countless bends is one of the few major problems in navigation
— like keeping a log of distance covered, having no location sense, and trying to stay alert.
So aim for an independent cross-check e.g. a landscape fix or a distance log.

Think like a Plant

E.g. Imagine you are a small ground-cover plant. In a desperate attempt to not be eaten or trampled to death you will probably branch and leaf out in all directions to simply survive... So don't expect much botanical asymmetry underfoot.

The idea is to become an amateur ecologist. For instance, you could work out the pollinators available, the dispersal mechanisms used, the biogeographical zones and overlaps, and the associations between species.
Therefore think about the geology, geoform, latitude, altitude and soils.

Temperature & Relative Humidity explain a lot of botany.
E.g. very cold morning winds are also very drying when they warm up later in the day.

Altitude, for instance, affects the species mix, so that with practice you can gauge your elevation above sea level.
You notice when sub-alpine species kick in e.g. Telopea (Waratah), then when true alpine species begin e.g. Richea (Scoparia).
Individual members of genera have altitude preferences and so are helpful to distinguish altitude
e.g. various Coffee berries (Coprosma) or Pink Berry (Cyathodes).
N.B. Since the prime effect of altitude is to drop the temperature, this is really a latitude-dependent temperature effect.
I.e. Going further polewards, the altitude preferences will change.

'**Aspect**' (i.e. South sloping or North facing) will be important in interpreting the presence of particular species.

Living stresses explain simple botanical systems.
E.g. In the desert, late afternoon is the baking-hot-sunlight side (cf noon in forest). Trees will be shiny leafed — to shed excess light (cf searching for more).
So consider the prevailing winds, salt spray, max. and min. temperatures, wet season, disturbance, trampling, fire, clearing, introduced competition or diseases, grazing, salinity, mineralisation.
Pioneer Species — These are the easiest to identify, because they are lonely.
Their preference is the saltiest, hottest, driest, highest, coldest, poorest, windiest, sand-blown, most flood-prone, …
If life becomes too easy, other species will smother them out. Hence, look for those that don't mind the hot baking Sun, etc.

There is a height-**succession** of plantforms grading from **lowest to tallest and deepest rooted**.
Ground lichen; moss; grass; herbs; creepers; shrubs; trees; ' emergents '
The depth of soil is crucial. The taller ones shade out the humbler ones, and their deeper roots seem to starve out the shallow-rooted ones from below. E.g. mosses need no soil and not much light, but will quickly be overgrown if these are provided.
Other factors keeping plants humble are trampling & grazing, fire frequency and dry season. Grasses survive all these stresses well.

Competition between species explains the more diverse vegetation.
E.g. Lichens overrun each other, and leave the rock surface, to become 'leafy' or fibrous or pendulous.
Lush growth betrays the most comfortable living conditions. E.g. rainforest; the underside of a rock in the desert; sheltered microclimates — crowded with species. This is opposite to pioneer species.
E.g. Where excess sunlight is the 'living stress', lush growth will show up on the shady side.

Draw the Skyline, the Ridge-lines, the Outlines
Label it with the 8 main directions

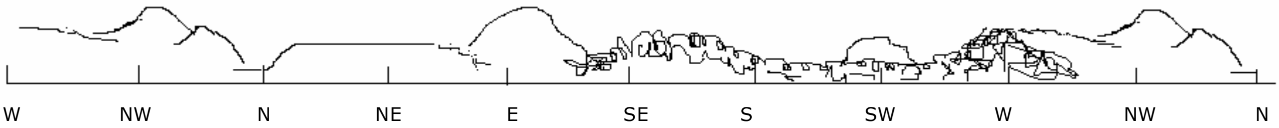

Talk the panorama into words. "Nor-Norwest Hillock, in front of a larger hill on the left, runs down into a gully behind NE– Plateau.
East Mountain is behind NE Plateau and drops away behind the vegetation south of us.
Distant Peak pokes up behind in the SW, and the closer skyline reappears in the West,
falling steadily to the NW Dip before the larger hill."

That took one minute. You do have time to look, notice, talk it through, and even draw it.
Marine navigators know that they have to.
Don't just imagine doing it, or do it silently, or do it half-heartedly — actually do it.
Unlike an avigator who uses a layout map to recognise roads, rivers, ranges, coastlines, etc, from the air,
a walker needs a nose-eye view…
MAPS DON'T DRAW YOU A NOSE-EYE VIEW — SO YOU HAVE TO!

Examine the Skyline BEFORE You Walk up to It

That might mean looking behind you, six hours ago, or to left or right, right now,
because the skyline ahead of you rapidly disappears behind foothills, just when you want a clear view of it!

A clear view of the skyline is uncommon, so you need to take advantage of any skyline view while you have it
— and that is often anywhere else except where you want to look — ahead.

Circuit walks and return trips mean that, if you haven't habitually looked at the surroundings abreast of you & behind you,
you won't be able to recognise what you are walking up to later.

One way to recognise the skyline later, is to use it
e.g. use it to keep track of bends; to keep track of progress; to disclose hidden valleys (by parallax or haze).

Another way is to name what you use e.g. "the hill behind the dam wall from camp 2—Dam Hill"

Mnemonic: "Down in the Mouth"

"Downstream" is "Down to the Coast". "Which coast? Which river mouth?"
Until you have answered that, you are disoriented.

You pass a creek-bed or a flowing stream: "Which way is it flowing?" is only half a question. "Where is it going?" is the other half.
You turn and head 'downhill': You think 'Down' — then say: 'Down in the Mouth — Which mouth?'

If you also ask "Upstream? To which ridge?" there are many possible answers because upstream is divergent.
Nonetheless it is good to know whether a river reaches to the very edge of the valley — whether it is a 'major' tributary.

Which way is the flow?

In dry, flat creek-beds, examine the tilt of the rocks… If any rock ever tilted up into the flow, it would be flipped over.
Examine the silt build up — on the sheltered downstream side of any obstacle.
Look for flood debris — matted against the upstream side of branches.
Consult your artificial horizon, if you have one.

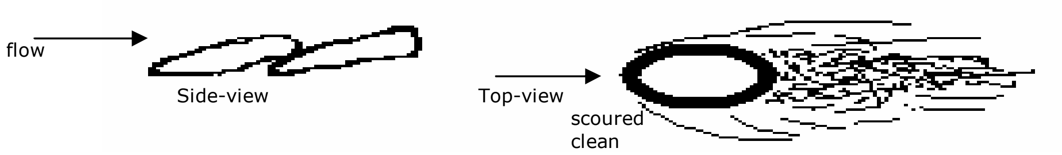

(This is much like grass tussocks in the wind.)

Hint

Don't think "this side" & "the other side". It only makes sense for the one main lowest riverbed.
Think: Up-catchment/down-catchment/cross-catchment, overall, & Port half/Starboard half of the whole catchment in question.
(as seen from the coast, looking upstream)
Think plural: "the other sides"/ "another side" in the network…

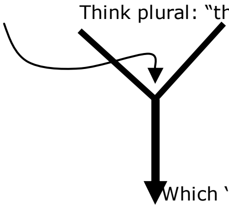

Which 'side' of the river are you now? "I have crossed to another part of the drainage system"
Another possible answer: Examine the map for the contour heights of each channel for the lowest one.
"We are now in the starboard half of the catchment"

The Difference between 'Direction' and 'Track'

'North' is a general orientation reference **direction**.	But	'North from here to there' is a specific route — a 'bearing', 'line' or **'track'**.
'Stick to a direction'/a 'heading' means e.g.'toward sunset'		'This road points toward that Gap' will never change — it is a fixed 'track'.
It can start from any point,	cf.	'Stay 5 miles off the lighthouse' is a (curved) line of position — a track.
and sidesteps around obstacles, without bothering to return to track		'From this stump, past that tree, to the rock' — sticks to a fixed straight line, despite detours.

So the final practical 'interpretation' of, say 270°T.,
is not 'West', not 'sunset', nor 'Pot-set', not 'THERE!', not a drawing in your notebook, Not 'How hot it must be in Baghdad!'
but is a navigational track... "From this place to… that particular tall tree" "Then we will get a line on another point directly West"

I.e. Interpret a 'compass course' , finally, into accurate, achievable 'legs'—'on track'.

Continuous bends

Riverside tracks are especially sinuous, yet slowly so, so that they are deceptive.
Beaches near headlands, and lake edges, are especially insistent in their constant bending.
Contours, constant horizontal angles, and constant distance-off, are all continuously curved. Aqueducts follow contours.
Consequently we are necessarily **continually disoriented**, in following them.
No sooner do we get our bearings straight, than we lose them.
Logically then, you need to orient, not by a direction e.g. the Sun, but if at all possible, to **orient by a 'track'**
i.e. a particular path on the ground — it doesn't have to coincide with your own track, but it often will
e.g. from headland to headland. (The beach will then lie wholly on one side of this shortest distance 'track'.)
A visible, noticeable, landscape track is best e.g. the biggest river valley around.
You now can feel continuously oriented to something fixed, something which you can see & 'feel in your bones'.
As a second best choice, choose a fixed compass course e.g. a rhumb line, and watch where it intersects the landscape.

Follow the bends by anticipating the trends

Before you drive off, or plunge into the twists and turns of endless undulations, examine the map for your average course
— to give you some overall reference track ahead of you. It may be mental, or visual, or by compass.
You can see when you turn to the left or right of it, direction-wise. You can feel it when you drift too far to one side, track-wise.
Interpret each leg meaningfully. E.g. "toward the morning Sun"
(the time of day doesn't matter).
E.g. "Keep Mt Wellington on my port quarter".
E.g. "From the pine trees to the river mouth".
Divide the journey into recognisable sections, if it helps…
Make this a first priority, before studying any details.
Preferably, mark them lightly on the map.

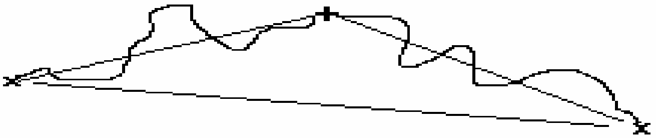

"Mosses Mop up Moisture"

◦ "Mosses grow best on the shaded side of a tree" or a rock, ◦ where there is not enough light for competitors
◦ where there is little soil to hold the moisture, ◦ and it doesn't dry out ◦ and often there is too much moisture for other plants.
Mosses actually monopolise the moisture
In hot places it is the late afternoon which is driest, so it is **the sunrise side, in a sunny climate**, which is the moister side.

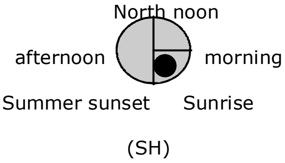

Note: Dead moss is black and brittle.

Examine fallen logs — the shady side has a sheltered dark crevice...

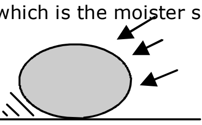

In a hot climate, it may be more accurate to say,
"Mosses grow best on the sheltered side"

Examine ditches — One side might be scorched, the other, lush.
Examine bumps in the ground — One side is greener, with more moss and taller grass.
Examine underneath shrubs — for the moister, mossier, greener side under the bush.
Examine intertidal rocks — for moss-like seaweed growths on the shady, wetter side.

"Lichens Like Light"

Lichens don't mind being dried out, so sometimes they are opposite to the mosses. However, not all lichens are so robust.
Lichens are fungi which live off algae entrapped and protected within ("Fungi Feed on Food"). They may live on mineralised substrates.
They are therefore very colourful. So the best way to learn lichen-likes is to open your eyes to the colours of each direction…

You may see white blotch lichens looking downSun but not upSun.
Yellow blotches may all be only looking uphill, not downhill. Leafy lichens like the shady enclaves, not the Sun-drenched flats.
Crusty greys and fibrous green might cohabit the underside of branches. "White paint" may only grow an the top of any rocks.
Orange granite might be just above high tide level, etc.
Diffuse lichens colonise the crumbly side of a sunbaked rock. Flat-topped rocks are trampled and won't have thick lichens.
Look closely at surface colours of rocks and barks and soil-crusts— they may be due to diffuse lichens.
Look on the branches of shrubby growth — the lichens will grow best on one side or the other.

The idea is to start looking locally, before you need them. It's a nice hobby. And hobbies keep you awake and observant.

"Fern Fronds Face the Front" i.e. the sunny side

Looking down-Sun, you'll see an array of green topside fronds flat on.
Looking up-Sun, you'll see more undersides & yellowness & edge-on fronds. It may not always show, but is clear when it does.
Tree ferns retain their lower branches longer on the sunlit side, and so seem to tilt toward the light.
(But in a windy place you will notice that the permanent lean, curve & frond-fall of ferns are downwind.)

Make a Mental Movie

Memorise a spectacular cliff looming closer overhead.
Watch the shape change. See the foreground shift. Notice the colours and contrasts. Look at the landscape flow.
Describe the silhouette. Peer into the deep shadows. Scan the skyline, looking at each interruption or change in trend.
Appreciate the hazy distance.
Follow a complete outline around, naming its shape.

Don't forget to "record" the sound-track!

Initially try a full production only for beautiful memorable scenes.
With practice you may find that your memories stick for more and more locations.

Take Mental Photographs

Familiarity and Recognition are primarily visual. Your visual impression is adequately vivid, but your recall of it is faulty.
So put handles on the memory and you will be able to revisualise e.g. "where we found the weather balloon".
To give your memory something to recall, take a detailed look.
The tricks are simply to give it a go, and to notice **the specifics**:

The colours, shapes, skyline, peculiarities, sequences — Talk about them.
Play a game: "Your turn: what else can you add to the picture"
Overlay some geometry on what you see — e.g. the log is parallel to the path and as far from it as the log itself is long.

Now close your eyes and see what you can reconstruct. Open them again and notice where you were wrong or were lacking detail.
Repeat this a few times until you could picture it in bed the next night.

Mental photographs are, after all, **your take-away memories**.
They are what names name; what "directions" talk about, what 'recognition' and 'familiarity' is based upon.

Take a Mental Photo of Each Change of Track

Where one subsection ends and another starts
— Look forward onto it, with the old terrain in the foreground and the new terrain joining on to it.
Now walk past the join, Look back, take in another view, with the old terrain now in the background.

This adds visual sequencing information to individual memories.

The Inside of a River Bend is Flatter Ground

The inside of a meander may be wider, sandier, marshier or of different vegetation.
On the outside of a meander, cliffs form, through ongoing undercutting. See also p.56a about the steeper North bank of a river (SH).
Consequently plan ahead and work out when to swap sides to find the easiest walking.

If the track becomes steeper underfoot, that would mean that you could be on the outside of a bend.
Check to see whether your heading has changed with the bending river.
If the track becomes flatter underfoot, this may indicate an inside bend (or a river junction)

Rivers Bend! Don't Forget to Look Backwards

Obstacles cause creeks to change direction. Outcrops of bedrock and creek junctions, for instance, cause bends, and this may be shown up ahead of time, by ridgelines pushing in closer, as you follow a riverbed.
Expect to meet a new creek clinging to the base of any ridge or shoulder which crosses your path, and also expect to change trend.

In addition, to that cause of bending, rivers follow contours and so usually are on a curved course anyway.
Furthermore, in the flats, rivers never follow a straight course but will 'meander'.

So if, while following a creek valley, you notice the slope underfoot changing, or the vegetation altering, watch out that the creek doesn't leave you, by bending away from you!

If you find yourself in the middle of a larger than normal river flat, suspect a river junction, and be very careful interpreting the trends.
E.g. you may meet 'the river' 'coming back toward you' when you were trying to go downstream!

So you can easily lose orientation while the winding river-track follows the lie of the land and gradually wends its way around corners!

Look ahead to catch the trend — and interpret it.
Look back, to see the extent of the curve.

Rivers Meander. Ridges Run Straighter

A ridgetop route will often prove shorter than following a winding river.

Keeping a Compass Course, by Legs

You can't see far ahead, but you can line up a rock or a bush which is in the right compass direction, then walk up to it.
It doesn't matter how you get there, as long as you get there.
Repeat the procedure, even if you can see far ahead — because, if you are merely following a compass needle or a distant landmark, you can "side-slip" off track.
Your eyes are freed from the compass needle while you walk.

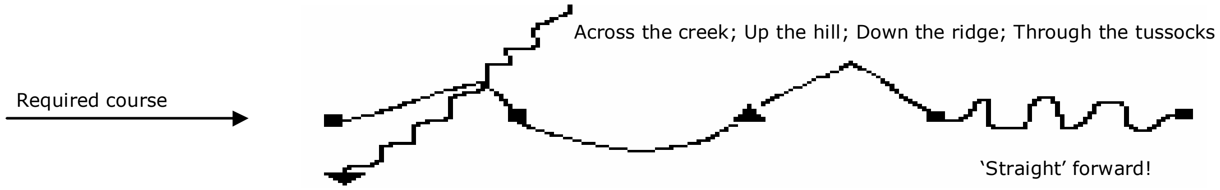
Across the creek; Up the hill; Down the ridge; Through the tussocks
Required course
'Straight' forward!

Turn Smooth Bends into Sharp Jumps

Your brain wants something to "latch on to". A smooth bend is just the opposite! So make a series of straight-line "legs".
You "let go" of the last leg, and make a definite jump, to "latch on to" the next leg, and your subconscious smiles, and remembers.

E.g. 'compass North'; then 'to the tree'; 'to the peak'

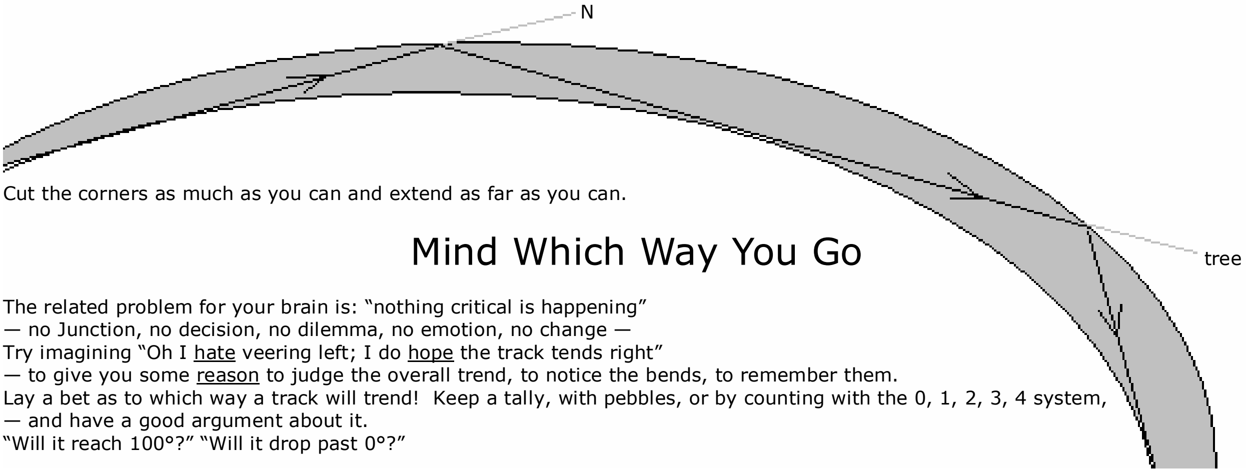

Cut the corners as much as you can and extend as far as you can.

Mind Which Way You Go

The related problem for your brain is: "nothing critical is happening"
— no Junction, no decision, no dilemma, no emotion, no change —
Try imagining "Oh I hate veering left; I do hope the track tends right"
— to give you some reason to judge the overall trend, to notice the bends, to remember them.
Lay a bet as to which way a track will trend! Keep a tally, with pebbles, or by counting with the 0, 1, 2, 3, 4 system,
— and have a good argument about it.
"Will it reach 100°?" "Will it drop past 0°?"

Bush Noon— The Sun is Always out, If You Know How to Look

Trees & shrubs live in the light, day in, day out. You would expect that they should respond to the light, and lean or reach out to the strongest light, and throw out more branches and more leaves to the bright side.
The disposition of branches and leaves stays in view even when the Sun isn't out, and it lasts longer than a cloudy spell.
Although the Sun is a huge factor in explaining foliage asymmetry, be aware of other factors…

- A lake or sea can double the light intensity on one side, pulling the foliage toward it.
- A clearing will let more light in on one side.
- Down slopes allow the trees to see more skylight on the down-side, and less skylight on the up-side.
- Other trees can shield a tree from the Sun. Look especially for recently **fallen** ones on the ground, to explain standing branches
- The tree itself shields its own branches, so that most growth is simply outwards.
- Unrelated factors like gravity, wind, and wood collecting, alter the shape of a tree.

However, if you **average out** all the other influences, you should be left only with the noon-side Sun's influence — "Botanical Noon".
With adequate practice you should never again be "lost in the bush" because it is the thick bush itself which will tell you where North is.
You may have an initial hopeless feeling of "I can't see any meaningful asymmetry!" but the feeling will be temporary (eventually!).

Of course, this works best only where the Sun is always one-sided—in nontropical latitudes.
Hint: If a tree confuses you by pointing in the wrong direction, go over to it to learn why.

False Asymmetry

The slope of the ground
may be forcing the asymmetry,
rather than the differential in the lighting.
In the second diagram, there is asymmetry:
A is ungrassed; B is grassed.

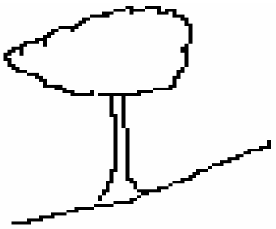

False Symmetry

C and D are both grassed.

True Symmetry

B and C are both grassed.

True Asymmetry

A and D are differentiated from each other.

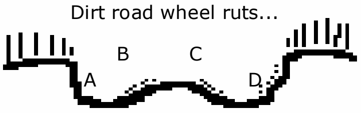

False ("Chalk and Cheese") comparisons abound in Nature. True asymmetry is there too
Your task is to sort it all out.

E.g. Shaded foliage sees mostly blue skylight, which should make it long and lanky, compared to the red-end light from the Sun

E.g. the West side shadow is frostier for longer and the East side shade is cooler in the p.m.

E.g. the North side soil will be warmer deeper than the South side of a thick hedge

Look on the ground beneath a shrub, or log, for asymmetrical shade and shadow effects.

Remembering the Sequence, Needs More Than Logic

Logic is logical and helpful e.g. The bridge was after lunch because we were worried about the crossing over lunch.
The red rock was after that, because the second creek was dry — it had all those beautiful red lichens in it.
Then the funny tree was where we had chocolate, i.e. 3 o'clock stop.
But logic by itself is not 'sexy' enough to be emotive and memorable.
So deliberately add some emotion…
E.g. "lunch", "worried", "beautiful", "red", "funny", "chocolate"
You may need to add a song and dance, a gimmick, a clue to the time or sequence, distance or linkage.
The trouble with a memory gimmick, is that you focus on the gimmick, not on the memory itself.
E.g. If you use an acrostic, remember to feel the sequence of it, as well.

DO Something NUMERICAL

E.g. Lean one stick against "Walking-Stick Cliff" at one o'clock.
Sing two songs at "Sing-Song Pond" at two o'clock.
Make three cairns at three o'clock at the "T(hr)ee Junction"
Make four arrows on "Four Arrow Slope" during four o'clock.
Do it. Don't just imagine yourself doing it.
E.g. Scratch numbers in the ground, in sequence, as you pass points of interest.
e.g. "5FF" (the 5th stop, with the flat floor); "6SS" (six sizzling sausages—the fireplace).

Make a SHAPE; Match the SHAPE

Every or location in a forest track or in open country can be recognised from its **peculiar shapes**. Every bend or fork can be **matched to** the forest, or the shrubs and stones, so both the bend or clearing and the surroundings can be remembered, with joy.
Locations: What does this fallen log **point to**? What is **parallel to** it. *Place* something so it *does* **make a shape**.
How many shrubs do these three trunks enclose? What **face** or **picture** can you make out of the bushes, rock, clearing and the creek?
Do you see that **line** of tussocks; that **clump** of trees half-way up the slope; those parallel gutters; that oval of sand-dunes? Where does it drain? What **direction** between those **paired** stags?
Bends: ~Look for a tree trunk **leaning** over at the same angle as the bend of the track, preferably leaning over the track or bend and in the right direction. The tree stands out from all the others because of that. (Any lean at any angle in any direction will suffice.)
~Look for a branch **pointing** in the new direction, or curling, or **bending** in a way which mimics the track-bend.
~Look for a log **lying parallel** to the new direction, or any ditch, or line of rocks, or outcrop. (It doesn't have to match.)
~Look for a large rock nearby which **fits the shape** of the bend, or orient it until it does. (Matching shapes are always the best)
~Look especially at the tree trunk, or the shrub which blocks the straight line of sight ahead, or the shape of the fallen-over trunk in the way. E.g. "I'll fall into that black hole in the trunk if I keep going straight ahead! I'll have to swerve". "I need to go around that round stump". A tree **right at the bend** will be visible from both directions.
~Look for a forked branch which **matches the angle of a fork** in the track. **A series of parallel branches** at the fenceline. A couple of trunks **leaning towards each other** over the gateway. A tri-parallel set for a creek with two banks.
~Look for a branch which **points downstream**. Watch it for the next 50 years as it grows. A tree recognised is a friend for life.
~At a noticeable uphill pitch, find a leaning tree trunk and imagine climbing up it.
~Some leaning trunks can be interpreted in **reverse**… from top to bottom, to indicate the new direction.
The more you latch on to a peculiar shaped feature and the more you study its peculiarities, the better your recognition will stick.
Try to **double up** on the first peculiarity to clinch the recognisability. A **matching pair** of branches. A second confirmatory log.
You will have to **look backwards** at each bend if you want it to work well for the return trip.
Every bend and feature is a good chance to recognise another couple of trees. You stop at the bend and look around for your "friends".

Land-Form

Work downwards from a 'Big Picture' understanding. Start with a small-scale geology map, showing overall elevations.
Continental-scale domes (anticlines) & depressions (synclines) explain the overall drainage and watersheds.
Fault lines — shown by escarpments & cliffs — and individual ranges, interrupt the overall scheme.
The geological layers may be visible connecting across large distances.
E.g.

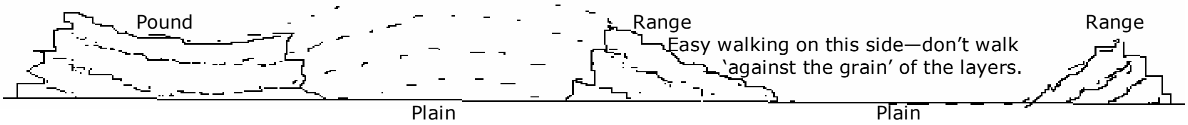
Pound
Range
Easy walking on this side—don't walk 'against the grain' of the layers.
Range
Plain
Plain

Layers become hardened and resistant to erosion when they are bent concavely upwards. They are compressed rather than fissured.
Convex layers are fractured and easily eroded, because they have been stretched. They will have been eroded away.
The present landform is a result of the erosion history on the previous landforms.

Hint

Most peaks have a steep side and a gentle side. Don't forget to make a mental note of that trend
in relation to the rest of the land and the other visible peaks
— you may notice a pattern, or simply keep your overall bearings better, or recognise the peak later, from afar.

What Makes a Cliff and an Overhang?

Resistant top layer

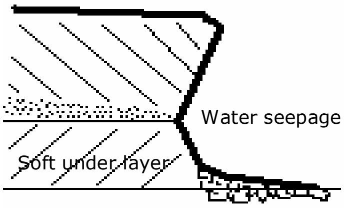

Reverse slope at the beach

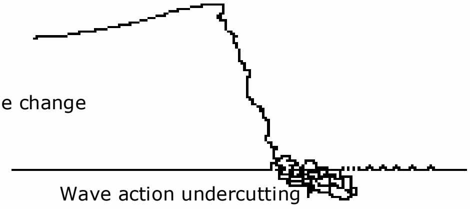

The outsides of river bends

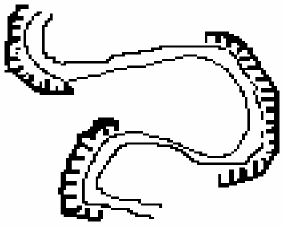

The cliffs migrate outwards

Waterfalls form at hard layers, at fault lines and at escarpments.
Look for outcrops striking across the land.

"End the Bend" i.e. Notice the Un-Bends

A bend is not simply a sharp corner, not just a change in direction: 'end of story!'.
It has a start, a curve, and **a finish, where it straightens out again**.
The curve needs to be **named** as a whole, **measured** as a whole, and thought of as **a whole bend**.
Look for the dividing points between successive curves

The arrows show where a curve to the right changes into one to the left,
(a Z bend), & vice-versa (an S bend). These are "points of inflection"

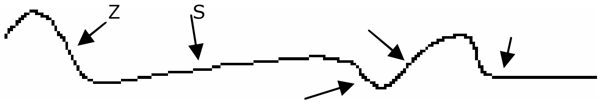

— a straightening out, and remaining straight, is the other possibility.
Incorporate half of the straight section into each adjoining 'bend'.
At the end of each bend, reorient yourself — look behind, at the compass, at the Sun and your shadow. Point to 'home'.

Name the Bend, By the Trend, of its End

Trend:	Left, toward Mt Michael	e.g. "Viewmont Bend"
	Right away from camp	e.g. "The Long Right Leg"
	Left into Lake Jordan	e.g. "Left it down the Drain"
Handedness:	Hard left around the spur	e.g. "The West Bank Left Bank"
	Half left from the creekbed	e.g. "The Left Contour"
	"Clockwise Climb"	
	"Right around the Bend"	Bends will often alternate.
Direction:	"Ending with Sunset in our face — Sunset Slap"	
	"Turning until the Full Moon Rises on our Right — Werewolf Way"	
Compass:	"North Now, to Alaska"	
	"Left-Sou-West"	

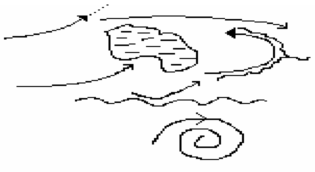

Add some description
to seal the uniqueness e.g... Skyline: "Lone Pine Ridge" "Three Tree Hill"
Detail: "Uphill; On the Left Bank; Just past the river crossing; at 10 a.m."

Notice and Name Perverse Bends

Unexplained bends, or ones which buck the trend, deserve a name.
Most bends are predictable — around obstacles — at creeks — around shoulders & outcrops — shortcutting
— contouring around sloping ground — at Y-junctions — joining with disused overgrown tracks
E.g. Seaside roads will hug the coast, go diagonally downhill but upstream (into a valley) and climb diagonally out of it, downstream — as a compromise between shortness and flatness.
E.g. For the same reason, tracks will often change direction at a creek crossing.

So name the curves that go against the flow…
E.g. "Sinister Switch" (Latin for left) "A Dextrous Diversion" (Latin for right) "Unorthodox Twist" (Greek: not "right" = left)
"A Wandering a Way" "Aimless Avenue" "It shouldn't've gone left!!" "The I'm-going-on-strike Uphill Bend"
"Mind Bender" "Morally Bent"

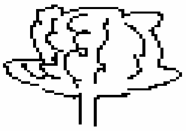

Higher Density of Leafiness on the Noon Side

From underneath, next to the trunk, you can look up in all directions and see which side is leafier.
Looking side-on you may see one solidly green side and the opposite side patchier in its cover.
Trying to look through the foliage to the back of a shrub, you will have most difficulty on the North.

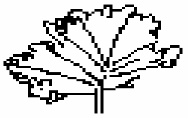

The Lowest Leafy Branches are on the Noon-Side

Looking from the sunny side you see a lower branch or even a wall of greenery, right to the ground.
Low branches on the shady side are often short, dead or only thinly covered with foliage.
Looking toward noon, you see more trunks, dead wood & grey-brown sticks.

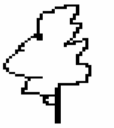

"It's the Shady Side that Shrubbery Shows up"

In practice you don't look for the noon side, but for the absence of leaves low on the pole side.
Looking toward noon, you see ill-developed shapes. Many plants will throw out branches and leaves in profusion in all directions, but you seek out the one direction which is different — the one less densely green — where the light is in short supply. It is **the 'light null' for solar direction finding**.
Look for where no vegetation leans: "No stems, plants, branches point in that direction".

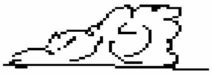

Tall, Steep and Ragged Backside

Shade-side branches often reach for the sky, and are more vertical on that side, making it steep.
The noon side will be more of a smooth continuous wall of greenery from low-down upwards.

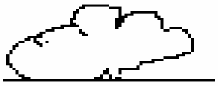

Wedge-Shaped

Looking away from noon, you see smooth and continuous foliage, slanted down toward you, often clinging right to the ground, like a wedge.
Looking toward noon you see flat tall backsides of the bushes.

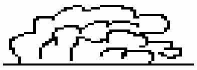

Hollow under the Shady Side

You can stand closer to the trunk, or at least see more of the trunk more clearly from one side.
Down low on the shady side is where the light matters most. There may be space enough for a branch but not light enough. On the sunny side you will see leafiness deep into the low recesses.

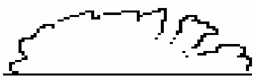

Imperfect Shape on the Shady Side

Many shrubs seem to adopt a perfect dome shape, until you look more closely.

Gaps on the Shady Half.

The backside of shrubs is often tall, untidy, and lacking leafy branches, close to the ground especially.

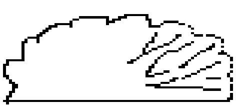

Dead Branches

The overall shape may be perfect, but one side is alive and green, while the other side lacks leaves.
"Dead" South. There is a scale-effect to dead-south: The larger bushes have more depth of foliage.
The more foliage up-sun, the darker down-sun. Only the larger or thicker bushes work best.

Use a Type of Memory which you are Good at

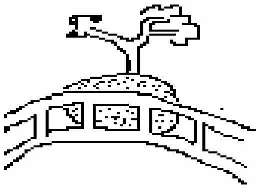

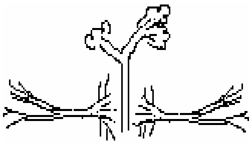

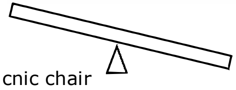

Combine the methods for the most memorable effect.
E.g. pictorial — join images. Make them touch or glue onto, squash or hold hands, follow logically or overlap
— use a stack of pictures to put incidents into sequence.
"After the bridge we came to a red rock and then sat down for a snack at the funny tree"
E.g. ditties — A Capital Tee of Tall Trees at Three O'clock.
E.g. invented "associations"— a vivid or absurd image like a Tasmanian Tiger eating that hill.
Paint it in partly related colours, with shapes and other senses involved, to give extra "handles"/clues.
E.g. Imagine successive scenes in an invented journey/building/furnished room/street scene/pattern points.
Emphasise movement, and links between stages. Make it visual, simple, exaggerated and appropriate.
e.g. "I went inside the cave and talked to the sleepy bear (Ranger) about the bearing to follow, and he said "Seek the morning sun", so coming back out I turned to the east until I met the keeper of the bridge who said "over my dead body lies the path to right; and left". So climbing up over him...
E.g. numbers — One Wooden Bridge, Two Red Rocks, Three Funny Tree… to reflect the hours of the day
Make up a story with rhyme, from e.g. 1, bun for lunch, 2, through the dew, 3, see the sea… 7 rhymes with leaden/Heaven/redden, 11 with Aladdin/pelican/caravan/a raven, 12 with elves/shelve/delve/selves
E.g. shapes — Make a 10-rock pyramid at the 10 o'clock stop; an 11 stick teepee at 11; a 12 leaf hexagon
— e.g. see-saw on a log at 7 and draw a see-saw in your notes—it has 7 corners.
E.g. You can weave a story, of invented images to go with the times of day.
e.g. each digit looks like a... 0 hole/ball/eye, 1 stick/pen, 2 cobra/sphinx/swan, 3 bosoms/camel, 4 yacht/picnic chair 5 measuring scoop/wheelchair, 6 bomb with a fuse/yo-yo, 7 arrow/step, 8 shotgun/glasses, 9 balloon/comma, 10 ball bouncing off a wall/golfball off its tee, 11 jail/stump/goalpost, 12 jail-break/caged lion.
E.g. acrostic — 7 a.m. Z. for Zero Rule Zero: Don't forget to pack the lunch!....7:15 A Bend A in the Creek
7:30 B Bridge across the River at Bend B. 7:45 C We didn't Cee the third bend on the map!
Using this method, elapsed hours happen at Z,D,H,L,P,T,X i.e. every four letters.

Use the Name of the Track for the Acrostic

L ying down tree
E ee! There's a bull!
E e-normous bracket fungus
S wim
P addled Across
A lternative Detour
D
D
O
C
K
S

*Every quarter of an hour, look for things beginning with that letter to remember that kilometre by.
*four letters make up one hour *Use stepped word pairs, 'Lying East-west; Exciting Exit; Extra Searches
*Do Something Alphabetical
E.g. Lie down at L; Scream at E; Examine the fungus at E; SPlash at SP
*Make all your names begin with that letter e.g. Pot-holes, Pines, Possum trees, Possible campsite, Photo
*Write them all down, but try to memorise the best one.
*You can attempt to improve the sequencing, by taking the letters two or more at a time e.g….
LE In the Lee of the lying down tree
EE EEE!
ES E.S.P. told us to find the track next to the E-normous bracket fungus
SP Splash Splash. Had a Swim
*Try also to relate picture to picture in sequence e.g….
Look at the Long horns on the Lying down tree; Enormous bull!; Enormous bracket fungus, dripping with water… Swim!

Acrostics combine individual memories with a clear sequence — firstly by letter, and secondly, by overlapping one pair of letters with the next, and thirdly by overlapping one picture with the next.
You have already memorised many long sequences of letters, e.g. the words of a song, which could be used as a unique acrostic.

Headlands, Bays, Prevailing Winds, Longshore Drift of Sand, Creeks

Headlands are ridges. Bays are valleys.
The ridges determine the headlands, bays, creeks, and — look for this — the position of offshore rocks and islands, especially where a 'point' marks a change in the trend of the coastline.

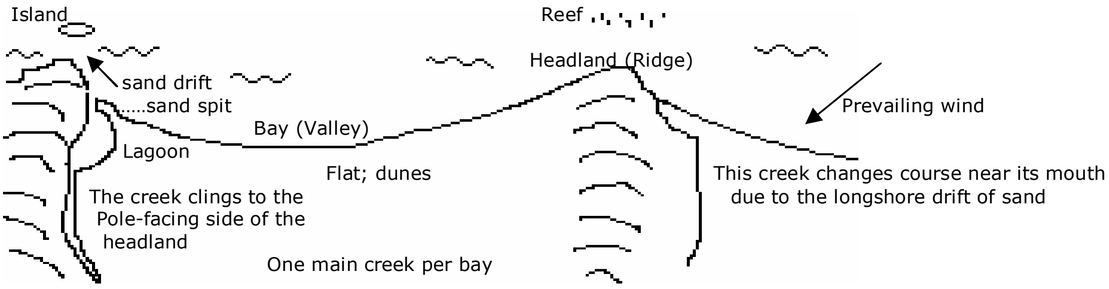
Island
Reef
Headland (Ridge)
sand drift
.....sand spit
Prevailing wind
Bay (Valley)
Lagoon
Flat; dunes
This creek changes course near its mouth due to the longshore drift of sand
The creek clings to the Pole-facing side of the headland
One main creek per bay

Beaches — Slope, Particle Size, and Wave Energy

Energetic erosion washes away any small particles.
A steep beach is a high energy beach with big particles.
When it is deep offshore the waves break with high energy.

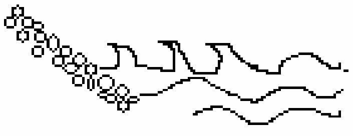

A flat beach reduces the energy of breaking waves.
A low energy beach is flat with small particles.
Sandbars and shallow sandy offshore seabeds make the waves break over a long-distance.

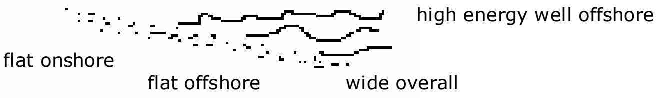
high energy well offshore
flat onshore
flat offshore
wide overall

Much the same seems to happen on land slopes — the flatter slopes have fewer rocks. I.e. the accumulated silt which is not washed away by surface water, covers up the larger rocks.

Sea-levels have been much higher in the past, so that there will almost always be buried boulders just above the present seashore.

Look Through the Trees, to the Sky and Skyline

This is a habit which stops you being befuddled by bends. You focus on the most distant thing you can see…

- E.g. the farthest tree trunk
- E.g. the skyline — ridges, gullies, peaks, light/dark division
- E.g. the clouds — but do stop to examine their movements occasionally!
- E.g. the Sun, the Moon — obviously the best, but they move too, and… don't look at the Sun!
- E.g. the dark blue polarisation band — Use Polaroid sunglasses if you are walking a forest track and can see only the sky.

Don't forget what is abreast of you e.g. a particularly tall tree. If you need to, look behind for some reference point.

Habits take practice to develop. This one is easy, because once you appreciate it, you won't want to give it up.
— It turns many small bends into a few major trends —

Periscope

Climb a tree. Climb a hill. Stand on somebody's shoulders. Or try this…
Attach a mirror (bigger = better) at 45° (try to be exact) to one end of a long stiff stick (e.g. a fishing rod)
— then scan the tree top horizon for landmarks.

Navigating by Trees

Tall trees on high ridges stand out against the sky and are visible for miles around.
Get to know the biggest, tallest, most visible ones.
Individual trees, gum trees especially, are wonderfully fingerprinted, by their branching pattern, outline, position & orientation.
E.g. look for the direction between the two largest branches. This stands out like a weathercock whenever you glimpse that tree again.
E.g. locate the relative positions of a handful of the best trees, so as to cover a wide area.
E.g. try to note the relative direction of pairs of trees.
Most big trees will have large branches to noon. So look for big branches which are *not* pointed to noon.

If you keep looking through the immediate green barrier for more distant (unmoving) skyline trees —
you will notice them and recognise them surprisingly often.

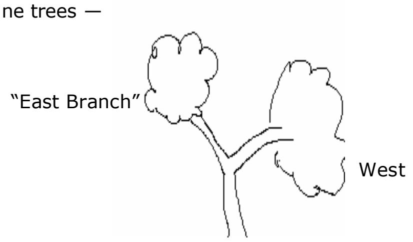
"East Branch"
West

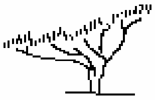

‘Solar-Panel’ Tilting of the Canopy Top

Some trees may be ‘flat-topped’, but they are not ‘level’. They tilt toward noon. Best seen from side on.

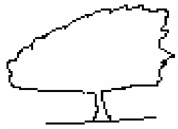

Wedge Shaped

Branches seem to know when to stop growing, and so manage to display
a fairly **uniform**, slanted **flat wall of leaves** toward the outside and especially toward the light.

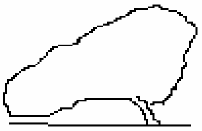

Lopsided Lean

Branches on the backside usually reach higher to bask in the skylight.
Branches on the front side are more horizontal, lower and longer.
Some trees don’t seem to try to compensate and simply leave their trunk behind, on one side.

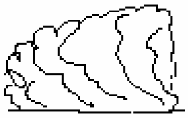

Half-Dome

— an extreme case of a steep back side

NOTE that all these signs work best in open woodland, not in forest understorey, but not always...

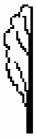

One-Sided Leafiness

E.g. Banksia trees
Such extreme asymmetry often only occurs in the understorey — where light is limited.
E.g. After a bushfire a gum tree revegetates from the trunk — most leafiness survives on the sunny side.

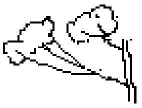

Individual Branchlets

— show similar shapes to whole trees
— tilted, asymmetrical twig patterns, one-sided leafiness, etc
The effect will show up best on the east and west sides of a plant, where the lighting is most asymmetrical.

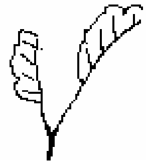

Bare Backbone

Look for it on individual branchlets, on the rear-side of main branches, and sometimes on whole trees.
You could run your finger up the entire length on one side of many shady-side branches.

Group Several Related Images into One

This is a standard method of making your memory ability efficient.
You have less items to remember—they become a few memory prompts, each to a handful of items.
E.g. 4 groups of 4… 6,7,8,9 a.m. / 10,11,12,1 / 2,3,4,5 / 6,7,8,9 p.m.
Write each group out visually shaped e.g. at the points of a square for a group of four
with 2 opposite squares for a group of eight
inside an infinity sign for a pair of items
List the hardest to remember items, first and last.
Group the shapes i.e. the groups, into another visual structure.
Use symmetry wherever possible—it makes it easy to recall.

The Start-Stop-Rest mnemonic (see Section 5. p.110) is a good example of this “clumping” technique—never more than a handful of things to remember, yet it leads you into hundreds of points.
Put the S for Sky at the top, and use a hexagon for the six start up points. Use a clock-face for TIME, and the compass-rose for NSEW.
Balance it all on a triangle, for the 6 a.m. start

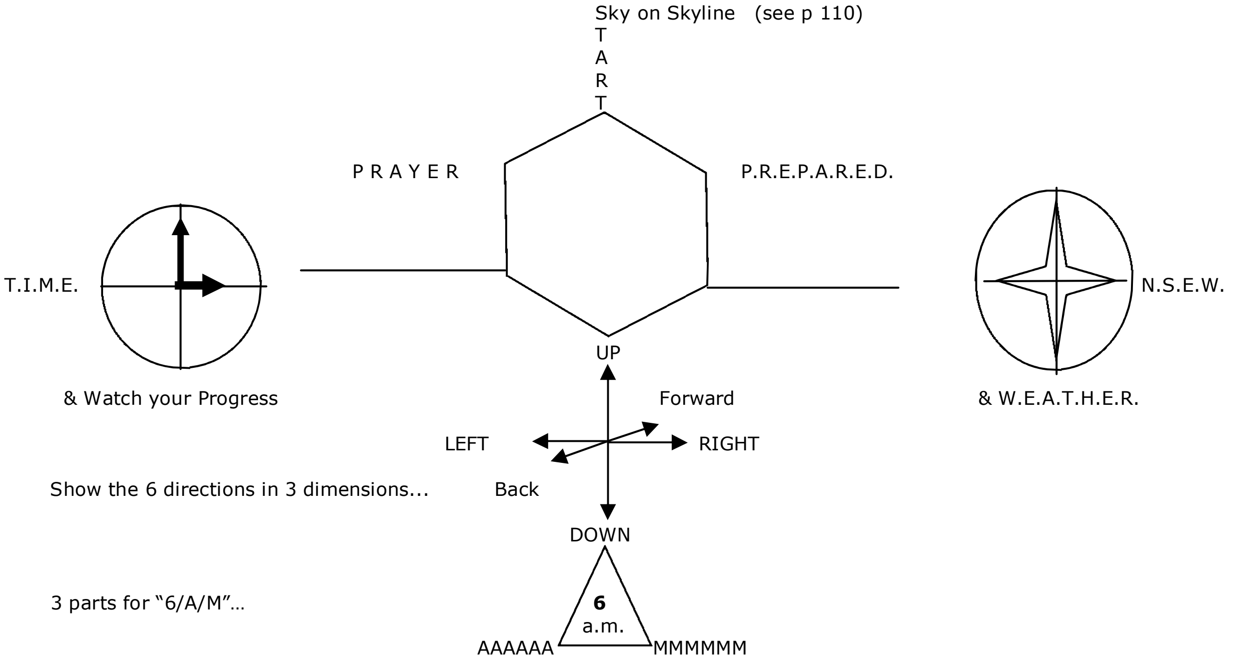

Rock Typing

Hardness — how easily eroded. Keep trying to scratch it until you find something just harder than it e.g. your pocket knife.
Colour — what colour silt. Scratch it against harder rock and look at the colour of the crumbs.
Sedimentary — the components are not of local origin. Layers — sedimentary origins. Melted sediments — crystalline bands
Sand content — grano-diorite — pale & silica rich. Clay content — basalt — dark & mineral rich. Mineral crystals in it.
Crystallisation — small crystals mean rapid (surface) cooling; large crystals means slow (deep earth—'plutonic') cooling.
Mineral content — what sort of soils does it produce? Flakiness. Behaviour in fire (Caution!)— it cracks, blisters or explodes! Melting pt.
Density, Magnetism, Fracture Pattern, Transparency, Permeability, Acid Solubility—evaporites and marine deposits may be limey.
Understanding these details is a key memory aid, since it gives you more meaning to look for, to memorise and recognise.
E.g. in basalt areas expect a more diverse & rich flora & fauna.
E.g. in the sandy plains, expect rivers to run underground, trees to have deep tap-roots and to line only the dry watercourses.
E.g. in dry sedimentary areas, look for a harder 'evaporite' layer underneath.
You don't have to identify the mineral, but you should characterise the rock type.

Soil Typing

1. Climate drives soil typing. What is the climate?
2. Latitude drives climate. What is the latitude?
3. High altitude mimics high latitude. What is the altitude?
4. How is the local climate modified? Look especially for temperature and rainfall factors.
   E.g. ranges upwind? A cold landmass? Warm ocean currents up-weather?
5. Evaporation produces salty limey evaporite deposits, whenever the evaporation rate exceeds the rainfall rate. High-temperature and wind, speed up evaporation. Flat inland areas receive less rain. Where rainfall exceeds evaporation, the soils are acidic, leached, and deep. Does the evaporation rate exceed the rainfall rate annually?
6. Climate drives vegetation as well as soil types. Soil types drive vegetation. The three go together. What is the vegetation type?
7. What is the parent bedrock? Granodiorite produces pale, sandy, friable, well-drained, mineral poor soils.
   Basalt produces dark, clay, dense, waterlogged, mineral rich soils.
   Warmth & rainfall aid chemical weathering and erosion rates of the parent bedrock, and so produce deep soils.
8. What are the relative constituents? Shake some soil and water together in a glass jar and leave it to stand until it is layered. Organics & charcoal float. Heavy minerals, pebbles & sand sink. Silts lie on top of them; then clays. Colloids remain suspended. Organic remnants accumulate in soil when it is warm enough to grow plants, but cold enough not to rot them away quickly. The decay of organic materials then provides acids for leaching, provided that high rainfall doesn't dilute the organic acids. Can you see from the sort of soils how cold and wet it has been?
9. What are the depths & layers of soil? It is leaching which transports poorly soluble salts to lower layers, until the water table dilutes the acidity. Subsoil clays also accumulate where leachates are deposited. Evaporites accumulate just below the surface and above the water table. Dry areas have poor soil depth due to slow bed-rock weathering and erosion rates.
10. What is the soil colour? Red indicates partly leached soil — iron and aluminium salts remain. Paler, ash-coloured soils are left behind after further acid leaching. Black soils result from poorly drained unleached soils. The importance is that leaching reduces fertility due to the lack of soluble plant foods. Leaching is aided by warmth (except in the case of limestone which dissolves better in colder climates).

Bush networks

The bush roadmap is made up of a network of interleaved ridges and rivers.
Instead of a major road or a minor road, you have a main ridge or river, shorter ones, and dead-end ones e.g. a 'shoulder'.
A 'cross roads' is where two creeks & two ridges meet — a saddle.
The seashore corresponds to a boundary fence; Peaks and islands act like townships — they are a focus of 'roads' and ring-routes.

Don't Navigate by Networks Only

Networks are fractal — big **creeks** have little creeks running into them; **walking tracks** have side tracks branching off from them;
big ridge lines have little **ridge lines** running off from them;
dirt roads have smaller dirt roads which have still smaller dirt roads joining in.
That is why it is difficult to recognise the scale of any map just by looking at the networks on it. The more detailed the map, the more detail there is to show up. The smaller scale the map, the more detail has to be left off.

So there will always be small creeks, unmarked roads, minor tracks and minor ridges to tempt you into thinking
"This one is what is marked on the map".
You will likely find one "exactly" where you look for it, but it may not be the right one.
Don't navigate simply by expectations of network junctions.
E.g. the next creek-line has a 50% chance of being "just as I expected it — a creek-line coming in from the right"
or if not expected, the 50% chance is that you will say "probably just not marked on the map"!
If by chance it is actually correct, it will only encourage you to continue navigating by network,
but you won't always pick the correct one.

Antidotes For Guessing at Network Junctions

- **Distance** covered. Accurate measurements of progress can eliminate some minor junctions which are "almost" where you expect the desired junction to be.
- Trend — of the junction branches. I.e. the **directions** must match your expectations too. E.g. "Each side of North"
- Landmark **bearings**. Try to separate identical looking junctions by cross-checking their position in relation to skyline features.
- Care.
- **Count**. Look-alike branches are nonetheless unique, when numbered.
- Recognition, from last time, if you had bothered to identify its peculiarities. I.e. bother now, for next time.

Branching Patterns — Examine Them for thickness, length, angle, curvature

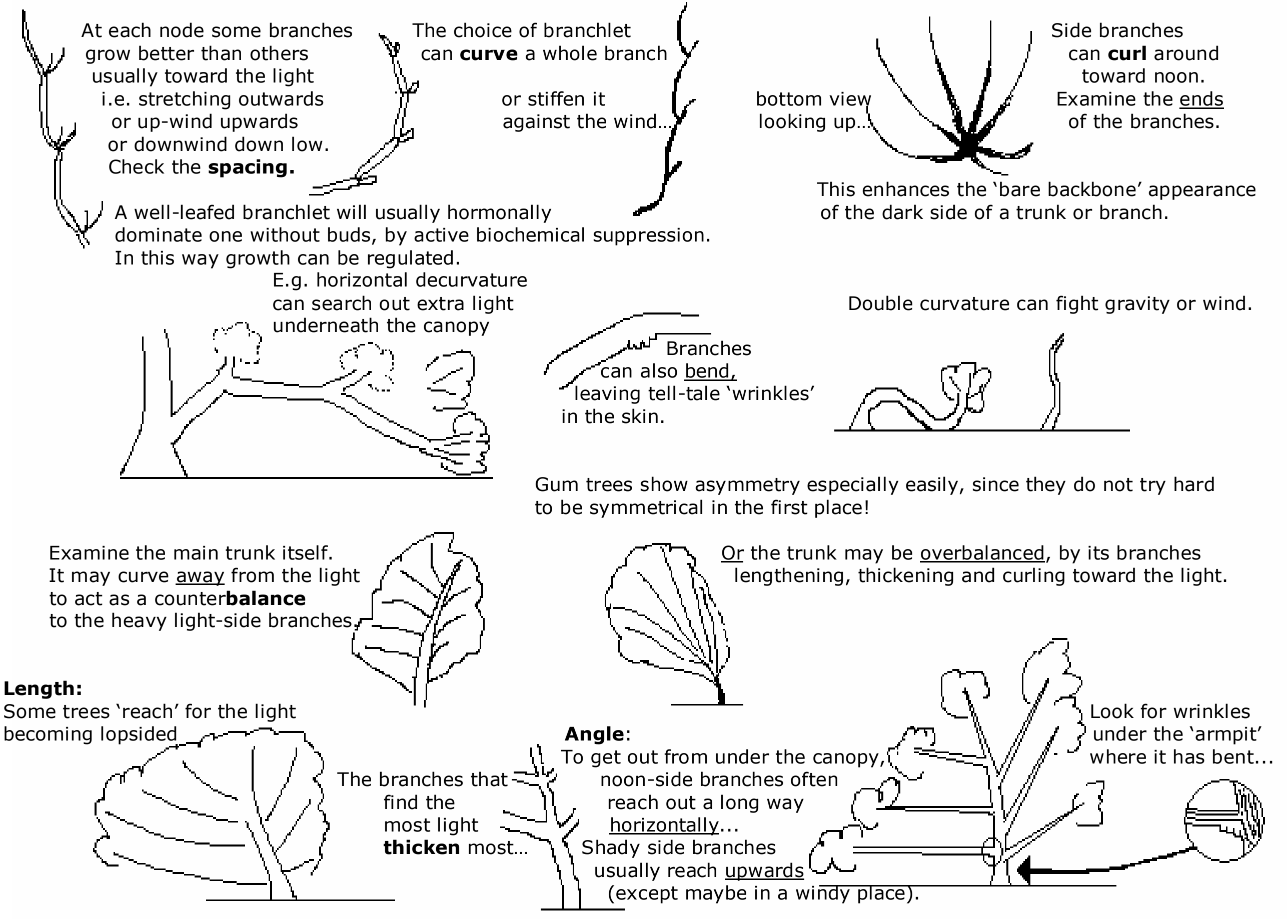

At each node some branches grow better than others usually toward the light i.e. stretching outwards or up-wind upwards or downwind down low. Check the **spacing.**

The choice of branchlet can **curve** a whole branch or stiffen it against the wind…

bottom view looking up…

Side branches can **curl** around toward noon. Examine the ends of the branches.

This enhances the 'bare backbone' appearance of the dark side of a trunk or branch.

A well-leafed branchlet will usually hormonally dominate one without buds, by active biochemical suppression. In this way growth can be regulated.

E.g. horizontal decurvature can search out extra light underneath the canopy

Branches can also bend, leaving tell-tale 'wrinkles' in the skin.

Double curvature can fight gravity or wind.

Gum trees show asymmetry especially easily, since they do not try hard to be symmetrical in the first place!

Examine the main trunk itself. It may curve away from the light to act as a counter**balance** to the heavy light-side branches.

Or the trunk may be overbalanced, by its branches lengthening, thickening and curling toward the light.

Length:
Some trees 'reach' for the light becoming lopsided

The branches that find the most light **thicken** most…

Angle:
To get out from under the canopy, noon-side branches often reach out a long way horizontally...
Shady side branches usually reach upwards (except maybe in a windy place).

Look for wrinkles under the 'armpit' where it has bent...

Working Names — Invent Them

E.g. Past the 'Little Castle Rock'
Up the 'Yellow-Crowned Tree' trail
Under the U-shaped rock face
Between the Big-Boulder Bluffs

"40 minutes in" is a handy name — when you recognise it, coming back out.

Stupid gawky names help you remember, but even better is to do something stupid to match.
E.g. Gallop like a galah, through the pass, and call it "Gallop Gap with Galahs Galore!"

Standard Name Plus Personal Name

Actual names last longer and are widely known. Use them.
But make up a personalised name too.

E.g. "Oh it is actually called 'Boomer Hill',	but we call it Porridge Hill, because it lies halfway between Oatlands and the Table Tier!
E.g. We call Wollongong "Kaus Australis"	'coz that's where we watched Kaus Australis go straight overhead.
E.g. Kaus Australis becomes	"The Wollongong Star"

Remembering Names

To help you later, when the name is 'on the tip of your tongue',
Count the syllables, consonants, short vowels, long vowels, now.
Rhyme it with something appropriate e.g. Condominium Creek has a pond of minimum seep.
Look at the spelt visual shape… for ascenders, descenders, double letters, symmetry.

Then later run through the alphabet for a guess at the first letter, then at the second, and so on.
Think about anything related, like why you went, what you ate, who you met, what you saw, the weather, the season, the news at the time…
Ask each person for these recollections. Recreate the full scene.

Contour Lines

Contour lines mark horizontal ground.
A simple trick to match up the map contour lines and the landscape, is to **point horizontally** parallel to the ground.
E.g. if you are in a valley, point along both creekbanks, using both hands. The steeper the valley sides, the narrower the contour angle.

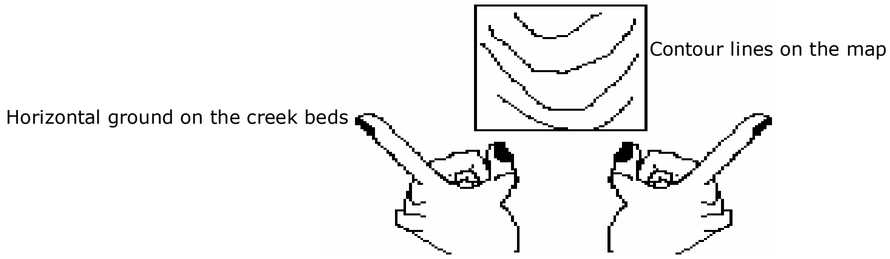
Contour lines on the map
Horizontal ground on the creek beds

Contour lines are like water-marks on the bank of a reservoir, lake or swimming hole, visible after the water-level has dropped.
But bigger like a global flood drying up 10m at a time, leaving water-marks.
like a photograph of it, from above.
like horizontal slices progressively taken off from the top of the mountains, as when you slice clay with a thin wire and you look at what is left.
like building up a model landscape one layer of cardboard at a time, from below, with progressively smaller shapes cut out according to the map contours.
Then you draw the results on a map.
They show the path you have to walk to go 'around the contour' so that you neither climb nor descend.

Contour Navigation: **Practise** It, **Before** you Need It

At night, or in fog, or in a blizzard, you may have no other clues, but the contours, on the map, and underfoot. Try to match them up.
Hint: Don't expect every little bump to show up on the map, but only the average smooth slope.
Hint: Keep good track of compass direction, time & distance as your best crosscheck. Assess the wind. Use an altimeter (,GPS, radar…).
Hint: **Delegate** tasks — map reading; course holding; distance logging; slope assessment; exploration; memory…
Hint: To go quicker, go slower — so as to be sure. Explore the slopes to left and to right. (But don't get lost or disoriented!)
Hold regular conferences to check your judgments.
Hint: Assess slopes by looking sideways, across them. They always seem so steep when you look up!

Boat navigators in a fog, can sail toward land until they reach a predetermined depth, then sail along parallel to the coast, using the depth contours marked on their chart, and some method of finding the depth. Marine depth sounders rely on a good echo, but you can always toss a weighted line over the side to find the actual depth. Make sure to make corrections for the depth of the depth-sounder, the depth of your keel, and the height of the current tide!

Sticking to the Bush Network

- Walking **up-ridge** may perch you on an isolated hillock.
- Walking **down-stream** must get you to salt water; or at least to water, civilisation, flats, roads; but also to thicker vegetation and a loss of view; and sometimes to impassable waterfalls and gorges.
- Walking upridge, uphill, upstream lands you in steeper, drier, more jagged terrain with more open vegetation, a better view and maybe an along-ridge track.
- Walking down-ridge will "take you to the lake", but simply walking downhill may lead you into a heavily vegetated creek-valley which then drains into the lake.

Natural Routes Naturally Converge

That means that, in reverse, they must also **diverge** — which is dangerous, open to mis-navigation.

- Provided you don't cross catchment divides, walking downstream or up-ridge is **convergent** and so is safe for a one-way trip, but **un-safe as a means of retracing your steps back to base**. E.g. it is easy to climb a hill, hard to retrace exactly.
- Walking down-ridge is **divergent — deceptive and dangerous**.
  You cannot easily see ridges peeling off below, and the map won't mark all the creeks you can see.
  If you are trying to retrace your steps or navigate from the map, you will have to be very careful.
- Ridges converge to hilltops — so walking **downridge** leads you widely astray.
- **Upstream** is divergent — you don't know where you will end up, but at least you can find your way back, because creeks converge toward the river mouth.
- Tracks & trails converge to: popular features, bridges, car parks, waterholes, homesteads, campsites, clearings, etc.
  So **leaving a focal point faces you with multiple choices** "How on earth do I get outta here!"

Coping with Divergence

The result of "directions" such as "You can't go wrong. You can't miss it" is that the truer that is, the easier it is to go wrong — after you have found it. Retracing your steps, or finding the desired way out, onward, can become a nightmare.

- Ask directions for the onward bit, not just how to get there.

For backtracking…

- Pay attention on the way in, despite your excitement at arriving. Note the direction of your arrival. Look back regularly. Mark the junctions you used, especially from minor out onto major paths. Retrace your own tracks.

Reverse logic

Be like yachtsmen, who should always explore harbours, shallows and tributaries at low tide "in case we do get stuck".

- Explore upstream — a divergent labyrinth — so that you can walk downstream to return.
  Provided you don't cross water-divides, or overshoot your starting point, "you can hardly go wrong".
  Ahah! There's the hint: Before you set off upstream, peek behind you, to see when to stop "coming back downstream"!
- Explore downridge, and climb back upridge later. Only don't climb down anything you can't climb back up.
- Explore away from a focus, so that you can converge back to it.

Isolated Vegetation Shows the Clearest Asymmetry

E.g. Look in any **clearing** for downwind debris/erosion,
and for up-Sun/down-Sun asymmetry at the edges.
E.g. an isolated tree in a paddock...
E.g. an open creek will give the bushes a clear view across the water...
E.g. a roadway will exaggerate the light asymmetry on one side
and suppress it on the other...
E.g. Look in the wheel ruts of bush tracks for asymmetries due to wind & light…

Examine **the tallest trees — The emergent ones** have a clear view of the Sun.

At the **edge** of a clearing, subtract the effect of wind…

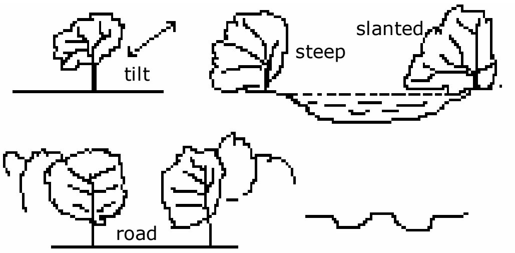

Closed Canopy — Look for Gaps

Each tree canopy might look uniform, but be blown back on the upwind side, leaving a gap there, with dead branches, where it *has tried* to fill the gap. But look for a gap without dead branches and without other explanation — It might be 'shady south', which gives the tree little incentive to cover. The tree *could have* spread that far by now, if it had wanted to, but hasn't.

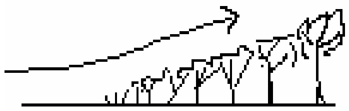

Tall and quick growing saplings **show the most asymmetry…**

They can be the best tell-tales of all, and the worst!
Because they respond to clearings, slope, water-shine, nearby trees — be careful!
Take extra care to average out their message.
Any one tree gives you too definite an impression!
Back-to-back saplings may show opposite asymmetry, by shading each other...
Treat the overall shape as one tree.
The sapling effect, arises from young forest trees shedding branches more often and more easily than mature trees.

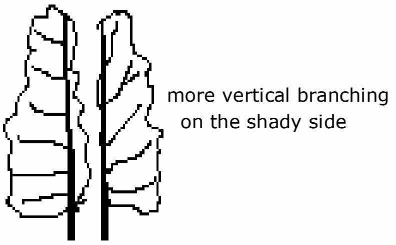

Roadsides & Campsites Reverse Some Effects!

Examine trunks for broken or lopped boughs before interpreting!
The bigger lower branches are easiest and best firewood pickings!

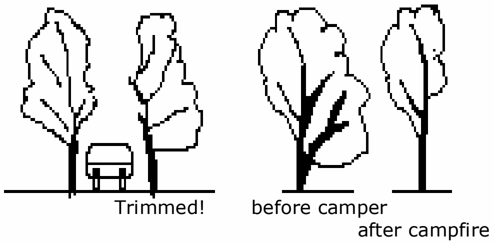

Name the Subsections

Heart-Attack Hill   Bust-me-gall Bend   Starve-gut-Ridge   : Be Vivid
On-the-Double Decline   Cross-the-Creeks Roller-Coaster   Blackberry Bank   : Be Descriptive
Left-hand Loop   Clockwise Climb   Right-around Contour   The Zed Bend   : Be Cheiral
N-E Straight   Into the Sun Slog   West Bank   Downwind Dunes   : Be Directional
Long Leg   Short Stretch   The Middle Mile   : Tell the Distance
Last-legs Leg   Dinnertime Creek   Sunset Strip   : Tell the Time
On the Left of Third Hill   Nine Mile Beach   Five Ways   Fourth Island   : Count
The Bush Blocks   The Rock Band   Turning Point   The Up-Part   The Trip down   : Be Clever
Rubble Ramble   Possum Passage   Flower Flats   : Alliterate
The Birdwatching Bit   The Deer Park   The Gloriously Green Glade   : Be Observant
The Monkey Saddle (three creeks)   The Big Bulge   Round Shoulder   Tight Turn Tongue   : Be Shapely
The Breakaway   The Hither Side of the Hill   Round the Rim   Unto the Brink   : Be Literary
Brown Beach Bay   Wobbly Shacks (= Mount Wobbly behind the shacks)   Zigzag Track   : Be Visual
Perverse Reverse   Inland Cusp   Round the Back   Out to Sea and Back   : Name Whole Curves

Name the Sequence

Fit Each Subsection into a Larger Scheme.

- Name the Links : The Cuvier Track   The Mount Barker Road   The East West Ridge
- Name the Junctions : Point Pass   Steppes Junction   Tee -tree Junction
- Contextualise : Beyond Burrowing Crayfish Ridge   Cross Catchment Climb
- Overlap Sections : Deer Sally to Sandy's Shore to All-Terrain Trek } these are simultaneous
  Sherwood Forest to Bumpy Flat to Uphill Haul } and overlapped 50%
  I.e. halfway through one section, think about how it is changing character
  into the next obvious section, and anticipate it.
- Time-Distance Clues : Lunchtime Lookout   The Approach   Thither Heather   Tea Junction   Ten-Mile Dam
- Lie-Linkage : Downstream from Dusty Waterfall   Wind-Shadow Plateau
- Numbers : Fourth Bump   Fifth Creek   Tee Junction # Three
- Acrostic : Short Steep Step in the Southern Subdivision after the Shallow Segment

Contour Curves: Your Job is to Imagine the Shapes into Life And to Match the Lines to the Landscape

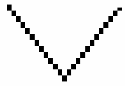

Deep Vee 'Ditches'
— they gather water.

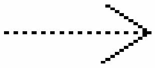

The **sharp** points, point **up**-river. (sharp/up rhymes in p)

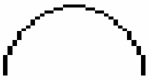

Round 'Shoulders' — they shed water.

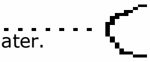

The **round** points point **down**-ridge. (round/down rhymes)

— being higher ground, they separate gullies

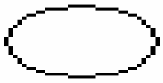

Peak or Depression — you can't easily tell which. Peaks are much more common.

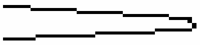

Hairpin 'Ridges' — they can look parallel, like 'a slope', if the edge of the map cuts them off.
Ridges usually show much rounder than this.

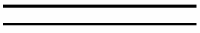

Parallel 'Slopes' —be careful about which way is 'up'. Check that they are not the same height—on a ridge or river

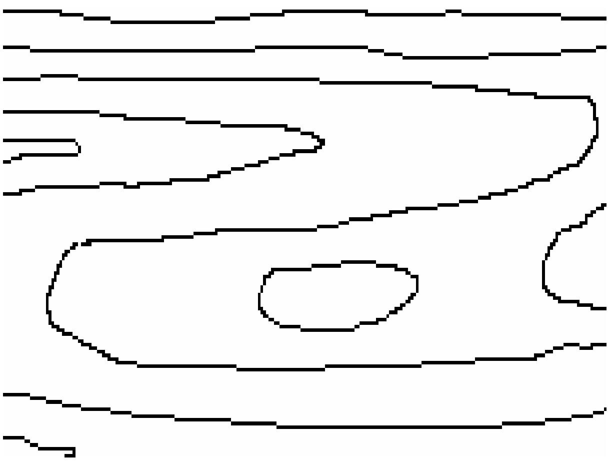

Here is a typical confusing array of parallel lines.
The first two are ambiguous, but seem to lead down into a valley…
The valley is shown as lines 3, 4 & 5…

The three corners here are all sharper than the corresponding ridges.
Downstream is to the left…
Imagine a deep-vee ditch.

The shoulders and peaks shown here are all much rounder…
..Uphill is ⟶ peak ⟵ saddle ⟶ ridge or peak ……….
Imagine a circular mound with a spur down to left and right.

Down here, upslope to higher ground is…↑

Trackless Navigation

When you leave the bush network to cut across the natural routes, you still navigate by them, by noticing when you cross them.
You choose a route in relation to the framework — of ridges, rivers, and the contours that cross them at right angles.
E.g. "I am now going diagonally right, uphill, on the left bank, to the Ridge, then…"
You describe the sort of route you are following, and **start a new subsection whenever that sort of route changes** in quality.
E.g. A dry shallow creekbed marks a change from downhill to uphill.
E.g. The major break of slope—on the "hillside"—marks the change between being on the "hill/crest/ridge" or in the "valley".
Don't forget to ask, as you cross each 'road', what that ridge or creek is doing, to the left & to the right e.g. joining, splitting, rising, falling, dividing major valleys.

The method of Names

The point of this method is to bring your normally subconscious memory into conscious focus.
You put a **name** i.e. an essential description, onto each subsection. This is in addition to naming the creeks and ridges you cross.
That is equivalent to supplying your own signposts & flashing lights to the landscape. But write down the other details also.
As with memorising any journey there are a handful of steps: Notice; Recognise; Describe; Name; Remember; Put it into Sequence.

Bush Networking Essentials

Answer up to six backbone questions about each leg, before deciding on an appropriate name for that subsection…

1. Hill-wise — What Peak? — Is your direction trend in relation to it Uphill or Downhill; towards or away?
2. Hillside-wise — Slope — Choose between Upslope/Downslope/Flat/Contouring.
3. Ridge-wise — Rise — Upridge/Downridge/Crossridge?
4. Drainage-wise — Lie — Upstream/Downstream/Crossstream?
5. Side-wise — using the standard names* — Left bank/Right bank (below the break of slope); Right flank/Left flank (above).
   *As a yachtsman going **up-**stream sees it — Port and Starboard. Use these terms in your names to remind you.
   As a traveler going **up**-ridge sees it.
   I.e. the standard directions are **up**-stream; **up**-ridge; **up**-north; **up**-wind; **up**-hill, **up**-slope.
   Be careful! The 'right bank' is the 'left flank'. You can climb to the right on the left flank (of a ridge which falls to your right).
   Walking diagonally 'downslope' can be 'upstream'.
6. Hand-wise — Tack — Clockwise around or Left around/On the left diagonal or Right diagonal?

Mnemonic: (With your right index finger) touch your **head** (the peak);
forehead (the slope);
bridge of your nose (the break of slope);
down the ridge of your **nose** (the ridge);
nostrils (the drainage);
right **cheek** (the side);
and now look at your **hand** in wonderment (What is my tack?)
Then **point** out some peculiarities...
Now **add the non-essential** — exact measurements e.g. direction/length/time/steepness/count —and descriptions, sequence, context.
(Remember: don't navigate only by networks.) See opposite, for descriptive names.

Flowers Respond to Sunlight

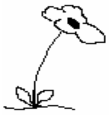

Examine the stalk-lean and flower-head tilt
e.g. of daisies.
Look up-Sun and down-Sun to see the striking difference, whenever the Sun is low.

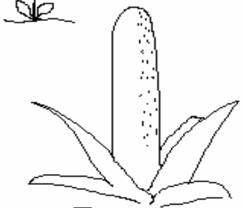

Look for the first flush of flowering on the sunny side of a flower spike
E.g. Yakka (Xanthorrhoea). The mid-morning sunlight direction gets the earliest flowers.

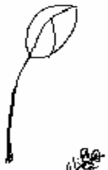

Examine closed flowers to see whether they betray where the Sun last came from.
Many flowers 'follow' the Sun around.

Some leafy stems seem to tilt to the noon, perhaps to point their terminal flowers into the Sun.
E.g. Sea Spurge

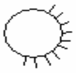

(Top view)
Banksia pistils sprout earlier on the sunny side.

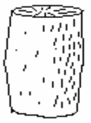

Banksia stamens deepen in colour on the sunny side.

Fruit Distribution and Ripening

Some bushes will fertilise more flowers in the sunlight because pollinators can see them better.

Fruit are fertilised flowers. So some species 'set' more fruit on the sunny side.
Look in all directions to see which aspect presents the most flowers or fruits to view.

'Ripe' colours are a result of sunlight e.g. apple, peach.
Sun-ripened colours aren't confined to fruits.
Mushrooms are often old enough to have seen some sunlight.
One mushroom in the open may not help you much, but they do come in flushes, so look for a *pattern*.

Remember the Sequence from AA to ZZ

— A Ready Made Acrostic of memorable unique items, with endless extension if needed. I recommend it as the pick of the mnemonics.

You may need to remember each tree in a forest walk, or every twist in a path, or every choice in a maze, or every square in a grid.
AA to ZZ supplies 676 *differently named and sequenced letter pairs*.

If you turn the letter pairs into appropriate *two word phrases*, there is no limit to the useful memory aids.
E.g. I could call this system the Alphabetic Zygote (of 'letters paired' A-Z).
As with all gimmicks, it only works best if you *study and* ***memorise*** **the object** ***receiving*** **the label**.
Examine the tree for anything which could attract the label DF (damp foliage; dendritic form; downpointing flowers),
or mentally *photograph* the MS munch-stop, or *actually* bird-watch at BW.

It's up to you to use the letter pair in a memorable way. Don't forget to use visual, emotional, sequenced and appropriate words.
Consistent rules help.
On a pure sequence walk, begin with double letters, rather than with A-Z single letters. This makes two word phrases the consistent rule.
In fact begin with *any* double letter, but especially **the initials of the walk**, e.g. BL for Blackmans Lagoon.
Two word Phrases e.g. AA = Ask Again, avoid the ambiguity about which other letter in the word or phrase might be emphasised.
E.g. is Aeroplane, AE or AR or AP or AN? Was 'Anabolic steroids' A or AS?
Add a third word to the phrase as needed, to spell out what it means, e.g. Backwater Island turnoff.

The trick is to make the mnemonic **match** the feature.
E.g. if there is a plague of mosquitoes which you attack with a stick, Aaron's Almond-tree walking stick might be appropriate for AA.
Then the need is to somehow link the previous with the next.
E.g. AB Aaron's Budded walking stick (if you know that story), might focus your attention on the buds in the next gum-tree.
Without a '**theme**' it is difficult to pluck out your memory what 'DF' might have been. But 'Doctor's Emergency' where you stubbed your toe, leads on to 'Doctor's Ferari' when you walk on quickly anyway, and on to 'Doctor's Gasoline' where you topped up with lunch.

If you do a 'grid-line walk', you use A-Z to begin with, and say you move from A0 to C4; you might go through A1, A2, B2, B3, C3, C4.
The memory aid goes A0, then Ahead on 1 leg, hopping, then Across, on 2 legs, Ban-tu, Ban-three, See the three cattle?, Count to 4...
'Minus', in cases of ambiguity, might come out as 'which was 4 metres deep', or 'where I left 4 pebbles behind', or 'and Not 4 blisters' or 'at anything but 4pm'.
When you get into AD you look for a-d **peculiarities** to remember like 'Add 4+6 wattles (to get 46)' or '33AD crucifixion tree' or 'Adder-Death, two score of them'.
As an alternative, name all grid squares with *letters*, e.g. D-G, and look for things to name, such as Dark Green; Discarded Gaiter.

In a maze, you could choose Long words for turning Right; Short words for turning Left, a mixed phrase for Straight Ahead. These match the meaning-words in length of sound.
Maybe you need some **variations**:
One, two, three or four syllables for left, right, through, reverse.
Long first vowel sound vs short vowel sound.
Begin with one *letter* and *end* with the other *sound*, e.g. Amanda, Abib, Ace, Add, Abbey…
Let sounds take priority where needed. E.g. few words start with K, but many start with a hard C. X is hard unless you allow Ex-.

Memorised sequences are most effective if revised almost on the spot, e.g. via a leap-frog walk (2 forward, one back), and on the return.
Equally, other methods are enhanced by adding this mnemonic, such as when marking a trail, writing track notes, or making a mudmap.

Contour Heights: Read the Numbers!

Determine **the Contour Interval** — and interpret which way is 'up' from the **numbers**… Note: On maps, the numbers are printed to be read as if looking **uphill.**

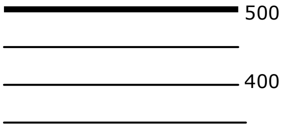

This first diagram is "every 50 m".
And 'up' is toward the top.

Zero means "sea level"
Read the absolute numbers!

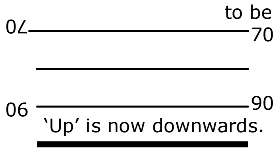

In the second diagram, the contour interval is every '10 m'

'Up' is now downwards.

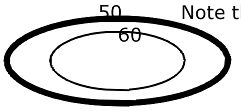

Note that **major** contours may be marked as thicker
e.g. every 50 or 100 m

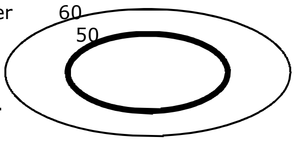

..The first is a peak; the second is a hollow...

Marine charts show spot-depths, 'soundings', in metres and decimeters, below lowest water, with offshore depth-contours.
If the "depth" is underlined, it refers to a "drying height" of intertidal rocks.
At sea, keep a record of the soundings as you sail in to land, preferably in a straight line. Plot them onto tracing paper or mark them on the edge of a ruler, and try to match the pattern to charted depths.

Contour Spacing: Close = Steep Slope

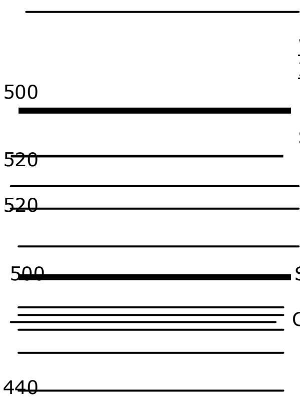

Wide spacing = flat land   Examples: mountain plateaus; valley floors; paddocks; dunes
Irregular contour shapes (meandering all over the map; not shown here) = very flat land

Slope

Ridge — carefully notice when numbers are identical.

The rule "Closer Equals Steeper" only holds for different height levels, not for hairpin contours around ridges or rivers.
In fact, "close" might equal "flat" i.e. on a ridgetop

Slope

Cliff — Very close = Very steep

Slope; have you worked out downhill?

Walking "Straight" through the Bush

- 'Straight' downstream; upstream; down-ridge; up-ridge; uphill; downhill
- 'Straight' around the contour
- 'Straight' along the path of least resistance
- 'Straight' through the obstacle course, dodging to left and right
- The Line 'of Best fit' — to a chosen trend cross-country

None of these are straight, but most are easier than 'straight across a slope' — crossing innumerable ups and downs!
Most are predictable.
The navigator simply goes last in line with a compass and logs the twists and turns, whatever course that takes.

Why last in line? It is a fact of life that following your nose through the bush quickly disorients the leader of the party, while those following behind maintain their bearings easily.
Also, to clearly show up the trend of a sinuous course, the last person can drag a long smooth string loosely behind him.

Straight around

At a…

- Constant Angle—to the (changing) Sun, Moon, star, planet, landmark e.g. keep the peak "ahead on your left" — a spiral course.
- Constant Distance Off — as judged by the rule of thumb — you keep a prominent landmark appearing to be a constant size — from base to peak. The course is circular, concentric to the landmark. E.g. offshore, around a lighthouse.
- Constant Horizontal Angle — between two prominent features. This gives a cusp-like circular course.

You can retrace these courses, but the going may well be difficult — crossing many streams and ridges.

How to Walk in an Unintentional Circle

Haven't you heard of people walking in a large circle when they are lost?
They thought that they were walking in a straight line, but some hours later they recognise the same scenery!
I've done it; it's no myth. Discover your natural drift...

Blindfold yourself. Try to walk straight across a flat. I bet you can't.
Most people have one slightly shorter leg, usually their left leg.

Now try to walk across a slope, blindfolded of course, in a straight line. I bet you drift.
But once you realise it, you are just as likely to head uphill too much, in overcompensation.

Don't trust your senses in a fog. Find out by trials at home — before you find out the hard way — how far 'off' you are.
Also find out how to properly walk in a straight line in a fog!

Stem and Leaf Colouring

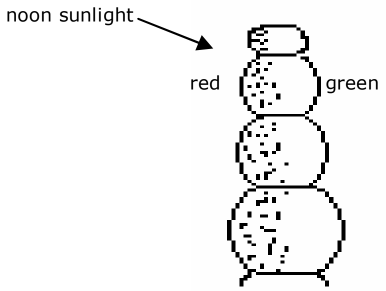

Samphire saltbushes
colour up
on the bright side

Many other plants
show red stems
or leaves etc
Look for a paler side.

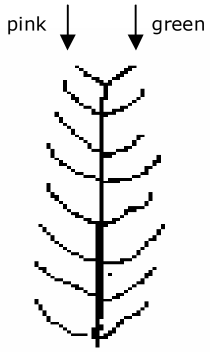

Look from the top
and sway side to side
to find the deeper colour

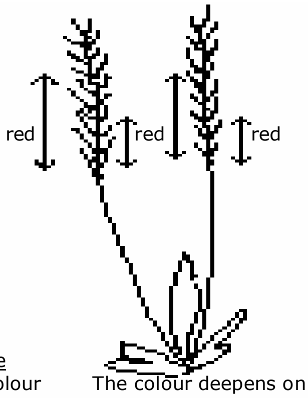

The colour deepens on
the bracts higher up
on the sunny side

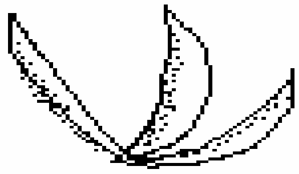

Succulent leaves redden up
on the bright side. Look one
way over the flats & it is
pinker than the other way,
especially just after sunset!

N.B. Older stems may show a reverse trend! The colour fades on the sunny side after having deepened more while young.

Beware: When flower-stalks twist as they grow, the colouring or fading effect is unreliable. You can detect the twisting *by* the fading.

Example: Blackberry stems and leaves redden in sunlight, staying green on the shady side.

New Growth is Colour Differentiated

Looking up- or down-Sun (on a cloudy day) will show up new growth as yellowish or brownish or bluish (depending on the species) in one direction more than the other.
The sunny side will 'green up' the new leaves more quickly and will look more normal coloured.
But test it out first — it may go the other way — sunlight might promote new growth, before it colours it up.
When new growth is distributed throughout the shrub, the sunny side will often show deeper green (i.e. more of the same pigment) and be more vari-coloured e.g. some leaves take on a red blush or red margins in response to the Sun.

Revise your Memories

..after a 10 minute distraction
..after an hour, a day, a week, a month, a year e.g. for an acrostic which you wish to remember permanently
..and whenever after, when you begin to have trouble remembering
Choose undistracted, mindless times
e.g. while trudging or being driven, waiting or relaxing, or last thing before sleep, to run over things in your head.
Overpractise your mistakes. Out loud is always better. Take any opportunity to tell someone.

To Remember Formulae such as those at p. 19b

(1) Make up memorable ditties, and refresh your memory, as above, before you lose them.
(2) 'Read' the formulae *for sense* — They are just shorthand for how things *must* behave. If you don't like maths, it's probably because you never understood the common sense behind it, and that is probably because you never had the opportunity to see it, or developed the skill to read formulae.
E.g. '**Sin, Sin, Sin! It's High Latitudes which Declines it**, for both Easterners and Westerners'
for: the three sine ratios in the formula… 'sin (**heigh**t E-W) sin (**latitude**) = sin (**declin**ation)'
i.e. as the latitude increases, the height must decrease, and vice-versa, to preserve any sin (dec) figure as constant.
The formula should return zero height for zero declination, and 90° height for dec=lat.

This next one assumes that you measure the angle away from East-West (rather than from the 'elevated pole', as on p 19b).
You can use 'sine of the variation from E-W' instead of 'cosine of the angle from the pole', because 'co-sine' says 'sine of the co-angle'.
'The sin of the increasing the sunset angle, by cause of increasing latitude, is a sign of decline'.
sin (sunset angle) (multiplied by) cos (lat) = sin (dec)
Read this as: For any specific declination, as lat increases (=the cos reduces), the angle must increase, to preserve a constant sin (dec).
You can also put in dec=0 and expect a 0° answer no matter what the lat.
At the Equator, declination directly determines the direction of set N or S of E-W.

'Why bother with the E-W Sun? Be**cause of its E-W *timing*!** It's **today or too late**' t for tan; d for dec; late for lat;
90° = 6hrs each side of noon cos (time of E-W) = tan D / tan L, *'or' for "o'er" for 'over'*
Lat = 0 throws the calculation into a spin, because any declination other than 0 can't be E or W. Dec = 0 must return 90° for any lat.
Dec = 90 must be impossible (it's N or S!), but so must any dec > lat, since cosine doesn't go above 1. Those stars stay polewards.

'One New Day at a Time? Not quite, **'cause for tomorrow's sunrise time, today's time's not too late**!'
It changes only slowly, as dec changes. cos (time of sunrise) = tan D x —tan L 15° = 1 hr. each side of noon
0,0 must return 90°. Opposite hemispheres must return a large positive cos and hence a smaller angle, and a short day.
Same hemispheres must return a negative cos, which means a long day.
Too much lat, and the tropical bodies disappear altogether, as expected. Too much dec, like a circumpolar star, and it won't even set.

Try **the Equation of Time**, p.17b.
The First FFive months, From FFebruary, Fence-sit Four Features (halF-each side). The next 7 *finish* at four more features.
–1**4**,0,**4**, 0, -**6**,0,1**6**, 0 Makes nice mirror patterns of the two sets of four features. Minus means a fast Sun, up early; a 'slow' clock.
FFFebruary Founders at FFFourteen; Marches up again; April crosses (its hot-cross buns); May turns 4; June double-crosses;
July *finishes* work at six; August rises augustly, to zero; SSSeptember and October get up to SSSixteen;
November and **Dec**ember **dec**line a lot; December crosses at Cross-mass; January Sun decends from the cross, fast, for a new year.

Contour Slopes: Work Out 1 mm and 1 cm Contour Spacing

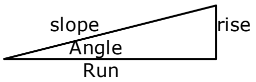

Example: a Rise of 10 m in a Run of 100 m is a Slope of one in 10 and an angle of 5.7°
[Tangent(angle)=rise/run   Angle=Arctan(rise/run)]
rise/run is technically called the "gradient" i.e. an up-slope.

1) Example of 1 mm contour spacing: A 1:100,000 map-scale says: 1mm equals 100m (See 'map scaling factors' p.92a)
with 50m contours the slope must be 50m every 100m or 1 in 2 or 26°½.
2) Example, with 1 cm contour spacing: 10 times less slope, in the same example, or one in 20, or 3°

Using these 2 reference spacings (**work them out again for every new map**) fit the actual contour spacing (eg 3mm) between them.
E.g. map scale: 1:25,000 (1mm = 25m); for 10m contours, 10m rise in the 1mm run, on the map (=25m in reality) is,
"1mm contour spacing = 1 in 2.5"; (10mm contour spacing) = 1 in 25; (20mm contour spacing) = 1 in 50;
So… 3mm means 1 in 7.5

Reference Slopes — To Imagine the Climb with!

One in one; 45°; One up / one across; On all fours!
One Degree is 'one in 57' (or approximately 1 in 60) So 10° is 'one in 5.7' — maximum for public **roads**
One in Four is 14° — a neat coincidence to memorise. So One in 40 is 1.4° — maximum for **railways**
36° slope — 8 in 11 — more or less maximum for sand, gravel, scree, or any loose material.
Small slopes may be multiplied with approximate linear accuracy, but in an **inverse** fashion…
Examples: 2° is one in 30 — from the one in 60 rule
1 in eight is 7° — from the one in four rule
5° is one in 11 — either from 57/5 or from 5.7 x 2 or 1/8 of 36° see also "sine-cos chart" p.106b

The Beaufort Wind Scale — Modified for Bushwalkers

One in 40	contour lines absent or rare; pleasant ambling; wind whistles through the lips; grey hair relatively common
One in 30	car engine quieter than road noise; perambulators plentiful; locomotives absent.
One in 20	car engine noise louder than road noise; walkers plentiful
One in 15	skin moist despite cold conditions; heavy breathing is heard above background noise
One in 10	eyes sting from sweat; bicycles experience difficulty staying upright
One in eight	flies are constant companions; car smoke exits parallel to the ground; mountain bikers walk
One in six	cigarette butts become less frequent; 2 wheel drives mostly absent
One in five	walkers experience difficulty making forward progress
One in four	cannot hear the sound of a cigarette being lit; leaf litter easily lifted off ground
One in three	sturdy walkers stagger and sway; frequent rests; slow progress; loose soil swept away
One in two	large trees tilt at a noticeable angle to the ground; rocks experience difficulty staying in place; walkers remain upright only by zigzagging
One in one	toenails begin to bleed; trail bikes begin to bleed
Two in one	fingernails begin to bleed; small trees are pulled up by their roots
Three in one	contour lines begin to merge; danger of severe falls; you are in danger of being completely… winded!

Walking an Exact Straight Line in Open Country

Assume the visibility is limited, or the terrain difficult, or the vegetation is closed in…

- Lengthen the Whole Party—In Single File. Shout to the leader, if he wanders to left or right. Stop still if you lose sight of the person behind you, i.e. do something to make sure that no one gets lost in the fog.
- Drag a flat object (or a person) on a long string. Tie your end of the string onto a stick, and hold it forward above your head. Alternatively, pivot the stick in your hand at its halfway point, and it will point in the right direction. (Watch that the wind doesn't blow it sideways; Stop the flat object rolling; Stop it sliding downhill; Don't let it drag at an angle)
Try just a slippery fishing line with no object.

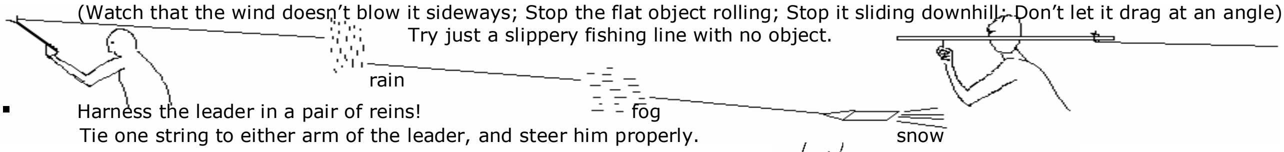

- Harness the leader in a pair of reins!
Tie one string to either arm of the leader, and steer him properly.
- Hold the rear end of the string over a compass.

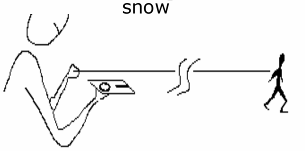

- Leave a trail of visible markers lined up exactly.
You need to look back frequently and drop a new marker.
— stones, sticks, stumps, tree trunks
Leapfrog with only three marker sticks — take them with you.
Leapfrog people!—leave someone behind; send someone ahead; line them up with each other & with a compass if available.
- Scratch a line into the dirt. Look back frequently. Readjust yourself
- Use the radio for radio direction finding.
- Orient by Sunset glow, the wind on your cheeks, the blue band, town lights, distant noise, etc.

To keep to a fixed "track", manoeuvre around obstacles in a rectangular fashion,

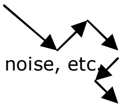

or reverse any diversion.

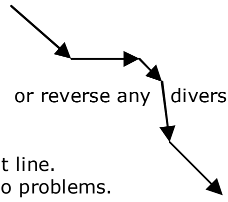

Leap-frog a Straight Line by Looking Ahead

A. A straight line course is easily achieved visually — like keeping a hundred fence posts in a dead straight line.
In clear open country, you simply walk toward a new mark halfway between you and your destination — no problems.
Keep the 2 lined up. When you get halfway choose another halfway point in line. If you can keep sight of the more distant destination, this method means you only use the magnetic compass once, to locate the distant destination—then keep a fixed 'track' visually.
When obstacles force you sideways off track, you can easily regain track, visually.
B. In more difficult circumstances, you cannot see far ahead and usually rely on looking back, as above.
But leaving markers behind you slows you down and keeps you looking around.
This method does not require you to look back, but requires a simple discipline — leapfrogging visible marks exactly ahead of you.
Step one: Choose two distinct marks exactly in line with your desired direction
Step two: **On reaching the first mark, choose a third, still in line. And so on.**
You can do this while following a formed track. The track will bend from the average straight line, but you'll notice that immediately.
I.e. when you don't get to the third mark, and have to start again, you are well aware of the change in trend.

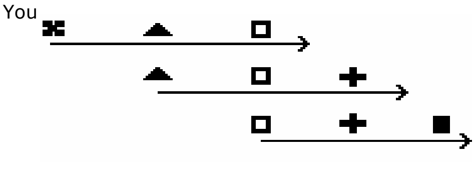

The end result accumulates errors
(despite the averaging process).

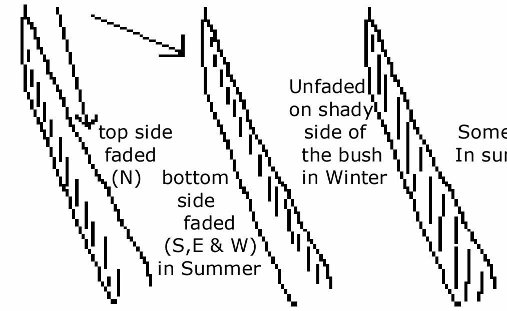

Noon-Sunlight is More Vertical

Examine needle bush (e.g. Hakea) needles, one by one.

Some effects will be more noticeable in winter, and in higher latitudes.
In summer, the 'shady' side is actually bathed in horizontal sunlight twice a day!

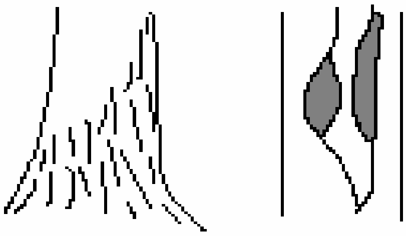

The Shedding of Bark on the Sun-Baked Side

The patchwork of coloured fresh new bark shows that the bark strips an extra layer deep on the sunny side.
Look also at the coloration of the freshly exposed bark, for asymmetry.
The bark 'stocking' clings higher on the shady side.
It is older, deeper, darker and allows more and older lichens.
Look for the 'curl' of bark ribbons — the pipe-like shapes peeling off — as there may be more curling shown up on the dry side, but it depends on the species.

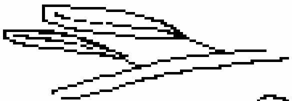

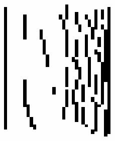

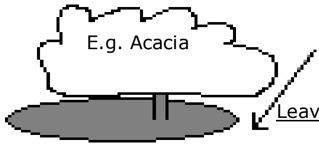

Where to Look for Fading or Lack of it

Turn leaf blades around to compare the greenness in each direction.
Find a pair facing the same direction, and twist one of them, then make a comparison to see which side is more faded.

Twist an *untwisted vertical reed stem*, to detect the more faded side.

Look at stumps and old posts, tree trunks and fallen sticks.
E.g. Sunlight bleaches the sunny side of bark and dries out any secretions
E.g. The dark side of the trunk will often be darkly stained
e.g. with cider-gum secretions.

Leaves dropped into the shady zone will not fade and dry out so quickly.
The shady zone is offset toward the pole,
because the sunlight penetrates deeper on the noon side.

The broken faces of lying timber may be more coloured and less bleached in the shade.

Bumps and slopes, diggings and burrows, may show dead dark moss on the hot side, or grass which has dried up.

Look for a very subtle pink glow on the shady side of dead timber, which is absent or greyer and paler on the exposed side.
Once your brain sees a hint of it, it becomes easier to see—so persevere.

Bracken fronds die off from their tips, but the fronds that retain their greenness the longest are on the lowest, shady side.
Dead thistle stems fade on the Sunny side.

The topsides of grass blades fade more due to being blown over flat by strong **wind**s, and so are exposed to sunlight, wind & drying.
Upwind edges of green bushes often show a brown wind-scorch.

Fresh diggings (echidna, bandicoot, wombat) expose soil to bleaching. Look at the steepest sides for the least faded soil.
Some fading is due to **oxidization** (exposure to air), some to **dessication** (loss of humidity), some to sun-**bleaching** (exposure to ultra-violet). Again it will pay to study the various effects before you need to rely on them.
Also be aware that diffuse lichens will give a covering colour to dirt, wood, rocks, etc, which may complicate the appearances.

Button-grass flower-stalks fade but darken on their exposed side, perhaps due to fungus. One side is dark grey; the other pale yellow.

Make a Clinometer — a Slope Measurer

One person holds the loop knots apart under slight tension. He barely touches both plumb-bobs onto the ground, to settle them.
Their identical lengths ensure a parallelogram.
The second person alters the middle string until it just touches the plumb-line.
The identical lengths ensure an isosceles triangle.
The 5 cm marks represent whole degrees of slope. Use 6 m of non stretchy cord. Measure and constru

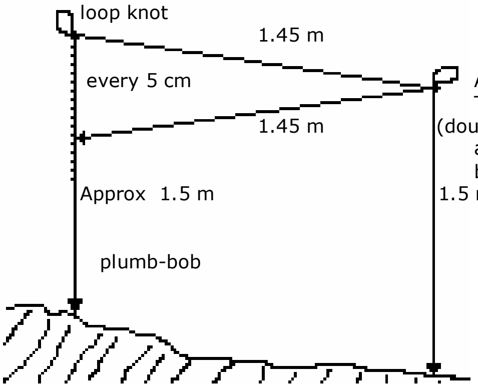

E.g. the slope is 11°½ here…

Accurate from 0° up to 20° or so
The method is based on the isosceles triangle rule
(double the half-angle sine; nearly linear for small angles)
and on 1 in 56. You can use 1.40 to 1.50,
but the lengths must be exactly equal.

The one-man version has two foot-loops.
Spread you feet and hold up the plumb bob.
See which side the plumb leans to.

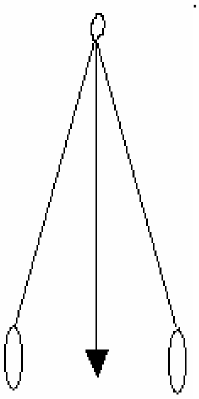

Use it occasionally, to refresh your judgment of slight slopes.
It may also be used to gauge the height of landmarks or stars. (Hang the second loop knot from a sky hook; Sight along the top string; Steady the bobs). Used horizontally, it yields horizontal angles. (Complete the string parallelogram; Sight along the two 'top' strings; and double the reading)
Of course this is only for a hobby, for interest's sake, for mathematical curiosity, not for absolute accuracy!
The more you play, the more you learn.

Just Draw It

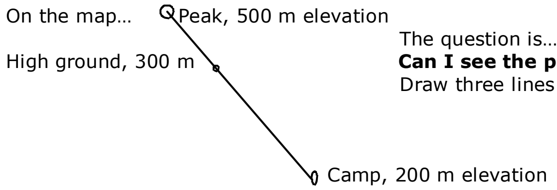

The question is…
Can I see the peak from camp?
Draw three lines of 5, 3, 2 units…

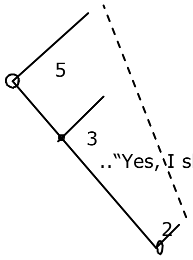

.."Yes, I should see over the ridge to the peak"

Radio Direction Finding — As Good As a Compass

AM stations — Amplitude Modulated — the older style, for medium distance transmissions
— also called Medium Wave; Medium Frequency — MW; MF

Your portable radio has an **internal aerial** for AM stations, which is direction sensitive.
It usually lies lengthwise inside the radio, as a long **ferrite rod** wrapped in a wire coil.
To make that ferrite rod horizontal, you probably need to lay the radio down flat.
Now twist the radio to change its horizontal direction.

When the internal ferrite rod points at the medium wave transmitter,
the reception drops off — you have detected a "null".

This occurs at two opposite orientations — **pointing away from the transmitter as well**.
(Good reception is by pointing the radio 90° away from that null direction.)

FM stations — Frequency Modulated — the newer style, of shorter distance transmission
— Very high, Ultrahigh frequencies — VHF; UHF — It does not use a ferrite rod.
With a whip aerial, or an extensible aerial, or a long wire aerial,
you can detect a null by pointing the aerial at the transmitter.
This is for horizontally polarised transmissions. Some stations are polarised vertically.

Shortwave stations — Amplitude Modulated — higher frequencies, for longer distance transmission
— SW; High-Frequency, HF — It usually requires an external aerial.

Use a **long wire aerial**; keep the ferrite rod perpendicular to the long aerial.
The null occurs when the long aerial does not point to the transmitter — but it is perpendicular to that direction. (Opposite to FM)
Point the long-wire aerial perpendicular to the transmitter, for a 'null'.

Weak AM stations can also be used, provided you have a long wire aerial to pull them in. The exact length sometimes matters—you will notice the volume rise and fall in cycles as you wind the aerial back in.

THE RADIO "NULL" ACTS AS A COMPASS NEEDLE — it keeps a constant angle to the distant transmitter tower or repeater.
You swing the radio aerial each side of the null to keep checking its direction.
To avoid having to give the radio constant attention, break your course into separate lines of sight ('legs') as with a compass.
Marine navigators know that radio signal direction can be distorted by mountains, or by the signal crossing the coast at an angle, or by reflections off the rigging, but bushwalkers seldom have to worry about that, because they only want a constant reference direction, not an exact bearing.

You could consider setting up a radio beacon so that you can always say: e.g. "The car is that way" or "Home-base is yonder"
End of Section 3.

Distance Guesstimation is a Major Problem for Walkers

Because distance measurement is a major problem, we find it easier to ignore.
We resort to some other method of navigation.
If we simply, out of laziness, give up on logging our pace & distance-already-come, we waste a sensible cross-check.

The first of these three sub-sections will look at the amount of distance you have covered — progress — distance made good…

- Elapsed distance — puts you on the map precisely — unless you are not very sure of your progress! or the map scale!
- Map scale — "It looked so easy on the map!" — Judging time, distance & difficulty is a question of reading the map-scale properly. "How far could we get by lunchtime?". Guessing at a curved distance is always a bit of a bluff. If you hear yourself saying "On the map that's **about** 5 km" stop and measure from the map-scale more accurately!
- Sideways drift — is a matter of trigonometry, when you discover an error in your direction heading
- Up/down progress — is in another dimension. You don't need an altimeter, but it helps. (If you do take a portable altimeter/barometer, make sure to test it to see whether or not it reacts falsely to temperature change—put it in the fridge then in a warm place; then don't forget to keep adjusting it according to the atmospheric pressure changes)

The second sub-section is about the rate of covering distance — pace — and distance-yet-to-go.

The third sub-section is about the distance you can see, but don't cover — estimating the 'range'. How Far to That Landmark?
E.g. You can identify a peak on the map — but only if you are certain enough of the distance from you to it.

Distance = Speed x Time

E.g. 4 km per hour x 3 hours = 12 km
Hint: Measure only elapsed, 'walking' time; Delegate someone to keep track of accumulated time out for stops and diversions.
Hint: Mentally bring forward your starting time e.g. "(as if) we started at 10a.m." (You started at 9, but had two, ½ hour stops)
E.g. foam takes ½ a second to pass your 10 m yacht; speed equals 10/.5 = 20/1= 20 m per second.
E.g. I seem to have drifted 12 km in one day = ½ kph current. E.g. The cloud shadows cover 5km in 2minutes = 150kph wind.

Keep a Log!

That's how mariners keep track—they write it down in the special book. Do you carry a notebook?

Adding Two Speeds

Convert speed to a length (e.g. 4 kph = 4 cm) and direction to an arrow.
Add the two arrows, tail to tip, in either order.
Complete the triangle or quadrangle tail-to-tip.

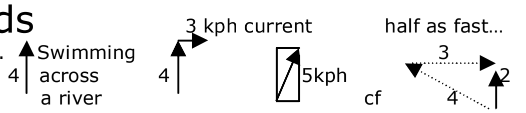

Note that it is faster to swim a river by pointing straight across and letting yourself be swept downstream, than by trying to stem the flow by pointing upstream. Only when the two speeds are perpendicular do they not affect each other e.g. the 4 km per hour is not affected by the current in the first example. Or e.g., If your boat speed is 4kph ahead, and the tidal flow is 5kph against you, your actual movement is backwards! (with any sideways component added in independently).

Example: Three air-speeds ↗ + wind ↘ = actual progress ↗ very fast across the ground, slightly to your right.

I.e. Your speed & direction through the medium (air/water), relative to the ground + the medium's movement, over the ground (the medium carries you when it moves) = your total speed and direction, over the ground.
Lastly, *to analyse apparent wind speed, to avoid confusion, draw only the three winds involved*. I move this way...
First convert your progress over the ground (top arrow) to an opposing apparent wind (right-hand arrow opposing it).
Don't try to add wind to your motion—that's not a sensible question/answer. The medium (air) doesn't carry you when it moves. Answer: I feel the actual ground-wind, a cross-wind (→) in my face, diagonally as shown…

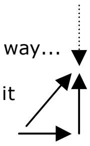

Range: Rule One: Guess

Then don't trust any calculation which doesn't agree with your initial guesstimation.
But remember that distance is deceptive over featureless terrain e.g. water/beach — we underestimate it.
Birds bob their heads to see how much an object shifts against foreground and background. It's worth some practice.

"Range" of Visibility — Two Halves to the Picture

"Extreme visual range" must be calculated in two bites — So Don't Forget To Add!
— e.g. when asking "Should we be able to see Ben Nevis from Brady's Peak?"
— e.g. at sea when you can first see the lighthouse lens, not just its loom.
From a kilometre high you can see 1° of latitude
before the curvature of the earth cuts your line of sight.
— e.g. you can see a 1 km high mountain from two degrees away
if you are on a 1 km high mountain — see diagram…

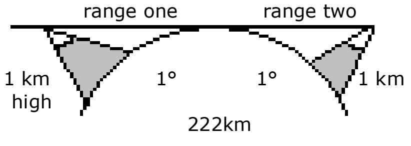

"Range Squared is Dee Height"

This is an easy mnemonic. "Dee" is dee Earth's Diameter i.e. Extreme Range = $\sqrt{(D \times H)}$ see p.92c
H is your height above the intervening horizon e.g. the wave tips, or a 250m plateau.
Since the Earth is 40,000 km around, divide by Pi to get **12,700 km diameter**.

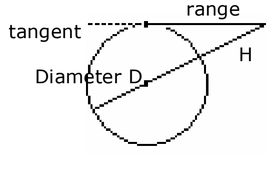

The formula is a theorem in geometry,
about the geometric mean (the tangent) of an extended diameter and the extension length (H).
The formula states: the square (range x range) = the rectangle (not shown: H x D+H).
(i.e. the range substitutes for both parts of the product, 'geometric-' or area-wise)
but since the height H, in reality, is very small, therefore **range^2 = D x H**

To “Read” a Map — Rule One: Don’t Lose Your Place!

Get the geometry fixed in your vision — of where you are on the map e.g. “in the top right grid square, next to that lake”.
And then don’t keep losing your place every time you look at the map again!
..Whether the map is now upside-down, side-on, or whatever.

Rule Two: Imagine Arrows, for the Sun, & Your Movement

Again, visualise them geometrically, in relation to the whole map shape, from one corner of the map, for the Sun, and from where you are on the map, for your movement. When your map is properly oriented, the movement arrow will point in front of you, ahead, matching the trend of your movement onwards.

Rule Three: Get Used To Any New Map Scale

Maps are “Fractal” They show “self-similarity across scale”
i.e. A river system will look the same on a small-scale or large-scale map — there is always finer detail to show.
A road network is the same — the smaller the scale, the more roads are left out, leaving a similar looking network.
A coastline shows such self-similarity. Magnify the details and it still looks like ‘a coastline’.
New Map-Scales are Hard to Adjust to
It takes time and testing to get used to a map scale, because they all look alike no matter what the scale.
To get used to the new ‘feel’ of a map, check up regularly on your map-progress, until you are no longer surprised by how far you have or haven’t come.

Rule Four: Don’t Give Up!

Don’t let the combination of difficulties bluff you!
E.g. You misjudge the scale, misjudge the Sun’s direction by optical illusion, you see turns & bumps not marked on the map, doubt the accuracy of the map, fail to assess the slope accurately,… and so eventually… give up in confusion!
Problems are normal, and **solving them by the map is normal too — don’t give up trying**.

—Also See “Map Checklist” p.113—

“L.M.N.Or P.”

L for Landscape; M for Map; N for Compass Needle: Or for Either/Or (the big point of the mnemonic)**; P for protractor**
Either/Or means: Each component of map&compass work can be used alone and independently for some tasks, if desired.
Plotting compass bearings can seem complicated. This mnemonic is designed to smooth the way.
A navigational compass has a protractor independent of its needle. The trick is to remember that!
The protractor has 2 parts too (p.93b). Note that some compasses have the magnet one-piece—glued onto—the protractor card.
If you get used to using each component separately, you will find ‘map and compass work’ to be simpler.
The trick is to isolate each of “LMN Or**P” in turn, to avoid confusion.** That includes the 2 moving parts of the protractor.
It is hard to use more than two at a time! so know which two you need to use at each stage.
E.g. A ‘prismatic’ compass will let you see both **L**and & numbers (**P**) at the same time—to take an accurate bearing e.g. at sea.
E.g. You can align the **M**ap to the **L**andscape, without ever using a number. Or you can point the **N**eedle to the horizon likewise.
E.g. The **P**rotractor may be used to assess slopes, to measure the Sun’s height, to take a map bearing, to draw an angle on the map, to take a horizontal angle between two landmarks, to sight a vertical angle like a sextant (this may need two people), to find North from Sun-time, or to set a course to the left of the Sun.
So don’t forget to use the **P**rotractor to its full potential, quite independently of that funny little piece of metal rattling around inside it. (**N**).
E.g. An emergency ‘compass’ is just a floating sewing **N**eedle, or a suspended magnet. So the protractor has little to do with orienting the map by the magnetic field, or with heading North by the compass needle.

If your navigational compass has a magnifying glass, a set-square and a graduated ruler, you should hardly get confused by thinking they had something to do with the magnetic field! — so don’t worry when you see some numbers marked around your compass needle — I think you learnt how to use a protractor in elementary school, didn’t you?

Working from land to compass is called “Taking a Bearing” (L+P1,N+P2) You use the protractor to put a number on the direction
..from map to compass is called “Getting a Heading”. (M+P). You then use the needle (P+N) to find then follow that heading (L).
Working from compass to map is called “Plotting a Fix”. (P+M). I.e. You use the protractor, ruler and set square to draw angles.

Maps Mean: Imagine Me! — Making Mental Maps

The meaning of using a map is that **it lines up**
— with what you see; with what you saw; with your mental maps; with what you know; with what you can’t see & don’t know.
E.g. You glimpsed the coastline as you drove in, the lie of the valleys, the drainage of the creeks, and of course you know the highway that you drove in on! You can no longer see them, but it is normal to orient yourself by imagining the invisible.
And then, your mental map is still incomplete, so that you need to use the map to add to your imagination…

“Map To Mind” Navigation

Transfer from the map both visible and invisible features back into your mental model and mental layout.
1. Start with the Invisible Big Pictures — visible on the map but not to you.
Look at the sky and the skyline as you interpret the map for what lies out of sight.
2. Add in the local visible features — visible both on the map and to you.
These two steps work together like lock and key — each is not very useful without the other.
Yet both steps are often neglected and seldom put together as a whole.
“The cars are now behind that notch in the skyline”
“The highway runs behind that hill then turns South-East”
“That low spot marks the river we want”
“Our camp tonight is in line between Mount Socks and the pine plantation yonder”
When you become temporarily bushed in, or blinkered by the rain, or shut in by the dark — then you see whether you can maintain proper orientation — by the invisible — or whether you just feel lost — “I can’t see where I am!”.
If you do not make a mental map and orient yourself by imagining the invisible—you will be disoriented,
even if you “can see where you are”.
♦ Part of the job is to look at the map in order to join your several local mental maps into one regional map.
Remember that a map is a great memory aid, or a great memory bypass — take your choice.
♦ Rather than “put yourself on the map”, put the map in you!

Use the Map Scale — Don’t Just Look at It!

Many errors of judgment result from guessing at the map distance, even when the map scale is clearly marked.
The scale is not much good down in the bottom corner of the map, fixed in place! So copy the map scale onto cardboard.
Somehow attach it to the map e.g. into a separate paper-pocket; onto a string; in the map bag.
The string is useful for measuring distance accurately. You place it over a curved path then stretch it out straight to measure it.
Point the corner of the cardboard at the starting position; position the string from there; then stretch the result along the scale.

The Jargon

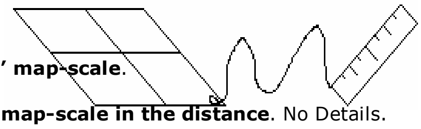

“Large-Scale” is like a “Full-Scale Model” It shows all the details.
Like a ‘large-scale simulation’ — it leaves out nothing. **Picture a magnified ‘Large’ map-scale**.
“Small-Scale” is like a “Small-Scale Model”. You can’t see the small details.
It is like a ‘small pilot study’ — not full scale. **Imagine a very hard to read small map-scale in the distance**. No Details.
“Full-Scale” is “1 to 1” like the Cosmos. “Half-Scale” is “1 to 2”. *“Small-Scale” is* “1 to a million” *like Cosmological diagrams.*
The bigger the second number, the smaller the fraction. The bigger the big number, the bigger the span on the map.

Map-Scaling-Factors are Meant to be Interpreted

1:100,000 means “One unit represents 100,000 units”
Split this into two parts: (a) 100 & (b) 1000 and **use centimetres i.e. 1/100 m** — 1 cm: 100,000 cm
Say (a) 1 cm represents 100 cm i.e. 1 m — **drop off two zeros and call centimetres ‘metres’**
And (b) 1 m, from part a, represents 1000 m i.e. 1 km — **drop off another three zeros and call metres ‘kilometres’**
So the answer is: “1 cm to 1 km”
1: 1,000,000 1 mm = 1 km I.e. **You can also drop off three (or 6) zeros and call millimetres ‘metres’ (or km)**
50,000 units makes you calculate ‘1/50th of a metre’ (=2cm), so 1:50,000 must be ‘2cm to 1km’. Or multiply to 2:100,000.
1: 250,000 Turn this into 4:1,000,000 by multiplying both sides by four. 4mm to a km (since 1mm is 1/1000m)
or (a) ÷100: cm to m (i.e. 4cm to 10,000m no longer cm to cm) (b) ÷10: cm to lots of 10m (c) ÷1000: cm to 10km
the grid should be marked in **10 km squares, of 4 cm each**
Your compass will be marked in centimetres, but perhaps also in special scales e.g. 1 inch to a mile (1: 63,360)

Pace Your Progress — Pre-Dict It, Don’t Post-Dict It!

To avoid looking at the map continually, to locate yourself continually, predict the estimated time of arrival **once**, and relax.
We are normally lazy about looking at our watch often, poor at calculating, on the hop, distance covered, and lazy about locating ourselves on the map often. Mental arithmetic is clearly noxious, so we try to avoid it. Working abstractly from time to distance to map, after the event is too hard for us. But we usually only resort to the watch after we begin to doubt our position. (See “the 11:59 syndrome” — we do doubt.)
Something needs to be done about this!

Consequently, the antidote to us refusing to log our distance, is to anticipate the timing instead
i.e. to work more concretely, from map (I can see that) to distance (I can visualise that) to time (I can calculate that).

- Examine the map beforehand — as a first step not a last step. You will have to do the calculation later anyway, on the trot! Why not do it now, calmly.
- Guesstimate the time needed to cover the distance ahead — to the next recognisable milestone. E.g. use 4kph, ¼hr/km
- Look at your watch, as the last step, when you near your goal.

In this way the **landscape ‘means’ time** to you; “When we reach the bridge, we turn left — and that should be in 25 minutes”
Your guesswork will get better too.
You will find yourself looking at your watch more frequently, with the question: “Should I be there yet?”.

Anticipation Is Exciting. Back-Calculation is as exciting as the washing-up.

“Range Squared is: Dee Height” “Dee” is Diameter

i.e. Extreme Range of Visibility = √(D x H) on this side of the horizon. **Don’t forget to add the ‘beyond the horizon’ range**.
Five round-number rules are highlighted. By playing around with some approximations, we can find some rules of thumb…
127 is close to 128 and 125 and 11.3^2; Pi (3.14) is close to √10 (3.16);… I hope these tricks will make the square roots easier

40,000/pi x 40/pi km = **400** km range from 12.7 km Notice the two related doubling sequences I have given.
40,000/pi x pi km high = **200** km range from 3.14km You can only see twice as far from four times the height
12,500 km x 4/5 km = **100** km range from 800 m high 125 is five cubed
113^2 x 9^2 / 10^2 = **100** km range from 810m 34 km from 90 m; 11.2 km from 10 m — 1/3 from 1/9 height
12,500 km x 1/5 km = **50** km range from 200 m high “Rule of (lesser) fives: **50 km from 1/5 km**”;
12,500 km x 1/20 km = **25** km range from 50 m high Multiply or divide H by four to double or halve the range
12½km range from 12½m e.g. from the top of a beach

12,700kmx.0127km=**12.7km from 12.7m—**from a ship’s deck, to the horizon, when estimating distances across open water.
12,700 km x 1.27 km = **127** km range from 1.27 km high 1/10 from 1/100. Remember the number 127 must be 400/Pi;
100,000/8 x 10/8 km = **125** km range from 1.25 km 0.125 is 1/8

12,800 km x 1/2 km = **80** km range from 500 m high or “Rule of (greater) fives: **50 miles from 500m**”
12,800 km x $1/2^3$ km = **40** km range from 125 m high You can also derive 160 km from 2 km; 800 km from 50 km
40,000/√10 x 0.4/√10 = **40** km range from 127 m high = “400km from 12.7 km” = 1/10 range from 1/100 the height.
12,800 km x $1/2^5$ km = **20** km range from 31 m high = 200 from 3.14 above.
12,800 km x $1/2^7$ = **10** km range from 7.8 m high ..if you bother to halve 1000m seven times!
12,800 km x 1/500 = **5 km range from 2m high** **From a dinghy/beach**. 1/5 range from 1/25 height cf 50m high.

21,600 nautical ml/pi = **150** nautical miles from Pi n.ml high approximately, because 150 squared is 22500.
83^2 n.ml x 1 n.ml. = 83 n.ml. **150km from 1 n.ml**. 60 x 360 / Pi is approximately 83 squared nautical miles
12,300 km x 1 km = **111 km** range from 1 km height **1° of latitude, from 1 km high**, approx. “**55** from 1/4” = 250m
111^2 km x 1/9 km **37 km** range from 111 m high 1/3 of the range from 1/9 of the height

Look Carefully at the Fine Detail on the Map

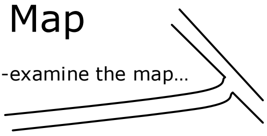

E.g. a kink in the track, just before a junction, may make it look "not like the map!" until… you re-examine the map…
"Oh yes, it is marked like that after all, **I just didn't look closely enough**"
E.g. that maze of fine blue watercourses marked — don't be a-maze-d by it — look at its detail!
There is usually a magnifying glass on the compass.
For a stronger **magnifying glass**, reverse your binoculars. For an even stronger magnifier, remove the eyepiece and use it.

Look Carefully at the Fine Print

E.g. date of compilation ..1864?
E.g. "adjust all longitudes by + 2 minutes of arc"
E.g. yearly magnetic drift
E.g. Contour interval — 10 m or 100 m?

Some Map Details are Not to Scale!

E.g. Two wheel ruts might look like a 100 m wide road on a map (if you took it seriously).
Consequently… 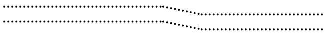 ..may be a 100 m Zig-zag! And… 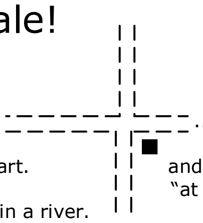 ..is not a cross-road and the hut is not "at the junction"
E.g. A "cascade" symbol… ..might equally mean two huge waterfalls, 300 m apart.
E.g. 'Smooth' bends ..probably hide many small extra sinuous ones. E.g. in a river.
Moral: Keep an Open Mind. **Keep your eyes open** for alternative interpretations.

Using the Protractor by Itself

The protractor has two sets of lines — the lines on the movable 0-360 protractor-circle, and the fixed "follow me" arrow, taken together with the edges. Line up one set with horizontal, or North, or whatever; line the other up with the slope or angle to be measured. The angle in degrees clockwise is the space between the zero and the "follow me" arrow, not the other way around.

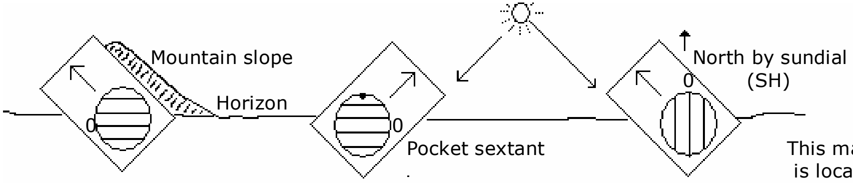

When the angle of the Sun (it moves around at 15°/hr) corresponds with the time of day, you must be facing North/Polaris.
This may be used when the magnetic field is locally anomalous or untrustworthy.
(Tilt the compass axis to Polaris)
Reverse this to get the time from North.

Point the edge of the compass body directly at the Sun until it casts no shadow. (Don't look at the Sun)
If your compass fluid has a bubble in it, it can be used as an artificial horizon.

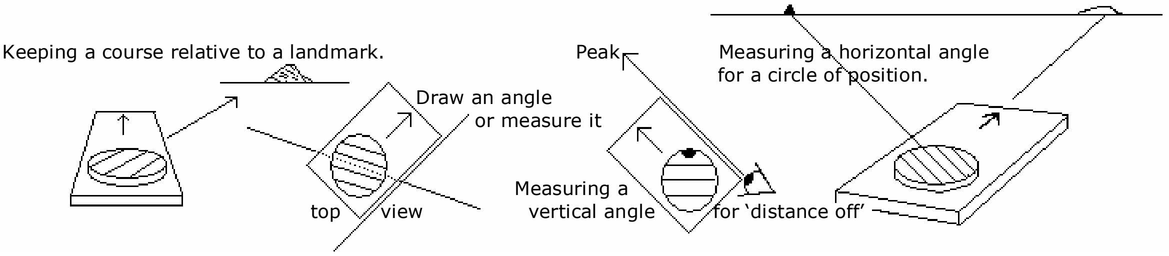
Keeping a course relative to a landmark.
Draw an angle
or measure it
top
view
Peak
Measuring a
vertical angle
Measuring a horizontal angle
for a circle of position.
for 'distance off'

Mental Networks

Our mental maps seem to run on networking — junctions and joins; choices and destinations; routes and 'directions'
Consequently mental maps are prone to logical-type blunders

- Missed junctions e.g. of creeks or paths
- Misidentifications e.g. of segments, features or forks
- Poorly judged measurements of lengths or direction e.g. hairpin bends which aren't noticed
- Mistakes of number

If you reread the section "Logical Blunders" (p.39b) you will notice that most can be seen as a **wrongly imagined network**.
That's one more reason to study the real map, in order to bring your mental map up to speed.
When we function on short term memory of mental branching directions, any shock or daze can be crucially disorienting.

The Most Valued Map

..Is one which you have made yourself. A do-it-yourself job makes it "yours". Practise with a MUDMAP, p46e.

When you copy important detail from other maps onto your map e.g. peak elevations; track notes; waterholes, then your map becomes more valuable than a replacement map.
You add extra details e.g. campsites; walking times; side-tracks & old tracks and suddenly you prefer the old dog-eared version to a brand-new map!
Sometimes it is even better to start with a blank sheet of paper.
But since it does you no good to pinpoint your position on a blank map, unless you are at sea, you proceed to add in the important local details so that you can tell where you are…

To copy a map: **Copy the Grid First**. ..if there is one. Add in many diagonals as guidelines.
The next rule is: **Get the Landmarks in their correct relative position first** **i.e. work top-down.**

The Psychological 'Half-Way Point'

You have to judge and juggle several factors before you decide that you are 'half-way'.
E.g. The "Unknown Territory" Scare-Factor (Ambition versus Safety Margin). In planning your time-budget, you can't relax in a one-way journey through unknown territory, until you know you have a reasonable safety margin in hand
So 2/3 the distance along, might be 'half-way through the test'. Other examples...
The "Point of No Return" — You tend to speed up, once committed to the next goal.
The "Downhill Run" — It is the quicker half.
The "Back-Track Return" — tends to be quicker i.e. the turnaround point can be after lunch.
The Morning Spurt-Afternoon Hurt Syndrome — puts 'halfway' before lunch.
The Side-Track Scheduling — if you don't explore side tracks on the way in, when you are fresh, you may not want to tackle them on the return journey.
The "My Feedbag is That Way" Home-run spurt on the last day.

When You Meet Other Walkers

"When did you leave this morning?" = how far to your next camp.
"Anything about the track that we should know?" E.g. slow sections
"Any deep water crossings" or other hold-ups
Other questions, less related to walking progress and pace, are…
Anyone else ahead of us? (Will there be room for us all?)
Do you carry a track guidebook? (Can I read it?) Do you know the forecast?
Any good water available on the way? Can you recommend a camp site?
Did you pass our tents? (Are they still there!)
Pluck up your courage to ask what you want to know, and be prepared to share what you know.

Advice about Walking Times — Double & Halve!

E.g. Brochures seem to add "return" or "one way" by guesswork sometimes.
So "6 hrs" means 3 to 12 hours until proven otherwise.
E.g. Fit young rangers saying "six hours" may mean "for me; double it for you" or "for you; halve it for me"
E.g. "walking time" may mean "excluding stops, rests and meals" — you could find yourself pushed for time.
E.g. "Not far, you can't miss it, it's easy all the way" could prove disastrous. E.g. ('provided you find the right track')

20 Minutes per Kilometre — Double & Halve — Easygoing Maths

It means **3 km per hour** —an easy-going overall walking pace when you include short rests, difficult vegetation (within reason) and lack of any hurry — tourists on track, walkers off track.
Obviously you need to multiply and divide by 20 to convert kilometres to minutes, or minutes to kilometres.

If the number is small — kilometres — multiply it up to the larger number — minutes — **double** it and add a zero.
If the number is big — minutes — divide it down to the smaller number — kilometres — **halve** it and subtract a zero.
You could also express it as a "half a kilometre every 10 minutes" but the "20" makes the maths easy;
And even simpler, to cope with the zero, express yourself in 'lots of 10 minutes'.
E.g. 5½ kilometres will take about… (double it, to get many minutes)… 11 **lots of 10 minutes**
E.g. "We've been going 50 minutes — (five lots of 10 minutes) — ; that's about… (halve it, to get a few km)… 2½ kilometres"

4 km per hour — Normal Walking Pace — Double and Halve!

You can easily double or halve your pace. Start with 4kph as a sustainable rate of progress, and use the doublings 1,2,4,8…
8 km per hour is someone tall walking in a hurry. 2 km per hour is someone small dawdling. 1kph is for very difficult going.
Coincidentally 4 km per hour is 10 miles in four hours or 2½ miles per hour.
Incidentally also, at 4 kph, a landmark 15 km on your left will change direction at 15° per hour as you walk past it.

Double, the Acute Angle, off Your Nose — For the Distance Off

This double-angle method puts numbers on a very "natural" method of gauging distance…
In the first diagrams you walk from the bottom left to the top, and watch a landmark, ahead on the right, pass you by.
The distance walked = the distance off if you walk far enough to double the original angle (away from straight ahead).
The initial angle must be acute. You need to maintain a straight course throughout.
The simplest case uses the diagonal of a square. 45° changes into 90°
The next simplest example is the equilateral triangle. 60° changes to 120°
Followed by the 30° isosceles triangle. 30° changes to 60°

But the general case is made clear by the geometry of two equal radii in a circle.
It is because the angle at the centre of a circle is double the angle at the circumference.
The angle at the centre is the final angle—double the initial one. top view…

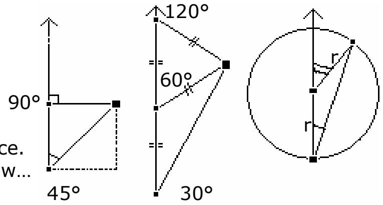

Ahead of you, the angle is always less than 90°.
In the second diagrams, below, the landmark may not be ahead of you and the initial angle may not be acute...

If you walk backwards, from top to bottom, using the top method in reverse, you get...

Halve, Any Angle, from your Stern—For Distance Off,

..i.e. In Retrospect The initial distance off, not the final distance off is what must be travelled.
The initial angle is now the double angle, and need not be acute.
I.e. turn the bow-angle diagrams upside-down…
E.g. 90° becomes 45°; or 120° becomes 60° becomes 30°…

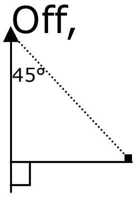

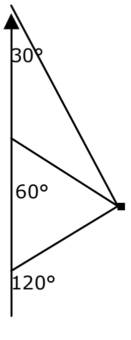

Permanent Landscape is More Trustworthy Than Surface Detail!

When what you see "doesn't match what is on the map" have the sense to distinguish between 100 m deep contours which would take an atom bomb to shift, and walking-track junctions which a fire could reshape.
Professional map-makers and satellite photos rarely make mistakes — in landscape positioning.
But when they add the local details, it is another story.
Time changes such details also. New roads, cleared bushland, second bridges — these will mislead and confuse you.
When your mind says: "The map is wrong!" remember that what the map says — about landscape — is not wrong.

Details Added on to a Map Are Prone to Error

Notorious for errors are: Map North; Map Scale; Roads.
So beware of anything other than a good-quality official government map.

Expect The Map To Be Unhelpful

Many important details will not be shown on the map — you just have to expect that.
You can often trust what is on the map; You can often trust that what you want, won't be on the map.

Maps Do Give You 10 Times More Detail Than You Can Notice

After you get lost and finally sort it all out on the map, you notice… "Oh yes, silly me, those contours are **up**slope not downslope! And they **do** show a creek-line, by those sharp bends! Oh, of course, there is **a peak** on the right, and that fenceline opposite — I missed both of those when I first studied the map. You **can** see Mount Bruce, lined up with the lake, from there, but not at all from over there. And that squiggle in the blue line, is the hair-pin bend that tricked me, but I should have realised that all rivers must drain to the north-west eventually…"

Try to reduce your ten-fold folly to five-fold folly for next time!

Using the Needle Alone — Emergency Magnetic Compass

- Magnetise your repair-kit sewing needle; It will probably already be magnetised somewhat. Stroke it with a South Pole toward the sharp end, so that the North-seeking pole is at the pointy end.

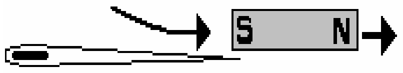

- Float the needle in calm water, on the surface tension (no detergent!)
  (a) lay the needle on toilet paper
  (b) lay both on a bowl of clean water
  The paper will wet and sink, leaving the needle afloat. You could also try sticking it through a piece of foam.

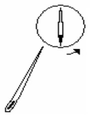

- Hang any magnet horizontally, from hair or fine thread. It helps if you know which end is North-seeking.
  In the field, test it against a Natural North, eg, Bush Noon. At home, test it on a compass. Like poles repel…

Hint Understand that **magnetic poles should be called 'North-Seeking' and 'South-Seeking'** and that **our "North Pole" is a South-seeking pole!**
(so that North-Pole seeking poles point to it — Unlike poles attract)

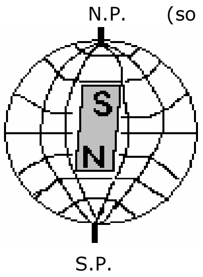

The Earth's internal magnet is as shown (It is slightly off-line with the rotation axis!)

A compass needle on the Earth's surface has a red end labeled 'N' i.e. North-Seeking

The white end is labeled 'S'

Strip Map Navigation — Follow-Your-Nose Navigation

Our mind definitely prefers a strip map — a sequence of "directions" to follow. (None-the-less, make a MUDMAP in conjunction)
Our eyes definitely prefer a strip map — "up" is "straight ahead"! (See "N.O.R.West/N.O.T.Least" "Nose Navigation" p31e)
A strip-map is simply a way of formalising your track notes—every navigator really ought to do it this way…
There are two good ways to make a strip map as you go: with a sheet of graph paper, working from bottom to top, and in a track notebook, working from top to bottom, but…
Both assume that **progress arrows are up-page**. You could turn your notebook side-on or upside-down to write in it.
In either case you set aside several **vertical** lines for each **type** of information,
and set aside one **horizontal** line for each **subsection** of the journey,
and a second line for descriptions and for comments,
and a line for overall information e.g. trend, drainage, season; totals: time, climbing and distance.
The trick is to use symbols. One vertical line for one sort of symbol e.g. compass needle direction
(See the **Symbol appendix** p 122) Another line for another sort of symbol e.g. slope-direction, and steepness
This way you can cram a lot of information into a small space, in a consistent and regulated manner, and be able to read it!
e.g. …

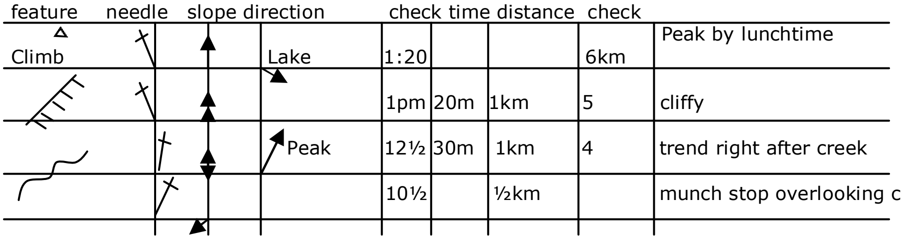

track feature	compass needle	up slope	landmark direction	time check	leg time	leg distance	distance check	comments
Climb			Lake	1:20			6km	Peak by lunchtime
				1pm	20m	1km	5	cliffy
			Peak	12½	30m	1km	4	trend right after creek
				10½		½km		munch stop overlooking creek

e.g...

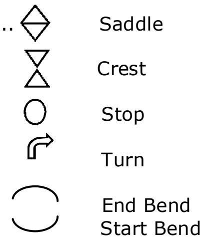

Distance

AIMING OFF by 10° — 10° is **175m per kilometre**

1° is one in 57.3, so one in 5.73 is about **10°**, (and 1 / 5.73 = .175) or **175 in 1000**. (For bigger angles use the 1 in 60 rule)
Alternatively, 2sin5° = 0.174 isosceles chord, yielding "0.174 per one" or 174 per 1000. (167m per 1000m / degree for >10°)
This is easier than working in sixes using "one in 60".

Hint: Express any ratio as "something in 1" or "in 10, 100, 1000"
E.g. '1 in 50' becomes '2 **in 100**' or '200 m in 10 km' (.4 in 20 etc)

Example: "The T-junction is somewhere ahead — about 017°, 5 km. But so that we can't miss it, we must aim off to the right e.g. 027°. We therefore expect to hit the road 875m to the right of the Junction" (5 lots of 175m).
This is how ocean navigators find an island in an ocean: Step 1: sail N/S to the right latitude. Step 2: sail E/W to the right long.
Don't just head straight toward something. If it doesn't show, you won't know which way to turn!
Remember that a 10° error is fairly easy to make when using a compass over rough terrain (or trackless shifting ocean)!

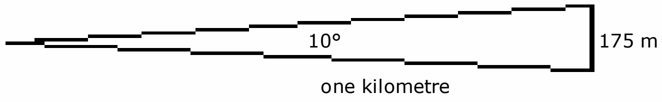

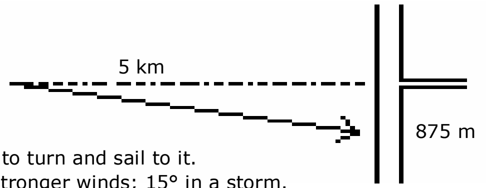

Aiming-off at Sea...

Aim-off to arrive up-current & up-wind of your destination—to make it easy to turn and sail to it.

- Aim-off 6° to counteract leeway when sailing into the wind; 10° in stronger winds; 15° in a storm. Measure the appropriate angle from the wake e.g. 6° is how much your pointing-finger-tip shifts when you change eyes
- Aim-off 6° (1 in 10) to counteract a 10% (1 in 10) cross-current. This refers to the perpendicular component, compared to your speed. The rule works +/−1° up to 36°—up to 60% cross-current. To avoid judging components...
- Aim-off 1°,2°,3°,4°,5°,6°, to counteract a 10% current at 10°,20°,30°,45°,60°,90° to course, either ahead or behind.
- Aim-off to avoid GPS-induced collisions caused by very accurate course-holding. You should plot a course off to one side of the obvious line on the chart, to avoid vessels coming the other way on the same line!

Aim Off 6° or 10%

Blink your eyes at the destination, and when you get there, pace off the 6° offset. Ten-tuple the result for the distance there.

---

1 Metre per Second is 3.6 kph

This is merely a logical reference point for your calculations. (And slightly faster than "20 minutes/kilometre" (3 km per hour))
60 minutes x 60 secs per min is **3600 seconds in one hour**, at 1 m per second, is 3600 m in one hour, or 3.6 km per hour.

One metre is Up to your belly button A child's double pace One strong fast stride

One second is the time it takes to pronounce "Alligator One" in a normal fast voice—don't let it drag. Follow it with "Parramatta two", "Oodnadatta three", (or "a thousand and four", but "alligator" keeps its speed more consistently when drowsy).
Get used to the rhythm of tapping your finger in time with the seconds — your mind subconsciously remembers the timing.
Take a Watch, Perhaps a Stopwatch
1 m per second is very nearly 2 knots.

"1 Hour Extra per 5 Hours"

You can normally walk at 5 km per hour, on a track, on a short walk.
But you will sustain an average 4 km per hour, when you include fatigue, and rest stops.

"1 Hour Extra per 500m easy climbing"

Gentle up-climbing is tiring—allow this extra time. (We are not concerned here with how long it takes to rock-climb!)
Down-slope walking may speed you up, or slow you down, depending on the slope.
Dense tussock country (e.g. snow grass; button grass) is the equivalent of a climb, in its up-down-up steps. Crossing the drainage can also be as tiring as climbing, in its regular climb-descend-climb-descend path.

Distance by Optical Rangefinder — Accurate up to a Kilometre or Three

We are woeful at guessing distance in that range, so there are focussing devices available which are designed for it.

Range by Parallax and a Piece of String — DIY Rangefinding

Animals shift their heads to see how much an object shifts. Blink you eyes to see the effect; walk sideways to practice it.
For more distant objects, the very slight sideways offset can be magnified then measured...
The landmark you are measuring appears in a slightly different direction from each of two different viewpoints.
Pack a long non-stretch string e.g. a polymer kite-string. Know its length. Mark it every 5 m before you go.
Use it as a giant pair of eyes—a long baseline—& to mark out a large square or rectangle on the ground, one side at a time.
Guess at the right-angles — they don't need to be accurate. Just make sure to use equal tension when you stretch the string.
But use a plumb-bob to mark the corners of the square on the ground accurately — with a pebble.
Use the plumb-line also for accurate sighting of the landmark. (You can pack up the string once you have placed the pebbles.)
Twice as big = four times as accurate (in measuring d) If you have enough room to move, aim for over 10 m sides.
Range = ab/d or a^2/d if a = b; from the similar triangles involved. Measure the shortfall, d, at the front, or the overfall, d, at the baseline.

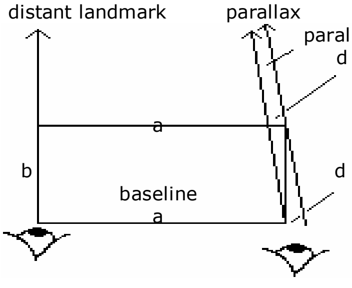

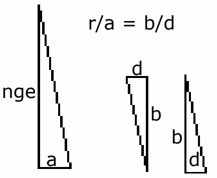

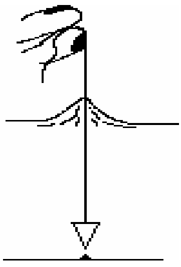

An easy way to remember the formula is to say that the area R d = the area a b because if you shear the ab rectangle, it becomes a long thin parallelogram R long, and d wide, preserving its area.

As a result, the formula becomes...
Range = Area / parallax offset, d
km sq. m mm

The Range=ab/d formula also *makes sense*
The range *should* so depend on a & b and d

“Maps Grip the Universe”

When you swing the map in your hands, around with you as you turn, the whole Universe doesn’t follow it! — it doesn’t twist!
Instead your map gets disoriented.
Leave the map behind, as you twist! — leave it lined up.
The map is designed to lock into the landscape, to ‘map onto’ it, and should be used like that, not twisted out of alignment.

You need to swing the map an *equal and opposite* amount to neutralise what you do.
It’s called “swinging the map” (when you leave it locked onto true North while you swing to left and to right).

It should not be called that!

It should be called “**NOT Swinging the Map** with you”

Map work includes ‘not swinging the map’. See “Magnets Grip Maps”. The most common method is to match the map to a compass needle p.103b

Maps mean landmarks — those landmarks there! Open your eyes — and line the map up with the visible landscape. p101a.
i.e. orient the picture of the land — the map — to map onto the actual land.
Line the Grid North up with North e.g. by compass or by Sun.

You may have to twist yourself, or put ‘North’ at the bottom left, but to always put ‘North’ up the top, is to divorce your eyes!
Yes, you may have to read names upside down! Yes, it is a bother, to keep changing its ‘up’ every time you turn.
But it means that you can just lift your eyes from the map to see what it points out. And can know what lies beyond that.
If your map is not consciously oriented, how can you be oriented, to other than what you can see in front of your nose?

Think “+/- 5°” For Compass Bearings; “+/- 10°” For Following a Compass; “+/- 15°” For Using a Compass!

Compass bearings “+/– 5°”
Test yourselves: Ask each person in turn to nominate where “magnetic North” is on the skyline
using one particular compass, handheld. **My results:+/- 5° “User Inaccuracy”**
Now try putting that compass on the ground. **My results: +/- 3° “Needle Instability, friction” etc**
Try stretching a fishing line over the needle. **My results: +/- 1° “We still can’t agree!”**
Next try that best method for different compasses. **My results: +/- 3° “Calibration and Zero Errors”**
Now try **not** being very careful, with an **untested** compass, at night, when you are tired!

Following a compass blind: Walk on a fixed compass bearing for 100 m or so in a fog i.e. eyes glued to the compass.
Repeat it e.g. turn around & see where you come back to. Give the next person a try at doing better.
I think it matters which eye you use and which hand you hold the compass in.
We don’t necessarily walk ‘straight ahead’ when we try. We don’t necessarily point the ‘follow me’ arrow straight ahead.
We don’t do very well at integrating and averaging out any diversions around obstacles.
Slopes play havoc with your sense of a ‘straight’ line ahead.

Try this…

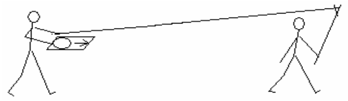

The front person finds the route.
The back person guides the route-finder.
The navigator logs the route directions taken.

Using a compass: You may plot a single line on the map, or as +/- 5°, but mentally allow an extra +/- 10°, i.e. think +/- 15° (for accumulated errors, possible blunders about magnetic variation, etc.). Keep the possibility of error in the back of your mind

---

Caving Strip-Maps — Come in Books

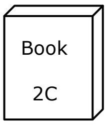

A “cave-directory” booklet records **each separate section of a cave on a separate page**.
E.g. Make your own…

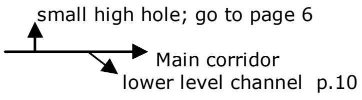

You may need separate booklets for different levels of the cave system.

Use the same symbols as before, for needle, slope, turns, etc.

Use Abbreviations: E1, E2, E3… for Exits; L1, L2, L3… for Light sources; Parentheses (L1) = “Can see L3 from here”

Make up names for each section e.g. “Chamber of Horrors”

Add copious notes e.g. “water level up to ½m deep”

Forethought and cooperation is required to quickly and effectively map a warren or a labyrinth.
Initially you make a muddy map, as you go, but then you add a neat copy into the book later.
(But keep your muddy copies, for next time.)

Up-down Progress

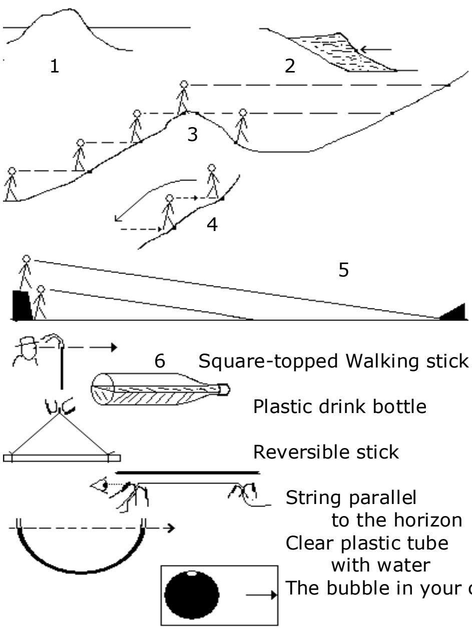

In the first diagram you can see that you are halfway up the mountain
alongside of you — **use the horizon to gauge your altitude**.
Note: the horizon drops below true horizontal as you go higher.

In the second diagram your shadow is obviously halfway up the shadow
of the hill you are on — **use the shadow of the mountainside.**

To find your tiny shadow:
Close one eye — to line up more accurately.
Look past the shadow of your head falling onto your hand
Look for a 'halo' — a bright spot directly down-Sun on the plain
Wave your raincoat around
Well, you know where the shadow should be by now!
So you don't even need to see your shadow!

Keep count of height in body-lengths
(3) Walk up to a stone which was horizontal from you before.
(4) Downhill can be counted by leapfrogging with two people.

Half-Way across? Provided you stay in one position...
Looking parallel from half the height gives you half the distance
(5) E.g. when someone is swimming across a lake

Artificial horizons allow you to maintain parallel gaze (6)
E.g. "The edge of the lake is 'two fingers low'

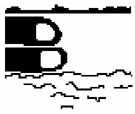

<horizon (or **any distant landmark**)

<lake edge (now, halve your height above water e.g. by squatting down)

Adjust Your Expected Pace

! Three hours up, two hours down — are typical proportions, except in very steep or loose descents — 6 hours up, 6 hrs down.

! You walk more briskly while fresh, in the morning, than after a whole morning's walk. Pace yourself with this in mind.

! Groups travel slower. Allow the slowest person to set the pace, because stragglers may battle or even panic if left behind.

! Retracing your steps is quicker, since it has fewer diversions, less interest, and no unknowns. Allocate less time to return.

! Carrying packs is not necessarily slower, on flats. If your pack is uncomfortable, you won't dawdle, or even want to sit down, and can't stop to look at insects or flowers or to chase birds.

! River crossings chew up enormous amounts of "fussing around" time — organising and reorganising afterwards.

! Beware of smoothed out "routes" marked on maps. The actual terrain may involve many obstacles and consequent slow going and detours. The words "route only" are trying to tell you that.

! Your pace quickens surprisingly when the party falls "into the groove" of walking (without talking, noticing things or relaxing often).

! Path finding can make you nervous especially in unfamiliar territory, which tends to quicken your pace without you knowing it.
The same person, following a leader, is relaxed — and slows down.
So when you take the lead and you wonder why the others seem to fall behind, realise that you are too keen.

Don't Overcompensate for Slopes And Meanders

The worst case for slopes is about 36° for rock scree. But that adds only a quarter to the distance shown on the map. It is far more relevant to say "It's the climbing and the awkwardness which will slow us down more".
The worst case for obstacles is grass-tussock negotiation — but that only adds a maximum of 50% to your distance covered.
An ordinary zigzag may only add 10%

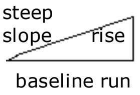

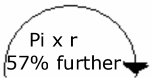

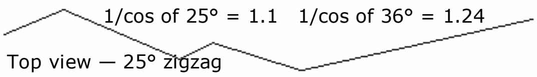

Distance by averaging Maximum & Minimum Guesses

Yes it is worth doing, because it reduces the slop in your individual guesses.
E.g. "More than a kilometre, for sure. Less than 3 km surely! Split the difference & say 2 km

Distance by Comparison with Some Length You Know

E.g. The central business district—1 square mile or whatever
E.g. old-fashioned 'chains' i.e. cricket pitch lengths
E.g. 100 m race; playing fields
E.g. "a day's walk"
E.g. that island which you can see, which the map says is 1 km long.

Distance by Cloud Shadows — Look for Them

- Notice how quickly the cloud shadows move here (use distance/time), then look how slowly they are moving over there
- Time how long they take to get there.
- See how large they look here, and how large there.
- Look how high the cloud base is here, then compare the size of hills in relation to that ground-to-cloud gap, over there.

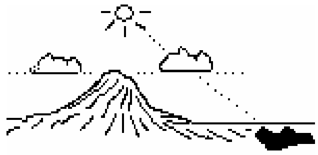

Hint: To see which shadow attaches to which cloud, hold a string,
to pass from Sun to shadow (past its cloud) but… Don't look directly at the Sun!

Sound Travels Three Seconds to the Kilometre

This tells you how close the lightning storm is. "(Flash), Alligator One, Alligator Two, Alligator Three, Boom!, One Kilometre!"
If you see a flash to ground, you can find the distance there.
Otherwise, the distance is to the nearest part of the lightning flash overhead.

Align the Track

Line the map up with what you see! Often that is the road ahead, or a well-marked track.
This is a great trick for the car navigator when using a road map — it's so easy to see the line!
It's equally helpful when following any marked track on the map with the map aligned to the landscape — it locates you on the map — on that bit of the map which lines up with the track.

If you were landing a fighter jet onto an aircraft carrier in the fog, you would not want to miss the runway!
Well, think of the real track as the runway on the aircraft carrier, the land as the sea around it, & the tiny map-picture of the track as the jet plane.
Twist and turn your 'jet plane' until it can come in to land properly. Even when you are in the fog!

Assuming you have a properly oriented map…

You can find where you are on it…

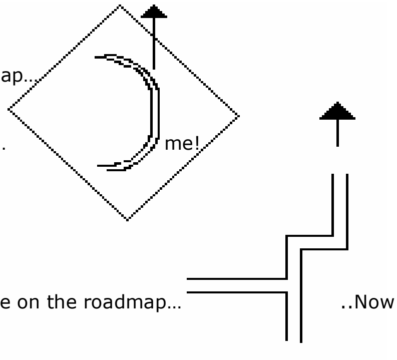

If you know where you are on the roadmap…

..Now the map is properly oriented.

Compass Errors — Be Impressed! Be Careful!

Electrical currents swing the needle e.g. reading the compass by torchlight; untwisted pairs of DC wires e.g. marine compass lights; Mobile phones; GPS; radio. "Impressed (DC) current" anti-corrosion systems in marinas.
Magnets e.g. other compasses, motors, loudspeakers, clocks, ironstone. Note: Iron-rich dolerite doesn't look rusty.
Iron & Steel e.g. your pack, pocket knife, keys, wrist watch, cutlery, steel-capped boots, glasses, geology hammer, batteries, torch, studs, zips, pens, trig points, shed, fence, car, ironstone, buried machinery, coils of old fencing wire half-buried in the ground, tools, fuel stove, internal springs, reinforced concrete, railway lines, power pylons, pipelines. Navigating in a mineshaft.
Many of these might have permanent magnetism too. Stainless steel is often magnetic.
User error
Parallax — viewing the needle from one side; viewing the scale from side-on.
Misreading — e.g. any scale marked every 2° is easily confused. Mis-setting the movable scale.
Miscalculations — especially applying the variation the wrong way — adding it instead of subtracting. Confusing left and right.
Out of date chart of deviation. Deviation is the general name for any compass error—apart from expected, normal variation.
Quality of Construction
Needle instability; under-damping; over-damping; motion. Let it settle. Put it on the ground.
Vibration — has unpredictable effects e.g. twisting the needle clockwise. Walking provides regular slow vibration! Motors.
Sticky pivot — Tap the compass Un-centred pivot
Zero-error — the needle doesn't match the 'zero' Mis-calibration of the protractor
Needle mis-magnetisation — not quite longitudinal; amateur attempts at re-magnetisation.
Reversed polarity — Poor quality needles can pick up remagnetisation from nearby strong magnets without you knowing it!
Variation The difference between True North and Magnetic North changes with time and place. The map is out of date.
Magnetic drift of the magnetic pole — we only have 'predictions' or 'trends' of an unpredictable variation. Are you up to date?
Latitude change — alters the magnetisation of surrounding fixed iron e.g. a car body or a boat's motor or hull.
Local anomalies e.g. Tasmania is full of ironstone dolerite. Many rocks preserve ancient magnetism.
External magnetic fields e.g. which way a boat moors, or who moors next to it. Competing fields change or capture each other.
Lightning strike after-affects.
Correcting mechanism out of **adjustment**.

Map Folding and Map Holding

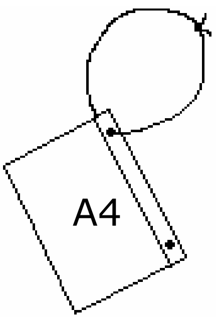

Maps aren't forever — they blow over waterfalls, tear & get unreadable.
Pack a spare
Cover the whole surface with (non-shrinking) clear plastic film
With good-quality laminating plastic & cutting, taping and sealing, maybe it will last forever!

Folds and Corners lose details, and wear through.
Cut along the fold lines cleanly e.g. to A4 size

Use a photocopy of your own map to use in the field, to keep your original map in good condition.
Wet maps lose details when rubbed.
Get an A4 plastic sheet protector for the map. Tape it closed. Hang it around your neck.

Mapwork requires bearing-lines to cross the whole map.
Tape the (A4 or other) sections back together with wide clear tape, leaving wide gaps, as shown.
— you can butt the edges closely when needed; the tape will crease out of the way
— you can refold it at will, as often as you like, and show the two most useful sides out of the many sections.

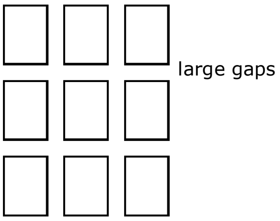

Distance by Double Pace — Test Yours

A ‘mile’ (or “thousand” in Latin) is a thousand double-paces, on forced march i.e. 5.28 feet (Romans were small people)
My dawdle double-pace is 5.25 feet. What is yours? In metres?
Answer: Test it. Test it again and again. Average it. Each person is different. Relax, don’t push yourself.
Test it again uphill, downhill, laden, unladen, tired, in a hurry.
Test it at one double-pace per second — walking by the clock.
It helps to make a 10 m cord. Use a tape measure to make the cord, then use the cord.
E.g. I know I use 12.5 **paces (single steps) per 10 m** i.e. alternating 12 & 13 double places per 20 m.
and I walk at 4.5 kph; 1 1/4 metres per second; eight **seconds per 10 m.** The timing varies more easily than the distance
While watching the seconds tick over on my watch, and striding out in time with them, I manage 1.5m/sec, get tired, and step further out.
You Can Be As Accurate As You Have Been Careful. If you are full-grown, working this out once can set you up for life.

Tie Your Ankles Together!

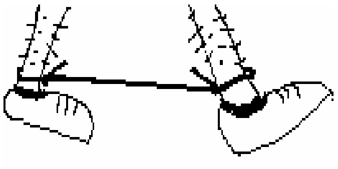

For a well-measured pace (heel to heel) use a 50 cm cord, for instance.
Count double-paces, and don’t lose count — see ‘finger math’

Test your pedometer

Pedometers work in different ways, so test yours on different terrain. Your knee-lift is higher on uphill sections, and your pace is shorter in rocky terrain. You can make your own pedometer from an old odometer, working off the flex of your boot, your hip, your walking stick, or whatever, or you can roll a small bicycle wheel along the ground, fitted with an odometer.

Don’t Lose Count! Use Finger-Math

Finger counting is a silent and **number-free** counting method — you only pronounce the number after you have finished counting. Count automatically, by placing your fingers down, each in turn, onto a hard surface.
A thumb counts for five fingers, and the **left-hand counts 10 per finger**.
When you run out of digits you have 99, so you next say aloud “A hundred” and start over again with no fingers down.
Place each finger down onto a surface, in a **definite sequence**: Right Index (“1”); Middle (two fingers down); Ring (3); Little (4); Lift all four and put down the thumb (“5”); Leave the thumb down and put and leave down the index finger (6); middle (7); ring (8); little (9); Lift all five and put down the left index finger (“10”); Then put down the right index finger again (11), and so on, until you need to put down the left middle finger in place of the other five (two ‘tens’ down).
With practice, you can count, add, subtract, multiply and divide — almost thoughtlessly — as quickly as with an abacus.
We are concerned here with **counting single or double paces**, e.g. in a fog, and doing so **in lots of 10m** (or 20m or 40m for double paces). I.e. you count up to about 12 or 25 single or double paces mentally, and add “one more lot” to your finger tally.
If you are in a group, you can check each other’s counting by stopping where you think every hundred metres is.

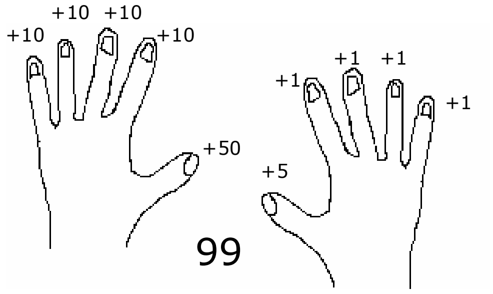

Lift all your fingers off
and nod your head down
for ‘100’, then start again.

100 lots of 10m equals 1km
100 lots of 20 m equals 2 km
100 lots of 40m equals 4 km

or 100 double paces is 1/10 of your mile

Lighting Conditions & Eyesight

Make allowances for optical effects…
Looking up-Sun, especially in the afternoon, you see haze (backlit dust & insects) which gives a false impression of ‘distance’.
Looking down-Sun on a clear morning after rain or after a cold change (i.e. minimal haze) things ‘look close enough to touch’.
As your eyesight begins to fail in mid-life, everything seems ‘distant’ but turns out to be surprisingly close
i.e. you have been judging distance by crispness of detail all your life…

Visual Acuity is about One Minute of Arc

..which means that you can spot a 1 m object at 3.6 km on a clear day (since one minute of arc is (1/60)° or ‘1 in 3600’).
I.e. You can see details several thousand times smaller than the distance to them.

This is how seamen can judge distances over open water, to land, even though there are no other clues to go by.
They ask: “Can I see trees? What about tree trunks? Or branches? Or leaves!”

Lighting, Contrast, Haze and Eyesight play havoc with accuracy; But experience is a great teacher.
Feed your mind and memory by deliberately looking through the haze to the finest details you can pick up.
Use binoculars as a cross-check. Have competitions with your family.
Children have better eyesight, but the idea is to judge the distance according to how it looks to you.

Align the Map to the Landmarks

Make the Map Match the Marks — if you know where you are on the map you need only one landmark…

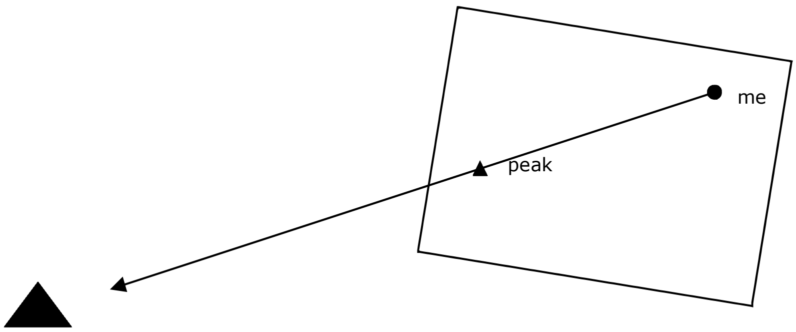

The easiest way to keep the map aligned to the landscape is by then using the direction of the Sun now, a very visual method.
The commonest method uses a magnetic compass to find North.
This one-landmark method can be used whenever you can see something in the distance.

If you don't know where you are on the map you can still orient it to match the landscape, but you will need two or more landmarks identified to do this.

Compass Use — Master It, but Be a Good Master.

- Use it to orient yourself, then put it away!
- Use it to orient the map or to choose a route, then put it away!
- Use it as a protractor. Use it as a ruler. Use it as a map scale measurer. Use it as an artificial horizon. As a set-square.
- Use the set square for a back-course or a side-course without altering the forward-set-course.
- Don't rely on it — use other methods of direction-holding.
- Let it point your eyes outwards, to the horizon.
- Use NORWEST/NOTLEAST, if you have to stare at it — e.g. in a cave.
- Hold it 1½m above rocks; 10m away from cars, sheds, etc; several item 'lengths' away from small items — test it.
- Let it point out a foreground mark as a destination, then follow a visual line of sight
  e.g. 'leapfrog' or manoeuvre by 'legs'
- Don't consult it again until you need to check or choose another point.
- Don't trust it, but trust yourself to it, when you need to.
- CHECK ITS ERRORS. Check it against the map by taking a bearing from a known point to a distant landmark
- Take a pair of forward and reverse bearings—to and from a chosen mark, to check for local magnetic field anomalies.
- Check your errors in using it, e.g. on the move; against someone else.
- Interpret it for orientation value, not just for route-finding.
- Consult it more often than you want to, when you are 'not using it'.
- Check your guesswork as to e.g. 'Gully North' or 'Wind-West' by it.
- Take it.

Layout Map — Blind

- This project requires a pedometer to measure every hundred, metres or paces or something. A stop-watch could do.
- Use graph paper, so that each square on the graph paper represents one 'leg' of 100 (or '1 minute's walking').
  (Hint: Graph paper is a handy thing when making a casual mud-map of your progress, for this very reason.)

Try it at least once. Without triangulation and cross-checks you can build up the layout just by reading the compass.
i.e. Beware: This is a bottom-up method, prone to inaccuracy. After one accurate practice-run, you can do a better casual job.
I achieved over 90% efficiency at coming back to my starting point after a 24 km circuit on my first try!
(In fact I know that the 8% error was mainly only due to a few blunders in marking the wrong direction.
Also, I only measured 'direction' to +/–15° since I was also testing the 'clockface method' of direction — to the nearest ½hr.)

When the compass read "2:30", I needed a reverse clockface on the map, and I needed to mark…

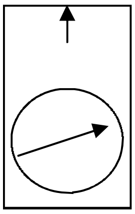

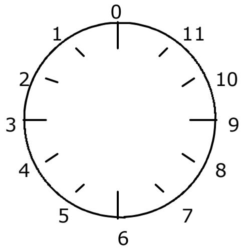

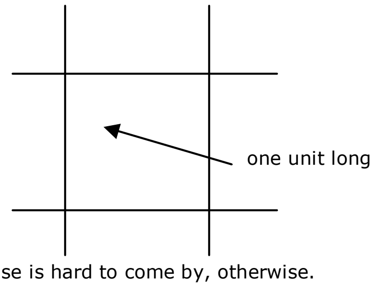

(Hence I made a few left-right blunders which I had no record of!)
You don't need a map for position-sense, if you make your own as you go. Position sense is hard to come by, otherwise.

Line of Sight Can Be Extremely Accurate

..But is often neglected, for far inferior methods.
E.g. locate yourself on a 'position line' as a means of marking your progress…
E.g. keep a straight line by walking to a point ½ way between you and your next destination
E.g. 100 fence posts can be positioned to look like one when seen end on.
E.g. fishermen can **locate** their offshore fishing spot again
— "The jetty is underneath Mount Lofty, and the tower is behind the Town Hall"
You can sometimes consult a map to fix your position...
E.g. You may notice that **"this part of the track runs directly toward"** a peak.

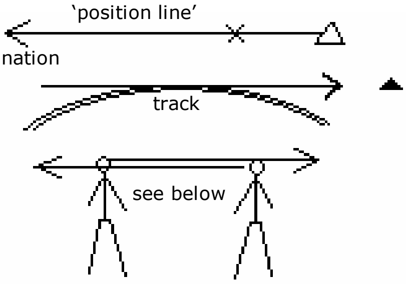

Un-mappable lines abound. Choose the **nearest** objects in front for best accuracy.
E.g. "That far rock lines up with this bush. **Remember** them!"

Use mapped features and the accuracy is still extraordinary.
E.g. "The peak lines up with the intersection, but if you step only three paces left you can see the difference!"
The alignment doesn't need to be accurate, provided you can memorise the exact inaccuracy e.g. 1° to the left of the hillside.
Two such alignments are needed for a fix. Look around: "What is in front of Mount William? What is behind the lighthouse?"

Two Opposite Landmarks

It **more than doubles** the alignments available to you if you have eyes in the back of your head!
With two people, separate some distance, for accuracy, then **look past each other's heads**.
Walk slowly around in a circle looking at the horizon for possible line-ups.
With one person, **place a rock on the ground then circle it**, scanning the horizon. Place a pebble to mark each feature.
The good thing about this method is that one landmark cannot 'blot out' another — all the visible features are available to you.
You may use this method to stick to a direct 'track' between two landmarks.

A Slow Kilometre Cancels out Three Quick Ones

Let's try to keep up an average of 4 km per hour…
First we travel 2 km in half an hour, but then we hit a slow section and manage only 1 km in the next half-hour (2 kph).
To counteract this we would need to triple our slow speed to 6kph for the next half-hour!
So we then decide to stop and catch our breath for half an hour.
To counteract that we would need to double our desired average speed (to 8kph) for the next ½hr!! Or 6kph for 6km.
i.e. a wasted hour takes several hours of catch-up, to regain.
(2kph for ½ hr + 6kph for ½ hr = 4kph; 0kph for ½ hr + 8kph for ½ hr = 4kph; 0kph for ½hr+ 6kph for 1hr= 6km in 1½hr)

Slow sections chew up the time available, so don't dawdle.
Quick sections can't easily repair the damage, because they don't last long.
You may try to be time-efficient and walk more quickly on the easy sections i.e. hurry along the flats,
but you can't gain much time that way.

To Progress Faster More Easily
Pace Yourself — Slow Down, but Don't Stop

i.e. Pace yourself so that you don't need rest stops, since they ruin your time budget, and it is time which determines pace.
It is better for your overall pace to keep walking slowly than to stop completely.
Go slowly on the hard sections. Pick up your pace, maybe, on long flat easy sections.
Shorten your rest stops e.g. don't remove your packs.
The old rule to cover a lot of ground quickly is: run 50, walk 50 (paces or double paces).
Catch the paradox of it — to keep going faster, go easier.

Approximate Figures for Visual Acuity

At 1:1,000 "I can just count my grandchildren"	You need to supply figures which work for you.
At 1:2,000 you can count individual items.	Multiply the size of the smallest objects you can **count** by 2,000
At 1:3,000 you **can't count** them. E.g. people/shrubs/birds	Multiply the size of the largest objects you **can't** count by 3,000
At 1:4,000 you can **see** items, but can't count them.	Multiply the smallest items you can **see** by 4,000.
At 1:5,000 you **can't see** them individually.	Multiply the largest details you **can't** see by 5,000.
At 1:6,000 you can **read unfamiliar letters**	Multiply the letter-stroke widths you can **read** by 6,000.
At 1:7,000 you can't **read unfamiliar letters**.	Multiply the letter-stroke widths you **can't** read by 7,000.
At 1:8,000 you can **read familiar signs**.	
At 1:9,000 you can't read familiar signs	
At 1:10,000 you cannot see that there are individual letters.	
At 1:15,000 you can see single objects e.g. trig points against the sky.	

Of course it's rough! I told you so in Rule 1: Guess!
Adjust the numbers to your eyesight e.g. "I can only count to 1500".
The idea is to build up your subconscious guesswork.
E.g. "I **can count** to 3000 and can count house windows" try 1 m x 3,000 = 3 km away
"I **can't quite see** heads" try 25 cm x 4,000 = 1 km away

Test Your Own Visual Acuity

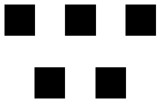

Walk up to a signpost until you can read it. Pace the remaining distance.
Measure the size of its letter strokes.
When do you lose sight of a dollar coin?! Can you count heads in a crowd at 600 m?
Set up a 1 cm test pattern of squares (for others) to count. Try yellow squares too.
Test yourself in different conditions — on an overcast day, in direct sunlight, in glare, on a hazy day, and at a large distance.

Align the Map to Identify Landmarks

Lift your eyes from the properly aligned map back onto the landscape… "There should be a peak in the distance over that way." "Look for an outstation in this direction" "Can you see the transmitter structure??"

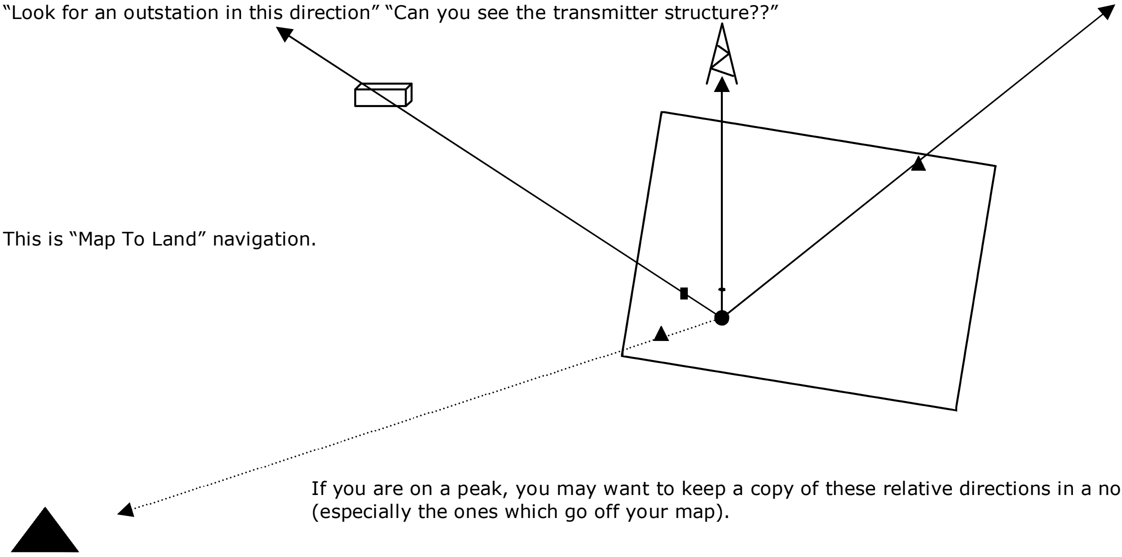

This is "Map To Land" navigation.

If you are on a peak, you may want to keep a copy of these relative directions in a notebook (especially the ones which go off your map).

"Magnets Grip Maps"

This mnemonic is meant to remind you to twist the map to the north — the map should be one piece with the compass needle.

- Place the compass on the map
- **Grip both the map and the compass in one hand**
- Turn the whole map, or yourself, or both — so that the magnetic needle points to the map's magnetic North

The result is that the map is aligned to the landscape.
Keep it that way!
Even while you walk, and twist, and turn… This is called "Swinging the Map".
A one-off effort is called "Orienting the Map".
Swinging all three of you is called "Orienting yourself" e.g. to North.

Align Your Map in the Dark

The commonest way to align the map to the landscape is by a magnetic compass.
The magnet — the compass needle — must 'grip the map', magnetic North to magnetic North, in order to keep the map oriented.
One method is to set the compass protractor to zero — true North — and
set the compass edge to grid North/true North (see diagram) and then
set the magnet needle to magnetic North — as marked on **the compass rose on the map**
Another method is to line the edge of the compass up with the magnetic North mark on the rose.
Then you simply keep the needle point pointing to zero on the protractor.
A third method is to set the compass to the correct compass heading then line the protractor edge up with the desired course on the map, and the needle to the protractor, then follow the compass arrow.
A fourth method is to set the compass to the correct true course on the map, then line the compass needle up with magnetic North on the map, then follow the arrow. But **always line up the two magnetic Norths!**

The Magnetic North marked on the map is crying out to be matched up with… ..the magnetic needle

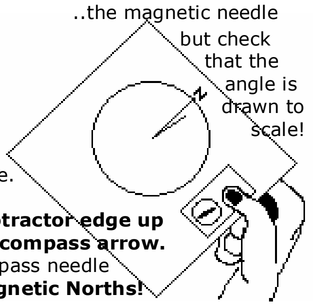

but check that the angle is drawn to scale!

A Compass Holder

Rather than hanging the compass around your neck, where you hardly ever see it, you can make it constantly visible.
For map making, it is indispensable.

Version 1: Aluminium wire around your neck and around the compass, like a bib or harmonica holder…
The compass circle is grasped by a circle of non-magnetic wire.

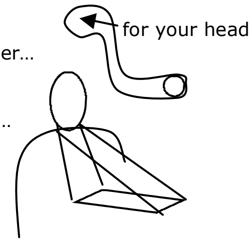

Version 2: Hang a flat map board from your neck, and blue-tack the compass on to the map board…
Drill holes in the four corners of a light board.
One string loop goes around your neck and through the two close corners.
Another string loop goes around your neck and through the other two corners and has an adjusting knot in it (rolling hitch).

The compass protractor is set to 0° i.e. Use Magnetic North
or it is set to x° i.e. Use True North
To orient the map, or yourself, turn around and face North.
For a back-bearing, turn around and use the white end of the needle to read your direction — where you have come from.
Use the "NORWEST; NOT LEAST" mnemonic to keep you in touch with normal orientation. See p.101c & 106b for how to build up a map, or keep track, blind, using such a compass holder.

"Changeover Lines" of Progress — Relative Apparent Sizes

All visual elements which surround you have an apparent size, each of which changes as you progress.
E.g. "The hut behind us seems so tiny now, poking out from the sea of snow"
E.g. "That hill now seems to loom **high** above the plain" (a vertical angle).
Two apparent sizes will, at some stage, become equal, at e.g. as you move away from one object, which shrinks, and toward another, which expands. Simply notice "the changeover" and you have 'a line of position' to log your progress.
"The hill is the **same size as** the leaning over gum tree" (a vertical angle, a diagonal angle, and a changeover line)
You don't need to remember what size you measured — just compare them for equality!
"The width of the patch of forest is much less than the island is **across**" (horizontal angles)
"But the island alongside us is about the **same width** as the **gap** between the two" (another "changeover point")
You don't get a single point fix, only a line of separation. The third and fourth diagrams each show an example.
A measurement without numbers is like saying "We are halfway there" instead of saying "4.5 km on"
or like waiting until you are "Exactly East of Mount Big" rather than measuring "079°C" at any time
or somewhat like "These two line up, if you put a hand span in the gap"
The closer the objects, the better. The more distant, the more careful you need to be!

Use the Landscape Itself to Measure & Mark Progress

Example: "Where I came out onto the beach,
Sloop Rock was one island to the right of Gannett Island"
A top view shows that this is simply a line of sight, locating a point next to the island.

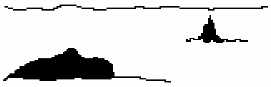

Island = gap;
An apparent-size changeover line

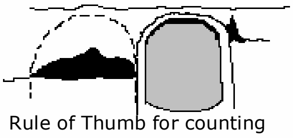

Rule of Thumb for counting

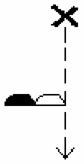

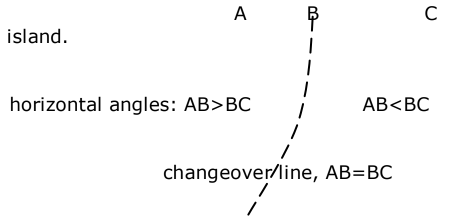

°

Distance-Off by Vertical Angle

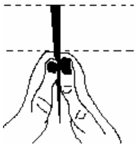

The simplest range finder is a piece of grass held at arm's stretch, (if your fingernails are too small).
Measure the apparent height of something which you can recognise later,
then measure how many millimetres of grass was needed!
You can relocate that distance-off later by its apparent height e.g. distance from a prominent peak;
from a lighthouse seen from offshore; from any recognisable tree; or from a cliff with a clear base to measure up from.
This is how our mind analyses the skyline heights — it is a very 'natural' method.
Yacht navigators will find out the height of a lighthouse from a chart,
measure the angle with a sextant, and calculate the range by trigonometry…

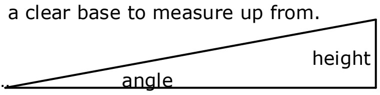

To fix your position requires two separate measurements e.g. two different vertical features
In theory, the plotting looks like two circular arcs intersecting, because any distance-off must give you a circle of possible positions.
The closer the features, the smaller the circles, and the more accurate the fixing of the location.
Confuse your friends with: "1 inch away from the windmill, and 13 mm from the pine tree"…

Safe Distance Off

You might be sailing around a headland or reef, and want to maintain safety…
Without needing to know where you are you can maintain a minimum distance off.
E.g. "no higher than 1° apparent height by vertical angle"
E.g. "keep the horizontal angle acute, between those two points"
E.g. "stay outside the apparent radar range of 5 km"
E.g. "deeper than 20 m"

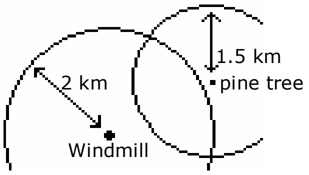

Discard one of the two possible positions.

Circle-of-Position by Horizontal Angle

Here you keep track of the horizontal space between two prominent points; or any noticeable gap; or the apparent width of an island. Use a string stretched across your line of sight, or a stick gripped between two fists.

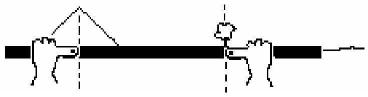

In this horizontal case though, unlike the vertical angle, the 'circle' of position is really only a cusp — part of a circumference, which must pass through the two points, and your position.
If you are closer — a wide angle — the arc is flatter; a narrower angle means you are further away — a larger part of the circle lies on your side of the landmarks.

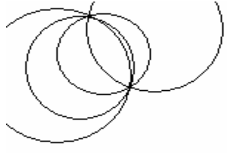

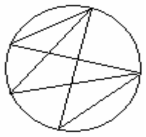

There is a constant angle
at anywhere on the circumference
when looking at the two points.

You don't need to draw this though.
Only memorise or note the angle.
To fix your position make two separate measurements e.g. one vertical angle and one horizontal angle too; or two horizontal angles—using three landmarks (or two vertical angles; or a radar range & direction from a single landmark; or use any other line(s) of position e.g. a compass bearing or line of sight or an advanced position line).
Now you can relocate that buried treasure, that fishing spot, waterhole, rare plant, good campsite… simply through making a couple of notches in your walking stick.

The best range-finder

..is a theodolite/sextant — designed to measure angles very accurately. You can use them horizontally.

Align the Map to Fix Your Position

Match the map to the landscape visually.
(The first result is to show you geographical (true) North.)
(The second result is to show up other landmarks.)
The third result is to show up **where you must be**.
(The fourth result will be to indicate which way to continue.)

I.e. Work **backwards** — from Land to Map.
One method is as follows…
Place a small pebble on the ground.
Place other pebbles in a circle around that,
each one lining up with some landmark.
Now place the map over that centre pebble.
Twist and shift the map until it lines up.
Get down behind the map and visually sight over it.
When everything lines up you must be at the centre of the circle
— over the central pebble.

Your position should **look** obvious!

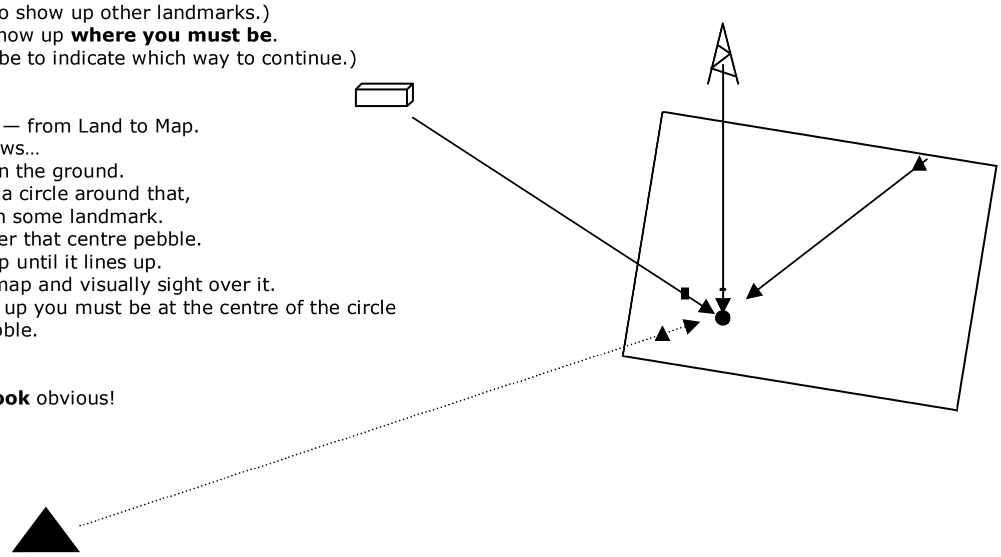

A Horizontal Angle between Features Yields a Position Circle

There is no need to draw this position circle (diagram 1), and it is somewhat awkward to construct, but there are two cases which are well worth drawing (permanently) on the map…

Diagram 2. A right-angle between two peaks — it means that those two peaks mark the **diameter** of the circle — easy to draw.
Diagram 3. The 'circumscribed' circle of three peaks — you can see whether you are inside or outside of that circle.
To find the centre of the circle, use perpendicular bisectors.

The angle seen from the circumference is constant.

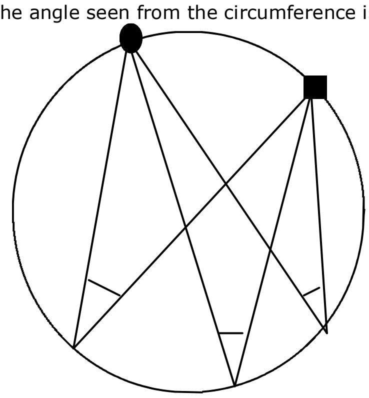

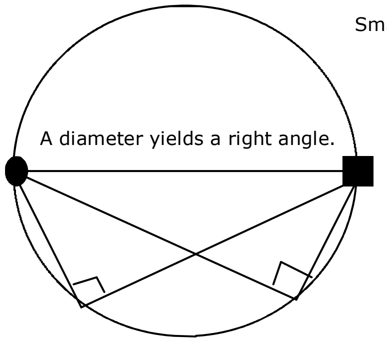

Horizontal angles are **larger** on the **inside**.

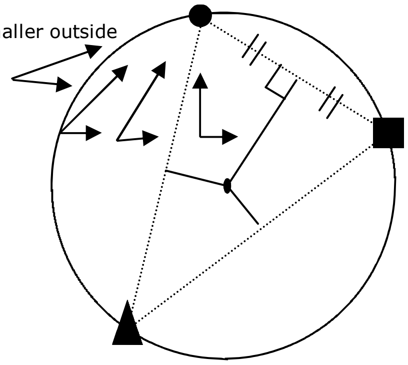

'Two Horizontal Angles' Usually Fixes Your Position

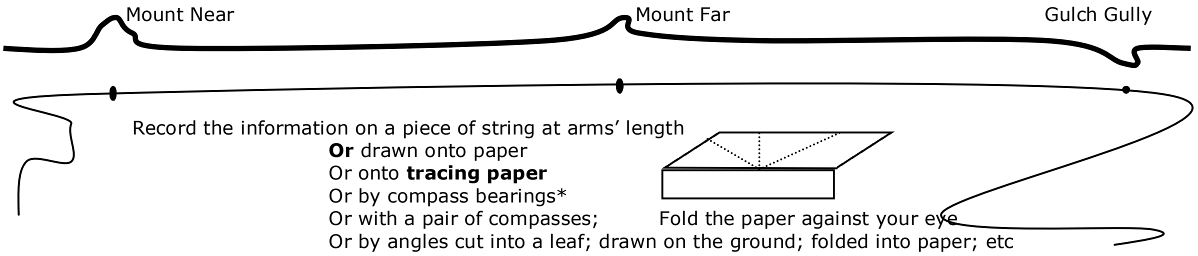

Record the information on a piece of string at arms' length
Or drawn onto paper
Or onto **tracing paper**
Or by compass bearings*
Or with a pair of compasses; Fold the paper against your eye
Or by angles cut into a leaf; drawn on the ground; folded into paper; etc
*you could plot back-bearings, or simply record the three directions for reference.

This is a quick, natural, accurate method which uses no numbers or circles. It is how our mind judges the horizontal layout.
Only one position on the map will produce such a pair of horizontal angles.
If you draw the angles onto tracing paper, you can **place it directly down over the map, to fit it on properly**.
If you make a three-arm trio of compasses with a locknut at the pivot, you can also place this directly onto the map.
Coastal yacht navigators find the method useful, and to measure angles very accurately they use a sextant horizontally.
In theory it plots as two intersecting circles — only one intersection makes sense.
The nearer the landmarks, the smaller the circles, and the better the fix.
Just don't choose three landmarks which are already on one circle along with you!

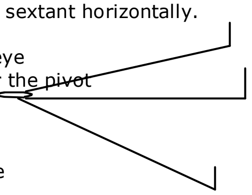

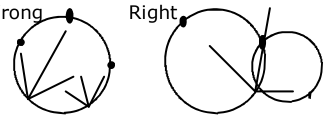

You don't have to draw these circles
Also, if you want to find an angle from the map
you can just read it from the map without drawing any circle

The Sine-Cosine Tally Method of Keeping Track, Blind

This project is only "for experience" — i.e. you are unlikely to want to do it twice! — but you should do it once.
The idea is to **keep a running total of your 'Northing' and 'Easting'** from a starting point.
I scored 97% efficiency in getting back home, on an 11 km circuit, on my first try, with an unreliable home-made pedometer!
It is a 'bottom-up' method — for when you are walking blind — when you cannot work down from the Big Picture.
With pen and paper you record your pedometer readings & directions, and update the numbers, say every hundred double paces.
E.g. "Current Pedometer Reading: 1400; (time-check 1 1/4 mins; pacing check 140 m) efficiency: 90% (slightly indirect path)
Recent Direction 005°; **adjustment needed +90, +8**; new tally: 419, -45; (comments: crossed a wet ditch)"

The chart of 'adjustments' follows this box, with an example.
200% efficiency would mean "I missed the last 100 reading"
50% = "I had a diversion" 110% ="I stopped at 110 — make the next reading 90% efficient"

◇ Set it all out with headings and long columns.
◇ Your judgment and arithmetic don't need to be accurate, because the errors are likely to cancel (!)
◇ Check your pedometer for each type of terrain, by time and by pacing.
◇ Avoid blunders e.g. check that the direction and the adjustments are similar between two legs.

It is quite pleasant to stop every minute or so — at every 100 on the pedometer — It keeps you alert, rested and oriented.
Making a full map (see 'Layout Map—Blind) is more satisfying and just as quick.

Mariners can (either draw their tacks, or, using this method) calculate an overall course-made-good from their many compass legs, by first adding up (distance x sin true course) East, and then adding up (distance x cos true course) North, from each individual distance and course.
Their overall distance is $\surd(\text{sum1}^2+\text{sum2}^2)$ at an angle of arc-tan(sum1/sum2). Then correct for magnetic variation.

Rise and Run Adjustments — Sine and Cos — North and East

Caution: These are COUNTER-Clockwise Bearings! — as seen from a compass holder…
I.e. The needle swings to left and to right while you point straight ahead.
E.g. when the needle points to "20° East of 0°", as in the diagram,
"N(orth) O(n) R(ight)-West" reminds you that you are heading NNW-ish,
not to the East. N20°W is your **anti-**clockwise heading.
Don't let this unusual method of measuring bearings confuse you
about normal clockwise bearings.
There are really only 10 pairs of adjustments used in the whole table
I have made East positive, only to match normal bearings and
Map grids, which use East as positive (West as negative).
Here, **West is negative; East is Positive.**
South is negative; North is positive
Examples: North East is + + N315°W;
South West is – – N135°W;
North West is +– N45°W;
South East is – + N225°W
The + and – signs tell you
how to change your tally of North and East.
E.g.
100 paces to the NNW ish, as in the diagrams is...
94 paces of Northing and 34 of Westing
Or +94, –34

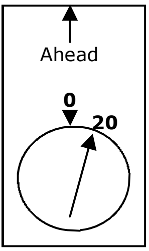

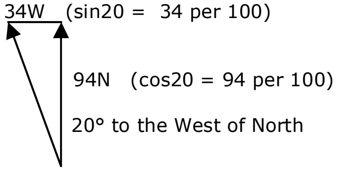

					Change	
Actual direction…	(NE	SE	SW	NW)	N	E
Adjustments…	+ +		− −		cos	sin
		− +		+ −		
Protractor Readings…	360	180	180	**000**	**100**	**0**
N nnn° W…	355	185	175	**005**	**100**	**9**
	350	190	170	**010**	**98**	**17**
	345	195	165	**015**	**97**	**26**
	340	200	160	**020°**	**94**	**34**
	335	205	155	**025**	**91**	**42**
	330	210	150	**030**	**87**	**50**
	325	215	145	**035**	**82**	**57**
	320	220	140	**040**	**77**	**64**
	315	**225**	**135**	**045**	**71**	**71**
The next figures rerun the previous ones...						
	310	230	130	050	64	77
	305	235	125	055	57	82
	300	240	120	060	50	87
	295	245	115	065	42	91
	290	250	110	070	34	94
	285	255	105	075	26	97
	280	260	100	080	17	98
	275	265	095	085	9	100
	270	270	090	090	0	100

e.g. As a different example, a tally of –640 +770 means that you are 640 paces South
and 770 paces East of your starting point, which means that you are now N230°W from home, according to the table.

Rule of Fingernails

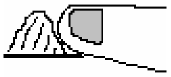

Your pointing fingernail should be about 1°, one in 60, when you stretch your arm.
Your thumbnail at arm's length, held upright, might be one in 50……
A pencil in your fist will be about one in 100.
Check it out with a tape measure and calculator. E.g. 1.2 cm (width) / 69 cm (distance from eye) = one in 57 = 1°
Adjust how you hold your arm (e.g. square in front, diagonally- or fully- on the side) and which fingernail, etc, until you find suitable combinations which you can remember.
But the trick to making this useful in range guesstimations is to guess the distance away first, then estimate the height of a mountain, as follows… 60 km away — use one in 60. One full nail means 1 km showing;
30km away—use 1 in 60 =½km showing; 50km away—use 1 in 50 = 1km showing; 10km away—use 1 in 100 = 100m showing
Reverse logic can give you the distance away, if you can see the base clearly e.g. from sea, and know the full height.

Rule of Tom Thumb

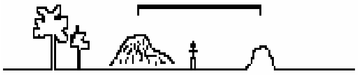

At various times you have to judge parallax.
And the trick is to project the needed sideways movement onto the intervening landscape.
Imagine yourself over there on that ridge — Are you as thick as that tree trunk? Are you as tall as that bush?
To make the two landmarks line up, imagine the distance over there in the distance.
"I would have to walk that far sideways over there" = "I would have to walk the same distance sideways over here"

Apparent Size

The size of people, cars, houses, as measured at arm's length, varies with distance. You know how big they are already.
To get a good gauge of very small angles, compare their width & height with a key or knife-blade or toothpick, held edge-on.

Which Way is North?

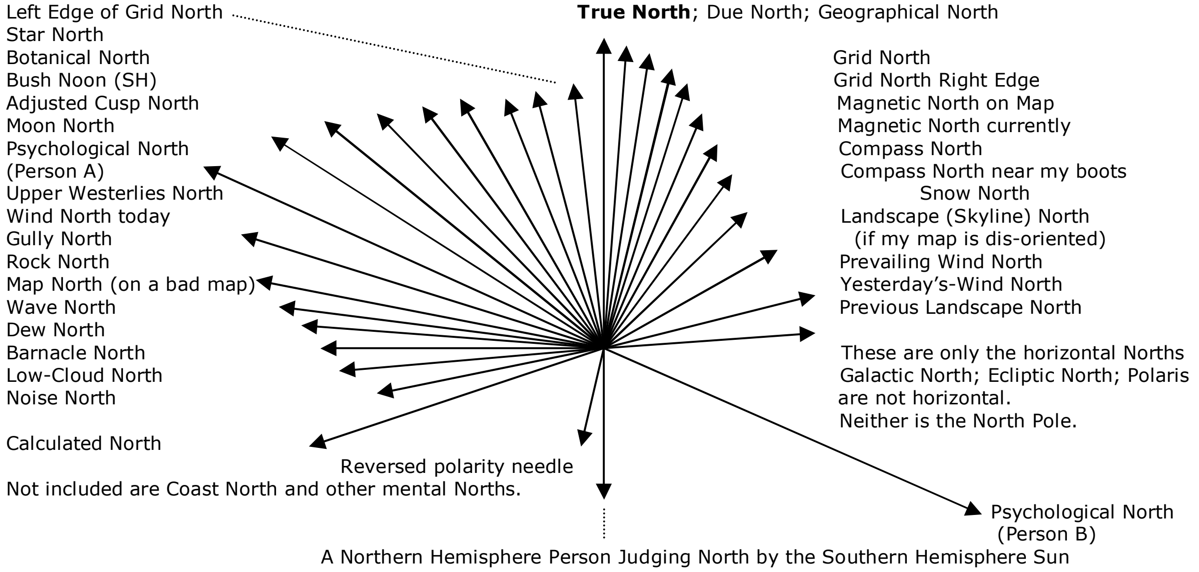
Left Edge of Grid North
Star North
Botanical North
Bush Noon (SH)
Adjusted Cusp North
Moon North
Psychological North
(Person A)
Upper Westerlies North
Wind North today
Gully North
Rock North
Map North (on a bad map)
Wave North
Dew North
Barnacle North
Low-Cloud North
Noise North
Calculated North
Reversed polarity needle
Not included are Coast North and other mental Norths.
True North; Due North; Geographical North
Grid North
Grid North Right Edge
Magnetic North on Map
Magnetic North currently
Compass North
Compass North near my boots
Snow North
Landscape (Skyline) North
(if my map is dis-oriented)
Prevailing Wind North
Yesterday's-Wind North
Previous Landscape North
These are only the horizontal Norths
Galactic North; Ecliptic North; Polaris
are not horizontal.
Neither is the North Pole.
Psychological North
(Person B)
A Northern Hemisphere Person Judging North by the Southern Hemisphere Sun

---

'East-least; West-Best' Rule — Tells You Which Bearing is 'Best'

This rule works only for horizontal directions, measured clockwise.
It is for shifting between True, Grid, Magnetic & Compass bearings 0-360°.
"Least" really means "Less than the other alternative number" **"Best" really means "More than the other choice"**
The following wording ("the North which is") is important, since you need to choose correctly between a right and wrong choice!
***The "North" which is East* (clockwise) of another "North" *gives the Least bearing*.**
Any *"North" to the West* (anticlockwise) of an alternative "North" yields the *Best* & Biggest numbers.
Specifically, it is most used to shift back and forth between magnetic North and true North bearings.
The wording may then be rendered: "Compass error West; compass heading best". Memorise that wording.
Because this mental arithmetic is a difficult ask, you need to draw a diagram to make the conversion clear.
Learn this reminder "Check-up-on-yourself!" diagram… And visualise it, on the real landscape in front of you.

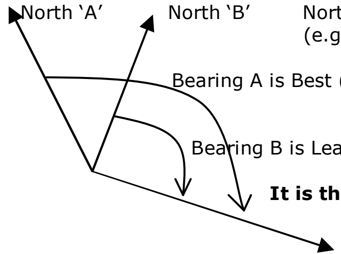
North 'A'
North 'B'
Landmark

North 'A' is **West** of North 'B': it yields the **Best** of the two bearings.
(e.g. if A is true North, and B is magnetic North, the 'magnetic variation' is 'To the East'
and "the compass heading is Least")

Bearing A is Best (It yields a bigger number — East-around)
(e.g. if the magnetic bearing is 085°, the true bearing is bigger… 120°)

Bearing B is Least North 'B' is **East** of North 'A': it gives the **Least** clockwise angle.

It is the *North* which is "East" or "West"; The *bearings from it* are "least" or "best".
Numbers are meant to make sense — your job is to look for that sense, visually.

E.g. Compass Deviation………If my compass *North* points too far East e.g. by 10°, 050°M = 040°Compass (-*bearing* is Least)
E.g. Grid Convergence Correction…If Grid N is 2° tilted to E of true N, 070°T = 068°G (Grid is Least, since Grid N is East)
E.g. Map to Compass………… 090° on the map means: set your compass to 100°M, when magnetic variation is 10° West (Best)
E.g. Compass to Map………… 205°M as a compass bearing = draw 220°T on the map, when magnetic variation is 15° Least
E.g. *Add* the GridtoMagnetic Convergence clockwise *E* angle to magnetic N to get *Grid* N "MAG=**M**agnetic, **A**dd(the GMC)=**G**rid"

String Map

This is a fun-project. As you un-do each knot later, you relive the journey!
A suitable string is builders' twine — long, thin & strong.
An extraordinary number of code variations are possible, provided you can "read" the different knots afterwards!
Keep it simple, for a start.
Example: Start with a slip knot. Half-hitch it every hundred paces, to keep track of distance.
Tie a special knot at any change — to remember it by.
E.g. a strangler hitch around some bark for "thick forest"!; a thumb knot around a feather for "Water Bird Lake"
Chain the slip loop to keep it going. Or thumb knot the loop to lock it all up.

A simple numerical code is: slip knot; + n half-hitches; chain on; m half-hitches; chain on; ongoing tally…
n might mean something like 'direction' by clockface
m could be 'terrain type'

Here are some variations…

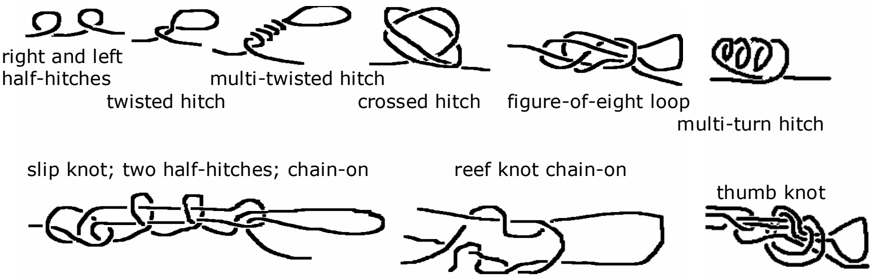
right and left
half-hitches
twisted hitch
multi-twisted hitch
crossed hitch
figure-of-eight loop
multi-turn hitch
slip knot; two half-hitches; chain-on
reef knot chain-on
thumb knot

Progressing A Position Line across the Map

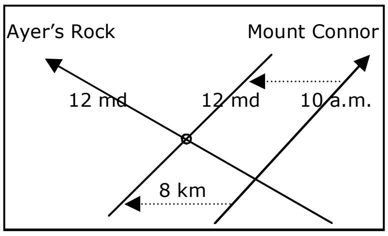

Situation: You catch sight of one landmark; plot a back bearing; then travel — say for two hours, West — before you catch sight of a cross-bearing. You can still plot a 'fix'! Example (fictitious): You lose sight of Mount Connor in the Northeast; and later see Ayer's Rock. Your pace is 4kph. From the bearing of Mount Connor, you plot a position line (by a back bearing running back from Mount Connor). You must label this '10 a.m.' Now **shift the whole position line** 2 hours to the West i.e. 8 km, and label it '12md'. To do this choose any two points on it; shift those points 8 km West; draw **a new line parallel to the old**. You are now ready to plot another back bearing — this time from Ayer's Rock — to cross the advanced position line, for a 'pinpoint fix'. Alternatively, you could 'shift' only Mt Connor on the map, first.

Pinpoint "Fixes" are a Myth!

(Rule: Don't suppress uncertainty, nor reject any unwelcome information!)

Any 'line' of position looks exact. But it comes with unavoidable uncertainty and should be plotted (mentally at least) as a wide band of 'likely position'. Potential inaccuracies abound. E.g. lining up the compass;
reading the compass; plotting that angle;
grid north variation across the map; map distortions;
errors in magnetic variation; protractor inaccuracy;…

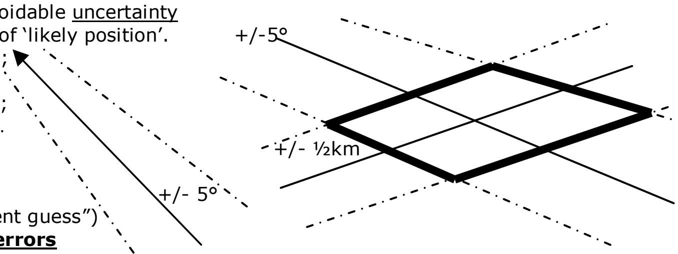

You can now advance your "area of uncertainty" across the map, expanding it with each blind move!
'Deduced Reckoning' navigation at sea (i.e. "intelligent guess") is more a matter of keeping track of **accumulated errors** due to tides, leeway, inaccurate course-holding/recording/plotting, inaccurate instruments. So too on land.

Getting Started in the Dark

The task is to know what direction to head off in, when you can't necessarily see the landscape…
You need a Map & Compass.

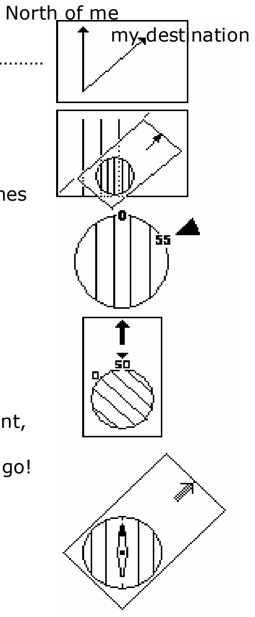

TAKE A MAP-BEARING — Find the desired angle, on the map, in relation to the North-South grid…………..
— "About 60°, as a **guess**"
— Use the edge of the protractor
— Measure the angle — only the protractor is needed for this!
Line up both sets of parallel lines — mapgrid lines, and the lines on the needle-holder-circle.
— The protractor will display the angle between the two sets of its parallel lines "55° E. of N." (It is marked clockwise)
(This is Called Working from **Map to Compass**)

SET THE COMPASS — Using 5° East magnetic variation, and ignoring compass deviation…
Compass bearings should be 'Least'
i.e. 050°C
— adjust the 'Follow Me' Arrow to point to 50° (to the East of North)
Twist the circle on its mounting.
— Don't alter the protractor again!

FOLLOW THE COMPASS — Now turn your attention to the needle
— Line it up with the 'zero' on the protractor, by twisting the whole instrument, And by using the grid lines on the movable circle as a guide.
— Twist yourself this time, to line up with the 'follow me' arrow. And off you go!

Mnemonic: **Map to Compass** (= Map to Compass Course) Means **M**ap Grid to Compass Zero;
Angle the **A**rrow to the course target;
Put on the correction; **To** Get it, use the East Least West Best Rule;
Compare the Compass Zero to the Compass Needle; **Pas**t the Follow Me Arrow is the Course Target.

Thinking Backwards

"Danger bearings" allow you to stay safe, without knowing your position accurately, but you need to think accurately…
As seen from the lighthouse, you must stay on the right of 315° (and to the left of 060°). That much is clear from the chart. Anything 'to the right', is clockwise from the observer, so for the keeper to keep something (e.g. you) 'on the right' means to keep its bearing bigger than ahead. 'Ahead' is parallel to the 315° track shown, since the lighthouse keeper is wishing to see you 'on its right'. He will be happy to see you on any bearing bigger than 315° and less than 060° i.e. 315 to 360, 0 to 60. From your point of view, that means, that you must *use back bearings,* of 135° and 240°, and then keep those tracks on your right and on your left respectively, as you look in. 'Ahead' is parallel again, but opposite to the track shown on the chart, if you are trying to keep the track 'on your right'. Looking in, moving yourself to the left will place the track further to the right, and make bearings to things on it bigger. That means, in numbers, the lighthouse will bear more than 135° and less than 240°, i.e. It is safe to approach at **anything between 135° and 240°**. Notice that the more/less sense is preserved with such use of backbearings.

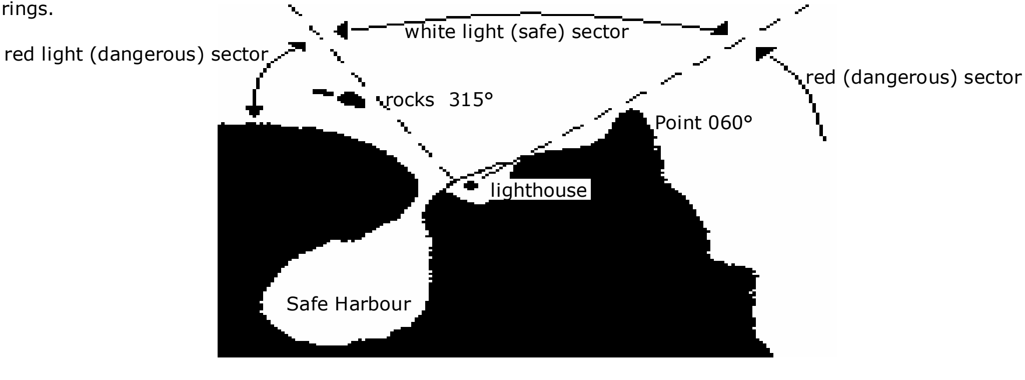

Taking a Bearing from the Landscape

- **Start with a guess** e.g. “That hill is about 060°T from me”
- Point the ‘follow me’ arrow to the landmark
- Grip and twist the movable protractor circle until the grid marks on it — i.e. the zero — line up with the compass needle
- Read the resulting angle off — between the zero (which is “North” —Compass North in this case), and the landmark.

The result should agree with your guess e.g. “The landmark lies at 065°C” i.e. East of Compass North i.e. clockwise from.
Take a B.E.A.Ring Mnemonic: **B**earings are Tracks; **E**xpect to have to plot a back-bearing;
Arrow to Landmark; **Ring** to Needle and **R**ead It off. (Think: “Arrowing” to remember A)

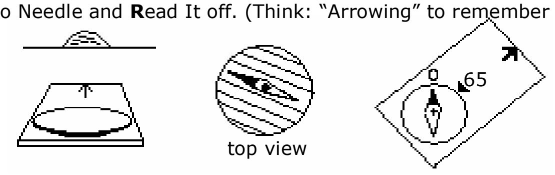

If it helps you sight the landmark, hold your compass string vertically upwards from underneath the tip of the follow me arrow.
The ‘Zero’ is meant to be used for ‘North’! But if you point the follow-me arrow to North and the ‘zero’ to the landmark, then (a) it is more awkward and (b) it is an anti-clockwise “bearing”.

If you are about to use a map, you must now convert the compass bearing to a ‘True bearing’.
Firstly mentally change the compass bearing to a magnetic bearing if you are aware of any compass deviation i.e. compass errors. (You usually do not worry about compass deviations when bushwalking.)
Next mentally convert the magnetic bearing to a true bearing using the East-least, West-best rule.
Example: “6° variation West” (The Magnetic *North* is West of True) “Compass Best” so… 065°C = “059° True”
Lastly, physically change the protractor to read 059° when you want to draw on the map, at an angle to the map’s True North.

See opposite & above

Plotting a (‘Compass-to-Map’) Back-Bearing

Draw a line on the map from the landmark in an opposite direction. Use the protractor to get the angle right.
If you see a landmark North of you, you must be South of it.
Or, what looks ‘South-West of me’ must mean that you are ‘North-East of it’.
Think backwards! See “Back bearings: +/-200° then -/+20°” p.18e E.g. 059°T from me to it…
Note that I have already converted Compass to True. 239°T from it to me.
One bearing yields **one** line of position (from a back-bearing plot — diagram 1)…
Two bearings from two different directions cross for a ‘fix’ — diagram 2…
Three bearings don’t cross. They form a ‘cocked hat’ and keep you humble — diagram 3…
In this case, you could be anywhere nearby (which is also true for any “Fix”).
Expand your possible positions, don’t try to pinpoint a “better” one.
Take more care next time, especially on the closer i.e. the more reliable landmarks.
Check your compass for fixed errors e.g. “always 3° too much”.
Choose directions which are quite different from each other, to get the clearest crossing.
Compass to Map Mnemonic: (Compass B.E.A.Ring, then) **Compass to** Grid (or True) Correction;
Make it a Back Bearing; **A**ngle to Grid; **P**encil to Paper—Draw It in. (Think: “Mangle” to remember A)

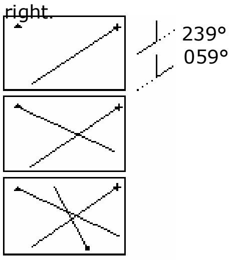

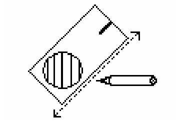

No-Number Back-Bearings

Method 1. Use the white-tipped (South) end of the compass needle, when you take the bearing.
Method 2. Point the “follow me” arrow back toward you, when you take the bearing or follow the compass.
Method 3. Line the protractor zero up with Map South, when you plot the bearing.
Method 4. Point the compass edge to the landmark on the map and draw your back-bearing line backwards from the landmark.
You don’t need to alter the compass setting at all (except for the magnetic variation or grid/magnetic convergence angle).
Warning 1. Don’t mix back-bearing shortcut-methods, or you may get a ‘double negative’!
Warning 2. Check all you do, against reality, for sense.
Warning 3. Don’t try to shortcut Compass-to-True conversions at the same time as shortcutting the back-bearing, or you will get confused.

Parallax While You Walk — JUST DRAW IT!

Most **hand-made maps** start with **a known baseline** scale — a straight segment of a known distance.
I.e. “I walked from here to there in a straight line for 1 km” — for this you may need a pedometer.
Or if you are Captain Cook or Matthew Flinders “We sailed on (this) bearing at (this) speed for (this amount of) time”

You can then **add cross-bearings at any time** to fix the position of distant landmarks — obviously at the beginning and end of the baseline are the preferred places, perhaps combined with one from the middle…
The longer the baseline the more accurate the fix on distant marks.

You really only need a protractor for this **running base-line triangulation**.
(And on a Mudmap, you just eyeball the direction and draw an arrow in freehand.)

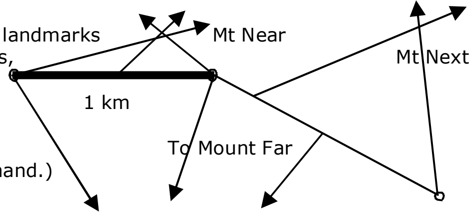

Here is an example of the easiest sort, and the reverse of a double-angle…
After a week at sea, with no sight of the sky to navigate by, I was understandably over-anxious about my position, and “needed” a “fix”! I decided to sail toward an isolated lighthouse-island, in the hope of passing close enough to see it, perhaps, if we weren’t too far off course. The winds were hurricane force, of course, and the sky was filled with salt-spray lifted off from the water’s surface. It was dusk. Suddenly, above the mist, dead ahead, the lighthouse turned on! Alarmingly high, and close. Too close!
Were we “found” or about to be “lost”!? (See “Aiming Off”!) After the cheers: “What do we do? How far off are we?!”
“Turn at right angles to it; Maintain that course; Take regular bearings until it is diagonal; Keep track of the distance run. Then resume your course. That way we can tell how far away it is.”
“How can you do that!” said the bewildered, inexperienced navigator.
I replied: “I’ll **just draw it**”

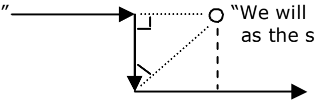

“We will pass as far away as the sideways distance we have run”

A "Start-Stop-Rest" Navigation Mnemonic Overview

S T A R T
"Start" has six points, as illustrated in the diagram below…
These mnemonics, e.g. S.T.A.R.T. or S.T.O.P., will be expanded in the rest of this section—they summarise the book.

- ◇ Start with a **Prayer** — "What more do I need to do to avoid getting dead?"
- ◇ **P.R.E.P.A.R.E.D.** — What have I forgotten to do or to take, to practise or remember? Start off oriented before you go!
- ◇ **S.T.A.R.T.** — with your start-off-trend-direction, which is the single most important navigational item. Start off aligned!
- ◇ Start on **T.I.M.E.** — Read your watch, and benchmark your trip. Say: "Time & Tide" as well, because sometimes tide levels are crucial.
  — Check pace & progress. Estimate the Range, because time is all about distance.
- ◇ Start time is **6 a.m.** — It's always time to **look** in **6** directions, plus 2—Time to Look **A**ll **A**round, **A**ppreciatively, plus time to look at the **M**ap
- ◇ Start with **N.S.E.&W.** —4 sets of 4 directions. The final W is for **W.E.A.T.H.E.R.**

The final R signals that you are **Ready** to go. Don't start until you **R**!

S.T.O.P. A **S**top is a **S**ignal—Interpret the Why? of your stop: Does it **S**ignal a change in watershed, catchment or view;
in **T**rend or **T**errain or '**T**ime & **T**ired';
in **O**rientation (embarrassment—if lost, rule one: Stop);
in **P**rogress/**P**osition/**P**lacement—campsite, junctions, features?

R E S T is best remembered as a mnemonic of 5 RE's…
R.E.-O.R.I.E.N.T. — using these letters as a mnemonic. E.g. **RI** re-Visualise the Invisible Big Picture, and **ENT** Relate yourself to Nature.
RE-RE.L.A.T.E. — re-interpret all directions to relate to each other and to the skyline, as explained below. Now don't forget...
RE-MEMBER — the WAY and the SEQUENCE and especially to NOTE down (in the NOTEBOOK) the NAMES, and DRAW what you need to.
RE-LAX — Play the Navigational Game—teach it—and 5 other games.
There is a final '-art' to 'stop and rest-'
RE-ST-ART 1.1 **re-**pray; 1.2 **re-**prepare; 1.3 **re-**s.t.a.r.t; 1.4 **re-**t.i.m.e.; 1.5 **re-**look (6 a.m.); 1.6 **re-**direct (NSEW), (which includes **re-**weather)
... e.g. 1.3 Re-S.T.A.R.T. after a rest, on the next leg—to relate it to the last & to the *start*-finish trend.

Summary: **"Start"** 6 times**; "Stop, and Re-st-"** (with 4 new **"Re-"**s)

(The Start mnemonic is a good example of the "clumping" technique
—never more than a handful of things to remember, yet it leads you into hundreds of points.)
Put the S for Sky at the top, where it should be.
Use a hexagon for the 6 START up points.
Use a clock-face for TIME,
and the compass-rose for NSEW.
Show the 6 directions in 3 dimensions.
Balance it all on a triangle
—for the "6 a.m. start" , since it has 3 parts.

Sky and Skyline
Toward what Target; through what Terrain
Away in what Anti-direction
Rain catchment; Region; Ranges
Trends

See the boxes to follow

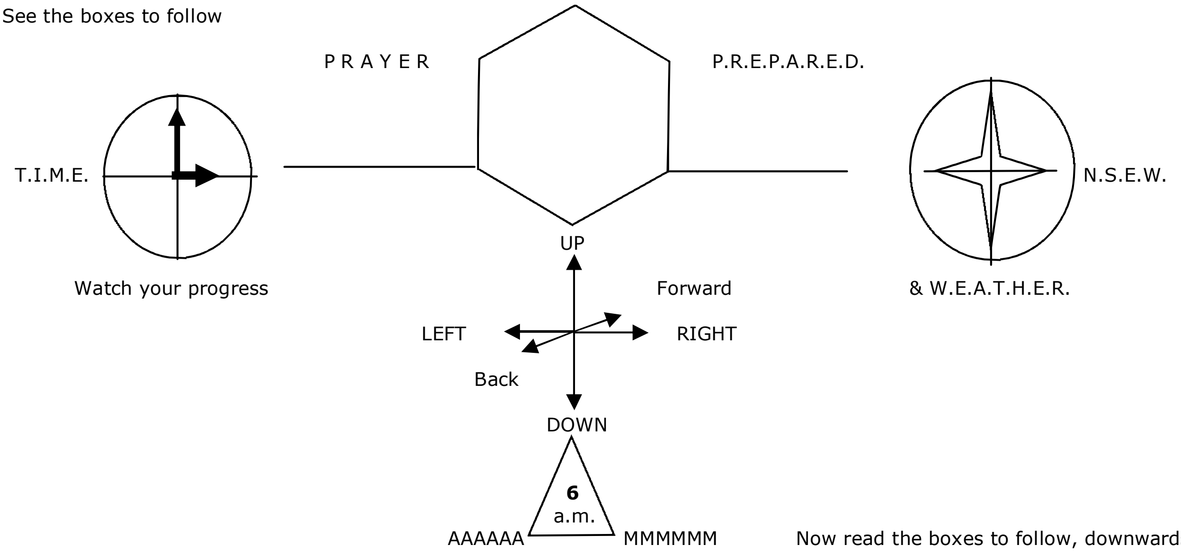

Now read the boxes to follow, downwards...

P.R.E.P.A.R.E.D.—A Before-You-Go Checklist ..so that most of your navigating is already done, from home!

P.ractise Global orientation. Get a blow-up globe. Teach others. Learn (what to learn, look for, do etc.). E.g. predict the Sun.
Learn the Mnemonics, Codes & standards e.g. S&R signals, compass directions
Pace your paces — metres per minute; steps per 10 metres. Left-right bias (blindfold).
R.esearch History. Ruins. Artifacts. Mines. Geology. Get a geological map from the library. Landform. Soils. Climate. Windy season. Prevailing winds. Wet season. Vegetation analysis. Rare species. Biogeography. Geography — global neighbours; time zones; local coasts & river mouths. Catchments. Cardinal features — places N,S,E,W from you. Land lines on map. Place names — their meanings and history.
E.quipment See the next box.
P.lace names See the box after that.
A.stronomical orientation Sleep on a map, and with a compass nearby. Point the map North. Times and directions and heights of Sunrise & set, Moonrise & set. Southern Cross. Pot. Sun at noon, at magnetic North/South, when East, when West, at magnetic East and West. What landscape markers (from the map) for these directions? The star opposite the Sun. Time zone. Local noon — its exact time. The Sun's coordinate predictions. Your longitude. The "Equation of Time". Moon-phase, -number and –day of week. Tide times.
R.eminder Note your intentions—and tell someone where you are going and when you expect to be back.
E.xpectations of the Weather Barometer reading. Get the Weather forecast. Follow the Changes. Buys-Ballot's law. Steering winds.
D.irections & Familiarity Ask someone who has been there. Read track guides. Look at an aerial photograph. Go beforehand! Find your old track notes. Look closely on the way in. Choose a route. Plan your strategies. Coordinate the cars, the car keys, any food drops, shuttles or codes

Equipment — A What-You-Might-Like-to-Take Checklist

(E. in PR**E**PARED, for "Equipment"…)

- **A Map** **A compass** **A clock**
- A notebook and pen — for track notes. A track guide. An aerial photograph. A navigator's handbook.
- For map making — graph paper, paper, pencil, a compass-holder. A map-holder (string around neck to a…), clipboard, stopwatch.
- Polaroid glasses — to orient by the dark blue sky band. Radio — for RDF, weather. Mobile phone. GPS. Camera.
- Binoculars. Pedometer. Rangefinder. Calculator. Altimeter. A sine-cosine table.
- A map cover. A spare map. A navigational protractor. A magnetised needle. A sundial compass. A two-flap cardboard compass.
- Marker tape; Chalk; Scissors (for cutting leaves!). A clear plastic tube. A clear plastic drinking bottle (for an artificial horizon).
- A long piece of string (e.g. for a range-finder). Short pieces of string (e.g. for a plumb-bob). A walking stick. Someone who knows the way.

(P in PRE**P**ARED) **P**lace-Names —Learn Them—You Navigate by the Local Features!

"Look. That's Big Split Rock and that's Little Split Rock. Wild Dog Tier is that range over there, and this plain drains out over there, into the Ouse River You can find your way anywhere around here from those."

Suddenly I was unforgettably oriented, just like he said he was, when his dad told him.
For me, the funny names on the map now meant something worth knowing.

Place-names are not just names you don't know; they name places, which you ought to know.

More accurately, place names 'place' you in the landscape, as well as 'orient' you to it.
A sense of location is almost independent from your sense of direction, and is rarer to come by.
Direction-sense places you in relation to very distant background references i.e. it doesn't 'place' you at all!
Location-sense places you in relation to the immediate foreground — visible landmarks, catchment basins, fixed 'tracks', etc.
Place Names Place You

S.T.A.R.T. off Oriented

— **S**tart-off Direction? (e.g. "upstream") — the **S**ingle most important navigational knowledge — it tells you the return path, and the search path.
Sky/**S**kyline i.e. especially according to the **S**ky (e.g. "toward the Moon; into the sunset") and related to the **S**kyline.
Start-off directions for the parallax method, pp. 51a,53a, e.g. to magnetic NSEW skyline or recognisable middle-ground points, or rhumb-line point and another set of such bearings, taken from the map, as seen from the finish, if you are on a one-way trip.
— **T**oward what **T**arget? Specify a
Track (e.g. "downhill to the lake") or a **T**rend. A rhumb-line bearing is best—from start to finish.
Through what **T**errain? (e.g. especially see what the land-shapes would look like without vegetation)
— Away from What Point? (e.g. "from the lookout carpark), and in what
Anti-direction (e.g. toward that other peak"—a skyline direction)
— **R**ain Catchment — In what major
Region? — as can be gauged by the surrounding
Ranges. (e.g. "on the main spur dividing North from South arm, downstream of…")
— Trend: **R**ead the **T**rend off the map, for starters; then continue to **R**ead the **T**rends & bends and the progress as you go on.
Read (your compass; map; watch) in **R**elation to the **T**rend, i.e. in order to follow the trend, which is the focus of this mnemonic.
Plus: **R**ead any other **T**rends, from any other places, from memory, in relation to the sky. Be able to revisualise them at will.
— Re-s.t.a.r.t. on the next leg, to relate it to the start-finish trend, and to relate it to the previous leg.

T.I.M.E. & Distance — that's the point

TIME to Look at your Watch — PACE YOUR PROGRESS — the point of this T.I.M.E. mnemonic is distance.
— "Watch" your steps—i.e. calculate your pace from the elapsed time;
— "Watch" the Sun without looking at it, because your watch tells you where it should be.
— TIDE times may be necessary.

Instrument Check — watch, stopwatch, pedometer, odometer, rangefinder — for progress
— compass: Double the bow angle, or halve the stern angle p.94c; parallax readings on landmarks.
— (and other instruments — barometer, altimeter, GPS, radio null, radar, depth-sounder)

Map-Scale Check — are you judging the scale well? — What is the next goal? (This M is not for Map in general, but only for distances.)

Estimate the Time of **E**xpected Arrival — guess each milestone ahead of time to set you free from a nervous TIME. Nominate an E.T.A.

Distance

PROGRESS: can also be gauged by lines of sight and changeover lines. Use a stopwatch, or pedometer. Read the scale.

PACE: distance divided by time; 20min/km, double and halve = 3kph; 1mps = 3.6kph; 4kph normal walking pace;
finger-math helps; pace yourself evenly.

RANGE: Use the letters of RANGE & FACULTIES as a mnemonic to remember the different ways of gauging distance in the distance...
Range squared = dee height; Area divided by offset; New direction parallax; Grass-leaf vertical angle, horizontal angles too; Educated guess
Familiar lengths, and Fingernail Rule; Acuity of vision and Apparent sizes; Cloud behaviour; Upper and lower guesses, averaged; Landscape lengths;
Tom Thumb Parallax; Intervening detail & haze alters your perception of distance; Extreme range comes in two bites; Sound travels at 1 km/3secs and
Sidewise offsets with blinking.

150km from 1 nautical mile high; 1° latitude from 1km; 50 miles from 500m; 50km from 1/5 km; 5km from 1/500 km (2 m); 12.7km from 12.7m.

Look in These 6 Directions — in 6 ways — Up and Down (Part of "Start at **6** a.m.")

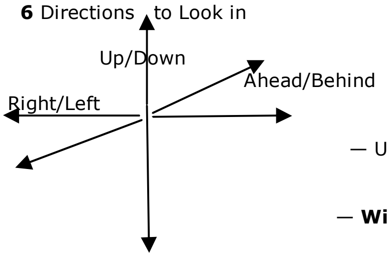

UP — **Astronomy** — Sun Moon Calendar-Stars
— **Sky** — Clouds, Contrails — Blue Band
— What's up with the Weather?
then DOWN —Geology — **Soil** Type; Rock Type
—Clues — Animal **Trails**; Bootprints; Fallen markers

— Up-**Sun** — haze & glare
— Over what tropical country? Then — **Down**-Sun — haze, shadows, colour changes

— **Wind**-wise — Up-Wind — Where on the skyline? — Down-Wind — Around what obstacle?
— Around what pressure system? (Wind in your face)
— From what coast/ocean? Listen; Smell
— Up-Breeze — To what cold surface? — Down-Breeze — To what warm surface?
— What's up-Weather? Where are the steering winds coming from?

— **Land**-wise
— Up-Hill — To what dominant peak?
— Point to the dominant peak, even when it is invisible,
— Down-Hill — antidirection?
turn around to see what lies away from the peak — the "anti-peak",
especially in relation to your shadow — this anti-direction has a lot to
do with a compass back-bearing—a visible line of position, you see
— Up-Ridge — To what hill/hillock? — Down-Ridge — Down which ridge? — To what headland?
— what offshore/underwater topography e.g. shallow reefs; islands?
Since this mnemonic is about direction, try to point to them.
— Up-Slope — To what ridge? — Down-Slope — Into what valley? Name them.
At what angle across the gradient?
— Where is the lowest upslope skyline, under the canopy? — Where is the highest downslope skyline, under the canopy?
— Up-**Stream** — To what distant divide? — **Down**-Stream — What stream? To what distant mouth?
— to what sedimented shelf; or deep-water channel?

Summary: **Above & Below** (twice each) + **Sun & Wind** (above eye-level) + **Land & Water** (below eye-level) = 6 ways for "Up & Down"

---

4 More Slopes

1 in 4, is 14°;

10° (aim off, 1 in 6) is 175 m per km;

1 in 60, is 1° (1 in 57.3 is the exact figure, but it doesn't multiply up linearly after a few degrees, whereas 1 in 60 does well after a few degrees)
& n° is n in 60
(1 in 6 is 10°; 1 in 10 is 6°);
The equivalent is 5/3 in 100, 5 in 300, 167 in 1000, for use with multiplying up with the isosceles triangle angle method.

36°, 8 in 11, is critically unstable.

Multiply them up, inversely.

Look in These 6 Directions — in 6 ways — Ahead and Behind

Look Back Regularly: Is bad weather sneaking up behind you? (p.46c) "Abracadabara Peek behind you" (p.63b)
How much has the track bent around behind you? Any acute track junctions? Can you remember this way back?
Compare Forwards with Backwards: Do you need a back bearing? Are you half-way? Is your starting point visible, to get an anti-track forwards?
Are there any changes in the terrain, to photograph, in your mind? Have you memorised the transition, from the far side?
Look forward — in **time** — the forecast! What is the Estimated Time of Arrival? Look Back — in time — How far have I come?
Look ahead and behind in **U**pper case **Time** — Could you recognise this place much later, in hindsight? — What is the history of this place?

Look AHEAD — Has the **view** changed — into a new catchment? Look BEHIND — Have you left the old view behind — crossed a divide?
— at the terrain, the track showing up ahead, and the difficulties — Do you remember the **way**? Could you recognise it again?
— Look for track **markers**, and for trails ahead — Look behind for hidden, reverse track markers.
— Line up two (new) marks exactly ahead to keep your direction sense. — Look for the old lined up mark left behind.
Is it time to choose another leg, and a **pair of marks** lined up to it? Do I need to find new ones behind me?
Mnemonic: R,S,T-U,V,W,X-Y-Z **R**egularly **R**everse your nose. **S**tart/anti**S**tart bearing & **S**napshot-**S**equence your transitions
time (+**U**pper case Time); The **V**iew; **W**ay (remember it?); **XYZ: X** marks the Track (Markers); the (**Y**&**Z**) Leapfrog Marks =6 ways for Ahead & Behind.

Look in These 6 Directions — in 6 ways — Right and Left

Look RIGHT Through! 1. — the bushes, to the land-form underneath; to the contours & bedrock
2. — the canopy, to the skylight beneath it
3. — the trees, to the most distant marks, including other trees
4. — the treetops, to the sky — the Moon, clouds and the blue band
5. — the foreground, to the most distant skyline horizon, whenever you can glimpse it
6. — the haze, to the skyline details beyond the land horizon—use binoculars too.
E.g. to barely visible mountain peaks, or a second skyline & its parallax, or to islands in the haze, or the rising Moon.
Look LEFT 1. — to the most distant thing you have LEFT BEHIND on each side! i.e. Have you left it behind yet? Is it time to choose another one?
2. — Have you left behind your: handbook; air photos; track guide; track notebook? — No, well, don't just carry them — Look at them!
3. — Have you left anything behind at the rest stop? It's an important habit, to ask this each time you move.
4. — Look to the right and left, for the contours, the natural 'tracks', and the behaviour of the network you are crossing and leaving behind.
5. — Things on the right have big bearings; things on the left are lesser; than straight ahead, even less as you leave them behind.
6. — NORWest & NOTLEast — North needle 'on the left', means you are heading roughly East. (N. On your Right = West-ish)
Summary: **Right Through & Left Behind**.
In summary for all the 6 directions make sure you **Check Each One Out—Thoroughly!** (6 ways each)

..Plus in 2 More Directions — (1.) All, Around, Appreciatively, And At 6 L's

1. **A**ll your senses: Listen, Smell. Look around with your other senses.
2. Look 360° **A**round for the best & worst showings of colour, shine, flowers, greenery, new growth, fern fronds, lichens, trunks, branches, dirt. Compare Ahead with Behind, & Left with Right, & upSun with down-Sun, & up-wind, uphill, etc. for direction-dependent light and colour effects.
3. **A**PPRECIATE — the scenery; the colours; the contrasts; the shapes, outlines and silhouettes; the foreground movements.
4. **A**n Environmental Notice Time (see later box E.N.T. in REORIENT)
5. **A**ny LOCAL PECULIARITIES to remember.
6. **A**nd **A**t — **the six L's...** — **Land**marks — distances, sizes (apparent Largeness), altitudes Use binoculars.
   - — Are any familiar, from other walks you have done elsewhere?
   - — Do the place names make any sense? — interpret them!
   - — **Lay**out — relative angles, compass bearings. Am I inside or outside of major landmark circles/triangles?
   - — **Lie** — slope and terrain, catchment and drainage, contours and altitude
   - — **Lines** Catchment Dividing Ridges and Rivers; Contours at right angles to the slope; Flight Tracks
   - — **Ligns** — 'Same direction' Lines of Sight, or nearly so; and 'Opposite directions' alignments: Am I in between?
     - — the fixed A-lignments . Am I left or right of major 'ligns'?
   - — **Sky-Line** — talk it through; draw it. Cross-relate all directions to each other, and to the skyline

Summary: 'A' for 'a.m.': 6 A's, with 6 L's

..(2.) Look at **the Map**: a Systematic Checklist. (What 'System'??)

The system is to examine **top-down**. From general to specific detail. From large-scale to small-scale. Look for the downward progression in these…
3-D location: Latitude, Longitude, Elevation. Centre of the time-zone. The inset map showing where you are on the continent. Scale & size. The overall width- and breadth- distance of the map coverage. Grid size e.g. 10 km squares. Where you are on the map, geometrically too. The Starting Point. The Relative Location—over the horizon landmarks; coastal towns N.S.E.W. of you; something NE,NW,SE,SW of you. Radio stations. Highways. Light sources. Powerlines. Pipelines. Place names. Visible Landmarks. False peaks. Lines of the land — between landmarks. Recognisable hill-hill reference distances. Recognisable elevations e.g. towers. Grid North; True North; Magnetic North; Magnetic Variation; Yearly Drift in Variation.
Catchments & Divides; Catchment Size; Adjacent Catchments; River Drainage. Land shapes i.e. interpret the slopes, contours, the contour interval.
Trends and Bends — of the coast; rivers; roads; ridges; ranges; the track. Notice any creeks which flow backwards i.e. against the trend, and why.
What rising ground will be visible in each direction? Perimeter features surrounding your walk e.g. a river, a range, a road, the coast, a forest. Escape routes — unplanned exits. What are the ridge-line limits of the sub-catchment which most encloses your possible location? What high points are on it? Possible routes. Clockwise or anticlockwise? Distances. Overall distance of your route. Milestone features. " Wake-Up " landmarks.
Initial bearings to landmarks. Initial angles to landmarks. Initial horizontal angles to landmarks. Final bearings. Rhumb–line trend from start to finish Around what Obstacles? Fine details of the route. Don't forget the fine print — the date and reliability of all the map details; the grid datum for GPS use; the legend of symbols; colour codes; contour line thicknesses.

Mark the Lines-of-the-Land On Your **Map**

6 M's for Map:
Map Checklist (above); Map-Lines (this box); (all below...) Map Alignment; East-Least West-Best Mnemonic; Map & Compass; Map-Making Methods.

Pre-prepare your map — **Mark in the navigational lines** which the mapmaker has left out. Most of these are permanent.

A. Cardinal Tracks — NSEW lines. Across the whole map, from the very prominent peaks; or simply covering a local area, from the local peaks
B. Alignments — e.g. from Peak to Peak; Tower to Town; Point to Point — You may see two in line, or walk in between two landmarks.
C. Directions to major distant features — (for orientation purposes) — e.g. capital cities; coastal towns
D. The Overall 'Track' — the RhumbLine — from start to finish. Mark this lightly; &/or two or three obvious overall 'legs'
E. Circles of Distance Off — e.g. 5, 10, 15 km from Mt. X.
F. Circles of 90° horizontal angle — i.e. make two landmarks the diameter of a circle
G. Circumscribed circles around (i.e. which just include) the triangles in the triangular grid of landmarks
H. Transmitter to Transmitter networks — i.e. a triangular network of major peaks; include the distances and bearings on each line.
I. Flight Paths — if you can get hold of an aeronautical chart — it forms a second triangular network.
J. Catchment divisions.

Mnemonic: Four Straight Lines; Three Circles; Two Sets of Triangles; And One Very Sinuous Set.

Mnemonics for **Map** and Compass Work

◦ Map Alignment

A.L.I.G.N. the **M.A.P:**
A… Align the map to Sun with an imagined **A**rrow, and to your course with **A**nother **A**rrow;
L… to the **L**andscape;
I… To **I**dentify Landmarks and to get a Fix;
G… **G**rip the Magnet to the Map, and the Map to the Universe;
N… **N**ot "Swinging the Map" As You Turn;
Magnetic Needle to **M**agnetic North (in four variations, "0°" to North or arrow to course, Magnetic or True);
Angles Horizontal, fit over the map;
P… Align the **P**ath like landing a **P**lane, or align the map by aligning the **P**ath.

◦ Mnemonic East least, West best

East is Least and West is Best, tells you the North which gives the bearing that is best. i.e. the biggest in number; or lesser in number.

◦ Map & Compass

LMNOrP: Landscape; **M**ap; **N**eedle; **P**rotractor; either one **O**r two at a time, but not all three.

M.A.P. To Com.Pass, = Map to Compass Course, Means…
Map Grid to Compass Zero; **A**ngle the **A**rrow to the course target; **P**ut on the correction;
To Get it, use the East Least West Best Rule;
Compare the Compass Zero to the Compass Needle; **Pas**t the Follow Me Arrow is the Course Target.

Take a B.E.A.Ring: Bearings are Tracks; (**E**xpect to have to plot a back-bearing); **A**rrow to Landmark; **R**ing to Needle and **R**ead it off.

Compass to M.A.P.; Compass bearing, then **Compass**-**to**-True Correction; **M**ake it a Back Bearing; **A**ngle to Grid; **P**encil to Paper—Draw it in

◦ MUDMAP:

Mudmap on paper; **M**ark the start point on one edge; **U**p-arrow for orientation (**U**pdate it); by a Reference **D**irection (Up**D**ate it);
Maintain a trend, a long straight leg; **A**lign the **A**rrow when you stop; **P**encil in the **P**rogress, and a new leg direction.

◦ Other **M**apmaking **M**ethods — Use the letters of COMPASSES:

Copy a Map Top-Down; Own Map is Valued; Mental Map—Map to Mind Navigation; Parallax onto a Running Baseline; Accumulate the Layout Blind; Sine-Cosine Tally of Accumulated Eastings and Northings; String Map; Exploring a Cave System Needs a Book, of Strip Maps.

“N.S.E.& W.” — the four cardinal directions, is a mnemonic of fours

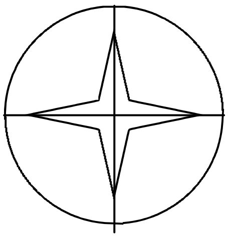

N for “**N**orth Coast” — all about visualising the invisible directions…

Name which coastal towns are N. S. E. W.
“**S**weetheart” features Name your favourite spots NSEW of you
Encircling horizon features — **Horizon** North; **Skyline** South; **Environmental** East; **Landscape** West.
Which-way-am-I-going trend (rhumb-line), from the map, then marked onto the terrain—e.g. to behind that hillock

S for “**S**hortest **Shadow S**outh” — all about the path of the **Sun** (and Moon)—the 1234 system
— and the **S**unny **S**kies quadrant system
— How the Sun, Blue Band trend & ends, and your Shadow, currently lie in relation to…

N Highest **Noon** (How High?)
S Shortest-**Shadow** Souths (SH) & the two-flaps South. Coldest South is the best reference in the SH. String South using the Sun’s Shadow.
E Sunrise (Where?) Anti-sunrise from a string-line to the horizon.
W Sunset (When?); Bisect them. Anti-sunset from a string-line in the afternoon.

E for **E**quinoctial **E**ast—i.e. other things astronomical, **E**tc.

Needle-North-Seeking-Northern-Canada—over Red Hot tropics!
South Pole Star (How High?) & Southern Cross (Where?) & Star-Chart octants & Star Compasses & Star Fixes
Ecliptic & Equator & Equinoctial Precession
W for the motion of the Pot, s**e**tting in the W**e**st (rhyme), at the latitude angle

W for **W**ind **W**est, i.e. Natural Directions.

N for Nature’s Norths: Rock Noon; Botanical Noon; Solar Panel Noon; Satellite Dish North
S for Snow South; Frost South-East; Steep Gully South; Mud-in-the-Shadow South; Dead South; Siesta South.
E for Moss East-South-East
Wind: Upwind: scorch, erosion & flats, Downwind: shelters, debris & steeps; Noise & colour from up-wind; Rumble & skyshine from downwind.
& **W**eather Directions (& see below)—Where is Wind West? What is Upwind? Upper Winds Steer the Weather (from the West?) What is Upweather?

Summary: 4 directions with 4 variations; Then finish with **W**.E.A.T.H.E.R. After a rest, re-direct yourself.

W.E.A.T.H.E.R. — Check for Change

W. for Wind. Has it changed — in strength, direction, gustiness?
for Wind — in Your Face — Low on Your Left. Low winds bring the rain, clockwise.
for “Weather systems come from the West” — in temperate latitudes — can you make a prediction?
for the upper level steering Westerly-Wave Winds — are they deviating much from average? Are they steering a storm toward you?
for “Wind-West” showing up in the vegetation, for when the prevailing wind is not blowing.
for the prevailing Westerlies (or whatever).
E. for Ephemerals — i.e. the clouds and contrails (don’t forget this E.!)
Have they changed? — in height, type, wateriness, thickness, darkness — in directions — steering wind directions? In relation to the Sun?
— Examine them carefully for dissipation (i.e. thinning, soft edges, layering) or for building up (hard edges and heaping).
A. for Atmospheric Pressure Change — imagine the isobars across the land from the forecast.
Sometimes you can see a lowering atmospheric pressure when marsh gas starts to bubble in cold weather i.e. not due to warming.
T. for Temperature — of the air-mass
H. for Humidity — the absolute one (measure the dew point — will it frost?). For the relative humidity try huffing to produce a fog, or onto metal.
Does the air mass humidity & temperature, speed & turbulence let you know which direction the wind must be from?
E. for Expectations — i.e. the weather forecast — test it — is it coming true?
R. for Rain — Will it? R.ight? Are you **R.eady** to move? Grab your R.aincoat and let’s go… Keep looking until it is time to S.T.O.P. for a rest…

Oriented? No?
Rule one: Stop! getting more lost.
Rule two: Start! getting unlost. Start navigating!
Rule three: Stay, or Steer a Straight course.
Rule four: Don’t do it again!

Position/**P**lacement/**P**rogress—has it changed?
Have you arrived somewhere?
Is it a named feature? A campsite? A junction?

Have you changed view?
Have you changed catchment?
Interpret the Why of your stop...

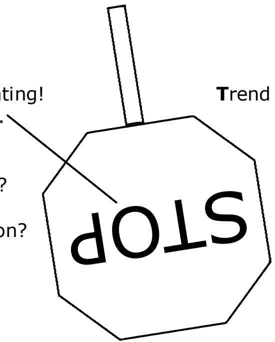

Trend And **T**errain—Has it changed?
Time & Tired—Does that explain your stop?

Stop -Signal: Why have you stopped?

REST and R.E.O.R.I.E.N.T yourself

Take 5; a 5 minute rest; REST is a mnemonic of 5’s.
RE-RELATE; **RE-ORIENT; RE-member**; **RE-lax;** (= the 4 new RE-’s before you) **RE-start**. = 5 RE’s

RE-RELATE to the skyline.
RE-RE for Re-Read the map,
L for Re-Late the Map to the Landscape
A for Astronomy—relate the Sky to the Skyline
T for Trends—imagine the Trend of invisible landscapes, in relation to the Sun, and relate your current trend to a point on the skyline
E for Each direction clue Re-lated to Each other direction clue, especially to the landscape/skyline shapes.
RE.O.RI.ENT — **RE** for Re-Orient the Map to the landscape,
— **O.** for **O**ops! Don’t make logical blunders; Have questions on your mind; notice; speak up; integrate; connect; interpret.
Observe the **Overall** trends and bends.
— **R.I.** for **R**e-Visualise the **I**nvisible (Next checklist; and see Section 1 Astronomical Overview; Include the clock-sense mnemonic here)
Re-late yourself to the Big Picture
— **E.N.T.** for an **E**nvironmental **N**otice **T**ime — Re-late yourself to Nature — coming up.

R.I. For Revisualise the Invisible (and then Relate It to the skyline!)

Work down from the Biggest Picture. Work most on the ones you can't yet visualise.

Astronomy — Visualise the paths of Sun/Pot/Moon/Stars; Point to the South Pole Star; North Pole Star; the **Clock-sense** of stars around each Pole
— The star opposite the Sun during the Day & during the night; the Southern Cross during the day; the Ecliptic; the Milky Way
— Earth rotation; Horizon rotation; horizon trails; clock-sense of Horizon rotation; **clock-sense** of low stars' movement around horizon
— the Moon-Sun, and Ecliptic line; The Moon phase explaining the present tidal force. See the overview at the end of section 1.
Global — Latitude; Polar distances; Equators; Earth Poles; Seasonal Sun height; Shadow South — visualise the aurorae; Antarctica;
— North Magnetic Pole in Canada, as a 'South-seeking' Pole; Aurorae; Unequal Tidal bulges. "Go Overseas" p.25e.
— Tropical countries — corresponding to sunrise, noon, noon height, sunset, the country under the Sun now, the time zones
— Light & Dark halves — Sunrise-/sunset-circle countries; Half the arc-distance; How the horizon is shifting due to your movement.
Continental — Weather map isobars—visualise them across the continent — the one in front of your eyes, and on a continental aerial photo.
— Latitudinal wind system and pressure system helices.
— North Coast; N-S-E-W townships; NW-NE-SW-SE features. "Walk across the Continent" p.29e. "Compass-needle Continent" p.9e.
Local — Catchments; Divides; Mouths; Mounts; Radio Transmitters; Night Lights
— Nearby landmarks & places & highways; Home; Car
— Perimeter features; Escape targets & Routes — Bird's-eye layout of your route.

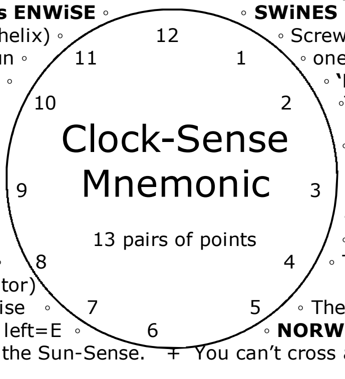

0 The standard viewpoint is "seen from above"; ◦ Clockwise is to your right; Compass Bearings Circle Clockwise;
1 Wrong Way Go Back! **Anti clockwise is ENWiSE** ◦ ◦ **SWiNES** That's how the Sun moves in the northern hemisphere
2 Unscrewing is an anticlockwise motion (on a right-hand helix) ◦ ◦ Screwing in is a 'right-hand' motion (for a right-hand screw)
3 24 hour clockfaces have only 15**°** per hour, like the Sun ◦ ◦ one hour = **30° clockwise**, on the 12 hr clock face method
4 High pressure systems circle **anti**clockwise in the SH ◦ ◦ **'Low' winds** come clockwise, in the SH.
5 Things on the **Left** have a **Lesser Bearing** ◦ ◦Things on the **Right** have a **Bigger Bearing**
6 All the solids rotate **anti-clockwise,** seen from the **N.** ◦ ◦ **Low stars** all around the horizon follow the Sun-sense
The ground under your feet shifts anti-clockwise in NH ◦ ◦ The ground under your feet turns clockwise (SH)
7a&b So the ho**RI**zon **N**ever **H**eads **RI**ght in the **NH** ◦ ◦ The ho**RI**zon **SH**ifts **RI**ght in the **SH**
8 Wind-shift due to ground friction opposes the Sun-sense ◦ ◦ Coriolis forces follow the Sun sense
9 **North lies between the SH Sun and the hour hand** ◦ ◦ The NH Sun is between the hour-hand & South
(actual noon points to Sun; clockface tilted parallel to Equator) (noon-time to Sun; axis to Polaris)
10 The SH Sun seems to move anticlockwise ◦ ◦ The Moon goes backwards, night by night (clockwise, in SH)
11 Nose Nav'n turns everything away **NOTLEast** Nth on the left=E ◦ ◦ **NORWest** — Needle-North on the right = heading West
12 **Shadows** Follow the Sun-Sense. + You can't cross a stream twice in the same flow-sense

---

E.N.T. for Environmental Notice Time —Orient yourself to Nature

Take **a top-down approach**...
Geology; geo-trends; geo-morphs i.e. landforms; geo-Norths
Dividing ridges; dominant peak; upstream catchment size; wind shadow; rain shadow
Season; climate type; prevailing wind; average rainfall
Rocks; bedrock; outcrops; parent rocks e.g. upstream
Soil types; nutrient availability
Soil Drainage — good or poor; Water Sources; groundwater? salinity
Aspect i.e. which direction does the slope face, in relation to high noon?
Living stresses? — salt, minerals, wind, exposure, frost, heat, low humidity, fire, flood, wet roots, pests, diseases, competition, grazing, trampling, disturbance, rock instability.
Vegetation type; the overall density of green, and of cellulose, and the height; dominant species; associations of plant types; diversity; individual species; individual plants. Pollinators available; seed dispersal mechanisms operating.
Mammals; Birds; Reptiles; Insects. Drinking water availability, cover for animals, hunting pressure
The 'Natural Norths' — Gully North; rock North; Bush North; insect North; Dew North; snow North; ripple North; prevailing wind North

---

RE-MEMBER to Note down the Names in the Notebook

Remember the **Way**, and the **Sequence**.
Remember to take a Notebook. Make a neat permanent copy of your notes i.e. in ink, maybe sometime later on, in your master track notebook.
Be Bit-wise!
Name the subsections, junctions & links. Name the peculiarities. In the name, try to include…
The Why, When & Where of your stop. Link & Sequence information — time, distance, context
Lie of the Land Words to indicate the direction trend, & curves — left or right
Visual words — colour, shape, alignments, curves Descriptive words. Collect the details. Enlist the others to describe and name it.
Give yourself something to remember — create an acrostic; craft a creative name; sit down and write up your track notes
Update any acrostic alliteration. And the count e.g. of the creek crossings. Do something numerical. Do something alphabetical.
Be vivid, clever, memorable. Do something Shapely or Geometrical there Do something Memorable there. Do something there.
Draw the shapes, outlines & skylines. Draw the directions of other landmarks from any high viewpoint. Take reference photographs.

Remember most of all to do it all — don't just recite the reminders to do it all!
Mnemonics of Fives: Remember the Way; the Sequence; Make Notes; Invent Names; Draw What You Need to.

Re-lax, and Play — Navigation Games and Activities (5 pairs to recall)

Walk & Talk — a night walk — a fog walk — a walk off the track — climb a hill with a small scale map & binoculars & compass.
"Track My Trail" ("If you want lunch, you'll have to find me — give me a quarter of an hour, then track me down").
Keep a compass course, blind — keep a straight course, blind — on a slope? "Let's all make an acrostic to remember this track by".
Try Orienteering — Try making a map of a maze of trails — then a string map — a strip map — or a sine/cosine tally map.
Look & Learn — "Go to the front of the line if you spot a track marker which the others have missed" In a windy place look for wind clues, etc.
Take it in turns to name a navigational or orientation clue which is visible, until you have to drop out.
"I notice… (fill in the blank). Your turn. What extra can you add (to the description, or to the clues)?"
Guess & Test — Each of you estimate the time of arrival; the direction to the car; the distance to a landmark — then check the map.
Guess the time from the Sun, Moon, Southern Cross, or other stars, then check your watch. Use your compass to check your guesstimation of North from the bush, gullies, rocks, a star-trail, a star-set etc. Predict the tide, or the time & direction of moonrise — then sit down and wait for it. Predict a star-set position. Find your 'shortest shadow' angle today & use it tomorrow.
Make & Use — an odometer from a wheel, a pedometer, a sundial/declination compass, a portable sundial, a string rangefinder, a slope-meter.
Measure — Try some star navigation observations — Rise & set bisection; Use a plumb-line; a bush quadrant; a bush sextant.
Test how accurately you can each use a compass. Test each compass against the others.
& Memorise — yesterday's acrostic. Remember all the mnemonics! Can you add to any of my checklists? Have a star identification evening.

Moon-Phase & Tide-Time Dial

This is a handy tide-predictor, based on the assumption that tides will follow the moon. It shows you where the moon will be, at high tide. It also tells you what the moon is doing if you know the date but haven't seen the moon for a while. The deluxe version can predict tides and phases years ahead of time. It consists of a series of dials rotating on a single pin.

◇ The idea is that many of these concentric cardboard/plastic dials rotate together once properly adjusted, and as quickly as the Moon orbits the Earth. They show the Moon Phase, the Sun, the Earth and the resulting astronomical tide-producing forces.

Each disk is loosely fixed to a neighbour with friction (e.g. a smear of dried silicone glue) or a clip or tape or a pin, to allow adjustment.

Set it up by observation, or by tide-table predictions. Add notes, as you accumulate information about locations, tidal range, etc.
It can be especially useful for travellers, e.g. when fishing an unknown coast.
If the dials are correct for any one tide (spring-tide is best), they should be correct for all the tides thereafter *
* give or take a couple of hours for the smaller effect of the position of the Sun.
The effect is 43% of that of the Moon, and can be estimated with practice—change the hour reading to toward the Sun, or the anti-Sun, a little, if the Moon is near the Sun-anti-Sun line (no more than 45° from). Local shallow water resonance effects will also alter the actual effect of the Sun. When in doubt, around neaps, try averaging the tide-time for one or two days ahead and 1 or 2 days after.
The theory is that slosh resonance, or the Moon alone, or the Sun alone, might predict the difficult tides better.

◦ The hanger (outline not fully shown) is hung from the wall and also folds down behind to receive the pin on which the dials rotate.
It doubles as an indicator window, and should mark five consecutive high tides (ignoring a.m./p.m.).
◦ The clockface is 24-hours i.e. 15° per hour clockwise, (markings not fully shown). Think "a.m./p.m." for the two high tides per day, no matter what the reading says.
Each new location visited may need the clockface to be twisted slightly to allow for local tide delays. Keep a note of such settings.
Daylight saving requires a +/– 1hr change — to relocate the Sun forward in summer, back to normal in winter.

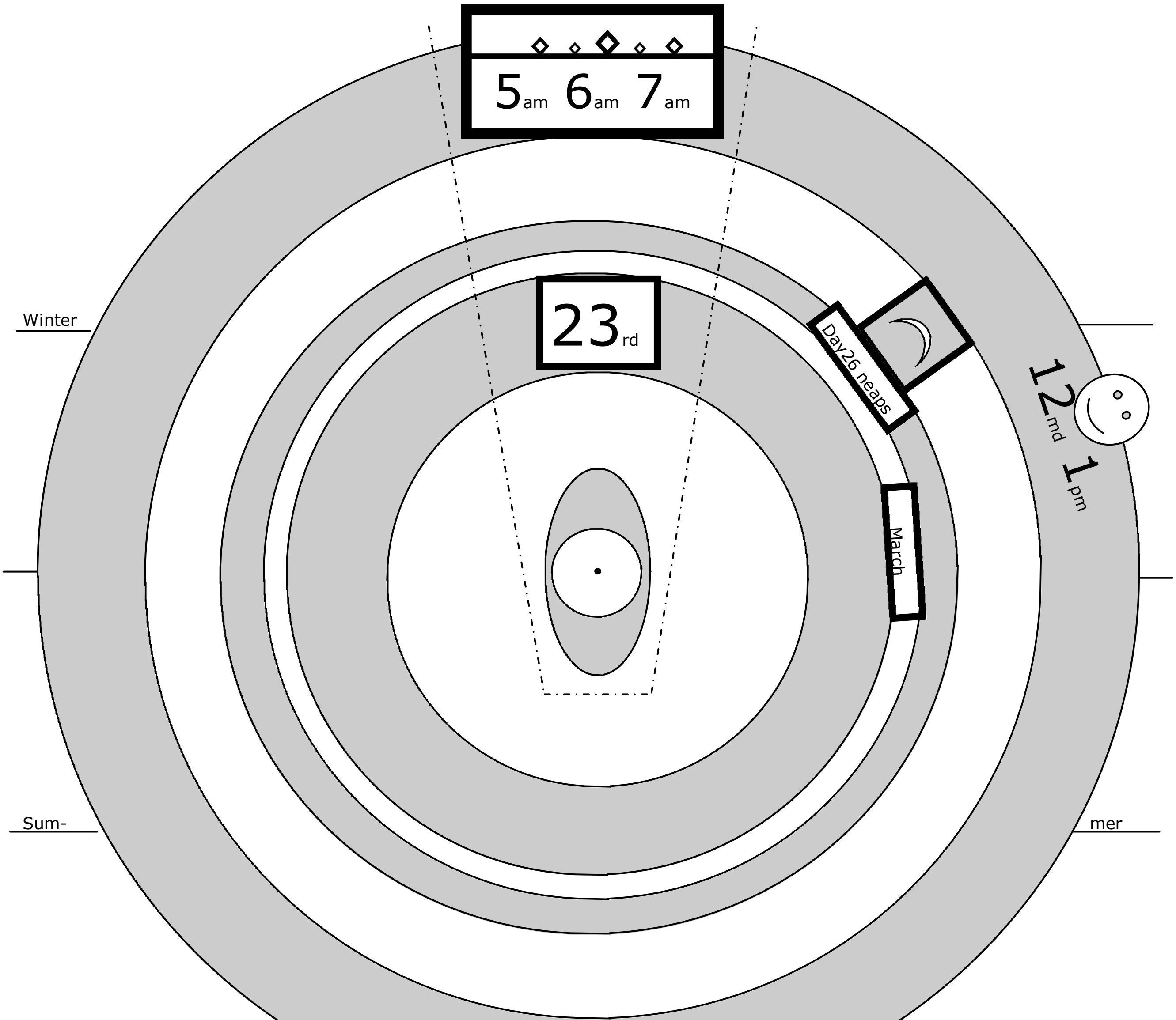

◦ Friction-fix the Moon-window dial (white with grey window) to the hanger, so that it maintains its position. It covers over all the other Moon phases.
◦ The Moon-phase diagrams (29 of them) underneath that, on a separate dial of the same diameter, should look right for the Southern Hemisphere—i.e. the Sun rising in the East, illuminating the Moon; all the phases parallel. Make the Moon-shape highly visible from a distance. Attach the Sun to the No-Moon phase, Day 0, via a transparent link; then set the Sun to the Clockface, to indicate true noon—when the Sun is exactly North.
◦ The optional tide-type ring (showing "neaps"), set to the Moon-phases, allows you to adjust the delay after new/full Moon, before spring tides.
If you also indicate which high tide is dominant—the one nearest the Moon or the one farthest—then that will need adjustment when the Sun crosses the Equator. I suggest that you set the top tide for the tide when the Moon is visible, to give you a visual feel for the tide producing forces.

◦ The date-dial is 29.5(30526805) days around—the complete lunar month—and marked clockwise from 1 to 32 (only one date is shown in the diagram). **Dial up the date, by rotating the Sun**. (For a 12-hour clockface, you would need to double the 12.19077473° per day). Since the next high tide will (if you are looking in the daytime) be on average that afternoon, I suggest you think of the date number as "the middle of the 23rd"—to avoid adjusting the date hand minutely for how far through the day the next tide will be.

◦ The month window is attached to the date dial, at Day 1 of the month, if you include a month setter.
◦ The optional but useful month-setter dial (markings not all shown) starts at March 1, to avoid leap-year adjustments; It makes March 32 = April 1, April 31 = May 1, and so on—spanning approximately 11 days for the year. Restart every March 1, and reset it every month. Calculate the monthly positions by the day number since March 1, avoiding an off-by-one error (17.9°;23.6;41.6;47.3;65.2;83.1;88.8;106.7;112.5;130.4;148.3). Set the month name to the first day of the date dial.
You can also calculate and indicate a one-year increment, 129°.6327779, a leap year, 141°.82355, and a four-year increment, 170°.7218865. To make an adjustment, make the old date indicated by one mark on the month dial, now line up with the increment mark selected, so as to indicate a date in a new month/year/leap year cycle.

◦ An optional Earth ring in the very centre, can show the entire Southern Hemisphere map (not illustrated, or NH on the reverse side), and be set to the right longitude. It is fixed to the hanger, and to the tide-bulge ring...
◦ A second ring just behind the earth can indicate two tidal bulges, one pointing upwards (or otherwise if you want to indicate the local tidal delay compared to the actual ocean bulge).

Mock it up in paper, until you understand the functioning.
Glue those paper discs onto cardboard to trial the thing for a while.
After deciding on improvements, make it neatly, to last for decades.
E.g. Try getting the computer to print the dial markings onto acetate, and glue that to sheets of stiff plastic.

You can experiment with different dial orders, e.g. the date-dial on the outside of the phase-dial.

You can make room for comments,
e.g. by expanding the dial area showing, to say when the moon rises and sets; e.g. "light from midnight onwards"—beneath each Moon-phase diagram
e.g. when and where the Sun rises and sets; Mark three horizontal lines in the background for Summer, Equinoxes and Winter.
I.e. mark them on the backfold of the hanger piece, or on the wall. Try to match midwinter sunrise time to where the "winter horizon" cuts the clockface, which will also suggest the approximate direction of sunrise.
e.g. when and where the first, second and third quarters set in each season; Full Moon opposite the Sun, i.e. 6 months different; Half Moon Waxing is one season ahead of the Sun; Half Moon Waning is one season behind the Sun (as far as which horizon line to use).

You will discover that an invisible connection needs to be made between the outer top window and the inner window and dials, so that they do not rotate with the other dials. One way to do this is to push the pin through a piece of rubber eraser, and glue that rubber to the dial. Another method is to have a square pin, and key the inner dials to it, while using round holes for the rotating dials. A third and simpler method is to have a transparent link from the centre, over the top of the other dials. You do not want the hanger to obscure any diagrams or dials which need to be visible.

The Time and Date dials are the indispensable ones. Instead of dialing up the date, you can use a normal clockface (12md always on top) and an hourhand to point out the next tide; you point the pointer at the date. On a 24hour clock-face it can be made like a + sign, to point out all 4 tides that day.

Here is the simpler alternative version, showing an 11:30 high tide on the 19th…

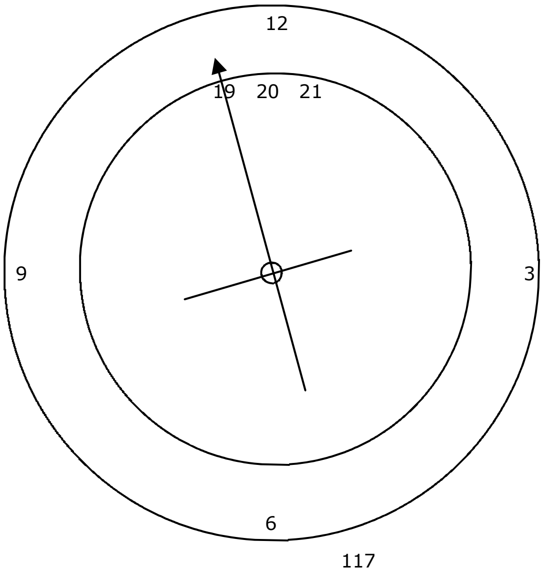

Divide the Star Globe into 8 Octants—and Learn the Stars in Each

The three dividing great circles are the Equator, the Ecliptic, and the Milky Way Galactic Plane.
A star may be North or South of each line, or on one or more of them. The lines on the backside of the globe are shown dashed.
The labels refer to the three lines in the order given. Six of the octants show in the front view with the two in parentheses on the back.
Verify each label before you start. E.g. NSN = North of the Equator, South of the Ecliptic, North of the Milky Way.
Compare with a good star chart e.g. out of a large atlas, as this is only a memory aid.

The Milky Way Crosses the Equator
at the centre of
Procyon-Sirius-Betelgeuse

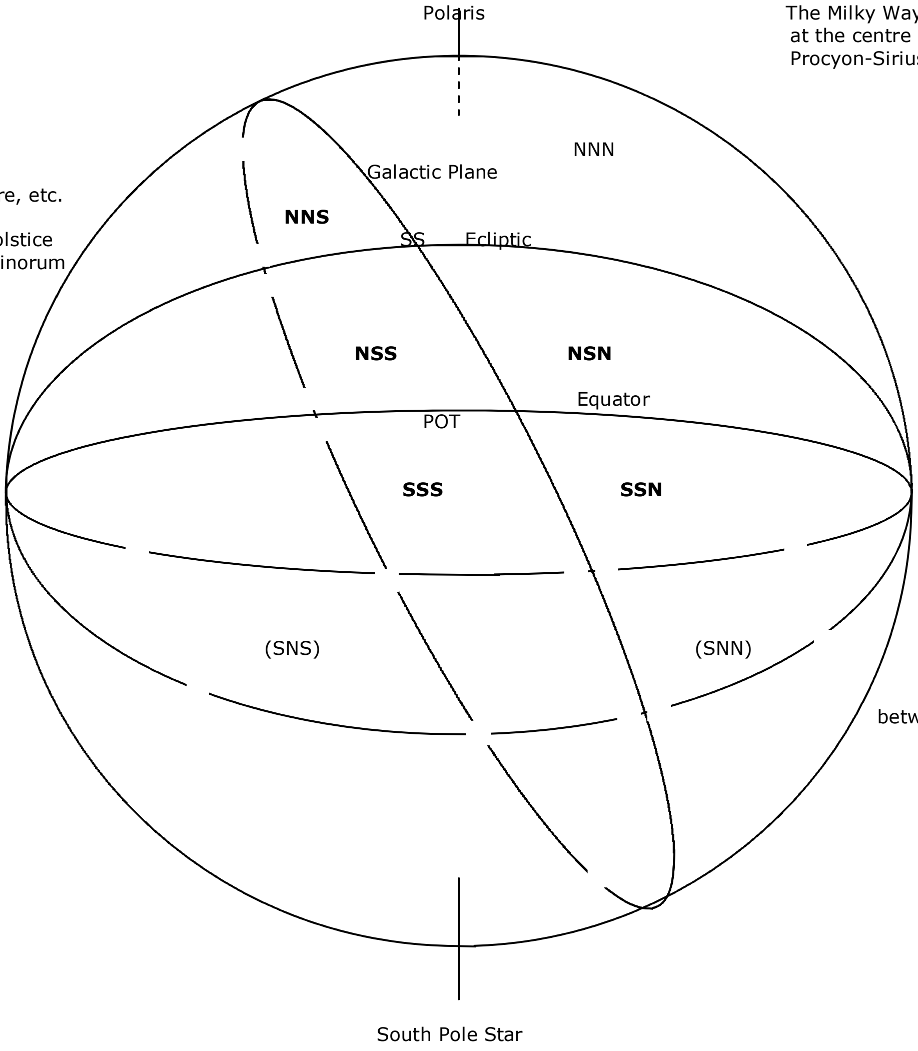

NNS—The Great Square, etc.

Summer (Northern) **S**olstice
Elnath to Gamma Geminorum

NSS
Aldebaran
Betelgeuse
Bellatrix
Menkar

SSS
Pot
Big Dog
Canopus
Diamond Cross
False Cross
Acamar
Achernar (central)
Magellanic Clouds
South Galactic Pole
South Pole Star
Diphda
Ankaa
Fomalhaut
The Crane
Peacock
The Archer
Scorpion's Tail
Atria

NSN
Procyon
Gamma Geminorum

SSN
The Centaur
Antares
Spica
Corvus
Suhail
Alphard

between SSS & SSN
Southern Cross
Southern Pointers

The Southern Figure of Eight—Two Star-Circles—joined by the backbone of Canopus—Achernar—Fomalhaut, a straight line

(not to scale; consult a star chart to verify each star, then find them in the sky)

Fomalhaut
Achernar
Canopus
Sirius |
Orion's belt | these 3 are also straight
Aldebaran |
Pleiades
Aries
Great Square

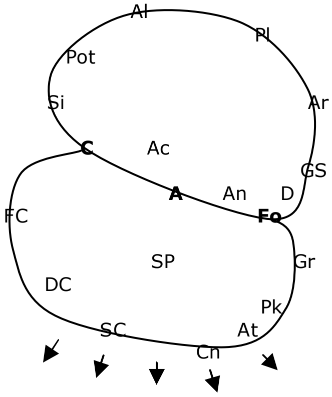

Canopus
Achernar (find Acamar on the way)
Fomalhaut (find Ankaa on the way)
Great Square (find Diphda)
Aries, the Ram
Pleiades, The Seven Sisters
Aldebaran, in the Bull
The Pot, Orion's belt
Sirius, in the Big Dog

Milky Way
Cn=Pointers (alpha and beta Centauris)
Southern Cross
Diamond Cross
False Cross

Around the South Pole Star
Canopus (find Acamar next)
Achernar (find Ankaa next) *Ach.-Fom.-Pck*
Fomalhaut *form a triangle*
Grus, the Crane, with Alna'ir *to find AlNair*
Peacock then Atria (find Scorpio) *& Ankaa*
Centaur, the two Pointers (find Arcturus)
(find Spica)
Southern Cross (find Corvus, the Crow)
(find Leo)
Diamond Cross, False Cross, Canopus

Visible in its entirety in October plus or minus a few months.

Divide the Star Globe into 8 Octants—and Learn the Stars in Each

The smaller groups, and the crossover points, are worth learning, because they show you the three lines best.

NNN
Alkaid (central)
Capella
Twins
North Galactic Pole
Big Dipper
Little Dipper
Polaris
Arcturus
Alphecca
Rasalhague
Draco
Vega
Deneb
Most of Leo
(Regulus adjoins NSN)

Summer **S**olstice
Between Aldebaran & Pollux

Autumnal **E**quinox
Between Spica
and Regulus

NNS
Altair
Enif
Great Square
Mirach
Almach
Algol, Mirfak
Aries
Pleiades
Elnath
(Caph, Schedar, Ruchbah)

Winter **S**olstice
Nunki to Sabik

SNN
Zubenelshemali
Sabik
(Zubenelgenubi adjoins SSN)

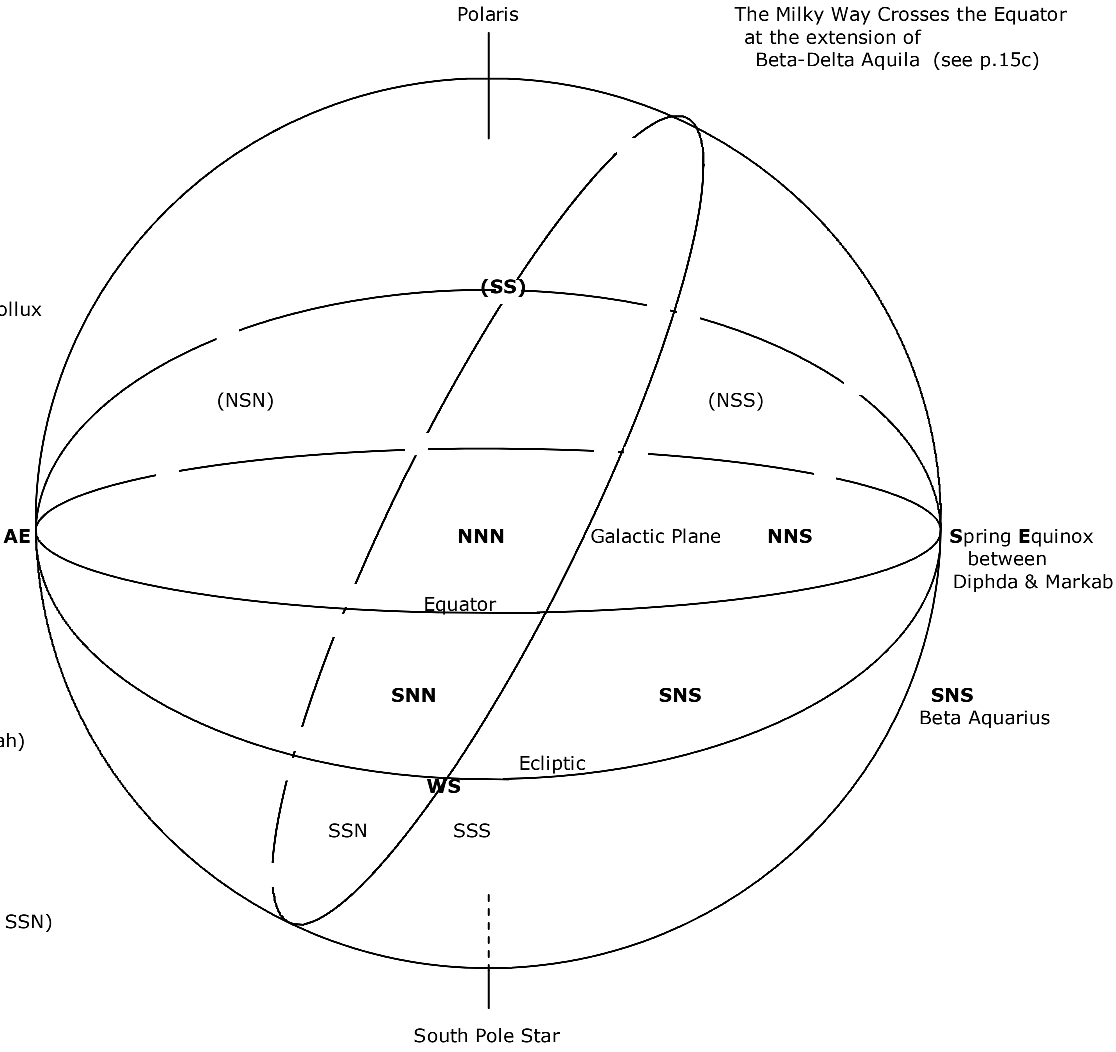

The Milky Way Crosses the Equator
at the extension of
Beta-Delta Aquila (see p.15c)

Spring **E**quinox
between
Diphda & Markab

SNS
Beta Aquarius

The Zodiac i.e. the constellations along the ecliptic. In order of appearance. Not all of them are bright or obvious in shape to look at.

Mnemonic:	Are	Tall	Twins	Cancerous	Leo?	Virginal	Liberated	Scorpions	Shoot	Capricious	Aquarium	Fishes.
i.e.:	Aries,	Taurus,	Gemini,	Cancer,	Lion,	Virgo,	Libra,	Scorpio,	Sagittarius,	Capricornus,	Aquarius,	Pisces
	Ram	Bull		Crab			Scales		Archer	Goat	Water-Carrier	

Morse Code & Rhythm Mnemonics

Dit/Dah/Pause is an extremely flexible system. Adapt it to torchlight, sounds, sewing stitches, body movements… A lot of fun for all the family.
if dit = one time unit; then Dah = three dits long; the di-Dah separator = one time unit; letter separator = three units long; word-separator = 6

	Say…		**Write…**	**Signal…**	Special Meanings (only for reference)
A	an A		. _	di Dah	
B	B bi-bi-bips		_ . . .	Dah di di dit	Fog Horn: I am under tow
C	C is Sea-sick		_ . _ .	Dah di Dah dit	Between ships manoeuvring: I a**c**knowledge your signal
	(Capital letters for a Dah syllable; lower case for a dit)				i.e. You may overtake on the side you indicated
D	Dee did it		_ . .	Dah di dit	Fog Horn: restricted manoeuvrability; deep draught; sailing; etc.
E	egg		.	Dit	I am altering my course to... **Starboard**
F	f’ f’ Forfeit	f’ f’ Fore foot	. . _ .	di di Dah dit	
G	Gee Gee git	Go Go get	_ _ .	Dah Dah dit	I intend to overtake you, to Starboard
H	h’-h’-h’-hit		. . . .	di di di dit	I intend to turn completely around to… Fog Horn: Pilot vessel on duty
I	is it		. .	di dit	I am altering my course to... **Port**
J	a J Goes Jay		. _ _ _	di Dah Dah Dah	
K	K for Kay		_ . _	Dah di Dah	
L	an ELL leph’nt /in ELLement /inELLegant		. _ . .	di Dah di dit	(draw out the Ell sounds)
M	My My		_ _	Dah Dah	Fog Horn: “under power, not making way”
N	No en		_ .	Dah dit	
O	Oh Oh Oh		_ _ _	Dah Dah Dah	
P	a P Goes pip		. _ _ .	di Dah Dah dit	
Q	Q Forms a Queue		_ _ . _	Dah Dah di Dah	(to the tune of “Here Comes the Bride”—Queue up to get married!)
R	an R dits		. _ .	di Dah dit	
S	s-s-s	ess ess ess	. . .	di di dit	I am under Astern propulsion
T	Tea		_	Dah	Fog Horn: “under power, making way”
U	it is U		. . _	di di Dah	
V	v’ v’ v’ Vee		. . . _	di di di Dah	(to the tune of Beethoven’s V)
W	a White Wine	a World War	. _ _	di Dah Dah	
X	EX mimics P	All ex’s Hurt	_ . . _	Dah di di Dah	(drawl the EX, to lengthen it in the first, not in the second)
Y	Why is Y Y?		_ . _ _	Dah di Dah Dah	(say the ‘is’ quickly)
Z	Zee Goes zi-zit	Zees Go zi-zit	_ _ . .	Dah Dah di dit	I intend to overtake you, to Port
Ü	it is U too	(this pairs to Zed)	. . _ _	di di Dah Dah	(U umlaut or U diaresis. Invent a meaning for it.)

Hint: For Learning, repeat each letter thrice.

By the time the other person recognises the first, the second has finished and he is waiting to check it out against the third.

0	No No No No No		_ _ _ _ _	
1	it Carts Four More Dahs		. _ _ _ _	
2	it is Two Two Two		. . _ _ _	
3	di di di Makes Three		. . . _ _	
4	di di di di Four	pitter patter 4 (your puppy’s feet)	. . . . _	
5	fit ‘em in a mit	(your five fingers, that is)	. . . . .	
6	Five si– si– si– six		_	
7	Five Seeks seven dits	(=567dits drawled on the 6)	_ _ . . .	
8	8 8 8 bi bits	(8 binary bits in a byte)	_ _ _ . .	
9	9 9 9 9 nips		_ _ _ _ .	
	Long dash	Daaaaaaaash	_______	e.g. (Fog Horn or car horn) before you enter a blind bend!
	Error	it isn’t a bit of it	. . . .	Between yachts: “I do not understand/agree with your intention/action”
:	Colon	Now Here Comes in a list	_ _ _ . . .	

Learn the next ones in pairs i.e. swap dit for dah

.	a STOP a STOP a STOP	(i.e. don’t pronounce ‘stop’ as short)	. _ . _ . _
;	SEMi SEMi SEMi	(i.e. drawl the first syllable)	_ . _ . _ .

These 2 shapes are convex to the right

,	Go Go comma Go Go	Dah Dah di di Dah Dah	_ _ . . _ _
?	did he Dah Dah did he?	did ’e Blah Blah did ’e	. . _ _ . .

These 2 come in pairs normally

“”	a Quote is a Speech bit	di Dah di di Dah dit	. _ . . _ .
()	Here’s the Blah Blah bit Here	There’s the Blah Blah bit There	_ . _ _ . _

These 2 are in the middle of a word ..hyphen and apostrophe

-	’Aitch in the middle Daash	(‘Hyphen’ starts with an ’aitch)	_ . . . . _
’	it Saves All Those Hard bits	di Dah Dah Dah Dah dit	. _ _ _ _ .

/ Slaash t’ diVide it (Drawl the first vowel) Slaash is a Slaash bit There was the Whole bit _ . . _ .
Á an A Sharp an A You could use it for “Repeat”: re Peat Whole re Peat . _ _ . _
É it is E a-cut(e) . . _ . . cute is cut short (You could use it for “Bad Reception”: cannot Hear the bits)
Ñ Nya Nya ny’ Nya Nya _ _ . _ _ (with thumb to nose, fingers waving) “ny” sound (I use enya for “Do You reCeive Me”)

SOS a bit of ‘Save Our Souls’ isn’t it? (no letter breaks; repeat regularly) . . . _ _ _ . . . . . . _ _ _ . . . (Life & death only)
XXX Come quickly Come, Come quickly Come, Come quickly Come All ex’s Hurt All ex’s Hurt All ex’s Hurt (When you Need urgent help)

Signals procedures, developed by hard experience, short circuit a world of miscommunications — stick to them. I.e. learn them too.

I’m Looking For Talk; CQ CQ CQ;	Calling Calling All Radios; Hear ye Hear ye All AuditOrs; Tune in Tune in Lend Me y’r Ears
From; D, E; De = French for ‘from’	Heard of it ; fr’m Dah-di-dit dit
Go Ahead; invitation for anyone to transmit; K	Call in Now
Break In B, K	Break in to it; Make a Break (“BK DE (identify yourself)” = “let me use this channel”)
Starts; Attention; Message Coming	Start it Start it Now Start of My bit Now _ . _ . _
End of Paragraph; Pause; new thought; double dash	Pause if it is New _ . . . _
Wait a Wait f’r a bit . _ . . .	You could invent “Speed up All Your Work” to pair with it.
Over; invitation for your contact to transmit	O ver Your Turn stop _ . _ _ . (to start a person-to-person contact: ###DE### Over)
End of Message in traffic; not end of contact; not clear of station	the End of My bit; con Clude con Clude it . _ . _.
R; Roger; Received and Understood all of Message	
End of Work; but still on air; clear of contact	it is the End of Work (to the tune of… here we go Out to Play) . . . _ . _
Clear of all contacts; shutting down; Clear of Station; C, L	Clear it Clear it, b’ Clear of it

Add your own signals for procedural or other matters. E.g. “Please repeat my message back to me for a check” “Adjust the beam left”

Ä	an A an A	. _ . _	an A ‘m Laut It pairs to C Invent a use for it like “i Don’t a-Gree” to be opposite to C, I Agree
Å	an A an A too	. _ . _ _	an A an Aang Strom
Xhi	Khi Khi Khi Khi	_ _ _ _	K-H-Like-H soft Kh sound (I use it for “to” i.e. T+O, after DE “from”) cf Ree Call All Souls
Ö	Oh Oh Oh dit	_ _ _ .	Goe– Goe– Goethe It pairs to V You could use it to say Sorry

North American Morse is somewhat similar. There is also a Japanese version. Morse is no longer supported as International.

Semaphore Signalling Code & Mnemonics for it

Semaphore is a clockwise system, as viewed by the decoder, and it starts from below.
When the two hands can't be seen, that is the neutral position, used for a… **Pause**
The black square represents the signaller's body; the line(s) are his arm(s).
The numbers 1-9 and 0 are the same as A-I and J. The first 7 single-arm positions are shown…
You can tell the difference, because the signal "numbers coming" precedes a number, and a "pause" follows the end of the numbers.

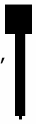

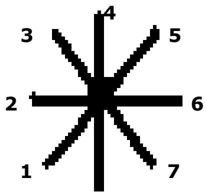

One arm, pointing up, therefore is either D or 4,
Think of God as the 4th Dimension to your life, and remember 'up' as "**4D**"
Or if it sticks in your memory, think perversely, that "**Up is fo(u)r D.own**!?"
Learn the letters in symmetrical pairs.

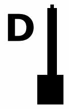

Here is a "**B F**lat" mnemonic. One arm flat is either a B (-flat) or an F (-flat)
"2 be or not 2 B?" and a "**F Sixth** chord" might help you remember the numbers.

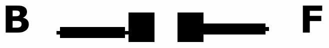

Wave the C to say "Replying". Wave the E to signal "Error"
(cf Morse "C" for "Acknowledge")
The *odd number* positions are diagonals.
1,3,5; 7… **One is an ACE** isn't it? And Dia**g**onal has a **G** in it doesn't it?
a G7 chord to follow the F **F**lat minor **6**
ABC = 123 is easy enough to see, and **5E follows 4D** makes the E position easy

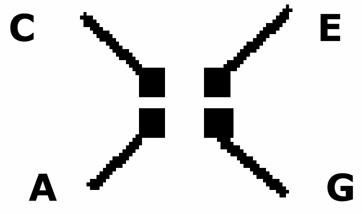

2 Hands…

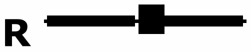

"Roger" is simply an indispensable signal to learn… it means "Yes, I understand, OK, Received"
For Learning: Acknowledge every letter understood, with a Roger

2 Hands Straight:
Think **"Long**, then **Left down = Slash down** or **L**eft **up = L**ift **up"** (referring to **their** Left arm)…

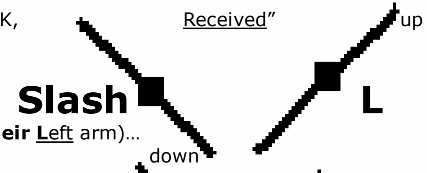

2 hands in a catching formation, above the head…
Shout **"Catch iT…"** and think **"..Tee—numbers coming"**

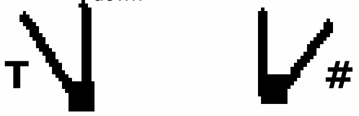

2 Hands in a slip-catching formation, to one side…
Shout: **"HOWZ"**at!

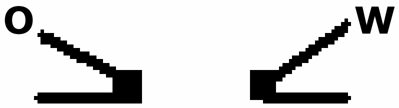

You are meant to wave black and white flags around,
but usually you make do with opening your palms.
Only be careful to position yourself against a good uniform background.

'aitch rhymes with eight

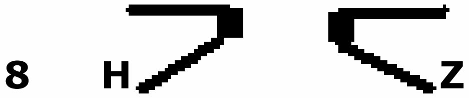

2 hands at a right-angle…
"Ya gotta have the **Right** **P**y**J**amas" ("PJ"s)
J, the tenth letter, is used for Zero ABCDE FGHIJ

PJ is not in alphabetical order, because in the Semaphore Alphabet J comes after #

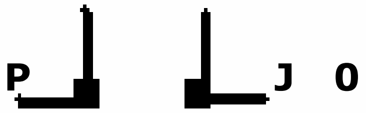

2 hands at a right angle, Vee-shaped, up or down…
Can you see the **shapes** of the **lower case letters n and u** in them?

To attract attention wave U, to say "Message Coming"

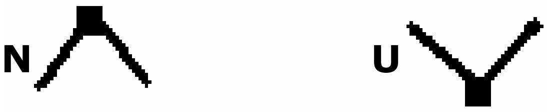

2 hands, like a Roman Numeral Five on its side…
I and **X** are **Roman Numerals too.** The I is used for numeral **9 (IX)** in semaphore
The X looks like **the right-half of the X**

So it sort of goes, left to right "nine, ten"

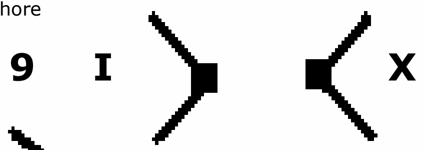

2 hands flapping and banking. I.e. a shallow Vee formation…
"MY Queue**S"** is the clockwise mnemonic
But YQ is not in alphabetic order (like PJ, the only two cases).

You can use Q for "Query" - What did you say/mean?

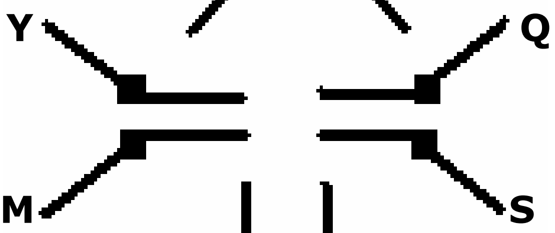

From the sky, down to earth come… **"K**ilo**V**olts" (or **"K**iller**V**olts")

You can follow the clockwise logic of Semaphore if you leave out the JVWXZ till last
Pause, ABCDEFG, HIKLMN, OPQRS, TUY/, #JV, WX, Z

Make each in turn to see the progression.

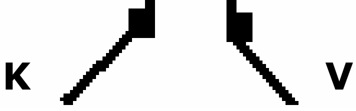

Appendix: More symbols for map-making Useful for mud-maps, or for track notes.

The first symbol in a strip map column should show the *reason* for the stop. This saves a lot of words and is instant to read.

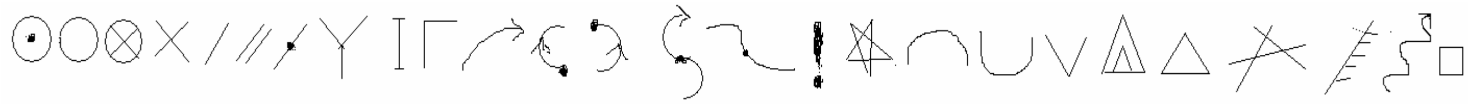

In turn these are:

A Context Entry; Surroundings; Overall Perspective. E.g. catchment and drainage. It's a chance for a general note.
A Stop. Expect a reason like 'Lunch'.
A Location. Expect a description.
Crossing. This might be a crossing of a ridge, river, road, canal, boundary. Expect notes on what is to left and right.
Leg. The first of the 2 main reasons for an entry: either a new subsection (leg) or the location *between* two sections in its own right (next symbols).
Overlap-section between legs. This is a memory aid, to sequence subsections, to remind you how individual sections join to each other.
Change of legs; new subsection; = notes on *the changeover*, rather than on the next section.
Junction (Y for Y junction, but used also for other junctions). This might be a 'natural' junction between ridges or creeks; or an intersection of paths.
Viewpoint; Change of view; First glimpse; Last glimpse (I for 'Eye')
Corner
Bend
Curved section beginning; Arrow shows clock-sense. This one is a Right-hand bend.
Curved section ending; Arrow should show same clock-sense as previous one. This one ends a left-hand bend.
Change from left-curve to right-curve
Break of slope
Feature
Special feature
Ridge; Crest
Saddle;
Valley; creek-bed; Gutter
Tent; Campsite
Peak
Fix
Cliff
Creek
New grid square

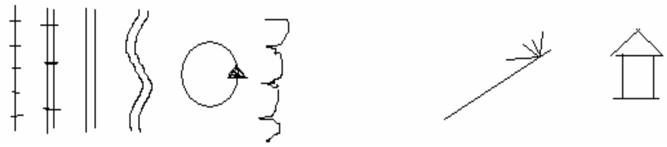

Boundary walks: fenceline; railwayline; road; path; circumnavigation; shoreline
Use symbols which work for you, e.g. Rise or Set direction. Add your own, e.g. Hut

The second symbol ought to show the main direction clue, followed by subsidiary ones as a third symbol

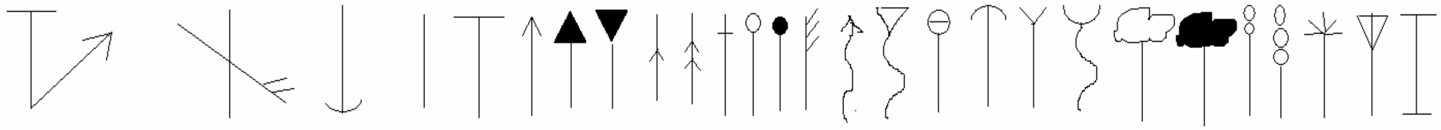

Trending half left from that landmark (write the name of the landmark next to the arrowhead). Useful when following a windy path.
Maintaining half right of downwind;
Walking the crest down ridge — Progress is assumed to be 'Up'wards on the page.
Leg.
Trend direction (for when the terrain is up and down, left and right) ('T' for Trend).
Named landmark; General purpose arrowhead; add the name and the bearing
Peak
Anti-peak direction (e.g. as mentally marked on the skyline opposite the peak)
Up-Slope.
One arrow for 'below average'; Two arrows for 'average'; three arrows for 'above average' — low; medium; high. Use this for other things too.
Compass needle North, modelled after the symbol for North
Up-Sun
Shadow; Down-Sun
Wind source; Upwind. Use the 1-2-3 code for strength. This one is 'above average'.
Down-stream river flow
Up-stream (to a peak) (useful when 'downstream' is behind you)
Moon direction
Up-Ridge; The direction of the highest skyline through the vegetation, if that is all you can see.
Down-valley. The direction of the lowest skyline when that is all you can see.
General Drainage trend (down the valley to the river mouth)
Up-Weather; Low cloud source direction; Cloud Source (when only one set of clouds is visible)
Down-clouds
Alto Clouds source direction (make the circles fluffier than I can show)
Cirro-clouds source direction
Glow at night
Radio-null (from the symbol for an antenna)
Trend of some feature like the coast — write in the details.

1,2,3 for size and importance of junctions…

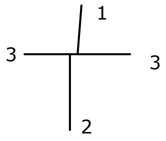

The notebook might be set out as follows: The first line across a narrow notebook is in informal columns; the second is for written notes.

Time ;	**Reason** ;	**Trend** ;	**Directions** ;	**Times** ;	**Distances** ;	**Grid**
Fill in a **number**	Use the top **symbols**	Use the main clue	make a rose of others	elapsed time	total elapsed	e.g. A5
if you don't				leg time	last leg length	
have a watch	Start with an 'Overall'			lost time	lost distance	
				total time out	total diversions	
Sequence is a				net walking time	net progress	
primary need						

(Use abbreviations, e.g. NWT, or dedicated columns)
(Do your calculations here at leisure)

Now write *across* the next line(s) the **name; mnemonic; description; comments**
The description of a cross-country leg could include the side or flank, hand, clock-sense, break, steepness, difficulty, curviness, hilliness, perversity, terrain, contours, topography, view, what you can't see, skyline, traffic intensity, track-markers, their age, rock type, soil type, vegetation, biology, colours, peculiarities, reference to another page of the notebook to a diagram.
The more you include in the symbols, the less you need to say, and the easier it is to read later.

Appendix: A List of Star Coordinates.

100 brightest stars
Declination is equivalent to Latitude. –nn = South; +nn = North
Siderial Hour-Angle is equivalent to Longitude West of Zero.

Common Name (s)		Dec.	S.H.A.	Constellation
Miaplacidus		-69.72	221.70	b Carinae
Atria		-69.03	107.83	a Trianguli Australis
Acrux		-63.10	173.35	a Crucis
Rigil Kentaurus	/Toliman	-60.84	140.10	a Centauri
Hadar	/Agena	-60.37	149.04	b Centauri
Mimosa		-59.69	168.07	b Crucis
Avior		-59.51	234.37	e Carinae
Aspidiske	/Tureis	-59.28	220.72	i Carinae
Achernar		-57.24	335.56	a Eridani
Gacrux		-57.11	172.21	g Crucis
Peacock		-56.74	053.60	a Pavonis
k Velorum		-55.01	219.46	k Velorum
d Velorum		-54.70	228.82	d Velorum
e Centauri		-53.47	155.02	e Centauri
Canopus	2nd brightest	-52.70	264.01	a Carinae
Muhlifain		-48.97	169.62	g Centauri
Regor		-47.34	237.61	g Velorum
Alna'ir		-46.96	027.95	a Gruis
b Gruis		-46.88	019.34	b Gruis
Suhail	/Al Suhail	-43.43	223.01	l Velorum
Akrab/ Sargas	/Rastaban Sc.	-43.00	095.67	theta Scorpii
Ankaa		-42.31	353.43	a Phoenicis
eta Centauri		-42.16	141.12	eta Centauri
Naos		-40.00	239.10	z Puppis
k Scorpii	(not Akrab)	-39.03	094.38	k Puppis
Shaula		-37.10	096.60	l Scorpii
pi Puppis		-37.10	250.71	pi Puppis
Menkent	(not Menkar)	-36.37	148.33	th Centauri
Kaus Australis		-34.38	083.96	e Sagittarii
e Scorpii		-34.29	107.46	e Scorpii
Ascella	/Nushaba	-29.88	074.36	z Sagittarii
Fomalhaut		-29.62	015.59	a Piscis Austrinis
Aludra		-29.30	248.97	eta Canis Majoris
Adhara		-28.97	255.34	e canis Majoris
Antares		-26.43	112.65	a Scorpii
Wezen		-26.39	252.90	d Canis Majoris
Nunki		-26.30	076.18	s Sagitarii
Dschubba		-22.62	119.91	d Scorpii
Diphda	/Deneb Kaitos	-17.99	349.11	b Ceti
Mirzam	/Marzim	-17.96	264.33	b Canis Majoris
Arneb	/Arnebo	-17.82	276.82	a Leporis
Sirius	/Dog (brightest)	-16.72	258.72	a Canis Majoris
Sabik		-15.72	102.41	eta Ophiuci
Spica		-11.16	158.70	a Virginis
zeta Ophiuci		-10.57	110.72	zeta Ophiuci
Saiph		-09.67	273.06	k Orionis
Zubenelschemali		-09.38	130.75	b Librae
Alphard		-08.66	218.10	a Hydrae
Rigel	not Rigil K.	-08.20	281.37	b Orionis
Alnitak	S. belt	-01.95	274.80	z1 Orionis
Alnilam	Mid belt	-01.20	275.94	e Orionis
Mintaka	N. belt	-00.30	277.00	d Orionis

Common Name (s)		Dec.	S.H.A.	Constellation
Polaris	N. Pole Star	+89.26	(322.06)	a Ursa Minoris
Kochab		+74.16	137.32	b Ursa Minoris
Alderamin		+62.59	040.35	a Cephei
Duhbe		+61.75	194.07	a Ursa Majoris
Navi	/Cih	+60.72	345.82	g Cassiopeiae
Ruchbah	/Ksora	+60.24	338.55	d Cassiopeiae
Caph		+59.15	357.70	b Cassiopeiae
Schedar	/Schedir	+56.54	349.87	a Cassiopeiae
Merak		+56.38	194.53	b Ursa Majoris
Alioth		+55.96	166.50	e Ursa Majoris
Mizar	(+Alcor nearby)	+54.93	159.01	z Ursa Majoris
Phecda		+53.69	181.54	g Ursa Majoris
Eltanin	/Etamin	+51.49	090.86	g Draconis
Mirfak		+49.86	308.92	a Persei
Alkaid	/Benetnash	+49.31	153.12	eta Ursa Majoris
Capella		+46.00	280.83	a Aurigae
Deneb	(not Denebola)	+45.28	049.64	a Cygni
Menkalinan		+44.95	270.12	b Aurigae
Almak	/Alamac	+42.33	329.02	g Andromedae
Algol		+40.96	312.96	b Persei
Sadar	/Sadir /Sadr	+40.26	054.45	g Cygni
Vega		+38.78	080.76	a Lyrae
theta Auriga		+37.22	270.07	th Aurigae
Mirach		+35.62	342.57	b Andromedae
Aljanah		+33.97	048.45	e Cygni
Castor	N. twin	+31.89	246.34	a Geminorum
Alpheratz	/Sirrah	+29.09	357.90	a Andromedae
Elnath	/Nath	+28.61	278.43	b Tauri
Scheat	/Schert	+28.08	014.06	b Pegasi
Pollux	S. twin	+28.03	243.67	b Geminorum
Izar		+27.07	138.75	e Bootis
Alphecca	/Gemma	+26.71	126.33	a Coronae Borealis
Hamal		+23.46	328.20	a Arietis
Sheratan	/Sheratain	+20.81	331.33	b Arietis
Zozma	/Zosma	+20.52	191.47	d Leonis
Algeiba		+19.85	205.00	g1 Leonis
Arcturus		+19.18	146.08	a Bootis
Aldebaran	(not Alderamin)	+16.51	291.01	a Tauri
Alhena		+16.40	260.56	g Geminorum
Markab		+15.21	013.82	a Pegasi
Denebola		+14.57	182.73	b Leonis
Rasalhague		+12.56	096.27	a Ophiuci
Regulus		+11.97	207.90	a Leonis
Enif		+09.88	033.96	e Pegasi
Altair		+08.87	062.31	a Aquilae
Betelgeuse		+07.41	271.20	a Orionis
Bellatrix		+06.35	278.71	g Orionis
Procyon		+05.22	245.17	a Canis Minoris
Vernal Equinox, defined as fixed		00.00	000.00	'the first point of Aries'

Figures are from Epoch 2000.0.

Precession of the equinoxes:

Stars drift about 1° of position relative to the (0,0) reference star, in **72 years**, because it moves to the **WNW along the Ecliptic**. See p 33b
The third figure will therefore become off by one after 10-20 years. E.g. Mimosa used to be –59.59; Atria used to be 108.3W
One degree of latitude or declination = 60n.ml. =111km; so a maximum change in **0.1° over 18** years represents 11km, or about 0.6km per year, max.
(The actual change in Declination depends on its relationship to the Northern Spring Equinox point — see below.)
One degree of SHA = 60n.ml. at the Equator; half that at 60° Lat. ; **1° reduction in 80 years** is about 3 seconds later in timing per year, or 1.4 km max, at the Equator.

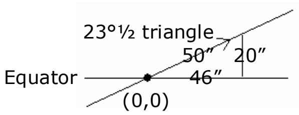

Ecliptic The first point of Aries moves along it at about 50 seconds of arc per year (5/6° in 60 Years)
Declination change of 1/3 minute of arc per year (20 secs of arc)
SHA change of 46 seconds of arc per year $20^2 + 46^2 = 50^2$

If you set your star watch by the actual visible stars each year, you eliminate E-W 'precession' component, and are left with only the drift *in declination*.
So it is of some worth to be able to predict how that rate of change in declination depends on SHA.
Obviously Stars to the North of Aries lose Northern dec, at the full rate. Stars to the South must *gain* Southern Dec.
Stars *opposite* Aries (SHA=180°) do the opposite. Stars exactly in between are *unaffected* (SHA=90° or 270°).
The rate for all cases follows **the cosine of the SHA**. E.g. stars at 60**°** SHA will have 50% of the full dec adjustment (of 1/3 n. ml. per year).
Take care with Polaris, since it is too close to the Pole.

Appendix: Drawing Great Circles

First, bisect a globe.
Two equal halves.

Put the halves back together.
Notice the ellipse.
The backside is dashed.

Draw it symmetrically around the centre.
Use *flat* ends to the ellipse.

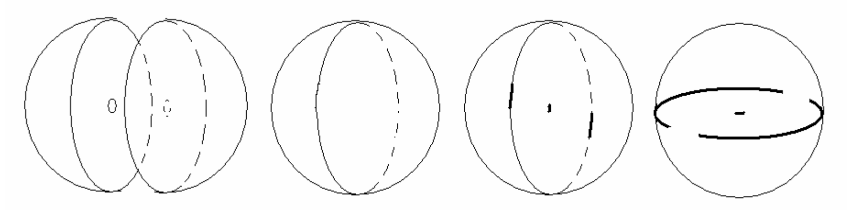

Choose the clearest perspective
E.g. 1. Rising/ setting is 90° away, so put yourself in the centre, surrounded by a 90° circle. Z is used for your Zenith — overhead.
E.g. 2. Look at the Equator side-on to make it and the latitude circles straight.
E.g. 3. Centre on the pole, to show up the polar distance. P is used for the Pole.
E.g. 4. Curve away from the centre. If in doubt, draw the whole ellipse and discard the unwanted parts.
5. A Great-circle can be drawn as a straight line. X is used to mark the position of the Star/ Sun/ Moon/ Planet
6. Draw the 'Vertex' of the course as a tangent to the maximum latitude small-circle.
7. The vertex can be drawn side-on.

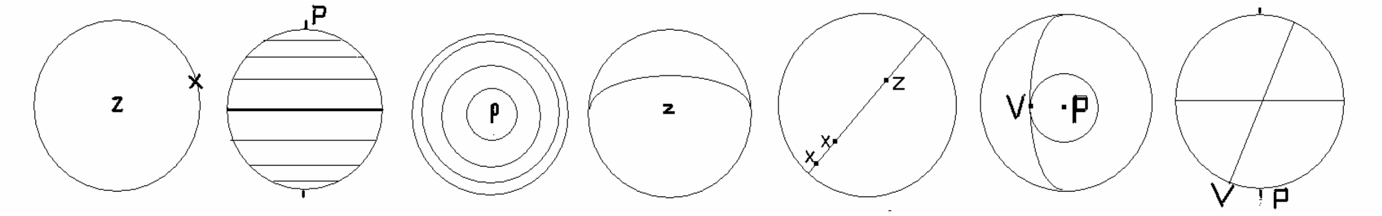

The PZX spherical triangle

8. Deliberately avoid straight lines to show up the curve of the triangles.
8a. ZX is the angular distance from you to the star, or the co-height.
8b. PX is the polar distance or co-declination to a point directly underneath the star.
8c. PZ is the polar distance or co-latitude of your position, 90° — lat.
8A. The angle ZPX, or 't', is a longitude difference — between the SHA of the star and your West longitude. A longitude difference is a time delay.
8B. The angle PZX is a horizontal direction relative to P, towards X. Technically called an 'azimuth'.
8C. The angle PXZ is similar but only meaningful if there were someone under the star, looking toward you.
9. A, B, C and a, b, c are used to label angles and sides of any spherical triangle; It's up to you to know which ones label what.
10. Draw a full 180° to show up the supplementary angles.
11. Draw the Equator cutting East and West of you, 90° away. Mark any right angles, and the complementary angles.

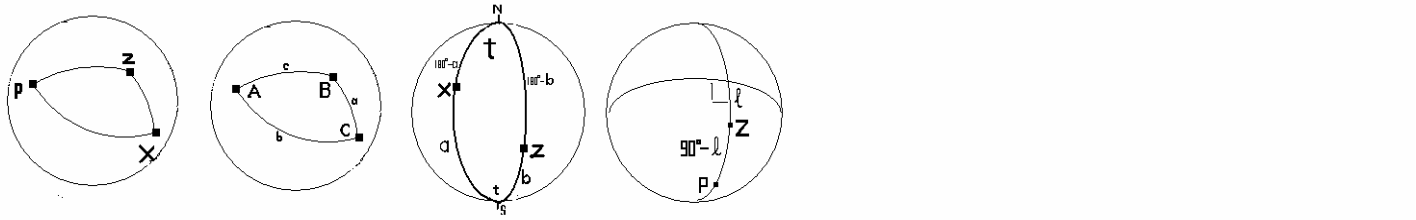

The PZX navigational triangle conforms to a handful of standard **fundamental spherical trigonometrical formulae**…

The Angle **Sum**: a+b and A+B are both greater than, or both smaller than, or both equal to, 180°
The **Sine** Rule: sina/sin A = sin b/sin B = sin c/sin C = constant. Ambiguity (about obtuse angles) can be resolved by the angle sum rule.
The **Polar** Triangle Rule: abcABC can all be swapped for the supplement of the opposite members.
E.g. 'a' becomes '180°—A'; 'B' becomes '180°—b', and the resulting formula is still valid. E.g. the sine rule simply inverts.
The **Cosine** Rule: cos a = cos b cos c + sin b sin c cos A
The Polar Cosine Rule: cos A = — cos B cos C + sin B sin C cos c
The **Four Parts** (in a row) Rule: e.g. AbCa or aBcA in the triangle aBcAbC or aCbAcB. You can find the 4th (outer) part from the first 3.
cos (**inner** side) **cos** (**inner** angle) = s**in** (**in**ner side) c**ot** (**ot**her **side**) — **sin** (**in**ner angle) c**ot** (**o**u**t**er **angle**)
If you want an inner part, use the sine rule and angle sum rule to find the 5th part, then go to the half-angle rules.
The **Five Parts** Rule: sin 1 cos 2 = sin 3 cos 5 + sin 5 cos 3 **cos** 4, where 4 is a side. It gives you any but part number 3, from the other 4.
The Polar Five Parts Rule: sin 1 cos 2 = **—** sin 3 cos 5 + sin 5 cos 3 cos 4, where 4 is an angle
The **Half-Side** Rule, to find the *middle* of 5 parts, when the middle part is a side: halve each side and angle first:
tan (half side) = tan (half sum) cos (half sum) / cos (half difference) ; tan refers to sides, sin and cos to angles
= tan (half difference) sin (half sum) / sin (half difference) but this introduces order-ambiguity because of the sines.
The **Half-angle** Rule to find the middle of 5 parts when it is an angle
cot (half angle) = tan (half sum) cos (half sum) / cos (half difference)
= tan (half difference) sin (half sum) / sin (half difference)

Make North and West positive, and East and South negative, and the formulae automatically work on PZX for angles up to 360°.

The formulae simplify where a side or an angle is 90°, since sin 90°=1 and cos 90°=0.
Napier's Rules…
The rules for **right-angled** triangles and **right-sided** triangles are simplified when the 5 other parts of the triangle laid out **in a pentagon**, and the 3 parts not adjacent to the right angle are changed to their **complement**.
In addition the part opposite a 90° *side* must have an extra **negative** sign.
E.g. if A is the right-*angle*, use (90°—a) as part 3, (90°—B) and (90°—C) for parts 2 and 4; if a is a 90° *side*, use A-90° for part 3.
Then… Sin of any one part = t**a**n t**a**n of **a**djacent parts or = c**o**s c**o**s of **o**pposite parts (in the pentagon of changed parts).
Ambiguity must still be resolved by special rules: Problems occur when there are 2 or more right-angles in the triangle.
'A' and 'a' (opposite parts in the triangle, with matching letter names) are both less than or both more than 90°. Likewise for Bb and Cc.
If one part and any part not opposite it (i.e. not A and a, as above) are on the *same* side of 90°, so are all the other mismatched pairs.
If one part and any part not opposite it (i.e. not A and a) are *not* on the same side of 90°, neither are any of the other mismatched pairs.
These refer to the triangle parts, not the altered pentagon parts.

Appendix: Calendars in the Sky — the first glimpse and last glimpse of stars

~ Firstly, What's wrong with **sunrise/sunset time**?

- To know what time of year it is, you could note the ordinary standard Sun time of sunrise/set at *that* place (e.g. from home).
  —But this is insensitive (maximum shift is about 1 minute a day), useless around solstice, latitude & longitude dependent & requires a watch

~ You need to use apparent *star* time; no watches, to **locate the Sun's position on the Zodiac relative to the stars**.
— i.e. 'what stars are up' at midnight, or at twilight, or what stars are invisibly up at mid-day or sunrise or sunset.

- The Sun creeps Eastward about 1° per day along the Ecliptic (through the middle of the Zodiac belt).
- That quadruples the sensitivity to 4 minutes change a day, steady year-round.
  It is only affected by the Equation of Time, which is a minor variability, 1/4 min a day, max.
- The Sun will 'blot out' each Ecliptic star for about a month at a time.
  That's a rough way of finding the month: 'I can't find The Twins' (July).
- 'Nautical twilight' is when the Sun is 10° below the horizon (about 1 hour, in Tasmania), dark enough for the bright navigational stars to be visible while the horizon still shows clearly enough to be used with a sextant;
  So bright stars 10-15° away from the Sun begin to be visible. That means that a bright Ecliptic star might only be invisible for 3 weeks.
  [Civil twilight is 5° below, requiring lighting; and astronomical twilight is 15° below, dark enough to be called 'night', and to use a telescope]

E.g. 'What stars are up at midnight', uses the 'apparent star time' of the anti-Sun, as it were. You need to bisect sunrise and sunset, and be up then

- Find which star is N-S of you (the 'local apparent star time') at 'local apparent midnight' (= opposite noon — i.e. the anti-Sun stars).

But that is difficult — you need to know the apparent noon (which is longitude dependent), not the clock noon — and it still requires a watch.
So how can you do better?

- You could note **the clock-time of any particular star event**. E.g. follow 'Rigel over Capella' during Summer. E.g. Antares at 10° high.
  This is *simple, and accurate,* but it would only work for that location, no other, since longitude alters the clock-time of appearances.
- You could note **the clock-time *delay*** between any particular star-event and either sunrise or sunset. **This is a good method**,
  since it pinpoints sunrise or sunset more accurately than saying 'At twilight, when the stars come out, or when they fade out'.
  For best results, choose which star-event directions to use, as explained below, so that people at all latitudes see the same appearances.
  - This pinpoints the star-time—it tells you where the stars are—in relation to the apparent Sun-time (by tying it to sunrise or set).
  - For Rise/Set observations, you need a good horizontal horizon (or other fixed height-of-star) for star-rise/set to be latitude-independent, and location independent; but at home you simply need a repeatable point of view of a given horizon.

~ But to get back onto the main topic: The use of a watch can be eliminated by using twilight 'glimpses'...

  - For Twilight observations, you need dependable eyesight; & a clear sky, with no Moon or light pollution, or haze, to be repeatable.

The idea of a 'glimpse' is that it is positively there or not. And it doesn't need a watch. There are 10 or so varieties...

- If you see a star *before dawn*, it's positively there, or if you can't see it *after dark*, it's not there.
  - Look to the East before dawn and you will see **a new star appearing** for the first time in the season — a first glimpse.
    You will then lose sight of it due to daylight. This is the standard star-calendar event.
  - Look to the West after sunset and a star will have **disappeared** for the rest of the season — a last sighting.
    You glimpse it briefly, before it moves into the sunset glow, and is lost to sight for weeks.
  - Examples: The last glimpse of Spica is just after the September Equinox; Regulus is first seen just ahead of the September Equinox;
    The Equinox is between the two, and both are on the Ecliptic, about 25° each side. That's 25 days for the Sun's movement.
    The first glimpse of Antares is around Christmas time, since it is about 20° ahead of the solstice;
    Aldebaran reappears around the winter solstice; The Twins disappear soon after it.
    The solstice is between these two, 25° away from each.

  You need to check the exact dates of appearance and disappearance because they vary with the exact latitude—this is a semi-local calendar.
- If *a star-pair* has tilted *out* of alignment by the time the stars come out, you simply *don't* see it occur;
  or if it doesn't align before dawn, you can't see it 'happen'.
  - Look pre-dawn for the first date on which you can glimpse **a star pair reach its proper alignment**, before dawn takes over.
  - Look post-sunset for the *last* date on which you can see a star pair attain its right alignment, after which it is lost for a month or so.
- Each day a Western star gets closer to the Western horizon from above. Each day a rising star gets earlier to the Eastern horizon, from below.
  - Look for the first glimpse of a star clearly **rising or setting** *pre-dawn*
  - Look for the last glimpse of a star clearly **rising or setting** *post-sunset*.

  If you see it setting, you know it; if you see it rise, you have seen it happen.
- Wait to see whether a star reaches a known fixed **height** (e.g. not too far above the horizon), to make observation easier.
  - Looking at a reflection of a star in smooth water or oil, will *double* its height 'above horizontal', if you want a method to work anywhere.
- Wait until a star coincides with a known fixed **direction** (e.g. 'over the tower'). This makes it location-specific, but may be useful.

These all have the effect of describing the height or tilt or direction of the horizon 'at twilight'.

- A star **overhead** at twilight will show up as a first/last glimpse. You look for it, but it may not 'happen'. This is latitude-specific.

~ You could now combine the 'glimpse' methods and 'crisp timing' to avoid the indistinctness of twilight (but this is more latitude dependent).
Look 'exactly 1 hour from sunset/rise' to crisp up the timing of the 'events'. Any exact time-span at least that long, would suffice.

~ To clinch the star-time, in a latitude-independent way, **use Great-Circle type directions** ('at right-angles to sunset/sunrise' 'a star near the Sun at twilight' 'opposite it') to ensure that people over a wide range of latitudes more or less agree in their observations.
The standard direction is not sunrise/sunset itself but 'where the Sun is *at twilight*' which is 1 hour to the left (SH); *the brightest part* of the horizon.

- identify star-pairs ***at right-angles to the Sun's* direction at twilight**.
- or examine the rise/set of **stars *near the Sun***, especially those ahead or behind it, in its rising and setting path.
- or look for stars rising, setting or simply being visible, ***opposite* the Sun** at twilight.
- You can use any other 'star direction', but it will only apply to that latitude. The best great-circle directions are between NW/SE & NE/SW.

~ But the sensitivity can be further improved, by **choosing *which* twilight** gives the most daily change…

- The nightly change (averaging 4 mins) is not the same amount through the year.
  It can be reduced to 3 or increased to 5 in a place like Tasmania, as the seasonal day-length varies by several hours.
  The change in sunrise or sunset time by up to a minute a day is overlaid onto the steady star progression.

Look at the time of day when the most rapid change is taking place, since you want the most accurate discrimination:

After mid-summer, look for 'first glimpses' in the pre-dawn ;

- Nights lengthen after the summer solstice — more stars rise in the East than set in the West as the night-time gradually dominates
  At dawn, the Sun lingers a little longer under the horizon allowing more stars to rise, in their creeping ahead of the Sun.
  But at sunset, the Sun goes down earlier, like the stars do, and you don't notice much change as to which stars are up.
- So when the nights lengthen, look **before dawn** and try to see any first glimpse this season.
  Luckily the pre-dawn in Summer/Autumn is mild, but it can be quite early. Use the clock-time delay method 'from sunset', not 'to sunrise'.

After mid-winter, look for 'last glimpses' post-sunset ;

- Conversely, after the winter solstice, the increasing day length dominates over the night sky — more stars set and fewer stars rise.
  You will lose sight of stars setting in the West, quite easily, as the Sun will linger before setting, because of the season,
  but find it hard to see a new star rising in the East for the first time that year, because the Sun pops up early.
- So when the days lengthen, to pinpoint the time of year, look **after sunset** for any last glimpse.
  That's not too late at night in Winter/Spring, and it should not be too cold. Or use the clock-time 'until sunrise', to maximise the daily gap.

The overall idea works in both hemispheres, but the optimal choices of when to look are tied to the seasons, not the date.
Use either method, when near to the solstice, or near the Tropics, or at any time you want to.

- Around mid-winter and mid-summer, you can happily use any of the 10 glimpses, or any of the clock-time methods.
- Near the Tropics, also, there is little differential advantage, compared with temperate latitudes, since day-length is fairly constant

~ Keep a special perpetual calendar (a 365 day chart with no days of the week specified) to note down your observations; with room for related seasonal changes, like bird migrations, frosts, plantings.
All this has little to do with navigation, directly, but gives you practice in *following the seasonal movements* of Sun and stars.

About the Author.

I have a big bushy beard from too much bushwalking and ocean sailing, but now that it has grown too grey, it is time to hang up my boots and write it all down.

My dad was a ’plane navigator during World War 2 and as a child I studied his manuals. Then at University I spent more time in studying naval navigational manuals and in bushwalking off track, finding my way over the snow or canoeing across lakes, than in studying for my exams. I used the library to learn spherical trigonometry, then, as soon as I finished Uni, set out across the trackless ocean. I soon found myself inventing techniques to plug the holes in my normal abilities. I returned home years later, not knowing what I was meant to do in life, but as I lay there on my old bed, I knew that if I was to be a navigator I would like to be one like Captain Cook. He experimented with new techniques, and with old — he described native sea-navigation expertise. How *did* they do it? Even now, no-one seemed to know. I immediately had a vision of one star above another, which kept me busy for years on great-circle navigation methods.

I had honed my skills trying to teach boaties and would-be ocean sailors, but my biggest challenge was to teach my wife and children. We became a feral family for a while and walked every track we came across, here and overseas. Only then did I appreciate how much more there is to orienting yourself *to your surroundings*, than simply having a compass.

My next inspiration came while walking the forest on a foggy day. I simply asked how the trees themselves would hold the secret to which way is which. Years of observation and testing later again, and this is the result. I don’t want it to die with me.

I hope you too can now enjoy the magic of the science and love the art of exploring, even teaching yourself, before you too have to stop still, inside the bounds.

Learn a little before it is too late.

Outline

"Navigation Down-Under" is your complete manual of bush navigation and outdoor orientation for when you go off the well-worn track. A free pocket guide to take with you bushwalking is part of the pack.
Do you know how to predict and find direction from the Moon by memorising just two simple numbers? Can you interpret the Australian bush plants for direction? You will find some never-before published new techniques, as well as all you need to know about traditional methods like Lie-of-the-Land.
If you come from the Northern hemisphere, you'll need the top-down global approach, and can try some special Southern Hemisphere methods of direction keeping.
Sailors will find here a valuable beginning into celestial navigation.
A lot of it is about how to navigate without a compass, how to stay oriented and how not to get lost. It is an essential course in basic bush awareness and orientation, for surety and safety, plus it also throws out the challenge that you can improve your already existing competence very quickly and easily.
For Scouts or emergency search-and-rescue workers, it is a thorough course, advancing beyond the map and compass. Or perhaps you are a tour-guide and need more than course and distance — You'll get something more, something more like biogeography or ecology.
Each page is a carefully arranged lucky dip of interesting tidbits.

Keywords

Navigation; Orientation; Celestial Navigation; Bushwalking; Map and Compass; Without a Compass

Section 1. BIG-PICTURE TOP-DOWN pp 3-33

General Introductory Points Big-Picture Rule One:

3 Locate a Distant Reference Direction First
3 Visualisation is your Key Skill
3 Big-Picture Orientation, a Mental Layout Acquired Beforehand
3 Compass Navigation Robs You of a Global Perspective
3 The Big Picture is Invisible — Your Job is to Visualise It!

Global Perspective; Earth

3 Compass Navigation Robs You of a Global Perspective
4 Global Re-orientation — Light-Half/Dark-Half
5 A mnemonic for global orientation: “True Polarity”
6 Mnemonic: “Time Tells Turning”
7 The sky turns 15° per hour —Use your thumb and forefinger
8 Global Orientation: Latitude and Polar Distance
9 The Angle of Rising and Setting—is Determined by Your Lat.
10 Global Orientation: Point to the South Pole…
..half its polar distance below horizontal
11 The North Pole is *Perpendicular to the South Pole*
12 Global Orientation: the South Pole Star, and Polaris
13 Pole Stars—“(Latitude)° above and below, South & North”
Point to 2 Points on the Equator — Halfway between the Poles
15 Perspective Distortion — The Sun Moves in a Straight Line!
16 Global Orientation: Half the Arc-Distance below Horizontal
17 Half the Arc-Distance below Horizontal—2 Explanations
18 Example: Point to Tahiti
19 Use Latitude for a Distance Scale (Not ‘Long’-itude)
20 The Horizon Rotates to Your Right in the S. Hemisphere
21 Global Perspective — Rotation of the Horizon
22 How to Shift Your Horizon —Its Timing and Height
23 Latitude without Numbers
24 Mnemonic “The Sun Travels the Tropics”
25 Tropical Countries and Places Under the Sun
— Arranged in Time Zones Relative to Eastern Australian Time
— 10 hrs ahead of Greenwich
26 North/South and Noontime by Bisection (of Rise and Set)
27 Finding Magnetic Variation by Bisecting Rise and Set
Find E/W by Delayed Bisection of Sun & Full Moon At Rise/Set
28 Find North from the Combination of the Sun and the Moon
29 Eastern Australian Standard Time Zone
10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun
32,33 Overview
128, 129 Tropical countries and places underneath the Sun

Sun Your Best Reference Direction:

4 Imagine the Sun at High Noon — Always
5 Mnemonic: **N**oon Sun Passes **N**orth of My **N**ose Down Under
6 Notice these, Yesterday — to Guess the Time Today
7 Mnemonic: The Southern Summer Sun Sets Somewhat South
8 Mnemonic: “Sun at High Noon — How High?”
8 On Overcast Days, Average the Brightest Sky over 1/4 hour
9 High Noon — How High?
10 Optical Illusion — Beware!
— Use Your Shadow Not the Sun, For Direction
11 The Dark-Blue Polarisation Band Perpendicular to the Sun
11 The Centre of the Rainbow is Down-Sun
12 Your Shadow Rotates with the same clock-sense as the Sun
Shadow-Rise and Shadow-Set is opposite direction to the Sun
12 Mnemonic: Shadow Rise is on the Right; Set is on the Left
12 Mnemonic: Your Shadow Passes South of Southerners
at Southern Noon
13 The Sun Passes Quickly Past North-South Nearer the Tropics
14 Shortest Shadow South—Opposite Highest Noon.
Always Imagine “Shadow-Noon” on your “Shadow South” Side
15 Locate Polaris (& Find N) by Imagining the Centre of Circling
16 “ Noon ” is Not at 12 O’clock
17 The “Equation of Time” Diagram
18 The Sun Oscillates Seasonally between +/- 23½° Latitude
19 Using the Declination of the Sun
20 Tell the Time by the Compass Protractor—Tilt to the Equator
21 Or Judge the Time By Sight—by Mirroring the Sun Around N
(To Make an Hour Hand on a 12 Hr Clockface)
22, 23 Walking in Phase with the Sun, Wind and Moon
24 Tell the Time of Night by the Star Opposite the Sun
25a Eastern Australian Standard Time Zone 10 hours ahead
25 Tell North from The Time By Using a Protractor
26 Longitude and Latitude from the Sun
27 A Bush Theodolite for Overnight Equal Height Observations
28 Find North by your (non-digital) Watch-Hands
29 Find Any Other Direction Too — Using the Hour Hand
29c String North
30 Locating Anti-sunrise Shortest Shadow Souths
30 Sun Path Norths — Summary
31 The 2 Flaps Method for N.—Lat. & Dec. Angles Separately
31 Seasonal Adjustments For Equinoctial East
32, 33 Overview
132, 133 Eastern Australian Standard Time Zone
10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun

Stars

4 The 4 Bright Stars of Orion Point out Where the Pot is Going
4d The Ecliptic
5 The Two Brightest Stars Point out the South Pole Star
6 The Spinning Earth Makes Circular Paths for the Sun, Moon
7 Stars Leave Star-Trails
8 RISE and SET Directions for a (Local) STAR COMPASS
9 Global Star-Compasses e.g. ‘Perpendicular to’ rise and set
10 Right Around the Horizon, Stars Move to Your Left, Down S.
11 Identify the Stars That Pass Directly Overhead
12 Ways to Locate the South Pole Star Or the N./S. Meridian
12a&13a Global Orientation: the South Pole Star, and Polaris
13 Pole Stars—“(Latitude)° above and below, South and North”
13 South Pole Star Chart
14 Find the Polar Axis by Star-Trail Bisection
15 The Eagle Clips the Equator, South of Altair
16 Equatorial Stars Rise and Set Due East and West
17 Rigel-to-Capella is a good S./N. Pair
18 Use the Southern Cross as an Hour Hand
18 Picture the Southern Cross during the Day
19 Perspective Distortion — Use the Correct Vertical
20 The Southwest Corner of the Great Square is N. S. E. W.
21 The Southern Summer Rectangle And a Triangle Pointing S.
22 Navigating by the Shape of the Night Sky
23a Latitude without Numbers
23 Matching Sky to Sky — Move Only Your Horizon
24b Stars Opposite the Sun
24 2 Different Back-and-Forth-Star-Directions for a Position Fix
25 For Land-Navigation, the Plumb-Line Star-Pair Method
26 Equal-Height Stars When the Sea Horizon is Visible
27 Great-Circle Reference Star Directions
28 Leo the Lion Looks like a Lion or Sphinx Squatting
It Has Four Lines which Show N./S. E./W.
29 String North Latitude without Numbers
30 “Star-Time” Gains Four Minutes a Day
31 Working the Star-Longitude Numbers
32, 33 Overview
118,119 Appendices: Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
The Southern Figure of Eight
The Zodiac
122 A List of Star Coordinates
124 Drawing Great Circles
125 Calendars in the Sky

Moon

4 Look for the Line Joining Sun, Moon, Planets & Zodiac Stars
— the Ecliptic
6 The Moon; Rule 4: Look for It
8 From the Moon, Locate the Sun-Moon Direct Line, Carefully.
10 *Mentally Rotate the Sun-Moon Line Around Polaris*
12 “Cusp North” — Approximate Direction from the Moon
14 The Rabbit Looks up Ahead at Full Moon Using the Ecliptic
16 The Moon Cusps Push and Pull the Hour Hand Backwards,
Down South
17 Watch the Clouds Cross the Moon
18 The Moon is Lit By the Sun
20 The Moon Rises and Sets 50 Minutes Later Each Night
22 The Moon is “in Phase” with the Sun — Watch for It
24a The Moon Moves Over the Tropics
24 Predict the Full Moon Path According to the Time of the Year
24d Finding North from the Moon, by its height & phase
25 Predict the Moon Phases by “ N minus M ”
28a Find E/W By Bisection of Sun & Full Moon At Rise or Set
28 To Predict the Moon’s Path, Visualise the Ecliptic
29a Find North from the Combination of the Sun & the Moon
29 A 4 Week *Position Cycle* of Moonrise/set, to Find E/West
29e Keeping Track of the Moon’s Motion along the Ecliptic
30 Adjusted Moon-Cusp North:
To the Left or Right, By Moving the Sun
32 Overview: The Moon
116,117 Appendix: Moon-Phase & Tide-Time Dial

Wind See also pp. 60a-68a

5 Mnemonic: “Wind? Weather Map!”
7 Mnemonic: “Wind in Your Face”
9 ‘Low’ Clouds Come Clockwise Down under
11 Weather Fronts Distort The (Circular) Wind Circulation
13 Wind-Direction-Change Rule for Buys Ballot’s Law
13 Steering Winds
15 Atmospheric Pressure Changes.
17 Watch the Clouds Cross the Moon
19 To Determine Upper Wind Direction from the Clouds
21 Clouds and Con-trails and Their Winds
23 Prevailing Winds on Creek Banks Tips for calm weather
27 Make a Bush Psychrometer
31 Topographic-Level Winds
33 Overview: Global Weather Patterns

52e The Aurorae
53e City Glow

Clockwise See also Map & Compass Work

4 Mnemonic: “Clock”wise Northern Hemisphere SWiNES
6 “VEERING” with the Sun is “Clockwise”
only in the Northern Hemisphere
“BACKING”, against the Sun’s movement, is “Anticlockwise”
— but only in the Northern Hemisphere.
8 Mnemonic: (when you say) “COMPASS” (say) “CANADA”
10 “Bears Circle Clockwise” Compass Bearings are Clockwise,
A Full Circle, and Eurocentric, Cee?
12 Mariner’s “Compass Points” work on Successive Halvings
14 “E. W. N. South” Of What!? Rule: Say What
16 Back-Bearings Killed Sleepy Pilots
18 Back-Bearing: Plus or Minus 200° Then Minus or Plus 20°
20 Turn 90°? “+/-100° -/+10°” !
22 The Clock-Face Method for Direction
24 One Degree is about One in Sixty
26 Anticlockwise is Your Best Spin-Reference Direction
28 Bearings & Tracks: “If it is right, it is big in my sight”
30 “By the Nose” Navigation
32 Overview

Anticlockwise

5 Mnemonic: Down Under, The Sun Moves ANTI-BigBENWiSE
i.e. anticlockwise
7 Southerners Need Anticlockwise Systems
9 A Compass-Needle Continent—Swinging the Continent
11 Interpret a Compass Needle as “RED HOT—WHITE FROST”
13 Four Reference Directions—Name Them “Rise, Noon, Set, Midnight”
15 “Rise-Noon-Set-Midnight” carries its meaning across the Equator!
N.S.E.W. doesn’t.
17 The 0, 1, 2, 3, 4 Reference Directions
19 The Sunny Skies Quadrant System
21 The “SWEETHEART” Method of Orientation
23 Mnemonic: “North; Coast”
23 Go Overseas!
27 Walk across Australia! English!
29 South is a Triply Better Reference Direction than North, in the Sth
31 Nose-Navigation Needs an ‘Anti’ Mnemonic: “Nor-West, Not-Least”
32 Overview
Revision, Overview and Explanations pp. 32,33
Top Down Orientation
Time is counted in cycles
Closer to Home
The Precession of the Equinoxes
The Moon
Global Weather Patterns
The Moving Horizon
The Spherical Earth
The Spinning Earth
Moving the Horizon

Section 2. THE INSIDE STORY—Mental Manoeuvres pp.34-55

Learn

34 Navigation is More Than Navigating Common Sense
36 The “Good” Navigator. You.
38 Don’t Just ‘Get By’ and ‘Make Do’
40 Walk Where the Navigators Talk — if you want to learn navigation
40 Explore It’s No Fantasy
42 Self-Motivation is Fundamental
44 Routine Mind-Shut-Down Procedures
46 Navigation is Mental Work — Stay Alert
48 To Work Harder, Rest Harder
50 A Three-Track Mind
52 Concentrate Outwards
54 Come to your Senses

Think; Interpret

34 Your Subconscious is Superior to Science!
36 Think out Loud
38 Memories are Made of Landscapes, Not Compass Needles!
40 Feed Your Subconscious through Your Conscious Focus
42 How to Feed Your Brain
44 Feed Your Subconscious Subconsciously
46 Instruments Instruct the Intelligent
48 Minds Work Best with Intelligent Questions on Them
50 Interpret!
52 SPELL Out Numbers — Into Words
52 PAINT your Words — Into Pictures
54 TRANSFER Your Pictures into Reality
54 Put That Experience in Your Memory Bank
55 Extend Visible Reality into the Invisible Picture

Angle:

34 Measuring angles at 1 cm per degree
35 Point and Blink = 10% Shift
36 Guesstimating Angles at 1 degree per cm
37 Distance using Apparent Size;
37 Visual Angle, Size and Distance & 6°; Side-wise Offset
(38, 39 Grid-line Walk & Mental Grid Maps)
40 Cyclical Motion Divided into 12ths
41 Approximate Sine-Wave Values

Look

34 Visual Assessment is Your Main Clue — Rely on It More
36 You Need Excuses to Stop and Have a Good Look
38 You Need a Road-Map To Climb a Hill!
40 Look into the Far Distance Look through the Haze Wait
40 Wait for a Change in Perspective...
42“Retracing is Re-preciating”
44 Look Back Regularly. Look Around
46 Look Through the Vegetation, to the Contours
48 Play the Field-Scientist
50 Look at the Ground
52 Look for Foreshortening & Overlap & Parallax
54 Look, with Vengeance Aforethought

Integrate

3 Visualisation is your Key Skill
34 Integrate: Jetstream, Earth Rotation & Heaven Rotation
36 Visualise Familiar Landscape Trends, Under the Sun
38 ‘Land to Sky — Sky to Land’ Navigation
40 Join the Invisible Big Picture To the Visible Landscape
42 Landmark to Landmark Network Navigation
44 Aim to Integrate…
46 Imagine an Aerial Photograph of the Continent
48 Connections to the Outside World
—You don’t need to feel isolated
50 Orientation is More Fundamental Than Position Sense
52 Make Common Sense of the Various Direction Clues
52 Don’t Degrade the Point of a Clue or System
54 Preparation is the Key to Difficult Navigation
42e Moveable Position vs Fixed Orientation
43e Name Each Clue — with solid ground
44e Notice and Match Up
45e In-Line Integration
50e Think in Pairs, in lines, trends and tracks
Orientation 35 Orientation = Direction Sense
37 The Get-Back Path is the Single Fundamental
38e Grid-Line Walk
39 Disorientation is More Dangerous Than Being Lost
39e Mental grid Map
41 Which Way Did I Come In? Which Way Do I Go on?
43 Disorientation — A Killer Syndrome
42e Movable Position vs Fixed Orientation
45 Un-Dis-Oriented;
47 Avoid Disorientation: OverNavigate
46e Making a M.U.D. M.A.P.-m.a.p.-m.a.p.
49 OverNavigate: Equals Over-Familiarise
51 Needle N, S, E & W. Identify them on the Landscape
51 Orientation to Start-Finish Rhumb Line—by sight or compass
51 Orientation to Your Finish Point, By the Parallax Method
53 Orientation to Your Start Point, By the Parallax Method
55 Look Backwards, Think Forwards

Disorientation

35 The Navigator’s Top 10 Problems; and Band-aid Solutions
37 How I Manage to Get Lost or Disoriented
39 Logical Blunders — Examples
41 Logical Presumption Leads to Utter Delusion
43 Casual Outings Catch You Out
45 Trapped! By Not Looking Back
47 How to Get Lost—Follow the Track—A Track is a Trap-Door
47e Disorientation in new territory
49 How to Get Lost—Follow the Leader
51 Exponential Anxiety — the 11:59 syndrome
53 Nervous Nellie Navigation — Not

Reorientation

35 Reorientation
37 Voice Your Thoughts — Don’t Suppress Important Clues
39 Trust the Untrustworthy
41 Trust Less and Prove More
43 Trust Yourself Less
45 Disoriented? Stop. Say So. Reorient!
47 Reorientation; And a Fresh Start
49 “As Solid As the Rocks in My Head”
51 Explore Junctions, While You Can
53 Look for the Essentials — the ‘Directions’
55 To Get There Quicker, Go Slower

Lost?

35 Lost? Rule One: Stop! ..getting More Lost
37 Back-track in Your Mind before You Back-track on Your Steps
39 Rule 2: Start! Getting Unlost. Find yourself. Regain confidence
38e Grid-line Walk — Explore in a Square Grid Pattern
39e Mental Grid Map
41 The Simplest Way to Get Unlost is to Retrace Your Steps
43 More is Better — Thought, Time Care
45 Lost? Rule 3: Do a Good Job—of walking out or staying put
47 How to Leave a Message
49 Lost? Rule 4: DON’T Do It Again!
51 Separated? An Agreed Plan Helps
53 A Whistle-Code for Contact Calls
55 Search and Rescue Signals — Noises Repeated Regularly
120 Appendix: Morse Code and Rhythm Mnemonics for it
121 Appendix: Semaphore Code and Mnemonics for it

Section 3. REALLY INTERESTING STUFF
—Bush Navigation pp56-89 See Also 'Wind'

Nature's Norths

56 Nature's Norths
58 Rock Noon
60 Wind West See also Wind, Section 1
62 Think Wind
64 Be Wind-Wise: Listen
64 Obvious Tell-Tales: Windblown Objects
64 Obvious Tell-Tales: Landscape
66 Light & Colour Effects
68 Botanical Tell-Tales
70 Combined Effects of Wind, Salt and Light
70 Botanical Asymmetry
72 Think like a Plant
72 "Mosses Mop up Moisture" "Lichens Like Light"
72 "Fern Fronds Face the Front" i.e. the sunny side
74 Bush Noon— The Sun is Always out, If You Know How to Look
74 False Asymmetry & Symmetry True Symmetry & Asymmetry
76 Higher Density of Leafiness on the Noon Side
76 The Lowest Leafy Branches are on the Noon-Side
76 "It's the Shady Side That Shrubbery Shows up"
76 Tall, Steep and Ragged Backside Wedge-Shaped
76 Hollow under the Shady Side & Imperfect Shape
76 Gaps on the Shady Half Dead Branches
78 'Solar-Panel' Tilting of the Canopy Top
78 Wedge Shaped Lopsided Lean Half-Dome One-Sided Leafiness
78 Individual Branchlets Bare Backbone
80 Branching Patterns
—Examine Them for thickness, length, angle, curvature
82 Isolated Vegetation Shows the Clearest Asymmetry
82 Closed Canopy — Look for Gaps; Saplings
82 Roadsides & Campsites Reverse Some Effects!
84 Flowers Respond to Sunlight Fruit Distribution and Ripening
86 Stem and Leaf Colouring New Growth is Colour Differentiated
88 Noon-Sunlight is More Vertical
88 The Shedding of Bark on the Sun-Baked Side
88 Where to Look for Fading or Lack of it

Memory & Naming Skills See Sect.5, Navigation Mnemonic

56 Don't Forget to Remember, Your Way!
56 Unlock Memorable Chapters in Your Life
56 Go for the Jugular. Zero in on What Matters
58 Don't Try to Memorise Details without an Overall Framework
58 Don't Try to Remember Anything You Haven't Memorised
58 Write down What You Memorise!
60 Anything is Memorable
60 Make Individual Memories Memorable!
60 Remembering Trees
62 The Difference between Looking at, and Seeing
62 Search-Image
62 We Recognise 'Sameness' But Need to Describe the Differences
64 Place Recognition
64 You Recognise What You See, By What You Can't See
64 Description Requires Words
64 Can You Describe All That You Can Recognise?
66 Discuss It
66 Collect the Details
66 Identify a Specific Place by its Specific Collection of Details
66 Include What You Can't See
68 Extend Your Familiar Range Systematically Hints
68 Leap-Frog Walk in New Territory
70 Draw the Skyline, the Ridge-lines, the Outlines
70 Label it with the 8 main directions
72 Examine the Skyline BEFORE You Walk up to It
72 Make a Mental Movie
72 Take Mental Photographs
72 Take a Mental Photo of Each Change of Track
74 Remembering the Sequence, Needs More Than Logic
74 Do Something NUMERICAL
74 Match the SHAPE
76 Use a Type of Memory which you are Good at
76 Use the Name of the Track for the Acrostic
78 Group Several Related Images into One See also p. 110
80 Working Names — Invent Them
80 Standard Name Plus Personal Name
80 Remembering Names
82 Name the Subsections
82 Name the Sequence
84 Revise your Memories To Remember Formulae
86 Remember the Sequence — from AA to ZZ

Lie of the Land

57 The "Lie" of The Land
57 Listen to the Pioneers Lie Has To Do with Water Drainage
59 Major Rivers & Ridges vs Minor Ones Sort it out on the Map
59 Catchment Divisions Snake from Coast to Coast
59 A Catchment-Division Separates Whole Valleys
61 Ridges Ridge Names Mnemonic: "Ridges Rise"
61 Uphill — Which Hill? Upstream — to which Divide?
63 Shoulders Stop. Shoulders are Not Easy (like Ridges) to Walk on
63 Saddle — A Flat between Hills, Sitting Astride a Ridge
65 Notice When You Cross a Catchment!
65 The Giveaway! A Change of View—Mnemonic: "Catch the View"
65 You Can't Cross a Creek Twice from the Same Side
67 Break of Slope/Make of Slope Scree Slopes — Tread Carefully!
69 Break of Slope — And Route-Finding
71 Mnemonic: "Down in the Mouth" Which way is the flow?
71Hint: Don't think "this side" & "the other side".
73 The Inside of a River Bend is Flatter Ground
73 Rivers Bend! Don't Forget to Look Backwards
73 Rivers Meander. Ridges Run Straighter
75 Land-Form. Hint. What Makes a Cliff and an Overhang?
77 Headlands, Bays, Prevailing Winds, Longshore SandDrift, Creeks
77 Beaches — Slope, Particle Size, and Wave Energy
79 Rock Typing Soil Typing

Contours and Slopes

81 Contour Lines
81 Contour Navigation: Practise It, Before you Need It
83 Contour Curves: Your Job is to Imagine the Shapes into Life
And to Match the Lines to the Landscape
85 Contour Heights: Read the Numbers!
85 Contour Spacing: Close = Steep Slope
85 Contour Slopes: Work Out 1 Mm and 1 Cm Contour Spacing
85 Reference Slopes — To Imagine the Climb with!
85 The Beaufort Wind Scale — Modified for Bushwalkers
87 Make a Clinometer — a Slope Measurer
89 Just Draw It

Tracks and Bends

57 Stick to the Track Reasons for Leaving the Beaten Track
57 Cross-Country Routes
59 Tracks are like Naughty Children
59 The same is true for creeks & ridges
59 Explore Branches While You Can
59 Convergent Tracks Lead to Civilisation
59 Down-Track vs Up-Track
61 Sympathy and Understanding with Track Markers
61 Old Track Markers
63 Track Markers When You Lose the Track-Markers
63 Notice (with Alarm!) When...
65 Mark Your Trail Mark Your Trail Non-Permanently
65 Tracking Someone Else
67 Mark Your Trail More Permanently and Ostentatiously
67 Mark Your Trail Invisibly
67 A pebble Code — Keep It Simple

Bends

69 Keeping Track of the Bends Desired Equipment
69 Logging Constant Diversions
71 The Difference between 'Direction' and 'Track'
71 Continuous bends
71 Follow the bends by anticipating the trends
73 Keeping a Compass Course, by Legs
73 Turn Smooth Bends into Sharp Jumps
73 Mind Which Way You Go
75 "End the Bend" i.e. Notice the Un-Bends
75 Name the Bend, By the Trend, of its End
75 Notice and Name Perverse Bend
77 Look Through the Trees, to the Sky and Skylines
77 Periscope Navigating by Trees
79 Bush networks Don't Navigate by Networks Only
79 Antidotes For Guessing at Network Junctions
81 Sticking to the Bush Network Natural Routes Naturally Converge
81 Coping with Divergence Reverse logic
83 Trackless Navigation The method of Names
83 Bush Networking Essentials
85 Walking "Straight" through the Bush Straight around
85 How to Walk In an Unintentional Circle
87 Walking an Exact Straight Line in Open Country
87 Leap-frog a Straight Line by Looking Ahead
89 Radio Direction Finding—As Good as a Compass

Section 4a. DISTANCE **pp. 90-108** left-hand pages
See also 'Angle'
Progress
90 Distance Guesstimation is a Major Problem for Walkers
92 Use the Map Scale—Don't Just Look at It The Jargon
92 Map-Scaling-Factors are Meant to be Interpreted
94 Psychological 'Half-Way Point' When You Meet Other Walkers
96 AIMING OFF by 10° — 10° is **175m per kilometre**
96 Aiming-off at Sea; Aim Off 10% or 6°
98 Up-down Progress
100 Distance by Double Pace—Test Yours Tie Your Ankles Together
100 Don't Lose Count! Use Finger-Math
100 Test your pedometer
102 Line of Sight Can Be Extremely Accurate
102 Two Opposite Landmarks
104 "Changeover Lines" of Progress — Relative Apparent Sizes
104 Use the Landscape Itself to Measure & Mark Progress
106 The Sine-Cosine Tally Method of Keeping Track, Blind
106 Rise and Run Adjustments — Sine and Cos — North and East
108 Progressing A Position Line across the Map
108 Pinpoint "Fixes" are a Myth!
Pace
90 Distance = Speed x Time Keep a Log! Adding two Speeds
92 Pace Your Progress — Pre-Dict It, Don't Post-Dict It!
92 Advice about Walking Times — Double & Halve!
92 20 Minutes per Kilometre — Double & Halve — Easygoing Maths
92 4 km per hour — Normal Walking Pace — Double and Halve!
94 1 Metre per Second is 3.6 kph "1 Hour Extra per 5 Hours"
94 "1 Hour Extra per 500 m easy climbing"
96 Adjust Your Expected Pace
96 Don't Overcompensate for Slopes And Meanders
98 Don't Lose Count! Use Finger-Math
100 A Slow Kilometre Cancels out Three Quick Ones
100 To Progress Faster More Easily Pace Yourself
— Slow Down, but Don't Stop
102 Distance-Off by Vertical Angle Safe Distance Off
104 Circle-of-Position by Horizontal Angle
105b A Horizontal Angle between Features Yields a Position Circle
105c 'Two Horizontal Angles' Usually Fixes Your Position
106 Rise and Run Adjustments — Sine and Cos — North and East
Range
24e One Degree is about One in Sixty
34-36e Estimating angles
37e Distance using Apparent Size & 6°
37e Visual Angle, Size and Distance Side-wise Offset
51a & 53a Parallax methods
90 Range: Rule One: Guess
92 "Range" of Visibility — Two Halves to the Picture
92 "Range Squared is Dee Height"
92 "Range Squared is Dee Height"
94 Double, the Acute Angle, off Your Nose — For the Distance Off
94 Halve, Any Angle, from your Stern—For Distance Off
96 Distance by Optical Rangefinder
96 Range by Parallax and a Piece of String DIY Rangefinding
98 Distance by averaging maximum & minimum Guesses
98 Distance by Comparison with Some Length You Know
98 Distance by Cloud Shadows — Look for Them
98 Sound Travels Three Seconds to the Kilometre
100 Lighting Conditions & Eyesight
100 Visual Acuity is about One Minute of Arc
102 Approximate Figures for Visual Acuity
102 Test Your Own Visual Acuity
104 Circle-of-Position by Horizontal Angle
105b A Horizontal Angle between Features Yields a Position Circle
105c 'Two Horizontal Angles' Usually Fixes Your Position
104 The best range-finder
106 Rule of Fingernails Rule of Tom Thumb Apparent Size

Section 4b. MAP and/or COMPASS pp. 91-109 right-pages
Map Use
54e Helpful Chart Symbols 122 Appendix: Symbols
38,39e Grid-line Walk & Mental Grid Maps
46e Making a M.U.D. M.A.P.-m.a.p.-m.a.p.
91 To "Read" a Map — Rule One: Don't Lose Your Place!
91 Rule Two: Imagine Arrows for the Sun, & Your Movement
91 Rule Three: Get Used To Any New Map Scale
91 Rule Four: Don't Give Up!
93 Look Carefully at the Fine Detail on the Map
93 Look Carefully at the Fine Print
93 Some Map Details are Not to Scale!
95 Permanent Landscape is More Trustworthy Than Surface Detail!
95 Details Added on to a Map Are Prone to Error
95 Expect The Map To Be Unhelpful
95 Maps Do Give You 10 Times More Detail Than You Can Notice
97 "Maps Grip the Universe"
99 Align the Track
101 Align the Map the Landmarks
103 Align the Map to Identify Landmarks
105 Align the Map to Fix Your Position
107 Which Way is North?
Magnetic Compass Use See also "Clockwise", Section 1
37e Chart symbols
91 "L.M.N.Or P."
93 Using the Protractor by Itself
95 Using the Needle Alone — Emergency Magnetic Compass
95 Hint Understand that magnetic poles should be called
'North-Seeking' and 'South-Seeking'
97 Think "+/- 5°" For Compass Bearings "+/- 10°"
For Following a Compass; "+/- 15°" For Using a Compass!
99 Compass Errors — Be Impressed! Be Careful!
101 Compass Use — Master It, but Be a Good Master.
103 "Magnets Grip Maps"
103 Align Your Map in the Dark
107 'East-least; West-Best' Rule—Tells You Which Bearing is 'Best'
108b Getting Started in the Dark
108c Thinking Backwards
109a Taking a Bearing from the Landscape
109 Plotting a ('Compass-to-Map') Back-Bearing
109 No-Number Back-Bearings
124 Appendix: Map Symbols
Mapmaking
38e Grid-line walk
39e Mental Grid Map
46e Making a M.U.D.M.A.P.-m.a.p.-m.a.p.
91 Maps Mean: Imagine Me! — Making Mental Maps
91 "Map To Mind" Navigation
93 Mental Networks
93 The Most Valued Map
95 Strip Map Navigation — Follow-Your-Nose Navigation
97 Caving Strip-Maps — Come in Books
99 Map Folding and Map Holding
101 Layout Map — Blind
103 A Compass Holder
106a The Sine-Cosine Tally Method of Keeping Track, Blind
106b Rise and Run Adjustments—Sine and Cos—North and East
107 String Map
109 Parallax While You Walk — JUST DRAW IT!

Section 5. Navigational Mnemonic **pp. 110-115**
78b Group Several Related Images into One
110 A "Start-Stop-Rest" Navigation Mnemonic
110 P.R.E.P.A.R.E.D.— A Before-You-Go Checklist
111 Equipment—A What-You-Might-Like-to-Take Checklist
111 Place-Names; Learn Them; You Navigate by Local Features
111 S.T.A.R.T. off Oriented
111T.I.M.E. & Distance — that's the point
111 Distance
111 Progress; Pace; Range
112 Look in These 6 Directions — in 6 ways — Up and Down
112 4 More Slopes
112 Look in These 6 Directions — in 6 ways — Ahead and Behind
112 Look in These 6 Directions — in 6 ways — Right and Left
113 ..Plus in Two More Directions
113 (1) All, Around, Appreciatively, And At 6 L's
113 (2) Look at the Map: a Systematic Checklist. What 'System'?
113 Mark the Lines-of-the-Land On Your Map
113 Mnemonics for Map and Compass Work
114 N.S.E.& W.—the 4 cardinal directions, is a mnemonic of fours
114 W.E.A.T.H.E.R. — Check for Change
114 Stop-Signal: Why have you stopped?
114 REST and R.E.O.R.I.E.N.T yourself
115 RI For Revisualise the Invisible & then Relate It to the skyline
115 Clock-Sense Mnemonic
115 ENT for Environmental Notice Time—Orient yourself to Nature
115 REMEMBER to Note down the Names in the Notebook
115 Re-lax, and Play — Navigation Games and Activities

Miscellaneous **pp. 34e - 55e**
52e The Aurorae
53e City Glow
54e Helpful Symbols for Standardising your Chart–Work

Appendices **pp. 116-121**
116 Moon-Phase & Tide-Time Dial
118 Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
118 The Southern Figure of Eight
119 Divide the Star Globe into 8 Octants
—and Learn the Stars in Each
119 The Zodiac
120 Morse Code & Rhythm Mnemonics
121 Semaphore Signalling Code & Mnemonics for it
122 Symbols for Track-Notes and Mud-maps
122 A List of Star Coordinates
128, 129 Tropical countries and places underneath the Sun
124 Drawing Great-Circles. The PZX navigational triangle
125 Calendars in the Sky
126,127 About the Author; Outline
Index pp. 128-129
Contents p. 2

Eastern Australian Standard Time Zone

10 hours ahead of Greenwich
Showing Tropical countries and places underneath the Sun

p24a,25a

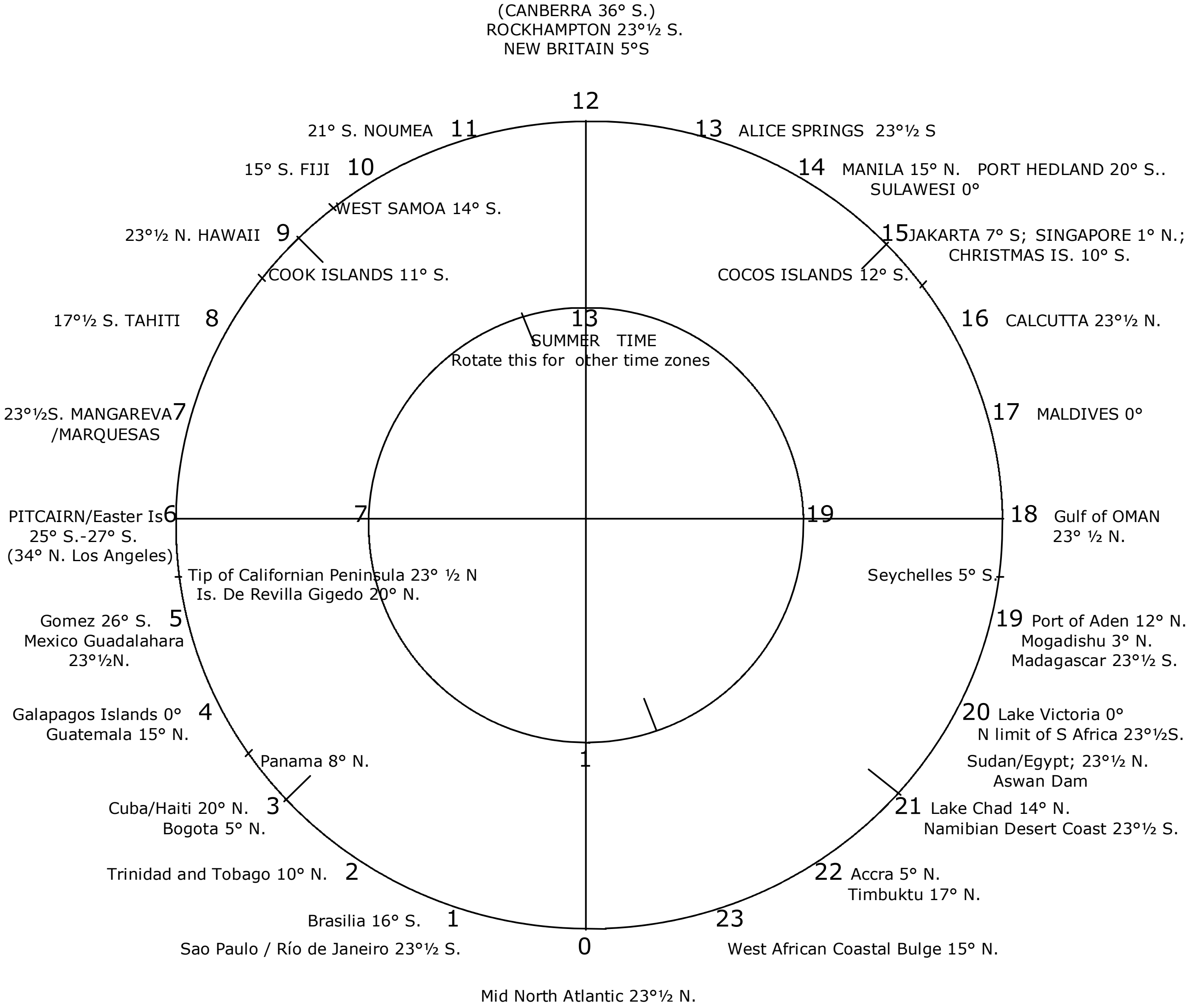
(CANBERRA 36° S.)
ROCKHAMPTON 23°½ S.
NEW BRITAIN 5°S
12
21° S. NOUMEA 11
13 ALICE SPRINGS 23°½ S
15° S. FIJI 10
14 MANILA 15° N. PORT HEDLAND 20° S..
SULAWESI 0°
WEST SAMOA 14° S.
23°½ N. HAWAII 9
15 JAKARTA 7° S; SINGAPORE 1° N.;
CHRISTMAS IS. 10° S.
COOK ISLANDS 11° S.
COCOS ISLANDS 12° S.
17°½ S. TAHITI 8
13
SUMMER TIME
Rotate this for other time zones
16 CALCUTTA 23°½ N.
23°½S. MANGAREVA 7
/MARQUESAS
17 MALDIVES 0°
PITCAIRN/Easter Is 6
25° S.-27° S.
(34° N. Los Angeles)
7
19
18 Gulf of OMAN
23° ½ N.
Tip of Californian Peninsula 23° ½ N
Is. De Revilla Gigedo 20° N.
Seychelles 5° S.
Gomez 26° S. 5
Mexico Guadalahara
23°½N.
19 Port of Aden 12° N.
Mogadishu 3° N.
Madagascar 23°½ S.
Galapagos Islands 0° 4
Guatemala 15° N.
20 Lake Victoria 0°
N limit of S Africa 23°½S.
Panama 8° N.
1
Sudan/Egypt; 23°½ N.
Aswan Dam
Cuba/Haiti 20° N. 3
Bogota 5° N.
21 Lake Chad 14° N.
Namibian Desert Coast 23°½ S.
Trinidad and Tobago 10° N. 2
22 Accra 5° N.
Timbuktu 17° N.
Brasilia 16° S. 1
23
Sao Paulo / Río de Janeiro 23°½ S.
0
West African Coastal Bulge 15° N.
Mid North Atlantic 23°½ N.

Places in capitals are less than 6 hours away from Australia. The others are on the other side of the world.
View this as if from above the North Pole—the Sun ‘moves’ clockwise, but the Earth spins anticlockwise.

Eastern Australian Standard Time Zone

10 hours ahead of Greenwich

Showing Tropical countries and places underneath the Sun

Cut this one out and post it on your wall until you have learned most of it.

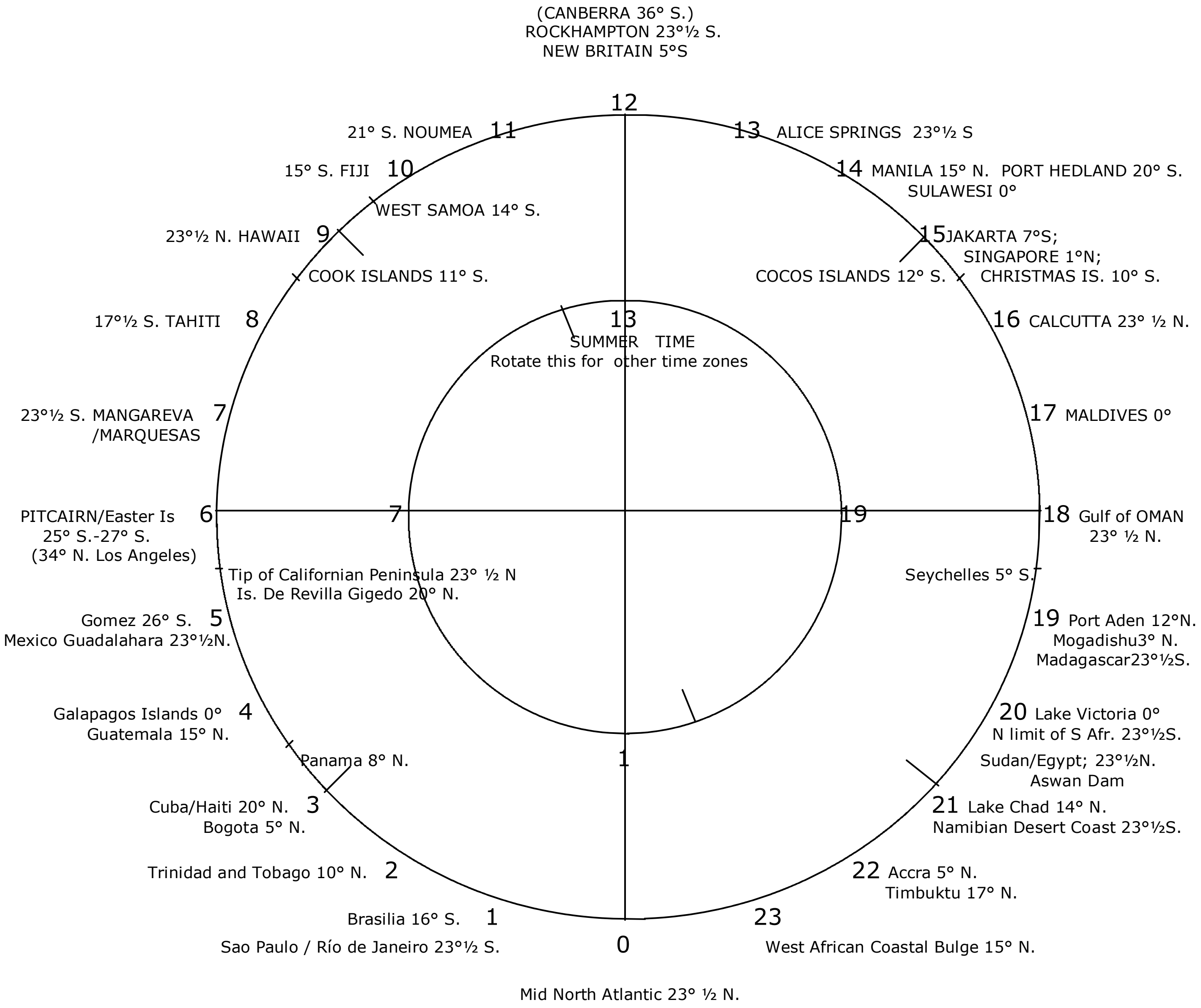
(CANBERRA 36° S.)
ROCKHAMPTON 23°½ S.
NEW BRITAIN 5°S
12
21° S. NOUMEA 11
13 ALICE SPRINGS 23°½ S
15° S. FIJI 10
14 MANILA 15° N. PORT HEDLAND 20° S.
SULAWESI 0°
WEST SAMOA 14° S.
23°½ N. HAWAII 9
15 JAKARTA 7°S;
SINGAPORE 1°N;
COOK ISLANDS 11° S.
COCOS ISLANDS 12° S.
CHRISTMAS IS. 10° S.
17°½ S. TAHITI 8
13
SUMMER TIME
Rotate this for other time zones
16 CALCUTTA 23° ½ N.
23°½ S. MANGAREVA 7
/MARQUESAS
17 MALDIVES 0°
PITCAIRN/Easter Is 6
25° S.-27° S.
(34° N. Los Angeles)
7
19
18 Gulf of OMAN
23° ½ N.
Tip of Californian Peninsula 23° ½ N
Is. De Revilla Gigedo 20° N.
Seychelles 5° S.
Gomez 26° S. 5
Mexico Guadalahara 23°½N.
19 Port Aden 12°N.
Mogadishu3° N.
Madagascar23°½S.
Galapagos Islands 0° 4
Guatemala 15° N.
20 Lake Victoria 0°
N limit of S Afr. 23°½S.
Panama 8° N.
1
Sudan/Egypt; 23°½N.
Aswan Dam
Cuba/Haiti 20° N. 3
Bogota 5° N.
21 Lake Chad 14° N.
Namibian Desert Coast 23°½S.
Trinidad and Tobago 10° N. 2
22 Accra 5° N.
Timbuktu 17° N.
Brasilia 16° S. 1
23
Sao Paulo / Río de Janeiro 23°½ S.
0
West African Coastal Bulge 15° N.
Mid North Atlantic 23° ½ N.

Places in capitals are less than 6 hours away from Australia. The others are on the other side of the world.
View this as if from above the North Pole—the Sun ‘moves’ clockwise, but the Earth spins anticlockwise.

Understanding, Building, and Using Baluns and Ununs

Theory and Practical Designs for the Experimenter

By Jerry Sevick, W2FMI

CQ Communications, Inc.

Library of Congress Control Number 2003103912
ISBN 0-943016-24-X

Editor: Edith Lennon, N2ZRW
Layout and Design: Elizabeth Ryan
Illustrations: Hal Keith

Published by CQ Communications, Inc.
25 Newbridge Road
Hicksville, New York 11801 USA

Printed in the United States of America.

Copyright © 2003 CQ Communications, Inc., Hicksville, New York

All rights reserved. No part of this book may be used or reproduced in any form, or stored in a database or retrieval system, without express written permission of the publisher except in the case of brief quotes included in articles by the author or in reviews of this book by others. Making copies of editorial or pictorial content in any manner except for an individual's own personal use is prohibited, and is in violation of United States copyright laws.

While every precaution has been taken in the preparation of this book, the author and publisher assume no responsibility for errors or omissions. The information contained in this book has been checked and is believed to be entirely reliable. However, no responsibility is assumed for inaccuracies. Neither the author nor the publisher shall be liable to the purchaser or any other person or entity with respect to liability, loss, or damage caused or alleged to have been caused directly or indirectly by this book or the contents contained herein.

This book is dedicated to my late wife, Connie.

Preface

I have often been asked the reason for spending so much time and effort on the transmission line transformer (TLT). It is obvious that many transformers have been constructed and measured. This whole subject resulted from a work transfer back to our Bell Labs Headquarters in Murray Hill, New Jersey, from our Western Electric Branch lab at Allentown, Pennsylvania.

Since my hobby was amateur radio, I decided to look at short vertical antennas and short radial ground systems. I found the TLT to be an ideal matching transformer for the low resonant-impedances encountered. Furthermore, my new assignment, which included arranging exchange visits between our technical staff and the outside world, gave me the opportunity to have my transformers measured on extremely accurate transmission test sets having an error of only plus or minus 1 millidb!

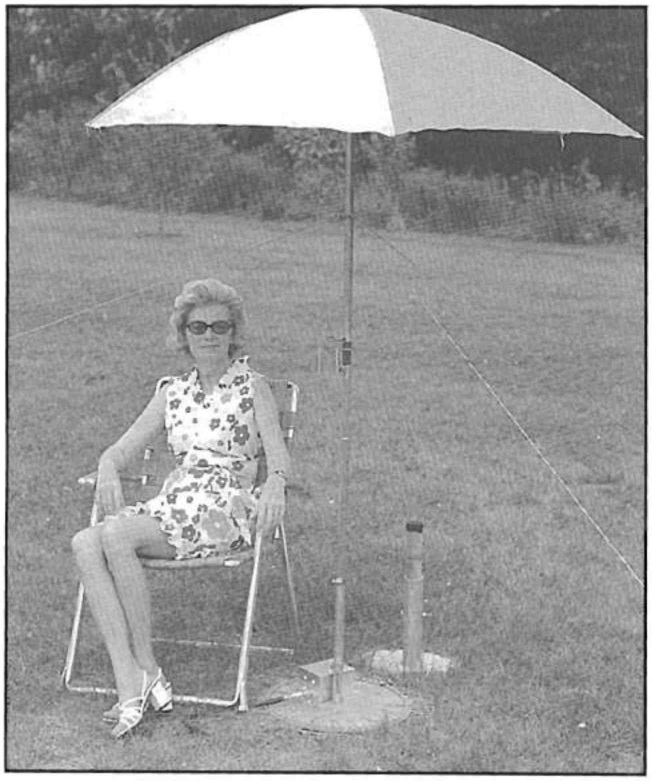

Photo 0-A. *The author's XYL sitting under the W2FMI 40-meter beach umbrella antenna.*

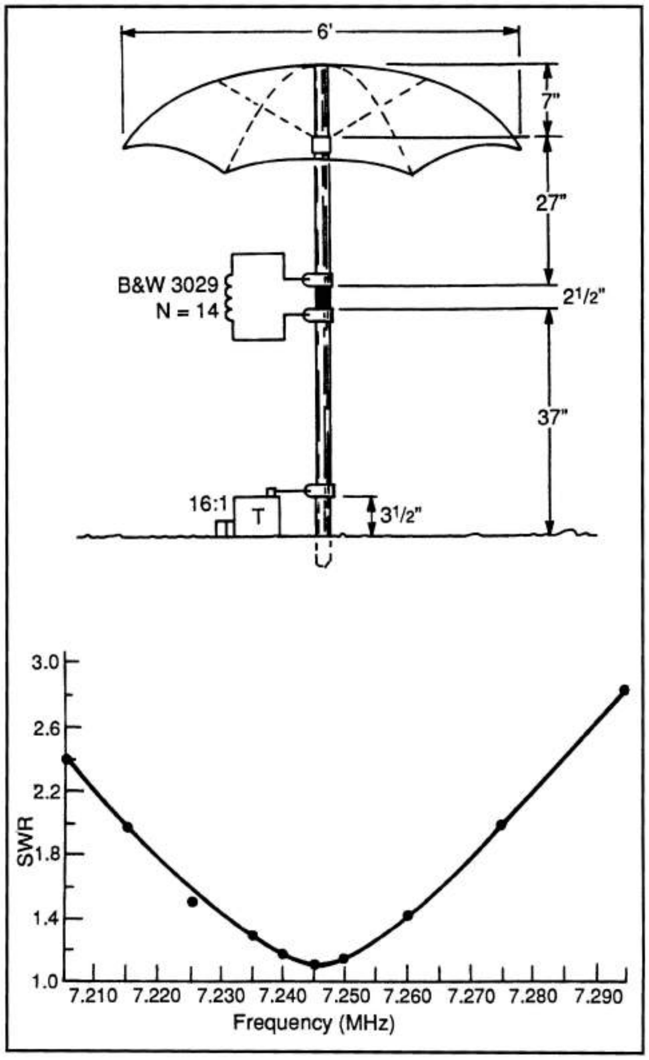

Figure 0-1. *The parameters and standing wave ratio as a function of frequency for the W2FMI 40-meter beach umbrella antenna.*

The ultimate results of my studies for a short vertical antenna in the 40-meter amateur radio band appear to be for a 6-foot-tall version with a large top hat. In shopping at a sports store, I saw the beach umbrella "antenna." It was a lot of fun describing my "beach umbrella antenna" to the radio hams on the 40-meter band. **Photo 0-A** shows a view of my late wife sitting under the beach umbrella resonant in the 40-meter amateur radio band. **Figure 0-1** shows the drawing and the response curve of this interesting, short vertical antenna.

Recently I was asked by Nobel Publishing to prepare a manuscript for a video/CD ROM tutorial on my book *Transmission Line Transformers*.[1]* Looking through this book as well as my second publication *Building and Using Baluns and Ununs*[2], I found that they mainly contain the design and performance of these broadband transformers. What theory they included was dispersed throughout the books and, in reviewing all this information, I found a very simple and unified approach to presenting the theory.

Briefly it involved classifying these transformers according to impedance ratios and applying Guanella's "basic building block" in their designs. In seeking permission from CQ Communications to use some of their artwork for the tutorial, the publisher, Dick Ross, not only obliged but recommended that I update *Building and Using Baluns and Ununs*. He also suggested that I break off the appendices on "Antennas and Ground Systems" so it could stand alone under its own cover, enabling the usual Internet search engines to readily locate it. As a result the two new books will be entitled *Understanding, Building, and Using Baluns and Ununs* and *The Short Vertical Antenna and Ground Radial*.

To understand and apply these transformers, one must be familiar with transmission line theory, conventional transformers, and general circuit theory. Graduate courses in these areas are recommended. Considerable work remains to be done on high-powered TLTs and low-impedance, low-ratio TLTs at frequencies beyond UHF.

In preparing this book I found it necessary to duplicate some of the diagrams, figures, and photos of this book in **Part II**. Since **Part I** is strictly on the theory of the broadband of the transformers, I realized that having some of the figures and photos near the explanations eliminated the chore of searching through half of the book for figures that relate to theory. The figures and photos are still necessary in **Part II**, however, which is on the building and using the TLTs.

I wish to thank four people who were instrumental in making this book possible. They are Dick Ross, Publisher, CQ Communications, Edith Lennon, Editor, and Kathy Collyer and Janette C. Bailey, my associates.

Jerry Sevick
Fellowship Village
Basking Ridge, NJ

---

* *There are separate References for **Parts I and II**. They are found at the end of each part.*

Introduction

The transmission line transformer (TLT) transmits the energy from input to output by a transmission line mode and not by flux-linkages, as in the conventional transformer. As a result, the TLT has much wider bandwidth and higher efficiencies than its conventional counterpart. With proper core materials and impedance levels of 100 ohms or less, bandwidths of about 100 MHz and efficiencies approaching 99 percent are possible today, matching 50 ohms up to 100 ohms and 50 ohms down to about 3 ohms. Future research and development, especially with impedance ratios less than 4:1, should make future TLTs operate at much wider bandwidths.

The transmission line transformer first appeared upon the scene in 1944 in a classic paper by George Guanella in the *Brown Boverie Review* (Brown Boverie is a company located in Switzerland). The title of his paper was "Novel matching systems for high frequencies."[3] The second classic paper was written by Clyde Ruthroff at Bell Labs and was published 15 years later in the *Proc. of the IRE* (Institute of Radio Engineers, which was later changed to the IEEE). His paper was fortuitously named "Some broad-band transformers."[4] In his paper he introduced the 1:4 Unun.

As seen in the **Table of Contents, Part I** includes a new approach to the theory of these transformers. **Chapter 3**, "TLTs with Ratios Less Than 1:4," is my contribution to the technology, and resulted from a need for a broadband TLT with ratios less than 1:4. Essentially, **Part I** includes the history, the theory, and the future of the TLT. The first two chapters of **Part I** present a review of 1:1 Baluns and 1:4 Baluns and Ununs (unbalanced to unbalanced transformers), as presented by Guanella and Ruthroff. Following that is my work on extending their approaches to TLTs with impedance ratios less than 1:4. **Chapter 4** contains information on TLTs with ratios greater than 1:4; in particular, 1:9, 1:12, and 1:16 are presented. Finally, important parameters, such as characteristic impedances for various types of conductors and efficiencies with various magnetic cores, are discussed, followed by a look at mistakes in the literature and the future direction of the TLT.

As was mentioned in the **Preface, Part II** presents the practical side of the TLT. It is essentially an update of my book, *Building and Using Baluns and Ununs*. Those chapters offer complete specifications and numerous photos of most of the devices.

Table of Contents

PREFACE v

INTRODUCTION vii

PART I—UNDERSTANDING BALUNS AND UNUNS

Chapter 1 TLTs WITH 1:1 RATIOS 3

Chapter 2 TLTs WITH 1:4 RATIOS 5

Chapter 3 TLTs WITH RATIOS LESS THAN 1:4 9

Chapter 4 TLTs WITH RATIOS GREATER THAN 1:4 15

Chapter 5 CHARACTERIZATION 17

Chapter 6 SUMMARY 21

REFERENCES FOR PART I 23

PART II—BUILDING AND USING BALUNS AND UNUNS

Chapter 7 THE 1:1 BALUN 27

Chapter 8 THE 4:1 BALUN 43

Chapter 9 BALUNS FOR ANTENNA TUNERS 55

Chapter 10 1.5:1 AND 2:1 BALUNS 65

Chapter 11 6:1, 9:1, AND 12:1 BALUNS 71

Chapter 12 THE 4:1 UNUN 83

Chapter 13 1.33:1, 1.5:1, AND 2:1 UNUNS 89

Chapter 14 DUAL-RATIO UNUNS 97

Chapter 15 MULTIMATCH UNUNS 101

Chapter 16 UNUNS FOR BEVERAGE ANTENNAS 107

Chapter 17 CONCLUDING REMARKS 111

REFERENCES FOR PART II 115

INDEX I-1

PART I—
UNDERSTANDING BALUNS AND UNUNS

Chapter 1

TLTs with 1:1 Ratios

The fundamental classification of the transmission line transformer (TLT) can be more easily categorized by impedance ratios than by Balun or Unun operation. This can be shown by the "basic building block" of Guanella in his classic 1944 paper[3] (see **Figure 1-1**), although he explained only the Balun operation when the center of the load, R_L, is grounded. In this case, Balun operation is assured when the inductive reactance of the top winding is at least 10 times greater than $R_L/2$. In other words, the reactance of the top winding should be five times greater than R_L at the lowest frequency of interest. When the characteristic impedance of the transmission line is equal to R_L, the upper frequency limit is mainly determined by self-resonance for coiled windings and the parasitics in the connections.

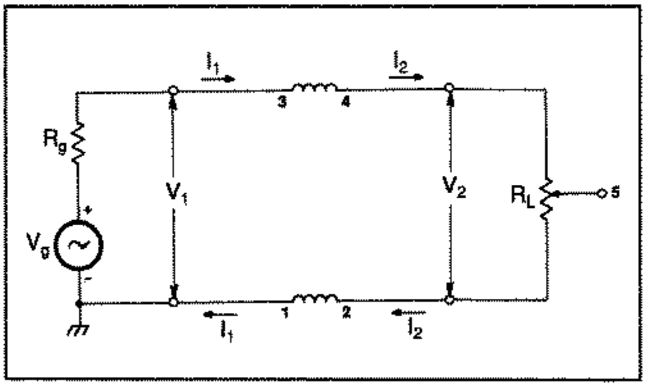

Figure 1-1. *The basic building block.*

With toroidal cores, the inductive reactance at the lowest frequency of interest can be easily calculated. It simply involves the number of turns, N^2, divided by the effective length, which is the circumference of the center of the cross section of the torrid in turn divided by the permeability, times the cross section area. This relationship is then multiplied by a constant, which brings it to the proper unit, called a Henry. With rod cores, there is no simple equation for calculating the inductance—then one has to rely purely on experimentation.

There are two other important functions of the basic building block which were not described in

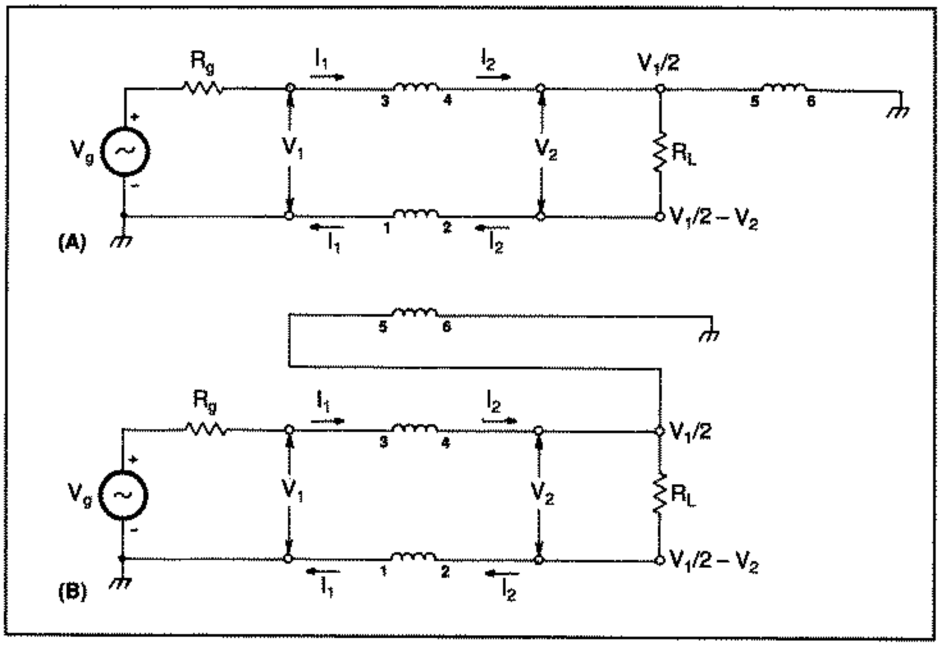

Figure 1-2. *(A) Ruthroff's 1:1 Balun high-frequency toroidal version model and (B) rod version.*

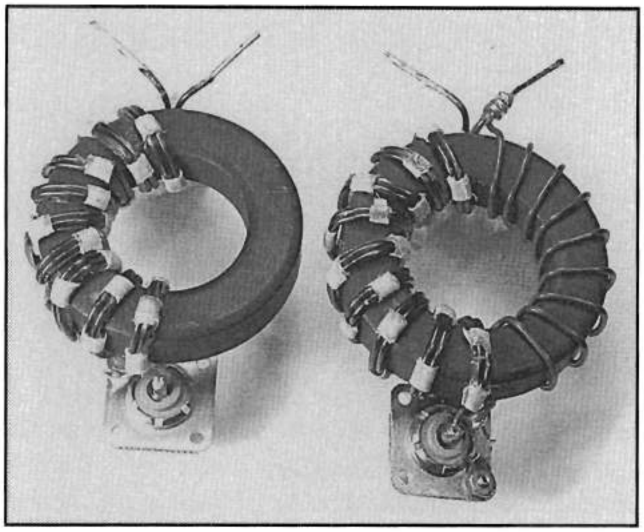

Photo 1-A. *Two basic forms of the 1:1 Balun.*

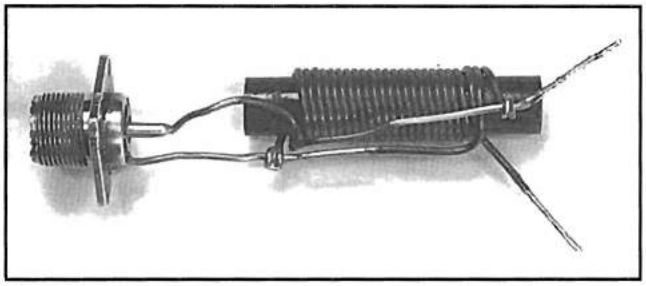

Photo 1-B. *A typical rod-type Balun.*

Guanella's classic paper. They are the phase inverter (when the top of R_L is grounded) and the delay line (when the bottom of R_L is grounded). In the delay line application, the voltage drop along the length of the transmission line is usually zero; therefore, there is no inductance and hence a core is required. Thus, Guanella's basic building block can be considered a 1:1 isolation transformer! In reviewing the Balun operation, the voltage drop is $1/2\ V_1$ and is $-V_1$ for the phase inverter. Because of the apparent isolation properties of the basic building block, it can also be used to change a Balun to an Unun, or vice versa. Since it has been shown that these transformers can exhibit such high efficiencies, especially when operating at low impedance levels, two or three TLTs can be connected in series to obtain the desired results.

Figure 1-2 shows two other versions of the 1:1 Balun which appeared after 1959. **Figure 1-2A** shows Ruthroff's model and **Figure 1-2B** shows a model by some unknown amateur using a rod core instead of a toroid. In Ruthroff's model, we see two inductances in series winding 3-4 and 5-6 forming a voltage divider resulting in terminal four being positive $V_1/2$ and terminal two being $V_1/2-V_2$. When R_L is grounded off its center, winding 5-6 becomes inoperative because of its very high reactance, resulting in a 1:1 Balun very much like Guanella's. But **Figure 1-2B** shows that winding 5-6 is always in the circuit and, not only does it reduce the characteristic impedance of windings 1-2 and 3-4, but it also couples the output impedance to the input impedance. In some cases, winding 5-6 is placed between the other two. With the trifilar nature of **Figure 1-2B**, the high-frequency response at the 50-ohm level can vary greatly at the high end. A suggested design would be simply the bifilar winding of 1-2 and 3-4 separated from its neighboring turns by about the diameter of the wires used. In this way the low-frequency response would be the same, but the high-frequency response would be much improved at the 50-ohm level. In other words, it becomes Guanella's form of the 1:1 (current) Balun.

Photo 1-A shows two forms of the 1:1 Balun. On the left is Guanella's and on the right is Ruthroff's. **Photo 1-B** shows the radio amateur's version of Ruthroff's model. As can be seen, the three windings are tightly coupled, yielding the undesirable results. **Photo 1-C** shows probably the latest amateur radio version of the 1:1 Balun. It consists of about 12 inches of 50-ohm cable threaded by ferrite beads. In order to obtain adequate isolation at the low frequency of 1.8 MHz, permeabilities in excess of 2000 are required. Manganese-zinc ferrite beads are, therefore, generally used in this case. As will be shown in **Chapter 5**, these beads exhibit excessive loss. In fact, beads have been known to literally fracture in this application because of the excessive heat.

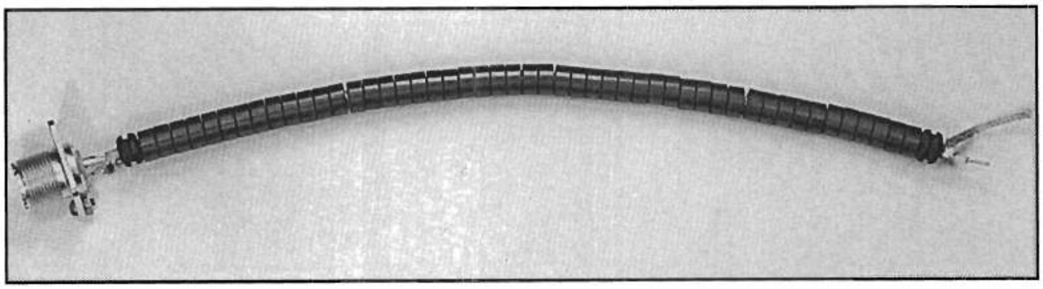

Photo 1-C. *The coaxial cable "choke" 1:1 Balun.*

Chapter 2

TLTs with 1:4 Ratios

The 1:4 TLT and the 1:1 TLT have been the most popular of the transmission line transformers. As Baluns, they have mostly been used in matching center-fed antennas to either coaxial cable or open wire transmission lines. The 1:4 Unun has mostly been used in broadband amplifiers.

The first version of the 1:4 TLT appeared in Guanella's classic 1944 paper[3]. **Figure 2-1** shows the high-frequency and low-frequency models of his Balun. In matching 50 ohms to 200 ohms, the transmission lines each see 1/2 of the load, R_L, and, therefore, the characteristic impedance of each transmission line should be 100 ohms. Since these two transmission lines are connected in parallel on the left side, the result is the desired value of 50 ohms. Since Guanella's TLTs sum voltages of equal delays, the high-frequency response is practically limited only by the parasitic in the connections. This assumes that the coiled transmission lines are not limiting the high-frequency response due to self-resonance.

The low-frequency model on the right in **Figure 2-1** shows that windings 3-4 and 5-6 are in series on the low-impedance side. If these windings are on separate cores their reactances are added in series. When operating at low-impedance levels, allowing for shorter transmission lines, the two transmission lines can be wound on a single toroid. This results in a gain of two times in the reactances in windings 3-4 and 5-6. In any event, the reactances of these two windings (or beaded lines) should be 10 times greater than the input impedance.

With reference to the basic building block discussed in **Chapter 1**, it should again be pointed out that Guanella's Balun can perform as a broadband delay line (Unun) when terminal 2 is grounded, and as a phase inverter when terminal 8 is grounded. It should be recognized that converting Guanella's 1:4 Balun to a 1:4 Unun yields exceptionally broadband widths and is an important design, especially for matching 50 ohms to 12.5 ohms.

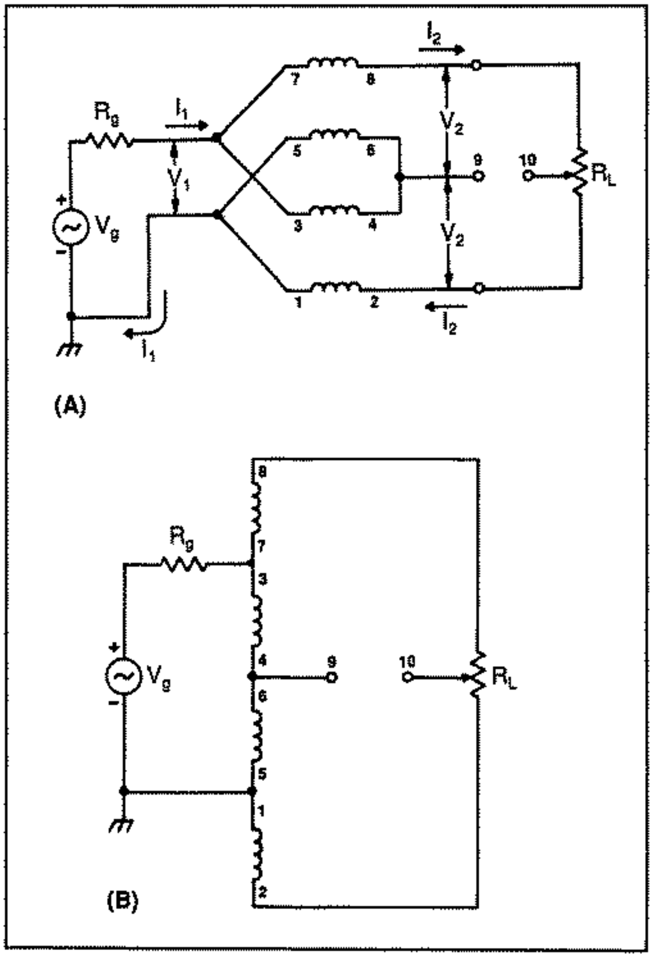

Figure 2-1. Electrical models of the Guanella 4:1 Balun: (A) high-frequency, (B) low-frequency.

In **Figure 2-2A** we see Ruthroff's high-frequency model for his 1:4 Unun and his 1:4 Balun in **Figure 2-2B**. In his 1:4 Unun he connects the high-voltage side of the input impedance to the end of the bottom transmission line. This lifts the whole transmission line by the voltage, V_1, resulting in a "bootstrap" effect. Since the transmission line sees 1/2 of the load, R_L, its characteristic impedance should also be 1/2. In the case of the Balun in **Figure 2-2B**, the

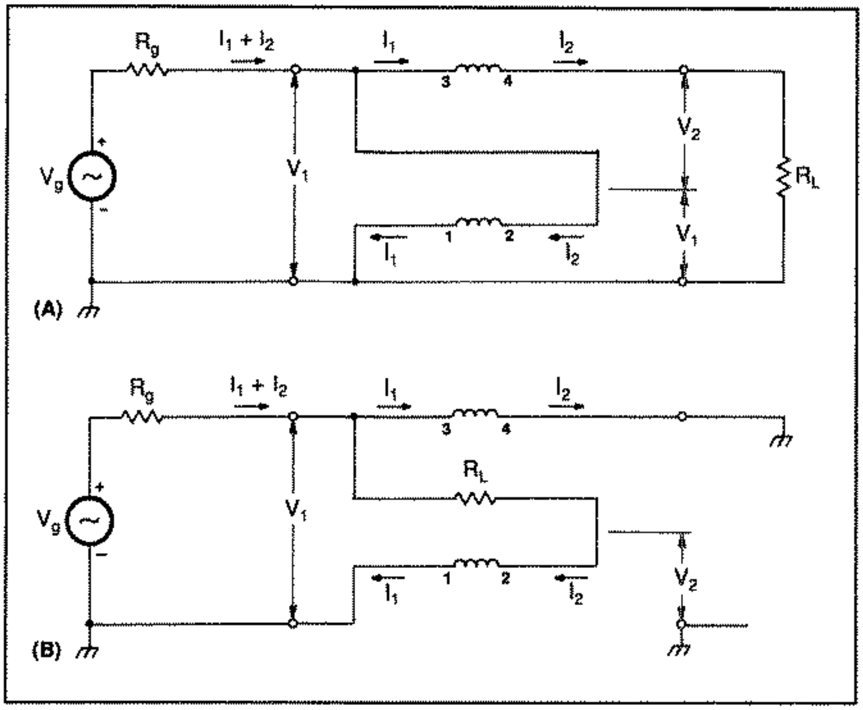

Figure 2-2. *The Ruthroff 1:4 transformers: (A) Unun and (B) Balun.*

basic building block is connected as a phase inverter. In this instance the characteristic impedance should also be 1/2 of the load, R_L, for maximum high-frequency response.

There are two major differences between Ruthroff's approach over that of Guanella's. It is somewhat simpler because it uses only one transmission line, but it also has a poorer high-frequency response since it adds a delayed voltage to a direct voltage. Thus, when the transmission line is a 1/2 wavelength long the output is zero. **Figure 2-3** shows the calculated losses of Ruthroff's 1:4 TLTs as a function of the transmission line length and characteristic impedance. As is shown when the transmission lines are 0.20 wavelength long, the loss is considerable. But for the radio amateur, usually operating below 30 MHz, 2/10 of a wavelength is equal to 2 meters in length, which is much longer than the lengths usually used in the TLTs.

Figure 2-4 shows the low-frequency models of Ruthroff's two TLTs. The model on the left is for his 1:4 Unun, which is identical to that of the autotransformer. The one on the right is for his 1:4 Balun.

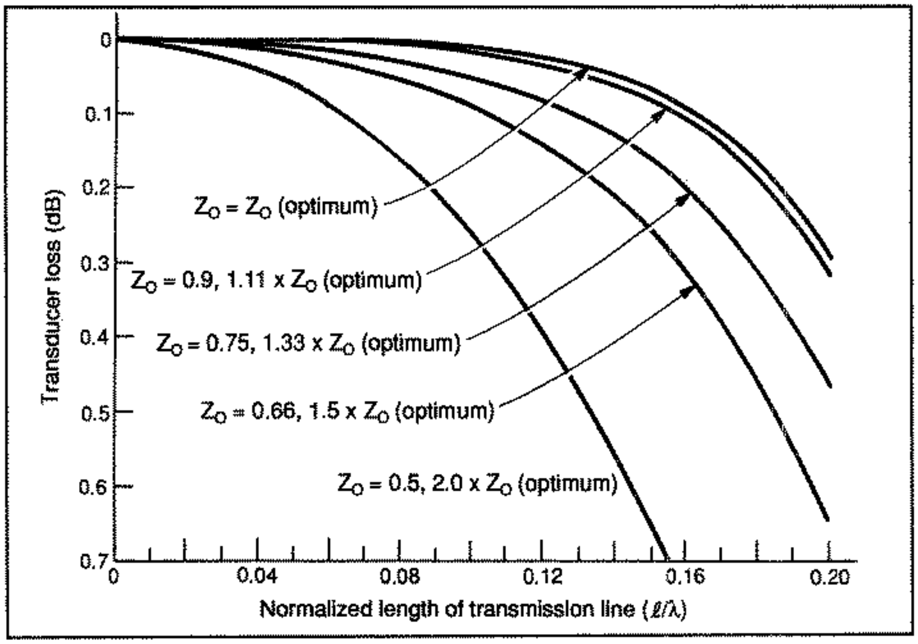

Figure 2-3. *Loss as a function of normalized transmission line length in a Ruthroff 1:4 Unun for various values of characteristic impedance, Z_O.*

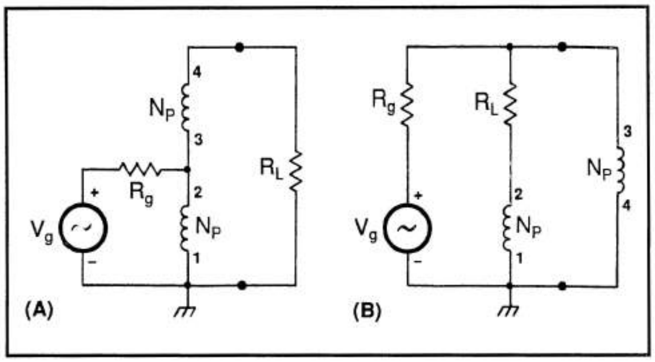

Figure 2-4. *Low-frequency models of the Ruthroff 1:4 transformers: (A) Unun and (B) Balun.*

Finally, **Photo 2-A** shows two versions of Guanella's 1:4 TLTs. The top TLT is a 1:4 Balun matching 50 ohms balanced to 12 1/2 ohms unbalanced. It uses tightly wound turns of No. 14 wire, yielding a characteristic impedance close to 25 ohms. The frequency response is flat from 1 1/2 MHz to about 50 MHz. The upper frequency response in this case is limited by the self-resonance of the windings. On the other hand, the bottom TLT shows a Guanella 1:4 Unun stepping down from 50 ohms unbalanced to 12 1/2 ohms unbalanced. It uses two 25-ohm coaxial cables, with one covered in ferrite beads. In this design, the response is flat from 10 MHz to well *over* 100 MHz!

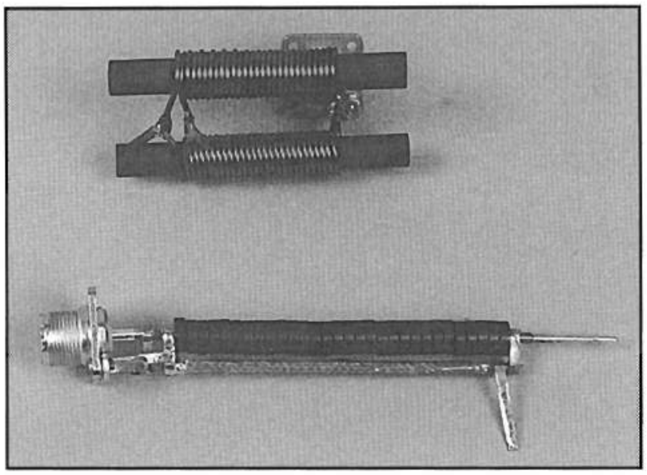

Photo 2-A. *Two Guanella 4:1 (50:12.5 ohm) Ununs: rod version (on the top), 1.5 to 50 MHz; beaded version (on the bottom), 10 MHz to over 100 MHz.*

The result of the two methods of obtaining 1:4 ratios can be categorized as follows: The Ruthroff bootstrap approach uses only one transmission line and can handle many applications when proper lengths of transmission lines are used. The Guanella approach always uses an added transmission line but, in turn, summons voltages of equal delays and, therefore, has a higher frequency response. The next chapter, which covers TLTs with ratios of less than 1:4, carries this same theme much further.

TLTs with Ratios Less Than 1:4

The first two chapters of this book essentially gave the history of the TLT until about the middle of 1970. Some comments were made regarding publications in the amateur radio literature but they were not in anyway additions to the theory.

In working more with these broadband-matching devices, it became apparent that a need existed for impedance ratios greater than 1:1 and less than 1:4. If this was possible, important ratios like 1:1.33, 1:1.5, and 1:2 could result in important applications in antenna uses and logic circuits operating in the usual 50-ohm environment. These low ratios in series with other known TLTs result in important ratios never before obtained. These will be discussed in the next chapter.

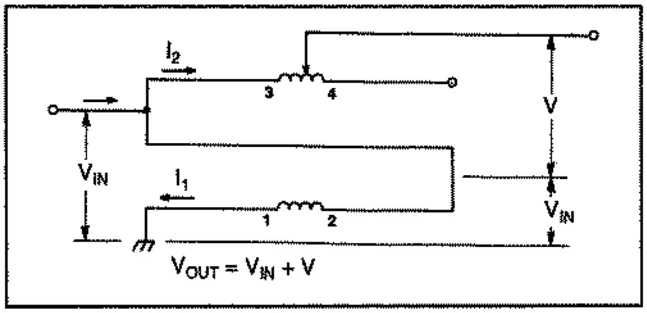

Figure 3-1. *Model for the analysis of the tapped bifilar transformer.*

My first attempt at obtaining a low-ratio TLT was in experimenting with the performance of a Ruthroff 1:4 Unun with a tap on its top winding. **Figure 3-1** shows the schematic. Two taps were employed: one at one turn from terminal 3 and the other at three turns. The experimental results are shown in **Figure 3-2**. There are seven total trifilar turns of No. 16 wire on a 4C4 toroid with an outside diameter of 1.5 inches and a permeability of 125. As is shown in the figure, the poorest performance occurs at the 24.5:50-ohm impedance level. The other two ratios exhibit wider bandwidth. Even though the bandwidths and efficiencies shown in **Figure 3-2** are quite poor, as will be shown later, they are still far superior to those of the autotransformer.

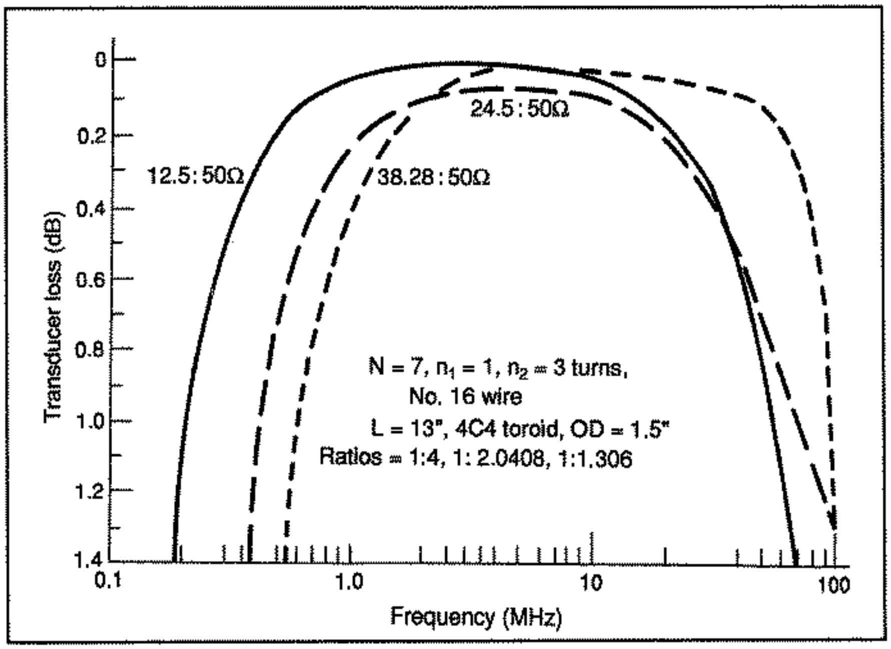

Figure 3-2. *Loss versus frequency for the tapped bifilar transformer.* ⟶

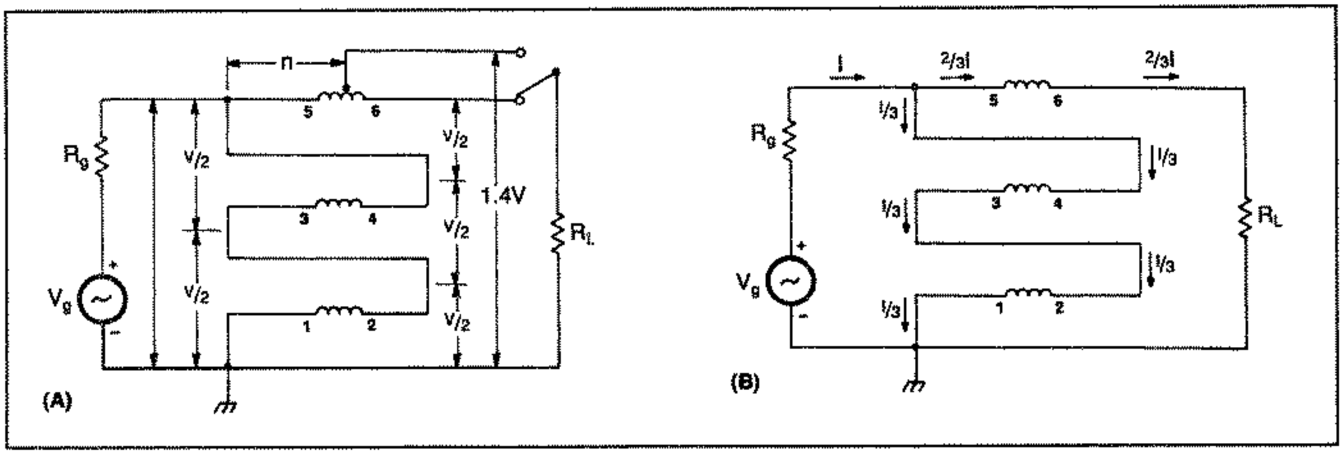

Figure 3-3. *Trifilar bootstrap Ununs: (A) voltages in 1:2 and 1: 2.25 ratios and (B) currents in 1:2.25 ratios.*

In order to improve upon the performance at around the 1:2 level, a trifilar connection that is an extension of Ruthroff's 1:4 "bootstrap technique" Unun is shown in **Figure 3-3**. **Figure 3-3A** shows the voltage relationships, including a tap on the top winding about 8/10 the distance from terminal 5 and yielding an output voltage very close to 1.4 times the input voltage, V. By connecting the load, R_L, directly to terminal 6, the output voltage is 1.5 times the input, yielding a natural ratio of 1: 2.25. This ratio should suffice in many 1:2 applications.

But the schematic in **Figure 3-3B** shows the currents that flow in this trifilar configuration. It is interesting to note that winding 5-6 in the 1:2.25 application carries 2/3 of the current, while the other two windings carry 1/3 each. This means a transmission line mode occurs such that winding 5-6 is shared by windings 1-2 and 3-4. Another view is that two transmission lines operate in parallel with one winding common to both. Physically this means that this very efficient broadband TLT is basically comprised of two of Guanella's basic building blocks in parallel! From this model it becomes clear that the performance of this devise is based upon the fact that there is a direct voltage of V/2 from terminal 2 to ground and *two delayed voltages of V/2*.

The highly efficient and broadband performance of this trifilar Unun is shown in **Figures 3-4** and **3-5**. **Figure 3-4** shows the trifilar winding on a 1 1/2-inch toroid with a tap yielding a 1:2 ratio. As is shown in

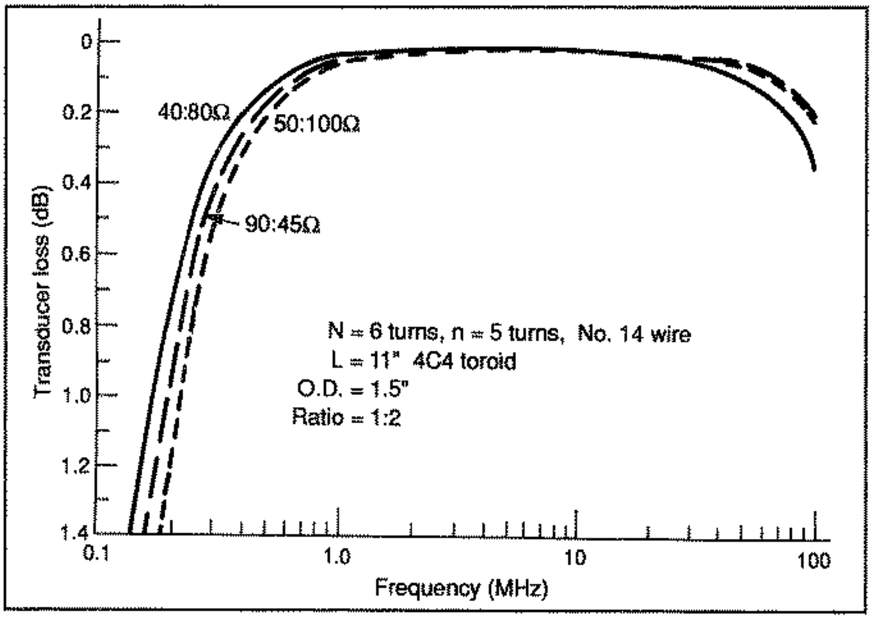

Figure 3-4. *Loss versus frequency for a tapped trifilar transformer at the 1:2 impedance ratio and at three different impedance levels.*

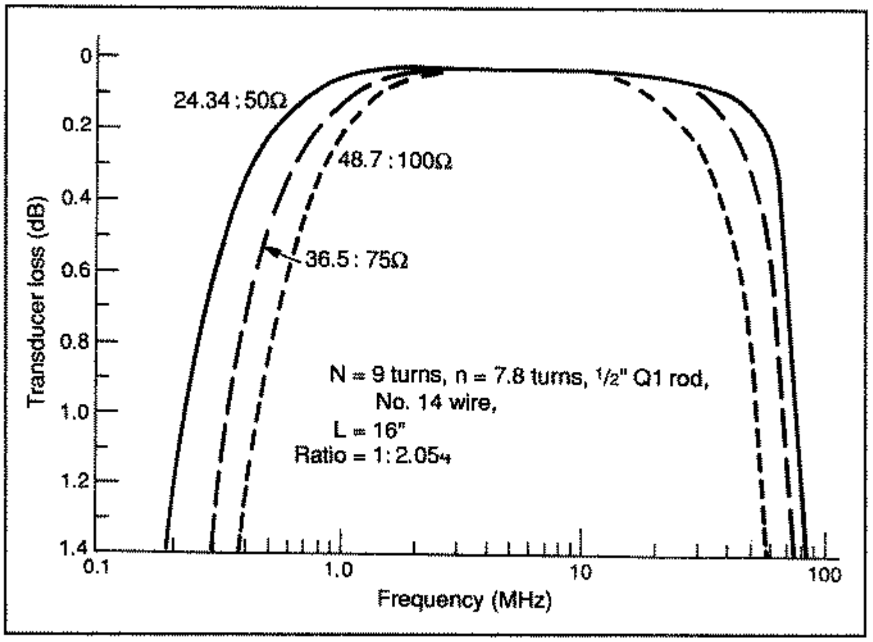

Figure 3-5. *Loss versus frequency for a tapped trifilar transformer using No. 14 wire on a 1/2-inch diameter Q1 rod and at three different impedance levels.*

the figure, the 50:100-ohm ratio is slightly better than the other two.

The efficiency and bandwidth shown here far exceed the 1:4 Ruthroff Unun since the voltage drops along the lengths of the windings are considerably less and the cancellation due to adding a direct voltage with the delayed one is considerably smaller. **Figure 3-5** shows the same trifilar winding on a 1 1/2-inch ferrite rod of similar permeability. Because the windings are tight to each other, the effective characteristic impedance concerning the top winding 5-6 is considerably lower. As is shown, the optimum performance occurs at the 24.34:50-ohm impedance level. This suggests that optimum performance at the 25:50-ohm impedance level for a toroidal core would probably occur if winding 5-6 in **Figure 3-3** were placed in between the other two. Examples are shown in **Part II** of this book.

Another revealing part of the story is illustrated in **Figure 3-6**, which shows two coaxial cables, connected in parallel and in Ruthroff bootstrap operation. The inner-conductor of the top coax is tapped at about 8/10 the distance from the left-hand edge, yielding a voltage ratio of 1.4 times the input voltage, V. **Figure 3-6A** shows the voltages and **Figure 3-6B** shows the currents at the 1:2.25 ratio. Since the voltage of each coax is 1/2 the input voltage, but the current in the top coax is twice that of the bottom one, its characteristic impedance should most likely be 1/2 that of the lower coax. For operations at the 25:50-ohm impedance level, the characteristic impedance of the top coax should be 25 ohms and that of the bottom 50 ohms. The application of coaxial cables looks very interesting.

Figure 3-7 shows a Guanella approach for a coiled wire TLT with ratios less than 1:4. Specifically, it shows a basic building block in parallel with a Guanella 1:4 Balun yielding a ratio of 2.25:1. The high-impedance side on the left shows the two TLTs in series and, on the right, the outputs in parallel. Since Guanella's TLTs are known to have equal delays for equal lengths of transmission lines, the high-frequency response is only limited by parasitics in the connections or self-resonance in coiled windings. In some instances, this form of TLT has been called an "equal-delay" TLT. If the impedance on the left is 100 ohms, the impedance on right side is 44.4 ohms. The optimum characteristic impedances of the transmission lines in this case are each 66 ohms. In matching 50 ohms to 22.22 ohms, the optimum characteristic impedances should be 33 ohms. For lower ratios, the bottom TLT in **Figure 3-7** should be changed to higher ratios. For example, with a 1:9 ratio the final overall ratio becomes 1.78:1, and for a 1:16 ratio in the bottom TLT the overall ratio becomes 1.56:1.

If the impedance on the left is 100 ohms, in the 1.78:1 ratio the ideal characteristic impedances should be 75 ohms for each transmission line; for the 1.56 ratio the characteristic impedances should be 80 ohms. When matching 50 ohms to lower impedances, the

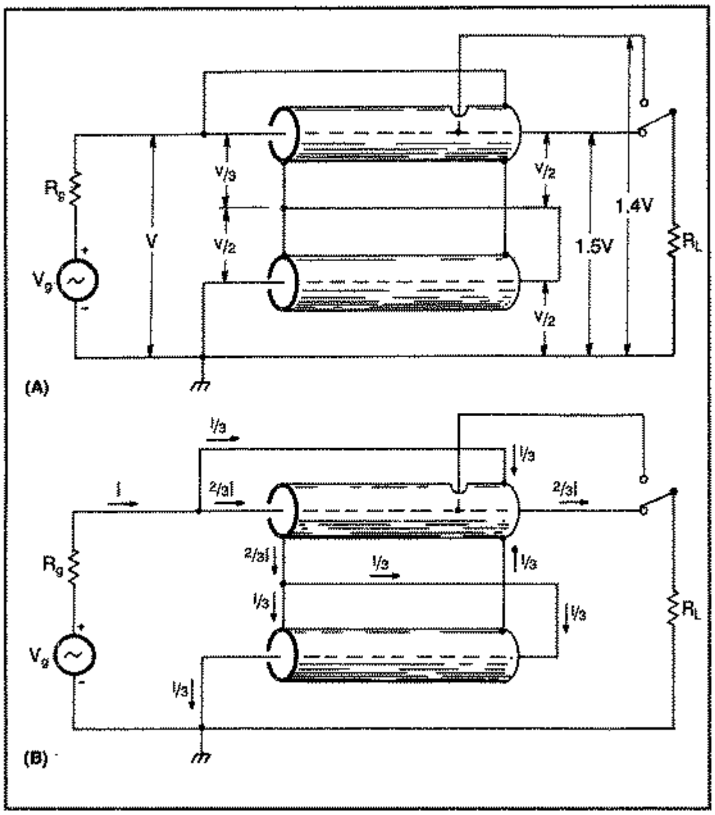
Rg
Vg
V
V/3
V/2
V/2
1.4V
1.5V
RL
V/2
(A)
I/3
I
2/3I
2/3I
I/3
I/3
2/3I
I/3
I/3
I/3
Rg
Vg
RL
(B)

Figure 3-6. *Parallel coax Unun: (A) Voltages in 1:2 and 1:2.5 ratios and (B) currents in the 1:2.25 ratio.*

optimum characteristic impedances of the windings (or beaded coaxes) should be 37 1/2 and 40 ohms, respectively. Further descriptions of these equal-delay transformers are described in **Part II, Chapter 10** of this book and in Chapter 18 of my other TLT book.[1]

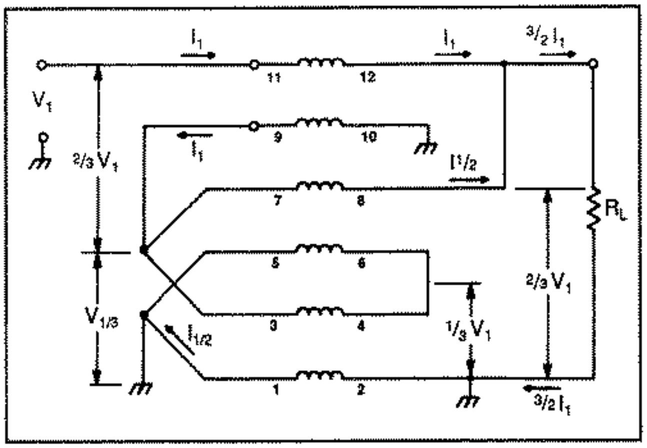
I1
I1
3/2 I1
V1
11
12
9
10
2/3 V1
I1
I1/2
7
8
RL
5
6
2/3 V1
V1/3
I1/2
3
4
1/3 V1
1
2
3/2 I1

Figure 3-7. *High-frequency model of the parallel-type 2.25:1 transformer. Connections shown are for Unun operation.*

Figure 3-8 shows two quintufilar windings of a 1:1.56 Unun. The figure on the left is optimized for matching 50 to 75 ohms and that on the right to matching 50 to 33 ohms. Although not shown here, winding 9-10 carries 4/5 of the input current and the other windings carry 1/5 each. The other point to make in both the trifilar and quintufilar windings is that their low-frequency models show considerably more windings in series than their 1:4 counterpart. This means that fewer turns and, hence, shorter lengths can be employed in these higher order windings. Proof of this is shown in **Figure 3-9**, which illustrates the incredible performance at the 64:100-ohm impedance level using schematic "A" in **Figure 3-8**. As is shown, only three

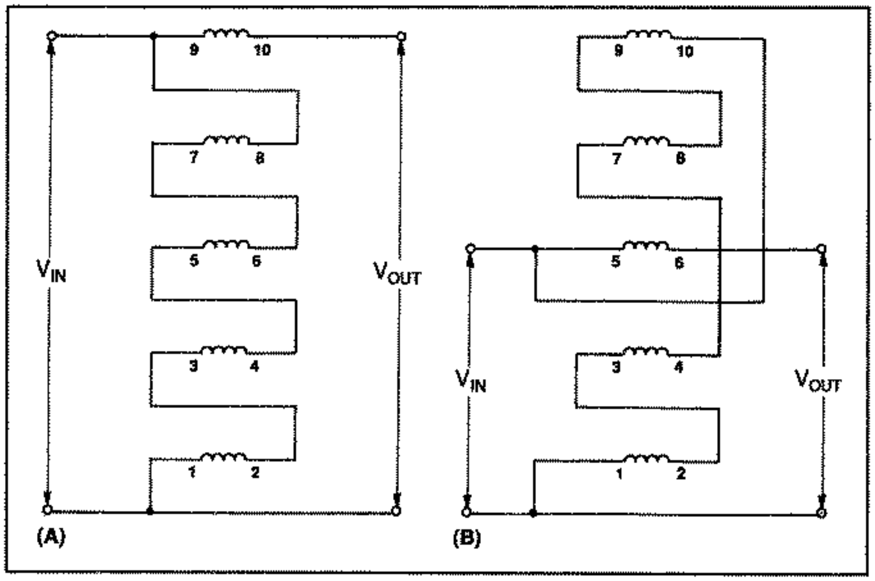
9
10
7
8
5
6
3
4
1
2
VIN
VOUT
(A)
9
10
7
8
5
6
3
4
1
2
VIN
VOUT
(B)

Figure 3-8. *Quintufilar transformers with impedance ratios of 1:1.56: (A) depicts a high-impedance operation and (B) displays windings configured for low-impedance operation.*

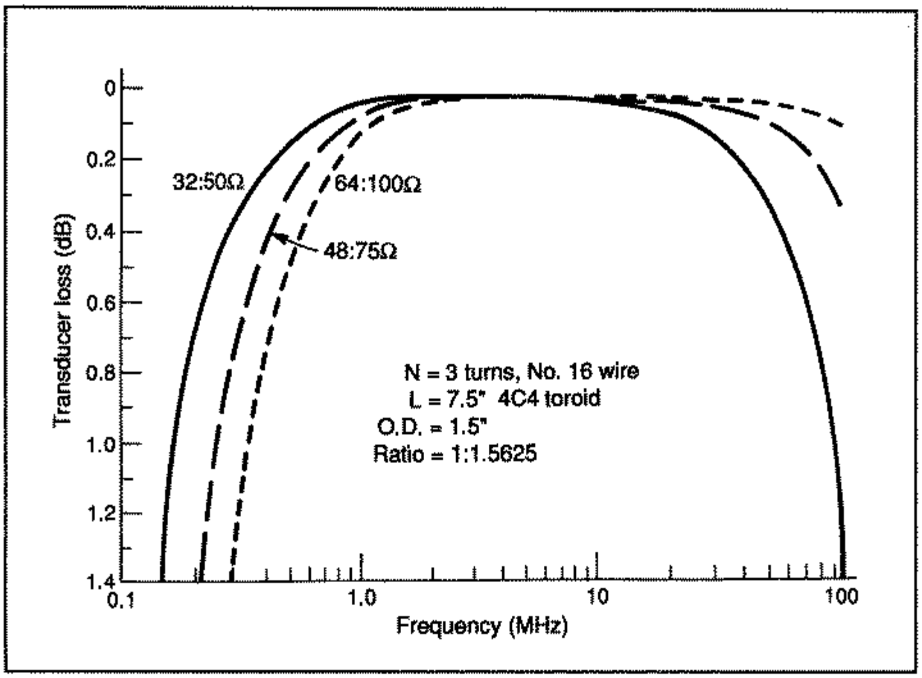

Figure 3-9. *Loss versus frequency for a five-winding transformer.*

turns of No. 16 wire is employed on the ferrite toroid with a permeability of 125. This performance even exceeds that of the trifilar one.

With shorter lengths of transmission line and higher permeabilities with smaller ferrite toroids, the bandwidths of these multi-conductor TLT's performances should be vastly improved. And, finally, tapping of the Guanella approach to these broadband low-ratio TLTs should be investigated. Since this technique is highly successful in the Ruthroff bootstrap approach, it is apparent that it should be also in the Guanella's equal-delay approach.

TLTs with Ratios Greater Than 1:4

From TLTs with ratios less than 1:4, as described in **Chapter 3**, it is now possible to easily achieve Baluns and Ununs matching 50 ohms to impedances as high as 600 ohms and as low as 3.125 ohms. Since TLTs can be developed with very high efficiencies and broadband widths, several TLTs can be connected in series with their outstanding features.

This chapter begins with the two TLTs that resulted in high ratios using only one transformer. The first, shown in **Figure 4-1**, is the Guanella 1:9 Balun. **Figure 4-1A** is the high-frequency model and **Figure 4-1B** is the low-frequency model. The design objectives when matching 50 ohms to 450 ohms, the characteristic impedances, Z_O, of the three transmission lines should be 150 ohms. As can be seen, each transmission line sees 1/3 of the load, R_L. In **Figure 4-1B**, the reactance of winding 5-6 in series with winding 3-4, in turn, in parallel with windings 9-10 and 7-8 should be 10 times greater at the lowest frequency of interest—50 ohms. This transformer can operate even better in matching 5.5 ohms to 50 ohms. At lower impedance levels, lower reactances are required. Further, higher core permeabilities can be employed since core loss is impedance level dependent. This transformer can also operate as an Unun when terminal 2 is grounded. In this application, a 1:1 Balun on the output of a low-impedance Balun on the input of a high-impedance Balun (matching 50 ohms to 450 ohms) would help in the low-frequency response.

Figure 4-2 is a schematic of a trifilar-wound TLT using Ruthroff's bootstrap approach for obtaining a

Figure 4-1. *Guanella's 1:9 transformer: (A) high-frequency model and (B) low-frequency model. It is assumed that* $Z_O = R_L/3$ *and, therefore, the output voltages of the transmission lines are each equal to* V_1.

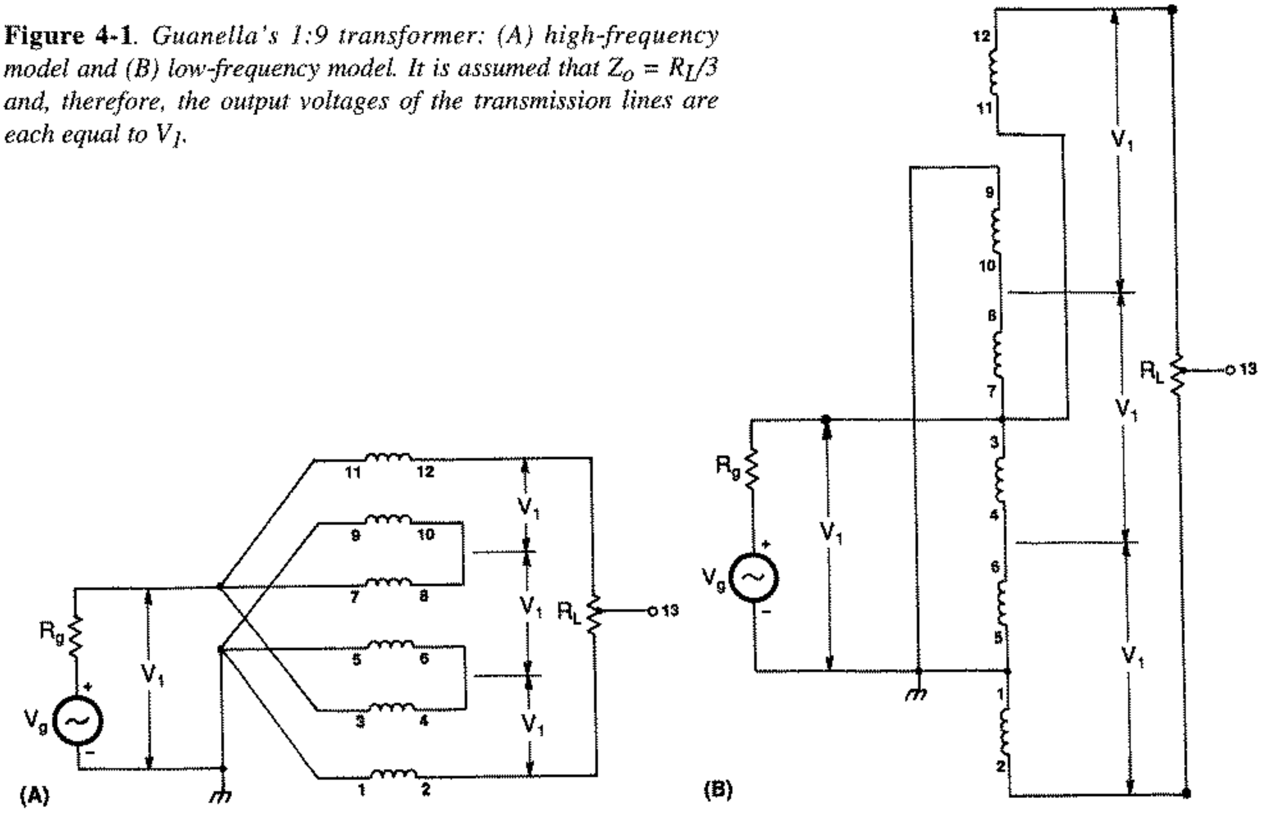

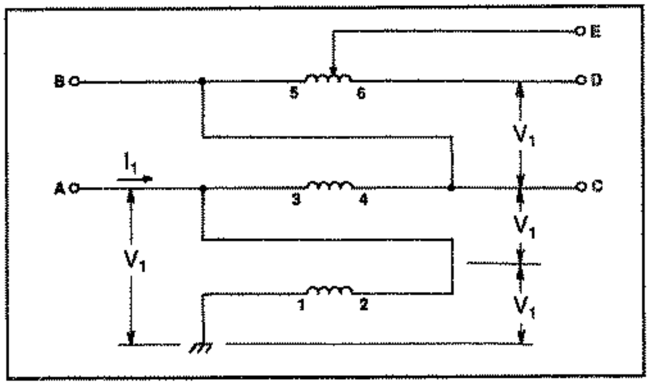

Figure 4-2. *A schematic diagram of a trifilar transformer for obtaining impedance ratios up to 1:9.*

variety of impedance-matching ratios from 1:2 up to 1:9. In the 1:9 ratio using terminals A-D, it shows a direct voltage added to a double delayed voltage. Therefore, the high-frequency response is much poorer than the other ratios. But at low-impedance levels, which require lower reactances and thus shorter lengths of transmission lines, this Unun makes a practical multi-match transformer for the HF band.

Figure 4-3 is a schematic for a 1:6.25 Balun matching 50 ohms to about 300 ohms. On the left is the extension of Ruthroff's technique to obtain an Unun with a ratio very close to 1:1.5, followed by a Guanella 1:4 Balun. By tapping the top winding in the Unun on the left, a ratio very close to 1:6 can be obtained. Further, by installing a 1:1 Balun in between these two TLTs and grounding terminal 2 in the 1:4 Balun, a broadband 1:6 Unun results.

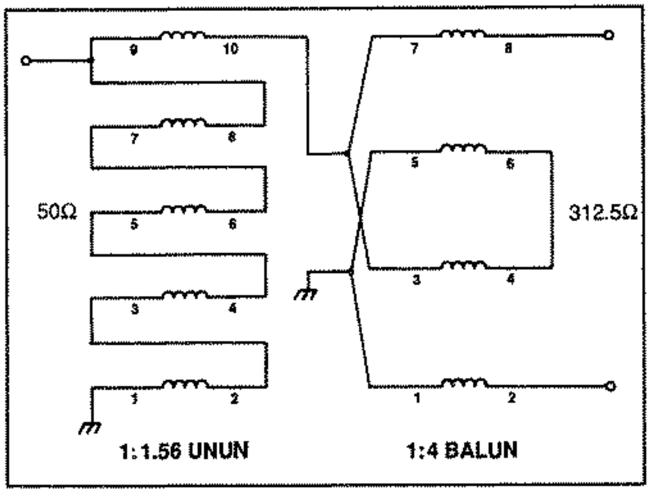

Figure 4-3. *Schematic diagram of the series-type Balun with a 1:6.25 ratio designed to match 50 to 312.5 ohms.*

Figure 4-4 is a schematic of a Balun matching 50 ohms to 600 ohms, yielding a ratio of 1:12. The Unun on the left is a tapped quintufilar winding, yielding a ratio of 1:1.33 in series with a 1:9 Guanella Balun. A septufilar-wound Unun can also be used. By installing a 1:1 basic building block in between these two TLTs and grounding terminal 2 of the 1:9 Balun, a broadband 1:12 Unun results.

Broadband ratios of 1:16 can also be obtained with the Guanella approach. It would contain four transmission lines in a parallel series connection. For the Ruthroff approach, it would be two 1:4 Ununs in series. With appropriate characteristic impedances, broadband operation is possible at these very high ratios.

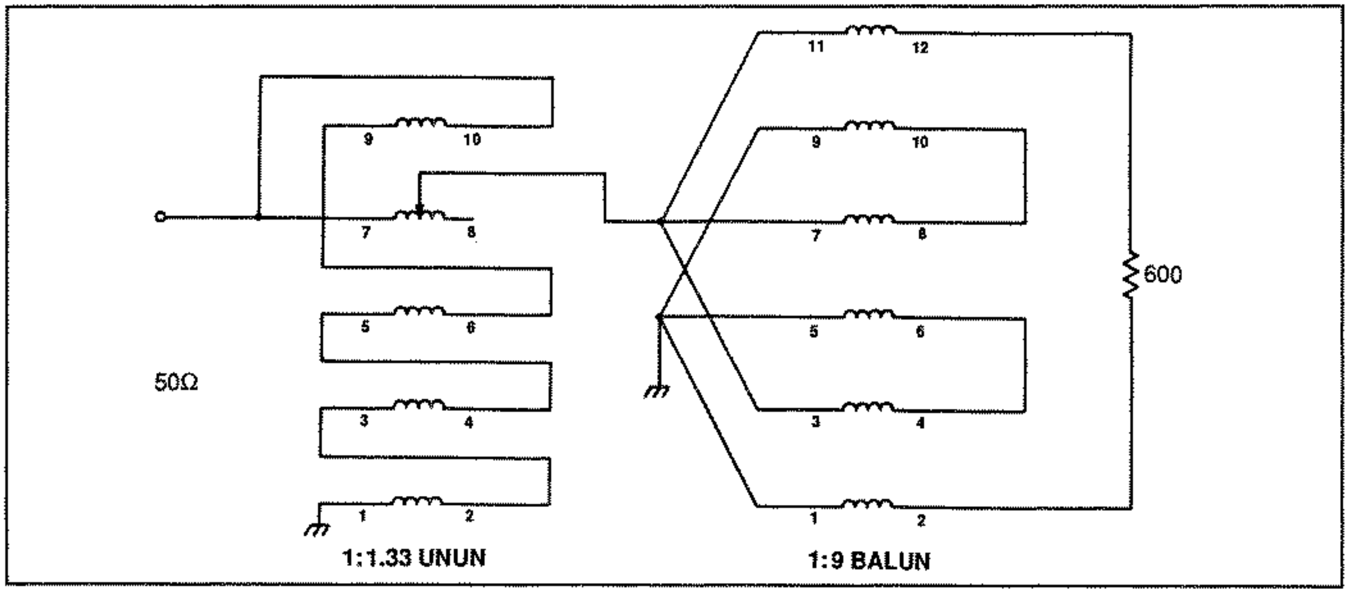

Figure 4-4. *Schematic diagram of the series-type 1:12 Balun using a 1:1.33 Unun in series with a 1:9 Guanella Balun.*

Chapter 5

Characterization

The word characterization is very familiar to those engaged in the research and development of Solid-State devices. It involves the study of the various components that make up the models of the particular devises. In the case of TLTs, we have two models of each transformer as previously mentioned: one is the high-frequency model and the other the low-frequency model. The major subject of interest with the high-frequency model is the characteristic impedance, Z_0, of the transmission lines employed. Having many types of transmission lines available leads to the major investigation to develop the proper values of characteristic impedance. It is assumed that the proper model has been chosen, whether it's Guanella's with its equal delays, or Ruthroff's with his bootstrap application. The make-up of the magnetic cores, which are usually ferrites, is also a study of interest. For high-power applications, the bulk resistivity becomes an important ingredient. In the low-frequency model, the reactance of the coiled or beaded transmission line at the lowest frequency of interest is the important consideration. Here, number of turns, permeability of the cores, and geometry of the cores all come into play.

Since this topic covers so much territory, only a few of the major subjects can be included in this chapter. The first topic deals with the geometry of the cores and the permeability of the material. **Figure 5-1** shows three TLTs with the same windings but with different cores. Two are toroids and one is a rod. One is a powered-iron toroid with a permeability of only 10. The other two use Q1 ferrite material with permeabilities of 125. As is shown, the powered-iron core with a permeability of 10 has the poorest low-frequency response, and the ferrite toroid with a permeability of 125 has the best low-frequency response of the three. Also clearly shown is the poorer response of a rod core of four inches in length compared to its counterpart, the toroidal core. Further, there is no simple formula for calculating the so-called "magnetizing

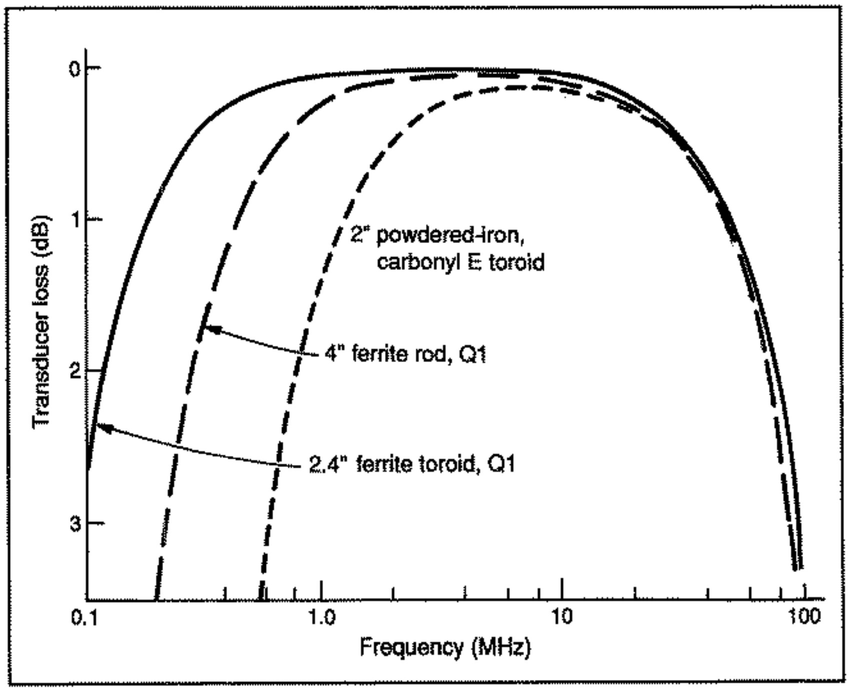

Figure 5-1. *Experimental results showing the low-frequency performance of the rod transformer versus the toroidal transformer.*

MATERIAL	SUPPLIER	PERMEABILITY	BULK RESISTIVITY (ohms-cm)
Q1 (NiZn)	Allen Bradley (formerly Indiana General)	125	10^8
G (NiZn)	Allen Bradley	300	10^6
Q2 (NiZn)	Allen Bradley	40	10^9
H (NiZn)	Allen Bradley	850	10^4 - 10^5
4C4 (NiZn)	Ferroxcube	125	10^7 - 10^8
3C8 (MnZn)	Ferroxcube	2700	10^2 - 10^3
K5 (NiZn)	MH&W Intl (TDK)	290	2 x 10^6
KR6 (NiZn)	MH&W Intl (TDK)	2000	10^5 - 10^6
CMD5005 (NiZn)	Ceramic Magnetics	1400	7 x 10^9
C2025 (NiZn)	Ceramic Magnetics	175	5 x 10^6
CN20 (NiZn)	Ceramic Magnetics	800	10^6
C2050 (NiZn)	Ceramic Magnetics	100	3 x 10^7
E (Powdered Iron)	Arnold Engineering, Amidon Associates	10	10 -2

Table 5-1. *Cores, suppliers, and specifications.*

inductance" when using a rod. The low-frequency results for this core are mainly determined experimentally. Rod cores have found favor in the past for amateur radio use because they are generally less expensive than their toroidal counterparts.

Table 5-1 shows a listing of the various core materials used in my investigation. Since this goes back to about 1975, it is expected that some of the manufacturers are no longer in existence and that new sources are available. A particular example is the K5 material, which is shown in succeeding figures. Since this material was manufactured in Japan, it has now been excluded because any compound using nickel is now considered a serious contaminant. Other countries still manufacture ferrites of nickel-zinc, however.

The two parameters that are important in TLTs are having ferrite materials that are both high in permeability and bulk resistivity. The importance of the latter parameter can be shown in the next two figures. **Figure 5-2** shows the response curve of two TLTs

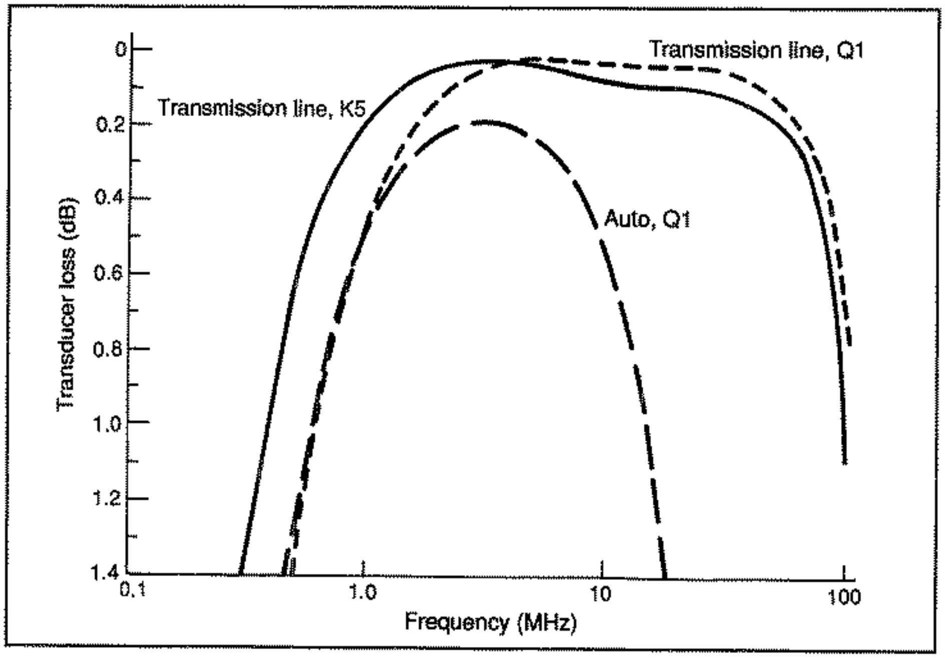

Figure 5-2. *Performance of 4:1 transformers operating at the 100:25-ohm level.*

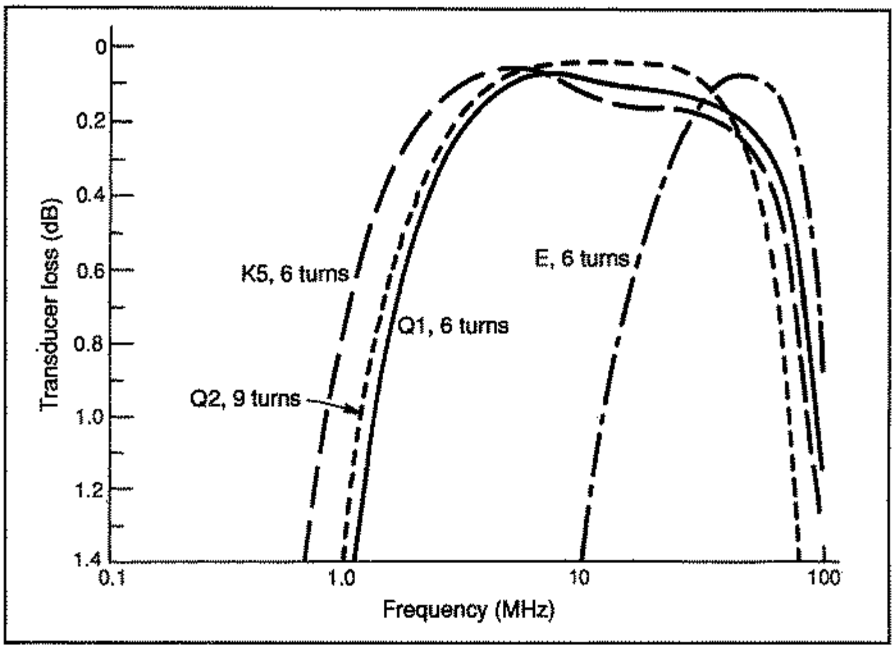

Figure 5-3. *Loss versus frequency of four 4:1 transformers at the 200:50-ohm level.*

using different permeabilities and compared to an autotransformer with the material (Q1) of one of the TLTs. As can be seen when matching at the 100:25-ohm level, the K5 material has a better low-frequency response but a little more loss above 10 MHz.

Figure 5-3 shows the comparisons at the higher impedance level of 200:50 ohms. The difference between K5 and Q1 becomes larger at the higher impedance level. Also shown in the figure are the results of Q2 with a permeability of 40 and E with a permeability of only 10.

Figure 5-4 clearly shows the expected performance with lower bulk resistivities. The 3C8 ferrite has a permeability of 2700 and a bulk resistivity of only 10^2–10^3. The other two curves are from KR6 with a permeability of 2000 and a bulk resistivity of 10^5–10^6, and H at a permeability of 850 and 10^4–10^5.

Experiments with various materials show the importance of investigating the losses of the materials, particularly as a function of impedance level. The 3C8 material has been used on the so-called "choke" Balun, a beaded 50-ohm coaxial cable, which explains why the

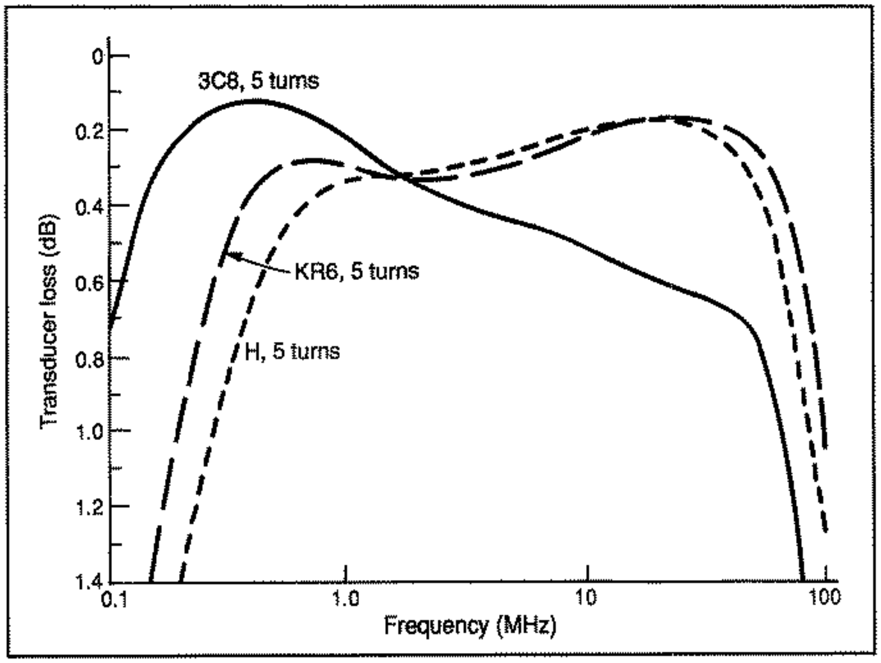

Figure 5-4. *Measurements of 3C8, KR6, and H materials at the 200:50-ohm level.*

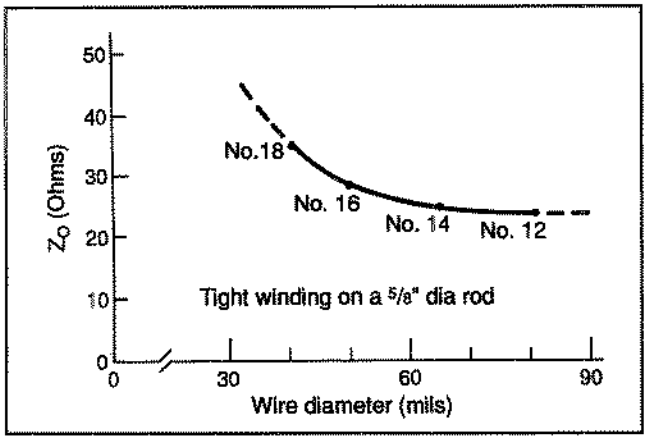

Figure 5-5. *Characteristic impedance, Z_O, versus wire diameter for a tightly wound transformer (tight winding on a 5/8-inch diameter rod).*

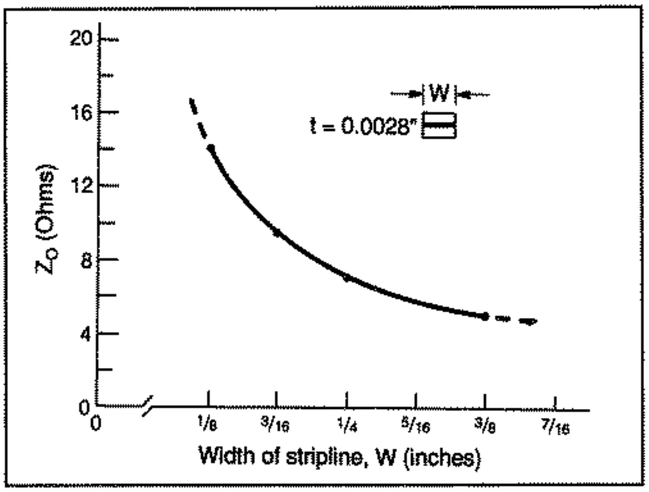

Figure 5-6. *Measured values of the characteristic impedance, Z_O, of striplines versus width.*

beads have been known to fracture due to excessive heat. By the use of ferrites with a permeability of 125 and very much larger cores for added dissipation, TLTs with 50:200-ohm ratios have been known to operate satisfactorily at the 10–20-kilowatt level.[5]

In the investigation of various transmission lines, two interesting results can be shown. **Figure 5-5** shows the variation of characteristic impedance for tightly wound transmission lines on a ferrite rod. The figure shows that No. 14 wire has a characteristic impedance of about 25 ohms in a tightly wound configuration. By separating the bifilar windings by about one wire diameter or winding it around a toroid, the characteristic impedance approaches 50 ohms. When working at impedance levels of 200:300 ohms, characteristic impedances of 100:150 ohms require extra spacing using Teflon™ sleeving and, generally, toroids are used instead of rods.

Figure 5-6 shows the experimental results of a stripline transmission line. As shown, characteristic impedances approaching 5 ohms are obtainable! With thinner insulation, low impedances can be obtained with less widths.

Figure 5-7 shows the results of two Ruthroff TLTs in series using stripline matching 50 ohms to about 3 ohms. By using two Guanella 1:4 Ununs, each using a transmission line as a delay line, bandwidths greatly exceeding those shown in **Figure 5-7** are expected.

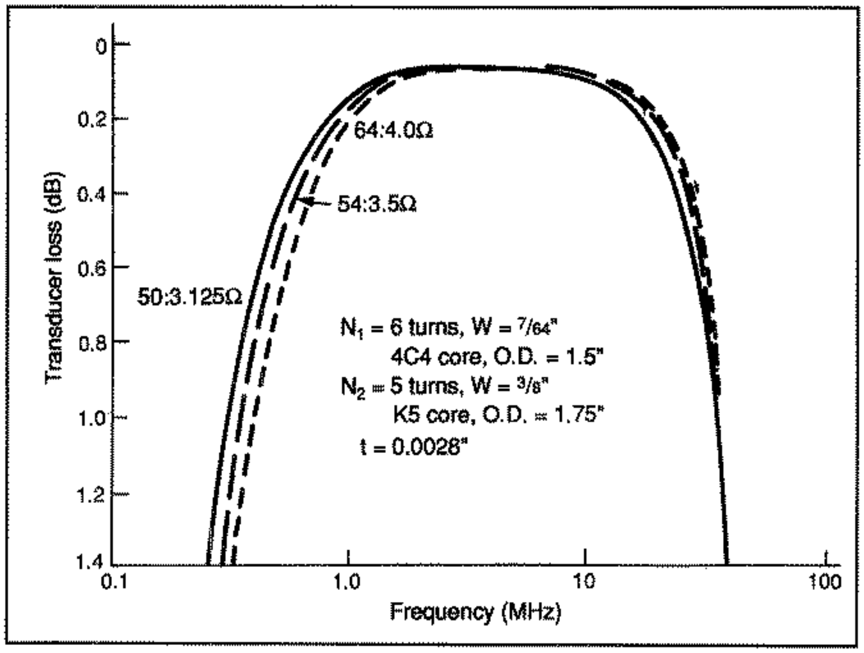

Figure 5-7. *Experimental results of the 16:1 transformer of Figure 4-4 at three different impedance levels. Note the best level is at about 54:3.5 ohms.*

Chapter 6

Summary

After many talks before amateur and professional groups, as well as during a recent literature search on diode mixers and power splitter/combiners, it became apparent that very few really understood transmission line transformer technology. Prime examples of this misunderstanding follow.

Figure 6-1 shows a schematic diagram of a diode mixer. The author states that the transformer, T-1, is a TLT. He assumes it is a TLT because its windings are twisted. But from the schematic it can be easily shown that T-1 has a primary and secondary winding like any conventional transformer in which it transmits the energy from input to output by flux linkages. The idea of twisting came from Ruthroff's approach. He worked with very small toroids and fine wires. In order to keep the characteristic impedance of the transmission lines under control he had to twist the wires.

Another author also employed a twisting of wires. **Figure 6-2** shows the transformer, which he called a TLT, used in his schematic. The input current, 2I, splits into two currents, which flow out of the two secondary terminals. The voltages of the three terminals are the same. Since it can be seen that the currents through the transformer are equal and opposite, there is virtually no voltage drop between its terminals because of flux canceling. Furthermore, since the input on the right side has an impedance half that of the two outputs, a 1:2 transformer is required to bring the circuit back to 50 ohms. The 1:2 transformer used in the author's schematic is an autotransformer, as was shown in **Chapter 5**, it is vastly inferior in bandwidth and efficiency to a TLT. And, finally, the theory of the flux-canceling transformer is certainly not that of a TLT and requires further study.

In reviewing the history of the TLT, we've seen that two approaches had been used until about 1970.

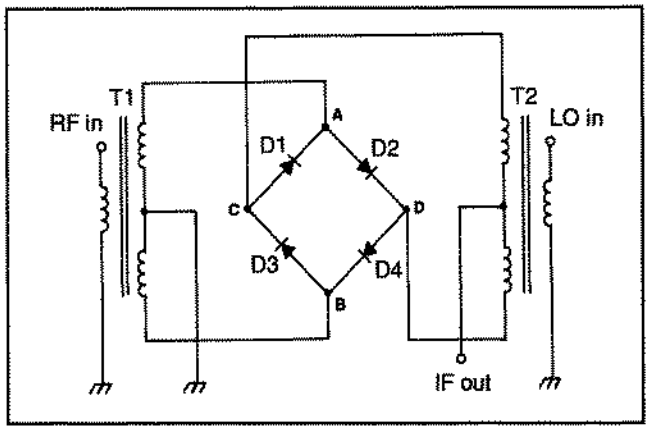

Figure 6-1. *Equivalent circuit of a double-balanced mixer using four diodes in a ring configuration.*

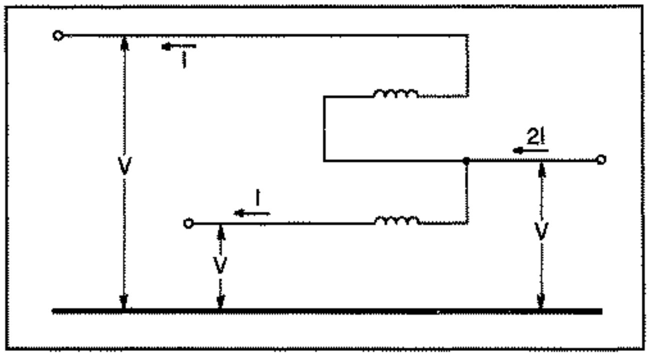

Figure 6-2. *Typical power combiner/splitter circuit.*

The first was Guanella's, which for 1:4, 1:9, and 1:16 TLTs used transmission lines connected in parallel at the low-impedance side and in series at the high-impedance side. Essentially he summed voltages of equal delays. Therefore, his TLTs, which were essentially Baluns, had high-frequency responses only limited by the parasitics. His 1:1 basic building block and 1:4 Balun are the Baluns of choice today.

Ruthroff, on the other hand, used Guanella's basic building block for obtaining a 1:4 ratio by connecting the basic building block in a bootstrap configuration. He was able to produce a broadband 1:4 Unun. By connecting the basic building block in a phase-inverter configuration he was able to obtain a broadband 1:4 Balun. But his 1:4 TLTs summed a direct voltage with a delayed voltage, which eventually ended up in a canceling mode, thereby limiting the high-frequency response. For many applications not requiring vary large bandwidths, his simple TLTs are applicable.

After 1970 advances where made, particularly with the 1:4, 1:2, and 1:1.5 Unun. The improved 1:4 Unun used Guanella's equal-delay approach by employing one of the basic building blocks as a delay line. Since this transmission line has no voltage drop across it, no magnetic medium is required. The other transmission line has the same voltage drop across it as the Ruthroff Unun, thus resulting in the same low-frequency response.

As was shown in **Chapter 3**, by using a trifilar bootstrap approach, improved bandwidth at the 1:2.25 ratio results. By tapping one of the conductors, a ratio of practically 1:2 can be obtained. By using five conductors in a quintufilar arrangement, broadband ratios of 1:1.5 are easily obtained. Since the delayed voltage is four times the direct voltage, this TLT has even much broader bandwidth than the 1:4 and 1:2 ratios. These ratios, again as shown in **Chapter 3**, can also be obtained by using Guanella's equal-delay approach. A 1:1 basic building block in series/parallel with a 1:4 Guanella TLT results in a very broadband 1:2.25 ratio. This is also called an equal-delay TLT. By using the basic building block in series parallel with Guanella's 1:16 TLT, a broadband equal-delay 1:1.56 ratio results. Although, to the best of my knowledge it has not been tried, tapping one conductor in the basic building block, which is connected in series parallel with the 1:4 equal-delay TLT, could result in a ratio close to 1:2. The choice one makes at the 1:1.5 ratio is still not clear.

The bootstrap approach has the advantage of using only one magnetic medium with all the conductors in a series aiding operation, which could make this device very practical in many applications. Undoubtedly, the equal-delay method suffers at the low-frequency end and requires *five* transmission lines, reducing its application. It is clear that the 1:4 and 1:2.25 equal-delay TLTs have much wider bandwidths than their counterparts.

When viewing the results of the various experiments performed on the TLT, it becomes apparent that this technology is not suited for the vacuum tube with its high impedances, but for the Solid-State devices of today with their lower impedances. With the application of the equal-delay approaches of Guanella and the low-impedance ratios of Ruthroff's bootstrap approach, the future is bright for the application of these broadband TLTs at low-impedance levels. With modern fabrication techniques, the TLT, in matching 50 ohms down to lower impedances, should have a capability greatly exceeding 100 MHz!

And, finally, for those interested in writing a comprehensive text like that of the renowned Frederick E. Terman,[6] I suggest including the simplified and unified approach to the theory of the very broadband, impedance-matching TLT shown in this book.

In closing, I feel it necessary to include a small tutorial on grammar. The "Balun" which is well known, is a match between a balanced impedance and an unbalanced impedance. The "Unun," a word I coined several years ago, came about because there was no acceptable term for a device matching unbalanced impedance to unbalanced impedance, and it became quite tedious to continuously type the whole expression. In discussing this device on the air, I found many called it a U-nun—obviously the right terminology is Un-un.

References

1. Jerry Sevick, *Transmission Line Transformers*, 4th edition, Noble Publishing Corporation, 2001.
2. Jerry Sevick, *Building and Using Baluns and Ununs*, 1st edition, CQ Communications Inc., 1994.
3. G. Guanella, "Novel Matching Systems for High Frequencies," *Brown-Boveri Review*, Volume 31, September 1944, pages 327–329.
4. C.L. Ruthroff, "Some Broad-Band Transformers," *Proceedings of the IRE*, Volume 47, August 1959, pages 1337–1342.
5. J. Terleski, <arraysolutions.com> website.
6. F.E. Terman, *Electronic and Radio Engineering*, 4th edition, McGraw-Hill Book Co., 1955

PART II—
BUILDING AND USING BALUNS AND UNUNS

Chapter 7

The 1:1 Balun

Sec 7.1 Introduction

In this chapter I will introduce the most popular broadband Balun in amateur radio use—the 1:1 Balun. This topic been discussed in the amateur radio literature since the publication of Turrin's 1964 article.[8] Although Turrin's Balun is really a version of Ruthroff's, which was introduced in his classic 1959 article,[9] the real beginning of the broadband 1:1 Balun dates back to Guanella's classic paper of 1944.[3] Guanella's objective was to design a broadband 16:1 Balun to match the balanced output impedance of 960 ohms of a push-pull, 100-watt vacuum-tube amplifier to the unbalanced load of a matched 60-ohm coaxial cable. Use of his approach for 1:1, 4:1, and 9:1 Baluns has produced the designs of choice. They are presently called current or choke Baluns.

This chapter begins with an introduction to the technology of transmission line transformers. Other topics discussed include: 1) when to use a Balun; 2) highlights of significant articles in the professional and amateur radio literature; 3) high-power, medium-power, and low-power designs; and 4) isolation transformers. The latter is presented here for the first time. The chapter closes with a brief summary of the significant points included within.

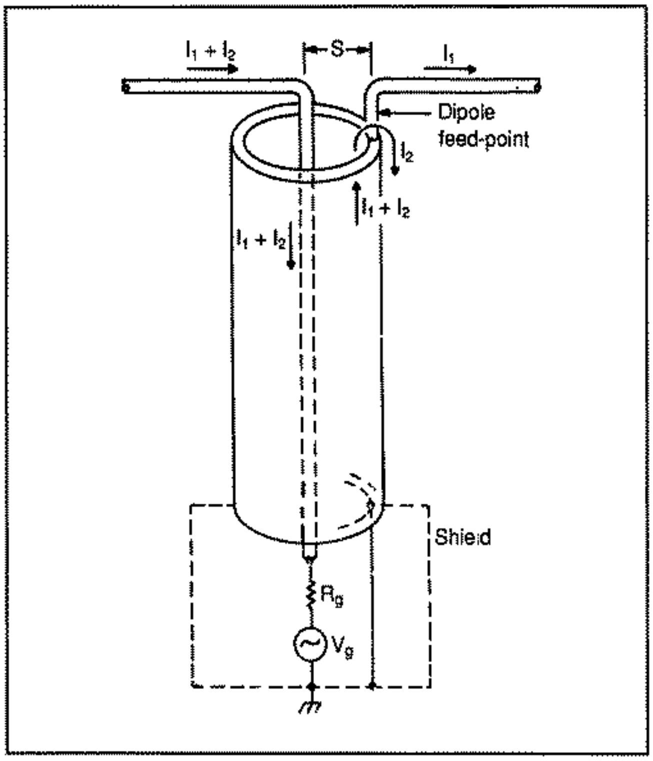

Figure 7-1. *An illustration of the various currents at the feedpoint of a dipole. I_1 is the dipole current and I_2, the inverted L (imbalance) current.*

Sec 7.2 When to Use a Balun

Baluns have taken on a more significant role in the past few decades with the advent of solid-state transceivers and Class B linear amplifiers with unbalanced outputs. That is, the voltage on the center conductor of their output chassis connectors varies (plus and minus) with respect to ground. In many cases, coaxial cables are used as the transmission lines from these unbalanced outputs to antennas like dipoles, inverted Vs, and Yagi beams that favor a balanced feed. In essence, they prefer a source of power whose terminals are balanced (voltages being equal and opposite) with respect to actual ground or to the virtual ground that bisects the center of the antenna. The question that is asked most frequently is whether a 1:1 Balun is really needed.

To illustrate the problem involved and to give a basis for my suggestions, I refer you to **Figure 7-1**. Here we have, at the feedpoint of the dipole, two equal and opposite transmission line currents with two components each—I_1 and I_2. Also shown is the spacing, s, between the center conductor and the outside braid. Theoretically, a balanced antenna with a balanced feed would have a ground (zero potential) plane bisecting this spacing. However, because a coax-feed is unbalanced and the outer braid is also connected to ground at some point, an imbalance exists at the feedpoint giving rise to two antenna modes. One lies with I_1, providing a dipole mode; the other lies with I_2, providing an inverted L mode.

If the spacing, s, is increased, the imbalance at the feedpoint becomes greater—giving rise to more current on the outer braid and a larger imbalance of currents on the antenna's arms. Several steps can be taken to eliminate or minimize the undesirable inverted L mode (that is, eliminate or minimize I_2). The obvious choice is to use a well-designed Balun that not only provides a balanced feed, but also minimizes (by its choking reactance) I_2—if the coaxial cable does not lie in the ground plane that bisects the center of the dipole. The other step is to ground the coaxial cable at a quarter-wave (or odd-multiple thereof) from the feedpoint. This discourages the inverted L mode because any radiating element will want to see a high impedance at these lengths instead of the low impedance of a ground connection.

I conducted experiments with Baluns on a 20-meter half-wave dipole at a height of 0.17 wavelengths, which gave a resonant impedance of 50 ohms. VSWR curves were compared under various conditions. When the coaxial cable was in the ground plane of the antenna (that is, perpendicular to the axis of the antenna), the VSWR curves were identical with or without a well-designed Balun—no matter where the outer braid was grounded. A significant difference was noted only when the coaxial cable was out of the ground plane. When the cable dropped down at a 45-degree angle under the dipole, a large change in the VSWR took place. This meant that the inverted L mode was appreciable.

It should also be mentioned that the direction of I_2, the imbalance current, can depend upon the side on which the coaxial cable is out of the ground plane of the dipole. For example, if the cable comes down under the right side in **Figure 7-1** (that is, the angle between the horizontal arm and the coax is less than 90 degrees on the right side and more than 90 degrees on the left side), then the direction of I_2 can be reversed by the imbalance in the induced currents on the outside of the braid. By the same token, by having the coaxial cable coming down on the other side, the value of I_2 is only increased in magnitude.

However, feeding a Yagi beam without a well-designed 1:1 Balun is a different matter. Because most Yagi designs use shunt-feeding (usually by hairpin matching networks) in order to raise the input impedance close to 50 ohms, the effective spacing (s) is greatly increased. Furthermore, the center of the driven element is actually grounded. Thus, connecting the outer braid (which is grounded at some point) to one of the input terminals, creates a large imbalance and a real need for a Balun. An interesting solution, which would eliminate the matching network, is to use a step-down Balun designed to match 50-ohm cable directly to the lower balanced-impedance of the driven element.[2]

In summary, if you concur with the theoretical model of **Figure 7-1**, my experiments performed on 20 meters, and the reports from radio amateurs using dipoles and inverted Vs without Baluns, then it appears that 1:1 Baluns are really needed for: a) Yagi beam antennas where severe pattern distortion can take place without one, and b) dipoles and inverted Vs that have the coaxial cable feed lines out of the ground plane that bisects the antennas, or that are unbalanced by their proximity to manmade or natural structures. In general, the need for a Balun is not so critical with dipoles and inverted Vs (especially on 40, 80, and 160 meters) because the diameter of the coaxial cable connector at the feedpoint is much smaller than the wavelength.

If my model, which assumes that a part of the problem when feeding balanced antennas with coaxial cable is related to the size of the spacing, s (shown in **Figure 7-1**), then the possibility exists for using Ununs for matching into balanced antennas with impedances other than 50 ohms and with small values of s. For example, half-wave dipoles at a height of about a half-wave, quads, and center-fed 3/2-wave dipoles—which all have impedances close to 100 ohms—could very well be matched to 50-ohm cable by a 2:1 Unun. As I will show, they are considerably easier to construct than 2:1 Baluns. Furthermore, Genaille[10] has recently shown considerable success using Ununs in this kind of application.

To wind up this section, I would like to comment on an article published by Eggers,[11] WA9NEW, concerning the use of a Balun with a half-wave dipole. While at North Carolina State University, he conducted an experimental investigation of pattern distortion without a Balun at 1.6 GHz in an RF anechoic chamber (which simulates "free space"). Briefly, his results showed that, with a Balun (bazooka type), the antenna radiation pattern compared very favorably with the classic "figure-eight." Without the Balun, the radiation pattern was severely distorted.

Even though the author expressed difficulty in obtaining accurate measurements at this very high frequency, I have a question regarding the validity of performing the experiment in the first place. From the photograph in the article, it appears that conventional coaxial cable and connectors were used in the experiment. If we assume an effective diameter of 0.375 inches for these components, then scaling up to 3.5 MHz (457.14 fold) results in a coaxial cable with a diameter of 14.28 feet! I am quite sure that the large spacing, s, of 7.14 feet would bring about a noticeable imbalance resulting in appreciable pattern distortion even at 3.5 MHz.

Sec 7.3 Highlights of Significant Articles on 1:1 Baluns

Although there have been many articles on 1:1 Baluns published in the professional and amateur literature, I have selected for review a few that I believe have had the most impact on 1:1 Baluns for amateur radio use. As you will see, even though I consider some of the amateur articles significant, their impact upon the use and understanding of these devices has not always been positive. In fact, in some cases, the opposite has been true.

In the professional literature

There are actually only two significant articles in the professional literature that provide the fundamental principles upon which the theory and design of this class of transformers are based. It can be said that succeeding investigators simply extended the works of the authors of these two articles.

The first presentation on broadband matching transformers using transmission lines was given by Guanella in 1944.[3] He coiled transmission lines forming a choke such that only transmission line currents were allowed to flow, no matter where a ground was connected to the load. His single, coiled transmission line resulted in a 1:1 Balun. It is shown schematically in **Figure 7-2A**. Prior to this, RF Baluns were achieved by the use of quarter and half-wave transmission lines, and as a result, had narrow bandwidths. Guanella then demonstrated broadband Baluns with impedance transformations of $1{:}n^2$ where *n* is the number of transmission lines he connected in a series-parallel arrangement.

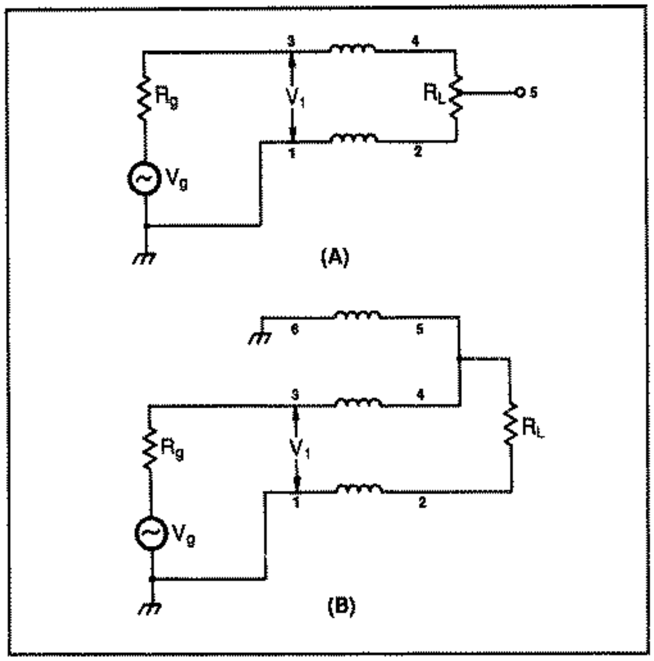

Figure 7-2. *Two versions of the 1:1 Balun: (A) the Guanella Balun and the basic building block; (B) the Ruthroff Balun as originally drawn.*

Several important points should be made regarding Guanella's 1:1 Balun shown in **Figure 7-2A**. With sufficient choking reactance, the output is isolated from the input and only flux-canceling transmission line currents are allowed to flow. With matched or very short transmission lines, the grounding of terminal 5 (actually or virtually like the center of a dipole) results in terminal 4 becoming $+V_1/2$ and terminal 2 becoming $-V_1/2$—creating a balanced output. This type of Balun has lately been called a "current" or "choke" Balun. A significant feature of this model is that a potential gradient of $-V_1/2$ exists along the length of the transmission line. This gradient, which exists on both conductors, accounts for practically all of the loss in these transformers because the loss mechanism is voltage dependent (a dielectric-type loss). All transmission line transformers have some sort of voltage gradient along their transmission lines and are, thus, subject to the same type of losses. Furthermore, the theory and loss mechanism is the same whether the transmission lines are coax or twinlead, or coiled around cores, or threaded through ferrite beads. Additionally, it was shown[2] that higher-impedance Baluns or Baluns subjected to higher VSWRs have more loss because the voltage gradients are also larger.

The second significant article on broadband transmission line transformers was published by Ruthroff.[9] His 1:1 Balun, which is shown as originally drawn in

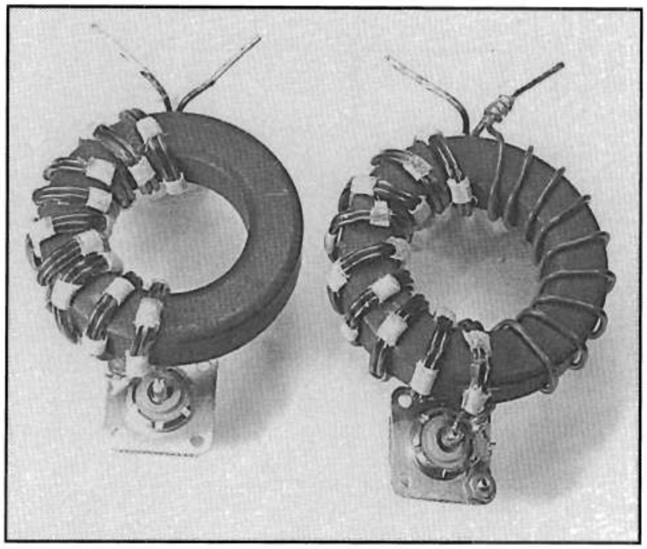

Photo 7-A. *The two basic forms of the 1:1 Balun that first appeared in the professional literature. The two-conductor Guanella Balun is on the left and the three-conductor Ruthroff Balun is on the right.*

Figure 7-2B, used an extra winding to complete (as he said) the path for the magnetizing current. Even though his schematic drawing appeared to look like a trifilar winding, his pictorial in the article clearly showed that the third winding (5-6) was on a separate part of the toroid. With an equal number of turns, it forms a voltage divider with winding (3-4) placing terminal 4 at $+V_1/2$ and terminal 2 at $-V_1/2$. In his classic paper, Ruthroff also presented his form of the 4:1 Balun (which is also different from Guanella's), a 4:1 Unun, and various hybrids. **Photo 7-A** shows the two basic forms of the 1:1 Balun that first appeared in the professional literature. The two-conductor Guanella 1:1 Balun is on the left and the three-conductor Ruthroff Balun is on the right. As was mentioned before, the Guanella Balun has also been called a "current" or "choke" Balun.

Before moving on to the significant articles in the amateur radio literature, some mention should be made of the differences between the two basic forms shown in **Photo 7-A**. Guanella's 1:1 Balun came to be known as the *basic building block* for this whole class of broadband transformers. This term was coined by Ruthroff as he showed its 1:1 Balun capability when the load was grounded at its center (terminal 5), and as a phase-inverter when the load was grounded at the top (terminal 4). By connecting terminal 2 to terminal 3 and connecting the bottom of the load to ground, Ruthroff demonstrated his very popular 4:1 Unun. I call this type of arrangement the bootstrap connection. By grounding terminal 2, there is no potential drop along the transmission line and, therefore, no need for magnetic cores or beads. This arrangement, which turns out to be an important function for extending the high frequency performance of this class of transformers, I call the "phase-delay" connection.

Thus, with the flexibility shown by Guanella's basic building block, a 1:1 Balun is now realized that not only presents a balanced power source to a balanced antenna system, but can also prevent an imbalance current (an inverted L antenna current) by its choking reactance when the load is unbalanced or mismatched or when the feedline is not perpendicular to the axis of the antenna.

Interestingly enough, except at the very low end of the frequency response of the Ruthroff 1:1 Balun where autotransformer action can take place, his Balun takes on the characteristics of the Guanella Balun. The reactance of the third winding becomes great enough to make it literally transparent. This is not the nature of the trifilar-wound (voltage) Balun, which is sensitive to unbalanced and mismatched loads over its entire passband because it is actually two tightly coupled transmission lines. This distinction was not recognized by most of those who published in the amateur radio literature.

In the amateur radio literature

R. Turrin, W2IMU—1964

The first presentation in the amateur radio literature on 1:1 Baluns using ferrite cores was by Turrin in 1964.[8] Turrin, who was a colleague of Ruthroff's at Bell Labs, took his small-signal design (which used No. 37 or 38 wire on toroids with ODs of 0.25 inches or less) and adapted it to high-power use. This was done by using thicker wire, larger cores, and (very importantly for high efficiency[2]) low permeability ferrite. Ruthroff used lossy manganese-zinc ferrites with permeabilities of about 3000 because efficiency was not a major consideration.

Figure 7-3 is a pictorial and a schematic of Turrin's design. As you can see, the third wire (winding 3-4) is placed between the two current-carrying wires (windings 1-2 and 4-5). **Photo 7-B** shows (on the left) his actual design using a ferrite core; a popular design (on the right) using a powdered-iron core is available in kit form from Amidon Associates, Inc. Both Baluns use 10 trifilar turns of a single-coated wire like Formex™ or Formvar™ on a toroid. Turrin's design

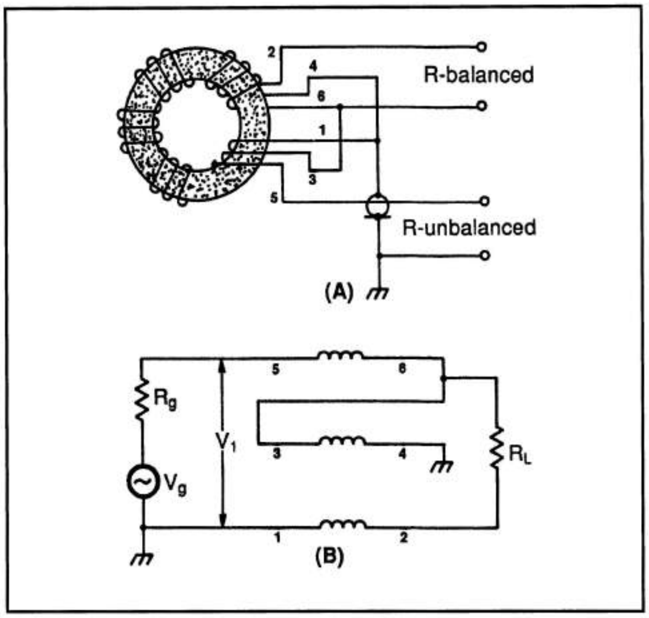

Figure 7-3. *(A) A pictorial of Turrin's 1:1 Balun, and (B) a schematic of his Balun.*

uses a ferrite toroid with an OD of 2.4 inches and a permeability of 40. The Amidon Associates' Balun uses a powdered-iron toroid with a 2-inch OD and a permeability of only 10. Both Baluns are specified to handle 1000 watts of power from 1.8 to 30 MHz.

Figure 7-4 shows the response curves for these two Baluns when terminated with 50-ohm loads. The response curve for a popular 1:1 rod-type Balun that uses the same schematic and wire is also shown. It has 8 trifilar turns, tightly wound on a rod of 0.5-inch diameter, 2.5 inches long, and with a permeability of 125. The rod-type Balun is shown in **Photo 7-C**.

Several important features should be brought out regarding the results shown in **Figure 7-4**. They are:

1. All Baluns had insufficient choking reactance and, hence, poor low-frequency responses. The powdered-iron version was especially poor. They all showed a drop in the input impedance and an inductive component at 2 MHz. This meant flux in the cores and an undesirable condition—especially for ferrite, which is a nonlinear material. Ferrite cores could not only suffer damage, but they could also generate spurious frequencies under these conditions. In fact, the same condition could occur at 4 MHz with a VSWR of 2:1! Therefore, I don't recommend any of these Baluns for use on 160 or 80 meters.

2. The major problem at the low-frequency end is the role of the third winding (3-4) in **Figure 7-3B**. It has been claimed[12] that the third winding improves the low-frequency response (over the two-conductor Guanella 1:1 Balun) because it enables autotransformer action at the low end. However, recent measurements I have made on two-conductor Guanella Baluns and three-conductor Ruthroff (or Turrin) Baluns, with loads grounded at their centers, show insignificant differences. This type of load approximates the actual condition when feeding a balanced antenna system. The negative feature of the third winding (3-4) is that, at the low-frequency end, there can be insufficient reactance to prevent harmful flux in the core because of a direct shunting path to ground. With the two-conductor 1:1 Balun, the only flux-inducing current is that of the imbalance current (the inverted L mode), which is usually far smaller.

3. Another important feature of the curves shown in **Figure 7-4** is the effect of the characteristic impedances of the coiled transmission lines. For example, a bifilar winding (wires tight together) on a toroid with spacing between adjacent bifilar turns exhibits a char-

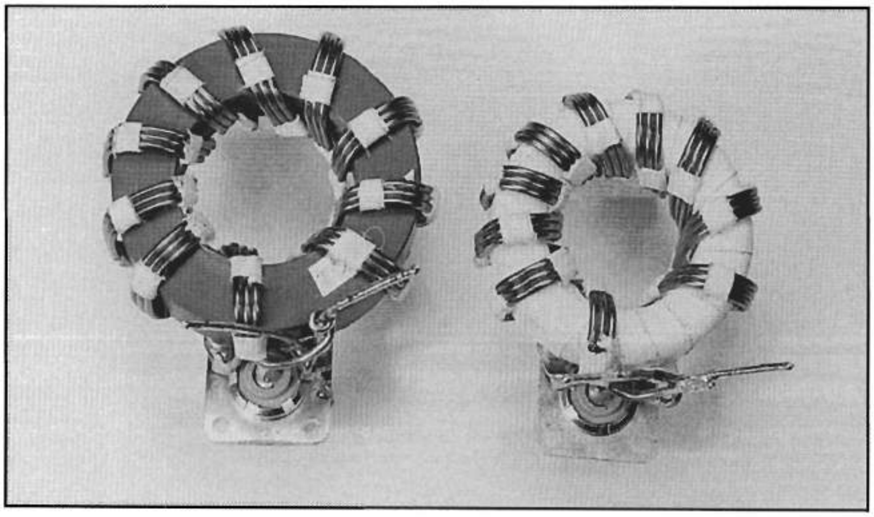

Photo 7-B. *Two versions of Turrin's design: On the left, the 1:1 Balun that has appeared in the amateur radio literature; on the right, a 1:1 Balun that has been readily available in kit form from Amidon Associates, Inc.*

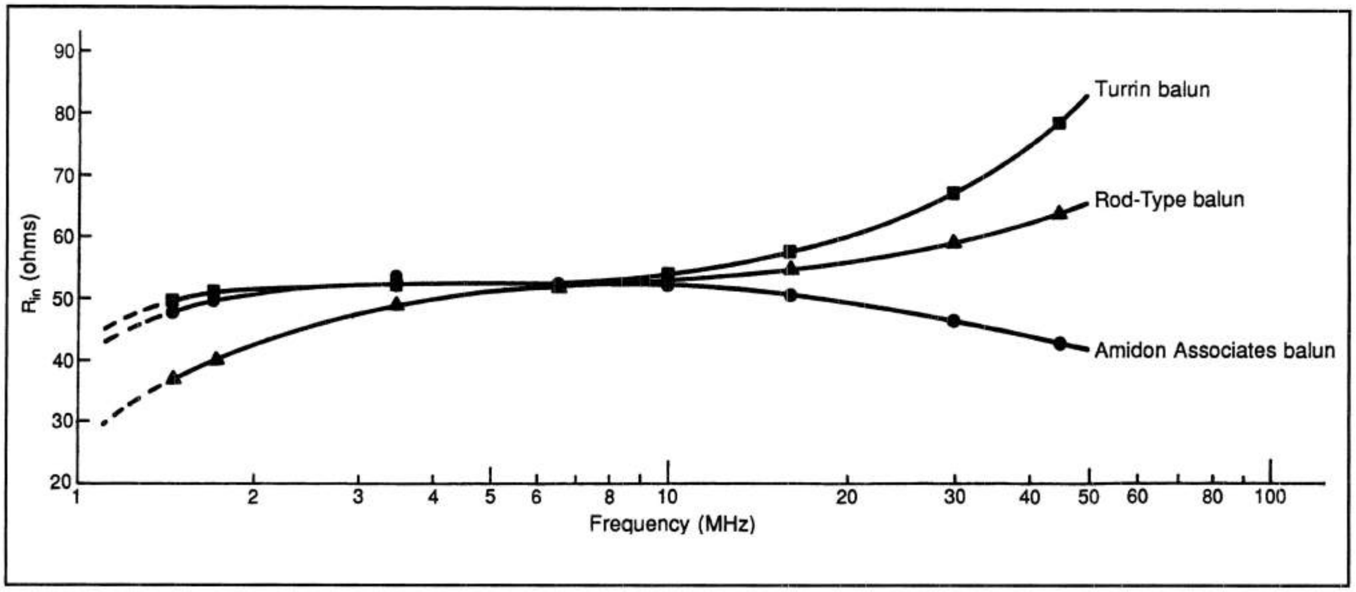

Figure 7-4. *The input impedance versus frequency, when terminated with 50 ohms, for the Turrin, typical rod-type, and Amidon Associates 1:1 Baluns.*

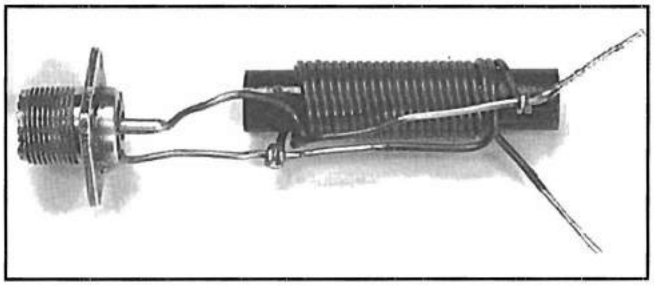

Photo 7-C. *A typical rod-type Balun.*

acteristic impedance of about 45 ohms. When wound on a rod with no space between adjacent bifilar turns, the characteristic impedance drops to about 25 ohms. With the third winding (3-4) between the other two as shown in **Figure 7-3B**, the characteristic impedance is raised to approximately 70 ohms in the toroidal case and to about 47 ohms in the rod case. If the toroidal Baluns were terminated in 70 ohms and the rod Balun terminated in 47 ohms, the high-frequency responses would be practically flat to at least 30 MHz. The difference in high-frequency response between the two toroidal Baluns (with 50-ohm loads) is due to the differences in the lengths of their transmission lines. The transmission line on the powdered-iron core is appreciably less because the OD, as well as the cross-sectional area, is smaller.

4. The trifilar-wound form of the 1:1 Balun also has an additional undesirable property. Its high frequency response is sensitive to unbalanced and mismatched loads. This is because the third wire now forms two tightly coupled transmission lines. It is unlike the Ruthroff version shown on the right in **Photo 7-A**. In his second article,[12] Turrin pointed out this important distinction.

J. Reisert, W1JR—1978

The next significant article on 1:1 Baluns was published by Reisert in 1978.[7] Reisert proposed winding some of the smaller (but still high-powered) coaxial cables around a 2.4-inch OD ferrite toroid with a permeability of 125. The windings also included a crossover, which is shown in **Figure 7-5** and **Photo 7-D**. In addition, he recommended various numbers of turns, depending upon the low-frequency requirement. For example, he suggested 12 turns to cover 3.5 MHz, 10 turns for 7 MHz, 6 for 14 MHz, and 4 for 21 and 28 MHz. Because the characteristic impedance of the coaxial cable is the same as the coax feedline, the Balun only introduces a foot or two of extra length to the feedline. This is true in the HF and VHF bands. The coaxial cables recommended in the article were RG-141/U, RG-142/U, and RG-303/U.

From the articles that followed in the amateur radio literature, it became apparent that few recognized all of the important features of Reisert's Balun. They are listed below.

1. An efficient, low-loss ferrite was used.
2. The Baluns had sufficient choking reactances for the various low-frequency requirements.

3. The characteristic impedance of the coiled transmission line was the same as that of the feedline, thus eliminating the extra transformer action of a length of transmission line with a different characteristic impedance.

4. The Balun is a form of Guanella's two-conductor 1:1 Balun, which is not prone to core flux and, hence, saturation and the generation of spurious frequencies. It is also not susceptible to mismatched and unbalanced loads as are the Turrin and "voltage" Baluns.

After constructing several of Reisert's Baluns and comparing them with other Guanella designs, I found that the crossover winding had virtually no effect up to 100 MHz (the limit of my equipment). In regards to Reisert's VSWR comparison with a rod-type Balun when feeding a triband Yagi beam on 20 meters, I found that his Balun had a lower VSWR (practically 1:1) at the best match point. The rod-type Balun had a best VSWR of about 1.3:1, but at a slightly higher frequency. He attributed the higher (and somewhat flatter) VSWR curve of the rod-type Balun to its greater ohmic loss. Because the rod-type Baluns I have investigated used the same low-loss ferrite that Reisert's did, I suspect that the differences in the VSWR curves were mainly due to the mismatch loss introduced by the rod-type Balun.

G. Badger, W6TC—1980

Badger published an in-depth, two-part series[13,14] in *Ham Radio* magazine in 1980 on air-core Baluns and Ununs. I am sure it was instrumental in advancing the technology of this class of wideband transformers. An article by Orr[15] also shows that there are many other radio amateurs who see the advantages of air-core transformers.

What are the claims for air-core Baluns over their ferrite-core counterparts? First and foremost, proponents of air-core Baluns claim they don't suffer the consequences of saturation that leads to spurious frequencies, heating, and ultimate damage. Secondly, it's said that they are not subject to arcing from the windings to the core.

What are the claims for the ferrite-core Baluns over their air-core counterparts? Simply put, they have wider bandwidths and are more compact.

After reading Badger's two-part series, I found that my curiosity was piqued by his experimental data on harmonic distortion due to saturation in a ferrite-core 1:1 Balun. Although many have expressed concerns regarding saturation in ferrite-core Baluns, Badger's data could very well provide the only results available. He used the two-tone test method, which combined two RF sources of 2.001 and 2.003 MHz, amplified it to 2 kW PEP and then fed it through a commercial 1:1 rod-type ferrite Balun. The data showed considerable distortion in the 3rd order and 9th order distortion products. In other words, appreciable nonlinearity took place at this high power level.

Several questions come to mind regarding these measurements. What was the low-frequency response of the commercial 1:1 rod-type Balun Badger used? From my measurements on a rod-type Balun (**Figure 7-4**), I found a drop in the input impedance and an inductive component at 2 MHz. This indicates flux in the core and a problem when using this Balun at 2 MHz. Because many rod-type 1:1 Baluns have been used over the years, it would have been instructive if he had also made these measurements at 4 and 7

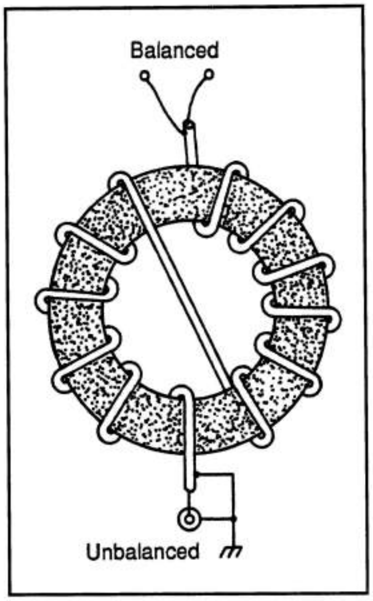

Figure 7-5. *A pictorial of the crossover used in Reisert's 1:1 Balun.*

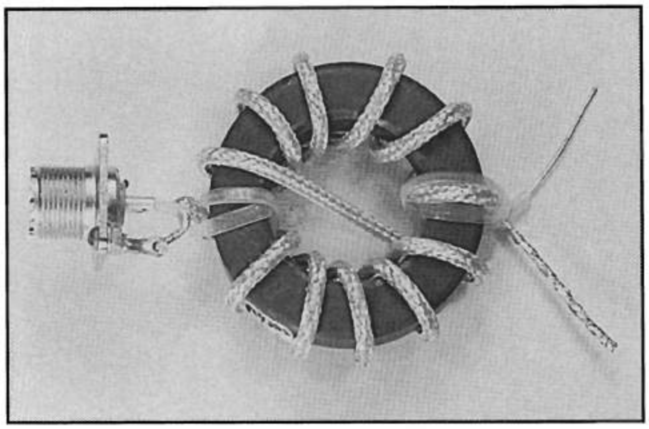

Photo 7-D. *A Reisert, W1JR, 1:1 Balun.*

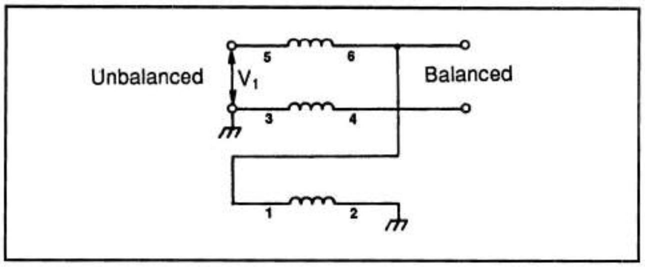

Figure 7-6. *Schematic of Badger's 1:1 Balun with a compensating winding (1-2). Winding (3-4) is the outer-braid of the coax and winding (5-6) is the inner conductor.*

MHz. They would have given the readers a safe lower-frequency limit for these Baluns.

I wondered why Badger didn't make similar measurements on Reisert's 1:1 Balun, which he included in his articles. As noted earlier, I consider Reisert's 1:1 Balun a very good design! I am sure that no distortion products would have been found at 2 MHz with this Balun. The end result is that Badger chose a poor ferrite-core design for making his comparisons. This helped to contribute to an undeserved reputation for the ferrite-core Balun.

Badger also suggested placing an insulated wire in parallel with the coax winding on Reisert's 1:1 Balun. He called this a compensating winding, which provided a superior balanced output. The third winding (1-2) is shown **Figure 7-6**. Winding (5-6) is the inner conductor and winding (3-4), the outer braid. Later experiments by myself and others have shown that a well-designed two conductor (Guanella) 1:1 Balun has a completely satisfactory balanced output for antenna applications. Furthermore, it does not suffer from an unbalanced and/or mismatched load and core saturation. Incidentally, Badger's schematic of **Figure 7-6** now adds up to four different versions of the 1:1 Balun. They are the two-conductor version of Guanella's and the three, three-conductor versions of Ruthroff's, Turrin's, and (now) Badger's.

Badger and Orr also mentioned the Collins Balun in their articles. This Balun is comprised of a dummy length of coax wound as a continuation of the original coiled coax winding. Interestingly, it is connected as a Ruthroff 1:1 Balun (**Figure 7-2B**), which also uses a third winding. Because there is appreciable coupling between the two coiled windings, the Collins Balun should also be susceptible to mismatched and/or unbalanced loads. Badger claimed it was, by far, the best 1:1 Balun he had ever used. Again, it would have been very informative if he had compared it with the Reisert Balun (without the compensating third wire).

M.W. Maxwell, W2DU—1983

One of the more significant articles on 1:1 Baluns was published by Maxwell[4] in 1983. Maxwell introduced, what he called, the "choke" Balun. It was formed by placing high-permeability ferrite beads over about one foot of small (but high-powered) coaxial cables similar to the ones used in the Reisert Balun. **Photo 7-E** shows the W2DU "choke" Balun removed from its plastic enclosure.

Maxwell compared his Balun with (what he termed) a "transformer-type" Balun by measuring the input impedances versus frequency when the outputs were terminated in 50 ohms. The "transformer-type" Balun didn't yield a true 1:1 impedance transfer ratio, which he claimed was due to losses, leakage reactance, and less than optimum coupling. Because Maxwell gave no description of the "transformer-type" Balun, I assumed it was the popular rod-type Balun shown in **Photo 7-C**. As you can see in **Figure 7-4**, this Balun has a poor low-frequency response. Furthermore, it is really optimized for a load of 47 ohms and not 50 ohms.

What Maxwell failed to realize was that his Balun was a form of Guanella's two-conductor type. That is, it is both a choke (a lumped element) and a transmission line (a distributed element). Additionally, Guanella's theory applies whether the transmission lines are coiled (about a core) or beaded, twin-lead or coaxial cable. From Ruthroff's classic paper,[9] which extended Guanella's work,[3] we became aware of the

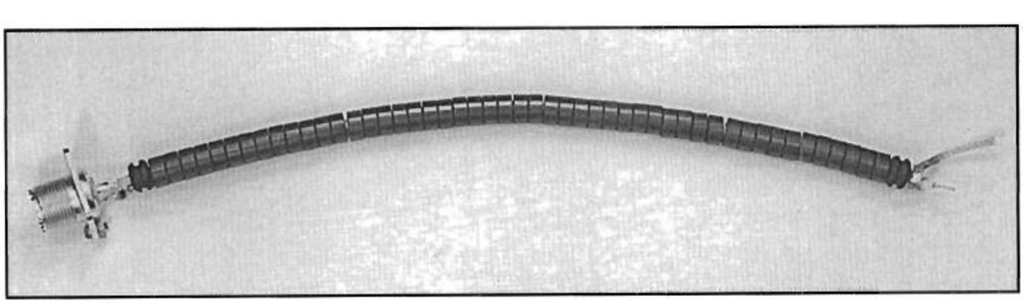

Photo 7-E. *The Maxwell, W2DU, "choke" 1:1 Balun.*

voltage drops along the lengths of the transmission lines. From very accurate insertion loss measurements,[2] we learned that the losses were mainly in the magnetic medium—and that they were related to the voltage levels and the permeabilities. Maxwell didn't take into account these latter findings. He used lossy high-permeability beads (2500) and assumed that the main loss was in the transmission line. He claimed that the CW power-handling capability of his Balun was 3.5 kW at 50 MHz and 9 kW at 10 MHz—the same as the coaxial cable itself. I seriously question these power ratings. Ironically, it is very likely that Maxwell's Balun had more real loss than the so-called "transformer-type" Balun!

R.W. Lewallen, W7EL—1985

There is very little doubt that Lewallen's interesting article[6] in 1985 contributed significantly to the better understanding and design of 1:1 and 4:1 Baluns. In it, he coined the (now very popular) terms "voltage" and "current" Baluns. The "voltage" Balun, a three-conductor type, has output ports which have voltages that are balanced to ground. It is brought about (see **Figure 7-6**) by the voltage-divider action of windings (5-6) and (1-2). Because we have two tightly coupled transmission lines in the passband with the same potential gradients, terminal 6 is at $+V_1/2$ and terminal 4 at $-V_1/2$ where V_1 is the input voltage. The "current" Balun, on the other hand, is a two-conductor Balun that produces equal and opposite currents on the output ports for any form of load impedance.

Lewallen conducted a series of experiments on 10 meters to compare the performances of "voltage" and "current" Baluns under balanced and unbalanced conditions. In the unbalanced (nonsymetrical) condition, the dipole was lengthened by five inches on one side and shortened by five inches on the other side. He then obtained a figure of merit for both Baluns (as well as for the case without a Balun) defined as the ratio of the average magnitude of the currents at the feedpoint over the magnitude of the imbalance (the inverted L) current. The magnitudes of the currents were obtained by current-probe toroids. Measurements were made at the antenna feedpoint and at a half-wave (physically) from it.

The "current" Balun consisted of 15 turns of very small RG-178/U coax on an FT82-61 core (a ferrite toroid with an OD of 0.825 inches and a permeability of 125). The "voltage" Balun had 10 turns of RG-178/U coax with a No. 26 wire in parallel (closely coupled) on the same toroid. The schematic is shown in **Figure 7-6**.

Lewallen concluded (and I agree) that his experiments clearly showed that the "current" Balun gave superior performance at every measured point in each experiment. However, the "voltage" Balun still improved the balance over the no-Balun case. He also concluded that other experiments should be performed in order to better compare the two forms of the Balun. One is to ascertain the difference when the feedline is placed nonsymetrical with respect to the antenna (to induce an imbalance current into the feedline). Others include determining the optimum point in the feedline to place the Balun, and testing the various kinds of core and beaded Baluns.

Although Lewallen's article pretty much speaks to Badger's proposal of adding a third wire to Reisert's Balun for better balance (i.e., avoid it), there are some comments and questions I have regarding his experiments and findings. They are:

1. Why didn't Lewallen use Reisert's Balun as the "current" Balun and Badger's suggested third-wire design as the "voltage" Balun? These would have been more realistic designs for comparisons. Instead, he used very small structures (which will only find use in QRP operations) and, as such, have higher frequency capabilities. Also, because the "voltage" Balun only had 10 turns (and hence a shorter transmission line and a poorer low-frequency response), it was favored in the comparisons on 10 meters. Had Lewallen used transmission lines of equal lengths on Reisert's cores, the differences between the two Baluns would have been even more dramatic.

2. It would also have been very useful if Lewallen had made comparisons between a "current" Balun that could handle the full legal limit of amateur radio power (again like the Reisert Balun) and "voltage" Baluns using Turrin's schematic (**Figure 7-3**) with rod and toroidal cores which have been readily available for nearly three decades.

3. Additionally, comparisons should not only be limited to 10 meters. Because 1:1 "voltage" Baluns are configurations of coupled transmission lines with various characteristic impedances, their performances with mismatched and unbalanced loads are more sensitive to the higher frequencies than their "current" Balun counterparts. Therefore, making similar measurements on 20 meters would also provide more useful information.

4. Even though Ruthroff's classic 1959 paper[9] has been the industry standard over the years, his 1:1

Balun design has been practically nonexistent in the amateur literature. Turrin mentioned its advantage over his first design in his second article.[12] But Turrin's first design has prevailed in our amateur literature. Because Ruthroff's design has the third conductor on a separate part of the toroid, it has the balanced output mentioned by Badger,[13] but still retains the flexibility of the Guanella[3] Balun. In other words, as the frequency is increased, the choking action of the third wire makes it practically transparent. This enables it to handle any form of load impedance. It would have been informative if Lewallen had pointed this out and also noted that Ruthroff's 1:1 Balun, although looking like a "voltage" Balun, is really a "current" Balun.

5. Lewallen and the others who have published in the amateur radio literature have failed to reference the first presentation on what are now known as "current" or "choke" Baluns, made by Guanella[3] in 1944. Even though Guanella used coiled transmission lines without a magnetic core, his theory on how these devices work is still applicable today.

J.S. Belrose, VE2CV—1991

The last article on 1:1 Baluns I considered worth mentioning was written by Belrose[5] in 1991. In it, he described the W2DU Balun by Maxwell and how his technique of threading coaxial cable through ferrite beads could be easily applied to 4:1 and 9:1 Baluns.

What immediately grabbed my attention in this article was the *deck head* (the editor's comments), which included highly complimentary remarks regarding the beaded-coax Balun. In essence, it said, "In this breakthrough article, W2DU's peerless 1:1 current-Balun design serves as the basis for excellent ferrite-bead-choke current Baluns capable of 4:1 and 9:1 impedance transformation."

However, if one reads the article carefully, it becomes apparent that this is not what Belrose said. His words were, "The *current Balun* of the *type* developed by Walt Maxwell, W2DU—a Balun consisting of ferrite beads slipped over a length of coaxial cable—is the best so far devised (italics mine)." He did not say that W2DU's Balun was "peerless." In fact, in the article he said just the opposite. He pointed out that the W2DU Balun's main disadvantage is that the beads are lossy at HF and heating becomes a concern when the transmitting power exceeds 125 watts! For high power (that is 1-kW CW), Belrose recommended Roehm's[16] designs, which use lower permeability (850) beads nearest the Balun's balanced output (where most of the heating takes place).

However, I do question two of the advantages he claims for the W2DU Balun. They are:

1. Its excellent power-loss and impedance-versus-frequency characteristics are much superior to those of a bifilar current Balun wound on a ferrite toroid.
2. It has excellent power-handling capability, and can function quite satisfactorily when working into highly reactive loads. This is so because the magnetic flux produced by currents flowing on this Balun's wires cannot saturate its ferrite beads.

Belrose obtained evidence for advantage number 1 by comparing the input impedance and power loss versus frequency of the W2DU Balun with a commercial Balun when they were terminated in 50 ohms. The commercial Balun was a bifilar wound toroidal type used in a differential-T tuner. What Belrose failed to realize was that the commercial Balun had heavily insulated wires, resulting in a characteristic impedance greater than 100 ohms. Thus, he was actually comparing a 50-ohm transmission line with a longer line that had a characteristic impedance in excess of 100 ohms! As expected, his input impedance versus frequency curve for the commercial Balun was even more severe than that of the Turrin Balun shown in **Figure 7-4**.

Advantage number 2 is based upon the premise that the magnetic flux produced by currents on the W2DU Balun's wires cannot saturate the ferrite beads, while the windings of a bifilar wound toroidal current Balun can. This is an incorrect assumption because the magnetic flux of a two-conductor type Balun, like the beaded-coax or the bifilar-wound toroidal Balun, is generated by the imbalance (inverted L) current and hence is *much* lower than the transmission line currents. This is especially true with sufficient choking reactances. The perception that the toroidal type Balun still transmits the energy to the output circuit by flux linkages, could very well lead to this mistaken impression.

For high power beaded-coax Baluns, Belrose referred to designs by Roehm,[16] who increased the power capability of this type of Balun by using lower permeability beads near the balanced output. He also increased the length considerably. For operation from 80 meters to 10 meters, Roehm used 28 inches of beaded coax. For 160 meters to 10 meters, he used 36 inches of beaded coax. Belrose's suggestion of connecting beaded coaxes in parallel on the low-imped-

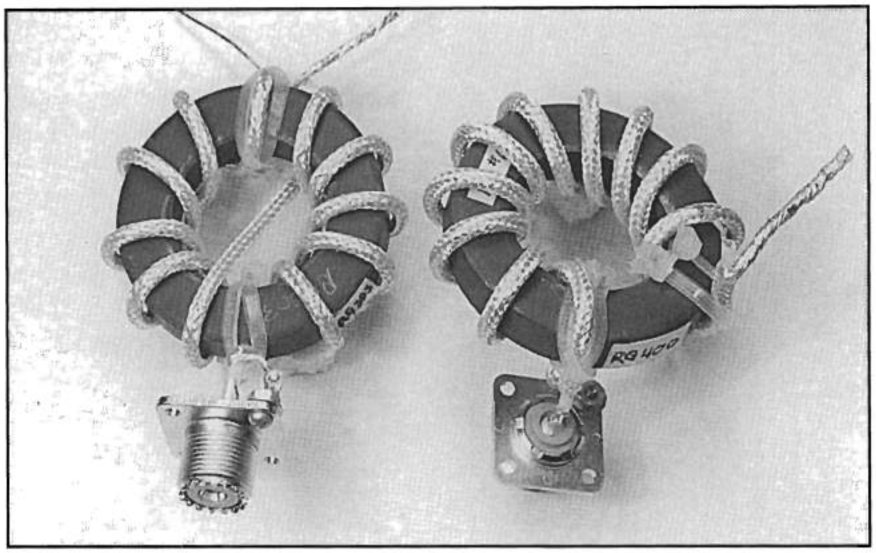

Photo 7-F. *Two versions of Reisert's 1:1 Balun. The Balun on the left uses the crossover shown in Figure 7-5. The Balun on the right is continuously wound. Both have the same electrical performance in the HF band.*

ance side and in series on the high impedance side to obtain a broadband 4:1 transformation ratio would require transmission lines with characteristic impedances of 100 ohms. This means, for a high power 4:1 Balun using beaded transmission lines, about 56 inches of beaded line would be required for the 80-meter to 10-meter operation and 72 inches for the 160-meter to 10-meter coverage. For a 9:1 Balun, it would be necessary to increase these lengths by 50 percent!

It remains to be seen what Belrose would have said or done if he had compared the W2DU Balun of Maxwell's with the W1JR Balun of Reisert's. He certainly couldn't claim the advantages listed in his article for the W2DU Balun. Would he still have claimed that the type of Balun developed by Maxwell is the best so far devised? Given the evidence, I doubt it.

Sec 7.4 High-, Medium-, and Low-power Designs

In this section I'll present my latest 1:1 Balun designs. Except for one Balun that appeared in the June 1993 issue of *CQ*, the others appeared for the first time in the magazine's April 1994 issue. Because I have favored Reisert's design throughout this chapter, the first Baluns described here are my versions of his technique of coiling small (but high power) coaxial cable around a low-permeability ferrite toroid. For my wire versions, I could have used all sorts of adjectives to describe them like Guanella, two-conductor, choke, and current. However, in the process of writing this section, I thought Belrose's adjectives were the most direct. Using his words, I call my wire versions of the 1:1 Balun simply—*bifilar toroidal Baluns*.

Photo 7-F shows two versions of Reisert's Balun. The one on the left uses the crossover shown in **Figure 7-5**. Because no difference in performance at HF was noticed without the crossover, a continuous-wound version is also shown on the right. The main advantage in the HF band with the crossover winding is purely mechanical. Having the input and output connections on opposite sides of the toroid is not only more convenient, but it also offers a much stronger method of mounting.

For operation from 1.8 to 30 MHz, 10 turns of small coax like RG-303/U, RG-142B/U, or RG-400/U are wound on a 2.4-inch OD ferrite toroid with a permeability of 250. If the use is limited from 3.5 to 30 MHz, then a permeability of 125 is recommended because it would yield a slightly higher efficiency at the high end. If one wants the highest possible efficiency and limits the operation from 14 to 30 MHz, then a permeability of 40 is recommended. With loads grounded at their centers, these conditions were found to give ample margins (handle a VSWR of 3:1 without any appreciable flux) at their low frequency ends.

For ease of winding, I found TY-RAP™ Cable Ties *very* useful. Two were used at each end. Removing the covering on the outer braid also helps. Because about 24 inches of cable is wound on the toroid, I recommend you start with at least 32 inches. Of the three cables noted above, I found RG-303/U cable the easiest to wind and connect. Although it only has a single-thickness braid (the others have double-thickness braids), its power rating is still the same—9 kW at 10 MHz and 3.5 kW at 50 MHz.

The next high power design is shown in **Photo 7-G** mounted in a 4 inch long by 3 inch wide by 2.25 inch

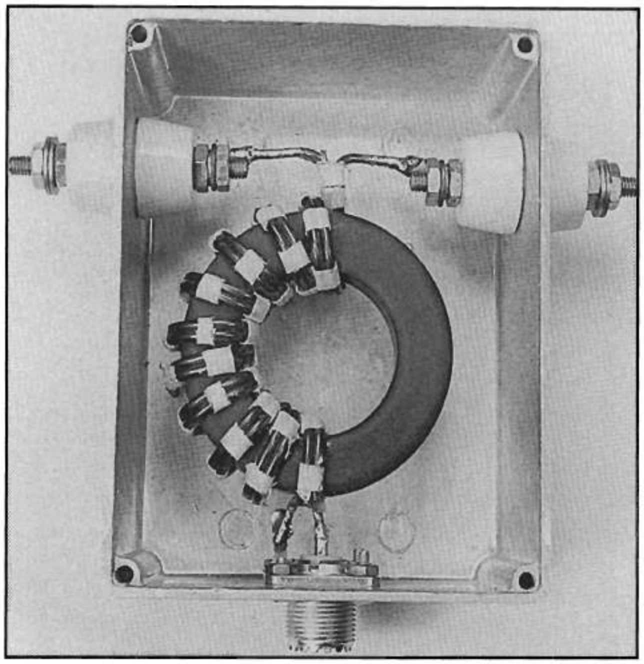

Photo 7-G. *My high-power design of a bifilar toroidal (Guanella/current) 1:1 Balun mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud aluminum box.*

high Bud CU 234 aluminum box. It has 10 bifilar turns of No. 12 H Thermaleze wire on a 2.4-inch OD ferrite toroid. As with the Reisert versions, a permeability of 250 is recommended for 1.8 to 30 MHz, 125 for 3.5 to 30 MHz, and 40 for 14 to 30 MHz. One wire is also covered with two layers of Scotch No. 92 polyimide tape in order to raise the characteristic impedance to 50 ohms. With this added insulation, the voltage breakdown of this twin-lead transmission line compares very favorably with RG-8/U cable (4000 volts). In order to preserve the spacing, the wires are also clamped together about every 1/2 inch with strips of Scotch No. 27 glass tape 3/16 inches wide and a little over 1 inch long.

Two "economy" versions of the high power bifilar toroidal Balun are shown in **Photo 7-H**. The one on the left shows the windings crowded on one-half of the toroid. The one on the right provides the same positions of the input and output connections by using the crossover. Their performances are identical. Both Baluns have 10 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid. The choices of permeability, which trade off bandwidth for efficiency, are the same as those used in the two previous high power designs. The word "economy" here refers to economy in labor—this Balun is actually very easy to construct.

This "economy" Balun, which also handles the full legal limit of amateur radio power, has a small tradeoff in high frequency response. Because no extra insulation is used, the characteristic impedance of two tightly clamped No. 14 H Thermaleze wires is 45 ohms. With one layer of Scotch No. 92 tape, it increases to 50 ohms. But for most of the HF band, the difference in performance between Baluns using transmission lines of 45 and 50 ohms should be negligible. Even without the extra insulation, the voltage breakdown should compare very favorably with the smaller, high-power coaxes used in the Reisert versions (1900 volts).

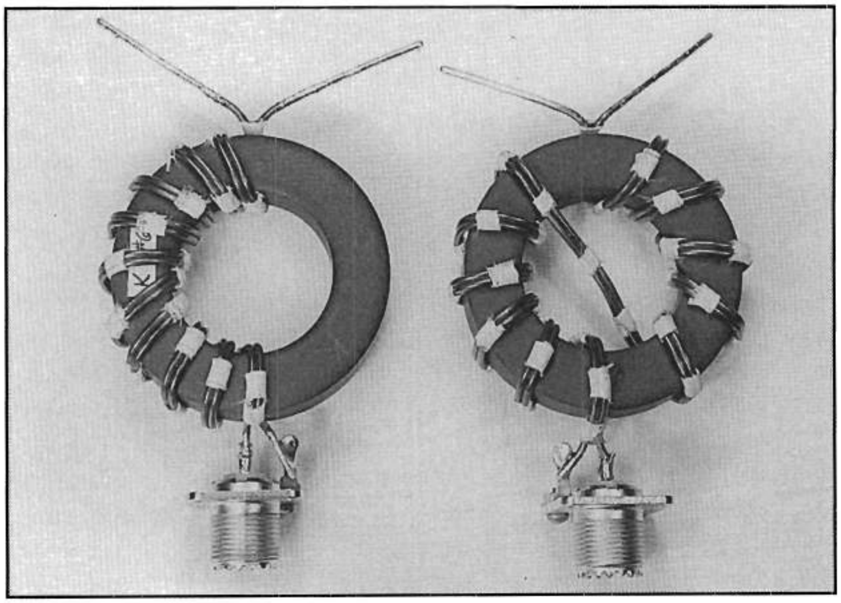

Photo 7-H. *Two "economy" versions of the high-power bifilar toroidal (Guanella/current) 1:1 Balun. The one on the right uses Reisert's crossover technique.*

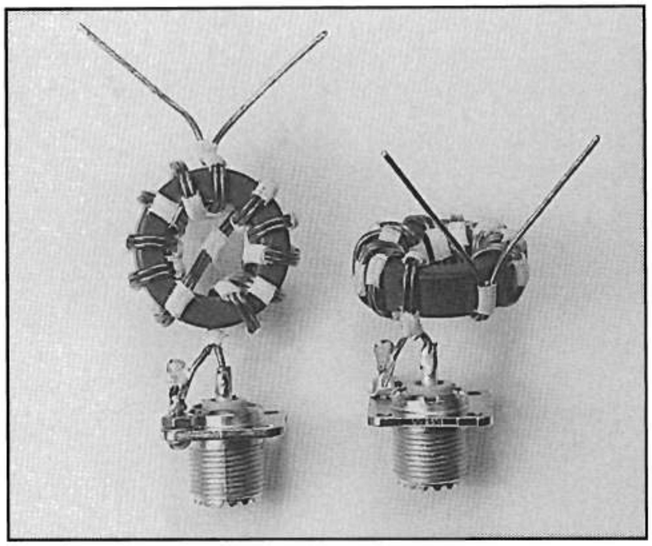

Photo 7-I. *Two low-power versions of the bifilar toroidal (Guanella/current) 1:1 Balun capable of handling the output of practically any HF transceiver. The Balun on the left has the crossover.*

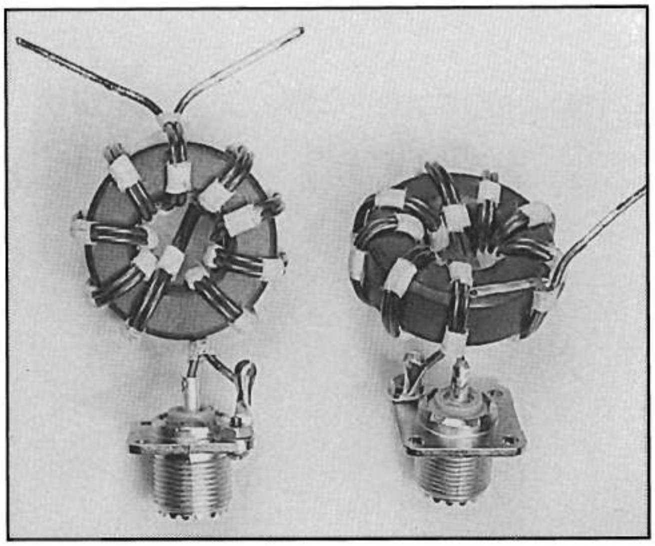

Photo 7-J. *Two medium-power versions of the bifilar toroidal (Guanella/current) 1:1 Balun capable of handling the full legal limit of amateur radio power when the VSWR is less than 2:1. The Balun on the left has the crossover.*

Photo 7-I shows two low-power versions of a bifilar toroidal Balun capable of handling the output of practically any HF transceiver. One has a crossover winding and the other, a continuous winding. They both have 10 bifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. Since efficiency is not a major problem in low-power use, I found no reason to suggest the other two versions, which use lower permeabilities. It is also interesting to note that two tightly clamped No. 16 H Thermaleze wires have a characteristic impedance close to 50 ohms. Therefore, this small Balun (particularly with its short leads) has a very good high frequency response.

Photo 7-J shows two medium-power versions of a bifilar toroidal Balun capable of handling the full legal limit of amateur radio power under controlled conditions—when the VSWR is less than 2:1. Being smaller than its larger (2.4-inch OD) counterpart, its heat-sinking capability and, hence, power rating is less. As before, one Balun uses a crossover while the other doesn't. Each has 8 bifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid. The ferrite permeabilities and expected bandwidths are the same as the other high-power Baluns. Because the average magnetic path length in the core is about two-thirds that of the 2.4-inch core, only 8 bifilar turns are required in order to produce a similar low-frequency capability. Even though the characteristic impedances of their bifilar windings are 45 ohms, their responses on 10 meters should be somewhat better than the "economy" models because the lengths of their transmission lines are shorter (18 compared to 24 inches).

And now a few words on what sort of efficiency one can expect in trading-off low frequency response by using lower permeability ferrite cores. From earlier studies,[2] it was found that the efficiency (with sufficient choking, so only transmission line currents flow) is related to the permeability, the voltage drop along the length of the transmission line, and the frequency. The higher the permeability and/or voltage drop, the greater the loss. Additionally, the higher the permeability, the greater the loss with frequency. It was also found that a permeability of less than 300 was necessary in order to obtain the very high efficiencies of which these devices are capable.

Here are some efficiencies that might be expected from ferrites under matched conditions, based on the results of the studies:

1. With 250 material, an efficiency near 99 percent at 1.8 MHz and 97 percent at 30 MHz.
2. With 125 material, an efficiency near 99 percent at 3.5 MHz and 98 percent at 30 MHz.
3. With 40 material, an efficiency of 99 percent at 14 MHz and at 30 MHz.

When a Balun is exposed to a high impedance resulting in a VSWR of 2:1, the voltage, and hence loss, increases by about 40 percent. With a VSWR of

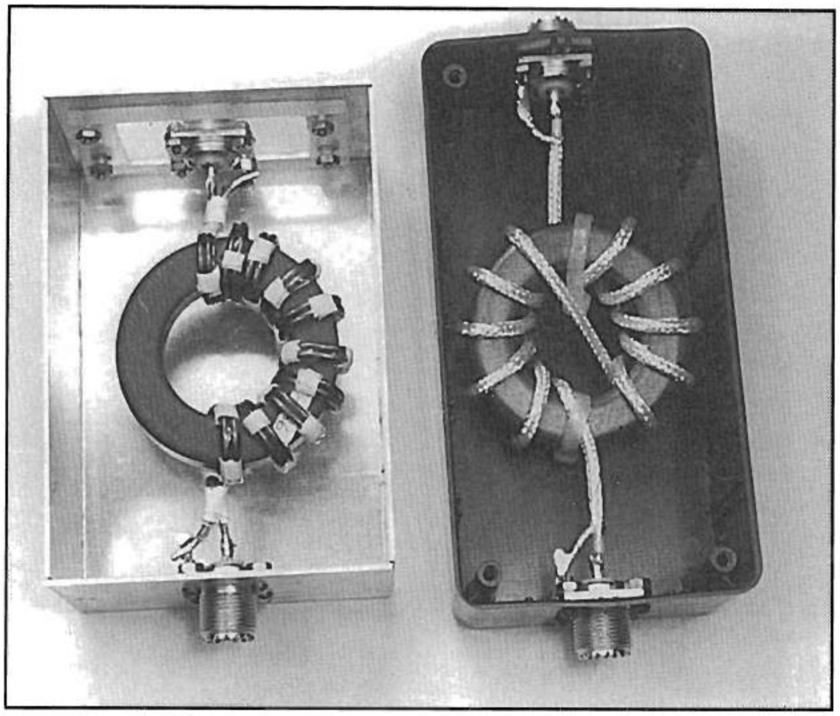

Photo 7-K. *Two examples of mounting the high-power 1:1 Guanella Balun in enclosures that allow isolation transformer operation.*

4:1, the loss doubles. With a VSWR of 10:1, the loss is more than three-fold. Since limited data was obtained in this study,[2] these increases in losses with increases in VSWR could very well be even greater.

Sec 7.5 Isolation Transformers

As was shown in the preceding sections, the 1:1 Guanella (current/choke) Balun is in reality a two-conductor RF choke, or an isolation transformer. However, practical Balun designs don't usually lend themselves to isolation use for mechanical reasons. They cannot be conveniently inserted into a coaxial cable system. Therefore, the objective is to mount the "Guanella Baluns" in enclosures that will allow them to perform as isolation transformers.

Photo 7-K shows two examples of enclosures that should accomplish the mission. The example on the left shows one of the high-power units mounted in a 5 inch long by 3.5 inch wide by 2.25 inch high aluminum enclosure. The output connector (on the top) is insulated by a 0.125-inch-thick piece of plastic. The one on the right shows another high-power unit mounted in a 6 inch long by 3 inch wide by 2 inch high plastic enclosure, which is available from RadioShack. The descriptions of the high-power units are given in the preceding section.

These isolation transformers can be used as Baluns when inserted in a coaxial cable one-half wavelength (physically) from a half-wave dipole[6] or between a coaxial cable and a balanced L-C tuner.[16] They should never be inserted in a coaxial cable system that presents a high impedance and, consequently, a high voltage. This could be harmful to the unit.

Note that the high-power unit on the left in **Photo 7-K** can also be used as a filament choke in a Class B linear amplifier. Its low-frequency response is much better than any rod-type, assuring 160-meter operation. Obviously, it would not be mounted in the aluminum enclosure shown in the photo. In addition, the two layers of Scotch No. 92 tape would not be required because the characteristic impedance of the winding is unimportant and the voltages involved are quite low.

By the way, the designs discussed above can be used as phase-inverters. Simply connect the hot lead on one side to the ground on the other side.

Sec 7.6 Summary

In investigating the 1:1 Balun, I was quite surprised to see the ferrite- and powdered-iron-core designs that have been available in the literature and off-the-shelf since 1964. They not only had poor low- and high-frequency responses, but they were also susceptible to flux in the cores at their low-frequency ends. Furthermore, since they only used single-coated wires, they were also prone to voltage breakdown. No doubt, these designs are responsible for the poor reputation that the Balun has had for many years.

It wasn't until 1978, when Reisert published his article, that a Balun which had all of the attributes of a good design became available. Namely:

a) It is efficient because it uses a low-permeability core.

b) It has sufficient choking reactance to meet its low-frequency requirement.

c) It is not prone to flux in the core (and, hence, saturation) because it has no third winding.

d) It has a 50-ohm characteristic impedance and maintains a 1:1 transformation ratio with a 50-ohm load.

e) It has a good voltage breakdown capability (1900 volts).

f) It can handle a mismatched and/or unbalanced load.

However, succeeding investigators failed to see the advantages of Reisert's design and proposed their own. Surprisingly, they belonged to two distinct groups. One group favored "air-core" Baluns and the other, "choke" (beaded-coax) Baluns.

The main argument given by the "air-core" followers was that their Balun would never experience problems with saturation, while the "ferrite-core" Balun would. However, the Reisert Balun is a current/choke type Balun that could only have flux in the core due to the imbalance (inverted L) current, which is much smaller than the transmission line currents. In fact, with any degree of choking reactance by the coiled transmission line, the imbalance current is essentially negligible. Therefore, saturation is not a concern with a Balun like Reisert's. In all fairness, it should be noted that it is a different story with the 4:1 current/choke and voltage Baluns. All three of these types of Baluns have a "magnetizing inductance" in their low-frequency models and hence a possibility of saturation with a poor design.

The advocates of the "choke" 1:1 Balun claim that their beaded-coax Balun can't saturate while the bifilar (current) toroidal Balun can. This is untrue; they are basically the same kind of structure, and neither has a third conductor which could allow a flux-causing current at the very low-frequency end. But of all of the attributes listed above for the Reisert Balun, the first one has the "choke" Balun at a disadvantage in the HF band. Because its transmission line is not coiled about a toroid, it does not have the multiplication factor of N^2 (due to mutual coupling), where N is the number of turns, while the toroidal Balun does. Therefore, higher-permeability beads are required in order to obtain sufficient choking reactance. This results in lower efficiency.

Finally, it should be pointed out that the 1:1 Balun using the small, but high-power, coaxial cable is capable of 5-kW operation in the HF band. Because of current crowding, the bifilar toroidal Balun, even with No. 12 wire, has shown excessive wire heating at this level. Therefore, No. 10 or No. 8 wire, with added layers of Scotch No. 92 tape in order to obtain a 50-ohm characteristic impedance, is recommended.

Chapter 8

The 4:1 Balun

Sec 8.1 Introduction

Chapter 7 in this book discussed the most popular of all Baluns—the 1:1 Balun designed to match 50 ohms unbalanced to 50 ohms balanced. It not only gave a review of the history, theory, and design of these broadband transformers, but also my viewpoint on published articles advocating later designs using coaxial cable wound around a toroid, threaded through ferrite beads, or just plain coiled in air (an air-core Balun). As was noted, I am in considerable disagreement with many of the claims advanced for these later 1:1 Baluns.

Chapter 8 deals with the next most popular Balun —the 4:1 Balun designed to match 50 ohms unbalanced to 200 ohms balanced. Baluns matching 50 ohms unbalanced to 12.5 ohms balanced are also included. It begins with a little information on the history and design of these Baluns, followed by high- and low-power designs and comparisons with other Baluns that have been on the market or in the amateur literature. Guanella's approach to the design of a 4:1 Balun is particularly noteworthy. It not only yields an excellent Balun design, but also an Unun (unbalanced-to-unbalanced transformer) with practically the same performance. His design could very well be called a *Balun/Unun*. As in the first chapter, this one also closes with a brief summary of the significant points.

Sec 8.2 A Little History and Design Information

There are really only two classic papers that have established the principles upon which the transmission line transformer (the Balun being a subset thereof) is based. The first one was written by Guanella in 1944. Guanella proposed the idea of coiling a transmission line to isolate the input from the output, resulting in the (now popular) *current* or *choke* Balun.[3] The second was by Ruthroff in 1959, whose analysis of these transmission line transformers is the present industry standard.[9] Ruthroff also introduced the Unun and the hybrid transformer.

Interestingly enough, both Guanella and Ruthroff had different approaches to their 1:1 and 4:1 Balun designs. Guanella used a two-conductor 1:1 Balun design, while Ruthroff used a three-conductor design. Ruthroff's third conductor (which was said to increase the low-frequency response over the two-conductor Balun[12]) lay on a separate part of a toroidal core. Investigators who followed failed to recognize this fact. Their comparisons were made with a three-conductor Balun that had the third wire in parallel with the other two, which then formed two coupled transmission lines. This gave rise to the term *voltage* Balun—an inferior design.[6]

However, the differences between Guanella's and Ruthroff's approaches to 4:1 Baluns were even more striking. Guanella connected coiled transmission lines in a parallel-series arrangement, so in-phase voltages were summed at the high impedance side. His Balun has been called a *current* Balun.[6] Ruthroff, on the other hand, obtained a 4:1 transformation ratio by summing a direct voltage with a delayed voltage that traversed a single transmission line in a phase-inverter connection (see **Chapter 7**). His Balun has been called a *voltage* Balun. The distinction between the operation of these two Baluns was also overlooked by practically everyone who followed.

This section reviews the two different approaches taken by Guanella and Ruthroff in obtaining 4:1 Baluns. Of particular importance are the descriptions of the potential drops along the lengths of the transmission lines when the loads are grounded at various points. These voltage drops are not only relevant to the ohmic losses in the Baluns, but also to their electrical performances. These descriptions were quite

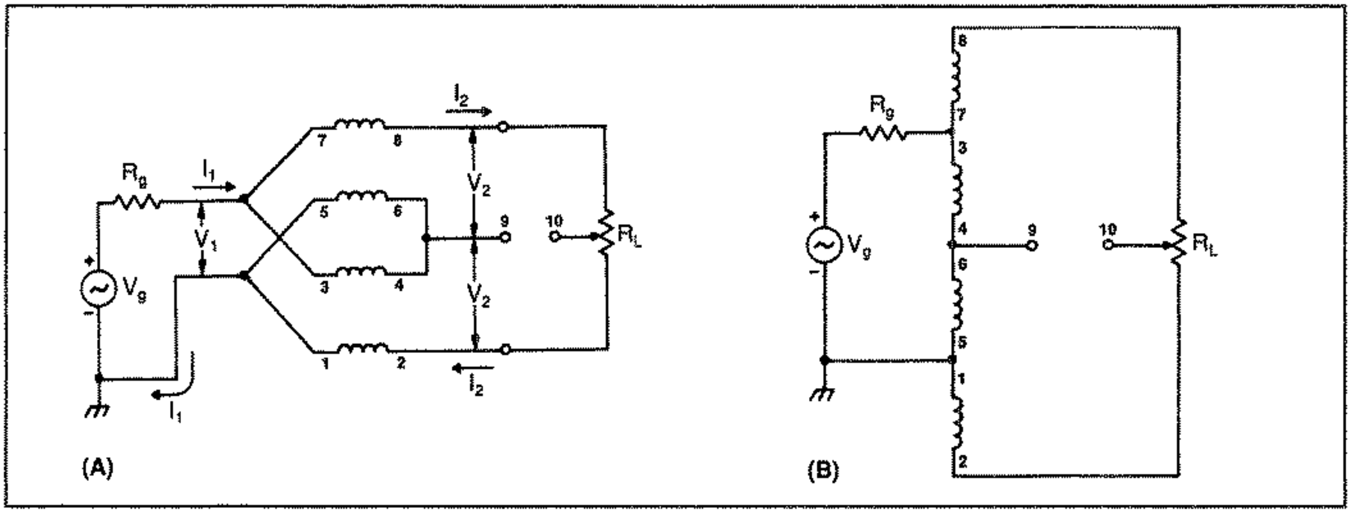

Figure 8-1. *Electrical models of the Guanella 4:1 Balun: (A) high-frequency, (B) low-frequency.*

possibly presented for the first time in the second edition of my book *Transmission Line Transformers.*[2]

Sec 8.2.1 Guanella's 4:1 Balun

Figures 8-1A and **B** show the high- and low-frequency models of Guanella's method of connecting transmission lines in parallel-series to obtain a 4:1 Balun. The high-frequency model (**Figure 8-1A**) assumes that the choking reactances of the coiled (or beaded) transmission lines are sufficient to isolate the input from the output, so only transmission line currents are allowed to flow. This occurs when the reactance of windings 3-4 and 5-6 (which are in series) is much greater than R_g (at least by a factor of ten.[2] If two cores are used, the reactance is the sum of the reactances of windings 3-4 and 5-6. If a single core is used, the reactance is twice as large because of the mutual coupling between the windings. The other advantage of Guanella's method (besides only using one core) is that shorter transmission lines can be used, resulting in better high-frequency performance.

As with all transmission line transformers, the objective is to have the transmission lines see loads equal to their characteristic impedances, resulting in "flat lines." This yields the highest frequency response. Because each transmission line in **Figure 8-1A** sees one half of the load, R_L, the optimum value of the characteristic impedance is $R_L/2$. Consequently, the input impedance, V_t/I_1, is simply the impedance of two identical transmission lines connected in parallel. It follows that the impedance transformation ratio is the load, R_L, divided by the input impedance.

Because the Guanella 4:1 Balun sums voltages of equal delays from identical transmission lines, his Balun is only limited in high-frequency performance by the deviation of the characteristic impedance of the transmission lines from the optimum values and the parasitics not absorbed into the characteristic impedance of the lines. I (and practically everyone else) had overlooked the simple and important statement, "a frequency independent transformation," which appeared in Guanella's 1944 article[3]—a fact that is evidenced by the scarcity of his designs in the literature. Another interesting aspect of the Guanella 4:1 Balun is the analysis of his Balun when the load is floating or grounded at different points. This leads to the determination of the voltage gradients that exist along the transmission lines and the various functions of which his 4:1 design is capable. Assuming a matched load or very short transmission lines resulting in $V_2 = V_1$, they are as follows.

Floating Load

With terminal 10 (which is at the center of R_L) floating, the potential gradient along the top transmission line in **Figure 8-1A** (windings 5-6 and 7-8) is $-1/2V_1$; along the bottom transmission line (winding 1-2 and 3-4) it's $-3/2V_t$. The voltage to ground on terminal 9, V_{90}, is $-1/2V_1$. Because the bottom transmission line (in **Figure 8-1A**) has a voltage drop along its length three times greater than the top transmission line, it results in three times more loss because losses in transmission line transformers are voltage dependent (dielectric-type losses).[2]

Even though a single-core Guanella 4:1 Balun maintains the voltages (stated above) when feeding a folded dipole (of about 200 ohms) which has a virtual-ground potential at terminal 10, it still feeds equal currents to each side of the antenna because of the series-connection at its output. Since the output voltages are not balanced to ground, a reactive component is probably introduced into the input impedance. Additionally, the choking reactance of the windings also prevents antenna currents from flowing on the outside of the coaxial cable feedline.

Load Grounded at Center

When two cores are used and terminal 10 (the center of R_L) is grounded, the voltage gradient along the top transmission line in **Figure 8-1A** is zero, and along the bottom transmission line it is $-V_1$. The voltage to ground on terminal 9 (V_{90}) is also zero. In fact, the core for the top transmission line isn't needed. It merely acts as a mechanical support for the top transmission line, which now only operates as a delay line. Also, all of the loss now occurs in the core of the bottom transmission line where a longitudinal potential gradient exists. Furthermore, the low-frequency response, as seen from **Figure 8-1B**, is now determined by the reactances of windings 1-2 and 3-4. This means that the low-frequency response with a floating load is better by a factor of two over the case where the load is grounded at its center.

But the single-core case is a different matter. Since the potential at terminal 9 (V_{90}) wants to be at $-1/2V_1$, connecting a ground directly to the center of R_L causes an imbalance that renders the single-core Balun unusable. If the ground were placed at a point 25 percent below terminal 8 (50 ohms from terminal 8 with a 200-ohm load), no difference would be noted from a floating load. This condition also exists when two cores are used.

Load Grounded at the Bottom

It is when the load is grounded at the bottom (at terminal 2), that we have what is probably the most interesting case. The 4:1 Balun (with two cores) is now converted into a very broadband Unun (unbalanced-to-unbalanced transformer). Because the bottom transmission line in **Figure 8-1A** has no potential drop along its length, it only acts as a delay line. The voltage to ground at terminal 9 (V_{90}) is $+V_1$, and the voltage gradient along the top transmission line is $+V_1$. This results in a voltage of $2V_1$ across the load. The low-frequency response is now determined by the reactances of windings 5-6 and 7-8. This is just the opposite of the Balun case when the center of the load was grounded. A single-core 4:1 Guanella Balun can also be converted to an Unun by putting a 1:1 Balun (for isolation) in series with the 4:1 Balun.[2]

There is a reason for claiming a very broadband response for a Guanella Unun (converted from his Balun)—the two in-phase voltages are now summed at the high-impedance side. The only other competition for a 4:1 Unun design is that of Ruthroff's,[9] where a direct voltage is summed with a delayed voltage that traversed a single transmission line (and, hence, had a built-in, high-frequency cut-off). In fact, very little information can be found in the literature on a *Guanella 4:1 Unun*.

Sec 8.2.2 Ruthroff's 4:1 Balun

Figures 8-2A and **B** show the high- and low-frequency models of Ruthroff's approach for a 4:1 Balun. The high-frequency model (**Figure 8-2A**) assumes that the

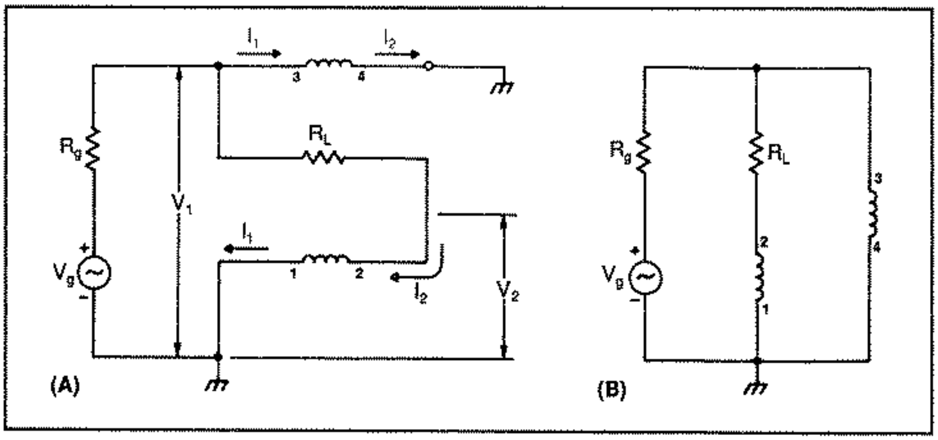

Figure 8-2. *Electrical models of the Ruthroff 4:1 Balun: (A) high-frequency, (B) low-frequency.*

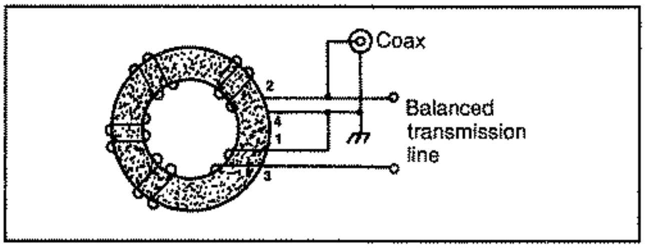

Figure 8-3. *A pictorial of the connections for a 4:1 Ruthroff (voltage) Balun.*

choking reactance of the coiled (or beaded) transmission line is sufficient to isolate the input from the output, so only transmission line currents are allowed to flow. This occurs when the reactance of winding 3-4 (or 1-2, because they are the same) is much greater than R_g (at least by a factor of ten[2]).

As **Figure 8-2A** indicates, the transmission line is connected in a *phase-inverter* function (see **Chapter 7**). That is, a $-V_1$ voltage gradient now exists along the length of the transmission line. Therefore, the voltage across R_L now becomes $V_1 + V_2$. Although Ruthroff analyzed his 4:1 Unun in his classic paper,[9] his results also apply to his Balun with a floating load, because both devices sum a direct voltage with a delayed voltage. In essence, he used loop equations on the input and output and transmission line equations to eliminate one set of variables (I_2 and V_2). Ruthroff also used a maxima technique (setting a derivative to zero) to solve for the optimum characteristic impedance of the transmission line. As in the Guanella case, he found the optimum value to be $1/2R_L$.

An inspection of **Figure 8-2A** shows that the left side of R_L (terminal 3) has a direct voltage (V_1) to ground, and the right side (terminal 2) a delayed voltage ($-V_2$) to ground, which traveled the length of the transmission line. Also note, that if the line is electrically one-half wavelength long, the output is *zero*. As a result, Ruthroff's design (which has a built-in cutoff) is sensitive to the transmission line length.

I have recently unearthed another interesting aspect of Ruthroff's design.[2] If the center of the load is grounded, the high-frequency performance is vastly improved. The built-in high-frequency cut-off is eliminated and the Balun appears to take on the character of a Guanella Balun that sums voltages of equal delays. A closer inspection reveals that the input impedance now consists of two impedances in parallel: one consisting of $R_L/2$, and the other of a "flat line" terminated by $R_L/2$. As a result, the currents are not in phase!

Surprisingly, when matching into a folded dipole with an input impedance of 200 ohms, Ruthroff's Balun exhibits a high frequency response that is much greater than expected (see later sections). Because the folded dipole has a *virtual* ground at its center, the Balun could very well be summing voltages of equal phases.

Sec 8.2.3 Amateur Radio History and Design

Looking back at old issues of amateur radio handbooks (I don't have a complete set), I found that the first presentation on broadband 4:1 Baluns appeared in the 1955 edition of *The ARRL Handbook.* The section was called "Coil Baluns." The schematic diagram was that of Guanella's, shown in **Figure 8-1A**. What surprised me was that this section appeared to use many of the important words contained in Guanella's article.[3] It mentioned that the choking action of the coiled transmission lines should be great enough to isolate the input from the output at the lowest frequency of interest. It also included the requirement on the characteristic impedance of the coiled transmission lines; namely, that the characteristic impedance should be equal to $R_L/2$, where R_L is the load.

However, the section also included two other statements which are not correct in light of today's design practices. One recommended that the length of the winding in each coil be equal to about a quarter wavelength. The other stated that the principal application is in going from a 300-ohm balanced load to a 75-ohm coaxial line. With magnetic cores, the lengths of the windings are now considerably shorter than a quarter wavelength, and the applications include a host of different impedance levels.

Recent issues of the handbooks now include the 4:1 broadband *coil Balun* (along with the same write-up that appeared in the 1955 issue), and one with windings on ferrite cores. They are now called 4:1 air-core current Baluns and "just plain" 4:1 current Baluns (ferrite cores being assumed). Unfortunately, what is lacking in the description of the 4:1 current Balun is information on the importance of the characteristic impedance of the windings and the value of the permeability of the ferrite cores. The literature states that 8 to 10 turns (of No. 14 Formvar-coated, close-spaced—I guess) on a toroidal core or 10 to 15 turns on a rod are typical values for the HF range. Ferrites with permeabilities from 850 to 2500 are also suggested. Nothing is mentioned regarding the dimensions of the cores.

In essence, there's very little information available today in our handbooks that would help one understand and construct the "popular" *current* Balun. Even the choices of recommended ferrites to be used are found wanting. Accurate loss measurements^2 have shown ferrites with permeabilities of 850 to 2500 to be lossy in Balun and Unun applications. Only when the permeabilities of ferrites are 300 or less will Baluns and Ununs exhibit the very high efficiencies of which they are capable.

Even though the 4:1 *voltage* Balun has actually had a shorter history than the *current* Balun, considerably more construction detail (including an actual photograph) has been available in the amateur radio handbooks. As far as I can tell, the first presentation took place between 1965 and 1968. In looking through succeeding issues (including the 1993 issue), I find the write-up hasn't changed much (if any) over the years. The 4:1 Balun from the handbooks and a commercial rod-type are described and compared with my designs in the succeeding sections.

Sec 8.3 4:1 Ruthroff Balun Designs

As mentioned above, the schematic of Ruthroff's Balun is shown in **Figure 8-2A**. A pictorial representation is shown in **Figure 8-3**. **Photo 8-A** (on the left) shows my construction of a design close to the one shown in the handbook's photographs. It has 10 close-spaced, bifilar turns of No. 14 Formvar-coated wire on a 2.4-inch OD ferrite toroid with a permeability of 40. **Figure 8-4** shows a plot of the input impedance versus frequency when the 200-ohm load is center-tapped-to-ground (which is close to the actual case when matching into balanced antenna systems). As you can see, when compared to a design that has the proper characteristic impedance of the winding and sufficient choking, the response is very poor. Although this Balun has been rated at 1000 watts of RF power from 1.8 through 60 MHz, I would suggest it not be used below 6 MHz for fear of excessive flux in the core (especially when the magnitude of the load is greater than 200 ohms). Also, above 14 MHz, the transformation becomes considerably greater than 4:1.

My design, on the right in **Photo 8-A**, has 14 bifilar turns of No. 14 tinned copper wire on a 2.4-inch OD ferrite toroid with a permeability of 125 or 250. The wires are threaded through No. 13 Teflon tubing with a wall thickness of 20 mils. As you can see by its excellent high frequency response in **Figure 8-4**, the characteristic impedance of the bifilar winding must be very close to the ideal value of 100 ohms. **Photo 8-B** shows two different views of my design mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud CU 234 aluminum box. The Balun, which is placed equidistant between the top, bottom, and sides of the enclosure, is securely mounted by soldering its leads to the two feedthrough insulators and the SO-239 chassis connector.

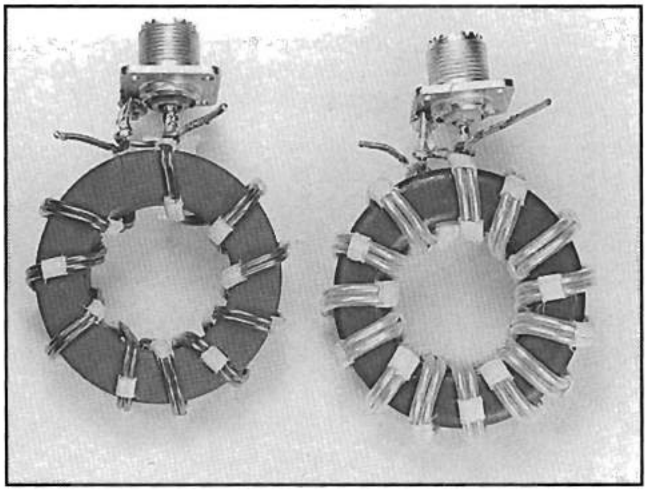

Photo 8-A. *Two designs of the 4:1 Ruthroff (voltage) Balun. The one on the left is taken from the amateur radio handbook. The one on the right is my improved version.*

It should be mentioned that if the Balun is to be used mainly on the lower portion of the HF band (including 160 meters), then the 250 permeability ferrite is recommended. Even though the difference in low frequency response between permeabilities of 125 and 250 doesn't show up in **Figure 8-4**, the 250 permeability would provide an extra safety margin (from flux in the core) at the low frequency end. The tradeoff lies in giving up a little in efficiency (about 1 percent) for an increase in the safety margin (a factor of 2) at the low end.

Incidentally, the handbooks also state that the Balun can be used between a balanced 300-ohm point and a 75-ohm unbalanced line. Because I suspected this statement as well, I again measured the input impedances versus frequency of both Baluns when terminated in a 300-ohm center-tapped-to-ground load. **Figure 8-5** shows the deterioration that takes place, especially at the high end. Even a Balun that is well designed for a 50:200-ohm impedance level is not recommended

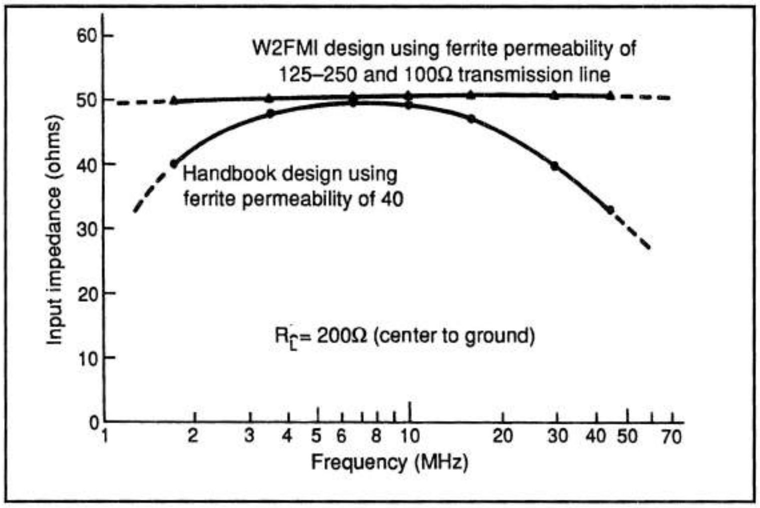

Figure 8-4. *The input impedance versus frequency for a 4:1 Ruthroff (voltage) Balun design from the amateur radio handbook and one optimized for the 50:200-ohm level. The load is grounded at its center.*

for the 75:300-ohm level. Because the length of the transmission line becomes significant beyond 10 MHz, standing waves then change the impedance ratio due to the mismatch with the Balun's transmission line. My design also shows more safety margin at the low end. I'm surprised that these simple measurements weren't made many years ago.

However, the Balun shown in the handbook does have one interesting feature. It uses a very low permeability ferrite (40), which has been shown by very accurate insertion loss measurements[2] to yield efficiencies in Baluns (and Ununs) of *99 percent* at the

Figure 8-5. *The input impedance versus frequency for the two Ruthroff Baluns of Figure 8-4, but with a 300-ohm load. Note the deterioration of the W2FMI design, which was optimized for the 50:200-ohm level.* ↓

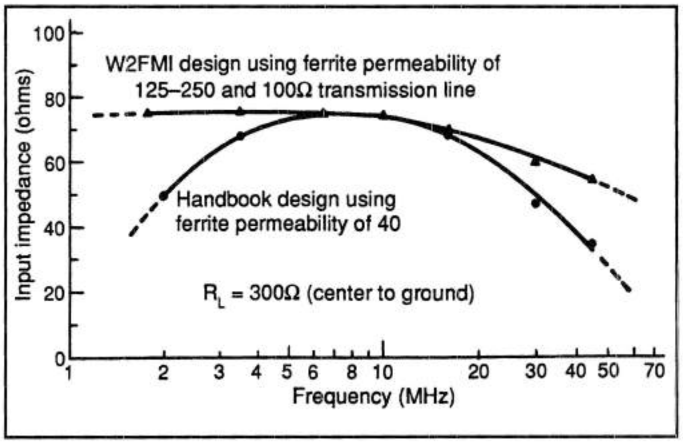

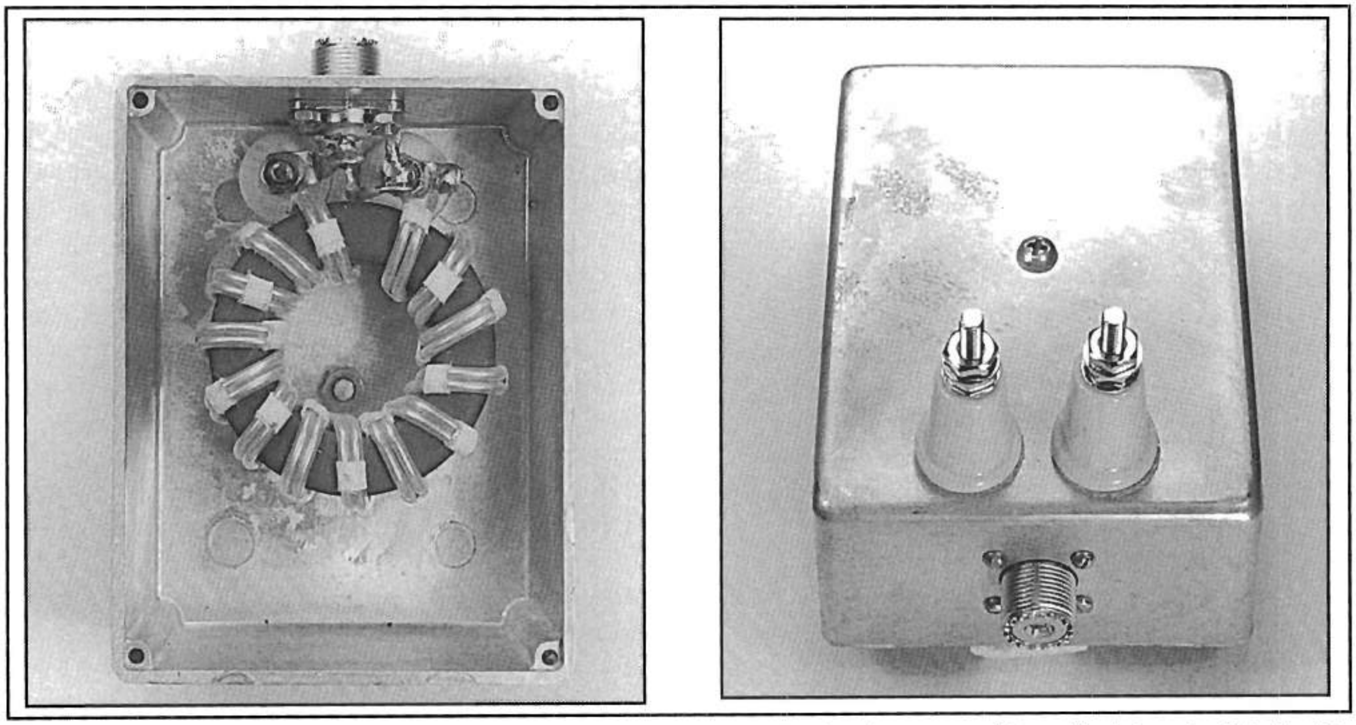

Photo 8-B. *Two different views of the optimized version of the 4:1 Ruthroff Balun mounted in a 4 inch long by 3 inch wide by 2.25 inch high Bud CU 234 aluminum enclosure.*

50:200-ohm impedance level! This is even a percent or two better than the ferrite with a permeability of 125. Because this ferrite permeability is so low, the major problem lies in obtaining sufficient choking reactance at the lowest frequency of interest, so that only transmission line currents are allowed to flow.

The design chosen (in order to exploit this very high efficiency) is shown **Photo 8-C**. It uses 14 bifilar turns of the same wire, as with my previous Balun shown on the right in **Photo 8A**, on two 2.4-inch OD cores (bound together with No. 27 glass tape) with permeabilities of 40. The unmounted view shows how the two cores are bound together by glass tape. The other views attempt to give an example for mounting the Balun. The Balun is supported by two acrylic end pieces which are, in turn, held fast to the enclosure by a long bolt. The Balun is placed equidistant between the top, bottom, and sides of a 5 inch long by 3.5 inch wide by 2.25 inch high aluminum enclosure. A few washers at the point where the bolt comes through the enclosure help to position the Balun between the top and bottom.

When matching 50 ohms (unbalanced) to 200 ohms (balanced), the response of this Balun is practically the same as mine, shown in **Figure 8-4** using a single core. From 1.7 to 30 MHz, it can certainly handle the full legal limit of amateur radio power with an efficiency close to 99 percent. However, if the operation of this Balun is restricted to the HF band only (that is, from 3 to 30 MHz), then it could be rated conservatively at 10 kW of peak power and 5 kW of average power. It would be an ideal Balun for a log-periodic beam antenna.

Finally, **Photo 8-D** shows three different views of a low power 4:1 Ruthroff (voltage) Balun designed to handle the output power of any HF transceiver easily. It has 10 bifilar turns of No. 18 hook-up wire on a 1.5-inch OD ferrite toroid with a permeability of 250. The enclosure is a 2.75 inch long by 2.125 inch wide by 1.625 inch high CU 3000-A minibox.

Section 8.4 4:1 Guanella Balun Designs

Photo 8-E shows two high-power Guanella 4:1 Baluns designed to match 50-ohm coaxial cable to loads of 200 ohms. They both use No.14 H Thermaleze wire with a covering of Teflon tubing giving characteristic impedances very close to 100 ohms (the objective). Their responses are flat from 1.5 MHz

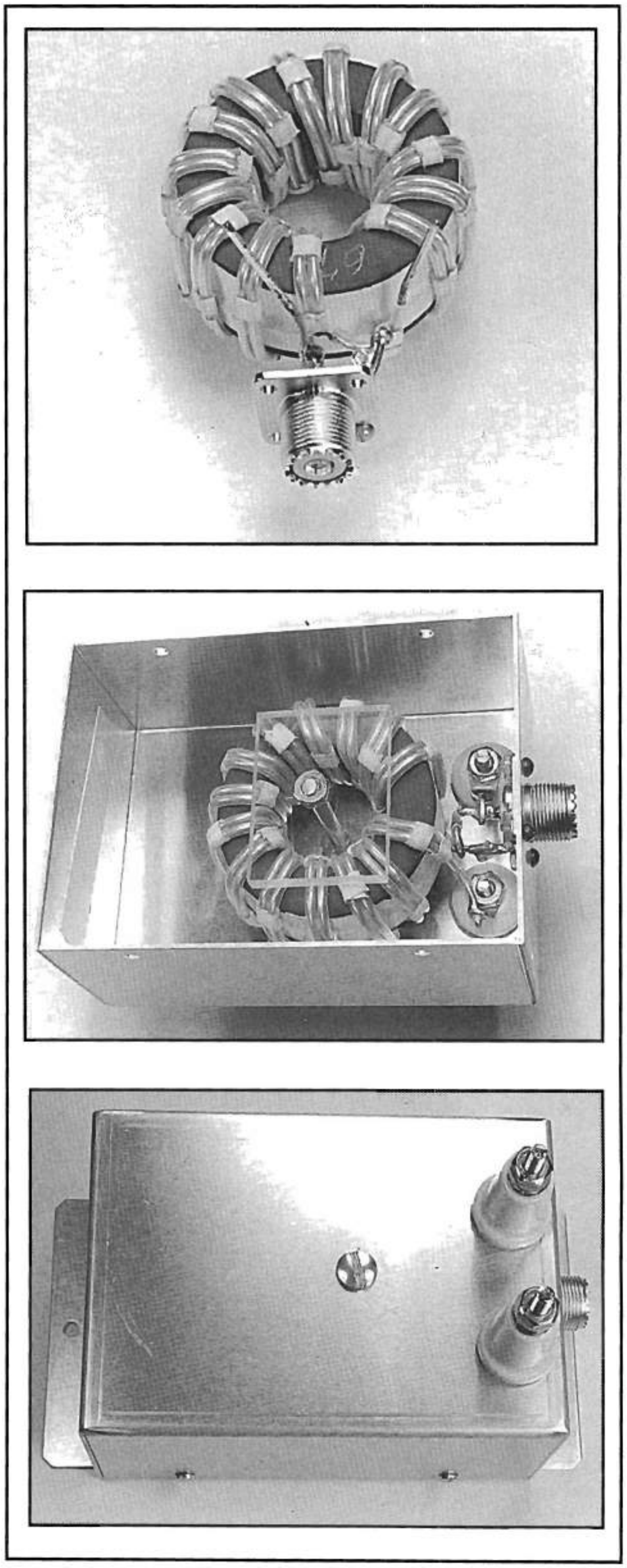

Photo 8-C. *Three views of the very high-power 4:1 Ruthroff Balun using two low-permeability (40) ferrite cores. Dimensions of the aluminum enclosure are: length, 5 inches; width, 3.5 inches; height, 2.25 inches.*

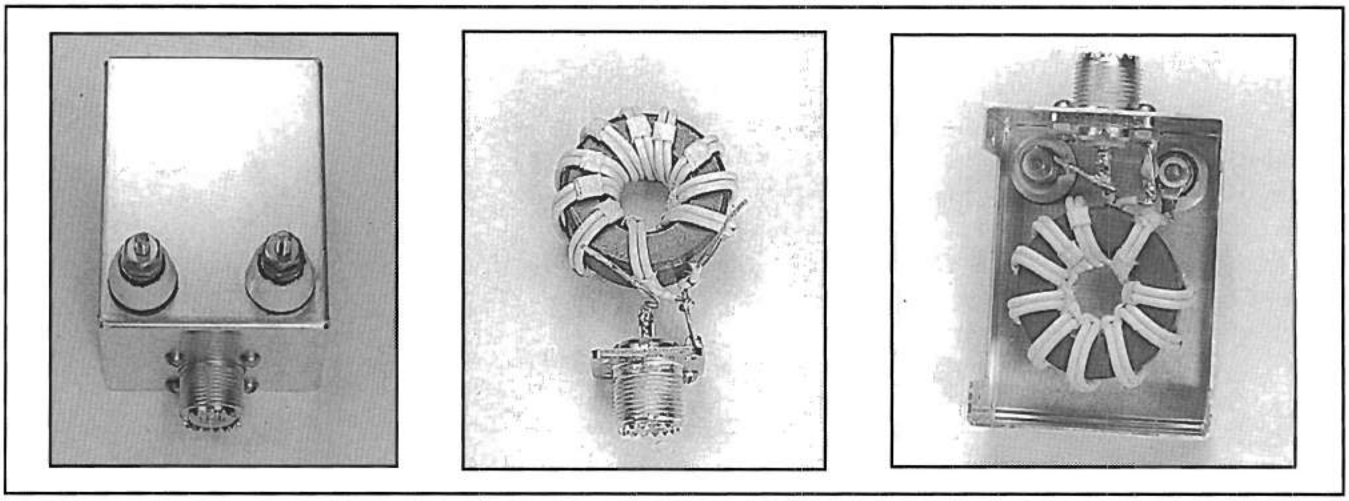

Photo 8-D. *Three views of the low-power 4:1 Ruthroff Balun designed to handle the output of any HF transceiver. The aluminum enclosure is a 2.75 inches high by 2.125 inches wide by 1.625 inches high CU 3000-A minibox.*

to well beyond 30 MHz. Both can easily handle the full legal limit of amateur radio power.

The single-core version (on the left) has 8 bifilar turns on each of its two transmission lines. The dual-core version (on the right) has 16 turns on each core. The wires are clamped together with strips of Scotch No. 27 glass tape placed about every 3/4 inch. The cores are 2.4-inch OD ferrite toroids with a permeability of 250. The connectors are on the low-impedance sides. For ease of connection, the dual-core version has one winding clockwise and the other, counterclockwise. Also, in the dual-core case, the spacing between the two cores (which isn't critical) can be as small as 1/4 inch.

These transformers can also be wound with ordinary No. 14 (solid) house wire. The several samples I tried yielded characteristic impedances close to 100 ohms (and, thus, were acceptable). The major difference lies in the voltage-breakdown capability. Units wound with Teflon-sleeved No. 14 H Thermaleze wire have been reported to withstand 10,000 volts without breakdown! Obviously, this is beyond the capability of ordinary house-wire.

Photo 8-F shows a Guanella (current) 4:1 Balun mounted in a 5 inch long by 4 inch wide by 3 inch high CU 3005-A minibox. It has 14 bifilar turns of No. 14 H Thermaleze wire on each of the two 2.4-inch OD ferrite toroids with permeabilities of 250. Each wire is covered with Teflon tubing, resulting in a characteristic impedance close to 100 ohms (the optimum). The windings also employ a crossover after the seventh turn, as shown in **Figure 8-6**. For ease of con-

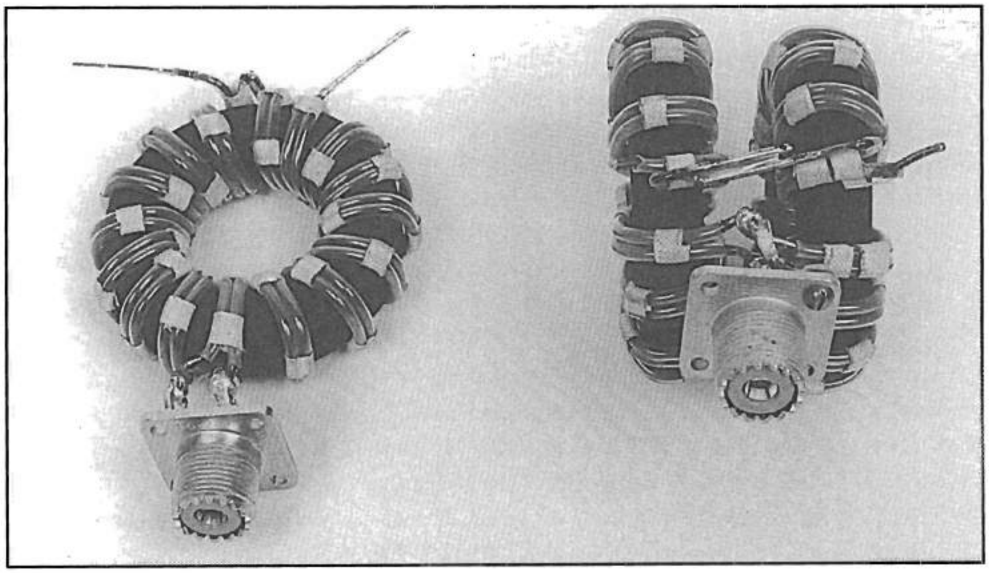

Photo 8-E. *Two high-power versions of the Guanella 4:1 Balun. The Balun on the left uses a single core while the one on the right uses two cores. The connectors are on the low-impedance sides.*

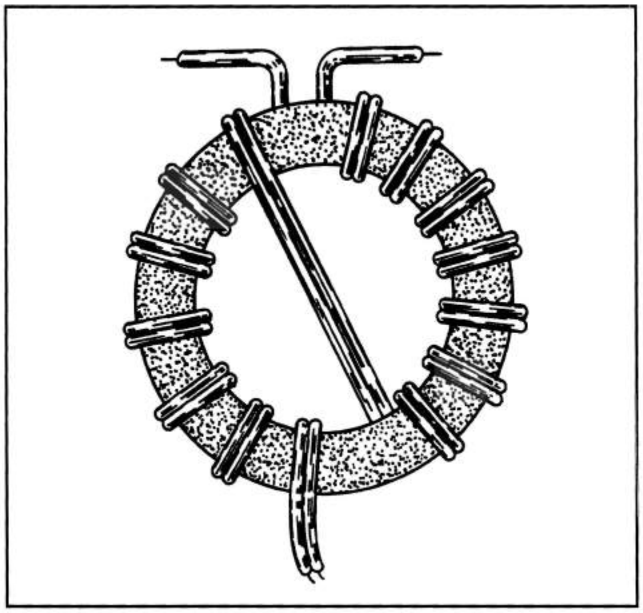

Figure 8-6. *Construction of a crossover placing input and output connections on opposite sides of the toroid.*

nection, one toroid is wound clockwise and the other, counterclockwise. The spacing between the toroids can be between 1/4 and 1/2 inch.

When matching 50-ohm cable to a balanced load of 200 ohms, the transformation ratio is constant (within 2 percent) from 1.5 to 45 MHz. This Balun can also handle the legal limit of amateur radio power. It would probably perform satisfactorily if wound with ordinary No. 14 house wire (solid), or with Teflon-covered No. 14 tinned wire. However, the design in **Photo 8-F** has withstood peak pulses of 10,000 volts! Considering this fact, it might be worthwhile to take the extra effort and use Teflon-covered No. 14 H Thermaleze wire.

A single-core version using two coiled transmission lines on a single core looks interesting and should be investigated. It results in balanced currents and unbalanced voltages. I would use two coiled transmission lines with 7 bifilar turns of the same wire on the same core as above.

Photo 8-G shows two low-power Guanella 4:1 Baluns designed to match 50-ohm coaxial cable to loads of 200 ohms. They both use No. 20 hook-up wire (solid) giving a characteristic impedance very close to the objective of 100 ohms. Their responses are flat from 1.5 MHz to well beyond 50 MHz. They are conservatively rated at 150 watts of continuous power and 300 watts of peak power. They have been exposed to 500 watts of continuous power (in a matched condition) for a considerable length of time with virtually no rise in temperature.

The single-core version (on the left) has 7 bifilar turns on each of its two transmission lines, while the dual-core version (on the right) has 14 turns on each core. The wires are clamped together about every 1/2 inch with strips of Scotch No. 27 glass tape. The cores are 1.25-inch OD ferrite toroids with a permeability of 250. The connectors are on the low-impedance sides. As above, the dual-core version has one winding clockwise and the other, counterclockwise.

Photo 8-H is a step-down version of the Guanella 4:1 Balun and uses two ferrite rod-cores 3/8 inch in diameter and 3.5 inches in length. Core permeabilities are 125. It uses the schematic of **Figure 8-1A,** but with the generator (which is grounded) on the right side and the load (ungrounded) on the left side. This 4:1 Balun is designed to match 50-ohm coaxial cable (on the right side) to a balanced load of 12.5 ohms. Each rod has 13.5 bifilar turns of No. 14 H Thermaleze wire. Again, for ease of connection, one rod is wound clockwise and the other, counterclockwise. The response is flat from 1.5 MHz to well over 30 MHz. This Balun is fully capable of handling the legal limit of amateur radio power. The connector is on the high-impedance (50 ohms) side. Beaded versions of Guanella's step-down 4:1 Balun also look very promising for operations on the VHF and UHF bands. The technique requires minimizing the parasitics in the interconnections. Some examples will be shown later.

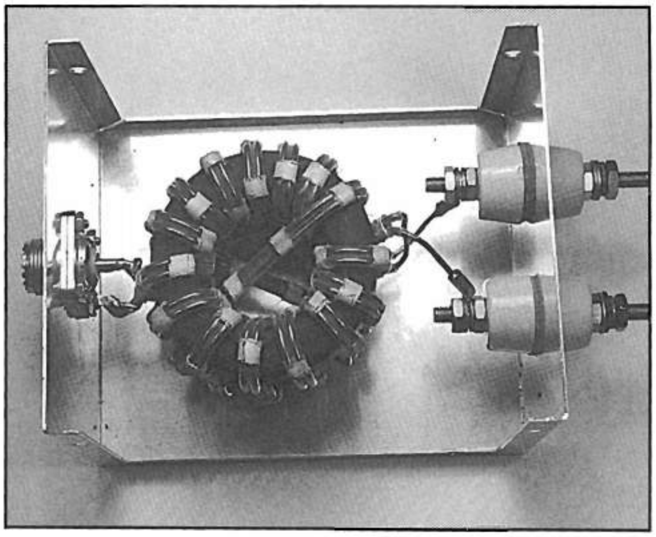

Photo 8-F. *A 4:1 Guanella Balun using the crossover connection of Figure 8-6 and mounted in a 5 inch long by 4 inch wide by 3 inch high CU 3005-A minibox.*

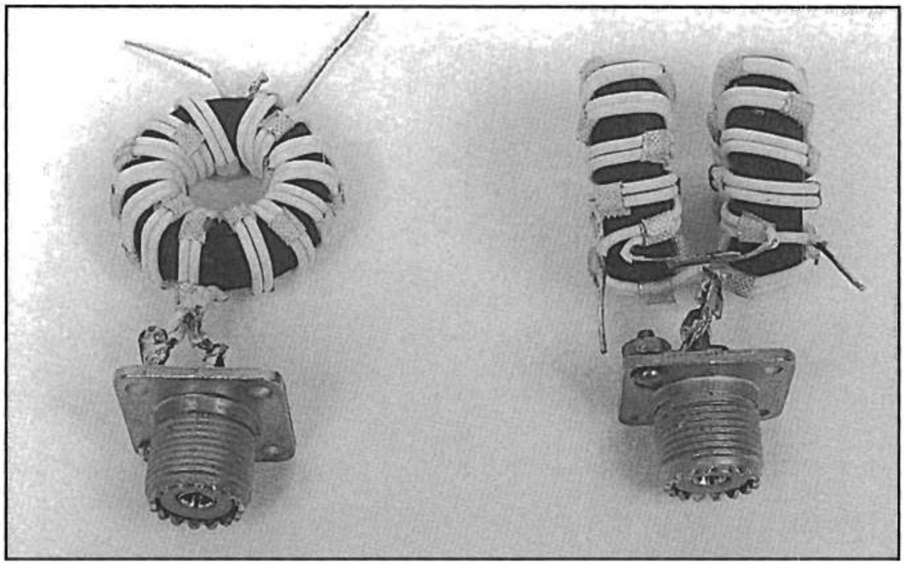

Photo 8-G. *Two low-power versions of the Guanella 4:1 Balun. The Balun on the left uses a single core while the one on the right uses two cores. The connectors are on the low-impedance sides.*

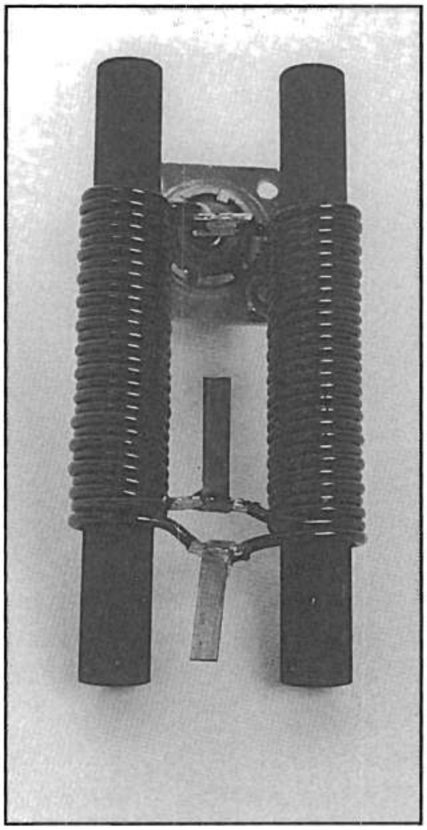

Photo 8-H. *A dual rod-core 4:1 Guanella step-down Balun designed to match 50-ohm cable to a balanced load of 12.5 ohms. The connector is on the 50-ohm unbalanced side.* ⟶

It should be mentioned again that the three dual-core Baluns above also make excellent broadband Ununs. They only sacrifice a little in low-frequency response. However, because of their conservative designs, they can still handle the 160-meter band.

Sec 8.5 Comparisons with Other Baluns

After completing the study on 4:1 Baluns, I thought it would be interesting to characterize other Baluns that are commercially available, or that have been recently described in the amateur radio literature. My findings are as follows.

The 4:1 Rod-type Ruthroff Balun

Photo 8-I shows a photograph of the typical rod-type 4:1 Balun, which was practically the only one available for the past three to four decades. The Balun in the photograph is the **HI-Q Balun.** It is the Ruthroff design (now called a *voltage* Balun[6]) with 10 bifilar turns of No. 14 wire on a 1/2-inch diameter ferrite rod 2 inches in length. In terminating this Balun with 200 ohms, the useful range was found to be from 7 to 15 MHz. Below 7 MHz, the input impedance showed a considerable inductive component—indicating autotransformer action and flux in the core (which could be harmful). Above 15 MHz, the transformation ratio increased and became complex. The optimum impedance level was found when matching 100 ohms to 25 ohms (indicating a characteristic impedance of the windings of only 50 ohms). The useful frequency range at this impedance level increased to 3.5 MHz to 30 MHz.

When matching 50-ohm coaxial cable to a 20-meter folded dipole at a height of 0.17 wavelengths (resulting in a resonant input impedance of 200 ohms), the VSWR curve was indistinguishable from that of the best Guanella 4:1 (current) Baluns.[6] This Balun also presented no difficulty in handling the full power limit. However, on 10 meters, the difference due to a very low characteristic impedance of the coiled transmission line became evident. **Figure 8-7** shows the poor VSWR curve of the rod-type Balun when compared to other Guanella and Ruthroff Baluns with characteristic impedances close to the optimum value of 100 ohms. The rod-type Balun with all of its inadequacies, which include voltage-breakdown, is certainly *not* recommended.

The high-power Ruthroff Balun is close to McCoy's design.[17] It uses 11 bifilar turns of No. 14 H Thermaleze wire on a stack of three T200-2 cores. The wires

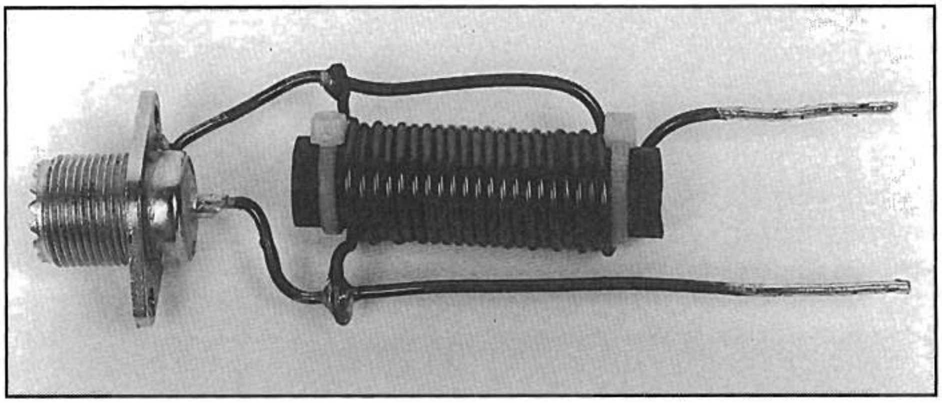

Photo 8-I. *A typical 4:1 rod-type Ruthroff (voltage) Balun (HI-Q).*

are also covered with 20-mil wall Teflon tubing yielding a characteristic impedance close to 100 ohms. I differ with McCoy's design because I feel the characteristic impedance of his transmission line could be closer to 50 ohms. His Balun and others that use powdered-iron cores will be described further in the next chapter on Baluns for antenna tuners.

The low-power units in **Figure 8-7** used 14 bifilar turns of No. 18 hook-up wire on 1.25-inch OD ferrite toroids with a permeability of 250. The transmission lines of the low-power Baluns were just under 20 inches, while those of the high-power Balun were 50 inches in length.

As can be seen in **Figure 8-7**, there are **very** small differences between the VSWR curves of the two low-power Baluns and the high-power Ruthroff (voltage) Balun. The differences could very well be attributed to the small variations in the characteristic impedances of the windings. Very likely, the most important information gleaned from **Figure 8-7** is that, when feeding a balanced dipole with a virtual ground-plain bisecting it, the Ruthroff Balun takes on the character of a two-core Guanella Balun. In other words, the Ruthroff Balun loses its built-in high-frequency cut-off!

4:1 Current Baluns

I also characterized several so-called *current* Baluns[6] that recently appeared on the market. These are my findings:

a) They are the dual-core (toroids) version of the Guanella Balun, which sums voltages of equal delays.

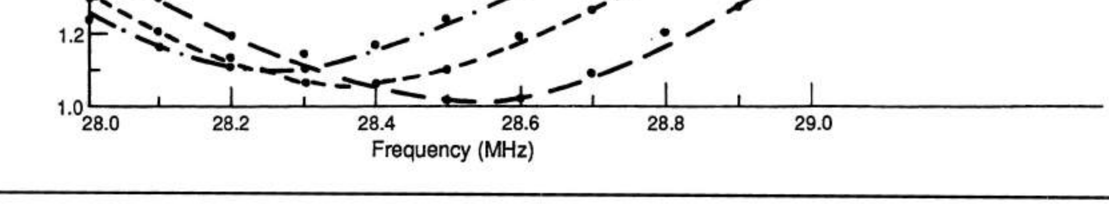

Figure 8-7. *Plots of VSWR curves on 10 meters for four different 4:1 Baluns. The high-power Ruthroff Balun is actually McCoy's design with a 100-ohm characteristic impedance winding (and is described in the next chapter). The Hi-Q Balun is shown in Photo 8-I. The comparisons show the importance of having the optimum value of the characteristic impedance of the windings and that the high-power Balun with its much longer transmission line indicates a Guanella-type operation.*

b) The electrical performances of these Baluns are vastly superior to the rod-type Balun described earlier.

c) These Baluns should meet their electrical and power-rating specifications.

d) My only criticism is that they could have more of a safety margin at the low-frequency end, where excessive core flux (due to higher than expected impedances) could take place. More inductance in the windings is recommended.

The Beaded-coax 4:1 Balun

A design in an amateur radio journal[5] advocated using beaded coaxial cable (of 100 ohms) in a 4:1 Guanella design. Various claims were advanced for this approach. I constructed one of these Baluns using No.14 wire with Teflon sleeving, resulting in the required 100-ohm characteristic impedance. Here are my findings:

a) The Balun had excellent margins at both the high- and low-frequency ends. The performance of this Balun verified the analysis (expressed earlier) with the high- and low-frequency models and the subsequent voltage gradients. In fact, the high-frequency performance exceeded the capability of my simple test equipment.[2]

b) The major disadvantage is in efficiency. Because high-permeability (2500) beads are required in order to obtain the required choking reactance in the HF band, this Balun had considerably more loss than coiled-type Baluns using low-permeability (less than 300) ferrite toroids.[2] A soak-test[2] (transformers connected back-to-back and about 500 watts applied into a dummy load) with the dual-core low-power unit in **Photo 8-G** showed that the smaller Balun ran considerably cooler! The beaded transmission line technique is recommended mainly for Baluns (and Ununs) operating at low impedance levels or on the higher-frequency bands.

Sec 8.6 Summary

Unlike the 1:1 Balun, the 4:1 Balun matching 50 ohms unbalanced to 200 ohms balanced has had no real standard for comparison in the amateur radio literature. As **Chapter 7** showed, Reisert's Balun in his 1978 article[7] had all of the attributes of a good 1:1 design. Therefore, he set a legitimate standard for others to follow or even attempt to exceed. Also shown were some of my variations in his design for increased efficiency and ease of construction.

However, **Chapter 8** has illustrated that the designs in the amateur literature (particularly the handbooks) are found lacking in bandwidth, or efficiency, or both. Even the 4:1 Baluns on the commercial market can be improved. This is *especially* true of the rod-type Balun that has been available for decades!

In the process of investigating the 4:1 Balun for my series of articles in *CQ* and *Communications Quarterly*, I have arrived at some designs that could provide the beginnings of standards for this device. They are included in this chapter and are:

1. For *balanced* applications like matching 50-ohm cable to the 200-ohm balanced input impedance of folded-dipole or log-periodic antennas, I recommend the single-core, Ruthroff design of **Photo 8-A**. It is capable of handling the full legal limit of amateur radio power. For a higher-power capability and a little less bandwidth, I recommend the Ruthroff design shown in **Photo 8-C**. These designs are presently called *voltage* Baluns.[6]

2. For unbalanced applications like the OCFD (off-center-fed dipole), or a dipole that could be unbalanced by surrounding structures or by construction errors, I recommend the two-core Guanella design of **Photo 8-E** or **Photo 8-F**. It is a much more flexible unit that can operate successfully as a Balun when the load is grounded at its center (**Figure 8-1A**), or as an Unun when the load is grounded at the bottom. Although not shown, it can even be grounded at the top, yielding a 4:1 phase-inverter.

3. For low-power 4:1 Baluns, there really have been no designs in the literature for comparisons. Therefore, by default, the designs in this chapter are suggested as standards. Applications of the single-core Ruthroff Balun and the two-core Guanella Balun are the same as their higher-power counterparts.

I am sure there are some who don't agree with the recommendations proposed above. The Guanella (current) Balun appears to be the main Balun of choice. The question is: Why use a two-core Guanella (current) Balun when a single-core Ruthroff (voltage) Balun will do? They both have the same power ratings! For those who disagree with my views, designs, or recommendations, I encourage them to (as the classic TV commercial used to say) *put it in writing*. Then we will all benefit from the new information.

Chapter 9

Baluns for Antenna Tuners

Sec 9.1 Introduction

The 4:1 Balun, matching 50 ohms (unbalanced) to 200 ohms (balanced), has found its most popular use in antenna tuners. Because the Balun rarely sees a resistive load of 200 ohms in this application, the primary objective is to take the balanced impedance (with respect to ground) of the input to an open-wire (or twin-lead) feedline and transform it into an unbalanced impedance which has one side grounded and can be transformed into 50 ohms by an L-C matching network. This was well described in two *CQ* articles by McCoy.[17,18]

As **Chapter 8** has shown, there are two different forms of the 4:1 Balun. One uses two transmission lines wound on separate cores (or threaded through ferrite beads in some cases) and connected in parallel at the 50-ohm side and in series at the 200-ohm side. This design was first presented by Guanella in 1944,[3] and is presently called a *current* Balun.[6] The other design, using a single transmission line wound around a core and connected in a phase-inverter configuration, was introduced by Ruthroff in 1959.[9] This design, which has recently been called a *voltage* Balun,[6] is now perceived as the inferior design.

This chapter presents another view of the 4:1 Balun. It not only includes the optimum design considerations for antenna tuner use, but also the design parameters for multiband antenna systems using center-fed dipoles with open-wire or twin-lead feedlines. Also included are my views on the G5RV antenna, the tuner using a 1:1 Balun before the L-C matching network, and the special cases of Ruthroff's 4:1 design matching into a load that is actually or virtually grounded at its center.

Sec 9.2 The Two Forms of Antenna Tuners

Three of the more common items in amateur radio jargon are VSWR, antenna tuners, and multiband antennas (especially the G5RV). These have appeared upon the scene because of the ease at which bands can now be changed and the narrow limits in the range of matching impedances with modern rigs.

The concept of using a wire antenna on many different bands isn't new. Designs have been around for more than six decades. In fact, satisfactory circuits have also been available which couple transmitters to balanced lines that present loads different than the transmitter output impedance. These were known as series and parallel-tuned circuits.[19] The transforming of a balanced impedance to an unbalanced impedance was accomplished by the isolation provided by magnetic coupling. Energy was transmitted from one circuit to the other by either having two coils in close proximity or by "link" coupling. However, these methods of coupling have fallen by the wayside, together with rock-bound rigs, plug-in coils, separate

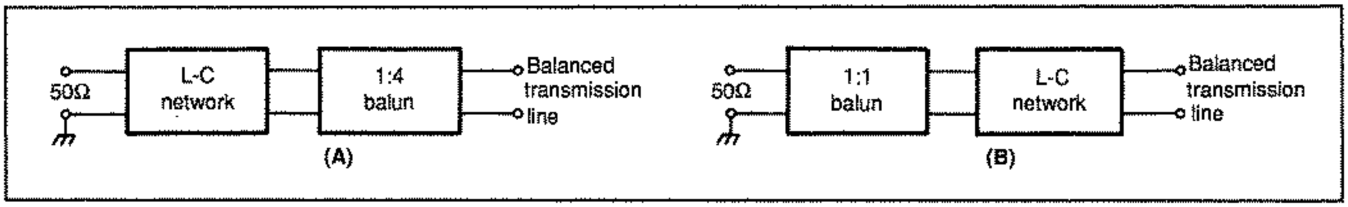

Figure 9-1. *The two basic forms of the Transmatch (antenna tuner): (A) the more popular design using an unbalanced L-C network and a 1:4 Balun; (B) a 1:1 Balun and a balanced L-C network.*

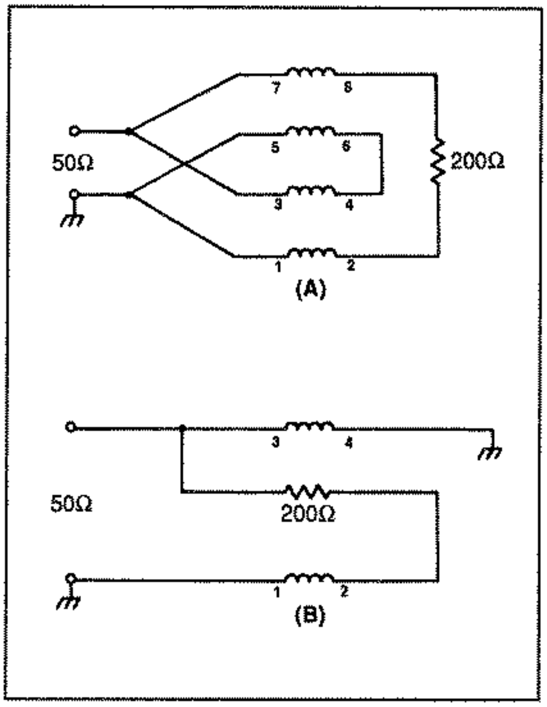

Figure 9-2. *The two basic forms of the 1:4 Balun: (A) the Guanella (current) Balun; (B) the Ruthroff (voltage) Balun.*

Figure 9-3. *The frequency response of a 4:1 Ruthroff (voltage) Balun with the load floating and with its center grounded. In the grounded case, the high-frequency response is similar to that of a 4:1 Guanella (current) Balun.*

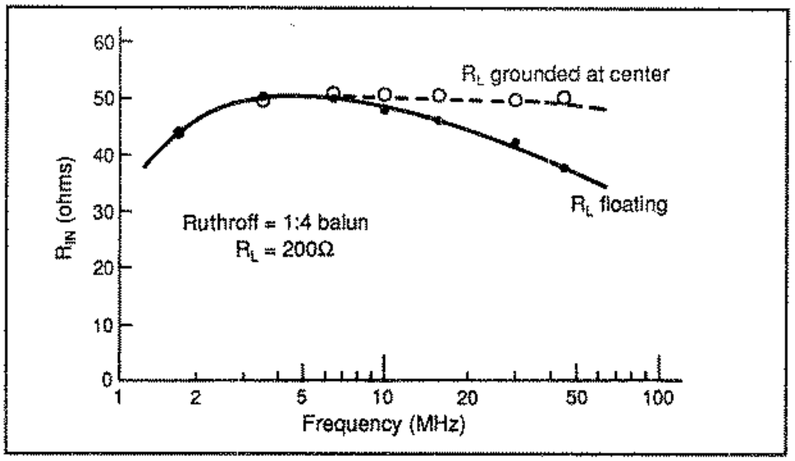

receivers and transmitters, and (sad to say) the exciting flashing of mercury-vapor rectifiers.

Today, the *transmatch* is most often used to convert the reactive/resistive load presented by an antenna system to a nonreactive, grounded 50-ohm load. The transmatch is also commonly known as an antenna tuner. The isolation role, that of converting a balanced impedance to an unbalanced one, is now provided by the Balun transformer.

There are two basic forms of the transmatch, and they are shown in **Figure 9-1**. **Figure 9-1A**, which shows a 4:1 Balun between the L-C network and the balanced transmission line, has been the most popular. In some designs, a 1:1 Balun has been used. In either case, this form of antenna tuner places the burden on the Balun, not on the L-C network. Depending upon the dimensions of the antenna and open-wire (or twin-lead) transmission line, the Balun can see very high impedances that may be harmful. In turn, the L-C networks are simple because of their unbalanced nature. Among the most popular networks are L, pi, T, Ultimate, and SPC types.[20–23] An added advantage to this approach is that the Balun can be placed outside the operating area and connected to the L-C network by a coaxial cable.[17,18]

On the other hand, **Figure 9-1B** takes the complexity out of the Balun and places it on the L-C network. With a balanced network, the 1:1 Balun should see a lower voltage drop along the length of its transmission line (and, hence, less loss[2]) because its load is always close to 50 ohms. Additionally, the choking requirements of a 1:1 Balun are considerably less than that of the 4:1 Balun in **Figure 9-1A**.

Roehm has addressed the problems related to this form of transmatch in an article.[16] In fact, he suggests a design using an unbalanced T network and a 1:1 beaded-coax Balun. Although a balanced L-C network is inherently more complex and costly, it would be interesting to see its comparison with Roehm's unbalanced design. Additionally, a comparison with a 1:1 Balun using 50-ohm twin-lead or coaxial cable wound around a ferrite toroid with a permeability of less than 300 would also be useful. Because the 1:1 beaded-coax (choke-type) Balun requires ferrite beads with permeabilities considerably greater than 300, it has more loss.[2]

In any event, the basic form of the transmatch using the design in **Figure 9-1B** looks promising and merits further investigation.

Sec 9.3 Another View of the 4:1 Balun

This section presents my views and the results of my work on a 4:1 Balun designed for use in the very popular "antenna tuner" shown in **Figure 9-1A**. Because the Balun may be exposed to harmful high voltage conditions in this application, the efficiency and

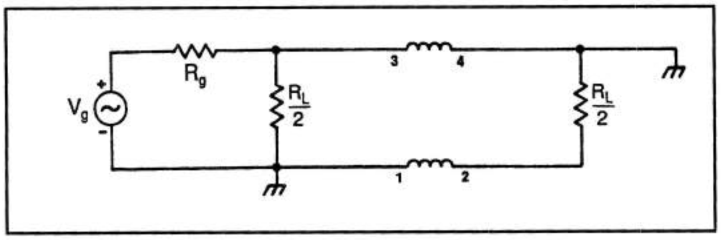

Figure 9-4. *Suggested model of the 4:1 Ruthroff (voltage) Balun when the load, R_L, is grounded at its center.*

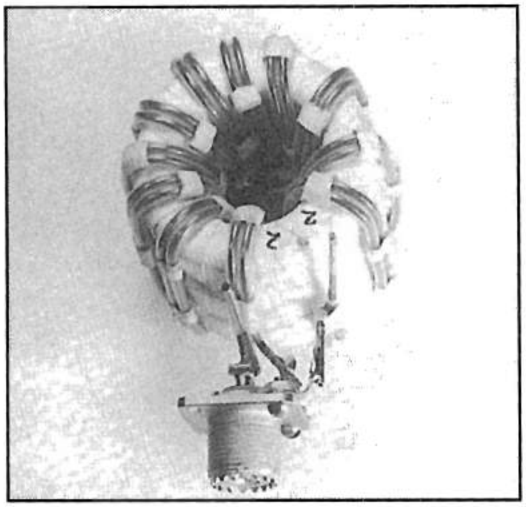

Photo 9-A. *The high-power Ruthroff 4:1 Balun used in the comparison with other Baluns (see Figure 9-7, and Figure 8-7 of Chapter 8). Except for some difference in the characteristic impedance of the bifilar winding, it is essentially the McCoy 4:1 Balun.*

ruggedness of the core materials are important considerations. Experiments have shown that losses in Baluns are related to the impedance level[2] (and, hence, voltage level), and the permeability of the materials. Therefore, the losses are of a dielectric-type and not of a current-type, as in conventional transformers. Moreover, it is well known that powdered-iron is a more rugged and linear material than ferrite. A very important question to ask is which form of the 4:1 Balun should be used in this application—Guanella's or Ruthroff's? My conclusions may surprise many readers.

As was mentioned at the beginning of this chapter and above, there are two basic forms of the 4:1 Balun. They are shown in **Figure 9-2A** and **B**. **Figure 9-2A** is Guanella's approach. It uses two coiled transmission lines (on separate cores) connected in parallel on the 50-ohm side and in series on the 200-ohm side. This has recently been called a *current* Balun.[6] In order to have "flat" transmission lines and obtain the highest frequency response, the characteristic impedance of the coiled transmission lines should be equal to the loads they see—namely, $1/2R_L$ and, in this case, 100 ohms.

As Guanella said in his classic paper,[3] this Balun is literally "frequency independent." At the low frequency end, the reactance of the coiled transmission line should be much greater than 100 ohms (in this case) in order to assure that the energy is transmitted from input to output by an efficient transmission line mode. Beaded transmission lines aren't recommended for use on the HF band at these impedance levels because of excessive dielectric loss.

Figure 9-2B shows Ruthroff's approach, which uses a single transmission line connected in the phase-inverter configuration (see **Chapter 7**). By grounding terminal 4, a voltage drop of $-V_1$ appears across the length of the transmission line. As a result, terminal 3 is at $+V_1$ and terminal 2 is at $-V_1$—a 4:1 transformation ratio.

It is when the load is actually grounded at its center that we see a very interesting feature of this approach. A simple impedance measurement will show that the high frequency response is *vastly* improved and that the Ruthroff (voltage) Balun appears to take on the character of a Guanella (current) Balun! **Figure 9-3** illustrates the measurements of the input impedance of a Ruthroff 4:1 Balun with the load floating, and when it is grounded at its center. These measurements (with a simple resistive bridge) were made on my design using a powdered-iron core, which will be described later.

A model for the Ruthroff 4:1 Balun, when the load is grounded at its center, is provided in **Figure 9-4**. If the characteristic impedance of the transmission line in the 4:1 Balun is 100 ohms (the optimum value if the load is 200 ohms), then the generator sees $R_L/2$ in parallel with a $R_L/2$ from a "flat" line. As a result, the generator sees its match of 50 ohms—even though the currents in the loads are not in phase!

However, when the Balun is connected to a center-fed, folded dipole, or log-periodic beam antenna with 200-ohm input impedances, the virtual ground-plane bisecting the antennas presents an interesting case. Because there is no metallic connection to the center of the load, the currents in both halves of the antennas are in phase. This is unlike the situation in **Figure 9-4**, where the current in the load on the right is delayed compared to the current in the load on the left. As is evident in **Figure 8-7** of **Chapter 8**, the VSWR curve

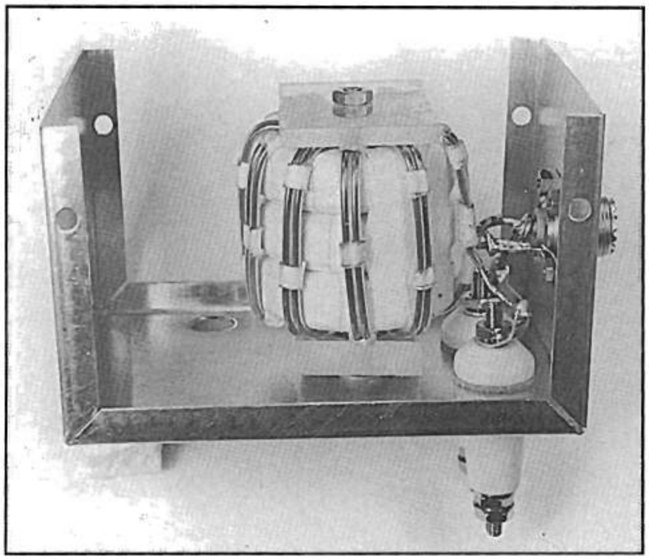

Photo 9-B. *The Balun of Photo 9-A mounted in a 5 inch long by 4 inch wide by 3 inch high minibox.*

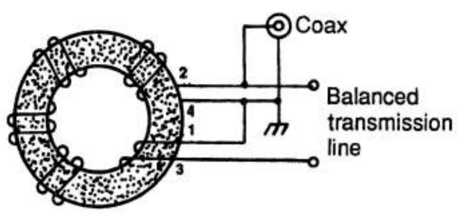

Figure 9-5. *A pictorial of the connections for a 4:1 Ruthroff (voltage) Balun.*

for the high-power Ruthroff Balun has practically the same shape as those of the two lower-power units—even though its transmission line is more than two and one half times longer.

Because the high-frequency response of a Ruthroff 4:1 Balun with a floating load is highly dependent upon the length of the transmission line, **Figure 8-7** suggests that, in the virtual ground case, the Balun acts as a Guanella Balun which sums voltages of equal phases. It also suggests that the effective electrical length of the transmission line is one-half of its actual length. I am quite sure that this model of Ruthroff's Balun was not proposed by him or by others that followed. However, it should be remembered that this condition only exists in the *balanced* case. With an unbalanced load, the Balun should introduce a reactive component that will limit the high frequency response. In the unbalanced case, the Guanella Balun with two cores is the Balun of choice.

Sec 9.4 Some "Hardy" 4:1 Designs

Photos 9-A and **9-B** show two views of a Ruthroff 4:1 Balun, much like McCoy's design, which has been used in his highly popular transmatch.[22] **Figure 9-5** shows a pictorial representation of the connections. It has 11 bifilar turns of No. 14 H Thermaleze wire on three (stacked) T200-2 cores. The cores are powdered-iron material with a permeability of 10. The OD

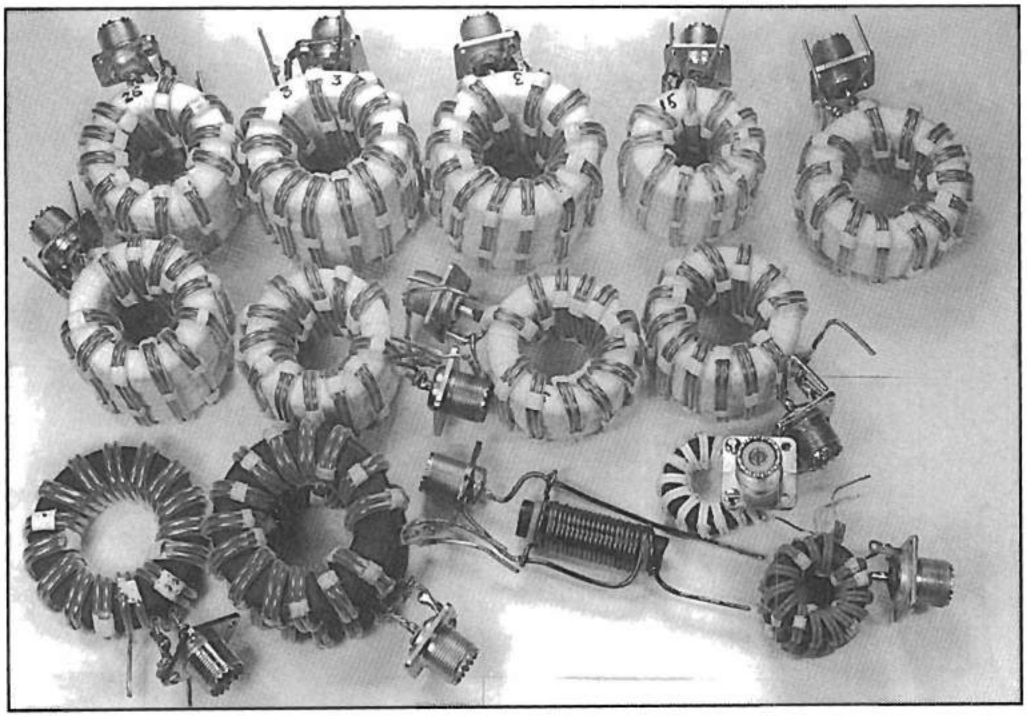

Photo 9-C. *The various Baluns used in the study of 4:1 Baluns for antenna tuners.*

is 2 inches. The wires are also covered with a 15-mil wall Teflon tubing, yielding a characteristic impedance close to the optimum value of 100 ohms.

I differ with McCoy's design here, because the characteristic impedance of his Balun could be closer to 50 ohms. **Photo 9-B** shows the Balun mounted in a 5 inch long by 4 inch wide by 3 inch high minibox. Because McCoy didn't use a thickly insulated wire, he wound a layer of Scotch No. 27 glass tape on each toroid before stacking. This was followed by another layer in the stacking process. With Teflon sleeving over the wire, the extra insulation provided by the glass tape could be dispensed with.

In order to improve the low-frequency response of McCoy's Balun, I made a study of higher permeability powdered-iron cores. **Photo 9-C** shows the various Baluns used in the study. The object of this study was to determine the best core material for a 4:1 Balun to be used in antenna tuners where they can be exposed to high impedances (and, hence, hostile environments). I knew, as the result of very accurate insertion loss measurements[2] that loss with ferrite materials was related to the voltage drop along the length of the transmission line and to the value of the permeability. Permeabilities of 40 (No. 67 ferrite) exhibited the lowest loss. The results taken on a single powdered-iron material—No. 2 material with a permeability of 10—also showed the very same low loss. Because powdered-iron material has been known to be more rugged and linear than ferrite material, this suggested that other powdered-irons with greater permeabilities should also be investigated.

I investigated four other powdered-irons with permeabilities of 20, 25, 35, and 75. Their designations were Nos. 1, 15, 3, and 26, respectively. Comparisons were made on input impedances (with the outputs terminated in 200 ohms) and temperature rises (when handling 500 watts of power). The power test showed, convincingly, that No. 26 material was not to be used because it showed a definite rise in temperature while the other three didn't. However, all four materials showed a definite lower input impedance than the No. 2 material, which has a permeability of 10. As expected, the higher the permeability, the larger the difference with No. 2 material. Although an input impedance measurement does provide some indication of loss because it appears as a shunting path to ground, a very accurate insertion loss measurement would provide a more precise indication of the trade-off that can be made in efficiency for low frequency response.

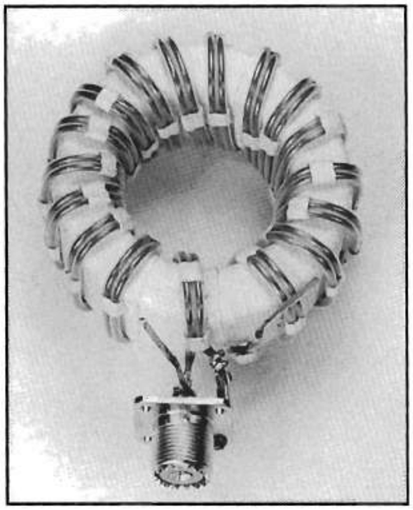

Photo 9-D. *An improved 4:1 Ruthroff design for antenna tuners.*

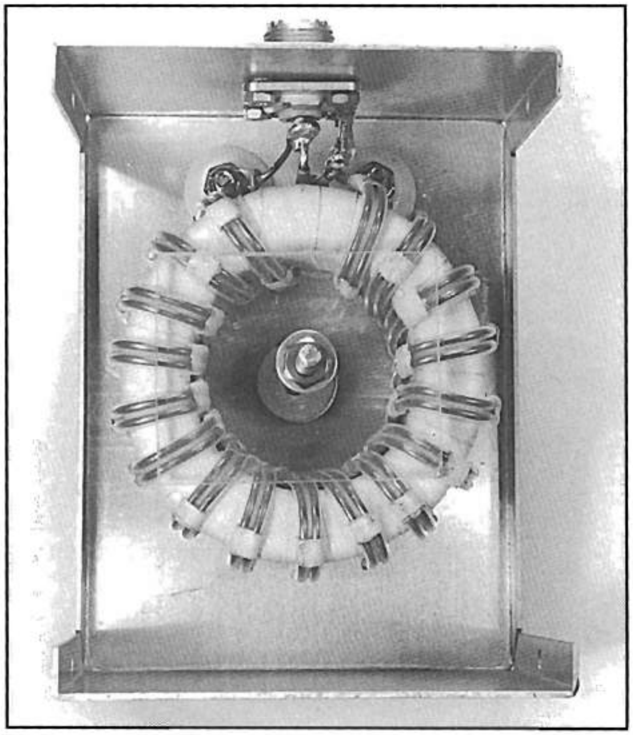

Photo 9-E. *The improved 4:1 Ruthroff Balun mounted in a 5 inch long by 4 inch wide by 3 inch high minibox.*

Because my simple loss measurements indicated that the higher permeability powdered-irons had more loss than the No. 2 material, I decided to design a 4:1 Ruthroff Balun using this material—but with a larger core and more turns than the McCoy[17,18] Balun. Although McCoy's design has enjoyed considerable success over the years, I felt that a larger inductive reactance was desirable in order to assure better performance on the lower frequency bands (particularly 160 meters).

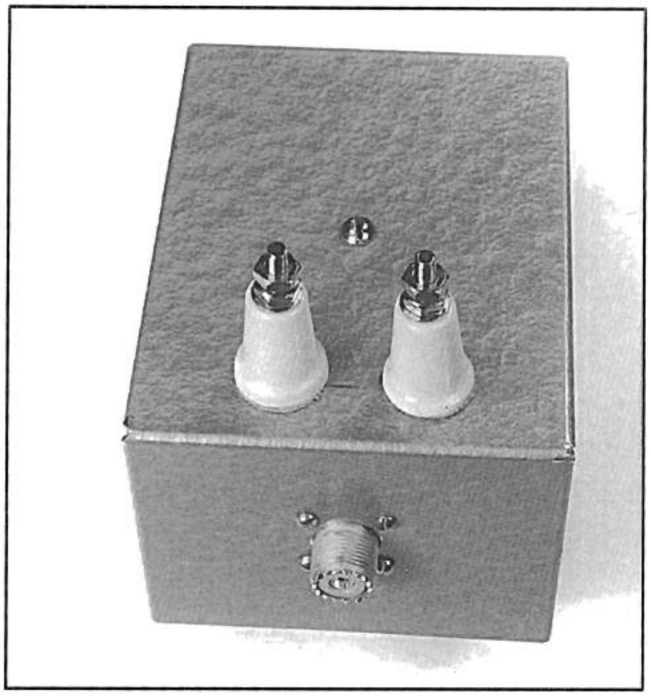

Photo 9-F. *The top view of the 4:1 Balun of Photo 9-E.*

The specific design is shown in **Photo 9-D**. It has 17 bifilar turns of No. 14 H Thermaleze wire on a T300A-2 powdered-iron core, which has an OD of 3 inches and a permeability of 10. With this number of turns and a larger cross section than the three T200-2 cores, the low frequency response improved by a factor of about two over the McCoy Balun, which has 10 to 12 turns on a stack of three T200-2 cores. Furthermore, the wires are also covered with 15-mil wall Teflon tubing, resulting in a characteristic impedance of 100 ohms (the objective). This well-insulated transmission line has been reported to handle 10,000 volts without breakdown. **Figure 9-3** illustrates the performance of this Balun (under a matched condition) when the load is floating and when it is center-tapped-to-ground. **Photo 9-E** shows the Balun mounted in a minibox 5 inches long by 4 inches wide by 3 inches high. **Photo 9-F** shows the top view of the mounted Balun.

Because the Balun with the larger core and more turns showed an improvement by a factor of two in the low-frequency response over the McCoy Balun, I constructed an even larger design. This is shown in **Photo 9-G**. It has 21 turns of the same wire on a T400A-2 core with an OD of 4 inches and a permeability of 10. Even though **Photo 9-G** shows the toroid wrapped with Scotch No. 27 glass tape, as mentioned earlier, this extra insulation isn't required because the wires are covered with Teflon sleeving.

Figure 9-6 shows the comparisons in the input impedances versus frequency of these three "hardy" Baluns when they are terminated in 200-ohm loads grounded at their centers. As can be seen, the low-frequency response of the Balun with the most turns and largest core is the best. Even though the Balun with the 3-inch core has a poorer low-frequency response, it is an improvement over the McCoy Balun and should find considerable use in transmatches. **Figure 9-6** also shows that the high frequency responses of these Baluns with the loads grounded at their centers are remarkably similar. This is especially interesting, as the length of the transmission line on the larger

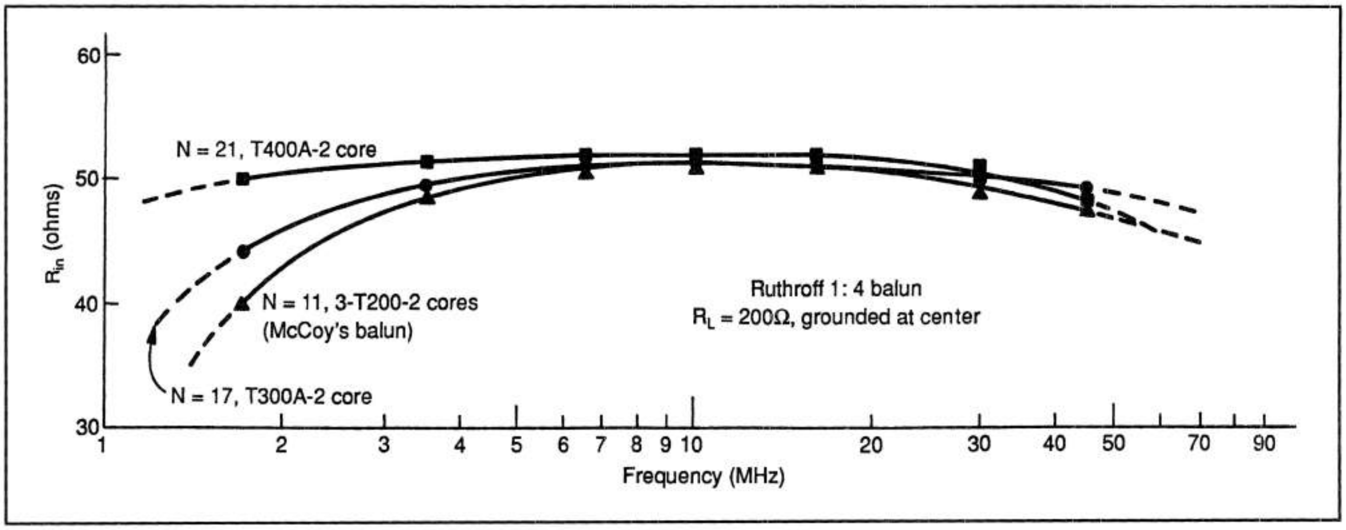

Figure 9-6. *A comparison of the performance of the McCoy 4:1 Balun with Baluns having larger cores and more turns.*

Balun (with the T400A-2 core) is more than twice as long as the other two (115 inches compared to 50 and 55 inches). If it is correct that my model of a Ruthroff (voltage) Balun feeding a balanced antenna or transmission line takes on the character of a Guanella (current) Balun because of the virtual ground, then this large Balun could have many applications.

Sec 9.5 Multiband Dipoles

Antenna tuners have been known to work well for some radio amateurs and not for others. This is due to the differences in the dimensions of their antenna systems. With high impedances seen by the 4:1 Baluns in the antenna tuners, the Baluns not only fail to provide a good balanced-to-unbalanced conversion, but they can also be damaged by excessive heating. The high current, low impedance condition seen by the Balun isn't a problem. Therefore, the object in multiband antenna design is to provide the most favorable impedances for the Baluns, especially on the three lowest frequency bands—40, 80, and 160 meters. Usually on the higher frequency bands, the impedances seen by the Baluns are not as high and the Balun's choking reactances are greater (assuring balanced-to-unbalanced conversion). This section discusses three cases of multiband center-fed dipole designs. They are 1) the "worst case" design, 2) a smaller design—the G5RV, and 3) a larger design. Others may have better designs for multiband operation, but it's clear that the "worst case" design ought to be avoided. **Figure 9-7** shows the symbols for the dimensions of the center-fed dipoles.

Sec 9.5.1 The "Worst Case" Design

Apparently, an 80-meter dipole with a quarter-wave open-wire (or twin-lead) feedline is a logical design for a multiband dipole. In **Figure 9-7** this would mean that $L_1 = L_2 = 59$ to 67 feet, depending upon the favorite operating frequency. If one were to use 450-ohm twin-lead with "open windows," L_2 would be diminished by 10 percent; with 300-ohm TV twin-lead, it would be diminished by about 20 percent. This would make a good antenna system on 160 meters. Because $L_1 + L_2$ is close to a quarter-wave, the current at the input to the feedline is at its highest value and the impedance at its lowest—a very favorable condition for the Balun. In fact, a 1:1 Balun at the feedpoint would probably do a good job.

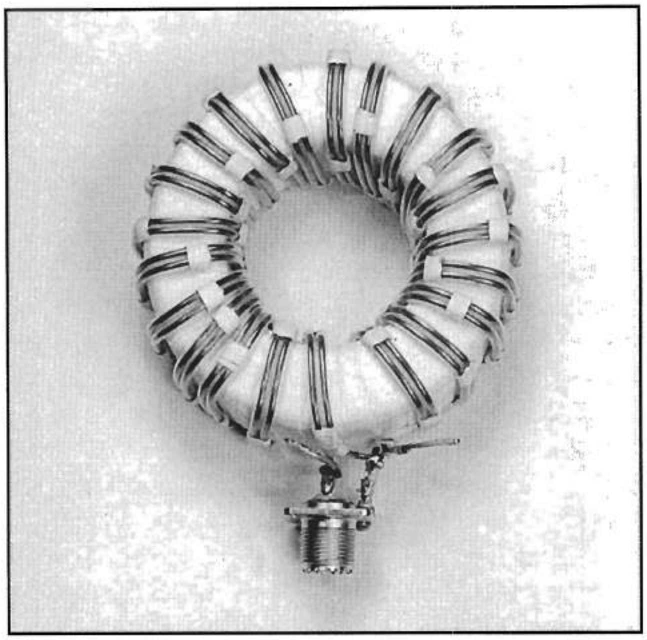

Photo 9-G. *A large 4:1 Ruthroff design using a 4-inch OD powdered-iron core and 21 bifilar turns.*

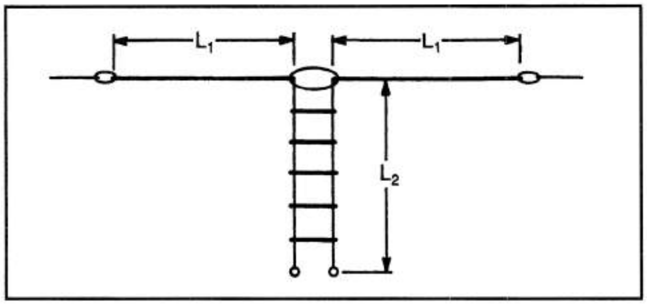

Figure 9-7. *The symbols used for the dimensions of a center-fed dipole with open-wire feeders (or twin-lead with appropriate consideration of velocity factors).*

However, what does the input to the feedline look like on 80 meters? If 450-ohm feedline is used, the 4:1 Balun sees a quarter-wave 450-ohm feedline terminated in approximately 50 ohms. Transmission line theory tells us that the Balun would see 4050 ohms—an impossible condition for most Baluns. The situation becomes worse on 40 meters, where we may have a center-fed, full-wave dipole with a half-wave transmission line (a 1:1 matching transformer). In this case, the Balun could see an impedance approaching 10,000 ohms—a more than impossible condition! Although high impedances would also be seen on 10, 15, and 20 meters, the conditions are not quite as severe because the Balun's choking reactances are usually greater and the impedances lower. In essence, the "worst case" design for a 4:1 Balun is the "best case" design (except for 160 meters), if one still uses inductive coupling to a parallel-tuned circuit.[19]

Sec 9.5.2 A Smaller Design— The G5RV

Varney,[24] G5RV, designed a multiband center-fed antenna system capable of operation on all HF bands from 3.5 to 30 MHz. In contrast to multiband antennas designed as half-wave dipoles on 80 meters (the "worst case" design), the full-size G5RV antenna was designed as a three-half-wave antenna on 14,150 MHz with a 1:1 transmission line matching transformer. It was possible to accomplish this using the dimensions of 51 feet for L_1 and 34 feet for L_2. For 450-ohm twin-lead with "windows," L_2 would be 31 feet; for 300-ohm TV ribbon, L_2 would be 28 feet. Consequently, the input impedance at the base of the matching transmission line was about 100 ohms on 14,150 MHz, and a manageable impedance for 50 or 80-ohm coaxial cable.

However, with a total length for $L_1 + L_2$ of 85 feet, the impedances at the input to the transmission line on 40, 80, and 160 meters are also manageable. Even though they have a reactive component on these bands, they aren't so high that a well-designed 4:1 Balun in an antenna tuner can't handle them easily. Varney also showed that the highest impedances occurred on the 18-, 21-, and 28-MHz bands. However, this doesn't present a problem with the design shown in **Photo 9-D** because it uses an efficient core material, and it also has the highest reactances of its windings at these frequencies.

Varney[24] also wrote, at considerable length, on the unsuitability of a Balun being used to connect the base of the 34-foot open feeders to a coaxial cable feedline. He stated that if a Balun is connected to a reactive load with a VSWR of more than 2:1, its internal losses would increase. Varney also mentioned heating of the wires and saturation of the core. Evidently Varney was not familiar with McCoy's design, which uses a powdered-iron core (with a permeability of 10) that can withstand VSWRs considerably greater than 2:1—without showing any temperature rise. Furthermore, the wire doesn't heat up; however, the core itself does via dielectric heating. Additionally, with sufficient choking reactance, Baluns can handle (equally) the resistive and reactive components of an impedance.

Finally, after observing the voltage and current distributions on all of the bands, it appears that a 2:1 (100:50-ohm) Balun might be an interesting one to try on the G5RV antenna. It could be that many of the bands would not require the added matching of an antenna tuner. If some of the bands require an antenna tuner in order to be used, then I would suggest using a "hardened" Balun. That is what I call McCoy's approach, which uses efficient and hardy powdered-iron cores. A 2:1 Balun, comprised of a 1:2 Unun in series with a 1:1 Guanella (current) Balun, could be easily designed and built (see **Chapter 10**).

Sec 9.5.3 A Larger Design

Even though the G5RV antenna can be made to operate on 160 meters with a suitable antenna tuner, an antenna system larger than the "worst case" design can provide better operation on the 40-, 80-, and 160-meter bands. As you might expect, an antenna system about twice as large as the G5RV offers these advantages. Suggested dimensions are L_1 = 80 feet and L_2 = 100 feet. If the feedline is 450-ohm twin-line with "windows," then L_2 = 90 feet. If it is 300-ohm TV ribbon, then L_2 = 82 feet. As with the G5RV, it's the total length of $L_1 + L_2$ that presents favorable or unfavorable impedances to the 4:1 Balun in the antenna tuner. Therefore, L_1 and L_2 could both be 90 feet, as well. Only very small differences in performance would be noticed between these two systems, particularly on the lower-frequency bands. Obviously, other combinations totaling 180 feet are also possible. In the G5RV case, it's 85 feet.

Sec 9.6 Summary

After reading this chapter, one might think that I have set this technology back a few years by advocating voltage Baluns and powdered-iron cores. I have even questioned the professional literature. However, my conclusions were based upon three experimental results. These were: 1) measurements with my resistive bridge on input impedances of 4:1 Ruthroff (voltage) Baluns with loads floating and center-tapped-to-ground, 2) VSWR measurements on folded dipoles with various 4:1 Baluns (large and small and, therefore, with many different lengths of transmission lines), and 3) McCoy's success with his 4:1 Balun. Also, as this chapter points out, it helps to have the dimensions of a multiband, center-fed dipole, and feeders favor the operation of a 4:1 Balun in antenna tuners. And, *yes*, there is a "worst case" antenna design!

As in many investigations, supplying answers to some questions can lead to others that appear to be important. Specifically, for powdered-irons, how would permeabilities in the 20 to 35 range perform? Simple impedance measurements showing lower values on input impedances, indicate that there is more loss than with a permeability of 10; but accurate insertion loss measurements are needed in order to tell the complete story—the trade-off in efficiency for low-frequency response.

Finally, low permeability ferrite-like No. 67 material with a permeability of 40 looks interesting for use in Baluns for antenna tuners. Accurate insertion loss measurements[2] have also shown the very same high efficiency that was exhibited by powdered-iron having a permeability of 10. With a sufficient number of turns on an appropriate size core, a Balun made of this material could be practical. Even though the amateur radio literature still refers to the problem of core saturation, there has been only one recorded case. This was on 2 MHz with a rod-type 1:1 Balun where insufficient choking reactance exists.[13,14]

Chapter 10

1.5:1 and 2:1 Baluns

Sec 10.1 Introduction

There are many applications for broadband Baluns with impedance transformation ratios close to 1.5:1 and 2:1. Two applications involve matching 50-ohm cable to balanced loads of 75 or 100 ohms, which are the input impedances of a half-wave dipole at heights of 0.22 or 0.34 wavelengths above ground. Another, is the matching of 50-ohm cable to the 100-ohm input impedance of a quad antenna. An interesting, and somewhat unexpected, application is the matching of 50-ohm cable directly to the input impedance of the driven element of a Yagi beam antenna of 33 or 25 ohms. This would eliminate the common hairpin matching network presently used to raise their input impedances to 50 ohms.

There are many versions of these two Baluns. They include: 1) high- and low-power designs, 2) designs matching 50-ohm cable to higher or lower impedances, 3) series- or parallel-type designs, 4) single- or dual-core designs, 5) dual-ratio designs, and 6) HF and VHF designs. The series-type Baluns use an Unun (unbalanced-to-unbalanced transformer) in series with a Guanella (current) Balun. More details on these Ununs are provided in later chapters. In this chapter, you'll read about many high-power designs capable of handling the full legal limit of amateur radio power. They are optimized for sufficient margins in choking reactance at their low frequency ends, and in efficiency throughout their passbands. Two of the 2:1 Baluns are specifically designed for 2-meter operation.

Sec 10.2 1.5:1 Baluns

In this section, I'll present two series-type 1.5:1 Baluns (actually 1.56:1, which should be close enough). They both use 1.56:1 Ununs in series with Guanella 1:1 Baluns. **Figures 10-1A** and **B** show their schematic diagrams. **Figure 10-1B** has an extra input (to a tap), which provides another ratio of 1.33:1.

The left-hand side of **Photo 10-A** shows a design using **Figure 10-1A** mounted in a CU 3006 minibox 5.25 inches long by 3 inches wide by 2.25 inches high (RadioShack has carried a similar enclosure). The 1:1.56 Unun has 4 quintufilar turns on a 1.5-inch OD

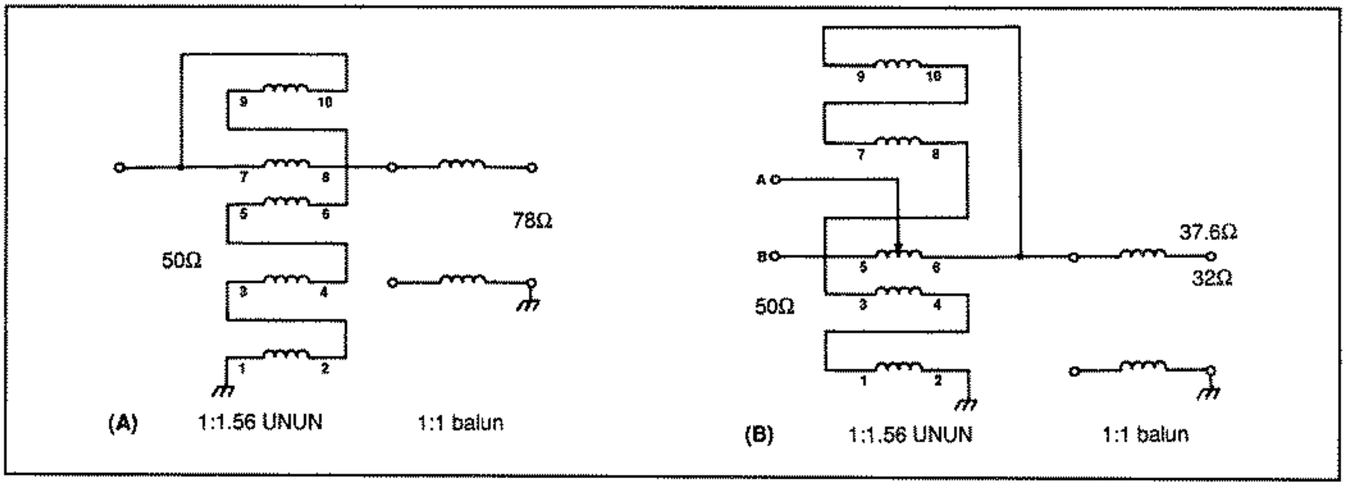

Figure 10-1. *Schematic diagrams of two 1.56:1 Baluns: (A) step-up, 50:78 ohms; (B) step-down, 50:37.6 ohms—connection A, 50:32 ohms—connection B.*

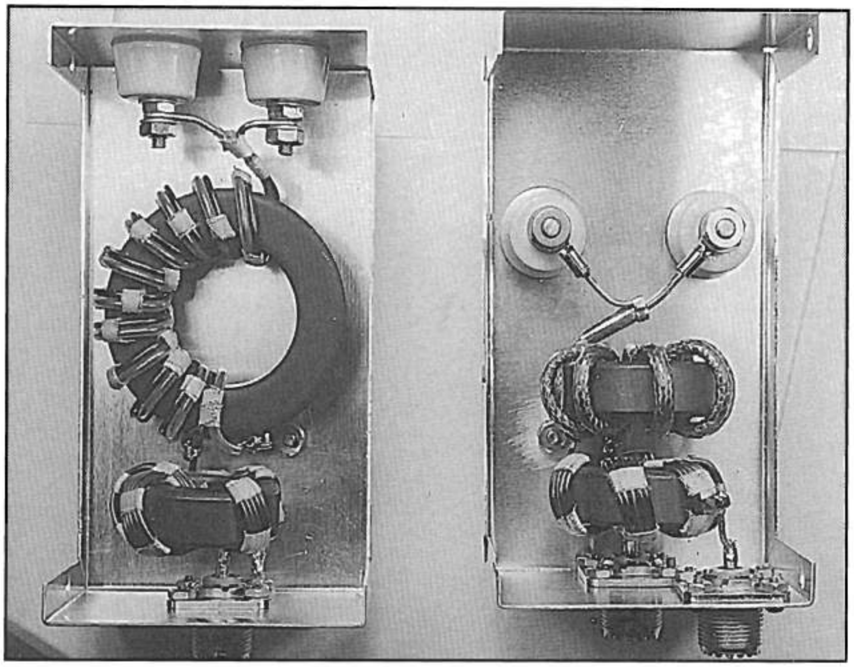

Photo 10-A. *Baluns using the schematic diagrams of Figure 10-1. Balun on the left matches 50-ohm cable to a balanced load of 78 ohms. Balun on the right matches 50-ohm cable to balanced loads of 37.6 or 32 ohms.*

ferrite toroid with a permeability of 250. Winding 7-8 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire.

The 1:1 Guanella Balun has 11 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid with a permeability of 250. One wire is covered with Teflon tubing, resulting in a characteristic impedance very close to 78 ohms (the optimum value).

When matching 50-ohm cable to a balanced load of 78 ohms, the impedance transformation ratio is literally flat (within a percent or two) from 1.5 MHz to 40 MHz! You might be interested to know that (separately) the 1:1 (75:75-ohm) Balun would make an excellent isolation transformer for 75-ohm hardline, and the 1.56:1 (78:50-ohm) Unun an excellent match between 75-ohm hardline and 50-ohm cable.

The right-hand side of **Photo 10-A** shows a design using **Figure 10-1B** mounted in a similar enclosure. The 1.56:1 Unun has 5 quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and is tapped at one turn from terminal 5. The other four wires are No. 16 H Thermaleze.

The 1:1 Guanella Balun has 7 turns of homemade coaxial cable on a 1.5-inch OD ferrite toroid with a permeability of 250. The inner conductor is No. 14 H Thermaleze wire and is covered with Teflon tubing. The outer braid, which is from a small coaxial cable or from 1/8-inch tubular braid, is also tightly wrapped with Scotch No. 92 tape to preserve the low characteristic impedance.

In matching 50-ohm cable to a balanced load of 37.6 ohms (connection A), or to a balanced load of 32 ohms (connection B), the response is essentially flat (within a percent or two) from 1.5 to 30 MHz.

Sec 10.3 2:1 Baluns

The 2:1 Balun lends itself to more choices in design than the 1.56:1 Balun. This is especially true because the parallel-type design, which provides a 2.25:1 Balun with the widest possible bandwidth, can easily be employed. The 1.56:1 Balun is at a disadvantage here. This section presents many Baluns using both series and parallel-type designs.

Sec 10.3.1 Series-type Baluns

Figure 10-2 shows circuit diagrams for two versions of the series-type Balun. **Photo 10-B** shows a design using **Figure 10-2A** mounted in a CU 3005-A minibox 5 inches long by 4 inches wide by 3 inches high. The 1:2 Unun has 7 trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. The output tap is located 6 turns from terminal 5. Winding 5-6 is No. 14 H Thermaleze wire and the other two are No. 16 H Thermaleze wire.

The 1:1 Guanella Balun has 14 bifilar turns of No. 14 H Thermaleze wire on a 2.4-inch OD ferrite toroid with a permeability of 250. Both wires are covered with Teflon tubing, which results in a characteristic impedance of 100 ohms (the optimum value). A crossover, placing 7 turns on one side of the toroid

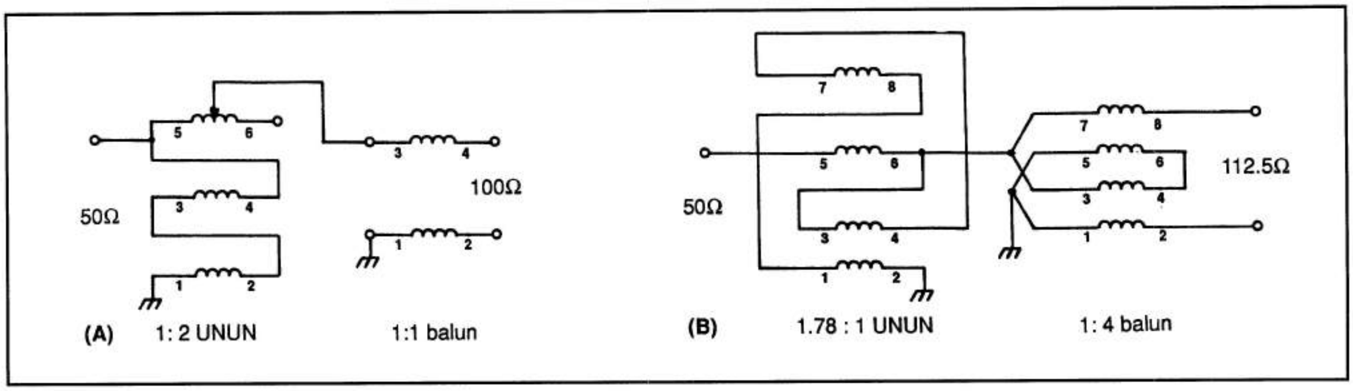

Figure 10-2. *Schematic diagrams of two versions of the series-type Balun: (A) 1:2 (50:100 ohms) Balun; (B) 1:2.25 (50:112.5 ohms) Balun.*

and 7 turns on the other, is used so the output and input are on opposite sides of the toroid. **Figure 10-3** is a drawing of the crossover. Although this technique has no electrical advantage at HF, the mechanical advantage is obvious.

When matching 50-ohm cable to a balanced load of 100 ohms, the response is literally flat (within 2 to 3 percent) from 1.5 to 30 MHz. By connecting the output of the Unun to terminal 6 instead of to the tap, the Balun would match 50-ohm cable to a balanced load of 112.5 ohms with about the same response.

Photo 10-C shows two slightly different versions of series-type 2.25:1 Baluns using the circuit of **Figure 10-2B**. Both have the same 1.78:1 step-down Unun, which has 5 quadrifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire, and the other three are No. 16 H Thermaleze wire. Each version also has 8 bifilar turns of No. 14 H Thermaleze wire on both of the 1.5-inch OD ferrite toroids, with a permeability of 250.

The differences are: 1) the Balun on the left in **Photo 10-C** has one layer of Scotch No. 92 tape on one of the wires in each bifilar winding and a crossover after the fourth turn, and 2) the Balun on the right has two layers of Scotch No. 92 tape on one of the wires on one toroid and no extra insulation on the wires of the other toroid. Therefore, one of the windings in the 1:4 Guanella Balun has a characteristic impedance a little less than 50 ohms and the other a little greater than 50 ohms, resulting in a canceling effect. Furthermore, the crossover isn't used in this design. The Balun on the left is mounted in a CU 3006 minibox 5.25 inches long by 3 inches wide by 2.25 inches high. The Balun on the right is mounted in a CU 3015-A minibox 4 inches long by 2 inches wide by 2.75 inches high.

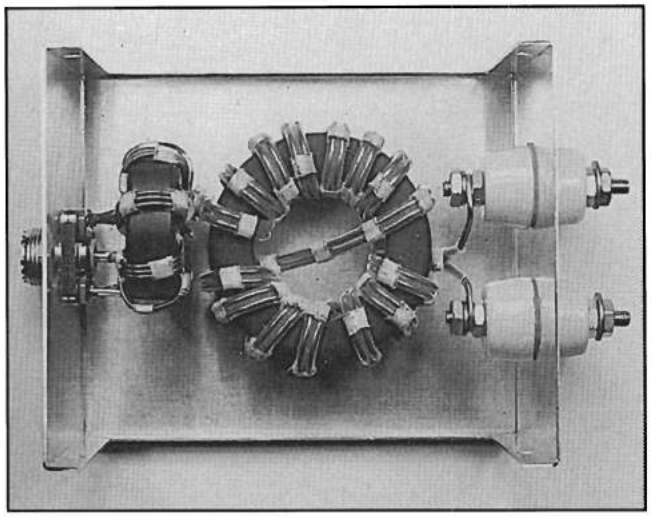

Photo 10-B. *A series-type Balun using the schematic diagram of Figure 10-2A designed to match 50-ohm cable to a balanced load of 100 ohms.*

The performance of these two Baluns is essentially the same. When matching 50-ohm cable to balanced loads of 112.5 ohms, the responses are essentially flat (within 2 to 3 percent) from 1.5 to 30 MHz.

From preliminary measurements on series-type 2:1 Baluns, the Balun in **Photo 10-B** is the one I'd recommend for matching 50-ohm cable to balanced loads of 100 ohms, while the Baluns in **Photo 10-C** would be best for matching to balanced loads of 112.5 ohms. Also, by replacing the 1.78:1 Unun in **Figure 10-2B** with a 2.25:1 Unun, and not adding any extra insulation to the windings of the 1:4 Balun, it's possible to obtain an excellent Balun matching 50-ohm cable to a balanced load of 89 ohms.

Figure 10-4 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to balanced loads of 25 or 22.22 ohms. **Photo 10-D**

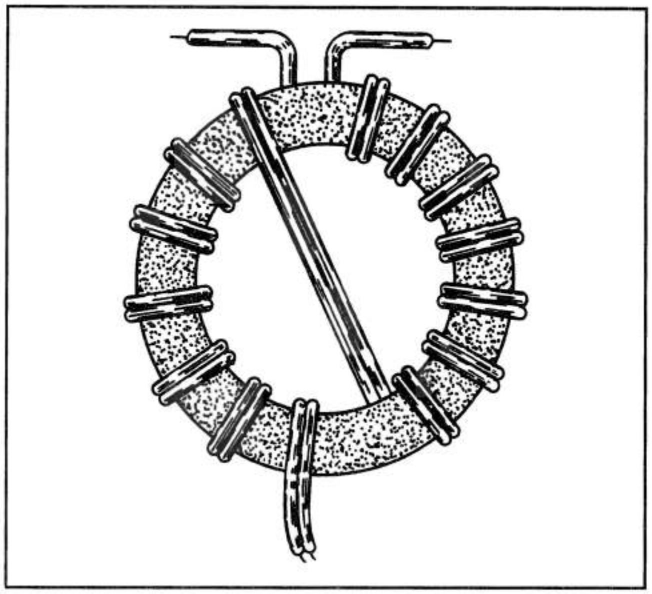

Figure 10-3. *Construction of a 1:1 Guanella Balun with a crossover placing the input and output terminals on opposite sides of the toroid.*

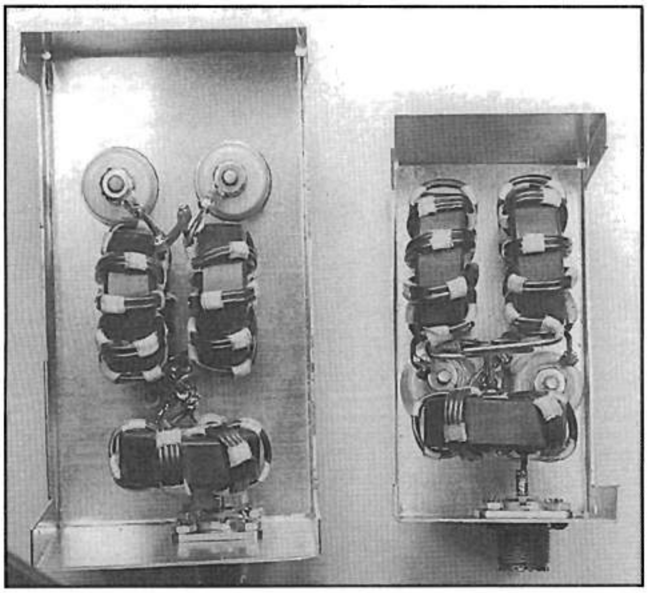

Photo 10-C. *Two series-type Baluns using the schematic diagram of Figure 10-2B designed to match 50-ohm cable to balanced loads of 112.5 ohms.*

shows two versions of this dual-ratio Balun. The Balun on the left, for its Unun, has 6 trifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 3-4 is tapped one turn from terminal 3, yielding the 2:1 ratio. The 1:1 Guanella Balun has 6 turns of homemade coaxial cable on a similar toroid. The inner conductor is No. 12 H Thermaleze wire and is covered with Teflon tubing. The outer braid, from a small coax or from 1/8-inch tubular braid, is tightly wrapped with Scotch No. 92 tape to preserve its low characteristic impedance.

In matching 50-ohm cable to balanced loads of 25 ohms (connection A) or to 22.22 ohms (connection B), the response is essentially flat (within a percent or two) from 1.5 to 30 MHz.

The Balun on the right in **Photo 10-D** has similar windings on a single 2.4-inch OD ferrite toroid with a permeability of 250. Its performance is quite comparable to the Balun on the left.

Sec 10.3.2 Parallel-type Baluns

As you saw in the previous section, the series-type Baluns presented here are combinations of Ununs with ratios of 1.33:1, 1.78:1, 2:1, and 2.25:1 in series with Guanella 1:1 or 4:1 Baluns. The Ununs, which are really an extension of Ruthroff's[9] bootstrap technique for obtaining a 4:1 Unun, sum direct voltages with delayed voltages that traverse a single transmission line. Therefore, the Unun eventually limits the high-frequency response of the series-type Balun.

On the other hand, the parallel-type Balun is an extension of Guanella's technique of summing voltages of equal delays. Instead of simply connecting transmission lines in parallel-series, the parallel-type Balun connects Guanella Baluns in parallel-series. As I noted in **Reference 25**, two 4:1 Guanella Baluns can be connected in parallel-series, yielding very broadband ratios of 6.25:1. This section shows how a 1:1 Guanella Balun can be connected with a 4:1 Guanella Balun in parallel-series, yielding a very broadband ratio of 2.25:1.

Figure 10-5 is the schematic diagram of the high-frequency model of a 2.25:1 Unun which is used for analysis purposes. Because the current through the load is $3/2I_1$, the transformation ratio is $(3/2)^2$, or 2.25:1. Therefore, if the impedance seen on the left side is 50 ohms, a matched impedance on the right side is 22.22 ohms. Because two thirds of the 50 ohms appears across the input of the Guanella 1:1 Balun, its optimum characteristic impedance is 33.33 ohms. Similarly, this is also the value of the optimum characteristic impedance for the windings of the 4:1 Balun. Because the 1:1 Balun wants to see 33.33 ohms on its output (a matched condition) and the 4:1 Balun wants to see 66.66 ohms, placing these two values in parallel results in the confirming value of 22.22 ohms.

If the 50-ohm generator is placed on the right side in **Figure 10-5**, the circuit becomes a step-up Unun matching 50 ohms to 112.5 ohms (on the left). If the ground is removed on the left side, the transformer becomes a Balun. A similar analysis as above, shows that the optimum characteristic impedance of the three bifilar windings now becomes 75 ohms.

Photo 10-E shows a parallel-type 2.25:1 Balun designed to match 50-ohm cable to a balanced load of 112.5 ohms. It has 9 bifilar turns of No. 14 H Thermaleze wire on each of the three toroids that have a 1.5-inch OD and a permeability of 250. Also, one of the wires on each toroid is covered with Teflon tubing, resulting in a characteristic impedance of 75 ohms (the optimum value). When operating as a Balun, the response is essentially flat from 7 MHz to over 45 MHz. As an Unun, the flat response is broadened from 1.5 MHz to over 45 MHz.

Because the coiled wire, parallel-type Balun didn't provide any real advantage over the series-type Balun (in fact, the low-frequency response was poorer), I investigated the beaded transmission line version for possible use in the VHF band. **Figure 10-6** shows a schematic diagram of one using coaxial cable. Obviously, twin lead could be substituted for the coaxial cable. **Photo 10-F** shows both versions.

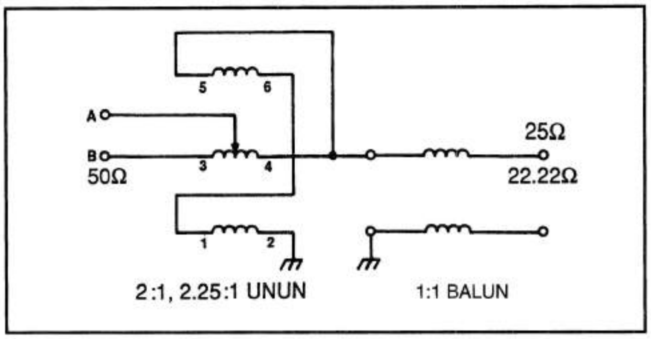

Figure 10-4. *Schematic diagram of a dual-ratio series-type Balun. Connection A matches 50-ohm cable to a balanced load of 25 ohms. Connection B matches 50-ohm cable to a balanced load of 22.22 ohms.*

The top design in **Photo 10-F** has 4 inches of 3/8-inch OD ferrite beads with a permeability of 125 on each of the three 75-ohm transmission lines. It is designed to match 50-ohm cable on the right to a balanced load of 112.5 ohms on the left (with the ground on terminal 1 removed). The transmission lines consist of two No. 14 H Thermaleze wires separated by the Teflon tubing covering one of them. When matching 50-ohm cable to a balanced load of 112.5 ohms, the response is essentially flat from 30 MHz to over 100 MHz (the limit of my bridge).

The bottom design in **Photo 10-F** also has 4 inches of 3/8-inch OD ferrite beads with a permeability of 250. However, they are now threaded by homemade coaxial cable with a characteristic impedance of 33.33 ohms. It is designed to match 50-ohm cable on the left to a balanced load of 22.22 ohms (with the ground on terminal 2 removed). The inner conductor of the coax is No. 14 H Thermaleze wire and is covered with

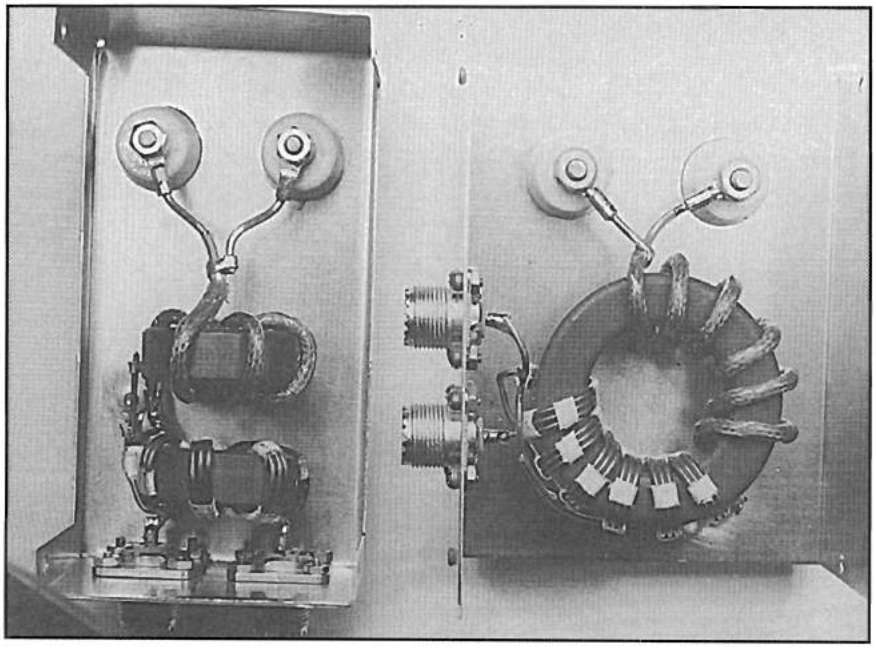

Photo 10-D. *Two series-type Baluns using the schematic diagram of Figure 10-4 designed to match 50-ohm cable to balanced loads of 25 ohms or 22.22 ohms.*

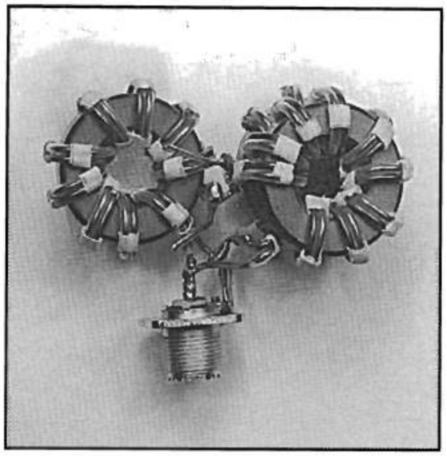

Photo 10-E. *Parallel-type 2.25:1 Balun matching 50-ohm cable to a balanced load of 112.5 ohms.*

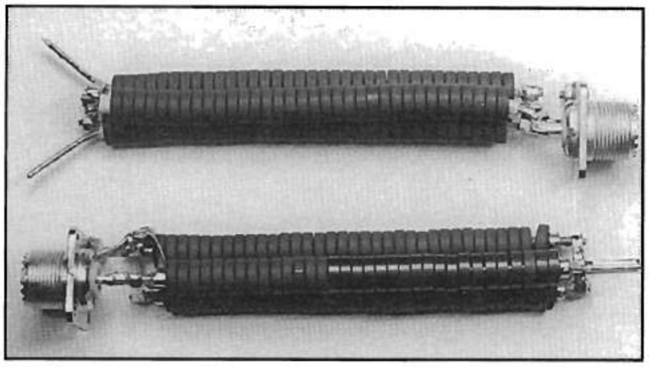

Photo 10-F. *Beaded transmission line versions of the 2.25:1 parallel-type Balun for operation in the VHF band. The top Balun matches 50-ohm cable to a balanced load of 112.5 ohms. The bottom Balun matches it to a balanced load of 22.22 ohms.*

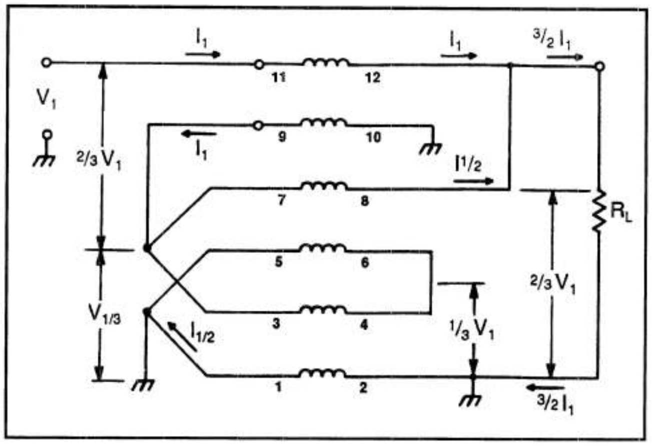

Figure 10-5. *High-frequency model of the parallel-type 2.25:1 transformer. Connections shown are for Unun operation.*

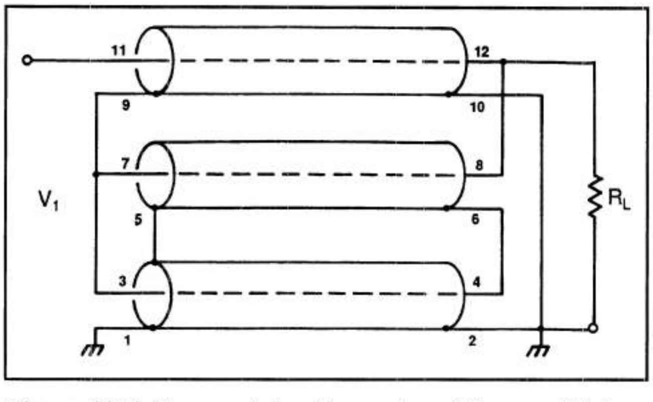

Figure 10-6. *The coaxial cable version of the parallel-type 2.25:1 transformer of Figure 10-5.*

Teflon tubing. The braid is from small coaxial cable or from 1/8-inch tubular braid. The braid is also tightly wrapped with Scotch No. 92 tape in order to preserve the low characteristic impedance. When matching 50-ohm cable to a balanced load of 22.22 ohms, the response is essentially flat from 14 MHz to over 100 MHz (again, the limit of my bridge).

Sec 10.4 Closing Comments

In closing, I'd like to make a couple of comments regarding parallel-type Baluns.

First, if you are interested in a 1.78:1 ratio, replace the 1:4 Balun in **Figures 10-5** and **10-6** with a 1:9 Guanella Balun (three transmission lines connected in parallel-series). This would yield an output current of $4/3I_1$ and a ratio of $(4/3)^2$, or 1.78:1. If you replace the 1:4 Balun with a 1:16 Guanella Balun (four transmission lines connected in parallel-series), the output current will be $5/4I_1$, with a ratio of $(5/4)^2$, or 1.56:1.

Second, because the parallel-type Balun is really an extension of Guanella's technique of summing voltages of equal delays,[26,27] the high-frequency response is mainly limited by the parasitics in the interconnections. Therefore, beaded transmission lines offer the best opportunity for successful operation on the VHF band. It is also recommended that the ferrite beads have permeabilities of 300 or less[2] in order to achieve the very high efficiencies of which these transformers are capable.

Chapter 11

6:1, 9:1, and 12:1 Baluns

Sec 11.1 Introduction

This chapter is a combination of two articles I wrote for *Communications Quarterly*. One, on the 6:1 and 9:1 Baluns, appeared in the Winter 1993 issue and the other, on the 12:1 Balun, ran in the Summer 1993 issue. I combined these two articles here because the three Baluns discussed have much in common. Commonalities include: 1) They are the most difficult to construct and are, therefore, the most expensive; 2) two of them, the 6:1 and 12:1 Baluns, use Ununs in series with Guanella Baluns; 3) two of them use the 9:1 Guanella Balun—the 12:1 Balun with a series 1.33:1 Unun; 4) they are generally associated with two-wire transmission lines with characteristic impedances of 300, 450, and 600 ohms, respectively; 5) when matching 50-ohm cable to higher impedances, they have more loss than the Baluns in the preceding chapters; and 6) the optimum application of these Baluns requires a more critical understanding of the trade-offs in low-frequency response for efficiency.

Like the 2:1 Balun in **Chapter 10**, the 6:1 (actually 6.25:1) Balun also comes in two forms: the series-type, which offers better low-frequency response in the HF band, and the parallel-type, which has a vastly greater high-frequency capability. The parallel-type 6:1 Balun, together with the 9:1 Guanella Balun (which is also a parallel-type), offer the potential for designs capable of efficient and broadband operation on the VHF and UHF bands.

Sec 11.2 6:1 and 9:1 Baluns

Many radio amateurs associate the use of the 6:1 and 9:1 Baluns with 300-ohm twin lead feeding folded dipoles, 450-ohm "ladder" line feeding single or multi-band antenna systems. However, what is neglected (in some cases) is the effect of the height of these antennas above earth and the length of the transmission lines feeding them.

Broadband 6:1 and 9:1 Baluns are considerably more difficult to construct than the more common 1:1 and 4:1 Baluns. This is especially true when matching 50-ohm cable to impedances of 300 and 450 ohms. Furthermore, there are some important trade-offs in low-frequency response for efficiency.

From what I could gather "on the air" or talking to radio clubs, I have determined what I believe are probably two of the most common misconceptions regarding the use of these Baluns:

1. 6:1 Baluns. In free-space, the folded dipole with 300-ohm twin-lead has a resonant impedance close to 300 ohms. The dipole also has this value at a height of about 0.225 wavelength above ground. However, it's only 200 ohms at a height of about 0.17 wavelength and 400 ohms (the maximum) and at 0.35 wavelength. In many cases, the 4:1 Balun would actually do a better job of matching.

2. 9:1 Baluns. Some are unaware of the relationship between the impedance at the input of a transmission line, the characteristic impedance of the line, and the impedance at the end of the line. Just because a transmission line has a characteristic impedance of 450 ohms doesn't necessarily mean that a 9:1 Balun will perform a satisfactory match to 50-ohm cable. Far from it. For example, if the line is terminated by a half-wave dipole with an impedance of 50 ohms, the 9:1 Balun would see 50 ohms when the line is a half-wave long and 4050 ohms when it's a quarter-wave long! Broadband Baluns cannot be designed to handle impedances as high as 4050 ohms. It's very likely that a well-designed 50:450-ohm Balun would experience (particularly on 80 and 160 meters) harmful flux in the core *and* excessive heating because of the large voltage drop along the length of its transmission lines. This problem of presenting very high (and harmful) impedances to Baluns is quite prevalent with multi-band antenna systems.

Clearly, there are many applications for 6:1 and 9:1 Baluns. They not only include matching 50-ohm cable

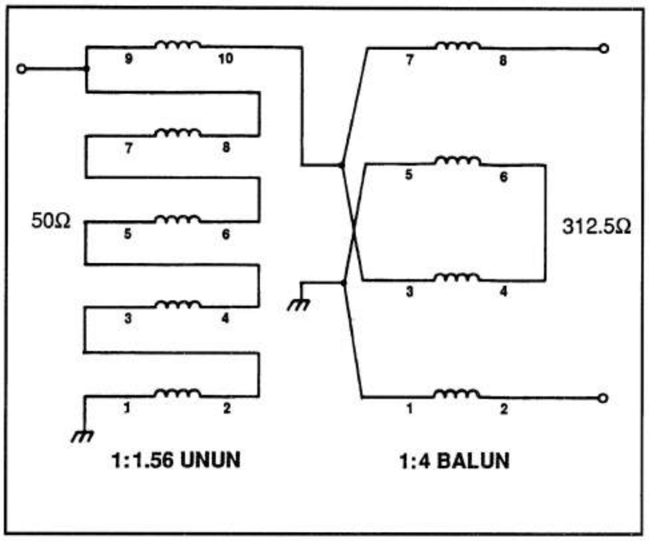

Figure 11-1. *Schematic diagram of the series-type Balun with a 1:6.25 ratio designed to match 50 to 312.5 ohms.*

to balanced loads of 300 and 450 ohms, but also to balanced loads of 8 and 5.6 ohms, as well. Furthermore, many of the designs in this chapter will perform almost as well in Unun (unbalanced-to-unbalanced) applications. The trade-off (which is usually very small) is in low-frequency response. Additionally, these Baluns could be used to exploit the low-loss properties of 300- and 450-ohm twin-lead where very long transmission lines are used. This is especially important at 14 MHz and above.

In the pages that follow, I'll present a variety of Baluns matching 50-ohm cable to 300 ohms (actually to 312.5 ohms, a 6.25:1 ratio) and to 450 ohms, as well as 50-ohm cable to 8 and 5.6 ohms. Also included are two different versions of 6.25:1 Baluns. One is a series-type using a 1.56:1 Unun in series with a 4:1 Guanella Balun and the other a series-parallel arrangement using two 4:1 Guanella Baluns. Because the series-parallel Balun adds voltages of equal delays, you'll find its high-frequency capability is much greater.

The 9:1 Balun is a conventional Guanella Balun with three transmission lines connected in series at the high-impedance side and in parallel at the low-impedance side. Therefore, it also sums voltages of equal delays. Some of the comparisons and analyses of these 6.25:1 and 9:1 Baluns were probably published for the first time in my Winter 1993 *Communications Quarterly* article.

Sec 11.2.1 6.25:1 Series-Type Baluns

Figure 11-1 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to a balanced load of 312.5 ohms. It consists of a 1:1.56 Unun in series with a 1:4 Guanella Balun. The overall ratio of 1:6.25 should satisfy most of the 1:6 requirements. **Photo 11-A** shows three examples. All three Baluns use the same step-up Unun that has four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 9-10 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. Because this Unun sums only one delayed voltage with four equal direct voltages, it has an excellent high-frequency response.[2]

The Balun on the left in **Photo 11-A** has eight bifilar turns of No. 18 hook-up wire on each transmission line of its 1:4 Balun. The wires are further spaced with No. 18 Teflon tubing providing a characteristic impedance close to 150 ohms (the optimum value). The ferrite toroid has a 2.4-inch OD and a permeability of 250. When matching 50-ohm cable to a floating

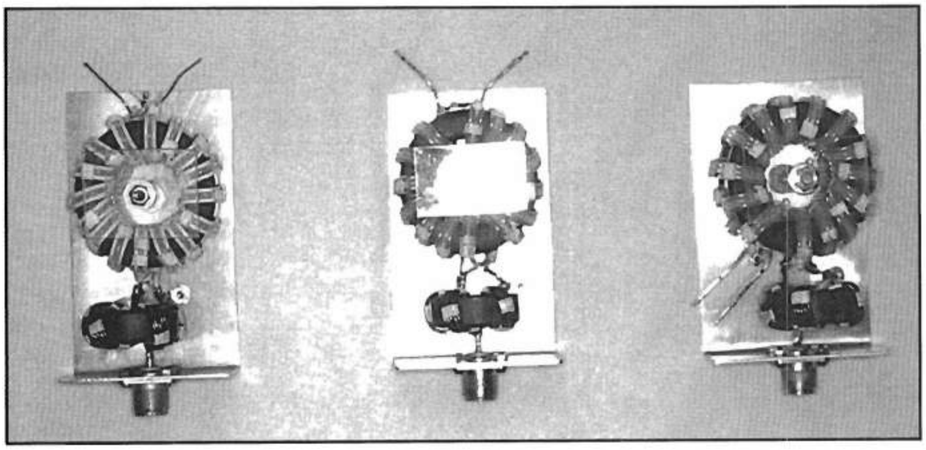

Photo 11-A. *Three examples of series-type 1:6.25 Baluns. The Balun on the right, with a double-core 1:4 Balun, has both a balanced voltage and current output. The other two only have balanced-current outputs.*

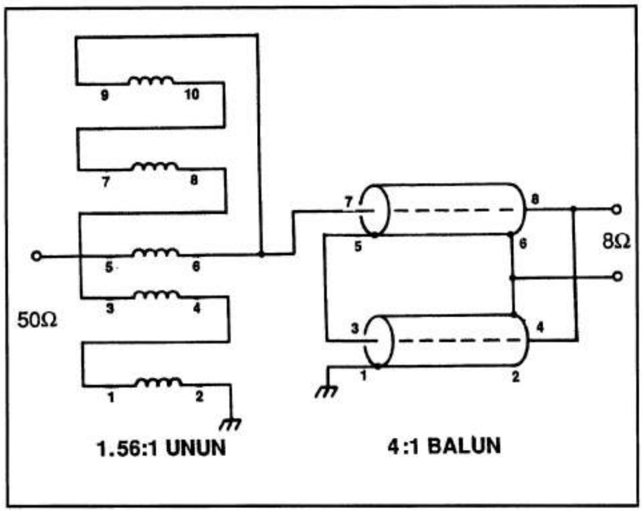

Figure 11-2. *Schematic diagram of the series-type Balun with a 6.25:1 ratio designed to match 50 to 8 ohms.*

load of 312.5 ohms, the response is essentially flat from 1.7 to 30 MHz. Under matched conditions, 500 watts of continuous power and 1 kW of peak power is a conservative power rating. Because the 1:4 Balun in this series-type 1:6.25 Balun uses only one core instead of two, this transformer should never be used when the load is grounded at its center. Also, it is not recommended for balanced antennas. This series-type Balun presents balanced currents, but does *not* present balanced voltages.

The Balun in the center of **Photo 11-A** has seven bifilar turns of No. 16 SF Formvar wire on each transmission line on its 1:4 Balun. The wires are covered with Telfon sleeving and further separated by No. 16 Teflon tubing. Like the Balun on the left, the characteristic impedance is also close to the objective of 150 ohms. The toroid also has a 2.4-inch OD and a permeability of 250. When matching 50-ohm cable to a floating load of 312.5 ohms, the response is essentially flat from 3.5 to 30 MHz. Over this frequency range, this Balun can easily handle the full legal limit of amateur radio power. Because this Balun also presents balanced currents and not balanced voltages, it should not be used when the loads are balanced to ground or grounded at their centers.

The Balun on the right in **Photo 11-A** has 14 bifilar turns of No. 16 SF Formvar wire on each of the two toroids of the 1:4 Balun. The wires are also covered with Teflon sleeving and further separated by No. 16 Teflon tubing. For ease of connection, one core is wound clockwise and the other counterclockwise. The two cores are spaced 1/4 inch apart with acrylic sections. When matching 50-ohm cable to a 312.5 load that is either floating, balanced to ground, grounded at its center, or grounded at its bottom (a broadband Unun), the response is essentially flat from 1.7 to 30 MHz. Under this matched condition, it can easily handle the full legal limit of amateur radio power. Furthermore, this is a true Balun because it presents equal currents and equal voltages. If one were to measure the voltages-to-ground, when matching into a balanced load, one would find them to be equal and opposite. The other two 1:6.25 Baluns using single-core 1:4 Baluns, would have equal currents but not equal voltages (see **Chapter 8**). Because they are easier to construct, it would be interesting to compare them with a true Balun.

Figure 11-2 shows the schematic diagram of a series-type Balun designed to match 50-ohm cable to a balanced load of 8 ohms (perhaps a short-boom Yagi). It consists of a 1.56:1 step-down Unun in series with a Guanella 4:1 step-down Balun. The overall ratio is 6.25:1. **Photo 11-B** shows two examples. Both Baluns use the same step-down Unun, which has four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. The interleaving of the wires is such that the performance is optimized for matching 50 to 32 ohms.

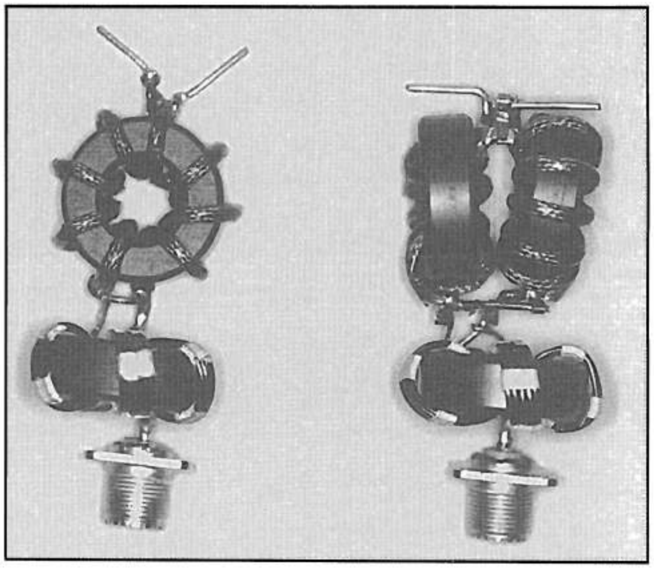

Photo 11-B. *Two examples of the series-type 6.25:1 Balun optimized at the 50:8-ohm level. The Balun on the left is designed to match into a floating 8-ohm load. The Balun on the right is designed to match into an 8-ohm floating, center-tapped-to-ground or grounded load (Unun).*

The Balun on the left in **Photo 11-B** has four turns of low-impedance coaxial cable on each transmission line on the single-core 4:1 Balun. The inner conductor is No. 14 H Thermaleze wire, and it has two layers of Scotch No. 92 polyimide tape. The outer braid is from a small coax (or 1/8-inch tubular braid) and is tightly wrapped with Scotch No. 92 tape to achieve the 17-ohm characteristic impedance (the optimum value). The ferrite toroid has a 1.5-inch OD and a permeability of 250. When matching 50-ohm cable to a floating load of 8 ohms, the response is flat from 1 to 40 MHz. In a matched condition, this Balun can easily handle the full legal limit of amateur radio power.

The Balun on the right in **Photo 11-B** has six turns of the same coaxial cable on each of the two cores of the Guanella step-down Balun. The cores also have a 1.5-inch OD and a permeability of 250. The performance of this Balun is practically the same as the above with the single-core 4:1 Balun. The important differences are that this 6.25:1 Balun also performs equally well whether the load is center-tapped-to-ground, balanced-to-ground, or grounded at the bottom (a broadband Unun). This is the one recommended for feeding a short-spaced Yagi beam antenna.

Sec 11.2.2 6.25:1 Parallel-Type Baluns

The 6.25:1 series-type Baluns described in the preceding section consisted of a 1.56:1 Unun, which is an extension of Ruthroff's bootstrap approach for Ununs,[9] in series with a Guanella 4:1 Balun.[2] The upper-frequency limit for this combination is really set by the Unun, which sums a delayed voltage with four direct voltages. The parallel-type 6.25:1 Baluns described in this section are really extensions of Guanella's approach, which sums voltages of equal delays. Therefore, the upper-frequency limit is mainly dependent upon the parasitics in the interconnections.

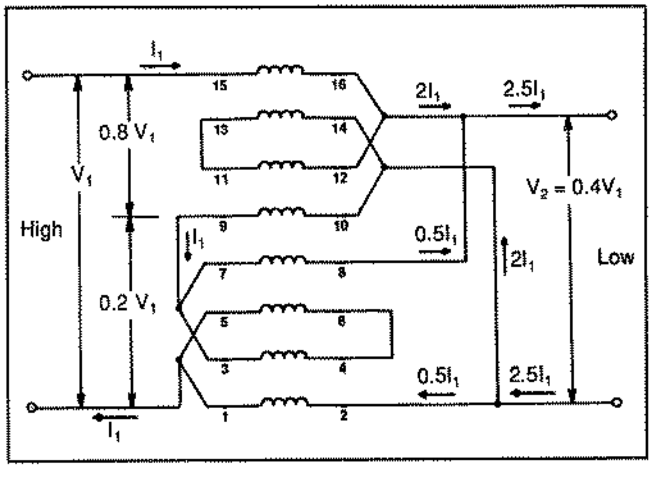

Figure 11-3. *Schematic diagram of the parallel-type Balun (and Unun) with a 6.25:1 ratio. The currents and voltages are shown for analysis purposes (see text).*

The 6.25:1 parallel-type Balun uses two 4:1 Guanella Baluns connected in parallel on the low-impedance side and in series on the high-impedance side. As you will see, one of the Baluns is reversed, giving the desired ratio of 6.25:1. Other combinations can produce different fractional-ratios (other than $1{:}n^2$ where n is 1, 2, 3, . . .), like 2.25:1 and 1.78:1. Because very little practical design information is available regarding this family of very broadband Baluns,[26,27] this section also includes my high-frequency analysis of the 6.25:1 parallel-type Balun. It should also be pointed out that very little sacrifice in performance occurs whether the load is grounded at its center or at the bottom (as an Unun).

Figure 11-3 shows the coiled-wire version of the 6.25:1 parallel-type Balun. For analysis purposes, the voltages and currents are also shown. As can be seen, the top 4:1 Balun is connected as a step-down Balun, while the bottom 4:1 Balun is connected as a step-up Balun. The Baluns are in series on the high-impedance side (on the left) and in parallel on the low-impedance side (on the right). As **Figure 11-3** illustrates, the lower 1:4 Balun adds a current of $0.5I_1$ to the load, resulting in a total current of $2.5I_1$. Thus, the impedance transformation ratio is 2.5^2, or 6.25:1.

For maximum high-frequency response, each transmission line should see a load equal to its characteristic impedance. In other words, they should be "flat" lines. If the high side on the left is 50 ohms, then 40 ohms appears on the input of the top Balun and 10 ohms on the input of the bottom Balun. Consequently, the optimum characteristic impedance for *all* transmission lines is 20 ohms. On the low-impedance side on the right, the top Balun wants to see 10 ohms, while the bottom Balun wants to see 40 ohms. Because 10 ohms in parallel with 40 ohms equals 8 ohms, each Balun conveniently sees its ideal load and a broadband ratio of 6.25:1 is obtained.

If the Balun is required to match 50-ohm cable (on the right side) to a balanced load of 312.5 ohms (on the left side), the same analysis shows that the optimum characteristic impedance of all the transmission lines is 125 ohms.

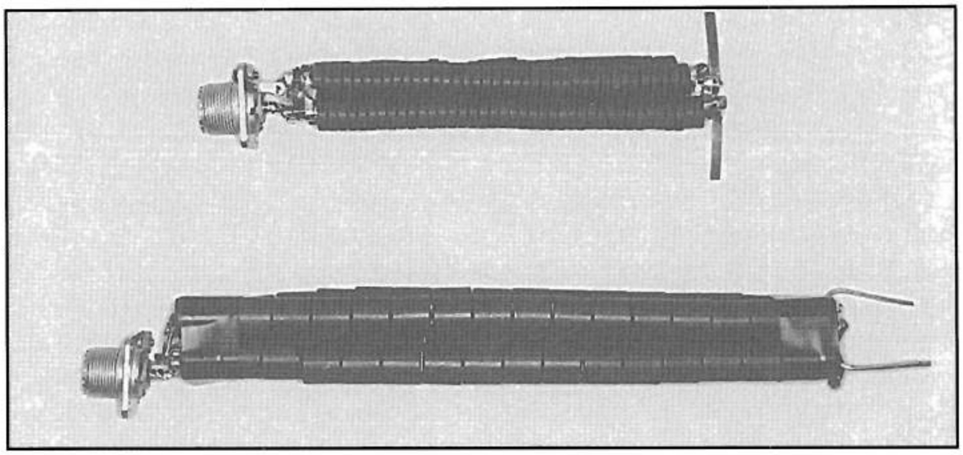

Photo 11-C. *Two beaded-line versions of the parallel-type 6.25:1 Balun (and Unun). The top transformer is designed to match 50-ohm cable to 8 ohms. The bottom transformer is designed to match 50-ohm cable to 312.5 ohms.*

Because the parallel-type Balun (or Unun) sums voltages of equal delays and, therefore, has no built-in high-frequency cut-off, it has a real advantage over the series-type Balun on the VHF bands and above. Furthermore, beaded transmission lines with low-permeability ferrite beads (125 and less) can be used, resulting in high efficiencies. On the HF band, where coiled windings are generally used, the series-type Balun is preferred because of its simplicity.

Photo 11-C shows two beaded-line 6.25:1 transformers. The top Balun, designed to match 50-ohm cable (on the left) to 8 ohms (on the right), uses low-impedance coaxial cable lines. The schematic diagram is shown in **Figure 11-4**. It has 5 inches of 0.375-inch OD beads (permeability 125) on four coaxial cables with characteristic impedances of 20 ohms. The inner-conductors of No. 12 H Thermaleze wire have two layers of Scotch No. 92 polyimide tape. The outer braids, from small coaxial cable or 1/8-inch tubular braid, are also wrapped tightly with the same tape in order to preserve the 20-ohm characteristic impedance. When matching 50 ohms to 8 ohms, the response is essentially flat from 10 MHz to beyond 100 MHz (the limit of my simple bridge). Under this matched condition, this Balun can easily handle the full legal limit of amateur radio power. Furthermore, it has practically the same performance when operating as an Unun (both terminals 1 and 2 grounded). In the Unun application, the bottom transmission line has no voltage along it and, therefore, requires no beads.

The bottom Balun in **Photo 11-C**, which is designed to match 50-ohm cable to a balanced load of 312.5 ohms, has 8 inches of 0.5-inch OD beads on 125-ohm twin-lead transmission lines. The ferrite beads also have a permeability of 125. The wires are No. 14 H Thermaleze wire and are covered with Teflon sleeving. They are further separated by No. 18 Teflon tubing. When matching 50-ohm cable (on the right side) to 312.5 ohms (on the left side), the response is essentially flat from 20 MHz to over 100 MHz. Under this matched condition, this Balun can also easily handle the full legal limit of amateur radio power. Additionally, this transformer performs practically as well when used as an Unun.

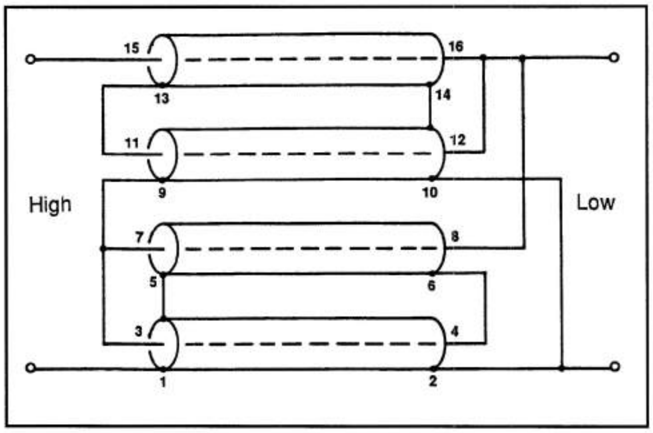

Figure 11-4. *Schematic diagram of the coaxial-cable version of the parallel-type 6.25:1 Balun (and Unun).*

Sec 11.3 9:1 Baluns

The broadband 9:1 Balun, matching 50-ohm cable to a balanced load of 450 ohms, is one of the most difficult ones to construct because high-impedance transmission lines (150 ohms) are required for maximum high-frequency response, and greater reactances are needed in order to isolate the input from the output. So one can appreciate the task at hand, this section also provides a brief review of the theory of these devices.[2]

Figure 11-5 shows the high- and low-frequency models of the Guanella 9:1 Balun that connect three transmission lines in series at the high-impedance side and in parallel at the low-impedance side. Because Guanella's Baluns (which can be easily converted to Ununs) sum voltages of equal delays, they offer the highest frequency capability.

The high-frequency model (**Figure 11-5A**) assumes there is sufficient choking reactance in the coiled (or beaded) transmission lines to isolate the input from the output and only allow transmission line currents to flow. Under this condition, the analysis is rather straightforward as it only involves transmission line theory. Simply stated—the maximum high-frequency response occurs when each transmission line is terminated in a load equal to its characteristic impedance, Z_0. Thus, the transmission lines in the 9:1 Balun have no standing waves. Because each transmission line sees one third of the load, the optimum value of Z_0 is $R_L/3$. Except for parasitics in the interconnections and self-resonances in coiled windings, Guanella's approach is literally "frequency independent."

On the other hand, the low-frequency analysis of the Guanella 9:1 Balun is most important because it reveals the major difficulty in designing them for low-loss, wideband operation. **Figure 11-5B** is the model for determining the low-frequency response. It assumes that no energy is transmitted to the load by a transmission line mode. Although the terminology and analysis is the same as that used for conventional autotransformers, the similarity ends when there is sufficient choking reactance to only allow for the efficient transmission line mode.

As with conventional transformers, one can analyze the low-frequency response of the 9:1 Balun from either the low- or high-impedance side. By putting the generator on the low-impedance side in **Figure 11-5B**, I've chosen to analyze it from that side. With the output open-circuited, the generator sees four coiled (or beaded) lines connected in series-parallel. The net result is that the generator sees the reactance of only *one* coiled (or beaded) line. To prevent a shunting current to ground (and/or autotransformer operation), the reactance the generator sees should be much greater than R_g (at least by a factor of 10 at the lowest frequency of interest). The inductance of the coiled or beaded line that prevents the unwanted currents is still known as the magnetizing inductance, L_M.

What's important to note here is that the low-frequency model of the Guanella 4:1 Balun does not have the series-parallel combination of coiled or beaded lines.[2] Only two lines, which are in series, exist in its model. Therefore, for a two-core Guanella 4:1 Balun having the same number of turns (and same cores) as a 9:1 Guanella Balun, its low-frequency response is better by a factor of two!

Another advantage that goes to the Guanella 4:1 Balun when matching 50 to 200 ohms, is in the number of turns that can be wound on the same cores. Since 4:1 Baluns require characteristic impedances of 100 ohms (instead of 150), the width of the transmission lines is considerably less, thus allowing for more turns. Also, as will be shown later, the efficiency of the 4:1 Balun is greater because the potential drops

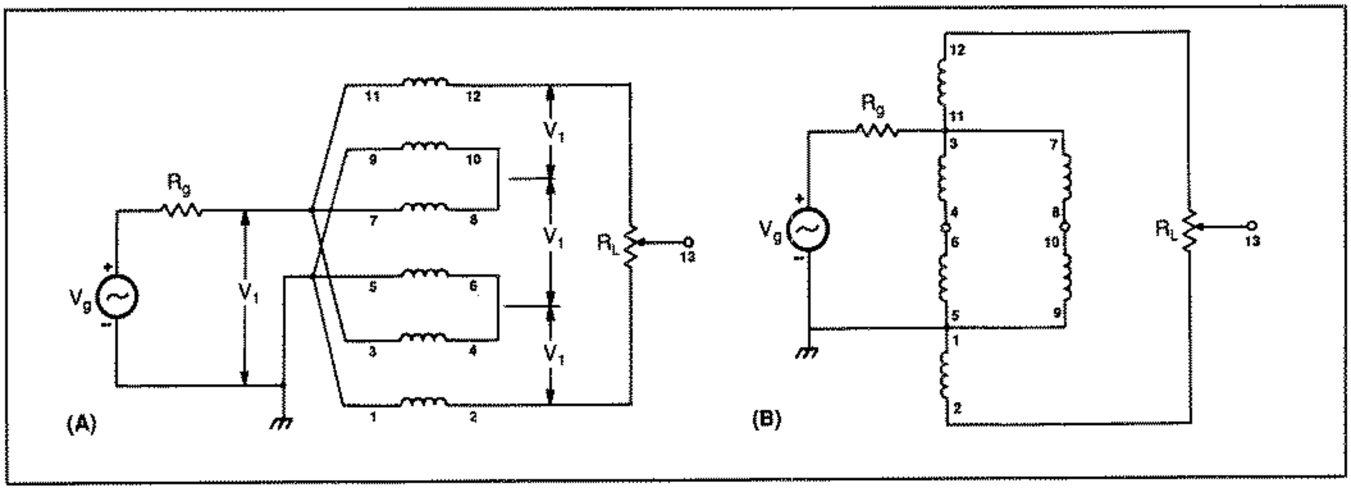

Figure 11-5. *Models of Guanella's 1:9 Balun. The high-frequency model, (A), assumes that* $Z_0 = R_L/3$*, and therefore* V_2*, the output of each transmission line, equals* V_1*. The low-frequency model, (B), assumes no energy is transmitted to the load,* R_L*, by a transmission line mode.*

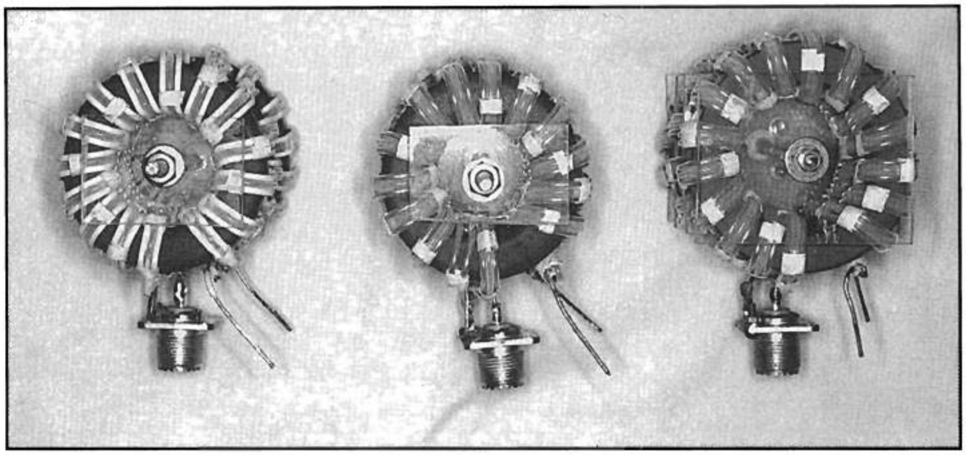

Photo 11-D. *Three broadband Guanella 9:1 Baluns designed to match 50-ohm cable to 450 ohms. The transformer on the left, using No. 18 hook-up wire, can handle 500 watts from 1.7 to 45 MHz. The transformer in the center, using No. 16 wire, can handle 1 kW from 3.5 to 45 MHz. The transformer on the right (with larger cores), using No. 16 wire, can handle 1 kW from 1.7 to 45 MHz.*

along the transmission lines are lower (less dielectric loss). Finally, as you can see from **Figure 11-5B**, by also grounding terminal 2 (Unun operation), windings 1-2 and 3-4 are both shorted—degrading the low-frequency response because L_M is reduced by one third.

Another interesting analysis with Baluns and Ununs concerns the potential gradients (voltage drops) along the transmission lines. Because the loss with these transformers, when transferring the energy via a transmission line mode, is a dielectric-type (voltage dependent), the higher the gradient, the greater the loss. The interesting cases occur when the load is: a) floating, b) grounded at the center, and c) grounded at the bottom (an Unun).

Floating load. With terminal 13 in **Figure 11-5A** ungrounded, the top transmission line has a gradient of $+V_1$ and the bottom transmission line has a gradient of $-V_1$. The center transmission line has a gradient of zero. Therefore, the center transmission line only acts as a delay line and doesn't require a magnetic core or beads. As a result, the top and bottom cores (or beads) account for the dielectric loss.

Load grounded at the center. With terminal 13 grounded at the center of the load, the top transmission line has a gradient of $+V_1$, the bottom transmission line has a gradient of $-V_1$, and the center transmission line has a gradient of $-V_1/2$. This configuration results in about 25 percent more loss because of the extra gradient along the center transmission line. Incidentally, this condition exists when matching into balanced systems like 450-ohm transmission lines or antennas because they have virtual grounds at the center of the loads they present.

Load grounded at the bottom. With terminal 13 grounded at the bottom of the load (an Unun), the top and center transmission lines have gradients of $+V_1$. The bottom transmission line has no gradient and, therefore, no loss. It only acts as a delay line and thus requires no magnetic core or beads.

Sec 11.3.1 Some Practical 9:1 Balun Designs

Photo 11-D shows three versions of the broadband Guanella 9:1 Balun designed to match 50-ohm cable to 450-ohm loads. The transformer on the left has 15 bifilar turns of No. 18 hook-up wire on each of the three ferrite toroids with a 2.4-inch OD and permeability of 250. The wires are further separated by No. 16 Teflon tubing, resulting in a characteristic impedance close to 150 ohms (the optimum value). The cores in this Balun, as well as the other two that follow, are spaced 1/4 inch apart by sections of acrylic. In matching 50 to 450 ohms, the response is essentially flat from 1.7 to 45 MHz. In this matched condition, this transformer can easily handle 500 watts of continuous power and 1 kW of peak power.

The transformer in the center of **Photo 11-D** has 14 bifilar turns of No. 16 SF Formvar wire on each of the three ferrite toroids with a 2.4-inch OD and permeability of 250. The wires are covered with Teflon sleeving and further separated by No. 16 Teflon tubing. The characteristic impedance is also close to the optimum value of 150 ohms. In matching 50 to 450 ohms, the response is essentially flat from 3.5 to 45 MHz. In this matched condition, this transformer can easily handle 1 kW of continuous power and 2 kW of peak power. **Photo 11-E** shows this Balun mounted in a minibox 6 inches long by 5 inches wide by 4 inches high.

The transformer on the right in **Photo 11-D** is designed to handle 1 kW of continuous power and 2 kW

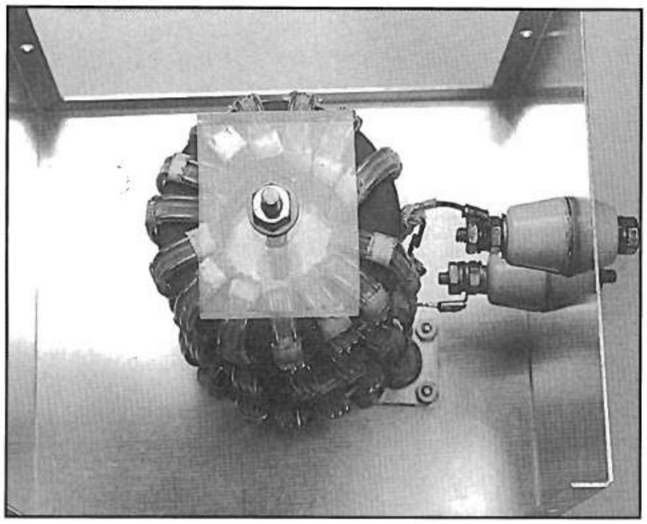

Photo 11-E. *The Guanella 9:1 Balun shown mounted in a large minibox.*

of peak power from 1.7 to 45 MHz. Because this transformer uses larger cores with slightly higher permeabilities (2.68-inch OD and 290 permeability), which are not as popular as those used in the other two Baluns, it is much more expensive to construct. It has 16 bifilar turns of No. 16 SF Formvar wire on each toroid. The wires are covered with No. 16 Teflon sleeving and further separated with No. 16 Teflon tubing. In matching 50 to 450 ohms, the response is flat from 1.7 to 45 MHz.

Finally, **Figure 11-6** shows the schematic diagram of a 9:1 Balun (or Unun) using coaxial cables wound around ferrite cores or threaded through ferrite beads. This form of the transformer is especially useful when matching 50-ohm cable to 5.6 ohms because the choking reactance of the magnetizing inductance, L_M, only need be much greater than 5.6 ohms. **Photo 11-F** shows two different designs.

The transformer on the top has 9 1/2 turns of low-impedance coaxial cable on each rod. The rods are 1/2 inch in diameter, 2 1/2 inches long, and have a permeability of 125. The low-frequency response of this Balun is quite insensitive to the length and permeability of the rods.[2] The inner conductors of the coaxial cables are No. 12 H Thermaleze wire with two layers of Scotch No. 92 tape. The outer braids are made from small coax (or 1/8-inch tubular braid). They are further wrapped with Scotch No. 92 tape to preserve the low-impedance of 17 ohms. In matching 50-ohm cable to 5.6 ohms, the response is essentially flat from 1.7 to 30 MHz. The optimum impedance level was found when matching 40 to 4.45 ohms. The addition of another layer of Scotch No. 92 tape would optimize this transformer at the 50:56-ohm level, and the high-frequency response would exceed 100 MHz. This transformer is very efficient and should handle the full legal limit of amateur radio power easily.

The bottom transformer in **Photo 11-F** is designed to match 50-ohm cable to a load of 5.6 ohms in the VHF band. It uses 3 1/2 inches of beads on three low-impedance coaxial cables. The ferrite beads have an OD of 3/8 inch and a permeability of 125. The inner conductors of the coaxes are No. 12 H Thermaleze wire with one layer of Scotch No. 92 tape. The outer conductors are from small coaxes or 1/8-inch tubular braid, and are also tightly wrapped with Scotch No. 92 tape to preserve the low characteristic impedance. In matching 50-ohm cable to 5.6 ohms, the response is essentially flat from 7 MHz to over 100 MHz (the limit of my test equipment). This 9:1 Balun (which can be used as an Unun) can handle the full legal limit of amateur radio power under matched conditions, because of the low-permeability beads and the low-voltage gradients along the lengths of its transmission lines.

Sec 11.4 Concluding Remarks on 6:1 and 9:1 Baluns

One of the most important properties of broadband Baluns and Ununs (which all use ferrites) is their capability of having extremely high efficiencies. Knowing the loss mechanism in these transformers and the trade-off in low-frequency response for efficiency allows one to optimize their applications. In the paragraphs that follow, I'll discuss the losses and

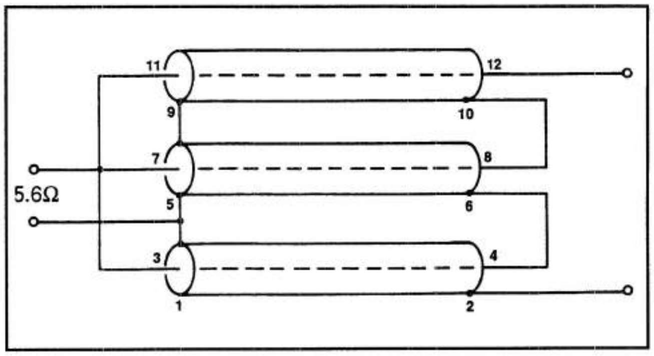

Figure 11-6. *Schematic diagram of a coaxial cable (beaded or coiled) 9:1 Guanella Balun (or Unun). This design is especially useful in matching 50-ohm cable to 5.6 ohms over a very wide bandwidth.*

trade-offs involved with the transformers presented in the preceding sections. The approach used here should be applicable to all forms of transmission line transformers. This section ends with a review of two articles that contained 9:1 Baluns. As you'll see, I have some rather different views on the claims made in these articles.

Accurate measurements on many broadband Ununs have found the losses to be related to the permeability and the impedance level.[2] Permeabilities greater than 300 resulted in excessive losses. Because these losses are unlike the conventional transformer whose losses are current-dependent, it can only be assumed that their losses are voltage dependent; in other words a dielectric-type loss. Therefore, higher-impedance transformers have higher voltage gradients along their transmission line and, thus, have greater losses. Additionally, it was found that the higher the permeability, the greater the loss with frequency. Taking into account the accurate measurements and the factors noted above, I offer these loss values for the transformers in the preceding sections:

1.56:1 Ununs. The 1.56:1 Ununs (either step-up or step-down) used in series with Guanella 4:1 Baluns to form 6.25:1 Baluns have the lowest potential gradients along their transmission lines. Voltage drops of only about 0.2 V_1, where V_1 is the input voltage, exist along their transmission lines. Accurate measurements have shown losses, in a matched condition, of only 0.04 dB. If the cores, which have a permeability of 250, were replaced with cores having a permeability of 125, the losses could be as low as 0.02 dB over much of the passband. The low-frequency response would still be acceptable at 1.7 MHz. This Unun is a natural for matching into 75-ohm hard line when long transmission lines are required.

6.25:1 and 9:1 Low-impedance Baluns. Baluns matching 50-ohm cable to 8 or 5.6 ohms, also have very low voltage drops along their transmission lines. Generally, they are about twice that of the 1.56:1 Unun. Therefore, the losses with these Baluns should be on the order of 0.1 dB in their passbands.

6.25:1 High-impedance Baluns. The losses in the series-type Baluns are mainly in the 1:4 Guanella Baluns, which have potential gradients of about 1.25 V_1, where V_1 is the input voltage. From previous measurements at this impedance level, the suggestion is that the losses (with ferrites of 250 permeability) should be about 0.1 dB at 7 MHz and 0.2 dB at 30 MHz. By using toroids with permeabilities of 125, the losses could be 0.07 dB and 0.15 dB, respectively. However, with a permeability of 40, the losses could be as low as 0.05 dB within the passband. However, one must consider the sacrifice in low-frequency response incurred when using these lower-permeability ferrites. With a permeability of 125, it's poorer by a factor of 2. With a permeability of 40, it's poorer by a factor of 8! The 6.25:1 parallel-type Balun in this article uses ferrite beads with a permeability of 125 and, therefore, should have losses similar to its series-type counterpart.

9:1 High-impedance Baluns. As was shown in the preceding section, the potential gradient along two of the transmission lines is V_1, where V_1 is the input voltage. The third transmission line, with a balanced load (or as an Unun), has no potential gradient and, consequently, no loss in its core. Because the loss with the series-type Balun mainly exists in one core, the loss with the 1:9 Balun should be a little less than twice as great. With ferrite cores of 250 permeability, the suggested losses are 0.2 dB at 7 MHz and 0.4 dB at 30 MHz. With cores of 125 permeability, the losses are about 0.14 and 0.28 dB, respectively. Again, by using cores with permeabilities of 40, the losses are practically negligible—approximately 0.1 dB within its passband.

As in the case of the 1:6.25 Baluns above, similar trade-offs occur in the low-frequency response. That is, if 125 permeability cores are used, the low-frequency response is poorer by a factor of 2; with 40 permeability cores, it's poorer by a factor of 8. The major difference here is that the low-frequency performance of the

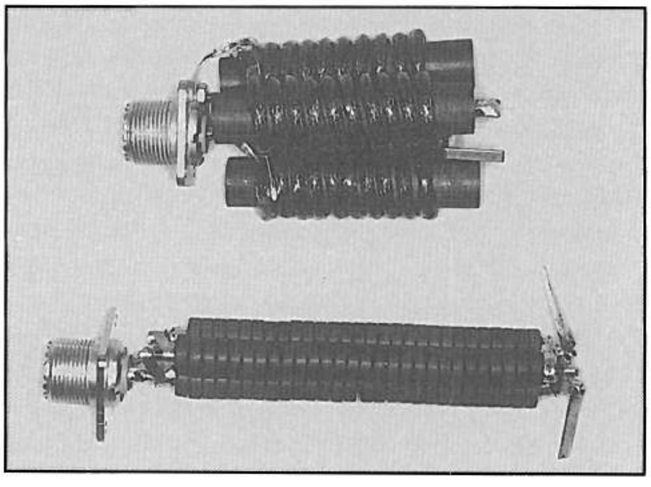

Photo 11-F. *Two versions of the coaxial-cable 9:1 Guanella Balun (or Unun) designed to match 50-ohm cable to 5.6 ohms*

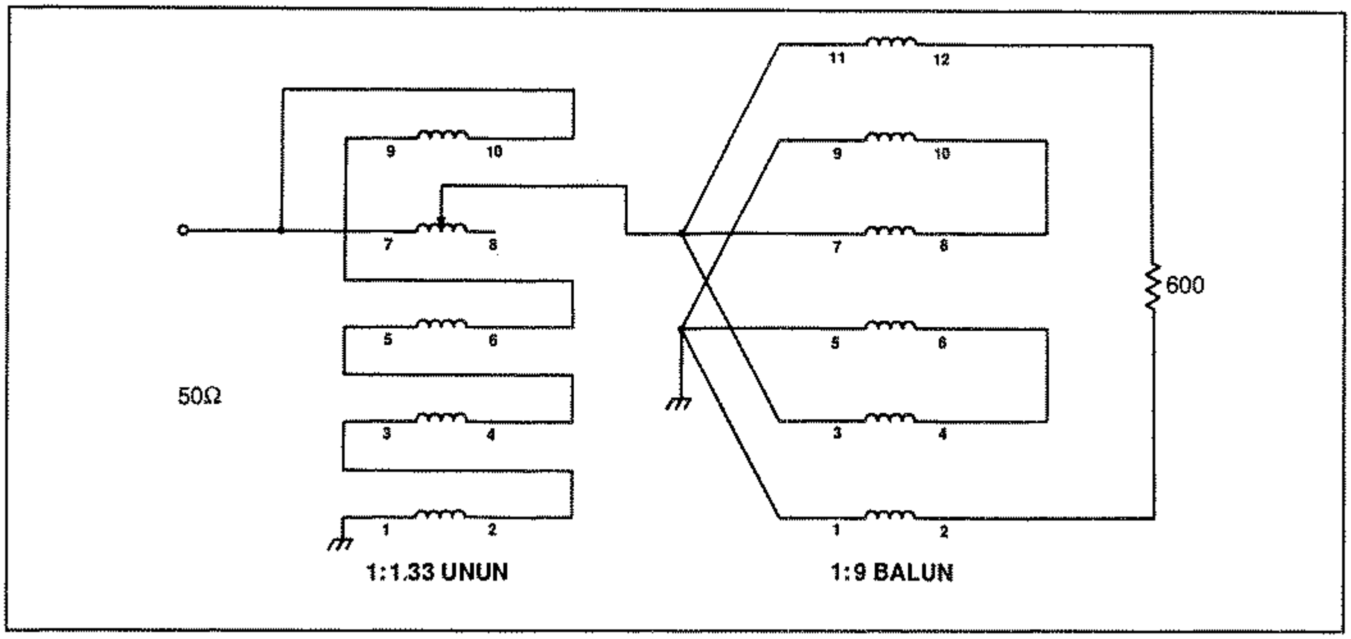

Figure 11-7. *Schematic diagram of the series-type 12:1 Balun using a 1:1.33 Unun in series with a 1:9 Guanella Balun.*

9:1 Balun, as seen by its low-frequency model, isn't as good as the 4:1 Balun that controls the low-frequency response of the series-type 1:6.25 Balun. Additionally, it should be pointed out that all of the suggested losses for the transformers in this chapter are for matched conditions—that is with VSWRs of 1:1. If the VSWR is 2:1 due to a load twice as large as the objective, the input voltage to the Balun increases by about 40 percent. Therefore, the losses should increase by close to the same percentage.

In closing this section, I would like to report my findings on two recent articles in amateur radio journals that also described 9:1 Baluns matching 50-ohm cable to 450 ohms. One[5] advocated using three 150-ohm coaxial cables threaded through high-permeability ferrite beads—a 9:1 Guanella Balun. Because of the low voltage-breakdown capability of the coaxial cable and the high loss found by accurate measurements on Ununs using these high-permeability ferrites, the design was suspect. I built a copy of the design and found it to be, as expected, unable to handle any appreciable power. The second article[28] advocated using 14 trifilar turns of "magnet wire" on a 2-inch OD powdered-iron core (permeability of 10). This Balun was also constructed and tested. Again, as was expected, when matching 50-ohm cable to a floating load of 450 ohms, the 9:1 Balun barely reached a true 9:1 ratio at 7 MHz. Above 7 MHz, the ratio became greater than 9:1 and also introduced a reactive component. Below 7 MHz, there was insufficient choking reactance to prevent flux in the core. My three objections to this design are: 1) a trifilar design has a poor high-frequency response because it sums a direct voltage with a delayed voltage that traverses a single transmission line and a delayed voltage that traverses two transmission lines, 2) the characteristic impedances of the transmission lines are only 50 ohms (the objective is 150 ohms), and 3) the low-frequency response is poor because of the low-permeability powdered-iron core. I do not recommend either of the designs in these two articles.

Sec 11.5 12:1 Baluns

Over the years, the broadband, 12:1 Balun has been of special interest to users of rhombic and V antennas. With the aid of this Balun, certain advantages over multi-element arrays can be fully exploited. Rhombics and Vs are easier to construct, both electrically and mechanically, and there are no particularly critical dimensions or adjustments. Furthermore, they give satisfactory gain and directivity over a 2-to-1 frequency range. These antennas have also been found to be more effective in reception. Because their designs can present input impedances of 600 ohms, and very long lengths of highly efficient 600-ohm open-wire line can be used between the shack and the antenna, an efficient and broadband 12:1 Balun is a natural for

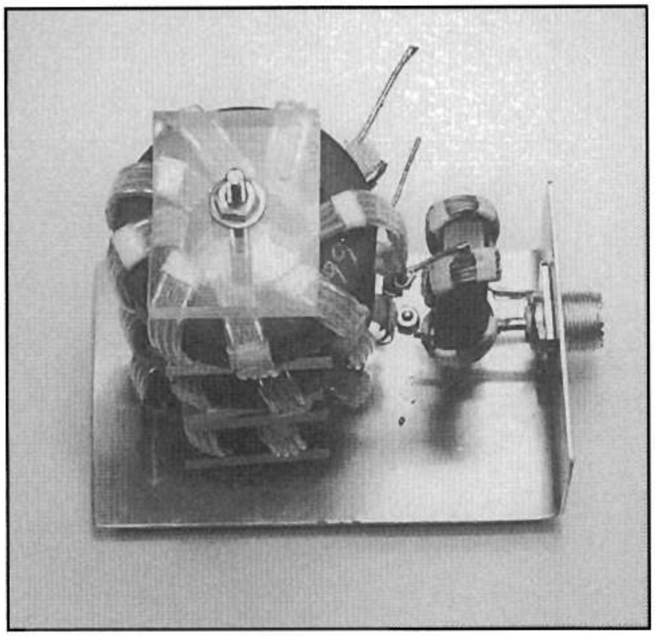

Photo 11-G. *Top view of the high-power 12:1 Balun.*

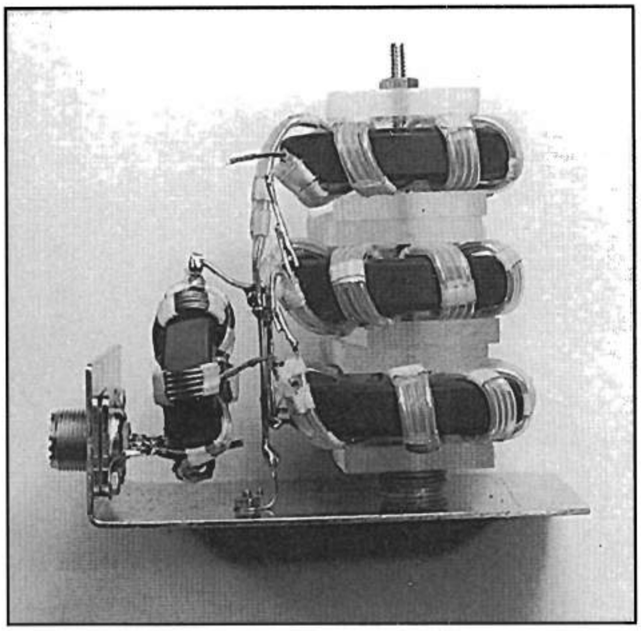

Photo 11-H. *Side view of the high-power 12:1 Balun.*

this application. However, to my knowledge, satisfactory Baluns have not been available for this use.

I will present two versions of a series-type Balun designed to match 50-ohm cable to a balanced load of 600 ohms. One is a high-power unit designed to handle the full legal limit of amateur radio power over a bandwidth of 7 to 30 MHz. The other is a medium-power unit capable of handling approximately one-half the legal limit of amateur radio power from 3.5 to 30 MHz. Both Baluns use a 1:1.33 Unun in series with a 1:9 Guanella Balun.

As you will see, these Baluns are not especially easy to design and construct. The major difficulties lie in trying to obtain sufficient choking reactances in the coiled windings to meet the low-frequency requirements, and large enough characteristic impedances of the windings to meet the high-frequency requirements. Because a coiled winding with a characteristic impedance of 200 ohms (the objective) is practically impossible to obtain with any reasonable wire size and number of turns, I used the compensating technique first described in my book.[2] Because the characteristic impedances of the 9:1 Guanella Balun are somewhat less than 200 ohms, a compensating effect (and hence higher frequency response) can be obtained by having a higher (than the normal objective) characteristic impedance of the windings in the 1:1.33 Unun. Earlier work (also described in my book) presented a 12:1 Balun using a 1:3 Unun in series with a 1:4 Ruthroff Balun. The Baluns presented in this section using a 1:1.33 Unun in series with a Guanella 1:9 Balun, are much improved designs.

Sec 11.5.1 A High-power 12:1 Balun

Figure 11-7 shows the schematic diagram of the series-type 12:1 Balun used in both the high- and medium-power versions. **Photo 11-G** shows a top view of the high-power Balun. **Photo 11-H** shows a side view.

The 1:1.33 Unun has 5 quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 7-8 is No. 14 H Thermaleze wire and the other four are No. 16 H Thermaleze wire. Winding 7-8 is also tapped at 3 turns from terminal 7.

The 1:9 Guanella Balun has 8 bifilar turns of tinned No. 16 wire on each of the three toroids. Each wire is covered with Teflon tubing and further separated by two Teflon tubings. The characteristic impedance of the windings is about 190 ohms (the objective is 200 ohms). The ferrite toroids have an OD of 2.4 inches and a permeability of 250. The spacing between the toroids is 1/2 inch.

In matching 50-ohm cable to a balanced load of 600 ohms, the response is literally flat (within a percent or two) from 7 to 30 MHz. Within this bandwidth, it is capable of handling the full legal limit of amateur

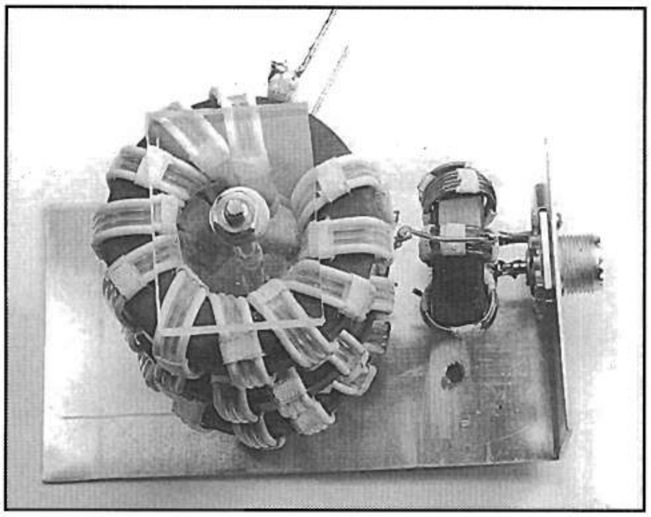

Photo 11-I. *Top view of the medium-power 12:1 Balun.*

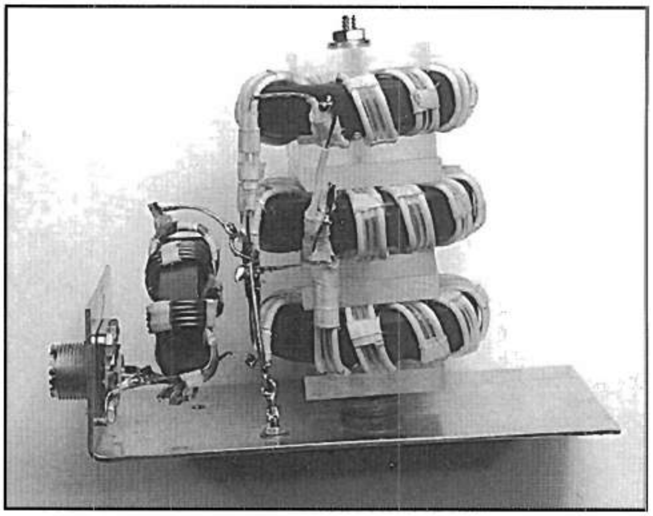

Photo 11-J. *Side view of the medium-power 12:1 Balun.*

radio power. In a matched condition, the expected insertion loss is about 0.25 dB.

Sec 11.5.2 A Medium-power 12:1 Balun

Photo 11-I shows the top view of the medium-power 12:1 series-type Balun. **Photo 11-J** shows the side view. This Balun also has the same 1:1.33 Unun as described earlier.

The 1:9 Guanella Balun has 11 bifilar turns of No. 18 hook-up wire on each toroid. The wires are further separated by two No. 18 Teflon tubings. The characteristic impedance of the windings is about 170 ohms. The toroids have an OD of 2.4 inches and a permeability of 250. The spacing between the toroids is also 1/2 inch.

In matching 50-ohm cable to a balanced load of 600 ohms, the response varies less than 5 percent from 3.5 to 30 MHz. Within this bandwidth, the Balun can handle about one-half the legal limit of amateur radio power. As with the high-power version, the expected insertion loss is also 0.25 dB.

Sec 11.6 Concluding Remarks on 12:1 Baluns

Many of the concluding remarks from the discussion on 6:1 and 9:1 Baluns (**Sec 11.4**) also apply to 12:1 Baluns; therefore, I won't repeat them here. But the following four remarks are specific to 12:1 Baluns and warrant mentioning:

First, high-impedance transmission line transformers like the 12:1 Balun are particularly sensitive to metallic enclosures. If a minibox were to be used, I would suggest the one shown in **Photo 11-E**, which is 6 inches long by 5 inches wide by 4 inches high. Smaller metallic enclosures would reduce the characteristic impedances of the windings and affect the high-frequency response. Even the spacing between cores had to be increased from 1/4 inch (for a 50:450-ohm Balun) to 1/2 inch. The subchassis shown in the photographs were used because they provided the necessary electrical and mechanical support.

Second, the 12:1 Baluns described in this section also make excellent Ununs, albeit with some compromise in the low-frequency response. I would suggest using the high-power unit only between 14 and 30 MHz and the medium-power unit only between 7 and 30 MHz.

Third, for the readers interested in VHF operation, I would suggest the parallel-type approach described in the earlier section on 6:1 Baluns. In this case, a 9:1 Guanella Balun is connected in series-parallel with a 1:4 Guanella Balun. This would produce a broadband ratio of 12.25:1. By using 170-ohm twin-lead (about 10 inches long) threaded through ferrite beads with a permeability of 125, it appears that it is possible to match 50-ohm cable to a balanced load of 612.5 ohms throughout the VHF band.

Fourth, by using torids with a permeability of 125 in the 1:9 Guanella Baluns (of the 12:1 Baluns), the insertion loss would be reduced by around one half (0.12 dB), with a trade-off in low-frequency response. The high-power unit would now cover about 10 to 30 MHz, and the medium-power Balun would cover about 7 to 30 MHz.

Chapter 12

The 4:1 Unun

Sec 12.1 Introduction

From an analysis standpoint, the 4:1 Unun can be said to have received the most attention in the literature. It began with Ruthroff's introduction and complete analysis of this device in his classic paper published in 1959.[9] Ruthroff's paper then became the industry standard for this class of devices known as *transmission line transformers*. These are devices that transmit the energy from the input to the output by an efficient transmission line mode, and not by flux linkages (as in conventional transformers).

However 15 years earlier, Guanella had introduced, in his classic 1944 paper,[3] the first broadband Baluns by combining coiled transmission lines in a series-parallel arrangement, yielding ratios of $1:n^2$ where n = 1, 2, 3,... and so on. It has also been shown that Guanella's technique also lends itself to Ununs as well.[2] In fact, in this chapter, you will see that his technique of summing voltages of equal delays promises to yield high-power designs capable of operating on the VHF and UHF bands.

The 4:1 Unun also exemplifies (more than any other transformer) the many choices that can be made in its design. These include: 1) Ruthroff's or Guanella's designs, 2) wire or coaxial cable transmission lines, 3) coiled or beaded lines, 4) rods or toroids, 5) low-power or high-power designs, 6) HF, VHF, or UHF designs, and 7) the trade-offs in efficiency for low-frequency response or for high VSWR. The 4:1 Unun is the most prevalent of all the Ununs. It finds extensive use in solid-state circuits and in many antenna applications involving the matching of ground-fed antennas—where impedances of 12 to 13 ohms must be matched to 50-ohm coaxial cable. This chapter provides information on many 4:1 Unun designs.

Sec 12.2 The Ruthroff 4:1 Unun

Figure 12-1 illustrates two versions of Ruthroff's approach to obtaining a 4:1 unbalanced-to-unbalanced transformer (Unun). As can be seen, one uses a coiled wire transmission line, while the other uses a coiled coaxial cable. Depending upon the frequency, beaded transmission lines may also be used.

Ruthroff's design uses a single transmission line connected in, what I call, the bootstrap configuration. That is, terminal 2 is connected to terminal 3, lifting the transmission line (at the high-impedance side) by the voltage V_1. If the reactance of the coiled winding or beaded line is much greater than R_G, then only flux-canceling transmission line currents are allowed to flow. It is also apparent that the output voltage is the sum of a direct voltage, V_1, and a delayed voltage,

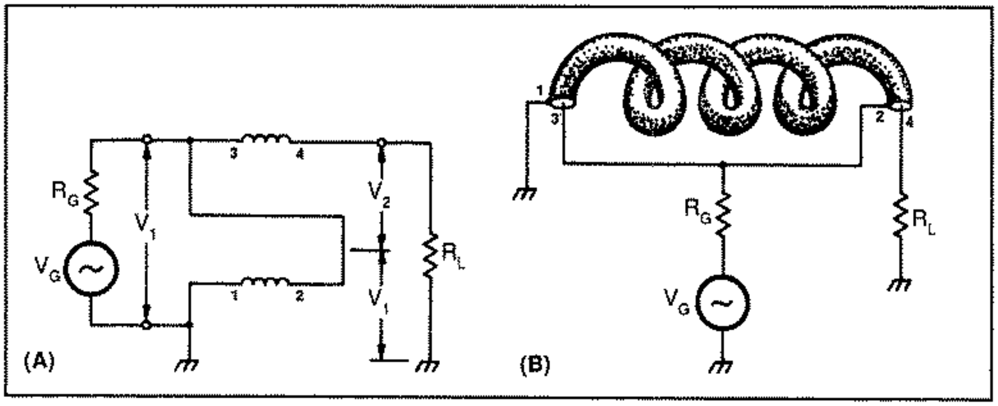

Figure 12-1. *The Ruthroff 4:1 Unun (R_L=$4R_G$): (A) coiled bifilar winding; (B) coiled coaxial cable.*

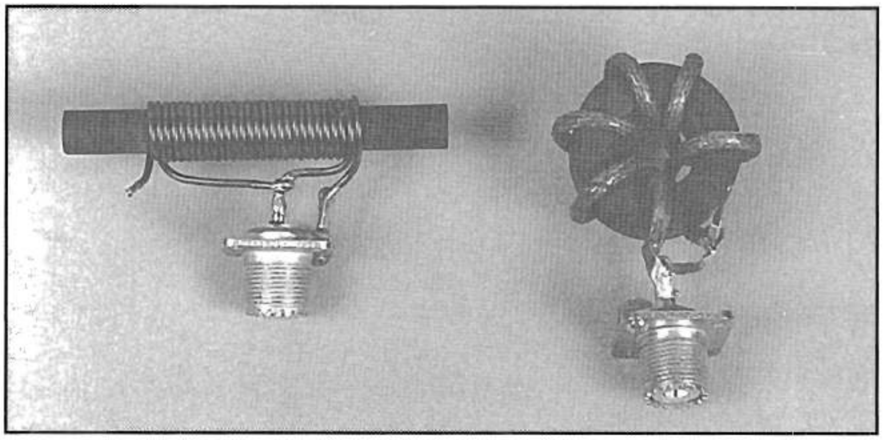

Photo 12-A. *Two versions of the Ruthroff 4:1 (50:12.5-ohm) Unun: coiled wire rod (on the left); coiled coaxial cable toroid (on the right).*

V_2, which traverses a single transmission line. This delay in V_2 eventually limits the high-frequency response. For example, if the electrical length of the line is a 1/2 wave, the output is zero. Ruthroff also found that the optimum value of the characteristic impedance of the transmission line (for maximum high-frequency response) is $R_L/2$.

Therefore, the electrical length and characteristic impedance of the transmission line play major roles in Ruthroff's design. Because his work was mainly concerned with small-signal applications, Ruthroff was able to obtain broad bands of a few tens of kilohertz to over a thousand megahertz. This was possible because he used a few turns (5 to 10) of fine wire (Nos. 37 and 38) on high-permeability toroids as small as 0.08 inches in OD. As a result, the phase-delay with these very short transmission lines was very small. However, large-signal (power) applications present an entirely different picture. For operation in the HF band (including 160 meters), transmission lines vary between one to three feet in length (depending upon impedance level). Consequently, phase-delay can play a major role, as will be seen in the following examples.

Sec. 12.2.1 50:12.5-ohm Ununs

Photo 12-A shows two examples of efficient and broadband 4:1 Ununs matching 50 to 12.5 ohms. The rod version (on the left) has 14 bifilar turns of No. 14 H Thermaleze wire on a low-permeability (125) ferrite rod 0.375 inches in diameter and 3.5 inches long. The connections are shown in **Figure 12-1A**. The cable connector is on the low-impedance side. The response is flat from 1.5 to 30 MHz. In a matched condition, this Unun can easily handle the full legal limit of amateur radio power. Because a tightly wound rod Unun yields a characteristic impedance very close to 25 ohms (the optimum value), this is quite likely the easiest one to construct that covers the above bandwidth.

The toroidal version (on the right in **Photo 12-A**) has 6 turns of homemade, low-impedance coaxial cable on a 1.5-inch OD ferrite toroid with a permeability of 250. The connections are shown in **Figure**

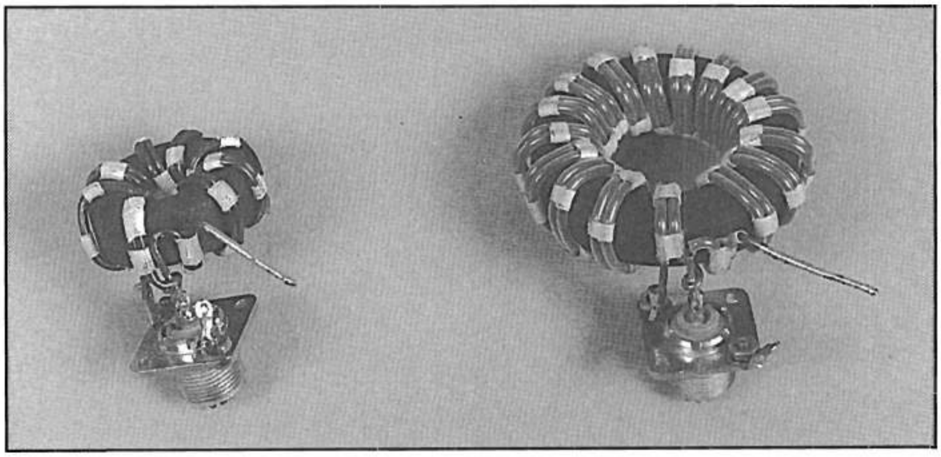

Photo 12-B. *Two higher-impedance Ruthroff 4:1 Ununs: 100:25-ohm (on the left); 200:50-ohm (on the right).*

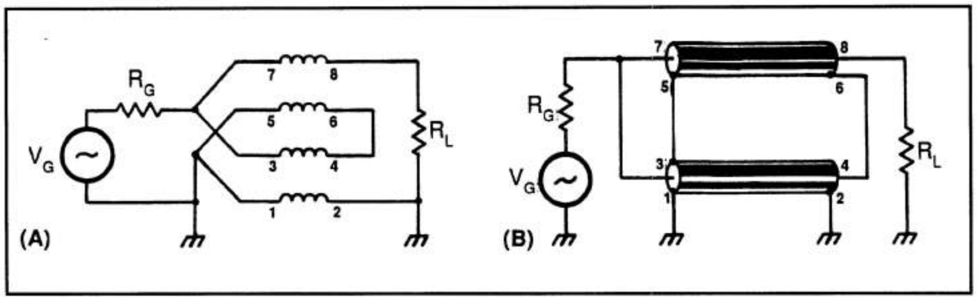

Figure 12-2. *The Guanella 4:1 Unun ($R_L=4R_G$): (A) coiled bifilar windings; (B) coiled or beaded coaxial cables.*

12-1B. The cable connector is on the low-impedance side. The inner conductor is No. 14 H Thermaleze wire and is covered with Teflon tubing. The outer braid is from a small coaxial cable (or from 1/8-inch tubular braid) tightly wrapped with Scotch No. 92 tape in order to obtain the desired characteristic impedance. In matching 50 to 12.5 ohms, the response is flat from 1.5 to 50 MHz. Because the current is evenly distributed on the inner conductor, this small Unun has an exceptionally high power capability—at least 5 kW of continuous power and 10 kW of peak power (in a matched condition).

Sec 12.2.2 100:25-ohm Unun

In some combiner applications, an Unun matching 100 to 25 ohms is required. The smaller toroidal version, pictured on the left in **Photo 12-B**, shows a Ruthroff design that can satisfy many of these requirements. It has 8 bifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. One wire is also covered with a single layer of Scotch No. 92 tape, providing a characteristic impedance close to the desired value of 50 ohms. In matching 100 to 25 ohms, the response is essentially flat from 1.5 to 30 MHz. This Unun can easily handle the full legal limit of amateur radio power.

Sec 12.2.3 200:50-ohm Unun

When dealing with this type of Balun, the Ruthroff approach cannot yield the broadband response of the lower-impedance designs shown above. Because more turns are required in order to obtain the necessary choking reactance, and a 100-ohm characteristic impedance that requires more spacing between the wires is used, the cores must be considerably larger. This results in longer transmission lines. Consequently, the high-frequency response is now limited by the greater phase delay of this high-impedance Unun.

The larger transformer, shown on the right in **Photo 12-B** is my optimized version of a Ruthroff 200:50-ohm Unun. It has 16 bifilar turns of No. 14 H Thermaleze wire on a low-permeability (250) 2.4-inch OD ferrite toroid. Each wire is covered with Teflon tubing, resulting in a characteristic impedance of 97 ohms. Because of the long transmission line (36 inches), the impedance transformation ratio (in matching 200 ohms to 50 ohms) varies from 4 to 4.44 from 1.5 to 30 MHz. A conservative power rating (under a matched condition) is 2 kW of continuous power and 4 kW of peak power. Because this higher-impedance Unun has a larger voltage drop along the length of its windings, its loss (a dielectric-type[2]) is a little greater than the lower-impedance Ununs described earlier. In a matched condition, the efficiency is about 97 percent, while the others experience efficiencies of 98 to 99 percent.

Sec 12.3 The Guanella 4:1 Unun

Even though Guanella's investigation[3] was directed toward developing a broadband Balun to match the balanced output of a 100-watt, push-pull, vacuum-tube amplifier to the unbalanced load of a coaxial

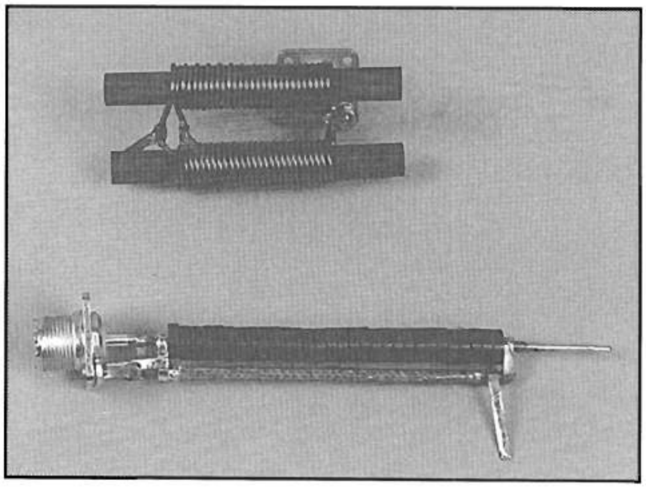

Photo 12-C. *Two Guanella 4:1 (50:12.5-ohm) Ununs: rod version (on the top), 1.5 to 50 MHz; beaded version (on the bottom), 10 MHz to over 100 MHz.*

cable, his technique of connecting transmission lines in a parallel-series arrangement has only recently been recognized as the design for the widest possible bandwidth in an unbalanced-to-unbalanced application.[2] Some have labeled his approach the "equal-delay network".[26] The major difference in Guanella's approach (from Ruthroff's) is that by summing the equal-delay voltages of coiled (or beaded) transmission lines, he minimizes the dependence of the high-frequency response on the lengths of the transmission lines. As was mentioned before, Ruthroff's method of summing a direct voltage with a delayed voltage that traversed a single transmission line has a limited application, especially with high-power, high-impedance Ununs (like 200:50 and 300:75 ohms).

Furthermore, Guanella's approach is also important in designing high- and low-impedance Baluns and Ununs with impedance transformation ratios other than 4:1. Connecting three transmission lines in parallel-series results in a 9:1 ratio, four in a 16:1. Also by connecting a fractional-ratio Unun in series with his Baluns, or by using various combinations of parallel-series transmission lines,[26,27] Ununs and Baluns are now available with a continuum of ratios from 1.36:1 to 16:1. Moreover, these ratios now make it possible to match 50-ohm cable to impedances as low as 3.125 ohms and as high as 800 ohms. A major factor in the success of these designs rests in the understanding of the low-frequency models of these various transformers.[2] This section looks at the 4:1 Unun using Guanella's approach. As in the Ruthroff case, the optimum value of the characteristic impedances of the transmission lines for a Guanella 4:1 transformer is also $R_L/2$.

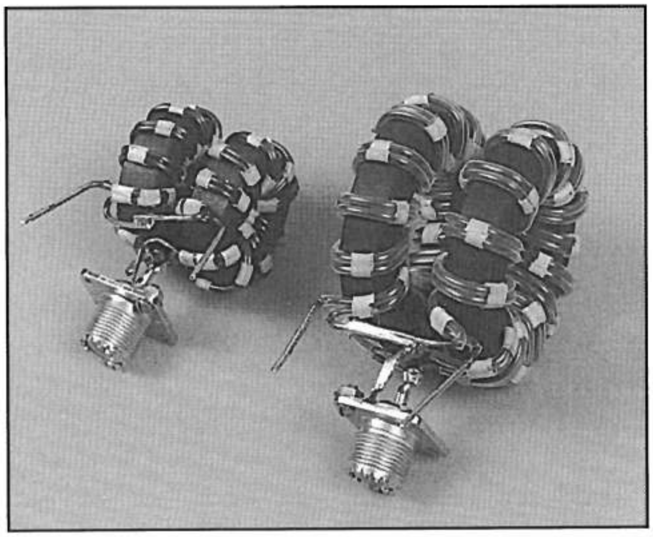

Photo 12-D. *Two higher-impedance Guanella 4:1 Ununs: 100:25-ohm (on the left); 200:50-ohm (on the right).*

Sec 12.3.1 50:12.5-ohm Ununs

Figure 12-2 shows the schematic diagrams of the coiled-wire and coaxial cable (coiled or beaded) versions of 4:1 Ununs using Guanella's technique of connecting transmission lines in parallel-series arrangements. As can be seen in **Figure 12-2**, the lower transmission lines are grounded at both ends and, therefore, have no potential drop along their lengths. Thus, the coiling or beading has no effect. The core only acts as a mechanical support and the beads can be removed. In essence, the bottom transmission line plays the important role of a delay line. In addition, the low-frequency response of this form of Unun is solely determined by the reactance of the top coiled or beaded transmission line.

The top Unun in **Photo 12-C** shows a rod version of Guanella's 4:1 Unun. There are 13.5 bifilar turns of No. 14 H Thermaleze wire on low-permeability (125) ferrite rods 0.375 inches in diameter and 3.5 inches long. For ease of connection, one winding is clockwise and the other is counterclockwise. The cable connector is on the high-impedance side. In matching 50 to 12.5 ohms, the response is flat from 1.5 to over 50 MHz! This Unun, in a matched condition, is capable of handling the full legal limit of amateur radio power. Furthermore, with the 50-ohm generator on the right (in **Figure 12-2A**) and a 12.5-ohm balanced load on the left (perhaps a Yagi beam), this transformer makes an excellent step-down Balun.

The bottom transformer in **Photo 12-C** shows a beaded-coax version of a 50:12.5-ohm step-down Unun designed for 2-meter operation. It has 3.5 inches of beaded coax on the top transmission line (**Figure 12-2B**) and no beads on the bottom transmission line. (Actually, the bottom rod in **Figure 12-2A** can also be removed with no change in performance.) The beads are low-permeability (125) ferrite. The inner conductor of the coaxial cable is No. 12 H Thermaleze wire with about 3.5 layers of Scotch No. 92 tape (two 0.5-inch tapes wound edgewise like a window shade), providing a characteristic impedance close to the optimum value. The outer braid is from a small coaxial cable (or from 1/8-inch tubular braid). This homemade coax is further wrapped tightly with Scotch No. 92 tape in order to preserve its low characteristic impedance. The cable connector is on the low-impedance side. The response of this Unun is essentially flat from 10 to 100 MHz (the limit of my bridge). It can also (easily) handle the full legal limit of amateur radio power.

Sec 12.3.2 100:25-ohm Unun

The Unun on the left in **Photo 12-D** is a Guanella version that matches 100 to 25 ohms. There are 8 bifilar turns of No. 14 H Thermaleze wire on each 1.5-inch OD low-permeability (250) toroid. One toroid is wound clockwise and the other is wound counterclockwise. One of the wires (on each toroid) is covered with one layer of Scotch No. 92 tape. The cable connector is on the low-impedance side. The response is flat from 1.5 MHz to well over 30 MHz. This Unun can also handle the full legal limit of amateur radio power.

It is interesting to note that when used as a Balun (the ground removed from terminal 2), and placed in series (on the left side) with a 1.78:1 Unun (see **Chapter 13**), this compound arrangement provides an excellent Balun for matching 50-ohm coaxial cable directly to quad antennas having impedances of 100 to 110 ohms.

Sec 12.3.3 200:50-ohm Unun

The transformer on the right in **Photo 12-D** is an excellent Unun (or Balun with terminal 2 removed from ground) for matching 50 to 200 ohms. It has 14 bifilar turns of No. 14 H Thermaleze wire on each low-permeability (250) toroid with a 2.4-inch OD. Each wire is covered with Teflon tubing, providing a characteristic impedance of 98 ohms (which is quite good because the optimum value is 100). Again, for ease of connection, one winding is clockwise and the other is counterclockwise. When operating as an Unun or a Balun and matching 50 to 200 ohms, the response is essentially flat from 1.5 to 30 MHz. A conservative power rating (in a matched condition) is 5 kW of continuous power and 10 kW of peak power. This transformer has been reported to handle peak pulses of 10,000 volts!

Summary

Since its introduction by Ruthroff in 1959,[9] the 4:1 Unun has been the most popular transmission line transformer matching unbalanced impedances to unbalanced impedances. As I mentioned at the beginning of this chapter, there are many choices to consider when designing these broadband and efficient transformers. One of the most important choices involves whether to use the Ruthroff or Guanella approach. In fact, the Guanella design should probably be designated a *Balun/Unun*. Recently, it has become the design of choice in the higher frequency bands. From the designs presented in this chapter, I offer the following recommendations:

1. For Ununs in the HF band with impedance levels of 100:25 ohms and lower, the Ruthroff approach is recommended because of its simplicity.
2. For high impedance levels in the HF band (like 200:50 and 300:75 ohms), the Guanella approach is recommended.
3. For low-impedance operation on the VHF band, the beaded-coax Guanella approach is recommended.
4. For high-impedance operation on the VHF band, the coiled-wire Guanella approach appears to be the preferred choice, and should be investigated first. Obviously, the number of turns should be reduced from the examples shown in this chapter because the reactance of the winding is proportional to the frequency.
5. For high-power use on the HF band, the Ruthroff Unun with low-impedance coaxial cable on a toroid (on the right on **Photo 12-A**) is recommended. It is easy to construct and can very likely handle more than 5 kW of continuous power.
6. Also, at high-impedance levels, one might consider using lower permeability ferrites for higher efficiencies. Look at permeabilities of 125 and 40.

Chapter 13

1.33:1, 1.5:1, and 2:1 Ununs

Sec 13.1 Introduction

Little practical design information has been available on Ununs with impedance transformation ratios of less than 4:1 (these are called *fractional-ratio* Ununs). However, many important applications can be found for efficient and broadband Ununs with ratios like 1.33:1, 1.5:1, and 2:1. Some examples include the matching of 50-ohm coaxial cable to: a) vertical antennas, inverted Ls, and ground-fed slopers (all over good ground systems), b) 75-ohm hardline cable, c) a junction of two 50-ohm coaxial cables, d) shunt-fed towers performing as vertical antennas, and e) the output of a transceiver or class B linear amplifier when an unfavorable VSWR condition exists.

These three Ununs also play an important role in making other useful Baluns possible. Examples given in earlier chapters include: a) connecting a 1.5:1 Unun (50:75 ohms) in series with a 1:1 Balun (75:75 ohms) results in a broadband 1.5:1 Balun (50:75 ohms); b) connecting a 2:1 Unun (50:100 ohms) in series with a 1:1 Balun (100:100 ohms) results in a broadband 2:1 Balun (50:100 ohms); c) connecting a 1.5:1 Unun (50:75 ohms) in series with a 4:1 Balun (75:300 ohms) results in a broadband 6:1 Balun (50:300 ohms), and d) connecting a 1.33:1 Unun (50:66.7 ohms) in series with a 9:1 Balun (66.7:600 ohms) results in a broadband 12:1 Balun (50:600 ohms).

It has been shown[2] that a continuum of ratios can now be obtained with Ununs matching 50-ohm cable to impedances as low as 3.125 ohms and as high as 800 ohms. In addition, by using higher-order windings (trifilar, quadfilar, etc.), Ununs can be constructed with two broadband ratios like 1.5:1 and 3:1, or 2:1 and 4:1. Furthermore, by tapping some of the windings of these higher-order Ununs, multimatch transformers can be constructed with many broadband ratios. As a result of this class of fractional-ratio Ununs, a continuum of Ununs and Baluns is now available to match 50 ohms unbalanced to unbalanced or balanced impedances as low as 3.125 ohms and as high as 800 ohms.

My first attempt to obtain ratios less than 4:1 was made by tapping one of the wires in a Ruthroff 4:1 bifilar Unun. My experiment met with only moderate success.[2] An adequate low-frequency response with a 1.33:1 ratio was difficult to obtain. Also, the 2:1 ratio had considerably greater loss than higher or lower ratios. Recently, I found that higher-order windings (trifilar, quadrifilar, etc.), some with taps, provide much wider bandwidths and higher efficiencies. This

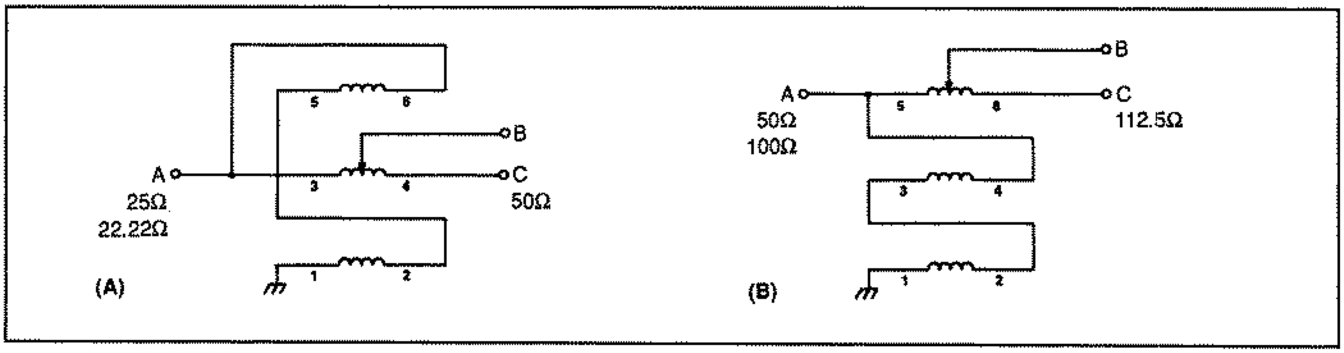

Figure 13-1. *Schematic diagrams: (A) matching 50 to 25 ohms (B-A) and 50 to 22.22 ohms (C-A); (B) matching 100 to 50 ohms (B-A) and 112.5 to 50 ohms (C-A).*

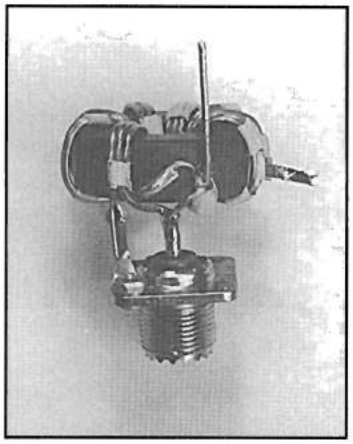

Photo 13-A. *Bottom view of the 2:1 Unun designed to match 50 ohms to 25 ohms or 22.2 ohms (Figure 13-1A). The connector is on the low-impedance side.*

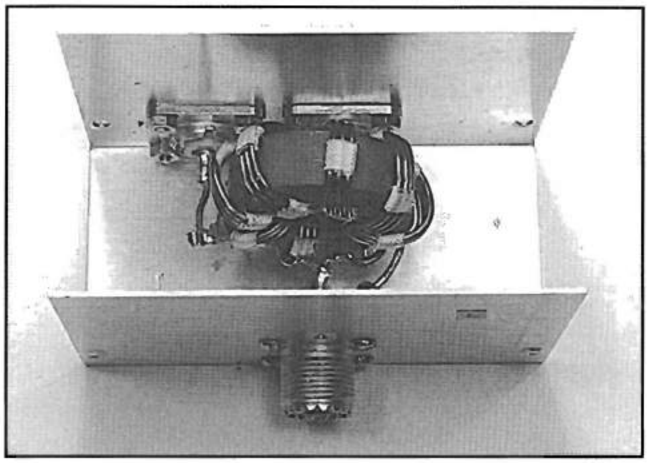

Photo 13-B. *The 2:1 Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

chapter describes fractional-ratio Ununs using these higher-order windings.

The next section discusses the practical aspects of the 2:1 Unun. This Unun is not only one of the more useful transformers, but it also serves as a good introduction to the trifilar and quadrifilar designs. What follows is an introduction to the most difficult fractional-ratio Unun—the quintufilar design, which results in very broadband 1.33:1 and 1.5:1 Ununs. A more complete discussion appears in **Part I** of this book.

This chapter closes with construction tips. As you will see, these Ununs can be difficult to construct.

Sec 13.2 2:1 Ununs

Figure 13-1A shows the schematic diagram of an Unun designed to match 50-ohm cable to an unbalanced load of 25 ohms (2:1 ratio with connections A-B) or 22.22 ohms (2.25:1 ratio with connections A-C). It has 6 trifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 3-4 is tapped at 5 turns from terminal 3. **Photo 13-A** is a photograph showing the various connections. The connector is on the low-impedance side. **Photo 13-B** shows the transformer mounted in a CU-3015A (4 inches long by 2 inches wide by 2.75 inches high) minibox. In matching 50 ohms to either 25 or 22.22 ohms, the transformation ratio is constant from 1 to 30 MHz.

Because the transmission lines are very short, this Unun does quite well as a step-up transformer. That is, when matching 50 ohms (on the left side) to 100 ohms (connections A-B) or 112.5 ohms (connections A-C) on the right side, the transformation ratio is constant from 1 to 15 MHz. Because of the extremely high efficiency of this transformer (98 to 99 percent under matched conditions), this small version can easily handle the full legal limit of amateur radio power.

Figure 13-1B shows the schematic diagram of an Unun designed to match 50-ohm cable to an unbalanced load of 100 ohms (2:1 ratio with connections A-B) or 112.5-ohms (2.25:1 ratio with connections A-C). It has 7 trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. The top winding 5-6 is No. 14 Thermaleze wire and is tapped at 6 turns from terminal 5. The other two windings are No. 16 H Thermaleze wire. **Photo 13-C** shows the various connections. The connector is on the low-impedance side. In matching 50-ohm cable to 100 ohms (A-B) or 112.5 ohms (A-C), the transformation ratio is constant from 1 to 30 MHz.

Again, because the transmission lines are very short, this Unun does quite well as a step-down transformer. In matching 50-ohm cable (on the right side) to 25 ohms (A-B) or 22.22 ohms (A-C), the transformation ratios are constant from 1 to 15 MHz. As above, this transformer can easily handle the full legal limit of amateur radio power.

Although the quadrifilar Unun shown in the schematic diagram in **Figure 13-2** and in **Photo 13-D** has an impedance transformation ratio of 1.78:1, it should also satisfy many of the 2:1 requirements. This Unun, which is designed to match 50-ohm cable to an unbalanced load of 28 ohms, not only has a very broadband response (1 MHz to over 50 MHz), but also offers other possible wideband ratios that will be covered in succeeding chapters.

Specifically, the Unun has 5 quadrifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and the other three are No. 16 H Thermaleze wire. Like the

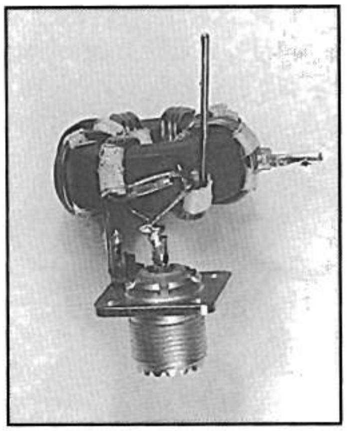

Photo 13-C. *Bottom view of the 2:1 Unun designed to match 50 ohms to 100 ohms or 112.5 ohms (Figure 13-1B). The connector is on the low-impedance side.*

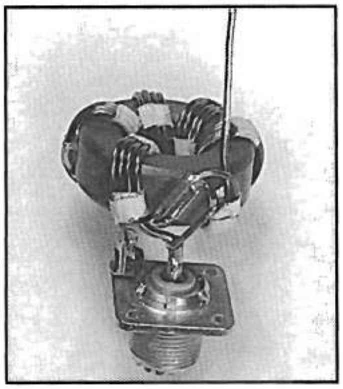

Photo 13-D. *Bottom view of the 1.78:1 Unun designed to match 50 to 28 ohms. The connector is on the low-impedance side.*

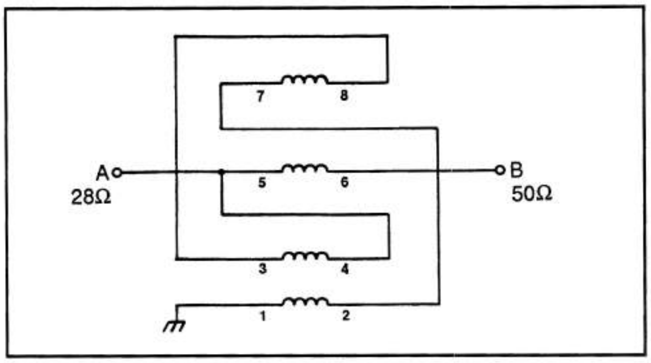

Figure 13-2. *Schematic diagram of the quadrifilar Unun designed to match 50 to 28 ohms (1.78:1 ratio).*

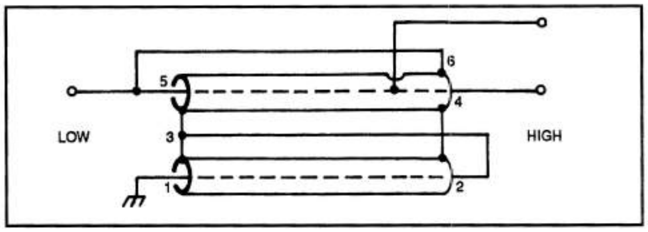

Figure 13-3. *Schematic diagram of a tapped-trifilar Unun that uses two sections of coaxial cable, yielding impedance ratios of 2:1 and 2.25:1.*

two 2:1 Ununs described above, this one also easily handles the full legal limit of amateur radio power.

As with most coiled Ununs that have little spacing between adjacent turns, current-crowding (between adjacent turns) can eventually limit the power-handling capability of these devices. It's possible to improve the ability for handling higher currents by using thicker wires, or by using coaxial cables where current-crowding is nonexistent. **Figure 13-3** shows the schematic diagram of a tapped-trifilar transformer that uses two sections of coaxial cable yielding impedance ratios of 2:1 and 2.25:1. **Photo 13-E** shows two trifilar toroidal transformers using low-impedance coaxial cables with their outer braids connected in parallel and acting as the third conductor. These transformers are conservatively rated at 5 kW of continuous power.

The smaller transformer in **Photo 13-E** has 7 trifilar turns of low-impedance coax on a 2-inch OD toroid with a permeability of 290. The No. 14 H Thermaleze wire inner conductors have four layers of Scotch No. 92 tape. The outer braids are made from small coaxial cables (or 1/8-inch tubular braid), and are also wrapped with Scotch No. 92 tape in order to preserve the low characteristic impedances. The inner conductor of the top coax in **Figure 13-3** is tapped at 6 turns from terminal 5. When matching 50 ohms to 22.22 ohms or 25 ohms, the impedance ratio is constant from 1 to over 50 MHz.

The larger transformer in **Photo 13-E** also has 7 trifilar turns, but on a 2.4-inch OD toroid with a permeability of 125. The inner conductors of No. 14 H Thermaleze wire now have a 15-mil wall of Teflon sleeving, yielding the low-impedance coaxial cable. The outer braids are the same. Because the ferrite permeability is lower and the lengths of the transmission lines are longer than those of the smaller unit, this transformer's bandwidth is not quite as good. When matching 50 to 22.22 ohms (this particular Unun doesn't have a tapped winding), the impedance ratio is constant from 1.7 to 30 MHz.

Although not shown, these very high-powered 2:1 Ununs can also be easily designed to match 50-ohm cable to unbalanced loads of 100 and 112.5 ohms. This is done by using small but high-powered coaxes like RG-303/U, RG-141/U, or RG-142/U.

Sec 13.3 1.5:1 Ununs

Figure 13-4 shows three basic forms of a quintufilar 1.56:1 Unun that should satisfy most of the 1.5:1 requirements. As can be seen, the only difference in

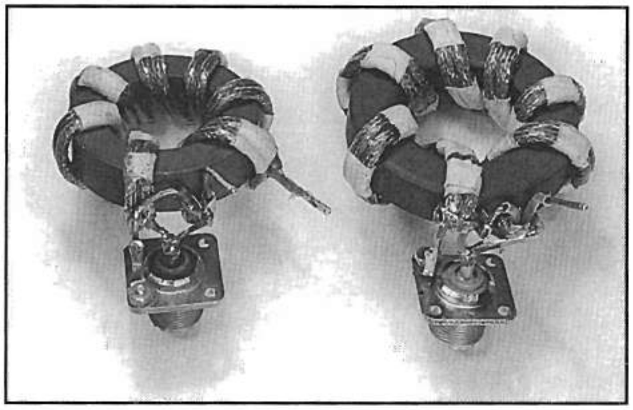

Photo 13-E. *Two toroidal transformers using coaxial cable and designed to match 50 to 22.22 ohms (2.25:1). The smaller transformer is also tapped, yielding a match of 50 to 25 ohms (2:1).*

the schematic diagrams is in the interleaving of the windings. This is done to optimize the performance of these Ununs at the various impedance levels. Schematic A is optimized for matching 50 to 75 ohms. Schematic B is optimized for matching 32 to 50 ohms. Schematic C, while optimized for matching 40 to 62 ohms, still yields quite broadband ratios at both 50:75 and 32:50-ohm levels. It should be a useful, general-purpose Unun.

Photo 13-F is a photograph of the bottom views (showing the connections) of the three different designs. They appear in the same order as the schematics of **Figure 13-4**; i.e., a) the Unun on the left is designed to match 50 to 75 ohms, b) the Unun in the center is design to match 32 to 50 ohms, and c) the Unun on the right is designed to work quite well at both impedance levels. The SO-239 connectors are all on the low impedance side of the Ununs.

All three transformers have four quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Their differences are:

1. **50:75 ohms** (on the left on **Figure 13-4** and **Photo 13-F**).

Winding 9-10 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. When matching 50 to 75 ohms (actually to 78 ohms), the transformation ratio is constant from 1 to over 30 MHz. In matching 50 ohms (on the right side in **Figure 13-3A**) to 32 ohms, it is still constant from 1 to 15 MHz.

2. **32:50 ohms** (in the center in **Figure 13-4** and **Photo 13-F**).

Winding 5-6 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. When matching 32 to 50 ohms, the transformation ratio is constant from 1 to over 30 MHz. In matching 75 ohms (on the right side in **Figure 13-3B**) to 50 ohms, it is still constant from 1 to 15 MHz.

3. **50:75 ohms; 32:50 ohms** (on the right in **Figure 13-4** and **Photo 13-F**).

Winding 7-8 is No. 14 H Thermaleze wire. The other four windings are No. 16 H Thermaleze wire. In matching 32 to 50 ohms, the transformation ratio is constant from 1 to 30 MHz. In matching 75 ohms (on the right side in **Figure 13-3C**) to 50 ohms, it is still constant to 21 MHz. This is quite a good general-purpose design.

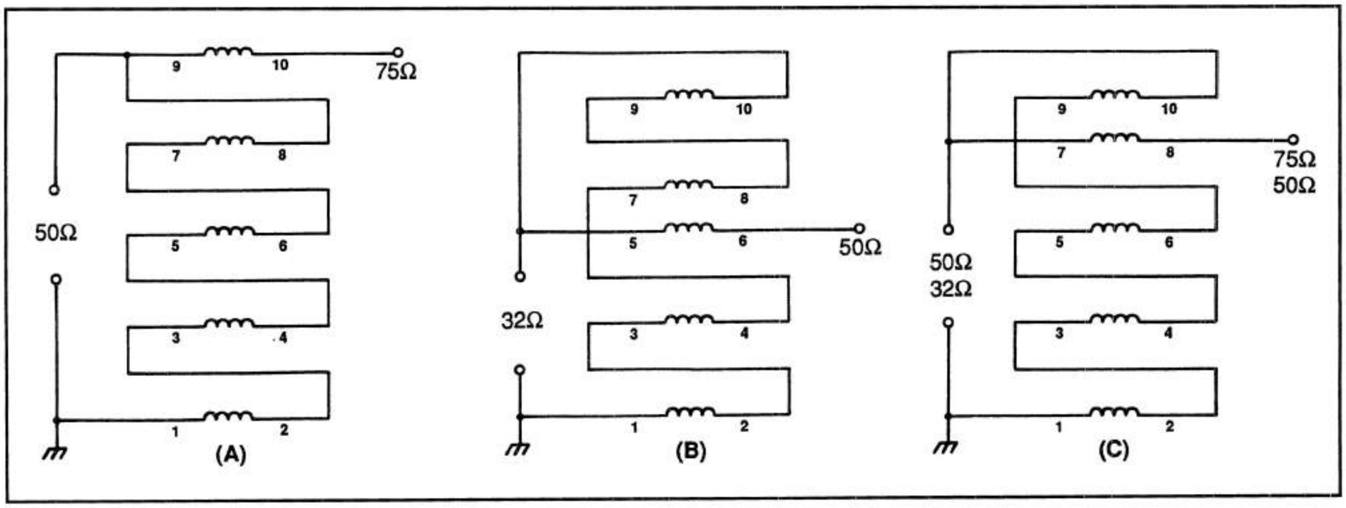

Figure 13-4. *Three basic forms of a quintufilar 1.56:1 Unun: (A) optimized to match 50 to 75 ohms, (B) optimized to match 32 to 50 ohms, and (C) optimized to match 40 to 62 ohms, resulting in a good general-purpose design.*

Even though a small toroid (with only a 1.5-inch OD) is used, these Ununs are still very sturdy transformers. Because their efficiencies are so high (98 to 99 percent), they can easily handle the full legal limit of amateur radio power.[2] Furthermore, the windings carrying the majority of the current (80 percent) are all No. 14 wire. Only when well-designed Ununs are subjected to very high VSWRs will excessive heating occur. Ununs (and Baluns) should never be exposed to these severe conditions.

Sec 13.4 A 1.33:1 Unun

The circuit shown in **Figure 13-5** evolved after many attempts were made at obtaining a broadband match of 50 to 66.7 ohms (1.33:1). **Photo 13-G** shows the bottom view of an actual design. The SO-239 connector is on the low impedance side. **Photo 13-H** shows the Unun mounted in a CU-3015A minibox.

Specifically, this Unun has five quintufilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is No. 14 H Thermaleze wire and is tapped at three turns from terminal 5 (**Figure 13-4**). It is also covered with one layer of Scotch No. 92 polyimide tape, optimizing the performance at the 50:66.7-ohm level. The other four windings are No. 16 H Thermaleze wire.

In matching 50 to 66.7 ohms (A-B), the transformation ratio is practically constant from 1 to 30 MHz. The ratio only decreases by 3 percent across the band. In matching 50 to 32 ohms (C-A), the transformation ratio is constant from 1 to 30 MHz. In matching 75 to 50 ohms (C-A), the ratio is constant from 1 to 15 MHz. In matching 50 to 37.6 ohms (B-A), the ratio is constant from 1 to 15 MHz. As you can see, the Unun has some useful broadband multimatches.

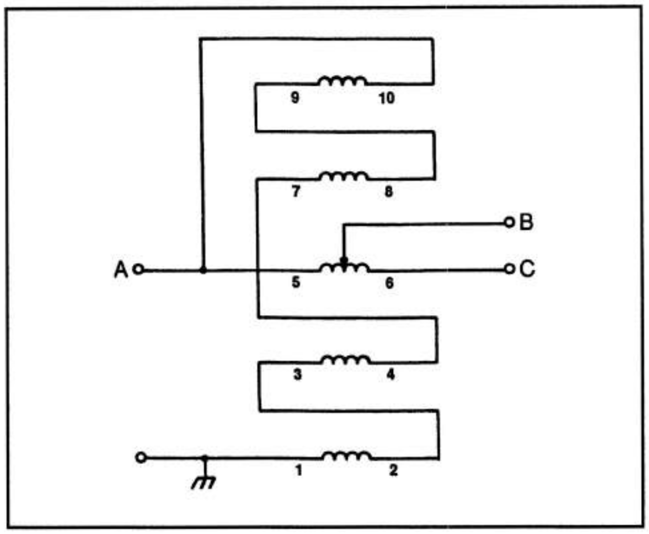

Figure 13-5. *Schematic diagram of a quintufilar Unun specifically designed to yield a broadband 1.33:1 ratio (66.7:50 ohms; connection B-A). Connection C-A also yields a broadband 1.56:1 ratio (50:32 ohms).*

As in the cases of the other three Ununs, this tapped-Unun also easily handles the full legal limit of amateur radio power. Like the 2:1 Unun, for higher power capabilities, thicker wires or a three-coax quintufilar design can be used.[2]

Sec 13.5 Construction Tips

Most of my Unun designs use the bootstrap connection that sums direct voltages (on the high-impedance side) with a delayed voltage, which traverses a single transmission line.[2] Therefore, in order to achieve the very wideband responses, small toroids (which allow the shortest transmission lines) are used. The small 1.5-inch OD toroids offer this advantage. Furthermore,

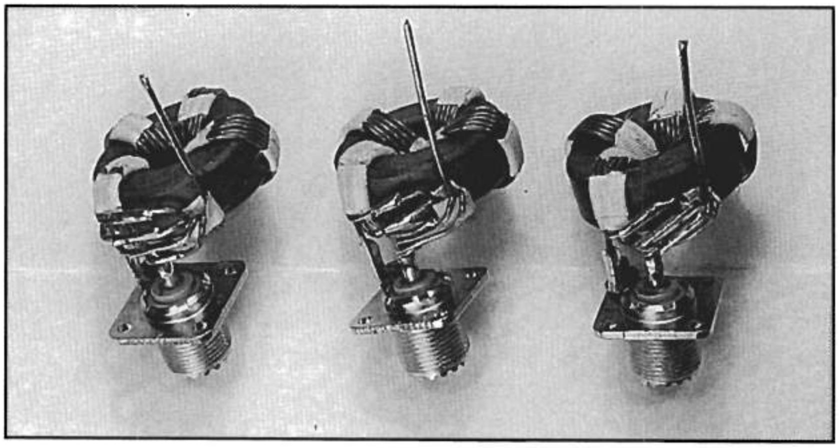

Photo 13-F. *Photograph of the three different Unun designs shown in Figure 13-4: A) on the left, 50:75 ohms, B) in the center 32:50 ohms, and C) on the right, a general purpose design matching both impedance levels quite well.*

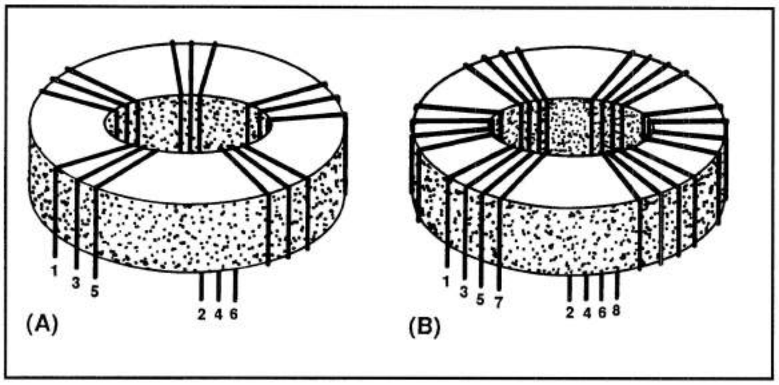

Figure 13-6. *Pictorials of higher-order windings: (A) trifilar and (B) quadrifilar.*

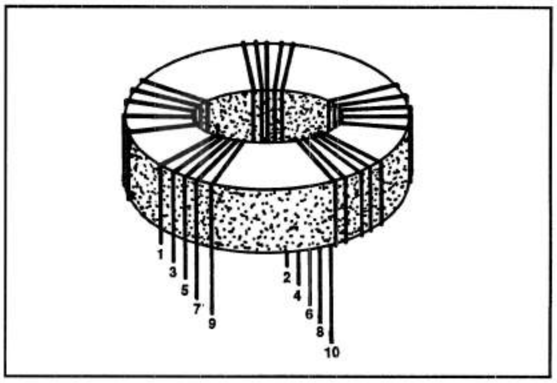

Figure 13-7. *Pictorial of a quintufilar winding.*

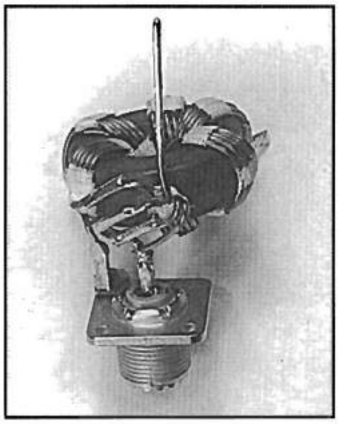

Photo 13-G. *Photograph of the 1.33:1 Unun (66.7:50 ohms). This transformer also has a broadband 1.56:1 ratio (50:32 ohms).*

quintufilar (and higher-order windings) Ununs are also eventually limited in their high-frequency responses by self-resonances. Shorter winding lengths keep these self-resonances well out of the HF band. For the Ununs in this chapter, they occur between 45 and 65 MHz, with **Figure 13-4A** having the higher value.

There is also a mechanical advantage in using the smaller toroids. You'll find that the popular CU-3015A minibox makes an excellent enclosure for the 1.5-inch OD toroid. Furthermore, because ferrite is a ceramic and, therefore, unaffected by moisture, no special precautions need be taken for out-of-doors use. Potting the transformer in plastic is unnecessary. One must only keep the Unun out of a pool of water.

Because well-designed transformers have virtually no flux in the core, their power ratings are mainly determined by the ability of the transmission lines to handle the voltages and currents. Furthermore, it can be shown that the losses in these transformers are related to the voltage gradients along the transmission lines.[2] Thus, they are dielectric-type ferrite losses. This means that the efficiency can be severely degraded with very high VSWRs since higher voltage gradients occur under these conditions.

Several suggestions can be made regarding the construction of Ununs using these higher-order windings (trifilar, quadrifilar, etc.). They are:

1) Make a ribbon out of the wires and wind them all at the same time. This keeps the wires as close as possible, resulting in the maintenance of the optimum characteristic impedance of the transmission lines. I found that strips made with 1.25- by 0.375- to 0.55-inch glass tape (Scotch No. 27), clamped about every 1/2 inch, hold the wires in place very well. The starting lengths of the wires should be about 5 inches longer than one would calculate knowing the number of turns and the length around each turn.

2) Because work-hardening of the copper wire takes place in coiling it around a toroid, a pair of pliers and

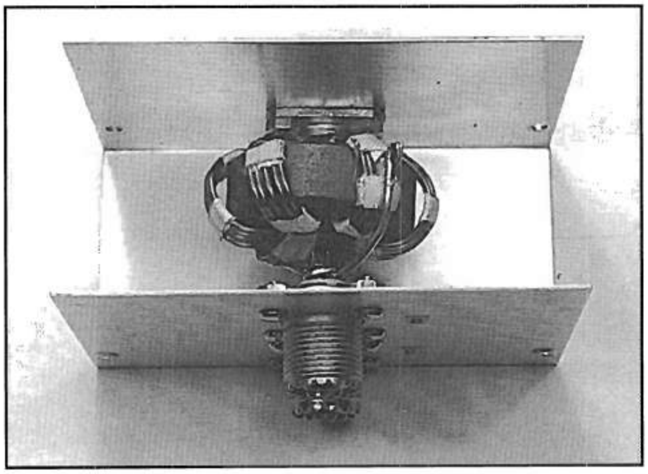

Photo 13-H. *The 1.33:1 Unun mounted in a CU-3015A minibox.*

a strong thumb (and arms) are indispensable tools. It takes considerable effort to wind these transformers. Also, because these designs have adequate margins at their low-frequency ends, some space between the windings and the toroids can be tolerated.

3) It is helpful to recognize the various patterns that appear at the ends of the windings. **Figure 13-6** shows a drawing of the trifilar and quadrifilar patterns and **Figure 13-7** shows the quintufilar pattern. Note that terminal 1 and terminal 6 or terminal 8 or 10 are the outside terminals of the patterns. Also, note that terminal 1 is always grounded in the schematic diagrams.

4) Tapping windings can be one of the more difficult tasks in constructing these transformers. Winding a tapped transformer is also more difficult. I found that the edge of a small, fine file does the best job in removing the insulation. About 1/8 to 1/4 inch is removed around the wire. It also helps to remove some of the copper. Then a flat 1/8-inch copper strip or No. 14 wire flattened on one end is soldered to the bare wire. The soldered connection is then rendered smooth using the edge of the file. Finally, two pieces of Scotch No. 92 polyimide tape are placed on the joint to insulate it from the neighboring turns. I also found that a tap placed one turn from the end of the winding is best made approximately 4 inches from the end of the wire (when the wire is straight).

Finally, a comment should be made about low-power Ununs. Practically all of the transformers in this chapter can be easily designed for low-power use. Designs capable of handing the full output of HF transceivers can be readily constructed. Cores with an OD of 1.25 inches are recommended. The same number of turns, but with one size smaller wire, is also recommended. Because smaller cores and thinner wires are used, these lower-power units are not only easier to construct, but they also have wider bandwidths due to the shorter lengths of the windings.

Chapter 14

Dual-Ratio Ununs

Sec 14.1 Introduction

Earlier chapters in this book on Ununs have basically covered single-ratio transformers. Broadband ratios of 1.33:1, 1.5:1, 2:1, and 4:1 were the design objectives. The 1.33:1 ratio was obtained by tapping a 1.5:1 (actually 1.56:1) quintufilar-wound Unun. The 2:1 ratio was obtained by tapping a 2.25:1 trifilar-wound Unun. Although these transformers can be considered to have two broadband ratios (1.33:1 and 1.56:1 or 2:1 and 2.25:1), their two ratios were not different enough for many practical applications. This is especially true of antennas where the input impedance varies with frequency.

My earlier work[2] and an article by Genaille[10] have shown that a host of ratios (less than 4:1) can be obtained by tapping the bifilar winding of a Ruthroff 4:1 Unun.[9] However, the bandwidths obtained using this technique are quite limited with each ratio and are highly dependent upon the impedance level. This is particularly true when a rod core is used because it requires more turns (resulting in longer transmission lines) in order to obtain the necessary choking reactance that isolates the input from the output. Furthermore, ratios around 2:1 exhibit more loss because autotransformer action also enters into the matching process.[2]

By using quadrifilar and quintufilar windings on small 1.5-inch OD cores, and connecting them in such a way that the characteristic impedances of the windings are near optimum, two very different and broadband ratios matching 50 ohms to lower impedances are obtained. Furthermore, because the transmission lines in these transformers are so very short (8 to 9 inches in length), these transformers do quite well in matching 50 ohms to higher impedances (as step-up transformers).

This chapter presents two Ununs which have two broadband ratios that differ by a factor of two! One has a 1.5:1 and a 3:1 ratio (actually 1.56:1 and 2.78:1). The other has a 2:1 (actually 1.78:1) and a 4:1 ratio. Also, this chapter introduces the novel tech-

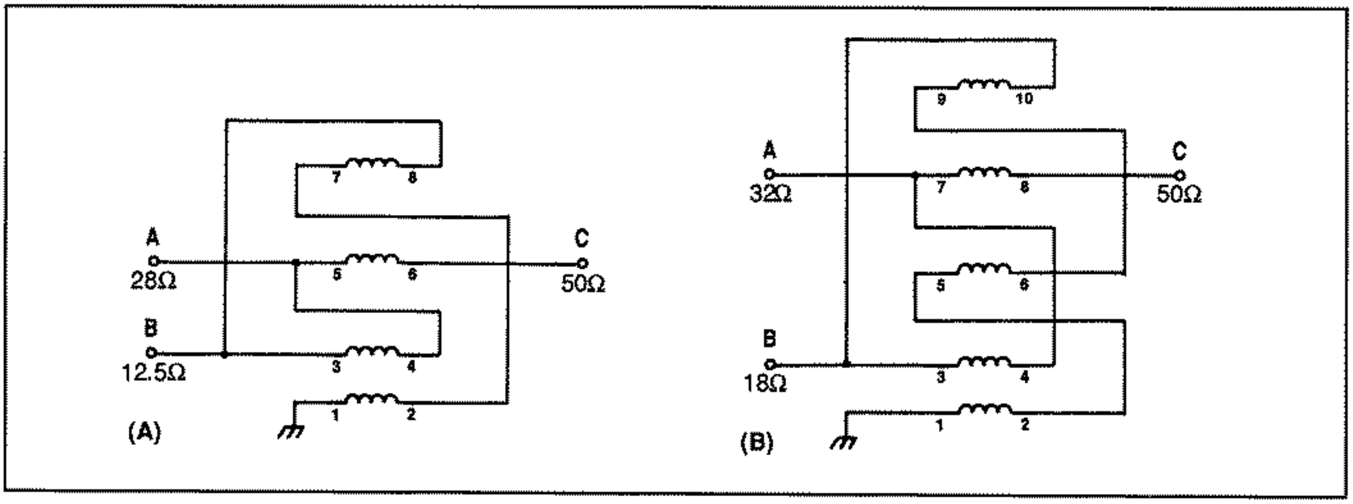

Figure 14-1. *Schematic diagrams of dual-ratio Ununs: (A) 1.78:1 connection C-A, 4:1 connection C-B; (B) 1.56:1 connection C-A, 2.78:1 connection C-B.*

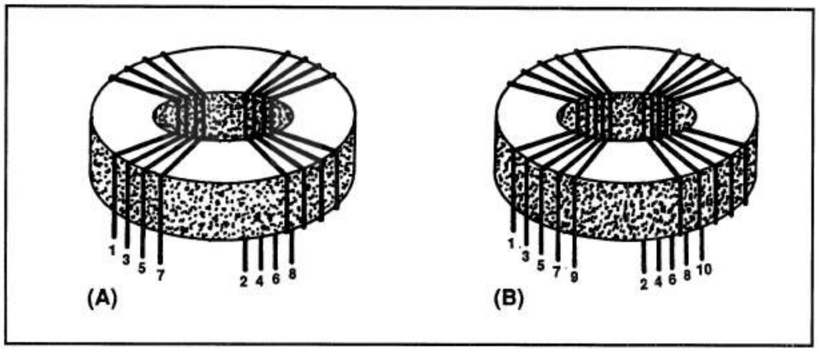

Figure 14-2. *Pictorials of higher-order windings: (A) quadrifilar, (B) quintufilar.*

nique of connecting these two transformers in parallel on their 50-ohm sides, resulting in four very broadband ratios.

Sec 14.2 2:1 and 4:1 Ratios

Figure 14-1A shows the schematic diagram of the quadrifilar Unun yielding ratios of 2:1 (actually 1.78:1) and 4:1. **Figure 14-2A** provides a pictorial of its windings. On the left in **Photo 14-A**, you see the bottom view of an Unun. The cable connector is on the 28-ohm (1.78:1) side. This Unun has four quadrifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD toroid with a permeability of 250.

In matching 50 to 28 ohms (connection C-A), the response is flat to within 1 percent from 1 to 30 MHz. From 1 to 50 MHz, it's flat to within 2 percent. In matching 50 to 12.5 ohms (connection C-B), the response is flat to within 3 percent from 1 to 30 MHz.

Because very short transmission lines are used, this transformer performs quite well as a step-up transformer. In matching 50 to 200 ohms (connection B-C), the response is flat to within 3 percent from 1.5 to 10 MHz. In matching 50 to 89 ohms (connection A-C), the response is flat to within 5 percent from 1.5 to 30 MHz!

Sec 14.2.1 Construction Tips

Start with about 14 inches of straightened wire. Form the wires into a ribbon with clamps of Scotch No. 27 glass tape every 1/2 inch. I found that strips 3/16 inch wide and about 1.5 inches long do a good job. The clamps should be long enough to go around the wires twice. After winding, connect terminals 2 and 7. Then connect terminals 3 and 8. Finally, connect terminals 4 and 5. Because work-hardening takes place quickly, you will find that a pair of pliers and a strong thumb (and arms) are necessary tools. You will also find that winding these exceptionally performing transformers is not easy. As in all endeavors, practice really pays off.

Sec 14.3 1.5:1 and 3:1 Ratios

Figure 14-1B shows the schematic diagram of the quintufilar Unun yielding ratios of 1.5:1 and 3:1

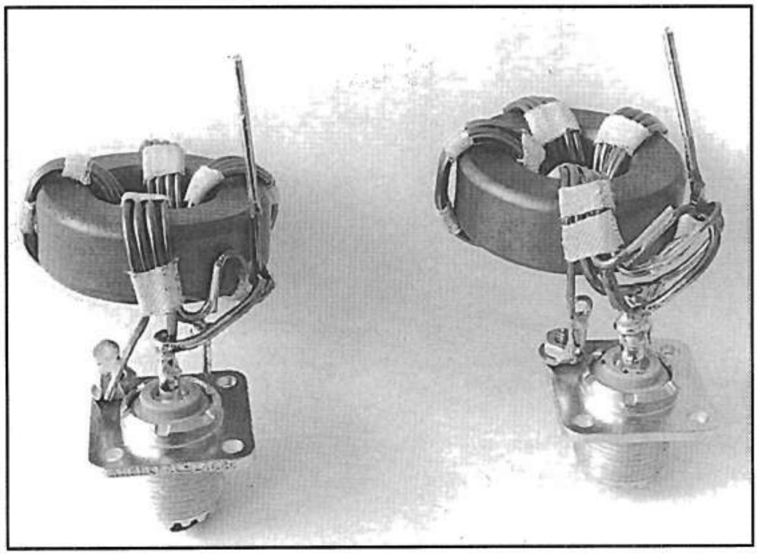

Photo 14-A. *On the left, a quadrifilar Unun with ratios of 1.78:1 and 4:1; on the right, a quintufilar Unun with ratios of 1.56:1 and 2.78:1.*

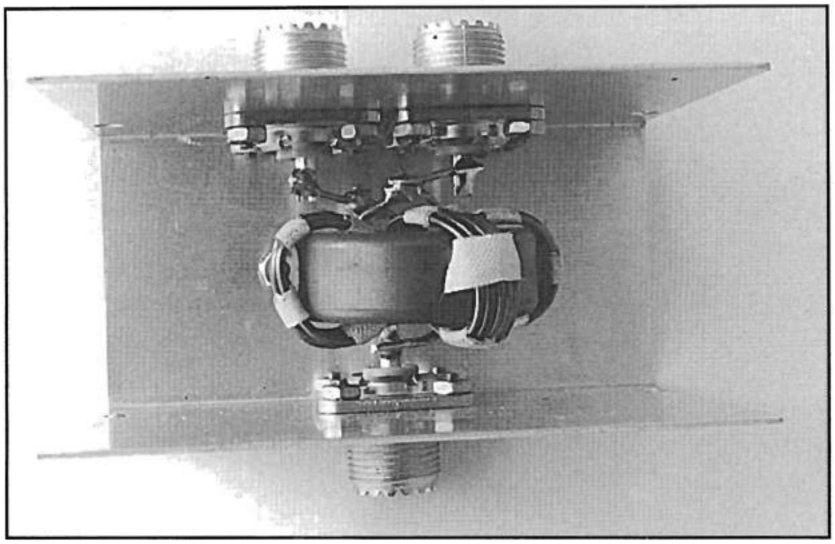

Photo 14-B. *The dual-ratio quintufilar Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

(actually 1.56:1 and 2.78:1). **Figure 14-2B** is a pictorial of its windings. On the right side in **Photo 14-A** you see a bottom view of an Unun. The cable connector is on the 32-ohm (1.56:1) side. This Unun has four quintufilar turns on a 1.5-inch OD toroid with a permeability of 250. Windings 3-4 and 7-8 are No. 14 H Thermaleze wire. The other three are No. 16 H Thermaleze wire. Winding 7-8 also has two layers of Scotch No. 92 polyimide tape, which optimizes the 1.56:1 ratio.

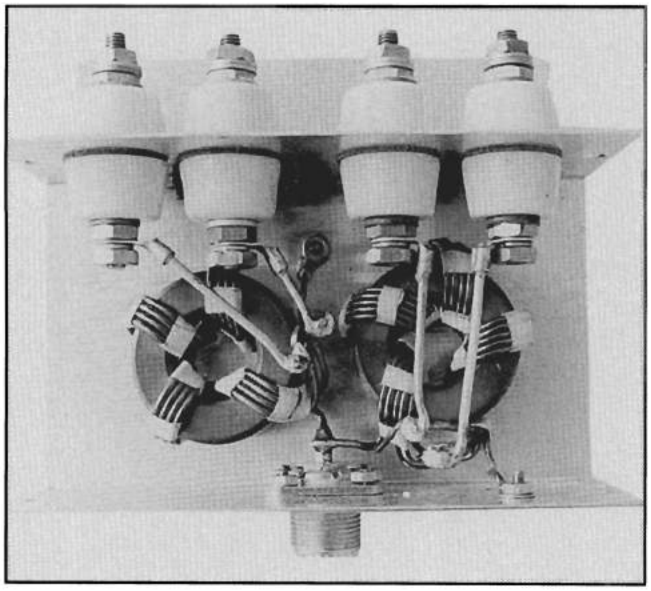

Photo 14-C. *The two dual-ratio Ununs connected in parallel on their 50-ohm sides providing four broadband ratios close to 1.5:1, 2:1, 3:1, and 4:1. The quadrifilar unit is on the left and the quintufilar unit is on the right. The enclosure is a 5 inch long by 3 inch wide by 2 inch high minibox.*

In matching 50 to 32 ohms (connection C-A), the response is essentially flat (less than 1 percent variation) from 1 MHz to over 40 MHz. Without the two layers of Scotch No. 92 tape, the response varies by 4 percent from 1 to 30 MHz. When used as a step-up transformer matching 50 to 78 ohms (connection A-C), and with the two layers of Scotch No. 92 tape on winding 7-8, the response is flat to within 5 percent from 1 to 15 MHz. Without the extra insulation on winding 7-8, the response is flat to within 5 percent from 1 to 7.5 MHz.

In matching 50 to 18 ohms (connection C-B), the variation in response is less than 3 percent from 1 to 40 MHz. The response is the same whether winding 7-8 is covered with the extra insulation or not. As a step-up transformer matching 50 to 139 ohms (connection B-C), the response is flat to within 3 percent from 1 to 10 MHz (with or without the extra insulation on winding 7-8). **Photo 14-B** shows this Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox. The two cable connectors on the low impedance side could be replaced with feedthrough insulators for antenna use.

Sec 14.3.1 Construction Tips

Prepare the ribbon as was described for the quadrifilar Unun. If you choose to use the two extra layers of Scotch No. 92 tape on winding 7-8, make sure this winding is on the outside position of the ribbon (refer to **Figure 14-2B**). I found the best order in which to connect the wires is as follows: first, connect terminal 2 to 5; second, connect terminal 6 to 9; third, connect terminal 3 to 10; and, finally, connect terminal 4 to 7.

As was mentioned before, because a small toroid is used in order to achieve the best response (due to shorter transmission lines), this transformer also requires considerable strength and patience in the winding process.

Sec 14.4 Parallel Transformers

One of the most pleasant surprises I received with these efficient and broadband transformers was to find that they can be connected in parallel on their 50-ohm sides and still possess the same performance levels. Because the loading effect of one transformer on the other is minimal (like a short length of transmission line), the transformer that is properly terminated takes the power, while the other one is transparent. This lets you obtain four very wideband ratios with the two dual-ratio Ununs described in this chapter. Obviously, this technique eliminates one transmission line. Furthermore, Baluns can also be connected this way for feeding, with a single coaxial cable, beams and dipoles with different resonant impedances.

Photo 14-C shows two transformers connected in parallel on their 50-ohm sides using the two dual-ratio Ununs described in this chapter. As was mentioned, it now yields four wideband ratios very close to 1.5:1, 2:1, 3:1, and 4:1. I have used this matching network to feed a host of ground-fed antennas (over a good ground system). In one case, I had a 10-, 15-, and 20-meter trap vertical, slopers for 40 and 160 meters, a 12-meter vertical, and an inverted L for 80 meters all matched to a single coaxial cable at the same time. It was a simple matter of connecting each antenna to the output terminal that presented the best match (lowest VSWR). This technique is actually an extension of connecting dipoles for different bands, in parallel. The antenna that presents the correct impedance takes the power, and the others are essentially transparent. In many cases, I found that only one transformer with two broadband ratios performed adequately.

Chapter 15

Multimatch Ununs

Sec 15.1 Introduction

Broadband multimatch Ununs capable of high-power applications have been the goal of many designers over the years. Some have resorted to using conventional autotransformers with tapped windings to obtain the many impedance transformation ratios. However, these attempts met with little success because of the device's limited bandwidths and efficiencies. Others (including myself)[2] have tried tapping a bifilar Ruthroff Unun.[9] Although these designs yielded the high efficiencies of transmission line transformers, they had limited bandwidths. Furthermore, their best bandwidths (for the various ratios) occurred at odd impedance levels. In other words, they didn't meet the objective of broadband operation with one of the input or output ports at 50 ohms.

Chapter 14 presented two Ununs which had two broadband ratios that differed by a factor of two. One had a 1.5:1 and a 3:1 ratio (actually 1.56:1 and 2.78:1). The other had a 2:1 (actually 1.78:1) and a 4:1 ratio. This chapter describes two multimatch designs that are capable of many more broadband ratios. For the most part, both are capable of broadband operation from 1.7 to 30 MHz.

One Unun has the following five ratios (which are close to): 1.5:1, 2:1, 4:1, 6:1, and 9:1. Because the two lower ratios work well in either direction (that is stepping up or down from 50 ohms), this design can match 50-ohm cable to impedances as high as 100 ohms (actually 112.5 ohms) and as low as 5.6 ohms over the frequency range. As a result, it has seven usable applications. Furthermore, because this is a transmission line transformer that cancels out the flux in the core, losses (in a matched condition) of only 0.04 to 0.08 dB can be expected.

The novelty in this design lies in the use of a trifilar winding (with one winding tapped) on a very small ferrite toroid, resulting in the shortest possible lengths of transmission lines. The windings are also connected in such a manner as to optimize their characteristic impedances from an overall standpoint.

I have used the adjective *ultimate* to describe the second Unun design. Although it might be risky business, I assume that this design will meet one of the most common definitions for this adjective—namely, *beyond which it is impossible to go.* For many of us, the classic use of this adjective was made by Lew McCoy in describing his popular transmatch.[22] Although there have been some improvements to Lew's design, his use of this definite (and strong) adjective can be said to have withstood the test of time. I hope my use meets with similar success.

While the tapped-trifilar design provides five broadband ratios and seven practical applications, the *ultimate* design presented in this chapter goes well

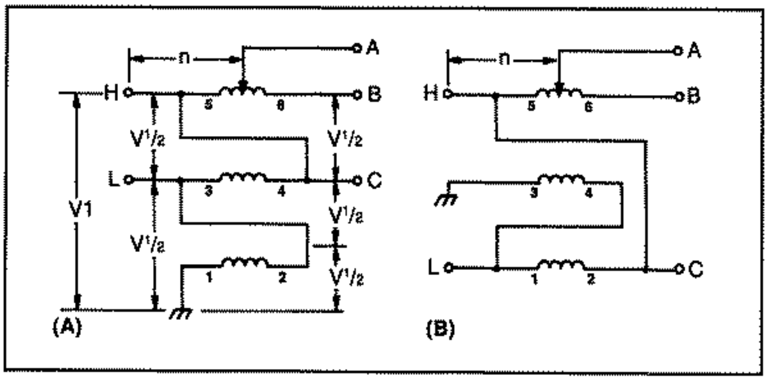

Figure 15-1. *Circuit diagrams for the 5-ratio Unun: (A) diagram for analysis; (B) transposed windings for best overall performance.*

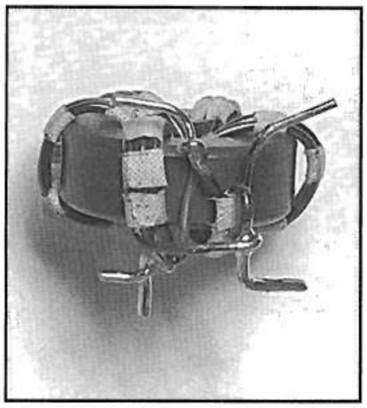

Photo 15-A. *Bottom view of the 5-ratio Unun of Figure 15-1B. The upper-left lead is terminal C. The upper-right lead is terminal B. The lower-left lead is terminal H. The lead pointing straight down is grounded (terminal 3). The lower-right lead is terminal L.*

beyond this number. It uses a tapped-quadrifilar design that yields the following 10 broadband ratios: 1.36:1, 1.56:1, 1.78:1, 2.25:1, 3.06:1, 4:1, 6.25:1, 9:1, 12.25:1, and 16:1. Because the four lower ratios also work quite well in either direction, this design offers *fourteen* applications in matching 50-ohm cable to impedances as high as 112.5 ohms and as low as 3.125 ohms. It also has the advantage of using a small low-loss toroidal core. Additionally, the windings are also interleaved in a pattern that optimizes their characteristic impedance.

However, this achievement comes at a price—*difficulty*. The 5-ratio Unun, which uses a trifilar winding, is considerably easier to wind. In addition, the quadrifilar 10-ratio Unun has two of its windings tapped, while the 5-ratio Unun has only one (see **Chapter 13** on tapping windings). If you have had little experience in winding Ununs or Baluns, attempt simplified versions of these two multimatch Ununs first. These versions eliminate the tapping of the windings. For the trifilar Unun, the remaining ratios would be: 2.25:1, 4:1, and 9:1. For the quadrifilar Unun, they would be: 1.78:1, 2.25:1, 4:1, 9:1, and 16:1.

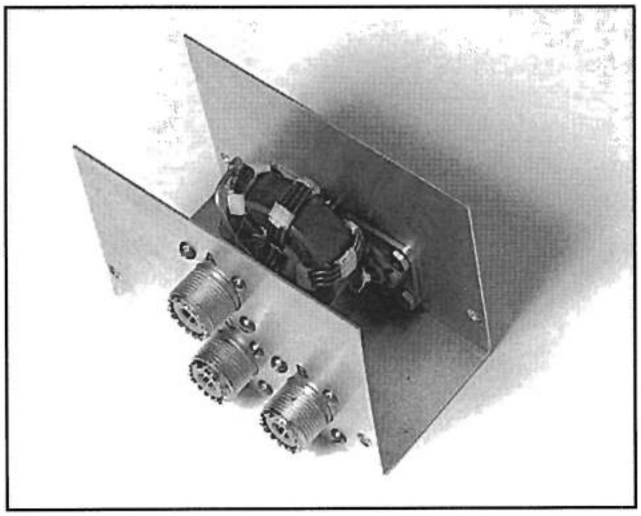

Photo 15-B. *The high-power unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

For those interested in the design considerations of these broadband multimatch transformers, a brief review is presented in each section. These sections are followed by others describing high-power designs capable of handling the full legal limit of amateur radio power. Finally, the remaining sections present low-power designs capable of handling the output of any HF transceiver. Because transmission line transformers can be made so efficient in matching 50 to 100 ohms or less, their small sizes will surprise many readers. Therefore, the combination of using small ferrite toroids with the maximum allowable permeability (less than 300) for high efficiency,[2] and with sufficient turns to meet the low-frequency objective, results in the excellent performance exhibited by the designs in this chapter.

Sec 15.2 The 5-Ratio Unun

Let's first look at **Figure 15-1A** because it is the easiest form of the trifilar-wound Unun to explain. For example, if the input voltage to ground, V_1, is connected to terminal H, the output terminal B, has a voltage to ground of $3/2V_1$. This results in a transformation ratio, g, of $(3/2)^2$ or 2.25:1. This should satisfy most 2:1 requirements. If the output is at terminal A to ground, then the output voltage is:

$$\begin{aligned} V_o &= V_1 + V_1(n/2N) \\ &= V_1(1 + n/2N) \qquad \text{(Eq 15-1)} \end{aligned}$$

where:
N = the total number of turns on the winding
n = the number of turns from terminal 5.

The transformation ratio, g, then becomes:

$$\begin{aligned} g &= (V_o/V_1)^2 \\ &= (1 + n/2N)^2 \qquad \text{(Eq 15-2)} \end{aligned}$$

If the input voltage to ground, V_1, is connected to terminal L, then terminal C has twice the voltage of V_1—resulting in a 4:1 ratio. Terminal B has three times the voltage resulting in a 9:1 ratio. With terminal A, the output voltage is:

$$V_o = 2V_1 + V_1(n/N)$$
$$= V_1(2 + n/N) \quad \text{(Eq 15-3)}$$

The transformation ratio, g, then becomes:

$$g = (2 + n/N)^2 \quad \text{(Eq 15-4)}$$

Sec 15.2.1 A High-power 5-Ratio Unun

After several attempts at rearranging the windings of **Figure 15-1A** for the best overall performance (optimizing the effective characteristic impedances of the windings), **Figure 15-1B** evolved. **Photo 15-A** shows the bottom view of an Unun, using the circuit of **Figure 15-1B**, capable of handling the full legal limit of amateur radio power. **Photo 15-B** shows the unit mounted in a CU-3015A minibox. It has five trifilar turns on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is tapped at two turns (n = 2) from terminal 5.

If the 9:1 ratio matching 50 to 5.6 ohms (connection B-L) is to be used at full power, then winding 3-4 should be No. 12 H Thermaleze wire. If not, then all windings can be No. 14 H Thermaleze wire.

A listing of the expected performance across the band from 1.7 MHz to 30 MHz, with the various ratios, is as follows:

9:1 (B-L); 50:5.6 ohms
Ratio is within 1 percent!
5.75:1 (A-L); 50:8.7 ohms
Ratio decreases by 5 percent.
4:1 (C-L); 50:12.5 ohms
Ratio increases by 15 percent (the greatest deviation of all the ratios).
2.25:1
a) (B-H); 50:22.22 ohms
Ratio decreases by 4 percent.
b) (H-B); 50:112.5 ohms
Ratio increases by 8 percent.
1.44:1
a) (A-H); 50:35 ohms
Ratio decreases by 10 percent.
b) (H-A); 50:72 ohms
Ratio increases by 2 percent.

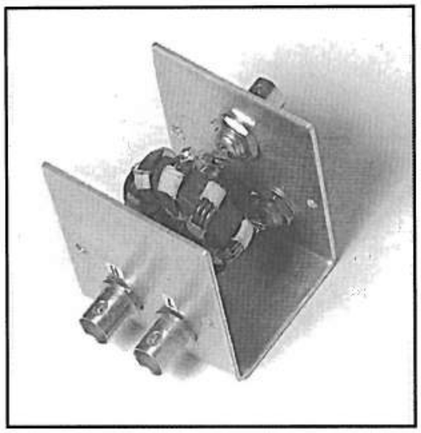

Photo 15-C. *The low-power unit mounted in a homemade 2 inch long by 1.5 inch wide by 2.25 inch high minibox.*

Several comments should be made regarding the expected results shown above. First of all, the greatest deviation from a flat response at any ratio occurs when matching 50 to 12.5 ohms (connection C-L; a 4:1 ratio). If an accurate insertion loss measurement was made at this ratio and impedance level, the result would show an insignificant difference across the band. Secondly, the major part of the deviations for all ratios occurs beyond 15 MHz (the effect of standing waves). Finally, the higher ratios should *never* be used to match 50 ohms to 450 ohms, 288 ohms, and 200 ohms, respectively. The characteristic impedances and choking reactances *do not* allow for broadband operation under these conditions.

Sec 15.2.2 A Low-power 5-Ratio Unun

Photo 15-C shows a low-power unit mounted in a homemade 2 inch long by 1.5 inch wide by 2.25 inch

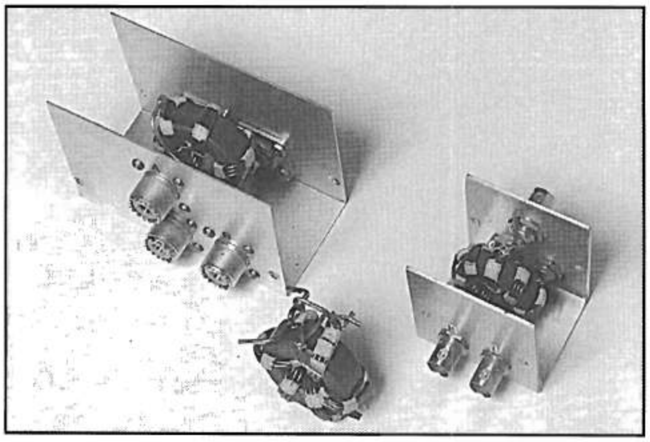

Photo 15-D. *The three 5-ratio Ununs together. From left to right, the high-power unit mounted and unmounted, the low-power unit.*

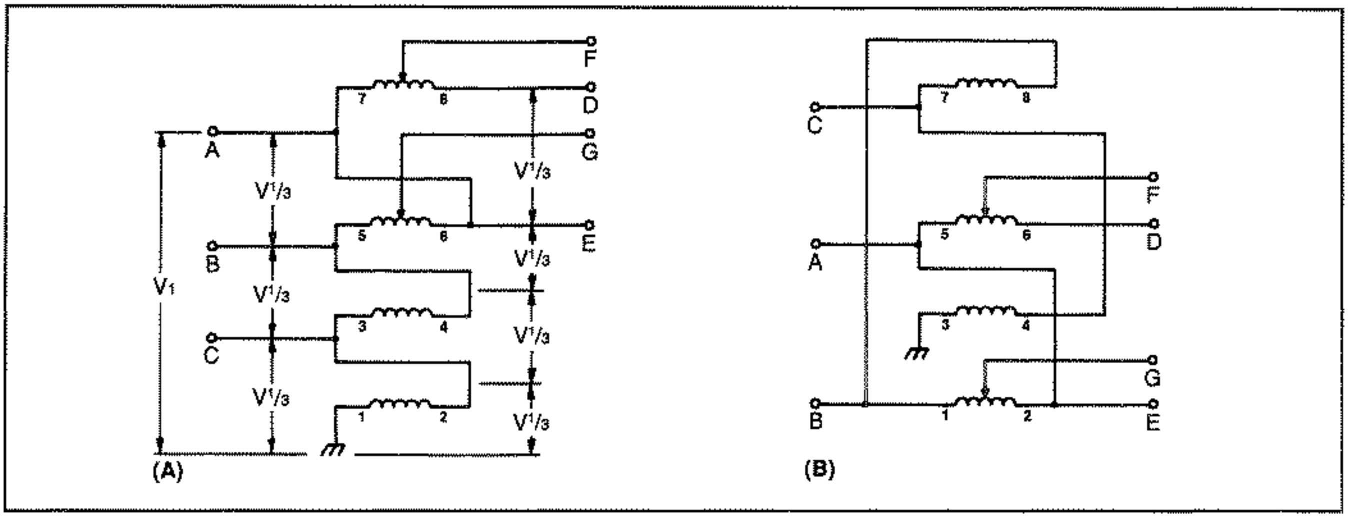

Figure 15-2. *Circuit diagrams for the 10-ratio Unun: (A) diagram for analysis; (B) transposed windings for best overall performance.*

high minibox. It has six trifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. The tap on winding 5-6 is located three turns from terminal 5, yielding ratios of 6.25:1 and 1.56:1 instead of the 5.75:1 and 1.44:1 ratios of the high-power unit. In actual use, these differences should be negligible.

Because this Unun has shorter transmission lines than its high-power counterpart, the deviations of the ratios across the band are even smaller. It is also interesting to note that, if No. 14 H Thermaleze wire was used in winding 3-4, this very small Unun could very well be rated at 500 watts of continuous power!

Photo 15-D shows all three 5-ratio Ununs together.

Sec 15.3 The 10-Ratio Unun—The Ultimate Multimatch

Figure 15-2A is presented here because it is the easiest form of a quadrifilar-wound Unun to explain. With the input voltage, V_1, connected to the various terminals on the left (the low-impedance side), and with very short transmission lines compared to the wavelength, we have the following transformation ratios:

1. V_1 connected to terminal A

a) At terminal D the output voltage V_o is $4/3V_1$. Therefore, the transformation ratio, g, with connection A-D is:

$$g = (4/3)^2 = 1:1.78 \qquad \text{(Eq 15-5)}$$

b) At terminal F the output voltage is:

$$V_o = V_1(1 + n/3N) \qquad \text{(Eq 15-6)}$$

where:
N = total numher of turns
n = number of turns from terminal 7.

The transformation ratio with connection A-F then becomes:

$$g = (V_o/V_1)^2 = (1 + n/3N)^2 \qquad \text{(Eq 15-7)}$$

2. V_1 connected to terminal B

a) At terminal E the output voltage is $3/2V_1$. Thus, the transformation ratio with connection B-E is:

$$g = (3/2)^2 = 1:2.25 \qquad \text{(Eq 15-8)}$$

b) At terminal G the output voltage is:

$$V_o = V_1(1 + n/2N) \qquad \text{(Eq 15-9)}$$

where n = number of turns from terminal 5.

The transformation ratio with connection B-G becomes:

$$g = (V_o/V_1)^2 = (1 + n/2N)^2 \qquad \text{(Eq 15-10)}$$

c) At terminal D the output voltage is $2V_1$. The transformation ratio with connection B-D is:

$$g = (2)^2 = 1:4 \qquad \text{(Eq 15-11)}$$

d) At terminal F the output voltage is:

$$V_o = V_1(3/2 + n/2N) \qquad \text{(Eq 15-12)}$$

where n = number of turns from terminal 7.

The transformation ratio with connection B-F, then, is:

$$g = (V_o/V_1)^2 = (3/2 + n/2N) \qquad \text{(Eq 15-13)}$$

3. V_1 connected to terminal C

a) At terminal E, the output voltage is $3V_1$. The transformation ratio with connection C-E becomes:

$$g = (3)^2 = 1{:}9 \qquad \text{(Eq 15-14)}$$

b) At terminal G, the output voltage is:

$$V_o = V_1(2 + n/N) \qquad \text{(Eq 15-15)}$$

where n = number of turns from terminal 5.

The transformation ratio with connection C-G is:

$$g = (V_o/V_1)^2 = (2 + n/N)^2 \qquad \text{(Eq 15-16)}$$

c) At terminal D, the output voltage is $4V_1$. The transformation with connection C-D becomes:

$$g = (4)^2 = 1{:}16 \qquad \text{(Eq 15-17)}$$

d) At terminal F the output voltage is:

$$V_o = V_1(3 + n/N) \qquad \text{(Eq 15-18)}$$

where n = number of turns from terminal 7.

The transformation ratio with connection C-F is:

$$g = (V_o/V_1)^2 = (3 + n/N)^2 \qquad \text{(Eq 15-19)}$$

Sec 15.3.1 A High-power 10-Ratio Unun

Figure 15-2B evolved after several attempts at rearranging the windings of **Figure 15-2A** for best overall performance (optimizing the effective characteristic impedances of the windings). **Photo 15-E** shows the bottom view of an unmounted Unun using the circuit of **Figure 15-2B**. The top-left lead is terminal E. The top-right lead is terminal D. The bottom-left lead is terminal B. The center lead (connected to the SO-239 connector) is terminal A.

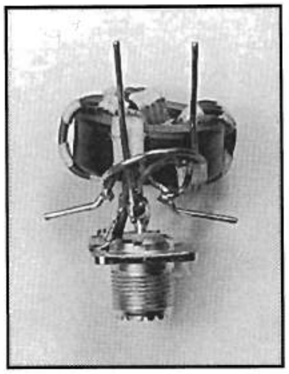

Photo 15-E. *Bottom view of the 10-ratio Unun. The connector is on terminal A.*

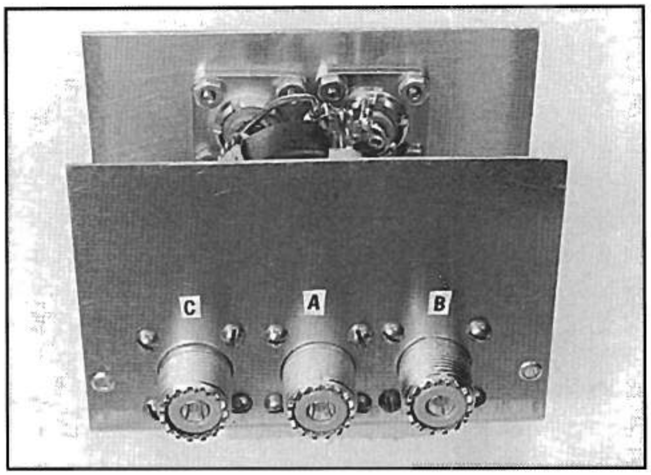

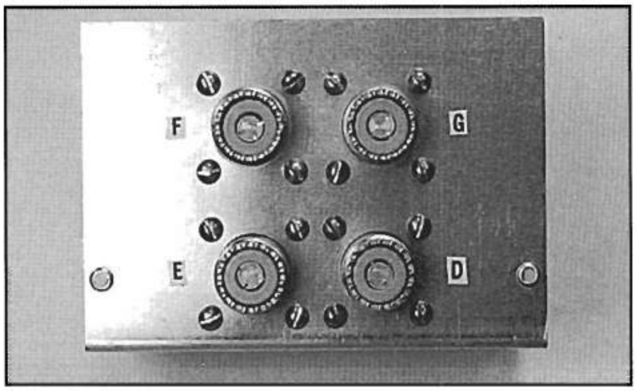

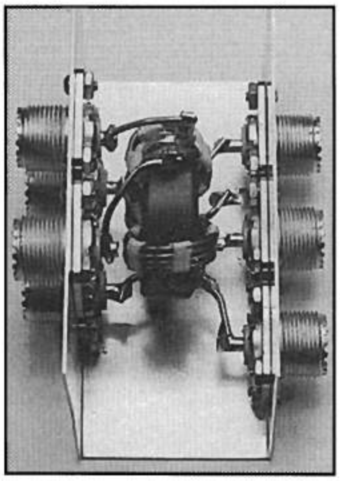

Photo 15-F. *Three different views of the 10-ratio Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high CU-3015A minibox.*

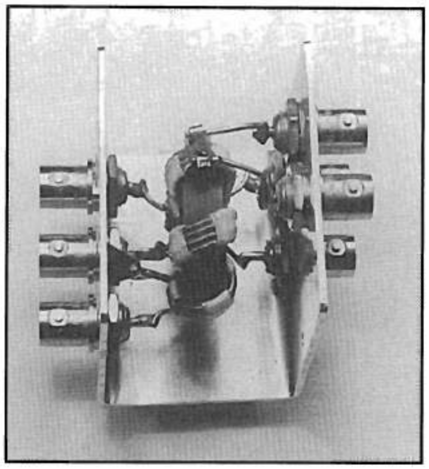

Photo 15-G. *The low-power unit mounted in a homemade 2.25 inch long by 1.5 inch wide by 2.25 inch high minibox.*

The bottom-right lead is terminal C. Below these three bottom leads is a ground connection (terminal 3 in **Figure 15-2B** to the SO-239 connector. **Photo 15-F** shows three different views of this high-power unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high CU-3015A minibox.

This 10-ratio Unun has four quadrifilar turns of No. 14 H Thermaleze wire on a 1.5-inch OD ferrite toroid with a permeability of 250. Winding 5-6 is tapped at 2 turns from terminal 5 and winding 7-8 is tapped at 2 turns from terminal 7.

If the 9:1 ratio (connection C-E), the 12.25:1 ratio (connection C-F), and the 16:1 ratio (connection C-D), are to be used at the full legal limit of amateur radio power, then I suggest that winding 3-4 be replaced with No. 12 H Thermaleze wire. If not, then these three ratios should be used at lower power levels (of 500 watts continuous and 1 kW peak). It should also be mentioned that using No. 12 wire for winding 3-4 adds a greater degree of *difficulty* to the construction process.

A listing of the expected performance across the band from 1.7 to 30 MHz, with the various ratios, is as follows:

16:1 (D-C); 50:3.125 ohms
Ratio is constant up to 21 MHz. It then decreases by 15 percent.

12.25:1 (F-C); 50:4.08 ohms
Ratio is constant.

9:1 (E-C); 50:5.56 ohms
Ratio increases by 5 percent.

6.25:1 (G-C); 50:8 ohms
Ratio is constant.

4:1 (D-B); 50:12.5 ohms
Ratio decreases by 5 percent.

3.06:1 (F-B); 50:16.3 ohms
Ratio decreases by 10 percent.

2.25:1
a) (E-B); 50:22.22 ohms
Ratio increases by 4 percent.
b) (B-E); 50:112.5 ohms
Ratio increases by 50 percent (the greatest deviation across the band of any of the ratios).

1.78:1
a) (D-A); 50:28.1 ohms.
Ratio is constant.
b) (A-D); 50:89 ohms.
Ratio increases by 15 percent.

1.56:1
a) (G-B); 50:32 ohms
Ratio increases by 10 percent.
b) (B-G); 50:78 ohms
Ratio increases by 40 percent.

1.36:1
a) (F-A); 50:36.8 ohms
Ratio decreases by 9 percent.
b) (A-F); 50:68 ohms
Ratio increases by 1.5 percent.

Sec 15.3.2 A Low-power 10-Ratio Unun

Photo 15-G shows a low-power unit mounted in a homemade 2.25 inch long by 1.5 inch wide by 2.25 inch high minibox. It has five quadrifilar turns of No. 16 H Thermaleze wire on a 1.25-inch OD ferrite toroid with a permeability of 250. The tap on winding 5-6 (**Figure 15-2B**) is at three turns from terminal 5 and on winding 1-2; it is three turns from terminal 1. Because the number of turns is different from the high-power unit, so are the ratios that use the taps. In this case, they are a little larger. Specifically, the tapped ratios are now: 1:12.96, 1:6.76, 1:3.24, 1:1.69, and 1:1.44. If the taps were at two turns from terminals 5 and 1, the ratios would be a little less than those of the high-power unit. You can play with the equations in the first section of this chapter and arrive at many different ratios.

Because this Unun has shorter transmission lines than its high-power counterpart, the deviations of the ratios across the HF band are generally smaller. Also, if winding 3-4 (in **Figure 15-2B**) were replaced with No. 14 H Thermaleze wire, this low-power unit could very well be rated at 500 watts of continuous power for all ratios!

Chapter 16

Ununs for Beverage Antennas

Section 16.1 Introduction

The Beverage antenna[19] is well known by 160-meter enthusiasts for enhanced signal-to-noise ratios when there are high levels of interference and atmospheric noise. If erected properly, Beverages also have excellent directivity. However, they are quite inefficient and, therefore, not generally suitable as transmitting antennas. Important considerations with Beverages are the terminating resistor (for the more common single-wire version) and the input matching Unun (unbalanced-to-unbalanced transformer). The terminating resistor and the impedance ratio of the Unun are determined by the characteristic impedance of the antenna acting as a long transmission line with one good conductor and one poor conductor (the earth). This line is generally between 400 and 600 ohms, and theoretically given by:

$$Z_0 = 138 \times \log(4h/d) \qquad \text{(Eq 16-1)}$$

where:

Z_0 = characteristic impedance of the Beverage
h = height of the wire above ground
d = diameter of the wire.

This chapter presents low- and high-power versions of multimatch Ununs designed to match 50-ohm cable to unbalanced loads from 450 to 800 ohms. The low-power unit, which is capable of handling continuous power levels up to 100 watts, is specifically designed for the Beverage antenna when it is performing as a receiving antenna. The high-power unit, which is capable of handling 1 kW of continuous power, can be used with the Beverage or any other traveling wave antenna when used as a transmitting antenna. Also presented are high-power designs capable of flat response, including the entire AM broadcast band. These multimatch Ununs could be of interest to designers of high-power amplifiers for the broadcast band. A little theory on how these devices are designed is also provided.

Sec 16.2 A Little Theory

Transmission line transformers[2] (the Unun being a subset thereof) are known for having greater bandwidths and efficiencies than their counterparts, the conventional transformers. Design considerations for the two types of transformers are also vastly different. Transmission line transformers use chokes and transmission lines, while conventional transformers use flux linkages.

High-impedance Ununs (and Baluns), which match 50 ohms unbalanced to impedances as high as 800 ohms, lie at about the edge of this technology's capability. The reasons are: 1) the windings require more turns because higher reactances are needed for isolating the input from the output, and 2) they require higher characteristic impedances in the transmission lines because the loads they see are greater. Therefore, when winding one of these devices, you'll *just* run out of space on your toroidal cores when trying to satisfy the low frequency and high frequency objectives. Incidentally, beaded transmission lines are not recommended at these impedance levels because of their excessive losses.

There are two methods for obtaining broadband operation at these high impedance levels. One uses Guanella's 9:1 and 16:1 Baluns, which are converted to Unun operation.[2] The other uses higher-order windings (quadrifilar in this case) on a single core, which is an extension of Ruthroff's bootstrap approach.[2] The Guanella approach, which uses coiled transmission line connected in series at the high-impedance side and in parallel at the low-impedance side, results in very broad bandwidths—but with difficulty in meeting low-frequency objectives. Low-frequency models[2] show that, with ratios above 4:1,

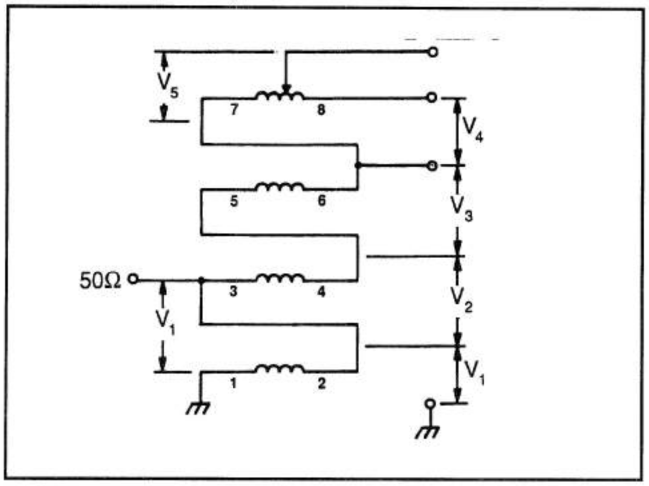

Figure 16-1. *Schematic diagram of the quadrifilar design, using Ruthroff's approach for high-impedance, low-frequency Ununs like the Beverage matching transformer.*

some of the coiled windings are connected in parallel, resulting in reduced reactances. However, with the Ruthroff approach, all of the inductances (at the low-frequency end) end up mutually aiding each other. However, Ruthroff's approach suffers at the high-frequency end because a direct voltage is summed with three voltages that traverse various lengths of transmission lines. As a result, Guanella's approach of summing voltages of equal delays is preferred for the higher frequency bands, and Ruthroff's approach is most often used for the lower frequency bands. This chapter presents designs using Ruthroff's approach.

Figure 16-1 shows the schematic diagram of a quadrifilar-wound Unun. If the lengths of the transmission lines are very short compared to the wavelength (therefore, phase delay and standing waves are negligible), then:

$$V_1 = V_2 = V_3 = V_4 \qquad \text{(Eq 16-2)}$$

at terminal 6,

$$V_o = V_1 + V_2 + V_3 = 3V_1 \qquad \text{(Eq 16-3)}$$

and the impedance ratio becomes:

$$g = (V_o/V_1)^2 = 9 \qquad \text{(Eq 16-4)}$$

At terminal 8, it becomes:

$$g = 16 \qquad \text{(Eq 16-5)}$$

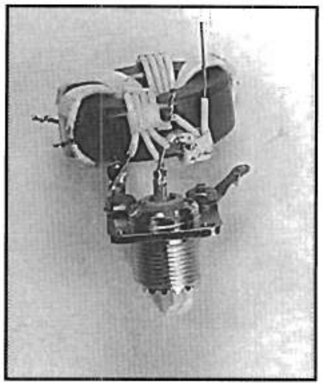

Photo 16-A. *The bottom view of the low-power Beverage antenna Unun.*

The voltage at the tap in winding 7-8 is:

$$\begin{aligned} V_o &= 3V_1 + V_5 \\ &= 3V_1 + n/NV_1 = V_1(3 + n/N) \\ &= V_1(3 + n/N) \qquad \text{(Eq 16-6)} \end{aligned}$$

where:
N = total number of turns
n = number of turns from terminal 7

The impedance ratio, using the tapped winding, becomes:

$$g = (V_o/V_1)^2 = (3 + n/N)^2 \qquad \text{(Eq 16-7)}$$

When the lengths of the transmission lines are significant, then important phase delays can occur and reduce the high frequency response. As you can see in **Figure 16-1**, V_2 travels one transmission line, V_3 travels two transmission lines, and V_4 travels three transmission lines. Additionally, the high frequency response is further diminished if the characteristic impedances of the transmission lines are not at their optimum values (which is hard to do at these impedances levels). Even with these major flaws, the Ruthroff approach is better for Beverage antenna use because this antenna's greatest advantages are on the lower frequency bands (80 and 160 meters).

Sec 16.3 A Low-power Design

Photo 16-A shows the bottom view of a 5-turn quadrifilar-wound Unun designed to handle 100 watts of continuous power with constant ratios from 9:1 to 16:1 in the 40- and 80-meter bands. It uses the Ruthroff approach of **Figure 16-1** and is shown here to give the reader a method for making the various interconnections. For operation on the 80- and 160-meter bands, I would use 6 quadrifilar turns on a 1.5-

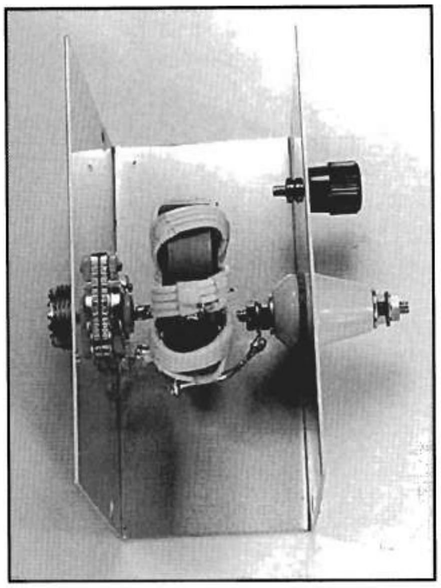

Photo 16-B. *The low-power Beverage antenna Unun mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox.*

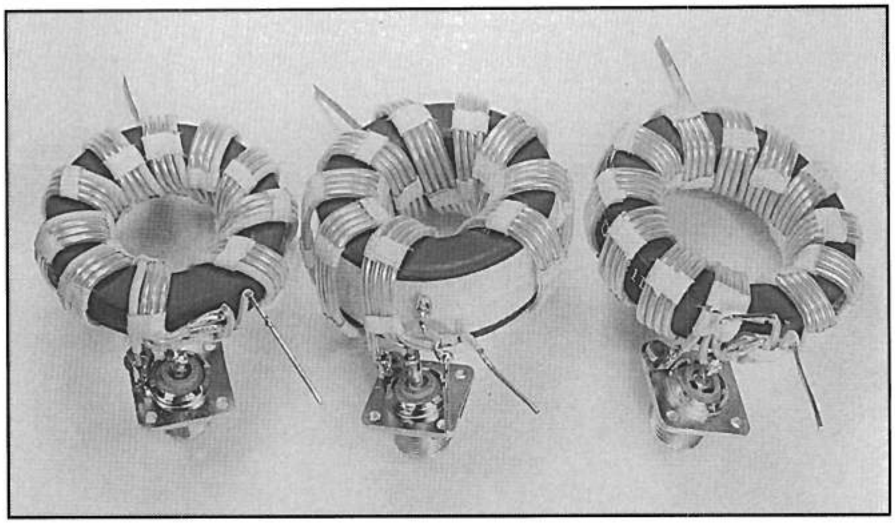

Photo 16-C. *Three high-power, low-frequency Ununs using a quadrifilar design with Ruthroff's approach. The one on the left is designed to cover the 80- and 160-meter bands. The other two are designed to cover the 160-meter and AM broadcast bands.*

inch OD ferrite toroid with a permeability of 250. The bottom winding is No. 20 hook-up wire and the other three are No. 22 hook-up wire. Winding 7-8 is tapped at 2 turns from terminal 7, yielding a 11.11:1 ratio, and at 3 turns from terminal 7, yielding a 12.25:1 ratio. Therefore, with outputs also at terminals 6 and 8, this Unun matches 50-ohm cable to loads of 450, 555.6, 612.5, and 800 ohms.

Photo 16-B shows the unit mounted in a 4 inch long by 2 inch wide by 2.75 inch high minibox. The output (the feedthrough insulator) is connected to one of the taps. A grounded binding post is also shown.

Sec 16.4 High-power Designs

Photo 16-C shows three high-power designs. The one on the left is specifically designed to cover the frequencies generally used with traveling wave structures like the Beverage antenna. This design has 10 quadrifilar turns on a 2.4-inch OD ferrite toroid with a permeability of 250. Winding 1-2 is No. 14 tinned copper wire, and the other three are No. 16 tinned copper wire. The wires are also covered with Teflon sleeving. Winding 7-8 is tapped at 5 turns from terminal 7, yielding a ratio of 12.25:1. When matching 50-ohm cable to loads of 450 ohms (terminal 6), 612.5 ohms (the tap), and 800 ohms (terminal 8), the variation in response is less than 5 percent from 1.5 to 4 MHz. At 6.5 MHz, the variation (which is an increase in the impedance ratio) increases to about 20 percent. **Photo 16-D** attempts to provide a better view of the connections.

The other two high-power Ununs in **Photo 16-C** are specifically designed to cover the broadcast and 160-meter bands. The one in the center has 9 quadrifilar turns (of the same wires as above) on a stack of two 2.4-inch OD ferrite toroids with permeabilities of 250. The tap on winding 7-8 is now at 4 turns from terminal 7, yielding a ratio of 11.86:1. When matching 50-ohm cable to 450 ohms (terminal 6), 593 ohms (the tap), or 800 ohms (terminal 8), the response is literally

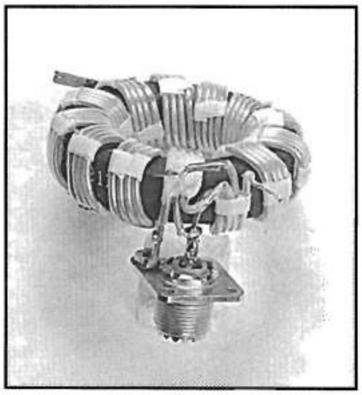

Photo 16-D. *The bottom view of the high-power Beverage antenna Unun.*

flat from 0.5 to 2 MHz. At 4 MHz, the ratios increase by about 6 percent. At 7 MHz, they increase by about 20 percent.

The Unun on the right in **Photo 16-C** illustrates another way of obtaining the same performance as above. In this case, the design has 12 quadrifilar turns (of the same wires as above) on a 2.68-inch OD ferrite with a permeability of 290. The tap on winding 7-8 is at 6 turns from terminal 7, yielding a 12.25:1 ratio (instead of 11.86:1 as above). Although the performance of this design is practically the same as the one above (using the two 2.4-inch OD cores), it is a much more expensive design because the 2.68-inch OD core is not nearly as popular. However, if a broadband, high-power and high-impedance Unun (or Balun) is required to cover 1.5 to 30 MHz, then these expensive 2.68-inch OD ferrite cores are very likely the *only* alternative!

Chapter 17

Concluding Remarks

Sec 17.1 Introduction

As I've often said, there is little information that deals with aspects of design and applications of practical hardware. Most companies are reluctant to publish their results for fear of giving away their hard-earned secrets. Textbooks only contain a few paragraphs on the subject. Therefore, very few people fully understand this technology, and, as a result, it is still far from reaching its full potential.

This book not only contains my designs, which have appeared in series written for *CQ* and *Communications Quarterly*, but also my views on other articles that have appeared in the amateur radio literature.

In the process of converting and combining the articles into appropriate chapters, one section in my article in the *CQ*, March 1993 issue entitled "Dual-Ratio Ununs," stood out as having broader applications. It is entitled "Reflections on Power Ratings" and is presented here. This information is followed by a section on misconceptions and one on the "state of the art."

After reading this book, some might think I was overly critical and didn't agree with any of the designs or explanations (or both) presented in the amateur radio literature. This is quite true. In taking this stand, I was hoping to provoke, in return, critical comment on my work. In this way, we can help our amateur friends—and perhaps even our professional friends—by advancing the understanding and application of these very useful transformers.

Sec 17.2 Reflections on Power Ratings

Power rating is one of the most controversial and least understood specifications for the transmission line transformer. As of today, no professional group has yet set the standard for this specification (as well as any other) for this popular class of matching transformers. In fact, manufacturers of ferrites, which are mainly the materials used with these devices, only specify them for their uses as conventional transformers and inductors, or microwave devices.

It is well known that power ratings for practically all conventional devices are based upon catastrophic failures (usually exceeding a voltage or current limit), and failures over a relatively short period of time due to an excessive rise in temperature.

With transmission line transformers, there are really two catastrophic-type failures that can occur. One is voltage breakdown. If the device is misterminated with a high impedance (especially an open circuit), a breakdown of the insulation can occur. This is particularly true of the 1:1 Balun (50:50 ohms) that is terminated with the very high impedance of a full-wave dipole or inverted V. Using heavily coated wires (but still maintaining a characteristic impedance close to 50 ohms), or small but high-power coaxial cable, can help under these conditions.

The second catastrophic failure occurs at the low-frequency end of the transformer's passband, when the energy is not completely transmitted to the output circuit by a transmission line mode. This takes place when the reactance of the coiled or beaded transmission line is not sufficient to prevent conventional transformer currents or shunting currents to ground. Under these conditions, harmful flux can take place in the core or beads. Nonlinearities can also occur if the flux becomes appreciable. The objective in design at the low-frequency end is to have a margin of safety such that, with a termination of about three times (hence VSWR of 3:1) that of a matched condition,[2] no flux will appear in the core or beads.

The failure due to an excessive rise in temperature is the least understood of the two because it involves the failure mechanism in transmission line transformers when only transmission line currents are allowed to flow. Unlike the conventional transformer whose losses are current dependent (wire, eddy current, and hystereis losses), the transmission line transformer's losses are voltage dependent (a dielectric-type). That is, the greater the voltage drop along the length of the transmission lines, the greater the loss. Furthermore, it can be shown that only low-permeability ferrites (less than 300) yield the extremely high efficiencies of which these transformers are capable.[2]

Because all transmission line transformers have voltage drops along their transmission lines, we must look at their high-frequency models to determine the magnitude of these drops and, hence, the temperature rise that can be expected. Here are some examples:

If V_1 is the voltage that appears on the 50-ohm side of the transformer, then:

1) For a quintufilar-wound Unun, the longitudinal voltage-drop is $V_1/5$ in matching to lower impedances (like 32 or 18 ohms) and $V_1/4$ in matching to higher impedances (like 78 or 139 ohms).

2) For a quadrifilar-wound Unun, the longitudinal voltage-drop is $V_1/4$ in matching to lower impedances (like 12.5 or 28 ohms) and $V_1/3$ in matching to higher impedances (like 89 or 200 ohms).

3) For a trifilar-wound Unun, the longitudinal voltage-drop is $V_1/3$ in matching to lower impedances (like 22.22 or 25 ohms) and $V_1/2$ in matching to higher impedances (like 100 or 112.5 ohms).

4) For a bifilar-wound Unun, the longitudinal voltage-drop is $V_1/2$ in matching to a lower impedance of 12.5 ohms and V_1 in matching to a higher impedance of 200 ohms.

Because the quintufilar-wound Unun has the lowest voltage drop, it is expected to have the highest efficiency. Furthermore, it can be seen that the highest efficiencies occur in matching to lower impedances. Very accurate measurements[2] have shown that 4:1 Ununs, using ferrite cores with permeabilities of 125, have exhibited losses of only 0.02 to 0.04 dB in matching 50 to 12.5 ohms from 1 MHz to over 30 MHz. Even though the Ununs in this book have mostly used permeabilities of 250, and should have slightly greater losses, many use higher-order windings (trifilar, quadrifilar, and quintufilar) and, hence, have lower longitudinal voltage drops. Therefore, they should have losses of only 0.02 to 0.04 dB, as well.

When matching at the 1 kW level, the figures above mean that only 5 to 10 watts would be dissipated in the Unun. As a heatsink, these small transformers should be able to handle this loss easily. In fact, they should be able to handle several times this level of continuous power. Also, because they use heavily coated wires, their peak power ratings should be greater by more than a factor of two!

Another important power rating consideration is to determine what happens when the transformers are misterminated. Because the losses being considered now are dielectric-types and, hence, voltage-dependent, the harmful terminations are greater than that for which the transformers were designed. For example, if the termination is three times greater (a VSWR of 3:1), the voltages along the transmission lines would increase by a factor of 1.73. This means the losses would practically double. The Ununs described in this series, when matching to impedances lower than 50 ohms, should easily handle this mismatch. Obviously, mismatches in the range of 10:1 would result in much lower efficiencies and should be avoided.

The analysis of the losses in Baluns follows the same pattern. The voltages are as follows: under matched conditions for a 1:1 Balun (50:50 ohms), $V_1/2$; for a 4:1 Balun (50:200 ohms), V_1; for a 9:1 Balun (50:450 ohms), $1.5V_1$; and for higher-impedance Baluns it could be $2V_1$. The higher voltage drops, together with high VSWRs, means that high-impedance Baluns (and Ununs) have more loss and require larger structures to dissipate the heat. It should also be pointed out that there is a tradeoff in efficiency for low-frequency response with Baluns (and Ununs) when matching 50 ohms to higher impedances like 200 ohms, 300 ohms, 450 ohms, and higher. This is done by using permeabilities of 125 and lower.

Finally, I thought it might be useful to give some general guidelines as to what efficiencies you might expect with Baluns and Ununs when using ferrite cores or beads with a permeability of 250. Here are some expected efficiencies when matching 50 ohms to various loads under matched conditions:

Loads	**Efficiency**
50 ohms or less	98 to 99.5 percent
50 to 100 ohms	97 to 98 percent
100 to 200 ohms	96 to 97 percent
200 ohms and above	93 to 96 percent

As I mentioned earlier, these efficiencies would be reduced by a percent or two with a VSWR of 3:1, which increases the loss by a factor of about two. Also, the efficiencies can be increased by a percent or two with high-impedance loads (greater than 100 ohms) by resorting to lower permeability ferrites that trade off efficiency for low-frequency response. In closing, I would like to say that high-permeability manganese-zinc ferrites should be avoided because of their much higher losses. Furthermore, their losses are highly frequency dependent, while low-permeability nickel-zinc ferrites are not.

Sec 17.3 Misconceptions

From recent discussions on the air and phone calls concerning Baluns, I think the most *expensive* misconception regarding Baluns is the assumption that a 9:1 (450:50 ohm) Balun would match 50-ohm cable (or the output of a linear or transceiver) to 450-ohm twin lead, without considering the effect of its termination. In truth, the 9:1 Balun would *only* see 450 ohms if the line were terminated in 450 ohms. In reality, if the line were terminated in a 50-ohm dipole, the Balun would see 50 ohms when the line is a half-wave long and 4050 ohms when is a quarter-wave long. The 9:1 Balun is clearly useless in this application.

By far, most misconceptions regarding Baluns are due to the many radio amateurs who perceive these devices as conventional transformers that transmit the energy from input to output by flux linkages and not as transmission line transformers, which transmit energy by an efficient transmission line mode. This is clearly shown by the writers who have compared their "new" coaxial cable (coiled about a toroid or threaded through ferrite beads) Baluns with Baluns using wire transmission lines coiled about a ferrite rod or toroid. They claim their Baluns are better because the others: 1) were limited by leakage inductance, 2) did not exhibit true 1:1 impedance transformations, 3) were prone to core saturation, 4) added a reactive component to the input impedance, 5) were susceptible to unbalanced and mismatched loads, and more importantly, 6) had *more* loss.

If the writers had accepted the correct model for these devices (given to us by Guanella and Ruthroff), which shows that they are really chokes (lumped elements) and configurations of transmission lines (distributed elements), then there are several parameters they should have considered in their comparisons. They are: 1) the characteristic impedances and lengths of the transmission lines (the high-frequency capability), 2) what form of the 1:1 Balun or 4:1 Balun is used by the other Balun, 3) the low-frequency capabilities (safety margins), 4) power capabilities, and finally 5) efficiencies.

Now, had the writers used the proper parameters in their comparisons, they would have found that mismatch loss was mistaken for real (ohmic) loss; high-frequency response was limited by standing waves, and not leakage inductance or shunting capacitance; the beaded-coax Balun had *more* loss than a well-designed Balun using wire or coax transmission lines coiled about a toroid; and that their comparisons were made with either the trifilar (voltage) 1:1 Balun or the Ruthroff 4:1 Balun, which are inferior designs.

In fact, the perception that the transmission line transformer is actually a conventional transformer is so prevalent, that a new name for this class of devices should be considered—*broadband transmission line matching networks*. This name (without the word *transformer*) would help in dispelling inaccurate perceptions and in standardizing the schematic diagrams. It would place the coiled or beaded transmission lines (in the high-frequency models) horizontally, and eliminate the phasing or polarity dots.

Sec 17.4 The State of the Art

Until very recently, the radio amateur had only two types of Baluns available in the literature and on the market. They were the so-called 1:1 and 4:1 "voltage" Baluns. As was shown in **Chapter 7**, the comparisons by others with new 1:1 designs using coaxial cable (called "current" Baluns) were made with an inferior trifilar-wound Balun, instead of Ruthoff's design that appeared in his 1959 paper and became the industry's standard. Ruthroff's third conductor on his 1:1 Balun was on a separate part of the toroid, thus giving it practically the same characteristics as the Guanella ("current") Balun. These articles on newer designs not only gave a new language to our Baluns, but also presented questionable statements regarding their performances. It would be interesting if the authors of these articles compared their Baluns with well-designed Ruthroff or Guanella Baluns using 50-ohm bifilar windings or coaxial cables on low-loss ferrite toroids (less than 300 permeability). I am quite sure their claims would be greatly diminished.

As was noted in **Chapter 8**, the 4:1 voltage Balun appeared in the amateur radio journals about 25 years ago (the same time as the "inferior" 1:1 voltage Balun). Considerable design information appeared in the handbooks of the time regarding the construction and performance of this Balun. Furthermore, this information also stayed the same over these many years. As was shown in **Chapter 8**, the design was found lacking. However, with some rather simple changes, like doubling the cross-sectional area of the core, increasing the number of turns from 10 to 14, and using extra insulation on the wires to increase the characteristic impedance of the coiled transmission line from about 50 to 100 ohms (the objective), a much better design emerged. In fact, for balanced antenna systems, this new design might well be described as "peerless."

A 4:1 Guanella (current) Balun has now appeared in our handbooks. This more flexible Balun uses two transmission lines wound on separate cores and connected in series at one end and in parallel on the other. Literally, no design information is given on its construction. What is offered are recommendations for the permeability of the ferrite cores. Values from 850 to 2500 are proposed. However, use of these high permeabilities would result in lossy Baluns.

I also found it interesting, in my work on these devices, that the classic papers of Guanella[3] and Ruthroff[9] are still the cornerstones of this technology known as *transmission line transformers*. To be sure, some of us have extended the work of these two by using better measuring equipment, creating more complicated configurations, and finding new applications. However, it is apparent from the articles published in the amateur radio journals and discussions on the air and at club meetings that most radio amateurs still perceive these devices as conventional transformers. They don't look at these devices as Guanella and Ruthroff did—as chokes and transmission lines. As a result, there has been a lack of good design information in our literature.

There are many new and useful designs possible with this technology, as discussed in **Part I** of this book. They include: higher power levels, applications on the VHF and UHF bands and above, and new Baluns and Ununs with ratios other than 1:n^2 where n = 1, 2, 3, . . . , etc. This book presents some designs* and suggestions for higher-power and higher-frequency applications.

I see two reasons for the lack of emergence in this technology. They are:

1) This subject is not adequately covered in any college textbook, and it generally has not been of interest to academics who rightfully view their role as basic research and not applications. As a result, there are few graduates with any skill in the design of transmission line transformers—in contrast to the areas of transmission line, waveguide, and antenna theory.

2) The professional societies don't receive enough application papers. Although much of the research and development work performed in industry is highly innovative, important to the advancement of the technology, and certainly publishable in scientific journals, corporations are often reluctant to allow publication for fear of "aiding" their competition. It has been stated[29] that in the past few decades, the submission of application papers to the technical journals of the IEEE has declined. In fact, a survey by one of the technical societies showed that 85 percent of the submissions now come from universities, not industry!

In order to assist technologies, like *transmission line transformers*, which are far from reaching their potential applications, IEEE has instituted a program called *Emerging Practices in Technology* (EPT). The object of the EPT program is to facilitate the development of new standards by disseminating and making available various EPT papers to the broadest possible audience worldwide. The papers on practices in various areas of technology are peer reviewed by relevant IEEE Technical Committees, and have the potential for standardization in the future. The papers (mine is **Reference 30**) are published by the IEEE Standards Press.

FOOTNOTE

*Kits and finished units available from Amidon Associates, Inc., 240 Briggs Ave., Costa Mesa, CA 92626.

Part II

References

1. Jerry Sevick, W2FMI, *Transmission Line Transformers*, 1st edition, Amateur Radio Relay League, 1987.

2. Jerry Sevick, W2FMI, *Transmission Line Transformers*, 4th edition, Noble Publishing Corporation, 2001.

3. G. Guanella, "Novel Matching Systems for High Frequencies," *Brown-Boveri Review*, Volume 31, September 1944, pages 327–329.

4. Walt Maxwell, W2DU, "Some Aspects of the Balun Problem," *QST*, March 1983, pages 38–40.

5. John Belrose, VE2CU, "Transforming the Balun," *QST*, June 1991, pages 30–33.

6. Roy Lewallen, W7EL, "Baluns: What They Do and How They Do It," *The ARRL Antenna Compendium*, Volume 1, Amateur Radio Relay League, 1985, pages 12–15.

7. Joe Reisert, W1JR, "Simple and Efficient Broadband Balun," *Ham Radio*, September 1978, pages 12–15.

8. Richard H. Turrin, W2IMU, "Broad-Band Balun Transformers," *QST*, August 1964, pages 33–35.

9. C. L. Ruthroff, "Some Broad-Band Transformers," *Proceedings of the IRE*, Volume 47, August 1959, pages 1337–1342.

10. Richard A. Genaille, W4UW, "How to Build a Multi-Tap Unun," *CQ* May 1992, pages 28–32.

11. Bruce Eggers, WA9NEW, "An Analysis of the Balun," *QST*, April 1980, pages 19–21.

12. Richard H. Turrin, W2IMU, "Application of Broad-Band Balun Transformers," *QST*, April 1969, pages 42, 43.

13. George Badger, W6TC, "A New Class of Coaxial-Line Transformers, Part 1," *Ham Radio*, February 1980, pages 12–18.

14. George Badger, W6TC, "A New Class Of Coaxial-Line Transformers, Part 2," *Ham Radio*, March 1980, pages 18–29, 70.

15. Bill Orr, W6SAI, "Radio Fundamentals: The Coax Balun," *CQ*, November 1993, pages 60–65.

16. Albert Roehm, W2OBJ, "Some Additional Aspects of the Balun Problem," *The ARRL Antenna Compendium*, Volume 2, Amateur Radio Relay League, pages 172–174.

17. Lew McCoy, W1ICP, "Let's Talk About Wire," Part 1, *CQ*, January 1993, pages 42–45.

18. Lew McCoy, W1ICP, "Let's Talk About Wire," Part 2, *CQ*, February 1993, pages 32–42.

19. Gerald Hall, K1TD, Editor, *The ARRL Antenna Book*, 16th edition, Amateur Radio Relay League, 1991, Chapter 25.

20. George Grammer, W1DF, "Simplified Design of Impedance-Matching Networks," in three parts, *QST*, March, April, and May 1957.

21. D.K. Belcher, "RF Matching Techniques, Design and Example," *QST*, October 1972, page 24.

22. Lew McCoy, W1ICP, "The Ultimate Transmatch," *QST*, July 1970, pages 24–27, 58.

23. Warren Bruene, W5OLY, "Introducing the Series-Parallel Network," *QST*, June 1986, page 21.

24. Louis Varney, G5RV, "The G5RV Muliband Antenna...Up-to-Date," *The Antenna Compendium*, Volume 1, Amateur Radio Relay League, pages 86–90.

25. Jerry Sevick, W2FMI, "6:1 and 9:1 Baluns," *Communications Quarterly*, Winter 1993, pages 43–51.

26. D. Meyer, "Equal-Delay Networks Match Impedances Over Wide Bandwidths," *MICROWAVES & RF*, April 1990, pages 179–188.

27. S.E. London and S.V. Tomeshevich, "Line Transformers with Fractional Transformation Factor," *Telecommunications and Radio Engineering*, Volume 28/29, April 1974.

28. Carl Markle, K8IHQ, "Wideband RF Baluns," *73 Amateur Radio Today*, September, 1992, pages 20–23.

29. S. Maas, D. Hornbuckle, D. Masse, "Applications Papers for the *MTT Transactions*," Summer 1992, IEEE MTT-S Newsletter, pages 3, 5.

30. Jerry Sevick, W2FMI, "Design and Realization of Broadband Transmission Line Matching Transformer," *Emerging Practices in Technology*, IEEE Standards Press, 1993.

Index

Amidon Associates, Inc...18, 30–32
  kits...30–31, 88
Antenna modes
  dipole mode...28
  inverted L mode...28, 31
Antenna tuners, Baluns for...55–63
Autotransformers...6, 9, 19, 21, 30–31, 52, 76, 97, 101

Badger, George...33–36, 89, 114
  two-tone test method...33
Baluns
  air-core...33, 41, 43, 46
  and dipoles...27–29, 35, 40, 45–46, 52–55, 57, 61–62, 65, 71, 100–111, 113
  and inverted Vs...27–28, 111
  and log-periodics...49, 54, 57,
  and multiband dipoles...61
  and quads...28, 65, 87
  and rhombics...80
  and Yagis...27–28, 33, 65, 73–74, 86
  attributes of a good design...41
  Balun/Unun...43, 87
  beaded-coax...36, 41, 54, 56, 86–87, 113
  bifilar toroidal...37–39, 40
  broadband...27, 29, 65, 71, 74, 78, 83, 85, 115
  coil...46
  Collins...34
  commercial...33, 36, 47, 52–54
  choke...4, 19, 27, 29–30, 34, 36–37, 40–41, 43, 56
  current...4, 27, 29–30, 35–36, 38–41, 43, 46–47, 50, 52–55–57, 61, 62, 65, 113–114
    bifilar current type...36
  dual-core...50–54, 65
  dual-ratio...65, 68–69
  economy...38–39
  ferrite core...30–31, 33–34, 39, 41, 46, 49, 78–79, 110
  for antenna tuners...55–63
  for 2-meter operation...65
  fractional-ratio...74
  "hardened"...62
  HF designs...65
  high-impedance...15, 79–80
  high-power...27, 37–40, 49–53, 57–58, 65, 81–82
  HI-Q...52–53
  low-power...27, 37, 39, 43, 50–54, 65
  medium-power...27, 37, 39, 81
  misconceptions about...113
  parallel-type...65, 66, 68–71, 74–75, 79, 82
  RF...29, 115
  rod-type...4, 31–34, 40, 47, 52–54, 63
  series-type...16, 65–69, 72–75, 79–82
  single-core...45, 50–51, 54, 73–74
  step-down...28, 51–52, 65, 73–74, 86
  step-up...65, 74
  three-conductor...30–31, 34–35, 43
  toroidal...32, 36–39, 41
  transformer-type...34–35
  trifilar-wound...30, 32, 113
  two-conductor...30–31, 33–37, 53
  VHF designs...65
  voltage...33, 35–36, 43, 47, 52, 54–55, 62, 113
  when to use...27–29
  "worst case" design...61–63
  1:1...2, 4, 15–16, 27–41, 43, 45, 54–56, 61–66, 68, 71, 89, 111–113
    articles on...29
    bifilar toroidal...37–39, 41
    in amateur radio literature...30–37
    trifilar-wound...30, 32, 113
    using ferrite cores...30, 112
  1:4...5–7, 11–12, 16, 21–22, 55–56, 67, 70, 72–74
  1.5:1...65–70, 89
  1.56:1...65–66
  1:6...72
  1:6.25...16, 72–73, 79–80
  1:9...15–16, 70–71, 76, 78–82
  2:1...28, 62, 65–71, 89
  4:1...5, 7, 27, 30, 35–37, 41, 43–62
    history...43–44
    beaded coax...54
    current...53
  6:1...71–82, 89
  6.25:1...71–75, 79
    high-impedance...74, 79
    low-impedance...74, 79

- 9:1...27, 36–37, 71–82, 89
  - high-impedance...79
  - low-impedance...79
  - practical designs...77–78
- 12:1...71–82, 89
- 16:1...27, 107
- 50:450-ohm...71, 82

Balun/Unun...43, 87
Beach umbrella antenna...Preface
Belrose, John...36–37, 115
Beverage antennas...107–110
Bootstrap...5, 7, 10–11, 13, 17, 21–22, 30, 68, 74, 83, 93, 107
Broadband matching transformers...29
Broadband transmission line matching networks...113

Characterization...17–20
Cores
- air...33, 41, 43, 46
- ferrite...30–31, 39, 46, 49, 78–79, 110, 112, 114
- geometry of...17
- low permeability...41
- magnetic...17, 30, 36, 46, 77
- materials...18
- powdered-iron...30, 32, 40, 53, 57–62, 80
- rod...3–4, 17–18, 51–52, 97
- toroidal...3, 11, 17, 35, 43, 46, 102, 107

Crossover...32–33, 37–39, 50–51, 66–68,

Delay line...4–5, 20, 22, 45, 77, 86
Dipole antennas...27–29, 35, 40, 45–46, 52–55, 57, 61–62, 65, 71, 100, 111, 113

Eggers, Bruce...28, 114

Ferrite
- Balun...33
  - beads...4, 7, 29, 34, 36, 43, 55–56, 69–70, 75, 78–80, 82, 113
- high-permeability...34–35, 54, 80
- core...30–31, 39, 46, 49, 78–79, 110, 112, 114
- efficiencies...39
- low-loss...32–33, 113
- low permeability...30, 37, 48, 54, 63, 75, 86, 112
- manganese-zinc...4, 30, 113
- nickel-zinc...18, 113
- rod...11, 20, 52, 84, 86
- toroids...13, 17, 31–32, 35–39, 47, 49–51, 53–54, 56, 66–68, 72, 74, 77, 81, 84–85, 90, 92–93, 101–104, 106, 109, 113

Flux canceling...21, 29, 83
Floating load...44–46, 58, 72–74, 77, 80
Fractional-ratio Ununs...86, 89–90

Genaille, Richard...28, 97, 115
Guanella, G.
- basic building block...3–5, 10–11, 16, 21–22, 29–30
- classic paper of 1944...3–5, 23, 27, 43, 57, 83, 114–115
- current/choke Balun...4, 27, 30, 36, 38–40, 43, 50, 54, 56
- dual-core Balun...52–53
- equal-delay approach...11, 13, 22, 86
- single-core Balun...44–45, 49–52
- two conductor...30–31, 33–34, 37, 40, 43
- 1:1 Balun...4, 29–31, 33–34, 38–40, 43, 65–66, 68
- 1:4 Balun...5–7, 11, 16, 21, 56, 67, 72, 79, 82
- 1:4 TLT...5–7, 22
- 1:4 Unun...5, 7, 20–22
- 1:9 Balun...15–16, 70, 76, 80–82
- 4:1 Balun...5, 7, 27, 43–45, 49–54, 56–57,68, 72–74, 76, 79, 114
- 4:1 Unun...7, 45, 83, 85–86
- 9:1 Balun...27, 71–72, 75–82, 107
- 16:1 Balun...27, 86, 107

G5RV antenna...55, 61–62, 115

Hairpin matching networks...28, 65
HI-Q Balun... 52–53

Imbalance current...28, 30–31, 35, 41
Inverted V antennas...27–28, 111
Isolation transformers...4, 27, 40, 66

L-C matching network...55
- L...56
- pi...56
- T...56
- Ultimate...56
- SPC...56

Lewallen, Roy...35–36, 115
- 1:1 Balun...35
- 4:1 Balun...35
- current Balun...35
- voltage Balun...35

Link coupling...55
Loads
- balanced...46, 51–52, 65–70, 72–75, 79, 81–82, 86
- center-tapped-to-ground...47, 60, 62, 73–74
- floating...44–46, 56–58, 60, 62, 72–74, 77, 80
- grounded at bottom...4, 45, 54, 73–74, 77
- grounded at center...3–4, 30–31, 37, 45–46, 54–55–57, 60, 73–74, 77
- mismatched...30, 32–35, 41, 113

  reactive...36, 62
  unbalanced...27, 30, 33–35, 41, 58, 85, 90–91, 107
Log-periodic antennas...49, 54, 57

Magnetizing inductance...17–18
Materials, core...18
Maxwell, Walt...34–37, 115
  choke Balun...34
  transformer-type Balun...34–35
  1983 article...34
McCoy, Lew...52–53, 55, 57–60, 62, 101, 115
Multiband antennas...55, 61–62
  dipoles...61
Multimatch transformers...89, 102

Orr, Bill...33, 115

Phase-delay...30, 37, 39, 73, 85, 87, 94, 108
Phase-inverter...4–6, 21, 30, 40, 43, 46, 54–55, 57
Pi network...56
Polarity dots...113
Power ratings...35, 54, 94, 111–112
  loads and efficiencies...112–113

Quad antennas...28, 65, 87

Reisert, Joe...32–35, 37–38, 41, 54, 115
  crossover...32–33, 37–39
Rhombic antennas...80
Roehm, Albert...36, 56, 115
Ruthroff, C.L.
  bifilar Unun...89
  bootstrap...5, 7, 10–11, 13, 15, 17, 21–22, 30, 68, 74, 83, 93, 107
  broadband transmission line transformers...22, 29
  classic paper of 1959...23, 27, 30, 34–35, 43, 46, 83, 113–115
  high-frequency model...5, 45
  hybrid transformer...43
  maxima technique...46
  phase-inverter configuration...21, 55, 57
  three-conductor Balun...30–31, 34, 43
  trifilar-wound Balun...15, 30, 32, 113
  voltage Balun...43, 52, 54–55, 62, 113
  1959 article...1, 4, 9, 10, 17, 29, 42, 57, 61, 87, 88
  1:1 Balun...3–4, 29–30, 34, 36, 43, 113
  1:4 Balun...5–7, 22, 56, 81
  1:4 TLTs...6
  1:4 Unun...5–6, 9, 11, 16, 21
  4:1 Balun...30, 43, 45–50, 53, 55–59, 61–62, 113
    rod-type...52–53
  4:1 bifilar Unun...89, 97
  4:1 Unun...30, 45–46, 68, 83–84, 87, 97
  100:25-ohm Unun... 84–85
  200:50-ohm Unun...84–85

Series and parallel-tuned circuits...55
Single-core Balun...45, 50–51, 54, 73–74
Step-down transformers...90
Step-up transformers...90, 97–99

Terman, Frederick E...22–23
Toroids...17, 20–21, 30, 51, 69, 73, 79, 81–82
  cores...3, 11, 35, 46, 107
  current-probe...35
  ferrite...13, 17, 31–32, 35–39, 47, 49–51, 53–54, 56, 66–68, 72, 74, 77, 81, 84–85, 90, 92–93, 101–104, 106, 109, 113
  powdered-iron...17, 31
  4C4...9, 18
Transformation ratios...37, 41, 43–44, 51–52, 57, 65–66, 68, 74, 85–86, 89–90, 92–93, 101–105
Transformers
  air-core...33
  Balun...56, 115
  bifilar, tapped...9
  broadband...30, 43, 100
    multimatch...89, 102
  conventional...21, 57, 76, 79, 83, 107, 111, 113–114
  equal-delay...12
  flux linkages...21, 83, 107, 113
  hybrid...43
  isolation...4, 27, 40, 66
  matching...29, 61–62, 108, 111
  multimatch...63, 76
  parallel...100
  step-down...90
  step-up... 90, 97–99
  toroidal...17, 91
  transmission line...3, 5, 21, 23, 27, 29, 43–44, 79, 82–83, 101–102, 107, 111–115
  unbalanced-to-unbalanced...43, 45, 65, 83, 107
  wideband..33
Transmatch...55–56, 58, 60, 101, 115
Transmission lines
  balanced...56
  beaded...17, 34, 37, 54, 57, 69–70, 75–76, 83, 86, 70, 107, 111, 113
  coax...29, 34, 83, 113
  coiled...5, 17, 29, 31, 33–34, 36, 41, 43–44, 46, 51–52, 57, 76, 83, 86, 107, 111, 114

  coupled...30, 32, 35, 43
  currents...27, 29, 36, 39, 41, 44, 46, 49, 76, 83, 111
  flat line...44, 46, 47, 57, 69, 74
  half-wave...29, 61
  high-impedance...75, 82
  ladder line...71
  matched...29, 62
  McCoy...53
  open-wire...5, 56
  quarter-wave...29
  theory...61, 76
  twin-lead...29, 34, 38, 56, 75
  very short...29, 44, 84, 90, 97–98, 104, 108
Transmission line transformers (TLTs)
  book by W2FMI...12, 44, 81, 115
  broadband...10, 13, 22, 29–30, 43, 100, 102, 113, 115
  catastrophic failures...111
  coiled wire...12
  equal-delay...12
  isolation transformers...4, 27, 40, 66
  low-ratio...9, 13
  power ratings...111–113
  trifilar-wound...15
  technology of...27, 113–114
  with 1:1 ratios...3–4
  with 1:4 ratios...5–7
  with ratios greater than 1:4...15–16
  with ratios less than 1:4...9–13
  1:1...5
  1:4...5–7, 21–22
  1:9...21
  1:16...21
Turrin, Richard
  articles...27, 30–32, 36, 115
  Baluns...27, 31–34, 36

Ultimate multimatch...104–105
Ununs
  air-core...33
  bifilar...89, 101, 112
  broadband ratios...89, 92, 97, 99, 101
  broadband...45, 52, 73–74, 79, 89
  coiled...91
  dual-ratio...97–100
    1.5:1 and 3:1 ratios...98–99
    2:1 and 4:1 ratios...98
  for Beverage antennas...107–110
    low-power design...108–109
    high-power designs...109–110
  for 2-meter operation...86
  fractional ratio...86, 89–90
  high-impedance...85–86, 107, 110
  high-power...83, 109–110
  input-matching...107
  low-power...83, 95
  multimatch...101–107
  parallel transformer...100
  parallel-type...74–75
  quadrifilar...67, 90–91, 94, 98–99, 102, 104, 106, 108–109, 112
    tapped...102
  quintufilar...12, 16, 65–66 72–73, 81, 90–94, 96–99, 112
  step-down...67, 73, 86
  step-up...69, 72, 79, 90
  trifilar...10, 66, 68, 89-90, 94, 97, 102
    tapped...91, 101–102, 112
  1:1.33...16, 80–82
  1:1.5...22
  1:2...22, 62, 66
  1:3...81
  1.33:1...71, 89, 93–94
  1:4...5–7, 9, 11, 16, 20–22
  1.5:1...89–93
  1.56:1...65–66, 72, 74, 79, 91–92
    quintufilar...66, 91-93, 97–98, 102
  2:1...28, 89–95
  4:1...7, 30, 45-46, 68, 83–87, 97, 112
  5-ratio...102–103
    high-power version...103
    low-power version...103–104
  9:1...78–79
  10-ratio...102, 104–105
    high-power version...105–106
    low-power version...106
    quadrifilar...102
  50:12.5-ohm...84–85
  100:25-ohm...85
  200:50-ohm...85

V antennas...80
Varney, Louis...62, 115
Vertical antennas...89
  shunt-fed towers performing as...89

Wideband transformers...33
Wire antennas...55

Yagi antennas...27–28, 33, 65, 73–74, 86

3 CRYSTAL SETS

The simplest type of radio receiver consists of an aerial, a tuned circuit and a detector–plus headphones to listen to the signals received by the tuned circuit. It is generally called a *crystal set* because it works on the same principle as the very earliest radio sets, using a lead-galena crystal and a 'cat's whisker' as a detector. The only real difference is that the crystal and cat's whisker have been replaced by the much more efficient point-contact *diode*.

A simple *tuned circuit* can be made by winding a coil of wire of specific diameter and length (number of turns) to match the capacitance range of a variable capacitor. The combination of coil (inductance) and capacitance can then be adjusted by turning the variable capacitor spindle to be resonant, or tune in to various broadcast frequencies. The signal which is tuned in is then passed to the diode which blocks the RF content, and rectifies and passes on the AF content to the headphones, which transform these signals into audible sounds.

The tuned circuit can be combined with the aerial by winding the coil on a *ferrite rod*. This increases the inductance of the coil (reducing the length of wire required), whilst the rod itself also acts as an aerial to pick up the radio signals. This is the type of aerial employed in most domestic radio receivers. However, with very simple sets working with very weak signals, an external aerial may also be required.

The performance of all simple, low-powered receivers is, in fact, very much dependent on the efficiency of the aerial and tuned circuit, which can be regarded as the critical part of the system. Performance can perhaps be improved by experimenting with different aerial/tuned-circuit systems, and by the addition of further circuitry, as explained in later chapters.

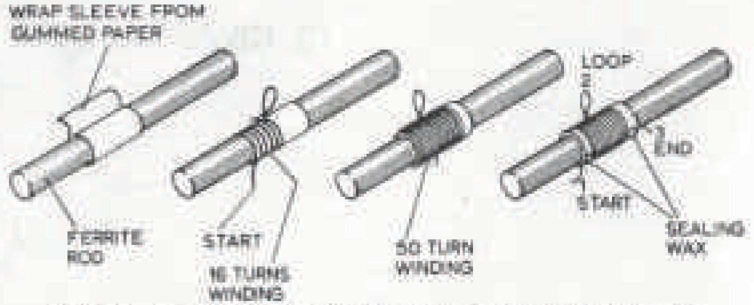

Fig. 3.1 Aerial tuning coil wound from 38-s.w.g. wire on a $\frac{3}{8}$-in. diameter ferrite rod.

Fig. 3.1 details the construction of a suitable coil for a tuned circuit, matched to a ferrite rod of $\frac{3}{8}$ in. diameter and about 4 in. long (see Chapter 4 for alternative coil windings on different rod sizes). If a rod is purchased longer than 4 in., it can be used as it is, or cut down by marking around with a file and then breaking off the surplus length.

Cut seven 1-in. lengths of gumstrip. Moisten one and wrap around the rod gummed side up. Now add about another half a dozen wrappings of similar length over the first, this time with gummed side *down* to form a reasonably rigid tube. Make sure that the paper tube is a *sliding* fit on the ferrite rod and leave to dry thoroughly (preferably removed from the rod so that it cannot become stuck to it).

When the paper tube is quite dry it should be rigid, when the coil windings can be applied. The wire to be used is 38-s.w.g. enamelled copper wire, the number 38 referring to the actual diameter of the wire according to the standard wire gauge (s.w.g.).*

Starting about $\frac{1}{2}$ in. in from one end of the paper tube, wind the wire carefully round the tube, with each turn tight against the one before it, until sixteen full turns have been completed. Then make a loop in the wire, as shown, and carry on winding, with succeeding turns touching, until fifty turns in all have been completed. The two loose ends of the coil (the start and finish) can be secured with a dab of sealing wax whilst the projecting loop can be twisted together (e.g. by putting a pencil through the loop and twisting up). Cut off the loop, leaving about $\frac{1}{2}$ in. protruding from the main coil, bare the wire ends and solder together. This forms point 2 on the coil; the start is point 1, and the end point 3—see Fig. 3.1. It will be easy to remember these

---

* American wire gauge (AWG) is about two sizes smaller than the British standard wire gauge (s.w.g.). For example, 38 s.w.g. is approximately the same as 36 AWG.

without marking since the loop or tapping point (2) comes much closer one end (1) than the other (3).

Cut a panel of *Paxolin sheet* to about the size shown in Fig. 3.2, using a hacksaw.* On this secure a *tag strip*, as shown, and drill a hole to mount a miniature or small-size 500 pF *variable capacitor*.

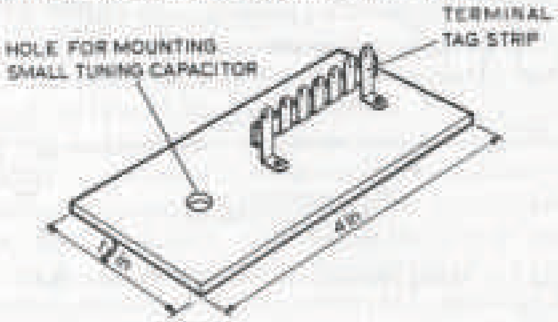

Fig. 3.2 Paxolin panel and tag strip for crystal set circuits.

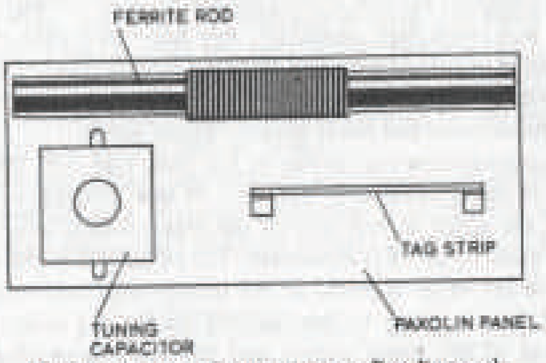

Fig. 3.3 Layout of components on Paxolin panel.

The aerial coil is then mounted on the panel as shown in Fig. 3.3, gluing the coil on to the Paxolin with two or three dabs of sealing wax, or some other suitable adhesive. *Note*: the ferrite rod must be free to slide in the paper tube for 'tuning' adjustments.

Virtually any miniature germanium or silicon diode will be suitable for the detector. Recommended types, which are readily available, are 1N34 and 1N914.

Earphones must be of *high-impedance* type, which need not be expensive to buy, and the higher the impedance the better the reception.

---

* A phenolic or fiberglass sheet may also be used.

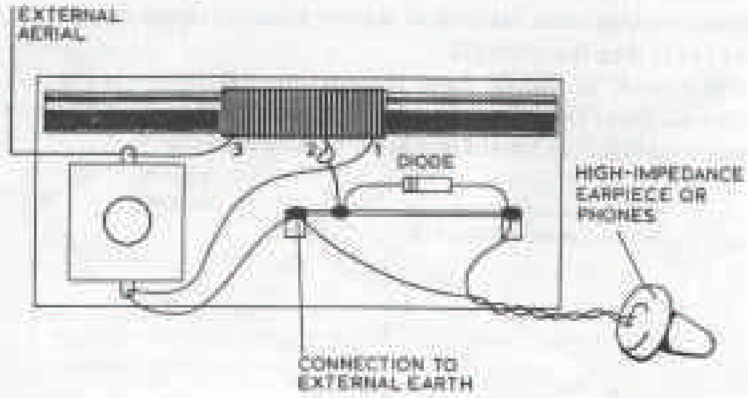
EXTERNAL AERIAL
3
2
1
DIODE
HIGH-IMPEDANCE EARPIECE OR PHONES
CONNECTION TO EXTERNAL EARTH

Fig. 3.4 Wiring connections to complete the crystal set.

Alternatively a deaf-aid type earpiece can be used; this will not give the same volume or quality of reproduction as headphones, but is a less expensive component. This should preferably be of the *high-impedance magnetic* type, with high sensitivity. Any high-impedance earpiece will suffice, but if of crystal type will require a resistor, connected across it, to complete the circuit. This will reduce the amount of current flowing through the earpiece and lower the strength of signal.

Wiring connections are shown in Fig. 3.4. End 3 of the aerial coil (the end of the 50-turn coil) connects to one terminal of the tuning capacitor, the aerial or 'hot' end of the tuned circuit, and the point to which an external aerial is connected. The other end of the coil (end 1) connects to the other terminal of the tuning capacitor, from which an additional wire is taken to the first tag on the tag strip. This is the 'earthy' end of the tuned circuit, and the point to which an external earth is connected. Leave plenty of slack wire between the coil and tuning capacitor.

The other connections are then as follows:

(i) Tapping point of the coil wire bared and connected to the second tag.
(ii) The diode also soldered to this same tag, and to any other free tag.
(iii) Headphone (or earpiece) leads to the 'earthing' tag, and to the 'free' tag to which the above diode has been connected. All connections should be made with *soldered* joints.

The wiring-up can be checked against the *circuit diagram* shown in Fig. 3.5 (ignoring the components shown with broken lines). The set should now be 'working'.

In areas of strong signal strength, no external aerial or earth connections should be necessary. Performance will, however, be improved in any area by attaching an aerial wire (which can be any thin wire, e.g. using the same wire as for the coil winding), of up to 160 feet. The longer the aerial the better the reception, provided it is led away from the receiver to as *high* a point as possible.

An earth connection may further improve aerial performance; by this we mean a connection to some conductor positively in contact with the ground (preferably buried), an excellent example being a metal water-pipe. Thus, if an earth connection is found to be necessary (or you want to try one to see how performance is affected), connect a wire from the 'earthing' tag on the receiver to a convenient water-pipe.

This question of obtaining a *good* aerial and earth is a most important one in areas of poor signal strength. Linking up to a television aerial is often a good plan, since TV aerials are also usually mounted as high as possible. If bare wire is used, it is also important that the upper (free) end of the aerial is not made fast to something which could produce an earth connection (e.g. a damp tree), or at least is suitably insulated from such a support. String is not an efficient insulator; that, too, can conduct when wet.

Quite good results are often obtained by using the springs of a bed as an aerial, in which case an earth connection is usually not necessary. Sometimes, too, when other attempts to yield a good signal strength in

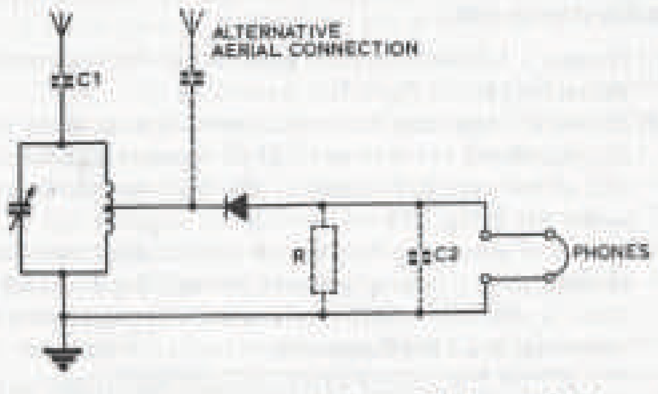

Fig. 3.5 *Circuit 1.* Circuit diagram of basic crystal set.

the aerial have failed, connecting the *aerial* side of the tuning coil to a *good earth* (a water-pipe) can produce better results, the normal earth connection being left off.

Tuning

The receiver is adjusted as follows. Turn the tuning capacitor to fully close the vanes, then open about half a turn on the spindle or knob. (If you are using a trimmer as a tuning capacitor, screw right down and then open half a turn.)

The tuning coil should now be slid up and down the ferrite rod (the coil leads were left a fairly long time to give the necessary freedom of movement) until BBC Radio 3 is heard.*

It may be necessary to slightly alter the adjustment of the variable condenser to tune in to this programme. Also, because of the ferrite-rod aerial the set will be directional, that is, the signal strength received will depend to some extent on the direction in which the aerial rod is pointing, so position the set to pick up the maximum volume.

Having established the best position of the tuning coil on the rod to receive BBC Radio 3, fix permanently with a dab of sealing wax. You should then find it possible to tune in to further stations by altering the setting of the variable capacitor–e.g. typically Radio 4 in about the middle of the capacitor travel and Radio 1 towards the other end.

Any reception you get will almost certainly be very weak and (unless you live close to a broadcast station) you can really feel satisfied if you get any station at all at audible strength. But it is surprising how, sometimes, even quite distant stations can be heard. Also you can often improve the reception and listening strength by quite simple modifications. Try these in order:

(i) Connect a 1,000-pF capacitor across the headphone (earpiece) connections (C2 in Fig. 3.5).

(ii) Instead of connecting the external aerial directly to the tuned circuit, connect one lead to a 220-pF capacitor, and the other end of the capacitor circuit to the 'hot' end of the tuned circuit (C1 in Fig. 3.5).

(iii) Instead of connecting the external aerial to the tuned circuit, connect to the tapping point of the coil (tag to which the diode is also connected). Try a direct connection, and also connecting in a 220-pF capacitor.

---

* Tune to an AM broadcast station near the lower end of the dial—close to 550 kHz.

(iv) Try connecting a 1.2-k resistor (or higher value) across the phone connection (R in Fig. 3.5). You may be using the wrong type of phones or earpiece, which do not provide a proper load or complete the circuit.

If there is a complete lack of response, check for faulty wiring-up. A more likely cause, however, is lack of an external aerial or earth connection in an area where these are strictly necessary for adequate reception; or an inefficient aerial (too short) or poor earth connection (bad electrical contact to a good earth point, or connection to a bad earth point).

Another possible cause of apparent failure may be too much outside noise entering the ear so that it is impossible to detect the very weak radio signal as it is being tuned in. Headphones are better than a single deaf-aid type of earpiece in this respect but, in any case, a really quiet room is virtually essential for initial setting up and tuning adjustments. Also, if your adjustment of the tuning control is too coarse, you may completely miss the setting for the station you are looking for, without realizing it.

Reception will also tend to vary with weather conditions. Some days it may be so poor that what was normally a strong station is hardly heard at all. The simple basic receiver has many limitations but, since it costs very little to construct and nothing at all to operate, this must be regarded as inevitable.

Providing you can hear something–even if too weak a signal to distinguish properly–you can certainly improve the performance of your basic set by further experimentation with tuned circuits (see Chapter 4) and/or the addition of amplification to the circuit. You can also try other types of basic crystal set, as described in the following projects.

Circuit 2 (Fig. 3.6) is identical to *Circuit 1* except that, instead of a diode, a transistor is used as a detector. Only two of the transistor leads are connected–the emitter (*e*) connection to the tapping part of the coil, and the base (*b*) to the 'earthy' end of the circuit. The collector lead of the transistor is ignored (bend it out of the way so that it cannot accidentally short out the other leads).

You can try almost any type of low-cost AF transistor; recommended types are OC42 2N370, 2N2925, 2N5088.

With the addition of two more components, Circuit 2 can be modified to work the transistor both as a detector and an amplifier, to give stronger signals through the headphones. Using the same transistor type

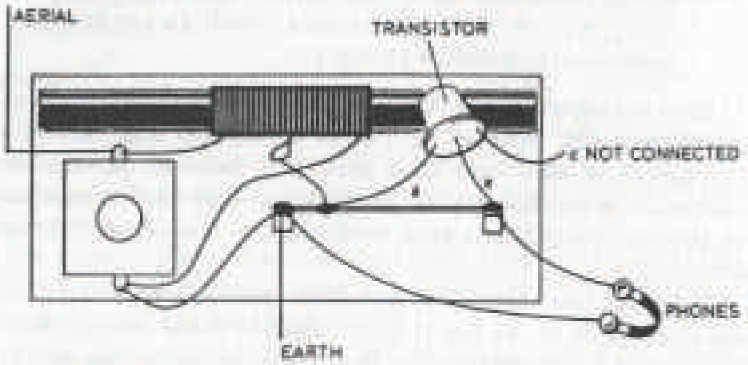
AERIAL
TRANSISTOR
e NOT CONNECTED
b
c
PHONES
EARTH

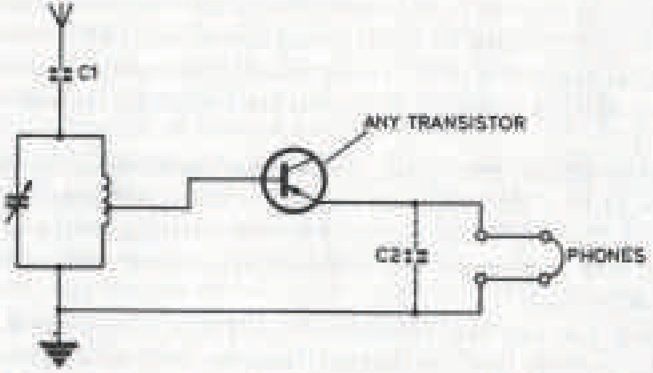
C1
ANY TRANSISTOR
C2
PHONES

Fig. 3.6 *Circuit 2.* Crystal set using transistor instead of a diode.

as above (or near-equivalent) resistor R should be 15 k and capacitor C3 1 µF or higher.

This time a battery is also required to supply power for the transistor to work as an amplifier. This can be from 1.5 V up to 9 V. Remember the rules for polarity of connection: those shown in Fig. 3.7 are for a p-n-p transistor; an n-p-n transistor would need the battery connected the opposite way round. Battery polarity also affects the connections of capacitor C3 (if an electrolytic or polarized type).

Experiment further by trying the effect of using additional capacitors in the circuit, e.g. C1, 220 pF; C2, 0.001 µF (try other values as well); C4, 0.001 µF (try other values as well).

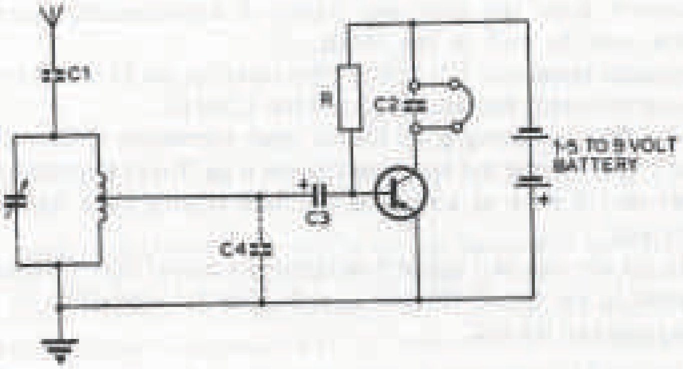
C1
R
C2
1·5 TO 9 VOLT BATTERY
+
+
C3
C4

Fig. 3.7 *Circuit 3.* Crystal set with amplification.

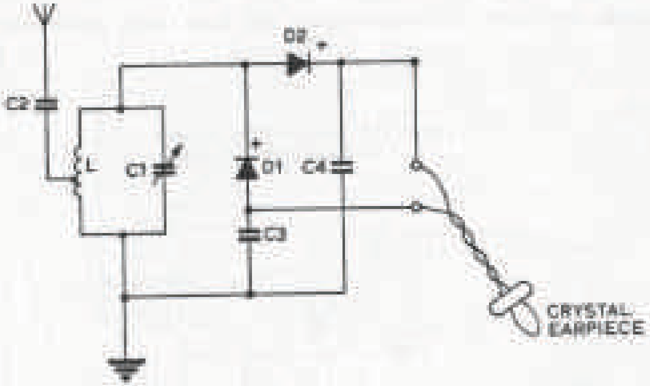
D2
+
C2
L
C1
+
D1
C4
C3
CRYSTAL EARPIECE

Fig. 3.8 *Circuit 4.* Double diode crystal set.

This circuit incorporates 'voltage doubling' to improve the signal volume and should give better performance than a single diode circuit. L and C1 are the usual tuned circuit, but an unusual feature is that the *aerial* is connected to the coil tapping point. Any type of germanium diodes can be used (they should preferably be the same), making sure to connect them the right way round. A high-impedance *crystal* earpiece must be used in this circuit.

Capacitor values are: C2, 220 pF (this capacitor can be omitted–try with and without in the circuit); C3 and C4, 1,000 pF.

The optimum tapping point for the aerial connection to the coil L is best found by trial and error (see Chapter 4 on 'Tuned Circuits'), but the set should work at a nominal one-third tapping point from the 'earthy' end.

This set can also be tried with conventional tuned-circuit coupling–i.e. aerial to the top of the coil L, and diode D2 connection to the tapping point on the coil.

4 MORE ABOUT TUNED CIRCUITS

The basic combination of a coil of certain *inductance* connected in parallel with a *variable capacitor* is the form of tuned circuit used in most radio receivers, from simple crystal sets upwards. The *tuning range* of such a circuit is determined by the respective values of inductance and capacitance used. At the same time, the *resistance* of the coil can have a significant effect on the *selectivity* of the circuit, so that a particular broadcast frequency is tuned into sharply, with other nearby frequencies rejected.

In the simple tuned circuit used for the crystal set of Chapter 3, the aerial is connected to the upper or 'hot' end of the coil, the other end of the coil being earthed; the diode is connected to a tapping point on the coil. Performance may well be improved by adjusting both the aerial-connecting point and the diode-connecting point, as shown in Fig. 4.1.

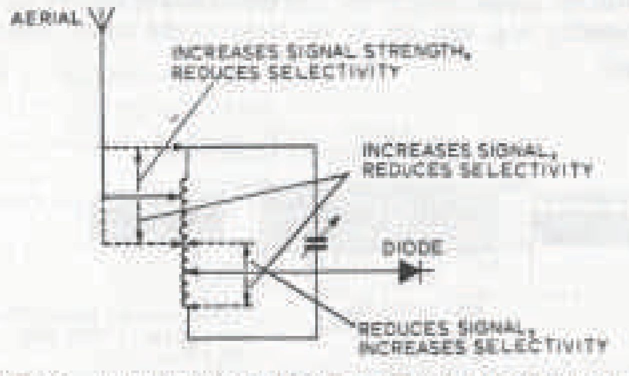
AERIAL
INCREASES SIGNAL STRENGTH, REDUCES SELECTIVITY
INCREASES SIGNAL, REDUCES SELECTIVITY
DIODE
REDUCES SIGNAL, INCREASES SELECTIVITY

Fig. 4.1 The effect of varying the tapping points on the aerial coil.

Connecting the aerial to the 'hot' end should give the loudest signal, but the *selectivity* of the circuit may be poor because of the coil *resistance* involved in the actual circuit to earth. Connecting the aerial to a tapping point lower down the coil (towards the 'earthy' end) should improve selectivity, but reduce signal strength. Experimentation will produce the optimum aerial-connecting point. There will also be an optimum tapping point for the diode.

Finding optimum tapping points for the aerial and diode connections is a bit tricky on a close-wound coil since each tapping point tried has to be bared by scraping off the enamel insulation, with the risk of producing shorted turns. The same *effect* can be provided by removing, or adding, turns at each end of the coil, using the original tapping point for the diode connection–e.g. see Fig. 4.2. Unless the *total* number of turns on the coil remain the same, however, the inductance of the complete coil will be altered, and thus the tuning range of the tuned circuit (see later). Then the *diode* tapping point will be shifted.

For this type of experiment an air-cored coil of relatively large diameter is easier to use, wound on a card or paper tube to which two $\frac{1}{4}$ in.^{3} lengths of hardwood or balsa strip have been cemented. The windings can be bared by scraping or sandpapering over this raised section, and individual wires carefully separated to avoid shorting–Fig. 4.3. Optimum tapping points can be established by pushing bared ends on the aerial- and diode-connecting wires into the raised section of the coil, soldering in place very carefully once they have been found. This should not result in shorting out more than a few turns at these points; which will not greatly modify the coil inductance if allowed for in the initial winding (e.g. add five or six turns to the design number of windings).

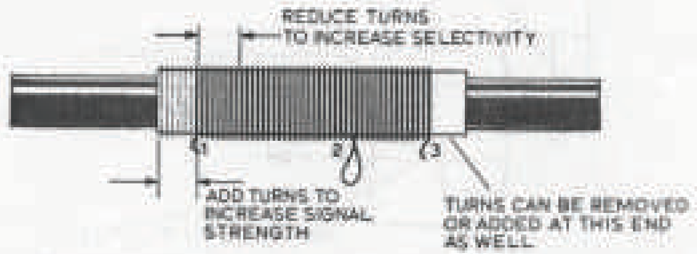

Fig. 4.2 The effective tapping point on a simple coil can also be varied by adding or removing turns on each end of the winding.

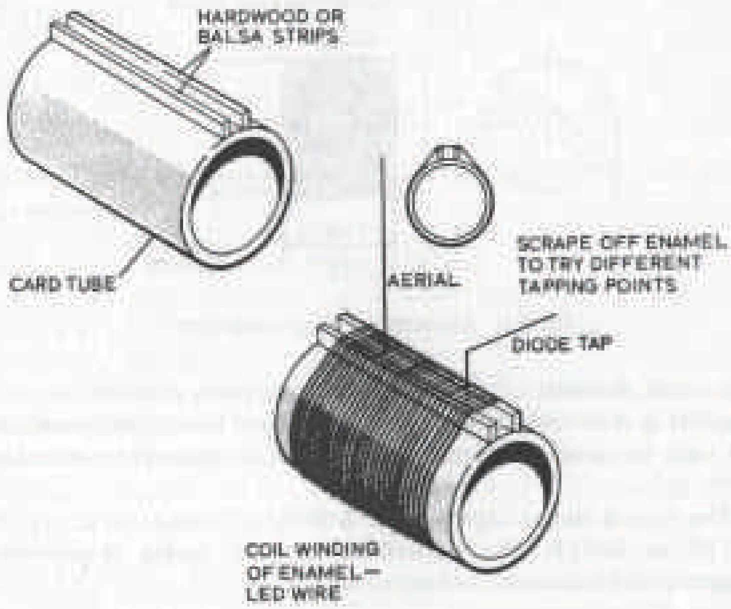

Fig. 4.3 Construction of an air-cored aerial coil.

In air-cored coils, the inductance can be calculated with reasonable accuracy from the diameter and number of turns (see also Fig. 4.4):

$$\text{L. inductance, microhenries} = \frac{R^2 \times N^2}{9R + 10l}$$

where R = coil radius in inches,
l = coil length in inches,
N = number of turns.

The more usual form of this formula is:

$$N = \sqrt{\left(\frac{(9R + 10l) \times L}{R^2}\right)}$$

(A worked-out example using this formula is given later.) This is based on the turns of the coil being close-wound, in enamelled copper wire.

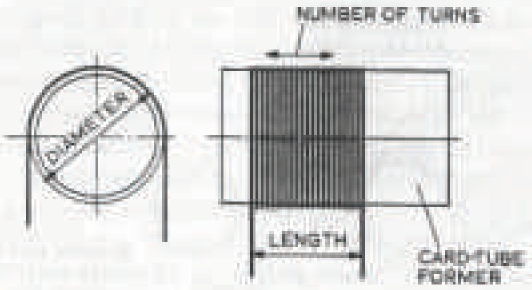

Fig. 4.4 Air-cored coil design parameters.

The actual diameter of the wire is not significant, provided the coil diameter is reasonably large (e.g. greater than 1 inch). The formula is not valid for smaller-diameter air-cored coils, or for coils wound on a ferrite rod.

The typical tuning capacitor has a rating of (about) 50 pF up to 350 pF or 500 pF. The relationship between tuning or *resonant frequency* and inductance and capacitance is:

$$\text{resonant frequency } (f) = \frac{1}{6.28\sqrt{(LC)}} \times 10^6 = \frac{0.16}{\sqrt{(LC)}} \times 10^6$$

where L = inductance in microhenries (μH),
C = capacitance in picofarads (pF).

This is more conveniently written as a solution for the *product* of inductance and capacitance required, i.e.

$$LC = \frac{0.025}{f^2} \times 10^{12}$$

For the medium waveband the range of frequencies to be covered is 500 to 1,500 kHz.* Thus, to encompass this range, the products of LC required are:

at 500 kHz
$$LC = \frac{0.025}{(500)^2} \times 10^{12}$$
$$= 0.001 \times 10^8 \text{ (approx.)}$$

---

* In the U.S.A. the AM broadcast band is from 530 to 1600 kHz.

at 1,500 kHz
$$LC = \frac{0.025}{(1{,}500)^2} \times 10^{12}$$
$$= 0.0001 \times 10^8 \text{ (approx.)}$$

At one end of the capacitor tuning range where C = 50 pF, inductance L required is:

$$\frac{0.001 \times 10^8}{\cancel{50}\ 500}$$
$$= 200\ \mu\text{H}$$

At the other end of the capacitor tuning range where C = 500 pF, inductance L required is:

$$\frac{0.0001 \times 10^8}{\cancel{500}\ 50}$$
$$= 200\ \mu\text{H}$$

Using a 1-in. coil diameter (0.5-in. radius) and a coil length of 1.5 in.

$$N = \sqrt{\left(\frac{(4.5 + 15) \times 200}{0.5^2}\right)}$$
$$= \sqrt{15{,}600}$$
$$= 125$$

In other words, a 125-turn air-cored coil of this size should be about right for covering the medium waveband, matching a 50–500 pF variable capacitor. Tapping point (or diode connection) would nominally be one-third of the coil length from the 'earthy' end.

Air-cored coils can also be designed in this way for long-wave coverage (more turns); or short-wave reception (less turns), which is not likely to be so good because of the low efficiency of the coils. Also, at higher frequencies the resistance of the *capacitor* is also significant, and so simple calculation of coil size is no longer valid.

The efficiency of a tuning coil can be expressed in terms of the magnification of the original signal received, when tuned to resonance, and decreases with increasing coil resistance. The magnification produced in a resonant circuit is referred to as the 'Q'; a high-Q coil is thus desirable for maximum performance and selectivity, and is essentially a low-resistance one. For this reason larger diameter air-cored coils, using thicker wire, are more effective than smaller diameter coils wound from thinner wire (since wire resistance increases with decreasing diameter).

Equally, the inductance of a wound coil can be increased by winding it over a magnetic-iron core, such as a ferrite rod. Much smaller diameter coils can be used, requiring less turns, to produce the inductance required—both reducing the length of wire and its resistance.

Because the actual inductance produced is dependent on the size and characteristics of the core material, windings for ferrite-rod aerials cannot be calculated, and are based on empirical results, or optimum windings established by trial and error.

Common sizes of ferrite rod available are:

$\frac{1}{4}$-in. diameter	length $3\frac{1}{2}$ in.
$\frac{5}{16}$-in. diameter	lengths 4, 5, 6 and 8 in.
$\frac{3}{8}$-in. diameter	lengths 4, 5, 6, 8 and 10 in.

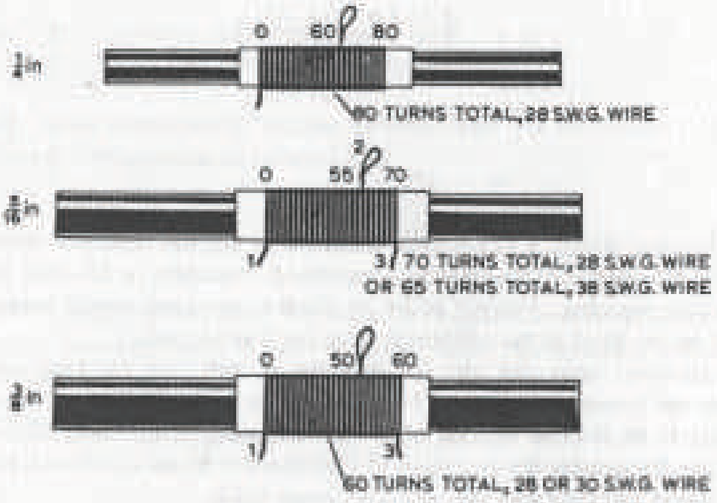
$\frac{1}{4}$in
0
60
80
80 TURNS TOTAL, 28 S.W.G. WIRE
$\frac{5}{16}$in
2
0
55
70
1
3
70 TURNS TOTAL, 28 S.W.G. WIRE
OR 65 TURNS TOTAL, 38 S.W.G. WIRE
$\frac{3}{8}$in
0
50
60
1
3
60 TURNS TOTAL, 28 OR 30 S.W.G. WIRE

Fig. 4.5 Medium-wave aerial coil windings on three sizes of ferrite rod.

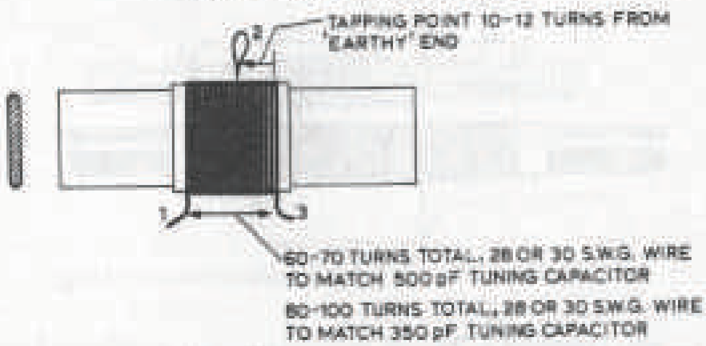
2
TAPPING POINT 10–12 TURNS FROM 'EARTHY' END
1
3
60–70 TURNS TOTAL, 28 OR 30 S.W.G. WIRE TO MATCH 500 pF TUNING CAPACITOR
80–100 TURNS TOTAL, 28 OR 30 S.W.G. WIRE TO MATCH 350 pF TUNING CAPACITOR

Fig. 4.6 Medium-wave winding for a ferrite slab.

A long length enables maximum adjustment of coil position (a way of 'trimming' the aerial) and, also, additional coils to be incorporated on the same rod, if the circuit needs them (e.g. for multi-waveband coils). But it is the diameter which really determines the winding specification.

Recommendations for medium-wave coil windings for these three standard ferrite-rod diameters are given in Fig. 4.5, all windings being enamelled copper wire. Normal tapping points are one-third from the 'earthy' end. It is a simple matter to experiment with different windings—e.g. adding more turns to extend the *wavelength* upwards (i.e. lower frequencies), or reducing the number of turns to extend the *wavelength* coverage downwards (i.e. for higher-frequency transmissions).

Ferrite-core material is also available in slab form, which is often easier to accommodate in a small receiver case. A typical ferrite-slab size is $2\frac{3}{4}$ in. long by $\frac{3}{4}$ in. wide, with a thickness of $\frac{1}{8}$ in. or less. A typical winding specification is shown in Fig. 4.6.

Generally, the long waveband should be covered by *doubling* the number of turns used for the medium waveband. *Reducing* the number of turns shifts the coverage out of the medium waveband down into the trawler band (80 m) and 'top-band' (160 m). For short-wave coverage very few turns indeed will be needed—perhaps less than half a dozen—and the optimum number can only be determined by careful experiment.

Separate coils for the various wavebands to be covered can be accommodated on the same ferrite rod and connected to a common tuning capacitor via a waveband switch—Fig. 4.7. This is the usual

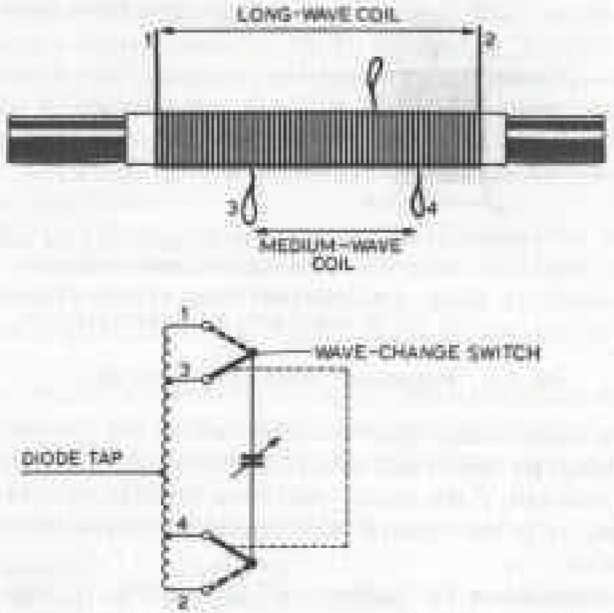
LONG-WAVE COIL
1
2
3
4
MEDIUM-WAVE COIL
1
WAVE-CHANGE SWITCH
3
DIODE TAP
4
2

Fig. 4.7 Long- and medium-wave aerial coil connected to wave-change switch.

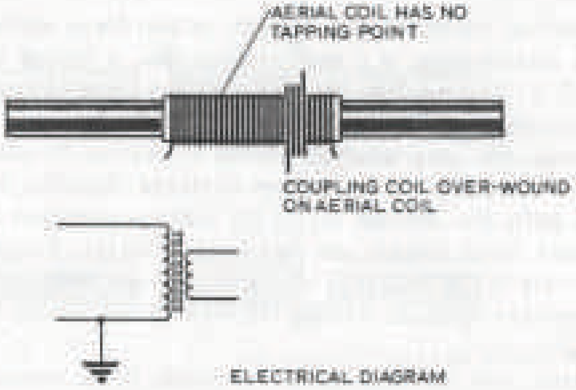
AERIAL COIL HAS NO TAPPING POINT
COUPLING COIL OVER-WOUND ON AERIAL COIL
ELECTRICAL DIAGRAM

Fig. 4.8 Inductively-coupled aerial coil.

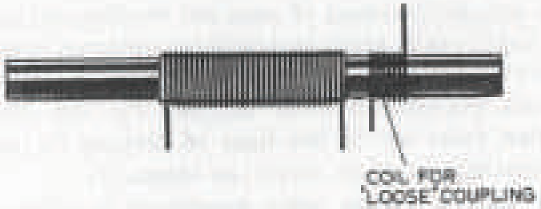
COIL FOR 'LOOSE' COUPLING

Fig. 4.9 'Loose' inductive coupling with separate coupling coil.

procedure with an elaborate receiver circuit and domestic radios, but is seldom satisfactory with the simpler circuits because of the tuned-circuit characteristics (i.e. the need for meticulous adjustment to get optimum results for any single waveband coverage).

All the coil constructions so far described are based on single windings (covering a specific waveband) with a tapping point for the output (diode connection in the case of simple receivers). The alternative form is to provide *inductive coupling* for the output in the form of an overwinding, which can comprise five or more turns of the same gauge wire used for the main winding–Fig. 4.8. The optimum number of turns for the overwinding depends on the degree of coupling required, which in turn is influenced by the characteristics of the following circuit. The position of the coupling coil can also affect the performance, although usually it is best positioned over the 'earthy' end

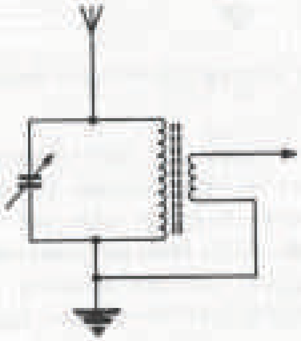

Fig. 4.10 Transformer coupling between tuned circuit and next stage of receiver.

of the main coil. Both number of turns and coupling-coil position are subjects for experimental adjustment to obtain optimum results.

Sometimes a coupling coil is wound on the ferrite rod, separate from the aerial coil, providing a 'loose' coupling–Fig. 4.9. Regenerative receivers often make use of this form of coupling for introducing 'feedback' into the front of the circuit (see Chapter 9).

Inductive coupling is also called *transformer* coupling, since the principle of passing current generated in one coil into another coil is the same. With transformer coupling (Fig. 4.10) the aerial coil has no tapping point, and is an alternative to direct coupling, where the next stage of the circuit is connected directly to a tapping point on the coil.

There is a third method of coupling which may be used between stages in a radio circuit–*capacity coupling*. Here, connection between the stages (e.g. to the tapping point on the aerial coil) is made through a capacitor, which isolates the two stages as far as d.c. flow is concerned (only a.c. will pass through the capacitor)–Fig. 4.11.

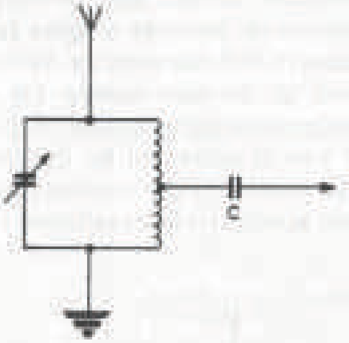

Fig. 4.11 Capacity coupling between tuned circuit and next stage of receiver.

Again the optimum value for the capacitor depends on the characteristics of the circuits being coupled. To couple an aerial coil to the next stage in a receiver, optimum values can vary from 200 pF to 10 µF. In general, though, a high value gives best results.

Direct, transformer, and capacity coupling are alternatives for connecting any two stages in receivers–not just the tuned circuit to the next stage.

5 AMPLIFIERS

It is a characteristic of a transistor that it works as an *amplifier* of signals. In the most widely used mode of connection of transistors for such duties with input and output circuits both connected to the emitter (common-emitter mode), the degree of amplification or current gain is called the 'beta' (β) of the transistor. This is given as h_{FE} (or static forward current ratio) in transistor characteristics data (see also Chapter 7).

A basic transistor amplifier circuit is very simple–and is identical for a p-n-p or n-p-n transistor, except for the battery polarity–see Fig. 5.1. Virtually any low-power AF transistor can be used in this circuit. The *bias resistor* (R) must have a value providing a collector current not exceeding the maximum specified rating for the transistor used, the actual current flowing in the collector circuit also being influenced by the voltage of the battery.

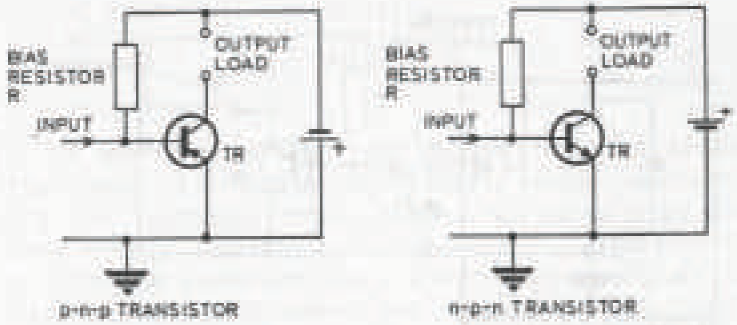

Fig. 5.1 Basic amplifier circuits with simple current biasing via a single resistor.

Knowing the transistor characteristics, a suitable value for R can be calculated as follows:

$$R = \text{gain} \times \frac{\text{battery voltage}}{\text{collector current}}$$
$$= h_{FE} \times \frac{\text{battery voltage}}{I_c}$$

where I_c is equal to, or preferably less than, the specification figure for I_c max.

Assuming that an OC72 transistor is being used, and the battery voltage is 9 V, specified figures are:

$$I_c \text{ max} = 250 \text{ mA}$$
$$h_{FE} = 30 \text{ to } 90$$

Taking the maximum gain, and 2.5 mA as a 'safe' working figure for the collector current:

$$R = 90 \times \frac{9}{0.0025}$$
$$= 324\text{k}\Omega$$

A suitable (preferred-value) resistor would thus be 330k.

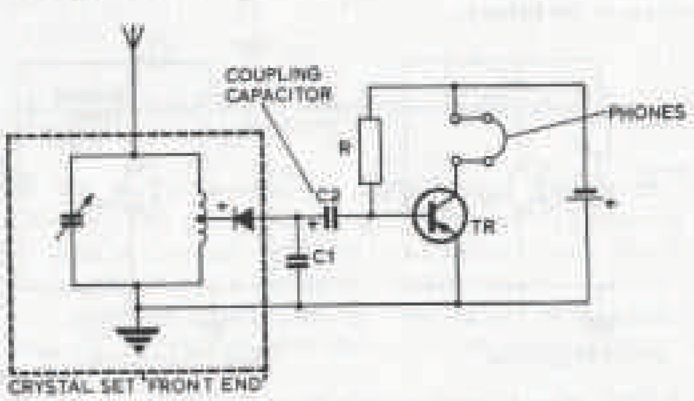

Fig. 5.2 *Circuit 5.* Crystal set with single stage of amplification. Any ~~RF~~ transistor can be used for TR.
AF

Fig. 5.2 shows the complete circuit for a simple AF amplifier of this type, coupled to the 'front end' of a basic crystal set. The detector (output) is coupled to the amplifier via capacitor C2, a suitable value for which would be 10 μF or higher (e.g. 20, 25, 30, 40 or 50 μF); capacitor C1 (0.001 μF) may not be necessary. The output load in the collector circuit is formed by high-impedance phones.

Note that the *polarity* of the battery used to power this circuit is important (as far as the transistor connection is concerned); also the connections to the *diode* and C2 (which will normally need to be an electrolytic type to provide the high-capacity value required). With an n-p-n transistor the battery polarity would be reversed; and also the diode and electrolytic capacitor connections.

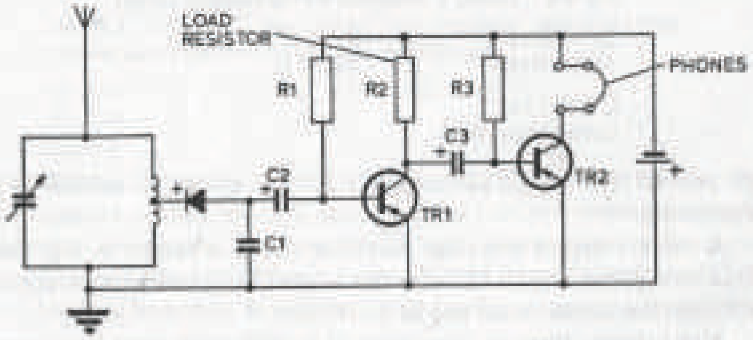

Fig. 5.3 *Circuit 6.* Crystal set with two stages of amplification. Component values matching OC71, OC72 or near equivalent for TR1 and TR2:

R1, 470 k
R2, 4.7 k
R3, 470 k

C1, 0.001 μF
C2, 8 or 10 μF
C3, 8 or 10 μF

Battery voltage 6 to 9 V.

The same type of circuit can be used to provide additional amplification, if required, simply by adding another amplifier stage (Fig. 5.3). The second amplifier stage can be identical to the first or based on a higher-power transistor taking the higher output present and providing even greater gain. The load in the output (collector) of the first amplifier stage is provided by a resistor (R2) (which should be about the same value as the phone resistance, 3.3 to 4.7 k), and the two stages coupled by a capacitor (C3). Value of R3 depends on the second transistor used, and may be anything from 47 Ω to 470 kΩ; capacitors

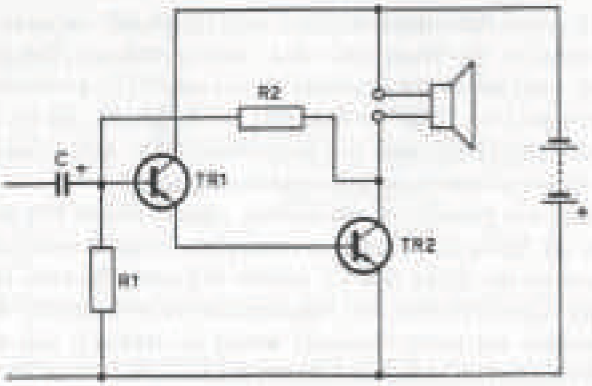

Fig. 5.4 *Circuit 7.* Amplifier for loudspeaker output

TR1, OC72 R1, 6.8 k
TR2, AD140 R2, 33 k
C, 8 or 10 µF
Loudspeaker, 80 Ω

C2 and C3 can be 10 µF or larger (20 µF up to 50 or 100 µF), electrolytic.

A more compact two-stage amplifier circuit is shown in Fig. 5.4. This may prove capable of operating a small 80 Ω loudspeaker direct, although the current drain will be quite high.

Alternatively, three or four stages of amplification using low-power transistors, following a basic 'front end' crystal set, should provide enough power to drive a small loudspeaker at the final output, through a suitable step-down transformer, to provide an impedance match (see Chapter 6).

Simple amplifier circuits of this type have the important limitation that performance of the transistor(s) will tend to vary with temperature. There is also the possibility of 'thermal runaway' developing, which can destroy the transistor(s), because as the external temperature increases the collector current also tends to increase, which in turn causes a further increase in junction temperature, so the effect is cumulative and goes from bad to worse, even to the point of ruining the transistor completely. It is possible to overcome this trouble by arranging for the circuit to be self-biasing or d.c. stabilized so that a constant operating collector current is provided, regardless of transistor type or temperature variations. In other words, the working point of the collector circuit is stabilized.

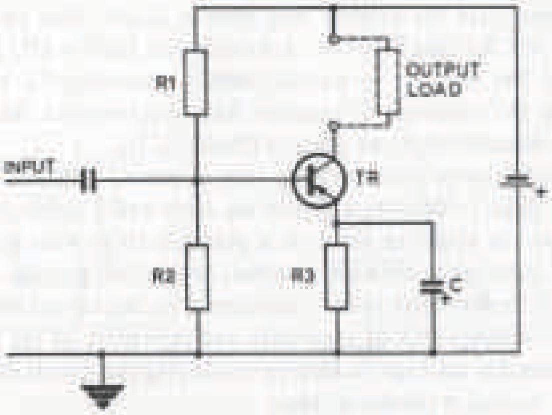

Fig. 5.5 Stabilized bias circuit for transistor amplifier stage. Typical values for low/medium-power transistors

R1, 22 k C, 8 or 10 µF
R2, 10 k
R3, 1 k

A further advantage of a stabilized circuit is that it makes the performance of the amplifier less dependent on the characteristics of individual transistors, which can differ appreciably even for the same type. Capacity coupling between stages should be used since this makes each stage independent as regards working, rather than inter-dependent.

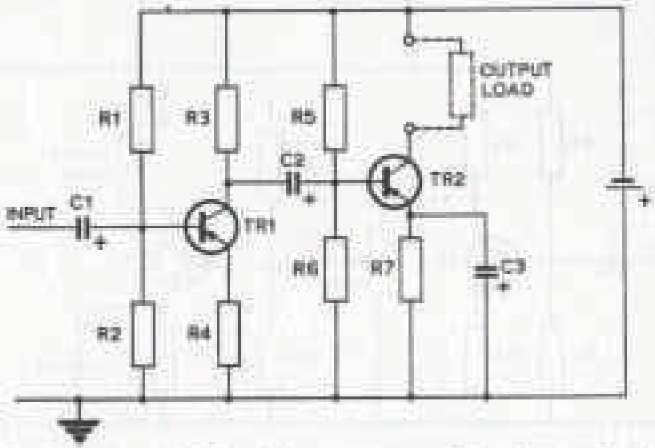

Fig. 5.6 *Circuit 8.* Practical two-stage amplifier using two low/medium-power AF transistors. (See Chapter 7 for determination of matching component values.)

To achieve this the original bias resistor is split into two separate values R1 and R2 (see Fig. 5.5). A further bias resistor (R3) is applied directly to the emitter, in parallel with a capacitor (C) to act as a by-pass for AF currents. (Transistor bias requirements, and data on transistor characteristics, are given in Chapter 7.)

A receiver design incorporating two stages of amplification with stabilized circuits following a crystal set front end is shown in Fig. 5.6. This, in fact, is about as far as it is practical to go with such a basic circuit as, although additional complete stages will provide more gain and greater final output power, deficiency in the circuit will also be aggravated–notably lack of sensitivity and selectivity in the front end. Such deficiencies are best tackled by modifying the circuit design from the simple 'crystal set detector' basis.

The amplifier stage(s) also provides a convenient point to insert a *volume control* into the receiver circuit. This takes the form of a potentiometer which can replace one of the bias resistors, or be placed in series with the flow of the current after the first stage of amplification–Fig. 5.7. This will introduce minimum distortion over the volume control range.

On low-power receivers which give satisfactory reception without an external aerial, a volume control can be an unnecessary refinement. Since the ferrite aerial has directional characteristics, merely turning the set one way and the other will produce changes in signal strength.

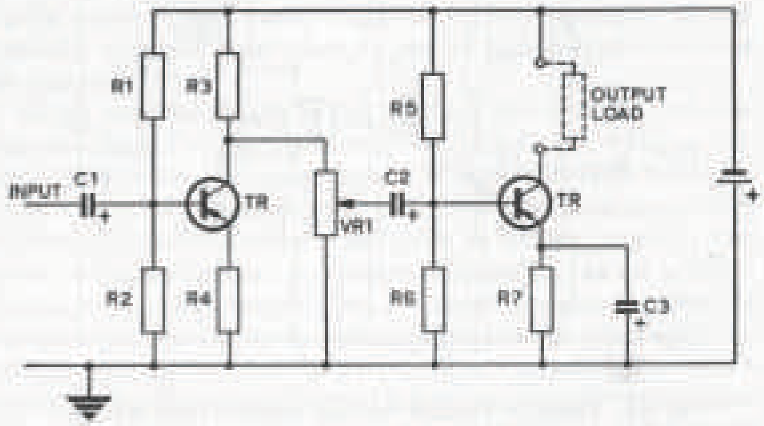

Fig. 5.7 *Circuit 9.* Two-stage amplifier with volume control.

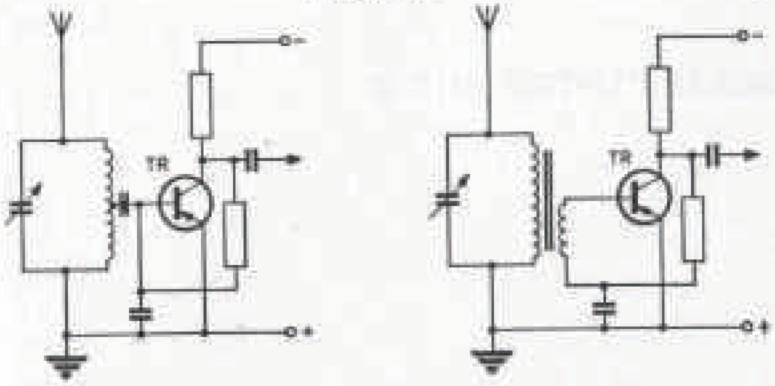

Fig. 5.8 Basic pre-amplifier circuits. The transistor (TR) must be an RF type; p-n-p transistors shown in these diagrams. If n-p-n transistors are used, the battery polarity must be reversed.

Pre-amplifiers

A pre-amplifier is designed to amplify signals between the tuned circuit and detector stages. The main difference is that the transistor(s) selected must be an RF type since it is handling RF signals at this stage. Otherwise the basic circuit is similar to that of an AF amplifier–see Fig. 5.8.

Phantom-Powered Active Loop Receive Antenna for 30 Metres

2010-03-27

This loop antenna was built for 30 metre QRSS reception, but tunes beyond 30 metres and might be useful for general narrow-band (fix-tuned) HF work.

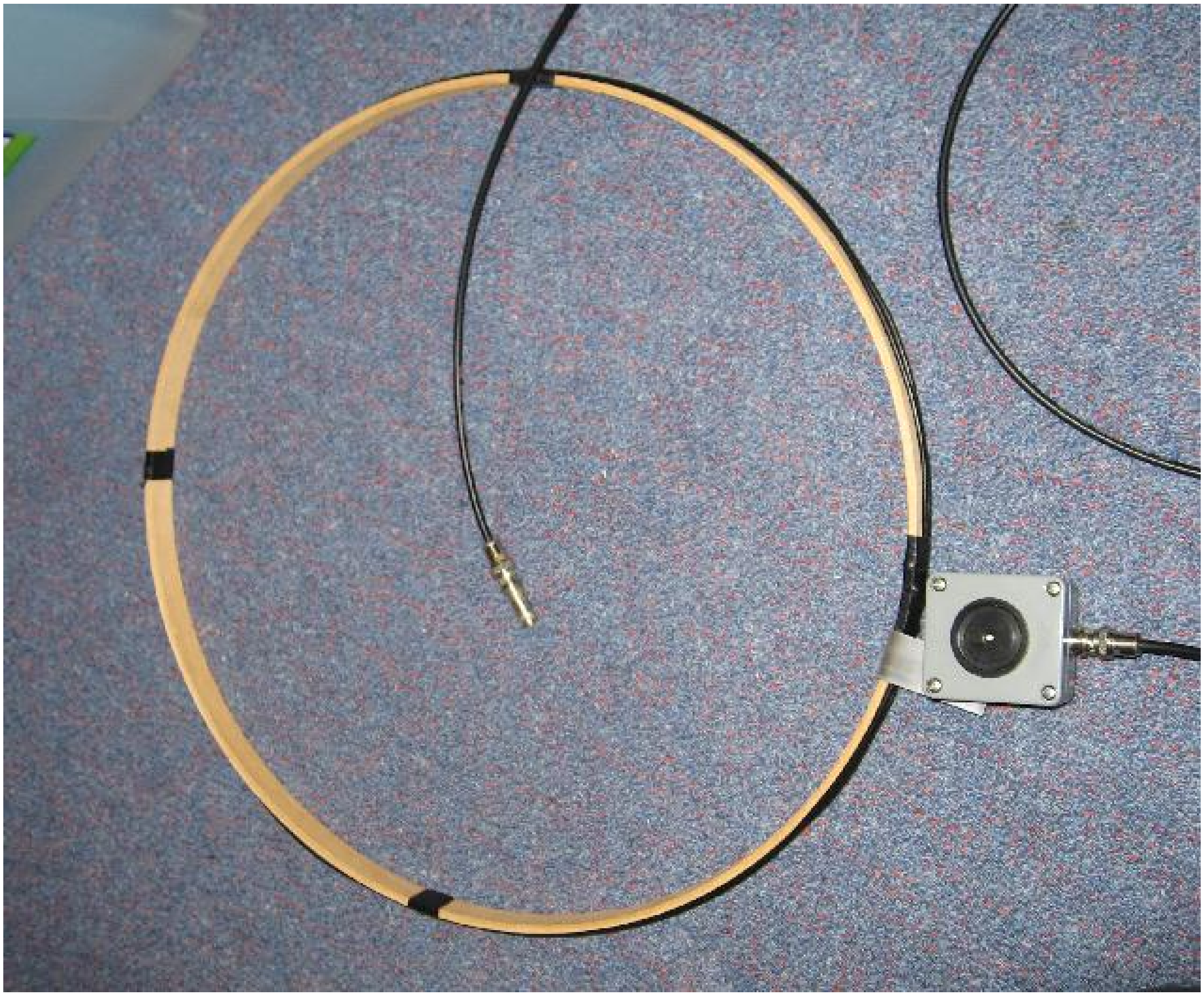

The Antenna

The loop is two turns of ~3 mm multi-strand hook-up wire, wound on a large (465 mm diameter) embroidery frame. The coil is centre-tapped, and referenced to "ground" at that point. A polyvaricon tunes the coil to resonance at the frequency of interest, and a push-pull JFET buffer amplifier transforms the very high impedance of the parallel resonant circuit down to something suitable for what is seen through the coax from the receiver (and also offers some power gain at the same time to offset feed-line loss). The buffer can deliver a relatively large amount of power, in excess of 0 dBm. This should offer good strong-signal handling, but I have not measured its IP3.

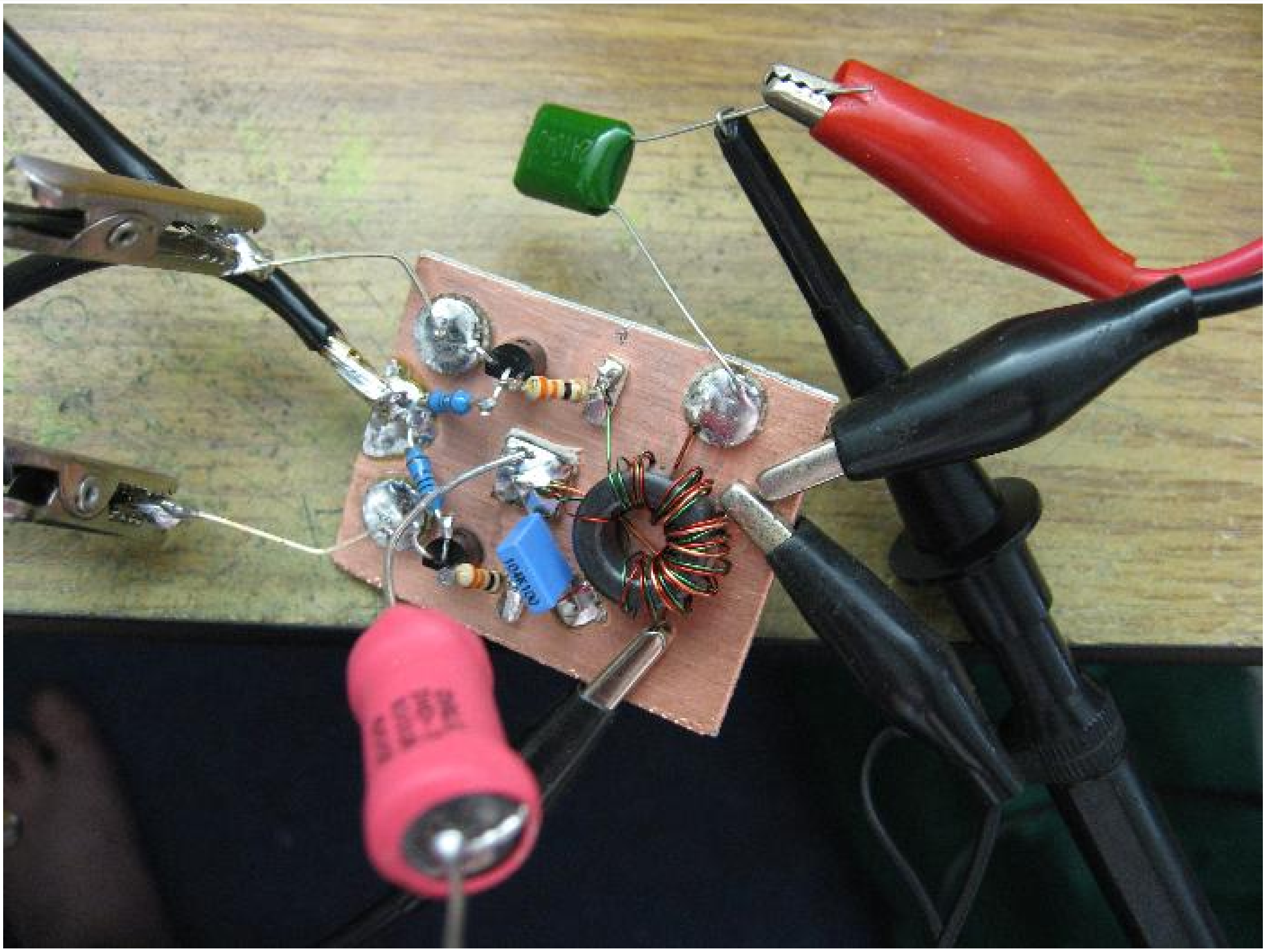

Each J310 stands about 2.5 mA. I hand-picked a pair of well matched J310s (by Vpp and Idss), this is likely unnecessary but it can't hurt to ensure each arm has similar biasing and gain to optimise the distortion cancelling effects. The 33 Ohm resistors in the drains help eliminate any tendency for spurious oscillation. The supply current feeds into the bifilar drain loads from the decoupled "cold"-end of the output winding, using it as a choke. This seems to work well in practice, with no visible distortion asymmetry (when over-driven) from the DC bias on the magnetics. (The standing currents in the bifilar winding cancel, but those in the output secondary do not, fortunately the supply current is only about 5 mA so this should be of no consequence. As long as the transistors saturate well before our ferrite core - an FT50-43 - we are in good shape.)

© 2000 - 2009 Alan Yates - All Rights Reserved.

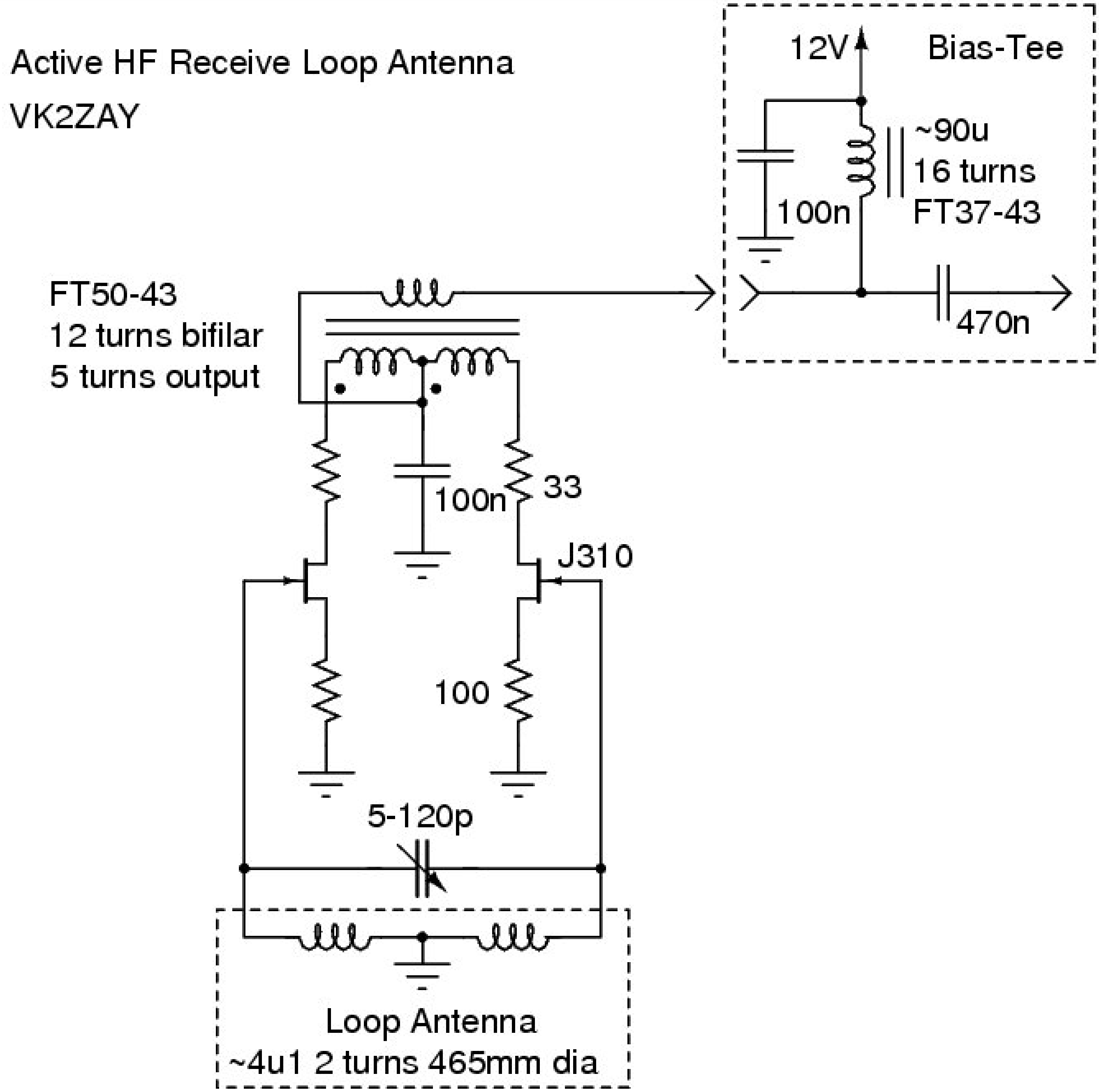
Active HF Receive Loop Antenna
VK2ZAY
12V
Bias-Tee
~90u
16 turns
FT37-43
100n
470n
FT50-43
12 turns bifilar
5 turns output
100n
33
J310
100
5-120p
Loop Antenna
~4u1 2 turns 465mm dia

More gain is available by bypassing the 100 Ohm source resistors. Oscillation can occur with excessive gain. You can set the gain at a higher, but stable level by adding resistance in series with the bypass capacitors.

The Bias-Tee

To feed the DC supply to the antenna at the shack-end of the coax a simple bias-tee was constructed. As the frequency of operation is only mid-HF a simplistic ferrite toroidal choke and two capacitor affair was constructed in a small die-cast box. I was confident this construction was sufficient at the frequency of operation, but curious about its performance elsewhere, as such it became one of the first test subjects for my experimental scalar network analyser.

© 2000 - 2009 Alan Yates - All Rights Reserved.

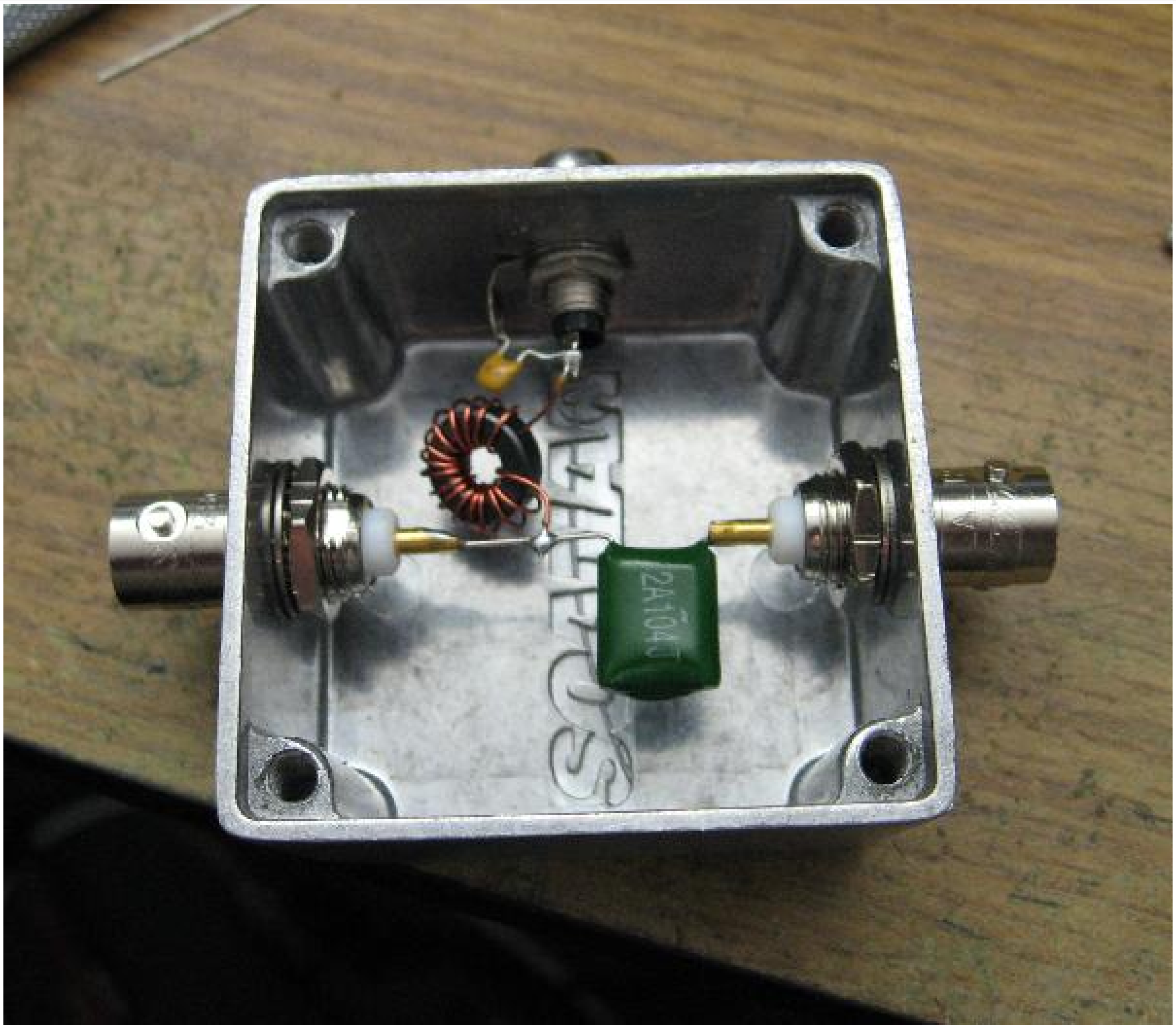

Despite having very poor "design hygiene" for high frequency response, sweeps of the bias-tee with the analyser suggest it offers acceptable performance across HF. There are some oddities in these measurements however. Fortunately they appear to be measurement equipment problems rather than excessively nasty behaviour of the device under test:

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

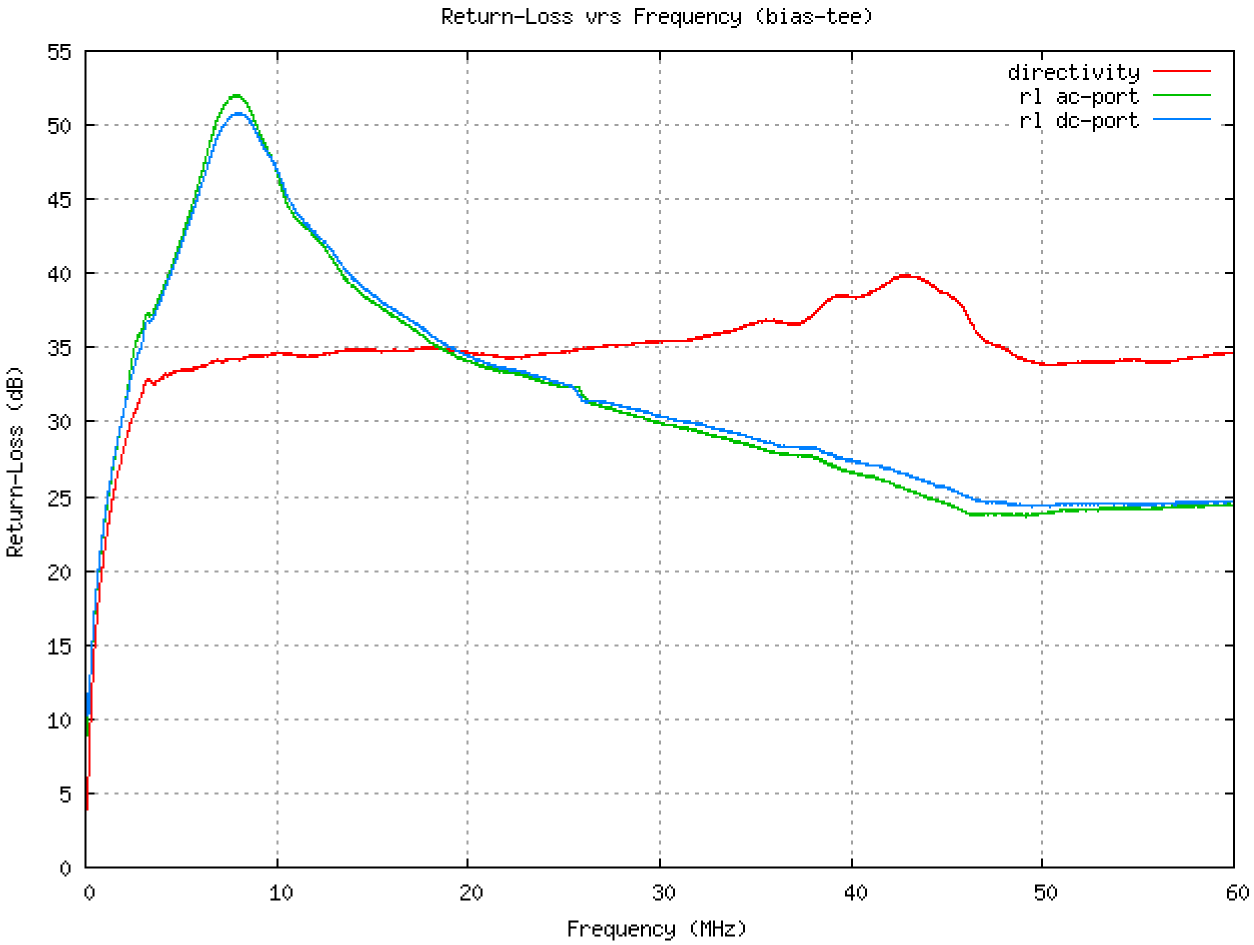

Note how the measured "return-loss" exceeds the bridge directivity below 20 MHz. This is of course an illusion, caused by the small stray reactances of the bias-T conjugate matching the bridge for better than calibration reference balance. Ideally I should change the graphing software to compute error bars based on the reflection signal magnitude compared to the directivity established by Open-Short-Load calibration at the same frequency. As the directivity of the bridge exceeds 30 dB above a few MHz all the way to 60 MHz we can safely say the bias-Tee is an reasonable match over the same range displaying a return loss exceeding 20 dB the entire way.

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

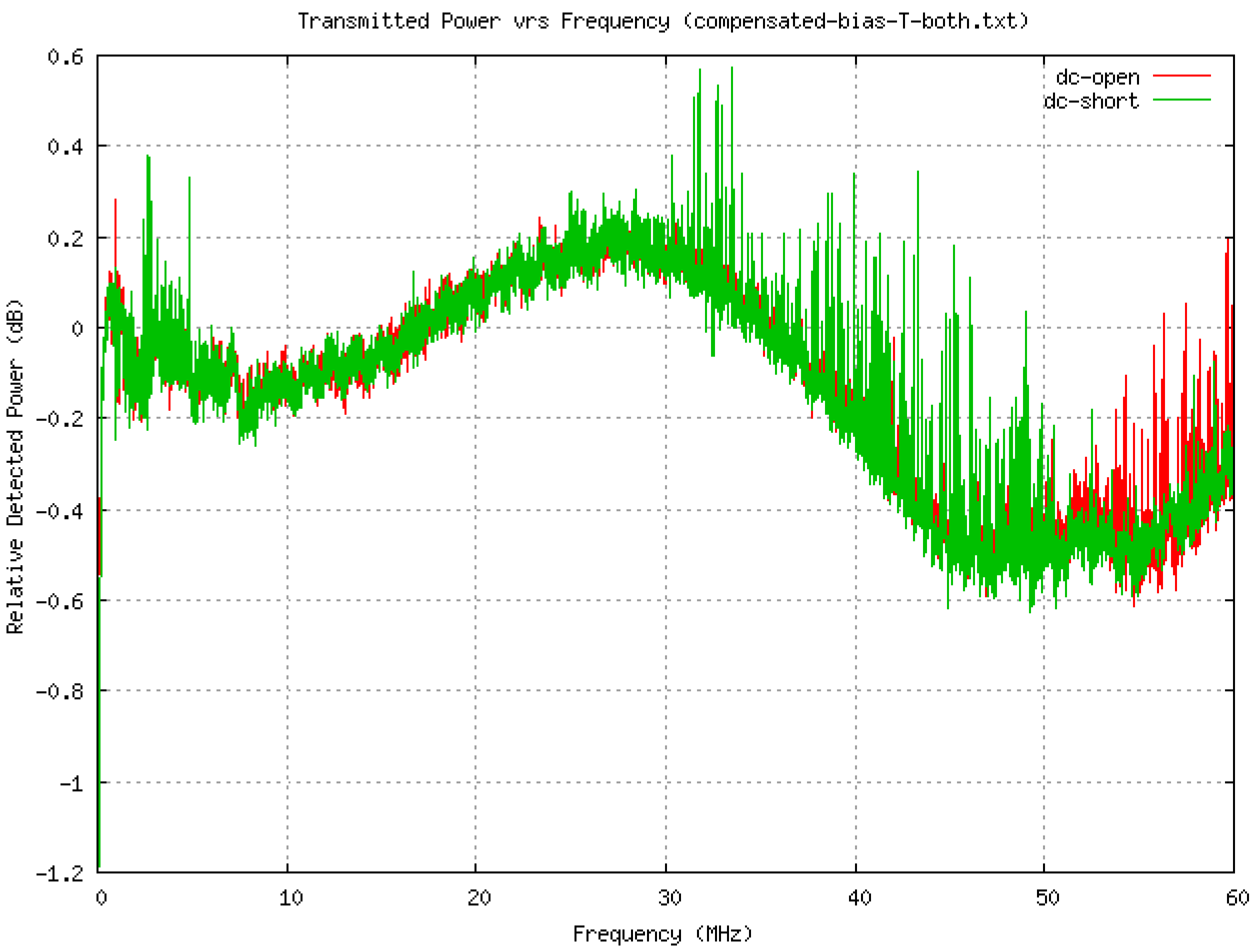

Transmission measurements similarly have some weirdness. Despite the test setup having 20 dB of attenuation in the signal path the bias-T performs the miracle of over-unity performance. Of course this is not real! Again it is just conjugate matching something in the system to make the reference calibration invalid. The variation is only a fraction of a dB so for all intents the bias-T is near-perfect across HF and into low VHF.

Lab Testing

It is rather difficult to lab test a loop antenna in a reasonable way. In particular it is extremely difficult to immerse it in a RF field of sufficiently controlled spatial uniformity and consistent amplitude with frequency to make absolute and repeatable measurements. For my initial testing a 100 mm diameter coupling loop was connected to the signal generator and loosely coupled to the antenna loop. Sufficient drive was applied to achieve a few dBm out of the loop buffer and into the power meter at peak of resonance. This allowed crude measurements of bandwidth and amplifier compression.

More detailed investigations were made by sweeping the unit with the same experimental scalar analyser used to test the bias-tee. Exact amplitude measurements made in this way are fairly meaningless, as the coupling loop is not well matched to the generator. 10 dB of padding was placed between the coupling loop and the generator but as the loop was placed almost orthogonally to the antenna loop to provide weak coupling (for Q estimation) the coupling loop does not see much of the antenna loop loss resistance.

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

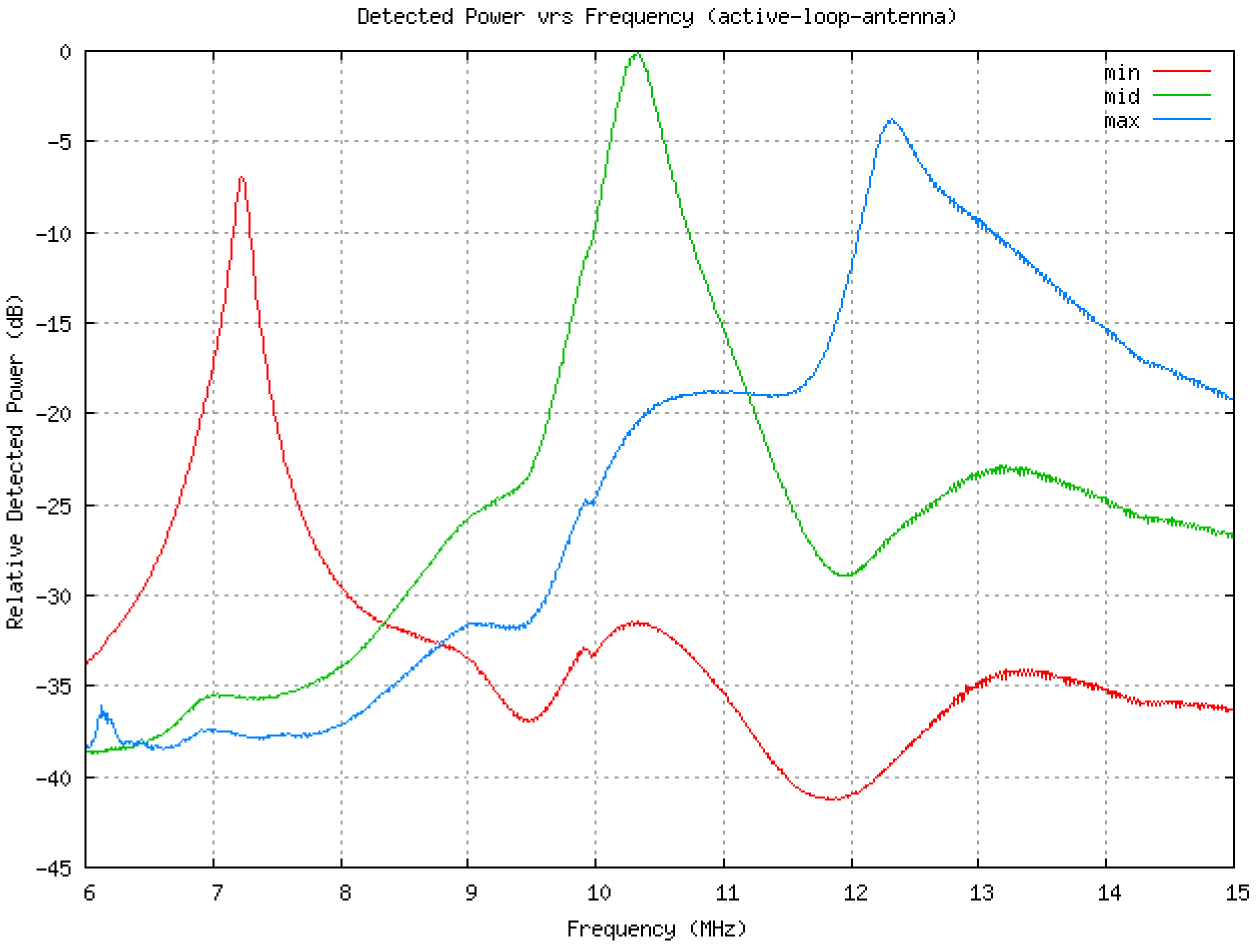

The loop tunes 7.23 MHz to 12.32 MHz with the polyvaricon used. The apparent loop Q drops with frequency, being 55.6 at 7.23 MHz, then 33.3 at 10.33 MHz and 28 at 12.32 MHz. The loop inductance is about 4.1 uH so the input impedances which match these Qs are 10 kΩ, 8.8 kΩ and 8.9 kΩ respectively. Wider sweeps show problems with the test set-up, in particular generator harmonic energy when tuned below the loop resonances. Neither sweeps have me feeling very comfortable about the quality of test set-up (or the loop construction for that matter). The HF feed-through is probably due to the unshielded housing of the buffer amplifier and stray circuit capacitances. Maybe a LPF should be added to the output to suppress these responses? A HF receiver should reject them with no dramas, but the number of hints of internal resonances and general "complexity" of the baseline above and around resonance doesn't make me too comfortable.

© 2000 - 2009 Alan Yates - All Rights Reserved.

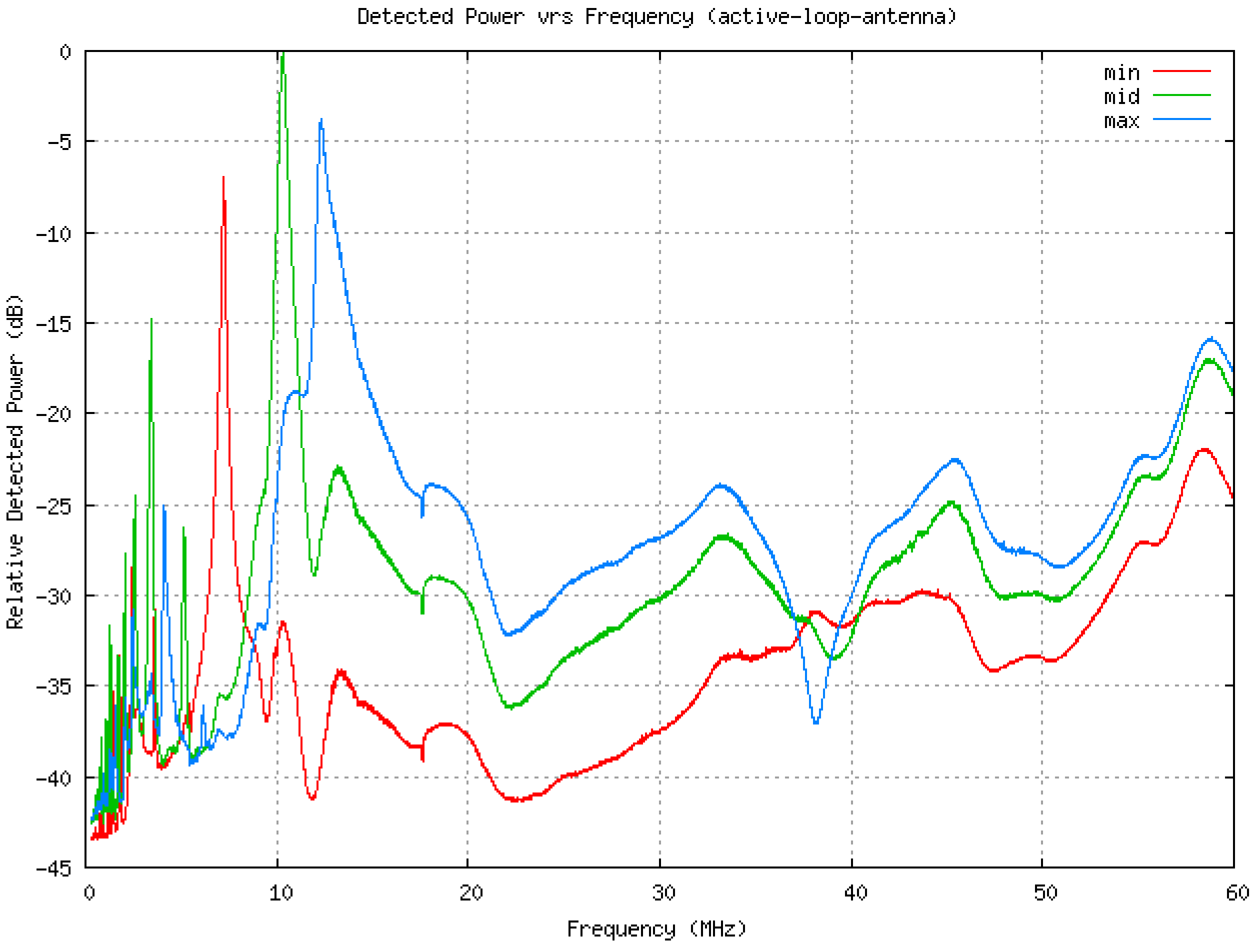

When the amplifier is left unpowered the loop leaks through RF at a lower amplitude, and the resonance is shifted down in frequency somewhat. This is immediately apparent when you connect a receiver to the loop, even without powering it on you can peak-up the background noise level by tuning the polyvaricon. However once power is applied the loop must be retuned (up somewhat) for maximum background noise. Loop Q is degraded quite significantly in the leak-through mode, with Qs of 39.3, 17.1 and 9.1 for the test frequencies discussed above. Loop Q in general could be improved by weaker coupling to the JFET gates, some simulation could optimise the values required if the resonator usable Q was measured and the FETs well characterised. Noise figure would be compromised by resistive DC gate biasing, maybe use chokes?

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

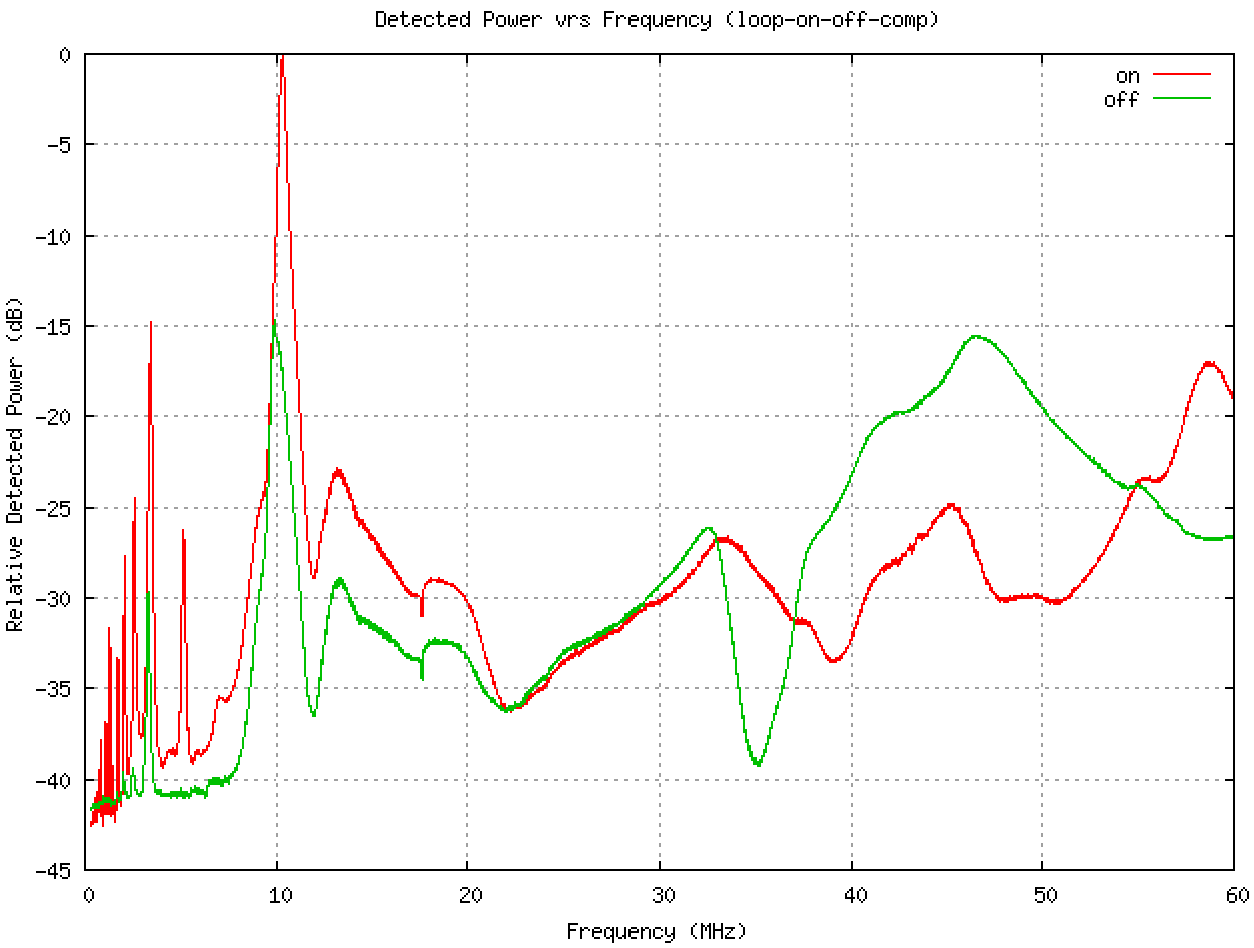

On thing I initially found rather disturbing about this particular sweep is the 2nd harmonic peak is absent from the "off" run. The 3rd harmonic peak is 15 dB different just as the fundamental is - but the 2nd harmonic "on" signal is smaller and 15 dB down from its level is into the -41 dB "leakage floor" we appear to be observing... All was revealed when I considered that the reference level here is about -12 dBm meaning the leakage is around -53 dBm, quite likely considering the unshielded nature of the amplifier in close proximity to the coupling loop.

Field Testing

For real RX testing I lashed-up the antenna on the balcony, mounted a little above railing height, and A/B compared it against the base-loaded vertical I use for QRSS transmission. This is *not* a very fair test, as the vertical is 3 metres tall, while the loop is less than half a metre in diameter and was mounted at the base of the vertical. Also, interaction between the antennas was *not* controlled during the experiment, and experience has taught me this is critical for meaningful results.

The loop is largely limited by its aperture (cross-section). Compared to my loaded vertical it is about 6 dB down, making it ineffective for its original design purpose (improving my QRSS RX noise floor). However, unlike the omnidirectional vertical it has well defined nulls which are useful for avoiding local interference. The nulls do help dodge some noise, but unless I make it larger or mount it higher the loop is simply not as good for QRSS RX.

For comparison purposes I have built a much larger passive tuned loop antenna. The square loop is root-2 metres on a side in a diamond configuration (for mechanical simplicity) and is matched to the coax using a ferrite transformer placed near the polyvaricon that tunes the loop winding to resonance. It was not possible to fit the loop in the shack for Q-determination and hence the matching is at best an educated guess. Its Q is likely dominated by whatever the receiver input impedance reflects into the loop tank through the matching transformer. Experiments continue, comparing the three antennas for relative performance. More experience and additional test instrumentation is required before fair comparisons can be made.

2 comments.

Attachments

title	type	size

© 2000 - 2009 Alan Yates - All Rights Reserved.

Active Loop Circuit Diagram Source	application/postscript	13.671 kbytes

© 2000 - 2009 Alan Yates - All Rights Reserved.

The Fredbox

2007-05-26

I first heard about The G3XBM Fredbox transceiver via Solder Smoke. As soon as I saw it I just had to give it a go. It is a very simple and elegant design. Of course it offers no bells or whistles, just a fixed TX and RX frequency, and a flea-power output on TX, but it has a special charm in its simplicity and the retro usage of AM on VHF.

Transmitter

I built the transmitter side first. BF199s were selected as a good candidate for the RF devices, but I lacked a crystal in the range specified in the original article. Instead, I used a common 16.384 MHz crystal and redesigned the circuit to be two triplers rather than two doublers. This crystal is a common "computer" crystal, but places the TX frequency in the high-end simplex segment. This is a bit too close to the pager-splattered end of the band for my liking, and doesn't match the band-plans. For now this is OK, I'll get a custom crystal cut eventually.

I didn't use shielded cans or variable inductors for the transmit coils (as specified by the article), rather I used fixed inductances wound on T37-6 toroid cores with bare 0.71 mm tinned copper wire. My LC resonance calculator and nH inductance meter were enormously helpful in making the selection and testing of the tripler and output stage resonators. Each stage was tuned with trimmer capacitors.

It was amazingly easy to get the TX-side working, I just built each stage from the crystal to the final amp in turn, testing as I went. Each stage is well behaved and peaks nicely.

Peter VK2TPM could hear the signal at his QTH several kilometres away when I connected the half-finished TX board into my flower-pot antenna. The DC input power was about 23 mW, and no special attempt was made to match the output into the load, in fact the series trimmer in the matching network was absent at this point, just a fixed 12 pF capacitor was used for DC blocking.

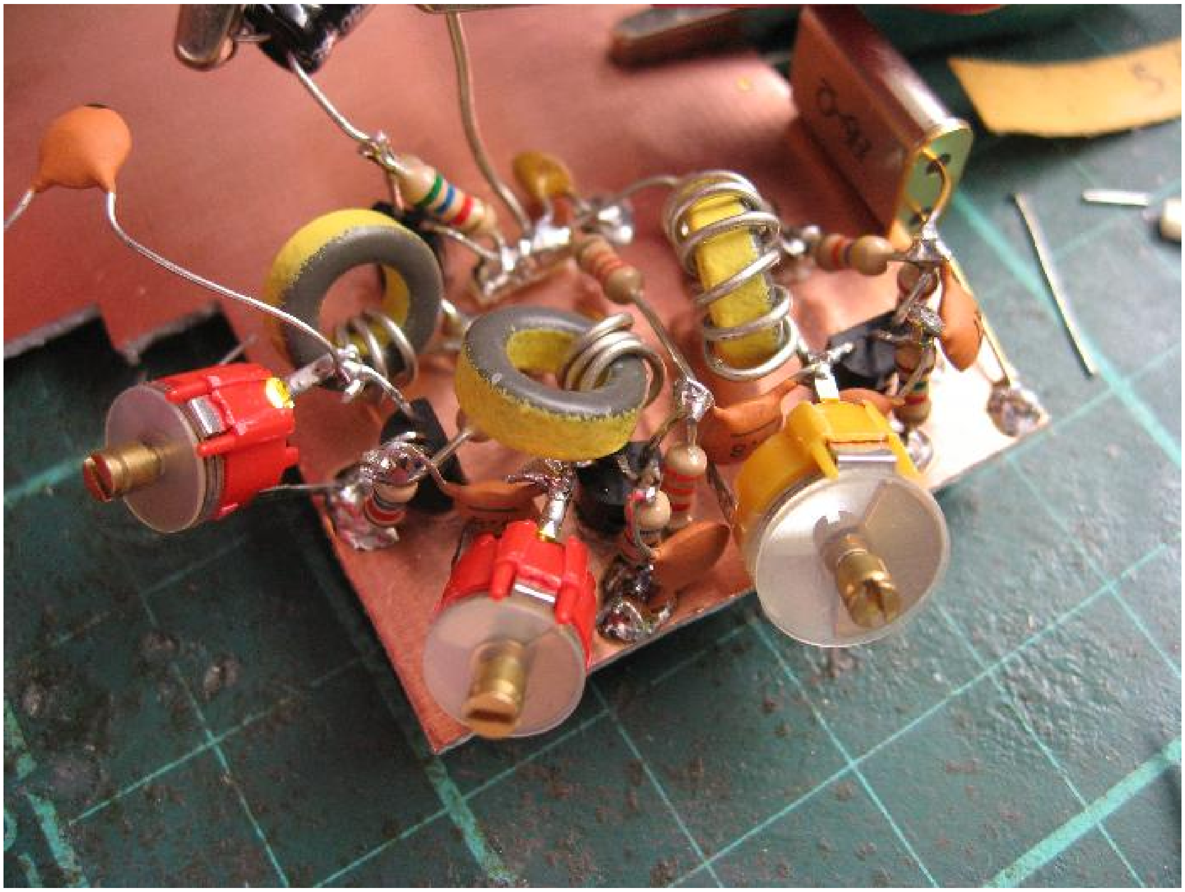

Upon finishing the TX circuit I did experience a bit of RF pick-up in the microphone amplifier 2nd stage. A 1 nF capacitor to ground discouraged its RF gain and eliminated the problem. The 2nd AF amp stage is located immediately adjacent the crystal oscillator stage and was picking up RF directly. The effect was not audible, but was visible on the spectrum analyser as weak 16.384 MHz sidebands either side of the carrier. This wasn't causing feedback, just high-frequency modulation of the signal. If nothing else it proved the bandwidth of the modulator, which is perhaps a surprise considering the 100 nF decoupling on the modulated rail, however the output impedance of the series modulator emitter is so low it could deliver a few tens of mVs of HF ripple into that kind of load.

Receiver

© 2000 - 2009 Alan Yates - All Rights Reserved.

As easy as the Transmitter was, the Receiver was hard. It fought me every inch of the way. I spent an entire day trying to work out why it was simply not super regenerating from about 130-160 MHz. It turned out to be a 10 nF decoupling capacitor on the cold-end of the detector resonator. The article specified 1 nF, and I originally intended to use this value, but I had a strip of 10 nF mono's on the bench, so I used them. At build time I did consider why 1 nF was specified in the first place, I figured it was avoid the exact problem that would befall my unit, decoupling resonance. (Lesson #1: Trust the original builder and your initial instinct.) When the unit wouldn't oscillate properly I assumed that I had damaged the capacitor on install - this is a pretty common fault, so I tested it in-place by ensuring it would shunt a HF signal (my standard decoupling cap test), it passed this test just fine. (Lesson #2: Test at the frequency of operation.)

My hubris about "modern components" being superior and likely "purely capacitive" at VHF turned out to be completely wrong. It took *hours* to work it out, but eventually I determined the entire decoupling network was resonant near the operating frequency. Much foul language later and I replaced the cap with a ceramic 1 nF, with its "flashing" broken off and scraped right back to the disc to minimise the lead length. This cured the problem.

For the longest time I had assumed it was the source coil - and in fact the first source RFC I used (a molded choke) was being operated above its self-resonant frequency and prevented any oscillation at all. I replaced it with a few turns on a ferrite bead which seems just fine now.

A couple of other minor annoyances/mistakes were worked through (like picking larger inductances and making the entire circuit so sensitive to stray capacitance it was a nightmare to tune - DUH!).

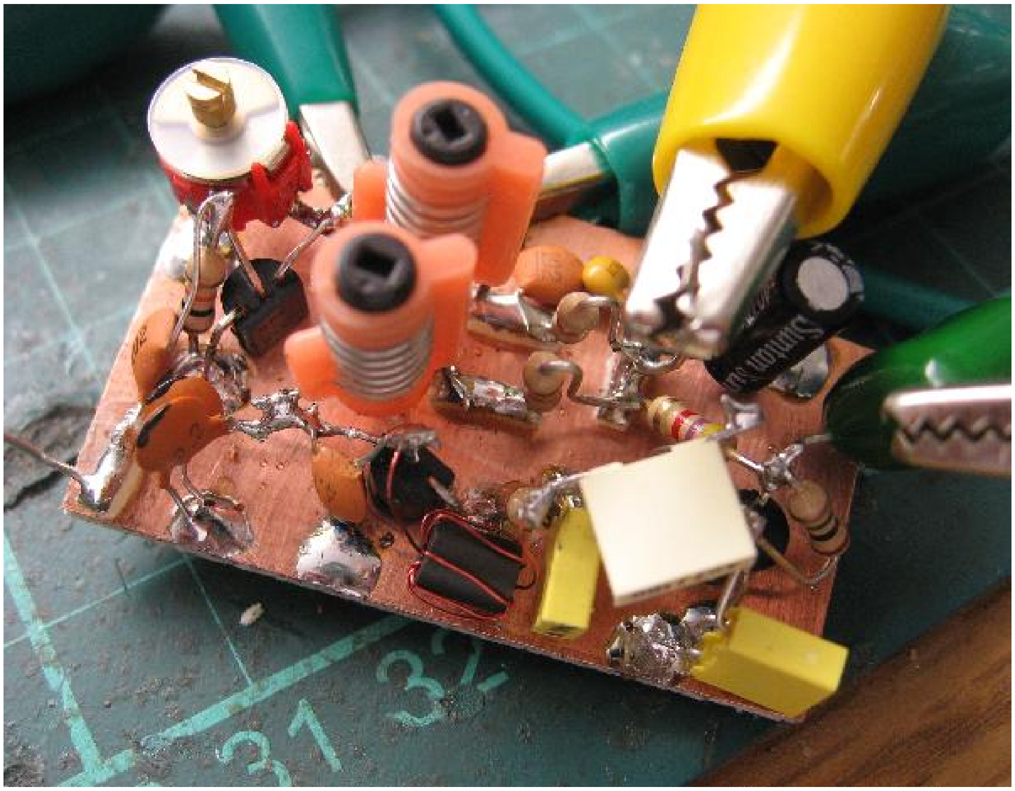

Eventually the unit super-regenerated right through the region of interest and the LNA stage was constructed. Initially I put the LNA drain coil too close to the detector coil and they over-coupled. This meant as I tuned through resonance on the LNA drain it would pull the detector so much it would shut down. Bugger! Moving the coils apart a little reduced this effect to acceptable levels, but the core of one still effects the other a little. I'm happy with the current coupling, and it is actually useful to help tuning the LNA resonator. As you rock it through resonance it will pull the detector, and by observing the wiggle on the spectrum analyser you can tell you've got it tuned up. The AGC action of the detector makes it hard to tune for maximum smoke otherwise, as the AF output doesn't change much at all even when the front-end isn't tuned up properly. Once you've got it nearly right you can use a weak signal to tune for best signal to noise.

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

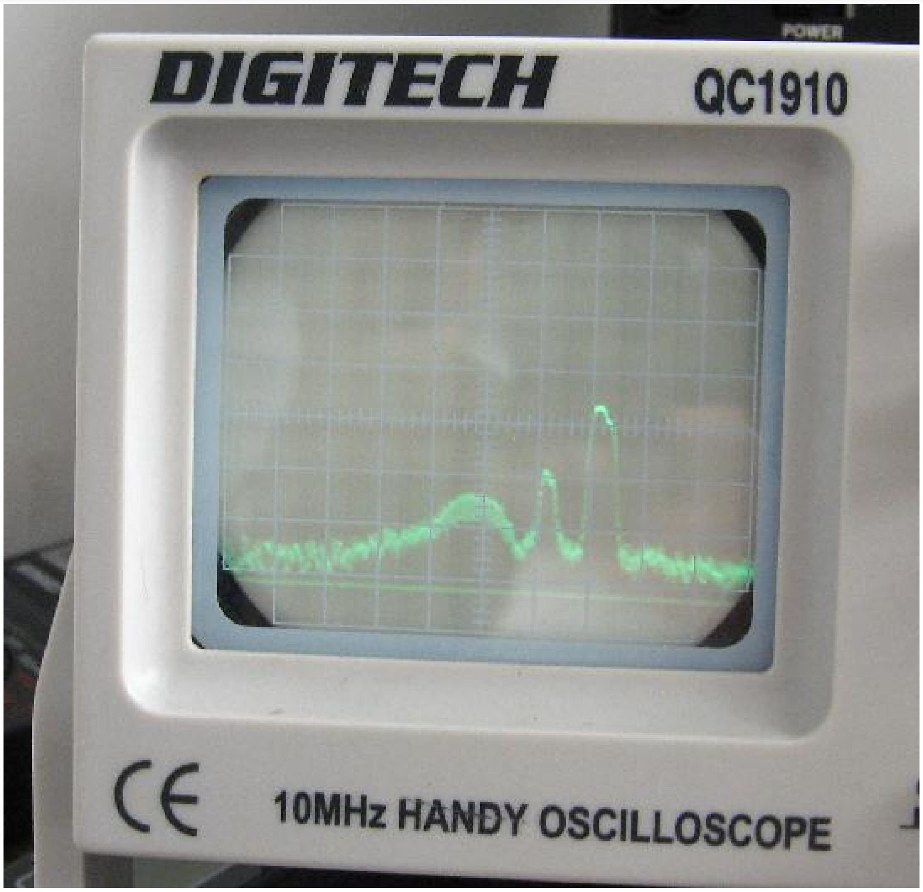

Note the pagers above the 2 metre band in this spectrogram. The hump in the noise floor is the receiver super-regeneration side bands. The smaller peak in between is the output of the Fredbox transmitter, the leakage from the unshielded prototype on my desk operating into a 47 Ohm load. It is rather disturbing that the pager signals are *larger* than this local signal just a foot or two from the spectrum analyser antenna.

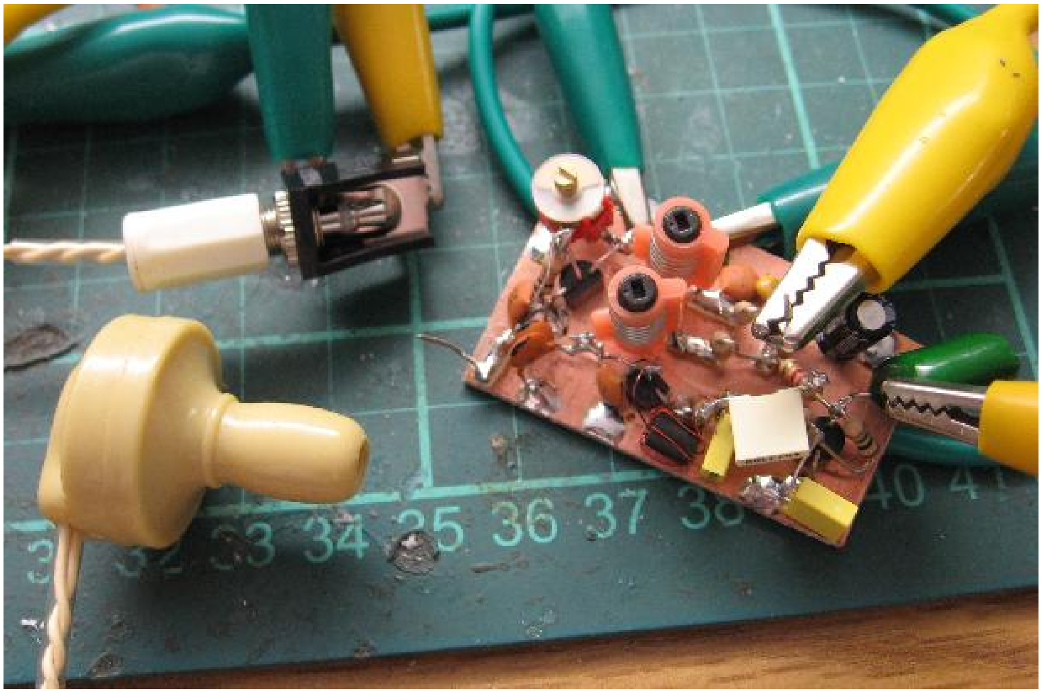

I used J310s for the receiver FETs, and a 2N3904 for the audio amp. Although the main design requirement for Roger appears to have been flea-power, I think a better AF amplifier that can drive modern low-impedance headphones would be preferable. Only Jaycar now carries crystal earphones with a nice soft silicone ear piece. The one I used comes from DSE and it is hard-plastic - not very nice on the ear. I really hate these kind of earphones anyway, I'll probably be rebuilding it for a low-Z output at some point, but it does work pretty well as-is.

BTW: While I was in the "Special Hell" of decoupling resonance I took the RX down to the FM broadcast band, and up to the VHF-hi TV band. It works wonderfully in both, which isn't a surprise. However, by adjusting the drain-source feedback (then a trimmer, now fixed) I

© 2000 - 2009 Alan Yates - All Rights Reserved.

Hall's regen on the FM band. The topology is more forgiving however, allowing grounding of the tuning cap. I also built several different oscillator topologies in desperation before I identified the decoupling fault, in one I got a Colpitts-like oscillator working with emitter coupled feedback to a tapped capacitor across the tank. This is something I should have thought about a *long* time ago, I'll probably build yet another FM broadcast regen using this topology for the detector, it seemed quite easy to control just by manipulation of the base voltage.

Boxing It Up

I am still tossing up between a cast Aluminium box, a custom box folded out of Aluminium sheet, or an Altoids tin. The circuit is small enough to just fit inside an Altoids tin but it probably won't fit with a battery. I'll pick up a centre-off momentary-one-side switch over the next week and finish off the radio one way or the other.

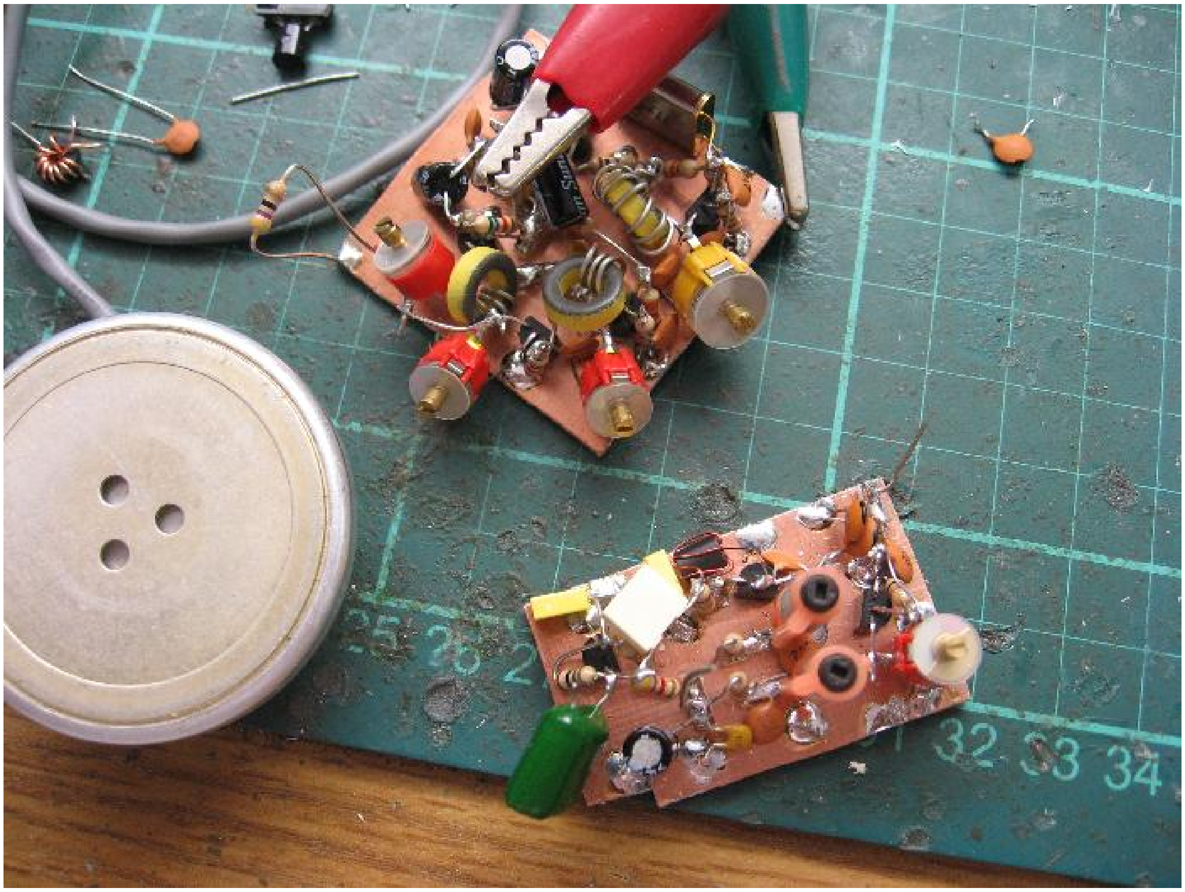

Note the use of an old telephone receiver as the microphone. It is nearly as large as the entire TX board. I'll have to find my electret mics, I know I have a bag of them somewhere that I got from a Rockby sale.

I'm strongly considering rebuilding the radio, perhaps through-the-hole to minimise its size. Although my prototype isn't too large as-is, it would be nice to neaten it up. Maybe I'll build it with fixed caps and variable inductors to save space too, although shielded cans are about the same size. It *might* be possible to tune the multiplier stages with stretched coils and fixed caps, this would make it much more compact and save money too if it ends up being kitted...

More TX Power

I am considering running the unit on 12 Volts to get a bit more RF out, and perhaps building in a small amplifier to get it up to 1 Watt region. This would probably involve a rebuild of the TX side to use a 2N4427 or similar final device, and a more robust series modulator transistor. This won't be efficient, but is probably easier to get going than a linear amp which would need careful drive adjustment.

I'll probably conduct some experiments around using a 2N7000 or VN10KM as the output device. The math suggests they may operate on 2 metres. I've already got them working on 6 metres in a brief experiment last month (must document that).

Comments

Working with VHF is fun! I find it a great learning experience, especially when you get problems like the resonant decoupling cap. That kind of thing really pushes your understanding of the physics and teaches you a lot.

I know a lot of HAMs won't touch anything above the bandwidth of their oscilloscopes. I can understand the frustration when something doesn't work, especially when you can't see why, but it really isn't that much worse at VHF. With just a diode probe you can achieve a lot. It does help enormously if you have VHF test equipment, for example I likely would have never noticed the HF modulation problem had I not had a spectrum analyser (although the HF was visible on the collector of the modulator drive amp, and most of us have a CRO that can see fine near 16 MHz).

A wavemeter can be helpful, if a bit retro, especially for making sure your multipliers are tuned up right and for looking for spurs. You can build one quite easily, it only needs to be a resonator with a detector and a LED or meter as a read-out. I'm a big fan of the biased 1N5711 through the decoupled bottom of the tank coil topology. For super-sensitivity you can use a MPSA18 as a DC amplifier. With a signal generator or dipper and a counter or scanner/receiver you can easily calibrate it. It can be used like a poor-man's spectrum

© 2000 - 2009 Alan Yates - All Rights Reserved.

My only advice when you get stuck is to trust the physics, do the math and follow your instinct when nothing is making sense. Rebuild parts of the circuit and test them independently. Measure what is actually happening and try to figure out what kind of misbehaviour in the circuit would cause the observed of behaviour. That will often solve an otherwise intractable problem.

Update 2007-06-05

I took the Fredbox boards to the local Homebrew Group meeting. I had both halves hooked up and talking across the room. Peter VK2TPM had brought along a digital audio recorder and did an interview with many of us in attendance. You can hear what I said about The Fredbox on Soldersmoke 62, and even a brief snippet of audio going through The Fredbox.

Also on the recording are Peter VK2EMU talking about the 80 meter challenge, John VK2ASU talking about his transmitter modules for the challenge (and an interesting diversion into IRF510 gate-modulation with some input from Brian VK2TOX - something I was thinking about back here, apparently Drew Diamond VK3XU has already produced a design doing just this). Mike VK2BMR also talks about his great VSWR/power meter project. His unit was absolutely beautiful, I was very much taken by the excellent job he did of cutting the PCB stock that made up the external directional coupler box, essentially flawless, perfectly square workmanship.

Update 2007-06-09

I've built a weak-signal source for aligning the receiver. One of the biggest weaknesses of the Fredbox receiver is that it drifts in frequency quite significantly with Vcc variations. Powered by a 9 Volt battery near the end of its life, it may drift enough to make the receiver completely off-frequency despite its poor selectivity. If I build this circuit again I'll probably put in a stabilized supply for the detector stage to avoid this problem.

© 2000 - 2009 Alan Yates - All Rights Reserved.

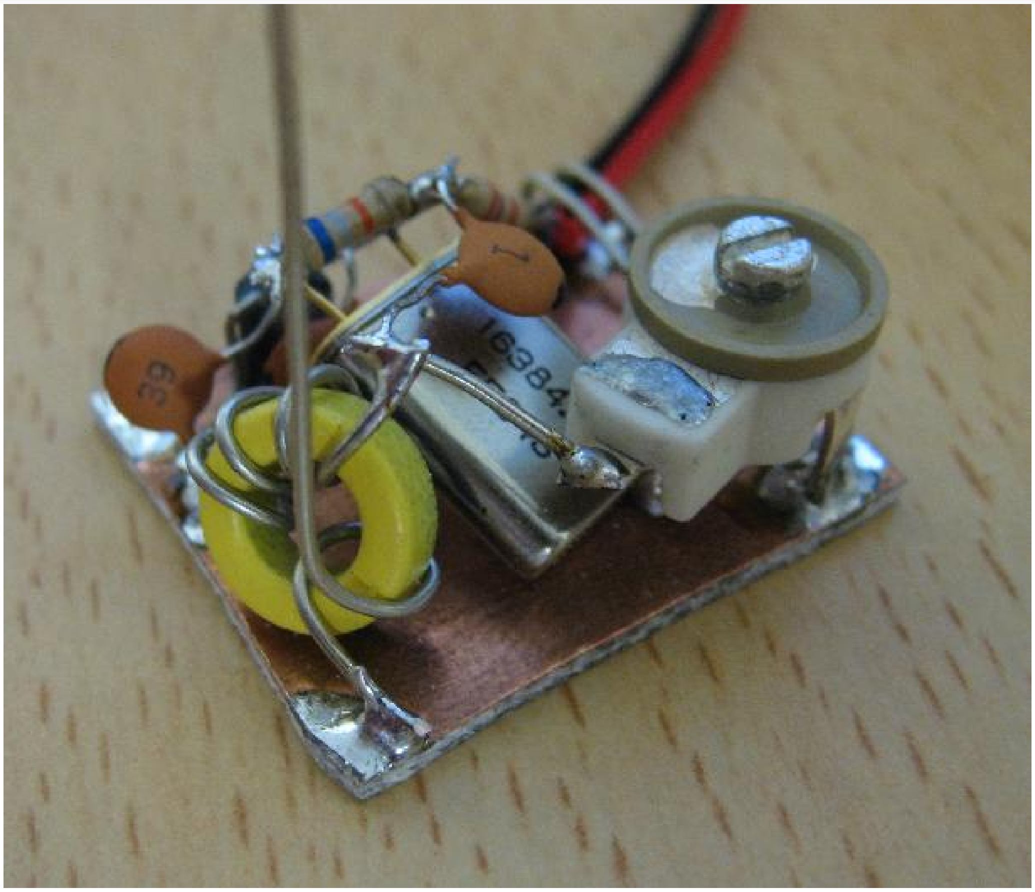

The signal source is a Pierce oscillator driving a tuned circuit which selects the VHF harmonic of interest. Because of the oscillator topology the crystal isn't pulled down as much as in the Fredbox circuit. The difference is fairly minor for my purposes, the poor selectivity of the receiver makes the difference in frequency of no real consequence. There is no active (or passive) multiplier, so the tuned circuit is merely extracting the harmonic energy from the oscillator. The harmonic energy available is very small, which is perfect for the application, giving an almost undetectable signal 1 metre away.

Update 2007-06-10

I've boxed up the Fredbox. As discussed earlier I went with the Altoids tin, despite this not allowing the battery to also fit inside the enclosure. I tossed up soldering an additional tin to the back to hold the batteries, but for now I've gone with the 9-volt battery snap just hanging out. It will work well with 6xAA battery holders which can just be held to the box with a rubber band.

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

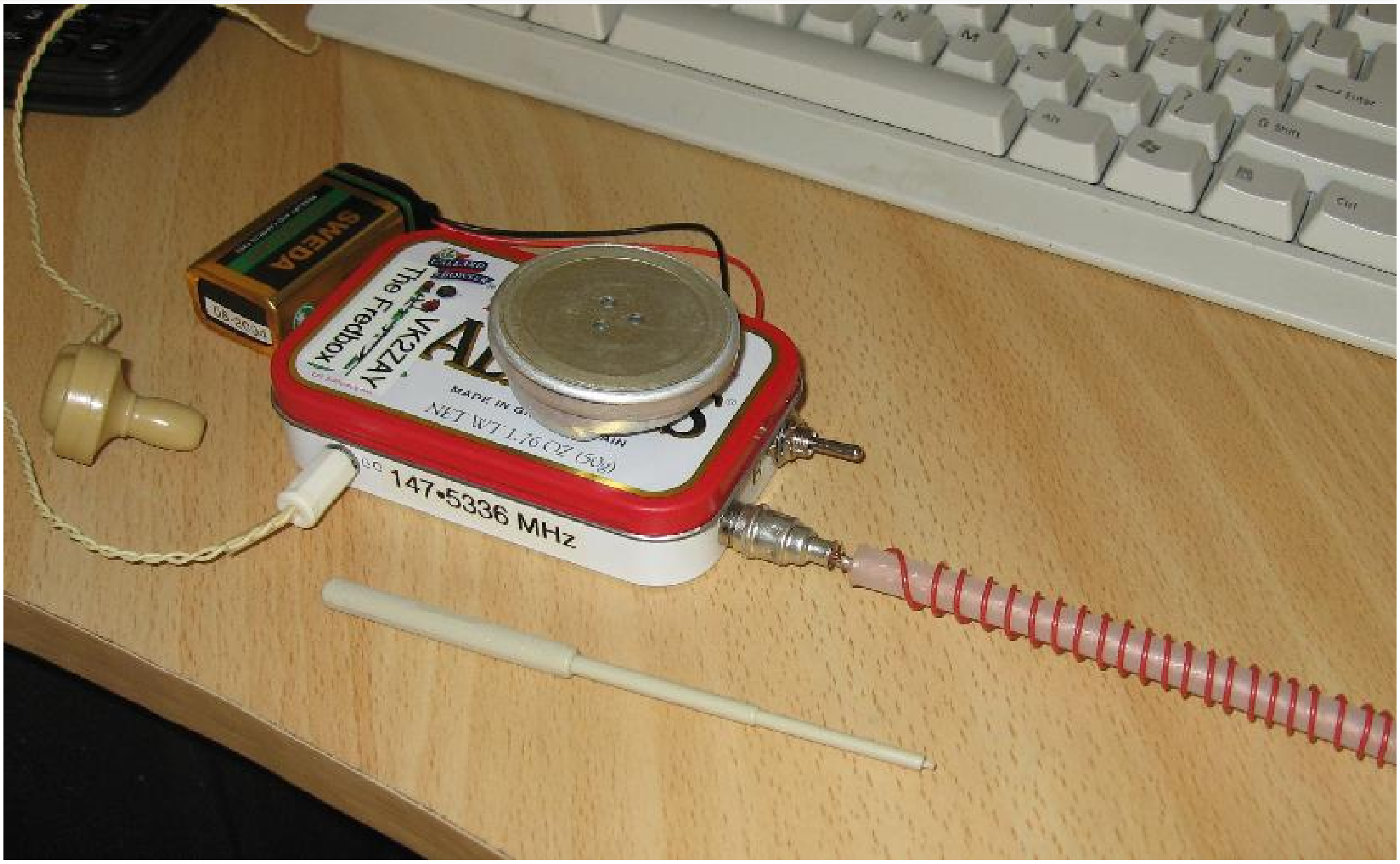

For the RF connector I chose an RCA. I can hear all the VHF engineers cringing, but for the purposes of this prototype it works just fine. The Antenna is a half-wave of wire helical which is quarter-wave resonant (with some pruning). The former is a piece of centre conductor and insulator from RG-213 coax. The centre conductor was left in place and is soldered into the RCA plug centre conductor, adding some capacitive loading and shortening the length of wire needed for resonance. I have no idea if this is good or bad, but it works.

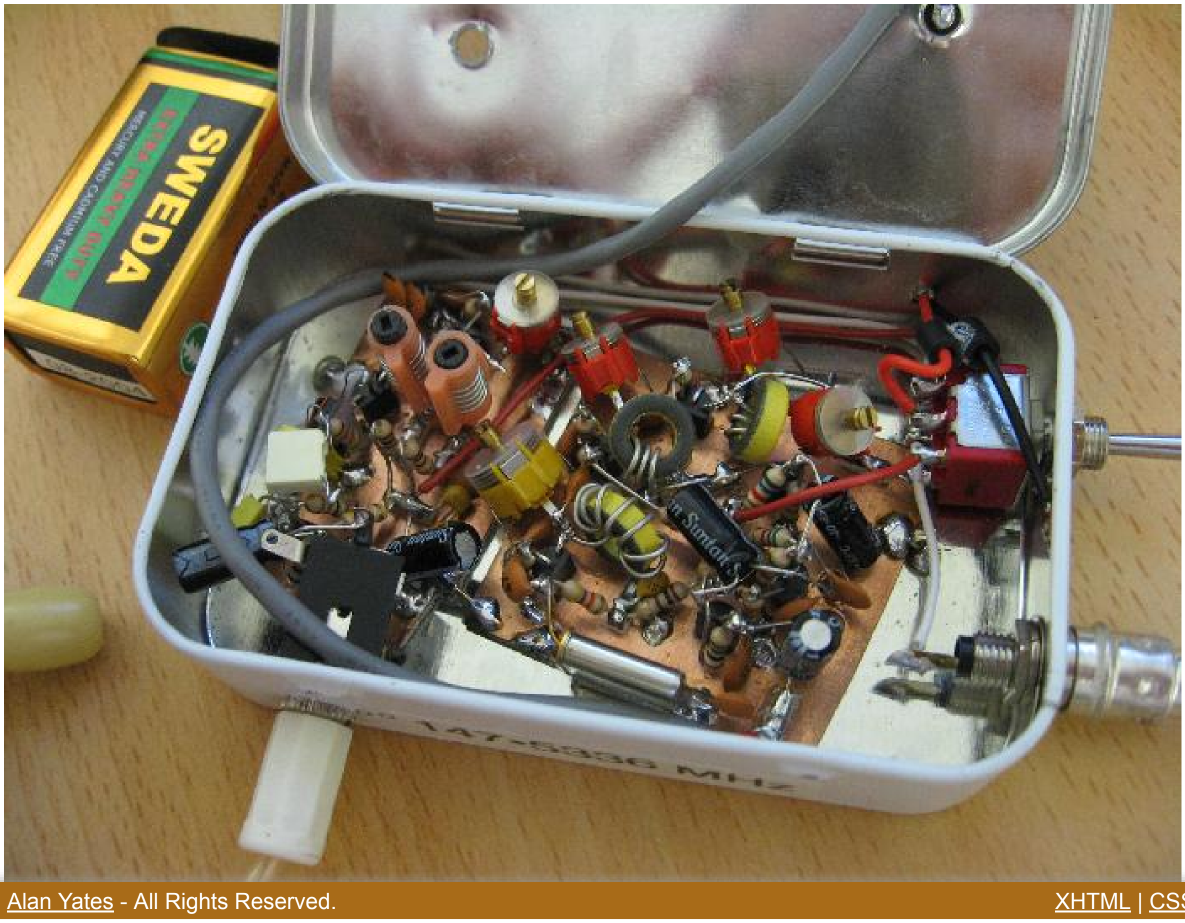

© 2000 - 2009 Alan Yates - All Rights Reserved.

I didn't find my stash of electret microphones, so I ended up just soldering the telephone receiver into the top of the Altoids tin. Ugly as hell, but also kinda unique. It looks somewhat like those Vietnam-war era VHF-low walkie-talkies.

I couldn't find a momentary-one-side toggle switch, so a centre-off on-both-sides switch was used instead. A special dummy load/diode peak-voltage probe was assembled for the final alignment. The carrier power ended up being near 25 mW on 12 Volts, on 9 Volts 10 mW just like Roger says in the article.

Update 2007-06-11

The 2N7000 on 2 metres experiment was a failure, it simply doesn't produce useful power beyond 90 MHz or so. However, it is very usable below 70 MHz. My input network was far from optimal, so perhaps with some more work it would be possible to get it working higher up, and I haven't tried an VN10KM in the same circuit.

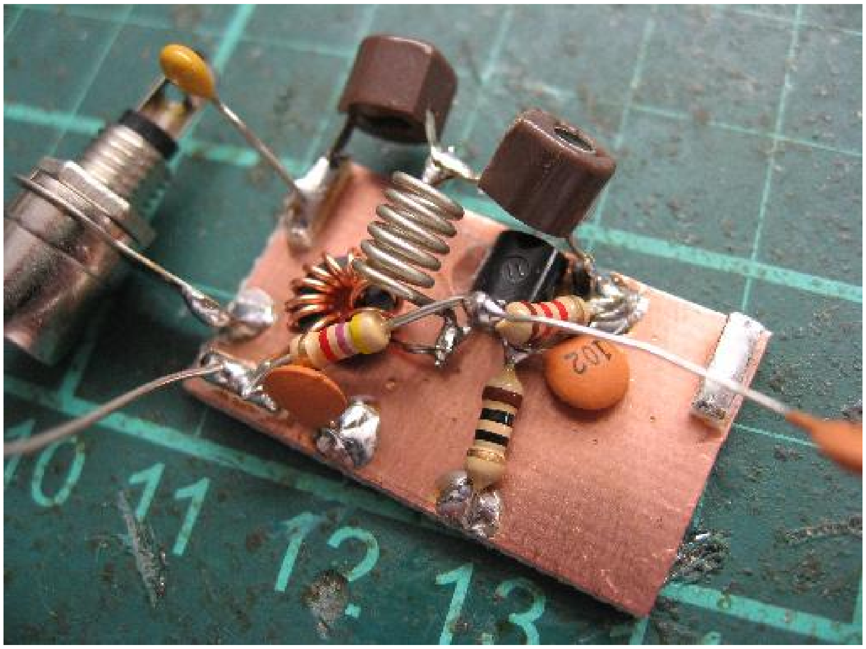

I've also been fiddling around with grounded-base class-C multipliers. (Not just decoupled, biased base, the base actually soldered directly to the ground plane.) At first this seems a little weird, but if you pull the emitter low with a link-coupling to the previous stage collector current will flow. The advantage is excellent reverse isolation, which might help with stability with less than ideal layouts and devices. Such a topology was apparently quite common years ago with the 2N918.

1 comment.

© 2000 - 2009 Alan Yates - All Rights Reserved.

AD8307 RF Power Meter

2010-02-01

At the latest ARNSW Homebrew Group meeting I had the opportunity to finally calibrate my AD8307-based RF power measurement head.

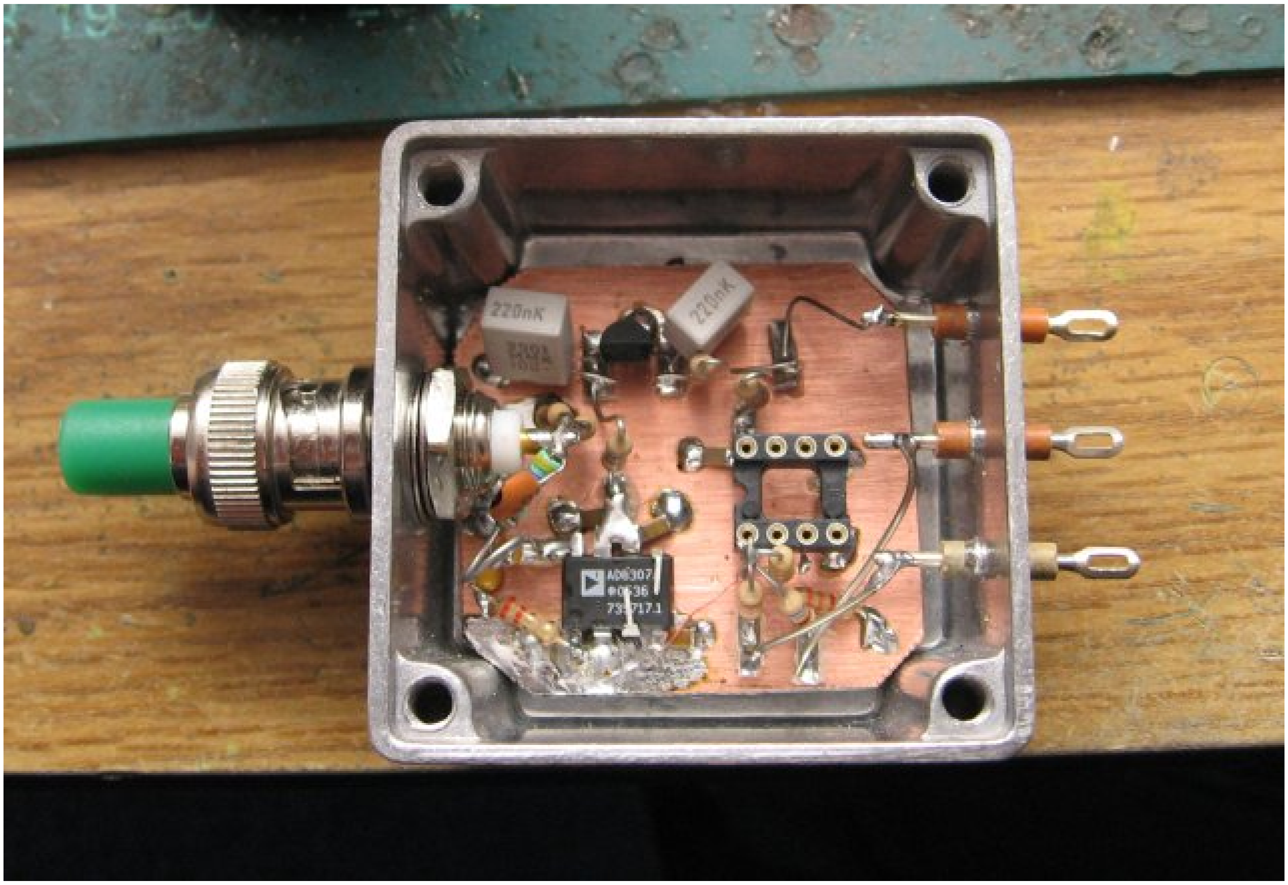

Mark VK2XOF brought along his power and frequency reference equipment and gave me a bunch of calibration points.

© 2000 - 2009 Alan Yates - All Rights Reserved.

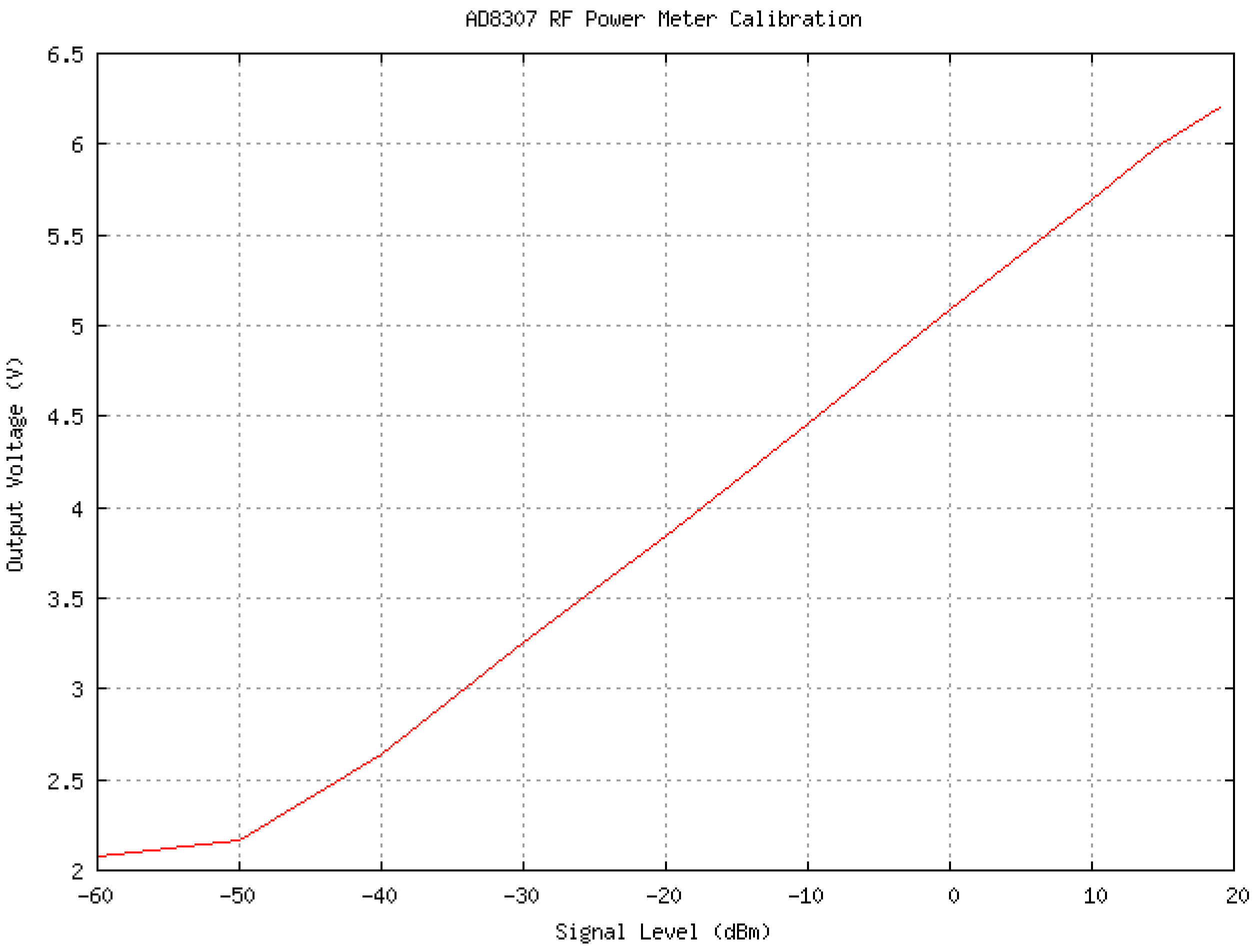

Data points above 13 dBm were given using the generator out of levelling-loop calibration so the slight compression approaching 20 dBm is to be expected. The levelling off below -40 dBm however is not. I know the noise floor of the meter (dummy-load terminated) is about 500 mV DC output, -40 dBm is about 2.65 V out, suggesting wideband noise from the generator was swamping the lower level spot calibration signals.

Regardless in between the data is very consistent and lets me compute a "slope + intercept" calibration relying on the excellent linearity of the AD8307. The 0 dBm intercept is 5.09 Volts and the slope is 16.469 dB per Volt (60.7 mV per dB). Based on the 500 mV noise floor this equates to about -75.6 dBm, or about 80 dB dynamic range - as expected for an AD8307-based unit.

Frequency response wise, my use of leaded resistors means it maintains its accuracy to about 250 MHz, which is sufficient for my immediate purposes. The circuit design itself is the W7ZOI design from EMRFD, page 7.7, figure 7.13.

3 comments.

© 2000 - 2009 Alan Yates - All Rights Reserved.

Antenna Experiments are Great Teaching Aids

2008-08-20

With the QRSS beacon now working pretty well it is down to getting the antenna system in better order. My limited space gives me little chance of a full-size dipole or vertical, so loaded compact antennas have been the mainstay of my station.

© 2000 - 2009 Alan Yates - All Rights Reserved.

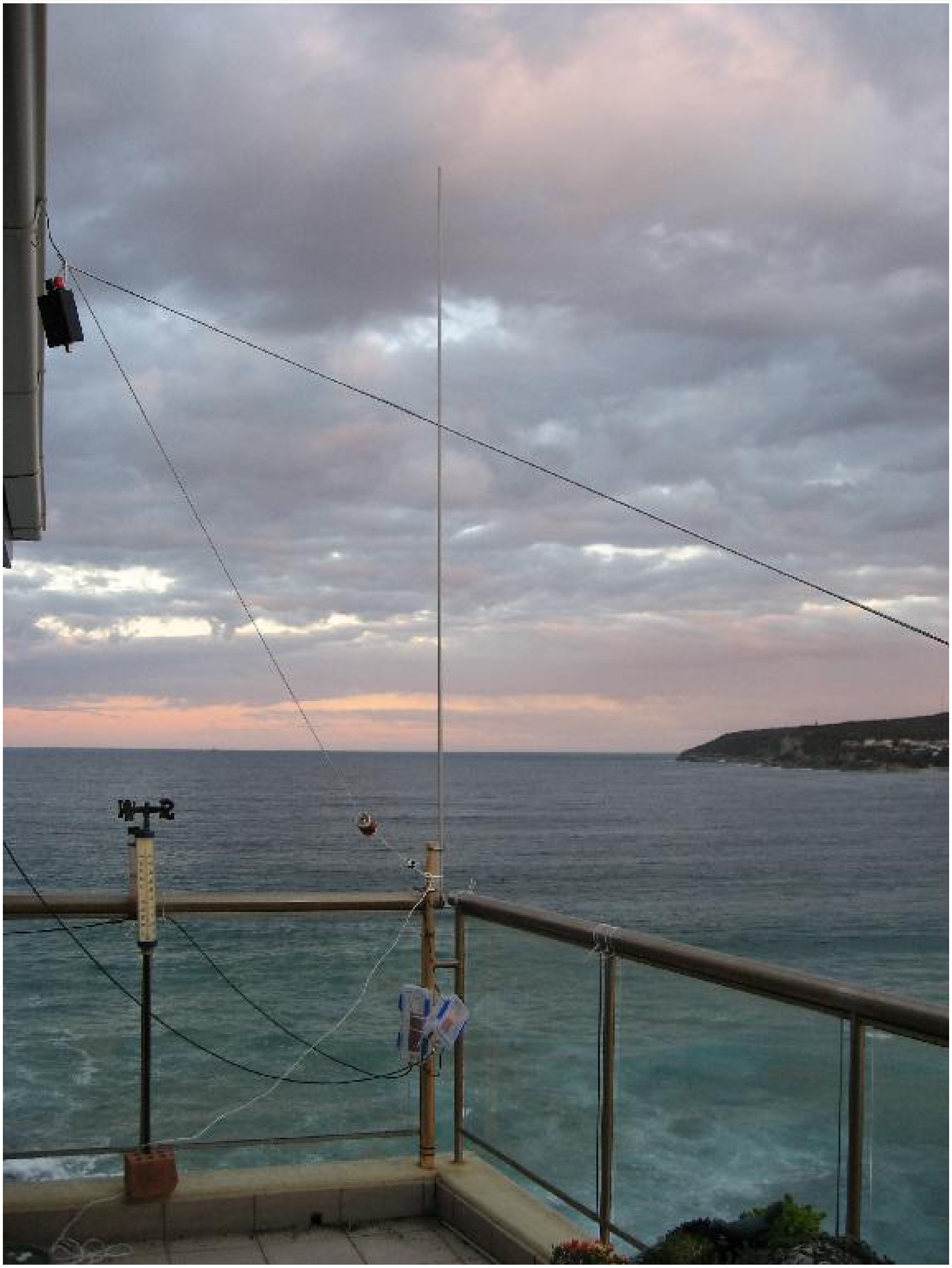

The Loaded Dipole

The 30 meter compact dipole is only about 5 metres long. It is currently erected sloping from the guttering to the balcony railing. Each arm has a 2 metre radiator then a ~27 uH loading coil wound on a piece of PVC pipe using 1.8 mm magnet wire, followed by a roughly 1/2 metre capacitive loading element which is rolled up at the ends to roughly tune the antenna. Fine tuning is achieved using the matching unit at the feed point.

© 2000 - 2009 Alan Yates - All Rights Reserved.

The Matching Unit

The matching unit is a conventional Z-match topology.

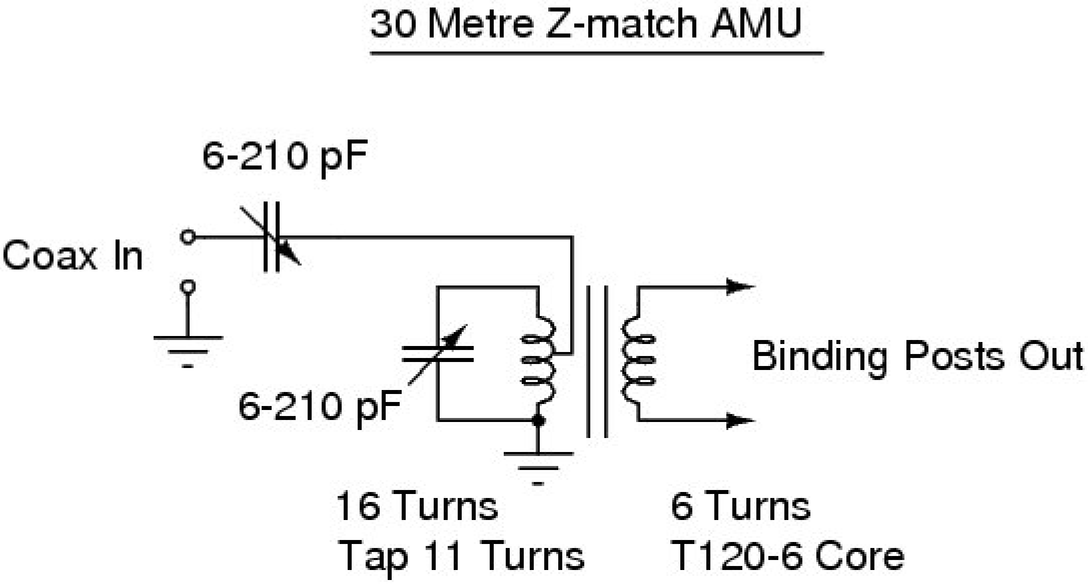
30 Metre Z-match AMU
6-210 pF
Coax In
Binding Posts Out
6-210 pF
16 Turns
Tap 11 Turns
6 Turns
T120-6 Core

The coil is 16 turns on a T120-6 core. The load winding is 6 turns over the main resonated winding. The coax centre feeds into the resonated winding via a 6-210 pF polyvaricon (both gangs together in parallel) at a tap 5 turns down from the "hot" side. The cold side is returned to the coax braid. Another polyvaricon is used to resonate the main winding, also using both gangs paralleled.

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

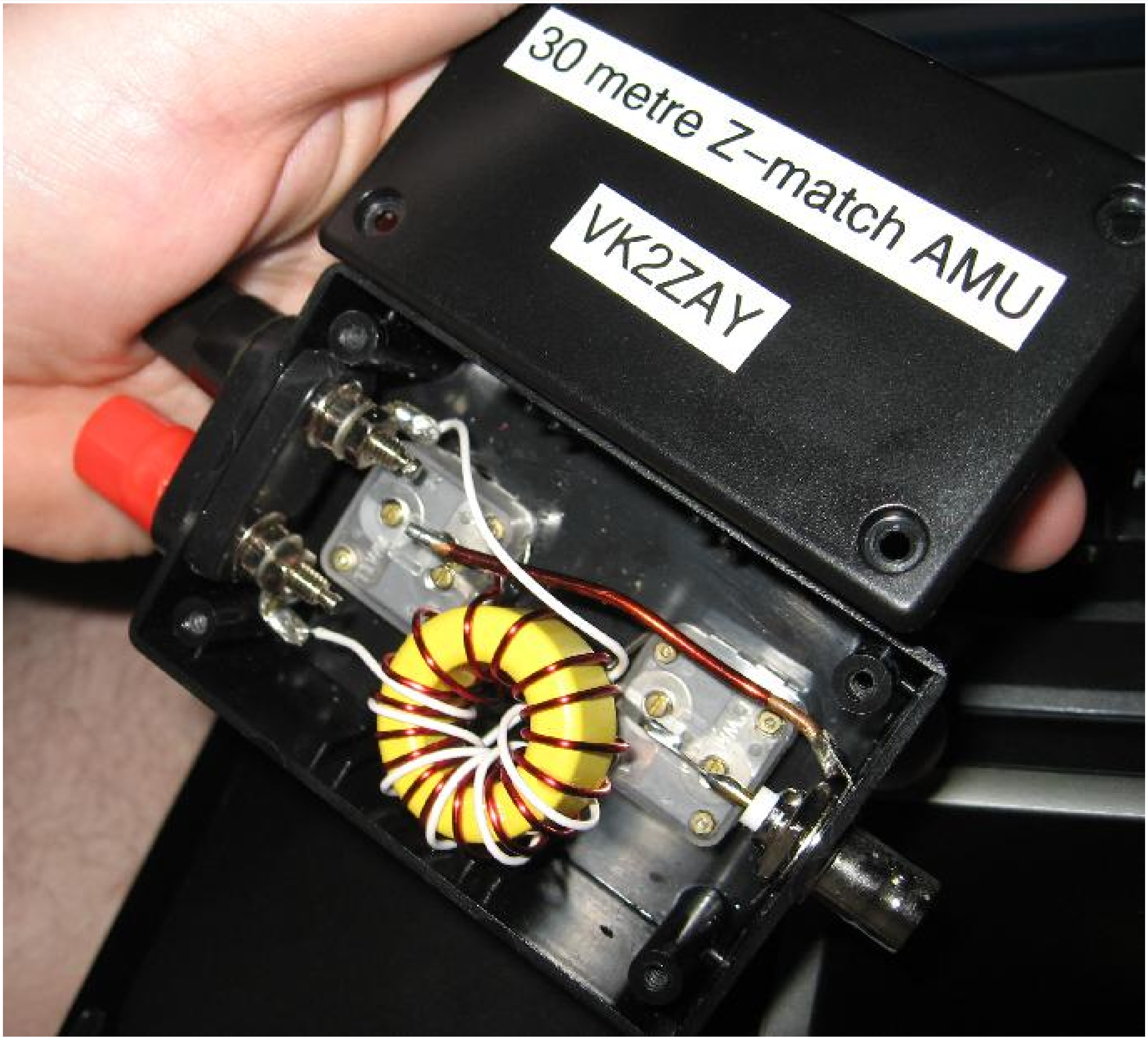

Achieving a match is an iterative process, but quickly finds very good return losses. I typically just hang the impedance bridge off the feed point while tuning up. No explicit balun is implemented, but current balance in the dipole appears pretty good (based on fluorescent tube experiments - see below) - that said no attempt to measure braid current has been made but it doesn't seem significant.

The Vertical

As detailed previously the 3 metre aluminium radiator used for the 80 metre beacon is matched simply with a polyvaricon (120 pF gang only) varying its effective capacitance as seen by the 7.8 uH loading inductor. The inductor is wound on a piece of ~1" OD PVC conduit.

© 2000 - 2009 Alan Yates - All Rights Reserved.

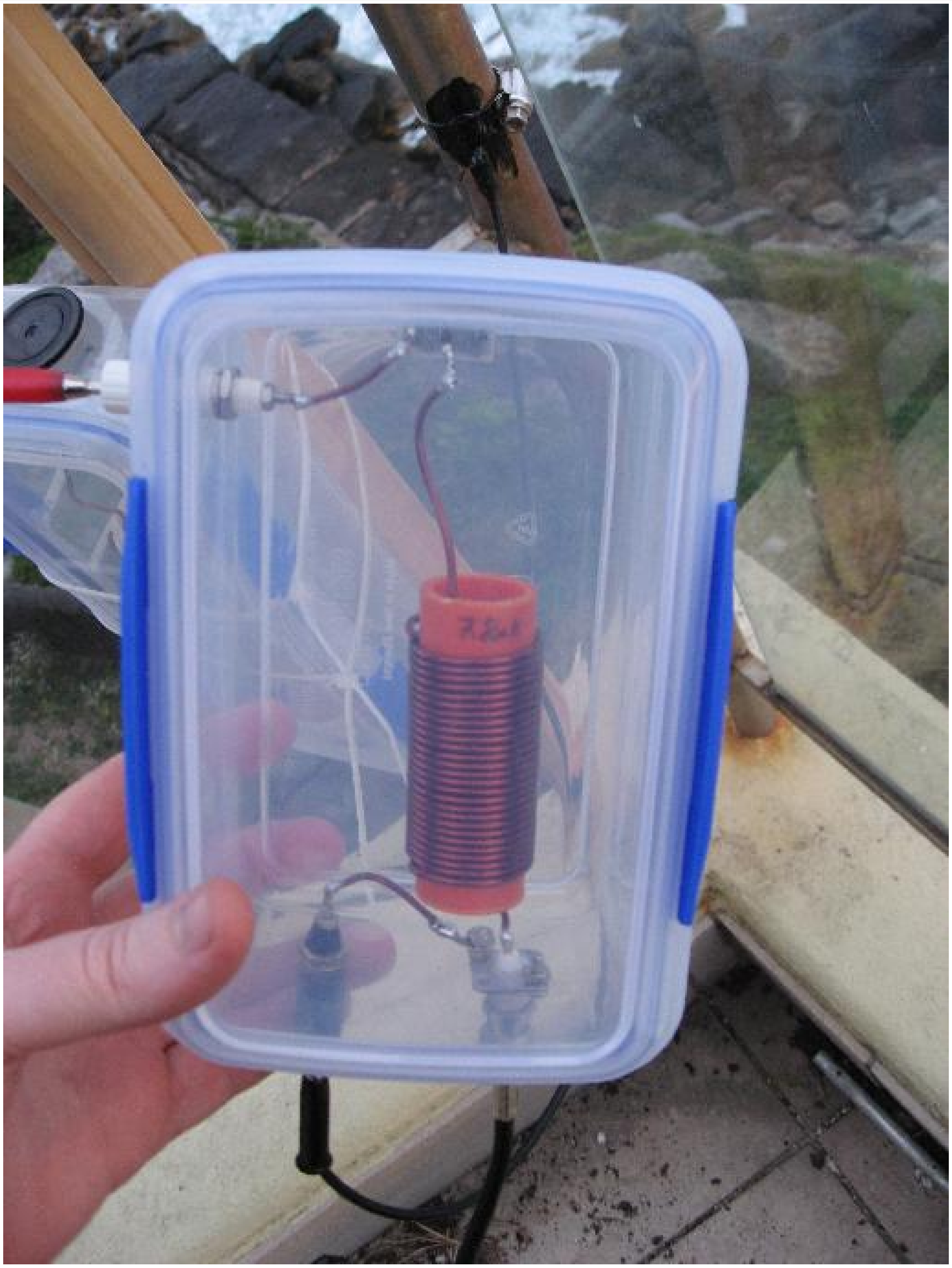

Once the capacitance of the whip was tuned out it appeared as almost exactly 50 ohms resistive. This took me by surprise, I expected a much lower value. At the time I blamed excessive ground losses, but even then that didn't completely make sense as on 80 metres the loss wasn't anywhere near that bad and the increase in radiation resistance with the higher frequency couldn't explain the high feed point impedance either. It worked, so I choose to ignore the odd result... Big mistake, science 101; when your results and model disagree check both!

Antenna Interaction

While initial QRSS transmitting experiments were carried out on the dipole (the vertical matching not yet attempted), all the recent transmissions have used the vertical. The signal seems better in practice, but I was curious how much the two antennas interacted, especially after the odd feed point impedance measurements of the vertical. When I switched over to using the vertical I simply disconnected the feed line to the dipole and plugged it into the vertical in close proximity. With the Z-match still tuned-up at the feed point all seemed fine, but I got a very good lesson in electromagnetics when I checked the current distribution using a fluorescent lighting tube!

© 2000 - 2009 Alan Yates - All Rights Reserved.

Field Around Dipole Loading Coil
(5.918 Mbytes)

With the vertical energised with only 1.5 Watts from the beacon TX I could easily light the tube in proximity to both dipole antenna loading coils. Clearly the antenna was parasitically absorbing a lot of the near-field of the vertical, no doubt re-radiating it and distorting the pattern of the vertical.

Field Around the Other Loading Coil
(3.862 Mbytes)

Experiments have just begun into the nature of the interaction, but detuning the Z-match or disconnecting it reduced the antenna current to negligible levels. Unfortunately this causes the feed point resistance of the vertical at resonance to drop enormously and offer a very poor match to the feed line. Remote signal levels seen in Grabbers are inconclusive, signals seemed to improve immediately after the change, but have since dropped off and now appear much worse! Only careful experiments with switching the resonance on and off could really tell for sure if the pattern produced by the interacting antennas is "better" for some destinations or not. Experiments such as this are complicated to implement because of the interaction itself and how much it changes the feed point impedance of both antennas. For an "apples to apples" comparison the antenna configuration in use must match the TX properly, accept all its power and not change the power the TX can generate (i.e. offer the same load).

Another complication was noted, the change in return loss of the vertical caused the beacon TX internal temperature to change, resulting in a visible drift of the beacon frequency by several Hz. A similar effect was noted when a freak sleet storm detuned the antenna enormously by covering it in icy slush. This is a form of drift I didn't predict after having gone to quite a length to minimise direct load pulling. I did have second thoughts during assembly about putting the amplifier inside the same box because of its heat generation, I had hoped it would just help keep the beacon above room temperature - which in fact it does quite well - but variations in dissipation with load condition is now something else to consider. Nature is often a very inconsiderate teacher, but a completely consistent one at least. No doubt people in the space program and other mission critical engineering applications have similar battles with what is totally obvious *after* it happens.

Fixing the Vertical

With the dipole no longer significantly interacting with the vertical I measured the input resistance of the vertical + loading coil at resonance as almost exactly 16 Ohms. An L-network was designed to match this to the 50 Ohm transmission line. The series inductance of the L network is small enough (370 nH) to absorb into the existing 7.8 uH loading inductance, leaving just the 458 pF shunt capacitance to be added. A 470 pF silver mica transmitting cap was soldered across the feed point allowing an excellent return loss from 50 Ohms to be achieved.

The feed point resistance once resonated on 80 metres was 22 ohms (must recheck this value my technique and instrumentation has improved since then). With a 500 milliohm radiation resistance on 80 metres this puts the antenna efficiency at about 2.3 % or about 16 dB down on a lossless 1/4 wave vertical. A similar calculation for 30 metres where the radiation resistance is about 4 ohms gives 25% efficiency or 6 dB down. 6 dB is one S-point - the loss is fairly trivial on 30 and not a massive deal on 80 either despite what might be called woeful efficiency.

How valid is this loss figure? If we take the 16 - 4 = 12 Ohm loss resistance for 30 metres and assume half of it is from the inductor finite Q then the Q figure that would give 6 Ohms loss is only 83 (XL ~= +j500). One hopes the inductor Q is better than this... Lets assume the inductor Q is at least 150 which seems reasonable, this is a

© 2000 - 2009 Alan Yates - All Rights Reserved.

ground loss seem reasonable? The balcony railing is Stainless Steel which means it has a higher resistivity than Copper, but also a deeper skin depth as a result. Physically it isn't large enough to be a "good" ground-plane, so 10 Ohms does sound fairly reasonable IMO.

For 80 metres where the XL is 1.4 kilo-ohms we have about 10 ohms more apparent loss. An inductor Q of about 117 gives 12 ohms of loss to add to the 10 ohms of estimated ground loss (likely not the same on 80). The geometry and construction of the 80 metre loading inductor could easily see its Q in this region. Measurement of the actual inductor Qs at the frequency of operation would be instructive to get a better estimate of ground loss, but it is obvious that the antenna efficiency is likely dominated by the ground losses on 30 metres and even the best efforts to improve loading coil Q on 80 won't improve the overall system efficiency all that much.

Other Notes

Because these are narrow-band devices and matching techniques, the return-loss seen on your instruments will rapidly tell you if you have a harmonic energy problem. A good return loss peak (or null if you prefer) will be almost impossible to achieve unless the transmitter is fairly spectrally pure.

My fluorescent tube ionises more easily than an NE-2 bulb - this was surprising. However its ionisation hysteresis seems larger. This is probably predictable as it is filled with an Argon/Mercury mixture rather than Neon and the pressure is likely quite different, but an interesting observation none the less.

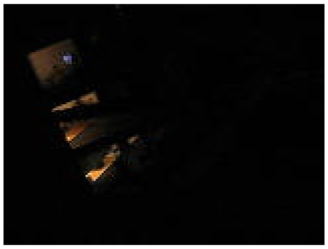

Ionisation Threshold Effect
(5.958 Mbytes)

The impedance of the 3 metre vertical is much lower on 30 metres than on 80, hence the Q is smaller and tuning up is easier and less touchy. Also the RF voltage produced at the same power level is smaller so Polyvaricons are quite suitable in the matching network. This also means my usual method of using an NE-2 bulb hanging off the mast clamp to tune for maximum output doesn't work. :(One of the big advantages of QRP is that you don't need insanely good components to withstand the voltages and currents associated with QRO operation, and you can't really burn yourself even from high-Z points.

Antennas aren't magic. I know that is stating the obvious, but like most Hams I think I am a little in awe of how a piece of metal the right size and shape can transduce electromagnetic energy with space. The laws of physics still apply and even your basic AC circuit theory is usable in the antenna domain. Radiation can make it seem like it is different, but if you model radiation much the same as loss (an extra resistance) it all works and makes sense. Antennas can be extremely confusing at times, and it is all too easy to omit the major radiator from your model! If experience has taught me anything about antennas (or RF in general) it is that you have to measure everything you can, confirm your assumptions and test your model. The more I work with antennas the more I am driven to build test equipment, especially impedance bridges, RF ammeters and (calibrated) field strength meters. To me this is what Ham Radio is all about, a technical hobby of life-long learning about the electromagnetic force of nature.

Leave a comment on this article.

Attachments

title	type	size
Z-Match Circuit Source	application/postscript	10.743 kbytes

© 2000 - 2009 Alan Yates - All Rights Reserved. XHTML | CSS | CSS toggle

End-Fed Half-Wave Antenna & Tuner

2007-03-11

For the on-going 2007 ARNSW Homebrew Group Challenge work I needed a simple 80 metre antenna. As a home-unit resident the size of a full half-wave is pretty impractical, but I decided to try it so I had a base-line to compare future shortened verticals to.

While centre-fed would be the simplest to get going and probably the easiest to install in a semi-permanent manner, I went with end-feeding. I would like to have the option of using the antenna portable, where a transmission line is just something else to carry and end-feeding is probably the easiest to set-up.

I read everything I could on the subject, the ARRL antenna book is good, but I found Steve Yates AA5TB's website especially useful. Steve explains the physics of the counterpoise well, and gave me confidence that it would indeed work as I had planned without stray-RF problems.

The Tuner

A matching box was constructed, containing both a matching circuit and a resistive 50 Ohm VSWR bridge that borrows heavily from Dan Tayloe N7VE (et al). The matching inductor value was picked to resonate the available variable capacitor gang just below 80 metres at maximum capacitance. The Hi-Z side floats to make it more versatile, a simple clip lead can be used to return the counterpoise side to the coax braid, if so desired.

The final device tunes a resistive load of 3-7 kilo-Ohms from 3.2 MHz to 12.6 MHz. 40 metres is covered (and 30 metres too, but I lack a transceiver for it currently), it is unfortunate that 20 metres could not be covered as well, without switching out some of the windings. W0CH switches his inductor taps, but this also changes the impedance match too much for my taste (Note that you can compensate somewhat for a few kOhms of resistive mismatch and an undetermined amount of reactive mismatch with the tuning). His circuit is otherwise very similar to mine (and just about everyone else's who has an internal bridge).

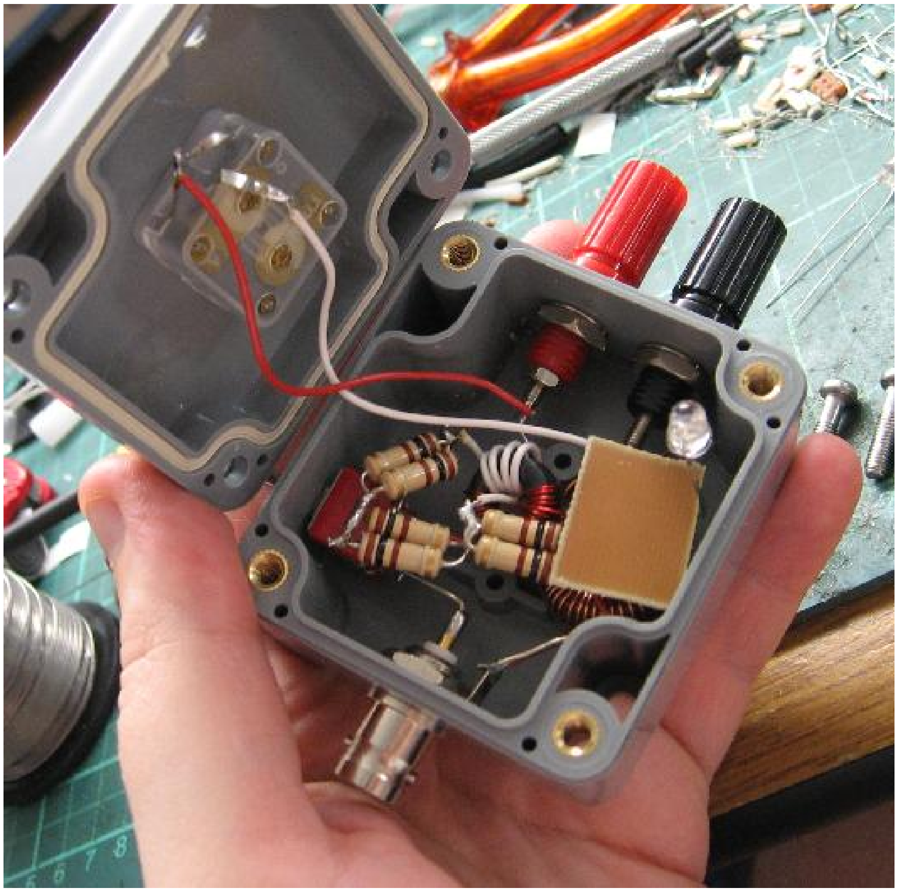

The unit was built in a small plastic box, and it was quite a squeeze to fit the tuner and the bridge inside. The resulting unit is very small and should work fine in portable operations. The tuner can be seen running about 1 Watt CW from the challenge transmitter into a 4k7 dummy load and a neon bulb which is getting stinking hot!

© 2000 - 2009 Alan Yates - All Rights Reserved.

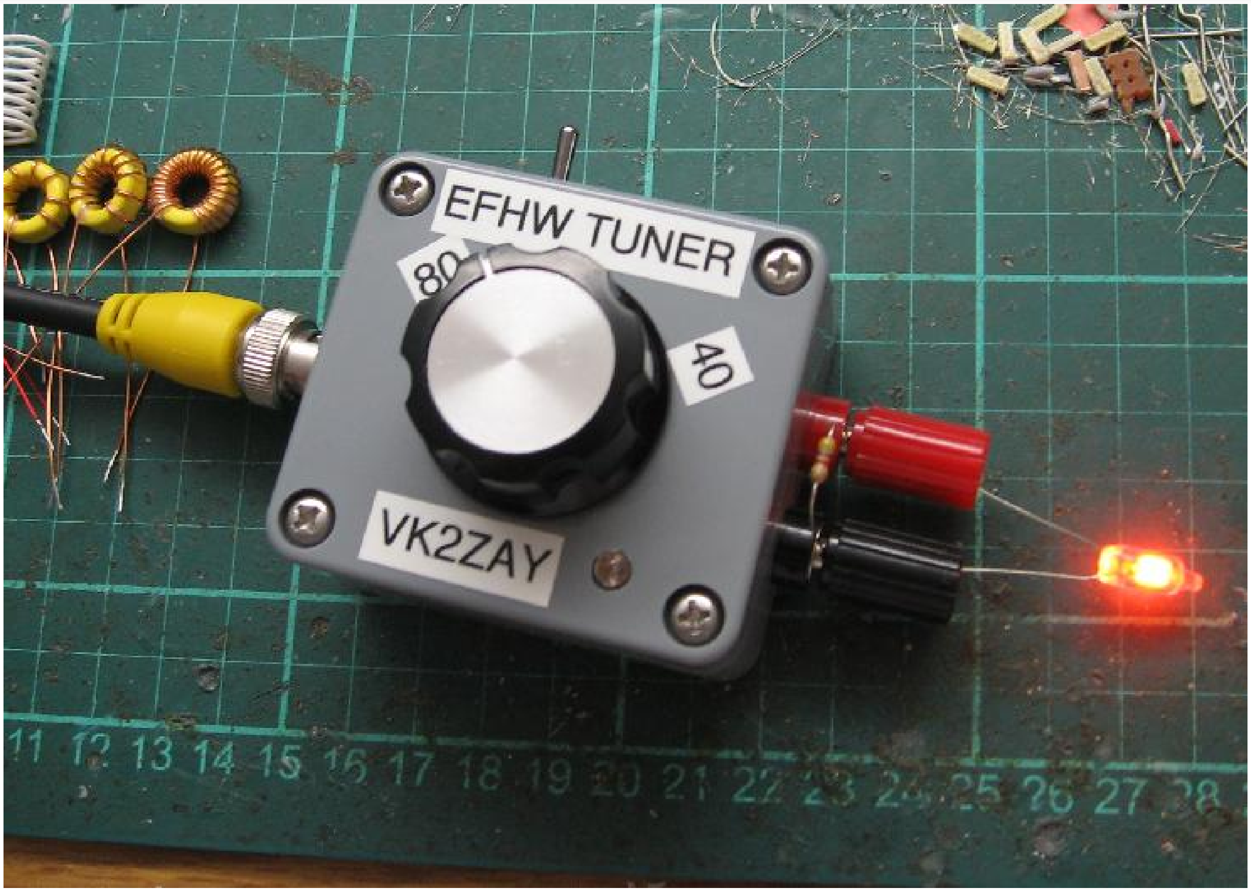

This gross abuse of a Neon bulb shows the voltage step-up of the tuner and proves it can feed a high impedance load fairly efficiently.

The Antenna

The radiator itself is simply a half-wave long piece of wire. I used an unzipped length of "zip cord", a full free-space half wave length (about 41.6 metres) which makes the antenna a bit reactive, but this is easily tuned out. Managing the wire while it was measured and unzipped was actually one of the more challenging parts of the project, but with the help of the XYL and the parents it took only 20 minutes or so. Black plastic egg insulators from the WICEN stand at Wyong were strung on the wire, two fixed at each end and two floating to allow various geometries for deployment. Banana plugs were added to each free end of the wire, so either might be used as a feed point.

© 2000 - 2009 Alan Yates - All Rights Reserved.

Testing

After some trial and error it was found possible to just fit the antenna's enormous length on my property. The centre was strung up at my 3rd story bedroom window and the ends sloped down to the extreme boundaries, one being the car park railing at the edge of a cliff, the other being the adjoining property's fence. The car park end was the most practical to feed from, so I took some gel-cells, a few radios and a fold-up chair down to the cliff edge and set-up there.

The antenna works much better than any I've used before. The inverted-V configuration may have something to do with this. I used the railing around the car park as a counterpoise, but it was possible to run using just the radio and its coax as a counterpoise too, by clipping the counterpoise connector to the coax socket outer with an alligator clip lead. It was very easy to tune in either way, on both 80 and 40 metres. The ARNSW morse practice beacon was full-scale on my VR-500, were normally is is barely S1 on my north-south horizontal 40 metre dipole. With a homebrew 80 metre transceiver I could hear lots of DX and local QSOs, and I listened to the WIA broadcast using my 80 metre VXO receiver for the first time at a good signal level. As the sun went down the mosquitoes got too bad to stand any longer, so packed up and went back indoors.

I had to take down the antenna, it is simply too big to leave up all the time, if I want to keep the neighbours on side anyway. The car park end is also a bit of pedestrian hazard, so this antenna will only be going up for special occasions. The experiment was otherwise a complete success, and I am very impressed with the antennas performance. I also now have good figures to shoot for with a shortened vertical.

5 comments.

© 2000 - 2009 Alan Yates - All Rights Reserved.

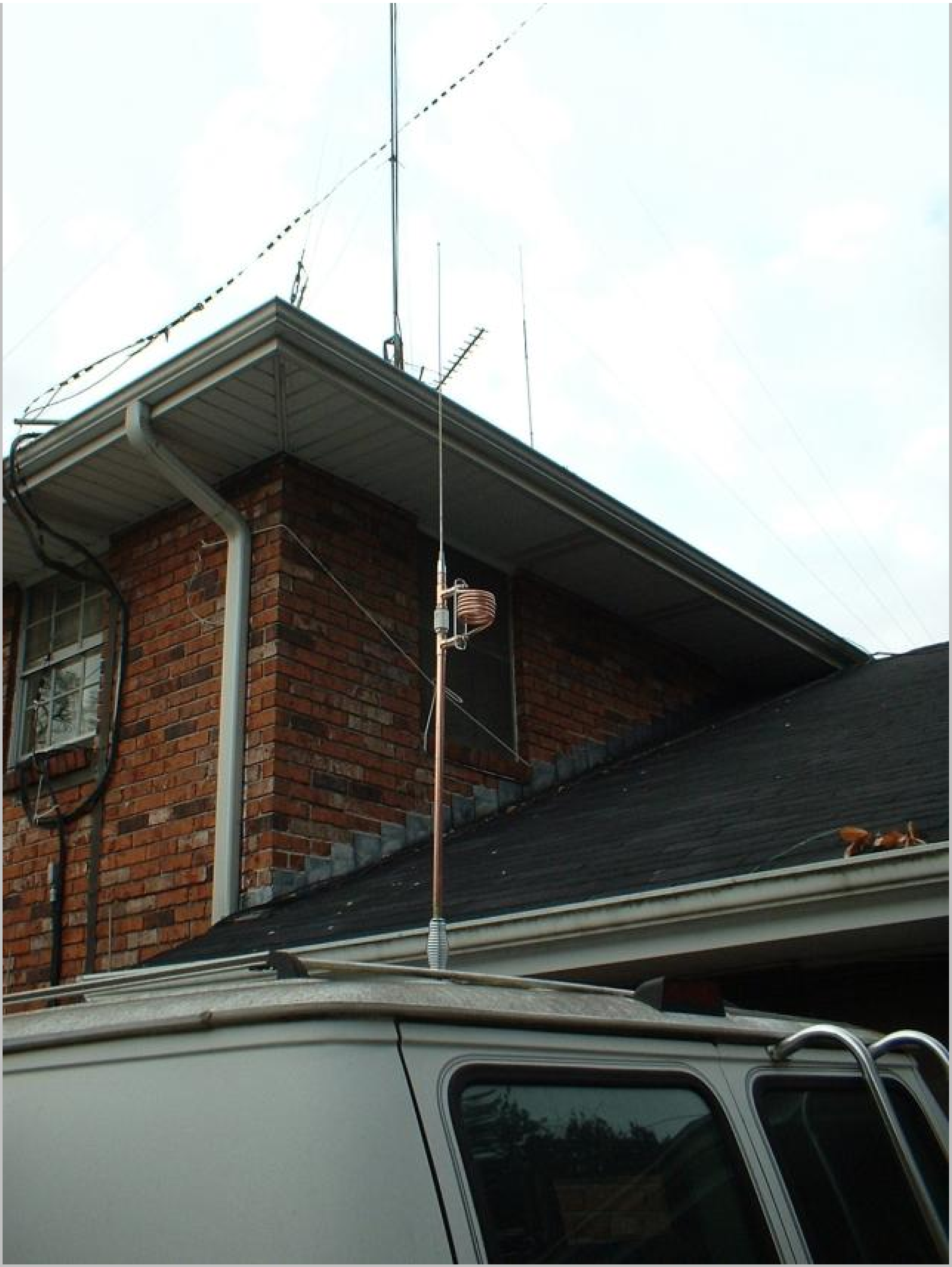

The original antenna was designed as a moderately short antenna for the 10 meter band.

I originally wanted an antenna that would not hit everything overhead while driving, since it was to be installed on a van. I came up with an antenna that extended only 1.65 meters or about 5 feet 4 inches above the top of the vehicle. It had a calculated efficiency of 42% when compared to a half-wave dipole. It was my hope to be able to install a series of different loading coils to be able to operate on the other HF bands. However, calculations indicated that the antennas was simply too short to work well as the frequency was lowered. Something better was needed.

The Evolution of the Antenna Design

My first thought was to increase the length of the antenna below the loading coil. My reasoning was that since that portion of the antenna carries the highest RF current, then adding more length there would increase the radiation efficiency of the antenna more than would adding the same length to the antenna above the coil. Unfortunately, since the antenna is mounted on top of a van which is a rather high vehicle, that would place the coil at a dangerous height with regard to tree branches and the like. Since the loading coils were going to be rather heavy due to their sturdy construction, placing a heavy coil high up on a flexible support did not seem like a good idea. I decided to investigate what increasing the length of the upper portion of the antenna would do to the radiation efficiency.

It quickly became apparent that very good results could be obtained by increasing the top whip length. As it turned out, this was due to two factors. First, the antenna was center loaded, and second, increasing the top whip length greatly reduced the coil losses due to the smaller number of turns required. This also meant that I could use larger wire for the coils, which further reduced the losses. It was almost like getting something for nothing!

Since I had once a standard CB-length whip plus the magnetic mount and base spring installed on the van, I knew that a total height of 3 meters was workable, even though it did hit quite a number of overhead obstructions. I ran calculations to determine how it would work. The results were very encouraging, and I promptly set out to build the antenna. I still had the CB whip, and figured that I could simply cut the top end off of it to get the total length of 3 meters for the completed antenna.

While laying out all the parts on the ground to see how they would fit together, I looked at the CB whip and realized that I would only have to cut off about 17 inches. That seemed like a waste, and I really hated to cut that small amount off of a perfectly good antenna.

Just for fun, I decided to repeat the calculations to see what would happen to the gain if I left the extra 17 inches on the antenna. I was surprised to find that the antenna efficiency increased between 19 and 31 percent, depending on frequency. The trade off, of course, was that I knew the antenna was going to hit a lot more objects overhead than if I trimmed the 17 inches off of it. As a test, I assembled the antenna full length and drove around with it for a few weeks to determined how much of a problem it would be. I decided that I could live with it, as a necessary price for the increased signal strength.

This table shows the results of the calculations. Notice the large jump in gain between the original short antenna and the 3 meter antenna. Longer IS better! <G>

Computed Radiation Efficiency of Center Loaded Mobile Antenna vs. 1/2 wave Dipole Antenna

TEST FREQUENCY, MHz	**Original 1.65 Meter Antenna**	**Proposed 3.0 Meter Antenna**	**Actual 3.42 Meter Antenna**	**% GAIN INCREASE 1.65 > 3.0 Meters**	**% GAIN INCREASE 3.0 > 3.42 Meters**
1.95	0.09	0.3	0.4	333	31
3.9	0.5	1.5	1.9	300	26
7.3	2.0	5.5	6.9	275	22
10	3.9	10	12.4	256	23

14.2	7.6	18.8	22.8	247	22
18.15	12	27.7	33	231	19
21.2	15.7	34.6	40.6	220	19
24.9	20.3	40.6	40.6	200	0
28.4	24.7	40.6	40.6	164	0

Construction Details

I was able to make use of the basic antenna design and most of the parts when constructed the final antenna. I'll show you in the following pictures how I did it. This is a very easy antenna to build, and you can do it with some simple tools and parts from the local home improvement or hardware store.

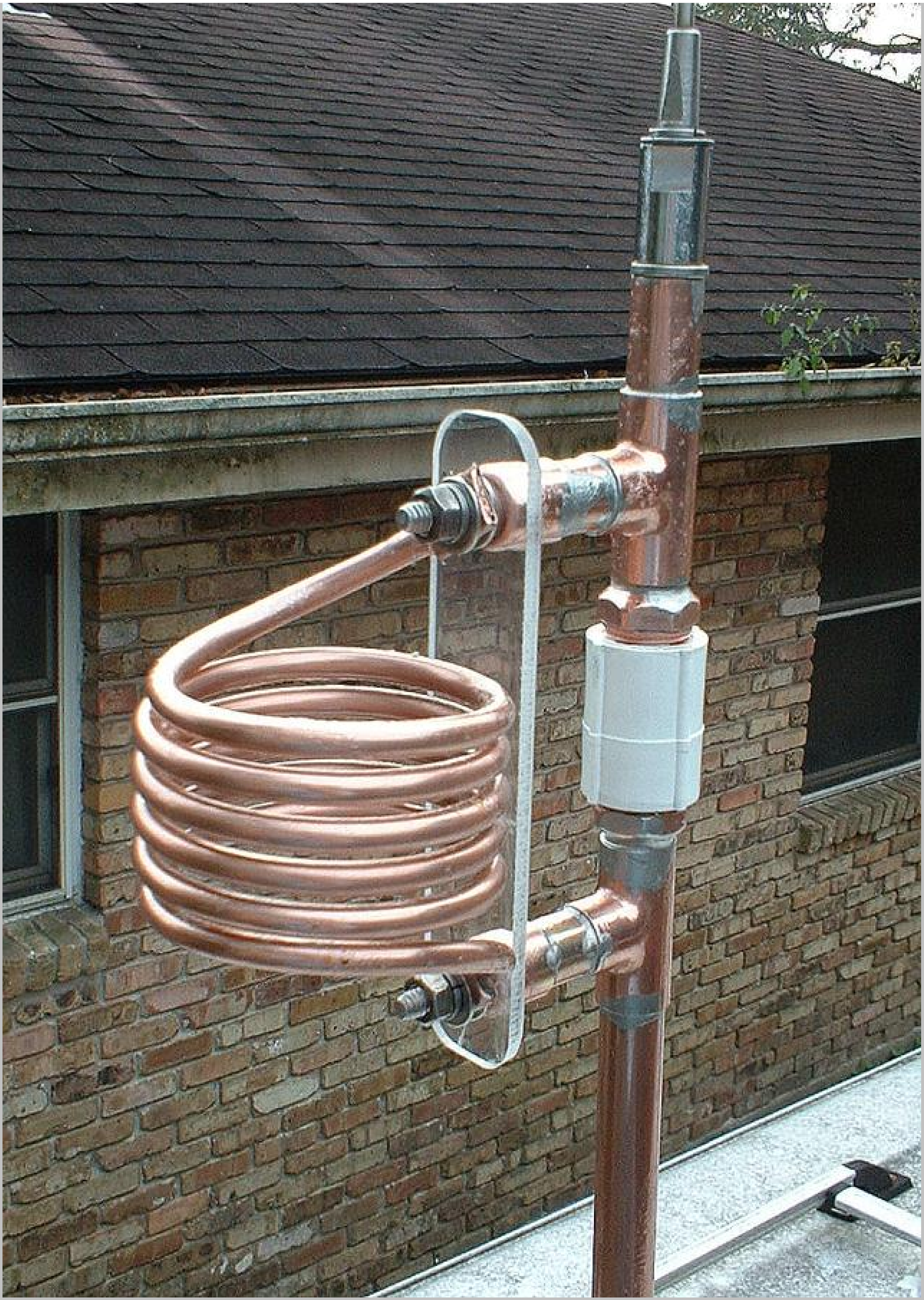

When I made the 10 meter coil for the first 1.65 meter long antenna, I used 1/4" OD copper tube for the coil. I inserted the ends of the coil into a short length of 3/8" OD copper tube and then flattened the ends of the tube. I then drilled a 1/4" diameter hole through the flattened ends so they could be attached to the antenna posts. This coil should have had one more turn on it. That way, the turns would have been spaced a bit further apart. This would have lessened the coil losses somewhat. However, the calculated loss in power was not enough to worry about, so I never bothered to make a new coil for the antenna.

See that length of clear plastic between the ends of the coil? When using a self-supporting coil like this one, you MUST use an insulating brace like the plastic strip shown here or the white insulator between the two antenna sections will fracture when you hit a big tree branch with the antenna!

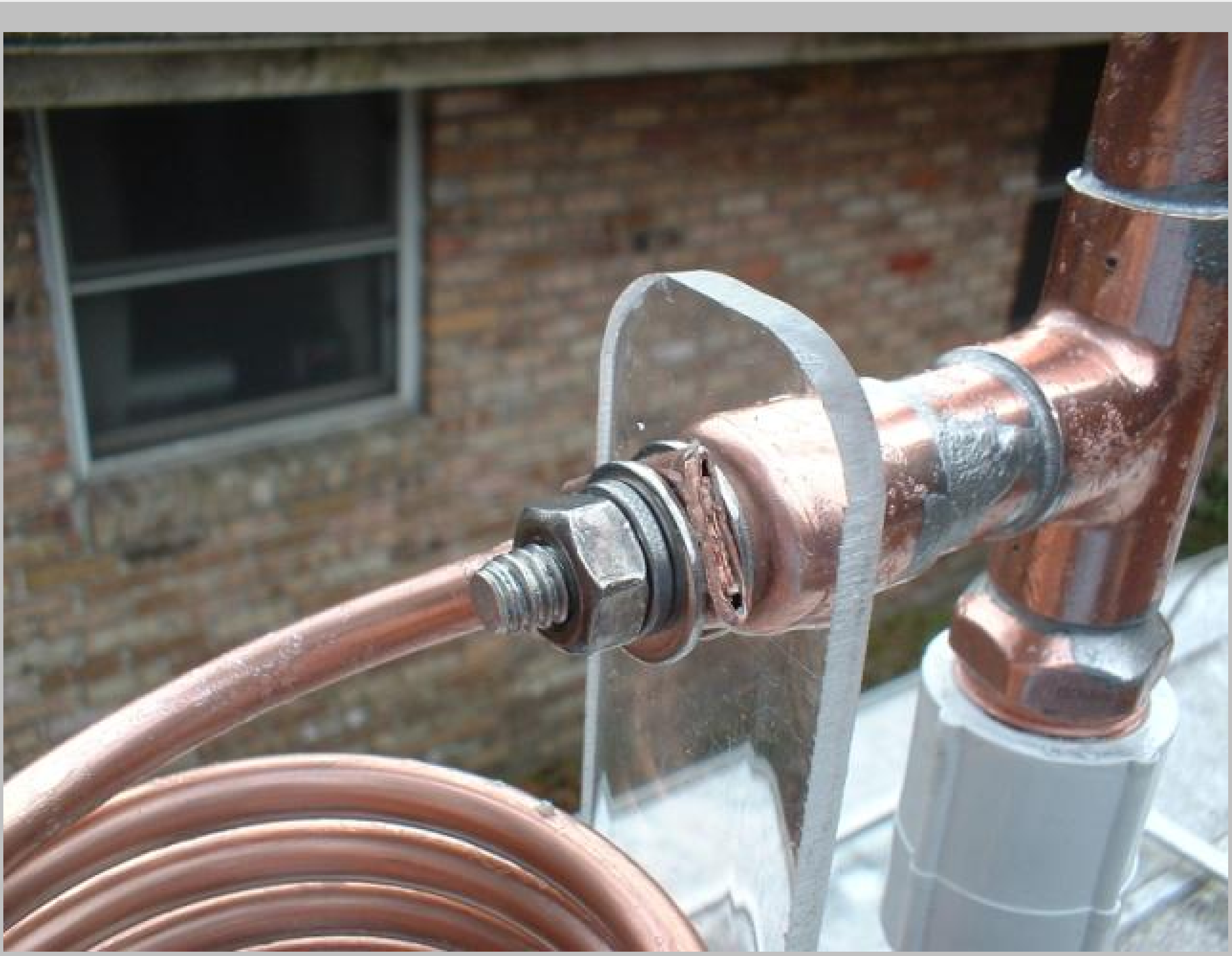

This is a close-up picture of one of the ends of the loading coil.

You can see the coil tube inside the larger flattened copper tube. The larger tube added structural strength to the ends of the coil and provided a larger area for drilling the mounting bolt holes through the coil ends.

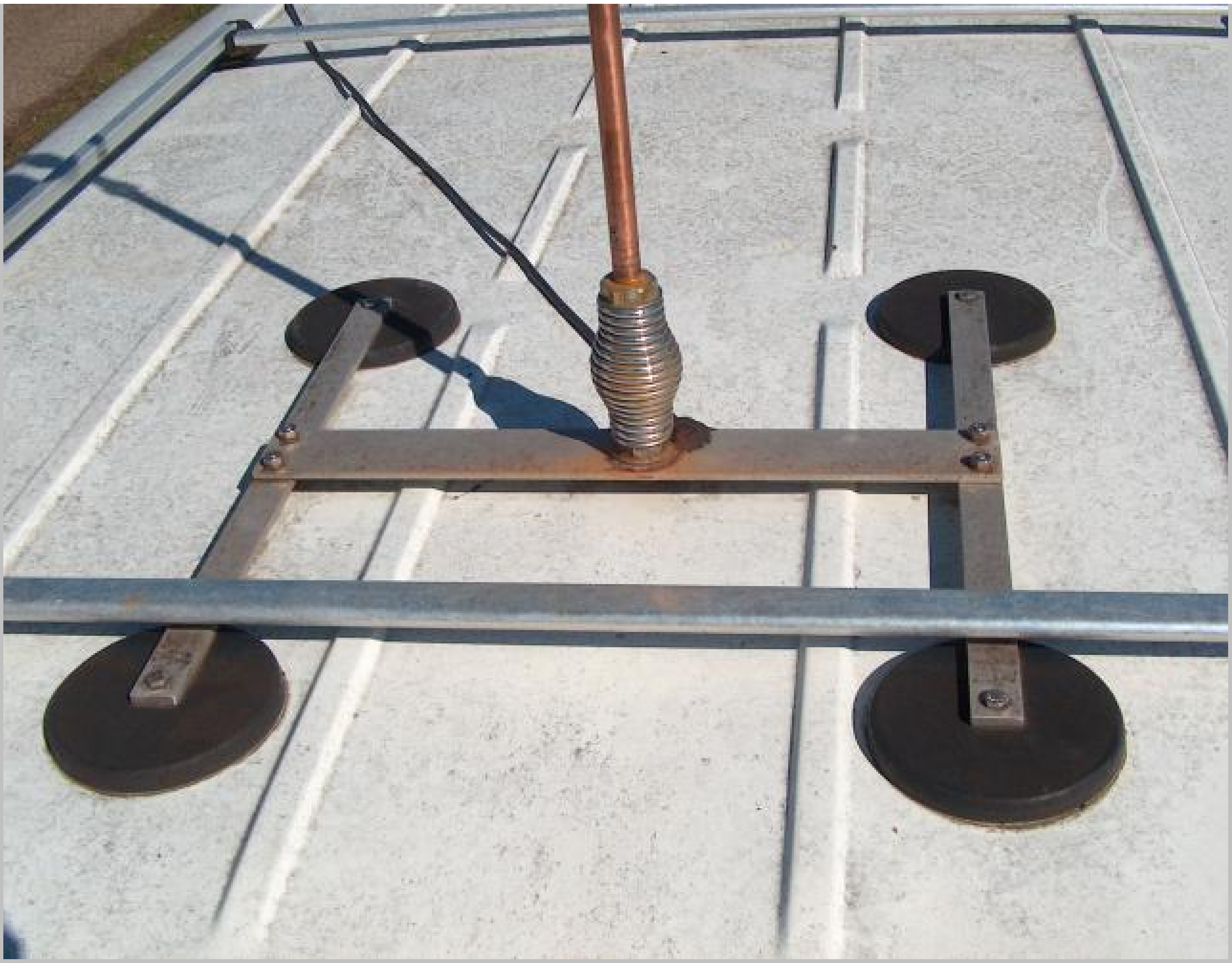

A big antenna needs a serious mag-mount!

I purchased this magnetic mount for my mobile antenna years ago over the Internet - I don't remember where I got it - but it holds to the roof like a barnacle on a ships' hull. You can fold the antenna over until it touches the roof of the van and the mag-mount stays put. It requires two hands and a short pry-bar to get it loose.

See all that rust on the horizontal bar? That's from a cheap steel quick disconnect I used for several years. It finally rusted to then point that it made intermittent contact and I decided to get rid of it. So far, I have not found anything strong enough to replace it. I have broken two fairly heavy brass quick disconnects so far. So, for now, I just either tie the antenna down to the roof of the van or unscrew it when I have to go into a low garage. Needless to say, I strongly favor outside parking spaces!

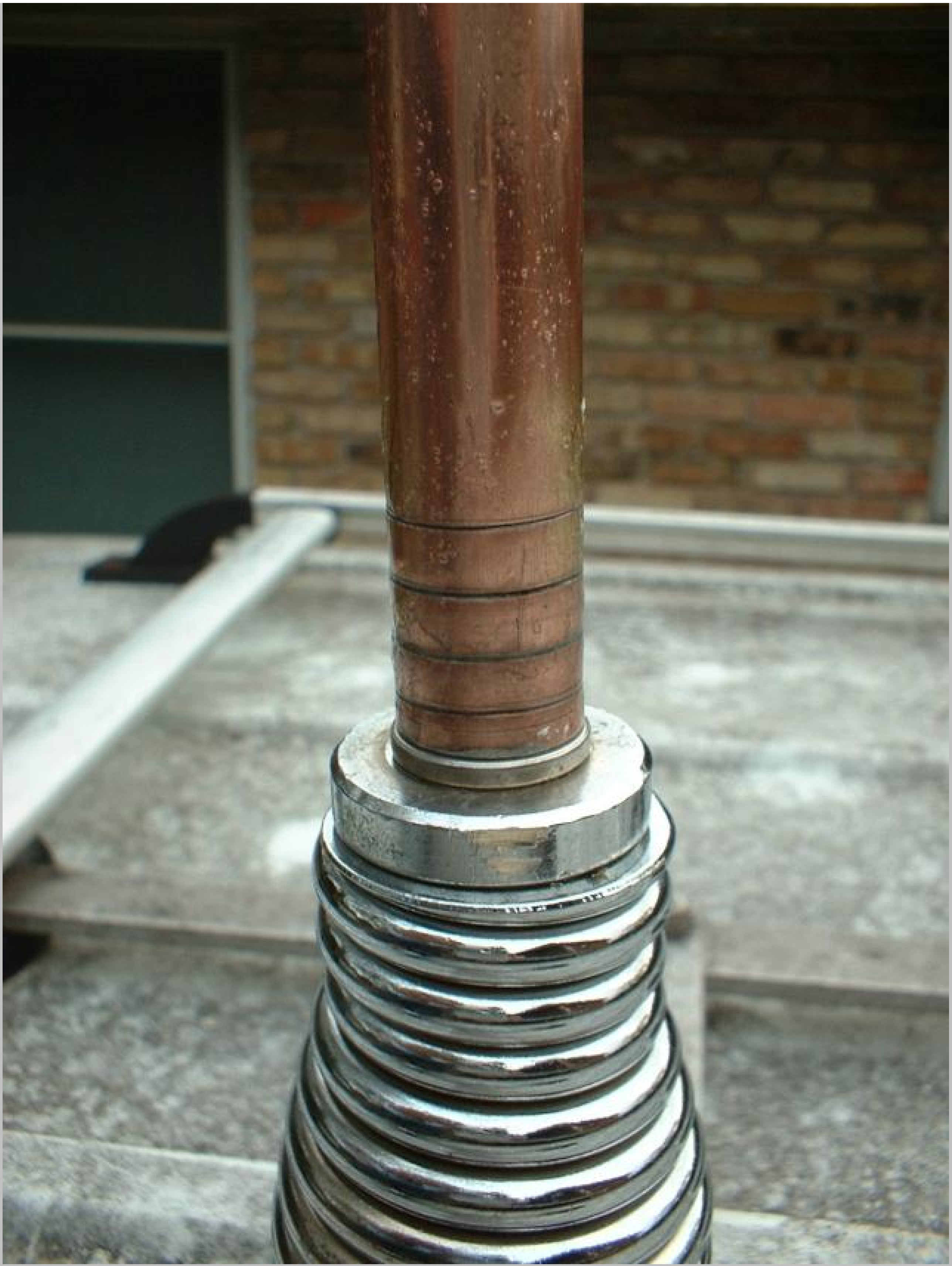

This picture was taken before I installed the quick disconnect. It shows the lower end of the bottom section of the antenna. The antenna mast sections are constructed from hard wall copper pipe, 1/2" in diameter. A brass plug with a 3/8" x 24 threaded hole was pressed into the end of the copper pipe and hard-soldered into place. To hold the plug in place during the soldering process, I rolled several grooves around the outside of the copper pipe as seen here. I was careful not to roll them too deeply and weaken the pipe.

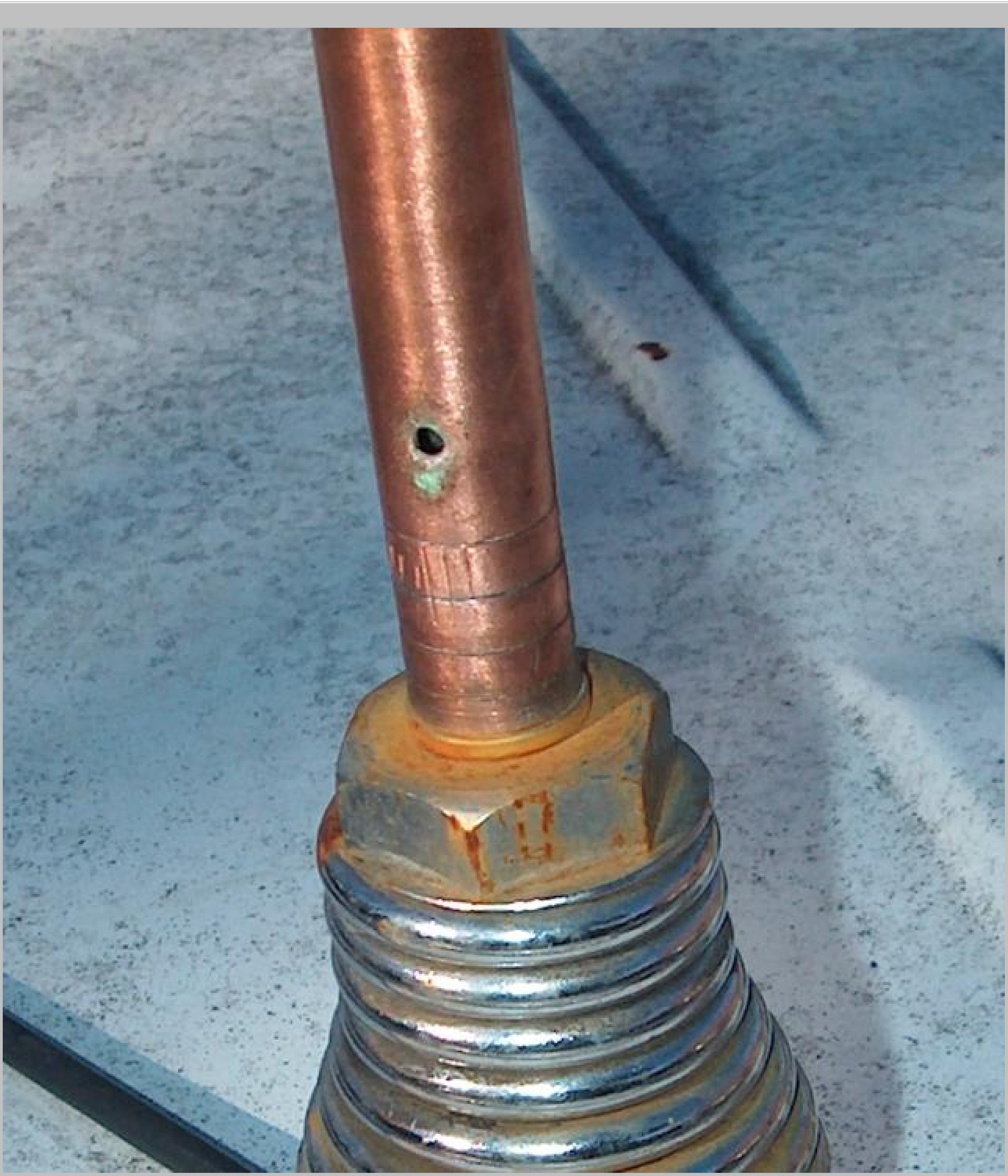

A small "weep hole" is drilled through the pipe wall just above the upper end of the brass plug that was soldered into the pipe. The hole allows water to drain out of the antenna and not sit inside and corrode the mounting bolt. Water WILL get into the antenna - you can't prevent it - so you might as well make provisions to allow the water to drain out.

To hold the antenna to the base spring, I used a length of threaded steel rod cut from a stainless steel bolt. After making sure that the threaded rod was the the correct length and that everything fit properly, I removed the mounting bolt from the base of the antenna and the base spring and used thread locking compound (OK, I used Super Glue, if you must know) to retain the bolt in the base of the antenna. That way I would not misplace the bolt when I had to remove the antenna from the base spring.

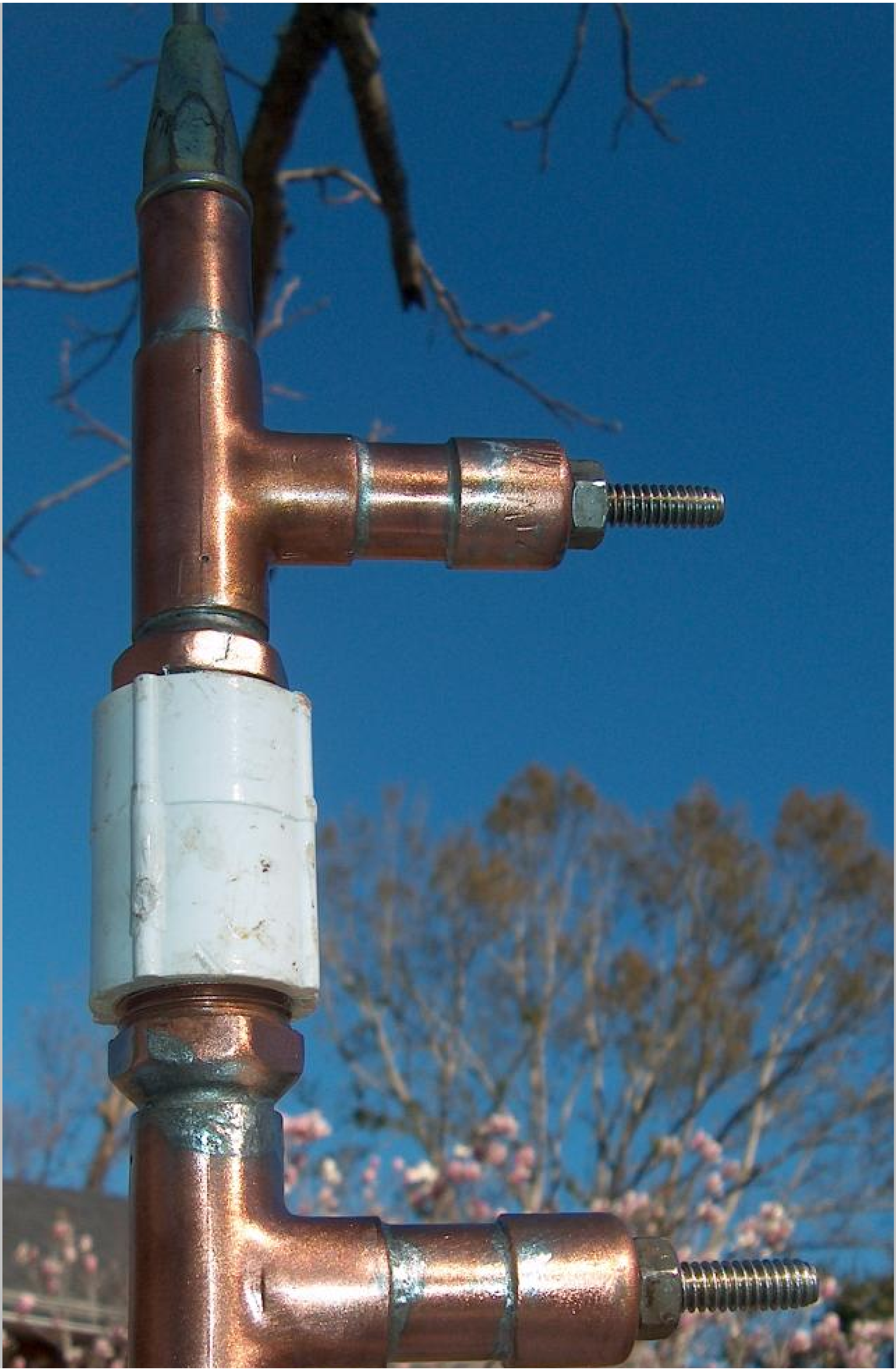

The Center Insulator and Coil Mounts.

The lower section of the antenna is a 1/2" diameter copper pipe about 60 cm (24 inches) long that extends upward to the lower coil support.

The support itself is made from a 1/2" sweat "T" fitting, a 2" length of 1/2" copper pipe and a 1/2" sweat cap.

The bolts holding the coil in place are 1/4" diameter x 1-1/2" long stainless steel bolts. A 1/4" hole is drilled through the center of each pipe cap before soldering it in place, and the bolts are temporarily held in place by gently tightening the nut visible in this picture.

The plastic center insulator is a 1/2" female-female threaded PVC coupler.

Each end of the antenna sections that thread into this coupler has copper adapters that go from 1/2" OD copper pipe to 1/2" iron pipe. These fittings have a male thread on them so they will screw into the plastic coupler which then becomes the antenna center insulator.

Using plumbers hard solder (not rosin core radio solder) and the proper cleaning flux, the copper components are soldered together.

Next, the interior of the copper sections is thoroughly washed with clean water to remove any soldering flux residue and then placed in the sun or some other warm place to dry.

When everything is dry, the nuts on the coil mounting bolts are firmly tightened, and the sections may be screwed into the center insulator.

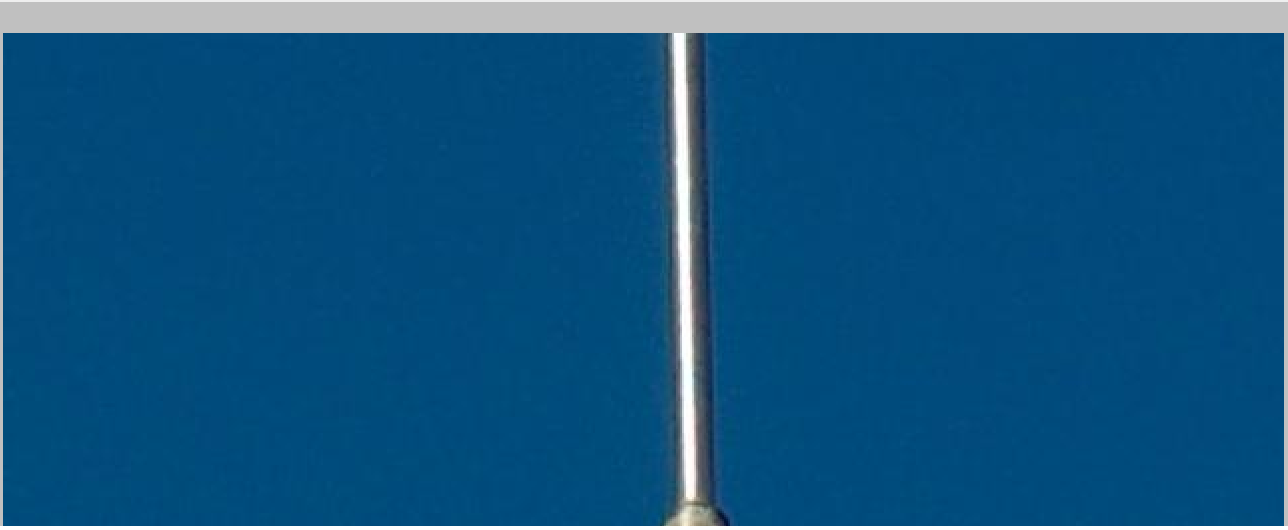

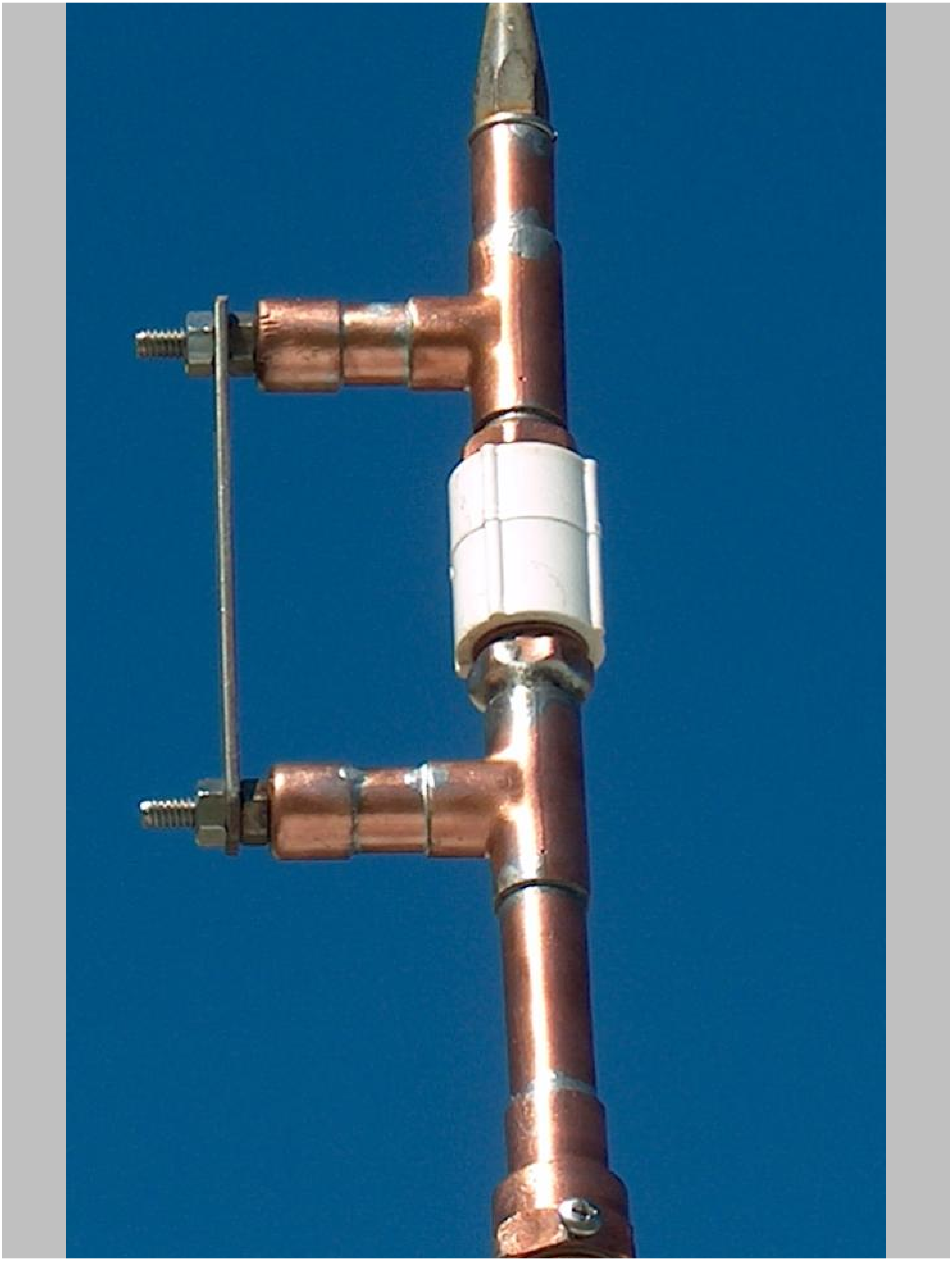

The coil support posts shown with the 10 - 12 - 15 meter shunt strap in place.

Note that with this shunt strap in place, the plastic strengthening strip is not needed. This is because the shunt strap is made from a length of heavy silver-plated copper stock.

The top support for the loading coil is fabricated in the same manner as the bottom support.

To hold the top whip, a 2" long length of 1/2" OD copper pipe is soldered into the "T" fitting. Another female threaded brass plug is inserted into the top end of the 2" long pipe section and soldered in place.

Hey! What's that extra set of pipe fittings doing, and why are they there? (Keep reading for the answer!)

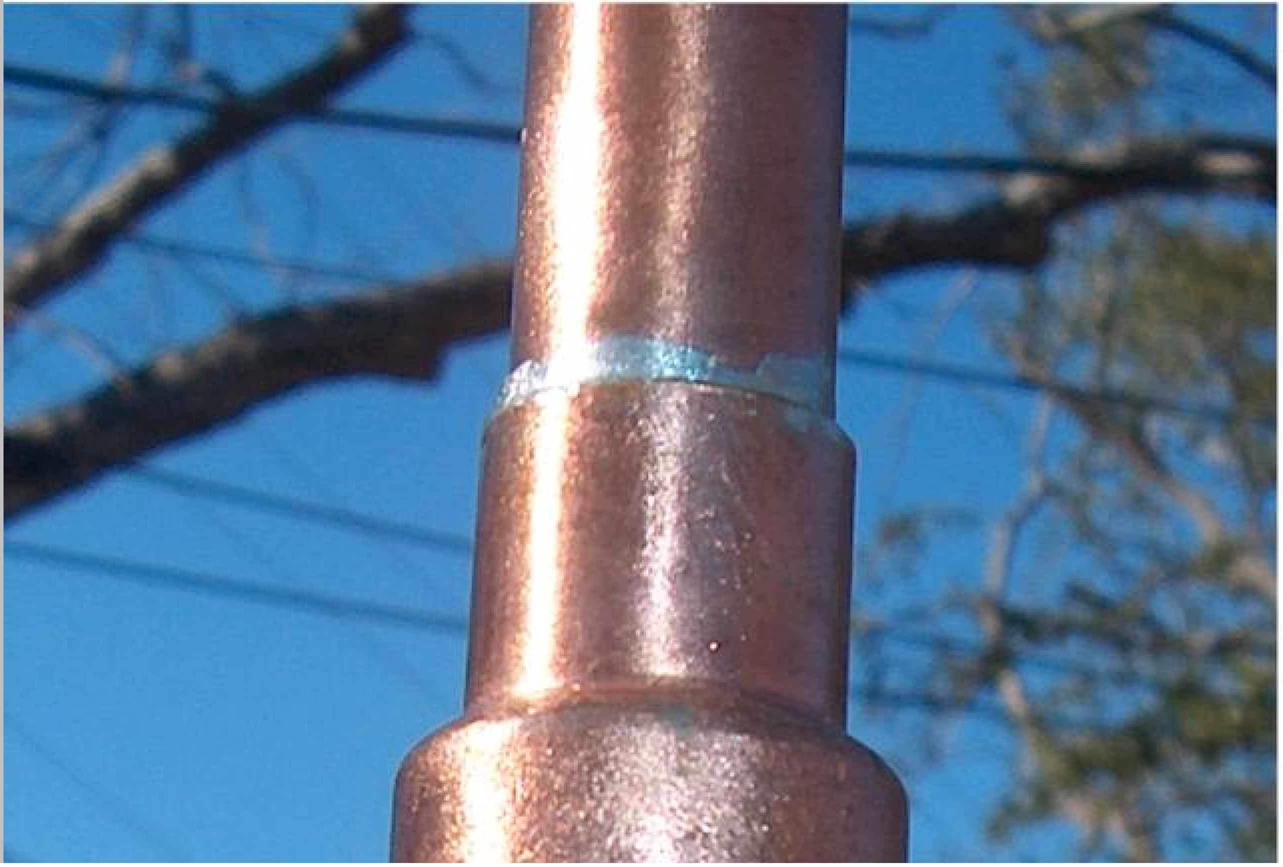

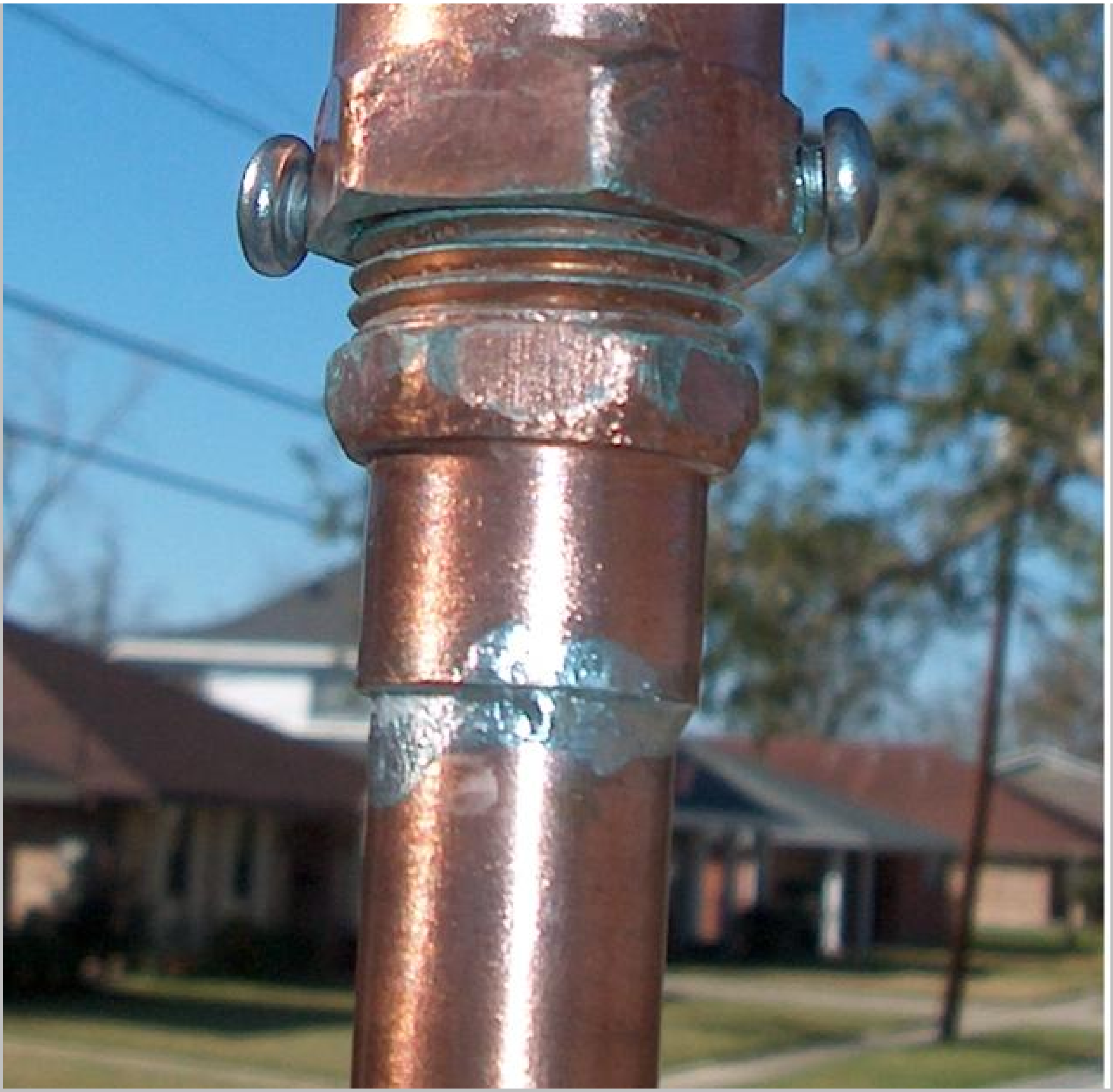

A (very) close up photo of those extra pipe fittings.

I always had a problem getting the antenna loading coil lined up "just so" on the mount. When you screw the antenna into the mount, you never know where the loading coil will be pointing when the mounting screw is tight. With coaxial mounted loading coils, this is not a problem, but this antenna has the loading coil mounted off-center from the mast. Provision needs to be made to adjust the coil position. How come?

Well, the loading coil must always "trail" the antenna mast; that is, the coil must be directly behind the antenna mast when the vehicle is moving. The reason is rather obvious - after you hit the first big low-hanging tree branch! If the coil happens to be in front of the antenna mast, the coil may become snagged on the tree branch and instantly becomes part of the local roadside litter. When the coil is mounted behind the mast, the mast simply slides harmlessly beneath the tree branch, the coil does not get hit, and all is well.

In order to accomplish this alignment without resorting to the use of shim stock, various thicknesses of washers, and all sorts of other chicanery, I decided I had to have a way to be able to rotate the coil around the mast in some way. Since everything was soldered together, I came up with the idea of using a pair of threaded mating fittings that I could simply twist to get the alignment exactly correct. I drilled and threaded a pair of holes through the outside (female) fitting so I could use a pair of stainless steel screws to lock the fittings in place after after the adjustment was complete.

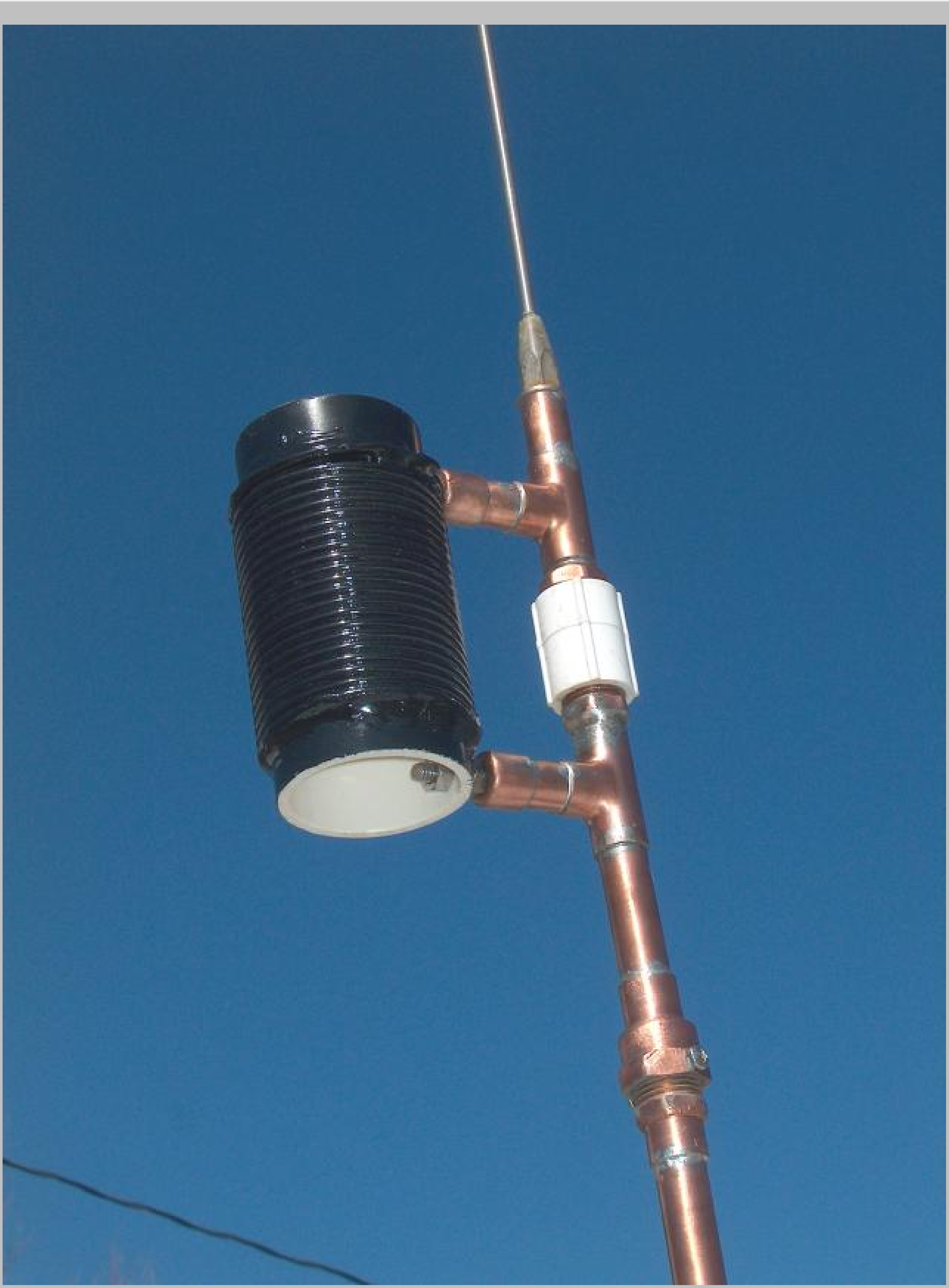

The 40 Meter loading coil in place on the antenna.

Notice the stainless steel nut at the bottom of the coil. This nit, and another one at the top of the coil holds the coil on the antenna. Changing coils is easy. Simply remove the nuts at the top and bottom of the coil, swap coils, and replace the nuts. Tighten firmly - but not excessively - and the PVC plastic acts as a lock nut to keep the coil in place. Still, it's a good idea to carry a few spare nuts in the vehicle in case you drop one while changing coils.

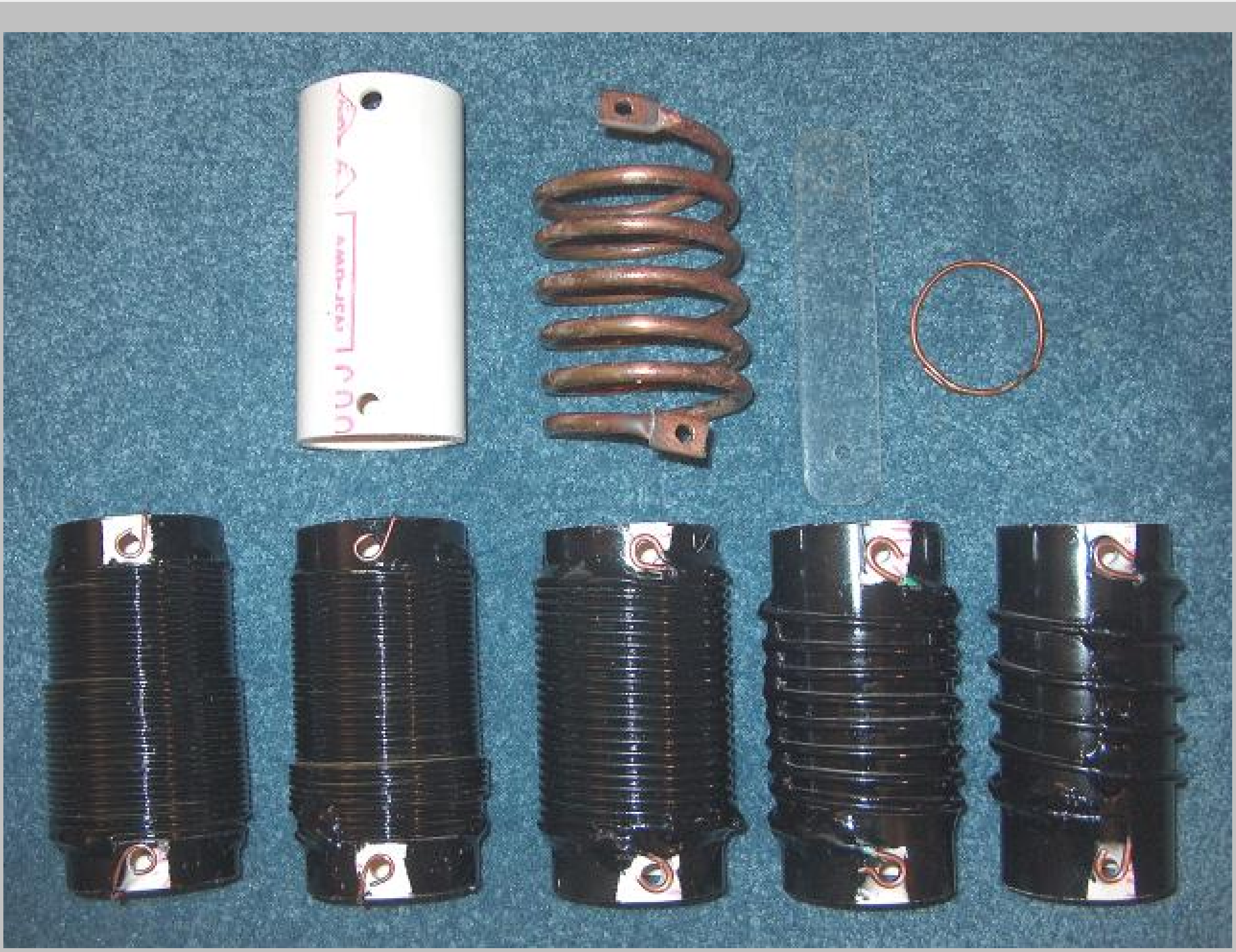

The complete Coil Set for the antenna.

From left to right -

Top Row: coil form made from 2" OD schedule 40 PVC pipe; old self-supporting 10-meter coil made from 1/4" copper tubing; plastic support spacer, used with self-supporting coils; copper shunt ring, used to tune loading coils.

Bottom Row: loading coils for 160-meters, 75-meters, 40-meters, 20-meters, and 17-meters.

The 160 and 75-meter coils are wound using #14 AWG Nylon insulated wire; all the rest of the coils are wound with #10 AWG THHN insulated wire. Note that no terminals are used on the coils - the wire ends are simply wrapped around the mounting bolts. When the coil is attached to the antenna, the wire loops are pressed against the copper coil mounts to make the electrical connection by the coil form when you tighten the 1/4" nuts from the inside of the coil form.

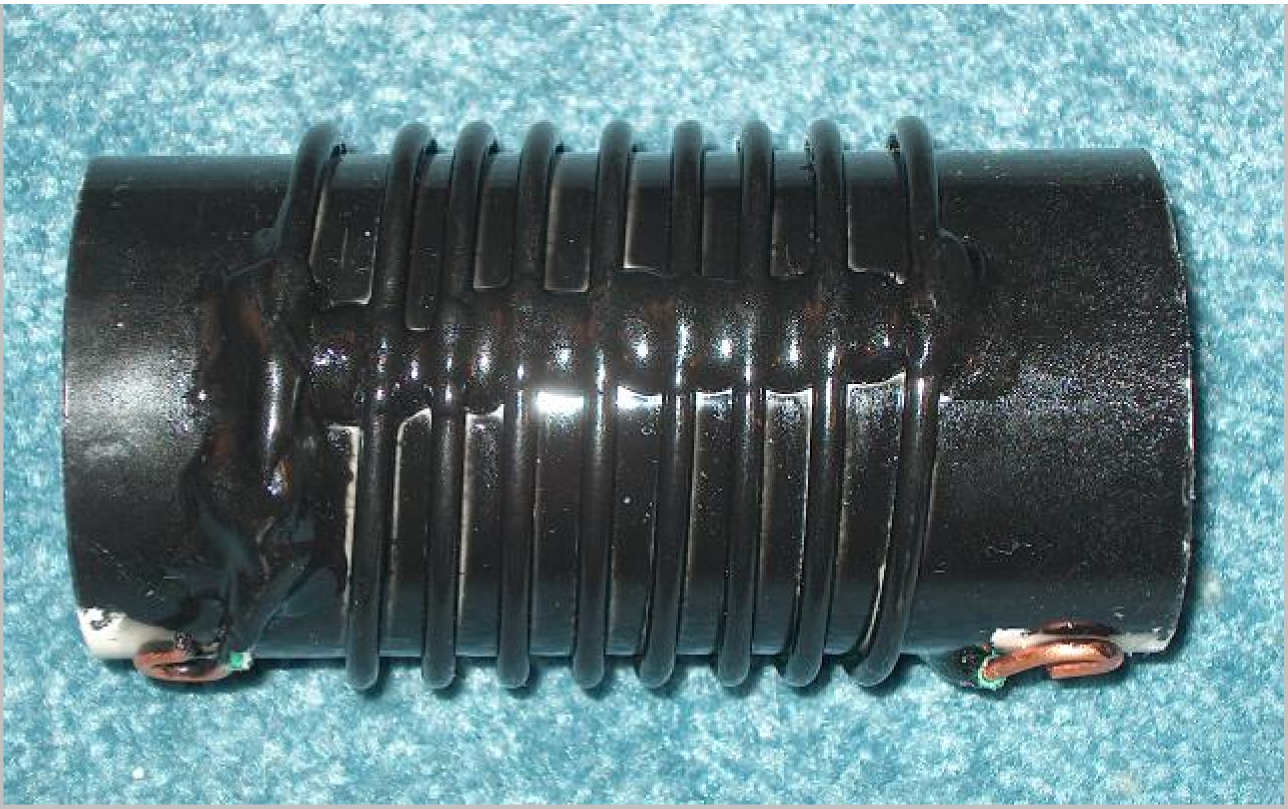

General Coil Construction

The coil form is made from a length of 2" OD PVC pipe. I cut each coil form about 3/4" longer on each end than the spacing between the mounting bolts. NOTE: Make sure that each coil form fits easily over the mounting bolts before winding on the wire or you'll have problems changing coils.

After winding the wire on each coil, an application of epoxy adhesive (J-B Kwik) was used to keep the coil turns in place. A coat of black spray paint was applied for appearance and to make the white plastic coil form less noticeable. NOTE: Apply the epoxy AFTER you have adjusted the coil to resonance.

If you just wind the wire on the coil forms directly, this will result in the coil "springing back" and becoming loose on the form when you release the wire after winding it. I wound several extra turns on each coil as I wound it on the form, then removed the coil from the coil form and then squeezed it down around a slightly smaller form (actually a spray can of insect repellent.) I then removed the now slightly smaller diameter coil and gently screwed it onto the coil form, where it remained tight enough to stay in place properly.

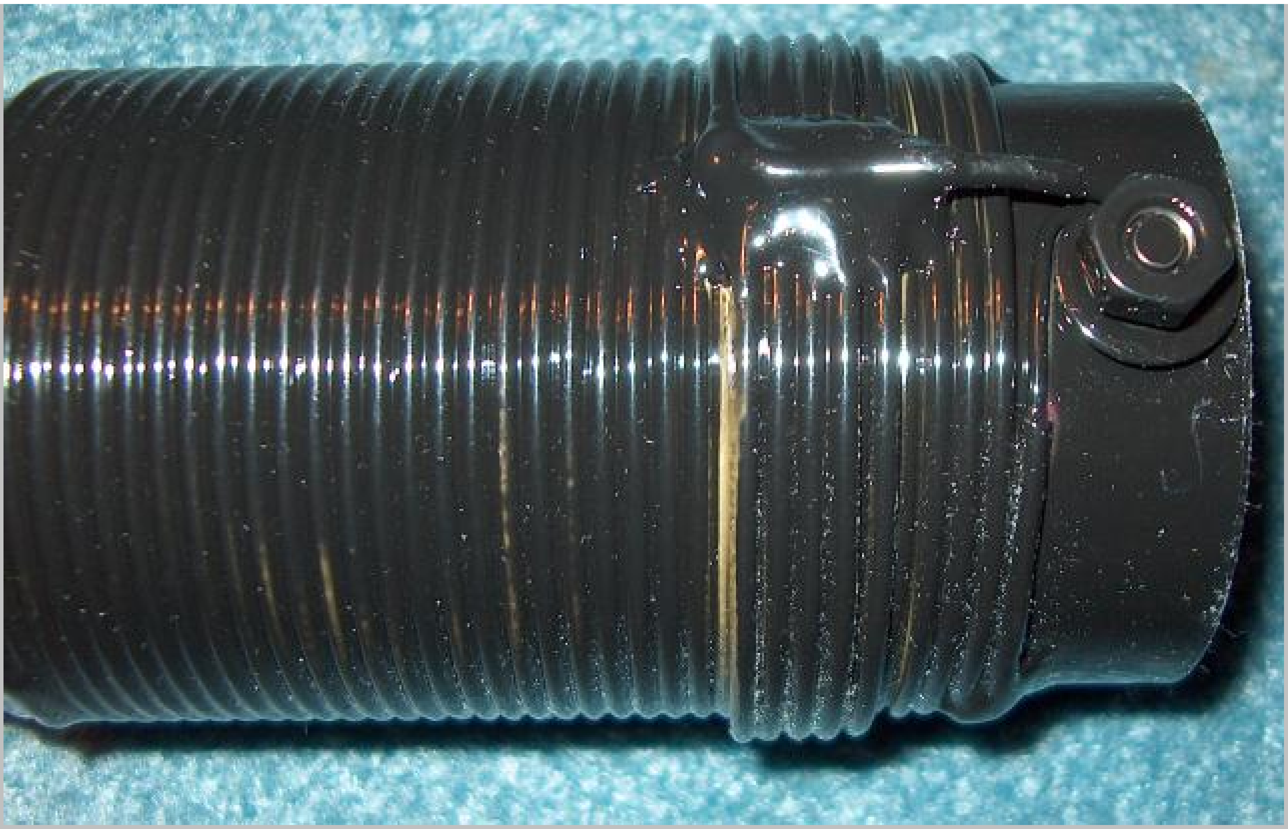

The 75-Meter coil

The calculations indicated that I would need to use #22 AWG wire to get enough turns on the coil in the space I had available. The losses would have been quite high, if I used that small size wire. So, this coil is wound with #14 AWG wire. Since using larger wire would not allow as many turns in the same length, I could either use a larger diameter form (too much wind resistance and negative "eye appeal",) or a longer form, continuing the extra turns below the lower mounting bolt (wind resistance, and higher RF losses.) Something better was needed.

Remembering my switching power supply transformer winding experience, I decided to try winding the loading coil for the antenna in the same manner, that is, complete the first layer of the winding in the usual manner, then "jump" the winding end back up towards the start of the first layer and then continue the second layer of the winding towards the end of the first layer of the winding. This is sometimes referred to as a "Z" winding.

You can see the black painted bolt in the right of the photo that holds the end of the first layer of the winding. From that bolt, the wire that begins the second layer "jumps" back to the left of the second layer of the winding. The end of the second layer is brought out to one of the mounting bolt holes. You can see the epoxy that holds the windings in place.

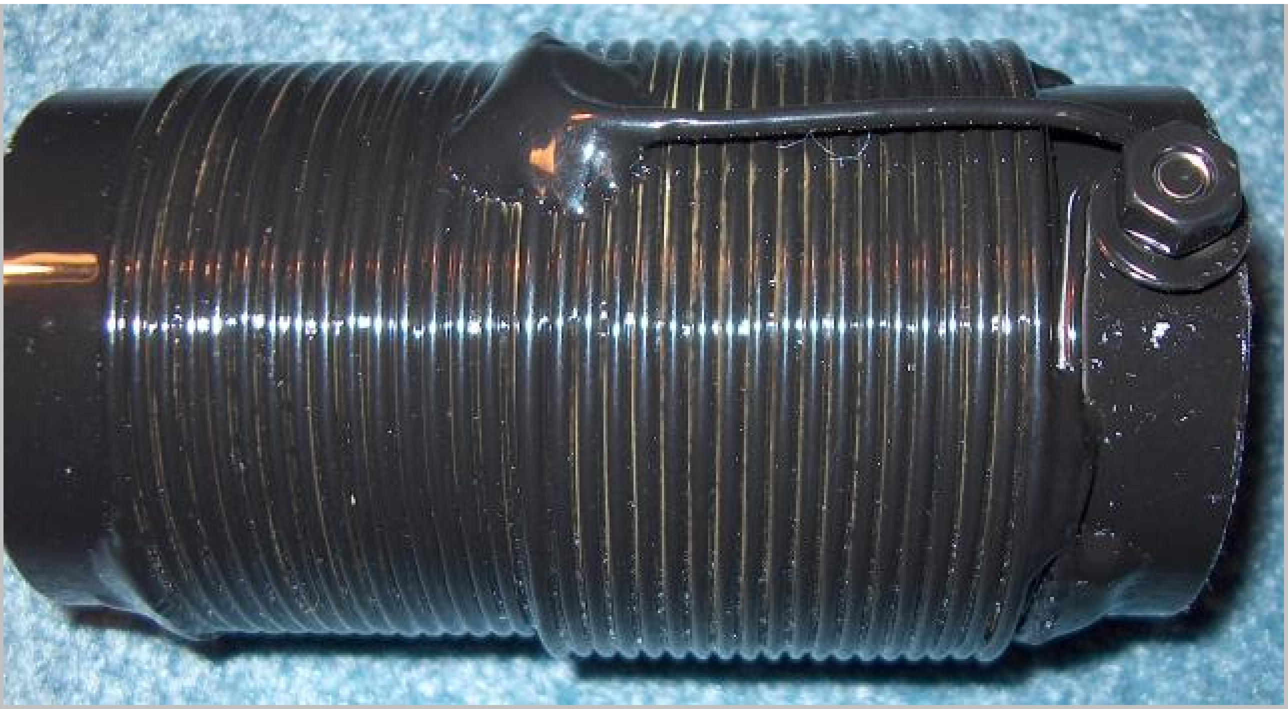

The 160-Meter Coil

Wound in the same manner as the 75-meter coil, the 160-meter coil required only 60% of the number of turns as would be needed if the windings were in a single layer. The trade off is that there are several stray resonances of the antenna system using this coil, but none of them cause any problems with normal operation. Since less wire is used in this coil than would be used in a single layer coil, the copper losses are less, but the dielectric losses are slightly higher due to the overlapped windings. However, the use of the "Z" winding method minimizes the dielectric losses as much as possible. The measured Rac loss of this coil is less than the originally calculated coil using #22 AWG wire.

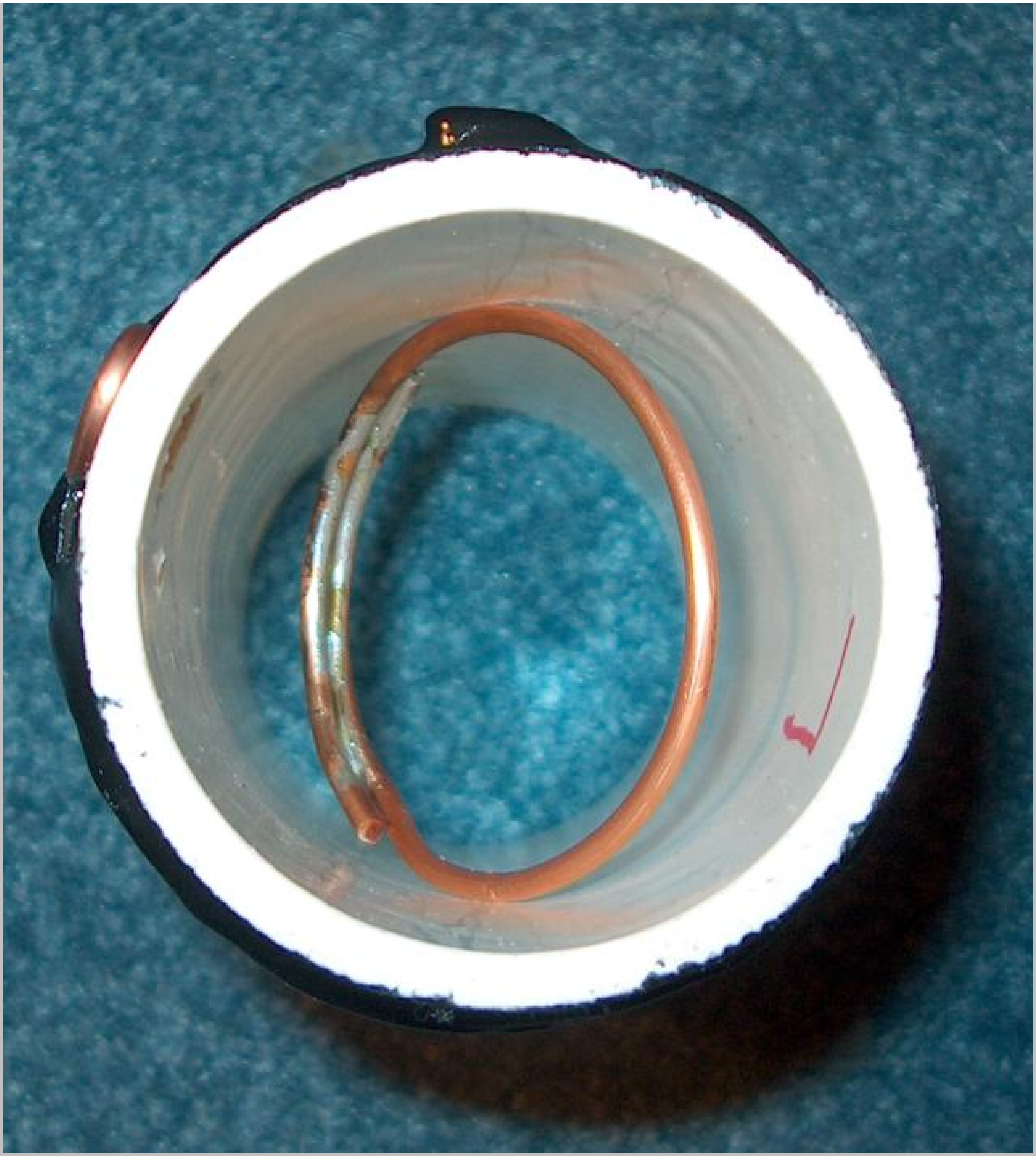

Tuning Ring in place inside the 40-Meter Coil

As it turned out, I didn't need to use it, but I found that I could insert a shorted copper ring inside the coil and adjust the antenna tuning plus or minus a turn or so on each coil. I also tried using various types of ferrite and iron tuning slugs, but found that the copper ring produced less extra loss in the coil than did the ferrite core. When the ring is placed parallel to the turns on the coil, it acts like a shorted turn and reduces the inductance of the coil. The ring fits in place by friction, and after adjusting it, it may be permanently attached with some epoxy. Further minor adjustments may then be made by bending the ring slightly.

ADDITIONAL LOADING COIL DATA

(This loading coil data was added on 24 July 2010)

Since this article went on-line, I have received quite a few questions about the exact construction of the loading coils. Although it is possible for a careful observer to look at the photos posted on this page and deduce the construction of the coils, it is probably a good idea for me to post a more complete description of the construction of the coils so the reader can more easily build them.

In the chart below, all the coils are wound on lengths of Schedule 40 white PVC pipe. The actual end-to-end length of the coil windings is shown in the chart. Due to the thickness of the insulation on the THHN wire, the actual diameter of the finished coil will be close to 2.5 inches. The inductance values are what I measured on my completed loading coils.

FREQUENCY BAND	INDUCTANCE	WIRE GAUGE	NUMBER OF TURNS	COIL LENGTH	COIL DIAMETER	COIL FORM
18 MHz	1.5 uHy	# 10 THHN	5	3.5 Inches	2.5 Inches	2" PVC sch 40 Pipe
14 MHz	3.8 uHy	# 10 THHN	9	3.5 Inches	2.5 Inches	2" PVC sch 40 Pipe
7 MHz	16.5 uHy	# 10 THHN	22	3.5 Inches	2.5 Inches	2" PVC sch 40 Pipe
4 MHz	55 uHy	# 14 THHN	Layer 1 = 34, Layer 2 = 6, overlaps layer 1	3.75 Inches	2.5 Inches	2" PVC sch 40 Pipe
2 MHz	100 uHy	# 14 THHN	Layer 1 = 34, Layer 2 = 20, overlaps layer 1	3.75 Inches	2.5 Inches	2" PVC sch 40 Pipe

Note that the coils for 160 and 75 meters have overlapping coil windings. This can be avoided by using smaller diameter wire or using a longer length coil form. Depending on where you place the top layer of wire on the first layer (near one end or near the center of the first layer of wire) the inductance of the coil will vary somewhat, and you may need to adjust the number of turns on the coil. Tuning on the lower frequency bands will be more critical, so you should expect to do some tuning as needed.

Please note that these exact coils may NOT work for you in your particular situation. Factors such as whip length, height above ground, size of the vehicle, etc., will require tuning the antenna, either by tweaking the number of turns on the coils or adjusting the length of the antenna's top whip slightly. In my case, "close enough" was good enough, because I planned to use an antenna tuner in my mobile station. In any case, the dimensions given in the chart above should get you "in the ball park", as it were.

Tuning the Antenna

A center loaded vertical antenna will not present a pure resistive load at the base of the antenna. Usually, a matching network is added at the bottom of the antenna to cancel the reactance and transform the lower than 50 Ohm feedpoint resistance to something close to 50 Ohms. Since this antenna was designed to operate over several HF bands, a single matching network is impractical. Instead, I chose to connect the antenna through a length of low-loss coaxial cable to an automatic antenna tuner (ATU) inside the vehicle within reach of the operator and out of the weather.

To make the best use of the antenna with this set-up, the loading coils for the antenna should be tuned to resonance at the high end of each band. The antenna will then look electrically "short" to the tuner, which will then be able to tune the antenna to the desired operating frequency. If the loading coil in the antenna is tuned to a frequency below the top of the band, then operation above that critical frequency will cause the antenna to look

ON6MU's "Vipormutant" Antenna
RE-AHFV14P

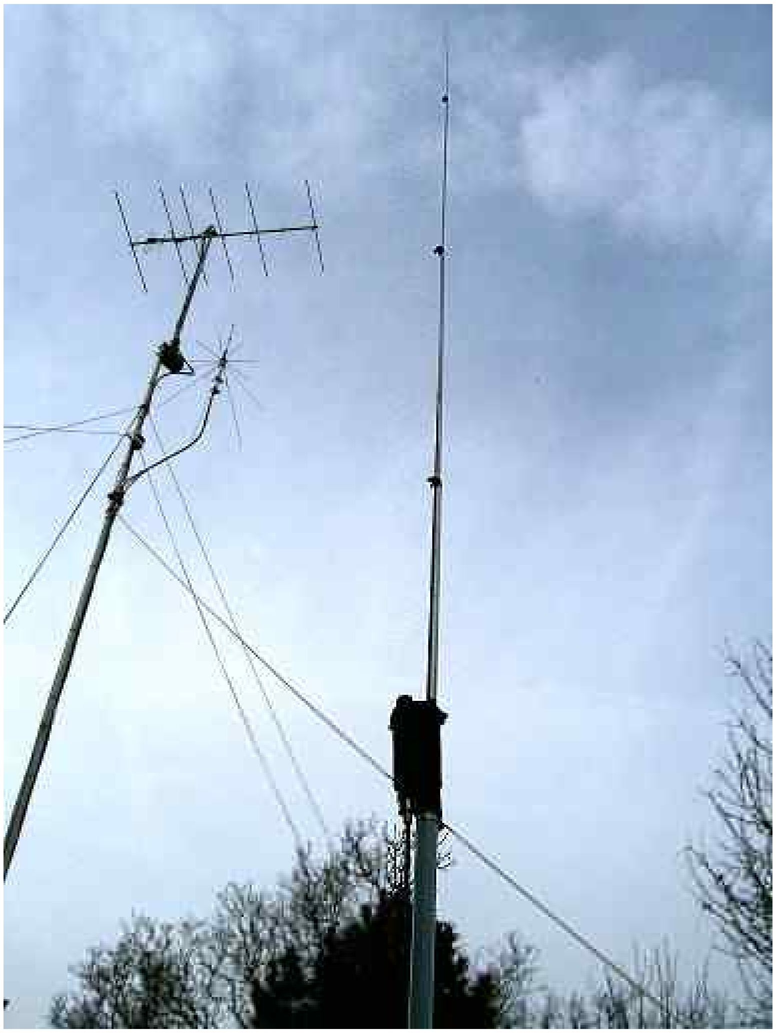

Versatile Inexpensive Portable Multi-band Tunable Antenna

de ON6MU
revision 2

AdChoices RF Antenna Design Antenna 2M 70CM Antenna

Features

- Only 3 meters high fully extended (effective radiating element height)
- Less then one meter inserted (no element is larger then one meter including the tunable section)
- Tunable without the use of an extra tuner (just switch till you get the best SWR)
- Covers all frequencies from UHF to 7 MHz without a tuner
- UHF (tuned by de- or increasing the length of the antenna) 3/4 2*5/8 4*5/8
- VHF (tuned by de- or increasing the length of the antenna) 1/4 5/8 2*5/8
- HF (tuned by switching)
  6 meters (1/4 1/2)
  10 meters (1/4)
  12 meters
  15 meters

  17 meters
  20 meters
  30 meters
  40 meters: with large counterpoise and/or with longer radiating element, or extra tuner
  80 meters: if radiating element is > 5 meters, or with large counterpoise and/or with extra tuner
- Works with or without cointerpoise
- Ideal as portable or balcony antenna
- Compact and extremly portable
- Not too critical on the material or sizes of the elements
- +- 50 watt input
- SWL's Note: tunable on all frequencies between the bands mentioned above

What you need to build the "Vipormutant"

- 5 (or more depending on how high you want your boom) alu tubes
- piece of hard insulating tube (+- 7 cm), examples: plastic, nylon, bamboo...

- some low loss RG174 50 Ohm coax

- carbon/ferrite bead or toroid (to act as a choke)

- a few meters of 0,75mm enamelled copper wire to make the coil

- SO239 (PL259 female)
- Paint, silicon, glue etc. to seal things up
- Plastic box to mount over the coil and where we'll put the switch and SO239 connector
- 12 position rotary switch
- a few innox hose clamps

About the "Vipormutant" antenna:

Well yes, one must have a name HI...It is nothing more then a base coil loaded antenna, but with a selector direct on the base to tune the antenna.
Most of us don't have the luxury of building a 1/4, 1/2 or even a 5/8 wavelength vertical antenna for HF. We have to settle for something a little shorter. (A lot shorter, in the case of people following the FCC's Part 15 rules, which limit them to 3 meters in size.) Shorter vertical antennas can give acceptable (but not spectacular) performance.
I needed a highly (HF) portable antenna to use with my FT-817 which should have the highest possible frequency range (also VHF) and still compact enough to take along almost anywhere! The antenna should be versatile enough to allow further experimenting, to allow being mounted on a balcony, caravan, outdoors etc... So I came up with a compact vertical (dismounted no higher then 1 meter) with a "tuner" directly connected to the antenna radiating element (the best possible place a tuner can be).
The "Vipormutant" tuning principle gets energy into the antenna on a wide range of frequencies, but the design of an antenna is what controls what happens to the RF energy from there. For some antennas, the antenna is simply not complete without a radial system, or at least a counterpoise. Other types of antennas need no RF ground system at all. Most reference books on antennas provide solid guidance on radials and counterpoises, but only for antennas cut to a specific frequency. When using the "Vipormutant" it will also act like tuner and at the best place a tuner should be: directly beneeth the antenna! So the rules have to change somewhat because the "Vipormutant" almost operates across the full range of HF frequencies unto UHF. It doesn't need a counterpoise to work, but the efficiency will increase when you do use it.

Considerations:

- If the length of the conductor is very short compared to a wavelength (< wave/4), the electric and magnetic fields will decrease dramatically within a distance of one or two wavelengths.
- It is impossible to make a small antenna to radiate as efficient like a big antenna.
- Ground losses affect radiation patterns and cause high signal losses for some frequencies. Such losses can be greatly reduced if a good conducting ground is provided in the vicinity of the antenna.

The coil/tuner

Wind 0,8mm enamelled copper wire around the isolator (+- 16mm diamter) and make a tap every xx turns (see fig. 2 and 3)

Fig.1

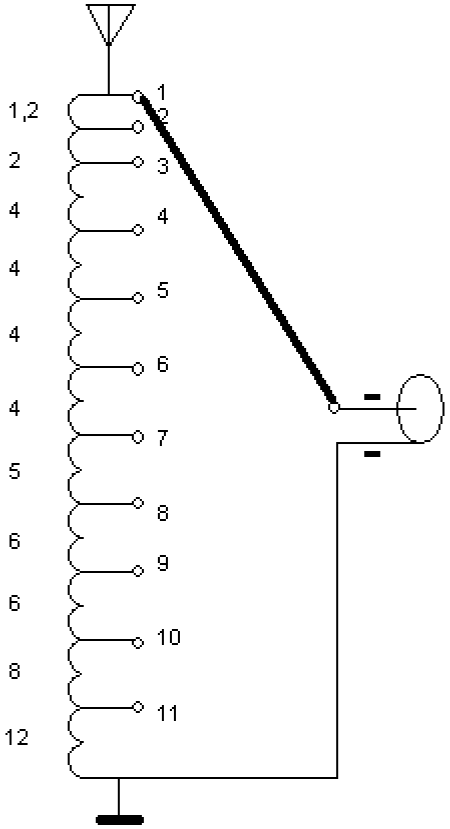

Fig.2

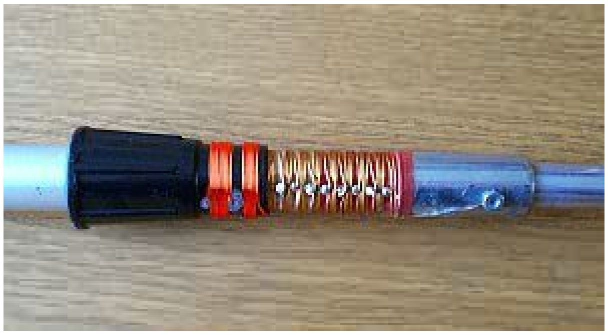

The coil dimensions isn't too critical.

Relatively short antennas behave like lossy capacitors and present a high impedance load to the transmitter due to the large amount of capacitive reactance that is present. The loading coil helps to tune out that reactance. Tuning out the reactance is important because a tuned antenna will accept and radiate much more power than a mismatched antenna.

When the loading coil is installed at the bottom of the vertical radiator, we call it a "base loaded" antenna. Base loading requires the smallest amount of inductance to achieve resonance.

The shoke

Is made out of miniature 50 Ohm coax (rg174) that goes a few turns through the carbon/ferrite bead or toroid. You can also use a Snap-Together Ferrite Choke Core.
If a ferrite is put over a cable which includes both signal and return lines, it will have no effect on the signal (differential-mode) current but it will increase the impedance to common-mode currents. This is because the differential currents, by definition, sum to zero in each wire pair and therefore there is no net magnetic field. If there is no field, the ferrite is invisible. But the common mode currents do produce a net magnetic flux and this flux is concentrated in the bulk of the ferrite, leading to an increased impedance for these currents only. The choke should prevent any mantle currents flowing and should decrease RFI.

The effectiveness can be increased by looping the cable several times through the core, but the benefit is limited at higher frequencies by the stray capacitance between the turns of the cable.

Fig 3.

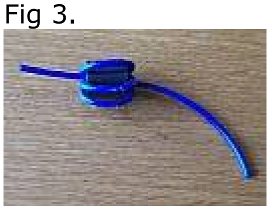

4 turns

Highlighted

Start Download - View PDF

Convert From Doc to PDF, PDF to Doc Simply With The Free Online App! download.fromdoctopdf.com

The base insulator

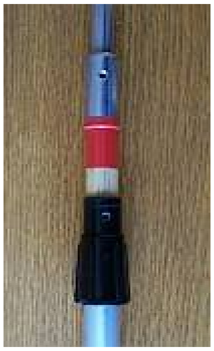

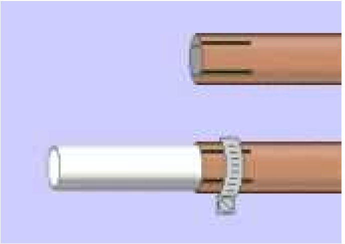

The vertical radiator "driver" element and tuning box

I used a plastic box of 130x70x40mm. On top and bottom I drilled a hole to fit the driver element (radiator of +/- 40cm length)and boom (also +/- 40cm length).

On the side I drilled a hole for the rotary switch and the SO239 connector, whilest on the opposite side I drilled a hole to fit a "ground" bolt where I can easily connect the counterpoise and/or ground to if needed.

Fig 4

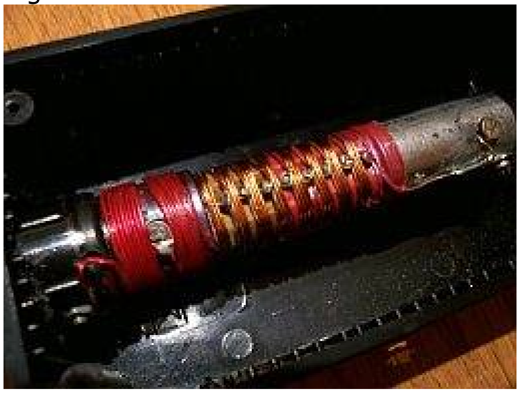

Fig 5

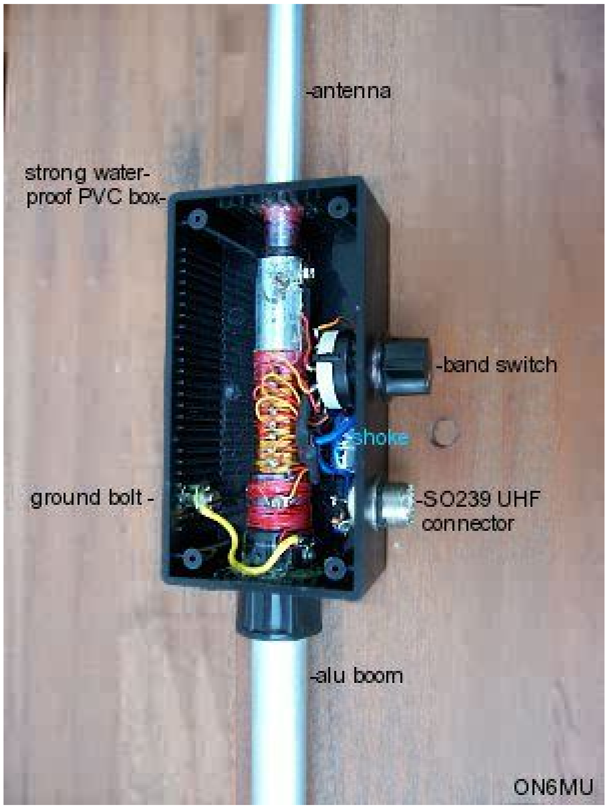
-antenna
strong water-proof PVC box-
-band switch
shoke
ground bolt -
-SO239 UHF connector
-alu boom
ON6MU

This is how things are connected inside the box:

Fig. 6

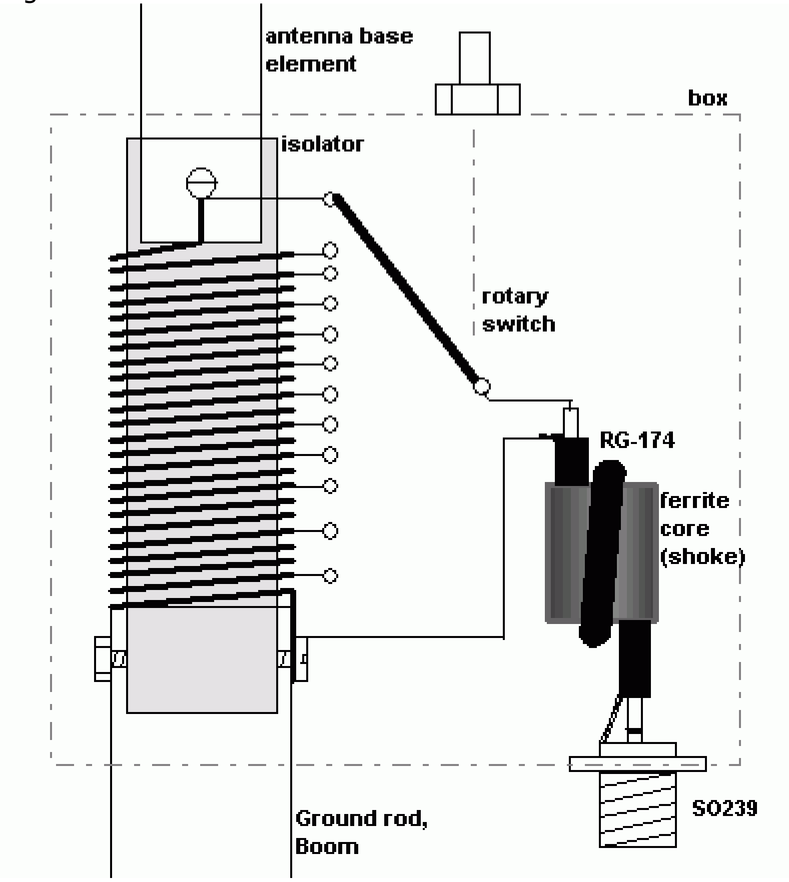
antenna base element
box
isolator
rotary switch
RG-174
ferrite core (shoke)
Ground rod, Boom
SO239

Example

The rotary switch is used for tuning the antenna on each band. The first position allows UHF/VHF ranges.
Tuning is done by sliding in/out of the elements.
The "lower" the switch (higher inductance) the lower the resonance frequency of the antenna.

Remember, and this is important too, to seal everything up so no moisture can penetrate the antenna! Because the radiating vertical antenna elements are made out of separate pieces that fits inside each other and are tightened by hose clamps, the construction isn't waterproof.
If you use a hollow isolating piece you need to to prevent moisture from getting inside the box (via the places where the elements are hold together). I've used a rubber "stopper" that fits snugly on the bottom of the driver element and glued tight.

Black paint finishes the job:

fig. 7

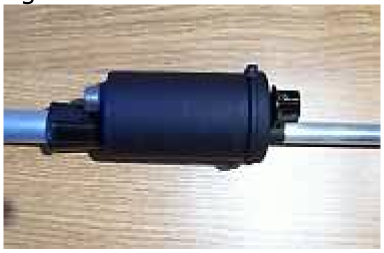

using a round box

In my first design I used a plastic box of 50mm diameter and 9 cm heigh. On top I drilled two holes: one for the driver element (radiator of 40cm) and a hole for the rotary switch (as was used in the first prototype).

fig.8: The main driver element and tuning unit finished

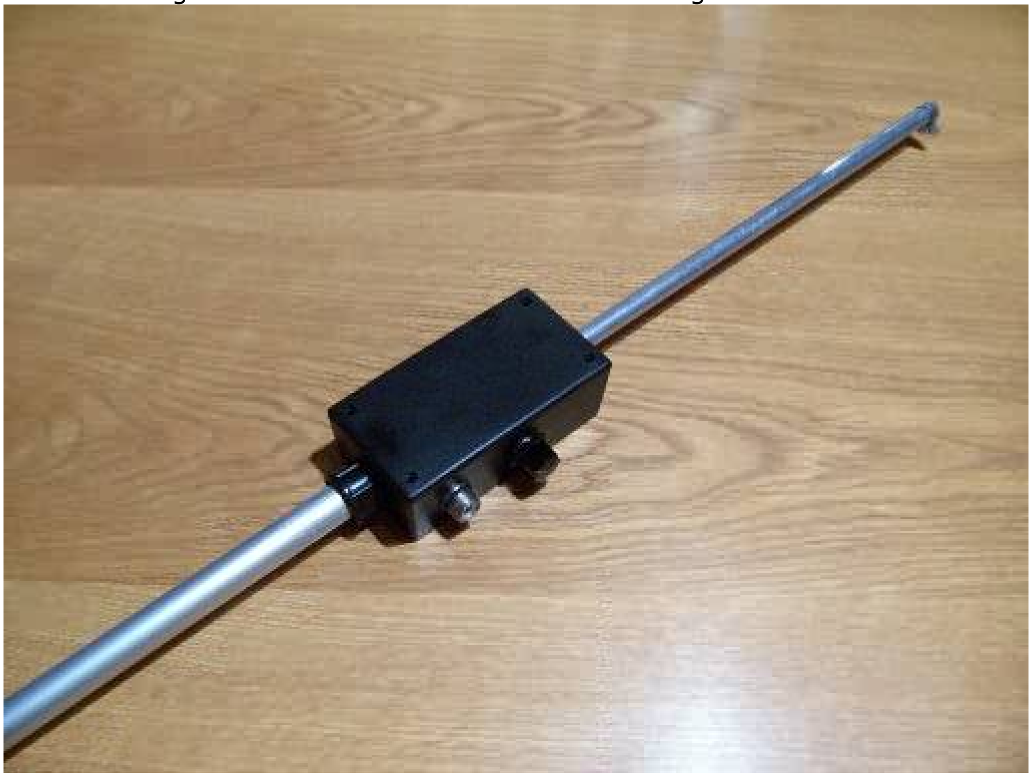

This allows it to be used on almost any boom or can be extended to use with or without vertical elements! Ideal for experimenting!

Featuring Today

The antenna construction specs

All elements are made out of aluminum.
Fig. 9

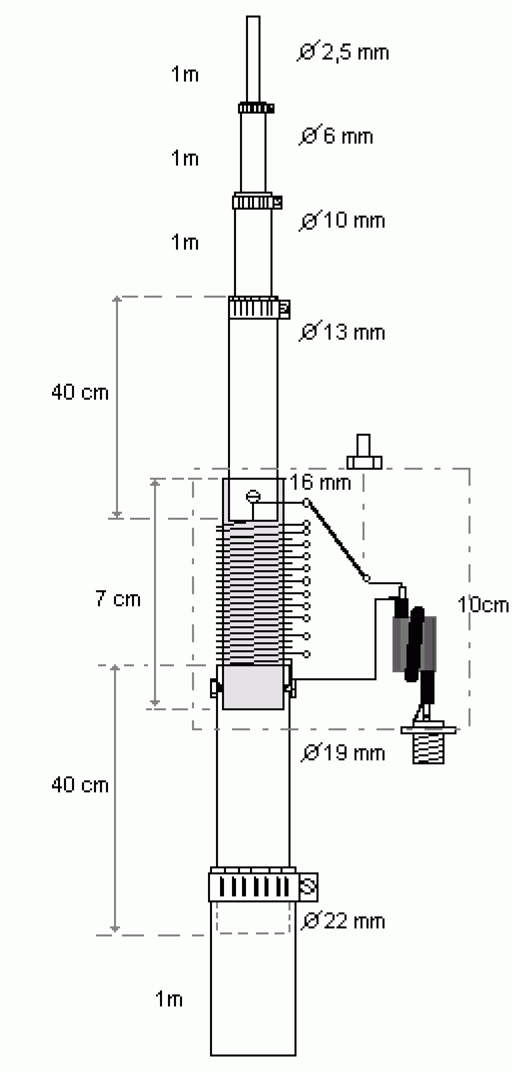
1m
Ø2,5 mm
1m
Ø6 mm
1m
Ø10 mm
40 cm
Ø13 mm
16 mm
7 cm
10cm
Ø19 mm
40 cm
Ø22 mm
1m

This makes the antenna effective radiating elements a total length of 3 meters. The boom elements can be chosen freely and on your needs. A short one (one element of a meter), a medium sized one of several 1 meter tubes or none at all! The bottom piece where the boom is "connected" too is 40 cm and can/could be put directly in the ground (if made pointed for sure). Or you could fix it in a umbrella stand. Use your imagination HI.

Examples of "Vipormutant's" utilization

Used it on a balcony:

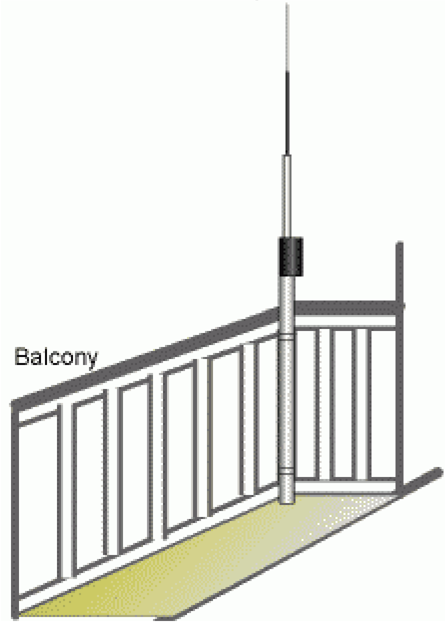

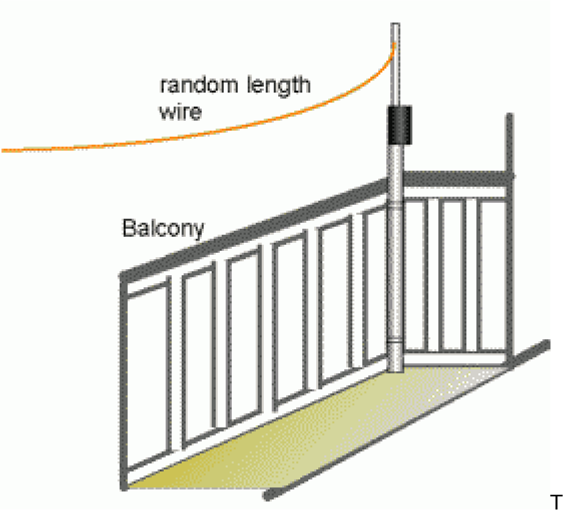
random length
wire
Balcony

This could cover even the lowest HF-bands!

Use it outdoors without grounding:

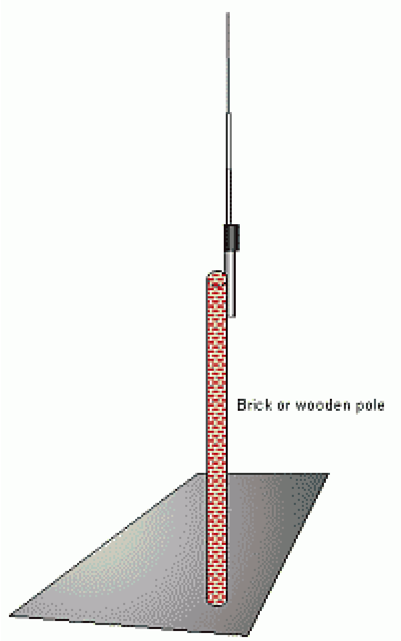
Brick or wooden pole

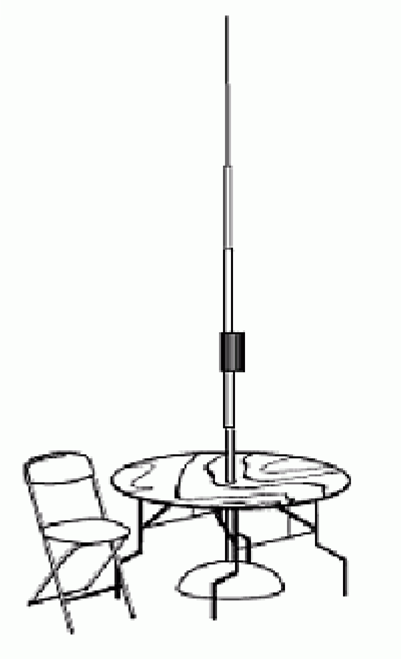

Use it outdoors as a shortened "dipole" balanced antenna (rotary switch set approx. in the middle):

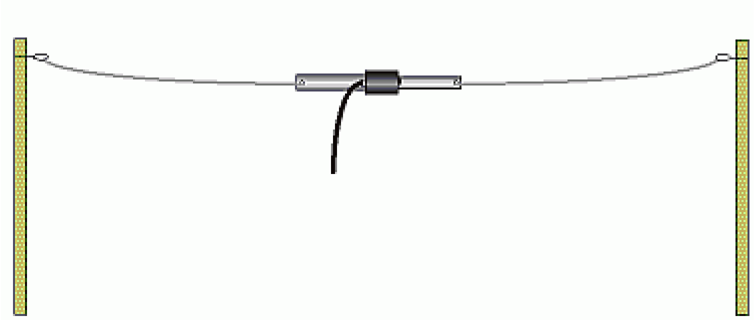

Frequencies below 7 MHz could easily be match 1:1 SWR if total length of the "dipole" > 4 meters
Further tuning can be done with selecting a different impedance using the rotary switch.
Different lengths of wire can be used (an example: one part 3 meter, the other 5 meter or more)

Use it outdoors with ground or counterpoise:

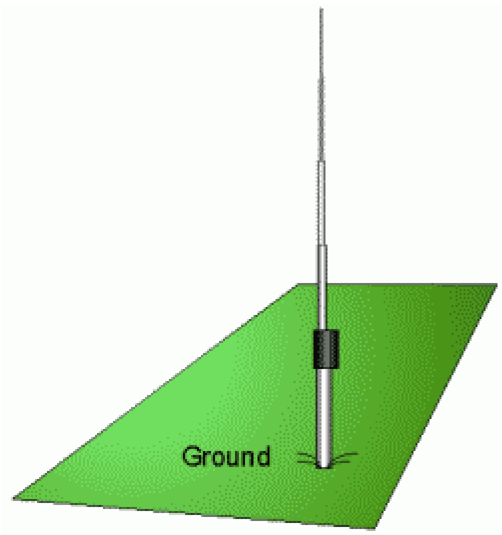

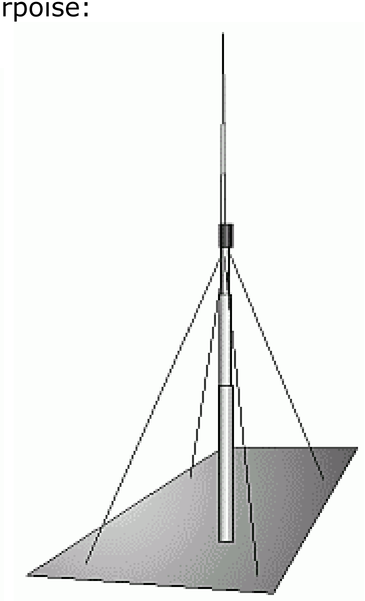

Radials and Counterpoises basic purposes:

1. To improve the RF ground conductivity for the ground current return path. Unless you live in a salt-water swamp, your ground conductivity makes a very poor path for the return of ground currents. This increases the ground losses and reduces the efficiency of an antenna that needs a good RF ground.

2. To provide a counterbalance for the feed point of the antenna to reduce RF radiation back to the radio room. The "Vipormutant" changes the rules because there is no single frequency that you will be operating on, so all of the thumb rules for 1/4 and 1/2 wavelength radials don’t apply. It is possible to be either a purist or a pragmatist in deciding what radials to put in place.

3. Number of radials: More is better, up to a point. In carefully controlled experiments, it has been proved that increasing the number of radials from 2 to 15, or from 4 to 16, produces significant increases in signal strength. Further increasing the number of radials to 60 only produces 1 to 2 dB of increase in field strength. Follow this link to see some of the empirical data.

4. Where to put the radials: For a semi-permanent installation, it is customary to bury the radials a few inches down in the soil. This makes it much easier to mow and walk in the area around the antenna. However, some experimenters have gotten an improvement in performance by raising the radials and the antenna base a few inches above the soil. Raising the antenna and ground system several meters above the earth, for example by installing the base of the antenna on a roof-top, can improve the antenna's performance by reducing capacitive earth losses.

While the "Vipormutant" will provide a good match with a poor RF ground system which will will able you to transmit, your antenna efficiency will be low. Nevertheless, by using a tuning circuit directly at the antenna radiating element losses are kept to a miminum. Getting the greatest efficiency out of your antenna system needs a proper RF ground unless you're using a balanced antenna system

The efficiency of the antenna increases by using a counterpoise. However, the antenna can be tuned perfectly without!

Don't forget to check these out:

ON6MU Homebrew projects
Radioamateur related projects

ON6MU Ham mods
Modifications of transceivers

ALLIED ELECTRONICS & AUTOMATION Allied Doesn't Take Lighting Lightly MW MEAN WELL SHOP NOW

73"

Have fun and my best 73"

Guy, ON6MU
http://www.qsl.net/on6mu

Comments, pictures or experiences with my antenna are always welcome!

ON6MU

9:1 Magnetic Longwire Balun / Unun

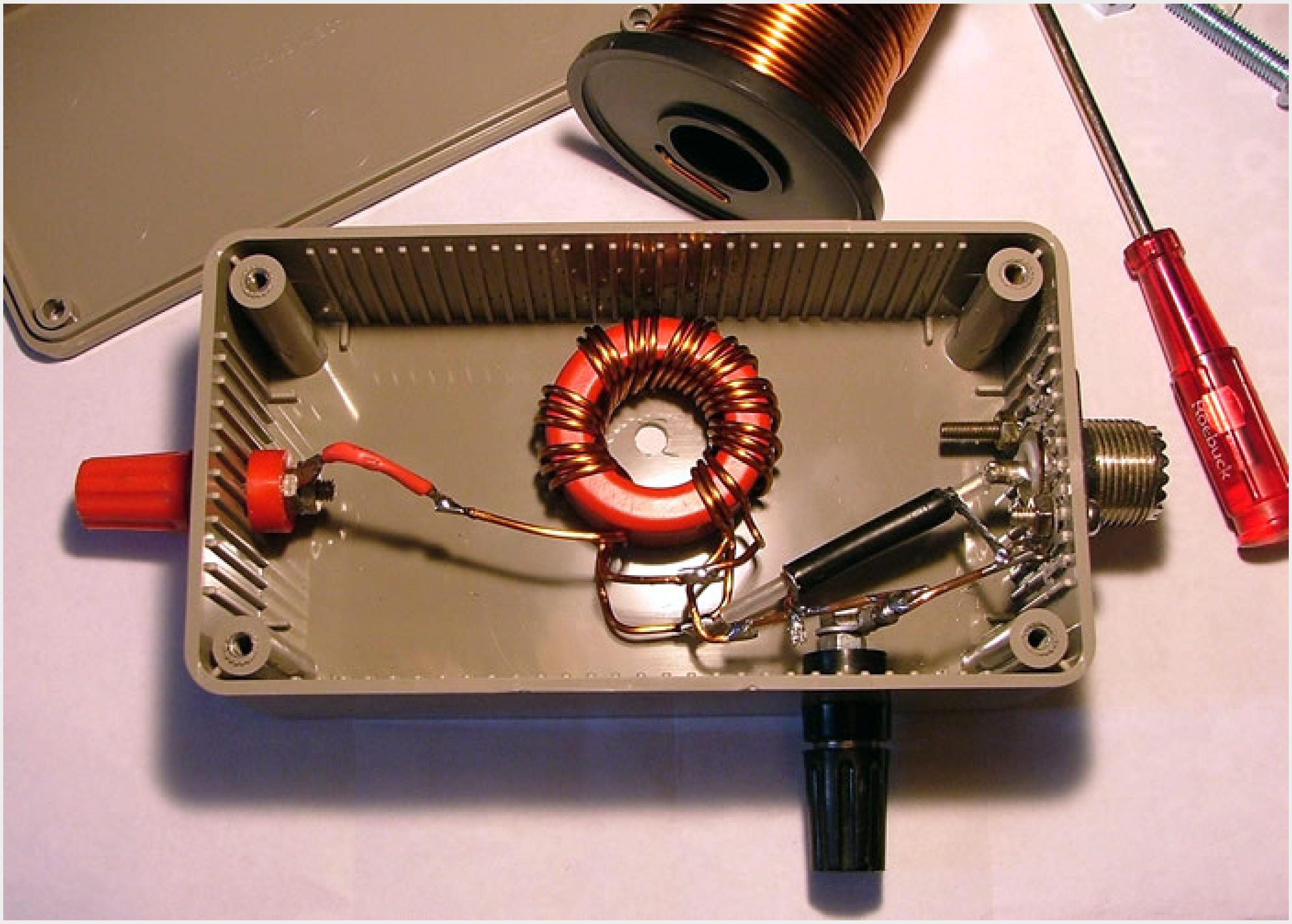

Completed 9:1 UnUn

These are sometimes known as a *Magnetic Longwire Balun*. Its really an impedance transformer (9:1) to feed a high impedance, end fed (unbalanced) random wire which is likely to be a few hundred ohms, and transform it into something closer to a 50Ω (unbalanced) coaxial input, hence UnUn. It can however also be used as a balun by not grounding the other half of the balanced output. The toroid is a T130-2 Iron Powder core, with 3 x 9 turns of 18SWG enammeled copper wire, and the connections can be seen below.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

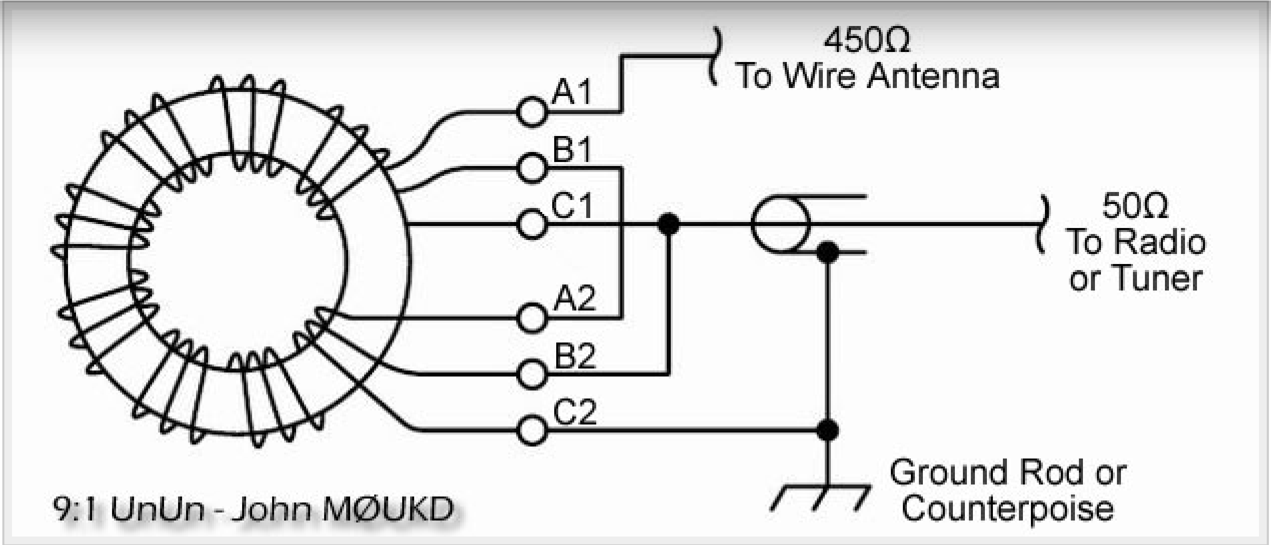
450Ω
To Wire Antenna
A1
B1
C1
A2
B2
C2
50Ω
To Radio
or Tuner
Ground Rod or
Counterpoise
9:1 UnUn - John MØUKD

The triple winding connections. 8 turns is shown here, but more turns is preferred. 10 would be ideal.

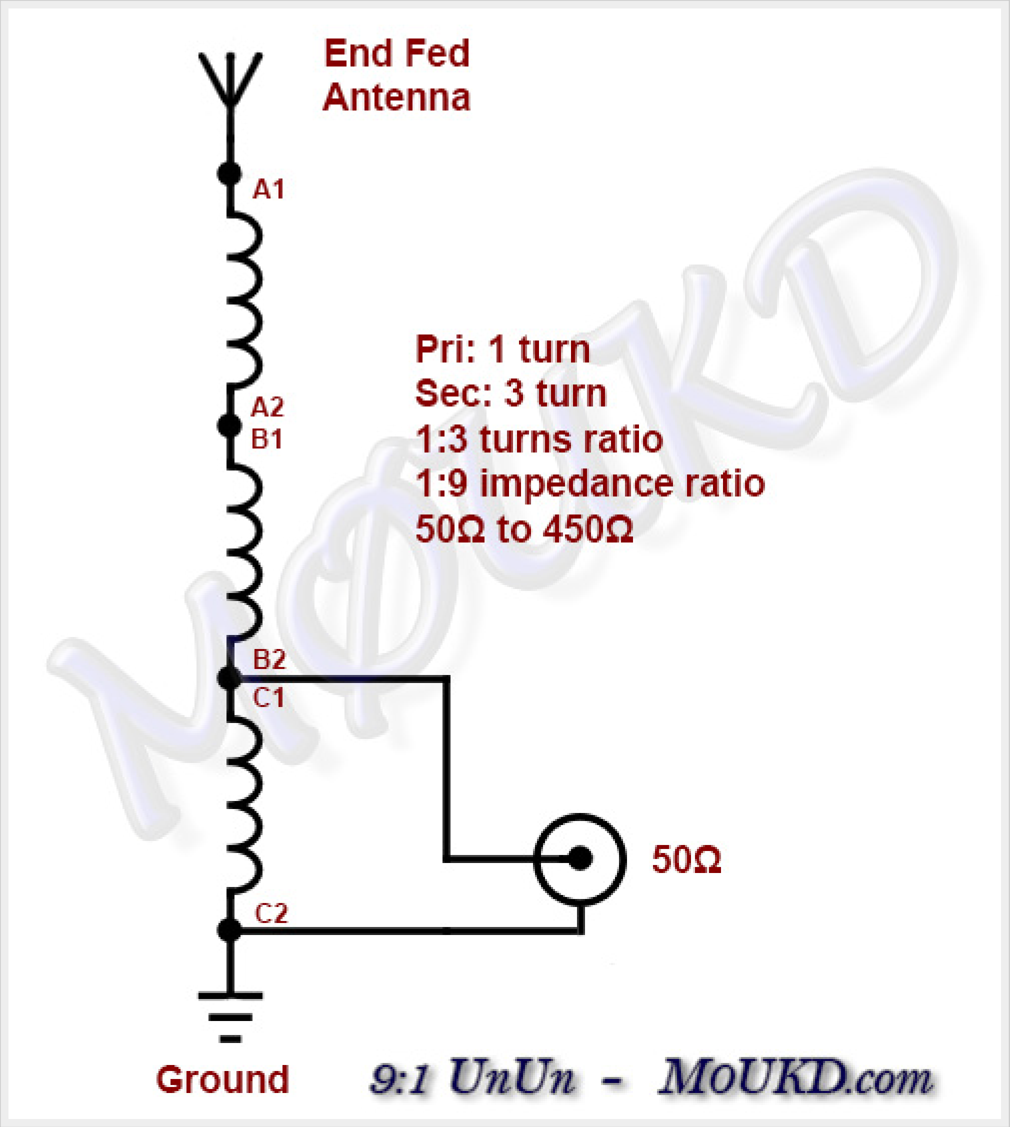
End Fed
Antenna
A1
A2
B1
B2
C1
C2
Pri: 1 turn
Sec: 3 turn
1:3 turns ratio
1:9 impedance ratio
50Ω to 450Ω
50Ω
Ground
9:1 UnUn - MOUKD.com
MØUKD

9:1 UnUn Schematic

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

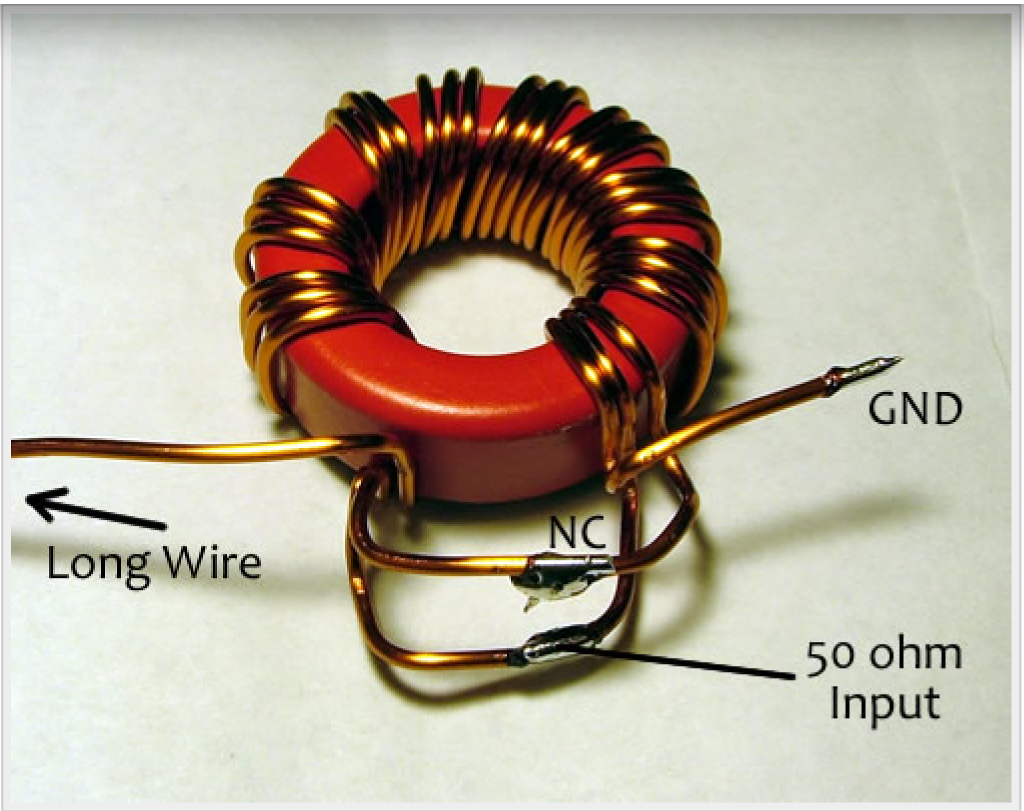
GND
Long Wire
NC
50 ohm
Input

Completed windings

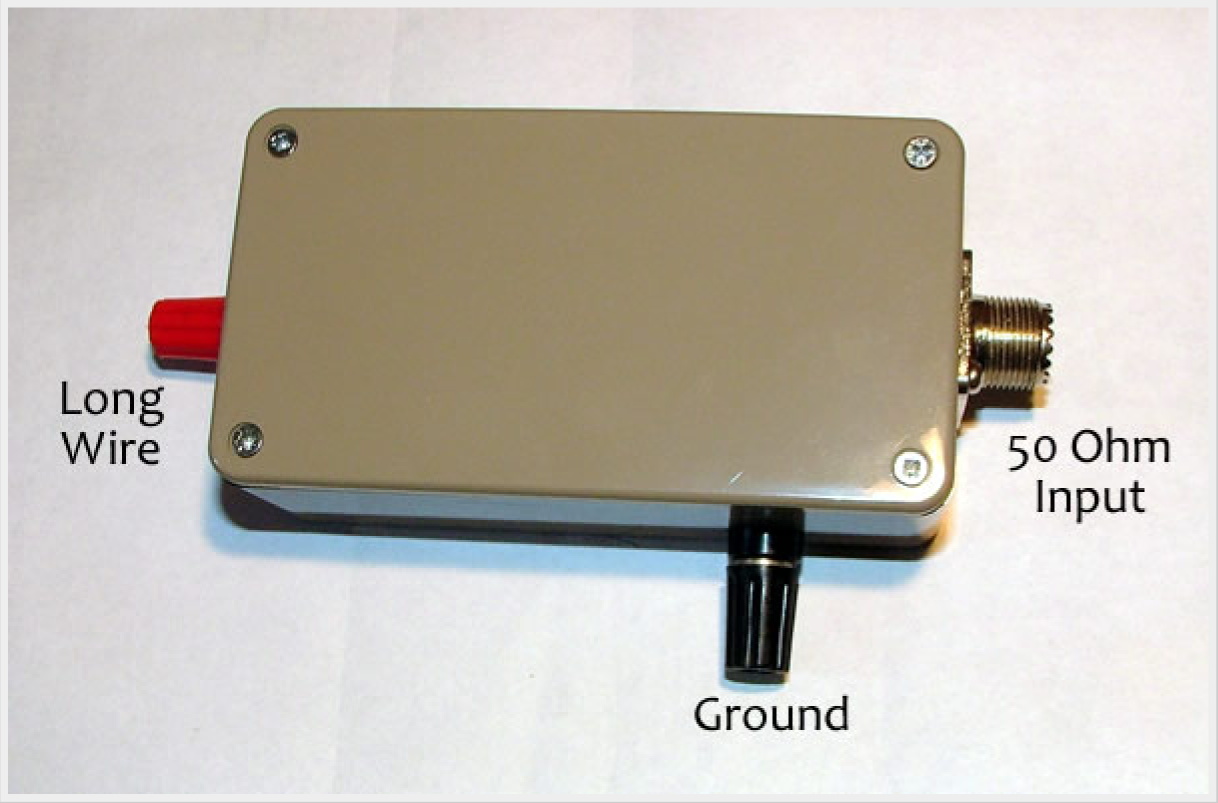
Long
Wire
50 Ohm
Input
Ground

The completed unit.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

Antenna Here is a Dipole

You've probably heard this phrase in many QSOs.
Just what *is* a dipole antenna, and why are they so popular?

By James W. ("Rus") Healy, NJ2L
Assistant Technical Editor

Dipole antennas have been widely used since the early days of radio. Simplicity and effectiveness for a wide range of communications needs are the reasons for this—and they're also the properties that make dipoles worthy of your consideration. There's more to building and installing an effective dipole antenna system than choosing the wire and insulators, as you'll see. Next month, I'll discuss choosing the right feed line for your dipole, and related subjects.

What is a Dipole?

The dipole gets its name from its two halves—one on each side of its center (see Fig 1). (In contrast, a *monopole* has a single element, usually fed against ground as a vertical.) A dipole is a *balanced* antenna, meaning that the "poles" are symmetrical: They're equal lengths and extend in opposite directions from the feed point. In its simplest form, a dipole is an antenna made of wire and fed at its center as shown in Fig 1. (This may look familiar: You may have received a QSL card with a sketched dipole, resembling Fig 1, to denote the antenna the other station used during your contact.)

To be *resonant*, a dipole must be electrically a half wavelength long at the operating frequency. A dipole's resonance occurs at the length at which its impedance has no reactance—only resistance—at a given frequency. As it turns out, that resonant impedance range is compatible with many common coaxial feed lines. Within limits, however, resonance isn't necessary for a dipole to be effective, as I'll explain a bit later. Resonant half-wave dipoles range in size from about 16 feet for the 10-meter band (28-29.7 MHz) to 260 feet for the 160-meter band (1.8-2 MHz). See Table 1.

The lowest frequency at which a dipole is resonant is known as its *fundamental resonance*. A dipole works best at and above its fundamental-resonant frequency. But if a total-length limitation imposed by property boundaries or the spacing of available supports keeps you from doing this, make the antenna as long as you can, even if it's not a half wavelength. Resonances repeat, for half-wavelength-long dipoles, at *odd multiples* of the fundamental-resonant frequencies. For instance, a dipole resonant at 2.5 MHz is also resonant at 7.5 MHz, 12.5 MHz, and so on. These higher-frequency resonances are known as *harmonic resonances*.

As I mentioned earlier, a dipole doesn't have to be resonant to work well. More important than resonance are good construction and efficient power transfer from the transmitter to the antenna. Using an antenna tuner, you can match dipoles that are far longer or shorter than resonant length, but the feed line plays an important role in efficient power transfer, as I'll discuss next month. If you're primarily interested in operating on only one band, a resonant dipole is a good choice. If you're interested in multiband operation with a single antenna, the picture is a bit different. In this situation, it's a good idea to make the antenna resonant at the lowest frequency you plan to use it on (that's where the antenna is longest, because antenna length is proportional to wavelength).

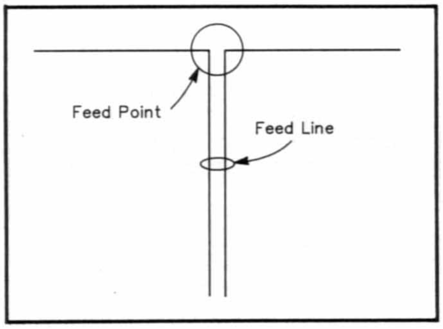

Fig 1—One of the simplest antennas used by hams, the dipole is also one of the most effective, considering the relatively small space it requires. In its simplest form, a dipole is a wire fed at its center.

Table 1

Approximate Lengths of Half-Wave Dipoles for the MF/HF Ham Bands*

Frequency	*Length*
28.4 MHz	16 ft, 6 in.
24.9 MHz	18 ft, 10 in.
21.1 MHz	22 ft, 2 in.
18.1 MHz	25 ft, 10 in.
14.1 MHz	33 ft, 2 in.
10.1 MHz	46 ft, 4 in.
7.1 MHz	65 ft, 11 in.
3.6 MHz	130 ft
1.8 MHz	260 ft

*General equation for half-wave dipole length: $\ell = 468 \div f$, where ℓ is length in feet and f is frequency in megahertz. This equation yields good starting points; you may have to lengthen or trim your antenna to achieve resonance. See the sidebar entitled "Dipole Construction and Adjustment."

Why are Dipole Antennas so Popular?

For almost any kind of MF/HF operation, dipoles are easy to build and install, and they give good results when put up at any reasonable height. "Reasonable heights" are anywhere from a few feet and up, depending on the band. A good general height guideline is half a wavelength or more, but that's impractical for many of us, especially on 40, 80 and 160 meters. At the least, a dipole should clear any surrounding buildings, and other large obstacles, for good performance.

Many hams do quite well with dipole antennas that are electrically low; for instance, an 80-meter dipole strung between two trees at 50 feet (less than a quarter wavelength) allows me to work Europeans regularly on 80-meter CW with a 100-watt transceiver. As a bonus, this antenna works even better at higher frequencies, where it's electrically much farther above ground, as I'll also discuss later.

Variations on the Theme

Part of the beauty of dipole antennas, like many other simple things, is their flexibility. Dipoles can be installed in an infinite number of configurations other than the classical flat-top arrangement (Fig 1). Some of the more common variations include the inverted **V** (sometimes called the drooping dipole—Fig 2A [page 26]); parallel-multiband dipole (Fig 2B); sloping dipole (*sloper*—Fig 2C); vertical dipole (Fig 1, rotated 90°); folded dipole (Fig 2D); and trap dipole (Fig 2E).

Inverted-**V** dipoles are probably more common than flat-top versions. As you might expect, the inverted **V** gets its name from its shape. The main advantages of inverted **V**s are that they need only one high support, and that you can get more total wire into the same horizontal space using

Dipole Construction and Adjustment

Dipoles can be made of almost any kind of wire or tubing, but how well they work and how long they last is determined by the quality of the parts and care you use in building them. With that in mind, here's what it takes to make a dipole antenna.

- *Wire.* Stranded and solid copper, and copper-plated steel (Copperweld), are the most popular types used for antenna construction. Don't use soft, solid copper (such as house wire) for long unsupported runs in resonant antennas; solid wire can stretch enough after installation to detune the antenna. Use reasonably large wire (no. 16 or larger) for permanent antennas. No. 18 or smaller wire is generally okay for low-profile or temporary antennas, or those cut for 14 MHz or higher frequencies. There's little difference between antennas made from insulated and uninsulated wire, although the presence of insulation generally lowers the resonant frequency of a given length of wire by a bit. Insulated wire is heavier and thicker, which, depending on its color, can make it more or less visible than its uninsulated equivalent. Davis RF (see "Where to Get the Pieces") carries uninsulated, stranded no. 14 copper that's made of 168 strands of wire. This is good stuff for dipoles (or home-brew balanced feed lines) because it's very flexible and easy to work with, doesn't stretch, and resists breakage well.

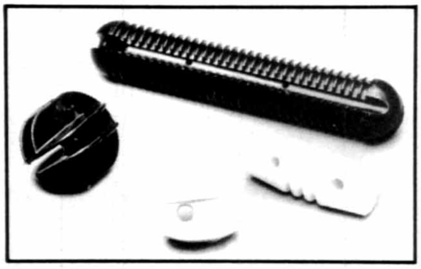

Fig A—Several varieties of insulators are commonly available to hams. The lower-center ceramic unit and the plastic insulator on the left are current Radio Shack products. The others, and close variants, are available from many mail-order sources (see "Where to Get the Pieces").

- *Insulators.* A dipole should have insulators (Fig A) at its center (feed point) and at each end, where ropes (or other restraints) support the antenna. The center insulator serves two primary purposes: (1) it separates and supports the feed-line conductors, making for strong mechanical wire joints that minimize stress on solder joints; and (2) it gives you a convenient place from which to support the dipole with a rope, if your dipole is center-supported. End insulators serve another purpose: They provide an isolated support point at the high-voltage ends of the antenna wire. (Potentials at the ends of a resonant dipole can be in the kilovolts—even with just a 100-watt transmitter!)

Plastic, ceramic or porcelain insulators are fine for most applications. Home-made varnished hardwood is also acceptable. There's no need to use a special center insulator at a dipole's feed point—the impedance, and therefore the RF voltage, there is low—although doing so can make construction easier.

- *Rope.* Nylon, Dacron and similar materials are best. Avoid polypropylene; it rapidly deteriorates in sunlight. Also avoid hemp and other natural materials. They tend to stay damp and rot in the center after getting rained on, with no outwardly visible signs of damage. If you select nylon rope, use the braided variety and carefully seal its cut ends by melting them with a match or cigarette lighter. This rope is very strong, but unlike Dacron, it also initially stretches as much as a few percent.

To make a dipole that'll stay up and that's easy to work on, using the right knots is important. Learn (or relearn) to tie a *bowline* (an unsliding loop); a *clove hitch* (for attaching rope to poles, tower legs, and such); a *taut-line hitch* (a sliding knot for adjusting the tension in a rope); and a *sheet bend* (for joining the ends of two ropes). These knots are simple to tie, and are described in every edition of the Boy Scout *Handbook* and *Fieldbook*, as well as some merit badge booklets (look under Boy Scouts of America in the Yellow Pages for local suppliers of BSA materials). Another good reason to be acquainted with these knots: When (if) you start working on towers (or other supports) and bigger antennas, these knots can save your *life*, not to mention your hardware.

- *Soldering.* To make permanent

this configuration. This is often an important advantage on the lower-frequency bands, where real estate and support height suitable for putting up a full-size dipole are at a premium. Inverted **V**s usually work almost as well as horizontal flat-top dipoles when the dipole's height is the same as the feed-point height of an inverted **V**. It's important to keep the antenna ends high enough above the ground so that people, vehicles and such can't come into accidental contact with them.

Another common dipole configuration is the multiband parallel version. In such an antenna (Fig 2B), multiple dipole elements are fed at the same point, with a single feed line, and supported by spacers attached to the longest dipole element. The main advantage of parallel dipoles is multiband coverage with resonant elements on each band, allowing the use of a single coaxial feed line for several bands without

Where to Get the Pieces

Here's a brief list of suppliers of wire, insulators, feed line, connectors, and other items you'll need for making and installing dipole antennas. For a more complete list, see the advertisements in this and other issues of *QST*.

- Radio Shack stores (antenna wire, coaxial cable, ceramic and plastic insulators, connector-sealant tape and other supplies). See 1991 catalog, page 137.
- Davis RF, PO Box 230, Carlisle, MA 01740, tel 508-369-1738 or 800-484-4002, extension 1356 (high-flexibility antenna wire, coaxial cable, several varieties of prefabricated balanced line [and parts to make your own], ceramic insulators, aluminum tubing and other supplies). Catalog available.
- Ocean State Electronics, PO Box 1458, Westerly, RI 02891, tel 401-596-3080, fax 401-596-3590 (antenna wire, coaxial cable and connectors, 300-Ω twinlead, insulators). Catalog available.
- Certified Communications, Rte 2, Pittman Rd, Landrum, SC 29356, tel 803-895-4195 (coaxial cable, balanced feed line).
- The Radio Works, PO Box 6159, Portsmouth, VA 23703, tel 804-484-0140, fax 804-483-1873 (antenna wire, coaxial cable and connectors, balanced feed line, baluns, insulators, rope, sealant and other supplies). Catalog available.
- Most major Amateur Radio equipment dealers (Amateur Electronic Supply, Ham Radio Outlet, and other *QST* advertisers) carry parts for dipole construction.—*NJ2L*

connections that will stand up to wind and weather, you'll need a soldering device that can quickly heat the wires to be joined. A 30-W iron probably won't do an acceptable job with wires larger than no. 16; a 100-W or larger soldering gun is best. If you're using relatively heavy wire and ceramic or porcelain insulators, you can use a propane torch (preferably with a soldering tip) for soldering. (If you do this, heat the joint first, then remove the flame and let the heat of the wires melt the solder into the hot junction.) Use 60/40 (or similar) rosin-core solder (available from Radio Shack and most hardware stores). Use caution when soldering close to plastic insulators; too much heat will damage them. Don't breathe solder or insulation fumes!

- *Tape, sealant and similar protective coverings.* Protect your solder joints and feed-line connections from the weather after soldering by tightly wrapping them with high-quality electrical tape (such as Scotch 33 or 88) or silicone adhesive tape, or covering them with Coax-Seal (or equivalent). Solder joints at end insulators can be sufficiently protected by spraying them with clear lacquer, available from most hardware and home-supply stores. Weatherproofing properly soldered end-insulator joints is optional, but *seal the center insulator* as if you were planning to use it underwater. The last thing you want is rainwater in your coax!

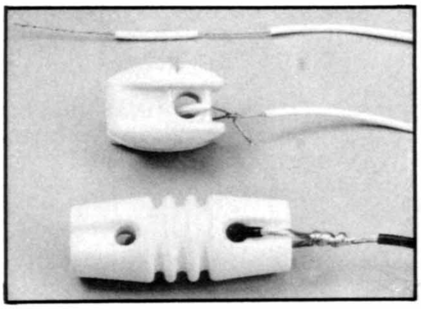

Fig B—Secure end-insulator connections are important to long antenna life. Top to bottom: Leave part of the insulation from the wire end to protect the wire where it goes through the insulator eye; wind several turns of wire back on itself to make a strong mechanical connection; solder the joint; and spray it with clear lacquer to prevent corrosion.

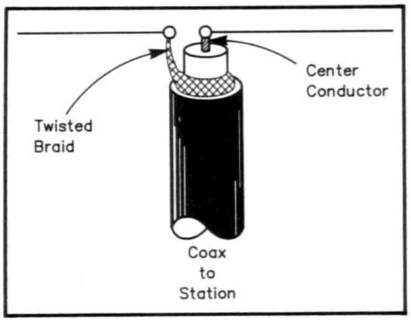

Fig C—Coaxial cable feeding a dipole antenna. The center insulator is omitted for clarity. When measuring and adjusting a dipole fed with coax, keep in mind that the stripped coax center-conductor and shield lengths *add to the antenna length*.

In dipole construction, it's important to make strong mechanical junctions in the wires. An antenna that depends solely on solder joints to handle wind stresses will surely fall down sooner than one that's made with good mechanical connections. RF interference can also result from deteriorated solder joints. Fig B shows how to make a solid mechanical connection at an insulator.

Bringing a Dipole "On Frequency"

To resonate a dipole that's fed with coaxial cable, select a band and cut the wire a couple of feet longer than the appropriate length shown in Table 1. Then, install center and end insulators and attach the coaxial feed line (Fig C shows how to attach coaxial feed line to a dipole). Support the antenna a few feet above ground, measure the SWR and trim the antenna (equally from each end) to raise the resonance to the desired frequency. If the antenna is too short, splice in some additional wire, or attach equal amounts at the end insulators and let it hang from them. Copper-and-brass split-bolt clamps, made for attaching copper ground wires together and available from most home-improvement stores, are great for making these additions.

Selecting supports and installing antennas is the subject of many other articles. See the references listed under "Further Reading" for a few of these.—*NJ2L*

Further Reading

For more information on antenna construction and installation, see the references listed below. The issues of *QST* cited here may be available at your local library; if not, contact the Technical Department Secretary at ARRL Headquarters (see p 3 of this issue) for any photocopies you need. (There is a nominal charge for this service.)

- G. Hall, Ed., *The ARRL Antenna Book*, 15th ed (Newington: ARRL, 1987). If you're aware of this book, you may think that it's a highly technical reference intended for those with lots of antenna knowledge and experience. In truth, it's intended for hams at all experience and license levels, and even those who aren't yet licensed. All of the subjects I've covered in this article are treated from basic to detailed levels in *The ARRL Antenna Book*, and it should be the key reference in your quest for antenna wisdom.
- C. Hutchinson and L. Wolfgang, Eds, *The ARRL Handbook for Radio Amateurs*, 1991 ed (Newington: ARRL, 1990). Chapter 17, "Antenna Fundamentals," provides a good technical introduction to dipole theory and feeding.
- D. Brede, "The Care and Feeding of an Amateur's Favorite Antenna Support—The Tree," *QST*, Sep 1989, pp 26-28, 40.
- W. Calvert, "The EZY Launcher," elsewhere in this issue.—*NJ2L*

the need for an antenna tuner. An inherent disadvantage of parallel dipoles, however, is narrower bandwidth than single dipoles provide.

Two other fairly popular dipole variations are the trap dipole and the folded dipole. Traps are tuned circuits (consisting of inductance and capacitance) that electrically isolate the inner and outer sections of the antenna at certain frequencies, providing multiband resonant coverage from a single antenna. At a trap's resonant frequency, it presents a high impedance and therefore isolates the outer segments of the dipole, making the antenna electrically shorter than it is physically. At frequencies below the trap's resonance, it has a low impedance, which makes it transparent to RF (ie, it doesn't isolate any part of the antenna).. Traps aren't used only in dipoles:

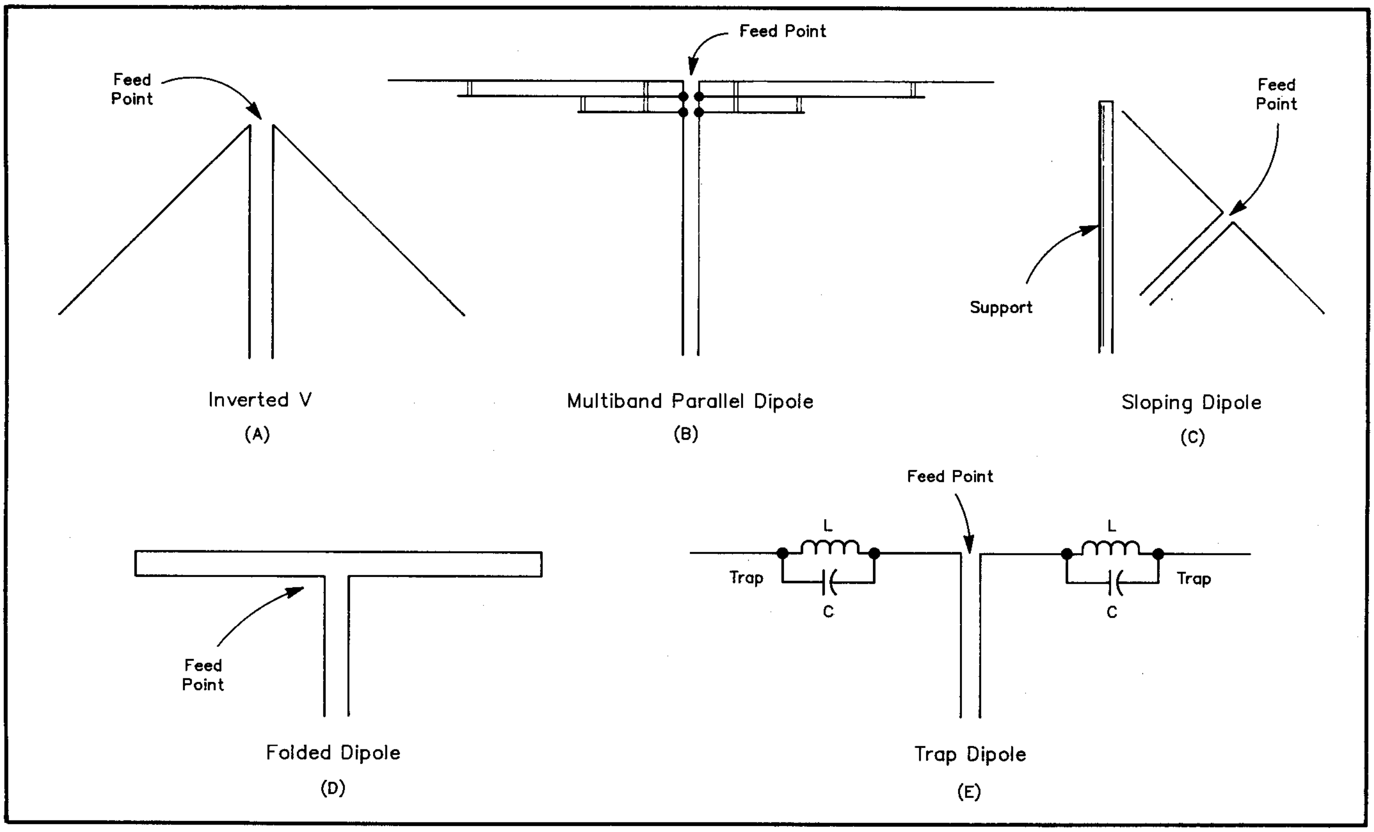

Fig 2—Variations on the dipole are numerous: at A, an inverted V; at B, a multiband parallel dipole; at C, a sloping dipole (sloper); at D, a folded dipole; and at E, at trap dipole. Dipoles of the multiband parallel, trap and folded varieties can be installed in sloping or inverted-V configurations.

Trap Yagi beams and verticals are also popular. Folded dipoles are a bit less common in Amateur Radio use; they use full-length parallel wires shorted at the ends, and have feed-point impedances that provide good matches to balanced feed lines. FM-broadcast receivers usually use folded dipoles made from TV twinlead. *The ARRL Antenna Book* and *The ARRL Handbook* cover trap and parallel dipoles in more detail (see "Further Reading").

The Dual-Band Dipole

Two popular ham bands, especially for Novice and Technician Class operators, are those at 7 and 21 MHz. As mentioned earlier, dipoles have harmonic resonances at odd multiples of their fundamental resonances. Because 21 MHz is the third harmonic of 7 MHz, 7-MHz dipoles are harmonically resonant in the popular ham band at 21 MHz. This is attractive because it allows you to install a 40-meter dipole, feed it with coax, and use it without an antenna tuner on both 40 and 15 meters.

But there's a catch: The third harmonic of the Novice 40-meter allocation (7100-7150 kHz) begins at 21,300 kHz; yet the Novice segment of 15 meters is 21,100-21,200 kHz. As a result of this and other effects, a 40-meter dipole does not provide a low SWR in the 40- *and* 15-meter Novice segments without a tuner.

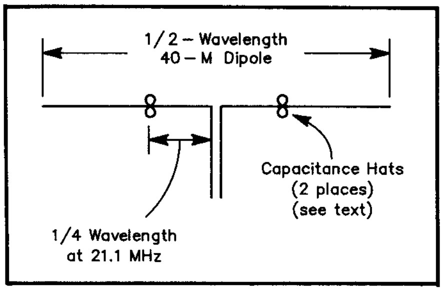

Fig 3—Figure-8-shaped capacitance hats, made and placed as described in the text, can make your 40-meter dipole resonate anywhere you like in the 15-meter band.

An easy fix for this, as shown in Fig 3, is to *capacitively load* the antenna about a quarter wavelength (at 21.1 MHz) away from the feed point in both wires. Known as *capacitance hats*, the simple loading wires shown lower the antenna's resonant frequency on 15 meters without substantially affecting resonance on 40 meters.

To put this scheme to use, first measure, cut and adjust the dipole to resonance at the desired 40-meter frequency, as described in the sidebar called "Dipole Construction and Adjustment." Then, cut two 2-foot-long pieces of stiff wire (such as no. 12 or no. 14 house wire) and solder the ends of each one together to form two loops. Twist the loops in the middle to form figure-8s, and strip and solder the wires where they cross. Install these capacitance hats on the dipole by stripping the antenna wire (if necessary) and soldering the hats to the dipole about a third of the way out from the feed point (placement isn't critical) on each wire. To resonate the antenna on 15 meters, adjust the loop shapes *(not while you're transmitting!)* until the SWR is acceptable in the desired segment of the 15-meter band. You can make all these adjustments with the dipole just a few feet off the ground; raising the antenna to its permanent height shouldn't shift the SWR much. Recheck the antenna's 40-meter resonance before raising it, though.

Feed-Line Considerations

The antenna wire and insulators, how you put them together and where you string them up is only part of a dipole-based antenna *system*. Next month, I'll cover selecting and using a feed line for your dipole(s). In the meantime, I suggest that you have a look at the two parts of Dave (WJ1Z) Newkirk's article, "Connectors for (Almost) All Occasions," in the April and May issues.

1:1 Voltage Balun for HF wire dipoles

Heres a neat 1:1 50 ohm balun for use on HF horizontal wire dipoles. It uses an AM radio ferrite rod, with 3×14 turns of wire. 10-14 turns should be good for 2-30MHz. I used 18SWG enammeled copper wire. It all fits in a small project box. Great for setting up portable. For 20 metres, I started with 6 metres of wire per element, and trimmed each down until it has minimum S.W.R. I managed to get a 1:1.1 S.W.R. on my desired frequency. Bandwidth from the resonant frequency is about 200KHz either way, any more needs an ATU. The wire and ferrite core I had already, the box, terminal plugs and SO-239 came to about £6. The balun can also be built on a ferrite toroid such as an FT140-43 or FT240-43.

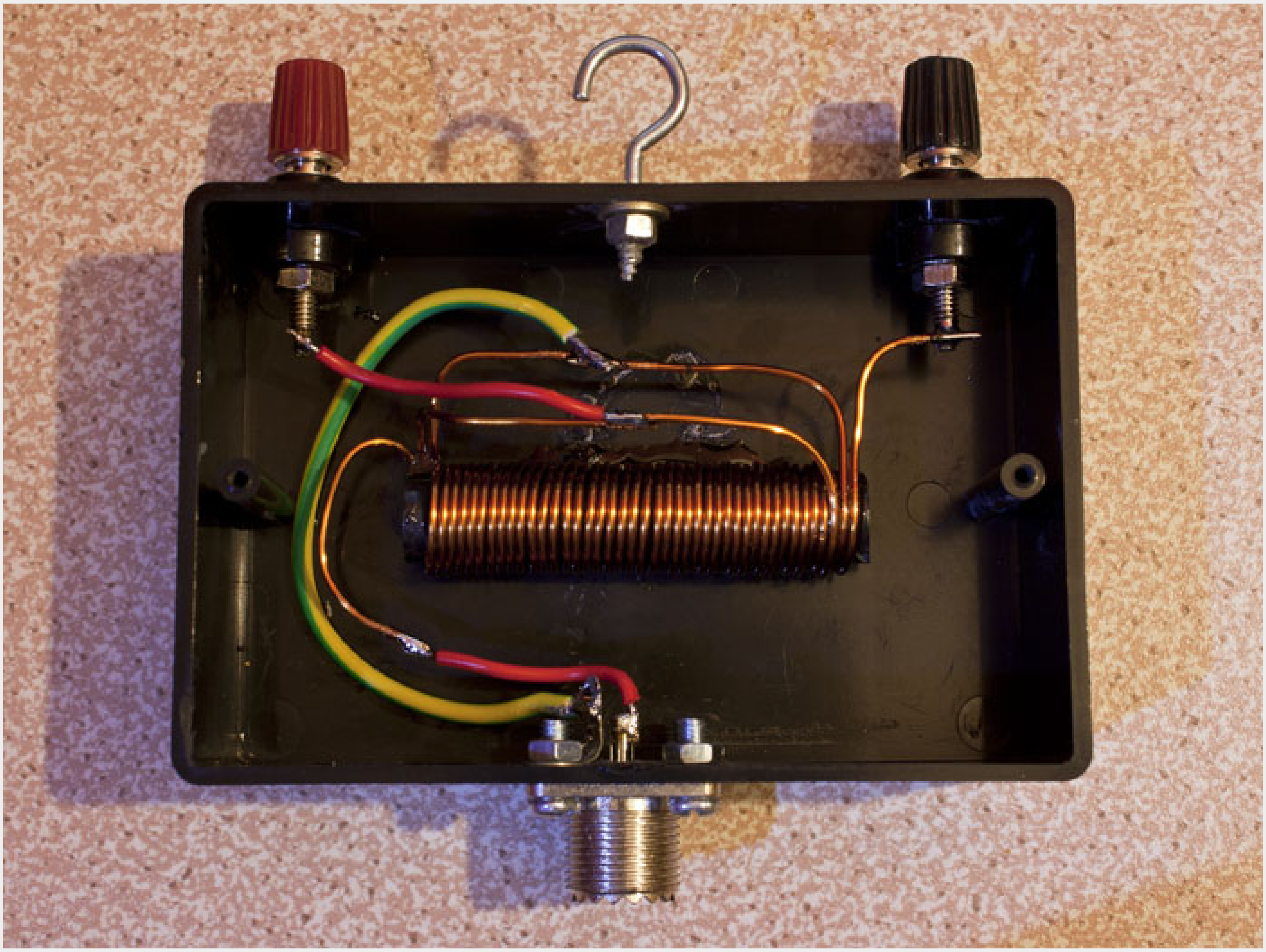

1:1 ferrite rod voltage balun

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Accept Read More

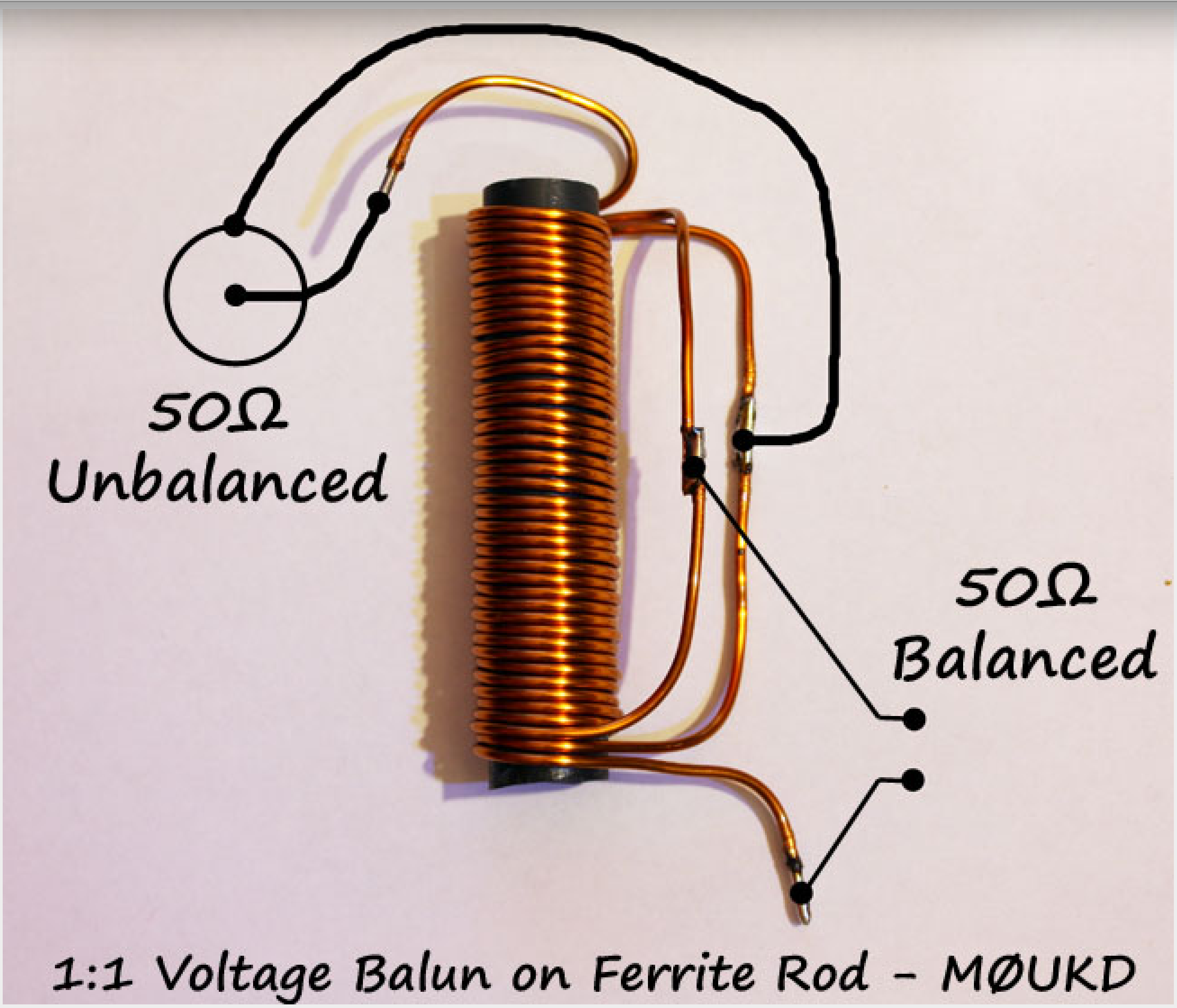
50Ω
Unbalanced
50Ω
Balanced
1:1 Voltage Balun on Ferrite Rod – M0UKD

Bandwidth may be increased by more spacing between each trifilar turn. I didn't have room on this tiny ferrite rod, but it performs OK.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

1:1 Voltage balun ("Ruthroff")

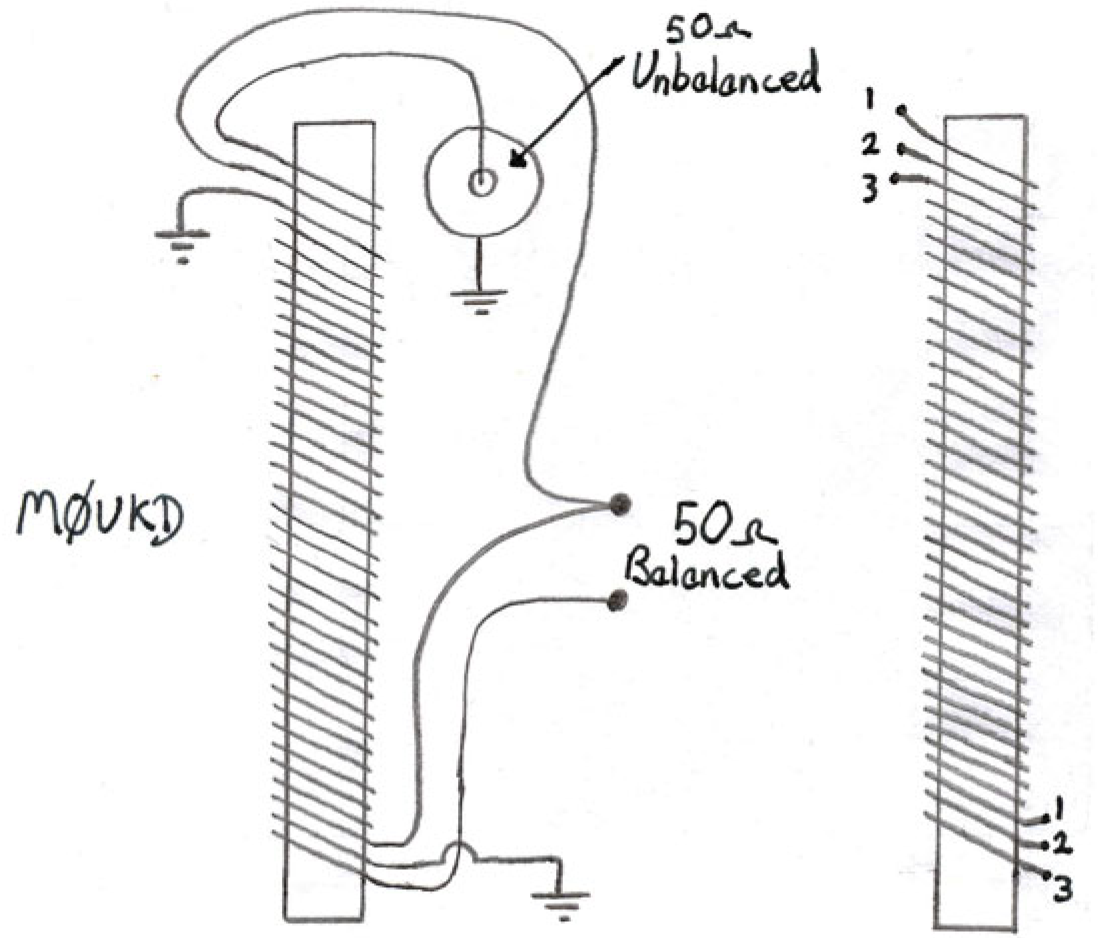
50Ω
Unbalanced
1
2
3
M0UKD
50Ω
Balanced
1
2
3

10 - 15 trifilar turns on ferrite rod.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

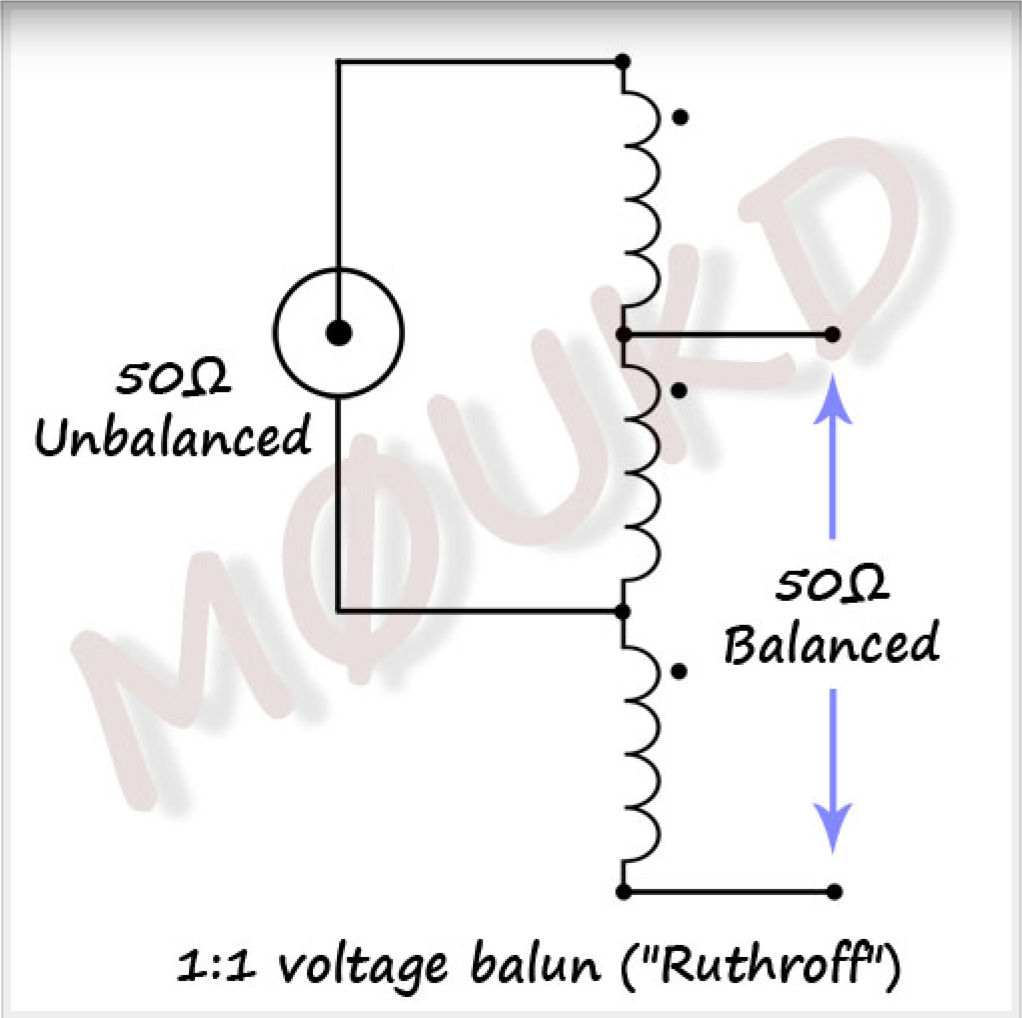
50Ω
Unbalanced
50Ω
Balanced
1:1 voltage balun ("Ruthroff")
M0UKD

Schematic Diagram

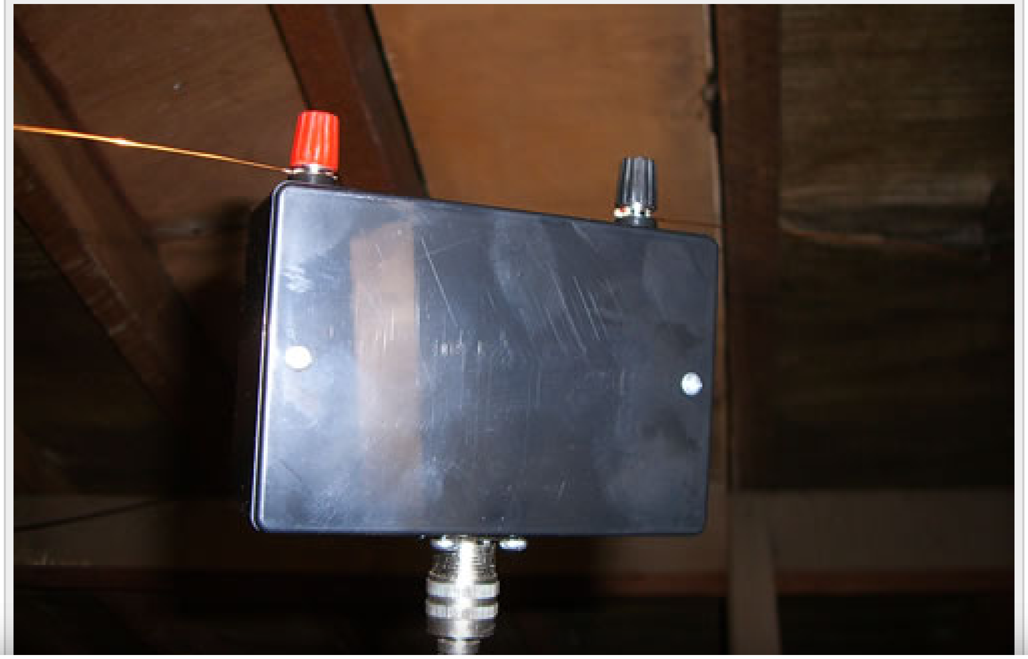

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

End Fed Half Wave Antenna Coupler (EFHW)

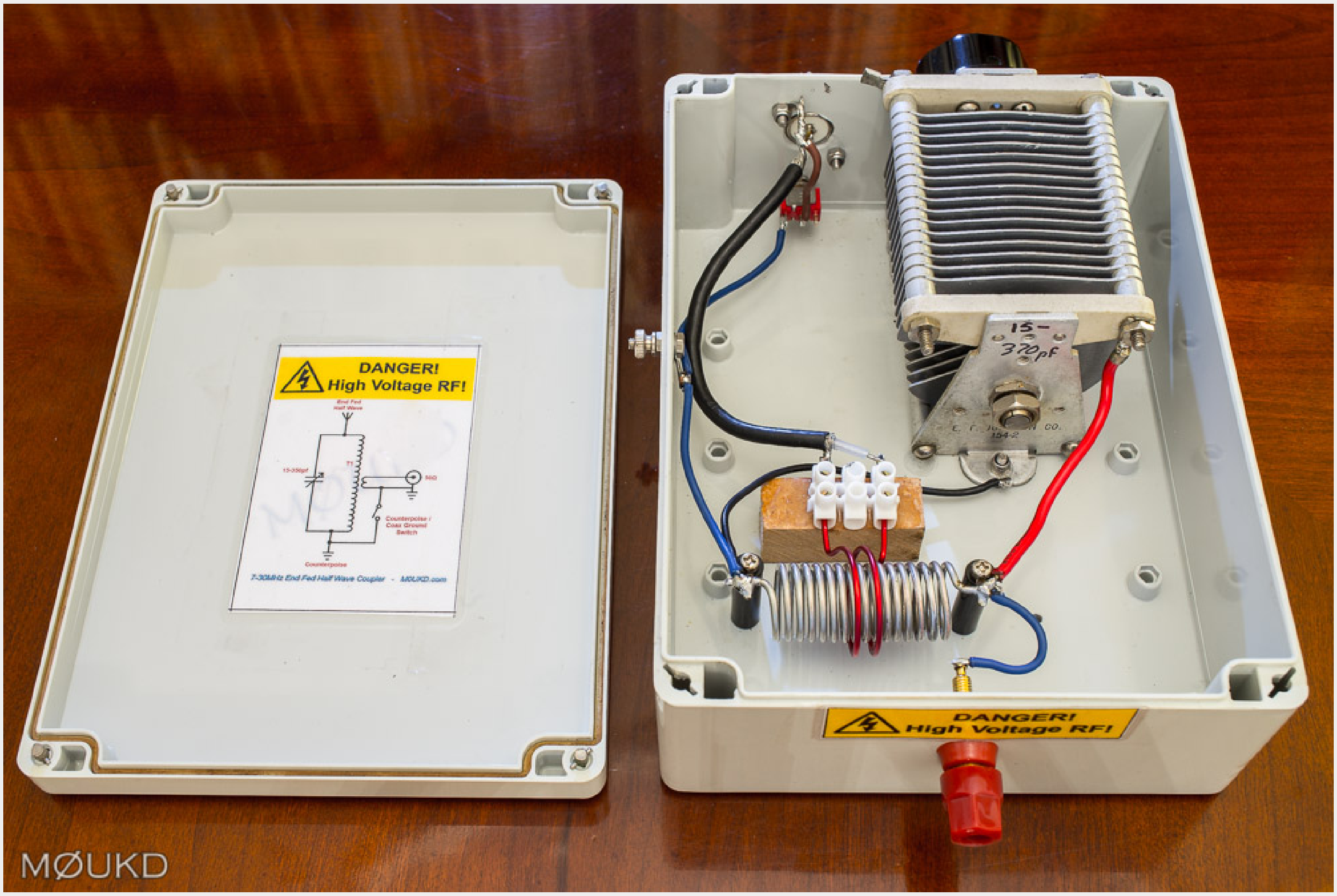

The finished End Fed Half Wave antenna coupler.

Centre fed half wave dipoles make great, simple and effective antennas for the HF bands. Sometimes however, the centre feed is not ideal, for example when you want to use it as a vertical. Being able to feed the dipole from one end gives you more options on how to erect an antenna and makes portable operation easier. A vertical, a sloper, a piece of wire hung in a hedge are all good examples. A ground mounted half wave vertical has a peak radiation angle of 20°, so it makes a good choice for DX.

I have been experimenting with feeding end fed half wave antennas matched by a parallel tuned circuit coupler. This article will explain my findings and reasons for constructing it the way I did.

A dipole can be fed anywhere along its length. A centre feed gives around 70Ω. A 'Windom' type off centre fed dipole (fed 38% along its length) provides around a 200Ω feed. The feed impedance at the very end of a half wave is thousands of ohms, usually somewhere between 2000Ω to 5000Ω, which we need to match to our 50Ω transceiver. The problem with end feeding a half wave is also its advantage. The high impedance means that the feedpoint has a very high voltage but low current, therefore very little ground is required. A very small counterpoise should be adequate from 7 to 30MHz, or you can even use the coax and transceiver

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Accept Read More

Below is the schematic of the final build.

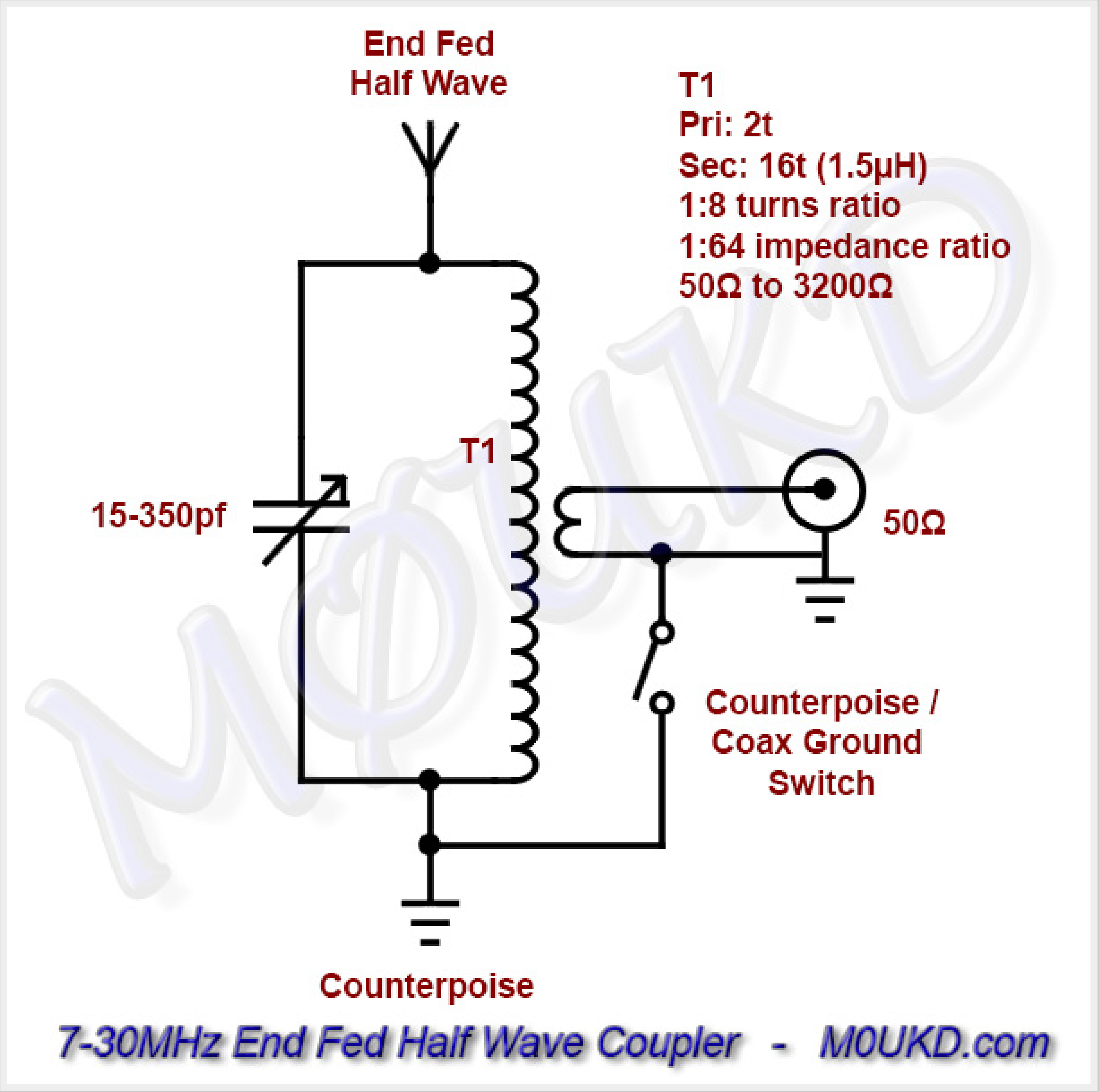
End Fed
Half Wave
T1
Pri: 2t
Sec: 16t (1.5μH)
1:8 turns ratio
1:64 impedance ratio
50Ω to 3200Ω
T1
15-350pf
50Ω
Counterpoise /
Coax Ground
Switch
Counterpoise
7-30MHz End Fed Half Wave Coupler - M0UKD.com
M0UKD

End Fed Half Wave Antenna Coupler Schematic – 7-30MHz

For calculating the length of a half wave in metres, I use 141÷*f* (MHz) for wire elements. These dimensions were derived from modelling the antenna in EZNEC. I have found this calculation to work well, however it depends on many factors such as wire used, location etc. I have made a javascript calculator below for simplicity with a 15m half wave ready to go!

Enter the frequency: 21.25 MHz

Calculate

Half Wave antenna is: Metres

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

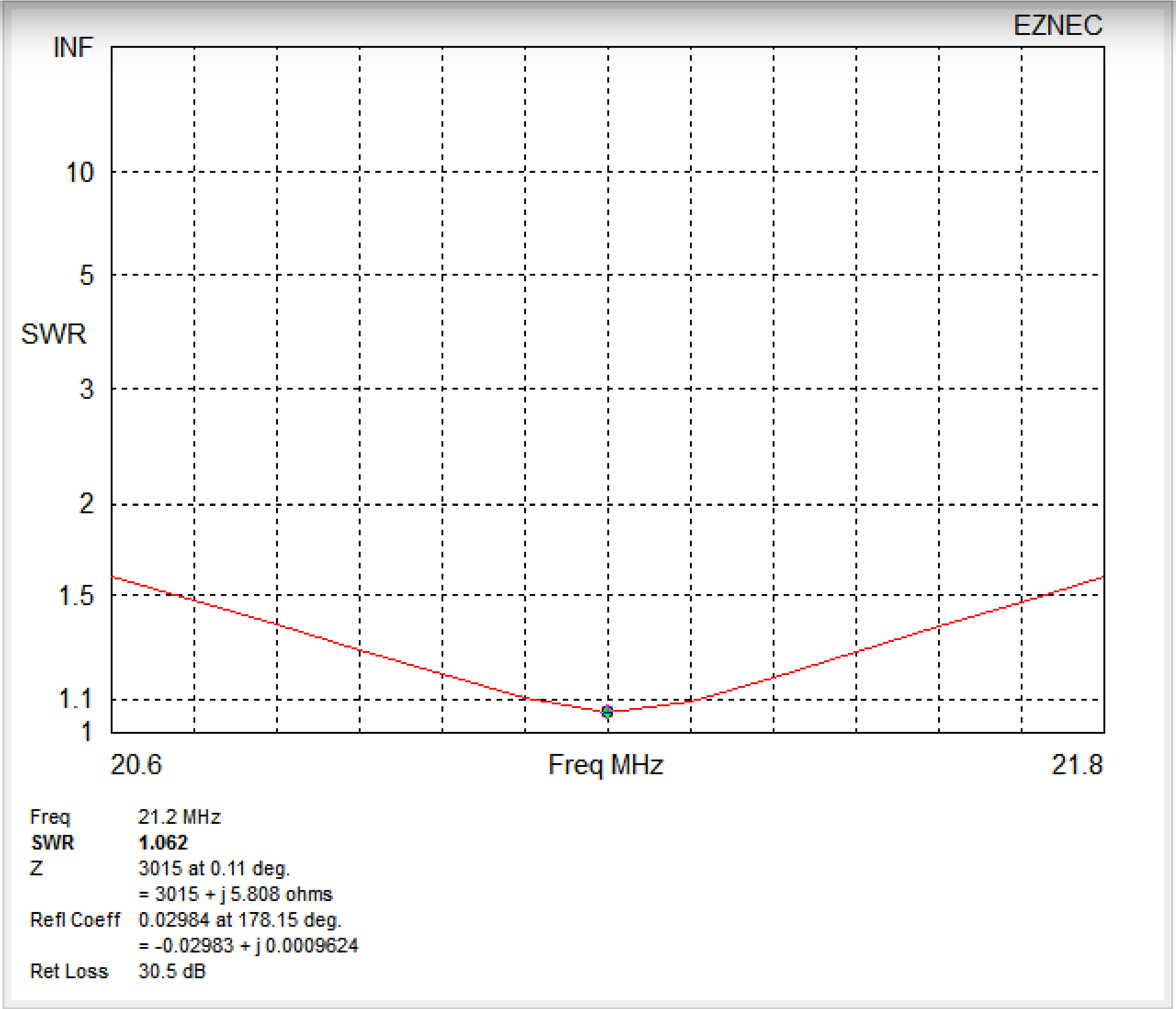

```
Freq        21.2 MHz
SWR         1.062
Z           3015 at 0.11 deg.
            = 3015 + j 5.808 ohms
Refl Coeff  0.02984 at 178.15 deg.
            = -0.02983 + j 0.0009624
Ret Loss    30.5 dB
```

End Fed Half Wave (6.65m @ 21.2MHz, ground mounted) SWR – Actual bandwidth on the 50Ω side of the coupler will be much narrower, due to the tuned circuit we are using. This is showing the feed point SWR at 3200Ω

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

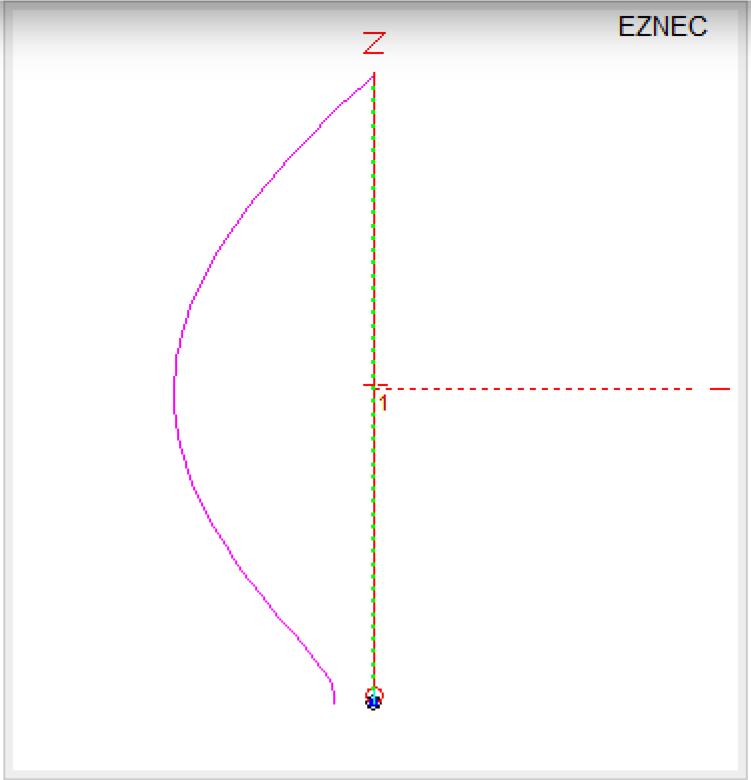

Current along the half wave. Notice how the feedpoint is not quite as high impedance as the top of the antenna.

First off, construction will be a juggle between available components and the desired band coverage. I had a capacitor that I wanted to use, which is a Johnson 154-2 air variable capacitor, with a range of 15-353pf. I would like to cover 7 to 30MHz with my coupler if possibe, so the next thing I needed to design was an inductor that would resonate in a parallel LC circuit at just above 30MHz when the capacitor is at minimum. Using this **Resonant LC Calculator**, I worked out that an inductance of 1.5µH will resonate at about 33MHz when the capacitor is at 15pf and 6.9MHz when the capacitor is at its maximum of 353pf. Sounds perfect to me! If you have a smaller value capacitor, for example 200pf maximum, it should still cover 10MHZ to 30MHz. A switch could be added to add a further capacitance of 150-200pf to include 7MHz operation.

So, lets build the secondary transformer inductor first. I wanted to transform 50Ω to around 3000Ω. This would require a 1:8 turns ratio. Impedance transformation is calculated by squaring the secondary turns ratio (note ratio, not actual turns) 8^2 is 64. 50×64=3200Ω (our input impedance x 8^2). I looked at using a **T200-6** iron powder toroid, but the problem was 1.5µH required *only* 12 turns. As I wanted a 1:8 turns ratio (for a 1:64 impedance ratio), I needed a primary of 1.33 turns, which was going to be impossible. I could have made it 16 turns and a 2 turn primary, but then the inductance would be too large for 30MHz.

So, I decided to use an air wound transformer. This way, I can build it with a 16 turn secondary, *and* at the inductance I wanted by altering the diameter and/or length of the coil. Using this **Air Inductance Calculator**, I worked out that a 19mm diameter, 52mm long, 16 turn inductor should give an inductance of 1.5uH, so this is what I built.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

drywipe marker
rizontally

Winding the secondary inductor

Once it was wound, I tested it on an LC meter, which confirmed it was around 1.5µH. You can expand or contract the inductor to fine tune the impedance to suit. The importance is that it remains 16 turns to match with the 2 turn primary. I then wound the primary coil, which had to be a slightly larger diameter so that it could fit over the secondary coil and provide inductive coupling. You can see the finished 2 turn primary and 16 turn secondary below.

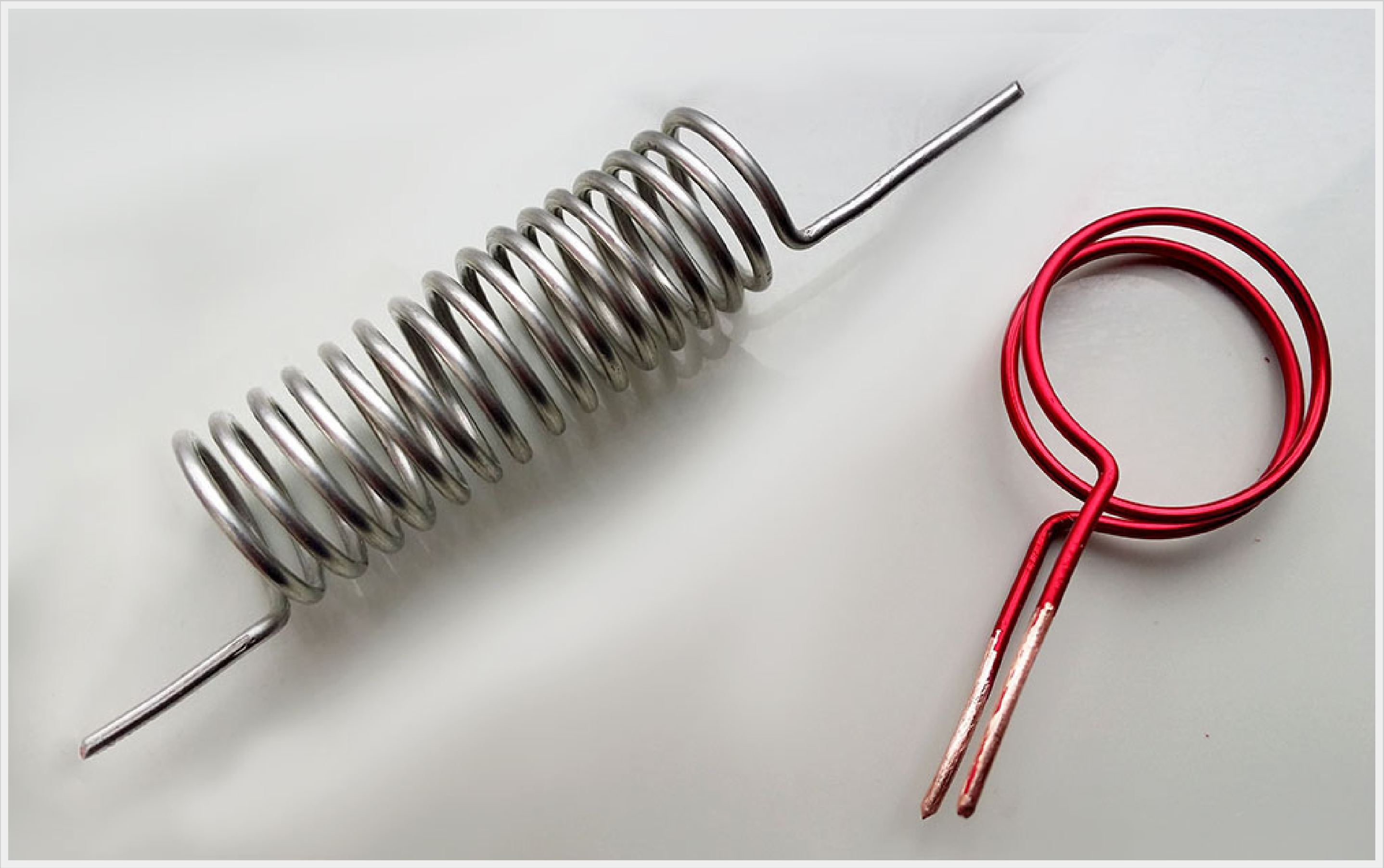

The completed inductors. They have their differences, but I will make them a happy 'couple'!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

secondary and a terminal block glued to a small block of MDF for the primary. The coax is soldered to the primary, the terminal block is just to hold it in place. There is a counterpoise connection on the side, the antenna connection on the front and the capacitor, SO239 and counterpoise/coax ground switch on the rear. Some photos of the completed unit are below.

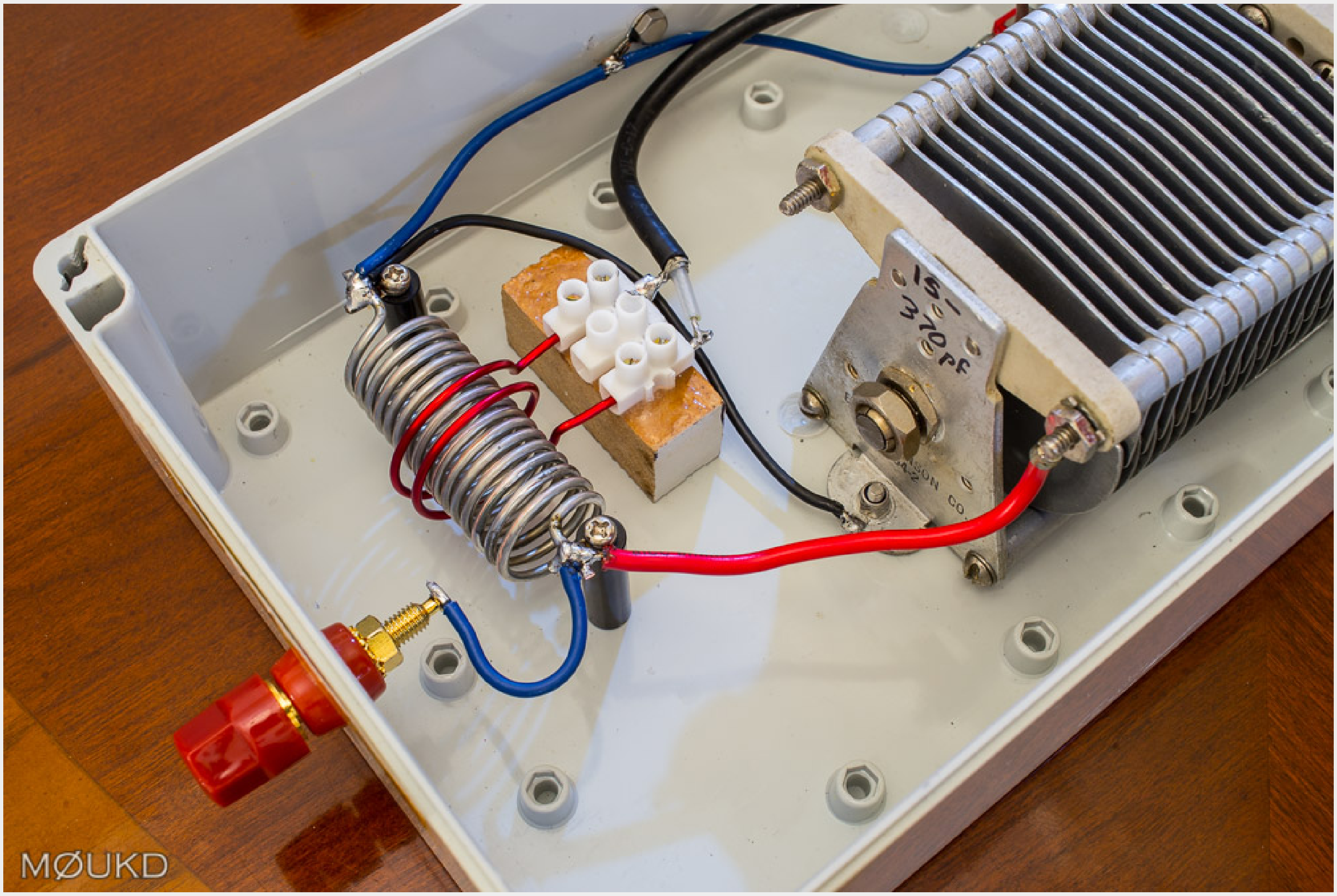

The completed End Fed Half Wave Coupler

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

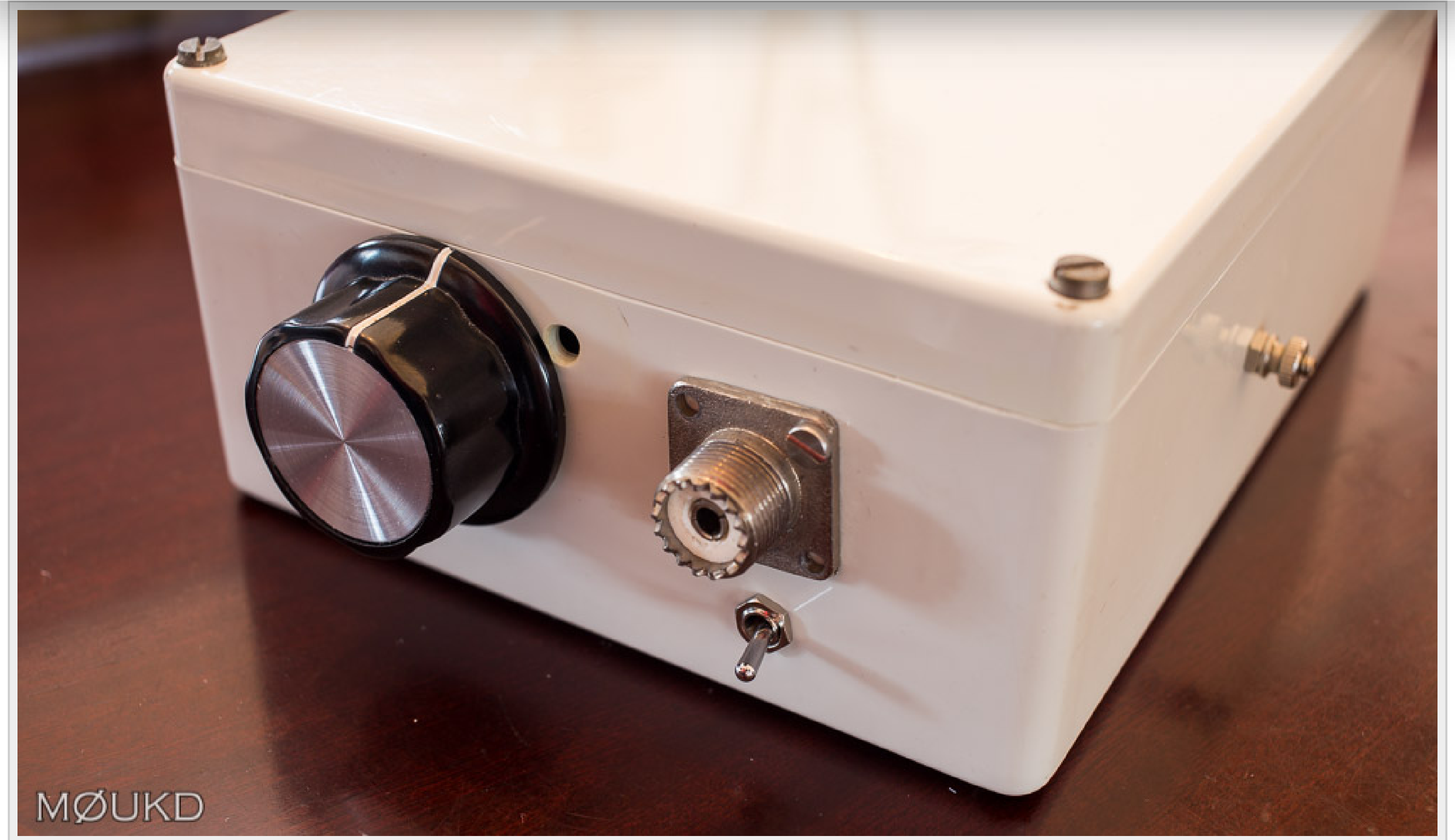

The tuning control, SO239, ground switch and on the right, the counterpoise connection.

DANGER!
High Voltage RF!
MØUKD

The antenna connection. The second hand box had a few holes in it, so I made this sticker to cover them up!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

this is due to extra stray capacitance. I tuned the coupler to resonance into a resistor (in my case 3.2kΩ) so a 1:1 SWR was obtained on my chosen frequency of 21.25MHz, then I removed the resistor and attached the half wave antenna (6.65m) and counterpoise (66cm) from the above calculator. Without adjusting the variable capacitor, I had a 1:1 SWR. This confirmed that I had a resonant half wave and that I was feeding it at or very close to the peak voltage point and the feed impedance was around 3200Ω. I found that a small counterpoise was required. Sometimes, the stray capacitance to ground was good enough on its own, especially on the higher bands and if the coupler was laying on the ground. If its all away from ground, I get consistent results using the above setup. I have had good success also by using the coax as the return, by connecting the 'bottom' of the primary and secondary together via a switch to use the coax shield and transceiver as a counterpoise.

You may be able to adjust the capacitor to match the antenna if its not a perfect half wave, but then it will be a higher current feed and the simple ground system will be inadequate. The coupler will also not be working as it was designed and will be inefficient, generating heat. The coupler is a tuned impedance transformer, not an antenna 'tuning' unit. Stick to the above, and it should be a sure thing!

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

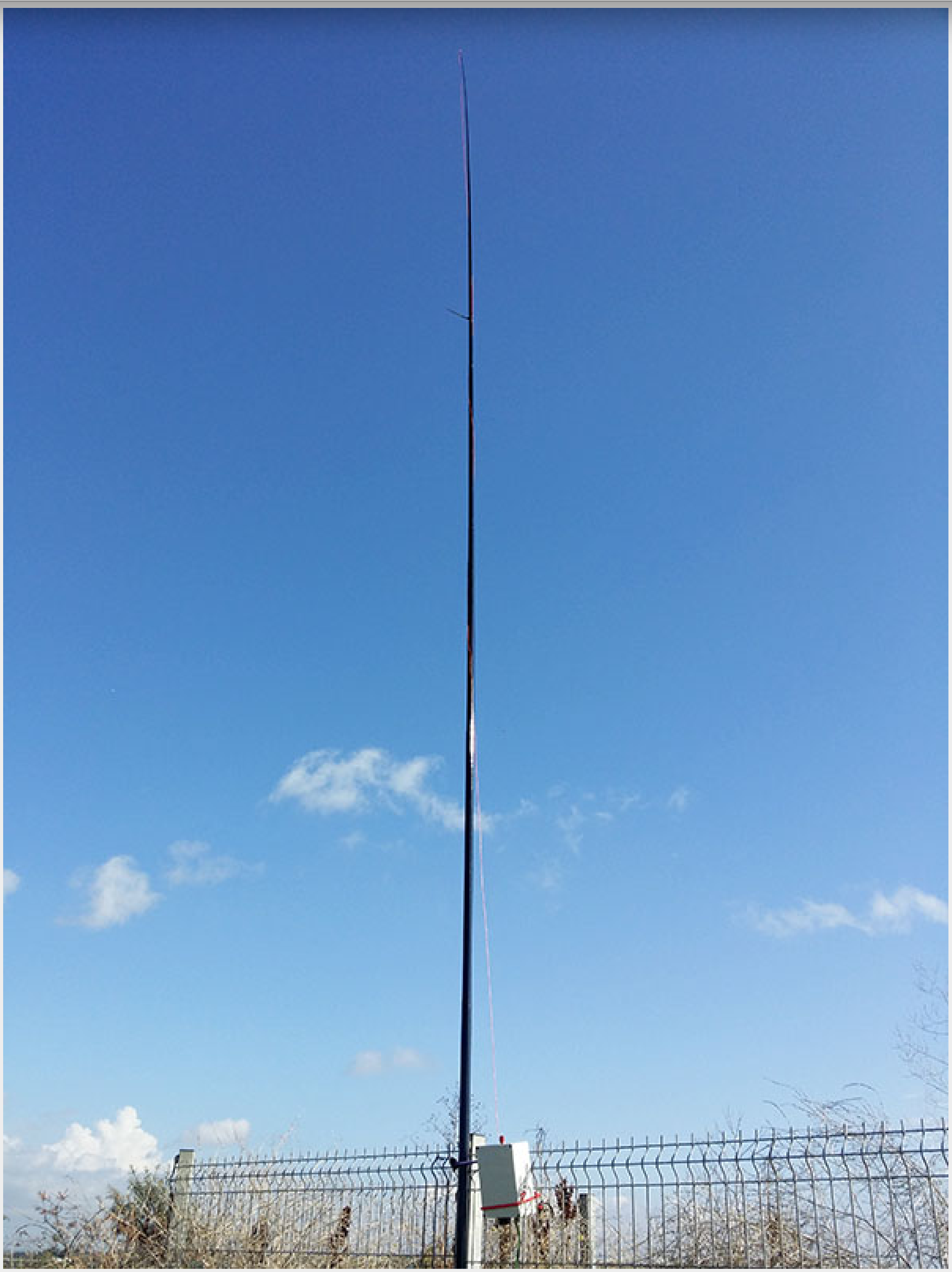

A half wave vertical for 21MHz. 6.65 metres of wire on a 9 metre fibreglass pole. The coupler is mounted 2 metres up and attached to the top of the fence. The 67cm counterpoise just hangs down in free air.

Below is a short video from my second test with the coupler, here on 17m.

This website uses cookies to improve your experience. We'll assume you're ok with this, but you can opt-out if you wish.

Read More

END FED 6–40 Meter Multiband HF Antenna

Introduction

This EARC project produces an inexpensive, multiband, end fed HF antenna matchbox that is quick and easy to setup and use. The end fed feature adds portable convenience, but does present another issue. The problem with an end fed half wave antenna is that the antenna presents high impedance, creating a significant miss match with the usual transceiver impedance of 50 ohms. This miss match is significantly greater than typical tuners can accommodate without a matching transformer.

This project creates a trifilar wound, 9:1 UNUN (unbalance to unbalance) toroid matching transformer that will match the high input impedance of an end fed antenna into the range where most antenna tuners can produce good performance. The matchbox handles 100 watts of power. ***This project requires an antenna tuner to achieve satisfactory SWR***. If you need an auto tuner, we recommend the LDG line of auto tuners, shown to work well with these matchbox antennas, and highly rated by users. They are reasonably priced and available at most ham stores.

The matchbox project uses readily available common hardware and materials listed below.

Matchbox Parts List

End Fed HF
Matchbox Antenna

1	small plastic enclosure (shown right)
1	powdered iron toroid T130-2
3	20” pieces of 18 AWG solid insulated copper wire in red, green, and black
2	# 8-32 x 3/4” hex head machine screws
2	# 8 lock washers
2	# 8 ring wire lugs
2	# 8 flat washers
2	# 8 lock washer/nut combination
2	# 8 wing nut
1	SO-239 panel mount connector to fit keyed enclosure opening

30’ # 18 AWG insulated stranded wire antenna with ring lug attached Small amount of clear silicone caulk to secure toroid in place

Preparing the MATCHBOX Plastic Enclosure

The enclosure needs one 5/8” hole for the SO-239 connector, and one 11/64” hole for the counterpoise wing nut on the lower side of the enclosure. An 11/64” hole on the upper right side of the box facilitates antenna connector mounting.

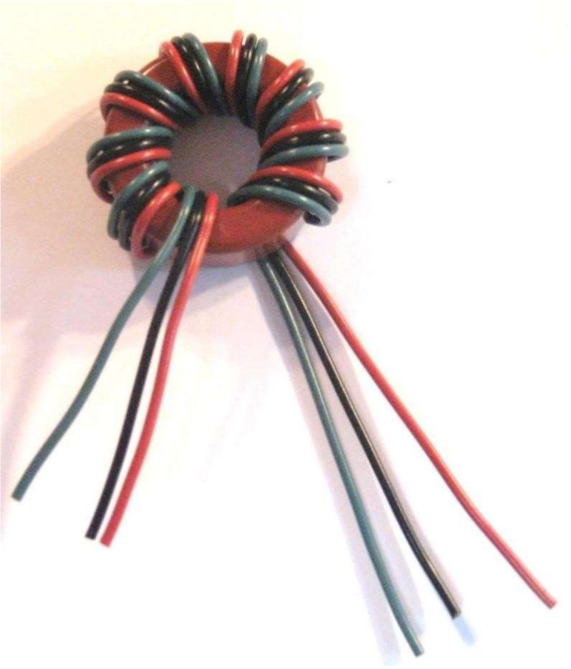

TOROID WINDING

First, wind the three 20” pieces of insulated solid wire onto the toroid. Place the wires as shown green-black-red, and wrap the toroid 9 turns so that it looks like the photo on the right. Notice there are three wires extending from the left winding and three wires extending from the right winding. As the connections are completed, the steps refer to the specific wires by left or right and color.

NOTE: Count turns by counting the number of times the wire goes thru the toroid center.

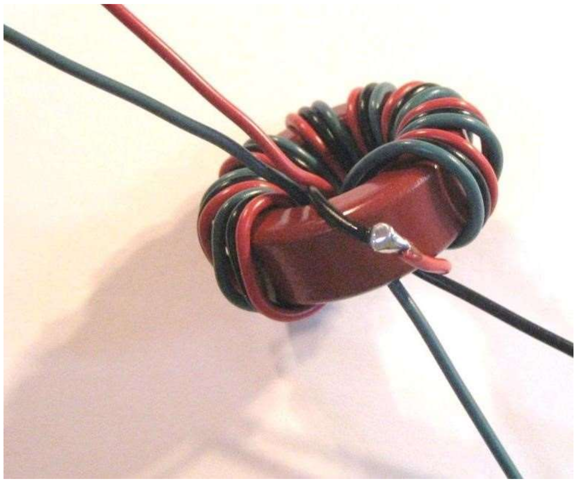

Crimp together and solder ***__the left black wire__*** with the ***__right red wire__*** When the step is completed, it will look like the right photo.

The next three steps should appear as shown in the first photo on page 3. Crimp and solder a # 8 lug to the ***__left red__*** wire about 2” from the toroid. The completed lug will later connect to the antenna connection bolt on the upper inside of the enclosure.

Twist the **__left green__** wire with the ***__right black__*** wire. Strip the ends of the two wires and twist together at about 2”. This twisted pair will solder to the center connection of the SO-239 connector in a later step.

Trim and strip the remaining ***__right green__*** wire at about 2”. Cut an additional 2” piece of green wire, and crimp and solder both wires to a # 8 lug. The 2” green wire will connect to the ground connector on the SO-239 already installed in the enclosure. Strip remaining green wire end 3/8” and bend into a hook for connection to the SO-239 ground connector.

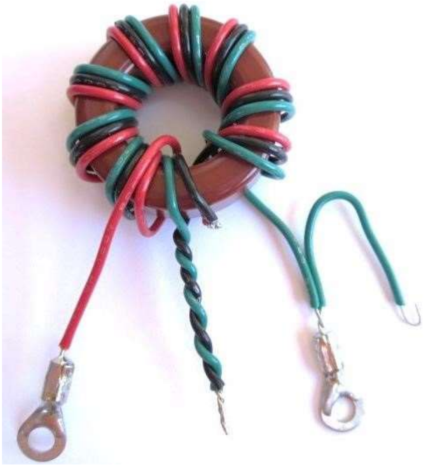

The SO-239 connector and ground lug should be installed through the keyed hole in the lower right with the mounting nut securely tightened.

Solder the green and black twisted pair to the SO-239 center connector.   Solder the green wire hook to the ground lug on the SO-239 connector.

From inside the box, place an 8-32 machine screw through a lock washer, the # 8 lug on the green wire, then through the lower 11/64’” mounting hole. Place a flat washer on the outside protruding machine screw followed by # 8 lock washer/nut and tighten securely. Place a # 8 wing nut on the machine screw to finish the **counterpoise connector**.

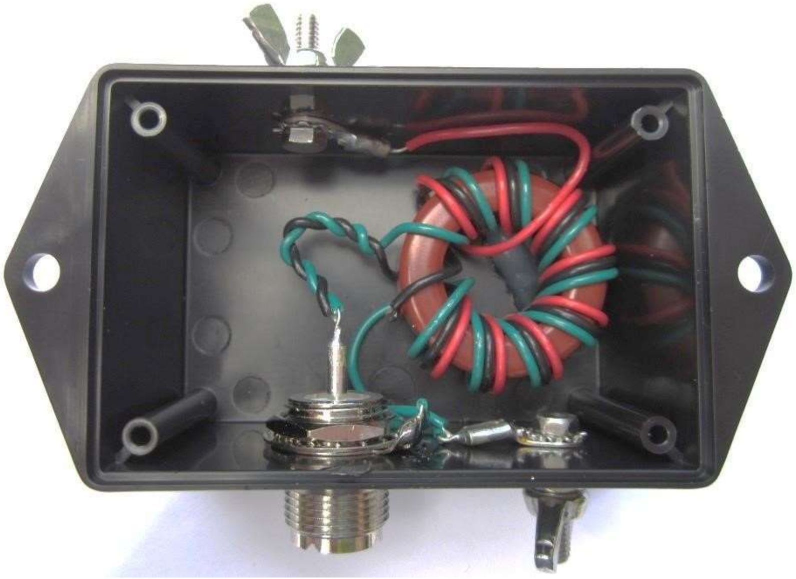

Position the toroid inside the box to allow connection of the red antenna wire lug to an 832 machine screw and lock- washer on the upper box side. Place a flat washer on the outside of the box followed with a # 8 lock washer/nut. Tighten the nut securely. Next, place the wing nut on the **antenna connector** and your project appears as shown above.

A small spot of clear silicone compound is used to secure the toroid from movement in the enclosure.   The final assembly step is to secure the box cover in place with four screws.

Preparing the Antenna Wire

Matchbox performance will be determined by two factors: The length of the antenna wire, and the capability of the tuner. The length of the antenna wire should be between 24 and 60 feet for best performance. Additional counterpoise is not normally required, as the coax shield provides the counterpoise function. Wires longer than 60 feet may have excessive impedance for some tuners to properly match. Wires shorter than 24 feet may not radiate as effectively. A 30 foot insulated 18 gauge stranded wire antenna and connecting lug is included with the project and should meet most requirements without need for any counterpoise.

Experience has shown that most external tuners and some internal tuners will tune 80–6 meters with an antenna length of 24' to 30'. If a longer antenna is desired, the provided antenna can be lengthened.

Some tuners, in particular internal tuners, may not tune the full 80-6 meter range. You may need to try different wire lengths to optimize your antenna configuration. If you are having difficulty getting your rig to tune, start with a 26' wire. This should produce good results on at least 40-6 meters using the narrowest performance range of internal tuners.

Best performance is achieved with a coax of 16' or longer, with the coax shield providing the counterpoise function. Additional counterpoise is usually not required in this design, although the lower wing nut provides a convenient counterpoise connector if needed. The end fed antenna system works well in horizontal, sloper, and vertical configurations.

Observe established safety practices when working with antennas, and avoid proximity to power or utility wires. Permanent installations should be equipped with appropriate static and lighting protection.

Keep amateur radio safe and fun!

End Fed HF
Matchbox Antenna

Toroid Wiring Schematic

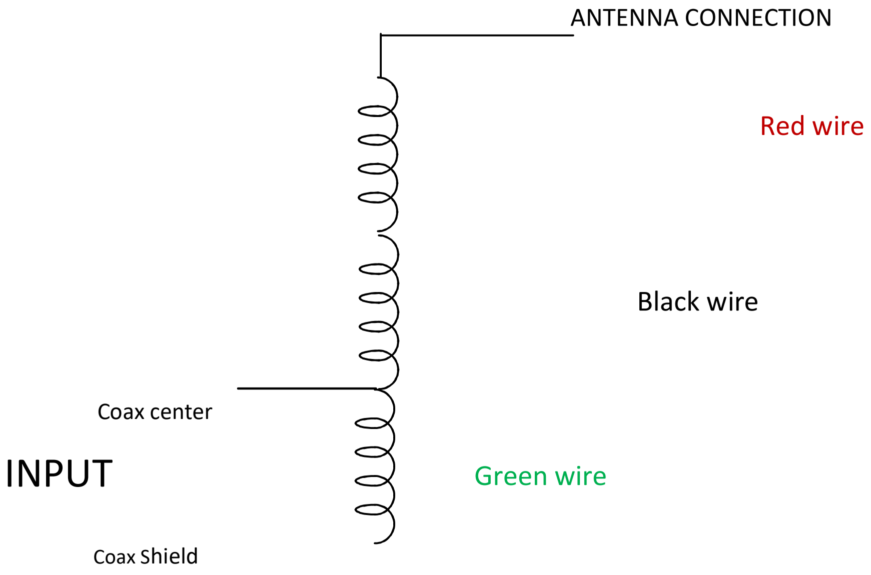
ANTENNA CONNECTION
Red wire
Black wire
Coax center
INPUT
Green wire
Coax Shield

Feedback on 6-40 End Fed Matchbox Antennas

The antenna is amazing. I receive fantastic signal report from DX stations as well as hams around the country. I previously used a short version GR5V that was in an inverted V configuration on the roof of our house. This antenna was noisy but worked. Your antenna out performs the prior antenna and is much quieter. Thanks again for the fine product that your club offers. ***May 2012***

It took me 10 minutes to install and WOW - within 10 minutes I made my first HF contact (from CT to GA). And in the span of 4 hours made 3 more !! I was transmitting just 5 Watts from Yaesu 857ND using this antenna. This antenna rocks ! Thanks a ton ! 73 ***May 2012***

The 6 - 40 Matchbox antenna works great! Easily tunes with the Z100 tuner - very low SWR's on 20m thru 10m. Antenna was up about 45 feet on tree branch - wire length was about 52.5 feet, slight slope, 55 feet of coax, line isolator at radio. ***Work Hawaii for the very first time. Guess the antenna wanted to call home :-) 57 report, 100 watts, 15 Meter band, Icom 718.*** A great portable antenna that I will use often at Montrose Harbor along the lakefront in Chicago. Thank you for a fine, well made product. ***April 2012***

It arrived today in fine shape. I got is up in the air this afternoon using a 31 foot radiator. Top is at 45 feet. Loads on all bands 80-10M with my LDG auto-tuner absolutely no problem, My first QSO was KH7X in the ARRL SS contest with 100W. Amazing. This thing rocks! It’s a great antenna! Nice job and a great buy. ***November 2011***

I set-up my jackite pole today and tried a 68' wire with the 9:1 unun matchbox. I set it up as an inverted L with about 26 feet vertical and 42 feet horizontal. With my IC-703 it tuned 80 through 6 all well below 1.5:1. I am going to leave the antenna up a few days and make a few contacts. I think the matchbox is terrific. ***October 2011***

My 6 – 40 meter end fed arrived through the UK holiday mail and I departed from my norm of a sloper and mounted it vertically on a 10m fishing pole. It is surprisingly effective and rewarded me with a surprise contact on 17m into the Falkland Islands at 20:30UT last night. ***January 2012***
When I attached it to a 10m fiberglass fishing pole and went vertical - wow! Easy 5/9 contacts out to 6,000+ miles when propagation was anything better than the worst. My simple ATU easily matches the antenna with SWR never greater than about 1.4 on any band it is built for. ***December 2012***

You can dither and procrastinate, but with this antenna, you'll be talking to the world in no time on any band that happens to be hot - and with your hard-earned cash hardly touched! ***January 2012***

Just wanted to report back that I have tossed the antenna into a tree about 20' up in an inverted vee shape and I am pleased to report that I can tune anywhere I need on 40-10M using the TS-590 internal tuner!! My first contact was on 40M within 10 minutes of erecting the antenna and was all the way to TX from my condo here in MI using just 10 watts. This is by far the best HAM related investment I have made in a long time. Living in a condo with strict association rules, I am able to conceal this wire in a tree out my bedroom window and talk all over. I'm very happy and would recommend this antenna to anyone in a similar situation. ***March 2013***

I have had your multiband end fed antenna installed here in Singapore for a few weeks now and am very pleased. I used the wire from my previous antenna which is 66 feet long. It works beautifully! I have it hanging vertically outside my 15th story apartment window on a 5 foot horizontal pole (to get it away from the reinforced concrete building). A couple of fish weights on the end keep it from blowing around too much. I would recommend this antenna to anyone who is facing a challenging QTH. ***April 2013*** I received my antenna yesterday, a quick 4 days since I ordered it. I attached it to a 31 foot pole and my IC-706MK2. Within minutes, I made my first contact with 9A4KW in Zagreb, Croatia on 20 meters. Being able to reach out over 4,000 miles gives me real confidence that I have found my portable antenna. Thanks for your rapid response and fine product. ***May 2013***

In only 2 weeks of casual on-air time. If you could see how I had to zig-zag it through a big tree above the roof of my RV, you'd be saying like I am, "how the heck can that work? I worked 90% of the stations I called, and just added Mexico City on 15m that gave me a 5x9. ***September 2013***

Just wanted to let you know how pleased I am with my ERAC matchbox antenna. I put the EARC antenna in the backyard with the feed point at 4 feet and the wire sloping up to 35 feet. I have 125 feet of coax feed line. My transceiver's built-in antenna tuner will adjust the SWR to below 1.2 to 1 on all bands 80-6 meters. Last weekend I worked 65 countries on 5 continents. I live in an antenna restricted community and this antenna has allowed me to enjoy ham radio again. ***February 2014***

I used the EARC Matchbox today. It works GREAT on 20m with a 30 ft radiator in vertical positioning, fed with 25ft of RG175/u coax. I've been talking up and down the east coast with 3W on SSB all afternoon! People ask me to repeat my power output again and again. I can't wait until I get the chance to use it at night, where I think it'll really shine! This has a permanent place in my portable kit. Thanks for the great product! ***April 2014***

I performed a simple install in Virginia making it vertical by using a 28 foot fiber glass kite pole. Within 30 minutes (NO KIDDING) I was receiving Lebanon on 17 meters! I also worked Russia and Eastern Europe that night on 20 meters. I love this antenna and I give it a 100% (5 STAR) rating. ***May 2014***

This little box is magic. NY, IA and NM from GA with ease on 100w ***June 2014***

Just wanted to let you guys know, this matchbox is wonderful. I built the matchbox, connected it to 55 feet of solid #14 about 20 feet in the air. My rig is a Yaesu FT-817ND connected to a MFJ-971 tuner with coax running between the tuner and the matchbox. So far I have worked Slovenia twice, Morocco, St. Helena Island, Cuba, Cost Rica, Vermont, Wyoming, Florida, North Dakota, Brazil, and Colombia. This setup works really well on the higher bands. Thank you for putting this design on the internet. I like it so much its become permanent here. ***September 2014***

This antenna continues to amaze me. It worked well U-shaped "indoors" with 50 watts, and now outdoors as a sloper. It easily loads 10-40m with my TS-570s internal tuner. I'm sold on it and own two of them now. I'm working at least 90% of the DX I chase. I have no RFI in the shack either. ***November 2014***

EARC Antennas are operating in virtually every state, these countries and more!

Argentina Australia Belgium Brazil Canada Chile Denmark Greece

Hong Kong India Ireland Israel Italy Mexico Mozambique New Zealand Russia Singapore
Sweden Tasmania United Kingdom

ORDERING A MATCHBOX ANTENNA FROM EARC

Volunteers of the Honolulu Emergency Amateur Radio Club (EARC) carefully assemble the 6-40 End Fed matchbox antennas for those who prefer not to build one. For a donation of **$56** (Including USPS priority mail shipping) club volunteers will build and deliver a 6-40 matchbox with a 30 foot antenna wire anywhere in the U.S., we do not ship internationally. Note that coax cable is not included.

Payment by PayPal

To pay via **PayPal**, make your payment to paypal@earchi.org using any major credit card. Include your name, mailing address, and email address with the order.

We will ship promptly and you will likely receive the order in the US in 7 days or less. If you have any questions about your order, please email webmaster@earchi.org

Thank you for your interest in the activities of Honolulu Emergency Amateur Radio Club and amateur radio.

Wire Antennas for Ham Radio

Iulian Rosu YO3DAC / VA3IUL,

http://www.qsl.net/va3iul

Note: *Dimensions of the antennas presented below are either in: feet ('), inch ("), meters (m), or centimeters (cm)*

01 - Tee Antenna
02 - Half-Lamda Tee Antenna
03 - Twin-Led Marconi Antenna
04 - Swallow-Tail Antenna
05 - Random Length Radiator Wire Antenna
06 - Windom Antenna
07 - Windom Antenna - Feed with coax cable
08 - Quarter Wavelength Vertical Antenna
09 - Folded Marconi Tee Antenna
10 - Zeppelin Antenna
11 - EWE Antenna
12 - Dipole Antenna - Balun
13 - Multiband Dipole Antenna
14 - Inverted-Vee Antenna
15 - Sloping Dipole Antenna
16 - Vertical Dipole
17 - Delta Fed Dipole Antenna
18 - Bow-Tie Dipole Antenna
19 - Bow-Tie Folded Dipole Antenna for RX
20 - Multiband Tuned Doublet Antenna
21 - G5RV Antenna
22 - Wideband Dipole Antenna
23 - Wideband Dipole for Receiving
24 - Tilted Folded Dipole Antenna
25 - Right Angle Marconi Antenna
26 - Linearly Loaded Tee Antenna
27 - Reduced Size Dipole Antenna
28 - Doublet Dipole Antenna
29 - Delta Loop Antenna
30 - Half Delta Loop Antenna
31 - Collinear Franklin Antenna
32 - Four Element Broadside Antenna
33 - The Lazy-H Array Antenna
34 - Sterba Curtain Array Antenna
35 - T-L DX Antenna
36 - 1.9 MHz Full-wave Loop Antenna
37 - Multi-Band Portable Antenna
38 - Off-center-fed Full-wave Doublet Antenna
39 - Terminated Sloper Antenna
40 - Double Extended Zepp Antenna
41 - TCFTFD Dipole Antenna
42 - Vee-Sloper Antenna
43 - Rhombic Inverted-Vee Antenna
44 - Counterpoise Longwire
45 - Bisquare Loop Antenna
46 - Piggyback Antenna for 10m
47 - Vertical Sleeve Antenna for 10m
48 - Double Windom Antenna
49 - Double Windom for 9 Bands
50 - Collinear Trap Antenna
51 - Short Dipole Antenna for 40m - 80m - 160m
52 - Center Fed-Zepp Antenna for 80m - 40m
53 - All Bands Antenna
54 - All-Bands Dipole Antenna
147 - KE4PT OCEF All-Band Dipole
148 - Wire Quad Antenna for 40m
149 - Inclined Dipole Antenna for 80m, 40m
150 - Pyramidal Wire Antenna for 80m
151 - Random Wire Antenna All Bands
152 - Multiband Dipole Antenna for 80m, 40m,15m, 10m
153 - Slim Jim Wire Antenna for 4m
154 - Delta Loop for 6m
155 - Re-Configurable Antenna for 160m and 80m
156 - Very Low Frequency Inverted-L Antenna
157 - Reduced Size Half Sloper Antenna for 160m
158 - Tree-Mounted HF Antenna
159 - Multiband Vertical Antenna for 80m, 40m, 20m
160 - Marconi Antenna for 136 kHz
161 - Simple Killer Antenna for 40m
162 - Stub-Directed V Antenna for 80m
163 - KT0NY Over-and-Under DX Antenna for 20m
164 - Horizontal Loop Antennas
165 - Ribbon J-Pole for 2m
166 - Dual Band Ribbon J-Pole for 2m, 70cm
167 - Square Vertical Loop Antenna for 40m
168 - Tri-Band Quad Antenna for 20m, 15m, 10m
169 - 3D Quad Antenna for 80m, 40m, 20m, 15m, 10m
170 - Sloping Wire Antenna for 30m, 20m, 17m, 15m, 12m, 10m
171 - Broadband Dipole Antenna with Coaxial Resonator for 80m
172 - Resonant Feedline Dipole Antenna for 80m
173 - Dual-Band Loading Wire Antenna for 80m, 30m
174 - Stub Matching Antennas
175 - J-Style Vertical Wire Antenna for 10m
176 - Dual Band Vertical with Zepp Feeders for 40m, 20m
177 - RCA Double Doublet for 40m to 12m
178 - RCA Spiderweb Antenna for 40m to 6m
179 - Folded Dipole with Shorted Straps
180 - Twin-Lead Marconi Antenna for 160m, 80m
181 - Broadband Antenna Quarter-Wave Balun for 80m
182 - Three-Quarter-Wave Folded Doublet for Dual-Band
183 - Three-Quarter-Wave Folded Doublet No-Switch for Dual-Band
184 - Wideband Omni-Directional Discone Antenna
185 - Wideband Rhombic Antenna for 40m to 10m
186 - Pre-Cut Linear Array Antenna 3dB-Gain for 40m
187 - X-Array Antenna 6dB-Gain for 20m, 15m, 10m
188 - Double-Bruce Array Antenna 5dB-Gain for 20m, 10m
189 - Bi-Square Broadside Array 4dB-Gain for 20m, 15m, 10m
190 - Six-Shooter Broadside Array 7.5dB-Gain for 20m, 15m, 10m
191 - Triplex Flat-Top Beam 4.5dB-Gain for 20m, 15m, 10m
192 - Dual-Band Tilt Antenna for 20m, 10m
193 - Super Space Multiband Dipole Antenna for 80m to 10m
194 - Bi-Square Beam Antenna Gain=5.5dB for 10m
195 - Cousin of G5RV Multiband Antenna for 40m to 10m
196 - Cayman Quad Antenna for 20m
197 - Hentenna Gain=3dB with Bazooka Match for 6m
198 - X-Beam Antenna Gain=3dB for 20m

55 - Multiband Z Antenna
56 - Multiband Dipole Antenna
57 - Five-Bands No-Tuner Antenna
58 - Dualband Full-wave Loop Antenna for 80m-40m
59 - Loop Antenna for 10m
60 - Lazy Quad Antenna for 10m
61 - Tri-band Delta Loop Antenna for 80m - 40m - 30m
62 - Dual-band Loop Antenna for 30m - 40m
63 - Wire-Beam Antenna for 80m
64 - Dual-Band Sloper Antenna
65 - Inverted-V Beam Antenna for 30m
66 - ZL-Special Beam Antenna for 15m
67 - Half-Sloper Antenna for 160m
68 - Two-Bands Half Sloper for 80m - 40m
69 - Linear Loaded Sloper Antenna for 160m
70 - Super-Sloper Antenna
71 - Tower Pole as a Vertical Antenna for 80m
72 - Clothesline Antenna
73 - Curtain Zepp Antenna for 160m, 80m, 40m
74 - Collinear Array Antenna for 40m, 30m, 20m
75 - 160m Inverted Delta Loop
76 - Half Rhombic Unidirectional Vertical for 20m to 6m
77 - Capacitance Loaded Vertical Antenna for 160m
78 - Fan Dipole Antenna for 80m to 6m
79 - Wire Ground Plane Antenna
80 - Inverted Delta Loop Antenna for 160m
81 - Inverted-L for 160m
82 - 300ohm-Ribbon Dual Band Dipole
83 - Tri-Band Beam for 20m, 15m, 10m
84 - Mini-Horse Yagi Antenna
85 - Backpack J-Pole Antenna for 10m, 6m, 2m
86 - Fan-Dipole Antenna for 80m, 40m, 20m
87 - Capacity Tuned Folded Loop Antenna for 20m
88 - Indoor Loop Antenna for 80m to 30m
89 - Indoor Loop Antenna for 80m
90 - Double-Delta Antenna 80m and 40m
91 - Inductance-Loaded Shortened Dipole for 160m
92 - V-Beam Antenna for 15m
93 - Picnic Vertical Wire Antenna
94 - Laid-Back Quad Antenna for 80m
95 - Phased Loop Antenna
96 - Loop Antenna for TX for 160m
97 - Morgain-Dipole Antenna for 160m and 80m
98 - ZL-Special for 20m, 15m, 10m
99 - Biconical Antenna
100 - Directive Delta-Birdcage Antenna for 20m to 10m
101 - Dual Polarization Antenna for 80m and 40m
102 - Directive 300-ohm-Ribbon Folded Dipole for 15m
103 - Miniature Directive Antenna for 10m
104 - Biquad Antenna 12dBi-Gain for 2.4GHz
105 - Dual-Rhomboid Antenna for 435MHz to 870MHz
106 - Double-Bazooka Antenna for 80m
107 - J-Style Antenna
108 - Vertical / Horizontal / Circular Polarization Antenna
109 - Coax Inverted-L Antenna for 80m
110 - Indoor Compact Loop Antenna for 80m
111 - Helix Antenna
112 - Novice Vertical Antenna for 80m, 40m, 15m, 10m
113 - Stub-Loaded Shortened Dipole for 80m
114 - Six-Band Wire-Stub Trap Antenna for 40m-10m
115 - Multiband Half-Wave Delta-Loop Antenna
199 - Twin Delta Loop Antenna Gain=6dB for 160m, 80m, 40m
200 - Inverted-V Beam Antenna for 30m
201 - FD4 Windom Antenna for 80m, 40m, 20m, 10m
202 - Two Elements Phased Delta Loop for 40m
203 - Three Elements Wire Yagi for 10m
204 - Low Radiation Angle Full-Wave Loop Antenna for 80m
205 - Collinear Antenna Gain=3dB for 2m
206 - X-Ray Monoband Antenna Gain=6.5dBd
207 - Broadside Collinear Curtain Array for 20m
208 - Shrunk Loop Antenna for 80m, 40m
209 - λ/3 Multiband Dipole for 40m, 20m, 17m, 15m, 13m, 10m
210 - K3LR Sloper Antenna Gain=3dB for 160m
211 - Mini Folded Vertical Monopole Antenna for 160m, 80m, 40m
212 - Three Band Dipole for 80m, 40m, 20m
213 - Stub Matched Dipole Antenna
214 - Inverted-Vee Antenna with Bazooka Match
215 - Multiband Long Wire Antenna for 40m, 20m, 15m, 10m
216 - Converted Vee Antenna for 80m, 40m
217 - Swiss Quad Antenna for 20m, 15m, 10m
218 - Japanese Quad Antenna for 20m, 15m, 10m
219 - Seven-Elements Triangle Beam Antenna for 20m
220 - Log Periodic Wire Antenna for 40m
221 - Three-Elements 90-degrees Wire Beam for 20m
222 - Two-Elements Wire Beam for 20m
223 - Suitcase-Fit Dipole Antenna for 80m, 40m, 20m
224 - Multiband Coaxial Dipole for 80m, 40m
225 - Vertical Log Periodic Antenna for 80m, 40m
226 - Inverted Vertical Log Periodic Antenna for 40m
227 - Double Extended Zepp Antenna Gain=7dB for 15m
228 - Double Lazy-H Antenna G=10dB for 70cm
229 - Wire Log Periodic Antenna G=12dB for 15m, 20m
230 - Five-Elements Vertical Log Periodic for 80m, 40m, 20m
231 - Sloping Delta Loop Antenna for 40m
232 - Doublet Antenna for 160m
233 - Log Periodic Wire Beam for 40m
234 - G3LDO Wire Beam Antenna for 20m, 15m, 10m
235 - Phased Delta Loop Array for 20m
236 - Dual Band V-dipole for 80m, 40m
237 - Collinear Array 5dB-Gain for 15m
238 - Two-Band Quad Loop Antenna 80m, 40m
239 - K7CW Quad Antenna
240 - Broadband Short Dipole for 80m
241 - Off-Center Fed Dipole for 40m, 20m, 10m
242 - Broadband Sloper Antenna for 80m
243 - Two-Frequency Dipole Antenna for 80m
244 - Top Loaded Vertical Antenna for 160m
245 - Two Loop Beam Antenna for 20m, 15m
246 - Corner-Fed Delta Loop Antenna for 80m, 40m, 20m
247 - Loaded Delta Loop Antenna for 40m, 20m
248 - Compact Size Dipole for 80m
249 - JAWS 3dB Gain Antenna for 40m
250 - Inverted-V Antenna for 160m
251 - Lazy-U Antenna for 160m
252 - AF3V Compact Antenna for 160m
253 - Hot Beam Wire Antenna for 40m
254 - Sloping Delta Loop for 40m
255 - K6FD Multiband Antenna with T-tuner
256 - Dual Band Dipole Antenna for 17m, 10m
257 - Compact Travel Dipole Antenna for 10m, 15m, 20m
258 - End-Fire Antenna for 160m, 80m, 40m, 20m
259 - Four Band Loop for 40m, 20m, 15m, 10m

116 - Hybrid Vee for 20m, 17m
117 - Six-Shooter Array Antenna, Gain=7.5dB
118 - Multiband Ground-Plane Antenna for 10m, 15m, 20m
119 - Wire Superbeam Antenna for 10m, 15m, 20m
120 - Two Elements Delta-Loop Antenna
121 - Sterba Curtain Antenna
122 - Half-Wave Vertical Zepp Antenna
123 - Lazy-Loop Antenna for 40m
124 - Terminated Folded Dipole for 80m, 40m
125 - Short-Fat Antenna for 15m
126 - Cobra Antenna for 80m
127 - Log-Periodic Wire Antenna for 20m, 15m, 10m
128 - 5-Element Log-Periodic Vertical Antenna for 80m, 40m, 20m
129 - 2m Vertical Wire Antennas
130 - Earth-Mover Inverted-V Antenna for 40m
131 - Coax-Cable Collinear Antennas
132 - Double Bobtail Antenna for 20m
133 - Collinear Zepp Antenna
134 - Taylor Vee Antenna for 20m
135 - Collinear Vertical Antenna 6dB-Gain for 2m, 1.3m, 73cm
136 - Bi-Loop Antenna for 20m
137 - Wire Beam 6dBd-Gain for 10m
138 - Sloping Diamond Antenna 4dB-Gain for 40m
139 - Twisted Loop Antenna for 160m
140 - DX RX Loop Antenna for 160m
141 - Hentenna 3dB-Gain for 10m, 6m, 2m
142 - VK2AAR Wire Antenna for 20m
143 - 2-Elements Quad Antenna for 6m
144 - Hula-Loop Bidirectional 6dB-Gain Antenna for 17m
145 - Moxon Rectangle Beam for 15m-10m
146 - Double-D Beam 4dB-Gain
260 - Carolina Windom Antenna for 80 to 10m
261 - Shortened Loaded Dipole for 40m
262 - Shortened Loaded Dipole for 80m
263 - Low Radiation Angle Inverted Delta Loop for 80m
264 - Compact Dipole Antenna for 40m, 15m
265 - Vertical Delta Loop for 40m to 10m
266 - Delta Loop Antenna for 17m
267 - Bi-Square Beam Gain=3dBd for 17m
268 -Trap Dipole Antenna for 17m,13m
269 - Hanging Unipole Antenna for 160m, 80m
270 - Wideband Sloping Vee Antenna - 30m to 6m
271 - 5/8 Wavelength Antenna
272 - C-Pole Vertical Antenna
273 - Bobtail Antenna for 30m
274 - Bobtail Antenna for 40m
275 - Two Elements Vertical Loop Antenna for 40m
276 - Double Extended Zepp Antenna for 40m
277 - Reversible Moxon Flip Beam for 40m
278 - Spitwire Antenna for 40m
279 - Fork Antenna for 40m
280 - Veebeam Antenna 3dB-Gain for All Bands
281 - Vertical Moxon Antenna for 10m
282 - Half-Square Antenna for 30m
283 - Rotatable Vertical Delta Loop for 20m
284 - Mini-Horse Antenna 10dB-Gain for 40m-to-2m
285 - Compact Moxon Antenna 5dB-Gain for 40m
286 - Two Elements Sloping Delta Loop 8dB-Gain for 10m
287 - Two Elements Phased Vertical Antenna 3dB-Gain for 40m

---

01 - Tee Antenna

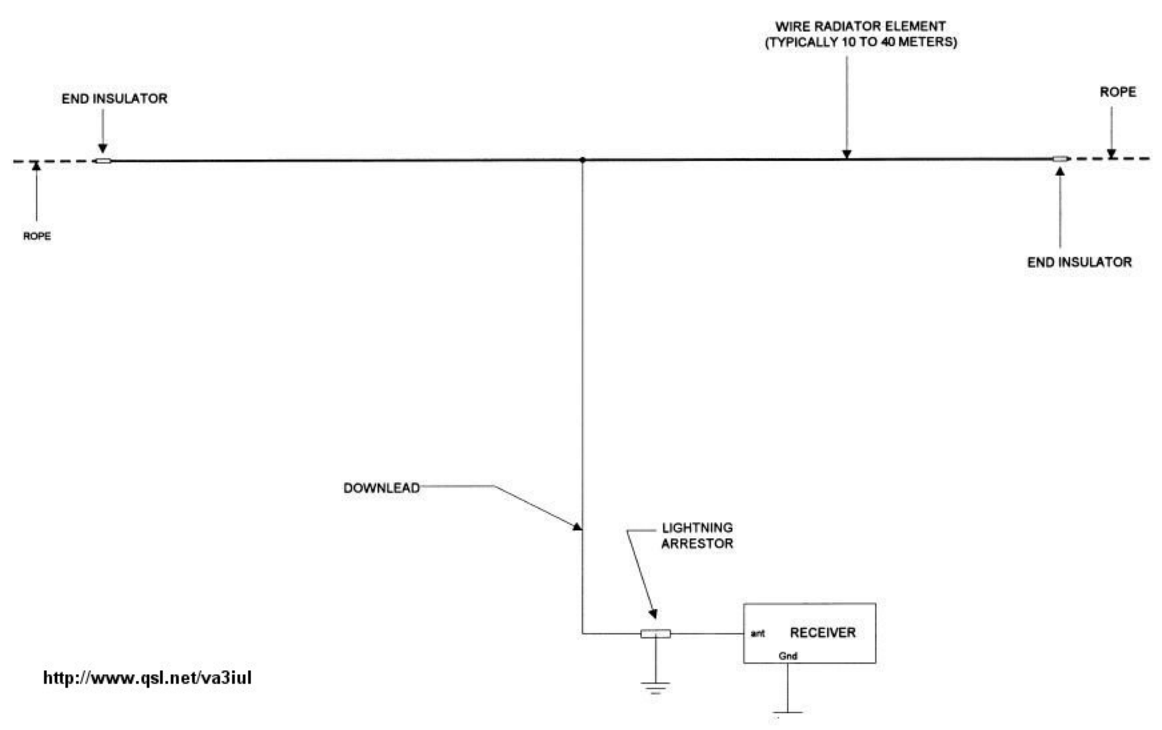
WIRE RADIATOR ELEMENT
(TYPICALLY 10 TO 40 METERS)
END INSULATOR
ROPE
ROPE
END INSULATOR
DOWNLEAD
LIGHTNING
ARRESTOR
ant
RECEIVER
Gnd
http://www.qsl.net/va3iul

---

02 - Half-Lamda Tee Antenna

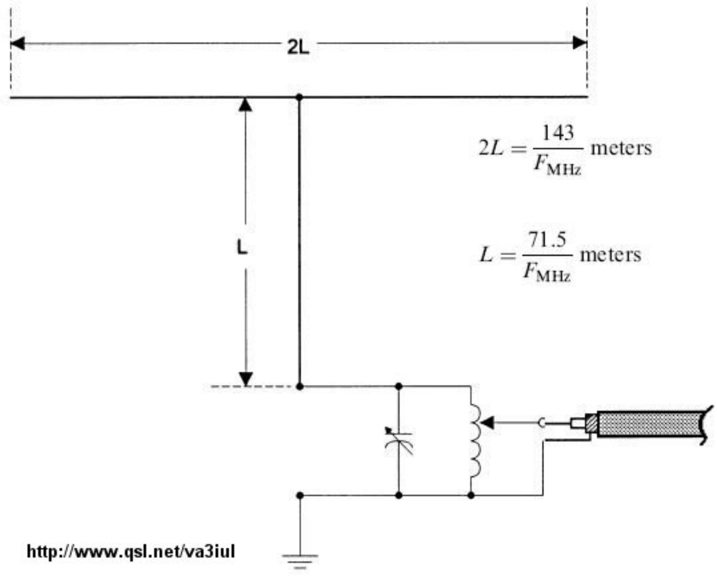
2L
L
$2L = \frac{143}{F_{\mathrm{MHz}}}$ meters
$L = \frac{71.5}{F_{\mathrm{MHz}}}$ meters
http://www.qsl.net/va3iul

---

03 - Twin-Led Marconi Antenna

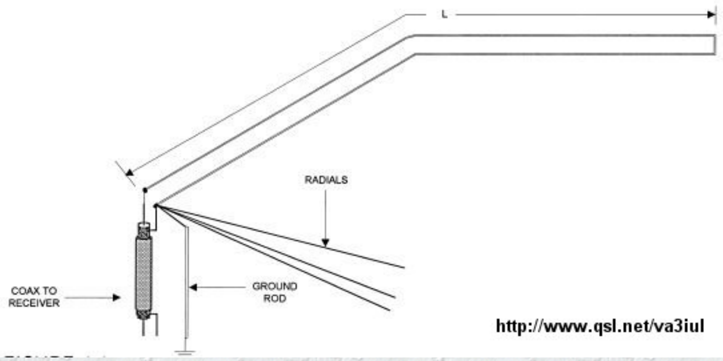
L
RADIALS
COAX TO
RECEIVER
GROUND
ROD
http://www.qsl.net/va3iul

---

04 - Swallow-Tail Antenna

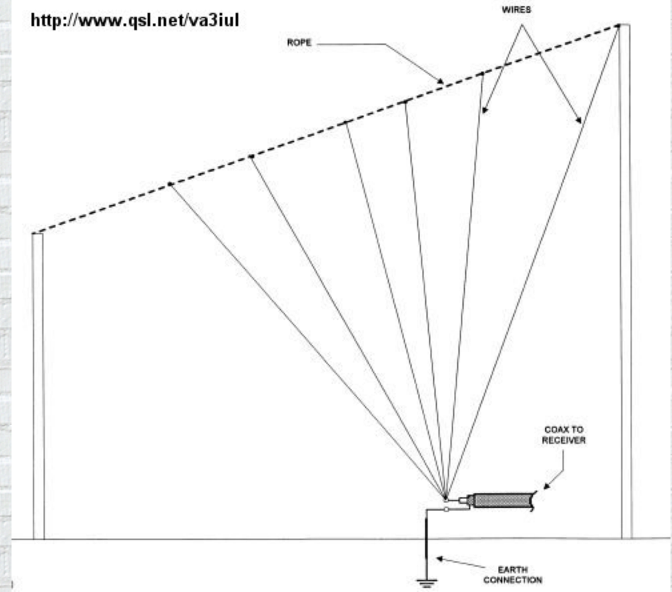
http://www.qsl.net/va3iul
WIRES
ROPE
COAX TO RECEIVER
EARTH CONNECTION

05 - Random Length Radiator Wire Antenna

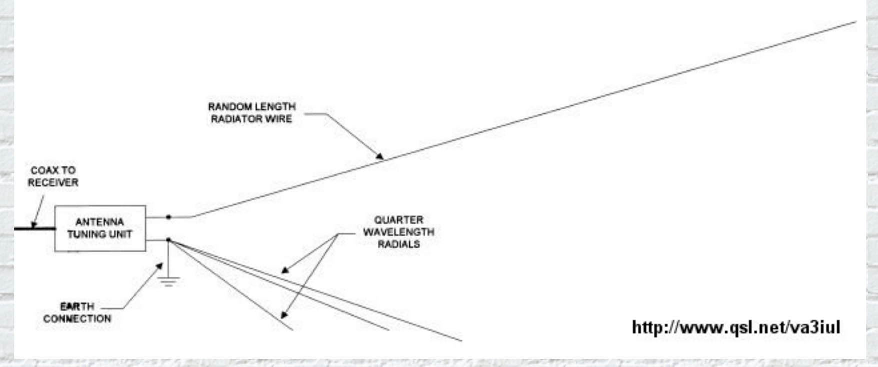
RANDOM LENGTH RADIATOR WIRE
COAX TO RECEIVER
ANTENNA TUNING UNIT
QUARTER WAVELENGTH RADIALS
EARTH CONNECTION
http://www.qsl.net/va3iul

06 - Windom Antenna

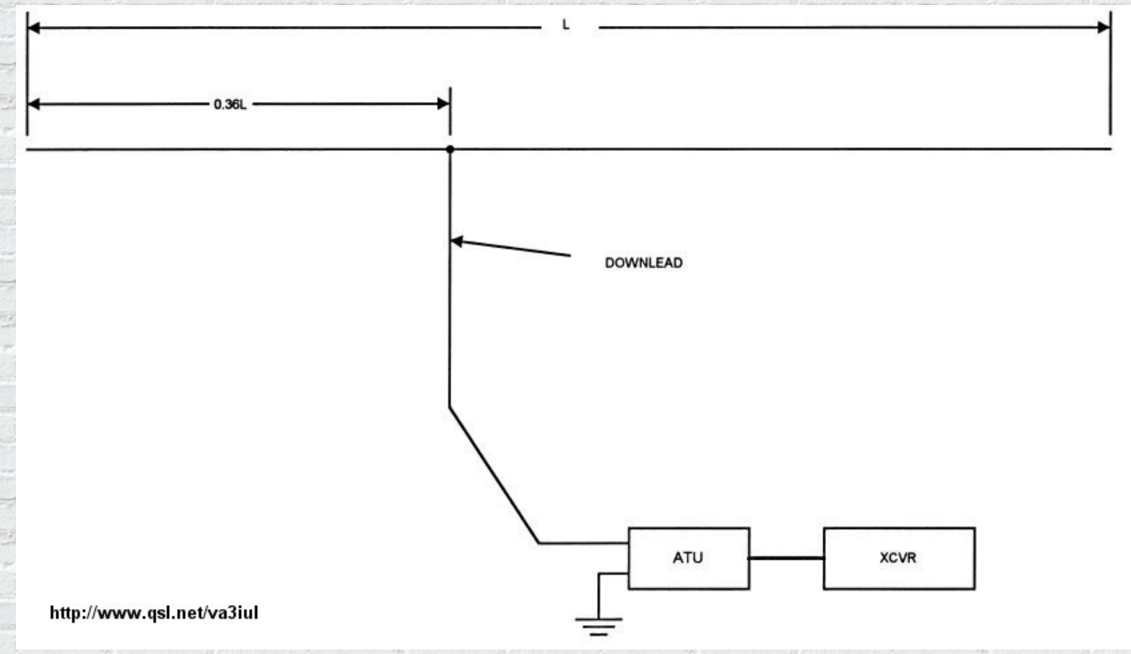
L
0.36L
DOWNLEAD
ATU
XCVR
http://www.qsl.net/va3iul

07 - Windom Antenna - Feed with coax cable

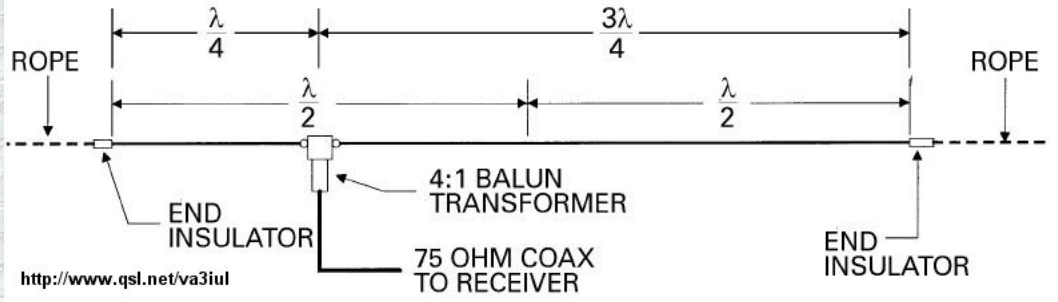
λ/4
3λ/4
ROPE
λ/2
λ/2
ROPE
4:1 BALUN TRANSFORMER
END INSULATOR
75 OHM COAX TO RECEIVER
END INSULATOR
http://www.qsl.net/va3iul

08 - Quarter Wavelength Vertical Antenna

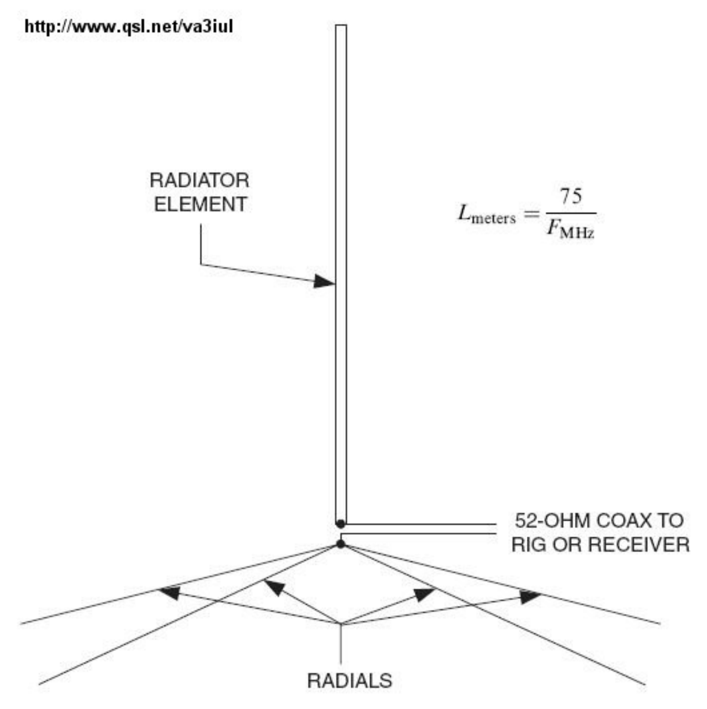
http://www.qsl.net/va3iul
RADIATOR ELEMENT
52-OHM COAX TO RIG OR RECEIVER
RADIALS

$$L_{\text{meters}} = \frac{75}{F_{\text{MHz}}}$$

09 - Folded Marconi Tee Antenna

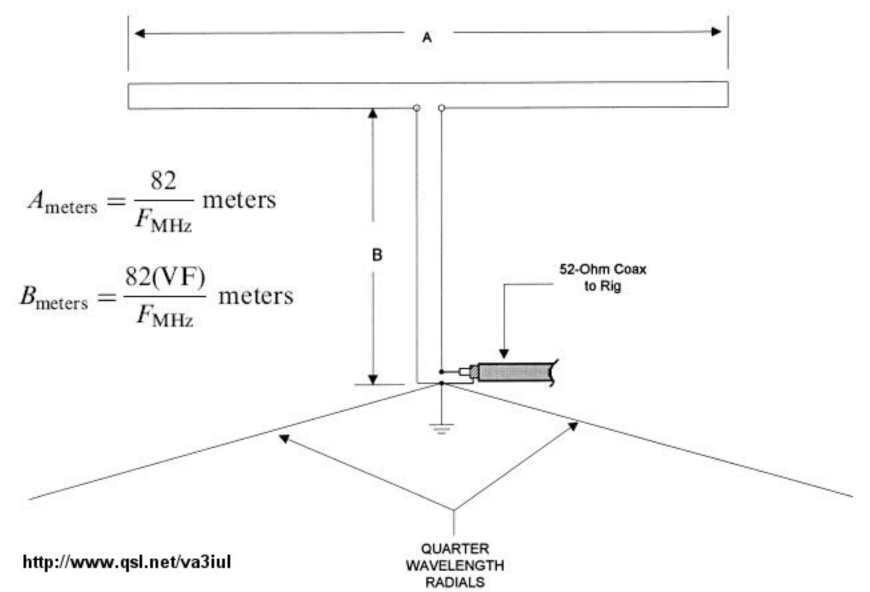
A
B
52-Ohm Coax to Rig
QUARTER WAVELENGTH RADIALS
http://www.qsl.net/va3iul

$$A_{\text{meters}} = \frac{82}{F_{\text{MHz}}} \text{ meters}$$

$$B_{\text{meters}} = \frac{82(\text{VF})}{F_{\text{MHz}}} \text{ meters}$$

10 - Zeppelin Antenna

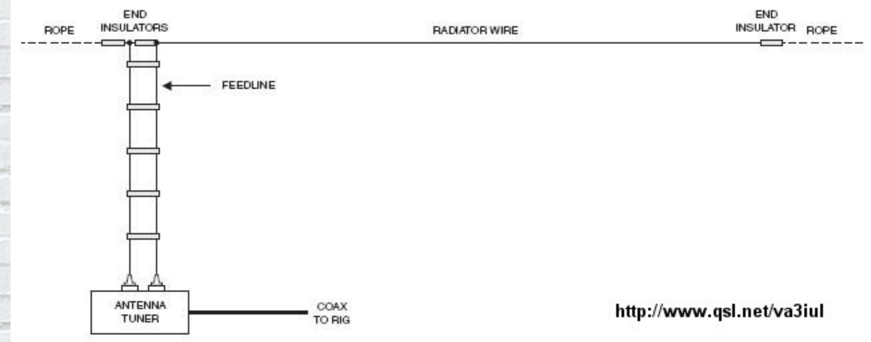
ROPE
END INSULATORS
RADIATOR WIRE
END INSULATOR
ROPE
FEEDLINE
ANTENNA TUNER
COAX TO RIG
http://www.qsl.net/va3iul

11 - EWE Antenna

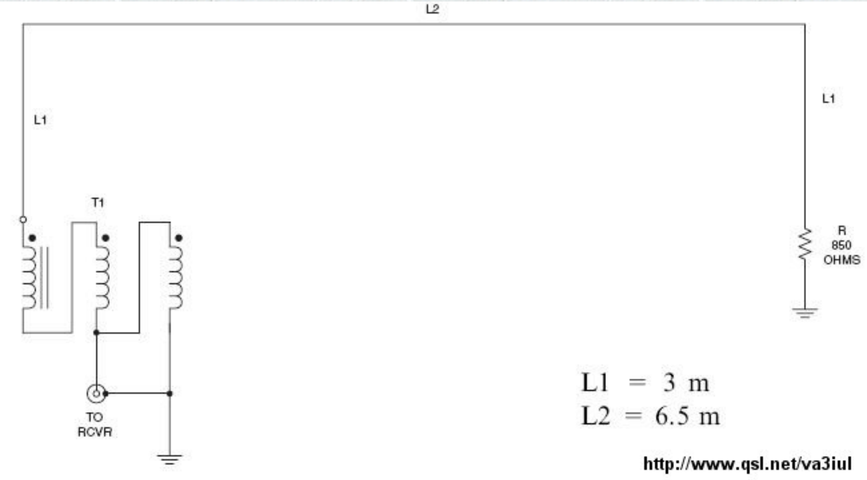
L2
L1
L1
T1
R 850 OHMS
TO RCVR
L1 = 3 m
L2 = 6.5 m
http://www.qsl.net/va3iul

12 - Dipole Antenna - Balun

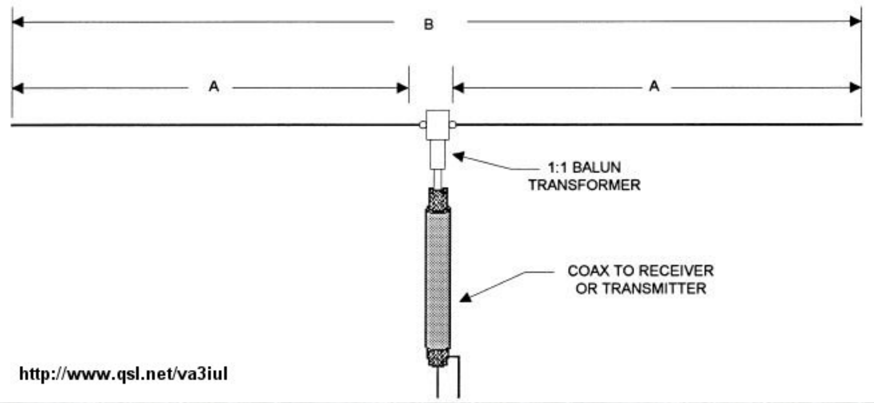
B
A
A
1:1 BALUN TRANSFORMER
COAX TO RECEIVER OR TRANSMITTER
http://www.qsl.net/va3iul

13 - Multiband Dipole Antenna

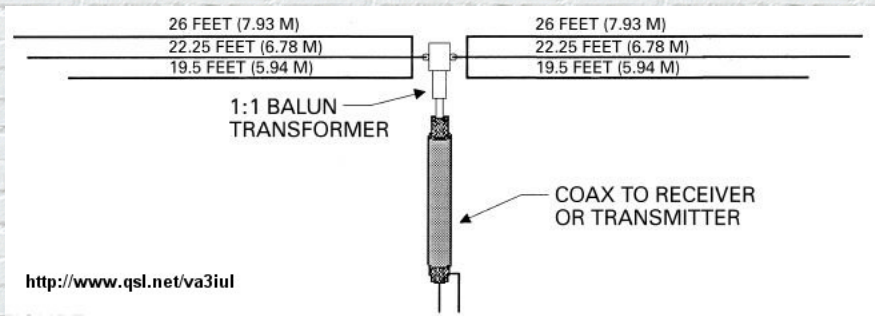
26 FEET (7.93 M)
22.25 FEET (6.78 M)
19.5 FEET (5.94 M)
26 FEET (7.93 M)
22.25 FEET (6.78 M)
19.5 FEET (5.94 M)
1:1 BALUN TRANSFORMER
COAX TO RECEIVER OR TRANSMITTER
http://www.qsl.net/va3iul

14 - Inverted-Vee Antenna

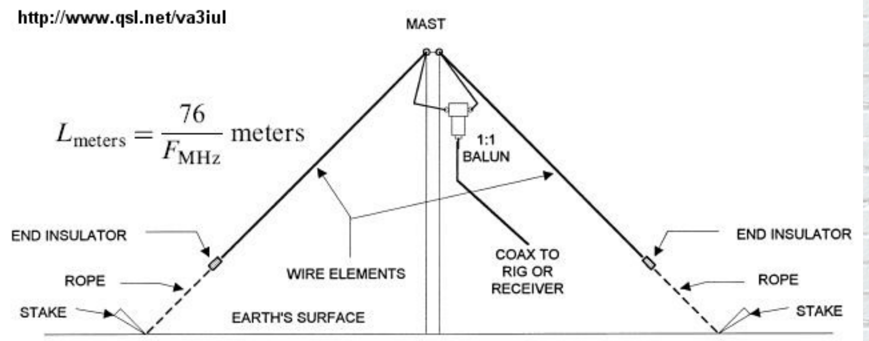
http://www.qsl.net/va3iul
MAST
$L_{\text{meters}} = \frac{76}{F_{\text{MHz}}}$ meters
1:1
BALUN
END INSULATOR
END INSULATOR
WIRE ELEMENTS
COAX TO
RIG OR
RECEIVER
ROPE
ROPE
STAKE
STAKE
EARTH'S SURFACE

15 - Sloping Dipole Antenna

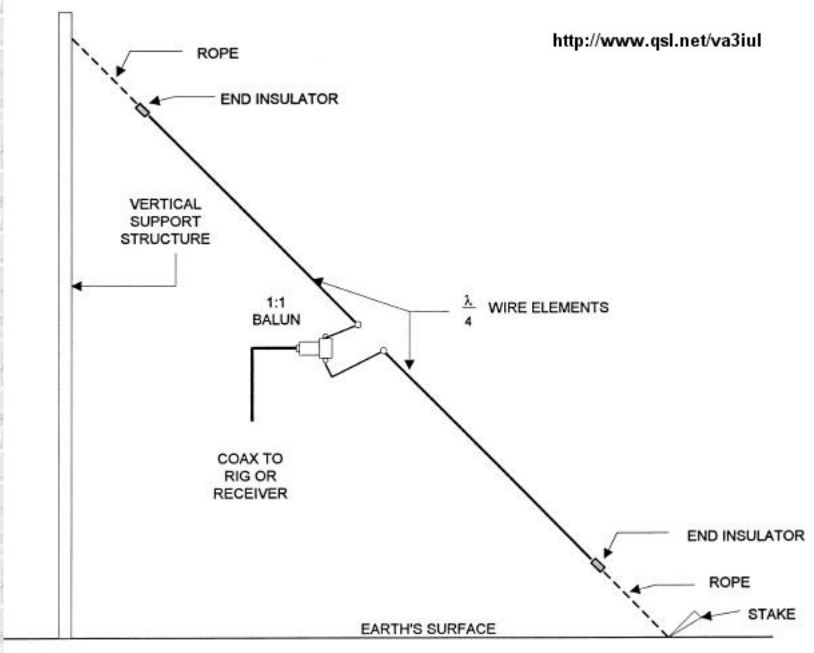
http://www.qsl.net/va3iul
ROPE
END INSULATOR
VERTICAL
SUPPORT
STRUCTURE
1:1
BALUN
$\frac{\lambda}{4}$ WIRE ELEMENTS
COAX TO
RIG OR
RECEIVER
END INSULATOR
ROPE
STAKE
EARTH'S SURFACE

16 - Vertical Dipole

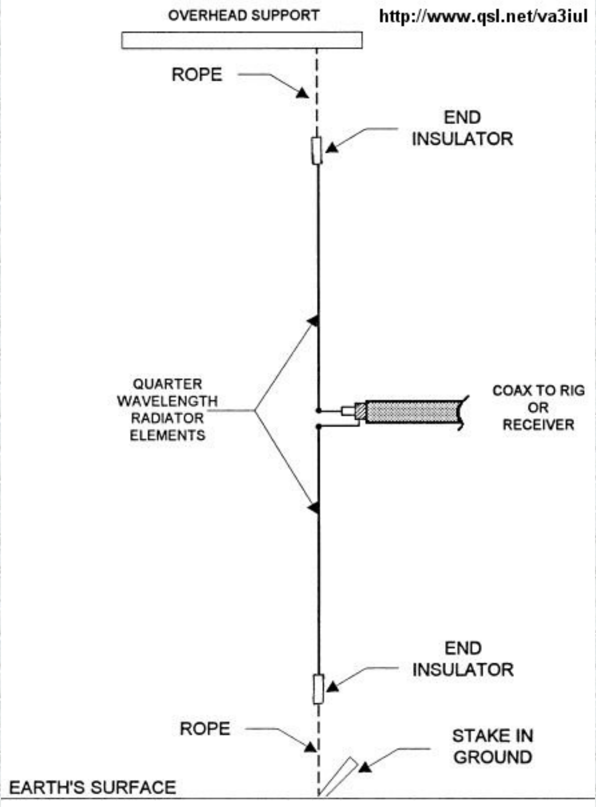
OVERHEAD SUPPORT
http://www.qsl.net/va3iul
ROPE
END INSULATOR
QUARTER WAVELENGTH RADIATOR ELEMENTS
COAX TO RIG OR RECEIVER
END INSULATOR
ROPE
STAKE IN GROUND
EARTH'S SURFACE

17 - Delta Fed Dipole Antenna

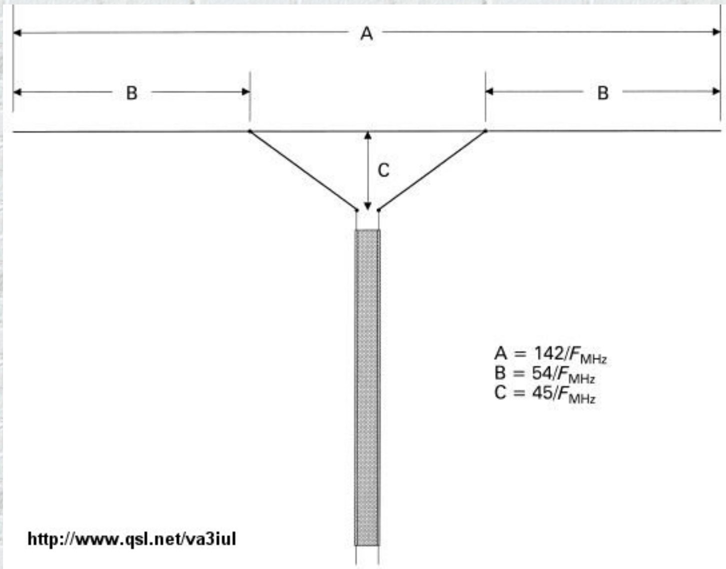
A
B
B
C
http://www.qsl.net/va3iul

$A = 142/F_{MHz}$

$B = 54/F_{MHz}$

$C = 45/F_{MHz}$

18 - Bow-Tie Dipole Antenna

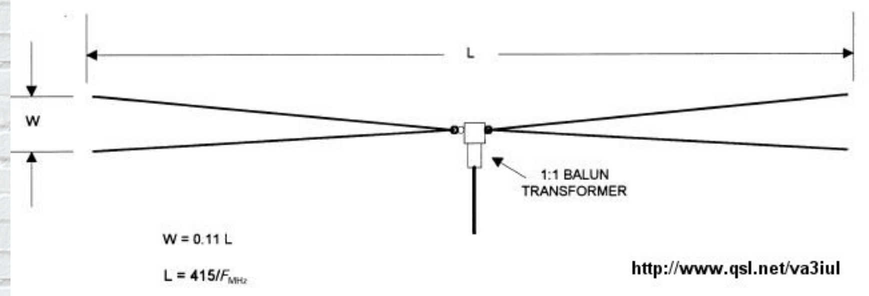
L
W
1:1 BALUN TRANSFORMER
http://www.qsl.net/va3iul

$W = 0.11\,L$

$L = 415/F_{MHz}$

19 - Bow-Tie Folded Dipole Antenna for RX

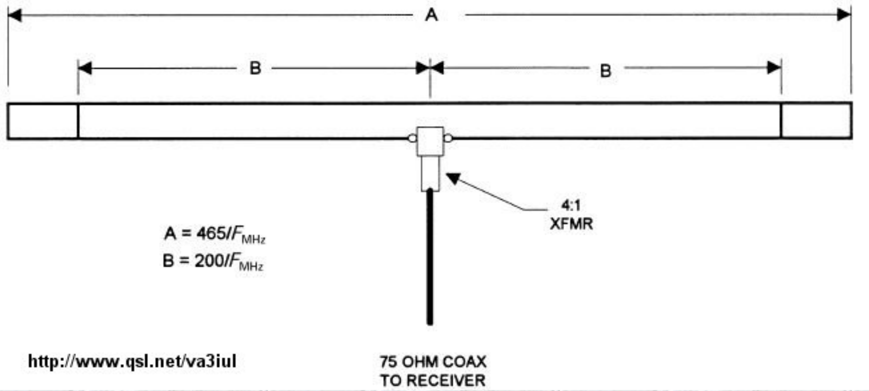
A
B
B
4:1
XFMR
A = 465/F_{MHz}
B = 200/F_{MHz}
http://www.qsl.net/va3iul
75 OHM COAX
TO RECEIVER

20 - Multiband Tuned Doublet Antenna

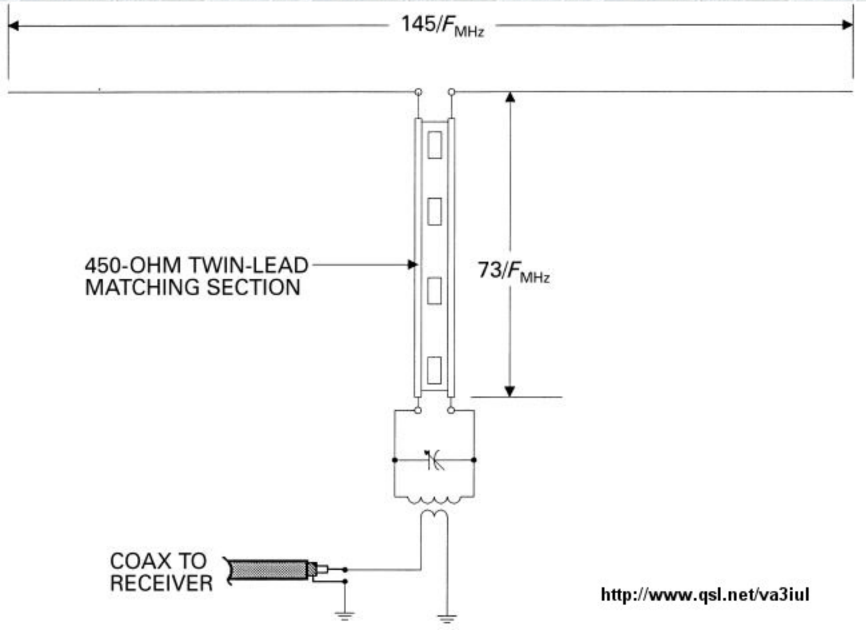
145/F_{MHz}
450-OHM TWIN-LEAD
MATCHING SECTION
73/F_{MHz}
COAX TO
RECEIVER
http://www.qsl.net/va3iul

21 - G5RV Antenna

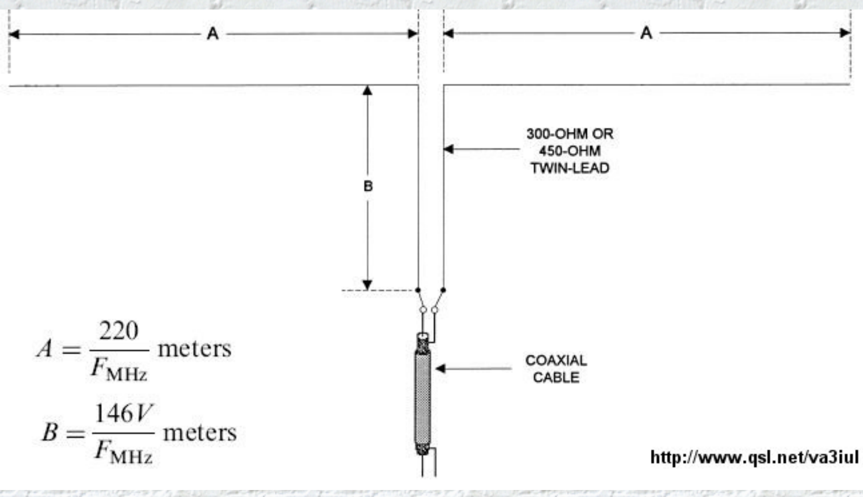
A
A
B
300-OHM OR
450-OHM
TWIN-LEAD
COAXIAL
CABLE
$A = \frac{220}{F_{\mathrm{MHz}}}$ meters
$B = \frac{146V}{F_{\mathrm{MHz}}}$ meters
http://www.qsl.net/va3iul

22 - Wideband Dipole Antenna

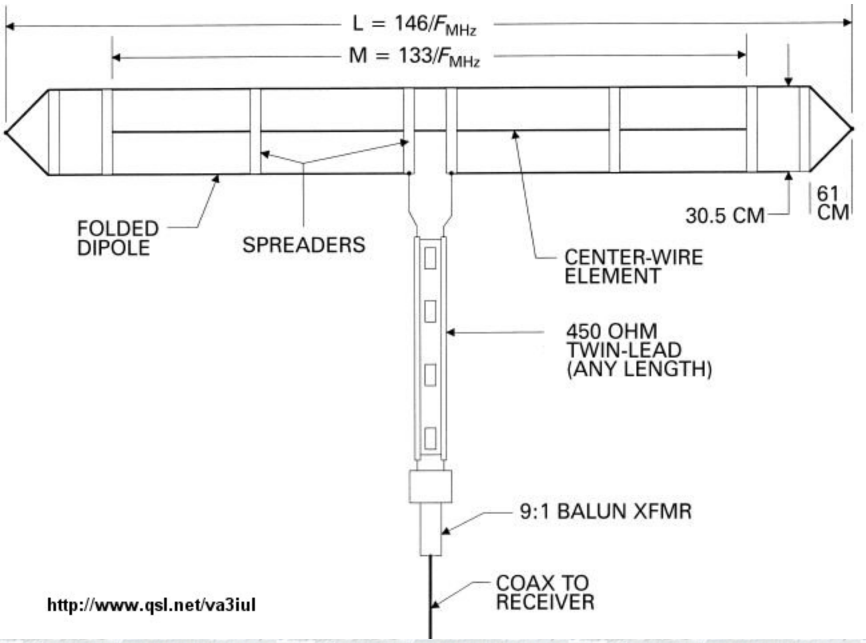
L = 146/F_{MHz}
M = 133/F_{MHz}
FOLDED DIPOLE
SPREADERS
30.5 CM
61 CM
CENTER-WIRE ELEMENT
450 OHM TWIN-LEAD (ANY LENGTH)
9:1 BALUN XFMR
COAX TO RECEIVER
http://www.qsl.net/va3iul

---

23 - Wideband Dipole for Receiving

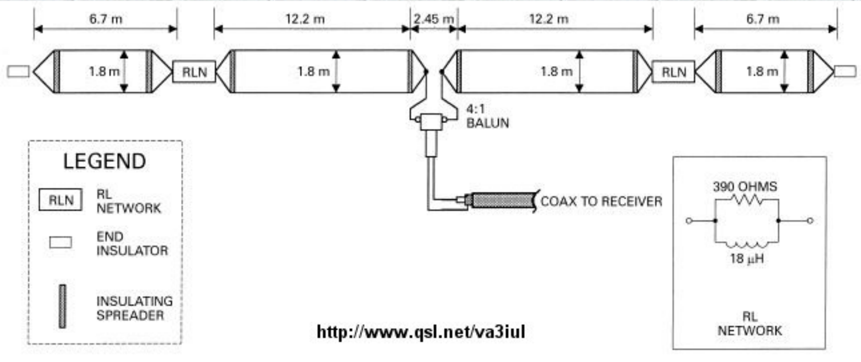
6.7 m
12.2 m
2.45 m
12.2 m
6.7 m
1.8 m
RLN
1.8 m
1.8 m
RLN
1.8 m
4:1 BALUN
LEGEND
RLN RL NETWORK
END INSULATOR
INSULATING SPREADER
COAX TO RECEIVER
390 OHMS
18 µH
RL NETWORK
http://www.qsl.net/va3iul

---

24 - Tilted Folded Dipole Antenna

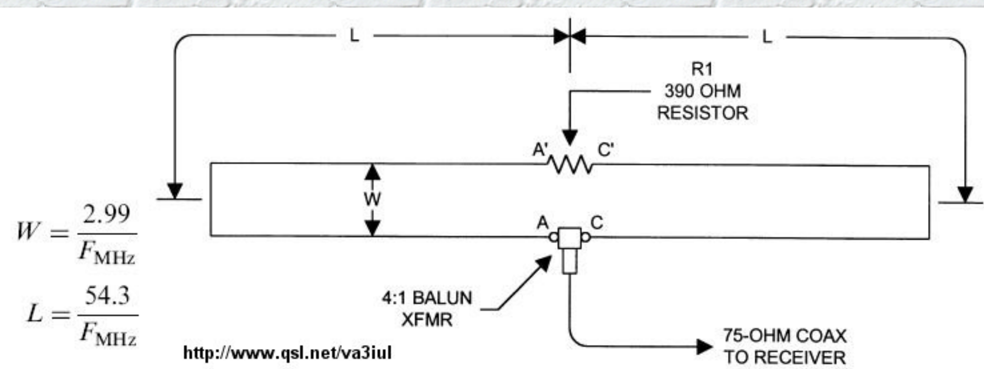
L
L
R1 390 OHM RESISTOR
A'
C'
W
A
C
4:1 BALUN XFMR
75-OHM COAX TO RECEIVER
$W = \frac{2.99}{F_{\mathrm{MHz}}}$
$L = \frac{54.3}{F_{\mathrm{MHz}}}$
http://www.qsl.net/va3iul

---

25 - Right Angle Marconi Antenna

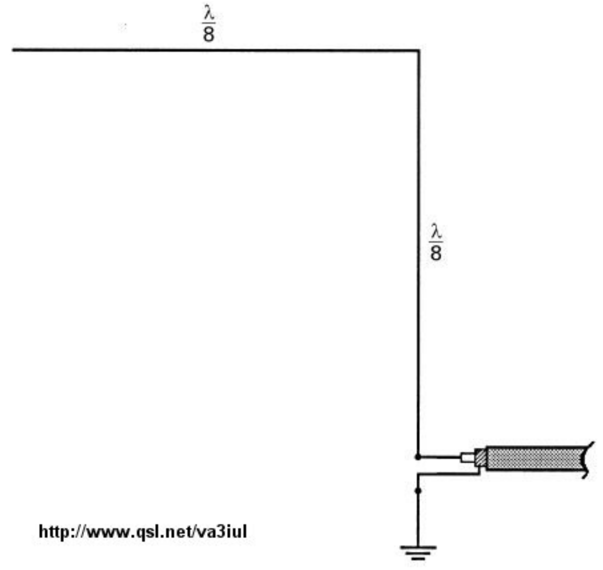
λ/8
λ/8
http://www.qsl.net/va3iul

26 - Linearly Loaded Tee Antenna

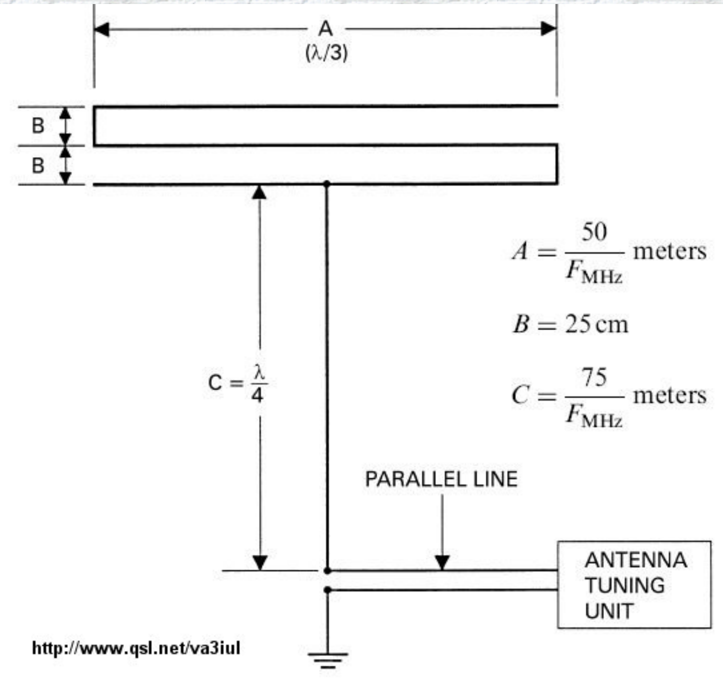
A
(λ/3)
B
B
C = λ/4
PARALLEL LINE
ANTENNA TUNING UNIT
http://www.qsl.net/va3iul

$$A = \frac{50}{F_{\mathrm{MHz}}}\ \text{meters}$$

$$B = 25\,\mathrm{cm}$$

$$C = \frac{75}{F_{\mathrm{MHz}}}\ \text{meters}$$

27 - Reduced Size Dipole Antenna

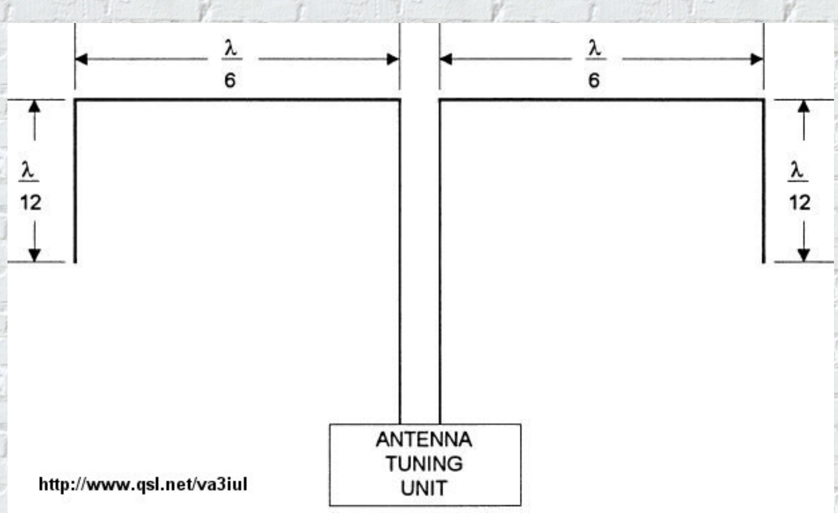
λ/6
λ/6
λ/12
λ/12
ANTENNA TUNING UNIT
http://www.qsl.net/va3iul

28 - Doublet Dipole Antenna

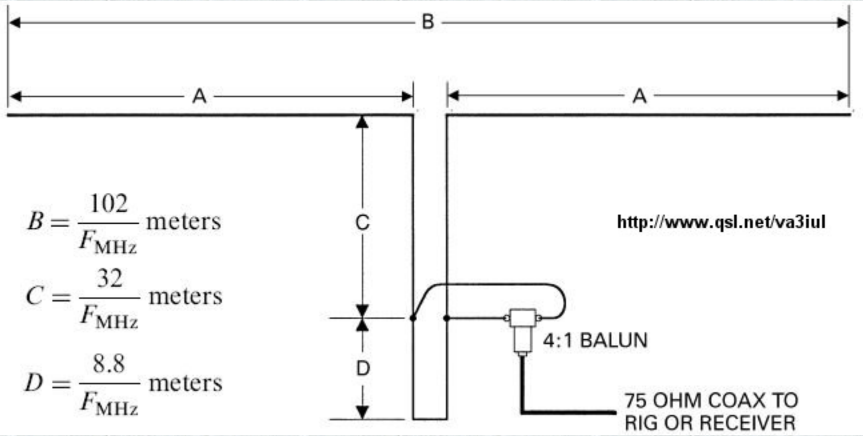
B
A
A
$B = \frac{102}{F_{\text{MHz}}}$ meters
$C = \frac{32}{F_{\text{MHz}}}$ meters
$D = \frac{8.8}{F_{\text{MHz}}}$ meters
C
D
http://www.qsl.net/va3iul
4:1 BALUN
75 OHM COAX TO
RIG OR RECEIVER

---

29 - Delta Loop Antenna

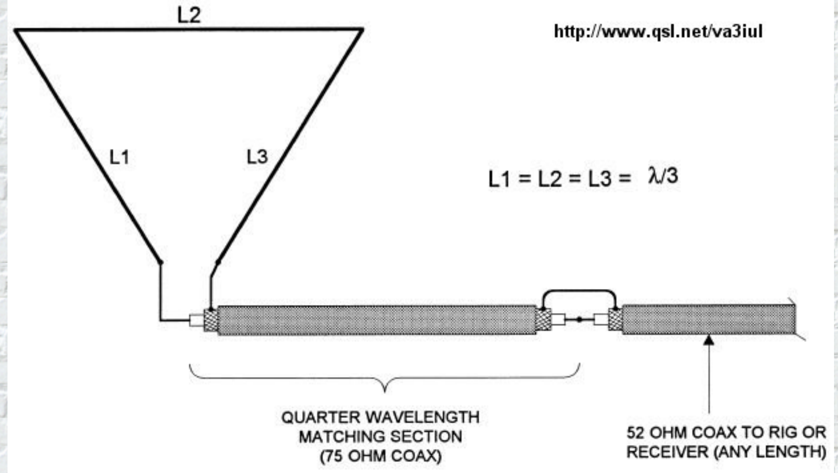
L2
L1
L3
http://www.qsl.net/va3iul
L1 = L2 = L3 = λ/3
QUARTER WAVELENGTH
MATCHING SECTION
(75 OHM COAX)
52 OHM COAX TO RIG OR
RECEIVER (ANY LENGTH)

---

30 - Half Delta Loop Antenna

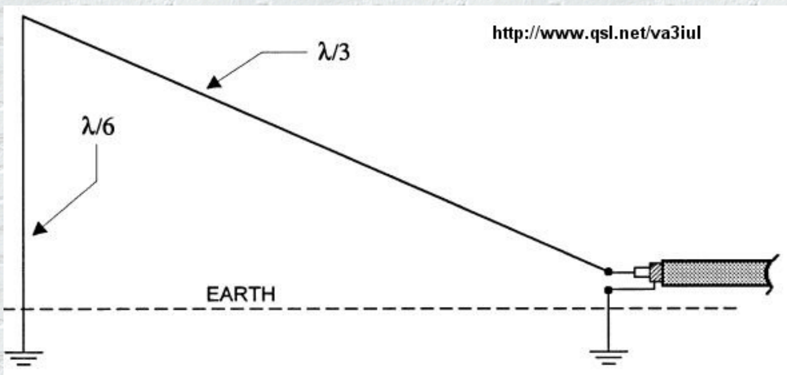
λ/3
http://www.qsl.net/va3iul
λ/6
EARTH

---

31 - Collinear Franklin Antenna

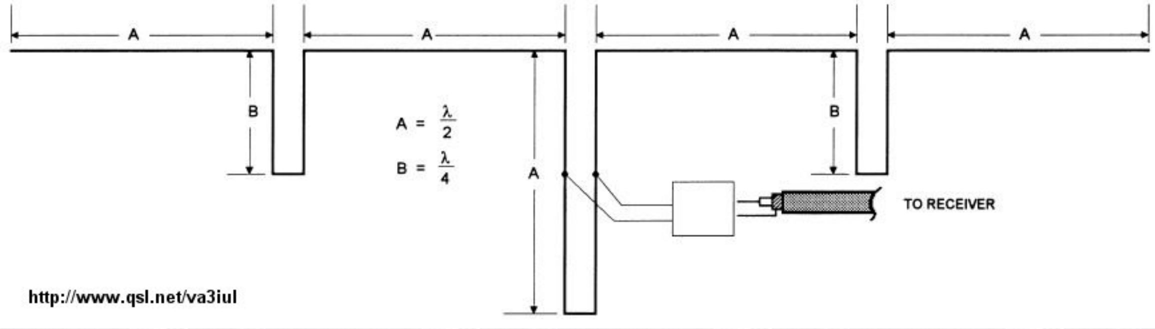
A
A
A
A
B
B
A = λ/2
B = λ/4
A
TO RECEIVER
http://www.qsl.net/va3iul

32 - Four Element Broadside Antenna

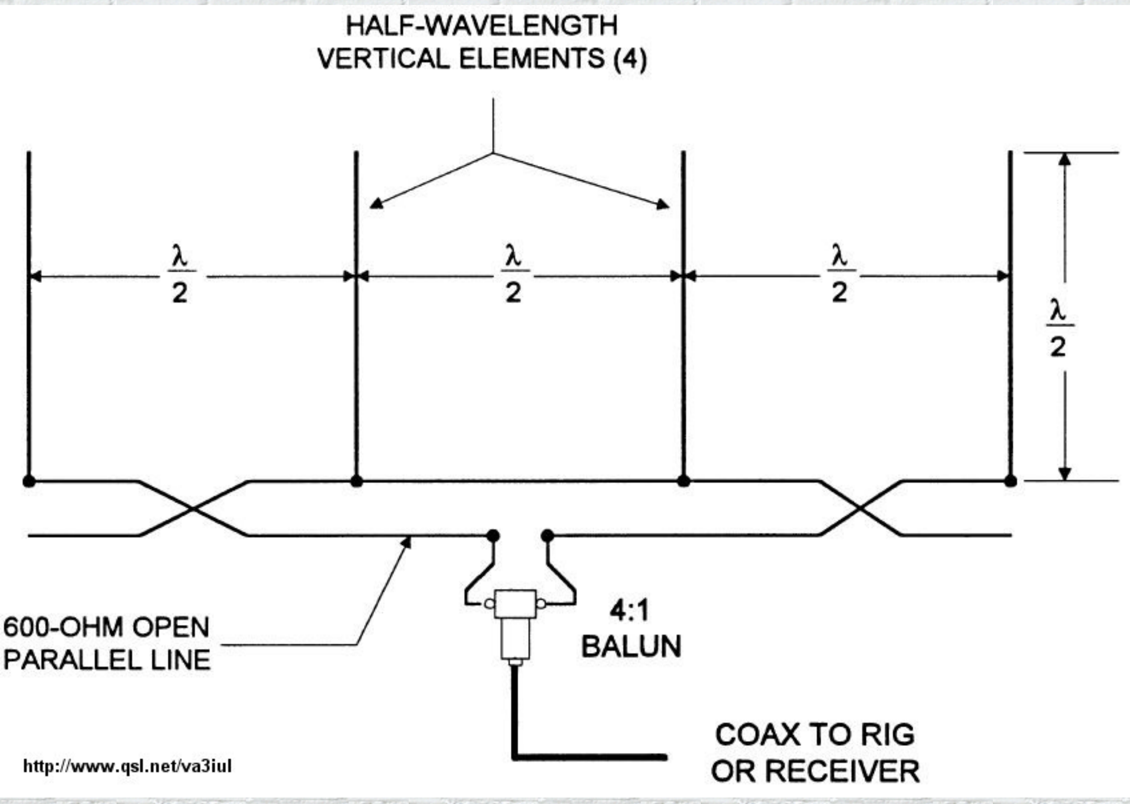
HALF-WAVELENGTH
VERTICAL ELEMENTS (4)
λ/2
λ/2
λ/2
λ/2
600-OHM OPEN
PARALLEL LINE
4:1
BALUN
COAX TO RIG
OR RECEIVER
http://www.qsl.net/va3iul

33 - The Lazy-H Array Antenna

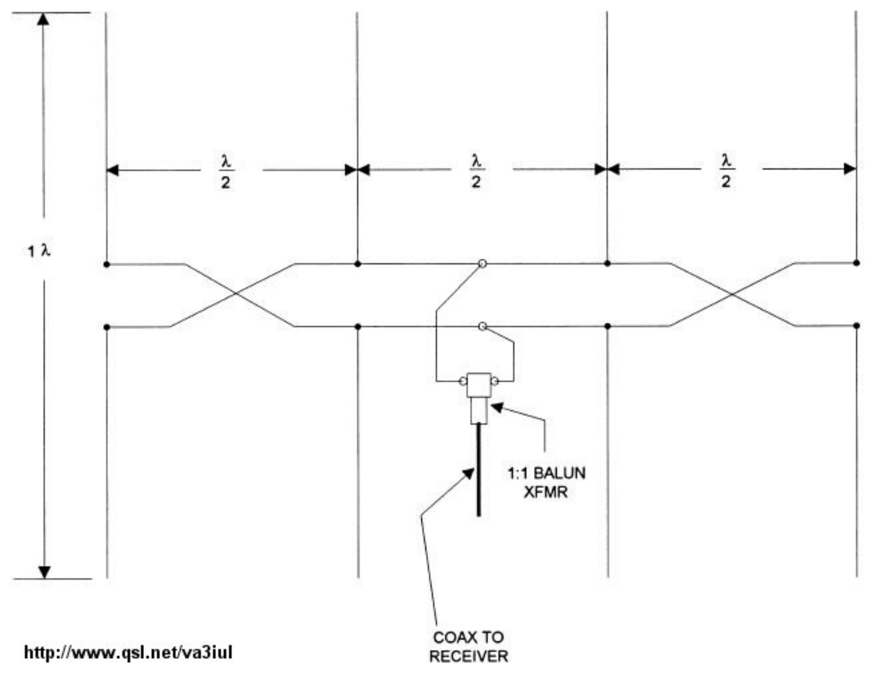
λ/2
λ/2
λ/2
1 λ
1:1 BALUN
XFMR
COAX TO
RECEIVER
http://www.qsl.net/va3iul

34 - Sterba Curtain Array Antenna

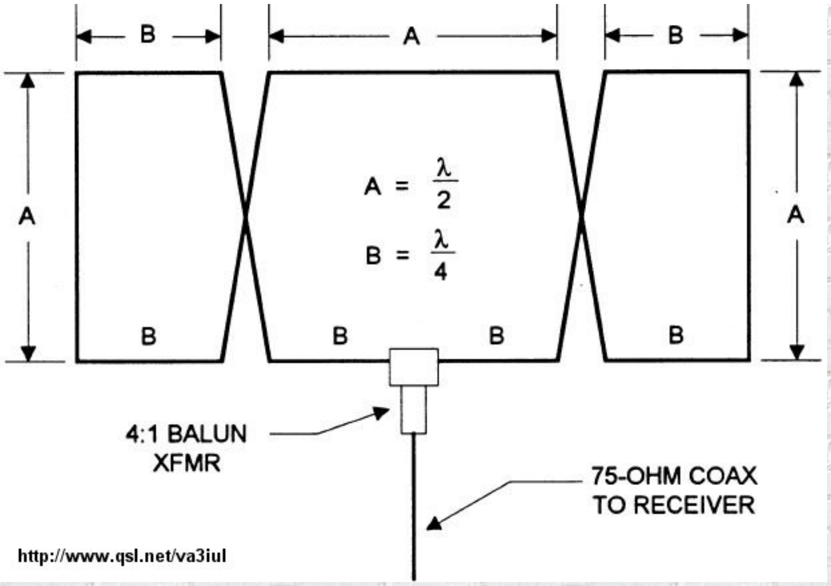
B
A
B
A
A
$A = \frac{\lambda}{2}$
$B = \frac{\lambda}{4}$
B
B
B
B
4:1 BALUN XFMR
75-OHM COAX TO RECEIVER
http://www.qsl.net/va3iul

---

35 - T-L DX Antenna

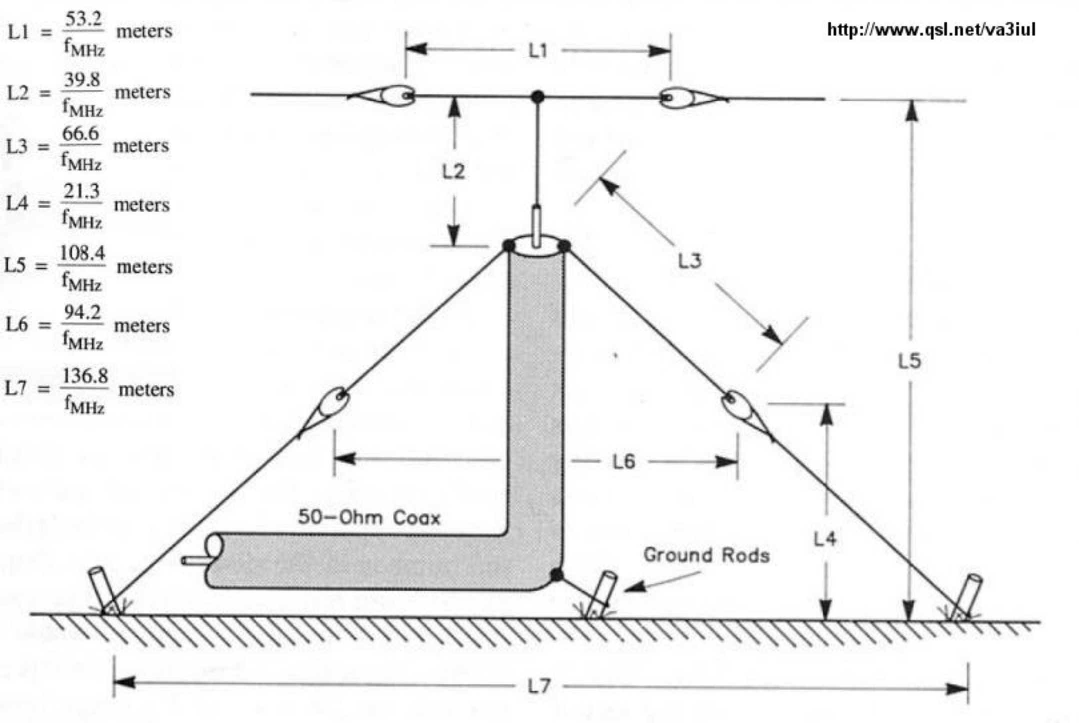
$L1 = \frac{53.2}{f_{MHz}}$ meters
$L2 = \frac{39.8}{f_{MHz}}$ meters
$L3 = \frac{66.6}{f_{MHz}}$ meters
$L4 = \frac{21.3}{f_{MHz}}$ meters
$L5 = \frac{108.4}{f_{MHz}}$ meters
$L6 = \frac{94.2}{f_{MHz}}$ meters
$L7 = \frac{136.8}{f_{MHz}}$ meters
http://www.qsl.net/va3iul
L1
L2
L3
L5
L6
L4
50-Ohm Coax
Ground Rods
L7

---

36 - 1.9 MHz Full-wave Loop Antenna

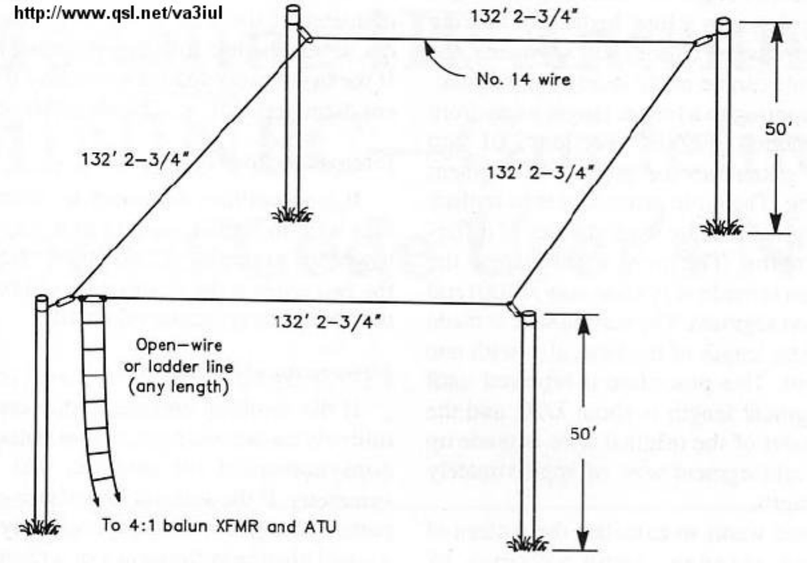
http://www.qsl.net/va3iul
132′ 2–3/4″
No. 14 wire
50′
132′ 2–3/4″
132′ 2–3/4″
132′ 2–3/4″
Open–wire
or ladder line
(any length)
50′
To 4:1 balun XFMR and ATU

---

37 - Multi-Band Portable Antenna

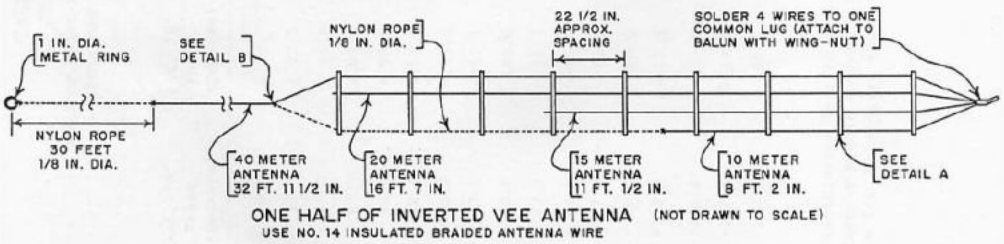
1 IN. DIA. METAL RING
SEE DETAIL B
NYLON ROPE 1/8 IN. DIA.
22 1/2 IN. APPROX. SPACING
SOLDER 4 WIRES TO ONE COMMON LUG (ATTACH TO BALUN WITH WING-NUT)
NYLON ROPE 30 FEET 1/8 IN. DIA.
40 METER ANTENNA 32 FT. 11 1/2 IN.
20 METER ANTENNA 16 FT. 7 IN.
15 METER ANTENNA 11 FT. 1/2 IN.
10 METER ANTENNA 8 FT. 2 IN.
SEE DETAIL A
ONE HALF OF INVERTED VEE ANTENNA
(NOT DRAWN TO SCALE)
USE NO. 14 INSULATED BRAIDED ANTENNA WIRE

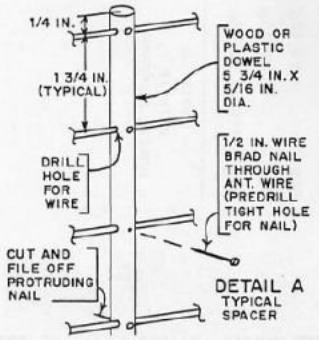
1/4 IN.
1 3/4 IN. (TYPICAL)
WOOD OR PLASTIC DOWEL 5 3/4 IN. X 5/16 IN. DIA.
DRILL HOLE FOR WIRE
1/2 IN. WIRE BRAD NAIL THROUGH ANT. WIRE (PREDRILL TIGHT HOLE FOR NAIL)
CUT AND FILE OFF PROTRUDING NAIL
DETAIL A
TYPICAL SPACER

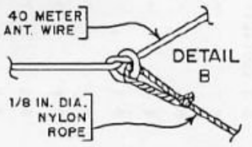
40 METER ANT. WIRE
DETAIL B
1/8 IN. DIA. NYLON ROPE

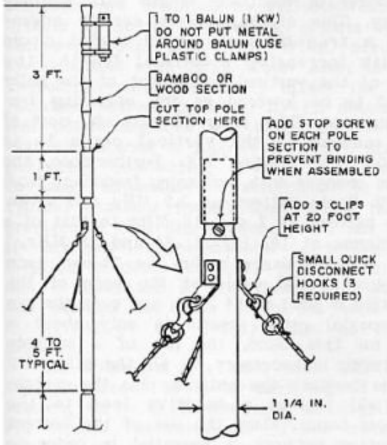
1 TO 1 BALUN (1 KW) DO NOT PUT METAL AROUND BALUN (USE PLASTIC CLAMPS)
3 FT.
BAMBOO OR WOOD SECTION
CUT TOP POLE SECTION HERE
ADD STOP SCREW ON EACH POLE SECTION TO PREVENT BINDING WHEN ASSEMBLED
1 FT.
ADD 3 CLIPS AT 20 FOOT HEIGHT
SMALL QUICK DISCONNECT HOOKS (3 REQUIRED)
4 TO 5 FT. TYPICAL
1 1/4 IN. DIA.

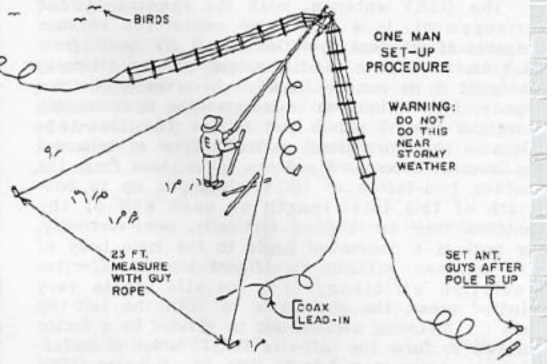
BIRDS
ONE MAN SET-UP PROCEDURE
WARNING: DO NOT DO THIS NEAR STORMY WEATHER
23 FT. MEASURE WITH GUY ROPE
SET ANT. GUYS AFTER POLE IS UP
COAX LEAD-IN

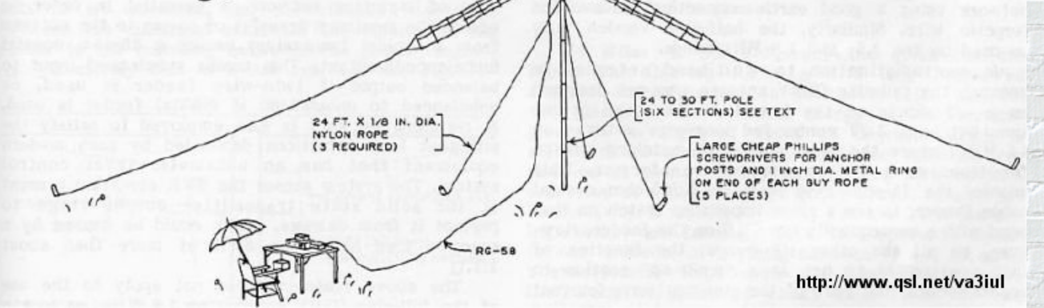
24 TO 30 FT. POLE (SIX SECTIONS) SEE TEXT
24 FT. X 1/8 IN. DIA. NYLON ROPE (3 REQUIRED)
LARGE CHEAP PHILLIPS SCREWDRIVERS FOR ANCHOR POSTS AND 1 INCH DIA. METAL RING ON END OF EACH GUY ROPE (3 PLACES)
RG-58
http://www.qsl.net/va3iul

38 - Off-center-fed Full-wave Doublet Antenna

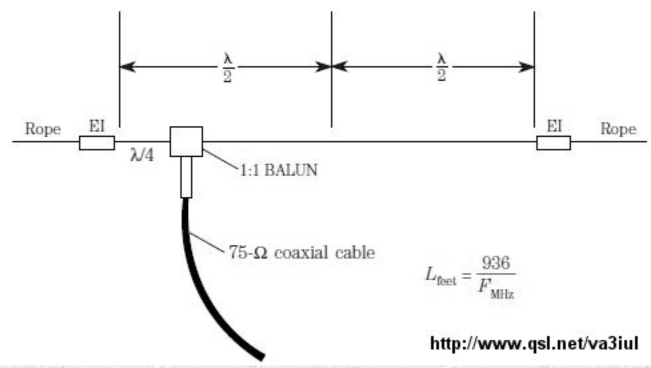
λ/2
λ/2
Rope
EI
EI
Rope
λ/4
1:1 BALUN
75-Ω coaxial cable
$L_{\text{feet}} = \frac{936}{F_{\text{MHz}}}$
http://www.qsl.net/va3iul

39 - Terminated Sloper Antenna

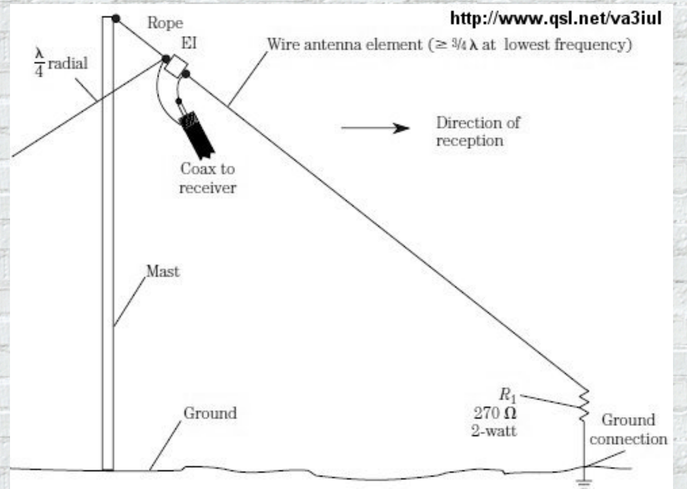
http://www.qsl.net/va3iul
Rope
EI
λ/4 radial
Wire antenna element (≥ 3/4 λ at lowest frequency)
Direction of reception
Coax to receiver
Mast
Ground
R1 270 Ω 2-watt
Ground connection

---

40 - Double Extended Zepp Antenna

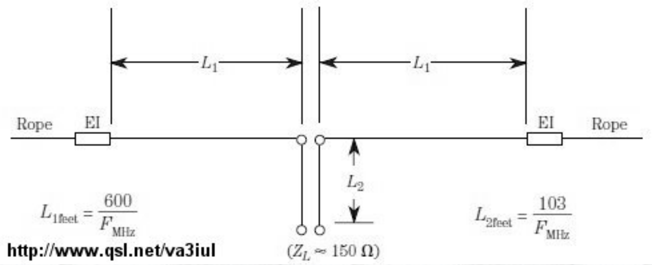
L1
L1
Rope
EI
EI
Rope
L2
L1feet = 600/F MHz
L2feet = 103/F MHz
http://www.qsl.net/va3iul
(ZL ≈ 150 Ω)

---

41 - TCFTFD Dipole Antenna

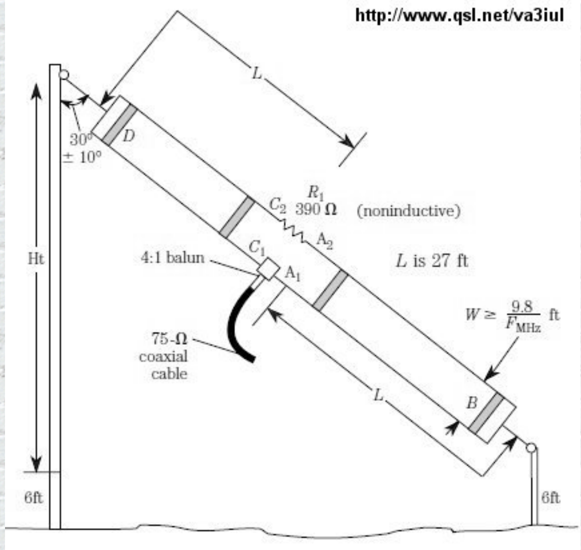
http://www.qsl.net/va3iul
L
30° ± 10°
D
C2
R1 390 Ω
(noninductive)
A2
C1
A1
4:1 balun
L is 27 ft
Ht
75-Ω coaxial cable
W ≥ 9.8/F MHz ft
L
B
6ft
6ft

---

42 - Vee-Sloper Antenna

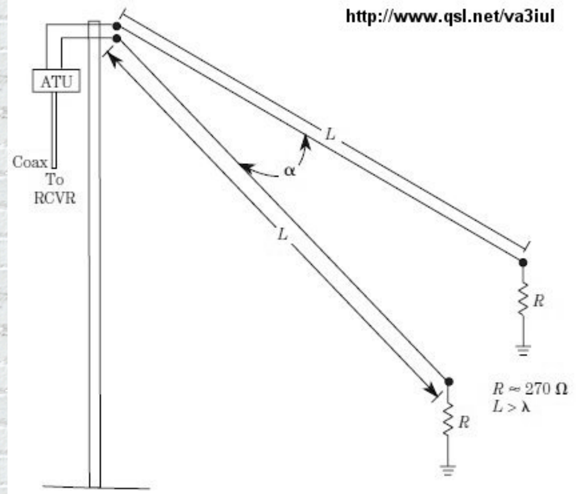
http://www.qsl.net/va3iul
ATU
Coax
To
RCVR
L
α
L
R
R
R ≈ 270 Ω
L > λ

43 - Rhombic Inverted-Vee Antenna

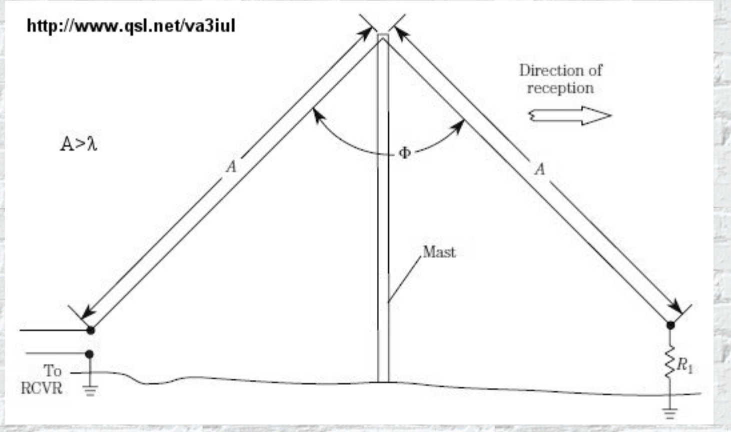
http://www.qsl.net/va3iul
A>λ
A
Φ
A
Direction of reception
Mast
To
RCVR
R1

44 - Counterpoise Longwire

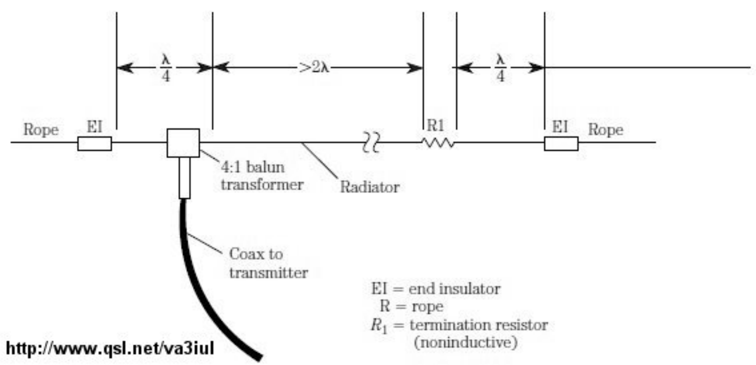
λ/4
>2λ
λ/4
Rope
EI
R1
EI
Rope
4:1 balun transformer
Radiator
Coax to transmitter
EI = end insulator
R = rope
R1 = termination resistor (noninductive)
http://www.qsl.net/va3iul

45 - Bisquare Loop Antenna

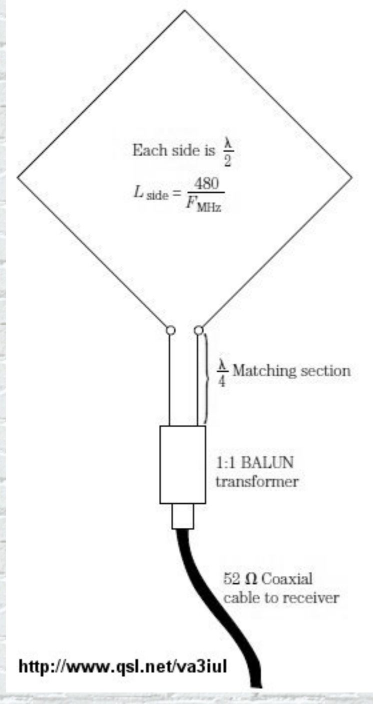
Each side is $\frac{\lambda}{2}$
$L_{side} = \frac{480}{F_{MHz}}$
$\frac{\lambda}{4}$ Matching section
1:1 BALUN transformer
52 Ω Coaxial cable to receiver
http://www.qsl.net/va3iul

46 - Piggyback Antenna for 10m

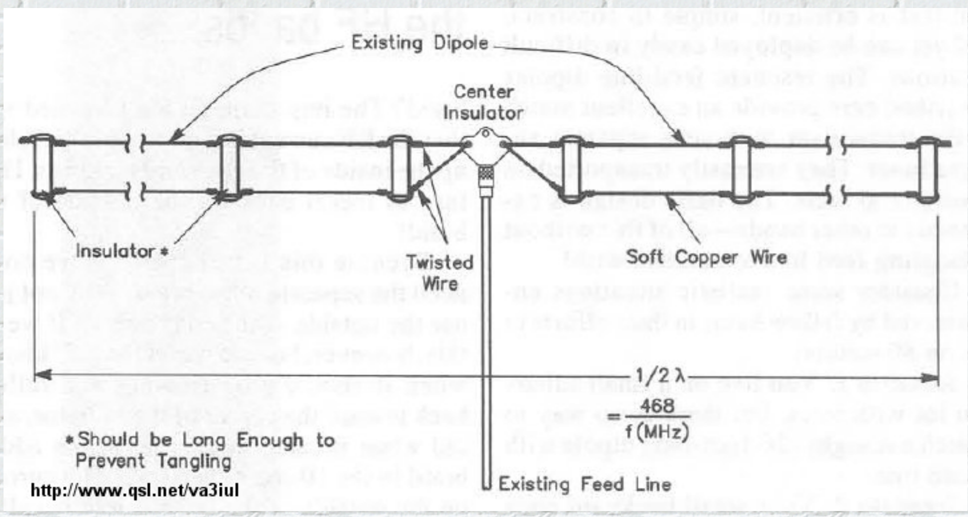
Existing Dipole
Center Insulator
Insulator *
Twisted Wire
Soft Copper Wire
$1/2\ \lambda = \frac{468}{f\,(MHz)}$
* Should be Long Enough to Prevent Tangling
Existing Feed Line
http://www.qsl.net/va3iul

47 - Vertical Sleeve Antenna for 10m

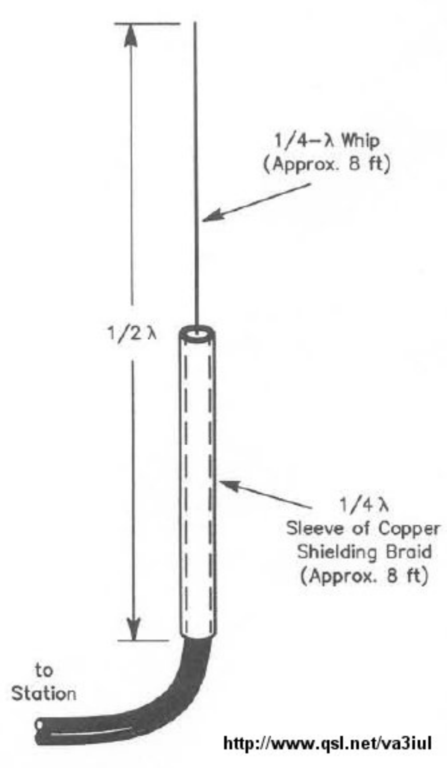
1/4−λ Whip
(Approx. 8 ft)
1/2 λ
1/4 λ
Sleeve of Copper
Shielding Braid
(Approx. 8 ft)
to
Station
http://www.qsl.net/va3iul

48 - Double Windom Antenna

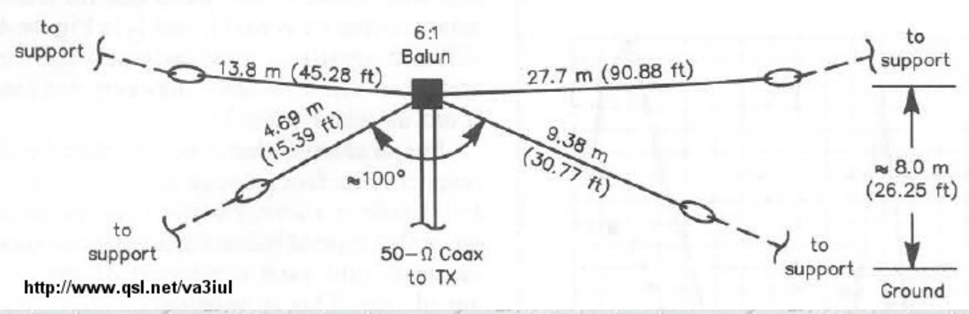
to
support
13.8 m (45.28 ft)
6:1
Balun
27.7 m (90.88 ft)
to
support
4.69 m
(15.39 ft)
≈100°
9.38 m
(30.77 ft)
≈ 8.0 m
(26.25 ft)
to
support
50−Ω Coax
to TX
to
support
Ground
http://www.qsl.net/va3iul

49 - Double Windom for 9 Bands

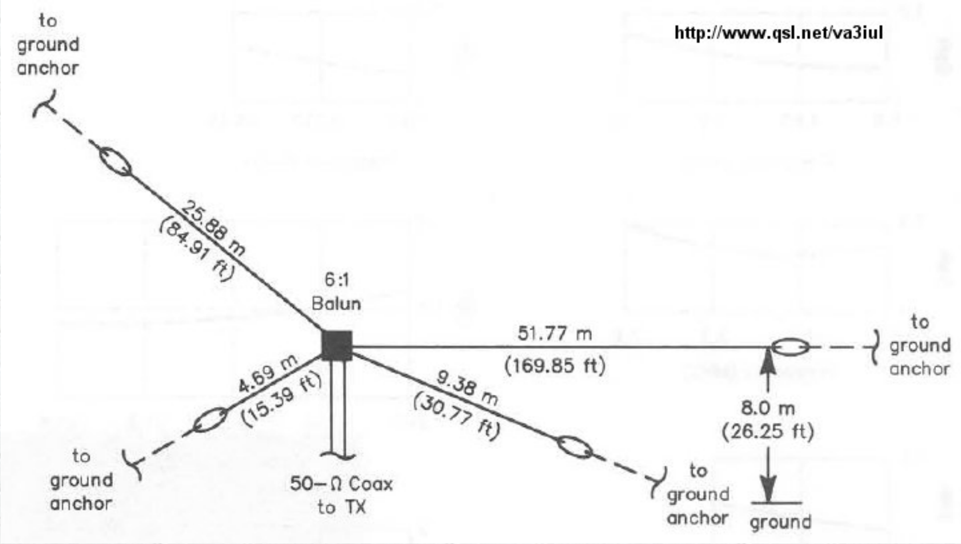
to
ground
anchor
http://www.qsl.net/va3iul
25.88 m
(84.91 ft)
6:1
Balun
51.77 m
(169.85 ft)
to
ground
anchor
4.69 m
(15.39 ft)
9.38 m
(30.77 ft)
8.0 m
(26.25 ft)
to
ground
anchor
50−Ω Coax
to TX
to
ground
anchor
ground

50 - Collinear Trap Antenna

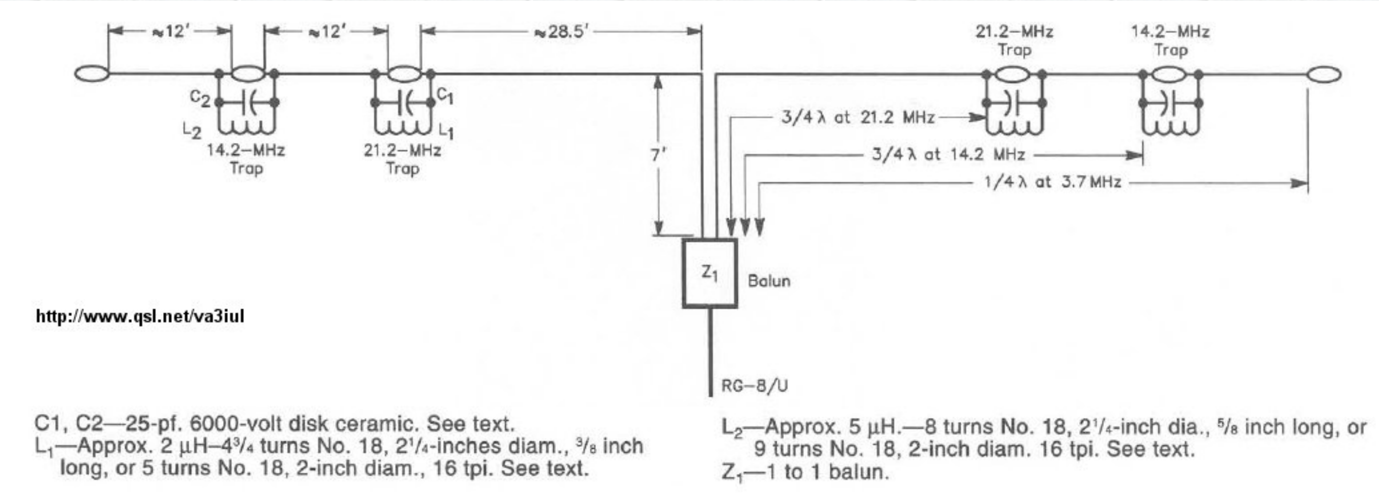
≈12'
≈12'
≈28.5'
C2
L2
14.2–MHz Trap
C1
L1
21.2–MHz Trap
7'
21.2–MHz Trap
14.2–MHz Trap
3/4 λ at 21.2 MHz
3/4 λ at 14.2 MHz
1/4 λ at 3.7 MHz
Z1
Balun
RG–8/U
http://www.qsl.net/va3iul

C1, C2—25-pf. 6000-volt disk ceramic. See text.
L_1—Approx. 2 µH—4¾ turns No. 18, 2¼-inches diam., ⅜ inch long, or 5 turns No. 18, 2-inch diam., 16 tpi. See text.

L_2—Approx. 5 µH.—8 turns No. 18, 2¼-inch dia., ⅝ inch long, or 9 turns No. 18, 2-inch diam. 16 tpi. See text.
Z_1—1 to 1 balun.

51 - Short Dipole Antenna for 40m - 80m - 160m

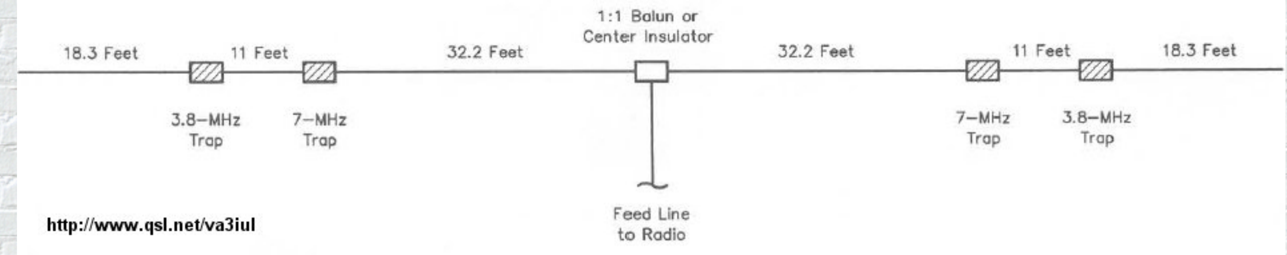
1:1 Balun or Center Insulator
18.3 Feet
11 Feet
32.2 Feet
32.2 Feet
11 Feet
18.3 Feet
3.8–MHz Trap
7–MHz Trap
7–MHz Trap
3.8–MHz Trap
Feed Line to Radio
http://www.qsl.net/va3iul

52 - Center Fed-Zepp Antenna for 80m - 40m

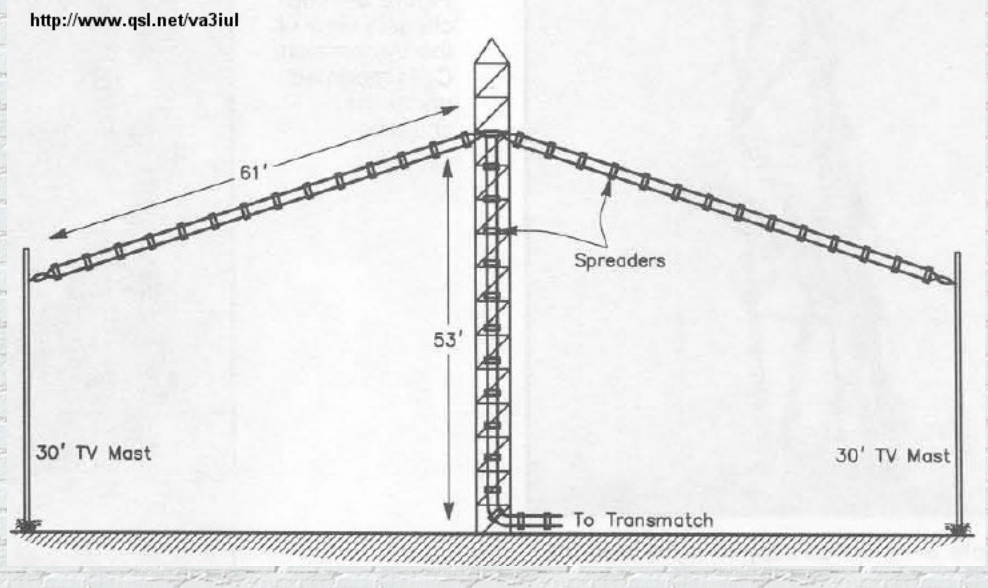
http://www.qsl.net/va3iul
61'
Spreaders
53'
30' TV Mast
30' TV Mast
To Transmatch

53 - All-Bands Antenna

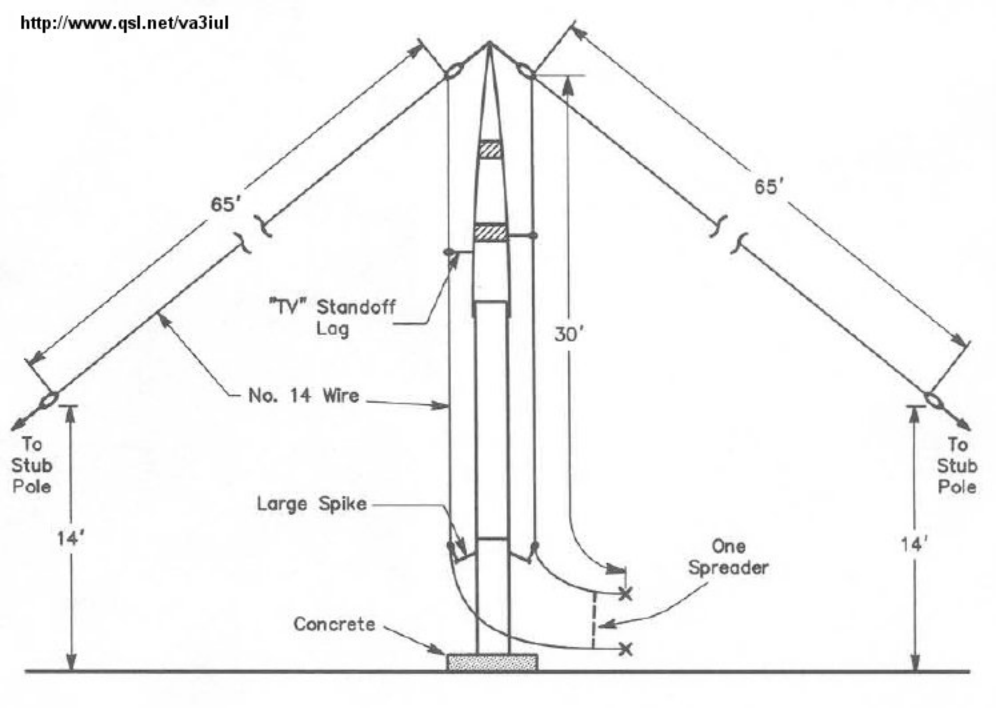
http://www.qsl.net/va3iul
65'
65'
"TV" Standoff Lag
30'
No. 14 Wire
To Stub Pole
To Stub Pole
14'
14'
Large Spike
One Spreader
Concrete

---

54 - All-Bands Dipole Antenna

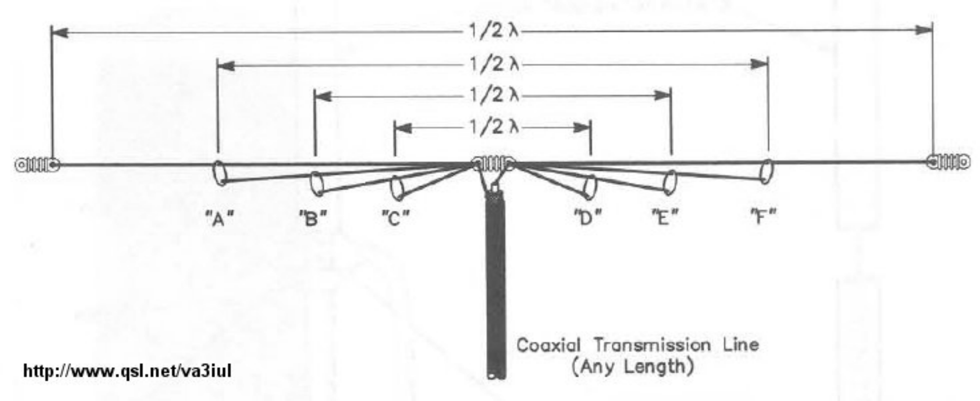
1/2 λ
1/2 λ
1/2 λ
1/2 λ
"A"
"B"
"C"
"D"
"E"
"F"
Coaxial Transmission Line (Any Length)
http://www.qsl.net/va3iul

---

55 - Multiband Z Antenna

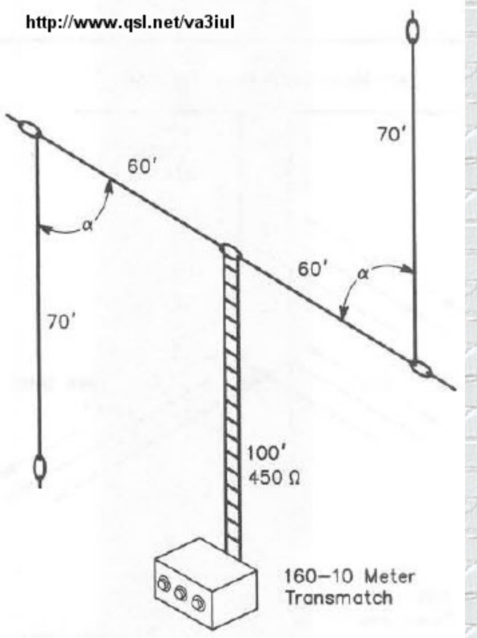
http://www.qsl.net/va3iul
70'
60'
α
60'
α
70'
100'
450 Ω
160–10 Meter
Transmatch

---

56 - Multiband Dipole Antenna

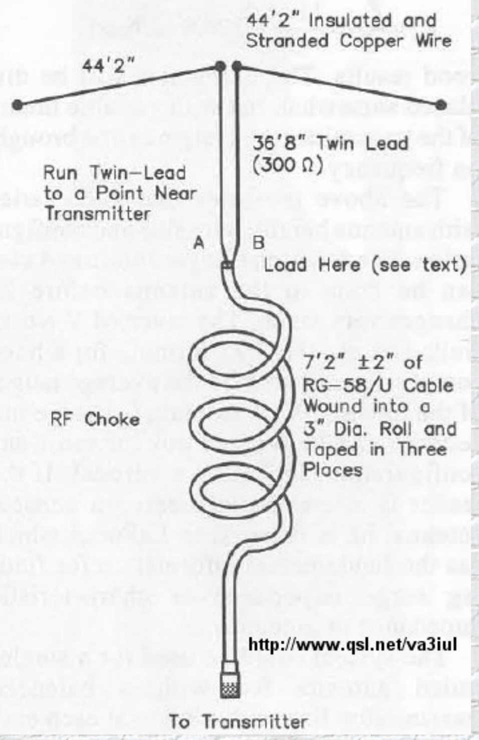
44'2" Insulated and
Stranded Copper Wire
44'2"
36'8" Twin Lead
(300 Ω)
Run Twin–Lead
to a Point Near
Transmitter
A
B
Load Here (see text)
7'2" ±2" of
RG–58/U Cable
Wound into a
3" Dia. Roll and
Taped in Three
Places
RF Choke
http://www.qsl.net/va3iul
To Transmitter

---

57 - Five-Bands No-Tuner Antenna

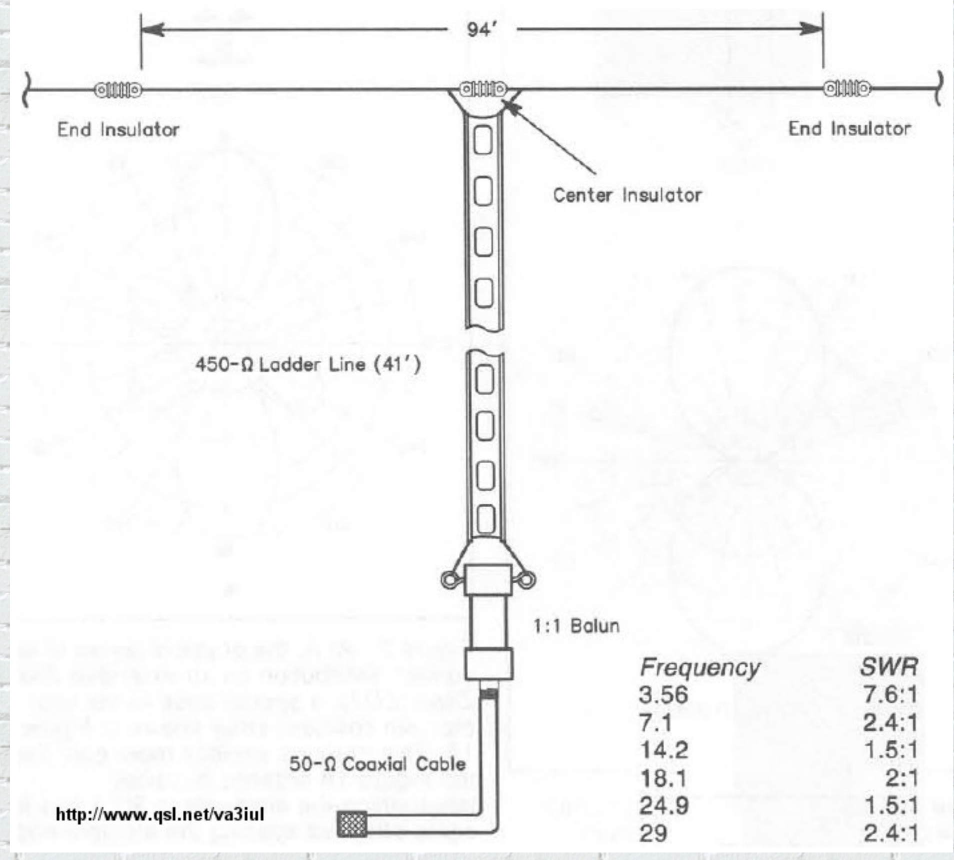
94'
End Insulator
End Insulator
Center Insulator
450-Ω Ladder Line (41')
1:1 Balun
50-Ω Coaxial Cable
http://www.qsl.net/va3iul

Frequency	*SWR*
3.56	7.6:1
7.1	2.4:1
14.2	1.5:1
18.1	2:1
24.9	1.5:1
29	2.4:1

---

58 - Dualband Full-wave Loop Antenna for 80m-40m

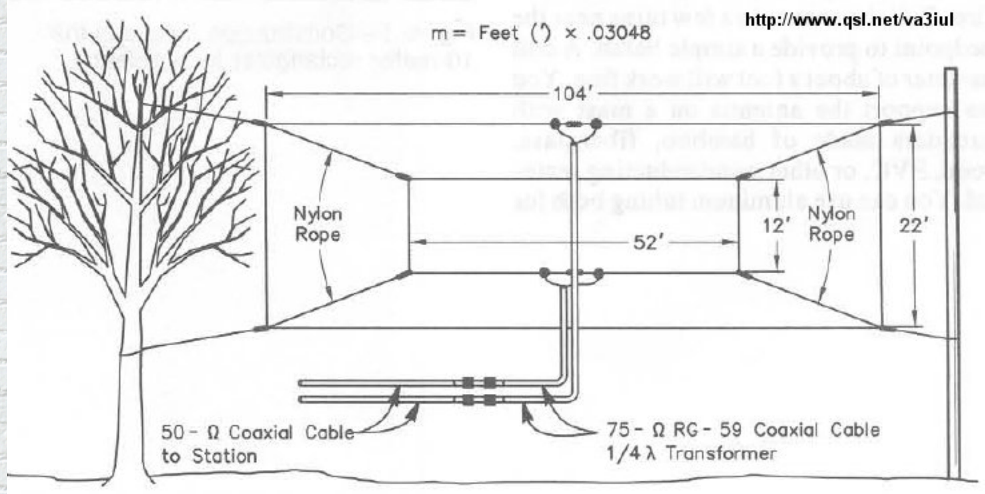
http://www.qsl.net/va3iul
m = Feet (') × .03048
104'
Nylon Rope
52'
12'
Nylon Rope
22'
50 - Ω Coaxial Cable to Station
75 - Ω RG - 59 Coaxial Cable 1/4 λ Transformer

---

59 - Loop Antenna for 10m

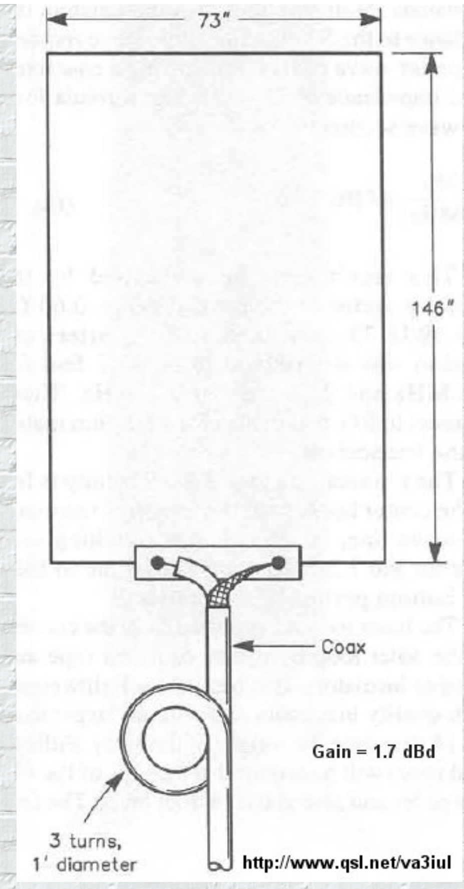
73"
146"
Coax
Gain = 1.7 dBd
3 turns,
1' diameter
http://www.qsl.net/va3iul

---

60 - Lazy Quad Antenna for 10m

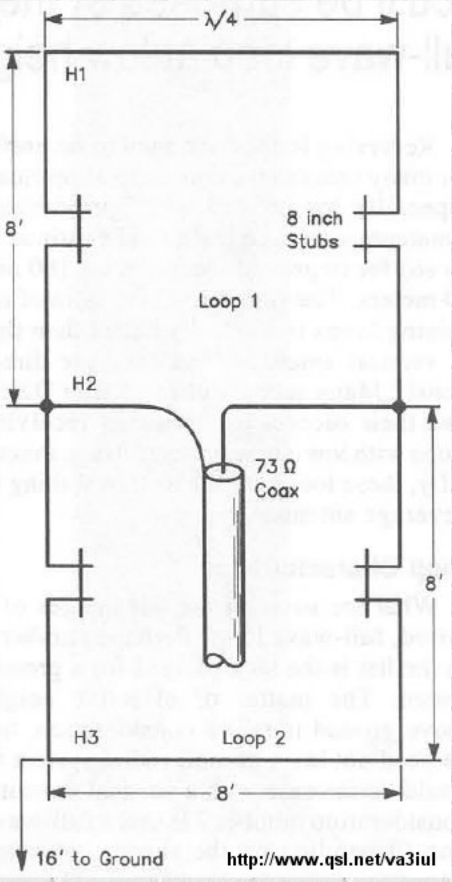
λ/4
H1
8'
8 inch
Stubs
Loop 1
H2
73 Ω
Coax
8'
H3
Loop 2
8'
16' to Ground
http://www.qsl.net/va3iul

---

61 - Tri-band Delta Loop Antenna for 80m - 40m - 30m

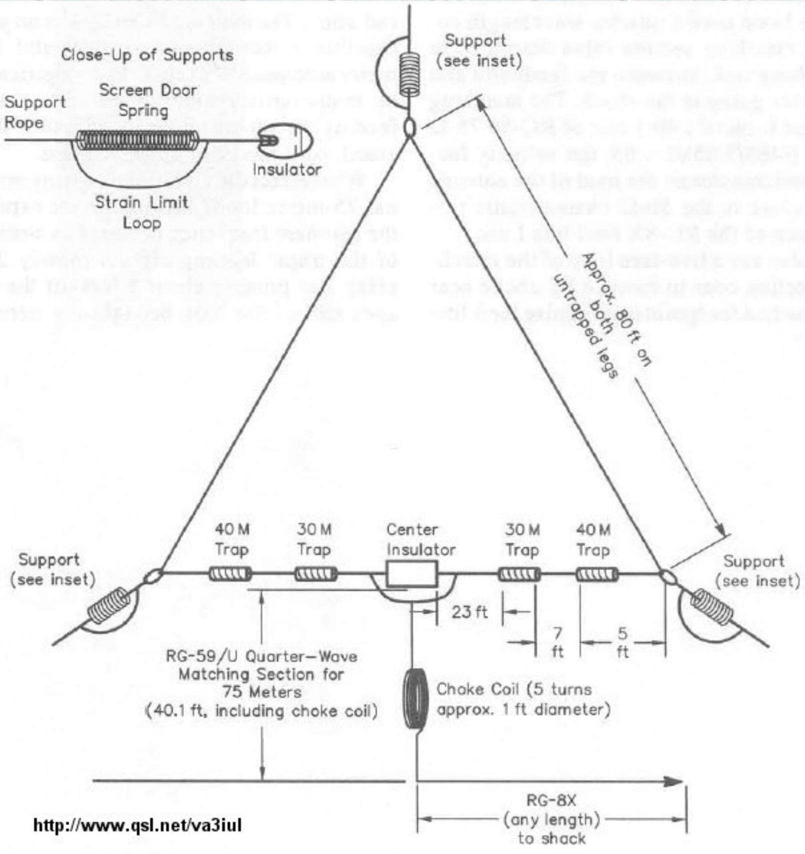
Close-Up of Supports
Support Rope
Screen Door Spring
Insulator
Strain Limit Loop
Support (see inset)
Approx. 80 ft on both untrapped legs
Support (see inset)
40 M Trap
30 M Trap
Center Insulator
30 M Trap
40 M Trap
Support (see inset)
23 ft
7 ft
5 ft
RG-59/U Quarter–Wave Matching Section for 75 Meters (40.1 ft, including choke coil)
Choke Coil (5 turns approx. 1 ft diameter)
RG-8X (any length) to shack
http://www.qsl.net/va3iul

---

62 - Dual-band Loop Antenna for 30m - 40m

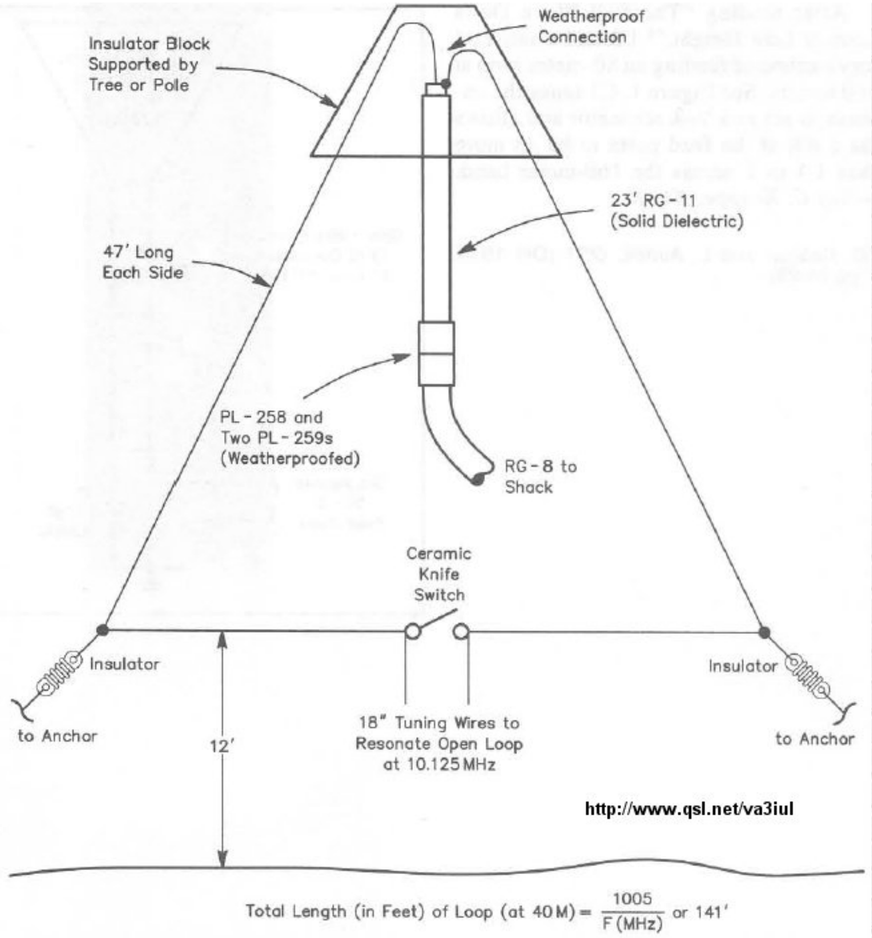
Insulator Block
Supported by
Tree or Pole
Weatherproof
Connection
23′ RG-11
(Solid Dielectric)
47′ Long
Each Side
PL-258 and
Two PL-259s
(Weatherproofed)
RG-8 to
Shack
Ceramic
Knife
Switch
Insulator
Insulator
to Anchor
to Anchor
12′
18″ Tuning Wires to
Resonate Open Loop
at 10.125 MHz
http://www.qsl.net/va3iul
Total Length (in Feet) of Loop (at 40 M) = 1005 / F (MHz) or 141′

63 - Wire-Beam Antenna for 80m

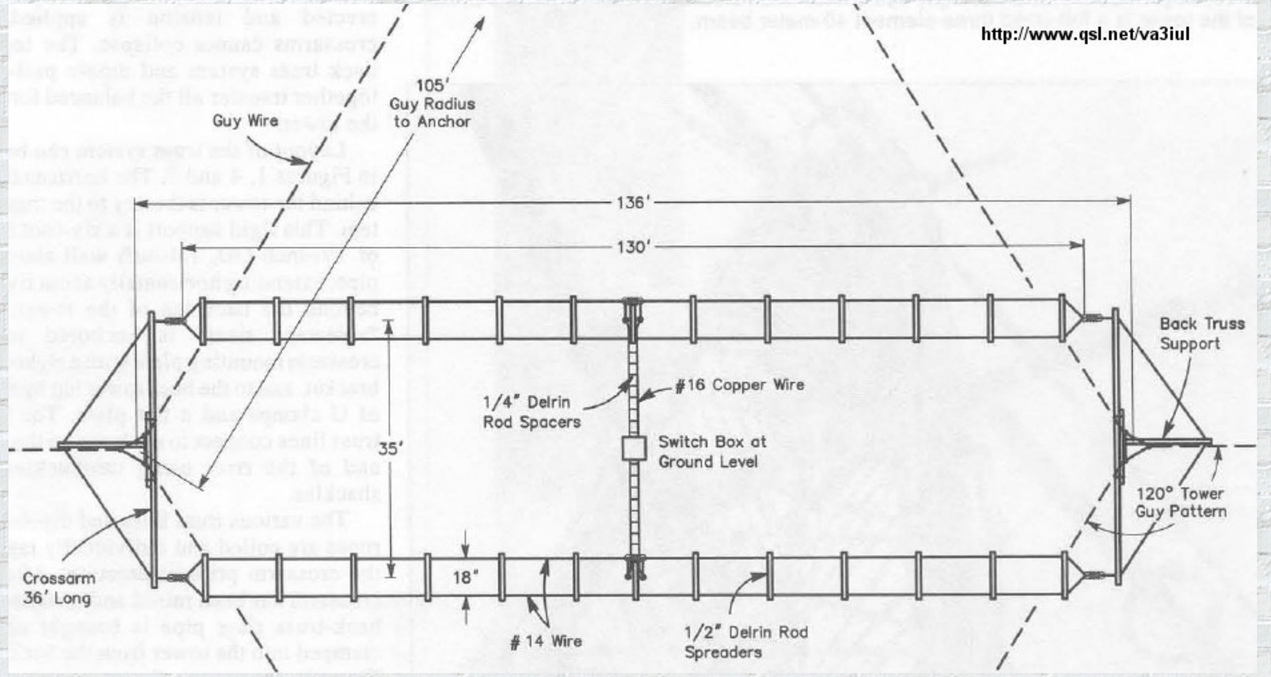
http://www.qsl.net/va3iul
105′
Guy Radius
to Anchor
Guy Wire
136′
130′
Back Truss
Support
1/4″ Delrin
Rod Spacers
#16 Copper Wire
Switch Box at
Ground Level
35′
120° Tower
Guy Pattern
Crossarm
36′ Long
18″
# 14 Wire
1/2″ Delrin Rod
Spreaders

64 - Dual-Band Sloper Antenna

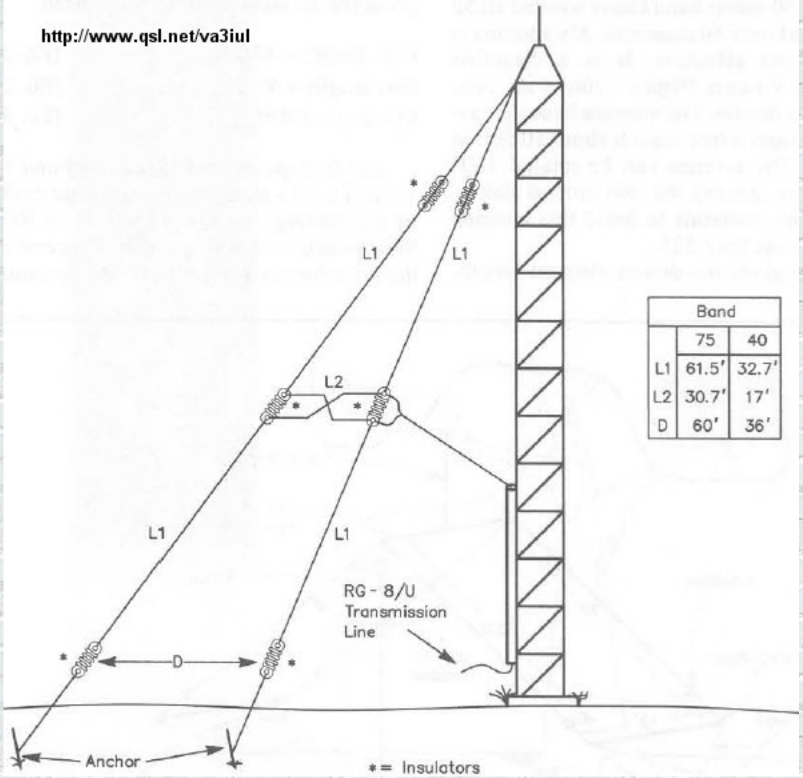
http://www.qsl.net/va3iul
L1
L1
L2
L1
L1
Band
75
40
L1 61.5′ 32.7′
L2 30.7′ 17′
D 60′ 36′
RG - 8/U
Transmission
Line
D
Anchor
* = Insulators

65 - Inverted-V Beam Antenna for 30m

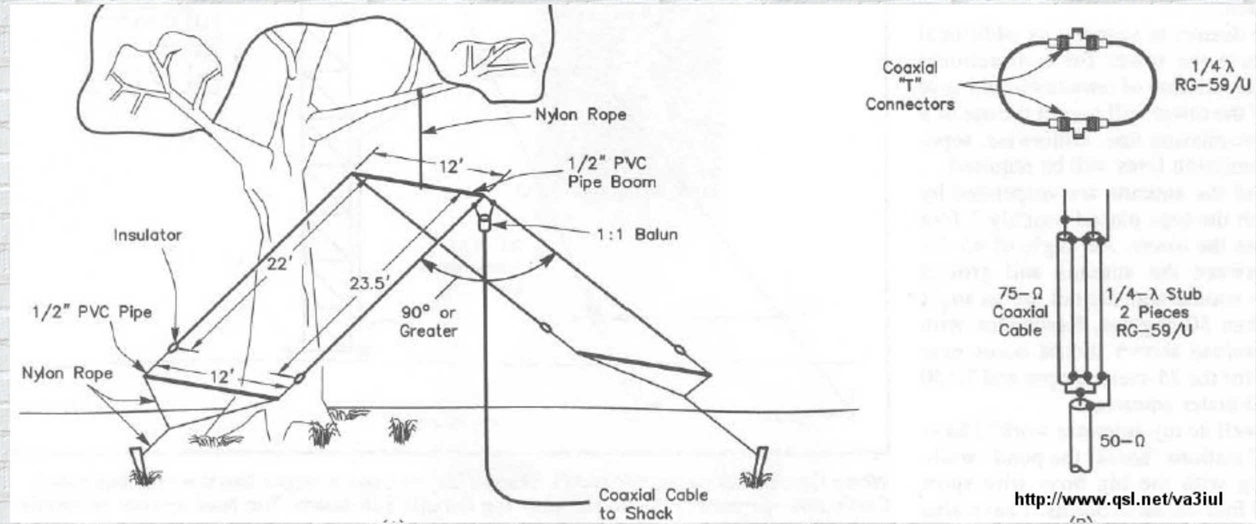
Nylon Rope
12′
1/2″ PVC
Pipe Boom
1:1 Balun
Insulator
22′
23.5′
90° or
Greater
1/2″ PVC Pipe
Nylon Rope
12′
Coaxial Cable
to Shack
Coaxial
"T"
Connectors
1/4 λ
RG–59/U
75–Ω
Coaxial
Cable
1/4–λ Stub
2 Pieces
RG–59/U
50–Ω
http://www.qsl.net/va3iul

66 - ZL-Special Beam Antenna for 15m

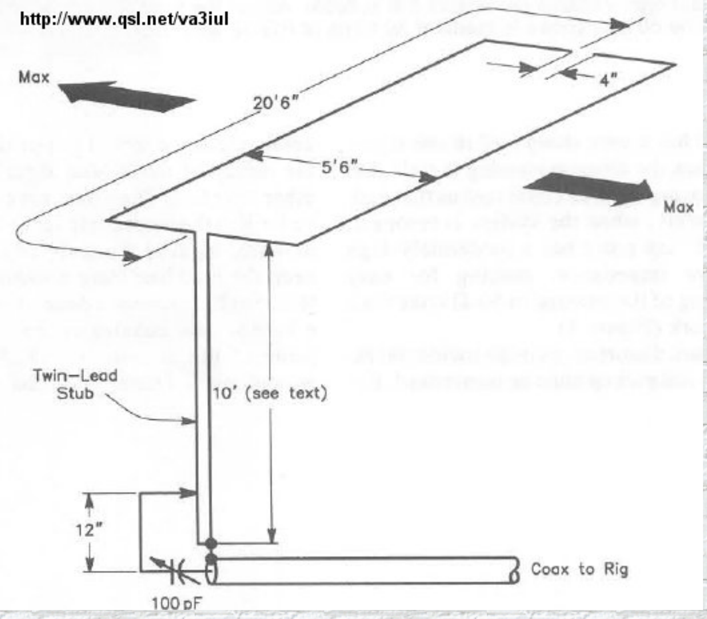
http://www.qsl.net/va3iul
Max
20'6"
4"
5'6"
Max
Twin-Lead Stub
10' (see text)
12"
Coax to Rig
100 pF

---

67 - Half-Sloper Antenna for 160m

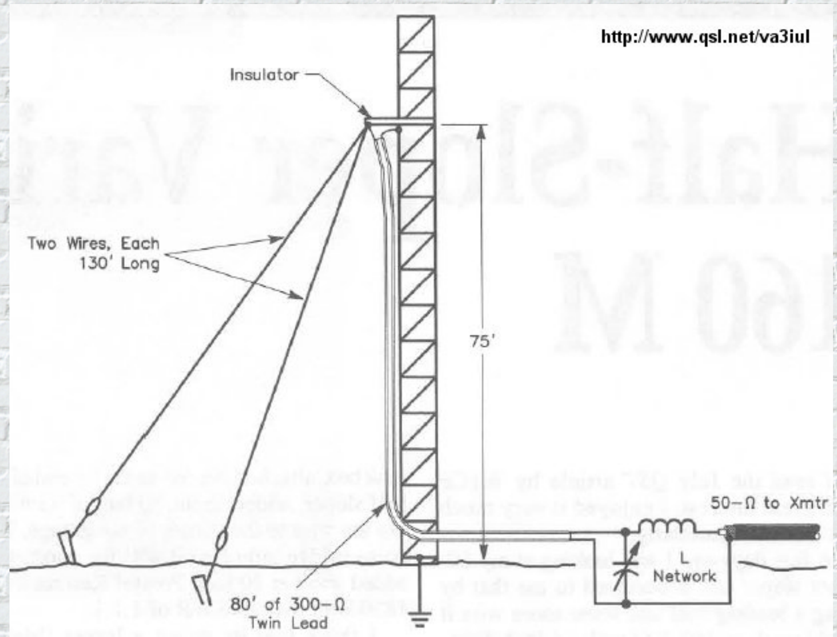
http://www.qsl.net/va3iul
Insulator
Two Wires, Each 130' Long
75'
50-Ω to Xmtr
L Network
80' of 300-Ω Twin Lead

---

68 - Two-Bands Half Sloper for 80m - 40m

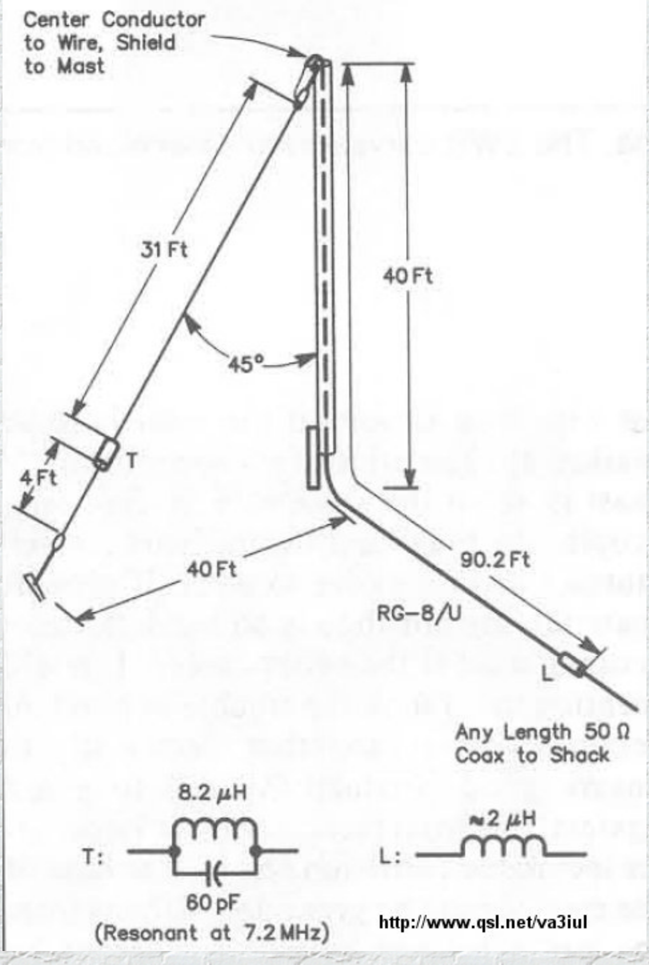
Center Conductor
to Wire, Shield
to Mast
31 Ft
40 Ft
45°
T
4 Ft
40 Ft
90.2 Ft
RG–8/U
L
Any Length 50 Ω
Coax to Shack
8.2 µH
T:
60 pF
(Resonant at 7.2 MHz)
≈2 µH
L:
http://www.qsl.net/va3iul

69 - Linear Loaded Sloper Antenna for 160m

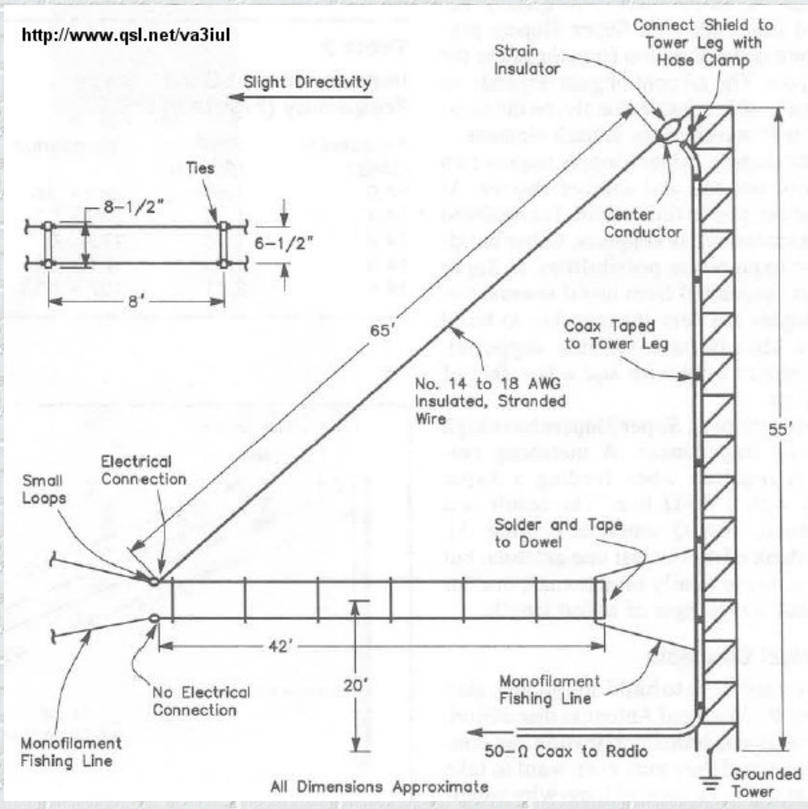
http://www.qsl.net/va3iul
Slight Directivity
Strain
Insulator
Connect Shield to
Tower Leg with
Hose Clamp
Ties
8–1/2"
6–1/2"
8'
Center
Conductor
65'
Coax Taped
to Tower Leg
No. 14 to 18 AWG
Insulated, Stranded
Wire
55'
Small
Loops
Electrical
Connection
Solder and Tape
to Dowel
42'
20'
No Electrical
Connection
Monofilament
Fishing Line
Monofilament
Fishing Line
50–Ω Coax to Radio
Grounded
Tower
All Dimensions Approximate

70 - Super-Sloper Antenna

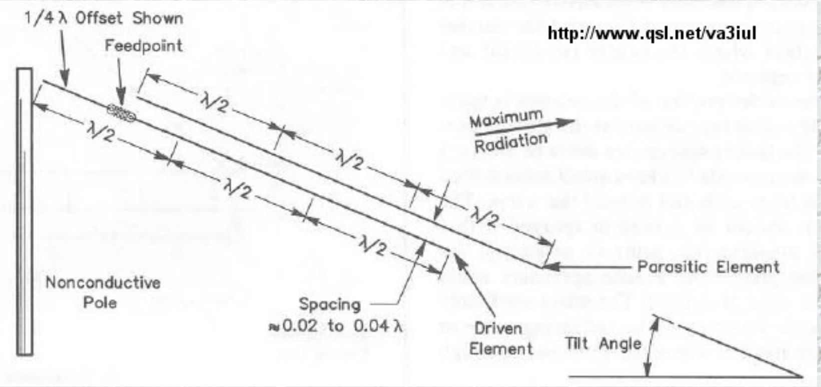
1/4 λ Offset Shown
Feedpoint
http://www.qsl.net/va3iul
λ/2
λ/2
λ/2
λ/2
λ/2
λ/2
λ/2
Maximum Radiation
Parasitic Element
Nonconductive Pole
Spacing ≈0.02 to 0.04 λ
Driven Element
Tilt Angle

---

71 - Tower Pole as a Vertical Antenna for 80m

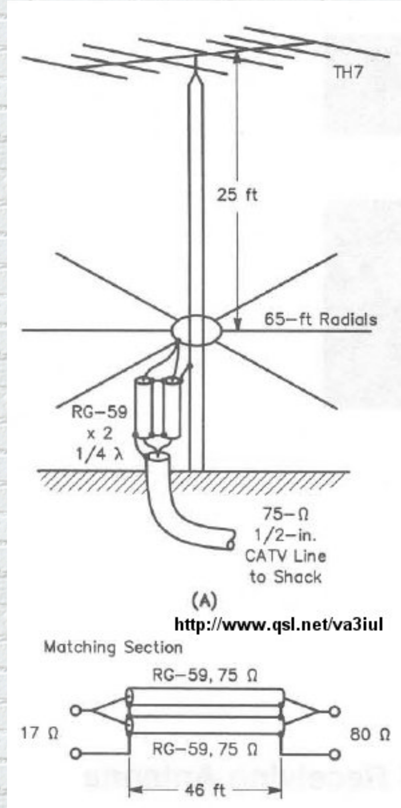
TH7
25 ft
65-ft Radials
RG-59 x 2 1/4 λ
75-Ω 1/2-in. CATV Line to Shack
(A)
http://www.qsl.net/va3iul
Matching Section
RG-59, 75 Ω
17 Ω
RG-59, 75 Ω
80 Ω
46 ft

---

72 - Clothesline Antenna

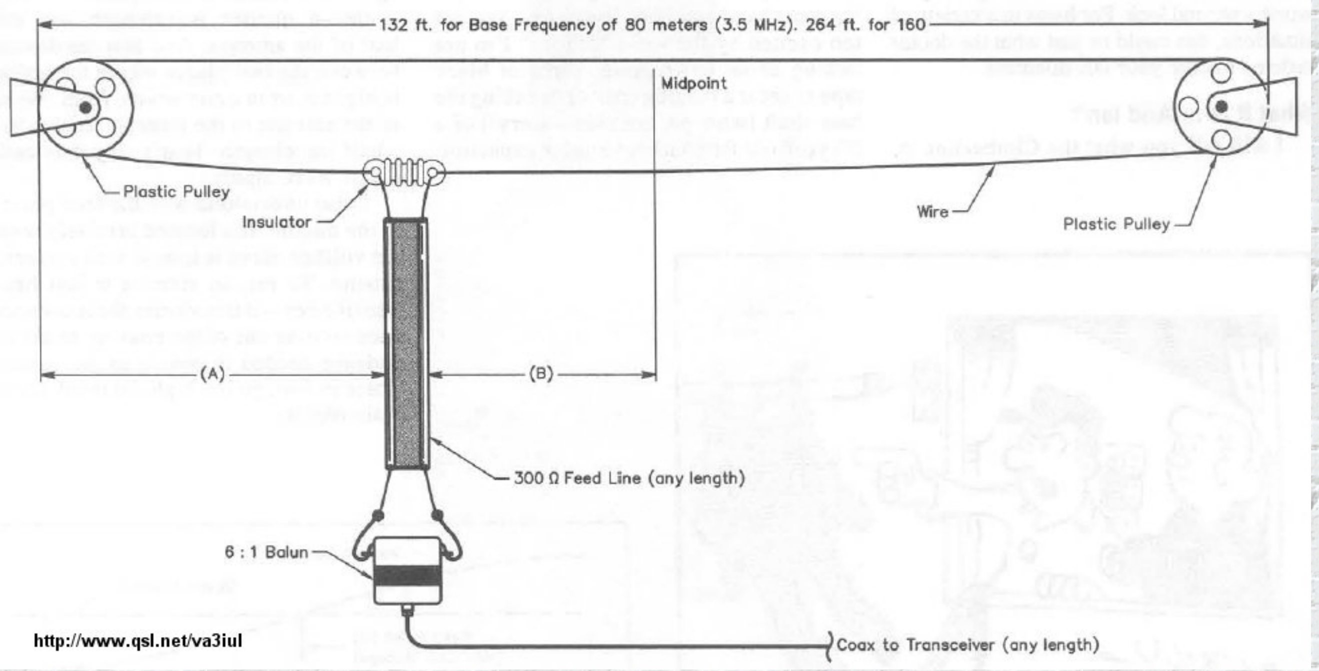
132 ft. for Base Frequency of 80 meters (3.5 MHz). 264 ft. for 160
Midpoint
Plastic Pulley
Insulator
Wire
Plastic Pulley
(A)
(B)
300 Ω Feed Line (any length)
6 : 1 Balun
Coax to Transceiver (any length)
http://www.qsl.net/va3iul

73 - Curtain Zepp Antenna for 160m, 80m, 40m

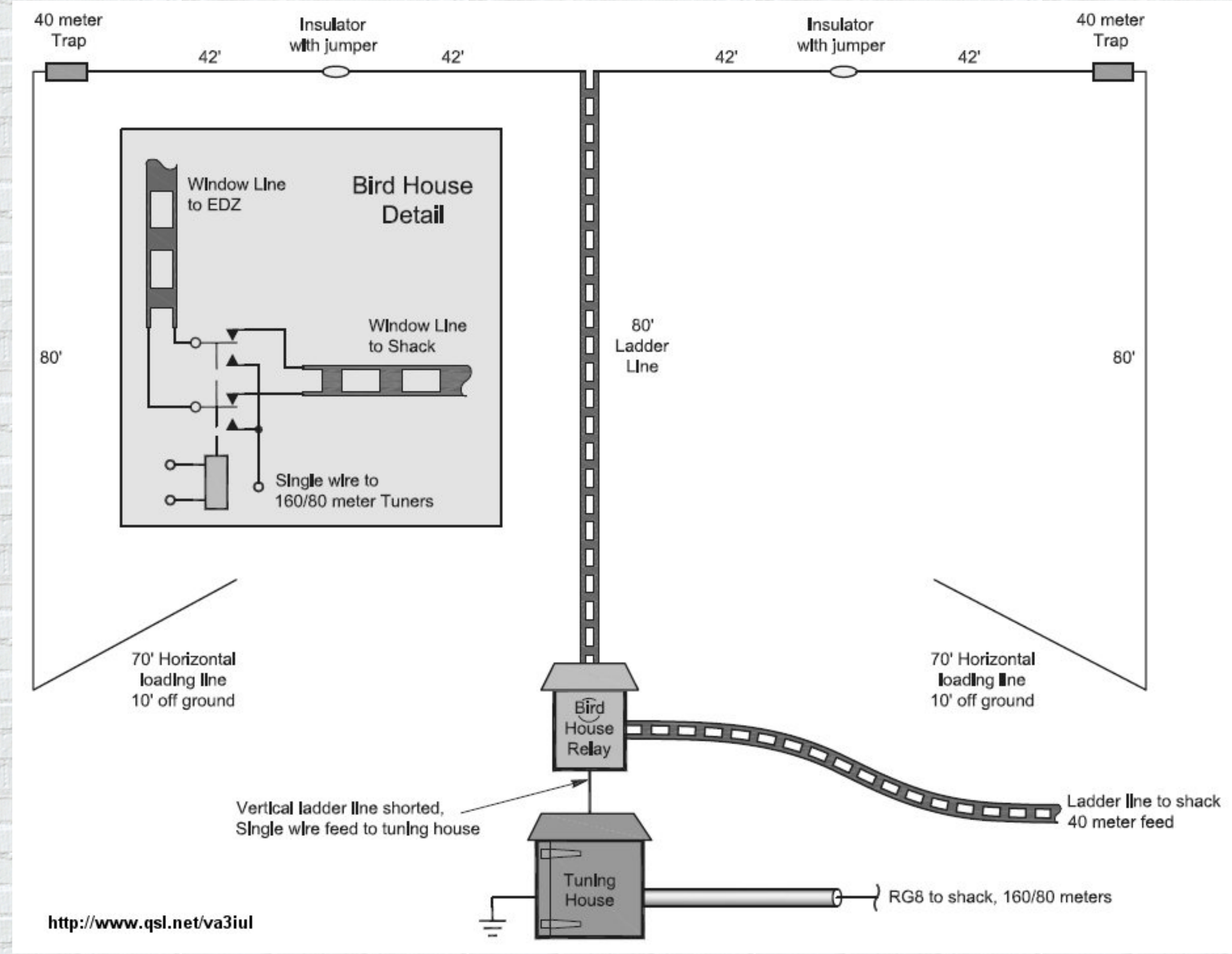
40 meter Trap
Insulator with jumper
42'
42'
42'
Insulator with jumper
42'
40 meter Trap
Window Line to EDZ
Bird House Detail
Window Line to Shack
Single wire to 160/80 meter Tuners
80'
80' Ladder Line
80'
70' Horizontal loading line 10' off ground
70' Horizontal loading line 10' off ground
Bird House Relay
Vertical ladder line shorted, Single wire feed to tuning house
Ladder line to shack 40 meter feed
Tuning House
RG8 to shack, 160/80 meters
http://www.qsl.net/va3iul

74 - Collinear Array Antenna for 40m, 30m, 20m

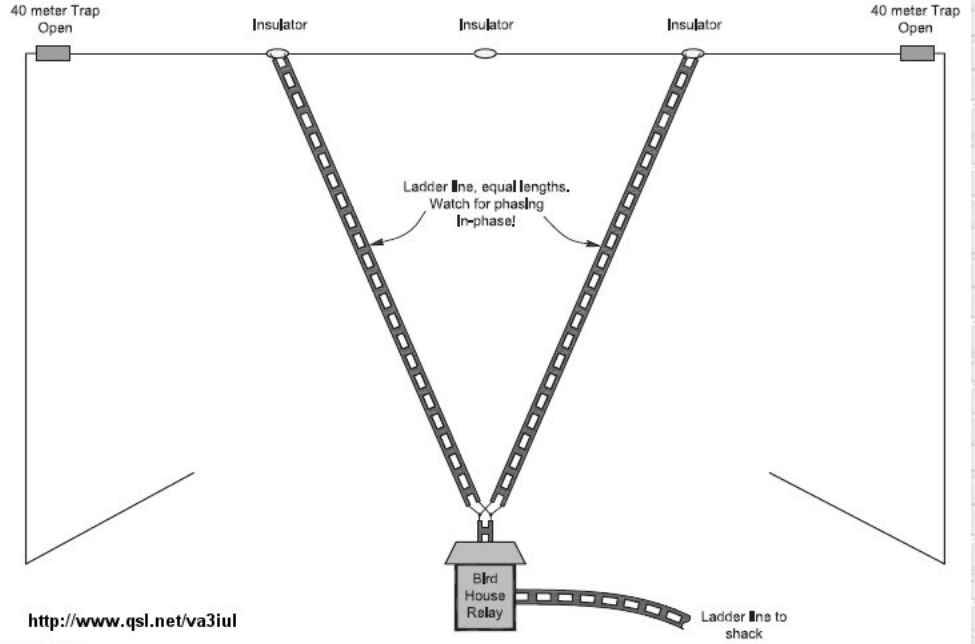
40 meter Trap
Open
Insulator
Insulator
Insulator
40 meter Trap
Open
Ladder line, equal lengths.
Watch for phasing
in-phase!
Bird
House
Relay
Ladder line to
shack
http://www.qsl.net/va3iul

75 - 160m Inverted Delta Loop

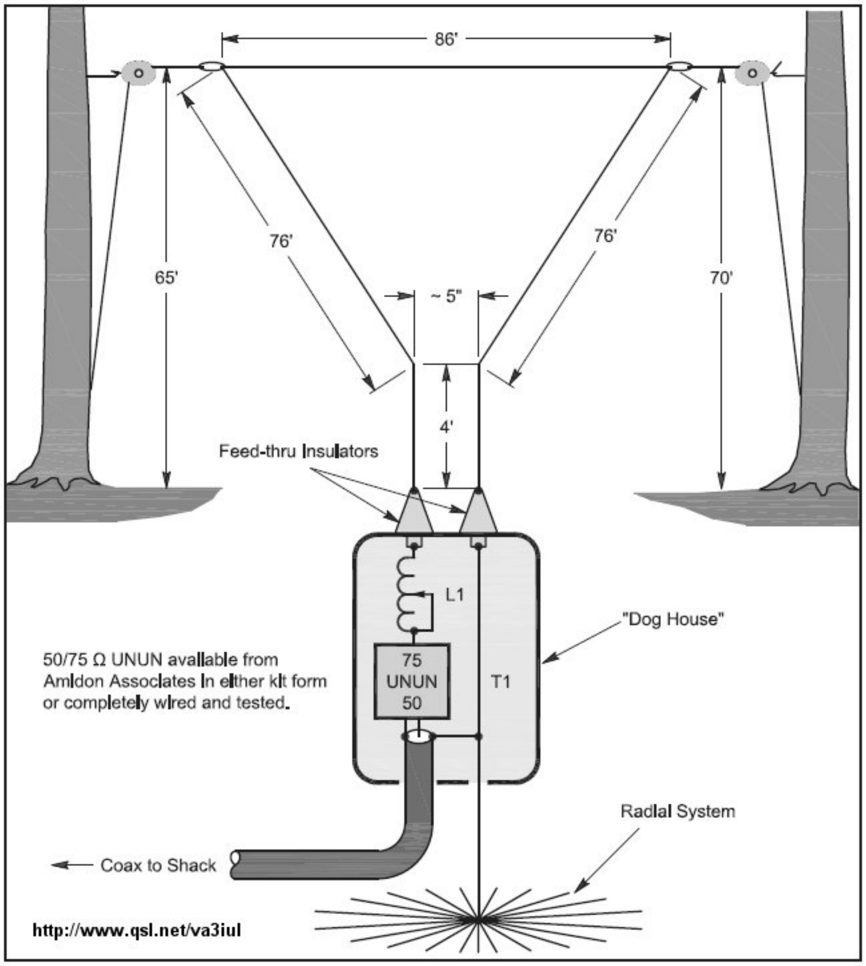
86'
76'
76'
65'
70'
~ 5"
4'
Feed-thru Insulators
L1
"Dog House"
50/75 Ω UNUN available from
Amidon Associates in either kit form
or completely wired and tested.
75
UNUN
50
T1
Radial System
Coax to Shack
http://www.qsl.net/va3iul

76 - Half Rhombic Unidirectional Vertical for 20m to 6m

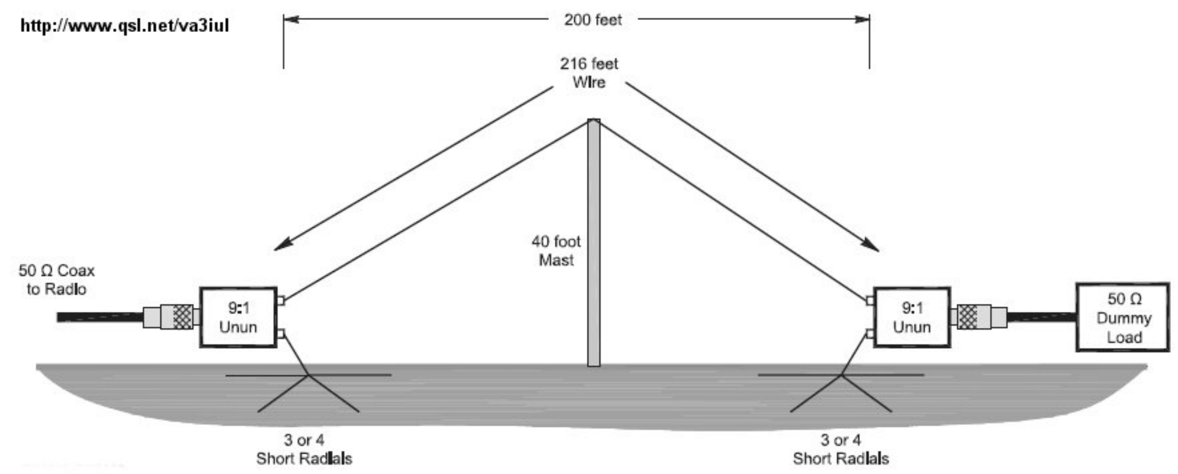
http://www.qsl.net/va3iul
200 feet
216 feet
Wire
40 foot
Mast
50 Ω Coax
to Radio
9:1
Unun
9:1
Unun
50 Ω
Dummy
Load
3 or 4
Short Radials
3 or 4
Short Radials

---

77 - Capacitance Loaded Vertical Antenna for 160m

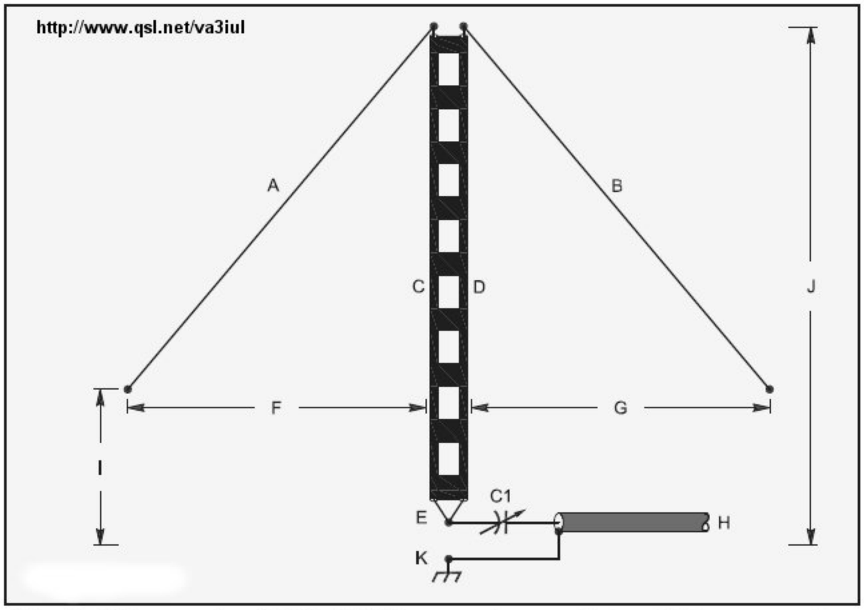
http://www.qsl.net/va3iul
A
B
C
D
J
F
G
I
C1
E
H
K

C1 — Variable capacitor, 0-1500 pF, 7 kV.

A, B — Sloping segments of 66 foot, 4 inch pieces of #16 AWG enameled copper wire soldered to the top of segments C and D and forming a 45° angle.

C, D — 59 feet of 450 Ω open wire or window line or, alternatively, two parallel lengths of #16 AWG enameled copper wire with a 1 inch separation.

E — 1 inch triangle formed by the bottom ends of the open wire line. Solder them tightly at the vertex and connect to C1.

F, G — Distance between the lower ends of segments A and B from segments C and D, 51 feet.

H — Any length of 50 Ω coaxial cable (RG-213 in this case). The central conductor must be soldered to C1, and the braid to the ground radials.

I — Approximately 16 feet from ground to lower ends of sloping segments A and B.

J — 60 feet from antenna top to bottom.

K — Ground radials, 60 ¼ λ (the more the better — 32 will form a reasonable ground).

---

78 - Fan Dipole Antenna for 80m to 6m

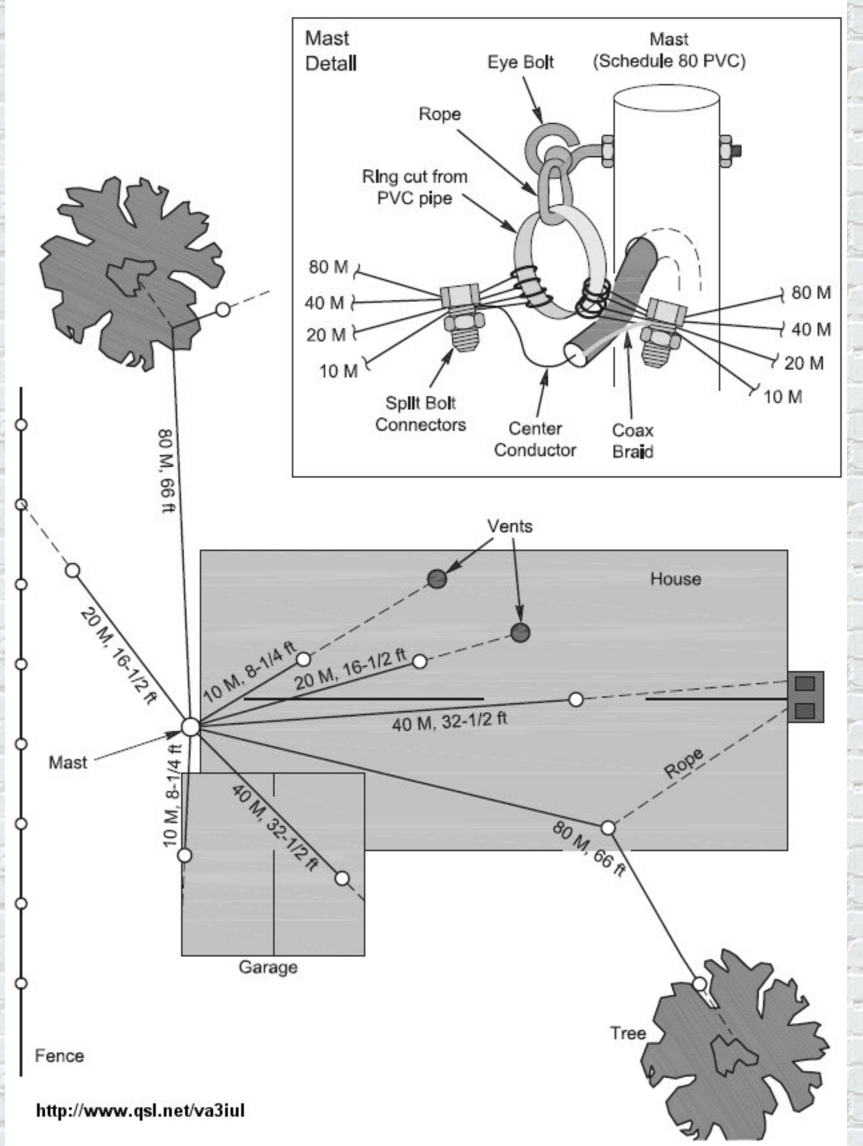
Mast
Detail
Eye Bolt
Mast
(Schedule 80 PVC)
Rope
Ring cut from
PVC pipe
80 M
40 M
20 M
10 M
80 M
40 M
20 M
10 M
Split Bolt
Connectors
Center
Conductor
Coax
Braid
80 M, 66 ft
Vents
House
20 M, 16-1/2 ft
10 M, 8-1/4 ft
20 M, 16-1/2 ft
40 M, 32-1/2 ft
Mast
10 M, 8-1/4 ft
40 M, 32-1/2 ft
Rope
80 M, 66 ft
Garage
Tree
Fence
http://www.qsl.net/va3iul

79 - Wire Ground Plane Antenna

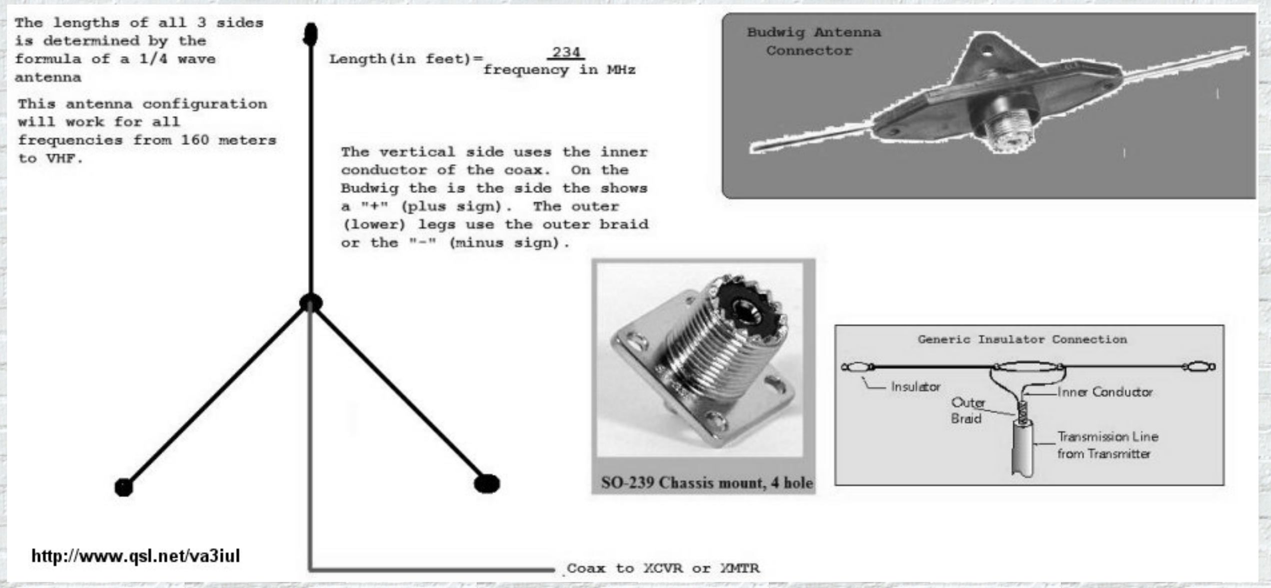

The lengths of all 3 sides is determined by the formula of a 1/4 wave antenna

This antenna configuration will work for all frequencies from 160 meters to VHF.

$$\text{Length (in feet)} = \frac{234}{\text{frequency in MHz}}$$

The vertical side uses the inner conductor of the coax. On the Budwig the is the side the shows a "+" (plus sign). The outer (lower) legs use the outer braid or the "-" (minus sign).

Budwig Antenna
Connector
SO-239 Chassis mount, 4 hole
Generic Insulator Connection
Insulator
Inner Conductor
Outer
Braid
Transmission Line
from Transmitter
http://www.qsl.net/va3iul
Coax to XCVR or XMTR

80 - Inverted Delta Loop Antenna for 160m

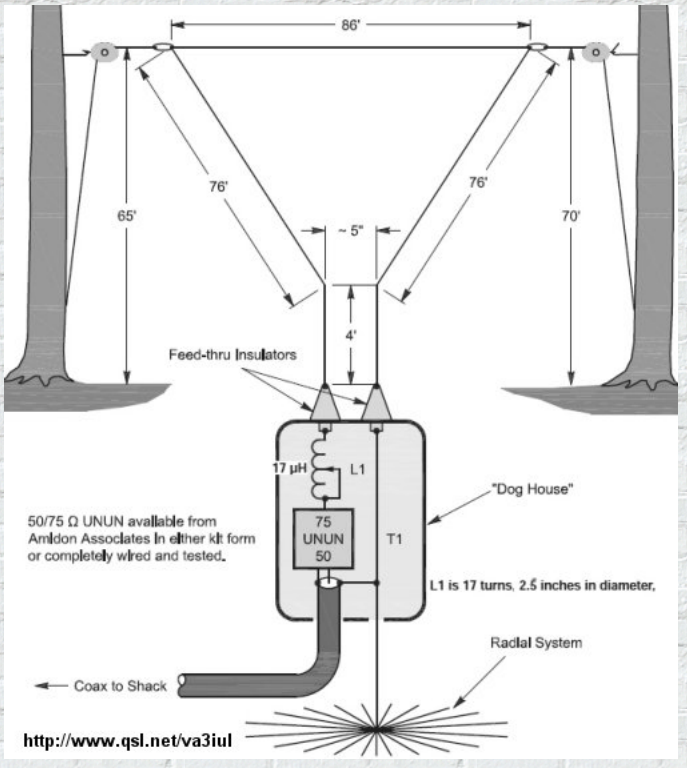
86'
76'
76'
65'
70'
~ 5"
4'
Feed-thru Insulators
17 µH
L1
"Dog House"
50/75 Ω UNUN available from Amidon Associates in either kit form or completely wired and tested.
75 UNUN 50
T1
L1 is 17 turns, 2.5 inches in diameter,
Radial System
Coax to Shack
http://www.qsl.net/va3iul

---

81 - Inverted-L for 160m

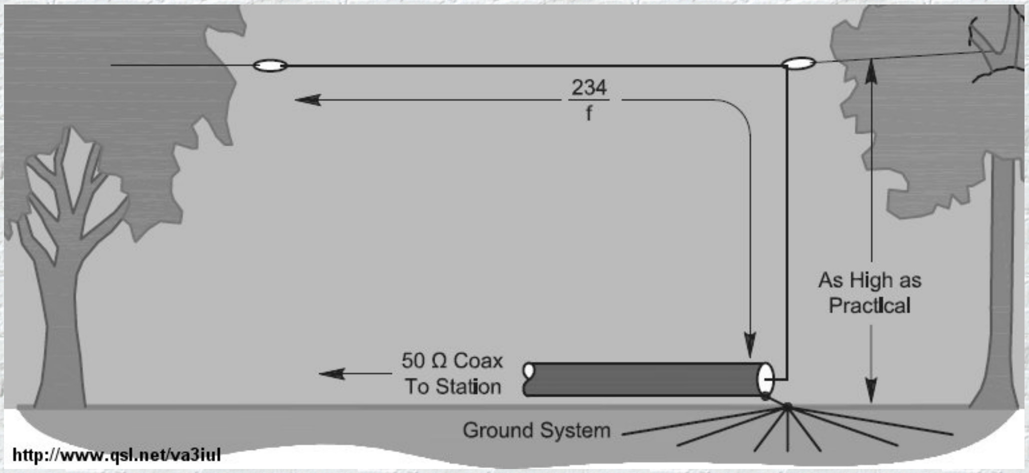
234 f
As High as Practical
50 Ω Coax To Station
Ground System
http://www.qsl.net/va3iul

---

82 - 300ohm-Ribbon Dual Band Dipole

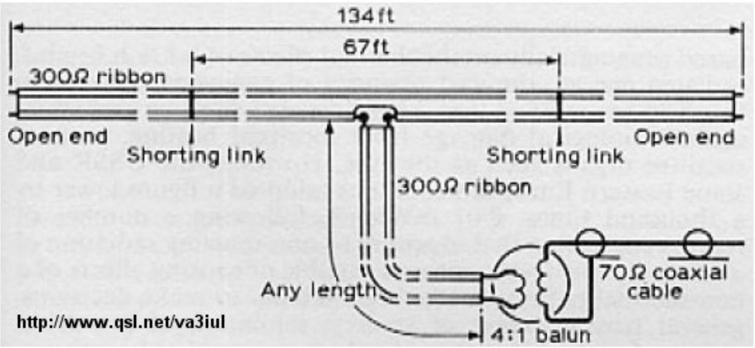
134 ft
67 ft
300Ω ribbon
Open end
Shorting link
Shorting link
Open end
300Ω ribbon
70Ω coaxial cable
Any length
4:1 balun
http://www.qsl.net/va3iul

---

83 - Tri-Band Beam for 20m, 15m, 10m

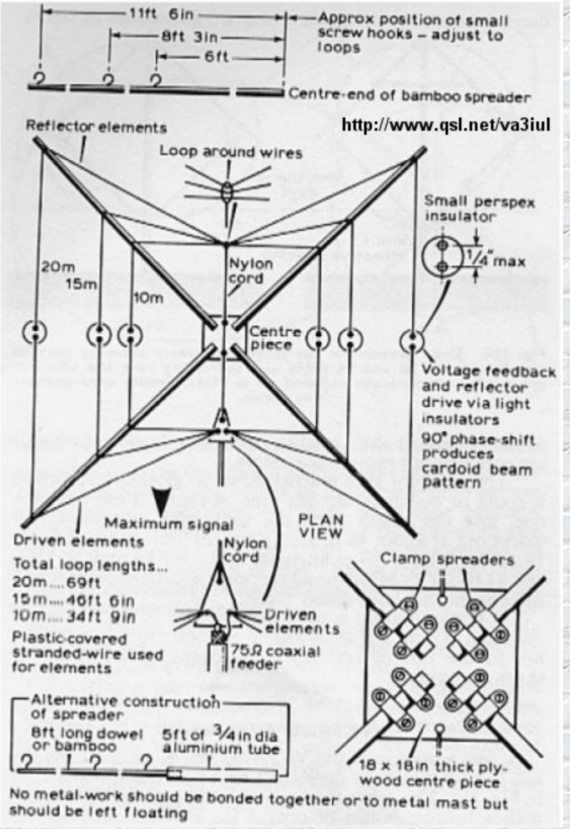
11ft 6in
8ft 3in
6ft
Approx position of small screw hooks – adjust to loops
Centre-end of bamboo spreader
Reflector elements
http://www.qsl.net/va3iul
Loop around wires
Small perspex insulator
1/4" max
20m
15m
10m
Nylon cord
Centre piece
Voltage feedback and reflector drive via light insulators
90° phase-shift produces cardoid beam pattern
Maximum signal
PLAN VIEW
Driven elements
Total loop lengths...
20m....69ft
15m....46ft 6in
10m....34ft 9in
Plastic-covered stranded-wire used for elements
Nylon cord
Driven elements
75Ω coaxial feeder
Clamp spreaders
Alternative construction of spreader
8ft long dowel or bamboo
5ft of 3/4 in dia aluminium tube
18 x 18in thick ply-wood centre piece
No metal-work should be bonded together or to metal mast but should be left floating

---

84 - Mini-Horse Yagi Antenna

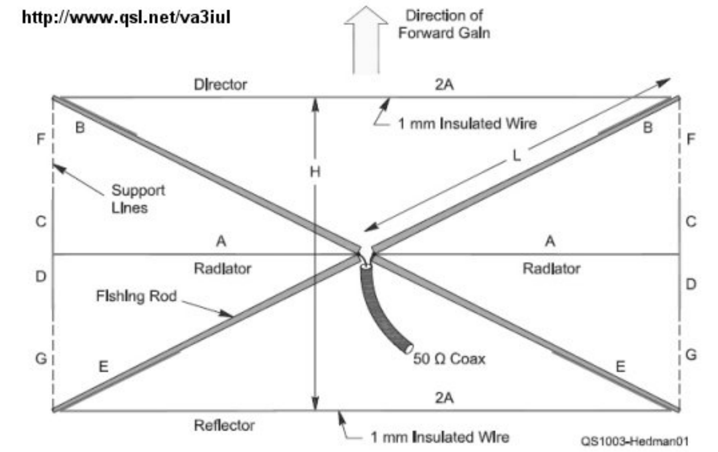
http://www.qsl.net/va3iul
Direction of Forward Gain
Director
2A
1 mm Insulated Wire
B
F
Support Lines
H
L
C
A
Radiator
D
Fishing Rod
50 Ω Coax
G
E
Reflector
1 mm Insulated Wire
QS1003-Hedman01

MHz	A	B	C	D	E	F	G	H	L
144	1' 2.2"	0' 4.7"	0' 3.5"	0' 1.8"	0' 7.1"	0' 4.7"	0' 6.5"	1' 4.5"	1' 4.3"
28	5' 11.6"	1' 10.8"	1' 6.9"	0' 9.4"	2' 11.8"	1' 10.8"	2' 8.3"	6' 11.4"	6' 11.0"
21	7' 11.6"	2' 6.3"	2' 1.2"	1' 0.6"	3' 11.6"	2' 6.3"	3' 6.9"	9' 3.0"	9' 2.6"
14	11' 11.3"	3' 9.7"	3' 1.8"	1' 6.9"	5' 11.2"	3' 9.7"	5' 4.6"	13' 10.9"	13' 9.7"
7.1	23' 10.5"	7' 7.3"	6' 3.6"	3' 1.8"	11' 10.5"	7' 7.3"	10' 9.1"	27' 9.8"	27' 7.4"

---

85 - Half-Rhombic Directive Antenna for 20m to 6m

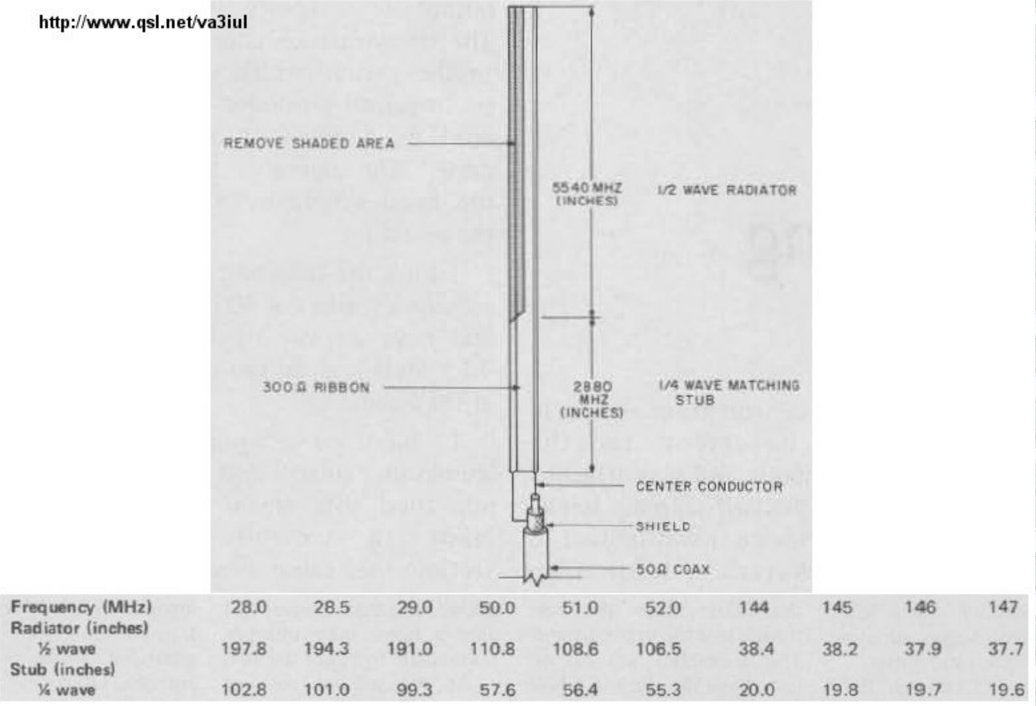
http://www.qsl.net/va3iul
REMOVE SHADED AREA
5540 MHZ (INCHES)
1/2 WAVE RADIATOR
300 Ω RIBBON
2880 MHZ (INCHES)
1/4 WAVE MATCHING STUB
CENTER CONDUCTOR
SHIELD
50Ω COAX

Frequency (MHz)	28.0	28.5	29.0	50.0	51.0	52.0	144	145	146	147
Radiator (inches)										
½ wave	197.8	194.3	191.0	110.8	108.6	106.5	38.4	38.2	37.9	37.7
Stub (inches)										
¼ wave	102.8	101.0	99.3	57.6	56.4	55.3	20.0	19.8	19.7	19.6

86 - Fan-Dipole Antenna for 80m, 40m, 20m

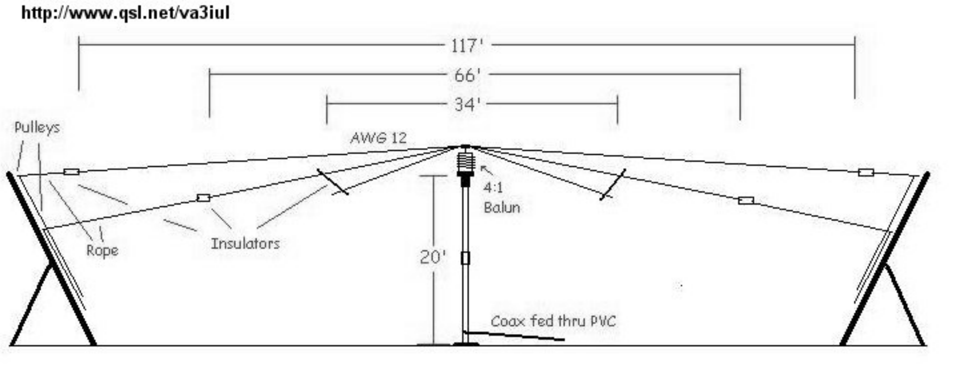
http://www.qsl.net/va3iul
117'
66'
34'
Pulleys
AWG 12
4:1
Balun
Rope
Insulators
20'
Coax fed thru PVC

87 - Capacity Tuned Folded Loop Antenna for 20m

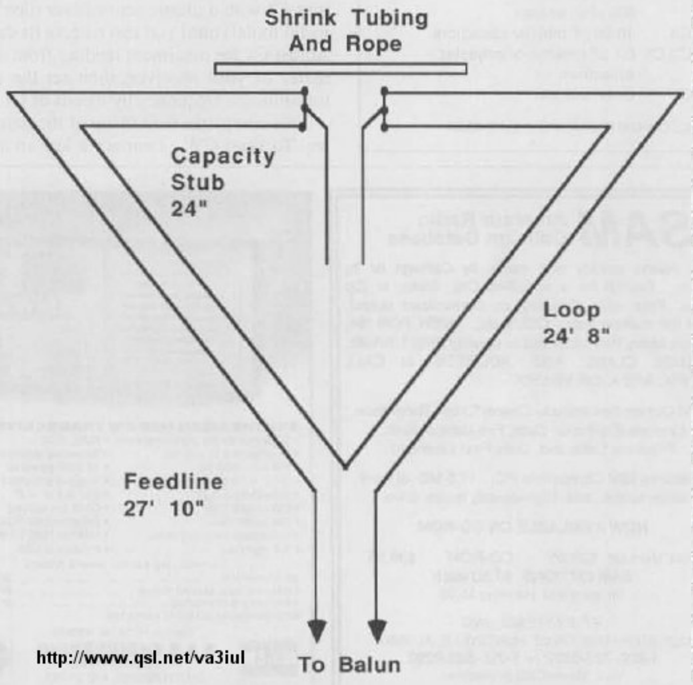
Shrink Tubing
And Rope
Capacity
Stub
24"
Loop
24' 8"
Feedline
27' 10"
http://www.qsl.net/va3iul
To Balun

88 - Indoor Loop Antenna for 80m to 30m

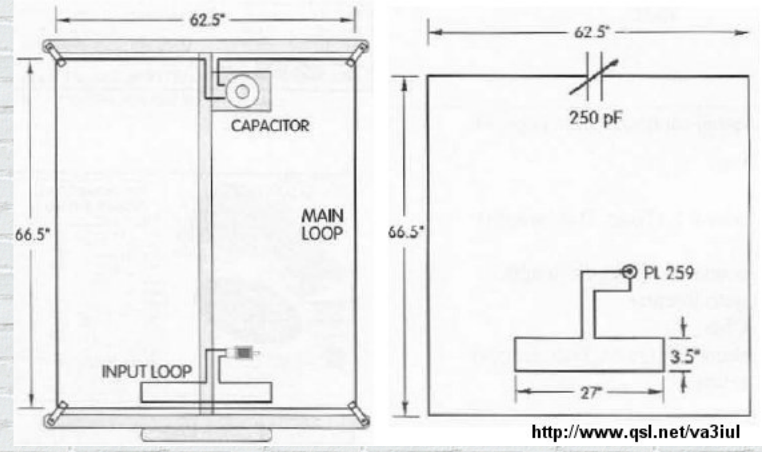
62.5"
CAPACITOR
MAIN
LOOP
66.5"
INPUT LOOP
62.5"
250 pF
66.5"
PL 259
3.5"
27"
http://www.qsl.net/va3iul

89 - Indoor Loop Antenna for 80m

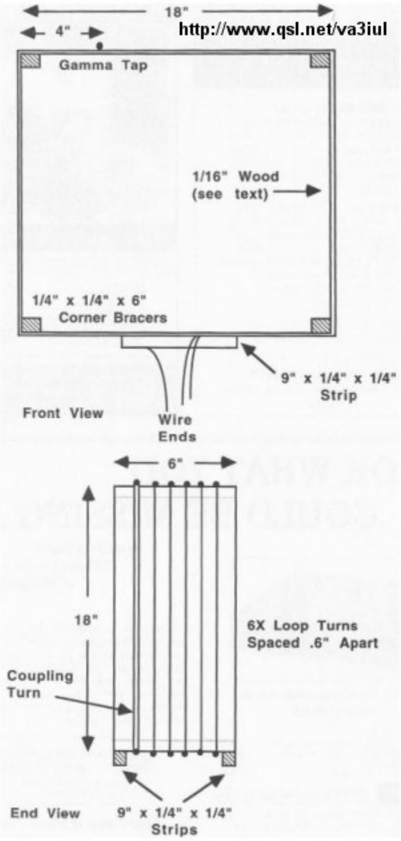
18"
4"
http://www.qsl.net/va3iul
Gamma Tap
1/16" Wood
(see text)
1/4" x 1/4" x 6"
Corner Bracers
9" x 1/4" x 1/4"
Strip
Front View
Wire
Ends
6"
18"
6X Loop Turns
Spaced .6" Apart
Coupling
Turn
End View
9" x 1/4" x 1/4"
Strips

90 - Double-Delta Antenna 80m and 40m

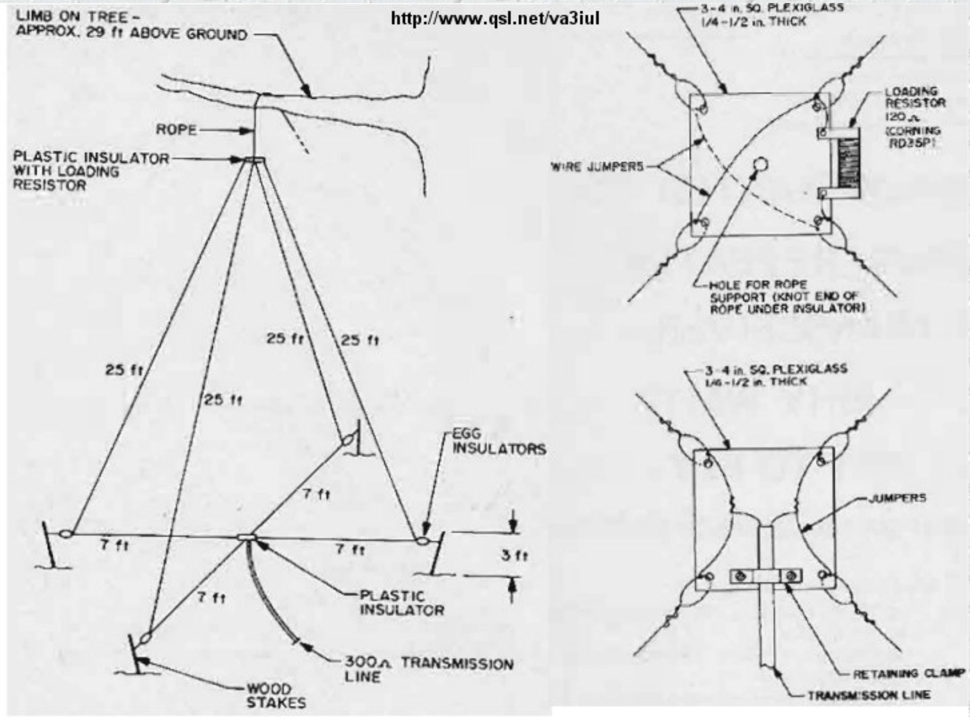
LIMB ON TREE - APPROX. 29 ft ABOVE GROUND
http://www.qsl.net/va3iul
ROPE
PLASTIC INSULATOR WITH LOADING RESISTOR
25 ft
25 ft
25 ft
25 ft
EGG INSULATORS
7 ft
7 ft
7 ft
7 ft
3 ft
PLASTIC INSULATOR
300Ω TRANSMISSION LINE
WOOD STAKES
3-4 in. SQ. PLEXIGLASS 1/4-1/2 in. THICK
LOADING RESISTOR 120Ω (CORNING RD35P)
WIRE JUMPERS
HOLE FOR ROPE SUPPORT (KNOT END OF ROPE UNDER INSULATOR)
3-4 in. SQ. PLEXIGLASS 1/4-1/2 in. THICK
JUMPERS
RETAINING CLAMP
TRANSMISSION LINE

91 - Inductance-Loaded Shortened Dipole for 160m

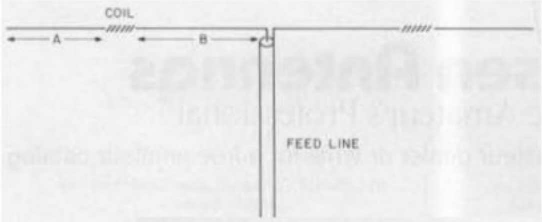
COIL
A
B
FEED LINE

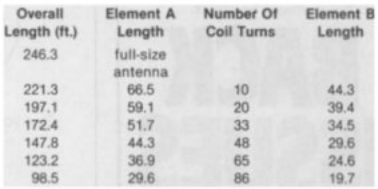

Overall Length (ft.)	Element A Length	Number Of Coil Turns	Element B Length
246.3	full-size antenna		
221.3	66.5	10	44.3
197.1	59.1	20	39.4
172.4	51.7	33	34.5
147.8	44.3	48	29.6
123.2	36.9	65	24.6
98.5	29.6	86	19.7

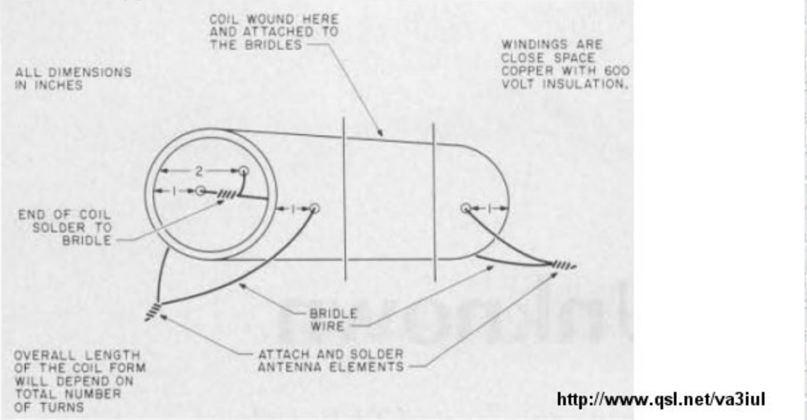
COIL WOUND HERE AND ATTACHED TO THE BRIDLES
ALL DIMENSIONS IN INCHES
WINDINGS ARE CLOSE SPACE COPPER WITH 600 VOLT INSULATION.
2
1
1
1
END OF COIL SOLDER TO BRIDLE
BRIDLE WIRE
OVERALL LENGTH OF THE COIL FORM WILL DEPEND ON TOTAL NUMBER OF TURNS
ATTACH AND SOLDER ANTENNA ELEMENTS
http://www.qsl.net/va3iul

92 - V-Beam Antenna for 15m

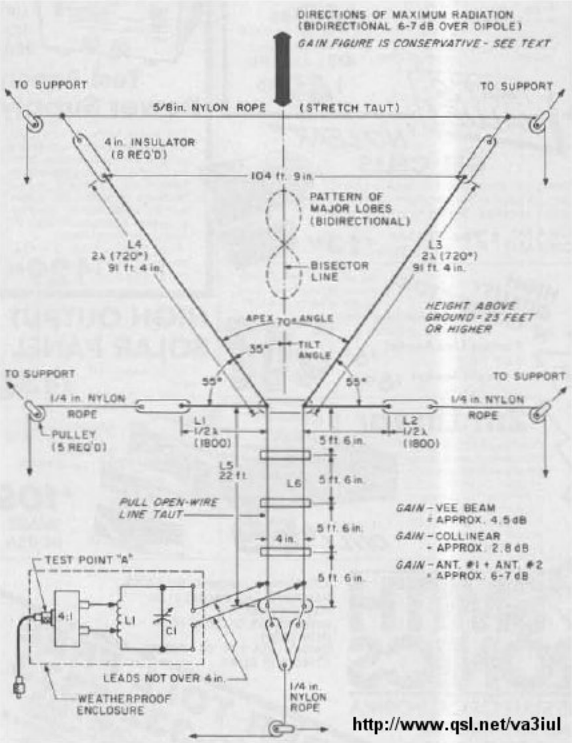
DIRECTIONS OF MAXIMUM RADIATION (BIDIRECTIONAL 6-7 dB OVER DIPOLE)
GAIN FIGURE IS CONSERVATIVE - SEE TEXT
TO SUPPORT
TO SUPPORT
5/8 in. NYLON ROPE (STRETCH TAUT)
4 in. INSULATOR (8 REQ'D)
104 ft. 9 in.
PATTERN OF MAJOR LOBES (BIDIRECTIONAL)
L4 2λ (720°) 91 ft. 4 in.
L3 2λ (720°) 91 ft. 4 in.
BISECTOR LINE
HEIGHT ABOVE GROUND = 23 FEET OR HIGHER
APEX 70° ANGLE
35° TILT ANGLE
55°
55°
TO SUPPORT
TO SUPPORT
1/4 in. NYLON ROPE
1/4 in. NYLON ROPE
PULLEY (5 REQ'D)
L1 1/2λ (180°)
L2 1/2λ (180°)
L5 22 ft.
L6
5 ft. 6 in.
5 ft. 6 in.
5 ft. 6 in.
5 ft. 6 in.
4 in.
PULL OPEN-WIRE LINE TAUT
GAIN - VEE BEAM = APPROX. 4.5 dB
GAIN - COLLINEAR - APPROX. 2.8 dB
GAIN - ANT. #1 + ANT. #2 = APPROX. 6-7 dB
TEST POINT "A"
4:1
L1
C1
LEADS NOT OVER 4 in.
WEATHERPROOF ENCLOSURE
1/4 in. NYLON ROPE
http://www.qsl.net/va3iul

93 - Picnic Vertical Wire Antenna

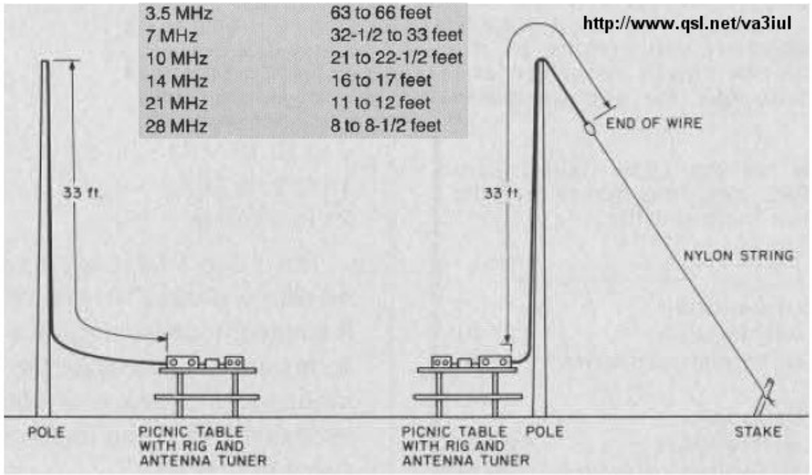

3.5 MHz	63 to 66 feet
7 MHz	32-1/2 to 33 feet
10 MHz	21 to 22-1/2 feet
14 MHz	16 to 17 feet
21 MHz	11 to 12 feet
28 MHz	8 to 8-1/2 feet

http://www.qsl.net/va3iul
33 ft.
33 ft.
END OF WIRE
NYLON STRING
POLE
PICNIC TABLE WITH RIG AND ANTENNA TUNER
PICNIC TABLE WITH RIG AND ANTENNA TUNER
POLE
STAKE

94 - Laid-Back Quad Antenna for 80m

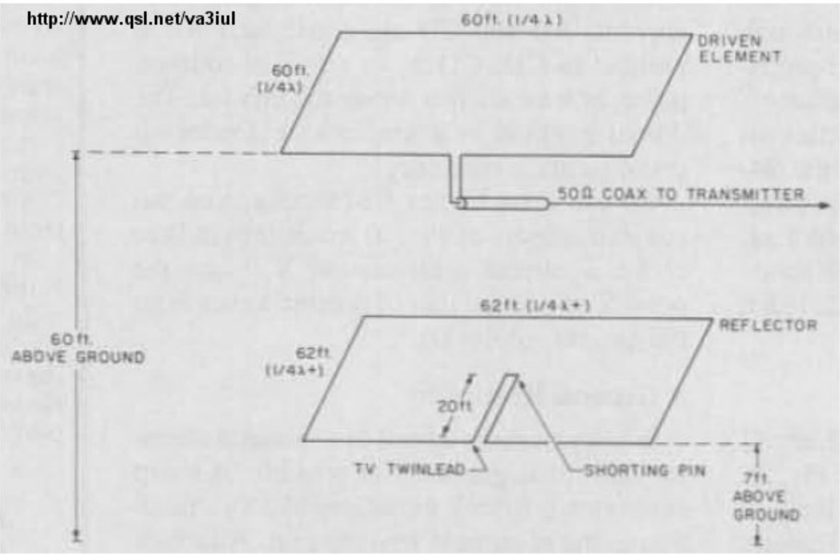

http://www.qsl.net/va3iul
60ft. (1/4 λ)
DRIVEN ELEMENT
60ft. (1/4λ)
50Ω COAX TO TRANSMITTER
60 ft. ABOVE GROUND
62ft (1/4λ+)
REFLECTOR
62ft (1/4λ+)
20ft
TV TWINLEAD
SHORTING PIN
7ft. ABOVE GROUND

95 - Phased Loop Antenna

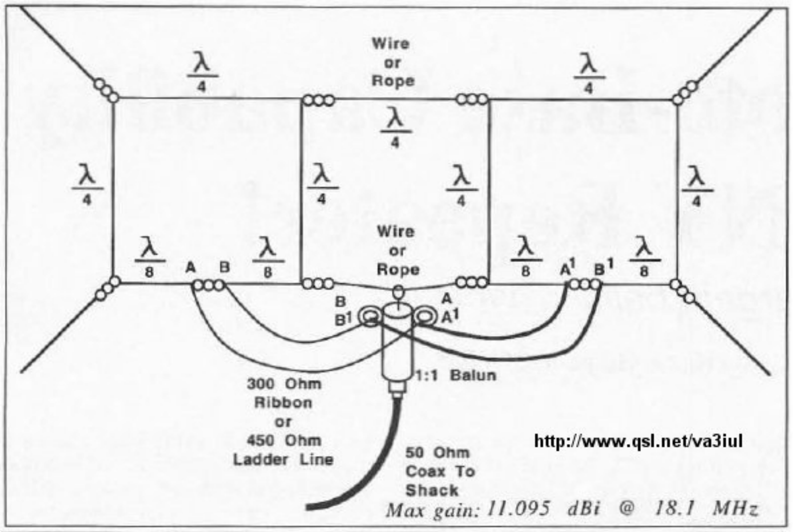

Wire or Rope
λ/4
Wire or Rope
λ/8
A
B
A¹
B¹
1:1 Balun
300 Ohm Ribbon or 450 Ohm Ladder Line
50 Ohm Coax To Shack
http://www.qsl.net/va3iul
Max gain: 11.095 dBi @ 18.1 MHz

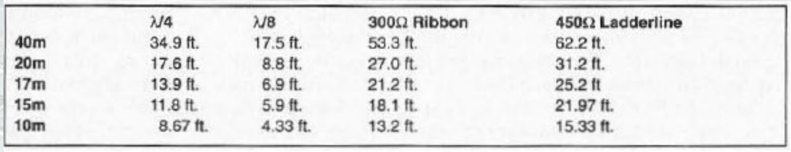

	λ/4	λ/8	300Ω Ribbon	450Ω Ladderline
40m	34.9 ft.	17.5 ft.	53.3 ft.	62.2 ft.
20m	17.6 ft.	8.8 ft.	27.0 ft.	31.2 ft.
17m	13.9 ft.	6.9 ft.	21.2 ft.	25.2 ft
15m	11.8 ft.	5.9 ft.	18.1 ft.	21.97 ft.
10m	8.67 ft.	4.33 ft.	13.2 ft.	15.33 ft.

96 - Loop Antenna for TX for 160m

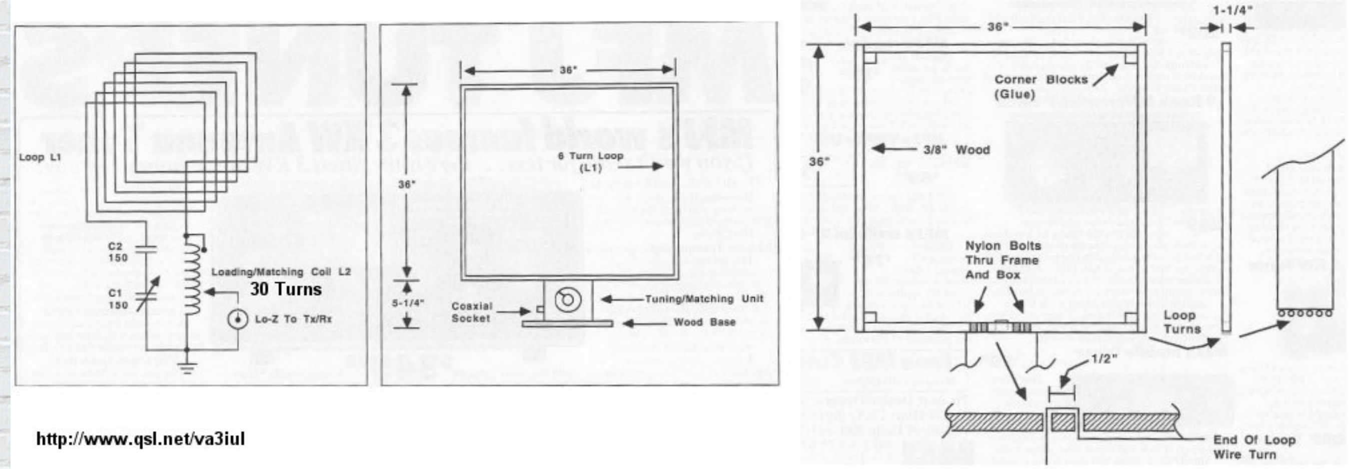
Loop L1
C2
150
C1
150
Loading/Matching Coil L2
30 Turns
Lo-Z To Tx/Rx
36"
36"
6 Turn Loop
(L1)
5-1/4"
Coaxial
Socket
Tuning/Matching Unit
Wood Base
36"
1-1/4"
Corner Blocks
(Glue)
36"
3/8" Wood
Nylon Bolts
Thru Frame
And Box
Loop
Turns
1/2"
End Of Loop
Wire Turn
http://www.qsl.net/va3iul

---

97 - Morgain-Dipole Antenna for 160m and 80m

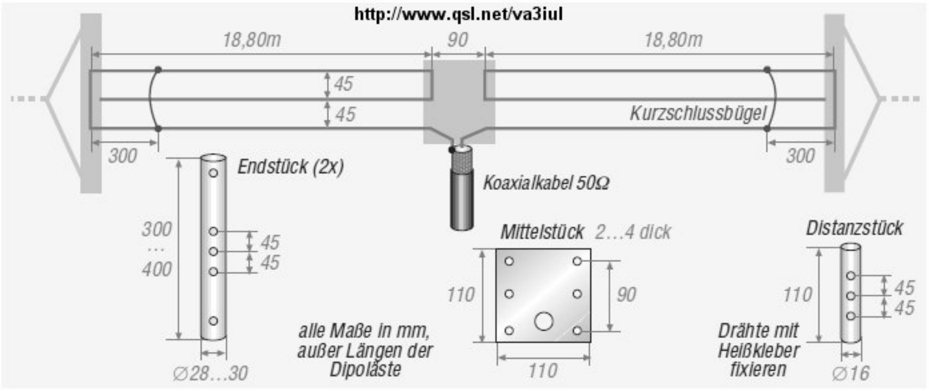
http://www.qsl.net/va3iul
18,80m
90
18,80m
45
45
Kurzschlussbügel
300
300
Endstück (2x)
Koaxialkabel 50Ω
300
...
400
45
45
Ø28...30
Mittelstück 2...4 dick
110
90
110
alle Maße in mm,
außer Längen der
Dipoläste
Distanzstück
110
45
45
Drähte mit
Heißkleber
fixieren
Ø16

---

98 - ZL-Special for 20m, 15m, 10m

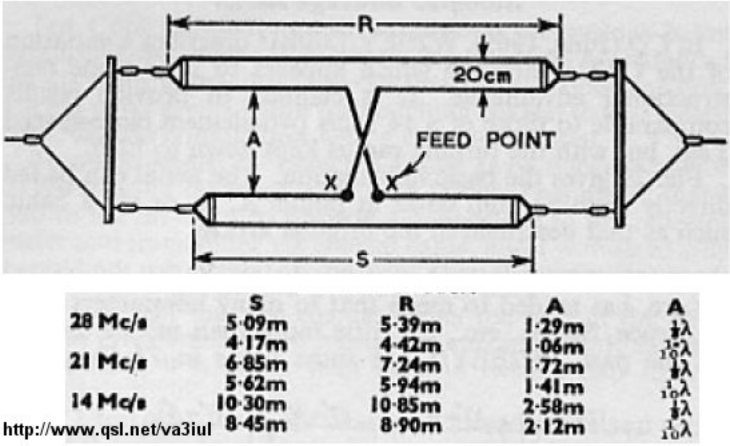
R
20cm
FEED POINT
A
X
X
S
http://www.qsl.net/va3iul

	S	R	A	A
28 Mc/s	5·09m	5·39m	1·29m	$\frac{1}{8}\lambda$
	4·17m	4·42m	1·06m	$\frac{1}{10}\lambda$
21 Mc/s	6·85m	7·24m	1·72m	$\frac{1}{8}\lambda$
	5·62m	5·94m	1·41m	$\frac{1}{10}\lambda$
14 Mc/s	10·30m	10·85m	2·58m	$\frac{1}{8}\lambda$
	8·45m	8·90m	2·12m	$\frac{1}{10}\lambda$

---

99 - Biconical Antenna

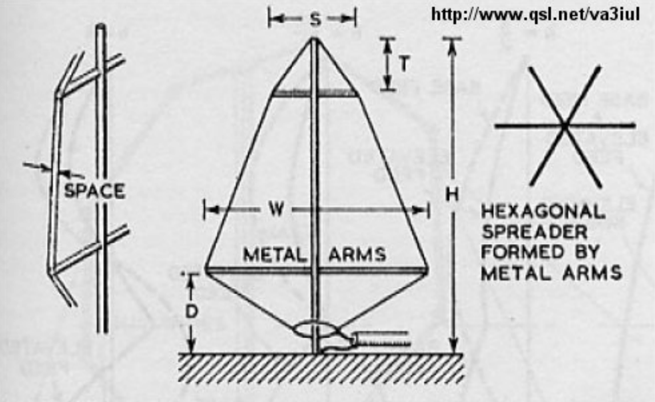
http://www.qsl.net/va3iul
S
T
H
SPACE
W
METAL ARMS
D
HEXAGONAL SPREADER FORMED BY METAL ARMS

	H	W	D	T	S	SPACE	WIRE
3·5 – 15 Mc/s	43′	17′-8″	16′-10″	2′-2″	5′-10″	3″	NR 8
7 - 28 Mc/s	23′	9′-6″	9′	12″	3′-2″	2″	NR10
14 - 56 Mc/s	12′	5′	4′-9″	8″	1′-8″	1″	NR12

100 - Directive Delta-Birdcage Antenna for 20m to 10m

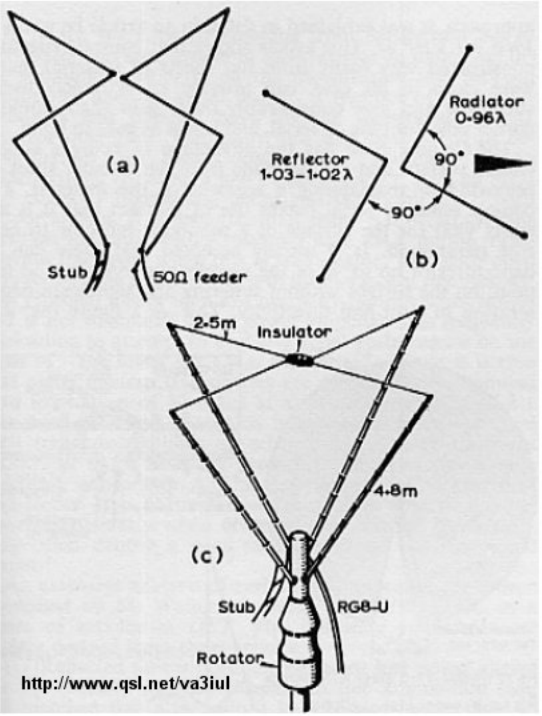
(a)
Stub
50Ω feeder
Reflector
1·03–1·02λ
Radiator
0·96λ
90°
90°
(b)
2·5m
Insulator
4·8m
(c)
Stub
RG8-U
Rotator
http://www.qsl.net/va3iul

101 - Dual Polarization Antenna for 80m and 40m

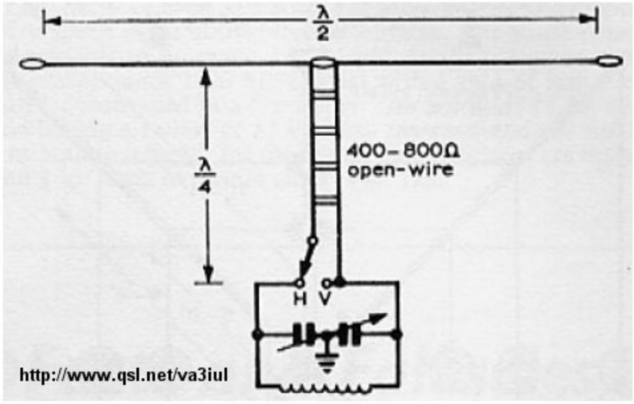
λ/2
λ/4
400–800Ω
open-wire
H
V
http://www.qsl.net/va3iul

102 - Directive 300-ohm-Ribbon Folded Dipole for 15m

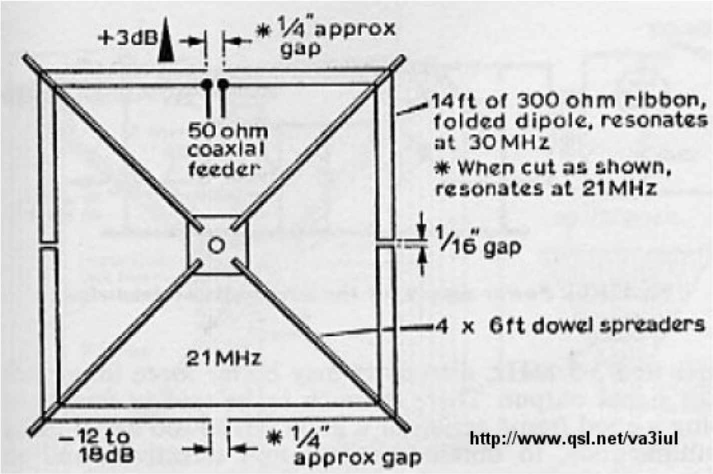
+3dB
* 1/4" approx gap
50 ohm coaxial feeder
14ft of 300 ohm ribbon, folded dipole, resonates at 30MHz
* When cut as shown, resonates at 21MHz
1/16" gap
4 x 6ft dowel spreaders
21MHz
-12 to 18dB
* 1/4" approx gap
http://www.qsl.net/va3iul

103 - Miniature Directive Antenna for 10m

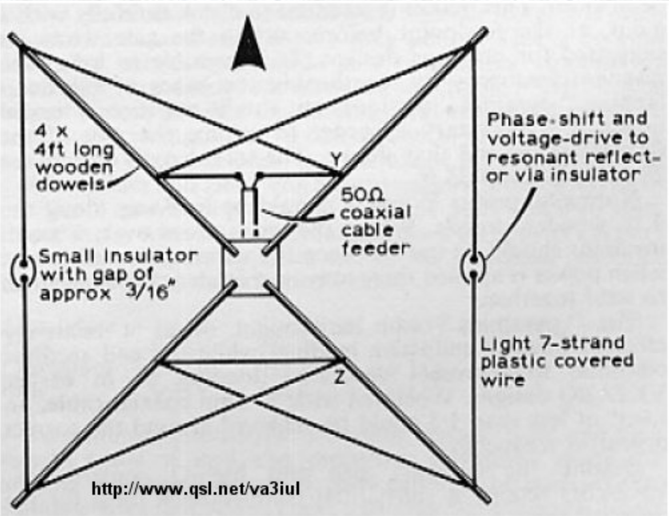
4 x 4ft long wooden dowels
Y
50Ω coaxial cable feeder
Phase-shift and voltage-drive to resonant reflector via insulator
Small insulator with gap of approx 3/16"
Light 7-strand plastic covered wire
Z
http://www.qsl.net/va3iul

104 - Biquad Antenna 12dBi-Gain for 2.4GHz

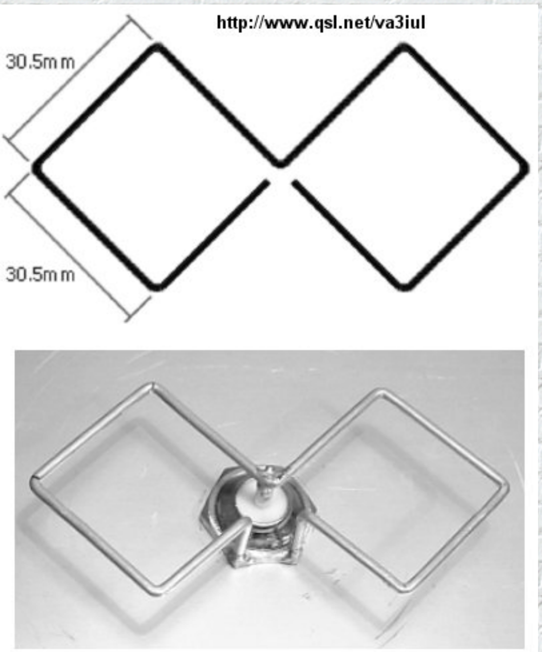
http://www.qsl.net/va3iul
30.5mm
30.5mm

105 - Dual-Rhomboid Antenna for 435MHz to 870MHz

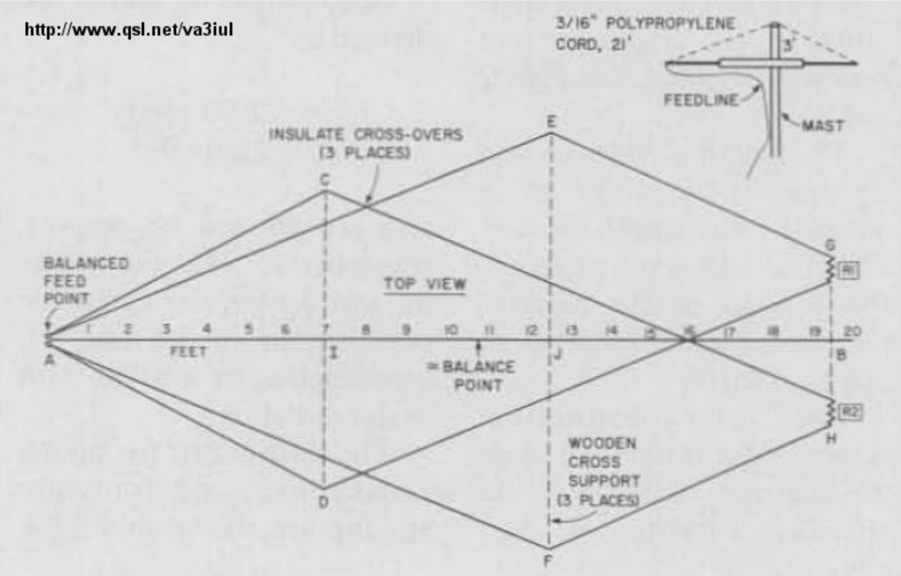
http://www.qsl.net/va3iul
3/16" POLYPROPYLENE CORD, 21'
3'
FEEDLINE
MAST
INSULATE CROSS-OVERS (3 PLACES)
E
C
G
R1
BALANCED FEED POINT
TOP VIEW
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
A
FEET
I
≈ BALANCE POINT
J
B
R2
H
WOODEN CROSS SUPPORT (3 PLACES)
D
F

Dual rhomboid antenna for 435-870 MHz. Beamwidth ≈ 10° H x 6° V. Gain over DP ≈ 26 dB. Boom length: AB = 19'6". Support spacing : AI = 7'; IJ = 5'6"; JB = 7'. Support length: CD = 7'3"; EF = 10'3"; GH = 3'0". Rhomboid sides: AC, AD, EG, FH = 7'9.5"; AE, AF, CH, DG = 13'5.5". Feedline: Wire needed: 14 AWG formvar, ≈ 86'0". Boom material: AI, JB = wood; IJ = metal. Cross support: CD, EF, GH = wood. Terminators: R1, R2 = 600 Ohms; Watts

---

106 - Double-Bazooka Antenna for 80m

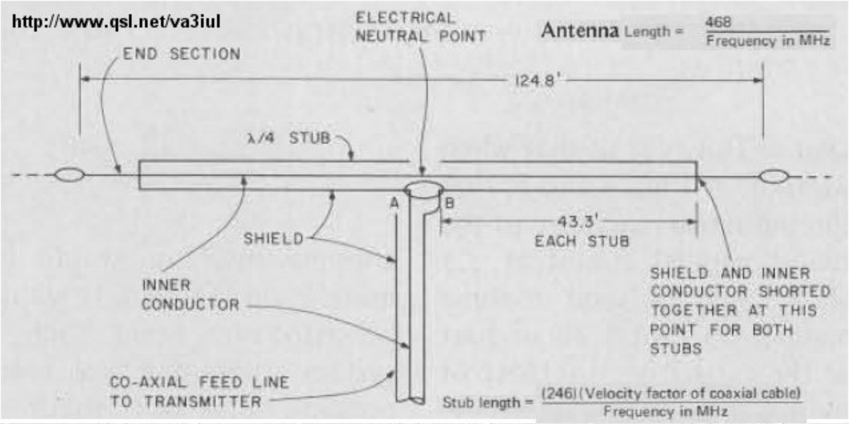
http://www.qsl.net/va3iul
END SECTION
ELECTRICAL NEUTRAL POINT
Antenna Length = 468 / Frequency in MHz
124.8'
λ/4 STUB
A
B
43.3' EACH STUB
SHIELD
INNER CONDUCTOR
SHIELD AND INNER CONDUCTOR SHORTED TOGETHER AT THIS POINT FOR BOTH STUBS
CO-AXIAL FEED LINE TO TRANSMITTER
Stub length = (246) (Velocity factor of coaxial cable) / Frequency in MHz

---

107 - J-Style Antenna

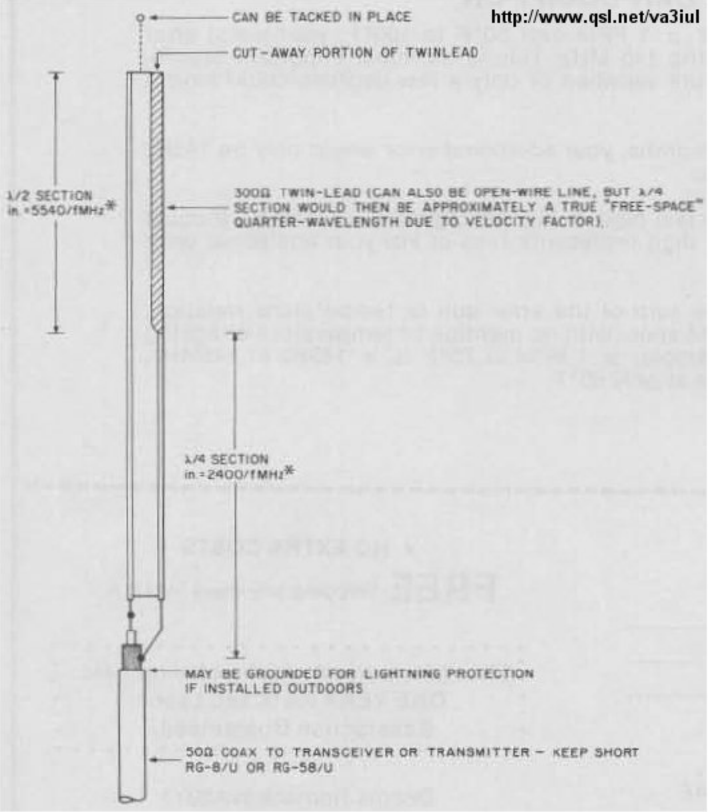
http://www.qsl.net/va3iul
CAN BE TACKED IN PLACE
CUT-AWAY PORTION OF TWINLEAD
λ/2 SECTION
in.=5540/fMHz*
300Ω TWIN-LEAD (CAN ALSO BE OPEN-WIRE LINE, BUT λ/4
SECTION WOULD THEN BE APPROXIMATELY A TRUE "FREE-SPACE"
QUARTER-WAVELENGTH DUE TO VELOCITY FACTOR).
λ/4 SECTION
in.=2400/fMHz*
MAY BE GROUNDED FOR LIGHTNING PROTECTION
IF INSTALLED OUTDOORS
50Ω COAX TO TRANSCEIVER OR TRANSMITTER - KEEP SHORT
RG-8/U OR RG-58/U

108 - Vertical / Horizontal / Circular Polarization Antenna

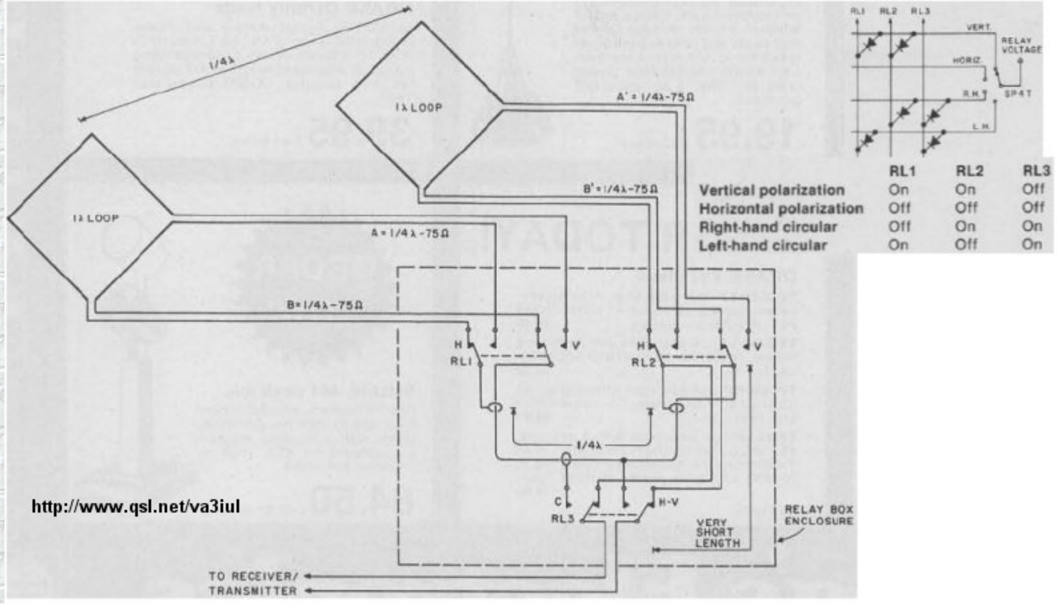
1/4λ
1λ LOOP
1λ LOOP
A' = 1/4λ-75Ω
B' = 1/4λ-75Ω
A = 1/4λ-75Ω
B = 1/4λ-75Ω
RL1 RL2 RL3
VERT.
RELAY VOLTAGE
HORIZ.
R.H.
SP4T
L.H.
H
V
RL1
H
V
RL2
1/4λ
C
H-V
RL3
VERY SHORT LENGTH
RELAY BOX ENCLOSURE
TO RECEIVER/ TRANSMITTER
http://www.qsl.net/va3iul

	RL1	RL2	RL3
Vertical polarization	On	On	Off
Horizontal polarization	Off	Off	Off
Right-hand circular	Off	On	On
Left-hand circular	On	Off	On

109 - Coax Inverted-L Antenna for 80m

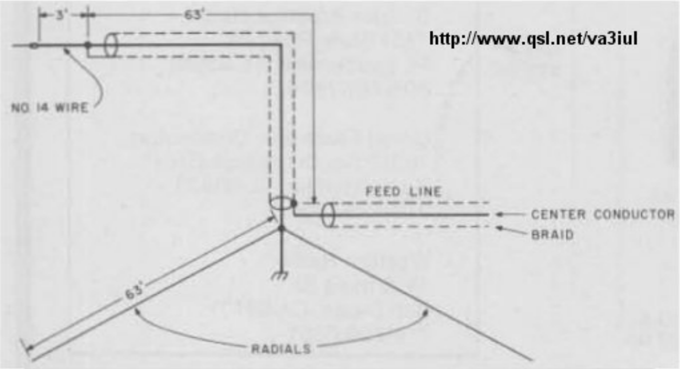
3'
63'
http://www.qsl.net/va3iul
NO 14 WIRE
FEED LINE
CENTER CONDUCTOR
BRAID
63'
RADIALS

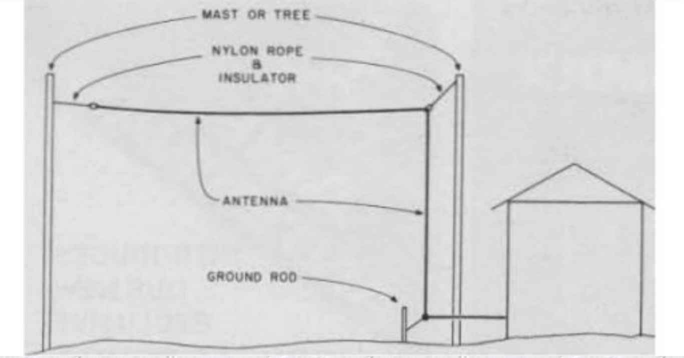
MAST OR TREE
NYLON ROPE
&
INSULATOR
ANTENNA
GROUND ROD

110 - Indoor Compact Loop Antenna for 80m

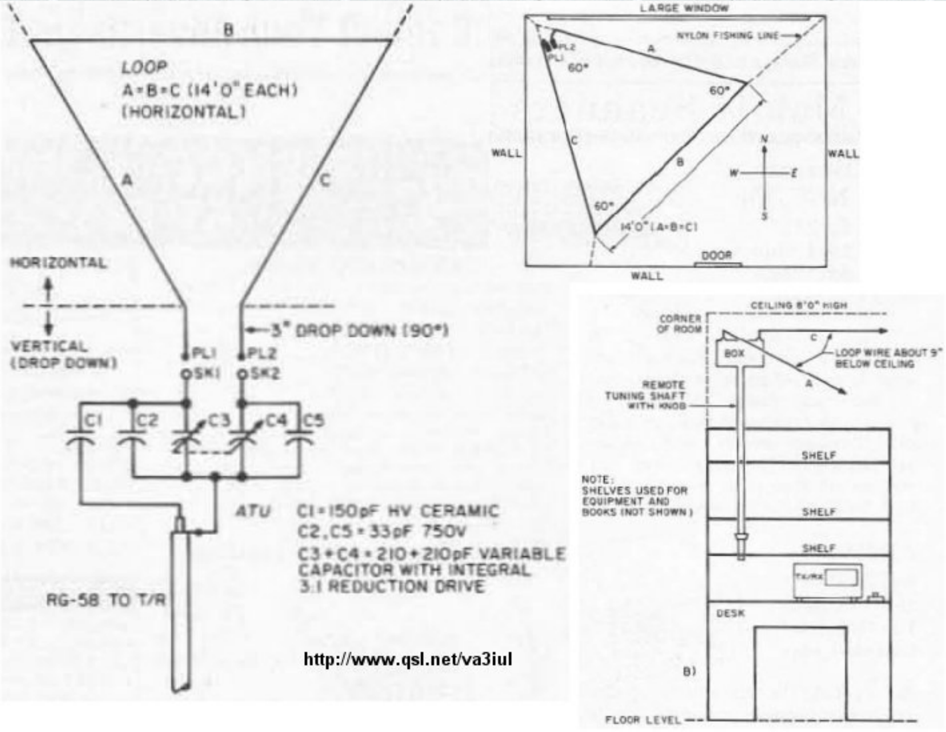
B
LOOP
A=B=C (14'0" EACH)
(HORIZONTAL)
A
C
HORIZONTAL
VERTICAL
(DROP DOWN)
3" DROP DOWN (90°)
PL1
PL2
SK1
SK2
C1
C2
C3
C4
C5
ATU
C1=150pF HV CERAMIC
C2,C5=33pF 750V
C3+C4=210+210pF VARIABLE
CAPACITOR WITH INTEGRAL
3:1 REDUCTION DRIVE
RG-58 TO T/R
http://www.qsl.net/va3iul
LARGE WINDOW
NYLON FISHING LINE
PL2
PL1
60°
60°
60°
WALL
WALL
WALL
N
W
E
S
14'0" (A=B=C)
DOOR
CEILING 8'0" HIGH
CORNER
OF ROOM
BOX
LOOP WIRE ABOUT 9"
BELOW CEILING
REMOTE
TUNING SHAFT
WITH KNOB
SHELF
SHELF
SHELF
NOTE:
SHELVES USED FOR
EQUIPMENT AND
BOOKS (NOT SHOWN)
TX/RX
DESK
B)
FLOOR LEVEL

111 - Helix Antenna

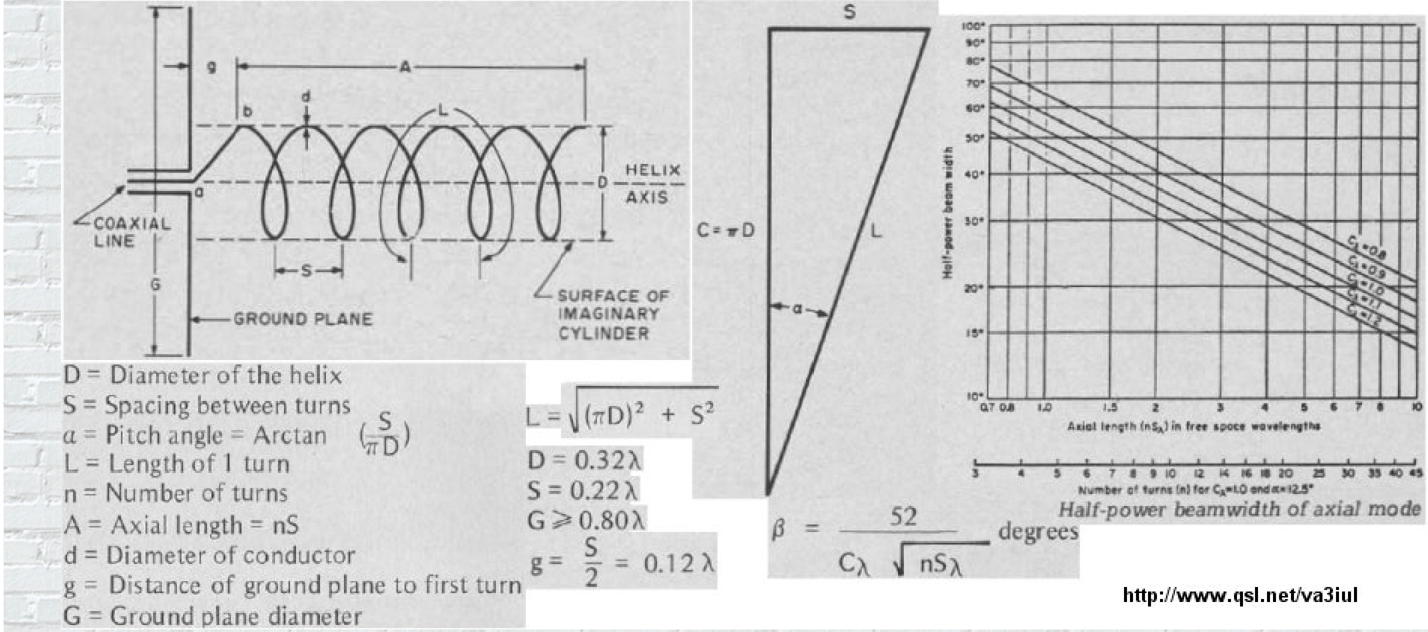
g
A
d
b
L
HELIX AXIS
D
a
COAXIAL LINE
S
G
GROUND PLANE
SURFACE OF IMAGINARY CYLINDER
S
C = πD
L
α
Half-power beam width
100°
90°
80°
70°
60°
50°
40°
30°
20°
15°
10°
0.7 0.8 1.0 1.5 2 3 4 5 6 7 8 10
Axial length (nS_λ) in free space wavelengths
3 4 5 6 7 8 9 10 12 14 16 18 20 25 30 35 40 45
Number of turns (n) for C_λ=1.0 and α=12.5°
C_λ=0.8
C_λ=0.9
C_λ=1.0
C_λ=1.1
C_λ=1.2

D = Diameter of the helix
S = Spacing between turns
α = Pitch angle = Arctan $\left(\frac{S}{\pi D}\right)$
L = Length of 1 turn
n = Number of turns
A = Axial length = nS
d = Diameter of conductor
g = Distance of ground plane to first turn
G = Ground plane diameter

$$L = \sqrt{(\pi D)^2 + S^2}$$

$D = 0.32\lambda$
$S = 0.22\lambda$
$G \geqslant 0.80\lambda$
$g = \frac{S}{2} = 0.12\lambda$

$$\beta = \frac{52}{C_\lambda \sqrt{nS_\lambda}} \text{ degrees}$$

Half-power beamwidth of axial mode

http://www.qsl.net/va3iul

---

112 - Novice Vertical Antenna for 80m, 40m, 15m, 10m

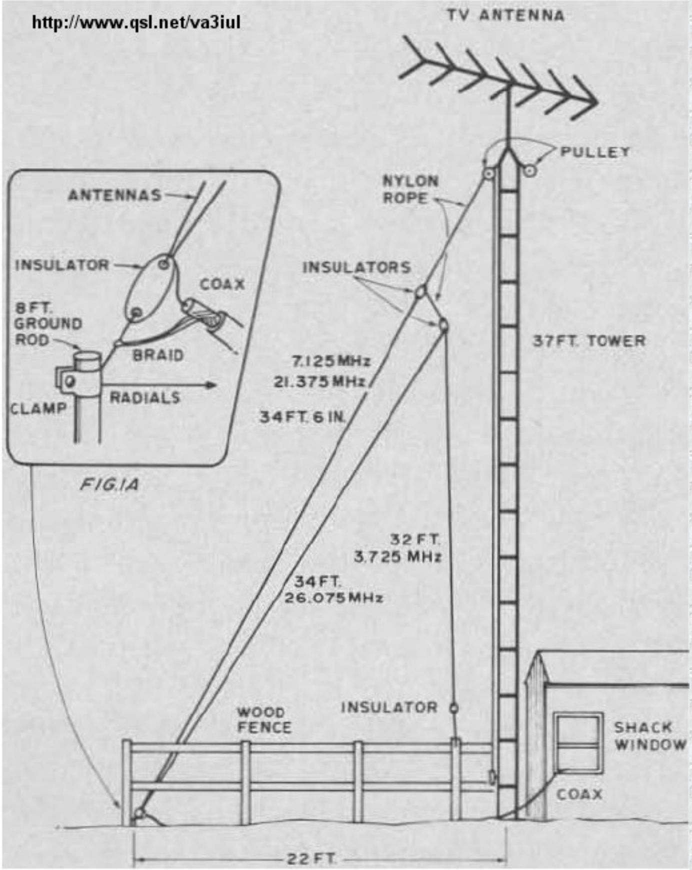
http://www.qsl.net/va3iul
TV ANTENNA
PULLEY
NYLON ROPE
INSULATORS
37 FT. TOWER
7.125 MHz
21.375 MHz
34 FT. 6 IN.
32 FT.
3.725 MHz
34 FT.
26.075 MHz
INSULATOR
WOOD FENCE
SHACK WINDOW
COAX
22 FT.
ANTENNAS
INSULATOR
COAX
8 FT. GROUND ROD
BRAID
RADIALS
CLAMP
FIG. 1A

---

113 - Stub-Loaded Shortened Dipole for 80m

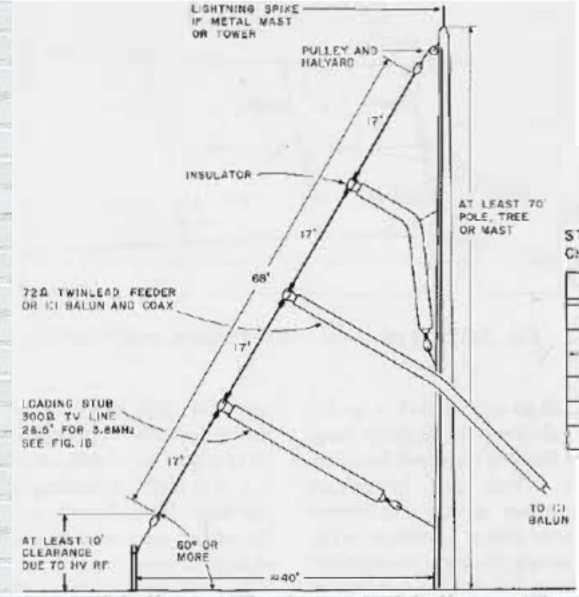
LIGHTNING SPIKE IF METAL MAST OR TOWER
PULLEY AND HALYARD
17'
INSULATOR
AT LEAST 70' POLE, TREE OR MAST
17'
68'
72Ω TWINLEAD FEEDER OR 1:1 BALUN AND COAX
17'
LOADING STUB 300Ω TV LINE 26.5' FOR 3.8MHz SEE FIG. 1B
17'
AT LEAST 10' CLEARANCE DUE TO HV RF
60° OR MORE
≈40'
TO 1:1 BALUN

http://www.qsl.net/va3iul

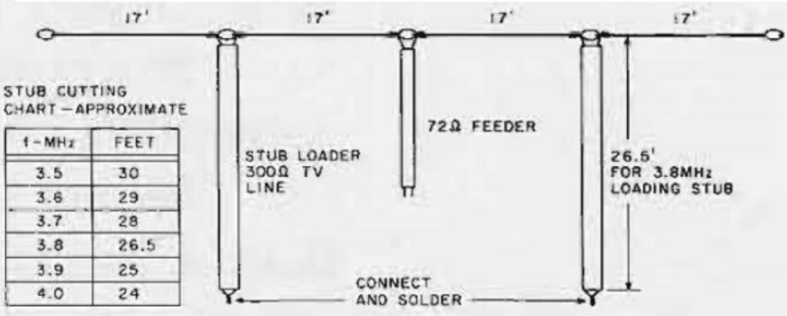
17'
17'
17'
17'
STUB LOADER 300Ω TV LINE
72Ω FEEDER
26.5' FOR 3.8MHz LOADING STUB
CONNECT AND SOLDER

STUB CUTTING CHART – APPROXIMATE

f-MHz	FEET
3.5	30
3.6	29
3.7	28
3.8	26.5
3.9	25
4.0	24

114 - Six-Band Wire-Stub Trap Antenna for 40m-10m

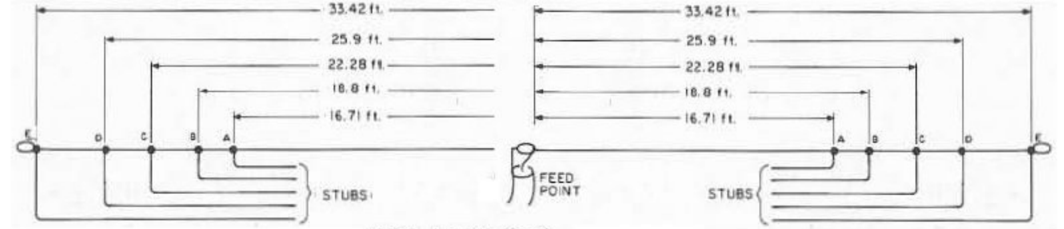
33.42 ft.
25.9 ft.
22.28 ft.
18.8 ft.
16.71 ft.
33.42 ft.
25.9 ft.
22.28 ft.
18.8 ft.
16.71 ft.
E
D
C
B
A
STUBS
FEED POINT
STUBS
A
B
C
D
E

Point	Length (feet)
A	8.37
B	9.40
C	11.14
D	12.95
E	16.72

40m
20m
17m
15m
12m
10m

http://www.qsl.net/va3iul

115 - Multiband Half-Wave Delta-Loop Antenna

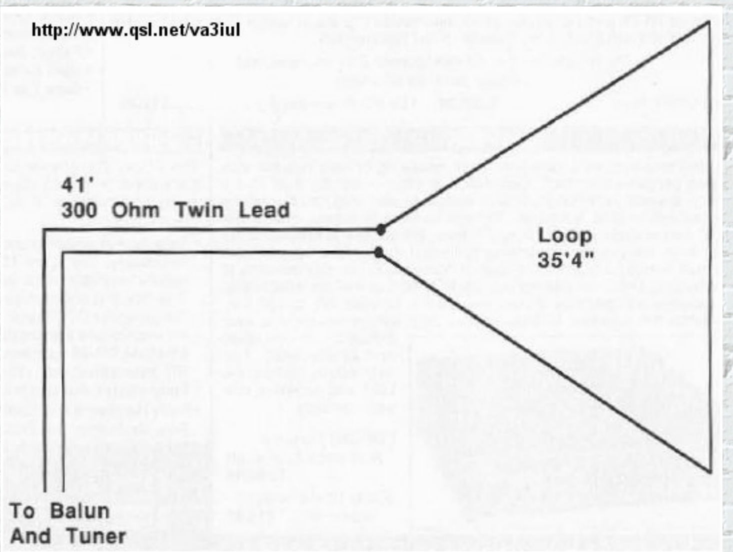
http://www.qsl.net/va3iul
41'
300 Ohm Twin Lead
Loop
35'4"
To Balun
And Tuner

116 - Hybrid Vee for 20m, 17m

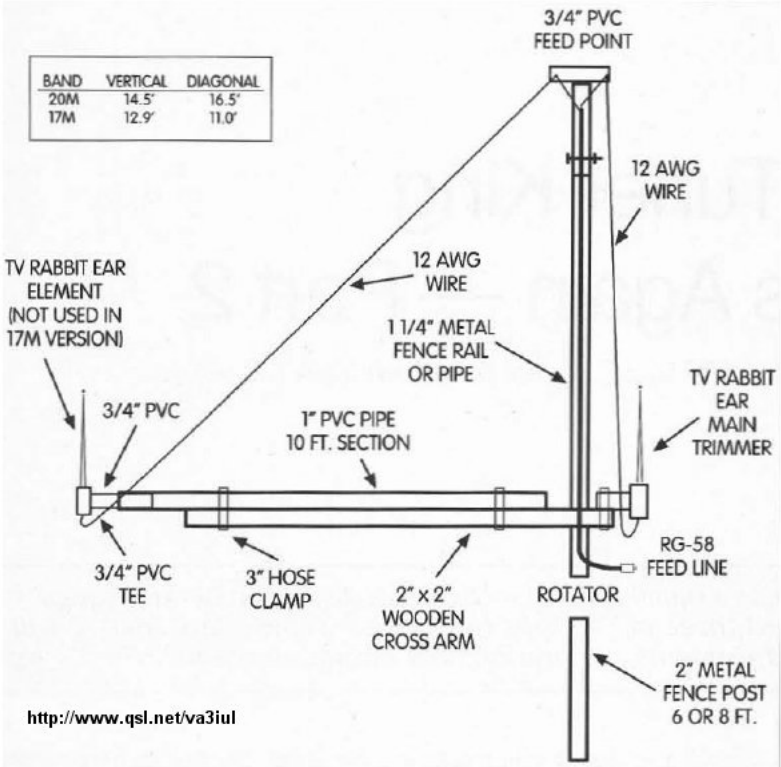
3/4" PVC
FEED POINT
BAND VERTICAL DIAGONAL
20M 14.5' 16.5'
17M 12.9' 11.0'
12 AWG
WIRE
12 AWG
WIRE
TV RABBIT EAR
ELEMENT
(NOT USED IN
17M VERSION)
1 1/4" METAL
FENCE RAIL
OR PIPE
TV RABBIT
EAR
MAIN
TRIMMER
3/4" PVC
1" PVC PIPE
10 FT. SECTION
RG-58
FEED LINE
3/4" PVC
TEE
3" HOSE
CLAMP
2" x 2"
WOODEN
CROSS ARM
ROTATOR
2" METAL
FENCE POST
6 OR 8 FT.
http://www.qsl.net/va3iul

117 - Six-Shooter Array Antenna, Gain=7.5dB

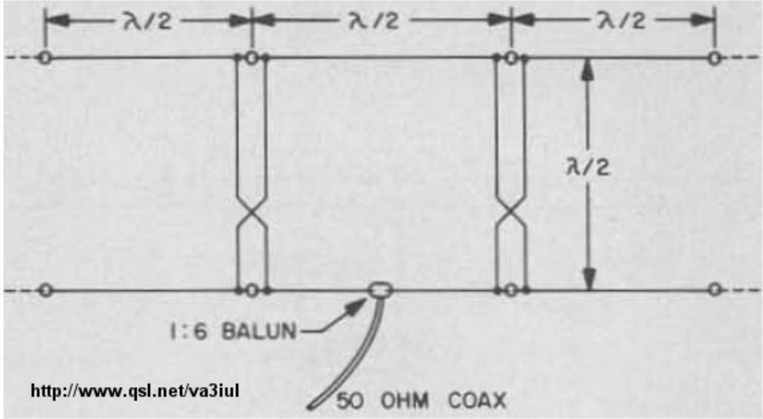
λ/2
λ/2
λ/2
λ/2
1:6 BALUN
50 OHM COAX
http://www.qsl.net/va3iul

118 - Multiband Ground-Plane Antenna for 10m, 15m, 20m

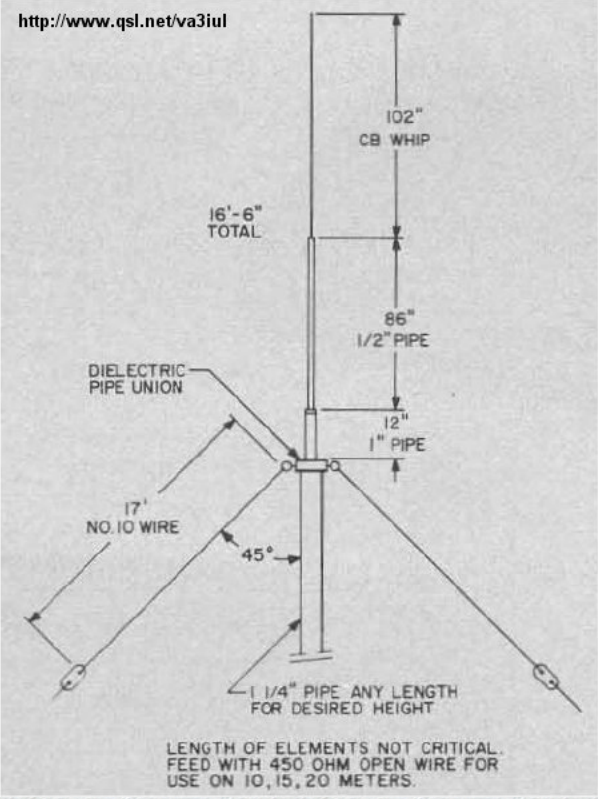
http://www.qsl.net/va3iul
102"
CB WHIP
16'-6"
TOTAL
86"
1/2" PIPE
DIELECTRIC
PIPE UNION
12"
1" PIPE
17'
NO. 10 WIRE
45°
1 1/4" PIPE ANY LENGTH
FOR DESIRED HEIGHT

LENGTH OF ELEMENTS NOT CRITICAL. FEED WITH 450 OHM OPEN WIRE FOR USE ON 10, 15, 20 METERS.

119 - Wire Superbeam Antenna for 10m, 15m, 20m

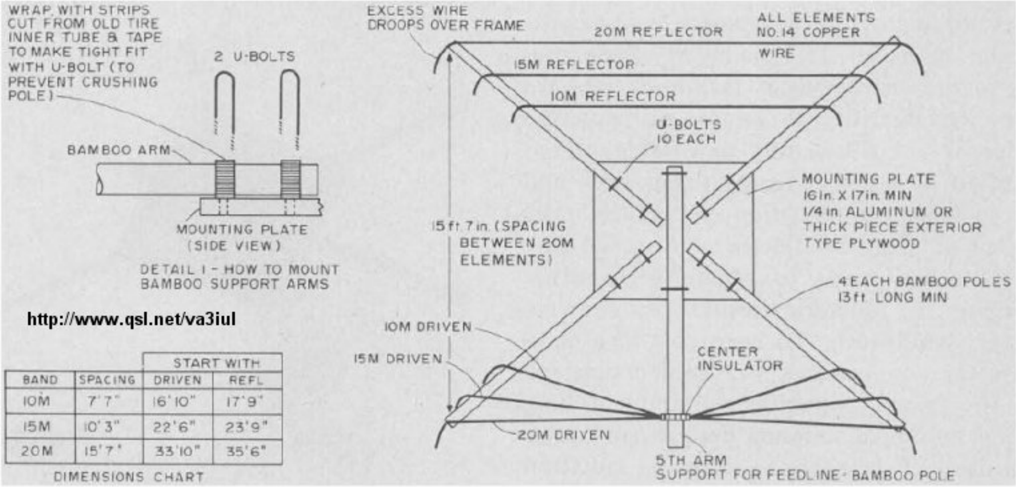
WRAP WITH STRIPS CUT FROM OLD TIRE INNER TUBE & TAPE TO MAKE TIGHT FIT WITH U-BOLT (TO PREVENT CRUSHING POLE)
2 U-BOLTS
BAMBOO ARM
MOUNTING PLATE (SIDE VIEW)
DETAIL 1 - HOW TO MOUNT BAMBOO SUPPORT ARMS
http://www.qsl.net/va3iul
EXCESS WIRE DROOPS OVER FRAME
20M REFLECTOR
ALL ELEMENTS NO. 14 COPPER WIRE
15M REFLECTOR
10M REFLECTOR
U-BOLTS 10 EACH
MOUNTING PLATE 16in. X 17in. MIN 1/4 in. ALUMINUM OR THICK PIECE EXTERIOR TYPE PLYWOOD
15 ft. 7 in. (SPACING BETWEEN 20M ELEMENTS)
4 EACH BAMBOO POLES 13 ft LONG MIN
10M DRIVEN
15M DRIVEN
CENTER INSULATOR
20M DRIVEN
5TH ARM SUPPORT FOR FEEDLINE - BAMBOO POLE

		START WITH	
BAND	SPACING	DRIVEN	REFL
10M	7'7"	16'10"	17'9"
15M	10'3"	22'6"	23'9"
20M	15'7"	33'10"	35'6"

DIMENSIONS CHART

120 - Two Elements Delta-Loop Antenna

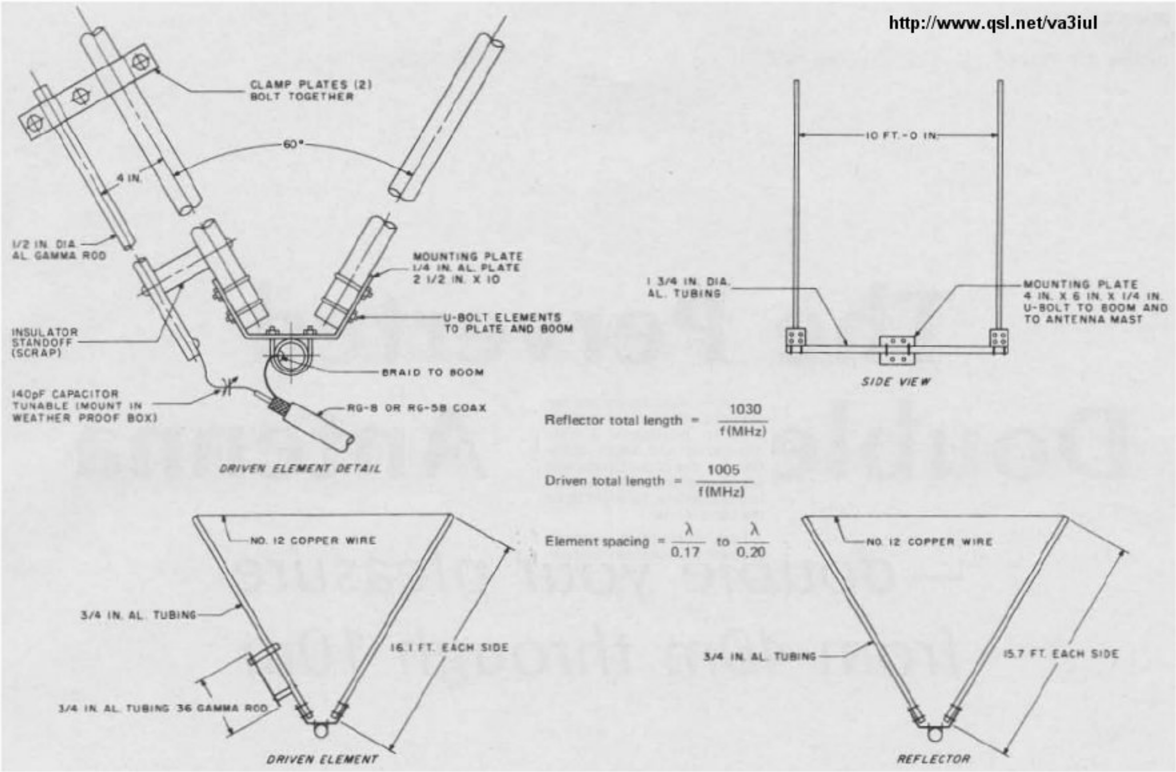
http://www.qsl.net/va3iul
CLAMP PLATES (2) BOLT TOGETHER
60°
4 IN.
1/2 IN. DIA AL. GAMMA ROD
MOUNTING PLATE 1/4 IN. AL. PLATE 2 1/2 IN. X 10
INSULATOR STANDOFF (SCRAP)
U-BOLT ELEMENTS TO PLATE AND BOOM
BRAID TO BOOM
140pF CAPACITOR TUNABLE (MOUNT IN WEATHER PROOF BOX)
RG-8 OR RG-58 COAX
DRIVEN ELEMENT DETAIL
10 FT.-0 IN.
1 3/4 IN. DIA. AL. TUBING
MOUNTING PLATE 4 IN. X 6 IN. X 1/4 IN. U-BOLT TO BOOM AND TO ANTENNA MAST
SIDE VIEW
Reflector total length = 1030 / f(MHz)
Driven total length = 1005 / f(MHz)
Element spacing = λ / 0.17 to λ / 0.20
NO. 12 COPPER WIRE
3/4 IN. AL. TUBING
16.1 FT. EACH SIDE
3/4 IN. AL. TUBING 36 GAMMA ROD
DRIVEN ELEMENT
NO. 12 COPPER WIRE
3/4 IN. AL. TUBING
15.7 FT. EACH SIDE
REFLECTOR

121 - Sterba Curtain Antenna

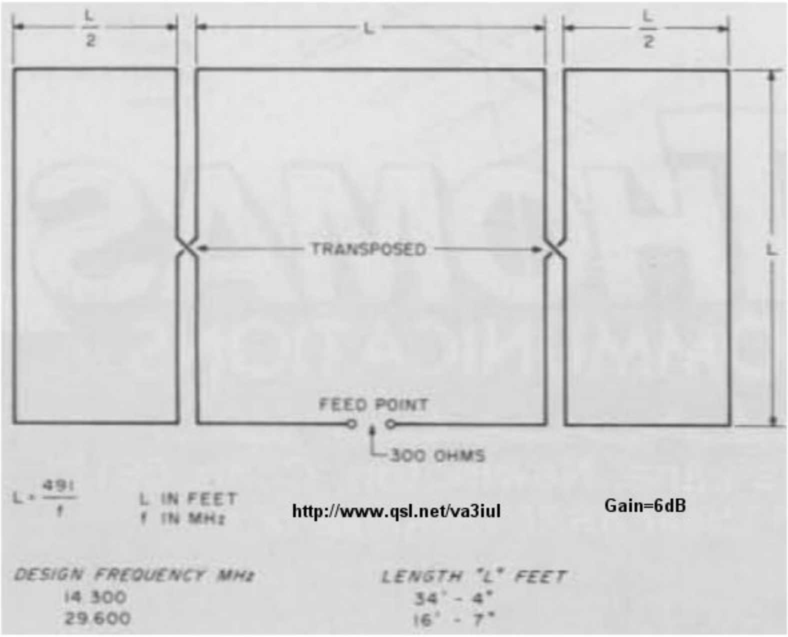
L/2
L
L/2
TRANSPOSED
L
FEED POINT
300 OHMS
L = 491 / f
L IN FEET
f IN MHz
http://www.qsl.net/va3iul
Gain=6dB
DESIGN FREQUENCY MHz
14.300
29.600
LENGTH "L" FEET
34' - 4"
16' - 7"

122 - Half-Wave Vertical Zepp Antenna

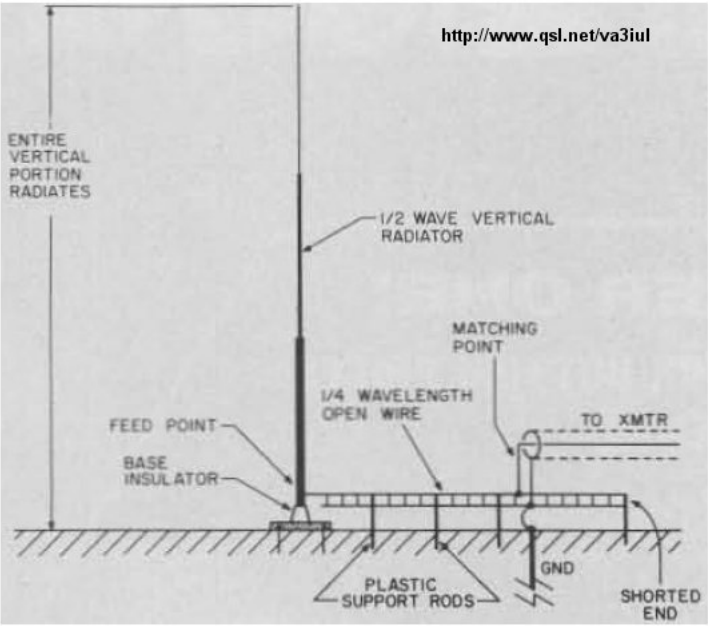
http://www.qsl.net/va3iul
ENTIRE VERTICAL PORTION RADIATES
1/2 WAVE VERTICAL RADIATOR
MATCHING POINT
1/4 WAVELENGTH OPEN WIRE
TO XMTR
FEED POINT
BASE INSULATOR
GND
PLASTIC SUPPORT RODS
SHORTED END

123 - Lazy-Loop Antenna for 40m

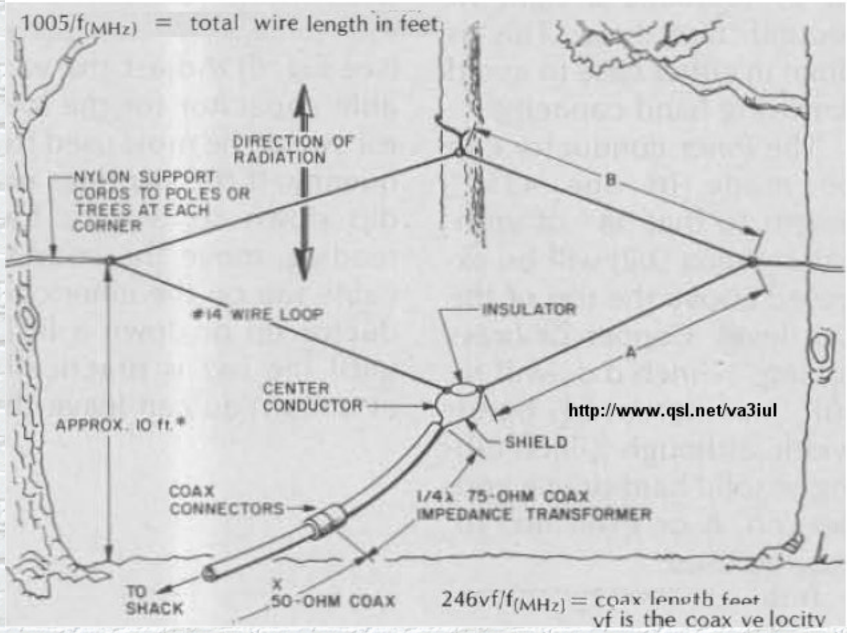
1005/f(MHz) = total wire length in feet
DIRECTION OF RADIATION
NYLON SUPPORT CORDS TO POLES OR TREES AT EACH CORNER
B
#14 WIRE LOOP
INSULATOR
A
CENTER CONDUCTOR
http://www.qsl.net/va3iul
SHIELD
APPROX. 10 ft.*
COAX CONNECTORS
1/4λ 75-OHM COAX IMPEDANCE TRANSFORMER
TO SHACK
X
50-OHM COAX
246vf/f(MHz) = coax length feet
vf is the coax velocity

124 - Terminated Folded Dipole for 80m, 40m

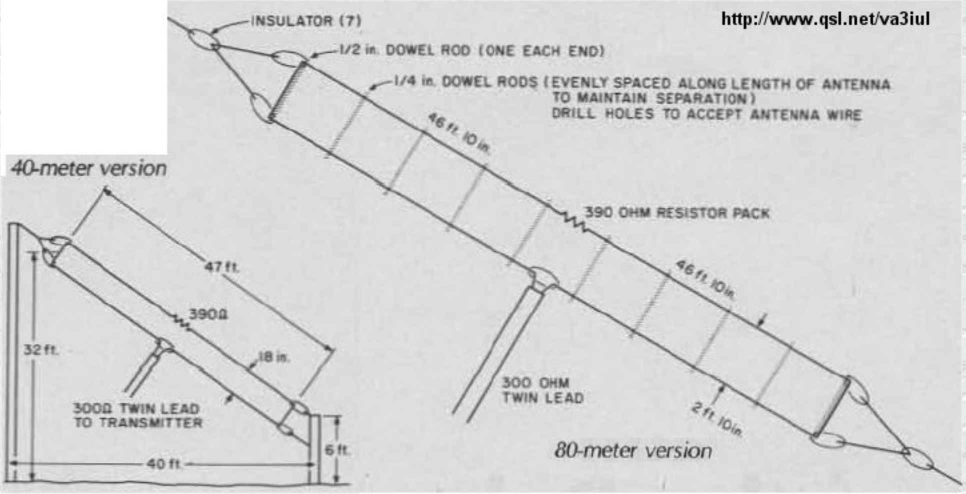
INSULATOR (7)
http://www.qsl.net/va3iul
1/2 in. DOWEL ROD (ONE EACH END)
1/4 in. DOWEL RODS (EVENLY SPACED ALONG LENGTH OF ANTENNA TO MAINTAIN SEPARATION) DRILL HOLES TO ACCEPT ANTENNA WIRE
46 ft. 10 in.
40-meter version
390 OHM RESISTOR PACK
47 ft.
390Ω
46 ft. 10 in.
32 ft.
18 in.
300Ω TWIN LEAD TO TRANSMITTER
300 OHM TWIN LEAD
2 ft. 10 in.
6 ft.
40 ft.
80-meter version

125 - Short-Fat Antenna for 15m

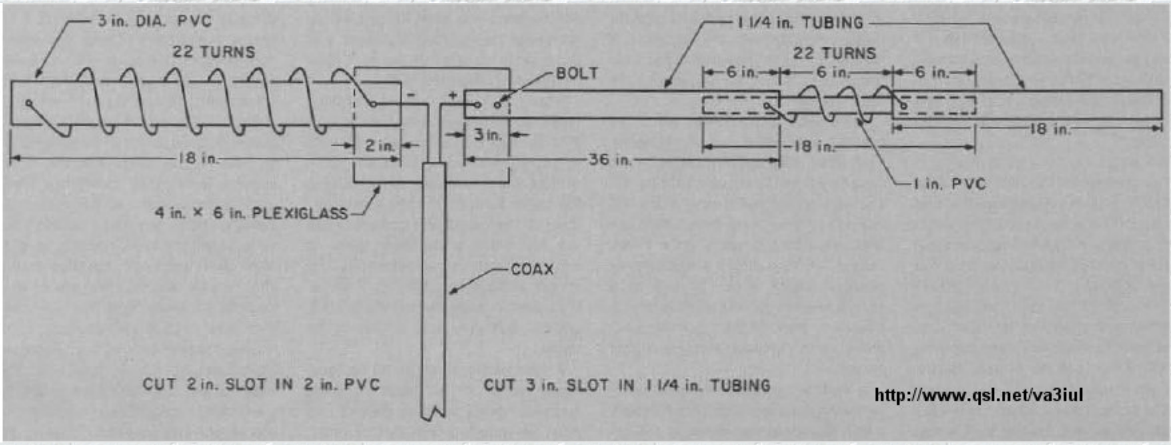
3 in. DIA. PVC
22 TURNS
18 in.
2 in.
4 in. × 6 in. PLEXIGLASS
BOLT
–
+
3 in.
36 in.
1 1/4 in. TUBING
22 TURNS
6 in.
6 in.
6 in.
18 in.
18 in.
1 in. PVC
COAX
CUT 2 in. SLOT IN 2 in. PVC
CUT 3 in. SLOT IN 1 1/4 in. TUBING
http://www.qsl.net/va3iul

126 - Cobra Antenna for 80m

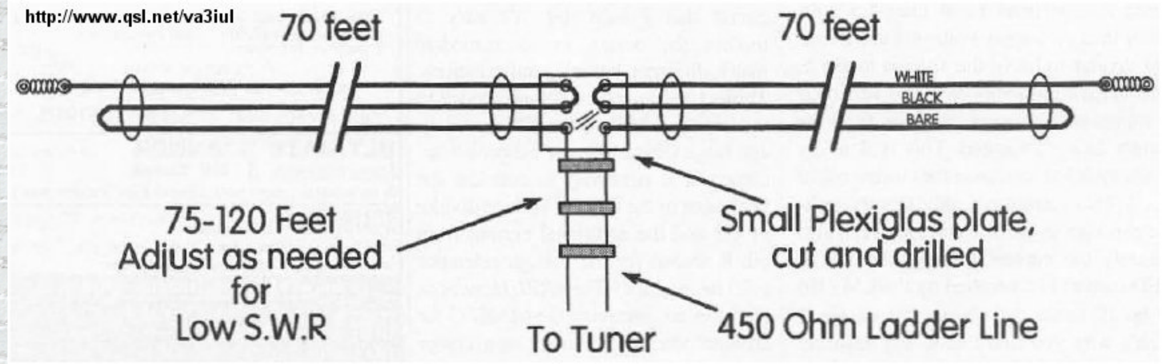
http://www.qsl.net/va3iul
70 feet
70 feet
WHITE
BLACK
BARE
75-120 Feet
Adjust as needed
for
Low S.W.R
To Tuner
Small Plexiglas plate,
cut and drilled
450 Ohm Ladder Line

127 - Log-Periodic Wire Antenna for 20m, 15m, 10m

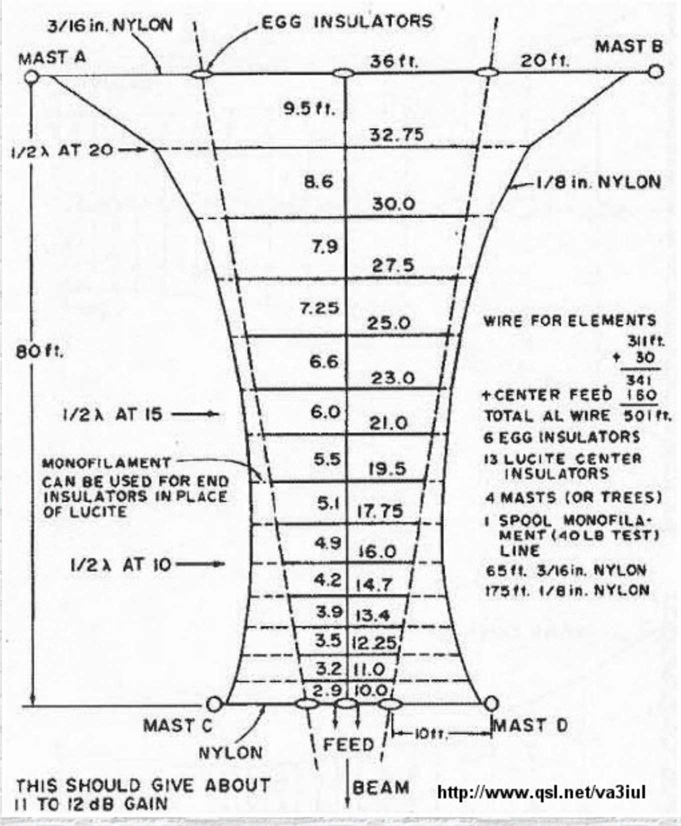
3/16 in. NYLON
EGG INSULATORS
MAST A
36 ft.
20 ft.
MAST B
9.5 ft.
32.75
1/2 λ AT 20
8.6
30.0
1/8 in. NYLON
7.9
27.5
7.25
25.0
WIRE FOR ELEMENTS
311 ft.
+ 30
341
+ CENTER FEED 160
TOTAL AL WIRE 501 ft.
6 EGG INSULATORS
13 LUCITE CENTER INSULATORS
4 MASTS (OR TREES)
1 SPOOL MONOFILAMENT (40 LB TEST) LINE
65 ft. 3/16 in. NYLON
175 ft. 1/8 in. NYLON
80 ft.
6.6
23.0
1/2 λ AT 15
6.0
21.0
MONOFILAMENT CAN BE USED FOR END INSULATORS IN PLACE OF LUCITE
5.5
19.5
5.1
17.75
4.9
16.0
1/2 λ AT 10
4.2
14.7
3.9
13.4
3.5
12.25
3.2
11.0
2.9
10.0
MAST C
NYLON
FEED
10 ft.
MAST D
BEAM
THIS SHOULD GIVE ABOUT 11 TO 12 dB GAIN
http://www.qsl.net/va3iul

128 - 5-Element Log-Periodic Vertical Antenna for 80m, 40m, 20m

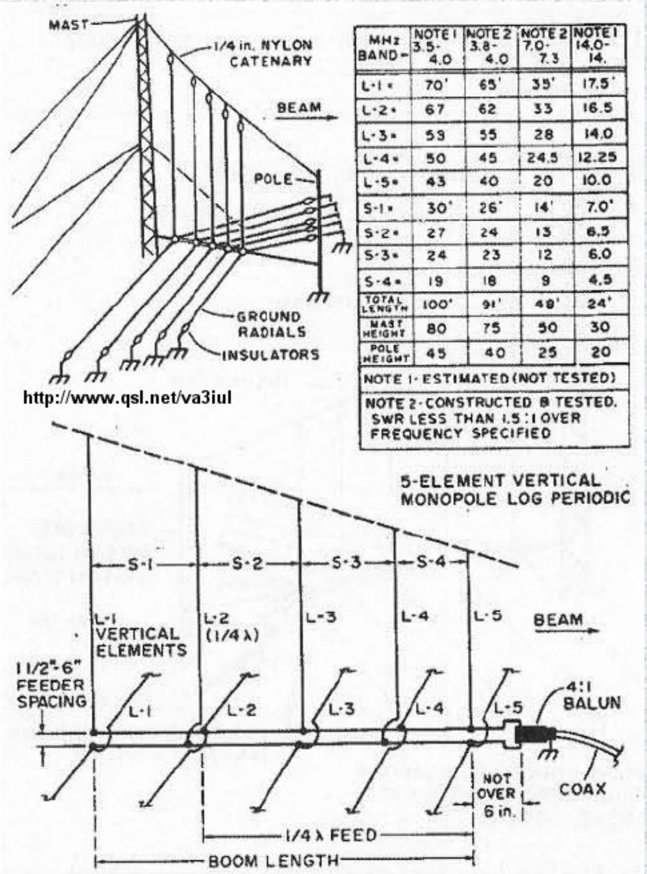
MAST
1/4 in. NYLON CATENARY
BEAM
POLE
GROUND RADIALS
INSULATORS
S-1
S-2
S-3
S-4
L-1 VERTICAL ELEMENTS
L-2 (1/4 λ)
L-3
L-4
L-5
BEAM
1 1/2"-6" FEEDER SPACING
L-1
L-2
L-3
L-4
L-5
4:1 BALUN
COAX
NOT OVER 6 in.
1/4 λ FEED
BOOM LENGTH

http://www.qsl.net/va3iul

MHz BAND –	NOTE 1 3.5-4.0	NOTE 2 3.8-4.0	NOTE 2 7.0-7.3	NOTE 1 14.0-14.
L-1 =	70'	65'	35'	17.5'
L-2 =	67	62	33	16.5
L-3 =	53	55	28	14.0
L-4 =	50	45	24.5	12.25
L-5 =	43	40	20	10.0
S-1 =	30'	26'	14'	7.0'
S-2 =	27	24	13	6.5
S-3 =	24	23	12	6.0
S-4 =	19	18	9	4.5
TOTAL LENGTH	100'	91'	48'	24'
MAST HEIGHT	80	75	50	30
POLE HEIGHT	45	40	25	20

NOTE 1 - ESTIMATED (NOT TESTED)

NOTE 2 - CONSTRUCTED & TESTED. SWR LESS THAN 1.5:1 OVER FREQUENCY SPECIFIED

5-ELEMENT VERTICAL MONOPOLE LOG PERIODIC

129 - 2m Vertical Wire Antennas

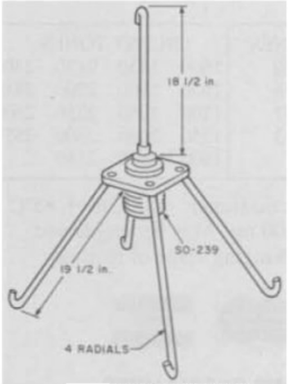
18 1/2 in.
SO-239
19 1/2 in.
4 RADIALS

Simple 1/4λ 2-meter ground-plane antenna.

http://www.qsl.net/va3iul

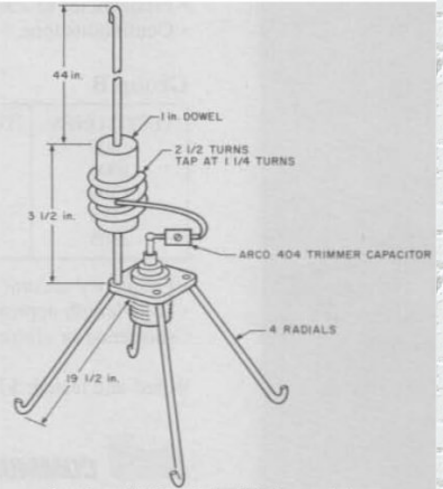
44 in.
1 in. DOWEL
2 1/2 TURNS TAP AT 1 1/4 TURNS
3 1/2 in.
ARCO 404 TRIMMER CAPACITOR
4 RADIALS
19 1/2 in.

Experimental 5/8λ vertical with 1/4λ ground plane.

130 - Earth-Mover Inverted-V Antenna for 40m

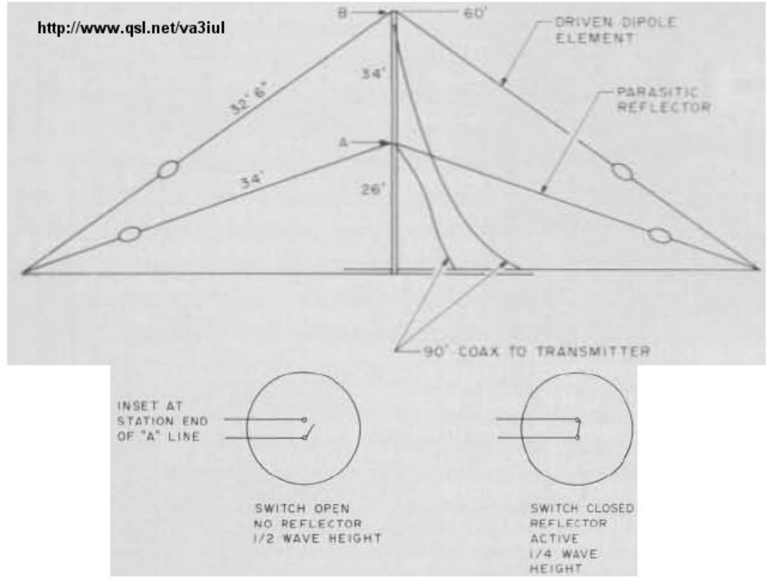
http://www.qsl.net/va3iul
B
60'
DRIVEN DIPOLE ELEMENT
34'
32' 6"
PARASITIC REFLECTOR
A
34'
26'
90' COAX TO TRANSMITTER
INSET AT STATION END OF "A" LINE
SWITCH OPEN NO REFLECTOR 1/2 WAVE HEIGHT
SWITCH CLOSED REFLECTOR ACTIVE 1/4 WAVE HEIGHT

131 - Coax-Cable Collinear Antennas

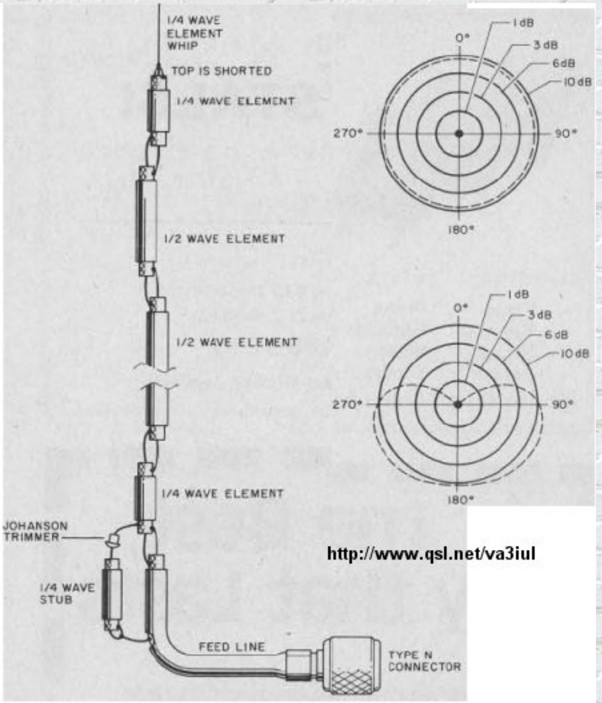
1/4 WAVE ELEMENT WHIP
TOP IS SHORTED
1/4 WAVE ELEMENT
1/2 WAVE ELEMENT
1/2 WAVE ELEMENT
1/4 WAVE ELEMENT
JOHANSON TRIMMER
1/4 WAVE STUB
FEED LINE
TYPE N CONNECTOR
0°
90°
180°
270°
1 dB
3 dB
6 dB
10 dB
0°
90°
180°
270°
1 dB
3 dB
6 dB
10 dB
http://www.qsl.net/va3iul

8 elements = 6 dB, 16 elements = 9 dB

132 - Double Bobtail Antenna for 20m

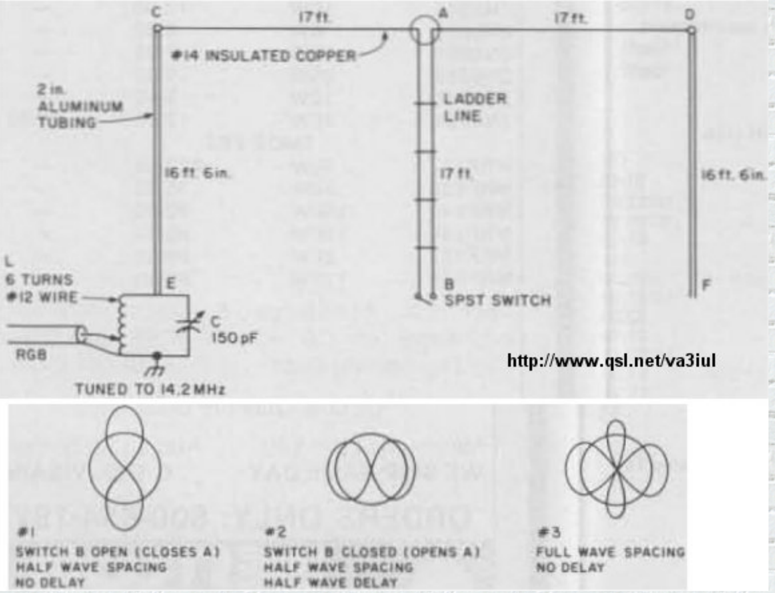
C
17 ft.
A
17 ft.
D
#14 INSULATED COPPER
2 in.
ALUMINUM
TUBING
LADDER
LINE
16 ft. 6in.
17 ft.
16 ft. 6in.
L
6 TURNS
#12 WIRE
E
B
SPST SWITCH
F
C
150 pF
RG8
TUNED TO 14.2 MHz
http://www.qsl.net/va3iul
#1
SWITCH B OPEN (CLOSES A)
HALF WAVE SPACING
NO DELAY
#2
SWITCH B CLOSED (OPENS A)
HALF WAVE SPACING
HALF WAVE DELAY
#3
FULL WAVE SPACING
NO DELAY

133 - Collinear Zepp Antenna

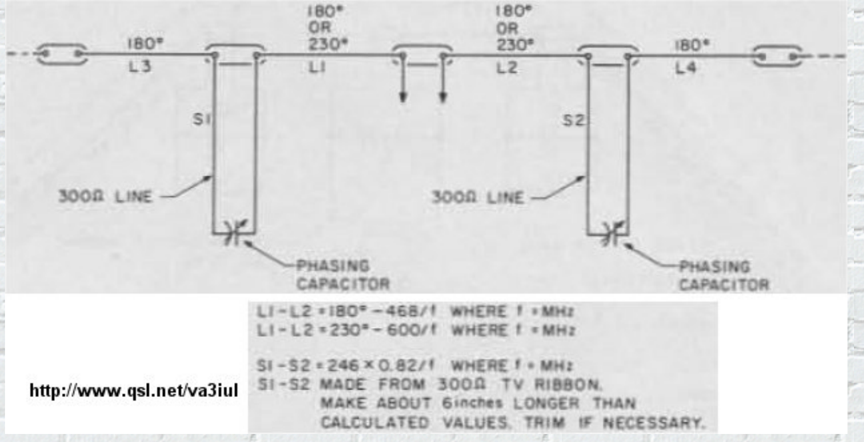
180°
L3
180°
OR
230°
L1
180°
OR
230°
L2
180°
L4
S1
S2
300Ω LINE
300Ω LINE
PHASING
CAPACITOR
PHASING
CAPACITOR
http://www.qsl.net/va3iul

L1-L2 = 180° - 468/f WHERE f = MHz
L1-L2 = 230° - 600/f WHERE f = MHz

S1-S2 = 246 × 0.82/f WHERE f = MHz
S1-S2 MADE FROM 300Ω TV RIBBON.
MAKE ABOUT 6inches LONGER THAN
CALCULATED VALUES. TRIM IF NECESSARY.

134 - Taylor Vee Antenna for 20m

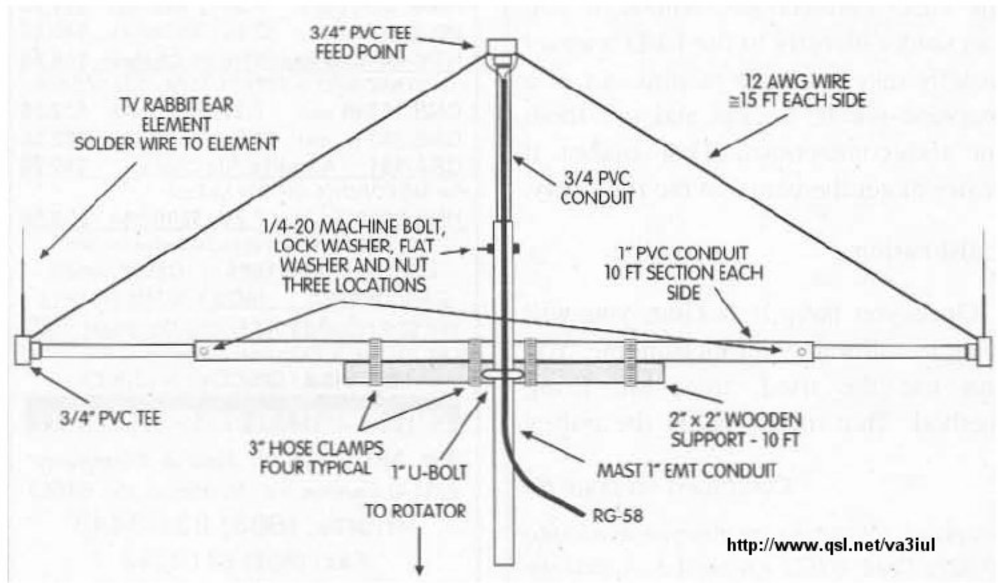
3/4" PVC TEE
FEED POINT
12 AWG WIRE
≅15 FT EACH SIDE
TV RABBIT EAR
ELEMENT
SOLDER WIRE TO ELEMENT
3/4 PVC
CONDUIT
1/4-20 MACHINE BOLT,
LOCK WASHER, FLAT
WASHER AND NUT
THREE LOCATIONS
1" PVC CONDUIT
10 FT SECTION EACH
SIDE
3/4" PVC TEE
3" HOSE CLAMPS
FOUR TYPICAL
1" U-BOLT
2" x 2" WOODEN
SUPPORT - 10 FT
MAST 1" EMT CONDUIT
TO ROTATOR
RG-58
http://www.qsl.net/va3iul

135 - Collinear Vertical Antenna 6dB-Gain for 2m, 1.3m, 73cm

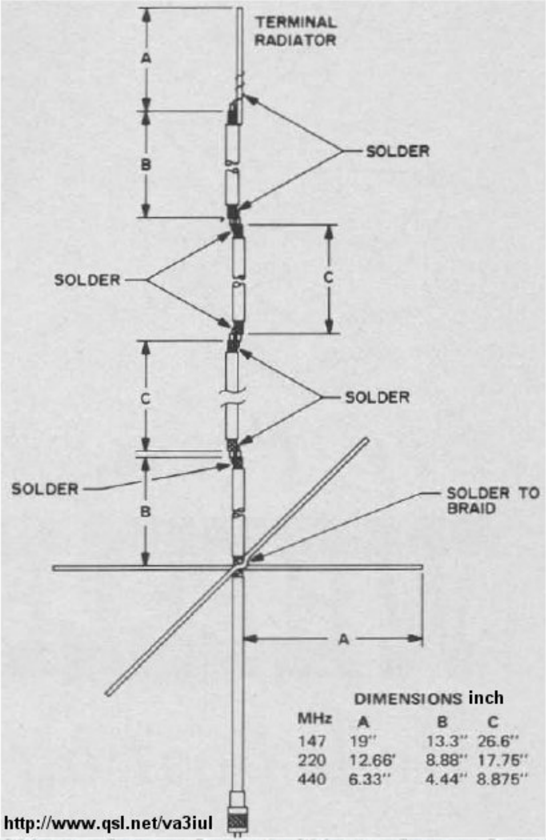
TERMINAL RADIATOR
A
B
SOLDER
C
SOLDER
C
SOLDER
SOLDER
B
SOLDER TO BRAID
A
DIMENSIONS inch
MHz A B C
147 19" 13.3" 26.6"
220 12.66' 8.88" 17.75"
440 6.33" 4.44" 8.875"
http://www.qsl.net/va3iul

136 - Bi-Loop Antenna for 20m

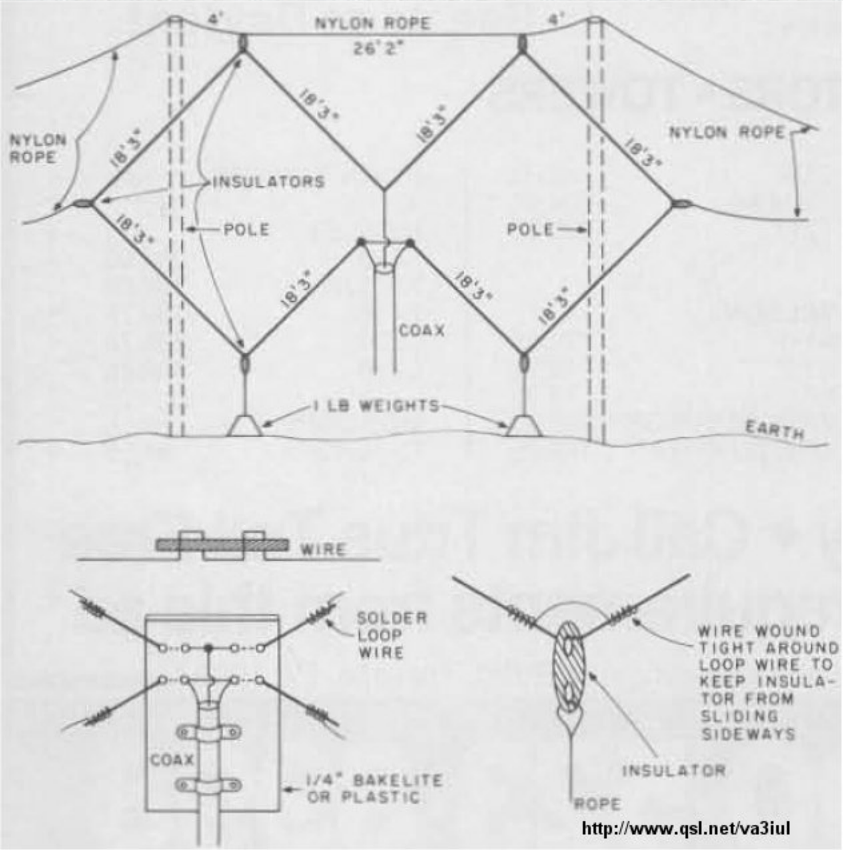
4'
NYLON ROPE
26' 2"
4'
18' 3"
18' 3"
18' 3"
NYLON ROPE
18' 3"
NYLON ROPE
18' 3"
INSULATORS
18' 3"
POLE
POLE
18' 3"
18' 3"
18' 3"
COAX
1 LB WEIGHTS
EARTH
WIRE
SOLDER LOOP WIRE
COAX
1/4" BAKELITE OR PLASTIC
WIRE WOUND TIGHT AROUND LOOP WIRE TO KEEP INSULATOR FROM SLIDING SIDEWAYS
INSULATOR
ROPE
http://www.qsl.net/va3iul

137 - Wire Beam 6dBd-Gain for 10m

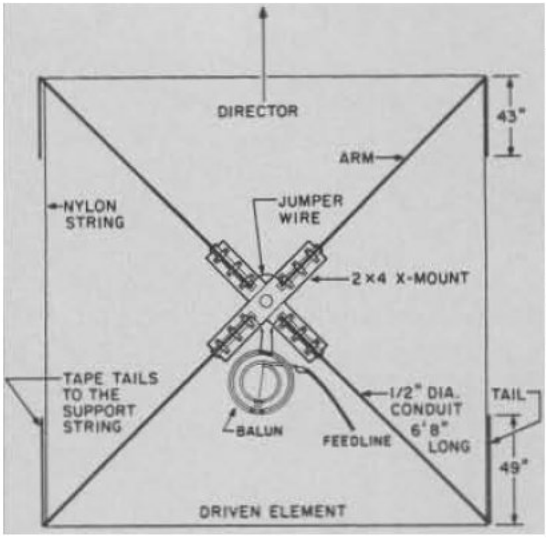
DIRECTOR
43"
ARM
NYLON STRING
JUMPER WIRE
2×4 X-MOUNT
TAPE TAILS TO THE SUPPORT STRING
BALUN
FEEDLINE
1/2" DIA. CONDUIT 6'8" LONG
TAIL
49"
DRIVEN ELEMENT

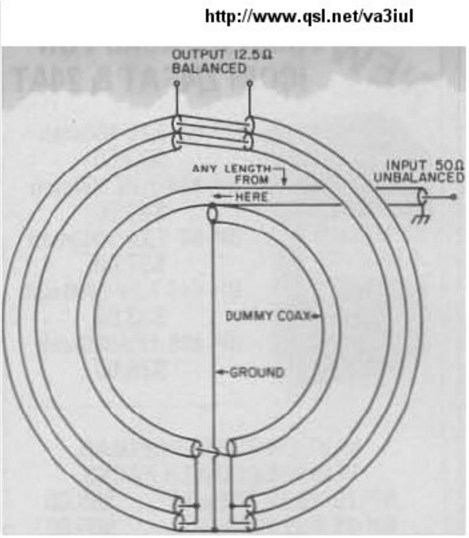
http://www.qsl.net/va3iul
OUTPUT 12.5Ω BALANCED
ANY LENGTH FROM HERE
INPUT 50Ω UNBALANCED
DUMMY COAX
GROUND

138 - Sloping Diamond Antenna 4dB-Gain for 40m

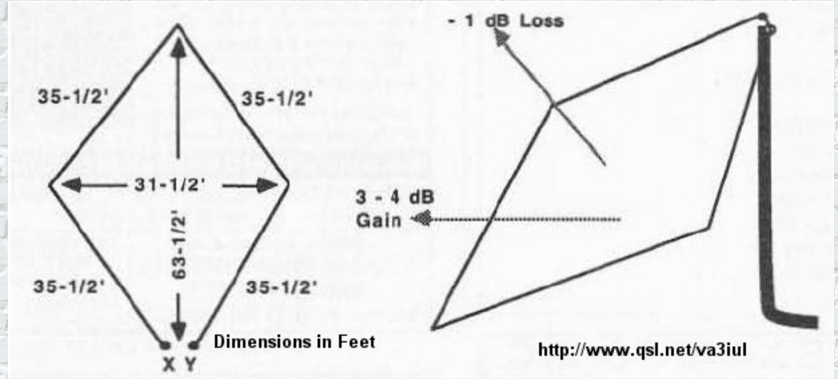
35-1/2'
35-1/2'
31-1/2'
63-1/2'
35-1/2'
35-1/2'
X Y
Dimensions in Feet
- 1 dB Loss
3 - 4 dB Gain
http://www.qsl.net/va3iul

139 - Twisted Loop Antenna for 160m

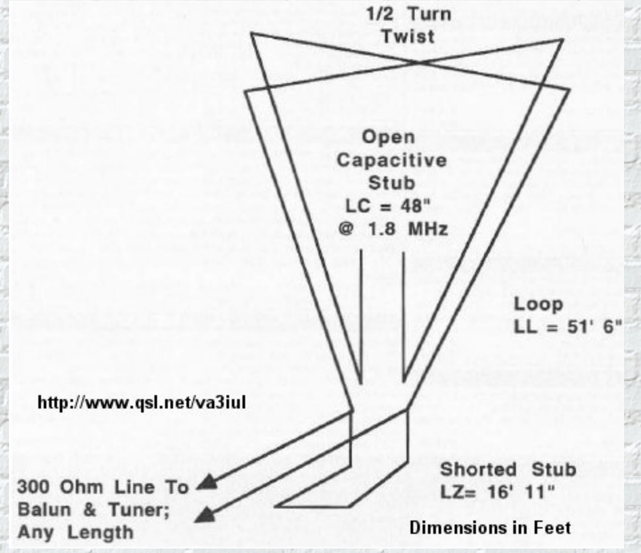
1/2 Turn Twist
Open Capacitive Stub LC = 48" @ 1.8 MHz
Loop LL = 51' 6"
http://www.qsl.net/va3iul
300 Ohm Line To Balun & Tuner; Any Length
Shorted Stub LZ= 16' 11"
Dimensions in Feet

140 - DX RX Loop Antenna for 160m

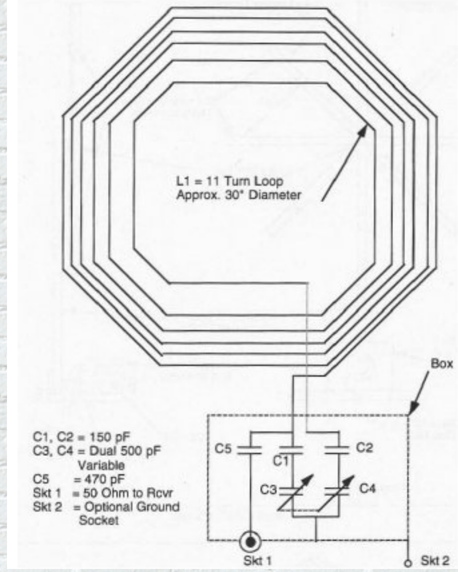
L1 = 11 Turn Loop
Approx. 30" Diameter
Box
C1, C2 = 150 pF
C3, C4 = Dual 500 pF Variable
C5 = 470 pF
Skt 1 = 50 Ohm to Rcvr
Skt 2 = Optional Ground Socket
C5
C1
C2
C3
C4
Skt 1
Skt 2

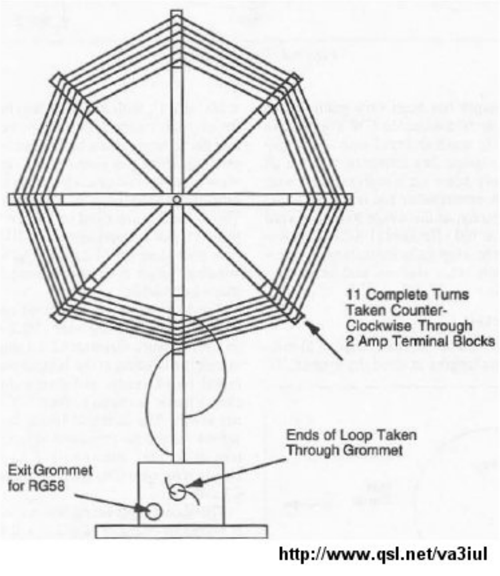
11 Complete Turns Taken Counter-Clockwise Through 2 Amp Terminal Blocks
Ends of Loop Taken Through Grommet
Exit Grommet for RG58
http://www.qsl.net/va3iul

141 - Hentenna 3dB-Gain for 10m, 6m, 2m

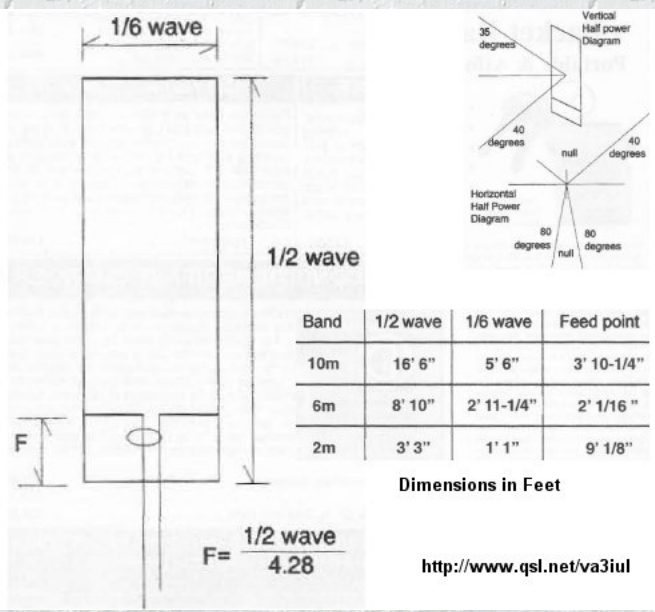
1/6 wave
1/2 wave
F
F= 1/2 wave / 4.28
Vertical Half power Diagram
35 degrees
40 degrees
null
40 degrees
Horizontal Half Power Diagram
80 degrees
null
80 degrees

Band	1/2 wave	1/6 wave	Feed point
10m	16' 6"	5' 6"	3' 10-1/4"
6m	8' 10"	2' 11-1/4"	2' 1/16 "
2m	3' 3"	1' 1"	9' 1/8"

Dimensions in Feet

http://www.qsl.net/va3iul

142 - VK2AAR Wire Antenna for 20m

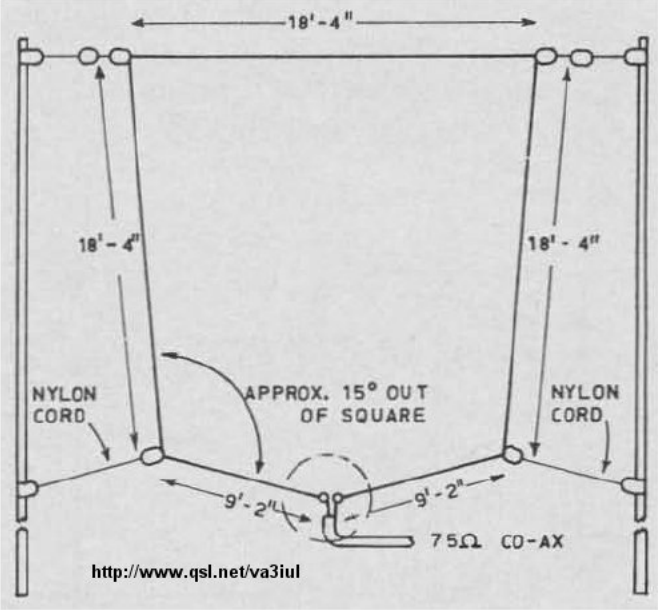
18'-4"
18'-4"
18'-4"
NYLON CORD
APPROX. 15° OUT OF SQUARE
NYLON CORD
9'-2"
9'-2"
75Ω CO-AX
http://www.qsl.net/va3iul

143 - 2-Elements Quad Antenna for 6m

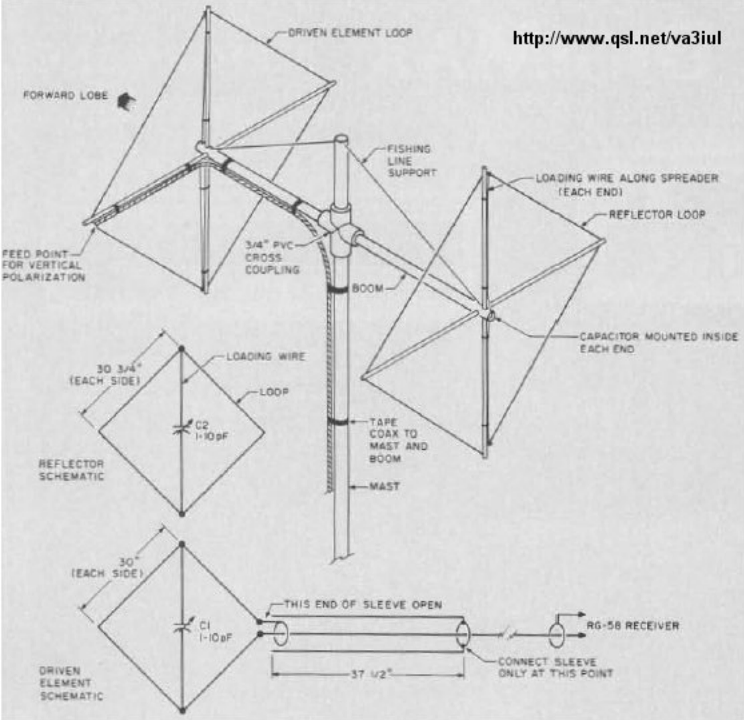
DRIVEN ELEMENT LOOP
http://www.qsl.net/va3iul
FORWARD LOBE
FISHING LINE SUPPORT
LOADING WIRE ALONG SPREADER (EACH END)
REFLECTOR LOOP
FEED POINT FOR VERTICAL POLARIZATION
3/4" PVC CROSS COUPLING
BOOM
CAPACITOR MOUNTED INSIDE EACH END
30 3/4" (EACH SIDE)
LOADING WIRE
LOOP
C2 1-10pF
REFLECTOR SCHEMATIC
TAPE COAX TO MAST AND BOOM
MAST
30" (EACH SIDE)
THIS END OF SLEEVE OPEN
C1 1-10pF
RG-58 RECEIVER
CONNECT SLEEVE ONLY AT THIS POINT
37 1/2"
DRIVEN ELEMENT SCHEMATIC

144 - Hula-Loop Bidirectional 6dB-Gain Antenna for 17m

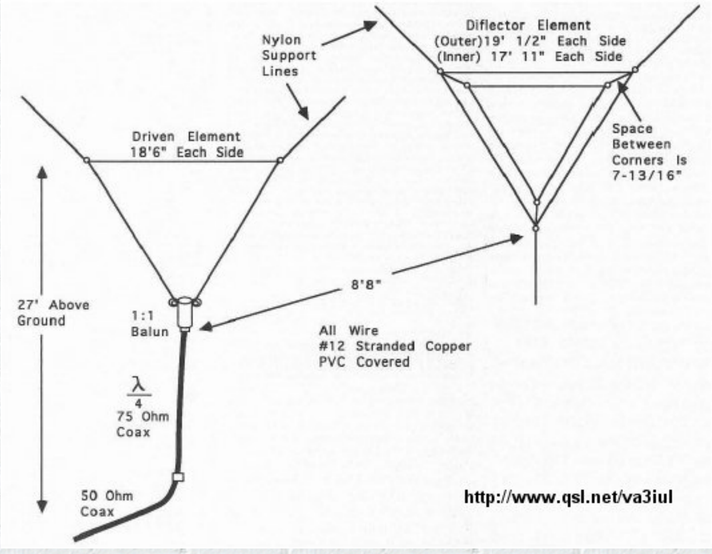
Nylon Support Lines
Diflector Element
(Outer)19' 1/2" Each Side
(Inner) 17' 11" Each Side
Driven Element
18'6" Each Side
Space Between Corners Is 7-13/16"
27' Above Ground
1:1 Balun
8'8"
All Wire
#12 Stranded Copper
PVC Covered
λ/4
75 Ohm Coax
50 Ohm Coax

http://www.qsl.net/va3iul

145 - Moxon Rectangle Beam for 15m-10m

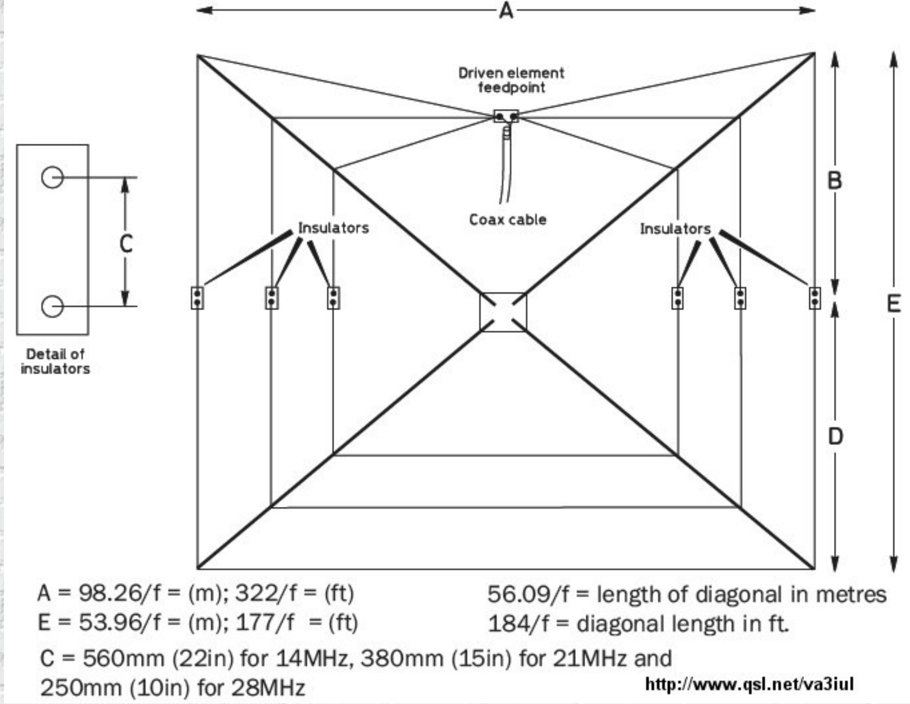
A
B
C
D
E
Driven element feedpoint
Coax cable
Insulators
Insulators
Detail of insulators

A = 98.26/f = (m); 322/f = (ft)
E = 53.96/f = (m); 177/f = (ft)
56.09/f = length of diagonal in metres
184/f = diagonal length in ft.
C = 560mm (22in) for 14MHz, 380mm (15in) for 21MHz and 250mm (10in) for 28MHz

http://www.qsl.net/va3iul

146 - Double-D Beam 4dB-Gain

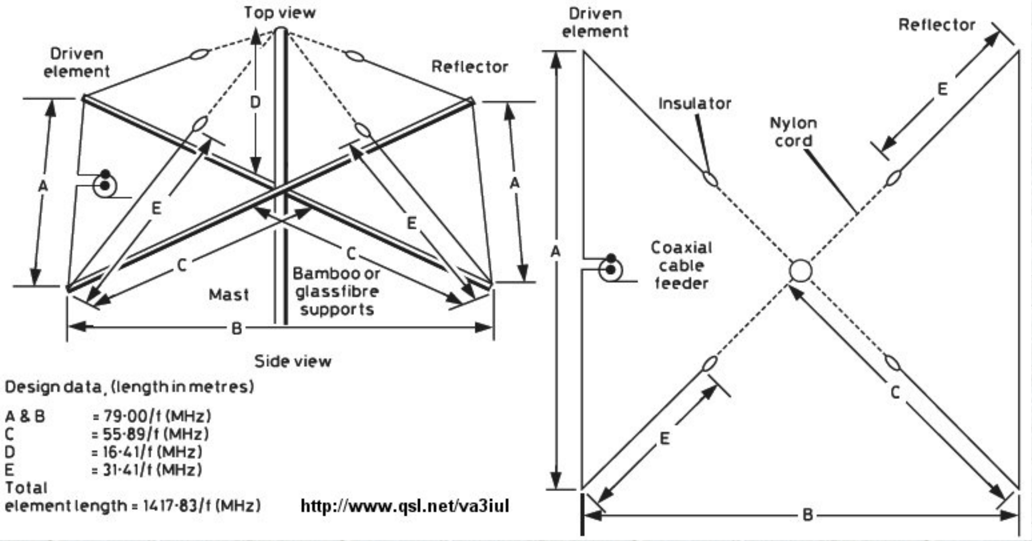
Top view
Driven element
Reflector
D
A
E
E
A
C
C
Mast
Bamboo or glassfibre supports
B
Side view
Driven element
Reflector
Insulator
Nylon cord
E
A
Coaxial cable feeder
C
E
B
http://www.qsl.net/va3iul

Design data, (length in metres)

A & B = 79·00/f (MHz)
C = 55·89/f (MHz)
D = 16·41/f (MHz)
E = 31·41/f (MHz)
Total element length = 1417·83/f (MHz)

147 - KE4PT OCEF All-Band Dipole

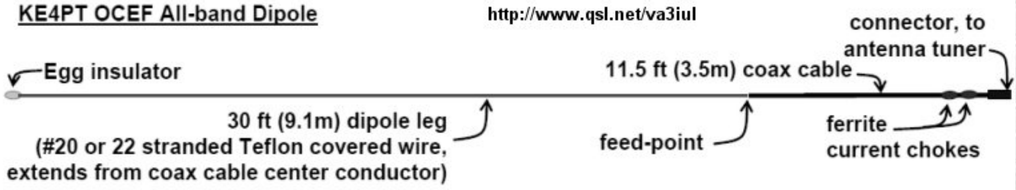
KE4PT OCEF All-band Dipole
http://www.qsl.net/va3iul
connector, to antenna tuner
Egg insulator
11.5 ft (3.5m) coax cable
30 ft (9.1m) dipole leg (#20 or 22 stranded Teflon covered wire, extends from coax cable center conductor)
feed-point
ferrite current chokes

KE4PT Off center end fed (OCEF) All-band Dipole

Lengths are not critical, as long as they are not near a half wavelength on any frequency used.
Dipole radiating length is 30+11.5=41.5 ft, or 12.6 m.
Coax cable may be 50 or 75 ohm on left side of connector, short 50 ohm coax to antenna tuner.
Use antenna tuner for operation on any ham band 40m – 6m.
May be hung by Egg insulator at one end, or draped and drooped over a fiberglass pole.

148 - Wire Quad Antenna for 40m

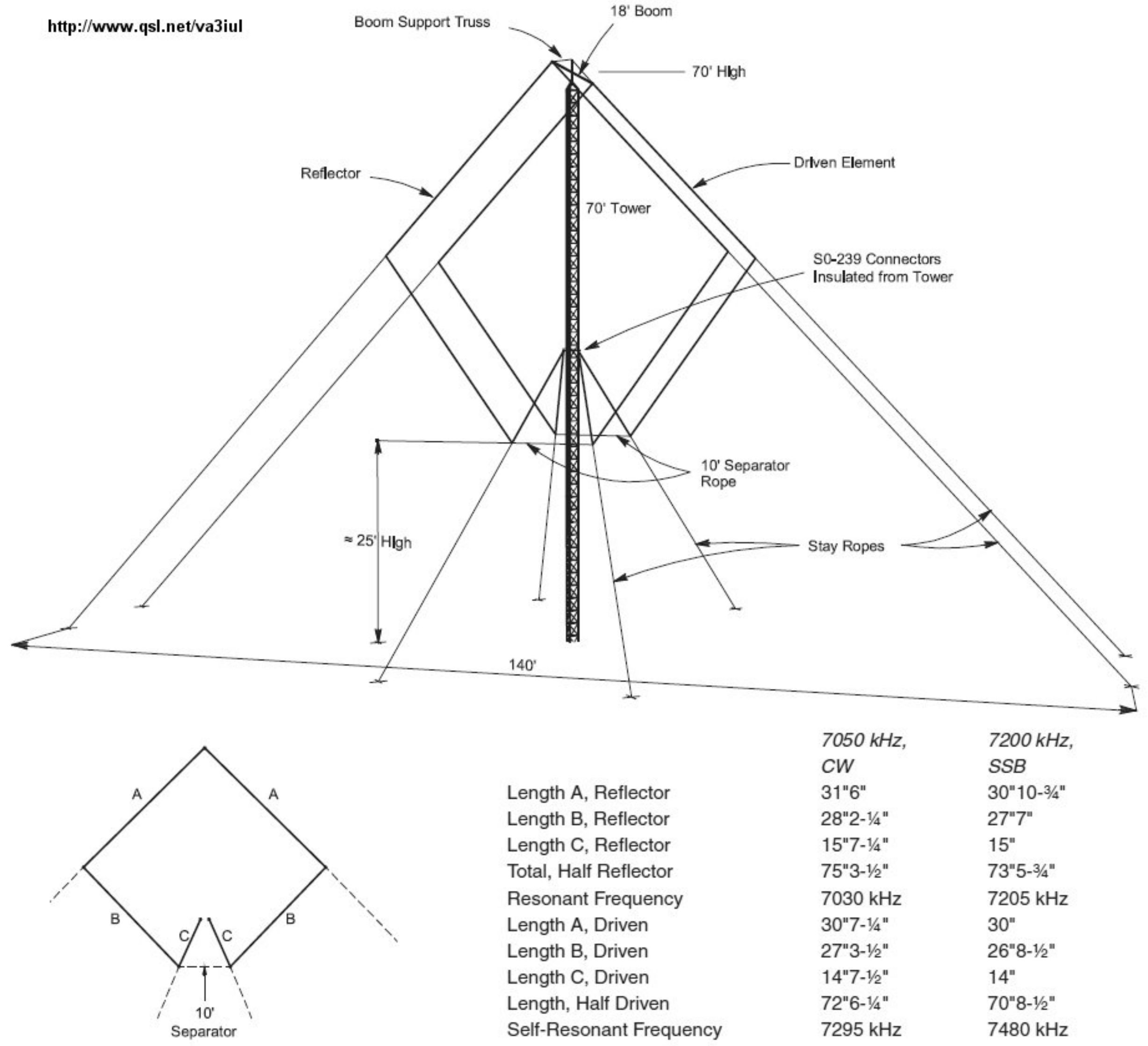
http://www.qsl.net/va3iul
Boom Support Truss
18' Boom
70' High
Reflector
Driven Element
70' Tower
S0-239 Connectors
Insulated from Tower
10' Separator
Rope
≈ 25' High
Stay Ropes
140'
A
A
B
B
C
C
10'
Separator

	7050 kHz, CW	7200 kHz, SSB
Length A, Reflector	31"6"	30"10-¾"
Length B, Reflector	28"2-¼"	27"7"
Length C, Reflector	15"7-¼"	15"
Total, Half Reflector	75"3-½"	73"5-¾"
Resonant Frequency	7030 kHz	7205 kHz
Length A, Driven	30"7-¼"	30"
Length B, Driven	27"3-½"	26"8-½"
Length C, Driven	14"7-½"	14"
Length, Half Driven	72"6-¼"	70"8-½"
Self-Resonant Frequency	7295 kHz	7480 kHz

---

149 - Inclined Dipole Antenna for 80m, 40m

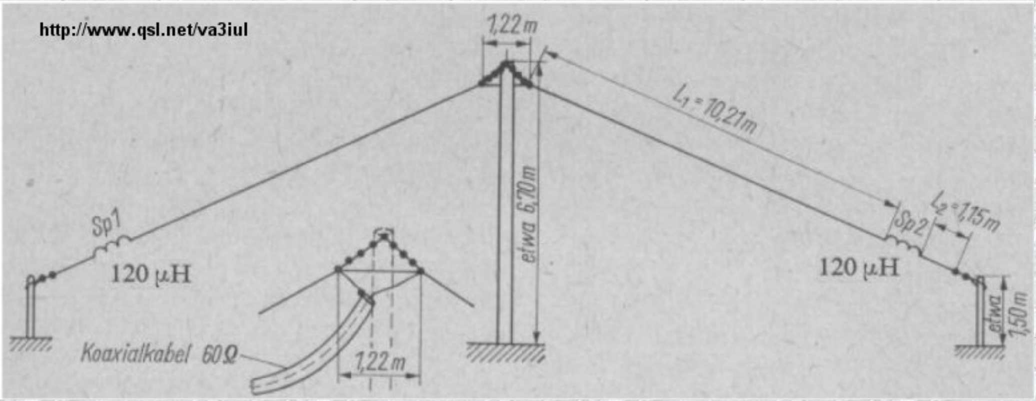
http://www.qsl.net/va3iul
1,22 m
L_1 = 10,21m
Sp1
120 µH
etwa 6,70 m
Sp2
L_2 = 1,15 m
120 µH
etwa 1,50 m
Koaxialkabel 60Ω
1,22 m

---

150 - Pyramidal Wire Antenna for 80m

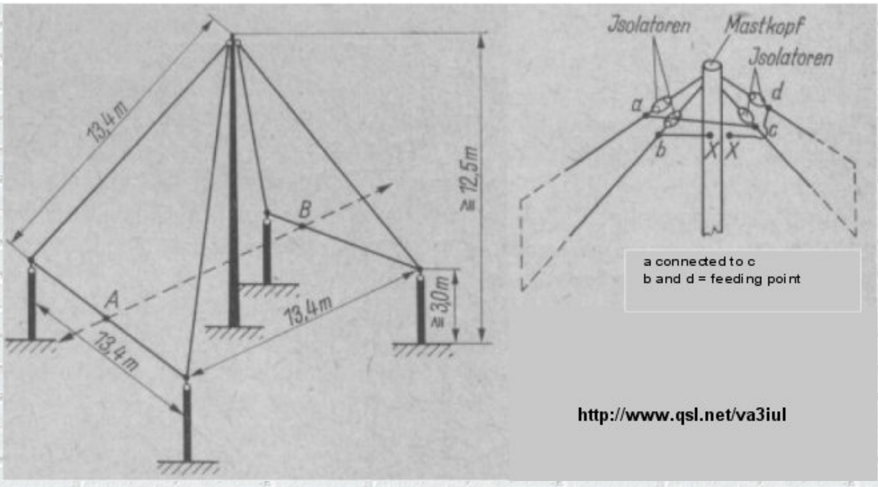
13,4 m
≈ 12,5 m
B
A
13,4 m
13,4 m
≈ 3,0 m
Jsolatoren
Mastkopf
Jsolatoren
a
b
c
d
X
X
a connected to c
b and d = feeding point
http://www.qsl.net/va3iul

151 - Random Wire Antenna All Bands

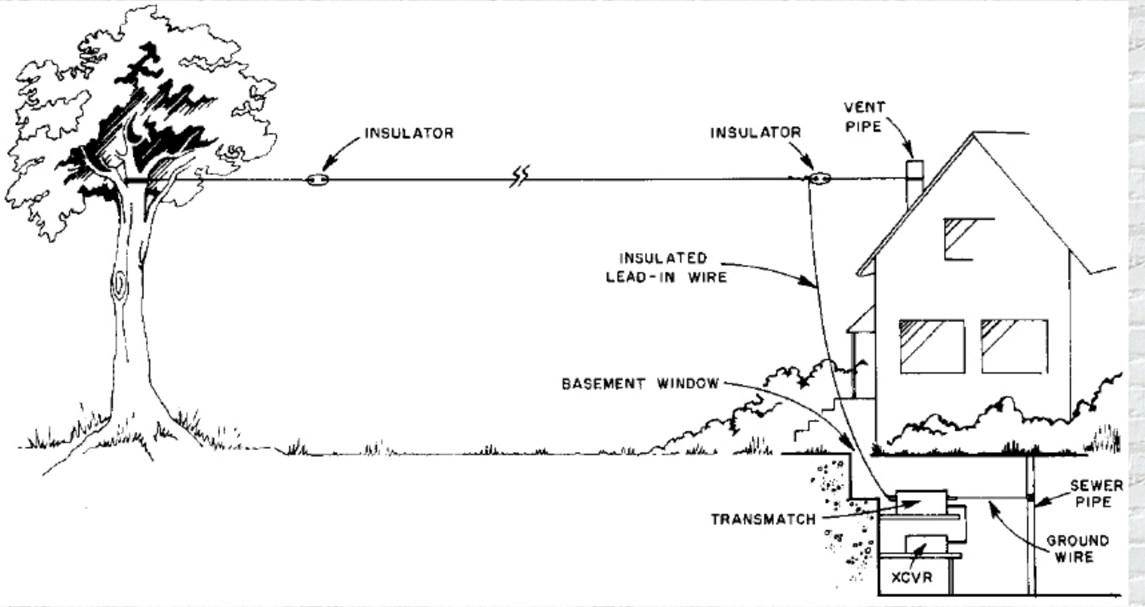
INSULATOR
INSULATOR
VENT PIPE
INSULATED LEAD-IN WIRE
BASEMENT WINDOW
TRANSMATCH
XCVR
SEWER PIPE
GROUND WIRE

152 - Multiband Dipole Antenna for 80m, 40m,15m, 10m

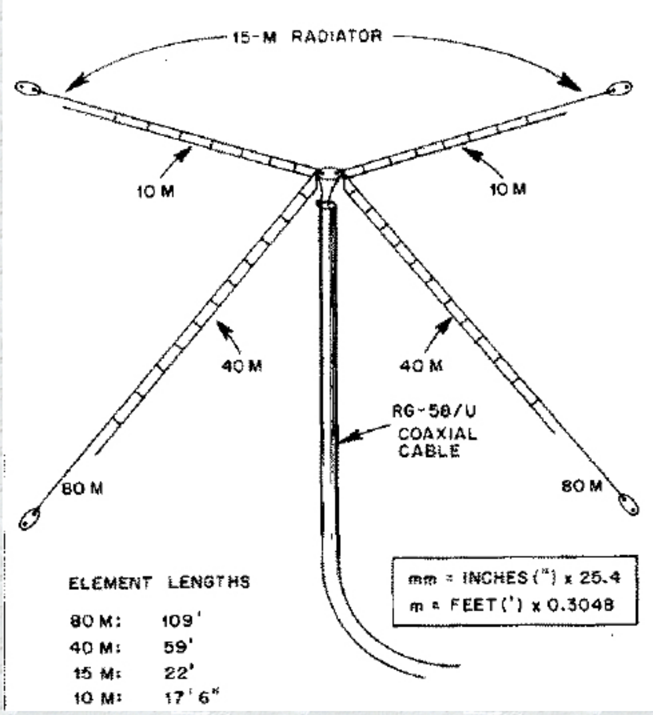
15-M RADIATOR
10 M
10 M
40 M
40 M
RG-58/U COAXIAL CABLE
80 M
80 M

ELEMENT LENGTHS

80 M:	109'
40 M:	59'
15 M:	22'
10 M:	17' 6"

mm = INCHES (") x 25.4
m = FEET (') x 0.3048

---

153 - Slim Jim Wire Antenna for 4m

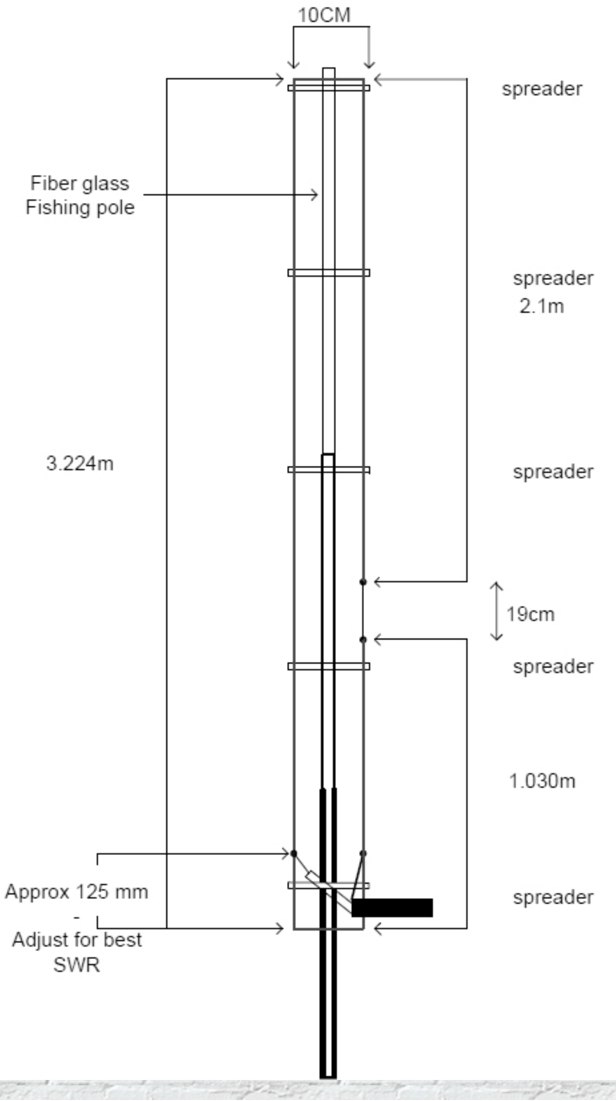
10CM
spreader
Fiber glass Fishing pole
spreader
2.1m
3.224m
spreader
19cm
spreader
1.030m
Approx 125 mm - Adjust for best SWR
spreader

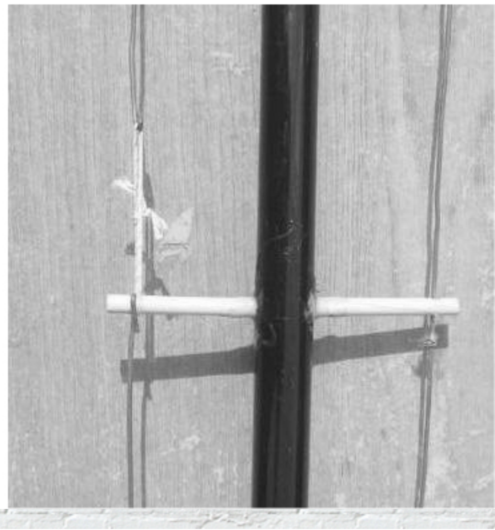

154 - Delta Loop for 6m

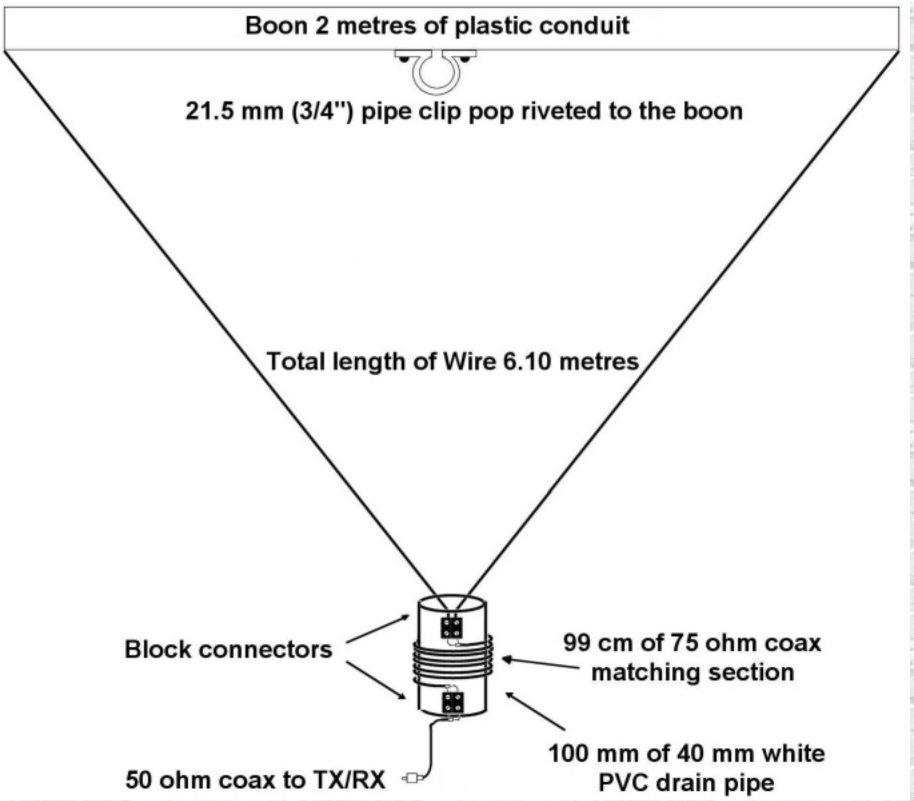
Boon 2 metres of plastic conduit
21.5 mm (3/4") pipe clip pop riveted to the boon
Total length of Wire 6.10 metres
Block connectors
99 cm of 75 ohm coax matching section
100 mm of 40 mm white PVC drain pipe
50 ohm coax to TX/RX

155 - Re-Configurable Antenna for 160m and 80m

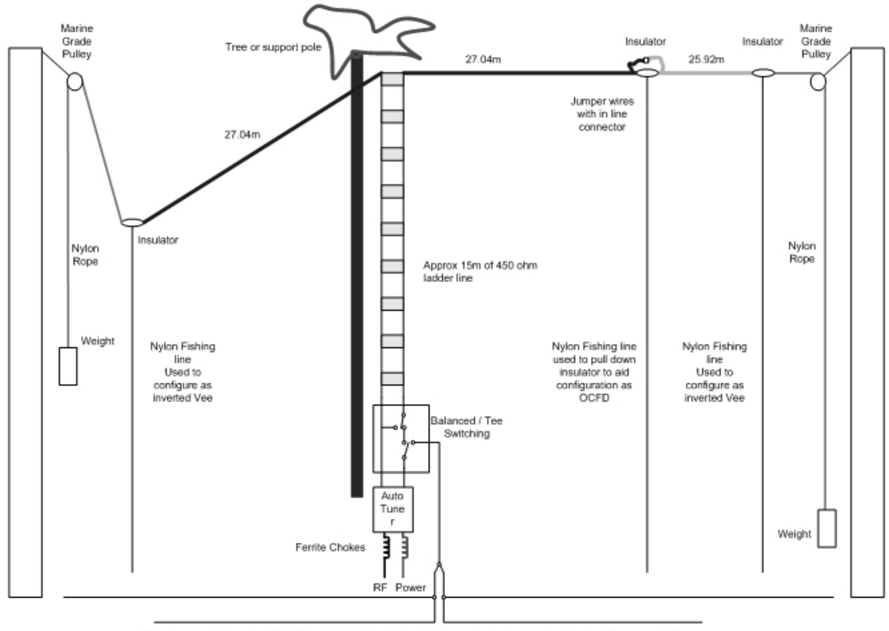
Marine Grade Pulley
Tree or support pole
27.04m
Insulator
25.92m
Insulator
Marine Grade Pulley
Jumper wires with in line connector
27.04m
Insulator
Nylon Rope
Approx 15m of 450 ohm ladder line
Nylon Rope
Weight
Nylon Fishing line Used to configure as inverted Vee
Nylon Fishing line used to pull down insulator to aid configuration as OCFD
Nylon Fishing line Used to configure as inverted Vee
Balanced / Tee Switching
Auto Tune r
Ferrite Chokes
Weight
RF
Power
Tuned ground planes for 160 & 80m approx 0.6 x ¼ Wavelength stapled to fence at 2m height of 2m spaced 0.5m apart

156 - Very Low Frequency Inverted-L Antenna

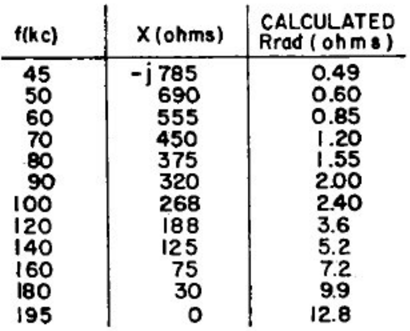

f(kc)	X (ohms)	CALCULATED Rrad (ohms)
45	-j 785	0.49
50	690	0.60
60	555	0.85
70	450	1.20
80	375	1.55
90	320	2.00
100	268	2.40
120	188	3.6
140	125	5.2
160	75	7.2
180	30	9.9
195	0	12.8

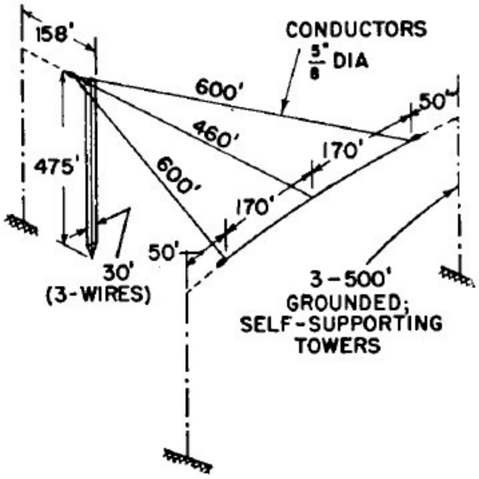
158'
CONDUCTORS
5" 8 DIA
600'
50'
460'
170'
475'
600'
170'
50'
30'
(3-WIRES)
3-500'
GROUNDED;
SELF-SUPPORTING
TOWERS

157 - Reduced Size Half Sloper Antenna for 160m

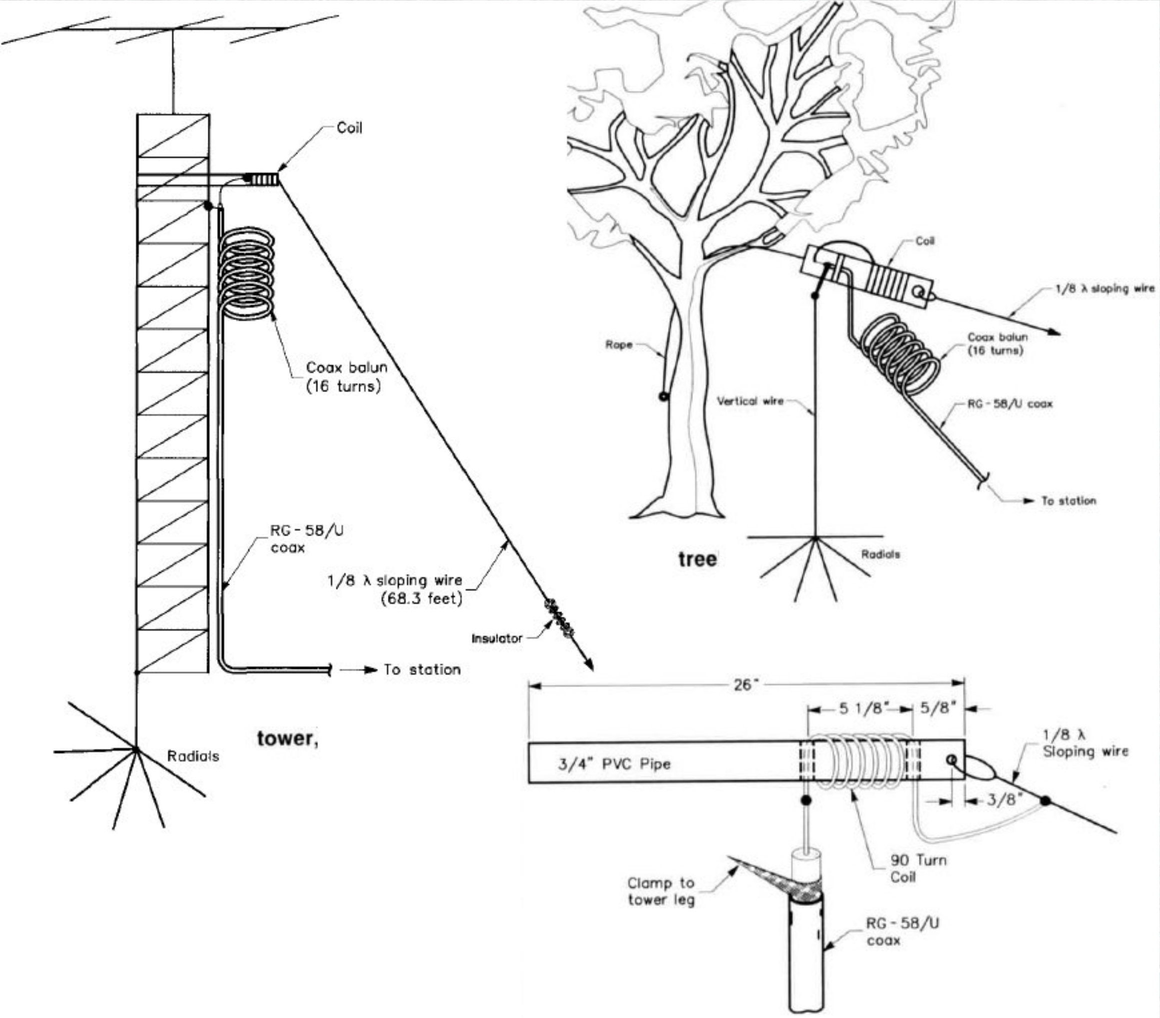
Coil
Coax balun
(16 turns)
RG-58/U
coax
1/8 λ sloping wire
(68.3 feet)
Insulator
To station
Radials
tower,
Coil
1/8 λ sloping wire
Rope
Coax balun
(16 turns)
Vertical wire
RG-58/U coax
To station
tree
Radials
26"
5 1/8"
5/8"
3/4" PVC Pipe
1/8 λ
Sloping wire
3/8"
90 Turn
Coil
Clamp to
tower leg
RG-58/U
coax

158 - Tree-Mounted HF Antenna

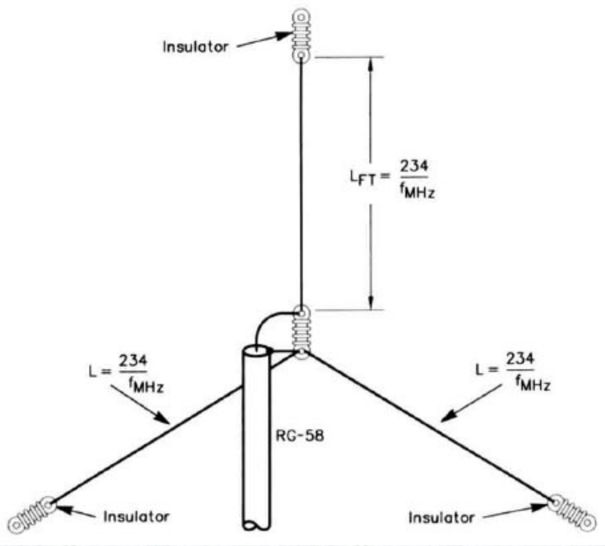
Insulator
$L_{FT} = \frac{234}{f_{MHz}}$
$L = \frac{234}{f_{MHz}}$
$L = \frac{234}{f_{MHz}}$
RG-58
Insulator
Insulator

159 - Multiband Vertical Antenna for 80m, 40m, 20m

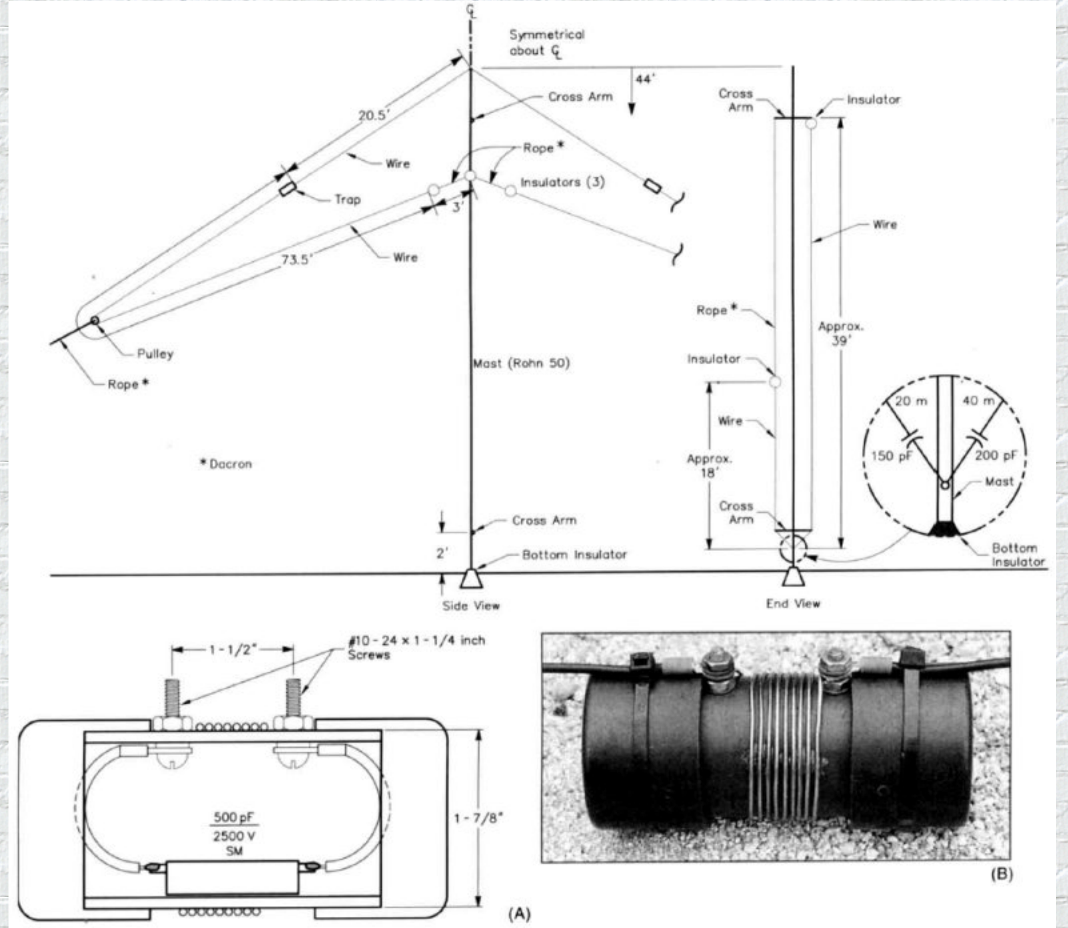
℄
Symmetrical about ℄
44'
Cross Arm
20.5'
Wire
Rope *
Insulators (3)
Trap
3'
73.5'
Wire
Pulley
Rope *
Mast (Rohn 50)
* Dacron
Cross Arm
2'
Bottom Insulator
Side View
Cross Arm
Insulator
Wire
Rope *
Approx. 39'
Insulator
Wire
Approx. 18'
Cross Arm
End View
20 m
40 m
150 pF
200 pF
Mast
Bottom Insulator
1-1/2"
#10-24 x 1-1/4 inch Screws
500 pF 2500 V SM
1-7/8"
(A)
(B)

160 - Marconi Antenna for 136 kHz

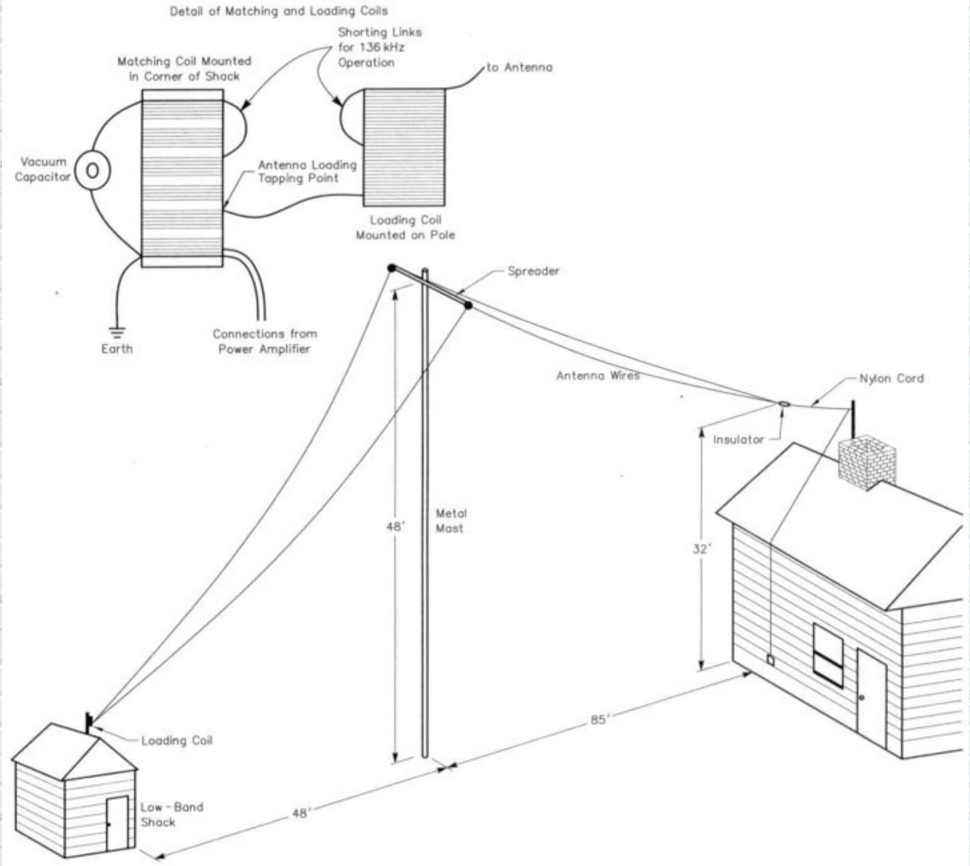
Detail of Matching and Loading Coils
Shorting Links for 136 kHz Operation
Matching Coil Mounted in Corner of Shack
to Antenna
Vacuum Capacitor
Antenna Loading Tapping Point
Loading Coil Mounted on Pole
Earth
Connections from Power Amplifier
Spreader
Antenna Wires
Nylon Cord
Insulator
48'
Metal Mast
32'
85'
Loading Coil
Low - Band Shack
48'

---

161 - Simple Killer Antenna for 40m

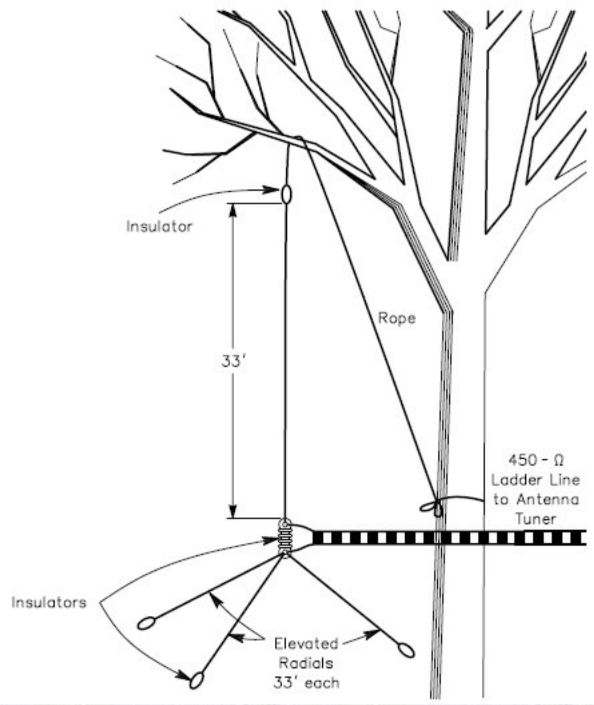
Insulator
33'
Rope
450 - Ω Ladder Line to Antenna Tuner
Insulators
Elevated Radials 33' each

---

162 - Stub-Directed V Antenna for 80m

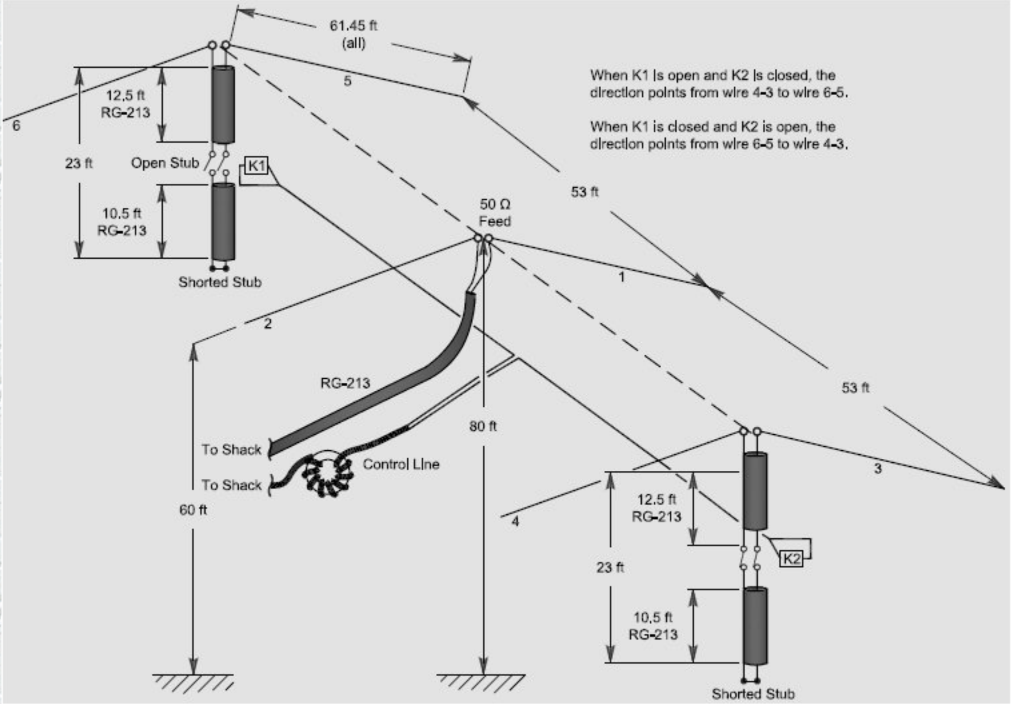
61.45 ft
(all)
5
6
12.5 ft
RG-213
23 ft
Open Stub
K1
10.5 ft
RG-213
Shorted Stub
When K1 is open and K2 is closed, the direction points from wire 4-3 to wire 6-5.
When K1 is closed and K2 is open, the direction points from wire 6-5 to wire 4-3.
53 ft
50 Ω
Feed
1
2
RG-213
To Shack
To Shack
Control Line
60 ft
80 ft
53 ft
3
4
12.5 ft
RG-213
23 ft
K2
10.5 ft
RG-213
Shorted Stub

163 - KT0NY Over-and-Under DX Antenna for 20m

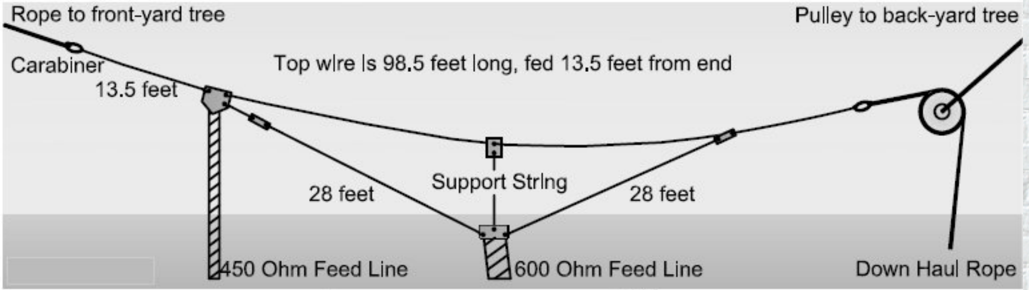
Rope to front-yard tree
Pulley to back-yard tree
Carabiner
Top wire is 98.5 feet long, fed 13.5 feet from end
13.5 feet
28 feet
Support String
28 feet
450 Ohm Feed Line
600 Ohm Feed Line
Down Haul Rope

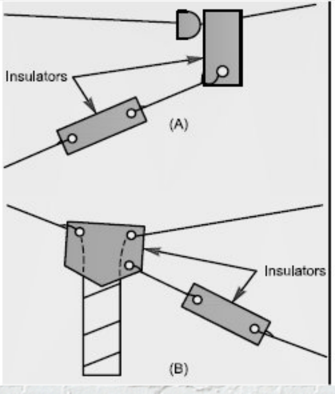
Insulators
(A)
Insulators
(B)

164 - Horizontal Loop Antennas

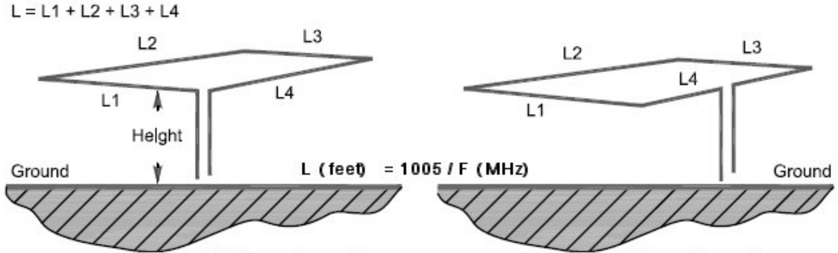
L = L1 + L2 + L3 + L4
L2
L3
L1
L4
Height
Ground
L (feet) = 1005 / F (MHz)
L2
L3
L4
L1
Ground

corner of angle-fed (AF) configuration

center-fed (CF) configuration

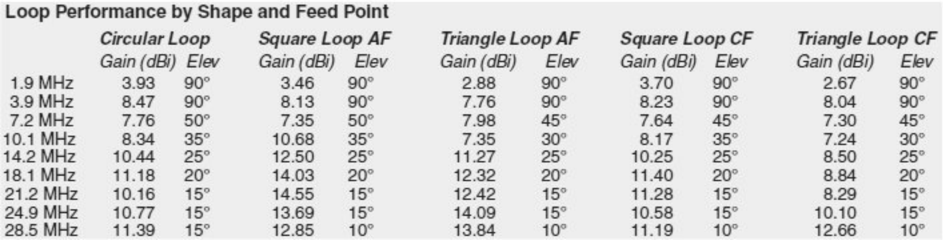

Loop Performance by Shape and Feed Point

	Circular Loop		Square Loop AF		Triangle Loop AF		Square Loop CF		Triangle Loop CF	
	Gain (dBi)	Elev	Gain (dBi)	Elev	Gain (dBi)	Elev	Gain (dBi)	Elev	Gain (dBi)	Elev
1.9 MHz	3.93	90°	3.46	90°	2.88	90°	3.70	90°	2.67	90°
3.9 MHz	8.47	90°	8.13	90°	7.76	90°	8.23	90°	8.04	90°
7.2 MHz	7.76	50°	7.35	50°	7.98	45°	7.64	45°	7.30	45°
10.1 MHz	8.34	35°	10.68	35°	7.35	30°	8.17	35°	7.24	30°
14.2 MHz	10.44	25°	12.50	25°	11.27	25°	10.25	25°	8.50	25°
18.1 MHz	11.18	20°	14.03	20°	12.32	20°	11.40	20°	8.84	20°
21.2 MHz	10.16	15°	14.55	15°	12.42	15°	11.28	15°	8.29	15°
24.9 MHz	10.77	15°	13.69	15°	14.09	15°	10.58	15°	10.10	15°
28.5 MHz	11.39	15°	12.85	10°	13.84	10°	11.19	10°	12.66	10°

---

165 - Ribbon J-Pole for 2m

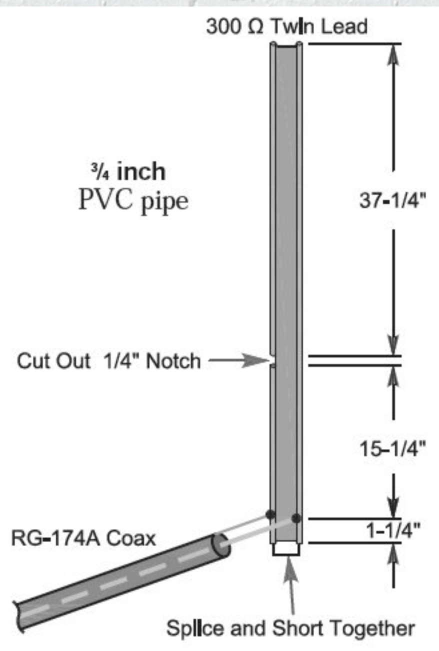
300 Ω Twin Lead
¾ inch
PVC pipe
37-1/4"
Cut Out 1/4" Notch
15-1/4"
1-1/4"
RG-174A Coax
Splice and Short Together

---

166 - Dual Band Ribbon J-Pole for 2m, 70cm

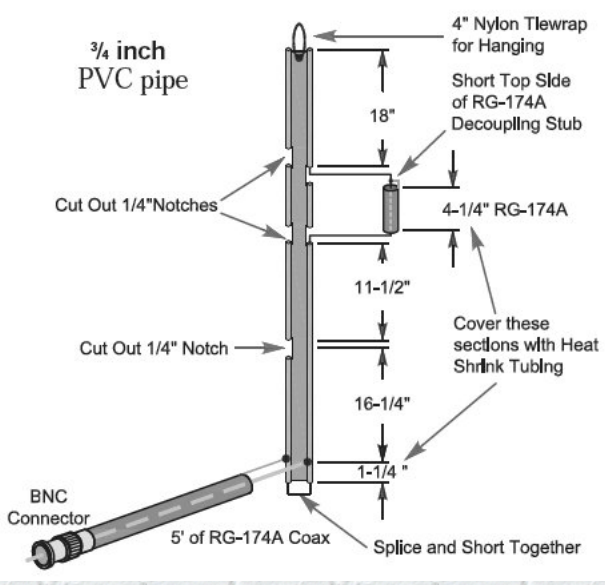
¾ inch
PVC pipe
4" Nylon Tiewrap
for Hanging
Short Top Side
of RG-174A
Decoupling Stub
18"
Cut Out 1/4"Notches
4-1/4" RG-174A
11-1/2"
Cover these
sections with Heat
Shrink Tubing
Cut Out 1/4" Notch
16-1/4"
1-1/4 "
BNC
Connector
5' of RG-174A Coax
Splice and Short Together

167 - Square Vertical Loop Antenna for 40m

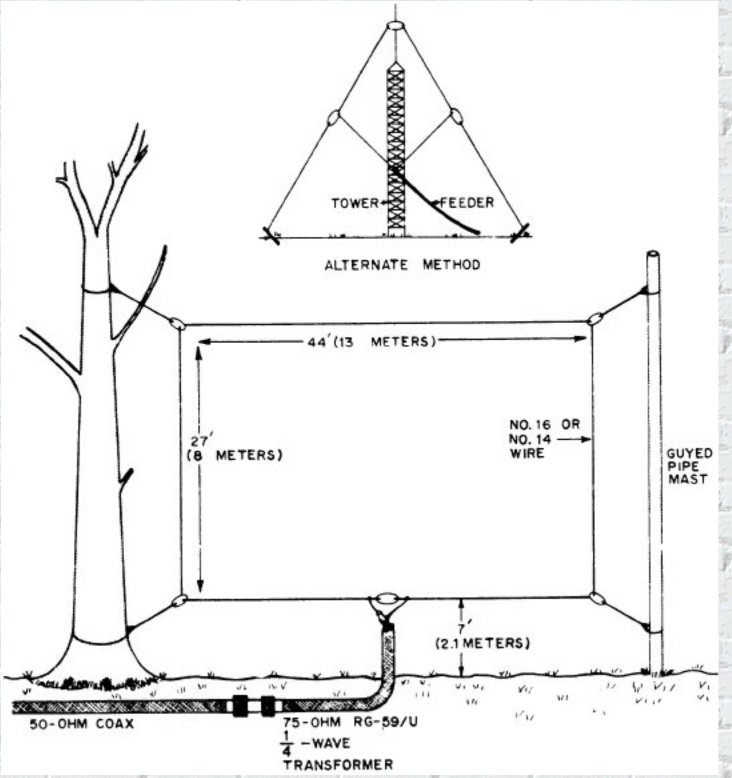
TOWER
FEEDER
ALTERNATE METHOD
44′ (13 METERS)
27′
(8 METERS)
NO. 16 OR
NO. 14
WIRE
GUYED
PIPE
MAST
7′
(2.1 METERS)
50-OHM COAX
75-OHM RG-59/U
$\frac{1}{4}$ -WAVE
TRANSFORMER

168 - Tri-Band Quad Antenna for 20m, 15m, 10m

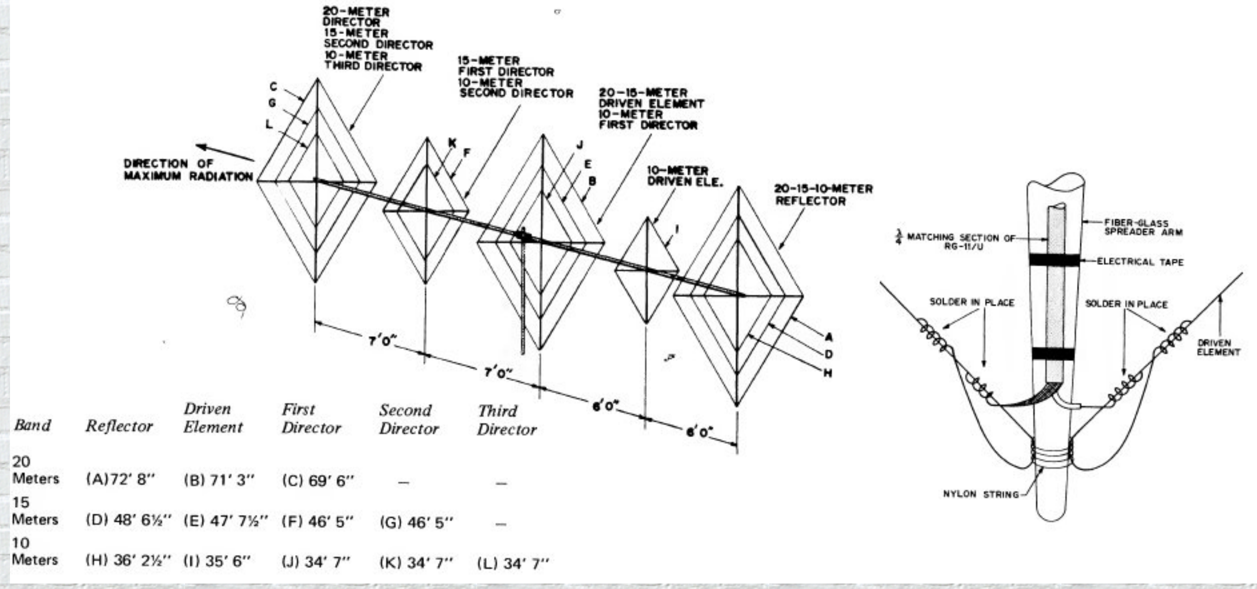
20-METER
DIRECTOR
15-METER
SECOND DIRECTOR
10-METER
THIRD DIRECTOR
15-METER
FIRST DIRECTOR
10-METER
SECOND DIRECTOR
20-15-METER
DRIVEN ELEMENT
10-METER
FIRST DIRECTOR
10-METER
DRIVEN ELE.
20-15-10-METER
REFLECTOR
DIRECTION OF
MAXIMUM RADIATION
C
G
L
K
F
J
E
B
I
A
D
H
7'0"
7'0"
6'0"
6'0"
¾ MATCHING SECTION OF
RG-11/U
FIBER-GLASS
SPREADER ARM
ELECTRICAL TAPE
SOLDER IN PLACE
SOLDER IN PLACE
DRIVEN
ELEMENT
NYLON STRING

Band	*Reflector*	*Driven Element*	*First Director*	*Second Director*	*Third Director*
20 Meters	(A)72' 8"	(B) 71' 3"	(C) 69' 6"	–	–
15 Meters	(D) 48' 6½"	(E) 47' 7½"	(F) 46' 5"	(G) 46' 5"	–
10 Meters	(H) 36' 2½"	(I) 35' 6"	(J) 34' 7"	(K) 34' 7"	(L) 34' 7"

---

169 - 3D Quad Antenna for 80m, 40m, 20m, 15m, 10m

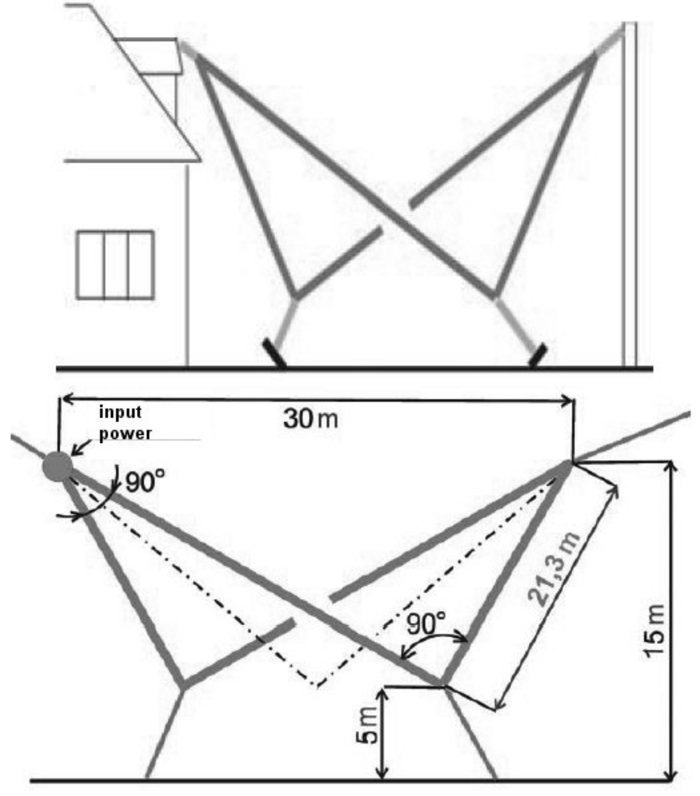
input
power
30 m
90°
90°
21,3 m
15 m
5 m

---

170 - Sloping Wire Antenna for 30m, 20m, 17m, 15m, 12m, 10m

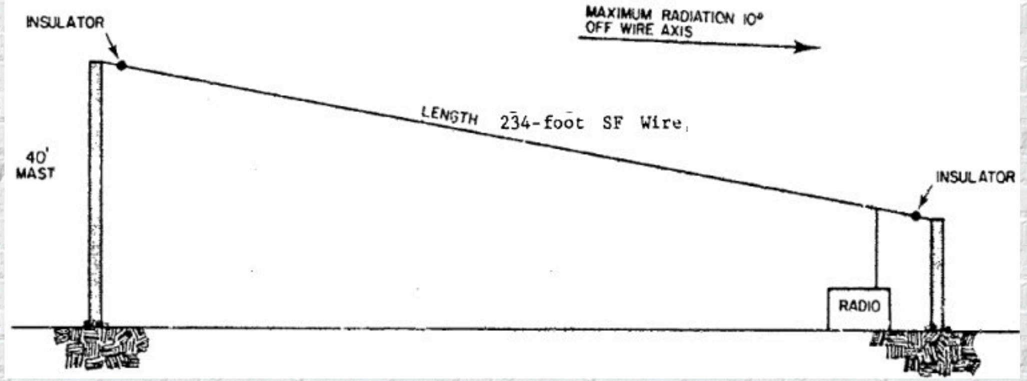
INSULATOR
MAXIMUM RADIATION 10°
OFF WIRE AXIS
LENGTH 234-foot SF Wire
40'
MAST
INSULATOR
RADIO

171 - Broadband Dipole Antenna with Coaxial Resonator for 80m

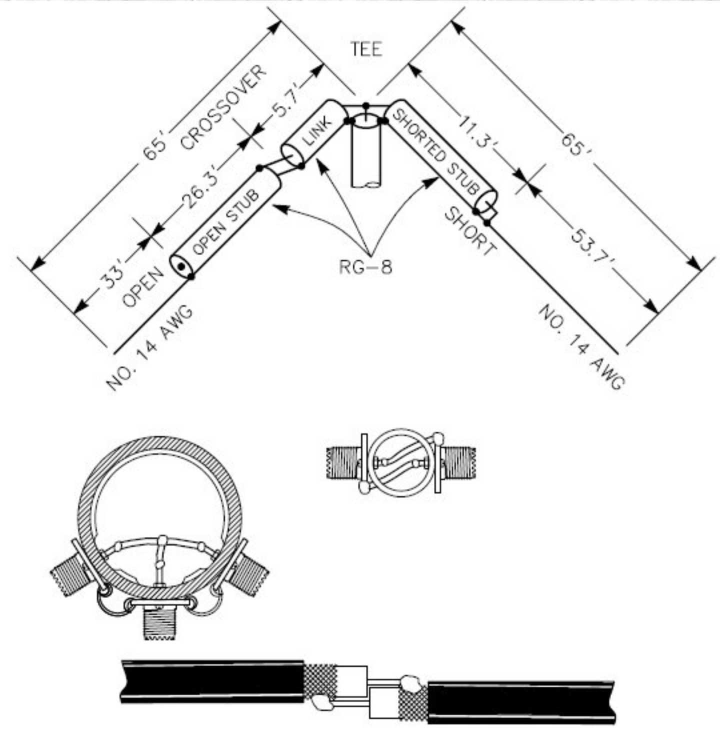
TEE
65' CROSSOVER
5.7'
26.3'
33'
LINK
OPEN STUB
OPEN
NO. 14 AWG
RG−8
SHORTED STUB
SHORT
11.3'
53.7'
65'
NO. 14 AWG

172 - Resonant Feedline Dipole Antenna for 80m

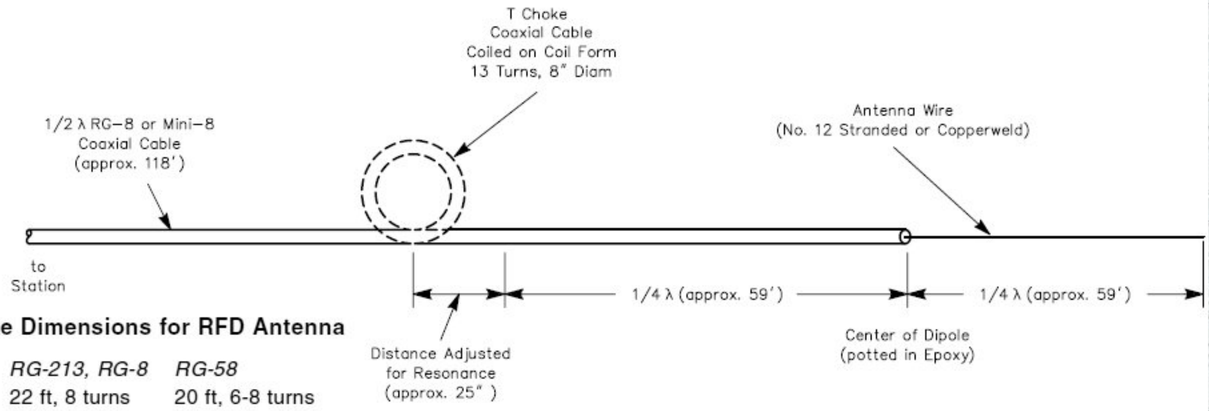
T Choke
Coaxial Cable
Coiled on Coil Form
13 Turns, 8" Diam
1/2 λ RG−8 or Mini−8
Coaxial Cable
(approx. 118')
Antenna Wire
(No. 12 Stranded or Copperweld)
to
Station
1/4 λ (approx. 59')
1/4 λ (approx. 59')
Distance Adjusted
for Resonance
(approx. 25")
Center of Dipole
(potted in Epoxy)

Choke Dimensions for RFD Antenna

Freq	*RG-213, RG-8*	*RG-58*
3.5	22 ft, 8 turns	20 ft, 6-8 turns
7	22 ft, 10 turns	15 ft, 6 turns
10	12 ft, 10 turns	10 ft, 7 turns
14	10 ft, 4 turns	8 ft, 8 turns
21	8 ft, 6-8 turns	6 ft, 8 turns
28	6 ft, 6-8 turns	4 ft, 6-8 turns

173 - Dual-Band Loading Wire Antenna for 80m, 30m

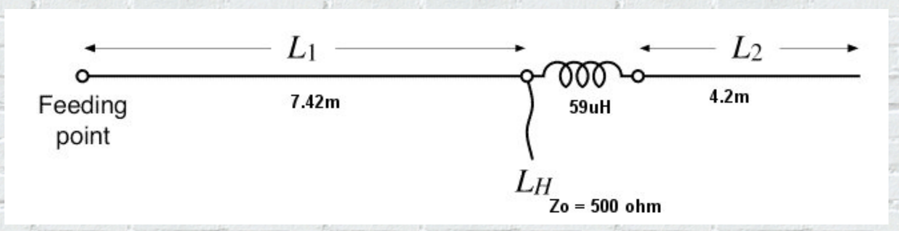
L_1
L_2
Feeding point
7.42m
59uH
4.2m
L_H
Zo = 500 ohm

---

174 - Stub Matching Antennas

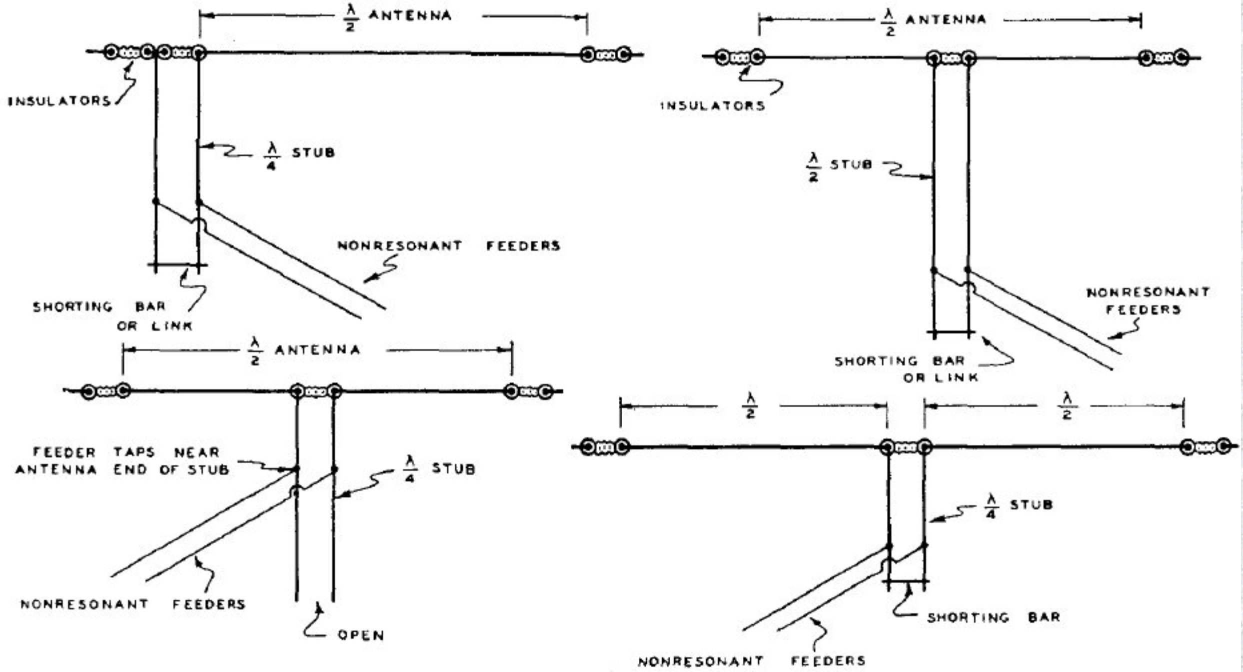
λ/2 ANTENNA
INSULATORS
λ/4 STUB
NONRESONANT FEEDERS
SHORTING BAR OR LINK
λ/2 ANTENNA
INSULATORS
λ/2 STUB
NONRESONANT FEEDERS
SHORTING BAR OR LINK
λ/2 ANTENNA
FEEDER TAPS NEAR ANTENNA END OF STUB
λ/4 STUB
NONRESONANT FEEDERS
OPEN
λ/2
λ/2
λ/4 STUB
SHORTING BAR
NONRESONANT FEEDERS

---

175 - J-Style Vertical Wire Antenna for 10m

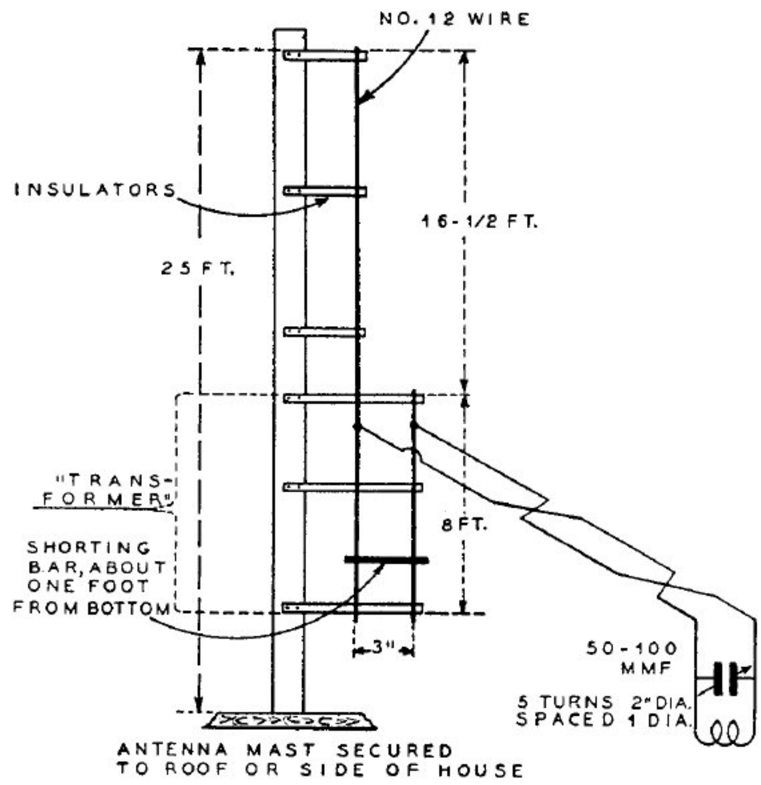
NO. 12 WIRE
INSULATORS
25 FT.
16-1/2 FT.
"TRANS-FORMER"
SHORTING BAR, ABOUT ONE FOOT FROM BOTTOM
8 FT.
3"
50-100 MMF
5 TURNS 2" DIA. SPACED 1 DIA.
ANTENNA MAST SECURED TO ROOF OR SIDE OF HOUSE

176 - Dual Band Vertical with Zepp Feeders for 40m, 20m

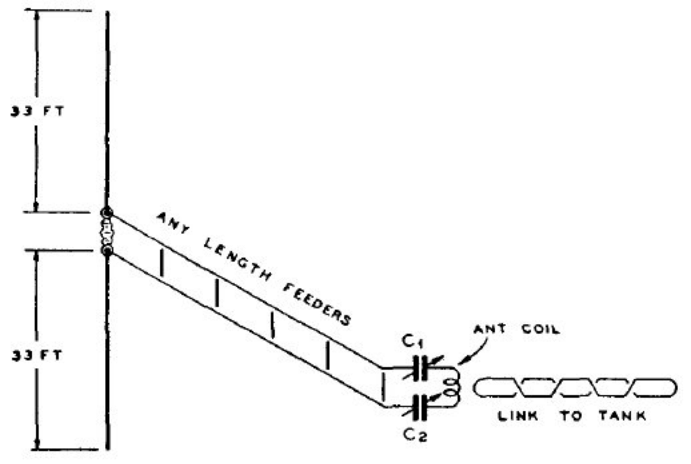
33 FT
33 FT
ANY LENGTH FEEDERS
C1
C2
ANT COIL
LINK TO TANK

177 - RCA Double Doublet for 40m to 12m

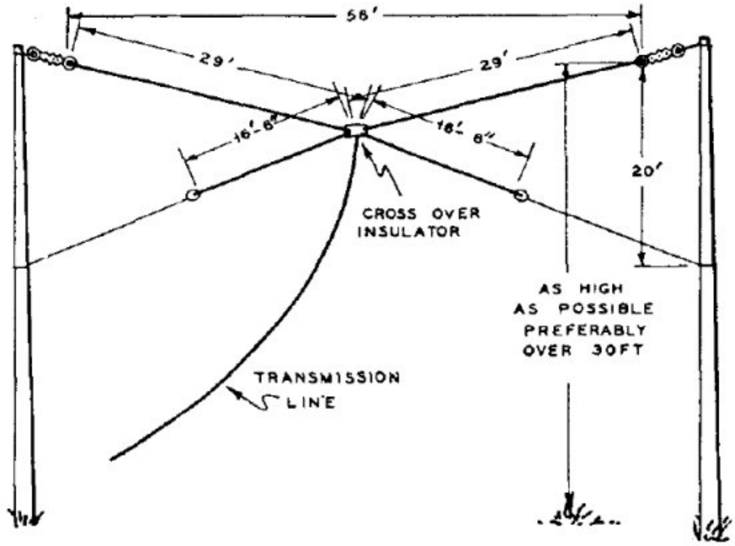
58'
29'
29'
16'-6"
16'-6"
20'
CROSS OVER INSULATOR
AS HIGH AS POSSIBLE PREFERABLY OVER 30FT
TRANSMISSION LINE

178 - RCA Spiderweb Antenna for 40m to 6m

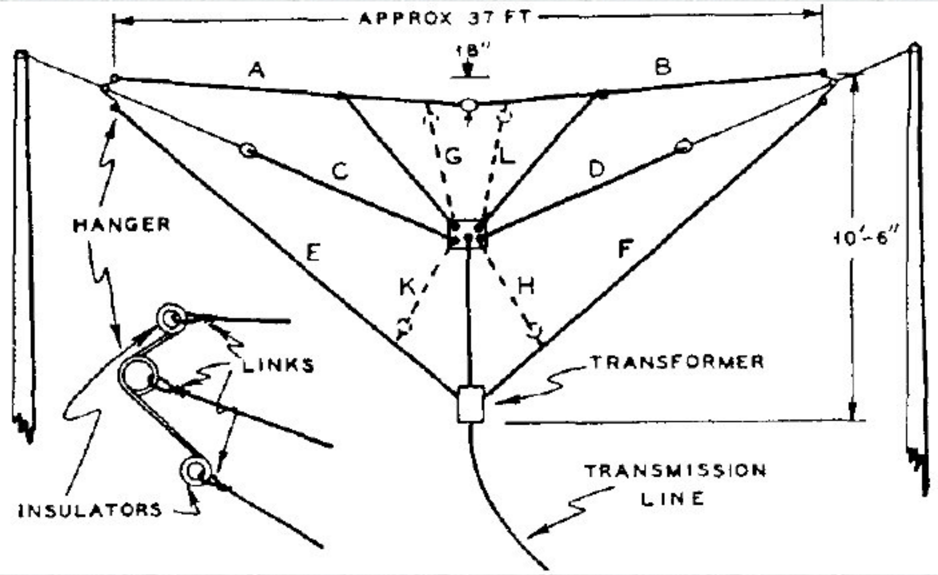
APPROX 37 FT
18"
A
B
G
L
C
D
HANGER
E
F
10'-6"
K
H
LINKS
TRANSFORMER
INSULATORS
TRANSMISSION LINE

179 - Folded Dipole with Shorted Straps

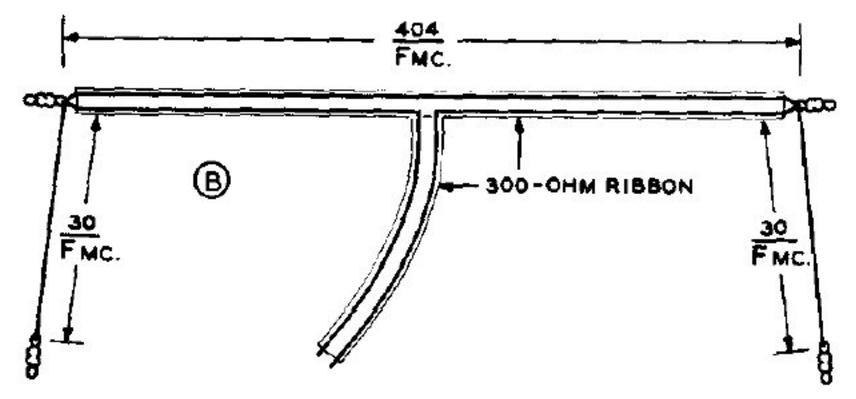
404 / FMC.
Ⓑ
300-OHM RIBBON
30 / FMC.
30 / FMC.

180 - Twin-Lead Marconi Antenna for 160m, 80m

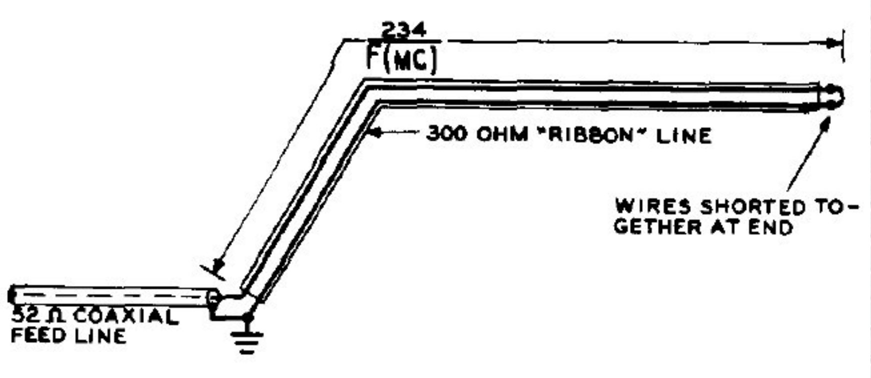
234
F(MC)
300 OHM "RIBBON" LINE
WIRES SHORTED TOGETHER AT END
52 Ω COAXIAL FEED LINE

---

181 - Broadband Antenna Quarter-Wave Balun for 80m

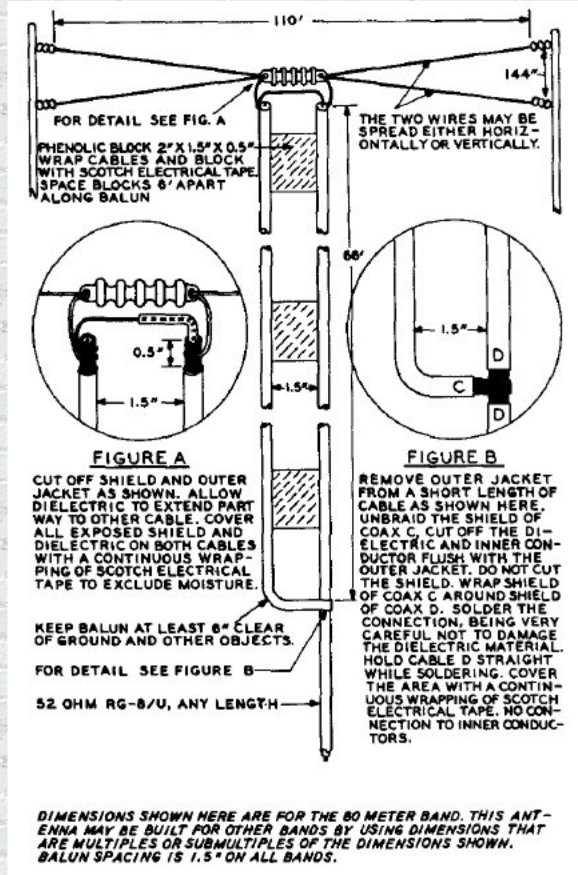
110'
144"
FOR DETAIL SEE FIG. A
THE TWO WIRES MAY BE SPREAD EITHER HORIZONTALLY OR VERTICALLY.
PHENOLIC BLOCK 2" X 1.5" X 0.5". WRAP CABLES AND BLOCK WITH SCOTCH ELECTRICAL TAPE. SPACE BLOCKS 6' APART ALONG BALUN
66'
0.5"
1.5"
1.5"
1.5"
C
D
D
FIGURE A
CUT OFF SHIELD AND OUTER JACKET AS SHOWN. ALLOW DIELECTRIC TO EXTEND PART WAY TO OTHER CABLE. COVER ALL EXPOSED SHIELD AND DIELECTRIC ON BOTH CABLES WITH A CONTINUOUS WRAPPING OF SCOTCH ELECTRICAL TAPE TO EXCLUDE MOISTURE.
FIGURE B
REMOVE OUTER JACKET FROM A SHORT LENGTH OF CABLE AS SHOWN HERE. UNBRAID THE SHIELD OF COAX C, CUT OFF THE DIELECTRIC AND INNER CONDUCTOR FLUSH WITH THE OUTER JACKET. DO NOT CUT THE SHIELD. WRAP SHIELD OF COAX C AROUND SHIELD OF COAX D. SOLDER THE CONNECTION, BEING VERY CAREFUL NOT TO DAMAGE THE DIELECTRIC MATERIAL. HOLD CABLE D STRAIGHT WHILE SOLDERING. COVER THE AREA WITH A CONTINUOUS WRAPPING OF SCOTCH ELECTRICAL TAPE. NO CONNECTION TO INNER CONDUCTORS.
KEEP BALUN AT LEAST 6" CLEAR OF GROUND AND OTHER OBJECTS.
FOR DETAIL SEE FIGURE B
52 OHM RG-8/U, ANY LENGTH
DIMENSIONS SHOWN HERE ARE FOR THE 80 METER BAND. THIS ANTENNA MAY BE BUILT FOR OTHER BANDS BY USING DIMENSIONS THAT ARE MULTIPLES OR SUBMULTIPLES OF THE DIMENSIONS SHOWN. BALUN SPACING IS 1.5" ON ALL BANDS.

---

182 - Three-Quarter-Wave Folded Doublet for Dual-Band

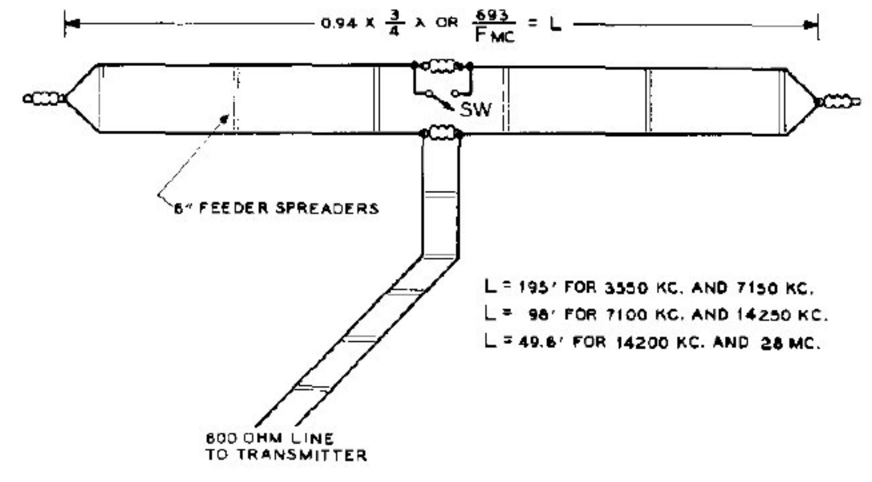
0.94 X 3/4 λ OR 693/F MC = L
SW
6" FEEDER SPREADERS
L = 195' FOR 3550 KC. AND 7150 KC.
L = 98' FOR 7100 KC. AND 14250 KC.
L = 49.6' FOR 14200 KC. AND 28 MC.
600 OHM LINE
TO TRANSMITTER

---

183 - Three-Quarter-Wave Folded Doublet No-Switch for Dual-Band

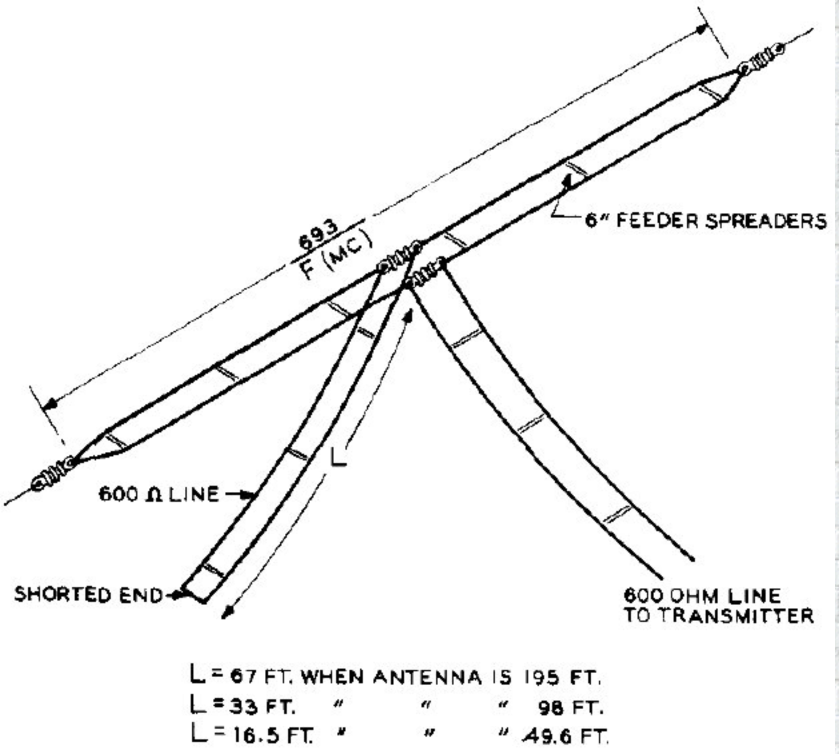
693/F (MC)
6" FEEDER SPREADERS
600 Ω LINE
L
SHORTED END
600 OHM LINE
TO TRANSMITTER
L = 67 FT. WHEN ANTENNA IS 195 FT.
L = 33 FT. " " " 98 FT.
L = 16.5 FT. " " " 49.6 FT.

---

184 - Wideband Omni-Directional Discone Antenna

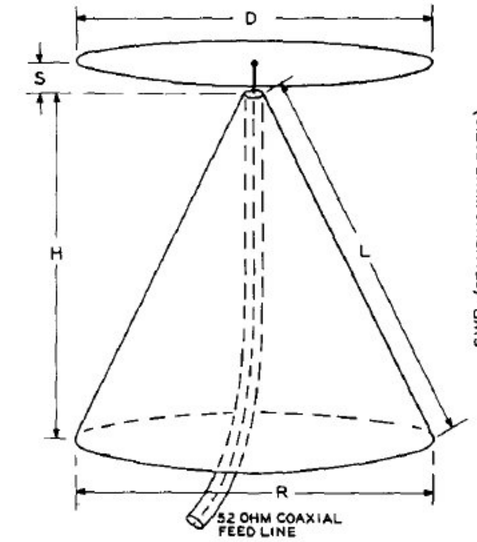
D
S
H
L
R
52 OHM COAXIAL FEED LINE

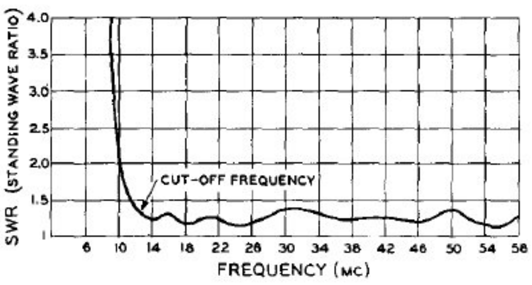

DIMENSIONS

20, 15, 11, 10, 6 METERS
D = 12′ L = 18′
S = 10″ R = 18′
H = 15′7″

15, 11, 10, 6 METERS
D = 8′ L = 12′
S = 6″ R = 12′
H = 10′5″

11, 10, 6, 2 METERS
D = 6′ L = 9′6″
S = 4″ R = 9′6″
H = 8′3″

185 - Wideband Rhombic Antenna for 40m to 10m

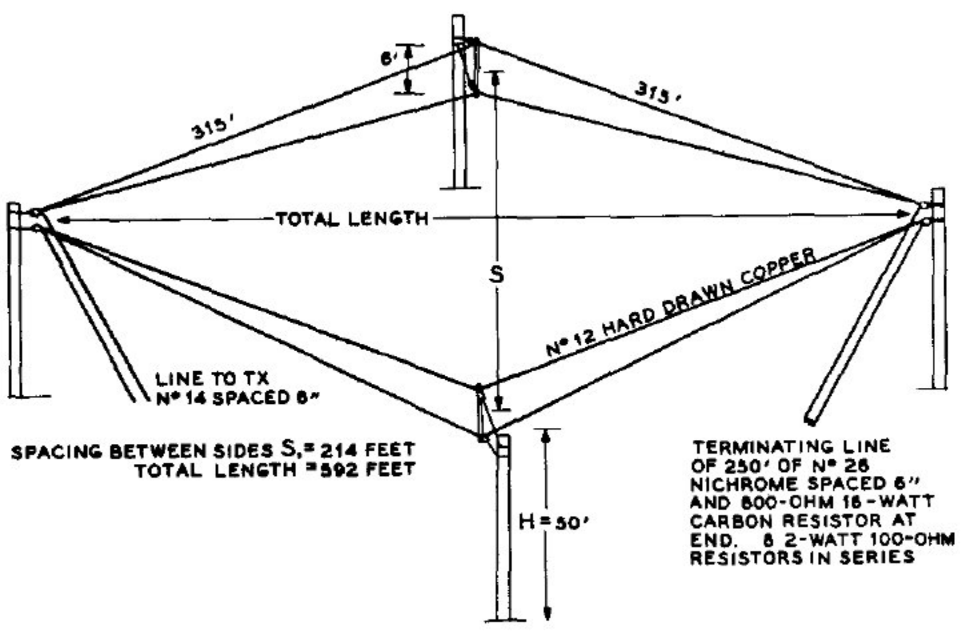
6′
315′
315′
TOTAL LENGTH
S
N° 12 HARD DRAWN COPPER
LINE TO TX
N° 14 SPACED 6″
SPACING BETWEEN SIDES S, = 214 FEET
TOTAL LENGTH = 592 FEET
H = 50′
TERMINATING LINE OF 250′ OF N° 26 NICHROME SPACED 6″ AND 800-OHM 16-WATT CARBON RESISTOR AT END. 8 2-WATT 100-OHM RESISTORS IN SERIES

186 - Pre-Cut Linear Array Antenna 3dB-Gain for 40m

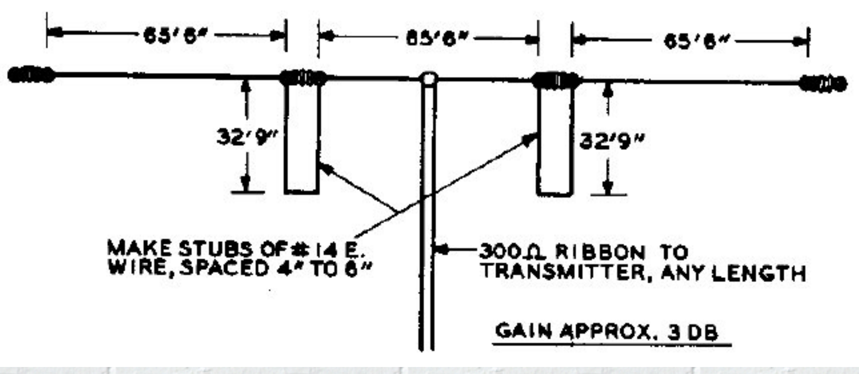
65′6″
65′6″
65′6″
32′9″
32′9″
MAKE STUBS OF # 14 E. WIRE, SPACED 4″ TO 6″
300Ω RIBBON TO TRANSMITTER, ANY LENGTH
GAIN APPROX. 3 DB

187 - X-Array Antenna 6dB-Gain for 20m, 15m, 10m

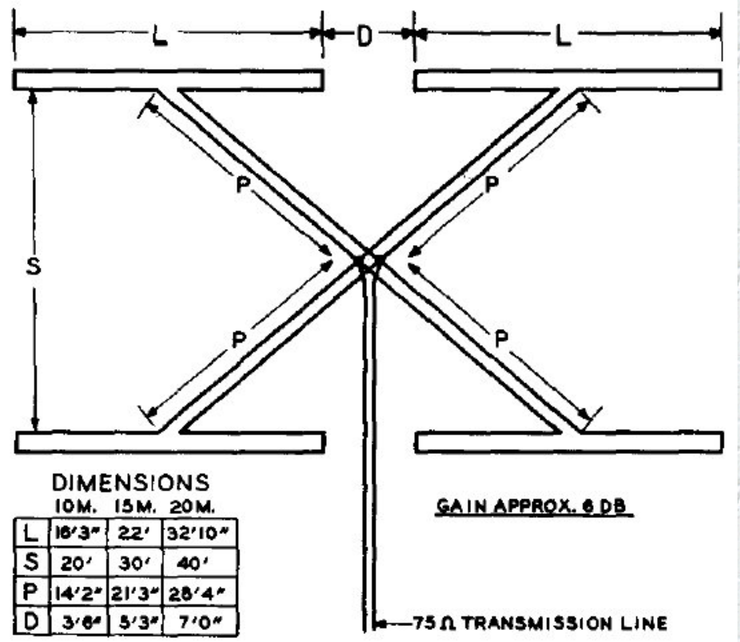
L
D
L
S
P
P
P
P
GAIN APPROX. 6 DB
75 Ω TRANSMISSION LINE

DIMENSIONS

	10M.	15M.	20M.
L	16'3"	22'	32'10"
S	20'	30'	40'
P	14'2"	21'3"	28'4"
D	3'6"	5'3"	7'0"

The entire array (with the exception of the 75-ohm feed line) is constructed of 300-ohm ribbon line. Be sure phasing lines (P) are poled correctly, as shown.

188 - Double-Bruce Array Antenna 5dB-Gain for 20m, 10m

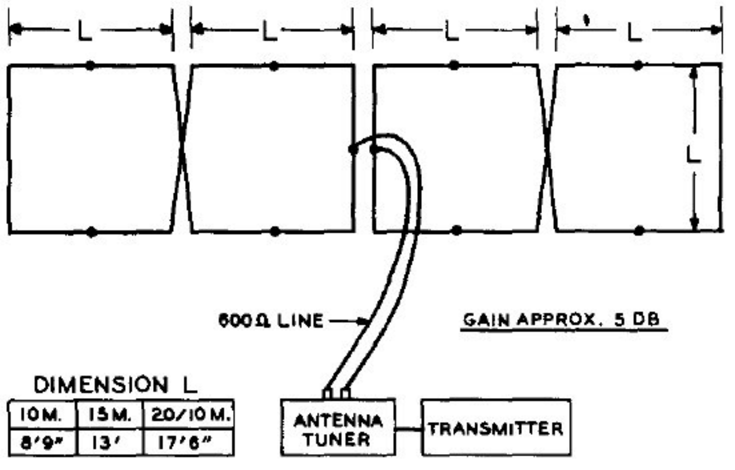
L
L
L
L
L
600 Ω LINE
GAIN APPROX. 5 DB
ANTENNA TUNER
TRANSMITTER

DIMENSION L

10M.	15M.	20/10 M.
8'9"	13'	17'6"

189 - Bi-Square Broadside Array 4dB-Gain for 20m, 15m, 10m

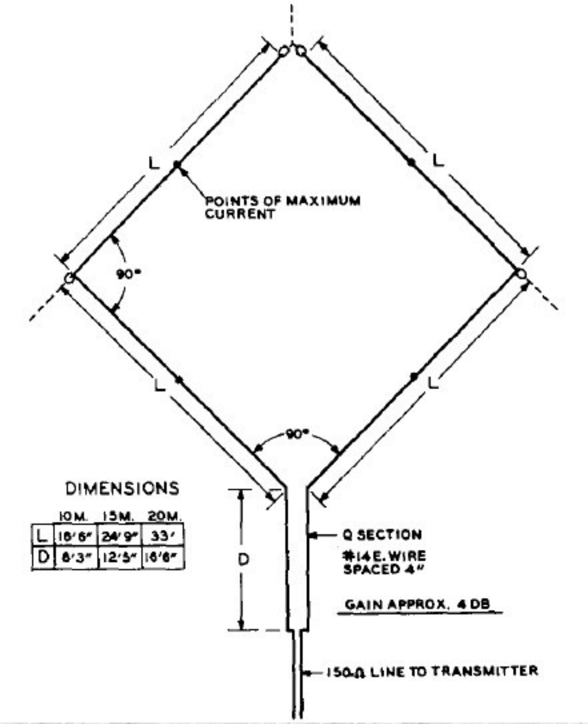
L
L
POINTS OF MAXIMUM CURRENT
90°
L
L
90°
D
Q SECTION
#14 E. WIRE SPACED 4"
GAIN APPROX. 4 DB
150-Ω LINE TO TRANSMITTER

DIMENSIONS

	10 M.	15 M.	20 M.
L	16'6"	24'9"	33'
D	8'3"	12'5"	16'6"

---

190 - Six-Shooter Broadside Array 7.5dB-Gain for 20m, 15m, 10m

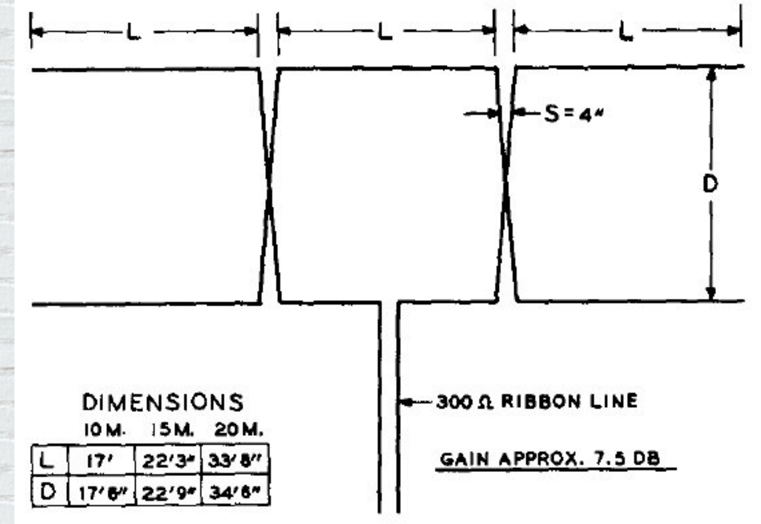
L
L
L
S = 4"
D
300 Ω RIBBON LINE
GAIN APPROX. 7.5 DB

DIMENSIONS

	10 M.	15 M.	20 M.
L	17'	22'3"	33'8"
D	17'6"	22'9"	34'6"

---

191 - Triplex Flat-Top Beam 4.5dB-Gain for 20m, 15m, 10m

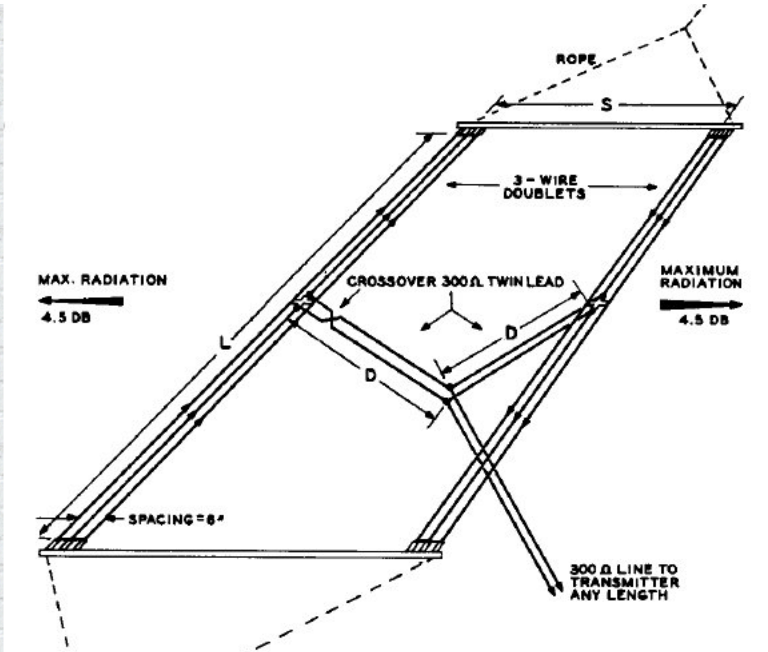
ROPE
S
3 - WIRE DOUBLETS
MAX. RADIATION
4.5 DB
CROSSOVER 300Ω TWIN LEAD
MAXIMUM RADIATION
4.5 DB
L
D
D
SPACING = 6"
300 Ω LINE TO TRANSMITTER ANY LENGTH

DIMENSIONS

	10M.	15M.	20M	MATERIAL
L	15'8"	21'5"	32'2"	THREE # 12 E. WIRES, SPACED 3"
S	5'6"	7'8"	11'	
D	7'2"	10'7"	14'4"	300 OHM RIBBON

---

192 - Dual-Band Tilt Antenna for 20m, 10m

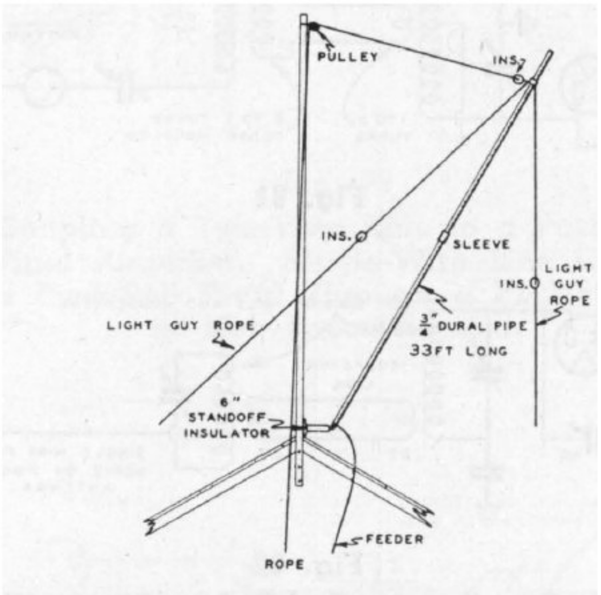
PULLEY
INS.
INS.
SLEEVE
LIGHT GUY ROPE
INS.
LIGHT GUY ROPE
3/4" DURAL PIPE
33FT LONG
6" STANDOFF INSULATOR
FEEDER
ROPE

TWO BAND (10-20 METER) TILT ANTENNA — VERTICAL FOR 20 METERS - TILT 54° FROM HORIZON FOR 10 METERS

---

193 - Super Space Multiband Dipole Antenna for 80m to 10m

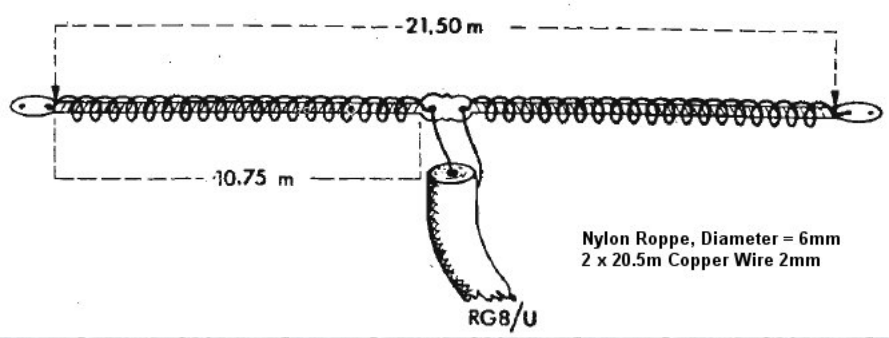
21.50 m
10.75 m
Nylon Roppe, Diameter = 6mm
2 x 20.5m Copper Wire 2mm
RG8/U

194 - Bi-Square Beam Antenna Gain=5.5dB for 10m

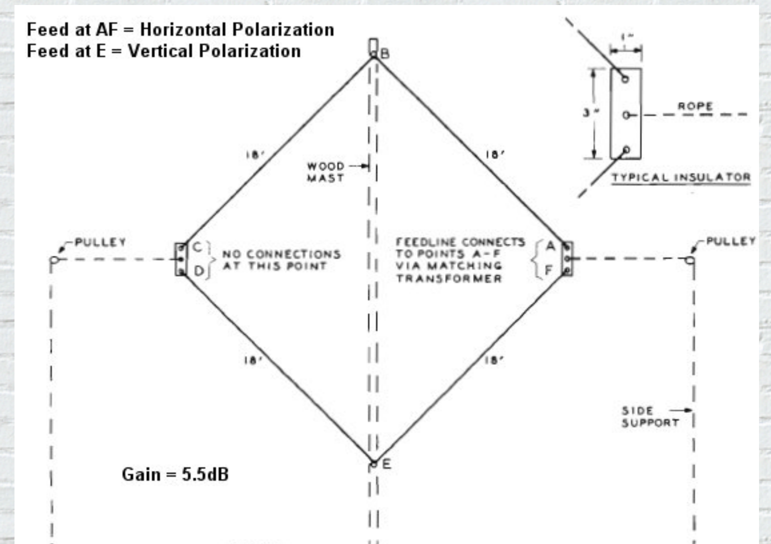
Feed at AF = Horizontal Polarization
Feed at E = Vertical Polarization
B
18'
18'
WOOD MAST
1"
3"
ROPE
TYPICAL INSULATOR
PULLEY
C
D
NO CONNECTIONS AT THIS POINT
FEEDLINE CONNECTS TO POINTS A-F VIA MATCHING TRANSFORMER
A
F
PULLEY
18'
18'
SIDE SUPPORT
E
Gain = 5.5dB
GROUND

195 - Cousin of G5RV Multiband Antenna for 40m to 10m

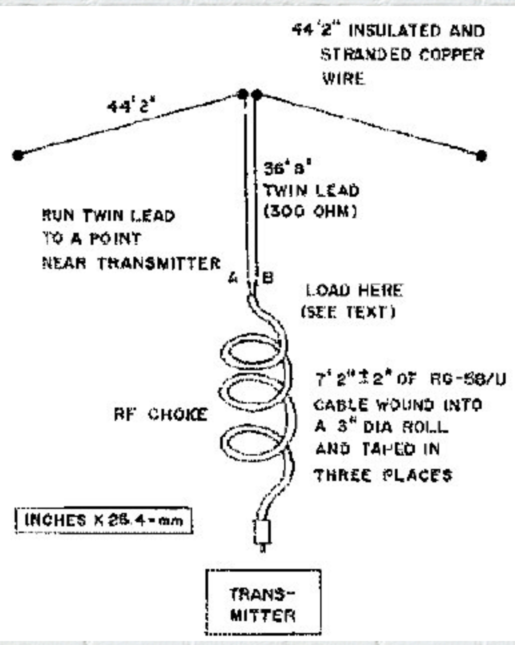
44'2" INSULATED AND STRANDED COPPER WIRE
44'2"
36'8" TWIN LEAD (300 OHM)
RUN TWIN LEAD TO A POINT NEAR TRANSMITTER
A
B
LOAD HERE (SEE TEXT)
7'2"±2" OF RG-58/U CABLE WOUND INTO A 3" DIA ROLL AND TAPED IN THREE PLACES
RF CHOKE
INCHES X 25.4=mm
TRANS-MITTER

196 - Cayman Quad Antenna for 20m

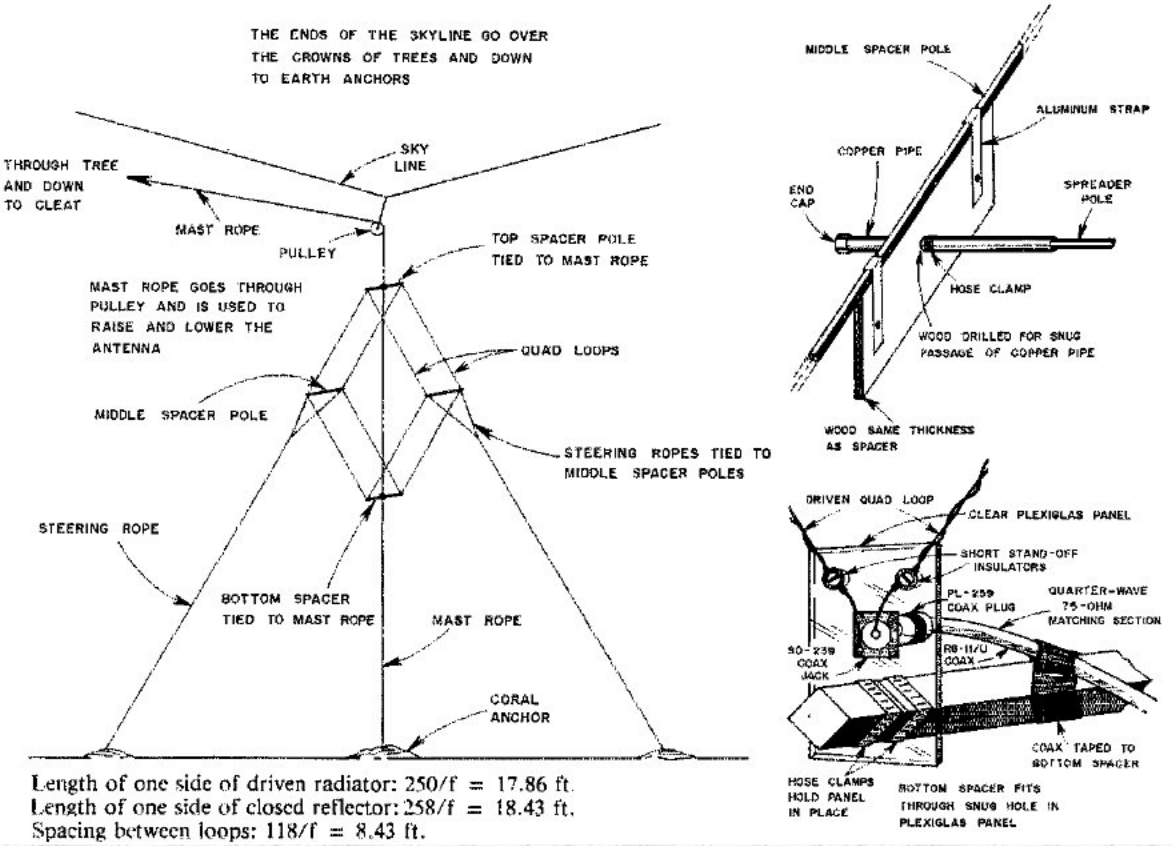
THE ENDS OF THE SKYLINE GO OVER THE CROWNS OF TREES AND DOWN TO EARTH ANCHORS
SKY LINE
THROUGH TREE AND DOWN TO CLEAT
MAST ROPE
PULLEY
TOP SPACER POLE TIED TO MAST ROPE
MAST ROPE GOES THROUGH PULLEY AND IS USED TO RAISE AND LOWER THE ANTENNA
QUAD LOOPS
MIDDLE SPACER POLE
STEERING ROPES TIED TO MIDDLE SPACER POLES
STEERING ROPE
BOTTOM SPACER TIED TO MAST ROPE
MAST ROPE
CORAL ANCHOR
MIDDLE SPACER POLE
ALUMINUM STRAP
COPPER PIPE
END CAP
SPREADER POLE
HOSE CLAMP
WOOD DRILLED FOR SNUG PASSAGE OF COPPER PIPE
WOOD SAME THICKNESS AS SPACER
DRIVEN QUAD LOOP
CLEAR PLEXIGLAS PANEL
SHORT STAND-OFF INSULATORS
PL-259 COAX PLUG
QUARTER-WAVE 75-OHM MATCHING SECTION
SO-239 COAX JACK
RG-11/U COAX
COAX TAPED TO BOTTOM SPACER
HOSE CLAMPS HOLD PANEL IN PLACE
BOTTOM SPACER FITS THROUGH SNUG HOLE IN PLEXIGLAS PANEL

Length of one side of driven radiator: 250/f = 17.86 ft.
Length of one side of closed reflector: 258/f = 18.43 ft.
Spacing between loops: 118/f = 8.43 ft.

---

197 - Hentenna Gain=3dB with Bazooka Match for 6m

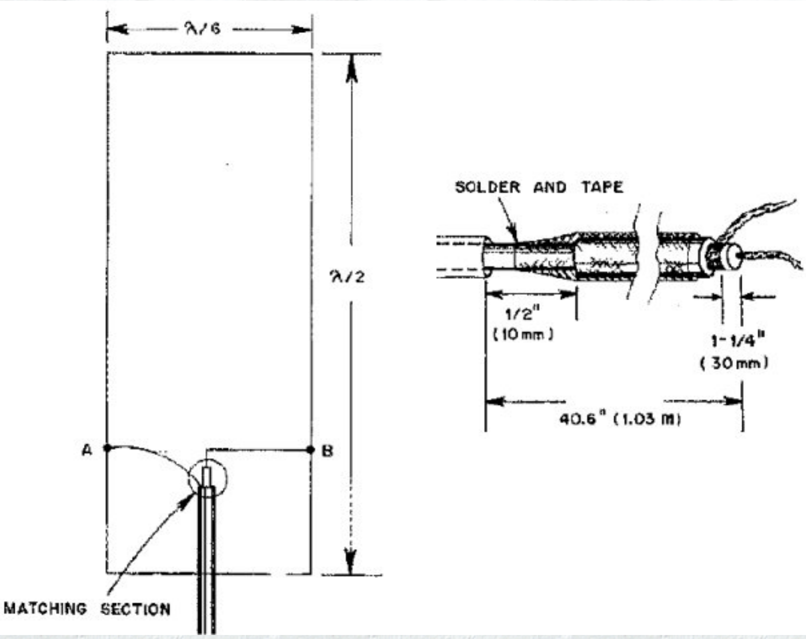
λ/6
λ/2
A
B
MATCHING SECTION
SOLDER AND TAPE
1/2" (10 mm)
1-1/4" (30 mm)
40.6" (1.03 m)

---

198 - X-Beam Antenna Gain=3dB for 20m

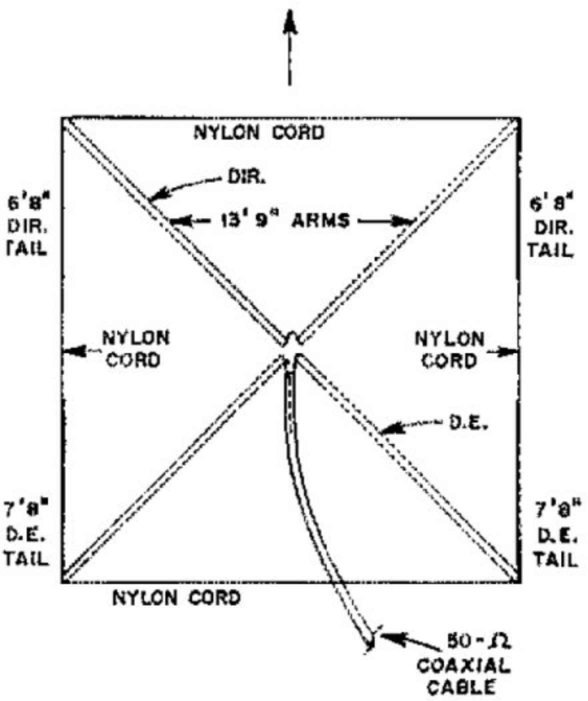
NYLON CORD
DIR.
13' 9" ARMS
6'8" DIR. TAIL
6'8" DIR. TAIL
NYLON CORD
NYLON CORD
D.E.
7'8" D.E. TAIL
7'8" D.E. TAIL
NYLON CORD
50-Ω COAXIAL CABLE

---

199 - Twin Delta Loop Antenna Gain=6dB for 160m, 80m, 40m

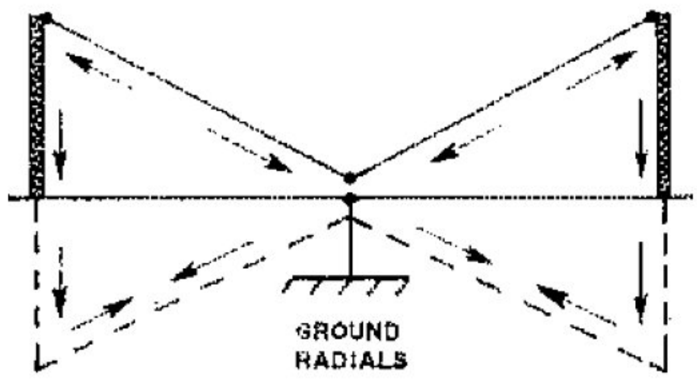
GROUND
RADIALS

Twin- Delta Loop, with the image in the ground plane.

Band	*Tower Height (ft)*	*Length of Sloping Wires (ft)*
160 m†	102.03	204.06
80 m	49.86	99.72
40 m	25.92	51.84

---

200 - Inverted-V Beam Antenna for 30m

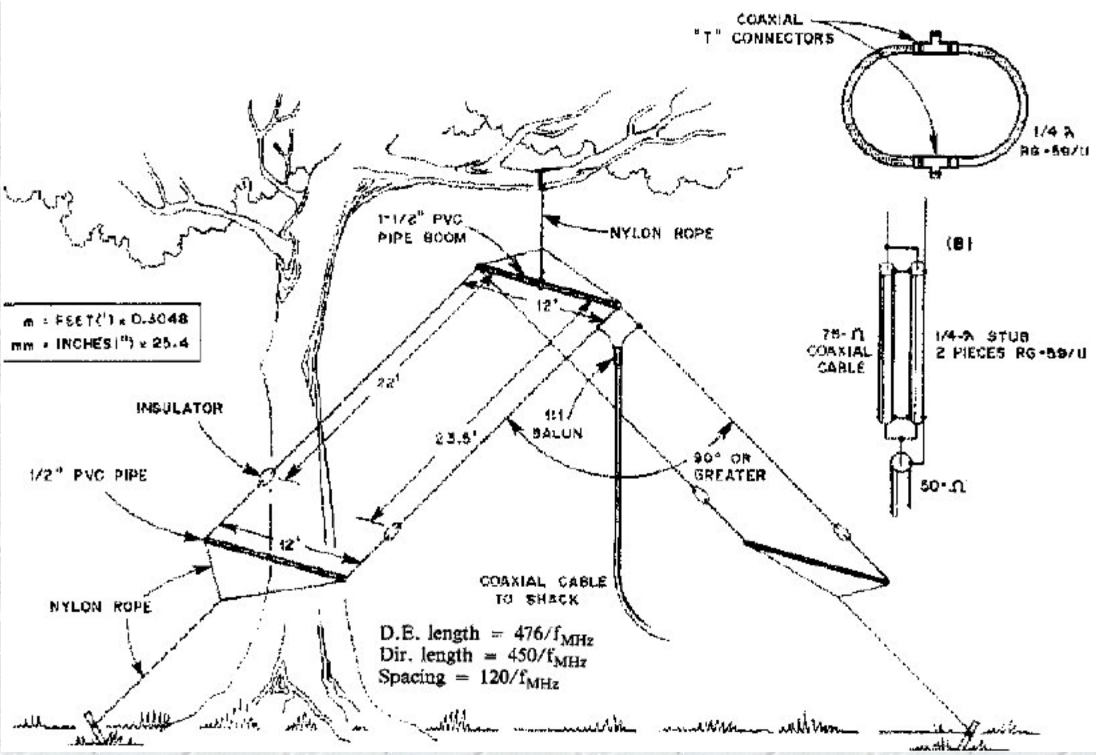
COAXIAL "T" CONNECTORS
1/4 λ RG-59/U
(B)
75-Ω COAXIAL CABLE
1/4-λ STUB 2 PIECES RG-59/U
50-Ω
1-1/2" PVC PIPE BOOM
NYLON ROPE
m = FEET(') x 0.3048
mm = INCHES(") x 25.4
12'
12'
22'
23.5'
INSULATOR
1:1 BALUN
90° OR GREATER
1/2" PVC PIPE
12'
NYLON ROPE
COAXIAL CABLE TO SHACK
D.E. length = 476/f_{MHz}
Dir. length = 450/f_{MHz}
Spacing = 120/f_{MHz}

---

201 - FD4 Windom Antenna for 80m, 40m, 20m, 10m

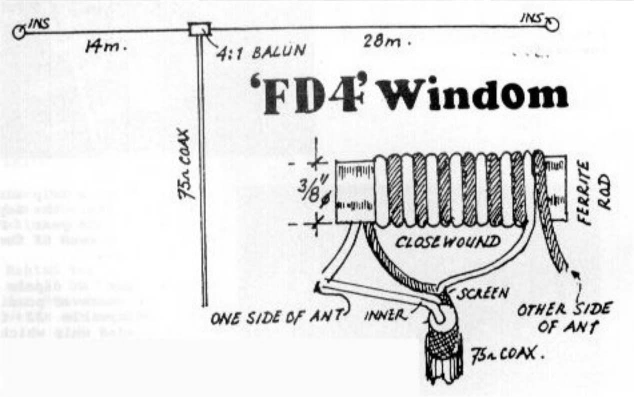
INS
14m.
4:1 BALUN
28m.
INS
'FD4' Windom
75Ω COAX
3/8" ⌀
CLOSEWOUND
FERRITE ROD
SCREEN
ONE SIDE OF ANT
INNER
OTHER SIDE OF ANT
75Ω COAX.

---

202 - Two Elements Phased Delta Loop for 40m

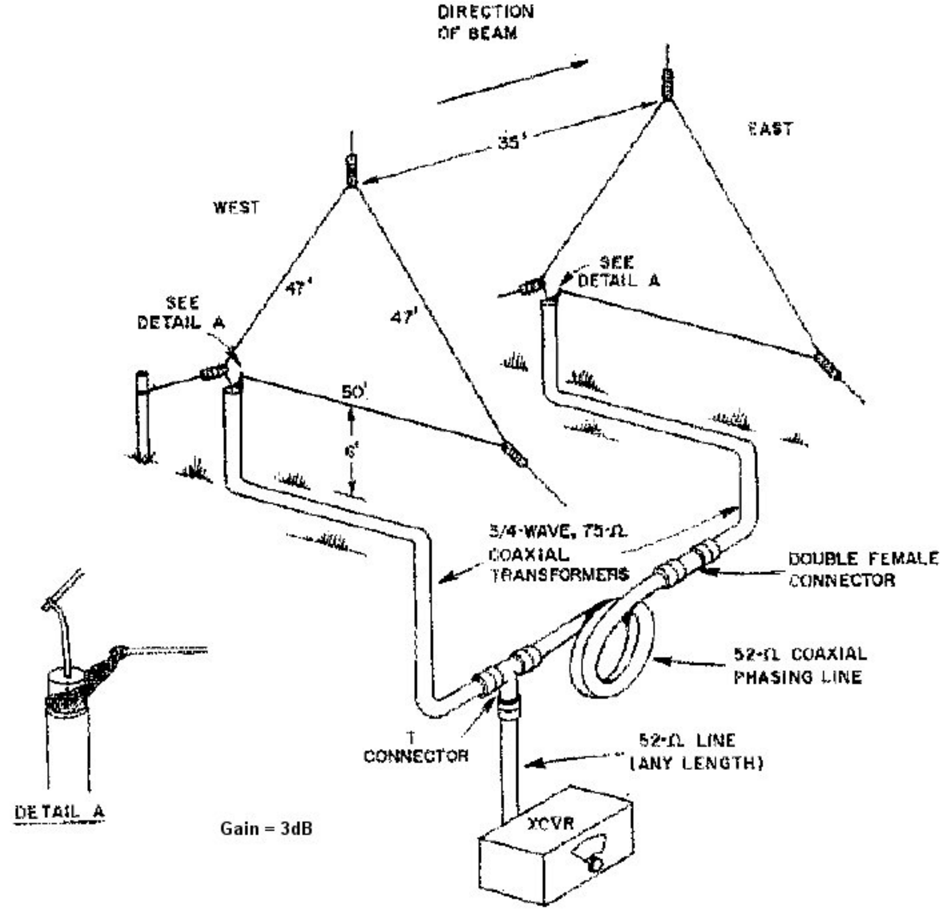
DIRECTION
OF BEAM
35'
EAST
WEST
SEE
DETAIL A
47'
47'
SEE
DETAIL A
50'
6'
3/4-WAVE, 75-Ω
COAXIAL
TRANSFORMERS
DOUBLE FEMALE
CONNECTOR
52-Ω COAXIAL
PHASING LINE
T
CONNECTOR
52-Ω LINE
(ANY LENGTH)
DETAIL A
Gain = 3dB
XCVR

203 - Three Elements Wire Yagi for 10m

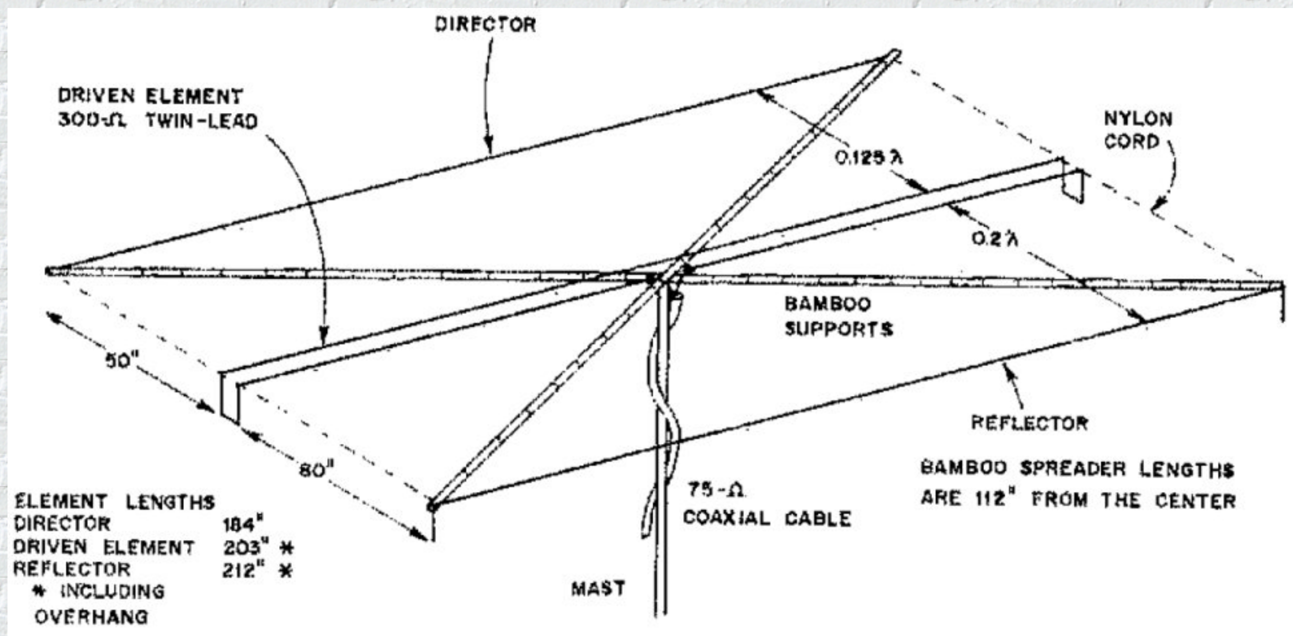
DIRECTOR
DRIVEN ELEMENT
300-Ω TWIN-LEAD
NYLON
CORD
0.125 λ
0.2 λ
BAMBOO
SUPPORTS
50"
80"
REFLECTOR
BAMBOO SPREADER LENGTHS
ARE 112" FROM THE CENTER
75-Ω
COAXIAL CABLE
ELEMENT LENGTHS
DIRECTOR 184"
DRIVEN ELEMENT 203" *
REFLECTOR 212" *
* INCLUDING
OVERHANG
MAST

204 - Low Radiation Angle Full-Wave Loop Antenna for 80m

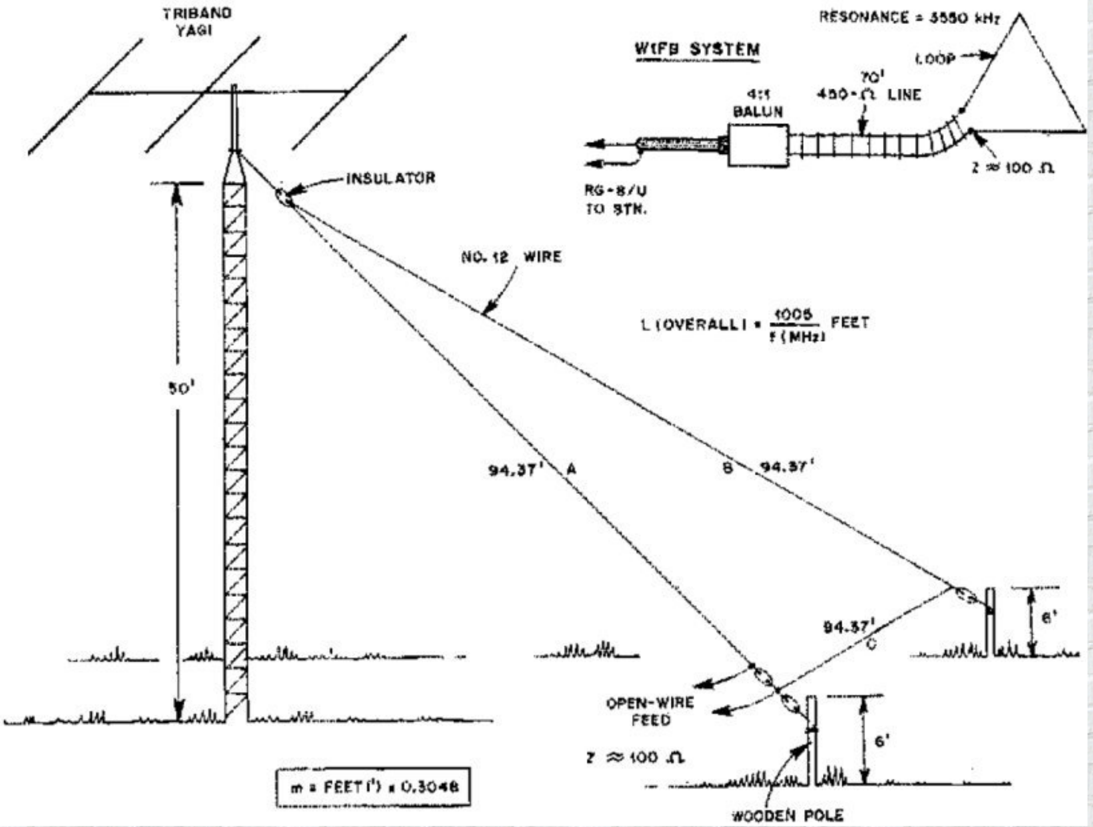
TRIBAND YAGI
INSULATOR
W1FB SYSTEM
RESONANCE = 3550 kHz
LOOP
70'
4:1 BALUN
450-Ω LINE
RG-8/U TO STN.
Z ≈ 100 Ω
NO. 12 WIRE
L (OVERALL) = 1005 / f (MHz) FEET
50'
94.37' A
B 94.37'
94.37' C
6'
OPEN-WIRE FEED
Z ≈ 100 Ω
6'
WOODEN POLE
m = FEET (') x 0.3048

205 - Collinear Antenna Gain=3dB for 2m

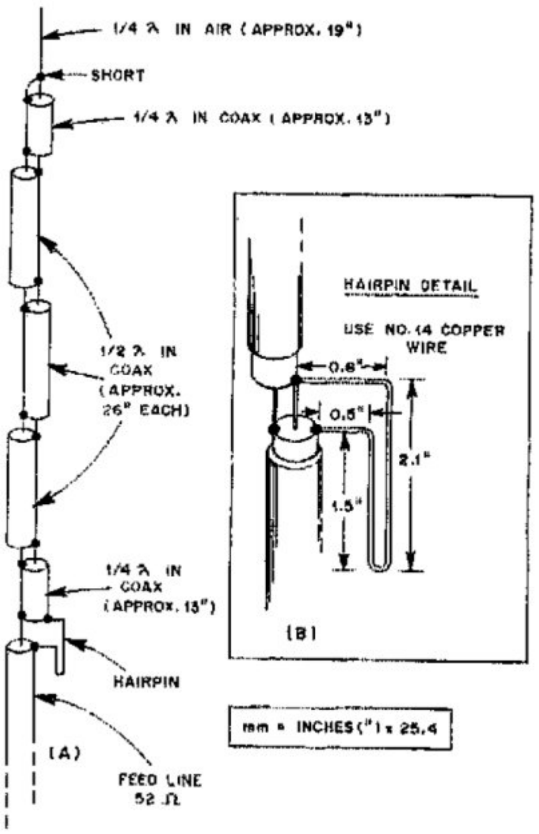
1/4 λ IN AIR (APPROX. 19")
SHORT
1/4 λ IN COAX (APPROX. 13")
1/2 λ IN COAX (APPROX. 26" EACH)
1/4 λ IN COAX (APPROX. 13")
HAIRPIN
(A)
FEED LINE 52 Ω
HAIRPIN DETAIL
USE NO. 14 COPPER WIRE
0.8"
0.5"
2.1"
1.5"
(B)
mm = INCHES (") x 25.4

206 - X-Ray Monoband Antenna Gain=6.5dBd

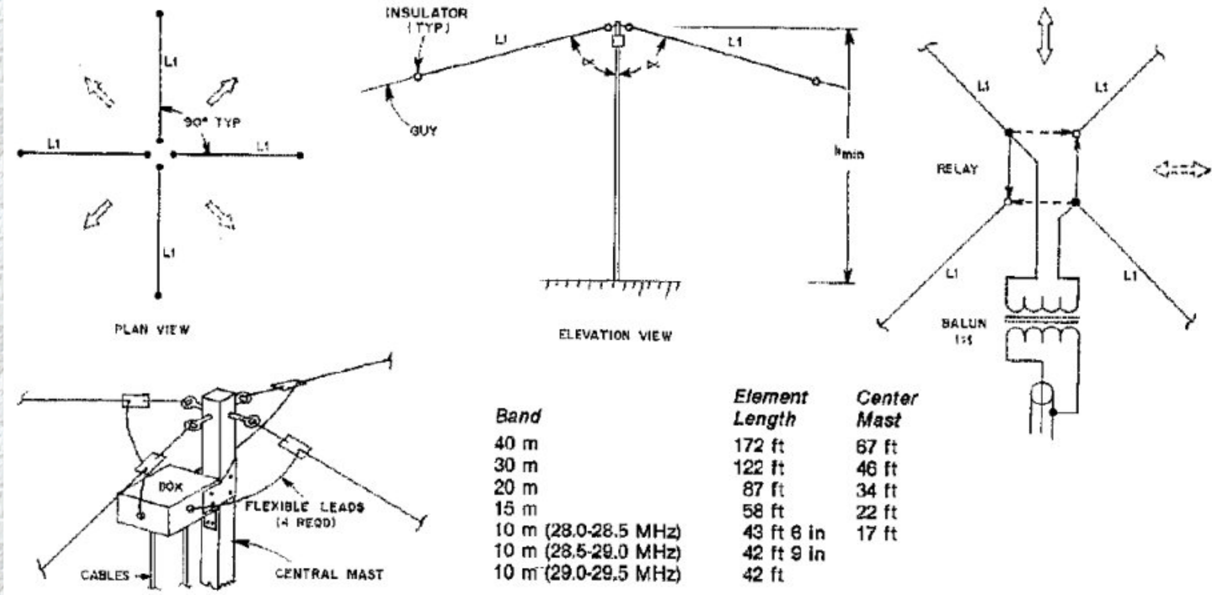
INSULATOR (TYP)
L1
L1
GUY
α
α
h_min
ELEVATION VIEW
L1
L1
L1
L1
90° TYP
PLAN VIEW
L1
L1
L1
L1
RELAY
BALUN 1:1
BOX
FLEXIBLE LEADS (4 REQD)
CABLES
CENTRAL MAST

Band	*Element Length*	*Center Mast*
40 m	172 ft	67 ft
30 m	122 ft	46 ft
20 m	87 ft	34 ft
15 m	58 ft	22 ft
10 m (28.0-28.5 MHz)	43 ft 6 in	17 ft
10 m (28.5-29.0 MHz)	42 ft 9 in	
10 m (29.0-29.5 MHz)	42 ft	

207 - Broadside Collinear Curtain Array for 20m

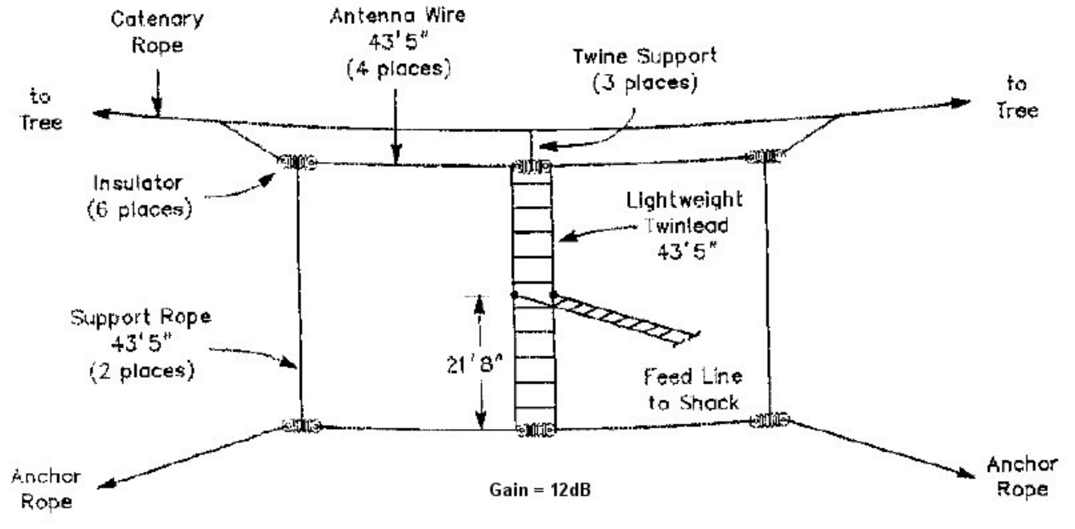
Catenary Rope
Antenna Wire 43'5" (4 places)
Twine Support (3 places)
to Tree
to Tree
Insulator (6 places)
Lightweight Twinlead 43'5"
Support Rope 43'5" (2 places)
21'8"
Feed Line to Shack
Anchor Rope
Anchor Rope
Gain = 12dB

208 - Shrunk Loop Antenna for 80m, 40m

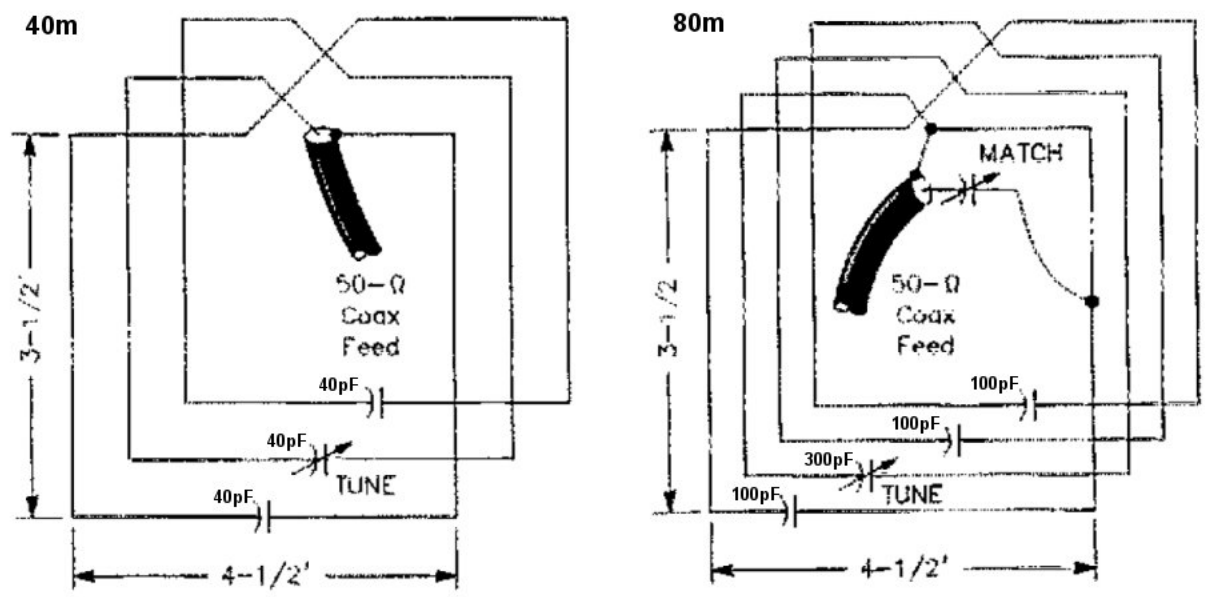
40m
50-Ω
Coax
Feed
40pF
40pF
TUNE
40pF
3-1/2'
4-1/2'
80m
MATCH
50-Ω
Coax
Feed
100pF
100pF
300pF
TUNE
100pF
3-1/2
4-1/2'

---

209 - λ/3 Multiband Dipole for 40m, 20m, 17m, 15m, 13m, 10m

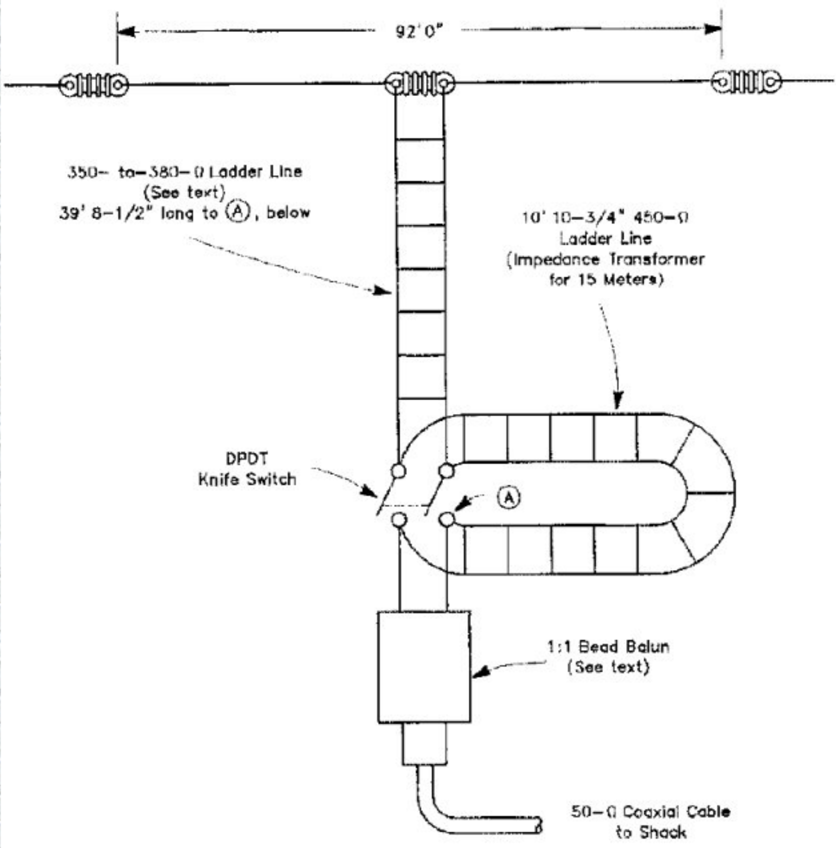
92' 0"
350- to-380-Ω Ladder Line
(See text)
39' 8-1/2" long to Ⓐ, below
10' 10-3/4" 450-Ω
Ladder Line
(Impedance Transformer
for 15 Meters)
DPDT
Knife Switch
Ⓐ
1:1 Bead Balun
(See text)
50-Ω Coaxial Cable
to Shack

---

210 - K3LR Sloper Antenna Gain=3dB for 160m

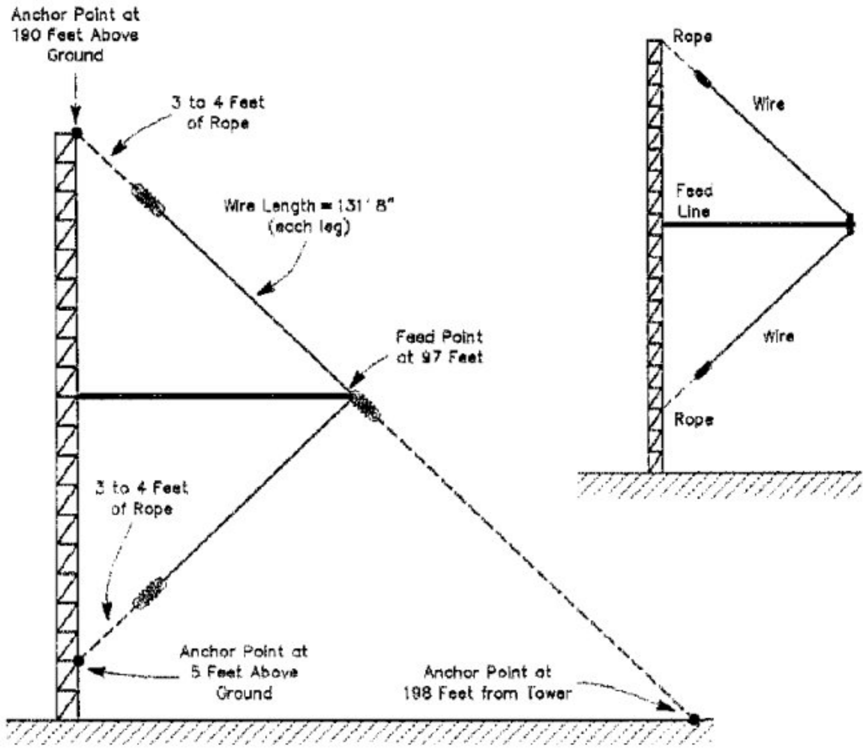
Anchor Point at 190 Feet Above Ground
3 to 4 Feet of Rope
Wire Length = 131' 8" (each leg)
Feed Point at 97 Feet
3 to 4 Feet of Rope
Anchor Point at 5 Feet Above Ground
Anchor Point at 198 Feet from Tower
Rope
Wire
Feed Line
Wire
Rope

211 - Mini Folded Vertical Monopole Antenna for 160m, 80m, 40m

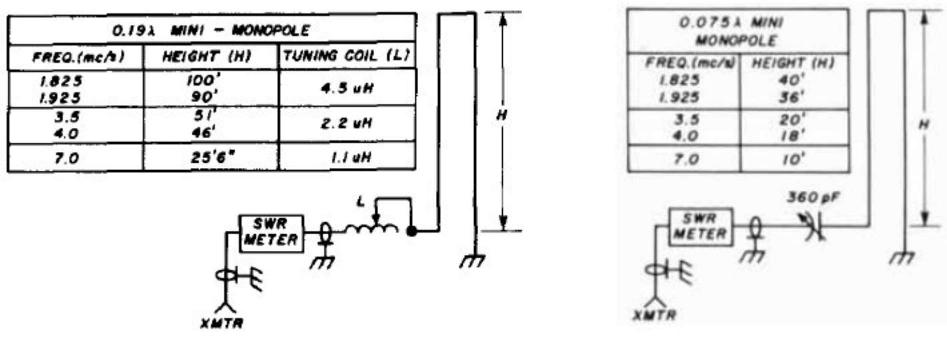
H
L
SWR METER
XMTR
H
360 pF
SWR METER
XMTR

0.19λ MINI - MONOPOLE		
FREQ.(mc/s)	HEIGHT (H)	TUNING COIL (L)
1.825 1.925	100' 90'	4.5 uH
3.5 4.0	51' 46'	2.2 uH
7.0	25'6"	1.1 uH

0.075λ MINI MONOPOLE	
FREQ.(mc/s)	HEIGHT (H)
1.825 1.925	40' 36'
3.5 4.0	20' 18'
7.0	10'

212 - Three Band Dipole for 80m, 40m, 20m

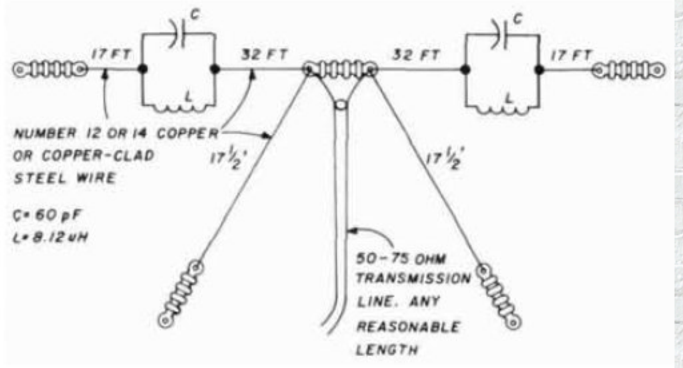
C
C
17 FT
32 FT
32 FT
17 FT
L
L
NUMBER 12 OR 14 COPPER OR COPPER-CLAD STEEL WIRE
17½'
17½'
C= 60 pF
L= 8.12 uH
50-75 OHM TRANSMISSION LINE. ANY REASONABLE LENGTH

213 - Stub Matched Dipole Antenna

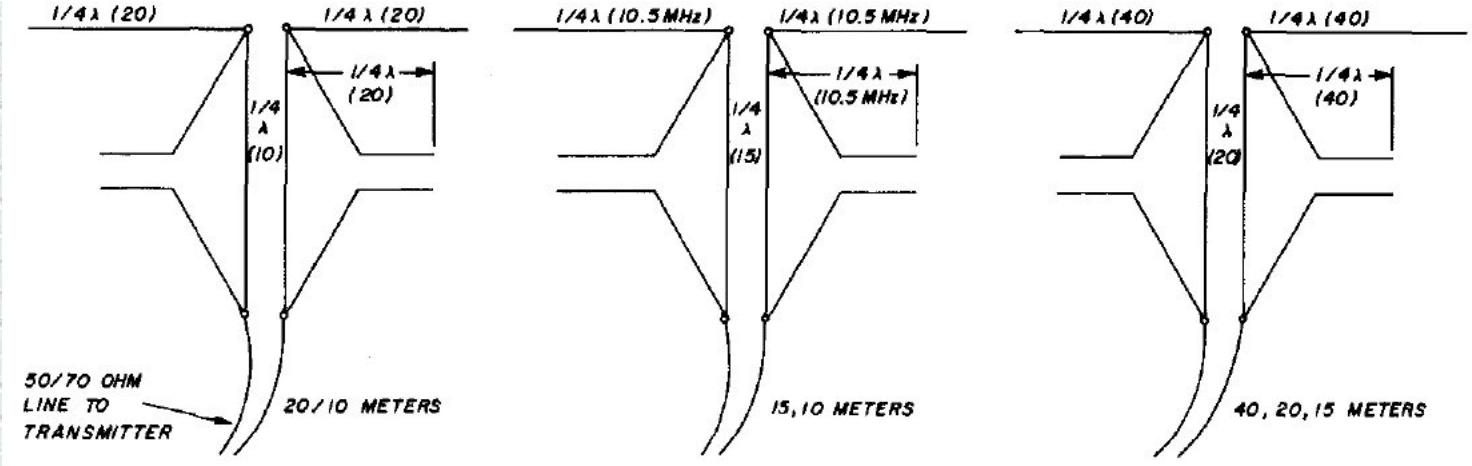
1/4 λ (20)
1/4 λ (20)
1/4 λ (20)
1/4 λ (10)
50/70 OHM LINE TO TRANSMITTER
20/10 METERS
1/4 λ (10.5 MHz)
1/4 λ (10.5 MHz)
1/4 λ (10.5 MHz)
1/4 λ (15)
15, 10 METERS
1/4 λ (40)
1/4 λ (40)
1/4 λ (40)
1/4 λ (20)
40, 20, 15 METERS

214 - Inverted-Vee Antenna with Bazooka Match

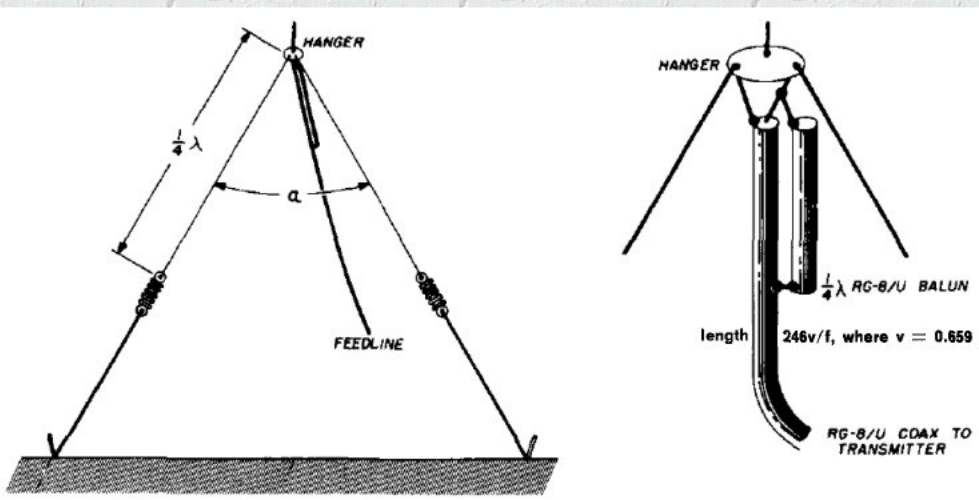
HANGER
$\frac{1}{4}\lambda$
a
FEEDLINE
HANGER
$\frac{1}{4}\lambda$ RG-8/U BALUN
length
246v/f, where v = 0.659
RG-8/U COAX TO TRANSMITTER

215 - Multiband Long Wire Antenna for 40m, 20m, 15m, 10m

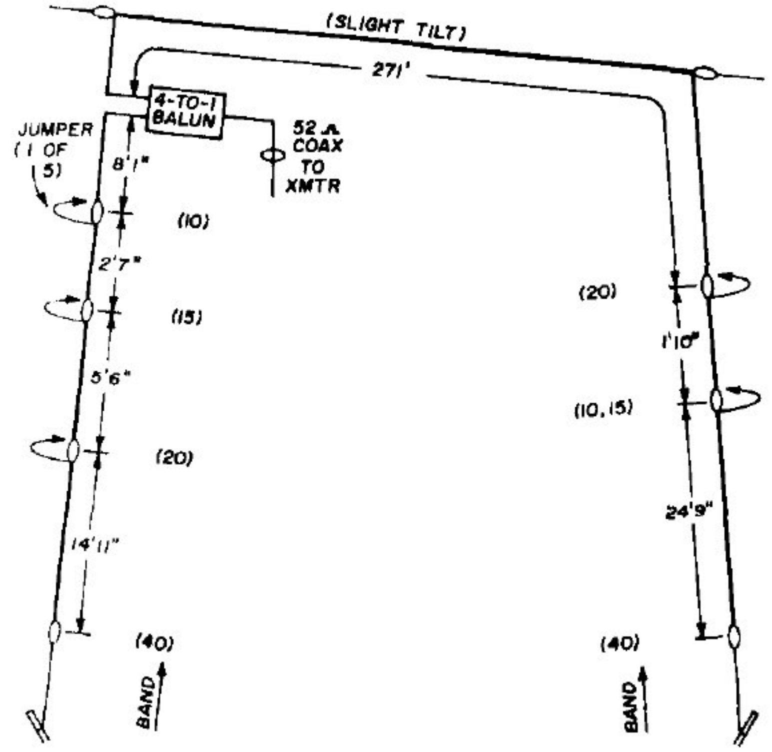
(SLIGHT TILT)
271'
4-TO-1 BALUN
52 Ω COAX TO XMTR
JUMPER (1 OF 5)
8'1"
(10)
2'7"
(15)
5'6"
(20)
14'11"
(40)
BAND
(20)
1'10"
(10, 15)
24'9"
(40)
BAND

216 - Converted Vee Antenna for 80m, 40m

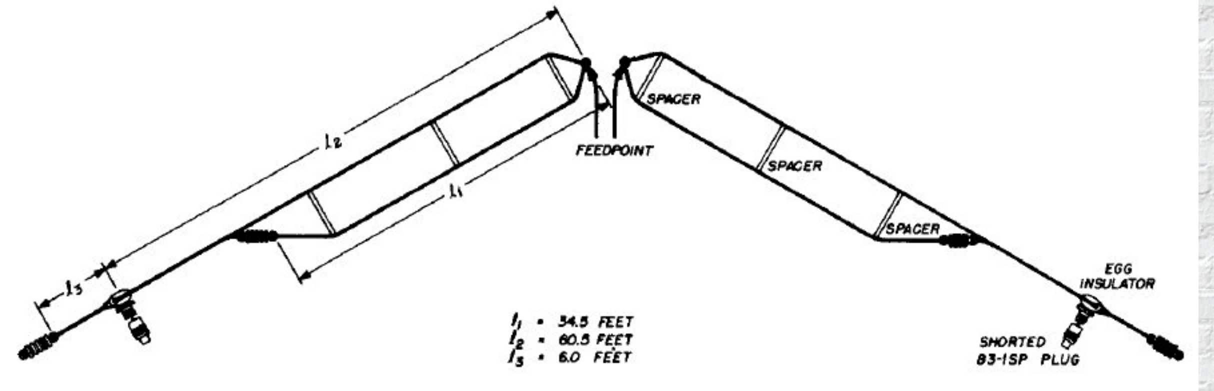
l_2
l_1
l_3
FEEDPOINT
SPACER
SPACER
SPACER
EGG INSULATOR
SHORTED 83-1SP PLUG
l_1 = 34.5 FEET
l_2 = 60.5 FEET
l_3 = 6.0 FEET

217 - Swiss Quad Antenna for 20m, 15m, 10m

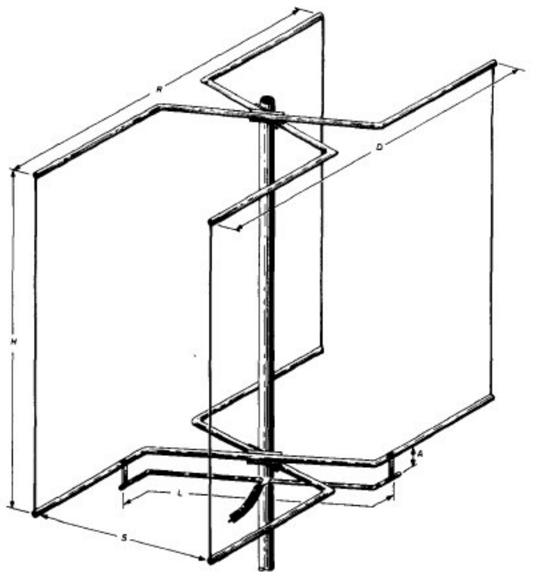

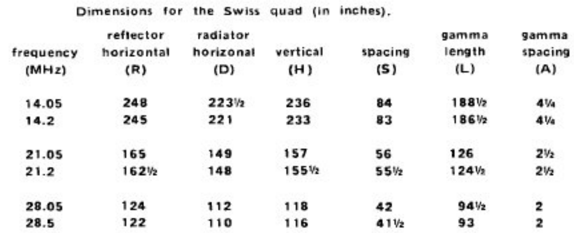

Dimensions for the Swiss quad (in inches).

frequency (MHz)	reflector horizontal (R)	radiator horizonal (D)	vertical (H)	spacing (S)	gamma length (L)	gamma spacing (A)
14.05	248	223½	236	84	188½	4¼
14.2	245	221	233	83	186½	4¼
21.05	165	149	157	56	126	2½
21.2	162½	148	155½	55½	124½	2½
28.05	124	112	118	42	94½	2
28.5	122	110	116	41½	93	2

218 - Japanese Quad Antenna for 20m, 15m, 10m

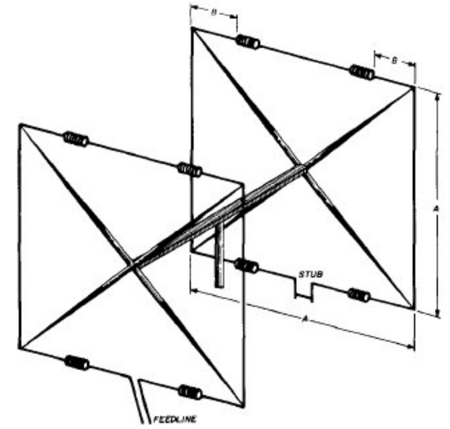
B
B
A
A
STUB
FEEDLINE

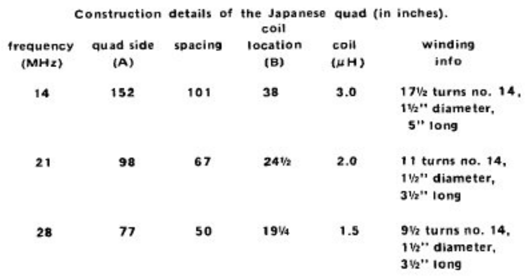

Construction details of the Japanese quad (in inches).

frequency (MHz)	quad side (A)	spacing	coil location (B)	coil (μH)	winding info
14	152	101	38	3.0	17½ turns no. 14, 1½'' diameter, 5'' long
21	98	67	24½	2.0	11 turns no. 14, 1½'' diameter, 3½'' long
28	77	50	19¼	1.5	9½ turns no. 14, 1½'' diameter, 3½'' long

219 - Seven-Elements Triangle Beam Antenna for 20m

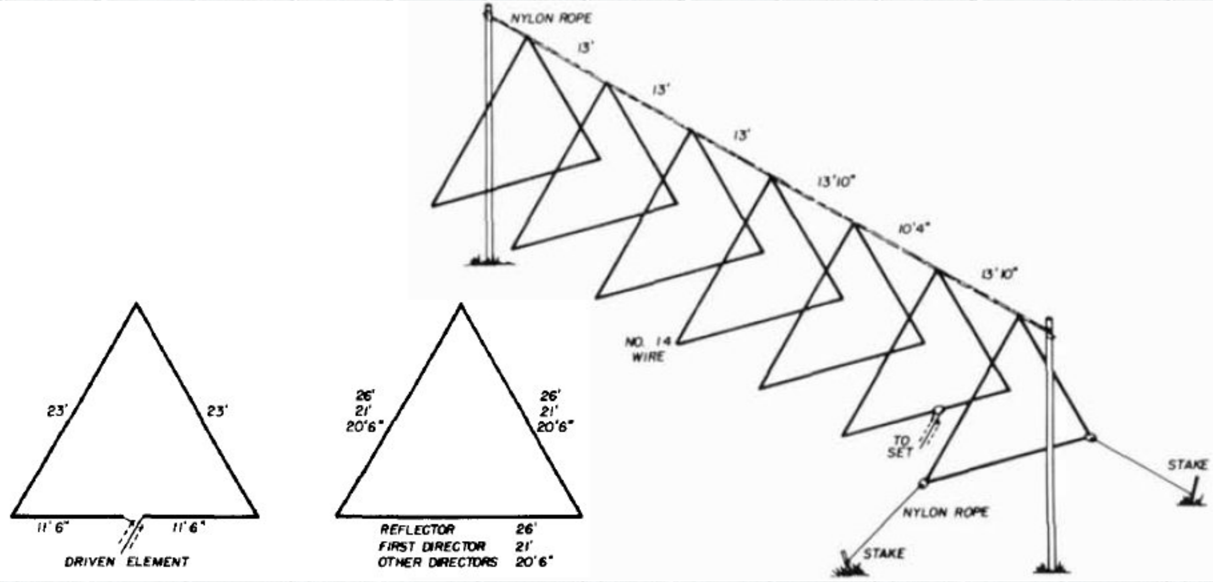
NYLON ROPE
13'
13'
13'
13'10"
10'4"
13'10"
NO. 14 WIRE
TO SET
STAKE
NYLON ROPE
STAKE
23'
23'
11'6"
11'6"
DRIVEN ELEMENT
26'
21'
20'6"
26'
21'
20'6"
REFLECTOR 26'
FIRST DIRECTOR 21'
OTHER DIRECTORS 20'6"

220 - Log Periodic Wire Antenna for 40m

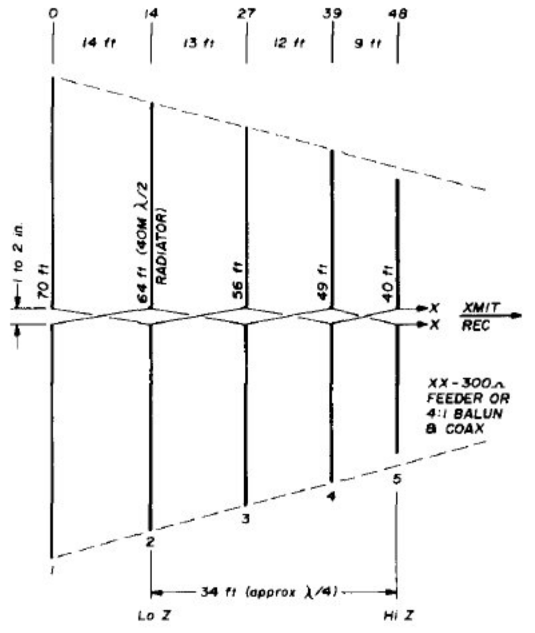
0
14
27
39
48
14 ft
13 ft
12 ft
9 ft
1 to 2 in.
70 ft
64 ft (40M λ/2 RADIATOR)
56 ft
49 ft
40 ft
X XMIT
X REC
XX - 300Ω FEEDER OR 4:1 BALUN & COAX
1
2
3
4
5
34 ft (approx λ/4)
Lo Z
Hi Z

221 - Three-Elements 90-degrees Wire Beam for 20m

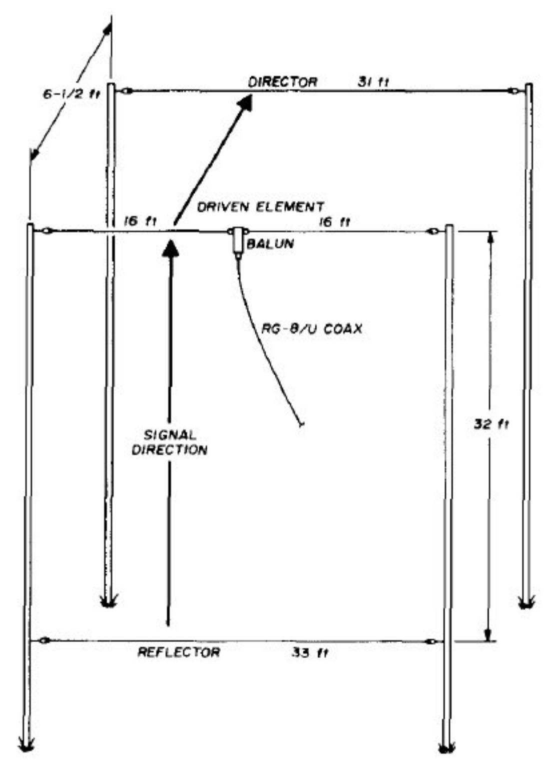
6-1/2 ft
DIRECTOR
31 ft
DRIVEN ELEMENT
16 ft
16 ft
BALUN
RG-8/U COAX
SIGNAL DIRECTION
32 ft
REFLECTOR
33 ft

---

222 - Two-Elements Wire Beam for 20m

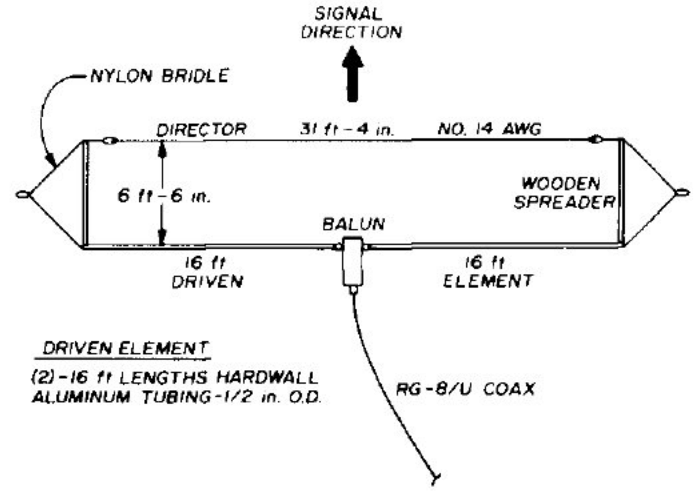
SIGNAL DIRECTION
NYLON BRIDLE
DIRECTOR
31 ft - 4 in.
NO. 14 AWG
6 ft - 6 in.
WOODEN SPREADER
BALUN
16 ft DRIVEN
16 ft ELEMENT
DRIVEN ELEMENT
(2)-16 ft LENGTHS HARDWALL ALUMINUM TUBING-1/2 in. O.D.
RG-8/U COAX

---

223 - Suitcase-Fit Dipole Antenna for 80m, 40m, 20m

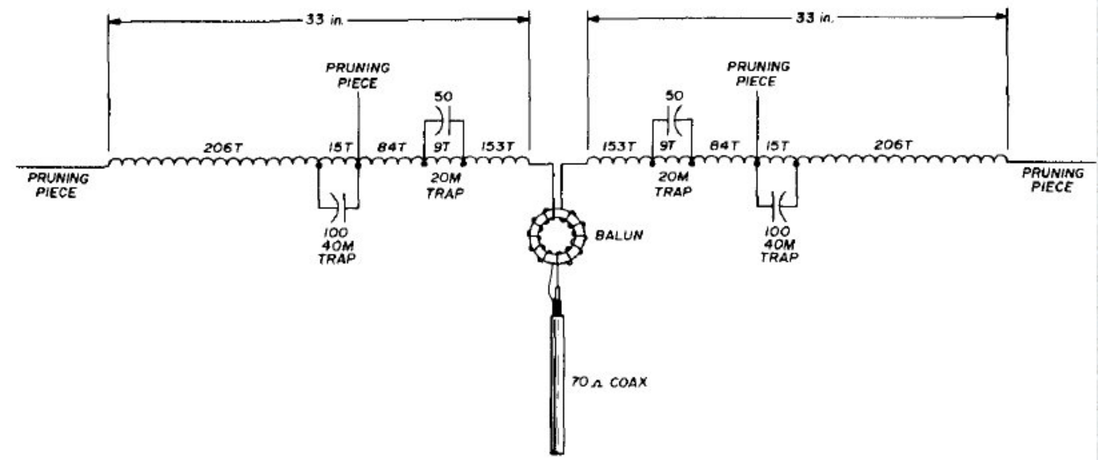
33 in.
33 in.
PRUNING PIECE
50
206T
15T
84T
9T
153T
20M TRAP
100 40M TRAP
BALUN
70Ω COAX
PRUNING PIECE

224 - Multiband Coaxial Dipole for 80m, 40m

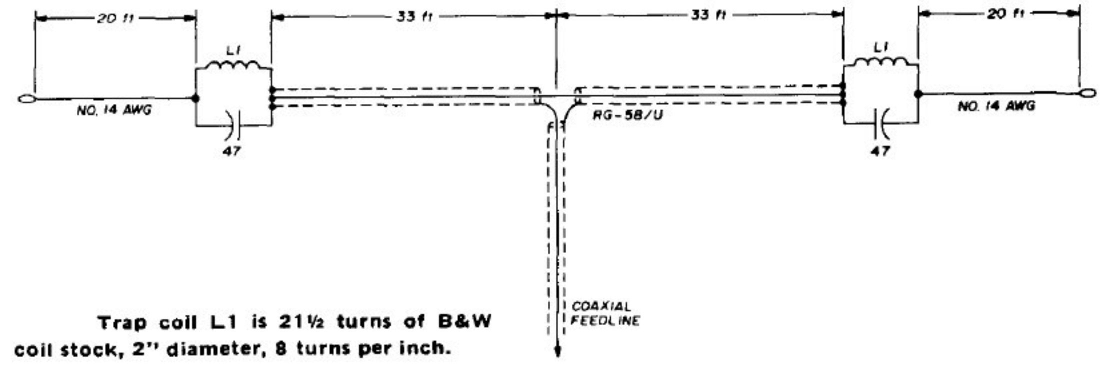
20 ft
33 ft
33 ft
20 ft
L1
NO. 14 AWG
47
RG-58/U
COAXIAL FEEDLINE

Trap coil L1 is 21½ turns of B&W coil stock, 2" diameter, 8 turns per inch.

225 - Vertical Log Periodic Antenna for 80m, 40m

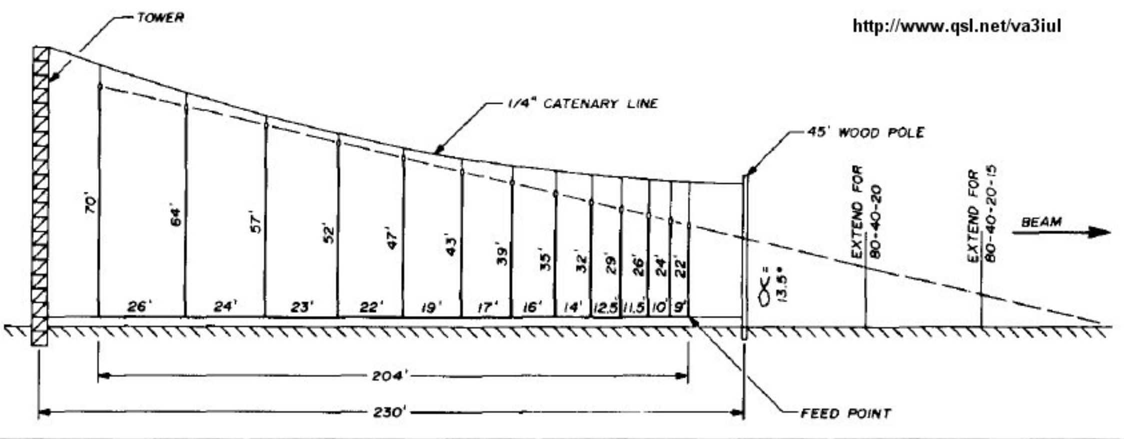
TOWER
http://www.qsl.net/va3iul
1/4" CATENARY LINE
45' WOOD POLE
70'
64'
57'
52'
47'
43'
39'
35'
32'
29'
26'
24'
22'
26'
24'
23'
22'
19'
17'
16'
14'
12.5
11.5
10
9'
α = 13.5°
EXTEND FOR 80-40-20
EXTEND FOR 80-40-20-15
BEAM
204'
230'
FEED POINT

226 - Inverted Vertical Log Periodic Antenna for 40m

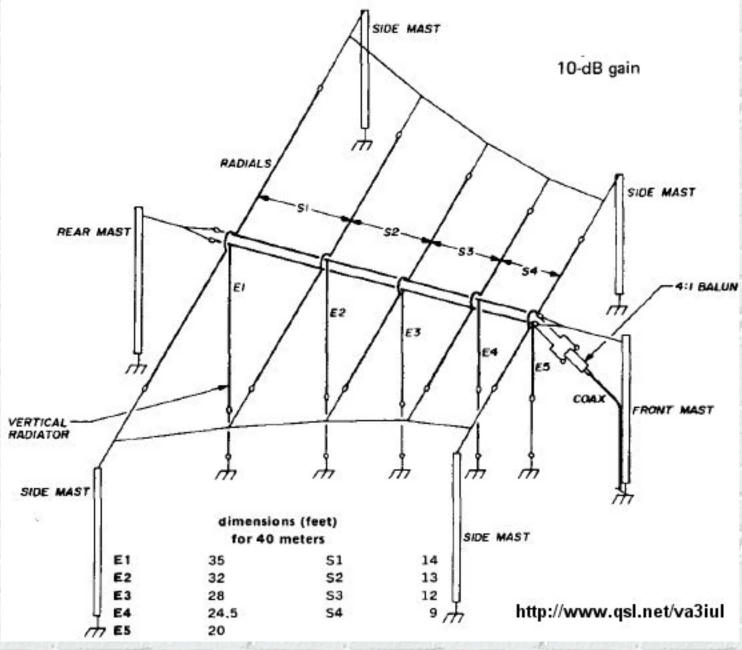
SIDE MAST
10-dB gain
RADIALS
SIDE MAST
REAR MAST
S1
S2
S3
S4
4:1 BALUN
E1
E2
E3
E4
E5
COAX
FRONT MAST
VERTICAL RADIATOR
SIDE MAST
SIDE MAST
dimensions (feet)
for 40 meters
E1 35 S1 14
E2 32 S2 13
E3 28 S3 12
E4 24.5 S4 9
E5 20
http://www.qsl.net/va3iul

227 - Double Extended Zepp Antenna Gain=7dB for 15m

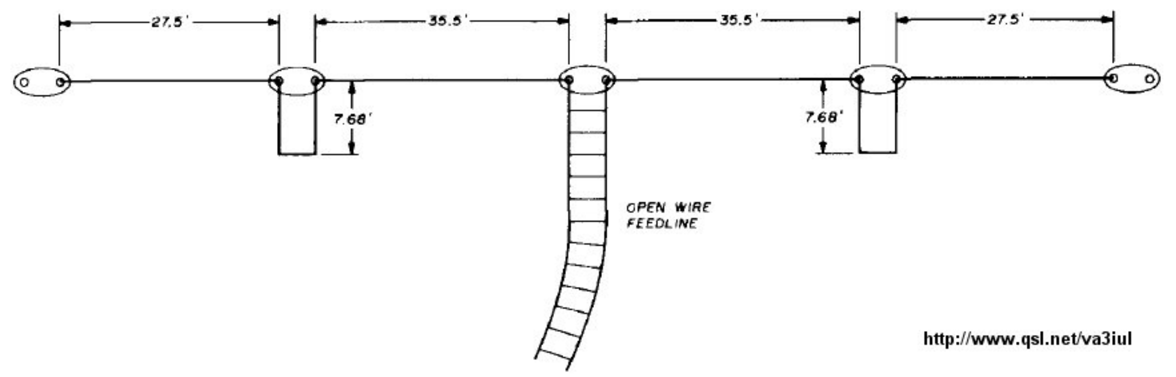
27.5'
35.5'
35.5'
27.5'
7.68'
7.68'
OPEN WIRE FEEDLINE
http://www.qsl.net/va3iul

228 - Double Lazy-H Antenna G=10dB for 70cm

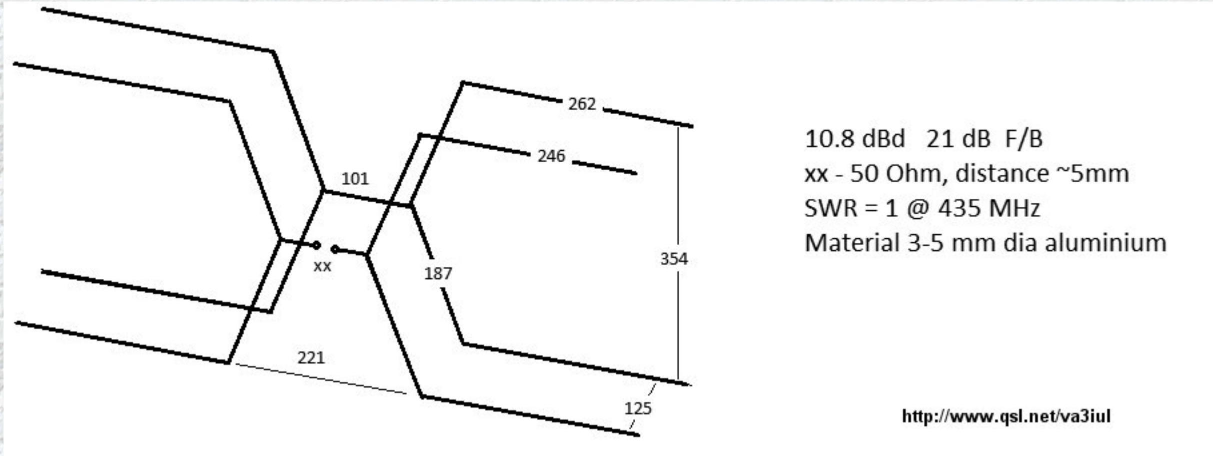
262
246
101
xx
187
354
221
125
10.8 dBd 21 dB F/B
xx - 50 Ohm, distance ~5mm
SWR = 1 @ 435 MHz
Material 3-5 mm dia aluminium
http://www.qsl.net/va3iul

229 - Wire Log Periodic Antenna G=12dB for 15m, 20m

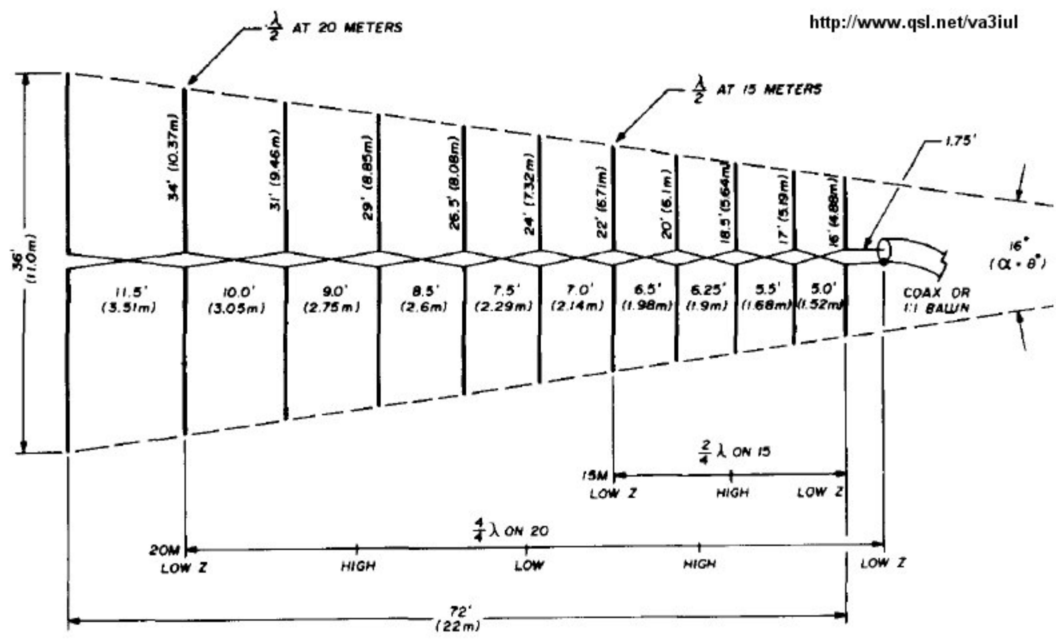
http://www.qsl.net/va3iul
λ/2 AT 20 METERS
λ/2 AT 15 METERS
1.75'
36' (11.0m)
34' (10.37m)
31' (9.46m)
29' (8.85m)
26.5' (8.08m)
24' (7.32m)
22' (6.71m)
20' (6.1m)
18.5' (5.64m)
17' (5.19m)
16' (4.88m)
11.5' (3.51m)
10.0' (3.05m)
9.0' (2.75m)
8.5' (2.6m)
7.5' (2.29m)
7.0' (2.14m)
6.5' (1.98m)
6.25' (1.9m)
5.5' (1.68m)
5.0' (1.52m)
COAX OR 1:1 BALUN
16° (α = 8°)
2/4 λ ON 15
15M LOW Z
HIGH
LOW Z
4/4 λ ON 20
20M LOW Z
HIGH
LOW
HIGH
LOW Z
72' (22m)

---

230 - Five-Elements Vertical Log Periodic for 80m, 40m, 20m

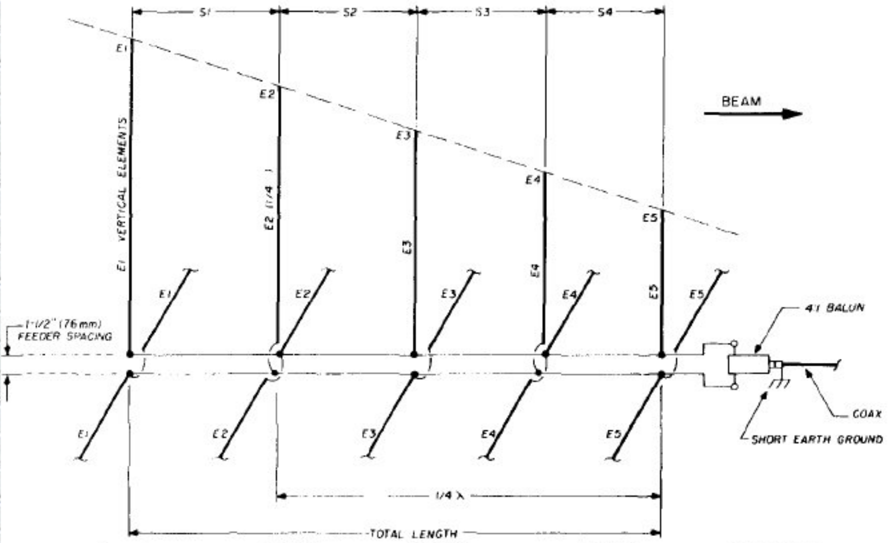
S1
S2
S3
S4
E1
E2
E3
E4
E5
BEAM
E1 VERTICAL ELEMENTS
E2 (1/4 λ)
E3
E4
E5
1-1/2" (76mm) FEEDER SPACING
4:1 BALUN
COAX
SHORT EARTH GROUND
1/4 λ
TOTAL LENGTH

element	3.5-4.0^1	3.8-4.0^2	7.0-7.3^2	14.0-14.35^1
E1	70' (21.4 m)	65' (19.8 m)	35' (10.7 m)	17.5' (5.3 m)
E2	67' (20.4 m)	62' (18.9 m)	33' (10.0 m)	16.5' (5.0 m)
E3	58' (17.7 m)	55' (16.8 m)	28' (8.5 m)	14.0' (4.3 m)
E4	50' (15.3 m)	45' (13.7 m)	24.5' (7.5 m)	12.2' (3.7 m)
E5	43' (13.1 m)	40' (12.2 m)	20' (6.1 m)	10.0' (3.0 m)
S1	30' (9.2 m)	26' (7.9 m)	14' (4.3 m)	7.0' (2.1 m)
S2	27' (8.2 m)	24' (7.3 m)	13' (4.0 m)	6.5' (2.0 m)
S3	24' (7.3 m)	23' (7.0 m)	12' (3.7 m)	6.0' (1.8 m)
S4	19' (5.8 m)	18' (5.5 m)	9' (2.7 m)	4.5' (1.4 m)
total length	100' (30.5 m)	91' (27.8 m)	48' (14.6 m)	24' (7.3 m)
mast height	80' (24.4 m)	75' (22.9 m)	50' (15.3 m)	30' (9.2 m)
pole height	45' (13.7 m)	40' (12.2 m)	25' (7.6 m)	20' (6.1 m)

http://www.qsl.net/va3iul

---

231 - Sloping Delta Loop Antenna for 40m

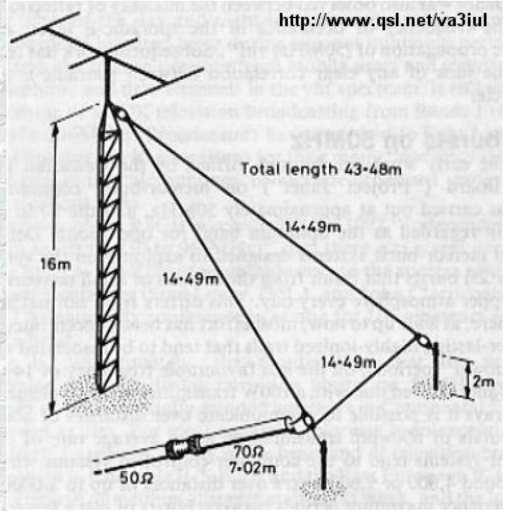
http://www.qsl.net/va3iul
Total length 43·48m
14·49m
16m
14·49m
14·49m
2m
70Ω
7·02m
50Ω

232 - Doublet Antenna for 160m

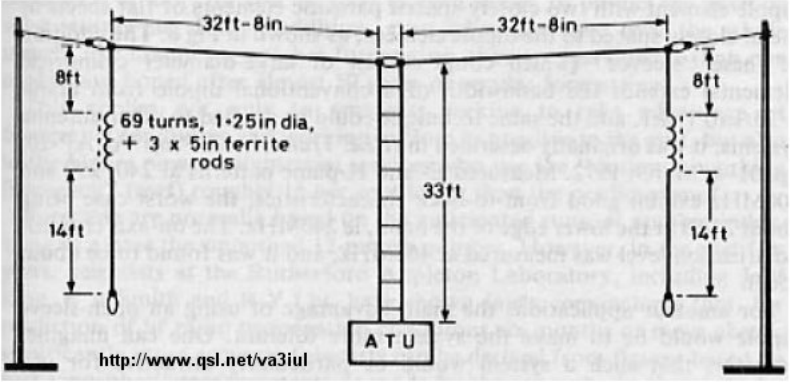
32ft-8in
32ft-8in
8ft
8ft
69 turns, 1·25in dia, + 3 x 5in ferrite rods
33ft
14ft
14ft
A T U
http://www.qsl.net/va3iul

233 - Log Periodic Wire Beam for 40m

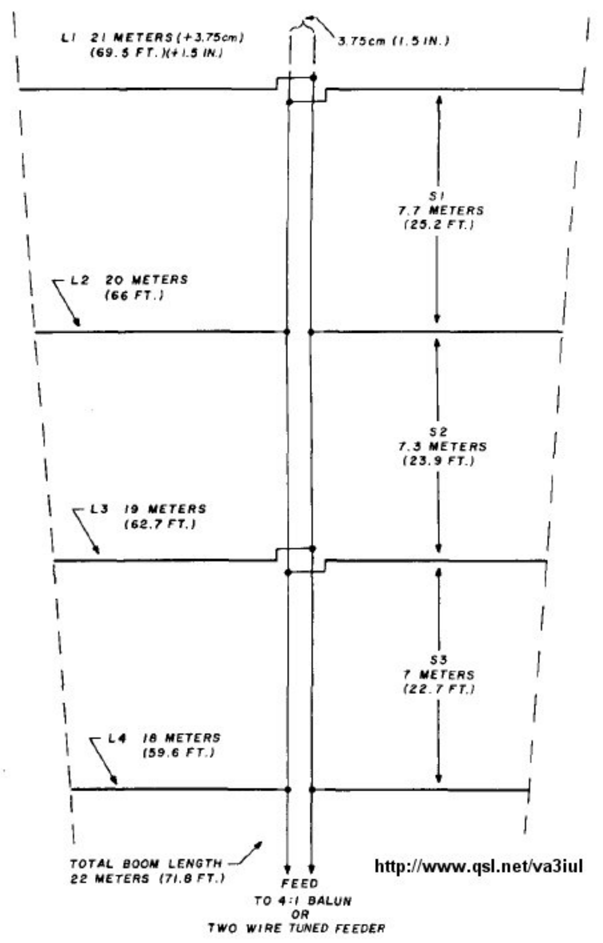
L1 21 METERS (+3.75cm)
(69.5 FT.)(+1.5 IN.)
3.75cm (1.5 IN.)
S1
7.7 METERS
(25.2 FT.)
L2 20 METERS
(66 FT.)
S2
7.3 METERS
(23.9 FT.)
L3 19 METERS
(62.7 FT.)
S3
7 METERS
(22.7 FT.)
L4 18 METERS
(59.6 FT.)
TOTAL BOOM LENGTH
22 METERS (71.8 FT.)
http://www.qsl.net/va3iul
FEED
TO 4:1 BALUN
OR
TWO WIRE TUNED FEEDER

234 - G3LDO Wire Beam Antenna for 20m, 15m, 10m

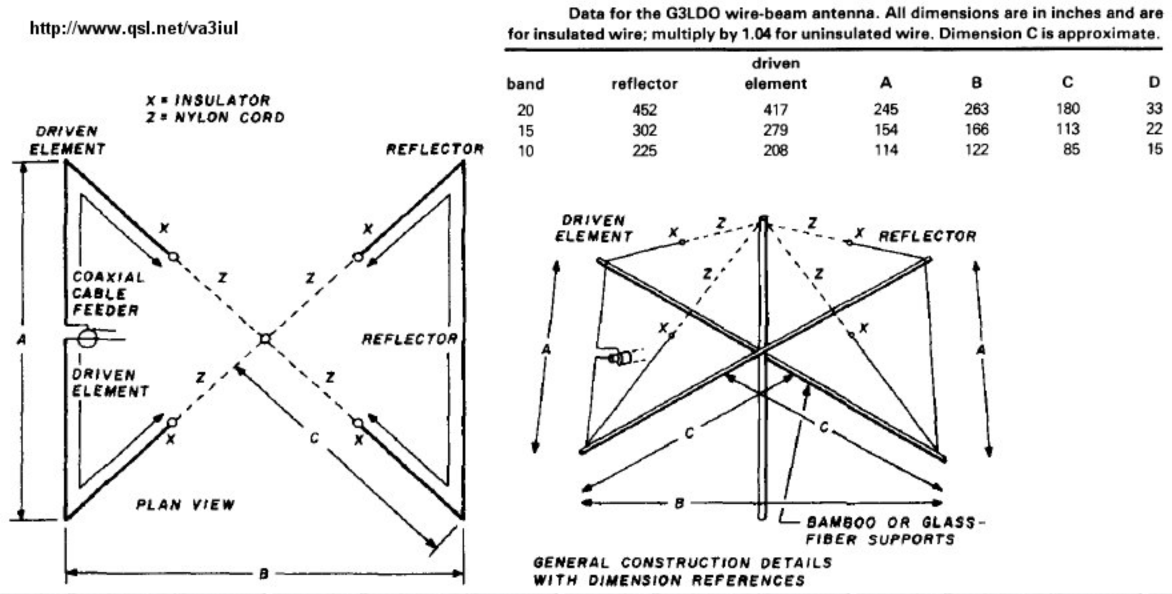
Data for the G3LDO wire-beam antenna. All dimensions are in inches and are for insulated wire; multiply by 1.04 for uninsulated wire. Dimension C is approximate.

band	reflector	driven element	A	B	C	D
20	452	417	245	263	180	33
15	302	279	154	166	113	22
10	225	208	114	122	85	15

http://www.qsl.net/va3iul
X = INSULATOR
Z = NYLON CORD
DRIVEN ELEMENT
REFLECTOR
COAXIAL CABLE FEEDER
DRIVEN ELEMENT
REFLECTOR
X
Z
A
B
C
PLAN VIEW
DRIVEN ELEMENT
REFLECTOR
BAMBOO OR GLASS-FIBER SUPPORTS
GENERAL CONSTRUCTION DETAILS WITH DIMENSION REFERENCES

235 - Phased Delta Loop Array for 20m

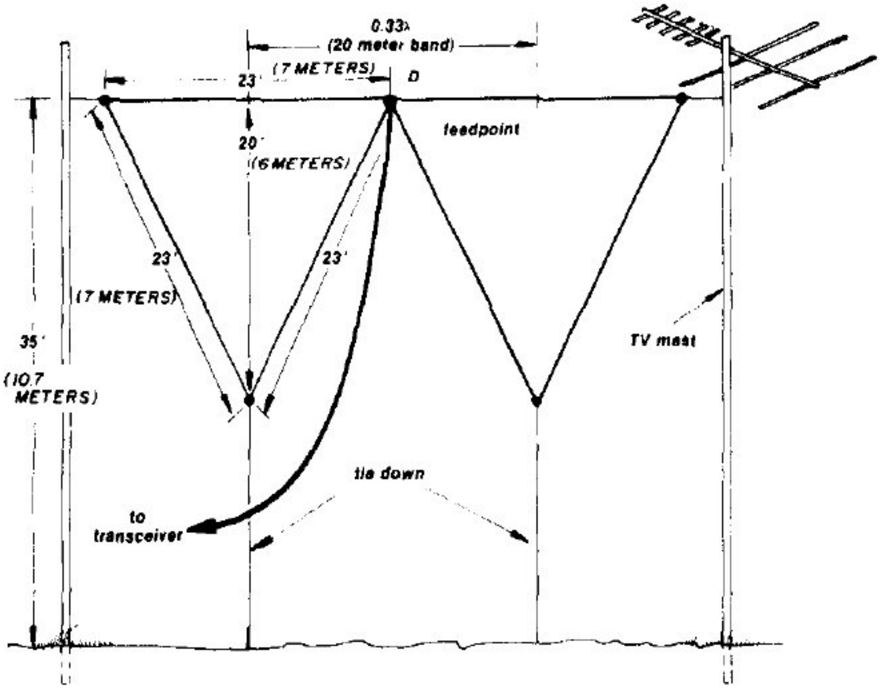
0.33λ
(20 meter band)
23'
(7 METERS)
D
20'
(6 METERS)
feedpoint
23'
23'
(7 METERS)
35'
(10.7 METERS)
TV mast
tie down
to transceiver

spacing wavelength	feet	(meters)	gain (dB)
0.33	23	(7)	2.0
0.40	27	(8)	2.8
0.50	34	(10)	4.0
0.65	44	(13)	4.8

http://www.qsl.net/va3iul

236 - Dual Band V-dipole for 80m, 40m

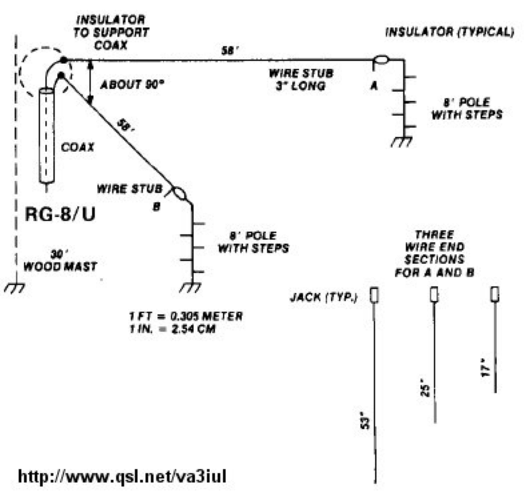
INSULATOR TO SUPPORT COAX
58'
INSULATOR (TYPICAL)
ABOUT 90°
WIRE STUB 3" LONG
A
8' POLE WITH STEPS
58'
COAX
WIRE STUB
B
RG-8/U
8' POLE WITH STEPS
30' WOOD MAST
THREE WIRE END SECTIONS FOR A AND B
JACK (TYP.)
53"
25"
17"
1 FT = 0.305 METER
1 IN. = 2.54 CM

http://www.qsl.net/va3iul

237 - Collinear Array 5dB-Gain for 15m

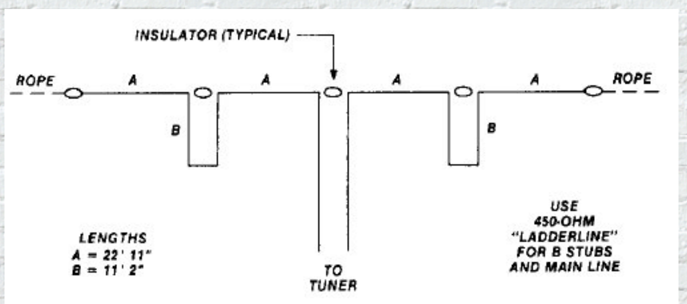
INSULATOR (TYPICAL)
ROPE
A
A
A
A
ROPE
B
B
LENGTHS
A = 22' 11"
B = 11' 2"
TO TUNER
USE 450-OHM "LADDERLINE" FOR B STUBS AND MAIN LINE

http://www.qsl.net/va3iul

1 FT = 0.305 METER
1 IN. = 2.54 CM

238 - Two-Band Quad Loop Antenna 80m, 40m

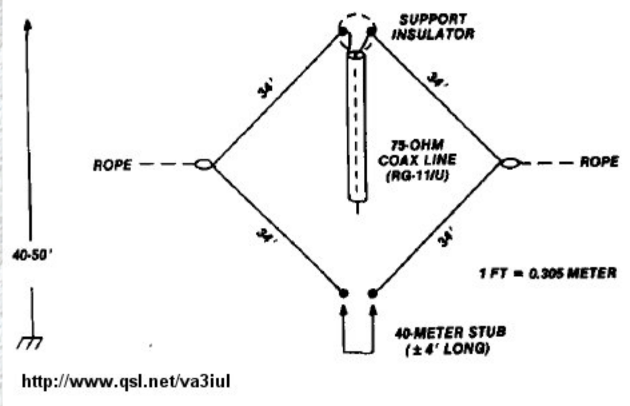
SUPPORT INSULATOR
34'
34'
34'
34'
ROPE
ROPE
75-OHM COAX LINE (RG-11/U)
40-50'
1 FT = 0.305 METER
40-METER STUB (± 4' LONG)
http://www.qsl.net/va3iul

239 - K7CW Quad Antenna

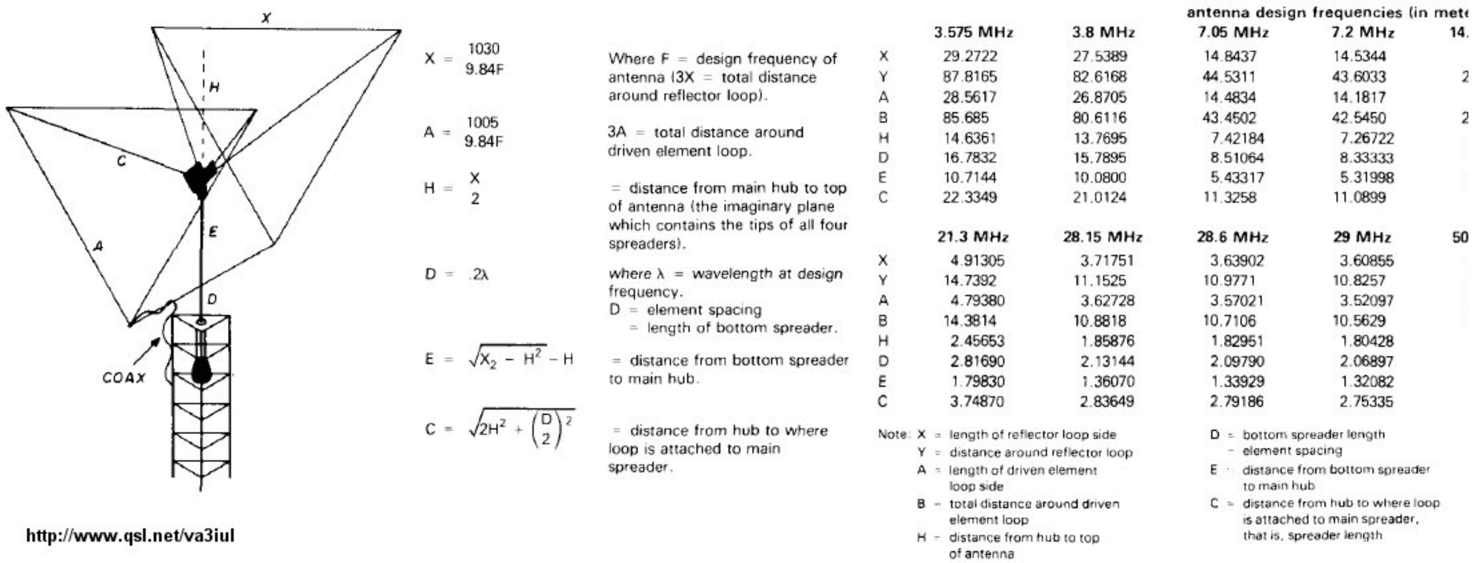
X
H
C
E
A
D
COAX
http://www.qsl.net/va3iul

$X = \frac{1030}{9.84F}$ Where F = design frequency of antenna (3X = total distance around reflector loop).

$A = \frac{1005}{9.84F}$ 3A = total distance around driven element loop.

$H = \frac{X}{2}$ = distance from main hub to top of antenna (the imaginary plane which contains the tips of all four spreaders).

$D = .2\lambda$ where λ = wavelength at design frequency. D = element spacing = length of bottom spreader.

$E = \sqrt{X_2 - H^2} - H$ = distance from bottom spreader to main hub.

$C = \sqrt{2H^2 + \left(\frac{D}{2}\right)^2}$ = distance from hub to where loop is attached to main spreader.

antenna design frequencies (in mete

	3.575 MHz	3.8 MHz	7.05 MHz	7.2 MHz	14.
X	29.2722	27.5389	14.8437	14.5344	
Y	87.8165	82.6168	44.5311	43.6033	2
A	28.5617	26.8705	14.4834	14.1817	
B	85.685	80.6116	43.4502	42.5450	2
H	14.6361	13.7695	7.42184	7.26722	
D	16.7832	15.7895	8.51064	8.33333	
E	10.7144	10.0800	5.43317	5.31998	
C	22.3349	21.0124	11.3258	11.0899	

	21.3 MHz	28.15 MHz	28.6 MHz	29 MHz	50
X	4.91305	3.71751	3.63902	3.60855	
Y	14.7392	11.1525	10.9771	10.8257	
A	4.79380	3.62728	3.57021	3.52097	
B	14.3814	10.8818	10.7106	10.5629	
H	2.45653	1.85876	1.82951	1.80428	
D	2.81690	2.13144	2.09790	2.06897	
E	1.79830	1.36070	1.33929	1.32082	
C	3.74870	2.83649	2.79186	2.75335	

Note: X = length of reflector loop side
Y = distance around reflector loop
A = length of driven element loop side
B = total distance around driven element loop
H = distance from hub to top of antenna
D = bottom spreader length = element spacing
E = distance from bottom spreader to main hub
C = distance from hub to where loop is attached to main spreader, that is, spreader length

240 - Broadband Short Dipole for 80m

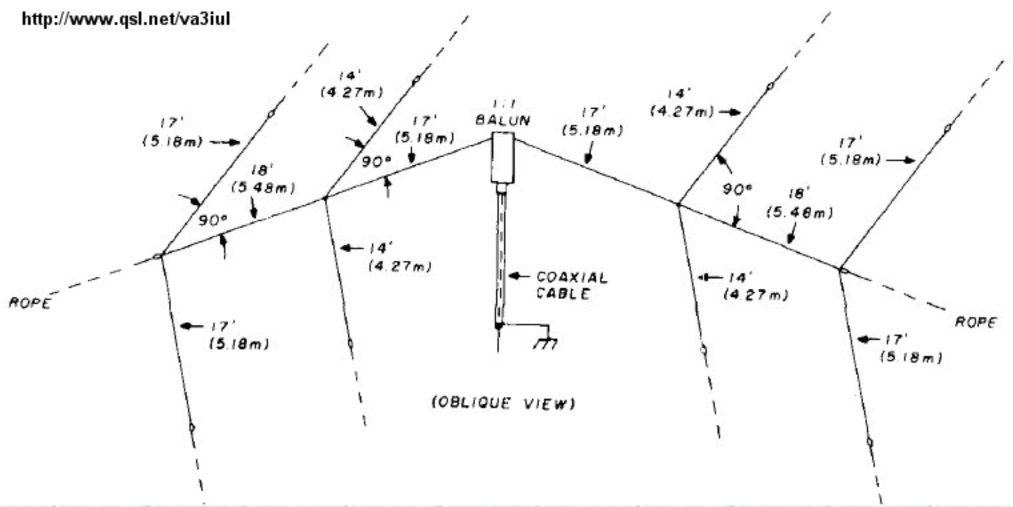
http://www.qsl.net/va3iul
14' (4.27m)
17' (5.18m)
17' (5.18m)
18' (5.48m)
90°
90°
1:1 BALUN
17' (5.18m)
14' (4.27m)
17' (5.18m)
90°
18' (5.48m)
ROPE
14' (4.27m)
COAXIAL CABLE
14' (4.27m)
ROPE
17' (5.18m)
17' (5.18m)
(OBLIQUE VIEW)

241 - Off-Center Fed Dipole for 40m, 20m, 10m

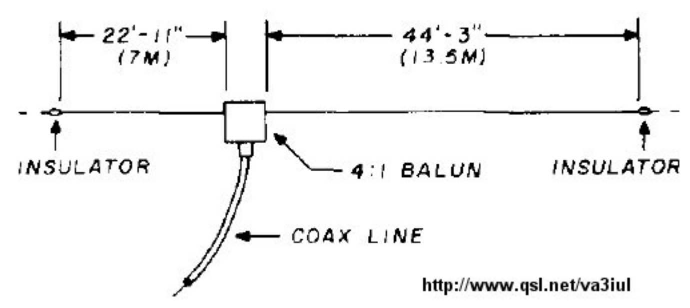
22'-11"
(7M)
44'-3"
(13.5M)
INSULATOR
4:1 BALUN
INSULATOR
COAX LINE
http://www.qsl.net/va3iul

242 - Broadband Sloper Antenna for 80m

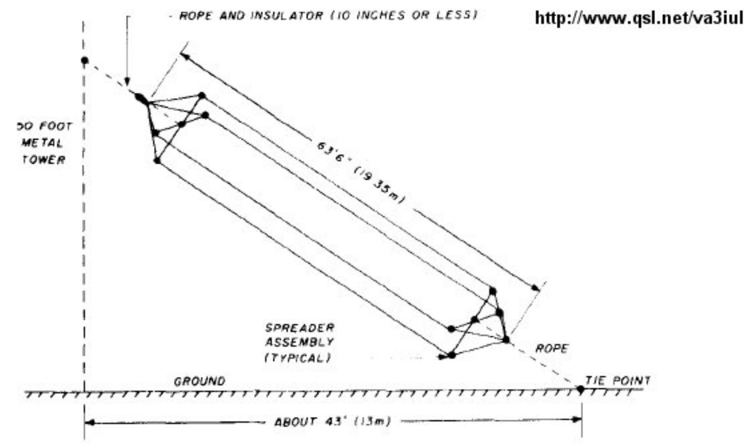
ROPE AND INSULATOR (10 INCHES OR LESS)
http://www.qsl.net/va3iul
50 FOOT
METAL
TOWER
63'6" (19.35m)
SPREADER
ASSEMBLY
(TYPICAL)
ROPE
GROUND
TIE POINT
ABOUT 43' (13m)

80-meter WD4FAB broadband sloper. Spreader assemblies are made of 36-inch long fiberglass tubes. The four antenna wires are tied together about 24 inches past each spreader assembly. A nylon line may be required from the crossover point of the spreader to the outer insulator to prevent bowing. The SWR plot of the antenna falls below 1.4-to-1 at 3.5 MHz; 1.25-to-1 at 3.8 MHz and 1.4-to-1 at 4.0 MHz. Feedpoint resistance of the antenna may be changed by altering the slope angle. The resonant frequency is adjusted by changing length — no climbing necessary!

243 - Two-Frequency Dipole Antenna for 80m

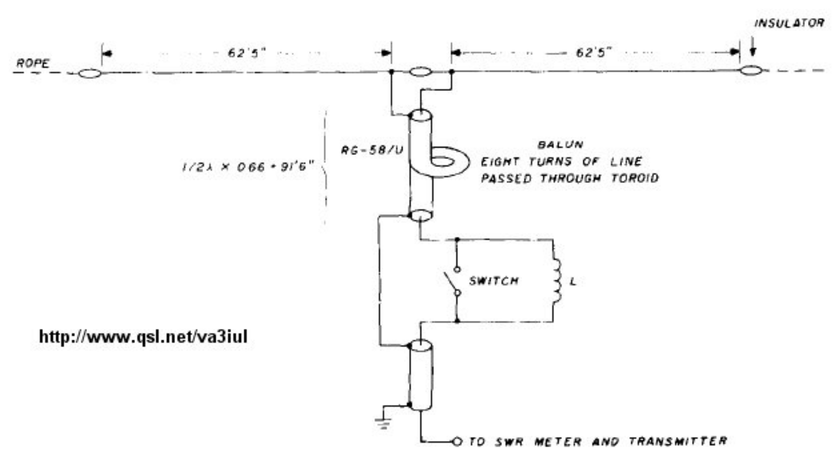
INSULATOR
ROPE
62'5"
62'5"
1/2λ × 0.66 = 91'6"
RG-58/U
BALUN
EIGHT TURNS OF LINE
PASSED THROUGH TOROID
SWITCH
L
http://www.qsl.net/va3iul
TO SWR METER AND TRANSMITTER

The W6TC two-frequency dipole for 80 meters. Antenna is cut for resonance at high frequency end of band. A single loading coil for operation at the low frequency end of the band is placed an electrical half-wavelength down the transmission line. The line is decoupled from the antenna by a simple balun placed directly at the antenna insulator. Toroid is powdered iron material T-200 (2″ OD) (1.25″ ID), 2-mix (red, μ = 10 (Amidon T-200-2), or equivalent.

244 - Top Loaded Vertical Antenna for 160m

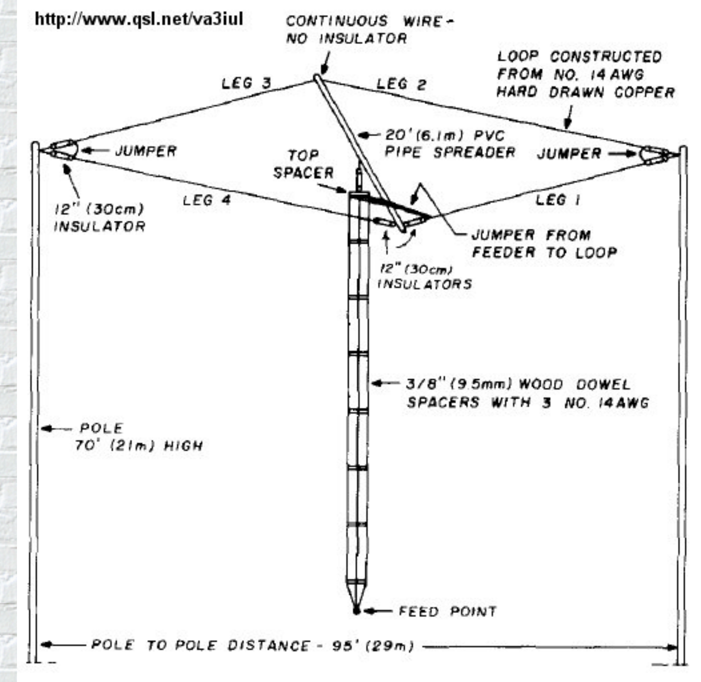
http://www.qsl.net/va3iul
CONTINUOUS WIRE - NO INSULATOR
LEG 3
LEG 2
LOOP CONSTRUCTED FROM NO. 14 AWG HARD DRAWN COPPER
20' (6.1m) PVC PIPE SPREADER
JUMPER
JUMPER
TOP SPACER
12" (30cm) INSULATOR
LEG 4
LEG 1
JUMPER FROM FEEDER TO LOOP
12" (30cm) INSULATORS
3/8" (9.5mm) WOOD DOWEL SPACERS WITH 3 NO. 14AWG
POLE 70' (21m) HIGH
FEED POINT
POLE TO POLE DISTANCE - 95' (29m)

245 - Two Loop Beam Antenna for 20m, 15m

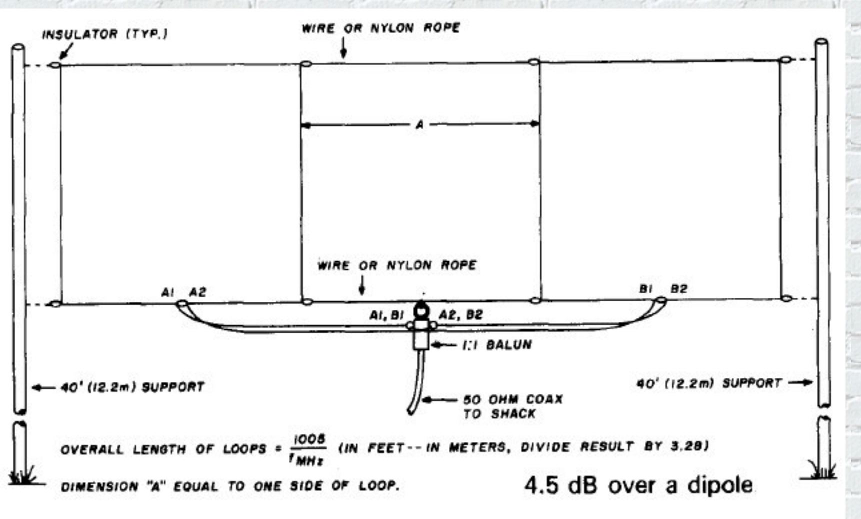
INSULATOR (TYP.)
WIRE OR NYLON ROPE
A
WIRE OR NYLON ROPE
A1 A2
B1 B2
A1, B1
A2, B2
1:1 BALUN
40' (12.2m) SUPPORT
40' (12.2m) SUPPORT
50 OHM COAX TO SHACK
OVERALL LENGTH OF LOOPS = 1005/fMHz (IN FEET--IN METERS, DIVIDE RESULT BY 3.28)
DIMENSION "A" EQUAL TO ONE SIDE OF LOOP.
4.5 dB over a dipole

The W2TBZ loop beam. The phasing lines are made of 300-ohm TV line. Length of phasing line = $\frac{468}{f_{MHz}}$ × velocity factor of 300-ohm line. Balun is supported at center of lower nylon rope.

http://www.qsl.net/va3iul

246 - Corner-Fed Delta Loop Antenna for 80m, 40m, 20m

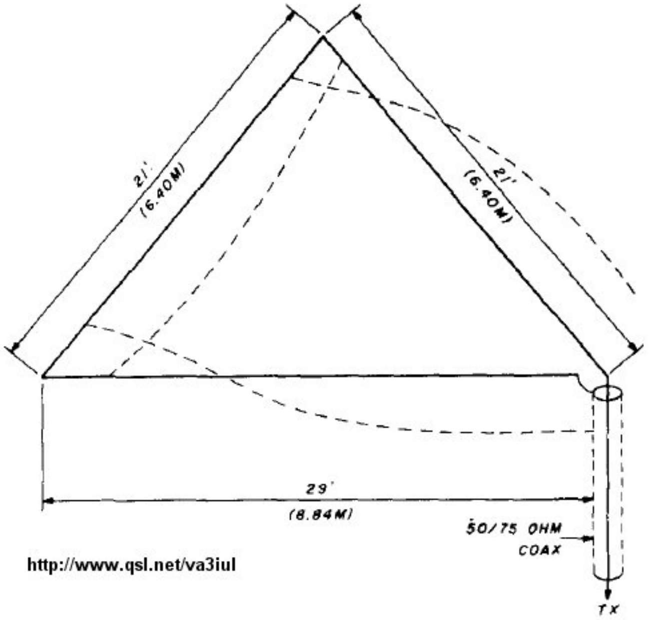
21'
(6.40M)
21'
(6.40M)
29'
(8.84M)
50/75 OHM
COAX
http://www.qsl.net/va3iul
TX

DIMENSIONS FOR FULL-SIZE CORNER-FED DELTA LOOPS

frequency (MHz)	total loop circumference (feet)	(meters)	length of each side (feet)	(meters)	length of base (feet)	(meters)
3.600	279 ft. 2 in.	(85.09)	82 ft.	(24.99)	115 ft. 2 in.	(35.10)
7.075	142 ft. 0 in.	(43.28)	42 ft.	(12.80)	58 ft. 0 in.	(17.68)
14.150	71 ft. 0 in.	(21.64)	21 ft.	(6.40)	29 ft. 0 in.	(8.84)

---

247 - Loaded Delta Loop Antenna for 40m, 20m

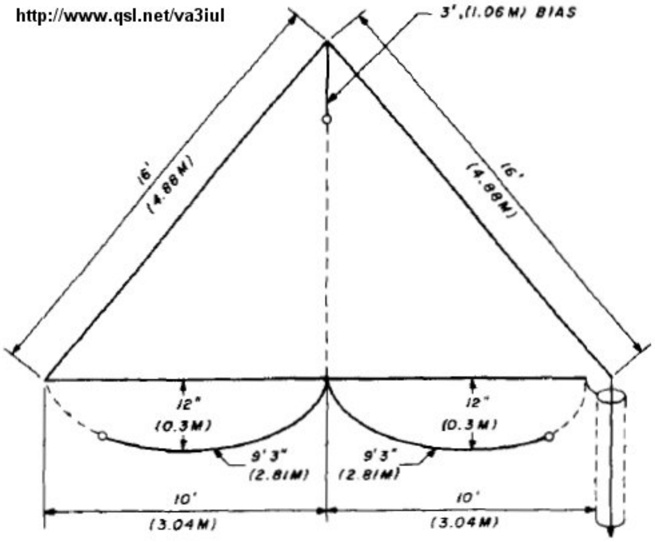
http://www.qsl.net/va3iul
3',(1.06M) BIAS
16'
(4.88M)
16'
(4.88M)
12"
(0.3M)
12"
(0.3M)
9'3"
(2.81M)
9'3"
(2.81M)
10'
(3.04M)
10'
(3.04M)

The final loaded loop for 14 MHz. All dimensions are doubled for operation on 7 MHz. Base loading wires are each 18 feet 4 inches (5.59 meters). The top loading wire is made taut by securing to base with thin light nylon cord.

---

248 - Compact Size Dipole for 80m

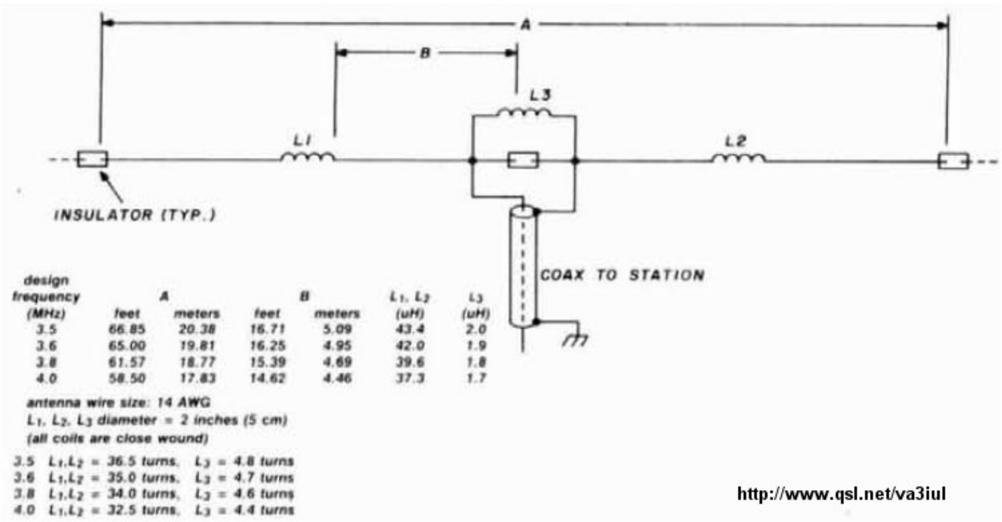
A
B
L3
L1
L2
INSULATOR (TYP.)
COAX TO STATION
http://www.qsl.net/va3iul

design frequency (MHz)	A feet	A meters	B feet	B meters	L_1, L_2 (uH)	L_3 (uH)
3.5	66.85	20.38	16.71	5.09	43.4	2.0
3.6	65.00	19.81	16.25	4.95	42.0	1.9
3.8	61.57	18.77	15.39	4.69	39.6	1.8
4.0	58.50	17.83	14.62	4.46	37.3	1.7

antenna wire size: 14 AWG
L_1, L_2, L_3 diameter = 2 inches (5 cm)
(all coils are close wound)

3.5 L_1,L_2 = 36.5 turns, L_3 = 4.8 turns
3.6 L_1,L_2 = 35.0 turns, L_3 = 4.7 turns
3.8 L_1,L_2 = 34.0 turns, L_3 = 4.6 turns
4.0 L_1,L_2 = 32.5 turns, L_3 = 4.4 turns

---

249 - JAWS 3dB Gain Antenna for 40m

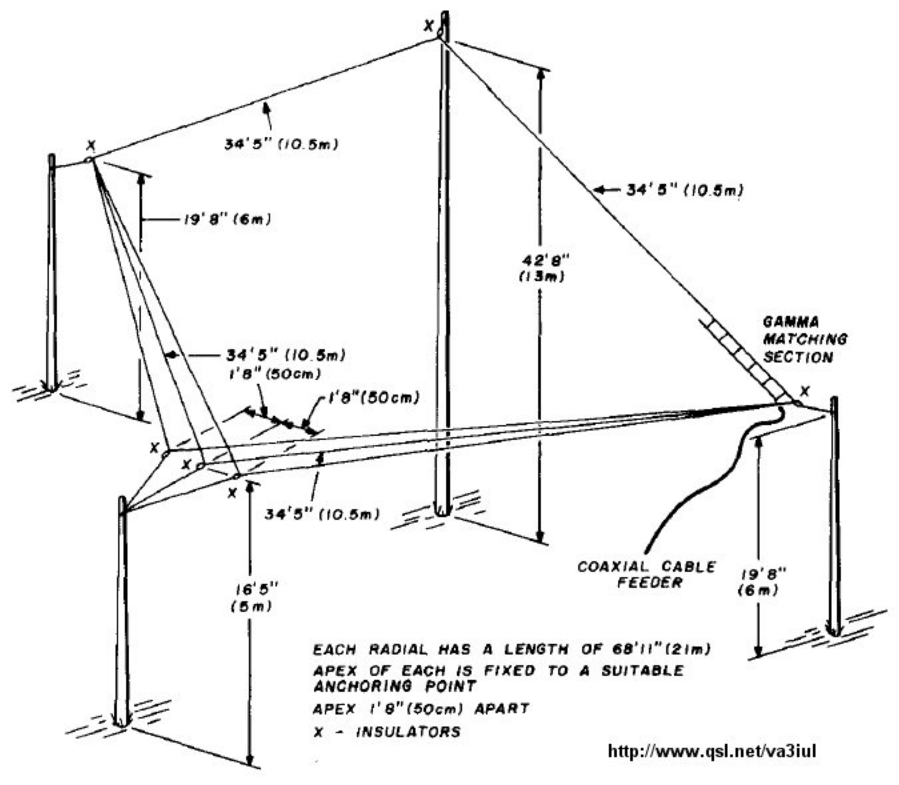
X
34'5" (10.5m)
34'5" (10.5m)
19'8" (6m)
42'8" (13m)
GAMMA MATCHING SECTION
34'5" (10.5m)
1'8" (50cm)
1'8" (50cm)
X
X
X
X
34'5" (10.5m)
COAXIAL CABLE FEEDER
16'5" (5m)
19'8" (6m)
EACH RADIAL HAS A LENGTH OF 68'11" (21m)
APEX OF EACH IS FIXED TO A SUITABLE ANCHORING POINT
APEX 1'8" (50cm) APART
X - INSULATORS
http://www.qsl.net/va3iul

---

250 - Inverted-V Antenna for 160m

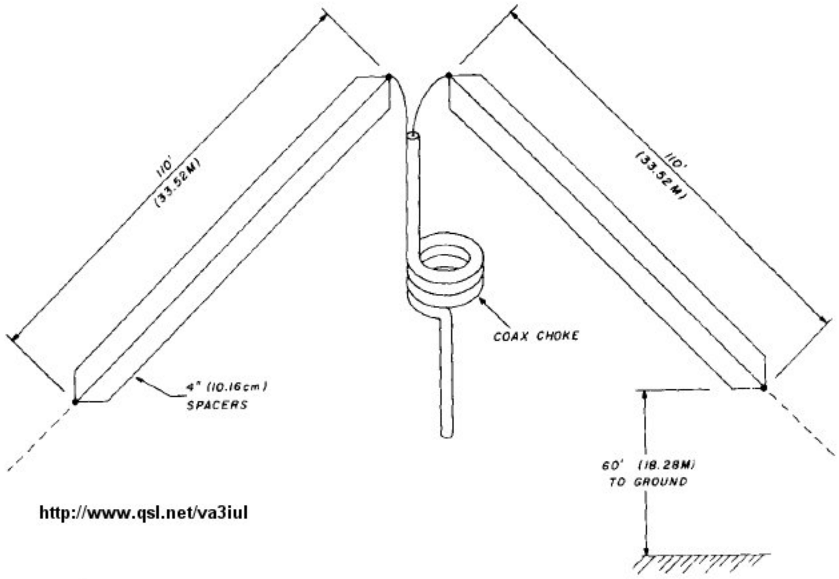
110'
(33.52M)
110'
(33.52M)
COAX CHOKE
4" (10.16cm)
SPACERS
60' (18.28M)
TO GROUND
http://www.qsl.net/va3iul

The inverted-V antenna for 160 meters at VE7BS. Wires all lie in the same plane. Apex is at 105 feet (32 meters).

---

251 - Lazy-U Antenna for 160m

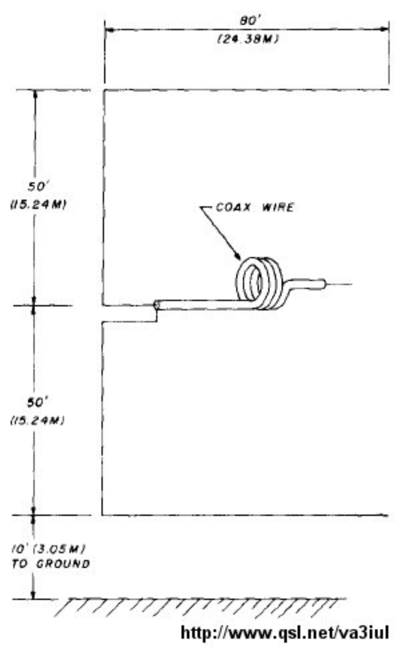
80'
(24.38M)
50'
(15.24M)
COAX WIRE
50'
(15.24M)
10' (3.05M)
TO GROUND
http://www.qsl.net/va3iul

---

252 - AF3V Compact Antenna for 160m

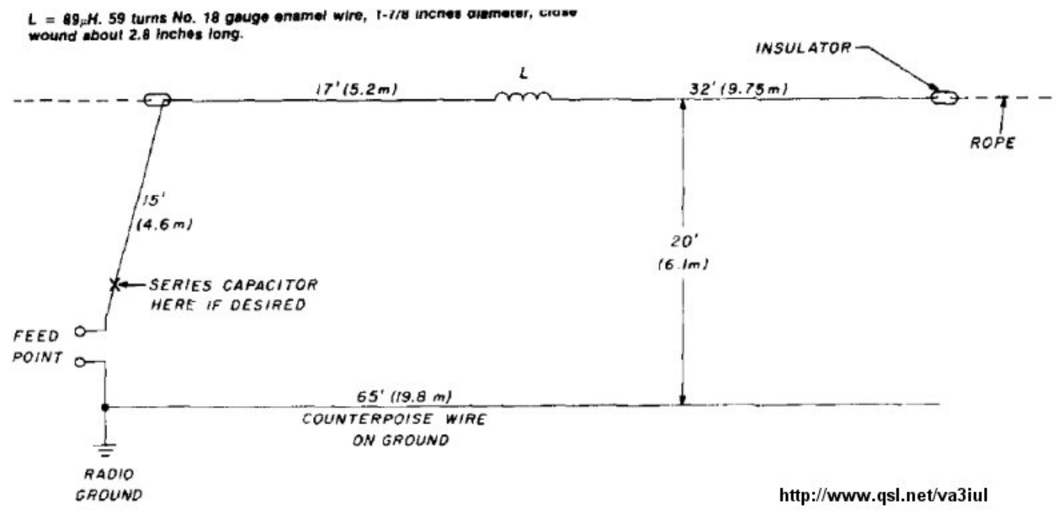
L = 89μH. 59 turns No. 18 gauge enamel wire, 1-7/8 inches diameter, close wound about 2.8 inches long.

INSULATOR
L
17' (5.2m)
32' (9.75m)
ROPE
15' (4.6m)
20' (6.1m)
SERIES CAPACITOR HERE IF DESIRED
FEED POINT
65' (19.8 m)
COUNTERPOISE WIRE ON GROUND
RADIO GROUND

http://www.qsl.net/va3iul

---

253 - Hot Beam Wire Antenna for 40m

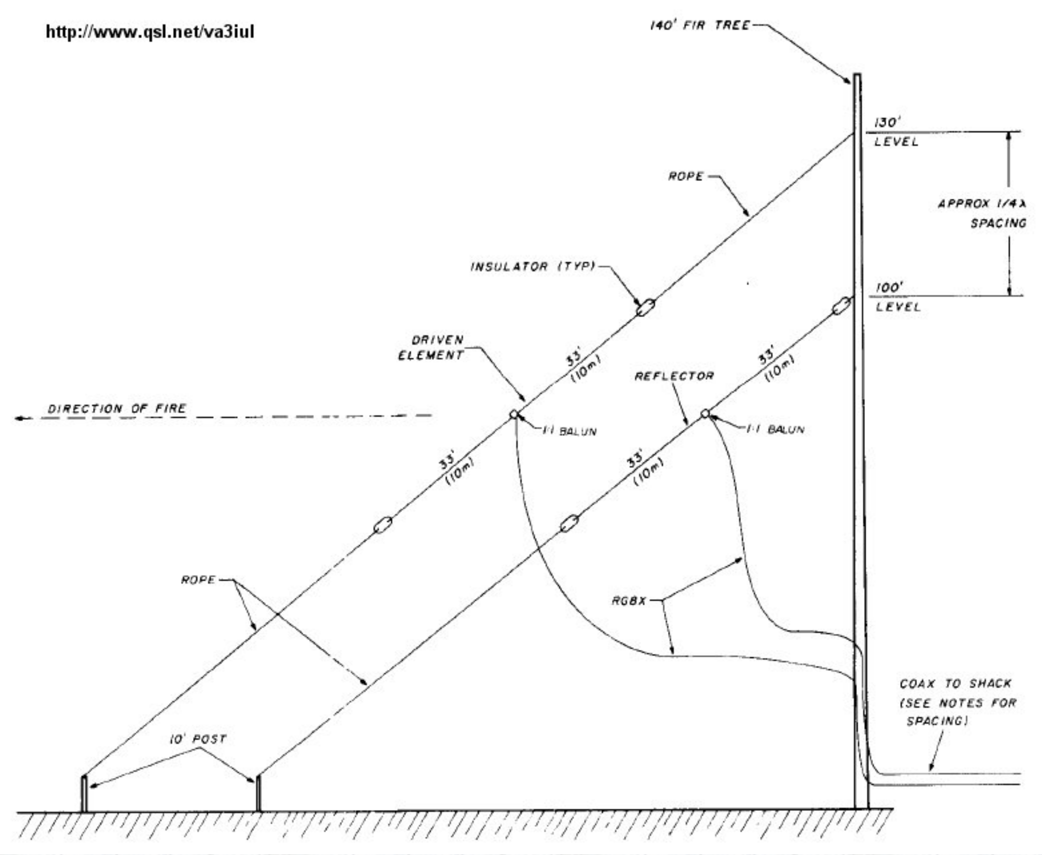
http://www.qsl.net/va3iul

140' FIR TREE
130' LEVEL
ROPE
APPROX 1/4λ SPACING
INSULATOR (TYP)
100' LEVEL
DRIVEN ELEMENT
33' (10m)
REFLECTOR
33' (10m)
DIRECTION OF FIRE
1:1 BALUN
1:1 BALUN
33' (10m)
33' (10m)
ROPE
RG8X
COAX TO SHACK (SEE NOTES FOR SPACING)
10' POST

---

254 - Sloping Delta Loop for 40m

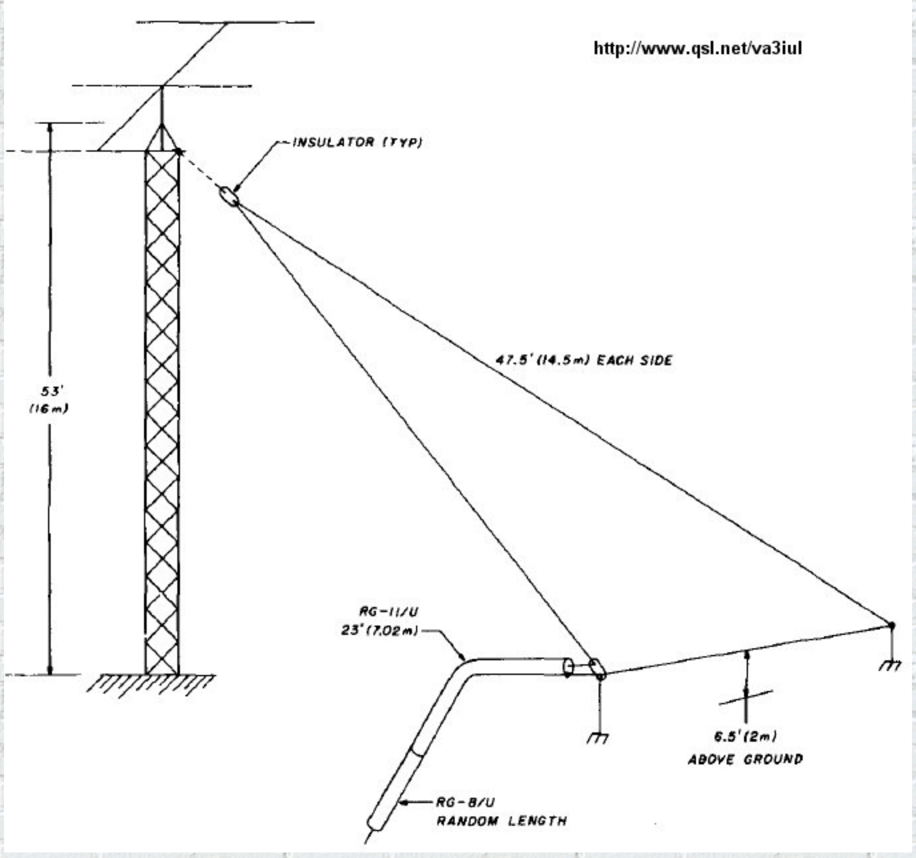
http://www.qsl.net/va3iul
INSULATOR (TYP)
47.5' (14.5m) EACH SIDE
53' (16m)
RG-11/U 23' (7.02m)
6.5' (2m) ABOVE GROUND
RG-8/U RANDOM LENGTH

255 - K6FD Multiband Antenna with T-tuner

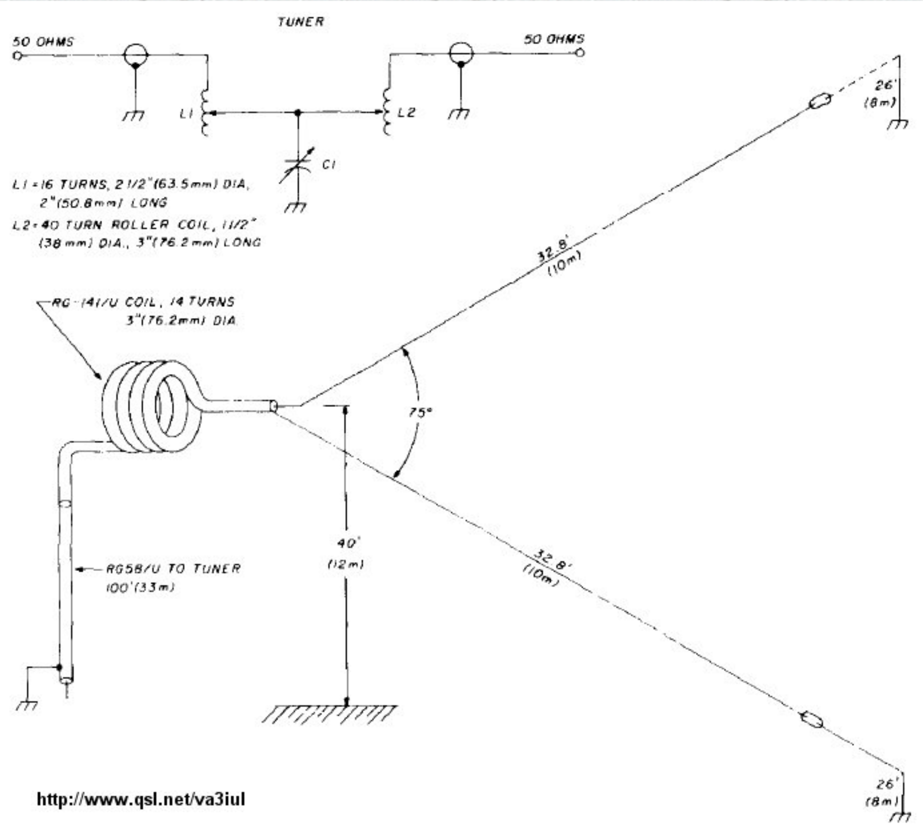
TUNER
50 OHMS
50 OHMS
L1
L2
C1
L1 = 16 TURNS, 2 1/2" (63.5mm) DIA, 2" (50.8mm) LONG
L2 = 40 TURN ROLLER COIL, 1 1/2" (38 mm) DIA., 3" (76.2 mm) LONG
26' (8m)
32.8' (10m)
RG-141/U COIL, 14 TURNS 3" (76.2mm) DIA
75°
40' (12m)
RG58/U TO TUNER 100' (33m)
32.8' (10m)
26' (8m)
http://www.qsl.net/va3iul

256 - Dual Band Dipole Antenna for 17m, 10m

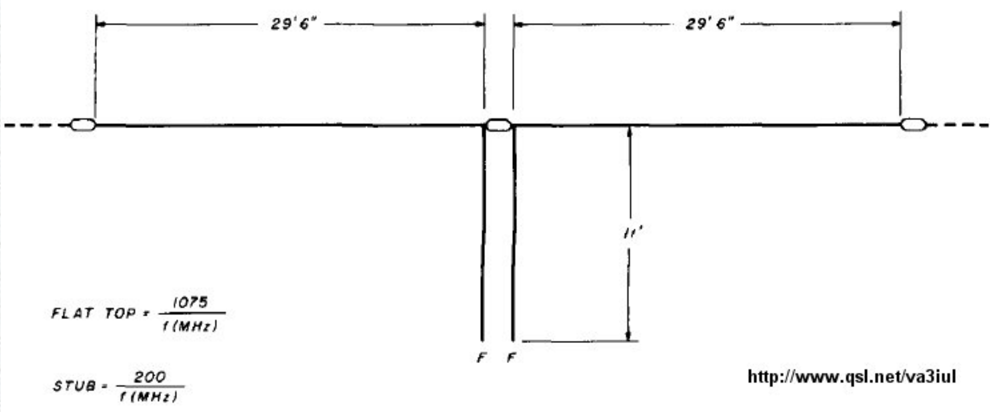
29'6"
29'6"
11'
F F
FLAT TOP = $\frac{1075}{f(MHz)}$
STUB = $\frac{200}{f(MHz)}$
http://www.qsl.net/va3iul

---

257 - Compact Travel Dipole Antenna for 10m, 15m, 20m

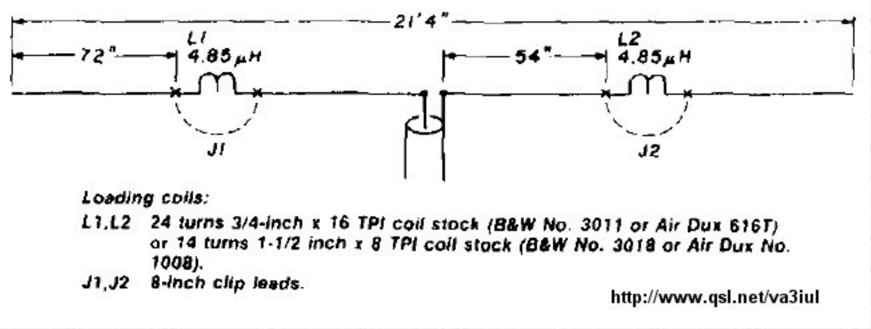
21'4"
72"
54"
L1
4.85 μH
L2
4.85 μH
J1
J2
http://www.qsl.net/va3iul

Loading coils:

L1,L2 24 turns 3/4-inch x 16 TPI coil stock (B&W No. 3011 or Air Dux 616T) or 14 turns 1-1/2 inch x 8 TPI coil stock (B&W No. 3018 or Air Dux No. 1008).

J1,J2 8-inch clip leads.

---

258 - End-Fire Antenna for 160m, 80m, 40m, 20m

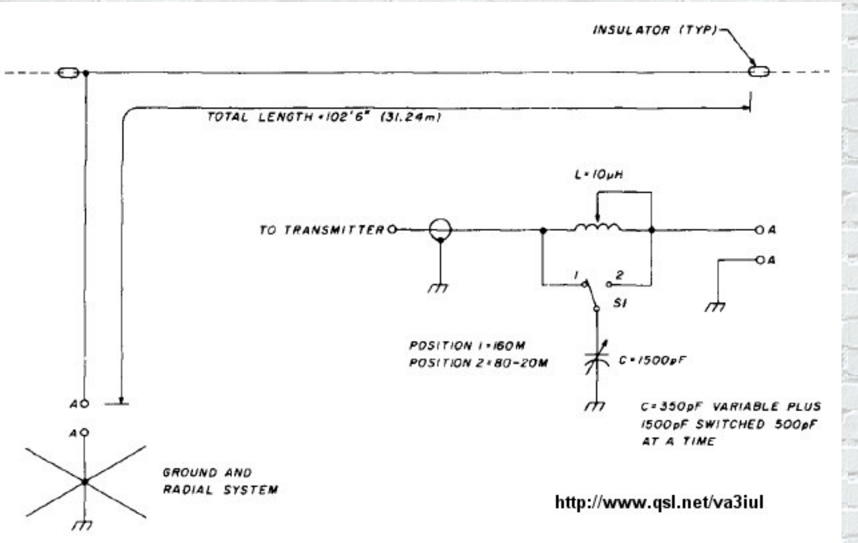
INSULATOR (TYP)
TOTAL LENGTH = 102'6" (31.24m)
L = 10μH
TO TRANSMITTER
A
A
1
2
S1
POSITION 1 = 160M
POSITION 2 = 80-20M
C = 1500pF
C = 350pF VARIABLE PLUS 1500pF SWITCHED 500pF AT A TIME
A
A
GROUND AND RADIAL SYSTEM
http://www.qsl.net/va3iul

---

259 - Four Band Loop for 40m, 20m, 15m, 10m

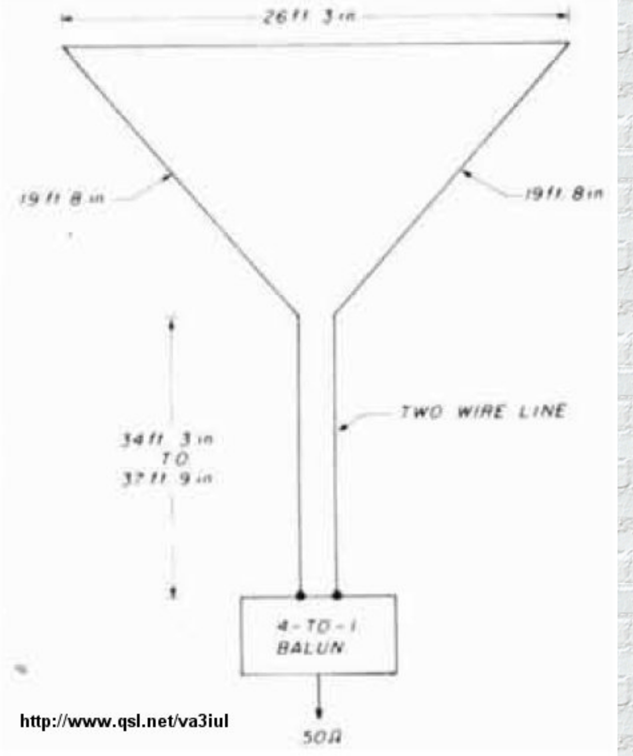
26 ft 3 in
19 ft 8 in
19 ft 8 in
34 ft 3 in
TO
37 ft 9 in
TWO WIRE LINE
4-TO-1
BALUN
50Ω

http://www.qsl.net/va3iul

---

260 - Carolina Windom Antenna for 80 to 10m

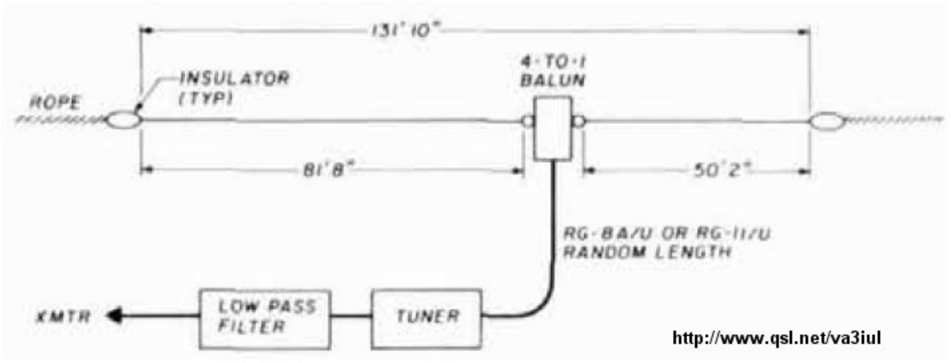
131' 10"
4-TO-1
BALUN
INSULATOR
(TYP)
ROPE
81' 8"
50' 2"
RG-8A/U OR RG-11/U
RANDOM LENGTH
XMTR
LOW PASS
FILTER
TUNER

http://www.qsl.net/va3iul

---

261 - Shortened Loaded Dipole for 40m

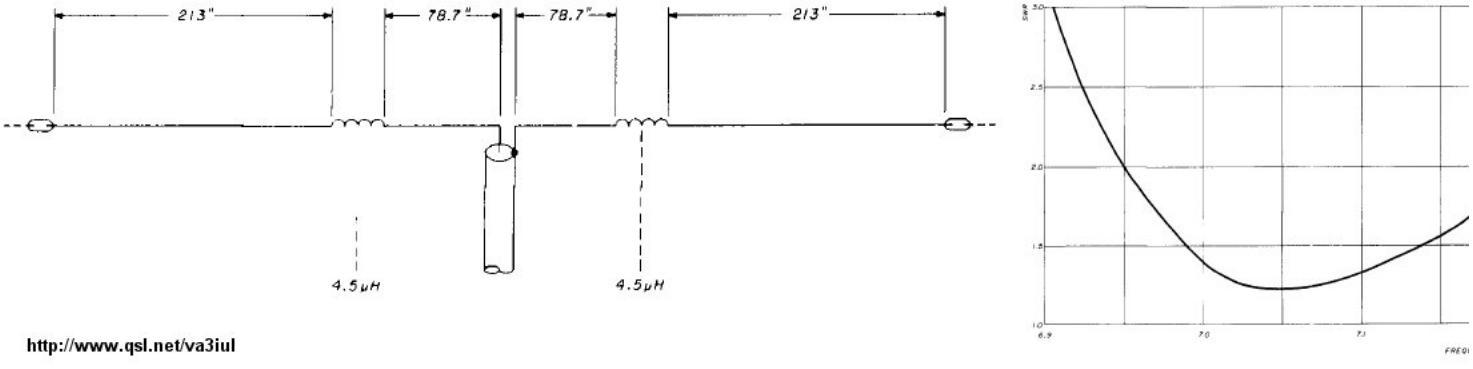
213"
78.7"
78.7"
213"
4.5µH
4.5µH
SWR
3.0
2.5
2.0
1.5
1.0
6.9
7.0
7.1
FREQU

http://www.qsl.net/va3iul

---

262 - Shortened Loaded Dipole for 80m

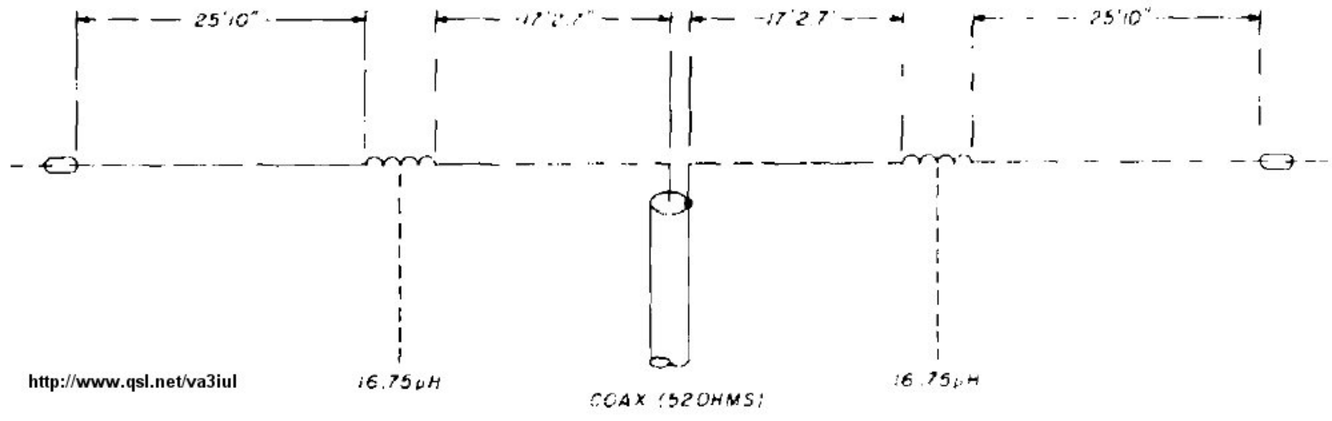
25'10"
17'2.7"
17'2.7"
25'10"
http://www.qsl.net/va3iul
16.75µH
COAX (52OHMS)
16.75µH

263 - Low Radiation Angle Inverted Delta Loop for 80m

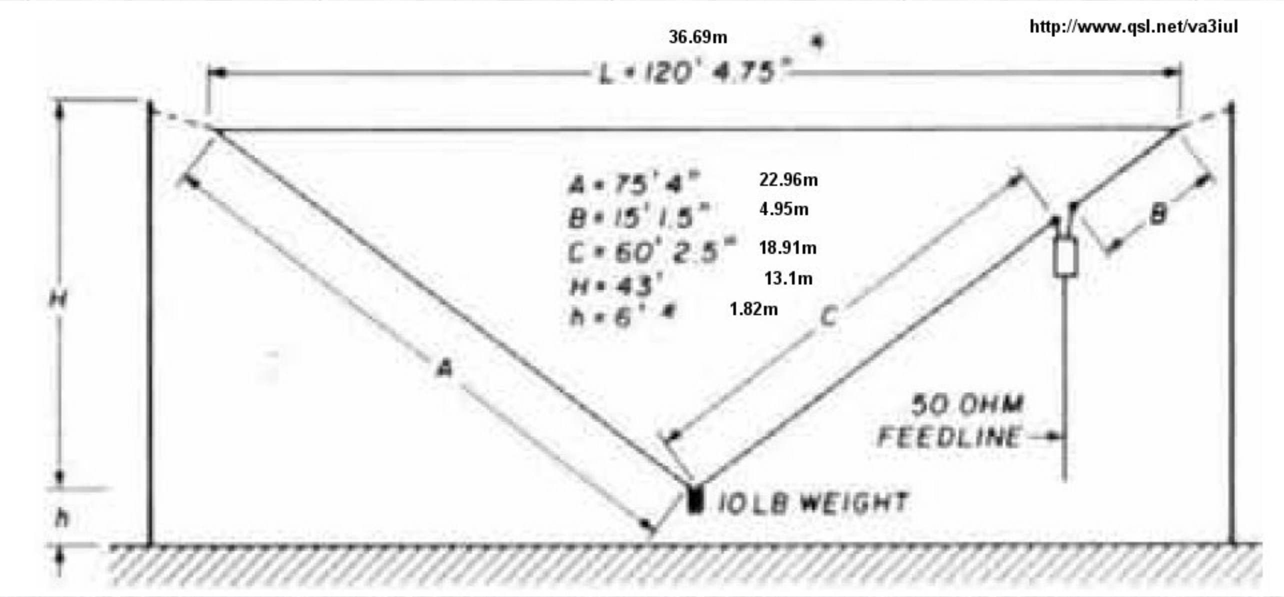
http://www.qsl.net/va3iul
36.69m
L = 120' 4.75" *
A = 75' 4" 22.96m
B = 15' 1.5" 4.95m
C = 60' 2.5" 18.91m
H = 43' 13.1m
h = 6' * 1.82m
H
h
A
B
C
50 OHM
FEEDLINE
10 LB WEIGHT

264 - Compact Dipole Antenna for 40m, 15m

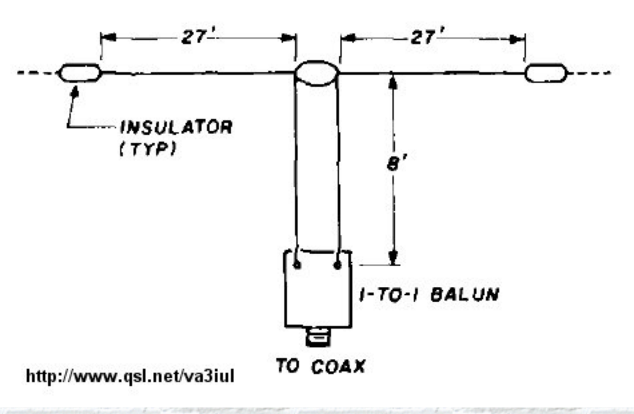
27'
27'
INSULATOR
(TYP)
8'
1-TO-1 BALUN
TO COAX
http://www.qsl.net/va3iul

265 - Vertical Delta Loop for 40m to 10m

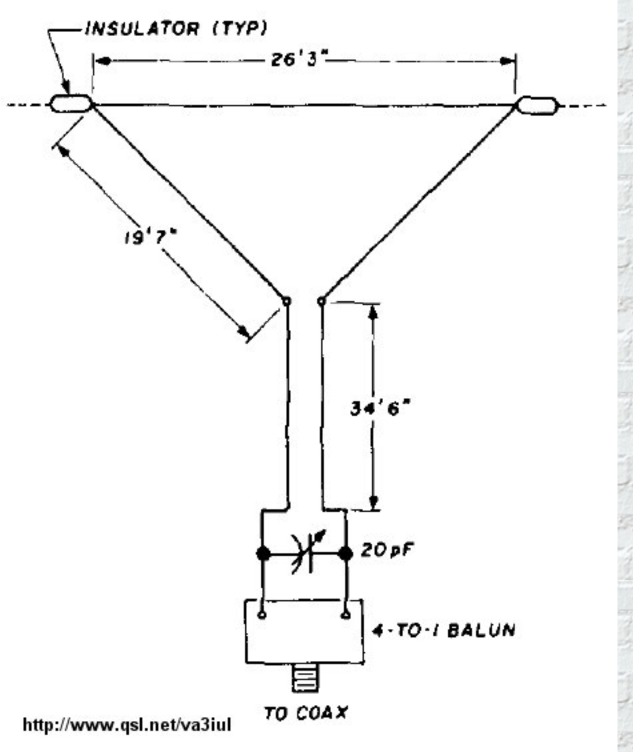
INSULATOR (TYP)
26'3"
19'7"
34'6"
20pF
4-TO-1 BALUN
TO COAX
http://www.qsl.net/va3iul

---

266 - Delta Loop Antenna for 17m

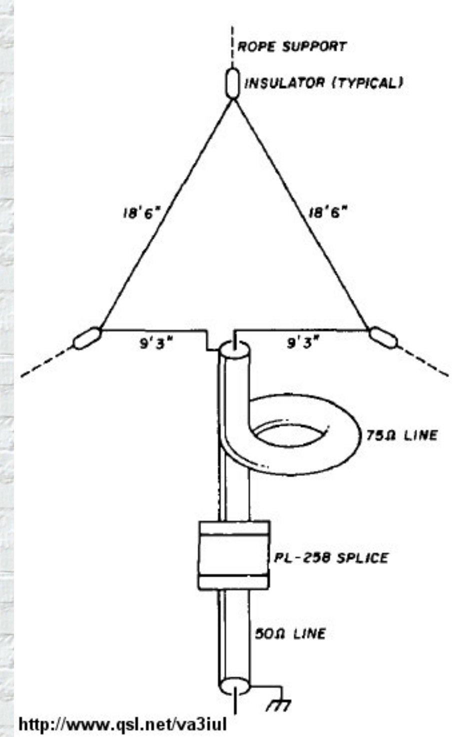
ROPE SUPPORT
INSULATOR (TYPICAL)
18'6"
18'6"
9'3"
9'3"
75Ω LINE
PL-258 SPLICE
50Ω LINE
http://www.qsl.net/va3iul

---

267 - Bi-Square Beam Gain=3dBd for 17m

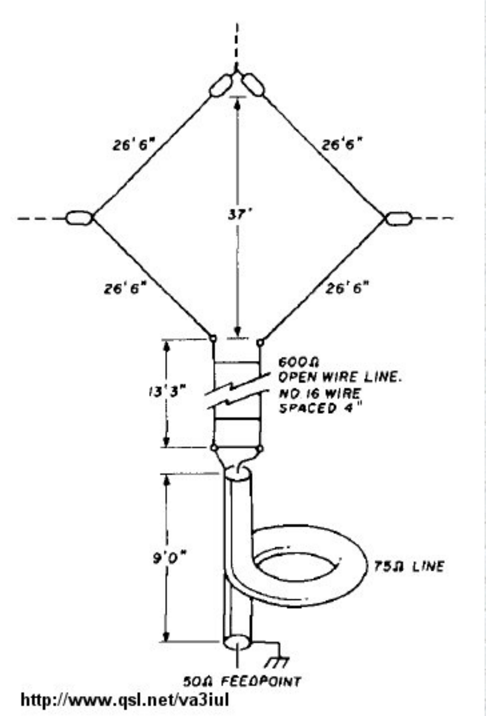
26'6"
26'6"
37'
26'6"
26'6"
600Ω
OPEN WIRE LINE.
NO 16 WIRE
SPACED 4"
13'3"
9'0"
75Ω LINE
50Ω FEEDPOINT
http://www.qsl.net/va3iul

---

268 -Trap Dipole Antenna for 17m,13m

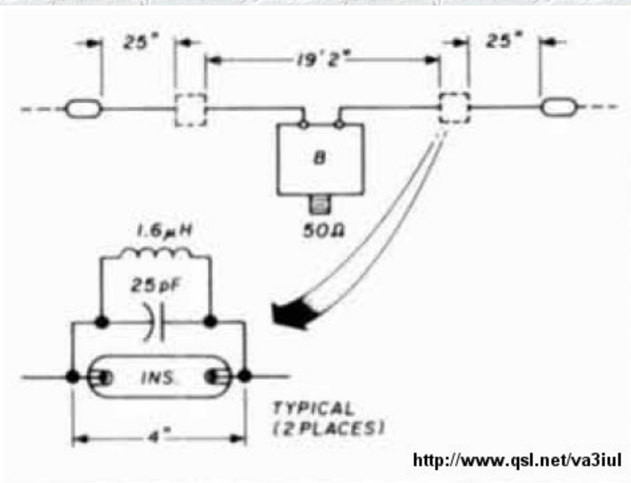
25"
19'2"
25"
B
50Ω
1.6μH
25pF
INS.
4"
TYPICAL
(2 PLACES)
http://www.qsl.net/va3iul

---

269 - Hanging Unipole Antenna for 160m, 80m

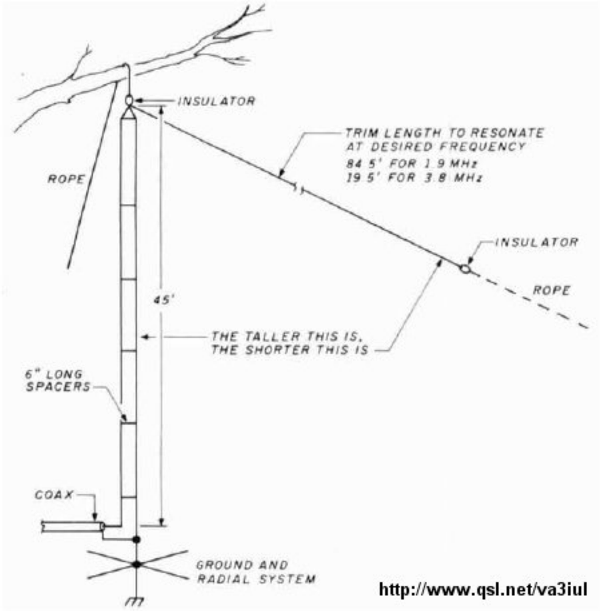
INSULATOR
TRIM LENGTH TO RESONATE
AT DESIRED FREQUENCY
84.5' FOR 1.9 MHz
19.5' FOR 3.8 MHz
ROPE
INSULATOR
ROPE
45'
THE TALLER THIS IS,
THE SHORTER THIS IS
6" LONG
SPACERS
COAX
GROUND AND
RADIAL SYSTEM

http://www.qsl.net/va3iul

---

270 - Wideband Sloping Vee Antenna - 30m to 6m

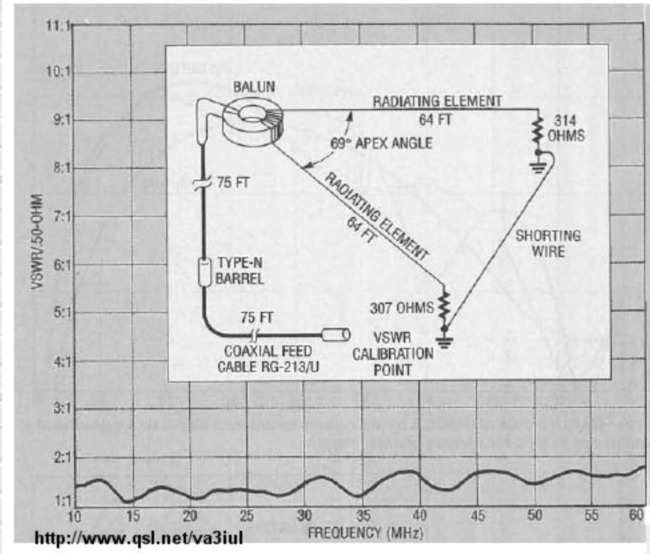

http://www.qsl.net/va3iul

SOURCES OF MATERIALS

The following companies are sources for materials and computer software for this project:

- Toroidal ferrite cores (Part No. FT240-43)—Radio Kit, Inc., P.O. Box 973, Pelham, NH 03076, (603) 635-2235
- Film power resistors—Power Film Systems, Inc., Yellville, AR 72687, (501) 449-4091
- Antenna design software—Phadean Engineering Co., Inc., P.O. Box 611, Shrewsbury, MA 01545, (508) 869-6077
- Phosphor-bronze wire—Astro Industries, Inc., Dayton, OH 43432, (800) 543-5810
- Fiberglass tubing—J. T. Ryerson Co., P. O. Box 1111, Boston, MA 02103. (617) 782-6900

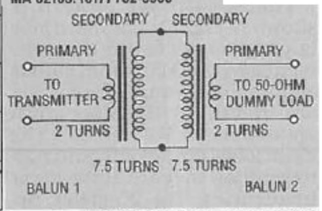
SECONDARY
SECONDARY
PRIMARY
PRIMARY
TO
TRANSMITTER
TO 50-OHM
DUMMY LOAD
2 TURNS
2 TURNS
7.5 TURNS
7.5 TURNS
BALUN 1
BALUN 2

-BACK-TO-BACK INSERTION LOSS TEST of baluns for sloping-vee antenna.

---

271 - 5/8 Wavelength Antenna

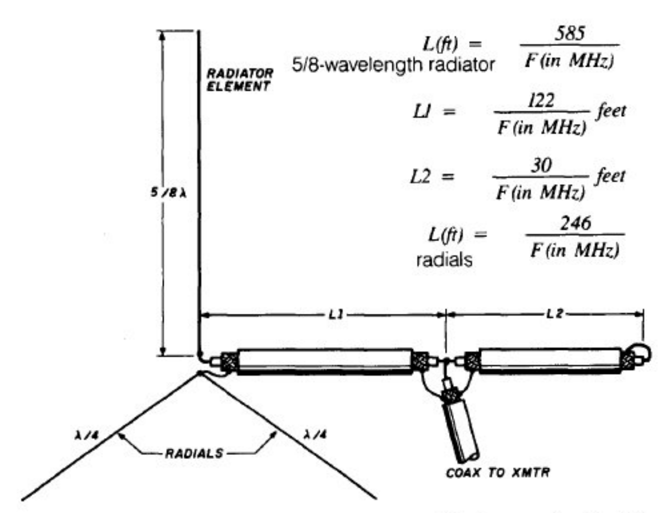
RADIATOR ELEMENT
5/8-wavelength radiator
L(ft) = 585 / F(in MHz)
L1 = 122 / F(in MHz) feet
L2 = 30 / F(in MHz) feet
L(ft) = 246 / F(in MHz)
radials
5/8 λ
L1
L2
λ/4
RADIALS
λ/4
COAX TO XMTR

http://www.qsl.net/va3iul

---

272 - C-Pole Vertical Antenna

http://www.qsl.net/va3iul

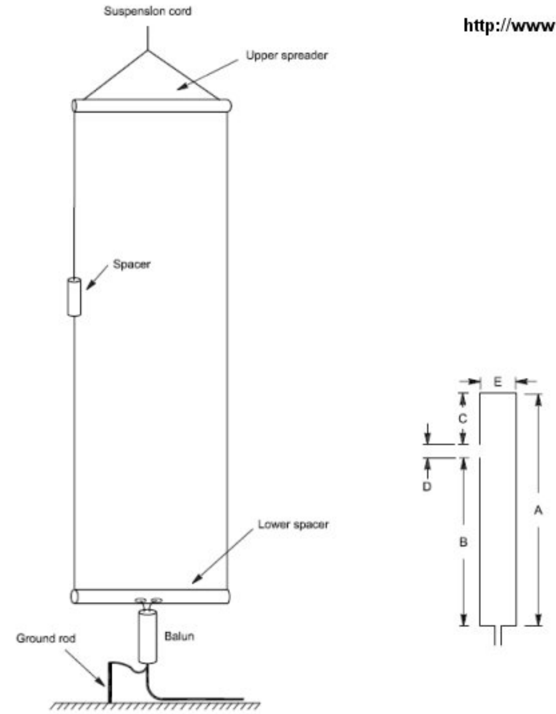
Suspension cord
Upper spreader
Spacer
Lower spacer
Ground rod
Balun
E
C
D
A
B

Band (meters)	A (inches)	B (inches)	C (inches)	D (inches)	E (inches)	2:1 SWR bandwidth (kHz)
160	1666	924	994	60	80	58
80	840	460	360	30	40	120
60	591	322	249	20	26	250
40	450	240	190	20	20	260
30	320	167	139	14	14	360
20	177	85	84	8	40	400
15	124	60	60	4	20	600
10	87	46	37	4	20	800

Wire diameter is 1/16-inch. Height of the lower horizontal wire is 12 inches to 24 inches and is non-critical.

---

273 - Bobtail Antenna for 30m

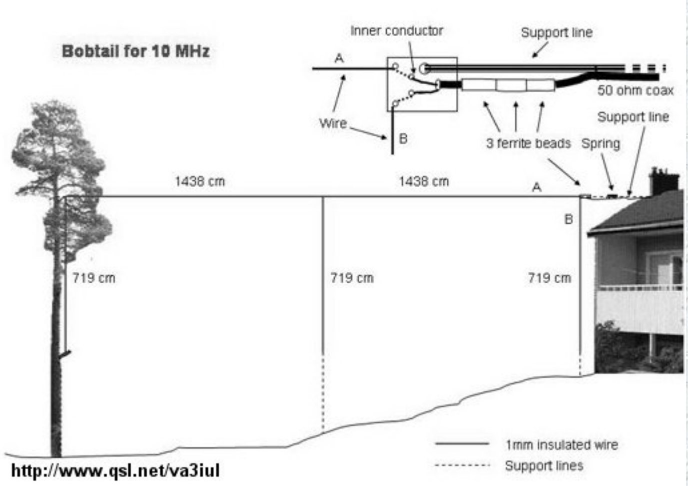
Bobtail for 10 MHz
Inner conductor
Support line
A
50 ohm coax
Wire
B
3 ferrite beads
Support line
Spring
1438 cm
1438 cm
A
B
719 cm
719 cm
719 cm
1mm insulated wire
Support lines
http://www.qsl.net/va3iul

274 - Bobtail Antenna for 40m

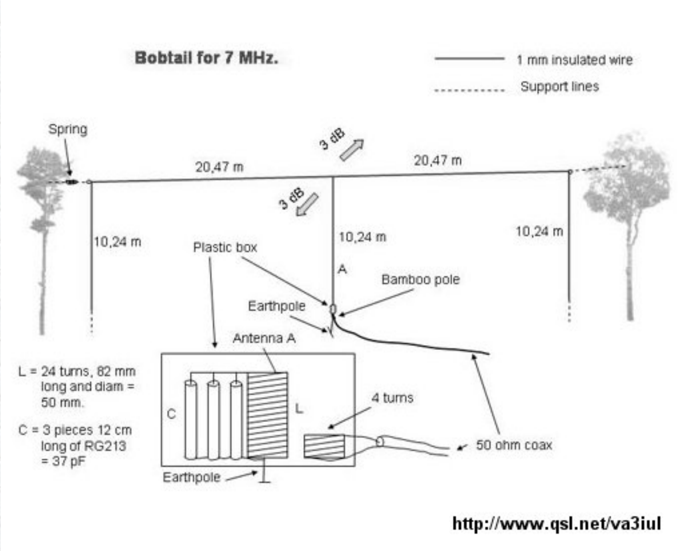
Bobtail for 7 MHz.
1 mm insulated wire
Support lines
Spring
3 dB
20,47 m
20,47 m
3 dB
10,24 m
10,24 m
10,24 m
Plastic box
A
Bamboo pole
Earthpole
Antenna A
L = 24 turns, 82 mm long and diam = 50 mm.
C = 3 pieces 12 cm long of RG213 = 37 pF
C
L
4 turns
50 ohm coax
Earthpole
http://www.qsl.net/va3iul

275 - Two Elements Vertical Loop Antenna for 40m

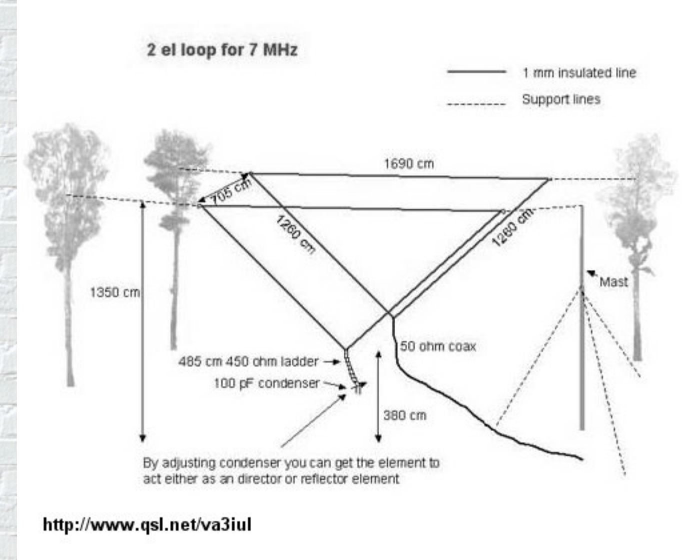
2 el loop for 7 MHz
1 mm insulated line
Support lines
1690 cm
705 cm
1260 cm
1260 cm
Mast
1350 cm
50 ohm coax
485 cm 450 ohm ladder
100 pF condenser
380 cm
By adjusting condenser you can get the element to act either as an director or reflector element
http://www.qsl.net/va3iul

276 - Double Extended Zepp Antenna for 40m

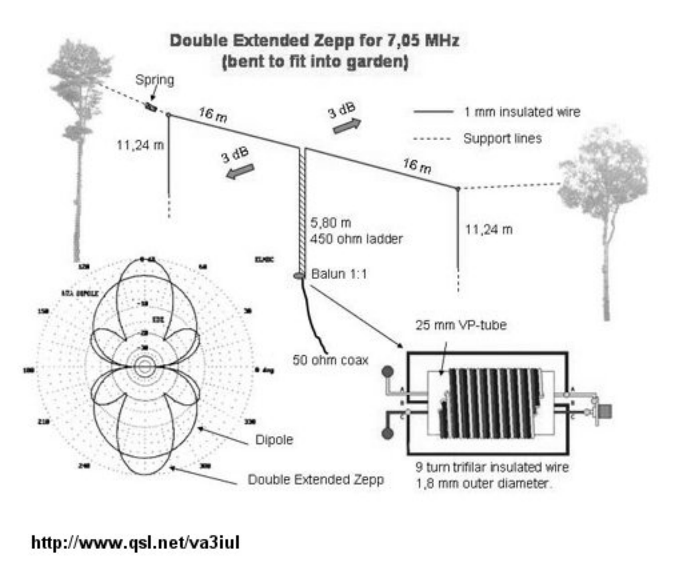
Double Extended Zepp for 7,05 MHz
(bent to fit into garden)
Spring
16 m
3 dB
1 mm insulated wire
Support lines
11,24 m
3 dB
16 m
5,80 m
450 ohm ladder
11,24 m
Balun 1:1
25 mm VP-tube
50 ohm coax
Dipole
Double Extended Zepp
9 turn trifilar insulated wire
1,8 mm outer diameter
http://www.qsl.net/va3iul

---

277 - Reversible Moxon Flip Beam for 40m

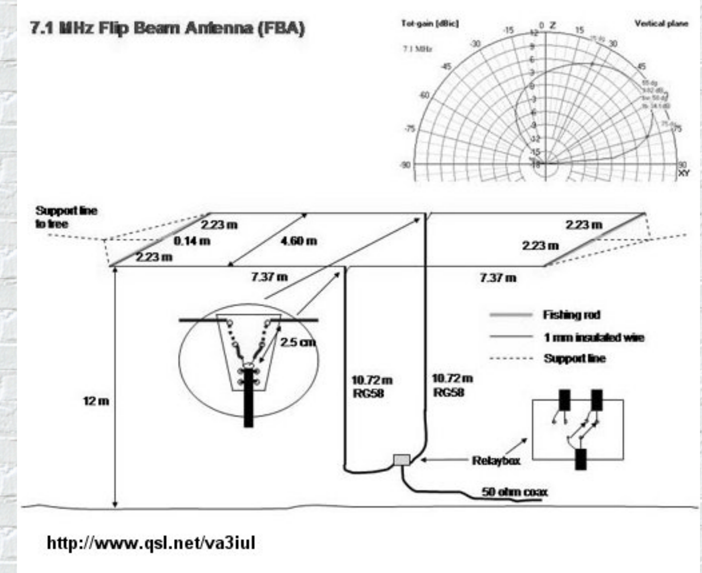
7.1 MHz Flip Beam Antenna (FBA)
Tot gain [dBic]
Vertical plane
7.1 MHz
Support line to tree
2.23 m
0.14 m
2.23 m
4.60 m
7.37 m
2.23 m
2.23 m
7.37 m
2.5 cm
Fishing rod
1 mm insulated wire
Support line
12 m
10.72 m
RG58
10.72 m
RG58
Relaybox
50 ohm coax
http://www.qsl.net/va3iul

---

278 - Spitwire Antenna for 40m

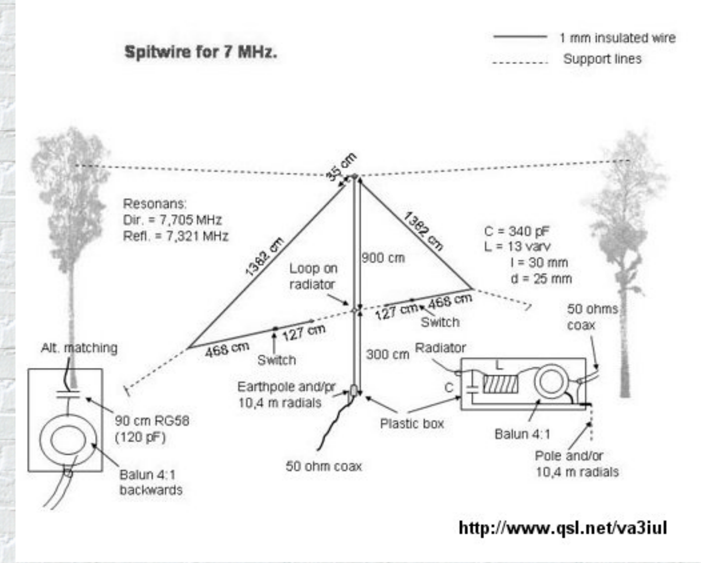
Spitwire for 7 MHz.
1 mm insulated wire
Support lines
35 cm
Resonans:
Dir. = 7,705 MHz
Refl. = 7,321 MHz
1382 cm
1382 cm
900 cm
Loop on
radiator
C = 340 pF
L = 13 varv
l = 30 mm
d = 25 mm
127 cm
468 cm
Switch
50 ohms
coax
468 cm
127 cm
Switch
Alt. matching
300 cm
Radiator
L
C
Earthpole and/pr
10,4 m radials
90 cm RG58
(120 pF)
Plastic box
Balun 4:1
Pole and/or
10,4 m radials
50 ohm coax
Balun 4:1
backwards
http://www.qsl.net/va3iul

279 - Fork Antenna for 40m

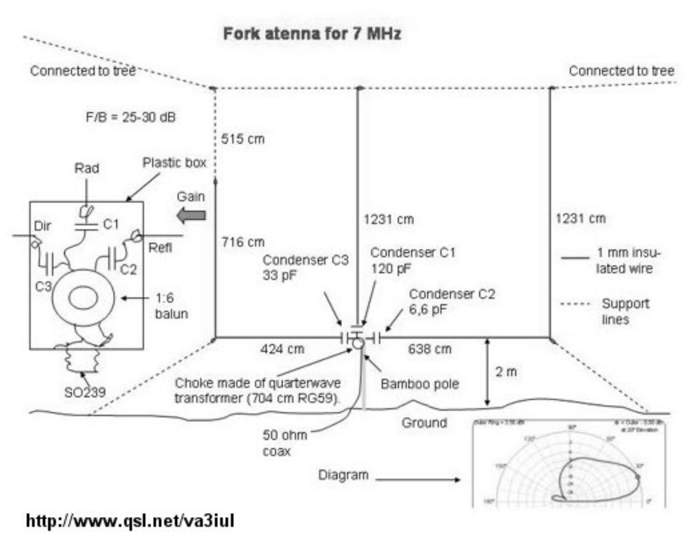
Fork atenna for 7 MHz
Connected to tree
Connected to tree
F/B = 25-30 dB
515 cm
Rad
Plastic box
Gain
Dir
C1
Refl
C2
C3
1:6
balun
SO239
716 cm
1231 cm
1231 cm
Condenser C3
33 pF
Condenser C1
120 pF
1 mm insu-
lated wire
Condenser C2
6,6 pF
Support
lines
424 cm
638 cm
2 m
Choke made of quarterwave
transformer (704 cm RG59).
Bamboo pole
Ground
50 ohm
coax
Diagram
http://www.qsl.net/va3iul

280 - Veebeam Antenna 3dB-Gain for All Bands

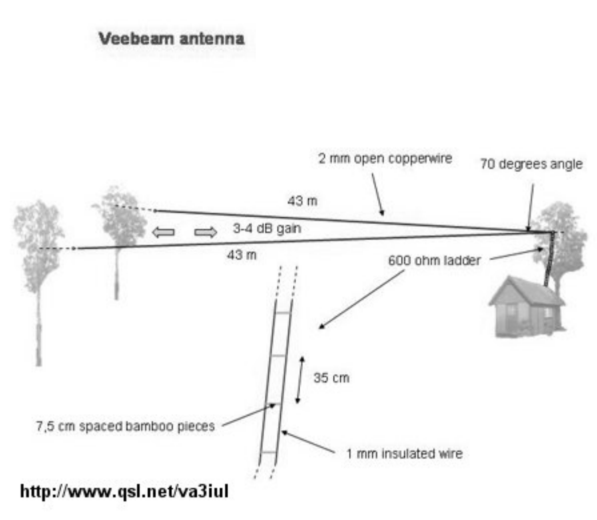
Veebeam antenna
2 mm open copperwire
70 degrees angle
43 m
3-4 dB gain
43 m
600 ohm ladder
35 cm
7,5 cm spaced bamboo pieces
1 mm insulated wire
http://www.qsl.net/va3iul

281 - Vertical Moxon Antenna for 10m

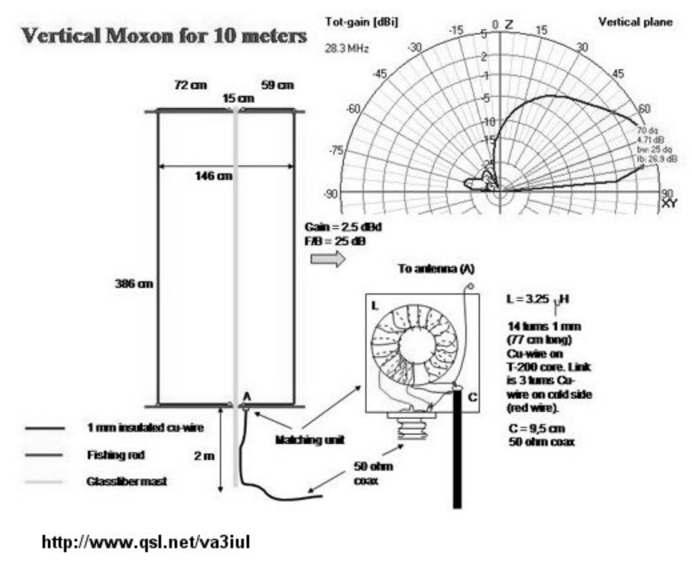
Vertical Moxon for 10 meters
Tot-gain [dBi]
28.3 MHz
Vertical plane
0 Z
-15
15
-30
30
-45
45
-60
60
-75
-90
90
XY
2
1
5
10
15
70 dg
4.71 dB
bw: 25 dg
fb: 26.9 dB
72 cm
59 cm
15 cm
146 cm
386 cm
A
2 m
1 mm insulated cu-wire
Fishing rod
Glassfibermast
Gain = 2.5 dBd
F/B = 25 dB
To antenna (A)
L
C
Matching unit
50 ohm coax
L = 3.25 µH
14 turns 1 mm (77 cm long) Cu-wire on T-200 core. Link is 3 turns Cu-wire on cold side (red wire).
C = 9,5 cm 50 ohm coax
http://www.qsl.net/va3iul

---

282 - Half-Square Antenna for 30m

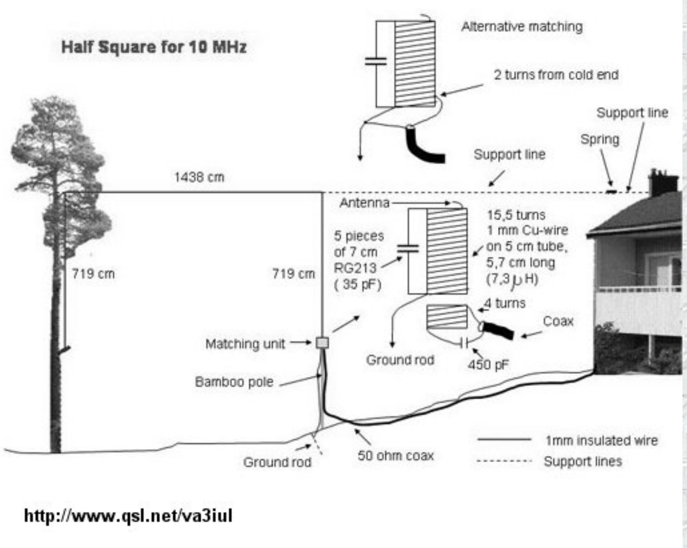
Half Square for 10 MHz
Alternative matching
2 turns from cold end
Support line
Spring
Support line
1438 cm
Antenna
15,5 turns 1 mm Cu-wire on 5 cm tube, 5,7 cm long (7,3 µH)
5 pieces of 7 cm RG213 (35 pF)
719 cm
719 cm
4 turns
Coax
Matching unit
Ground rod
450 pF
Bamboo pole
Ground rod
50 ohm coax
1mm insulated wire
Support lines
http://www.qsl.net/va3iul

---

283 - Rotatable Vertical Delta Loop for 20m

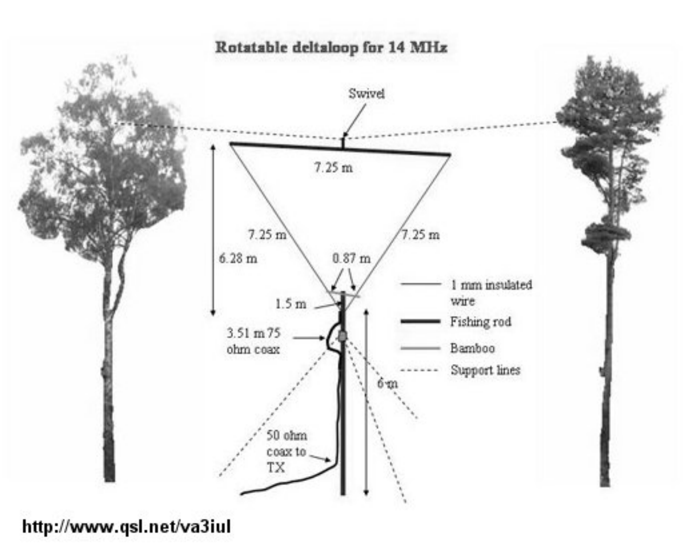
Rotatable deltaloop for 14 MHz
Swivel
7.25 m
7.25 m
7.25 m
6.28 m
0.87 m
1.5 m
3.51 m 75 ohm coax
6 m
50 ohm coax to TX
1 mm insulated wire
Fishing rod
Bamboo
Support lines
http://www.qsl.net/va3iul

---

284 - Mini-Horse Antenna 10dB-Gain for 40m-to-2m

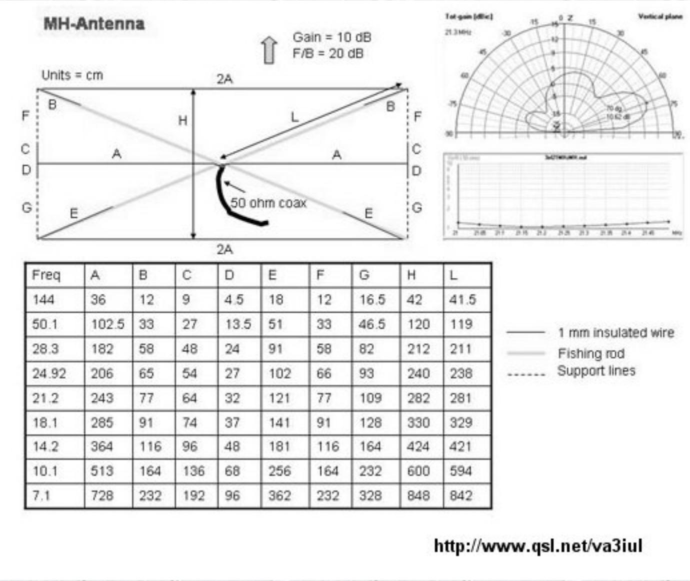
MH-Antenna
Gain = 10 dB
F/B = 20 dB
Units = cm
2A
B
H
L
B
F
F
C
D
A
A
C
D
G
E
50 ohm coax
E
G
2A
1 mm insulated wire
Fishing rod
Support lines
http://www.qsl.net/va3iul

Freq	A	B	C	D	E	F	G	H	L
144	36	12	9	4.5	18	12	16.5	42	41.5
50.1	102.5	33	27	13.5	51	33	46.5	120	119
28.3	182	58	48	24	91	58	82	212	211
24.92	206	65	54	27	102	66	93	240	238
21.2	243	77	64	32	121	77	109	282	281
18.1	285	91	74	37	141	91	128	330	329
14.2	364	116	96	48	181	116	164	424	421
10.1	513	164	136	68	256	164	232	600	594
7.1	728	232	192	96	362	232	328	848	842

---

285 - Compact Moxon Antenna 5dB-Gain for 40m

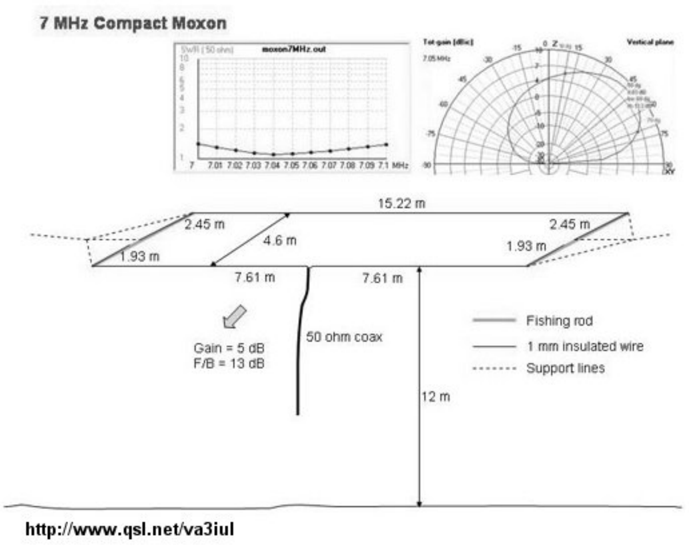
7 MHz Compact Moxon
moxon7MHz.out
7 7.01 7.02 7.03 7.04 7.05 7.06 7.07 7.08 7.09 7.1 MHz
Vertical plane
15.22 m
2.45 m
2.45 m
4.6 m
1.93 m
1.93 m
7.61 m
7.61 m
Gain = 5 dB
F/B = 13 dB
50 ohm coax
12 m
Fishing rod
1 mm insulated wire
Support lines
http://www.qsl.net/va3iul

286 - Two Elements Sloping Delta Loop 8dB-Gain for 10m

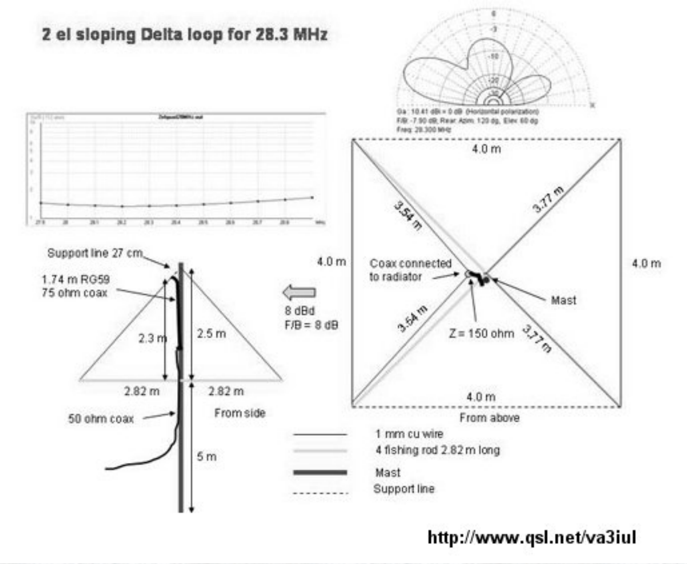
2 el sloping Delta loop for 28.3 MHz
Support line 27 cm
1.74 m RG59
75 ohm coax
2.3 m
2.5 m
2.82 m
2.82 m
From side
50 ohm coax
5 m
8 dBd
F/B = 8 dB
4.0 m
4.0 m
4.0 m
4.0 m
3.54 m
3.77 m
3.54 m
3.77 m
Coax connected
to radiator
Mast
Z = 150 ohm
From above
1 mm cu wire
4 fishing rod 2.82 m long
Mast
Support line
http://www.qsl.net/va3iul

---

287 - Two Elements Phased Vertical Antenna 3dB-Gain for 40m

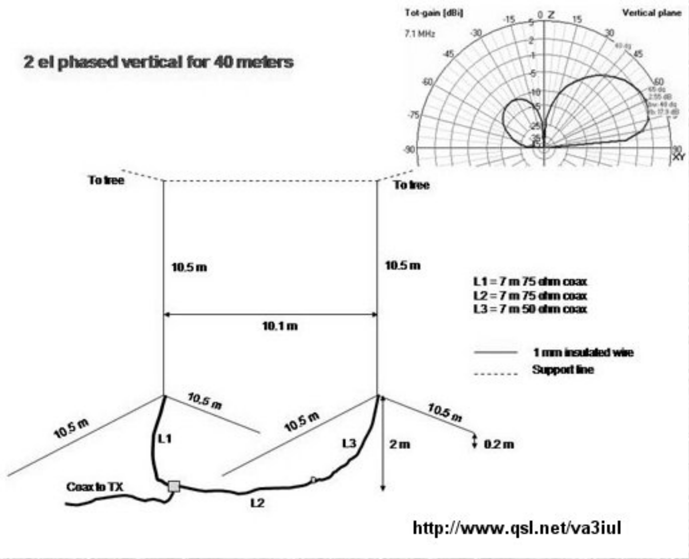
2 el phased vertical for 40 meters
Tot-gain [dBi]
7.1 MHz
Vertical plane
To tree
To tree
10.5 m
10.5 m
10.1 m
L1 = 7 m 75 ohm coax
L2 = 7 m 75 ohm coax
L3 = 7 m 50 ohm coax
1 mm insulated wire
Support line
10.5 m
10.5 m
10.5 m
10.5 m
L1
L3
2 m
0.2 m
Coax to TX
L2
http://www.qsl.net/va3iul

References:

Antenna Toolkit - J. Carr
Practical Antenna Handbook - J. Carr
More Antenna Classics - C. Hutchinson
Amateur Radio Techniques - P. Hawker
The Radio Handbook - W. Orr (editor), 15th edition
The Radio Antenna Handbook - Engineering Staff of "Radio"
Jones Antenna Handbook - Frank C. Jones
SM0DTK - http://www.sm0dtk.se/antennas.htm
ARRL Antenna Handbook 1990-2007
ARRL Handbook 1937-2009
ARRL Antenna Compendium
73 Magazine 1970-2002
Funkamateur Magazine 1990-2011
RSGB Handbook - 2007 - 2010
QST Magazine - 1980 - 2010
Ham Radio Magazine - 1969 - 1990
Radio-Electronics Magazine - 1986 - 1990

home: ***http://www.qsl.net/va3iul***

Live Traffic Feed

Watch Live | Get FEEDJIT

Richmond, California, United States arrived 0 secs ago.

United Kingdom arrived 2 mins ago.

Kundan, Kashmir, India arrived 3 mins ago.

San Francisco, California, United States arrived 9 mins ago.

Rostock, Mecklenburg-Vorpommern, Germany arrived 13 mins ago.

Russia arrived 16 mins ago.

Abu Dhabi, United Arab Emirates arrived 32 mins ago.

United States arrived 35 mins ago.

Menlo Park, California, United States arrived 58 mins ago.

Québec, Canada arrived 59 mins ago.

Radio and Electronics Cookbook

Radio and Electronics Cookbook

Edited by
Dr George Brown, CEng, FIEE, M5ACN

RSGB

Newnes
OXFORD AUCKLAND BOSTON JOHANNESBURG MELBOURNE NEW DELHI

Newnes
An imprint of Butterworth-Heinemann
Linacre House, Jordan Hill, Oxford OX2 8DP
225 Wildwood Avenue, Woburn, MA 01801-2041
A division of Reed Educational and Professional Publishing Ltd

A member of the Reed Elsevier plc group

First published 2001

© Radio Society of Great Britain 2001

All rights reserved. No part of this publication may be reproduced in any material form (including photocopying or storing in any medium by electronic means and whether or not transiently or incidentally to some other use of this publication) without the written permission of the copyright holder except in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, England W1P 0LP. Applications for the copyright holder's written permission to reproduce any part of this publication should be addressed to the publishers

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7506 5214 4

RSGB
Lambda House
Cranborne Road
Potters Bar
Herts
EN6 3JE

Composition by Genesis Typesetting, Laser Quay, Rochester, Kent
Printed and bound in Great Britain

PLANT A TREE
BTCV
British Trust for Conservation Volunteers
FOR EVERY TITLE THAT WE PUBLISH, BUTTERWORTH-HEINEMANN
WILL PAY FOR BTCV TO PLANT AND CARE FOR A TREE.

Contents

Preface ix

1. A medium-wave receiver 1
2. An audio-frequency amplifier 4
3. A medium-wave receiver using a ferrite-rod aerial 9
4. A simple electronic organ 12
5. Experiments with the NE555 timer 17
6. A simple metronome 21
7. What is a resistor? 24
8. Waves – Part 1 27
9. A beat-frequency oscillator 31
10. What is a capacitor? 34
11. Waves – Part 2 38
12. An LED flasher 41
13. Waves – Part 3 44
14. Choosing a switch 46
15. An aerial tuning unit for a receiver 49
16. A simple 2 m receiver preamplifier 52
17. Receiving aerials for amateur radio 54
18. The Colt 80 m receiver – Part 1 58
19. A crystal radio receiver 62
20. The varactor (or varicap) diode 64
21. A portable radio for medium waves 65
22. The Colt 80 m receiver – Part 2 70

23. A simple transistor tester 73
24. An introduction to transmitters 77
25. The Colt 80 m receiver – Part 3 81
26. A two-way Morse practice system 88
27. The Colt 80 m receiver – Part 4 91
28. A simple crystal set 95
29. A crystal calibrator 100
30. A simple short-wave receiver – Part 1 104
31. A fruit-powered medium-wave radio 106
32. A capacitance bridge 109
33. A simple short-wave receiver – Part 2 113
34. A basic continuity tester 117
35. A charger for NiCad batteries 119
36. An 80 metre crystal-controlled CW transmitter 123
37. A solar-powered MW radio 129
38. A receiver for the 7 MHz amateur band 133
39. Diodes for protection 137
40. An RF signal probe 140
41. An RF changeover circuit 142
42. A low-light indicator 146
43. A J-pole aerial for 50 MHz 149
44. Measuring light intensity – the photometer 153
45. A 70 cm Quad loop aerial 156
46. A UHF field strength meter 160
47. Christmas tree LEDs 162
48. An audio signal injector 166
49. Standing waves 168
50. A standing-wave indicator for HF 170
51. A moisture meter 174
52. Simple aerials 177
53. A breadboard 80 cm CW transmitter 182

54. A 7-element low-pass filter for transmitters 186
55. Radio-frequency mixing explained 189
56. A voltage monitor for a 12 V power supply 192
57. A 1750 Hz toneburst for repeater access 196
58. A circuit for flashing LEDs 201
59. Digital logic circuits 205
60. A resistive SWR indicator 210
61. An audio filter for CW 213
62. An electronic die 215
63. The absorption wavemeter 222
64. An HF absorption wavemeter 224
65. A vertical aerial for 70 cm 228
66. A UHF corner reflector aerial 230
67. A switched dummy load 234
68. A simple Morse oscillator 238
69. A bipolar transistor tester 240
70. The 'Yearling' 20 m receiver 245
71. Adding the 80 metre band to the Yearling receiver 251
72. How the Yearling works 255
73. A field strength meter 258
74. Preselector for a short-wave receiver 261
75. An audible continuity tester 265
76. An experimental 70 cm rhombic aerial 268
77. Water level alarm 272
78. A delta loop for 20 metres 275
79. A simple desk microphone 279
80. Morse oscillator 284
81. A simple 6 m beam 287
82. An integrated circuit amplifier 291
83. A novice ATU 293
84. CW QRP transmitter for 80 metres 297

85. An audio booster for your hand-held 303
86. A grid dip oscillator 306
87. A CW transmitter for 160 to 20 metres 312
88. Matching the end-fed random-wire aerial 315

Preface

Although we are surrounded by sophisticated computerised gadgets these days, there is still a fascination in putting together a few resistors, capacitors and the odd transistor to make a simple electronic circuit. It is really surprising how a handful of components can perform a useful function, and the satisfaction of having built it yourself is incalcuable.

This book aims to provide a wide variety of radio and electronic projects, from something that will take a few minutes to a more ambitious weekend's worth. Various construction techniques are described, the simplest requiring no more than a small screwdriver, the most complex involving printed circuit boards.

Originally published by the Radio Society of Great Britain, the projects were all chosen to be useful and straightforward, with the emphasis on practicality. In most cases the workings of the circuit are described, and the projects are backed up by small tutorials on the components and concepts employed. In the 21st century it may seem strange that few of the published circuits use integrated circuits (chips). This is intentional as it is much easier to understand how the circuit works when using discrete components.

Anyone buying the *Radio and Electronics Cookbook* will find that it will lead to hours of enjoyment, some very useful and entertaining gadgets, and increased knowledge of how and why electronics circuits work, and a great sense of satisfaction. Beware, electronic construction is addictive!

WARNING: This book contains construction details of transmitters. It is illegal to operate a transmitter without the appropriate licence. Information on how to obtain an Amateur Radio Licence can be obtained from the Radiocommunications Agency, tel. 020 7211 0160.

1 A medium-wave receiver

Introduction

Let us start off with something that is really quite simple and yet is capable of producing a sense of real satisfaction when complete – a real medium-wave (MW) radio receiver! It proves that receivers *can* be simple and, at the same time, be useful and enjoyable to make. To minimise the confusion to absolute beginners, *no* circuit diagram is given, only the constructional details. The circuits will come later, when you have become accustomed to the building process. In the true amateur spirit of ingenuity and inventiveness, the circuit is built on a terminal strip, the coil is wound on a toilet roll tube (as amateur MW coils have been for 100 years!), and the receiver is mounted on a piece of wood.

Putting it together

Start by mounting the components on the terminal strip as shown in **Figure 1**, carefully checking the position and value of each one. The three capacitors are all the same, and so present no problem. They (and the resistors) may be connected either way round, unlike the two semiconductors (see later). The resistors are coded by means of coloured bands. You can refer to Chapter 7 if you have difficulty remembering the colours and their values.

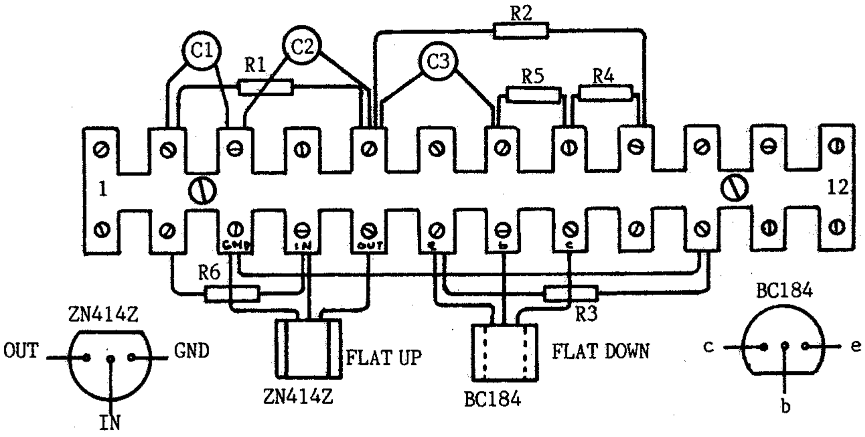

Figure 1 Terminal strip – position of components

1. Brown, Black, Yellow	100 000 ohms	(R1, R5, R6)
2. Green, Blue, Brown	560 ohms	(R2)
3. Red, Violet, Brown	270 ohms	(R3)
4. Brown, Black, Orange	10 000 ohms	(R4)

The integrated circuit (the ZN414Z) and the transistor (the BC184) *must* be connected correctly. Check Figure 1 carefully before fitting each device.

Now wind the coil. Most tubes are about 42 mm diameter and 110 mm long. Don't worry if your tube is slightly different; it shouldn't matter. Make two holes, about 3 mm apart, about 40 mm from one end, as shown in **Figure 2**. Loop your enamelled wire into one hole and out of the other, and draw about 100 mm through; loop this 100 mm through again, thus anchoring the wire firmly. Now wind on 80 turns, keeping the wire tight and the turns close together but not overlapping. After your 80th turn, make another two holes and anchor the wire in the same way as before. Again, leave about 100 mm free after anchoring. Using another piece of enamelled wire (with 100 mm ends as before), loop one end through the same two holes which contain the end anchor of the last winding, wind two turns and anchor the end of this short winding using another pair of holes. Figure 2 shows the layout.

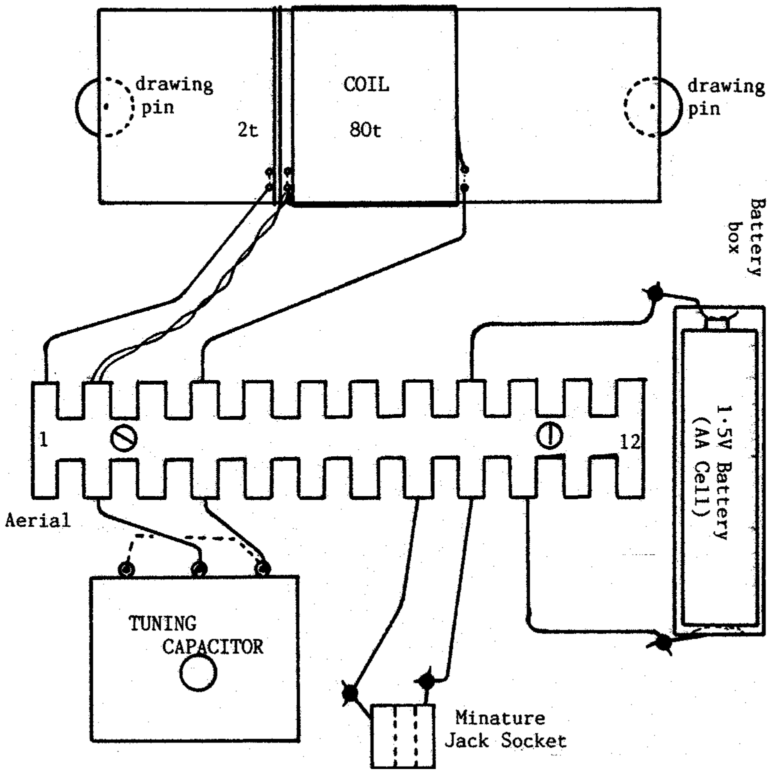

Figure 2 The layout of the parts on the wooden base

With some glass paper, remove the enamel from the ends of both pieces of wire which go through the same holes (i.e. the bottom of the large coil and the top of the small coil), then twist these bare ends together. Remove the enamel from the remaining ends of the coil. The coil is now finished!

The baseboard can be any piece of wood about 150 mm square. Fix the coil near the back edge using drawing pins and connect the wires from the coils to the terminal strip as shown in Figure 2. Using short pieces of PVC-insulated wire (and with assistance if you have never soldered before), solder one piece across the two outer tags of the variable capacitor, shown by the dotted line in Figure 2, and then two longer pieces to the centre tag and one outside tag. Connect these to the terminal strip. Then solder two more insulated wires on to the jack socket (into which you will plug your crystal earpiece), the other ends going to the terminal strip. The last two wires (one must be red) need to be soldered on to the battery box, their other ends going to the terminal strip also. Make sure the red wire goes to the *positive* terminal on the battery, and is connected to terminal 9. The other connection to the battery goes to terminal 10.

Attach the terminal strip to the baseboard with small screws or double-sided sticky tape. The other parts can be mounted the same way.

Listening is done ideally with the recommended crystal earpiece. Don't be tempted to use your Walkman earpieces; they are not the same and will not perform anything like as well. The receiver should work without an extra aerial, but one can be attached to terminal 1 if necessary. A long piece of wire mounted as high as possible is ideal. The *Audio-frequency Amplifier* project will enable you to use a loudspeaker with your receiver, using the signal from the jack socket. No circuit modifications will be needed!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R5, R6	10 kilohms (kΩ)
R2	560 ohms
R3	270 ohms
R4	10 kilohms (kΩ)

Capacitors

C1, C2, C3	100 nanofarads (nF)
	500 picofarads (pF)

Semiconductors

ZN414Z, BC184

Additional items

- 12-way 2 A terminal strip
- 22 metres of 28 SWG enamelled copper wire
- A few short pieces of coloured PVC-insulated wire
- Crystal earpiece
- 3.5 mm jack socket
- 1.5 V AA-size battery and box
- Toilet roll tube
- Double-sided sticky tape or selection of screws

Tools required

Small screwdriver, soldering iron.

2 An audio-frequency amplifier

Introduction

This simple amplifier can be built by anyone who is able to solder reasonably well. It doesn't require any setting up and, provided our instructions are followed exactly, will work very well. The circuit diagram is included for the benefit of our more advanced readers, but it is not needed in the construction process. Please practise your soldering before you start, and don't use a printed circuit board (PCB) until you are confident that your soldering is up to scratch.

The amplifier can be used with other projects; it will provide plenty of sound from the MW Radio or from the Morse Sounder projects. It will usually be built into other pieces of equipment, so a box is not supplied with the kit. There is no reason why it shouldn't be put into a box and used as a general-purpose amplifier to help test other projects.

The components

Before you start, you should check that you have *all* the components to hand. A list and some helpful hints are given below.

1. PCB. The plain side is the *component side* and the soldered side is the *track side*. **Figure 1** shows the track side full size. Make the PCB from the pattern given in Figure 1. Otherwise, build the circuit on a matrix board.
2. Three resistors. Locate the gold or silver band around the resistor, and turn the resistor until this band is to the right. There are three coloured bands at the left-hand end of the resistor. Find the resistor whose colours are YELLOW, VIOLET, RED, and look at the resistor colour code chart which you will find in Chapter 7. From this, you will see that YELLOW indicates the value 4, VIOLET the value 7, and RED the value 2. The first two colours represent real numbers, and the last value is the number of zeros (noughts) which go *after* the two numbers. So, the value is 47 with two zeros, i.e. 4700 ohms. In this way, the resistor coloured BROWN, GREY, BROWN has a value of 180 ohms, and the last one, BROWN, RED, GREEN, has a value of 1 200 000 ohms. The ohm (often written as the Greek letter omega (Ω)) is the unit of resistance. If you do not yet feel confident in identifying resistors by their colours, use the Resistor Colour Codes given in Chapter 7.
3. Four capacitors. The two small 'beads' are tantalum capacitors and will be marked 4.7 μF or 4μ7, with a '+' above one lead. A tubular capacitor with wires coming from each end should be marked 220 μF, with one end marked '+' or '–'. This is called an *axial* capacitor because the wires lie on the axis of the cylinder. This is in contrast to the final capacitor, where both wires emerge from the same end. This is a *radial* capacitor, and will be marked 47 μF. Again, one lead will be marked '+' or '–'. Capacitors marked like this are said to be *polarised*, and it is vital that these are placed on the PCB the right way round, so take notice of those signs!
4. Two diodes. These are tiny glass cylinders with a band around one end, and may be marked 1N4148; this is their type number. Like polarised capacitors, they *must* be put on the PCB the correct way round!

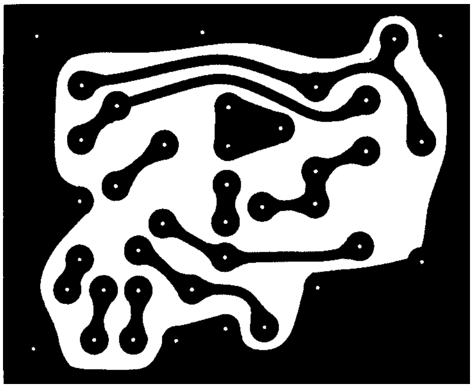

Figure 1 The toil pattern of the PCB – looking from the track side

5. Three transistors. One should be a BC548 (or a BC182), the other two should be BC558 (or BC212).
6. One volume control with internal switch.
7. One loudspeaker. This is quite fragile – don't let anything press against the cone.
8. One PP3 battery clip with red and black leads.

Putting it together

Lay the PCB on a flat, clean surface with the track side downwards. It is always useful to compare the layout with the circuit diagram, given here in **Figure 3**. Although you can't see it, the D-i-Y Radio sign should be at the top. Compare the hole positions with those shown in **Figure 2**. Bend the resistor wires at right angles to their bodies so that they fit cleanly into the holes in the PCB. Push each resistor towards the board so that it lies flat on the board. Then supporting each one, turn the board over and splay out the wires *just* enough to prevent the resistor falling out. Then, solder each wire to its pad on the PCB, and cut off the excess wire. When you have more confidence, you can cut of the excess wire *before* soldering; it often makes a tidier joint.

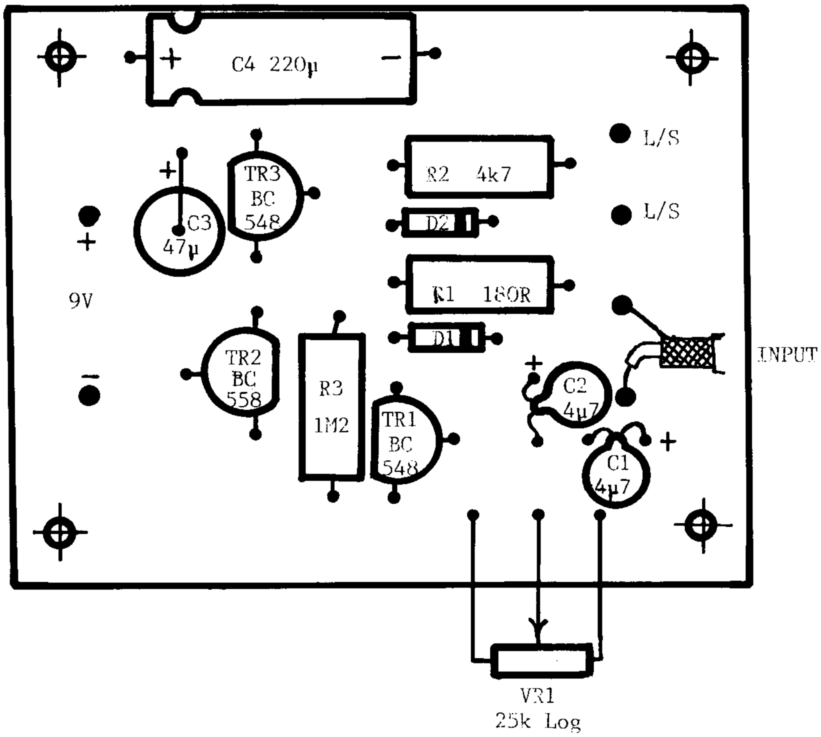

Figure 2 Positions of the components on the printed circuit board (PCB)

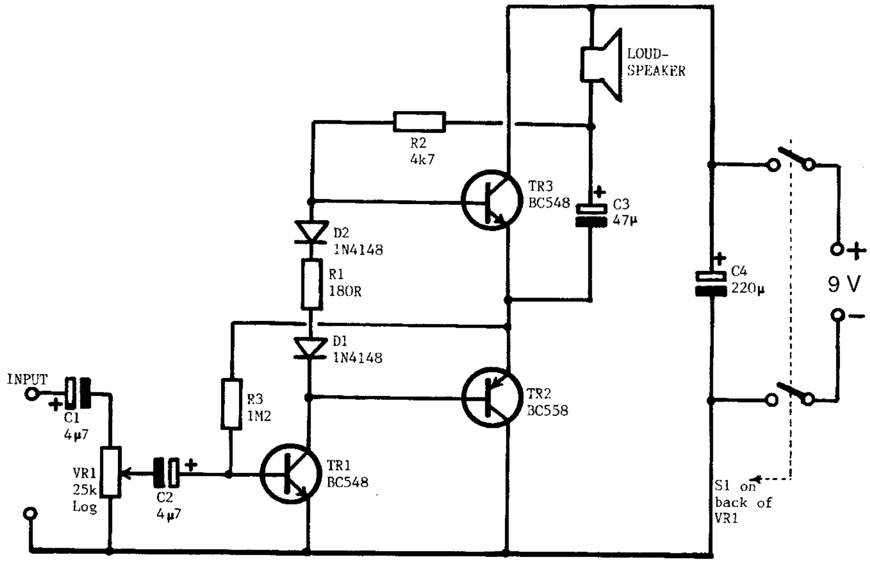
LOUD-
SPEAKER
R2
4k7
TR3
BC548
C3
47µ
D2
1N4148
R1
180R
D1
1N4148
C4
220µ
+
9 V
−
INPUT
C1
4µ7
R3
1M2
TR2
BC558
VR1
25k
Log
C2
4µ7
TR1
BC548
S1 on
back of
VR1

Figure 3 The amplifier's circuit diagram

Now fit the four capacitors. Each must be connected the right way round, so look at each component, match it up with the diagram of Figure 2, bend its wires carefully and repeat the soldering process you performed with the resistors, making sure that the components are close to the board and not up on stilts! Fit the two diodes the correct way round, and solder then as quickly as you can – they don't like to be fried!

Mount the transistors about 5 mm above the PCB. Make sure the correct transistors are in the correct places, and that the flats on the bodies match up with those shown in Figure 2.

Mount the volume control so that the spindle comes out from the front of the board. Use a piece of red insulated wire to the pad marked + on the PCB, and a black piece to the pad marked –, and solder these to the tags on the back of the control, as shown in **Figure 4**. Connect the two leads from the battery clip to the other tags on the switch; Figure 4 will help you. Finally, use two pieces of insulated wire about 100 mm long, twisted together, to connect the loudspeaker to the PCB.

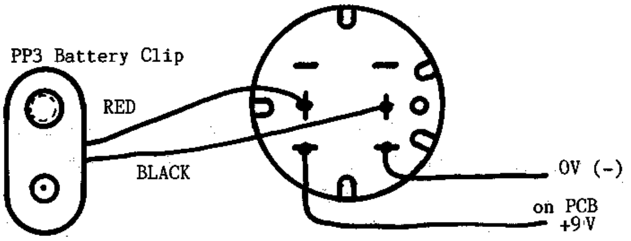
PP3 Battery Clip
RED
BLACK
0V (-)
on PCB
+9V

Figure 4 Connections to switch on back of VR1

Box clever!

If you wish to put the amplifier into a box, there is no problem; almost any box that is big enough will do. All that is needed is one hole big enough to accept the bush of the volume control; the PCB will be supported by the volume control. The prototype was not fitted into a box, but mounted on an odd piece of aluminium, bent into an L-shape and screwed on to a wooden base. The loudspeaker was mounted on the aluminium panel by two small pieces of aluminium with 3 mm holes drilled in them, which acted as clips around the edge of the speaker. Drill a few holes in the panel in the position of the speaker to let the sound get out!

Your input signal can be connected to the amplifier with two short pieces of wire, but if the connection needs to be long, use screened cable, with the braid connected as shown in Figure 2.

If you decide to use a different loudspeaker, make sure that its impedance (the resistance value marked on the back of the magnet) is at least 35 ohms. Anything lower may damage TR2 and TR3, and will certainly run down your battery very quickly. You will be surprised at the uses you can find for this little amplifier!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	180 ohms (Ω)
R2	4.7 kilohms (kΩ)
R3	1.2 megohms (MΩ)
VR1	25 kilohms (kΩ) log with DPST switch

Capacitors: all rated at 25 V minimum

C1, C2	4.7 microfarads (μF)
C3	47 microfarads (μF)
C4	220 microfarads (μF)

Semiconductors

TR1, TR3	BC548 npn
TR2	BC558 pnp
D1, D2	1N4148

Additional items

- PCB
- Speaker >35 ohms
- PP3 battery clip and battery

3 A medium-wave receiver using a ferrite-rod aerial

Introduction

This design came from the Norfolk Amateur Radio Club, and enables you to build a simple Amplitude Modulation (AM) receiver for frequencies between 600 kHz and 1600 kHz. It should take you around 2 hours to build, and can be used with Walkman-type earpieces. **Figure 1** shows the circuit diagram.

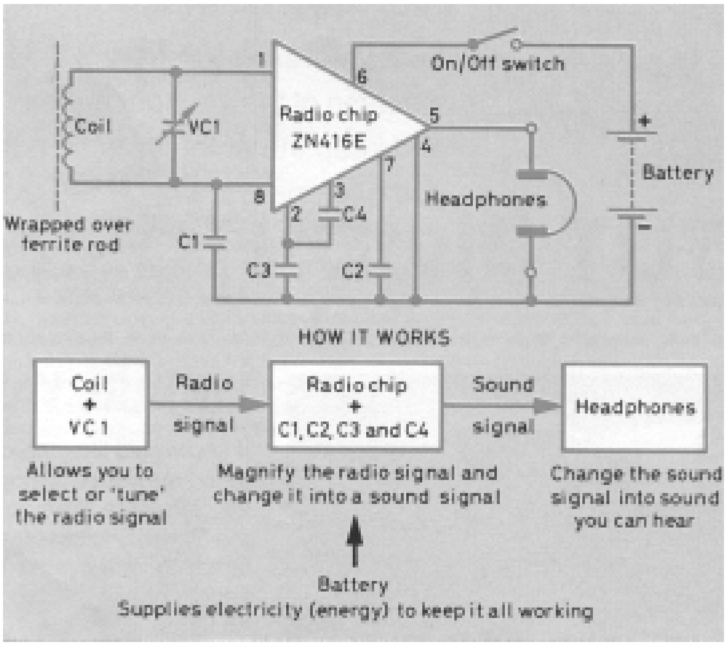
1
6
On/Off switch
Radio chip
ZN416E
5
4
7
Coil
VC1
+
Battery
−
Headphones
8
3
2
C4
Wrapped over ferrite rod
C1
C3
C2
HOW IT WORKS
Coil + VC 1
Radio signal
Radio chip + C1, C2, C3 and C4
Sound signal
Headphones
Allows you to select or 'tune' the radio signal
Magnify the radio signal and change it into a sound signal
Change the sound signal into sound you can hear
Battery
Supplies electricity (energy) to keep it all working

Figure 1 Circuit and block diagrams of the radio

Description

The whole circuit is built on a 50 mm by 50 mm printed circuit board (PCB) designed to fit on the inside of the lid of a plastic box, and is stuck there using sticky pads, the shaft of the variable capacitor going through a hole in

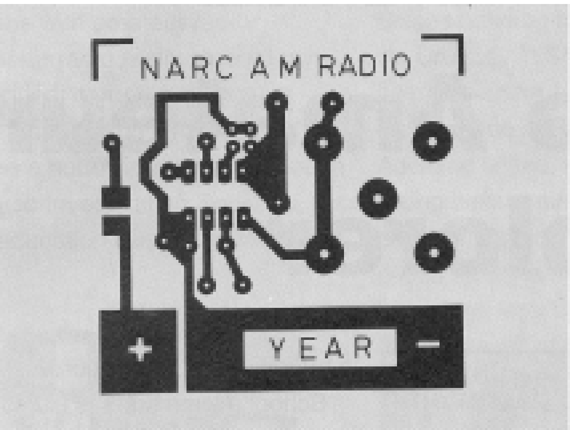

Figure 2a The PCB, solder side

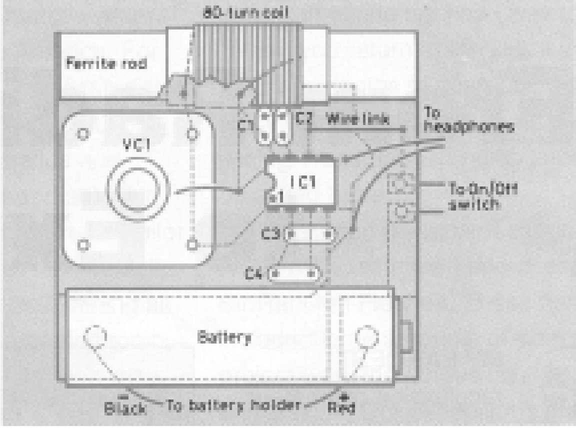

Figure 2b The PCB, component side

the lid. Only two pairs of leads are soldered to the board – one pair goes to the 1.5 V battery in its holder, and the other to the earphone socket. **Figures 2a** and **2b** show the printed circuit and the component layout double size for clarity. You are not obliged to build the circuit on a PCB.

Building it

1. **Check and identify components**. Tick the parts list.
2. **Carefully unwind the wire.** Use paper to make an insulating tube (called a 'former') around the centre of the ferrite rod and secure it with Sellotape. Now, close-wind *all* the wire (leave no gaps between adjacent turns) around the paper former. Secure the winding with more Sellotape, leaving 50 mm of wire free at each end for connection to the circuit. See **Figure 3a.**
3. **Solder in VC1.**
4. **Solder in the integrated circuit holder.** There is a notch in one end of the holder; this should face VC1. Solder also the wire link and the capacitors. Be careful to avoid solder 'bridges' between adjacent tracks on the PCB.
5. **Solder the battery leads.** These must be connected properly – the red battery lead to the + (positive) area and the black lead to the – (negative) area.
6. **Strip bare 1 cm of insulation from the ends of two wires**. Solder them between the PCB and the headphone socket (see **Figure 3b**). Use the end tabs on the socket. Using another pair of insulated wires connect the ON/OFF switch to the PCB tabs shown in Figure 2b.

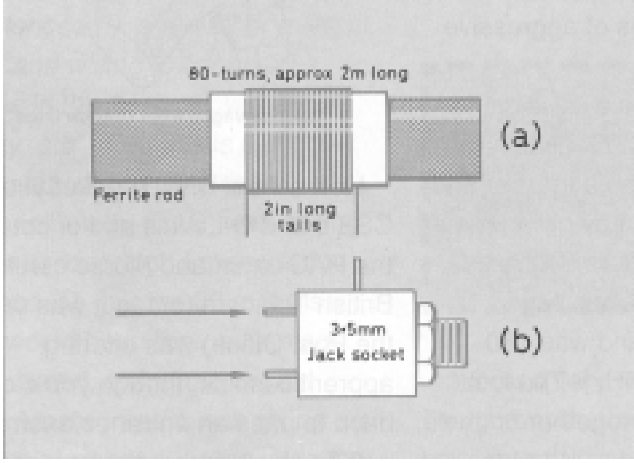

Figure 3 Details of coil and headphone socket

7. **Fix the elastic band.** This goes through the holes at the top of the PCB, with the ferrite rod being slipped through the two end loops. (Note: although the coating on the copper wire is designed to melt away during soldering, it is quite common for difficulty to be experienced in obtaining a good soldered joint; to be on the safe side, remove the coating *before* soldering (with a small piece of sandpaper).) Carefully place the wire ends of the coil through the PCB just above VC1, and solder on the track side.
8. **Fit IC1 into its holder.** This should be done carefully, making sure that *all* the pins are located above their respective clips *before* applying any pressure! Make sure also that the notch on the IC (as shown in Figure 2b) matches the notch in the holder, and faces VC1.
9. **Put battery in its holder.** Listen for some noise in the headphones as VC1 is rotated. Make sure the headphone plug is fully inserted into its socket.
10. **Fix the working board to the lid.** Use the sticky pads and apply *gentle* pressure. Fit the tuning knob, the ON/OFF switch and the earphone socket.
11. **Test again.** If all is still working, fit the lid screws and admire your completed radio!

Parts list

Capacitors

C1, C2	0.01 microfarad (μF)
C3, C4	0.1 microfarad (μF)
VC1	500 picofarads (pF)

Semiconductor

IC1	ZN416E

Additional items
- Plastic box (recommended size 76 × 64 × 50 mm internal)
- 8-pin DIL socket for IC1
- Printed circuit board
- Tuning knob for VC1
- Wire link for PCB
- 2 m of 30 SWG copper wire, self-fluxing
- Piece of paper 25 × 50 mm, to make the coil former
- Ferrite rod 70 mm long by 10 mm diameter, approximately
- Battery, AA size 1.5 V, with holder and attached wires
- Miniature earphone socket (3.5 mm stereo jack)
- ON/OFF switch (push-button SPST latched or slide switch)
- 4 off 100 mm insulated connecting wires, for jack socket and ON/OFF switch
- Pair Walkman-type earphones
- Elastic band, to attach ferrite rod to PCB
- 4 off sticky pads for securing PCB to box lid

Kits

Ready-made PCBs may be available from Alan J. Wright, G0KRU, Hewett School, Cecil Road, Norwich NR1 2PL.

4 A simple electronic organ

Introduction

This project has nothing to do with radio but, let's admit it, *any* electronics project is good experience! Why not build this little organ – it will keep the children amused at least! It uses the popular NE555 integrated circuit, which contains a circuit which will periodically switch the voltage on the output pin between the supply voltage and zero. Just how frequently this switching occurs depends upon the components external to the integrated circuit. If this switching occurs several hundred or thousand times a second, the change in voltage produced will generate a musical note when connected to a small loudspeaker. The circuit is shown in **Figure 1**.

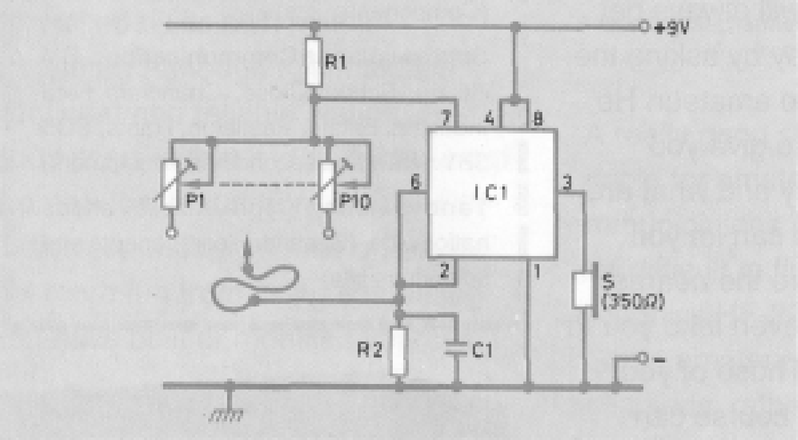
+9V
R1
7
4
8
P1
P10
IC1
6
3
2
1
S
(350Ω)
R2
C1
−

Figure 1 Circuit diagram

Putting it together

(a) Using a PCB. The job is very simple. The placement of components on the unsoldered side of the board is shown in **Figure 2** and the design on the copper track is illustrated in **Figure 3**. Put each component, in turn, on the board, making sure that it lies flat on the board with its tags or wires going cleanly through the holes provided for it; then, solder the wires to the board, cropping them before or after the soldering,

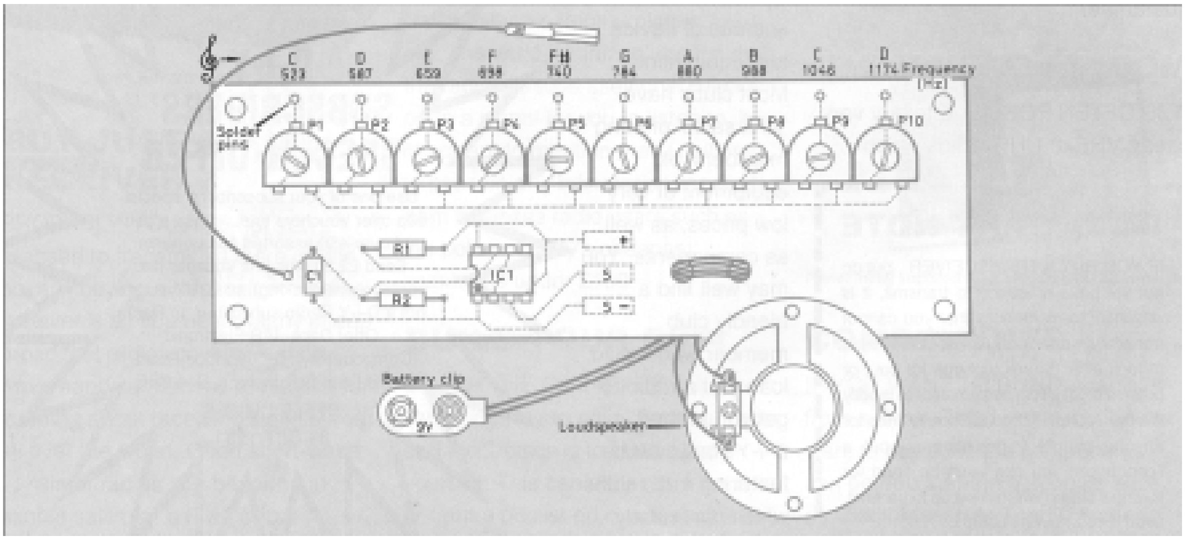
C 523
D 587
E 659
F 698
F# 740
G 784
A 880
B 988
C 1046
D 1174
Frequency (Hz)
Solder pins
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
R1
R2
IC1
C1
+
S
S −
Battery clip
9V
Loudspeaker

Figure 2 Position of components on the printed circuit board (PCB)

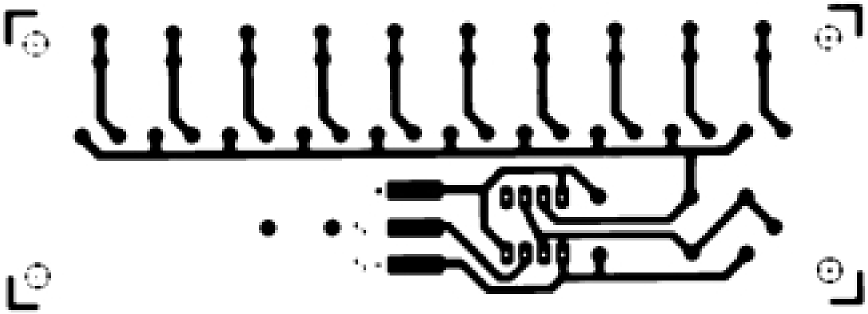

Figure 3 The connections

depending on your preference. If you choose to use a holder for your integrated circuit (highly recommended if your soldering is less than perfect), make sure that the end with a notch in it faces R1 and R2, as shown in Figure 2. Solder the two leads to the speaker to the tabs marked S (either way round), having looped them through the two holes to the right of the tabs in Figure 2. Looping them through the holes acts as a strain relief, ensuring that the soldered joints are not subjected to pulling and bending as you move the wires about. Do the same with the battery leads, the red lead going to the + tab and the negative lead to the – tab (which also has one speaker lead already attached to it). **Figure 4** shows this in detail. Treat the loudspeaker with care – the cone is quite fragile and must not be touched.

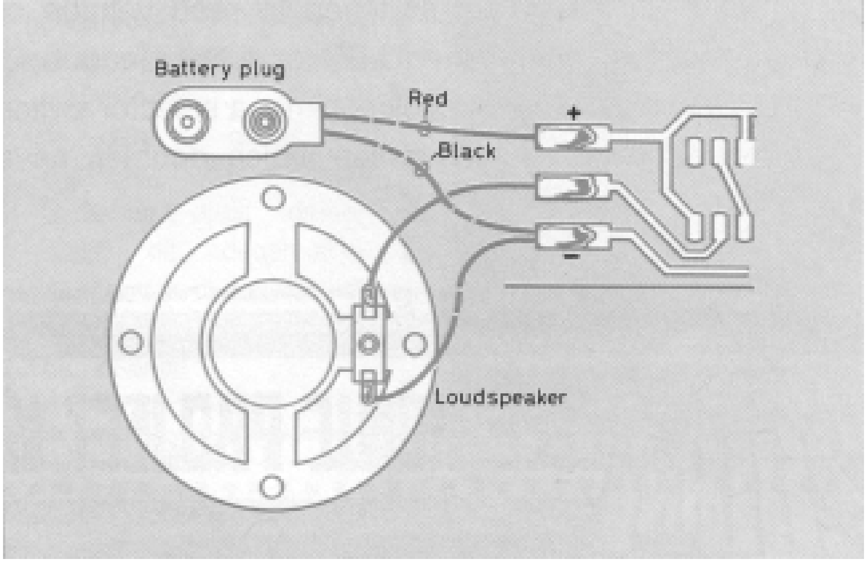

Figure 4 Battery plug and loudspeaker connections

(b) Without a PCB. This is more difficult, and you may need to enlist some help. Using some matrix board (such as Veroboard) is probably the best way of replacing the PCB. You could arrange your circuit in exactly the same way as in the PCB in Figure 3, using wires to replace the copper track.

A simple ‘keyboard’

The keyboard is a row of solder pins along the rear edge of the PCB, one for each note covering the range shown in Figure 2. A flying lead with a small spade on it is provided to touch any of the pins in turn, producing any one of ten different notes.

Testing

Check first that each component is in the correct place. When inserting the NE555 chip, first make sure that the end carrying the notch lies over the end of the holder with the notch; then, make sure each pin of the chip lies directly above the hole into which it fits, before pressing *gently* to insert the chip into the socket. Make sure the battery connections are correct, and insert the battery into the clip. Nothing should happen, except for a click from the loudspeaker; touching the spade on any of the pins should produce a coarse note from the speaker. If nothing happens, check everything again; don’t *assume* that wires go where you think they go!

After you get the first note, all the others should work, too, but they will sound off-tune at first. The organ needs tuning up by adjusting the 10 preset variable resistors P1 to P10. The approximate frequency to which each note should be tuned is given in Figure 2; if you can beg, borrow or steal a frequency counter, setting up is easy. If you have a piano, the organ can be tuned by comparison of the notes with those on the piano. The frequencies are given in Hertz (abbreviation Hz), and represent the number of times the IC switches on and off every second. If the sound coming from the loudspeaker is too loud or very distorted, then try putting an 330 Ω resistor (colour code orange, orange, brown) in series with the loudspeaker. This is done by taking the resistor and cutting its leads to about 5 mm; then, disconnect one speaker lead from the tab on the PCB (it doesn’t matter which). Solder one end of the resistor to the vacated speaker tab, and the free speaker lead to the other end of the resistor. This will limit the volume of sound from the speaker, and lengthen the life of your battery. If it is still too loud, try a resistor of a larger value, or use a smaller resistor to make it louder.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	2.7 kilohms (kΩ)
R2	1 Megohm (MΩ)
P1–P5	Preset resistor 100 kilohms (kΩ)
P6, P7	Preset resistor 50 kilohms (kΩ)
P8	Preset resistor 25 kilohms (kΩ)
P9, P10	Preset resistor 10 kilohms (kΩ)

Capacitor

C1	100 nanofarads (nF) or 0.1 microfarad (μF)

Integrated circuit

IC1	NE555 timer chip

Additional items

S	Loudspeaker >60 ohms (Maplin)
	1 off battery clip (for PP3 battery)
	1 off spade terminal
	12 off solder pins 'Veropins'
	3 off 10 cm lengths of 'hook-up' wire

(This article is based on projects originally designed by Radio Scouting, Netherlands.)

5 Experiments with the NE555 timer

Introduction

Several of the projects in this book use the NE555 timer, an integrated circuit which is at the heart of many circuits whose processes are determined by time intervals. **Figure 1** shows the circuit diagram of an audio oscillator using the 555. The timing voltages (governing the frequency of oscillation) are produced by R1, R2 and C2; a voltage appears at pin 3 which 'switches' at this frequency between zero and a voltage close to the supply voltage, which in this case can be anywhere between 6 V and 14 V. The output current, when applied through R3 to a small loudspeaker, produces an audible tone, provided that there is a DC path between the two test leads.

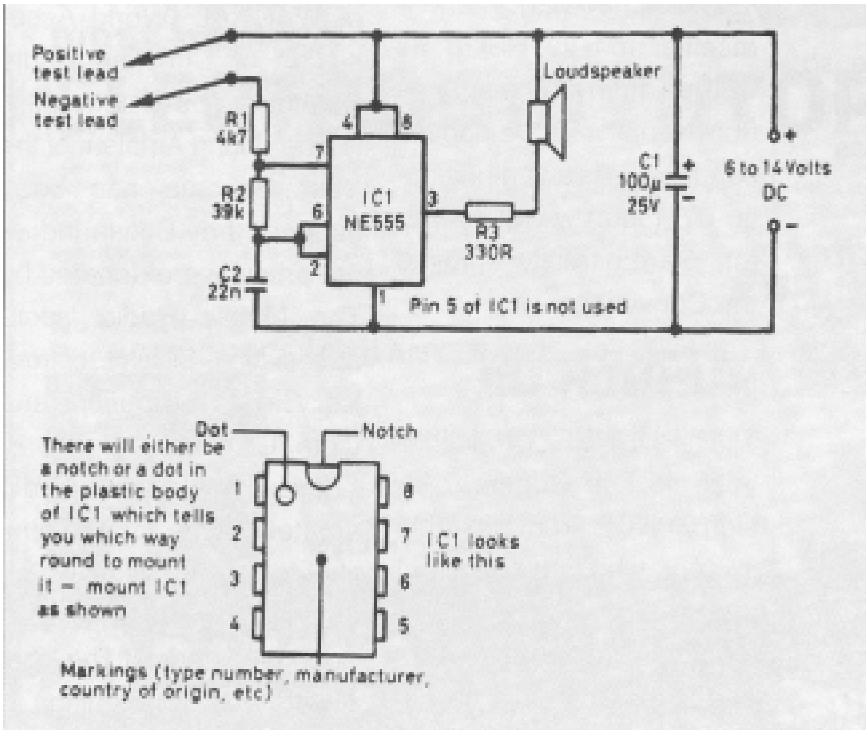
Positive test lead
Negative test lead
R1 4k7
R2 39k
C2 22n
4
8
7
6
2
1
3
IC1 NE555
R3 330R
Loudspeaker
C1 100μ 25V
+
−
6 to 14 Volts DC
Pin 5 of IC1 is not used
Dot
Notch
There will either be a notch or a dot in the plastic body of IC1 which tells you which way round to mount it – mount IC1 as shown
1
2
3
4
8
7
6
5
IC1 looks like this
Markings (type number, manufacturer, country of origin, etc)

Figure 1 The circuit diagram

Construction

The simplest way to mount the components is on a piece of matrix board (Veroboard), available from any of the good suppliers. The prototype of this circuit used the type of board with copper strips along the underside; these strips are used like the copper tracks on a PCB, to join components together. Firstly, cut the four strips between the positions of the pins of the IC socket, as shown in **Figure 2**. You can buy a tool for this purpose, but a small twist drill (about 3 mm diameter) is just as good. Turn it between your fingers – if you use a drill you will end up with holes right through the board! Then solder in the IC socket (with the notch in the position shown), followed by the four links made with single-conductor insulated wire. Put in each component as shown, ensuring that C1 (an *electrolytic* or *polarised* capacitor) is connected correctly. When all the components have been soldered in, take the 555 chip and lay it on its socket, with its own notch lying above that of the holder. Then, making sure that each pin lies *directly* above its corresponding socket, press down *gently* on the chip, with the board supported on a flat surface.

Testing

Connect the circuit to a battery or small power supply, ensuring that the positive and negative leads are the right way round. Always use red and black leads here, then you are less likely to get it wrong! Switch on. Nothing

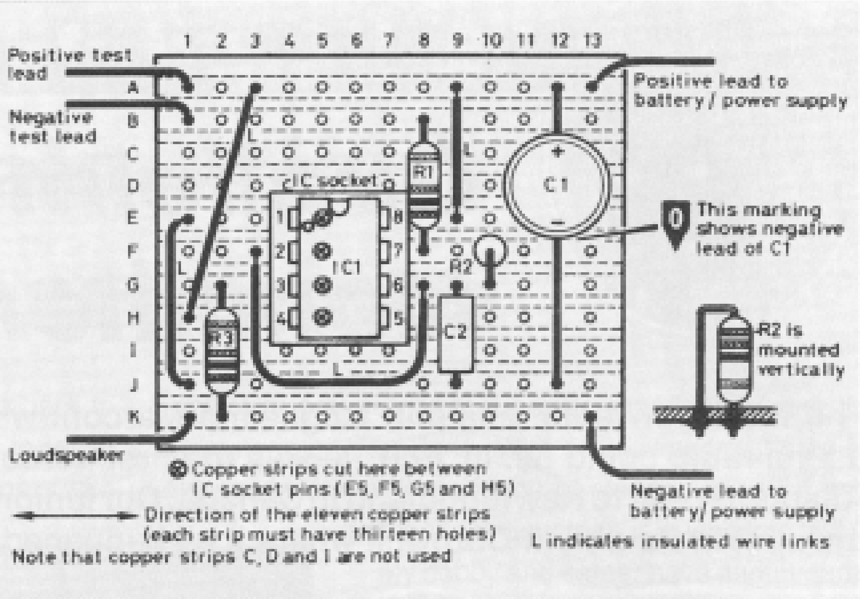
1 2 3 4 5 6 7 8 9 10 11 12 13
A
B
C
D
E
F
G
H
I
J
K
Positive test lead
Negative test lead
Loudspeaker
Positive lead to battery / power supply
IC socket
IC1
1
2
3
4
5
6
7
8
R1
R2
R3
C1
+
–
C2
L
This marking shows negative lead of C1
R2 is mounted vertically
Negative lead to battery/ power supply
Copper strips cut here between IC socket pins (E5, F5, G5 and H5)
Direction of the eleven copper strips (each strip must have thirteen holes)
Note that copper strips C, D and I are not used
L indicates insulated wire links

Figure 2 Veroboard layout. If you can read a circuit diagram, the project can be built using other methods

should happen until you short together the two test leads, when there should be a note from the loudspeaker. If this doesn't happen, switch off, disconnect the circuit and check your wiring and soldering. Is it exactly like Figure 2? Are the soldered joints round and shiny? If any are dull, then 'sweat' them briefly with a hot soldering iron until the solder runs, remove the iron, and check that they are as shiny as the rest. Check that there are no solder 'bridges' between adjacent tracks by holding your board up to a strong light. Then, reconnect, switch on and touch the test leads together. All should now work!

Uses of your circuit

1. As a Morse practice oscillator. Simply connect the two test leads to your key and, each time the key is pressed, you should hear a note from the speaker. The frequency of the note may be altered by putting a resistor in series with the key. To do this, remove one test lead from the key and select a resistor; connect one end of the resistor to the free test lead and the other to the empty terminal on your key. Selecting the value of resistor that you need will be a useful experiment in itself.
2. As a continuity tester. You can check fuses and lamp bulbs by connecting them across your test leads. If the speaker remains silent, the fuse or bulb has blown.
3. To indicate changes of resistance. Hold the ends of the test leads in each hand; you should hear a low note, because of the high resistance of your body. Squeeze the ends harder, and the frequency of the note should rise, because you are now making better contact. Repeat this with damp hands and the frequencies will be higher still.
4. As a thermometer. Connect the test leads to a thermistor (a device whose resistance changes with temperature) and warm it with a hair-dryer, or even in your hands, and you will hear the pitch changing with the temperature of the thermistor. A suitable 'bead' thermistor is available from Maplin (order code FX21).
5. As a diode tester. Use any diode, and connect the negative test lead to the end of the diode marked with the ring. This is the cathode of the diode. The other end, the anode, should be connected to the positive test lead, and a note should be heard from the speaker. This does not necessarily mean that the diode is working – yet. Reverse the connections and *nothing* should be heard. If this is the case, the diode *is* working.
6. As a light meter. Use a photoconductive cell (a device whose resistance changes with light intensity) connected between the test leads. A note should be heard. Shading the device with your hand will increase its resistance and the note should decrease in frequency. A suitable device is the ORP12 cell from Maplin (order code HB10).

There are many more applications. **Do not connect the test leads to other circuits that are switched on.** Your circuit, or the circuit you are connecting it to could be damaged. Think of a *passive* device or circuit (i.e. one *not* requiring a power supply or battery) where changes of resistance occur, and you have found another application!

Parts list

Resistors: all 0.25 watt carbon film types

R1	4.7 kilohms (kΩ) – yellow, violet, red
R2	39 kilohms (kΩ) – orange, white, orange
R3	330 ohms (Ω) – orange, orange, brown

Capacitors

C1	100 microfarads (μF) 25 V radial electrolytic
C2	22 nanofarads (nF) or 0.022 microfarad (μF) polyester type with 10 mm lead spacing

Integrated circuit

IC1	NE555

Additional items
- Miniature loudspeaker (preferably 35, 40 or 80 ohm)
- 8-pin DIL socket for IC1
- 0.1 inch Veroboard (‘stripboard’), size 11 strips by 13 holes
- PVC-covered stranded wire for test leads, loudspeaker and battery connections
- PVC-covered solid wire for links on the board
- A power source of between 6 V and 14 V, such as a 9 V battery (PP3)

Component sources

Cirkit
Tandy – many high street shops

6 A simple metronome

Introduction

A metronome is a device used by musicians to indicate the tempo of a piece of music. Until electronics came on the scene, this 'beating of time' was achieved in much the same way as a clock keeps time, i.e. with a pendulum device, the clicking of the escapement indicating the beats of the music.

Those of you who have already built the Morse Key and Buzzer from the designs in this book, will recognise the circuit of this metronome – it is exactly the same as was used to produce the note of the buzzer. This circuit is shown in **Figure 1**.

The circuit

Three components determine the speed at which the circuit oscillates – the speaker (LS), the resistors (VR1 + R1) and the capacitor (C1). VR1 is a variable resistor, so that the speed at which the oscillator operates can be varied. Compared with the component values of the Morse Buzzer (which operated at around 800 Hz), these components now give an oscillation frequency of around 1.25 Hz, which is far too low to be heard as a note. What we *do* hear, however, is a series of clicks, as the voltage across the speaker changes quickly from 0 to 9 V and back again.

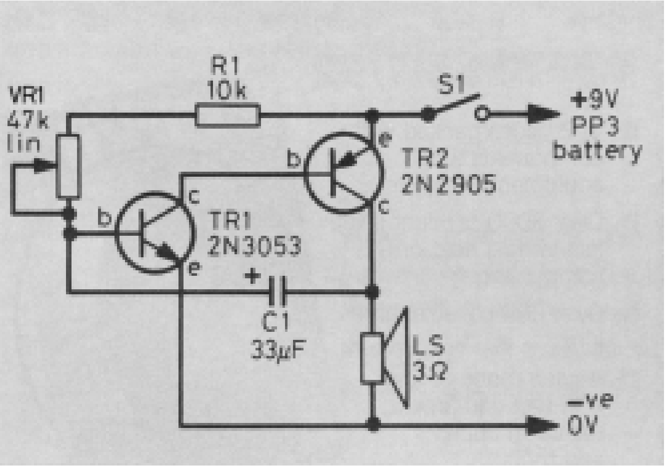

Figure 1 The metronome circuit is rather like the Morse oscillator

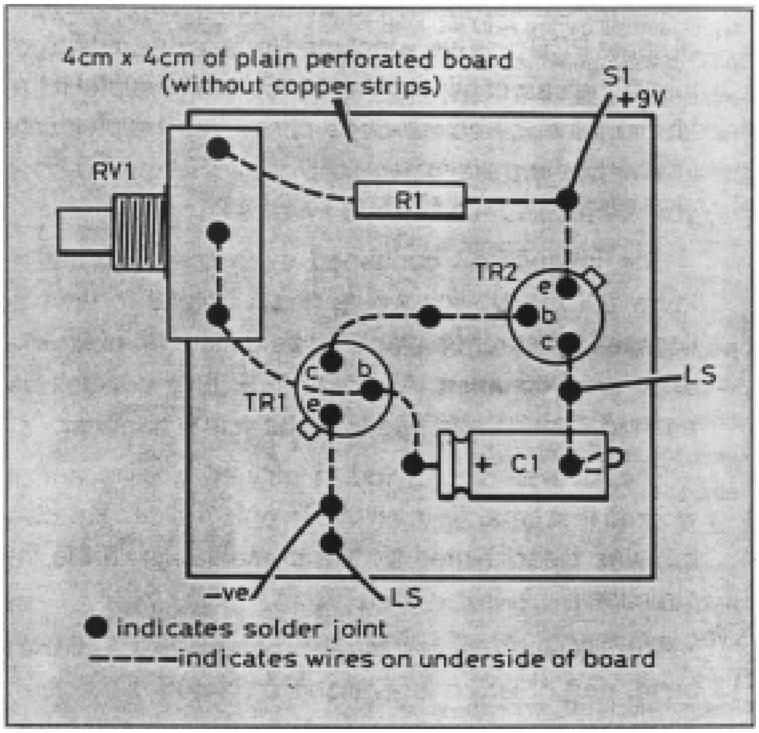
4cm x 4cm of plain perforated board
(without copper strips)
S1
+9V
RV1
R1
TR2
e
b
c
LS
c
b
TR1
e
+ C1
–ve
LS
● indicates solder joint
– – – –indicates wires on underside of board

Figure 2 The component wires are pushed through holes in the circuit board and joined together underneath

Variation of speed could be achieved by varying resistance or capacitance. However, as you may already know, variable capacitors have values in the picofarad range, not the tens of microfarads used here, so it is very simple to employ a variable resistor (potentiometer) to control the oscillator. You could use a multi-way switch to switch in one of several capacitors, as well as having the variable resistor, but this was found to be an unnecessary complication. This design operates between about 100 clicks per minute and 200 clicks per minute.

Making the prototype

A single piece of plain matrix board (no copper strips) measuring about 40 × 40 mm is sufficient to hold all the components except the potentiometer and switch (see later). The case can be plastic or aluminium, and one measuring 65 × 100 × 50 mm is about right. Make sure there are holes in the case beside the speaker cone to let the sound out, and larger holes for the potentiometer and switch. If a potentiometer is used with a combined ON/OFF switch, then the extra hole for the switch is not necessary! It is advisable to construct the circuit *before* putting it in the box, so that it can be tested to ensure that everything is working. If it is, then you can exercise your ingenuity in mounting the speaker, battery and board inside the box. A final test can be made before starting the calibration process.

Calibration

There is no ‘easy’ way to do this. The frequencies involved are too low to be measured with the average frequency counter, so you will need to resort to using a stopwatch and counting the number of clicks per minute.

Parts list

Resistors: 0.25 watt, 5% tolerance

R1	10 kilohms (kΩ)	
VR1	47 kilohms (kΩ) linear potentiometer	

Capacitor

C1	33 microfarads (μF) electrolytic	

Transistors

TR1	2N3053	npn
TR2	2N2905	pnp

Additional items

S1	SPST	ON/OFF switch
LS	3 ohms (Ω)	loudspeaker
	Knob with pointer for VR1	
	PP3 battery and connector	
	Aluminium case, 65 × 100 × 50 mm	
	Matrix board (plain), 40 × 40 mm	

7 What is a resistor?

Introduction

Materials that carry electricity easily are called **conductors**. They include all metals and salt water, for example. We use wire as a conductor, and the ease with which it passes an electric current depends upon the material, its thickness and its length. Silver (symbol Ag), gold (Au), copper (Cu) and aluminium (Al) are the best metallic conductors. Most wires are made of copper, although the best conductor, weight for weight, is aluminium.

Materials that *don't* carry current (or, at least, do so very badly) are called **insulators**, and they include dry wood, rubber, plastic and glass among their number. Wires are often coated with a layer of insulator to prevent adjacent wires touching and causing an accident.

Resistors

If there wasn't such a thing as resistance, the subject of electronics wouldn't exist; only infinite currents would flow and voltages wouldn't exist either! We need to reduce the flow of current if we are to make current do something useful for us. Components that resist the flow of current are called **resistors**, and they are said to have a **resistance** which is measured in ohms (Ω), named after Georg Ohm, who formulated the law (also named after him) by which the voltage and current through a conductor are related. His law gave rise to the formula *everyone* remembers:

$$I = \frac{V}{R},$$

where I is the current flowing, measured in amps,
V is the voltage across the conductor, and
R is the resistance of the conductor, measured in ohms.

From this equation, you can see that, for a constant value of voltage, V, if the resistance goes up, the current will go down, and vice versa. The circuit symbols for resistors are shown in **Figure 1**. You will find the upper symbol in older magazines; it is still preferred by many engineers. The lower symbol is the prevalent standard symbol.

Resistors are made in several ways, the cheapest using carbon; another type is usually made from a ceramic cylinder (used only as a support) on which

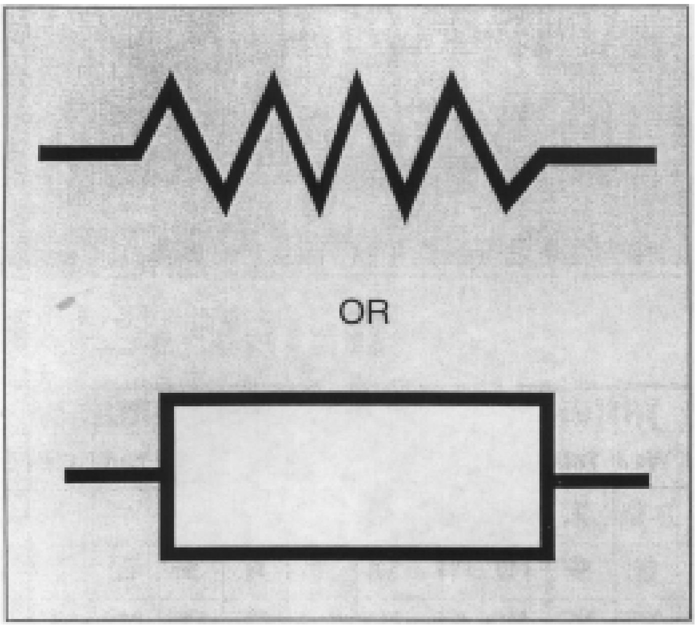

Figure 1 Circuit symbols for resistors

is placed a very thin film of metal – the thinner the film, the greater the resistance. All resistors are coated with a thin film of insulation, for the same reason we discussed earlier.

The colour code

Each resistor has coloured bands on it which enable us to see what value of resistance it has. There are normally three (but sometimes four) at one end, and a single one at the other (see **Figure 2**). The colours indicate figures, according to the list below.

Colour	*Value*	*Colour*	*Value*
Black	0	Green	5
Brown	1	Blue	6
Red	2	Violet	7
Orange	3	Grey	8
Yellow	4	White	9

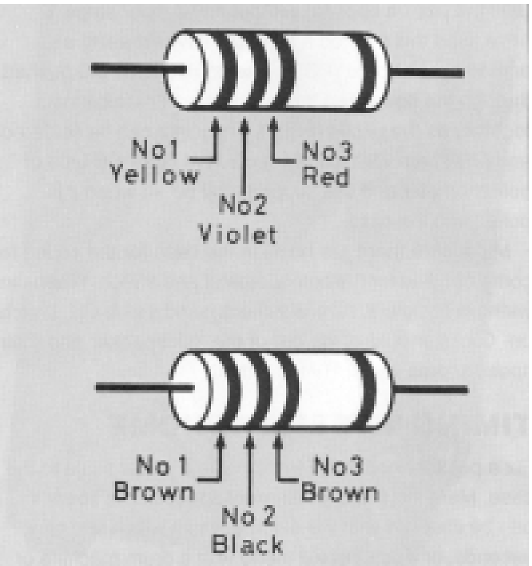

Figure 2 Some examples of resistor colour codes; top 4700 Ω (4.7 kΩ) and bottom 100 Ω

Using the colour codes is easy, once you see the logic behind it. Hold the resistor so that the single band is towards the right. The three colours on the left are read in the normal order from left to right. The first two bands *always* indicate numbers; the third band gives the number of zeros to add to the right of these two numbers. So, looking at the top resistor in Figure 2, yellow, violet, red means 4, 7, and two zeros, giving 4700 ohms! Looking at the lower resistor, brown, black, brown means 1, 0, and one zero, giving 100 ohms.

Remembering the order of the colours may be difficult at first. The colours from red to violet are the colours of the rainbow, in order, so if you know those, you're almost there! Around those colours are black and brown below the red, and grey and white above the violet, which you can imagine as getting brighter from black to white (well, almost!). It won't be long before you don't need to remember them at all.

The isolated band on the right-hand side is not part of the resistor's value; it indicates its *tolerance*, i.e. how close it might be to the indicated value. A brown band indicates ±1%, a red band ±2%, a gold band ±5% and a silver band ±10%. For example, a resistor marked as being 100 ohms with a ±5% tolerance will have an *actual* value somewhere between 95 ohms and 105 ohms.

8 Waves – Part 1

Introduction

Waves are responsible for most of the processes in life where energy is transferred from one place to another. Heat and light energy from the sun, for example, come to us as *electromagnetic waves*. Sound travels through the air as a wave; it is not the same sort of wave as light or heat, but it obeys many of the same properties. Damage is caused to coastal margins by the waves of the sea – again, another type of wave, but still obeying many of the same properties. Radio waves are of the same type as heat and light waves, as are gamma rays, X-rays, ultra-violet waves and infra-red waves. So, once we begin to understand what radio waves do, we are also learning about a huge chunk of physics at the same time! All these waves are part of the *electromagnetic spectrum*. The word 'spectrum' simply means a 'range', so what we have is a range of electromagnetic waves – that's all!

Sensing things

Light waves are invisible, but our eyes can detect the *effect* they have on different materials because the waves produce an effect on the retina of the human eye which the brain can interpret. We cannot see heat waves either, but we can feel the *effect* they have on our skin. Gamma rays and X-rays are also invisible, but their detrimental *effects* on human tissue are well known. It is not surprising, then, that we cannot see radio waves. We cannot sense them, either, until we produce a device upon which they have an *effect*. That device is a **radio receiver**; it is able to process certain characteristics of radio waves, and make these characteristics audible by generating sound waves from the loudspeaker or headphones. Other characteristics of the same waves may be turned into light as a TV picture on a cathode-ray tube, or as a fax image on a sheet of paper.

Visible waves

Let's start our description with some waves that we can actually see! When a small stone is thrown into a pond, we see circular water waves *radiating* from the point where the stone fell into the water, as **Figure 1** shows. (Notice that we use the word *radiating*, even with water waves; it is not a radio term, but one which describes any motion where the *radius* of a circle is increasing. In this case, it is the radius of the circular waves which is

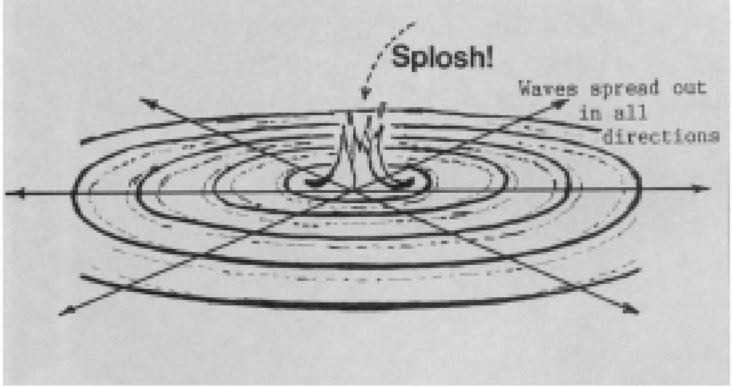
Splosh!
Waves spread out in all directions

Figure 1 The waves move out from the point where the stone landed

increasing.) If you were to watch the water waves down at the water level, perpendicular to the direction in which they are travelling, you would see something like the illustration in **Figure 2**.

The horizontal line represents the water level before the wave started, and the vertical line represents the direction in which the water is displaced at any instant.

All waves are described in the same way

'Freezing' the motion of the water in this way allows us to define two very important characteristics of a wave, characteristics which we talk about every day – *wavelength* and *amplitude*. The wavelength of a wave is simply the distance (measured in metres) from any point on one wave to the same point on the adjacent wave. Look at the diagram and you will see what is meant by 'the same point on the adjacent wave'. The amplitude of a wave

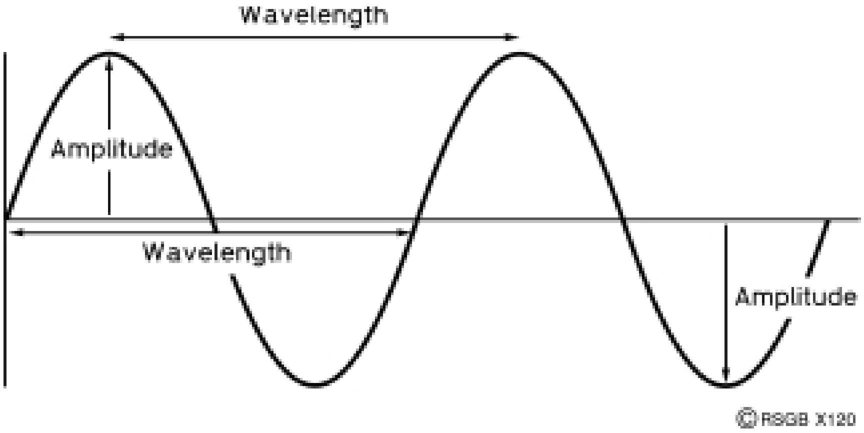

Figure 2 A water wave, viewed in cross-section

is always measured from the centre (undisturbed) position of a wave to the peak (or the trough) of the wave. Both these positions are shown, the arrow indicating that the measurement is taken *from* the centre *to* the peak or trough. The amplitude of a wave is defined as the maximum displacement of the wave from the centre position – the direction (up or down) of that displacement does not matter. Waves of greater amplitude carry more energy with them.

If we now 'unfreeze' the wave, we will see it travel from left to right (or right to left, depending on where we are looking). The speed at which it moves is called its *velocity*, and is measured in metres per second.

Another useful word is *propagate*; it means *travel*. We talk about radio waves propagating from a transmitter to a receiver. This velocity of propagation (for electromagnetic waves) is very fast indeed – they will cover 300 million metres in one second. This is virtually incomprehensible, so think of a radio wave travelling around the earth – it can travel 7½ times round the earth in one second! We use the symbol *c* for the velocity of radio waves (which is the same as the velocity of light, of course – all electromagnetic waves travel at this speed through air and space).

The last thing we need to know about the wave is its *frequency*. Imagine a cork floating on the water in the path of the wave; it will bob up and down. If we were able to count the number of times it went through its highest position in one second, then that number would be its frequency. Any periodic motion like this is said to go through one *cycle* each time one complete wave passes a point (in this case, our cork). We are thus counting the number of cycles per second of the cork's motion. The unit of frequency is thus 'cycles per second'; this unit is now named after Hertz, a radio pioneer, and is abbreviated to Hz.

Our description of the wave is now quite simple – we need only four quantities:

(a) **Frequency** symbol f – unit, hertz (Hz)
(b) **Wavelength** symbol λ (Greek letter lambda, pronounced 'lamb-da') – unit, metre (m)
(c) **Amplitude** symbol a – unit depends on application
(d) **Velocity** symbol c – unit, metres per second (m/s or ms^{-1})

The basic formula

Whatever may happen to a wave while it travels through different media (vacuum, air, brick, wood, etc.), one thing and **only** one thing remains constant – its frequency. Its wavelength, amplitude *and* velocity may change, but its frequency never does. Three of the four characteristics already identified are connected by the simple relationship

$$c = f \times \lambda.$$

Remember that *c* is constant if the wave travels in air or in a vacuum. This means that waves having higher frequencies (*f* large) must have smaller wavelengths (λ small) and vice versa. You can imagine frequency and wavelengths being on opposite ends of a see-saw!

Divisions of units

Because the frequencies of radio waves are so high (despite them having the lowest frequencies in the electromagnetic spectrum!) we have a problem with writing them down. Do you write in your log book that you have just heard a station on 14 100 000 Hz? Of course not, you write it as 14.1 MHz, knowing that the prefix mega (M) means 'one million'. The prefixes which you need to know (when applied to frequency) are:

kHz	kilohertz	meaning	1000 Hz
MHz	megahertz	meaning	1 000 000 Hz
GHz	gigahertz	meaning	1 000 000 000 Hz.

Notice that the 'k' in kilohertz is a lower case letter. It is **incorrect** to write it as an upper case letter. 'K' is a computer-related prefix meaning **not** 1 000 but 1 024!

When we come on to discuss heat and light waves, we will use wavelengths rather than frequencies, because of the see-saw effect – as the frequencies get larger and larger, the wavelengths get smaller, and hence are numbers which are more manageable, both to talk about and to write down!

Bands

Gamma rays, X-rays, ultra-violet waves, light waves, infra-red waves are all part of the electromagnetic spectrum, but we divide them up because they have different properties. This is why we divide up our radio frequencies into different bands. The radio waves of top-band signals (around 2 MHz) have completely different properties compared with those in the 20 metre band, so we are dividing up the *radio spectrum* in the same way – by property.

9 A beat-frequency oscillator

Introduction

Many readers will know that, although they have a short-wave radio which covers at least one of the amateur bands (e.g. 7 MHz or 14 MHz), they are unable to listen to SSB or Morse signals. This is because the receiver lacks a Beat-Frequency Oscillator (BFO). We need the ‘carrier’ frequency of a BFO to replace the carrier that has been removed from the signal at the transmitter. When listening to Morse signals, the BFO signal ‘beats’ with the incoming signal to produce a note in the loudspeaker. If you are a musician, you will be familiar with the method of using ‘beats’ to tune one musical instrument from another; in the BFO, the beat frequency produced is the tone signal you hear.

In the more complex amateur radio receiver, a BFO is incorporated as part of the whole system. In our model, it is an external circuit that sits alongside your radio. The circuit diagram is shown in **Figure 1**.

Construction

Built on a small piece of matrix board about 80 × 50 mm, the circuit can be fitted inside a small plastic box. For once, we *don’t* want to screen the circuit to prevent it interfering with other equipment; we *want* it to interact

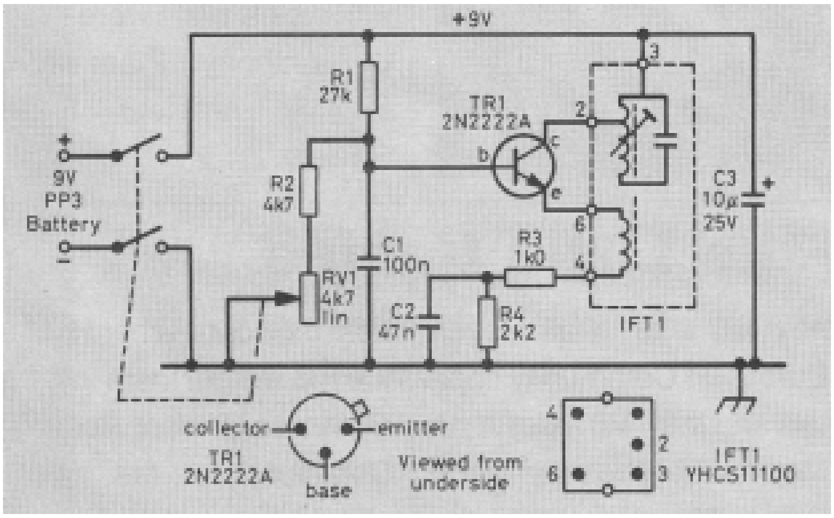
+9V
3
R1
27k
TR1
2N2222A
2
b
c
e
C3
10µ
25V
+
9V
PP3
Battery
R2
4k7
6
R3
1k0
4
C1
100n
RV1
4k7
lin
C2
47n
R4
2k2
IFT1
collector
emitter
TR1
2N2222A
base
Viewed from
underside
4
2
6
3
IFT1
YHCS11100

Figure 1 Circuit diagram of the BFO

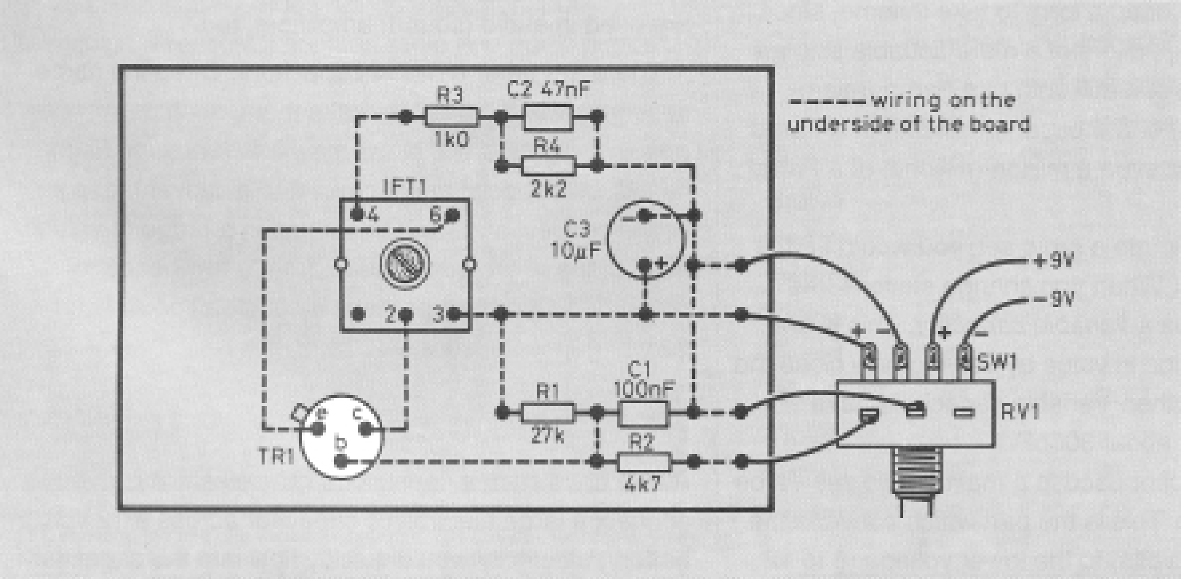

Figure 2 Matrix board layout shown from the component side. Adjust IFT1 carefully for the best results

with our receiver! This is why we use a plastic box. Maplin Electronic Components supply a suitable box, complete with the matrix board to fit inside (order code YU46).

Look at **Figure 2** carefully before you start to build the circuit, so that you can position the components correctly. Firstly, inspect the intermediate frequency (IF) transformer, IFT1, and remove its metal screening can very carefully. Again, this is to allow some signal to escape from our circuit and enter our radio. Having done this, solder the components, using the matrix board as a support. Underneath the board, the components are linked by single-conductor, insulated wire. Take particular care with the polarity of the electrolytic capacitor, C3, and the connections to the transistor, TR1 (see Figure 1).

The variable resistor, VR1, has a switch mounted behind the control itself, and the insulated leads to it from the battery should be about 10 cm long. Connect these before fitting VR1 into the case, so that the BFO can be calibrated (adjusted) correctly.

Calibration

After a final check that all the components have been fitted and soldered correctly, connect the battery, switch on, and hold the transistor between your fingers, to check that it is not getting hot. Place the circuit close to your receiver, and set RV1 to mid-position. Tune your receiver to find an amateur SSB transmission; the frequencies listed below will help you in knowing where to look. It may sound very strange, but don't worry. *Slowly* turn the core of IFT1 with a small, non-metallic screwdriver or with the correct 'trimming tool'. The core into which the blade fits is very fragile, so attempt this process with care. When the speech sounds as natural as you can get it, leave the core at this position, and use RV1 to make the speech sound natural.

Using the BFO

For best results, you may have to move the BFO nearer or further away from your radio. At the lower end of most bands (for instance just above 7.000 MHz or 14.000 MHz) you should be able to resolve Morse code (CW) signals. If you find that the BFO signal is a little weak, solder a 15 cm length of insulated wire to pin 2 of IFT1, and place it alongside your radio. This should improve signal intelligibility. When you are happy with the performance, switch off, drill a 10.5 mm hole in the box and fit RV1, followed by the matrix board assembly. Screw the base to the box, fit the knob, and the BFO is complete!

Where to listen

Band	**Frequencies (MHz)**
15 m	21.000–21.450
17 m	18.068–18.168
20 m	14.000–14.350
30 m	10.100–10.150
40 m	7.000–7.100
80 m	3.500–3.800

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	27 kilohms (kΩ)
R2	4.7 kilohms (kΩ)
R3	1 kilohm (kΩ)
R4	2.2 kilohms (kΩ)
VR1	4.7 kilohms (kΩ), linear, with DPST switch

Capacitors

C1	100 nanofarads (nF) or 0.1 microfarad (μF), ceramic
C2	47 nanofarads (nF) or 0.047 microfarad (μF), ceramic
C3	10 microfarads (μF), 25 V radial, electrolytic

Additional items

TR1	2N2222A npn
IFT1	Toko type YHCS11100
Box	plastic, approximate size 100 × 70 × 45 mm
Board	matrix, to fit inside the box
Connector	for PP3 battery
Knob	for RV1

10 What is a capacitor?

Introduction

A capacitor is a device that stores energy in the form of electricity. Much less energy than a battery, and for a shorter time, however. The simplest form of capacitor takes the form of two flat metal sheets separated by air; connections are made to each plate, as **Figure 1** shows. If you imagine a pair of these plates, 30 cm square and separated from each other in air by 1 mm, the capacitance of this device would be almost exactly 80 picofarads (pF), i.e. 80 million-millionths of the unit of capacitance, the farad. Now this is quite a small value, you will agree, and it comes about because the farad is such a large unit. Nevertheless, as you will probably know, we may have capacitors of value 10 000 microfarads (μF) in our radio equipment, and they can be smaller than your little finger, so they are obviously not made the same way!

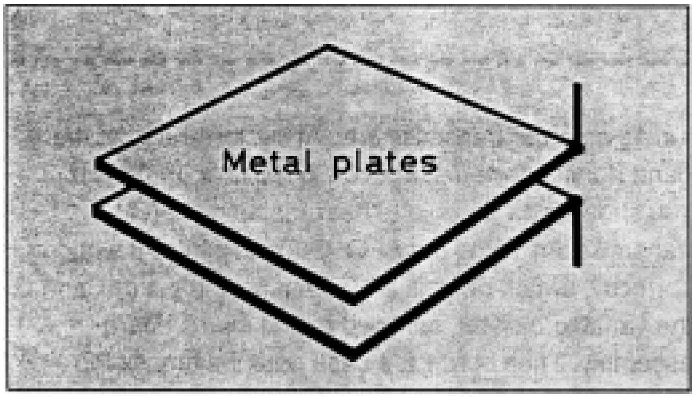

Figure 1

Large and small

We cannot go into the manufacture of capacitors here – after all, we are users of the devices, not the designers! First of all, beware of incorrect statements; the words ‘capacitor’ and ‘capacitance’ are *not* the same. For example, a large capacitor would be a description of one the size of a toilet roll. It need not have as large a capacitance as one the size of your little finger. A ‘large capacitor’ is one which is physically *big*, a ‘large capacitance’ refers to a capacitor which can store a larger amount of energy when a certain voltage is applied between its plates. The capacitors in a mains power supply are usually big **and** have large capacitances. High-power RF amplifiers may have large capacitors with small capacitances!

Electrolytics . . .

Electrolytic capacitors usually have capacitances of 1 μF or above. They differ from other capacitors in that they **must** be connected the right way round (i.e. they have positive and negative connections, just like a battery). They may explode if the connections are reversed!

. . . and the others

Other capacitors may be connected either way round, despite their names. We have *polystyrene*, *ceramic*, *silver-mica* and *tantalum*. Each has its own advantages and disadvantages, and the parts list for a project will always tell you which type is best.

Storing energy

If you were to connect a large capacitance across a 12 V power supply, nothing would appear to happen. Removing the capacitor from the supply and connecting it to a voltmeter would show that the capacitor has 12 V between its ends. This shows that, while the capacitor was connected to the supply, energy flowed from the supply into the capacitor. We say that the capacitor was *charged up* by the supply. If you are using an *analogue* voltmeter (i.e. one with a meter and pointer), you will notice that the indicated voltage slowly drops until, eventually, there is no voltage across the capacitor. This is because the capacitor has *discharged* its energy into the voltmeter. If you had used a smaller capacitance, the same would happen, except that the voltage would drop to zero more quickly – the capacitor stores a smaller amount of energy because its capacitance is smaller. Capacitors behave like other things in life – a small car can move more quickly than a large bus – a small piccolo emits a higher note than a flute – the voltages in a circuit containing a small capacitance will change more quickly than those in a circuit with large capacitance.

Varying the capacitance

Some capacitors are capable of having their capacitance changed manually; these are called *variable capacitors*. They work like the basic capacitor of Figure 1. Imagine moving the top plate of the pair a little to one side; the capacitance is determined, not just by the size of the two plates, but by their *area of overlap*. As this decreases, so does the capacitance. Such devices are limited in their capacitance, about 500 pF being the maximum value.

AC and DC

Because the plates of a capacitor do not touch each other, a direct current (DC) cannot pass between them. However, an alternating voltage on one plate can induce an identical alternating voltage on the opposite plate, and thus a capacitor appears to pass an alternating signal, even though currents as such, do not pass between the plates. This property of passing AC and not DC is very important, and a capacitor used in this way is called a *DC blocking capacitor* or, simply, a *blocking capacitor.* A blocking capacitor can be used at the same time, to couple a signal from one circuit to the next; here it would be known as a *coupling capacitor. Decoupling capacitors* are to be found where the capacitor is employed to remove an AC signal while retaining a DC component.

Finally . . .

Unlike resistors, the manufacture of capacitors renders them susceptible to excess voltage, so if you find a capacitor labelled 10 μF 16 V, it means that operating it above 16 V may fatally damage the device (and the circuit around it). This voltage is called the *working voltage* of the capacitor; on some electrolytics, you may find it expressed as *volts working* (i.e. 8 μF 450 V WKG).

Many smaller capacitors have their properties marked on them in a colour code, like resistors. **Figure 2** shows these codes, and their meaning, and the table below summarises the values of the colours.

Table 1

Colour	*Value*	*Voltage (tantalum capacitor)*	*Voltage (polyester capacitor)*
Black	0	10	–
Brown	1	–	100
Red	2	–	250
Orange	3	–	–
Yellow	4	6.3	400
Green	5	16	–
Blue	6	20	–
Violet	7	–	–
Grey	8	25	–
White	9	3	–

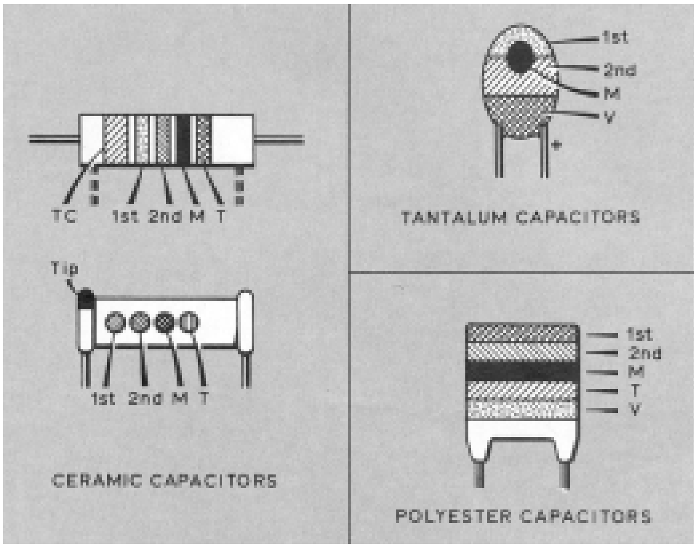
TC
1st
2nd
M
T
1st
2nd
M
V
+
TANTALUM CAPACITORS
Tip
1st
2nd
M
T
CERAMIC CAPACITORS
1st
2nd
M
T
V
POLYESTER CAPACITORS

Figure 2 Some capacitors have coloured bands or stripes, rather like resistors. The colour code, which is the same as the resistor code, is shown in Table 1. The band shown on the chart as '1st' is the first number of the capacitor's value in pico-Farads, '2nd' is the second number and 'M' is the Multiplier or number of noughts. For example, a capacitor reads from the top: Brown, Black, Yellow, Black, Red. Its value is One, then Nought, then Four more noughts = 100 000 pF (also referred to as 0.1 μF or 100 nF). Its tolerance (Black) is 20% and the working voltage (Red) is 250 V. The 'V' means the maximum working voltage. The band marked 'T' shows the tolerance, just like resistors, and the one marked 'TC' is only used on special capacitors designed to change their value with temperature

11 Waves – Part 2

Introduction

We left Part 1 with the concept that radio waves are divided up into bands which have different properties. Not *all* the properties are different, though. We need to discuss several wave properties, so we will start with what happens to waves as they propagate over long distances.

Getting weaker

Imagine a torch battery connected to a bulb by wires about 1 metre long. The bulb lights normally. If we now take the bulb 100 metres away from the battery and wire it up, we would expect the bulb to be somewhat dimmer, which is exactly what would happen. It happens because of the resistance of the wires – the wires do not form a perfect conductor. A similar situation occurs with radio waves.

All waves suffer from *attenuation* – they get weaker the further they travel. In cases of extreme attenuation, we need to apply some *amplification* before the attenuated wave can be used in a receiver.

Carrying information

When we speak over the telephone, the range of frequencies in our voices extends from very low frequencies up to about 15 or 20 kHz. In audio terms, this is quite a large *bandwidth* (meaning a wide band of frequencies). For communications purposes, however, most of this bandwidth is not needed, and in the telephone system (and in our transceivers), this is cut down so that it extends from about 200 Hz to 3 kHz, a reduction in bandwidth from 20 kHz to about 3 kHz. A bandwidth of 3 kHz has been found to be sufficient to convey speech intelligibly which, after all, is just what we need!

The radio waves coming from an amateur transmitter convey our speech signals over long distances. By themselves, the speech signals do not travel very far, so they have to be combined with a radio signal that *will* travel long distances. This extra signal is called the *carrier wave* (or just the *carrier*), because its job is to *carry* the speech signals along with it! The process of combining the speech (or Morse code) with the carrier is called *modulation*.

Wider and wider

Sending Morse code is achieved simply by switching (or *keying*) the carrier on and off. The bandwidth of the transmitted signal is only about 100 Hz. Speech, with its reduced bandwidth of 3 kHz, will produce a *single-sideband* (SSB) transmitted signal with a bandwidth of 3 kHz. If the same speech signal were used to produce an *amplitude-modulated* (AM) signal from the transmitter, it would have a bandwidth of about 6 kHz. Perhaps you can now understand why the bandwidths needed to produce hi-fi broadcasts need to be so large. TV signals need bandwidths running into tens of megahertz!

Waves need aerials

Radio waves are produced whenever changing currents flow through a wire, and when that wire is made in such a way as to maximise the *radiation* from the wire, it is called an *aerial* or *antenna*. The same piece of wire will receive radiation from other aerials; an aerial will transmit and receive. This is an important property of the aerial: when a current flows through it, electromagnetic waves are launched into the air; when electromagnetic waves in the air encounter the aerial, currents are produced in it.

From the simplest transmitting aerial, waves travel in all directions, like the waves on the pond that we considered in Part 1. They will travel a long way through air and space before they become too week to be received. They won't travel very far into the earth, however! The earth will *reflect* some of the wave and will *absorb* the rest. That portion of the wave which is reflected will again travel through air and space until it is totally attenuated.

Look at **Figure 1**; A represents a radio transmitter, with B1, B2 and B3 being receiving stations. The two arrows pointing 'downwards' from A represent two of the waves from A which just graze the earth's surface. Waves above

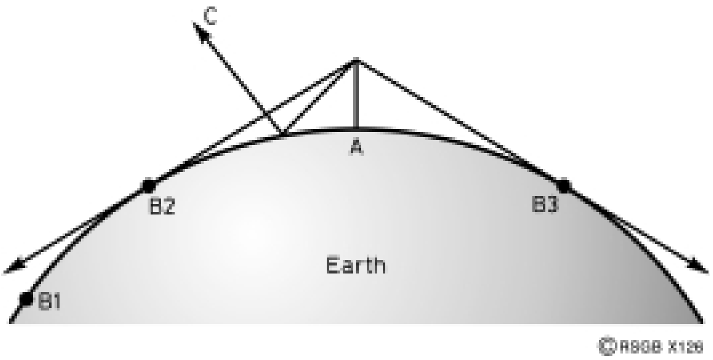

Figure 1 Stations between B2 and B3 receive the *ground wave*

these will travel on into space; waves below them will either be absorbed and reflected by the earth or received by aerials. B1 will not receive any signals from A, because it is below A's horizon. B2 and B3 can receive A's signals because they are *just* on A's horizon. Any stations between B2 and B3 will also receive A's signals, which are known as *ground-wave signals*. The wave at C represents one which is reflected by the ground and travels into space.

This description begs the question of how signals are received from stations well beyond the ground-wave range.

Mirrors in space

Suppose that there was something, out in space, that would reflect radio waves. Waves from A that travel out into space could be reflected off it and return to earth, enabling stations such as B1 to receive A's signals. The situation just described is illustrated in **Figure 2**.

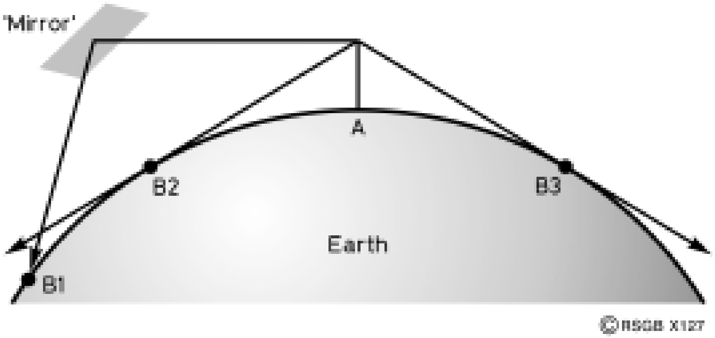

Figure 2 Reception of A's signals at B1 using the *sky wave*

Just such a mirror in the sky really *does* exist. It is not man-made, of course! The earth is surrounded by the atmosphere, a mixture of many gases, such as nitrogen and oxygen. The energy contained in the radiation from the sun is more than sufficient to *ionise* these gases, thus making them into electrical conductors. When a gas is ionised, some of its electrons are physically stripped out of the atoms and are free to move about, just as the electrons of a metal do in a wire. Consequently, we can regard this part of the atmosphere (the part illuminated by the sun) as acting like a sheet of metal, which *reflects* radio signals! It is not a perfect reflector, but is sufficient to produce long-range (DX) radio propagation via the *sky wave* under the right conditions. (A more down-to-earth example of ionised gases conducting electricity can be found in the fluorescent tube and the neon sign – many gases glow when they are continuously ionised.)

This conducting region at the extremity of the atmosphere is called the *ionosphere*, and it exists in layers between 60 km and 700 km above the earth's surface. When the ionosphere is sufficiently ionised, it glows; this is the natural phenomenon known as the *aurora borealis*, or the *northern lights*.

The property of the ionosphere that enables radio waves to be reflected does not act in a uniform way; it is very selective about which waves it reflects, and which waves go straight through it and into outer space. In general, it reflects only those waves with frequencies below about 30 MHz – the HF bands!

In Part 3 we will look at families of waves.

12 An LED flasher

Introduction

The LM3909 is an integrated circuit (IC) which will flash a light-emitting diode (LED). Using only two extra components and a battery, the circuit is cheap and has a very low current drain from a 1.5 V cell. The circuit can be used as a novelty flasher, an indicator for a dummy alarm bell box, or it could be attached to a torch so that it could be found easily in the dark! The simple circuit is shown in **Figure 1**.

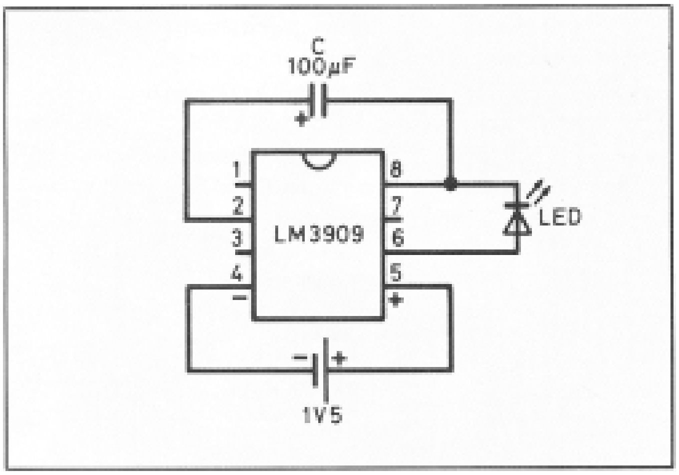

Figure 1 Circuit diagram of the LED Flasher. Pins 1, 3 and 7 of the IC are not used

Assembly

The circuit can be built on a small piece of Veroboard (the piece shown in **Figure 2** measures 15 holes by 10 strips). Using such a board, follow these instructions.

1. Depending on how far away you want the LED from the circuit board, solder a length of insulated wire to each lead of the LED. Use different colours of insulation – say, red and black, connecting the *red* lead to the *anode* (a) lead (the longer one) of the LED, and the black one to the *cathode* (k). Figure 2 shows these leads.
2. Cut the copper tracks as shown in Figure 2, using a 3 mm (⅛ inch) diameter drill, rotated between thumb and forefinger, or use the proper tool. Make *absolutely* sure that the tracks are completely broken!
3. Fit the IC holder in the correct position, using the cut tracks as guides, and make sure the small notch is facing towards the top of the board. Solder the pins to the copper tracks.
4. Mount the capacitor, positive end to the left, so that the positive lead is soldered to track F, which connects it to pin 2 of the IC; the negative lead is soldered to the right-hand side of track E, this being connected to pin 8 of the IC.
5. Solder on the battery leads, positive to the right, and the extended LED leads, positive downwards.

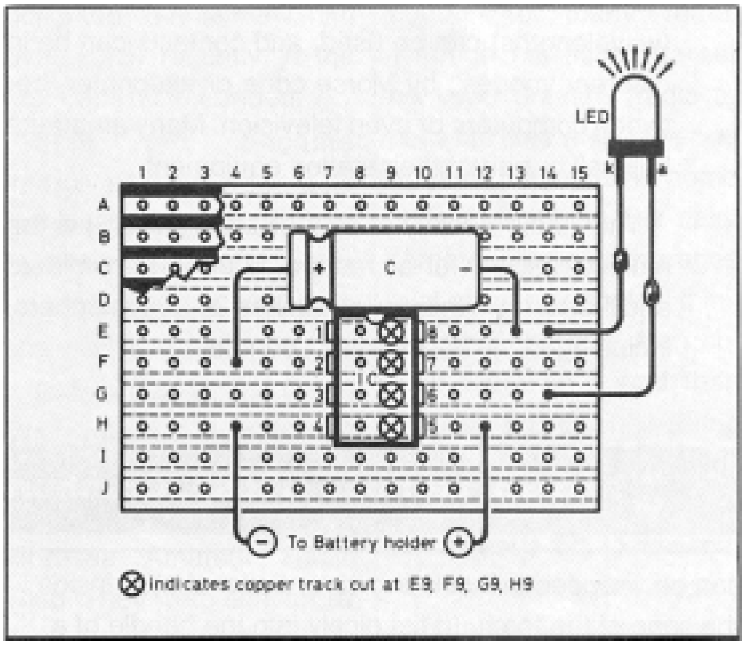

Figure 2 Board layout viewed from the component side. The tracks are cut under the board where shown

6. Check the circuit, and hold up the board to a bright light and look carefully for solder bridges between the tracks and pieces of copper swarf which may have escaped your inspection in 2 above! Remove whatever you find.
7. When all seems well, put the IC into the socket, ensuring that the notch or dot on the upper surface of the IC lines up with the notch on the holder. Line up each pin on the IC with the hole below it before pressing *gently* on the IC with the board supported on a firm surface.
8. Connect the battery; the LED should start to flash. The circuit is complete and working!

If you prefer, the whole circuit (battery included) can be mounted in a small plastic box, with the LED mounted on a clip and protruding through the panel. There are many other possibilities, and it is up to you to find an application for your own use.

Parts list

		Maplin code
LM3909	Integrated circuit	WQ39N
IC socket	8-pin DIL	BL17T
LED	5 mm diameter	WL27E
100 microfarad (μF)	Electrolytic capacitor (10 V)	FB48C
Battery holder	For AA-size cell	YR59P
Battery	1.5 V AA cell	

Small piece of Veroboard (15 holes by 10 strips)
Small plastic box (if required)
LED clip (if required)
Two lengths of coloured, insulated wire for LED (as required)

Availability

All parts can be obtained from Maplin Electronics Ltd.

13 Waves – Part 3

Introduction

When we talk about the spectrum being divided up into bands, this is just for our convenience; there are no natural divisions although, as we have seen, some of the properties of different bands really *are* different! Let's look at **Figure 1**, and see how the frequencies are divided up.

The divisions

- **Very low frequencies** (VLF) cover the range from a few kilohertz up to 30 kHz. Very long-range communication is possible, but at *very* small bandwidths. It is used for special purposes.
- **Long waves** (LW) are used for medium-distance commercial broadcasting and have frequencies from 30 kHz to 300 kHz.
- **Medium waves** (MW) are used for commercial broadcasting, and use frequencies from 300 kHz to about 1.5 MHz (1500 kHz). Typical range is about 200 km.
- **Short waves** (SW) encompass both the low-frequency (LF) and high-frequency (HF) amateur radio bands. There are nine narrow amateur bands in the SW spectrum between 1.8 MHz and 30 MHz. Some of these bands give round-the-world communication.
- **Very high frequencies** (VHF) span the range between 30 MHz and 300 MHz. Relatively short-range communication is possible. They were once used for broadcast TV before it moved to UHF. There are now three

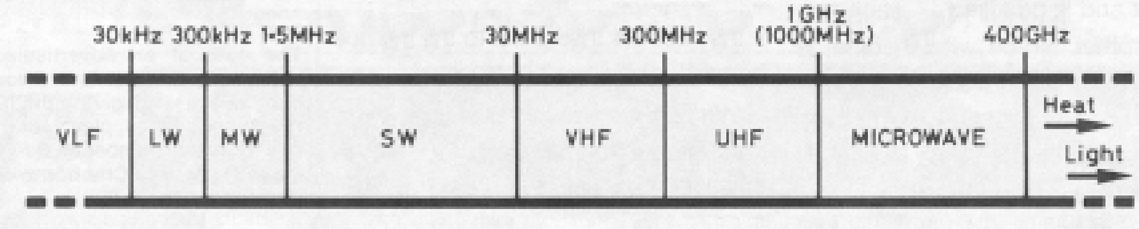

Figure 1 Diagram of radio frequency spectrum

amateur bands here – 6 m, 4 m, and 2 m. Repeaters are used to extend the usable range of mobile stations. VHF waves are not usually reflected by the ionosphere, but when they are, ranges of several thousand kilometres are possible. Weather affects these waves on a regular basis, however. In addition to amateur users, the VHF part of the spectrum is also used by the police, the fire and ambulance services, weather satellites, and many others,

- **Ultra-high frequencies**, sometimes called *centimetre waves*, cover the range from 300 MHz to 1000 MHz (or 1 GHz). The only amateur band in this range is the 70 cm band, and we share it with radar, TV and cellular telephone users as well.
- **Microwaves** begin at 1 GHz and extend to about 400 GHz. They are *never* reflected by the ionosphere, are partially attenuated by buildings, and are reflected from aircraft and cars. Microwave absorption in the atmosphere is quite significant, and rain and fog can attenuate microwaves quite heavily.
- **Heat, light** . . . Above 400 GHz we run into the infra-red bands and on into the visible light and ultra-violet bands. We generally take 400 GHz as being the limit of what we class as radio waves.

Bandwidth again

Complex signals need more bandwidth than simple signals. Even if it were possible, we would not be able to transmit a single television channel in the whole of the MW broadcast band! When TV used part of the VHF band, only five channels were possible in the range from 45 MHz to 68 MHz. By moving TV to the UHF band, we now have 47 channels between 470 MHz and 855 MHz!

It's your choice!

The number of permutations you have amongst all the modes and all the bands is enormous! Only you can decide what you are interested in and what you want to learn about. That is the attraction of amateur radio!

14 Choosing a switch

Simple, but be careful!

A switch is the simplest electronic component, but care is needed in choosing the correct one for the job. Here is a list of the main characteristics of a switch, which will help you to select what you want.

1. **Rating**. This gives the maximum voltage and current that a switch can handle. For example, a 250 V 1.5 A switch will switch mains voltage at a current not exceeding 1.5 A. If the current is greater than 1.5 A, the switch may get hot and fail. If the voltage is too great, the switch may arc each time it is switched off, thus wearing away the contacts.
2. **Number of poles**. A single switch can control many circuits; the number of poles tells you *how many* different circuits it can handle. See **Figure 1**.
3. **Number of throws**. This tells you the number of positions each pole can have. This is best illustrated in Figure 1. The simplest ON/OFF switch is a Single-Pole, Single-Throw (SPST) switch. A Single-Pole,

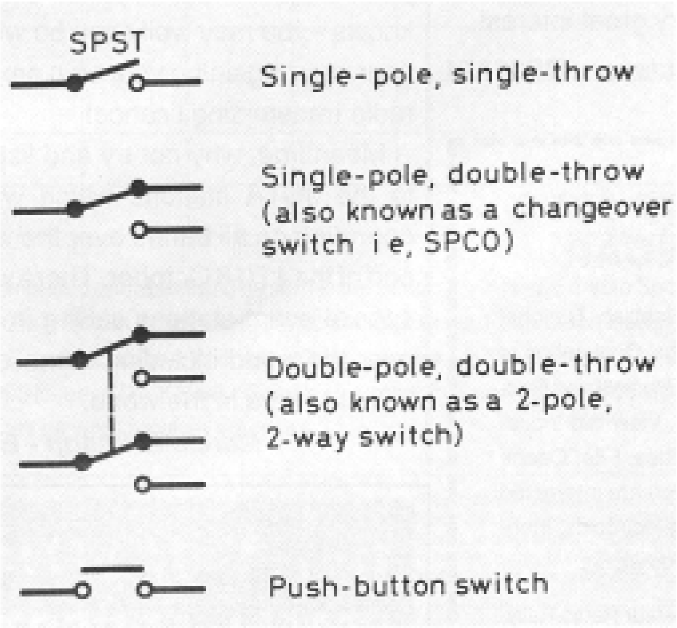
SPST
Single-pole, single-throw
Single-pole, double-throw
(also known as a changeover
switch i e, SPCO)
Double-pole, double-throw
(also known as a 2-pole,
2-way switch)
Push-button switch

Figure 1 Some circuit symbols for different types of switch

Double-Throw (SPDT) switch may also be used as an ON/OFF switch, but is used mainly to change between two parts of a circuit, and is commonly known as a Single-Pole Change-Over (SPCO) switch. Two or more SPCO switches can be operated at once; Figure 1 shows an example. The two switches are said to be *ganged*. See also 'Types of switch' below.

4. **Number of ways**. When a switch has more than one throw, we tend to use the word 'ways' instead. This means that if a switch has one pole and six throws, we would normally call it a '1-pole 6-way' switch. Such switches tend to be rotary switches, as are described below.

Types of switch

(a) **Push-button**. These are found on calculators, telephones, electronic games and most equipment with a digital display.

(b) **Rotary**. These are switches controlled by a knob, and are turned instead of moved up and down. **Figure 2** shows the rear of such a switch and its circuit symbol. The commoner types of rotary switch are: 1-pole, 12-way; 2-pole, 6-way; 3-pole, 4-way; 4-pole, 3-way; 6-pole, 2-way. All these switches have 12 click positions, as you may have guessed, but each one comes with an adjustable end-stop so that you can set the correct number of ways according to the contacts on the switch.

(c) **Slide**. This switch is common on the cheaper types of radio, mainly as an ON/OFF or band-changing switch. They are not very rugged, but are small and cheap to produce. Very small types are manufactured for use on PCBs.

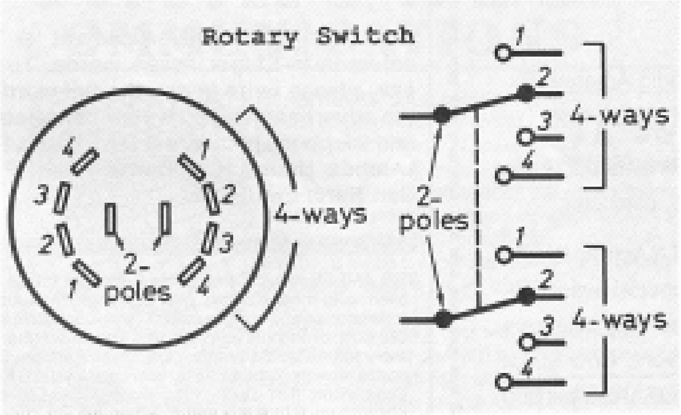

Figure 2 The rotary switch can select several different circuits at once

(d) **Toggle**. Available as two-way types or three-way, with a centre-off position. Can be manufactured to handle very high currents.

(e) **Micro-switch**. These are devices which are usually operated indirectly, such as when the cover is removed from a high-voltage power supply, or when the door of a fridge is opened. The ‘micro’ part of the name doesn’t refer to the size of the switch, but to the *small* movement that is required to activate it.

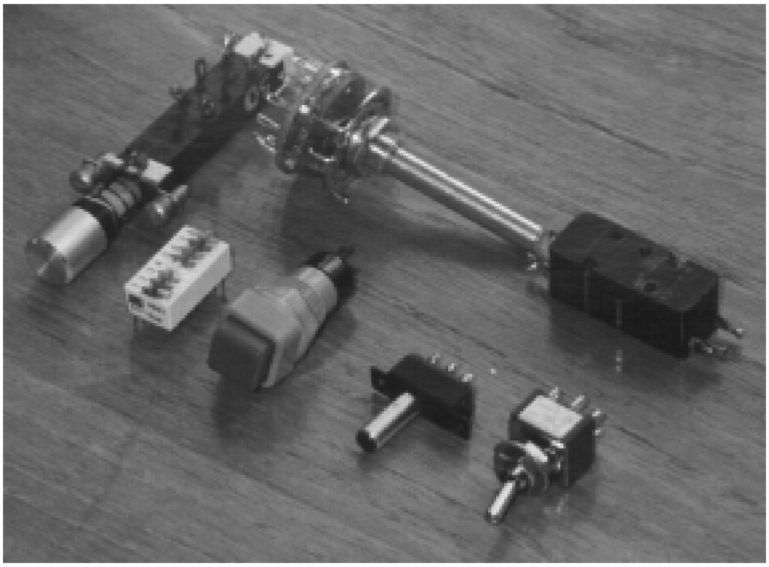

15 An aerial tuning unit for a receiver

Introduction

Any length of wire will act as an antenna (or aerial) but, to get the best results from a transceiver or receiver, an aerial tuning unit (ATU) is required. This matches the *impedance* of your aerial to the impedance at the aerial socket of your radio. Impedance is like resistance, and is measured in ohms, but it is used for alternating currents, and hence is common in audio and RF engineering. Most receivers have an impedance at the aerial socket of about 50 ohms (Ω); aerial impedance, on the other hand, can be anywhere between 20 Ω to over 1000 Ω, depending on its length and its height above ground.

Fortunately, you don't *need* to be able to calculate your aerial's impedance; all you need is a device that will perform the matching operation for aerials with a large range of impedances, and this is exactly what is described here! The subject of *matching* is a complex one, but all you need to know is that most signals will become clearer, and that there will be less noise and interference. Stations will become louder, so you will probably be able to reduce the setting of your RF gain control (always a good thing to do).

This design of ATU covers all amateur and broadcast bands from 10 m (28 MHz) to 80 m (3.5 MHz), and is very easy to build. The circuit is shown in **Figure 1**.

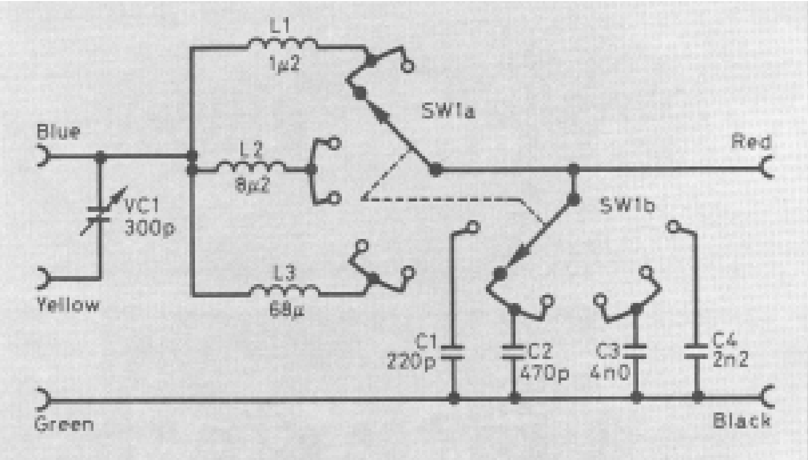

Figure 1 Circuit diagram of the Antenna Tuning Unit, showing the use of a 2-pole, 6-way rotary switch to select inductors (L) and capacitors (C)

Construction

1. Firstly, you will need a simple plastic case in which to house the ATU. The size should be approximately 85 × 145 × 50 mm.
2. Start by drilling two 10.5 mm holes in the front of the case; these are for the 6-way switch and the tuning capacitor.
3. Drill three 8 mm diameter holes in the left-hand side of the box, for the three sockets, coloured blue, yellow and green.
4. On the right-hand side of the box, drill two 8 mm diameter holes for the red and black sockets.
5. Now, fit the 6-way switch (SW1), the tuning capacitor (VC1), and all the sockets to the case. Check that the vanes of the capacitor rotate smoothly when the shaft is turned.
6. Wire up the inductors (coils). Figure 1 and the wiring diagram of **Figure 2**, will help with this. As you can see, each one side of each coil is connected to two switch connections, the other end going to VC1.
7. Solder in the fixed capacitors. One end of each goes directly to the ground socket (black), and the other end goes to the switch.
8. Solder a wire between the green and black sockets. The output from the ATU comes from the red socket, and this is connected to the two tags in the centre of the switch, as Figure. 2 shows clearly.
9. Finally, connect the blue and yellow sockets to the tuning capacitor, and the ATU construction is complete.

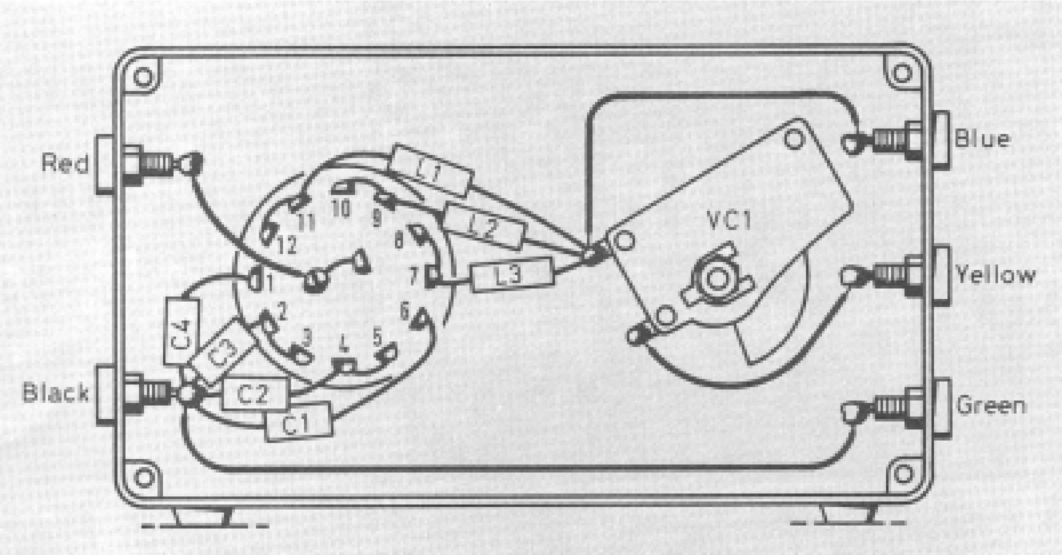

Figure 2 The internal view of the case shows the main tuning capacitor, VC1. This is a solid dielectric type, which has adjustable brass plates

In practice . . .

Figures 3 and 4 show two different ways of connecting your aerial to your ATU. In each case you will need to select each switch position in turn, and rotate the tuning capacitor through its full range while listening to a station. You should find that one switch position enables VC1 to produce a peak in the signal strength in the loudspeaker. At this point, your aerial and receiver are said to be *matched*. Stations in the same band will probably peak with VC1 at the same setting of SW1, but different bands will almost certainly require different positions of SW1.

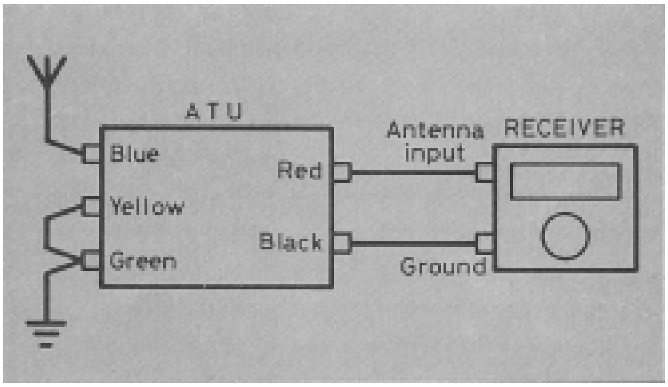
ATU
Blue
Yellow
Green
Red
Black
Antenna input
RECEIVER
Ground

Figure 3 You may find that parallel tuning gives best results with your antenna

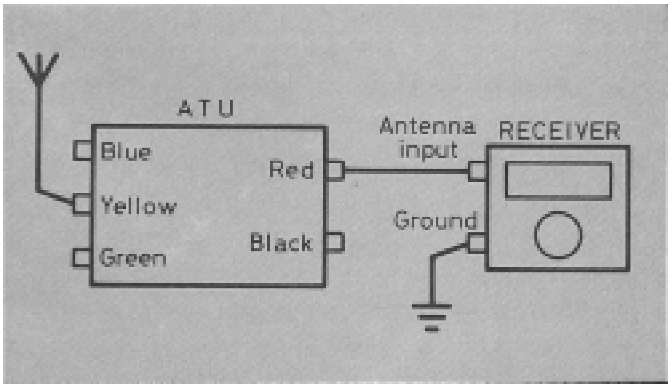
ATU
Blue
Yellow
Green
Red
Black
Antenna input
RECEIVER
Ground

Figure 4 In other cases, series tuning could be the most effective arrangement

Parts list

Capacitors (all rated at 16 V or more)

C1	220 picofarads (pF) polystyrene
C2	470 picofarads (pF) polystyrene
C3	1000 picofarads (pF) or 1 nanofarad (nF) polystyrene
C4	2200 picofarads (pF) or 2.2 nanofarads (nF) polystyrene

Inductors

L1	1.2 microhenries (μH)
L2	8.2 microhenries (μH)
L3	68 microhenries (μH)

Switch

SW1	2-pole 6-way rotary

Sockets
  4 mm type, one each of red, black, yellow, blue, green

Additional items
  Plastic or metal case, e.g. Maplin type YU54
  Two large knobs for SW1 and VC1

16 A simple 2 m receiver preamplifier

Introduction

Designed specifically to complement the modified air-band portable (also described in this series), this can be used with some success on many receivers suffering from 'deafness' on VHF.

The circuit

An RF preamplifier is a device which improves the input signal to an existing receiver, enabling it to work more effectively. Because of the noise which is added to the signal by the preamplifier, *very* weak signals may not be usefully enhanced, but stronger signals will be improved considerably.

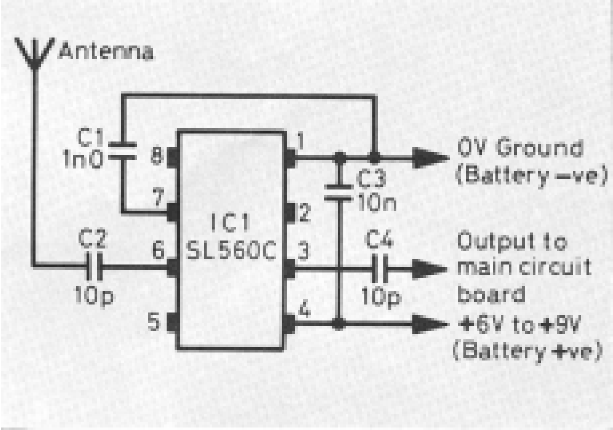

Figure 1 The integrated circuit is mounted upside down. Make sure you identify pin 1 which has a small dot next to it

This little circuit (shown in **Figure 1**) uses a GEC/Plessey integrated circuit type SL560C. With the addition of four capacitors, it is used between your aerial and the aerial input of the radio's PCB.

Putting it together

1. Use a small piece of prototype (matrix) board about 25 mm square. Use an 8-pin DIL socket for the integrated circuit (don't risk soldering the chip – it is seldom a risk worth taking). Figure 1 shows the connections to the socket, *looking from underneath*.
2. Make special note of the pin numbers, so that you know how to put the chip into the socket when you have finished. The positive and negative connections to the circuit are taken from the main PCB *after* the ON/OFF switch – so that the switch operates the preamplifier, too.
3. Unsolder the lead to the radio's telescopic aerial and connect it to the free end of C4, as shown. Then solder a short lead between the telescopic aerial and C2.
4. Sometimes it is possible to cajole your little preamplifier into the radio's plastic case, provided there is room and that you make sure that none of the soldered joints on your little board touch any of the metal inside the case when you replace the back and screw it on again.
5. If there is not enough room inside, then put the preamplifier into its own box, with battery and switch, and its own aerial. Keep the lead from the preamplifier to the aerial connection of the radio as short as possible – perhaps using screened cable.

Because your preamplifier is untuned, you will find not only that it helps with reception on 2 metres, but also that reception on the FM broadcast band is improved!

Parts list

Integrated circuit

IC1	GEC/Plessey SL560C

Capacitors

C1	1 nanofarad (nF) ceramic
C2, C4	10 picofarads (pF) ceramic
C3	10 nanofarads (nF) ceramic

Additional item

Prototype broad approx. 25 × 25 mm

17 Receiving aerials for amateur radio

Introduction

For any radio receiver to work well, it must have some form of antenna, or aerial. In almost all domestic transistor radios, the aerial is built into the set, either as a ferrite rod (which looks like a rod of dark grey metal) on which are wound coils of wire, or as a chromium-plated telescopic metal rod. Some radios have both forms of aerial, using the ferrite rod aerial for long waves (LW) and medium waves (MW) and the metal rod for very high-frequency (VHF) stations using frequency modulation.

There's broadcast reception . . .

No aerial is perfect, and these two types are *far* from perfect! As in most mass-produced equipment, *they serve their purpose*, which is not critical, and they allow the radio to be carried around easily, because they are not big and bulky. The broadcast stations are very powerful, usually quite close, and the circuits in the radio are quite sensitive, so the need for large aerials disappears.

. . . and there's amateur DX

Here we have an application which, in comparison with the broadcast situation, couldn't be much more different. The stations do not use much power, they may be half the world away, and the requirement for good aerials and very sensitive receivers is paramount. The aerials must be large and they must be as high as possible, away from buildings and trees, which cause reduction in signal strength, and away from man-made sources of interference.

You may not have thought about this but, in general, the larger an aerial becomes, the longer the cable (or 'feeder') must be in order to reach your shack and the receiver inside it. Cables reduce the received signal, so what your aerial gains by being large, the feeder (if you're not careful) will lose!

A simple aerial . . .

We now know that a 'good' aerial is essential. But what is a 'good' aerial? It depends on your purse, your property and your enthusiasm! One of the simplest (and, incidentally, one which is *not* subject to the cable loss problem discussed above) is the **Long Wire**, shown in **Figure 1**. This is, quite literally, a long piece of wire going from a chimney stack to a tree or pole

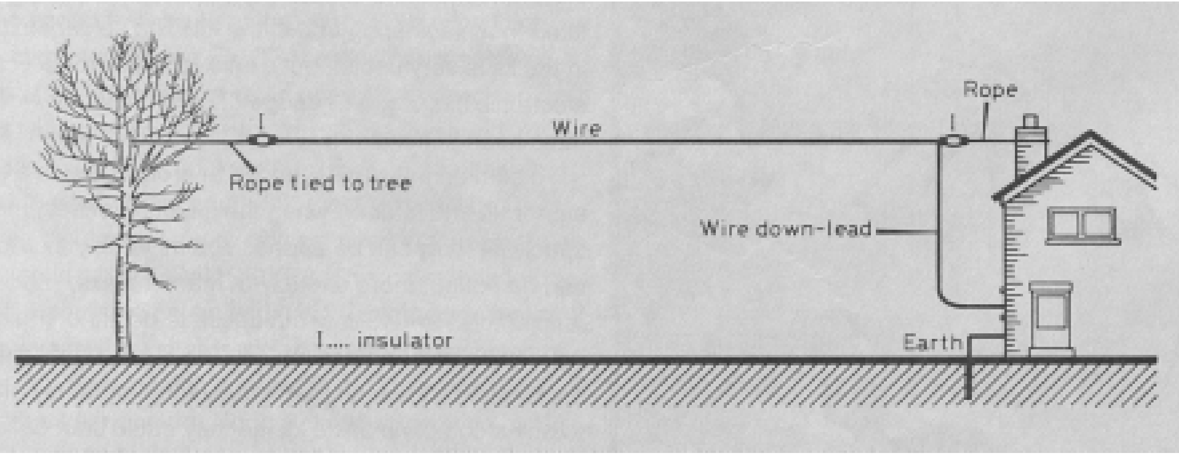

Figure 1 A tree at the bottom of the garden can provide a useful antenna support. Insulators can usually be obtained from Tandy stores or TV aerial suppliers

at the end of the garden. The longer it is, the better. Notice that the wire itself is *not* used to loop around the tree or chimney. Rope is used for both, and is secured to the end of the wire using an insulator of the 'dog bone' or 'egg' variety, to be found in profusion at rallies.

The wire is brought into the house through a window into the shack where the receiver is situated. The long wire aerial is best used with an aerial tuning unit (ATU), which is also described in this book. One of the advantages of the long wire is that it can be used on several frequency bands.

. . . and more complicated ones

Next up the ladder of complexity is the **dipole** (meaning two poles, or two elements). One form of dipole for lower frequencies would take the same basic form as the long wire, except that the feed to the receiver is taken, not from the end, but from the centre. The wire is essentially cut into two halves, and the two ends at the centre are connected to one end of a coaxial cable, which is then taken to the receiver. A smaller form of this is shown in **Figure 2**, which is conveniently mounted in a house loft. Aerials should always be mounted outside for best results, but will work when mounted inside, and the loft space is the logical situation.

The total length of the dipole should ideally be one-half of the wavelength of operation – hence the term 'half-wave dipole'. For example, a dipole for use on 20 m should be about 10 m long. Dipoles are thus 'single-band' aerials. They can be modified for use on several bands, and then become known as 'trapped dipoles', having coils and capacitors at certain points along their lengths.

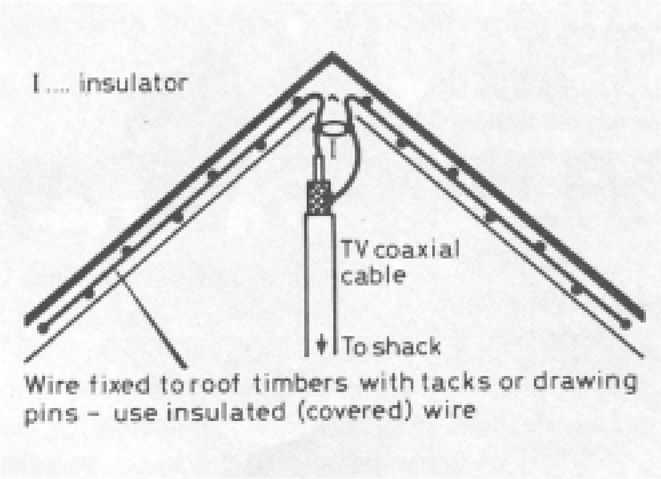
I .... insulator
I
TV coaxial cable
To shack
Wire fixed to roof timbers with tacks or drawing pins – use insulated (covered) wire

Figure 2 An indoor antenna should be mounted as high as possible

Indoor aerials

For flat- or apartment-dwellers, an indoor aerial is often the only solution, in which case the dipole of Figure 2 may be of great use. Other popular types include the **loop aerial**, of which several types are shown in **Figure 3**. The shape of the loop is of secondary importance, but it should be as large as possible.

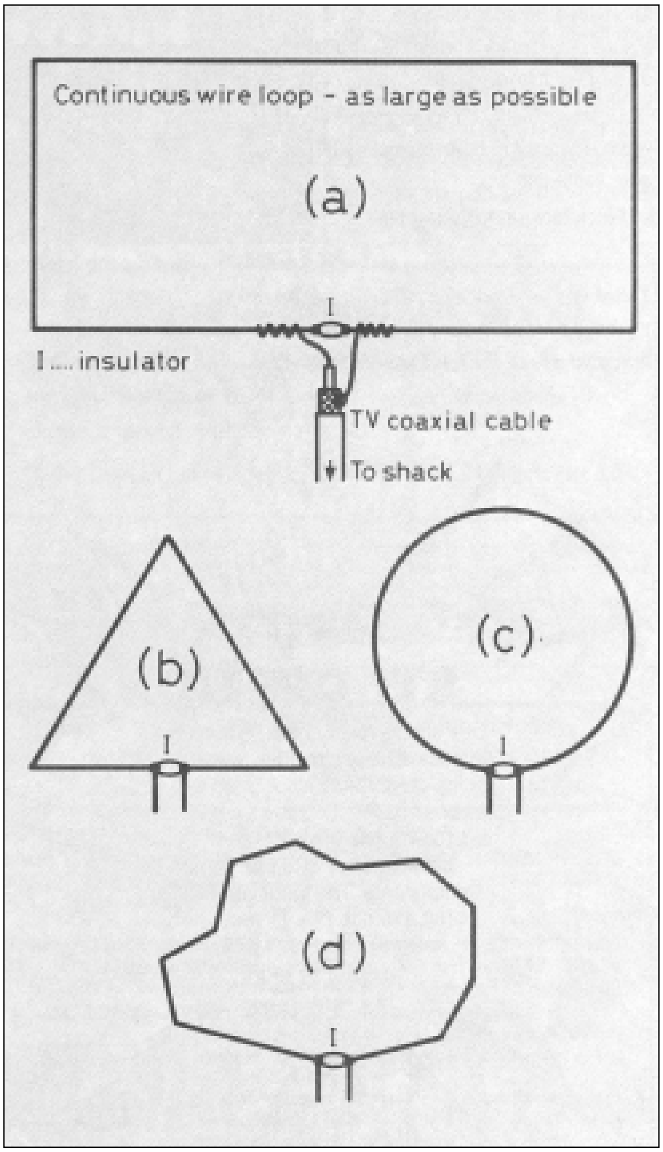
Continuous wire loop – as large as possible
(a)
I
I .... insulator
TV coaxial cable
To shack
(b)
I
(c)
I
(d)
I

Figure 3 Try different shapes of loop antenna to see which works best

Ingenuity is called for when considering indoor aerials. A dipole can be mounted along a picture rail or pelmet. A vertical loop of aluminium foil can be stuck to an inside wall and covered with wallpaper! The designs are limited only by your imagination, and will give you hours of listening pleasure

Outdoor aerials

If you want to venture beyond the simple long-wire discussed earlier, then the erection of outdoor aerials can be a bit tricky; they can be more than a one-man job, and can involve seeking planning permission. Get some help advice from someone who has done it before – what about your local radio club? Plenty of designs for aerials appear in *RadCom*, so you should never be short of ideas!

18 The Colt 80 m receiver – Part 1

Introduction

This is the first of four parts detailing the design and construction of a simple radio capable of good reception of amateur radio signals on the 80 metre band. If you can solder and have some basic hand tools, at the end of Part 4 you will have a working receiver, in which you can take pride, and start some serious listening on 80 m! For testing, all you will need is another receiver and a multimeter.

Description

The radio will be built in three modules, or stages, as illustrated in **Figure 1**. Each of these is built on a printed-circuit board (PCB) or matrix board and can be tested in its own right. The case will need some holes drilled, and care will be needed to mount the PCBs in the case.

Get to it!

Surprisingly enough, when you are building a radio section by section (which is *always* the best way), it is easiest to start at the output and work

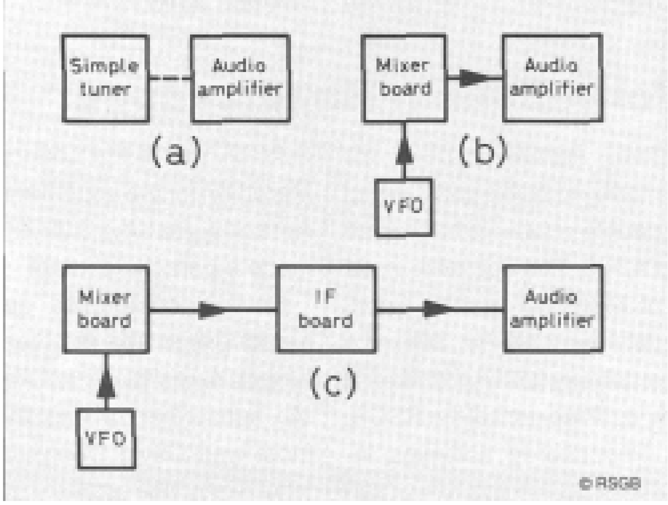
Simple tuner
Audio amplifier
(a)
Mixer board
Audio amplifier
(b)
VFO
Mixer board
IF board
Audio amplifier
(c)
VFO
© RSGB

Figure 1 Build your Colt and watch it grow. A simple crystal set (a) becomes a direct conversion receiver (b) and finally an 80 metre amateur band superhet (c)

towards the input, and you can test what you have done stage by stage. You will see what this means as you progress with your construction.

Every receiver needs some audio frequency (AF) amplification to make the sound signals big enough for you to hear. The circuit for the AF amplifier is shown in **Figure 2**. It uses the TDA7052 integrated circuit (IC) plus a handful of extra components. R1 in conjunction with C2 and C3 *decouple* the battery supply, preventing any audio signals getting through to it and affecting other parts of the radio, when they are connected. C1 acts to prevent high frequencies (above the 3 kHz bandwidth) going into the amplifier input. A volume control, VR1, is connected across the amplifier input, so that the amplifier can accept signal inputs over a wide range. **Figure 3c** shows the connections, which are made with screened cable. The

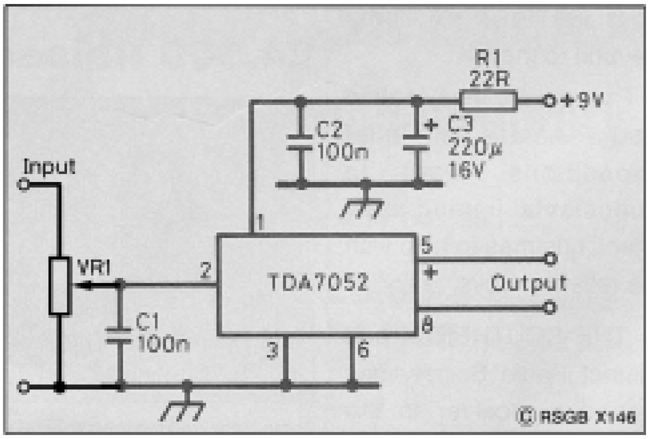
R1
22R
+9V
C2
100n
+ C3
220µ
16V
Input
1
VR1
2
TDA7052
5
+
Output
8
C1
100n
3
6
© RSGB X146

Figure 2 The Philips TDA7052 integrated circuit (IC) used in the audio amplifier needs very few extra components. It has a signal voltage gain of 100 times and the output is suitable for a loudspeaker or headphones

centre conductor of the left lead goes to the point marked ‘input’ in Figure 2, the braid being connected to the amplifier earth (0 V) tag. The right lead to VR1 goes to whatever signal source you have for testing – see later. VR1 is *not* mounted on the PCB.

The circuit is constructed on a small PCB or matrix board. Make sure that the electrolytic capacitor, C3, is soldered into the board the correct way – its positive and negative connections are shown in **Figure 3b** for reference. If you are **at all** concerned about soldering the IC into the board, enlist some help, *or*, obtain an 8-pin DIL socket, which you can solder in and then carefully insert the chip into the socket, making sure that is the correct way round. The markings on the chip are shown in **Figure 3a**.

The output leads from the amplifier go to a plastic 6.3 mm (¼ inch) mono jack socket, so that *neither* output lead is connected to the metal case. The amplifier will drive a pair of headphones or a small 8 Ω loudspeaker.

The battery leads must be the right way round also; the battery itself can be a PP3 or PP9, or you can use a small DC power supply.

Testing

First, check that all the components are in the right places, that your soldering is good, and that you have headphones or a loudspeaker connected. Set VR1 about halfway along its travel. Connect the battery.

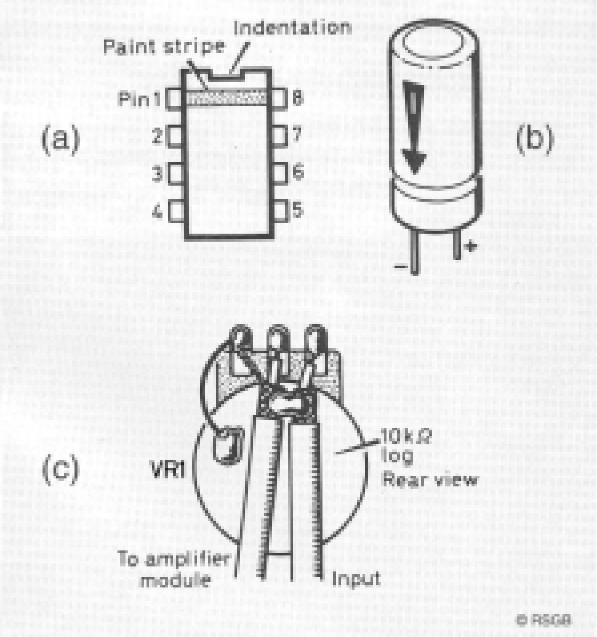

Figure 3 It is important to check the component connections carefully. The diagram shows (a) top view of the IC, (b) electrolytic capacitor and (c) volume control (VR1) connected across the input of the amplifier

A slight hissing noise should be heard; touching the input lead to the amplifier (the centre of the three connections on VR1) should produce a loud buzz. Touching the shaft at the same time will make the buzz quieter. This is the quickest way of confirming that your amplifier *seems* to be working. The only real test is to give it something meaningful to amplify! See the design of our *Crystal Radio Receiver* for full details.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	22 ohms (Ω)
VR1	10 kilohms (kΩ) log

Capacitors

C1, C2	0.1 microfarad (μF)
C3	220 microfarads (μF) electrolytic 16 V

Integrated circuit

IC1	TDA7052 audio amplifier

Additional item

PCB	(see below)

Component suppliers:
Maplin

The next part . . .

The metal case will be marked out ready to receive the completed modules.

19 A crystal radio receiver

Introduction

Here's a quick project to fill in a winter's evening! It was originally designed as a piece of test gear for the *Colt* receiver, but can be successfully used by anyone as a first radio project. Another use for it would be as test gear for any audio amplifier project requiring an audio input for test purposes; this is ideal, as it does not require any power supply!

Details

The initial rough-and-ready test for the audio amplifier of the *Colt* receiver will have whetted your appetite; you will want to prove more conclusively that your amplifier works, and in a way that others in the household will appreciate. Buzzing noises are not convincing in this respect! You are going to put together a simple crystal set – the simplest type of radio that there is – and use it as a signal injector for your amplifier. In this way, you build a real medium-wave (MW) receiver which drives a loudspeaker, as an intermediate product of the construction of an 80 metre amateur band receiver!

The circuit diagram is shown in **Figure 1**. It has a 60-turn coil mounted on a small piece of paper or card wrapped round a ferrite rod. The coil has a connection made to its centre-tap (the middle turn of the coil). The tuning capacitor, VC1, is the most expensive part of the circuit but don't worry, it will be used in the final design of the receiver also! Connect the diode, D1, from the centre-tap to the input to the potentiometer of the amplifier circuit of the *Colt*. Be careful to connect the aerial to the *vanes* of VC1, and not to its *frame*, or you will experience some strange effects when you are tuning. If your amplifier is working correctly, you should be able to receive local stations on medium-waves quite well.

If you think the crystal set will be of use to you in the future as a signal injector, all you will need will be another variable capacitor! If you do not intend to use the amplifier, a small crystal earpiece will allow you to listen. Walkman-type headphones will *not* work!

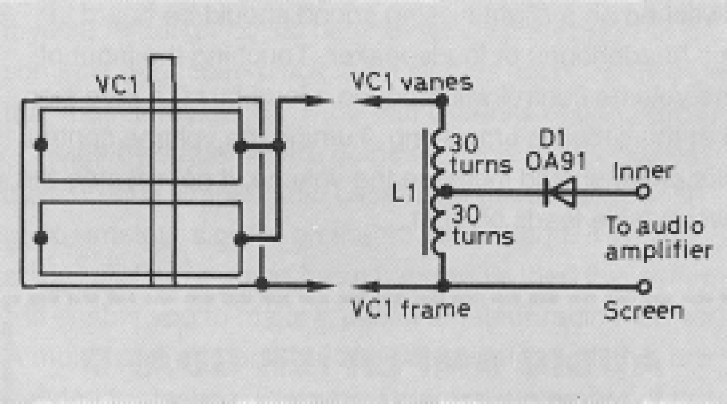
VC1
VC1 vanes
30 turns
D1 OA91
Inner
L1
30 turns
To audio amplifier
VC1 frame
Screen

Figure 1 Both sets of moving vanes are joined as shown. L1 is wound on a ferrite rod with 32 SWG wire and centre-tapped. A single winding (no tap) can be used, joining D1 to the top of the winding

Parts list

Variable capacitor	
VC1	125 + 125 picofarads (pF) twin-gang
Ferrite rod	
About 100 mm to 140 mm long	
Diode	
D1	OA91 germanium diode
Additional items	
Wire	2 m of enamelled copper wire, between 23 SWG and 32 SWG
Earpiece	High-impedance crystal type (only needed if you are not using the amplifier)

20 The varactor (or varicap) diode

Introduction

Many of the circuits for receivers and transmitters presented in this series rely upon the variable capacitor as a means of tuning. Another method of varying capacitance (without any moving parts) is provided by the *varactor diode*, sometimes called a varicap diode. This is a component which changes its capacitance as the voltage across it is varied.

The details

Figure 1 shows how a varactor diode might be connected to demonstrate its operation. Its symbol is that of an ordinary diode, with a capacitor symbol next to it. A variable voltage is applied across it in such a way that the diode is *reverse-biased*. This means that virtually no current passes through it – the positive voltage is applied to the cathode. Varactors are cheaper than variable capacitors, and they are tiny in comparison, very suitable for today's miniature circuits. If A and B were connected across the tuning coil in a simple receiver (with a series capacitor to block the DC from the battery reaching the coil), the tuning operation would be accomplished by turning the knob on the 10 kilohm potentiometer.

Varactors are available with different values, from less than 20 picofarad (pF) for VHF applications to 500 pF for medium-wave radios. They are

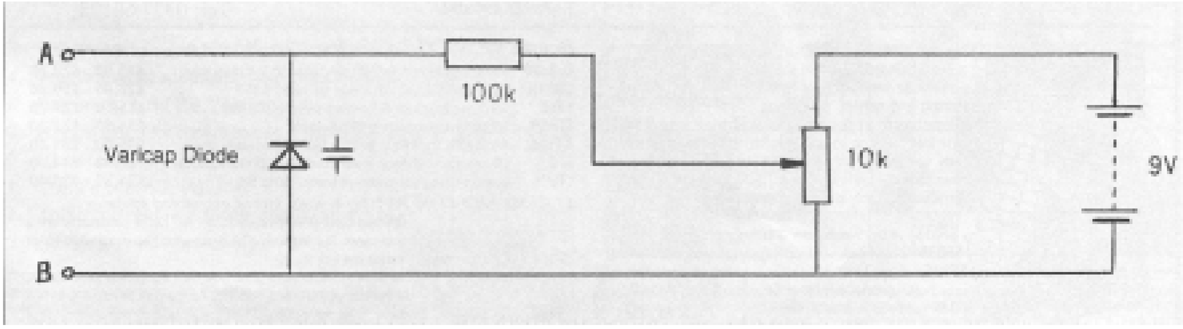

Figure 1 The capacitance of the varicap diode (between A and B) increases as the voltage is reduced, using the variable resistor

tuned usually by voltages between 2 V and 9 V. For a real application of varactors, you should consult the circuit diagram of the *Yearling* 20 metre receiver, elsewhere in this book.

In some circuit designs, several circuits are all tuned to the same frequency in order to improve the overall *selectivity* (the ability of the circuit to reject signals very close in frequency to the wanted signal). Special dual- and triple-varactors are available for circuits like this. Having been made at the same time from the same materials makes their individual characteristics virtually identical. Like all other diodes, they must be correctly wired into the circuit – their *polarity* is important.

Changes in temperature will cause the capacitance to change which, if it were part of an oscillator circuit, would cause the oscillator to drift – you would have to keep retuning the radio! This can be corrected by using a special integrated circuit called a *phase-locked loop* (PLL). Modern TV sets and satellite receivers use varactors and PLLs in this way.

Some useful varactor types

Type No.	*Tuning range*		*Description*
	pF/V	*pF/V*	
BB204B	42/2.0	15/12	Dual VHF
BB212	560/0.5	22/8	AM tuning
KV1235	450/2.0	30/8.5	Triple AM
KV1236	450/2.0	30/8.5	Dual AM
MV1404	120/2.0	9/10.0	HF tuning

21 A portable radio for medium waves

Introduction

The ZN415E integrated circuit (IC) can be used to make a very efficient AM portable MW broadcast radio with a built-in loudspeaker. Here's how!

The circuit

Figure 1 shows the circuit diagram of the portable radio. It's not as complicated as it may appear, especially after you have got started. L1 is a coil of wire mounted on a ferrite rod, acting as an aerial; VC1 is a variable capacitor which works, with L1, to tune in different stations. IC1 contains circuits of its own which boost the selected signal and it includes a *detector* which extracts the audio signal from the incoming RF signal. Earphones could be connected to the output of IC1 (between pins 4 and 5), but the output would not be powerful enough to drive a loudspeaker.

More sound

This is where IC2, an LM386 comes in. This is a small audio *power* amplifier which produces audio signals with enough power to drive a small

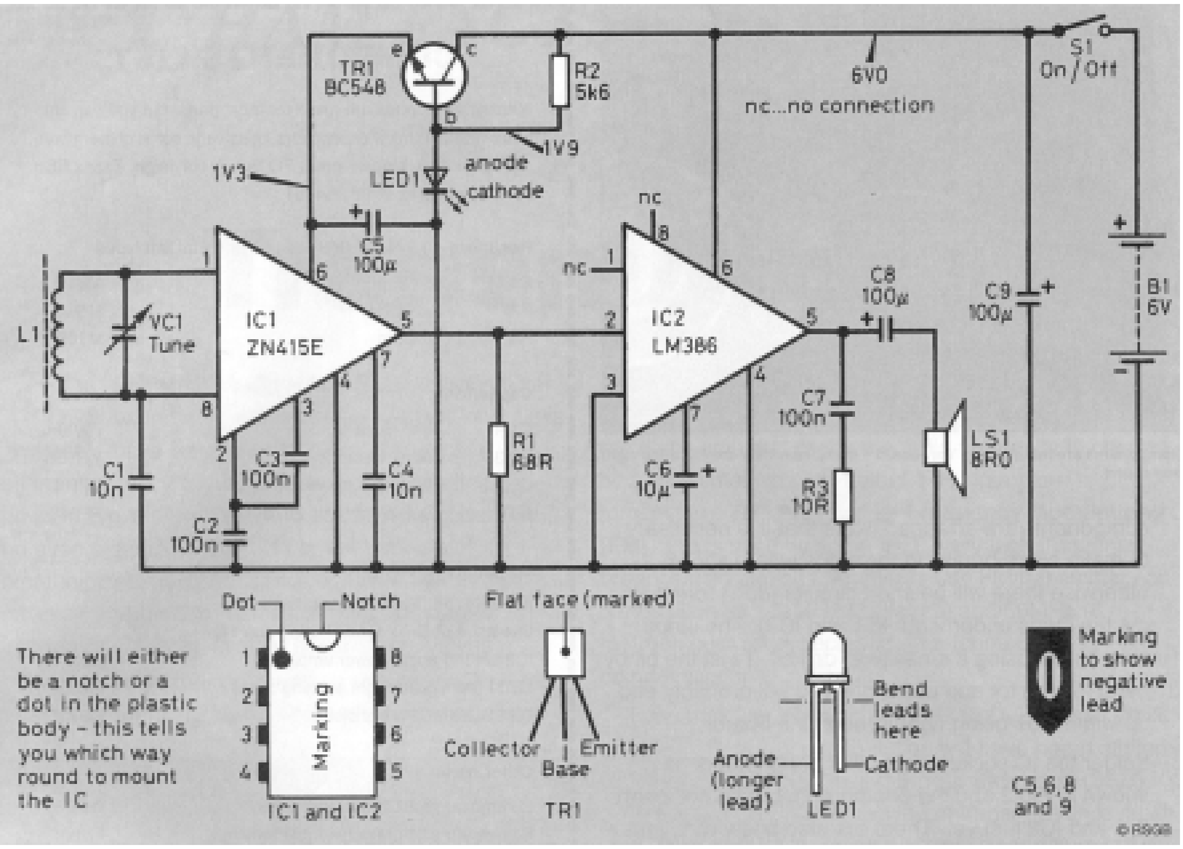
TR1
BC548
e
c
b
R2
5k6
6V0
nc...no connection
S1
On / Off
1V9
anode
cathode
1V3
LED1
C5
100µ
nc
nc
IC1
ZN415E
L1
VC1
Tune
C1
10n
C2
100n
C3
100n
C4
10n
R1
68R
IC2
LM386
C6
10µ
C7
100n
R3
10R
C8
100µ
LS1
8R0
C9
100µ
B1
6V
Dot
Notch
Marking
IC1 and IC2
Flat face (marked)
Collector
Base
Emitter
TR1
Bend leads here
Anode (longer lead)
Cathode
LED1
Marking to show negative lead
C5,6,8 and 9
©RSGB

There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC

Figure 1 The circuit diagram of our easy-to-build portable radio. Take care to mount the ICs and LED the correct way round

loudspeaker, LS1. The radio uses a 6 volt battery, which is made by connecting four 1.5 volt AA cells in series (4 × 1.5 V = 6 V) using a battery holder designed for this purpose. Although a 6 V supply is ideal for IC2, it is far too great for IC1, which needs only about 1.3 V. This lower voltage is provided from the 6 V supply by TR1 (an npn transistor), R2 and LED1 (a light-emitting diode). When current passes through an LED (see the description of the LED in this series) a reasonably constant voltage of 1.9 V appears between the anode and the cathode. Because of the voltage (0.6 V) that *always* exists between the base and emitter of a working transistor, the voltage on the emitter is about 1.9 V – 0.6 V = 1.3 V, and this is used as the power supply for IC1.

To keep the radio as simple as possible, no volume control has been fitted. Instead, you can use the directional properties of the ferrite rod aerial (see the information on ferrites in this book) to reduce the volume by rotating the set about a vertical axis using the handle provided.

Putting it all together

1. Start by covering the ferrite rod with Sellotape, or alternatively wrap a piece of paper tightly around it, and secure it with Sellotape. Then, with at least 2 metres of 24 SWG enamelled copper wire, wind 75 turns tightly around the rod. To be safe, leave about 50 mm of wire at the ends of the coil, then wrap the whole coil with Sellotape to hold the turns in place, leaving only the ends free. Then, using a small piece of sandpaper, remove the enamel from the last centimetre of each end of the coil.
2. Most of the components are mounted on a piece of Veroboard (the type with parallel copper strips on one side). The piece used on the prototype measured 32 holes by 10 strips, as **Figure 2** shows. Before you start fitting components, cut the copper strips as shown. It is easier to do it now than when the board is littered with components! The strips may be cut with a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger. Resist the temptation to use a hand drill – the idea is just to cut the copper, **not** to drill right through the board!
3. Solder the IC sockets and the other components on the board as shown in Figure 2. Make sure that the IC sockets are fitted with their notches towards the *top* of the board, as viewed in Figure 2. Do not insert the chips yet. Always keep the wires left over from cropping resistors and capacitors, they will come in handy at times like this: make the wire links that are clearly shown in Figure 2. Connect the electrolytic capacitors (C5, C6, C8 and C9), the transistor and the LED the correct way round; then check it again when you have done it!
4. Finally, solder lengths of stranded insulated wire to act as 'flying leads' for future connection to L1, VC1, LS1, S1 and the battery connector.
5. Apart from the battery holder, everything is mounted on the case *lid*. This makes assembly and testing much easier, and eases fault-finding if

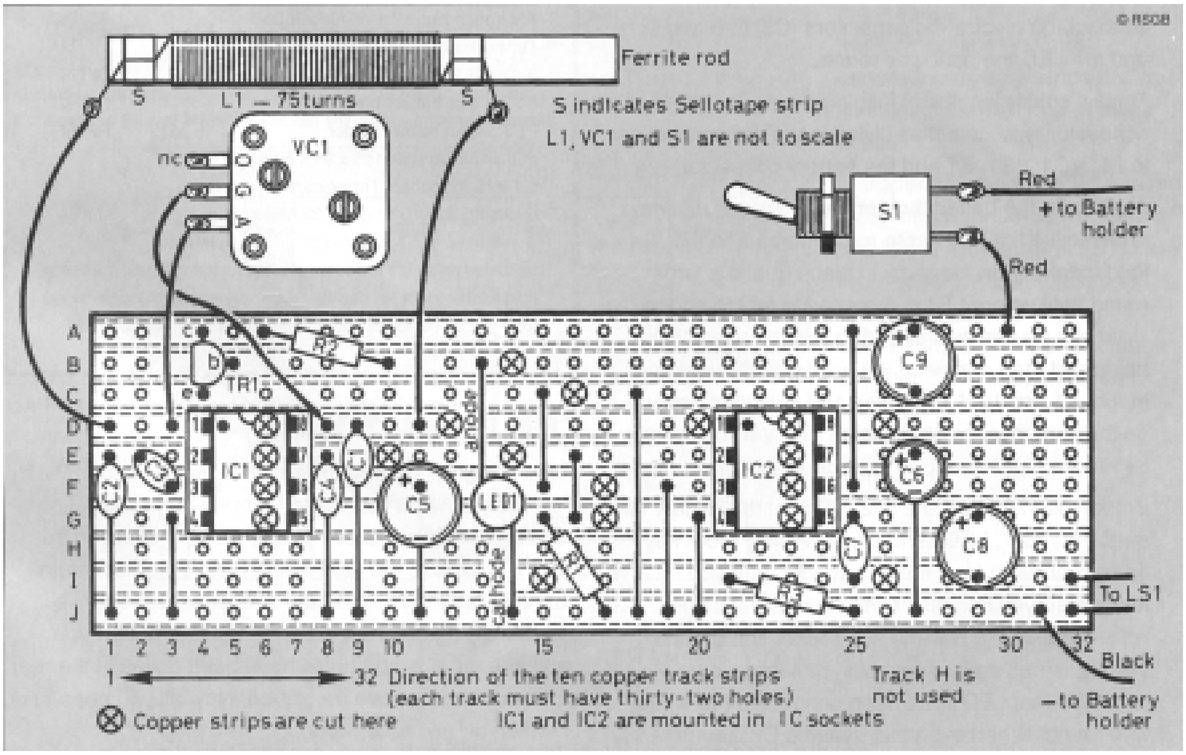
© R508
Ferrite rod
S
L1 — 75 turns
S
S indicates Sellotape strip
L1, VC1 and S1 are not to scale
VC1
nc
O
G
A
S1
Red
+ to Battery holder
Red
A
B
C
D
E
F
G
H
I
J
c
b
e
TR1
R2
IC1
C2
C3
C4
C1
C5
anode
LED1
cathode
R1
IC2
C7
R3
C9
C6
C8
To LS1
1 2 3 4 5 6 7 8 9 10 15 20 25 30 32
Black
− to Battery holder
1 ◄——► 32 Direction of the ten copper track strips (each track must have thirty-two holes)
Track H is not used
⊗ Copper strips are cut here
IC1 and IC2 are mounted in IC sockets

Figure 2 Veroboard layout for the Portable Radio. Make sure that all the wire links are included

the need arises! At the speaker position, make one large hole or a series of small holes to let the sound out. The ferrite rod may be stuck to the lid, as may the loudspeaker. Drill holes of the correct size to fit the particular types of variable capacitor (VC1) and switch (S1) that you are using. The Veroboard may be held in position by Blu-Tack or double-sided sticky tape.

6. Before inserting IC1 and IC2, connect the battery and switch S1 on. The LED should glow dimly (you may have to shield it with your hand in order to see it). If you have a test meter, check that there is about 1.3 V between pins 6 and 4 of IC1. If the reading is around 6 V or there is no glow, you may have connected the LED the wrong way round! When everything seems normal, switch off and disconnect the battery. Insert IC1 and IC2, making sure that the pins are straight and lie immediately above their corresponding holes in the sockets, and that the notches line up with the notches in the holders. Then push *gently* downwards on each IC in turn until the chip is firmly seated in its socket.
7. Switch on! By rotating the tuning capacitor, VC1, you should now be able to tune in many stations, rotating the radio to give you some volume control.

Final touches . . .

The handle was made with part of an old leather belt, secured to the case with 'number plate' nuts and bolts from Halfords. The loudspeaker grille is the lid from a *pot-pourri* container, and some extra holes were drilled in the back to improve the sound. See what you can find to finish off your radio!

Parts list

Maplin order codes are given for most of the parts, but you should get used to using the 'beg, borrow or steal' technique, or to use your ever-expanding junk box.

		Maplin code
Resistors: all 0.25 watt, 5% tolerance		
R1	68 ohms (Ω)	M68R
R2	5.6 kilohms (kΩ)	M5K6
R3	10 ohms (Ω)	M10R
Capacitors		
C1, C4	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic	BX00A
C2, C3, C7	100 nanofarads (nF) or 0.1 microfarad (μF) ceramic	YR75S
C5, C8, C9	100 microfarads (μF) electrolytic, at least 10 V	FF10L
C6	10 microfarads (μF) electrolytic, at least 25 V	FF04E
Semiconductors		
IC1	ZN415E radio chip	
IC2	LM386 audio power amplifier	UJ37S
LED1	3 mm green LED	WL33L
TR1	BC548 npn transistor	QB73Q
Additional items		
LS1	Miniature 8 ohm loudspeaker	WB08J
S1	Miniature SPST toggle switch	FH97F
Ferrite rod	Length approx. 100 mm	YG20W
	24 SWG enamelled copper wire	BL28F

Additional items (*continued*)

VC1	Tuning capacitor 140 to 300 picofarads (pF)	FT78K
	Tuning knob	FK41U
	8-pin DIL IC sockets (two required)	BL17T
	4 × AA-size battery holder (long)	HF94C
	PP3-type clip for battery holder	HF28F
	Plastic box approx. 158 × 95 × 54 mm	LH51F
	0.1 inch Veroboard, min. size 32 holes × 10 strips	JP46A

Plus

Stranded insulated conductor for flying leads
Multicore solder
Materials for handle and speaker grille
Double-sided sticky tape or Blu-Tack
Sellotape
Glue
Four AA-size 1.5 V batteries

22 The Colt 80 m receiver – Part 2

Introduction

In Part 1 we constructed the audio amplifier module for the system and tested it in a very simple way. If you did what was suggested and built the simple Crystal Set to use as a signal source, you will know just how well the amplifier works.

The case

Metal cases for the project are available from Maplin, telephone 01702 554 161 (code XB67). From the photograph on p. 72 you can see the way the components are mounted. The audio amplifier is seen at the top right of the base, to the right of the tuning capacitor VC1. The next in this series will deal with the variable-frequency oscillator (VFO) and VC1. The current part deals with preparing the case to receive the components.

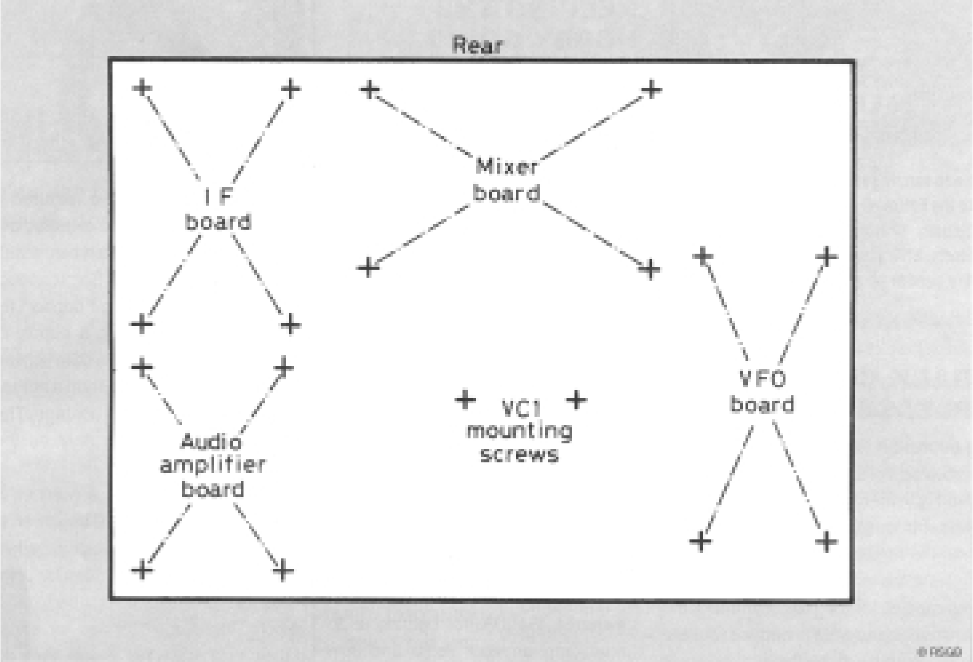
Rear
I F board
Mixer board
VFO board
Audio amplifier board
VC1 mounting screws
© RSGB

Figure 1 Fixing holes for each module are best measured from each printed circuit board or matrix board

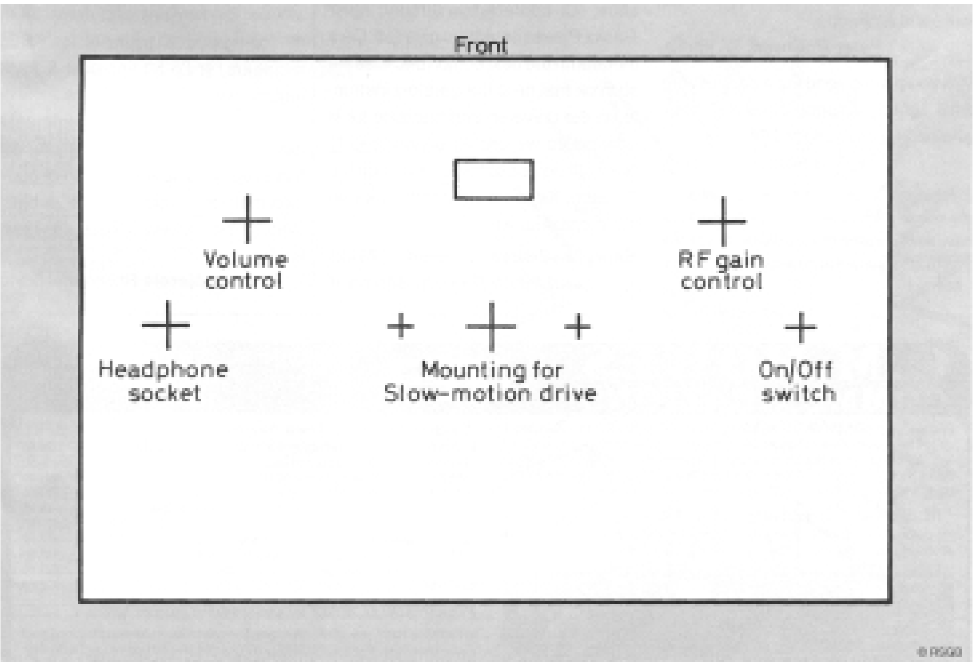
Front
Volume control
RF gain control
Headphone socket
Mounting for Slow-motion drive
On/Off switch
© RSGB

Figure 2 Position the slow motion drive to allow viewing of the tuning dial

Figures 1 and **2** show the markings for preparing the front panel and base. All the circuit boards and the tuning capacitor are mounted on the base using 10 mm stand-off pillars with 6BA bolts. The board locations are shown in Figure 1. The front panel control positions are shown in Figure 2, together with the small rectangular hole for viewing the tuning dial.

The best way to mark out the holes for the boards is to lie the boards flat on the base (before you've started soldering the components in) and marking the base through the holes in the boards. This minimises the scope for errors!

A *reduction drive* is used between the tuning knob and the capacitor shaft. This is simply a gear mechanism that slows down the capacitor shaft by a factor of six compared with the tuning knob, and makes tuning very much easier. The recommended variable capacitor also has a pulley wheel mounted on the shaft. Glued to this wheel will be a scale marked with frequency and is visible through the rectangular hole in the front panel.

The next part . . .

The variable-frequency oscillator and mixer will be added to the project.

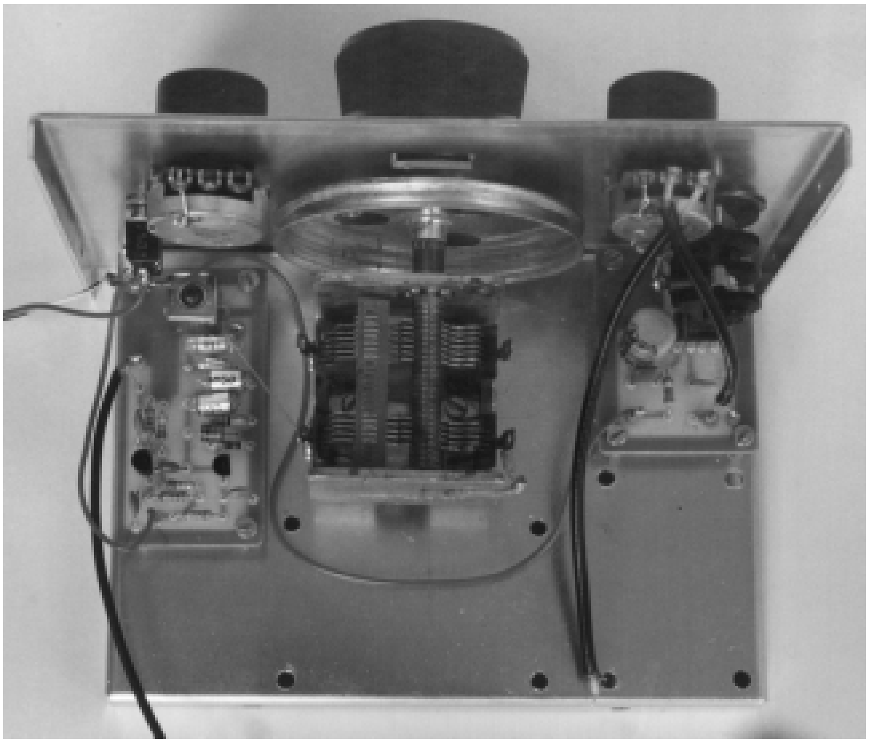

23 A simple transistor tester

Introduction

Although transistors aren't used as much as they were before integrated circuits came along, a transistor tester is still a useful piece of test equipment to have around the shack. This design is about the simplest possible and will produce an indication of whether a transistor is giving any current gain; this does not necessarily mean that the transistor is perfect but that it is working to a certain extent. This tester will **not** test field-effect transistors (FETs). If you buy a bag of transistors at a rally, this tester is useful for giving a yes/no indication of which ones go straight in the bin and which are kept for further use.

How it works

Figure 1 shows the simple wiring circuit. In order to explain the working of the circuit, a circuit diagram is shown in **Figure 2**, with the npn transistor, TR1, under test shown as part of the circuit. The pnp/npn selector switch, SW1, is omitted for clarity.

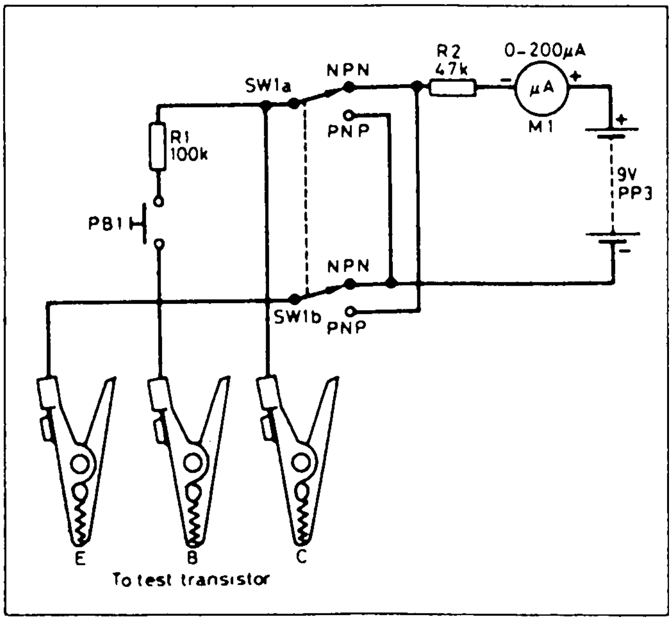

Figure 1 Circuit diagram of the transister tester

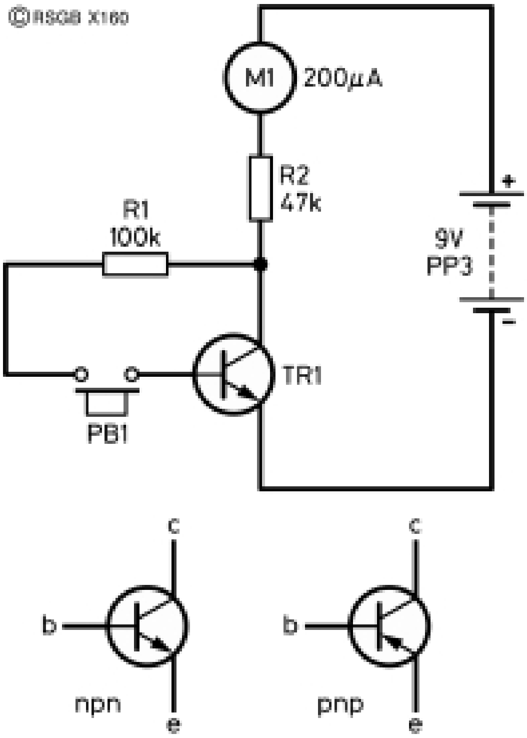

Figure 2 Testing an npn transistor

Any current that flows through the transistor, TR1, must flow from the battery, through the meter, M1, and through the protective resistor, R2. R2 prevents excessive current flowing through the meter and damaging it. Even if there is a short-circuit between emitter and collector, the maximum current that will flow is given by the simple equation

$$I = \frac{V}{R},$$

where I is the current flowing in amps,
V is the battery voltage, and
R is the total circuit resistance in ohms.

Putting in the correct values, gives

$$I = \frac{9}{47\ 000} = 0.000191 \text{ A, or 191 microamps } (\mu\text{A}).$$

The resistance of the meter, M1, will cut this down a little more, but it is within the indicating range (200 μA) of the meter.

As shown, with the push-button switch, PB1, *open*, a good transistor will not draw any current from the battery, and M1 will thus remain at zero. When PB1 is *pressed*, a *very small* current is injected into the base of the transistor. If the transistor is working, it will produce a much larger current between the collector and emitter, and this current will also flow through M1 and R2, giving a significant reading on M1, showing that all appears to be well. If an appreciable current flows when PB1 is *open*, then your transistor is suspect.

Don't be put off by the apparently complicated switch, SW1. It is there to allow the other type of transistor, the pnp type, to be tested. All it does is reverse the battery connections, so that the emitter goes to the *negative* battery terminal for testing an npn transistor, and to the positive terminal for a pnp type!

Most transistors in common use are of the npn type, which is why Figure 2 shows the testing of an npn type. The connections to the different transistor *encapsulations* (shapes) are given in any good component catalogue. Avoid the trial-and-error method to discover the connections to a transistor. This is unscientific, and can be very frustrating, particularly if the transistor is faulty in the first place!

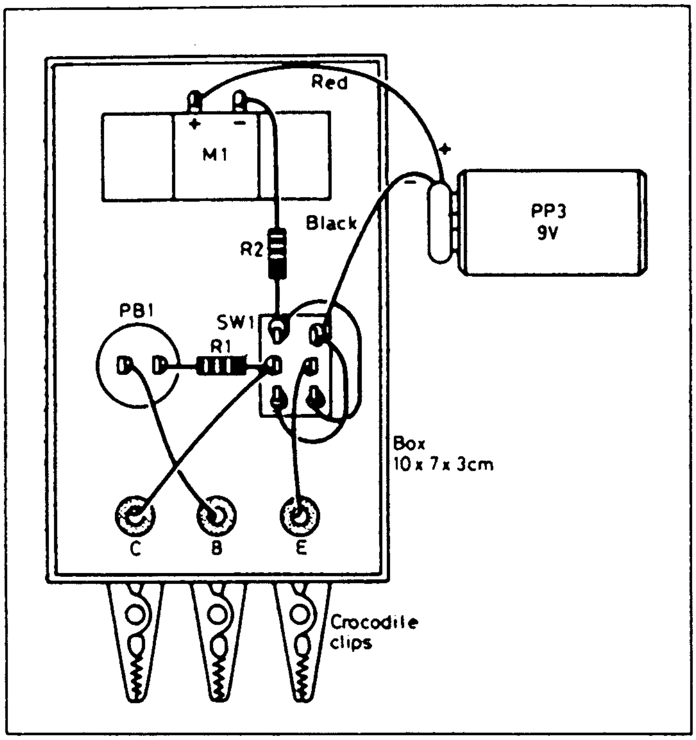

Figure 3 This shows how to wire up the tester

Construction

The unit can be built into a small plastic box, the components being soldered directly to the fixed terminals of the meter and the two switches; no circuit board is necessary! Use different colours of wire for the three test leads, and make sure you know which is which! Check the connections against the wiring diagram of **Figure 3**. When you are confident that all is correct, connect the battery and make sure there is no reading on the meter when nothing is connected to the crocodile clips!

Find any transistor for which you know the connections and the type (npn or pnp). Set the npn/pnp switch accordingly. Connect the three clips, making sure that they do not touch each other. A small reading on the meter at this stage means the transistor is suspect; a large reading means it is not working and should be thrown away! Press PB1 and watch the meter; a reading greater than half of full-scale indicates a good transistor. If it is less than half, you may have a transistor with ‘low gain’; it may be usable for non-critical applications, but if you are in any doubt – throw it away!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	47 kilohms (kΩ)

Additional items

PB1	Push-button switch
SW1	DPDT (double-pole double-throw) switch
M1	Micro-ammeter – *not* less than 200 μA full-scale deflection
	Crocodile clips (3 needed)
	PP3 battery and connector
	Plastic box about 10 × 7 × 3 cm

Common types of transistor

BC108, BC109, 2N2369A

These are small-signal npn types, used in audio amplifiers. They are in metal cases (called TO18) and have a tab next to the emitter lead. Common types have a B or C suffix (e.g. BC109C). The C suffix indicates a higher current gain than those with a B suffix. The 2N2369A is specially designed for radio use at high frequencies.

2N3703, BC212L, BCY71
These are pnp transistors, and so must be used with the collector and base *negative* with respect to the emitter. These three types are used in small amplifiers and audio oscillators. The first two have plastic encapsulations (TO92), while the BCY71 has a metal case (TO18).

BFY50, BFY51, BFY52
For slightly higher powers, these are ideal. They have been used in novice transmitters up to 600 milliwatts (mW). The TO5 case is a scaled-up version of the TO18 case. They are all npn types.

2N3055, 2N3773, TIP35C
These are high-power transistors in bigger encapsulations. The thick metal TO3 case of the 2N3055 and 2N3773 is designed to bolt to a heat sink, a large piece of metal which conducts the heat into the air more rapidly than the transistor itself can. The TIP35C is made of plastic but has a thick metal tab by which it, too, can be bolted to a heat sink.

24 An introduction to transmitters

Introduction

We usually think of a transmitter as being a 'black box'. However, that is the form a transmitter takes for our use on the amateur bands. Many electrical circuits are transmitters, even though transmitting may not be their primary function!

Anything that emits electromagnetic energy at *any* frequency is a transmitter, from radio at the low-frequency end of the spectrum, to gamma rays at the high-frequency end. We all know that a magnet will attract certain metals and that a comb rubbed on your coat sleeve will pick up small pieces of paper. The former is an example of the effect of a magnetic field, the latter of an electric field. Electromagnetic fields are combinations of both types of field, and are produced whenever an electromagnetic wave is transmitted.

What frequency?

Many everyday objects have a natural frequency of oscillation. This is called their *resonant frequency*. A wine glass will ring when struck gently; an empty wine bottle will sound if you blow across the top; a guitar string will vibrate when plucked. These are all examples of resonance, and the resonant frequencies will **not** change unless the objects themselves are changed physically in some way. These are resonances in *sound*; we are primarily interested in electrical resonances.

The basic electrical resonant circuit is the combination of an inductor (coil) and a capacitor, as shown in **Figure 1**.

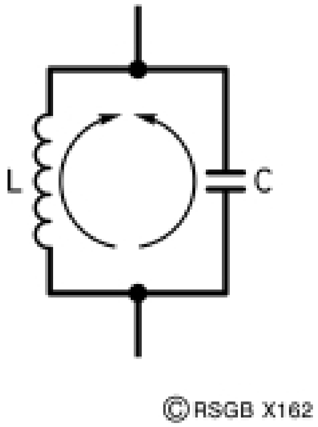

Figure 1 The basic electrical resonator. The energy in the circuit alternates between the inductor and the capacitor

A pulse of energy applied to this *tuned circuit* will make it *ring* (oscillate) at its resonant frequency. The energy in the circuit transfers between the inductor and the capacitor every cycle of the oscillation. Just like the wine glass, its oscillation dies away because it is losing some energy to its surroundings – it is transmitting! The frequency of the resonance depends on the values of L and C.

Keeping it going

If we want to keep the circuit oscillating, rather than having it die away, we must supply the circuit with just enough energy to replace the energy lost both by radiation and by losses in the circuit itself. Because of this, you will find in all oscillator circuits, a transistor, valve or FET working with the tuned circuit to provide this extra energy.

As it stands, of course, even with its transistor, our oscillator will not radiate very far. Connecting an aerial to it, and a Morse key to interrupt the power supply, it would become a very low-power CW transmitter. Add a couple more transistors to form a radio-frequency (RF) amplifier, and you have the basis of a simple low-power (QRP) transmitter.

Resonant circuits can also be made using quartz crystals; these work at the crystal frequency only, and this is marked on the crystal case.

A tiny spark transmitter

This is a simple piece of test gear that will increase your knowledge and understanding of resonance. You can use it to estimate the resonant frequency of most of the inductor/capacitor (LC) tuned circuits that you build. The circuit is shown in **Figure 2**. It operates around a relay. Any relay that operates from a 6 V to 9 V source and has contacts which are normally closed (i.e. closed when the battery is not connected). Fit the relay, a toggle switch and the battery in a metal box, and connected up as shown in the diagram. Some foam rubber inside the box may help to reduce the escaping noise of the relay. A small hole in the side of the box enables the 2-turn loop to emerge. This should be about 40 mm diameter, made with insulated wire. Switch on; there should be a loud buzzing noise from the relay. If not, you have probably chosen the wrong contacts on the relay!

When it is working, bring the loop close to the aerial of a radio – it should produce a loud noise from the speaker!

How it works

When you switch on, current flows through the relay contacts and through the relay coil. The relay operates and opens the contacts, causing the relay to 'drop out'. When it does, the circuit is completed again and the contacts are opened, and the cycle repeats. Each time the relay contacts open, there is a small spark between them, causing very rapid current surges through the wire loop. This makes the loop transmit RF energy, very briefly. In the early days of radio, this type of circuit was known as a *spark transmitter.*

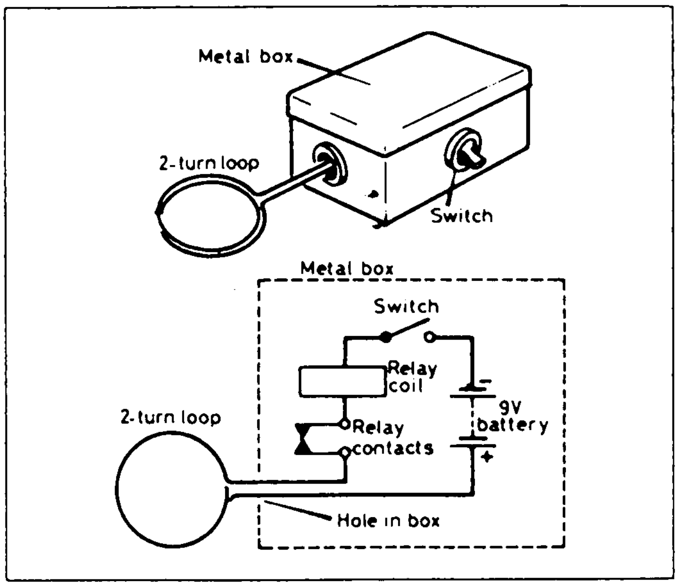

Figure 2 The current path is interrupted when relay is energised as shown above

Make a tuned circuit

Use a discarded toilet-roll centre, and wind about 10 turns of enamelled copper wire round it, keeping each turn close to the next. Scrape off the enamel for about 1 cm at each end, and solder a 100 picofarad (pF) capacitor (or a variable capacitor of about the same value) between the ends. The resonant frequency should be about 10 MHz. If you have used fewer turns or a smaller capacitor, the frequency will be higher.

Measuring the resonant frequency

Set up the buzzer as shown in **Figure 3**, with the loop around one end of your coil. Then make a similar loop, solder it to the end of a piece of coaxial cable going to the aerial socket of a calibrated receiver. Set the buzzer going, tune the receiver around 10 MHz, and search for the maximum noise level from the speaker. When you have found it, move the two loops as far away as possible from the main coil. This is called reducing the *coupling* between the coils, and it may result in a slightly different, but more accurate, resonant frequency.

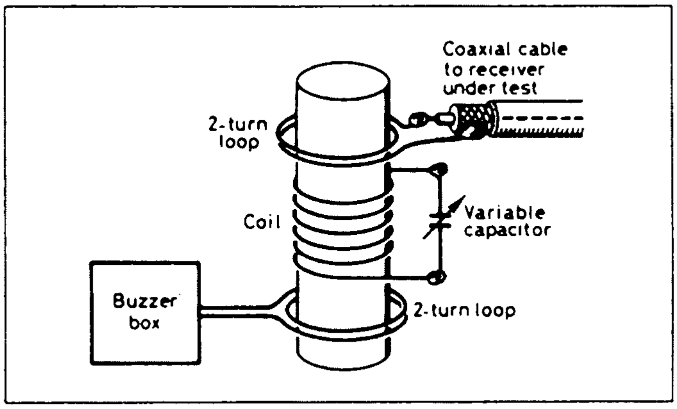

Figure 3 Experiment with the spacing between the loop and the tuned circuit

Parts list

Any small relay which operates between 6 V and 9 V
Metal box – do not use a plastic box!
9 volt battery and connector
On/off (SPST) toggle switch
Plastic foam, as required

25 The Colt 80 m receiver – Part 3

Introduction

In Part 2 we marked out the case ready for installation of the modules when they are completed. In this part, the building of the variable-frequency oscillator (VFO) and mixer will be described. This will produce a type of receiver known as *direct-conversion*, because it converts the radio-frequency (RF) signal directly into an audio-frequency (AF) signal which we can hear in a loudspeaker after amplification. A block diagram of the system is shown in **Figure 1**. The modifications needed to make a full superheterodyne receiver will be left until later.

The direct-conversion receiver covers the 80 metre amateur band and will receive both Morse (CW) and speech (SSB) signals. The audio amplifier was covered in Part 1, so your *Colt* is rapidly taking shape! By the time your construction has reached the end of this part, you will have a receiver ready to use, even if the project is not yet complete!

The direct conversion process

Like most things in radio, the principles of direct conversion are not difficult. From the aerial, the signal we want to hear is selected by the *tuned filter*, which rejects the signals we don't want. The signal then enters the

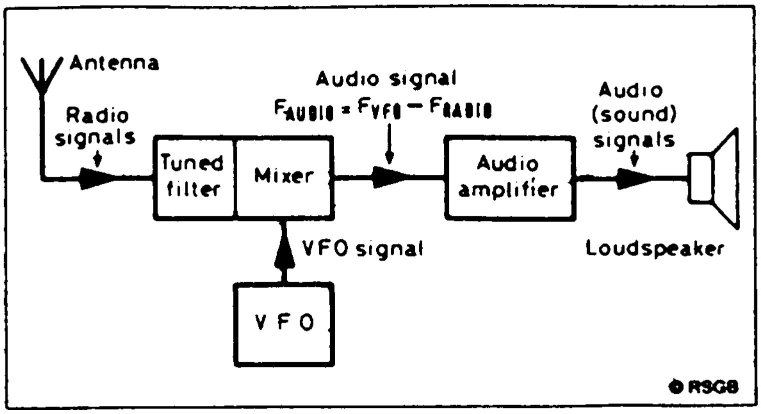
Antenna
Radio signals
Audio signal
$F_{AUDIO} = F_{VFO} - F_{RADIO}$
Audio (sound) signals
Tuned filter
Mixer
Audio amplifier
VFO signal
Loudspeaker
VFO
© RSGB

Figure 1 Stages of direct-conversion receiver

mixer, along with the signal from the *VFO*. The VFO produces a sine wave whose frequency can be varied across the whole of the 80 metre amateur band (3.5–3.8 MHz), by turning the knob on the tuning capacitor, VC1. The mixer produces, at its output, *two* signals; one signal is at the frequency of the **sum** of the signal and VFO frequencies, the other is at the **difference** of the two frequencies. It is the latter that we want. Let's look at the numbers involved. If the signal is at 3650 kHz and the VFO is at 3651 kHz, then the sum frequency is 7301 kHz, and the difference frequency is 1 kHz. If we feed the output of the mixer into our audio amplifier, the 7301 kHz signal is automatically removed (it is far too high to be considered an audio signal!) and the resulting 1 kHz signal is amplified and fed to the speaker, producing a note which we can hear!

Building the VFO

Figure 2 shows the circuit of the VFO. It is a tried and tested circuit, and should work first time. It uses a *field-effect transistor* (FET) for TR1, the oscillator itself. RFC is a *radio-frequency choke*, a coil of wire which will pass a direct current (DC) but which will prevent radio-frequency (RF) signals getting through.

All VFOs have a tuned circuit which, in this case, is formed by the coil L1 and the capacitors C1 and VC1. The frequency will also be affected to some extent by C2, C3, C4 and C5. Transistor TR2 is an *emitter follower*, a stage which gives no voltage gain but provides a good *buffer* stage, isolating the VFO from the effects of the stages that follow it. When building a VFO, the

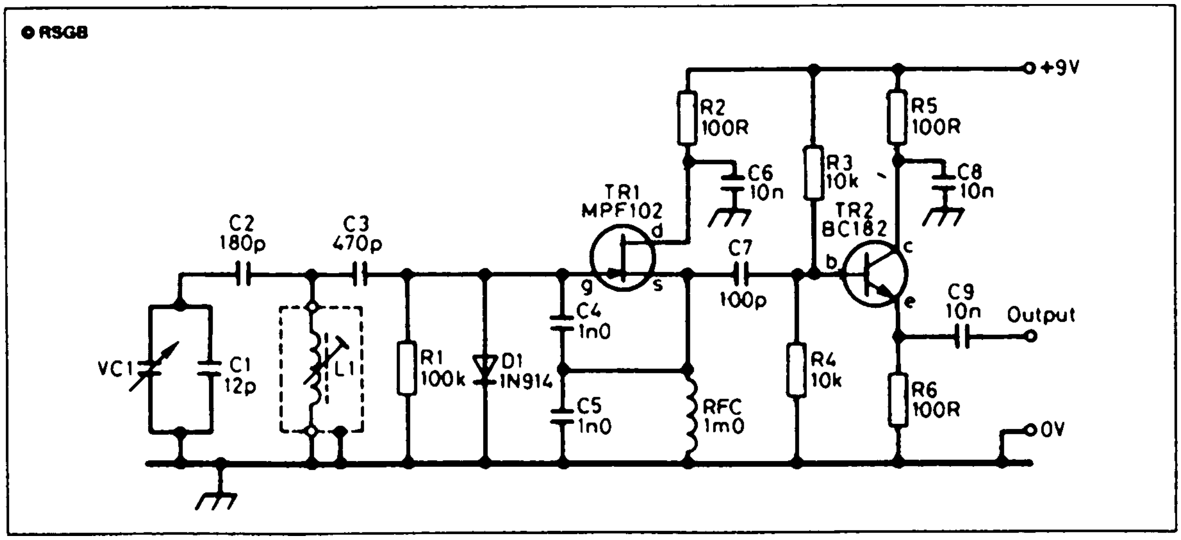

Figure 2 The variable-frequency oscillator uses a field-effect transistor (FET)

parts must be securely mounted. If components move, so does the frequency! At worst, the oscillation will become unstable and the VFO will be useless. Keep the component leads as short as possible – this improves their mechanical stability as well as their electrical stability!

Mount the parts on the printed-circuit board (PCB) or matrix board and, when completed, the VFO should look like the one shown on the left in the photograph on p. 72. Make sure that TR1, TR2 and D1 are the right way round.

On completion, check the component positions then mount it in the case as shown in the photograph, to the left of the tuning capacitor when viewed from the rear. The VFO coil, L1, will need some adjustment, but that will have to wait until the mixer is built. Connections to the other boards are made with screened cable.

The mixer board

So far, we have an audio-frequency (AF) amplifier and a VFO; the addition of a mixer board gives us a complete direct-conversion receiver for 80 m. The mixer circuit diagram is shown in **Figure 3**. Let's follow the signal path.

- From the aerial, the RF signal goes to the gain control potentiometer, RV1. This reduces very strong signals, to prevent them overloading the mixer.
- To select the required band of signals, a *bandpass* filter is made up of RF transformers T1 and T2; these are tuned by C1 and C3, and are coupled together by C2.

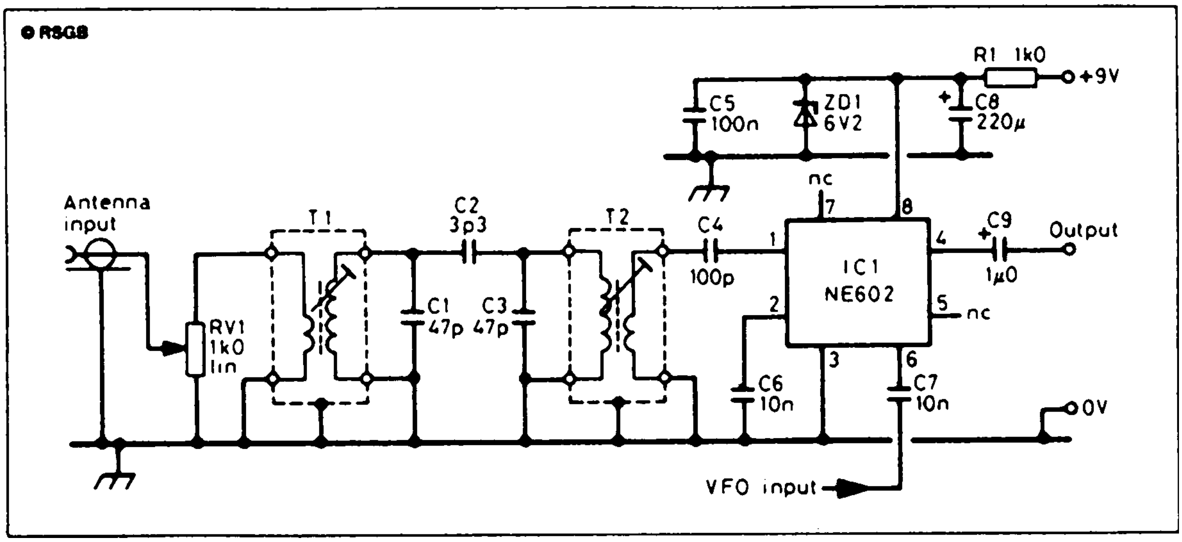

Figure 3 The mixer board has a bandpass filter and stabilised supply

- After the filter, the signal is coupled into the integrated circuit (IC) mixer type NE602, by the capacitor C4. Capacitors C5 and C8 *decouple* the supply line, to prevent unwanted signals on the supply from disturbing the VFO operation. The use of the term *decouple* is exactly the opposite of *couple*; when two circuits are *coupled* together, the signal passes from one to the other; when two circuits are *decoupled*, signals **cannot** pass from one to the other.
- The NE602 works with a 6 V supply; it is produced here from the 9 V supply by the *Zener diode* ZD1 and the resistor R1. ZD1 operates at 6.2 V, and gives a steady output for the mixer. The audio output from the mixer appears at pin 4 or IC1. This is taken to the audio amplifier board via C9 and the volume control (see Part 1).

Care must still be taken to insert some components the right way round. These are the electrolytic capacitors, C8 and C9, the Zener diode, ZD1, and the integrated circuit, IC1. Check all component positions and make sure all your soldered joints are bright and shiny.

Putting it together

The interboard wiring, shown in **Figure 4**, uses screened cable; ideally, this should be thin coaxial cable, but screened microphone cable is suitable. The diagram shows how the two controls, the RF Gain and Volume, are connected to the boards. The leads marked '+9 V' are all connected to the battery supply via a miniature on/off toggle switch. Double check all connections before connecting the battery.

Setting the VFO

Very little adjustment is needed to get the receiver going. Firstly, the VFO must be adjusted to cover the required band, in this case 3.5–3.8 MHz.

If you have a frequency counter, connect it to the output of the VFO. If you haven't, read this part anyway so you understand the process, then another means of setting the VFO will be given especially for you! Rotate VC1 until the vanes are fully meshed. Very carefully, adjust the core of the VFO coil (L1) with a plastic trimming tool so that the frequency approaches and settles at 3.500 MHz. When you rotate VC1, the frequency should increase to **at least** 3.800 MHz at the far end of its travel.

In the absence of a frequency counter, borrow a communications receiver, set it for SSB reception (USB or LSB) on exactly 3.500 MHz. Set VC1 with the vanes fully meshed and turn the core of L1 in both directions until you hear a whistle in the communications receiver. Rotate the core so that the whistle reduces in frequency. It will eventually fade out at around 200 Hz;

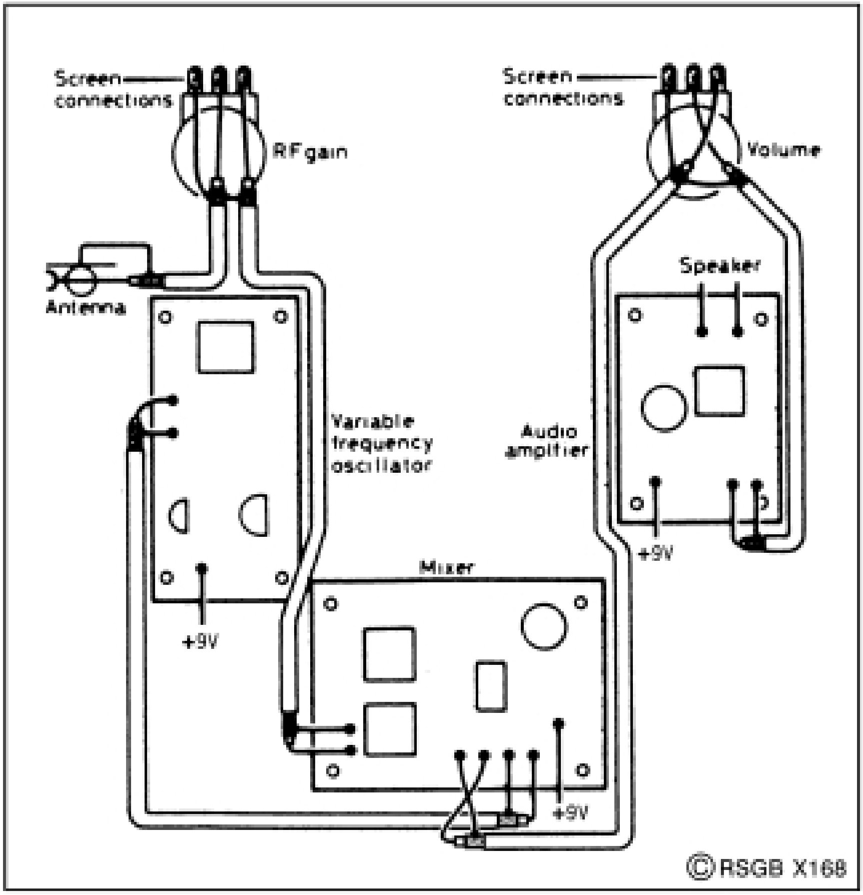
Screen connections
RF gain
Antenna
Variable frequency oscillator
+9V
Screen connections
Volume
Speaker
Audio amplifier
+9V
Mixer
+9V
©RSGB X168

Figure 4 Make sure that interconnections between the boards are correct, including cable screens

turn the core a little further and then leave it at that position. Rotate VC1 right to the other end of its travel, and search for the whistle with the tuning knob of the communications receiver. Check that the VFO frequency is at least 3.800 MHz.

Then, whichever method you are using for frequency measurement, mark the dial with frequency steps of 50 kHz. Setting and calibration are finished!

Setting the mixer

Again, there are various ways of doing this. If you have a signal generator, inject a signal at a frequency within the 80 m band, and adjust the cores of T1 and T2 sequentially for maximum output.

If you haven't a signal generator, connect an aerial to the mixer input, set the RF Gain to maximum (fully clockwise), and find a consistent signal. Adjust the volume control to a comfortable level. Rotating the core of T2 with your trimming tool, maximise the output. Then do the same with T1, although this will have much less effect. Find another station, and check that the positions of the cores aren't too different for a maximum signal.

You may find that your receiver benefits from the insertion of an aerial tuning unit (ATU) between the aerial and the input, to compensate for the impedance of your aerial not being 50 Ω. A design for such an ATU is presented in another part of this series. If the signals are still weak, connect the ATU to the junction of C1 and C2 via a 100 pF capacitor.

Try listening!

Remember that 80 m is a variable band. During daylight hours, your will hear Morse signals at the lower end of the band, and some British and closer continental stations between 3.7 and 3.8 MHz. In the evenings, stations up to 1000 miles away should be heard. Look for Novices around 3.7 MHz!

Parts list – VFO board

Resistors: all 0.25 watt, 5% tolerance		
R1	100 kilohms (kΩ)	
R2, R5, R6	100 ohms (Ω)	
R3, R4	10 kilohms (kΩ)	
RV1	1 kilohm (kΩ) linear	
Capacitors		
C1	12 picofarads (pF) polystyrene	
C2	100 picofarads (pF) polystyrene	
C3	470 picofarads (pF) polystyrene	
C4, C5	1 nanofarad (nF) polystyrene	
C6, C8, C9	10 nanofarads (nF) polystyrene	
C7	100 picofarads (pF) min. ceramic	
VC1	140 + 140 picofarads (pF) variable	
Semiconductors		
TR1	MPF102	FET
TR2	BC182	npn
D1	1N914	silicon
Inductors		
L1	Toko KANK3334	
RFC	1 mH	RF choke

Parts list – mixer board

Resistors: 0.25 watt, 5% tolerance

R1	1 kilohm (kΩ)
VR1	1 kilohm (kΩ) potentiometer (linear)

Capacitors

C1, C3	47 picofarads (pF) min. ceramic
C2	3.3 picofarads (pF) min. ceramic
C4	100 picofarads (pF) min. ceramic
C5	100 nanofarads (nF) min. ceramic
C6, C7	10 nanofarads (nF) min. ceramic
C8	220 microfarads (μF) electrolytic 16 V
C9	1 microfarad (μF) electrolytic 16 V

Integrated circuit

IC1	Philips NE602 or NE602A

Additional items

T1/T2	Toko KANK3333
	On/off switch
	Miniature toggle switch

The next part . . .

The IF amplifier and the Beat-Frequency Oscillator will be added to convert the *Colt* into a superheterodyne receiver.

26 A two-way Morse practice system

Introduction

It is said that the pleasure of sending and receiving Morse code more than compensates for the learning effort needed. Transmitters for Morse code can be much simpler than those needed for any of the speech modes.

Many people choose to learn sending and receiving on their own. It can be much more fun if you have someone to learn with you, and it is for this reason that the following project arose. It comprises just one small circuit, and you build an identical circuit for each of the people who want to learn with you. All the individual circuits are connected with their output leads in *parallel*, as the two-way circuit of **Figure 1** shows.

A simple circuit

This diagram shows two identical circuits, as would be used if two people wanted to learn together. The circuit centres around our old friend the NE555 integrated circuit (IC). As it is connected here, it works as an *astable multivibrator*, a daunting name for what is essentially an oscillator. Each circuit is self-contained, having its own battery, Morse key, sounder and

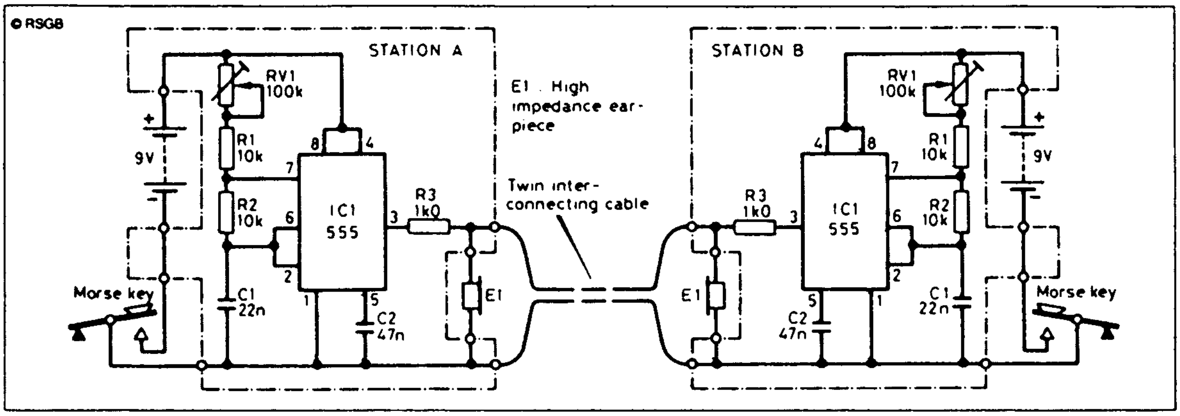
© RSGB
STATION A
STATION B
RV1 100k
R1 10k
R2 10k
9V
IC1 555
R3 1kΩ
C1 22n
C2 47n
E1
Morse key
E1 . High impedance ear-piece
Twin inter-connecting cable

Figure 1 The circuit diagram shows two stations, but you could connect several more if needed

plastic box. The Morse key makes and breaks the power supply to the circuit, thus turning on and off the note emitted by the earpiece. The frequency of the note is controlled by the variable resistor RV1, resistors R1 and R2, and the capacitor C1. The range given by RV1 is approximately from 500 Hz to 2200 Hz.

R3 protects the output of the IC against accidental shorting either to the 0 V or to the 9 V supply rails. For best results, a crystal earpiece or high-impedance headphones should be used. You could try Walkman-type headphones, but the volume may be too low for you.

The units may be interconnected with thin twin cable (the sort used for wiring doorbells) and can be comfortably separated by 25 m, more than enough to communicate between rooms, or even with a neighbour!

In use

When an operator presses his key, the note is heard by himself and by everyone else who is connected. If more than one key is pressed at the same time, two or more notes are heard by everyone! It helps if each person adjusts his own potentiometer, RV1, to give a note which is different from the others. When one operator stops sending, the other can start immediately, without pausing to change from transmit to receive. In practice, this technique is known as *full break-in*.

The current drawn by the circuit is only 10 milli-amps (mA) when the key is down, meaning that the life of a typical PP3 battery will be virtually its shelf-life! No switch is needed, because the key acts as the switch. The current drain will be even less if the NE555 is replaced by an ICM7555 IC. It has exactly the same pin connections as the NE555, so no circuit modifications are needed.

The circuit board

The circuit is constructed easily on a small piece of Veroboard of the copper-strip type. **Figure 2** shows the layout of the prototype, the board measuring 18 holes by 10 strips. **Be aware that there is no row 'I' when you are transferring mental images of where the parts are to real positions on the board!** Break the copper tracks where shown with a 3 mm (⅛ inch) twist drill, rotated carefully between thumb and forefinger. Hold the board up to a bright light to make sure that the tracks are completely broken and that there are no fragments of copper swarf shorting adjacent tracks. Then, solder in the wire links shown in Figure 2, followed by the resistors, RV1, and the capacitors, C1 and C2. Solder the battery connector leads to tracks A and G, G being positive. Use insulated wire to connect from the board to the jack sockets used as connectors for the key and the earphones. If you use

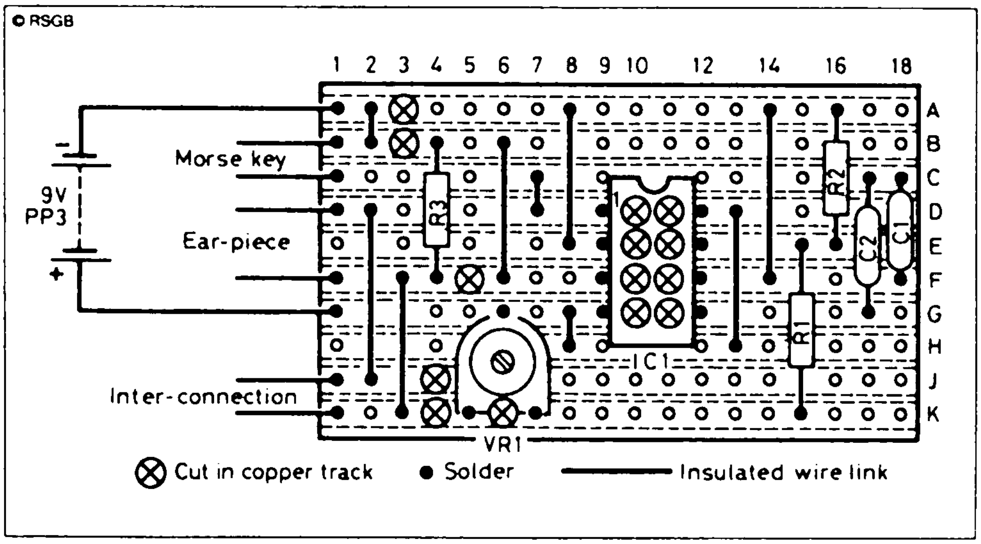
© RSGB
1 2 3 4 5 6 7 8 9 10 12 14 16 18
A B C D E F G H J K
9V PP3
−
+
Morse key
Ear-piece
Inter-connection
R3
R2
R1
C2
C1
1
IC1
VR1
Cut in copper track
Solder
Insulated wire link

Figure 2 The Morse Duet is an easy project to build on stripboard (Veroboard)

different sized jack sockets (i.e. 2.5 mm and 3.5 mm) you won't plug the earphones into the key socket! A two-screw terminal strip is useful to connect the cable runs between operators.

The key

The circuit is designed to be used with a *straight key* (one that moves up and down). It is a requirement of the UK Morse test that a straight key is used, so it is very sensible to learn sending with a straight key before you try anything more complex! You will find many to choose from at rallies.

Parts list

For each board, you will need:		**Maplin order codes**
Resistors: all 0.25 watt, 5% tolerance		
R1, R2	10 kilohms (kΩ)	M10K
R3	1 kilohm (kΩ)	M1K0
RV1	100 kilohm (kΩ) min. preset (horizontal mounting)	UH06G
Capacitors		
C1	22 nanofarads (nF) ceramic, 25 V	WX78K
C2	47 nanofarads (nF) ceramic, 25 V	RA47B

Integrated circuit	
IC1 NE555	QH66W
Additional items	
8-pin DIL socket	BL17T
Jack plug and socket 2.5 mm	HF76H, HF78K
Jack plug and socket 3.5 mm	HF80B, HF82D
Battery type PP3	
Battery connector	HF28F
Plastic box with lid 114 × 76 × 38 mm	LH14Q
Veroboard 0.1 inch with copper strips 18 holes × 10 strips	JP46A
Terminal strip (two section)	FE78K
Crystal earpiece	LB25C
Morse key	

27 The Colt 80 m receiver – Part 4

Introduction

In Parts 1 to 3 of this series, the design of an 80 metre direct-conversion receiver has been described. In this final part, we are going to change the circuit to operate as a superheterodyne receiver, or *superhet*. Most radio receivers are superhets, and they change the incoming signal to another frequency, known as the *intermediate frequency*, or IF, before producing an audio signal. The use of a superhet in a good receiver is mandated by the requirements for good *sensitivity* and *selectivity.*

Sensitivity and selectivity

Figure 1 shows the block diagram of the receiver. If you look closely, you will see that it is two similar circuits, one after the other. The incoming

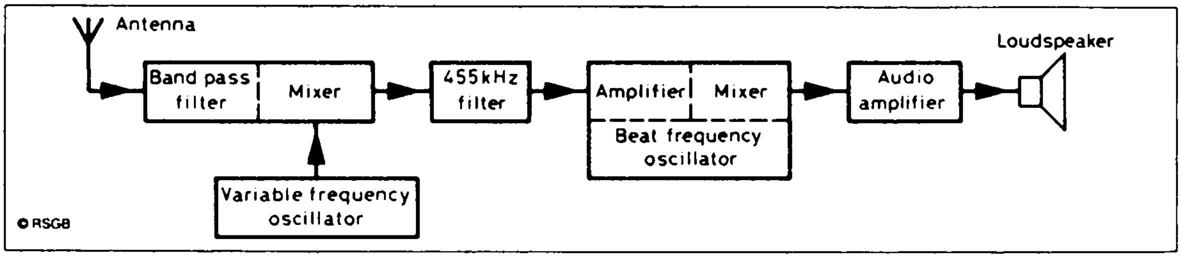

Figure 1 This is how the different stages of the Colt go together to make a superhet receiver

signal is filtered and fed to a mixer where it is combined (mixed) with the signal from the variable-frequency oscillator (VFO). This oscillator operates at a frequency which is 455 kHz higher than the incoming signal from the aerial, and the mixer output is therefore at a frequency of 455 kHz. If you are not sure about this, please refer to the section 'The direct conversion process' in Part 3 of this project. This new frequency is called the *intermediate frequency*, or IF. This frequency doesn't change; the tuning is accomplished by the VFO, and the mixer output is *always* at 455 kHz. The extraction of the audio signal from the IF signal is identical with the direct conversion process which is used in your existing receiver.

This may seem a long-winded way of doing things, but it has its advantages. A receiver must have a good *sensitivity*, or gain, so that it can receive very weak signals. In very general terms, it is easier to handle low-frequency signals than it is to handle high-frequency signals. We are changing our signal frequency from around 3.6 MHz down to 0.455 MHz (455 kHz), which is much lower and can be filtered and amplified relatively easily. A second advantage is that it is easier to provide gain at a fixed frequency than at a variable frequency. Remember that the IF is fixed, and providing gain is, again, relatively simple.

Our receiver also needs good *selectivity*, the ability to separate (or select) one station from another very close to it in frequency. This requirement is significantly simplified by the fact that the IF is fixed, and a good filter in the IF circuits can do wonders for the rejection of adjacent-frequency stations! Several stages of IF amplification and filtering are possible in more adventurous designs.

The filtered signal at 455 kHz passes to another mixer which has an associated oscillator, usually called a *beat frequency oscillator* (BFO). When receiving CW (Morse) signals, the BFO is usually tuned about 1 kHz above or below the IF (i.e. at 454 or 456 kHz) to produce a 1 kHz beat note as the audio signal. This signal is then amplified and fed to a loudspeaker or headphones.

The circuit

Figure 2 shows the circuit of the IF section shown in the photograph on p. 94; this section is added to the existing circuit to make it a superhet. The existing mixer board (described in Part 3) is used as the first mixer. Between it and the audio amplifier is connected the new IF board.

The signal output from the first mixer (at 455 kHz IF) is fed into a *crystal filter*, the most expensive part in the whole receiver. It provides the selectivity which makes the Colt such a good receiver. Another NE602 mixer/oscillator chip follows the filter. The oscillator section is controlled by the tuned circuit in T1, the frequency of which can be altered by rotating the core inside the coil. Once set, it remains fixed.

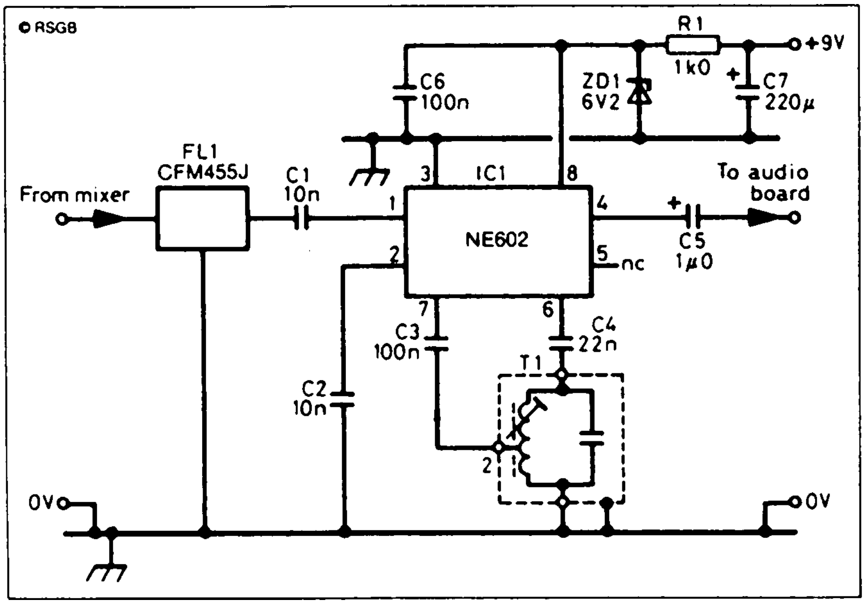

Figure 2 The IF board is connected between the mixer and audio amplifier

At this stage you should have built the first mixer and the audio amplifier, and proved that they *both* work by using the circuit as a direct-conversion receiver. When you have finished constructing and checking the IF board, you will need to add it to your existing circuits.

Adjusting and testing

There are three pairs of connections to the IF board – the 9 V supply leads, the input leads and the output leads. The IF board should be mounted on the metal baseplate, along with the other boards. **In making the following connections, make absolutely sure that the braid of each piece of coaxial**

cable is soldered to the correct connection on each board. The same applies to the polarity of the 9 V leads. Disconnect the screened lead which presently goes from the first mixer board to the AF amplifier at the amplifier end. Connect it instead to the *input* of the IF board. Connect a new piece of screened cable from the IF output to the vacated AF input. Now connect the 9 V supply leads to the same points as the supply leads from the other boards. Check these newly made connections.

You should have confidence at this point that things should be right. After all, you *have* tested the direct-conversion process and you *know* it works. All that you are now testing is the IF board. This is the attraction of building a receiver in modules, working from the speaker backwards, and testing as you go!

The VFO and BFO need to be correctly adjusted. Mesh the vanes of the tuning capacitor and, using the same frequency measurement method as you did originally, set the VFO frequency to 3.955 MHz (which is 3.500 MHz + 0.455 MHz, if you hadn't guessed!). The BFO can be set using a frequency counter, but it is just as good to set it by listening to SSB or CW signals. A high-pitched hissing sound should be heard in the speaker. As you rotate the core, the pitch should reduce, go through a minimum, then increase again. Set the core at the minimum pitch position. You may want to readjust the two cores a little as your listening skills improve but, once you are happy, they will never need to be altered again!

In conclusion . . .

You should now have built a superhet receiver capable of excellent results. It uses the same type of circuit as that found in far more advanced receivers.

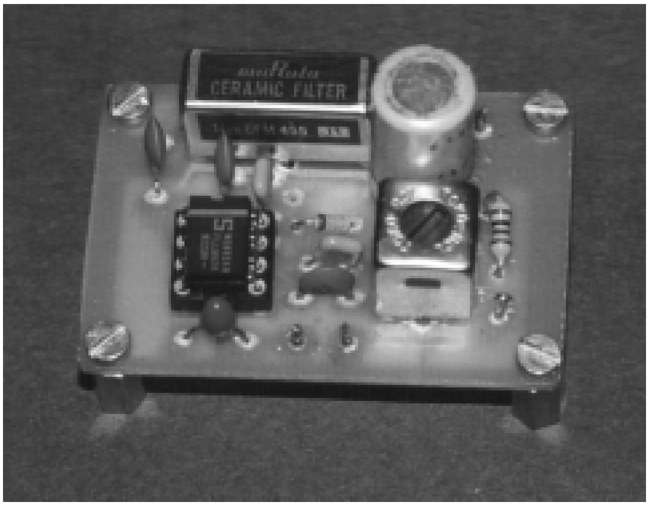

The superhet is far more sensitive than the direct-conversion type, and can weed out those elusive DX stations. You may have found that a station will appear at two places on the dial of the direct-conversion receiver; you will have no such problem with the superhet.

Parts list

Resistor

R1	1 kilohm (kΩ), 0.25 watt, 5% tolerance

Capacitors

C1, C2	10 nanofarads (nF) min. ceramic
C3, C6	100 nanofarads (nF) min. ceramic
C4	22 nanofarads (nF) min. ceramic
C5	1 microfarad (μF) 16 V electrolytic
C7	220 microfarad (μF) 16 V electrolytic

Integrated circuit

IC1	NE602

Additional items

ZD1	Zener, 6.2 V 0.5 watt
FL1	Crystal filter, Murata CFM455J
T1	Tuned inductor, Toko YHCS11100AC

28 A simple crystal set

Introduction

There are many designs of crystal set – they are all ‘simple’, and even the most seasoned radio amateur will build one of these every so often because it is something that never ceases to amaze! Although using modern components, the design is a period piece!

Design and construction

The circuit diagram of **Figure 1** shows the simplicity of the circuit. The prototype was constructed on a wooden baseboard with an aluminium front panel, although this could be made of wood if you prefer.

Winding the coil always causes the most groans but when finished, gives the most satisfaction. The construction of the coil is detailed in **Figure 2**. A discarded toilet roll centre is an ideal *former* on which to wind the coil. The coil will take up about 7 cm of the length of the tube, so assuming the tube is about 11 cm long, the ends of the coil will be about 2 cm in from each end of the tube.

The aerial coil, the *primary winding* of this radio-frequency (RF) transformer, is made from about 30 turns of 26 SWG enamelled copper wire, wound on the matchstick ribs. These ribs should be about 2 cm long, and are glued on **top** of the end of the secondary coil *after* the secondary coil has been wound. The *secondary winding* has 140 turns with *taps* every 10 turns for 70 turns. This coil covers most of the length of the cardboard tube. To make your coil-winding easier, here are some tips for completing the coil and still having some hair left at the end!

- Support the reel of wire on a dowel rod or pencil, clamped in a vice or held rigid in a vertical position by some other means. Don't leave it to trail around on the floor.
- Have a small piece of sandpaper handy to remove the enamel from the copper wire, starting with the end coming from the reel. Remove about 1 cm only, until the shiny copper is visible all round the end of the wire.

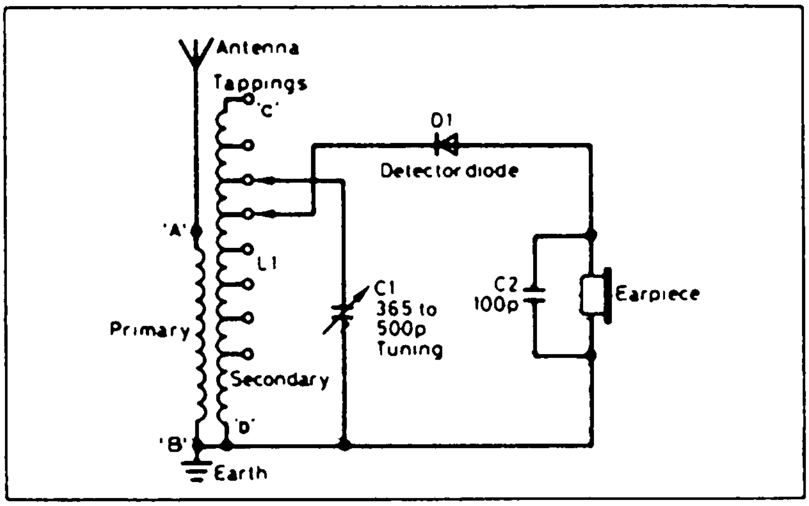

Figure 1 Circuit diagram of the crystal set

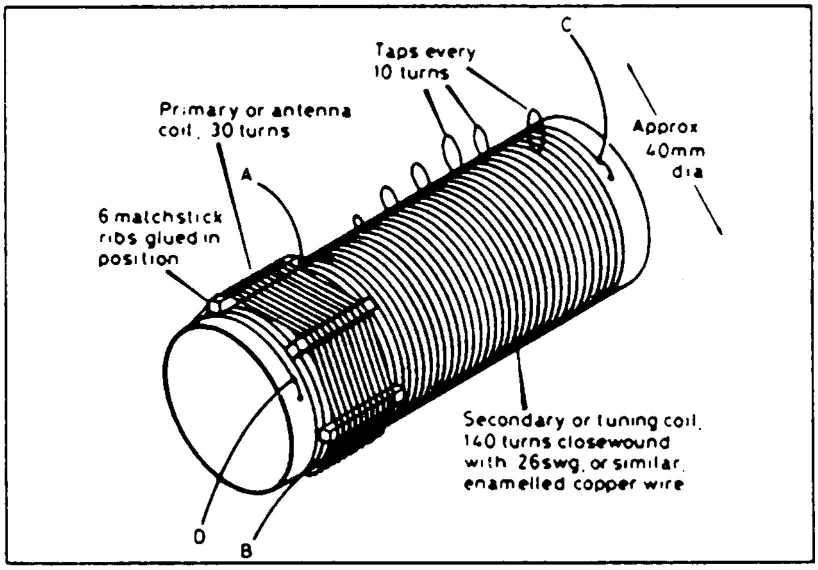
C
Taps every 10 turns
Primary or antenna coil, 30 turns
Approx 40mm dia
A
6 matchstick ribs glued in position
Secondary or tuning coil, 140 turns closewound with 26swg, or similar, enamelled copper wire
D
B

Figure 2 The coil is mounted on a cardboard tube as shown. Exact size is not too important. Try different taps for best results

- This part is the most difficult of all. Seasoned coil winders may have their own favourite ways of doing this, but for first-timers, play it safe and do it this way! Have a second bobbin (or a piece of stout card) ready, on which you will need to wind the wire loosely while you prepare the taps. Then, starting from the free end of the wire, measure 125 cm and, at this point, remove about 2 cm of the enamel with the sandpaper. Bend the wire firmly back on itself in the middle of the bared area, and *tin* (cover with solder) the exposed copper. This is your first tap. Then repeat this six more times at 125 cm intervals, thus giving you seven taps; wrap the wire as you go, on to the second bobbin or piece of card. After the seven taps, you are half-way down your coil so, to begin winding it in earnest, wind the wire carefully from the second bobbin or card **back** on to the original reel of wire.
- Looking at Figure 2, make two small holes about 3 mm apart, about 2 cm down from the top of the tube, where wire C will be entering. Poke the end of the wire into one hole, then bring it out again through the other; leave about 10 cm of wire on the free end. Loop this wire in and out once more, thus anchoring the wire firmly.
- Then, begin winding; keep the wire tight and make adjacent turns touch neatly. As you pass each tap, make sure it sticks outwards while avoiding flexing the wire too much; the wire is inherently weak at each tap.
- When you reach the last 10 cm of wire, stop. Put two more holes, like the first, beside your stopping point, and anchor the end, D, of the coil in the same way as you did with end C. The secondary winding is complete. Take a break *after* you have completed the next step!

- Prepare the six matchsticks for supporting the primary coil and glue these every 60 degrees around the end of the coil at D. Remove the enamel from the free end of the new copper wire on your reel and, using your favourite quick-drying cement, put a drop over the wire on one of the matchsticks, leaving 10 cm free as before. If you are using 26 SWG enamelled copper wire, its diameter is 0.46 mm, so 30 turns should occupy 14–15 mm of the 20 mm matchstick length. This will allow you to position the glued point such that the coil will be roughly in the centre of the matchsticks. When the glue has set, wind the 30 turns and anchor this end of the coil in the same way. Then remove the enamel at the end of the 10 cm lead.

Assembly

The coil former can be screwed to the baseboard at each end, *before* the front panel is screwed to the baseboard! As the photograph on the next page shows, the tuning capacitor is mounted in the centre of the front panel. Five solder tags (or drawing pins as a last resort) should be screwed into the baseboard. The two to the left of the photograph are the connections to the earphones, the two along the rear edge have the diode, D1, soldered to them. One end of the diode (and, for once, you can connect the diode either way round!) is connected to one earphone tag. The other end has a flying lead of about 10 cm attached to it and terminated in a crocodile clip which is used to connect to one of the taps on the coil. The fifth tag secures the end, A, of the primary coil, to which you will attach the aerial. The earth tag on the tuning capacitor, C1, serves as an anchor point for the ground connections to B and D and to the other earphone tag.

Use

The longest piece of wire you have available to use as an aerial should be connected to the aerial tag just mentioned. If you can connect the earth tag of the tuning capacitor to a good electrical earth, this will help also. You should be able to hear something as you turn the tuning knob. Try adjusting the tapping connections on the coil – change only one at a time, or you will never find the optimum positions. You now have a fully operational crystal set!

Simple it may be, but this circuit illustrates some important principles which are used even in the most expensive receivers. Firstly, the coil, L1, and the capacitor, C1, form a *tuned circuit* which *resonates* at the frequency of the station you have tuned in. This *selects* the signal you want to hear. The *modulation* on this signal is removed by the *detector*, D1, and fed to the earphones, which act as *transducers*, turning the electrical energy into sound energy which you can hear. It doesn't need a battery or other power supply either!

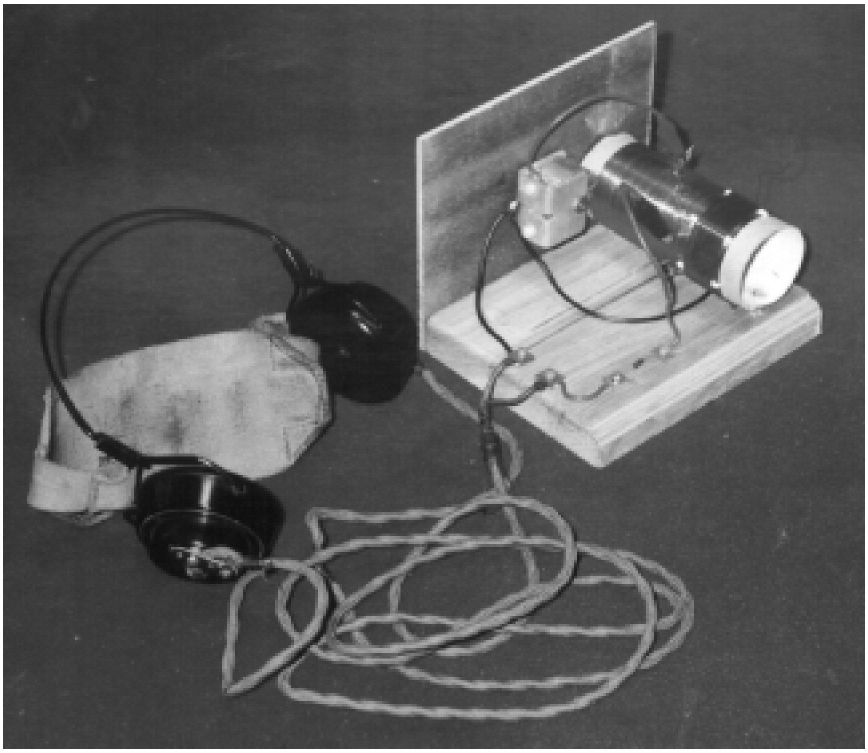

Parts list

Capacitors	
C1	Variable capacitor of between 200 and 500 picofarads (pF) maximum
C2	100 picofarads (pF) min. ceramic
Additional items	
D1	Germanium, type OA47 or OA90
	High impedance crystal earphone (not Walkman type)
	Tuning knob
	Wooden base, approx. 110 mm square
	Aluminium (or wood) front panel, approx. 80 × 100 mm
	Cardboard tube (toilet roll centre), about 110 mm long, 40 mm diameter
	Reel of 26 SWG enamelled copper wire for both windings of L1

29 A crystal calibrator

Introduction

A crystal calibrator is a device that produces oscillations of a precise frequency which are rich in *harmonics*. A harmonic is an integer multiple of the fundamental frequency; for example, if our fundamental frequency was 100 kHz (as it is in this circuit), there would be harmonics of this frequency at 200 kHz, 300 kHz, 400 kHz, and so on. These harmonics can be used, when picked up by any receiver, to calibrate that receiver, as the harmonics are quite accurate in frequency (see later for an assessment of accuracy). However, a gap of 100 kHz between harmonics is rather wide for most purposes, so we reduce this frequency to 25 kHz, so that the harmonics are then 25 kHz apart, thus producing a much more useful set of *marker points*. A calibrator such as this is often called a *crystal marker*, producing these marker points from 25 kHz to beyond 30 MHz.

The circuit

The circuit diagram is shown in **Figure 1**. The circuit around TR1 is the fundamental oscillator, and its frequency is controlled by the quartz crystal, X1. Even crystal oscillators are not 100% accurate, and the small trimmer capacitor, TC1, is able to ‘pull’ the frequency to one which is nearer the

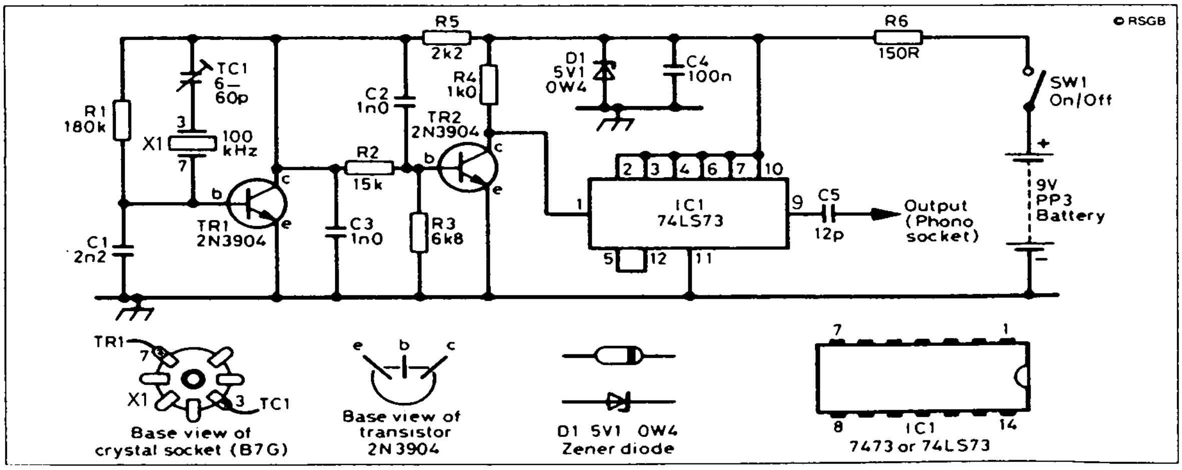
© RSGB
R5
2k2
R6
150R
TC1
6–
60p
R1
180k
X1
3
7
100
kHz
C2
1n0
R4
1k0
TR2
2N3904
D1
5V1
0W4
C4
100n
SW1
On/Off
R2
15k
b
c
e
TR1
2N3904
C1
2n2
C3
1n0
R3
6k8
2
3
4
6
7
10
1
IC1
74LS73
9
5
12
11
C5
12p
Output
(Phono
socket)
+
9V
PP3
Battery
−
TR1
7
X1
3
TC1
Base view of
crystal socket (B7G)
e
b
c
Base view of
transistor
2N3904
D1 5V1 0W4
Zener diode
7
1
8
IC1
14
7473 or 74LS73

Figure 1 Circuit diagram of the calibrator which uses easily obtainable components

correct one. TR2 is a *buffer* stage, which isolates the oscillator from the rest of the circuit. It acts as a switch, and applies a good signal (switching between 0 and 5 V) to the input of IC1.

IC1 is an integrated circuit which can be connected to do several things. It is connected here to divide the incoming frequency by a factor of 4, producing on pin 9 an output frequency of 25 kHz.

The combination of R6, C4 and D1 produces a supply of 5 V for IC1; it would be damaged if the battery voltage were applied to it. You will no doubt be ready to assemble the circuit, so here is some information for you.

Construction

If possible, always build a circuit in individual stages, which you can test as you go along. It is not always easy in small projects like this, but even the crystal calibrator can be split into two for construction and testing.

It is an ideal project for assembly on Veroboard of the copper-strip type; the prototype layout is shown in **Figure 2**, on a board measuring 11 holes by 24 strips. First of all, remove the copper strip at the locations shown, using a 3 mm drill rotated between the fingers. Hold the board up to the light to ensure that there are no pieces of copper swarf bridging adjacent strips, and that the copper is completely removed where it should be! Assemble the circuit from left to right, but do not wire up anything around IC1. Be aware that the diagram shows the board *from the component side*. Although you

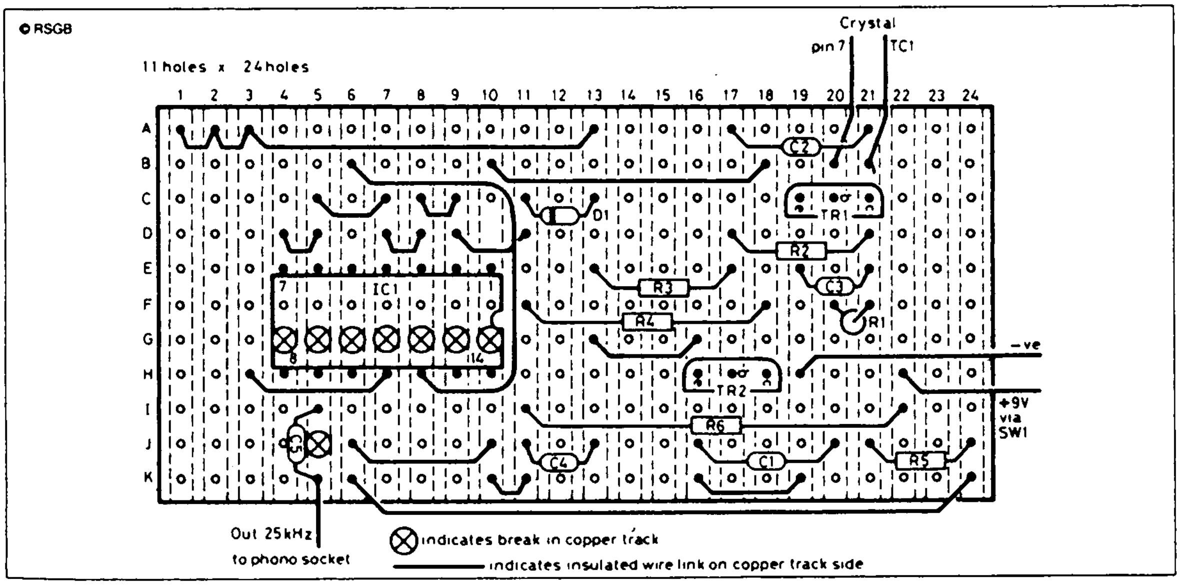
© RSGB
Crystal
pin 7
TC1
11 holes x 24 holes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
A B C D E F G H I J K
IC1
7
8
14
D1
C2
TR1
R2
C3
R1
R3
R4
TR2
R6
C1
R5
C4
−ve
+9V via SW1
Out 25kHz to phono socket
indicates break in copper track
indicates insulated wire link on copper track side

Figure 2 Stripboard, such as Veroboard, should be cut as shown to approx. 62 × 28 mm

can use virtually any kind of crystal, you will need to mount it firmly. If you use a small one, it can be soldered directly in position on the board. With one in a valve envelope, you will need a B7G valveholder. Be prepared to mount the circuit in a **metal** container; the prototype used an (empty) tin of tuna! The valveholder mounts well on any metal case.

Disconnect the aerial of your receiver, and replace it with about 30 cm of wire laid near your oscillator circuit. Connect the battery and switch on. Tune the receiver around until you can hear a whistle on SSB. Rotate the tuning knob to reduce the frequency of the whistle and, as the whistle becomes too low to hear, you have reached one of the calibration markers, and your frequency will be an integer multiple of 100 kHz. If you are already using a calibrated receiver, you will be able to verify this. Going up or down in frequency should locate another marker 100 kHz away, and so on right through the receiver's tuning range. If the signals coming from the circuit are very weak, switch off the calibrator and connect another 30 cm piece of insulated wire to the collector of TR2, and lay it close to (but *not* touching) the wire from the aerial connector. Switch on an try again; you should not now have a problem! If you have access to a multimeter, check that the voltage at the collector of TR1 is very close to 5 V. If it is close to 9 V, you have connected the diode, D1, the wrong way round!

Having verified that the oscillator is working, you can now wire up the integrated circuit socket, being careful to put the notched end of the socket pointing towards R4, the collector resistor for TR2.

Check your connections around IC1, and when you are satisfied that they are correct, line up IC1 with its socket, making sure that the notched ends are together, and press down *gently* to insert the IC into its socket. Insert the 30 cm piece of insulated wire into the output socket, connect the battery and switch on. You should still hear whistles in your receiver, but now they should be 25 kHz apart, rather than 100 kHz apart.

All that now remains to be done is to mount the circuit rigidly inside whatever casing your have chosen. Make sure that none of the connections under the board touch the metal case, and secure the valveholder, on/off switch and output connector to the case. You now have a completed crystal calibrator.

Calibration

The simplest way to calibrate your circuit is with a frequency counter. Most clubs will have one of these and, if not, will know someone who has! Connect it to the collector of TR2, where the frequency should be 100 kHz. Do *not* connect it to any part of the circuit around TR1, or you may alter the frequency you are trying to measure! If the frequency is not exactly 100.000 kHz, rotate TC1 until it is (or is as close as you can get it). Now your calibrator is as accurate as the counter with which you have calibrated it.

Accuracy

Despite your best efforts at calibration, by whatever means, your crystal will *never* have a constant frequency. Such a thing is a scientific impossibility. It is usual to express the accuracy of a crystal in parts-per-million (ppm), and it is governed by many things, principally its temperature. You will be *very* fortunate if your circuit maintains an accuracy better than about ±10 ppm. Expressed in figures, it means that the true frequency can be anywhere between 99.999 kHz and 100.001 kHz, i.e. within 1 Hz of the correct frequency.

Although this may seem more than adequate, it is as well to remember that the accuracy degrades as the frequency increases. At 1 MHz the error will be ±10 Hz and at 10 MHz it will be ±100 Hz. At 30 MHz it will be 300 Hz. Even so, this should be acceptable for most non-critical applications.

Parts list

Resistors: all 0.25 watt, 5% tolerance		**Maplin code**
R1	180 kilohms (kΩ)	M180K
R2	15 kilohms (kΩ)	M15K
R3	6.8 kilohms (kΩ)	M6K8
R4	1 kilohm (kΩ)	M1K0
R5	2 kilohms (kΩ)	M2K0
R6	150 ohms (Ω)	M150R
Capacitors		
C1, C2, C3	1 nanofarad (nF) or 1000 picofarads (pF)	WX68Y
C4	100 nanofarads or 0.1 microfarad (µF)	YR75S
C5	12 picofarads (pF)	WX45Y
TC1	60 picofarads (pF) trimmer	WL72P
Semiconductors		
TR1, TR2	2N3904 npn	QR28F
IC1	7473 or 74LS73N dual JK flipflop	YF30H
D1	5.1V 500 mW Zener	QH07H
Additional items		
100 kHz crystal		
B7G valveholder (found at most rallies)		
14-pin DIL IC socket		BL18U
Phono socket		YW06G
PP3 battery connector		HF28F
PP3 battery		
On/off switch		FH00A
Veroboard		JP47B
Metal case as required		

30 A simple short-wave receiver – Part 1

Introduction

No apologies for yet another design of short-wave radio. This is the beauty of our hobby – there is always something new to try. Not all circuits operate in the same way; not all circuits work equally well; sometimes a simple design suits the operator better than a complicated design. So here is another one for your consideration; it's a good project for a novice, *and* a good one-evening project for someone more experienced.

A basic description

Figure 1 gives the block diagram of the receiver, which employs a *regenerative detector*, one of the earliest techniques by which excellent *selectivity* (the ability to separate two stations very close in frequency) could be combined with good *sensitivity* (the ability to pick up very weak stations) using a very simple circuit. The receiver falls into the category known as *tuned radio frequency* (TRF), meaning that the whole circuit (prior to the extraction of the modulation from the carrier) operates at the incoming radio frequency. In other words, it is *not* a superhet.

As you can see, the regenerative detector has what is called a *feedback loop*, which feeds a small amount of the output signal back to the input. You have heard the effects of feedback with a public address system, when the microphone gain is too great – everything becomes very loud and then bursts into a deafening squeal! This is exactly what the feedback does here, except that it is carefully controlled, thus providing both gain and selectivity. The resulting audio is then amplified for use with headphones.

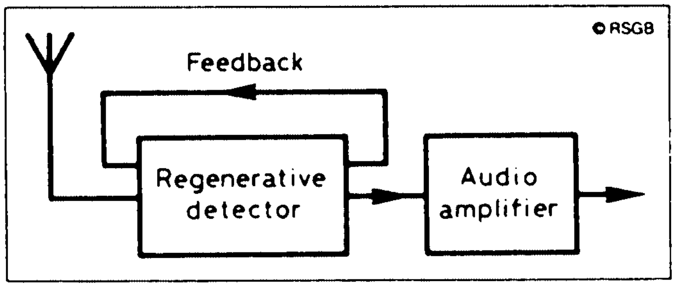

Figure 1 Just two simple stages make up the circuit

The circuit

The circuit of **Figure 2** shows the complete system. TR1 is an untuned *field-effect transistor* (FET) stage, and is used to match the aerial to the next stage. Occasionally, a regenerative detector produces unwanted signals, and TR1 also prevents them from reaching the aerial and being transmitted! The smaller of the two windings on T1 (the *primary* winding) will match a low-impedance aerial, the capacitor input matching a high-impedance aerial.

Ignore TR2 for the moment – the next stage in the signal path is TR3, the detector, using another FET. The tuned circuit is formed by T2, VC1 and VC2 (remember TR1 is untuned). The primary winding on T2 taking the output from TR1. The reason for having two variable tuning capacitors, one large, the other small, is that the large capacitor is the main tuning capacitor, while the small one is used as a *bandspread* control (for very fine tuning). The tuned signals are then detected (converted to audio frequencies) by TR3.

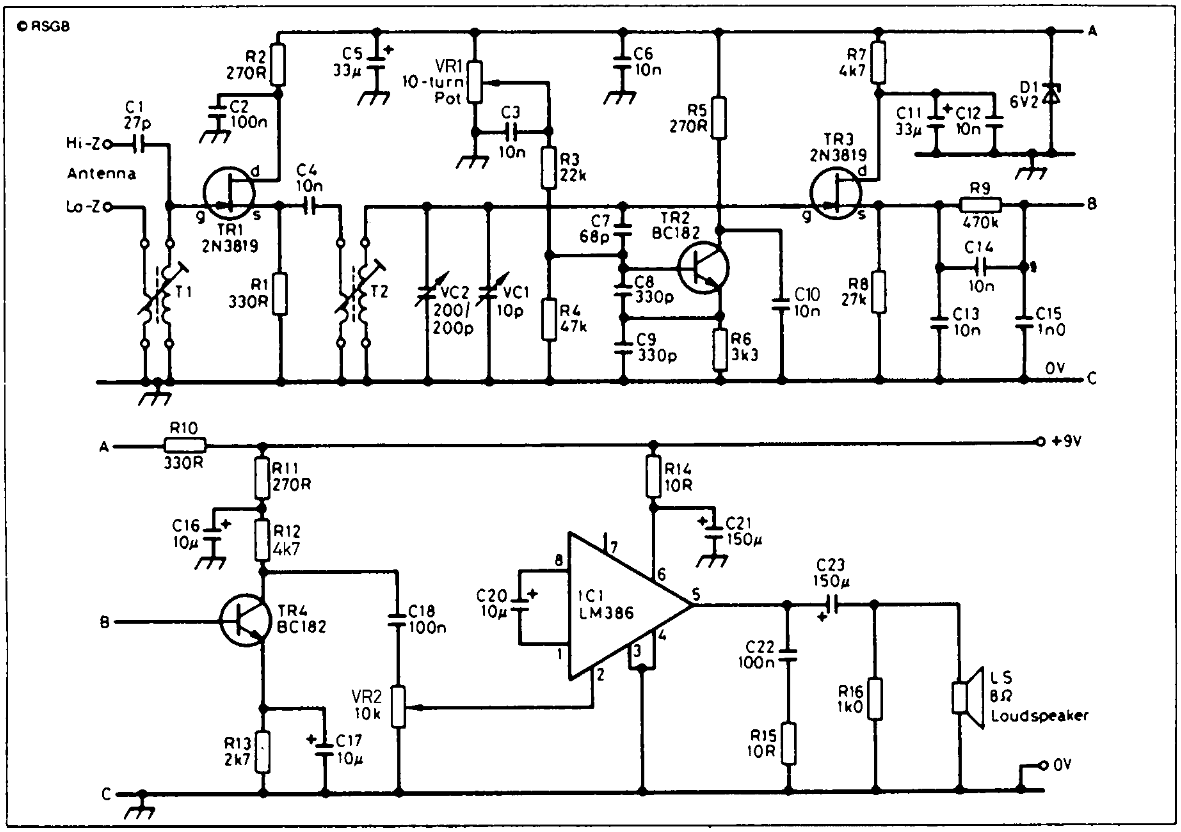
© RSGB
C1
27p
Hi-Z
Antenna
Lo-Z
R2
270R
C2
100n
TR1
2N3819
g
d
s
R1
330R
T1
C4
10n
T2
C5
33μ
VR1
10-turn
Pot
C3
10n
R3
22k
VC2
200/
200p
VC1
10p
R4
47k
C6
10n
C7
68p
TR2
BC182
C8
330p
C9
330p
R5
270R
R6
3k3
C10
10n
R7
4k7
TR3
2N3819
C11
33μ
C12
10n
D1
6V2
R8
27k
R9
470k
C14
10n
C13
10n
C15
1n0
A
B
0V
C
R10
330R
+9V
R11
270R
C16
10μ
R12
4k7
TR4
BC182
C18
100n
R13
2k7
C17
10μ
VR2
10k
C20
10μ
IC1
LM386
1
2
3
4
5
6
7
8
R14
10R
C21
150μ
C23
150μ
C22
100n
R15
10R
R16
1k0
LS
8Ω
Loudspeaker
0V

Figure 2 The circuit diagram showing the 10-turn potentiometer VR1. This is the regeneration control for the circuit

TR2 is a *Q-multiplier* stage. It is a Colpitts-type oscillator which uses C8 and C9 to give feedback and produce the oscillation. C7 couples the tuned circuit (and the input to TR3) to TR2. The tuned circuit thus controls the frequency of oscillation of TR2 and the signal passed on to TR3.

The secret of the ease of operation of this receiver is VR2, a 10-turn potentiometer (or *helipot*). The resistance wire is wound in the form of a helix, giving a much greater wire length than in a normal potentiometer, and the shaft must be turned ten times to cover the whole length. Helipots are very useful when very fine adjustments have to be made. Here, VR2 sets the regeneration (or *reaction*) level, depending on the type of signal you are receiving, as will be discussed in Part 2.

TR4 provides the first stage of audio amplification and, after the volume control, VR2, the audio amplifier integrated circuit, IC1, will drive a small loudspeaker or headphones.

In Part 2, the construction will be discussed, together with the choice of aerial, the parts list, and advice on using the receiver.

31 A fruit-powered medium-wave radio

Introduction

This is a one-evening project that will result in a working medium-wave (MW) radio, and will also teach you a little about the way electricity can be generated from the right metals and a little (safe) acid. All you need are three lemons or other citrus fruit, three pieces of copper and three pieces of zinc (or galvanised metal) for your power supply.

Construction

Figure 1 shows the circuit and **Figure 2** its layout on a simple ‘plug-in’ prototype board. The six pieces of metal are connected as shown, to wire the three lemons in series; use ordinary wire between each lemon and the next. If you have a meter to measure the total voltage, it should be about 1.8 V. Use a standard ferrite rod, and wind on it about 40 turns of single-conductor PVC-insulated wire.

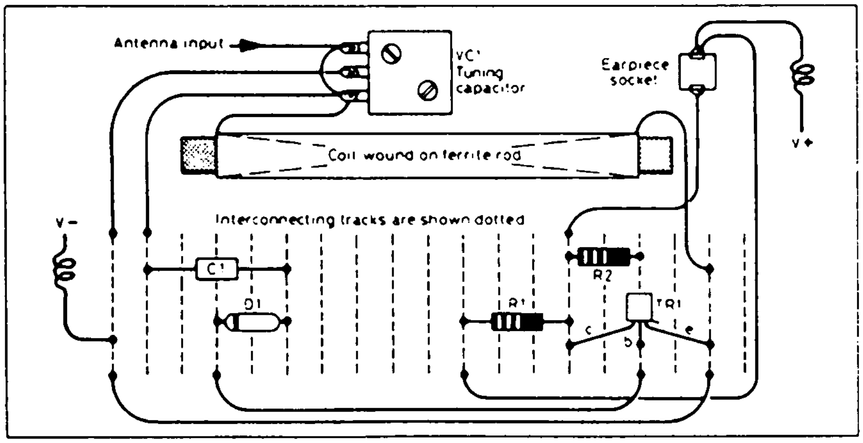
Antenna input
VC1
Tuning
capacitor
Earpiece
socket
Coil wound on ferrite rod
Interconnecting tracks are shown dotted
V−
V+
C1
D1
R1
R2
TR1
c
b
e

Figure 1 Most parts are plugged into the board as shown – soldering is not required

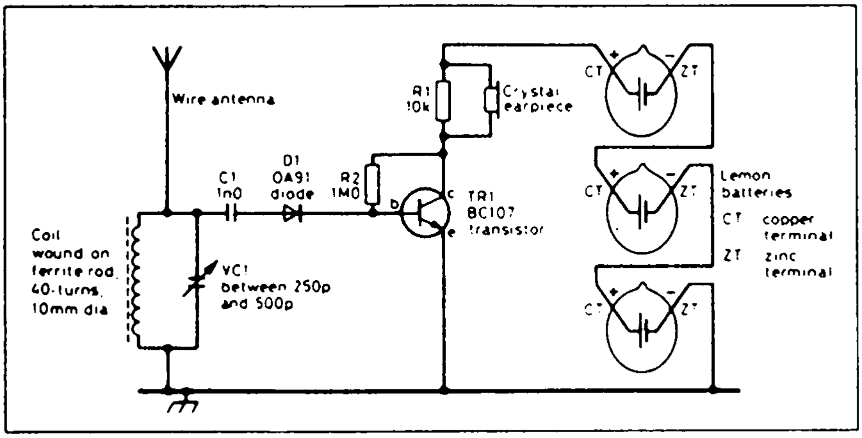
Wire antenna
R1
10k
Crystal
earpiece
CT
ZT
+
−
D1
OA91
diode
C1
1n0
R2
1M0
b
c
e
TR1
BC107
transistor
Lemon
batteries
CT copper terminal
ZT zinc terminal
Coil
wound on
ferrite rod
40-turns,
10mm dia
VC1
between 250p
and 500p

Figure 2 Three lemons power the radio, which gives good results for such a simple circuit

Wire up the circuit on the board as illustrated in Figure 2. Soldering is not required with this type of board – just plug in the components and the wires. Only one transistor is needed. The tuning capacitor, VC1, selects the station you want to hear, and D1 helps to remove the carrier from the RF signal. The resulting audio signal is fed to TR1, a small transistor audio amplifier, which makes the signal big enough to drive a crystal earpiece comfortably. Walkman-type earphones will not work, so invest in a crystal earpiece which you can use in several other projects, too! If you use a smaller capacitor than that specified for VC1, you will need more turns on the aerial coil.

Tests on the prototype indicated that the radio will run for about a week on three lemons!

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	1 megohm (MΩ)
R2	10 kilohm (kΩ)
Capacitors	
C1	1 nanofarad (nF) min. ceramic
VC1	250–500 picofarad (pF) variable
Semiconductors	
TR1	BC107 npn (or BC108, BC109C)
D1	OA90, OA91 germanium (not silicon)
Coil	
L1	2 metres of single-conductor insulated wire on a standard ferrite rod

Additional items
- Plug-in prototype board, e.g. Maplin YR84F
- Wire aerial at least 3 m long
- Crystal earpiece
- Three juicy fruits

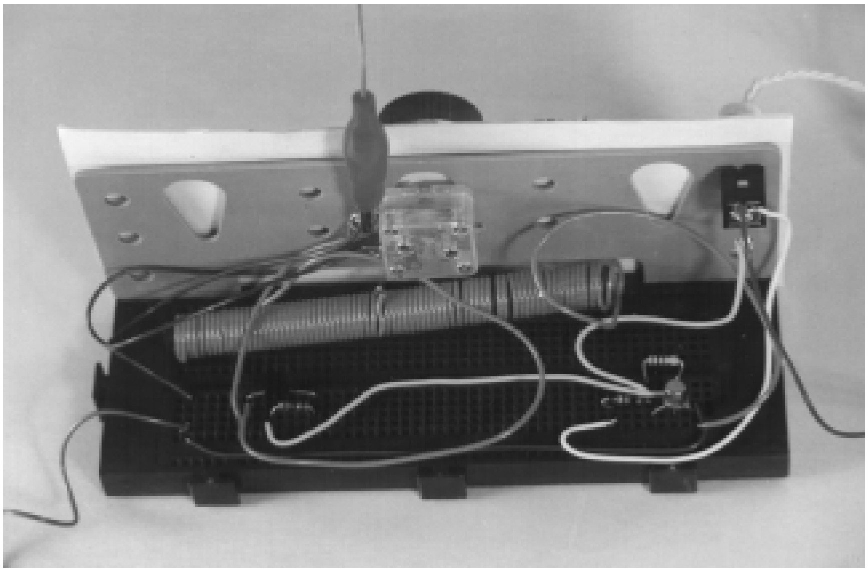

32 A capacitance bridge

Introduction

It is always valuable to have a means of measuring capacitors amongst your test gear. Universal LCR bridges (i.e. systems that can measure inductance, capacitance and resistance) can be found at most rallies, and you can work your way up to owning one of these. In the meantime, a piece of home-made equipment gives you experience as well as resulting in a useful measuring instrument.

The circuit described here is called a *capacitance bridge*, because it balances the effects of one resistor/capacitor pair against another; if one capacitor has an unknown value, then the other can be calculated. The basic bridge circuit is shown in **Figure 1**. To avoid having calculations to perform, this instrument will be calibrated by using capacitors of known values. The bridge is a useful way of performing measurements, because a knob is turned until there is a null in the signal from an earpiece or loudspeaker. The ear is very precise in being able to perceive nulls, which makes the bridge easy to use and reasonably accurate. At the null, the bridge is said to be *balanced*.

How does it work?

Figure 2 is the circuit diagram for this capacitance bridge. Transistors TR1 and TR2 form an oscillator. This is the audio oscillator shown in Figure 1, and produces an alternating voltage which is fed to the bridge. RV1 (in the collector lead of TR2) replaces both R1 and R2 in Figure 1 – that part of

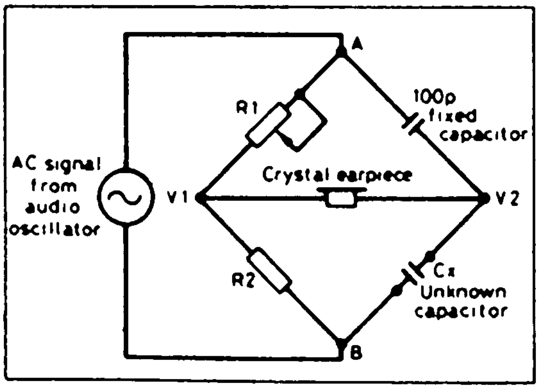

Figure 1 Simplified circuit of a capacitance bridge. R1 is adjusted for minimum sound

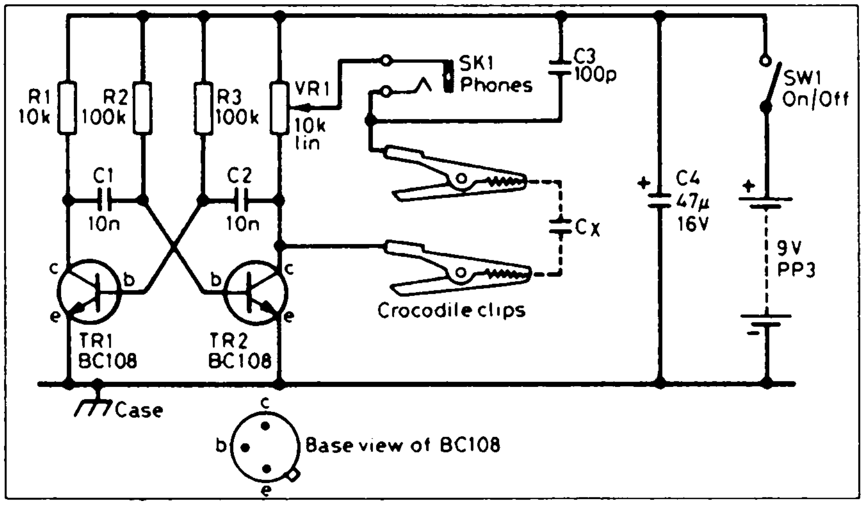
R1
10k
R2
100k
R3
100k
VR1
10k
lin
SK1
Phones
C3
100p
SW1
On/Off
C1
10n
C2
10n
C4
47μ
16V
+
Cx
9V
PP3
+
-
c
b
e
b
c
e
TR1
BC108
TR2
BC108
Crocodile clips
Case
c
b
e
Base view of BC108

Figure 2 Transistors TR1 and TR2 give an audio signal which is adjusted by variable resistor RV1

RV1 *above* the wiper represents R1, while that part *below* the wiper represents R2. The voltage on one side of the earpiece is determined by the ratio of these values, and is adjusted by rotating RV1. The voltage on the other side of the earpiece is determined by the ratio of C3 to Cx, where Cx is the unknown-valued capacitor. When these two voltages are the same, the bridge is balanced and there is no current through the earpiece.

Figure 3 shows the layout of the components on the matrix board of the prototype. It measures 10 holes by 14 holes, and is of the plain type, i.e. no copper strips. All earth connections are taken to a single solder tag. When

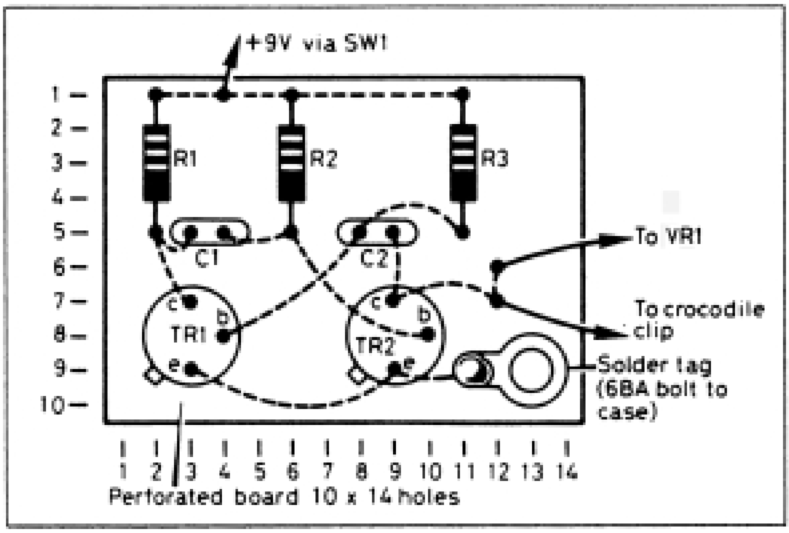
+9V via SW1
1
2
3
4
5
6
7
8
9
10
R1
R2
R3
C1
C2
To VR1
c
b
TR1
e
c
b
TR2
e
To crocodile clip
Solder tag
(6BA bolt to case)
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Perforated board 10 x 14 holes

Figure 3 Components are soldered together on a small prototype board

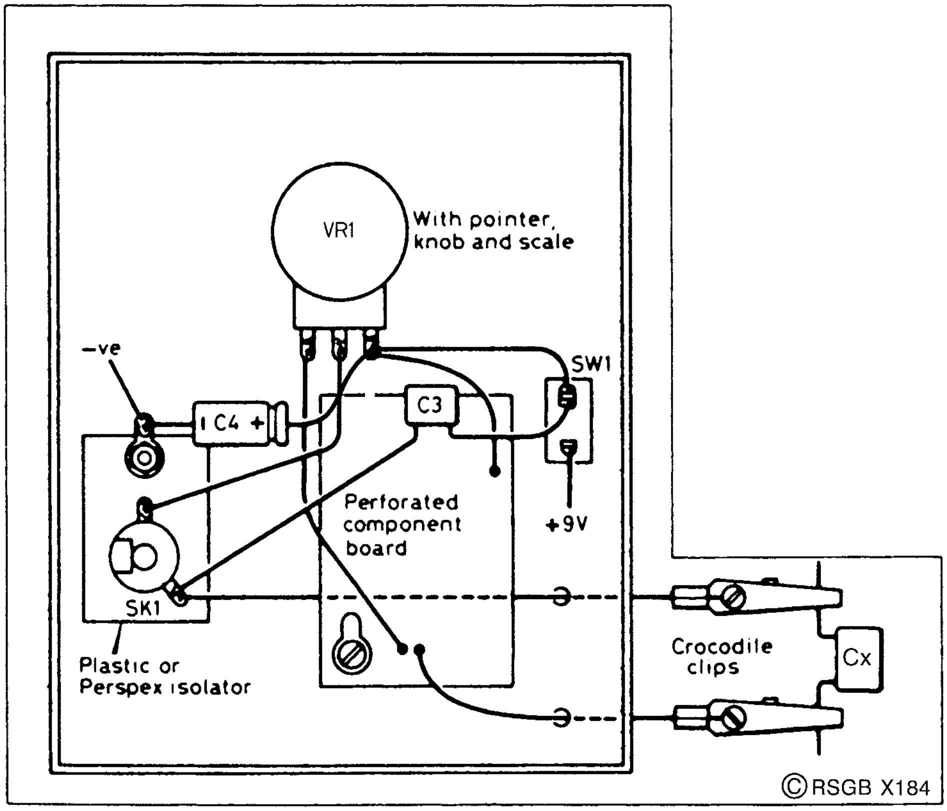
VR1
With pointer, knob and scale
–ve
C4
C3
SW1
Perforated component board
+9V
SK1
Plastic or Perspex isolator
Crocodile clips
Cx
©RSGB X184

Figure 4 Layout of the parts inside an aluminium box. The earpiece socket is insulated from the case

mounting the board inside a metal box, a long screw is used with extra nuts, to earth the board to the case while providing a stand-off, thus preventing any unwanted short-circuits between the board connections and the case. Be aware that a *crystal* earpiece must be used; low-impedance headphones of the Walkman type are not satisfactory. Solder two flying leads as shown in **Figure 4**, about 15 cm long, terminated in small crocodile clips for attaching to the unknown capacitor.

Calibration

After checking the circuit carefully, attach the battery and switch on; a buzzing noise should be heard in the earpiece. This is the first sign that everything should be OK. If there is no buzz, switch off and recheck the connections.

You will now need a range of close-tolerance (1%) silver-mica capacitors covering the range 10 to 1000 picofarads (pF). Arrange the capacitors in ascending order and connect them to your bridge in sequence. After having prepared a neat piece of card or paper mounted behind the knob on the front panel, mark the dial at the positions of the nulls for all the capacitors. You have now calibrated your capacitance bridge. If you are more likely to want to measure larger capacitors, replace C3 by a 1 nanofarad (nF) capacitor, and the bridge will measure up to 10 nF approximately.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	10 kilohms (kΩ)
R2, R3	100 kilohms (kΩ)
RV1	10 kilohms (kΩ) linear
Capacitors	
C1, C2	10 nanofarads (nF) ceramic
C3	100 picofarads (pF) silver mica or polystyrene
C4	47 microfarads (μF) 16 V electrolytic
Semiconductors	
TR1, TR2	BC108 npn
Additional items	
SW1	SPST on/off switch
	Battery connector, PP3 type
	Earpiece, crystal type
SK1	3.5 mm jack socket for earpiece
	Battery, 9 V PP3
	Matrix board, approx. 10 holes by 14 holes
	Aluminium case, approx. 10 × 8 × 5 cm

33 A simple short-wave receiver – Part 2

Introduction

In Part 1, the design of this receiver was discussed in some detail. Now we are going to put it all together and see how it works. The receiver is laid out on a printed-circuit board (PCB) or on a matrix board.

Construction

The layout of the components is shown in **Figure 1**. Identify each part separately, insert it into the holes in the board and solder carefully. Long leads may be cropped before or after soldering, depending on your skill and preference. All electrolytic capacitors, T1, T2 and IC1, must be connected correctly. The front-panel controls are connected to the PCB terminals shown in Figure 1.

The layout of the controls and the placing of the board inside the case are matters of personal preference. The size of the prototype front panel is shown in **Figure 2**. The prototype had a slow-motion drive fitted to VC2, the main tuning capacitor. This required the capacitor to be fitted on its own small panel. The bandspread control does not need any form of

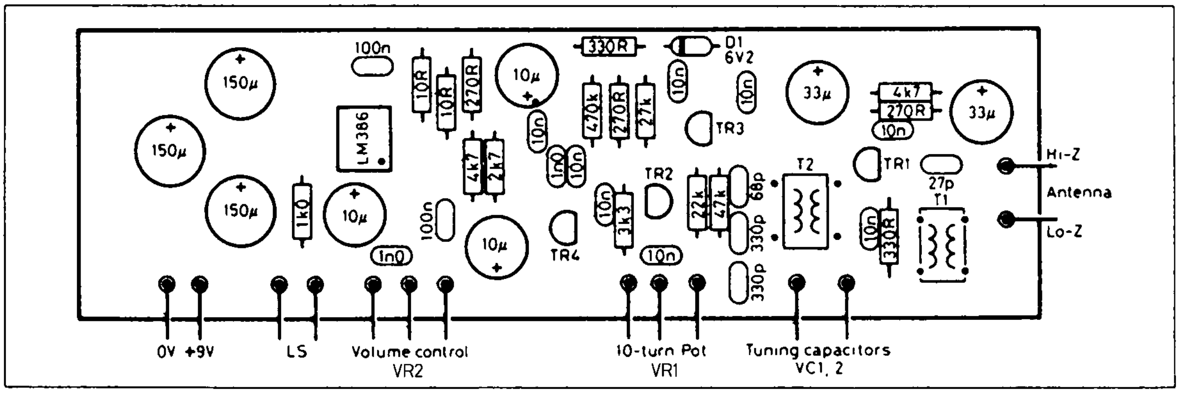
150µ
150µ
150µ
100n
10R
10R
270R
10µ
330R
D1
6V2
10n
10n
33µ
4k7
270R
10n
33µ
LM386
470k
270R
27k
TR3
10n
1n0
10n
4k7
2k7
TR2
T2
TR1
27p
T1
Hi-Z
Antenna
Lo-Z
1k0
10µ
100n
10µ
1n0
TR4
10n
3k3
10n
22k
47k
68p
330p
330p
10n
330R
0V +9V
LS
Volume control
VR2
10-turn Pot
VR1
Tuning capacitors
VC1, 2

Figure 1 Layout of the receiver circuit board

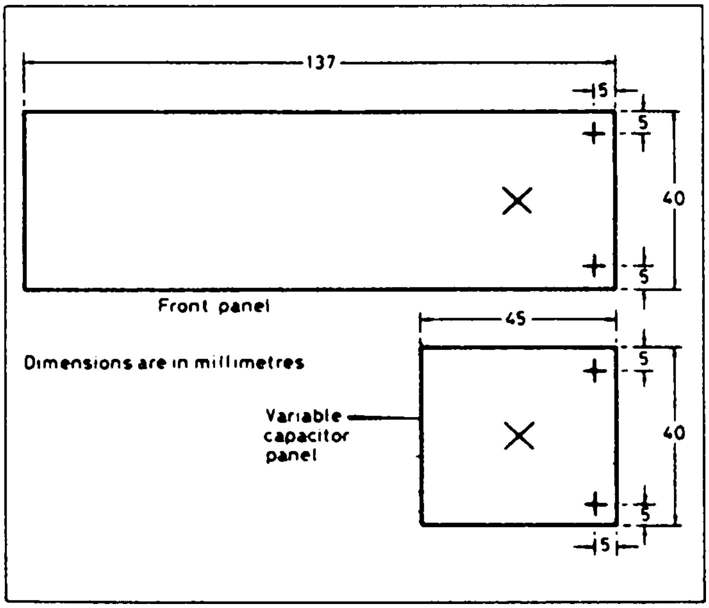

Figure 2 The case can be metal or made from PCB material

slow-motion drive. A tuning dial may be fitted behind the tuning knob on VC2 if required. The volume control, VR2, has its own on/off switch for the battery, which can be mounted behind the back panel, on which are mounted the aerial input sockets and the headphone/speaker socket. Check your soldering carefully, together with all the wiring between the board, the potentiometers and variable capacitors. When you are confident that everything is perfect, connect the battery and switch on.

The first tests

If you have a good aerial and are using an aerial tuning unit (ATU), use the low-impedance input. Excellent results are possible, though, with about 3 metres of wire connected to the high-impedance input. Set the volume control to give a gentle hiss in the headphones or speaker. Advance the reaction control to give a definite hiss. As you tune in an AM station, the hiss will change to a whistle; back off the reaction until the oscillation *just* stops. The receiver is now correctly set for AM reception.

On the amateur bands, the stations will be SSB or CW, and the reaction needs to be set *just above* the oscillation point, and will need slight adjustment from time to time for different qualities of signal. Juggling with the volume, reaction and tuning is part of the pleasure of using regenerative receivers!

In action

Practice is needed for best results. The regenerative receiver is renowned for its versatility in being able to be set up *exactly* for all types and strengths of signal. The basic receiver tunes from 6.5 MHz to 11 MHz approximately; this includes two amateur bands at 7.0 MHz and 10.1 MHz and two broadcast bands.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R10	330 ohms (Ω)
R2, R5, R11	270 ohms (Ω)
R3	22 kilohms (kΩ)
R4	47 kilohms (kΩ)
R6	3.3 kilohms (kΩ)
R7, R12	4.7 kilohms (kΩ)
R8	27 kilohms (kΩ)
R9	470 kilohms (kΩ)
R13	2.7 kilohms (kΩ)
R14, R15	10 ohms (Ω)
R16	1 kilohm (kΩ)
VR1	10 kilohm (kΩ) linear 10-turn potentiometer
VR2	10 kilohm (kΩ) log potentiometer

Capacitors

C1	27 picofarads (pF)
C2, C18, C22	100 nanofarads (nF)
C3, C4, C6, C10, C13, C14	10 nanofarads (nF)
C5, C11	33 microfarads (μF) 16 V electrolytic
C7	68 picofarads (pF)
C8, C9	330 picofarads (pF)
C15	1 nanofarad (nF)
C16, C17, C20	10 microfarads (μF) 16 V electrolytic
C21, C23	150 microfarads (μF) 16 V electrolytic
VC1	10 picofarads (pF) variable
VC2	200 picofarads (pF) variable

Semiconductors

TR1, TR3	2N3819
TR2, TR4	BC182
IC1	LM386
D1	6.2 V 0.5 W Zener

Inductors

T1	3 turns primary, 15 turns secondary, wound on 2-hole ferrite bead, with 28 SWG wire
T2	2 turns primary, 17 turns secondary, wound on a T68-2 toroidal former, with 28 SWG wire

Additional items
- Printed-circuit board
- Battery connector
- PP3 battery
- 8-pin DIL socket
- 8 ohm speaker or headphones (Walkman type)
- Case to suit

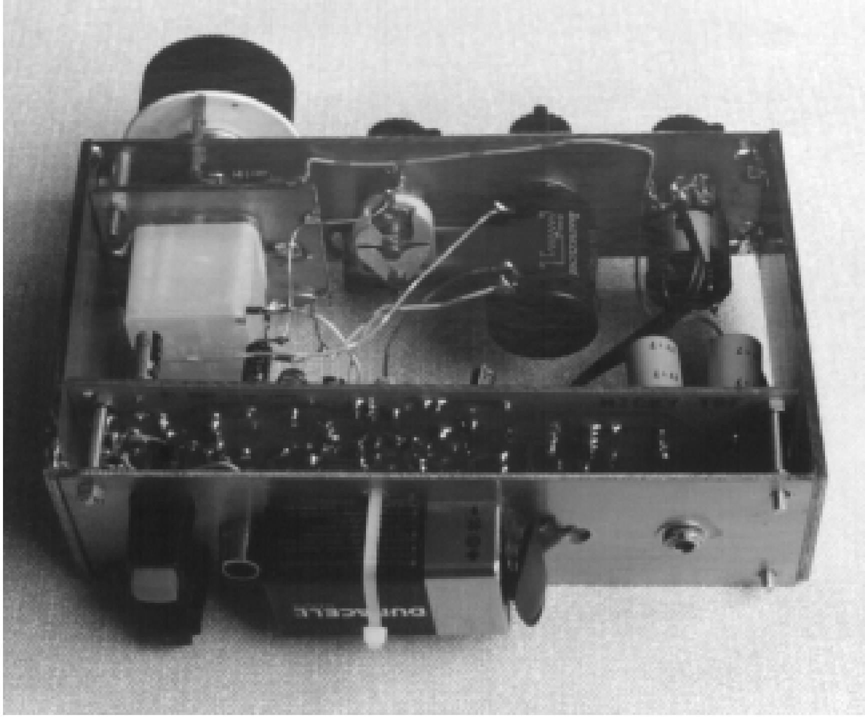

34 A basic continuity tester

Introduction

This little device can be built on a plug-in breadboard, and is ideal for testing fuses and cables, as well as doubling as a signal source for testing amplifiers, etc. It can even be used for testing npn transistors by replacing either of the transistors in the circuit, and seeing if the circuit still works!

Simple and quick to build

Using a plug-in breadboard, this circuit is so simple it could almost be built when you need it, and then dismantled again! If you want the circuit to make a different sound, the components to change are R2, R3, C1 and C2. Always make sure that R2 = R3 and C1 = C2, or the sound may be excessively ‘edgy’ and lacking in volume.

Basically, the circuit is an oscillator which drives a little loudspeaker directly. The fuse or cable you are testing is connected between the crocodile clips. If there is a current path between the two clips, the current also flows through the circuit, thus operating the oscillator and producing a sound from the loudspeaker.

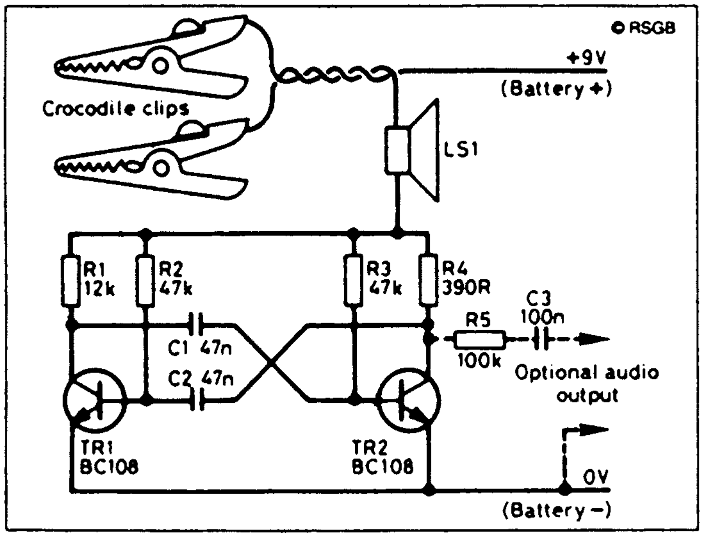
© RSGB
+9V
(Battery +)
Crocodile clips
LS1
R1 12k
R2 47k
R3 47k
R4 390R
C3 100n
R5 100k
C1 47n
C2 47n
Optional audio output
TR1 BC108
TR2 BC108
0V
(Battery –)

Figure 1 The circuit diagram of the continuity tester

If you want to use the circuit to produce a signal to test an amplifier, for instance, connect the croc clip which comes directly from the battery to the junction of R1, R2, R3 and R4. This supplies current to the circuit while, at the same time, bypassing the loudspeaker. Use the two components R5 and C3 between the oscillator and your amplifier. The loudspeaker is bypassed because you will want to listen to the output from your amplifier, *not* from the oscillator!

Remember, NEVER test any equipment which is still connected to the mains electricity supply. Avoid testing anything which is switched on and has its own power supply. You may damage both your tester ***and*** the circuit you are 'testing'!

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	12 kilohms (kΩ)
R2, R3	47 kilohms (kΩ)
R4	390 ohms (Ω)
R5	100 kilohms (kΩ) optional

Capacitors

C1, C2	47 nanofarads (nF) ceramic
C3	100 nanofarads (nF), 0.1 microfarads (μF) ceramic optional

Semiconductors

TR1, TR2	BC108 npn

Additional items

LS1	Loudspeaker (8 Ω)
	2 crocodile clips
	Prototype circuit board
	PP3 battery connector
	PP3 battery

35 A charger for NiCad batteries

Introduction

The cost of replacing dry batteries can be alleviated to a great extent by the use of rechargeable batteries such as those containing nickel and cadmium (called *NiCads*). They are more expensive than the batteries they replace, but they can be charged hundreds of times and thus prove cheaper in the long run.

NiCads produce only 1.2 V per cell, compared with the 1.5 V of ordinary cells, so before you rush out and buy lots of these, please make sure that the equipment on which you plan to use them will work on the reduced voltage! For example, if you are using four cells to give you a 6 V supply, NiCads will give you only 4.8 V, which is quite a reduction. Using six cells to replace a 9 V battery will give you only 7.2 V. Not all equipment is happy with these reductions!

However, we all use them when we can, and they save substantial amounts of money. Here is the circuit of a charger to keep them in prime condition.

Charging NiCads – the ampère-hour

NiCads require charging at constant current, which means that connecting one across a normal power supply (constant *voltage*) is useless and can destroy it. They need pampering to the extent of needing a long charge (around 16 hours) at a rate dependent upon the *capacity* of the battery. By the capacity of a battery, we mean how much energy it can store. You will probably know that energy is measured in joules. For the purposes of storing energy in batteries, the joule is *not* the ideal unit, so we use one that is! This unit is the ampère-hour (Ah), and must be interpreted with some realism. For example, if the battery is rated at 2 Ah, it will deliver a current of 0.5 amp for 4 hours, or 0.25 amp for 8 hours. Provided the current is not too high, the product of the current (in amps) and the time for which it will flow (in hours) before the battery is flat will always be around 2 Ah. A workable ‘rule of thumb’ for calculating the charging current is that its value should be around one-tenth of the numerical value of the capacity; so, for our 2 Ah battery, a charging current of around 200 mA (2 ÷ 10 = 0.2 A or 200 mA) would be used.

Constant voltage to constant current

Many integrated circuit (IC) chips are available for use as voltage regulators, i.e. they supply a constant voltage. Most of these can be persuaded to become constant current supplies with one external resistor!

The voltage regulator IC usually has only three connections – ‘input’, ‘output’ and ‘common’. It is designed (in the case of the LM7805) to produce a constant 5 V output between the ‘output’ and the ‘common’ connections, at currents of up to 1 A. If a resistor is connected between these, the IC will maintain 5 V across it. If you look at **Figure 1**, you will see the circuit performing this conversion.

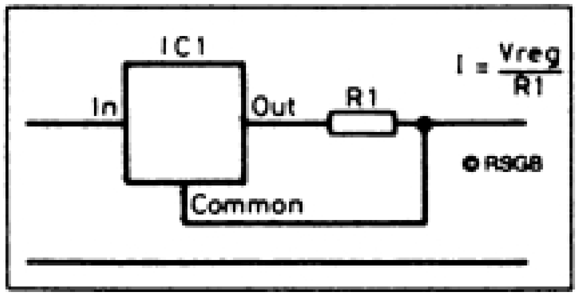

Figure 1 Voltage regulator arranged to produce a constant current

For the previous example, we derived a charging current of 200 mA, so we now need to calculate the value of resistor that will produce this current. Using the equation which is derived from Ohm’s law:

$$I = \frac{V}{R}, \text{ from which } R = \frac{V}{I},$$

where R = R1, the resistance in ohms that we are calculating,
V is the voltage across R1 (5 V), and
I is the current flowing (200 mA).

So,

$$\text{R1} = \frac{5}{0.2} = 25\,\Omega.$$

25 ohms is not a ‘common’ or ‘preferred’ resistor value, so we must choose the next *largest* value, which is 27 Ω. This reduces the current, but only slightly – it is now 185 mA. When calculating resistor values in power supply circuits, we must *always* check the power that they dissipate and make sure we specify and fit a suitable resistor.

Power (in watts) is the product of the voltage across and the current through a device, so in this case it is given by:

$$\text{Power} = V \times I = 5 \times 0.185 = 0.925\,\text{W}.$$

Rather than use a 1 watt resistor operating very near its limit, it is safer to use a 2 watt resistor operating well within its limits.

Looking again at Figure 1, we now have a constant-current source producing 185 mA, when R1 is a 27 Ω, 2 W resistor. For use with NiCads requiring charging currents other than 200 mA, you will need to repeat the two equations above, using a new value for *I*.

The full circuit and its assembly

This is shown in **Figure 2**, and can be broken down into two parts. The first is the AC to DC conversion produced by the mains transformer, T1, the bridge rectifier, BR1, and the smoothing capacitor, C1. The second is the constant-current section already discussed.

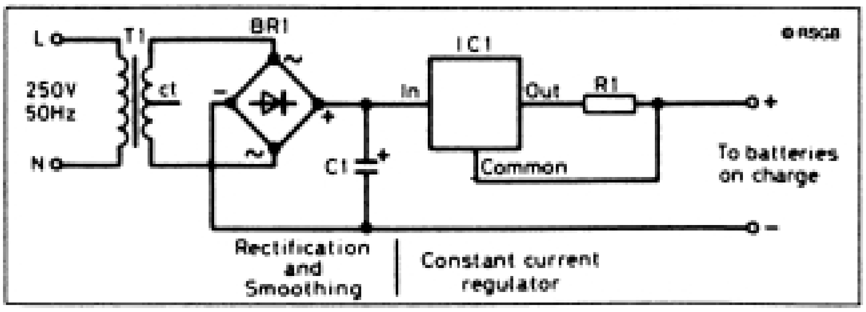

Figure 2 The complete circuit including mains components

The prototype was assembled on matrix board measuring 18 holes by 12 strips, although, as **Figure 3** shows, this is much larger than is strictly necessary. No strip cutting is needed, but make sure that IC1 and C1 are inserted correctly.

Warning! Before you attempt to wire up the transformer and the bridge rectifier, be aware that you will eventually be connecting the circuit to the mains supply, so there are three possibilities for you: (1) get a qualified friend to supervise your completion and testing of the circuit; (2) get your qualified friend to complete and test the circuit for you; (3) replace the transformer and bridge rectifier with a mains adapter.

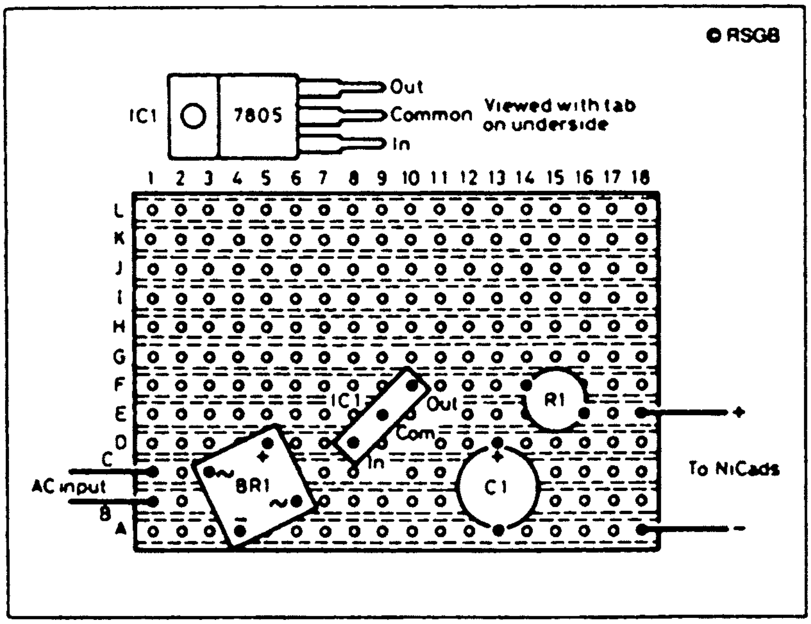

Figure 3 The charger can be built on Veroboard

If you decide to use the mains adapter, its output is connected directly across C1, because T1 and BR1 are now no longer needed. Make sure the polarity (positive and negative) is correct and that the adapter output is set to 12 V.

A quick test

If you have built the mains version, make sure all connections are correct, and that there are no soldered joints which will touch other parts of the circuit. The box must be securely closed before tests begin. The RSGB cannot be held responsible for damage to equipment or batteries! The version using the mains adapter need not be closed during tests.

Switch on. With nothing connected to the output, the unit should run cold. If this is not the case, switch off and recheck your circuit. If all is well, switch on again and connect a DC multimeter (on the current range) across the output. It should indicate only a slight difference from the calculated value of 185 mA. You can now charge your NiCads!

Other charging currents can be set by having different values for R1, perhaps selectable by a rotary switch. Remember to make sure that the values of both resistance *and* power dissipation are correct, and don't exceed the 250 mA rating of the transformer (or the 1 amp rating of the IC if you are using a bigger transformer).

Parts list

		Maplin code
Resistor		
R1	As required – see text	
Capacitor		
C1	1000 microfarads (μF) 35 V electrolytic	FF18U
Semiconductors		
IC1	LM7805 5 V, 1 A regulator	CR14Q
BR1	W005 50 V, 1 A full-wave rectifier	QL37S
Transformer		
T1	9–0–9 V, 250 mA sub-min. transformer	YN15R
Additional items		
Case		FG41U
Veroboard, 18 holes by 12 strips		
Plug to suit NiCads		
Double-sided sticky tape as required		
Insulated wire for battery connections		

36 An 80 metre crystal-controlled CW transmitter

Introduction

A simple transmitter is ideally suited to anyone venturing into our marvellous hobby for the first time. If you are put off by the complexities and prices of the 'black boxes' then this is the transmitter you've been looking for!

The circuit

The complete circuit is shown in **Figure 1**. It is a single-transistor crystal oscillator which is *keyed* (switched on and off) by the Morse key in the +12 V supply rail. The circuit first appeared in *QST* (the American equivalent of *RadCom*) in 1982, and has since appeared in a modified form in the *ARRL Handbook*. The circuit can produce about 2 watts on 80 m, and can be constructed on a piece of plain matrix board; the prototype board measured 7 cm by 3 cm, and its layout is shown in **Figure 2**.

Note that a bare copper wire runs along the bottom edge of the board to act as an earth wire for the relevant components. The only 'difficult' part of the construction is the winding and wiring of T1. The main winding is 38 turns of 26 SWG enamelled copper wire, and there are two link windings of four turns each. Make sure that all windings are wound the correct way round the toroid – Figure 1 shows this and should be studied carefully. If the windings do not have the correct *sense* (i.e. a clockwise coil has been wound anticlockwise, or vice versa), or have been connected incorrectly to the rest of the circuit, the oscillator will not work!

When putting components on the board, wire in the crystal socket *without the crystal in it*. Crystals do not like to be subjected to the horrors of a soldering iron, so keep your crystal to one side during the construction process!

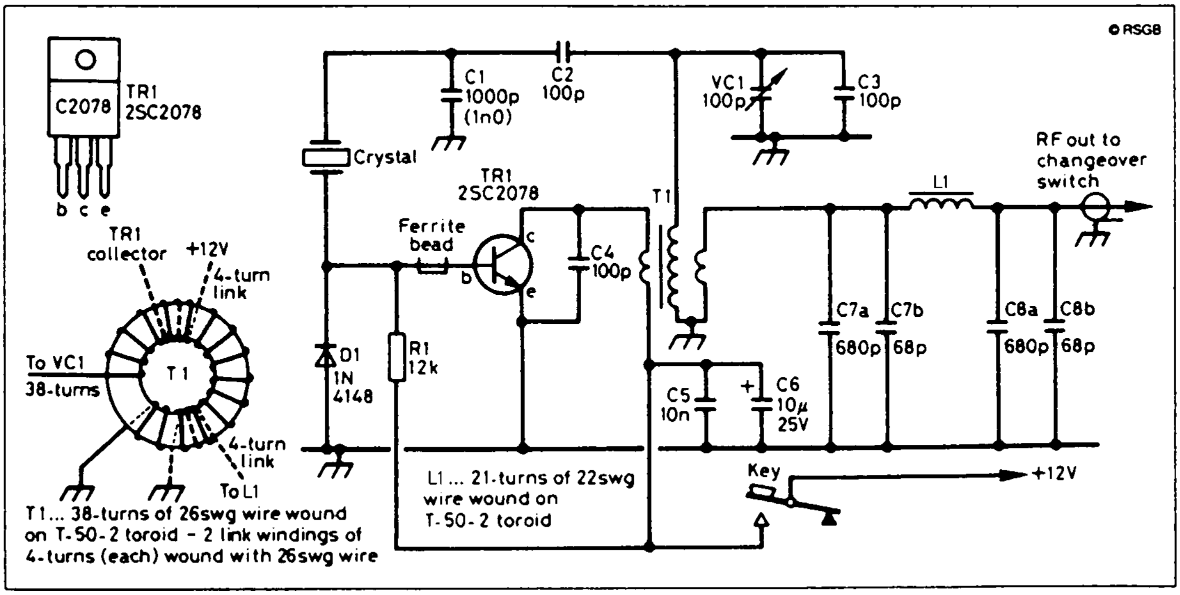
©RSGB
C2078
TR1
2SC2078
b c e
TR1
collector
+12V
4-turn
link
To VC1
38-turns
T1
4-turn
link
To L1
T1... 38-turns of 26swg wire wound on T-50-2 toroid – 2 link windings of 4-turns (each) wound with 26swg wire
C1
1000p
(1n0)
C2
100p
Crystal
TR1
2SC2078
Ferrite
bead
b
c
e
C4
100p
D1
1N
4148
R1
12k
VC1
100p
C3
100p
T1
L1
RF out to
changeover
switch
C7a
680p
C7b
68p
C8a
680p
C8b
68p
C5
10n
+
C6
10µ
25V
Key
+12V
L1 ... 21-turns of 22swg wire wound on T-50-2 toroid

Figure 1 The simplest form of transmitter is a keyed crystal oscillator. Note that L1, C7 and C8 make up a *low-pass filter* which reduces unwanted *harmonics* (outputs at the transmitter frequency multiplied by 2, 3, 4, etc.)

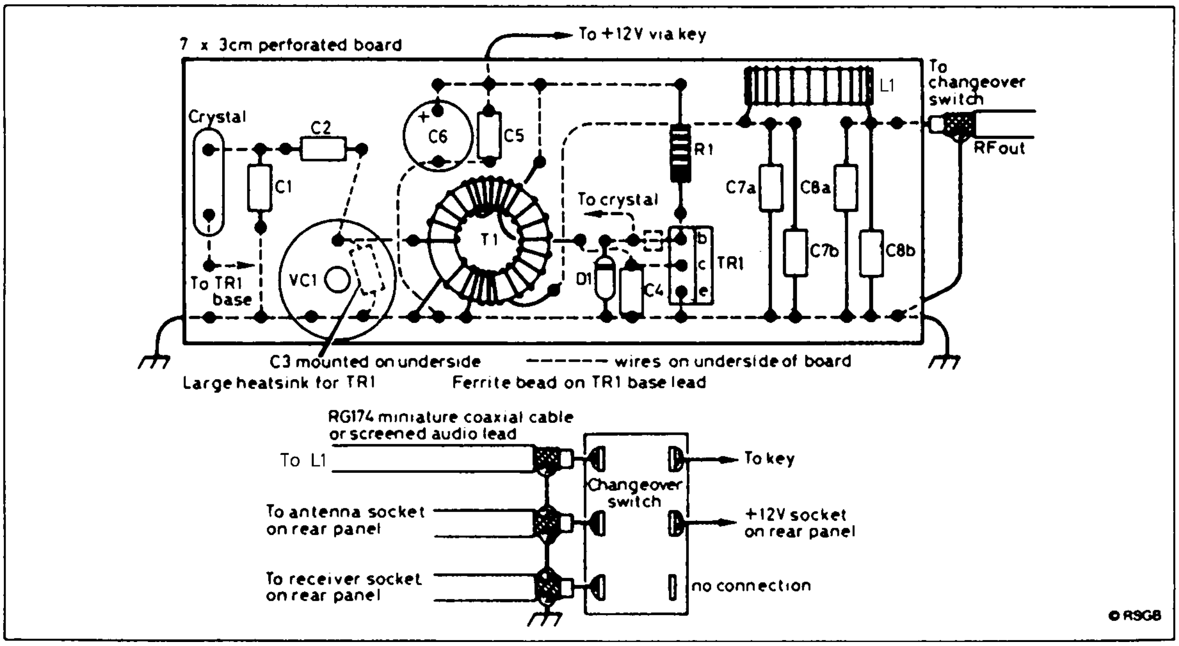
7 x 3cm perforated board
To +12V via key
Crystal
C2
C1
C6
C5
R1
L1
To changeover switch
RFout
C7a
C8a
C7b
C8b
To crystal
T1
TR1
b
c
e
D1
C4
VC1
To TR1 base
C3 mounted on underside
------ wires on underside of board
Large heatsink for TR1
Ferrite bead on TR1 base lead
RG174 miniature coaxial cable or screened audio lead
To L1
To antenna socket on rear panel
To receiver socket on rear panel
Changeover switch
To key
+12V socket on rear panel
no connection
© RSGB

Figure 2 Component layout is straightforward on a 'matrix board' (it has holes but no copper strips). The dotted lines are the connecting wires. The lower part shows the connections for the transmit/receive switch

The output filter, which comprises C7, C8 and L1, is a *low-pass* filter, which helps reduce any *harmonics* present in your signal. Harmonics are integral multiples of your transmitter frequency, so if you are transmitting on a frequency, *f*, harmonics will be present at frequencies 2*f*, 3*f*, 4*f*, . . . and so on. L2 is another inductance using 22 SWG enamelled copper on a ferrite toroid. The changeover switch is external to the transmitter board, and is used to switch your aerial between the transmitter and the receiver; its wiring is shown in Figure 2.

Use a dummy load

A dummy load enables you to test your circuit without actually transmitting a signal. If you haven't such a thing already, it is easy to construct one to use with this transmitter. Don't use it for transmitters of more than 2 watts output, though. Use two 100 ohm, 1 watt resistors, connected in parallel across the end of a short piece of coaxial cable, terminated in a BNC, PL259 or N-type free plug, as shown in **Figure 3**. Plug this into the aerial socket on your transmitter, plug in your crystal and connect the transmitter to a 12 V supply. Have another receiver switched on and tuned to the crystal frequency. Although the radiation from your dummy load is minimal, it will

Figure 3 A simple 2 W dummy load can be made from two 100 Ω (ohm) resistors in parallel. The plug should match the socket on your transmitter

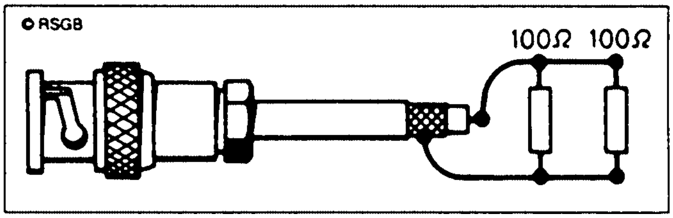

be enough to be picked up by a receiver in the same room. Send dashes with the Morse key, and adjust VC1 until the received note is *clean*. It should not sound rough, or have a *chirp* (change its frequency during a dash or dot). Avoid tuning for maximum power; this is seldom the correct setting!

You will need to put your completed transmitter in a metal box, using sockets for the power supply, aerial, receiver and Morse key. The sockets can be chosen to match your existing equipment.

Figures 4–7 are taken from the RSGB book *Practical Antennas for Novices*, and may give you some ideas on the type of aerial to be used with your transmitter.

Figure 4 A simple dipole can be very effective. For the 3.5 MHz band, length L is 40 metres and height H should be as large as possible. The far support S can be a tree, pole or building. Insulators I may be home made from strong plastic and the feeder F should be 50 Ω (ohm) coax cable

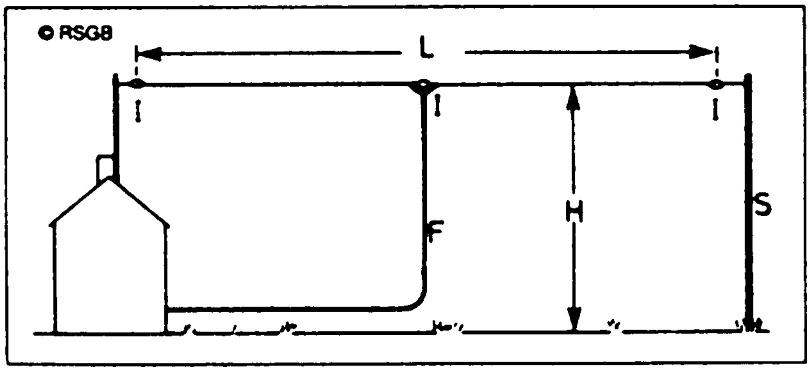

Figure 5 Your signal is radiated mostly from the centre of the dipole so the ends can droop or even be bent but the length may need shortening by a few centimetres because the ground and the bends will detune the dipole. Cords C are best made from strong plastic rope from a sailing or camping shop

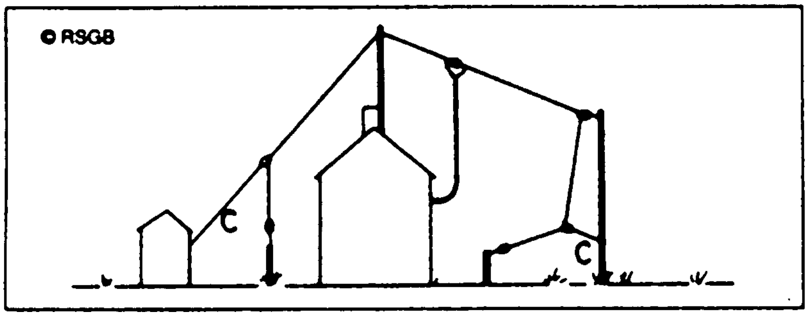

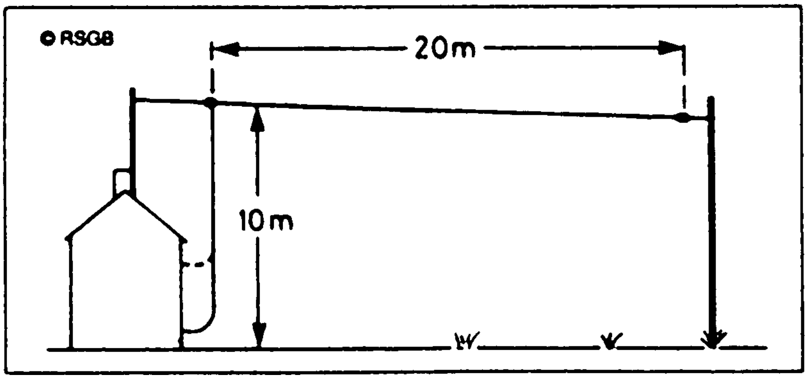
© RSGB
20m
10m

Figure 6 An 'inverted-L' takes up less space than a dipole and doesn't need coax cable. Like the dipole, the end can droop or be bent to save space as in this case most of the radiation comes from the area around the top of the vertical part

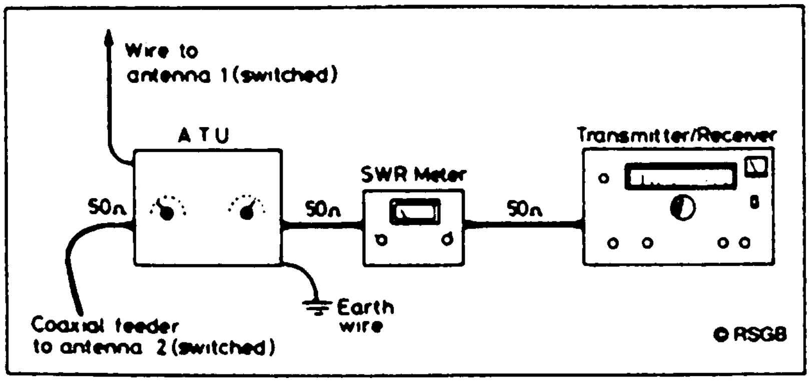
Wire to antenna 1 (switched)
ATU
50Ω
50Ω
SWR Meter
50Ω
Transmitter/Receiver
Earth wire
Coaxial feeder to antenna 2 (switched)
© RSGB

Figure 7 Almost any length of wire more than 10 m or so long will work (though it will work better the longer and higher it is) but an Aerial System Tuning Unit (ASTU or ATU) will be needed

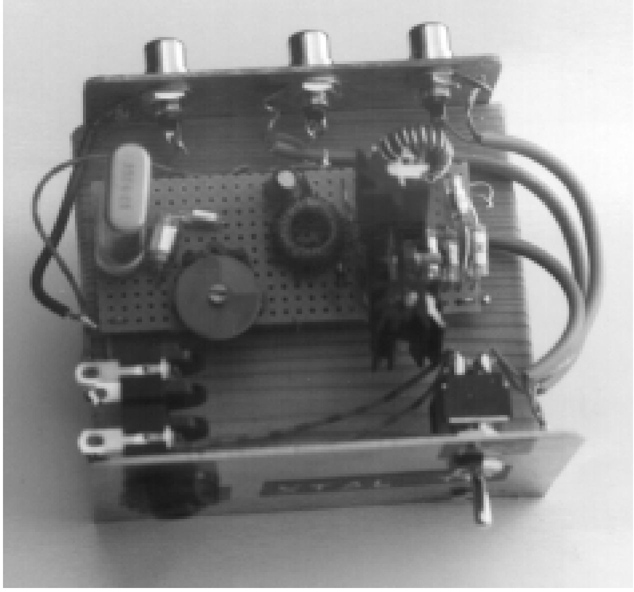

Parts list

Resistor

R1	12 kilohms (kΩ), 0.25 watt, 5% tolerance

Capacitors

C1	1000 picofarads (pF) polystyrene
C2	100 picofarads (pF) polystyrene
C7a, C8a	680 picofarads (pF) polystyrene
C7b, C8b	68 picofarads (pF) polystyrene
C3, C4	100 picofarads (pF) ceramic
C5	10 nanofarads (nF) polyester
C6	10 microfarads (μF) electrolytic 25 V
VC1	100 picofarads (pF) trimmer

Semiconductors

D1	1N4148
TR1	2SC2078 (see sources list)

Inductors

T1	38 turns 26 SWG enamelled copper on T–50–2 toroid, with two link windings of four turns
L1	21 turns 22 SWG enamelled copper on T–50–2 toroid

Additional items

- Ferrite bead
- Crystal (e.g. 3.579 MHz) and holder
- Metal box
- Socket for Morse key.
  This must be totally isolated from the metal of the box, as *both* connections can be at +12 V.
- Sockets for 12 V supply, aerial and receiver
- Switch – DPDT
- Heat sink for TR1
- RG174 miniature coaxial cable for signal leads (see Figure 2)

Component sources

Special components

2SC2078	Cricklewood Electronics Ltd, 40 Cricklewood Broadway, London, NW2 3ET.

37 A solar-powered MW radio

Introduction

What could be more ecologically friendly than a radio powered by the sun's energy? This design is quite standard, and if you have built any of the other medium-wave radios in this series, then this one should present few problems.

The solar panel

The solar panel is to the solar cell as the battery is to the cell; in other words a solar panel is several solar cells connected in series. The solar panel quoted for this radio will generate about 9 V at a current of around 30 mA on a sunny day. The circuit will operate on a supply of around 2 V, so bright sunshine is *not* necessary for satisfactory operation. The volume will be less, of course.

The circuit

The radio uses the ZN415E integrated circuit (IC), connected as shown in **Figure 1**, the complete circuit diagram. The signal is tuned in by the combination of L1 and VC1. L1 is made by winding about 35 turns of 24 SWG enamelled copper wire on a ferrite rod. A standard ferrite rod of 10 cm length and 1 cm diameter is used.

Signals selected by the tuned circuit are passed to IC1, which amplifies the signals and removes the audio component, which is then amplified further by IC2 for driving a small loudspeaker. The removal of the audio component is the process we call *detection*. In addition to this, IC1 provides *automatic gain control* (AGC), which helps to keep the audio signal constant, even when the incoming RF signal may vary due to fading.

The prototype board

Veroboard (also known as matrix board or stripboard) is ideal for the construction of the radio. The layout is shown in **Figure 2**. The board size is 11 strips by 30 holes (**please note that there is no row 'I', so take care with your counting!**). Using a 3 mm ($\frac{1}{8}$ inch) twist drill rotated between

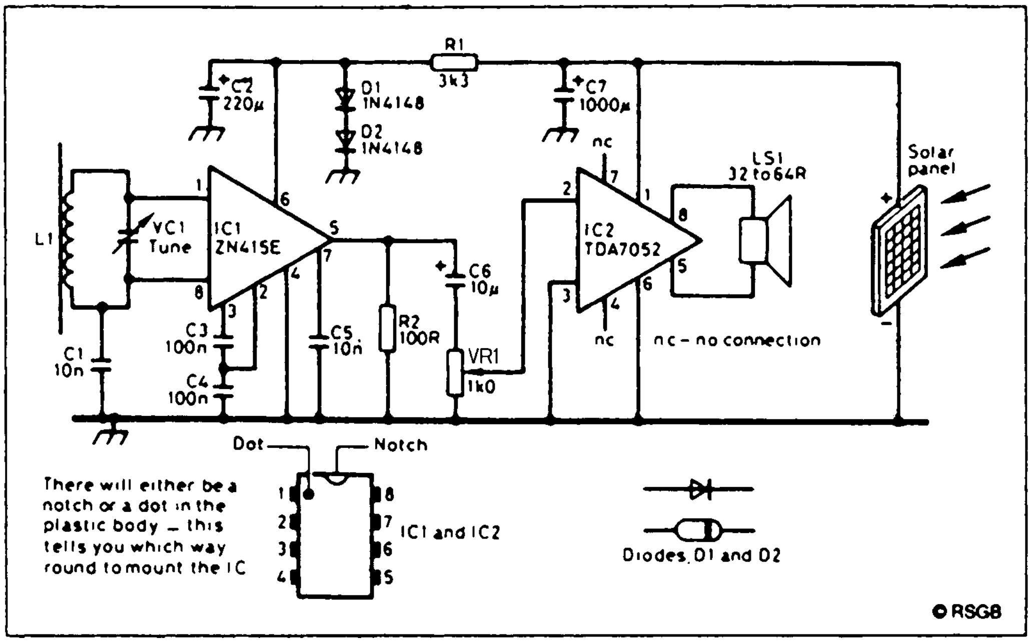
R1
3k3
C2
220µ
D1
1N4148
D2
1N4148
C7
1000µ
nc
LS1
32 to 64R
Solar panel
L1
VC1
Tune
IC1
ZN415E
IC2
TDA7052
C1
10n
C3
100n
C4
100n
C5
10n
R2
100R
C6
10µ
VR1
1k0
nc
nc – no connection
Dot
Notch
There will either be a notch or a dot in the plastic body – this tells you which way round to mount the IC
IC1 and IC2
Diodes D1 and D2
© RSGB

Figure 1 The radio uses just two integrated circuits (chips) and operates at any voltage from 2 to 9 V

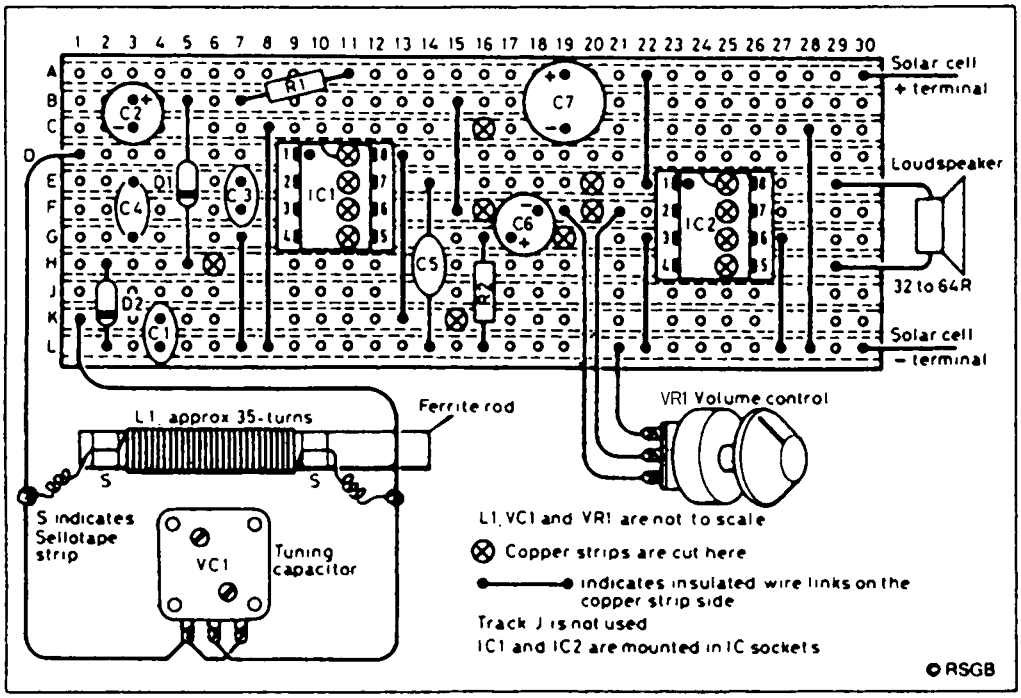
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
A B C D E F G H J K L
R1
C2
C7
D1
D2
C1
C3
C4
C5
C6
IC1
IC2
R2
Solar cell + terminal
Loudspeaker
32 to 64R
Solar cell – terminal
VR1 Volume control
Ferrite rod
L1 approx 35-turns
S
S
S indicates Sellotape strip
VC1
Tuning capacitor
L1, VC1 and VR1 are not to scale
Copper strips are cut here
indicates insulated wire links on the copper strip side
Track J is not used
IC1 and IC2 are mounted in IC sockets
© RSGB

Figure 2 Follow the layout carefully, making sure that all connections are neatly soldered

your thumb and forefinger, break the strips at the points marked with a cross inside a circle. Hold the board up to the light to ensure that the tracks *have* been broken and that there is no copper swarf shorting tracks together.

Firstly, solder in the 8-pin DIL sockets for the ICs, making sure that the notches in the sockets are facing upwards, as shown on the diagram. Then, solder in the wire links, resistors, capacitors and diodes, making sure that the electrolytic capacitors (C2, C6 and C7) and the diodes (D1 and D2) are connected the correct way round. Use different colours of wire for the connections to the volume control, VR1, to avoid incorrect connections. Note the wiring of the tuning capacitor (VC1) shown in Figure 2; a two-section type is used, and both sections are wired in parallel to give twice the capacitance of a single section.

There is no on/off switch – just turn the volume down when you are finished using the set! The solar panel can be mounted parallel to the top of the case, or angled to receive the maximum energy from the sun, as shown in the photograph. You could have a battery available as a standby source to use the radio after dark; any battery of between 6 V and 9 V will do. Wire it with a simple changeover switch, so you can switch between solar and battery power. Ask a friend for help with this if you are not sure how to do it.

You may need to adjust the number of turns on L1 to get the best results, but it should be possible to receive at least five stations at good volume with your sun-powered radio!

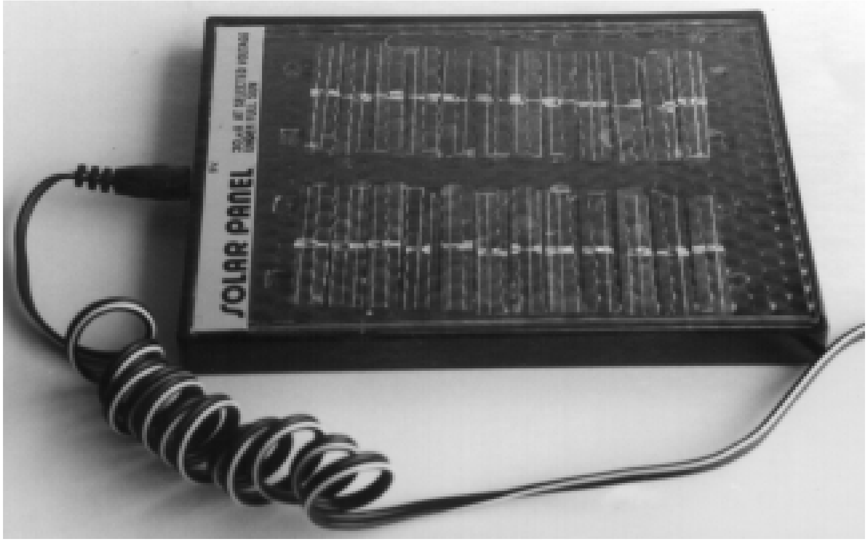

Parts list

		Maplin code
Resistors: 0.25 watts, 5% tolerance		
R1	3.3 kilohms (kΩ)	M3K3
R2	100 ohms (Ω)	M100R
VR1	1 kilohm (kΩ) log	FW21X
Capacitors		
C1, C5	10 nanofarads (nF) ceramic	BX00A
C2	220 microfarads (μF) electrolytic (10 V)	FB60Q
C3, C4	100 nanofarads ceramic	YR75S
C6	10 microfarads (μF) electrolytic (10 V)	FB22Y
C7	1000 microfarads (μF) electrolytic (10 V)	FB81C
VC1	140–300 picofarads (pF)	FT78K
Semiconductors		
IC1	ZN415E (or ZN416E)	UR70M
IC2	TDA7052	UK79L
D1, D2	1N4148	QL80B
Solar panel	9 V at 50 mA	RK23A
Additional items		
LS1	32–64 ohm miniature loudspeaker	YT28F
	Ferrite rod	YG20W
	24 SWG enamelled copper wire	BL28F
	Plastic box, approx. 220 × 140 × 70 mm	YN39N
	Veroboard, cut to size	JP47B
	8-pin DIL sockets, 2 required	BL17T
	Knobs, 2 required	FD67X
	Material for speaker grille	
	Connecting wire	

38 A receiver for the 7 MHz amateur band

Introduction

Listening on the 40 metre band (from 7.0 to 7.1 MHz) can be very rewarding – it is a popular haunt for HF Special Event stations, and at night there are signals to be heard from all over Europe. This receiver is designed purely for the 40 m band, and is ideal for those who have built the simpler receivers and are looking for something a little more challenging. The more experienced constructor may prefer to build this on prototype board.

The circuit and its construction

Figure 1 shows the circuit diagram. The receiver will work well with headphones or loudspeaker. Walkman-type headphones and speakers are ideal for use here.

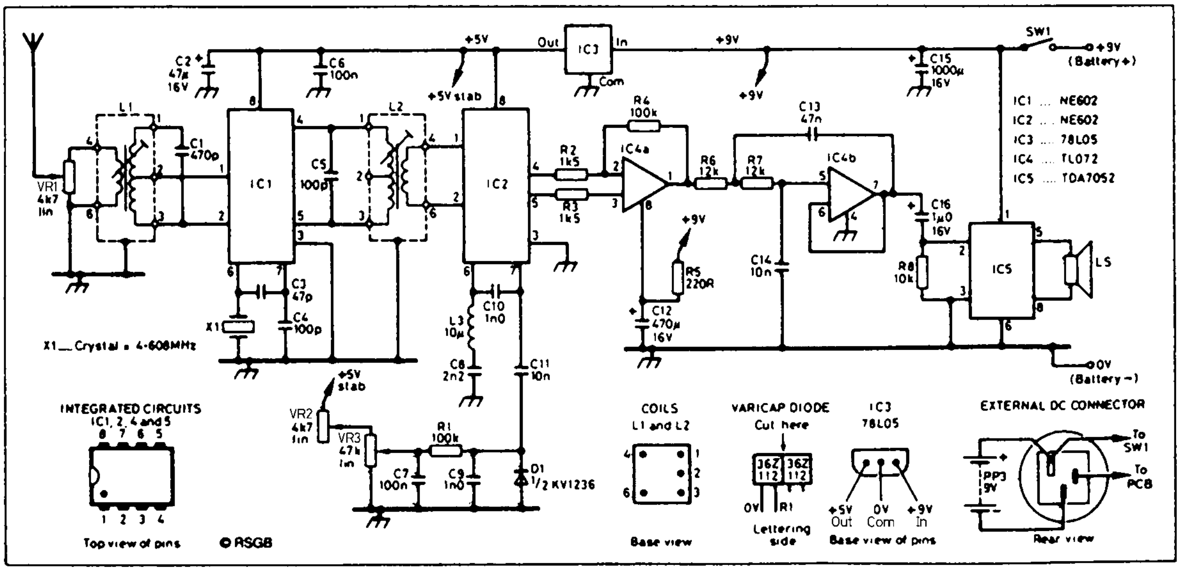
C2 47µ 16V
C6 100n
+5V
Out
IC3
In
Com
+5V stab
+9V
+9V
C15 1000µ 16V
SW1
+9V (Battery+)
L1
L2
C1 470p
VR1 4k7 lin
IC1
C5 100p
IC2
R2 1k5
R3 1k5
R4 100k
IC4a
R6 12k
R7 12k
C13 47n
IC4b
+9V
R5 220R
C14 10n
C12 470µ 16V
C16 1µ0 16V
R8 10k
IC5
LS
0V (Battery −)
IC1 ... NE602
IC2 ... NE602
IC3 ... 78L05
IC4 ... TL072
IC5 ... TDA7052
C3 47p
C4 100p
X1
X1 — Crystal = 4·608MHz
L3 10µ
C10 1n0
C8 2n2
C11 10n
+5V stab
VR2 4k7 lin
VR3 47k lin
R1 100k
C7 100n
C9 1n0
D1 ½ KV1236
INTEGRATED CIRCUITS
IC1, 2, 4 and 5
8 7 6 5
1 2 3 4
Top view of pins
© RSGB
COILS
L1 and L2
Base view
VARICAP DIODE
Cut here
36Z 112
36Z 112
0V
R1
Lettering side
IC3
78L05
+5V Out
0V Com
+9V In
Base view of pins
EXTERNAL DC CONNECTOR
PP3 9V
To SW1
To PCB
Rear view

Figure 1 The receiver gives good performance on the 7 MHz amateur band as well as being simple to construct and align

Signals arriving at the aerial are coupled into IC1 via gain control VR1, which also functions as the on/off switch. Tuning is provided by varying the voltage on the *varactor diode* (or varicap), D1. VR3 is the main tuning control, and VR2 is the *bandspread* (fine tuning) control. The varactor diode is supplied as a dual device, which must be cut down the middle *carefully* with a sharp knife; with the lettering upwards, the ground lead (0 V) is on the left-hand side, as Figure 1 illustrates.

Solder in the IC sockets first, followed by the coils. After this come the links, resistors, capacitors and varactor diode. Ensure that IC3, the voltage regulator, is wired correctly, and check the polarity of the electrolytics. The crystal, X1, is very fragile, so take extra care with it. The wiring of the three controls is shown in **Figure 2**.

Before putting the ICs in their sockets, connect up the battery and check the following voltages with the negative voltmeter lead connected to the negative terminal of the battery:

Pin 8	IC1	5 V
Pin 8	IC2	5 V
Pin 8	IC4	9 V
Pin 1	IC5	9 V

When all these have been found to be correct, switch off and put the ICs carefully into their sockets. Use wire of different-coloured insulation to wire up the front-panel controls.

The case can be a small plastic box of size 22 cm by 15 cm by 8 cm, with three 10.5 mm holes drilled in the front and two 8 mm holes in the side for the aerial and earth connections. On one side are a 6 mm hole for the speaker socket and an 11 mm hole for the optional external power supply.

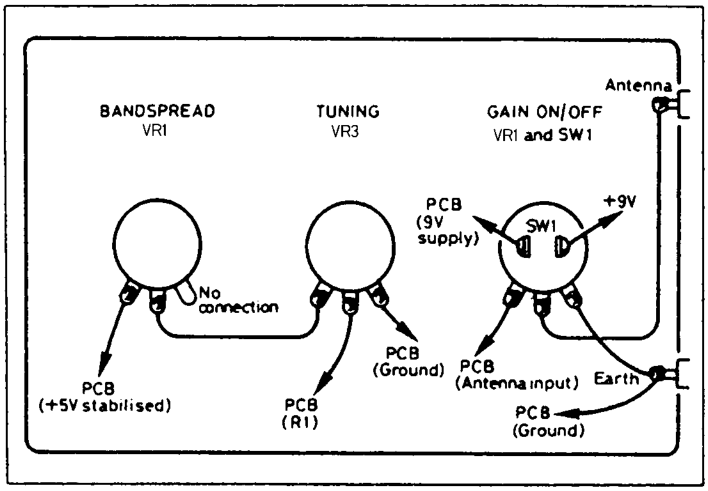

Figure 2 Rear view of the variable resistors. Check the connections carefully to make sure the wires are fitted correctly

Testing and tuning

The aerial for the receiver should be between 30 and 70 feet of wire, mounted as high as you can make it, away from trees and buildings if possible. Connect the battery and switch on. Adjust L1 and L2 for the best results. Tune slowly with VR3; you should find CW stations at the lower end of the band (anticlockwise) and SSB stations at the upper end (clockwise). You may find that it is easiest to make these adjustments *before* mounting the board in the case with double-sided sticky tape or pads. If you are planning to use an external DC supply, make sure it is a safety approved stabilised 9 V type, and **disconnect the battery before you use such a supply!**

If you suspect that the tuning doesn't quite cover the lower CW end of the band, try increasing C9 to 1200 pF. If it is the upper SSB end which is missing, decrease C9 to 820 pF.

It is always advisable to use an aerial tuning unit (ATU) between your aerial and the receiver. A suitable design of ATU is included as a project in this book.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	100 kilohms (kΩ)
R2, R3	1.5 kilohms (kΩ)
R5	220 ohms (Ω)
R6, R7	12 kilohms (kΩ)
R8	10 kilohms (kΩ)
VR1	4.7 kilohms (kΩ) linear, with SPST switch
VR2	4.7 kilohms (kΩ) linear
VR3	47 kilohms (kΩ) linear

Capacitors: all rated 16 V or more

C1	470 picofarads (pF) polystyrene 5%
C2	47 microfarads (μF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4, C5	100 picofarads (pF) polystyrene 5%
C6, C7	100 nanofarads (nF) ceramic
C8	2.2 nanofarads (nF) polystyrene 5%
C9, C10	1 nanofarad (nF) polystyrene 5%
C11, C14	10 nanofarads (nF) ceramic
C12	470 microfarads (μF) electrolytic
C13	47 nanofarads (nF) ceramic
C15	1000 microfarads (μF) electrolytic
C16	1 microfarad (μF) electrolytic

Inductors

L1	Toko KANK3335R
L2	Toko KANK3333R
L3	10 µH 5%, e.g. Toko 283AS–100

Semiconductors

IC1, IC2	NE602 or NE602A
IC3	78L05 5 V, 100 mA
IC4	TL072
IC5	TDA7052

Additional items

D1	Toko KV1236 cut into two sections (one half used)
X1	4.608 MHz (available from Cirkit)
	3 × silver knobs, one with pointer
	Plastic case approx. 22 × 15 × 8 cm
	Speaker 8–32 Ω, or headphones
	4 × 8-pin DIL sockets for IC1, IC2, IC4, IC5
	2 × 4 mm sockets (red and black) for aerial and earth
	3.5 mm chassis-mounting jack socket for speaker
	DC power socket for external supply (if required)
	Prototype board

Kits

A complete kit is available from JAB Electronic Components.

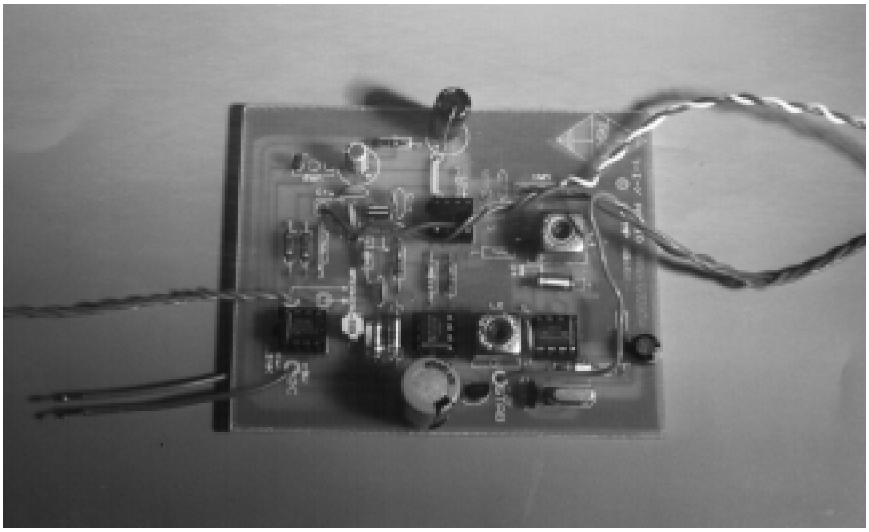

39 Diodes for protection

Introduction

Many semiconductor devices can be destroyed in an instant if their supply is reversed. With the use of batteries as power sources, it is quite common for the battery to be connected 'the wrong way round', and the scope for damaging equipment is significant and very real. Diodes can protect equipment in several ways, and you may do worse than to consider one of these approaches to protect your next expensive project.

Choose wisely!

All the diode circuits given here are so simple as to invite calamity. The circuits are not foolproof but, with a little care, will work first time. Remember that a semiconductor diode has a forward voltage drop of between 0.5 V and 0.7 V, depending on its type and the current flowing. This will be mentioned later.

The first thing you need to do is to insert a good multimeter *in series* with the circuit you want to protect; set the range to Amps DC, and switch on. Check that the circuit works properly. Then, decrease the current range on the meter until a good reading is indicated. Make a note of this current, as it is the normal running current of your circuit.

To choose a diode, you must consult the catalogues or data sheets and find one where the quoted *maximum forward current* exceeds the current you have measured; preferably it should be at least twice your measured current. Secondly, the diode will have a *peak inverse voltage* (PIV); this is the maximum voltage it can withstand when the cathode is made positive with respect to the anode, i.e when it is *reverse-biased* and not conducting. This voltage must be greater than your battery voltage, again by a factor of about 2. Except in the case of the bridge rectifier (see later), these criteria will enable the selection of a suitable diode to be made easily.

The series diode

The simplest and most obvious way to protect equipment is to insert a diode in the positive supply lead, as shown in **Figure 1**, with the diode passing current *only* when the supply is of the correct polarity. Because of the 0.6 V that exists across the diode, your equipment will normally operate on a

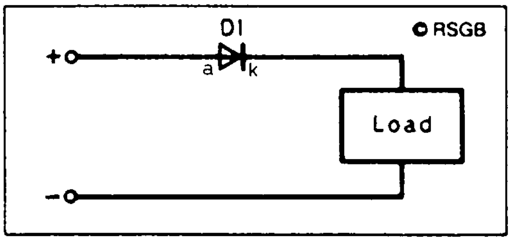

Figure 1 Series diode protection

slightly lower voltage. If you imagine reversing the supply to this circuit, you will see that the negative terminal of the battery is connected to the anode of the diode; the diode becomes reverse-biased and will not conduct any current, thus protecting the load. Your equipment will not operate when the battery connections are reversed.

The parallel diode

The circuit of **Figure 2** overcomes this voltage drop. It places the diode in parallel with the load (your equipment) but in a normally reverse-biased condition so that it draws no current when the battery is correctly connected. Reverse the battery connections, however, and a very large current will flow through the diode, thus blowing the fuse! For this technique to work successfully, the current drawn by the diode when the battery connections are reversed *must* be much greater than the maximum current drawn by the equipment, in order to blow the fuse. This is usually not a problem, however.

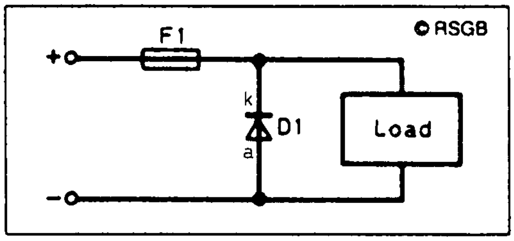

Figure 2 Parallel diode and fuse

The diode bridge

For sheer elegance, the circuit of **Figure 3** takes the biscuit! It uses four diodes connected in the form known as a *bridge rectifier.* Such rectifiers exist, and do not have to be made up from four discrete diodes.

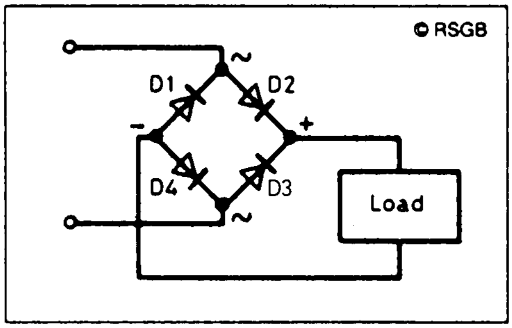

Figure 3 The diode bridge

Follow the current round the circuit from the supply, assuming initially that the top wire is positive and the bottom wire negative. It flows from the positive supply

(a) through D2
(b) through the load (top to bottom)
(c) through D4

and back to the negative of the supply.

Now assume that the bottom supply lead is positive and the top lead negative. The current flows from the positive supply

(a) through D3
(b) through the load (top to bottom)
(c) through D1

and back to the negative of the supply.

So, whichever way round the battery is connected, the current will *always* flow the same way through your equipment!

The circuit does have a drawback, however. Whichever way round the battery is connected, there are always **two** diodes conducting the current at any time. In the first case it is D2 and D4; in the second case, D1 and D3. This means that there is a total voltage drop of about 1.2 V. If your equipment can tolerate that reduction in voltage, then you will not have a problem.

Decoupling

Whenever the supply rail to a piece of equipment, or even to an individual stage of a circuit, is broken for the insertion of a device that will drop voltage, strange things can happen. This is because the supply for any circuit is assumed to have a low resistance to DC and a low *impedance* to AC. (Impedance is the AC equivalent of DC resistance.) These two are not the same, and the insertion of a diode or diodes is certain to make a big difference to them both. To overcome any irregularities in the operation of the circuit you are protecting in any of the ways described previously, the protected supply needs to be *decoupled* with a capacitor.

If the circuit handles audio frequencies only, placing an electrolytic capacitor directly across the load in all three circuits should solve the problem. The parallel circuit of Figure 2 is less at risk than the other circuits. The size of the capacitor will be determined by the current taken by the circuit, and may need to be chosen within the range 100 μF to 10 000 μF, with a working voltage greater than the supply voltage.

If the circuit is mainly handling RF currents, placing a capacitor of 0.01μF across the load should prevent any problems. A second capacitor, also across the load, of between 10 μF and 100 μF may be needed. Again, the parallel circuit of Figure 2 is less at risk than the other two.

Don't be afraid to experiment, but confine your experimenting (at first) to small equipment and low currents, until you get a 'feel' for the technique.

40 An RF signal probe

Introduction

A radio-frequency (RF) diode probe is a simple device which, when used with a conventional multimeter, enables the measurement of RF voltages in a circuit. When constructed, this will be one of the most useful pieces of test equipment for the experimenter who revels in the construction of transmitters and receivers.

The circuit

Figure 1 shows the simple circuit diagram. It is *almost* the same as a common diode rectifier circuit, but with a simple change to make it more sensitive. The circuit is known as a *voltage-doubler*, and is often found in high-voltage supplies, with beefier capacitors and diodes, of course! Because we are dealing with high frequencies and smaller voltages, the diodes and capacitors can be physically very small. The diode circuit of D1 and D2 *rectifies* or *detects* the RF from the probe, and any remaining AC is removed (short-circuited to ground) by C2. This produces, at the output, a constant voltage proportional to the peak-to-peak RF voltage present at the input; the output voltage is fed to an ordinary multimeter (on a voltage range).

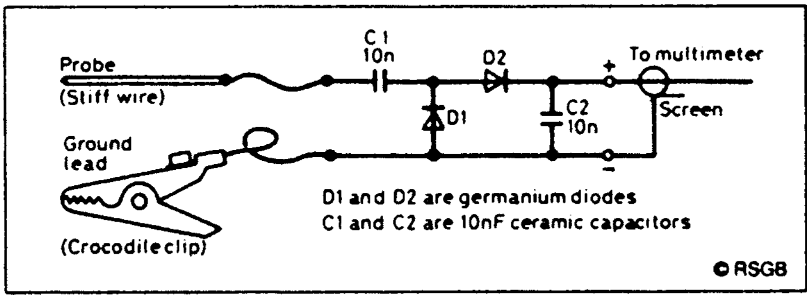
C1
10n
Probe
(Stiff wire)
D2
+
To multimeter
D1
C2
10n
Screen
Ground
lead
−
D1 and D2 are germanium diodes
C1 and C2 are 10nF ceramic capacitors
(Crocodile clip)
© RSGB

Figure 1 The diode probe can be made from four electronic components

Construction

Although the circuit layout is not critical, a description of the prototype is given here for your information. **Figure 2** has the details. The components are soldered to square pieces of copper-clad printed-circuit board (PCB) glued to a larger piece of PCB. The larger piece serves as a ground connection. Use a stiff copper wire as the probe, and an insulated flexible wire with a crocodile clip to connect the probe to the ground of the circuit under test.

Cut a piece of plain PCB, 30 mm by 45 mm and another of 15 mm by 45 mm. Cut the smaller piece into three measuring about 15 mm square. Stick the three small pieces to the larger piece, ensuring that there are small gaps between each, as Figure 2 shows. Solder the components in place. The connection to the multimeter should be thin coaxial cable or screened microphone cable. If you use unscreened cable, there may be RF pickup here which can lead to false readings.

Simple to use

Using the probe is simple. Connect it to the multimeter and set the meter to around 10 V DC – you may need to reduce this, depending upon the magnitude of the RF voltage you are trying to measure. Hold the probe by

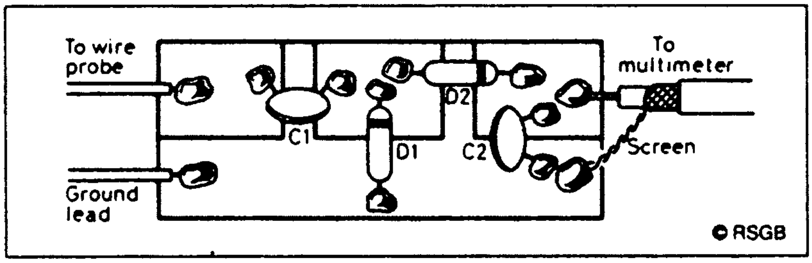
To wire
probe
C1
D1
D2
C2
To
multimeter
Screen
Ground
lead
© RSGB

Figure 2 The components are mounted on a copper-clad PCB

its base, being careful *not* to touch any of the components. Connect the croc clip to a ground point on the circuit and touch the probe on the circuit point to be tested. If an RF signal is present, there should be a reading on the multimeter. That's all there is to it!

Parts list

Capacitors	
C1, C2	10 nanofarads (nF), or 0.01 microfarad (μF) ceramic
Semiconductors	
D1, D2	Any *germanium* signal diode, e.g. OA91 or AA119

Additional items
- Stiff copper wire
- Insulated flexible wire
- Copper-clad PCB
- Crocodile clip
- Coaxial or screened cable

Source

All components are available from Maplin.

41 An RF changeover circuit

Introduction

The simplest entry route to operating an HF station is to buy or build a receiver, and to add a simple CW (Morse) transmitter. In theory, this sounds so simple, yet in practice there is one significant hurdle to be overcome. How can they both share the same aerial? A manually operated changeover switch is the obvious solution, but it takes time to perform the switching operation, by which time someone else has squeezed in before you and contacted the DX station! This small circuit accomplishes the changeover automatically, as soon as the transmitter is keyed.

How it's done

The circuit is shown in **Figure 1**, and uses the signal from the transmitter to operate a relay. A relay is a switch which is operated electrically in the following way. A coil of wire with an iron core is used as an electromagnet. When current flows through the coil, it produces a magnetic field which, in turn, is used to pull a set of switch contacts. These contacts will be used to switch the aerial from transmitter to receiver and vice versa. The relay used here has a double-pole changeover (or DPDT, or DPCO) switch. If you are confused by the different types of switch, look at the basic descriptions elsewhere in this book. The *pole* of a switch is the part that doesn't move; in Figure 1, one pole is connected to L1 and the other to the aerial output. In a circuit diagram, the switch contacts are always shown in their *normal* state, i.e. when *no* current flows in the relay coil. The changeover switch is normally in the receive position, so we say the receive switch is *normally closed*, and the transmit switch is *normally open*. We choose to have the circuit in the receive position normally, because everyone spends a lot more time receiving than transmitting. The relay is energised only when you are transmitting.

There are three RF sockets – one for the transmitter, one for the aerial, and one for the receiver. When the transmitter is not in use, there is a direct connection (via the Rx normally closed contacts of the relay) between the aerial and the receiver. The transmitter (which is not keyed at this time) is connected to a 50 ohm dummy load, R1.

When the transmitter is activated by pressing the Morse key, an RF signal appears at the transmitter socket. The wire carrying the signal to the relay

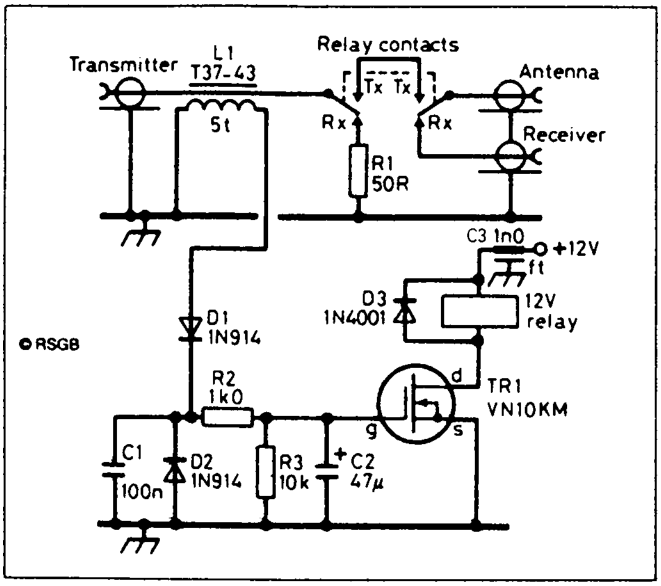

Figure 1 RF activated relay

Tx contacts passes through a toroid (a ferrite coil former, looking like a black Polo mint), which has a five-turn coil wrapped around it. This five-turn coil picks up the signal, which is then detected by D1 and D2, and converted to a steady voltage. C1 removes any remaining RF and C2 provides *hang*. Without it, the circuit would detect the gaps between every dot and dash being sent, and switch the aerial over very rapidly and very frequently! This is *not* what we want. We need the relay to remain in the transmit condition at normal Morse keying speeds, and then return to receive when the sending is complete. The combination of C2 and R3 achieves this.

The voltage appearing across C2 is fed to a VMOS field-effect transistor (FET), TR1, which acts as an electronic switch. When a voltage appears across C2, the FET switches on and passes a current through the relay coil, changing the contacts from receive to transmit. The diode, D3, across the relay winding protects the FET from being damaged by the large reverse-voltage spike which occurs across the coil when the FET switches off.

Construction

This is very simple – a circuit board is not required. The prototype was built into an old 50 g tobacco tin. All ground leads are soldered directly to the tin, thus supporting the components automatically. Phono sockets were used for the aerial, receiver and transmitter; you could use whatever connectors matched the rest of your station. The 12 V supply is fed through the tin using a 1000 pF feed-through capacitor, C3.

The ‘dummy load’, R1, must be able to withstand the RF output power of the transmitter. For novice use, 4 watts is adequate. Two suggestions for making up R2 from standard resistors are shown in **Figure 2**. Six 330 ohm, 1 watt resistors in parallel have a combined resistance of 330/6 = 55 Ω at 6 W. Two 100 ohm, 2 watt resistors in parallel have a combined resistance of 100/2 = 50 Ω at 4 W.

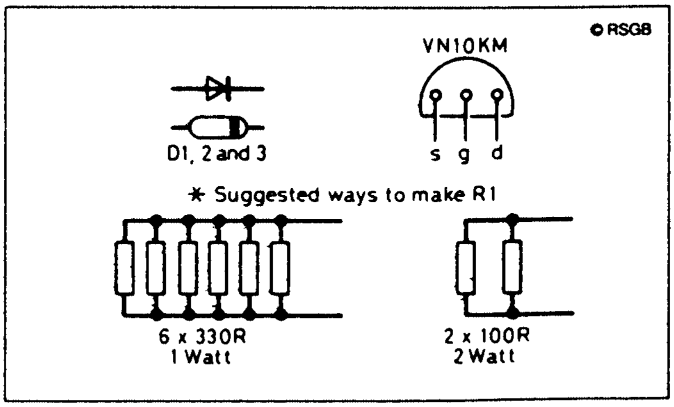

Figure 2 Component details

Make sure all the contacts to the relay are correct; also ensure that TR1, C2, D1, D2 and D3 are correctly wired.

Winding the toroid, L1, is very simple. It is made from PVC-insulated hook-up wire. Each time the wire passes through the centre of the toroid counts as one turn. The wire from the transmit socket to the relay simply passes through the centre! Use thin 50 Ω coax for the leads to the three sockets.

What happens to the receiver?

On transmit, the receiver is not connected to the aerial, but is only the separation of the switch contacts (about 0.5 mm!) away from the transmit lead. The receiver *will* pick up the transmitted signal, and there is a possibility that this signal will be enough to damage the receiver's sensitive input circuits. This can be prevented with the simple addition shown in **Figure 3**. Two diodes are connected back to back across the receiver socket. These act as a *limiter*, reducing the amount of signal that can enter the receiver. Solder these directly between the receiver socket and the tin.

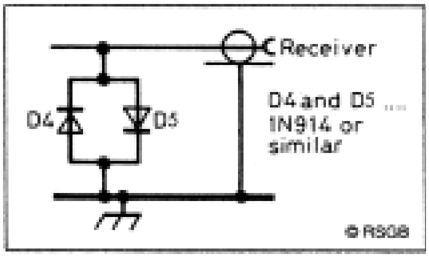

Figure 3 Receiver input protection

Warning

The wiring **must** be thoroughly checked. If the relay switch connections are wrong, there is a **great probability** that the full RF power will be applied to your receiver! You can check these connections as follows. With the 12 V supply connected, check that the relay energises (clicking sound heard) when the FET drain (d) is shorted to earth. Check the aerial and receiver connections with an ohmmeter or continuity tester. Disconnect the 12 V supply, and check that there is continuity between the centre pins of the aerial and receiver sockets. Reconnect the 12 V supply, short the FET drain to earth again (this will not damage the FET) and check that there is continuity between the centre pins of the aerial and transmitter sockets. If these tests show correct operation, disconnect the drain shorting wire (*most* important!) and your aerial auto-changeover is ready for use! Avoid using the device without the lid fitted.

Parts list

Resistors: 0.25 watt, 5% tolerance, unless otherwise stated

R1	50 ohms (Ω) – see text and Figure 2
R2	1 kilohm (kΩ)
R3	10 kilohms (kΩ)
Capacitors	
C1	100 nanofarads (nF), 0.1 microfarad (μF)
C2	47 microfarads (μF) 15 V electrolytic
C3	1000 picofarads (pF) feedthrough
Semiconductors	
D1, D2, D4, D5	1N914
D3	1N4001
TR1	VN10KM
Additional items	
L1	T37–43 toroid
	Relay 12 V DPDT (DPCO) relay
	Three sockets to suit (phono, SO239, etc.)
	Thin coax cable and hook-up wire

42 A low-light indicator

Introduction

This is a simple one-evening project that can be built for the pure fun of it, or to use as the basis of a more complex project to switch your shack lights on when it gets dark! In its prototype form, it simply flashes an LED when the ambient light level drops to a preset point.

Operation

The heart of the circuit shown in **Figure 1** is a *photo-conductive cell*, also called a *light-dependent resistor* (LDR), a device whose resistance changes according to the amount of light falling on it. In bright light, the resistance is low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting material known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given

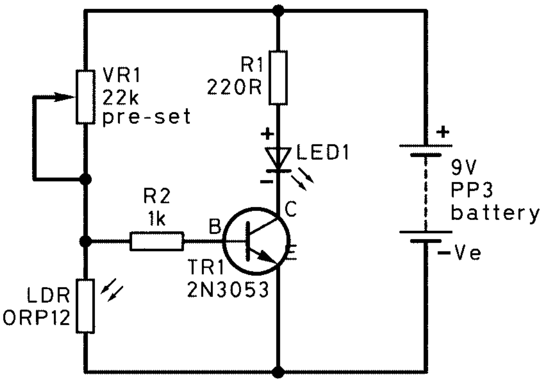
VR1
22k
pre-set
R1
220R
+
LED1
-
9V
PP3
battery
-Ve
R2
1k
B
C
E
TR1
2N3053
LDR
ORP12

Figure 1 Circuit diagram

©RSGB DY107

length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern, as can be seen diagrammatically in **Figure 2**. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

In this circuit, the LDR is connected across the 9 V supply in series with a variable resistor, VR1. In this arrangement, the voltage that exists across the LDR will be determined by the light level. As the light intensity increases, the resistance of the LDR falls, dropping a smaller voltage across it. The

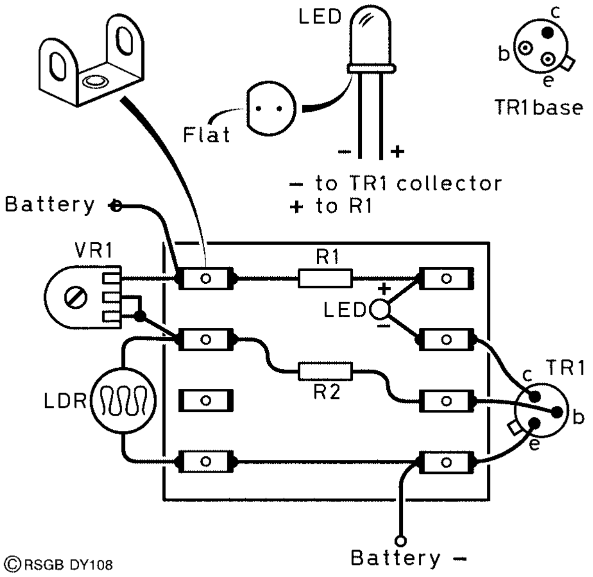
LED
Flat
- +
- to TR1 collector
+ to R1
c
b
e
TR1 base
Battery +
VR1
R1
+
LED
-
LDR
R2
c TR1
b
e
Battery -

Figure 2 Component layout on tagstrip

©RSGB DY108

reverse happens when the light intensity falls – the voltage across the LDR increases. This voltage is used to drive an npn transistor, TR1, connected as an electronic switch. As the voltage on the base (B) rises, it will reach a point where TR1 will suddenly start to pass current, just as an ordinary switch does when pressed. The current flowing through TR1 also flows through a flashing light-emitting diode (LED), D1 and its series resistor, R1. D1 can be a steadily glowing type, if preferred. If you want to make the circuit switch the LED on at a different ambient light level, adjust VR1.

Construction

The prototype was made on an eight-tag tagboard (Figure 2). The resistors R1 and R2 can be laid on the tagboard for soldering, the rest of the components lying above or to the side of the board. Check the circuit after you have soldered everything on, then connect the battery. The LED should flash if you put your hand over the LDR, and VR1 can be adjusted to vary the point at which the LED lights. The project can be housed in a small plastic box with holes provided for the LDR and LED and an on/off switch if you want one.

Parts list

Resistors: all 0.25 ohm, 5% tolerance

R1	220 ohms (Ω)
R2	1 kilohm (kΩ)
VR1	22 kilohms (kΩ) linear pre-set

Semiconductors

LDR	ORP12
TR1	2N3053 npn
LED1	Flashing LED

Additional items

- Plastic box
- Tagboard with 8 tags
- PP3 battery connector

43 A J-pole aerial for 50 MHz

Introduction

For FM communication (i.e. voice and data) on the VHF bands, a vertical aerial is used to give all-round (non-directional) coverage. This is a half-wave aerial which can be fed at the end, thus removing the principal problem with the conventional vertical centre-fed half-wave dipole, which is that the feeder should leave the dipole at right angles. This is no problem when the dipole is horizontal, but can be difficult for the vertical dipole.

Basic facts

A feeder must be connected to an aerial at a point where the *impedance* (AC 'resistance') of the aerial closely matches that of the feeder. The difference between the two impedances gives rise to the voltage standing-wave ratio (VSWR), which is unity only when the two impedances are the same. With 50 Ω feeders, the feed point of a half-wave aerial is at the centre, where the aerial impedance is around the same value. At the end of a half-wave aerial, the impedance is high, so it is *not* a suitable point to connect a 50 Ω feeder.

Connection at this point can be effected using an RF transformer. RF transformers act in the same way as ordinary transformers, except that they are much smaller, and usually comprise wires of particular lengths adjacent to each other. **Figure 1** is a good starting point. It shows the aerial in its diagrammatic form. Notice that the aerial is in the form of an elongated letter 'J'; this shape gives rise to its nickname – the *J-pole*. The quarter-wave RF transformer is the lower 'U' section below the half-wave element. At the bottom of the U section, the impedance is zero (this may become clearer later) and at the top of the U section it is high, thus matching the aerial impedance. The coaxial feeder cable is connected part-way up the U section, where the impedance is around 50 Ω.

The practicalities

Figure 2 shows the aerial as constructed. As it is about 4.5 metres high, it may be too high for the average house loft, but is ideal for mounting outside, supported by a non-metallic pole or hung from a tree branch. The upper half-wave section is made from 1.5 mm insulated copper wire, as used in domestic mains wiring. The quarter-wave transformer below the half-wave section is made from 300 Ω balanced line ('ribbon cable'). The

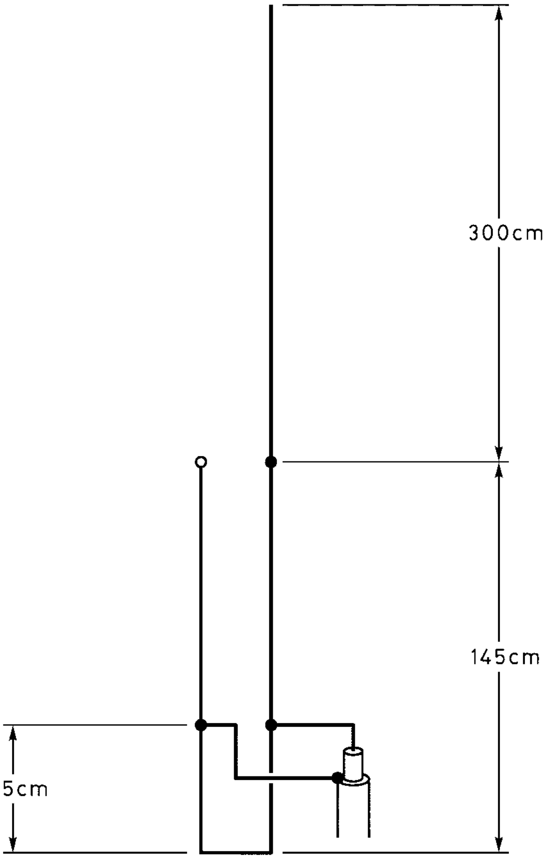

Figure 1 Overall dimensions of the 6 metre J-antenna

wires at the bottom of the transformer section are stripped of their insulation, twisted and soldered. At the upper end, only one wire of the balanced line is soldered to the bottom of the half-wave section. The other wire of the pair is not connected and is left insulated.

At the feed point of the transformer, the insulation needs to be carefully stripped from the balanced line. You will need a standing-wave meter (VSWR meter) in the coaxial line between your transmitter and the aerial, and you will need to adjust the position of the feed point. 45 cm from the bottom was the best point on the prototype, but this position is dependent upon the immediate surroundings of the aerial, and *must* be done when the aerial is in its final operating position. **Warning: *Never* make adjustments to the feed point when the transmitter is on. Make a VSWR measurement, switch off, move the feed point, switch on again, make another measurement, and so on.** You will need to aim for the lowest VSWR you can – certainly better than 2:1. Having found the best position, wrap all the exposed wires with self-amalgamating tape, to seal them against the ingress of moisture.

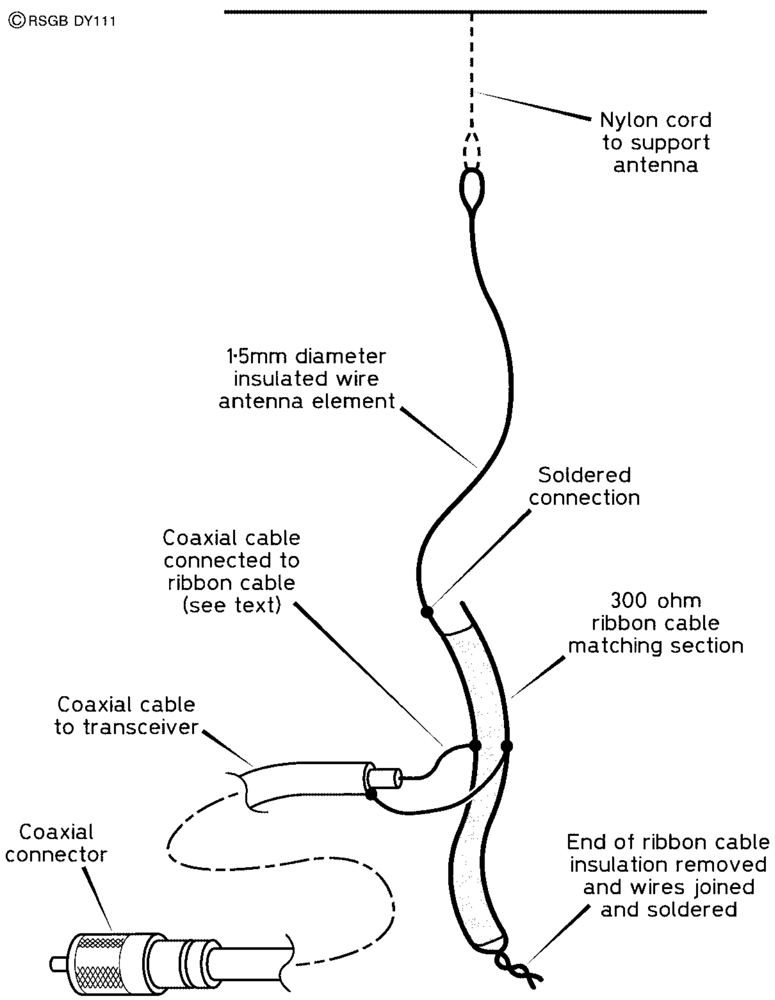
©RSGB DY111
Nylon cord to support antenna
1·5mm diameter insulated wire antenna element
Soldered connection
Coaxial cable connected to ribbon cable (see text)
300 ohm ribbon cable matching section
Coaxial cable to transceiver
Coaxial connector
End of ribbon cable insulation removed and wires joined and soldered

Figure 2 Construction details

How it performs

Figure 3 shows a computer prediction of how the J-pole radiates. It is called a *polar diagram*, and shows the distribution of your transmitted power when viewed ‘from the end of your garden’. Most of your signal is sent at a fairly small angle to the horizontal; very little signal goes upwards, which is a good thing, of course. This also shows why the J-pole (or any other vertical aerial) should not be called ‘omnidirectional’, which means it radiates in all directions. It is omnidirectional *only* in the horizontal plane.

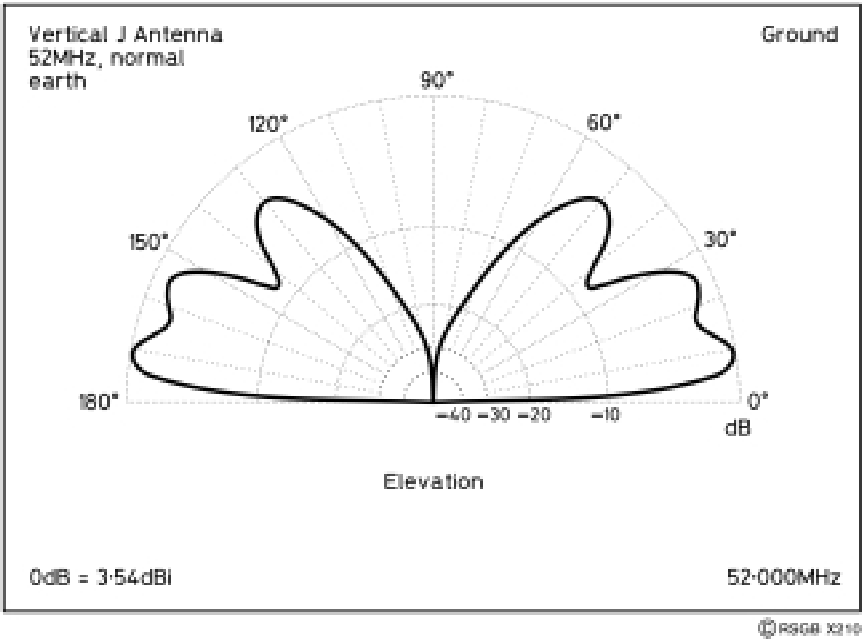

Figure 3 This polar diagram computed for the J-pole

Safety

Where you mount your aerial is a matter of personal preference and the restrictions of height and space, but the following safety rules *must* be applied.

1. Never fix an aerial where it may come into contact with power lines or telephone lines.
2. When climbing a ladder to put up an aerial outside, make sure the ladder is safe and that it is secured.
3. Don't do this alone. Preferably have someone with you. If this is not possible, make sure someone knows where you are.

Parts list

3.00 metres	1.5 mm insulated copper wire
1.50 metres	300 ohm balanced line
As required	50 ohm coaxial cable
As required	Self-amalgamating tape

44 Measuring light intensity – the photometer

Introduction

Before the days of automatic 'point-and-shoot' cameras, a photographer would use a light meter or *photometer* to measure the light level, then manually convert this reading into shutter speed and lens aperture settings to ensure a correctly exposed negative. Modern cameras have quite sophisticated photometers, which control the shutter speed and iris settings automatically.

A short explanation

A simple photometer circuit is shown in **Figure 1**, and is based on a device called a *light-dependent resistor*, or LDR. As its name tells us, its resistance depends upon the amount of light falling on it. In bright light, the resistance is relatively low (about 1 kΩ), whereas in the dark, its resistance is very high (up to 10 MΩ). The cell is made from a semiconducting substance known as cadmium sulphide (CdS), and is enclosed in a small plastic container. The semiconductor is laid on a flat insulating surface in the form of a small flat ribbon. The ribbon construction gives a good area of surface for a given length of ribbon, and the length of the ribbon is maximised by laying it out in a zig-zag pattern. In the dark, CdS is an insulator; when light falls on it, electrons are released inside the CdS, making it conduct. The more light there is, the more electrons there are, and the resistance falls.

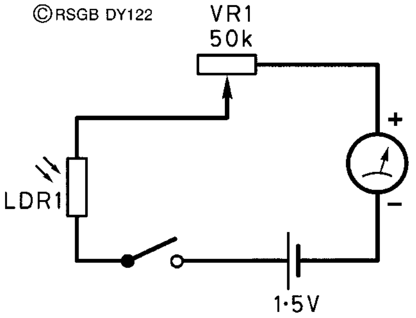

Figure 1 Circuit diagram of photometer

The photometer circuit

The circuit is simply a series connection of four things – the battery, the LDR, a variable resistor and a meter. A switch is also provided. The combination of the LDR resistance and that of VR1 determines the current flowing, which is indicated on the meter. Altering the resistance of VR1 sets the *sensitivity* of the photometer – you may want full-scale deflection of the meter needle for a bright light, or for a dim light.

It is simple to make, and a plug-in type matrix board is ideal to test the circuit, so that you can decide if you want to make a permanent version. Connect all the components in series; the only change you may want to make is the connection to VR1. If you find that the sensitivity control seems to work ‘backwards’, simply unsolder the wire from the end tag of VR1 and solder it to the opposite end tag. Problem solved!

In use

As soon as you connect up the battery, you will probably have a meter reading because of the daylight falling on the LDR. Shading it with your hand should reduce the reading. If the meter needle is hard over against the end-stop, turn VR1 until it indicates about half-scale. The LDR is very sensitive, and will read zero only in a dark room. If you put on a torch to see what the meter reading is, the LDR will detect the torch light, and will give a reading!

Here is a simple project where you can use the photometer in an experiment which has an analogy in radio. Draw a circle on a large (A3 or bigger) sheet of paper and divide it up into 30-degree sectors, as shown in **Figure 2**. Draw a series of smaller circles which divide the maximum radius into five. Look at the figure if you’re not sure about this. Bring the LDR away from the circuit by using two long, flexible wires. The experiment must be performed in a darkened room (preferably in total darkness). Prepare a table with two columns, the left-hand one headed ‘Angle (degrees)’ and the right-hand one ‘Meter reading’. Fill in the left-hand column 0, 30, 60 . . . and so on up to 360°.

Place the torch in the position shown, with its lens at the centre of the circle and pointing along the 0° line; switch it on. Place the LDR facing the torch and adjust VR1 until you have full-scale deflection of the meter needle. Suppose the meter indicates 10 units at this point. Enter this into your table in the 0° row. Keeping the torch the same distance from the circle centre, and pointing at it, move the LDR round all 30° positions and record the meter readings. Switch off the torch and take the sheet of paper and your tabulated results into daylight!

Lay the large sheet of paper on a table with your results beside it; then, at each 30° interval, plot the point along the radius corresponding to the meter

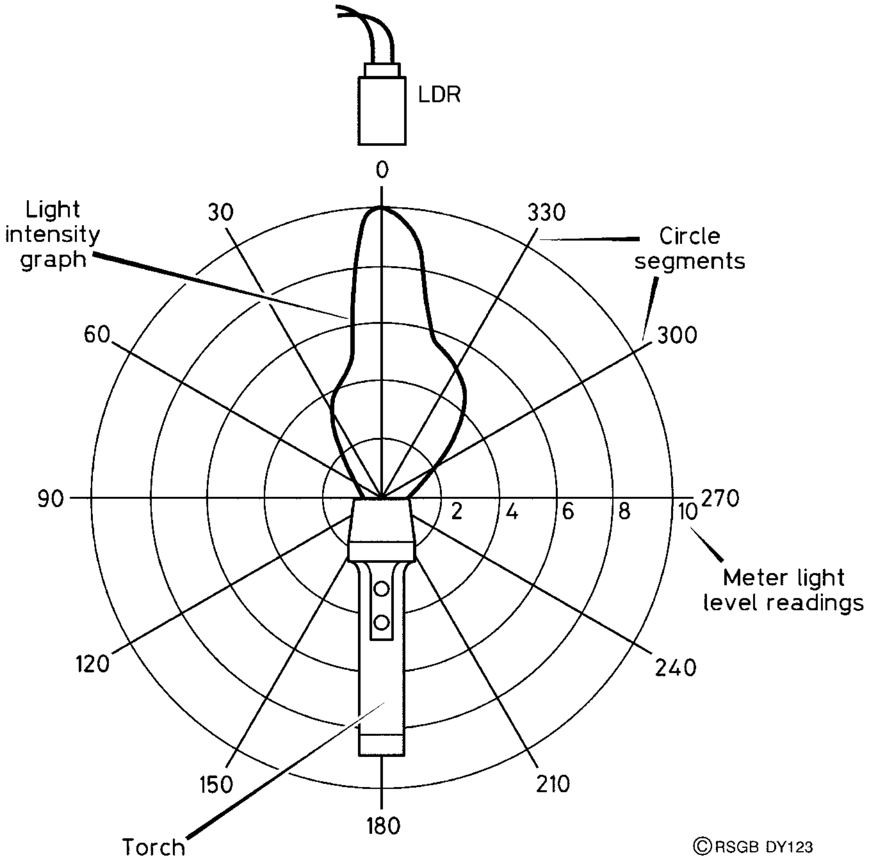

Figure 2 Torch light intensity pattern plotted at RSGB HQ

reading. Then, join up all the points and you have what is called a *polar diagram* of light intensity. The use of the word 'polar' implies that the readings have been taken in a circle and plotted that way.

Light waves and radio waves are both examples of *electromagnetic radiation*. The torch is designed to 'beam' its radiation in a particular direction, just like an aerial does. Hence the use of the word 'beam' for a directional aerial. If a similar polar diagram is drawn for a Yagi-type aerial, it will show the same general characteristics as does Figure 2, namely a main direction (or 'lobe') where most of the energy is concentrated, with evidence of *sidelobes*, indicated by 'lumps' on the otherwise smooth main lobe.

Also in this book you will find a project for the construction of a UHF field-strength meter, which you could use to carry out a measurement of the polar diagram of a UHF aerial.

Parts list

Resistor

VR1	50 kilohms (kΩ) linear

Semiconductor

LDR1	ORP12

Additional items
Meter, 50 or 100 μA
Battery clip for single AA cell
AA cell

45 A 70 cm Quad loop aerial

Introduction

This is a description of how to make an aerial for the 70 cm band which has gain compared with the 'rubber duck' or the dipole aerial. It can easily be dismantled and reassembled, making it ideal for contest use.

The principles

Most aerials which comprise several similar *elements* arranged along a *boom* are variations of the design originated by Yagi and Uda, and which takes the name (for historical reasons) of the former, and is know as the Yagi array. Let us suppose we have a Yagi aerial beaming left to right in front of us. The elements get progressively shorter from the left (look at almost any TV aerial to see this). All the elements on a Yagi aerial are classified as follows:

- The *reflector* – the leftmost element as we look at the array. It is the longest. Next to it is:
- The *driven element* – this is the element connected to the feeder, which in turn runs down to the transceiver. It is slightly shorter than the reflector.
- All the elements beyond the driven element are called *parasitic elements*, or *directors*. They are shorter than the driven element and usually get progressively shorter as we progress to the right along the boom. The directors are mainly responsible for the *directivity* (or *beamwidth*) of the array.

The progression from a simple dipole (a driven element in isolation) to a Yagi array is simple, but is nevertheless important. To make an aerial of two elements, a reflector (*not* a director) is added to the driven element. For three or more elements, directors are added to the two-element design. Adding more and more directors soon becomes impractical, the reduction in beamwidth (such as it is) does not warrant the extra expense, weight and wind resistance that is incurred.

Instead of using linear (straight) elements, as in the generic Yagi, this design uses loops. Designs using squares of wire instead of loops are known as *Quad* aerials, and HF designs require large X-shape frames to support the large squares of wire. At 70 cm, however, the use of wire loops is easier, and they are self-supporting.

Construction

This is quite simple. Any type of material (metal, plastic, wood) can be used for the *boom* (the support for the elements) and for the mast. The elements are made from 14 SWG enamelled copper wire. 16 SWG hard-drawn aerial wire, which is not enamelled, has also been used with success. Thinner wire might result in a rather 'floppy' aerial! The separate parts of the aerial are held together with *jubilee clips* (hose clips).

The driven element is secured to the boom with a jubilee clip and a three-connector plastic connector block as shown in detail in **Figure 1**, and in the photograph. Cut the wire for the driven element 70 mm longer than the 700 mm indicated in Figure 1. Then, using sandpaper, remove the enamel from one end to a distance of 20 mm, and from the other to a distance of 50 mm. After forming the loop of the driven element, bend both stripped ends through 90°, and insert them into the first to holes of the plastic connector block (Figure 1). Do not tighten the screws yet. Push both ends into the block as far as they will go, then bend the 50 mm end back on itself and pull the ends back through the connector block so that the end you have just bent goes into the third hole in the block. Now tighten the screws in the block and in the jubilee clip.

Each director and the reflector should be made 40 mm longer than the circumferences shown in Figure 1. Strip the enamel, as before, from the last 20 mm at each end. Form the wire into the loop, slip the ends under the jubilee clip (Figure 1) and tighten it. You may find that it helps to solder the stripped ends together before securing the jubilee clip.

The boom is fixed to the mast using jubilee clips and wire, as shown in Figure 1. Solder the feeder cable to the driven element, with the braid soldered to the end of the driven element which is connected to the boom (this applies to metal booms; with plastic or wooden designs, the feeder connections are not critical).

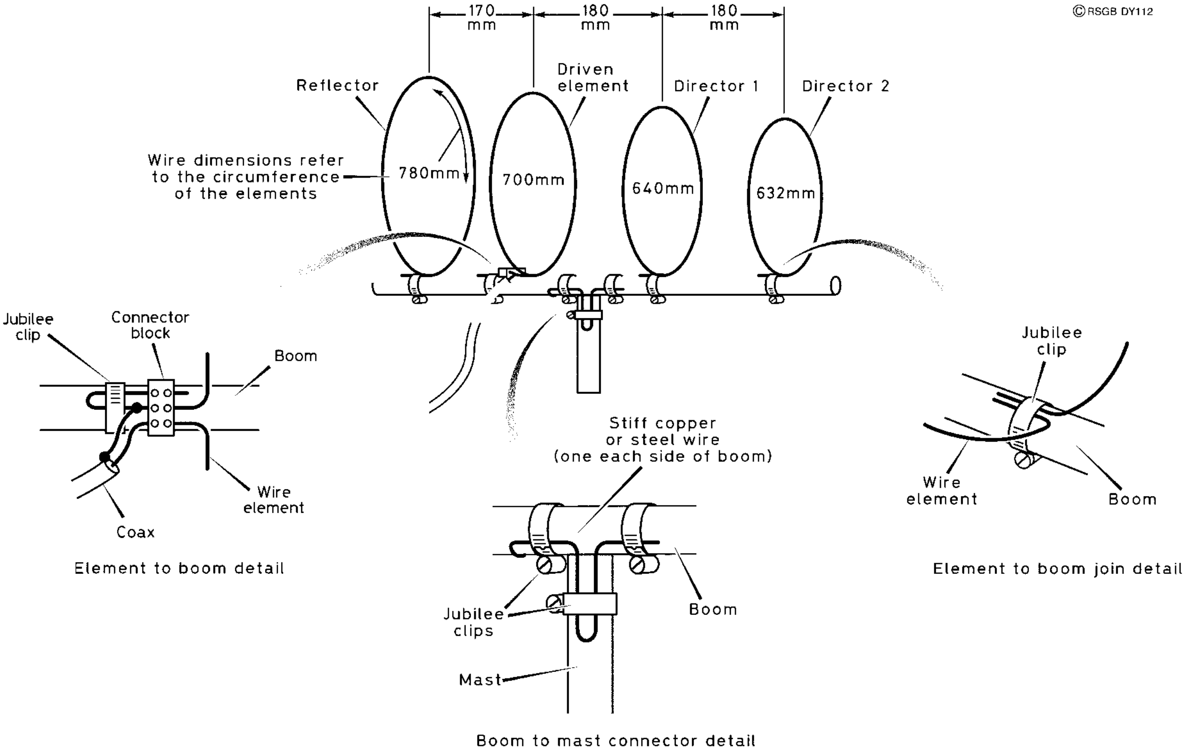
©RSGB DY112
170 mm
180 mm
180 mm
Reflector
Driven element
Director 1
Director 2
Wire dimensions refer to the circumference of the elements
780mm
700mm
640mm
632mm
Jubilee clip
Connector block
Boom
Wire element
Coax
Element to boom detail
Stiff copper or steel wire (one each side of boom)
Jubilee clips
Boom
Mast
Boom to mast connector detail
Jubilee clip
Wire element
Boom
Element to boom join detail

Figure 1 70 cm 4 element quad construction, with detail of how hose clamps are used in the construction

You will need a plug on the shack end of your feeder to suit the transceiver, aerial tuning unit (ATU) or the standing-wave meter (SWM) you are using. Always connect these in the order: transceiver – SWM – ATU – aerial.

Testing

Always test aerials outside and away from buildings (if possible!). This avoids getting misleading results.

Use a rubber duck, or whatever aerial you usually use, and tune around to find a repeater or beacon signal which is consistent. Note the reading on the S-meter. Then, connect your new Quad loop. Rotate it to give the strongest S-meter reading (don't forget it is directional). Verify that the meter reading varies as you rotate it. How does the S-meter reading compare with the original reading? Much depends on the siting of your original aerial; if you are comparing your Quad loop at ground level with a vertical on the chimney, you wouldn't expect your new aerial, even with its gain, to outperform a vertical which is well elevated!

Now a test on transmit is called for. The use of an SWR meter is essential here. Find a clear frequency and check that it *really is clear* before announcing your presence and carrying out the test. A reading of 1:1 is excellent, but any value less than about 1.8:1 would be acceptable. You can measure the directivity of your aerial using the field strength meter, also described in this series.

Materials

4 metres of 14 SWG enamelled copper wire
Material for boom and mast
15 amp connector block
7 Jubilee clips

The enamelled copper wire is available from AA&A Ltd, Sycamore House, Northwood, Wem, Shropshire SY4 5NN. Everything else is available from most hardware stores.

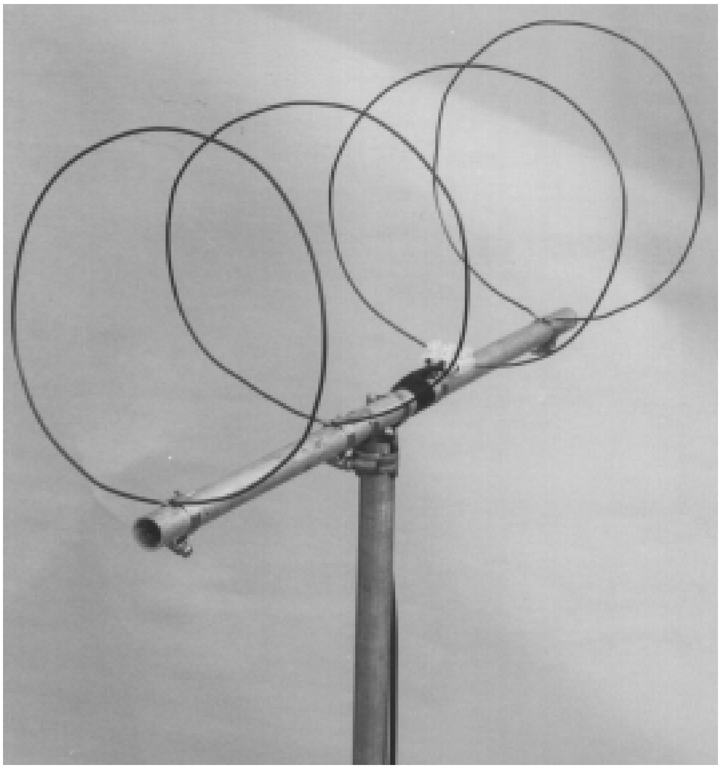

46 A UHF field strength meter

Introduction

It is always interesting, and often useful, to know just *where* the radiation is going from your aerial. How much of your radiated energy is going in the general direction of the station you are in contact with, and how much is being effectively wasted? Some of these questions can be answered with the use of a *field strength meter.* A field strength meter is simply a receiver, stripped down to its bare essentials, such that it responds only to the magnitude of the carrier. The use of a field strength meter assumes that the aerial under test is radiating a continuous carrier. Don't forget to find a clear frequency and identify your transmissions at least every quarter-hour, in order to comply with the terms of your licence.

Description

Two types of field strength meter are shown in **Figure 1**. You will recognise both circuits (Figure 1b particularly) as being types of 'crystal set' with a meter replacing the headphones. Figure 1a is a broad-band HF design (there is no tuning provision) and Figure 1b is tuned in the same way as the crystal set; with a loop of wire as an aerial, it will perform well in the VHF/UHF range.

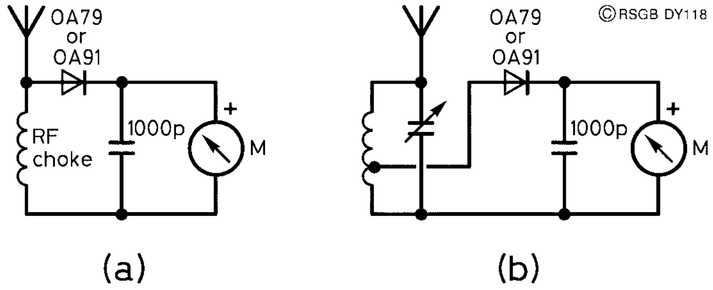

Figure 1 Construction of the UHF field strength meter

A tuned field strength meter also doubles as an *absorption wavemeter* if it is well calibrated. Such devices are useful for detecting transmitter harmonics also. This design is intended for use in the UHF band, so it will have to be sensitive around 432 MHz.

Construction

The field strength meter comprises a loop of wire, 600 mm long, which acts both as aerial and as the tuning inductance, a diode, a capacitor, a connector block, a meter and a length of twin wire. All the components, with the exception of the meter, are fixed to a pole with a jubilee clip, as shown in **Figure 2**. The meter should have a sensitivity of between 50 μA and 100 μA, or a multimeter can be used. The multimeter is more flexible, as you can select different current ranges, giving you a range of sensitivities.

Using the field strength meter

Connect your meter to the ends of the twin wire from the pole. Place a hand-held transceiver about 2 metres away, and press Transmit. If there is no meter reading switch off the transmitter and check the wiring. If the needle attempts to go negative, simply reverse the wires to the meter. If the reading is too high, either move the transmitter further away, or increase the current range on the multimeter. Try changing the orientation of the transmitter aerial, and note how the signal varies.

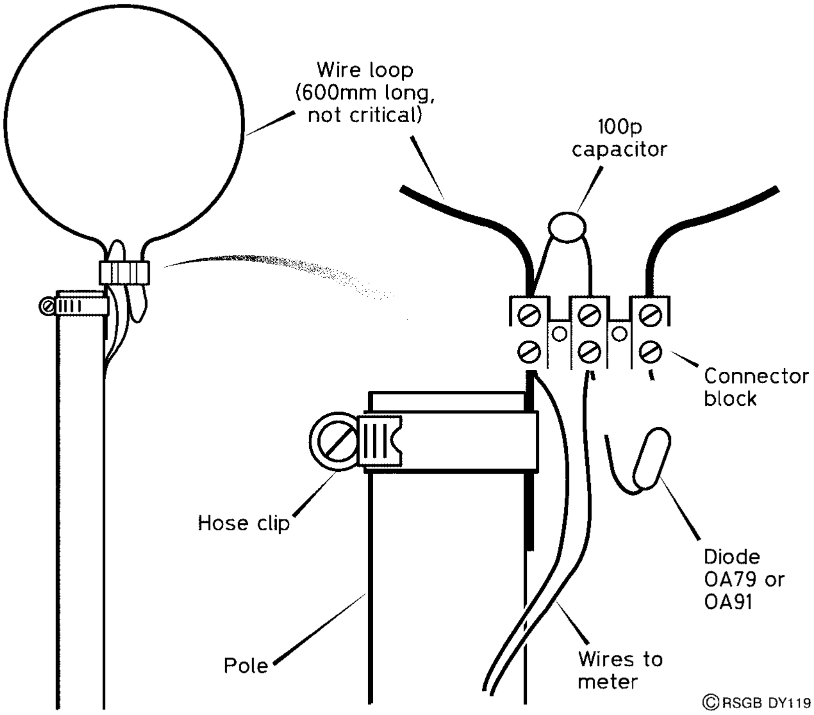
Wire loop
(600mm long,
not critical)
100p
capacitor
Connector
block
Hose clip
Diode
OA79 or
OA91
Pole
Wires to
meter
©RSGB DY119

Figure 2 Assembly of the field strength meter

To measure the *polar diagram* (a graph of the signal strength against the angle between the aerial boom and the field strength meter, plotted on polar axes) of a beam aerial, mount the loop as far away from the aerial as you can, preferably at the same height, and rotate the aerial, say 15 degrees at a time, and record the signal strength, until the aerial has been turned through 360 degrees. If you need help plotting the graph, enlist the help of a friend who has done it before, or one who knows about polar graphs! If you have already built the Photometer project, you will have measured and plotted the polar diagram of the light intensity from a torch. Now you will see the great similarity!

Parts list

Capacitor	100 picofarads (pF)
Diode	Germanium, OA79 or OA91
Connector block	10 A, 3-way
Wire	600 mm of 16 SWG enamelled copper
	Length of twin cable for meter connection
Clip	Jubilee (hose) clip

47 Christmas tree LEDs

Introduction

A novelty ideal for the festive season, this circuit causes one LED at a time to light up around a small cardboard Christmas tree.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which each IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**
3. **Repeat the process with the second IC.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there are no supply decoupling capacitors across each IC.

Description

Figure 1 shows the layout, and how the wires are connected from the circuit board to the LEDs around the tree. The circuit, shown in **Figure 2**, is quite complicated, so you need to be confident in your logical approach to circuit-building before you attempt this one! It uses two common integrated circuits (ICs). IC1 is simply an oscillator which provides timing pulses for IC2, which 'counts' up to a maximum of 10. The outputs of the counter are indicated by light-emitting diodes (LEDs); red, orange, yellow and green are common colours which you can use.

The circuit is built on a single piece of Veroboard measuring 29 holes by 12 strips. **Be aware that there is no strip labelled 'I', so don't make mistakes in your counting!**

First of all, cut the tracks in the positions shown in Figure 1 using a 3 mm ($\frac{1}{8}$ inch) twist drill rotated between thumb and forefinger. Then, solder in the IC sockets, the notches facing row M. Wire up and solder in the links and then the Veropins for the connections to the LEDs and battery. Having done this, the resistors and the capacitor should be fitted. When wiring the LEDs, each cathode (the lead adjacent to the 'flat' on the LED encapsulation) is connected to R4, and the anodes go to separate pins on IC2.

Testing

Hold up your circuit board to the light and check carefully for solder bridges between adjacent tracks. Then check again that the wiring is correct. Place the ICs in their holders, with the notched ends lining up with those on the holders. Connect the battery, and the LEDs should illuminate in sequence.

If you have no success, you are now wishing you had checked the circuit more carefully! Learn something from your mistake and it will not have

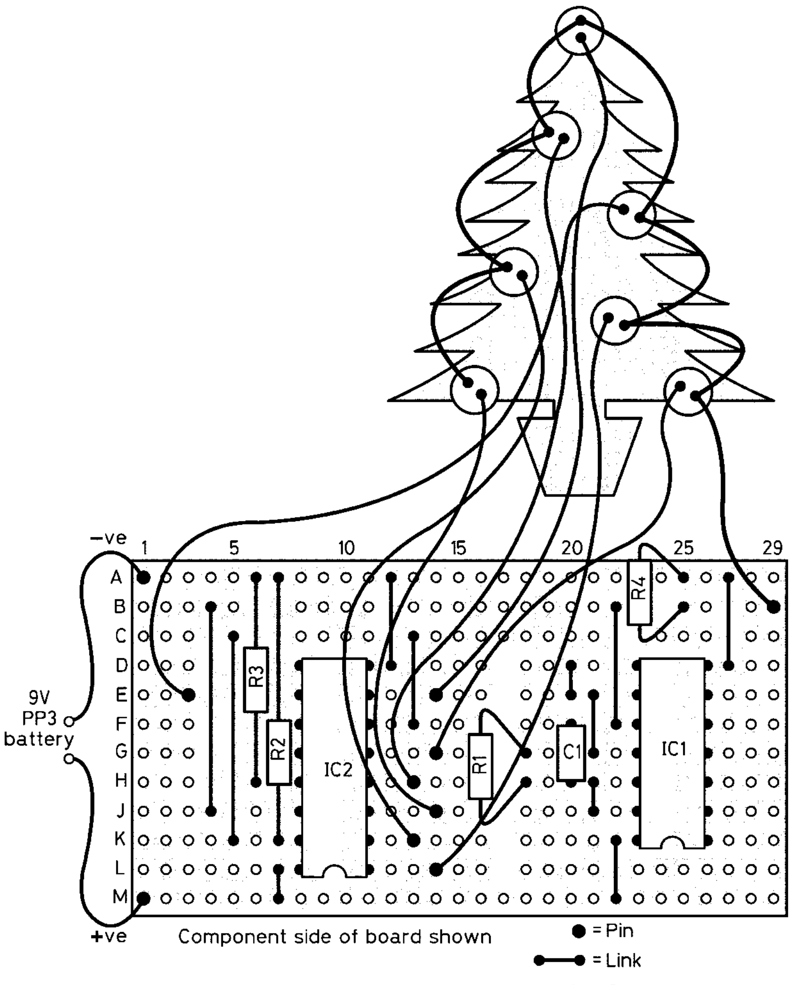
−ve
1
5
10
15
20
25
29
A
B
C
D
E
F
G
H
J
K
L
M
9V
PP3
battery
+ve
R3
R2
IC2
R1
C1
IC1
R4
Component side of board shown
= Pin
= Link
= Cut in copper track

©RSGB DY137

Figure 1 Christmas tree, component layout

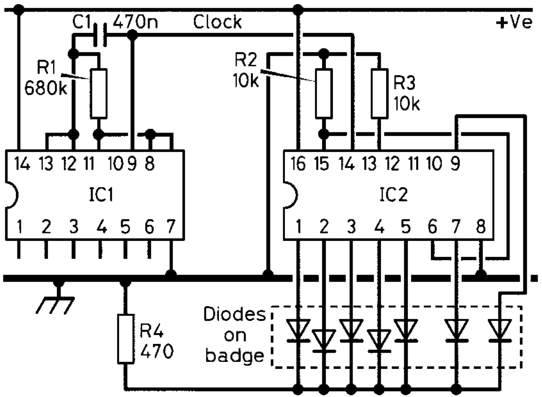
C1
470n
Clock
+Ve
R1
680k
R2
10k
R3
10k
14 13 12 11 10 9 8
IC1
1 2 3 4 5 6 7
16 15 14 13 12 11 10 9
IC2
1 2 3 4 5 6 7 8
R4
470
Diodes
on
badge
D6 D2 D1 D3 D7 D4 D5

©RSGB DY138

Figure 2 Christmas tree, circuit diagram

been in vain! The first thing to check is that the LEDs are the correct way round. If that is OK, then check the positions where the tracks are broken (intentionally!). After that, are the wire links in the right places – had you forgotten there is no row ‘I’? Check *all* the wiring, then check *again* for solder bridges, and switch on again. One of these tests should have revealed a fault. If it *still* refuses to work, perhaps one (or both) of your ICs were damaged by static electricity, despite your precautions – or did you choose to ignore them?

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	680 kilohms (kΩ)
R2, R3	10 kilohms (kΩ)
R4	470 ohms (Ω)
Capacitor	
C1	0.47 microfarad (μF) min. metallised polyester film
Semiconductors	
IC1	4011
IC2	4017
Additional Items	
D1–D7	5 mm LEDs in choice of colours
	14-pin DIL socket for IC1
	16-pin DIL socket for IC2
	PP3 battery clip and battery
	Veroboard, Veropins
	Insulated wire for links

48 An audio signal injector

Introduction

An audio signal injector is a device used to test audio frequency circuits. It is simply an oscillator running at a frequency in the audio range, so that when its output is fed to the input of an amplifier, it will produce a sound in the loudspeaker if the amplifier is working. The oscillation is so rich in harmonics that the signal can also be heard (sounding rather different) when injected into an RF circuit.

The design

The circuit is shown in **Figure 1**, and is a basic *astable multivibrator*, a free-running oscillator producing a roughly rectangular-wave output. The two transistors, TR1 and TR2, operating as switches, switch on and off alternately at a frequency around 500 Hz. The prototype was constructed on plain matrix board (no copper strips), as illustrated in **Figure 2**.

Both transistors are type BC108, which are only a few pence each new, and can be found at almost any rally. You can add an on/off switch, or simply disconnect the battery when you are finished using it. To make the unit in one piece, the battery can be taped to the board, as the diagram shows.

The probe itself is made from a short piece of stiff insulated wire, soldered to a tag on the board; an earth lead is also soldered to the board, and terminated in a crocodile clip to attach to the ground lead of the equipment under test.

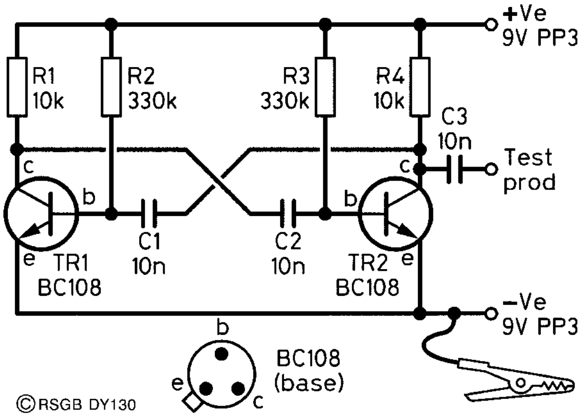

Figure 1 Circuit diagram of the signal injector

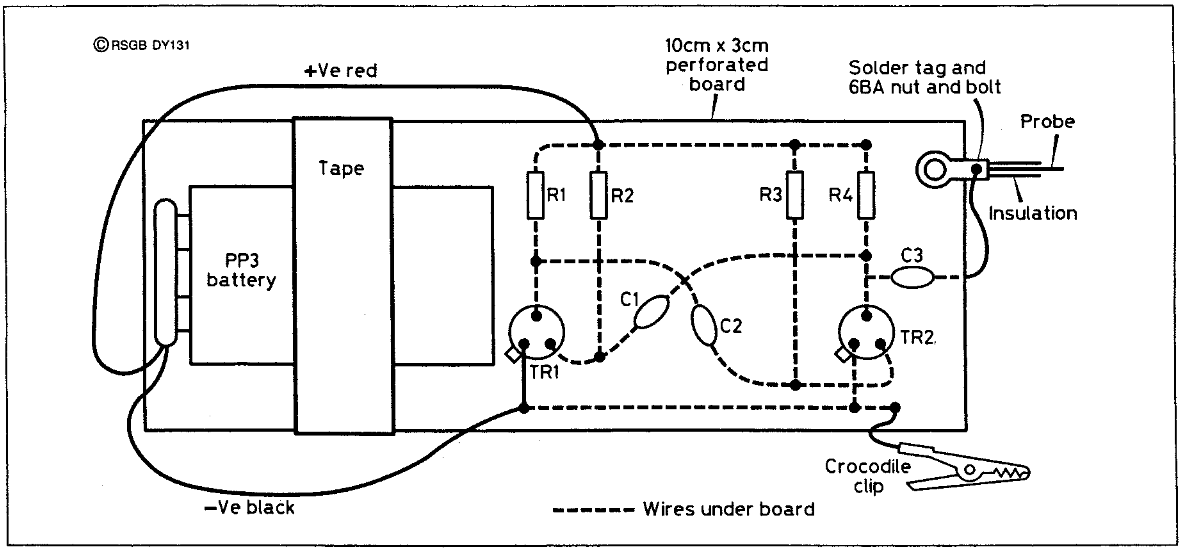
©RSGB DY131
+Ve red
10cm x 3cm perforated board
Solder tag and 6BA nut and bolt
Probe
Tape
PP3 battery
R1
R2
R3
R4
Insulation
C3
C1
C2
TR1
TR2
Crocodile clip
−Ve black
------ Wires under board

Figure 2 Component layout of the signal injector

Using it

After checking the wiring of the circuit, connect the battery. Find a transistor radio that is known to work. Carefully remove the case, switch on and apply the probe to the centre tag of the volume control. A *very* loud buzz should be heard from the speaker, indicating that the audio circuits of the radio are working. Using the injector to fault-find equipment you have made yourself is rather more instructive and rewarding, because you *know* where to inject the signal, and you should *know* what to expect when you do.

Warning

Do *not* work on any equipment connected to the AC mains. Work only on battery-powered circuits, for your own safety.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1, R4	10 kilohms (kΩ)
R2, R3	330 kilohms (kΩ)

Capacitors

C1, C2, C3	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic

Semiconductors
TR1, TR2 BC108

Additional items
- PP3 battery and connector
- Matrix board 10 cm by 3 cm
- 6BA solder tags
- Thick insulated wire for probe
- Crocodile clip

49 Standing waves

Introduction

Everyone has, or *should* have, a standing-wave ratio (SWR) meter as part of his/her array of test gear. Most people know how to use it, but what does it really do?

Before attempting to answer that question, we need to look at some aerial fundamentals. An aerial is a *transducer*, the word meaning ‘to lead across’. We use it whenever one form of energy is converted into another form. A bulb is a transducer; it converts electrical energy in the filament to radiated heat and light energy. **Figure 1** shows the situation.

An aerial, or antenna, is also a transducer; it converts radio-frequency (RF) energy in the feeder (or *transmission line*, to give it its proper name) into radiated electromagnetic energy, in the manner shown in **Figure 2**. The aerial has resistance, just like the bulb filament, and if the filament had zero resistance, there would be no radiated energy. In the same way, if the aerial had no resistance, there would be no radiation from it.

Transmission lines should convey RF energy from transmitter to aerial with the minimum power loss. A common form of transmission line is the *coaxial*

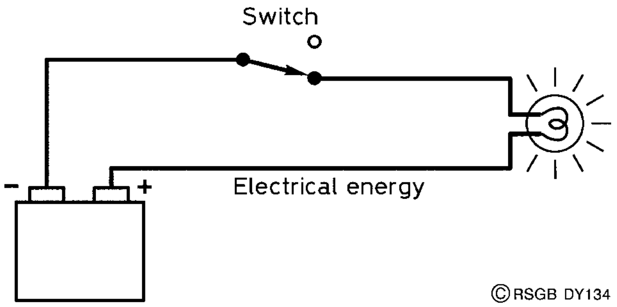

Figure 1 The light bulb converts electrical energy to electromagnetic light energy

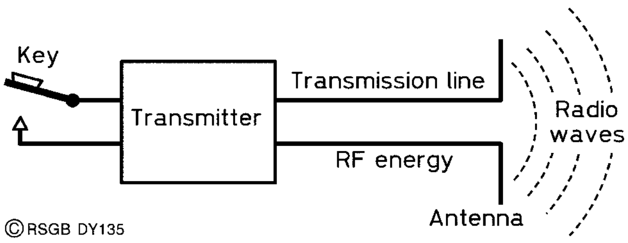

Figure 2 The antenna converts RF current to electromagnetic radio waves

cable. As you might expect, any feeder cable has a DC resistance caused by the resistance of the copper wire from which it is made. It also has an AC resistance, caused by the capacitance between the centre conductor and the braid, and by the inductance of the cable itself. This means that the feeder has an *impedance*, which is constant for the particular type of cable. This is what we call the *characteristic impedance* and, for the cables used in most amateur radio applications, it is 50 Ω.

If this impedance can be made the same as that of the aerial (it is already the same as the impedance of the transmitter output), then the transfer of energy will occur with minimum loss. If the aerial and cable impedances are *not* the same, then there is a *mismatch*, which causes some of the RF energy to be reflected back towards the transmitter. We now have a situation where RF energy is flowing along the cable from the transmitter to the aerial (the *forward* wave) and, at the same time, flowing from the aerial to the transmitter (the *reflected* wave). The two waves interact along the cable and form a stationary pattern of voltage and current. The pattern is known as a *standing wave*, and can be visualised from the waveforms in **Figure 3**. The ratio of the maximum voltage to the minimum voltage on a given wave defines the *voltage standing-wave ratio* (VSWR), or just *standing-wave ratio* (SWR) for short.

The SWR meter is easy to use. It is positioned in the feeder between the transmitter and the aerial tuning unit (ATU) if there is one, or between the transmitter at the aerial otherwise. Most meters have a single meter and a four-position switch. The transmitter is first keyed and, with the switch in

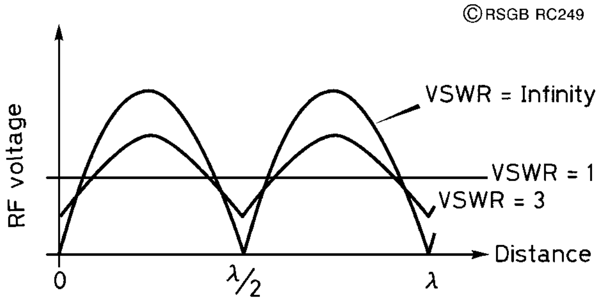

Figure 3 Standing waves on a transmission line

the 'calibrate' position, the sensitivity control is adjusted to give full-scale deflection on the meter. The switch is then changed to 'forward' to read the forward power, to 'reflected' to read the reflected power, and to 'SWR' to read the value of the standing-wave ratio. When there is no reflection (see Figure 3), the meter should read 1:1 or, simply, 1.

Most SWR meters remain in the feeder line while the transmitter is operating, so the condition of the aerial and feeder can be constantly monitored. Problems with the aerial (such as water entering the feeder at its junction with the driven element) are immediately shown up. Without the use of the SWR meter, the situation would slowly deteriorate over several months and you would be left wondering why so few stations were answering your calls!

50 A standing-wave indicator for HF

Introduction

The standing-wave ratio (SWR) meter shows how well the aerial system, including the feeder, is *matched* to the output of the transmitter. This design does not *measure* SWR, but it gives an indication of when the SWR is minimum for a given system of aerial and feeder. The design is usable on the HF bands from 1.8 to 28 MHz, and can be used at 50 MHz with reduced sensitivity.

How it works

There are two types of wave in any feeder: the *forward* wave, which travels from the transmitter to the aerial; the *reflected* wave, which travels back to the transmitter from the aerial. The presence of a reflected wave is evidence that some of your transmitted power is not being radiated, but is being returned to the transmitter to be lost as excess heat. When aerial and feeder are perfectly matched, there is *no* reflected wave, and all the power from the transmitter is radiated.

Referring to the circuit of **Figure 1**, a tiny fraction of the signal is removed by the transformer, T1, and by the capacitors, VC1 and C1. It is then detected by the germanium diodes, D1 and D2, and any residual RF removed by the capacitors, C2 and C3. The currents through the diode and meter (depending on the position of switch, S1) represent the forward and reflected signals. VR1 acts as a sensitivity control for the meter.

It pays to shop around for a suitable meter. Surplus types from tape recorders and hi-fi equipment are usually ideal for this purpose. A new one would cost several pounds. The more sensitive the meter, the more sensitive your indicator will be. Meter sensitivity is given by the current required to give full-scale deflection (FSD) of the pointer. One with an FSD of between 50 and 200 micro-amps (μA) is suitable for this circuit. The higher the FSD, the less sensitive the circuit.

Construction

The meter circuit and the sampling transformer (see **Figure 2**) are built and mounted on Veroboard of the copper-strip variety. It simplifies construction

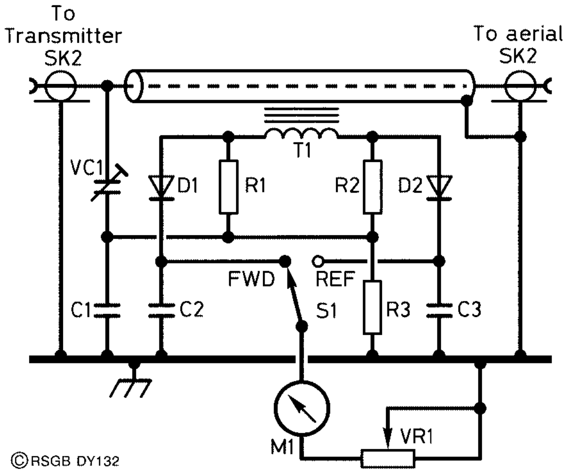

Figure 1 Circuit diagram of the SWR meter

©RSGB DY132

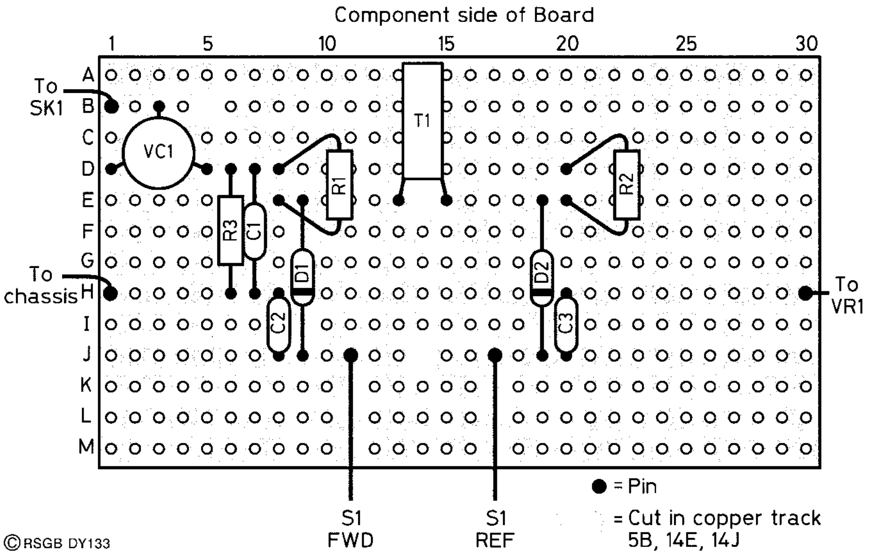

Figure 2 Component layout of the SWR meter

but reduces the operational range of the meter to below 30 MHz because of the capacitive coupling between strips. The board has 13 strips by 30 holes, although you can reduce this if you have a smaller case.

Firstly, cut the tracks at the three points shown. Then insert and solder Veropins for connections to the external components, the switch, variable resistor and the meter. Solder in the components starting with the resistors and followed by the capacitors and the diodes, ensuring that the diodes are inserted correctly.

Now you have to wind the transformer, T1, on a small toroidal ferrite core. Wind the secondary with 15 turns of 36 SWG enamelled copper wire, spaced evenly over about two-thirds of the former. The turns should not overlap, and considerable care must be taken; the wire is very thin, will kink easily and will break if you apply too much tension. The ‘primary’ is an 8 cm length of 50 Ω coaxial cable which passes through the toroid on its way between the input and output connectors. The braid of the cable is connected to the case at only *one* of the connectors (see Figure 1); this prevents the screen and the metal case between the two sockets forming a single, shorted turn.

The ends of the secondary winding must be carefully stripped of their enamel with sandpaper, before attaching the toroid to the board with cotton or nylon fishing line. On no account must wire be used for this!

Solder the secondary connections of T1 to the board and thread through the coaxial cable ready for soldering to the connectors.

The case used was an aluminium box (Maplin LF02C), but any suitable metal box could be used. Aluminium is preferable, as it is easily drilled with simple tools. Use standoff insulators to mount the board in the case. Once this has been done, the leads from the board to the chassis-mounted components can be soldered. So can the coaxial cable passing through the toroid. Make the lead from the input socket to VC1 as short as possible.

Setting up

You will need a 50 Ω dummy load and a transmitter to set up your indicator. Connect the transmitter to SK1 and the dummy load to SK2. Set the toggle switch, S1, to *forward* and the sensitivity control, VR1, to mid-travel. Switch on the transmitter, and set VR1 for maximum meter deflection. Switch to *reflected* and adjust VC1 until the reading is minimum (ideally zero). This completes the setting up!

Using the indicator

For setting up an aerial, connect your circuit between the transmitter and the cable leading to the aerial. With S1 in the *forward* position, key the transmitter and adjust VR1 for maximum reading on the meter. Switch to *reflected*, and then adjust your ATU to give minimum reflected power. If your adjustments are to be made to the aerial itself, to give minimum reflected power, you must make a note of the reflected reading, switch off the transmitter, make a change to the aerial, key the transmitter, and note whether the reflected power is greater or less than before. Then, make more changes to the aerial. **Never adjust your aerial with the transmitter on**. Make your adjustments on an unused frequency, and do it as quickly as possible, thus avoiding (or minimising) interference to other stations.

Parts list

Resistors: all 0.25 watt, carbon 5% tolerance
(or Maplin 0.6 watt metal film)

R1, R2	27 ohms (Ω)
R3	2.2 kilohms (kΩ)
VR1	10 kilohms (kΩ) linear

Capacitors

C1	220 picofarad (pF) disc ceramic 50 VDC
C2, C3	0.1 microfarad (μF) disc ceramic 50 VDC
VC1	20 picofarad (pF) trimmer

Semiconductors

D1, D2	OA91 germanium

Additional items

S1	Single-pole changeover (SPDT or SPCO)
SK1, SK2	Coaxial sockets to suit station standards
	Veroboard – 13 strips by 30 holes
	Veropins (7 off)
	Amidon FT 50–43 ferrite toroid
	Meter, less than 200 μA FSD
	36 SWG enamelled copper wire
	Short length of UR43 or RG58 coaxial cable
	Insulated stranded wire
	Aluminium box
	Standoff insulators for mounting the board
	Knob for the sensitivity control

51 A moisture meter

Introduction

Dry rot (*Merulius Lacrymans*) can strike havoc in buildings, causing the timbers to decay and crumble to dust – hence the term *dry rot*. Wood is attacked only if its moisture content rises above 20%.

Construction

The circuit of the moisture meter is shown in **Figure 1**. The two probes touch the wood, and the current that flows between them depends on the moisture content of the wood. If the moisture is sufficiently high, the current, after amplification, will be enough to light the LED.

The meter can be made on a piece of plain matrix board (no copper strips), as **Figure 2** shows. The board is big enough (10 cm by 2.5 cm) to accommodate the PP3 battery, taped on. No case is needed, unless you want

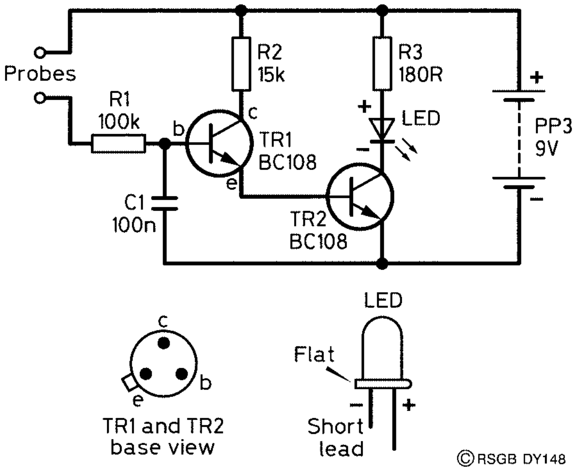
Probes
R1
100k
b
c
R2
15k
TR1
BC108
e
C1
100n
R3
180R
+
LED
–
TR2
BC108
+
PP3
9V
–
LED
Flat
– +
Short lead
c
e
b
TR1 and TR2 base view
©RSGB DY148

Figure 1 Circuit diagram of a moisture meter

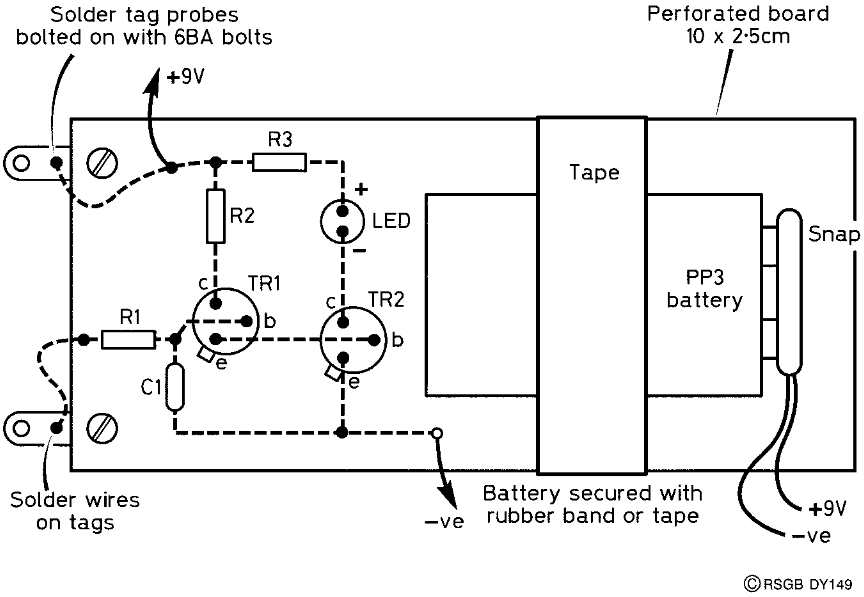
Solder tag probes bolted on with 6BA bolts
+9V
Perforated board 10 x 2·5cm
R3
Tape
+
R2
LED
–
Snap
TR1
c
TR2
c
R1
b
b
PP3 battery
e
e
C1
Solder wires on tags
–ve
Battery secured with rubber band or tape
+9V
–ve
©RSGB DY149

Figure 2 Moisture meter, component layout

to leave the meter in a damp location for a protracted period! Make sure that the transistors and LED are mounted correctly. In Figure 2, the connections as shown to TR1 and TR2 are illustrated as if the transistors were transparent. An on/off switch is not really necessary, as only a very small current flows when the probes do not touch anything. Use solder tags, screwed to the board, to act as probes.

After the assembly is completed, check your circuit one final time, and then connect the battery. Nothing should happen at first. If you lick your forefinger and hold it across the probes, the LED should light.

Using it

The prototype was compared with a commercial moisture meter, and the LED lit when the moisture was around 20%. This was quite fortuitous, as the point at which the LED lights depends both on the separation of the probes and on the gain of the two transistors.

In addition to searching for dry rot, the instrument may be used to monitor the moisture in the soil of household plants. In this case, probes made of 16 SWG copper (*not* enamelled) should be soldered on to the two tags, and should penetrate the soil to a depth of several centimetres, and R1 may require adjusting so that the LED extinguishes if the soil is too dry, and lights if the soil is sufficiently moist. If you wanted to leave the meter with the probes in the soil, an on/off switch *would* be necessary.

Parts list

Resistors: all 0.25 watt, 5% tolerance	
R1	100 kilohms (kΩ)
R2	15 kilohms (kΩ)
R3	180 ohms (Ω)
Capacitor	
C1	0.1 microfarad (μF) polyester
Semiconductors	
TR1, TR2	BC108
LED	Any shape or colour will do

Additional items
- PP3 battery and connector
- Solder tags (2 off) for probes
- Matrix board 10 cm by 2.5 cm

52 Simple aerials

Introduction

The performance of *any* receiver or transceiver, no matter how expensive it is, is limited by the aerial that feeds it. Two of the most frequently asked questions are:

- Which is the best sort of aerial to use?
- Where is the best place to locate an amateur radio aerial?

To answer these questions, you must ask yourself what sort of operation you want to do. Are you interested in local, chatty contacts on the lower bands or VHF, or are you more disposed towards long-distance (DX) contacts, and on what band?

A house with a moderately sized garden is assumed in the diagrams here, to illustrate the configurations of some simple aerials. You would not need all these aerials festooned around your house, because one or two would be sufficient for your needs. The problems incurred by properties with more restricted space will be covered later.

VHF aerials

For VHF operation, the aerial should be mounted as high as possible, either on a mast or on a chimney. For all-round coverage on FM and the local repeaters, a vertical colinear is a good choice. For SSB and CW DX operation, a horizontal rotatable beam is needed. If satellite working is envisaged, you will need to contemplate mounting an elevator on top of your rotator, so that your beam can point in *any* direction, including vertically upwards! An advantage of satellite working is that the aerials do not *necessarily* have to be up in the air, provided you have a relatively uncluttered site. Your rotator and elevator can be at ground level, which is good!

If the VHF aerial is mounted on the chimney, use a double mounting bracket, particularly if you have a beam and rotator. Keep the TV, broadcast FM and amateur aerials as far apart as possible, and keeping the feeders separated is also a good plan.

The dipole aerial

One of the simplest types of aerial for single-band operation is the half-wave dipole. (The name ‘dipole’ simply means ‘two poles’ or ‘two elements’, and

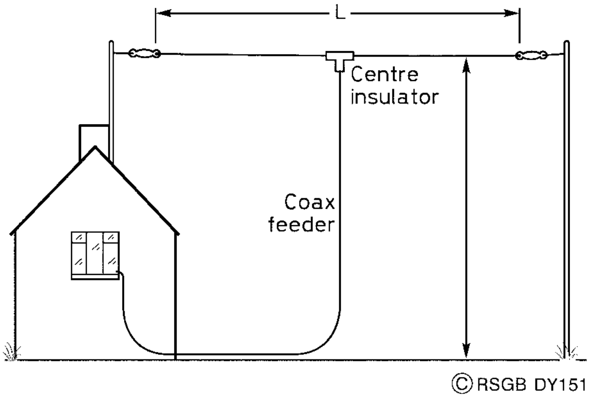

Figure 1 Layout for a dipole aerial

in this case the total length of the dipole is approximately half a wavelength at the operating frequency.) It is usually fed in the centre by coaxial cable as shown in **Figure 1**. The length of the dipole for the lowest frequency in each band is shown in **Table 1**. Normally, the length of the aerial will be 'trimmed' to be tuned to the centre frequency of the part of the band in which you will operate. This is done using the data in the right-hand column of Table 1. As an example, suppose you wanted your aerial to be resonant at 3.7 MHz. The table gives an overall dipole length of 42.86 m for 3.5 MHz. To resonate the aerial 200 kHz *higher*, then this length must be shortened by 2 × 0.595 m = 1.190 m. Your dipole would thus be 41.67 m long. Remember to allow extra wire for fixing the dipole ends to the insulators.

Table 1 Dipole lengths for lowest frequency of each band and the length to be trimmed from each to raise the resonant frequency by 100 kHz

Band (MHz)	*Dipole length (m)*	*Trim each end (mm/10 kHz)*
1.8	83.33	2190
3.5	42.86	595
7	21.43	150
10	14.85	70
14	10.71	35
18	8.33	20
21	7.14	15
24	6.03	12
28	5.36	10
50	3.00	6

On the lower-frequency bands, the lengths become rather large. In this case, you can ‘bend’ your dipole, as illustrated in **Figure 2**. The length of wire required to give an acceptable value of SWR (less than 2:1 on transmit) may need to be different from the calculated value, so be prepared to experiment!

Dipoles are single-band aerials, although they will often work acceptably on the third harmonic of their design frequency: a 7 MHz dipole often operates reasonably well on 21 MHz. It is possible to operate several dipoles in parallel, as **Figure 3** shows. Interaction between the elements can occur if the spacing between them is less than about 10 cm. A multi-band dipole, as shown in Figure 3, has the elements separated with plastic spacers, and drooping ends to produce maximum spacing between the elements’ ends.

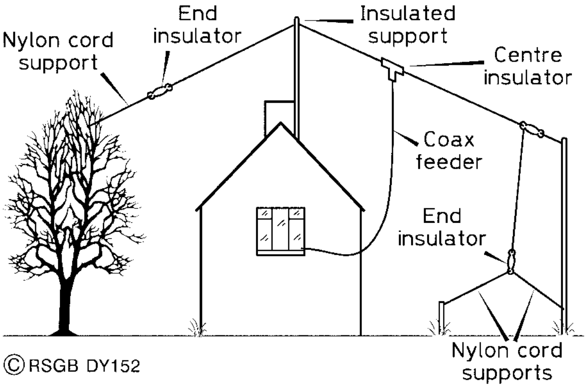
End insulator
Insulated support
Nylon cord support
Centre insulator
Coax feeder
End insulator
Nylon cord supports
©RSGB DY152

Figure 2 Possible layout for a dipole aerial in a confined space

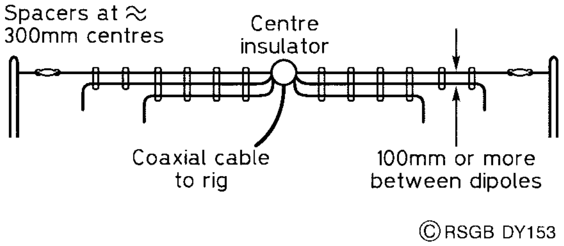
Spacers at ≈ 300mm centres
Centre insulator
Coaxial cable to rig
100mm or more between dipoles
©RSGB DY153

Figure 3 Multi-band dipole aerial

The long-wire aerial

This aerial is simple, cheap, easy to erect, and suits most houses and gardens, as **Figure 4** shows. Using an aerial tuning unit (ATU), an end-fed long wire can function on several bands when used with a set of radials or a counterpoise. **Figure 5** illustrates the setup. The length of the aerial will determine the bands which will be covered.

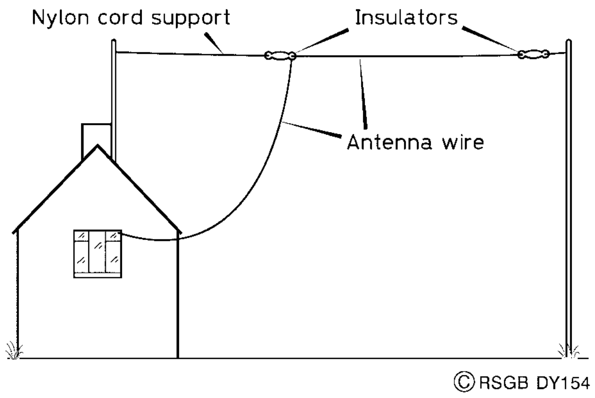
Nylon cord support
Insulators
Antenna wire
©RSGB DY154

Figure 4 Long-wire or inverted-L aerial

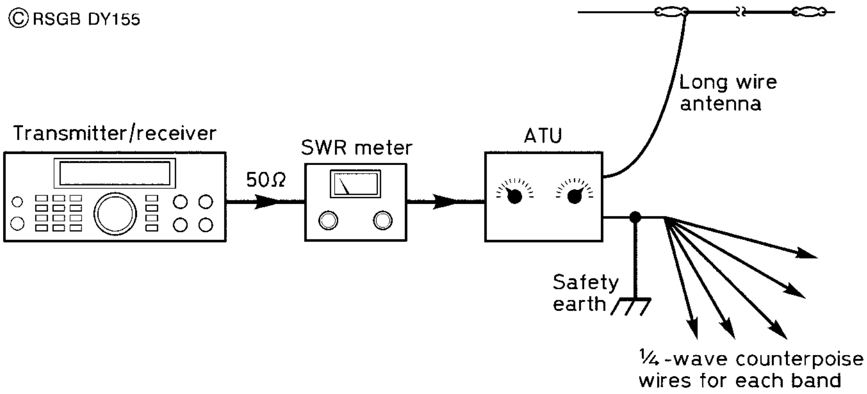
©RSGB DY155
Long wire antenna
Transmitter/receiver
50Ω
SWR meter
ATU
Safety earth
¼-wave counterpoise wires for each band

Figure 5 How to connect a radio to a long-wire aerial

A wire length of 10.5 m will work on the 40, 30, 17, 15 and 12 m bands.

A wire length of 15.5 m will work (with an ATU) on the 80, 40, 20 and 12 m bands and possibly (depending on your ATU) on the 17 and 15 m bands.

A wire length of 26.5 m will operate on all bands, but may be difficult to load on 10 m.

The wire lengths given here may need some adjustment because of the geometry of your particular house and garden. For receive-only purposes, the lengths are far less critical.

In general, you cannot get a good radio-frequency (RF) earth from a first-floor (or higher) shack. Unless a good RF earth exists within a small fraction of a wavelength of the transceiver, an *artificial ground* comprising a single $\frac{\lambda}{4}$ radial or counterpoise will be needed. You will need one counterpoise for each band you intend to use, and the wire can be concealed around the skirting-board of the shack, or under the carpet. Make sure that the free end

Table 2 Lengths of elements for vertical antennas, radials for verticals and counter-poises for end-fed long wire antennas

Band (MHz)	*Element length (m)*
1.8	39.66
3.5	20.40
7	10.20
10	7.14
14	5.1
18	3.96
21	3.4
24	2.95
28	2.55

of each counterpoise is well insulated; this point can carry a *very high voltage* when you transmit; anyone coming into contact with this can suffer very severe RF burns. Counterpoise lengths can be read from **Table 2**.

The vertical aerial

The single-band vertical aerial is sometimes used by DX operators because it has a low angle of radiation, which favours long-distance propagation. However, it must be sited clear of obstructions and must have a *good* counterpoise or radial system. Illustrations of the vertical aerial are shown, and the lengths of the vertical and radial sections are given in Table 2. The centre of the coaxial feeder is connected to the vertical section, and the braid to the counterpoise or radial system, which is made up of four or more wires buried just below the surface and joined together near the base of the aerial.

Cable entry to the house

Bringing coaxial cable into the house by an open window must be regarded as a temporary measure. Wooden window frames can be drilled, one hole for each feeder. Make the holes slope downwards from inside to out to prevent rain entering, and treat these with wood preservative. Leads from long-wire and inverted-L aerials should be kept separate from other cables.

Alternatively, a plastic pipe large enough to take all your feeders could be fitted into the brickwork (again, sloping downwards towards the outside). You may want to let a friendly builder do this for you.

53 A breadboard 80 m CW transmitter

Introduction

In the early days of radio, many circuits were built on a wooden baseboard, the parts being screwed down on the board. This was called breadboard construction, because it was a breadboard that was frequently commandeered for the process! Wives have always been generous in this respect, it appears!

This circuit was originally designed by GM3OXX, and became known as the *Oner*, because it was built on a circuit board one inch square! The circuit appeared in the G QRP Club journal *Sprat* and, since that time, many hundreds of *Oner* circuits have been built and used on the air. It is a well-proven circuit.

The transmitter has no tuned circuits in the power amplifier (PA) and thus has a rather high *harmonic* content. It **must** be used with the ***low-pass filter*** described elsewhere in this book. Without the low-pass filter, interference will be caused to other stations.

Simple aerial changeover switching is provided, which allows this circuit to be used with any of the 80 m receivers, such as the *Colt*, described in this book. It can also be used with any kit or commercial receiver for the 80 metre band.

The circuit

The transmitter circuit is shown in **Figure 1**. TR1 is a crystal oscillator, the frequency of which is controlled by crystal X1. A small trimmer capacitor, TC1, is added to allow the frequency of X1 to be varied by a small amount. If adjustment of this trimmer is made possible from the front panel, it is useful to adjust the transmit frequency to avoid other stations already on the crystal frequency. The collector load resistor, R2, of the oscillator transistor, TR1, determines the power output; a value of 3.3 kΩ seems to work well in producing an output of 3 watts.

TR1 is directly coupled to TR3, a VMOS transistor (a type of field-effect transistor (FET)). This acts as the power amplifier (PA) stage. TR3 should

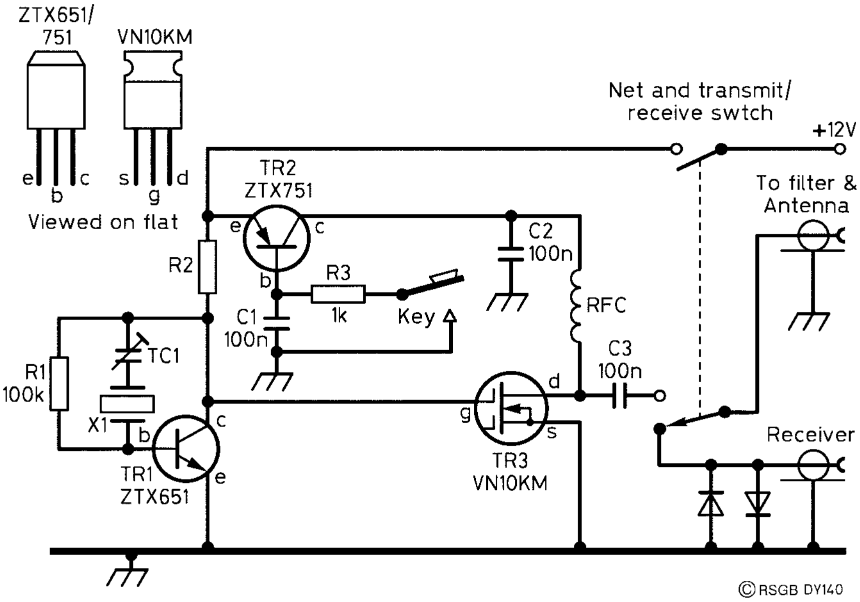
ZTX651/
751
VN10KM
e
b
c
s
g
d
Viewed on flat
Net and transmit/
receive swtch
+12V
TR2
ZTX751
e
c
b
C2
100n
To filter &
Antenna
R2
R3
1k
Key
RFC
C1
100n
TC1
R1
100k
X1
b
c
TR1
ZTX651
e
C3
100n
d
g
s
TR3
VN10KM
Receiver
©RSGB DY140

Figure 1 Circuit diagram of the breadboard transmitter

give about 3 W output, which is then coupled to the output by C3. The radio-frequency choke (RFC) providing the drain load of TR3 is simply a few turns of wire on a ferrite bead.

TR2 is an interesting addition to the circuit. It is used as a switch to 'key' the PA, TR3. The transmitter *could* be built without TR2, just placing the Morse key between the top of the RFC and the 12 volt supply. Adding TR2 is helpful, because it means that one side of the Morse key can be grounded (always a good thing), and some degree of *shaping* of the output RF waveform is provided by R3 and C1. This makes the transmission sound a little better and reduces the possibility of spurious frequencies being generated and transmitted. TR2 is a pnp transistor; note that it is the *emitter* of this transistor which is connected to the positive side of the supply.

Some form of changeover switching is needed for the aerial. A double-pole changeover toggle switch can be used. See the chapter on switches, later on in this book. One pole is used to switch the aerial between transmitter and receiver; the other pole is connected in the 12 volt supply line, and is labelled RECEIVE/TRANSMIT – NETTING. Its use will be described later. In this simple circuit, the PA cannot work when the key is open, because the key breaks its supply (via the RFC). When the key closes, TR2 switches on and applies the 12 volt supply to the top of the RFC. C2 is a *decoupling* capacitor, which prevents any residual RF signals at the top of the RFC reaching TR2.

Building

The prototype was put together as follows. Take a piece of plain printed-circuit board (PCB) measuring 5 cm by 4 cm. Then, with a *new*, *sharp blade* in a junior hacksaw, draw the blade horizontally across the surface of the copper in order to make a pattern of 6 squares along the 5 cm side and 5 squares along the 4 cm side. No more pressure should be applied than is necessary to cut through the copper! All the parts will be soldered on these pads in a form of surface-mount construction. To do this, each active pad (i.e. one that is going to have a component soldered to it) needs to be *tinned*. This means coating the pad's surface with solder, and is carried out as follows. Place the hot tip of the soldering iron on to the pad, and hold it there for a second or so. Then, with the tip still in place, touch the end of your reel of solder *on the pad*, not the tip of the iron. The solder should flow evenly all over the pad, and you can remove the iron. The solder should solidify in a rounded, shiny blob! This provides a good surface for making soldered joints.

To join component leads to the pads, cut each lead about 1 cm long, and then bend the last 2 mm at right angles to the rest of the lead. As you did before, tin the 2 mm length of each lead. Place the tinned portion on to the pad, and place the tip of your iron *on the pad*, close to the lead. The solder on the pad and on the lead will melt and run together; remove the iron and *hold the component still* until the solder solidifies. When the joint has cooled, give the lead a gentle tug to make sure you have a good joint (a good *mechanical* joint is usually a good *electrical* joint, too!). Each transistor straddles three pads, so the centre lead will need to be shorter than the other two. Take care here to get the lead lengths right – if you do, you will be surprised how much more firmly the transistor is held than if you just botched the lead lengths by bending them to fit! Make sure the connections to the transistors are correct.

Winding the RFC is quite simple. Seven (or more) turns of thin (32 SWG) enamelled copper wire are threaded through a small ferrite bead. This requires care, because the bead is small and the wire is thin. Trim the ends to within about 1 cm of the bead, remove the enamel carefully with sandpaper and tin the bare ends, prior to soldering the choke to the board.

After completion of the wiring, check the circuit against Figure 1. Bread-boarding a circuit like this has its advantages, but it can have disadvantages, too. One of these disadvantages is that it can make circuit checking difficult. For a simple circuit like this, it is not too bad! Check that no solder has run between the pads. Plug in your crystal for the 80 m band, and connect up the 12 V power supply. Do not connect the Morse key yet and do not switch on the power.

Testing and operating

Clear your workbench of all metallic objects, slivers of copper, bits of wire, etc., switch on your power supply. With an external receiver, listen on and around your crystal's frequency for a signal. Remember that the oscillator runs all the time and, because you haven't yet connected the Morse key, your receiver is close enough to pick up the signal from the oscillator. This confirms that your oscillator is running. Switch off.

Connect the station aerial to the transmitter's aerial socket, and the receiver to the transmitter's receiver socket. Connect the Morse key, put the Receive/Transmit–Netting switch in the receive position and switch on. You should be able to hear stations in the normal way. Now put the switch in the Transmit–Netting position. Signals in the receiver should almost disappear, as the circuit has disconnected the receiver's aerial.

Tune the receiver until you can hear your own crystal oscillator signal. This is known as *netting*, tuning your receiver and transmitter to the same frequency. Pressing the key will now transmit your signals when the switch is in the Transmit–Netting position; switch back to the receive position to listen for stations answering your call. As soon as you are happy that your circuit is functioning properly, **you must build the low-pass filter circuit** before using the transmitter regularly.

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	100 kilohms (kΩ)
R2	See text
R3	1 kilohm (kΩ)

Capacitors

C1, C2, C3	100 nanofarads (nF), or 0.1 microfarad (μF)
TC1	3–60 picofarads (pF) trimmer

Semiconductors

TR1	ZTX651
TR2	ZTX751
TR3	VN10KM

Additional items

RFC	7 turns of 32 SWG enamelled copper on a ferrite bead
Switch	Double-pole changeover (DPDT or DPCO) toggle
Crystal	For 80 m band
Crystal holder	HC25 type
Sockets	According to station fittings

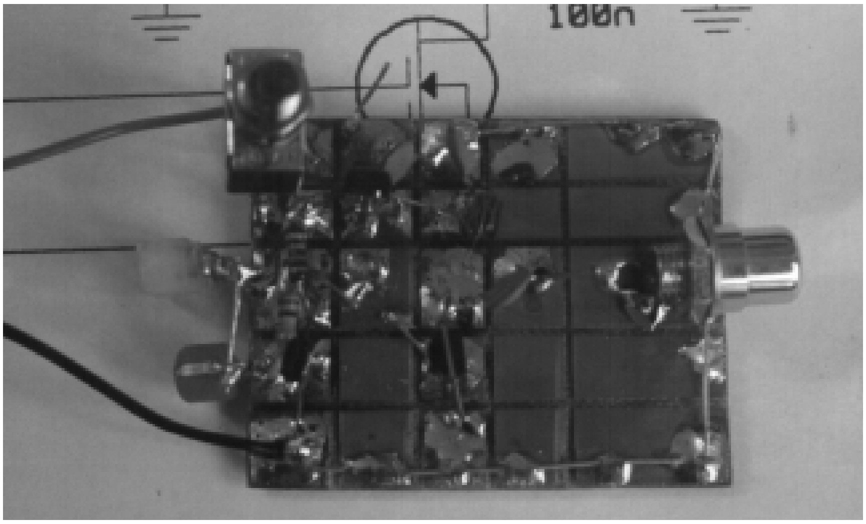

54 A 7-element low-pass filter for transmitters

Introduction

The simpler the transmitter, the more likely it is to radiate *harmonics* of its fundamental frequency. Harmonics are integral multiples of the frequency on which the transmitter is *designed* to operate. If you think you are transmitting on a frequency, *f*, for instance, you will also be radiating the harmonics of *2f*, *3f*, *4f*, . . . and so on. This results in your signals being heard on *several frequencies*, spread over a very wide frequency range. You are also contravening the terms of your licence. To avoid this, it is always advisable to use a *low-pass* filter between your transmitter and aerial. This is a filter which will *pass* your signal frequency, *f*, and all frequencies *below* it, but will *not pass* (attenuate) frequencies above *f* to any significant extent. The value of *f* is known as the *cutoff frequency* of the filter.

A design of 7-element low-pass filter

A 7-element low-pass filter (LPF) is so called because it has seven components, as the circuit diagram of **Figure 1** shows. Filters containing any odd number of elements are possible: a 3-element filter would

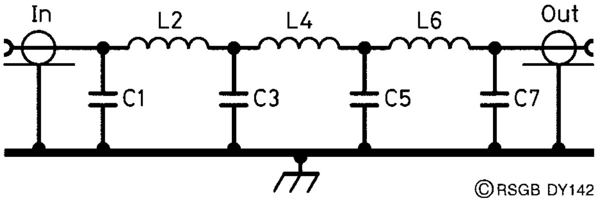

Figure 1 Circuit diagram of filter

comprise C1, L2 and C3 only, and is sometimes called a pi-network because the element disposition resembles the Greek letter pi (π); a 5-element filter would comprise C1, L2, C3, L4 and C5 only, and so on. In general, the more elements the filter has, the more effectively it attenuates signals above *f*.

The circuit of Figure 1 is designed to have an input and an output impedance of 50 Ω, which means that it can be placed in the aerial feed of any common transmitter. Filter design is a very complex business, and is best left to the experts. One such expert is W3NQN, who produced a number of computer designs of LPF using commonly available (*preferred value*) capacitors, and aimed specifically for use on amateur frequencies. The results of this work are condensed into **Table 1**. The inductors are wound on standard toroidal cores, and their details are included in the table.

Table 1 Filter component values for each brand

Band metres	*C1,7 pF*	*C3,5 pF*	*L2,6 turns*	*L4 turns*	*Core type*	*Wire SWG*
80	470	1200	25	27	T37-2	28
40	270	680	19	21	T37-2	26
30	270	560	19	20	T37-6	26
20	180	390	16	17	T37-6	24
15	82	220	12	14	T37-6	24
10	56	150	10	11	T37-6	22

Making the filter

The filter was made originally as an adjunct to the *Breadboard 80 m CW transmitter*, which you will also find in this book. It uses the same constructional technique, based on a single piece of plain, copper-clad

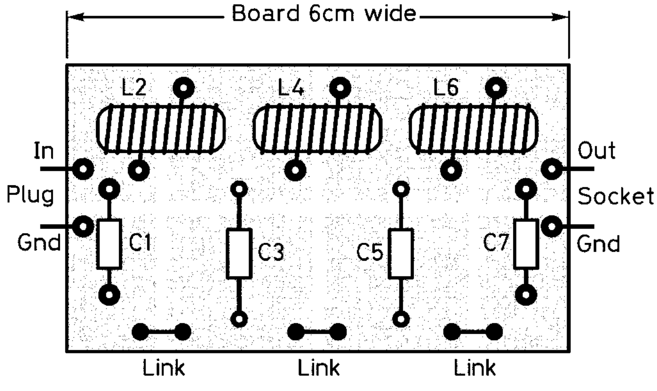

©RSGB DY141

Figure 2 Layout and construction of filter

PCB, with 'pads' created by using a sharp blade in a junior hacksaw. The cuts in the copper are shown in the layout diagram of **Figure 2**. There are two ways of mounting the components: the first way is to drill small holes in each pad, as shown in Figure 2, and mount the components through the holes in the normal PCB manner; the second way is to solder the components directly to the pads, in the way that was described for the *Breadboard transmitter.*

Winding the inductors is quite simple. All you need to remember is that each time the wire passes through the core counts as one turn. Cut off the spare wire at each end of each coil to about 1 cm, scrape off the enamel with sandpaper, and tin the exposed copper. See the transmitter description if you are unsure of how to do this. Note the wire links between each of the lower pads, forming a solid 'ground' for the elements. The prototype had a plug and socket on the ends, to match the transmitter and aerial terminations.

The type of capacitor used in the design is not critical; the polystyrene type works well.

55 Radio-frequency mixing explained

Introduction

Mixers find widespread use in electronic circuitry. Many of the projects in this book, together with every TV set and radio in the home, contain mixer circuits – a good indication of their usefulness.

Confused?

Audio mixers (as used in recording studios and radio broadcast stations) are used to **add** or 'balance' the signals from various sources such as microphones, CD players, etc. These have nothing whatsoever to do with radio-frequency (RF) mixers, and should never be confused with them.

RF mixers and beat frequencies

Instead of **adding** signals (as in the audio mixer), the RF mixer **multiples them together.** As you might expect, this has an entirely different effect. The two signals entering the mixer *beat* or *heterodyne* with each other to produce *signals on other frequencies*. One example of this occurs in sound, when two musical notes of almost the same frequency are heard together. Instead of hearing two separate notes, the listener hears one note whose *intensity* (*loudness*) appears to increase and decrease. This intensity variation is called a *beat*, and its frequency is equal to the difference in the frequencies of the two original notes. The technique is used by musicians to tune their instruments. If one note is known to be a correct frequency, the other can be tuned to it by making the beat frequency as close to zero as is possible.

Multiplying together

The process of mixing presupposes that we have a device which will automatically multiply two signals together. Fortunately, this is easy; so easy, in fact, that it often occurs when we do not want it! Multiplying is achieved by any device which is *non-linear*; this means a device whose output is not a constant factor larger than its input, something that can be achieved by many electronic devices and circuits.

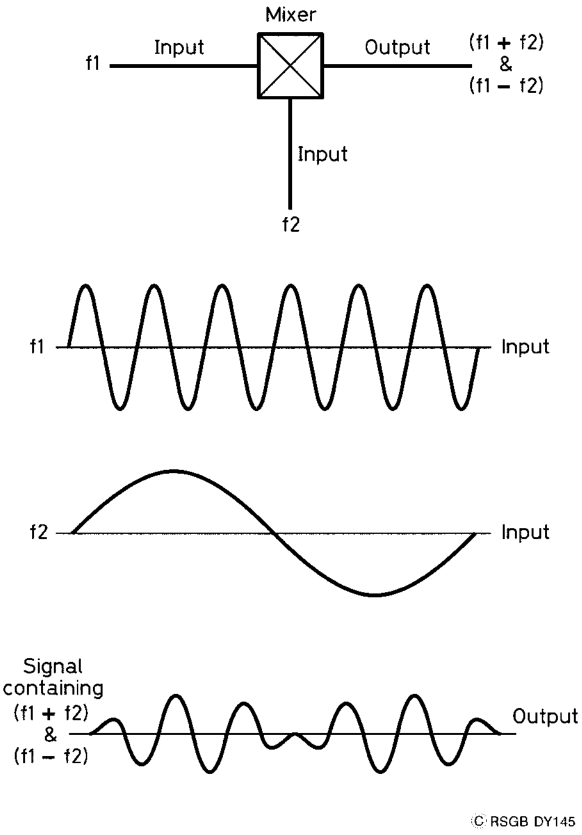

Figure 1 The effect of multiplying (or mixing) two signals together

Let us look now at what a mixer does in concrete terms. Suppose two signals, of frequencies *f*1 and *f*2 go into our mixer. These signals are shown in **Figure 1**. Putting numbers in, to make the situation clearer, suppose *f*1 is 1.000 MHz and *f*2 is 160 kHz. The beat frequency is the *difference* of these: 1.000 MHz – 0.160 MHz = 0.840 MHz, or 840 kHz. A mixer *also* produces an output at the *sum* of these frequencies; in this case the new frequency would be 1.000 MHz + 0.160 MHz = 1.160 MHz.

Suppose you fed the output of your mixer, operating with these input frequencies, into a receiver and tuned around to find what frequencies were present. You would find two signals, one at 840 kHz and one at 1.160 MHz, showing that the two 'new' frequencies were very real!

In addition to drawing out the waveform of the resultant signal, as in Figure 1, we can draw the inputs and outputs on a frequency axis, to form a *spectrum* of the signal components. This is done in **Figure 2**. The top two diagrams show the input signals at *f*1 and *f*2. The bottom diagram shows the output signals in relation to the input signals. Depending on the type of mixer used, one or both of the input signals would be removed.

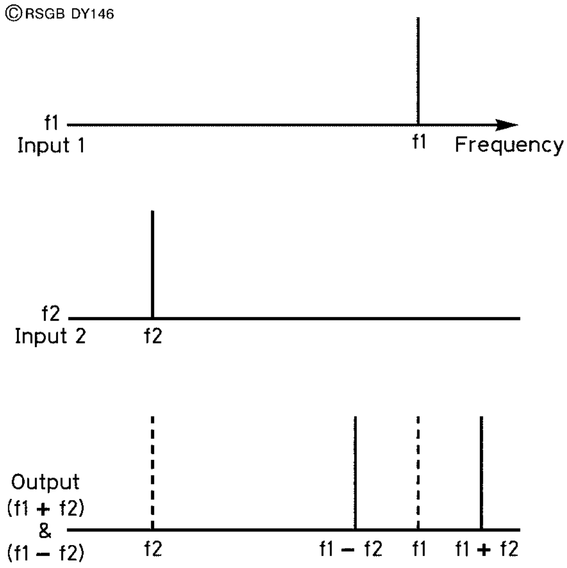

Figure 2 The result of multiplying or mixing two signals together as seen on a spectrum analyser

A mixer in every radio

Basically, a mixer is used to change a signal from one frequency to another, something it does without altering the characteristics of the incoming signal. If the incoming signal is amplitude modulated (AM), then the frequency-changed signal would be AM also. The same applies to FM, SSB, CW, and all other modulation forms you can think of. This explains why mixers are often called *frequency changers*.

Frequency changing is the key process in the type of radio known as a *superheterodyne* (or *superhet*). By mixing the incoming signal with a variable-frequency *local oscillator* as **Figure 3**, shows, the signal can be converted to the fixed frequency of a filter and amplifier. This is useful

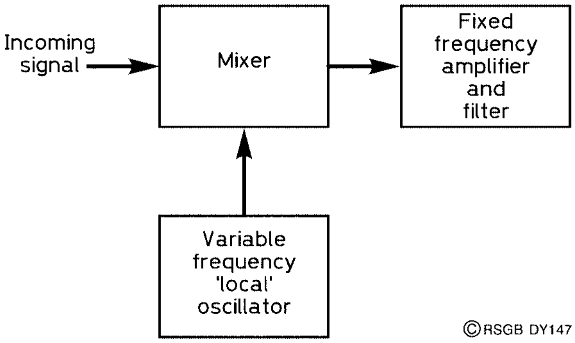
Incoming signal
Mixer
Fixed frequency amplifier and filter
Variable frequency 'local' oscillator
©RSGB DY147

Figure 3 The basic idea of a superhet receiver

because it is easier to make a very high-quality filter and amplifier at a single fixed frequency, than at a variable frequency.

All TV receivers and virtually all radio receivers (and transmitters) use mixers. Both the *Yearling* and *Colt* receivers (see the relevant projects) are superhets, and use mixers.

56 A voltage monitor for a 12 V power supply

Introduction

If for any reason, the stabilisation of your main 12 volt power supply unit (PSU) breaks down, it is possible that a voltage much higher then the nominal 13.8 V will be applied to your precious equipment. If you would like to know the instant that this occurred, and hence be able to switch things off before it was too late, then this circuit is what you need. It will give audible *and* visual indications if the voltage rises above 14.4 V, and a visual indication only if the voltage is reduced.

The circuit

The circuit uses three ICs and is shown in **Figure 1**. The circuit is powered by the PSU whose output is being monitored, and the circuit's immunity to supply line variations is secured by the 6 volt regulator, IC1. The heart of the circuit is IC2, an LM3914; it is a *bargraph driver*, which operates ten LEDs in a display resembling a thermometer – the string of lit LEDs increases in length as the voltage on pin 5 increases.

The input voltage range on pin 5 is 1.2 V maximum, making each LED correspond to one-tenth of this, which is 0.12 V, the *step size*. R1 and R2 act as a voltage divider, so that voltages of up to the maximum of 14.4 V may be applied to R1 without exceeding 1.2 V at pin 5. R3 sets the brightness of the LEDs and R4 determines the step size.

IC3 is an *opto-isolator*, a device containing an LED and a phototransistor in one package. This enables the piezoelectric sounder to operate without affecting the operation of the bargraph driver. The input to IC3 is provided by the voltage on D8, so that if any of the LEDs at or above D8 are lit, the

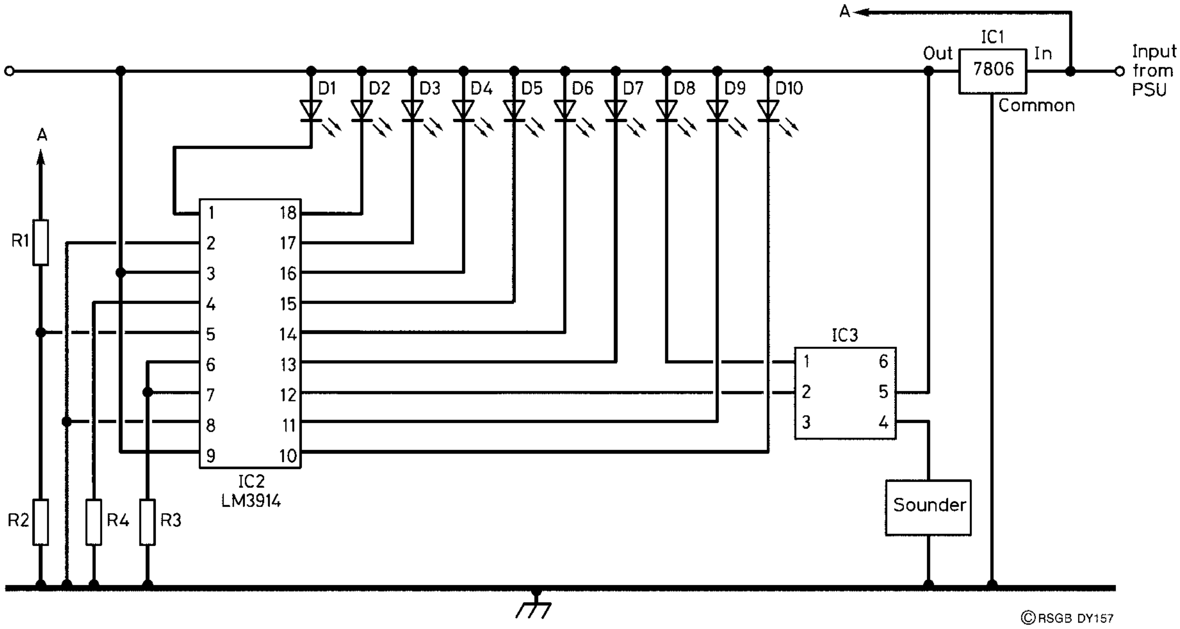
A
IC1
Out
In
7806
Input
from
PSU
Common
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
A
R1
R2
R4
R3
1
2
3
4
5
6
7
8
9
18
17
16
15
14
13
12
11
10
IC2
LM3914
IC3
1
2
3
6
5
4
Sounder
©RSGB DY157

Figure 1 PSU monitor, circuit diagram

sounder will operate, indicating overvoltage. Despite the fact that some of your other equipment might be damaged by this overvoltage, the monitor circuit itself is unaffected.

Construction

The prototype was built on two pieces of Veroboard of the copper-strip variety. The main (circuit) board measured 15 strips by 25 holes, and the display (LED) board measured 4 strips by 30 holes. Track cuts are necessary in each board. The correct places are shown for the main board in **Figure 2**, but there is sufficient flexibility in the layout of the display board for a prescriptive layout not to be needed. All the anodes of the LEDs are connected to the same strip, which makes things comparatively simple!

For the main board, insert the Veropins and the wire links first, and solder them to the copper strips. Then fit the IC holders and the resistors. Fit the IC holders with their notches towards the top of the board. When fitting the voltage regulator, IC1, note than the centre lead (the ‘common’ lead in Figure 1) does *not* go to position A3; the track should be cut at A3, and the common lead soldered at B3. This is indicated in Figure 2.

Wire up the LED board with its ten LEDs and 11 connecting wires, each about 10 cm long. This length depends on how far away from the main

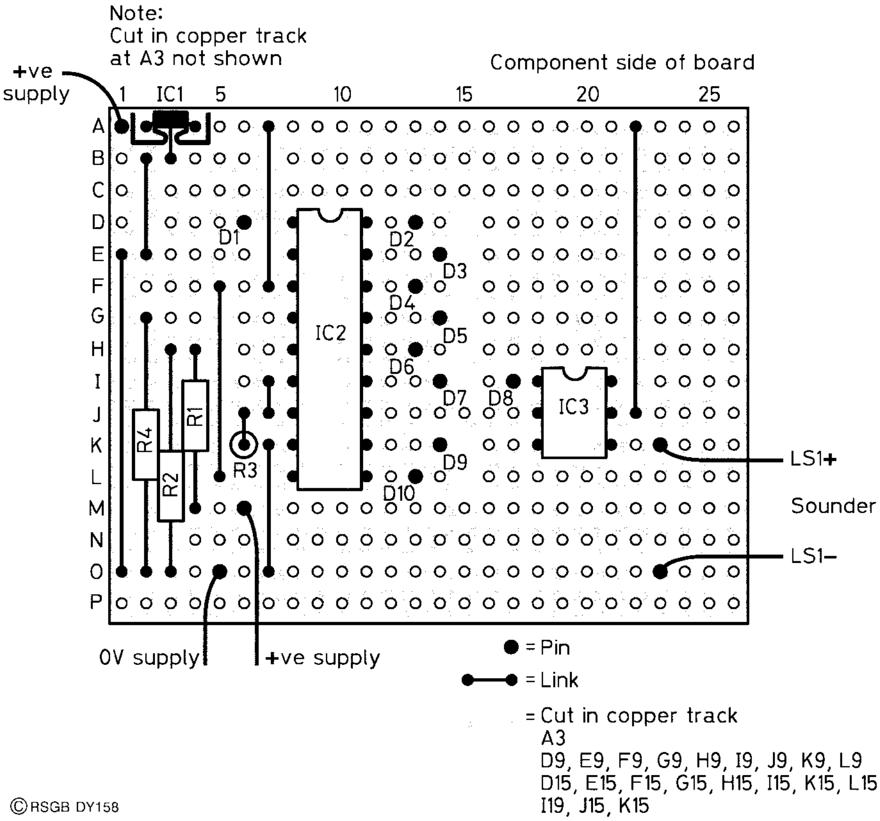
Note:
Cut in copper track
at A3 not shown
Component side of board
+ve supply
1
IC1
5
10
15
20
25
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
IC2
IC3
R1
R2
R3
R4
LS1+
Sounder
LS1–
0V supply
+ve supply
● = Pin
●—● = Link
= Cut in copper track
A3
D9, E9, F9, G9, H9, I9, J9, K9, L9
D15, E15, F15, G15, H15, I15, K15, L15
I19, J15, K15
©RSGB DY158

Figure 2 PSU monitor, PCB layout

board you are planning to mount the display. The LEDs should have different colours: three orange, four green and three red, to indicate 'low', 'medium' and 'high' voltage.

Now solder the other ends of these leads to the Veropins on the main board, making sure that the order is correct. Finally, connect the piezoelectric sounder to the main board; the polarity is important, so make sure the red lead goes to pin 4 of IC3 and the black lead to the ground rail. If you were careful to fit the IC holders with their notches in the correct positions, match these up with the notches on the ICs before pushing home the ICs *gently*. Check your circuit for solder bridges and unwanted pieces of copper swarf before screwing the small heat sink to the voltage regulator, IC1.

Testing

(a) **With a variable-voltage PSU**. Set the PSU for 10 V, connect the circuit and switch on. Increase the voltage slowly, and check that each LED lights up after the one before it. As the voltage exceeds about 14.4 V, the first red LED should light and the sounder should operate. If an LED does not illuminate, you should immediately suspect either a dry joint or an incorrect LED polarity. The voltage at which the first red LED lights can be adjusted by varying R1; increase R1 if the LED comes on too early; decrease R1 if the LED comes on too late.

(b) **Without a variable-voltage PSU**. For your 'variable supply', you can use several AA-type 1.5 V cells (or 1.2 V NiCad cells) in series. The voltages produced by a range of cells is shown in **Table 1** – because of the lower voltage of NiCad cells, more of them are needed to produce a given voltage.

Once the operation of the circuit has been checked, it can be fitted into a plastic or metal box. Only two connections are needed for the PSU. You may want to drill some holes in the case to increase the apparent loudness of the sounder.

Table 1 Test voltages available from batteries in series

No. of batteries	*Voltage, nicads*	*Voltage, dry cells*
7	8.4	10.5
8	9.6	12
9	10.8	13.5
10	12	15
11	13.2	–
12	14.4	–
13	15.6	–

Parts list

Resistors: all 0.25 watt, 5% tolerance

R1	11 kilohms (kΩ) – see text
R2	1 kilohm (kΩ)
R3	1.2 kilohms (kΩ)
R4	18 kilohms (kΩ)

Semiconductors

IC1	L7806
IC2	LM3914
IC3	Opto-isolator – Maplin code WL35Q
D1–D3	3 mm LED, orange
D4–D7	3 mm LED, green
D8–D10	3 mm LED, red

Additional items

LS1	Piezoelectric sounder, wire leads
	6-pin DIL socket for IC3
	18-pin DIL socket for IC2
	Veroboard – two pieces for main and display boards, see text for sizes
	Veropins
	Heat sink for IC1
	Single-core insulated wire for links
	Insulated stranded wire for interconnecting the boards
	Case as required

57 A 1750 Hz toneburst for repeater access

Introduction

Repeaters across the UK and much of Europe need an access tone to switch the transmitter from standby ready for use. Commonly, this is a 1750 Hz tone of duration no greater than half a second. Although many UK repeaters may now be accessed using the *continuous tone-coded squelch system* (CTCSS – see the *RSGB Yearbook*), you may wish to access a repeater whose CTCSS frequency you don't know; in this case, using the universal 1750 Hz tone will gain you access. Commercial transceivers are usually fitted with an automatic toneburst, but if you are using a home-made design, then you may want to incorporate this little circuit.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current

and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery.

Circuit description

The circuit is shown in **Figure 1**. The tone is generated by an integrated circuit oscillator (IC1), whose frequency is controlled by a *ceramic resonator*, XL1. Its frequency is very high, and is divided down to the 1750 Hz needed by the same chip. The ceramic resonator is designed to operate at 455 kHz, the intermediate frequency of many receivers. Because all divider circuits use powers of 2, we need the oscillator to run at 448 kHz so that when it is divided by 256 ($256 = 2^8$), we end up with 1750 Hz. Try it on your calculator:

$$\frac{448\ 000}{256} = 1750.$$

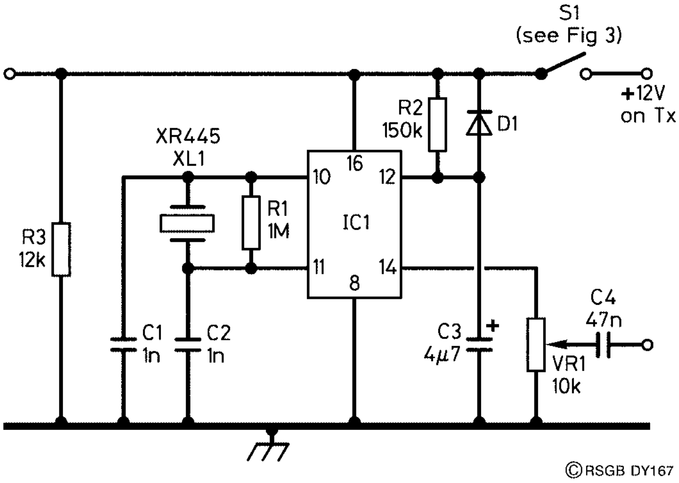

Figure 1 Toneburst module, circuit diagram

We use C1, C2 and R1 to *pull* the frequency of the oscillator away from 455 kHz to 448 kHz. The divider chain has eight *counters* in it, and each counter divides the frequency of the signal it sees by two, giving the final division of 256.

This counting process can be stopped at any time by taking the voltage on the *reset* pin (pin 12) up to the supply voltage. When the circuit is switched on by closing S1, pin 12 is at 0 V because C3 is discharged. The oscillator runs, producing the output frequency of 1750 Hz. As time progresses, C3 charges up through R2 and the voltage on pin 12 rises. When this has risen sufficiently, and in a time determined by the values of C3 and R2, the counter resets and stays in the reset state; no division takes place and there is no output. The duration of the toneburst is thus governed by C3 and R2.

When S1 is opened, the circuit is switched off, and C3 is discharged through D1 and R3, ready for the next toneburst. If you have used a repeater, you will know that a toneburst is needed **only** to activate a repeater in the standby condition; it is not needed once a contact has been established.

VR1 adjusts the amplitude of the tone fed to the microphone, and C4 prevents any voltage that may be present on your microphone connector from damaging the integrated circuit.

Construction

The prototype circuit was built on Veroboard of the copper-strip variety, measuring 20 holes by 14 strips. The layout is shown in **Figure 2**. Make the track cuts first, and check that there are no slivers of copper wedged between adjacent tracks. Then, solder in the IC socket (with the notched end

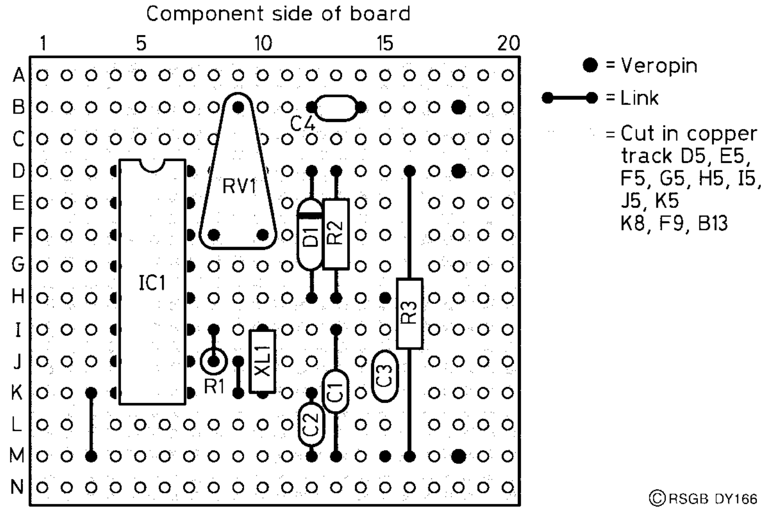

Figure 2 Veroboard component layout

facing towards track A), the wire links and the three Veropins. Having done this, solder in the resistors, capacitors and diode, making sure that D1 and C3 are the right way round! Using your best soldering technique, solder in the ceramic resonator quickly, to prevent heat damage. Recheck your circuit, check for solder splashes and bridges, and then gently insert IC1 into its socket, matching up its notch with that of the socket.

Testing

Set VR1 to half-way and connect a crystal earpiece to the output; apply power to the circuit. You should hear the tone, lasting for about half a second. If there is no tone, disconnect your circuit from the power supply, and check for dry joints in the vicinity of pin 12. Is the diode, D1, the correct way round? Is C3 the correct way round? Did you choose to ignore the CMOS safety precautions given earlier?

Once the circuit is working, you need to decide how you are going to connect it to your transmitter. Two options are shown in **Figure 3**. If you have access to a point in your transmitter circuit that has between 9 V and 12 V positive on it during transmit, you can use this to power your circuit. As the toneburst is needed only for repeaters, the switch, S1, disconnects it when not needed, as shown in Figure 3a. If you want the circuit to be self-powered, then a 9 volt PP3 battery may be used; Figure 3b shows this configuration.

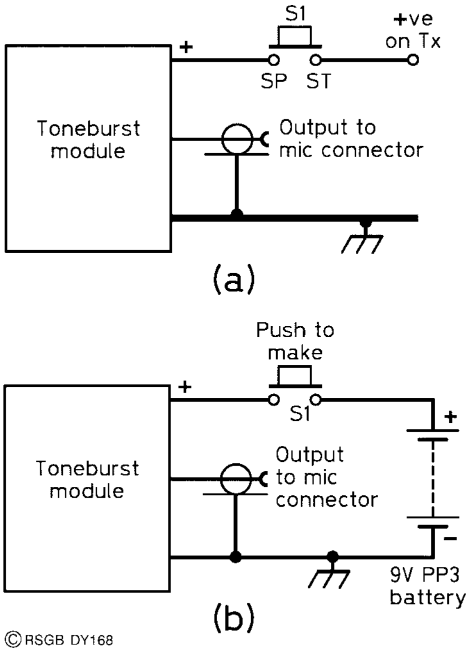

Figure 3 Toneburst module: two alternative switching arrangements

The output from the circuit board is fed directly into the microphone socket, in parallel with the microphone itself; use thin coaxial or screened cable for this lead, or you may induce hum into the microphone circuit and suffer from RF breakthrough into the audio circuits. To adjust the setting of VR1, start with it at the zero output position and connect a dummy load to your transmitter. Slowly, increase the output while monitoring your transmitted signal on another nearby receiver. Make sure you do **not** increase the output so far that the signal sounds distorted. If you would prefer that the tone was on continuously while you made this adjustment, simply connect a wire across C3 remembering, of course, to remove it as soon as you have completed the test!

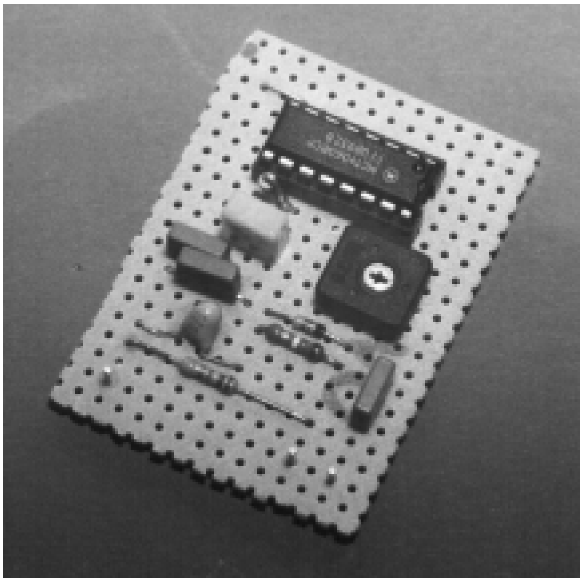

Parts list

Resistors: all 0.25 watt, 10% tolerance (or better)

R1	1 megohm (MΩ)
R2	150 kilohms (kΩ)
R3	12 kilohms (kΩ)
VR1	10 kilohms (kΩ) horizontal preset

Capacitors: all 16 V WKG or higher

C1, C2	1 nanofarad (nF) disc ceramic
C3	4.7 microfarad (μF) tantalum bead
C4	47 nanofarads (nF) disc ceramic

Semiconductors	
IC1	4060
D1	1N4148
Additional items	
XL1	XR455
	Veroboard (see text for size)
	Veropins (3)
S1	Switch (momentary action push-to-make SPST)
	16-pin DIL socket for IC1
	Single-core insulated wire for links
	Coaxial or screened cable for microphone connection

58 A circuit for flashing LEDs

Introduction

There are many occasions when one's attention needs drawing to the fact that something important has happened. A single red light coming on is seldom sufficient to attract attention, particularly if it is surrounded by other lights and indicators. The eye is known to be very sensitive to *changes* in its peripheral vision; such changes can be brought about by movement or by differences in light level – a flashing light, for example. So, a circuit that flashes a single LED or a pair of LEDs finds plenty of uses in the amateur station.

Warning

This circuit uses a member of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton

destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from destruction while it is connected to the battery. However, when the battery is removed, the same care should be exercised with its handling, because there is no supply decoupling capacitor across the IC.

Basic description

LEDs can be made to flash (switch on and off) by driving them from sources that switch on and off. Such a source is an *astable multivibrator*. If you have built or read about *A basic continuity tester*, elsewhere in this book, you will have come across such a beast before. That circuit used an astable multivibrator made from two transistors. This new circuit achieves the same behaviour from a single integrated circuit, the CMOS 4011. To give it its full description, the 4011 is a *quad 2-input NAND gate*. Quite a mouthful, but all it means is that inside the chip are four NAND gates, each with two inputs.

A NAND gate needs a positive voltage (known as a *logic 1*) on *both* inputs in order to produce zero volts (known as *logic 0*) at the output. Two NAND gates can be connected, as are A and B in **Figure 1**, to make our astable multivibrator. The combination of A and B has been described as the most perverse circuit in electronics; as soon as the output goes to logic 1, the circuit decides that it would prefer to have a logic 0 there, and switches over. With logic 0 at the output, the circuit now prefers to have logic 1 there, and so it goes on! We are going to use this continuous switching backwards and forwards to flash two LEDs. The rate at which A and B 'change their minds' is the frequency at which our LEDs will flash, and is controlled by the charging and discharging times of C1 through R2 and C2 through R1. As the values of R1 and R2 are the same, and those of C1 and C2 are the same, the ON and OFF states of the circuit are the same.

Gates C and D do not contribute to the flashing action; they act as *buffers* to isolate the LEDs from the multivibrator circuit itself. You will find in electronics that an oscillator is seldom used to drive another device *directly*; there is usually a buffer between it and the stage it drives.

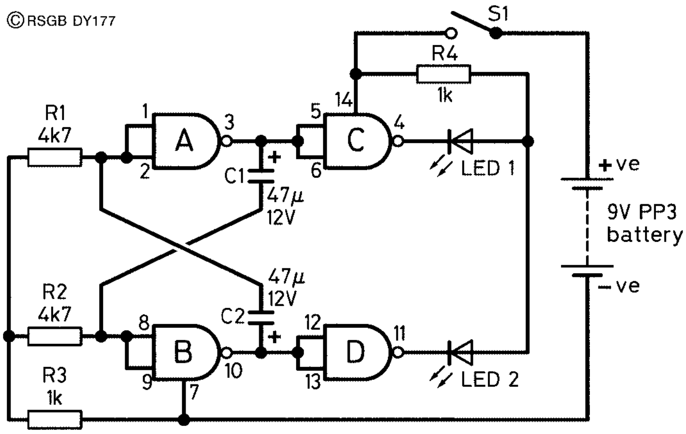

Figure 1 Flashing LEDs, circuit diagram

A characteristic of all multivibrators, astable or not, is that they have two outputs. In this case, those outputs are at pins 3 and 10, which are then buffered and appear at pins 4 and 11, respectively. When one output is at logic 1, the other is at logic 0, and vice versa. This means that LED1 is off when LED2 is on, and LED1 is on when LED2 is off, the two states switching backwards and forwards at the frequency of the oscillator.

Construction

Read the warning at the beginning of this article again. It is not intended to scare you off from building this, but is a genuine piece of advice which can save you time and irritation when all your labours result in a circuit that doesn't work! That extra bit of care can make all the difference!

Veroboard (the copper strip type) is used for the layout, shown in **Figure 2**. It measures 20 holes by 12 strips. **Be aware that there is no row 'I' in the layout, so don't miscount when you are placing components on the board!**

Firstly, cut the tracks using a 3 mm (⅛ inch) twist drill held between thumb and forefinger; check that there are no slivers of copper bridging any of the tracks, and that the tracks have been completely cut by the drill. Solder in the components carefully. Leave the IC in its carrier for the time being, and solder in the IC socket, with the notched end facing row A. On completion, check the circuit carefully. If you are happy that it is correct, follow the instructions given earlier and fit IC1 into its socket, matching up the two notches. Connect the battery and switch on. The two LEDs should flash on and off alternately. If only one LED flashes, you have probably connected the other one the wrong way round. Switch off, check and correct if necessary. If neither LED flashes, you must have a significant error in your circuit, which will need checking *again*! Or did you choose to ignore the handling precautions for the CMOS chip?

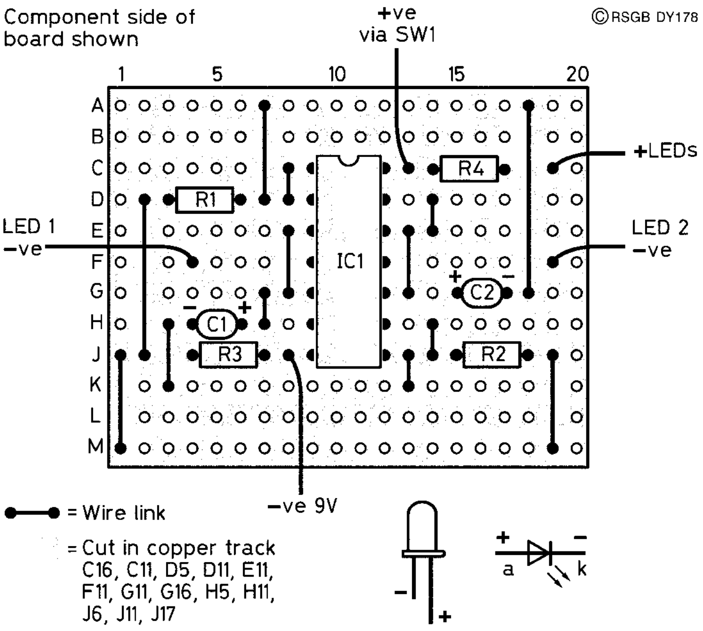
Component side of board shown
+ve
via SW1
©RSGB DY178
1
5
10
15
20
A
B
C
D
E
F
G
H
J
K
L
M
R1
R4
IC1
C1
C2
R3
R2
+LEDs
LED 1
−ve
LED 2
−ve
−ve 9V
= Wire link
= Cut in copper track
C16, C11, D5, D11, E11,
F11, G11, G16, H5, H11,
J6, J11, J17
−
+
+
−
a
k

Figure 2 Flashing LEDs, component layout

The board can be mounted near to the point where you want your flashing LEDs to be seen, although long leads to the LEDs are acceptable. The LEDs can be different colours – it's all up to you now!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better	
R1, R2	4700 ohms (Ω)
R3, R4	1 kilohm (kΩ)
Capacitors	
C1, C2	47 microfarads (µF) 12 V WKG
Semiconductors	
IC1	4011
Additional items	
LED1, LED2	Any size of LED, any colour
S1	SPST on/off
	Plastic box if needed, 8.5 by 5 by 2.5 cm

Source

Components are available from Maplin.

59 Digital logic circuits

Introduction

Logic circuits form the backbone of even the most advanced computer, yet their basic operation can be demonstrated by a couple of switches, a battery and a bulb.

Logic using switches

Everyone reading this article will look at **Figure 1** and know immediately how it works and be able to write down something like ‘When switch A and switch B are closed, the light will come on’. Without knowing it, you have written down a logic statement involving the so-called **AND** operation; the light comes on only when switches A **AND** B are ON. Below the circuit in Figure 1, is a table showing the only possible positions of the two switches and the state of the bulb for each position. This is called a *truth table*, and is frequently used in logic analysis.

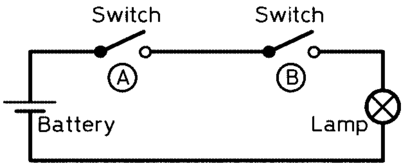

A	B	Lamp
Off	Off	Off
On	Off	Off
Off	On	Off
On	On	On

©RSGB DY182

Figure 1 Switches and lamp AND gate

Figure 2 shows a different circuit. Here, the two switches are in parallel rather than in series, as was the case in Figure 1. Again, if you analyse the circuit in words, you would say that the light will be on when switch A **OR** switch B is ON. This is an example of the **OR** operation, and its truth table is shown in Figure 2. The statement above is not complete, however; can you see why? The truth table will show you. The light comes on if A is ON, **OR** if B is ON, ***OR*** if A **AND** B are both ON. That third condition is easy to miss, but don’t worry about it!

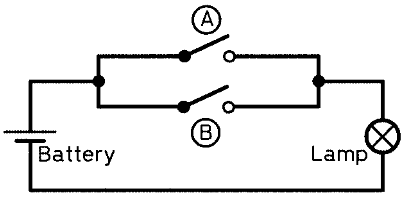

A	B	Lamp
Off	Off	Off
Off	On	On
On	Off	On
On	On	On

©RSGB DY184

Figure 2 Switches and lamp OR gate

Believe it or not, some very complicated logic is possible (in theory) using switches and lights, but it is highly impractical and would be very slow. This is where electronic logic circuits come in.

Switches with no moving parts

You may have come across projects in this book where a statement is made such as '. . . the transistor is being used as a switch . . .'. Transistors *can* be used as switches, as were thermionic valves in the world's first programmable computer *Colossus*, at Bletchley Park. However, technology has moved on from valves, through transistors to *logic gates*, combinations of electronic switches designed specifically to perform logic functions.

These act on voltage levels as their inputs and produce changes in voltage levels as their outputs. A positive voltage is called *logic 1*, and corresponds to a switch being ON in our previous descriptions; a zero voltage is called *logic 0*, and corresponds to a switch being OFF. The output from a logic gate (normally labelled Q) is also logic 1 or logic 0, corresponding to our light being ON or OFF, respectively, in our switch analogy.

Many logic devices operate from a stabilised 5 V supply, and this determines the *ideal* voltages corresponding to the two logic states:

logic 0 ≡ 0 V,

logic 1 ≡ 5 V.

The world isn't an ideal place, so the *real* voltage ranges used by the logic gates are:

logic 0 ≡ 0.0 to 0.4 V,

logic 1 ≡ 3.0 to 5.0 V.

The **AND** circuit of Figure 1 is now called an **AND gate**, and requires logic 1 inputs on A **AND** B to produce a logic 1 at the output. **Figure 3** shows this. The truth table is identical with that of Figure 1 – logic 1 replaces ON and logic 0 replaces OFF. Now compare Figure 2 with **Figure 4** – circuits and truth tables for the **OR** function. Again, we have exact similarity.

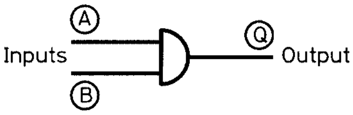

Inputs		Output
A	B	Q
0	0	0
0	1	0
1	0	0
1	1	1

©RSGB DY183

Figure 3 Electronic AND gate

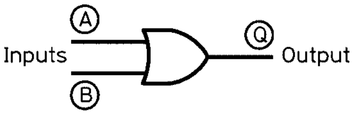

Inputs		Output
A	B	Q
0	0	0
0	1	1
1	0	1
1	1	1

©RSGB DY190

Figure 4 Electronic OR gate

There is another very common logic gate, which performs the **NOT** function. It is easy to understand. Just ask yourself the question 'What is **NOT** logic 0?', and the answer is obviously 'logic 1'. Similarly, logic 0 is **NOT** logic 1. A **NOT** gate simply changes the logic state of the input; it is also known (because of this behaviour) as an *inverter*. Its symbol and truth table can be found in **Fig 5**. Note the little circle on the output of the gate in Figure 5. In logic circuits, this symbol always implies inversion, or the presence of a **NOT** gate. Keep an eye open for it!

So far, the logic functions we have discussed have all been words which we use in everyday language, which has made the electronic interpretation of them relatively easy. Now we must introduce a function for which there is *no* analogy in normal speech – the **NAND** function. This means a

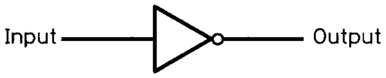

Input	Output
1	0
0	1

©RSGB DY185

Figure 5 Electronic NOT gate

combination of an **AND** gate and a **NOT** gate, and the sharp-eyed reader will have spotted the little circle added on to the normal **AND** gate symbol in **Figure 6**!

To make things easier to understand, the truth table in Figure 6 has four columns, not three as in previous tables. The third column is the standard **AND** output – compare it with the third column in Figure 3. That is the output from the **AND** gate *before* it encounters the little circle that inverts it, so the final output from the **NAND** gate is the output of the **AND** gate, inverted! The third and fourth columns are the inverse of each other.

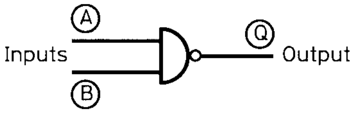

Inputs		and	Output
A	B		Q
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

©RSGB DY186

Figure 6 Electronic NAND gate

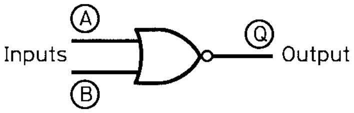

Inputs		or	Output
A	B		Q
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

©RSGB DY187

Figure 7 Electronic NOR gate

We can add the inverting operation to the output of an **OR** gate also, producing a **NOR** gate! The symbol and its truth table are shown in **Figure 7**. Again, notice that the final output is that of the ordinary **OR** gate, inverted!

These new functions of **NAND** and **NOR** are used more than the **AND** and **OR** functions because it makes other circuits easier to design using combinations of these gates.

A taste of Boolean algebra

The design of circuits using combinations of logic gates usually begins with a little mathematics, where the functions to be implemented are analysed. The mathematics used is surprisingly simple, and is a slightly changed version of ordinary algebra called *Boolean algebra*, which allows manipulation of logic functions to be made. Normal algebra has operations in it such as addition, subtraction and multiplication and division. The mathematician Boole found that the logical **AND** operation could be handled by the algebraic operation of *multiplication* (symbols × or •), and the **OR** operation by the algebraic operation of *addition* (symbol +). The **NOT** operation involved a new symbol, that of a bar over the input being inverted, such as $\overline{A}$.

So, our five basic logic operators can now be written in a mathematical form:

AND	$Q = A \times B$
OR	$Q = A + B$
NOT	$Q = \overline{A}$
NAND	$Q = \overline{A \times B}$
NOR	$Q = \overline{A + B}$.

Using logic operations in a mathematical form enables the most complex logic to be designed, simplified and converted into circuit diagram form in a very efficient and rapid way. There are more logic operators than the five we have considered here but, in general, they can all be broken down into combinations of these gates alone!

60 A resistive SWR indicator

Introduction

When a transmitter produces some output power, we want to make sure that as much as possible of this power is radiated by the aerial. This often requires the use of an *aerial tuning unit* (ATU), which matches the aerial impedance to that of the transmitter.

How do you know when this matching has been achieved? The most usual way is to use a *standing-wave ratio* (SWR) indicator. If the impedance of the aerial does not match that of the transmitter output, some of your transmitter power (also known as the *forward* power) is reflected back along the aerial feeder and back into the transmitter, where it causes excess heating. The forward and reflected waves interact along the feeder to produce a wave whose position remains constant, and which is therefore called a *standing wave* or a *stationary wave*. An SWR meter simply indicates forward power and reflected power, and adjustments are made to your ATU until the reflected power is as small as possible (ideally zero, of course). If there is no reflected power then, by a process of elimination, all your forward power is reaching the aerial!

Sampling the RF

Whatever type of SWR indicator you use, it must use some sort of sampling circuit to pick up the forward and reflected waves. The project *A standing-wave indicator for HF*, elsewhere in this book, uses a toroidal transformer to separate the readings for the forward and reverse waves. This design differs in that it measures the voltages across resistors through which the RF current is passing. Its advantages are:

(a) it uses cheap parts – four resistors, two capacitors and a diode, together with a rotary switch, a surplus meter, a preset potentiometer and two sockets;
(b) within this SWR indicator, there is always a resistive path for the RF current from the transmitter, formed by R1, R2 and R3; this can prevent damage to simpler home-made transmitters, which may be damaged during adjustment of the ATU when using more conventional SWR indicators.

The only disadvantage of this form of indicator is that it must be switched out of circuit once the ATU has been adjusted for a particular band.

Construction

The SWR indicator is very simple to build, as most of the components can be mounted on the back of the 3-way rotary switch. This is shown in **Figure 1**. The switch is a 4-pole, 3-way rotary type, of which only two poles are used.

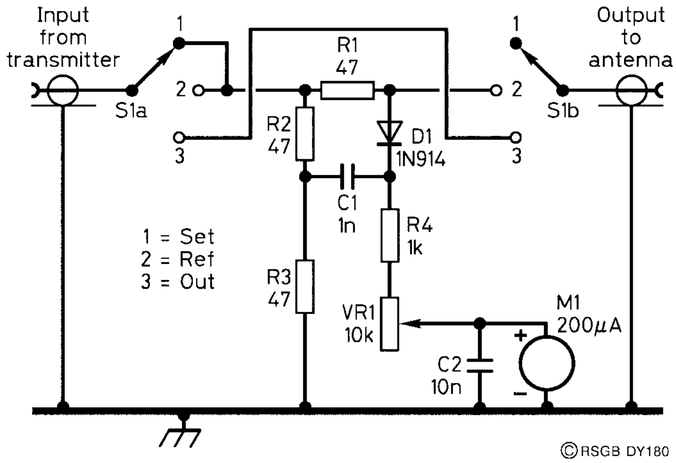

Figure 1 Resistive SWR meter, circuit diagram

Because the other switch contacts are not being used, they can be employed as support tags for other components. The ground wires are all soldered on to the metal frame of the switch. If your switch frame is of all-plastic construction, then a 12 SWG copper wire run around the switch will make a good earth connection to the metal case for the leads shown in Figure 1. The preset potentiometer used to control the sensitivity of the circuit can be mounted directly on the meter tag.

Resistors R1, R2 and R3 handle the RF power during the tuning-up process. If you have them, use 1 watt resistors; otherwise, you can use two 100 ohm half-watt resistors in parallel for each of R1, R2 and R3. The meter, M1, can be any DC type of sensitivity around 200 μA.

In use

First, find a clear frequency, and without the indicator in circuit, check that the frequency really *is* clear by asking and identifying yourself. If it is, connect the indicator between the transmitter and the ATU which, in turn, is connected to your aerial. Turn SW1 to the SET position and key the transmitter. Adjust VR1 until the meter reads full scale. Switch off the transmitter. Turn SW1 to the REF position and key the transmitter again. Adjust the ATU until the lowest reading is obtained on the meter. Switch off the transmitter. For the chosen frequency, you have adjusted your ATU for minimum reflected power and hence the lowest SWR. You will need to repeat the process when you change bands, and possibly when you change frequency within the same band. Switching SW1 to the OUT position, you are ready to transmit. You may have noticed that it is good practice to switch the transmitter off when operating SW1. Get into that habit!

Parts list

Resistors	
R1, R2, R3	47 ohm (Ω) 1 watt (or 2 × 100 Ω, ½ watt, see text)
R4	1 kilohm (kΩ) 0.25-watt
VR1	10 kilohms (kΩ) preset
Capacitors	
C1	1 nanofarad (nF)
C2	10 nanofarads (nF)
Semiconductors	
D1	1N914 or similar
Additional items	
SW1	4-pole 3-way rotary switch, of which only 2 poles are used
M1	200 µA DC
	Case

61 An audio filter for CW

Introduction

This is a simple passive circuit (it has no power supply) that adds some audio selectivity for Morse code reception and also includes a very simple noise limiter that gives a visual indication of when noise spikes are being removed!

The circuit

Figure 1 shows the complete circuit. The tuned circuit of C1 and L1 resonates very close to 800 Hz, so initially you will have to tune a signal in carefully until it sounds loudest – you will soon be able to do this without thinking. The two LEDs connected back to back across the signal path act as a noise limiter, reducing the amplitudes of static crashes and noise from car ignition systems, etc. The LEDs blink when they conduct – this is not necessary to the operation of the circuit, but adds a little colour to your listening! The noise limiter *does* make listening more comfortable, though.

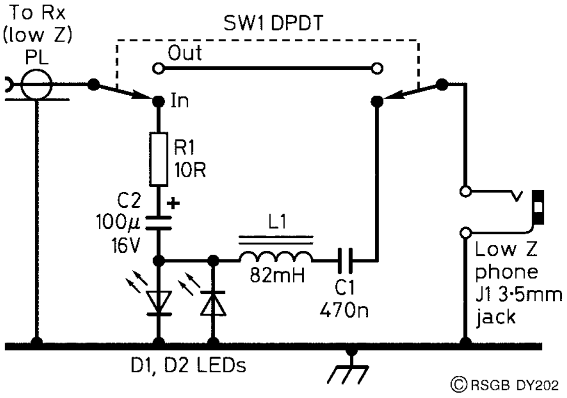

Figure 1 CW filter, circuit diagram

Construction

The circuit layout is shown in **Figure 2**. Point-to-point wiring is used, with a small tag-strip being the only item used for the extra support of C2 and R1. The LEDs and SW1 support the other components. An aluminium box (**Figure 3**) is used to make the circuit tidy and usable.

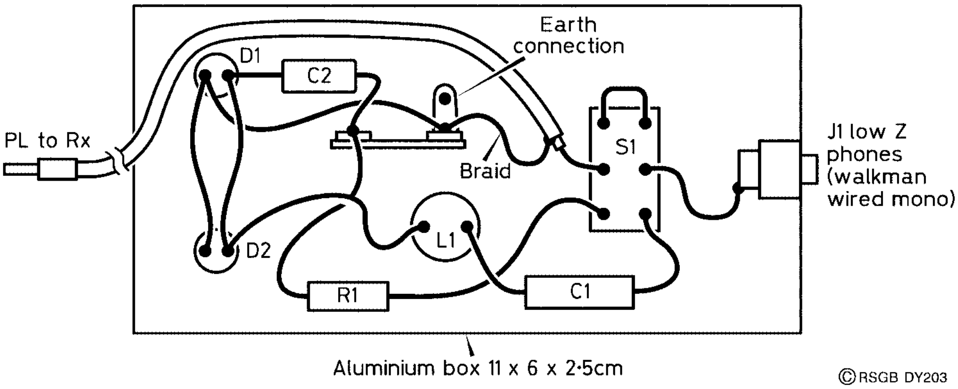
Earth connection
D1
C2
PL to Rx
Braid
S1
J1 low Z phones (walkman wired mono)
D2
L1
R1
C1
Aluminium box 11 x 6 x 2·5cm
©RSGB DY203

Figure 2 CW filter, component layout

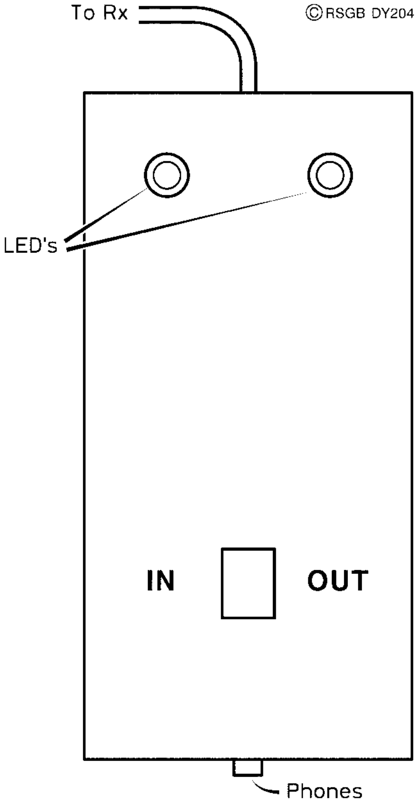
To Rx
©RSGB DY204
LED's
IN
OUT
Phones

Figure 3 CW filter, front panel

You must use low-impedance headphones for the circuit to perform properly. Plug the completed unit into the headphone socket of your receiver and adjust the volume so that the LEDs are just *not* blinking on normal audio. The unit should be switched out of circuit for speech listening; this is the purpose of the toggle switch, SW1.

Parts list

Resistor

R1	10 ohms (Ω), 0.25 watt, 10% tolerance

Capacitors

C1	470 nanofarads (nF)
C2	100 microfarads (μF) electrolytic, 16 V WKG

Inductor

L1	82 millihenries (mH)

Semiconductors

D1, D2	LEDs

Additional items

SW1	DPDT
J1	3.5 mm jack socket
PL	Jack plug to suit receiver
	Aluminium box, approx. 11 by 6 by 2.5 cm

Source

Components are available from Maplin.

62 An electronic die

Introduction

Throwing a die is a venerable way of generating a 'random' number between 1 and 6. A die is easy to lose, so here is an electronic die-throwing circuit which brings the technique up to date and serves as another application of logic circuits. If you need a reminder of the basics of logic, refer to *Digital logic circuits*.

Warning

This circuit uses members of the integrated circuit family known as CMOS (complementary metal-oxide semiconductor). These use very little current and can be completed destroyed if they come into contact with the

magnitudes of static electricity that most of us carry about when we walk on carpets and wear rubber shoes. You will never know if this wanton destruction has happened – all you will discover is that your circuit doesn't work and that you have tested *everything*. To avoid this problem do the following things:

1. **Before you open the little packet in which the IC is supplied, touch something which you *know* to be earthed – the metalwork of any equipment which is mains-earthed, for example. Then open the packet.**
2. **Let the IC fall gently on the bench – don't pick it out with your fingers. Touch your earthed metalwork again. Pick up the IC and insert it gently into its holder.**

The circuit is safe from damage while it is connected to the battery.

Description

The circuit is shown in **Figure 1**. IC1a and IC1b form an *astable multivibrator*, similar to that used in 'An LED Flasher' also in this book. It runs constantly, and its output is fed to IC2 via a single NAND gate, IC1c. The other input of IC1c is held at 0 V by R3 and C3. Whenever one input of a NAND gate is logic 0 (or 0 V), there can be *no output* from the gate, irrespective of what is happening at other inputs. So, despite the fact that the oscillator (or *clock*) is running all the time, its square-wave output never

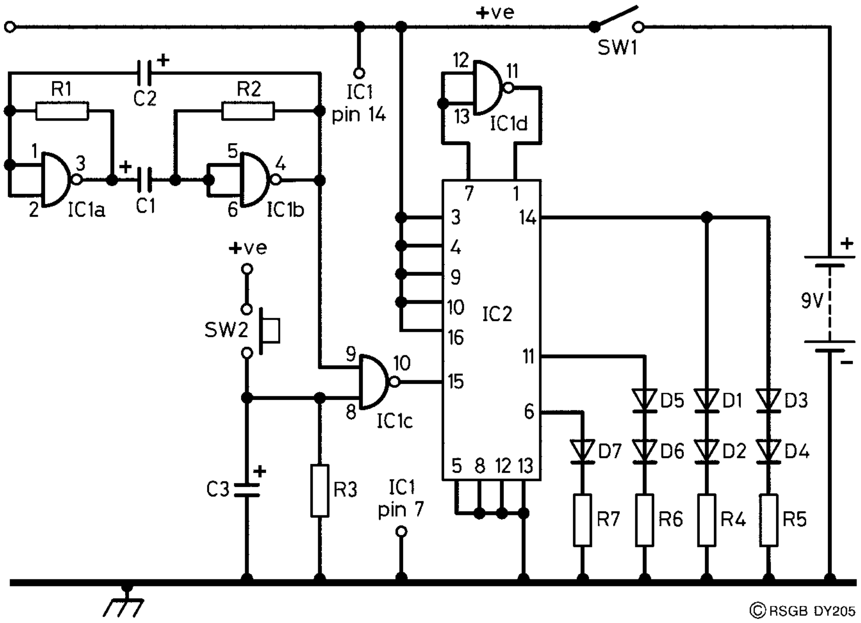
+ve
SW1
R1
C2
R2
IC1 pin 14
12
11
13
IC1d
1
3
2
IC1a
C1
5
4
6
IC1b
7
1
3
14
4
9
10
IC2
16
11
15
6
5 8 12 13
+ve
SW2
9
10
8
IC1c
C3
R3
IC1 pin 7
D5
D1
D3
D7
D6
D2
D4
R7
R6
R4
R5
9V
©RSGB DY205

Figure 1 Electronic dice, circuit diagram

reaches IC2 until the 'other' input of IC1c is brought away from logic 0. This happens when SW2 is pressed. Pressing SW2 momentarily, as is normally done, charges up C3 to 9 V, putting a logic 1 on the second input of IC1c and allowing the clock signal through to IC2. As the switch is released, C3 begins to discharge through R3, gradually lowering the voltage on pin 8 of IC1c. As this voltage crosses 4.5 V (half the supply voltage), IC1c then treats this input as logic 0, which cuts off the clock signal from IC2 again. Thus, when SW2 is momentarily pressed, the clock signal is fed to IC2 for a short interval of time, before being blocked again.

IC2 is a binary counter, and the outputs that concern us are from pins 14, 11 and 6. Pin 14 is the *most significant bit* and pin 6 the *least significant bit*. See the panel for an explanation of what happens here.

To understand these outputs, you should be able to count in binary, using three bits. At the start of the counting process, all the bits have zero values, i.e. 000. The left-most bit is called the *most significant bit*, and the right-most bit is called the *least significant bit*. (When we write numbers normally, a number 1 in the left-hand position represents one hundred, 1 in the middle column represents ten, while the number 1 in the right-hand position means one. One hundred is a more significant number than one, hence the nomenclature.)

The three bits of our binary number have values, from left to right, of 4 ($=2^2$), 2 ($=2^1$) and 1 ($=2^0$). This means that a number 1 in the left-hand column signifies the normal number four, in the middle column it would represent two, and in the right-hand column, it would represent one. This should help you understand the patterns of bits which emerge as the clock waveform is counted, as the next paragraph explains.

As each cycle of the clock enters IC2, it increments its internal counter and the values of that counter are shown by the states of pins 14, 11 and 6. After the first pulse, these three pins would have states corresponding to 001; after the second, 010; after the third, 011; after the fourth, 100; after the fifth, 101; after the sixth, 110. This sequence of 3-bit numbers represents a binary count from 1 to 6 in 'normal' parlance. These six states are used to illuminate the conventional pattern of dots on a die, using LEDs.

As the clock cycles are counted, the LEDs flicker as the die is 'rolled'. Resistors R4 to R7 are used to limit the current through the LEDs. When the counter stops, it can be in any of the positions shown in **Table 1**. Because of

Table 1 Dice number relative to counter output

Step	*Pins*			*Dice number*
	14	*11*	*6*	
1	0	0	1	4
2	0	1	0	2
3	0	1	1	6
4	1	0	0	1
5	1	0	1	5
6	1	1	0	3

the way in which the dots are grouped on the faces of a die, the wiring of the LEDs is simpler than it might otherwise be. You can see from Figure 1 that there are really only three sets of connections to the LEDs – one from each bit of the counter output. When pin 14 is at logic 1, four LEDs are lit, corresponding to the number four. For the number five, pins 14 and 6 will be at logic 1. For the number one, only pin 6 is at logic 1. For three, pins 11 and 6 are at logic 1, for six, pins 14 and 11 are at logic 1, and for two, only pin 11 is at logic 1. These conditions are summarised in Table 1.

Construction

Two pieces of Veroboard (of the strip type) are needed – one for the main circuit (**Figure 2**) and the other for the display LEDs (**Figure 3**). The display board is easier to build, so we will do that first. It measures 20 holes by 18 strips.

Cut the tracks as shown in Figure 3, using a 3 mm (1⁄inch) twist drill rotated between thumb and forefinger. Insert and solder the Veropins, resistors and the wire links. Then solder the LEDs in place, making sure their polarities are correct. The LEDs are mounted proud of the board (**Figure 4**, which will help when you come to fix the board into a case.

The main board measures 32 holes by 18 strips. As before, remove the tracks in the places shown in Figure 2. Note that track ‘L’ is *not* broken under the position for IC2. Solder in the IC holders and the links. Solder in R1 and R2 vertically, and R3 horizontally. Then solder in C1, C2 and C3; these all being electrolytics, check their polarities. Solder in the Veropins and prepare the connections to those components not on the main board, using stranded, insulated wire. Check the circuit carefully, and read the handling precautions given earlier before carefully inserting the two ICs.

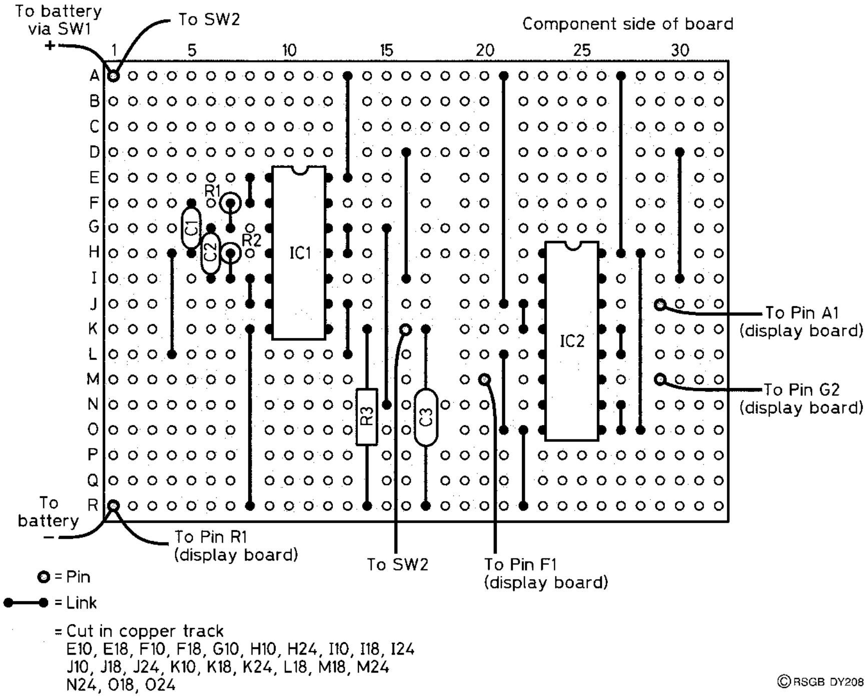
To battery via SW1
+
To SW2
Component side of board
1
5
10
15
20
25
30
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
R1
R2
C1
C2
IC1
IC2
R3
C3
To Pin A1 (display board)
To Pin G2 (display board)
To battery
−
To Pin R1 (display board)
To SW2
To Pin F1 (display board)
= Pin
= Link
= Cut in copper track
E10, E18, F10, F18, G10, H10, H24, I10, I18, I24
J10, J18, J24, K10, K18, K24, L18, M18, M24
N24, O18, O24
©RSGB DY208

Figure 2 Main board, component layout

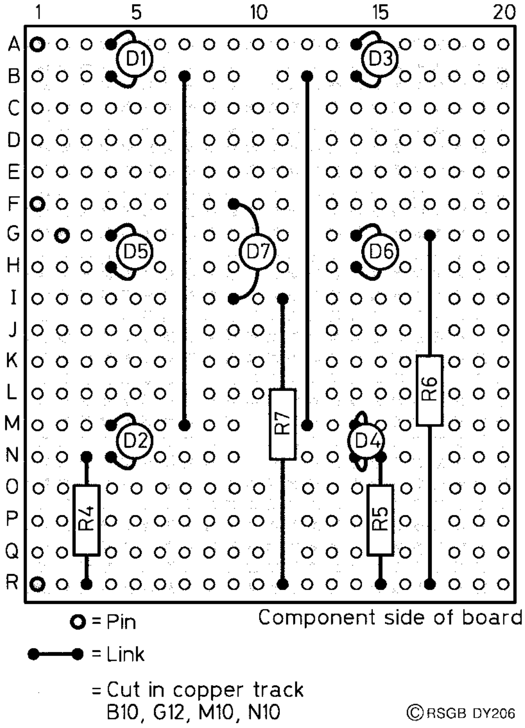
1
5
10
15
20
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
D1
D2
D3
D4
D5
D6
D7
R4
R5
R6
R7
= Pin
= Link
Component side of board
= Cut in copper track
B10, G12, M10, N10
©RSGB DY206

Figure 3 Board for the six LEDs

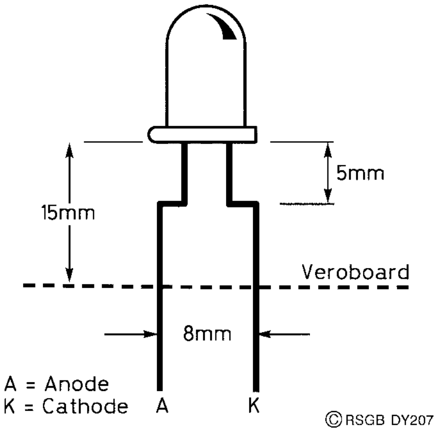

Figure 4 D7 lead bending

Testing and use

When the circuit is switched on, it is likely that the centre LED only will be lit. Press SW2, the 'throw' switch. All the LEDs should flicker, even after the switch is released, but only for a second or so. Then a standard pattern should show, signifying numbers between one and six. If this does not happen, switch off and check your circuit again. Are there any missing links? Are there any obvious dry joints (usually dull instead of shiny)? Can you see any solder bridges between tracks? Did you choose to ignore the handling precautions? If none of these results in a working circuit, try some fault-finding.

- Disconnect the lead to D7 on the main board (marked 'to pin F1' on Figure 2). Touch the lead on to the positive supply rail. D7 should light. Now transfer this lead on to pin 4 of IC1. If it flickers, then the oscillator is running (which it should).
- Transfer this lead to pin 10 of IC2. D7 should flicker when SW2 is pressed, but should be on permanently when it is released. This means that the correct signals are reaching IC2 from IC1.
- If one of the chains of LEDs (i.e. D1 and D2, D3 and D4, D5 and D6) does not light, you may have one *or both* LEDs the wrong way round.

When the die is working, you may care to experiment with the values of R1 and R2, but you should keep their values the same, i.e. R1 = R2. The larger the values of these, the more slowly the die will appear to 'roll'.

Finishing touches

On completion of the project, you will want to mount it in a smart case; any plastic box is suitable for this, with the display board mounted so that the

LEDs protrude through the top and are fixed to it using LED clips. The way in which you mount both boards to the case is entirely up to you.

Parts list

Resistors: all 0.25 watt, 10% tolerance, or better

R1, R2	5600 ohms (Ω)
R3	15 kilohms (kΩ)
R4–R7	470 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG
C3	68 microfarads (μF) electrolytic, 16 V WKG

Semiconductors

IC1	4011
IC2	4029
D1–D7	LEDs, any size and colour

Additional items

SW1	On-off switch SPST toggle
SW1	Push-to-make, non-latching
	Veroboard, 2 pieces, see text for sizes
	Veropins, 10
	Stranded insulated wire for general wiring
	Single-core insulated wire for links
	PP3 battery clip
	PP3 battery
	Plastic box to suit
	LED mounting clips (7)
	Means of mounting boards to box

Source

Components are available from Maplin.

63 The absorption wavemeter

Introduction

The purpose of the absorption wavemeter is to check that a transmitter is radiating within the correct waveband, and to detect any spurious *harmonic* emissions. A harmonic frequency is an integral multiple of the carrier frequency – e.g. if the carrier is at a frequency *f*, the harmonics are at frequencies 2*f*, 3*f*, 4*f* . . . and so on.

It can also be employed as a relative field-strength indicator, being used in experimentation with aerials. What it *cannot* do is to measure the transmitted frequency accurately; all it can do is to place the signal within a particular band, say the 7 MHz or the 18 MHz band (or their harmonics).

How it works

You may recognise the circuit of **Figure 1** as a crystal set with a meter replacing the headphones. It is simply the parallel combination of an inductance with a variable capacitance, which constitutes a resonant circuit with a variable frequency. A dial on the variable capacitor can be roughly calibrated with frequency. The fixed capacitor across the meter helps to improve the meter reading by making it read the peak value of the carrier wave rather than the average value.

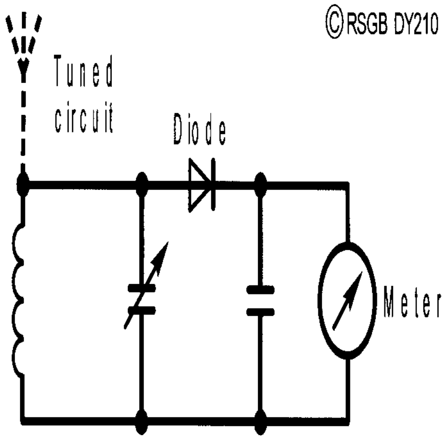

Figure 1 Absorption wavemeter circuit

When the wavemeter is situated in the vicinity of a relatively strong radio-frequency (RF) field, the inductor absorbs a small amount of energy from that field. If the circuit is resonant at the same frequency as that of the transmitter, an RF voltage is produced across the coil and is proportional to the strength of the RF field. The RF voltage is *rectified*, or *detected* (turned from AC into DC) by the diode and the second capacitor, and is displayed by the meter.

The tuned circuit

The formula for calculating the resonant frequency of a coil and capacitor is relatively simple. It is

$$f = \frac{1}{2\pi\sqrt{LC}},$$

where f is the resonant frequency (Hz),
L is the inductance (H), and
C is the capacitance (F).

If this formula puts you off, there is this slightly easier version:

$$f = \frac{159}{\sqrt{LC}},$$

where f is the frequency in **MHz**,
L is the inductance in **microhenries** (**μH**), and
C is the capacitance in **picofarads** (**pF**).

When a variable capacitor is specified, it is usual to quote its maximum capacitance, i.e. the value when the vanes are fully meshed. When the value of the capacitor is small, it is usual to quote its minimum capacitance also, the value when the vanes are fully open. If you use a variable capacitor of 10–100 pF with a 100 μH inductor, you will use the value of 10 pF to calculate the maximum resonant frequency, and the value of 100 pF to calculate the minimum resonant frequency, giving a tuning range of 1.6 MHz to 5 MHz. Check it yourself!

64 An HF absorption wavemeter

Introduction

An absorption wavemeter is basically a simple tunable detector circuit, such as would be used in a crystal set, but with the headphones replaced by a meter, in order to indicate the strength of the received signal.

The circuit

Figure 1 shows the circuit of the wavemeter. It is designed to cover all the HF bands from 1.8 MHz to 28 MHz in four switchable ranges. There is no built-in method of amplification, so a sensitive meter is needed in order to indicate sufficiently using the available absorbed energy. The meter does not need calibrating, it serves to produce only an indication of the absorbed power, *not* its absolute value.

An absorption wavemeter does not have outstanding selectivity, mainly because of the loading effect of the meter. This problem is ameliorated here (but only slightly) by the inclusion of R1 in series with the meter. It gives

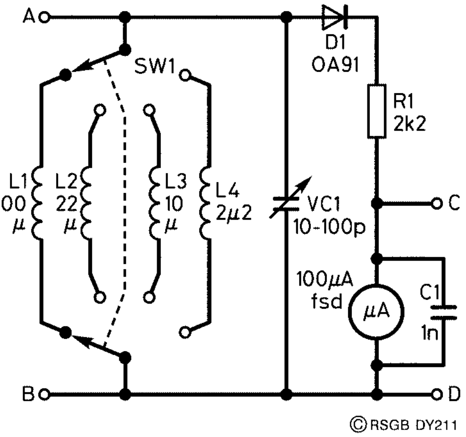

Figure 1 HF absorption wavemeter, circuit diagram

improved selectivity at the expense of a slight loss of sensitivity. The appearance of capacitor C1 across the meter is discussed in *The Absorption Wavemeter*, in the previous section.

Coils

These can be purchased (see the parts list) or hand-wound on short lengths of PVC water pipe or conduit, and the coil-winding details are given in **Table 1**. In theory, the 100 pF tuning capacitor would cover all the HF bands with only three coils, omitting the 22 μH coil. Sample calculations are explained in the previous article. However, its inclusion avoids the common problem of having some bands cramped at the extreme edges of the tuning scales.

Table 1 Close-wound coil values

L (μH)	No of turns	Span (mm)	SWG
PVC FORMER OF OUTSIDE DIAMETER 25 mm			
2.2	9	9	20
10	24	24	20
22	44	44	20
100	110	60	26
PVC FORMER OF OUTSIDE DIAMETER 20 mm			
2.2	10	10	20
2.2	9	5	26
10	32	32	20
10	24	12	26
22	63	62	20
22	41	21	26
100	134	70	26
PVC FORMER OF OUTSIDE DIAMETER 19 mm			
2.2	11	11	20
2.2	9	5	26
10	35	34	20
10	25	13	26
22	68	67	20
22	44	22	26
100	157	79	26

Construction

Hand-wound coils, using 19–25 mm PVC formers, have single-layer windings, with their ends secured by threading the wires through two small holes drilled in the formers at each end of each winding.

You may have wondered about the use of the two-pole switch SW1, as all the lower ends of the coils are grounded anyway. The answer lies in the ease of construction. Each coil can be mounted on the tags of the switch for support. If you are using the commercial coils, they are small enough to be accommodated on a miniature rotary switch; a larger rotary switch is required for hand-wound coils. The specification of a 6-way switch allows for the addition of extra frequency ranges at little extra cost. Rotary switches such as the 2-pole, 6-way variety have an adjustable end-stop which allows the number of ways to be set anywhere between 2-ways and 6-ways. Remove the nuts from the switch, and a ‘washer’ will fall out. It is not a washer, really, and has a little piece of metal bent over at right angles to the washer. This fits into one of 12 holes in the body of the switch, and prevents the shaft from turning through 360° and selecting the number of ways.

Calibration

This process needs a commercial amateur-bands transceiver, if it is to be done quickly and accurately.

- Connect the transceiver to a dummy load, and connect a piece of flexible wire from point A in Figure 1 and wrap two or three turns around the cable between the transceiver and the dummy load.
- Turn the transceiver power level control to minimum, and set the frequency at 1.81 MHz and the mode to FM or AM. Switch to transmit.
- With the wavemeter set to its lowest frequency range (i.e. 1.8–3.5 MHz), rotate the tuning capacitor until maximum deflection is obtained from the meter needle. Write the frequency on the inner ring of the wavemeter scale at that point.
- Switch off and retune the transmitter to 3.5 MHz. Repeat the above process.
- Repeat, using switch range 2 for 3.5, 7 and 10 MHz, range 3 for 7, 10 and 14 MHz, and range 4 for 14, 18, 21, 24 and 28 MHz.

An alternative source of calibration signal could be an HF signal generator, with a single-turn loop of wire at the remote end of its cable, and the wire from the wavemeter brought close to the loop.

Remember that the marks on the scale represent frequency *bands*, not precise frequencies.

Extending the range

If you would like to experiment with increasing the range to the lower VHF band, then 50 MHz should be achievable with an additional self-supporting coil (no former). Try three turns of 20 SWG enamelled copper wire of 25 mm ID (internal diameter). Wind it on some 25 mm PVC pipe, as before, then slide out the pipe! The same result could be achieved with four turns of 19 or 20 mm ID, or 12 turns of 7.1 mm ID.

Parts list

Resistor	
R1	2200 ohms (Ω) 0.25 watt, 10% tolerance
Capacitors	
C1	1 nanofarad (nF) min. ceramic
VC1	100 picofarad (pF) variable
Inductors	
L1	100 microhenries (μH)
L2	22 microhenries (μH)
L3	10 microhenries (μH)
L4	2.2 microhenries (μH)
Semiconductors	
D1	Germanium OA91 or 1N4148
Additional items	
SW1	Rotary 2-pole, 6-way
	Moving-coil meter, 50 or 100 μA FSD (any meter of this sensitivity will do)
	Plastic box, approx. 150 by 80 by 50 mm
	Connecting wire, coloured

65 A vertical aerial for 70 cm

Introduction

If the range of your hand-held transceiver is very limited, what you need is a vertical aerial mounted outside. This design is a half-wave dipole fed at the end instead of the middle. Because the impedance of the dipole is high at its ends (and low in the centre), a matching circuit is needed so that this high impedance can be matched to that of the low-impedance coaxial cable. All that is needed is a coil, which increases the *electrical length* of the aerial to ⅝-wavelength. (Note that the *electrical* length (i.e. the length as it appears to an RF signal) is not necessarily the same as the *actual* length.)

Construction

The aerial element and the coil are made from a single piece of 1.5 mm welding (brazing) rod, and the dimensions are given in **Figure 1**. Wind the coil around a 4 mm rod or the shank of a twist drill. The lower end is filed to a point and then soldered into the centre conductor of a 4-hole panel-mounting BNC socket. (Try to obtain a good-quality BNC socket with PTFE insulation – the insulation of cheaper sockets is easily damaged.) Trim the element to 427 mm (top of element to top of coil) *after* the wire has been soldered to the socket.

The base coil causes the aerial to be rather ‘whippy’, so a piece of 5 mm plastic knitting needle can be cut to the length of the coil and then forced into it.

The radials are made from four lengths of 3 mm welding rod. These are bent and soldered into the four mounting holes of the socket, and then cut to the lengths shown in Figure 1.

Testing

Using a standing-wave-ratio (SWR) meter connected between your aerial and the transceiver, measure the SWR (or obtain some indication of the reflected power) on transmit. If it is greater than 1.5, switch off the transmitter, trim about 3 mm off the end of the element, and try again. Repeat the process until the measured SWR (or the reflected power) is as low as you can get it.

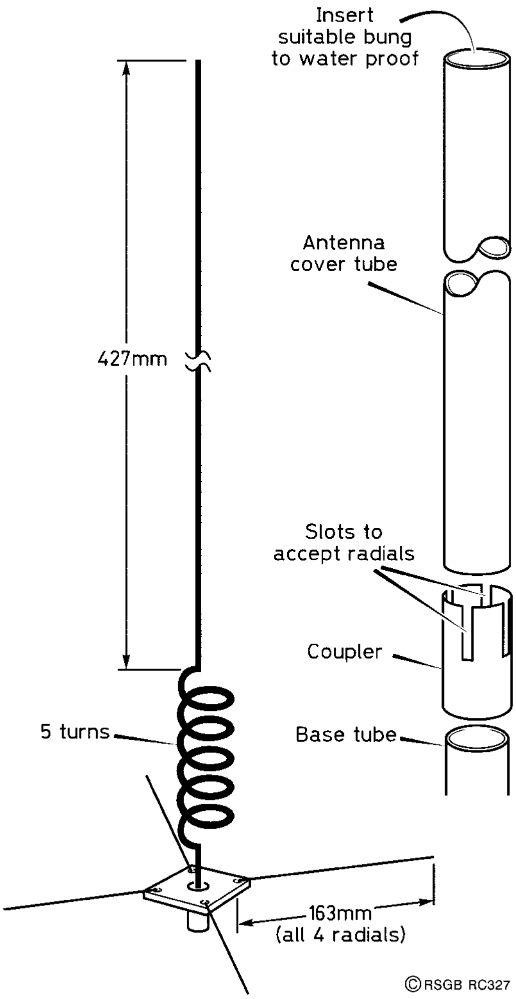
Insert suitable bung to water proof
Antenna cover tube
Slots to accept radials
Coupler
Base tube
427mm
5 turns
163mm (all 4 radials)
©RSGB RC327

Figure 1 Construction of the 70 cm antenna

Installation

The aerial is made waterproof by enclosing it in a 22 mm diameter PVC waste water pipe. It is 'weldable', and available at plumbers' merchants, usually by the metre. A coupler is slotted to take the radials (Figure 1). File the BNC socket as required, so that it slides inside the coupler until the radials poke out of the slots. Cut a length of plastic tubing which is 30 mm longer than the aerial, and push it into the coupler. You will then need a plastic bung or screw-top to waterproof the top end.

Plastic welding solution is now applied to the joints, sealing the aerial inside the tube. The coupler has been weakened as a result of making the slots, so it is worthwhile applying PVC tape around this joint until the welding solution sets.

A support for the aerial is made from an off-cut of tube pushed into the lower end of the coupler and held with a self-tapping screw.

Parts list

1	4-hole panel-mounting BNC socket
1 m	3 mm brazing rod (may be available from a small garage)
1 m	1.5 mm brazing rod
1 m	22 mm PVC waste water pipe
1	22 mm straight coupler
	Plastic welding solution

66 A UHF corner reflector aerial

Introduction

The corner reflector is a well-known design and is capable of good performance on the VHF and UHF bands. At UHF, the practical implementation of the corner reflector is an ideal constructional project.

Some details

Expressed quite simply, the aerial consists of a $\frac{\lambda}{2}$ dipole (where λ is the standard symbol for wavelength, making a '$\frac{\lambda}{2}$ dipole' a half-wave dipole).

Nothing new in that, you might say. However, the interesting feature is the reflector, which is not the usual single element, but a 90° metal 'corner', acting rather like a parabolic dish as used for satellite signal reception. The wind resistance of this type of reflector makes it impractical so, to reduce the 'windage' quite significantly, we make the 'corner' from closely spaced rods, as illustrated in **Figure 1**.

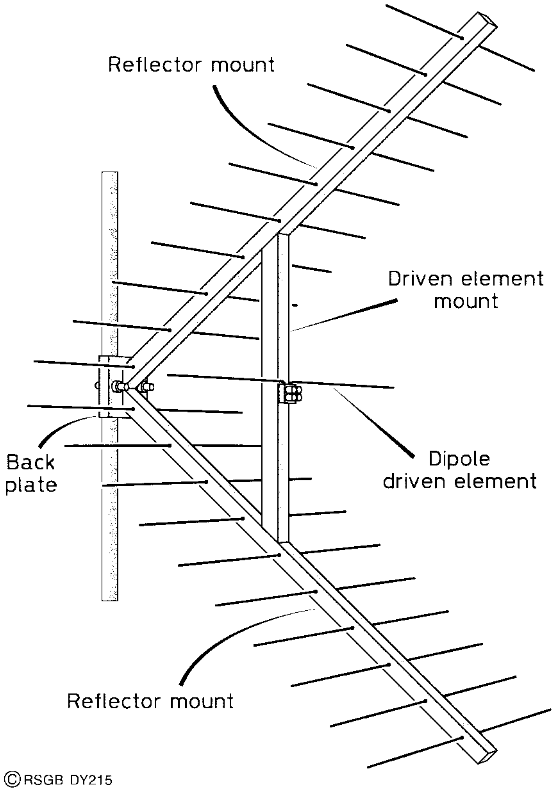

Figure 1 70 cm corner reflector antenna

The reflector consists of a number of 0.6λ rods, spaced from each other by 0.1λ. The aerial frame can be made of metal or wood, but wood is easier to work with, and mounting the elements to the frame is simpler. The prototype was made with wood of 20 mm by 15 mm cross-section, as **Figure 2** shows. The wood was varnished for protection. The elements were made from 1.5 mm diameter copper wire, because a large reel of the wire happened to be available. The wire diameter is not critical; tubing could be used just as successfully. 14 SWG hard-drawn copper aerial wire would be even better than that used in the prototype.

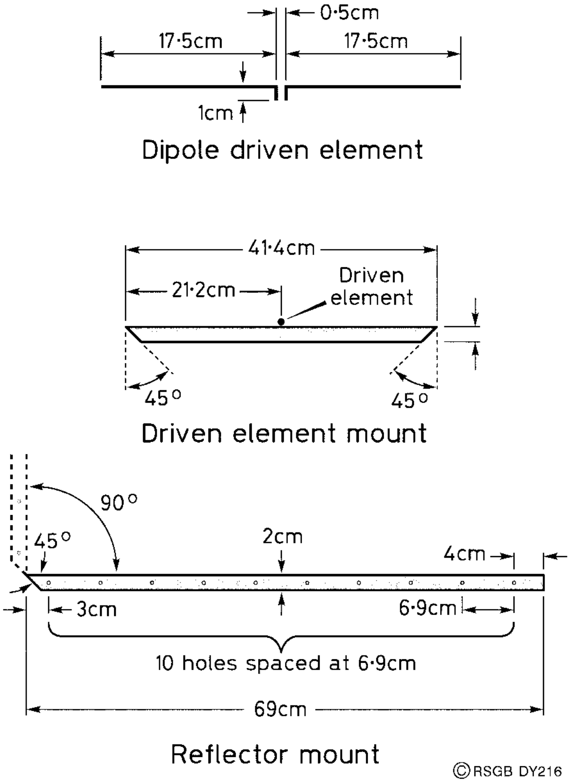
0·5cm
17·5cm
17·5cm
1cm
Dipole driven element
41·4cm
21·2cm
Driven element
45°
45°
Driven element mount
90°
45°
2cm
4cm
3cm
6·9cm
10 holes spaced at 6·9cm
69cm
Reflector mount
©RSGB DY216

Figure 2 Driven element dimensions, together with boom dimensions for driven element and reflectors

Construction

This project is just as much a woodworking project as a radio project! Follow the instructions carefully, and you should have little trouble.

- Cut the booms for the reflectors, as shown in Figure 2. A mitre block is invaluable here in producing the 45° corners.
- Using the dimensions given on the diagram, mark the hole positions for the reflector elements, and then drill holes of a size which holds the elements firmly.
- Cut the driven element boom according to the diagram, and mark the point midway along the longer side, which will assist you later in positioning the driven element.
- Cut the back plate to size (about 120 mm by 80 mm). You may need to alter this size depending on the size of the U-bolt you will be using to clamp the aerial to the mast.

- If you want to be extra cautious in your construction, use the belt-and-braces approach, commonly known as 'screw-and-glue' to fix the booms to each other and to the back plate.
- Fix the reflector booms to the back plate first, then slide in the driven-element boom until it will go no further, then apply the wood glue and screw the two ends tightly to the reflector booms. Leave for the period prescribed by the glue manufacturers for the glue to harden.
- Varnish the whole structure.
- Cut the driven element to the correct size plus a couple of centimetres (the reason for this will be evident in the *Testing* section), and fix it to the centre of its boom (at the position you marked earlier) with a 'chocolate block' connector to which the coaxial feeder cable will eventually be connected.
- Cut and fix the reflector elements in place. If you find that these are a loose fit in the holes then, *for each element*, drill a pilot hole through the boom to intersect the hole for the element. File off the point of a woodscrew, and screw it gently into the pilot hole until it meets the element and grips it in place. You will now see why the point was filed off! Alternatively, you can glue the elements in place.

Testing

Place the aerial on a mast, clear of obstructions. Connect it to a transceiver with a length of coaxial cable, with a standing-wave-ratio (SWR) meter in circuit. Find a clear frequency, identify your transmission and ask if the frequency *really is* clear. If so, key the transmitter again and note the SWR. **Do not stand in front of any aerial when it is radiating!** The length of the driven element must be adjusted to obtain an SWR of less than 2. If you have to shorten the dipole, bend the ends over rather than cut them off. That way, if you go too far, your can lengthen them again! The dipole was initially cut too long intentionally, to allow for adjustment here! Bending the ends over also reduces the risk of physical damage to clothing, skin and eyes. You may like to consider applying the same technique to the reflector elements for that reason alone.

Moving on . . .

Once you have warmed to the idea of the corner reflector as an aerial, you might like to ring the changes regarding the reflector. How about replacing the 20 reflector elements with a wire mesh, such as garden centres sell as 'chicken wire'? Choose the finest mesh if there is a choice. Some extra support may be needed around the edges of the mesh, but you could go on to make a comparison of aerial gain between the two types, using the *UHF Field Strength Meter* described elsewhere in this book.

Materials

Stiff wire or thin-walled tubing for dipole and reflector
Frame – wood, 15 mm by 20 mm cross-section, lengths given in text
Back plate – stout plywood, dimensions given in text
U-bolt to suit mast
Wood screws
Wood glue
50 W coaxial cable for feeder
2-terminal ‘chocolate block’ for dipole connection to feeder
Varnish

67 A switched dummy load

Introduction

A dummy load is a pure resistor of value 50 ohms which can replace your transmitting aerial and enable you to operate the transmitter for test purposes without radiating a signal. It sounds simple enough, but there are two main problems. Firstly, it is impossible to delve into your junk box and emerge with a resistor that will dissipate 100 W PEP and still retain its marked resistance value. Secondly, a ‘pure resistance’ is very difficult to achieve. A pure resistance is a device which has resistance but no reactance. All common resistors have significant reactance at radio frequencies, particularly the wire-wound varieties, which have a helical (i.e. wound like a coil) construction. This is particularly annoying, because wire-wound construction is normally used for large-wattage resistors.

Although all resistors have *some* reactance, not all are quite as bad as the wire-wound type. Carbon film resistors are made by depositing a thin film of carbon on the surface of a small, hollow ceramic cylinder, the thickness of the film of carbon determining the value of the resistor. Provided the lead lengths are kept short, these resistors have a tolerably small reactance, and will be used in this project.

Bearing the load

A 2 W carbon film resistor is hardly going to withstand our 100 W PEP of SSB, so it is obvious that the design of our dummy load must be a little more complex than a single resistor and a switch, despite what **Figure 1** might suggest! In fact, it uses 20 resistors, each of value 1000 Ω (1 kΩ). How does this solve our problem?

Perhaps a little theory is in order here, but no more than is required by the Radio Amateurs' Examination.

When two *equal* resistors of value *r* are combined in parallel (i.e. side by side), the total resistance, R_T, is given by:

$$\frac{1}{R_T} = \frac{1}{r} + \frac{1}{r}.$$

Adding 1/*r* to 1/*r* gives 2/*r*, therefore:

$$\frac{1}{R_T} = \frac{2}{r}, \quad \text{i.e. } R_T = r/2.$$

So, by connecting *two* equal resistors in parallel, we get a combined resistance which is *half* the individual resistances. If we combine *three* in parallel, we get a *third* of the resistance, and so on.

Here, we are connecting *20* resistors of 1 kΩ in parallel, so we will produce an overall resistance of *one-twentieth* of the individual resistance, i.e. 1000/20 = 50 Ω, which is what we set out to achieve!

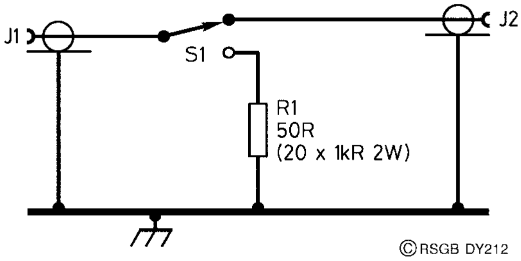

Figure 1 Basic circuit diagram

This is not the only advantage, though. Each resistor is capable of dissipating 2 W, so 20 of them will safely dissipate 40 W, for short periods at least. This power dissipation is approximately the same as 100 W PEP of normal speech, so the design should be capable of use in an 'average' amateur station. The power-handling ability of any dummy load can be improved by providing a 'heat sink' which helps to conduct the heat energy away from the resistors, thus lowering their temperature. One popular heat sink is a can of transformer oil, into which the resistors are immersed. This design uses a rather more mundane heat sink, but which is adequate for the job in hand.

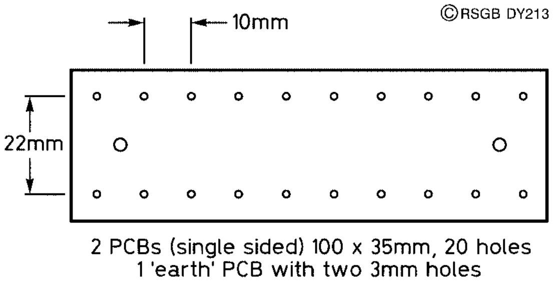
10mm
©RSGB DY213
22mm
2 PCBs (single sided) 100 x 35mm, 20 holes
1 'earth' PCB with two 3mm holes

Figure 2 PCB resistor supports

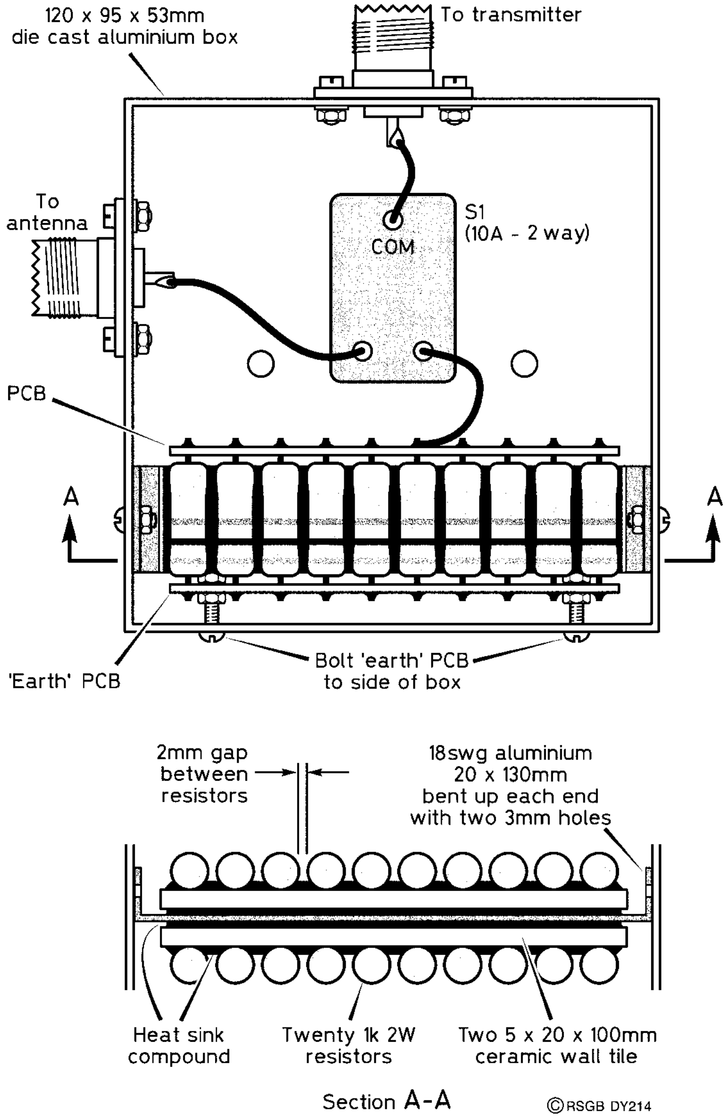
120 x 95 x 53mm
die cast aluminium box
To transmitter
To
antenna
S1
(10A - 2 way)
COM
PCB
A
A
'Earth' PCB
Bolt 'earth' PCB
to side of box
2mm gap
between
resistors
18swg aluminium
20 x 130mm
bent up each end
with two 3mm holes
Heat sink
compound
Twenty 1k 2W
resistors
Two 5 x 20 x 100mm
ceramic wall tile
Section A-A
©RSGB DY214

Figure 3 Switched dummy load, component layout

Pouring the heat away

Two basic PCBs are needed, the details being given in **Figure 2**. Both measure 11 × 35 mm. Each has a set of 20 holes of diameter about 1 mm to take the ends of the resistors. Only one of them has two 3 mm holes which are used for mounting the completed load. **Figure 3** shows how the heat sink is assembled. Solder one end of each resistor into the 'earth' PCB, and then mount this to the box using the two bolts as shown in the diagram. Between each rank of resistors is a 'sandwich' consisting of an aluminium strip and two pieces of ceramic wall tile, to act as a heat sink for the resistors. Heat sink compound is used to provide good thermal contact between the resistors and the tiles, and between the tiles and the aluminium strip. This is shown in the lower part of Figure 3. Thread the loose wires of the resistors through the holes in the second PCB, solder into place, and crop the protruding wires.

Switching

A changeover switch must be used so that the transceiver can be switched between the dummy load and the aerial without unscrewing connectors. For most purposes, an ordinary 10 A 230 V changeover switch will suffice, as found in many electrical shops and DIY stores.

Wire this into the circuit as shown in Figure 1 and Figure 3. Check your wiring. Put S1 in the 'dummy load' position. If you have a multimeter which includes an ohmmeter, measure the resistance across the socket, J1, before connecting it to the transceiver. It should be very near 50 Ω. With S1 in the 'aerial' position, there should be an infinite resistance across J1. Move your ohmmeter to read the resistance across J2. It should be infinite for both positions of S1. If all seems correct, close the box and your dummy load is ready for use!

Parts list

Resistors
- 1000 Ω (1 kΩ) 2 W carbon film, 20 required

Additional items
- 10 A 230 V changeover (SPDT) switch
- SO–239 sockets, 2 required
- Aluminium box 120 × 95 × 53 mm or similar
- Ceramic wall tile cut as required
- Aluminium strip, 18 SWG
- Nuts and bolts as required

68 A simple Morse oscillator

Introduction

This is an excellent project which uses the ‘junk box’ as its source of components. If you have trouble in finding the bits for this one, a good source of the components for this and many other similar projects is to be found with the parts list at the end of the project.

The circuit

This is shown in **Figure 1**, and uses an ‘unknown’ Plessey chip, which makes the overall circuit extremely easy to build. A 0.1 μF capacitor is connected between pins 7 and 8, a speaker (in the popular 8 to 25 Ω impedance range) is connected between pins 8 and 9. If a 9 V battery is connected with its positive terminal to pin 8 and its negative terminal to pin 1, 3 or 5, a tone will be produced in the speaker.

To make this circuit into a good Morse practice oscillator, it is necessary only to insert a Morse key into the supply rail from the battery.

However, there is another refinement which you may care to build into the circuit. The tone from the loudspeaker is different, depending upon which of pins 1, 3 or 5 you use. In the prototype, a single-pole changeover switch was used to select the tones from pin 1 or pin 3, and the Morse key would be connected to the circuit via a small jack socket. There is no need for an on/off switch, as the Morse key performs that function. The switch and the jack socket can be seen in the photograph.

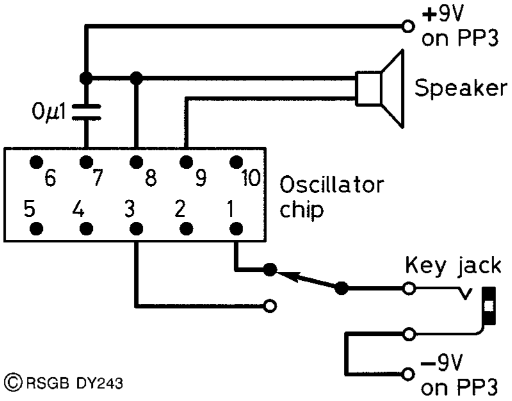

Figure 1 Morse oscillator, circuit diagram

The case

Almost any small loudspeaker will do; the higher the impedance the better. The 'impedance' is the figure usually printed on the rear of the speaker magnet. It is a value given in ohms (Ω). For simple circuits like this, you will usually find that the higher the impedance (within reason), the louder the sound it will produce. Speakers from old transistor radios will work, although their impedances can be rather low sometimes.

Any case big enough to house the components and the battery will be suitable. The prototype used a 'Walkman'-type speaker and case, and is shown in the photograph.

To use the circuit, simply attach a 3.5 mm jack plug to your Morse key, and insert it into the socket. Nothing should happen until you press the key, when a tone should be heard from the loudspeaker.

Another application

Try soldering the two wires of a 'twisted pair' to the 3.5 mm jack plug. Touching the wires together produces a tone from the speaker. This simple circuit can then be used as a 'continuity tester'. Touch the two wires to the ends of a fuse. If there is no sound, then the fuse is blown. There are many other such tests you could perform with this device – to check whether there really *is* a connection between one end of a wire and the other, for example.

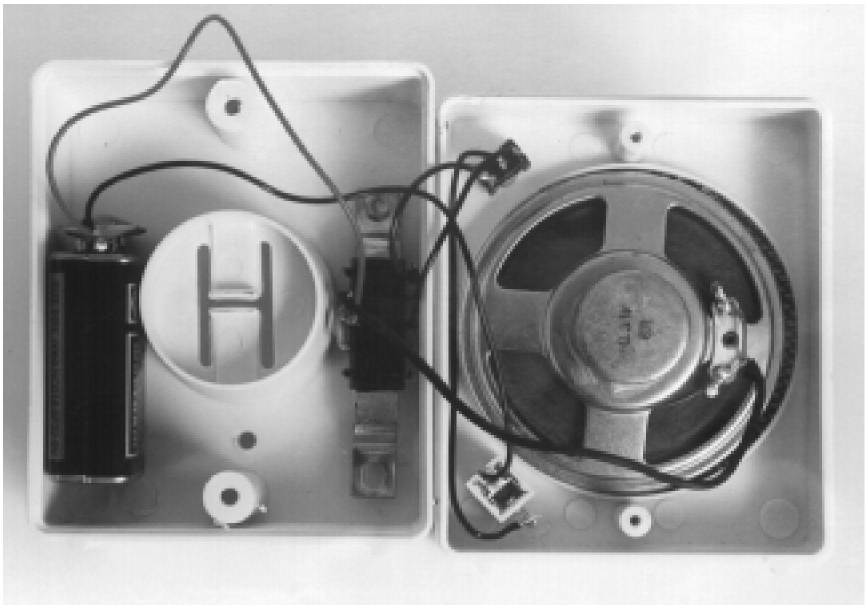

Warning

You must **never** make such tests on any equipment which is connected to the mains supply, even if it is switched off. If you want to make any such tests, make sure you are supervised by someone who understands what you are doing and is competent to advise and supervise you.

Parts list

Plessey oscillator chip
0.1 μF capacitor (disc ceramic)
Loudspeaker 8 to 25 Ω
3.5 mm jack socket (and 3.5 mm jack plug if needed)
Single-pole changeover switch

The Plessey oscillator chip is available from J. Birkett, 25 The Strait, Lincoln LN2 1JF, tel: 01522 520 767.

John Birkett may also provide a kit of parts (the chip, capacitor and loudspeaker).

69 A bipolar transistor tester

Introduction

This is a circuit which will test normal transistors, i.e. npn or pnp. It has the advantage of being able to test devices while they are still connected in their original circuits. However, when such tests are made, **the circuit containing the transistor under test must not be switched on**.

The circuit and how it works

The circuit runs from a 9 V battery such as a PP3 or six AA-type 1.5 V cells. Alkaline cells are to be preferred, as their electrolyte leakage properties are better. The circuit shown in **Figure 1** uses a single CMOS integrated circuit type 4001 or 4011. CMOS circuits require special handling precautions which are described in the project *Christmas Tree LEDs*, elsewhere in this book.

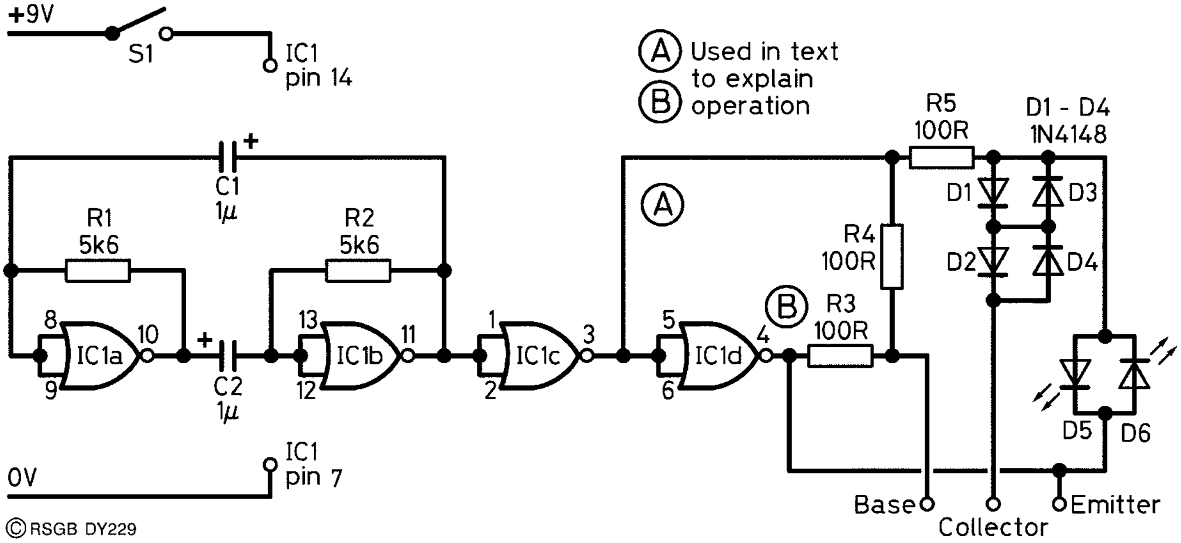
+9V
S1
IC1 pin 14
Ⓐ Used in text to explain
Ⓑ operation
R5 100R
D1 - D4 1N4148
C1 1µ
R1 5k6
R2 5k6
R4 100R
R3 100R
D1
D2
D3
D4
D5
D6
IC1a
IC1b
IC1c
IC1d
8
9
10
13
12
11
1
2
3
5
6
4
C2 1µ
IC1 pin 7
0V
Base
Collector
Emitter
©RSGB DY229

Figure 1 Transistor tester, circuit diagram

Inside the IC are four logic gates (see *Digital Logic Circuits*) which are all connected as inverters, which means that the output signal is always the logical 'opposite' of the input. The first two gates are connected as an oscillator; the circuit being the same as that used in *An Electronic Die*.

The output of the oscillator, at pin 11, is connected to the input of a buffer stage, IC1c, which helps to isolate the oscillator from the circuit that follows it. The buffer output appears on pin 3, which we shall label as *test point A* for future use. Another inverter, IC1d, follows this, its output at pin 4 being labelled *test point B*.

There are two LEDs connected back to back at the circuit output. These are D5 and D6, D5 being red and D6 being green. Whatever the output of the oscillator at any instant, one of the LEDs must be lit and the other unlit. With point A positive and point B zero, the red LED is lit, when A is zero and B is positive, the green LED is lit. Because the oscillator output is repeatedly switching from one polarity to the other, the lit LED is alternately red and green. They switch between the two colours much faster than we can see, so what we think we see are both LEDs lit together.

The two 100 ohm resistors, R3 and R4, provide the bias to the base of the transistor under test. When A is positive and B is zero, the base-emitter junction of the transistor will be forward-biased, and the transistor will switch on (if it is a working npn type). When the transistor is on, it effectively short-circuits D5 (the red LED) and it extinguishes. When point A is zero and B is positive, an npn transistor will be switched off and the green LED (D6) will light. Thus, for a working npn transistor, only the green LED is lit.

If the transistor under test is a pnp type, it will switch on when A is zero and B positive, thus short-circuiting the red LED (D6). When A is positive and B zero, the transistor is off and the red LED (D5) is lit. Thus for a working pnp transistor, only the red LED is lit.

To summarise, the states of the LEDs indicate the following conditions:

- Both LEDs apparently lit: no transistor connected, or transistor permanently open circuit.
- Neither LED lit: a collector-emitter short-circuit is almost certain.
- Red LED alone: pnp transistor in working order.
- Green LED alone: npn transistor in working order.

Construction

The prototype tester was built on a piece of Veroboard measuring 15 strips by 18 holes. Cut the tracks using a track-cutting tool or a 3 mm ($\frac{1}{8}$ inch) twist drill, as shown in **Figure 2**. Notice that, in this diagram, there is no

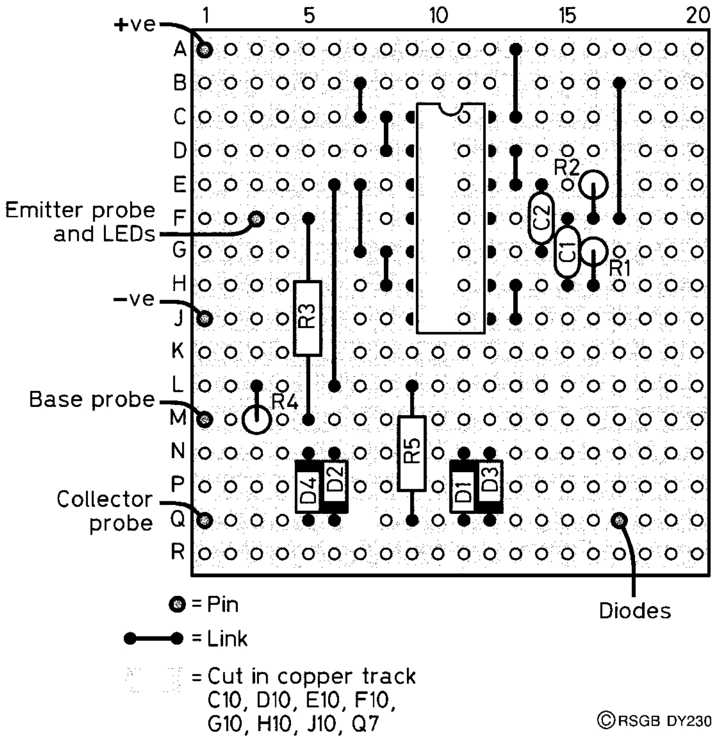
+ve
1
5
10
15
20
A
B
C
D
E
F
G
H
J
K
L
M
N
P
Q
R
Emitter probe and LEDs
−ve
Base probe
Collector probe
R2
C2
C1
R1
R3
R4
R5
D4
D2
D1
D3
Diodes
= Pin
= Link
= Cut in copper track C10, D10, E10, F10, G10, H10, J10, Q7
©RSGB DY230

Figure 2 Transistor tester, component layout

track ‘I’, so try to avoid miscounting when you translate diagram positions to real positions on the board. Hold the board up to a strong light to ensure that there is no copper swarf shorting adjacent tracks together, and that you have made the cuts in the correct places.

Having done this, insert and solder Veropins for all the connections to be made to components not on the board itself. Then solder in the IC holder and the wire links. Then insert the other components in the order resistors, capacitors and diodes D1–D4. Some resistors are mounted vertically so that their connections are on adjacent tracks. Double check the diode polarities – it is easy to make a mistake when wiring diodes in anti-parallel! Next, connect up the off–board components, again making sure that the LEDs have the correct polarity. The probe leads for the emitter, base and collector should be made from different colours of wire and terminated in probe clips (small insulated crocodile clips).

Check carefully for dry joints and errant blobs of solder. Plug the IC into its holder, ensuring that it is inserted the right way round, as shown in Figure 2.

Testing

Without a transistor in circuit, and the battery connected, both LEDs should be lit. Connect a known good npn transistor and verify that the green LED lights. Now simulate two transistor faults: disconnect the base lead and both LEDs should light; remove the transistor and connect the emitter and collector leads together. Neither LED should light.

Repeat the tests with a known good pnp transistor. The results should be the same, except that the correct indication should now be a lit red LED. On your computer, make a small label of the bulleted list above, to fit on your tester showing the states of the LEDs and what they mean. It will act as a useful *aide mémoire* when you use the tester in future.

Using

The circuit will test transistors in isolation or in an existing circuit, i.e. prior to use. You can check the lead identifications in component catalogues such as the Maplin catalogue. The tester is ideal for going through the large bags of unmarked transistors that you can buy for a song at rallies. You can sort them into three piles – npn, pnp and dud!

Parts list

Resistors: all 0.25 watt, 10% tolerance or better

R1, R2	5600 ohms (Ω)
R3, R4, R5	100 ohms (Ω)

Capacitors

C1, C2	1 microfarad (μF) electrolytic, 16 V WKG

Semiconductors

D1–D4	1N4148 general-purpose silicon diodes
D5	Red LED
D6	Green LED

Integrated circuit

IC1	CMOS 4001 or 4011

Additional items

- Veroboard, 15 strips by 18 holes
- Veropins
- PP3 battery and connector (or 6 × AA cells in PP3 clip holder)
- Switch, SPST
- Connecting wire
- One each of known working npn and pnp transistors for test purposes.

Source

Components are available from Maplin.

70 The ‘Yearling’ 20 m receiver

Introduction

Published to celebrate the first anniversary of *D-i-Y Radio*, this excellent receiver design forms a suitable ‘second receiver project’ for those who have successfully completed the MW receivers earlier in this series. The receiver is powered from a PP3 battery or from a mains adaptor, and can be built with the help of an experienced constructor, on a prototype board. The circuit diagram and some of the components used are shown in the separate diagram. Headphones or a loudspeaker can be used and, once the radio is completed, a few simple adjustments will make the *Yearling* spring to life!

Building the receiver

Before starting the constructional process, start by identifying all the parts. One by one, tick them off against the parts list. Are their values correct? The varactor diode is a twin type (see circuit diagram overleaf), and must be cut *carefully* down the middle, producing two devices, D1 and D2, with two wires each.

First, solder the IC sockets, followed by the coils (inductors); L1 is pink inside the top, and L2 is yellow inside. Then, solder in the varactors; the lettering on D1 should be next to coil L1, and the lettering on D2 should face resistor R7. After those, the capacitors, wire links and resistors should be soldered to the board. Take care to wire the voltage regulator, IC3, correctly. Solder in the crystal X1 as quickly and deftly as you can – crystals do *not* take kindly to having their leads bent and being fried with a soldering iron! Make sure that the electrolytic capacitors C2, C12, C15 and C16 are fitted the right way round. Most electrolytic capacitors have only the negative lead marked.

Figure 1 shows the rear of the front panel, illustrating the connections from the board and antenna socket to the controls. All normal connecting wires are 22 SWG or thereabouts, with insulation. Their lengths should be about 15 cm, except for the battery lead to the switch, which is about 8 cm. It is recommended that you use different-coloured wire for each connection to a control. **Figure 2** shows this. The variable resistor section of VR5, the AF gain control, uses single screened cable connected to 0 V (ground) at the

Amateur Radio

D-i-Y

Year

Antenna Socket

Variable Inductor

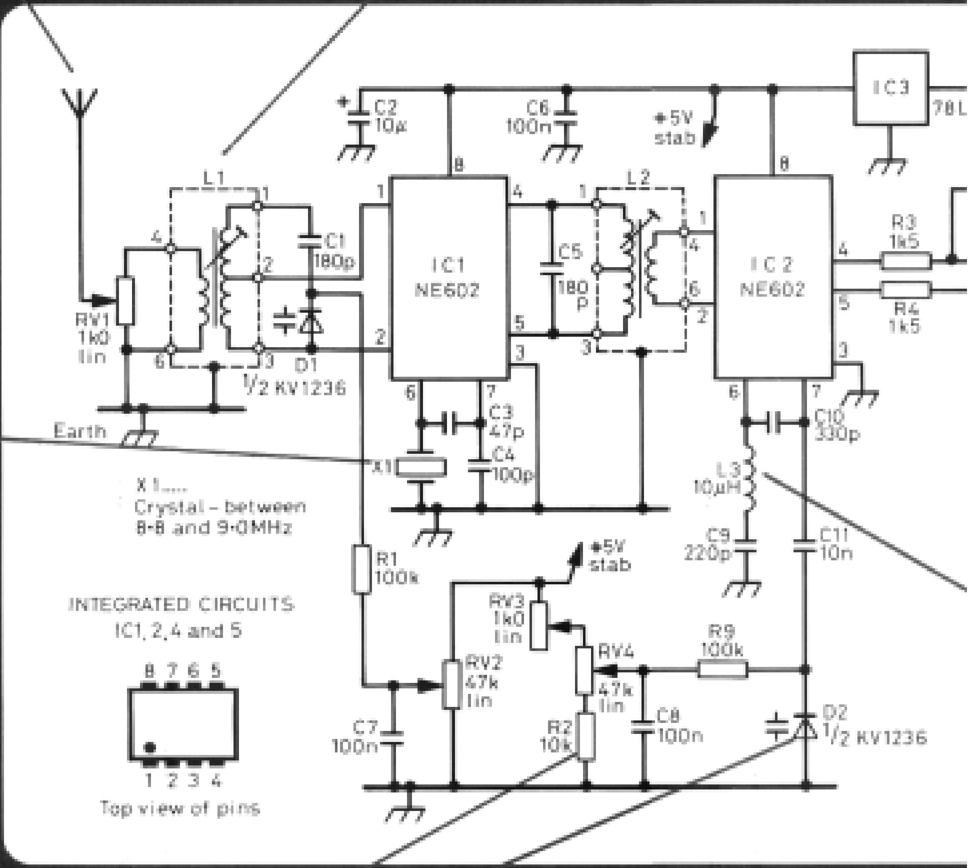
C2
10µ
C6
100n
+5V
stab
IC3
78L
L1
L2
C1
180p
IC1
NE602
C5
180
P
IC2
NE602
R3
1k5
R4
1k5
RV1
1k0
lin
D1
1/2 KV1236
Earth
C3
47p
C4
100p
X1
C10
330p
L3
10µH
C9
220p
C11
10n
X1.....
Crystal - between
8·8 and 9·0MHz
R1
100k
+5V
stab
RV3
1k0
lin
RV4
47k
lin
R9
100k
INTEGRATED CIRCUITS
IC1, 2, 4 and 5
8 7 6 5
1 2 3 4
Top view of pins
RV2
47k
lin
C7
100n
R2
10k
C8
100n
D2
1/2 KV1236

Crystal

Resistor

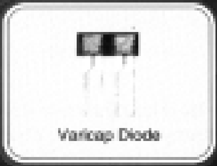

Varicap Diode

**Most radio amateurs enjoy building par
is Do-it-Yourself (D-i-Y) Radio, and it is
successfully built yourself. Licensed
allowed (by the Government) to use ho
can see how radio components are dra
circuit diagram for the**

D-i-Y Radio, RSGB, Lambda House, Cra

– Do It Yourself

RADIO

rling

Polyester Capacitor

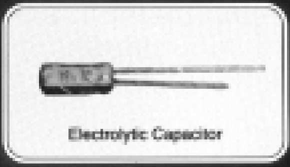
Electrolytic Capacitor

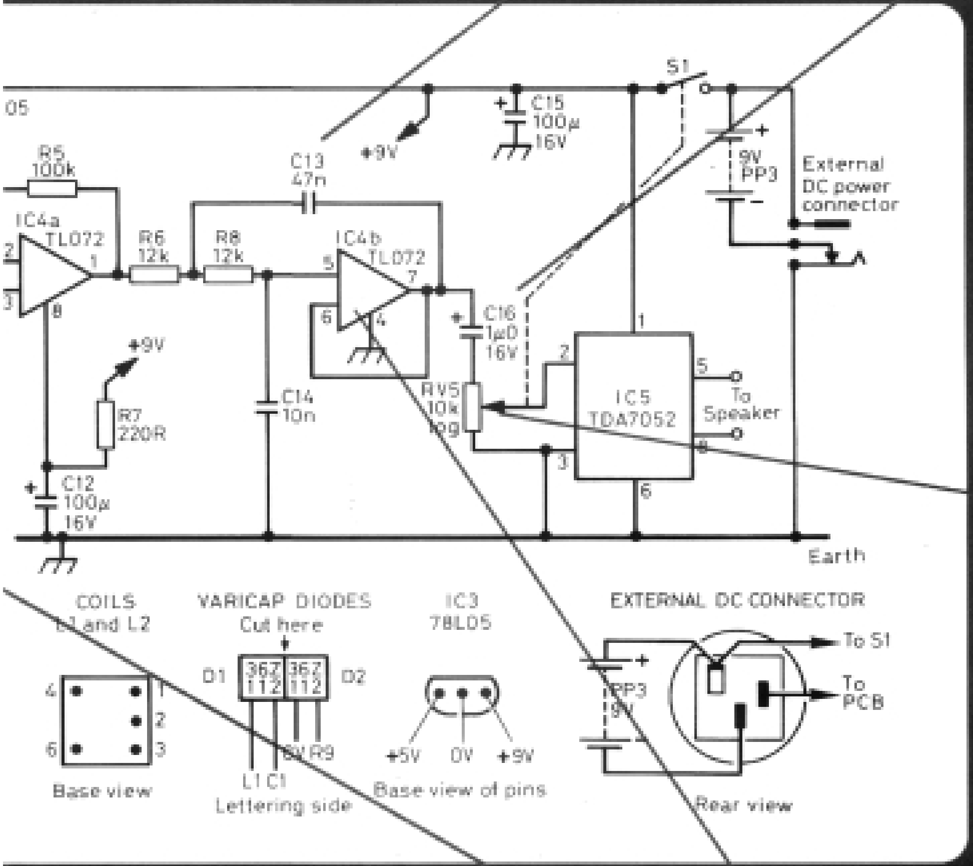

05
R5
100k
IC4a
TL072
2
3
1
8
R6
12k
R8
12k
C13
47n
+9V
C15
100µ
16V
S1
+
9V
PP3
–
External
DC power
connector
IC4b
TL072
5
6
7
4
+9V
R7
220R
C12
100µ
16V
C14
10n
C16
1µ0
16V
RV5
10k
log
1
2
3
5
6
8
IC5
TDA7052
To
Speaker
Earth
COILS
L1 and L2
4
1
2
6
3
Base view
VARICAP DIODES
Cut here
D1
367
112
367
112
D2
0V
R9
L1
C1
Lettering side
IC3
78L05
+5V
0V
+9V
Base view of pins
EXTERNAL DC CONNECTOR
To S1
To
PCB
PP3
9V
Rear view

AF Gain Control

of their radio stations themselves, this
very satisfying to use a radio you have
radio amateurs are the only people
me-built transmitters. On this page you
wn on a circuit diagram. This is the full
D-i-Y Radio Yearling.

Fixed Inductor

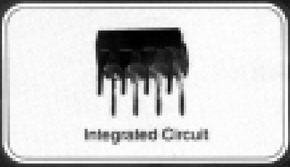
Integrated Circuit

nborne Road, Potters Bar, Herts. EN6 3JE

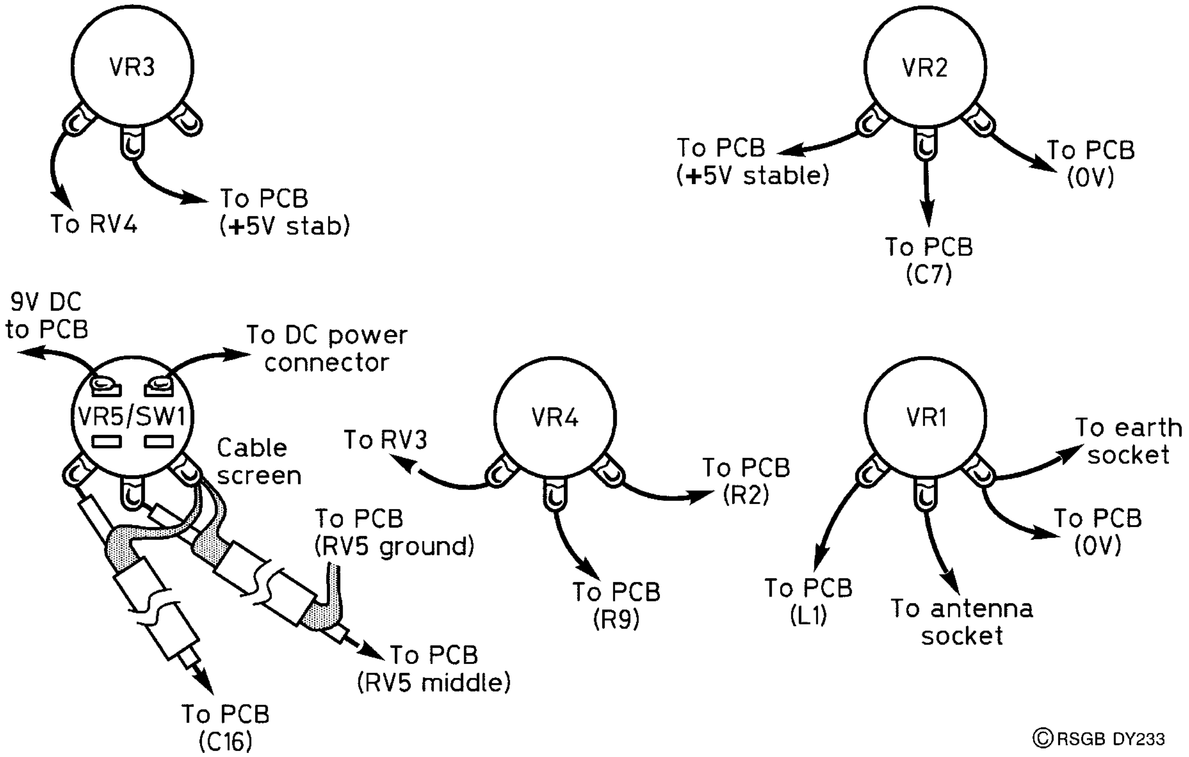
VR3
To RV4
To PCB
(+5V stab)
VR2
To PCB
(+5V stable)
To PCB
(0V)
To PCB
(C7)
9V DC
to PCB
To DC power
connector
VR5/SW1
Cable
screen
To PCB
(RV5 ground)
To PCB
(RV5 middle)
To PCB
(C16)
VR4
To RV3
To PCB
(R2)
To PCB
(R9)
VR1
To earth
socket
To PCB
(0V)
To PCB
(L1)
To antenna
socket

©RSGB DY233

Figure 1 Rear view of the variable resistors. Check the connections carefully to make sure the wires fit the correct holes

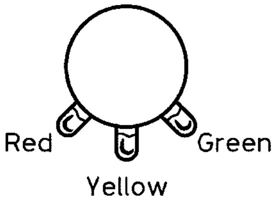
Red
Green
Yellow

©RSGB DY234

Figure 2 It is helpful to use different colour insulated connecting wires. Wires between each variable resistor and the board should be twisted together to give a neat wiring outfit.

board end. Now fit the ICs into their sockets. Make sure that they are the right way round (see circuit diagram) and that each pin lies directly above its corresponding socket before applying *gentle* pressure with the back of the board firmly supported.

Check that all the connections are correct (don’t *assume* this – check the ends of each wire against the circuit diagram) and that all your soldered joints are shiny. Lastly, drill the five 10.5 mm diameter holes for the main controls. On the side of the case, a 6.3 mm diameter hole is needed for the speaker socket, and 8 mm holes for the antenna and earth sockets. The external power socket requires an 11 mm diameter hole.

Adjusting it

Before you can make adjustments, and in order to hear *anything* on your *Yearling*, you will need to connect an antenna (aerial) to the antenna socket. About 8 metres of wire, preferably outdoors and as high as possible, is all you need to connect to the socket. Connect the 3.5 mm jack socket for the speaker, and a 3.5 mm jack to your speaker leads. Connect a battery and switch on.

1. Using a very small screwdriver or, better still, a non-metallic ‘trimming tool’, *gently* screw in the core of L2 as far as it will go, but *don’t force it*. Then unscrew it by three turns anticlockwise.
2. Set VR1, VR2 and VR4 to mid-position and rotate the core of L1 anticlockwise until the hissing noise you hear reaches a maximum intensity. Then adjust L2 for maximum noise.
3. If you now tune carefully with the main tuning control, VR4, you should hear some amateur Single-Sideband (SSB) speech signals. You may have to adjust the bandspread (fine tuning) control, VR3, to make the speech sound normal.
4. Having verified that everything is working, switch off and mount the controls on the front panel and the sockets on the side. To do this, it is much safer to disconnect all the controls and sockets, mount them in their final positions, and then wire them up again.
5. Fit the front panel knobs, connect your aerial and switch on again, checking that everything is working. Then, locate the cluster of CW (Morse) signals to be found at the bottom of the 20 metre band, set the bandspread control to mid-position and slacken off the main tuning knob. Turn the knob (but *not* the control!) until the pointer lies a little clockwise of 14.0 MHz. The SSB signals should now lie roughly between the dial centre and 14.35 MHz. Tighten up the knob.
6. Finally, fix the board to the rear panel, and secure the battery (if you are using one). Attach the rear panel to the back of the box, and you are finished!

Listening!

Make a habit of keeping the bandspread control in its centre position when searching for stations; then you can adjust it either way to make the signals readable. Unless you have a very big aerial, it is best to have the ‘RF Gain’ control, VR1, at maximum. Use the ‘Antenna Tune’ control, VR2, to give the best signal, and control the volume with the ‘OFF/AF Gain’ control. You will find some excellent DX stations with your *Yearling* receiver, and it will serve you well.

Parts list

Resistors (all 0.25 W, 5%)

R1, R5, R9	100 kilohms (kΩ)
R2	10 kilohms (kΩ)
R3, R4	1.5 kilohms (kΩ)
R6, R8	12 kilohms (kΩ)
R7	200 ohms (Ω)

Capacitors (all rated at 16 V or more, tolerance *at least* what is quoted)

C1, C5	180 picofarads (pF) polystyrene 5%
C2	10 microfarads (μF) electrolytic
C3	47 picofarads (pF) polystyrene 5%
C4	100 picofarads (pF) polystyrene 5%
C6, C7, C8	100 nanofarads (nF) or 0.1 microfarad (μF) ceramic
C9	220 picofarads (pF) polystyrene 2%
C10	330 picofarads (pF) polystyrene 2%
C11, C14	10 nanofarads (nF) or 0.01 microfarad (μF) ceramic
C12, C15	100 microfarads (μF) electrolytic
C13	47 nanofarads (nF) or 0.047 microfarad (μF) polyester, 5%
C16	1 microfarad (μF) electrolytic

Variable resistors

VR1, VR3	1 kilohm (kΩ) linear
VR2, VR4	47 kilohms (kΩ) linear
VR5	10 kilohms (kΩ) log with switch

Inductors

L1	Toko KANK3335R
L2	Toko KANK3334R
L3	10 microhenries (μH), 5%

Semiconductors	
IC1, IC2	Philips/Signetics NE602 or NE602A
IC3	78L05 5 V 100 mA regulator
IC4	TL072 Dual Op-Amp
IC5	Philips TDA7052 audio amplifier
Additional items	
D1, D2	Varactor diode Toko KV1236 (cut into two sections – see text)
X1	Crystal 8.86 MHz type (from Maplin, etc.)
4 off	8-pin DIL sockets for IC1, IC2, IC4 and IC5
2 off	4 mm sockets aerial (red) and earth (black)
1 off	3.5 mm chassis-mounting speaker jack socket
1 off	DC power socket for external supply (if required)
4 off	Red knobs with pointers
1 off	Tuning knob with pointer (e.g. 37 mm PK3 type)
1 off	Printed-circuit board or prototype board
1 off	Plastic case approx 170 × 110 × 6 mm (e.g. Tandy number 270-224)
1 off	Speaker 8–32 Ω impedance (or headphones)

71 Adding the 80 metre band to the Yearling receiver

Background

You will have noticed that your *Yearling* receiver has a dial which shows coverage of the 80-metre amateur band (3.5–3.8 MHz). This band is used for local contacts during the day, and contacts up to about 1600 miles in darkness. Longer distances are possible, particularly in the middle of winter.

The modifications

Only a few extra parts are required, as you may have noticed from the parts list. A low-pass filter, FL1 (one which passes low frequencies and rejects higher frequencies), is switched into the circuit on 80 m. The circuit of the

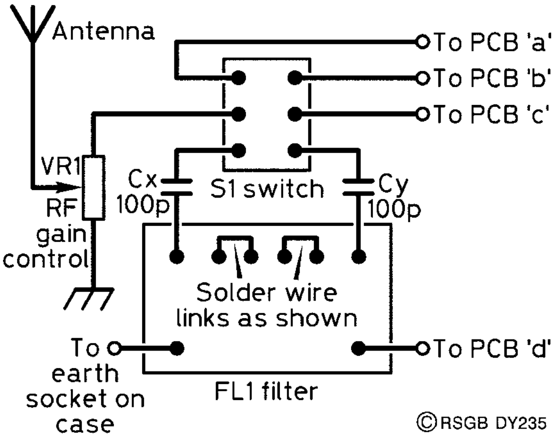
Antenna
To PCB 'a'
To PCB 'b'
To PCB 'c'
VR1
RF
gain
control
Cx
100p
S1 switch
Cy
100p
Solder wire
links as shown
To
earth
socket on
case
To PCB 'd'
FL1 filter
©RSGB DY235

Figure 1 The circuit diagram shows the extra components for 80 m operation. Note the connections to the PCB 'a', 'b', 'c' and 'd'

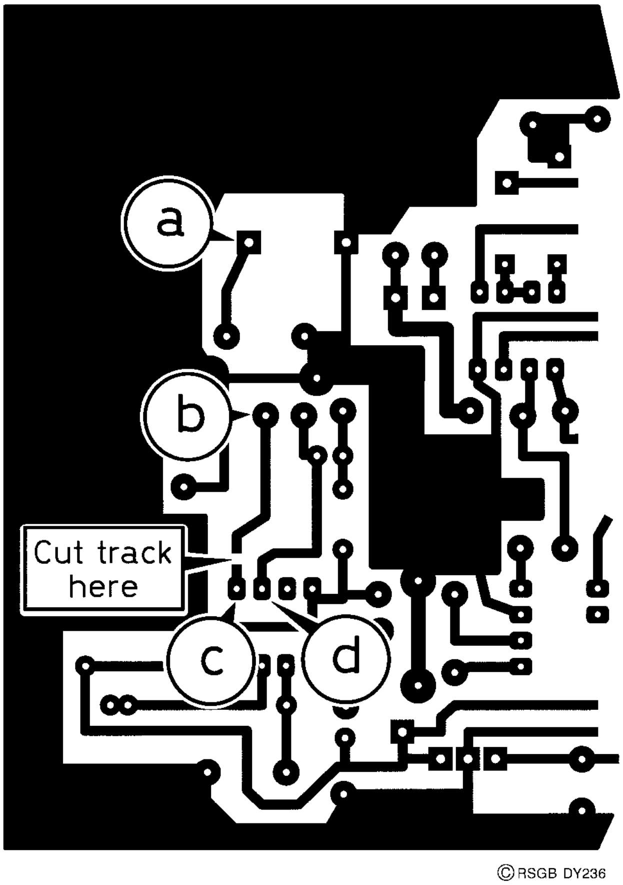
a
b
Cut track
here
c
d
©RSGB DY236

Figure 2 The underside of the PCB. Wires are connected from the switch and filter as shown

switch and its connections is shown in **Figure 1**. Before making the modifications, dismantle the receiver so that you have easy (and safe) access to the case and the track side of the PCB.

1. The first thing to do is to drill a 6.5 mm diameter hole in the side of the case into which the switch fits.
2. Then, using a sharp Stanley knife or scalpel, carefully cut the track on the PCB as shown in **Figure 2**, making a gap about 1 mm wide.
3. Using 10 cm lengths of different-coloured insulated wire, make the four connections, a, b, c and d, to the PCB, as shown in Figure 2.
4. Solder the two links on the filter, and then make the connections to the switch, capacitors and PCB. You will have to disconnect the existing wire between the RF gain control, VR1, and pin 4 of L1 on the PCB.
5. Lastly, check your new connections carefully, then mount the switch in the new hole and fix the filter to the bottom of the case with a little glue, as shown in **Figure 3**. Reassemble the circuit, and replace the back of the case.

More testing!

Firstly, switch your new switch, S1, into the 20 m position, to check that the original circuit still works! If you find that the ‘Antenna Tune’ control peaks at a slightly different position, don’t worry.

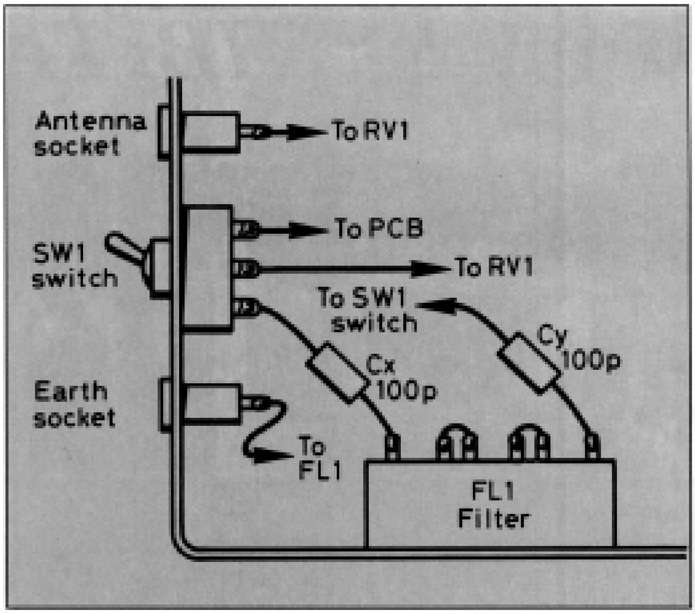

Figure 3 Internal view of the Yearling case. The filter FL1 is attached to the base with glue

Now switch to 80 m, and tune around the anticlockwise end of the dial; you should hear some SSB stations, particularly in the evenings and at weekends, when may people are on the air. At the other end of the travel of the tuning control, you should hear CW (Morse) stations.

The Radio Society of Great Britain broadcasts amateur radio news every Sunday morning on or about 3.65 MHz; the table below has the details. Finally, a good antenna is more important than ever for 80 metre reception – aim for more height and length, and then consider the project concerned with making an Antenna Tuning Unit (ATU)!

Parts list

Capacitors (all rated at 16 V or more, tolerance 10% or better)

Cx, Cy	100 picofarads (pF) polystyrene

Filter

FL1	Toko 237LVS1110 low-pass filter

Additional items

SW1	2-pole 2-way (changeover) toggle switch
7 off	Short lengths of insulated wire of different colours

The RSGB news broadcasts, GB2RS – Sunday mornings

Frequency (MHz)	*Local time*	*Reception area*
3.650	0900	SE England
3.650	0930	Midlands
3.650	1000	SW England
3.650	1100	Yorkshire
3.640	1130	Aberdeen
3.660	1130	Glasgow

The Midlands transmission is repeated at 1800 (6pm) local time on 3.650 MHz. All frequencies are approximate in order to avoid interference, and use lower sideband (LSB). If you also have a 40 metre receiver, there are GB2RS news broadcasts on 7.048 MHz at 0900 local time from Northern Ireland and from 1100 local time from the north midlands.

72 How the Yearling works

Introduction

The *Yearling* was designed to provide an introduction to *Amateur Radio* on the 20 m amateur band. Let's look at how the different sections (or 'stages') of a radio work, and how they fit together to form a complete receiver. **Figure 1** shows a block diagram which you can follow and compare with the circuit diagram of your *Yearling* receiver.

The antenna (or aerial)

Connected to your receiver, it will pick up not only amateur signals, but all other signals as well! This means that the receiver has to select the one signal that interests you, while rejecting all the others. The following stages do just that.

The RF filter

This stage (centred around L1) selects the band of radio frequencies (RF) containing the signal you want, in this case, those having wavelengths around 20 m. Signals from the 40 m band, for example, would not get through.

The crystal oscillator

This is an oscillator circuit designed around a quartz crystal (X1), and has a very stable frequency. It produces a single, very pure frequency to feed into the mixer. A crystal having a frequency between 8.800 MHz and 9.000 MHz is suitable for this circuit. The oscillator and mixer functions are both carried out inside IC1.

The first mixer

Yes, this stage 'mixes' two signals together. In this case, the two signals are (i) from the aerial via the RF Filter, and (ii) from the crystal oscillator. Two

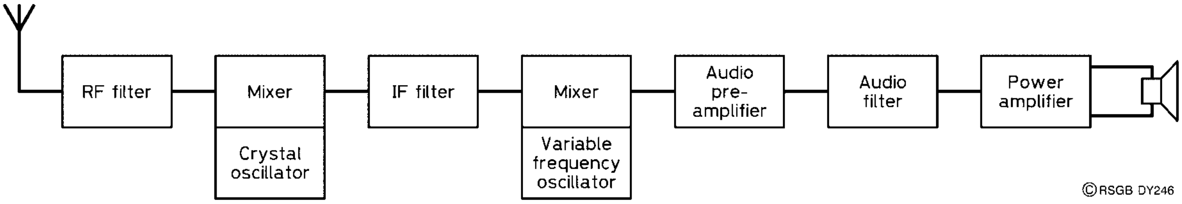

Figure 1 Block diagram of the Yearling, showing how the various stages fit together to make a complete radio receiver

bands of signals emerge from the mixer. The first is centred upon a frequency equal to the incoming signal frequency **added** to the crystal frequency, and the second is centred upon a frequency equal to the incoming signal frequency **subtracted** from the crystal frequency. Look at an example – if the signal is at 14 MHz and the oscillator at 9 MHz, then the mixer outputs will be 14 + 9 = 23 MHz and 14 – 9 = 5 MHz.

Intermediate frequency (IF) filter

It is the purpose of the IF filter (centred around L2) to select only one of these two bands of frequencies emerging from the mixer. In this case, it is the lower band of frequencies (around 5 MHz) which we select. This is because, in general, lower frequencies are easier to handle than higher ones.

Variable-frequency oscillator (VFO)

The VFO (part of IC2) enables us to tune into a particular station, and operates over a band of frequencies between 5 MHz and 5.35 MHz in this receiver. You will notice that IC1 and IC2 are the same type of chip, so that you will be expecting another mixer stage to be associated with the VFO. You are quite right!

The second mixer

This mixer obeys exactly the same rules as those of mixer 1. Sum and difference frequencies are produced, like this. Mixing is between the incoming IF signals (around 5 MHz) and the VFO signals (around 5 MHz), producing output frequency bands centred upon 10 MHz and 0 MHz. The use of the words ‘band of frequencies’ throughout this explanation is intentional. If all the signals were pure, there would be no bands; the bands are produced because of one thing – the modulation imposed on the pure frequencies at the transmitter. So, the ‘bands’ contain the one thing that we want to extract from the signal, and that is the speech or Morse code that the signal contains. The band of frequencies at 0 MHz is just that – the audio frequencies we want in the loudspeaker. Because of this, the audio frequency output of the second mixer is selected and passed on for amplification.

The audio preamplifier

Preliminary amplification of the minute audio signal which emerges from the second mixer is provided by IC4a, which will respond only to audio signals, automatically rejecting the 10 MHz signal.

The audio filter

The bandwidth of normal speech when transmitted by an amateur station is around 3 kHz, so there is no advantage to be gained in amplifying frequencies greater than this. IC4b is known as a low-pass filter, because it passes (lets through) lower frequencies and rejects higher ones.

The power amplifier

IC5 produces the final audio amplification and provides enough power (about 350 milliwatts (mW) to drive a small speaker.

How does it work on 80 m?

If you have fitted the 80 m modification to your receiver, you are probably wondering how the circuit works at this different frequency. Firstly, the filter which you fitted selects the 80 m band instead of the 20 m band. The only other slight difference lies in the way the first mixer stage works. Its job is to produce sum and difference frequencies from the incoming signal and crystal frequencies. On 20 m, it did this by subtracting the crystal frequency (9 MHz) from the incoming frequency (14 MHz) to produce an IF output of 5 MHz. On 80 m, the incoming frequency (3.5 MHz) is subtracted from the crystal frequency (9 MHz) to produce an IF output of 5.5 MHz, which is still within the tuning range of the VFO in the next stage.

73 A field strength meter

Introduction

Having a reliable field strength meter is always useful. There are two basic types – the tuned variety (also known as an absorption wavemeter) and the broad-band variety. A project centred on the tuned variety can be found in this book (*An HF absorption wavemeter*), and the way it works is also presented (*The absorption wavemeter*). This project describes an untuned or broad-band type which is every bit as useful as the tuned type, when it is the level of RF in which you are principally interested.

With the possible exception of the meter, you may have all the parts needed in your junk box. The prototype was distinguished by its meter, found at a rally, and marked ‘Safe’, ‘Dangerous’ and ‘Explosive’ on its scale.

How it works

The basic field strength meter uses the circuit of a crystal set, but with a meter replacing the headphones. A better design, which is used here, is that of a voltage doubler, giving more sensitivity. **Figure 1** illustrates the voltage doubler circuit.

The AC input shown will be our RF input, which will be explained soon. Diode D1 will pass the positive half of the signal and use it to charge up C1 to the peak value of the signal. D2 uses the negative half of the signal to charge up C2 to the same value. Because C1 and C2 are in series, the peak voltage appearing across *both* of them (which is the DC output voltage) is equal to *twice* the peak input voltage, hence the name voltage doubler. If you’re wondering why DC is present at the output when AC comes in at the

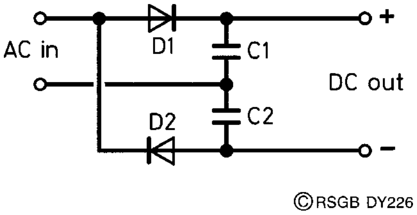

Figure 1 Voltage doubler circuit

input, remember that capacitors pass AC (RF) but block DC; thus the RF is shorted out through C1 and C2, but the steady voltage (DC) remains across the two capacitors. The voltage is thus proportional to the size of the RF signal applied at the input.

The circuit

The voltage doubler is converted into a field strength meter using the circuit of **Figure 2**. A piece of wire serves as the aerial to provide an RF signal across the radio-frequency choke (RFC). A choke is an inductor which is large enough to prevent the RF passing through it – it ‘chokes’ the RF. This produces the maximum RF signal at the input to the voltage doubler, and the DC output from it is measured on the meter.

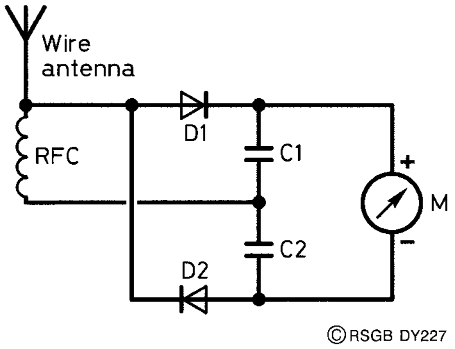

Figure 2 Field strength meter

The parts used are not very critical. The RF choke can have any value between about 1 mH and 2.5 mH. Almost any common diode such as the 1N914 or 1N4148 can be used for D1 and D2. The two capacitors could be any value between 1 nF and 100 nF (0.001 μF and 0.1 μF). The meter should be reasonably sensitive, with a full-scale deflection (FSD) in the range 50 μA to 100 μA. Look for VU meters at rallies – these are ideal.

Construction

The prototype circuit was made on matrix board. If you don’t want to use pins with the board, simply push the component leads through the holes and make the connections on the underside of the board, either with the excess component leads themselves, or with ordinary connecting wire.

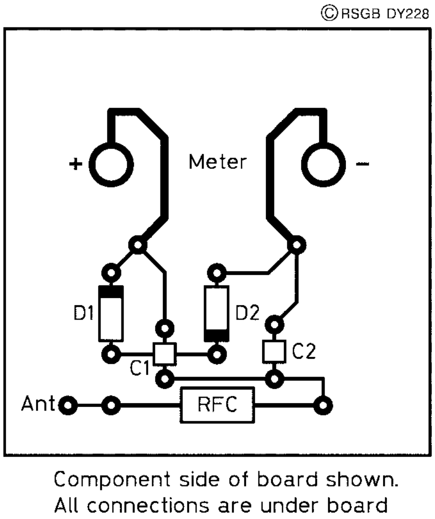
©RSGB DY228
+
Meter
–
D1
D2
C1
C2
Ant
RFC
Component side of board shown.
All connections are under board

Figure 3 Board layout and interconnections

Figure 3 shows how the parts are placed and connected. The matrix board can be mounted directly on the meter using its terminal bolts!

In use

Needing no power supply other than an RF signal, just connect it all up and leave it to work! A short length of insulated wire is enough to pick up some RF and display it on the meter. Using a 200 μA meter, about 3 metres of wire gave a good deflection on the meter. To increase deflection, the wire can be wrapped around the aerial lead, provided that the wire is PVC covered and doesn't come into contact with the aerial wire.

This requires a little experimentation. Try a long piece of wire first and adjust its position until the meter needle kicks whenever there is a transmission. It is very reassuring to see the meter moving during a transmission. Although SWR meters also indicate power, they are usually set to read reverse power, and show little or no movement during transmission.

Parts list

D1, D2	1N914, 1N4148 or similar
C1, C2	10 nF disc ceramic
RFC	Miniature axial choke (1 mH)
Meter	Surplus VU meter or similar
Matrix board or similar	

Components are available from Maplin.

74 Preselector for a short-wave receiver

Introduction

A preselector is a simple RF tuned amplifier which is inserted between the aerial and the receiver. It provides some extra gain and may improve the overall performance of the receiver. This project uses a Field-Effect Transistor (FET) amplifier in grounded-gate mode.

The design has a tuned circuit at both the input and output which, with excessive gain and poor construction, would produce only one thing – oscillation! So, to avoid this happening, we will have only a low gain, and use a circuit which provides good isolation between input and output. The grounded-gate FET amplifier fulfils both these criteria. It will also cover a frequency range from about 7 to 30 MHz, which includes most of the HF amateur bands.

The circuit

This is shown in **Figure 1**. The signal from the aerial arrives at an RF transformer, the secondary of which is tuned with capacitor VC1a. The output from the tuned circuit is taken from a tap on the secondary to the source of the FET. The gate is grounded (earthed) and the amplified signal appears at the FET drain, which is then fed to the primary of another RF transformer, which is tuned by VC1b. The output to the receiver comes from the secondary of the RF transformer.

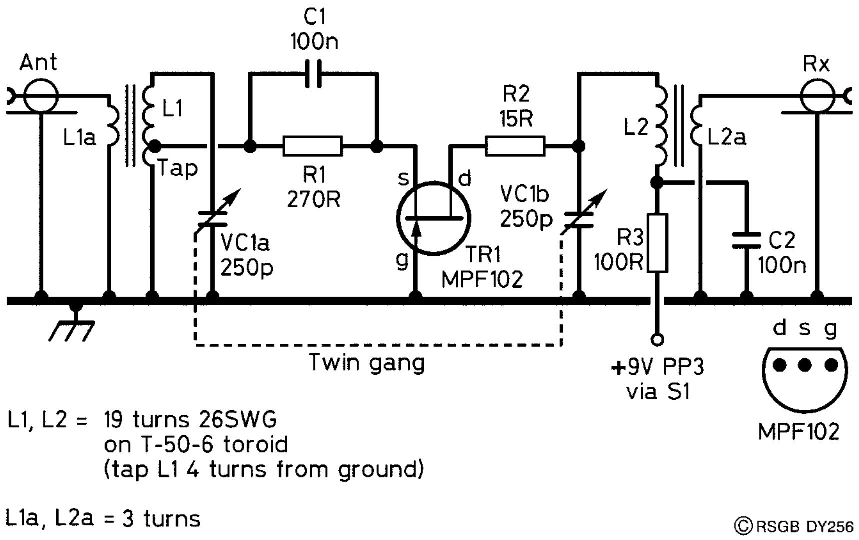
C1
100n
Ant
Rx
L1
L1a
Tap
R1
270R
R2
15R
L2
L2a
s
d
g
VC1b
250p
VC1a
250p
TR1
MPF102
R3
100R
C2
100n
Twin gang
+9V PP3
via S1
d s g
MPF102
L1, L2 = 19 turns 26SWG
on T-50-6 toroid
(tap L1 4 turns from ground)
L1a, L2a = 3 turns
©RSGB DY256

Figure 1 HF preselector, circuit diagram

Notice that the two RF transformers are identical, but they are used 'back to back', with the secondary of the first and the primary of the second being tuned. They are tuned with identical capacitors, fitted on the same shaft of a variable capacitor. We say that the two capacitors are 'ganged'. Because L1 and L2 are the same, and VC1a and VC1b are the same, both RF transformers should be resonant at the same frequency, no matter what that frequency is.

Construction

The final layout should look something like that shown in **Figure 2**. The external connectors and controls being two SO-239 sockets for connection to your receiver and aerial, a tuning control and its associated scale, and an on/off switch.

The circuit can be put together on a plain matrix board, using pins to anchor the components, or simply by pushing the component leads through the board and making connections on the underside. The layout of the prototype is shown in **Figure 3**. Mounting the board to the aluminium box is accomplished with bolts, solder tags and stand-off insulators.

Check your construction against the circuit diagram and against the layout diagram. Wire in the PP3 battery clip, put the switch in the 'off' position, and fit the battery. Testing can be carried out without fitting the top of the box.

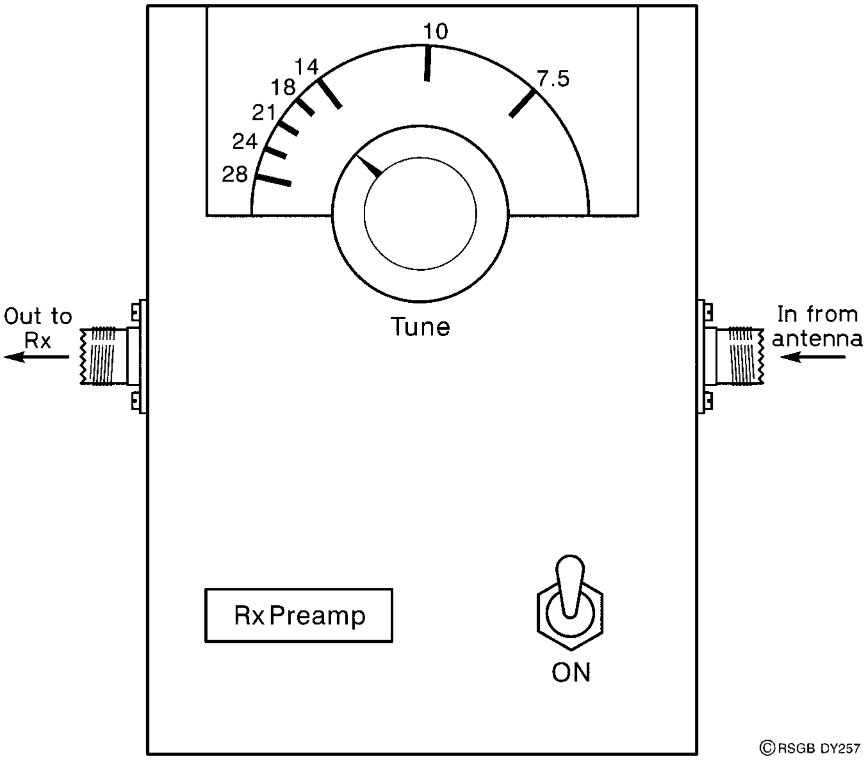

Figure 2 HF preselector, front panel layout

Testing

Don't fit your preselector yet. Tune your radio to a broadcast station, preferably a fairly weak one. Disconnect the aerial and fit your preselector between the aerial and receiver. Switch it on and rotate the tuning knob slowly. You should find a position where your original station is received more clearly than before. If it doesn't work at all, recheck your wiring. Is there a positive voltage on the drain of the FET? If not, work back towards the positive battery terminal. Is there a voltage at the junction of L2 and R2? Is there a voltage at the junction of L2 and R3? Is there a voltage at the junction of R3 and the battery lead? If there isn't a voltage at that point, then you have probably mounted your switch upside down, and it is off, not on! It's a common mistake.

Calibration

This is not obligatory, surprisingly enough. However, if having a frequency scale appeals to you, then using an RF signal generator (or using the services of a friend who has one) is the simplest solution. Feed in a weak modulated

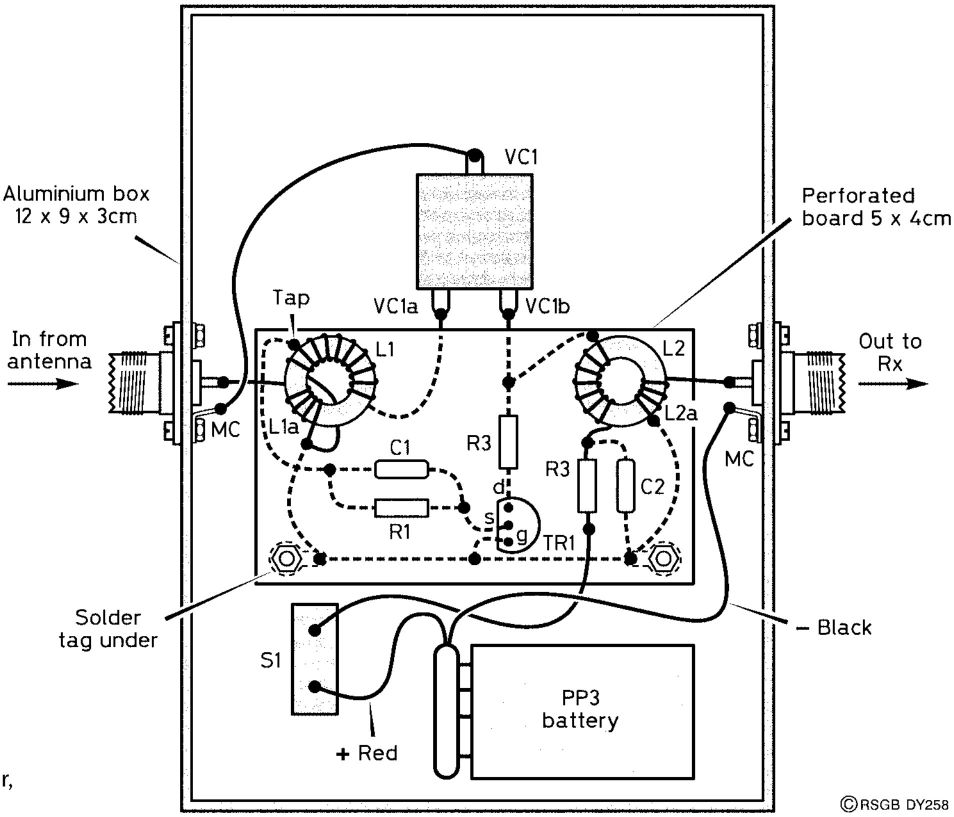
VC1
Aluminium box
12 x 9 x 3cm
Perforated
board 5 x 4cm
Tap
VC1a
VC1b
In from
antenna
L1
L2
Out to
Rx
MC
L1a
L2a
C1
R3
R3
C2
MC
d
s
g
TR1
R1
Solder
tag under
– Black
S1
PP3
battery
+ Red
©RSGB DY258

Figure 3 HF preselector, component layout

signal from the generator to the preselector, and rotate VC1 until the signal is maximum. Mark this frequency on your dial. Repeat the process for the frequencies shown on the dial in Figure 2.

If you have a commercial transceiver, feed its output into a dummy load (such as the type described in *A Switched Dummy Load*, elsewhere in this book) using a distinctive modulating signal such as an idling RTTY signal. Set the receiver to the same frequency with the preselector out of circuit. Insert it into the aerial lead, and search for the signal with VC1 until it gives the maximum deflection on your S-meter, then mark the frequency on your dial.

Parts list

Resistors: all 0.25 W carbon film or better

R1	270 Ω
R2	15 Ω
R3	100 Ω

Capacitors

C1, C2	100 nF (0.1 μF) ceramic
VC1	250/250 pF polyvaricon

Inductors

L1, L2	19 turns 26 SWG enamelled copper on T.50.6 toroid L1 has a tap 4 turns from ground end L1a and L2a – 3 turns wound over previous winding

Semiconductors

TR1	MPF102 FET

Additional items

Matrix board	To fit aluminium box (see Figure 3)
Aluminium box	12 × 9 × 3 cm
Battery and connector	PP3 9 V
S1	SPST on/off switch
Knob	As required
SO239	Sockets – 2 required

75 An audible continuity tester

Introduction

This is not the only continuity tester in this book. This alone attests to their use, so you may well want to experiment with several designs, then come up with one of your own! The very simplest form of continuity tester is probably a battery and a bulb in series, with the circuit being closed by connecting it to a fuse or other object being tested for continuity. The bulb could be replaced by a buzzer to give an audible indication. The current taken by the buzzer could damage some components, however. An ohmmeter can also be used, and is very popular for the purpose, as it indicates whether the circuit is low, medium or high resistance. This project has the advantages of indicating whether there is no continuity, some resistance or good continuity, and making an audible sound, so that you don't have to move your eyes from the circuit while making the test.

The circuit

Using only one integrated circuit, four components and a battery, this is a particularly simple circuit, as **Figure 1** shows. IC1 is usually used to flash an LED from a 1.5 V source, and to have a low current consumption. By changing the component values, IC1 is made to oscillate at audio frequencies, and we can hear these through the loudspeaker, LS1. Low-impedance speakers can be used, but result in an increased current drain, so the higher the impedance, the better.

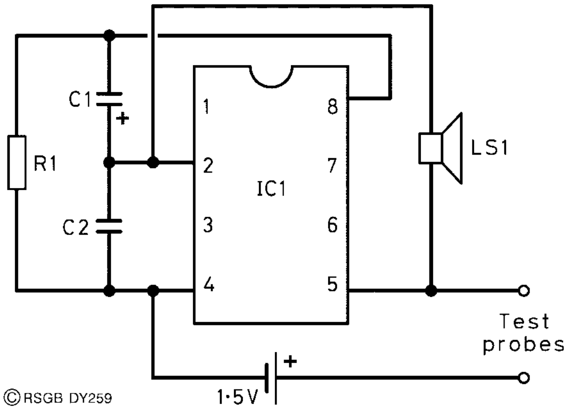

Figure 1 Continuity tester, circuit diagram

Construction

The layout of the circuit on Veroboard measuring 7 strips by 15 holes is shown in **Figure 2**. Start by making the four track cuts which will lie underneath the IC. Use a track cutter or a 3 mm (⅛ inch) twist drill rotated between thumb and forefinger for this. Solder the Veropins in place, followed by the wire link. Then solder in the IC holder, the resistor and the capacitors, making sure that the electrolytic capacitor, C1, is connected the correct way round.

Check, with the board against a bright light, that there are no shorted tracks, either by large blobs of solder or by copper swarf from the track-cutting process. Then insert IC1 into its holder the right way round. The probes can be ordinary connecting wire, the free ends being tinned with solder to prevent wire whiskers from touching components other than the one you are testing.

If you are happy that the circuit and the wiring appear to be correct, fit the battery into its clip. Nothing should happen until the probes are touched together, when you should hear a note from the loudspeaker. If nothing happens, all you can do is to recheck your circuit, as there is nothing else to go wrong!

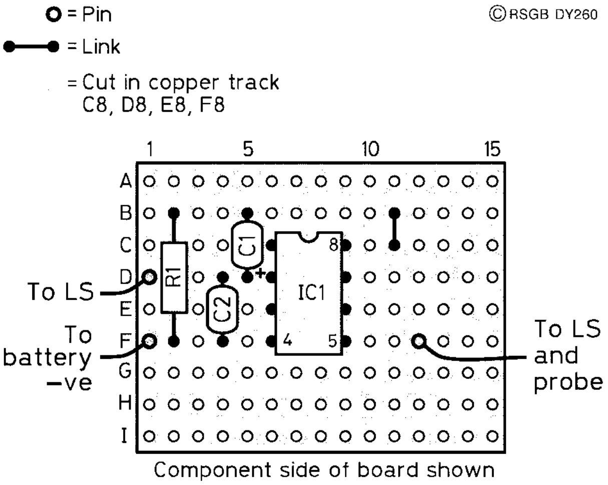
= Pin
= Link
= Cut in copper track C8, D8, E8, F8
©RSGB DY260
1
5
10
15
A
B
C
D
E
F
G
H
I
R1
C1
C2
IC1
8
4
5
+
To LS
To battery −ve
To LS and probe
Component side of board shown

Figure 2 Continuity tester, component layout

Any box can be used; there is no justification for a metal case, unless you want to make use of something, which is to hand, such as a tobacco tin. Alternatively, any suitable plastic box will do.

Use

The tester will give different pitch notes for resistances of different values. The higher the resistance, the higher the pitch from the speaker. Try it with small inductors, and you will learn to recognise the different tones produced by the IC.

Safety notice

Using the continuity tester on components *in situ* is not advisable, as the results could be misleading. It can be dangerous to make measurements *in situ* on equipment, which is operating. If you must make such tests, always disconnect the equipment form its power source first.

Parts list

Resistor

R1	1000 Ω (1 kΩ), 0.25 W carbon, 10% or better

Capacitors

C1	10 µF electrolytic, 16 V
C2	0.1 µF subminiature polyester or ceramic

Integrated circuits	
IC1	LM3909N
Additional items	
	Veroboard 7 strips by 15 holes
LS1	64 Ω miniature loudspeaker
	IC holder
	Veropins
	Connecting wire
	Test leads
	1.5 V AA alkaline battery

76 An experimental 70 cm rhombic aerial

Introduction

Most commonly used aerials can be classed as *resonant* or *standing-wave* aerials. There is another class known as *non-resonant* or *travelling-wave*. Resonant aerials, such as the dipole, are narrow-band; this occurs because resonance occurs only over a narrow band of frequencies. Travelling-wave aerials, on the other hand, can operate over a wide band of frequencies.

The *rhombic* is an example of a non-resonant or travelling-wave aerial. It is often employed for fixed commercial and military short-wave radio links. Made with wire, it has a diamond shape when looking down on it from above. The four corners are supported on four masts. It is a very effective aerial, and has good gain, a quality which can be judged from the *polar diagram* shown in **Figure 1**. This has been obtained from a computer program, and so is the perfect shape for a rhombic aerial.

Theory (but only a little)

A polar diagram shows graphically the ability of an aerial to radiate (or receive) more effectively in one direction at the expensive of the radiation in other directions. Figure 1 shows the polar diagrams of two aerials, a simple 70 cm dipole and the rhombic described in this project. The dipole has the

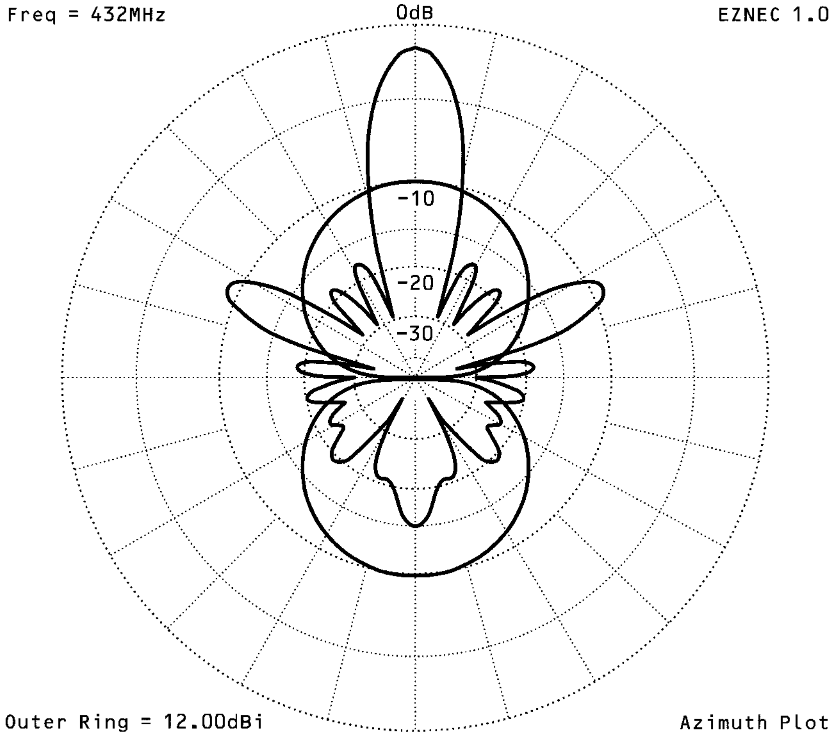

Figure 1 The horizontal polar diagrams of the small rhombic antenna and a dipole compared

©RSGB DY262

well-known figure-of-eight shape, showing a symmetrical angular behaviour about the direction of the aerial and about a direction perpendicular to the aerial. That of the rhombic, on the other hand, is quite irregular by comparison, but is still symmetrical about one axis only, not two. It's this asymmetry that gives the rhombic its gain, by virtue of its *front-to-back ratio*. This is the ratio of the power radiated forwards to that being radiated backwards. Notice the large *lobe* (lump) at the top of the polar diagram; this is the direction in which the aerial transmits best. The lobe in the opposite direction has been reduced significantly, allowing more power to be directed forwards, not backwards

Problems

The rhombic is not found in every amateur's back garden, despite its attractions. To work best each edge of the diamond shape should be about two wavelengths long. Hence, a rhombic for the 20 metre band could be about 80 m from tip to tip! Another disadvantage is that, because of its size, it cannot be rotated.

It will operate well over a range of frequencies. An HF rhombic could work on the 7, 10, 14, 18, 21 and 28 MHz bands. One for the lower VHF frequencies could operate on the 50, 70 and 144 MHz bands.

Its large size is less of a problem at UHF. A portable rhombic can be made for 70 cm which can be used with a hand-held transceiver. The aerial design to be described here will give a gain of up to 9 dB relative to a dipole. This is written as 9 dBd, the second ‘d’ meaning ‘relative to a dipole’. This is equivalent to improving your signal by 1.5 S-points at the receiver or (and you may be surprised by this) by fitting a linear amplifier to your transceiver that would take 5 W input and produce 40 W output to a normal dipole! Consider the relative costs of the two approaches to producing the same received signal. You will also receive everyone else’s signals 1.5 S-points better than before!

Unlike the popular Yagi aerial, this design has no critical dimensions and can be folded up for transport by car or bicycle. It has a fairly high input (or feed) impedance, being fed usually by balanced twin feeder. To feed it with standard 50 Ω coaxial cable, a matching transformer in the form of a *balun* is required. A balun will convert the *bal*anced (symmetrical) aerial impedance to the *un*balanced and lower impedance of the coaxial cable. The details of how to build the balun, which in this case is a half-wave transformer, are included in the constructional details.

Construction

The aerial frame is made up of 1 cm × 2 cm strips of wood fixed to a plywood centre using 30 mm long M4 bolts, as shown in **Figure 2a**. The outer bolts fixing the front and side supports can be removed for folding prior to transportation, as shown in **Figure 2b**.

The wires are fixed to the front and rear supports using screw connectors, sometimes called ‘chocolate block’ connectors. Detail X and Y of Figure 2 show how this is done. The side supports have holes in the ends through which the wire is threaded.

The aerial must be mounted in the horizontal plane using a small shelf bracket attached to the centre plate, using the same bolts that hold the rear support. The other half of the bracket may be mounted to a vertical mast using screws or jubilee clips (hose clips).

The balun

Cut a 23 cm length of coaxial cable; this is to be our half-wave transformer. Cut and remove 2 cm from each end of the sheath. Make the braid into a pigtail at each end. Cut and remove 1 cm of the inner insulator from each

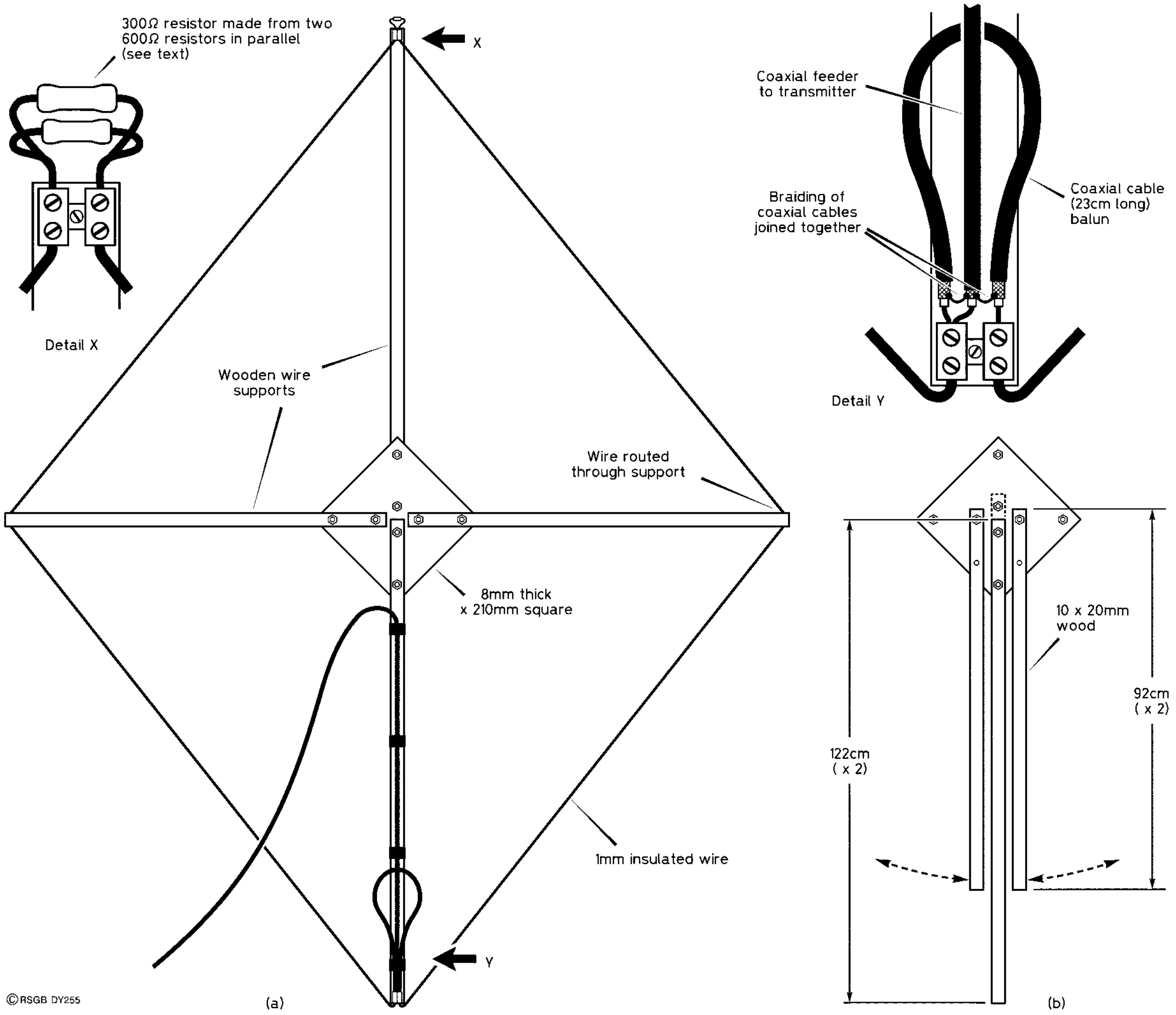
300Ω resistor made from two
600Ω resistors in parallel
(see text)
Detail X
X
Coaxial feeder
to transmitter
Braiding of
coaxial cables
joined together
Coaxial cable
(23cm long)
balun
Detail Y
Wooden wire
supports
Wire routed
through support
8mm thick
x 210mm square
10 x 20mm
wood
92cm
(x 2)
122cm
(x 2)
1mm insulated wire
Y
©RSGB DY255
(a)
(b)

Figure 2 Construction of the UHF rhombic antenna

end, leaving 1 cm of the centre conductor exposed. Fasten this piece of cable to the chocolate block connector together with the coaxial feeder cable as shown in Detail Y of Figure 2. The braids of all three prepared ends are soldered together, but are not connected to anything else. The inner conductor of the feeder from the transceiver is soldered to one end of the centre conductor of the 23 cm piece, and is connected to one end of the rhombic by one side of the chocolate block. The other end of the half-wave transformer also goes to the chocolate block, where it is connected to the other side of the rhombic loop.

The opposite corner of the rhombic, as shown in Detail X of Figure 2, shows that the loop is broken at the chocolate block, and is 'terminated' by two 600 Ω resistors in parallel. This makes the aerial a broad-band travelling-wave device, and gives it its directivity and gain. There is a rule of thumb, which includes a safety factor, which says that the terminating resistor must be able to absorb one-half of the maximum transmitter output power. So, if you use two 2 W 600 Ω carbon resistors (*not* wire-wound resistors) in parallel, you can use a transmitter with an output of 8 W, which is more than adequate for 70 cm hand-helds.

Using the rhombic

Fit the aerial to a pole or mast in the horizontal plane, if you intend to use CW or SSB, but in the vertical plane if you want to concentrate on FM work. In the latter case, a wood or fibreglass pole is mandatory. Tune to a local repeater, whose signal strength you know. Rotate the aerial to face the repeater, and you should see that the signal strength is greatly improved! Verify the directional properties of the aerial by rotating it and observing the changes in signal strength on your S-meter.

77 Water level alarm

Introduction

This is a simple device which, when you have built it, you begin to wonder how you ever managed without it! It will sound a buzzer when the level of water reaches a particular point, which you can set and vary at will.

Detecting water

Water is not too difficult to detect electrically, because it is a conductor of electricity. Not a good one, but sufficiently so to carry enough current to control a warning circuit.

The circuit is shown in **Figure 1**. If you imagine a resistor placed between the ends marked 'water detector', the circuit is then recognisable as a transistor used as an electronic switch. Whether the switch is off or on depends on the value of this resistor. If the resistance is low, the transistor switches on and the buzzer sounds. If the resistance is high, the transistor switches off and the buzzer stops.

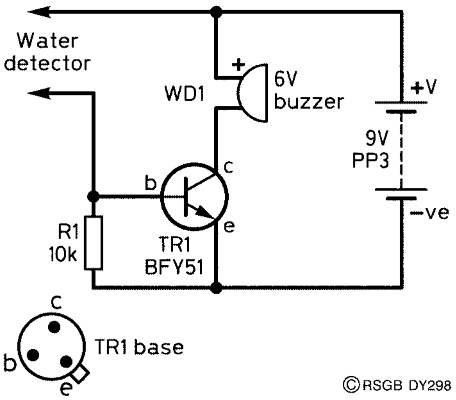
Water
detector
WD1
+
6V
buzzer
+V
9V
PP3
−ve
b
c
e
R1
10k
TR1
BFY51
c
b
e
TR1 base
©RSGB DY298

Figure 1 Circuit diagram of the water level detector

Construction

The unit is made in two separate sections – the detector board (**Figure 2**) and the main board (**Figure 3**). The main board is plain perforated board (without copper strips). The components are pushed through holes in the main board and connected together on the underside. The battery is held on with sticky tape. No on/off switch is used, as the circuit consumes virtually no current when it is not sounding the buzzer.

Unlike some buzzers, this type is polarised, and the red lead must go to the supply as shown in Figs 1 and 3.

The detector board is a piece of standard matrix board – the type with copper strips along one side. It measures 9 strips by 15 holes, and is wired in such a way that alternate strips are connected together, as Figure 2 shows. This gives a greater surface area for the water to touch, thus increasing the sensitivity of detection.

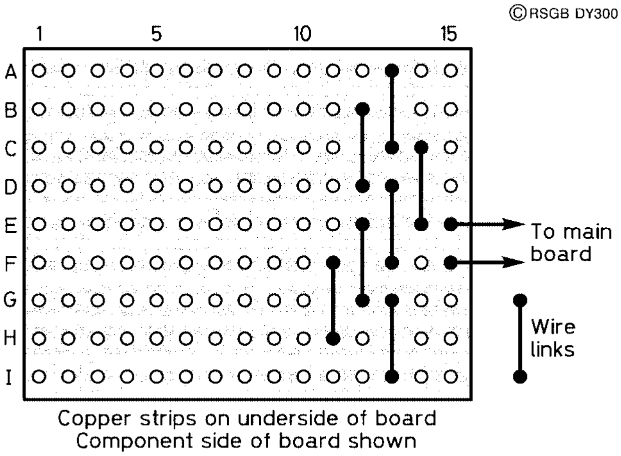
©RSGB DY300
1
5
10
15
A
B
C
D
E
F
G
H
I
To main
board
Wire
links
Copper strips on underside of board
Component side of board shown

Figure 2 Layout of the main board

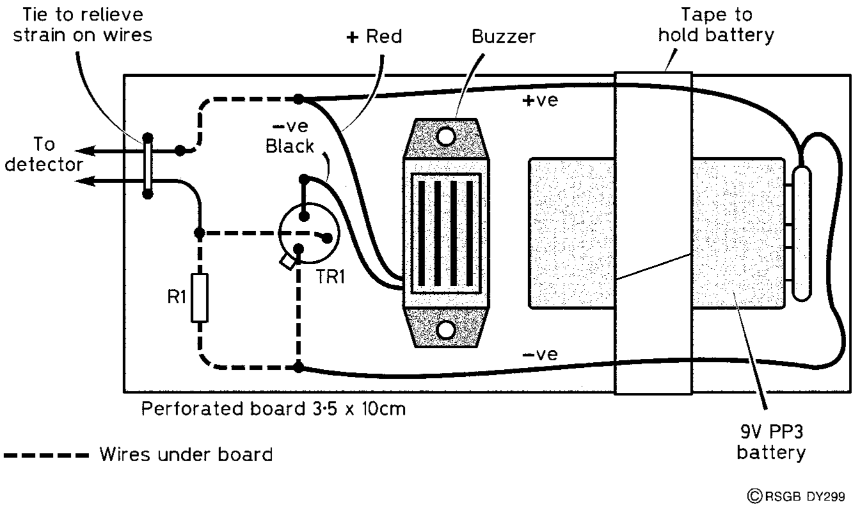
Tie to relieve strain on wires
+ Red
Buzzer
Tape to hold battery
+ve
To detector
−ve Black
TR1
R1
−ve
Perforated board 3·5 x 10cm
9V PP3 battery
Wires under board
©RSGB DY299

Figure 3 Water detector constructed from stripboard

The two boards are connected by ordinary twisted pair, enabling the detector to hang over the edge of the water container, with the buzzer circuit safely out of the way. When positioning the detector board, the left-hand edge (as shown in Figure 2) should face downwards, with the connecting-wire edge facing upwards. When the water level rises, the water causes current to flow between adjacent strips, and the higher the water, the lower the resistance between them. Eventually, the resistance will be low enough to turn on the transistor and sound the alarm.

Parts list

Resistor

R1	10 000 Ω (10 kΩ), 0.25 W

Semiconductor

TR1	BFY51

Additional items

WD1	6 V miniature buzzer (polarised)
	Battery PP3
	Battery connector
	Strip board 9 strips by 15 holes
	Perforated board 3.5 × 10 cm

78 A delta loop for 20 metres

Introduction

HF aerials take up a lot of room when they are straight. Space can be saved by bending them and, at the same time, giving them properties which are quite different from their original linear forms. The delta loops shown in the accompanying diagrams have the shape of the Greek upper-case letter delta (Δ) upside down. It can be upright, except that it is more practical to have the feed point at the bottom than at the top.

Putting the loop together

This is the design for a loop to be used on the 20 metre band. It is a very popular band and carries most of the amateur radio DX traffic. The aerial is light in weight and operates best if the top section is at least 30 ft above the ground, with the feeder point being about 6 ft above ground. All you need is some good wire, some polypropylene rope, some insulators and a method of matching the output of this 75 Ω balanced aerial to the 50 Ω unbalanced coaxial input of your transceiver or receiver.

Each of the three sides must be 7.2 m long. Using tent pegs or six-inch nails, mark out the three corners on the lawn, making sure that each pair of pegs or nails is exactly 7.2 m apart. Put the wire around the nails, together with two insulators on the wire for the two top corners. These insulators can be of the ‘dog bone’ variety, or of the home-made type, using a piece of flat plastic with holes in each end. The use of these is shown in **Figure 1**. Located at the positions of two of the tent pegs or nails, they must be secured at those points by any convenient means.

Now attach a support rope to the free end of each of the insulators, ready for suspension, as shown in the diagrams, from any convenient supports (house, tower, tree, etc.). The top section of the loop could be taped to a continuous piece of rope between the two supports. You might like to try this if you think it is easier. Then, when you find some dog-bone insulators, you can change the design and see if there is any noticeable change of performance.

Connecting to the radio

If you are the proud possessor of an aerial tuning unit (ATU) with a balanced output, all you will need is a length of twin cable soldered to the

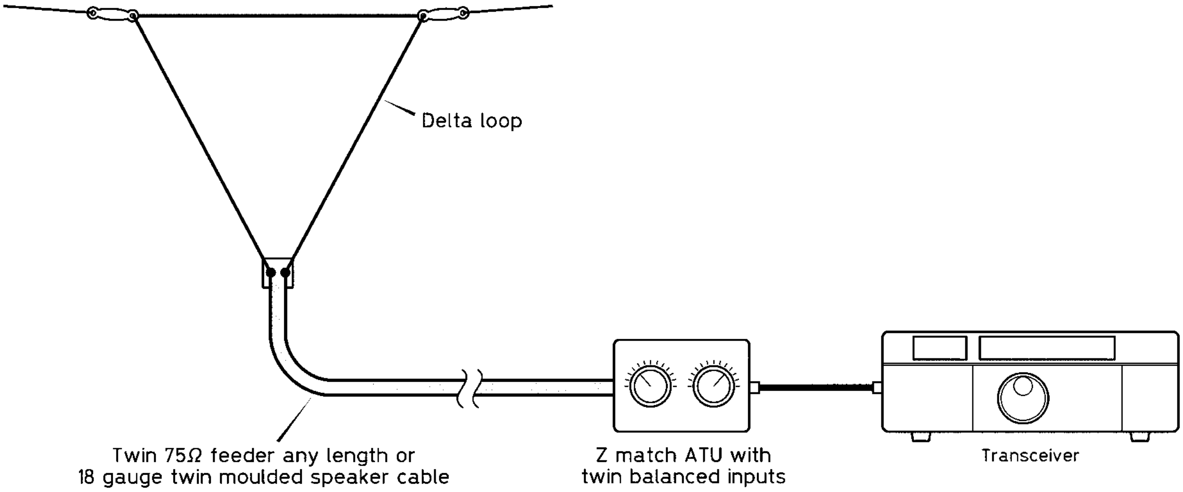
Delta loop
Twin 75Ω feeder any length or
18 gauge twin moulded speaker cable
Z match ATU with
twin balanced inputs
Transceiver

©RSGB DY302

Figure 1 Multiband version of the delta loop connected to the ASTU using 75 Ω twin feeder

ends of your delta loop and connected to the balanced output terminals on the ATU, as Figure 1 shows.

Another way is to buy a ferrite-cored 1:1 balun, and use it as indicated in **Figure 2**. This produces an aerial which will operate over the whole range of amateur and commercial short-wave frequencies, when used with an ATU.

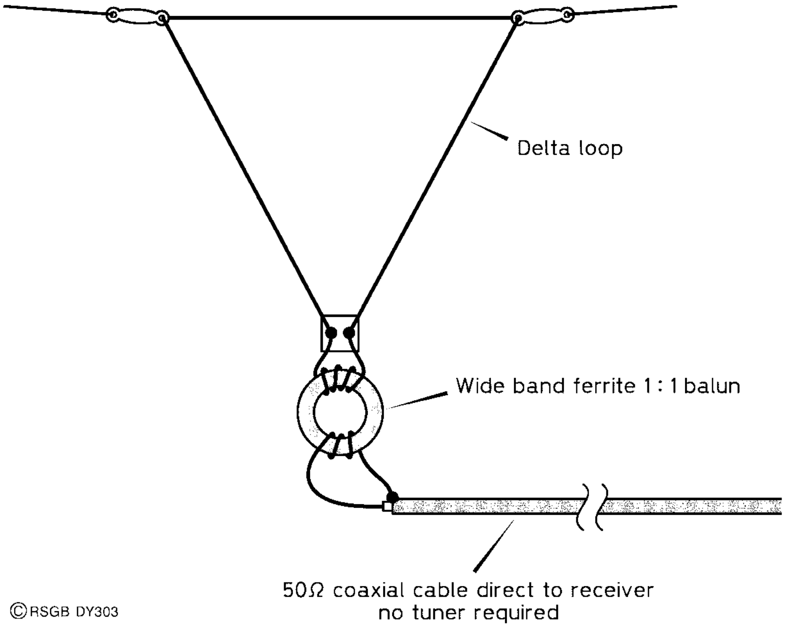
Delta loop
Wide band ferrite 1 : 1 balun
50Ω coaxial cable direct to receiver
no tuner required

©RSGB DY303

Figure 2 Multiband version of the delta loop using 75 Ω coaxial cable. This arrangement requires an ASTU if used with a multiband antenna

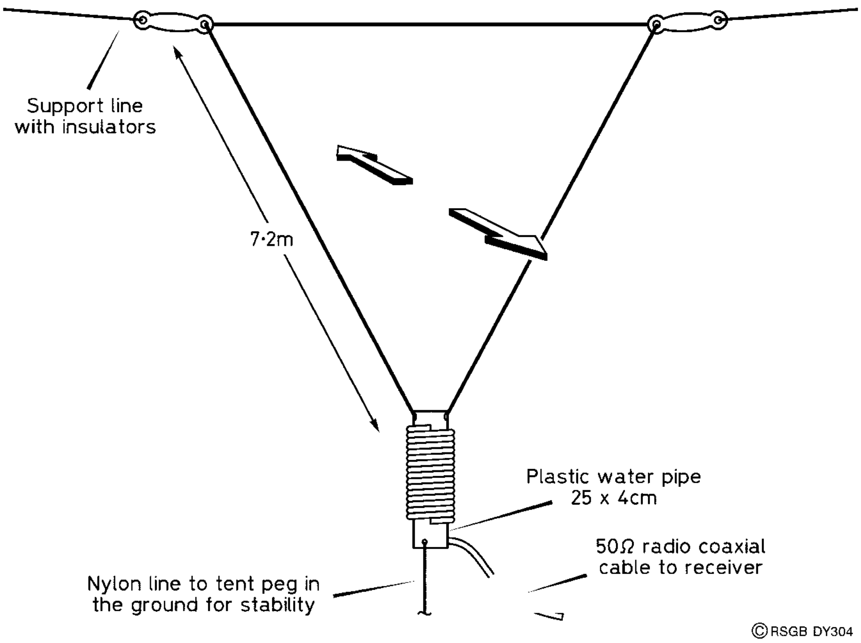

Figure 3 Single band version of the loop antenna using a coaxial balun. The direction of maximum signal strength is indicated by the arrows

However, since this aerial is designed to be a single-band type, there is a simpler answer to the problem of matching the aerial to your 50 Ω coaxial cable – **Figure 3** shows this. Make a tuned balun with a length of plastic water pipe, 25 cm long and 4 cm diameter and a piece of good-quality TV aerial cable. TV aerial cable has a characteristic impedance of 75 Ω, compared with the 50 Ω of the coaxial feeder that we usually use with amateur radio equipment. Use the type with a brown sheath and a closely knit earth braid, not the type having an earth foil inside the sheath.

The length we need for the balun is 3.8 m, but always allow between 3 and 4 cm extra for preparing the ends. Drill two small holes diametrically opposed in the top of the tube; these will be used to anchor the two ends of the delta loop, as shown in Figure 3. Drill another single hole in the bottom end of the tube, which will be used to anchor a nylon line going to the ground to add stability to the loop. Then, after drilling a pilot hole, drill a 5 or 6 mm hole near the top end of the tube, as shown in **Figure 4**. Prepare both ends of the coaxial cable, then feed one end into the tube, far enough for its ends to be soldered to the ends of the loop when the assembly reaches that stage.

Now, close-wrap the cable around the tube until only about 3 cm remain. Holding the cable tightly, drill another 6 mm hole beside the free end of the cable and feed it into this hole. At this point, feed the two ends of the loop into the top two holes and twist it back on itself. Figure 4 shows how this is done. Then, solder the ends of the coaxial balun to the ends of the loop.

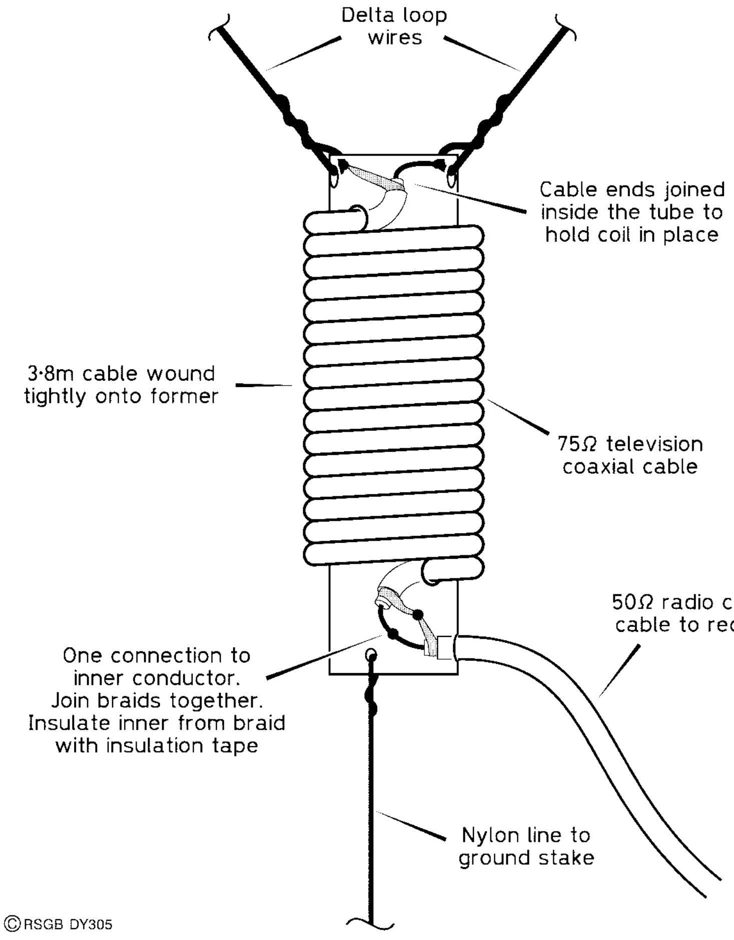
Delta loop wires
Cable ends joined inside the tube to hold coil in place
3·8m cable wound tightly onto former
75Ω television coaxial cable
50Ω radio coaxial cable to receiver
One connection to inner conductor. Join braids together. Insulate inner from braid with insulation tape
Nylon line to ground stake
©RSGB DY305

Figure 4 Construction of the balun

The bottom ends of the balun are then soldered to the 50 Ω coax which goes to your shack and to the transceiver.

Hoist the aerial into position carefully, being careful not to pull too hard on the support lines. Then, take the nylon line from the bottom of the balun to a peg in the ground. This adds stability to the aerial.

Using the delta loop

It is a directional aerial, as Figure 3 shows. It produces maximum power (and has maximum receive sensitivity) along a direction perpendicular to its own plane. Don't be too concerned with which direction it is pointing at first. Give it a try 'on the air' and see how it performs. Then you can contemplate how to point it in your favourite direction, to the USA, or Australia, for example.

Experimenting

You may want to enclose the balun in some sort of weatherproof container. Plastic ice cream containers are favourites for this sort of job. Seal all the holes where wires enter it with silicone sealant or self-amalgamating tape.

If you have used ordinary single-strand or multi-strand wire for your aerial, it will stretch over time under its own weight and that of the balun. Its operating frequency will fall slightly as a result. If you notice a significant difference, then dismantle it, remeasure and fix the sides and erect it again, perhaps facing a different direction. You can buy pre-stretched or hard-drawn wire for such purposes, if you feel that periodic tweaking of your aerial is a chore.

You can make a delta loop for different frequencies simply by scaling the lengths of wire for the loop and for the balun according to the design frequency. If your maths isn't quite up to this, enlist the help of someone well versed either in maths or aerial design!

79 A simple desk microphone

Introduction

A hand-held transceiver is usually the first type used by most novices. It is ideal when used out of doors, in the shack or mobile. However, using one in the shack with an outside aerial is not the best of solutions.

This article describes the design of a desk microphone which can be used either for transceivers using a common microphone and push-to-talk (PTT) line, or for those using a separate PTT line.

It is a simple design, using the barrels of two plastic pens, the lid of an old coffee jar, and the plastic sheath of a DIN plug.

Construction

The end caps and the ink tubes should be removed from two ball-point pens, leaving only the two plastic shells. One of them is used for the vertical microphone support, while the other is used to hold the microphone element.

The vertical support should have the pointed end shaped so that it holds the microphone support at a convenient angle, as **Figure 1** shows.

The base should be prepared next – see **Figure 2**. Four holes are required. Two, A and B, are small holes for the cable to be fed through. A third, C, is slightly larger and is used to take one of the plastic end caps. The fourth hole, D, is determined by the type of switch that you want to use for the PTT function.

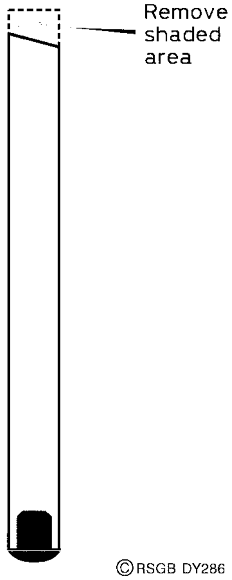
Remove
shaded
area
©RSGB DY286

Figure 1 Vertical support made from ball-point pen shell

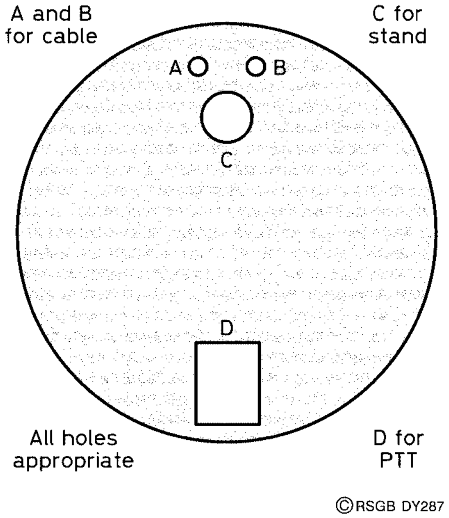
A and B
for cable
C for
stand
A
B
C
D
All holes
appropriate
D for
PTT
©RSGB DY287

Figure 2 The microphone stand base

Some epoxy resin such as Araldite Rapid is used to mount the separate parts. Mix only a small amount at a time, and glue each part of the base assembly and allow to harden before moving on. Be sure to read and follow the instructions for the use of the glue to prevent accidents.

Firstly, glue the end cap to the base, feeding it underneath and pushing it up through hole C. Then, the vertical pen body is placed over the protruding end cap and glued into place, making sure that the top (shaped) end is pointing in the right direction to hold the other plastic tube carrying the microphone. While the glue hardens, place the DIN plug sheath on the second pen body. Place and glue this pen body on the shaped top of the vertical support, making sure that the whole assembly is supported while the glue sets. What you should have now ought to resemble **Figure 3**, but without the microphone and wiring. Leave everything to set for 24 hours.

Any switch of the press-to-make variety can be used for your PTT switch, but choose one which does not apply too much force to operate! If you find the whole assembly is too light and slides over your table, some plasticine can be pushed into the base when all the wiring has been completed.

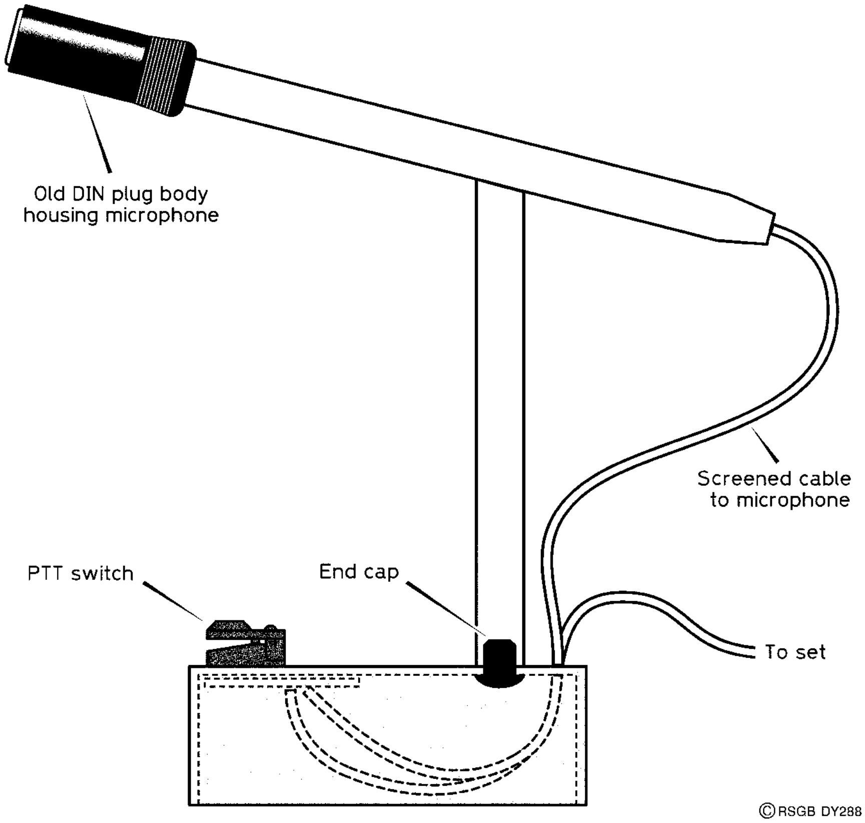
Old DIN plug body housing microphone
Screened cable to microphone
PTT switch
End cap
To set
©RSGB DY288

Figure 3 Complete microphone stand assembly

The circuit

A small electret condenser microphone is used, powered by a DC supply. Most hand-held radios use a single screened lead for both the PTT line and the microphone audio lead. In such cases, the circuit of **Figure 4** will operate the PTT function. Note that the PTT switch is in series with the audio lead from the microphone. If you find that the PTT switch does not switch your radio into transmit, reduce R1 from its normal value of 33 kΩ to 27 kΩ. You should find that the original value works with most hand-helds.

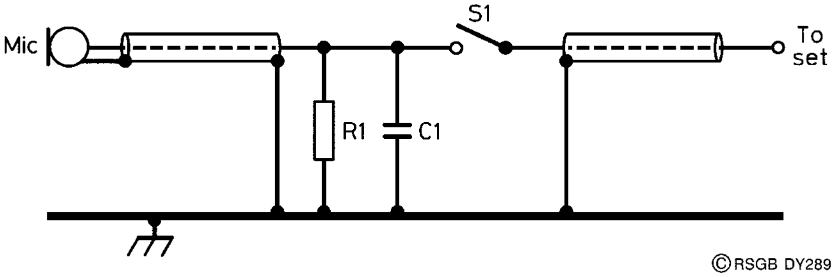

©RSGB DY289

Figure 4 Wiring diagram for hand-held radios

For radios with a separate PTT lead, the circuit of **Figure 5** is used. However, a power supply is needed for the electret microphone. This can be a PP3 battery, or can (in most cases) be derived from the microphone socket on the transceiver. You will need to consult the makers' handbook for this information. Figure 5 shows the PTT switch wired in a 'ground-to-transmit' configuration, which is correct for most base-station transceivers. If you're in any doubt about what your radio needs, consult an experienced friend.

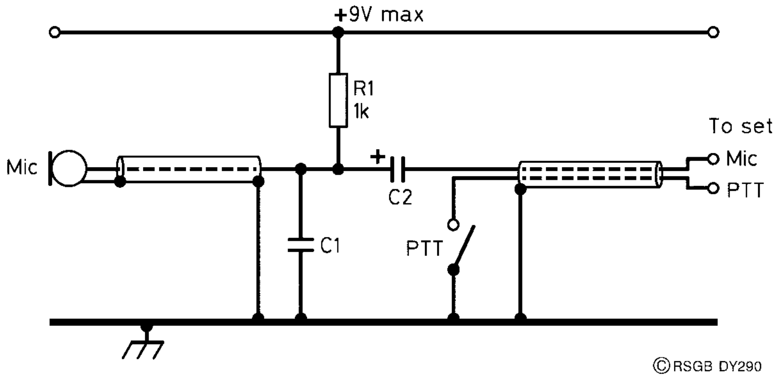

©RSGB DY290

Figure 5 Wiring diagram for radios which use separate PTT lines

The last connections

The microphone element should be connected with screened cable. Tape (masking or insulating) should be wound around the electret insert until it is a snug fit inside the DIN plug sheath. Feed the screened cable through the barrel of the pen until it emerges from the far end. Poke it through hole B and make the connections under the coffee jar lid. The components can all be mounted on the PTT switch or on a small piece of Veroboard mounted under the top surface. The output lead emerges through hole A and is of sufficient length to reach your transceiver. A suitable plug needs to be fitted to it.

Parts list

Resistors: all 0.25 W carbon, 5% tolerance

R1 (Figure 4)	33 000 Ω (33 kΩ)
R1 (Figure 5)	1000 Ω (1 kΩ)

Capacitors

C1 (Figure 4 and 5)	0.1 microfarads (0.1 μF) disc
C2	10 microfarads (10 μF) electrolytic

Additional items

S1	Push-to-make
Mic	Electret microphone (Maplin type FS34W)
Screened cable	As required
Plug	As required by your radio
Veroboard	If required
PP3 battery and clip	If required

Two old plastic ball-point pens
Lid of coffee jar
Plastic sheath from DIN plug
Araldite or similar epoxy resin glue

80 Morse oscillator

Introduction

This is not the simplest Morse oscillator to build, but it differs from the simple circuits in that it produces a pure note, not a coarse, rasping sound. People who have practised Morse using a non-sinusoidal oscillator sometimes find that they have trouble copying Morse code with a pure tone. As the pure tone is the correct way to receive Morse code, it is important that you should learn to listen to the code from a pure oscillator – so here's one!

The twin-T

There are many oscillator circuits, and there are many variations of the twin-T oscillator that we are going to use. **Figure 1** shows one version of a very useful circuit. All oscillators must have positive feedback in order to work. The feedback determines the frequency of the note produced by the oscillator.

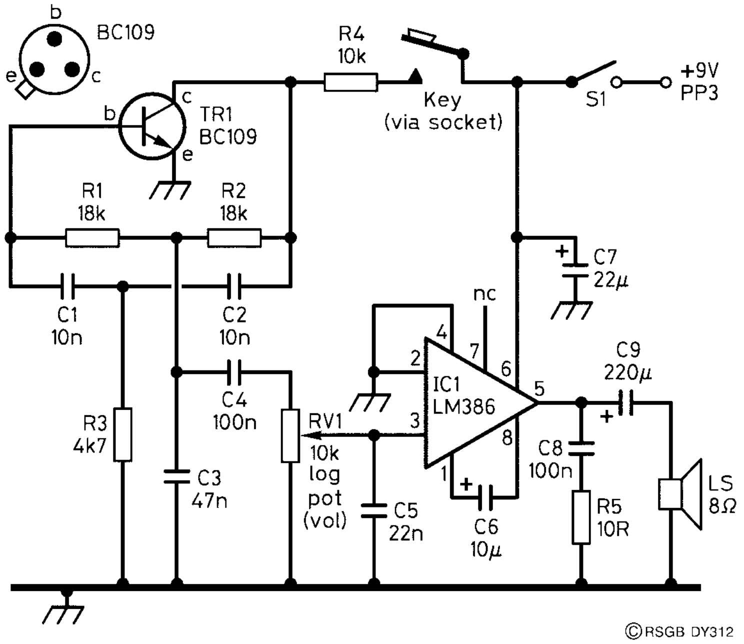
b
BC109
e
c
R4
10k
Key
(via socket)
S1
+9V
PP3
b
c
TR1
BC109
e
R1
18k
R2
18k
C1
10n
C2
10n
C7
22µ
nc
4
7
2
6
IC1
LM386
5
C9
220µ
3
8
C4
100n
RV1
10k
log
pot
(vol)
R3
4k7
C3
47n
C5
22n
1
C6
10µ
C8
100n
R5
10R
LS
8Ω
©RSGB DY312

Figure 1 Circuit diagram of the sinewave oscillator and amplifier

Here, the feedback circuit looks like two letter Ts. If you look at Figure 1, one T is formed by R1, R2 and C3, the other by C1, C2 and R3 – hence the name ‘Twin-T’. Notice that the two Ts are connected in parallel between the collector and base of TR1, so any signal appearing at the collector is fed into both Ts. What emerges is then fed back into the base, producing in turn a signal at the collector. And so it goes on, producing a sine wave output.

The oscillator output is fed into an integrated circuit amplifier for output via a small loudspeaker.

Putting it together

The prototype was constructed on plain matrix board (the type *without* the copper strips), as shown in **Figure 2**. The components have their leads pushed through the holes in the board, and connections are made underneath.

Build the amplifier circuit first, using a socket for IC1. Connect the 9 V supply and touch pin 3 with an ordinary piece of wire. If a buzz is heard from the speaker, all should be well. If not, check your circuit and make changes until it does.

Build the oscillator circuit, and connect its output to the volume control VR1 via C4. Set VR1 half-way along its travel and switch on. A note should be heard from the speaker when the Morse key is depressed. The component values making up the twin-T determine the frequency of the note. Try varying them if you think your note is too high or too low. Whatever changes you make, either to the resistors or the capacitors, always ensure that R1 = R2 and that C1 = C2.

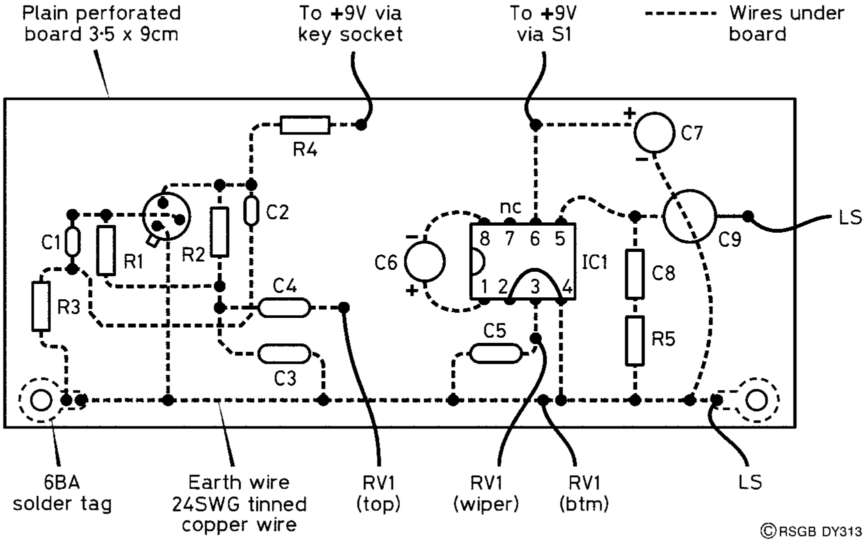
Plain perforated board 3·5 x 9cm
To +9V via key socket
To +9V via S1
------- Wires under board
R4
+
C7
−
C2
nc
C1
R1
R2
8 7 6 5
C6
IC1
C8
LS
C9
−
+
1 2 3 4
R3
C4
C5
R5
C3
6BA solder tag
Earth wire 24SWG tinned copper wire
RV1 (top)
RV1 (wiper)
RV1 (btm)
LS

Figure 2 Component layout and interconnection diagram

©RSGB DY313

Parts list

Resistors: all 0.25 W carbon film

R1, R2	18 000 ohms (18 kΩ)
R3	4700 ohms (4.7 kΩ)
R4	10 000 ohms (10 kΩ)
R5	10 ohms (10 Ω)

Capacitors

C1, C2	10 nF (0.01 μF)
C3, C8	47 nF (0.047 μF)
C4	100 nF (0.1 μF)
C5	22 nF (0.022 μF)
C6	10 μF electrolytic, 16 V WKG
C7	22 μF electrolytic, 16 V WKG
C9	220 μF electrolytic, 16 V WKG

Semiconductor

TR1	BC109

Integrated circuit

IC1	LM386

Additional items

VR1	10 kΩ log potentiometer with switch
LS	8 Ω loudspeaker

Matrix board 3.5 × 9 cm
Small jack socket for key input
Box
PP3 battery and clip

81 A simple 6 m beam

Introduction

The attraction of building your own aerials is an abiding feature of our hobby. You can *buy* almost any shape or size of aerial, but one you have made yourself can often work every bit as well as a commercial device costing ten times as much.

The design

This aerial, designed for use on the 6 m band, is essentially a two-element Yagi, with the elements bent in order to reduce the physical size. It is known

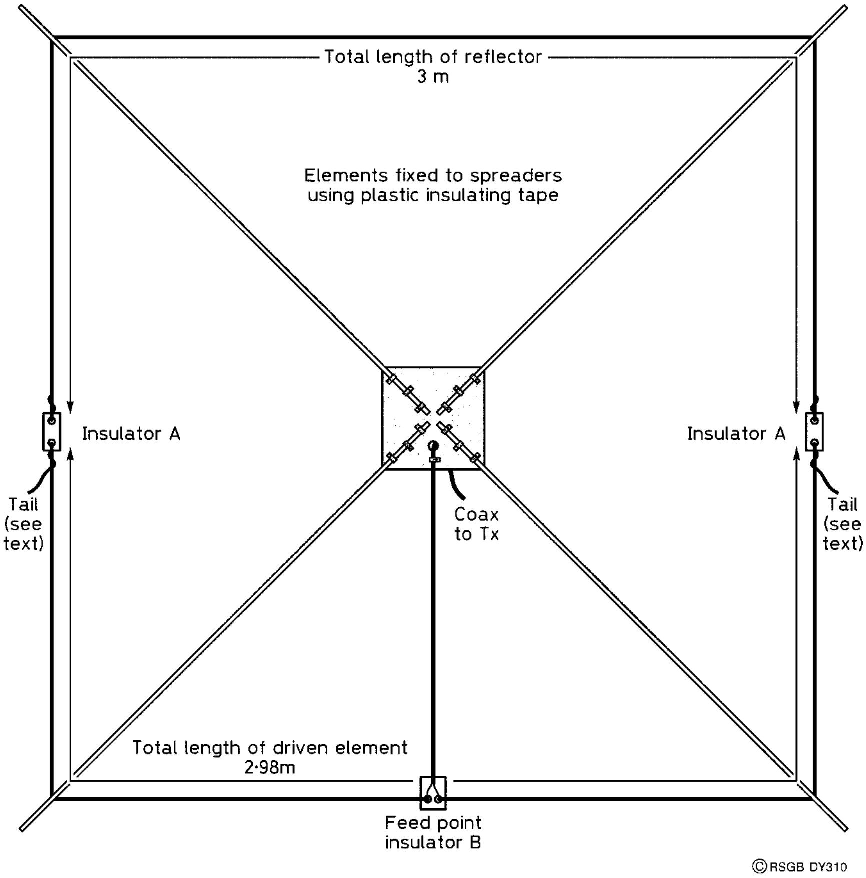
Total length of reflector
3 m
Elements fixed to spreaders
using plastic insulating tape
Insulator A
Insulator A
Tail (see text)
Tail (see text)
Coax to Tx
Total length of driven element
2·98m
Feed point
insulator B
©RSGB DY310

Figure 1 Plan view of the complete 50 MHz VK2ABQ antenna

as the VK2ABQ beam, and was designed originally for the 20, 15 and 10 m bands, principally because of its space-saving qualities. It is made using a wooden frame and wire elements, and is ideal for portable operation.

Tools ready?

The beam is shown in **Figure 1**. The driven element is the one whose centre is fed by the coaxial cable, and lies between the two insulators marked A and the feedpoint at B. The reflector is also anchored at the points A, and lies over the upper half of the frame.

The wooden centrepiece is used to support the cross-pieces and to mount the aerial on the mast, using a common shelf bracket. The cross-pieces, known as *spreaders*, can be wooden canes or dowelling, and are mounted to the centrepiece using cable clips and adhesive. **Figure 2** shows how this is done.

If the aerial is to be a permanent installation, the spreaders should be weatherproofed using a good-quality exterior varnish. The wire elements are PVC covered and fed through holes in the spreaders.

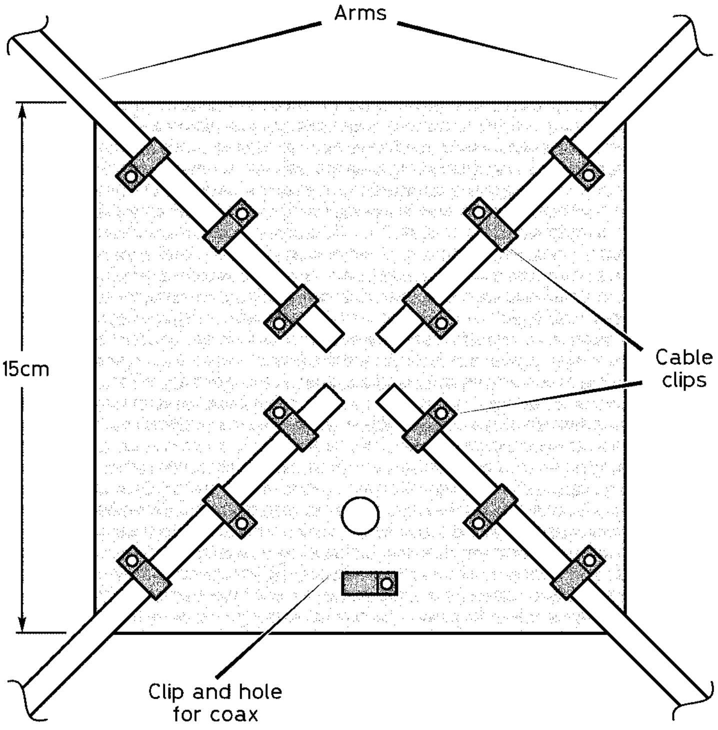

Figure 2 Centre support piece

©RSGB DY309

The end insulators are made of drilled perspex, and the wire passed through the two holes and twisted, as shown in **Figure 3**.

Note: if you have not drilled holes in perspex before, take care! The drill bit must be well-lubricated because it generates a lot of heat (enough to melt the perspex and jam the drill). Never turn the drill for more than a few seconds at a time, and moisten the bit in between. Start with a pilot hole and use bits of increasing size until the hole is the size you want.

Another perspex insulator is used to secure the feeder to the aerial, as shown in Figure 3. The feeder then passes directly to the centrepiece, where it is fastened with a cable clip and then passes down the mast.

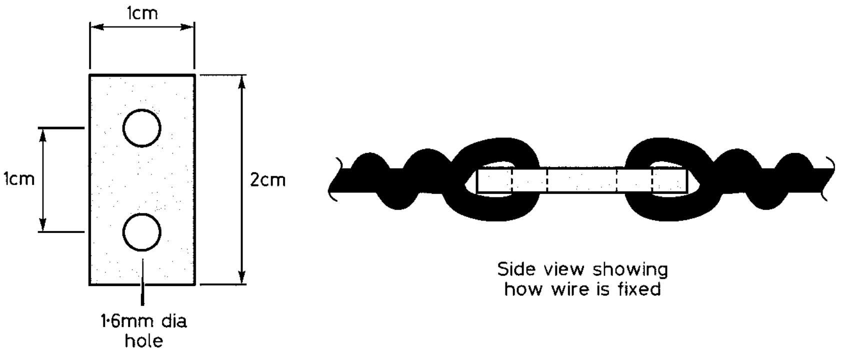
1cm
1cm
2cm
1·6mm dia hole
Side view showing how wire is fixed

Insulator A

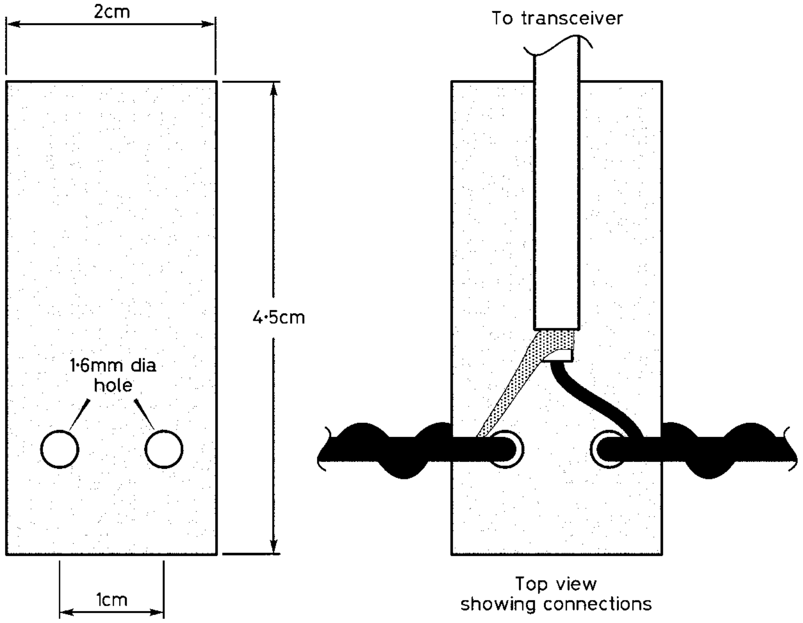
2cm
To transceiver
4·5cm
1·6mm dia hole
1cm
Top view showing connections

Insulator B

©RSGB DY311

Figure 3 Details of insulators

Adjustment

The driven element (A–B–A) in Figure 1 is fixed to the end insulators in such a way as to have 'pigtails' which are about 10 cm long. Using an SWR meter between the aerial and the transmitter, trim the pigtails equally at each end for minimum SWR in that portion of the 6 m band in which you plan to work. Make sure that the transmitter is off when you trim the ends, as high voltages can be present there. Always listen on the frequency before you transmit and, when you do, ask if the frequency is in use and identify yourself. Use as little power as possible.

The prototype had an SWR of 1.2 at 50.2 MHz and performed well. If you look at Figure 1, which is a view of the aerial from above, you will immediately see that when the aerial is horizontal, it radiates with horizontal polarisation in a direction from the top of the page to the bottom. A metal pole or mast can be used for horizontal polarisation, but if you intend to use the aerial for vertical polarisation, it is much better to use a wooden or fibreglass pole.

Portable use

If you plan to operate portable with this aerial, the only real modifications you need are to the centrepiece and how it supports the spreaders. Instead of using glue and cable clips, nuts and bolts through the spreaders and centrepiece would allow the spreaders to be 'hinged' closed for transport.

Materials

Centrepiece	Hardwood, 15 × 15 × 25 mm
4 spreaders	110 cm long (cane or 6 mm dowel)
Wire	1.5 mm PVC-covered copper
50 Ω coaxial cable	RG58 or similar
Cable clips	6 mm (12 off)
Varnish	Polyurethane for waterproofing
Tape	Self-amalgamating, to waterproof all soldered joints
Insulators	See text

82 An integrated-circuit amplifier

Introduction

A simple audio-frequency amplifier is a very useful building-block in many more advanced electronic circuits. It is also a very useful piece of test equipment. To have one spare in the shack can be a life-saver at times.

Planning

Get into the habit of planning your project. How do you want it to look when it is finished. Do you want it in a box? Do you want it 'open-plan', with all the components on view?

The 'minimalist' approach to any project is simply a front panel and a baseboard, on to which all the components are fitted. In this case, the front panel would accommodate the loudspeaker, the volume control and the input and output jack sockets, and the baseboard would support the circuit board. Once you have decided these things, and know the size of the board, speaker and volume control, you can decided how big the panel and baseboard should be.

The amplifier

Instead of building the amplifier from discrete components (i.e. transistors), as was done in the project *An audio-frequency amplifier* (which you will find elsewhere in this book), we are going to use an integrated-circuit (IC) amplifier. External resistors and capacitors are still needed but, compared with the number of components *inside* the chip, these are very few indeed!

The IC we are going to use is the TBA820M. The circuit may look complicated, but with the use of a matrix circuit board it becomes quite simple. **Figure 1** shows the circuit diagram and **Figure 2** the layout on the board. The external connections to the PCB are shown in **Figure 3**.

To avoid having a separate switch to switch off the power supply when the amplifier is not being used, a volume control which incorporates a switch is used.

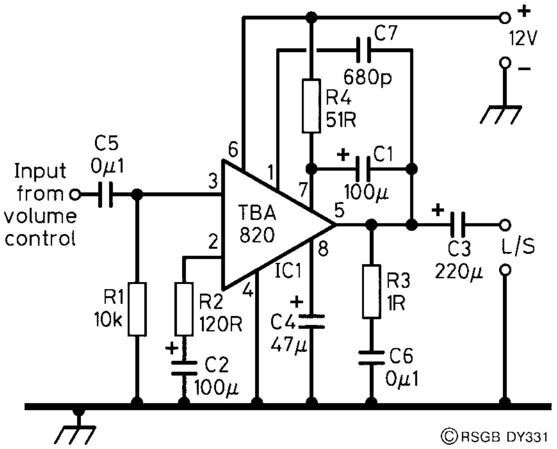
+
12V
−
C7
680p
R4
51R
C5
0µ1
Input
from
volume
control
6
3
1
7
+
C1
100µ
TBA
820
5
2
8
IC1
+
C3
220µ
L/S
R1
10k
R2
120R
4
+
C4
47µ
R3
1R
+
C2
100µ
C6
0µ1

Figure 1 Audio amplifier, circuit diagram

©RSGB DY331

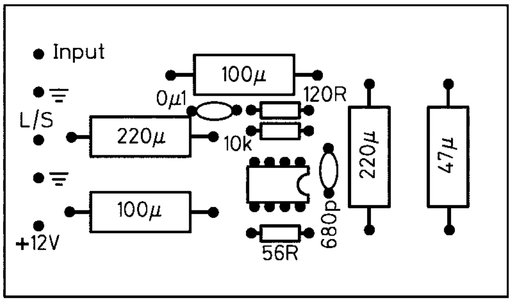
Input
L/S
+12V
100µ
0µ1
120R
220µ
10k
100µ
680p
56R
220µ
47µ

Figure 2 Audio amplifier, component layout

©RSGB DY332

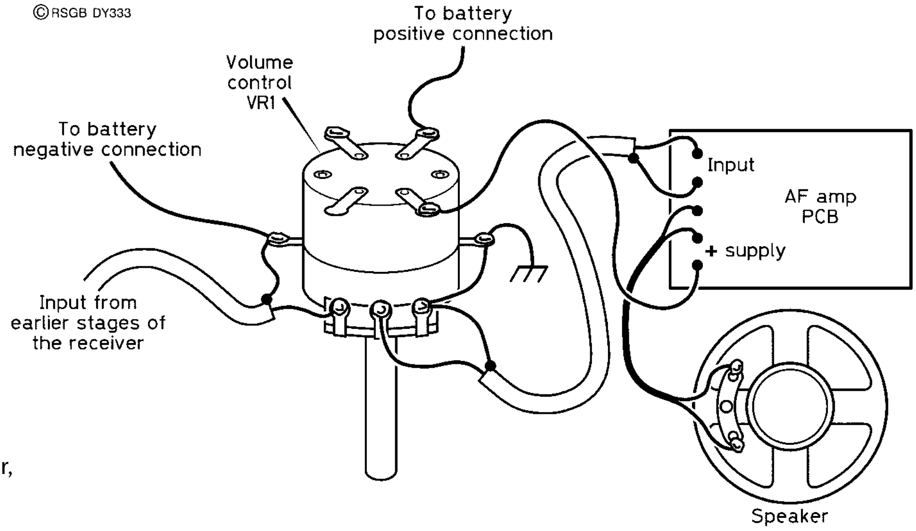
To battery
positive connection
Volume
control
VR1
To battery
negative connection
Input
AF amp
PCB
+ supply
Input from
earlier stages of
the receiver
Speaker

©RSGB DY333

Figure 3 Connecting the circuit board to the speaker, volume control/switch and battery

What power supply?

If the amplifier is to be used only for short periods, as a test instrument, for example, then it can be run from a PP3 battery, which can be mounted on the baseboard. If you intend to use it often, then a connection to an external power supply is preferable. For this option, you may want to consider fixing two terminals to the baseboard for this connection.

83 A novice ATU

Introduction

Having an aerial tuning unit (ATU) is always useful. It is used for adjusting the aerial impedance, as 'seen' by the transceiver, to be the same as that of the transceiver itself, usually 50 Ω. This process is called *matching*, and ensures that the receiver and the power amplifier (PA) stage of the transmitter work efficiently.

This design is due to the late Doug DeMaw, W1FB, and uses readily available components. It will handle up to about 5 W, and operates over the frequency range 1.8 to 30 MHz.

Circuit evolution

The basic circuit of one type of ATU is shown in **Figure 1**. On transmit, the input signal at L1 is coupled into L2, which forms a resonant circuit with C1. The signal across the resonant circuit is fed to the aerial by C2. The combination of L2 and C1 helps to remove signal harmonics, because they are not at the resonant frequency and are shunted to earth. On receive, only those signals which are within the pass-band of L2 and C1 will pass into the receiver. This improves receiver performance by rejecting many out-of-band signals which the simpler receiver doesn't like.

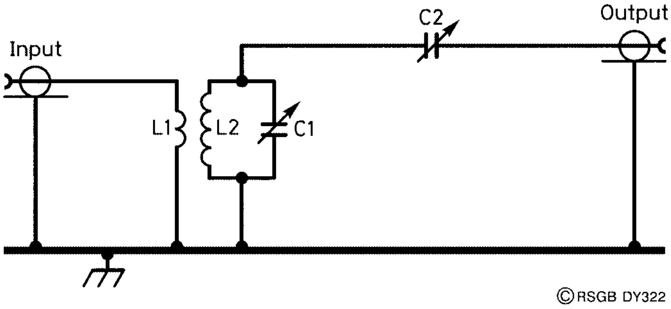

Figure 1 Basic circuit of the ATU

For this design, the all-important resonant frequency is chosen to be in the 80 metre band. If we want to make the circuit operate on other bands, we need to change either L2 or C1. It is easy to change L2 by using a set of switched inductors (L3–L5) as shown attached to S1 in **Figure 2**. Forget about S2 and L6 for the moment.

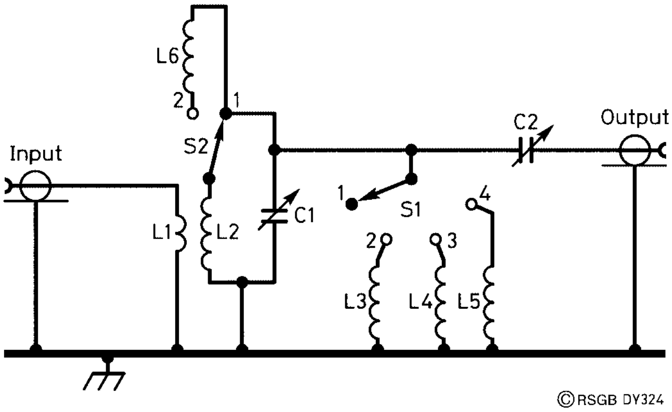

Figure 2 L6 added for 160 m

When S1 is in position 1, the circuit reverts to that of Figure 1. S2 should be brought into circuit only when S1 is in position 1. Looking at the way the circuit is drawn, you can see that when S2 is in position 2, we have two coils in series, which is equivalent to adding more turns to L2. More turns means more inductance, which lowers the resonant frequency still more, and gives coverage of the 160 m band.

Rather than having to remember to flick S2 to the correct position *and* move S1 to position 1 when we want to operate on 160 m, both functions can be combined into one rotary switch with two wafers. This circuit is shown in **Figure 3**.

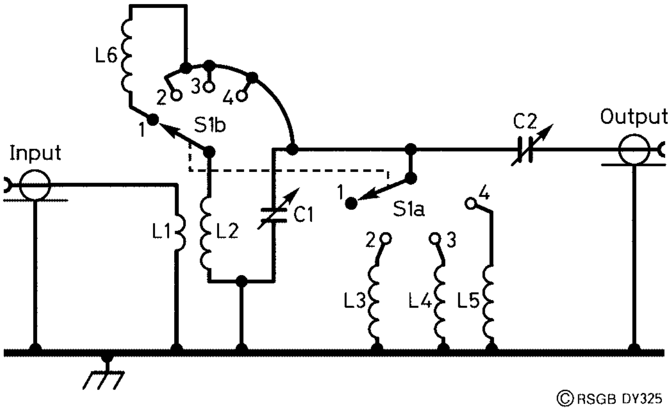
L6
2
3
4
1
S1b
Input
L1
L2
C1
C2
Output
1
S1a
4
2
3
L3
L4
L5
©RSGB DY325

Figure 3 Two-pole switch for all bands

In all positions of S1a *except* position 1, the extra coil, L6, is shorted out. The two halves of the switch, S1a and S1b, move together, as the shaft is turned, so both halves are in position 1 at the same time, position 2 at the same time, and so on. The two switches, S1 and S2, are said to be *ganged*.

Construction

L1 is formed by winding four turns of 22 SWG enamelled copper wire over L2, as in **Figure 4**. L2 already exists on the purchased former. After scraping the enamel off the ends of the wire (with a sharp knife or sandpaper), one end of L1 must be soldered to one end of L2, as shown in Figure 4. The free end of L1 goes to the transceiver aerial socket.

All components except the capacitors are assembled on the switch. Note that the rotor of C1 is earthed, but neither side of C2 is earthed. This means that the metal shaft of C2 is not earthed, and touching it will detune the ATU. Using a plastic knob for C2 will minimise this effect. If you decide to use a metal box, take precautions to ensure that no part of C2 is in electrical contact with the box.

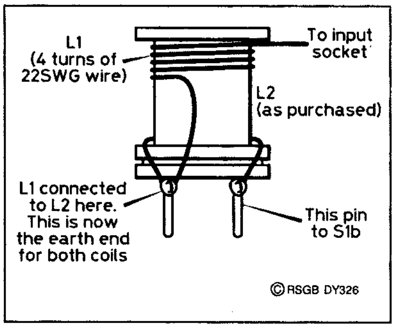
L1
(4 turns of
22SWG wire)
To input
socket
L2
(as purchased)
L1 connected
to L2 here.
This is now
the earth end
for both coils
This pin
to S1b
©RSGB DY326

Figure 4 L1 is wound over L2 as shown

The capacitors are mounted using M2.5 screws. Make sure that the screws do not foul the vanes of the capacitor. If your screws are too long, a few washers between the box and the capacitor will solve the problem!

In use

The best indication of a good match is obtained with a standing-wave meter between the ATU and the transceiver. The controls are adjusted alternately to 'feel' your way to a better and better match.

For receive-only use, the same alternate adjustments are used, watching for the maximum signal strength on the S-meter or, for a very weak signal, making the signal from the loudspeaker as large as possible.

Parts list

Capacitors

C1, C2	350 pF variable

Inductors

L1	See text
L2	27 μH
L3	10 μH
L4	2.2 μH
L5	1 μH
L6	65 μH

Additional items

SO239 sockets (2 off)

S1	2-pole 6-way rotary
Box	as required

Plastic knobs (2 off)
Stick-on feet (4 off)
M2.5 screws for capacitors
Screws, nuts and washers for mounting the input and output sockets.

84 CW QRP transmitter for 80 metres

Introduction

This is a relatively simple transmitter design having an output of 1 W. The design is not new, having been described before in other amateur radio publications. The components are all available new and the total cost should not exceed £15.

The circuit

Like other simple transmitters (see *An 80 Metre Crystal-Controlled CW Transmitter* and *A Breadboard 80 m CW Transmitter* elsewhere in this book) this one is crystal controlled. This assures frequency stability, but limits the usefulness of the transmitter. The key to increased frequency coverage without a conventional Variable Frequency Oscillator (VFO) is the use of a low-cost 3.58 MHz *ceramic resonator*. The 'pulling' range of a 3.58 MHz ceramic resonator covers the UK novice 80 m sub-band and some of the CW segment below 3.525 MHz.

A ceramic resonator is like a crystal, but not quite as stable in frequency. Its main advantage is its large pulling range.

The block diagram is shown in **Figure 1**. It is very similar to a crystal-controlled transmitter, and includes an oscillator, buffer and final amplifier.

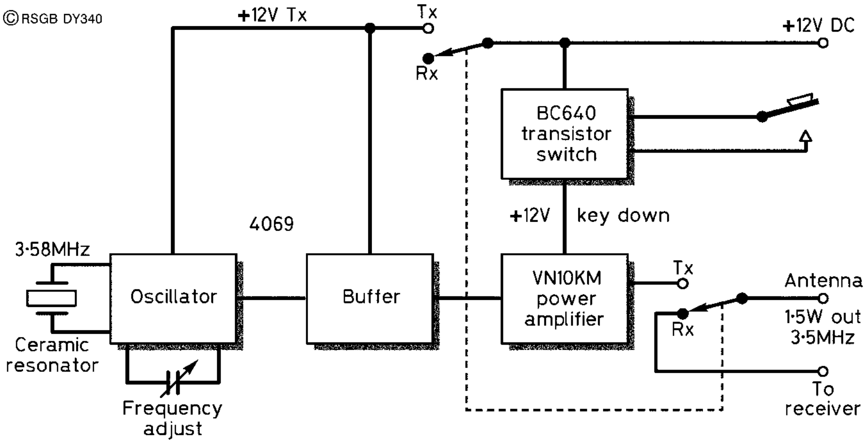
©RSGB DY340
+12V Tx
Tx
Rx
+12V DC
BC640 transistor switch
+12V
key down
4069
3·58MHz
Ceramic resonator
Oscillator
Buffer
VN10KM power amplifier
Frequency adjust
Tx
Rx
Antenna
1·5W out 3·5MHz
To receiver

Figure 1 Block diagram of the ceramic resonator controlled 80 m CW transmitter

This amplifier is keyed, the oscillator remaining running all the time. This improves frequency stability because the oscillator is not being continuously stopped and started by the key. It is switched off while receiving, though, to avoid interference with the received signal. Transmit/receive switching is accomplished by a panel-mounted switch controlling both the aerial, oscillator and buffer switching.

Figure 2 shows the transmitter circuit diagram. An unusual aspect of this transmitter is the use of a digital CMOS integrated circuit (IC) type 4069 for the buffer and oscillator stages. The IC houses six inverters, four of which are used in the circuit. One is used as the oscillator, two are used for the buffer stage, and the fourth provides an output for a direct-conversion receiver, should one be added at a later date.

The frequency of the oscillator is changed by varying the capacitance in the ceramic resonator circuit. This is provided by VC1.

The power amplifier (PA) is a small MOSFET (metal oxide semiconductor field-effect transistor), TR1. This is capable of providing an output power of 2 W but, in this circuit, it is run conservatively to give 1.5 W. The output can be varied by changing the resistance (R5 + R6) in the gate circuit. Attempts

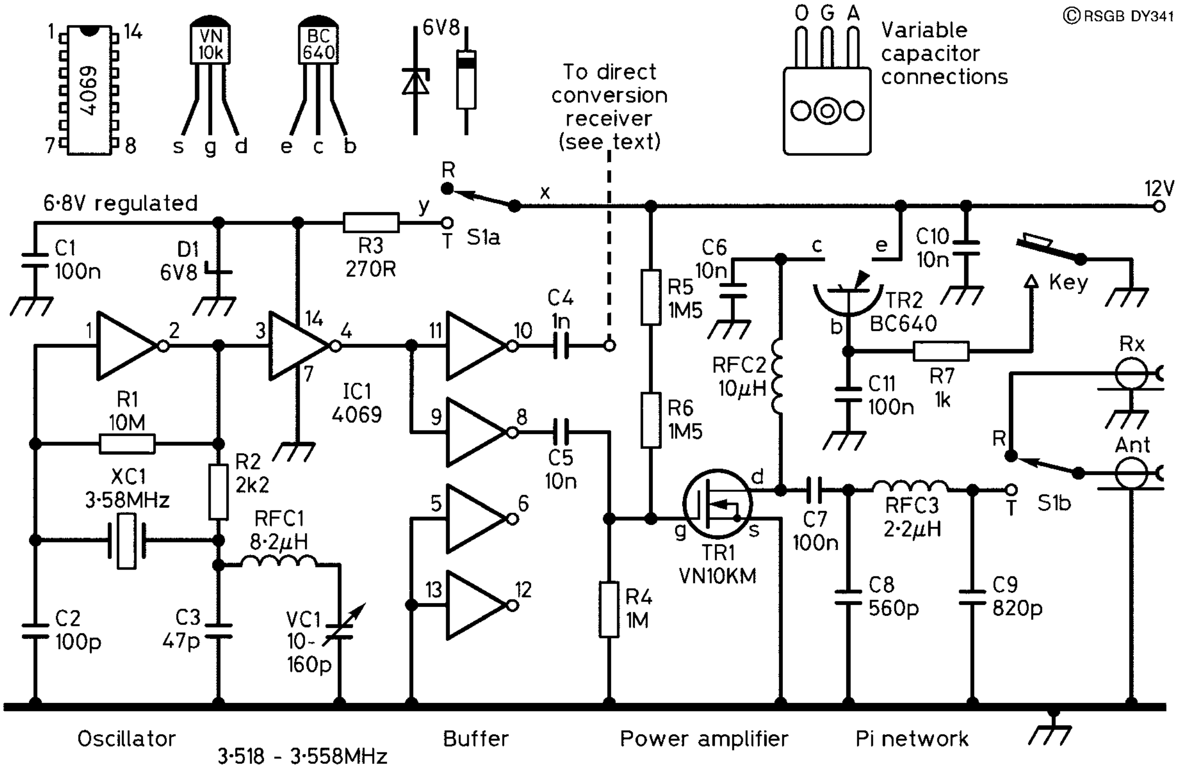
©RSGB DY341
4069
VN 10k
s g d
BC 640
e c b
6V8
To direct conversion receiver (see text)
O G A
Variable capacitor connections
6·8V regulated
C1 100n
D1 6V8
R3 270R
R
y
x
T
S1a
12V
C4 1n
R5 1M5
C6 10n
c
e
C10 10n
TR2 BC640
Key
b
R7 1k
C11 100n
RFC2 10μH
Rx
Ant
R
S1b
T
R6 1M5
d
g
s
TR1 VN10KM
C7 100n
RFC3 2·2μH
C8 560p
C9 820p
R4 1M
IC1 4069
R1 10M
XC1 3·58MHz
R2 2k2
RFC1 8·2μH
C2 100p
C3 47p
VC1 10-160p
C5 10n
Oscillator
3·518 - 3·558MHz
Buffer
Power amplifier
Pi network

Figure 2 Circuit diagram of 80 metre transmitter

to raise the output power by decreasing the values of these resistors may result in immediate MOSFET failure.

A pi-network (C8, RFC, C9) provides impedance matching to 50 Ω, together with harmonic suppression. Like all inductors in this transmitter, the pi-network inductor is a pre-wound RF choke. A pi-network is so called because the components are arranged in the shape of the Greek letter pi (π).

Keying is carried out by a pnp transistor switch, TR2. Closing the key earths the base and supplies 12 V to the collector of TR2 and to the drain of the MOSFET, TR1, allowing the PA to operate.

Construction

You must house your transmitter in a metal box, to avoid hand-capacity effects and the radiation of spurious frequencies. Size is not important, provided it is large enough to accommodate the transmitter without cramping the components. You may want to allow space for future additions such as a direct-conversion receiver, break-in keying, sidetone or a small power amplifier. A good size is 5 × 15 × 15 cm. You can make your own box, buy it, or even use a biscuit tin!

Front and rear panel connectors can be fitted first. The choice of these is a personal matter, but a good working choice would be:

(a) Power socket – 2.1 mm panel socket – centre pin positive.
(b) Key socket – ¼ inch jack socket.
(c) Aerial and receiver connectors – panel-mounting SO239 type.

Particular attention must be paid to the mounting of the variable capacitor, VC1. Make sure the hole for the shaft is amply big enough, and if you use screws to mount the capacitor on the front panel, then make sure they are not too long, otherwise they will touch the vanes of the capacitor! Mounting can be by means of glue, sparingly applied and kept well away from the shaft.

A board size of about 6 × 10 cm is adequate. Component layout on the board is suggested in **Figure 3**. The prototype used ordinary matrix board, which is preferable to stripboard for a design like this; stripboard has undue capacitance between adjacent strips. Component leads are fed through holes in the board and are connected underneath. Make sure that leads and connections are rigid because, if they can move, there is always the danger of short-circuiting, and capacitance changes.

To facilitate construction, servicing and testing, it is advisable to use Veropins for connections to the variable capacitor, transmit/receive switch,

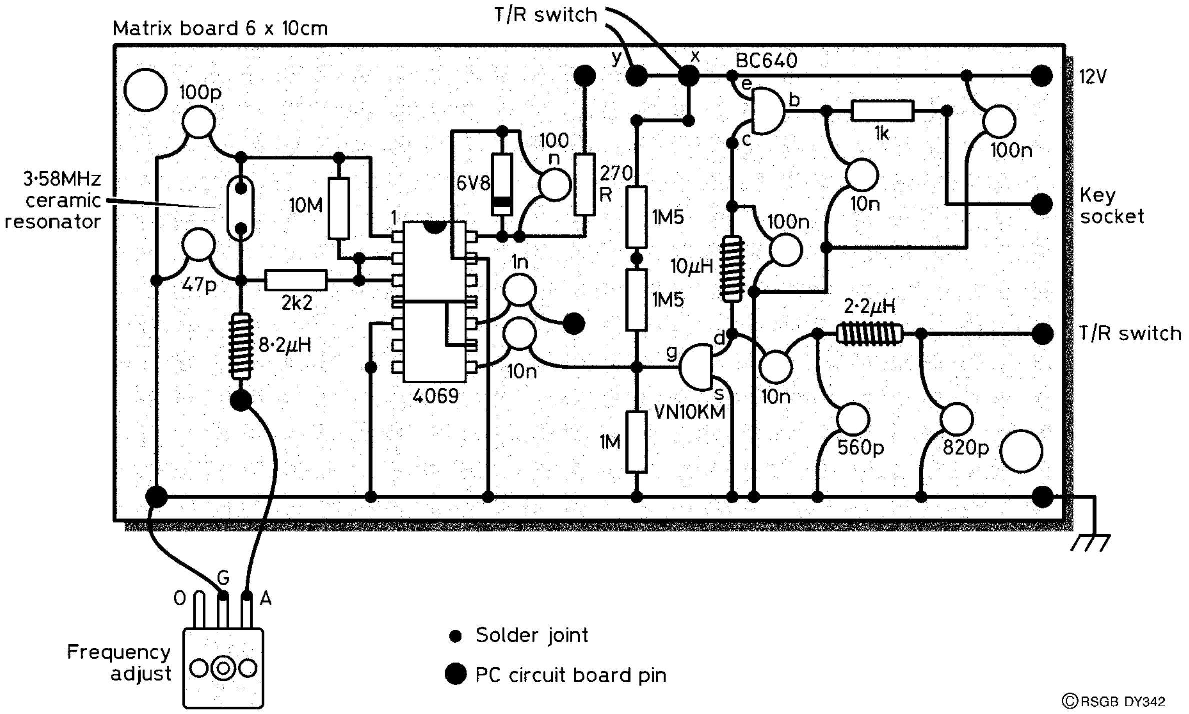
Matrix board 6 x 10cm
T/R switch
y
x
BC640
e
b
c
12V
100p
3·58MHz
ceramic
resonator
10M
1
6V8
100
n
270
R
1M5
1k
100n
10n
Key
socket
100n
10µH
47p
2k2
1n
1M5
2·2µH
T/R switch
8·2µH
4069
10n
g
d
s
VN10KM
10n
1M
560p
820p
O
G
A
Frequency
adjust
Solder joint
PC circuit board pin
©RSGB DY342

Figure 3 Component layout of the 80 metre transmitter. The transmit/receive switch is not mounted on the board and is not shown in this diagram

aerial and power sockets. Use screws and spacers to mount the circuit board to the box. Mounting the board horizontally assists troubleshooting.

Use a socket for the IC, and observe the CMOS handling precautions given in *An Electronic Die*, elsewhere in this book. When soldering the leads to the ceramic resonator, do it quickly – excessive heat damages the device. The earth lead running acoss the bottom of the board must be connected to the metal case by a short length of stout wire.

Testing

After carefully checking your wiring, both against the circuit diagram and the layout diagram, it is time to test your circuit. You will need a multimeter, an 80 m SSB receiver and a 50 Ω dummy load. A good design of dummy load can be found in the project *A Switched Dummy Load*, also in this book. An RF power meter and frequency counter will also be useful, although if your receiver has a digital frequency readout and S-meter, the latter two items are not really necessary. You will also need a 12 V 1 A power supply unit (PSU) to power the transmitter.

Switch the transmitter to receive and switch on the transmitter. No current should be consumed. Switch to transmit and check that pin 14 of IC1 is 6.8 V positive. With the dummy load connected to the aerial socket, press the key. The voltage on TR2 collector should now be 12 V, dropping to zero when the key is released.

Now check the operation of the oscillator. In transmit mode, you ought to be able to find a strong carrier signal with the receiver, even though the dummy load is connected. Adjusting the variable capacitor should change the frequency. At the lower end of the frequency range, you may find that the oscillator is unreliable in starting, because the circuit is attempting to pull the resonator too low in frequency. If this is the case, set the trimmer at the back of VC1 to minimum capacitance. If your version of VC1 has two trimmers, and you don't know which one to set, set them *both* to minimum capacitance. If there is still a problem, reduce the value of RFC1 to 6.8 or 4.7 μH.

In all probability, the unmodified circuit of Figure 2 will not require any of the changes outlined here. A coverage of 3.518 to 3.558 MHz should be possible, while preserving good frequency stability and reliable oscillation.

A signal probe (see *An RF Signal Probe*, elsewhere in this book) is useful for checking the operation of the oscillator and PA. Alternatively, an RF power meter or the receiver's S-meter can be used. With the PA running, the unit should draw between 200 and 300 mA. If TR1 becomes too hot to touch after a few seconds of transmitting, increase R5 or R6 to limit the transistor's heat dissipation. A small 6.3 V bulb connected across the aerial output is a simple way to check that the PA is working. An orange/white glow when the key is pressed is indicative of correct operation.

The final test is to monitor keying 'quality'. With your dummy load connected, press the key and listen to the note on the receiver's loudspeaker. Then operate the key, sending a string of dits, for example. What you hear should be free of chirps and clicks, as well as being stable in frequency. This test is sometimes better performed with no aerial connected to the receiver, thus preventing receiver overload and its associated plops. No problems should be encountered here.

Frequency tuning

A peculiarity of ceramic resonators is that, every now and again, their frequency changes abruptly by 100 Hz or so, then remains stable for some time. This is certainly noticeable in the received signal, but does not detract from the QSO and no characters are lost as a result. Try to keep the area around the ceramic resonator cool, to avoid temperature variations.

Parts list

Resistors: all 0.25 W, 5% tolerance	
R1	10 megohms (MΩ)
R2	2200 ohms (2.2 kΩ)
R3	270 ohms (270 Ω)
R4	1 megohm (1 MΩ)
R5, R6	1.5 megohms (1.5 MΩ)
R7	1000 ohms (1 kΩ)
Capacitors	
C1, C7	100 nF
C2	100 pF
C3	47 pF
C4	1 nF
C5, C6, C10	10 nF
C8	560 pF
C9	820 pF
VC1	10–160 pF variable
Inductors	
RFC1	8.2 μH
RFC2	10 μH
RFC3	2.2 μH
Semiconductors	
IC1	4069
TR1	VN10KM
TR2	BC640
D1	6V8 Zener

Additional items
- Matrix board (see text)
- 14-pin DIL socket
- Pointer knob
- Sockets (see text)

85 An audio booster for your hand-held

Introduction

The audio output from many hand-held transceivers and receivers usually leaves much to be desired, so this little amplifier was designed to increase the output at minimal expense.

All that is needed is a separate amplifier and bigger loudspeaker. This is accomplished using a single integrated circuit (IC), a few components, and a loudspeaker from the junk box. This circuit will enable the output from your hand-held to be heard easily in a car.

The circuit

This is shown in **Figure 1**. It uses only those components necessary to operate the IC amplifier. VR1 is the preset volume control, and varies the signal coming from the 'External speaker' jack on the hand-held before feeding it into the IC for amplification. C1 blocks any constant voltage present on the input.

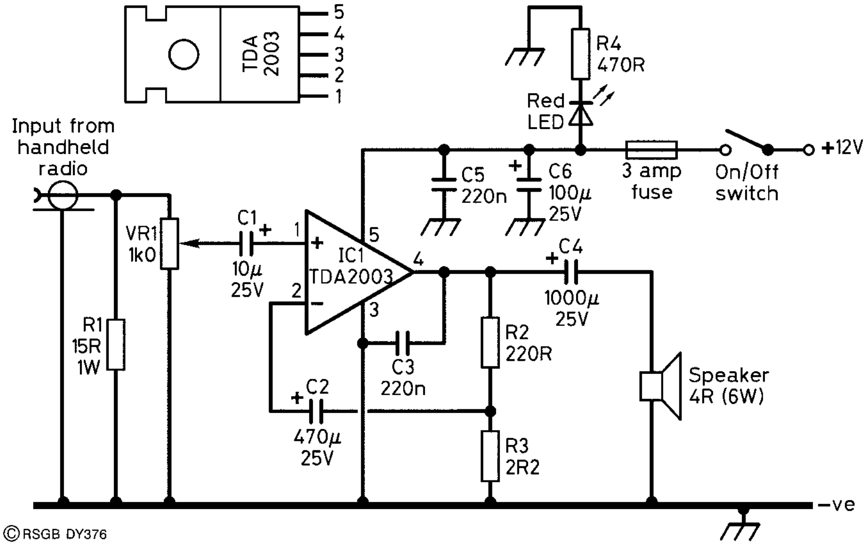
TDA 2003
5
4
3
2
1
R4
470R
Red LED
+12V
Input from handheld radio
C5 220n
C6 100µ 25V
3 amp fuse
On/Off switch
VR1 1k0
C1 10µ 25V
IC1 TDA2003
C4 1000µ 25V
R1 15R 1W
R2 220R
C3 220n
C2 470µ 25V
Speaker 4R (6W)
R3 2R2
−ve
©RSGB DY376

Figure 1 Circuit diagram of the audio amplifier. The power is derived from the cigarette lighter socket and the fuse is in-line with the lead

The IC output comes from pin 4 and is fed via the electrolytic capacitor, C4, to the loudspeaker. The circuit is provided with an on/off switch, fuse and LED to indicate when the circuit is switched on.

Construction

The box is made of aluminium. This is necessary to help to dissipate some of the heat generated by the IC. Do not build the circuit inside a plastic box unless you take special precautions! The IC has a metal mounting tab with a hole, specifically designed to be mounted to a metal box or other metal heat sink. Apply plenty of heat sink compound between the tab and the box, tighten the nut and bolt, and then wipe off any excess compound. The box will get slightly warm in operation.

The size of the box is not specified. You may want to decide on this when you find a loudspeaker. Choose one which will be able to handle 6 W

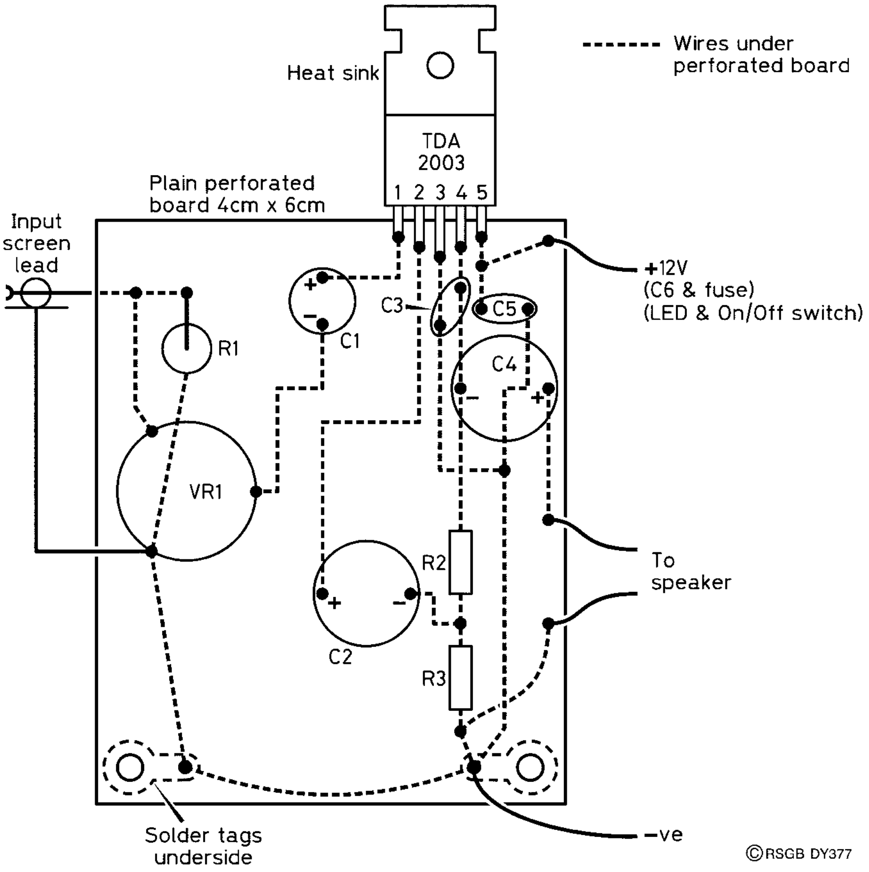
Heat sink
Wires under perforated board
TDA 2003
1 2 3 4 5
Plain perforated board 4cm x 6cm
Input screen lead
+12V
(C6 & fuse)
(LED & On/Off switch)
R1
C1
C3
C5
C4
VR1
R2
C2
R3
To speaker
Solder tags underside
−ve
©RSGB DY377

Figure 2 Layout of the components within the box

output. Drill all the holes in the box first. Holes for the speaker, input phono socket and the LED. The amplifier can be constructed on ordinary matrix board, which can be mounted inside the box with screws and spacers.

The layout is shown, for your guidance, in **Figure 2**. The components are mounted by pushing their leads through the holes in the board and making connections on the underside. The preset volume control, VR1, is set such that the hand-held's volume control is sufficient to control the final output over a good volume range. Use a screened lead from the 'External speaker' jack socket to the phono plug.

An external power supply is needed for this circuit. The normal dry battery which we usually use for small projects in this book will *not* work here, so you will need a proper mains power supply producing a stabilised 12 V. If you are going to use the unit principally in a car, then the cigar lighter socket can supply this voltage easily. Do make sure that the polarity is correct before you switch on!

When you plug the jack plug from your booster into the 'External speaker' socket on your hand-held, its internal speaker will be muted, so don't think that something dire has gone wrong! Adjust VR1 for a good volume range on your booster, when the volume control is turned on the hand-held.

Parts list

Resistors: all 0.25 W, 5% tolerance, unless otherwise stated

R1	15 ohms (15 Ω) 1 W
R2	220 ohms (220 Ω)
R3	2.2 ohms (2.2 Ω)
R4	470 ohms (470 Ω)
VR1	1000 ohms (1k Ω)

Capacitors

C1	10 μF 25 V
C2	470 μF 25 V
C3, C5	220 nF (0.22 μF) Mylar
C4	1000 μF 25 V
C6	100 μF 25 V

Semiconductor

LED	5 mm Red

Integrated circuit

IC1	TDA2003

Additional items
- Heat sink compound
- Nuts and bolts
- Loudspeaker 4 Ω 6 W
- 3 A fuse
- On/off switch (SPST)
- Matrix board 4 × 6 cm
- Solder tags
- Plugs and screened cable for connecting lead
- Aluminium box

86 A grid dip oscillator

Introduction

When an inductor and a capacitor are connected, whether in series or parallel, they form a circuit with a natural (or *resonant*) frequency. The circuit stores energy, and this energy is being constantly shifted from the inductor to the capacitor and back again.

The dip oscillator is a simple instrument used to measure the resonant frequency of a tuned circuit without having to make any direct connection to the circuit. The circuit is more commonly known as the *grid dip oscillator* (GDO), from the days when the active device in the circuit was a valve. The FET or Field-Effect Transistor operates in a way which is very similar to that of the valve, so it is not quite a misnomer to call this instrument a grid dip oscillator, too.

The circuit

The GDO uses a calibrated, tunable FET oscillator in the circuit of **Figure 1**. It has a frequency range of 1.6 to 35.2 MHz in four ranges using a set of plug-in coils, shown in **Figure 2**. When the oscillator coil, L1, is placed near an external resonant circuit, some of its RF energy is coupled into the external circuit. A gain in energy of the external circuit must mean a loss of energy in the GDO circuit, resulting in a change of current through TR1, which is measured by the meter, M1.

The current through TR1 is of the order of 5 to 8 mA, but the change of current may be only a few microamps. To measure a very small

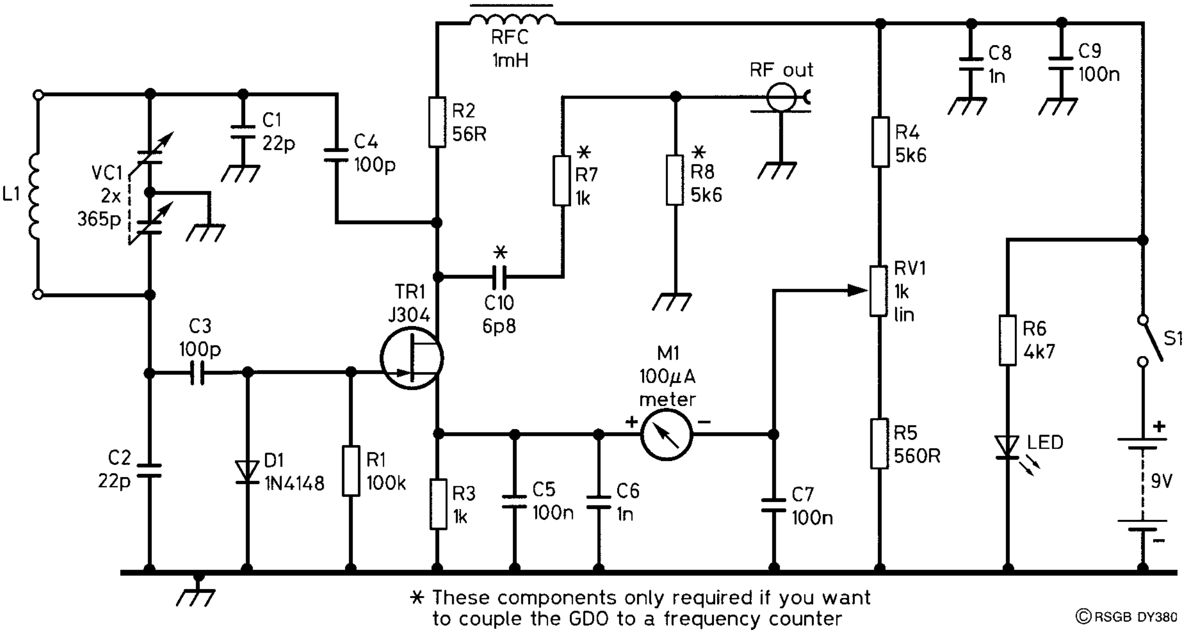
RFC
1mH
RF out
L1
VC1
2x
365p
C1
22p
C4
100p
R2
56R
R7
1k
R8
5k6
R4
5k6
C8
1n
C9
100n
C10
6p8
TR1
J304
C3
100p
RV1
1k
lin
M1
100μA
meter
R6
4k7
S1
C2
22p
D1
1N4148
R1
100k
R3
1k
C5
100n
C6
1n
C7
100n
R5
560R
LED
9V
* These components only required if you want to couple the GDO to a frequency counter
©RSGB DY380

Figure 1 Circuit of an FET GDO. The coils are wound on DIN speater plugs, which provide both a plug-in base and a coil former

Figure 2 Details of coil construction:
Range 1: 1.6–4.0 MHz 55 turns of 30 SWG
Range 2: 3.3–7.9 MHz 27 turns of 30 SWG
Range 3: 6.3–4.0 MHz 14 turns of 26 SWG
Range 4: 11.9–4.0 MHz 7 turns of 24 SWG.

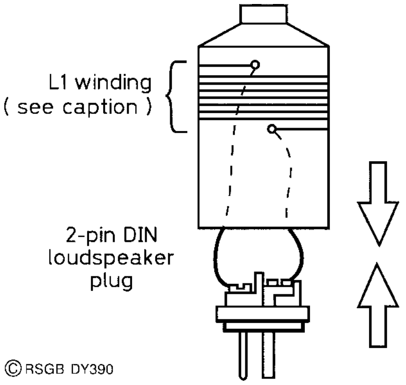
L1 winding
(see caption)
2-pin DIN
loudspeaker
plug
©RSGB DY390

change superimposed on a much larger standing current, the method of *offset* can be used.

One connection to the meter goes to the source of the FET, while the other goes to a variable offset voltage set by VR1. M1 has a full-scale deflection (FSD) of 100 μA. If the current through TR1 changes, the voltage across R3 changes. When there is no resonance, the voltage at the wiper of VR1 is set to be very slightly greater than that across R3, and there is a 75% FSD meter deflection. When the voltage across R3 decreases very slightly, due to external circuit resonance, a significant 'dip' in the meter deflection is produced, hence the name of the instrument.

The circuit is not difficult to make on standard matrix board. Provided you can follow a circuit and translate it into a good component layout, then this project is probably only an evening's work.

The most important part of the GDO is the tuning capacitor and its associated frequency-calibrated dial. New, air-spaced tuning capacitors can cost you up to £20, so it is worth delving around in junk boxes, or scouring the tables at a local boot sale or rally. The tuning capacitor from an old transistor radio should be perfect. It may even have a slow-motion drive and a dial which can be remarked for the project.

Choose a coil plug and socket arrangement that is practical. Think about crystal holders or phono plugs and sockets. The prototype shown in Figure 1 and Figure 2 used 2-pin DIN plugs, with the coil wound on the outside of the plastic plug cover. Figure 2 shows the coil construction and the winding details.

If you use a variable capacitor, VC1, with a value different from that shown in the parts list, then the frequency ranges will be different. This does not matter, as it will be taken into account during calibration.

Position VC1 so that the dial will be easy to see and to operate, while locating the coil socket as close to it as possible. **Figure 3** shows the traditional layout of the GDO.

Calibration

Because the GDO also radiates a very small amount of energy, a general coverage receiver can be used to calibrate the dial. Don't try to aim for great accuracy and clutter the dial with marks and figures! If you include C10, R7 and R8 in your circuit, you can connect a frequency counter directly to the GDO and leave it in circuit all the time.

The GDO in use

Always try to place the external tuned circuit with its coil coaxial with the plug-in coil, as shown in Figs 1 and 3. If the coils are at right angles, the GDO may not produce any resonance. Set the offset control to give about 75% FSD and slowly tune L1 through its whole range. If no dip occurs, you may have the wrong coil plugged in. When you eventually find a dip, move the external coil further away until only a minute dip is seen. You may have to retune the dip meter as you do this, but it gives a much more accurate reading of frequency.

Remember that you cannot 'dip' a coil by itself – there must always be a capacitor present.

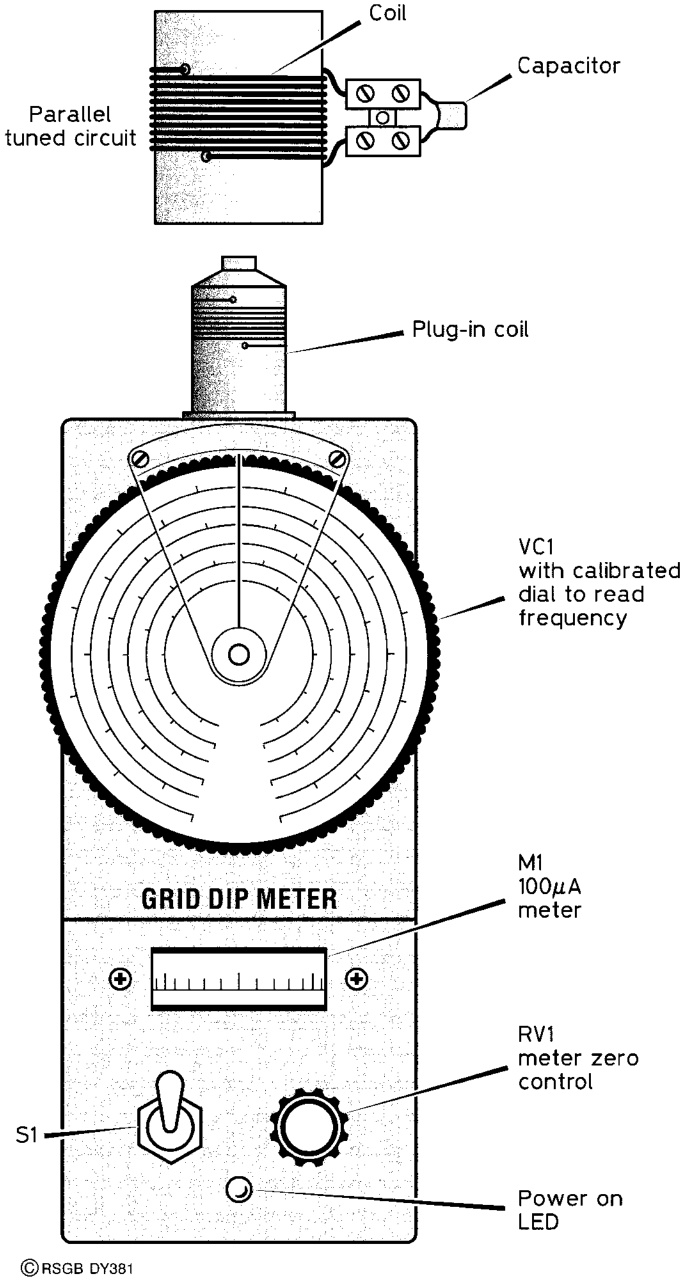
Coil
Capacitor
Parallel tuned circuit
Plug-in coil
VC1 with calibrated dial to read frequency
GRID DIP METER
M1 100μA meter
RV1 meter zero control
S1
Power on LED
©RSGB DY381

Figure 3 Layout of a typical GDO. The dial, meter and the location of the coil to the circuit under test can all be viewed at the same time. It is shown measuring resonance of a tuned circuit

Aerial resonance

Figure 4 shows how to check the resonance of a dipole aerial. Disconnect the coax feed at the aerial and place a short piece of wire, terminated with crocodile clips, across the centre insulator to short together the two ends of the aerial. By placing the GDO close to the shorting link, a dip should be seen on the meter while VC1 is turned. Alternatively, a loop in the element can be made around the coil, as in Figure 4, or the shorting link can be made long enough to loop over the coil. The latter method does not require tampering with the mounting and tensioning of the dipole wires.

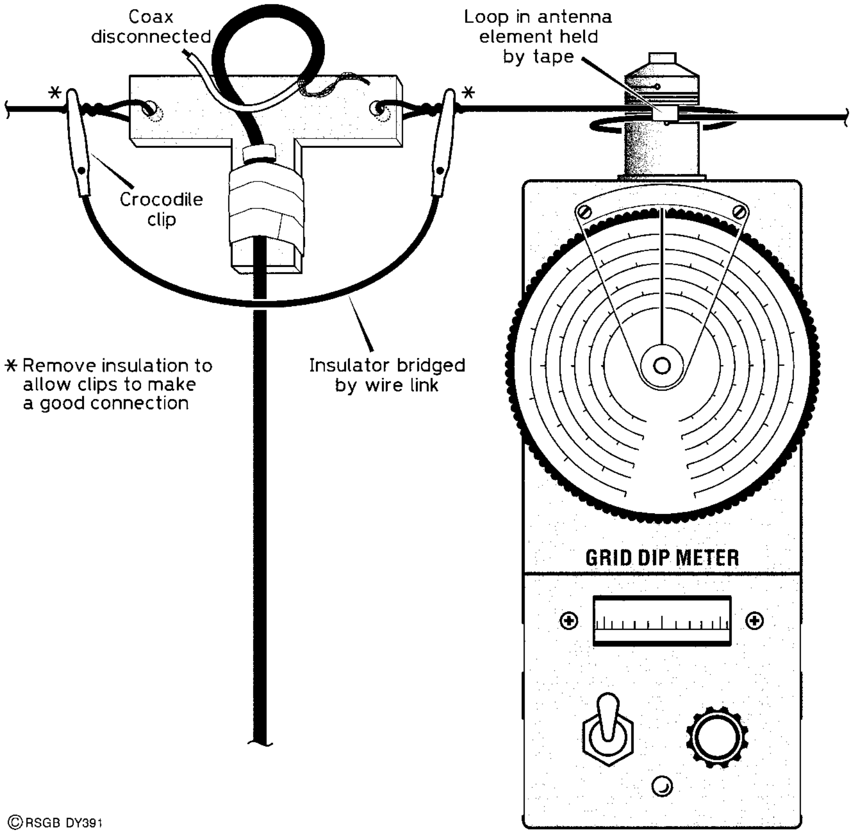
Coax disconnected
Loop in antenna element held by tape
*
*
Crocodile clip
GRID DIP METER
* Remove insulation to allow clips to make a good connection
Insulator bridged by wire link
©RSGB DY391

Figure 4 A wire antenna element can be looped into a single turn coil for increased coupling to the GDO

Parts list

Resistors: all 0.25 W, 5% tolerance

R1	100 000 ohms (100 kΩ)
R2	56 ohms (56 Ω)
R3	1000 ohms (1 kΩ)
R4	5600 ohms (5.6 kΩ)
R5	560 ohms (560 Ω)
R6	4700 ohms (4.7 kΩ)
R7	1000 ohms (1 kΩ)
R8	5600 ohms (5.6 kΩ)
VR1	1000 (1 kΩ) linear

Capacitors

C1, C2	22 pF
C3, C4	100 pF
C5, C7, C9	100 nF (0.1 μF)
C6, C8	1 nF (1000 pF)
C10	6.8 pF
VC1	2 × 365 pF

Semiconductors

TR1	J304 or similar
D1	1N4148
LED	

Additional items

L1	See Figure 2
RFC	1 mH
S1	SPST
M1	100 μA

Source

Components are available from Maplin.

87 A CW transmitter for 160 to 20 metres

Introduction

This very small transmitter is designed to work on any band from top band (160 m) to 20 m, with an RF output of 1 W. It will work on higher frequencies but with a reduced output.

The circuit

The three-transistor circuit is shown in **Figure 1**. It comprises a crystal oscillator using a BC182 transistor which drives a 2N3866 power amplifier (PA) keyed by a ZTX750 PNP transistor. The oscillator and PA are coupled by a capacitor and resistor; this provides a very small amount of positive bias to the PA.

The oscillator can be used as a basic crystal oscillator but, by including a variable series capacitor as shown in Figure 1, the crystal frequency can be 'pulled' slightly, making the oscillator a *variable crystal oscillator* (VXO).

Construction

The PCB layout is shown in **Figure 2** and in the photograph. Although the prototype was built around a PCB, this circuit is equally amenable to

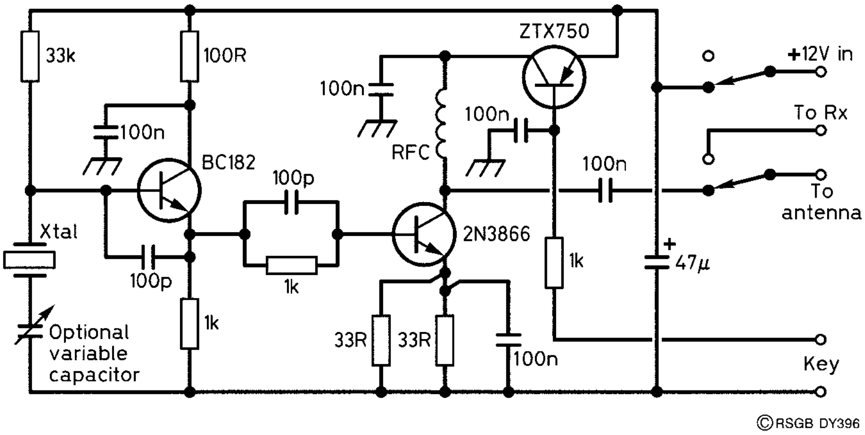
33k
100R
ZTX750
+12V in
100n
100n
To Rx
100n
BC182
RFC
100p
100n
To
antenna
Xtal
2N3866
1k
47µ
100p
1k
1k
Optional
variable
capacitor
33R
33R
100n
Key
©RSGB DY396

Figure 1 Transmitter circuit diagram

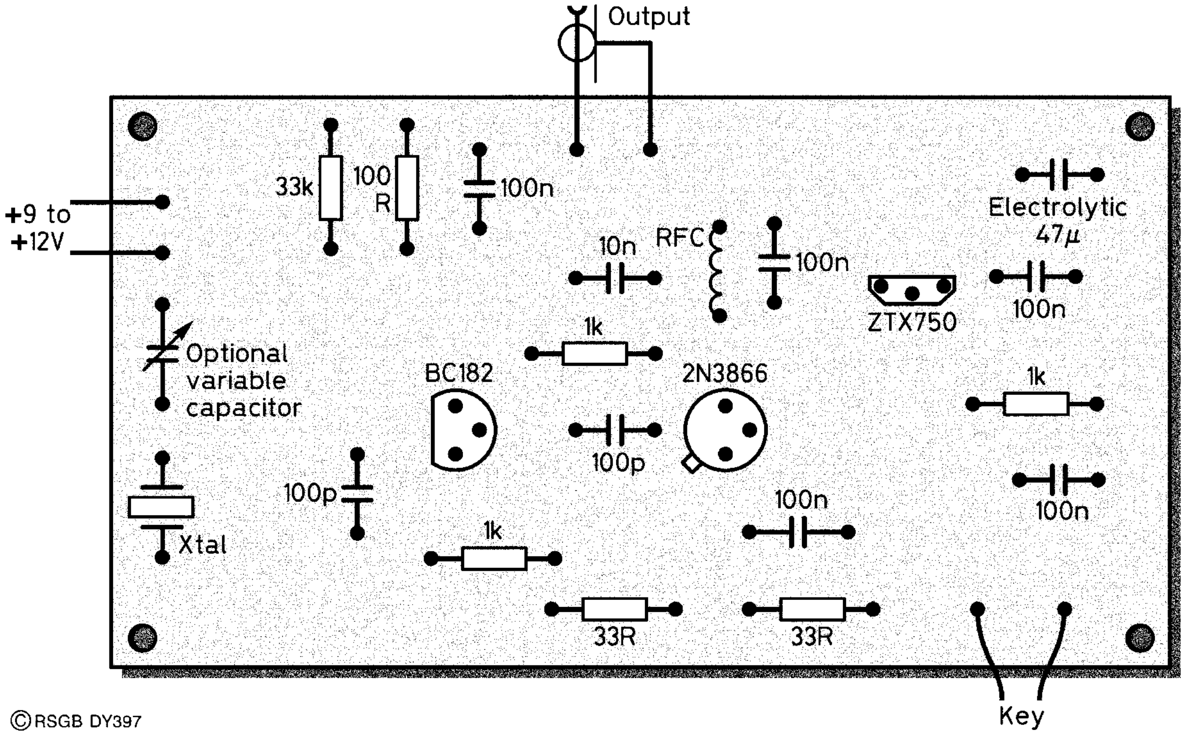
Output
+9 to
+12V
33k
100
R
100n
10n
RFC
100n
Electrolytic
47μ
100n
ZTX750
Optional
variable
capacitor
1k
BC182
2N3866
1k
100p
100p
100n
100n
Xtal
1k
33R
33R
Key
©RSGB DY397

Figure 2 Component placing on the PCB. The external connections are also shown

construction on a matrix board. Populating the PCB is very simple, and you can expect to be able to do this in about one hour. The radio-frequency choke (RFC) is made by winding 10 turns of 33 SWG wire on a ferrite bead. The enamel coating of the wire is intended to vaporise when soldered into the board, thus obviating the need to remove the enamel manually with a knife or sandpaper. However, if you do have problems with the PA either not working or keying intermittently, it is suggested that you investigate the RFC connections immediately!

If you decide to make the VXO version, you will have to cut the track between the crystal and earth, and connect the variable capacitor (250 pF) across the break.

In use

After performing the usual checks on the accuracy of your circuit building and the wiring of the external components, it is time to connect a 12 V battery between the points shown in Figure 2. Do not switch on yet. An aerial needs to be connected to the output via an ATU and a crystal, of frequency matching that of the aerial, fitted. The variable capacitor should give you a tuning range from about 14.058 to 14.064 MHz. 'Netting' (the process of tuning your transmitter to the same frequency as that of a received station) is achieved simply, because the oscillator is always running, and the leakage of the signal (despite the fact that the PA is not powered) is sufficient to bring the two frequencies to zero beat.

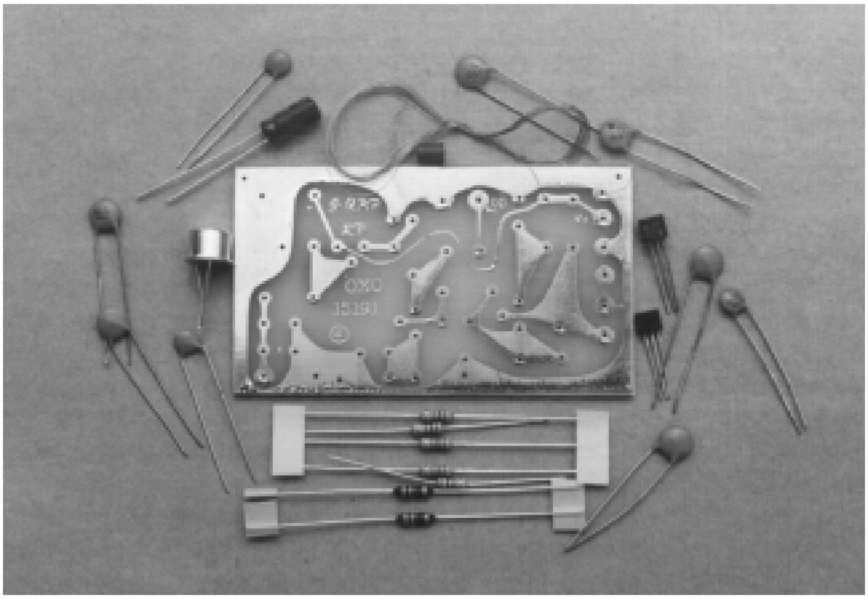

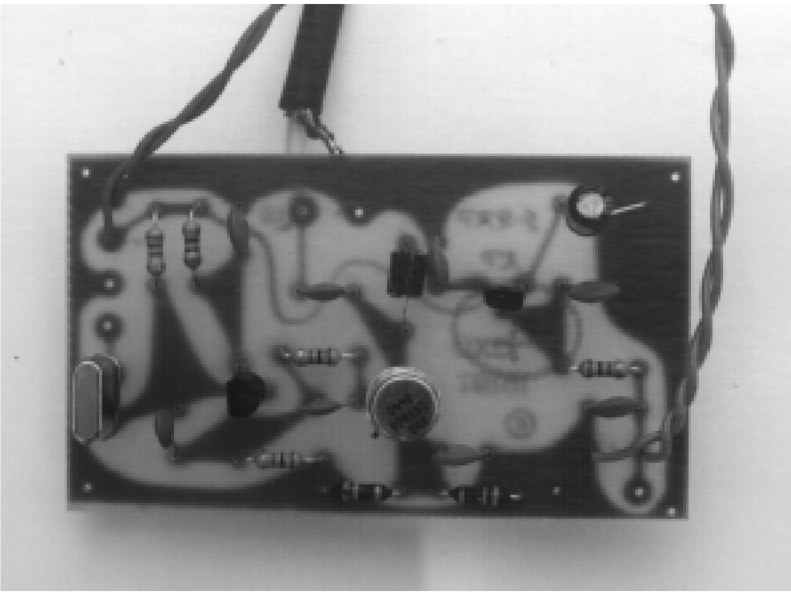

Warning

Note that the transmitter has no filtering; harmonics are not suppressed. It is strongly recommended that you use this transmitter in conjunction with the excellent low-pass filter described in the project *A 7-element low-pass filter for transmitters*, described elsewhere in this book.

88 Matching the end-fed random-wire aerial

Introduction

Many amateurs who do not have the space (or money) for a multi-band beam aerial, make use of the simplest possible alternative – the longest piece of wire that they can erect, with its end connected to the transceiver or receiver by an aerial tuning unit (ATU).

The length of the wire is not of major importance. Any length between 10 m and 80 m, with bends if necessary, will suffice. A good earth connection to the radio is just as important. Bends in the aerial wire can have some interesting effects on the directional properties of the aerial; V- or L-bends, or even a square shape are permitted. The only thing *not* permitted is to fold the wire back on itself in a tight hairpin bend!

Longish wire aerials

The term ‘long wire’ is usually used (incorrectly) to describe an end-fed aerial. How long is a piece of string? It depends what you mean by ‘long’. In aerial parlance, it means ‘long with respect to one wavelength’. Again, this depends on the band you are using. A long wire at 20 m is somewhat different from a long wire at 160 m. However, if you have sufficient real estate for a long wire on the 160 m band it must, by definition, be a long wire on all the other bands, too!

This should put you on your guard when analysing published data about feed-point impedance and the directional properties of a long wire aerial. Such theoretical data relate to a *real*, *ideal* long wire which is straight, horizontal, very high above perfect (conducting) ground, and not obstructed in any way. So your aerial doesn't quite match these criteria? Join the club!

Don't let this dampen your ardour when it comes to evaluating what the longish wire can do for you. The following should explain why.

Feed-point impedance

The impedance at the end of a longish-wire aerial can vary from a few tens of ohms to several thousand ohms, depending on the frequency in use and the wire length. It is also affected by factors such as bends, height above ground, proximity to buildings and wire diameter.

The actual value doesn't matter, provided we can make the aerial *appear* to have a 50 Ω impedance at the aerial socket of the transmitter. This process is what we call *impedance matching*, or simply *matching*. It maximises the power transfer from the transmitter to the aerial, and from the aerial to the receiver.

That is why an aerial tuning unit, or ATU, is almost (but not necessarily) obligatory.

The ATU

Many commercially produced HF receivers and transceivers have single 50 Ω coaxial sockets as their one and only means for connecting an external aerial. This means that an external aerial should have a 50 Ω feed impedance if it is to work efficiently, and it rules out most of the aerials being used by amateurs on the HF bands. Some means is necessary to change the aerial feed impedance to 'match' that of the transceiver. Such impedance-matching, or Z-matching (because Z is the symbol for impedance, just as L is the symbol for inductance) is the rôle of the ATU.

These can be bought and will accommodate either an end-fed or a coax-fed aerial. They can be bought ready for use or in kit form. Whether you want one for receiving only, or for use with a low-power (QRP) or high-power (QRO) transmitter, will determine what you need and how much you pay.

A simple single-band ATU

The simplest form of ATU is shown in **Figure 1**. It is simply a parallel LC (coil and capacitor) circuit, resonant at the chosen frequency, with taps on the coil for the aerial and the coaxial feed to the transceiver. If we assume the circuit is resonant, a high impedance exists at the top of the coil, and a low impedance at the bottom.

We said earlier that the end-fed longish wire presented an impedance which was high (or at least higher than 50 Ω). This explains why the aerial is tapped to the coil near the top, where the impedance is high, and the 50 Ω coax is tapped near the bottom, where the impedance is low.

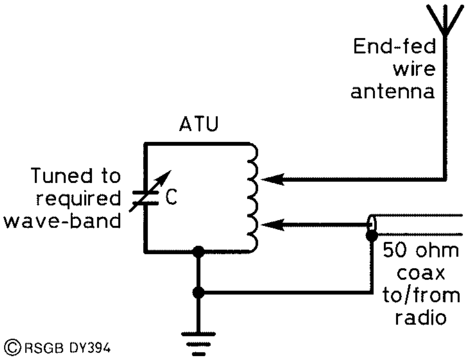

Figure 1 Parallel tuned circuit as single-band ATU for end-fed longish-wire antenna

Because the feed-point impedance of the aerial changes with frequency, so must the point at which the aerial is tapped to the coil to achieve impedance matching. The value of C must be changed also, to ensure that the circuit is resonant, and the 50 Ω tap will require tweaking also.

Setting up an ATU is quite simple. Make up an LC parallel-tuned circuit consisting of 50 turns of enamelled copper wire on an empty 35 mm film plastic container (or similar), tuned with a 500 pF variable capacitor. Make sure the enamel is removed from the ends of the wire before soldering.

Solder the inner wire of the coaxial cable from the radio to the first or second turn of the coil from its grounded end. Then solder the braid to the grounded end. Connect the aerial about one-third of the way down the coil from the top, removing the enamel at the connection point.

Adjust the variable capacitor for maximum noise or signal strength in the receiver. Then, try different tapping points from the aerial, to maximise the signal again. This matches the aerial impedance to that of the tuned circuit. Repeat the process with the coax tap, thus matching the impedance of the radio and the cable to that of the tuned circuit.

You will no doubt find that the tapping process on the coil was not easily accomplished, especially when the enamel must be removed at each tapping point without shorting adjacent turns together. It is therefore logical to produce a design where the taps have been prepared during the winding of the coil, and are selected with a rotary switch. To this end, the following multi-band design is described.

A multi-band ATU

The same type of coil design around a 35 mm film container is used (see **Figure 2**). The tapping points can be prepared in advance by a little judicious planning. Dismantle your original coil and measure the length of wire on it. It will be a little more than a calculation of $50 \times \pi D$ would suggest (where D is the diameter of the film container), due to the lack of tension in the coil and the wire diameter itself. You will need aerial tapping points at turns 1, 2, 3, 5, 10, 15, 20, 25, 30, 35, 40 and 45, corresponding to all the bands from 28 MHz to 1.8 MHz. The coaxial cable tap is fixed at turn 2 (an acceptable compromise). All the turns are counted from the earthy end.

Cut another piece of enamelled copper the same length as you used originally. Then, with the aid of an ordinary calculator, work out the positions of the points where the enamel must be removed for the taps. For example, turn 15 will have to be made $\frac{15}{50}$ of the way along the wire, turn 20 tap made at $\frac{20}{50}$ of the way along. So, if the length of wire is, say 4.7 metres, the two taps in question will be made at $\frac{15}{50} \times 4.7 = 1.41$ m and $\frac{20}{50} \times 4.7 = 1.88$ m from one end. This must be repeated for each of the tap positions, and the enamel removed ready for the wire to be soldered to it. With a

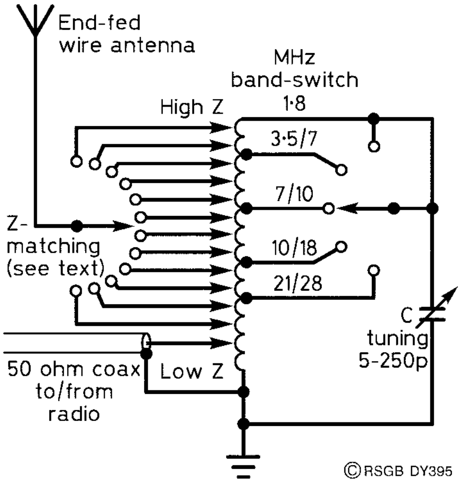

Figure 2 Simple multi-band ATU for end-fed longish-wire antennas

soldering iron, tin each tap point while the copper is shiny, thus ensuring a good, low-resistance connection. Do not solder on the taps yet.

The coil can now be wound as before. The taps can be soldered on, taking the lead from each one to the wafer of a single-pole 12-way rotary switch, the pole being connected to the aerial.

The tuning of the ATU is carried out by the same 250 pF capacitor, with a single-pole 5-way rotary switch used to select the band. Its tapping points will need to be chosen manually, using the method described earlier. Don't attempt to make new tapping points on the coil for this – use the taps available on the wafer of the other rotary switch, and find which is optimum for each band.

Notes

- For the aerial, use PVC-covered stranded tinned copper, of size 16/0.2 mm or 24/0.2 mm.
- Make the wire as long as possible, but anything over 10 m should be OK.
- Keep the wire as high as you can, in the clear and away from obstructions.
- Don't worry about bends, but don't use hairpin bends.
- Use a good insulator to attach your aerial – anything plastic will do.
- Anchor the wire near the point of entry to the building, but use a U-bend to prevent ingress of water.
- The wire can be brought in through the corner of a window, the PVC acting as an insulator. If you must drill a hole in the brickwork, make sure it slopes upwards from outside, so that water is deterred from entering.
- Use a good RF earth (as opposed to an electrical earth) such as half a dozen bare copper wires buried under the lawn in a fan shape. They should be joined together at the point of the fan and strapped to the earth connector of your ATU.

$4.95

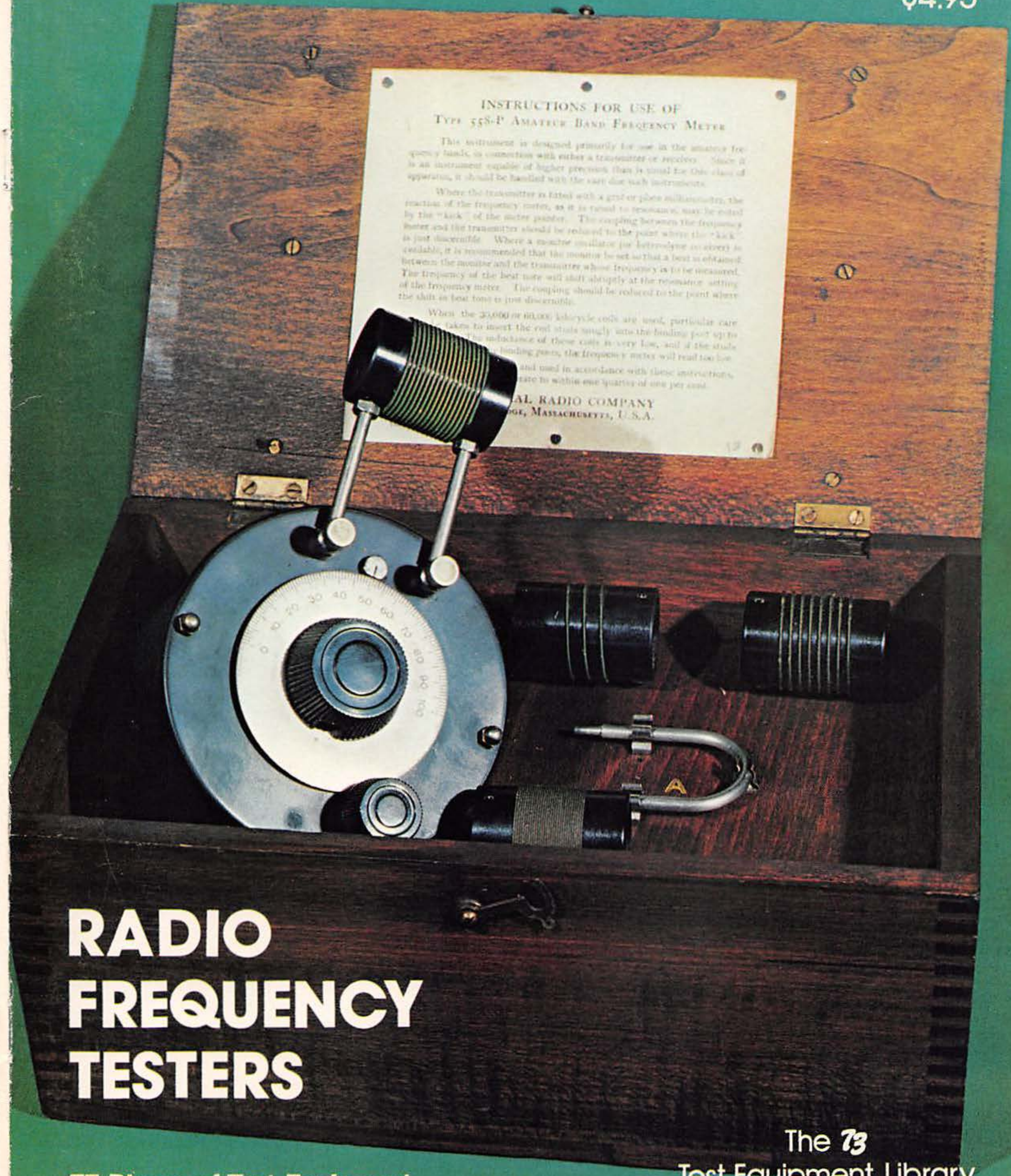

INSTRUCTIONS FOR USE OF
TYPE 558-P AMATEUR BAND FREQUENCY METER

AL RADIO COMPANY
GE, MASSACHUSETTS, U.S.A.

RADIO FREQUENCY TESTERS

The 73
Test Equipment Library
volume III

77 Pieces of Test Equipment You Can Build!

the 73 TEST EQUIPMENT LIBRARY

Volume III: Radio Testers

Edited by Sandy Cole W1PUF
and
Rich Force WB2QYV/1

A product of 73 Inc.

EDITOR'S COMMENTARY

This is Volume III in the *73 Magazine* series of Test Equipment books. The first two volumes covered component testers and audio frequency testers. This book, which deals with radio frequency testers, is by far one of the most important of the series.

The generation of radio frequency waves is the one common denominator through all of ham radio. No matter what your specific interest in this fascinating hobby, without the correct generation and radiation of radio waves, your enjoyment will be frustrated. In this book we have attempted to put together some of the best articles which have appeared in *73 Magazine* all aimed at assisting you in achieving the best possible on-the-air signal. Many hours have gone into its preparation by scanning past issues for projects which would be of the most use, be up-to-date and be easily constructed with available parts. We think we have achieved this end.

It must be noted that some of the equipment described in this book appeared in vintage issues of *73 Magazine*, yet the design and value of the equipment has not aged and it is as up-to-date today as it was when it was first published.

We at *73 Magazine* hope this book will help to make your effort to put out a "respectable" signal easier and more enjoyable.

73, Rich Force WB2QYV/1

Copyright 1976
73 Inc., Peterborough NH 03458

All Rights Reserved
Printed in U.S.A.

Table of Contents

The 73 Test Equipment Library

Volume III: Radio Frequency Testers

Chapter I: What is Your SWR, OM?

SWR Power Meter for 80 Thru 10m Bands W6VAT 5
An Etched Circuit SWR Bridge WA5SWD 6
Checking Your SWR Indicator W5EHC 8
Home Brew Bridge Calibration W4OAB 8
Mini SWR Bridge W2EEY 8
VSWR Supreme W4BRS 9
A Simple VHF SWR Meter W2EEY/K3EZ 10
Directional Coupler and VSWR Bridge for VHF and UHF WA6SXC 13
A Slotted Line for 1250 MHz WA8CHD 14
Some Directional Wattmeters and a Novel VSWR Meter G3PDM 16
SWR Bridges Using Zero-Center Meters W2EEY/K3EZ 18

Chapter II: How to Measure RF Impedance

The Antennascope – An Effective Tool VE3CEA 21
How to Build an Antennascope WA1CCH 22
The Mark III RF Impedance Bridge W6CRT 22
Measure RF Impedance with Your SWR Bridge W2AOO 26
Use Your GDO and Z Meter W0BMW 27

Chapter III: Measuring RF Power Output

Measuring RF Output WA6CPP 30
Light Bulbs as RF Indicators WB2GQY 30
Build Yourself a Light Wattmeter Henry 32
Low Power Light Wattmeter K1CLL 33
Yet Another Light Wattmeter K5CXN 34
A Low Cost RF Wattmeter WA3AJR 35
RF Power Measurement Using Hot Carrier Diodes W6AJF 36
VHF Dummy Load Wattmeter K9STH 38

Chapter IV: Measure Your Field Strength and Frequency

The RF Sniffer K5JKX 39
A Useful Accessory for the Ham Shack W2WHX 40
Two Thru Twenty FSM W7OE 40
An Amplified, Calibrated Signal Strength Meter VE3CES 41
Frequency Measuring Equipment at Microwave Frequencies WA9VFG 42
Frequency Meter-1 to 10 GHz Amateur Microwave K1CLL 43
Lambda Lines W2DXH 46

Chapter V: RF Signal Generators

Zero-Beating with a Frequency Meter WB4TBX 48
Hand Calibrate the BC-221 Frequency Meter Henry 48
High Accuracy VHF Frequency Measurements W5WGF 49
Transistorized LM Meter W5SOT 50
An I-F Spotter W5WGF 52
430-470 kHz Sweep Frequency Generator WA5SWD 53

An Armstrong Sweeper W2EMF 54
A New Broom K6JHJ 54
Low Cost Signal Source VE3GFW 55
Receiver Tweaker K6UAW 56
Citizen Band Alignment Aid WA5SWD 56
FET Chirper K6QKL 57
UHF Signal Source K1CLL 58
Crystal Controlled Signal Source with Infinite Attenuator for 144, 432 and 1296 MHz K1CLL 59
Low Cost 220 MHz Signal Generator K1CLL 60
UHF Signal Generator K6MIO 62
Lab Type IF/RF Sweep Generator Using ICs K4DHC 63
100 kHz Thin-Line Pulse Generator W2DXH 70

Chapter VI: Crystal Calibrators

All Band Band-Edge Marker WA2HRZ 74
The Multical K9VXL 74
100 kHz Marker Generator W7CJB 75
All Band Frequency Marker W1KNI 75
Poor Man's Universal Frequency Generator W2EEY 76
Calibrate that Calibrator W2KPE 78
A Simple Frequency Deviation Meter K6BJ 78

Chapter VII: GDOs

The Indicating Oscillator KH6AF 80
The Little Gate Dipper W5ETT 81
The Great Dipper WA0AYP 83
The Great Dipper-Modification WA0AYP 85
Etched Circuit UHF Dipmeter W1JJL 86
UHF Grid Dipper WA6BSO 87
Wide Range VHF-UHF Dipper K1CLL 91
Extra Services from Your Grid Dip Oscillator WA4UZM 94
GDO to Find C G3OGR 94
GDO Coil Extension WA0ABI 94

Chapter VIII: Noise Generators

Simple Noise Generator K9ONT 96
A Crystal Diode Noise Generator W5TEV 96
Temperature Limited Diode Noise Generator Thomas 97
Calibrated Noise Generator for 432 W6GXN 98
Another Way to Measure Noise Figure K5JKX 99

Chapter IX: Attenuators and Dummy Loads

Low Power Attenuators for the Amateur Bands WB6AIG/WA6RDZ 102
S-Unit Attenuator WA5SWD 103
The Ever-Useful T-Pad K5JKX 103
Combination Dummy Load/Attenuator Network W2EEY 104
The Minican W6JTT 106
The Tiny Terminator K5JKX 106

Chapter I

What is Your SWR, OM?

SWR POWER METER FOR 80 THRU 10m BANDS

James Lee W6VAT

With the declining sunspots, usage of the lower frequency bands is definitely increasing. The resulting increased QRM is helpful to no one and only efficient equipment and operating procedures will result in a maximum number of QSO's. The SWR/PWR meter described here won't make you a better operator, but it can help you be sure that you are delivering the most rf to your antenna from your rig. Fig. 1 shows a front view of the meter.

The basic circuit is a directional coupler switched to sample either forward or reverse voltage and a voltmeter to read this voltage.

This type of coupler has an output proportional to length, power, and frequency. The longer it is, the more output it gives. Since it puts a small impedance "bump" in the line, the length of the coupler should be limited to not over 1/20 wavelength at the highest frequency, or it may begin to contribute noticeably to the SWR itself. For a given power, if the frequency of the rf flowing through the coupler is reduced, the maximum coupler output is reduced. This means that you can get full scale readings on 10 M with a lot less power than on 80 M. The meter described here gives half scale deflection on 40 M, at maximum sensitivity, with about 350 watts of forward power. If your rig is a KW this meter will fill the requirements for SWR/PWR measurements nicely. It can be left in the line at all times to monitor SWR or Power delivered to the antenna or other load. If desired, a Barker and Williamson type 551A coaxial switch can be used to insert the meter in the line for test purposes and then switch it out during operating periods.

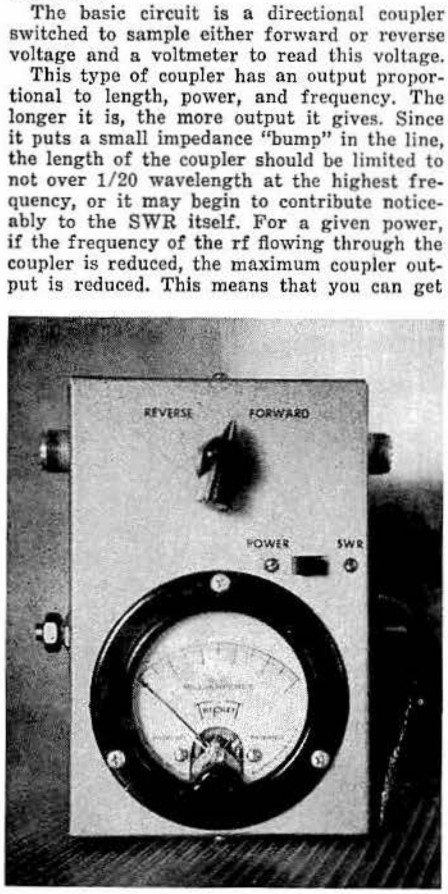

Construction

The unit is built in a gray hammertone LMB type 141 box. The dimensions of this box are 3" x 4" x 6". Fig. 2 shows the parts placement and should answer any questions concerning layout.

The coax directional coupler is made from a 14" length of RG-8/U. The outer covering is slit lengthwise with a knife and peeled off. Take care here not to cut into the woven braid. The woven braid is then bunched toward the center to loosen it. Next, a length of #22 enameled wire is passed through the braid at about 2¼" from one end and run under the braid next to the inner insulation. It is brought out at the other end, again 2¼" from the end of the shield braid. When this is done, smooth the braid back to its original position carefully to avoid scratching the enamel on the #22 wire. The #22 enameled wire should lay as straight as possible under the braid. It should have no slack nor should it twist around the inner insulation to any great degree. The ends of the shield braid are trimmed back far enough to be soldered to Amphenol 83-1H hoods. The inner insulation is trimmed off so as to expose about 3/16" of the inner conductor. The inner conductors are then soldered to Amphenol 83-1R female type chassis mount coax connectors. The coupler may now be set aside and the rest of the meter constructed.

Potentiometer R_1 is mounted on 1/16" x 1" x 4" piece of bakelite and positioned approximately in the center of the box to minimize capacity to ground. The metal rear cover is removed from the pot for the same reason. The pot ground lead is to the center of the coax and is wire in last. The diode D_1 should be protected from damage by heat when soldering it in the circuit. Long nose pliers gripping the leads near the body of the diode are satisfactory. The directional coupler is the last item soldered in place. As can be seen in the photos—the coax connectors are mounted from the inside of the box. When the #22 enameled leads are soldered to S_1 they should be the same length, and again, take care not to scratch the enamel or a short circuit may occur and you'll have to do it all over again.

Calibration

There are any number of ways to calibrate the SWR/PWR meter, but the way most hams will use is their own rig and a suitable dummy load. Though very limited in power, a 2 watt 50 ohm resistor such as made by Ohmite, IRC and others, mounted inside an Amphenol 83-1SP male coax plug makes a very good dummy load. Although not completely noninductive, this dummy load is far superior to such real unknowns as light bulbs, electric iron heating elements, etc. This particular load is 50 ohms shunted by 6 mmfd over the range of 3-30 mc. It is a good dummy load—though a low power one.

With the back cover off, attach the dummy load to J_2, the load jack. Set R_1, the Sensitivity Control, to maximum and S_1 to FORWARD. Next apply power by hooking your rig (or

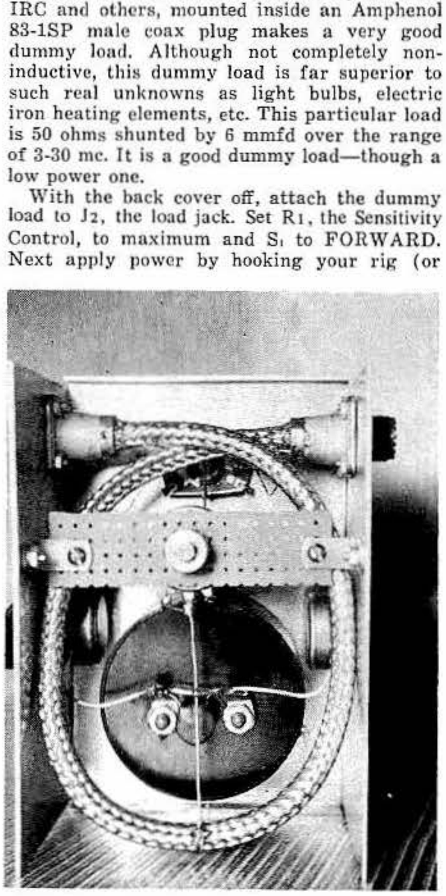

other rf source) to J_1, the transmitter jack. Make sure that the rf applied is 28-30 mc, or, the highest frequency you operate in the 3-30 mc region. This meter is sensitive to frequency. Stray capacities and other unbalances will have their greatest effect at the highest frequencies. In any event, calibrate the unit on the highest frequency your rig will tune in the 3-30 mc region. With power applied, set R_2 so the meter reading is at least half scale and switch S_1 to REVERSE. This will result in a lower meter reading. R_1 should then be adjusted for a minimum reading on the meter. Using the suggested load, you will not get a complete null, but the null should not be much more than 50 ua for half scale deflection in the FORWARD position on S_1. Don't put too much power into the dummy load since excessive dissipation can ruin it and change its characteristics greatly. If you have a higher power dummy load whose characteristics you know accurately, by all means use it, but remember a light bulb is not a good load. Once nulled, lock the nut on R_1 taking care not to disturb the setting. Replace the back cover and using the dummy load, recheck the null to make certain that it has not shifted.

Using The Meter

To make SWR measurements you need only insert it in the line and set S_1 to FORWARD and S_2 to SWR. Adjust the sensitivity control for at least a half scale reading. Then switch S_1 to REVERSE and read the value. SWR is then calculated by the following:

$$\mathrm{SWR} = \frac{I_{fwd} + I_{rev}}{I_{fwd} - I_{rev}}$$

For example:

let $I_{fwd} = 500$ ua and $I_{rev} = 50$ ua, then

$$\mathrm{SWR} = \frac{500 + 50}{500 - 50} = \frac{550}{450} = 1.22{:}1$$

The meter is useful for tune-up purposes where exact SWR is not needed. Just keep the FORWARD reading at a constant value and tune for minimum REVERSE readings. The exact SWR can be calculated when you have found the lowest REVERSE position.

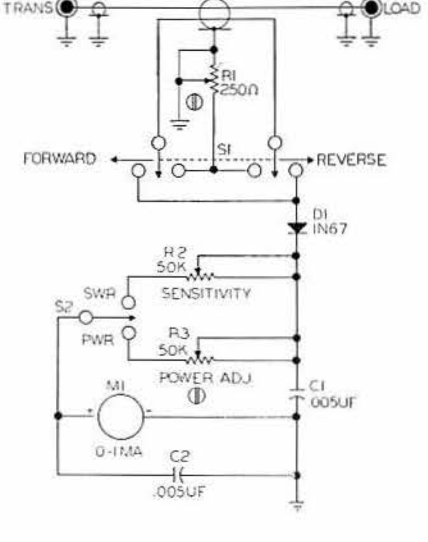

The POWER position may appear to be essentially the same as the SWR position and it is. When measuring power into a load of known fixed value you only need to know the voltage across (or the current through) the load. The POWER position on S_2 is used with S_1 set to FORWARD. R_3 is merely set to a scale reading that is convenient for all bands if only a relative reading is used. If a good 50 ohm high power dummy load is available, you may make accurate calibrations by using a VTVM plus high frequency detector probe and measuring the actual voltage across the load. Just put a T-connector on the load jack J_2. Put the dummy load on one arm of the "T" and read the voltage at the other arm. 200 volts across a pure resistive load of 50 ohms is equal to 800 watts of power. If your rig delivers a key down 800 watts to a load then you could set R_3 at 0.8 ma on the scale, etc. The scale will not be precisely linear, particularly at low powers, but if enough resistance is used at R_3 the effects should be minimum. Since the coupler voltage is frequency sensitive one setting of R_3 will not hold for all bands. If desired, R_3 could be replaced with a switch and a number of selected resistors (one for each band). This would keep the scale factor constant between bands. This could be done for R_2 as well, but complicates an otherwise simple device

Operation

Once calibrated, the meter is very simple to use. Just hook it in the line at some convenient point and apply power. With S_1 set to FORWARD and S_2 set to SWR, adjust R_2 for at least a half scale reading. Switch S_1 to REVERSE and you will then be able to continuously monitor your reverse power. For power measurements, set S_1 to FORWARD and S_2 to PWR. Apply power and the meter will be monitoring your forward power continuously. R_3 is a screw driver adjusted pot. It can be set, and locked at a point which allows operation on all the bands you operate. The scale reading on each band will be different, but once set, these readings can be jotted down in your log book and any change quickly noted. This meter is also an excellent device for the antenna "tinkerer," or if you have coax coupling between stages in your rig it can be used to provide proper power transfer. Best of all, it is not expensive nor difficult to construct. Try one and you won't be without one again.

AN ETCHED CIRCUIT SWR BRIDGE

Ed Lawrence WA5SWD

Several of these boards have been etched and units assembled, and all have been satisfactory. I have used mine, with a 200 microamp meter, from a kw on the low bands up to a "Twoer," with no problems with sensitivity.

Since Bud Miniboxes are commonly available, I scaled this board to fit into the Bud

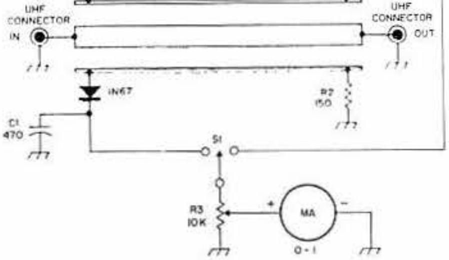

Schematic diagram of SWR bridge.

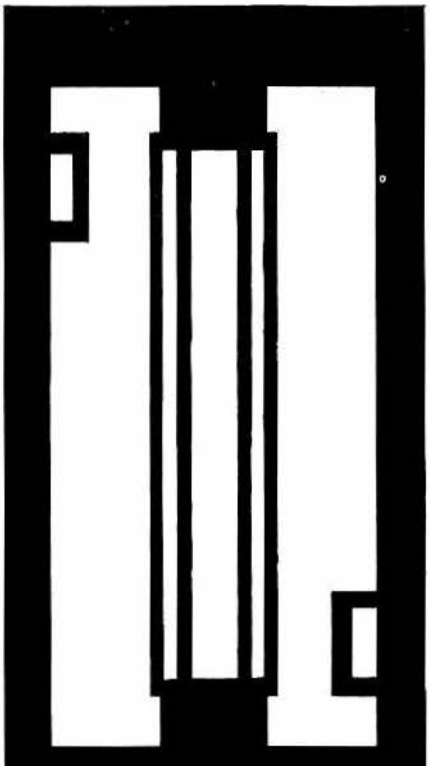

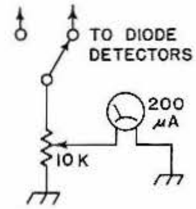

Fig. 1. Metering circuit for the etched circuit SWR bridge.

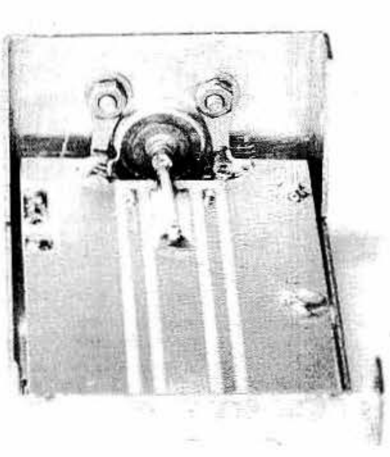

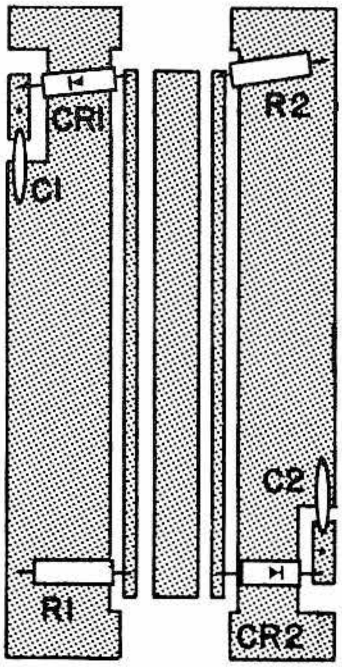

Fig. 2. Layout for the printed circuit board.

Minibox CU 2102A or CU 3002A. If you want to include the metering in the same package, the CU 2103A or CU 3003A should be used to give more room.

If you use the CU 2102A, center the coax connectors on the ends. If the CU 2103A, mount the connectors .8 inches from the open end. I used solder lugs bent at right angles to mount the P. C. Board to the chassis. The photograph shows the mounting much better than 10,000 words.

The meter sensitivity control circuit shown has a wider control range than the one shown in the *ARRL Handbook*, since the pot shunts the meter at low settings. This action could be accentuated by putting a fixed resistance in series with the meter movement, at the expense of sensitivity.

Although the virtues of operating transmission lines with low standing wave ratios (SWR) have been discussed many times in the past, evidently the economics of maintaining low SWR's are not readily apparent, particularly if the frequency of operation is low and the transmission line short. This has been reflected in various pseudo-technical QSO's where many have been led to the utter disregard for standing wave ratios. Most members of the amateur fraternity exist on limited budgets at best and when a significant portion of that precious transmitted power is eaten up by transmission line losses and misinterpreted standing wave ratios for naught, something should be done.

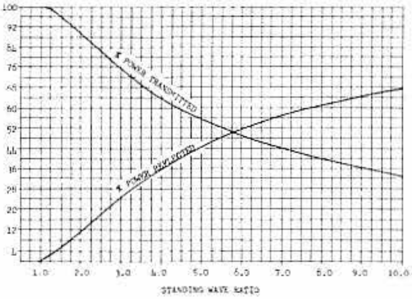

SWR vs % Power Transmitted and Power Reflected

Figure 1

A look at the graph in Fig. 1 will show you the percentage of power reflected for various standing wave ratios. For instance, if you are presently tolerating an SWR of about 5.8:1 (not uncommon in many ham shacks), 50% of the power which reaches the antenna is actually reflected back down the transmission line, heating up the final tank and causing TVI. Nor is only the transmitted signal effected, a high SWR will similarly degrade the received signal. This is particularly important in the reception of the extremely low level signals often encountered in DX and VHF operating. Stereophonic buffs should take heed too. A recent report by the IEEE (Institute of Electronic and Electrical Engineers) Professional Group on Broadcasting noted that a high SWR on receiver antenna inputs causes a reduction of stereo quality.

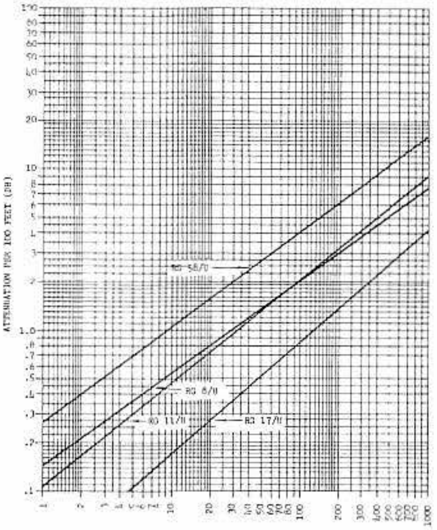

Figure 2

Attenuation vs frequency.

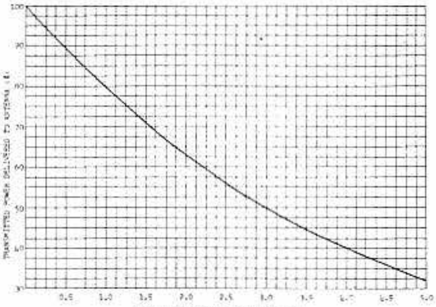

Figure 3

Attenuation vs Power transmitted.

Many of the amateur stations on the air today make use of RG8A/U coaxial cable. Its excellence is proven out by its extensive use by the military, but a look at the loss graph (Fig. 2) for this cable indicates that it is not completely lossless! Even at 4 mc it has approximately 0.3 db loss per 100 feet, and on six meters there is a loss of 1.4 db for the same length. A look at Fig. 3 indicates that on 75 meters only 94% of the transmitted power is delivered to the antenna if 100 feet of RG8A/U is in use. At 50 mc the loss has sky rocketed to 26% for the same length of line. However, there is one big hooker for these conditions to exist: the SWR must be 1:1. For any other value of SWR there will be further line losses as shown in Fig. 4 because standing waves have the property of multiplying attenuation. This graph indicates that if a transmission line is operating at an SWR of 3.7:1, the line loss will be multiplied by a factor of two. For the previously mentioned situation on 50 mc, an additional 24% loss could be expected with an RG8A/U line operating at an SWR of 3.7:1.

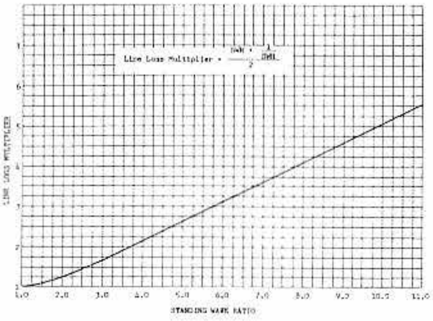

Figure 4

Line loss multiplier vs SWR.

It should be obvious by now that the use of an SWR bridge in the line at all times is very advantageous in the maintenance of a low SWR at the operating frequency. However, contrary to popular belief, the SWR bridge does not tell all. Since there is loss or attenuation in any length of transmission line, the reflected wave will be attenuated in the same matter as the transmitted or incident signal. Because the standing wave ratio is the ratio of the incident wave to the reflected wave, attenuation of the reflected wave will give erroneous SWR measurements when the SWR bridge is conveniently located at the transmitter. In this location the bridge will see the full power of the transmitter, but only a portion of the reflected signal. In some cases where the length of the transmission line is excessively long, the reflected wave will be attenuated to such a degree that the SWR will appear to be very close to 1:1, while in reality it will be a good deal higher. This fact is graphically represented in Fig. 5.

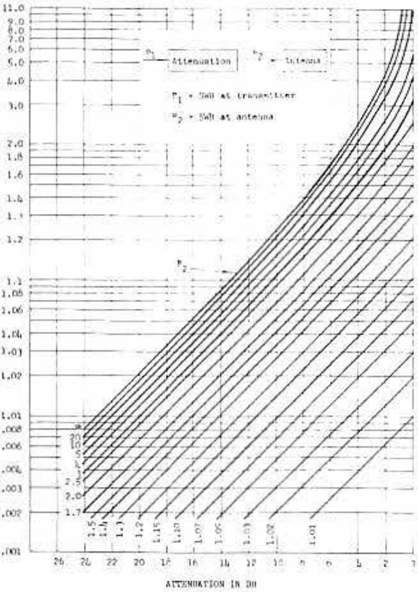

Figure 5

SWR vs attenuation.

For example, 143 feet of RG8A/U at 50 mc would result in approximately 2 db attenuation. If an SWR bridge inserted in the line at the transmitter indicated an SWR of 2:1, this graph shows that an SWR of 3:1 exists at the antenna. A look at Figs. 3 and 4 will indicate that a 3.3 db loss (2 db times 1.65 multiplier) occurs, amounting to 47% loss of transmitted power in transmission line losses. Of the remaining 53% power arriving at the antenna, 24% will be reflected back down the line. A little simple arithmetic will show that of the total power transmitted, only 30% will be radiated! This simple mathematical fact should make the merits of low standing wave ratios immediately obvious if we wish to get the most out of our equipment. By keeping transmission lines short and by insuring that the SWR is as close to 1:1 as practicable, line losses will be minimized, maximum power will be delivered to the antenna and more successful and reliable radio communications will result.

CHECKING YOUR SWR INDICATOR

Carl Drumeller W5EHC

Many articles have been published on how to build and even some on how to calibrate a VSWR indicator. The calibration instructions usually tell you to terminate the indicator's output with a purely-resistive 52-ohm load and then to adjust the device so that a maximum forward and minimum reflected meter deflections are obtained. Sometimes they'll go further and tell you to reverse the device and recheck for opposite indications.

This is all very well. It assures you that the VSWR indicator will be telling you the truth when it says "All's well!" while looking into an utterly-flat transmission line. It doesn't tell you a thing about what the indicator will have to say when it gets tangled up with a line that has a wildly-mismatched termination.

As most transmission lines, in actual practice, are terminated in loads which are not only mismatched in the matter of resistance but also in the inclusion of a considerable magnitude of reactance, it would be well to explore the indications you'll get under realistic circumstances. After all, these are the situations under which you'd want to take corrective steps. Accurate indications of undesired conditions, therefore, are imperative if intelligent remedial actions are to be taken.

Fortunately, some quite enlightening tests are made easily. All you'll need are some lengths of coax transmission line (the same as you're using in your antenna feedline) equipped with male fittings at each end and a few female-to-female junctions. Select the frequency at which you want to make the test. Usually it's wise to make the test on the highest frequency band you plan to use. With this in mind, make up three one-eighth wavelength sections of transmission line and mount the male fittings on the ends of each section.

If your antenna presents an unmatched load to your transmission line, you may elect to skip over this paragraph and go directly to the next one. If it does not (Ah, you dreamer!), you'll need another piece of transmission line. It should be fairly long, perhaps a half wavelength. Put a male fitting at one end and attach a termination, which is deliberately made to be a sad mismatch, at the other end. Don't just mismatch it by using too high or too low a value of resistance. Throw in some reactance, too! You might use a resistor with an inductor in series. Or, you might try a capacitor in series with the resistor. Or, you could use either an inductor or a capacitor in parallel with a resistor. In fact, it would be best to experiment with all four!

Now that you have a transmission line available that you know is mismatched, you're ready to start the test. The first check (the "control", you might call it) is made with everything normal. That is, you'll have the transmitter feeding directly into the VSWR indicator's transfer box and the transmission line (either the one to your antenna or the substitute line to the mismatched load) attached to the output of the transfer box. Note the VSWR indicated. Also note your transmitter; insure that it's tuned to resonance and is adjusted to a power you can maintain throughout the first portion of the tests. Jot down these indications. Now, insert an eighth-wave section between the transfer box and the transmission line and without making *any* other changes or adjustments, note the VSWR. Repeat these steps, adding an additional eighth-wave section each time until you've used all three. Did you detect any change in VSWR? If there was even the slightest change, your VSWR indicator is not trustworthy!

Now for two more checks. Try varying the transmitter power output. Does this have any effect upon the indicated VSWR? If it does, your VSWR indicator is not trustworthy! Then try varying the transmitter output tuning, deliberately throwing the stage out of resonance. Does this have any effect upon the indicated VSWR? If it does, your VSWR indicator is not trustworthy!

Few VSWR indicators under the $150 class will pass these basic tests. If yours doesn't, don't be perturbed. You have an instrument that still has a useful field of application. You can use it as a comparative indicator. For instance, if you're adjusting the gamma match at an antenna, it'll serve quite well; in this application, you're holding all of the significant variables constant, with the exception of one (the gamma match), the effect of which you want to observe. Your tests will have shown you the parameters you'll have to hold constant for any other than simple comparisons. In all probability, you will have found that measurements taken with different (electrical) lengths of transmission line are invalid. Also, it's probable that, owing to the non-congruity of diode curves, measurements will have to be taken at precisely the same level of rf power if accurate comparisons are to be made.

If you'll keep its very real limitations in mind, you'll find that even an inexpensive VSWR indicator has excellent potentials for useful measurements. But don't ask it to perform at levels that even its expensive siblings can't attain!

HOME BREW BRIDGE CALIBRATION

Clifford Honess W4OAB

This table is useful for all homebrew SWR bridges and gives the % reflected power and the % full scale reading in the reflected mode, when the meter is set at full scale in the forward mode.

SWR	% Pwr. Refl.	Refl. Rdg. in % of Full Scale
1.1	.2%	4.8%
1.2	.8	9.1
1.3	1.7	13.1
1.4	2.8	16.7
1.5	4.0	20.0
1.6	5.3	23.1
1.7	6.7	25.9
1.8	8.2	28.6
1.9	9.7	31.1
2.0	11.0	33.3
2.1	12.6	35.5
2.2	14.0	37.5
2.3	15.5	39.4
2.4	17.0	41.2
2.5	18.4	42.9
2.6	19.7	44.4
2.7	21.1	45.9
2.8	22.5	47.4
2.9	23.8	48.7
3.0	25.0	50.0
4.0	36.0	60.0
5.7	49.0	70.0
9.0	64.0	80.0
19.0	81.0	90.0

MINI SWR BRIDGE

John Schultz W2EEY

Most SWR meters today are of the coupled variety which can be left in a transmission line while a transmitter is operated at full power. However, for a number of prolonged tune-up operations, involving antenna matching systems, for instance, such couplers have several disadvantages.

On 160 and 80 meters, especially, a reasonable amount of power is necessary to produce full deflection—up to 100 watts with some configurations. With a very low-powered transmitter, making adjustments at this power level certainly may damage the output tube or tank-circuit components with a high SWR. This will not be the case with higher-powered circuits but, in any case, a signal strong enough to cause needless QRM will be radiated.

Another disadvantage of the coupled SWR meter, if it is home constructed, is that it must be carefully calibrated since its response is very dependent upon the mechanical configuration of the coupling circuit. This is unlike the bridge-type SWR meter (described in this article) where a standard SWR curve may be used with a good degree of accuracy.

The above factors, plus the fact that I didn't need an SWR meter continuously in the transmission line, led me to construct the little resistance type SWR bridge shown in the photograph. It is just about as simple and inexpensive a unit as can possible be built.

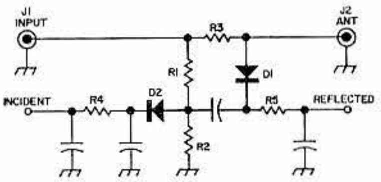

Fig. 1. SWR Bridge Circuit. See text for values of R_1, R_2 and R_3. R_4 and R_5 may be any matched value from 10 k to 47 k. D_1 and D_2 are IN34, IN54, or similar types. All capacitors are disc ceramic, .005 MF, 100V.

Construction

Two SO-239 coax chassis connectors are joined back to back by two 1¼ inch threaded hex spacers. The two four-lug terminal strips are mounted at the ends of one of the spacers. The wiring of diodes D_1 and D_2 as shown in **Fig. 1**, should be such that the incidental voltage-measuring point appears on the terminal strip mounted on the "input" SO-239 connector, in order to avoid confusion in measurement. Short leads, of course, should be used but hardly anything else is possible with only 1¼ inches between connectors.

Some attention must be paid to the components used if accurate readings are to be obtained. Resistor R_3 must closely match the impedance of the coaxial line used (52 or 75 ohms). For 52-ohm lines, a suitable resistor (within ½ to 1 ohm) can usually be found from a group of standard 10% tolerance, 47-ohm resistors; and for 75-ohm lines, from a group of 68-ohm resistors. Resistors R_1 and R_2 can have any value from about 30 to 100 ohms, but it is important that they are as closely matched as possible. One trick which may be used to affect very small resistance changes is to file "V" notches in a composition resistor to raise its resistance. Two-watt units are suggested for these resistors because of their longer-term stability and endurance in case too much input power is applied.

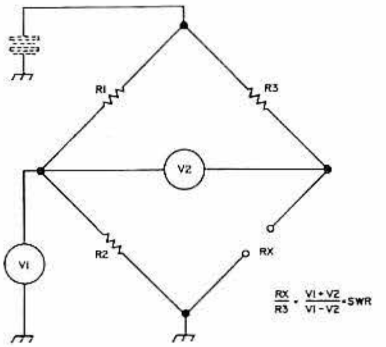

Fig. 2. Simplified diagram of SWR Bridge. V_1 represents incident voltage and V_2 the reflected voltage.

Resistors R_4 and R_5 serve as linearizing resistors so that almost any meter with a basic movement of 1 mA or less can be used as an indicator. The lower dc voltage ranges on almost any VOM will work fine. These resistors as well as diodes D_1 and D_2 should be checked to see that they match reasonably well (the resistors within a few precent and the diodes within a few percent for their forward and reverse resistance readings).

Calibration

There are really no adjustments that can be made to the bridge, and calibration really consists of checking the balance. **Fig. 2** is the dc circuit of the bridge (a simple Wheatstone bridge with resistance arms). If the balance of the bridge is good, V_2 should be the same when points R_X are opened or shorted so long as V_1 is held constant. This can be checked on the actual bridge by applying an input at the highest frequency of interest (6 or 10 meters), shorting J_2, and checking that incident and reflecting voltages are the same. The same is done with J_2 open. If the voltages are not equal, the difference can be taken as an indication of how accurate the SWR readings will be. If the difference is too great, R_1 or R_2 will have to be changed for a better match or the mounting of the components changed to reduce stray couplings.

A further check is to connect a known 52 or 75 ohm resistor across J_2. The reflected voltage should, of course, read zero.

Operation

As noted in **Fig. 2** the actual SWR is a simple function of the incident and reflected voltage readings. **Fig. 3** presents this function in graphical form. The incident voltage is simply adjusted for some convenient value, say 10 volts reading on the dc scale of a VOM (possible with most SSB transmitters by adjusting the carrier balance control with no audio input). The reflected voltage is then read as a percentage of the incident voltage and the SWR found from **Fig. 3**. The input power required to operate the bridge is essentially independent of frequency, being about 1-2 watts maximum.

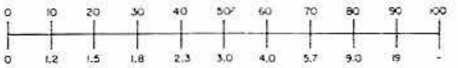

Fig. 3. SWR values for selected reflected voltage readings taken as % of incident voltage reading.

It should be remembered that such a bridge can measure only the resistive portion of an impedance. When using it to adjust a circuit, if a SWR minimum null but not a zero reading for reflected voltage can be obtained, it indicates some reactive component must still be present.

VSWR SUPREME

E. L. Klein W4BRS

One of the most valuable tools used by the amateur is the Voltage Standing Wave Ratio Meter. It ranks with the grid-dipper and the plate current meter as an indispensable instrument around the ham shack. We know that the VSWR meter is very handy in indicating relative power output when tuning a transmitter, particularly when the plate dip is not too discernable. It is most useful, however, in proving that the last available watt has reached the antenna where it can do some good.

Why a good match?

Although a good copper connection is made all the way to the antenna, an efficient transfer of power may not be achieved because of a mismatch between the characteristic impedances of the various portions of the transmission system. The interesting thing is that the "match" is different for each frequency because the antenna is essentially a single frequency device.

We can appreciate the importance of a proper match between the transmitter and the antenna when we are told, for example, that a VSWR of 3 to 1 causes a power loss of nearly 3 dB for 200 feet of RG-8/U coaxial line at 30 MHz. The table below provides

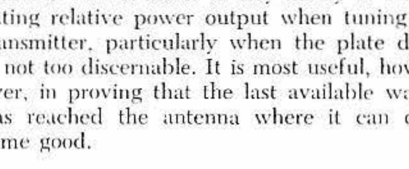

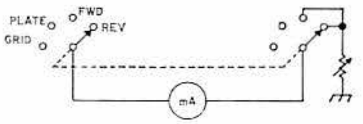

Fig. 1 Schematic of basic VSWR Meter and associated switching circuits.

Table I

VSWR	Power Loss (dB)	Transmitter Power Needed to Provide 1 kW at the Antenna
1.5:1	2.1	1600 watts
2:1	2.3	1700
3:1	2.8	1900
4:1	3.3	2000
5:1	3.7	2300
7:1	4.5	2800
10:1	5.3	3400

Additional power needed to compensate for a poor impedance match between transmitter and antenna. Figures are based upon 200 feet of RG-8/U cable at 30 MHz.

the real reason why we should be concerned with the impedance match. Notice how much the transmitter power would have to be increased to make up for a poor match between the transmitter and the antenna. Incidently, this match involves each and every part of the total transmission system including connectors, antenna relay, low-pass filter, balun, etc., as well as the transmission line itself and that particularly critical point at which it is connected to the antenna.

A new approach

Most VSWR meters today are an external accessory to the transmitter. But this practice is not good. Coaxial connectors are expensive and cause unwarranted mismatch and power loss. Meter faces usually end up behind the transmitter or in some other inaccessible location. When switching from forward to reverse, the little accessory box scoots across the table leaving scratches and a distraught operator.

The transmitter plate current meter is no longer a plug-in accessory. Why should the VSWR meter be? (Believe it or not, plate meters used to be plugged in with phone jacks.) Using the simple design described here, the home constructor as well as the commercial manufacturer can now build the VSWR meter into the transmitter in the smallest possible space and at only pennies of cost.

The circuit

Nothing is new about the circuit. It has been adequately described in the past in magazine articles and handbooks. However, for the convenience of the reader, the VSWR meter circuit is reproduced in **Fig. 1** for handy reference. Terminating resistors R1 and R2 should be 33 ohms for a 50-ohm transmission line when the physical configuration, as shown here, is used. One-half watt or smaller size resistors may be used. Diodes CR1 and CR2 are any matched pair of silicon diodes or germanium. The types which are enclosed in glass cases are the easiest to use because of their small size. The ohmmeter can be used to select and match the diodes of the ten-cent surplus variety found in advertisements in ham magazines such as *73*. Bypass capacitors C1 and C2 are 1500 pF Centralab type FT-1500.

Physical components

Parts used in this VSWR meter are illustrated in **Fig. 2.** The brass tubing is about 5 to 7 inches long and of 5/16 inch outside diameter. This size tubing fits snugly around the inner polyethelyne insulation from RG-8/U coaxial cable. About 10 inches of coax is stripped of its outer jacket and braid. The inner insulation is trimmed to extend 1/8 inch past each end of the brass tubing. Two large solder lugs are selected to fit over the 1/4 inch threaded shank of the bypass capacitors. These lugs should be of the long variety so they may be shaped and soldered to the brass tubing as shown in **Fig. 3.** Two 8-inch pieces of #22 enameled copper wire are also required.

Assembly

After soldering the lugs to the brass tubing about 5/8 inch in from each end, the bypass capacitors are assembled to the lugs. Place several fiber washers under the ring nuts prior to tightening them down on the threaded shank of the capacitors. This permits careful soldering of the capacitors to the lugs without danger of also soldering the nuts in place.

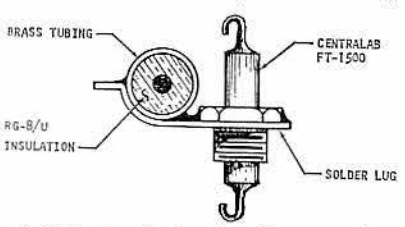

Fig. 3 Solder lug is shaped to fit one-quarter way around the brass tubing.

Two small grooves are now cut 180 degrees apart for the total length of the polyethelyne insulation. A small wood carving gouge or carefully manipulated razor blade can be used for this purpose. These grooves provide a space for the enameled copper wire which is held in place when assembling as shown in **Fig. 4.** Prior to this operation, the wire should be stretched and work-hardened by jerking it between two pairs of pliers. Be sure that the plane described by the two wires lies at right angles to the chassis on which the unit is mounted. This permits all resistors and diodes to have equal lead lengths.

When the inner assembly has been tugged and shoved into place within the brass tubing, the #22 wire ends are trimmed, stripped and soldered to their respective resistors and diodes. Much care should be exercised at this point to prevent melting the insulation or damaging the near zero-length component leads.

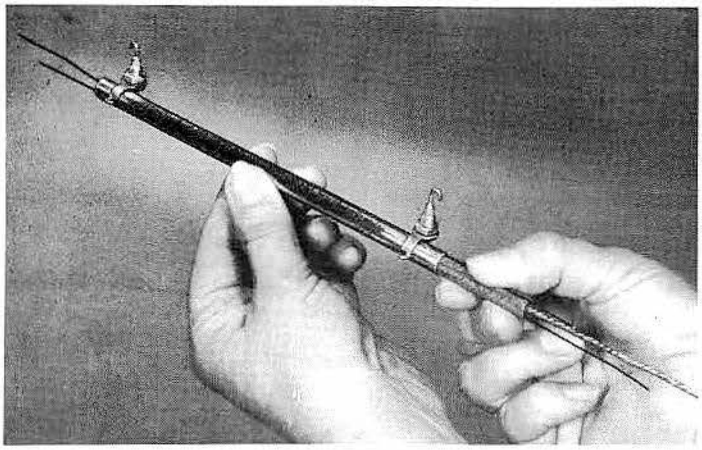

Fig. 4 Assembling the inner components into the brass tubing.

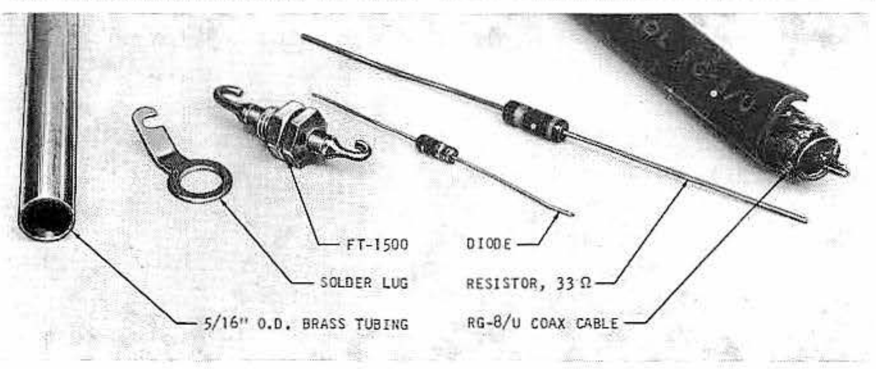

Fig. 2 Component parts used in making the VSWR sensing unit.

Application

A completed sensing unit for the VSWR meter is shown mounted on a typical chassis in **Fig. 5.** It will be noted that no conventional box or housing is used because the total outside of the unit is at ground rf and dc potential, save for the component connections at each end. By mounting the bypass capacitor in the chassis, the low-voltage rectified current fed to the meter switch is

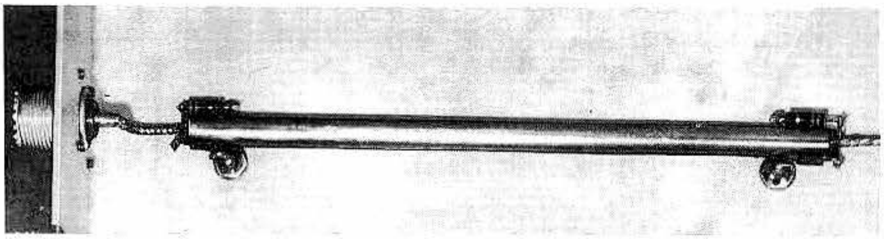

Fig. 5 Finished VSWR sensing unit mounted on a typical chassis.

isolated from high-power rf on the other side of the chassis. It can readily be seen that the finished sensing unit occupies no more space than would be used by a coaxial lead running from an antenna relay to the antenna connector on the chassis.

A further refinement is shown in **Fig. 6.** Complete isolation of the high-power rf is provided by the coaxial hood. Impedance discontinuity is also minimized by use of the hood, which was designed for this purpose and is readily available.

Length of sensor

The dimensions given for the length of the sensor element, including its outer tubing and inner conductor, are not critical. They are, however, directly related to the power of the transmitter with which the VSWR meter is used. For example, with a one kilowatt high-frequency CW transmitter and a 0-1 milliammeter as the indicating meter, the length of the sensor can be as short as 2-3 inches. A sensor which is constructed approximately 7 inches long, as illustrated in this article, will work fine with the same meter on a 25-200 watt high frequency transmitter. If meters with higher current ratings are used, a longer sensor is required, and, conversely, a more sensitive meter would provide adequate full-scale deflection with a shorter sensor element. Obviously, it is impracticable to vary the length of the sensor element in order to vary the sensitivity of the VSWR meter as a whole. It is for this reason that the adjusting resistor is provided in series with the meter. For VHF use, the sensor can be shorter.

All that has been said above can be depicted graphically. **Fig. 7** shows the generalized relationship between the sensor length and transmitter power with which it is used. The two curves represent different meter sensitivities. A 0-1 milliammeter offers a good compromise. With a 5-7 inch long sensor and the proper series resistor, all powers normally encountered in amateur work can be handled. However, if space requirements so dictate, a shorter sensor unit can be employed with some small sacrifice in accuracy.

Terminating resistor

Small variations in mechanical construction and lead dress will have an effect on the value of the terminating resistors, R_1 and R_2. Also, a carbon resistor does not display the same reactance at high frequencies as its measured resistance at dc. The value of the 33-ohm resistor was therefore determined empirically.

To verify the proper value of the terminating resistors, the test set-up shown in **Fig. 8** is used. A radio-frequency source of approximately 10 to 20 watts is required. A transmitter exciter stage operating on the 10-meter band is preferred for this purpose. Ten meters, or even fifteen meters, will provide better accuracy than one of the lower frequency bands. A dummy load is also required. This load must be capable of dissipating the power of the radio-frequency source used in making the test. Three or four 2-watt carbon resistors of the proper value in parallel to provide 52 ohms will suffice if the power is not left on continuously.

In making the test, the selector switch is first placed in the "forward" position. With power applied, immediately adjust the sensitivity control so that the meter reads full scale. Upon switching to the "reverse" position, the meter should read near zero and be at or below the 1:1 calibration point on the meter scale. Several resistors may be substituted until the proper value is found. The important thing to remember is that both of the resistors should be simultaneously substituted and that they must be as near identical as possible as measured on a reasonably good ohmeter. Lead lengths should also be as short as possible and of identical length.

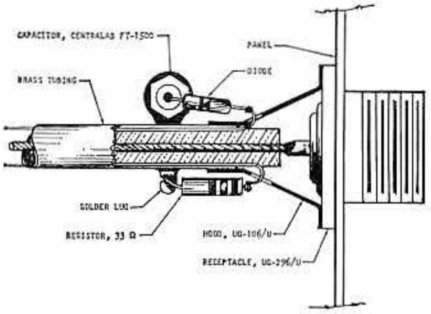

Fig. 6 Recommended chassis connection for output of VSWR meter sensing unit.

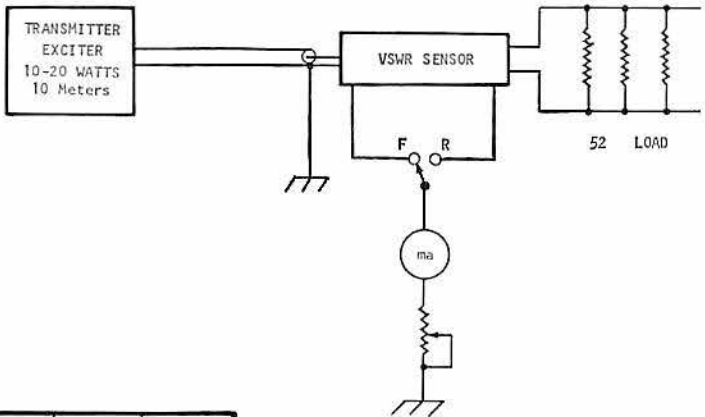

Fig. 8 Test set-up for verifying the proper value of the terminating resistors which are a part of the sensor unit. The value of the dummy load should match the characteristic impedance of the sensor unit and have a total wattage rating nearly equal to the source power.

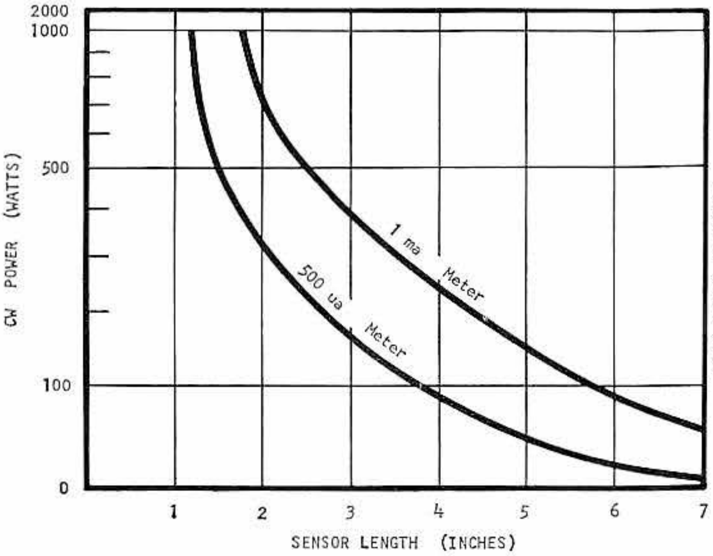

Fig. 7 The approximate relationship between sensor length and transmitter power is shown for two commonly used meter movements. Other meter values may be used as discussed in the text.

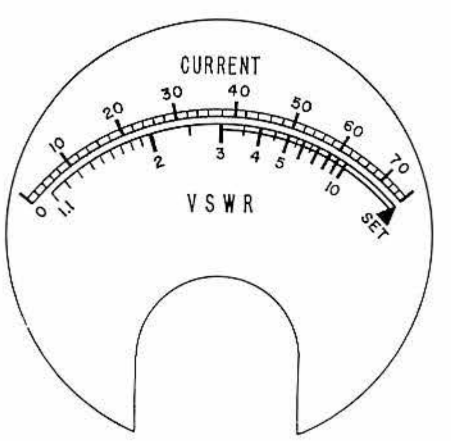

Fig. 9 Full-size photomaster of the VSWR meter dial. A multi-purpose meter was used in the author's transmitter so that the grid and plate currents could also be read on the upper scale.

The dial scale

Using the standard formula for calculating VSWR, it is possible to calibrate the meter face as follows:

$$\text{VSWR} = \frac{\text{forward} + \text{reverse}}{\text{forward} - \text{reverse}}$$

Fig. 9 is a full-scale illustration of a meter face used with the *VSWR Supreme*. This scale fits the Triplett Model 327, as well as a number of other meters of the same size category. A word of caution—don't assume that the scale calibration, or linearity will be the same for all makes of meters. The individual meter movement selected should be checked by using the above formula and marking off radials representing 4-5 different VSWR values. With the scale from your meter at the center of an oversized radial(s) drawing, it is possible to verify the angular placement of each VSWR calibration point.

The *VSWR Supreme* is truly a novel approach to an old standby. Using the construction methods outlined in this article, it is possible to fabricate the sensor unit so that it occupies the smallest possible space. This sensor can now be built into a transmitter and take up no more room than the coaxial lead which it replaces.

A SIMPLE VHF SWR METER

John Schultz W2EEY/K3EZ

I recently wanted to do some work on a 2 meter antenna and since no other instrument was handy, started to use an SWR meter manufactured for use on the high frequency bands. After some erratic results, it was decided to check the SWR meter accuracy on 2 meters with some carbon resistors to simulate different SWR's. The results readily confirmed that the SWR meter was useless at VHF unless one didn't care whether a SWR was really 1:2 or 1:5. Rather than purchase another SWR meter, it was decided to construct one that would render reasonable results, within 10% accuracy or so, on the VHF bands, particularly 144 and 220 MHz.

There is nothing basically new in the circuitry of the SWR meter to be described. What is different about it is that it utilizes a particularly simple and inexpensive method of construction that yields reasonable results. It can be constructed as a completely self-contained SWR meter or only the pickup element can be constructed and used with an external meter. The circuitry as shown here for the meter utilizes two meter movements so one can read forward and reflected power simultaneously and avoid the annoyance of having a forward-reflected switch arrangement for a single meter.

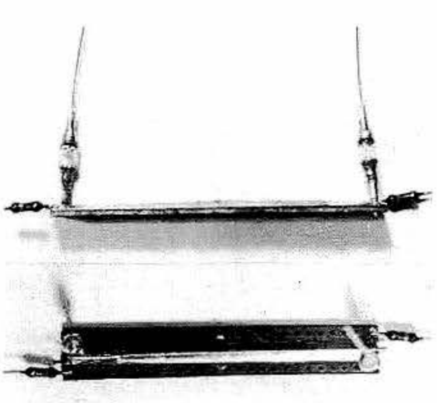

The heart of the SWR meter is a carefully constructed pickup element. Details of construction are discussed in the text but the photo shows how closely the diodes and terminating resistors must be soldered to the pickup element.

Pickup Element

The "heart" of any SWR meter of the reflectometer type is the pickup element. Many elaborate forms for such elements have been devised which involve complicated mechanical construction. Such complicated construction does become necessary if very accurate results are desired and if the meter is to maintain linearity over a very wide frequency range. However, over a smaller frequency range and with some minor sacrifice in accuracy, the construction of a pickup element can be greatly simplified. Basically, the pickup element should not cause any discontinuity in the transmission line section in which it is inserted but yet be long enough so enough voltage can be picked up in both the forward and reflected directions to make the meter usable with even low power transmitters.

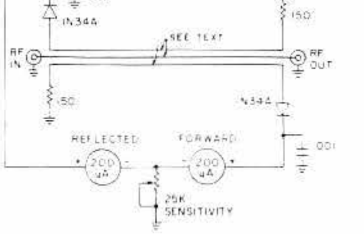

Fig. 1. SWR meter circuit. Two identical meters should be used (current range and internal resistance).

The pickup element I devised is shown in the photo. It is a 2-7/8" long piece of alternate grid pre-punched perf-board stock. The board is about 7/16" wide and within this width contained 4 separate copper strips spaced about 1/16" or less apart. The center two strips are soldered together to act as the inner conductor continuation of a coaxial line section. To solder the two inner strips together tack solder a bare piece of hookup wire between the two strips and then cover the entire two strips with solder.

Without the use of pickup wire, it will be nearly impossible to develop a smooth solder flow between the strip. Each outer strip acts as a pickup element for the SWR meter circuit shown in Fig. 1. The terminating resistor and diode are soldered to each end of the outer strip as shown in the photo and with minimum excess lead length to the strip. The use of a heat sink on the diode is necessary to prevent damage during soldering.

Mounting The Pickup Element

The pickup element made from the board stock is mounted between two approximately spaced SO-239 chassis connectors. The center strip of the board is soldered at each end to the center post of the SO-239 connector. The terminating resistor at each end is grounded as directly as possible to a ground lug held in place by one screw of the SO-239 mounting hardware. These details are shown in Fig. 2. It is important that the terminating resistor be grounded in this manner with minimum lead length. The enclosure in which the pickup element is contained should just be wide enough to accommodate the SO 239 connectors so that when the enclosure is secured together, the pickup element is boxed in by a metal surface on each side except directly above it. Many chassis or enclosure types are suitable for this purpose and the overall size of the enclosure will depend, of course, on the meter used and sensitivity control placement. These details are not covered here because they can be made as desired. They will not affect the basic accuracy of the meter as long as the pickup element is properly mounted and enclosed. The bypassing of the pickup rectifier diodes must also be done with as short leads as possible on the bypass capacitor. As shown in Fig. 2, a two lug terminal strip (one lug grounded) mounted on the side wall of the enclosure immediately at the cathode end of the diode will perform this task very well. The length of the IN34A diodes is such that the bypass capacitor cannot be connected to the same ground lug used for the terminating resistor. The leads on the capacitor would be too long and it will be ineffective.

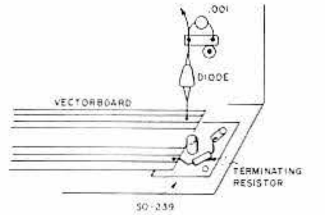

Fig. 2. Center strip of vector board is soldered to center post of SO-239. Other components are mounted at each end of board as shown (only one end shown here).

Operation and Results

The dual meter circuit of Fig. 1 reads forward and reflected power simultaneously. The sensitivity potentiometer is set to read full scale on the forward meter and the SWR read directly from the reflected meter. The latter meter can be calibrated for various SWR's by the use of small carbon resistors (100Ω to simulate a 1:2 SWR in a 50Ω line, etc.). Usually, it is only necessary to calibrate the reflected meter for SWR's of 1.5, 2 and 3 at the frequency of interest. Calibrated in this manner, the accuracy of measurement will be roughly 10% and is certainly good enough for most general antenna work. A particularly nice meter display can be made if one can find a two meter movement in one enclosure. I purchased a surplus stereo VU meter which had dual 200 μA movements and used it in the SWR meter.

The sensitivity of the SWR meter is such that transceivers of the 1-2 watt output class can easily be used with it on the VHF bands. The basic meter, of course, can also be used on the lower frequency bands as well and it will retain good accuracy. The only problem on the lower frequency bands is that the pickup strips are so short that more transmitter power has to be used to activate the meter than is usually convenient to use during antenna experiments. No exact tests were made but probably 70–100 watts would be needed to use the meter on as low a band as 80 meters. The meter was used and checked, however, on 40 meters. The accuracy of the meter remained very good and full scale deflection of the forward level meter required a power level of 60 watts. Being an in-line type meter, it can be left permanently in line when used on any band with a minimum of loss.

DIRECTIONAL COUPLER AND VSWR BRIDGE FOR VHF AND UHF

Bob Kolb WA6SXC

The aerospace industry has fostered the development of many new components and materials. Hams, being the kind of people they are, are quick to see practical applications for these materials that never occur to design engineers. I have often heard the criticism that it is impractical to publish articles or design ham gear with these new or expensive materials, because most OM's don't have access to them. Yet I've often been dismayed when I learn of an application for a piece of surplus equipment after it is no longer available. For this reason I feel that we should publish any application that is practical regardless of how immediately it can be used. Sooner or later, the material will show up on the surplus market and then we'll know what use can be made of it.

Reliable test equipment for the VHF-UHF bands is difficult to come by on a low budget. The literature is full of "relative" measuring devices but few pieces of homebrew gear are engineered for repeatable performance. Several directional couplers have been built according to the description in this article and each has measured to within ±.5 dB of the designed value. This is due in part to the mechanical rigidity and close tolerance of RG-141 "coaxitube." No special tools are required outside of a cheap pair of vernier calipers. The tools used to make the original coupler were an Xacto knife, file, soldering gun, vernier calipers and a vise. Don't let the calipers scare you. If you're not after a closely calibrated device they may be omitted.

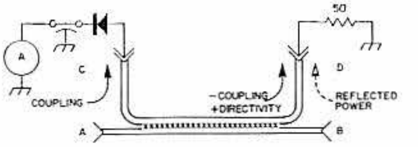

Fig. 1. A typical directional coupler. This device is the heart of a VSWR bridge, and can also be used for many other applications.

The design goal was a directional coupler with about 30 dB directivity in the passband with a low insertion loss. Each milliwatt measured at the coupling arm equals one watt through the main line. Such a device is the heart of a good quality VSWR Bridge. The measured values were 30.3 dB coupling and 38 dB directivity at 432 MHz. Data presented in the graph was taken using HP608C and 614A signal generators and a General Microwave R. F. Power Meter. The measured insertion loss was 0.2 dB.

Resolution of the smallest possible VSWR is limited by directivity. Few of the handbook VSWR bridges or the low cost type attractive to the CB trade achieve as much as 20 dB directivity. Thus the minimum discernable VSWR is approximately 1.7:1. With 38 dB directivity, 1.02:1 VSWR's can be accurately measured.

Directivity may be defined as the isolation of arm D from arm A, over and above the coupling as shown in the **Fig. 1.** Coupling is achieved by removing part of the jacket between adjacent coax conductors. If the input is at arm A, incident power can be sampled 30 dB down at arm C but appears −68 dB at arm D. Reflected power entering arm B is sampled −30 dB at arm D while at arm C it is −68 dB. It stands to reason if the directivity is low, one cannot tell with certainty if he is measuring incident or reflected power. Port D may be used as the dc return for a detector at port C and vice versa.

This device will have its fundamental passband where the length exposed between the two lines is $\lambda/4\sqrt{\epsilon_r}$. It will also have a passboard at $(2n - 1)\ \lambda/4\sqrt{\epsilon_r}$ or at three, five, seven, etc., times the frequency for which it is a quarter wave. Hence a coupler designed at 432 MHz is usable at 1296 MHz.

This coupler has also been used to measure relative power and modulation at 2 meters where its coupling factor for incident power is approximately −40 dB but the directivity is poor, hence arm D must be terminated in 50 ohms. It's a real aid for tune up and will give a good indication of increased power with AM modulation right in the r.f. line. RG141 will handle 500 watts of rf up to 2000 MHz.

This UHF directional coupler is very simple to make, yet offers excellent performance on the 70 and 23 cm bands.

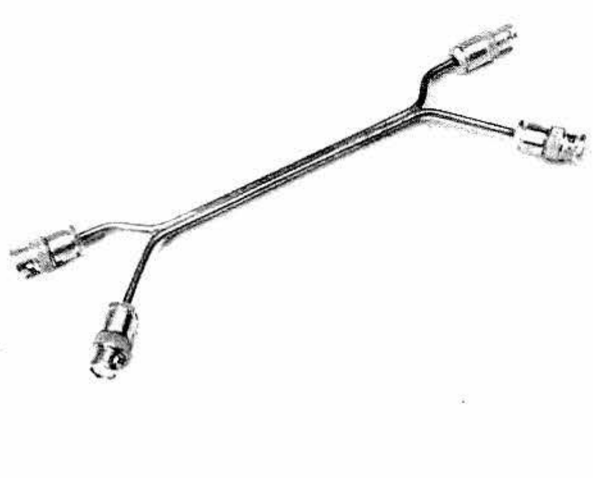

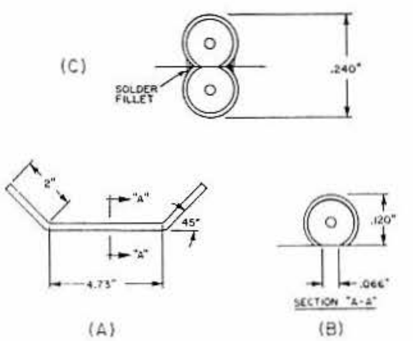

Fig. 2. Details of the construction of the UHF directional coupler.

The formula for determining coupling length is

$$\frac{c}{4f_o\sqrt{\epsilon_r}} = \frac{\lambda c}{4\sqrt{\epsilon_r}} \text{ or } \frac{300 \times 10^8 \text{ cm}}{4 \times f_o \times \sqrt{2.1} \times 2.54 \ ^{\text{cm}}/_{\text{in}}}$$

$$= \frac{\lambda \text{ c inches}}{4\sqrt{\epsilon_r}}$$

$$\sqrt{\epsilon_r} \text{ for Teflon} = \sqrt{2.1} = 1.449$$

From these calculations $\frac{\lambda\text{coupling}}{4}$ at 432 MHz is 4.73 inches. With an Xacto knife cut two pieces of line 8.73 inches long and carefully bend them so that they form the shape shown in **Fig. 2A.**

Clamp the bent coax into the vise and file away the copper jacket taking care that the filed surface is smooth and flat. A belt or stationery disc sander works well too. A cross section of the filed piece should look like **Fig. 2B.** Next fit the two pieces together so that a cross section would look like a figure 8 and secure in a vise. Heat with a soldering gun only. Do not use a torch. Avoid excessive heating. Flow solder between the two lines as shown in **Fig. 2C.** The "arms" can now be bent into any convenient configuration provided enough allowance is made at the ends for connector assembly. A good rule to follow is a minimum bend radius of half an inch although a quarter inch radius is permissible. The arms should be approximately two inches long.

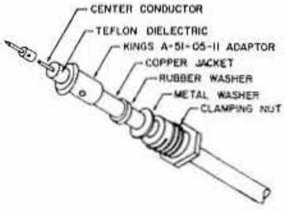

Fig. 3. Use of Kings A-51-05-11 adapter for using GR-141 with standard BNC connectors.

RG 141 has the same cross section as RG 58/U *without* the vinyl jacket therefore any connector that will accept RG 58/U can be used on RG 141 provided a sleeve is made up to make a snug fit in the clamping nut. A special adaptor is made by Kings for this purpose and sells for 45 cents. The connector assembly is shown in **Fig. 3.** Three RG 88E/U and one RG 89C/U connectors were used on the coupler shown in the photo.

Fig. 4 gives the measured directivity and coupling for this type of directional coupler at both 70 and 23 cm. You can see that performance is quite satisfactory.

Fig. 5 lists a number of applications for a directional coupler. The detectors in the measuring instruments should be suitable for use at 500 or 1300 MHz.

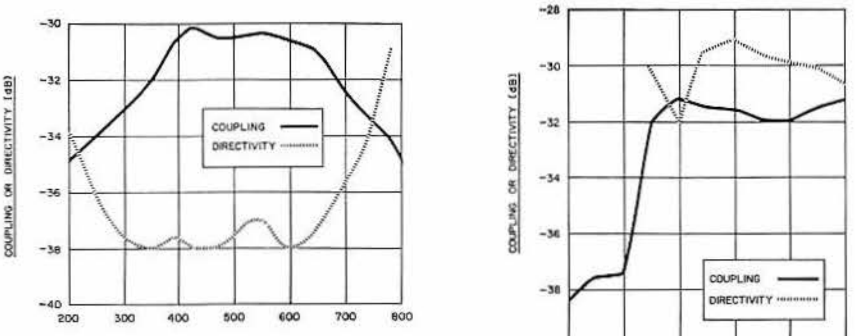

Fig. 4. Coupling and directivity for a directional coupler similar to the one discussed in the text. This device used a coupling wavelength of 4.635 inches rather than the 4.73 inches specified in the text. The only effect of the longer wavelength is to center the curves on 432MHz instead of about 500 MHz.

A SLOTTED LINE FOR 1250 MHz

Silas Smith WA8CHD

The SWR bridge is a very useful—and sometimes badly neglected—tool. Especially on 1250 mc with only two or three watts of power, you should measure your SWR and do something about it if necessary. I recently joined the 23 cm boys, though I have only worked 1234 mc so far. My first attempt to communicate with W8VKQ on the other end failed but after two evenings of diligent work with a meagre amount of test equipment we made contact. It would have been a lot easier with test equipment good in the 1250 mc range since most hams are not equipped to measure frequency and adjust their rigs at these frequencies.

After that experience, I made the slotted line indicator described in this article. The cost is next to nothing and it's easy to build, but it does a good job. The unit is built around a one inch thin wall copper tube 10¼" long (Fig. 1). This tube has a ⅛" slot cut lengthwise for 7½", about ¾ wavelength at 1250 mc. This is long enough to get a fair sampling of the standing wave or null points.

Three other pieces of 1" copper pipe are needed. Two are 5/16" long and one is 1⅜" long. The two 5/16" pieces each have ⅛" cut out and are reshaped to fit inside the 1" tube. These two pieces were each soldered to 1/16" plates as in the detail in **Fig. 1.** This process makes two cup-like structures which should

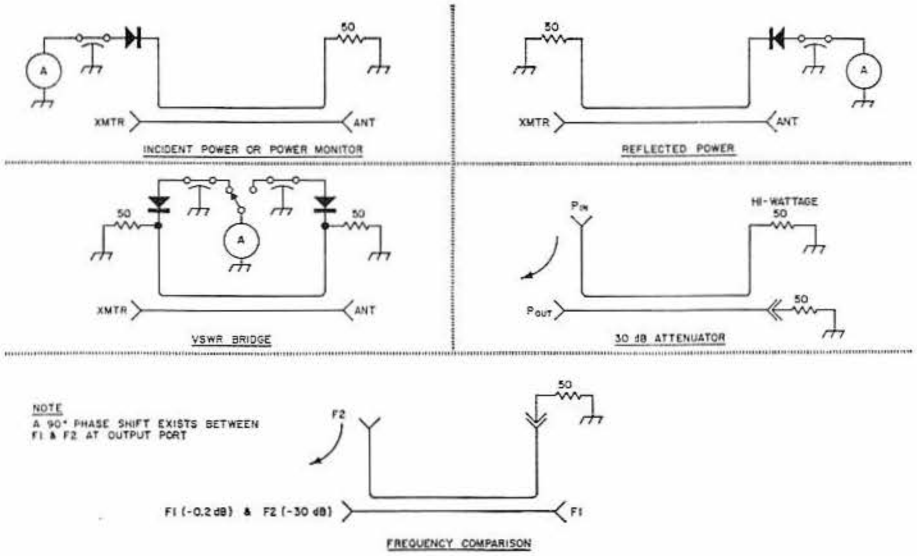

Fig. 5. Applications of the directional coupler described in the text. Unlike most pieces of ham-made test equipment, this one is good at 450 and even 1300 MHz.

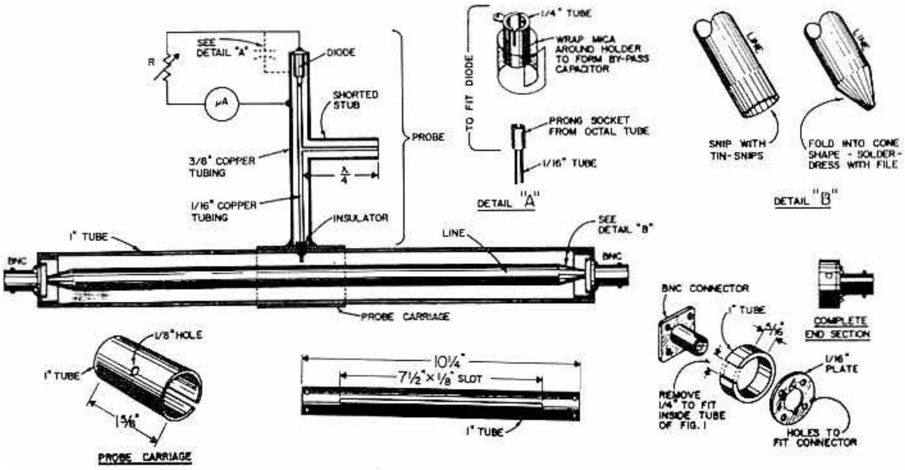
SEE DETAIL "A"
R
μA
DIODE
SHORTED STUB
PROBE
3/8" COPPER TUBING
λ/4
1/16" COPPER TUBING
INSULATOR
1" TUBE
BNC
LINE
PROBE CARRIAGE
1/8" HOLE
1" TUBE
1⅜"
PROBE CARRIAGE
10¼"
7½" × 1/8" SLOT
1" TUBE
TO FIT DIODE
1/4" TUBE
WRAP MICA AROUND HOLDER TO FORM BY-PASS CAPACITOR
PRONG SOCKET FROM OCTAL TUBE
1/16" TUBE
DETAIL "A"
SEE DETAIL "B"
BNC
LINE
LINE
SNIP WITH TIN-SNIPS
FOLD INTO CONE SHAPE - SOLDER - DRESS WITH FILE
DETAIL "B"
BNC CONNECTOR
1" TUBE
5/16"
COMPLETE END SECTION
1/16" PLATE
REMOVE 1/4" TO FIT INSIDE TUBE OF FIG. 1
HOLES TO FIT CONNECTOR

Fig. 1. The slotted coaxial line and probe for 1215 to 1300 mc described in this article.

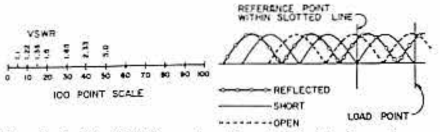
VSWR
1.1
1.22
1.35
1.5
1.85
2.33
3.0
0
10
20
30
40
50
60
70
80
90
100
100 POINT SCALE
REFERENCE POINT WITHIN SLOTTED LINE
REFLECTED
SHORT
OPEN
LOAD POINT

Fig. 5, Left. VSWR scale. Fig. 6, Right. Impedance measurements.

fit within the end of the 1" pipe. Drill nine holes in each cup: four holes for attaching the cup to the slotted line, four to fit the BNC connector mounting plate and one ⅜" hole to pass the main body of the connector. File two notches for the connector ears. You can fasten the cups to the slotted tube with small sheet metal screws, threaded holes or nuts soldered to the back of the holes.

The 1" x 1⅜" piece of pipe is for the probe carriage. It is cut lengthwise on one side, slipped over the 1" tube, centered, and a ⅛" hole drilled through it over the slot in the 1" tube.

The probe is built from ⅜" and ⅛" brass tubing from your hobby shop. At the probe end is an insulator from a coax fitting and on the other end is a jack made from an octal tube socket. A small piece of ¼" brass tubing is used to make the diode socket (see detail A). A piece of 22 gauge wire is inserted in the probe end of the 1/16" tubing and soldered. A 1N21 diode works very well.

Use **Fig. 2** to choose the proper size for the line in the center. Each end of this line should be tapered and soldered to the coax fitting as in detail B. The center conductor should be 7/16" for 50 ohms or 5/16" for 73 ohms.

The meter used should have a range that permits a full scale reading and should be calibrated with some scale.

Uses: Measuring frequency

Frequency can be measured with the slotted line in two ways: the distance between null points at one half wavelength ($\lambda/2$) or the distance between peaks. See **Fig. 3.** Measuring between the nulls is preferable since the peaks are very broad. If your line is flat, the peaks and nulls may be very small, so you may have to induce a mismatch in the line. Likewise, you may have to induce a mismatch to get the two nulls to fall within the 7½" of the line. The easiest way to induce a mismatch is a short in the line.

Measure the distance between the nulls carefully with a centimeter rule. Twice this distance divided into 30,000 will give you your frequency in megacycles. It would be a good idea to check this against a standard or at least avoid operating too close to the edge of the band.

Measuring VSWR

Measuring voltage standing wave ratio (VSWR) with the slotted line is easy, too. We are looking for a ratio of E_{min}, the minimum voltage on the line, and E_{max}. The ratio:

$$\frac{E_{max}}{E_{min}} = \text{VSWR}$$

However, there is a slight complication of vectors in E_{max} and E_{min}, so that to find them, you must first determine the highest voltage read on the meter, E_1 and the lowest voltage read on the meter, E_2 with the same setting of the shunt resistor, as the probe carriage is slid along the line. Now, $E_{max} = E_1 + E_2$ and $E_{min} = E_1 - E_2$ so that

$$\text{VSWR} = \frac{E_1 + E_2}{E_1 - E_2}$$

An even simpler method is to set the meter to full scale (100) at the highest reading, and reading the VSWR directly from **Fig. 5.**

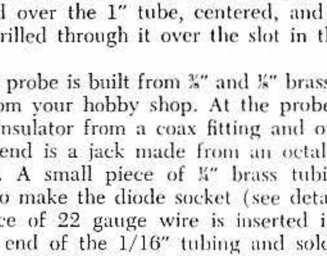

Fig. 2. Ratio of outer tube to inner line for various impedances.

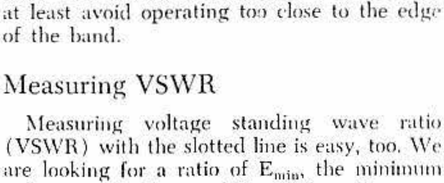
GENERATOR
SLOTTED LINE
LOAD
$\frac{\lambda}{2}$
STANDING WAVE
$\frac{\lambda}{2}$
GENERATOR
SLOTTED LINE
LOAD
FORWARD WAVE
REFLECTED WAVE

Fig. 3, Left. Measuring frequency. Fig. 4, Right. Measuring VSWR.

Impedance measurements

It's a little harder to measure impedance, but even it's not bad if you do it step by step. First, consider the high impedance on your line when your antenna is open or shorted. Perhaps you don't think of it as a change in impedance. This is what we are looking for. Remember that there is a definite relationship between impedance and frequency. **Fig. 6** shows the voltage relationship between a short and open. On a short set the probe for a null point, usually the first null from load end of the slotted line. This will be our reference point. You will note that when the load point is open there is a shift of 90° or a quarter wavelength either side of the reference point. This adds up to one half wavelength. Any impedance between infinity and zero ohms will lie somewhere along the half wavelength of the line. It will be noted that an open is equidistant from the reference point toward the generator and toward the load, however the line has moved toward the load (capacitive reactance). To measure this reactance, we replace the load that we measure with a short. We set our probe to the first minimum reading from the load end of the slotted line. Mark this spot with a scribe. This is our reference point. Now we remove the short and add our load under measurement. It will be noted that the voltage has moved upward. Move the probe toward the load. If the voltage goes down, you are heading in the right direction. If the voltage starts to go up change direction toward the generator. Move the probe as above to the new null point and note the direction that the probe has been moved.

Here's a step-by-step example. First, you'll need the frequency and wavelength, VSWR, direction of probe, distance between reference point and new null, and impedance of the slotted line.

Let's say as we measured the frequency, the distance between the null points was 12 cm. This is half the wavelength, so the wavelength is 24 cm. The frequency is 30000/24 or 1250 mc.

Let's say we read 80 as a maximum and 20 as a minimum on a point scale in measuring the VSWR:

$E_{max} = E_1 + E_2 = 80 + 20 = 100$
$E_{min} = E_1 - E_2 = 60$

$$\text{VSWR} = \frac{100}{60} = 1.66$$

Or set E_1 to 100 by adjustment of R, then E_2 would read 25 on the meter, a VSWR of 1.66 by **Fig. 5.**

As in **Fig. 6,** the probe was moved toward the load.

Let's assume the distance from the reference point to the new null was 2 cm. What part of a wavelength is 2 cm? Wavelength is 24 cm, so 2/24 = .083 wavelengths.

This is a movement of .083 wavelengths toward the load.

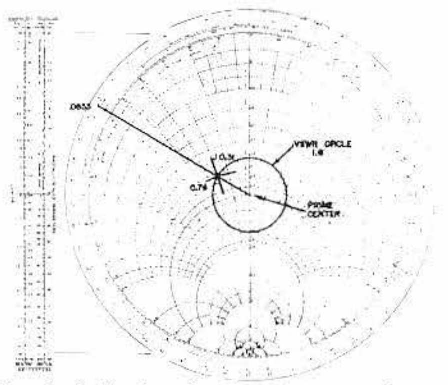

Fig. 7. Smith chart for determining impedance.

The slotted line used has an impedance of 50 ohms.

We are now ready to plot this information on the Smith chart (**Fig. 7**, better get out your magnifying glass).

Draw the VSWR circle at the prime center with a radius of 1.6.

Draw a line from .083 wavelength toward the load from the prime center.

At the intersection of the circle and line, follow the constant resistance circle to .74 and the capacitive reactance circle to .31. Our load impedance is then:

$Z_L = Z_o\,(.74 - .31j)$
$Z_L = 50\,(.74 - .31j)$
$Z_L = 37 - 15.5j$

Well, there are some of the things your slotted coax line can do besides telling you that you're on the air. A commercial version would cost $750. I guarantee that you'll get your dollar's worth out of this one as it costs less than $1 to build. I hope that it helps you get on 1250 mc.

SOME DIRECTIONAL WATTMETERS AND A NOVEL VSWR METER

P. G. Martin G3PDM

Most conventional reflectometers cannot be used for accurate power measurement because their sensitivities are frequency dependent. This is due to the use of combinations of reactance and resistance in the sampling circuits which detect the transmission line current and voltage.

This basic problem can be solved by the use of conventional lumped components instead of the distributed parameters of a transmission line. The line voltage can be sampled by two resistors or two capacitors used as a voltage divider, rather than one resistor and some distributed capacitance. The line current can be monitored by a properly designed current transformer instead of an inductance and resistance. High frequency current transformers consist of primary and secondary windings on a ferrite or iron dust toroidal core, with a low value of load resistance across the secondary winding.

All SWR bridges and directional wattmeters need to generate two dc voltages proportional to the forward and reflected voltages or currents of the transmission line. To achieve this one has either the current detector or the voltage detector providing two antiphase signals so that addition and subtraction can be performed.

A Frequency-Independent Directional Wattmeter

M. B. Allenson G3TGD, has designed a wattmeter using the above principles, where the low resistance in the current transformer secondary circuit is split into two equal parts. The center connection is taken to the voltage sampling point so that sum and difference voltages are available at the ends of the transformer secondary winding, see Fig. 1.

With two meters, this circuit can be used as a versatile calibrated directional wattmeter over the freqnecy range 100 kHz to 70 MHz, with an accuracy of about 10 per cent. Precise calculations of SWR and transmitter efficiency can be made.

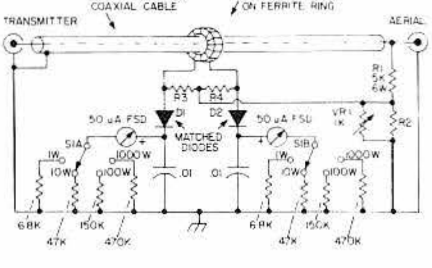

Fig. 1. Circuit of the basic frequency-dependent directional wattmeter due to G3TGD. The two meters indicate forward and reflected powers.

Maximum sensitivity with a 50 μA meter is less than five milliwatts, but with the multiplier resistors specified in Fig. 1, full scale deflection corresponds to power of 1, 10, 100 and 1000 watts. Calibration is non-linear, because the instrument samples voltage, and power is proportional to voltage squared.

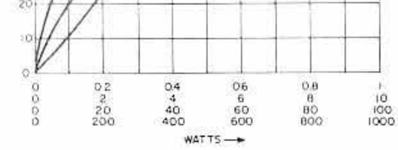

Fig. 2. Calibration curves for the instrument described in Fig. 1.

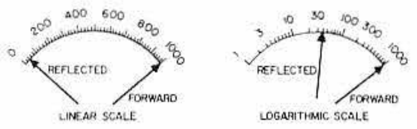

Fig. 3. Linear and logarithmic scales. The inherent advantages of the logarithmic form are immediately obvious.

Unfortunately, two transmission line impedances are in common use in coaxial systems: 50Ω and 75Ω. As it is not possible to design instruments whose sensitivities are independent of line impedance, some component values must depend on the impedance in use. For simplicity, only one of the voltage driver resistors need be changed, but instrument calibration will be different. By changing the current transformer resistors as well as one of the voltage divider resistors, the calibration is the same for both line impedances. This technique has been adopted here, and the calibration curves in Fig. 2, are correct for 50 or 75Ω lines provided the resistor values in Table I are used.

The Logarithmic Wattmeter

The basic instrument can be simplified by including a logarithmic network so that the power range switch is redundant and a single meter scale can be used for powers from, say, one watt to 1000 watts. A logarithmic scale has the 1, 10, 100 and 1000 watt points equally spaced (see Fig. 3).

The advantage of a logarithmic instrument is that one can measure very low reflected powers and very high forward powers simultaneously with the same percentage accuracy, without having to switch meter ranges.

It is simple to add a reasonably accurate *wide-range* logarithmic network to the meter in Fig. 1 (see Fig. 5). The basis of its

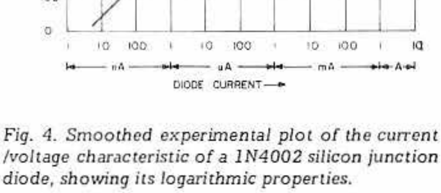

Fig. 4. Smoothed experimental plot of the current /voltage characteristic of a 1N4002 silicon junction diode, showing its logarithmic properties.

operation is that the voltage dropped across a forward-biased p-n junction diode is proportional to the logarithm of the current passing through it (see Fig. 4). To reduce the potential dynamic range of the circuit, a relatively insensitive meter is used, and a small resistance is added in series with the logarithmic diode to restore a logarithmic form to the scale (see Fig. 6).

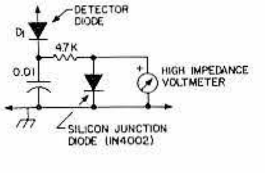

Fig. 5. Basic wide-range logarithmic converter.

An experimental logarithmic wattmeter is shown in Fig. 7. Figure 8 gives the calibration scale for 50 or 75Ω lines, provided the correct resistors are used (Table I).

A Direct-Reading SWR Meter

A particularly useful device would be an instrument giving a direct measurement of the standing wave ratio on a transmission line, independent of the absolute power levels or the frequency in use. Such an instrument, with its single meter, would be ideal for incorporation into transmitters and transceivers (especially with the physically small sampling circuits associated with it).

The swr can be expressed in terms of the forward and reflected voltages according to:

$$\text{swr} = \frac{E_f + E_r}{E_f - E_r} \qquad (1)$$

We wish to generate this function electronically, so that outputs of the two detectors can be used to generate a meter current proportional to SWR. This would be rather tedious, though not impossible.

Conveniently, manipulation of equation (1) shows that:

$$\frac{E_f}{E_r} = \frac{SWR + 1}{SWR - 1} \qquad (2)$$

which although not proportional to SWR, is a mathematical function of it only. Electronic division of E_f by E_r is easily done by taking logarithms and subtracting. That is:

$$\log \frac{E_f}{E_r} = \log E_f - \log E_r$$

Table I

	Ω	Ω
Line impedance	50	75
R3 and R4	27	33
R2	220	180

Values for R2, R3 and R4 to be used in 50 and 75Ω transmission lines.

In Fig. 9, the two silicon diode voltages are proportional to the logarithms of their currents, which in turn are proportional to the forward and reflected voltages. The two diode voltages can be subtracted directly by connecting a meter between them, rather than from each one to chassis.

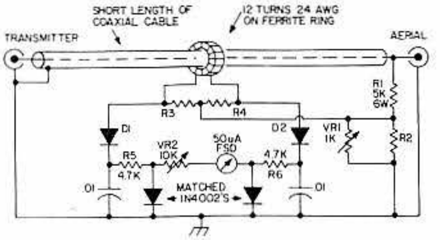

Fig. 6. Circuit of the logarithmic directional wattmeter. D3 and D4 are matched (see text).

The meter cannot be calibrated linearly in SWR, because of equation (2), and because the circuit does not take anti-logarithms after subtracting the logarithms. The outcome of this is beneficial: the SWR meter is increasingly sensitive as the standing wave ratio approaches 1:1. This is where one wants most sensitivity: to make the final adjustments to aerial arrays, to measure the variations in SWR over a band, and so on. Fig. 10, shows a calibration curve for SWR meters. Naturally the meter sensitivity cannot be completely independent of the power level in use. Accuracy falls when the reflected power is less than about half a watt (this corresponds to an SWR of 1.05:1 when the forward power is 1 kW).

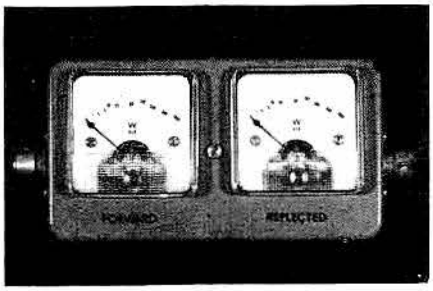

Fig. 7. An experimental logarithmic wattmeter.

A differential amplifier could be added to the circuit of Fig. 9, to enable a less sensitive meter to be used.

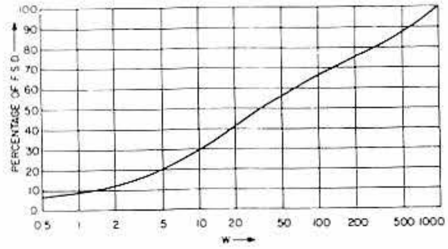

Fig. 8. Calibration curve for logarithmic wattmeters.

Construction of the Instruments

Layout of the sampling circuits is fairly critical, see Fig. 11. The input and output sockets should be set a few inches apart, and connected together with a short length of coaxial cable. The coax braid must be grounded at one end only, so that it acts as an electrostatic screen between the primary and secondary windings of the toroidal transformer. Twelve turns of 24 AWG enamelled wire, equally spaced around the circumference of the ring, form the secondary winding. The primary is formed by simply threading the ring onto the coax.

A suitable ferrite ring is the Mullard FX1596, made in England, although other types are suitable. The FX1596 has an outside diameter of half an inch, and is designed for wideband rf applications between 5 and 20 MHz. The main requirement is that the ferrite material should maintain a high permeability over the frequency range in use.

Other components in the sampling circuits should have the shortest possible leads. R1, R2 and R must be non-inductive solid carbon types: for high power levels (about 100 watts) R1 should consist of two or three 2 watt carbon resistors in parallel. VR1 should be a miniature skeleton potentiometer to keep stray reactance to a minimum, although it may be dispensed with by trying various fixed resistors for R2 until the reflected indication under matched conditions is zero.

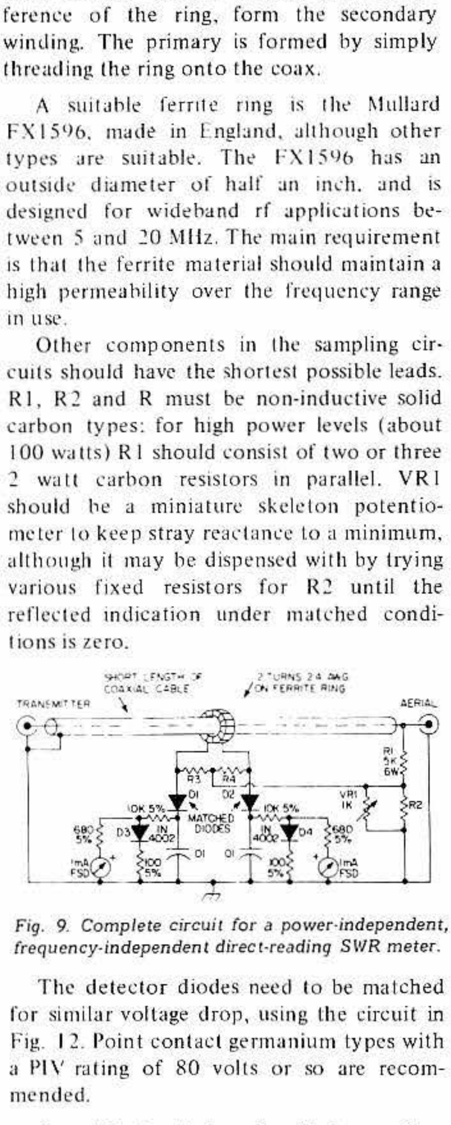

Fig. 9. Complete circuit for a power-independent, frequency-independent direct-reading SWR meter.

The detector diodes need to be matched for similar voltage drop, using the circuit in Fig. 12. Point contact germanium types with a PIV rating of 80 volts or so are recommended.

Logarithmic diodes should be modern medium-current silicon junction types, such as conventional rectifier diodes. The 1N4002 is specially recommended for its good logarithmic properties. Log diodes should also be matched with the circuit in Fig. 12.

The 0.01 µF decoupling capacitors should be a disc ceramic type.

In designing a toroidal transformer different to that specified, several factors should be borne in mind. As the number of secondary turns increases, the self-capacitance rises and causes the response to fall at high

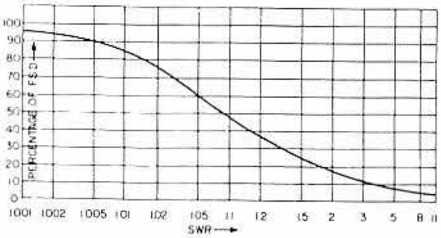

Fig. 10. Calibration curve for SWR meters of the type described in Fig. 9.

frequencies. Failure of this nature causes the reflected power indication to rise; in other words the directivity of the instrument falls. If the 27Ω resistors are raised appreciably in value, the instruments will eventually become frequency sensitive.

The ratio of the voltage sampling resistors (R1 and R2) in the HF designs is determined by the sensitivity of the current sensing circuit, and the two sampling voltages must be equal in magnitude under matched conditions. VR1 provides fine adjustment of the ratio. Absolute values of the resistors can be varied considerably, bearing in mind that as their values increase the stray capacitance across them may need to be compensated for.

Useful Equations

Let the line current be I amps, the line voltage be V volts, and the characteristic impedance of the transmission line be Z_o. Then $V = IZ_o$.

If the current transformer has a ratio of 1:n, and each of the resistors in its secondary circuit has a value of RΩ, then the rf voltage across each of them is given by:

$$V_{(i)} = \frac{IR}{n} \qquad (3)$$

The voltage detector output is obviously

$$V_{(v)} = \frac{VR_2}{R_1 + R_2} = \frac{R_2}{R_1 + R_2} IZ_o$$

Which is, to a good approximation,

$$V_{(v)} = \frac{R_2}{R_1} IZ_o \qquad (4)$$

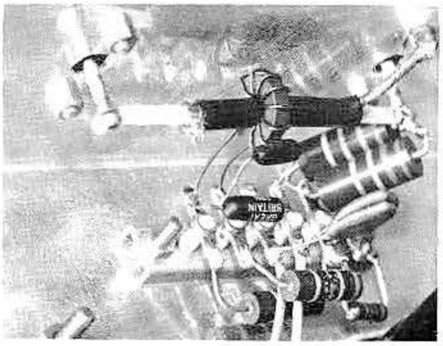

Fig. 11. Photograph showing layout of sampling circuits used in an experimental swr meter.

The main design equation for all the HF instruments is therefore:

$$R_2 = \frac{R.R_1}{n.Z_o}$$

where the value for R2 includes the effect of VR1, if fitted.

The dissipation of some of the components specified is quite high. For those planning different circuits, the following equations express the dissipation of R1 and the current transformer resistors R:

$$W_{(R1)} = \frac{Z_o.W}{R_1} \quad \text{watts,}$$

where W is the transmitter output power.

$$W_{(R)} = \frac{W.r}{n2.Z_o} \quad \text{watts.}$$

In the instruments described, $W_{(R1)}$ is about 5 watts, and $W_{(R)}$ 2 watts for a transmitter power of 500 watts.

Calibration

If any of the instruments are built exactly as described, and used in systems of the correct impedance, the calibration given in Figs. 2, 8 and 10 will be sufficiently accurate for most purposes. For those designing their own circuits, the following procedure is recommended.

Test equipment needed includes a high power rf source (a transmitter) and an rf voltmeter. The instruments can be calibrated with less accuracy without the rf voltmeter. The wattmeters are calibrated by feeding power through the meter into an appropriate dummy load (50 or 75Ω). VR1 is set for minimum reflected power indication, and the power scale is marked according to the rf voltage appearing across the load. If an rf voltmeter is not available, a peak-reading type can be made with a diode, capacitor and dc voltmeter. As the detector output is equal to the peak rf voltage applied to it, equation (4) leads to:

$$V_{(det)} = 2.8\ V\ \frac{R_2}{R_1} = 2.8 \sqrt{WZ_o \frac{R_2}{R_1}}$$

It would be difficult for most amateurs to obtain sufficient high power carbon resistors to calibrate an SWR meter by means of deliberate mismatching. An indirect method is therefore recommended.

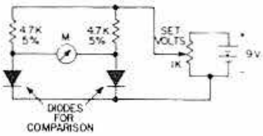

Fig. 12. Hookup circuit for matching detector diodes for equal forward voltage drop, and silicon junction diodes for similar logarithmic properties. The meter should be as sensitive as possible (say 50 µA fsd), and should not deflect appreciably as the voltage is varied between zero and nine volts.

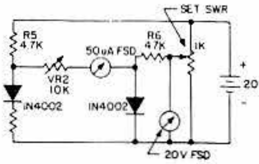

Fig. 13. Circuit used to calibrate SWR meters(see text).

Disconnect R5 and R6, Fig. 9, from the detectors, and connect them instead as shown in Fig. 13. One voltage is fixed at about 20 volts, and the other is varied between zero and 20 volts. The ratio of these voltages corresponds to a definite SWR which can be determined from equation (1). Before carrying out this procedure, however, VR2 should be set for full scale deflection of the meter under matched conditions at the highest power level to be used.

Conclusions

All of the instruments described in this article have been tested under actual operating conditions, on all amateur bands between 1.8 MHz and 30 MHz. Power levels used varied from 100 to 1200 watts. With the components specified, the instruments will sustain power levels well above the kilowatt level for periods of tens of seconds.

It is hoped that by introducing frequency independent directional wattmeters, one will be able to make useful comparisons of absolute power levels and accurate assessments of standing wave situations. The logarithmic scales are an added convenience, and the direct-reading SWR meter offers a saving in meters.

SWR BRIDGES USING ZERO-CENTER METERS

John Schultz W2EEY/K3EZ

The basic coaxial cable swr meter, which samples both the forward and reflected voltages on a transmission line, is an essential instrument in setting up an amateur station. Although a low transmission line swr in itself doesn't guarantee that an antenna system will work, knowing what the swr is remains an essential bit of information in evaluating the status of the transmitter/antenna interface.

The virtues of the usual coaxial line swr meter – economy and simple circuitry – are offset by the awkwardness involved in using the instrument: that is, switching back and forth between "forward" and "reflected" switch positions and adjusting a sensitivity control for sull-scale deflection in the forward position. This awkwardness of operation is partly relieved by dual-instrument swr meters so the forward/reflected switch is no longer necessary – but the cost of the instrument increases. The awkwardness of operation can be completely relieved by special dual movement meters with crossing pointers and/or special circuitry, but then the cost really soars.

Another approach is the usage of some different circuitry to detect the swr which can then activate a zero-center meter so one can peak up the antenna coupler, etc., for minimum swr on a transmission line with the same ease as one tunes an FM entertainment receiver for a zero-center reading on the discriminator meter. In the end analysis, practically no line is set up for any specific swr, but simply for the lowest swr that can be achieved. So, why not make the process as simple as possible? Circuitry to do this has actually been available for many years and was especially developed for use with the automatic servo-tuning of military transmitters. The zero output when tuning conditions were correct, and the plus or minus output when they were not, allowed, after sufficient power amplification, the activation of motors to tune transmitter PA task circuits, coupling circuits, etc. Now, with the economical availability of toroids and imported zero-center meters, zero-center swr type circuits can be applied to amateur usage and meet the criteria of being easy to construct and economical. They offer a real challenge to the conventional type of swr meter, especially when one is using a multiband antenna and tuner arrangement, or just leaves an swr meter in a transmission line to verify that the line swr is holding constant. In the former case, one usually knows the approximate transmitter/tuner settings on each band, and the swr meter is used to peak up the adjustments. Such peaking-up, as mentioned before in the case of an FM receiver, is certainly facilitated by just having to look at a zero-center meter. In the latter case, a zero-center meter will more readily reveal slight changes in transmission line swr from an established norm. Perhaps the only place where a conventional swr meter is still to be preferred is when one is dealing with experimental antennas where extreme swrs exist and a conventional swr meter can help to more quickly obtain "ballpark" antenna coupler or loading coil settings. The zero-center type circuits operate on a slightly different basis than the conventional swr meter circuits, and just a brief review of transmission line operation will be helpful to understand the different basis for the functioning of the two types.

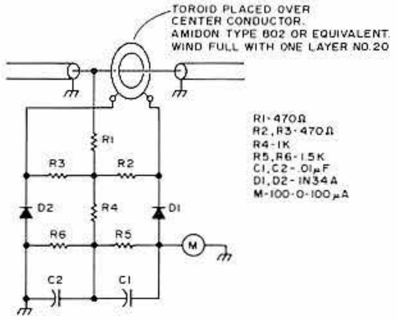

Figure 1.

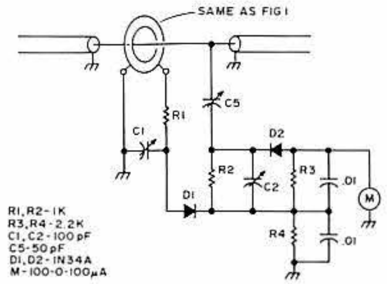

Figure 2.

If a transmission line is attempting to deliver power to a load, both voltage and current values exist along the line. If the antenna load impedance is resistive and equals that of the line impedance, all the power feed into the line reaches the antenna less any power loss in the line itself because of the line's inherent loss. If the load is not equal to the line impedance or reactive, part of the power the line tries to deliver is rejected by the load and reflected back along the line. This sets up standing waves of voltage and current which are not in phase with the original voltage and current waves the transmitter is trying to "pump" into the line.

Conventional swr meters sample a portion of the voltage or current wave traveling towards the load and a sample of that traveling back from the load. The relative amplitude of the two samples is the swr and, of course, the desired condition is that there be no reflected wave from the load. If there is no reflected power from the load, the line is often referred to as being "flat." This term is a bit misleading. Although the reflected power may be zero under matched conditions, the power flowing to the load via the line will still set up normal sine wave distributions of voltage and current along the line.

Another way to sense that the load is matched to the line is to determine first of all that the phase difference between the voltage and current on the line is zero and that the ratio of the in-phase voltage and current is correct for the impedance of the line being used. If there is no reactance present in the load, there will be no phase difference between voltage and current, as in any ac circuit feeding a resistive load. If the load is not only non-reactive but also of the correct resistance, the ratio of voltage and current on the line will have a definite relationship correct for a given line impedance. The zero-center meter circuits are designed to monitor both of these line conditions.

Practical Circuits

Two separate circuits are used to monitor the line conditions just described. One could switch a separate zero-center meter between the two line monitor circuits but with imported zero-center meters available for as low as $1.50, it is hardly worthwhile to do so. (A good source of the 100-0-100uA meters needed is Edlie Electronics, 2700 Hempstead Turnpike, Levittown NY 11756. Their meters with order number DA798 or DA792 cost only $1.50 but they have a $7.50 mail order minimum.) Also, for many installations, it is not necessary to use both circuits. For instance, in a home installation feeding a set of various fixed antennas, it would probably be of primary interest to have the impedance monitor circuit (voltage/current ratio). Once the antennas have been set up and operate satisfactorily, one would primarily just be interested to know that some impedance condition on the line has not changed. In a mobile installation, where some reactance cancelling L or C needs to be adjusted with frequency changes, it would be of prime interest to have the phase monitor circuit once the installation has been initially set up for a correct impedance match, and the reactance cancelling component only needs adjustment during actual operation.

Phase Monitor

Fig. 1 shows the circuit of the phase monitor. It is a basic discriminator circuit. The voltage on the transmission line to ground is sampled via R1. The current on the line induces a voltage in the windings on the toroid placed in the line which, of course, is 90° out of phase with the voltage sample. The voltage proportional to the line voltage is divided across R1 and R4 (C1 and C2 are rf bypass capacitors). The voltage proportional to line current (across the toroid winding) is divided between R2 and R3. Referenced to the center point of R2 and R3, these voltages are 180° out of phase with respect to each other and either one is 90° out of phase with respect to the voltage across R4. The voltages across R2 and R4 are rectified by D1 and, because of the diode polarity, produce a voltage with the polarity indicated across R5. Similarly, R3 and R4 and D2 produce a voltage of opposite polarity across R6. When the transmission line current and voltage are in phase, the voltages across R5 and R6 are equal and opposite, and the resultant voltage output which drives the meter is zero. For an out-of-phase condition, some output voltage exists whose magnitude indicates the amount of the phase difference and whose polarity would indicate whether the line voltage leads or lags the line current.

One could actually calibrate the output meter for different magnitudes and directions of phase difference, but this would be very tedious and not of real use for just checking the adjustment of a line for reactance cancellation. The only calibration that has to be checked is to excite the circuit into a resistive dummy load and to be sure that the zero-center feature works. If it does not, it indicates that some unbalance has been introduced during construction or some of the components (R2/R3 and R4/R5) are not closely enough matched. If the meter excursions off zero center become too violent with a high power transmitter, some additional resistance can be introduced in series with the meter to limit the maximum current flow.

Impedance Monitor

Fig. 2 shows the circuit of the impedance monitor, which is also a form of discriminator circuit. Samples of the transmission line voltage and current when they achieve the desired ratio produce an output voltage from the circuit which is zero. As in the previous circuit, a toroid placed around the center conductor of the transmission line develops by transformer action a voltage proportional to the line current. A sample of the line voltage to ground is taken via the variable capacitor C5. Unlike the previous circuit, both the sampled voltages are rectified, producing voltages across load resistors R3 and R4, without paying any attention to the phase differences between the sample voltages. It is only the amplitude of the sample voltages which determines the amplitude of the voltage across R4 (for the line voltage) or across R3 (for the line current). Because of the diode polarities, the voltages across R3 and R4 are of opposite polarity. They will cancel when equal, producing zero output voltage, and since C5 is variable, this condition can be set to occur at any desired line voltage to current ratio.

The difficulty with this circuit is that the sample voltages generated by the toroid transformers and the C5 coupling are frequency dependent. The rest of the components in the circuit (R1/C1 and R2/C2) are there to help eliminate the frequency sensitive output of the coupling elements over a reasonable range.

C5 is set with C1 and C2 at mid-range to provide a zero voltage output from the circuit on 20m when the line is terminated in a resistive load equal to the line impedance. Going to 10m, C2 is adjusted to restore balance and on 80m, C1 is used to restore balance. The adjustment has to be gone through several times before a good balance is achieved over the entire HF range. It would be possible to calibrate the meter with different value load resistors for different swrs. For high value loads, the meter would move off center in the opposite direction for low value loads. But, such calibration would only hold true for nonreactive loads.

Construction

Many conventional swr bridge circuits have been described over the last year or two in amateur literature where a toroid transformer, rather than a parallel wire to the center conductor of a coaxial line, has been used as a pickup element. The construction of such units applies equally well to the circuits shown here. It is only the circuitry following the pickup element that is really different, plus the added R1 pickup in Fig. 1, or the C5 pickup in Fig. 2. Note that it doesn't matter on which side of the toroid these pickups are placed. However, as in any rf circuit, all leads must be kept as direct and as short as possible. The circuits shown are for the HF range, but by changing the toroid and values of the bypass capacitors, they can be extended, by experienced experimenters, into the VHF range.

Chapter II

How to Measure RF Impedance

THE ANTENNASCOPE – AN EFFECTIVE TOOL

W. R. Carruthers VE3CEA

There are two types of antennas, commercial and amateur. A commercial antenna is generally designed for one frequency, has many acres of ground around it, no obstructions and miles of heavy copper cable buried underground to provide an "effective" ground. These antennas work as designed – very well. The amateur antenna, on the other hand, is just that – an amateur design and construction.

This antenna is subject to all ills, roof tops, buildings, trees, TV masts, house electric wiring, telephone wires and what not. It's a wonder they work at all! But they can be made to work and thousands of amateurs make them work. They make them work by pruning or lengthening the feeder cable and by using an antenna coupler. These are always empirical steps, the "let's cut and try and see what happens" method. How much better it would be, and a time saver too, if we tested our antenna systems electrically and *knew* what was happening and then could take intelligent action to put the whole antenna system into resonance.

This fact is well known – an antenna can only accept power and radiate properly when it is operating at its resonant frequency. This is no problem for the commercial people who operate at one frequency. The amateur, however, wants to "roam the band" and may wish to operate over frequencies hundreds of thousands of cycles wide, even megacycles wide. How can he do this with a fixed antenna system? The answer is, he can't! But he can construct an antenna system for a certain frequency and take the penalty of reduced radiation when he moves far away from it. However this actually works very well, because each amateur has his own particular part of a band in which he likes to operate – and his friends tend to stay there too. On this particular spot, the amateur works diligently to "put out a good signal."

The question arises – how can we make sure our antenna system is radiating well at the particulaf frequency we wish to use? One answer is to use electrical test equipment to show us what is happening on the whole antenna system, which includes the antenna and the feed line.

One of the most useful devices for this purpose is the rf bridge, generally called the Antennascope. It is a simple device, inexpensive to construct and very effective in results. It is usually powered by a grid dip oscillator. Such bridges should be used at the junction of the feed line and the antenna and will show the resonant frequency of the antenna itself and the radiation resistance at the feed point.

Making such measurement up in the air is a difficult thing for the average amateur and impossible for those whose antennas are supported at the ends. If we are willing, however, to accept a small degradation in results, we can use the *rf* bridge at the station end if we have a half wave, or multiple of a half wave, feed cable. At every half wave point on a feeder cable the voltage and current vectors are in phase, which simply means that the electrical condition seen at the end of the cable is repeated every half wavelength in the cable. We can use the *rf* bridge then, at the station end of the feed line, if we are willing to agree that the results will not be 100% but reasonably close to it. The results will be affected by all the various factors that affect amateur antenna resonance and these effects may give us some peculiar results, but they can be overcome and the final results may be quite valuable to us.

Let me give you an example to illustrate what I'm talking about and to show you how effective the use of the *rf* bridge can be: –

A friend of mine constructed a 40 meter inverted V antenna, held at the feed point 40' up on his beam tower, 66' legs down to supports which held the ends about 8' off the ground. Feed line was 100' of Twin Amphenol cable, velocity factor .68. The antenna was difficult to feed, swr was high, radiation was poor. He asked me to have a look (electrical) at it. I took my grid dip meter, *rf* bridge and vtvm.

The first thing done was to check the feed line length. 1/2 wave length at 7.1 MHz was 492 x .68/7.1 or 47.1 feet. Two 1/2 wave lengths (to get into the station) would be 94.2 feet.

The first conclusion was that the feed line was 5.8 feet too long.

Next Test No. 1 was made using the *rf* bridge with results as shown in **Fig. 1**, the results being shown in table form and also plotted in graphical form.

It was obvious from this graph that the antenna system was resonating outside the

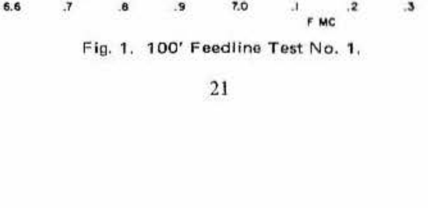

F MC	R OHMS
7.0	55
7.1	80
7.2	110
7.3	115

Fig. 1. 100' Feedline Test No. 1.

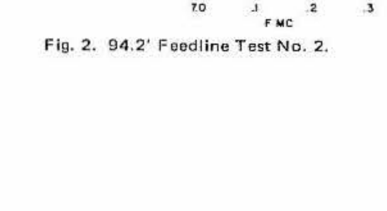

F MC	R OHMS
7.0	48
7.05	51
7.1	69
7.2	80
7.3	115

Fig. 2. 94.2' Feedline Test No. 2.

band as shown by the dotted lines. This test was repeated and the results were taken down to 6.4 MHz. They showed the system to be resonant at 6.6 MHz.

Test No. 2 was made next using the feed line cut to 94.2 feet. **Fig. 2** shows the results.

It was obvious the resonant point of the system was rising.

Test No. 3 was made next, cutting the feed line to 91.2 feet long. **Fig.** 3 shows the results.

F MC	R OHMS
7.0	43
7.1	46
7.2	76
7.3	81

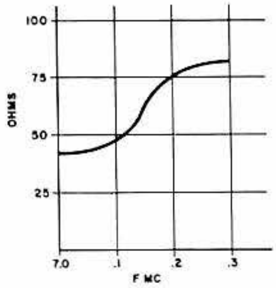

Fig. 3. 91.2′ Feedline Test No. 3.

The resonant point was rising, but not far enough yet.

Test No. 4 was made using the feed line cut to 88.2 feet long. **Fig. 4** shows the results. It was obvious that we were very close to the resonant frequency of 7.1 MHz which my friend wished to use.

F MC	R OHMS
7.0	45
7.05	47
7.1	51
7.2	60
7.3	75

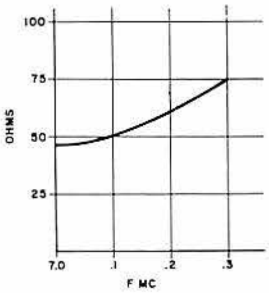

Fig. 4. 88.2′ Feedline Test No. 4.

Test No. 5 was with 85.2 feet in the feed line. **Fig.** 5 shows the results.

F MC	R OHMS
7.0	78
7.1	61
7.2	60
7.3	66

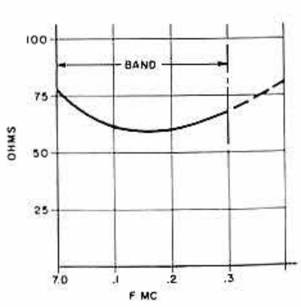

Fig. 5. 85.2′ Feedline Test No. 5.

Test No. 6 was with the transmitter (300 watts CW) and antenna coupler connected. There was no trouble in loading and no trouble in balancing the coupler to obtain an swr of 1 to 1 ratio.

The results on the air were interesting, 5/9+ reports to the Eastern half of the U. S. A., 5/8 reports to Germany etc. *Conclusion:* The results shown above are not precise, nor can they be expected to be precise. There are too many unknown factors entering the electrical picture, such as those which required a shortening of the feed line, in this example, to somewhat less than a half wave length. But the bridge showed us the overall picture and suggested what was required to be done. The on-the-air results show that it was giving us a good picture and a result that was very satisfactory for my friend's needs.

Why not construct an *rf* bridge and check you own antenna system? I suggest it will pay off and be very informative to you, showing you what your antenna system looks like electrically and what to do to bring your whole system to the resonant frequency you wish to obtain. . . . VE3CEA

HOW TO BUILD AN ANTENNASCOPE

Paul Franson WA1CCH

A type of simple bridge used for measuring antenna impedance is called the Antennascope, shown in Fig. 1. This bridge is designed for low power operation – a grid dip meter usually gives plenty of power. It should be built very compactly with short leads. The potentiometer should be of high quality; an Allen-Bradley Type J is fine. The bridge can be calibrated with regular composition resistors. Simply connect the resistors in turn to the antenna terminal and adjust the pot until the meter reading dips to zero. Then mark the value of the resistor by the pot pointer. In use, the meter reading will not null completely except for resistive loads, so it will not read zero for reactive antennas. Nevertheless, the minimum reading will occur at the approximate impedance reading. Remember that all antenna bridges should be used between the antenna and the transmission line.

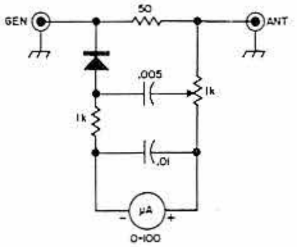

Fig. 1. The Antennascope is a simple antenna impedance bridge. It should be constructed compactly for best high frequency use.

THE MARK III RF IMPEDANCE BRIDGE

Mark Cholewski W6CRT

NEED to measure the input impedance of that new beam? Or maybe to find out just what is the Q of the coils in your final? Or even to determine how much signal is being soaked up by your coax? If you ever want to do these, or any similar jobs, then the Mark III RF Impedance Bridge is the thing for you.

You can build it for a total cost of about $30 (exclusive of sheet metal) and, if you follow instructions closely, it will be accurate to closer than 10 percent throughout its operating range. Unlike the more common resistive-bridge and reflectometer methods of measuring impedance, the Mark III operates equally well at resonance or far away. It will measure both resistance and reactance present in resistors, capacitors, inductors, antennas, and transmission lines at any frequency between 2 and 30 mc.

Before we start into the actual construction of the Mark III, one thing must be emphasized. Accuracy can be assured only if the components, circuit, and parts layout are absolutely duplicated. The original instrument's calibration was obtained through tedious laboratory techniques. If you make any changes, the calibration curves will no longer apply. However, if instructions are followed to the letter you need have no worries about accuracy. A test model, built by W6BJU following these instructions, checked out to 2 percent accuracy at 2 mc and 10 percent at 30 mc.

Construction of the Mark III divides into three major sections: Preliminary metalwork, actual wiring, and calibration. Each will be described separately. Ready? Let's go!

Preliminary Metalwork

1. Cut, drill, and bend to shape from soft aluminum shields S1, S2, and S3 as shown in Figs. C1, C2, and C3.
2. Cut, drill, and tap plexiglas insulators I1, I2, and I3 from bulk rod stock as shown in Figs. C4 and C5. When tapping plexiglas, use water as lubricant.
3. Cut, drill, bend, and solder tubular shields S1A, S2A, and S3A as shown in Fig. C6. Copper or brass may be used; aluminum should be avoided because of solding difficulties.
4. Assemble shielded resistor assembly R2/S4 as shown in Fig. C7. The copper tubing must be drilled out to clear the body of R2. When soldering, hold the assembly in a vise to protect R2 from excessive heat.
5. Cut, drill, bend to shape, and solder box shields S1B, S2B, and S3B as shown in Fig. C8.
6. Drill S5 (a 3x4x6 LMB unpainted chassis box) as shown in Fig. C9.
7. Cut, drill, and bend to shape shield partition S5A as shown in Fig. C10.

Shielded Transformer

While classified under the "preliminary metalwork" section for reasons which will become obvious, construction of the shielded transformer is the most critical part of the entire project. Before proceeding, read and re-read steps 8 through 24 and be sure that you understand them fully. Take special care when soldering—three transformers were built for the original instrument before a non-shorted one was achieved.

8. Cut to length, drill, and tap transformer mounting insulator I4 from 5/16-inch plexiglas rod.

9. Cut and drill two bobbin-end washers as shown in Fig. C12 and solder them to a length of copper tube as also shown. Then cut halfway through the bobbin with a hacksaw.

10. Pull the shielding from an 18-inch length of RG58/U. Save 8 inches for step 19 and use the rest in step 11.

11. Solder one end of the 10-inch shielding into the U-shaped slot on the bobbin end, using an aluminum rod or small drill to keep the inside of the shielding open. It must pass

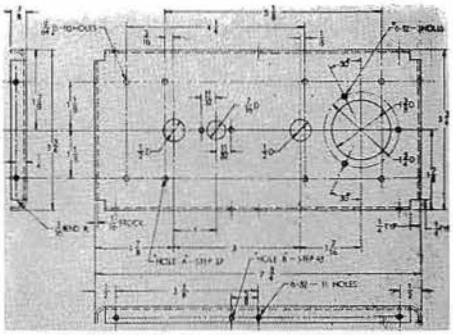

C1—Shield S1 (1 REQ), Mat'l Aluminum

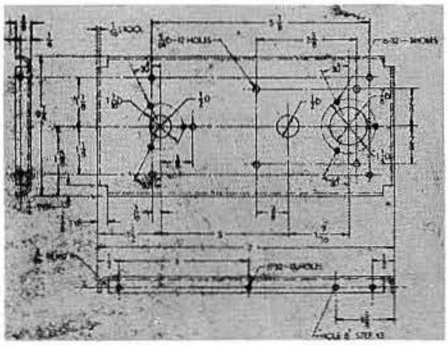

C2—Shield S2 (1 REQ), Mat'l Aluminum

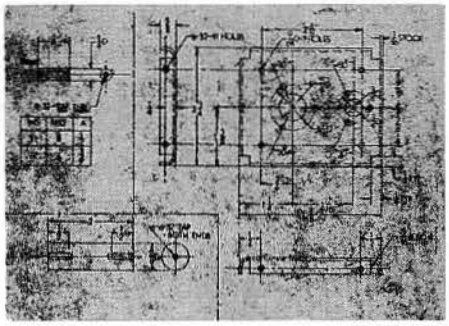

C4 — Insulators, Mat'l Plexglas rod

C5 — Insulator I3 (8 REQ), Mat'l Plexiglas rod

C3—Shield S3 (1 REQ), Mat'l Aluminum

No. 26 plastic-covered wire. Clean off all metal burrs and solder splatters.

12. Wind 48 turns of No. 26 plastic-covered hookup wire on the bobbin in three lays of 16 turns each. Solder the start of the winding at point A (see Fig. C12) and wind in the direction shown by the arrow.

13. The last turn will end at the shielding attached in step 11. Feed the free end of the wire into the shielding, draw the turns tight, and secure the winding with plastic tape.

14. Cut a piece of brass or copper shim stock as shown in Fig. C13 to a length which will wrap around the bobbin but will not allow the ends of the shim stock to touch each other.

15. Tin the shim stock along the edges.

16. Place the wound bobbin in a vise, wrap the shim stock around it, lining up the free ends of the shim with the slot cut in step 9, and solder the shim to the bobbin ends. Caution. *Do not overheat the winding; the plastic covering melts easily and a short is almost impossible to detect.*

17. Connect an SO-239 coax connector and UG-177/U hood to the shielded primary lead as shown in Fig. C15.

18. Wind one turn of ¼-inch diameter half-hard copper tubing around a 1-inch diameter form. Saw the tubing as shown by "phantom lines" in Fig. C14. Drill as shown in Fig. C14 and clean off all burrs.

19. Locate the 8-inch piece of shielding left over from step 10.

20. Using same technique as in step 11, solder one end of the shielding into the ⅛-inch hole. Clean off all solder splatter and burrs.

21. Solder one end of another length of No. 26 plastic-covered hookup wire to point A (see Fig. C14) and wind three turns inside the tubing in the direction shown.

22. Feed the free end of the hookup wire through the shielding. Bend the tubing into final shape as shown in Fig. C14. Pull up the three turns snugly, making sure that the plastic coating is undamaged.

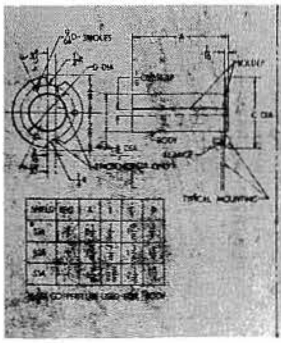

C6—Tube shield, Mat'l .010 Brass

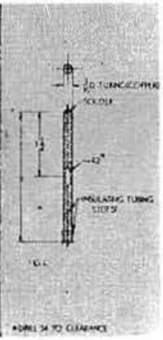

C7 — Shield Assembly S4

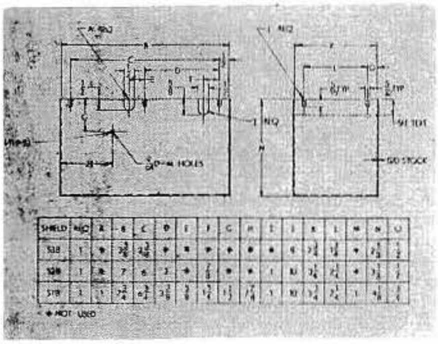

C8—Shield can, Mat'l Aluminum

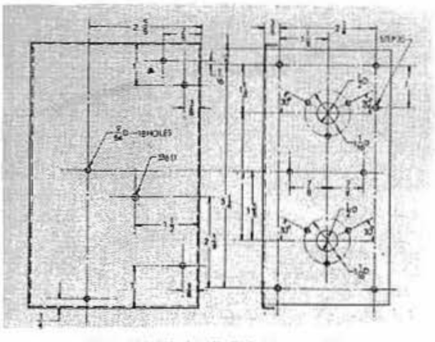

C9—3X4X6 LMB Shield can S5

23. Connect a male phono plug to the free end of the shielded wire as shown in Fig. C15. Length of the wire is critical.

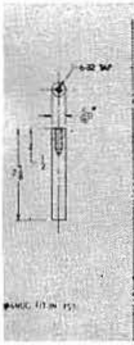

C11 — Insulator I4, Mat'l Plexglas

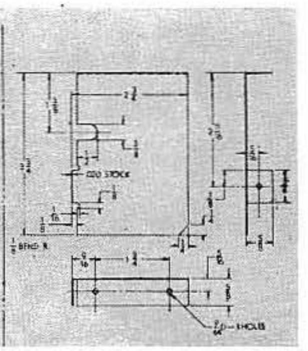

C10—Shield Partition S5A, Mat'l Aluminum

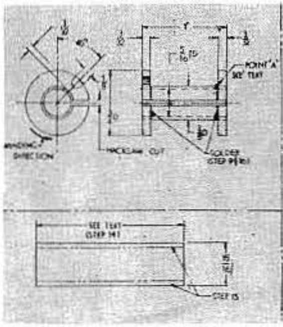

C13—Shield PS1A, Mat'l .005 Brass or Copper

C12 — Primary bobbin PS1, Mat'l Brass or Copper

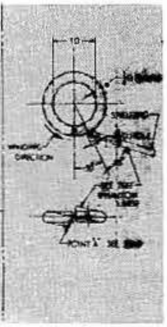

C14—Shield PS2, Mat'l copper

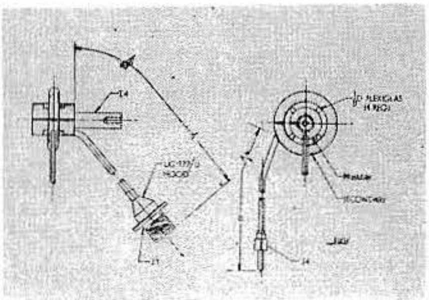

C15—Transformer assembly

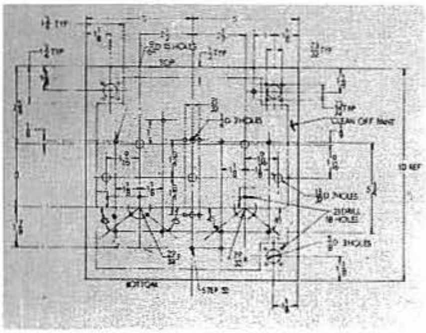

C16—Inside view of panel

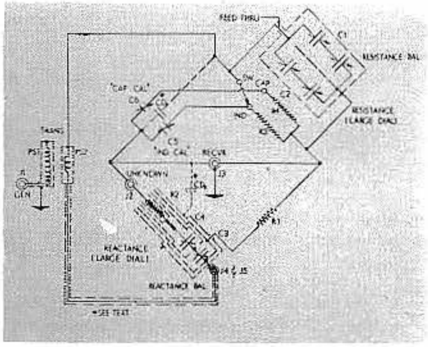

Fig. 1—HF. RF. Bridge Mark III Schematic

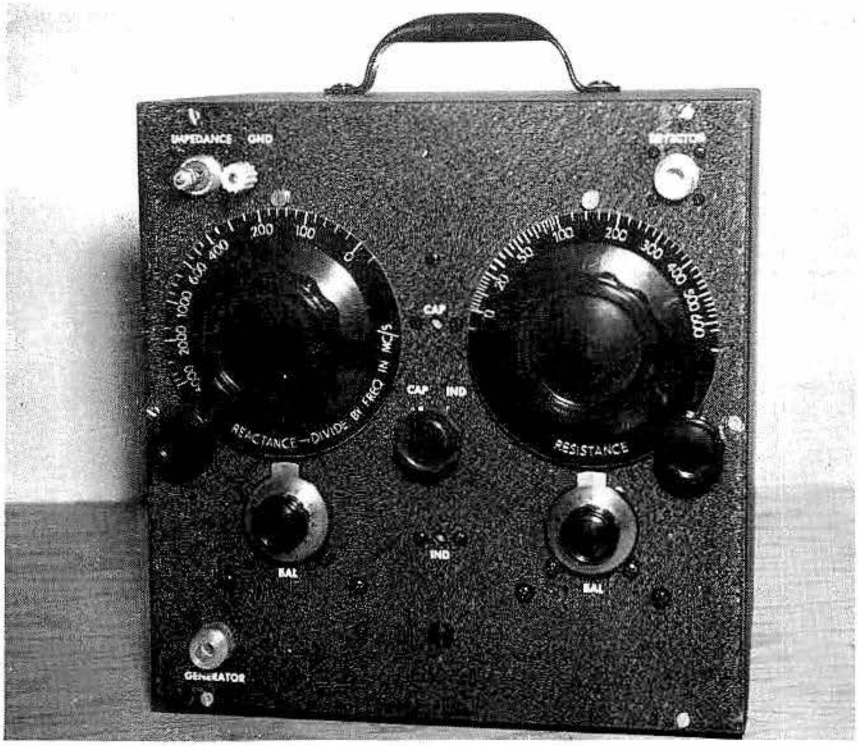

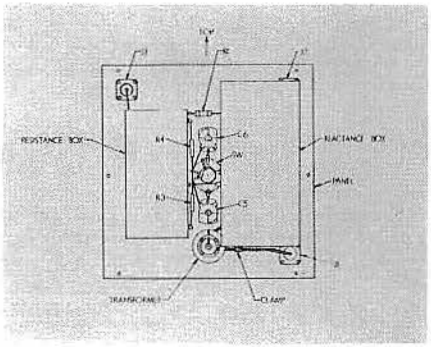

Fig. 2—Inside view

24. Using four pieces of ⅛-inch diameter plexiglas rod as spacers, assemble the primary and secondary shielded windings as shown in Fig. C15. Attach insulator I4 to the transformer by cementing it into the bobbin hole with Duco. Cement both windings to spacers with Duco and allow to dry overnight. This completes the transformer.

25. Remove top and bottom from the 8x10x10 utility box. Remove all paint from flanges; clean to bare metal to provide adequate rf shielding on reassembly.

26. Remove paint from inside of bottom plate for ½ inch in from each edge.

27. Remove paint from inside of top plate as shown in Fig. C16 by "phantom lines."

28. Drill the top plate as shown in Fig. C16.

Note that drawing shows INSIDE surface of plate.

29. Cut shafts of all four variable capacitors to ⅝-inch length. Remove all trimmer capacitors.

30. Tap the threeholes on the face of each capacitor, using a 6-32 tap. Take care not to damage the first stator plate; a bottom tap may be necessary. Attach three type I2 insulators to C1 and C2, using ¼-inch-long 6-32 set screws as shown in Fig. 3. This completes preliminary metalwork.

Actual Wiring

31. Attach C1 and C2 to shield box S5 using six ¼-inch-long 6-32 machine screws. Connect stator lugs of C1 to those of C2 with No. 18 tinned wire, as shown in Fig. 3.

32. Mount two soldering lugs as shown in Fig. 3 and connect remaining stator lugs to them, using No. 18 tinned wire.

33. Press the Erie CF-408 feed-thru into the 0.136-inch diameter hole in S5. Solder a short No. 18 tinned lead from the inside terminal of this insulator to the wire installed in step 31, as shown in Fig. 3.

34. Attach shield partition S5A, using three ¼-inch-long 6-32 machine screws. One of the screws installed in step 32 must be temporarily loosened and removed.

35. Attach four type I3 insulators to shield box S5 as shown in Figs. 3 and C9.

36. Attach three more soldering lugs to S5 as shown in Fig. 3, using ¼-inch-long 6-32 machine screws.

37. Attach four type I3 insulators to shield platform S1, using the 4¼-inch-spaced holes in S1 and 6-32 screws. Attach S1A to S1, using 6-32 screws from the inside of S1. Attach the female phono socket to S1 in the 7/16-inch diameter hole. Attach four type I1 insulators, using ¼-inch-long 6-32 screws. Do not tighten the screw in the hole marked "Hole A" in Fig. C1; this screw will hold a cable clamp later. See Fig. 4 for details of insulator placement.

38. Attach four type I1 insulators to shield platform S2, using the 2⅜-inch-spaced holes in S2 and ¼-inch-long 6-32 screws. Attach S2A to S2. Attach capacitor C3 with 1-¼-inch-long 6-32 screws, using two nuts on each screw as shown in Fig. 4. Mount a soldering lug under one nut as shown. Align the capacitor by adjustment of the mounting screws and nuts.

39. Solder a No. 18 tinned wire to the female phono socket and pass the wire through the corresponding hole in S2. Attach S2 to S1 with ¼-inch-long 6-32 screws going into the type I1 insulators attached to S1 in step 37.

40. Attach C4 to shield platform S3 using ⅝-inch-long 6-32 screws with dual nuts (same as in step 38). Attach S3A to S3.

41. Connect two of the stator lugs of C4 with No. 18 tinned wire as shown in Fig. 4. Slide shield assembly S4 into S3A. Center assembly S4 in S3A, using a piece of ¼-inch-long insulating tubing. Make certain that opposite ends of S4 and S3A are even as shown in Fig. 4, and cement tubing in place with Duco. Solder the shorted end of resistor R2 (which is in S4) to the wire connecting stator lugs of C4.

42. Attach shield platform S3 to S2, using ¼-inch-long 6-32 screws going into the type I1 insulators installed on S2 in step 38.

43. Place a ¼-inch-long 6-32 screw in the flange of S3 as shown in Fig. 4, with a soldering lug. Connect this lug to the stator lug of C3 with No. 18 tinned wire. Using ¼-inch-long 6-32 screws, mount a soldering lug in Hole B (see Fig. C1) of S1 and another in S2 as shown in Fig. C2. These lugs are mounted in a direction opposite to that of the platform flanges.

44. Attach the "unknown" ground lug to the panel next to the "unknown" coax connector hole, as shown in Fig. C16. Attach the "receiver" coax connector to the panel, from the inside. Attach the "IND-CAP" switch to the center of the panel, using an extra nut to position the switch as far as possible from the panel. Orient the switch as shown in Fig. 5. Place a soldering lug under each nut. Mount C5 and C6, using ⅜-inch spacers between the capacitor frames and the panel. Connect the

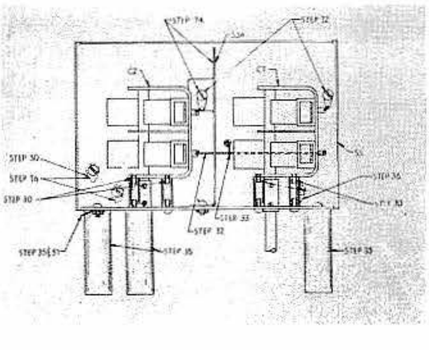

Fig. 3—Resistance box—Inboard view

rotors of C5 and C6 to the lugs of the switch with No. 18 tinned wire as shown in Fig. 5.

45. Solder a 2½-inch length of No. 18 tinned wire to the "COMMON" terminal of the switch. Connect the stators of C5 and C6 to the remaining switch terminals, as well as resistors R3 and R4. Complete connections are shown in Fig. 5.

46. Attach the "unknown" coax connector to the panel, placing a soldering lug under one mounting screw. Connect the "unknown" ground lug to this soldering lug to provide a good bond.

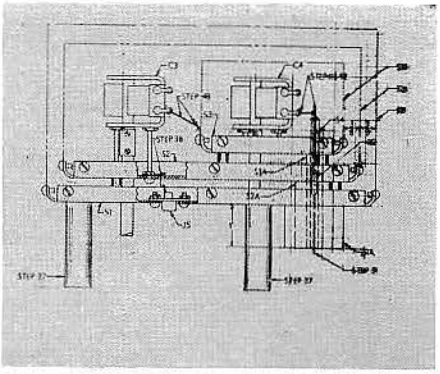

Fig. 4—Reactance Assembly—Inboard view

47. Mount panel bearings for capacitors C2 and C4. Mount the special panel bearings furnished with the Johnson Vernier Dial assemblies in place. Attach the two large dials to dummy shafts and mount the dial indicator in the position you prefer. Remove the large dials and dummy shafts after placing the dial indicators.

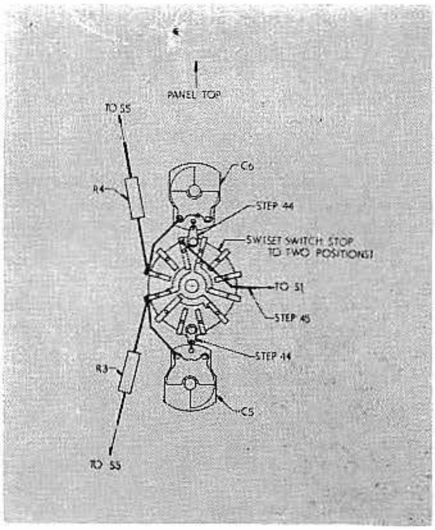

Fig. 5—Switch Assembly

48. Mount S5 at the left side of the panel (as shown in Fig. C16) and mount the assembly of S1, S2, and S3 at the right side. See Fig. 2.

49. Attach shaft couplers to the four capacitors. Cut the plexiglas shafts to length and mount them in place. Attach the large vernier dials. Set the dial of C2 so that it reads "0" at minimum capacity. Set the dial at C4 so that it reads "100" at maximum capacity. Mount the two Calrad dials on the panel, setting them so that they both read "0" at maximum capacity. Attach the knob to the "IND-CAP" switch.

50. Connect resistor R1 (270-ohm deposited-carbon) from the soldering lug on S2 to the lug in line on S5. Connect R3 (220 ohms) and R4 (100 ohms) to the soldering lugs on S5 which were installed in step 36. R3 will be the resistor nearest the bottom of the panel.

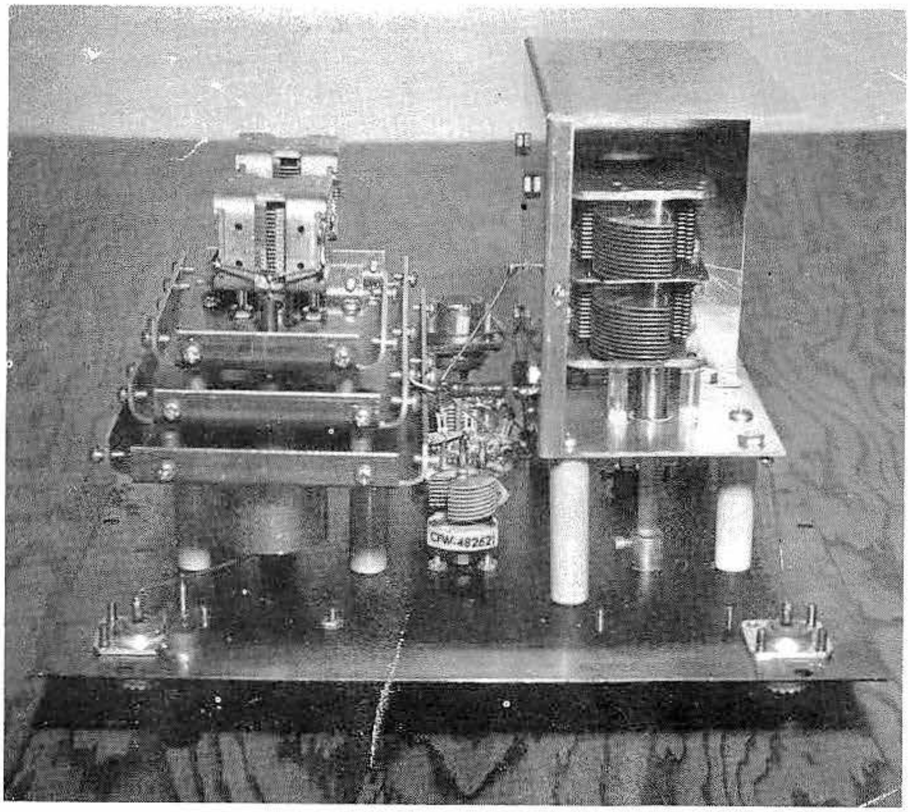

51. Connect a No. 18 tinned-wire lead from the "detector" coax connector to the lug on S5. Connect a No. 12 (note different wire size) lead from the "unknown" connector to the free end of R2. *Make sure that R2 is not shorted to any shield.*

52. Attach the transformer, completed in step 24, to the bottom of the panel as shown in Fig. 2. Attach the coax connector connected to the transformer to the panel in the "signal generator" hole. In the hole between the transformer and the connector, mount a cable clamp to hold the shielded primary lead. Secure the shielded secondary lead with another clamp held by the loose screw installed in step 37.

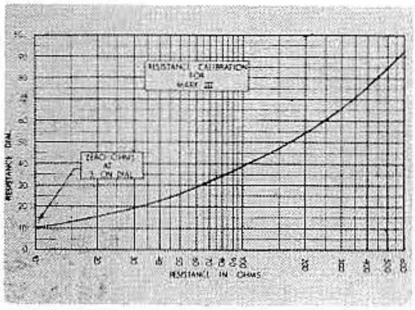

Fig. 6

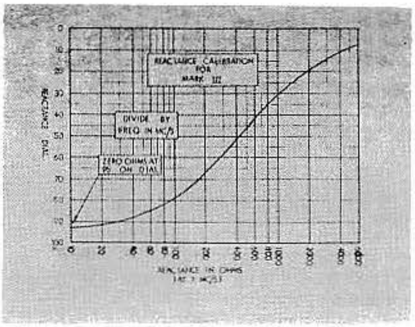

Fig. 7

53. Mount a soldering lug to S1B with a ¼-inch-long 6-32 screw, placing the lug in the direction of the box opening.

54. Place ¼-inch-long 6-32 screws in all remaining tapped holes in S1, S2, and S3. Starting with S3, place all shield boxes in place and secure screws. Connect a No. 18 tinned lead from the soldering lug installed in step 53 to the Erie feed-thru on S5. Set C5 and C6 to mid-capacity.

55. Attach four rubber feet to the bottom plate of the case and four more to the bottom side of the utility cabinet. Restore the front plate in place and secure with the sheet-metal screws provided. This completes construction of the bridge. After calibration, it will be ready for use.

Calibration and Use

Since the Mark III is a null-type instrument (adapted from the Schering bridge circuit) it can only be used with a signal generator and a detector. Both must be shielded; however, a Heath SG-8 will do nicely as the signal generator and any decent communications receiver will serve as the detector. For best results, it should be calibrated with the signal generator and receiver with which it will be used.

56. Connect the bridge to the signal generator and the receiver, using coax cable from the panel connectors to each.

57. Set both the signal generator and the receiver to 2 mc.

58. Short the "unknown" terminal and ground terminal using a banana plug in the connector or a coaxial short made by soldering the pin of a PL-259 connector to the shell through a copper disc.

59. Set the "IND-CAP" switch to "CAP."

60. Set the "Reactance" dial to 15.

61. Set the "Resistance" dial to 5.

62. Tune the signal generator, only, for maximum signal in the receiver.

63. Using the two small "balance" dials, reduce the signal in the receiver to the lowest level possible. This is the "null" or "balance" condition. The dials will interact, and multiple adjustment will be necessary. However, all receiver and signal generator adjustments must be left alone during this step.

64. Replace the short across the "unknown" connector with the 620-ohm deposited-carbon "test" resistor.

65. Using only the large "resistance" and "reactance" dials, null out the signal once more. Record the final reading of the "resistance" dial. If this reading is 95, you are extremely lucky and the first half of the calibration is complete. If not, proceed with step 66.

66. If the "resistance" dial reading is less than 95, increase the capacity of C6 slightly and repeat steps 60 through 65. If the reading is larger than 95, decrease the setting of C6 and repeat steps 60 through 65. Continue this process until the reading comes out at 95.

67. When capacitive calibration is complete, set the "IND-CAP" switch to "IND," and the "reactance" dial to 30.

68. Repeat steps 60 through 65; if the final reading of the "resistance" dial is other than 95, proceed with step 69.

69. If the reading is less than 95, increase the capacity of C5 slightly and repeat steps 60 through 65. If the reading is larger than 95, decrease the setting of C5 and repeat. Continue until the reading is 95. This completes calibration, and the curves shown here can be used for readings.

Using the Mark III

To use the Mark III, set the instrument up as described in step 56 of the calibration procedure. Set the signal generator to the desired frequency and short out the "unknown" terminal of the bridge as described in step 58.

If you're measuring an inductive impedance, set the "IND-CAP" switch to "IND," the reactance dial to 95, and the resistance dial to 5. Null the signal with the small balance dials.

Now remove the short and connect the unknown. Rebalance the bridge using the large dials, note the readings, and convert the readings to ohms by use of the calibration curves.

These initial values must be corrected. The reactance value is corrected by dividing the value read from the curve by the frequency (in mc) at which the reading was taken. The resistance value is corrected by obtaining the correction factor due to frequency from Fig. 8 and multiplying the reading by this factor.

If you're measuring a capacitive impedance, set the switch to "CAP." Then estimate the reactance of the unknown and set the large reactance dial to a value just larger. Set the resistance dial to 5 and short out the "unknown" connector. Null the signal with the short and again with the unknown, the same as for an inductive impedance.

Correct the resistance reading in the same manner as for an inductive impedance. However, the reactance reading is corrected differently: Subtract the value taken from the calibration curve from the value originally set on the reactance dial. Now divide this remainder by the frequency (in mc) at which the reading was taken. The result is the true value of capacitive reactance.

A few minutes' practice will make operation of the Mark III far more simple than the detailed directions would indicate; in practice, you can make a reading in less time than it takes to read these paragraphs.

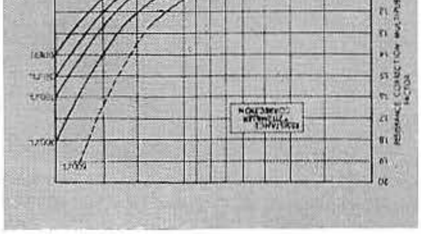

Fig. 8

Parts Substitutions and Design Changes

Since few hams are content to build a "Chinese copy" of someone else's design, a few words on the effect of changes are necessary.

Naturally, the Mark III doesn't represent the only possible—or even necessarily the best—way in which such a bridge can be built. Any part, or all, may be changed. However, any such change will invalidate the calibration curves, and is not recommended unless the builder has access to laboratory equipment.

Even then, before making any substitutions, these three design articles should be read and fully understood:

"A Radio-Frequency Bridge for Impedance Measurements From 4000 KC/S to 60 MC/S," D. B. Sinclair, Proceedings of the I.R.E., November, 1940, pages 497-502.

"A High Frequency Model of The Precision Condenser," D. B. Sinclair, General Radio Experimenter, October-November, 1938, pages 1-7.

"The Effect of Stray Capacitances to Ground in Substitution Measurements," M. Reed, Wireless Engineering, May, 1936, page 284 ff.

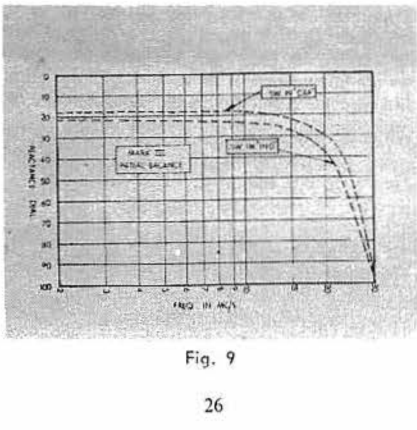

Fig. 9

Parts List

Quan.	Item	Description	Cat. No.
1 ea.	C1, C2	Miller, 10-365 mmf/sect.	2112
1 ea.	C3, C4	Miller, 10-365 mmf	2111
1 ea.	C5, C6	50 mmf air padder (good quality)	APC-50
1 ea.	J1, J2, J3	SO-239 Panel coax connector	
1	J4	Phono plug	
1	J5	Phono socket	
1	R1	Aerovox 1% 1/2W 270 ohm "Carbofilm"	Type-CP
1	R2	330 ohm 5% 1/2W Carbon Resistor	
1	R3	220 ohm 5% 1W Carbon Resistor	
1	R4	100 ohm 5% 1W Carbon Resistor	
1	Test	Aerovox 1% 1/2W 620 ohm "Carbofilm"	Type-CP
1	SW	Centralab 3p; 1 Sect; 2-3 Position (reset stop to two positions only)	2506
*2	Dials	E. F. Johnson Vernier (0-100 CW)	116-285-1
2	Brgs	E. F. Johnson Panel Bearings	115-255
4	Cplrs	Metal ¼ shaft couplers	
3	Rod	¼ inch diameter plastic rod x 12"	
1	Rod	⅛ inch diameter plastic rod x 6"	
2	Rod	⅜ inch diameter plastic rod x 12"	
2	Dials	"Calrad" 1-13/32 inch Vernier Dial	VD-36
1	BP	Binding Post (ground lug)	
1	Case	Bud carrying case 10"x10"x8"	CC-1100
1	Box	LMB box (plain) 3" x 4" x 6"	
1	Hood	UG 177/U coaxial cable hood	
1	Insul.	Erie feed-thru insulator	CF-408
1	Short	PL-259 coaxial male connector	
2	Clamps	Walsco clamp for ⅛ to 3/16 cable	7502-N

Of the above list the critical parts are:
C1, C2, C3, C4, J2, R1, Test resistor, Case, LMB box, UG 177/U, Erie feed-thru.

---

* This author hand engraved these dials per Figs. 6 and 7.

MEASURE RF IMPEDANCE WITH YOUR SWR BRIDGE

Warren McDowell W2AOO

So far, the great majority of antenna impedance bridges that we have found in construction articles are devices that function only with a low power rf source. The run-of-the-mill impedance bridge is designed to operate with a grid dipper as the source of rf excitation. Operation with tube type dippers is generally intended as the transistor dippers produce an rf level that is too low for excitation of this bridge type.

The conventional antenna bridge cannot be left in the transmission line continually as excessive rf energy would soon destroy the device. This means that each time measurement of antenna impedance is desired, the transmission line must be opened and the bridge inserted and grid dip excitation applied. Grid dippers are not necessarily the most accurate rf source for a specific frequency in an amateur band . . . therefore the station receiver must monitor the dipper output for any bridge accuracy. A low power bridge will not often present the true operating impedance of the antenna . . . especially antennas with parasitic elements. A bridge that operates under full transmitter power will present a much more accurate picture of your antenna system at a specific frequency.

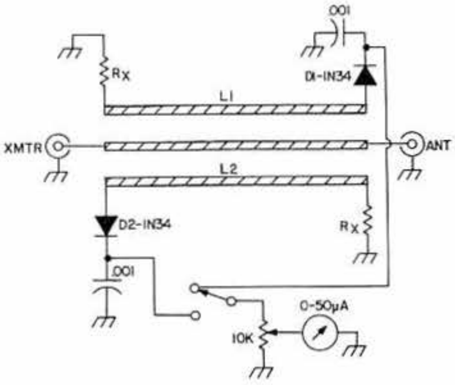

Fig. 1. Conventional swr bridge.

Inspiration for the "In Line" full power bridge came from information concerning the standard swr bridge. Just about every amateur has in his possession some sort of swr bridge and the great majority are of the type illustrated in Fig. 1. This bridge consists of a section of transmission line near which are placed two inductors. These inductors are actually two bridges along with their associated diodes and resistors. One of the bridges reads forward power and the other reflected power. The resistors (Rx) at the end of the inductors L1 and L2 are critical for accurate bridge null (balance) and therefore must be the proper value for the specific transmission line used. For the average swr bridge the value for Rx is 100Ω for 75Ω line and 150Ω for 50Ω transmission line. Considering that resistor Rx is critical for the impedance of the line in use, varying the value of Rx and devising a system of calibration for Rx would enable us to determine the impedance of a line when a null is achieved on the bridge meter.

The "reflected" inductor which is L1 in Fig. 1 is the portion of the bridge circuit we are interested in for impedance measurements. The value of Rx and the transmission line must balance the bridge for a null to be realized. Any variation from the above parameters will mean changing the value of Rx so that the bridge again balances at a new impedance value.

By experimenting with various values of resistance at Rx, it was determined that a 1000Ω potentiometer represents a fair value. The 1000Ω potentiometer is inserted in place of Rx on inductor L1 (see Fig. 2). This is the inductor with the diode pickup located toward the load or antenna end of the swr bridge.

Make sure that all leads to the 1000Ω potentiometer are short and that the metal case (shell) of the potentiometer is well grounded. Excessive lead length or inductance will create inaccuracy of the device. The position of the potentiometer will be determined by the physical layout of your particular swr bridge. It must be set at a point where the shaft can be extended through the front panel of your swr bridge. Allowance must also be made for a dial or other indicating device which can be calibrated in ohms (impedance) on the front panel. It might even be desirable to mount your entire present bridge in another larger case so that all functions can be accommodated.

Calibration of this in-line bridge was the major problem. An ordinary grid dip meter will not provide sufficient excitation for readings. With full power applied, especially a kilowatt, it becomes difficult to find resistive dummy loads of various values to calibrate the bridge. Even with 100W of rf, proper resistive load values are not common.

The solution to the calibration problem came to us in the form of an (ouch!) CB transmitter. A CB transmitter is fortunate if it is able to put out 3W of rf and at the same time is well within the frequency range of an swr bridge. The most important fact is that a CB transmitter will provide adequate excitation for calibration of the bridge with ordinary 5W 5% carbon (garden variety) resistors. For calibration, a good assortment of these resistors is necessary. Use values such as 5, 27, 47, 75, 100, 150, 220 and 470Ω. Intermediate values can be then interpolated on your scale. The calibration procedure is simple . . . first borrow your neighbor's CB, then attach the 5W resistors across the antenna coax connector of the bridge and excitation of the CB transmitter is applied to the remaining connector on the bridge. The bridge sensitivity should be set for a middle scale reading of the meter and the 1000Ω potentiometer is varied until you reach a null on the meter. Mark the value of the calibration resistors on the potentiometer scale (dial). Do this for all of the available resistors and your bridge will be in fair calibration.

At this point we should mention that this system does not measure reactive components in the antenna system. If your antenna is reactive, either inductive or capacitive, the meter will present a shallow, poorly defined null at the operating frequency. A sharp, well defined null will indicate a purely resistive impedance.

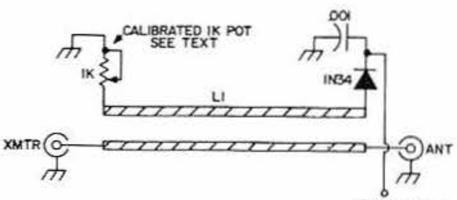

Fig. 2. The modified bridge leads to 1K pot should be as short as possible and shell (case) of pot grounded.

When using the bridge in its former function as an swr bridge, set the resistance dial to the value of your transmission line. When measuring impedance, vary the dial for maximum dip on the meter and read the resistance (impedance) directly.

As a final point, it is wise to insert the swr/impedance bridge at a half-wave or an even multiple of a half-wavelength from your antenna. At half-wave points from the antenna, the antenna impedance is repeated. This will enable your measurements to be much more accurate. When determing half-wavelength points, take into consideration the velocity factor of your particular coax.

USE YOUR GDO AND Z METER

Denys Fredrickson WØBMW

The GDO is one of the most versatile pieces of test equipment available. Yet there are many hams who don't know how or when to use one. The writer will try to describe and explain some of its various functions.

The GDO is basically a variable high frequency oscillator with a frequency range of approximately 550 kc to 250 Mc. It may also be used as a diode detector or wave meter. The GDO gets its name from the fact that a meter measures the grid current and when the oscillator circuit is coupled to a resonant circuit a reduction in grid current is obtained. This is called the grid dip. However, when it is used as a wavemeter and coupled to an rf source, an increase in current is obtained at resonance.

The GDO and impedance meter can be used to accomplish the following:

1. Determine the resonant frequency of tuned circuits, including antennas.
2. Determine the impedance of circuits, receiver inputs and antennas.
3. Determine the length of half-wavelength or quarter-wavelength transmission or tuning stub lines.
4. Determine the "Q" of a circuit or component with the aid of a VTVM.
5. Determine the resonant frequency of individual coils, capacitors or crystals that are within the range of the GDO.
6. Determine the rf frequency of energized circuits.
7. Monitor a radiated rf signal with the aid of headphones.
8. Neutralize rf stages.
9. Locate parasitic oscillations.
10. To align receivers and television sets.
11. Determine where BCI and TVI is entering the radio or television receivers.
12. Determine unknown inductance.
13. Determine unknown capacitance.

Now if you will step into the lab we will try to demonstrate how these instruments can be put through their paces. Let's begin with the simple functions and then gradually creep up to those which are more complex so they don't scare us before we get started.

An oscillator-detector

Simply plug in a pair of headphones (if GDO has facilities for them) and "zero-beat" with the radiating signal. This then will be the frequency of the radiating signal.

Crystal frequencies

Connect a one turn loop of wire across the crystal and couple the GDO close enough to get a dip of the meter when resonance is obtained. It is always wise to check lower frequencies to be sure it is the fundamental frequency that is being indicated.

Frequency determination

Generally the GDO has a switch which is used to remove the plate voltage from the tube. The tube will then serve the function of a diode and the meter as a diode load. When a peak deflection of the meter is obtained this will indicate the frequency of the radiating signal.

Resonance of an RF choke

When an rf choke is used as a parallel or shunt fed circuit, it must be free of self resonance over the operating frequency range of that circuit or it may burn up. The popular pi tank circuit is an example. Place a short circuit across the choke and then determine its self resonant frequency by coupling the GDO close enough to indicate a dip on the meter when the resonant frequency is obtained.

Neutralization

Apply plate power to the exciter stages and filament power only to the stage being neutralized. Use GDO as a wavemeter and couple close to the tank coil in the stage being neutralized. Vary the frequency of the GDO until maximum reading is obtained and then adjust the neutralization for minimum GDO meter reading. The circuit being neutralized may have to be retuned and the above procedure repeated with a closer coupling of the GDO to the Tank coil.

IF alignment

Tune the GDO to the desired frequency and couple it close to the if coil to be aligned. Adjust the if coil until a dip is observed on the meter. The if coil will then be tuned to the desired frequency.

Inductance and capacitance checking

To determine the value of an unknown capacitor, connect it across a known inductance and use the GDO to find the resonant frequency of the circuit. With these known values a reactance chart will give the value of the capacitor. Some GDO's supply a chart which corresponds to the coils supplied with the GDO as the known inductances. To determine the value of an unknown inductance, connect a known capacitance across the coil and use the GDO to find the resonant frequency. Again, the reactance chart may be used or the following formula (which may be used for either inductance or capacitance) for resonant circuits:

$$L = \frac{1}{39.48\ (f^2)\ C} \quad \text{or} \quad C = \frac{1}{39.48\ (f^2)\ L}$$

Where f = cycles per second
L = inductance in henries
C = capacitance in farads

The inductance of an air core coil can be estimated by the following formula:

$$L = \frac{(rN)^2}{9r \neq 10w}$$

Where L = inductance in microhenries
N = number of turns
r = radius of coil in inches
w = length of coil in inches

Q measurements

Connect a condenser across the coil so the tank circuit resonates at the desired frequency. Connect a VTVM across the tuned circuit and tune the GDO until maximum reading is obtained on the VTVM. The GDO coupling may be changed until a convenient value is obtained on the VTVM and then it must not be moved during the remainder of the test. Note the resonant frequency f_c then detune the GDO to a lower frequency until the VTVM reads 70.7 percent of its original or peak value and call this frequency f_1. Now detune the GDO to a higher frequency until the VTVM again reads 70.7 percent of its original or peak value and call this frequency f_2. The Q is then calculated by using the following formula:

$$Q = \frac{f_c}{f_3}$$

Where f_c = is the center of resonant frequency
f_3 = the difference between f_1 and f_2

Parasitic oscillations

By using a pair of headphones with the GDO, the parasitic oscillation frequency may be determined. Turn the power off of the stage being checked and then use GDO to find the circuit which resonates at the parasitic frequency by moving the GDO slowly around the wiring. When a "dip" is observed, moisten the finger and touch an ungrounded point of the circuit. If a change in the dip is observed, it indicates that it is the portion of the circuit that would be a likely suspect.

BCI and TVI locator

Most of the BCI and TVI problems can only be resolved at the receiver, either by installation of filters, resistors or condensers or a combination of all three. The problem is—where is the rf entering the receiver? Use the GDO tuned to the frequency which produces the greatest amount of interference. Probe around with the GDO until the most sensitive spot is located, which is indicated by watching or listening to the receiver interference. After the point of entry is determined then the appropriate corrective action can be accomplished.

Antenna measurements

Space does not permit to discuss all types of antennas and adjustments so only a few will be mentioned to give some idea on the use of the instruments. At this point it should be mentioned that inductive type coupling should be used between the GDO and antenna when checking near the current maximum point and capacitive coupling when checking near the voltage-maximum point.

The beam antenna has gained tremendous popularity in recent years plus many headaches for those striving to obtain the maximum effectiveness. Most of the headaches can virtually be eliminated by using the GDO and Impedance meter (Z-meter). Let's take a look at a 3 element yagi and see what has to be done to obtain a good adjustment. The element lengths must be physically adjusted or electrically loaded to obtain resonance at the desired frequencies and the feed point impedance must match the impedance of the transmission line. These two points are not the only considerations for beam adjustment but they are the most important factors. The GDO can be inductively coupled to each element and the elements adjusted until each one is resonant at the desired frequencies. It is best to make the measurements while the antenna is in operating position. This is very difficult to do in many cases but let's assume you can. After the elements have been adjusted, the feed point must be adjusted to match the line. The Z-meter and GDO will be used to accomplish this adjustment. The Z-meter is basically a resistance type bridge with a calibrated potentiometer as one of the bridge arms. Connect the Z-meter directly to the antenna feed point. Couple the GDO to the Z-meter inductively thru a couple loops of wire connected to the other terminals of the Z-meter. Tune the GDO to the resonant frequency of the beam and adjust the Z-meter to the dip or null. If the impedance indicated by the Z-meter is not the same as the transmission line then readjust the matching network and redip the Z-meter until the impedances are equal.

Now—if you can't adjust the antenna in the operational position you still can determine the resonant frequency and the impedance by standing on the good old Terra-Firma. The procedure is a little more involved but effective. First—we must have a means of electrically connecting the instruments to the antenna. This is best accomplished by a transmission line a half-wave or a multiple of a half-wave in length. Determine the height of the antenna above ground and calculate how many half-wave lengths of line will be required by using the following formula for a half-wave length of line:

$$L = \frac{(492)\ (K)}{f}$$

Where L = feet
f = megacycles
K = propagation constant
(RG/8 is .66)

This is an approximate length so be sure and cut it extra long because now we will find the exact physical length. Why an exact physical length? A halfwave length of transmission line will reflect the resistance placed across the output at the input end of the line, i.e. if a 50 ohm non-reactive resistor is placed across one end of a half wave or multiple length thereof, the GDO and Z-meter will indicate 50 ohms at the other end of the line. Cut the line somewhat longer than calculated above, short one end and connect the Z-meter to the other end of the line. NOTE: Keep twin lead off the ground and away from metal objects. Set the Z-meter to zero impedance and couple the GDO inductively to the Z-meter. Adjust GDO frequency until the fundamental frequency causes the Z-meter to dip or indicate a null. The frequency indicated should be lower than the desired frequency. Simply cut a few inches of cable off, short the end again and readjust the GDO. Repeat this procedure until the desired frequency (which should be the same as the resonant frequency of the antenna) is obtained. You will then have an electrically halfwave length of line or a multiple thereof.

Coax or twin-lead may be used for the halfwave length line when checking the impedance of the antenna. Connect the line to the antenna, hoist the antenna up to its operating position and adjust both the Z-meter and GDO for the null indication. If the antenna is not resonant at the desired frequency, the driven element should be readjusted a measured amount and then note the frequency change. This will give you an idea how much the resonant frequency changes with a corresponding element change. Now adjust the matching network to the desired impedance. This will be accomplished when the Z-meter dips at the desired impedance with the GDO set at the resonant frequency of the Antenna.

What would you do if your 100 foot coax cable developed a short someplace along the line? Replacing the whole line would be too expensive. Simply connect the Z-meter to one end of the line, adjust the Z-meter for zero impedance and then adjust the GDO for lowest frequency which will produce a null on the Z-meter. Use this frequency in the formula given for a halfwave length line and carefully calculate the length which will be the distance from the input end to the short.

A quarter wave length tuning stub can also be determined by using the procedures just outlined for the halfwave length line except a quarter wave line reflects a short at the input when the output end is electrically open. Now that we have mentioned the quarter wave length line, some may be wondering just what useful purpose does it serve. The quarter wave tuning stub (as it is sometimes called) may be used for antenna matching, TVI elimination or matching two units which have different impedances. The quarter wave matching stub can be used as a matching device on antennas which is explained in most antenna handbooks. It may also be used to eliminate an interfering frequency from entering the TV. This is accomplished by connecting a quarter wave stub to the TV antenna terminals which is a quarter wave in length at the interfering frequency.

Another use for the quarter wavelength matching stub is to permit maximum signal transfer between the source and a load which have different impedances. If the signal source impedance was 100 ohms and the load impedance was 52 ohms, a 72 ohm quarter wavelength of line would give a good impedance match. Hold it just a minute, how in the world did we come up with that 72 ohm business? Simple—another formula will give us this information.

$$Z_o = \sqrt{Z_s Z_a}$$

where Z_o = Impedance of quarterwave matching stub
Z_s = Impedance of the source
Z_a = Impedance of the load

Very little has been said concerning the various methods of GDO coupling. Actually—only two types of coupling are used; inductive and capacitive. Capacitive type coupling may be used on shielded coax cable, the ends of antenna elements and generally where the voltage maximum exists. To obtain the greatest accuracy, the GDO should be loosely coupled. Parallel coupling to inductors can be used to obtain maximum coupling.

Chapter III

Measuring RF Power Output

MEASURING RF OUTPUT

Paul Schuett WA6CPP

So your new Bandjammer 5000-Q is rated at 752W peak power. Big deal. How much of this is getting out where it counts?

It's easy to find out by inserting an rf ammeter in series with the line. A more exact reading would be to have the rf ammeter at the antenna input terminals, but that might be impractical when it comes to reading the meter (although you could put a diode there and a remote-reading meter in the shack).

Recently I found an rf ammeter at one of the mail-order surplus houses for $2.95. I installed this in a little cabinet, put two coax connectors on the back (in and out) and now can read rf current in the line, into the dummy antenna, or wherever it is going.

Remember the formula $P=I^2R$? Square the reading on the meter, multiply by the impedance of the line you're using, and you have the power past that point. At the antenna, you could determine the antenna resistance and take the current reading at that point.

For those who have difficulty with mental computations and can't find a pencil, the following chart fives the computed power levels present in matched 50 and 70 Ohm coaxial lines for various levels of rf current:

RF Current Amperes	Power Output in Watts 50 Ohm Line	70 Ohm Line
0.5	12.5	17.5
1.0	50	70
2.0	200	280
3.0	450	630
4.0	800	1120
5.0	1250	1750

When tuning up the rig, place the ammeter on the antenna side of any tuning or matching devices and tune for maximum current.

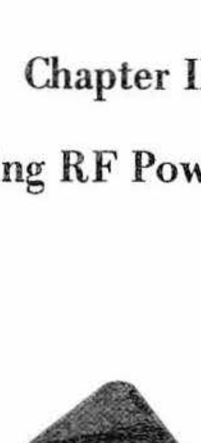

Inside wiring is extremely simple. Note ground wire installed to insure continuity of the shield circuit.

The completed instrument showing the coax connectors on the rear.

It's amazing to learn how much (or how little) current these rigs produce. For instance, my Swan 250-C puts 1.35A into a 50Ω line. My friend's SB101 from Heathkit put out 0.7A until we worked on the antenna – then it put out about 1.25A. My Heathkit SB401 puts anywhere from 2 to 0.6A into a 50Ω Cantenna, depending on what band it's on.

Commercial stations determine their power by the antenna current. Knowing the antenna resistance at the operating frequency, they multiply that by the current squared. If the antenna resistance is 62½Ω, 4A rf would be 1 kW into the antenna. 126.48A would be 1 MW.

Remember the ammeter does insert a little reactance in the line (I never leave it in all the time), and the calibration changes with frequency, although you can tell what side of the ballpark you are on. An rf ammeter in the shack makes a nice piece of test equipment.

LIGHT BULBS AS RF INDICATORS

John Houser WB2GQY

The major appeal to the amateur – as well as some commercial applications – of light bulbs as rf power indicators is low cost. To this must be added the universal availability of bulbs and screwbase sockets for pennies.

As low cost is of primary interest to well over 50% of those interested in any project, and as I have always had an insatiable desire to find out the why's and wherefore's of standard light bulbs as rf power indicators, I decided it might be the opportune time to do a resarch project and determine once and for all just which bulbs might be suitable and which might not be, and also to determine whether light bulbs would make *good* rf power indicators, or poor, and to find out what precautions might have to be taken if one decided he was going to take this low-cost path of determining his transmitter output power rather than go for a more expensive power output meter.

Also, power output meters in the higher wattage ranges become quite expensive compared to the $2 to $5 which might be expended in a light bulb indicator. In general, porcelain screw bases are available for from 12¢ to 25¢ each, and bulbs from 15¢ to 65¢ each, and not more than four of each are necessary for up to 3 KW power indication.

Table 1 lists most of the common types of electric light bulbs readily available. One look at this table immediately reveals why such light bulbs might not be such good rf power indicators as some folks may have thought they were in the past. It also reveals that some very special precautions have to be taken in using them, or the user may find he has overloaded his transmitter and burned up a few components which might be expensive to replace.

Table 1

Variation in Resistance, Cold to Hot State, Common Variety of Electric Light Bulbs.

Bulb Rating Watts At 115V	Cold Filament Resistance	Hot Filament Resistance	Ratio Cold to Hot Filament Res. (Approx.)
7.5	166	1750	1 to 10
25	40	529	1 to 13
40	27	331	1 to 12
60	20	219	1 to 11
100	9	132	1 to 15
150	6	83	1 to 14
200	4.5	65	1 to 14
250	3.5	53	1 to 15
500	2	26	1 to 13
750	1+	17.7	1 to 15

Table 2

Possible Configurations for Various Power Outputs at Various Impedance Terminations

A. – Nominal 70 to 73Ω Impedance Loads:		
175W Load:	3–60W bulbs in parallel	(73Ω)
	OR	
	7–25W bulbs in parallel	(70Ω)
3,000W Load:		
	4–750W bulbs in series	(71Ω)
B.) Nominal 50 to 55Ω Impedance Loads:		
250W Load:	4–60W bulbs in parallel	(54.9Ω)
500W Load:	2–150W bulbs in series, both paralleled by 1–150W bulb	(55Ω)
1,000W Load:	2–500W bulbs in series	(53.6Ω)
2,250W load:	3–750W bulbs in series	(53.1Ω)

The extremely high ratio of cold to hot filament resistance in all types of these bulbs immediately struck me as being the most undesirable factor in using them.

It is very easy to see, for instance, that if one wished to use a 250W bulb for indication on a 250W transmitter, and he computed the resistance at 250W to be 53Ω, (which it is, but only when *hot*), he would assume he had just about a perfectly matched indicator to plug in in place of his 52.5Ω feed line.

However, from this table, it is apparent that this 53Ω resistance is attained *only* at full brilliance and wattage, and the actual cold resistance is only 3.5Ω. In other words, if the bulb were connected to the antenna terminals of the transmitter, and the transmitter keyed full power, the transmitter would be looking into *not* 53Ω, but 3.5Ω, which is a lot of difference, and an extremely low value for any pi network to match.

For a few seconds, until the filament attained full brilliance, the transmitter would be subjected to a terrific overload, due to this impedance mismatch.

Therefore the first precaution which might be emphasized in using light bulbs would be *not* to key the transmitter at full power with a cold bulb, but to gradually bring the power from some lower value to full power as the bulb attains full brightness (and hot, matching resistance).

Not until I got into this project did I realize the very high ratio of resistance of these filaments from the cold to hot state; I don't suppose very many people do. It also brings to mind how the house electric meter must jump every time a bulb is snapped on in the house. This is not an ad for those light dimmers being sold at all the electrical stores, but it sure brings to mind that power bills could be cut appreciably through their use, i.e., bringing the bulb gradually to full brilliance instead of just snapping on a switch.

Getting back to the bulbs, Table 2 gives in various configurations series, and/or parallel combinations which would be most likely to give the amateur a load for a particular transmitter power output, in nominal impedances near 52 and 72Ω. If the configuration mentions 200W, this does not mean that it would be suitable for indicating the output of a 100W output transmitter, because at half brightness, the resistance offered by the bulb is not identical to that at full brightness.

While a differance of an ohm or two would not be serious, nor would a difference of as much as five, or even ten watts, at high power levels, at low power levels less than 100W, for instance, such differences would be seen to become increasingly serious from the matched impedance standpoint. The configurations given match quite a variety of standard line impedances and a wide range of power outputs. Matches can be obtained for RG-8, 11, 17, 13, 58 and 59 type cable.[1]

One may not realize without measurement that the lead length of the filament support wires alone inside the 25–150W bulbs is very close to 18 cm. Even though they are coiled on a 2 to 1 ratio, the filament is inductive in every sense of the word. At higher frequencies, the filament support wires would appear inductive, and to these factors must be added the parallel capacity of the screwbase shell and the central base contact wafer. Even though such capacity is small, it would become significant at most amateur frequencies above the 30 MHz range. Though the 22 cm total wire path would perhaps indicate a bulb could be used up to 300 MHz, such is not at all the case.

It is easy to see that the sometimes suggested trick of using a capacitor in series with a light bulb as a load should be approached with caution, for it would be very easy indeed to run into a series resonant circuit which might result in damage to the transmitter to which such circuit were connected.

In the course of my preparation of this article, I discussed the ramifications with a number of interested hams. Some of them suggested I extend the research to include the use of the smaller types of indicator (pilot) bulbs as loads for testing out transmitters with power outputs in the 1W to 20W range, not only just for amateur applications, but also with a view to using them as loads in testing FM transmitters.

When one considers that there are well over 100 types of these small bulbs, rated from .001W to 2W, and if all of these were to be considered individually, it could take a vast amount of time – and eventually one would end up with perhaps only five or so of these bulbs that would be at all suitable, so such research was not included in this article. However it did open up a field in which there may be a demand for information and may be the subject of a subsequent article.

Frequency Ranges

The use of standard screw-base ceramic or steatite porcelain light bulb sockets is entirely feasible for all of the configurations shown and will handle all amateur bands, 160 through 10. Naturally the leads from socket to socket should be as short as possible in either the series and/or parallel configurations. I found these leads can be kept to approximately 2 cm for such interconnections. Likewise, the coax termination lead should be kept to 2 cm or less.

If extra precautions as to lead lengths are observed, and the bases of the bulbs removed to enable connections directly to the stem wires, it would appear reasonable to suspect that these bulbs might be used for 6, 5, and perhaps 2 meter bands, but it is also quite evident the 2 meter band would be the practical limit.

One should be able to conjecture that light bulbs as power rf indicators are not quite the equal of well-designed power output meters which maintain their rated impedances over a very wide power output range – bulbs do not – but then, they *are* cheap in comparison.

Visual comparison of brightness is completely satisfactory for comparison purposes.

For instance, a 500W bulb connected to the 115V mains should show the same brilliance as one of the 500W bulbs as used in the 1 KW load.

Actually a transmitter supposedly putting out 2,000W PEP is putting out something less than 1,000W with average voice modulation; it would be more of the order of 500–750W average power. Remember that the light bulb is only going to show average power output, not peak, and as ham transmitters are limited to 1,000W dc input to the final amplifier, one cannot expect much more than 500–750W output (average) unless the efficiency of the final amplifier stage approaches 85% which is very unusual, although I am hearing lately that certain high-power transistors are in development which will deliver such high efficiency figures; a bit above that which heretofore has been obtainable with tubes. You should be hearing a lot more about these super-efficiency transistors in the near future; and I expect them to be appearing in certain ham transmitters within a year or so.

Naturally a CW transmitter with the final operated Class C may deliver as much as 850W with 1,000W dc input, while a DSB transmitter on phone could not be expected to deliver more than 650W with Class A or B modulation.

The research and conclusions I reached on this project brought to mind the old subject of using light bulbs in series with primaries of transformers to reduce the secondary output voltages, which is a trick which has been used for years by hams and others. The information contained herein indicates they are not only quite suitable for such usage, but in fact make quite ideal voltage regulators of a sort.

In fact, the question immediately arises as to why bulbs would not make rather ideal voltage regulators for high voltage supplies if used as a variable-resistance dc regulator in the dc leg. This again opens up a field which might bear intense investigation.

[1] Solid Dielectric RF Transmission Lines, W8LUQ, *Radio News* Oct. 1946.
Line Matching: Table of Power and Voltage Loss in DB, *Radio News* Feb. 1947.

BUILD YOURSELF A LIGHT WATTMETER

Carl Henry

Every week brings something new in the hectic field (or pasture) of electronics. Attempting to solve old problems with new components is an interesting pastime for electronics enthusiasts, but they must be careful not to put their foot in the wrong thing. One of the new components is the cadmium sulphide/selenide photocell. A semiconductor sensitive to light is not an entirely new concept, since selenium cells have been around for some time, but the degree of sensitivity makes the cadmium cell stand out.

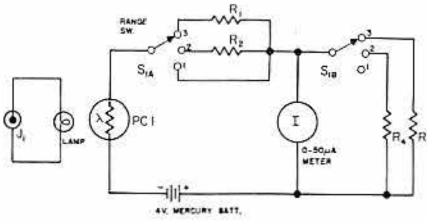

Fig. 1. Circuit of Light Wattmeter.

In measurements especially, there are many possible applications for cadmium photocells. One measurement in particular is usually difficult for the amateur, and this is power measurement. There is a way of using cadmium photocells to measure power, and I call this circuit a "light wattmeter". Operation is just as the name implies, that is, the power is used to generate light which is measured by the photocell.

You have probably realized by now that we are going to use a light bulb as a load. Now this is frequently done in amateur circles, but no one will go out on a limb as to its accuracy. Except me. We know that the ordinary lamp filament has a positive temperature coefficient. Fig. 5 illustrates the variation of a typical lamp filament resistance with input power. By keeping this in mind, fairly good accuracy can be had. Of course, if you use a different lamp, the curve will still apply, but the resistance will be different.

You can run your own graph, however, by using the setup shown in Fig. 2. Either dc or ac can be used, and since the lamp has little inductive or capacitive effect, the readings taken will be good to better than 100 mc. After you have a graph on the lamp you are using, you can effect any kind of impedance match you wish.

Using a lamp as a load simplifies the problem of power measurements because loads for this service are hard to come by. Power resistors are too inductive, and when they approach 1000 watts, they become downright expensive. However, even a 1000 watt lamp is not too expensive.

Fig. 1 shows the circuit that I used as the basis of this article. A standard 150 watt lamp was used here as a load. The photocell is mounted about five inches from the bulb. A wooden box houses the wattmeter, completely sealed internally against extraneous light. Figs. 3 and 4 illustrate the variation of resistance of the photocell with variation of power applied to the lamp. Even very small amounts of power are measureable, if the cell

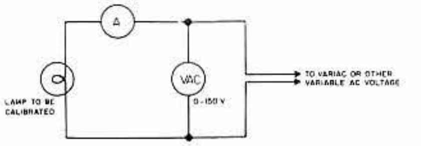

Fig. 2. Method of calibrating lamp and determining resistance.

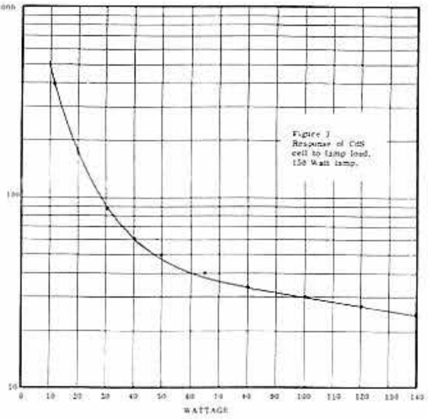

Fig. 3. Response of CdS cell to 150 watt lamp.

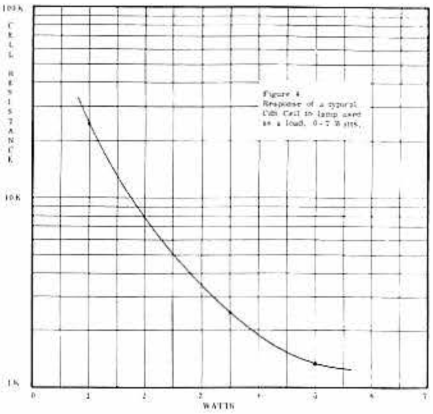

Fig. 4. Response of cell to 7 watt lamp.

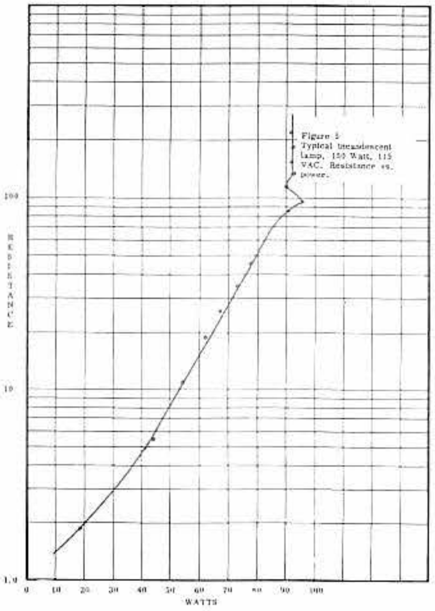

Fig. 5. Resistance versus power input, 150 watt lamp.

is shielded from external light.

The wires from the input connector to the lamp are kept as short as possible by removing the lamp base and soldering the lamp wires directly to the coaxial connector. Switch S1 selects three ranges, which can be set by the builder to anything he desires. In my case I used three ranges which cover from 0.5 watt full scale to 250 watts full scale. With some photocells a 1.35 volt mercury battery can be used instead of the 4 volt battery shown. Also, I used a 50 microamp meter, because of convenience (mine), but even a 10 ma meter will work. Don't exceed the rated power dissipation of the photocell, and remember this may derate with increasing ambient temperature. One thousand watts in a box can be a lot of ambient temperature.

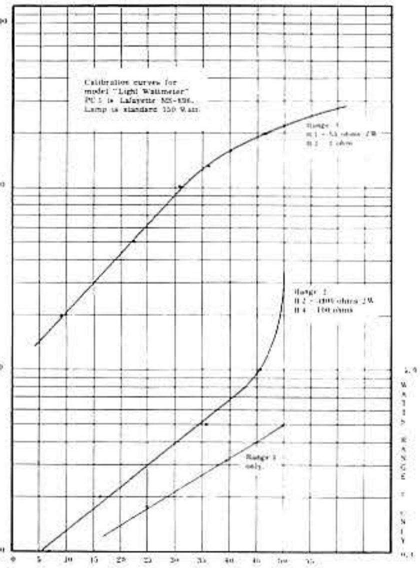

Fig. 6. Calibration curve for Light Wattmeter.

Getting back to the accuracy, if I were to tell you what the meter would read with such and such a lamp, and a given power input, your calibration would be off ten to twenty per cent when the construction was finished. The best way of getting a good calibration is something you've heard before; "if you want a job done right around here you've got to do it yourself".

Now obviously you can't calibrate the wattmeter at 14 mc, because this is what you are trying to measure to start with. Luckily the light from the lamp filament is primarily a power function, and it doesn't matter whether this is dc power, or 100 mc power.

So, to calibrate the wattmeter, apply ac or dc as you wish. Measure the input voltage and current, and read the meter. Make a graph of input power vs. meter reading, and there you have it. Fig. 6 is the graph that I use with the circuit in Fig. 1. Keep the impedance variation in mind, and select a lamp with the power impedance so that standing waves won't eat up lots of your power. As a dummy load and approximate power indictor this is not too critical, of course. But if you wish accurate power measurements, your impedance should be approximately matched.

LOW POWER LIGHT WATTMETER

Bill Hoisington K1CLL

This article describes a very useful gadget for determining the rf power output of solid-state VHF–UHF transmitters in the difficult range to measure, from about 10 mW up to 5 watts. It does not read watts directly; but by a simple comparison of calibrated pilot light brilliance, it will tell you how many watts you are putting out, to within less than 5%. It allows you to check power increases and estimate your efficiency quite close.

Principle Involved

We'll start right in with this part because, while this unit is not by any means a "trick," it does not read rf directly. You first light a pilot light as a good dummy load, matching it into the rf tank circuit of your transmitter by the normal means, also noted here.

You then switch on a second bulb of the same type by means of a battery, controlling the light output with a $1.30 wirewound potentiometer in series, as shown in Fig. 1. This pot must be previously calibrated in milliwatts, as by the method of "volts times milliamperes equals milliwatts." You then match the brilliance of the bulb lit up with rf or its dull glow at some 18 to 25 milliwatts if you're just getting your transmitter going, and read the watts on the wattmeter dial. It's astonishing how well it works, how repeatable it is, and how you wouldn't be without it once you build and calibrate it.

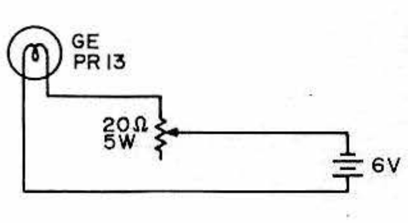

Fig. 1. The circuit supplies the brilliance "standard" for comparison. When the "standard" lamp is mounted adjacent to the dummy load, the pot permits variation of the standard to match the load. If the resistance is panel-marked in watts, a good power indication is achieved.

Brilliance Standard

Figure 2 tells almost the whole story at a glance. You can, of course, put as much calibration on the dial as you have time for. It is quite important to orient the bulb filaments in the same relation to your eyes for best matching. There isn't much in back of the panel except one 6V battery which can be obtained in any hardware store.

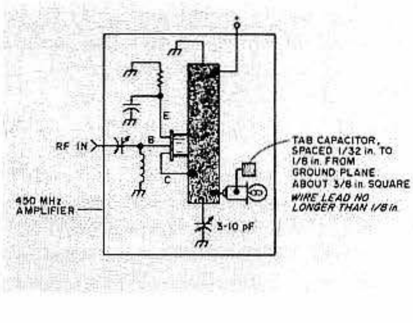
RF IN
E
B
C
TAB CAPACITOR, SPACED 1/32 in. TO 1/8 in. FROM GROUND PLANE. ABOUT 3/8 in. SQUARE
WIRE LEAD NO LONGER THAN 1/8 in.
450 MHz AMPLIFIER
3-10 pF

Fig. 3. Matched pilot light load for the UHF version.

RF Matching

Not that it is particularly critical, but be sure and note the need for a large range of series capacitors for the rf pilot lights as you go up in frequency. This can be seen clearly in Figs. 2 and 3. The block diagram, Fig. 2, shows a 6 meter setup. As you go up in frequency the series capacitance drops. A good matched load on 432 MHz can be obtained as shown in Fig. 3. I sometimes remove the tin base from the bulbs, but this is not an absolute neces-

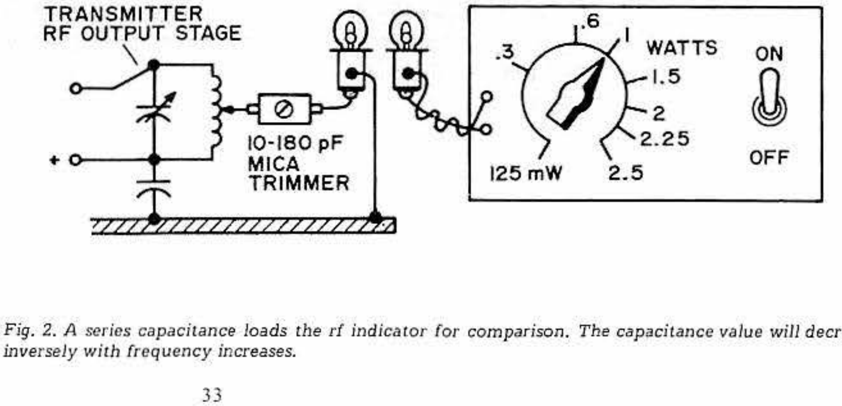

Fig. 2. A series capacitance loads the rf indicator for comparison. The capacitance value will decrease inversely with frequency increases.

sity. It is important to vary the amount of coupling, and thus the series capacity, by spacing the tab capacitor closer or further away from the ground plane, as detailed in Fig. 3. You also can use as many bulbs as you can solder onto a tuned rf inductor, even though they don't all light up with the same brilliance. You can match them all up, but you don't have to. Just check the wattage, or milliwattage, of each one and add them up for the total.

The number 48 or 49 bulb, listed at 2V and 60 mA, is rated at 120 mW, and glows dim at about 12 to 15 mW; so it can be used for low-power receiver oscillators, etc. With two other bulbs found in hardware stores, connected and matched to the rf inductor, such as the PR13 (5V at 500 mA), you can read correctly up to 5W. From there on up you're on your own, although a good variable 115V dc supply can be made up to work around 50 to 100W. I generally use a variety of 115V bulbs of different wattage, light them up with rf, and use their rated wattage.

YET ANOTHER LIGHT WATTMETER

John Meisner K5CXN

Many VHF operators would like a cheap, accurate instrument to measure rf power. The same desire is frequently expressed by operators of the HF bands. Here is a very simple wattmeter which when used with a 50Ω transmission line or load has all of the following desirable characteristics:

1. Easily calibrated to good accuracy (±5%) with your multimeter and a variable dc source;
2. Perfectly flat from dc to 450 MHz;
3. Insertion VSWR less than 1.05
4. Power readings in the 2 to 50 watt range (higher power can be measured with slight design changes);

The operating principle of this wattmeter is stark simplicity. A pilot lamp across the rf line senses a small portion of power in the line and glows brighter with increasing power. An appropriately located photovoltaic cell connected to a microammeter measures the light output which is proportional to the power flowing in the line. Of course a good deal of nonlinearity is involved in the various elements – both the lamp's resistance and its spectrum output change with heating; output of the photovoltaic cell varies considerably with both the amount and frequency of the light shining upon it. Some of these factors tend to cancel out however, because the photovoltaic cell produces some current with only infra-red lamp output at low power levels before the lamp even produces a visible glow, and the cell tends to saturate, increasing its output quite slowly at more intense illumination levels.

Probably the biggest single requirement of any wattmeter is that it must be capable of being inserted into a transmission line without disturbing the operating conditions in the line (low insertion VSWR). It might be argued that hanging a lightbulb across a transmission line will seriously affect the line impedance. Ordinarily, this is true, but the undesirable changes can be minimized and indeed approach an insignificant level if the resistance of the lamp filament is very large when compared to the line impedance. In general, a factor of 15 or more times the line impedance is sufficiently large to produce negligible effects. In the case of the suggested 10V, .014 Ampere pilot lamp, the mismatch produced in the line gives a VSWR of 1.05:1 at the power level of 1 watt. This mismatch decreases rapidly with increasing power. It falls to well below 1.01:1 at 50 watts. Another point of interest with regard to this particular choice of lamp is that it is a long-life type with a life expectancy of 10,000 hours. This implies two advantages: (a) The lamp will operate at well over its rated voltage without burnout (50.0V at 50 watts), and (b) the interior of the lamp envelope will resist darkening which would negate the wattmeter calibration. In addition, the filament structure of this type of lamp is a single-strand straight tungsten wire. The coiled type of filament structure introduces undesirable inductance into the circuit which can distort wattmeter readings in the UHF range.

The wattmeter shown is constructed in two boxes for the sake of convenience. The sensor can be located in the transmission line at any point and the meter can be placed beside the transmitter. For test work, the whole unit could easily be put into a single box. Coaxial fittings and cable were used for interconnection, but since dc only flows in this circuit, any type of wiring would be satisfactory.

The photovoltaic cell used in this wattmeter is a unit obtained from the local Allied/Radio Shack store. (Catalog No. 276-115.) Output of the cell is rated .5V at .6 mA in sunlight.

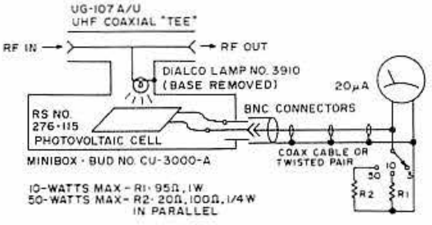

Fig. 1. Schematic diagram.

Since the cell generates considerably more than 20 mA under moderate illumination, a switching arrangement is incorporated into the wattmeter to shunt it into progressively higher current ranges with increasing power. Alternatively, a 50 or 100 milliammeter could be used with less switching at a sacrifice of sensitivity in the 1–3 watt range.

Calibration of the wattmeter is a simple process. By solution of the formula $P = V^2/R$ for voltage, the following tabulation is made for a 50Ω line impedance:

VOLTAGE (rms or dc)	POWER (Watts into 50Ω)
7.07	1
10.00	2
12.24	3
14.14	4
15.71	5
17.32	6
18.71	7
20.00	8
21.21	9
22.36	10
31.62	20
38.71	30
44.71	40
50.00	50

Using the tabulation, fasten a metered variable dc supply into the wattmeter according to the following diagram. Now simply note the reading on your wattmeter for each of the selected voltages in the table and tabulate this reading with the corresponding power in watts in a table of your own. It may even be possible to remove the front of the meter case and mark new calibrations directly on the dial. This was not possible with some hermetically sealed meters. It is best to disconnect the wattmeter from the antenna feedline for this calibration. If the meter is left with a transmission line attached and the antenna happens to be fed through a balun device with near zero resistance, the power supply, and perhaps the balun, will suffer.

It must be pointed out again that this power meter is intended for use either with 50Ω coaxial line systems with low VSWR or with 50Ω dummy loads. A coaxial line that is not "flat" (unity VSWR) or a dummy antenna that does not look like" 50Ω, which is the case with

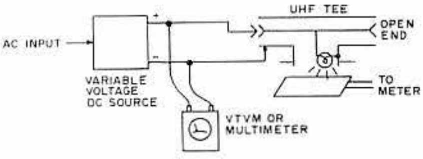

Fig. 2. Calibration diagram.

most HF loads used at VHF or UHF, may cause distorted power readings. If you have doubt about your transmitter's power output when using this meter, you should check it while using a known 50Ω load rated at the transmitter's output frequency.

I mentioned before that this power meter was easily adaptable to higher power readings. To upgrade the meter, the only change required is to insert the correct higher voltage pilot lamp having low current, long-life specifications. Suggestions for some of these are the following Dialco Lamps:

Part No.	Max Power Level (Watts)
24CS	100
48CS	200
60PSB	340
120PSB	800

A LOW COST RF WATTMETER

Mark Leavey WA3AJR

You say you just finished building that 2 meter rig and want to find out how much power you're running but can't afford to buy a wattmeter? You've been calling "CQ 80 QRP" all day with your quarter-watt wonder and nobody is answering? Then get yourself up, go down into the workshop, and build yourself a neato-keeno handy-dandy wattmeter.

If we are going to build a wattmeter, let's consider what we want. Accuracy and ease of calibration, as well as simplicity in construction are prime requisites. The meter described here is as accurate as components allow, and it's easy to build. The calibration is logarithmic, which means that a simple graph is possible, and easier than changing the meter scale.

For the mathematicians, I will present the formulas upon which this device is based, and ways of modifying it; for those of you who avoid math whenever you can, look at the graphs and skip these few paragraphs.

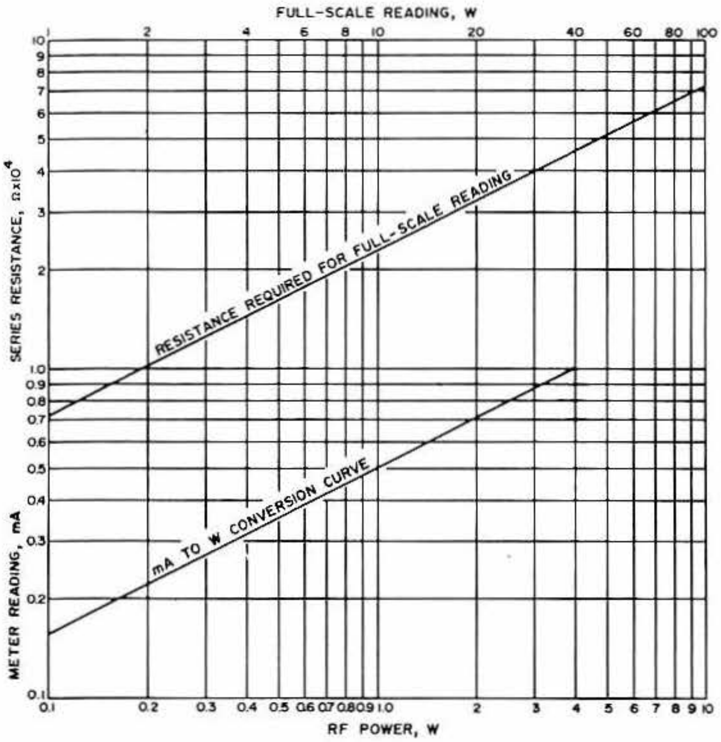

Fig. 1. Logarithmic plots for determining power. The upper curve gives resistance values for determining what the full-scale meter deflection will be (remember to multiply the series resistance value shown on the chart by 12 kΩ). The lower curve will allow you to determine your precise power out if you use a 0–1 mA meter.

To spare undue complexity, assume 51Ω line – other values can be dealt with later. Perhaps the easiest parameter to measure, and one that is proportional to power, is rf voltage. A voltmeter can be made most easily with a series resistor and a 0–1 mA meter. Now let's plunge into the actual calculations.

Assume W is the full-scale meter reading in watts, Z is the line impedance, E is the voltage measured, I is the full-scale meter reading in amps of the basic meter, and R is the value of the series resistor in ohms. We know that the voltage (IR) is equal to the square root of "impedance times power," or 14.270V. Now, since the voltage and the current (.001A because the full-scale movement is 1 mA) are known, simple division yields 14,270. The resistor value, then, is 14.270 kΩ. The upper portion of Fig. 1 is a graph that will enable the nonmathematician to choose the value of the resistor for full-scale readings up to 4 kW, with 51Ω line and a 0–1 mA meter.

Why 4 kW with an amateur power limit of 1 kW? A look at the bottom half of Fig. 1 will explain. Although this is the calibration of the prototype, for 4W full scale, it will double for 40, 400, or 4000W. A half-scale reading, 0.5 mA, corresponds to 1W (1 kW, etc.). This spreads out the range below 1 kW for ease of reading and measuring.

Now get out that soldering copper and gas pliers, and build it. As the schematic (Fig. 2) shows, the circuit is a basic rectifying type rf voltmeter. The prototype was built in a small can of the plug-in-module variety that was scrounged from the junkbox. About the only critical part is the series resistor. The capacitors in the prototype were mica, but ceramic disks would work as well. The diode can be a 1N34A, 1N270, 1N52, 1N38A, or just about anything else. Use the old ham's rule of thumb: "When in doubt, try it out!"

Two sockets might prove more convenient rather than one with a coaxial tee as shown. Conventional minibox construction or building into a new or existing rig will be more than adequate. Point-to-point wiring is used to permit compactness and reduce lead length.

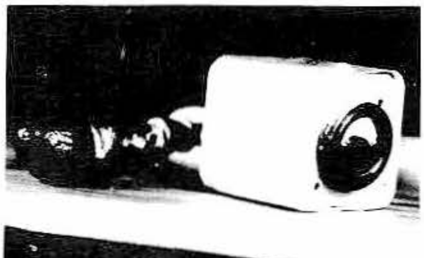

Front view of unit.

"Fine," you say, "but I don't have a huge mound of test equipment. How do I calibrate it?" That is the beauty of it – you don't! If the series resistor is accurate, the meter will be self-calibrating to a log scale. Remember, you know R and Z, and the full-scale W. Now assume a half-scale

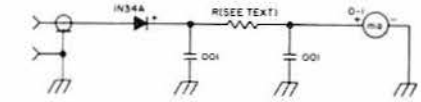

Fig. 2. Schematic diagram of the simple, accurate, and easy-to-build rf wattmeter.

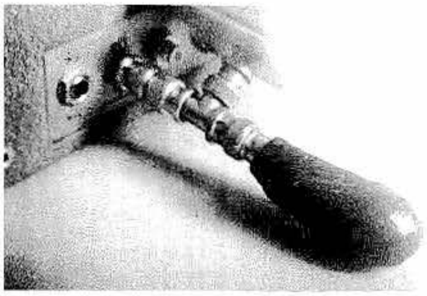

Connection to transmitter with dummy load (see text).

reading, I = 0.0005, and calculate W for half-scale. Plot these two points at 1.0 and 0.5 mA on Fig. 1, and connect by a straight line, which you may extend the length of the graph.

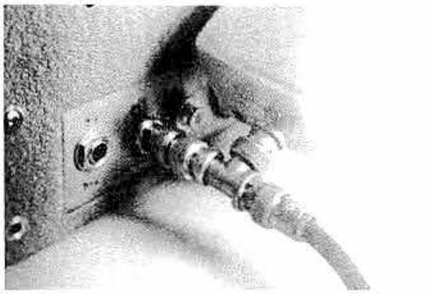

Connection to transmitter with antenna connected.

Install the meter through a coaxial tee at your antenna connector, or through some other predetermined means, and terminate with a dummy load. The one seen in the picture is three 150Ω resistors in parallel, dipped in epoxy, shielded with a copper braid, and installed on a BNC plug. Apply power and read the meter. That's it! The meter can be used with an antenna if your swr is below about 1.2:1.

So what did I promise? A low-cost, rf wattmeter that is inexpensive enough for the Novice, practical and useful enough for the General, and "Extra" accurate. Go raid the junkbox, and add a worthwhile piece of gear to your shack.

RF POWER MEASUREMENT USING HOT CARRIER DIODES

Frank Jones W6AJF

Two rf wattmeters are shown here, one with a range of 25 mW to 10W and the other covering the range of 5 to 300W. Both are useful from low radio frequencies on up through 450 MHz.

The low-power version (Fig. 1) makes use of a 20W Sierra dummy antenna built into the meter case, though the metering circuit only goes up to 10W. If the maximum is to be 20W, the reference meter reading could be about 45 μA instead of 30. The minimum power reading would be doubled. In this wattmeter, the power range potentiometer is calibrated and only a reference line on the meter is used when making rf measurements. The dummy 50Ω antenna resistor is rated up to 1000 MHz so is excellent from 450 MHz down.

The range potentiometer had an audio (nonlinear) taper. By connecting the "high" resistance end to the diode, the watt range scale is spread out quite well in the 0.1–10W range. The hot carrier diode, an HP 2900, has a 10 PIV rating, which means that the rms rf voltage across it should be less than 3V for safe operation. At 10W of rf power, the rms voltage would be a little over 22V, which means a voltage divider is needed to keep the applied diode voltage down to about 2V. An HP 2800 diode with a 75 PIV rating would be more desirable, especially if the meter was to be calibrated for 20W maximum. This diode is about $1 and has a little higher capacitance, which would require a different shunt capacitance across parts of the resistor divider to make the device work with the same power range calibration.

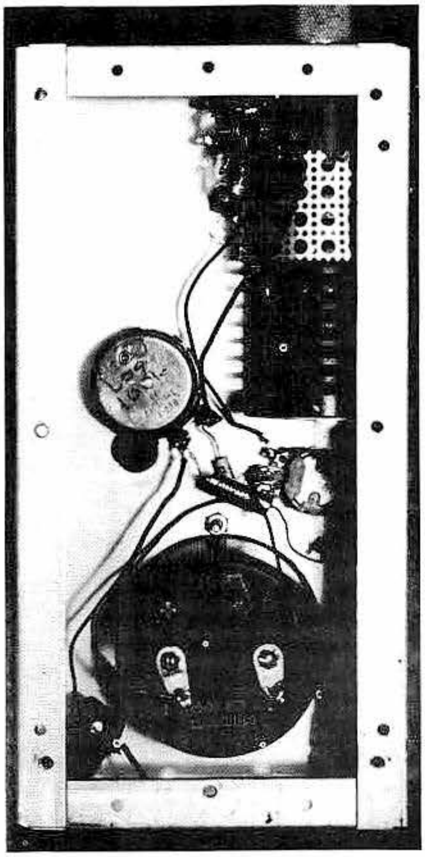

Bottom view of 10W unit with the rf dummy antenna clamped in one corner.

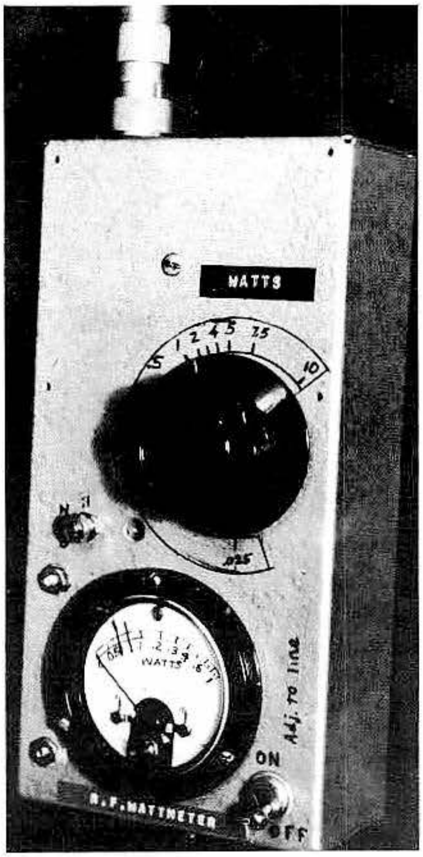

Top view of low powered rf wattmeter covering .025 to 10W. Built into a 8x4x2 chassis with wire screen bottom plate for ventilation.

The divider should use ½W resistors of the carbon or metal film type, since these units are part of the rf circuit. It is better to use three ½W 300Ω resistors in the string rather than a single 900Ω 2W resistor: this is because the rf resistance characteristic is usually better in ¼ or ½W types in certain ranges of resistance. Every resis-

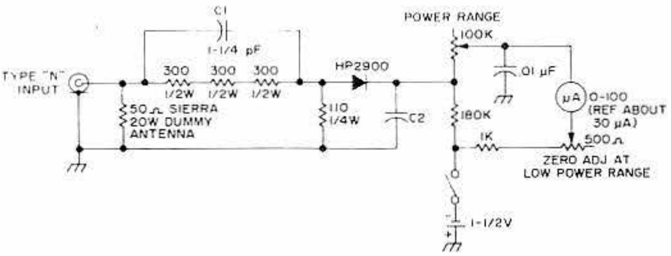

Fig. 1. .025 to 10W RF wattmeter.

tor has some inductance and shunt capacitance which becomes part of the voltage divider. The diode shunt capacitance is in parallel with that of the 110Ω ¼W resistor in Fig. 1. However, nearly any combination of resistor sizes can be equalized within 10 to 20% over the desired frequency range. This divider is across the 50Ω dummy antenna, so should not shunt the value down to less than 49 or 48Ω. This divider has to dissipate a little rf power also. Its total resistance should be at least 20 times as high as the dummy antenna load resistor.

The values shown in Fig. 1 are just about the minimum that should be used. Too high values makes it more difficult to extend the frequency range to the upper end, though it can be done, as was discovered in the higher-powered wattmeter of Fig. 2.

All diodes are poor rectifiers at applied rf voltages below their forward bias values of 300–700 mV (peak). By using a forward dc bias voltage to make the diode conduct at least 5 or 10 mA, the detection sensitivity is increased as much as 5 or 10 times. This requires a small battery, a couple of fixed-value resistors, and an adjustable pot to balance this current out of the meter when measuring rf powers below 100 mV. If the power range is limited to a minimum of ¼ or ½W, no bias circuit is needed in this 10W instrument. The range scale in either case has to be hand calibrated.

A low-powered radio transmitter or exciter can be used as a 10W power source when calibrating the power range pot scale. The transmitter can use stage detuning to reduce power outputs down to the lower values needed. Many swr meters have watts of power calibration and one of these can be put in the coax line to the rf wattmeter for calibration service. A more accurate calibration can be made by comparing the power readings against some reliable commercial rf wattmeter within its frequency range and calibration charts. This scheme is usually necessary for checking the calibration at VHF or UHF. Another method is to use an accurate rf voltmeter across the dummy antenna connection to ground and read the power values in watts = E^2/R. For example, 5V (rms) squared is 25; and divided by 50Ω is equal to 500 mW.

The Sierra 50Ω dummy antenna has no connection available at the high end of the resistor, which terminates in a type N fitting. The metering circuit has to connect to this point as close as possible by getting into the inner conductor of a coax fitting, or by drilling a 3/8 or ½ in. hole through the shell of the dummy antenna close to the rf fitting end. This can be done and the first 300Ω resistor in the voltage divider soldered to the inner connection to the large 50Ω resistor. A long 1/8 in. diameter soldering iron tip is needed. The divider resistors, diode, and four .001 μF stud-mounted bypass capacitors were all mounted around this large hole in tapped 6-32 holes for the four capacitors. Larger values of bypass capacitors can be shunted across these 0.001 μF values to ground to extend the frequency range down to low rf or even af values. For example, a .02 μF capacitor shunt would allow operation to 2 MHz. A miniature 50 or 100 μF electrolytic shunt would function at audio frequencies down to 300 Hz. The diode must have a low-impedance path to ground over the desired frequency range to function as a peak rectifier and get as much dc output voltage as possible for the meter circuit. The microammeter in series with a variable range resistor is simply a dc voltmeter. The diode rectifier converts rf voltage to dc, so the diode should be equally efficient over the whole rf range.

The 5–300W unit was built to use with a large dummy antenna rated up to 500 MHz, which is a massive unit external to the box shown in the photographs. Quite a bit of rebuilding went into this device to make one calibration of the range potentiometer fit all frequencies from 450 to 2 MHz. The input and output coax fittings had to be finally mounted so the inner conductor tips could be soldered together and the resistor divider connected to this point. The latter consisted of two 4300Ω 2W carbon resistors and a 68Ω 1W resistor in series to a copper sheet inside of the aluminum box.

The watt range variable resistor was a 500 kΩ linear potentiometer which was limited to a lower value by shunting it from the moving arm to the diode connection end with a 220 kΩ resistor. This gave a maximum power reading of 300W when the reference line was drawn on the meter face at 12 μA. The import, low priced, 0–30 μA meter had a large meter scale. A smaller 0–50 μA meter would have been usable, since the meter is used only as a reference. The range pot knob is adjusted when rf power is applied to run the meter reading up to the line drawn on the meter scale face.

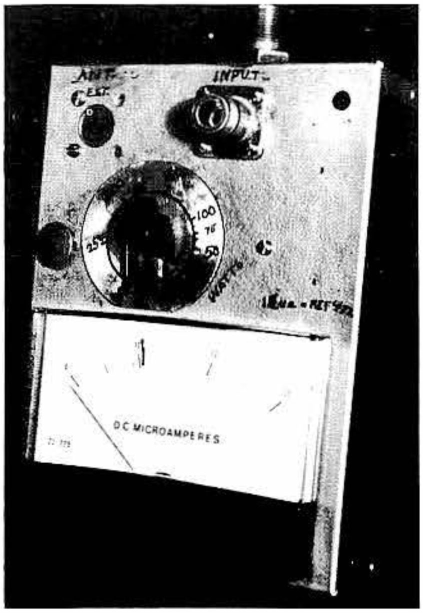

Top view of 5 to 300W metering circuit for use with external high powered dummy antenna.

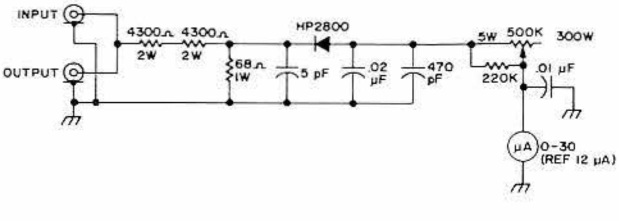

Fig. 2. 5 to 300W RF wattmeter metering circuit. External 300 or 400W during antenna load.

The circuit shown in Fig. 2 was equalized to within about 15% error over the range of 2 to 450 MHz by shunting a 5 pF capacitor across the 68Ω resistor in the rf divider.

Calibration of this device was made at 144 MHz using a transmitter having up to 400W available carrier output. The meterizing unit was connected to a large Bird rf wattmeter at the external fittings of the latter. Several thermocouples had to be used to cover the wide range of power for the calibration. This required reading a chart curve for each Bird wattmeter reading and using correction factors for frequency in order to obtain the actual watts of rf power. Now, the large unit is used without the thermocouples, charts, rf choke, etc. simply as a dummy antenna. The new metering circuit connects directly into the antenna fitting, with a few feet of 50Ω coax over to the transmitters being tested.

This power measuring device can be used in any 50Ω coaxial line to monitor the actual power going toward the antenna. The swr in the line should be low, or near unity, in order for the calibration to be reasonably accurate.

VHF DUMMY LOAD WATTMETER

Glen Zook K9STH

The unit

The unit described herein is similar to some 60 watt units which may be found around many commercial two-way radio shops. This dummy load has provision for connection to an external relative output meter. This external output meter may become an accurate wattmeter if the following criteria are met:

1. Frequency bandwidth of ±10% of calibration frequency.
2. RF output kept within power dissipation of dummy load.
3. Accurate initial calibration.

These criteria may be easily met in amateur vhf operation if only one band is considered for each set of calibration data. Since most vhf amateurs operate on 50 mhz, 144 mhz or 432 mhz, the ± 10% frequency limitations may be easily met. This limitation gives a 10 mhz bandwidth at 50 mhz, 29 mhz bandwidth at 144 mhz, and 86 mhz at 432 mhz. The limitation to the power ratings of the dummy load is only common sense, for if a resistive network is overloaded, the impedance may be drastically increased, caused by damage to the load resistors. The calibration limitation may be overcome if a standard, previously calibrated unit, or commercial unit is used.

The unit consists basically of 16 220 Ohm resistors in a series parallel arrangement. The metering circuit consists of a germanium diode pickup with necessary rf filtering. The meter movement is generally a vom, but any 50 ua meter movement should suffice. Exact physical layout is not extremely critical, but it is suggested that the layout be made similar to the unit shown in the accompanying photographs. This unit is acceptable for 60 Watt output transmitters without modification. The power capability may be increased to about 200 Watts if the resistive network is suspended in 1 quart of oil. If this is done, care must be taken to keep the metering circuit out of the oil. The lead from the diode to the resistive network must, of course, be partly submerged, but keep the diode itself out of the oil. I do not personally use this arrangement, but I know of two units which have been in use at a large Southeastern two-way radio shop for several years. The same shop has incorporated a range switch with several meter shunts for various maximum scale power readings. This feature is especially useful to the amateur vhf FM operator who may be working with equipment of from ½ to 250 Watt outputs. The schematic appears as Fig. 1, and the basic circuit for various meter shunts as Fig. 2.

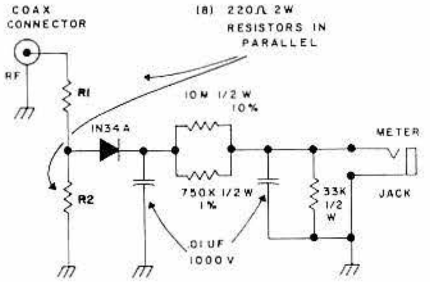

Fig. 1. Schematic. R1 and R2 consist of eight resistors each, in parallel.

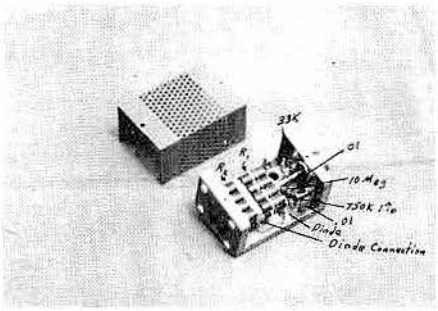

Parts layout and interior view of Dummy Load – Wattmeter

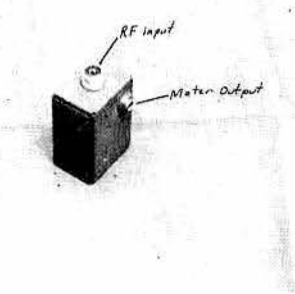

External view.

Calibration

Calibration is best accomplished by using a Bird "Thruline" or similar commercial vhf inline wattmeter. Second choice is a Bird "Termaline" or similar dummy load-wattmeter. In both cases, a graph should be created by plotting meter divisions on the horizontal axis, and power on the vertical axis. The meter shunt should be placed at minimum resistance and increased to give maximum reading at the desired power level (this holds primarily true for units using the metering circuit of Fig. 2) or, if a vom or vtvm is being used the range switch should be placed on a high voltage setting and reduced a setting at a time until the desired reading is obtained. The transmitter should be adjusted for various power levels on the standard wattmeter and the voltage or current reading on the new meter recorded on the graph. In the case where the standard meter is of the dummy load-wattmeter type, it will be necessary to switch the coax from one unit to the other. Do not retune the transmitter, for each unit will present almost the same load to the transmitter (50 ohms). Take the reading and record as with an inline type of meter. The points on the graph should now be connected with a smooth curve (use of a draftsman's

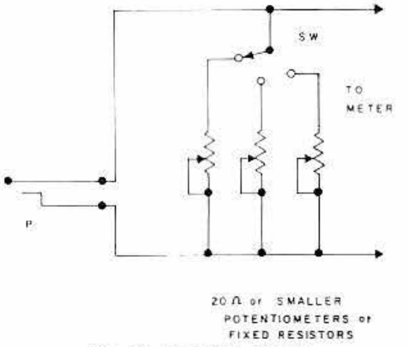

Fig. 2. Metering shunts.

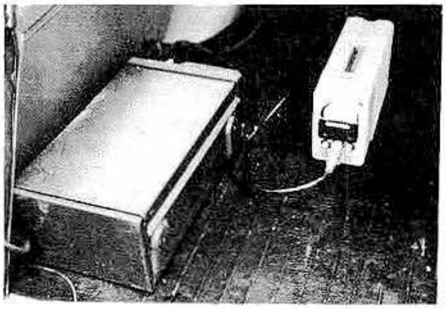

Using the Dummy Load – Wattmeter to check output of FM unit.

range switch is used, it will be necessary to calibrate for each switch position. Also, if multi-band use is expected, the graphs must be made for each band. Use of the wattmeter now requires only the connection to the transmitter, setting of range switch to the proper level, and reading the graph.

Uses

The uses of this dummy load-wattmeter are as varied as the amateur mind can devise. One very important use is determining the losses of 50 ohm coax. Measure the output of the transmitter at the transmitter. Then measure the output at the end of the length of coax. The losses in the line become apparent. The loss in db may be calculated by the standard power ratio formula, 10 $\log_{10}$ Power out of coax/Power into coax.

Another use is the determination of efficiency of final amplifier stages. This efficiency may be calculated by Power out (measured by dummy load-wattmeter)/Power in (measured by plate current/plate voltage meter) x 100%. A third use is determining once and for all which amateur really has the most output. This list may be expanded by the builder to suit his own tastes.

Conclusion

This dummy load-wattmeter is not a Bird "Termaline" nor should it be regarded as a substitute for any other laboratory equipment. However, with a little care in calibration, (assuming a 5% accuracy standard is used for initial calibration) the accuracy should be within 10%, and this, my friend, is not bad for a wattmeter costing less than $10.

Chapter IV

Measure Your Field Strength and Frequency

THE RF SNIFFER

Jim Kyle K5JKX

Every now and then there's a need to know if any rf is present in a circuit. Frequency isn't so important—the question is simply. "Is there rf here?"

Your grid-dipper can frequently answer this, if used in the wavemeter mode, but occasionally it's not sensitive enough—particularly if you're working with a receiver oscillator where power is measured in microwatts.

Here's an rf Sniffer which will indicate the slightest trace of rf in a circuit. In addition to checking receiver oscillators, it's a perfect gadget to ensure perfect neutralization of a transmitter final.

Connect the components as shown in the schematic. Use long-nosed pliers as a heat sink between the diode and the solder joint when wiring, to prevent diode damage. Note that the pickup loop of 14 gauge wire is insulated with a strip of spaghetti.

Mount capacitor and diode on back of meter with shortest possible leads. Attach pickup loop directly to negative meter terminal; it's stiff enough to do without other mechanical support.

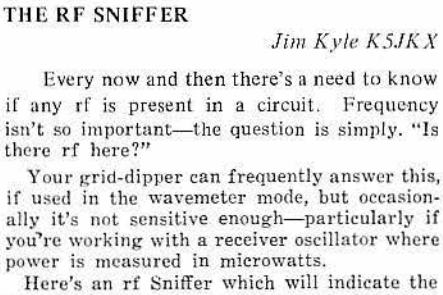

Amplifier Neutralization—Couple the Sniffer to the antenna terminal with a temporary two-turn link around the pickup loop. Remove plate and screen voltage from the final amplifier. Apply drive. Adjust neutralization for minimum indication on the Sniffer—but don't expect to be able to get it down to zero.

Oscillator Checking—Place the pickup loop near the oscillator coil. If the oscillator's working, the Sniffer will indicate rf. Touching either the grid or plate lead (use an *insulated* tool for this test, not your fingers) should reduce the Sniffer's indication.

Receiver Troubleshooting—Check the oscillator as described above. If it's okay, next check the mixer plate coil by placing the Sniffer pickup loop near it. If you get an indication here, move to the first if stage and place the pickup loop near the plate pin of the tube socket. Proceed through the receiver until you lose the indication. The trouble is somewhere between the last indication and the point at which it disappeared.

Field Strength Meter—Couple a short antenna to the pickup loop by two turns of wire around the loop. Field strength will be indicated in a comparative manner by the meter. It cannot be calibrated, but proves useful in tuning mobile or beam antennas, etc.

SWR Measurement—(Parallel lines only). Move the Sniffer long the line. Mark maximum reading and minimum reading over a half-wavelength. Divide minimum into maximum. The quotient is, roughly, your VSWR. This method is by no means exact, but will indicate whether the line is under or over a 2:1 SWR.

UHF Frequency Measurement—Set up Lecher wires. Couple the rf Sniffer lightly to the tank circuit instead of using a flashlight bulb. Use Lecher wires in normal fashion, reading Sniffer indications for maximum and minimum. This is much more exact than the normal methods.

Improvised Grid-Dipper—If you have a signal generator available, it can be used with the rf Sniffer to serve as a "grid-dip" meter to locate resonance for any tank circuit. Couple both the generator and the Sniffer lightly to the unknown tank. Vary generator frequency. A sharp rise in Sniffer indication indicates the resonance point.

THE PINK TICKET REJECTOR

A. D. Taylor GW8PG

Hams sometimes get into trouble with the FCC because they have made a mistake in tuning a frequency multiplier or PA stage. For example, if the multiplier stage following a 3.5 MHz oscillator is accidentally tuned to 10.5 MHz instead of 7 MHz, the PA can be loaded on 10.5 MHz; and if this is done even a short transmission is likely to produce an unwanted "QSL" from the FCC! Mistakes like this can be prevented by using the simple, easily built absorption wavemeter shown in Fig. 1. Only one coil is required in this meter, the low frequency range being obtained by switching a padding capacitor in parallel with the variable tuning capacitor. With switch S1 open the tuning range is approximately 6.8 to 30 MHz, and with S1 closed it is approximately 3.5 to 6.8 MHz.

Construction of the wavemeter is simple. It can be put together in a metal utility box, on a wooden panel and baseboard, or even in a cracker can of the right size. The coil can be wound on any type of ¾ in. diameter form. If nothing else is handy, a short length of dowel could be used. The coil winding consists of 7 turns of 20-gage wire, close wound. Both the tuning capacitor (C1) and the padding capacitor (C2) should be 500 pF components, but the values of the other components are not critical.

Any meter having a full-scale deflection between 100 µA and 1 mA can be used. If the wavemeter is built in a metal box, L1 must be mounted outside the box. A slow-motion drive is not needed. If C1 is fitted with a pointer type knob and a cardboard scale is cemented onto the front panel, the calibration points can be written on the scale in ink.

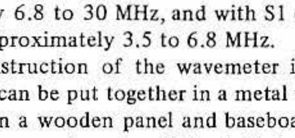

Fig. 1. L1 – 7 turns 20 gage enameled copper wire, close-wound on a ¾ in. form; C1 – 500 pF variable capacitor; C2 – 500 pF fixed capacitor; C3 – Fixed capacitor, any value between 1000 pF and 0.01 µF; D1 – Silicon or germanium; M1 – Moving coil-meter (100 mA–1 mA).

When it came to calibrating the wavemeter I thought of a simple method that I have not seen described before. I soldered about 2 in. of wire onto one end of L1, and about 10 ft of wire onto the other end. I then connected the short lead to the antenna terminal of my receiver and strung the 10 ft of wire up as a temporary antenna. I then tuned the receiver to a steady signal at each frequency at which I wanted a calibration point, and adjusted C1 until the strength of the received signal suddenly dropped sharply, indicating that the wavemeter was tuned to the frequency of the signal. I was then able to mark this frequency on the cardboard tuning scale. The drop in signal strengh was very sharp so calibration was easy. Once enough calibration points had been obtained the temporary wires were unsoldered and the wavemeter was ready for use.

The only frequency which may require a little adjustment is 30 MHz. If the wavemeter will not tune as high as this, push the top turn of coil L1 about 1/8 in. away from the other turns. This should reduce the inductance sufficiently; a spot of cement will hold the turn in its new position.

To use the wavemeter for checking a transmitter, bring coil L1 close to the tank coil of the stage being checked, apply power to the stage and rotate C1 until a maximum reading is obtained on the meter. The output frequency of the stage can then be read off from the wavemeter tuning scale.

The wavemeter can also be used as a radiation meter for tuning up single-wire antennas. If it is tuned to the transmitter output frequency and placed near to the antenna wire, maximum reading on meter M1 will indicate maximum output power from the transmitter.

Many readers will have realized that the calibration method I suggest uses the principle of the rejector circuit. That is why I have called the little gadget "the pink ticket rejector"!

A USEFUL ACCESSORY FOR THE HAM SHACK

Rex Morris W2WHX

BECAUSE we are dealing with something we cannot see, namely electrons and electromagnetic radiation (we only see their effects), we must acquire the ability to use test equipment, in order to understand and find our way in this invisible realm.

The piece of test equipment about to be described is one of the more useful instruments that the amateur should have in the shack. While this instrument serves primarily as a field strength meter, it will also serve as a phone monitor, neutralization indicator and a sensitive wavemeter.

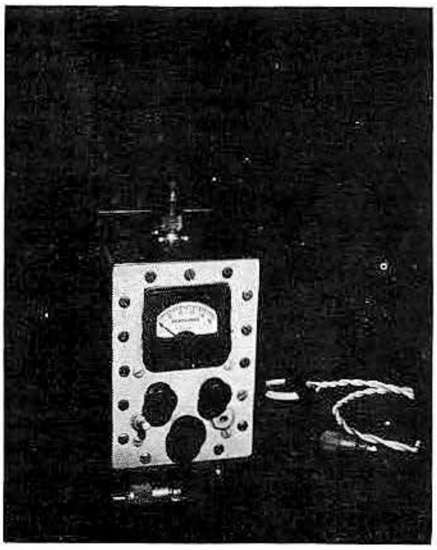

Referring to the circuit diagram, meter M1 is a sensitive instrument which indicates the pressure of rf when the device is used as a field strength meter, wavemeter or neutralization indicator.

To use this instrument as a wavemeter, a pickup loop is substituted for the short whip antenna at the top. Switch S1, the band switch on the right of the panel, is placed on the proper position. Coil L1 tunes the vhf range of approximately 90 mc to 170 mc (2 meters). L2 tunes 28 mc to 100 mc (10 meter and six meter bands). L3 tunes to 7 mc to 30 mc (40, 20, 15 and 10 meter bands). L4 tunes from 2 mc to 7 mc (80 and 40 meter bands). With switch S1 on the proper tap, condenser C2, on left of panel, is used to peak the reading on meter M1. Using a calibrated dial, frequency may be read directly. Toggle switch S2 is a high-low range switch for meter M1, providing a means of keeping the meter on scale and protecting it against burn-out. Potentiometer R1 (center of panel) is a vernier shunt control, also for keeping the meter on scale.

For use as a phone monitor, rf should be fed into the input jack with a link or pick-up wire. Once again the LC circuit is resonated to the frequency we desire. Earphones inserted in jack J2 will open the meter circuit and allow you to monitor the signal. Switch S2 is placed in the Hi position.

Field strength readings can be taken by using a short pickup antenna. Again the LC circuit should be tuned to resonance. Meter M1 will give an indication of field strength.

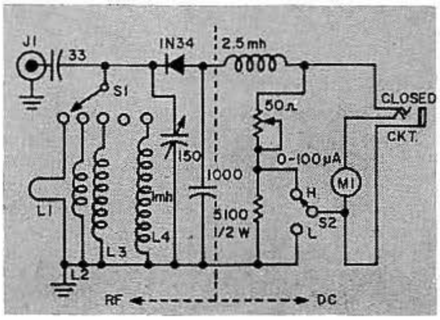

With switch S2 in the Hi position the meter is very sensitive and potentiometer R1 is a variable shunt providing much range of scale adjustment. For use as a remote reading field strength meter an external microammeter (with up to 200 feet of wire) may be plugged into J2. For making transmitter adjustments this is a very desirable feature.

Now notice that one position on the bandswitch S1 is vacant, this vacant position provides a very broad band—low sensitivity position for those extremely high rf fields where even R1 and S2 cannot provide enough attenuation.

Neutralization measurements are made by coupling the instrument through a pick-up link to the tank coil involved, with S2 in the Lo position. When the instrument is tuned to resonance it becomes a very sensitive rf indicator. It is so sensitive that it will readily be seen that complete neutralization exists in theory only.

As in all simple gadgets there are a few simple construction techniques which make the difference between gadget and instrument. In this case the important thing to keep in mind is that from the antenna to the crystal is the rf portion, and from the crystal to the jack J2 is the dc portion (with audio superimposed). With this in mind, construction is such that the two parts are separated, thereby giving some measure of protection from rf energy to the very sensitive microammeter M1. Also note the extensive use of sheet metal screws on the aluminum case. The only rf we want to enter the case is the rf we are attempting to measure via the antenna jack. Note again, we have here an instrument for detecting electromagnetic radiation, from approximately 170 mc to 2 mc. We now have an instrument for visualizing what cannot be seen.

The usefulness of this instrument is limited only to one's ability to apply it and interpret the results it gives. These only come with experience, trial and error and determined application.

Coil Data

L1—One turn hairpin loop.
L2—5 turns of #18 enamel wire space wound 1" dia.
L3—24 turns #22 cloth covered wire, close wound, ⅝" dia.

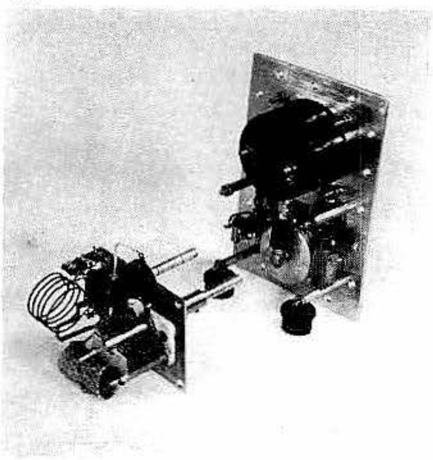

TWO THRU TWENTY FSM

Howard Pyle W7OE

I used a small LMB aluminum meter cabinet with a hole for a 2" meter; any equivalent cabinet can of course be used. With a small enclosure such as this (4"x4") the meter occupies the face of the cabinet and all controls are on the rear. Using a somewhat larger housing both the meter and the control knobs can be placed on a front panel if you prefer.

Rather than wind up the three coils I found that the J. W. Miller Company RF chokes with minor modifications were admirably suited for a compact job. One each of their catalog numbers 4606, 4588 and 4580 were required to cover the three frequency spreads from 2 through 20 meters. The #4606 coil should have five turns carefully removed, to cover the 10-20 meter bands; remove three turns from #4588 for six meters and two turns from the #4580 coil to handle 2 meters. These can all be resonated with a 50 pf midget variable tuning capacitor; a Hammarlund HF-50 or the equivalent is excellent. Any of the small multi-point selector switches can be used for band switching. A JBT lever switch type SS-14-ILS will handle the job or, if you prefer a rotary switch you can use a Mallory type 3215J as I did, leaving two positions unused. (I might want to add a couple of coils and try 40 and 80 later!).

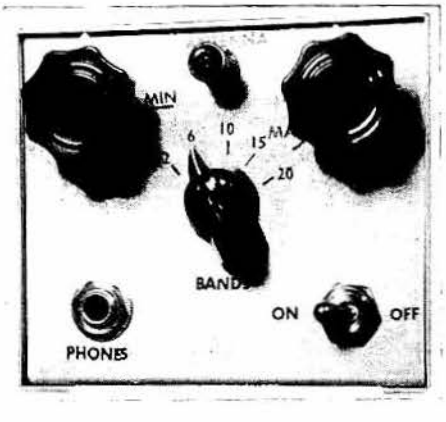

Resistors R-1 and R-2 shown in the schematic, need be only ½ watt. A single flashlight cell will serve for the battery although I chose a 1.4 volt mercury transistor battery to conserve space. Other items shown on the schematic are obvious and all parts are readily available from most electronic parts distributors as well as from the electronic mail order houses.

Build this little F/S meter and know what your VHF outputs are doing. If you want to check your modulation quality, a phone jack may be added as shown, making this little gadget really versatile.

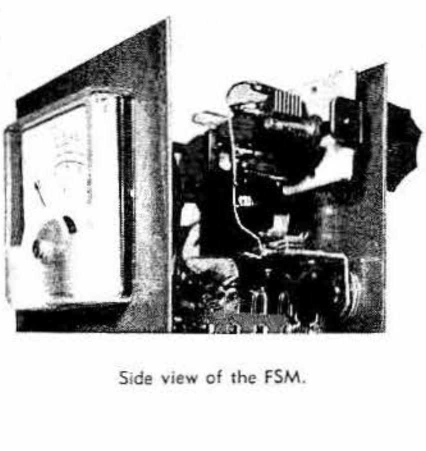

Side view of the FSM.

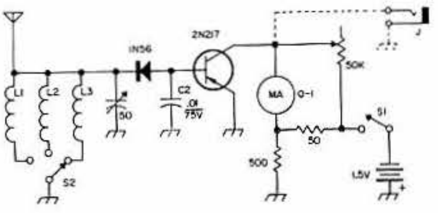

Fig. 1. The bandswitching, transistorized field strength meter for two through twenty meters.

AN AMPLIFIED, CALIBRATED SIGNAL STRENGTH METER

J. L. Iliffe VE3CES

Recently, I had the problem of tuning a four element quad. As you may or may not be aware, these beasties are supposed to be tuned from the rear for minimum signal. A quick check showed none of my friends had a signal strength meter, so I prepared to degrade myself and buy one. A look at a few prices convinced me to build.

Since I wanted some other information on the quad, like front to back ratio and the effect of more or fewer elements, I decided to add a calibrated attenuator and enough gain to make a fairly wide input range. It also had to be cheap!

The result is shown in Fig. 1.

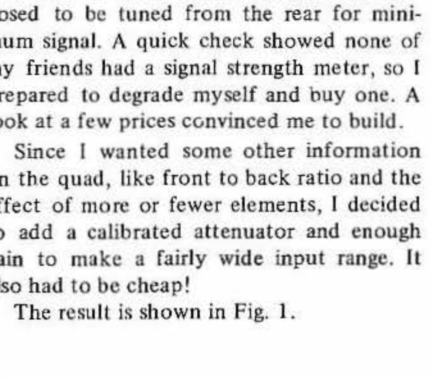

Fig. 1. Diagram of the field strength meter.

To use the normal rf attenuator method of switched T-pads requires complicated shields and quite a few resistors. Also the attenuator has to be terminated in its characteristic impedance to read correctly. To bypass this problem I first detect the rf, then attenuate the dc. This has the added advantage that the circuit is no longer frequency sensitive.

The incoming rf is tuned by C1-L2. C1 can be any small variable. I used both sections of a dual 15 pF because my local surplus store has them for 60¢. For VHF use only 1 section.

L2 is wound on a plastic pill bottle about 1" in diameter and tapering to 7/8". To cover 13–24 MHz, I used 11 turns spaced over about an inch. L1 is 2 turns over top of L2. I tried bandswitching with another pill bottle fastened on the other side of the shield from L2. The idea was to bandswitch another frequency range but I find it more convenient to wind on coils as needed. Use the grid dipper to get you in the ballpark. I have used this meter as high as 72 MHz without trouble. D1 can be any diode. I used a 1N34 because I could then specify it and know it would work, but I tried a computer type which also worked. If you prefer the meter to peak rather than dip, reverse D1.

Rectified rf from D1 is put on the top of R1, the calibrated attenuator. R2 in series with R1 gives the 0 dB point at its junction. For a 30 dB range, R2 is 47K if R1 is 1 MΩ This doesn't quite fill the range but is close. Changing the value of R2 will change the range but 30 dB is considerably more than the F/B ratio of most beams.

Q1 is a 2N5033 FET. The high input impedance of Q1 allows us to set the calibration of R1 directly by dc voltage measurements on the VTVM since it does not draw any base (gate) current. It is a p-channel device. If you use an n-channel type you will have to change the entire biasing of the circuit and also reverse D1.

Q2 and Q3 form a differential amplifier to drive the meter. Q2 is necessary to avoid loading Q1 and I had quite a bit of trouble balancing the meter against battery voltage changes until I added Q3. It will now operate from 8.5–9.2 volts with no trouble. The 10K pot in Q3's base centers the meter. The meter I used is a 250 μA tuning meter with no markings on it except a red/white/blue bar. This is all you need since we calibrate on R1, not the meter.

Operation

The meter is quite sensitive and with a two foot antenna I could get a reading several hundred feet behind my quad at 60 watts input. First tune the input (which is quite sharp) with the attenuator set at zero. This is the least sensitive position. Now set the meter for a convenient reading near the center scale with the incoming signal still on using R3. Adjust your antenna. When you feed it power again the reading will not be quite on scale on the meter but turning up R1 will allow you to put the meter back to the original position. Do not touch R3. The reading on R1 is now the increased gain in dB needed to bring the signal back to its original strength. In other words, the decrease in signal strength.

Note that during measurements you need a received signal to use R1. With no signal the meter will be off scale.

I have also used the meter to align oscillators and doublers in my two meter receiver. A probe can be made for this from two turns of wire on the end of a piece of coax. The high gain available allows the pickup loop to be quite far away which reduces detuning. Adjustments show up well on the meter.

FREQUENCY MEASURING EQUIPMENT AT MICROWAVE FREQUENCIES

Silas Smith WA9VFG

This article is not intended to give the theory, but rather a practical solution to the building and use of wavemeters at microwave frequencies.

In microwave work, frequency is one of the most important measurements. It must be understood the wave length in the devices described here is not the exact frequency wave length. A well-constructed wavemeter that has been calibrated can be very precise. They can be within 1.5 MHz at 10 GHz or less than half of a MHz at 1250 MHz. Temperature has some effect on the frequency. Most commercial wavemeters are constructed of Invar, a metal that changes very little with temperature. Some parts are of bi-metal construction to compensate for temperature. For the average experimenter, brass and copper will have to suffice. Although silver plating is desirable, it isn't an absolute necessity. Frequency at microwave frequencies can be measured by three methods: wavemeters, slotted lines, and frequency comparisons. All of these methods are used commercially. The frequency comparison is usually used in the laboratory to calibrate the wavemeter and the slotted line. As a general rule, any method of frequency measurement used at lower frequencies can also be used at the microwave frequencies, but are not always practical. The resonate cavity as a wavemeter is used in microwave measurements.

There are three types of cavity wavemeters: the transmission type, Fig. 1A and 1B, the reaction type, Fig. 1C, and the assorption or absorption type, Fig. 1D. All are resonate cavities. The way in which the wavemeter is used determines the type.

All wavemeters are adjusted for maximum readings except the assorption type. The assorption is adjusted for a dip in power output. The most popular wavemeter used by the beginner is the open circuited transmission line type, Fig. 2. This type of wavemeter is the equivalent of lecher wires. (Open circuit refers to the standing wave within the cavity, not the physical construction except as it pertains to the frequency wave length.) The practical physical dimensions are not many. The inner circumference of the main tube should be less than one wave length at the highest frequency to be measured. The rod should be small compared to the tube. If inductive coupling is used, the inductive coupling should be close to the shorted end. For probe coupling, the probe should be close to the middle. The open circuited transmission line is generally used in two ways. This type can be used "in line' (Fig. 1B) as it has very little loss when it is resonated. However, it should be removed from the line before transmitting, as it will act as a narrow band filter. The half wave length is the measurement between the two successive points at which the generator will load to maximum, as the rod is inserted or withdrawn.

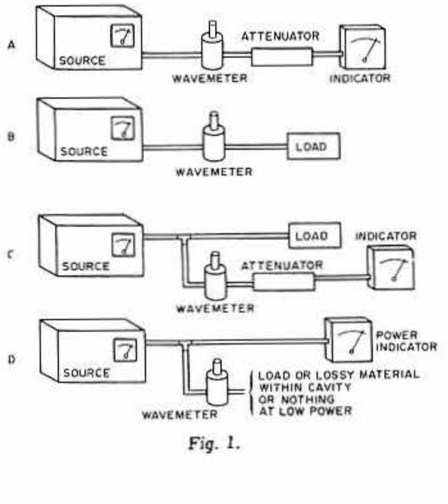

Fig. 1.

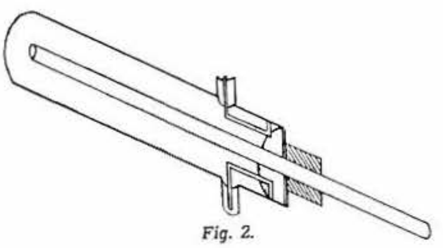

Fig. 2.

Another method in the use of the open circuited transmission line calls for the use of an additional circuit, as in Fig. 3. The circuit is a simple crystal diode detector connected to a microampmeter. The diode and condenser are usually built into the connector, as the leads should be kept as short as possible. The half wave length measurement is made on the rod between two successive maximum readings on the meter, as the rod is inserted or withdrawn. See Fig. 1A and 1B for the setup.

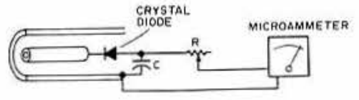

Fig. 3.

The quarter wave coaxial cavity is actually a shorted coaxial line one quarter wave length long (Fig. 4, 5, 6). As illustrated in these figures this type makes a very good cavity to use as a standard. To calibrate, a chart is made of the micrometer settings at different frequency wave lengths from a calibrated source generator. Don't tell them so, but the Public Relations Department of the Telephone Company may help you here, if they have any microwave technicians close by. The quarter wave coaxial cavity can be either physically open or closed. If closed, the closed end should extend at least a quarter of an inch beyond the center conductor at its lowest frequency. The closing of the end will lower the resonate frequency. Probe (capacitance) coupling as used for coupling in Fig. 5 will shorten the center conductor, and loop (inductive) as used in Fig. 4 will lengthen the conductor. In Fig. 4 we change the length of the center conductor to change its one quarter wave length. In Figs. 5 and 6 the center conductor remains the same, and we change the resonate frequency by capacitance. This method makes it necessary to construct the center conductor very short as compared to the full quarter wave length as in Fig. 4. These devices are not longer, so caution must be used when calibrating. The closed wavemeter as indicated in Fig. 7 is a shorted coax line at each end. The wavemeter uses a shorting plunger which is movable along part of its length. If used as the quarter wave coaxial cavity, the center conductor must be longer than a quarter wave length.

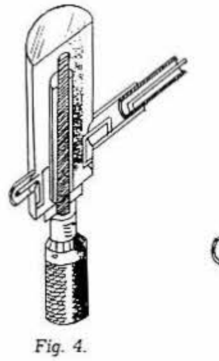

Fig. 4.

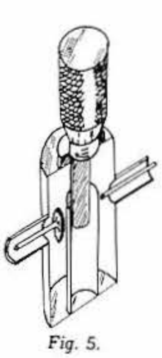

Fig. 5.

Up to this point we have covered most of the wavemeters that could be used from around 144 MHz up to approximately 3000 MHz. 1000 MHz to 3000 MHz are usually called the lower microwave frequencies. If the inner circumference of the outer tube is kept less than one wavelength, these wavemeters will operate in the desired TM mode.

There are four ways to couple energy into a wavemeter, loop (Fig. 4), probe (Fig. 5), direct (Fig. 6) and slit (Fig. 8). The most commonly used is the loop, as it has very little effect upon the electric field. The usual methods for changing loop coupling is to change the size and orientation of the loop. Loop coupling is usually placed in the high current area of the wavemeter. Capacitative coupling is changed by the size of probe and the distance from the center conductor. Capacitative coupling is usually placed at the high voltage portion of the wavemeter. As in Fig. 5, a small probe – say ½ in. piece of No. 22 wire, for example – may require an external voltage amplifier. The smaller the probe, the less effect on the resonate fre-

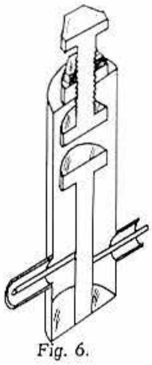
Fig. 6.

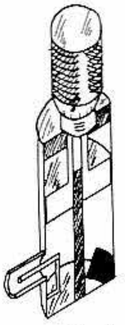
Fig. 7.

quency of the wavemeter. In direct coupling as in Fig. 6 the primary concern is impedance. To increase the impedance, move the coupling up the line away from the shorted end. To decrease the impedance, move the coupling down the line toward the shorted end. Slit coupling (Fig. 8) can be a small hole or a series of small holes or a slit. Its purpose is to allow a certain amount of leakage. In all forms of coupling, it is desirable to use loose coupling, as the wavemeter will have less effect on the system, and the Q of the circuit will be higher.

Above 3000 MHz, usually only the tunable wavemeter is used. The cavity is one quarter wave length long. The cavity can be coupled in three ways: loop, probe and slit. Because of the high frequencies, the slit is usually used, and the meter is most often kept as an assorption meter. In the assorption method, the wavemeter should be detuned when not in use. Some of the cavity wavemeters have a little lossy material added to absorb some of the energy, as in Fig. 7. Lossy material can be made from graphite impregnated cloth in epoxy.

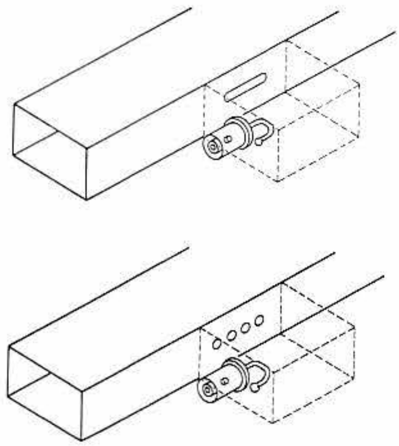
Fig. 8.

There is one other type of wavemeter that can be briefly mentioned; it is the reference wavemeter. The reference wavemeter is of any design as described, but would be constructed more like Fig. 6. It can be locked when adjusted to a selected frequency and used as a reference standard. The micrometer assemblies can be made from any micrometer with additional parts welded on. I constructed one using an oversized tube over the main cavity, and dimpled it at various places around its circumference until it fit smoothly over the cavity, and I used a piece of 3/8 threaded brass pipe as the main adjusting screw. I am sure you can come up with a good one without any backlash. This wavemeter spread the 1250 MHz band out to over 100 inches by rough measurements. I haven't calibrated it, so I can't say for sure just how far. There are three nice veeder root counters in the APX 6 which would make excellent wavemeters plus sliding contact material. One could even use the entire cavity.

The last method of microwave measurement that we will look at is the slotted line (Fig. 9). The slotted line is a section of coax line along which is cut a slot. A probe, which is a simple crystal detector with a one quarter wave length shorted stub for a dc return path, is moved along near the center conductor of the slotted section. In this case we are looking for two successive minimum readings along the line. The distance between these readings is one half wave length. A slotted line should also be calibrated. If calibrated at one spot near the intended frequency to be measured, a chart will not have to be made – just a K factor obtained. The distance between two successive readings times the K factor should equal the frequency half wave length. There are a few accessories that can be either built in or used externally. One is a coax attenuator shown in Fig. 4. It will slide in and out of the other half. It too can be calibrated if one wishes. A 50Ω resistor can be used for an impedance match for 50Ω lines if inductive coupling is used such as in Fig. 10 and when the loop is small. The line stretcher (Fig. 11) is useful with the slotted line. It merely consists of two coax sections, one sliding into the other.

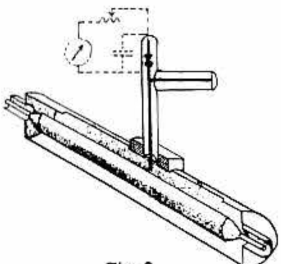
Fig. 9

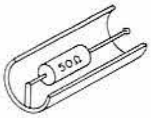

Fig. 10

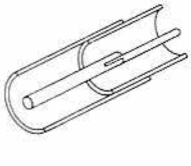
Fig. 11.

FREQUENCY METER – 1 to 10 GHz AMATEUR MICROWAVE

Bill Hoisington K1CLL

A quarter wave coaxial cavity is used up to about 5 ghz, and from there to over 10 ghz the three quarter mode is used. A complete explanation of these types of operation is given.

The same type of unit can be used as a very good tuned mixer from 1 to 10 ghz.

The Coaxial Cavity

The basic circuit of the coaxial cavity is shown in Fig. 1. A cylindrical outer cavity wall encloses a round rod some 4 inches long which is the center conductor–this center conductor is grounded at one end.

The Shape of the Cavity

The exterior shape of the cavity is shown in Fig.2. and is seen to be rectangular in cross section, with two thin walls and two thick side walls. Believe me, this configuration was not arrived at in one day! De-

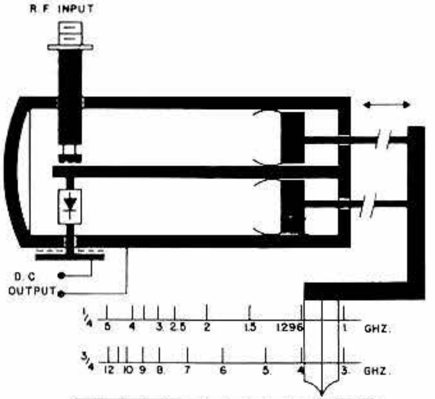
R F INPUT
D.C OUTPUT
1/4
5 4 3 2.5 2 1.5 1296 1 GHZ.
3/4
12 10 9 8 7 6 5 4 3. GHZ.
C M. 1 2 3 4 5 6 7 8 9 10 11 12 13
Fig. 1. Basic coaxial cavity.

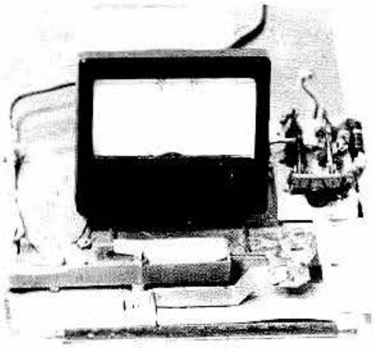

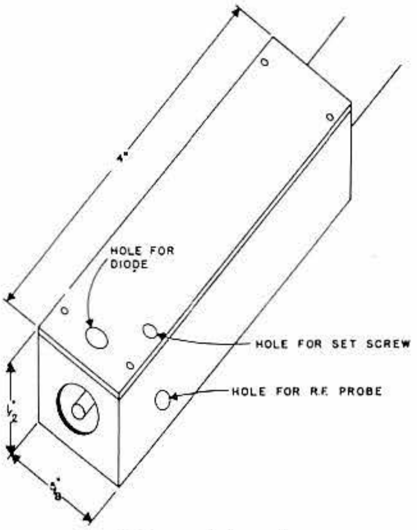

Fig. 2. Shape of the cavity.

signing tuners for X Band, I gaily started in with sections of thin-wall round pipe, the way I'd always done on uhf. The first thing you run up against is, how do you make the diode bypass capacitor? Machine out a curved saddle piece to fit exactly over the outer wall? Possible, but too expensive. And then how do you introduce the rf probe coupling into the cavity? Add on a "saddle" with a hole in it? These considerations and others, such as mounting (more saddles?) led to the abandonment of the pipe as a shape for microwave cavities; but not until a lot of time had been spent on the above mentioned items.

Diode holder and capacity

Looking at **Fig. 3.** you will see the first answer arrived at; but only after weeks and weeks of making different types and shapes. The center conductor is slightly flattened and drilled out to fit the diode prong. An 8/32 copper machine screw is drilled out to fit the other prong, then slotted with a fine jeweller's saw, and then compressed slightly to an inside diameter a shade less than the OD of the diode prong. In this way the copper screw will hold the diode as you insert it into the cavity. Believe me, that helps!

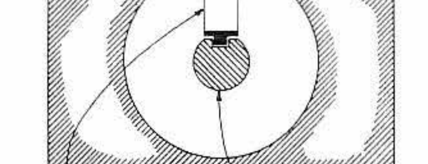

Fig. 3. Diode holder and capacitor.

The second answer is also evident from **Fig. 3.** as the diode bypass capacity can now be made efficient at X-Band. As mentioned before, you cannot "buy" a capacitor "good for X-Band. You can make it though, as shown in **Fig. 3.**, if the cavity body has been designed correctly for it. One of the thin wall sides of the cavity is drilled out (or machined out) just wide enough to clear the diode and it's holder, which is the 8/32 copper screw. The copper capacitor plate, which is thick enough to take at least a half dozen 2/56 threads, is drilled and tapped for the 8/32 screw, and clearance drilled in the corners for the 2/56 mounting screws. A soldering lug for the dc connection is used under one of these, and a three mil (three thousandth of an inch) thick sheet of fiberglass cut out to fit, larger than the plate. This helps to keep metal particles from lodging inside the tiny crack that might be there if the fiberglass sheet did not extend out beyond the plate all the way around. You can begin to see some of the detail needed at X-Band.

Further reasons for the rectangular cross-section now show up in **Fig. 4.**, which details the rf probe connections. This item was also very troublesome in first models using pipe walls, where "more saddles" was the only solution. All "saddles" are eliminated by the rectangular shape. Small semi-rigid cable is used for the connector. I have some short lengths with X-Band antennas connected to them for use as "In-Space" pick-ups, feeding directly into the wavemeter cavity. There is at times an advantage in this type of "energy collection" (antennas) which will be taken up later.

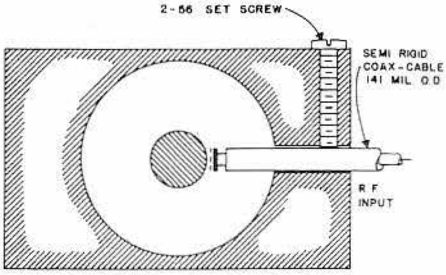

Fig. 4. RF probe connector detail.

Fig. 5. shows detail of the treatment of the cavity end of the rf cable, or probe. The outer conductor is cut away for about one quarter inch in length and removed. About a sixteenth or so of the Teflon is left, which is then removed from the center conductor. A thin copper washer (which I generally cut out of sheet copper since the hole to solder the center conductor is quite small) is then soldered to the center conductor, making the "capacity probe", as shown in **Fig. 5.**

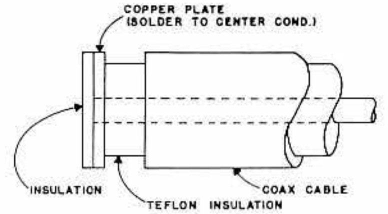

Fig. 5. RF probe detail.

Mylar tape or other good insulation is fastened to the side of this washer facing the center conductor. With this insulation in place you can push the probe all the way in, while testing, and still not have a dead short. Different thicknesses of fiberglass sheet can also be cemented on, to make up more permanent types of fixed capacitors, of different values.

For some uses, particularly in this one as a wavemeter, loose coupling is desired, but it must be securely locked with the set screw, otherwise your dial calibration and frequency reading will suffer.

Plunger fingers

Here is the most difficult item. It is hoped to have stock pieces made up for this work that you can purchase at reasonable cost. The fingers should be made of tempered beryllium-copper, which is not easy to work with.

Fig. 6. shows some details of the plunger and fingers. I assume, having been told so by "well-informed sources" (mechanical engineers) that these units should be made in a machine shop by competent machinists. Maybe so, as the ones I have made here in the shack by hand tend to lose their tension if not handled carefully.

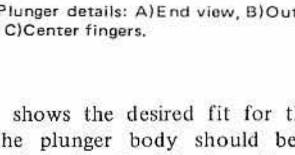

Fig. 6. Plunger details: A)End view, B)Outer fingers, C)Center fingers.

Fig. 7. shows the desired fit for these fingers. The plunger body should be an easily slide-fit *inside* the ¼ inch cavity, and the center hole in the plunger after the fingers should also be an easy fit *over* the center conductor.

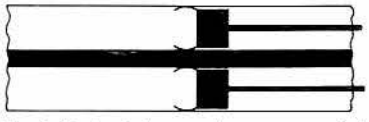

Fig. 7. Desired shape and curvature of the plunger fingers.

Two steel push rods lead back from the plunger through small holes in the back end of the cavity (see **Fig. 1.**); these terminate in the brass block which is furnished with a pointer for the frequency scale. Maximum extension of the plunger should be up against the end piece, as a positive reference point for the dial, in case of trouble after calibration. This point should be indicated on the scale as "minimum frequency" in order to reset the pointer if it should ever become displaced after calibration.

The diode

At present, the diode used is an X-Band "pill package," with a prong at each end as shown in **Fig. 3.** These are point-contact diodes, like the famous 1N23 ceramic cartridge types of World War II fame, only a lot smaller. Referring again to **Fig. 3.** always make sure that the ceramic part of the diode is, as nearly as possible, in the open space between the inner and outer conductors. This space is where the rf is! It is also important to make sure that there is as much metal surface continuity as possible along the cavity wall, across the fiberglass sheet X-Band capacitor insulation onto the diode capacitor plate, and from there over to the diode holder and onto the metal end of the diode.

The rf is at a maximum between the inner and outer conductors, which is an air space of a sixteenth of an inch. and that is where the diode should be.

The diode rf bypass capacitor, formed by the diode plate and the flat top of the cavity body, need only have a capacity which is relatively small; anything over about 20 pF is sufficient. What it *must* have is the proper *lack* of inductance! The details of how this act has been covered in previous paragraphs, and if you follow those details you will find little or no rf on the *outside* of the diode capacity plate or the dc lead from it.

X-Band is not just short waves. It is really short; like a quarter wave at X-Band equals 9/32 of an inch as you can plainly see, if you get one (or more) of those little plastic millimeter rulers in a stationery store for 5 or 10¢. Be sure and get some, by the way, if you're going to do anything above two meters.

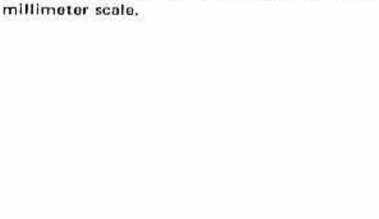

Fig. 8. Full waves for S, C and X-Bands on millimeter scale.

Fig. 8. shows the millimeter scale, with s, C, and X-Band plainly showing.

A handy wavelength-frequency chart is included here for your convenience, which is useful from the khz range way up *above* X-Band. See **Fig. 9.** Get to know the easy reciprocals, like 1 centimeter equals 30,000 mhz, 3 centimeters equals X-Band, 10 centimeters equals S-Band (3,000 mhz), 1,000 mhz equals 30 centimeters, etc. Very useful!

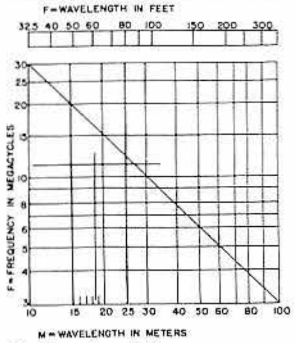

Fig. 9. Wave-length/frequency converter. Use of multiplying factors such as those at the bottom of the graph will cover any portion of the electromagnetic-wave spectrum.

The 3/4 mode and harmonics

Don't worry about that word "mode." Generally when something odd takes place in a cavity or waveguide, certain types of engineers tend to fall back on obscurantism (I seem to have fallen for that $64 word. It just means covering up). They say, "It jumped mode", or, "Spurious showed up."

Here's the straight dope. **Fig. 10** shows the quarter wave "mode" of operation. Starting at 1 ghz you will find *one* point of maximum dc output. If the oscillator under measurement is "running hard" with lots of 2nd and 3rd harmonic energy content, these will be found at 2 and 3 thousand megahertz, and possibly higher ones, which should drop steadily in power as you go up. The diode itself may cause some of these if hit too hard with the rf input.

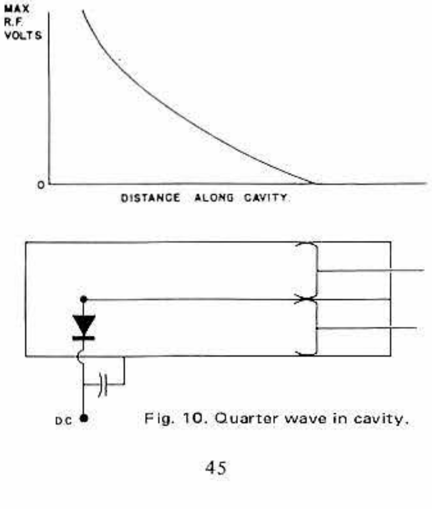

Fig. 10. Quarter wave in cavity.

Fig.11. shows the 3/4 wave mode, which is a very "natural" type of operation. Don't forget that in an instrument of this kind you are looking for standing waves and you want them to be of the greatest amplitude possible (within reason). So, if you tune the cavity by the plunger so that it measures some three quarter waves on it (allowing for length-loading of the diode on the first quarter), you will find two peaks on the meter due to the situation shown in **Fig.11.** The higher the Q, and the lower the losses along the line, the more quarter waves can be found. For the 4 inch cavity shown, three quarter waves at S-Band are the longest that will fit.

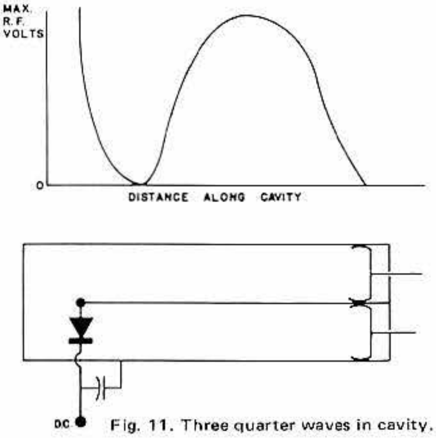

Fig. 11. Three quarter waves in cavity.

A check on this operation is easy. Using the millimeter scale on the "dial", take several readings between maximums, for example, 22, 37, 51, and 67, add the spacings together, which comes to 45 millimeters, divide by three (the number of samples), and you will find an average of 15 millimeters for the waves which are standing on the center conductor (or "along the cavity", if you prefer) and there you are, 15 millimeters for the half wave, 3 centimeters for the full wave. Which is X-Band at 10,000 mhz or 10 ghz.

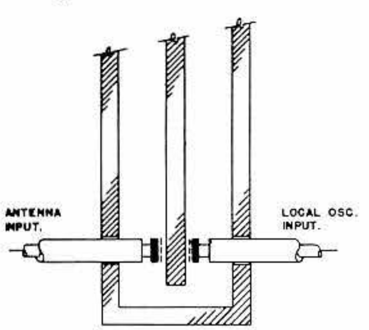

Fig. 12. Double input detail.

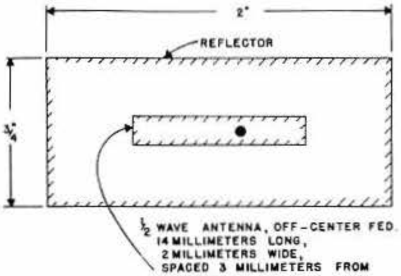

Fig. 13. Test antenna–X-Band "two-element".

If you find numbers which are not well known, you can find the frequency on the chart, at least close enough to put you in one of the microwave amateur bands, such as 5,650 or 10,500 mhz.

Use as a Microwave Mixer

This same type of cavity can be used from 1 to 10 ghz as a mixer for the front end of a superhet receiver covering those frequencies.

This application will only be touched on briefly here as the whole receiver is detailed in another article in *73 Magazine*.

Fig. 12. shows how to do it, so you can plan on this use, and make more than one, if you wish to.

Looking at Fig. 12., you can see how useful it is to have *two* thick sides on the cavity, one for rf input and one for the local oscillator input.

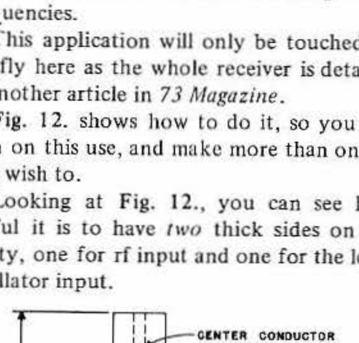
1⅜"
1⅜"
CENTER CONDUCTOR INSIDE INSULATION.
SLEEVE CONNECTED ONLY TO OUTER CONDUCTOR.
SEMI-RIGID COAX.

Fig. 14. Test antenna–S-Band. Dimensions suitable for amateur S-Band 2,400 mhz (omni-directional).

Conclusion

That about covers the details and some uses. The whole unit can be mounted on a piece of copper-clad, along with a 50 ma meter, the dial scale, and the centimeter-frequency chart. I broke down on this one and used a "regular" small microwave input connector for the rf. (Instead of an "RCA Phono Jack".) For connections to other units, such as oscillators and multipliers, small flexible cable may be used.

Fig. 13 shows a test antenna for X band, not the best in the world but good enough for a starter. With a lens in front it really picks up signals. Fig. 14 shows an S band antenna for the 2,400 MHz amateur frequencies.

LAMBDA LINES

James Ashe W2DXH

Over the past few years, new kinds of commercial gear and military surplus have made it quite easy to generate VHF frequencies. But the problem of measuring the frequencies has not become any easier for amateurs short on calibrated instruments. Of course this means the newcomer to VHF! What can he do to find what ballpark a circuit is radiating in, if there are no accurate devices available? When this problem came up recently, a simple solution appeared quite by accident. It was so simple, in fact, that its simplicity must be the feature that has kept it out of the ham publications. Another first for 73!

The traditional solution to the rough frequency measurement problem is to make up a Lecher Wire system. There is some question about the value of one of these in the modern ham shack. Narrow-band crystal-controlled techniques guarantee frequency and stability once the multipliers are tuned properly. In the old modulated oscillator days things were not that stable . . . so why go to all that carpentry and construction work for what fairly well promises to be a use-it-once gadget? Particularly when a little reflection (pun intended!) may bring out a cheaper, faster and better arrangement?

Ham and commercial builders of VHF gear have been using tuned stubs for years to match impedances, tune out frequencies, tune in others, etc. Yet it seems to have occurred to very few workers indeed that it might be possible to cut stubs to length accurately enough to serve as frequency standards. Apparently this can be done, with an accuracy of about 5%! This compares very favorably indeed with the performance of lower-frequency grid dip meters and some signal generators. It's pretty good for a pencil and yardstick operation: the only other items required are some understanding of how it works and a piece of 300 ohm twin lead. Belden #8235 recommended. You can calibrate that new GDO for 432 at a cost of just a few cents!

Theory

Many kinds of things show a property of tuning sharply to a certain frequency. This property is called resonance. We hear it when a struck piece of metal rings, and see it in the pendulum of an old grandfather clock. The grid dip meter shows a drop in grid current of an oscillator when a nearby resonant circuit steals energy from the oscillator. And it is the nearby resonant circuit that is the subject of this article.

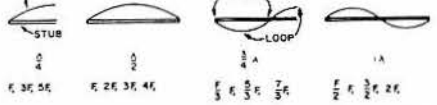

Fig. 1. Four kinds of tuned stubs, taking their fundamental resonances as the frequency producing the illustrated voltage distribution. This shows that none of them has a unique resonant frequency.

The basic circuit is the quarter-wave stub. A little browsing around in the handbooks and earlier issues of 73 and other ham magazines will tell you lots about quarter wave stubs. The important practical points are that the stub resonates at certain frequencies, and that at these frequencies it can be dipped at its shorted end in the same way as any other resonant circuit.

But the term 'quarter-wave' has to be taken with a grain of salt. The tuned stub will be shorter than a free-space quarter wave, because the dielectric has a slowing-down effect on the rate at which the RF bounces end-to-end along it. Suppose you laid out a mile or so of twin-lead and transmitted a signal, at the same time sending off a reference signal by space wave. The reference signal would arrive at the other end first, in about 5.35 microseconds. The twin-lead signal would arrive a full microsecond later, about a 20% delay. Since this applies even to short lengths of twin-lead, the delay must be taken into account for accurate measurements. Also, there is a considerable difference in velocity factors between different brands and qualities of twin-lead.

Crystals are often used in overtone oscillators for generating stable VHF frequencies. The various modes of oscillation are pictured in the handbooks. Tuned lines will also show overtone resonances, and in the case of large uncertainty, it might just happen that reasonable errors could lead to a consistent but very wrong result. A halfwave line will resonate at a frequency f, and also at 2f, 3f, and so on. Note both odd and even multiples! All other resonant lines have a similar overtone resonance property. The problem is slightly aggravated by the convenience of using relatively long lines at the higher frequencies because they are easier to handle. The solution is to cut a pair of lines whose collections of resonant frequencies have only one resonance in common. The recommended lengths are a halfwave and a three-quarter wave line.

Fig. 1 shows four basic tuned lines. Just which resonance is an overtone and which is not depends somewhat on the application. The simplest way out of this problem in semantics is to say that the three-quarter wave line really doesn't have that resonance at f/3, ignore the quarter and fullwave lines, and stick to the remaining two for test work.

At 432 MHz a wave in free space is about 27.3 inches long. Suppose we are using Belden #8235 twin lead, which Belden says has a velocity factor or propagation constant of 0.77. The twinlead wavelength than is 27.3 times 0.77 or 21 inches. The halfwave stub must be 10.5 inches long, shorted on both ends; and the three-quarter wave stub 15.75 inches long, shorted at one end. These are convenient lengths, not too long to use on the workbench, nor so short that percentage accuracy in cutting becomes a big question.

Using the stubs

The commercially available grid dip meters are not noted for accuracy. It's commonly estimated that the scale calibration can be

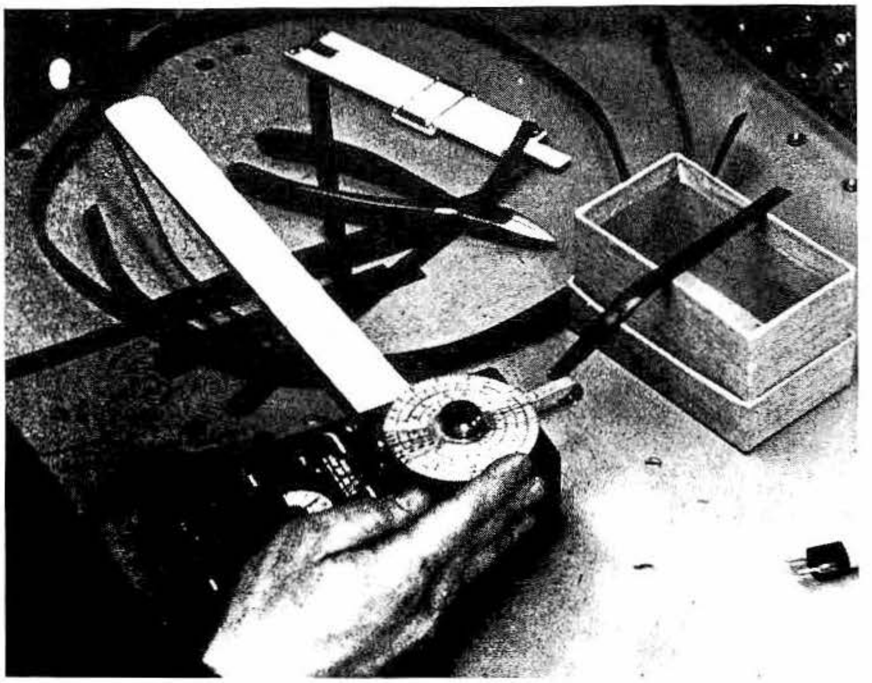

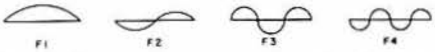

Fig. 2. Some overtone resonances of a halfwave stub (shorted at both ends). F1, F2, F3 and F4 should be 1F, 2F, 3F, 4F.

trusted to within about 20%. With some care, calibration points taken from twin-lead resonators appear to be good to about 5%. The first precaution is accurate construction. Cut the strips slightly long, short one end of each, and cut the three-quarter wave line to length. Then go more carefully at the other end of the halfwave line, which must be shorted at both ends. It should not be too hard to get the correct lengths within one sixteenth inch.

When making frequency checks, the lines must be held off the workbench an inch or two. Use small boxes or pieces of cardboard. Probably the better part of a foot distance is in order if the workbench is of metal or has a copper surface. At two meters, a perceptible change in calibration can be detected if the line is laid out on a wood surface! It's very good practice to make up lines for two meters or lower, and practice dipping them. Some refinement of technique will certainly be required before a halfwave and a three-quarter wave line can be made to dip at the same point on a standard dip meter. Once the trick is mastered, it can be carried up to the higher frequencies.

The lines are dipped in the same way as any coil. Because they have a very high Q. there will be a tendency for the dip oscillator to pull, or to seem to give different readings when tuning down to frequency and tuning up to frequency. The remedy is less coupling: move the dip meter a little further away from the line and try again. Dip the stub at its shorted end!

But what was that trick for calibrating a dip oscillator, mentioned earlier? Can't make up a pair of lines for each frequency. No need to! That's simply the reliable way for finding the right ballpark. When you're there, you can set the lines aside, make up another three-quarter wave resonator cut for the lowest frequency, and after marking that point on the scale, trim the stub up to the next calibration frequency. Throw the remainder away when done calibrating.

The half-wave stub is also useful as a tuned coupler. Suppose you want to tune an oscillator to a particular frequency but have nothing to indicate at that frequency. Loosely couple the RF into one end of the half-wave stub, and take it out the other end with a hairpin loop, through a diode to a 50 μA meter. You will only get a reading at the resonant frequency of the half-wave stub. Simple!

Chapter V

RF Signal Generators

ZERO-BEATING WITH A FREQUENCY METER

Jim Harrison WB4TBX

After putting an LM-18 in service, I found it a nuisance to plug in headphones every time it was necessary to adjust the corrector control to produce Zero Beat at one of the crystal check points. I hooked up a phone plug to the plate winding of an audio output transformer, and then connected a small loudspeaker to the secondary winding. I leave the phone plug inserted in the LM phone jack at all times. After using this setup for a while, it occurred to me that the last few cycles either side of zero beat might not be audible, so I tried a parallel hook-up off the primary side of the transformer. The other end of the cable was connected to the Vertical input of an oscilloscope. Now, when adjusting the corrector control for zero beat, you will see low amplitude sine waves even after you can no longer hear the beat note. You tune out the sine waves until you have a perfectly straight reference line on the scope. The Heterodyne oscillator is then corrected to calibration much more accurately than trusting to the ear alone. The principle is not new, yet the idea may not have occurred to some people. Any old scope will do, as the frequencies are in the audio range.

Another bit of frustration occurred when trying to adjust the corrector knob. I would pass the zero beat point time and time again, due to the stiffness of the control. Remember, these units are built for ship-board use, and loose controls cannot be tolerated. I dug up a small vernier tuning drive, similar to the Jackson Planetary-Vernier drive, (about 5 to 1 ratio). I mounted this drive on a right-angle aluminum bracket and attached it to the LM using the two screws on the upper right side of the unit. Bore holes slightly oversize to allow for accurate alignment of the control shaft so that it will not bind. Cut out a portion of the bracket so the "High-Low" knob can be moved. The precise Zero beat adjustment can now be made very easily with the reduction drive.

HAND CALIBRATE THE BC-221 FREQUENCY METER

Carl Henry

At present many amateurs have an opportunity that may not repeat itself. I refer to the many army and navy surplus frequency meters now available, through MARS and surplus sales. Many of these meters have the original calibration chart missing, and are considered virtually useless. This is demonstrated graphically by prices, the frequency meters with charts costing $100, the meters without charts selling for one-third this amount. Meters without charts are worth more than this would seem to indicate. It is no great job to calibrate one of these meters, but it is time consuming. A little care is required, but you do not need to be a super technician or electronic engineer to do the job. Neither is elaborate equipment required.

Essential to the calibration are several items not commonly available around the shack, but easy to build. Figure 1 illustrates the first requirement. This is a 10 kc multivibrator, which will operate from a 100 kc xtal calibrator, and give 10 kc markers up to 20 mc. I am assuming here that you have or can beg, borrow, or steal a 100 kc calibrator, this being a common item. In fact, some receivers have them built in. Using this 10 kc marker source, the high band of the frequency meter can be calibrated. Interpolation to 1 kc points is then possible, and if care is used an accuracy of better than 0.005% will result.

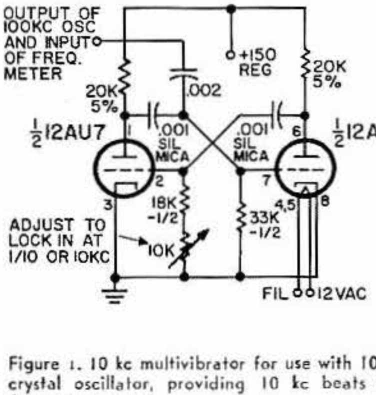

Figure 1. 10 kc multivibrator for use with 100 kc crystal oscillator, providing 10 kc beats with frequency meter.

Fig. 2 shows a simple beat detector that can be used with earphones to get an exact zero beat. It took about ten minutes to wire up a 1629 (army surplus magic eye tube) temporarily for this purpose, and it is certainly worth while in the interest of increased accuracy.

With your equipment assembled, connect the frequency meter to its power supply. Allow at least one hour for proper warm-up so that all equipment will stabilize, including the 100 kc calibrator and 10 kc multivibrator. Set the CALIBRATE control on the frequency meter to center, and be especially careful not to move it until the calibration is complete. After the warm-up period the 100 kc calibrator should be zeroed against WWV, at as high a frequency as possible. The 10 kc multivibrator must be checked to see that it gives nine beats between each 100 kc beat. This can be done with your communications receiver, and it will be easiest to do it on as low a frequency as possible, say 600 to 700 kc. Now type several sheets of paper, listing the high range of your frequency meter by 10 kc points. If you are in doubt as to the range, check it with your receiver. To do this set the dial to the low end of its range, turn on the internal xtal, and zero in on the strongest beat at or near the low end. Turn off the internal xtal oscillator and find the signal from the frequency meter with your receiver. Note the frequency. Tune your receiver until you reach another harmonic from the frequency meter. The difference between the first and second readings is the frequency of the meter at the low end

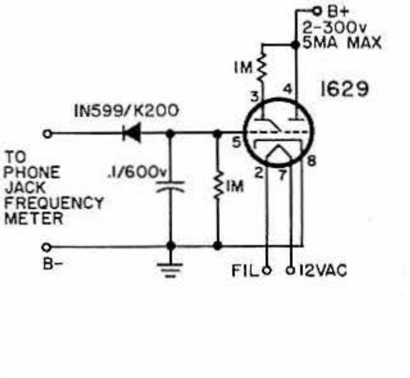

Figure 2. A simple zero-beat indicator for use in conjunction with headphones.

of its dial. Now do the same thing at the high end, and you have the primary coverage of your frequency meter. Most have a range of 125 to 250 kc on the low band and 2 to 4 mc on the high band.

When you write up the frequency range by 10 kc points, leave a space between notations, since you will be adding information here later. Now, using the 100 kc xtal, check the high band of the frequency meter at 100 kc points, listing the dial readings. Now go back and check the dial at 10 kc points, listing all the dial readings. The checks should agree with the 100 kc checks at every tenth point. This is a handy method of cross checking yourself. You might note that on equipment of this type, always approach the final reading from below. If you pass the zero beat point, don't jockey the dial back and forth for zero beat. Go back below the beat by 5 kc or so and approach again from below. This procedure will help to eliminate error from dial backlash.

After your 10 kc points are all listed, interpolate to 1 kc points. This is the hardest part of the job, and is very time consuming. A big pot of coffee and a patient and loving XYL will be a big help here. The patient XYL can be replaced with an adding machine if you have one available, but you'll still need the coffee.

The difference between each 10 kc point must be listed. Note this in the space you left on your sheets. Each 1 kc will be 10% of this, so add 10% to the 10 kc listing for 11 kc, 10% more for 12 kc, 10% more for 13 kc, and so on. When you reach the next listed 10 kc point, the calculated listing and the measured listing must agree, another cross-check. This will take 2000 individual additions.

After the high band is finished, you may wish to calibrate the low band in the same fashion. Referring to Figure 1, change the grid-plate capacitors to 0.01 mfd. This will put the multivibrator at approximately 1 kc intervals. If you have trouble syncing the multivibrator, you may have to build another to operate at 10 kc, and sync the 1 kc from this. No trouble should be had with a strong output xtal calibrator, however. Calibrate the multivibrator as before, except with 1 kc intervals instead of 10 kc.

A calibration book can be prepared when you finish. It is a good idea to file all your original calculations and papers, should the book ever be destroyed. The frequency meter is now as good as any with original calibration book, at a good saving of money.

HIGH ACCURACY VHF FREQUENCY MEASUREMENTS

Howard Burgess W5WGF

Accurate frequency measurements can be a problem for the amateur with a limited budget. However with some home construction and careful operation the average ham can make VHF frequency measurements to an accuracy better than .00015% at two meters. This is the equivalent of measuring the distance from New York to Los Angeles with an error of only 25 feet. Many commercial units cannot equal this figure. The same method can also be used for HF and UHF measurements.

There are many ways to measure frequency but few of them are satisfactory for use at the very high frequencies. The well known hetrodyne frequency meter becomes unstable when its oscillator is operated at VHF. It can no longer be held or read to any degree of accuracy. The oscillator can be operated at a low frequency and one of the harmonics used at VHF, but any error in the oscillator will be multiplied by the number of the harmonic used. A frequency meter that can be held to within 200 hertz at 4 MHz will be off by 7.4 kHz at the 148 MHz harmonic.

A second method of much greater accuracy uses low frequency crystals which are referenced to a known standard such as WWV. The harmonics of these oscillators will be quite accurate and useful far into the UHF region. However this system has its limitations. Even when used with multivibrators and harmonic amplifiers it produces only spot frequencies.

Although neither of these two methods is satisfactory when used alone, they can be combined to make an accurate and versatile system. If you haven't guessed it by now, the system works like this. A crystal oscillator operates on 5 MHz. This oscillator can be kept to zero beat with WWV with very little effort. With a simple harmonic amplifier following it, strong markers are available every 5 MHz far into the UHF region. To fill in between the 5 MHz points, and get full tuneable coverage, all that is required is to add the output of a stable low frequency VFO to the proper marker. Example: To measure 146.25 MHz just add 1.25 MHz from a calibrated tuneable oscillator to the 145 MHz harmonic of the crystal. The same results can be had by using the 150 MHz marker and substracting 3.75 MHz.

The tuneable low frequency oscillator of this hetrodyne system can be any stable, calibrated, oscillator that will give the desired frequencies. A good signal generator can be used but better yet is the old faithful BC-221 frequency meter. The crystal oscillator that supplies the 5 MHz markers should be designed for high stability. However, even simple crystal-controlled units can be kept zero beat with WWV for periods long enough to make most measurements.

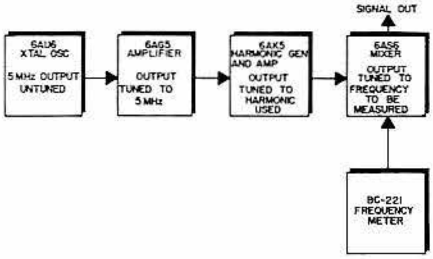
SIGNAL OUT
6AU6
XTAL OSC
5 MHz OUTPUT
UNTUNED
6AG5
AMPLIFIER
OUTPUT
TUNED TO
5 MHz
6AK5
HARMONIC GEN
AND AMP
OUTPUT
TUNED TO
HARMONIC
USED
6AS6
MIXER
OUTPUT
TUNED TO
FREQUENCY
TO BE
MEASURED
BC-221
FREQUENCY
METER

Fig. 1. Block diagram of the VHF frequency meter.

Earlier we quoted a figure of .00015% or better for the accuracy of this system. Perhaps we should show how this is possible. The crystal oscillator can be held to near zero beat with WWV but due to propagation errors in the signal of WWV, we can never be sure that our crystal is closer than 2 parts in 10 million. This would be 2 hertz of error at 10 MHz or an uncertainty of 29 Hz in the 145 MHz marker. The BC-221 is normally considered to be a .05% instrument. This would be an error of about 1.75 kHz at 3.5 MHz. However with care in calibration, and reading it is not difficult to reduce this value to 200 hertz or less. In a hetrodyne system the error of the VFO is not multiplied at VHF but is just added to the error of the crystal marker used.

The total error at 2 meters is 29 Hz contributed by the crystal and 200 Hz by the VFO for a total of 229 Hz. This is a little more than 1.5 hertz per million hertz for a tuneable system. Of course these values are approximate and with careful operation they can be reduced by 50% or more.

In the 146.25 MHz example used earlier, the VFO was required to furnish less than 1% of the total output. To put it another way, the only wobble is in the smallest cog and its contribution is so small it can't shake up the machinery too much.

The circuit shown in **Figs. 1 and 2** has been used for monitoring MARS, CAP, and several other services. The crystal oscillator is quite stable but can be tuned enough to zero with WWV. Tuning is done with C1. One stage of harmonic amplification is sufficient to give strong signals well above 150 MHz. The plate circuit of this amplifier stage is tuned to the harmonic to be used. This feeds one input

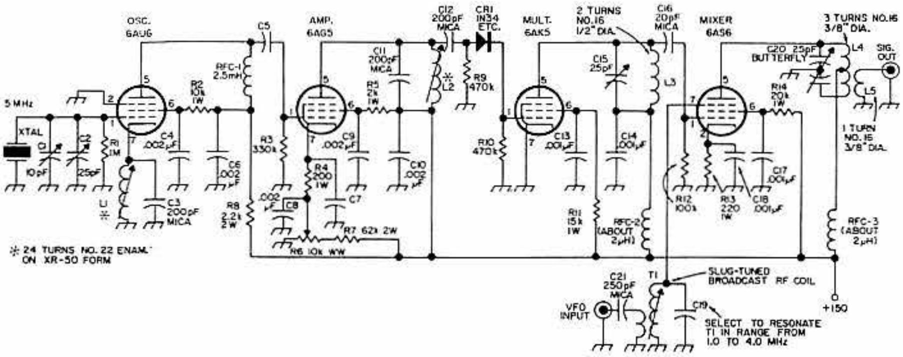
OSC.
6AU6
5 MHz
XTAL
C1
10pF
C2
25pF
R1
1M
L1 ✻
C3
200pF
MICA
C4
.002μF
R2
10k
1W
C6
.002μF
RFC-1
2.5mH
C5
R3
330k
R8
2.2k
2W
AMP.
6AG5
.002μF
C8
C9
.002μF
R4
200
1W
C7
R5
2k
1W
C11
200pF
MICA
C10
.002μF
R7 62k 2W
R6 10k WW
✻ 24 TURNS NO. 22 ENAM. ON XR-50 FORM
C12
200pF
MICA
✻ L2
CR1
1N34
ETC.
R9
470k
MULT.
6AK5
R10
470k
C13
.001μF
R11
15k
1W
2 TURNS
NO.16
1/2" DIA.
C15
25pF
C14
.001μF
L3
RFC-2
(ABOUT 2μH)
C16
20pF
MICA
MIXER
6AS6
R12
100k
R13
220
1W
C18
.001μF
C17
.001μF
R14
20k
1W
C20 25pF
BUTTERFLY
3 TURNS NO.16
3/8" DIA.
L4
SIG.
OUT
L5
1 TURN
NO. 16
3/8" DIA.
RFC-3
(ABOUT 2μH)
+150
SLUG-TUNED
BROADCAST RF COIL
T1
C21
250pF
MICA
VFO
INPUT
C19
SELECT TO RESONATE
T1 IN RANGE FROM
1.0 TO 4.0 MHz

Fig. 2. Schematic of the oscillator, multipliers and mixer for using a BC-221 on the VHF ham bands.

grid of the mixer. The other grid of the mixer is driven by the output of the BC-221 frequency meter. The tuned circuit shown in this grid resonates broadly in the 2-4 MHz range of the BC-221. This helps to keep the higher harmonics of the BC-221 out of the mixer.

The output of the mixer is resonated to the desired operating frequency. This will be either the sum or difference of the two input signals. The level of the output signal can be controlled by R6.

Operation of this system is simple. The "cook book" would read as follows:

1. Couple the output to the antenna of the VHF receiver.
2. Determine the crystal harmonic and VFO frequency that will give the frequency of the signal to be checked.
3. Tune the VHF receiver to the signal to be checked.
4. Tune the BC-221 until the output of the frequency monitor zero beats the received signal.
5. If required, peak the tuned circuits in the monitor for maximum output and adjust R6 as needed.
6. The frequency of the received signal will be the crystal harmonic plus (or minus) the reading of the BC-221.

Many details cannot be covered in one story due to lack of space. The operator will have to determine the most effeceive method of coupling to this particular receiver. He will also have to explore the many combinations of frequencies which can be used. These and many other questions cannot be included at this time. However those who require such a system as this will probably be capable of filling these details.

One word of caution is in order. With two oscillators that are rich in harmonics, there can be many unwanted "birdies." These present no problem after the operator has gained experience but the new user should be very cautious. Many times an unwanted beat can be eliminated at a critical spot by changing the two frequencies that are being mixed (shift from sum to difference).

Perhaps we should emphasize that this system is a "trade off" where the amateur can trade his skill and patience for highly accurate measurements with simple equipment.

A TRANSISTORIZED LM METER

Charles Landahl W5SOT

The last word may never be written about the BC-221 and LM frequency meters. The LM is particularly attractive because it is in the smaller package. With transistors replacing tubes, it has features most everyone wants – it is rugged, portable, and accurate, to name a few. I will describe a conversion of an LM-15 frequency meter in which field-effect transistors replace tubes; the power supply becomes a standard 9V transistor radio battery and the current drain is less than 3 mA when all functions are energized. In addition, I offer calibration information which will be of interest to anyone having a BC-221 or LM without the official calibration book. I bought an LM-15 for a temptingly low price (Fair Radio Sales Co., Lima, Ohio, $14.95). The set is sold in the "as is" condition with tubes and crystal but without calibration book. It is a good idea, but not necessary, to start with a set which is working before making the change to FETs. Resistance measurements will show if the circuits are complete. Important values are marked on the schematic of Fig. 1.

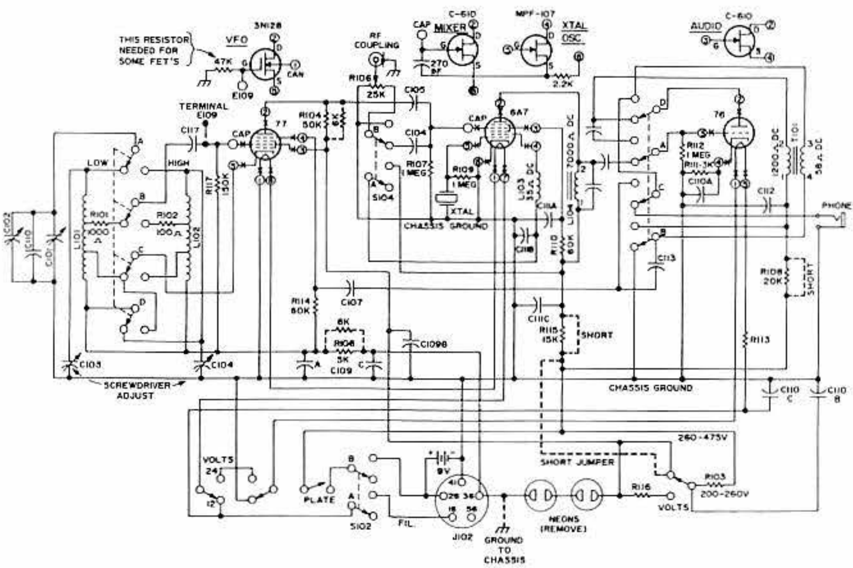
THIS RESISTOR NEEDED FOR SOME FET'S
VFO
3N128
47K
G
D
S
CAN
E109
TERMINAL E109
C117
CAP
77
R104 50K
RF COUPLING
R106
25K
C105
C104
R107 1 MEG
S104
R109 1 MEG
XTAL
CHASSIS GROUND
C-610
CAP
MIXER
270 pF
MPF-107
XTAL OSC
2.2K
6A7
L103 35Ω DC
L104 7000Ω DC
C111A
C111B
R110 60K
C-610
AUDIO
76
R112 1 MEG
R111-3K
C110A
C112
T101
1200Ω DC
58Ω DC
PHONES
LOW
HIGH
R101 1000Ω
R102 100Ω
L101
L102
R117 150K
C102
C115
C101
C103
SCREWDRIVER ADJUST
C104
R114 60K
C107
6K
R108
5K
C109
C109B
C111C
R115 15K
SHORT
C113
R108 20K
SHORT
R113
CHASSIS GROUND
C110 C
C110 B
260-475V
SHORT JUMPER
R103
200-260V
R116
VOLTS
VOLTS
24
12
PLATE
S102
FIL.
9V
J102
GROUND TO CHASSIS
NEONS (REMOVE)

Fig. 1. Modified schematic of frequency meter.

Smash Tubes

The most difficult part of my conversion was getting up the courage to smash the tubes! I wanted the bases for mounting transistors. Place the tubes one at a time in a paper sack, hold the top closed and with a metal object, strike the glass through the paper. The flying glass is caught and collected for disposal. Scrape and clean the mastic from the inside of the tube bases. Should you choose to mount a transistor socket in the wall of each tube socket, you can use the original wires; otherwise, unsolder the old wires and replace the needed ones with about 2 in. of sturdy new tinned wires. The appropriate tube base pin connections are shown in Fig. 2. Actually there is no preferred mounting scheme. Use whatever appeals to you.

Check for clearance between socket and walls. My conversion used transistor sockets mounted on metal plates which were bolted to the wall of the salvaged tube bases. This allowed FET substitution to determine which ones would work best in the several circuits of the LM. All FETs used are N-channel.

Modification

With cover removed and the LM in the upright position, front panel toward you, on the left side wall, look through two oblong machined slots and see mounted on a phenolic board a 50 kΩ plate resistor. Parallel it with about 6 kΩ. Turn the LM upside down, panel toward you. On the underside, two resistors must be shorted and a jumper wire made up and connected. Short R115, which is a 15 kΩ wirewound resistor, quite visible on a phenolic board at the left of the 1000 kHz crystal can. Run an insulated wire from a terminal of this shorted resistor to the 260–470V tap contact of the link switch. This wire can be about 6 in. long and conveniently passes through a wall slot behind the crystal socket. The link switch and its terminals are on a phenolic board in the compartment aft of the crystal socket. The jumper wire will cross near the grid resistor, R109, of the crystal oscillator. While there, change the 100 kΩ (R109) to 1 MΩ. Next, unfasten the screws holding the phenolic board located to the left of the power plug. Tip up the board and short across R108. This is a 20 kΩ composition resistor which is in the plate voltage line to the audio amplifier. Also, at the power plug, locate pin 36. Short it to chassis ground. On most sets pin 36 is the ground return for the vfo cathode. The circuit was closed through external connections in a power supply. You have completed the surprisingly few changes needed to make the LM work on FETs and a 9V battery.

The VFO

With FET source connected to pin 5, drain to pin 2, can to pin 1 (if needed), plug the FET into the vfo socket. Connect a solid wire between terminal E109 and the gate of the FET. (Terminal E109 held the grid cap wire for the vfo tube.) Connect a 9V battery to the power plug pins. PLUS to 26 and MINUS to terminal 41. If you have a milliammeter in the battery lead, it should read about 1.5 mA when you turn on the FIL and PLATE switches. Provided you were fortunate in the choice of FET, you should hear a clear CW signal in your receiver. Set your receiver to 2 MHz or 4 MHz. You may need to connect a wire from the rf coupling post on the front of the LM to your receiver antenna. Rotate the LM dial between 0300 and 0600 on the readout. Your vfo will be on the low end of 125–250 kHz or 2–4 MHz depending on the position of the low or high band switch. The XTAL and MOD switches should be off. The FET selected for the vfo may require a 47 kΩ resistor between gate and chassis ground. I found this to be true for the RCA 3N128, 3N142,and *one* of the two 40559A FETs. On the other hand, one RCA 40559A and one of several 2N3085 silicon N-channel FET from Poly-Paks worked beautifully without adding 47 kΩ to the gate.

Apparently junction and insulated-gate field-effect transistors have slightly different characteristics which show up in this peculiar vfo circuit. My own choice is the 3N128 with the additional resistor on the gate. You may find it necessary to "tune" the source, drain, and gate resistances in order to have your vfo working well with a particular FET. I used potentiometers across the various elements to arrive at the recommended values. My vfo works reliably from 10V down to 6V and the maximum drain current is 1.5 mA.

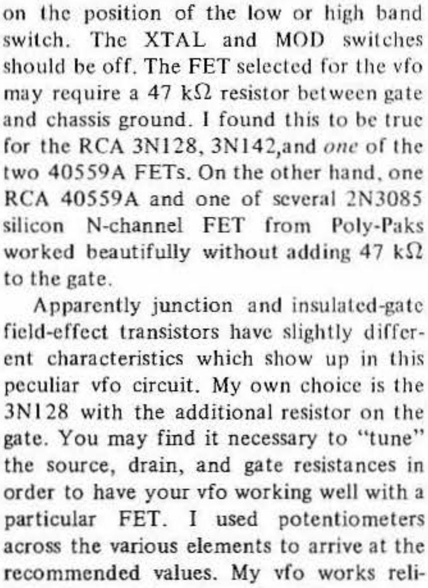

Fig. 2. Suggested semiconductor hookups for tube sockets.

Caution

Other oscillator hook-ups may occur to you, and they will work – but the tuning range and linearity of the vfo will suffer! Linear tuning is most important, so stick with that shown.

The AFO and AF Amplifier

The audio oscillator and amplifier wasn't as fussy as the vfo. I used a Radio Shack 276-664 FET – it is said to replace a C-610 or 2N3088. I found that the Poly-Paks "hobby" FET 92CU588 will work equally well. In making up the socket, gate goes to tube pin 3, source to cathode pin 4, and drain to plate pin 2. That is all there is to this one. Plug in the FET. When you next turn on the 9V power, the milliammeter will barely increase a few hundred microamps as you switch on the modulation control. At this moment a rather pleasant 500 Hz tone will appear on the vfo frequency no matter which harmonic you have tuned in on receiver. Your modulator is finished. The audio amplifier is too, for that matter. You just won't hear anything in the headphones until you complete the crystal oscillator and the mixer circuits.

Crystal Oscillator

The reference oscillator is not much trouble. You have already changed the gate resistor from 100 kΩ to 1 MΩ. Actually this change may not be necessary because some crystals are more active and will oscillate well with the original resistor. Mine went into oscillation better with the higher value. The FET you select for this circuit can be one of several. Mine is a Motorola MPF-107. I found the Radio Shack 276-112 and the Poly-Paks 2N3085 also work, but draw more current. Whichever you choose, the gate connects to base grid pin 5, source to cathode pin 6 *through* a 2.2 kΩ resistor. Drain hooks to plate pin 4.

Now, when 9V is turned on, MOD off, XTAL on, you should hear the crystal oscillator signal every 1000 kHz on your receiver. The milliammeter should increase about 1.5 mA or less when XTAL is turned on. If you don't hear the crystal frequency, bring the receiver antenna wire close to the crystal FET. We still haven't made the connection which adds the crystal-oscillator signal to the rf coupling post on the front of the LM. Assuming you have all circuits in working order up to this point, we move to the mixer.

Mixer

There is no single FET substitute for a pentagrid converter tube. The dual-gate MOSFET comes closest; however, use of one would have defeated my goal of simplest conversion. Therefore, four N-channel FETs are needed to do the work of three tubes, but what a saving in power supply! The mixer concerns itself only with beat frequencies occurring between the reference oscillator, vfo, or an external signal – all audio work. Thus, a hobby FET was selected. I used the Poly-Paks N-channel FET. A Radio Shack C-610 replacement will also work. Connect source to chassis ground through pin 6 of the tube socket; the drain connects to plate pin 2, gate to mixer grid cap wire through the lug in tube socket wall.

Finally, connect a capacitor (200–300 pF) from the top of the 2.2 kΩ crystal oscillator source resistor to the gate of the mixer FET (grid cap wire). You are in business.

With 600Ω phones plugged into the LM, you should hear all the necessary beat frequencies occurring between the vfo and the crystal oscillator as you tune the vfo through its range. XTAL must be on and the MOD switch off. Otherwise, the audio amplifier becomes the modulator and you hear nothing in the LM phones.

What Next?

With the beat notes loud and clear you are ready to calibrate. This is quite the most fun part of the work because the linear tuning rate of the LM is almost unbelievable. The slow rate is due to the series combination of the A section of C109, C101, tuning L101 or L102. The amount of the matter is that one revolution of the 100-division circular dial produces about 3 kHz change on 125–250 kHz range, and about 50 kHz per revolution on the 2–4 MHz range. The actual calibration of my unit was 2.89 kHz and 45.17 kHz per revolution.

The linearity can be checked by how little you need to vary the "corrector" for each zero-beat checkpoint. Each LM or BC-221 will be slightly different. Now, when you consider that the vernier allows you to split one division into tenths, then it is clear that you can set a frequency to better than 0.5 kHz over the range of the frequency meter. May I repeat:

1 dial revolution of
100 div = 45.17 kHz
1 division = 0.4517 kHz
1/10 div = 45 Hz

Therefore, all you need is a checkpoint at which to zero the vfo and start counting revolutions, divisions and tenths of divisions to accurately set any frequency within the two ranges of the vfo. I found it useful to construct graphs on K&E 358 11L graph paper. The grid is 10 X 10 (per 0.5 in.). The paper has 20 units vertical and 30 horizontal. This allows graphing 100 division and leaves room for 10 vernier divisions on the right hand end of the paper. Use the crystal checkpoints listed in Table I to locate *your* dial settings. Once graphed, a frequency can be selected directly from the chart, or, depending on the accuracy desired, interpolated between checkpoints.

Table I. Crystal Checkpoints

KHz	VFO	XTAL	Approximate Dial Settings
Low Band			
125	8	1	0320
150	20	3	1192
166.667	6	1	1750
200	5	1	2935
222.222	9	2	3700
250	4	1	4647
High Band			
2000	1	2	0396
2250	4	9	0945
2500	2	5	1493
2750	4	11	1050
3000	1	3	2606
3250	4	13	3150
3500	2	7	3711
3750	4	15	4262
4000	1	4	4812

It is obvious from this discussion that the low band of the LM is fabulous. You can squeeze down to about 3 Hz by use of the vernier scale. By the way, hidden behind two cover plates just beneath the corrector knob, are "high" and "low" padder capacitors. These were used when vfo tubes were replaced to bring calibration book values into usefulness. The padders should be set near the middle of their range.

Make it Handy!

Fasten a handle to the case, strap on a 9V battery, go forth and have fun with your rejuvenated frequency meter. I use mine for its intended purposes as well as a band-edge marker and keying monitor.

AN I-F SPOTTER

Howard Burgess W5WGF

If you have ever tried to find the *if* frequencies of unfamiliar and inoperative pieces of surplus gear with no schematic, it is a waste of time to tell you how rough it can be. Even a single *if* transformer from the junk box can be a problem if it has no part number or identification.

Of course in some cases a grid dip meter can be used to find the operating frequency. However, few grid dippers cover the important *if* frequencies below 2 mc. To complicate things, if a dipper is used on a shielded transformer above 2 mc, the resonant frequency of the transformer may be shifted if the shield is removed.

If these problems sound familiar to you, we would like to suggest a little gadget that can help solve them. With just two resistors, two capacitors, and a tube, don't expect it to give a digital read-out to all your questions, but it can put you in the ball park.

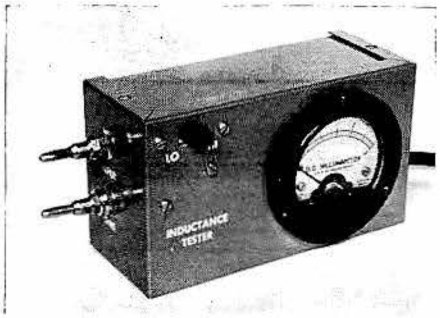

The if spotter. The switch for changing the coupling capacitor is at the upper left marked "HI" and "LO." The posts on the end are X1 and X2.

The principle of operation is as simple as the construction. The tuned circuit in question is merely made to oscillate at its resonate frequency. The frequency can then be determined by tuning in its radiated signal on the ham receiver. To set the unknown coil into oscillation requires the use of a simple "two terminal" oscillator. Such an oscillator is shown in the schematic of Fig. 1. When any tuned circuit is connected to the two points marked X, the circuit will oscillate at its resonant frequency.

In this oscillator the twin triode is a tube such as the 12AT7. The section V2 furnishes the necessary feedback and eliminates the need for extra coils or feedback connections.

The construction is simple. The unit could have been built in a larger case with its own power supply and would have become a nice piece of bench equipment. However, due to the few parts required and the small amount of plate power used (3 mils at 90 volts) it was built as an overgrown probe. The power is robbed from another piece of test equipment or the receiver.

As a probe it can be used on the work bench to test individual coils and transformers or it can be held in contact with the various transformers in a receiver.

There is only one point of caution that should be observed in construction. The lead from the grid contact of VI to the XI post should be kept as short and direct as possible with the least capacity to ground. This lead becames part of the oscillating tuned circuit and limits the upper frequency to which the unit will operate.

The coupling capacitor from the grid of V1 to the plate of V2 furnishes the feedback required to maintain oscillation. To reduce the loading on the tuned circuit, this capacitor should be held to the smallest value that will sustain oscillation. Because of the wide range over which this instrument operates, a switch is provided to change values. With the capacitors shown, operation is possible from about 60 mc down to well below 50 kc. The larger value is used at the lower frequencies and is switched in only when required. With coils of medium Q the switching point is around 3 to 5 mc.

The connectors X1 and X2 can be almost any kind of posts. The ones shown on the unit here are banana plugs. These can be used as test points, or alligator clips can be slid over them for use in clipping to coil leads.

The meter shown is a three mil meter and is used to read the total plate current of both halves of the tube. This will indicate when the circuit is oscillating. When the tester is not oscillating the meter will indicate a current of about 1 mil (with 90 plate volts). Under oscillating conditions the meter will rise to as much as 3 mils with a high Q coil. The actual amount of current is not so important as the fact that the upward shift indicates that the coil is not open and is oscillating. A 5 or 10 mil meter will serve the purpose just as well.

As the pictures show, this unit was built in a small 5" x 3" x 2" box. The unit will work just as well if it is built on a small piece of peg board with a couple of leads run out to the multimeter. This is for the man in a hurry.

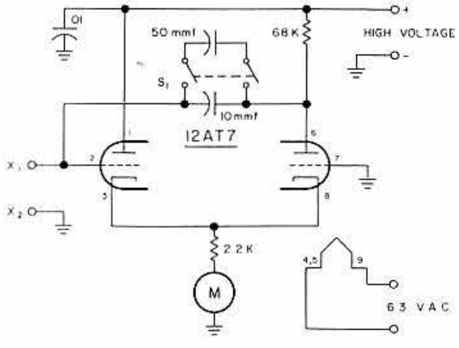

FIGURE 1

Circuit diagram of the if spotter

.01 Ceramic capacitor
50 pf Mica capacitor
10 pf Ceramic
68 K ½ watt carbon resistor
22 K ½ watt carbon resistor
Milliammeter in range of 3 to 10 ma
Slide action Double Pole-Single Throw switch (S1)
2 Banana plugs (x1 and x2)
12AT7 tube
Minibox (size depends on type of meter used)
9 pin miniature tube socket

When the unit is finished, apply power with the "X" points open (no coil across them). Because of the open grid of V1, the meter will drift about. After the tube has had time to warm up, short the X posts with a heavy piece of bus or copper. The meter will now come to rest somewhere around 1 mil. This is the "no oscillation" current and should be kept in mind as a reference point for future use. Now remove the short and connect almost any kind of an LC circuit across the posts. The meter reading will now rise from the "no oscillation" value indicating that the coil is oscillating. Do not use the large coupling capacitor unless the circuit refuses to oscillate with the smaller value.

To check a single *if* transformer, all that is necessary is to hook one of the *tuned* coils

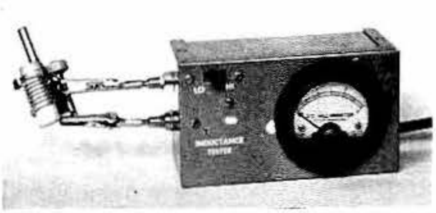

Checking a ten meter tank circuit

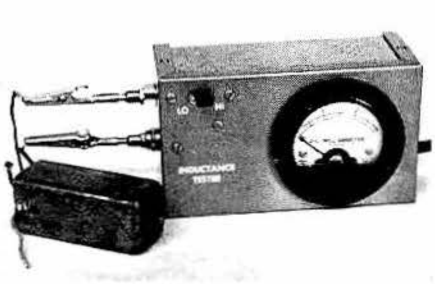

Checking an if transformer

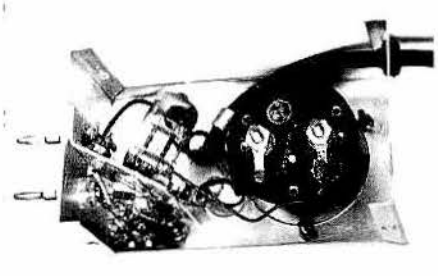

Inside view showing extreme simplicity of the "Spotter." The tube is mounted at an angle to keep the grid lead to the "unknown" as short as possible.

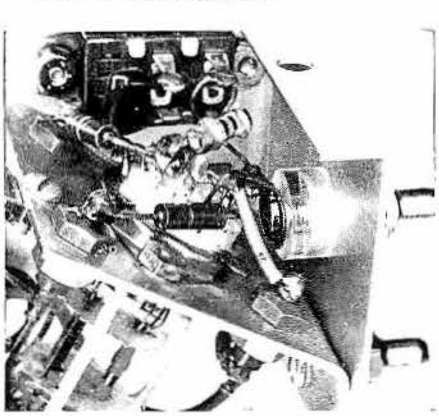

Close up of tube socket showing parts placement.

to the input terminals of the tester. Some transformers have a number of terminals which may not go directly to the tuned circuit inside. To obtain oscillation there must be both a dc and an rf path between the two pots.

If the transformer is one whose frequency falls in the range below the broadcast band it is quite convenient to have one of the surplus receivers that covers the range down to 200 kc. However the check can still be made with a regular broadcast receiver. All that is necessary is to find the harmonics of the tester as they fall in the broadcast band. They will be separated by a value equal to the frequency of the coil being tested. As an example, if a signal is spotted at 900 kc and the next one higher is found at 135 kc, it is a pretty good bet that the transformer is operating on 450 kc ($1350 - 900 = 450$).

To find the operating frequency of an *if* stage it is not necessary to have the amplifier in operating condition or the tubes hot. Just connect the two contacts across the primary or secondary of the transformer in question and watch for signs of oscillation on the meter. Some transformers have a portion of the bias system inside of the can. This can usually be overcome by connecting the tester from grid to ground of the tube in the stage being tested.

In addition to checking transformers it can also be useful in testing the range over which a transmitter tank will tune. Just make sure that the high voltage is turned off and connect the probe across the tank to be tested. Now you can tune the tank and follow its entire usable range with the receiver. If the tank being checked happens to be the final, the meter on the probe will indicate when the antenna is brought into resonance.

This little tester was built to do just one thing—sort out some old *ifs* in the junk box. After we tried it we found that it would do a lot of useful chores around the ham shack. With proper care and feeding, it can probably learn to do tricks that we haven't even thought of.

430-470 kHz SWEEP FREQUENCY GENERATOR

Edward Lawrence WA5SWD

Here is a simple sweep frequency generator for aligning the most common *if* strips. The unit has only one transistor because the sweep voltage is taken from the oscilloscope it is used with. Most general purpose oscilloscopes have a sawtooth output jack. By using this sawtooth, the frequency is locked to the position of the trace.

The oscillator is made to deviate in frequency in step with the voltage applied to the base bias circuit, either the sweep voltage or the dc voltage applied at the control point. If no sweep or control voltage is applied, the oscillator runs at the center frequency, and may be used as a conventional signal generator.

The 2.5 mh choke and the two 150 pf capacitors form a very broad tank circuit, so it is easy to FM the oscillator without a great change in amplitude. As to how the change in base voltage causes the frequency shift, I am not quite sure. All I know is that it is quite linear and is a positive shift (**Fig. 3**). That is, an increase in base voltage causes an increase in frequency. Also, a change in collector supply voltage will shift the frequency, so be sure to use a stable supply.

Since an rfc is used as part of the tank, you may have to compensate for a variation in center frequency by changing the 150 pf capacitors to some other value to get 455 khz as your center value.

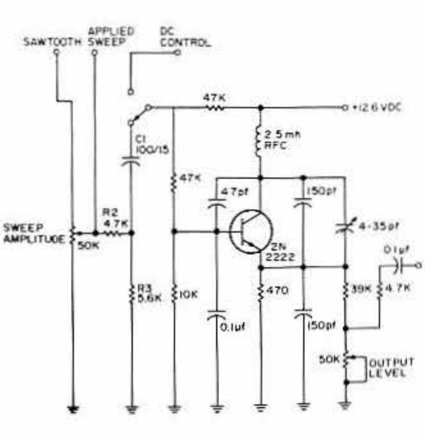

Fig.1. Sweep generator circuit diagram.

If you do, then run a plot of dc control voltage vs frequency and adjust the voltage divider R2 and R3 so that 1 volt P-P at the sweep pot wiper gives plus and minus 5 khz (10 khz total). This is not as hard as it sounds. Just listen for the harmonics on a broadcast band radio. For instance, the second harmonic of 430 khz is 860 khz, and the second harmonic of 470 khz is 940 khz, both handy on the BC radio. Note the dc voltages at the control point required to obtain these frequencies. Take the difference and divide by four. The answer is the P-P voltage required to shift the oscillator plus and minus 5 khz. Then apply a low, known ac voltage to the wiper of the sweep pot with the switch in the *sweep applied* position. Adjust the value of the divider resistors to get the proper fraction. In my case it was 0.43.

It is advisable to run the sweep rate as slowly as possible, in order to display the response curve as accurately as possible. The sharper the skirts, the more slowly the generator must pass through the bandpass. With this generator, the amount of frequency deviation is controlled by the P-P amplitude at the sweep pot arm. If you are looking at the response of a regular *if*, you would set the sweep amplitude high to see the entire response curve. As far as an accurate display is concerned, this is fine, since the slope of the skirts is shallow. But if you were looking at the response of a sharp filter, you could not tolerate such a wide sweep, because the fast rate-of-change would tend to skew the display. To correct this, reduce the sweep amplitude to reduce the frequency deviation down to the edges of the skirts of the response curve. This reduces the rate of change and minimizes the skewing of the display. Also, it is better to display the *if* before detection, if possible, to prevent the detector time constant from possibly distorting the display.

If you have the sweep generator set up as described, then it is easy to set the sweep for a known deviation, and then procede to read the 3 or 6 db points from the face of the scope. Be sure to disable the AGC for this test.

For those who aren't familiar with the setup for obtaining the response curve display, refer to **Fig. 2.** and the following outline.

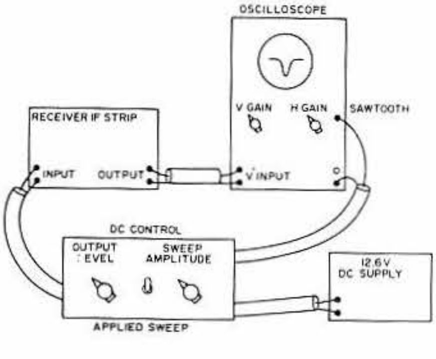

Fig.2. Interconnections of sweep generator.

First, hook up the equipment as in Fig. 1 Set *horz gain* for desired sweep width. Adjust *sweep amplitude* pot on generator for desired frequency range. Set *output level* to mid-range. Adjust the *vert gain* for the desired pattern height.

If we wish to change the total frequency deviation and the horizontal display width at the same time, use the scope's *horz gain* control. This presumes that the amplitude of the sawtooth output is also varied by the *horz gain* control.

If we wish to change the frequency spread and not change the width of the scope display, adjust the *sweep amplitude* control. This allows you to take a better look at the sidelobes or any ripples in the passband, depending on how the controls are adjusted.

You don't have to use a sawtooth to sweep with if it isn't handy.

Sixty hz can also be used, but it will probably skew the passband you are trying to display, so I don't recommend it. One note here: when the sweep generator is first turned on, or the setting of the sweep amplitude is changed, the frequency will drift for a few seconds. This is due to the charge on C1 changing to a new level. C1 is large to couple the low frequency sweep with as little distortion as possible.

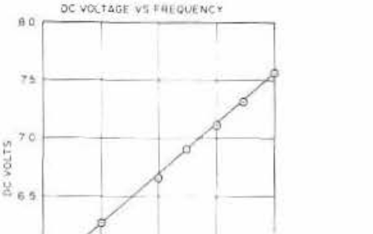

Fig.3. Sweep frequency vs dc voltage.

DC CONTROL VOLTS	FREQUENCY
5.86	430
6.27	440
6.69	450
6.90	455
7.12	460
7.33	465
7.55	470

AN ARMSTRONG SWEEPER

Al Donkin W2EMF

Sometime ago while experimenting with SSB crystal filters at 480 kc, it occurred to me that a manually swept Armstrong oscillator with a pot coupled to the tuning capacitor for 'scope sweep voltage would save point by point frequency measurements with a BC221 while adjusting the filter for minimum in-band ripple. A circuit was hurriedly developed, and the results were even better than expected. Simple though it may be, I'm sure anyone who is working with crystal filters will appreciate its usefulness. The effects of tuning and "diddling" were immediately visible in the in-band ripple and the skirts down to about 20 db could be observed with a few twists of the knob.

With this encouraging experience, a sturdier model was constructed. The circuit was assembled on a small piece of Vector-board, with the tuning capacitor and batteries mounted in a 5 × 7 × 3" aluminum chassis. An aluminum bracket was bent to hold the "sweep" pot and mounted behind the tuning capacitor, spaced to accomodate the mechanical shaft coupling. The Vectorboard was also mounted on the bracket.

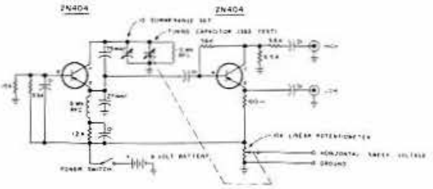

The oscillator circuit is simple, using a 1 mh rfc for the tank inductor. The range of the circuit as shown is about 400 kc to 500 kc, at center frequency. The tuning capacitor is a Hammerlund MC-50-M with two of the rotor plates removed, resulting in a range of approximately 7-25 mmfd. An untuned amplifier provides isolation of the oscillator from the reactance of the load. I used 2N404 transistors but most PNP *if* transistors will work.

In addition to sweeping crystal filters, I have found this little box handy as a general purpose low frequency oscillator and sweeper for receiver *if's*. The persistency of most oscilloscopes produces a much better display than might be expected, and it certainly beats the point by point method of plotting filter responses.

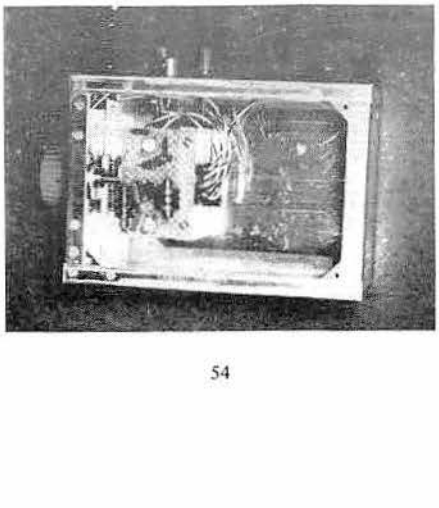

A NEW BROOM

Dick Gridley K6JHJ

This unit is a sweep generator of simple design, with many uses. Besides being useful for filter alignment, as shown in the diagram, it is useful for optimizing receiver and transmitter performance.

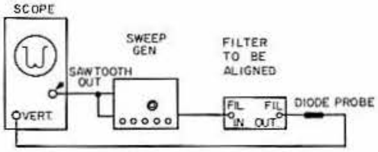

The heart of the rig is an inexpensive item called the Vari-Cap by International Rectifier Corp. Two are discussed here, the 6.8SC20 and the 100SC2. The 6.8SC20, according to the manufacturer, will vary its capacitance from 50 mmfd at .lv to 2.5 mmfd at -200v. The 100SC2 will vary from 600 mmfd at .lv to 100 mmfd at -10v. These units each cost less than $3.00. The construction is not critical in any way, nor is layout a problem—just use good ham practices. Any power supply delivering 150 v dc at 60 ma regulated will do. The oscillator tank must be grounded and I did have some trouble until I removed the unused winding from the 455 *if* transformer used for the oscillator tank.

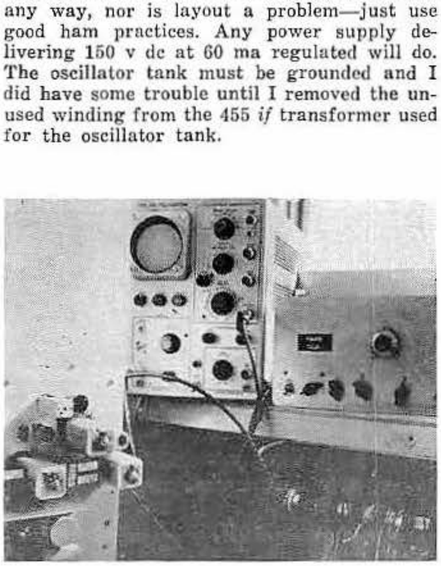

Sawtooth energy is applied to R1 which regulates the amount to be amplified by V1. R2 linear control merely puts a portion of regulated voltage on the Vari-Cap and thus sets the capacity of Vari-Cap close to its center, so the sawtooth pulse creates a linear frequency swing. This circuit gives about 6 kc sweep. I have used TV horizontal coils for L1 and increased C1 and C2 to .003 for 85 kc and 50 kc operation. However, to get enough swing at low frequency, the Vari-Cap was changed to a 100 SC2. For 200 kc to 10 mc the 6.8SC20 gives plenty of sweep.

A note here about scopes. Most of the average do not go below 15 cycles and the sweep rate should be about 10 cycles. Some do not have a sawtooth output jack. The sawtooth output can be picked off the sync circuit. I found the best point is on the feed side of the sync pot through a .25 mfd capacitor. Some scopes have a couple of jack pins to add capacity externally, some don't. You will find little trouble adding about .25 mfd across the slowest speed.

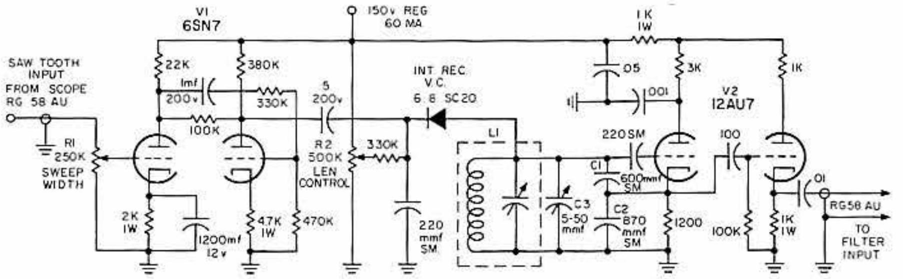

L1 STANDARD I F TRANSFORMER WITH ONE WINDING REMOVED.
SM-INDICATES SILVER MICA.

The use of the rig is simple. Plug the lead from R1 into the scope's sawtooth output. Couple the output from the cathode follower section of V2 to the filter input through a .001 mfd capacitor. The vertical input of the scope is connected to the filter output through an rf probe. Adjust R1 and R2 about ½ open. With C3 about ½ meshed, adjust the trimmer in the *if* can to put you in the ball park, and then use C3 for fine adjustments of frequency. Additional minor adjustments of R1 and R2 may be necessary. Of course, when used to align a receiver, the diode probe is unnecessary if the input to the scope is picked off of the diode detector. It goes without saying that it is necessary to plot an index for the scope bezel so that a ready reference is available for band pass width and depth of skirt.

Some one now says "This is just like setting ye old TV—*if* band pass." That it is. The only difference here is the frequency and the sweep generator cost was junk and $5.00.

The unused switch, shown in the photo, was installed to select one of three frequencies. However, this rig has been used on about six. I don't think I ever will decide which three to use. The unidentified tube is a 6U8 used as a 1 kc audio oscillator with cathode follower output.

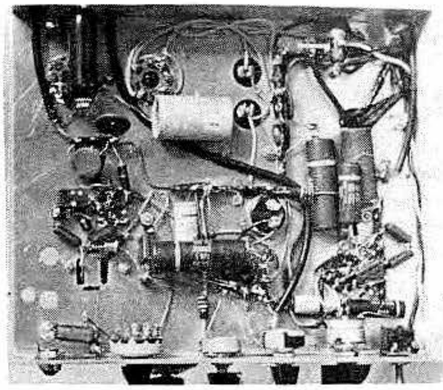

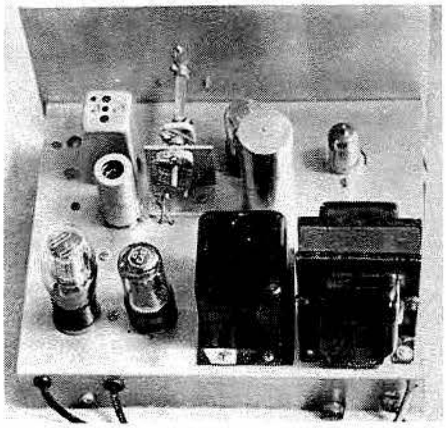

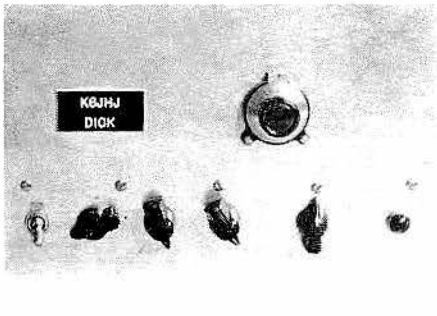

LOW COST SIGNAL SOURCE

Howard White VE3GFW

For many years the Measurements Corporation Model 80 signalgenerator has been the industry standard for tuning rigs. Many hams have wanted to have a laboratory-grade signal source of this type but the $800 price tag is prohibitive. This article describes a signal source that has many of the same features as the 80 except that it only costs 1% as much. That's why I call this the Model 0.8.

You wonder at some of the features of the marvelous little device? To briefly list them, it has:

- Variable output from about 80 nV to 50 mV of rf power.
- Frequency range from 1.8 to 450 MHz so you can cover 160 through ¾ meters.
- Crystal-control frequency stability.
- Fairly clean output signal.
- A 51Ω antenna load.
- Safety feature to prevent the destruction of the device in case a transmitter is accidentally loaded into it.

The Circuit

The circuit is the combined brainchild of many ham engineers and technicians in our local 2m FM club – the Toronto FM Communications Association. More than 300 of these units have been built in the past two years, so the circuit has been exceptionally well tested in the field.

The unit consists of a single 2N5130 NPN transistor oscillator configuration. Other transistors, such as the 2N706, can be used in this circuit with slightly different biasing arrangements; however, the 2N5130 appeared to give the cleanest output signal.

Output level can be varied from less than 80 nV (the best we could measure) to 50 mV. It is controlled by a 1 kΩ pot in the transistor collector circuit. Power is obtained from a 1.5V penlight cell bypassed for rf by a 0.1 μF capacitor.

My present "0.8" signal source has been used intermittently for more than a year without any degradation of output. However, remember to shut the unit off when you're finished with it; the battery doesn't last forever.

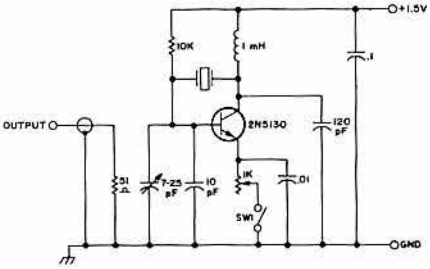

Fig. 1. Schematic diagram of the "Point Eight" Signal Source, the "poor man's Model 80."

The unit is built on a PC board. The parts layout is almost exactly the same as that of the schematic shown in Fig. 1. The electronic components are placed on the copper side of the board. The on-off switch-level control, frequency adjustment, output jack, and crystal socket are on the other side (Fig. 2). The unit can be packaged in a minibox.

Almost any crystal in the 1.8–12 MHz will oscillate in this unit. The output is rich in harmonics to 450 MHz, so you can tune up any receiver from 160 to ¾ meters.

The output impedance of the signal source is 51Ω, a value which simulates a perfect antenna load to the receiver. This is the manner in which this is accomplished: A 51Ω, ½W resistor is connected across the terminals of an RCA phono jack. No physical connection is made to the circuit except through the common

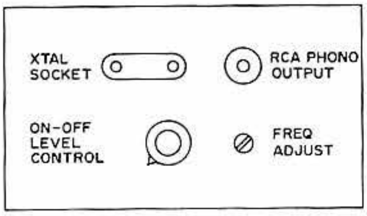

Fig. 2. Minibox layout of the signal source panel.

ground of the PC board. The maximum output level depends on the lead lengths of the resistor. (The unit has plenty of output when the lengths are about ½ in. long.) This design has an added safety feature, too; if by accident you load a transmitter into the unit, all that is destroyed is the ½W resistor.

Operation and Uses

These units were originally designed to tune up 2m FM receivers. The procedure is quite simple. You connect the signal source to the antenna input on your receiver. Plug in a transmit crystal, and adjust the frequency control for a zero reading on the discriminator. Adjust the output level to the desired signal strength (below first limiter saturation) and tune up the receiver.

There are a myriad of other uses for the signal source. Using a 3.5 MHz crystal, you have a band edge marker. With the transmit crystals on a Twoer or any other transmitter you have instant frequency spotting without modifying the circuit of the transmitter. Of course, the signal source can be used to tune up any receiver, peak tuned circuits, be an rf source for an antenna noise bridge, and so on.

RECEIVER TWEAKER

Malcolm Oakes K6UAW

I've found that the singlemost piece of needed test equipment by the amateur on FM is a receiver alignment generator. Most of us, however, do not have access to a signal generator. (Come on now, you wouldn't really call that *TV thing* a signal generator would you?)

The unit to be described is a very functional device that will allow you to scrape every ounce of sensitivity from your receiver. The generator can be built so small you can carry it around in your shirt pocket. Compactness, combined with its battery-powered portability make it ideal for servicing mobile receivers. Stability? It's crystal controlled and is as good as the rock you plug in (you obtain the rock from your transmitter). Output reactance? Nearly zero degrees, allows proper tuning of rf stage. The attenuator shown does not have a great deal of dynamic range due to distributed capacity of the potentiometer. However, I said before, the unit was "functional" and it is just that. The signal can be attenuated into the noise and brought up to approximately a 30 uv level; overly sufficient for a normal alignment.

Construction

The circuit layout is not particularly critical, and if laid out in a manner similar to the schematic, no problems should be encountered. The attenuator section should be completely shielded away from the oscillator and multiplier stages, so rf leakage will not be a problem. A crystal socket should be provided so the unit is versatile for any frequency. However, if you plan on using it on only one channel and can spare a crystal, build it with the crystal inside. Two glass piston screwdriver adjusted trimmers must be provided (to "rubber the crystal" and tune the output to resonance) as front panel controls.

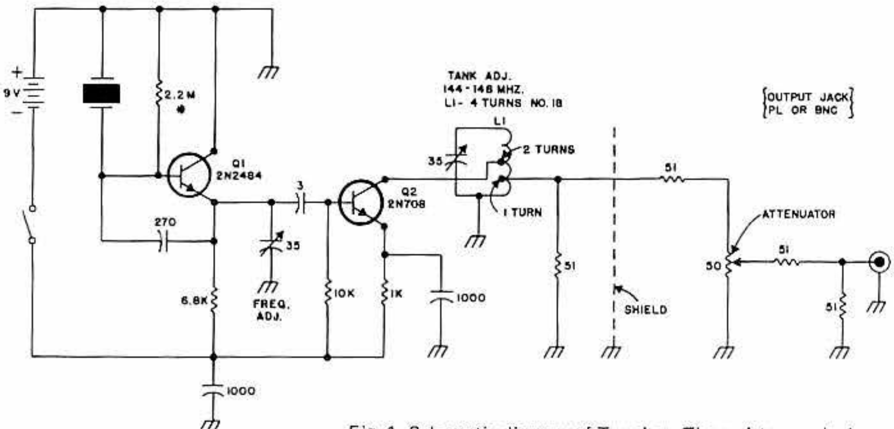

Fig. 1. Schematic diagram of Tweeker. The registor marked * may have to be changed to one of lower value if the recommended transistor is not used.

Circuit

Transistor Q1 in the first stage is a crystal oscillator which is very loosely coupled to Q2 the multiplier. This stage is biased into "class C" so as to multiply the 6 mhz crystal 24 times, up to the two-meter band. (3 mhz crystals will also work at a multiplication of 48.) Another important function of this stage is to attenuate the oscillator output (about 6-7 volts peak-to-peak) to a level usable at the two-meter frequency for alignment – 30 to 40 uv. The potentiometer adjusted attenuator takes the signal down to a level as desired by the operator.

I developed this unit for use on two meters. Several have been built and all work fine. But, for those of you who need a six-meter generator, this same circuit should work fine with the only modification needed being the final tank frequency. (A few more turns on the coil and a slightly larger trimmer capacitor.)

Adjustment and operation

Connect the output of the generator to your rig. Plug in a crystal and tune for zero (with frequency adjustment) on the discriminator (center frequency). Now, looking at the first limiter voltage, peak the final tank. Attenuate the generator as necessary to avoid saturation of the limiter. Repeat as necessary until a definite peak is reached.

If the generator will not go down into the noise with the attenuator control, the final tank may be detuned as necessary to provide the desired range on the attenuator.

CITIZENS BAND ALIGNMENT AID

Edward Lawrence WA5SWD

It is obvious that the front end of a receiver should be tuned to respond equally well across the band, and thus have consistent sensitivity at any frequency in the band, if maximum usefulness is to be obtained. The usual method of insuring reasonable results is to tune at the middle, and check the ends. This is rather spotty, at best, as it would be better to check all channels. This is involved since you must set both frequency and amplitude 23 times in a single pass.

Naturally, a good sweep generator could be used, and the rf-mixer tracking set with it instead, but this was not available to me at the time. The circuit in Fig. 1 was originally a super-regenerative receiver taken from the GE Transistor Manual. These, as is well known, tend to spray a lot of rf back out the antenna. With this in mind, I listened with a communications receiver and noted strong modulation products about 100 kHz above and below the "carrier." Deducing this to be caused by the quench frequency, I brilliantly decided to drop the quench frequency into the audio range, by increasing the capacitance across the base bias resistor. Lo and behold! Suddenly the entire Citizens Band was alive and jumping with a mass of "line noise," which at first appeared to be like

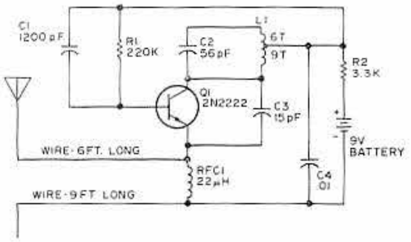

Fig. 1. Schematic of the alignment aid. L1 is 15 turns No. 22 tapped at 9 turns from the collector end. No switch is shown, so remove battery when not in use.

73 LIBRARY

NEW REPEATER ATLAS

Hundreds of new listings . . . listed by both location and frequency . . . dual listing, invaluable for your car . . . find those repeaters as you travel. This is the ONLY complete list of repeaters being published. Almost 3000 repeaters listed in this issue . . . repeaters from all over the entire world. only $1.95

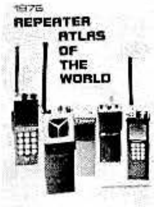

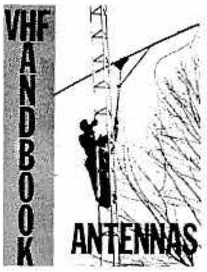

VHF ANTENNA HANDBOOK

The NEW VHF Antenna Handbook details the theory, design and construction of hundreds of different VHF and UHF antennas . . .

A practical book written for the average amateur who takes joy in building, not full of complex formulas for the design engineer. Packed with fabulous antenna projects you can build. $4.95

73 VERTICAL, BEAM AND TRIANGLE ANTENNAS by Edward M. Noll W3FQJ

Describes 73 different antennas for amateurs. Each design is the result of the author's own experiments: each has actually been built and air-tested. Includes appendices covering the construction of noise bridges and antenna line tuners, as well as methods for measuring resonant frequency, velocity factor, and standing-wave ratios. 160 pages. $5.50

73
Vertical,
Beam, and
Triangle
Antennas

73
Dipole and
Long-Wire
Antennas

73 DIPOLE AND LONG-WIRE ANTENNAS by Edward M. Noll W3FQJ

This is the first collection of virtually every type of wire antenna used by amateurs. Includes dimensions, configurations, and detailed construction data for 73 different antenna types. Appendices describe the construction of noise bridges, line tuners, and data on measuring resonant frequency, velocity factor, and swr. 160 pages. $5.50

2M FM HANDBOOK

Contains almost every conceivable circuit that might be needed for use with a repeater. All circuits explained in detail. All aspects covered, from the operator to the antenna. $5.95

The 2-Meter FM Handbook

Practical Test
Instruments You
Can Build

Practical Test Instruments You Can Build

37 simple test instruments you can make — covers VOMs, VTVMs, semiconductor testing units, dip meters, wattmeters, and just about anything else you might need around the test lab and ham shack. $4.95

IC OP-AMP COOKBOOK

by Walter G. Jung. Covers not only the basic theory of the IC op amp is great detail, but also includes over 250 practical circuit applications, liberally illustrated. 592 pages, 5½ x 8½, soft bound. $12.95

IC
op-amp
COOKBOOK

RF AND DIGITAL TEST EQUIPMENT YOU CAN BUILD

RF burst, function, square wave generators, variable length pulse generators — 100 kHz marker, i-f and rf sweep generators, audio osc, af/rf signal injector, 146 MHz synthesizer, digital readouts for counters, several counters, prescaler, microwavemeter, etc. 252 pages. $5.95.

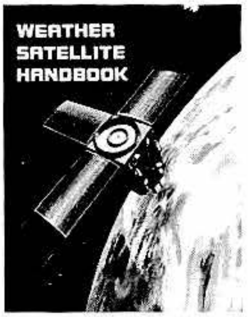

WEATHER SATELLITE HANDBOOK

Simple equipment and methods for getting good pictures from the weather satellite. Antennas, receivers, monitors, facsimile you can build, tracking, automatic control (you don't even have to be home). Dr. Taggart WB8DQT $4.95.

73
HANDBOOK
SLOW
SCAN
TELEVISION
HANDBOOK

SSTV HANDBOOK

This excellent book tells all about it, from its history and basics to the present state of the art techniques. Contains chapters on circuits, monitors, cameras, color SSTV, test equipment and much more.
Hardbound $7 Softbound $5

The Fascinating World of
RADIO COMMUNICATIONS

FASCINATING WORLD OF RADIO COMMUNICATION

Interesting stories of the history of radio pioneering and discovery. Also includes the fundamentals of broadcast band DXing. A must for every radio amateur. $3.95.

73
TEST
EQUIPMENT
LIBRARY
Component Testers

TEST EQUIPMENT LIBRARY

A multi-volume anthology that you won't find anywhere for three times the price! The editors of *73* present a complete guide to testing that is both basic and complete and aimed at everyone from beginners to amateur Extras. Everything from how to test diodes to frequency counters. Plus a cumulative index . . . all in *73*'s all new TEST EQUIPMENT LIBRARY Vol. I — Vol. III.

Vol. I — Component Testers — $4.95
Vol. II — Audio Frequency Testers — $4.95
Vol. III — Radio Frequency Testers — $4.95

1001 PRACTICAL ELECTRONIC CIRCUITS

Tab's new 1001 circuits is available for only $9.95 ppd. The next time you want a circuit for just about anything, eat your heart out that you didn't send for this book the first time you read about it. You'd better order the book right away, before they run out. $9.95

Master
Handbook of
1001
Practical Electronic Circuits

VHF PROJECTS FOR AMATEUR AND EXPERIMENTER

A must for the VHF op. Opening chapters on operating practices and getting started in VHF, both AM and FM, followed by 58 chapters on building useful test equipment, modifying existing and surplus gear. $4.95

73 MAGAZINE • PETERBOROUGH NH 03458

Hewlett Packard 6384A power supply

SPECIFICATIONS:

DC Output: 4 to 5.5 Vdc @ 8 Amps
Load Regulation: 2 mV no load to full load
Line Regulation: 2 mV for line changes between 104 and 127 Vac at any output voltage and current within specifications
Ripple & Noise: Less than 5 mV peak to peak
Crowbar, Factory Set at 6.25 V, Automatic cutback limit circuit and overvoltage crowbar protection.

HEWLETT PACKARD 6384A Power Supply $99.95

Power design 1210 regulated power supply

Output: 0 to 12 V @ 0 to 10 Amps
Regulation: 0.01% or better.
Ripple and Noise: Less than 500 uV rms. Voltage Controlled, Metered.

POWER SUPPLY 1210 - Regulated Power Supply - $149.00

a full range, 5-function 3½ digit multimeter

DC CURRENT: 0 to ±1.999/19.99/199.9/1999 mA
AC CURRENT: 0 to 1.999/19.99/199.9/1999 mA
RESISTANCE: 0 to 199.9 Ω through 19.99 Megohms (6 ranges!)

ACCURACIES

DC VOLTS: ±0.2% F.S. ±0.2% of reading
(1000V range: ±0.5% F.S. ±0.5% of reading).
AC VOLTS: ±0.7% F.S. ±1% of reading
(1000V range: ±0.7% F.S. ±2% of reading).
DC CURRENT: ±0.5% F.S. ±0.8% of reading
(1000 mA range: ±0.5% F.S. ±1.8% of reading).
AC CURRENT: ±0.5% F.S. ±1.3% of reading
(1000 mA range: ±0.5% F.S. ±2.3% of reading).
RESISTANCE: ±0.5% F.S. ±0.8% of reading
(10 megohm range: ±0.5% F.S. ±1.8% of reading).

RANGES

DC VOLTS: 0 to ±1.999/19.99/199.9/1500 Volts
AC VOLTS: 0 to 1.999/19.99/199.9/1000 Volts

DATA PRECISION MODEL 134 — $169.00

INSTRUMENTS & COMPONENTS

AST/SERVO SYSTEMS, INC.
20 Republic Road, North Billerica MA 01862
(617) 667-8541

NEW 73 BOOKS

TEST EQUIPMENT LIBRARY
Vol. III, Radio Frequency Testers

Radio frequency waves are the common denominator of Amateur Radio so here is a book for all hams. No matter what your specific interest, such items as SWR, antenna impedance, line impedance, RF output and field strength should interest you. This book not only gives detailed instructions on testing these items but includes sections on signal generators, crystal calibrators, grid dip oscillators, noise generators, dummy loads and much more. It's a must for all up-to-date shacks.

Test Equipment Library
Vol. III, Radio Frequency Testers $4.95

TEST EQUIPMENT LIBRARY
VOL. II, Audio Frequency Testers

HOW'S YOUR SPEECH POWER? YOUR SHIFT? YOUR SYNC?

You can find out easily with a little time and a junk box full of parts. It's all right there in the new Volume II of the 73 Test Equipment Library . . . Audio Frequency Testers . . . jam packed full of all kinds of audio frequency test equipment. So if you're into SSB, RTTY, SSTV, etc., this book is a must for you . . . good book for hi-fi addicts and experimenters too!

Test Equipment Library
Vol. II, Audio Frequency Testers $4.95

BUILD YOUR OWN TEST EQUIPMENT!

SAVE A BUNDLE!

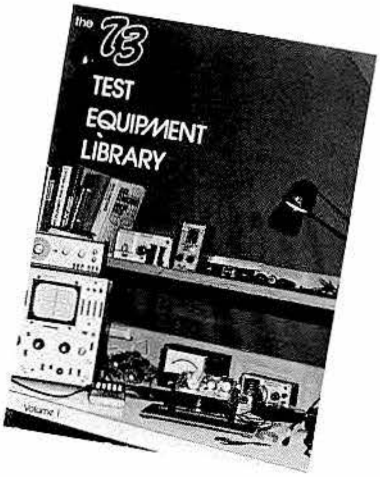

TEST EQUIPMENT LIBRARY
VOL. I, Component Testers

Just out is Volume I of the 73 Test Equipment Library . . . how to build transistor testers (eight of 'em), diode testers (3), IC testers (3), voltmeters and VTVMs (9), ohmmeters (8 different kinds), inductance (3), capacity (9), Q measurement, crystal checking (6), temperature (2), aural meters for the blind (3) and all sorts of miscellaneous data on meters . . . using them, making them more versatile, making standards, things like that. Invaluable book, ridiculously low priced.

Test Equipment Library, Vol. I, Component Testers $4.95

Send me:
☐ Vol. I . . . . Component Testers @ $4.95
☐ Vol. II . . . . . Audio Frequency Testers @ $4.95
☐ Vol. III . . . . . Radio Frequency Testers @ $4.95

☐ Cash enclosed ☐ Check enclosed

☐ BankAmericard ☐ Master Charge ☐ American Express

Card # ____________ Interbank # ____________ Expiration date ________
Name ____________ Signature ____________
Address ____________
City ____________ State ____________ Zip ____________

73 magazine • Peterborough NH 03458

Meshna

New England's #1
Surplus Dealer For Over 20 Years.

Send for our latest catalog.

Meshna PO Box 62 E. Lynn, Mass. 01904

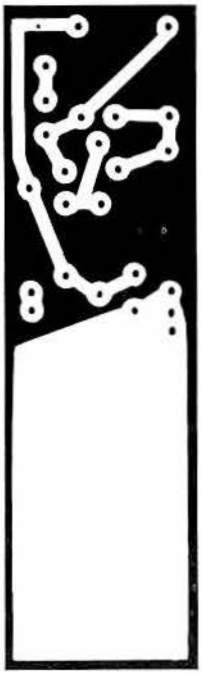

Fig. 2. Full size PC board layout (foil side).

natural noise. Then I turned on the BFO, which showed it to be many, many, many closely spaced "carriers." I presume the lower quench frequency develops a much larger bias swing on the base of the transistor, FM'ing it across the entire band. This stuff is concentrated in the region of 27 MHz, with very little spill-over into 10 or 15 meters, although the level across the design range is rather constant. The nearly constant rf level is what I needed to simplify tuneup of the front ends of the units I was working on.

Figure 2 is the layout of the P.C. board I made to facilitate reproduction of this useful circuit for a few interested friends and Fig. 3 is the component layout.

I used 6 ft of wire as an antenna connected to the emitter side of the rfc, with 9 ft of wire on the battery side of the rfc as a counterpoise. Since this made it a bit large, I carefully hung it vertically in a handy tree nearby. Of course, since this put a strong signal on every channel into every CB receiver within 300 yards, I limited the use of this device to avoid unnecessary interference to the deserving users of the band. However, one fellow accidentally launched his into a tree too high to recover with an over-ambitious heave, and it ran for about a month in the dead of winter in a nameless northern city, helping every CBer for miles around align their sets! It only draws about half a milliamp from a 9 volt battery.

Since I am sure many others would like to align their 11 meter receivers for optimum results, I am presenting this circuit as a very low cost and worthwhile aid.

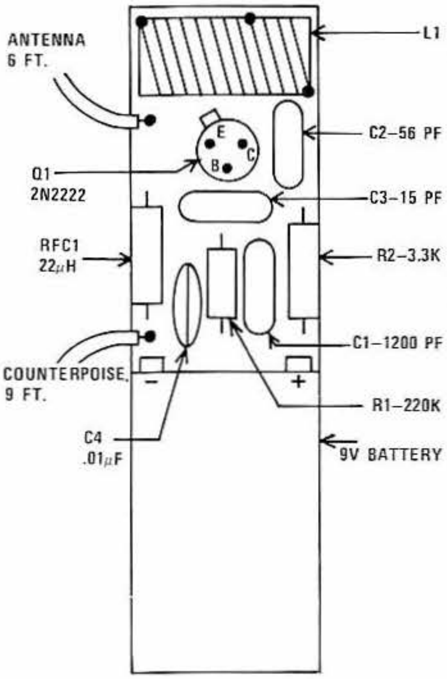

Fig. 3. Component side parts placement.

FET CHIRPER

Chuck Hines K6QKL

The Chirper is an automatically keyed, crystal controlled, signal source which may be used to optimize the signal-to-noise ratio of a receiving converter. Homebrew or commercial, converters are a common thing around an amateur station. And, most of the VHF Tribe have read thru a jungle of esoterica dealing with low noise front ends, the velvet beauty of FET's on Two, noise generators and eternal truth, and how to copy 20 db below the noise by the selective use of liquid helium. With a kind of relentless evolution converters have been getting better and better, noise figures become lower, and the prices of suitable front end devices are dropping by the hour. But when it comes to aligning these converters the scene is one of wretchedness. A black art at best, the job is taken up with an enduring combination of blunt instrument and myth. The latter have a certain charm. Are you convinced your converter is in top notch condition because you can "hear noise" when you attach the antenna – or better yet, when you place a 50 ohm resistor across the input? Try putting a complete short across that same input. Shorts aren't much good as noise sources. You'll find the short gives about the same change in noise level as the 50 ohm termination. What has changed is the impedance the front end "sees". The same is partially true of the noise from the antenna. Neither is indicative of the performance of the converter. Peaking the system up for maximum on either a weak signal or on noise gets you nowhere. The diode noise generator which every VHF book of substance describes is a good and useful tool when used properly. The assumption is that everyone already knows full well how to use it and does so. Few in fact do.

I'm sure you've read of it before in many places, but a little redundancy is in order. The noise with which you are concerned is the noise generated *internally* by the first tube, transistor, or other active device the signal encounters upon its arrival at your converter. By fiddling with the external reactances, adjusting the voltage and current and otherwise manipulating the things soldered to the device, one may minimize the internally generated noise. At the same time the reason the front end exists is to amplify the signal. One usually desires as much amplification possible, short of smoke and oscillation. Minimum noise and maximum amplification is the game. Though the two are not quite mutually exclusive a certain amount of compromise takes place. Thus, the signal to noise ratio. When aligning a converter's first stage every adjustment effects *both* signal and noise. Given a constant signal source coupled into the converter thru an appropriate impedance, the job is finished when the front end has been adjusted for the greatest *difference* between signal and noise of which it is capable.

The Chirper is designed to help you do all this by letting you see what effect each adjustment has on both signal and noise. The TIS34 oscillates at a frequency controlled by the crystal. With the constants shown, that can be anywhere between 8.2 and 36 MHz. The variable capacitor must be adjusted for resonance. It isn't particularily critical but its setting peaks the *rf* output at either the fundamental or some harmonic. For 6 meters an 8.35 MHz crystal is used. A 9.0 MHz rock will pin the meter when the

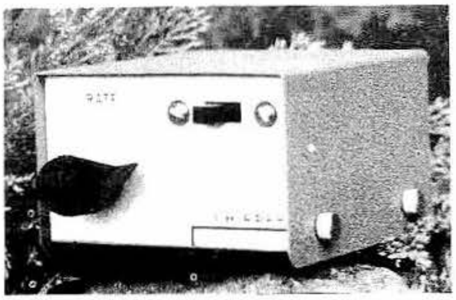

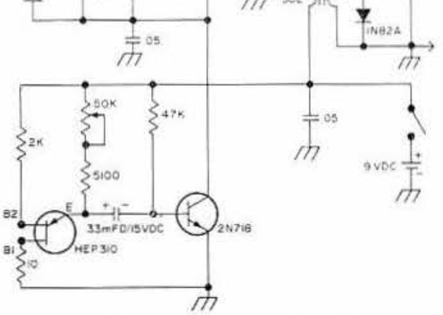

Fig. 1. Schematic of the FET chirper.

Chirper is connected into a 2 meter converter. The Amidon toroid is wound with No. 30 enameled, 40 turns for the primary and 5 turns for the secondary. After it is wound, spread the turns to fill the toroid and paint it with Q dope. The diode and 5 pF coupling capacitor are connected with the shortest possible leads, the diode being grounded at the *rf* connector. The harmonic output is excellent and quite useable at 1200 Mhz.

The oscillator is turned on and off by a multivibrator combination of unijunction and NPN transistor adapted from the *G. E. Transistor Manual.* The rate at which the multivibrator cycles is determined by the large value capacitor, in this case 33 mF. The polarity of the capacitor is critical. Observe it. To increase the cycling rate, decrease the capacitance; and, to decrease the rate, put in a larger value. Mine cycles a little under once per second. A value somewhere between 30 and 40 mF should suit your needs. You are better off scrounging some odd value from a defunct computer board because of the tolerance problem. If it says 33.2 mF, its probably pretty close to that value. Otherwise you're dealing with tolerances of plus 100% and minus 50% or something equally grotesque. The 50K pot determines the portion of the cycle during which the oscillator is On and is mislabeled rate on the Chirper shown. The HEP-310 is generally available and inexpensive. Other unijunctions were not tried. On the other hand almost any NPN of reasonable quality will work in place of the 2N718. A number of 2N388 and 2N3478's were tried and behaved well. It's a good place to use those transistors you've replaced with FET's. Use something with a Beta of 50 or better for best results. The 5100 ohm resistor in series with the pot is for current limiting. It's deletion will increase battery drain with no increase in Chirper performance. Normal current from the 9 volt battery is around 5 mA.

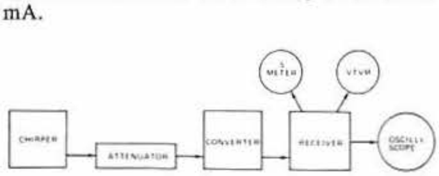

Fig. 2. Test set-up for converter alignment.

Construction is non-critical and pretty much a matter of taste. Mine is built on a piece of vector-board and mounted in a Suzurando box, model M-1N. It measures 3¼ X 2 X 4" and there is still room inside for additions. It sells for 330 yen, about 92c. A slide switch is used to turn the power on and off. Paint one well of the slide switch with red paint – Testor's Pla, a model plastic paint, is good – and the switch will indicate its position. Red for On and black for Off. It saves batteries. Check your work and the polarity of the large capacitor. Re-check the connections to all the semi-conductors. With four different kinds of devices things can become confused. Set the pot to the middle

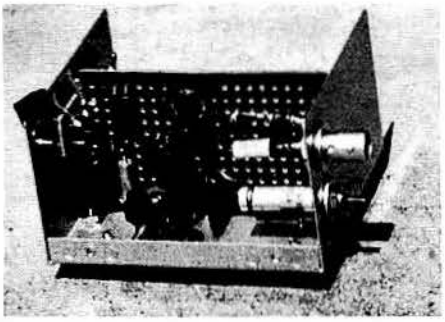

of its range. Insert a crystal in the socket. Connect the Chirper to your converter, turn the switch on and adjust the variable capacitor for the highest reading on your S-meter. The oscillator will turn on and off. Varying the pot will extend or diminish the amount of time the oscillator is on. Whatever you do, *don't connect the chirper to an external antenna.* The harmonic content is high and even at this power level is sufficient to cause severe interference to television receivers within a two block radius.

For converter alignment, the test set-up is illustrated in Fig. 2. The Chirper is fed to the converter thru an attenuator for two reasons. First, the power output of the Chirper is too high on six and two. You don't want to align with a forty over nine signal. Something around S-5 to S-7 is desired. Second, the attenuator maintains a 50 ohm termination for the converter. A converter cannot be aligned with a floating input impedance. Fixed and variable attenuators of excellent quality are available thru surplus and homebrew data is available. See *73,* January 67, p. 40 for one that will do the job. Turn the receiver avc off. The read-out options are diverse. The best is probably a scope connected to the *if.* A vtvm can be used, connected to the audio output. And, the S-meter can be used with the avc on *fast.* This will vary with the receiver and it's particular time constants. What needs to be avoided is avc pumping that interferes with your readings.

Turn the Chirper on, adjust the attenuator for a convenient signal level. When the oscillator is on, you're reading signal. When the oscillator is off, you're reading noise (on the scope, vtvm, S-meter, etc.). As you make adjustments on your converter, observe the effect on both signal and noise. Adjust for the greatest difference between the two. Turn the Chirper off and re-check the converter neutralization. If necessary, reneutralize the converter and go thru the whole thing again. Talking about it makes it seem somewhat complex. It really isn't and the whole business won't take long once you've done it. It will become quickly apparent that highest signal level and lowest noise level do not coincide. You can vary the bias, voltage, etc., and observe the effects of each on the signal-to-noise relationship. You can, in short, optimize your converter's performance.

A number of things can be done with the Chirper. There is room in the box to build another oscillator section connected to the transistor collector, operating in parallel and simultaneous with the first oscillator. By appropriate choice of crystal and attenuation, both signals can be introduced into the converter in order to adjust the mixer for minimum cross modulation.

Or, instead of using an oscillator at all, you can use the switching section of the Chirper to key a noise generator on and off. This has a certain attraction where an integrating network is used prior to a vtvm. In this case noise is used as a signal.

In spite of it's name, the Chirper is remarkably stable. Chirp becomes apparent from two meters or so, but is no problem. Build one and take the myths out of your converter.

UHF SIGNAL SOURCE

Bill Hoisington K1CLL

The UHF experimenter learns early that the UHF ham bands don't always furnish a signal when you need them. When I build receivers and converters for 432 and 1296, I find that I need a small signal generator for alignment and band spotting. Not much power is needed for this work, so transistors are the ideal choice for generating the signal. This signal source is in three parts, a 432 source, a tripler to 1296 and a modulator.

The 432 Driver

I started with a low-cost 27 mc crystal that I found in my junk box. (Never mind how it got there!) I used my usual phase-reversing crystal circuit followed by a bunch of doublers. Lower-priced transistors were used in the early stages and the UHF ones saved for later. You really need good ones on 1296! Each doubler is biased from rectification of rf from

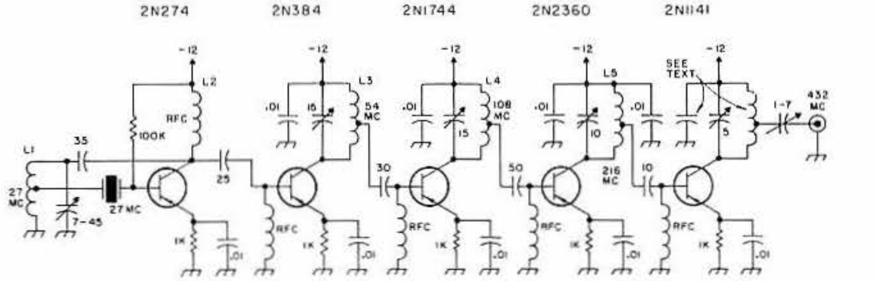

Fig. 1. Schematic diagram of the 432 mc signal source. The transistors aren't critical in most cases and other UHF and VHF ones will work fine.

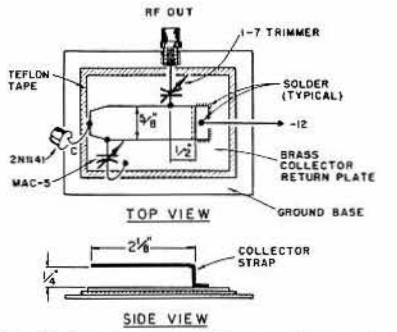
RF OUT
I-7 TRIMMER
TEFLON TAPE
SOLDER (TYPICAL)
3/8"
1/2"
-12
2N1141
C
MAC-5
BRASS COLLECTOR RETURN PLATE
TOP VIEW
GROUND BASE
2 1/8"
1/4"
COLLECTOR STRAP
SIDE VIEW

Fig. 2. Details of the 432 mc collector circuit.

the previous stage. If you need more output than this circuit gives you, use less than a 1 k resistor in the emitter, but watch out for high collector current. **Fig. 1** shows the schematic of the 432 mc generator and **Fig. 2** gives details of the 432 mc collector circuit.

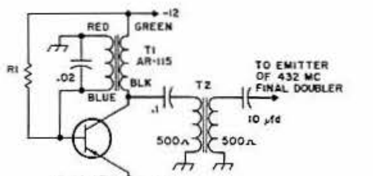
-12
RED
GREEN
T1
AR-115
R1
.02
BLK
T2
BLUE
TO EMITTER OF 432 MC FINAL DOUBLER
.1
10 µfd
500Ω
500Ω
2N247
1K
.1

Fig. 3. The tone oscillator for the 432 mc signal source.

The tone modulator

This is a crude modulator (**Fig. 3**), but it works. The modulation transformer is not absolutely necessary, but seems to improve results. You can apply the modulation almost anywhere for this application, but modulating across the 1 k emitter resistor gave the cleanest sound with the doubler used.

Now to 1296

I built this tripler to 1296 just for the fun of it. But it worked quite well. I normally don't hold with triplers at this frequency, but it's an easy way to get 1296 mc energy from the 432 mc driver. The transistor I used was a Motorola 2N1141. It's several years old and there are better ones that are far cheaper now. But it does work on 1296. I couldn't get it to work with grounded emitter, but grounded base is fine. **Fig. 4** gives the schematic but Fig. 5 gives the details, which are vital. The input on 432 tunes very nicely. But I had to reduce the emitter resistor in the doubler to 432 to get enough drive. It ended up at 200 ohms. The collector circuit is short, but tunes smoothly. The 1296 mc output registers 100 a in the 1296 cavity in the May 73.

You might try a small amount of modulation on the 1296 mc tripler. Also a waveguide attenuator. Be seeing you on 23 cm.

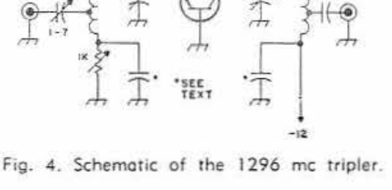
2N1141
MAC-5
432 MC
1-7
1K
*SEE TEXT
1296 MC
-12

Fig. 4. Schematic of the 1296 mc tripler.

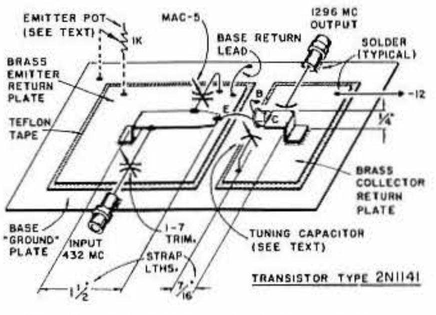
EMITTER POT (SEE TEXT)
1K
MAC-5
1296 MC OUTPUT
BASE RETURN LEAD
SOLDER (TYPICAL)
BRASS EMITTER RETURN PLATE
-12
B
E
C
TEFLON TAPE
1/4"
BRASS COLLECTOR RETURN PLATE
BASE "GROUND" PLATE
INPUT 432 MC
1-7 TRIM.
TUNING CAPACITOR (SEE TEXT)
STRAP LTHS.
TRANSISTOR TYPE 2N1141
1 1/2"
7/16"

Fig. 5. Pictorial layout of the 1296 mc tripler.

CRYSTAL CONTROLLED SIGNAL SOURCE WITH INFINITE ATTENUATOR FOR 144, 432 AND 1296 MHz

Bill Hoisington K1CLL

One of the most useful test-equipment gadgets the homebrewer can build is a signal generator. The one described here is of commercial quality and it can be completely contained inside a waveguide. Positioning, by sliding along the waveguide, provides a variable-strength stable signal of one millivolt, one microvolt, one nanovolt, or less, dropping down gradually to a true zero. It does this in a perfectly smooth fashion without steps or jumps so that every fraction of a decibel in lower noise figure shows immediately on the slide dial. What's more, the slide can be calibrated so that FM'ers can use the device for directly measuring receiver sensitivity in tenths of a microvolt.

In building a 6 meter receiver recently for maximum absolute sensitivity I naturally had to check especially on the first-stage rf transistor and circuit for minimum noise figure. (For this type of work you *must* have a signal generator capable of being attenuated out of sight with any receiver you can buy for any money.) The usual generators on the market under $100 do not do this. And many of the very expensive generators get so leaky that they have to be used 200 ft from the receiver. At any rate, the generator described here can be made up quickly and at low cost, and it *is* stable, reliable, and infinitely variable.

Waveguide

The only possible difficulty might be in obtaining the piece of waveguide needed. The piece I used is 4¼ in. wide by 2 1/8 in. high, and is 24 in. long. If you have a choice, get a piece a little longer. You could make up this item out of brass or copper if you had to, because in this case it is not used to *carry* energy but to *attenuate* it, so the worse you make it the better!

The waveguide *must not* have any holes in it and should be reasonably smooth inside; otherwise your dial would not read smoothly in attenuation. You *could* use copper or aluminum drain pipe, although I have not tried them yet. Working directly on the rf, this attenuator is good for any kind of modulation, including SSB, FM, pulse, or what have you.

Construction

Figure 1 shows the basic idea. When the signal generator plate is close to the

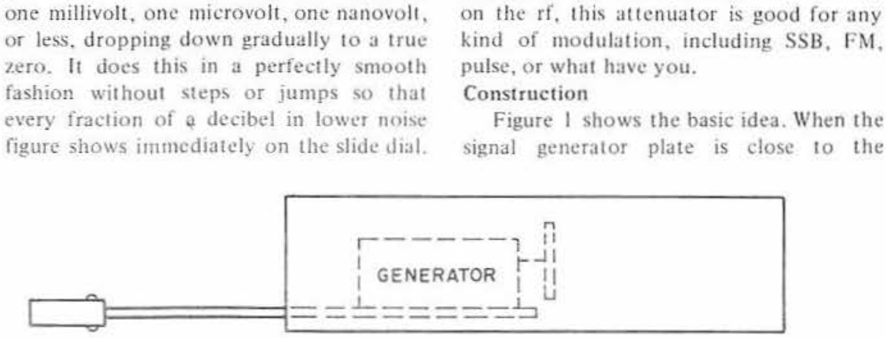
GENERATOR
WAVEGUIDE CLOSED AT REC. END, OPEN AT THE OTHER END
COPPER END CAP MUST BE TIGHT!
DOWEL
B A T
A F
M O D
O S C
RCA JACK J1
CARDBOARD OR PLYWOOD BASEBOARD, SLIDES IN & OUT
SIGNAL GENERATOR PLATE
RECEIVER PICKUP PLATE

Fig. 1. Sketch shows plan-view layout of unit inside waveguide attenuator. The oscillator unit is mounted on a flat wood or cardboard strip that can be calibrated to give accurate indications of output signal.

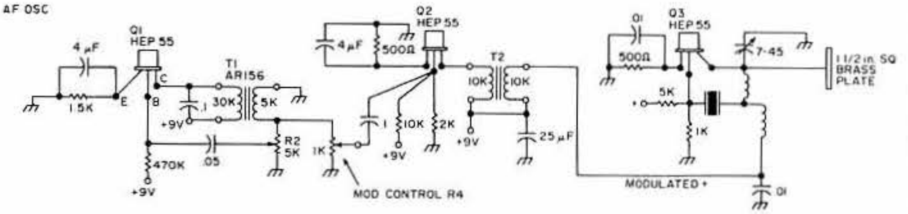

Fig. 2. Schematic of generator. This entire assembly about 3 x 4 in. including small 9V battery and switch must be entirely inside the waveguide. No wire or metal of any kind can be brought outside.

receiver pickup plate, you can get about 100 mV of signal into the receiver, and it is handy for checking diode receivers. When the two plates are about 8 in. apart, the signal is just detectable on a good receiver. Additional spacing between plates amounts to "waveguide beyond cutoff." I do not believe that there is any receiver in the world that can pick up the signal much beyond the 8¾ in. point.

Pretty soon in your receiver "peaking" work you get to that signal that may be but a tenth of a microvolt or so, and you begin dreaming about cryogenic front ends, masers, and such. As mentioned, every fraction of a decibel lower in noise figure, every improvement in sensitivity comes out rigorously and relentlessly on that slide dial. You can easily check which of your low-noise transistors is really low, whether that MOSFET will do a better or worse job for you, and in which circuit.

As you go up in frequency you may have to make smaller and smaller oscillators in order to fit in smaller waveguides to get the cutoff effect. (That will not be a problem if you read 73; the May issue described a "postage-stamp-sized" rf generator that is an ideal candidate for the signal source.)

Circuit

A crystal oscillator, an af oscillator, and a simple class A modulator do an excellent job to start with. Figure 2 shows the present unit as used on 6 meters. It must be stressed again that no wire or other piece of metal may be allowed to reach the outside from this assembly. I'm making up another for 2 meters soon (still my favorite band) and will try one on 450 a little later.

Audio

A controlled-feedback transformer-coupled af oscillator does a good job in furnishing a sine wave. A Motorola HEP55 is used for the oscillator, with feedback to the base from the collector through transformer T1, controlled by resistor R2. Audio output is taken off the 5 kΩ winding of T1, is fed through R4 the modulation control, and then to the base of af modulator Q2. Transistor Q2 is set up for low-power class A operation because not much modulation is needed for the signal generator. Transformer T2 is an old 5W unit from "tube-type portable" days. The secondary of T2 feeds a modulated +9V signal to Q3, the crystal-controlled 50 MHz oscillator.

This rf oscillator is one of my negative-feedback jobs with phase reversal in the crystal. A 1½ in. square plate is tied onto the collector, radiating energy to the receiver pickup plate facing it inside the waveguide. This energy is rapidly attenuated as you move the plates apart, and should be impossible to detect after some 9 or 10 in. of separation.

Once again, do *not* bring any wires or any other metal or conductor out from the oscillator assembly. If you want an outside controlled switch or other control, bring it out as a wooden dowel handle.

That's about it. Tune everything up outside the waveguide on the bench; when you're satisfied, plug your best 6 meter receiver into J1, push the oscillator plank along the waveguide (or rather I should say pull it along) away from J1. You'll get a surprise! Hope this helps you with your low-noise receiver work. It did a lot for me.

LOW COST 220 MHz SIGNAL GENERATOR

Bill Hoisington K1CLL

This article describes the design and construction of an easy to build, inexpensive, crystal controlled signal generator for the 220 MHz band, including a very low cost attenuator that goes from a quarter volt down through 1/20th of a microvolt and on to a real zero (of rf power). It is very useful for receiver front-end tuneup, low noise tests, and as a portable field generator for overall antenna tests through the receiver. For signal identification purposes, af and FM modulation are included.

If you really want to fight for a low-noise front end, this piece of equipment will be of great assistance to you, because the attenuation really is infinite and without any difficult bypassing or shielding.

Design of the attenuator.

Infinite attenuation is achieved here by the use of a 50¢ piece of aluminum tubing, as shown in Fig. 1. You cannot drive 220 MHz signals more than a few inches down inside of a piece of aluminum tubing. By putting *everything* – battery, on-off switch, circuit and all – on the movable generator strip and sliding it in and out of the tubing, you avoid all touchy, difficult and expensive bypassing, costly attentuator pots, shielding, etc., and provide a simplified means of varying the attenuation with stable, smooth, easy repeatability. Calibration is of the slide rule variety and also simple as far as writing down the microvolts on the scale is concerned.

This principle is older than radio tubes; in fact Sir Oliver Lodge used it in his 1890 microwave work.

Attenuator Construction Details

Figure 1 tells most of the story, with details in Figs. 2 through 9. An adequate rf seal can be made at the pickup end of the aluminum tubing, standard TV masting, 1¼ O.D., by 2 or 4 tabs in one end as in Fig. 3 and bending them back as shown, then cutting off the excess tubing. Install the pickup, plate, output jack, and end plate as shown in Fig. 1. I used time-saving external mounts for fastening it down to the wood baseboard as shown in Figs. 2 through 5. Drill a ¼ in. hole for a screwdriver as in Fig. 2, and use angles for the pickup end. Figure 2 also shows the scale in use for attenuation settings, and Fig. 3 shows pointer details.

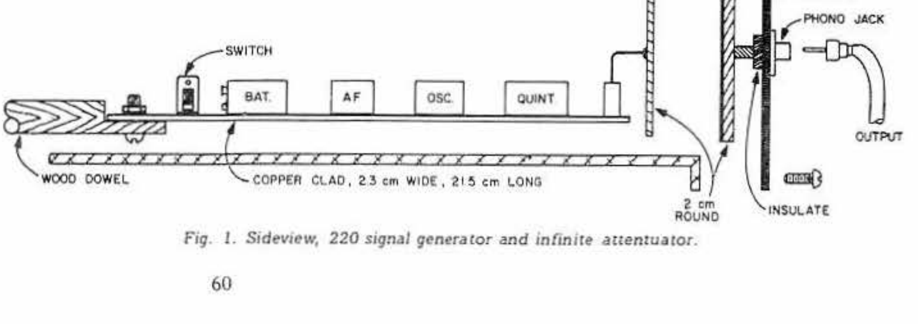
STANDARD ALUMINUM MAST, 1 1/4 INCH O.D., 46 cm LONG
COPPER
END PLATE
COPPER
PHONO JACK
SWITCH
BAT.
AF
OSC.
QUINT
OUTPUT
WOOD DOWEL
COPPER CLAD, 2.3 cm WIDE, 21.5 cm LONG
2 cm
ROUND
INSULATE

Fig. 1. Sideview, 220 signal generator and infinite attentuator.

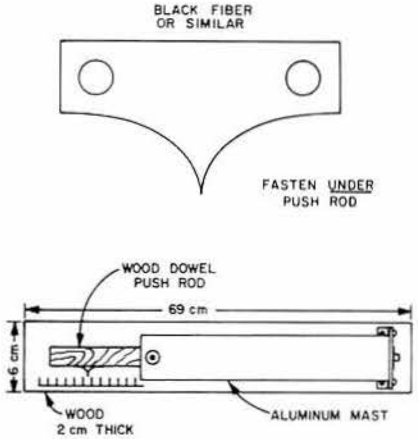
BLACK FIBER OR SIMILAR
FASTEN UNDER PUSH ROD
WOOD DOWEL PUSH ROD
69 cm
6 cm
WOOD 2 cm THICK
ALUMINUM MAST

Fig. 2. Top view of the 220 signal generator and details of the calibration pointer.

Figures 6 and 7 show pictorials of the layout, top view and side view.

The Generator

Nothing too fussy here, but attention to details will assure reliable af and rf oscillation at low power and low battery drain and good frequency multiplication. Figure 8 shows the schematic with the details of the two oscillators, the crystal in the 44 MHz range, and the quintupler. The af uses a standard circuit which, however, has one item to watch. Contrary to a transformer coupled circuit, which is seldom mentioned, this twin-T job has a nasty trick of not starting every time. However a small cap from collector to ground cures this and makes it 100% reliable in that respect. The emitter being grounded, I suppose this establishes the correct in-phase relation with the collector, in which both of these elements should be in phase. You can put a small trim pot af gain control between the modulator and the oscillator if you wish, watching out for dc voltages of course. As shown here, there is plenty of modulation for signal identification, both AM and FM.

Referring to Fig. 8, at the left is the af oscillator. It is not down symmetrically, but you can note the two 22K resistors and the two .02 frequency settling caps, along with the .05 and the 2.2K terminating the lines. All of these set the frequency, and to change the frequency you should vary all of them in at least their approximate present ratios. It is around 500 cycles as shown. Do not forget the "starting cap" from collector to ground.

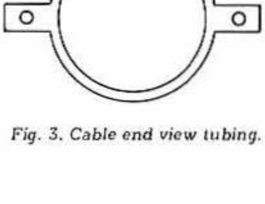

Fig. 3. Cable end view tubing.

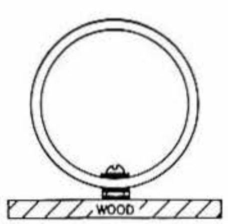

Fig. 4. Open end view.

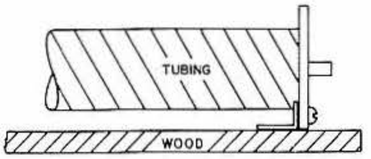
TUBING
WOOD

Fig. 5. Side view.

The output goes to the antenna plate on the forward end of the baseboard plank. When this plate is moved all the way in so that it is only ¼ in. from the cable pickup plate in the end of the aluminum tube, a full scale reading may be obtained on a 50 microamp meter on the output of a tuned diode detector (see Fig. 9). My first model here has marks on the scale (see Fig. 2) for 1 rf stage; 2 rf stages; mixer (feeding into a good i-f strip); mixer plus 1 rf, etc. At the 1 rf plus mixer, you begin to hear all the repeaters within 100 miles or so. With 2 rf plus mixer (followed by a sensitive i-f of course) you are really getting sensitivity. This is where you put a scope in line in order to have an electronic comparison point for signal coming out of the noise purposes, and then can really get into the low-noise bit, if you have a hermit location. Just happens I have one here. When I hear the noise of a car, it's a visitor!

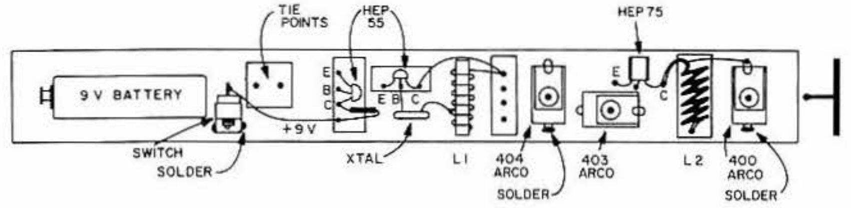
TIE POINTS
HEP 55
HEP 75
9 V BATTERY
E
B
C
E B C
+9V
E
C
SWITCH
SOLDER
XTAL
L1
404 ARCO
SOLDER
403 ARCO
L2
400 ARCO
SOLDER

Fig. 6A. Layout, top view.

This audio is fed to the base of the multiplier where it provides some AM and some FM modulation for signal identification purposes. When working with receiver oscillators, and in particular with high-ratio multipliers, this is very important. It may also be locked into any old scope sync for noise figure and sensitivity comparisons. The scope sync gives a reference point where the signal to noise ratio will always be the same, without resorting to guess work. The crystal oscillator is my old tried and true crystal phase-reversing job, which uses negative feedback from the collector coil, which, after going through the crystal, reverses phase and becomes positive, thus assuring oscillation but *only* on the crystal frequency. A HEP 75 (similar to the famous 3866) is used for the quintupler. A lot more output is noted with this powerful but smooth operating old faithful, still good to 450 MHz.

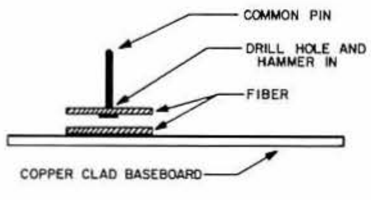
COMMON PIN
DRILL HOLE AND HAMMER IN
FIBER
COPPER CLAD BASEBOARD

Fig. 6B. Tie-point construction.

Output

You will see for yourself as soon as you start testing that the attenuator is smooth-working and stepless, and that true infinite attenuation is at hand. An rf input state (pre-amp) with a noise figure a fraction better than another shows right away on the scale. For example, adjustment of the fixed bias voltages on the two gates of a 3N200 or 3N201 shows right away on the scale as the

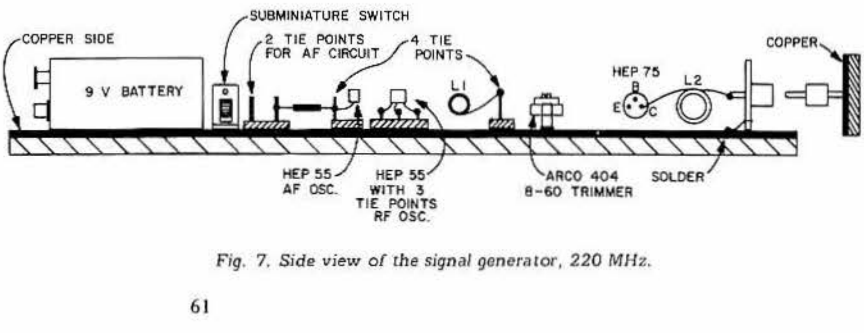
SUBMINIATURE SWITCH
COPPER SIDE
2 TIE POINTS FOR AF CIRCUIT
4 TIE POINTS
COPPER
9 V BATTERY
L1
HEP 75
B
E
C
L2
HEP 55 AF OSC.
HEP 55 WITH 3 TIE POINTS RF OSC.
ARCO 404 8-60 TRIMMER
SOLDER

Fig. 7. Side view of the signal generator, 220 MHz.

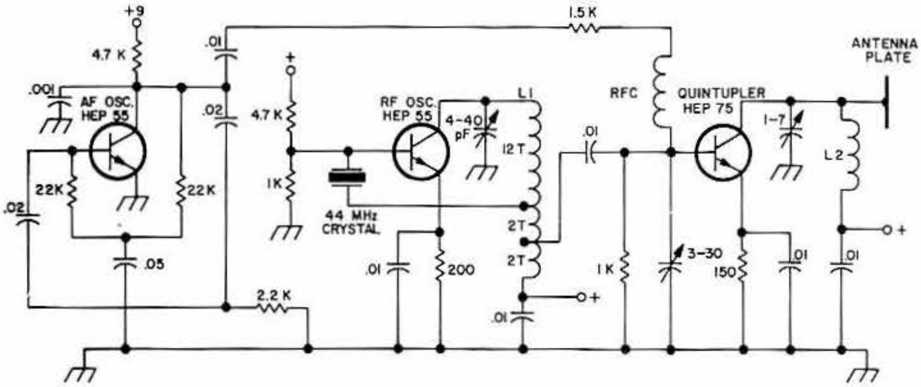
+9
4.7 K
.001
AF OSC.
HEP 55
22K
.02
22K
.05
.01
.02
2.2 K
+
4.7 K
1K
44 MHz
CRYSTAL
RF OSC.
HEP 55
4-40
pF
L1
12 T
2T
2T
.01
.01
200
.01
O+
1.5 K
RFC
1K
QUINTUPLER
HEP 75
3-30
150
1-7
.01
L2
O+
.01
ANTENNA
PLATE

Fig. 8. Schematic. L1 = 16 turns No. 26 output tap at 2 turns, crystal feedback tap at 4 turns, from cold end. Wound on phenloic form .6 cm O.D. L2 = 6 turns No. 18 bare, air wound, .6 cm O.D., 2.5 cm long between tie points. RFC = about 40 turns No. 40, on phenolic form .3 cm O.D., 1 cm long (not critical).

push-rod is moved in and out and the signal is locked onto the scope. This work you can do right on the bench and at low cost.

There is quite a bit of mechanical work in the unit, depending on just how much "finish" you want it to have. You can also bring a dowel rod for "on and off" use. Do not, under any circumstances, bring out a *conductor*. You *can* do this, but only with an extreme amount of filtering, which is not part of this article. Be sure to set up the baseboard, antenna plate and battery first, and get them working mechanically. With a drawn-out shape like this I generally start with a longer piece of copper-clad than needed, build from one end, and then cut off what is left over. Understandably, once you have made the first one you can always see, after it is done, many ways of improving it. However, someone has to make the first one, and that's generally my job.

Antenna and Field Tests

Out of the tubing, and with a small antenna connected via a one turn link around the quadrupler coil and then to ground, returning L2 for maximum output, this little rig puts out a lot of signal on 220. Especially if you reduce the oscillator emitter resistor! Up to several volts of rf can be obtained in a tuned diode receiver if you push things along, which is around 5 to 10 mW. If you place this generator out in a field several hundred yards or more away, you can then line up your antenna on the car or house, check antenna cables, antenna input alignment, and match or mismatch for lowest noise figure, etc.

Front end alignment should first be done with a relatively broadband i-f strip on 10.7 MHz. Be sure nothing metallic on the generator strip protrudes enough to touch the inner wall of the tubing, or "scratch" will occur in the high gain receiver. A piece of thin fiberglass or other insulating sheet wrapped around the whole generator movable plank is a good precaution.

Once again I include a 220 MHz tuned diode detector, which is an absolute must for frequency multiplication, especially quintupling and such, where the other unwanted harmonics are as little as 20% away from the desired frequency. Figure 9 shows this piece of test equipment in pictorial schematic form. Remember, shape is of considerable importance as you go from VHF into UHF. It is quite easy to make the square trough line out of an old piece of copper clad. Or even a new piece! And this particular one described and shown in Fig. 9 goes very well to over 450 MHz and thus is very nice for the next band also. If you make it just as I've shown it, it will do a good job for you.

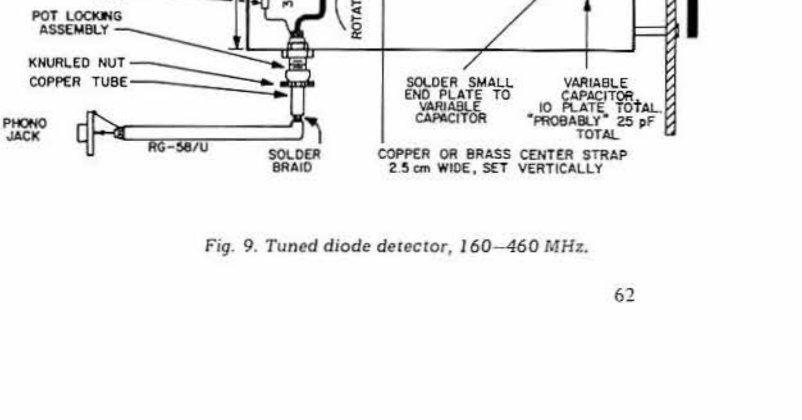
AF
DC
OUTPUT
"PHONO" JACK
ADAPTER
(EXTERNAL)
1.4 cm SPACERS
COPPER-CLAD
COPPER INSIDE
TWO SMALL
LAFAYETTE
.001 CAPS
IN295
DIODE
10 X 10 cm
PANEL
10.2 cm
POINTER
KNOB
5 cm
7.6 cm
INSIDE TEFLON
OF RG-58/U
2.2 pF
3 cm
ROTATES
POT LOCKING
ASSEMBLY
KNURLED NUT
COPPER TUBE
PHONO
JACK
RG-58/U
SOLDER
BRAID
SOLDER SMALL
END PLATE TO
VARIABLE
CAPACITOR
VARIABLE
CAPACITOR,
10 PLATE TOTAL
"PROBABLY" 25 pF
TOTAL
COPPER OR BRASS CENTER STRAP
2.5 cm WIDE, SET VERTICALLY

Fig. 9. Tuned diode detector, 160–460 MHz.

Calibration will present some difficulties, so line up some other lads around who are already on these bands and get your calibration that way.

So good luck, friends, more coming – lot's more! Keep reading.

UHF SIGNAL GENERATOR

Jim Kennedy K6MIO

Several requirements were the basis for the design of this signal generator: operation on both 432 mc and 1296 mc, high stability e.g., crystal control, variable rf amplitude, provision for insertion of various types of modulation, and lowest possible cost consistent with satisfactory operation.

I had a 27.005 mc overtone crystal left over from a brief period of disillusionment about CB and, since 27 times 6 is 432 which times 3 is 1296, this seemed a likely place to start.

Fig. 1 shows the final result. The overtone oscillator is voltage regulated and is left operating at all times when the generator is on. The crystal is mounted underneath the chassis to protect it from rapid temperature changes from drafts, etc., further enhancing the stability.

The use of diode multipliers at 432 mc and 1296 mc greatly simplifies developing UHF rf.

Direct coupling the modulation into the cathode of the last vacuum tube multiplier provides a modulation input that will accomodate almost any signal from audio to video, or even pulse.

A four position mode switch on the front panel allows the choice of carrier, no carrier, carrier with 60 cycle modulation and finally carrier with external modulation.

Construction

Building the unit presents no special problems except getting it all under the chassis. The 5 x 9½ x 2½ chassis doesn't leave much useful

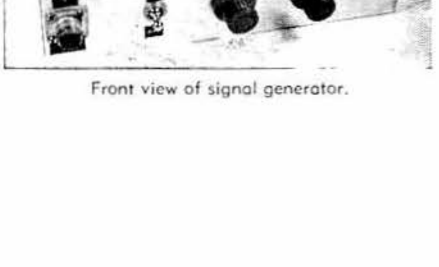

Front view of signal generator.

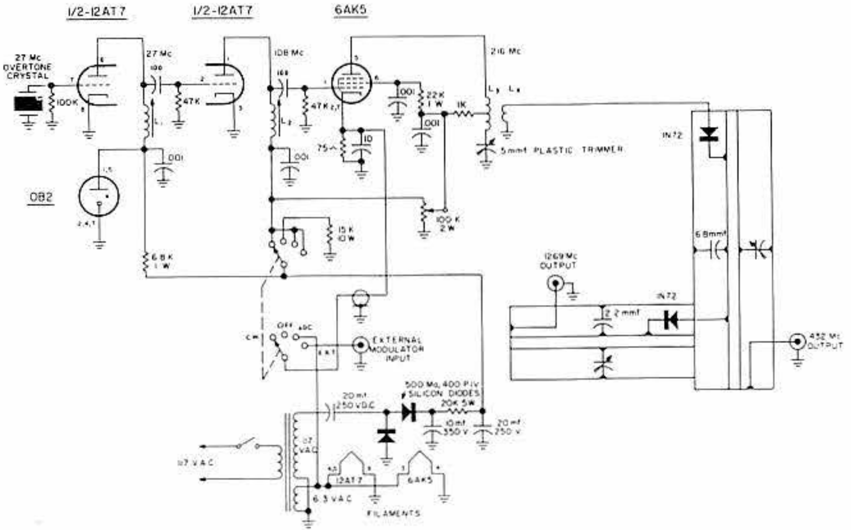

FIGURE 1

L1—10 turns # 28, 3/8" slug form.
L2—3 turns #28, 3/8" slug form.
L3—10 turns #18, 1/2" dia.
L4—3 turns #18, 1/4" dia.

room to spare, but it does make for a very compact device.

The lengths of the diode multiplier tuned lines were determined primarily by the space available and bear only the slightest relation to the wave lengths involved; hence, it was necessary to pad both lines with additional capacitances in order to resonate them properly.

Both lines were folded from 1/32 inch brass sheet stock. The inner conductors are made from ¼ inch brass tubing. The 432 line is 1" x 1" x 7". The input diode is tapped onto the inner conductor 1¼ inches from the other end. The signal output loop is about 1¼ inches long. The output diode is tapped 2 inches from the loop end of the line.

The tuning capacitor is made by soldering a 10-32 nut over a hole in the middle of the line shell. This provides the threads for the 10-32 screw which actually serves as the capacitor.

Round ½ inch disks are soldered to the bottom of the center of the inner conductor and to the end of the turning screw to provide additional range.

The 1296 line is similar except it is only 3⅜ inches long. The multiplier diode is tapped on the inner conductor ⅝ inches from one end. The signal output loop is about ¾ inches long and is located at the other end of the line. The tuning capacitor is identical to the one used in the 432 line.

The 15K 10 watt resistor in the carrier OFF position limits the no load voltage from exceeding the ratings on the filter capacitors.

Operation

The tune up of the tube stages can be easily accomplished with a grid dipper. The UHF stages should only need to be peaked with a converter and receiver.

Shields on all tubes and a bottom cover on the chassis are important requirements if stray leakage at the signal frequency or lower harmonics is to be eliminated as a possible source of measurement error. These were removed for the photographs but are always used in practice.

For best results it is a good idea to employ a 50 ohm (or 75 ohm as the case may be) pad between the signal generator and the equipment under test. This will insure that the generator will appear as a resistive load as much as possible. This is necessary if the generator is to duplicate the conditions presented to the receiver by a good antenna.

This pad can take many forms. The two easiest ways to make one would be to use sufficient lengths of some lossy coax. At 432 mc higher, 20 or 30 feet of RG-58 (or RG-59) seems to be quite adequate for most applications. The other approach would involve building a conventional resistor pad into a male and female BNC connector soldered back to back. Fig. 2 shows such a pad.

Aside from receiver testing, this device has several other uses to which it can be put. It is also useful as a signal source for antenna tests and it may be used as frequency standard.

This last use can be of great value at UHF because of the persistent problem of crystal tolerance versus frequency multiplication. A converter can easily be 40 or 50 kc or more off of the indicated frequency at 1296 mc. This can be quite a problem if you are hunting for a moonbounce signal through a 5 cps filter.

Though not shown on the diagram or in the pictures, it was found that a 50 mmfd variable capacitor connected across the crystal was sufficient to reduce 27.005 mc to the 27.000 . . . mc required to put the generator on 432.00 . . . and 1296.00 . . . mc. It will probably be necessary to retune the oscillator plate circuit for best operation; some loss in output will be encountered but this should not be a serious problem. However, if such operation is contemplated, it might be of some advantage to obtain a 27.002 mc crystal and employ a smaller variable capacitor.

In order to make the frequency standard really something of a standard, a 1 mc calibration oscillator should be employed. Some harmonic of the oscillator (5 mc, 10 mc, etc.) should be zero beated against WWV. The signal generator should then be zero beated against the 27 mc harmonic of the 1 mc calibrator—presto!—a 432-1296 mc frequency standard.

Though there are many refinements which can be made, the unit described has done a yeoman service in all the varied tasks asked of it.

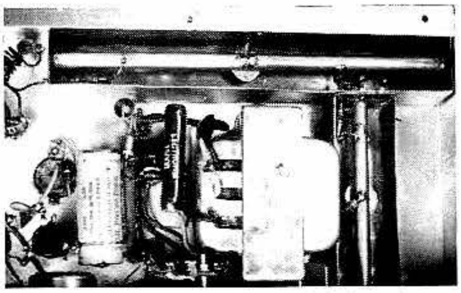

Closeup view showing details of diode multipliers lines.

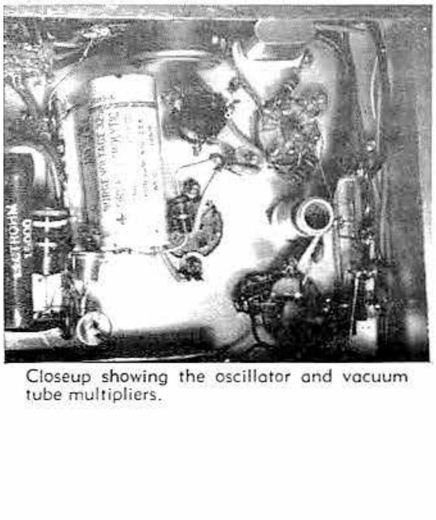

Closeup showing the oscillator and vacuum tube multipliers.

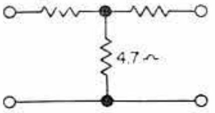

FIGURE 2

Fig. 2. 20 db 52 ohm pad. Resistors mounted in BNC connector.

LAB TYPE IF/RF SWEEP GENERATOR USING ICS

Ray Megirian K4DHC

Just about every kit-type sweep generator on the market starts life out at around 3 or 4 MHz and ends up at 200 MHz or so. This always made analysis or display of 456 kHz i-f amplifier response characteristics a tedious step-by-step plotting procedure or required time out to rig up a temporary gimmick to do the job. No more. The sweep generator to be described here will do the job with ease and precision. Even if you do not intend to duplicate the instrument, you may find parts of it useful for other applications.

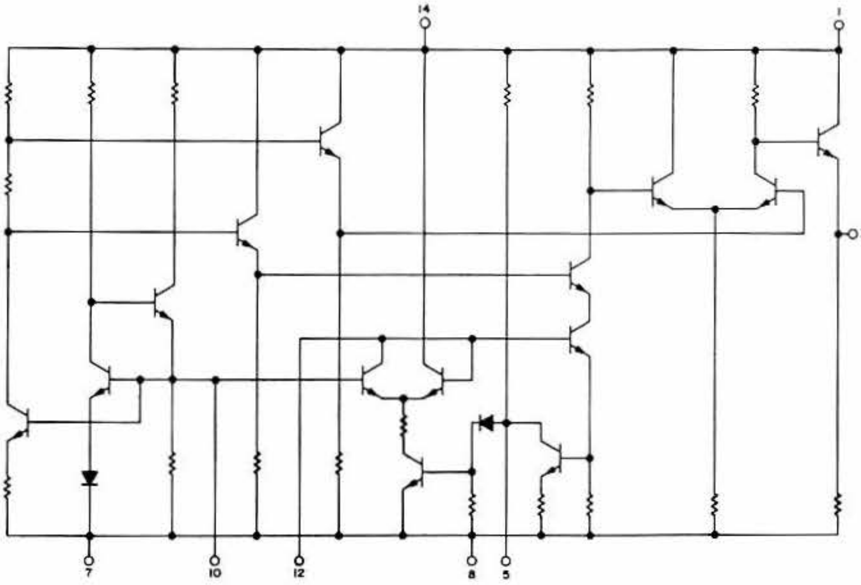

Fig. 1. Schematic of the MC1648 IC.

The generator covers a frequency range of from 400 kHz to 30 MHz in five bands. It can be operated in the CW mode as well as swept, thus allowing it to be used as an ordinary signal generator. Maximum output is 350 mV p-p across 50Ω. When sweeping, the return trace may be blanked or not, as desired. Two calibrated dials are provided for setting the start and stop frequencies and the maximum sweep width would be the entire band in use. The frequencies covered by the five bands are: 400 to 900 kHz, 850 kHz to 2.3 MHz, 2 to 6 MHz, 5 to 15 MHz and 10 to 30 MHz. Sweep time is variable between 20 ms and 6 seconds per sweep. A step attenuator in conjection with a vernier control provide a maximum attenuation of the output of 120 dB. An input is provided for a post-injection marker system built into the unit. A synchronous ramp with gain control is provided for driving the oscilloscope horizontal sweep. Blanking pulses are also brought out to a connector in case they are needed for synchronizing external equipment.

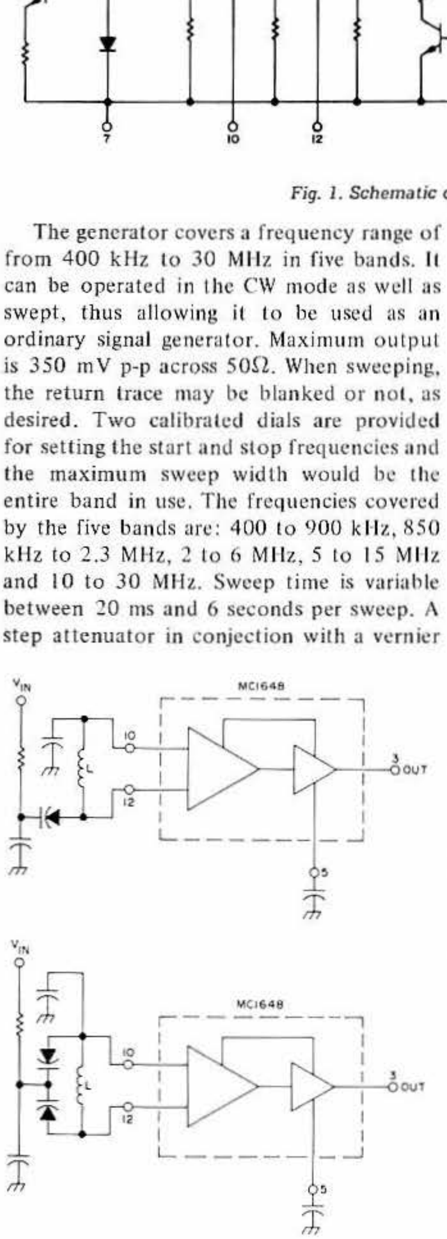

Fig. 2. Tuning the MC1648 with a single diode (A) and using back-to-back diodes (B).

A New IC

The heart of the generator is a new Motorola IC, the MC1648 emitter-coupled oscillator. It was intended for use in phase-locked loop systems operating in excess of 150 MHz, but may be used in many other applications such as this one. The device provides output of high spectral purity and incorporates an internal agc system which simplifies design of the sweep generator by eliminating the need for external leveling. A buffer amplifier and emitter-follower output are also incorporated on the chip, eliminating the need for external amplifiers. Figure 1 is a schematic of the MC1648 IC. Figure 2 illustrates two methods for tuning the oscillator. The device is packaged in a 14-pin DIP and requires a 5V dc supply. Since it was intended to be used with Motorola MECL III logic, either polarity is permissible. A positive supply is used here, connected to pins 1 and 14 with 7 and 8 grounded.

All the information covered by the data sheet for the MC1648 dealt with operation from 10 MHz up to about 180 MHz. Since I was interested in going as low as 400 kHz, I had to do some experimenting with tank circuits. My best results were obtained with the use of cup cores for the two lowest bands. I also found that ordinary molded iron core rf chokes of the miniature variety did an excellent job the rest of the way. Use of these tiny components made possible a very compact 5-band assembly.

The VCO

In order to facilitate shielding and simplify construction, electronic band switching is employed. Figure 3 is the schematic for the rf portion of the generator. The 2N4391 J-FET has a low 'on' resistance and works very well as a switch. One of these transistors is placed in series with each tank circuit

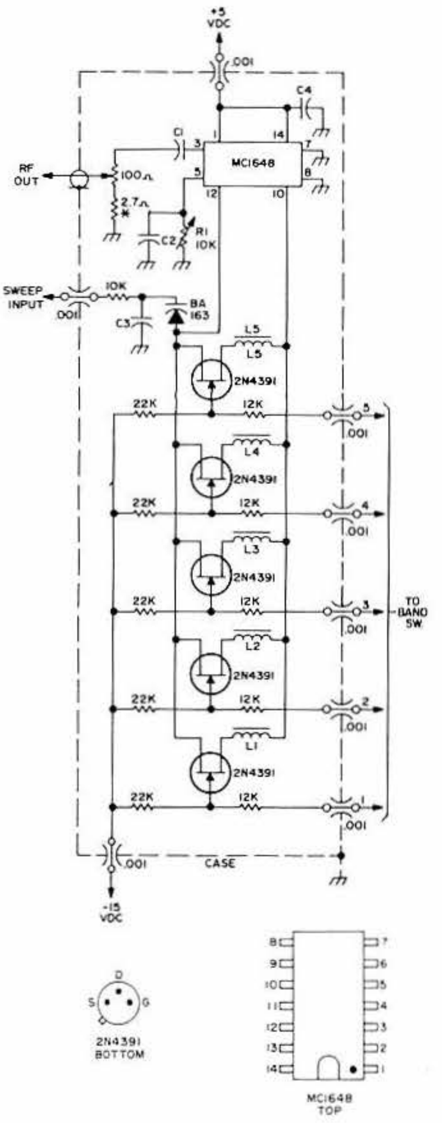

ALL RESISTORS-1/4W, 5%
L1-28 TURNS NO. 30 AWG ENAMEL.
L2-12 TURNS NO. 28 AWG ENAMEL.
(L1 & L2 MOUNTED IN CUP CORE-See Text)
L3-18 μH MINIATURE MOLDED RFC.
L4-4.7 μH MINIATURE MOLDED RFC.
L5-1.1 μH MINIATURE MOLDED RFC
C1, C2-TYPE 150D TANTALUM-2.2 μF/20V.
R1-BECKMAN 89PR10K 15-T TRIMMER.

*-SELECTED PART-See Text.

Fig. 3. Schematic of the rf assembly using dc band switching.

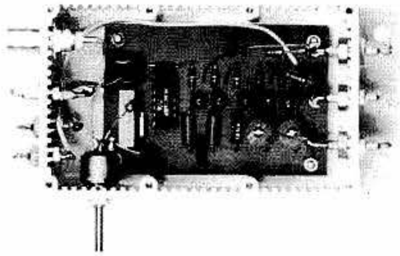

Completed rf assembly ready for mounting in the cabinet. A plastic version of the 2N4391 was used with a considerable savings in parts cost.

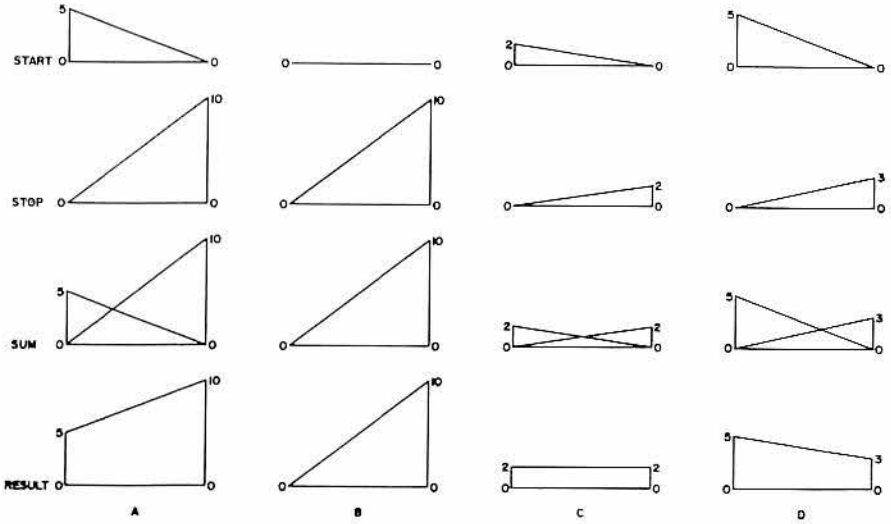

Fig. 5. Pictorial demonstration of ramp generation.

and can be turned on by application of a positive voltage at the gate. This allows the use of a strictly dc-operated remote switching arrangement. The entire assembly is built on a 2 x 3 in. pc board and mounted in a Pomona Model 3306 enclosure with a Model 3328 bottom mounting plate.

A BA163 tuning diode by ITT is used to sweep the oscillator. Although it is intended for bc band use, it performs admirably at these higher frequencies and its high capacitance ratio allows wide sweep excursions. If you are planning to use this oscillator for VHF applications, a more suitable diode would be in order. According to the data sheet on the MC1648, typical maximum output frequency is 225 MHz.

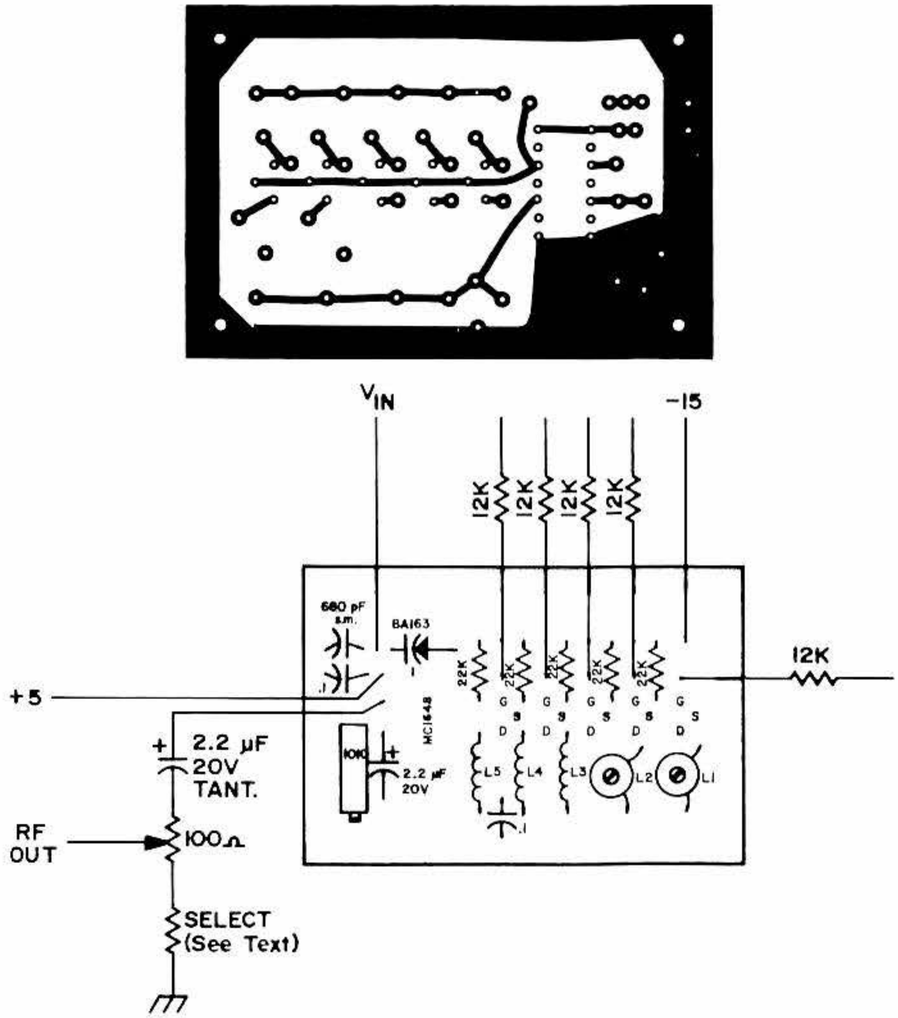

Fig. 4. (A) Foil side of rf pc board. (B) Location of parts on component side.

The vernier attenautor is a 100Ω RV6 style potentiometer mounted inside the enclosure. A coupling and extension shaft are used to bring the control out to the front panel. The value of the resistor at the bottom end of the pot will have to be selected so that a range of 20 dB is provided with full swing of the control. Since end resistance varies from one control to another, an exact value will have to be arrived at experimentally.

The output from the VCO is a square wave. If sine wave output is desired, the agc characteristic may be modified by introducing resistance between pin 5 and ground. The small trimmer incorporated on the board is used for this purpose. It was my experience, however, that this may introduce instability with some MC1648s. Although the trimmer was left in, the one in my unit is turned to maximum resistance where it has no effect and left there. It would do no harm to leave it out altogether since no other changes would be needed.

If any readers are contemplating exact duplication of this generator, you will find the cup cores I used are no longer available. These were Ferrox-cube part number 332P133B4-3C and are obsolete. I had a large number of these on hand left over from another project and saw no sense in buying more. I am sure some of the presently available cup cores will make excellent substitutes. Those used here are about 3/8 in. in diameter and are ungapped. You may have to experiment with the number of turns in the coil to obtain the coverage you require. One handy trick you can pull with a cup core is to rotate the slots in each half so that fractional turns are produced when the coil wires are brought out separately through the

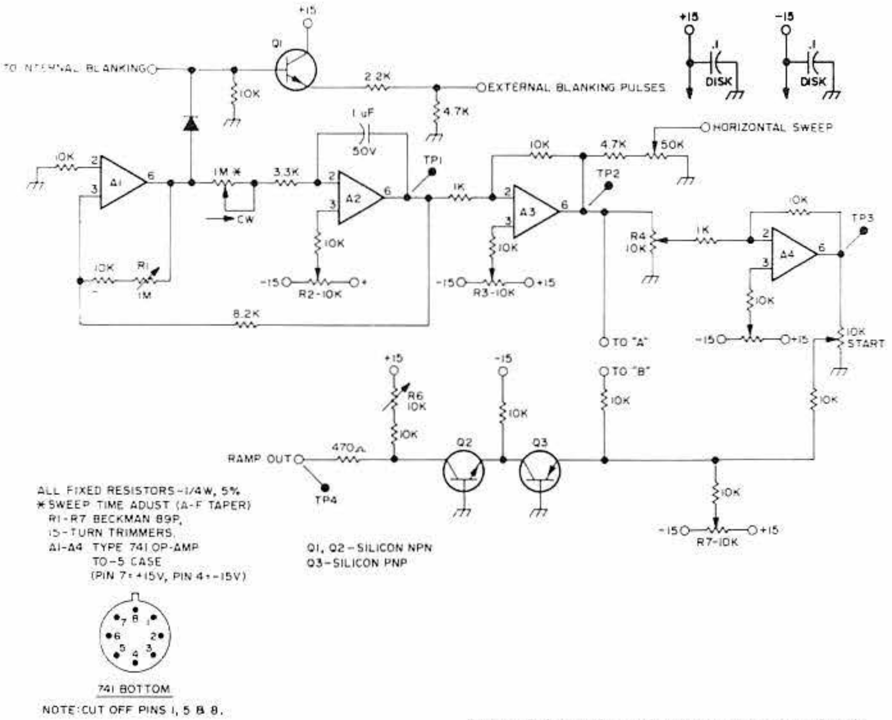

Fig. 6. Schematic of the ramp generator circuit.

two spaced holes. Exerting pressure on the core by means of the mounting screw will also shift the frequency and is almost like having a slug to tune. Increasing pressure seems to raise the frequency. The rf chokes used for the higher bands may also vary slightly from piece to piece and several may have to be tried to get the desired coverage. There isn't much we can do to alter the frequency where the chokes are used. Of course, slug-tuned coils may be used if the pc layout is modified. I do not recommend their use on the two lowest bands, however, since my results were rather unsatisfactory with this sort of tank circuit.

One final point would be in order before leaving the VCO. Note that the 10 KΩ resistor going to the tuning diode, the 2.2 μF rf output capacitor and each of the 12 KΩ resistors going to the transistor gates are not mounted on the board. One end of each has a hole provided on the board and then the component itself is used to make the connection to its final destination. This saves board space and eliminates separate wires. The small resistor at the bottom of the vernier attenuator is similarly mounted between the board and the bottom lug on the control. These points are more clearly seen in the photo. Also note that the cup cores are fastened directly to the board by means of 1-72 screws into threaded holes.

The Ramp Generator

In contrast with kit-type sweep generators that use the 60 Hz line for sweeping the oscillator, laboratory instruments have internal circuitry designed for this purpose. In our case, as with most sweep generators, a voltage ramp is used to control the oscillator during the sweep period. The circuitry has been arranged so that we can adjust the starting point (dc level) of the ramp independently of the stopping point. This simply means that we can set the frequency at which the sweep starts as well as the frequency at which it stops. We can also control, over a fairly wide range, the time it takes for a complete sweep. In this case, the sweep time is continuously adjustable between 20 ms and 6 seconds per sweep.

To get a better idea of how the ramp generator produces the results we've described above, let's look at Fig. 4. Note that a negative-going linear ramp (start) is combined with a positive-going linear ramp (stop) to produce the resultant ramp being applied to the varactor diode. If either input ramp is zero, the resultant will be the same as the single ramp since adding zero to anything will not change its value. Naturally, if both are zero, the output will be zero and the oscillator output frequency will be constant during the sweep period. The same is true whenever both ramps are of equal amplitude, except that the output will be a steady voltage other than zero. Also note that the resultant ramp could have a negative slope if the 'stop' input is lower than the 'start' input, and the oscillator would obviously sweep down in frequency rather than up.

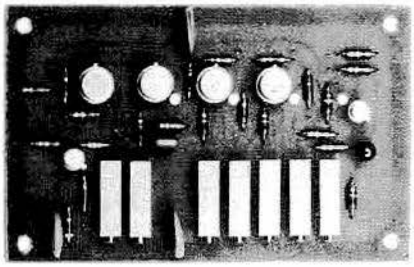

The ramp and blanking generator pc assembly. The leads going to external points will go in the holes visible on the board.

The schematic for the ramp generator is shown in Fig. 6. Operational amplifiers A1 and A2 form a triangle wave generator with A1 acting as a threshold detector and A2 as an integrator. One half of the triangle (negative-going) represents the sweep period while the other half represents the retrace time. Since a square wave is generated at the output of A1, we have a convenient source of blanking voltage built right in. This signal is positive during the sweep period and is connected to the band switch when Swept 1 (blanked retrace) operation is selected. During retrace the square wave drops to zero and actually disconnects the tank circuit until the next sweep starts. An emitter-follower, Q3, buffers the square wave output for external use.

The output of inverting amplifier A3 is a positive-going ramp during the sweep period and is used for both the 'stop' signal and horizontal sweep for the oscilloscope. Since an additional inversion takes place in the output of A4, the negative-going ramp at this point is used as the 'start' signal. These two ramps are applied across front panel controls which have 6:1 reduction drives and are fitted with dials calibrated in frequency for the five operating bands. The outputs from these two pots are fed to summing amplifier Q1, where the output will be the resultant ramp we discussed above. Because the output Q1 is in the negative region (PNP transistor), a second common-base (NPN) amplifier is used to shift the output back to where it will always be positive.

The ramp and blanking generator is constructed on a pc board 2.6 x 4.1 in. in size. For convenience, all the trimmer resistors were mounted along one edge. As seen in the photo, this board was mounted by means of two small brackets in a vertical position with the trimmers facing up. To simplify pattern layout, unused pins 1, 5 and 8 on each of the four 741 opamps were clipped off right at the case. Color-coded wires were connected to all necessary points on the board before it was mounted, with the leads made long enough to reach their destinations. For those interested in building this unit, a detailed procedure for setting up the ramp generator will be given later.

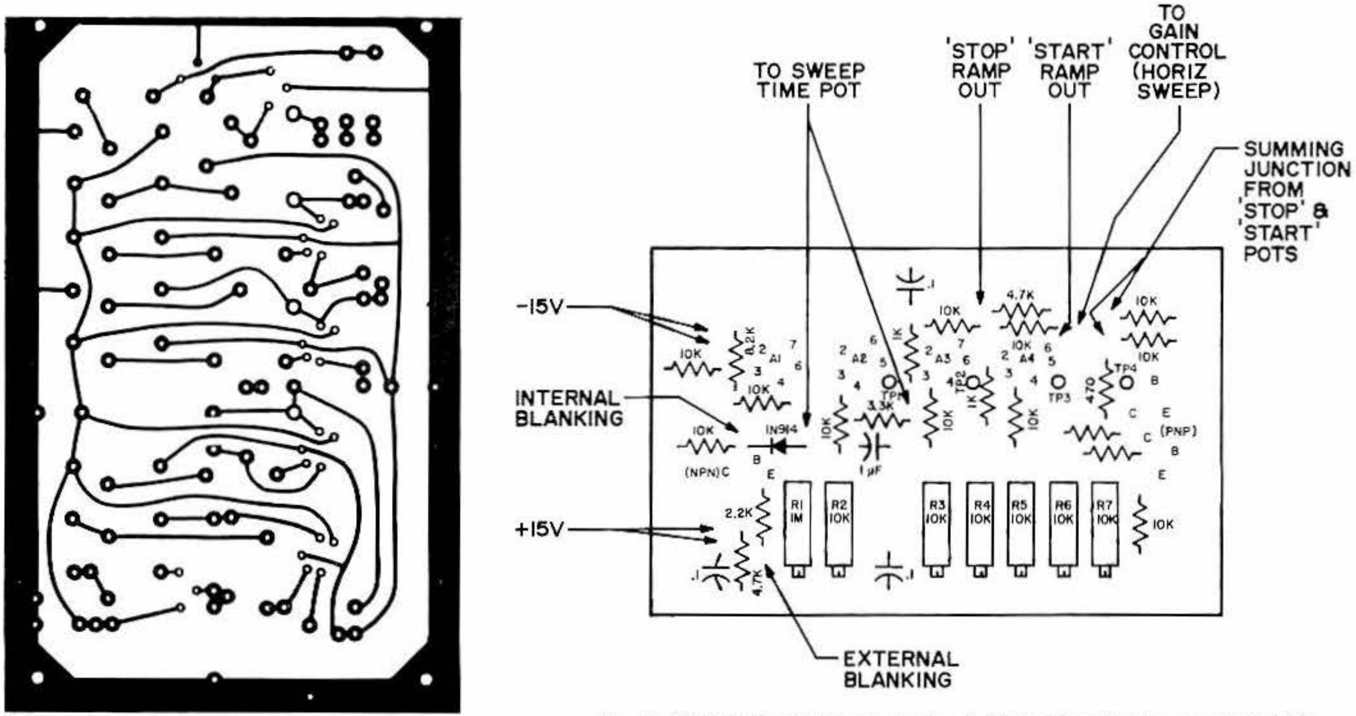
TO SWEEP TIME POT
'STOP' RAMP OUT
'START' RAMP OUT
TO GAIN CONTROL (HORIZ SWEEP)
SUMMING JUNCTION FROM 'STOP' & 'START' POTS
-15V
INTERNAL BLANKING
+15V
EXTERNAL BLANKING
10K
8.2K
A1
A2
A3
A4
1K
4.7K
3.3K
1N914
(NPN)C
B
E
1 µF
TP1
TP2
TP3
TP4
470
(PNP)
2.2K
R1 1M
R2 10K
R3 10K
R4 10K
R5 10K
R6 10K
R7 10K
.1

Fig. 7 (A) Foil side of ramp generator board. (B) location of parts on component side.

The Mixer/Amplifier

In order to provide a means for displaying frequency markers, a mixer and amplifier circuit was built into the sweep generator. The schematic for this circuit is shown in Fig. 6. A marker generator was not included inside the sweeper because other rf generators were available for use as markers and the added expense seemed unnecessary. Because the mixer board was to be installed inside a shielded box, dc gain control was resorted to as with the band switching discussed earlier. A 2N4391 is used as a variable resistor to control the signal input to the amplifier stage. This in turn controls the size of the marker being displayed on the curve. A 2N3819 FET is used in a simple mixer circuit which worked the best of several tried. Incidentally, the PNP and NPN transistors used here and on the ramp generator board are unknown surplus silicon units.

The pc board for the mixer is 1.5 x 2.9 in. and is mounted in a Bud CU-124 die-cast aluminum box. Note that here again some of the components are mounted between the board and their external connectors. These include both the 10 pF and 100 pF input capacitors to the 2N3819 and the 10 KΩ resistor going to the gate of the 2N4391.

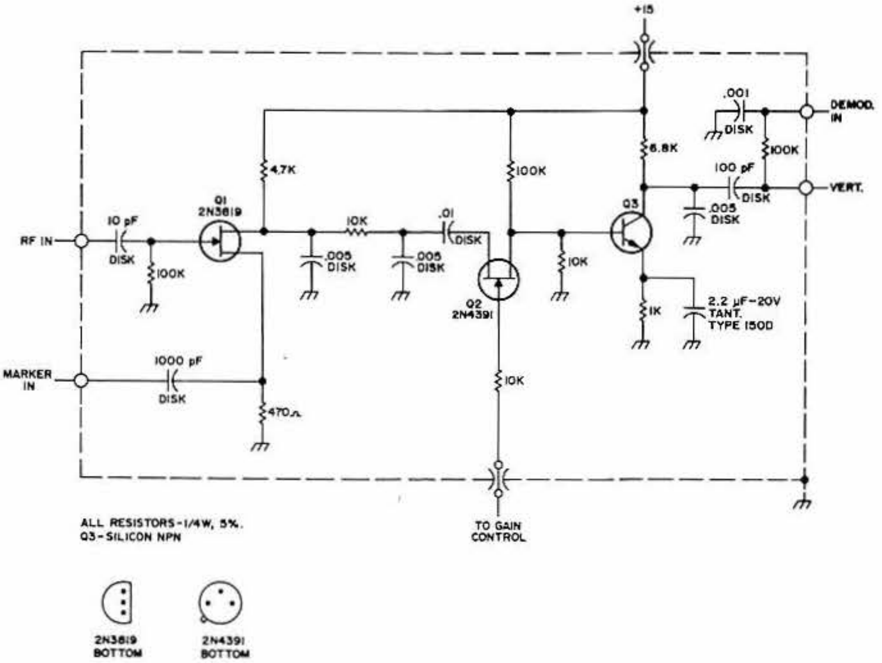
+15
RF IN
10 pF DISK
100K
Q1 2N3819
4.7K
10K
.005 DISK
.01 DISK
100K
Q2 2N4391
10K
Q3
6.8K
.001 DISK
100 pF DISK
DEMOD. IN
VERT.
1K
2.2 µF-20V TANT. TYPE 150D
MARKER IN
1000 pF DISK
470Ω
TO GAIN CONTROL
ALL RESISTORS-1/4W, 5%.
Q3-SILICON NPN
2N3819 BOTTOM
2N4391 BOTTOM

Fig. 8. Post-injection marker circuit with dc gain control of marker size.

The Power Supply

Figure 7 is a schematic for the power supply. Voltage requirements for the sweep generator are +15V, −15V and +5V dc. The current drawn by any of the above circuits runs no more than 30 or 40 mA maximum,

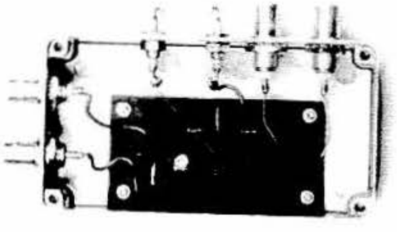

Completed mixer/amplifier board mounted in the die-cast housing.

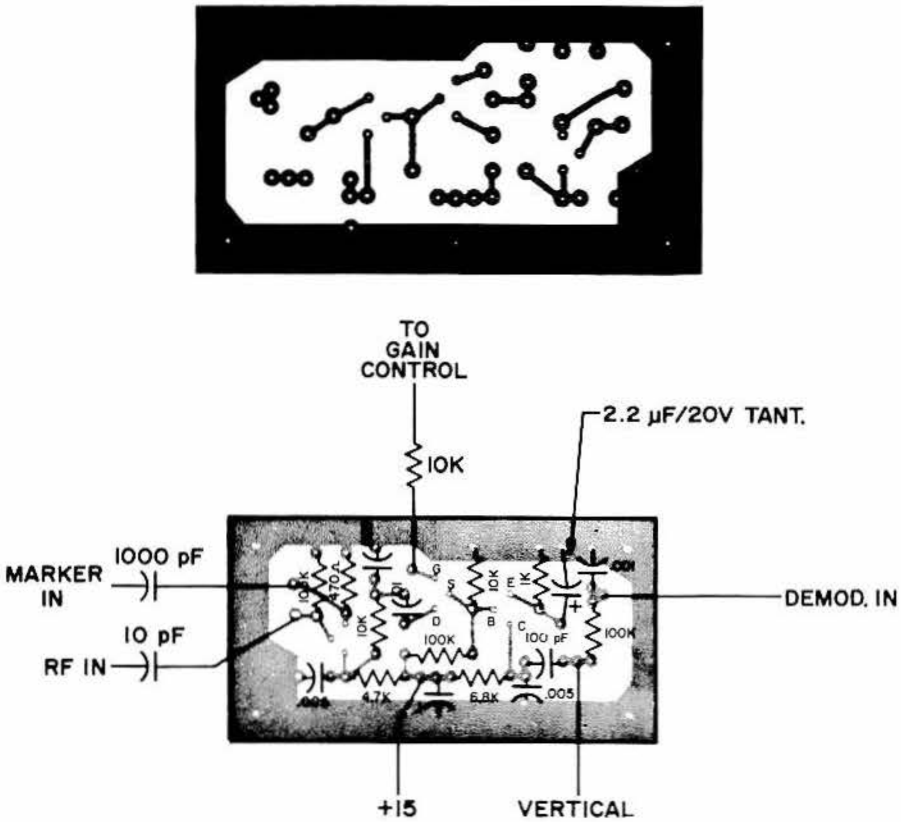
TO GAIN CONTROL
2.2 µF/20V TANT.
10K
MARKER IN
1000 pF
10 pF
RF IN
DEMOD. IN
10K
470Ω
10K
100K
100 pF
100K
.001
4.7K
6.8K
.005
.005
+15
VERTICAL

Fig. 9. (A) Foil side of mixer/amplifier pc board. (B) Location of parts on component side.

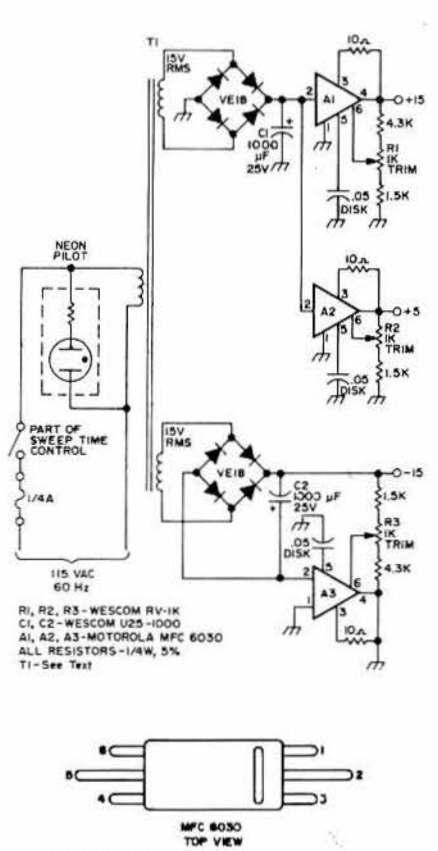
T1
15V RMS
VE18
C1 1000 µF 25V
A1
10Ω
+15
4.3K
R1 1K TRIM
.05 DISK
1.5K
NEON PILOT
A2
10Ω
+5
R2 1K TRIM
.05 DISK
1.5K
PART OF SWEEP TIME CONTROL
1/4A
115 VAC 60 Hz
15V RMS
VE18
C2 1000 µF 25V
.05 DISK
-15
1.5K
R3 1K TRIM
4.3K
A3
10Ω
R1, R2, R3 - WESCOM RV-1K
C1, C2 - WESCOM U25-1000
A1, A2, A3 - MOTOROLA MFC 6030
ALL RESISTORS - 1/4W, 5%
T1 - See Text
MFC 6030 TOP VIEW

Fig. 10. (A) Foil side of power supply pc board. (B) location of parts on component side.

making the use of small IC voltage regulators ideal. Motorola MFC6030 (plastic) regulators are used in each supply and are overload protected against accidental short circuit. Varo type VE18 molded bridge rectifiers, together with 1000 µF filter capacitors, supply the dc input to the regulators. The +15V and +5V regulators are both fed from the same source. The power transformer was a surplus unit with a 30V CT secondary. The center tap was uncovered and the two leads separated so as to provide two independent 15V windings. A small trimmer is provided in each supply for voltage adjustment.

The pc board for the power supply is 2.6 x 4.1 in. All components but the transformer are mounted on the board. The transformer is mounted directly on the chassis. Relative placement of the various components making up the complete generator can be seen quite clearly in the photo.

Control Circuits

Figure 8 is the wiring diagram for the control circuitry. Note that for CW operation of the generator, a dc voltage is applied to the varactor by way of the Start pot. In order for the dial calibration to be valid in either mode, the dc level applied to the control must be exactly the same as the peak amplitude of the ramp during swept operation. As will be explained in the set-up procedure, the ramp will vary from a starting point of approximately +1.4V and peak at +12V. Resistor R1 in series with the +15V line is selected for a +12V level at the high end of the pot. R2 is selected for a level of +1.4V at the low end. The reason for the offset of 1.4V at the low end is because the pin to which the tuning diode is connected at the MC1648, sits at this level. The source of this bias is the drop developed across two forward-biased junctions within the IC and not from any external source. Once these dc levels have been set for CW with the fixed resistors, the ramp can be made to match by means of the trimmers in the ramp generator circuit.

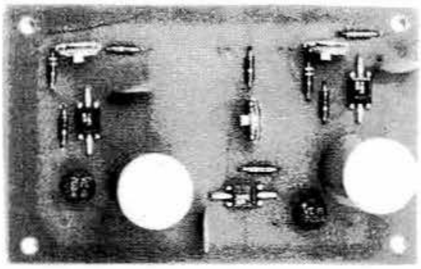

The power supply pc assembly.

The rest of the control circuitry is quite straightforward. In position 2 (Swept 1) of the mode switch, positive pulses from the ramp generator are fed to the band switch during the sweep period. When retrace occurs, the pulse drops to zero and the oscillator shuts off until the next sweep starts. In the CW and Swept 2 positions of the mode switch, a steady +15V is applied to the band switch and the oscillator runs continuously.

The final function performed by the mode switch is to route the dc for CW or the ramp for swept operation to the tuning diode. The two fixed resistors are mounted point-to-point behind the panel since all points are readily accessible within every short distances. The fixed resistors associated with the marker gain control are also mounted the same way behind the panel. The switches are ordinary rotary types and the potentiometers are all ordinary carbon controls. The Start and Stop controls as well as all gain controls have linear tapers. The Sweep Time pot is an audio taper type with built-in switch for ac power.

Setting Up the Ramp Generator

For best results, a calibrated dc scope is required to properly adjust the ramp generator. Test points have been provided on the pc board and short pieces of bare wire connected to each of these points make excellent tie points for the scope probe. The

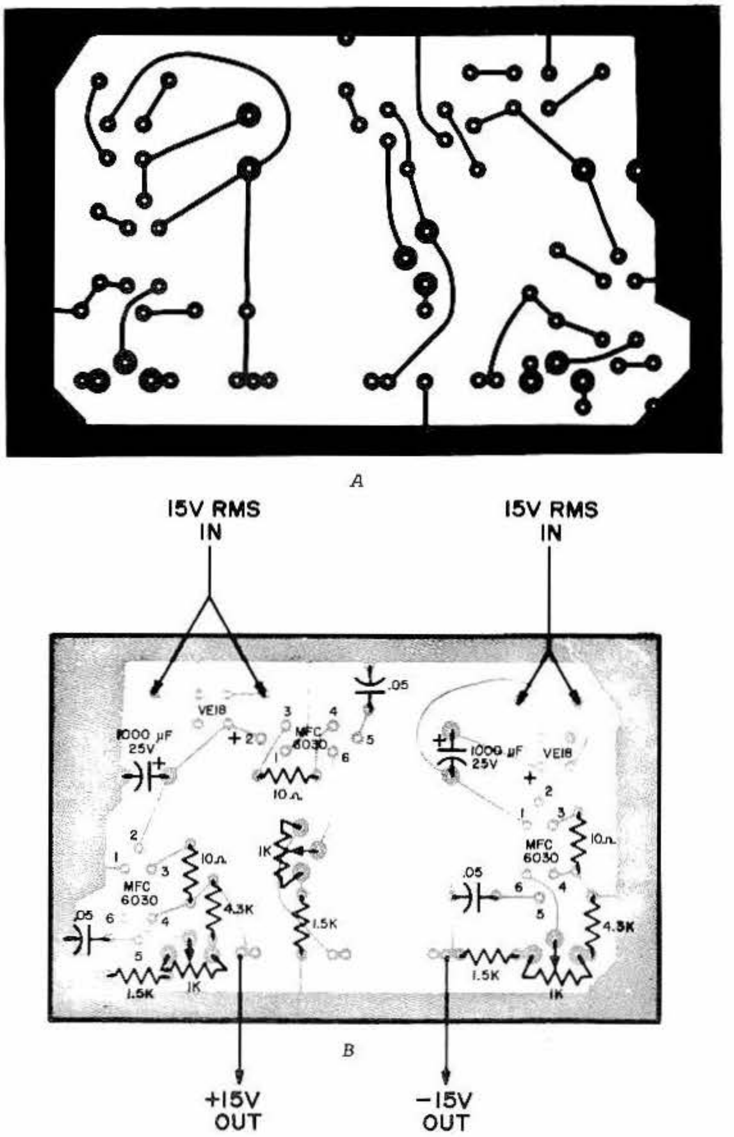

Fig. 11. (A) Foil side of power supply PC board. (B) Location of parts on component side.

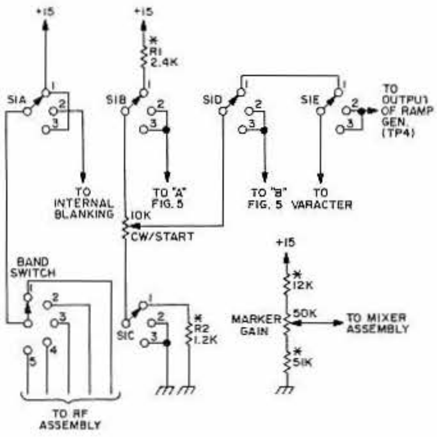

S1-MODE SWITCH

POS.	FUNCTION
1	CW
2	SWEPT 1 (BLANKING)
3	SWEPT 2 (NO BLANKING)

* SELECTED-See Text

Fig.12 Control circuit wiring schematic.

circuit may be aligned either before or after installation. The Start, Stop and Sweep Time pots can be connected at the ends of their respective leads if the unit is checked outside the cabinet.

Set all seven trimmers to mid-range. Connect the scope to TP1 and adjust the horizontal for a full sweep of 20 ms. Set the Sweep Time control to minimum resistance and apply power. Some sort of triangular wave should be displayed. Adjust R1 for a 3V p-p amplitude of the waveform. Turn R2 cw until the negative-going portion of the triangle is 20 ms long. Since R1 and R2 interact, you will have to stop occasionally and reset R1 for proper amplitude. Once the ramp is set at 20 ms with the Sweep Time pot at minimum, the slow speed end will automatically be about 5 or 6 seconds with a 1 MΩ pot.

Transfer the scope probe to TP2 where a positive-going ramp should be seen. Set the starting point of the ramp to zero volts by means of R3. Amplitude of this ramp must be 10V. If it is not, go back and adjust R1 slightly until it is. If necessary, reset R2 for 20 ms trace length. As soon as a 10V, 20 ms, zero-based, positive-going ramp has been achieved at TP2, go on to TP3.

At TP3 there should be a negative-going ramp. Once again we require a zero base line or starting point. Adjust R5 to accomplish this. Amplitude should once again be 10V and is controlled by R4. Once you have a 10V, zero-based, negative-going ramp, move on to TP4.

Connect the probe to TP4 and set the Stop control to full cw. The Start pot should be at minimum setting. A positive-going ramp should be present at TP4. By means of trimmer R7, move the ramp up or down until the starting point is approximately +1.4V. Ordinarily the ramp will not go below zero but will flatten out as the positioning control is adjusted. This is due to clipping in the output amplifier. If the ramp generator is going to be used with another type of VCO, the ramp can be set to exactly zero if desired. In our case we can't as explained previously.

Because a substantial amount of capacitance change occurs in the tuning diode up to 12V, the final ramp amplitude is set to 12V by means of R6. Because R6 stretches the signal in a negative direction, R7 will have to be used to reset the base line at 1.4V. Other than for checking tracking between the CW and swept modes, this completes the alignment procedure for the ramp generator.

When the generator is finished, you can check the tracking by observing actual signals at the input to the VCO. Connect the scope probe to the feedthrough capacitor on the VCO assembly that goes to the tuning diode and set the Start control to maximum. Switch rapidly between the CW and Swept 1 positions of the mode switch. The CW input will be a straight line sitting at +12V while the swept input will be a ramp peaking at +12V. Next, turn the Start control to minimum and again compare levels. If the Stop control is also at minimum, both signals will be straight lines at +1.4V. If there is any substantial difference between the two modes, use R6 and R7 on the generator board to make the ramp agree with the dc levels. That completes all phases of the ramp set-up procedure.

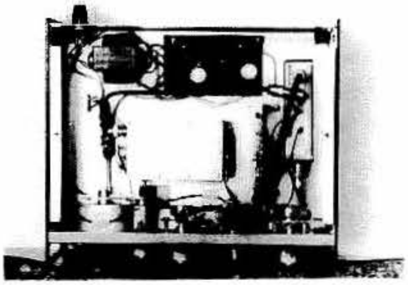

This bird's eye view of the interior shows the location of all major assemblies.

Construction

The cabinet used here is one manufactured by Sorensen Electronics in their Mod-U-Line series. These are the most reasonably priced instrument enclosures I've come across so far and I've used them for several projects. This one is a Model MCH-5129 with a CP-129 chassis plate. Dimensions are: 5¼ in. high by 12 in. wide by 9 in. deep.

The 5-step attenuator was picked up surplus from Fertik's Electronics for about $10. It is well made and designed for 50Ω systems operating up to 1 GHz. It has an integral female BNC output connector on the front face along with four threaded mounting holes for ease of installation. The input connector is a BNC male at the end of a short piece of coax.

All rf assemblies are interconnected inside the cabinet by means of coax cables. This is clearly evident in the photo.

The Start and Stop pots were mounted on brackets behind the front panel so that Jackson Brothers type 4511/DAF reduction drives could be installed for easier tuning. The two circular dials are slightly under 2 in. in diameter and were cut from sheet plastic. While operating in the CW mode, one of these was calibrated in pencil to provide a pattern for the finished product. A master was laid out using Rubylith® film and rub-on numbers. A negative of this was then made using 3M reversing film. From the negative a finished set of dials was printed on aluminum material of the presensitized variety. These were cut out and stuck to the plastic by means of their own pressure-sensitive adhesive backing. The nameplate was made from the same material. The two index pointers are clear plastic with hairlines scored on the inside surface. They are mounted on spacers directly over each dial. All remaining labeling was done with tub-on lettering.

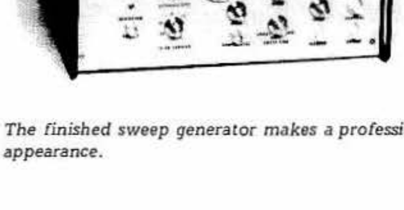

The finished sweep generator makes a professional appearance.

Vendor Addresses:

Fertik's Electronics, 9th & Tioga Sts., Philadelphia PA 19140.

Sorensen Electronics Co., Inc., 418 Queens Lane, San Jose CA 95112.

Wescom, P.O. Box 2436, El Cajon CA 92021.

100 KHZ THIN-LINE PULSE GENERATOR

James Ashe W2DXH

Ordinary 100-kHz frequency standards are usually audible up to a few tens of megahertz. A good one might be usable at 50 MHz. The circuit described here uses a dual NAND gate to generate a 100 kHz signal whose harmonics are usable to 432 MHz or higher. And it can be built without benefit of special instruments and knowledge.

The thin line pulse

One rather surprising result of higher mathematics is that all repetitive signals are composed of harmonically related sine and cosine waves. For example, the familiar square wave is composed of a fundamental frequency, which sets its basic repetition rate, and of odd harmonics only of its fundamental, which contribute to its square corners. If the harmonics' amplitude or phase relationship is upset, the square wave is distorted. This feature makes the square wave very useful for amplifier testing, but its harmonic content is not very good for frequency standard applications.

Now suppose that we start adding up signals of F, 2F, 3F, and so on, phased in so that they all reinforce each other once per cycle. Let's say thay are all the same amplitude. What would we get? See **Fig. 1A.**

The five equal amplitude sine waves peak simultaneously at the beginning of the fundamental's cycle. Everywhere else, until near the end of the cycle, they are more or less out of phase. Trying to see what will happen, we try adding the first two frequencies. **Fig. 1B**, the result, might suggest something to a mathematician.

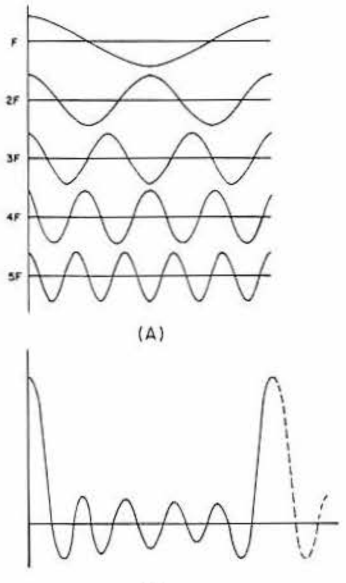

Fig. 1. Five sine waves (A) and the waveform as a result of point-by-point addition (B).

As the number of frequencies is increased, their amplitudes tend to average to zero everywhere except at the beginning of the cycle. Here, they all add up to a short, sharp pulse. It follows that a short, repetitive, one-sided pulse should contain odd and even multiples of the fundamental frequency.

An ideal thin line pulse has infinite frequency content.* No real signal could meet

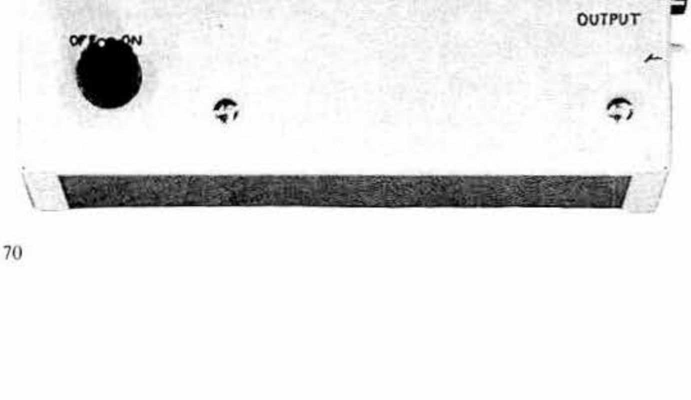

External view of the 100 Hz thin-line generator.

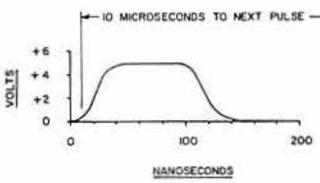

Fig. 2. Real circuit output as seen by a Tektronix 545A oscilloscope. A faster scope shows shorter rise-time and sharper corners.

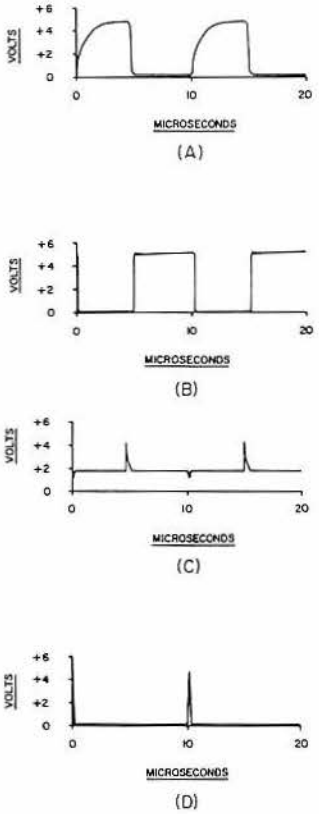

Fig. 4. Signals at four critical points in the generator, as displayed on a Tektronix 545A oscilloscope. They are shown in time coincidence.

this spec, but a fast digital IC can produce a very workable approximation. Fig. 2 shows a Tektronix 545A view of the generator output and tests with other scopes indicate the real pulse has better rise time and sharper corners than shown here.

If this pulse is viewed on a low-performance service variety scope, its appearance will be greatly changed. There will be an apparent loss in amplitude, since the pulse occurs and terminates before the slow circuitry can properly respond. The apparent duration is increased, also because of the slower viewing circuitry. And the fast pulse may excite circuit resonances, so that the thin line pulse appears as a damped oscillation. But these problems do not interfere with constructing the generator, because the very simple NAND gate circuitry contains no critical elements or adjustments.

How it works

There are four circuit sections, shown in Fig. 3. A 100-kHz crystal-stabilized oscillator sets the basic frequency, and a dual NAND gate circuit converts the oscillator output to a thin line pulse. A 1-Hz astable generates the output marking signal. A 6 volt dc power source is provided by a voltage doubler zener-regulated supply.

Multivibrator oscillators are not ordinarily very stable frequency sources. But if the oscillator is designed to run slightly below required frequency, and an appropriate crystal is connected between transistor base terminals, oscillations are stabilized at the crystal frequency.

The crystal does not change the multivibrator's style of operation. It synchronizes the astable to its own frequency, by triggering the OFF transistor into conduction shortly before normal RC turn-on. The output is a squarish wave with good fall time, but a long rise time as shown in **Fig. 4A.**

In passing through the first NAND gate the pulse is squared up and becomes slightly unsymmetrical. See **Fig. 4B.** A differentiating network, C7 and R11, converts the square wave into the pulses shown in **Fig. 4C.** These pulses, applied to the second NAND gate, reappear as the thin line pulses shown in **Fig. 4D.**

Since one CW signal sounds just like another and there may be several in the vicinity of a check point, a marker feature is required. This is provided by the 1-Hz astable, which paralyzes the second NAND gate part of the time. Its base bias resistors are unequal, giving a distinctive duty cycle to the output signal. A switch disables the astable if a continuous signal is required. **Fig. 5** shows the output when the second astable is operating: the output is locked in the up condition during half of each 1-Hz astable cycle.

Sometimes an astable oscillator will refuse to start oscillating when it is turned on. It does not start because both transistors are in saturation. This reduces loop gain so that available noise cannot be amplified around the loop. It would never start without some strong, outside interference.

A pair of diodes, D1 and D2, provide a reliable remedy. The diodes are arranged so that base bias must come from whichever collector is at the higher voltage. If both transistors are in saturation, their collectors are at perhaps 1 volt, which cannot provide enough base current to keep the transistors in saturation. This contradictory situation does not arise in the real circuit, which starts reliably.

Additional diodes, D5 through D8, appear in the base circuit of the 1-Hz astable. These are protective diodes. The collector swing at turnoff of about 5 volts is conveyed powerfully to the opposite base through the large coupling capacitors C5 and C6. The reverse B-E breakdown voltage of these transistors is not known, so the diodes are provided to prevent the turnoff voltage exceeding 2 volts or so.

DC power for the Generator circuitry comes from a voltage doubler supply based on a low-current filament transformer. Its design is conventional, but a large capacitor, C12, is provided across its output to mini-

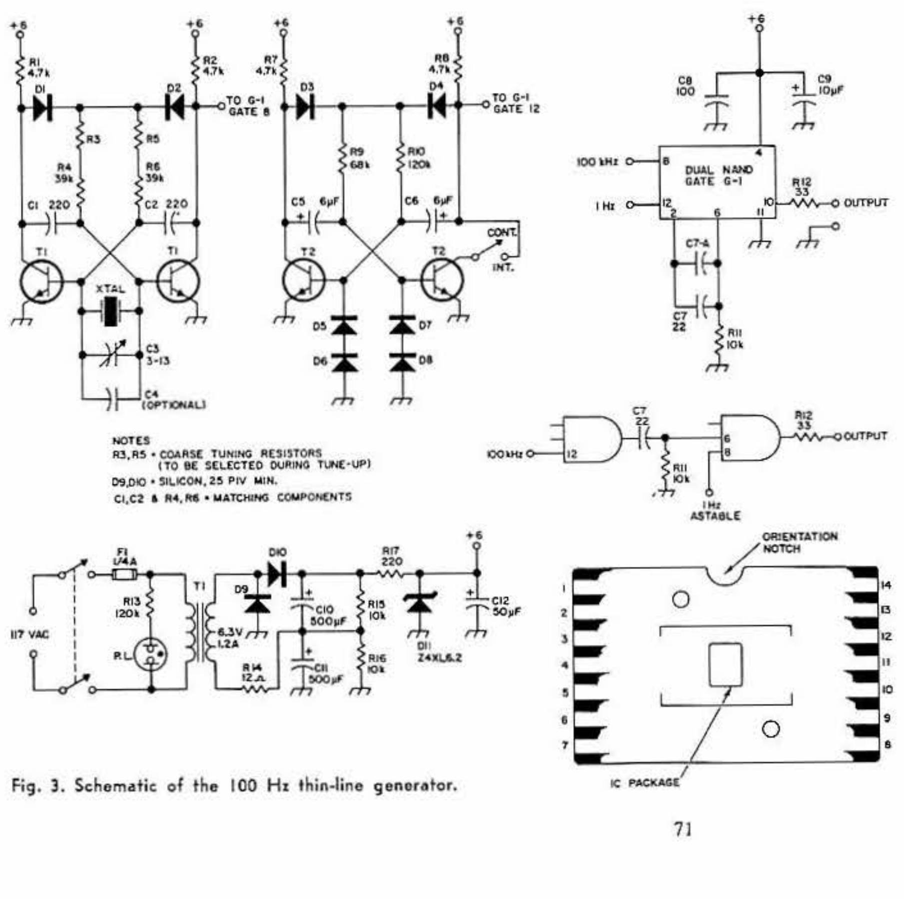

NOTES
R3,R5 - COARSE TUNING RESISTORS (TO BE SELECTED DURING TUNE-UP)
D9,D10 - SILICON, 25 PIV MIN.
C1,C2 & R4,R6 - MATCHING COMPONENTS

Fig. 3. Schematic of the 100 Hz thin-line generator.

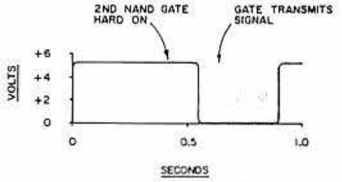

Fig. 5. The second NAND gate locks in its up position part of the time to produce an intermittent output.

mize noise on the supply line. The supply could be replaced with some batteries, shunted by a 50 μF or larger capacitor to absorb transients. The original breadboard ran very well, powered by four flashlight batteries.

Construction

The generator is built in a Premier #PMC 1008 3x5x7 inch heavy aluminum box. Its top cover was refinished in light green enamel, and four ⅜ inch grommets in the bottom piece serve as protective feet.

Inside the box, the 6.3-volt transformer and cheater cord connector are mounted on the left-hand wall. A pilot lamp, fuse, and two switches are mounted on the horizontal panel, at the extreme left. This leaves just enough open space for the two circuit boards which occupy most of the box. Two banana jack output connectors are placed on the right-hand side, just below the panel.

The circuit boards are cut to 4½ x 5 inches, from Vector ³⁄₃₂ inch pattern A stock and mounted parallel to the panel. The upper board is spaced an inch from the panel, and carries both astable oscillator circuits. The other board is mounted one half inch below, and carries the digital IC and the power supply circuitry. Assembled, the two boards make a sandwich with wiring sides together.

Both boards are mounted on the same four centers. These are through the second hole diagonally inward from each corner. The 1 inch 6-32 internally threaded spacers are modified by adding a short length of 6-32 threaded shaft to one end of each, simplifying assembly.

Component assembly on the boards is largely a matter of plugging in Vector T9.4 lugs. The finished product looks much better if some thought is given to facing the lugs in one of two directions. Mounting and transistor holes should be drilled and reamed to size before installing lugs.

The general arrangement puts all wiring on one side of the board and practically all components on the other side. This approach seems a little inflexible but is straightforward and looks good.

Possible board orientation problems may be overcome by working out a handling and wiring procedure that doesn't require constant reference to actual components. A good approach assumes that the board is only turned over an imaginary hinge at its bottom edge, so that top down when one side is up becomes bottom up when the other side is down. This preserves left-right relationships. Another useful convention is that all supply wiring goes to left-hand end of components.

Wiring is carried out one network (plus supply lines; ground lines; interstage lines, etc.) at a time, with prearranged color coding. Bare wire goes for short runs and where there is no chance of a short. Solder each lug when convenient. #22 solid wire fits the T9.4 lugs well, but flexible stranded wire is used for the four lines from one board to the other.

Transistors precede other components into the board, because they are convenient position markers. They are placed in their mounting holes in the board from the component side, and their leads brought to the T9.4 lugs. Then the other components are mounted on the boards. Diode and electrolytic capacitor mounting polarity should be double checked. The T9.4 lugs may need a little bending before they will take a good grip on the components, but no component soldering is done until everything is installed.

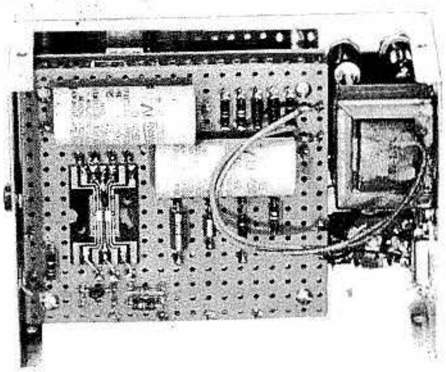

Inside the assembled thin-line generator showing the component side of the power supply and IC board.

Trimmer capacitor C3 is mounted on its tabs just under the top panel. Then a small screwdriver access hole is drilled over it in the panel, before painting, for vernier frequency adjustment after final assembly.

Certain components are matched before installation. An ohmmeter and a capacitor checker will do a satisfactory job of selecting C1 and C2, and R4 and R6, for equal values. These components are chosen alike for best symmetry of the 100-kHz oscillator operation. It might be good planning to leave these components unsoldered until tuning is completed, but everything else can be soldered to the board at this point. Note that the R3 and R5 sites do not get resistors until later.

Two optional capacitor sites are included. These are for C4, an additional and probably unnecessary padder across the crystal; and C7A, which can be added to increase the width of the thin line pulse.

Apparently, the digital IC comes in a specially designed package for testing before use. To mount the IC, solder a ⅜ inch piece of #22 wire in each of the T9.4 lugs carrying supply and signal voltages to the IC. Place the IC between the two rows of lugs, bend the wires against the proper terminals, and solder. No other mounting is required.

The original breadboard showed a lot of transient noise in its supply circuit. This originated from the IC, which was trying to get big chunks of current to manufacture pulses. Since the IC cannot deliver frequencies not available from the supply lines, very careful bypassing is indicated.

High-frequency bypassing consists of C9, a .01μF disc ceramic capacitor across the IC supply terminals on the wiring side of the board, and C10, a 100 picofarad capacitor soldered directly between supply terminals on the IC. The capacitor leads are provided with spaghetti insulation and placed for minimum open space between the capacitor leads and the IC's supply leads.

Testing before final assembly is very easy, because the odd appearing board layouts go together giving a structure that opens out like a book. The hinge is the four leads between boards. Leave transformer leads long, so that the circuit may be tested well free of its cabinet.

The upper half of the Premier box is prepared by a powerful cleaner which removes its original paint. After thorough removal of the cleaner, the metal is roughened with wet sandpaper, rinsed in vinegar solution and then clear water, leaving a very good surface that does not require priming for excellent paint adhesion. Watch out for greasy fingerprints.

Rustoleum #868 Green applied from a convenient spray can gives a fine finish. Follow instructions on the can. After drying, the fresh, clean enamel will take waterproof India ink, applied with a Leroy drafting pen. When the ink is thoroughly dry, a final coat of Rustoleum #717 Clear finishes the job. The enamel is soft at first, but hardens into a coat durable in normal lab use.

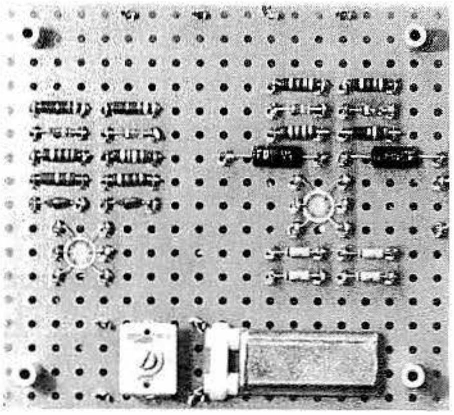

View of the component side of the astable oscillators board.

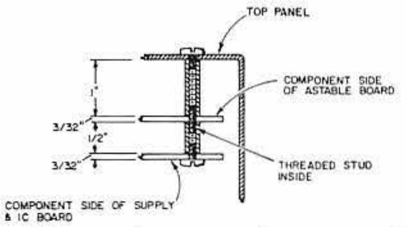
TOP PANEL
COMPONENT SIDE OF ASTABLE BOARD
1"
3/32"
1/2"
3/32"
THREADED STUD INSIDE
COMPONENT SIDE OF SUPPLY & IC BOARD

Fig. 6. Mounting dimensions and spacer assembly diagram.

Table of special parts

Crystal: 100 kHz parallel resonant 32 pF. shunt capacitance normally designed quartz crystal.

The following parts were obtained from Solid State Sales, P. O. Box 74, Somerville, Mass. 02143.

T1 & T2: 2N2060 type dual NPN transistor

D1, D2, D3, D4: fast point-contact Germanium diodes coded 1N59

D5, D6, D7, D8: fast point-contact Silicon diodes marked S284GM

G1: surplus digital integrated circuit Solid State Sales type G1. (comes with data sheet)

Tuning up

The generator should be zeroed to frequency before installation in its case. This is a two-step process. First, the 100-kHz astable base resistances are adjusted by choosing resistors for R3 and R5 to bring the oscillator frequency within trimmer range of 100 kHz, perhaps a few hundred cycles high at 15 MHz. Then the trimming capacitor brings the frequency to accurate coincidence with WWV.

To roughly zero the generator, set the trimmer capacitor, C3, at minimum capacitance. Identify WWV on a short-wave receiver, and tune around a bit to familiarize yourself with what's happening in the vicinity. It would be nice if things are fairly quiet.

Then put 4.7k resistors into the astable board at the R3 and R5 sites, turn on the generator, and look around for the signal. Depending upon actual values of C1 and C2, the signal may be on either side of WWV but is likely to be on the high side. If so, try again with resistors one size larger, which will lower the frequency. You should shortly find resistors that bring the frequency near enough to WWV for final zeroing with the capacitor. Verify tuning range on both sides of WWV.

Correct values for R3 and R5 may be approximated very quickly if a good triggered scope is available. Try selecting resistors for a period of 11.4 microseconds with the crystal removed.

Using the thin line generator

A breadboard test showed that (as might have been expected) there should be some way to distinguish generator signals from other CW signals. The continuous/intermittent feature provides the marking, and once the correct signal is located the generator can be switched to "continuous" for accurate work.

At low frequencies, the generator output and behavior resembles a conventional 100-kHz standard. The signal simply is not as strong. A greater difference appears at higher frequencies: the original model yields an audible beat note at 80-MHz from a diode mixer through an inexpensive audio amplifier. And another test shows a usable signal at 432 MHz: the 4,320th harmonic.

Some connection to the receiver or other detector is required. This is a natural consequence of a circuit design that puts the signal where it belongs, rather than spraying it all over the lab. A few picofarads coupling capacitance is sufficient at all frequencies.

Perhaps this circuit can be used for purposes other than a frequency standard. Its moderate amplitude but wideband output should be ideal for detecting changes in receiver sensitivity over a broad tuning range. In fact, with a little decoupling of the input leads and provision of a coax output connector the generator should do well as a stable, reliable small-signal source. A piece of adjustable waveguide-below-cutoff would make an excellent attenuator for work not requiring exact measurements. Another thought that occurs is possible further development by provision of some arrangement for detecting which harmonic is actually being heard.

Chapter VI

Crystal Calibrators

ALL BAND BAND-EDGE MARKER

Charles Berner WA2HRZ

This calibrator is complete with its own ac supply, eliminating the need for taking power from the receiver. If ac outlets are at a premium at your shack, the on-off switch and the line plug can be eliminated and the ac line connected across the receivers ac input so that the calibrator comes on whenever the receiver is turned on. A high density selenium rectifier was used and this contributes greatly to the compactness of the unit. The whole thing runs very cool and even after 24 hours operation, still isn't hot.

Wiring isn't too critical, but keep all leads short and direct. Using an octal socket for a xtal holder lets you use the unused pins as tiepoints and is recommended.

A Petersen type Z-2 xtal was used because it has a tolerance of .002%. This is one reason why no provision is made for "zeroing" the calibrator against a standard. The other is that there isn't any such standard as WWV for use. However, if this provision is desired, C1, the 56 mmfd capacitor, connected to pin 1 of the 6BH6 thru the 18K resistor, can be made variable. Also, if band edge markers are wanted for 50 mc and up, a 5 mc xtal may be substituted when this function is wanted and C1 made variable for "zeroing" against WWV.

This gadget can also be used as a xtal activity checker by simply replacing the marker xtal with the xtal to be tested. Tuning your receiver should get you a strong signal at the xtal's fundamental if it is OK.

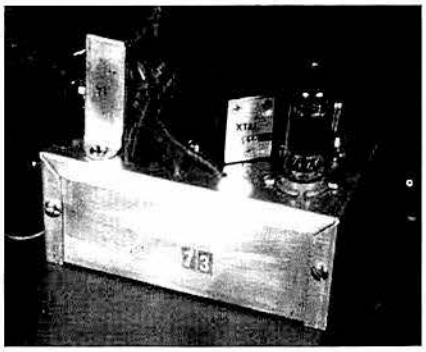

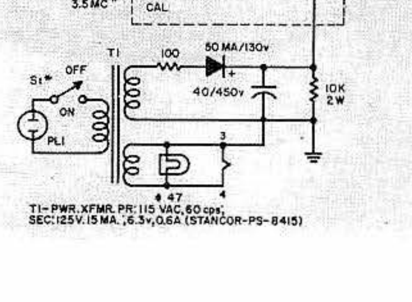

Table 1:

	Harmonic		
Band	3.5 mc	100 kc	Marker Frequency
80 meters	1	35	3.5 mc
40 "	2	70	7.0 mc
20 "	4	140	14.0 mc
15 "	6	210	21.0 mc
10 "	8	280	28.0 mc
6 "	15	525	52.5 mc
2 "	42	1470	147.0 mc
1¼ "	63	2205	220.5 mc
1¼ "	64	2240	224.0 mc

THE MULTICAL

E. R. Davisson K9VXL

What is the "Multical"? As the name implies, "multi" would suggest several uses, and "cal" might infer a calibrator of some sort.

Well, that's right, but there is slightly more significance to the name. "Multi" is also a short form term used to describe flip-flop circuits known as multivibrators.

By combining the basic characteristics of a free-running multivibrator (astable) with crystal control, you have a simple, stable, virtually-insensitive-to-temperature-changes, crystal calibrator for that receiver you have been wondering about.

The circuit uses no inductors and depends upon the crystal for the proper feedback for oscillations. Temperature stability is partially due to the absence of capacitors.

Transistor stage Q_2 operates with unity gain, whereas transistor Q_1 operates at considerably more gain. Both stages are operating as feedback amplifiers. The harmonic generator diode D_1 is a 1N128. Any general purpose diode may be used.

By using the multivibrator circuit, the waveform obtained is comparatively rich in harmonics and could be used without any further refinements. However, to insure useful harmonics through 30 MHz starting from a 100 kHz crystal, a harmonic generator consisting of R_6 and D_1 shown in **Fig. 1** was added. The capacitors C_1 and C_2 are used strictly for coupling and have no effect on frequency stability.

Crystals from 100 kHz up to 1 MHz may be used in the Multical with no changes. The circuit will oscillate from voltages as low as 2 volts and can be operated safely from voltages as high as 20 volts. This wide range of voltage operation allows the source to be obtained from virtually any place.

Output from the calibrator may be fed directly into the receiver's input, or may be coupled to a short whip antenna. With a whip antenna, close coupling to the receiver's input

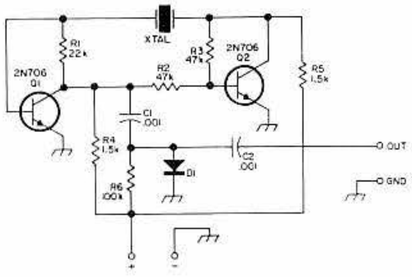

Fig. 1. Schematic of the Multical.

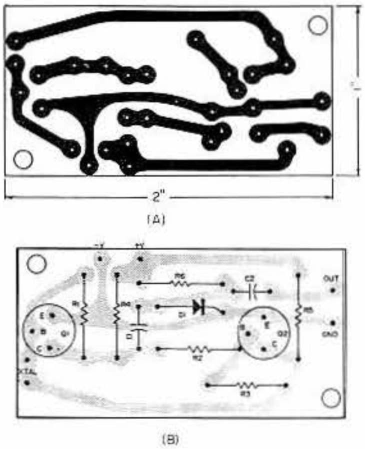

(A)

(B)

Fig. 2. Suggested printed circuit board layout for the Multical. A gives the copper side, B the component side. A board for the Multical is available for $1 from the Harris Company, 56 E. Main Street, Torrington, Conn.

may be required at higher frequencies. (Especially at the lower voltage levels.)

For the more ambitious builders, **Fig. 2** shows the printed circuit board layout for the Multical. Due to its small physical size (1" x 2"), room can probably be found even in the most compact of receivers. **Fig. 2A** shows the foil side, and **2B** shows the parts placement.

So the next time you wonder about the accuracy of your receiver calibration, give this simple circuit a try and you'll know for sure.

100 KHZ MARKER GENERATOR

W. W. Davey W7CJB

This useful piece of equipment generates usable harmonics from 100 kHz to 225 MHz. It is completely self contained and portable which makes it convenient not only to use in the ham shack, but also in the mobile unit or at a field day location. Its use lies mainly in accurately spotting band edges and 100 kHz calibration points throughout the ham bands.

Most modern day home receivers are equipped with calibrators, but these calibrators are of little use when needed to spot frequencies on VHF and UHF converters or portable equipment.

The generator is constructed in a 2¼ x 4¼ x 1½ inch handi-box. The parts are mounted on a vector board, and the entire unit is powered by one #216 nine volt battery or its equivalent.

Hints on construction First obtain some vector board. The piece I used was cut from the board supplied in a "GE experimenters aid hobbyist kit." The board must be cut to size before construction and will measure 3½ x 2 inches. This will allow room for the 9 volt battery in the end of the handi-box. Make sure the newly cut vector board will fit inside the handi-box before you start mounting parts. It might save a lot of trimming at a later date.

The parts layout is not critical. Components may be arranged as shown in the photo or in any other arrangement suitable to the components you may be using. I used sockets for the transistors, as I wanted to be able to experiment to see which transistors from my junk box would give the most output in the VHF and UHF bands. I ended up using the 2N404 for the oscillator and a 2N384 in the multiplier stage. I also found that Japanese 2SA83 transistors which had been removed from the *if* stages of a junked transistor radio, would work equally well in both sockets.

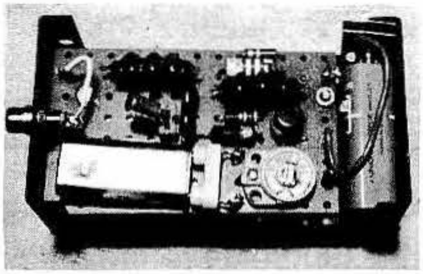

All components are mounted on the top of the Vector board with the exception of C4 and C5. For the most part, wiring can be completed with existing leads on components. The push-in terminals furnished with the GE experimenters kit were used for the battery connections, antenna connection and for mounting the crystal socket. The circuit board can be mounted to the handi-box with three 11/16 inch bushings. This leaves room for a slide switch to be mounted on the cover of the handi-box. Two of these bushings were purposely placed at the end of the board to form a sort of socket to hold the 9 Volt battery. The antenna output connector which is mounted on the handi-box is a switchcraft #3501FP phono jack. A small hole may be drilled in the bottom of the handi-box through which a screwdriver may be inserted for adjusting C1. For extreme accuracy C1 is adjusted to zero beat with WWV.

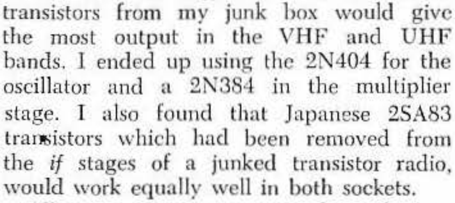

A 36 inch piece of insulated wire soldered into a phono plug may be inserted into the phono jack and used as a test antenna. The intensity of the markers may be varied by moving this test antenna near your receiver antenna lead-in. If you are using a coax lead-in you can couple by drilling a small hole in your coax relay so that the test antenna can be inserted near the relay armature. You will find that most SWR meters provide an easy method of coupling to the center conductor of the coax. As a last resort you can always couple to the receiver or converter antenna coil.

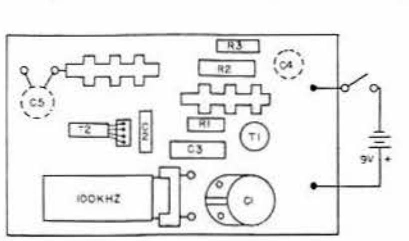

I have made very good use of this gadget to spot frequencies in the 144 and 220 MHz bands. It was well worth the time and effort it took to build it.

ALL BAND FREQUENCY MARKER

Kenneth W. Robbins W1KNI

Crystal controlled marker generators are useful adjuncts in any frequency determining situation requiring high accuracy, such as locating band edges, sub-bands and calibrating receivers. If you've been entertaining thoughts about construction of one, a version is described here which uses the new C/MOS integrated circuits powered by a 9 volt transistor radio battery. And instead of the usual rotary harmonic selector switch, a multi-pin IC connector strip and three test plugs serve as a miniature patch panel to enable various divisions of the reference crystal, with a maximum countdown of 256. "Rocks" from 100 kHz to 4 MHz oscillate readily in this circuit. In this model an FT241 xtal set to 400,000 Hz has been chosen for control and has usable receiver calibration divisions down to 2.5 kHz. The harmonic spectrum extends to at least 160 MHz, the tuning limit of a transistor superregen used in testing. When used in densely occupied HF bands, an AM beeper can be switched on as an identification aid.

Divide by	Output, kHz
1	400
2	200
4	100
8	50
10	40
16	25
20	20
40	10
80	5
100	4
160	2.5

Table 1.

Referring to Fig. 1, one third of a hex inverter makes up a crystal controlled oscillator and buffer, another third is a slow rate pulser and the two remaining units function in the dividing section. These are all standard circuits described in RCA's COS/MOS Data Book #SSD-203. An emitter follower minimizes loading on the IC outputs, speeds up rise time to increase harmonic content, and provides a low impedance output. The AM

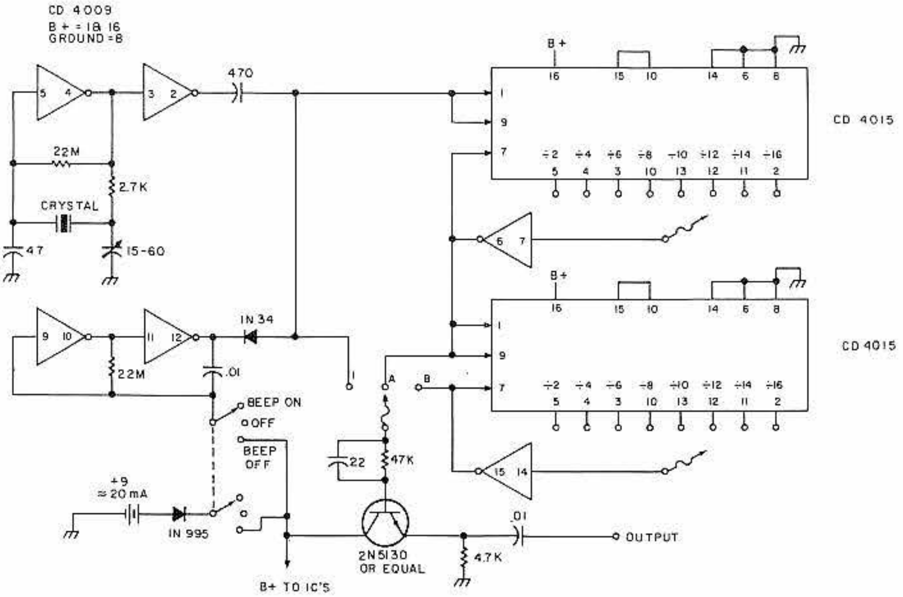

Fig. 1. Schematic.

beeper is a simple clamp that gates rf on or off to following stages.

Photo 1 shows all components mounted on Vector P pattern perf-board that fits inside a Bud minibox. Sleeving 3/8" (10mm) long is slipped over the wire trap terminals of the contact strip to space it up from the board. A DPDT center-off miniature toggle switch acts as one board to panel spacer. Diagonally across from it, a 4-40 threaded rod conducts emitter follower output up through the front panel via a 1/2" (13mm) insulating spacer and plastic shoulder washers. Two regular 4-40 screws and spacers complete the four corner mounting. This spacing allows the contact strip to project partly through a panel cutout so that it is mechanically secure without fastening.

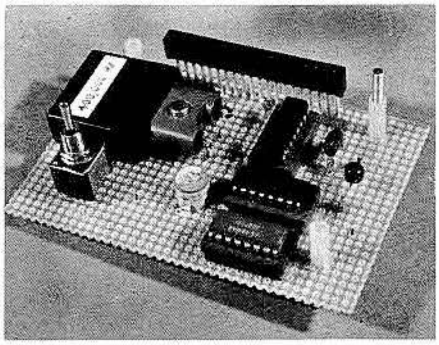

Photo 1.

Photo 2 (completed assembly) shows a stick-on label with patching connection call-outs for various division ratios. If only one crystal is employed, labeling could indicate most used frequencies instead. A typical frequency vs division listing for this model is shown in Table 1. You can easily make up a complete table of all possible ratios, remembering that each CD4015 shift register divides by even numbers ONLY, starting at 2 and ending at 16.

Uses to which a marker generator may be put have been described before: i-f alignment, BFO, scope linearity, etc. A type that divides down to the audio range like this one is especially useful in checking superhets. A

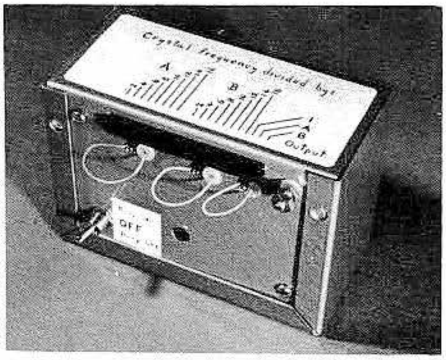

Photo 2.

very broad and flat spectrum of overlapping signals is generated and an audio tone will be heard no matter where the set is tuned. If its tracking and sensitivity are top-notch, the S meter will hold steady over the tuning range. Tracking adjustment amounts to tweaking for maximum meter reading or loudest audio tone. Then patch for 100 KHz markers and check calibration. It's a lot faster and easier than using a conventional signal generator.

1	CD4009	1	15/60pF trimmer
2	CD4015	1	.01 uF
1	1N34	1	xtal; see text
1	1N995	1	xtal socket
1	2N5130 or equiv.	1	Alco #MST205P switch
1	2.7k	1	Bud #CU-2115HG minibox
1	4.7k	1	Vector #44P29-062 perfboard
1	47k	1	#216 battery
2	22M	1	Battery connector
1	22pF	3	16 pin IC sockets
1	47pF	3	Augat patch pins or equiv.
1	100pF	1	20 pin contact strip
1	470pF		SAE #Series 7000

Table 2. Parts List.

POOR MAN'S UNIVERSAL FREQUENCY GENERATOR

John Schultz W2EEY

As precise frequency control and measurement becomes more and more a part of the amateur radio game, the need develops for test instruments that deliver a wide range of both rf and af signals of high accuracy. It would be ideal if everyone could have a frequency counter and a synthesizer type rf and af generator but that is hardly the case. Most amateurs must utilize their basic station gear along with selected accessory items to test out and adjust equipment. This article describes a very useful accessory item that for a modest cost goes a long way toward having some of the expensive test equipment just mentioned. The item to be described is somewhat like a grid-dip meter in that it is basically a simple type of oscillator but as one gets to know and use it, new uses for it are found and its versatility constantly expands.

Circuit Description

Figure 1 shows the circuit diagram of the test generator. Basically, it consists of a string of SN7490 decade counters which are used to divide down a selected input signal by a factor of 10 or 2. The input signal can come from a 1 MHz master oscillator, a special crystal oscillator for externally used crystals or from any external sine-wave source. The special crystal oscillator which uses a SN7400 will operate with almost any basic or overtone crystal in the hf range. It can be used for crystals in the low frequency and lower VHF range also by a simple modification. One gate of the SN7400 crystal oscillator is used to drive a LED which will indicate that the crystal is oscillating so it serves as a crystal activity indicator as well. When an external sine-wave source is used, it is first coupled through a SN74121 multivibrator. This stage squares off the sine wave so it can better drive the subsequent frequency divider chain.

The frequency divider chain is fixed, although one could easily switch the individual SN7490 units to divide by different ratios when desired. This should be obvious by noting the wiring of the divide by 2 SN7490 with that of the divide by 10 units. However, the variety of frequencies which can be generated then with different input sources becomes confusing and more than would normally be needed.

The fixed divider chain follows the sequence: divide by 10, divide by 2, divide by 10, divide by 10. A separate branch after the first divide by 10 unit goes through two other divide by 10 stages. In the case of the divider chain being driven by the 1 MHz master oscillator, this results in the following output frequencies being simultaneously present: 1 MHz (basic oscillator output), 100 kHz, 50 kHz, 10 kHz, 5000 Hz, 1000 Hz and 500 Hz. With any other frequency input source you can easily calculate what frequency outputs the divider chain will bring in both the rf and af regions. Many surplus crystals will produce interesting frequencies of high stability in the af region that can be used for test purposes.

When using the special crystal oscillator, the LED will glow to indicate that oscillation is taking place. As shown with a 150 pF capacitor from one side of the crystal oscillator circuit to ground, the oscillator will work satisfactorily with hf crystals. Its range of oscillation can be extended to lf as well as high frequency overtone crystals by changing this capacitor. The value of capacitor required in picofarads is 500 divided by the frequency of the crystal in MHz. This value need, however, to be only approximate unless you require an absolutely square wave output from the unit.

When using the multivibrator input about a 1½ to 2V peak input, either sine-wave or approximate square wave is required.

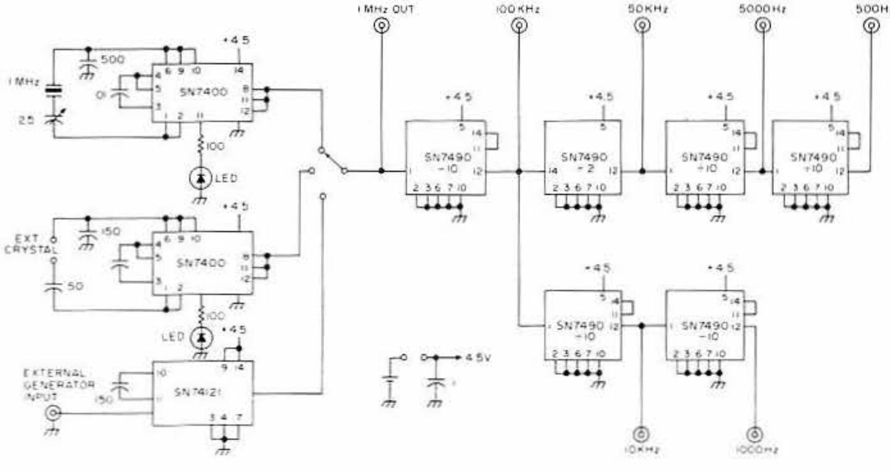

Fig. 1. Diagram of universal frequency generator. Output frequencies shown are for using 1 MHz oscillator.

Construction

The whole unit can be constructed on a piece of perforated board about 3 x 2 in. and made completely portable if powered by a 4½V battery (Burgess No.532) or just three D cells in series. This arrangement does not provide the absolutely best stability for the 1 MHz master oscillator but unless you intend to use the unit for marker frequency generation in the VHF range, it is a perfectly satisfactory arrangement. Alternatively, one could power the ICs from any standard 5.5V regulated supply used for IC digital circuitry.

I constructed my unit for battery powered operation and enclosed the unit in a small aluminum mini-box. The output of each divider was brought to a pin jack on the front panel of the unit.

One simple way to wire the relatively small number of ICs involved is to purchase perforated board which has hole spacing to fit standard DIP and preferably with a copper pad still left around each hole. The ICs are then placed on the board and the appropriate pins which either go to ground or to the 4.5V line bent in different directions. The ground line is run along one side of the IC and the 4.5V line along the other side and bare wire used to connect the appropriate pins to either line. Figure 2, illustrates the wiring for one of the divide by 10 ICs. When one starts this process on the board, it will be surprising how fast the wiring is completed. Individual insulated wire jumpers are used to make the input/output connections between ICs. The wiring is not critical and using a receiver to hear the markers, or an audio amplifier for the lower frequency outputs, one should be able to determine quickly if the circuit is working. The frequency of the 1 MHz master oscillator may be brought exactly on frequency using the 25 pF trimmer in the circuit and checking against WWV with a harmonic of the oscillator or by using a counter.

Applications

As I mentioned before, the applications that you can find for the generator really begin to unfold only after you have had it around the shack for awhile. Some of the applications would be:

1. A frequency marker generator for receiver calibration. The markers are usable up into the VHF range.

2. To extend the range of present rf or af signal generators into lower frequency ranges than they presently cover.

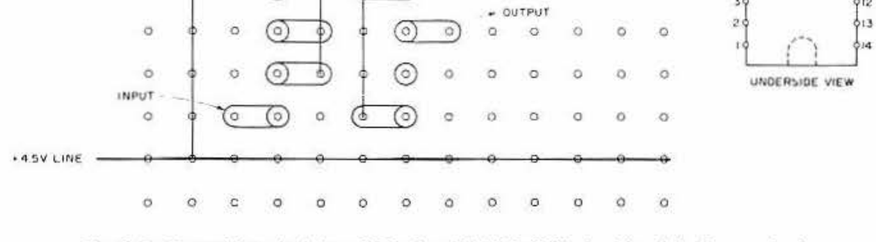

Fig. 2. Perforated board wiring of ICs. One SN 7490 divide by 10 unit is shown wired.

3. To perform stability checks on high frequency variable oscillators. The divider chain will always perform precisely and you can monitor the change in frequency of a higher frequency oscillator with a stable low frequency receiver.

4. A frequency generator to generate precise rf or af square wave signals at any frequency desired by choosing the proper crystal.

5. A crystal activity checker.

6. By taking two or more of the simultaneous outputs together via mixing diodes and a series tuned circuit resonant at the desired frequency, you can also mix the divider outputs to generate a variety of intermediate frequency outputs.

CALIBRATE THAT CALIBRATOR

Mitchel Katz W2KPE

Most modern receivers and transceivers in use today rely upon a 100 kHz crystal oscillator to calibrate the tuning dial. While some of the calibrators are built in, others come as outboard accessories. In any event the operation of each is the same.

By this time we all probably know what "zero beating" is. The 100 kHz oscillator in order to serve as a calibrator must be "zeroed" to some standard frequency such as the WWV carrier frequency on 5, 10, 15, 20, etc. MHz. With a CW or AM receiver, we can very easily tune through the zero beat point. On SSB receivers because one of the sidebands is missing we can only hear the one side as we approach zero. The other side of zero is greatly attenuated and may possibly not be heard at all. To further complicate matters for us in trying to calibrate the 100 kHz oscillator, as we can only hear down to about 20 Hz, we can't zero in any closer than this. Leaving the receiver at this point, we next turn on the 100 kHz calibrate oscillator. After a suitable warm up period, we turn the tune control of the oscillator and again adjust for a zero beat condition against the WWV frequency. With this method of calibration we have several possible sources of error. First in zeroing WWV with the receiver beat frequency oscillator and then zero beating the 100 kHz calibrator against the bfo. Each of these adjustments is limited to the lower limit of our hearing range, as well as the fact that we are obtaining the zero beat at a relatively low i-f frequency.

A more accurate method of calibrating the 100 kHz oscillator will now be discussed. After the receiver and calibrator oscillator have been warmed up for about 30 minutes, tune in WWV on a frequency that produces a fairly good, steady signal. Adjust the tuning for maximum reading on the S-meter. Having tuned in WWV, turn *off* the beat frequency oscillator. Now turn on the 100 kHz oscillator that is to be calibrated. If a harmonic of this oscillator is fairly close to the WWV frequency, a beat note will be heard. At this time adjust the calibrator "crystal tune" control and the S-meter will start pulsing from a maximum to a minimum value. The closer you get to dead center the slower the pulsing action will become. It is fairly easy to come down to 1 pulse per second with this method. If your receiver doesn't have a meter, you can also hear this pulsation very clearly. In any event you would always tune for the slowest pulse rate.

Note that with this method we have adjusted the calibrator frequency harmonic directly to the WWV carrier rather than to a low i-f. We have eliminated one zero beating step, and this, together with the fact that we are obtaining the zero beat at a much higher frequency, will provide greater accuracy.

Having described the method, here are a few points of general interest:

1. Before attempting any calibration let the equipment heat up for at least a half hour to stabilize.

2. After tuning in WWV, wait until the 400 Hz modulating tone goes off before adjusting the calibrator. If not, you may find later that you zero beat the 400 Hz instead of the carrier frequency!

3. The levels produced in the receiver by WWV and the calibrator oscillator should be about equal to produce a good beat between the two frequencies.

4. Use the highest WWV frequency that will produce a good, stable signal in the receiver. Certainly a 1 pulse per second beat at 20 MHz will provide greater calibration accuracy than 1 pps beat at 5 MHz or better yet than 455 kHz! The accuracy will be considerably greater and it is no more difficult to come by.

A SIMPLE FREQUENCY DEVIATION METER

John Reinartz K6BJ

A frequency-deviation meter allows one to read the deviation plus or minus that a received signal is off frequency. Depending on the meter range desired and used, a deviation of ten or less cycles can be read either high or low. Such a device is especially useful when used on MARS nets or when frequency checks are desired of any incoming signal and the answer must be in cycles low or high of a desired frequency. Those grinding their own crystals or desiring to compare crystals will find this device especially useful.

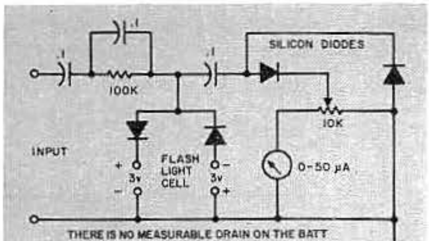
.1
.1
100K
.1
SILICON DIODES
10K
INPUT
+
3v
-
FLASH
LIGHT
CELL
-
3v
+
0-50 μA
THERE IS NO MEASURABLE DRAIN ON THE BATT
WHEN METER IS NOT IN USE.

Fig. 1

Two fundamental circuits were investigated, one using diodes only and the other using transistors only. These are shown in Fig. 1 and 2 respectively. A 0-50 microampere meter should be used for the diode type and a 0-1 milliampere meter will serve nicely for the transistor type, although a 0-50 or 0-100 microampere meter will also serve nicely in the transistor type frequency-deviation meter system. Silicon diodes were used for the diode type and 2N123 for the transistor type. An input voltage of 25 is needed for the diode type and 7 volts or less for the transistor type depending on the meter sensitivity, being 2 volts when a 50 microampere meter is used.

Whatever scale reading is desired, be it 250 cycles low or high or 500 cycles low or high, the meter cover is removed and new figures are added below the meter scale with a zero in the center of the scale and maximum readings at each end of the scale as appropriate. Pencil markings will do. The 250 cycle can be read to 10 cycles per division and the 500 cycle scale can be read to 20 cycles per scale division. Each can be read to half these values or 5 and 10 cycles respectively.

In use, you set your frequency meter, LM or 221 either 250 cycles or 500 cycles lower than the frequency to be checked. If the frequency to be checked is right on, the frequency-deviation meter will read zero at the center of the scale on the meter; if the frequency is low, the meter will read low and if higher, the meter will read higher. The answer in cycles will be the value indicated by your new markings. In use the frequency-deviation meter is connected across the high impedance output of your receiver in the case of the diode type and across the low impedance output in the case of the transistor type.

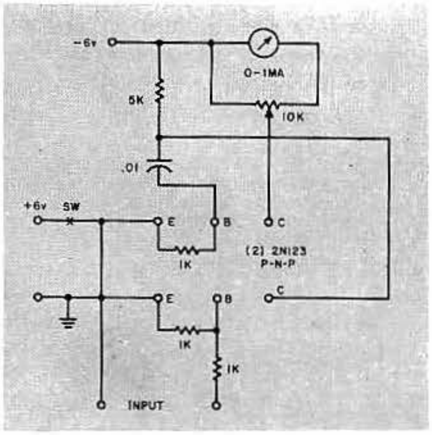

Fig. 2

In those cases where a definite frequency will be under observation, it will be found advantageous to grind or obtain a crystal that is adjustable to 250 cycles low or 500 cycles low, as appropriate, and to use it in the transistorized oscillator shown in Fig. 3. Any crystal holder that has an adjustable air gap will do. Some of the TCS surplus crystal holders have a three point adjustable top plate and are about the best obtainable. Since your best and probably only method of adjusting the crystal is by the use of your LM or BC221, be sure that your frequency standard is accurate. It is best to use the low frequency position and with the 1 mc crystal switch on, tune the meter to that portion of the desired frequency less the mc part. For instance, to set the LM for a reading of 2,732,000 cycles, set the LM on the low frequency for 732,000 cycles or 732 kc. The 1 mc crystal will furnish the mc part of the reading. In my case I set the LM to 731.5 kc and adjust the crystal to that frequency in the adjustable TCS holder. The transistorized oscillator holds the frequency to such a close tolerance it has not been necessary to make adjustments in weeks. A hand held push switch connected into the positive battery lead allows the oscillator to be turned on as needed to check the frequency of a MARS station on 2732 kc to an accuracy of plus or minus 10 cycles of that frequency. The frequency-deviation meter is of course checked against the 440 or 600 cycle tone of WWV, no other check is necessary since the scale is linear.

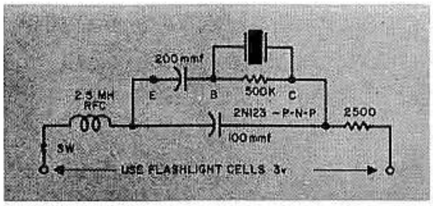

Fig. 3

Chapter VII

GDOs

THE INDICATING OSCILLATOR

Ken Brown KH6AF

Relatives of this little gadget have been around a long time! Even the transistorized versions, which usually leave something to be desired. With an FET, however, we are back in business as with tubes, but with many advantages.

The range of this oscillator is fantastic. That is, without any circuit tricks or special handling. The low frequency end was carried down to the 1 MHz in order to cover the lowest ham band. The high end takes care of 250 MHz easily. This can be extended with a little more effort.

Use of a field-effect transistor (FET) allows operation more nearly like the tubes with which we are, perhaps, more familiar. However, we are not tied to the power lines, which alone makes it worthwhile.

A 2N4221 FET was used in this indicating oscillator. Very likely other FETs will work also. The MOSFETs also should be as good, if not better. The new ones with built-in diode protection would be much easier to handle.

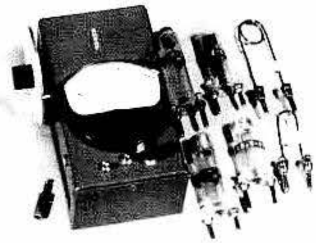

Standard banana plugs spaced ¾ in. make ideal bases for the coils. The lower frequency coils are wound on polystyrene forms.

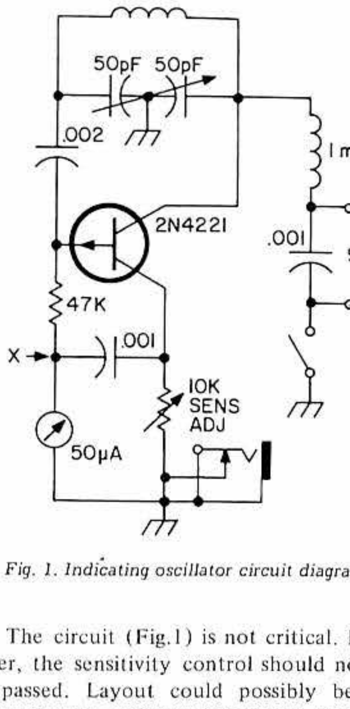

Fig. 1. Indicating oscillator circuit diagram.

The circuit (Fig.1) is not critical. However, the sensitivity control should not be bypassed. Layout could possibly be improved with a slightly larger box, allowing the dial to be placed on the face with the meter. It is a good idea to keep the layout symmetrical as far as possible, particularly the tuned circuit. This can be seen in the photo of the inside view. The box used was an LMB 532 EL with the cover reversed to allow for coil mounting insulator (polystryrene or other good rf insulating material). The meter should be a 50 μA movement; otherwise, a meter amplifier such as the one shown schematically in Fig. 2 will be necessary. This is no problem, as there is room for this amplifier on the circuit board.

A thumbwheel from a BC-375 tuning unit could be used very nicely as a dial. Three sides have been left clear for ease in placement of unit when in use. Plug-in coils makes for easy bank change and application to the job at hand. Standard-spaced banana pins allow for use with other accessories. Use 5/8 in. polystryrene tubing and stud-type banana pins.

A dual banana plug can be used during construction for setting the spacing accurately. The coils shown have the pins wired in place for stronger mechanical assembly. Drill a hole for about 24 AWG copper wire on each side of the pin studs which will lie along each side of the poly tubing. Use a number 59 or 60 drill. One wire is enough on each pin. Form a hairpin with about an inch of wire, push it through the drilled holes from the outside. Now twist tightly with longnose pliers, cut it off short, but not so short as to allow the wire to untwist! Then apply several coats of liquid "poly" cement. Be sure to move the coils frequently during the hardening period to make sure the liquid "poly" flows evenly over the stud and forms a slight fillet with the tube. Epoxy doesn't seem to work well with polystyrene. Neither Allied nor Newark list liquid polystyrene any longer, but your neighborhood hobby shop should be well stocked.

The lowest frequency coil (number 1, 0.95–2.2 MHz) was made from a Miller

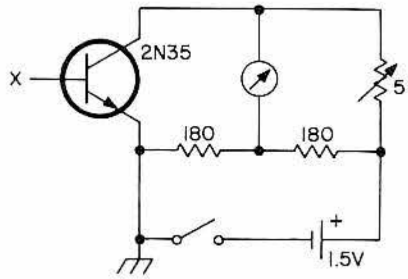

Fig. 2. This meter amplifier will increase the level of the signal so that a less-sensitive meter than 50 µA may be used.

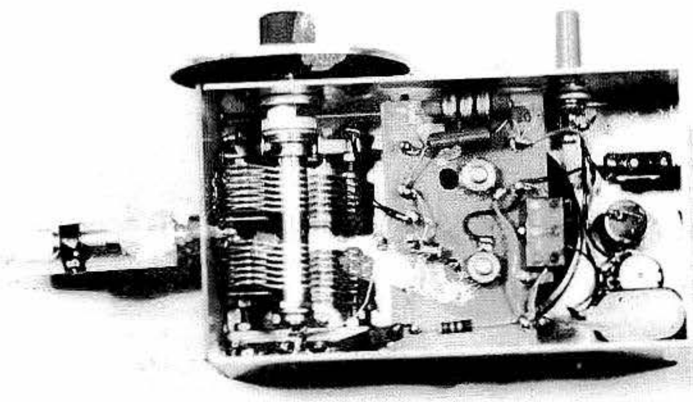

Interior view of oscillator with coil attached shows construction of the author's version.

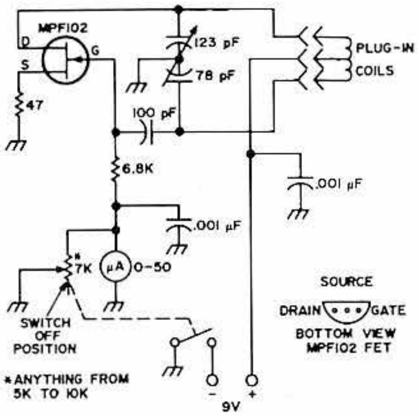

Fig. 1. Schematic diagram of the little gate dipper.

951 ferrite 0.5 mH choke with 6 or 8 turns removed – just enough to slip inside the poly tubing. Some reaming may be necessary. The number 2 coil (2.2–5.4 MHz) was made from a Miller ferrite antenna unit with the slug permanently installed in the top end and all lugs and mounting hardware removed. This coil was also mounted inside the poly tubing. The number 3 coil (5.4–13.5 MHz) consists of 32 turns of 28 AWG enameled wire, close-wound on the outside of the poly tubing. All coils are 2½ in. long with windings as near the end as practical. The number 4 coil (27–50 MHz) consists of 10 turns of 24 AWG enameled wire, close-wound. The number 5 coil (45–100 MHz) consists of 2 turns of 14 AWG enameled wire, self-supporting. The number 6 coil (60–270 MHz) is one hairpin loop 3/8 x 1 in.

A jack is provided for headphone use. The screwdriver-adjusted miniature pot is for zero adjustment of the meter amplifier. It can be seen in the photograph.

THE LITTLE GATE DIPPER

John Aggers W5ETT

Possibly you now own a grid dipper, but is it small, easy to handle, and cordless, making it completely portable? If not, you will want to build this gate dip meter. The cost is extremely low – only about $7. All parts are readily obtainable and construction is simple. The plug-in coil forms, using battery plugs and polystyrene tubing, are easy to make.

The Circuit

An MPF 102 FET is used in a modified Colpitts circuit. Except for the #1 coil, where a choke is used, the B+ is fed to the centertap of the coil. This is necessary to obtain a fairly constant gate current as the oscillator is tuned to its end frequencies. Drain current varies from 4 to 1 mA proceeding from 225 to 1.7 MHz. At the same time the gate current varies from 20 to well over 50 μA.

From this, it is apparent that the stronger the oscillations the smaller the drain current and the larger the gate current. In gate dip operation, as power is drawn from the oscillator the drain current will increase and the gate current will decrease or dip.

Limited wavemeter operation, obtained by switching off the B+, is accompanied by a slight shift in calibration. When the circuit picks up rf, the FET suddenly goes into oscillation using the rf as its battery. Thus, the amount of rf picked up must be large enough or there will be no oscillation and no meter indication. However, despite these deficiencies, it is still considered a useful mode of operation and for that reason has been included. It is only necessary to wire the sensitivity control so that the resistance is maximum when the switch is in the off position.

Construction

A natural finish aluminum minibox (4 x 2-1/8 x 1-5/8 in.) is used for the meter case. The variable capacitor came from an old transistor radio and measured 1 3/8 x 1/2 x 1 in. The shaft was already squared and tapped for a small screw. Since those listed in the catalogs have a plain or flat shaft, you will have to use a collar with setscrew, or drill and tap the shaft. The trimmer capacitors are not used and should be removed.

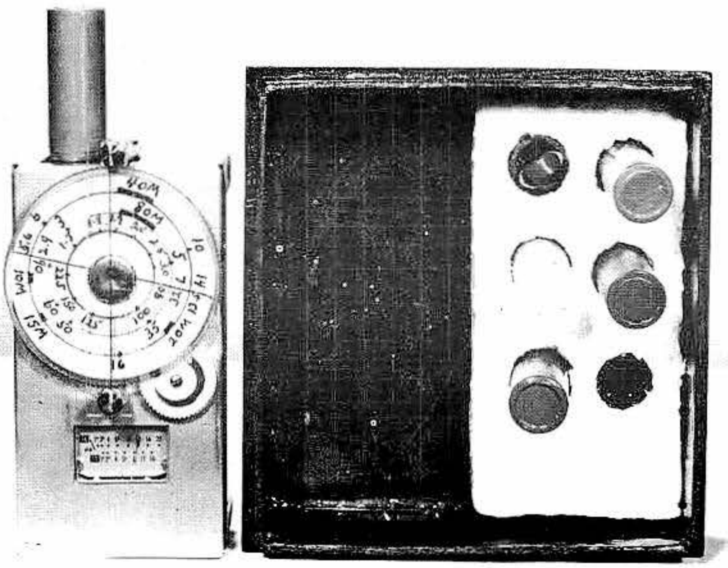

The little gate dipper with spare coils.

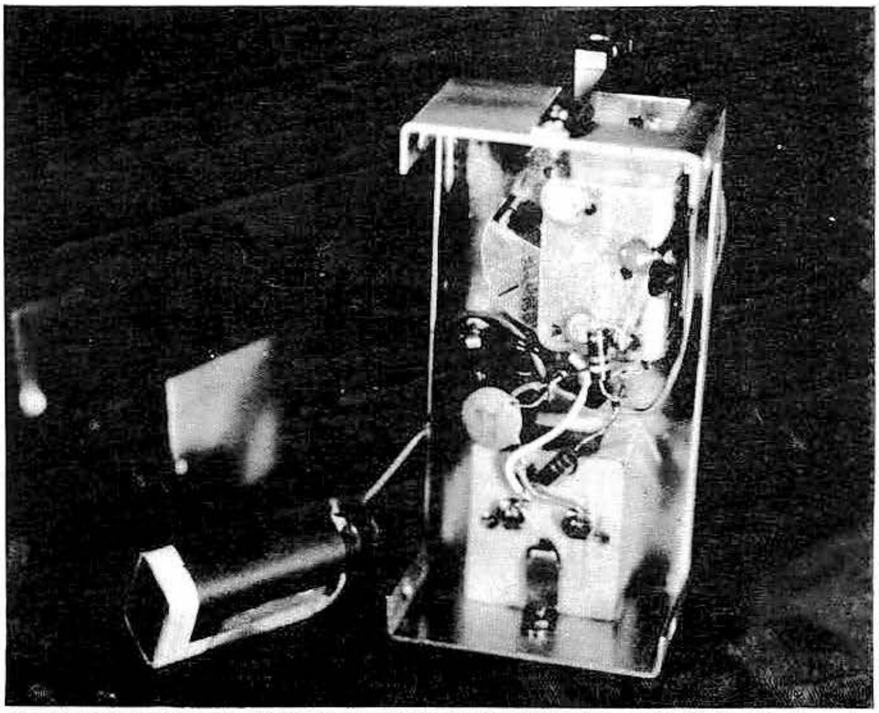

Meter is held against the front panel by a small bracket. The FET is the small black object in the center.

To make the coil socket you will need three pin receptacles from an octal socket, two pieces of 1/8 in. Plexiglas approximately 7/8 x 3/4 in., and one battery plug for a pattern. The pins of the battery plug form a triangle. I shall refer to the holes at the base as the outside holes. Drill holes in one piece of plastic to match the pins of the battery plug. Match the two pieces of plastic, clamp in a vise, and drill the two outside holes in the second piece. Bend the lug part of each socket pin to a right angle. Slip one over each outside pin of the battery plug. Using this as a jig, solder the lug portions to the stators of the variable capacitor. Remove the plug, and the pieces of plastic should fit down over the variable capacitor. The lug part of the center socket pin is brought out between the two layers of plastic.

File a small notch in the bottom piece to accommodate the lug. Before cementing the two pieces together and to the frame, make each hole slightly larger than the diameter of the socket pins. This will allow for expansion when the plug is inserted.

The dial is made of 2-1/4 in. diameter 1/8 in. Plexiglas. To give the dial a rough edge, for good thumb traction, I heated an old gear wheel and rigged up an arrangement to rotate the dial against it. The gear should have rather coarse teeth and rotate with the dial, or you will create flat spots.

The variable capacitor can now be mounted in the case. Position it so that the top and sides of the dial will be just about even with the edges of the case.

The dial marker is mounted on square aluminum posts. The top post (2 in. long) has 1½ in. of its length filed down to a 1/8 in. thickness to reduce its bulky appearance. To make the hairline, scribe a line in a ½ in. wide piece of plastic and fill in with a ballpoint pen.

The sensitivity control I used was already prepared for the knob shown. If you don't have one like it, use a dime-size pot and a setscrew knob. Any resistance from 5 to 10 kΩ will be fine.

Keystone light meters are available from Olson Electronics in a package of five ($3.99) or Transistors Unlimited Co. (75¢ each). Some modification of the meter is necessary. Remove the light cell and series resistor. Drill two holes, spaced ½ in. apart in the back of the case to pass 4–40 machine screws for easy soldering, make sure the heads and nuts are clean and free of any nickel plating. The screws should be filed even with nuts in order to make room for the battery. Solder the leads from the meter movement to the terminals, but be quick because the plastic case tends to melt in a hurry.

Wiring is just a here-to-there proposition, requiring no terminal boards or terminal lugs. The FET is soldered in place supported by its own leads. With reasonable care you should not damage it. A battery holder was found unnecessary; however, it is a good idea to wrap a layer of tape or stiff fiber paper around the battery to prevent the metal case from shorting out the meter terminals.

Coil Construction

Figure 2 and the photo give the necessary dimensions and show the parts needed to make the coil forms. The battery plugs are listed in the catalogs to fit #482 and

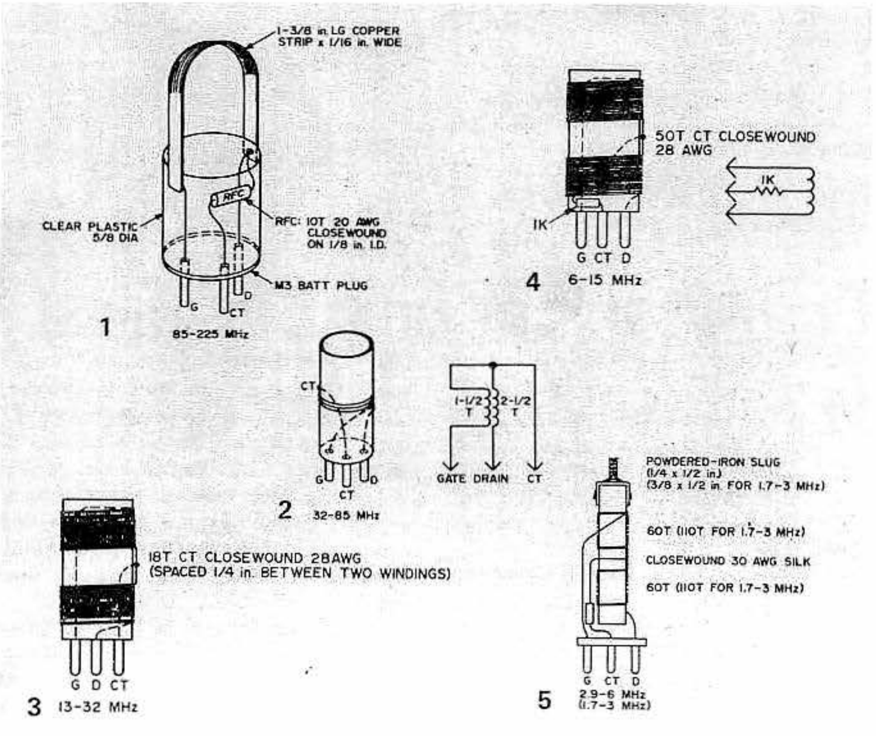
1-3/8 in. LG COPPER STRIP x 1/16 in. WIDE
CLEAR PLASTIC 5/8 DIA
RFC
RFC: 10T 20 AWG CLOSEWOUND ON 1/8 in. I.D.
M3 BATT PLUG
G
CT
D
1
85-225 MHz
50T CT CLOSEWOUND 28 AWG
1K
1K
G CT D
4
6-15 MHz
CT
G
CT
D
2
32-85 MHz
1-1/2 T
2-1/2 T
GATE
DRAIN
CT
18T CT CLOSEWOUND 28AWG (SPACED 1/4 in. BETWEEN TWO WINDINGS)
G D CT
3
13-32 MHz
POWDERED-IRON SLUG (1/4 x 1/2 in.) (3/8 x 1/2 in. FOR 1.7-3 MHz)
60T (110T FOR 1.7-3 MHz)
CLOSEWOUND 30 AWG SILK
60T (110T FOR 1.7-3 MHz)
G CT D
5
2.9-6 MHz (1.7-3 MHz)

Fig. 2. Coil configurations for various frequencies of resonance.

M3 batteries. The center pin should be filed slightly shorter to make the plug seat evenly in the socket. While you are at it, file the nickel plating from the ends of all plug prongs. This will make for easier soldering.

Complete coil information is given in Fig. 2. However a little explanation may be in order. The irregular method of winding the #2 coil is necessary to reach 85 MHz and still maintain oscillation. With 4 turns close-wound the highest was too low. With the 4 turns spaced, oscillation ceased at the highest frequency. The 30 AWG silk wire was taken from a TV flyback transformer. The resistors in the centertap of the last three coils improve the meter's sensitivity slightly. They are mounted right next to the coil winding. With a slight groove filed inside the insulating sleeve, it should slip over the resistor.

The #6 coil is layer-wound as space permits and scrambled wound the rest of the necessary turns. The top winding of all coils should end near the very edge of the coil form. This will make for easier coupling to a tuned circuit. After the coils are checked out the insulating sleeves may be glued to the plug base.

Allow the glue to dry for several days before plugging the open ends of each coil with a small cardboard disk. The coils are painted with colored lacquers. Colored paper between the coil and the insulating sleeve will probably work just as well.

Calibration

For calibration purposes, you will need another indicating oscillator or dip meter. Operate it in the diode or wavemeter mode and loosely coupled to the gate dipper. I calibrated only 5 points on each scale plus any ham bands which appeared. Remember the dipper is not a frequency meter but something to get you in the ballpark.

Conclusion

The little gate dipper was checked against a well known commercial tube equivalent and, as near as I could tell, they were just about even. The battery should last for a long time because the current drain is extremely low.

THE GREAT DIPPER

John E. Boyd WA0AYP

Test equipment is essential in the hamshack, as those of us have found when we attempted to get that new piece of homebrew perking for the first time. One of the most useful pieces of test equipment is the grid dip oscillator or simply, the GDO; besides being relatively inexpensive, it is particularly versatile. Need an indicating absorption wave meter? The GDO will do that. How about a modulated signal source? It handles that too. If you are interested in discovering its whole variety of uses, why not purchase one of several books on the subject.

Every item used in this GDO was selected with an eye toward the average home builder. There are no parts which must be specially purchased from the West Indies Export Company or similar outfit. Nearly all the parts, except for the meter, miniature pot, and mode switch, were obtained from the junkbox, or rather, from several junkboxes. If you insist on buying all new parts, total cost of the project will be about $20.

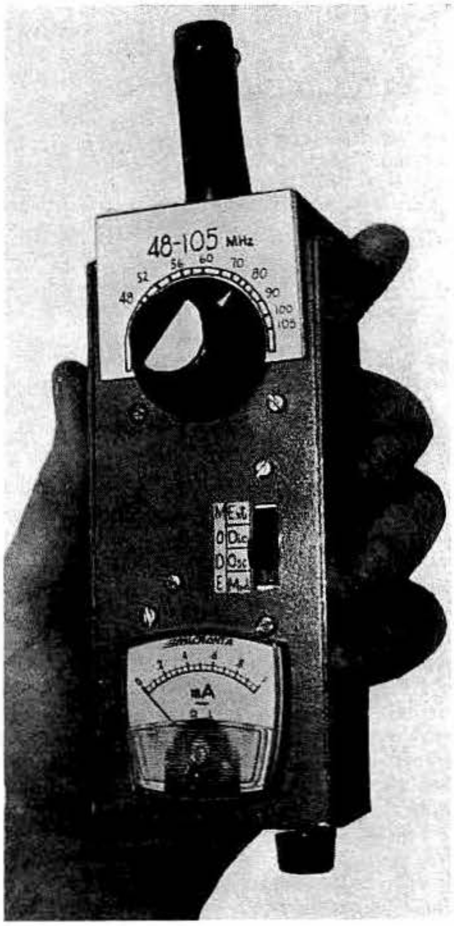

Circuit description

The grid-dip oscillator, in this case more properly termed an emitter-dip oscillator, gets its name from the fact that emitter current in transistor Q_1 decreases when the tuned circuit C_1-L_1 is in resonance with a nearby circuit. This decrease is easily seen by the dip of the meter indicating pointer.

When switched to the diode position, B+ is removed from the oscillator and the incoming rf is rectified by diode D_x; the voltage developed across the 2k resistor is amplified by the meter amplifier and monitored by the 0-1 mA meter. In switching to the signal position, B+ is removed from the meter amplifier but applied to the modulator, and a 1 kHz tone is available from one of the output jacks. In the modulated oscillator position, B+ is reapplied to the oscillator, and the oscillator is modulated by the 1 kHz tone.

Like a patch-work quilt, this GDO was built using circuits from already published articles or books and modified where necessary. The whole circuit is composed of three separate entities–oscillator, meter amplifier, and audio tone generator. The circuit is not particularly critical, but lead lengths and dress in the oscillator must follow good VHF practice, if stable VHF oscillation is to be maintained.

My operating time is spent on the various bands from 28 mHz to 432 mHz. Quite naturally, when I discovered that the GDO would not oscillate satisfactorily over the entire range from 2 mHz to 200 mHz, I juggled values so that it would oscillate well at 216 mHz (for tuning frequency doublers to 432 mHz); then I tried to get as low in frequency as possible. Oscillation was vigorous to about 20 mHz. Coils and scales were then made to cover the respective ranges. It you don't do any homebrewing on the VHF bands perhaps you will find it necessary to

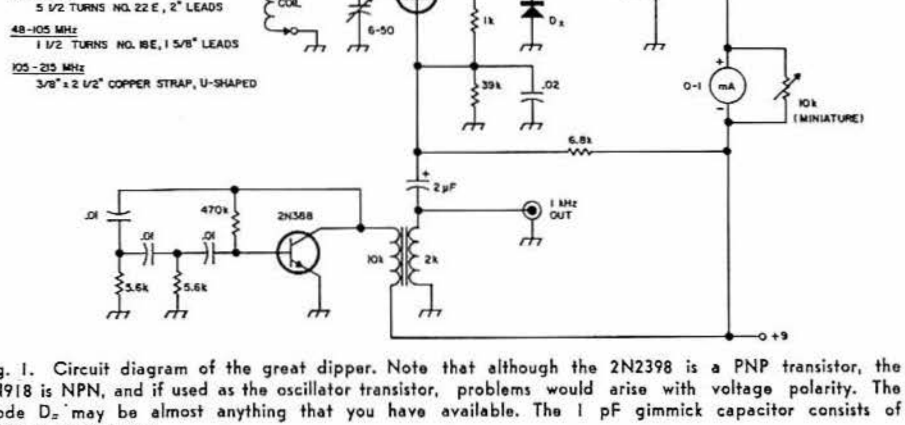

Fig. 1. Circuit diagram of the great dipper. Note that although the 2N2398 is a PNP transistor, the 2N918 is NPN, and if used as the oscillator transistor, problems would arise with voltage polarity. The diode D_x may be almost anything that you have available. The 1 pF gimmick capacitor consists of 1½" of twisted wire.

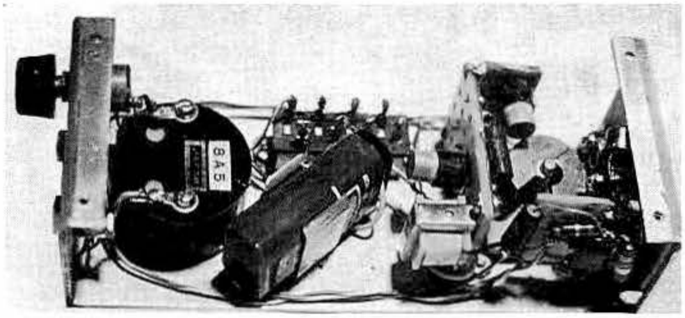

Internal construction of the great dipper. The modulator and oscillator boards are to the right—the oscillator transistor is mounted right next to the coil jacks.

change the value of the emitter-collector feedback capacitor, and to juggle the emitter and base resistor values in order to sustain oscillation at your desired frequencies.

Construction

Vector board was used, mainly because I wanted to experiment with component values; however, a printed circuit would be just as good, especially for something such as a club project. It can be seen from the photograph that the meter amplifier and audio oscillator are built on separate boards. This is due to the fact that I built several different amplifiers; the layout would look neater if they were on the same board. Positioning of the rf oscillator and capacitor C_1 as shown in the photograph is recommended, but the placement of other parts is not critical.

Fiberglass board is used as an insulator for mounting the banana jacks and plugs. It cuts and drills easily and appears to work fine. Three banana jacks were used, the third jack being used merely to provide mechanical rigidity. It could also be used, if necessary, to shunt additional capacitance across the emitter and collector on the lower frequencies.

Because a shear and a brake were available, I constructed my own chassis, consisting of two U-shaped pieces of 1/16" aluminum. Using the GDO is a breeze, for it fits the hand very comfortably; if placed on the workbench, it doesn't roll off each time it is bumped. The completed case (1¾" H x 2½" W x 6¾" L) is exceptionally rigid and imparts a reassuringly solid feel when handled. Commercially available miniboxes could be used if you don't have facilities available for rolling your own.

In building the rf oscillator, keep all the rf leads as short as possible; especially the short lead from C_1 to Q_1 and from circuit ground to chassis ground. It was found that false dips could be completely eliminated if a copper strap ¼" wide was added from the capacitor ground lug directly to chassis ground. Apparently the ground on the variable capacitor C_1 is not quite good enough at frequencies above 100 mHz. Various transistors were tried in the oscillator; the PNP type 2N2398 was found to be a good performer, as was the NPN type 2N918. However, the use of the NPN type could lead to problems with battery polarity. Capacitor C_2 is a 1½" length of twisted wire positioned near the collector lead of Q_1. This slightly modified oscillator circuit is from a book describing, among other things, a transistorized GDO which you may wish to use as a reference.

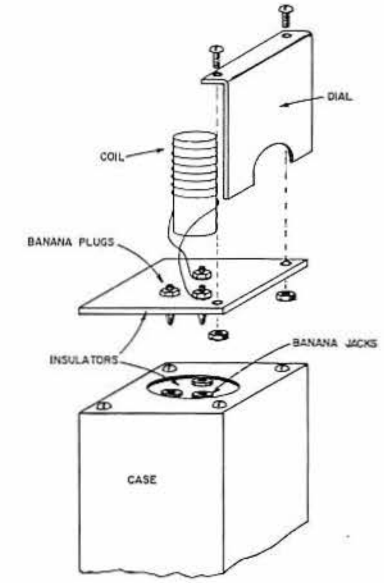

Fig. 3. Construction of the plug-in coil assemblies. The coil forms were made from the plastic containers which hold Polaroid print coater.

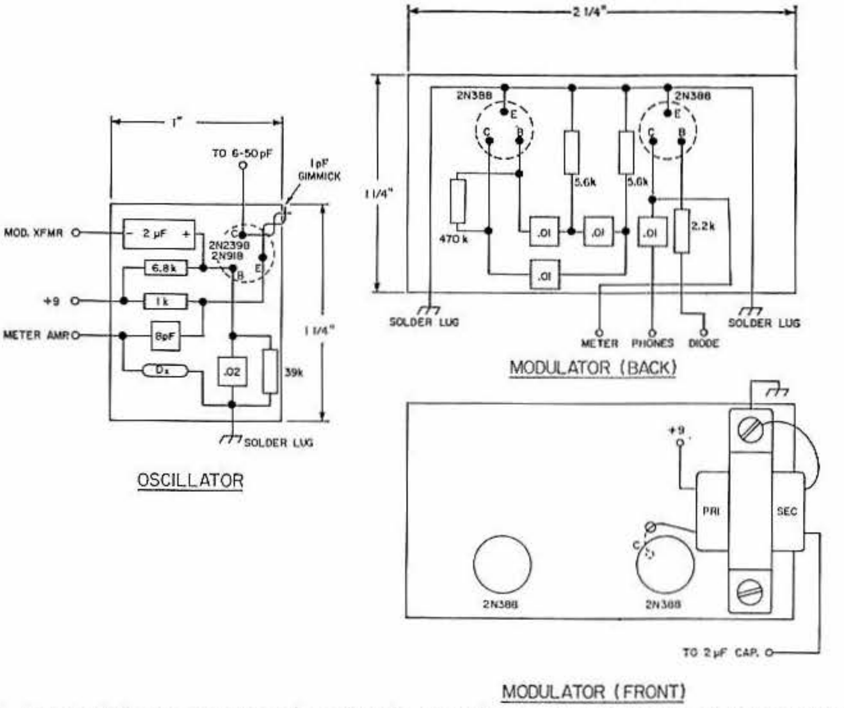

Fig. 2. Layout of the two circuit boards used in the great dipper. Although two boards were used in this case, the circuit could be easily adapted to one board, and even to printed circuitry.

To keep cost low, a 0-1 mA meter was used in conjunction with a simple meter amplifier. If you happen to have a 0-50 μA meter lying in the junkbox, that would work equally well, and the circuit could be simplified accordingly. Several circuits were built for the meter amplifier; the one chosen was a compromise between cost and performance. A germanium transistor was used because it requires less voltage to turn it on. Leakage is low, the pointer of the meter resting just off zero when no coil is in place. Further information can be obtained from January 1966 *73 Magazine*, which was the source for this circuit.

A transistorized audio tone generator is coupled by a 2 μF capacitor to the base of the oscillator transistor for modulation. This modulated oscillator allows the GDO to be used as a versatile signal source. An output jack is included on the panel to allow the 1 kHz tone to be used without turning on the oscillator.

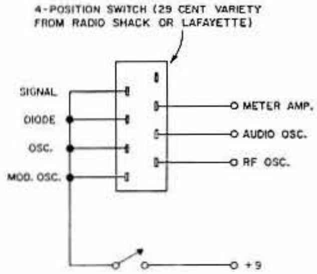
4-POSITION SWITCH (29 CENT VARIETY FROM RADIO SHACK OR LAFAYETTE)
SIGNAL
DIODE
OSC.
MOD. OSC.
METER AMP.
AUDIO OSC.
RF OSC.
+9

Fig. 4. Wiring the four-position slide switch for the great dipper.

The three plug-in coils for the Great Dipper. Three ranges cover from 28 to 216 MHz. The second harmonic of 216 MHz may be used for tuning up 432 MHz converters and such.

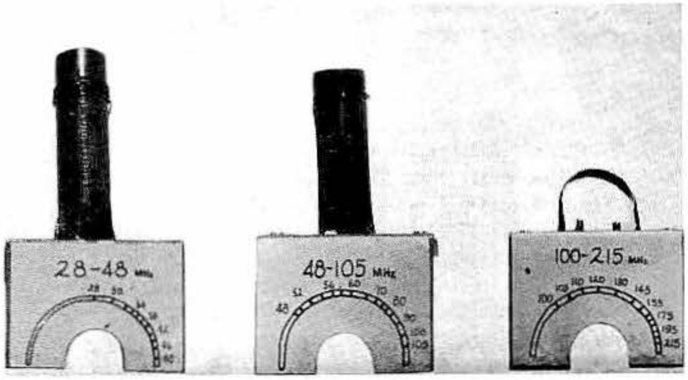

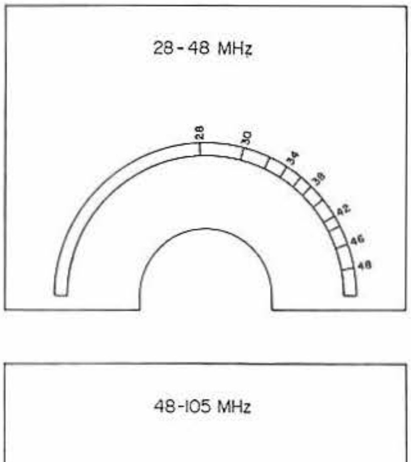

Fig. 5. Full-scale dials for the great dipper. If the construction shown in the photographs is followed closely, the calibration of these dials should be within several percent.

There are a couple of components which not everyone will want to duplicate. One, the sub-miniature 10k pot with SPST switch, was chosen because a very limited amount of space was available; if the unit is built on a larger chassis, the more commonly available *Midgetrol* could be used. The other, a four position switch used to select the desired mode, is a 29c variety available from Lafayette or Radio Shack. It has a peculiar switching arrangement and if you duplicate this project, several hours of experimenting could be eliminated by following the pictorial diagram included in this article. A disadvantage of this particular switch is that the meter does not indicate in the modulated oscillator position.

Using individual scales on each plug-in coil assembly greatly enhances scale legibility, reducing the chance of reading error and speeding frequency identification. This scheme; however, is not original. It was described in a 1957 issue of *Short-wave Magazine*, and is currently being used on a commercial GDO. It requires little additional effort to build the coil assembly in this manner and is, to me, well worth that extra effort. For want of anything else, the coil forms were made from the plastic tubes which contain the film coater supplied with each roll of Polaroid film.

Lastly, ease of tuning is accomplished largely through the use of a 1" skirted knob. Small knobs are simply too difficult to use comfortably.

Calibration and operation

It is best to calibrate this GDO by listening for the oscillator, modulated by the 1 kHz tone, on a general coverage receiver. An alternate method is to use another GDO, placing one in oscillate and the other in diode, tuning for either peak or dip. The scales which were used on this GDO will serve if parts and layout are followed closely.

To use this unit as a dipper, place the mode switch in oscillate, and place the dipper coil next to the coil under test. The turns of both coils should be parallel, and not at right angles to each other. To keep from pulling the oscillator frequency, keep the two coils separated as much as possible, while still maintaining a meter dip. This assures that dial accuracy will be kept high. If a coil is inaccessable, twist a pair of wires together, forming a two turn coil on each end; slip this coupling link over the two coils. Keep in mind that a coil, when it is in a circuit, may not dip at the same frequency as when it is out of this same circuit.

GREATER DIPPER – MODIFICATION

John E. Boyd WA0AYP

Introduction

To the uninitiated, a grid dip oscillator is neither glamorous nor exotic, even if it does happen to be transistorized. But to those who have ever tried to put a coil on frequency, one is worth its weight in microcircuits.

The original model of the *Great Dipper* proved to be a versatile, if unexciting, grid dip oscillator. Intended to be used mostly at VHF, the dipper was limited to frequencies above 28 MHz. To those who seldom operate above 10 meters, the lack of frequency coverage below 28 MHz was a definite handicap. Expanding the frequency coverage downward extends the dipper's usefulness considerably.

Component changes made were not extensive, nor were the physical dimensions of the unit changed. The schematic diagram of the modified circuit is shown in **Fig. 1**, and the pictorial diagram of **Fig. 2** is included to supplement the photographs printed in the original article.

Modification

First, the oscillator circuit board should be rotated 90 degrees from its original position so that transistor substitution or replacement can be made easily. A microwave diode might be substituted for the original glass computer-type diode; this results in better performance at VHF, but its effect cannot be accurately predicted. A 1N21 or 1N23 diode could be used in place of the D4900 called for in the schematic. Best oscillator performance over a wide frequency range results from using a 5 pF capacitor as the collector feedback coupler, but this was a compromise value. If operation is likely to be confined to either the HF or VHF segments of the radio spectrum, some experimentation with this capacitor should result in improved performance over that particular range. For experimentation, use a 3-12 pF or 4-30 pF trimmer.

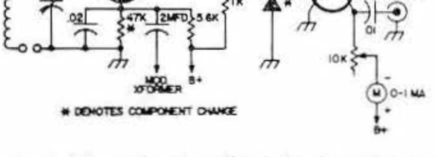

*Fig. 1. Schematic of modified circuit, with component changes indicated by *. L1: 13-23 MHz, 14½ turns No. 20; 22-44 MHz, 5¼ turns No. 24, 45-90 MHz, 1½ turns No. 24; 90-195 MHz, 2½ x 3/8 in.*

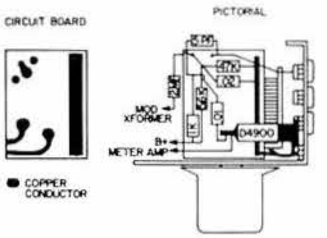

Fig. 2. Pictorial view of the dipper and its circuit board.

To perk up meter amplifier performance, the 2N388 meter amplifier is replaced by a 2N918, a relatively high gain, low leakage transistor.

Although the change was not made in this unit, a 6-100 pF (MAPC-100) variable capacitor could be substituted for the existing 5-50 pF unit. As can be seen from the full-size scales, frequency coverage is narrow on the 13-23 MHz plug-in coil. Those who frequently construct HF gear might benefit from the substitution.

Operation

Operation of the dipper is the same as before, except the frequency range is greater. Another coil must be wound for the 10-25 MHz range, but existing coils were not modified. Calibration is now different from before; full-size scales are included with this article for those who do not have a general-coverage receiver or another grid dip oscillator on hand for calibration.

Oscillation is still vigorous at 13 MHz, which indicates lower frequencies may be reached with a suitable coil plugged in. Similarly, oscillation is still noticeable at 195 MHz, so operation at higher frequencies may be possible.

Finishing Touches

Appearance of the dipper was improved after enlisting the help of a draftsman. The dial and meter scales were professionally inked. **Figure 3** provides full-size copies of these scales.

To improve the legibility of the labels and scales, they were inked on yellow cards rather than on conventional white paper. It's a small point, but for those who use test equipment frequently, the black-on-yellow technique results in less eye strain.

ETCHED CIRCUIT UHF DIPMETER

Bob Corbett W1JJL

I made three models of the dipper. They cover 130-175 MHz, 175-250 MHz, and 250-480 MHz. The first two models use the same size inductor, with the 130-175 MHz model using a larger capacitor for tuning with a ceramic trimmer across it. This trimmer is not shown in the schematic; its value is 7-45 pF and it should be adjusted to cover the proper range.

The 250-480 MHz dipper uses a smaller inductor than the others. It also has a copper jumper (shown in the layout) that the lower frequency dippers don't have.

Each dipper is complete (including the battery) except for the meter. The meters were omitted to save space and money, but can be included if you wish to use a slightly larger case.

Each dipper uses one RCA 2N3478 NPN silicon transistor. These transistors cost only $1.90 apiece, but it's likely that other transistors that are even cheaper could be used. The 2N3478 has odd basing—the only reference is the short case-shield lead—so don't shorten any leads until you're sure that you can keep track of the connections.

The copper side of the board for the dipper is shown in Fig. 2 with the component side in Fig. 3. Use glass or Teflon based board. Paper or bakelite board probably wouldn't be satisfactory. Trim the board to the proper shape with a nibbler. The inductor should be coated with coil dope to keep from shorting it when you use it.

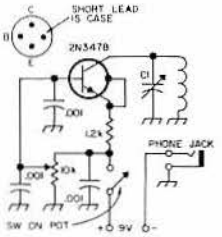

Fig. 1. The etched circuit dipper is very simple. The circuit is almost identical to the one described by WA1CCH in the December 1965 73, but the construction is quite different. C1 is Johnson 160-104 (9 pF) for the two higher frequency dippers, and 160-107 (14 pF) for the 130-180 MHz model. There is a trimmer across C1 in the 130-180 MHz model; see text.

To mount the boards, you'll have to cut a thin slot near the edge of the Minibox used as a case. One way to do it is to drill a number of holes of the proper size in a row, then use a file to finish the slot. You'll have to bend that side of the Minibox out to get the board in. It's held in place by an extra set of nuts on the shafts of the potentiometer and the tuning capacitor. Be sure to trim the leads projecting from the copper side of the board so that they won't touch the metal of the case. The battery is held in place with a simple clip made from scrap metal.

The dipper is very easy to use. But before we get to that, let's check it out and calibrate it.

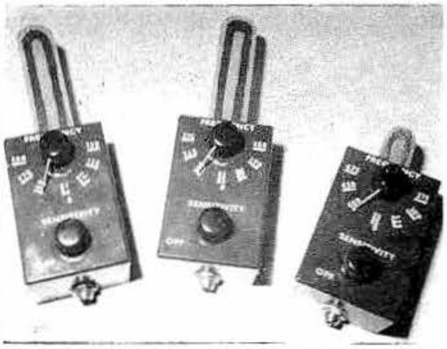

The three dippers shown here cover 130-480 MHz.

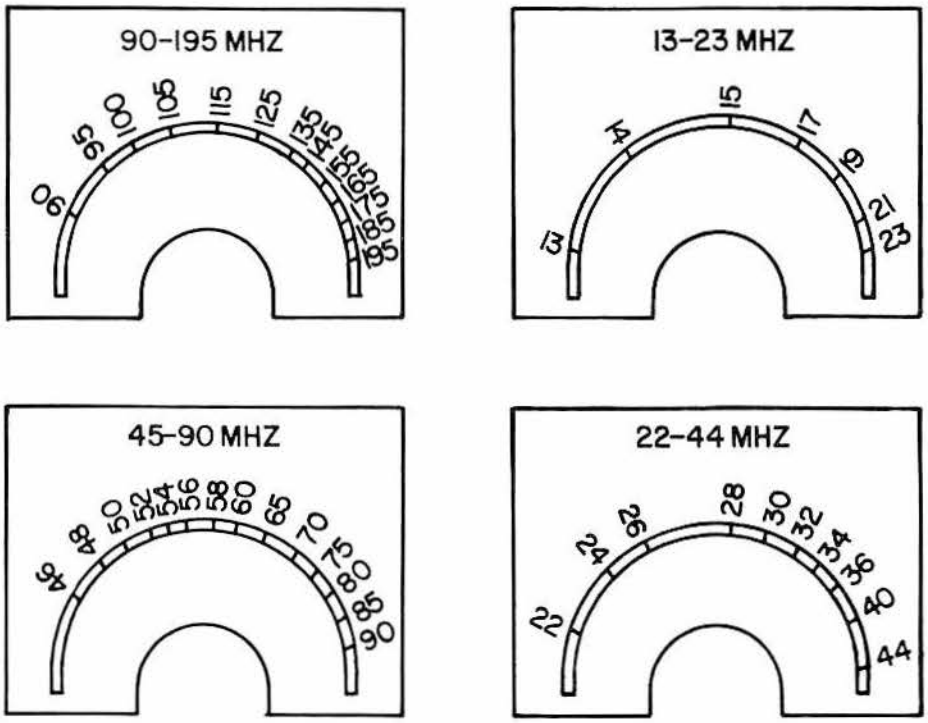

Fig. 3. Dial designs for the various MHz ranges, full size.

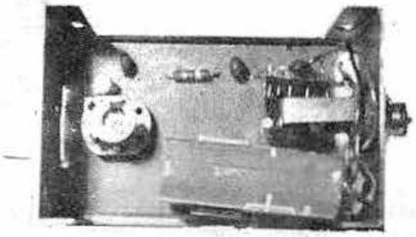

Here's the inside of one of the dippers.

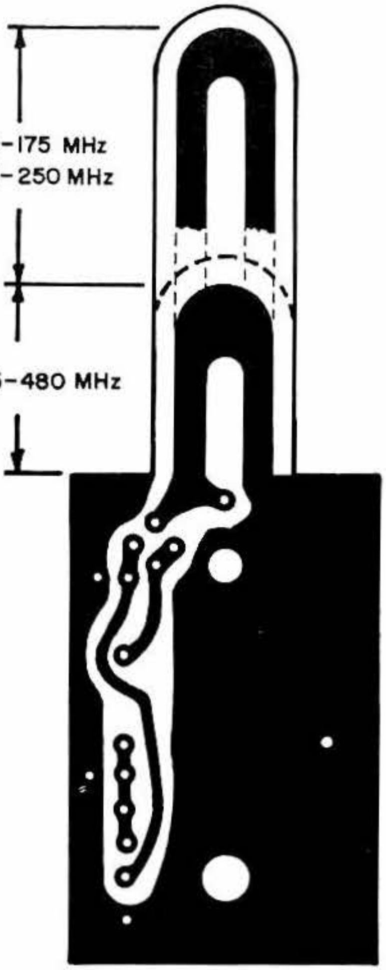

Fig. 2. The copper side of the etched circuit board used in the dipper. This layout is full size. Use board suitable for these frequencies: fiber glass or Teflon.

Plug a 2 to 5 mA meter into the meter jack. You can use a more sensitive meter if you shunt it with a resistor that gives the proper scale. Put the resistor across the meter jack terminals in the dipper if you use the meter for other things.

Turn on the dipper by twisting the potentiometer knob clockwise until it clicks. The meter should show very low current. As you turn the pot, the current should suddenly jump to about 1 mA. That means that the transistor is oscillating. If you touch the coil, the meter reading should drop and the dipper may stop oscillating completely. Now tune the capacitor through its range. There should be a little variation in current, but not too much.

Now you're ready to calibrate the dipper. The easiest method is a sensitive wave meter that covers the range, but it's quite easy to do the job with a TV set. A TV set covers 176 to 216 MHz (channels 7-13) for the low calibration. Then the second harmonics of the dipper tuning 235 to 445 MHz can be received on a UHF TV set (470-890 MHz). If you have a two meter receiver, that gives you another maker at 146 MHz. You can put on the panel markings with Ami-Tron or Datak rub-on lettering.

The dipper should be complete now, and ready for use. Bring the dipper near a resonant circuit in the dipper's range and tune the frequency control. You should get a prominent dip in current when both circuits are tuned to the same frequency. The amount of dip depends on the setting of the pot in the dipper, the distance from the tuned circuit, the Q of the circuit, and the type of coupling.

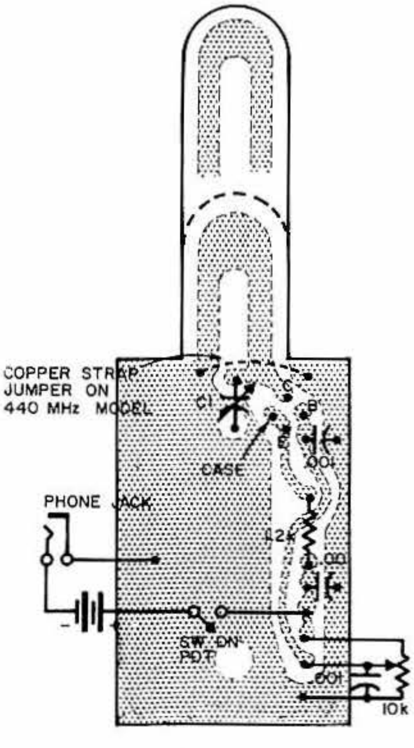

Fig. 3. Component side of the dippers. There is a 7-45 pF ceramic trimmer across C1 in the lowest frequency model. See the text.

In many cases it's easiest to leave the dipper stationary and tune the other circuit.

The dipper can also be used for monitoring AM transmitters by plugging a set of headphones in the meter jack and adjusting the tuning and pot. You can also use the dipper for determining the frequency of another oscillator. Simply tune the oscillating dipper with headphones plugged in until you hear a slight click. You probably won't be able to get a zero beat at these frequencies.

Be careful that you don't use the dipper around an energized transmitter of more than a few watts output or the dipper may be damaged.

These dippers are simple, inexpensive and non-critical to build. After you've build them, you'll wonder how you ever tried to build UHF equipment without a good dipper.

UHF GRID DIPPER

J. Fisk WA6BSO

When building or testing equipment for the 420 mc band, amateurs invariably run into the problem of, "Where am I?". It becomes a little difficult to tell whether you are actually in the band or somewhere nearby. The uninitiated will counter that you should be able to figure close enough, after all the band is 30 mc wide; but even the experienced old timer will confirm that this just isn't so.

The 420-450 mc amateur band falls between the VHF and UHF television assignments and there is very little to use for a frequency reference point. It is nearly impossible to tell exactly where you are without resorting to expensive commercial gear or lecher wires. Nor is the problem confined to the ¾ meter band; when building crystal controlled converters for 432 or 1296, it's nice to know that the last tripler is really tripling and not doubling or quadrupling. Even with a 45 or 50 mc crystal, it's quite easy to tune up on the wrong harmonic. Remember that a 40 mc change at 400 mc is analogous to a 400 kc change on 75 meters.

There are several commercial instruments that fill this requirement nicely, but the cost of the least expensive of these would buy a pretty respectable all band receiver. Occasionally suitable test equipment appears on the surplus market, but again, the price is prohibitive. The simple UHF grid-dipper described in this article was designed specifically to economically fill this need. It covers the frequencies

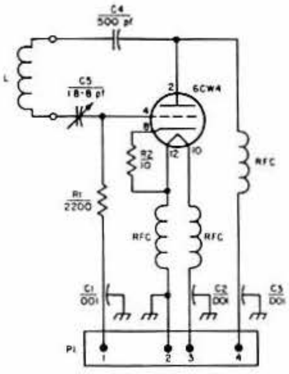

C1, C2, C3	STANDOFF BUTTON BYPASS CAPACITORS
C4	BUTTON CAPACITOR (SEE TEXT)
C5	E F JOHNSON TYPE 160-104 VARIABLE
L	PLUG-IN COIL
P1	AMPHENOL 91-MC4M MALE CABLE PLUG
RFC's	8-1/2" No 2" ENAMELED WIRE WOUND ON 1/2 W, 100K COMPOSITION RESISTORS

Fig. 1. Schematic of the rf head of the UHF grid dipper.

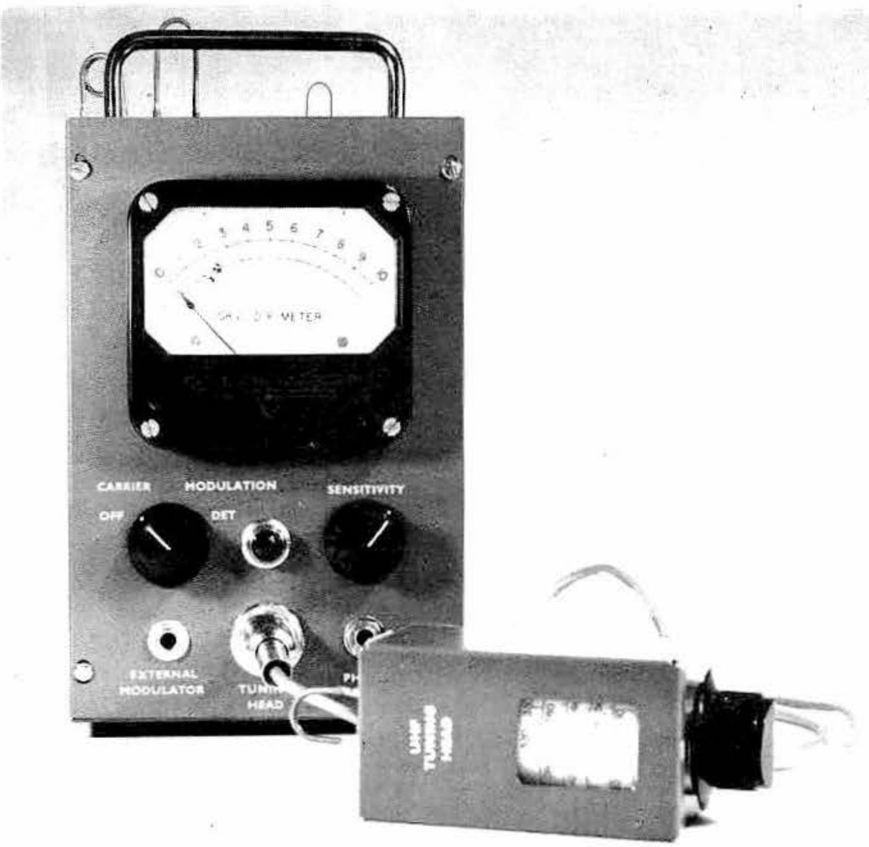
GRID DIP METER
CARRIER
MODULATION
SENSITIVITY
OFF
DET
EXTERNAL MODULATOR
TUNING HEAD
UHF TUNING HEAD

from 300 to 680 mc at a total cost that is comparable to a low-frequency kit-type grid-dipper. By using junk-box parts and smart horsetradesmanship this cost may be substantially reduced. I should hasten to point out that substitution of parts in the tuning head should be avoided if accurate dial calibration is desired.

This particular grid-dipper has been duplicated several times and in each case calibration has been within several percent of the original unit; at least as accurate as the familiar kit-type grid-dipper. This is helpful to the ham who does not have access to commercial equipment for calibration purposes. By using the layout and parts described in this article, the precalibrated dial illustrated in **Fig. 8** may be used with a minimum of error.

Although an accessory power supply, modulator and indicator are included in this design, an enterprising ham could use the power supply and meter presently available in any lower frequency tube type grid-dipper. Although using a modified low-frequency dipper is simple and inexpensive, the power/indicator box specifically designed for the UHF tuning head has several features that are not available in inexpensive kit-type units. These include internal 1,000 cycle modulation, provision for external modulation, and voltage regulation which provides the necessary stability at the ultra high frequencies.

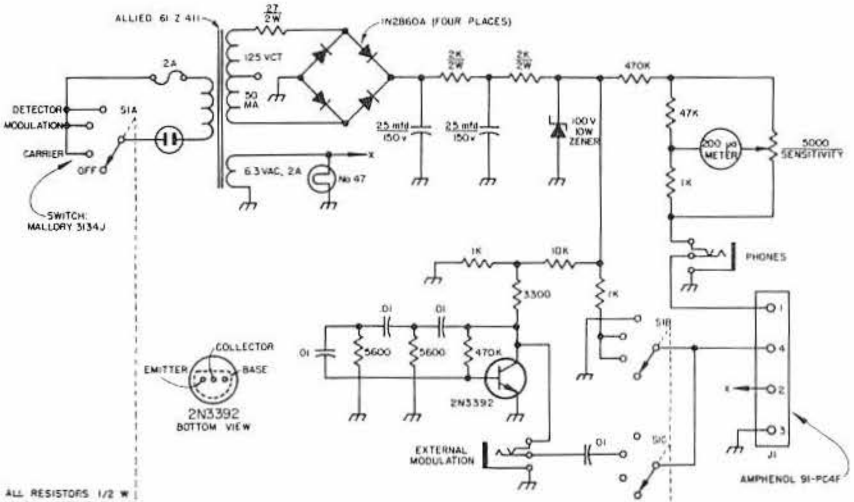
ALLIED 61 Z 411
27 2W
1N2860A (FOUR PLACES)
2A
125 VCT
50 MA
2K 2W
2K 2W
470K
DETECTOR
MODULATION
CARRIER
OFF
S1A
25 mfd 150v
25 mfd 150v
100V 10W ZENER
47K
200 µa METER
5000 SENSITIVITY
6.3VAC, 2A
No 47
X
1K
SWITCH: MALLORY 3134J
1K
10K
PHONES
3300
1K
.01
.01
.01
S1B
5600
5600
470K
COLLECTOR
EMITTER
BASE
2N3392 BOTTOM VIEW
2N3392
1
4
2
3
X
J1
EXTERNAL MODULATION
.01
S1C
AMPHENOL 91-PC4F
ALL RESISTORS 1/2 W UNLESS NOTED OTHERWISE

Fig. 2. Power supply and modulator for the UHF grid dipper.

The rf tuning head

The heart of this instrument lies in the tuning head itself. There is nothing particularly new or different about the circuit, but at these frequencies stray inductance and capacitance in circuit layout and construction will seriously affect the end result. All the lead lengths must be as short as humanly possible and physical circuit layout must follow standard UHF practice. One of the big problems in building tunable oscillators at UHF is to obtain an oscillator that will tune from one end of the range to the other with no "holes," frequency jumping or instability. The series tuned 6CW4 nuvistor oscillator shown in **Fig. 1** fills these requirements.

Insofar as possible, all the tuning head wiring is done on a point-to-point basis with the components mounted directly to the 6CW4 tube socket or button capacitor lugs as illustrated in **Fig. 3.** To obtain the desired accuracy with the precalibrated dial, this diagram should be followed as closely as possible.

To keep stray circuit inductance to a minimum, all wiring between the coil and tube socket is done with one-quarter inch wide strips of thin copper sheet. In addition, the rotor of the variable capacitor is connected to the 6CW4 grid (pin 4) with a short strip of the same material. The only tricky part of this wiring is the installation of the series capacitor C4. This capacitor is a 500 picofarad button type mounted as shown in **Fig. 5.** The "S" shaped bracket ("A") is made from a one-quarter inch wide copper strip, one and one half inches long and soldered to the mounting stud of the capacitor. Connection to the 6CW4 plate (pin 2) is made with the soldering lug on top of the button.

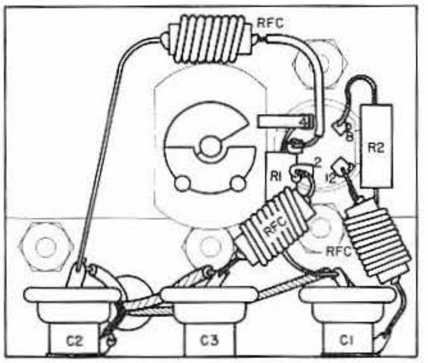
RFC
R2
R1
RFC
RFC
C2
C3
C1

Fig. 3. Critical wiring around the 6CW4 socket.

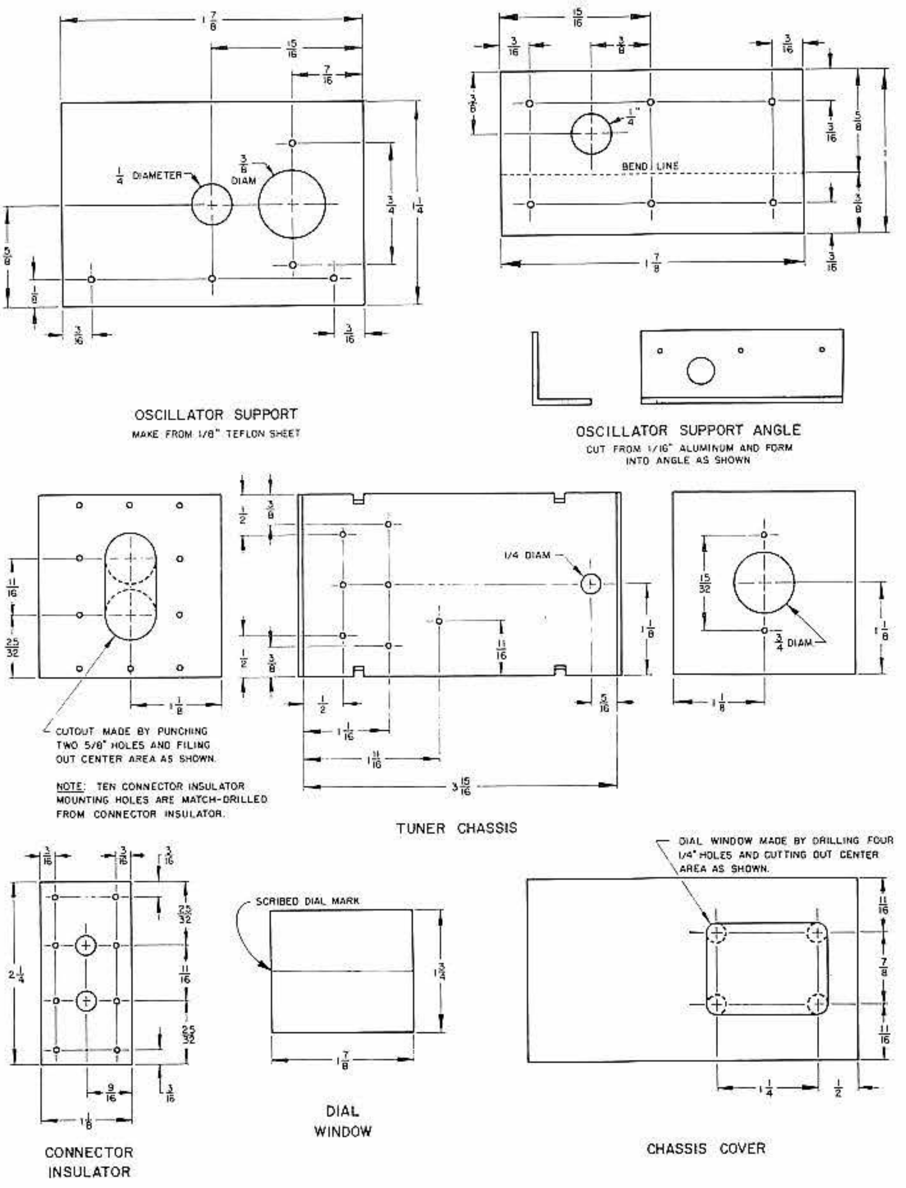
OSCILLATOR SUPPORT
MAKE FROM 1/8" TEFLON SHEET
1/4 DIAMETER
3/8 DIAM
OSCILLATOR SUPPORT ANGLE
CUT FROM 1/16" ALUMINUM AND FORM INTO ANGLE AS SHOWN
BEND LINE
CUTOUT MADE BY PUNCHING TWO 5/8" HOLES AND FILING OUT CENTER AREA AS SHOWN.
NOTE: TEN CONNECTOR INSULATOR MOUNTING HOLES ARE MATCH-DRILLED FROM CONNECTOR INSULATOR.
1/4 DIAM
3/4 DIAM
TUNER CHASSIS
3 15/16
CONNECTOR INSULATOR
SCRIBED DIAL MARK
DIAL WINDOW
DIAL WINDOW MADE BY DRILLING FOUR 1/4" HOLES AND CUTTING OUT CENTER AREA AS SHOWN.
CHASSIS COVER

Fig. 4. Layout of the tuner chassis, chassis cover, dial, connector insulator, oscillator support and oscillator support angle.

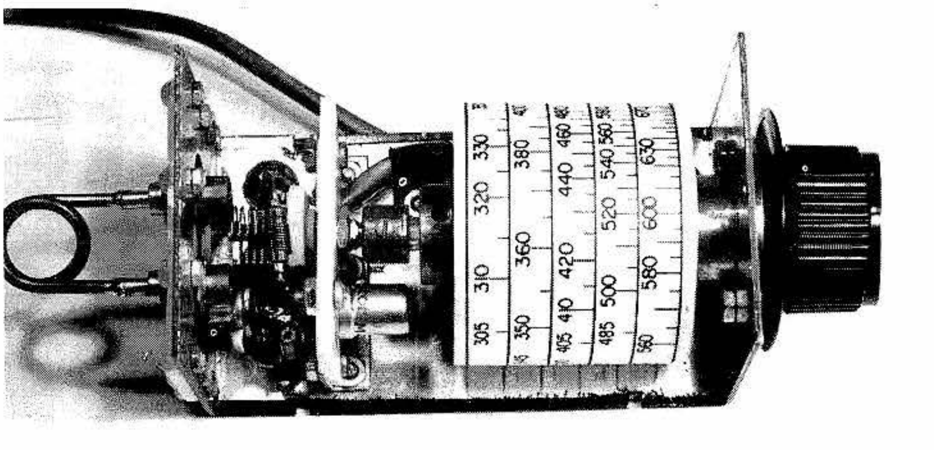
305
310
320
330
350
360
380
405
410
420
440
460
485
500
520
540
560
580
600
630

Connection to the stator of the variable capacitor is accomplished with another short strip of thin copper as shown in **Fig. 5** ("B"). This piece of copper is bent so it touches both stator mounting pins when the unit is assembled; then it is soldered in place.

The oscillator and tuning mechanism are housed in a standard 2¼ × 2¼ × 4 inch chassis box laid out as shown in **Fig. 4.** Although the author's unit is based on an LMB type 107 chassis box, other manufacturers have similarly sized boxes which are equally suitable. The layout of the enclosure is straight-forward and no difficulty should be found in duplicating it. The dial and coil socket cutouts are made by drilling or punching round holes and then cutting out the area between them as shown in the drawing. This is easily done with an Adel "nibbling" tool. All of the small screw holes are drilled with a standard ⅛ inch drill to pass 4-40 screws; 5/32 inch holes will be required if 6-32 screws are used.

The dial window is cut from a sheet of 1/16 inch clear plastic to the dimensions shown in **Fig. 4.** A hairline is scribed in the center of this window; if a more distinct line is desired, this scribe mark may be filled in with black india ink. After the tuner enclosure is completed, this dial window is epoxied in place over the dial cutout.

The coil socket is made by installing two banana jacks (E. F. Johnson type 108-740) on 11/16 inch centers in the Teflon connector insulator illustrated in **Fig. 4** (if Teflon is not available, Polystyrene may be used). This "socket" is installed over the large oblong hole cut in the end of the chassis. The screw holes used for mounting the connector insulator are match-drilled to the holes in the insulator itself. In this way it is properly mated to the enclosure. Although nylon attaching screws were used in the original model, they are not necessary and regular metallic screws will not alter any of the oscillator's characteristics.

To reduce stray capacity to a minimum, the oscillator circuitry is mounted on an insulating sheet. Teflon was used for this purpose in the original unit, but epoxy board with the copper peeled off would be perfectly suitable. Sheet polystyrene is not too desirable in this location because of its susceptibility to heat. The envelope of the 6CW4 gets very hot and the ambient temperature within the confines of the small chassis is quite high.

The general layout of the Teflon oscillator support and associated support angle are also shown in **Fig. 4.** The angle is cut out from a piece of 1/16 inch aluminum sheet and bent in a vise to form the angle. When these two pieces are mated together, the metal angle will probably interfere with the lower nuvistor socket mounting screw. It must be drilled out using the hole in the oscillator support as a guide. This additional hole is not shown because its exact location will vary from unit to unit and depends upon the accuracy with which the parts are laid out.

The dial mechanism is not complicated, but it is hard to ascertain from the photographs exactly how it is put together. The exploded drawing of Fig. 6 should help in this respect. The vernier mechanism is an Eddystone 10:1 planetary drive that provides both smooth

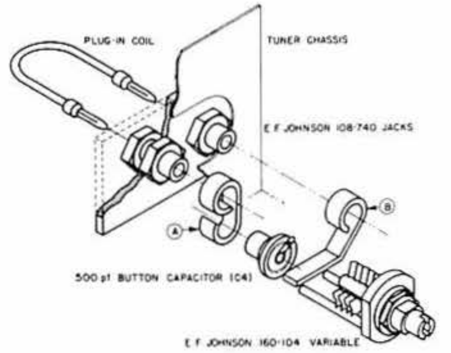
PLUG-IN COIL
TUNER CHASSIS
E F JOHNSON 108-740 JACKS
B
A
500 pf BUTTON CAPACITOR (C4)
E F JOHNSON 160-104 VARIABLE

BOTH THE BUTTON CAPACITOR AND VARIABLE CAPACITOR CONNECTING STRAPS (A AND B RESPECTIVELY) ARE MADE FROM 3/16" WIDE THIN COPPER STRIPS AND FORMED INTO SHAPE AS SHOWN. NOTE THAT THE STATOR SUPPORTING PINS ON THE VARIABLE CAPACITOR HAVE LENGTHENED SOMEWHAT FOR PURPOSES OF THIS ILLUSTRATION.

Fig. 5. Exploded view of the connections between the 6CW4 and the coil socket.

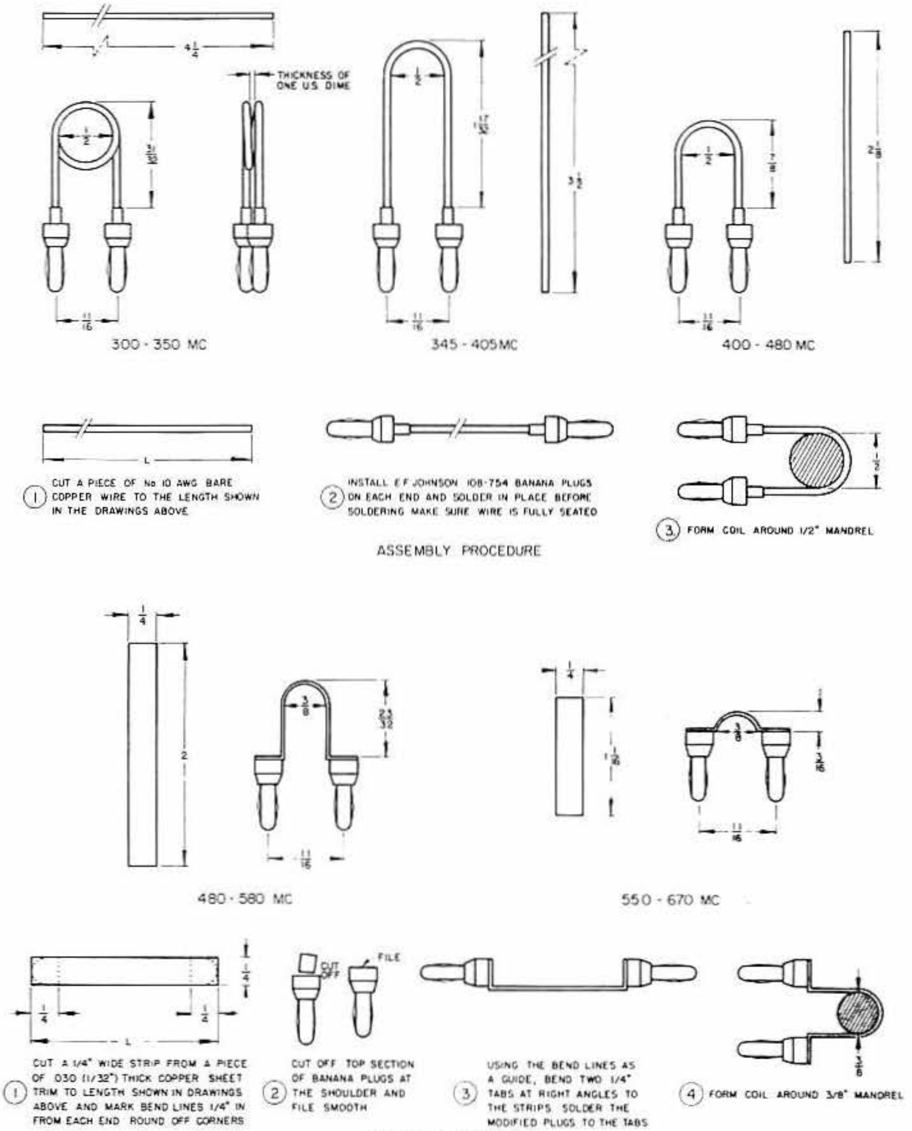

① CUT A PIECE OF No 10 AWG BARE COPPER WIRE TO THE LENGTH SHOWN IN THE DRAWINGS ABOVE

② INSTALL E F JOHNSON 108-754 BANANA PLUGS ON EACH END AND SOLDER IN PLACE. BEFORE SOLDERING MAKE SURE WIRE IS FULLY SEATED

③ FORM COIL AROUND 1/2" MANDREL

ASSEMBLY PROCEDURE

① CUT A 1/4" WIDE STRIP FROM A PIECE OF .030 (1/32") THICK COPPER SHEET. TRIM TO LENGTH SHOWN IN DRAWINGS ABOVE AND MARK BEND LINES 1/4" IN FROM EACH END. ROUND OFF CORNERS

② CUT OFF TOP SECTION OF BANANA PLUGS AT THE SHOULDER AND FILE SMOOTH

③ USING THE BEND LINES AS A GUIDE, BEND TWO 1/4" TABS AT RIGHT ANGLES TO THE STRIPS. SOLDER THE MODIFIED PLUGS TO THE TABS

④ FORM COIL AROUND 3/8" MANDREL

ASSEMBLY PROCEDURE

Fig. 7. The coils for the UHF grid dipper.

action and repeatability in a small package. Although this unit is manufactured in England, it is available from many of the larger electronics parts houses in this country.

Substitution of similar drives should be perfectly satisfactory as long as they don't extend more than one inch beyond the front panel of the tuner chasssis.

The vernier drive is connected to the variable capacitor through a ¼ inch polystyrene shaft 1-¼ inches long and the usual "¼ to ¼" shaft couplers. Polystyrene or some other insulator must be used here because the rotor of the capacitor must be isolated from ground. Because of the space limitations inside the enclosure, the coupler at the variable capacitor end of this shaft is only one-half of a standard coupler. A standard "¼ to ¼" coupler is sawed in two and one-half is epoxied to the end of the polystyrene shaft. Save the other half; it will be used for the dial drum hub.

The 8 pf variable capacitor was designed for screwdriver adjustment and its 3/16 inch shaft must be made compatible with the standard coupler. This is accomplished with a bushing made from sheet copper. A short piece of ¼ inch wide, 1/32 inch thick copper strap is formed around the capacitor shaft and takes up the slack between the shaft and the coupler.

The drum dial in the original unit was made from the metal top of a Johnson's Shoe Shine Kit (49¢ at the local grocer's), but any similar closed cylinder 1-15/16 inches in diameter and about 1-11/16 inches long should be suitable; other diameters will void the accuracy of the precalibrated dial. The "skirt" or bottom rim is cut off the metal can at the circumferential notch and a ¼ inch hole is drilled in the center of the top. The remaining half of the ¼ inch shaft coupler that was left over from the polystyrene shaft is then epoxied in place over the hole to provide a dial drum hub. Another hole is drilled in the side of the can 11/16 inches from the top; this provides access to the shaft coupler on the rear end of the vernier drive.

The precalibrated paper dial may now be cemented in place. It's a good idea to cement a piece of white paper the same size as the dial between the dial and the drum, otherwise the label on the can will show through the paper dial. Rubber cement is recommended at this point to prevent excessive wrinkling and distortion of the dial.

When all of the dial parts are completed, they are put together as shown in **Fig. 6**. It's a little crowded in the small box, but all of the parts *will* fit. However, in order to get all of the dial machinery into the box in the right order, a correct assembly sequence must be followed. First the polystyrene shaft and coupler are inserted into the dial drum from the rear. Next insert the Eddystone drive assembly through the hole in the front panel and mate it with the end of the polystyrene shaft. Place the bushing over the capacitor shaft, attach the polystyrene shaft and tighten the coupler; also tighten the coupler at the back end of the Eddystone drive thru the access hole provided in the dial drum. Install the vernier drive mounting screws. Now completely mesh the capacitor plates, center the low edge of the dial in the window and tighten up the dial drum hub. Disassembly must be accomplished in reverse order.

The tuning head is attached to the indicator/power unit through a four-conductor cable three feet long. This cable is attached to the box with a plastic cable clamp mounted in a hole provided immediately adjacent to the oscillator support angle and exits through a rubber grommeted hole at the front end of the enclosure. There is not enough room at the rubber grommet to use a cable clamp, so the cable is epoxied to the box at this point. Before installing the cable, however, check for sufficient clearance between it and the dial drum. It will probably be necessary to route the cable along the corner of the chassis to gain enough clearance.

Power indicator modulator

The power/indicator unit is housed in a standard 9 x 6 x 5 inch utility box (Bud type AU-1040 or equivalent). The layout of this

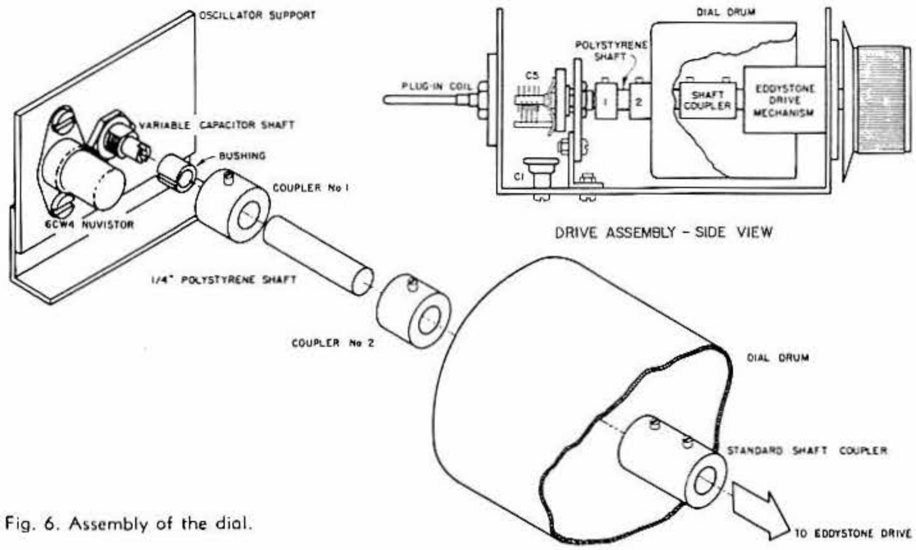

Fig. 6. Assembly of the dial.

circuitry is not at all critical, and just about anything that suits the builder may be used. The only particular caution that must be observed is with the transistorized 1 kc phase shift oscillator. This unit is built on a piece of perforated epoxy board (Vector 32AA18) 1-3/16 inches wide and 1-11/16 inches long. To preclude any 60 cycle pickup, this board is situated on the opposite side of the chassis from the power transformer.

In the author's case all the power/indicator components were mounted on a 4 x 5 x 1 inch aluminum chassis. This chassis was then mounted to the front panel of the utility box with the phone jack and power plug mounting nuts. Two large diameter holes (1-⅜ inch) are punched in the rear panel to provide access to the fuse holder and to pass the AC power plug. A chassis handle (Bud H-9168) on the top and rubber feet on the bottom just about complete the unit. There is one other addition however; five pairs of 5/32 inch holes, drilled on 11/16 inch centers along the rear edge of the top of the box provide convenient storage for the five frequency determining coils. A coat of spray paint and some Datak "Letraset" dry transfer labels are the finishing touches.

Calibration and operation

Without access to existing 420 mc equipment with known frequency characteristics, exact calibration in this band is impossible. However, this grid-dipper may be checked on the other ranges with the aid of an all-band television receiver. If the circuit layout and construction techniques described are closely followed, good correlation can be obtained and reasonable accuracy insured. The bands that are likely to be the furthest off are the lowest and the highest. In the lowest the coil spacing is quite critical and in the highest a slight change in length will move the frequency several megacycles. An accuracy of 2% at 600 mc is plus or minus 12 mc; close attention to the specified dimensions should provide accuracy better than this.

Operation of this grid-dipper is exactly the same as any lower frequency unit. It may be used in determining circuit resonance, detecting parasitic oscillations or as a signal generator. Because of its extended range, it has been found to be very useful in determining the series resonance point of rf chokes and ceramic bypass capacitors.

In the detect mode, this unit will indicate rf voltages as low as 50,000 microvolts. More sensitive operation may be obtained if it is used as an oscillating detector. In this case headphones are used and an audio beat note will be heard when the grid-dipper is tuned to the oscillator being checked. In the upper frequency ranges, it is usually difficult to obtain an actual beat note, and only a "tick" will be discernible when you tune by the frequency of the unknown energy.

When the oscillator is tuned between 605 and 650 mc, there is sufficient second harmonic energy to provide a strong reference signal in the 1215 mc amateur band. This is particularly useful in the initial tune-up of converters for this band. Use of the internal 1000 cycle modulation aids in distinguishing the grid-dipper signal from other rf sources that are present throughout the spectrum.

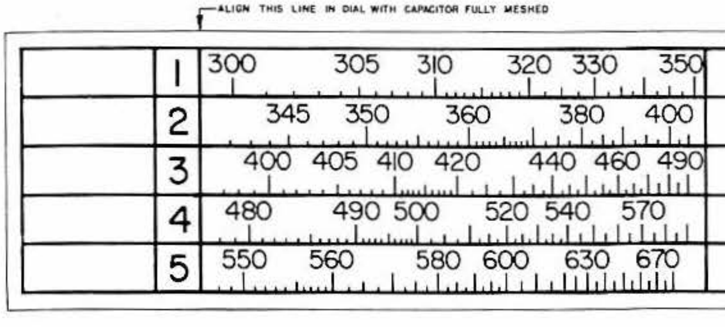

Fig. 8. Full size dial scale.

WIDE RANGE VHF-UHF DIPPER

Bill Hoisington K1CLL

Most dippers for amateurs that I have seen so far, not counting the $400 ones, stop around 200 MHz just as you are about to enter the fascinating UHF region. We do have the 432 and 1296 bands, so let's become more familiar with them.

After all these years of "grid-dipping" we find ourselves without a grid, so it just becomes a "dipper". To retain the prestige of a hyphenated name we can call it a "dipper-generator". Most grid-dippers have been used as generators, but this one has built-in modulation, variable input-output coupling, controlled Q, and several other interesting features. Best of all, it goes all the way up to 1296 MHz.

When this little unit is completed it may be used as a dipper for determining the resonant frequency of VHF and UHF circuits, as an indicating frequency meter with an adjustable Q-multiplier, a field strength meter and modulation monitor, a sensitive regenerative receiver, or a CW and MCW signal generator. You can also use it as a harmonic monitor or as a frequency transfer unit from one transmitter to another.

Several circuits must be considered when building a wide band instrument such as this. For example, you should change circuits around 100 MHz and again at 600 MHz, give or take a few hundred. Below 100 MHz coils are good; from there to 600 MHz you can use ¼ wave resonators, and after that the ½ wave job becomes rapidly the best method, up to 1300 MHz.

Plug-in rf heads

I have made no attempt to cover the complete range from 130 to 1300 MHz with one oscillator. By using plug-in tuners you may vary the components to suit the frequency. On 50 MHz for example, you may use a low cost transistor, a coil, and a 25 or 50 pF capacitor. From 100 to 600 MHz you use a better transistor, a ¼ wave strap, and a 10 or 15 pF capacitor. In the microwave region up to 1296 MHz you use the best transistor you've got, ½ wave lines,

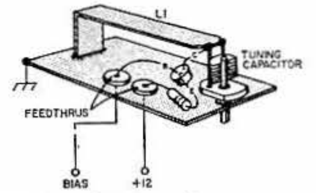

Fig. 1. Basic VHF/UHF oscillator circuit.

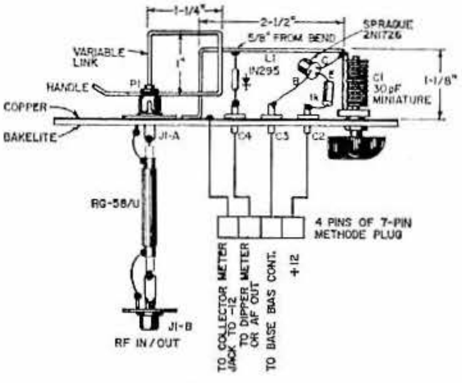
1-1/4"
2-1/2"
SPRAGUE 2N1726
5/8" FROM BEND
VARIABLE LINK
1"
L1
1N295
C
B
E
1k
C1 30pF MINIATURE
1-1/8"
HANDLE
P1
COPPER
BAKELITE
J1-A
C4
C3
C2
RG-58/U
4 PINS OF 7-PIN METHODE PLUG
TO COLLECTOR METER JACK TO −12
TO DIPPER METER OR AF OUT
TO BASE BIAS CONT.
+12
J1-B
RF IN/OUT

Fig. 2. 130 to 300 MHz tuning head.

and a small butterfly capacitor of 3 to 5 pF.

If you break the circuit at the right point, it simplifies things—then the two halves may be connected through a miniature 7-pin socket and plug as shown in Fig. 2. All four leads are reasonably dead to rf. You can leave out some of the audio if you like, but it's very handy to have a modulated signal. If you're running triple or quadruple conversion, it's nice to know by it's modulation which is the signal and which might be a *birdie*. As far as dials are concerned—it makes calibration and reading a lot easier to have only one band or range per dial.

130 to 300 MHz oscillator

Fig. 1 shows the basic ¼ wave circuit; Fig. 2 the complete rf unit with control, af output and modulation.

The circuit itself is very simplified, as seen in Fig. 1; there being only one inductance, L1, and no choke coils. This should make for a flat tuning oscillator without power dips as it is tuned over a 2 to 1 range in frequency, and it does just that. With a 2N1726 in the circuit there is a smooth power output curve from about one volt rf at 130 MHz down to ½ volt at 300 MHz.

The rf coupling jack J1 couples the rf energy both in and out. This is because L1 acts as either a detector resonator or an oscillator resonator, as required. Actually this rf jack can be used as shown in Fig. 2. P1 is a variable link to L1 and is plugged into J1; J1A has a few inches of cable between the white ABS plastic front panel and the copper clad bakelite sub-panel. Because the phono plug is rotatable, a nice variation in rf coupling can be obtained. The coax cable and J1B get the rf out to the front panel for easy use with antennas, probes, cables, etc.

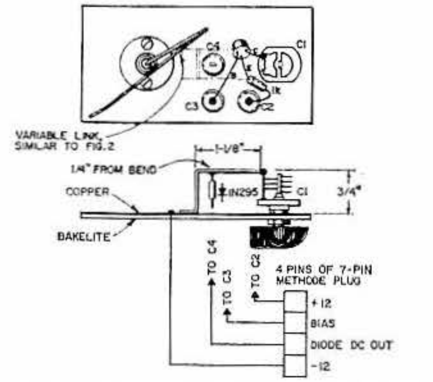
C4
C1
C
E
B
1k
C3
C2
VARIABLE LINK, SIMILAR TO FIG.2
1-1/8"
1/4" FROM BEND
COPPER
1N295
C1
3/4"
BAKELITE
TO C4
TO C3
TO C2
4 PINS OF 7-PIN METHODE PLUG
+12
BIAS
DIODE DC OUT
−12

Fig. 3. 300 to 600 MHz oscillator with variable link.

The emitter goes to a 1K resistor then through a coaxial bypass capacitor which gets the dc in and out and leaves the rf behind. These feed-through type bypasses are very necessary—do *not* skimp on this item.

300 to 600 MHz unit

Fig. 3 shows that this unit is essentially the same as the last, except for dimensions. I used a 2N1141 here although many others will work too. It tunes smoothly from 300 to 600 MHz; use the variable link feature as in Fig. 2.

900 to 1100 MHz

For this frequency range we need a little different approach. From Fig. 4 we can see that we now have two ½ wave lines on which low-voltage points can be found to attach the base and collector resistors. Most of the ¼ wave portion of the lines on the transistor end are actually inside the case. The places where the base resistor and the 500 ohm collector resistor are attached to the ½ wave lines can be found, or checked, by watching the rf meter and touching the lines with a pencil. At the proper point no change occurs in the rf output; sometimes it even increases.

The diode circuit of Fig. 6 is not ideal but it works. I have several of these around the shack and they work very well for detecting 1296 Mhz energy. Even ordinary hook-up wire will support the assembly of Fig. 6 about ¼ inch below the rf lines; you will soon find the best spot with the unit oscillating. The rf input jack and associated loop L3 are fastened so that L3 is in place over L1 and L2, and it's coupling can be varied in a semi-fixed fashion.

At this point we should mention that as a "dipper" the circuit is still working fine; also as a signal generator. It also serves as an rf detector but as the frequency gets up into the microwave ranges it is not quite as good as the tuned rf detectors featured in another article in 73 Magazine. Ideally, you should use the dipper on microwaves as a modulated generator and couple it into the unknown circuit; then a probe attached to *another* tuned detector should be coupled into the unknown circuit. There are quite a few variations using the dipper as an oscillator that you will find useful if you use a little ingenuity.

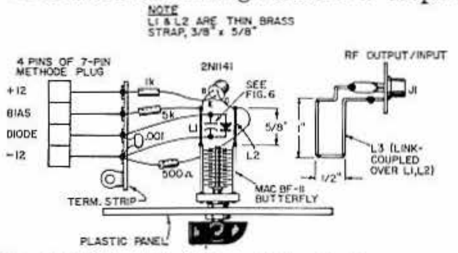
NOTE
L1 & L2 ARE THIN BRASS STRAP, 3/8" x 5/8"
4 PINS OF 7-PIN METHODE PLUG
+12
BIAS
DIODE
−12
2N1141
1k
5k
.001
500Ω
B
C
E
SEE FIG.6
L1
L2
5/8"
1"
RF OUTPUT/INPUT
J1
L3 (LINK-COUPLED OVER L1,L2)
1/2"
MAC BF-II BUTTERFLY
TERM. STRIP
PLASTIC PANEL

Fig. 4. 900 to 1120 MHz oscillator circuit.

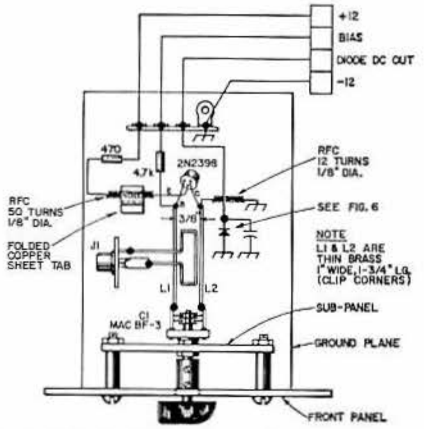
+12
BIAS
DIODE DC OUT
−12
470
4.7k
2N2398
E
B
C
RFC 12 TURNS 1/8" DIA.
RFC 50 TURNS 1/8" DIA.
3/8"
SEE FIG. 6
FOLDED COPPER SHEET TAB
J1
NOTE
L1 & L2 ARE THIN BRASS 1" WIDE, 1-3/4" LG. (CLIP CORNERS)
L1
L2
C1 MAC BF-3
SUB-PANEL
GROUND PLANE
FRONT PANEL

Fig. 5. 1200 to 1300 MHz oscillator and layout.

In the microwave detector line, my experience indicates that the plunger tuned coax cavity line is the best, the tuned trough line next, and the circuit of Fig. 5 next best. As a dipper, generator and regenerative receiver it is still good at 1296 MHz. Just to check, I plugged an antenna into J1, put an audio amplifier across the diode and copied a small transistor oscillator across the room. The base bias control works as a very smooth regeneration control. Smooth regeneration, as we will see later, is very important for maximum sensitivity when looking for harmonics and weak signals.

1200 to 1300 MHz unit

Fig. 5 shows the 1296 unit; I have used this circuit for many months as a dipper, variable-frequency generator, modulated-oscillator source, and as a regenerative receiver for 1296. In this circuit I used a negative dc grounded collector return. Don't short the base plate to the modulator base. Note that one end of the diode is tied to the base plate; this lead is brought out as the minus 12 volt lead. You can also use it ungrounded as in Fig. 4—you can use a 5th lead in the 7-pin plug and keep the diode isolated from the minus 12 volts. Suit yourself, just remember that *all* units have to use the same leads, as they all plug into the single modulator rf unit.

I just plugged a little 12 element Yagi antenna into this dipper and it works nice and smooth as a regenerative receiver. Please note, this is only for test purposes around the shack. You can hear with it, but not *that* good!

I had to put a choke in the cathode lead on this one, and tune it (the choke) with a piece of copper foil. A choke was needed in the collector lead too; after all this is the L-band microwave region.

The rf input jack J1 is mounted on a bakelite upright. Be careful of vertical metal pieces attached to the base plate; they only need to be two inches long or so to become Marconi antennas on 1296! Bring the base resistor and the collector choke away from the lines in a perpendicular fashion—it helps.

The total length of the diode and it's two leads, from ground to the tiny .001 capacitor C2 is about 1⅜ inches; it is spaced about 3/16 from the ground plane. The bottom edges of the lines are about ⅛ inch from the ground plane.

The transistor presently in the unit is a selected 2N2398; about half of the dozen or so I have here go to 1300 MHz, a couple go to 1400, and the rest to 1100 or 1200.

Don't be alarmed that L1 and L2 are longer than those of Fig. 5, the smaller butterfly capacitor does that. You can make a choice as to capacity, length of brass, and desired rf range. You can use a 5 or 10 pF capacitor for C1, shorter lines, and tune over 1300 MHz. In fact, I have reached 1600 MHz with this circuit!

Modulation and control

Fig. 7 shows the circuitry for bias control, modulation and audio. Don't let it scare you. It's just the same old deal of doing what has to be done for control purposes, and from then on just turning the knobs to get what you want.

I have found that a very good plug and jack can be made by using an ordinary 7 pin socket and a Methode 7 pin bakelite plug. Unfortunately, I have never found a miniature tube with a bakelite base; they are always made of glass, so you will probably have to buy the 7 pin plug.

The minus 12 volt lead goes to the collector meter jack and then to the minus 12 volt of the power supply. This puts the rf panel ground at minus 12 volts and the audio af panel at plus 12. Of course, you don't have to ground the plus 12 on the audio panel, I just have that habit.

The base return goes through a 5K resistor and then to the bias control potentiometer. The diode dc/af output goes to af jack J2 and the meter M1 through the meter sensitivity control. This potentiometer is selected to suit the meter; I have used a 10 K unit with a 500 microampere meter. Note that part of this resistor should be used when af output is desired, otherwise the meter shorts out the af.

Audio modulation

You might think that just about everyone knows how to build an audio oscillator. Mine did oscillate, but the tone! And the wave-shapes, hoo-boy! I used an af transformer with the collector on one side and the base on the other. So, once again to the handbooks and once again practically zilch. I did get the idea for a phase shift oscillator out of one of them, even if the circuit didn't work at first. After considerable experimentation I can recommend the circuit shown in Fig. 7. It works! In addition, the modulation may be adjusted to exactly 1000 MHz. This is very useful as many microwave test amplifiers have built-in narrow band af audio filters centered on 1 KHz. There is also a modulation gain control. This helps if a nice tone is desired.

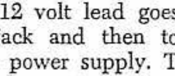

Fig. 6. RF detection loop for half-wave oscillators.

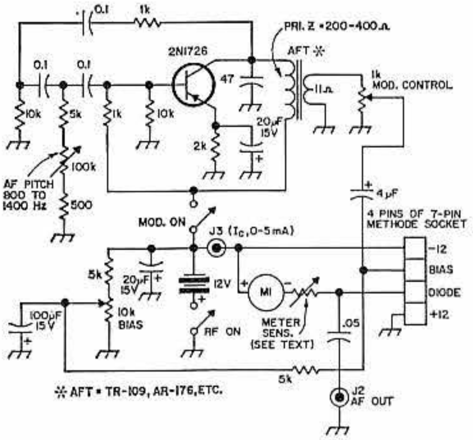

Fig. 7. Power control, modulation and audio circuitry. This circuit is used with all the rf heads.

Almost any small transistor output transformer will do the job for the af transformer, but don't go over 400 ohms impedance in the collector winding. Note the 1K resistor between the collector and the phase shift network; this reduces feedback to the base and may have to be increased or decreased depending on the gain of your transistor at 1000 Hertz.

I usually run the rf current (collector dc) between one and two mils with the 2N1726 transistors; other transistors may take more. A number 48 bulb in the collector lead may save you a $3 transistor. With the 2N1141 the rf output keeps climbing up to 4 or 5 mA collector current. You will readily find the best place to operate. When the collector mils keep climbing and the rf output starts to drop off, back off!

Operation

Dipper

As a general rule "dipping" is easier in the VHF region and gets more difficult as you go into the microwave region. On HF you couple one coil end on to the other; on VHF you bring it near, and on UHF you have to work to get the necessary coupling or use a probe.

With nothing near the dipper, swing the dial through the range to make sure it has no dips of it's own. The VHF/UHF units without chokes described here do not generally have such dips. Unwanted dips can be caused by chokes, resonant feed wire lengths, metal supports, and rf links and cables among other things.

When you do get a dip after coupling to the unknown circuit be sure and change the resonance of that circuit for a final check while watching the dipper meter. If the test circuit is a tuned circuit vary the tuning and see if the dipper will follow it—it *should*. If all else fails, use the dipper as a generator. Since it is very difficult to get the far end of a cable matched exactly over much of a frequency range, *expect* to find external dips in the dipper when using a cable.

Indicating frequency meter

Always keep the transistor plugged in and the base bias at zero so the diode is doing the work. When you advance the bias, collector capacity will cause the dipper frequency to change a little—more with some transistors than others.

For finding a weak signal you can use regeneration by turning up the base bias, but watch out for slight frequency changes. This regeneration can be very handy for finding weak oscillators or hard-to-find rf energy. Use the rf input loop with care; the least coupling is the best. Remember that some cables and terminations will detune L1.

Field strength meter and modulation monitor

The first part is obvious; use a small antenna or probe, get some signal in, and go ahead. Do not use any base bias to start with. If you are working with a very weak signal you might have to push the bias up for regeneration.

The modulation monitor is simply our often-described system of diode detector, transistor amplifier, and padded ear-phones. You can actually hear what your own transmitter sounds like to others. I use it on every new rig and after every circuit change. You don't need much of an antenna or probe when listening to your own rig; don't overload the diode when checking modulation. In fact, use light rf coupling and plenty of audio gain to hear yourself as others hear you.

Regenerative Receiver

Plug an audio amplifier into J2, Fig. 7, advance the bias control, and tune. I have heard several UHF TV stations from Massachusetts up here in Peterborough, so it is really sensitive. One nice feature of this circuit is that the regeneration turns into oscillation very smoothly. Stability is good too. You can heterodyne a crystal controlled two meter signal and copy CW with it. Not bad for a 144 MHz blooper!

To transfer signals from one transmitter (A) to another (B), just tune in A, then shut it off; listen for B and tune it in. That's all. Harmonic monitoring is easy: just tune over the suspected range in the regenerative condition. It is particularly good because only *one* frequency is present in the receiver. This is *not* the case when using a super-het receiver for monitoring harmonics.

Signal generator

One of the big features of this circuit is the presence of an rf meter right in the proper place circuit-wise. The modulation also helps, especially when running triple or quadruple conversion in a receiver. The modulation control is very convenient, at full on its spreads the signal across 20 or 30 kHz on a selective receiver. For checking a difficult to get at circuit, use a cable and probe, either capacitive or inductive, to get the signal into the unknown circuit.

I often use one of these units for antenna and receiver tests. I just plug a little two element beam into the rf jack and set it out away from the shack; often one or two hundred yards away. There is nothing like tuning up pre-amps with your antenna system connected. For antenna tests it is used in reverse.

EXTRA SERVICES FROM YOUR GRID-DIP OSCILLATOR

W. B. Cameron WA4UZM

One of the handiest instruments around the average ham shack is the grid dip oscillator, but many hams do not get all of the advantages this little instrument can offer. Most people know that it is useful for finding the resonant frequency of an LC circuit, or for generating a signal someplace in the range, although it is not quite a substitute for a well calibrated signal generator in this respect.

Many people do not realize that the grid dip oscillator makes an excellent crystal calibrator as well. The typical grid dip oscillator uses a two-terminal coil and a two-section variable capacitor, connected in the form of a Colpitts oscillator circuit, with the variable capacitor providing the voltage division for proper feedback. This same circuit, with a quartz crystal replacing the coil, becomes a very satisfactory crystal oscillator. In this case, the variable capacitor becomes a trimming capacitor capable of putting the crystal on the proper frequency with considerable accuracy and also allowing a fairly wide latitude higher or lower for checking bandpass.

The meter reading gives an index of crystal activity. What is required is an adapter to match the pins of the crystal holders to the coil socket of the grid dip oscillator. I have made up such an adapter for my particular unit which consists of a 1½ inch by 4½ inch piece of masonite which carries two banana plugs to insert into the oscillator socket, and several different sizes of crystal holders, all connected in parallel, to accommodate any crystals I happen to have around the shack. I carry in my tool kit various assorted crystals. These give me spot frequencies for band edges and other reference points on all of the amateur bands.

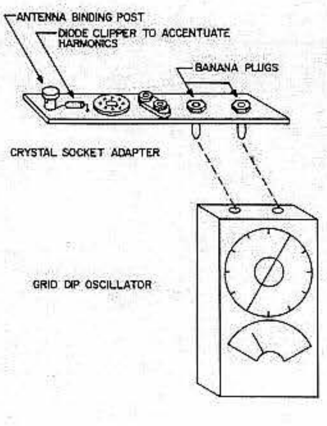

Crystal socket adapter for grid dip oscillator extends the unit's capability by enabling it to read crystal activity.

In checking the higher frequency bands, I find it useful to add a diode clipper and have one permanently wired on the same adapter board. This is a 1N69 which simply goes from one side of the crystal to a terminal point at which I can attach an antenna or a lead to the input jack of the receiver. By simply attaching a short length of wire approximating a quarter wavelength at this frequency, I can provide a strong enough signal from the shack to enable me to check out my direction finding loop in the automobile fifty feet away.

In tuning up my 2 meter FM gear, I plug in a 3061.25 crystal (which is the transmitter oscillator crystal for my Link 2m FM unit) and then couple from the 1N69 to the input of the receiver. This provides a signal stronger than most of the local signals on the band, and adequate for initial tune-up. For finer adjustment of the i-f strip, I clip the lead from the 1N69 to the chassis close to the antenna input jack and the leakage provides a signal of the order of .2 μV, which is useful for final adjustment of the critical tuned circuits, the squelch control, and others. For best results, I leave the oscillator and the receiver on for a half hour or so to warm up, and then with the variable capacitor in the grid dip oscillator, I zero-beat the incoming signal of a station generally considered to be on frequency. Once this is done, I have a signal source on the bench with which to check the receiver for passband balance as indicated on the microammeter connected to the discriminator output.

GDO TO FIND C

F. C. Rayer G3OGR

Most amateurs have a grid dip oscillator lying around, and an easy accessory can allow it to be used to read the values of unknown capacitances up to about 1000 pF. This is useful for unmarked surplus, those with obliterated markings, or to find the swing of small variables. Or we may check the best value found in some circuit position with a preset or variable, then measure this and substitute a near value fixed capacitor.

Figure 1 is the circuit. C1 and C2 are on the lid of an insulated box, carrying also spring terminals for CX. The capacitors have good knobs with pointers. Coil L can be half a dozen turns of stout wire, self-supporting, or anything which comes within a convenient range of the GDO (say 2.5–10 MHz) with both variables fully closed.

To calibrate, close C1 and C2 fully. Tune the GDO for the usual dip. Note the frequency on the box for future use. Take a few 1% capacitors, such as 100 pF, 200 pF, and so on, up to a total of about 1000 pF. Clip one to CX. Open C2 to restore the dip on the GDO. Mark the capacitor value on C2 dial. Series and parallel capacitors give more values. For example, 100 pF plus 200 pF in parallel gives 300 pF, while 500 plus 200 gives 700 pF, and so on.

Restore C2 to its fully closed mark. Repeat to calibrate C1, this time using capacitors such as 5 pF, etc., up to about 50 pF. When you see how the scales mark, estimate intermediate markings, to fill in.

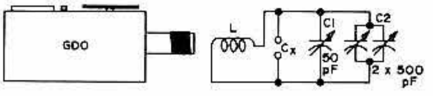

Fig. 1. GDO to find C.

Once calibration is finished, it is easy to find a capacitance value from about 2.5 to 1000 pF. Close C1 and C2. Put the GDO near L and tune the GDO for dip. Clip the unknown capacitor across CX. Open C1 or C2, as appropriate, to restore the dip. Read off the value from the scale. That's all there is to it!

GDO COIL EXTENSION

Bill Turner WA0ABI

A grid dipper is a great piece of test equipment and belongs wherever rf equipment is to be designed, constructed, or serviced. All writings on the subject seem to be prefaced with this comment, and I hesitate to break tradition. Now that the formalities are over, we will proceed with the subject matter.

Probably the one most inconvenient feature of all grid dip meters is the physical size of the instrument. This holds true for the two unit commercial models as well as the more common self-contained types. It is not so much that they are large but rather that the head will not always fit into the nooks and crannies which seem to be designed into most equipment. This situation will more than likely become more severe as integrated circuits come into more general use. At lower frequencies the coil assembly is usually long enough to reach to within shouting distance of the desired circuit, but as the frequency is increased the dipper coil shrinks to perhaps an inch protruding from the case. Try getting this even near a circuit buried 2 or 3 in. deep in a chassis and you will fully appreciate the magnitude of the problem.

There is a way of relieving the difficulty which requires neither a great deal of time nor expense. It is only necessary to put to use some basic theory which each of us was required to know in order to get a ham license. Required are perhaps a 3 ft length of small coax (RG 174/U), 6 in. of hookup wire and a ballpoint pen casing. The innards of the pen are removed, and the business end reamed slightly to accept the ends of a ½ in. diameter one turn link formed of hookup wire. Inside the barrel of the pen the ends are soldered to the coax, insulated, and cemented in place. The remaining wire is formed into a 2 turn link which fits snugly on the coil form of your grid dipper. This link is soldered to the opposite end of the coax and secured. Shrinkable tubing is perfect to cover the joint.

The next time you face this perplexing problem hold your head high and fear not. Merely slip the link over the coil form, place your probe so as to allow coupling into the desired circuit and proceed as usual. The only difference will be a more or less fixed load on the oscillator which doesn't affect its ability to indicate resonance. As always, best results are obtained with minimum coupling.

Chapter VIII

Noise Generators

SIMPLE NOISE GENERATOR

George Rubis K9ONT

As anyone knows that has done any work at all on receivers, whether it is a conversion or simply substituting a "hotter" tube in the front end, we get to the point where we begin to wonder if the adaptation was worth while or have we been fooling ourselves.

A noise generator using one of the noise diodes (IN21 or IN23) can give an indication if any improvement has been made.

The circuit is straight forward, but with one addition that others that I have seen do not have. The voltage is regulated by a Zener diode.

The reason is obvious to anyone who has worked with the simpler type of noise generator. The results are not always consistent from measurement to measurement and from day to day. The voltage and current vary with the setting of the variable resistance and due to the normal aging of the battery.

The Zener Diode eliminates this by maintaining a constant line voltage. In our particular instrument it is six volts. Of course we must use a battery in excess of six volts. Nine volts is a good value. I have found that used transistor radio batteries still have enough life in most instances to last for many tests.

One of the main requisites of a noise generator is that it must be shielded throughout. Therefore we must give some thought as to the placement of the various components.

A Mini-box 2¼ x 2¼ x 4 is an ideal size. As for a connector I used the SO-239 coaxial. I find that this connector allows more flexibility than any other. If a direct connection to the receiver is desired merely attach it through the double connector type DKF-2 made by Dow Key. On the other hand if it is desired to have the controls of the noise generator close at hand merely connect a length of Coax of eighteen inches or so. I haven't been able to discover that it has affected any measurements to any degree.

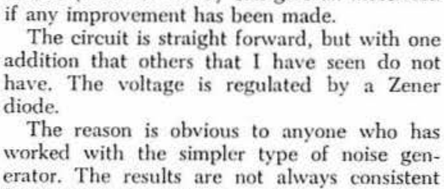

To avoid excess wear and tear on the zener diode and the battery a 200 ohm resistor should be inserted between the 9 v battery and S1.

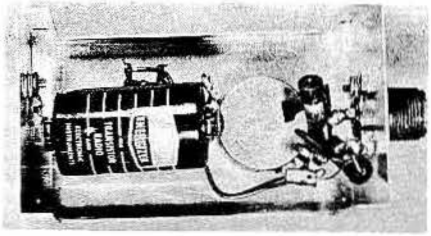

In the construction of this noise generator just remember a few basic rules. Keep all connections as short as possible. The noise diode and bypass condenser and resistor (50 or 75 ohms as the case may be) as close as possible to the output plug. Remember to use pliers to absorb the heat when soldering the leads of the diodes.

To mount the silicon diode, which has one large end and one small, we must improvise to a certain extent. For the small end a lug from one of the old tube sockets will do. For the large end use a small fuse clip.

Don't be too fussy about the variable resistor. For most purposes any value from 10M up to 50M can be used.

The battery you choose will determine the manner of mounting.

No need to give detailed instructions as to the use of this noise generator. There are ample instructions to be found in various magazines as well as handbooks.

. . . K9ONT

Parts List

2—¼ x 2¼ x 4 Minibox
Bt—9 volt battery
Cr1—6 Volt Zener Diode
Cr2—IN 21 or IN 23 Silicon Diode
R1—10M-50M Variable
R2—51 ohm or 75 ohm (according to your line)
C1—.001 to .005 disk ceramic
S1—S.P.ST. this may be on your variable resistance
P1—So-239

A CRYSTAL DIODE NOISE GENERATOR

Karl Tipple W5TEV

Most amateurs who do very much operating at the higher frequencies where receiver sensitivity rather than QRM often determines success eventually come to the conclusion that their present receiving equipment is not as sensitive as would be desirable and that a change is in order. And once that change has been made, whether it takes the form of a new converter or modification of existing rf stages, one nearly always wonders whether the new really *is* better than the old. Comparisons of signal strength with new and old equipment are erratic and unreliable at best and do not necessarily give the desired information since the absolute output level or "S" meter reading is not the problem. What is actually of concern is whether the signal to noise ratio of the receiving system has been improved. Or in other words, does a given rf signal produce an audio output signal from the new receiving equipment that is louder in comparison to the background noise than did the same input signal with the original receiving equipment. If it does, the noise figure of the new receiving equipment is smaller (better) and it should be possible to copy a weaker signal than before.

Unfortunately few of us own or have access to a signal generator of the proper frequency range, with an accurately calibrated attenuator and shielded output, for making the type of measurements mentioned above. For the majority of amateurs something less expensive and more readily available is needed. A crystal noise generator, which can be constructed for a few dollars and which will enable anyone to make before and after comparisons of noise figure, is such a device.

Briefly, the theory behind the use of the noise generator follows. The sensitivity of a receiver is highly dependent on the ability of the rf stages to amplify a weak signal as much as possible while adding a minimum amount of noise to the signal. Noise of this type will

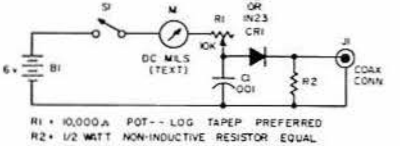

R1 = 10,000Ω POT -- LOG TAPEP PREFERRED
R2 = 1/2 WATT NON-INDUCTIVE RESISTOR EQUAL TO LINE IMPEDANCE

FIG. 1

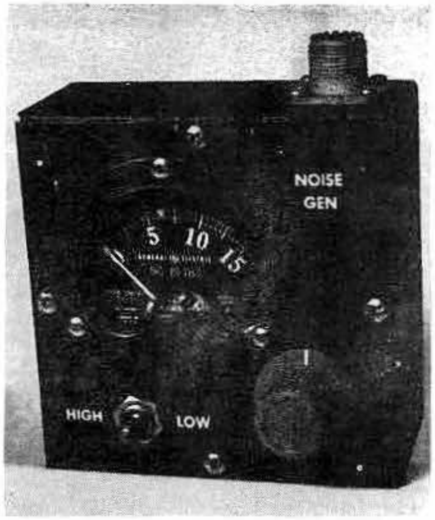

be heard at the receiver output as a familiar hiss or "rushing" sound. Now, if an external source of broadband or white noise (random noise independent of frequency) is connected to the receiver antenna terminals, an increase in noise or hiss will be noted at the receiver output. And if the noise generator output is adjusted to provide exactly twice as much receiver output noise as existed before the application of the noise generator, then the generator must be providing a noise signal equal to that generated in the receiver. Thus, if the receiver can be modified so as to reduce the noise generator signal required to double the receiver noise output, the noise figure of the receiver will have been improved.

A noise generator of the type described in this article can not be used for measurement of absolute noise figure without additional calibrating equipment. It can, however, be used for *comparison* of noise figures. It enables one to determine whether a receiver really has been improved and it can be used to determine which of a number of possible modifications will yield the best results in terms of minimum noise figure (maximum sensitivity). Use of the device will be discussed in more detail later.

A basic noise generator circuit is shown in Fig. 1. The noise is produced by CR1, a silicon diode of the 1N21 or 1N23 variety. These diodes have been widely available on the surplus market for several years. They are also readily available new at small cost; however as will be explained later, it may be desirable to use a surplus diode rather than a new one.

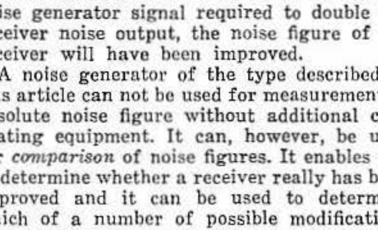

FIG. 2

R1 = 10,000 ohm pot—log taper preferred
R2 = 1/2 watt non inductive resistor equal to line impedance
Other parts same as Fig. 1

The current supplied to the diode from the battery and therefore the noise generated by the diode is determined by the setting of R1. The best value for R1 is necessarily a compromise since in its maximum resistance position it must limit the current to a low value but it must also not change resistance too rapidly at the low resistance end of the range or the normal operating range of the instrument will be limited to only a few degrees rotation of the R1 shaft. A 10,000 ohm log taper pot should prove satisfactory.

The value of R2 should be about the same as the transmission line impedance. A good value might be 56 or 68 ohms, although if it is desired to use the noise generator with equipment which normally operates from higher impedance lines, a larger value should be selected for R2. In this case one might compromise with a value of 100 ohms. C1 serves as bypass capacitor and may be either mica or disc ceramic.

A more elaborate arrangement was used in the construction of the unit shown in Photo. 3. A schematic of this generator appears in Fig. 2. Here a tapped battery (2 pencells) was used to minimize the range to be covered by R1. It should also be noted that the diode is shown reverse biased in Fig. 2. Many of the surplus diodes available have very poor reverse resistances, which will allow reverse bias operation, and when operated under these conditions will generate considerably more noise for the same bias current than when operated in the more conventional configuration shown in Fig. 1. The diode must have a reverse resistance of only a few thousand ohms or less if it is to work successfully in the reverse bias connection. Therefore unless the reader can obtain some surplus diodes and select one with a low reverse resistance, it will probably be necessary to operate the diode forward biased and to use the higher battery voltage of Fig. 1.

The meter range depends upon the battery voltage and the resistance of the diode in whichever connection it is used. The meter used in Fig. 2 was a 0-4 ma surplus meter. However, if the diode is operated in the forward biased connection a meter of somewhat higher current rating will probably be needed. Therefore, it is suggested that a low range meter be obtained and if more than full scale current is required for the diode to develop sufficient noise, a shunt resistance can be placed across the meter terminals to extend the meter range. An inexpensive meter can be used since all that is needed is a relative indication of current.

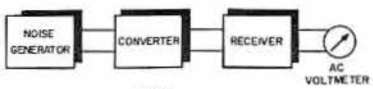

FIG. 3

For those not familiar with the IN21-1N23 type diodes it might be mentioned that they have no wire leads. The anode connection is a brass shell on one end of the unit and the cathode connection is a brass pin at the other end. Although external connecting wires might be soldered to the brass ends, there is danger that the diode will be over heated and destroyed by such techniques. It is safer to make connections by clamps made from small plate or grid caps or fahnstock clips.

It should be evident from the above discussion that there is nothing critical about the construction of a noise generator of this kind and that a number of different circuit variations are possible. Even if the device were constructed entirely of new components the cost should be less than $8 and with a little ingenuity the cost can be reduced substantially through the use of surplus parts.

Perhaps use of the noise generator can best be illustrated by considering an example of noise generator application. Suppose that you want to use either one of two available 6 meter converters with your low band receiver and that you wish to determine which converter will provide the best weak signal reception. A test set-up is shown in Fig. 3.

The ac voltmeter is used to measure the audio output of the receiver and may be either a multimeter or a VTVM. It can be connected across the speaker terminals or across the headphone line. Higher voltages can generally be obtained from the headphone output and this connection will probably have to be used if a multimeter without a very low ac range is used.

The measurements can be started by measuring the ac noise output voltage from the receiver with the noise generator turned "off". Then the noise generator should be turned "on" and the generator level adjusted until the output meter reading has doubled. At this point a note should be made of the noise generator setting. The above procedure can now be repeated with the second converter connected to the receiver. Whichever receiver-converter combination requires the smaller noise generator current setting to double the noise output is the combination that will provide the best weak signal reception.

The noise passed by the receiver is a function of the bandwidth of the receiver. Also the reading given by a particular output meter is dependent on its bandwidth. Therefore the bandwidth of the receiver should not be changed during measurements and the same output meter should be used thoughout a set of measurements.

The technique described above can also be used to determine which of a number of rf amplifier tubes will give the best performance. The noise generator can be used for alignment of rf stages.

For the man who constructs his own converters or receivers the device is a tremendous boon since such adjustments as determining the optimum antenna to grid coil coupling can be made simply by adjusting for maximum increase in noise when the generator is switched "on". (This technique gives the same results as adjusting for minimum noise generator setting to double noise output and it is a bit easier to use when adjustments must be made.) If the input circuit uses a single tapped coil instead of two coils, the best tap position can be determined.

The uses mentioned above are only a few of the possible applications of this noise generator, and for the amateur who constructs his own receiving equipment the device is indispensable.

TEMPERATURE LIMITED DIODE NOISE GENERATOR

F. L. Thomas

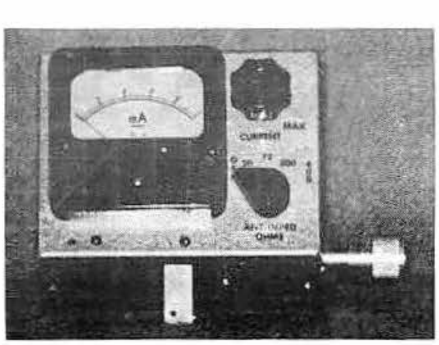

Recently I was confronted with the necessity for a good noise generator. All of the noise generator designs available in any of the amateur publications at hand were of the crystal diode type. The disadvantage of this type of instrument lies in the fact that the current through the diode has no simple relation to the noise output. Unless expensive calibration equipment is on hand this type of generator is useful only on a comparative basis. Consequently it was decided that a temperature limited diode noise generator would be built.

The noise output of a temperature limited diode noise generator is simply related to the current flowing through the diode.[1] The noise figure of a receiver may be calculated directly from the magnitude of the current by the following equation:[2]

Noise figure in db.$=10 \log (20\ I\ R)-(1)$

Where I=current through temperature limited diode required to make the noise output power of the receiver double the value it was with no current through the diode.

R=antenna impedance

The actual noise generator is quite simple, consisting of half a 6AL5 with a milliammeter in the cathode leg and the appropriate resistance for the antenna circuit connected to the plate by means of a selector switch. The heater current is controlled by a 20 ohm potentiometer in series with the heater. The unit is battery operated for convenience. High current capacity, small sized mercury cells are used throughout. The whole unit is contained in a 3" x 4" x 5" Minibox.

Construction

The resistors are mounted on the switch, making the leads as short as possible. The tube socket is mounted on a small bracket screwed onto the side of the chassis, placed so that the distance from the switch is a minimum. The meter calibration potentiometer and the heater supply battery holders are mounted on another bracket placed over the meter as shown in the photograph.

Operation

To use the generator it is connected to the receiver antenna terminals, the AVC is turned off, and the audio output is measured with the generator off. The generator is then turned on and the current through the diode is increased until the power output is double what it was before. At this point the voltage output will only be 1.41 times the original. The current necessary to give this noise increase is noted and the noise figure calculated according to equation (1), or estimated from the figures given in Table I.

Noise Figure in db.	Current, ma., for antenna impedance of:			
	50Ω	72Ω	300Ω	600Ω
2	1.40	1.15	-	-
4	2.25	1.80	0.42	-
6	3.55	2.85	-	-
8	5.65	4.50	1.05	0.53
10	8.85	7.15	-	-
12	14.2	11.3	2.65	1.32
16	-	-	6.65	3.30
20	-	-	16.6	8.42

The unit described is useful for noise figures up to 13 db at 72 ohms. For higher noise figures a higher current capacity diode and a higher heater voltage and current supply is required, necessitating an ac operated power supply. With a suitable power supply a 6X4 may be satisfactorily operated up to a noise figure of about 22 db. In both cases the maximum noise is obtained by passing more than the allowable average current through the diode. Consequently readings at these extremes should be made and the current lowered again in only a few seconds, or damage to the diode will result.

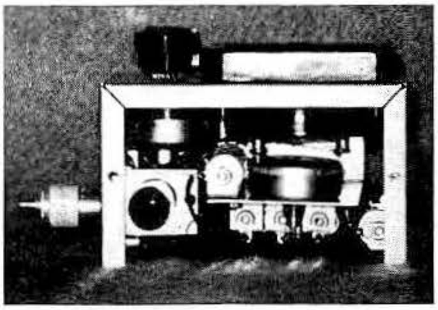

Table I shows the current readings for various noise figures with different antenna impedances.

The generator will operate satisfactorily up to at least 50 mc, and probably considerably higher.

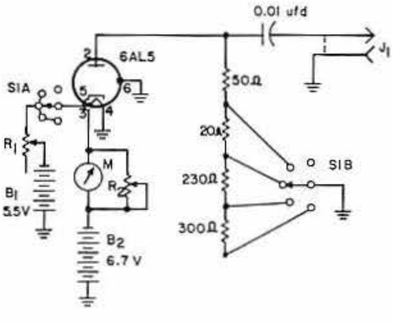

B1—four Mallory RM12 or RM12R in series (lifetime of this circuit should be greater than 10 operating hours)
B2—Mallory TR-135R
M—0-1 ma (Lafayette TM-60 or equivalent)
R1—20 ohm potentiometer
R2—20 ohm Potentiometer
J1—coax connector
S1—DP 5 pos miniature rotary switch

Editor's Note: By substituting flashlight batteries for the mercury cells and a 20 mA meter instead of the 1 mA meter with a shunt across it, the cost of construction can be greatly reduced.

REFERENCES

1. **Radiotron Designers Handbook**, fourth edition, page 1307, reproduced and distributed by RCA Victor Division, Radio Corporation of America, Harrison, N. J., 1953.
2. **Radio Engineering Handbook**, fifth edition, page 19-10, Edited by Keith Henney, Published by McGraw-Hill Book Company, 1959.
3. Terman, F. E., **Electronic and Radio Engineering**, fourth edition, Published by McGraw-Hill Book Company, 1955.

CALIBRATED NOISE GENERATOR FOR 432

Hank Olson W6GXN

As pointed out in previous articles—the temperature-limited noise diode is the best tool for measuring the sensitivity of one's receiver. The measurement made with this device is "noise-figure" or "noise-factor" and this is the magic number by which you can compare your VHF or UHF receiver with Joe's down the street, or Sam's in Massachusetts. These measurements will be comparable between any two receivers because bandwidth, type of detector, and other miscellaneous features (different for each receiver) do not affect the measurement, if it's carefully made (that is, if one's measurement of 3 db power increase, when the diode is turned on, is true).

Past amateur articles on noise-diodes and their use in noise-factor measurement have only shown how to construct units for up to about 200 mc. The noise generators described previously, for amateur use, have used the 5722, the diode wired 24G (3C24), or the diode wired 801A, as their tungsten-filament diodes. The 15E and 01A have also been used occasionally as diode-wired triodes, too. Commercial UHF noise generators use a variety of tungsten-filament tubes, all of which are a bit expensive for most hams. If you can lay your hands on one of these tubes, used or otherwise surplus, by all means employ it in your noise generator. Table I is a list of such gems.

Philco:	L1262A
RCA:	R6212
Bendix:	6144
Marconi:	CV2171

Table I

The tube I used is a triode, a Western Electric 708A, originally designed for grounded-grid UHF amplifier service. The grid and filament are used as the diode elements, ignoring the plate altogether; the plate, if used, could only increase transit time and shunt capacity. The tube is used, as it was intended to be used, with the metal shell (grid) grounded. The 50 Ω load is connected in the filament circuit; the filament power is fed in by means of a concentric inductor, which also tunes out the stray capacity of the tube.

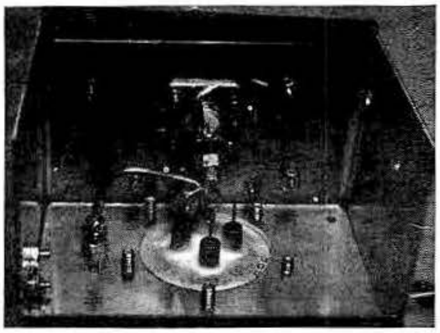

RF Circuitry.

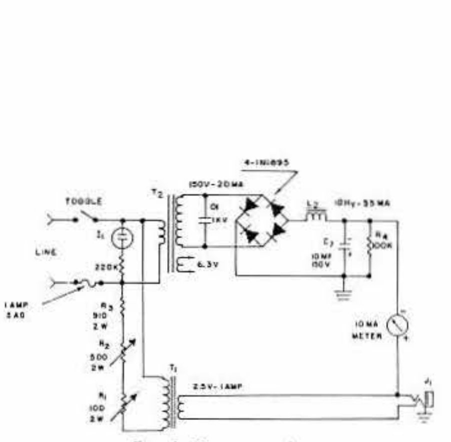

Fig. 1. Power supply.

C_1, a small "tweaker," adds in a tiny additional capacity to make adjustment to 432 mc easier; it makes tuning to anywhere in the 420 to 450 mc band possible. L_1 is constructed of a 1⅞" length of ⅛" copper tubing and has a piece of No. 20 teflon insulated wire inside it. The use of teflon insulated wire is only necessary because teflon will withstand the heat of soldering.

The effective circuit, then, is as in Figure 2.

At 432 mc, if $Cgk + C_1 = 3$ mmfd, then L must be 0.04 μh to be parallel resonant. The reactance of either $Cgk + C_1$ or L_1 is about 100 Ω, so the system has a Q of ½, and hence will be rather broad in its noise output spectrum—just as we want it to be.

The WE 708A tube was recently available from a Los Angeles surplus emporium at the price of 39 cents each or ten for a dollar. We bought a buck's worth, figuring some would be NG, but all were perfect and saturated well. One was lost in initial test, when we applied too much filament voltage; it was subsequently hack-sawed open to find out the details of its construction and to confirm connections. The details learned are presented in Figure 3 along with its saturation curve.

The filament is a single, fine, straight, tungsten wire through the grid helix. The grid helix is perhaps 1/16" diameter and is welded every turn to the shell. All this adds up to: good cylindrical diode configuration, close cathode-grid spacing to cut down transit time, and low grid to case inductance. In short, we have a nearly ideal noise diode for a dime a piece.

Construction details: The rf section of the generator is constructed on an aluminum plate bent into an L, the WE 708A protrudes through a 1⅞" round hole to expose its filament pins next to where the UG58A/U connector is mounted on the other side of the L. The WE 708A is held in place by five 8-32 binding-head screws that are tapped into the plate. The UG58A/U has four 200Ω, ½ w resistors soldered to it each at 90° to its neighbors to form a less inductive load, approximating a resistive sheet. These resistors are soldered to the UG58/U before it is mounted to the aluminum sheet to make soldering easier (less heat required). Then L_1 is formed and soldered from the WE 708A filament pin to C_3. The center filament lead of L_1 is fed through and soldered to the other filament pin and to C_4. C_2 is then soldered in; be sure this is the type called for or a similar low inductance stand-off ceramic. The rest is straight forward. Diode and power supply were each built in an LMB 141 box chassis. The details of wiring the diode circuitry are shown in Fig. 4.

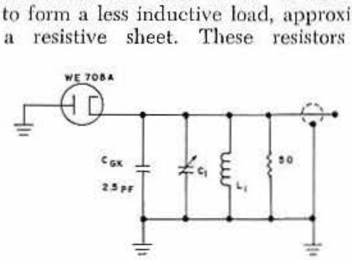

Fig. 2. Basic noise generator circuit.

A word about R_1 (the "fine" adjust) is worthwhile. Make *sure* this one is a 2w type A.B. (ohmite) molded carbon pot, if not both R_1 and R_2. This will make smooth diode-plate current adjustment easy; a wire wound pot will cause the plate current to vary in steps because of the effect of the pot's sliding contact sequencially contacting each wire (the same applies if you use a Variac).

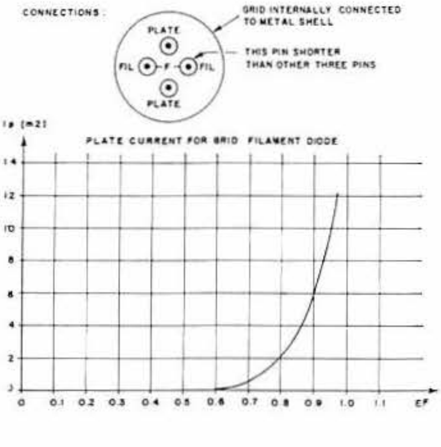

FIG 3

Fig. 3. Connections of 708A and graph of relationship between "plate" current and filament voltage.

To align on 432 mc, C_2 is temporarily removed and a UHF grid dip meter coupled to L_1, loosely. C_1 is adjusted for a dip at 432 mc. Then C_2 is reinstalled, and we should be ready for receiver checks.

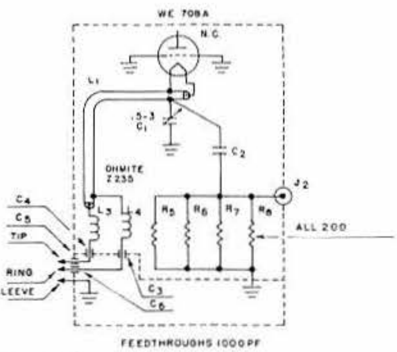

Fig. 4. Circuit of diode noise generator.

The above noise generator was compared with a Hewlett Packard 343A noise diode using my own 432 mc converter as the "to be measured" device. The results showed less than 0.5 db difference.

The author wishes to thank Gene Howell, W4RLU, for his photography of the unit.

ANOTHER WAY TO MEASURE NOISE FIGURE

Jim Kyle K5JKX

Many times it's been said, but it can always bear repeating. Noise figure is probably the least-well-known measurement in any amateur station.

At least a part of the difficulty with noise figure lies in the means usually employed to measure it. While it's not too difficult to figure out the power input to your final, or even the power *output* (often a surprisingly different figure), measurement of receiver noise figure tends to be a complicated and somewhat inaccurate process at best. It requires special equipment, and even then may be no more accurate than plus-or-minus 100 percent.

The classic means of measuring noise figure is to use a noise generator and crank in additional noise until receiver output is doubled. This means, of course, that the noise generator output is then exactly equal to the original noise, and if the noise-generator output is accurately known then the original noise is also known. All this has been gone into in detail in another article.

However, a noise generator with accurate calibration isn't so easy to come by, and an inaccurate noise generator doesn't do much good for measurement purposes (although it's fine for tune-up).

There *is* another way to do it, which is actually much more in line with amateur practice. This other way also requires some test equipment, but it might be more easy to come by.

Before we get into the details of the "other way" to measure noise figure, let's take another look at the reason for using noise-figure measurement as a yardstick for receiver sensitivity in the first place.

To start, we're really interested in the answer to the question "How weak a signal can I hear?" In the 3-30 mc range, the question can be answered directly—how many microvolts must the receiver have to give readable output?

As the vhf region is entered, though, the fractional microvolts become alarmingly small.

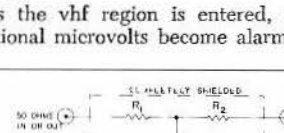

DB Loss	R1, R2	R3
3	8.2	150
6	16	68
9	24	39
20	39	10

All resistors composition, ½ watt, 5%

Fig. 1. Construction of T-pads

Somebody figured out that most of the problem lay in the receiver's own internal noise, and came up with the idea of a "perfect" receiver which would have no noise at all. This is a noise figure of 0 db. Now by comparing existing receivers to this perfect ideal, and comparing the internal noise in db, we had a way of discussing receiver sensitivity.

Since we're now talking about noise, which is equally present at all frequencies throughout the spectrum, we can see that the amount of noise present in a receiver's output is at least partially determined by how much of the spectrum we are looking at. A broad receiver has more noise output than a narrow one, all other things being equal. If you don't believe it, fiddle with the selectivity switch on your own rig and listen to the change of noise output.

This dependence of noise on bandwidth is another reason for using noise figure as a comparison. The *actual* amount of noise is cancelled out in the comparison, leaving only the relative amounts of noise in the "perfect" receiver and the receiver under test to be measured.

When all this became established, nobody was paying much attention to receiver bandwidth and it was felt that a true determination of the effective noise bandwidth of a receiver was much more complicated than the comparison measurement. However, in these days of SSB and special filters, that's not so true any more.

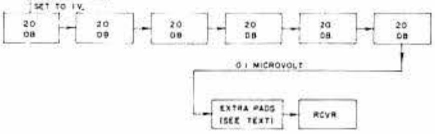

Fig. 2. Test set up.

As you may have guessed by now, the "other method" of determining noise figure depends on a microvolt measurement and knowledge of the receiver's effective noise bandwidth. The only reason for converting the results back to noise figure is to allow comparison with measurements made in the more conventional manner.

With typical ham measurement techniques, the results won't be of National-Bureau-of-Standards accuracy. However, if you're reasonably careful, results using this method will be at least comparable in accuracy to those made with a homebrew noise generator. Ready? Let's go:

You'll need two items of test equipment (only one if you're really lucky). These are an rf signal generator covering the desired frequency range on fundamental output, and an rf VTVM reasonably accurate at the desired frequency. If you have access to a "microvolter" or similar laboratory signal generator, you won't need the VTVM.

In addition, you'll need a whole handful of 50 ohm T-pads; these can easily be put together in a hurry by following the schematic in Fig. 1. You'll probably need about 8 20-db pads, as well as one each in 3-db, 6-db, and 12-db values.

Turn on both the receiver and the signal generator and let them warm up. For protection against any leakage from signal-generator to receiver through the power lines, it's best to supply them from separate circuits and to use a power-line filter such as that used to eliminate rf interference between the power line and the unit.

Connect a string of six 20-db pads to the signal-generator output as shown in Fig. 2, and adjust output of the signal generator to 0.1 volt. If you have a microvolter or equivalent, use only one 20-db pad and set generator output to 1 microvolt.

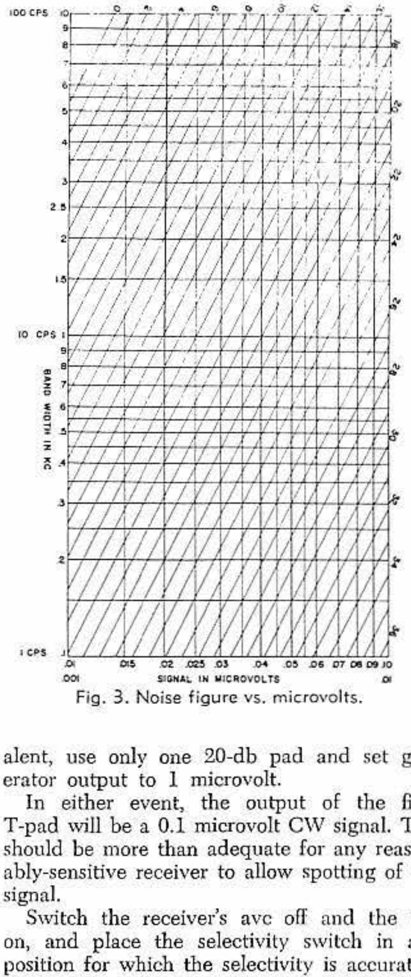

FIG. 3

Fig. 3. Noise figure vs. microvolts.

In either event, the output of the final T-pad will be a 0.1 microvolt CW signal. This should be more than adequate for any reasonably-sensitive receiver to allow spotting of the signal.

Switch the receiver's avc off and the bfo on, and place the selectivity switch in any position for which the selectivity is accurately known. The selectivity marked on the front panel will not be the effective noise bandwidth, but you can use it as a starting point to guesstimate the noise bandwidth. If your receiver uses a mechanical filter or other device with approximately the same skirt selectivity, effective noise bandwidth will be about 1½ times the bandwidth marked on the front panel. If it is one of the older types with reasonably broad skirts, noise bandwidth will be about 3 times the marked value. Both these correction factors are approximate, of course; if you have any means of measuring effective noise bandwidth, use it instead.

For a start, use a fairly broad selectivity position; this requires more signal and makes things a bit easier.

Now tune in the signal from the generator, leaving gain controls at maximum but tuning for maximum signal strength just as if it were the new state you need. The 0.1-microvolt signal should be easy to find.

Next step is to reduce the generator output by hooking in additional T-pads until you locate the point of "minimum discernible signal." The 3, 6, and 12 db pads may be hooked up in series in any combination to give you from 3 to 24 db additional attenuation in 3-db steps. Using another 20-db pad will give you from 20 to 44 db more attenuation, and the signal is sure to become too weak to copy before you reach 44 db below one-tenth of a microvolt!

The point of MDS is approximately equal to a 0-db signal-to-noise ratio for most of us, and is considerably easier to determine than would be a true output S/N ratio. When you find this point, record the db below 0.1 microvolt and the selectivity (in kilocycles) used.

Now switch to a different bandwidth on the receiver and repeat the test. Record its results also. For maximum accuracy, repeat each of the tests 10 to 12 times and average the result.

The signal level in microvolts corresponding to db below 0.1 microvolt is given in Table I. Locate it there and move to Fig. 3, the graph of signal versus bandwidth by noise figure.

Enter the graph from the side with effective noise bandwidth, and move across until you intersect the line corresponding to signal level in microvolts. The diagonal lines are noise figure; if one passes through the intersection point, read noise figure in db from it. If not, interpolate between the lines.

In reading Fig. 3, use the 10 kc-100 cps scale with the .01-.1 microvolt scale, and the 100 cps-1 cps scale with the .001-.01 microvolt scale. If your bandwidth-signal level combination falls off the graph to the left, use the *lower* signal-level scale with the *higher* bandwidth scale and subtract 20 db from the resulting noise figure.

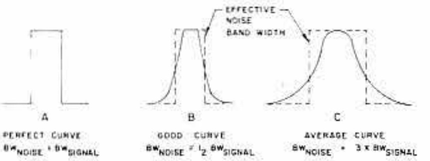
EFFECTIVE NOISE BAND WIDTH
A
PERFECT CURVE
$BW_{NOISE} = BW_{SIGNAL}$
B
GOOD CURVE
$BW_{NOISE} = \frac{1}{2} BW_{SIGNAL}$
C
AVERAGE CURVE
$BW_{NOISE} = 3 \times BW_{SIGNAL}$

Fig. 4. Relation of noise and signal bandwidths.

In the happy event that all your errors (and our approximations) cancel out, you'll find the noise figure to be the same at both the narrow and the broad bandwidth positions. However, it's more likely that you'll measure different noise figures at different positions of the selectivity control. It's safest to take the highest noise figure measured as being closest to correct, but you can average them if you prefer. Either way, you will probably be within 1 db of the real figure—and this is as accurate as most noise-generator techniques can be, also.

That completes the measurement, but before we wind this up let's take a more detailed look at the idea of "effective noise bandwidth" which is such a key part of this measurement technique.

Most of us are familiar with the idea of a "perfect" curve for receiver selectivity such as that shown at A in Fig. 4. Here the receiver has equal response over the desired band, and response drops to zero at the band edge. Such a curve is said to have a shape factor of 1, and is of course impossible to achieve in practice.

Now back to noise; it's spread out equally over the spectrum. A noise bandwidth of 1000 cycles per second contains 10 times as much noise as one of 100 cps. Thus "noise bandwidth" inherently has a shape factor of 1.

Since such a shape factor is impossible to achieve, it follows that "noise bandwidth" and actual receiver bandwidth must differ. If receiver bandwidth is measured at the —60 db points, the noise bandwidth will always be smaller than this receiver bandwidth. If receiver bandwidth is measured at the points where response drops 1 db below peak, the noise bandwidth will always be greater.

The mathematical expression for noise bandwidth is an integral equation involving differential gain, which is a cumbersome thing to solve. In general, the noise bandwidth of a receiver is said to be approximately equal to the bandwidth between points which are 3 db down from peak response.

In practice, if the shape factor (6 to 60 db) of the receiver is 2, the effective noise bandwidth will be approximately 1.3 times the 6-db bandwidth. If shape factor is between 2 and 10, noise bandwidth will be approximately equal to the square root of the shape factor (6 to 60 db) times the 6-db bandwidth. Few receivers have shape factors greater than 10.

db BELOW 0.1 MICROVOLT	MICROVOLTS
3	.07
6	.05
9	.035
12	.025
15	.018
18	.013
20	.01
21	.009
23	.007
24	.0063
26	.005
29	.0035
32	.0025
35	.0018
38	.0013
41	.0009

Table I.

The approximations quoted earlier (1.5 times marked bandwidth for SSB-selectivity receivers, 3 times marked bandwidth for others) are based on these relations. If you're really interested in calibrating your receiver's noise bandwidth for using this measurement technique, however, you might take a converter and have it measured for noise figure by the generator technique, then run this technique backwards to determine the effective noise bandwidth of your receiver in each position of the selectivity control.

The technique described here, incidentally, assumes that no audio filters are used following the detector. If they are, all results are off, since the effective noise bandwidth will have been changed in an unpredictable manner by the audio filters.

However, you can remove the audio filters from the hookup for measurement purposes, determine noise figure, then return the audio filters to the circuit and run the measurement backward to find out your effective noise bandwidth with filters present. Don't be surprised if it comes out in the region from 1 to 10 cycles per second; a good audio filter can work wonders with weak-signal reception.

For additional details on this technique of measuring noise figure, you can consult Reference Data for Radio Engineers, 4th edition, published by IT&T and available from Radio Bookshop, or any good radar text.

Chapter IX

Attenuators and Dummy Loads

LOW POWER ATTENUATORS FOR THE AMATEUR BANDS

George Daughters WB6AIG
Will Alexander WA6RDZ

In the evaluation of rf amplifiers, filters and many other devices, a variable attenuator is indispensable. This article describes attenuators built and tested by the authors. These attenuators are flat from dc to over 50 MHz and usable to over 450 MHz. They use low cost parts, are very simple to build, and are more accurate than ordinarily required in amateur applications.

The basic attenuator section is the symmetrical pi shown in **Fig. 1**. Resistance values are given by the relations:

$$R_1 = R_s\,(\sqrt{K} + 1)/(\sqrt{K} - 1)$$
$$R_2 = R_s\,(K - 1)/(2\sqrt{K})$$

where R_s is the characteristic impedance of the pad (equal to the source and load impedance) and K is the attenuation factor, P_{in}/P_{out}.

Resistor values for the most commonly used impedance (50 ohms) are shown in **Table 1.**

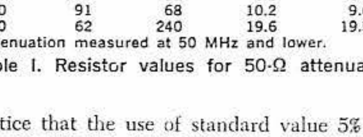

Fig. 1. Basic pi-network attenuator section.

Nominal atten-uation in dB	R1 in ohms	R2 in ohms	Calculated atten-uation in dB	Measured atten-uation in dB
1	910	6.2	1.1	1.1- 1.2
2	430	12	2.1	2.1- 2.3
3	300	18	3.0	3.2
6	150	39	6.2	6.3- 6.6
10	91	68	10.2	9.6-10.1
20	62	240	19.6	19.5-19.7

Annenuation measured at 50 MHz and lower.

Table I. Resistor values for 50-Ω attenuators.

Notice that the use of standard value 5% half watt composition resistors allows accuracy within 1 dB of the calculated value of attenuation and within 1 dB of the desired nominal value.

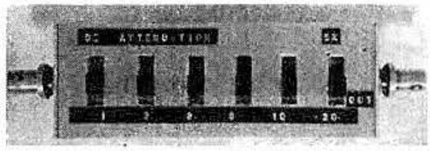

Amateur low-power attenuator made from inexpensive slide switches, 5% resistors and fiber glass, copper-clad board.

The attenuators are built in small channel boxes made of copper-clad etched circuit board material. Aluminum channel boxes commercially available would probably work equally well. Small, inexpensive DPDT slide switches (H. H. Smith No. 518 cr equivalent) are soldered directly to the copper board and the resistors are soldered to the switch terminals (which have been cut short) with the shortest leads possible.

Two wiring variations have been tried, one with the series resistors (R_2) mounted between the switch wiper contacts (type B), and one with all resistors connected to the attenuator in terminals (type A). See **Fig. 2** and the photo of the interior of the attenuator. It was found that the latter arrangement, type A,

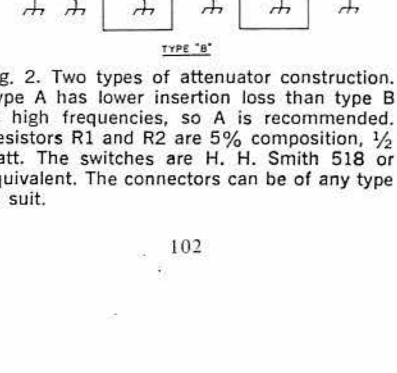

Fig. 2. Two types of attenuator construction. Type A has lower insertion loss than type B at high frequencies, so A is recommended. Resistors R1 and R2 are 5% composition, ½ watt. The switches are H. H. Smith 518 or equivalent. The connectors can be of any type to suit.

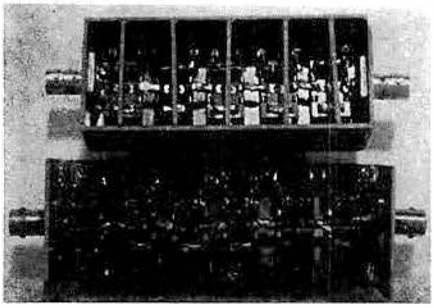

Details of attenuator construction. The top style, with complete shielding, is recommended.

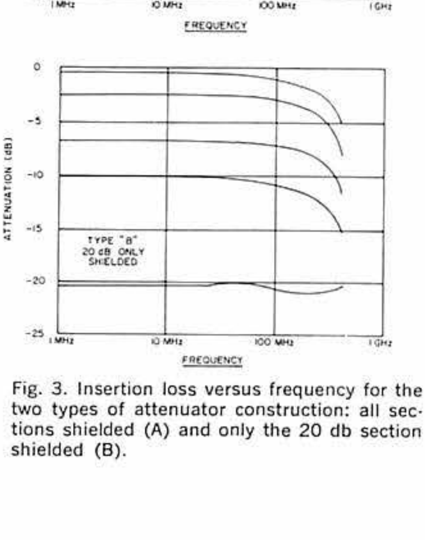

Fig. 3. Insertion loss versus frequency for the two types of attenuator construction: all sections shielded (A) and only the 20 db section shielded (B).

provided less insertion loss than the former at high frequencies, so this type of construction is recommended. Also note that dividing shields are desirable between input and output elements of a single section. These shields prevent capacitive feedthrough at the high frequencies, and are desirable on the high attenuation sections (anything over 10 dB) even at low frequencies. On the low attenuation sections, very little difference is evident below VHF. See **Fig. 3** for the attenuation of the attenuators up to over 450 MHz. Building attenuators with greater than 20 dB attenuation per section by this method is not recommended for high frequency use.

S-UNIT ATTENUATOR

Ed Lawrence WA5SWD

Since the topic of "S meters" is a popular one among radio amateurs, a lot of time is spent describing these devices, usually along the lines of how generous or "Scotch" the meters are at the QTH of the parties in the QSO. After a few such QSO's, I decided to build an attenuator, calibrated in "S" units. My aim was to attain an accuracy of 1 db or better, using 5% ½w resistors and simple construction so it would be easy to duplicate.

As a sidelight, I started out by calculating both "tee" and "pi" pads, and used "pi" because all values of resistance are close to standard values, but (especially for high attenuation pads) the values for "tee" pads can get quite small; and expensive.

I figured the values required from the tables in the Allied's "Electronics Data Handbook", page 8, 5th edition. (Allied Radio, 75c, full of good info.)

Since "S" units are supposed to be 6 db, I figured data for steps of 1,2,4 and 8 times that amount, or 6,12,24 and 48 db. With these steps, any number from 0 to 15 "S" units of attenuation could be selected. However, 8 "S" units proved to be too much for one step, as shown by the lowered attenuation at 30 MHz, due to the inherent shunt capacitance of the resistor used in the series leg, plus the stray capacitance of the switch. So I removed the 8 "S" unit step and installed another 4 "S" unit step. This allows selected steps of attentuation from 0 to 11 "S" units.

Here are the values I calculated, and the actual values used, based on 51 ohms. The steps are switched in series, as required for the desired attenuation.

Resistance Values for 51 Ohm Attenuator:

		R1		R2	
"S" units	DB	ideal	actual	ideal	actual
1	6	154	150	38	39
2	12	85	82	96	100
4	24	58	56	405	390
8	48	51.5	51	6400	6800

After the attenuator was completed, the attenuation was measured at 3 kHz and at 30 MHz. With the test equipment available it was possible to measure more accurately at 30 MHz than at 3 kHz. Below is the data from the tests.

Atten Step		Predicted	Measured	Measured
"S" units	DB	atten DB	at 3 kHz	at 30 MHz
1	6	6.2	6.0	6.02
2	12	12.3	12.2	12.16
4	24	23.3	24.2	24.05
8	48	48.5	47.5	39.11

Now if we want to make an educated guess as to how far up we can expect good results, say 1 db error out of 24 db, then we can use the measured error in the 48 db step to calculate the capacitance across the series leg, and from that calculate the frequency where the 1 db error will occur. Go through the math if that is how you get your kicks, or take my word for it. It comes out to about 2 pF. And this will cause a reduction of 1 db at about 220 MHz. And since the resistor is of a lower value for the smaller steps, they should hold their values to even higher frequencies, but I expect other factors would get into the act along the line somewhere. I will state that still works well at 2 meters.

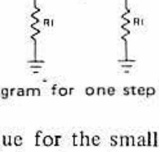

Fig. 1. Diagram for one step in the attenuator.

If you want to get fancy, you can always figure the values for 1, 2 and 3 db steps and have from 0 to 72 db attenuation in 1 db steps.

Referring to the photos; you can see I built my attenuator in a Bud Minibox CU-2102-A, 4" X 2-1/8" X 1-5/8". Four steps is the maximum in this size box, unless different switches are used. Mine are Cutler-Hammer 7592K6. The shielding was made from transformer strap, but could be any soft copper available. Try a Hobby Shop and get the thin sheet that is used for embossing if all else fails.

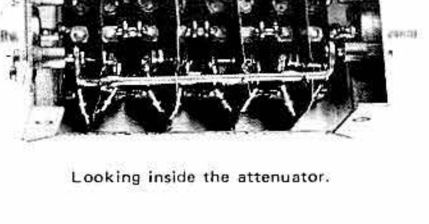

Looking inside the attenuator.

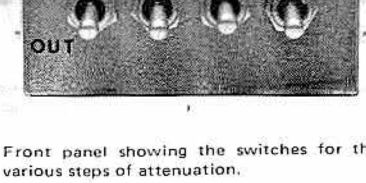

Front panel showing the switches for the various steps of attenuation.

Here are some of the uses an attenuator of this type is suited for:

Checking receiver "S" meter calibration.

Attenuating signals to aid in peaking receivers and converters.

Calibrating diode voltmeters for *rf* measurements.

Checking antenna gain. Or gain of that outboard *rf* stage.

P. S. My "S" meter lies, just as I thought!

THE EVER-USEFUL T-PAD

Jim Kyle K5JKX

Most of us have transmitters; let's hope that an equal number of us have receivers also. Antennas and microphones are usual station accessories, with a few determined diehards here and there clinging to the trusty old key. But how many of us have much in the way of test equipment?

Now and then somebody pops up with a VOM, and occasionally you can even find an operator who uses a scope. But the kind of special-purpose test gear you find in a well-equipped laboratory is almost always absent in the ham shack.

Which is more or less as it should be, since we're hams, not laboratory technicians. But with the present trends toward VHF, at least some specialized test gear is necessary. Otherwise, the regular station equipment can't be tuned for maximum performance.

One of the simplest such items is a noise generator, for getting the VHF receiver in perfect tune. This gadget has been described many times before, so we won't repeat it again —but we do have something which transforms the usual noise-generator lashup from a so-so item to an instrument capable of laboratory accuracy.

Before going into detail, let's look at the normal method of using a noise generator: you connect the generator to the antenna input, tie the converter to the receiver, turn the avc off, connect a voltmeter to the detector load resistor in the receiver (or put an ac voltmeter across the speaker leads), and measure the voltage produced by just noise. Then you turn on the generator and adjust it for a 3 db (1.4 times the voltage) increase in output; the object is to achieve the 3 db increase with the minimum amount of current flowing in the noise generator.

However, this technique of using the noise generator is pretty sloppy, since it assumes that the receiver's detector is absolutely linear for small signals—and this assumption is almost always incorrect.

A far better technique is to hook things up just as before, except now you place three T-pads in the line between converter and receiver. The T-pads on each end of the string serve merely to clamp the line impedance at 50 ohms, but the one in the center is built for precisely 3 db loss.

Now take the 3 db pad out and take your reading as before with the noise generator off; you don't have to turn off avc or hook up a voltmeter. The receiver's S-meter can be used instead, since we're not measuring anything with the meter itself. It merely serves as an indicator so we can come back to the same point.

Then replace the 3 db pad in the line, turn on the generator, and crank it up until you get the same meter reading as before. Since you now have 3 db of added loss between converter and receiver you must have increased the noise power output of the converter by that same 3 db, and you could care less about the linearity of the detector!

This whole method is far from new, but previous descriptions of it have left something to be desired in the way of telling how to build the T-pad. The gadget is so simple that it must have seemed obvious to previous writers—but it does have its tricky points too.

For instance, since a T-pad consists merely of 3 resistors, it is pretty easy to just wire them up by their leads. But they are hanging in the receiver antenna lead, and present-day receivers are rather sensitive. If you happen to find a 20-meter signal, it's going to foul up your measurements!

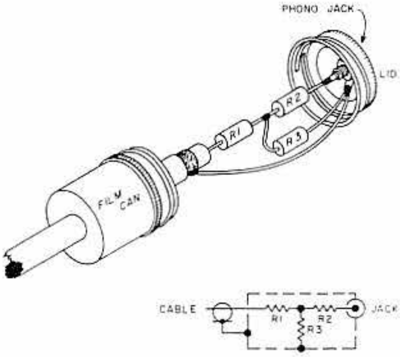

T-Pad construction details

One of the quickest ways to sidestep this problem is to shield the pad against all outside influences—but how do you shield anything so tiny?

The answer here is to use discarded 35-mm film cans, which all photographers who shoot 35-mm cameras have in abundance. The Kodak kind seem to work best. This type has a threaded cap, with a flat spot in its center just right for drilling a ¼ inch hole to take a single-hole-mounting phono jack. At the other end, a ⅜ inch hole can be punched and lined with a rubber grommet for coax cable to enter.

Next step is to place the jack in the hole in the lid, with its solder lug on the inside (be sure to clean the paint so that a good electrical contact will result). The T-pad can be assembled as shown in the drawing with shortest possible leads, and supported by its lead connecting to the jack. Thread a short length of RG-58 through the grommetted hole, and connect its center conductor to the other lead of the pad; the shield and shunt lead of the pad connect to the solder lug of the jack.

All that's left is to wire-brush the threads on the film can for good contact, and screw the lid down tight. Presto, a shielded T-pad. A phono connector should be put on the free end of the cable.

You can make up a whole bunch of these in various loss values, and get virtually any amount of loss you want by stringing them together. And this has a whole lot more use than just using with a noise generator . . .

For instance, when you want to test an antenna, arrange for a steady signal, strong enough to register on your S-meter with the back of the antenna pointed at it. Then swing the antenna in small steps, and bring the S-meter back to the same reading by inserting additional loss between converter and receiver with the T-pads. The difference between this and S-meter indications may amaze you.

Or if you are called upon to measure the difference in signal strength between two stations, the same approach can be used. Note the S-meter reading of the weaker, then knock the stronger one back to the same reading by putting T-pads in the converter-to-receiver line. Read the db off the pads and add them up.

You can even use this in place of an S-meter if you really want to know the signal-to-noise ratio of an incoming signal with accuracy; take a reading on noise, then knock the signal back to the same point. Total up the db, and there's your answer.

Though resistance values in the pad must be precise for absolute accuracy, the pad has an inherent tolerance of small errors and you should have better than 2 percent accuracy if you use 5-percent resistors. Specifically, a 10-percent error in the resistance value of any one arm produces less than ¾ db error in the pad loss, and less than 4 percent error in impedance. Using 5-percent resistors would, of course, cut these error limits in half.

If you want to follow the approach of using a string of these pads for all purposes in the shack, it's best to make them up on a "binary" approach since this gives you the maximum number of db values with the minimum number of parts. A basic assortment might be two 1 db pads, and one each of 2 db, 4 db, 8 db, 16 db, and 32 db. Using them in series in various combinations, you can get any whole number of db from 0 to 64, which pretty well covers the range of values you may ever need. For an example, to get 50 db you would use the 32 db pad, the 16, and the 2. For 60 db, you would use the 32, the 16, the 8, and the 4.

If the ability to increase loss in 1 db steps seems a bit exotic to you in view of the fact that 3 db is only half an S-unit, then you can use a binary progression in 3 db steps; this takes one 3 db pad, one 6 db, one 12 db, and one 24 db. The range is from 0 to 45 db, 3 db at a time.

For clamping a line's impedance, it's a good idea to use at least a 3 db pad and a 6 db unit might be even better. The lower-loss pads pads may not have the ability to swamp out impedance variations on their other sides.

About all that's left to make this complete is a chart of resistance values for various loss figures. Here it is; all are for use with a 50 ohm line; to use at any other impedance, multiply these values by the ratio of the new impedance to 50 ohms:

LOSS AMOUNT	R1, R2	R3
1 db	2.7	430
2 db	5.6	220
3 db	8.2	150
4 db	11	100
6 db	16	68
8 db	22	47
12 db	30	27
16 db	36	16
24 db	43	6.2
32 db	47	2.2

COMBINATION DUMMY LOAD/ ATTENUATOR NETWORK

John Schultz W2EEY

There are many instances when it is desired to use an existing transmitter as an exciter unit for a high-power linear amplifier. Many such linear amplifiers require a drive level that is only a fraction of the transmitter's output. To some degree, the transmitter can be detuned in order to reduce its output level, but this procedure is rarely possible when several orders of magnitude reduction in the power level are necessary. In such a case one can either internally modify the transmitter for a lower output level or use an alternator network between the transmitter and linear amplifier. In the latter case, the transmitter can be operated at its normal power input level and with its tuning controls at their normal settings.

The unit described in this article functions as both an *rf* attenuator and as a dummy load. The latter capability allows a transmitter to be properly tuned alone for correct operation before it is used to drive a linear amplifier. An optional wattmeter circuit is included which when calibrated allows direct reading, in watts, of the full transmitter output or of the drive level supplied to the linear amplifier.

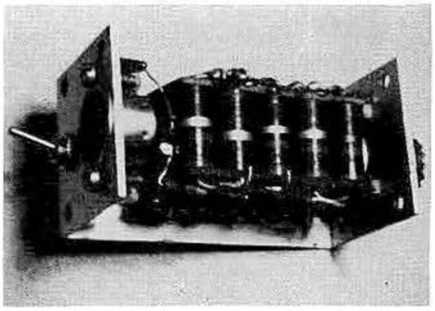

A simple method of construction is employed. Based mainly on "sandwiching" the resistors used between two pieces of vectorboard. Details are given in the text. SO-239 is used as coax input connector. The circuit function switch is located below the output connector.

The unit described was built for use with a nominal 100 output transmitter used primarily for SSB service. The construction used, however, can be extended to other power levels for transmitters operating on 80-10 meters. Also, using the information

Power Reduction	Resistor Factors a	Resistor Factors b
1/4 (6db)	.3	1.3
1/5 (7db)	.4	1.1
1/10 (10db)	.5	.7
1/20 (13db)	.6	.5

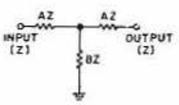

Fig. 1. Approximate resistor factors for "T" network attenuators over the ranges normally desired for exciter power output reduction.

supplied, the same type of attenuator/ dummy load can be designed for other than 50 ohm transmission line systems. The attenuator was not designed as a precision network in order to allow the use of inexpensive resistors. However, the attenuation characteristics are quite satisfactory for the intended usage.

Besides its application as a power reducer when driving a linear amplifier, the unit can be used with a transmitter whenever a quick, known level of power output reduction is needed for operating purposes, approximate gain measurements, etc.

Circuit

Fig. 1 shows the circuit values for a generalized T network attenuator that can be used in any impedance unbalanced transmission line. The scaling factors are only shown for those power reduction levels most likely to be needed when driving a linear amplifier with a 75-200 watt transmitter, in order to avoid unnecessary detail. Factors for intermediate power reduction values can be found by interpolation to a satisfactory degree or one can consult an electronics handbook. The basis of the attenuator/ dummy load network is to find the combination of resistor arms that will provide the desired attenuation and still be able to be connected together to form a dummy load of the correct value. Fig. 2 shows one possible combination. Each resistor bank has a value of about 20 ohms (5 resistors of 100 ohms each in parallel). In one position of the DPDT switch, the resistor banks are formed into a "T" network attenuator. In the other switch position, all three banks are placed in series as a dummy load connected across the input only. The resistance values which result are not exactly those shown in **Fig. 1** for a 10db alternator. However, they are close enough to be effective and some tailoring of the individual legs is possible since each of the resistor bank values vary by a few ohms due to the tolerance of the resistors used. An optional voltmeter circuit is also shown in **Fig. 2** connected to one pole of the DPDT switch. It can be used as a relative power output indicator or if calibrated, as described later, actually measure the power output of the transmitter and of the attenuator.

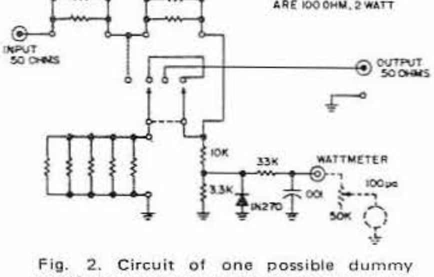

Fig. 2. Circuit of one possible dummy load/attenuator network providing about 10 db power reduction. Optional wattmeter circuit is also included.

Many variations of the basic idea are possible. **Fig. 3** shows the use of four banks of 100 ohm resistors. All four are used to form an attenuator that comes reasonably close to the values required for 7db attenuation in a 50 ohm system. Only three are used in series for the dummy load function. In this case only a simple SPST switch is necessary to disconnect the output. The same *rf* voltmeter circuit as used in **Fig. 2** may be added if desired. The switch, in fact, could be eliminated entirely if one were willing to disconnect the output termination in order to use the dummy load feature.

Whatever combination of resistance banks are used in order to achieve a desired attenuation value and the correct dummy load resistance, care must be taken that each resistance bank has sufficient power dissipation capability. The dissipation in each leg of the "T" network varies according to the attenuation level and can be calculated by Ohms Law. In general, a continuous power rating for a resistor bank equal to about one-third of the SSB peak power rating seems to suffice, including for quick tune-up on CW. For keyed CW service, the power rating should be increased to at least one-half the key-down power level.

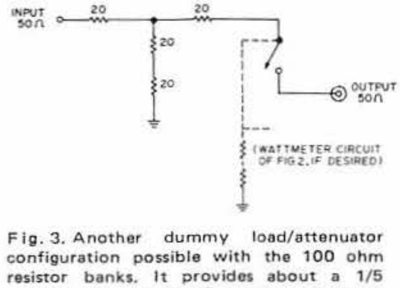

Fig. 3. Another dummy load/attenuator configuration possible with the 100 ohm resistor banks. It provides about a 1/5 power reduction (7db) when used as an attenuator.

Construction

The approach of using a relatively large number of 2 watt composition resistors is far less expensive than using specific value *rf* non-inductive resistors of 10-30 watts power rating. In quantities of more than 10, IRC type RC-2, 2 watt, 10% tolerance resistors cost about 9 cents each. So, one can achieve a 40 watt unit for less than $2 resistor cost. Banks composed of these resistors work well up to 30 mc as long as the interconnecting leads are kept short.

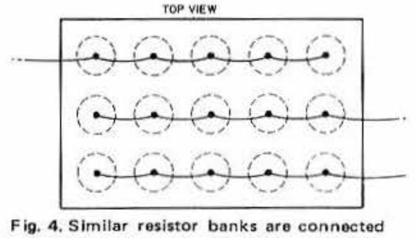

Fig. 4. Similar resistor banks are connected together on the underside of the assembly.

The photograph shows the construction used by the author for the circuit of **Fig. 2.** Similar construction can be used for larger size units as well. As shown in the photograph, the 15 resistors in rows of 5 each are sandwiched between two 1-1/8" x 1-7/8" pieces of vectorboard. None of the resistors physically touches. The wiring is done using the resistor leads. This construction is somewhat compact to expect full, continuous power dissipation from the unit but suffices for intermittent use. The frame measures 2¾" x 1½" x 1½". A cover is not absolutely necessary since the minor radiation that takes place is not important in this application. If a cover is used, it certainly should be of a perforated type to allow maximum air flow. A SD-239 connector is used at one end of the frame for the input. A dual connector is used at the other end, but normally one would use two RCA type phono jacks–one for the output and one for a meter circuit. The switch is located immediately below the output connector–a miniature Alco MST type.

Calibration

If it is desired to calibrate the voltmeter circuit as a wattmeter, it is necessary to use a probe and VTVM. Using the unit as a dummy load, the *rf* voltage is measured at the input and the power calculated. The 50K ohm potentiometer is used to set the meter at full scale for the highest power level used. The *rf* voltage is measured and the power level calculated in order to calibrate the meter for lesser power levels leaving the potentiometer at its "set" value. The same procedure is followed to calibrate the meter for the output power level by measuring the output *rf* voltage when the unit is used as a "T" attenuator and connected to a regular dummy load. The calibration should be made on the lowest frequency band used and rechecked on the highest frequency band used. If the readings differ significantly on the highest frequency band from those established, it may be necessary to add a few mmf capacitance across the diode in the voltmeter circuit in order to compensate for the slightest reactance present in the circuit.

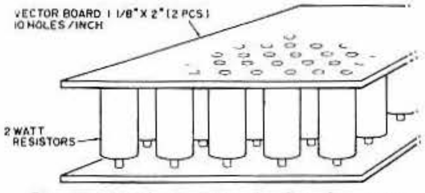

Fig. 5. Sketch showing details of resistor "sandwich" assembly.

Operation

When used between a transmitter and the 50 ohm input of a linear amplifier, the unit is first used as a dummy load for tune-up of the transmitter. The unit is then switched (with the transmitter unkeyed) to its attenuator position. In most cases, no returning of the transmitter should be necessary unless the input of the linear amplifier is particularly reactive.

Summary

The unit described is not intended as a precision attenuator or power measuring device. However, it will perform very well for its intended applications and costs far less than more sophisticated units performing the same functions.

THE MINICAN

Sam Kelly W6JTT

The Heathkit "Cantenna" has proven to be a major breakthrough in the field of dummy loads for ham use. Unfortunately, it isn't the most convenient thing to use on a small work bench with low power rigs! Borrowing their idea, I built this load for use with transmitters in the 5 to 15 watt range. The parts are few: a Campbell soup tin can, four one watt resistors, a UG-254-A connector, a short piece of 5/16 inch brass tubing and transformer oil.

Fig. 1 is a sketch of the assembled unit. The 50 ohm resistance was made up of three 15 ohm and one 5 ohm one watt carbon resistors.

First sand a can lid from a larger size can until it is free of paint. Drill a ½ inch hole through the center of the lid, and a ¼ inch hole on the perimiter. Mount the coaxial connector through the center hole. Solder a 1 inch length of 5/16 in. brass tubing over the ¼ inch hole. Solder the resistors as shown. Center the lid on the can and solder the lid to the can. Use a file to remove all rough edges. Mask the connector with masking tape and paint the can to prevent rusting.

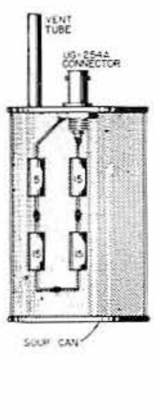

Fig. 1. Construction of the Minicon. Main parts are a soup can, coax connector and resistors.

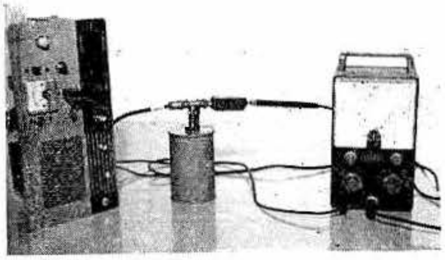

The Minicon in use with the companion detector.

Fill the can with transformer oil. A good grade of automatic transmission fluid was used in one load with no degredation in performance. However, it probably is not advisable as the fluid has a relatively low ignition temperature and might create a fire hazard.

The load was tested by running it for five hours with an input of 15 watts of 50 MHz rf. The can became warm, but the resistors showed no signs of overheating.

A maximum VSWR of 1.5:1 was obtained at 234 MHz. The measurement was made at this frequency because an automatic test set was available.

A companion rf detector unit shown in Fig. 2 was built in a two inch section of ¾ inch square extruded brass stock. A Dage 394-1 BNC connector is mounted on one end for connecting to the RF circuit, while the DC output to the VTVM is a tip jack.

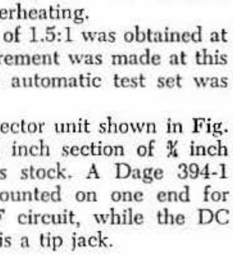

Fig. 2. Rf detector for use with the Minicon.

THE TINY TERMINATOR

Jim Kyle K5JKX

HAVE you ever felt the need for a dummy load which would also indicate accurately the actual rf output power of a transmitter?

You can buy such a beastie, you know. The Bird *Termaline*, standard of the two-way communication industry, shouldn't set you back much more than a C-note . . .

But you can also build one at considerable less outlay, and that's what this article is all about.

As the photographs show, there's almost nothing to the "Tiny Terminator"—nothing, that is, but a 50-ohm load for any transmitter, which will absorb 20 watts for days on end and will handle 40 watts for brief periods (a minor modification can double these ratings), will *not* radiate rf into the ether after the fashion of the "standard" light bulb load, will allow accurate measurement of output power up past 225 mc, and can be built in less than two hours for less than $5.

You can see by the schematic that there's nothing to the device electrically; the tricks are all mechanical. Electrically, the Tiny Terminator consists of 20 (that's what we said) 1000-ohm 1-watt resistors connected in parallel, with a crystal diode running from the hot end of the composite resistor to the meter jack and a 27 mmfd capacitor bypassing the meter.

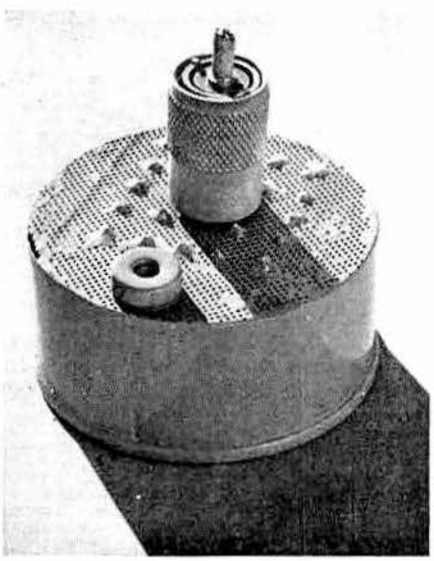

Mechanically, the big problem with a unit such as this is the problem of keeping the load resistive regardless of frequency. Resistor leads have inductance (about 25 millimicrohenries per inch) so we get rid of the leads as completely as possible. This is done by punching holes in a flat plate, passing the lead through the hole until the resistor is flush with the plate, and soldering rapidly with a hot iron (an Ungar 47-watt 1100-degree tip was used in building the model shown) so that the resistor body won't cook before the leads and plate are joined.

The flat plate has capacitance. This is avoided by keeping the plate shielded from outside influences, and separated as far as possible from its shield.

The shield, incidentally, is the bottom of 1½ inches of a Canada Dry cola, soldered all around the edge to the front resistor plate. This confines all rf inside the terminator.

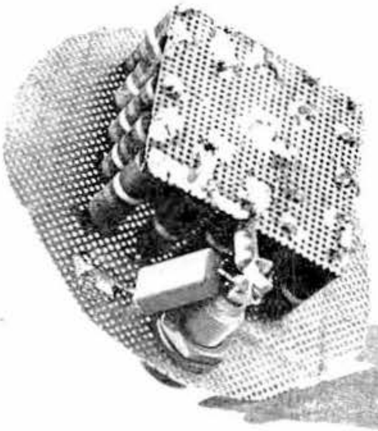

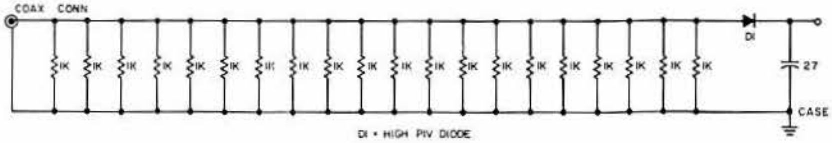

Addition of the crystal diode (a 1N34 was used, but a higher-voltage unit is recommended if you ever expect to measure power higher than 8 watts) and the bypass capacitor provide the power-measuring feature. These two components, in conjunction with an external voltmeter of at least 1000 ohms per volt sensitivity, provide you a peak-reading ac voltmeter. The voltage indicated on the meter will be equal to the peak value of rf voltage present across the load. Squaring the voltage and dividing by 50 (the resistance) will give you the *peak* power. Most power ratings are in rms values rather than peak; multiplying the indicated voltage by 0.707 before squaring will give you the rms power output.

The unit shown in the photos used perforated brass stock for the resistor plates; this happened only because a length of the brass was on hand in the junkbox at the time. A cut-out tin can lid will work equally well and will be much less expensive.

Not visible in the photos is the means of connecting the hot end of the resistor plate. A hole was punched in the middle of the hot plate and a length of No. 14 bare wire was soldered in, then filed off flush on the rear side. The bare wire was threaded through the coax connector and soldered to the center contact after checking for possible shorts.

The coax connector itself is held in place by a solder joint all around the rim of the cable-end aperture (see photo). This joint, if well made, provides plenty of strength.

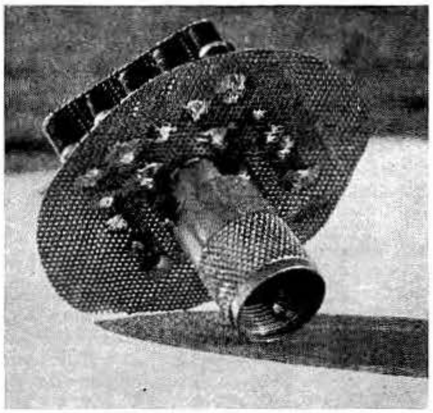

Earlier, we mentioned that a minor modification would double the power rating. That modification is this: instead of 1-watt resistors, use the 2-watt variety. If you really want to go high-power, use 50 2700-ohm 2-watt resistors and have a terminator which will absorb 100 watts continuously and 200 watts ICAS. However, one that big will cost more—and you'll probably have to use a coffee can for the shield instead of the cola can used here.

Volume III
Test Equipment

INDEX

Absorption Wave Meter 39-40
Antenna Impedance Bridge 21-22
Antenna Measurements 28
Antennascope 21
Attenuator, RF 58, 101-6
Audio Frequency Oscillator 50

Capacitor Checking 28, 93
Citizen Band Alignment Aid 55
Crystal Calibrators 73-78

Dummy Loads 38, 101-6

Field Strength 39-46
Frequency Generator, Sweep 52-53
Frequency Measurement 28, 39-48

Grid Dip Oscillators 27, 79-94

Hot Carrier Diode RF Meter 36

Impedance Meter 27
Intermediate Frequency
  Alignment 28
  Spotter 51
  Sweep Generator 62

Lambda Lines 46
Light Bulb as Power Indicators 30
LM-15 Frequency Meter 49

Microwave Frequency Measurement 42-6

Neutralization 28
Noise Figure Measurement 95-100
Noise Generators 95-100

Oscillator Detector 28

Parasitic Oscillation Checking 28
Pulse Generator 69

"Q" 27-8

Radio Frequency
  Attenuation 58-9
  Impedance Bridges 21-9
  Power Meters 5, 16, 30-8
  Pulse Generator 69
  Signal Generators 47-72
  Sniffer 39
  Sweep Generator 62
Ramp Generator 67
Receiver Tweaker 55

Signal Generators 58-9
Slotted Line 14
Standing Wave Ratio 5-20
S-Unit Attenuator 102
Sweep Frequency Generator 52, 62

T-Pad 102
Transmission Line Attenuator 7
TVI Testing 28

UHF
  Dipmeters 85-93
  Frequency Measurement 42-6
  Noise Generator 97
  Signal Generator 57, 61
  SWR Bridge 13

VHF
  Dipmeters 90-3
  Dummy Load 38
  Frequency Measurement 46, 48
  SWR Meter 12-6
  Wattmeter 38
VSWR 5-20
Vacuum Tube Voltmeter 27

NOW FROM TEXAS INSTRUMENTS...
three machines in one.

programmable
slide-rule calculator
SR-52

TEXAS INSTRUMENTS
SR-52
LUNAR LANDING GAME
ON

- 10 user defined keys
- 224 program storage locations
- 23 preprogrammed key functions
- 8 preprogrammed condition statements
- 20 independent addressable memory registers
- Permanent program storage on magnetic cards

It took TEXAS INSTRUMENTS to invent the SR-52 calculator. It took C & S MARKETING ASSOCIATES to offer it at a price you can afford, now only $249.95. With such versatility and such an affordable price, you cannot afford to be without the problem solving power of card programability. Now solve problems in seconds that would take hours with any ordinary calculator or slideruler if they could be done at all.

For more information or the answer to any question you may have about the SR-52 calculator, call toll free (800-251-6771)*. Tenn. residents call (800-262-6706). Other TEXAS INSTRUMENT models available from $49.95.

Each TEXAS INSTRUMENT calculator comes with a 1-year warranty. Should your unit prove defective within 60 days, just return it for a new unit! Finally should you be dissatisfied with your calculator return it within 15 days for a prompt refund.

C&S MARKETING ASSOC.
P.O. Box 165 Algood, Tenn. 38501

QTY.________________PRICE $249.95 ea.

☐ CHECK ☐ M.O. ☐ C.O.D.

☐ LITERATURE ☐ INFORMATION

DIAL (800-251-6771)

IN TENN. DIAL (800-262-6706)

"New England's
FRIENDLIEST HAM STORE"

complete stock

of the following **lines:**

- ATLAS
- VHF ENGINEERING
- STANDARD
- REGENCY
- BOMAR CRYSTALS FOR 2MTR FM
- 73 MAGAZINE PUBLICATIONS
- QST PUBLICATIONS
- NEWTRONICS
- TEN-TEC

Largest inventory of used equipment in the Boston area.

TUFTS Radio Electronics
386 Main St., Medford MA 02155
Phone: 617-395-8280

South Com

YOUR BEST THEFT PROTECTION!

FOILER

PATENT APPLIED FOR TM

Quick Release Adapter for Trunk Lip Antennas

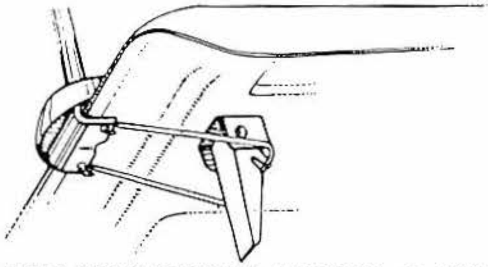

USES YOUR EXISTING ANTENNA & MOUNT

MOUNTS EASILY INSIDE TRUNK WITH ONLY 2 SCREWS

DON'T ADVERTISE YOUR RADIO INSTALLATION TO THE RADIO THIEF

$4.95

IF YOUR DEALER DOESN'T HAVE THE FOILER IN STOCK, ORDER DIRECT. ADD 50¢ POSTAGE & HANDLING. TEXAS RESIDENTS ADD 5% SALES TAX. SEND CASHIER'S CHECK OR MONEY ORDER FOR PROMPT SHIPMENT.

– DEALER INQUIRIES INVITED –

SouthCom, Inc.
P.O. BOX 11212 FORT WORTH, TEXAS 76109

South Com

Tired of "CB-Quality" Cables?

INEXPENSIVE ASSURANCE OF THE MOST CRITICAL LINK IN YOUR 2-WAY SYSTEM!

Regardless of how much you spend on your radio and antenna system, they can't do their best job unless all of your output power gets from the radio to the antenna, and all of the weak signals get from antenna back to radio. The road both power and signals use for this trip is your coaxial cable. Obviously, then, if the cable is broken or contains losses, you won't get the performance you paid for, yet, very few radio operators pay any attention to the quality of the cable they buy, nor do they replace it periodically. Even the best of cable weathers, ages, and develops losses. The coax cable is the least expensive part of your system. The difference in the best and the cheapest you can buy is only a couple of dollars at worst and yet it can literally "make or break" your performance.

Most cable available today in ready-made lengths, and most cable furnished with new antennas, has three less-than-desirable characteristics.

First, most such cables have a single center wire, which breaks with bending, twisting, and day-to-day handling. SouthCom's cables have a 19-strand center wire, so all 19 strands must break before the cable loses connection.

Second, most cable does not have full shielding. You can peel back a small section of the outer black insulation, and you can see the inner insulation through "holes" in the braid. SouthCom's cable is fully shielded!

Third, most cable uses bare copper conductors, and you know how quickly bare copper can corrode. SouthCom's cables use tinned copper to resist corrosion.

Connectors vary widely. Solder-on types are the most common, yet they require a lot of heat which can partially melt the inner insulation, and distort the shape of the cable. In addition, you can't tell how much of the braid has actually been soldered through the four tiny holes in the fitting. SouthCom uses AMP, Inc. air-crimped connectors, which require no heat, and assure a solid bond to all of the braid. Because of the inner construction of the connector, it is almost impossible for a SouthCom cable to short.

In seven months of production, we have had only one report of a defective cable! Quite a record. Part of this record can be attributed to our quality control which checks every cable for SWR and power loss, not just continuity. We think the quality we build in, and the care with which we put our cables together makes them worth a little bit more. Since you're surely entitled to the performance you paid for when you bought your radio and antenna, you should insist on SouthCom cables. If your dealer won't get them for you, we will ship to you from the factory. Add $1.00 postage and handling. Texas residents add also 5% sales tax. Send cashier's check or money order for immediate shipment.

58-3	3-ft. cable/PL-259's	$3.98
58-6	6-ft. cable/PL-259's	4.85
58-12	12-ft. cable/PL-259's	5.65
58-20	20-ft. cable/PL-259's	6.50
58-20L	20-ft. cable, PL-259-solder lugs	5.95
58-50	50-ft. cable/PL-259's	9.95

REPLACE WITH SOUTHCOM CABLES TODAY...
GET THE PERFORMANCE YOU PAID FOR!

SouthCom, Inc.
P.O. BOX 11212 FORT WORTH, TEXAS 76109

S. D. SALES CO.

YOUR SOURCE

for

CHIPs, CLOCKs, etc.

Send for Flyer

S. D. SALES CO.

P. O. BOX 28810
DALLAS, TEXAS 75228

Exclusively Yours

Call your BANK AMERICARD or MASTER CHARGE
order in on our
CONTINENTAL UNITED STATES TOLL FREE WATTS:

Your
BANKAMERICARD
welcome here

1-800-527-3460

TEXAS RESIDENTS CALL COLLECT:

WE HONOR
master charge
THE INTERBANK CARD

214/271-0022

1 MORE TIME

Please call between 8:30 AM and 6:00 PM C.S.T. – Monday through Friday. You may also call to check stock or just ask a question. However, only B.A.C. and M.C. orders will be accepted. We do not ship C.O.D.

AST/SERVO SYSTEMS INC.

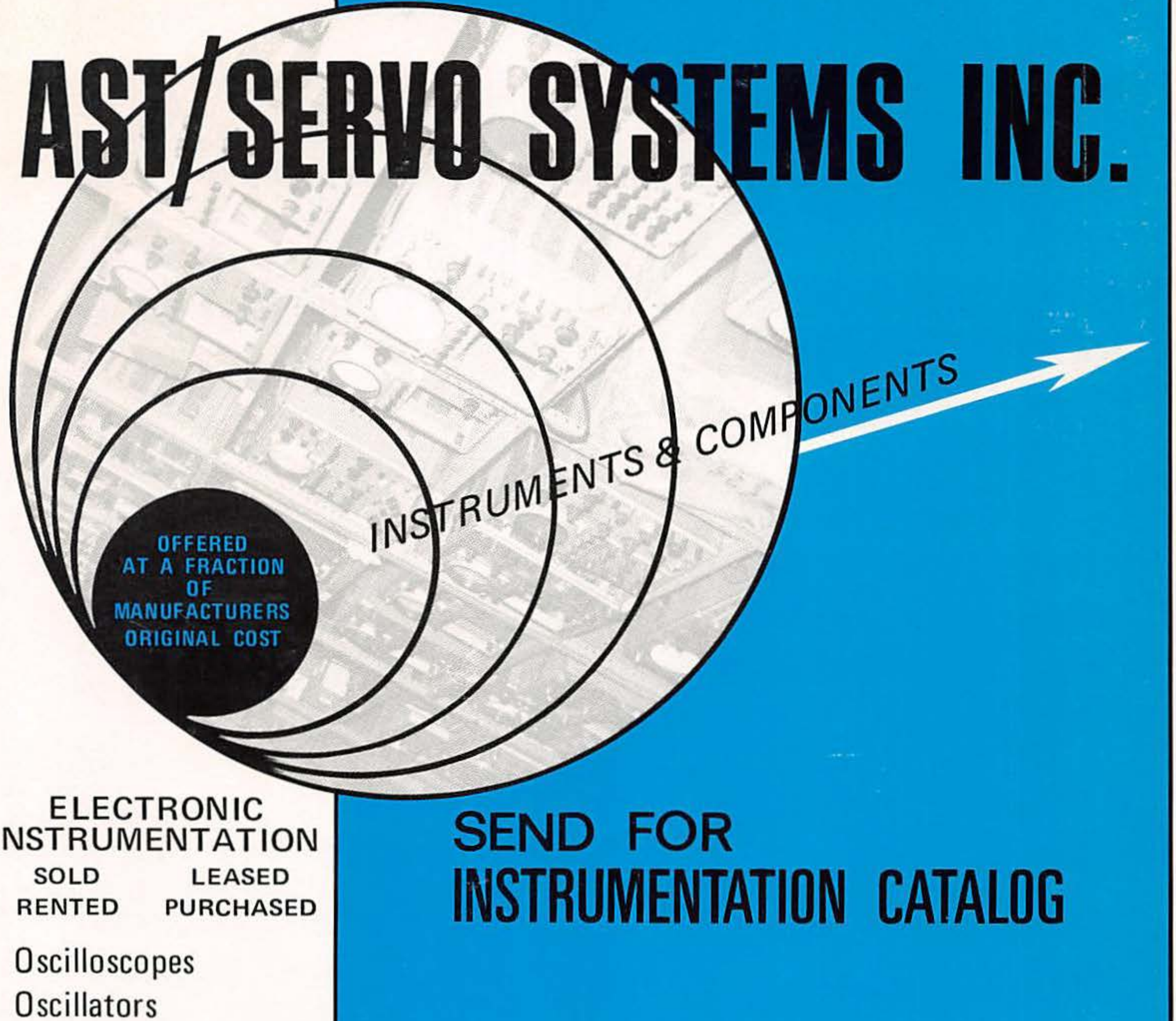
INSTRUMENTS & COMPONENTS
OFFERED AT A FRACTION OF MANUFACTURERS ORIGINAL COST

ELECTRONIC INSTRUMENTATION

SOLD LEASED
RENTED PURCHASED

- Oscilloscopes
- Oscillators
- DVM
- R.F. Volt Meters
- Power Supplies
- Transducers
- Recorders
- Synchros
- Servo Motors
- Generators
- Related Components & Systems

SEND FOR INSTRUMENTATION CATALOG

Name____________________

Address__________________

City______________________

State________ **Zip**________

AST/SERVO SYSTEMS INC., 20 Republic Road, North Billerica, Mass. 01862 617-667-8541

ARRL's

$19.95

RF Amplifier Classics

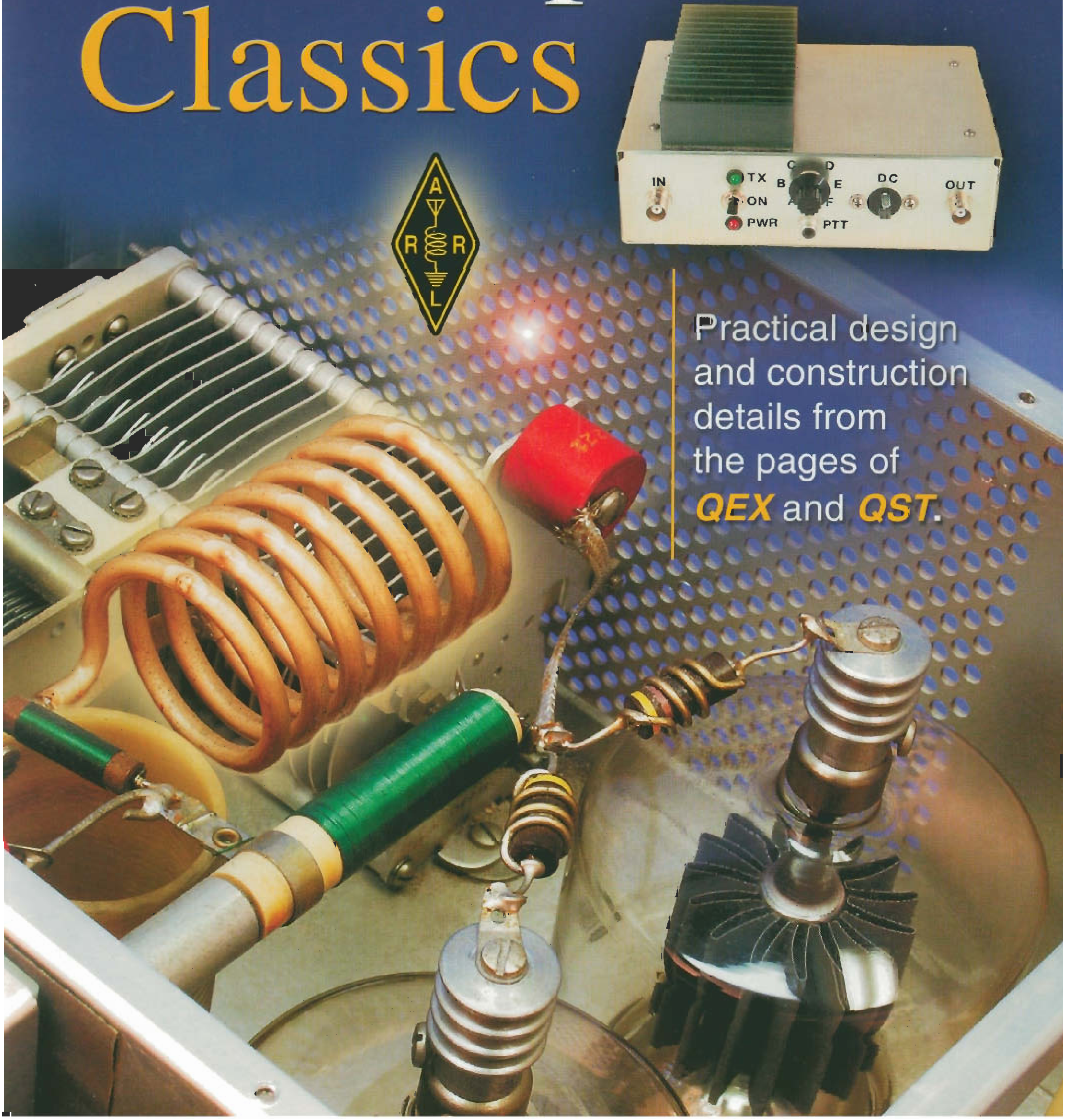
IN
TX
ON
PWR
C
D
B
E
A
F
PTT
DC
OUT

Practical design
and construction
details from
the pages of
QEX and ***QST***.

Contents

HF/MF/VHF Amplifiers (1 to 54 MHz)

An Easy-to-Build 25-Watt MF/HF Amplifier	Gary Breed, K9AY
A Compact 1-kW 2-50 MHz Solid-State Linear Amplifier	H.O. Granberg, K7ES/OH2ZE
A Broadband HF Amplifier Using Low-Cost Power MOSFETs–Parts 1 and 2	Mike Kossor, WA2EBY
A 1.8 to 54 MHz 5-Watt Amplifier	Zack Lau, KH6CP
An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz	Joel Paladino, N6AMG
An All-Band, 1500-Watt-Output 8877 Linear Amplifier–Parts 1 and 2	Jerry Pittenger, K8RA
High-Efficiency Class-E Power Amplifiers–Parts 1 and 2	David Rutledge, KN6EK, et al.
A 100-W MOSFET HF Amplifier	William E. Sabin, W0IYH
The FARA HF Project	Jim Valdes, WA1GPO

VHF/UHF Amplifiers

A Compact "Brick" for 6 Meters	Pat Bunn, N4LTA
A 300-W MOSFET Linear Amplifier for 50 MHz	Richard Frey, K4XU
A No-Bandswitch, Dual-Band VHF Desktop Amplifier	Paul Hewitt, WD7S
An 8-Watt, 2-Meter "Brickette"	Bob Larkin, W7PUA
903-MHz Linear Amplifiers–Parts 1 and 2	Dave Mascaro, WA3JUF
25-Watt Linear Amplifiers for 144 and 220 MHz	Dave Mascaro, WA3JUF
A High-Power 2-Meter Amplifier Using the New 3CX800A7	David Meacham, W6EMD
A Quick Powerhouse	Russ Miller, N7ART
A Cathode-Driven Tetrode for 6 Meters	David Munyon, W7DVB
A UHF Amplifier—from Scratch	John Reed, W6IOJ
A Solid-State 6-Meter Linear Amplifier You Can Build	Tarmo Tammaru, WB2TMD
Build a 6-Meter "Mini-Lini"	Wilson Hoag, WA5OLT
A Grounded-Grid Kilowatt Amplifier for 432 MHz	Stephen J. Powlishen, K1FO

Microwave Amplifiers

A Quarter-Kilowatt 23-cm Amplifier–Parts 1 and 2	E.R. "Chip" Angle, N6CA
A 2-W 13-cm Amplifier	Zack Lau, KH6CP
1296-MHz Solid-State Power Amplifiers	Al Ward, WB5LUA

Amplifier Maintenance

Amplifier Care and Maintenance	Ward Silver, N0AX

Foreword

If antenna gain alone is not sufficient to bridge the path between two stations, the alternative is to increase RF power. Hams have been aware of this fact since the earliest days, and that is why RF power amplifier projects have always been popular.

In *RF Amplifier Classics* we have assembled a collection of articles published in *QST* magazine, and its sister technical journal, *QEX*. The collection spans the early 1980s through 2003 and includes many prominent authors. In those few instances where an author provided a design revision (or correction) after the article was published, that revision is *included* in the article as presented in this book.

See the leatest issue of *QST* for other ARRL RF design-related publications, or visit our on-line bookstore at **www.arrl.org/catalog**. Please take a few minutes to give us your comments and suggestions on this book. There's a handy Feedback Form for this purpose at the back, or you can send e-mail to **pubsfdbk@arrl.org**.

Our thanks to the many authors whose work appears in this book. Without their willingness to share their knowledge with the amateur community, *RF Amplifier Classics* would not be possible.

Dave Sumner, K1ZZ
Executive Vice President
Newington, Connecticut
August 2004

By Gary Breed, K9AY **From *QST*, February 1994**

An Easy-to-Build 25-Watt MF/HF Amplifier

Do you need a medium-power linear amplifier for SSB or CW? Congratulations—you just found it!

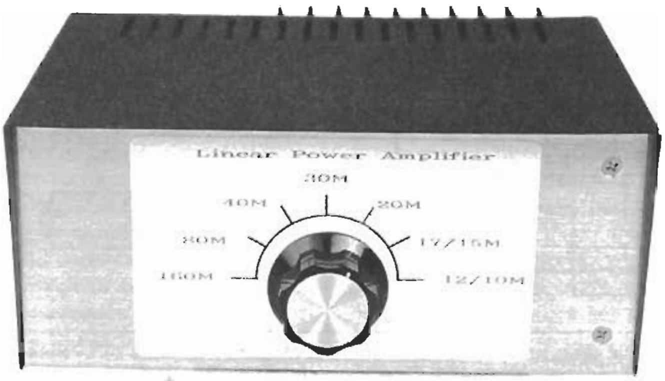

Here's a 25-W, 1.8- through 30-MHz, class-A linear power amplifier that's simplicity itself. What makes it simple is the use of a self-biased transistor module requiring few external components. To control harmonic output, a set of five-section low-pass filters is included. Power-supply requirements are +28 V at 2.5 A and —5 V at 200 mA.[1] With a gain of about 13 dB, a 1- to 1.4-W driving signal is all that's needed to deliver 25 W output. Gain is flat within ±0.75 dB across the covered frequency range.

If 25W isn't enough for you, it's easy to directly apply the design information to build a 50-W amplifier—all you do is use a larger transistor module! Another step toward project simplicity is the availability of kits. Each kit contains all the major components for either a 25- or 50-W version.[2]

Amplifier Design

When designing a power amplifier, the first step is to select the right transistor(s). Excellent bipolar-junction transistors (BJTs) and field-effect transistors (FETs) are available from well-known companies such as Motorola, M/A-COM PHI, SGS-Thomson, Philips, Mitsubishi and others. A number of smaller companies also make power transistors, usually for more-specialized applications. MicroWave Technology, Polyfet RF Devices, and Directed Energy may be company names unfamiliar to you, but they all make power transistors for MF and HF applications.

In this amplifier, I use the SLAM-0111 from MicroWave Technology.[3] I didn't choose it because of its gain, its efficiency, or even its price; I selected it because it's very easy to use. The device consists of two power JFETs (the particular specialty of MicroWave Technology), operating in push-pull. Since JFETs behave similarly to triode vacuum tubes, the company dubbed them *Solid State Triodes*. SLAM (Solid-state-triode Linear Amplifier Module) devices include thick-film bias resistors in the package with the transistors. These resistors set the gate bias for class-A operation, and establish a 50-a input impedance. At the rated power and supply voltage, the push-pull output impedance is also 50 Ω!

With such convenient input and output impedances, matching the devices to a 50-Ω system merely requires 1:1 balun transformers at the input and output. Because the bias voltage is internally generated, the only other external circuitry required is a suitably bypassed and isolated 28-V power supply.

Circuit Description

The amplifier schematic is shown in Fig 1. The balun driving the gates of the push-pull transistors is a conventional transformer. The primary and secondary windings are each three turns of #28 wire, wound on a two-hole ferrite balun core of 73 material ($\mu_i = 2500$). These transformers are broadband enough to provide 1.8- to 30-MHz operation and offer dc isolation with no additional components. The input-transformer primary is center-tapped and bypassed to provide access to the gates for external dc bias (more on this later).

The output transformer is constructed in the same manner as the input transformer—it's just larger. Two ferrite beads of 77 material ($\mu_i = 2000$) make a two-hole core, with primary and secondary windings of three turns each, using #24 hookup wire. The primary (transistor side) is center-tapped and bypassed to provide dc voltage to the drains. Feeding dc through a center-tapped transformer eliminates the need for the usual bifilar RF choke seen in push-pull amplifiers—another reduction in the component count. Multiple bypass capacitor values (0.01, 0.1 and 10 µF) are used to cover the MF/HF range. That's the basic amplifier block: two transformers, a SLAM device, and a few bypass capacitors!

Class-A Operation Notes

By definition, transistors operating in class A conduct over the entire 360 degrees of the signal (that's all the time, of course). This operational mode assures that the transistor is always operating in the linear region of its input-to-output transfer characteristic. To do this, the device must be biased to handle the maximum signal at all times.

Obviously, this class of operation is pretty inefficient, since full current is drawn whenever the amplifier is on. A "perfect" transistor operating class A can only be 50-percent efficient, and real transistors

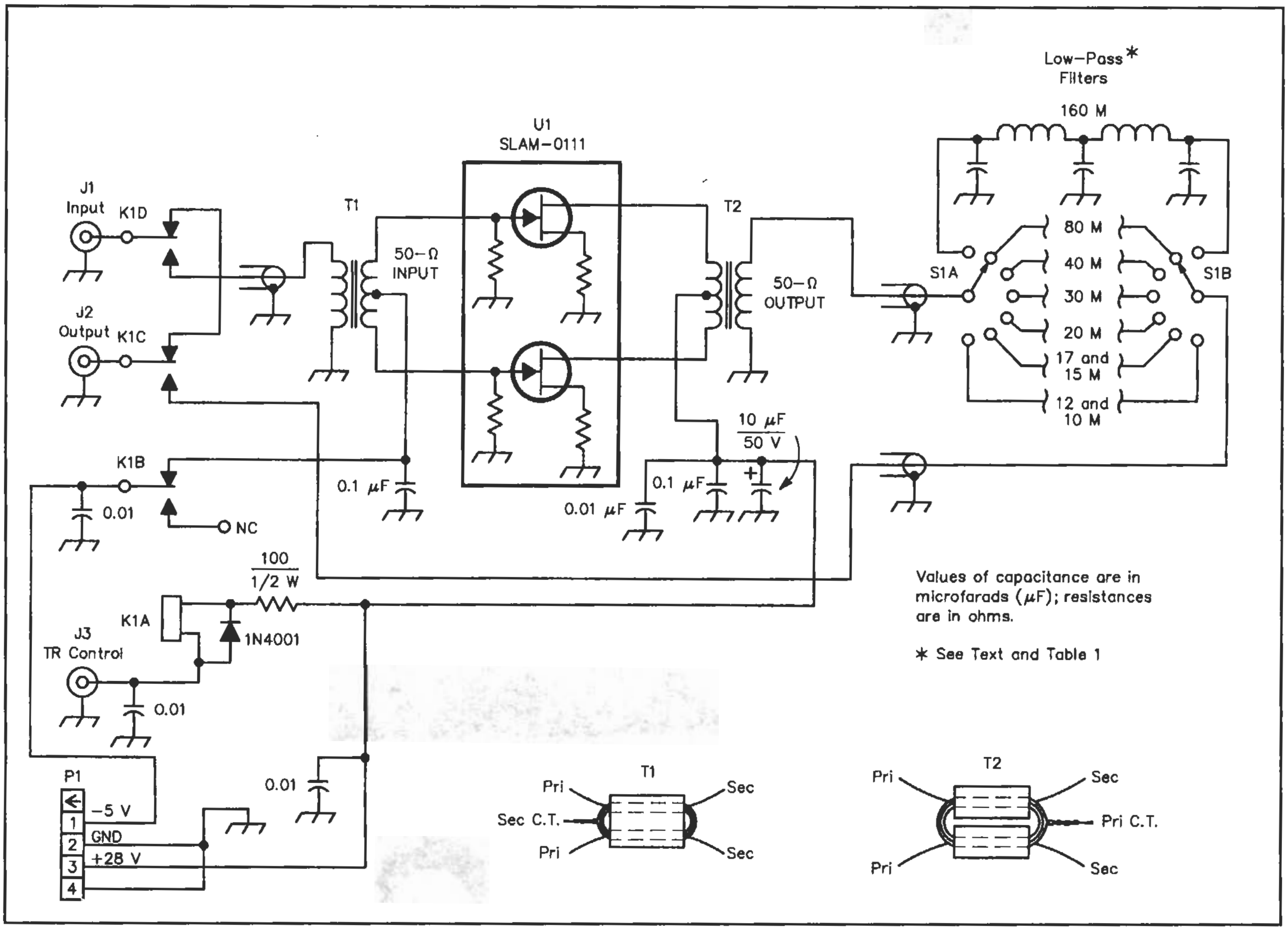

Fig 1—Schematic diagram of the 25-W class-A amplifier. Unless otherwise specified, resistors are $\frac{1}{4}$-W, 5%-tolerance carbon-composition or film units. Equivalent parts can be substituted.

J1—Panel-mount BNC socket.
J2—SO-239 connector.
J3—Phono jack.
P1—4-pin male Jones plug.
K1—3PDT relay with a 24-V dc coil. A surplus Potter & Brumfield KHP series 4-pole relay is shown in Fig 6; one pole is unused. (All Electronics catalog number 4PRLY-24N [$4] or Ocean State Electronics R12-17D3-24 [$10.90] are suitable. See the Part Suppliers List on pp 35-40 of *The 1994 ARRL Handbook* for addresses and telephone numbers. —Ed.)
S1—2-pole, 7-position ceramic rotary switch. My switch is made from two surplus CRL 11-position switch wafers and an indexing assembly providing selectable stops. The wafers are spaced about $1\frac{1}{2}$ inches apart. CAL PA-200 series switch wafers and PA-300 series shaft and indexing assemblies are suitable (switches are available from Newark Electronics; tel 312-784-5100, fax 312-784-5100, ext 3107, to locate your nearest Newark distributor).
T1—Primary: 3 turns #28 AWG; secondary, 3 turns #28 AWG, center-tapped. Core: Fair-Rite #2873002402 balun (Amidon BN 73-2402).
T2—Primary, 3 turns #24, center-tapped; secondary, 3 turns #24. Core: two Fair-Rite #2677006301 beads (Amidon FB 77-6301).
U1—SLAM-0111 ultralinear 25-W, class-A, self-biased power FET module or SLAM-0122, 50-W version (Microwave Technology, 4268 Solar Way, Fremont, CA 94538, tel 510-651-6700, fax 510-651-2208).
Misc: RG-174 coax, enclosure ($3\frac{1}{2}$ x $\frac{7}{8}$ x $5\frac{1}{4}$ inches [HWD]), heat sink (3 x $4\frac{1}{8}$ x $1\frac{3}{8}$ inches [HWD]), PC-board material, knob, mounting hardware.

do no better than about 40 percent. This amplifier draws 2.5 A from a 28-V power supply for an input power of 70 W. When it is providing 25 W, it's 36-percent efficient. (When there is no input, it's 0-percent efficient!)

To help reduce the heat generated by an amplifier that requires 70 W, a negative bias can be applied to the gates when not transmitting. A bias of —5 V results in a 0.25-A standby drain current instead of the full 2.5 A. The internal bias resistors are about 50 ~ on each gate, and dissipate a maximum of I W. Under these biasing conditions, the resistors each dissipate 0.5 W. Don't try to cut off the transistors completely with greater bias voltage! You'll risk burning out the resistors.

Some may ask, "If class A is this power hungry, why use it?" In a word: linearity. If you want excellent linearity (which means minimum distortion caused by harmonics or intermodulation), class A is the way to go. For example, all small-signal amplifiers for receivers and low-level transmitter stages operate class A because they must handle signals without distortion. However, they operate at very low power, so power dissipation is rarely an issue. This power amplifier further minimizes distortion by using push-pull operation, which cancels even-order distortion products in the output and makes the next part of the design easier than usual.

Harmonic Filter Design

As mentioned previously, the amplifier uses several low-pass filters to cover the nine MF/HF amateur bands. Each filter was initially designed for a cutoff frequency 20 percent higher than the upper end of their respective 160, 80, 40, 30, 20, 15 and 10-meter ham bands. The 15-meter filter is also used for 17 meters, and the 10-meter filter for 12 meters.

With no filtering, even-order harmonics (2nd, 4th, etc) are more than 40 dB below the carrier, the result of good push-

Table 1

Filter Circuit and Comparison of Ideal and Final Component Values.

	Ideal Filter Values			*Actual Filter Values*		
Cutoff Freq. *(MHz)*	*C1, C5* *(pF)*	*C3* *(pF)*	*L2, L4* *(µH)*	*C1, C5* *(pF)*	*C3* *(pF)*	*L2, L4* *(µH)*
2.40	1521	2620	4.55	1470 (1000 + 470)	2880 (2200 + 680)	4.41 (30 t on T50-2)
4.80	761	1310	2.27	830 (560 + 270)	1430 (1000 + 430)	2.37 (22 t on T-50-2)
8.76	417	718	1.25	430	820	1.25 (16 t on T-50-2)
12.18	300	516	0.900	300	560	0.960 (14 t on T-50-2)
17.22	212	365	0.634	220	370 (270 + 100)	0.706 (12 t on T-50-6)
25.74	142	244	0.424	150	240	0.460 (10 t on T-50-6)
35.64	102	176	0.306	100	180	0.314 (8 t on T-50-6)

In some cases it is necessary to parallel two smaller-value capacitors to obtain the proper values of capacitance for C1, C2 and C5

The inductors are wound on T-50-2 or T-50-6 cores. Inductors for the 160- and 80-meter filters are wound with #26 AWG wire in order to fit all turns on the cores; the other inductors are wound with #22 wire.

pull balance using factory-matched transistors. The 3rd and 5th harmonics are more than 15 dB down. To reduce the 3rd harmonic to at least 50 dB below the carrier, a five-section Chebyshev filter with low passband ripple is an appropriate choice. This type of filter has a good SWR in the passband, and a smooth roll-off characteristic. The design process began by creating ideal designs using a public-domain filter design program.[4]

Ideal designs rarely correspond to standard capacitor or inductance values that can be realized with a discrete number of turns on common toroid cores. Using a circuit analysis program,[5] the ideal designs were analyzed to see the effects of such realworld limitations on harmonic rejection and SWR performance.

First, the ideal component values were entered into the program, and varied ±20 percent to see which ones had the greatest effect on performance. Cl and CS (see Fig 2 and Table 1) were found to be least sensitive to variations, L2 and L4 were moderately sensitive; varying C3 had the greatest effect on both passband and stopband performance. The ideal capacitor values were then replaced with standard capacitor values or—in some cases— parallel combinations of two common capacitor values. Inductors were given the nearest value available for coils wound on either T-50-2 or T-50-6 toroid cores. The final filter designs are the result of trade-offs between inductance, capacitance and filter performance. Table 1 shows the filter topology, along with a comparison of the original ideal filter component values and the values selected for the finished unit.

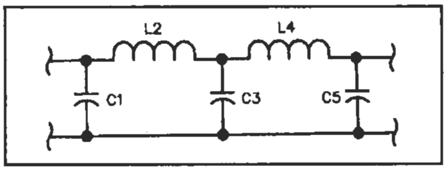

Fig 2—Schematic of the filter used for each band.

Construction

I built my amplifier and low-pass filter modules on single-sided PC boards, using pads to mount the components. No holes are drilled (except for mounting screws) and all leads are attached by soldering them to the pads. The PC-board patterns for the amplifier and filters are available (see Note 2).

Fig 3 shows the amplifier-assembly parts. This assembly is mounted to a heat sink (see Figs 4 and 5) capable of dissipating more than 40 watts without excessive temperature rise. (This assumes a worst case of 50-percent transmitting time, and 7-watts dissipation in standby.) A cutout in the middle of the amplifier board allows placement of the SLAM device. The PC board leaves a conducting path around the ends of the SLAM to maintain a ground potential across the entire board. Four mechanical components make up the amplifier assembly. The first is a 0.1875-inch-thick aluminum base plate to which the SLAM is mounted. Next are two aluminum 0.1-inch-thick spacers, which are placed between the base plate and the circuit board. These spacers set the proper distance from the base plate to SLAM leads. The SLAM is installed through the top of the PC board, and its leads are soldered to the traces on top of the board.

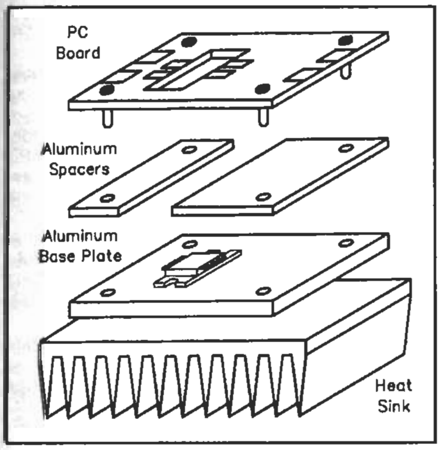

Fig 3—Mechanical assembly of the amplifier-module PC board, aluminum spacers and heat sink.

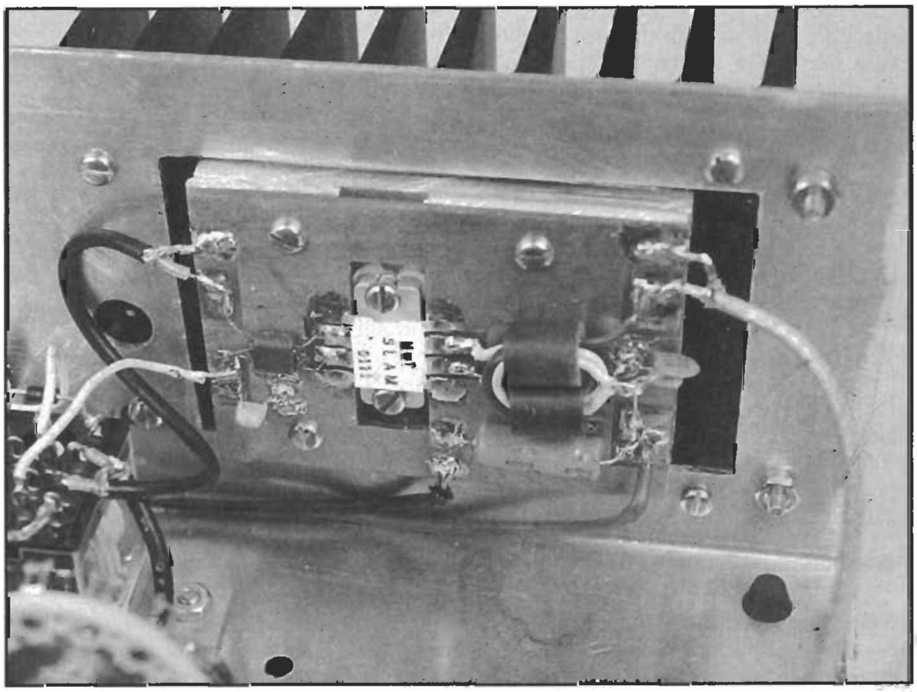

Fig 4—The assembled amplifier-module PC board in position and secured to the heat sink (see Fig 5, next page).

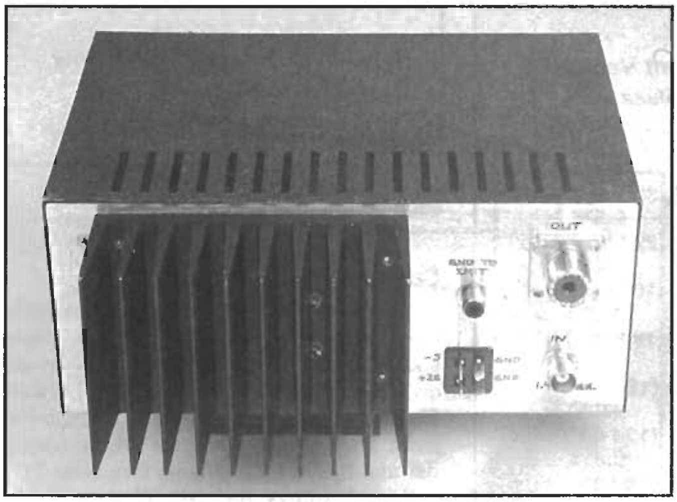

Fig 5—Rear view of the completed amplifier showing the hefty heat sink.

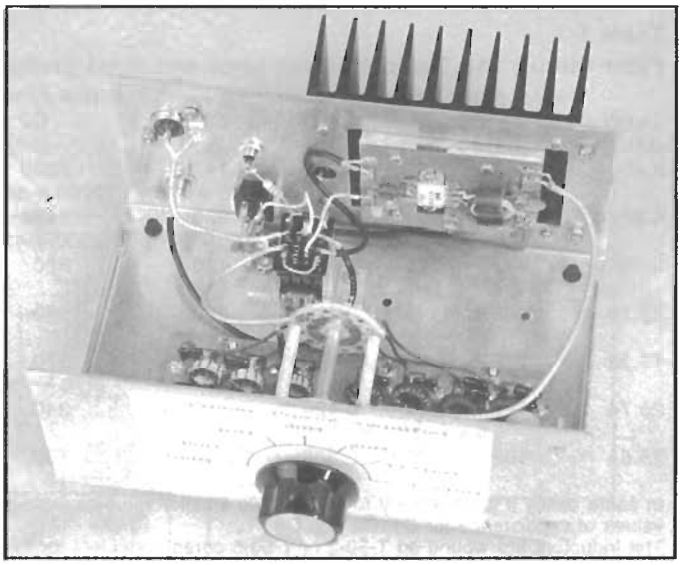

Fig 6—This interior view of the amplifier shows its simple and clean layout. The band switch is centered on the front panel. Immediately beneath the band switch is the filter assembly. Behind the switch and to the left is the TR relay, K1. A four-pin Jones plug power connector is mounted on the rear panel behind and to the left of the relay. On the bottom, near the outside lip of the rear panel, is J1. Above it is J2, with J3 to its right. Most of the rear panel—from its middle to the right lip—is occupied by the SLAM IC PC board and the aluminum spacers secured to the heat sink mounted on the rear panel's exterior. Rubber feet on the cabinet bottom help prevent scratching the supporting surface beneath and keep the amplifier from sliding. The band-switch knob center section is 1⅛ inches in diameter; the skirt flares to a diameter of 1½ inches.

Construction is easiest if the transformer connections to the SLAM are not soldered until after the SLAM is installed. This eliminates the possibility that the transformer connections will get in the way when you try to solder the SLAM into place. As with any power device, place a thin coating of thermal compound between the SLAM and the base plate, and between the base plate and the heat sink. Solder bypass capacitors directly to the transformer center tap and to the ground plane, with the minimum possible lead lengths.

The low-pass filter board is constructed one filter at a time. First, install the capacitor at the center (C3), then the inductors L2, L4), and finally the end capacitors (C1, C5). All inductors are wound with even spacing over three-quarters of the core circumference. Simply solder the capacitors to the pads and ground plane. Silver-mica capacitors were used in the prototype because they were on hand. Ceramic-disc capacitors with 200- to 500-V ratings will work equally well.

If the band switch is located close to the filter board (see Fig 6), short lengths of hookup wire can connect the filters to the switch wafers.

A spacious box houses the filter and amplifier assemblies, along with a TR relay that also switches the standby bias. Power and relay control leads are bypassed where they enter the enclosure.

Before final assembly, I gave the panels of the case a brushed look using a sanding block with oiled sandpaper. Band markings for the switch (see the title-page photo) are drawn on a large, adhesive-backed label attached to the front panel.

Performance

Amplifier gain ranges from 12.5 to 14 dB between 1.8 and 30 MHz. The gain flatness is basically a function of the input and output transformers. (It's possible to make the amplifier gain flat within 1 dB from 1 MHz to 100 MHz using transmission-line transformers and frequency compensation.) The required drive power for 25 watts output is 1.0 to 1.4 watts.

On-the-air performance is excellent. Besides low distortion in the SSB mode, a small advantage of linear amplification is a complete absence of rise and fall distortion of a CW waveform, which sometimes occurs in class-C amplifiers.

Summary

This project shows how new RF products can make home construction of amateur equipment very easy. Home-brewers can benefit from a growing trend in RF product engineering: reducing development time by using "super components" that require few external components and little engineering time to design them into a product.

A secondary purpose of this project is to show how even simple software tools can be used to speed up design. The programs used to design the amplifier's low-pass filters are inexpensive, and accurate at frequencies in the MF/HF bands. In this case, they made it possible to examine tradeoffs among standard-value components for seven different filters, without having to build, measure and tweak each one.

The result is a linear power amplifier with good gain and performance. Its uncomplicated design leaves little room for error, and no fancy test equipment is needed to successfully build it. Projects this easy can make an old-timer forget about the "simpler" days of vacuum tubes!

Notes

[1] Power supplies are available from Marlin P. Jones & Assoc, Inc, PO Box 12685, Lake Park, FL 33403-0685, tel 407-848-8236; fax 1-800-432-9937.

[2] Parts kits for this project are available from Crestone Engineering, PO Box 3702, Littleton, CO 80161, tel 303-770-4709. Each kit includes all electronic and mechanical components for the amplifier module and lowpass filter assembly, including circuit boards, heat sink and rotary band switch. Kits do not include an enclosure, connectors, or TR relay. A 25-W kit using the SLAM-0111 is $115; a 50-W kit using the larger SLAM-0122 is $190. Add $6 per kit for shipping. Payment may be made by check, money order, VISA, MasterCard or American Express.

PC board patterns for the amplifier and filters are available free from the ARRL. Send your request to the Technical Department Secretary, ARRL, 225 Main St, Newington, CT 06111. With your request for the BREED AMPLIFIER PC-BOARD TEMPLATE, enclose a business-size envelope with one First-Class stamp.

[3] Microwave Technology, 4268 Solar Way, Fremont, CA 94538. Their products are distributed by Richardson Electronics, 40W267 Keslinger Road, LaFox, IL 60147, tel 708-208-2200.

[4] Mike Ellis, "A Comprehensive Filter Design Program," RF Design, July 1991. The program is available from the RF Design Software Service, PO Box 3702, Littleton, CO 80161-3702, tel 303-770-4709 (part #RFD0791, $15 postpaid).

[5] NOVA, a shareware program by Robert Stanton, also available from the RF Design Software Service (part #RFD-0391 ,$15 ppd).

By H.O. Granberg, K7ES/OH2ZE

From *QEX*, July 1990

A Compact 1-kW 2-50 MHz Solid-State Linear Amplifier

Solid-state high-power linear amplifiers are becoming more and more popular in the field of ham radio as the prices of HF power transistors continue to fall. 250-W devices are now available for almost half the price they were selling for a few years ago. RF power FETs are still more expensive, but eventually their prices will also fall, although not as fast since they are still novelty items and the manufacturing yields are low due to ESD problems and requirement for cleaner facilities for wafer processing.

General

It is much easier to design wideband power amplifiers with FETs than bipolar transistors mainly due to their higher input impedances at least up to VHF. Their input impedance also varies less with frequency than that of bipolar devices and changes in the output load line are reflected back to the input to a lesser degree because of the much lower value of feedback capacitance (collector to base vs drain to gate). Practically all RF power FETs on the market today are of the enhancement MOS type, meaning that positive voltage at the gate in respect to the source is required to turn the device on.

The 1-kW amplifier described here would be difficult, if not impossible, to design to cover four and a half octaves with comparable performance using inexpensive bipolar transistors. In addition, a series of power splitters and combiners would be required to reach high power levels. Biasing to class AB linear operation is also much simpler with FETs since the gate does not draw any dc current, whereas a current equal to I_C(peak)/h_{FE} must be supplied to the base of a bipolar device. One example of this and the splitter-combiner complexity is presented in the Application Note AN-758 by Motorola, Inc.

This article features a state of the art extremely compact design using a pair of FETs rated for 600 W of power output each. It would be capable of a power output of 1.2 kW as a push-pull circuit, but with the output matching employed, which is optimized at around 800 W, the unit starts saturating at around 1 kW at a 50-V dc supply, resulting in high IM distortion. Similarly at a 40-V supply, it would be usable up to 800 W. The type output matching transformer employed allows only integers as 1:4, 1:9, 1:16, etc. The 1:16 impedance ratio transformer would make the output matching optimized at 1400 W, which would result in a poor efficiency at 1200 W and lower power levels. The only way to compensate for this would be to adjust the supply voltage accordingly, in this case 45-46 V. However, the 1:16 ratio transformer of this type is physically much more difficult to fabricate than the lower ratio ones, and may not be available in the commercial market.

The Bias Regulator

The gate bias regulator (IC1 in Fig 1) allows the main supply voltage to be varied or the use of an unregulated supply while keeping the gate bias voltages and the FET idle currents constant. Since the maximum operating voltage of the regulator IC is only 40 V, a Zener diode (D1) is employed to keep it at a safe level. The regulator supply terminals are separated from the main power supply permitting the use of a separate bias supply if desired. There is also an option for a thermistor connection to stabilize the idle currents against temperature changes. The thermistor should be in a physical contact preferably with a mounting flange of one of the FETs. The gate voltages are individually adjustable (R1, R2) making gate threshold voltage matching of the devices unnecessary. In case of a device failure, such as a drain-gate short, D2 and D3 block the full supply voltage from being fed back to destroy the regulator. R10, R11 and C3, C4 are merely AC filters to protect the regulator from possibly strong RF fields. To set the idle currents, R1 and R2 must be adjusted to minimum. R3 is then adjusted for a regulator output voltage of about double the FET gate threshold voltages (IC1, pin 3). The current is monitored at the main supply voltage point while adjusting Ri for a desired idle current, typically 800 mA-1.0 A. R2 is then advanced until the current is doubled, resulting in equal idle currents for both devices. After this procedure, the settings of R1 through R3 should remain until one or both FETs must be replaced.

The RF Path

The amplifier is designed to operate into the industry standard 50-ohm input and output interface. The impedance matching to the low impedance levels of the FETs is accomplished with broadband RF transformers. Both the input transformer (T1) and the output transformer (T2) are of the so-called conventional type in contrast to transmission line transformers.[1,3,4] Both employ only one turn in the low impedance winding. T2 is far more critical than T1 because it determines the efficiency and the high frequency end gain characteristics, plus it must be able to handle a large amount of RF power. For increased bandwidth characteristics, its low impedance, one turn winding consists of three paralleled 10-ohm coaxial cables, resulting in a tight and controllable coupling between the primary and secondary. According to formulas given in Reference 2, approximately twice the present 4.7 cm^2 ferrite cross sectional area would be required in order for the core not to saturate with the calculated 127 gauss flux density. The saturation mainly occurs at the lowest frequencies, in this case at 2-3 MHz. Unfortunately most ferrite manufacturers do not give information on saturation flux densities that applies to applications such as this. However, it is known that high permeability ferrites, in

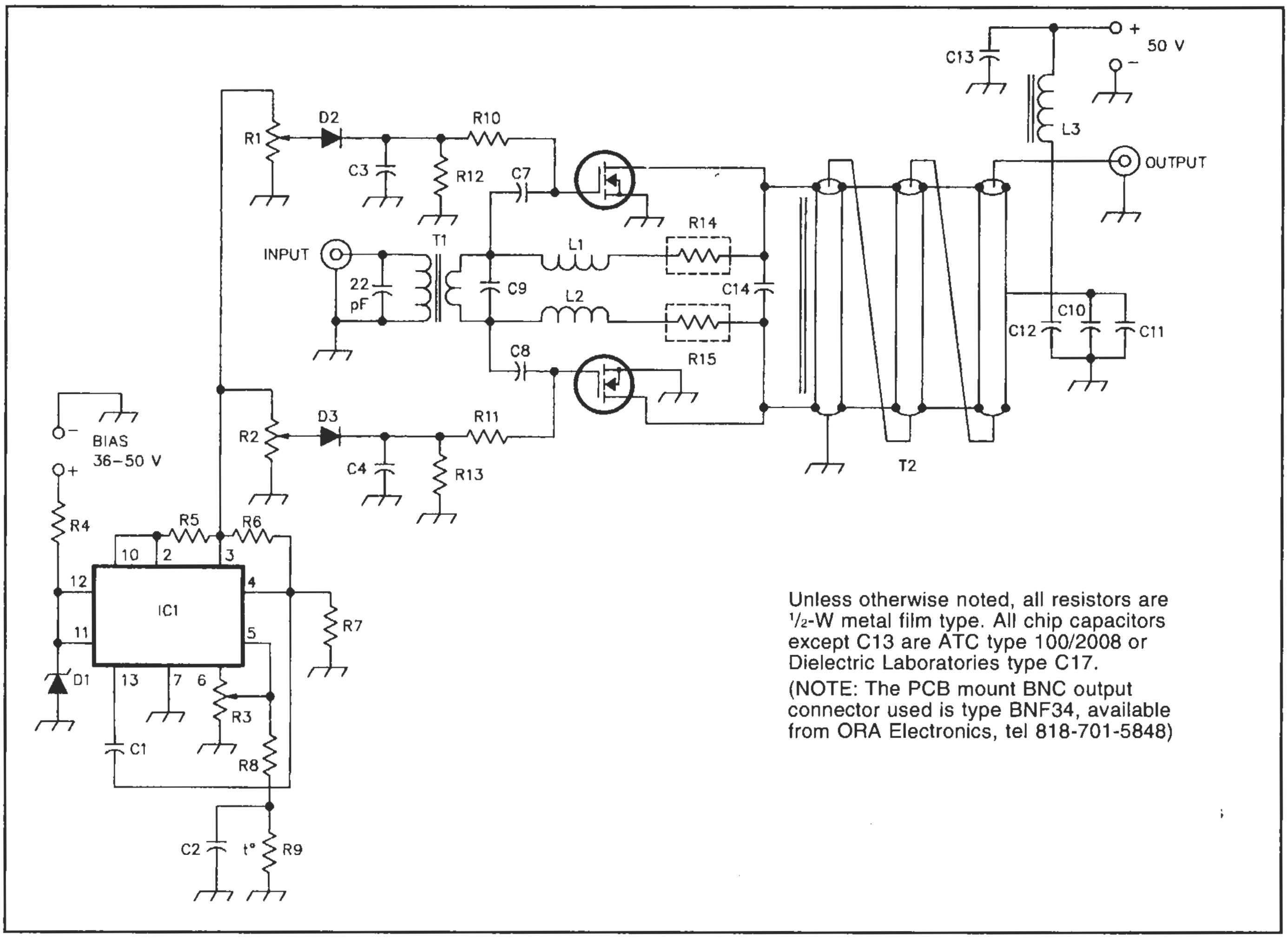

Unless otherwise noted, all resistors are ½-W metal film type. All chip capacitors except C13 are ATC type 100/2008 or Dielectric Laboratories type C17.

(NOTE: The PCB mount BNC output connector used is type BNF34, available from ORA Electronics, tel 818-701-5848)

Fig 1—Circuit Diagram—2 to 50 MHz Amplifier 2-50 MHz Amplifier Components List

R1,R2—1 kΩ single-turn Trimpots
R3—10 kΩ single-turn Trimpot
R4—470 Ω, 2 watts
R5—10 Ω
R6,R12,R13—2 kΩ
R7—10 Ω
R8—Exact value depends on thermistor R9 used (typically 5-10 kΩ)
R9—Thermistor, Keystone RL1 009-5820-97-Di or equivalent
R10,R11—100 Ω, 1 W carbon
R14,R15—EMC Technology model 5308 or KDI Pyrofilm PPR 870-150-3 power resistors, 25 Ω
D1—1 N5357A or equivalent
D2,D3—1N4148 or equivalent
IC1 —MC1723 (723) voltage regulator
C1—1000-pF ceramic disc capacitor
C2,C3,C4—0.1-μF ceramic disc capacitor
C5—0.01-μF ceramic chip capacitor
C6,C12—0.1-μF ceramic chip capacitor
C7,C8—Two 2200-pF ceramic chip capacitors In parallel each
C9—820-pF ceramic chip capacitor
C10,C11—1000-pF ceramic chip capacitor
C13—0.47-μF ceramic chip capacitor or two smaller values in parallel
C14—Unencapsulated mica, 500 V. Two 1000-pF units is series, mounted under T2.
L1,L2—15 ηH, connecting wires to R14 and R15, 1.5 cm each #20 AWG
L3—10 μH, 10 turns #12 AWG enameled wire on Fair-Rite Products Corp ferrite toroid #5961000401 or equivalent
T1,T2—9:1 and 1:9 impedance ration RF transformers, types RF800-3 and RF2067-3 R, respectlvely (RF Power Systems, 3038 E Corrine Dr, Phoenix, AZ 85032)

general, saturate easier than low permeability materials. Thus, the lowest permeability material should be selected that will satisfy the minimum inductive reactance requirement at the lowest frequency of operation. The formula to calculate this is $NX_L = 2R_{S(L)}$, where: X_L=inductive reactance for one turn, N = number of turns, $R_{S(L)}$ = source or load impedance. Low permeability material is also less lossy at high frequencies, resulting in less heat generated in the transformer. T1, which must handle only 8-12 W of power, is made of higher permeability ferrite. This makes it possible to make the unit physically small as well. In Ti, the secondary consists of metal tubes (see Ref 1), where three turns of the primary wire is threaded through. Metal tubes are also used in T2, but only to hold the structure mechanically together.

At high-power levels generated with solid-state devices, which operate at relatively low voltages, the impedance levels automatically become low. This creates a problem for finding passive components, especially capacitors to handle the high RF currents involved. In vacuum tube circuits a similar problem exists, but in the form of high voltages. In this design, C14 gets the roughest treatment. It must be able to carry RF currents in excess of 10 amperes at the higher frequencies, although the voltage across it is only 75 V rms. At first, several good quality ceramic chip capacitors were tried in parallel, but temperature excursions caused them to crack resulting in AF arcs that burned the circuit board in the area as well. Finally, two unencapsulated mica capacitors (brand names such as Unelco, Underwood, Standex, Elmenco and Semco) were soldered in series by attaching the terminal tabs together, making it a sym-

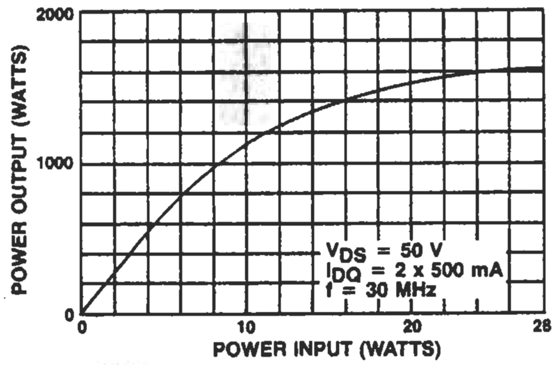

FIGURE 2 — POWER OUTPUT VERSUS POWER INPUT

Fig 2

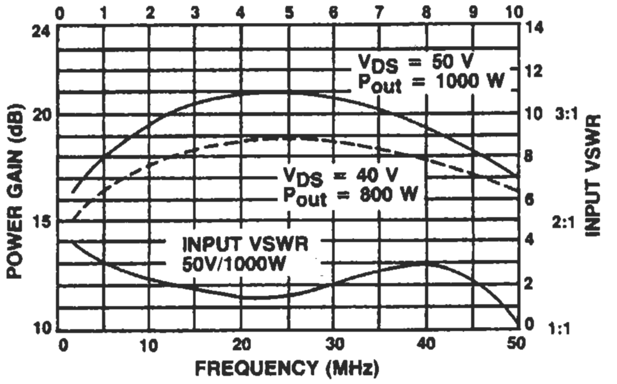

FIGURE 3 — POWER GAIN AND INPUT VSWR VERSUS FREQUENCY

Fig 3

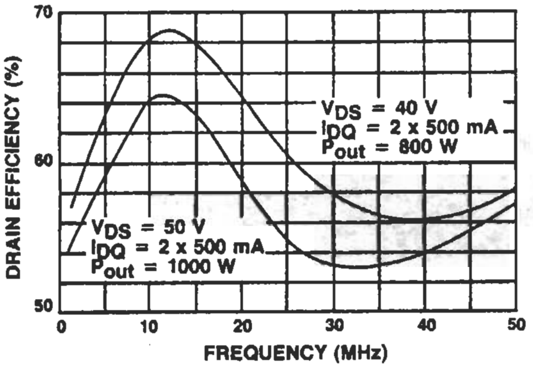

FIGURE 4 — DRAIN EFFICIENCY VERSUS FREQUENCY

Fig 4

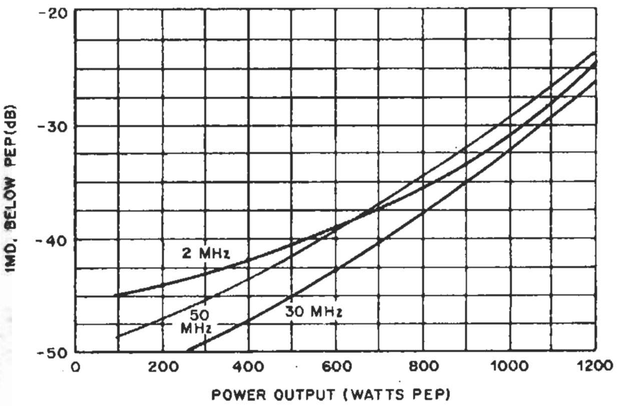

Fig 5

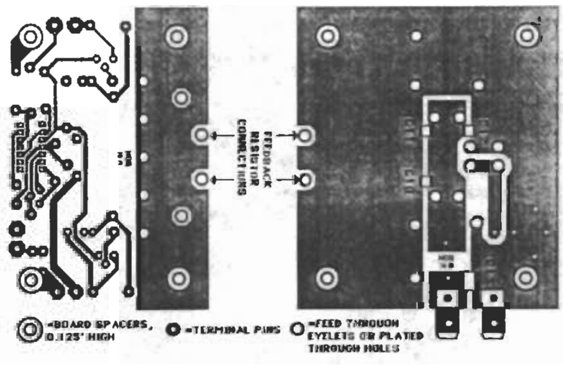
FEEDBACK RESISTOR CONNECTIONS
=BOARD SPACERS, 0.125" HIGH
=TERMINAL PINS
=FEED THROUGH EYELETS OR PLATED THROUGH HOLES

COMPONENT LAYOUT — BOTTOM SIDE

Fig 6

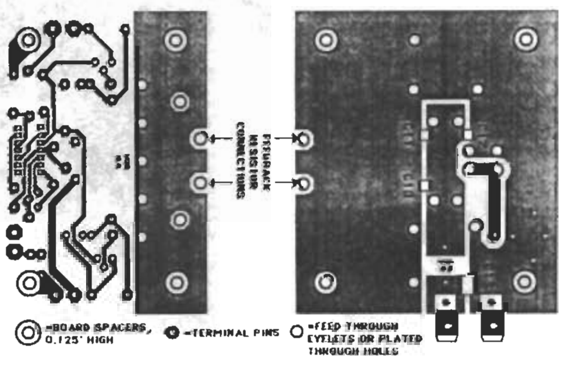
FEEDBACK RESISTOR CONNECTIONS
=BOARD SPACERS, 0.125" HIGH
=TERMINAL PINS
=FEED THROUGH EYELETS OR PLATED THROUGH HOLES

COMPONENT LAYOUT — BOTTOM SIDE

Fig 7

metrical structure. Since each is double the total value required and with double the number of plates, this increases the RF current carrying capability and provides a larger area to be soldered to the board metal foil to make the cooling more efficient. The low impedance winding terminals are then soldered to the tops of the capacitor metal casings, leaving the effective capacitance across the winding. For further fine tuning, an Arco (Elmenco) #469 or Sprague #GM-40900 compression mica trimmer can be soldered to the fronttop terminals of the transformer. Slot openings in the metal foil (Fig 7) located on each side of the output transformer, next to the drain terminals, were provided to increase the series inductance for certain highfrequency narrowband applications. This tunes out some of the FET output capacitance, resulting in increased efficiency. At lower frequencies (below 80 MHz) however, they only add to the IR loss and should be shorted. The location of C9 is also critical and should be placed approximately as shown in Fig 7. This will affect the input VSWR at frequencies above 30 MHz.

Bypass capacitors C10 through C12 must also be of good quality. The center tap of T2 should be free of AF if the circuit is balanced. This may not always be the case, in which case these capacitors will aid this function. L3 and C13 form an additional filter, ensuring that no RF energy is being fed back to the power supply. Switchmode power supplies especially are sensitive against RF and may actually get damaged from it.

Negative feedback is provided through the networks L1-R14 and L2-R15. Its purpose is to produce a relatively flat power gain versus frequency response. It also

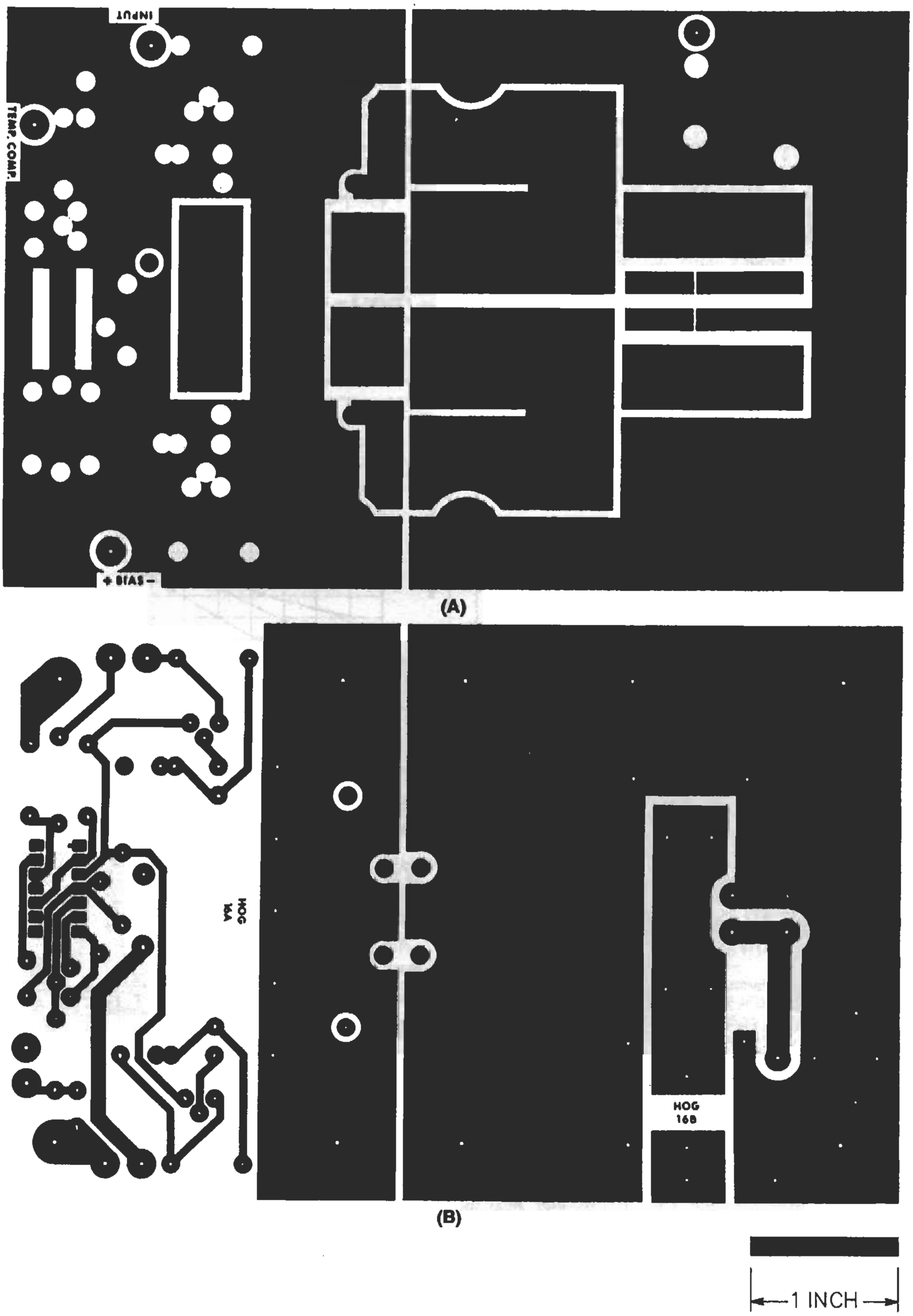
INPUT
TEMP. COMP.
+ BIAS –
(A)
HOG 16A
HOG 16B
(B)
1 INCH

Figs 8A and 8B–Circuit Board–Top Side and Bottom Side

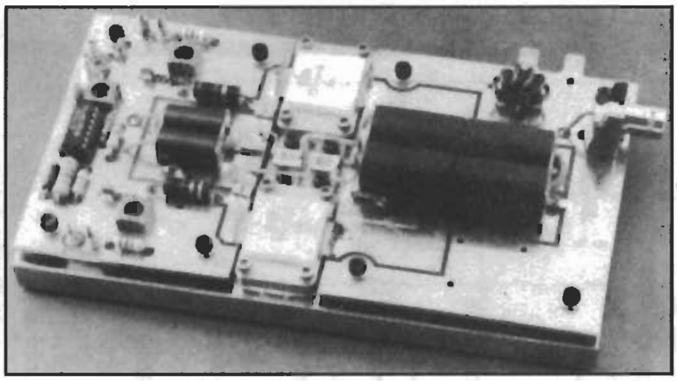

Fig 9—Amplifier mounted to the Heat Spreader.

improves the input return loss and helps to stabilize the amplifier at low frequencies, where the power gain would be 25-30 dB without it. The feedback is at its minimum at the high frequency end and at maximum at low frequencies, where most power is dissipated in R14 and R15. This power is roughly the difference in power input without the feedback between 2 and 50 MHz assuming a constant power output (in this case 25-30 W). A simple formula for calculating the feedback resistor values as well as their dissipation ratings is given in Reference 5. Reference 5 also includes information on physical construction of RF transformers such as used here.

Thermal Aspects

Assuming a 50% worst case efficiency for the unit, each FET dissipates 500 W of heat in an area of 1 x 1.5 inch. It is imperative that the transistors are mounted on the surface of a material with low thermal resistance such as copper. This is called a heat spreader as it is then attached to a heat sink made of material with poorer thermal resistance. It should extend about one inch beyond the edges of the FET mounting flanges at least on three sides. It is even more practical to make the heat spreader as large or larger than the amplifier itself. This would allow all circuit-board spacers to be an equal height of 0.125 inch. The thickness of the heat spreader should be a minimum of 0.375 inch. The heat spreader is then separately attached to the actual heat sink, which can be a 12-inch length of Wakefield Engineering type 4559 extrusion or equivalent[6]. Heat sink compound must be applied to all thermal interfaces and the recommended transistor mounting procedure should be followed, including the screw torque. Fig 9 shows the amplifier mounted to the heat spreader. Although the heat sink is not shown, one must be used for continuous operation and for test periods longer than a couple of minutes. For continuous operation, two 5-inch muffin fans under the heat sink will suffice. They will keep the device case temperature at below 80°C, and the die temperature, which equals to device thermal resistance x power dissipation + case temperature = 0.13 x 500 + 80, at less than 145°C, which is well below the 200-degree maximum recommended value. We must realize that the 500-watt dissipation is only valid when the unit is operated into a 50-ohm load. Under mismatched conditions, depending on the phase angle, the dissipated power may be lower or higher than this value.

Performance

Some of the amplifier performance characteristics are shown in Figs 2 through 5. Although at 30 MHz and above all harmonics are 25 dB or more below the fundamental, an output filter is required to comply with FCC regulations. However, it can be a simpler one than required for the low frequencies, where the third harmonic may be only attenuated 12-15 dB. In push-pull amplifiers, the even harmonics are not usually a problem since they are attenuated by the balanced operation of the circuit. Information on high power low-pass filters for applications as this can be found in Reference 7. These filters are automatically relay switched with BCD code available in most modern transceivers.

References

The circuit boards and other components for this design are available from Communication Concepts, Inc, 508 Millstone Drive, Xenia, OH 45385, tel 513-429-3811/ 220-9677.

[1] Motorola, Inc, Semiconductor Sector Application Notes AN-749 and AN-i 035.

[2] Hilbers, A.H., "Design of HF Wideband Power Transformers," Amperex (Philips) Application Laboratory Report EC06907 and ECO7213.

[3] Blocksome, Aoderick K., "Practical Wideband RF Power Transformers, Combiners and Splitters," *Proceedings of RF Expo*, February 1986.

By Mike Kossor, WA2EBY **From *QST*, March 1999**

A Broadband HF Amplifier Using Low-Cost Power MOSFETs

Part 1—With only 1 W of drive, you'll get over 40 W out—from 160 through 10 meters!

Many articles have been written encouraging experimenters to use power MOSFETs to build HF RF amplifiers. That's because power MOSFETs—popular in the design of switching power supplies—cost as little as $1 each, whereas RF MOSFET prices start at about $35 each!

Over the years, I tucked away several of these articles, waiting for an opportunity to experiment with them. That opportunity came when I received a call from Al, W2OBJ. Al wanted a low-cost linear amplifier to use with his 5 W QRP transmitter when band conditions got poor. Ideally, the amplifier would generate at least 25 W on all the HF bands. Al's inquiry renewed my interest in the topic and provided the motivation I needed to get my project underway.

Al provided me with an extensive list of RF-amplifier construction articles that use power MOSFETs.[1-8] These articles provided useful information about MOSFETs and general guidelines for working with them, including biasing, parasitic-oscillation suppression, broadband impedance-matching techniques and typical amplifier performance data. It was clear from the performance data that Al's desire to get 25 W output from power MOSFETs on 1.8 to 30 MHz was going to be a challenge! The RF output power of most of the amplifiers described in the articles drops off to 10 W or less as frequency increases just to 14 MHz.

An Idea Brews

After hundreds of hours of experimentation, I came up with a design that exceeds our original objective: One watt of input power produces over 40 W of output (after harmonic filtering) from 160 through 10 meters. To the basic amplifier, I added an RF-sensed TR relay and a set of low-pass filters designed to suppress harmonic out-

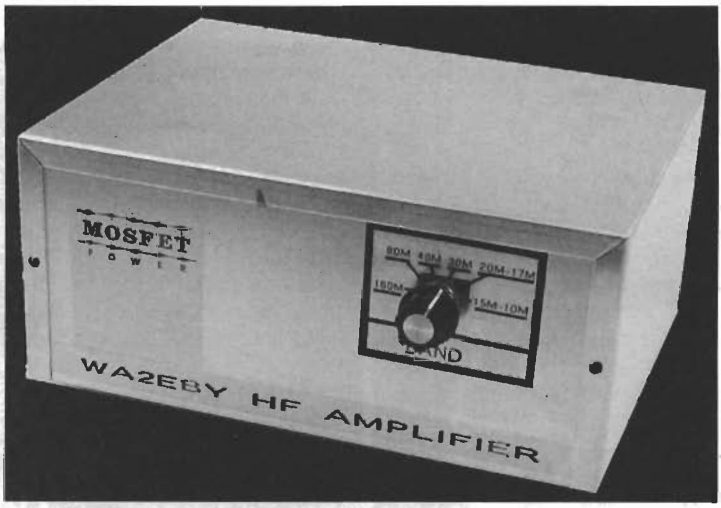

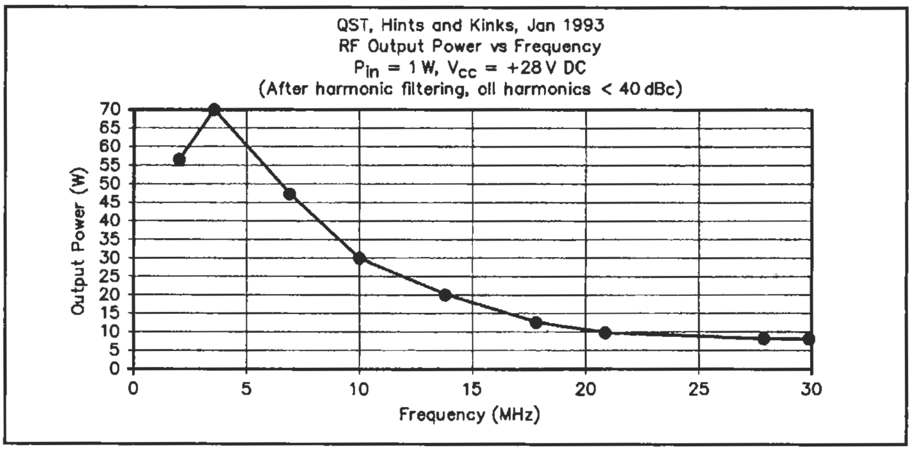

Figure 1—Jim Wyckoff, AA3X, "1 W In, 30 W out With Power MOSFETs at 80 M," Hints and Kinks, *QST*, Jan 1993, pp 50-51.

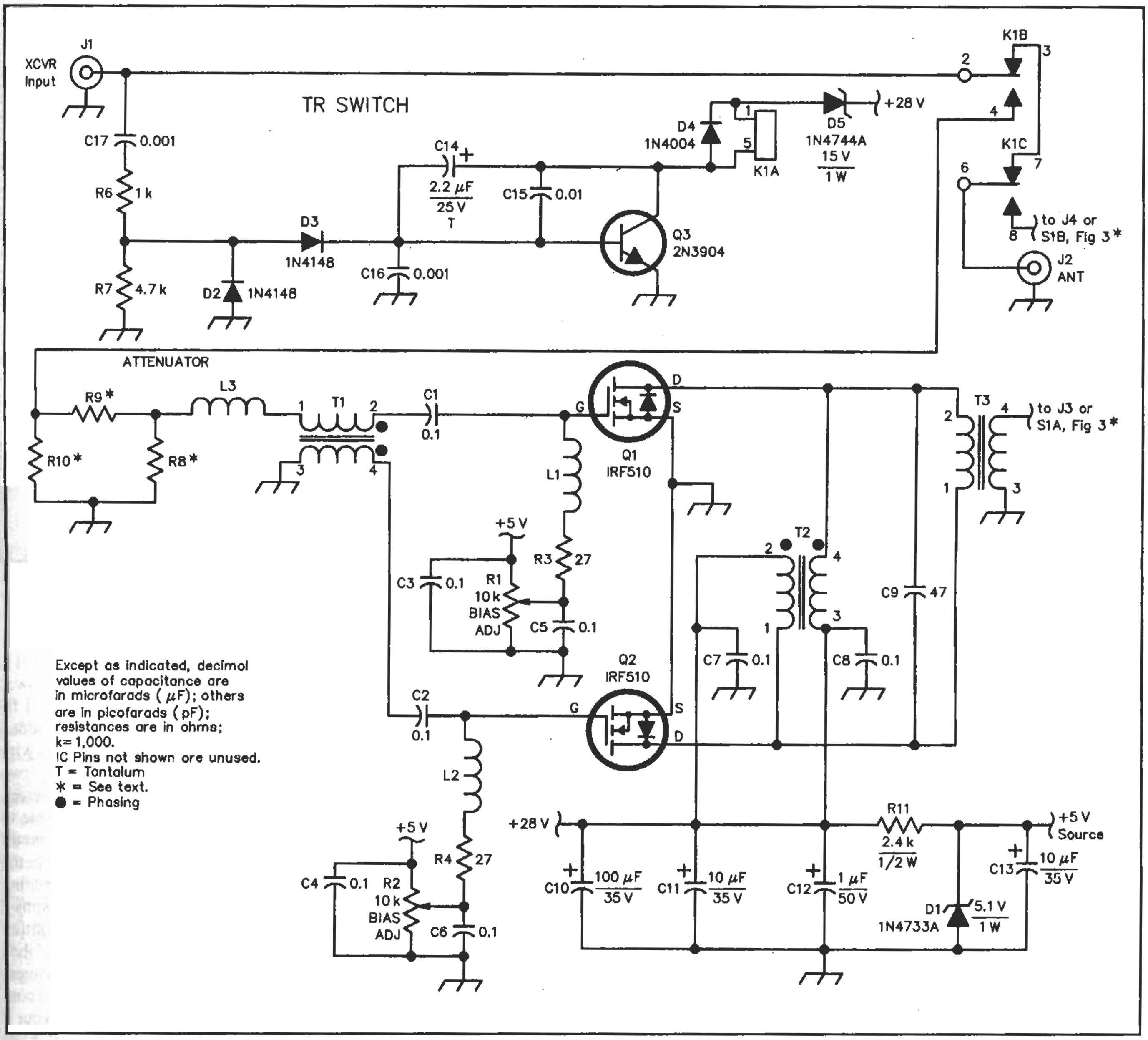
J1
XCVR Input
TR SWITCH
C17 0.001
R6 1 k
R7 4.7 k
D2 1N4148
D3 1N4148
C14 2.2 µF 25 V T
C16 0.001
C15 0.01
Q3 2N3904
D4 1N4004
K1A
D5 1N4744A 15 V 1 W
+28 V
K1B
K1C
to J4 or S1B, Fig 3*
J2 ANT
ATTENUATOR
R9*
R10*
R8*
L3
T1
C1 0.1
Q1 IRF510
L1
+5 V
R3 27
C3 0.1
R1 10 k BIAS ADJ
C5 0.1
T3
to J3 or S1A, Fig 3*
T2
C9 47
C7 0.1
C8 0.1
Q2 IRF510
C2 0.1
L2
R4 27
C4 0.1
R2 10 k BIAS ADJ
C6 0.1
+28 V
C10 100 µF 35 V
C11 10 µF 35 V
C12 1 µF 50 V
R11 2.4 k 1/2 W
+5 V Source
C13 10 µF 35 V
D1 1N4733A 5.1 V 1 W
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k= 1,000.
IC Pins not shown are unused.
T = Tantalum
* = See text.
● = Phasing

Figure 2—Schematic of the MOSFET all-band HF amplifier. Unless otherwise specified, resistors are ¼ W, 5% tolerance carbon-composition or film units. Equivalent parts can be substituted. Part numbers in parentheses are Mouser (Mouser Electronics, 968 N Main St, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422, fax 817-483-0931; **sales@ mouser.com**; **http://www.mouser.com**); see Note 9.

C1-C8—0.1 µF chip (140-CC502Z104M)
C9—47 pF chip (140-CC502N470J)
C10—100 µF, 35 V (140-HTRL35V100)
C11, C13—15 µF, 35 V (140MLR35V10)
C12—1 µF, 50 V (140-MLRL50V1.0)
C14—2.2 µF, 35 V tantalum (581-2.2M35V)
C15—0.01 µF chip (140-CC502B103K)
C16, C17—0.001 µF chip (140-CC502B102K)
D1—1N4733A, 5.1 V, 1 W Zener diode (583-1N4733A)
D4—1N4004A(583-1N4004A)
D2, D3—1N4148 (583-1N4148)
D5—1N4744A, 15 V, 1 W Zener diode (583-1N4744A)
J1, J2—SO-239 UHF connector (523-81-120)
K1—12 V DPDT, 960 Ω coil, 12.5 mA (431-OVR-SH-212L)
L1, L2—9½ turns #24 enameled wire, closely wound 0.25-in. ID
L3—3½ turns #24 enameled wire, closely wound 0.190-in. ID
Q1, Q2—IRF510 power MOSFET (570-IRF510)
Q3—2N3904 (610-2N3904)
R1, R2—10 kΩ trim pot (323-5000-10K)
R3, R4—27 Ω, ½ W (293-27)
R6—1 kΩ chip (263-1K)
R7— 4.7 kΩ chip (263-4.7K)
R8—130 Ω, 1 W (281-130); for 7 dB pad (5 W in, 1 W out)
R9—43 Ω, 2 W (282-43); for 7 dB pad (5 W in, 1 W out)
R10—130 Ω, 3 W (283-130); for 7 dB pad (5 W in, 1 W out)
R8, R10—300 Ω, ½ W (273-300); for 3 dB pad (2 W in, 1 W out)
R9—18 Ω, 1 W (281-18); for 3 dB pad (2 W in, 1 W out)
R11—2.4 kΩ, ½ W (293-2.4K)
T1—10 bifilar turns #24 enameled wire on an FT-50-43 core.
T2—10 bifilar turns #22 enameled wire on two stacked FT-50-43 cores.
T3—Pri 2 turns, sec 3 turns #20 Teflon-covered wire on BN-43-3312 balun core.
Misc: Aluminum enclosure 3.5∞8∞6 inches (HWD) (537-TF-783), two TO-220 mounting kits (534-4724), heat-sink compound (577-1977), amplifier PC board (see Note 9), heat sink (AAVID [Mouser 532-244609B02]; see text), about two feet of RG-58 coax, #24 enameled wire and #20 Teflon-insulated wire.

put and comply with FCC requirements. The amplifier is built on double-sided PC board and requires *no tuning*. Another PC board contains the low-pass filters. Power-supply requirements are 28 V dc at 5 A, although the amplifier performs well at 13.8 V dc.

Several of these amplifiers have been built and exhibit similar performance. Al has been using his amplifier on each of the HF bands, logging well over 500 contacts in 18 months. Signal reports indicate a noticeable improvement in readability (about two S units on average) over his 5 W rig. No indications of in-stability, CW key clicks or distortion on SSB have been reported. To make it easy for you to duplicate this project, PC boards and parts kits are available, all at a cost of about $100![9]

An Overview of MOSFETs

MOSFETs operate very differently from bipolar transistors. MOSFETs are voltage-controlled devices and exhibit a very high input impedance at dc, whereas bipolar transistors are current-controlled devices and have a relatively low input impedance. Biasing a MOSFET for linear operation only requires applying a fixed voltage to its gate via a resistor. With MOSFETs, no special bias or feedback circuitry is required to maintain the bias point over temperature as is required with bipolar transistors to prevent thermal runaway.[10] With MOSFETs, the gate-threshold voltage increases with increased drain current. This works to turn off the device, especially at elevated temperatures as transconductance decreases and $R_{DS(on)}$ (static drain-to-source *on* resistance) increases. These built-in self-regulating actions prevent MOSFETs from being affected by thermal runaway. MOSFETs do not require negative feedback to suppress low-frequency gain as is often required with bipolar RF transistors. Bipolar transistor gain increases as frequency decreases. Very high gain at dc and low frequencies can cause unwanted, low-frequency oscillation to occur in bipolar transistor RF amplifiers unless negative feedback is employed to prevent it. Low-frequency oscillation can damage bipolar transistors by causing excess power dissipation, leading to thermal runaway.

MOSFET Limitations

Of course, MOSFETs do have their limitations. The high gate impedance and the device structure make them susceptible to electrostatic discharge (ESD) damage. Some easily applied precautions prevent this: Use a soldering iron with grounded tip; use a wrist strap connected to ground through a 1 MΩ resistor to bleed off excess body charge while handling MOSFETs and do all work on an anti-static mat connected to ground via a 1 MΩ resistor.

The sensitivity of a MOSFET's gate to static and high-voltage spikes also makes it vulnerable to damage resulting from parasitic oscillation. This undesired self-oscillation could result in excessive gate-to-source voltage that permanently damages the MOSFET's gate insulation. Another MOSFET limitation is gate capacitance. This parameter limits the frequency at which a MOSFET can operate effectively as an RF amplifier. I recommend reviewing the referents of Notes 1-3 if you are interested in more detailed information about MOSFETs.

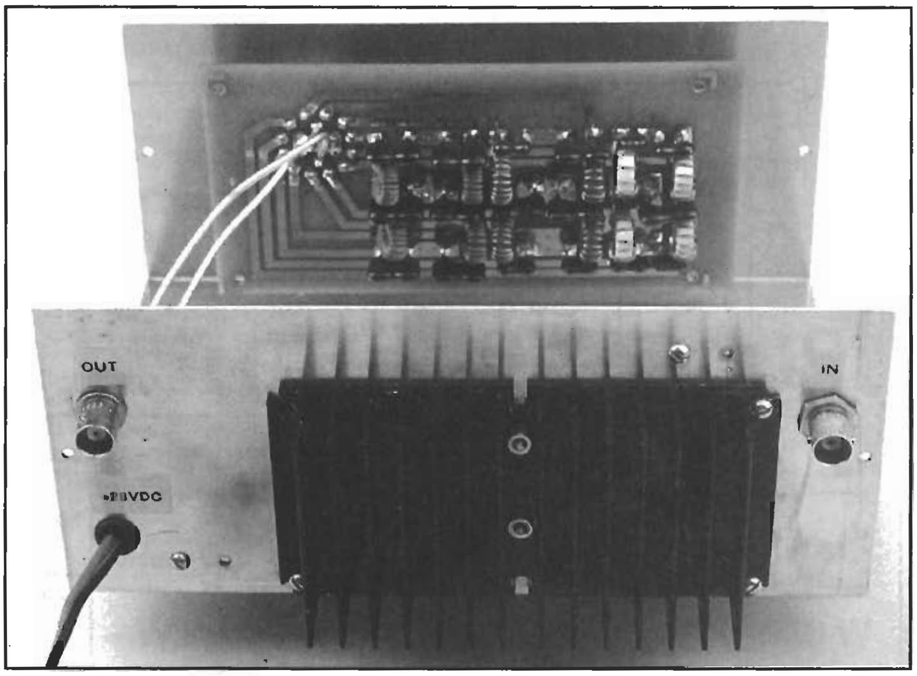

A rear panel view showing the heat sink.

Power MOSFET RF Amplifiers

Of the several power MOSFET amplifiers I built to check their performance, the one providing the best performance is the push-pull design described by Jim Wyckoff, AA3X, in *QST* (see Note 3). I used IRF510 power MOSFETs rather than the IRF511s specified. The performance of this power MOSFET amplifier design is summarized in Figure 1; its basic design is very similar to another amplifier described in the referent of Note 4, written 10 years earlier. That amplifier uses a pair of more-expensive MRF138 MOSFETs designed specifically for RF applications.

As Figure 1 shows, the Hints and Kinks amplifier performance is excellent from 1.8 MHz to 7 MHz and far exceeds the published figure of 30 W output on 3.5 MHz. As frequency increases above 10 MHz, however, output drops off rapidly, falling below 10 W above 21 MHz. (These levels were measured after harmonic filtering.)

Although the amplifier is identified as stable, my first attempt at duplicating the amplifier resulted in oscillations that destroyed one of the IRF510s. I was puzzled by this. At first, I thought the problem was caused by my substitution of the slightly more robust IRF510 MOSFETs for the called-for IRF511s. That idea proved wrong when my second attempt to power up the amplifier with IRF511 MOSFETs installed also resulted in a blown IRF511. (Thank goodness these are $1 power MOSFETs, not $35 RF MOSFETs!). I finally achieved good stability when I added a small amount of inductance in series with the MOSFET source to ground (just two turns of #24 wire, 0.125 inch diameter). With this added inductance, I was able to remove the ferrite beads from the circuit without any sign of instability. I believe the substitution of the IRF510 and minimizing source lead inductance are the reasons I obtained significantly higher RF output power and wider bandwidth than described in the referent of Note 3. This experiment underscores the need to observe *exact* construction techniques and physical layout if similar performance is to be expected. Even though I used PC board construction, I got significantly different results because my layout was not the same as the author's.

Modifying the Design

Although the amplifier performed better than expected, its bandwidth was significantly less than desired. Considerable experimentation (and I do mean considerable!) resulted in the circuit shown in Figure 2. This amplifier consists of two power MOSFETs operating in push-pull and employs an RF-sensed TR relay.

During receive, TR relay K1 is deenergized. Signals from the antenna are connected to J2 and routed through K1 to a transceiver connected to J1. (This path loss is less than 0.3 dB from 1.8 MHz through 30 MHz.) In transmit, RF voltage from the transceiver is sampled by C17 and divided by R6 and R7. D2 and D3 rectify the RF voltage and charge C16. Q3 begins conducting when the detected RF voltage across C16 reaches approximately 0.7 V. This energizes K1, which then routes the transmitted RF signal from J1

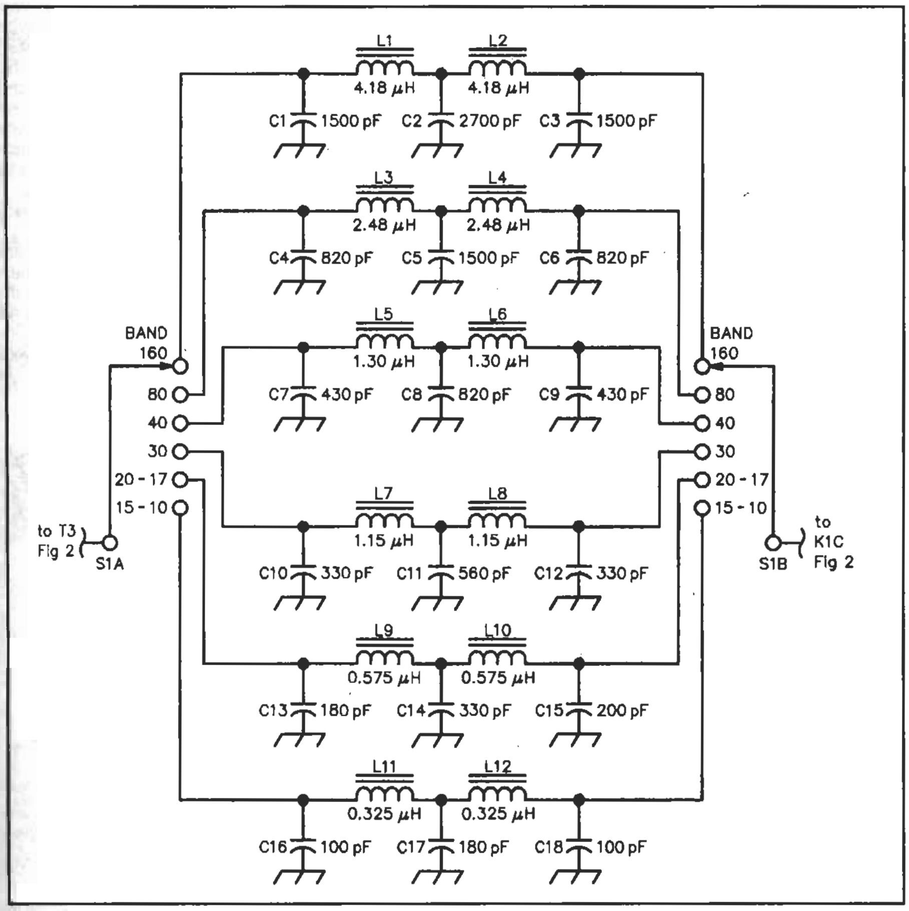

Figure 3—Low-pass filter schematic. In some cases, the actual filter component values differ from the calculated values of a standard 50 Ω-input filter. Such differences improve the impedance matching between the amplifier and the load. Capacitors are all dipped mica units.

C1, C3, C5—1500 pF (5982-19-500V1500)
C2—2700 pF (5982-19-500V2700)
C4, C6, C8—820 pF (5982-19-500V820)
C7, C9—430 pF (5982-15-500V430)
C10, C12, C14—330 pF (5982-19-500V330)
C11—560 pF (5982-19-500V560)
C13, C17—180 pF (5982-15-500V180)
C15—200 pF (5982-15-500V200)
C16, C18—100 pF (5982-10-500V100)
S1—2 pole, 6 position rotary (10YX026)
Misc: low-pass filter PC board (see Note 9)

to the input of the amplifier and sends the output of the amplifier to the antenna at J2. RF-sensed relay response is very fast. No noticeable clipping of the first CW character has been reported.

I made provisions to include an RF attenuator (consisting of R8, R9 and R10) to enable adjusting the amplifier input power to 1 W. (The parts list contains resistor values to reduce the output of 2 or 5 W drivers to 1 W.) The 1 W signal is then applied to the primary of T1 via an input impedance-matching network consisting of L3. T1 is a 1:1 balun that splits the RF signal into two outputs 180 degrees out of phase. One of these signals is applied by C1 to Q1's gate. The other signal is routed via C2 to Q2's gate. The drains of Q1 and Q2 are connected to the primary of output transformer T3, where the two signals are recombined in phase to produce a single output. T3 also provides impedance transformation from the low output impedance of the MOSFETs to the 50 Ω antenna port. Dc power is provided to the drains of Q1 and Q2 by phase-reversal choke, T2. This is a very effective method to provide power to Q1 and Q2 while presenting a high impedance to the RF signal over a broad range of frequencies. The drain chokes for Q1 and Q2 are wound on the same core, and the phase of one of the chokes (see the phasing-dot markings on T2) is reversed. C9 increases the bandwidth of impedance transformation provided by T3, especially at 21 MHz.

The 5 V bias supply voltage is derived from 28 V by Zener diode D1 and current-limiting resistor R11. Bypass capacitors C3, C4, C5, C6 and C13 remove RF voltages from the bias supply voltage. Gate bias for Q1 and Q2 is controlled independently. R1 adjusts Q1's gate-bias voltage via R3 and L1. R2 works similarly for Q2 via R4 and L2.

At low frequencies, the amplifier's input impedance is essentially equal to the series value of R3 and R4. L1 and L2 improve the input-impedance match at higher frequencies. The low value of series resistance provided by R3 and R4 also reduces the Q of impedance-matching inductors L1 and L2, which improves stability. Dc blocking capacitors C1 and C2 prevent loading the gate bias-supply voltage.

C14 keeps transistor Q3 conducting and K1 energized between SSB voice syllables or CW elements. Without C14, K1 would chatter in response to the SSB modulation envelope and fast keying. Increasing the value of C14 increases the time K1 remains energized during transmit. The reverse voltage generated by K1 when the relay is deenergized is clamped to a safe level by D4. D5 drops the 28 V supply to 13 V to power 12 V relay K1. D5 can be replaced with a jumper if K1 has a 28 V dc coil or if you intend to operate the amplifier with a 13.8 V dc supply.

Harmonic Filtering

Although biased for class AB linear operation, this amplifier (like others of its type) exhibits some degree of nonlinearity, resulting in the generation of harmonics. This push-pull amplifier design cancels even-order harmonics (2f, 4f, 6f, etc) in the output transformer, T3. Odd-order harmonics are not canceled. Second-order harmonics generated by the amplifier are typically less than 30 dBc (30 dB below the carrier) whereas third-order harmonics are typically only 10 dBc. FCC regulations require all HF RF-amplifier harmonic output power to be at least 40 dBc at power levels between 50 to 500 W. To meet this requirement, it is common practice for HF amplifiers to use low-pass filters. Separate low-pass filters are needed for the 160, 80, 40 and 30 meter bands. The 20 and 17 meter bands can share the same low-pass filter. So, too, the 15, 12 and 10 meter bands can share a common low-pass filter; see Figure 3.

Switching among the six filters can be a messy wiring problem, especially on the higher-frequency bands where lead lengths should be kept short for optimum performance. This problem is solved by mounting all six low-pass filters on a PC board. A two-pole, six-position rotary switch (S1) mounted directly on the same PC board manages all filter interconnections. One pole of S1 connects the amplifier output to one of the six filter inputs, while S1's other pole simultaneously connects the corresponding filter's output to the TR relay, K1. Only two coaxial-cable connections are required between the RF amplifier and the low-pass filter board.

Next Month

In Part 2, I'll wrap up with amplifier construction and adjustment, and discuss the amplifier's overall performance. See you then!

Notes

[1] Doug DeMaw, W1FB, "Power-FET Switches as RF Amplifiers," *QST*, Apr 1989, pp 30-33. See also Feedback, *QST*, May 1989, p 51.

[2] Wes Hayward, W7ZOI, and Jeff Damm, WA7MLH, "Stable HEXFET RF Power Amplifiers," Technical Correspondence, *QST*, Nov 1989, pp 38-40; also see Feedback, *QST*, Mar 1990, p 41.

[3] Jim Wyckoff, AA3X, "1 Watt In, 30 Watts Out with Power MOSFETs at 80 Meters," Hints and Kinks, *QST*, Jan 1993, pp 50-51.

[4] Doug DeMaw, W1FB, "Go Class B or C with Power MOSFETs," *QST*, March 1983, pp 25-29.

[5] Doug DeMaw, W1FB, "An Experimental VMOS Transmitter", *QST*, May 1979, pp 18-22.

[6] Wes Hayward, W7ZOI, "A VMOS FET Transmitter for 10-Meter CW," *QST*, May 1979, pp 27-30.

[7] Ed Oxner, ex-W9PRZ (SK), "Build a Broadband Ultralinear VMOS Amplifier," *QST*, May 1979, pp 23-26.

[8] Gary Breed, K9AY, "An Easy-to-Build 25-Watt MF/HF Amplifier," *QST*, Feb 1994, pp 31-34.

[9] Parts for this project are available in five modular kits. The following three kits are available from Mouser Electronics (Mouser Electronics, 958 N Main St, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422, fax 817-483-0931; **sales@mouser.com**; **www.mouser.com**): Amplifier components (Mouser P/N 371-HFAMP1) consisting of the amplifier PC board and all PC-board-mounted components (except for the ferrite cores). Price: $35, plus shipping. Amplifier hardware kit (Mouser P/N 371-HFAMP2) consisting of the aluminum enclosure, two UHF connectors, two TO-220 mounting kits, AAVID heat sink and one container of heat sink compound. Price: $30 plus shipping. Low-pass filter kit (Mouser P/N 371-HFAMP3) consisting of the low-pass filter PC board, rotary switch and all PC-board-mounted capacitors (inductor cores are *not* included). Price: $35, plus shipping. Part-placement diagrams accompany the PC boards.

PC boards only are available from Mouser Electronics: HF amplifier board (#371-AMPPWB-2); filter PC board (#371-LPPWB-2). Price $15 each, plus shipping.

The following two kits are available from Amidon Inc (Amidon, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; tel 1-800-898-1883, 714-850-4660, fax 714-850-1163): Amplifier ferrite kit (Amidon P/N HFAFC) containing the ferrite cores, balun core and magnet and Teflon wire to wind the transformers for the HF amplifier. Price: $3.50 plus shipping. Low-pass filter cores kit (Amidon P/N HFFLT) containing all iron cores and wire for the low-pass filters. Price: $4.50 plus ship.

[10] See Motorola Application Reports Q1/95, HB215, *Application Report AR346*.

Thermal runaway is a condition that occurs with bipolar transistors because bipolar transistors conduct more as temperature increases, the increased conduction causes an increase in temperature, which further increases conduction, etc. The cycle repeats until the bipolar transistor overheats and is permanently damaged.

By Mike Kossor, WA2EBY **From *QST*, April 1999**

A Broadband HF Amplifier Using Low-Cost Power MOSFETs

Part 2—Let's put the finishing touches on this all-band HF amplifier!

Last month,[11] I covered the history and development of this 40 W (average) amplifier. I'm sure you're anxious to get your *amplifier finished and on the air, so let's get going!*

Amplifier Construction

The amplifier is constructed on a double-sided PC board with plated through holes to provide top-side ground connections. I used chip resistors and capacitors to simplify construction, but leaded capacitors may work if lead lengths are kept short. First, assemble all chip capacitors and resistors on the PC board. Tweezers help to handle chip components. Work with only one component value at a time (chip caps and resistors are very difficult to identify!). Chip capacitor and resistor mounting is simplified by tinning one side of the PC board trace with solder before positioning the capacitor or resistor. Touch the soldering iron tip to the capacitor or resistor to tack it in place. Finish mounting by soldering the opposite side of the component. *Don't apply too much heat to chip capacitors.* The metalized contacts on the capacitor can be damaged or completely removed if too much heat is applied. Use a 15 to 20 W soldering iron and limit soldering time to five seconds.

Mount axial-leaded resistors, diodes and remaining capacitors next. To avoid damaging them, mount inductors and transformers last. L1 and L2 are wound on a 0.25-inch drill-bit shaft. By wrapping the wire around the shaft 10 times, you'll get 9½ turns. The last turn arcs only a half-turn before entering the PC board. L3 is wound on a 0.190-inch diameter drill bit with 3½ turns wound the same way as L1 and L2. Mounting K1 is simplified by first bending all its leads 90° outward so it lies flat on the PC board. Use a wrist strap connected to ground through a 1 MΩ resistor to bleed off static body charge while handling MOSFETs, and do the work on an anti-static mat connected to ground via a 1 MΩ resistor. The gate input can be damaged by electrostatic discharge!

When winding T3, wind the primary first and add the secondary winding over the primary. Be sure to use Teflon-insulated wire for T3's windings; the high operating temperatures encountered will likely melt standard hook-up wire insulation.

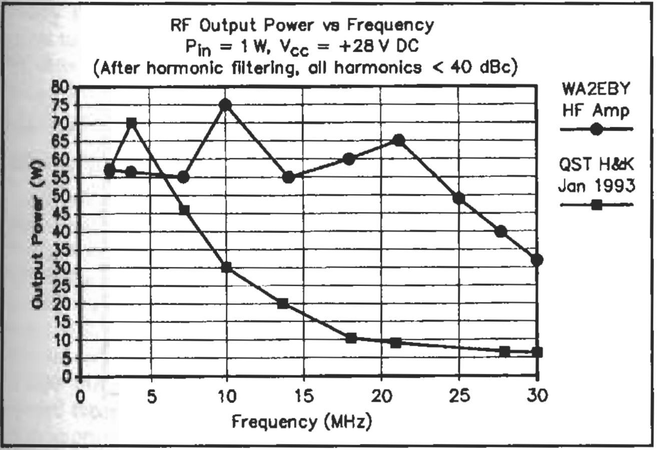

Figure 4—RF output power comparison of the Hint and Kink amplifier and this design.

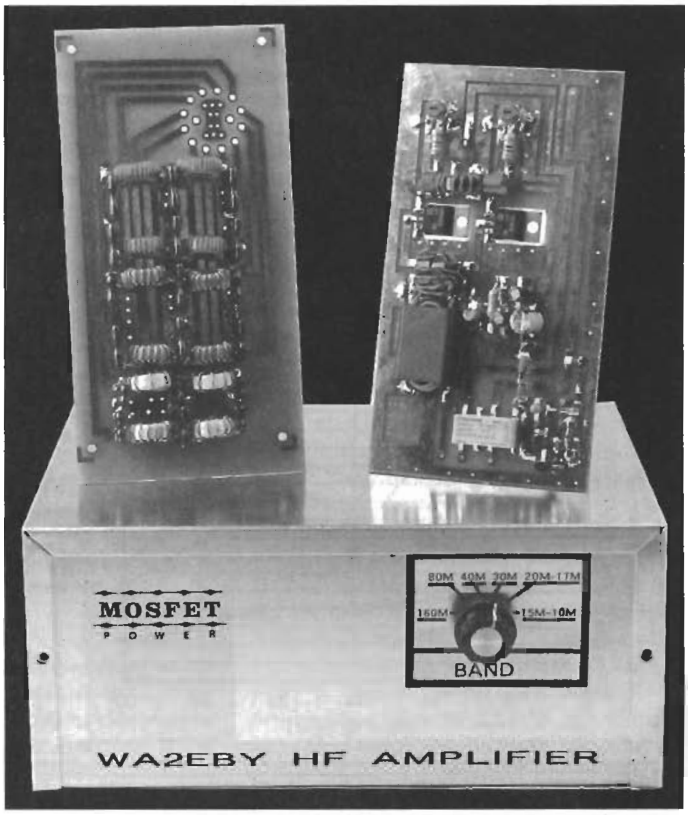

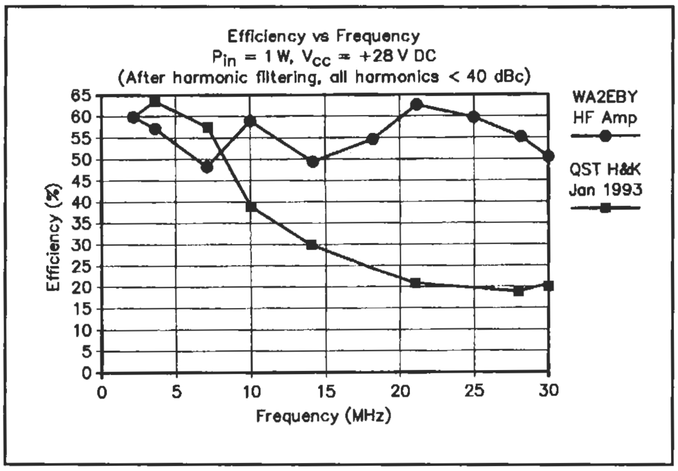

Figure 5—Efficiency comparison of the Hint and Kink amplifier and this one.

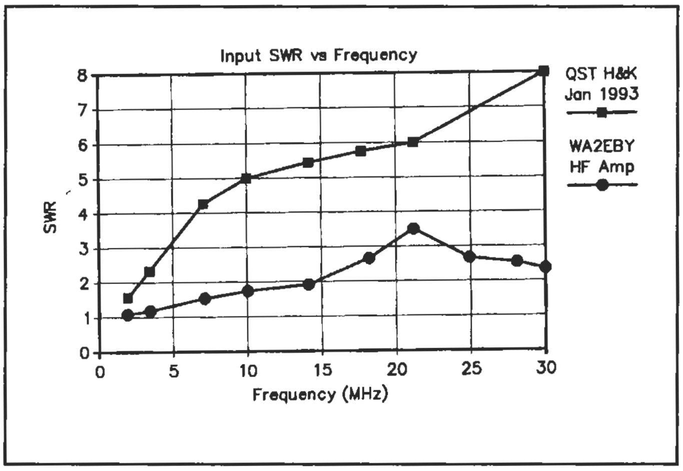

Figure 6—Input SWR comparison of the two amplifiers.

Heat Sinking

Together, Q1 and Q2 dissipate up to 59 W. A suitable heat sink is required to prevent the transistors from overheating and damage. I used an AAVID 244609B02 heat sink originally designed for dc-to-dc power converters. The amplifier PC board and heat sink are attached to an aluminum enclosure by two #4-40 screws drilled through the PC board, enclosure and heat sink at diagonally opposite corners. A rectangular cutout in the enclosure allows Q1 and Q2 direct access to the heat sink. This is essential because of the large thermal impedance associated with the TO-220 package (more on this topic later). Mark the locations of the transistor-tab mounting-hole location in the center of the heat sink in between the cooling fins. Disassemble the heat sink to drill 0.115 inch holes for #4-40 mounting screws, or tap #4-40 mounting holes in the center of the heat-sink fins.

Use mica insulators and grommets when mounting Q1 and Q2 to prevent the #4-40 mounting screws from shorting the TO-220 package drain connections (tabs) to ground. Coat both sides of the mica insulator with a *thin* layer of thermal compound to improve the thermal conduction between the transistor tab and the heat sink. Be sure to install the mica insulator on the heat sink *before* assembling the amplifier PC board to the enclosure and heat sink. The mica insulators are larger than the cut outs in the PC board, making it impossible to install them after the PC board is mounted.

Low-Pass Filter Construction

Inductor winding information for the low-pass filters is provided in Table 1.

Single Band

A PC-board trace is available on the amplifier PC board next to amplifier output (J3) to allow the installation of a single-band low-pass filter between the terminals of J3 and K1's input, J4. This is handy if you intend to use the amplifier on one band only. The input inductor of the low-pass filter connects from J3 to the single PC trace adjacent to J3. The output inductor connects in series between the single PC trace to J4. The three filter capacitors connect from J3, J4 and the PC-board trace near J3 to ground. *This single trace is not used when multiple filters are required.* Remember to remove the single trace adjacent to J3 on the amplifier PC board before attaching the amplifier board between the RF connectors on the enclosure's rear panel.

Multiple-Band Filters

Using the amplifier on more than one band requires a different approach. A set of

Table 1
Low-Pass Filter Inductor Winding Information

(Refer to Figure 3 in Part 1))

Inductor Number	*No. of Turns*	*Core*
L1, L2	30 turns	T-50-2
L3, L4	22 turns	T-50-2
L5, L6	16 turns	T-50-2
L7, L8	14 turns	T-50-2
L9, L10	11 turns	T-50-6
L11, L12	8 turns	T-50-6

Note: All inductors are wound with #22 enameled wire except for L1–L4, which are wound with #24 enameled wire.

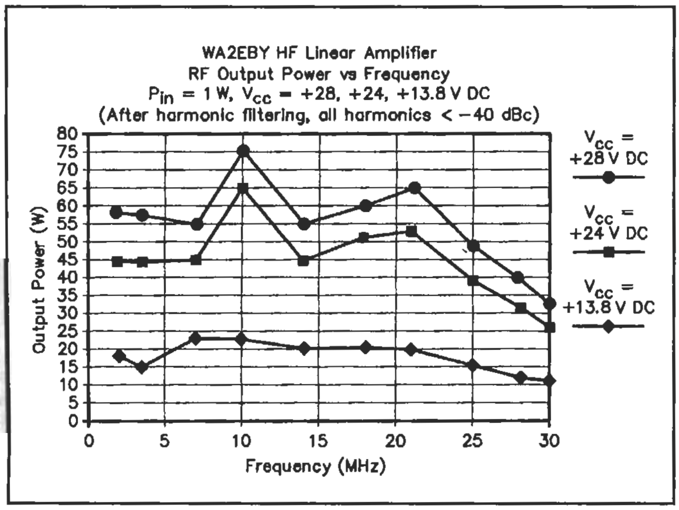

Figure 7—RF output power versus supply-voltage of this amplifier.

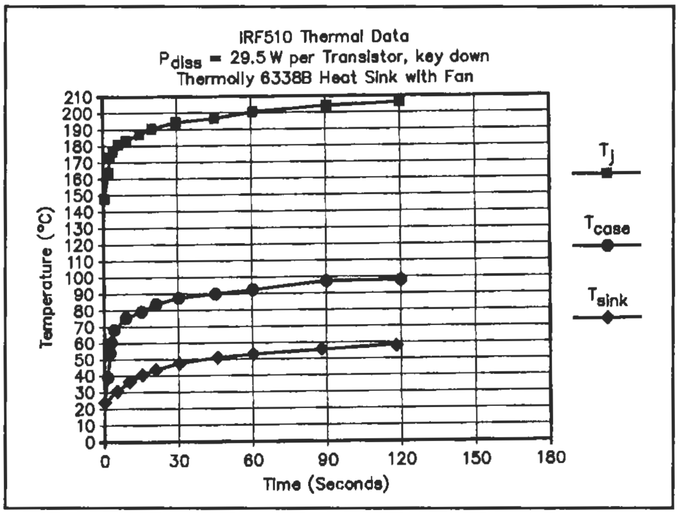

Figure 8—Thermal performance of the amplifier during key-down conditions.

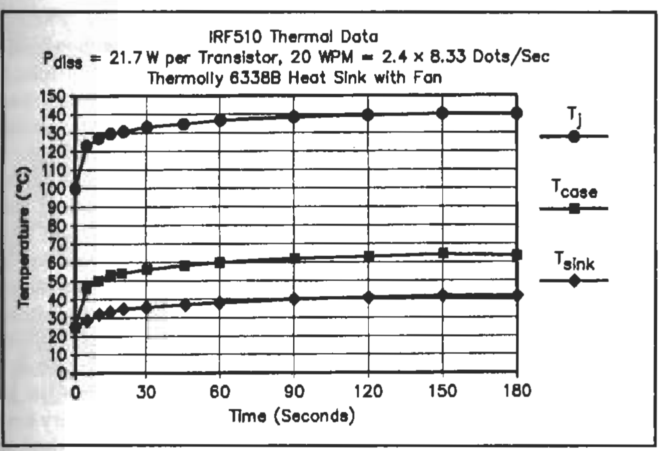

Figure 9—Thermal performance of the amplifier during simulated CW conditions.

six low-pass filters is built on a double-sided PC board with plated through holes to provide top-side ground connections. A PC-board mount, two-pole, six-position rotary switch does all low-pass filter selection. Silver-mica, leaded capacitors are used in all the filters. On 160 through 30 meters, T-50-2 toroids are used in the inductors. T-50-6 toroids are used for inductors on 20 through 10 meters. The number of turns wound on a toroid core are counted on the toroid's OD as the wire passes through the core center (*The ARRL Handbook* [12] provides complete details for winding toroids). Assemble one filter section at a time starting with the 160, 80, 40-meter filter, then the 30-meter filter. With the switch mounting position at your upper left, the filter input (C1) is near the top edge of the board and the filter output (C3) is near the bottom edge. *The last two filters are out of sequence;* the 15-10 meter filter comes *before* the 20-17 meter filter) and the inputs/outputs are reversed to simplify the PC-board layout. The input capacitors, C13 and C16, are mounted on the board *bottom edge*, and output capacitors, C15 and C18, are on the *top edge*.

Use care when assembling the rotary switch. All 14 terminals must fit through the PC board without damaging or bending the pins. Make sure there are no bent pins before you attempt assembly. Insert the rotary switch into the PC board. Do *not* press the rotary switch all the way into the PC-board holes flush with the ground plane! If you do, the top flange of the signal pins may short to the ground plane.

Bias Adjustment

The biasing procedure is straightforward and requires only a multimeter to complete. First, set R1 and R2 fully counterclockwise, (0 V on the gates of Q1 and Q2). Terminate the RF input and outputs with a 50 Ω load. Next, connect the 28 V supply to the amplifier in series with a multimeter set to the 0-200 mA current range. Measure and record the idling current drawn by the 5 V bias supply. The value should be approximately 9.5 mA (28 – 5.1 V) / 2.4 kΩ = 9.5 mA). Set Q1's drain current to 10 mA by adjusting R1 until the 28 V supply current increases by 10 mA above the idling current (9.5 + 10 = 19.5 mA). Next, adjust R2 for a Q2 drain current of 10 mA. This is accomplished by adjusting R2 until the 28 V supply current increases by an additional 10 mA (to 29.5 mA).

Amplifier Performance

With a 28 V power supply and 1 W of drive, the RF output power of this amplifier exceeds 40 W from 1.8 MHz through 28 MHz. Peak performance occurs at 10 MHz, providing about 75 W after filtering! A performance comparison between this amplifier and my modified version of the Hint and Kink amplifier mentioned earlier is shown in Figure 4.

As shown in Figure 5, this amplifier achieves an efficiency of better than 50% over its frequency range, except at 7 MHz where the efficiency drops to 48%. In contrast, the Hint and Kink amplifier delivers greater efficiency between 1.8 and 7 MHz, but it drops rapidly to only 20% as frequency is increased.

Figure 6 compares the input SWR of the two amplifiers. The Hint and Kink amplifier's SWR is acceptable (< 2:1) only at 1.8 MHz. This amplifier is better, however it, too, exceeds 2:1 above 14 MHz. The input SWR of this amplifier can be improved to better than 2:1 on all bands by adding a 3 dB pad (R8-R10 of Figure 2) at the input and supplying 2 W to the pad input. This keeps the amplifier drive at 1 W.

Figure 7 graphs this amplifier's RF output power as a function of drain supply voltage. During this test, the amplifier RF drive level was kept constant at 1 W. As you can see, even when using a 13.8 V dc supply, the amplifier provides over 10 W output (a gain of more than 10 dB) from 1.8 to 30 MHz.

Operation

The amplifier requires no tuning while operating on any HF amateur band. You must, however, *be sure to select the proper low-pass filter prior to transmitting.* If the wrong low-pass filter is selected, damage to the MOSFETs may result. Damage will likely result if you attempt to operate the amplifier on a band with the low-pass filter selected for a lower frequency. For example, driving the amplifier with a 21 MHz signal while the 1.8 MHz low-pass filter is selected will likely destroy Q1 and/or Q2.

The amplifier can also be damaged by overheating. This limitation is imposed by the TO-220 packages in which Q1 and Q2 are housed. The thermal resistance from junction to case is a whopping 3.5°C/W. This huge value makes it virtually impossible to keep the junction temperature from exceeding the +150°C target for good reliability. Consider the following conditions: key down, 1 W input, 53 W output on 7 MHz (worst-case band for efficiency). The amplifier consumes 28 V × 4 A = 112 W, of which 53 W are sent to the antenna, so 59 W (112 W – 53 W = 59 W) are dissipated in Q1 and Q2. Assuming equal current sharing between Q1 and Q2, each transistor dissipates 29.5 W. To keep the transistor junction temperature below +150°C requires preventing the transistor case temperature from exceeding 46.8°C (150 – [3.5 × 29.5]) while dissipating 29.5 W. Also, there is a temperature rise across the mica insulator between the transistor case and heat sink of 0.5°C/W. That makes the maximum allowable heat-sink temperature limited to 46.8 – (0.5 × 29.5) = 32°C. In other words, the heat sink must dissipate 59 W (29.5 from each transistor) with only a 7°C rise above room temperature (25°C). Even if the junction temperatures were allowed to reach the absolute maximum of 175°C, the heat sink temperature must not exceed 57°C. Accomplishing this requires a heat sink with a thermal resistance of (57 – 25) / 59 = 0.54°C/W. This is far less than the 1.9°C/W rating of the AAVID 244609B02 heat sink I used. The situation may seem bleak, but all is not lost. These calculations make it clear that the amplifier should not be used for AM, FM or any other continuous-carrier operation. The amplifier should be used only for CW and SSB operation where the duty cycle is significantly reduced.

Thermal performance of the amplifier is illustrated in Figure 8. Data was taken under dc operating conditions with power-dissipation levels set equal to conditions under RF operation. A RadioShack brushless 12 V dc fan (RS 273-243A) blows air across the heat sink. Key down, the maximum rated junction temperature is reached in as little as five seconds as illustrated in Figure 8. Prolonged key-down transmissions should be avoided for this reason.

Under intermittent CW conditions, the situation is very different. Transistor-case temperatures reached 66°C after operating four minutes under simulated CW conditions at 20 WPM (60 ms on, 60 ms off). The corresponding junction temperature is +141°C (based on an equivalent RMS power dissipation of 21.7 W per transistor). This keeps the junction temperature under the 150°C target (see Figure 9). One simple way to reduce power dissipation is to reduce the power-supply voltage to 24 V. RF output power will decrease about 10 W from the maximum levels achieved with a 28 V supply.

From a thermal standpoint, the IRF510 power MOSFET is a poor choice for this RF amplifier application. Although I must say I am impressed with the robustness of these devices considering the times I spent testing them key down, five minutes at a time, without failure. Q1 and/or Q2 may need to be replaced after a year or so of operation because of the compromise in reliability. Considering their low cost, that is not a bad trade-off.

Stability

High gain, broad bandwidth and close input/output signal routing (within the TR relay) all work against stability. With a good load (< 2:1 SWR) the amplifier is stable from 1.8 MHz through 39 MHz. Oscillation was observed when the transmitter frequency was increased to 40 MHz. The output load match also affects stability. Oscillation was observed on 27.5 MHz when the load SWR was 3:1. This should not be a problem since the frequency is outside the ham bands. I spent a great deal of time trying to make this design unconditionally stable even with loads exceeding 3:1 SWR without sacrificing output power (gain) at 28 MHz without success. I did identify some reasonable compromises.

One of the easiest ways to improve stability and the input SWR seen by the RF source is to add an RF attenuator (pad) at the amplifier input. An attenuator is absolutely required if the transmitter (driver) provides more than 1 W to the amplifier. R8, R9 and R10 form an RF attenuator that attenuates the transmitter drive level, but does not attenuate received signals because it is only in the circuit when K1 is energized. To drive this amplifier with a 2-W-output transmitter requires use of a 3-dB pad. The pad improves the amplifier input SWR and the isolation between the amplifier's input and output. The drawback is that 1 W is wasted in the pad. Likewise, a 5-W driver requires use of a 7-dB pad, but 4 W are wasted in the pad. (Values for R8, R9 and R10 to make a 3-dB pad and a 7-dB pad are given in the parts list.) Installing a pad requires cutting the PC-board trace *under R9*, otherwise R9 would be shorted out by the trace. Make a small cut (0.1 inch wide) in the trace under R9 before soldering R9 in position. R8 and R10 have the same values, but may have different power ratings. Connect R10 between the RF input side of R9 and ground. Install R8 between the amplifier side of R9 and ground.

An impedance mismatch between the output of a 1-W-output driver and the amplifier input can be a source of instability. (Obviously, if the driving transmitter's output power is only 1 W, you can't use a pad as described earlier.) If you encounter stability problems, try these remedies: Place a resistor in parallel with L1 and L2 to decrease the Q of the amplifier matching network (try values between 50 and 220 Ω). Try reducing the value of L3 or eliminating L3 entirely. Both of these modifications improve stability, but reduce the amplifier's output power above 21 MHz.

Summary

This project demonstrates how inexpensive power MOSFETs can be used to build an all-band linear HF power amplifier. Frequency of operation is extended beyond the limits of previous designs using the IRF510 and improved input-impedance matching. Long-term reliability is recognized as a compromise because of the poor thermal performance of the low-cost TO-220 package.

If you have been thinking about adding an amplifier to your QRP station, this project is a good way to experiment with amplifier design and is an excellent way to become familiar with surface-mount "chip" components. I made arrangements with Mouser Electronics and Amidon Inc to provide parts kits for this project at a discounted price (see the parts list in Part 1). These parts kits make it very easy to get started and more economical to "homebrew" this project.

Acknowledgments

I want to thank the following individuals associated with this project: Harry Randel, WD2AID, for his untiring support in capturing the schematic diagram and parts layout of this project; Al Roehm, W2OBJ, for his continued support and encouragement in developing, testing, editing and publishing this project; Larry Guttadore, WB2SPF, for building, testing and photographing the project; Dick Jansson, WD4FAB, for thermal-design suggestions; Adam O'Donnell, N3RCS, for his assistance building prototypes; and my wife, Laura, N2TDL, for her encouragement and support throughout this project.

Notes

[11] Mike Kossor, "A Broadband HF Amplifier Using Low-Cost Power MOSFETs—*Part 1*," *QST*, Mar 1999, pp 40-43.

[12] R. Dean Straw, N6BV, *The 1999 ARRL Handbook for Radio Amateurs*, (Newington: ARRL), 76th ed, pp 25-23*ff*.

By Zack Lau, KH6CP/1 **From *QEX*, May 1992**

A 1.8 to 54 MHz 5-Watt Amplifier

Need a rugged and stable amplifier for your multiband QRP rig? Not only has the design been optimized on a pricey computer program called *Touchstone* (by eesof) for unconditional stability, it has actually survived a variety of poor loads—it was used to sweep filters with 5 W of RF. The gain of the two stage amplifier was measured to be between 28 and 30 dB in the amateur bands, though there is another dB of gain around 37 MHz.

For ruggedness and ease of design, a Motorola MRF 137 was selected as the final transistor. While the MRF 138 may be more linear, insufficient design information was available to ensure a stable design. While some amateurs will balk at the high cost of these devices ($24 in November 1991), such savings are easily lost if the cheaper device has a habit of blowing up. Also, one picks up a real clean SSB signal—the high-order IMD products are way down compared to typical bipolar amplifiers. For instance, the worst IMD on 3.5,7,14 and 28 MHz was –38 dB on 28 MHz, with the 5th order products 61 dB down (relative to PEP). The device was putting out 5-W PEP while being biased at 0.5 A (28-V supply).

Perhaps the biggest flaw is the power requirement— these FETs really like to see high voltages for best performance, and the MRF 137 is no exception. I biased the MRF 137 for 0.55 A and 28.2 V. It drew 0.6 A when putting out 4.6 W at 28 MHz. The driver runs off your normal 12-V supply.

The input amplifier shown in Fig 1a is pretty straightforward—a bipolar 2N5109 with the feedback networks adjusted to compensate the gain of the MRF 137. A series network of a 470-a resistor and a 12-pF capacitor was tacked between the collector and ground to ensure stability at all frequencies. The MRF 137 rolls off a few dB at 54 MHz, but the bipolar amplifier adequately compensates for this gain deficiency. The input return loss is better than 18 dB between 1.4 and 29.9 MHz, but degrades to 12 dB at 50 MHz. The input SWR was not tested with poor loads.

By itself, the MRF 137 amplifier stage shown in Fig 1b (see next page) makes an excellent 16-dB gain block between 1 and 32 MHz, having less than 0.5 dB of gain variation. The transmission line transformer on the input seems to help the input return loss/SWR, keeping these numbers above 18/below 1.3 to 1 between 1 and 50 MHz. I suppose that putting another transmission line transformer on the output could be used to get a more powerful amplifier with less gain over a similar frequency range, but this variation has not been investigated.

The simplest circuit board I could think of was used— I cut two pads in a piece of double-sided circuit board for the gate and drain leads. Then I wrapped the edges of the board with copper tape and soldered it down for good grounding. After making holes for the MRF 137 transistor and the mounting screws in the board and a spacer, made of 0.050-inch aluminum, I attached the spacer, the circuit board, and the MRF 137 to a heat sink tapped with 4-40 screw holes. Standard ground-plane construction was used to attach the other parts. The 2N5109 amplifier was built on its own ground plane—RF amplifiers work better if there isn't too much gain in one place. Three additional amplifiers were built by Mike Gruber, WA1SVF, for use in the lab. He noted that R8 had to be changed from 4.7 kΩ to 1 kΩ to bias the amplifiers at 0.5 amps. Apparently, the MRF 137s he used have a higher gate threshold voltage. Otherwise, performance was as expected.

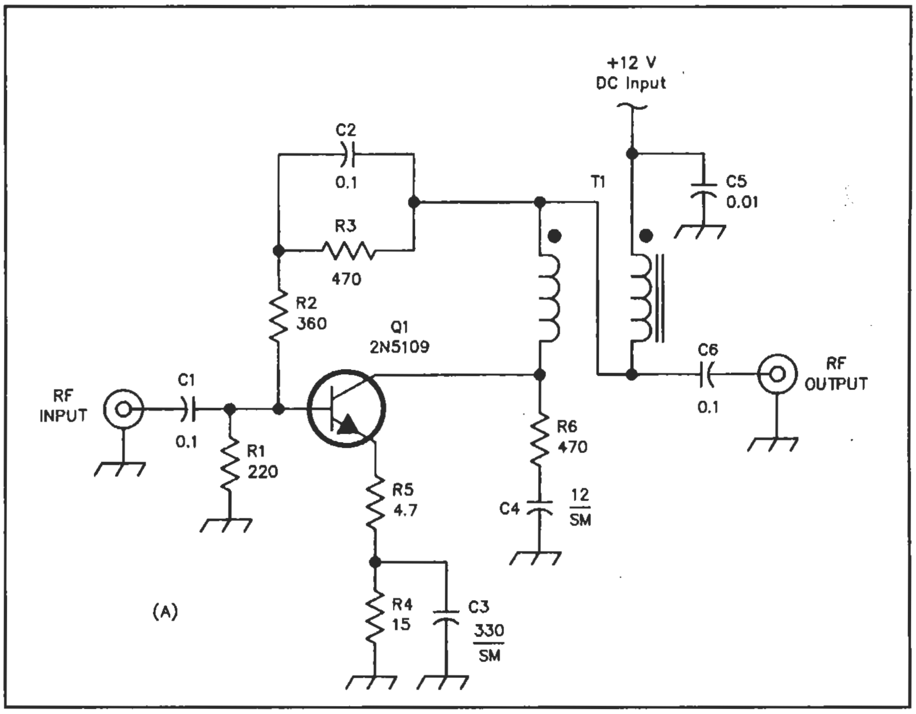

Fig 1a—Low-level amplifier designed to compensate for the gain rolloff from the power amplifier.
Q1—2N5109, 2.5-W heat-sinked RF transistor, f_T is 1200 MHz.
T1—15 turns bifilar #28 on FT-37-43 toroid core.

Fig 1b—5-W TMOS power amplifier.
L1—26 turns no. 26 enameled wire on T-44-2 toroid, 3.9 µh.
Q2—MRF 137 transistor.
R9—10-kΩ turn potentiometer for bias setting.
RFC1—21 turns of no. 26 enameled wire on FR-37-67 toriod.
T2—4 turns 25-Ω coax on FT-50-43 toroid core. The 25-Ω coax is actually two 50-Ω coax run side-byside. The prototype used RG-196/U coax.
U1—78L05 5-V regulator.

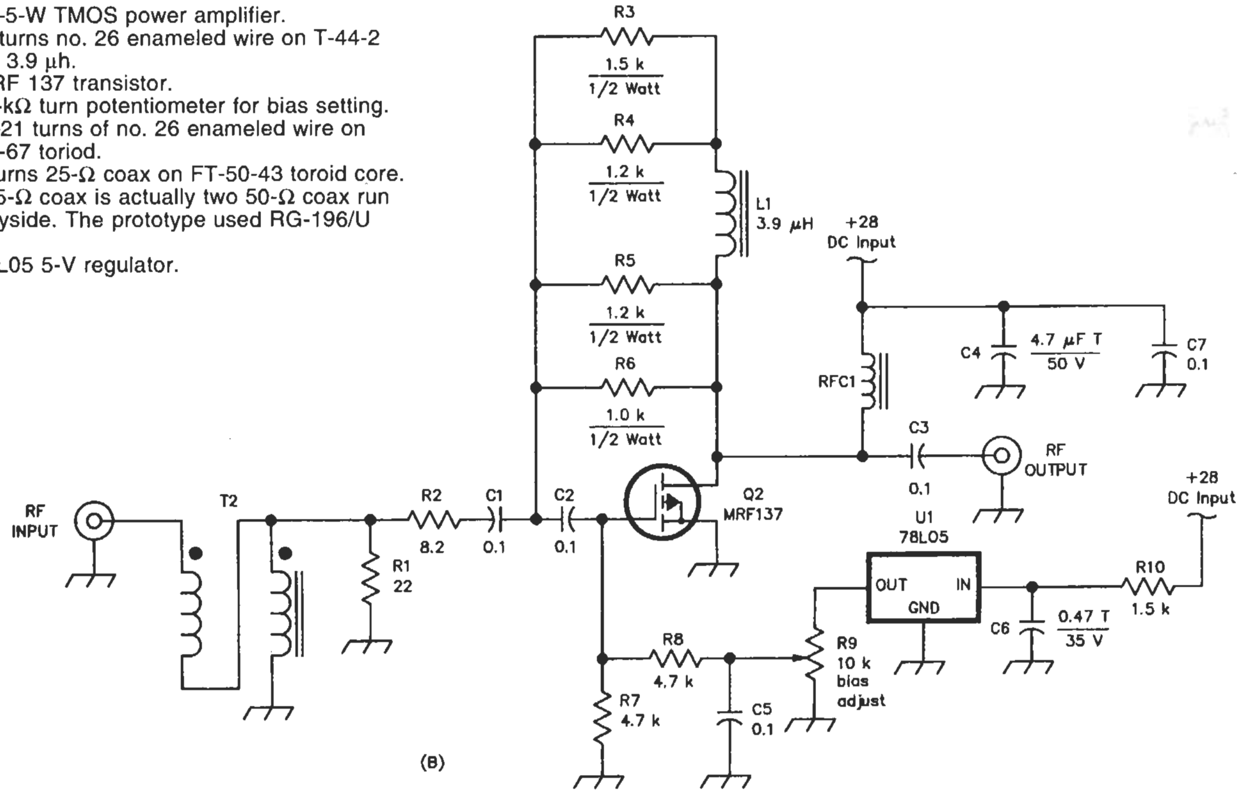
R3
1.5 k
1/2 Watt
R4
1.2 k
1/2 Watt
L1
3.9 µH
+28
DC Input
R5
1.2 k
1/2 Watt
R6
1.0 k
1/2 Watt
C4
4.7 µF T
50 V
C7
0.1
RFC1
C3
0.1
RF
OUTPUT
RF
INPUT
T2
R2
8.2
C1
0.1
C2
0.1
Q2
MRF137
R1
22
U1
78L05
OUT
IN
GND
+28
DC Input
R10
1.5 k
C6
0.47 T
35 V
R8
4.7 k
R9
10 k
bias
adjust
R7
4.7 k
C5
0.1
(B)

By Joel Paladino, N6AMG **From *QST*, September 1992**

An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz

Many kilowatt amplifiers could anchor a small boat, and don't cover 6 meters. Combined with its power supply, *this* kilowatt weighs less than 35 pounds.

This article describes progress toward achieving a well-defined goal: building the smallest possible MF/HF/VHF amplifier capable of at least 1000 watts output. Nicknamed the Solid State Kilowatt (SSKW), the project has its roots in an article by Helge Granberg of Motorola in October 1986 *RF Design*.[1] Two water-cooled amplifiers of that design, built by Mike Staal (K6MYC) of M2, had many problems. Some of them were device-related; others, power-supply related. (More about these issues later.) In February 1990 I rebuilt one of the water-cooled units using the old design and two new transistors. Optimized for 50 MHz, this amplifier could just reach the 1000-watt level. After using it for three weeks of South Pacific DXpeditioning, I decided to construct an air-cooled unit. I spent the spring, summer and fall of 1990 building the SSKW and getting it ready for another DXpedition in the fall.

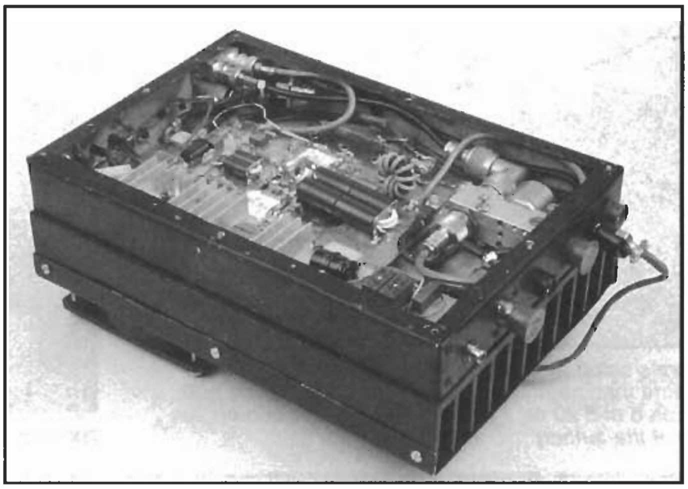

Circuit Description

Motorola's article reprint AR-347 describes the basic amplifier design. July 1990 QEX also carried an article about it.[2] Fig 1 shows the amplifier schematic. Briefly, the circuit consists of two MRF154 RF power MOSFETs in push-pull. Each of these transistors is capable of 600 watts output up to 100 MHz. The input and output transformers, 9:1 and 1:9, respectively, use cores of #67 ferrite material. The output transformer's 1:9 ratio is a compromise that is optimum at about 800 watts.

Incorporating the Motorola building block into a DXpedition-ready package required experimentation and problem-solving, as I described at the 1991 Central States VHF Conference.[3] Here's where the system stands today.

TR Switching

The SSKW includes TR relays that bypass the Fig 1 circuitry in receive mode. I designed their control circuitry to eliminate the possibility of the high RF fields in the amplifier compartment causing relay falsing problems, and to simplifies field repair by keeping parts count low. The relays are sequenced to allow the output relay to close before the amplifier puts out power. Because relays take a few milliseconds to operate, sequencing is necessary to keep the relay from hot-switching the amplifier output. (Hot-switching 1 kW will destroy a relay rapidly!)

This design has two small problems. For the short time it takes the amplifier's input relay to close on switching from receive to transmit, the exciter operates without a load. This can be remedied by delaying the exciter keying. The other problem is the relay's closure time changes as the relay heats. This can be solved two ways: Use external electronics to do the delay, or temperature-compensate for the resistance change.

SWR Protection

I haven't yet built SWR protection into the amplifier. In January 1983 *QST*, Helge Granberg described an SWR-protection circuit for a 2- to 30-MHz amplifier.[4] A phone conversation with Helge indicated that this circuit will work to 54 MHz, so it should suffice. How much mismatch the MRF154s can tolerate is unknown, however, so I don't

6-Meters: DXpeditioning with a Difference

Making the most of 6 meter DXpeditioning depends on knowing where and when to go If you want to work Europe from the Canary Islands summertime is best because sporadic E propagation is best then If you want to work the US from EA8 go when fall shifts into winter to take advantage of east/west F2 layer propagation It s always more exciting to go to a rare place but not necessarily more fun'

Once you ye decided on a DXpedition site plan your travel on the assumption that you ll want to stay on site for at least two weeks This gives you about a 50% better chance of having good openings within a given month

Pay particular attention to the RF prospects of your accommodations Ideally you d be able to see ocean in all directions If that s not possible all s not lost A direct oceanic path isn t always the best propagation may be better over the backscatter path You ll want to get your antenna up as high as possible to minimize radiation angle and AFI And consider EMI and NI possibilities carefully Some areas of the world use NTSC channel 2; others, PAL on channel 1—right in the middle of the band! If your RF gets into a resort's CATV system you ll drive 60 televisions crazy rather than one or two.

Chapter 6 of *The ARRL Operating Manual* covers DXpedition power, health, licensing and logistical concerns in detail so see that book for more about those topics Ill add one thing though It helps to find a local ham with FAX capability as well as HF who can help you with questions on licensing and living accommodations

This can take 80% of the surprises out of the expedition!

Equipment

What I bring and use on 6-meter DXpeditions is based largely on the experiences of Jim Treybig, W6JKV, in his many years of 6 meter DXpeditioning. The core of the approach is pretty much this: Run as much power as permitted and bring as big an antenna as possible. There are practical limits to this, but let your ingenuity direct your thoughts. I consider 100 W as a minimum, 500 watts good and 1000 watts as optimum. High power lets you take advantage of scatter paths that simply won't work at 100 watts. It's very frustrating to hear a well equipped station that you cannot work!

Which Antenna?

A DXpedition antenna must be compact, light and reproducible The Yagi I use has six elements on a 30 foot boom. Optimized by Brian (K6STI) Beezley's Yagi optimization program *YO* it's fed with a T match and a half wave balun. Its boom folds down to two 40 inch sections, each 2 inches in diameter. Most of the element pieces fit inside the two 40-inch sections. The antenna is light enough to lift with one hand.

A rotatable antenna is *mandatory.* You'll need to turn your antenna to find the best direction of propagation. The antenna mast, 20 feet long overall, consists of 40 inch sections of 2-inch aluminum tubing joined with internal sleeves. The rotator is at the mast bottom and there's a slip ring at the mast top just below the Yagi (Having the rotator at the bottom makes putting the antenna up easier by minimizing weight at the upper end of the mast.)

I use Dacron rope for guying. How much rope should you

The author's setup at CN2JP, Rabat, Morocco. Efficiency counts more than neatness while you're busy making 900 contacts on 6 and 20 on 2-meter EME! *(DXpedition photos courtesy of the author)*

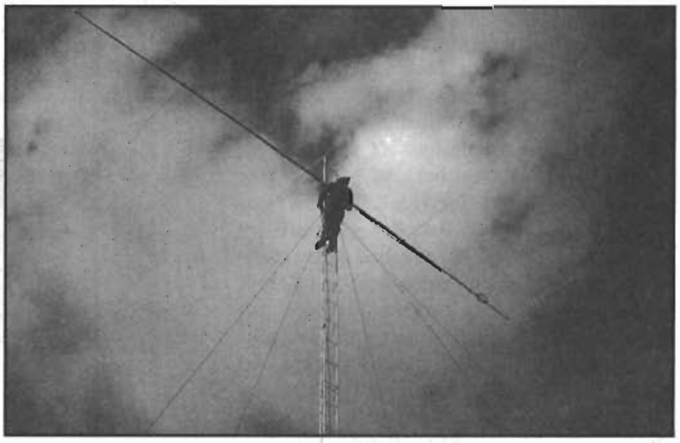

DXpeditioning would be no fun at all without antenna work!

want to use the SSKW to calibrate its own SWR-protection circuitry!

The amplifier's output can be reduced two ways: by turning down the exciter power via amplifier-generated ALC, or by reducing the amplifier bias. (These are enhancement-mode FETs, so positive bias is necessary to turn them on.) The ALC option is better because it does not affect the amplifier's linearity. A no-output-load condition must shut down the amplifier as quickly as possible.

Overdrive Protection

Overdrive can destroy power MOSFETs instantly, so a drive limit control is essential. A threshold detector should be incorporated. The SSKW's input includes a 5-dB attenuator so 100-watt exciters can drive the amplifier without damaging it. If the 5-dB pad fails to attenuate, or if the exciter puts out too much power, the amplifier must be shut down rapidly.

The amplifier must also be shut down if the transistors' flange temperature exceeds 40°C. Shutdown can be done two ways, depending on how the amplifier is used. The first way is to just turn down the bias, which changes the amplifier's linearity. The second way is to totally turn off the amplifier. This must be done gracefully so the amplifier does not shut down while producing full power. The following shutdown sequence is essential: (1) Turn down the bias; (2) open the input relay; and (3) open the output relay.

Bias

The SSKW's bias circuitry is straightforward. R8's value must be tailored to the particular MRF154s used. If the bias decreases too fast with rising temperature, increase the value of R8. The LM723's input voltage must never exceed 40 volts.

Output Filtering

Filtering must be added to make the amplifier comply with FCC signal-purity regulations. The reactance of these filters away from their intended passbands must be taken into account. Absorptive low-pass filters should be considered. Such filters dissipate harmonic energy as heat rather than reflecting it back to the amplifier transistors.

Packaging

I intended to make this amplifier as small and light as possible. Most of the amplifier's weight is its aluminum heat sink take? Simple: Estimate what you need, multiply by 2.5 and you might make it! I usually bring two diameters, $^{3}/_{16}$ and $^{1}/_{8}$ inch.

Radios

Bring at least two radios: one for HF liaison (usually at 10 meters) and another for 6 meters. I recommend maximizing redundancy by taking two radios that both cover 10 and 6. At least one of the 6-meter rigs should include an excellent noise blanker for power-line problems and motorbikes. It must also be able to drive a kilowatt amplifier to full output on 6. (I use one Kenwood TS-680S and one ICOM IC-575.) Receive coverage between 10 and 6 meters is another plus. Where it's allowed, listening to nonamateur services between 45 and 50 MHz can be very important in determining the direction of propagation.

Additional Gear

Bring 12-volt supplies capable of operating at the line voltage and frequency at your DXpedition destination. The supplies should also be able to handle wide line-voltage variations around nominal. Carry two of them—one for each radio. I modify mine by adding line filtering and fusing. Parallel a standard receptacle of some kind with your supplies' rig plug or cable. Two-prong, polarized Jones plugs work well. You can use them on everything for quick setup and breakdown.

Bring a memory keyer for use as a beacon and in normal CW operation. Also consider taking a laptop computer. I now use a computer for logging, taking notes, calculating beam headings, and determining the footprint of the sun.

Operating

Several operating subjects are of prime importance. Ten-meter liaison comes first. You must use 10 meters to get up-to-date propagation information and find out who's being heard. When you're feeling your way through a pile of stations, you can sometimes miss very rare propagation opportunities if you don't check 10 meters often.

Six-meter contact procedures come next. It's very difficult making everyone happy on 6 meters when you're the DX. You must set rules and stick to them. Pick a frequency and do your best to keep it when the band gets busy. Try to not work the same stations over and over again, especially during pileups. Your goal is to give as many people as possible a new country. Work each station as quickly as possible. Don't discuss names and grid squares—the time it takes just gives fewer people a chance to work you. Leave that info to the QSL cards! Which should you operate—phone or CW? The answer is "Probably both." CW penetrates weak conditions better than phone, and it gives more people at the edge of propagation a chance at working you. Then, when you go to phone, everyone who can will worK you again. I favor CW DXpedition operation because of my experiences on the non-DX end.

DXpeditioning is also about meeting people. Here's Tarik Skiredj, CN8ST, and friends.

Should you operate on a single frequency or split? Go to split when there's a massive opening and you're being QRMed by people calling you. This usually happens during intense backscatter openings. Hams on the 10-meter liaison frequency can tell you when QRM gets too heavy.

Conclusion

Six-meter DXpeditioning is a blast. And the SSKW makes it even more fun: I can devote more weight to antennas!—*N6AMG*

and copper heat spreader. How much heat sinking is needed depends on the duty cycle required. I decided that a duty cycle of 50% or less was acceptable. The resulting weight for the package, 14 pounds, is very acceptable. The SSKW is 8 inches wide, 12 inches long and 5 inches high.

The key to making the package light was to use the smallest possible heat spreader. I decided to use enough ¼-inch-thick copper to surround each MRF154 with at least 1 inch of spreader. I machined the spreader flat and attached it to the heat sink, also machined flat. After attaching the copper to the aluminum, I resurfaced the copper again at the transistor contact points to insure flatness. I spread a thin layer of heat-sink compound between the copper and aluminum, and another layer between the copper and the transistors.

With two high-pressure, 24-volt fans blowing on the sink, the resultant duty cycle limit is a little less than 50%. These fans are in series across the amplifier's 50-volt supply. The small fan in the compartment is a 12-volt unit that just barely fits. (The compartment needs to be a little deeper.) A fan is necessary here, however, because compartment airflow dramatically keeps the output transformer's ferrite from getting hot.

A heat sink with more fins and a thicker base, but the same fin depth, would raise the duty-cycle limit. (The copper spreader should be made larger as well.) These changes would also make the amplifier heavier.

Construction Techniques

A milling machine is required for surfacing all of the amplifier's heat-conductive interfaces. Physical flatness is essential for maximum heat conductivity and to avoid warping the transistor cases. The Motorola bulletin specifies the torque of the screws on the MRF154s' flanges. Tightening these screws uniformly assures that the flanges won't warp during temperature cycling.

R14 and R15, the feedback resistors, are flange-mounted. Their beryllium copper leads cannot be flexed very many times before they break. (According to their manufacturer, that may be only once!) These two resistors just barely fit. So, make sure the two MRF154s are separated properly.

Power Supply

To keep the amplifier small and light, I use a switching power supply. A Lambda

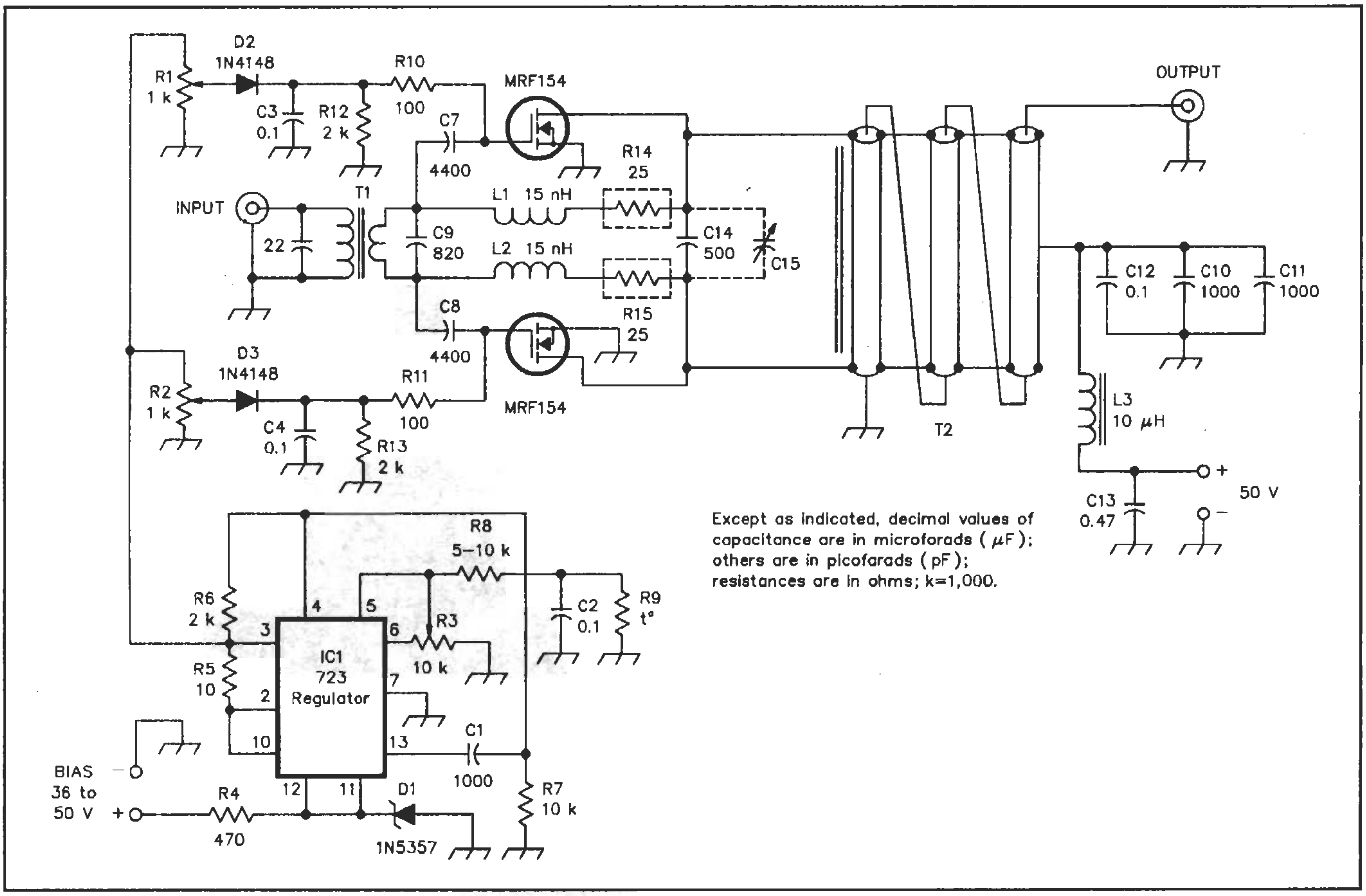
D2
1N4148
R1
1 k
R10
100
MRF154
OUTPUT
C3
0.1
R12
2 k
C7
4400
R14
25
INPUT
T1
L1 15 nH
22
C9
820
L2 15 nH
C14
500
C15
C12
0.1
C10
1000
C11
1000
C8
4400
R15
25
D3
1N4148
R11
100
MRF154
R2
1 k
C4
0.1
R13
2 k
T2
L3
10 µH
+
50 V
−
C13
0.47
R8
5–10 k
R6
2 k
IC1
723
Regulator
R3
10 k
C2
0.1
R9
t°
R5
10
C1
1000
BIAS
36 to
50 V
R4
470
D1
1N5357
R7
10 k

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.

Fig 1—The basic SSKW circuit can produce at least 1 kW from 2 to 54 MHz. Duplicating this circuit requires additional information not given here. Output filtering is also necessary to ensure compliance with FCC emission-purity rules. See the text, Motorola Application Note AN-287, July 1990 *QEX* and the *1991 Central States VHF Conference Proceedings* for details. The *Central States* write-up includes SSKW performance graphs and spectrograms in addition to a diagram of the amplifier's relay-control circuitry.

LFS-50-48 (net weight, 20 pounds) does the job, with good results. Sized at 15 inches long, 7.5 inches wide and 5 inches high, it can source 50 amperes at 50 volts. This power supply produces significant radio noise up to 30 MHz, most of it radiating from the supply's ac-line leads. I EMI-filtered the supply's leads with good success. (We need better, RF-quieter switching supplies!) Because linear supplies generate little or no RF EMI, they are superior to switching supplies for fixed-station use.

Overshoot is a regulated power supply characteristic that's particularly troublesome in switching supplies. Overshoot occurs when a regulator responds too slowly to keep its output voltage down in response to short-duration, high-current loading. In a 50-volt switching supply, and depending on the load, overshoot transients of more than *100 volts* may result. I've seen overshoot destroy expensive transistors! The SSKW supply must be able to safely handle the variable power-supply demand that occurs during SSB and CW transmission.

Power-supply RF sensitivity is another consideration. Some power supplies are sensitive to RF. Their output voltages may vary with the presence and amplitude of RF on their input and output leads. The SSKW's supply must be free of such effects.

Hints and Kinks

Be careful when applying dc to the amplifier for the first time. Do not use a high-current supply for initial tests. Use a

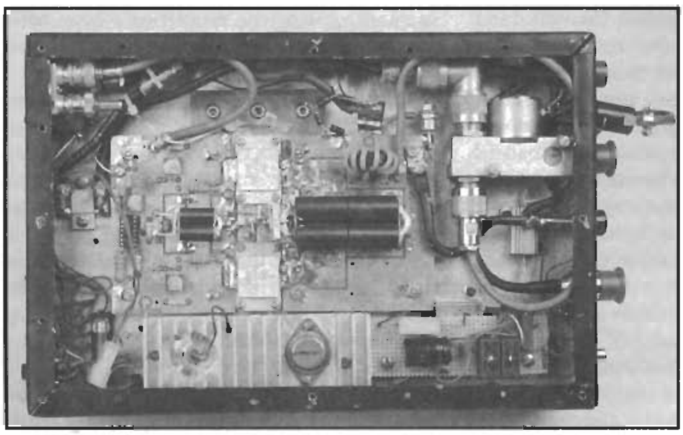

Two Motorola MRF154 RF power MOSFETS in push-pull make the SSKW perk. A 20-watt wxciter drives the amplifier to full output. (amplifier photos by Kirk Kleinschmidt, NT0Z)

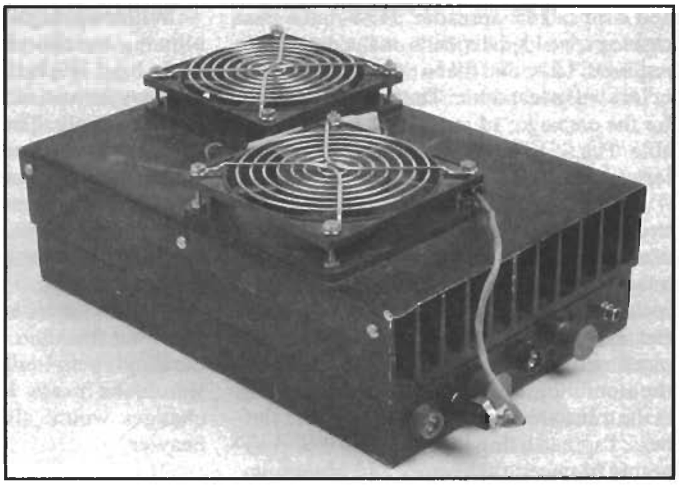

Without its two 24-volt fans, the SSKW's heat sink and spreader would have to be much larger.

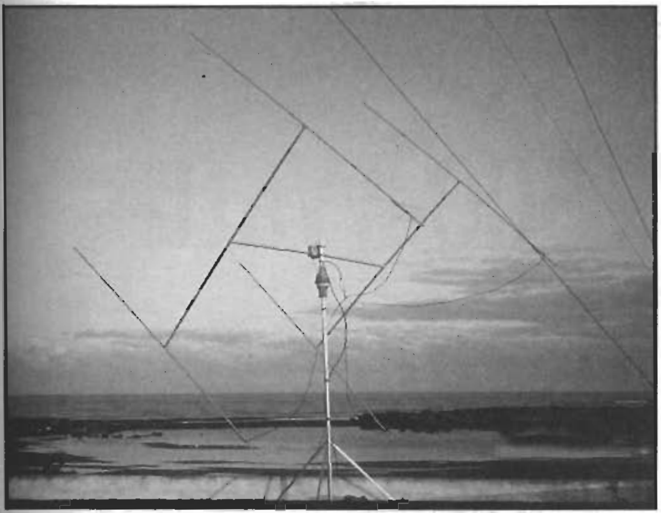

What luggage weight the SSKW amplifier conserves can be put to other uses—2-meter moonbounce gear, for instance. Operating as CN2JP, the author completed 20 EME contacts—Morocco's first—in November 1991. On the same trip, the SSKW performed like a heavyweight through 900 contacts with 56 countries on 6 meters.

Table 1
SSKW Performance Versus Frequency

Freq (MHZ)	*Drive Power (W)*	*Power Output (W)*	I_D *(A)*	*Gain (dB)*	*Efficiency (%)*	*Input SWR*	V_{cc} *(V)*
2	24.5	1000	39.1	16.1	51.2	1.41	50
4	21.0	1000	33.4	16.8	59.9	1.39	50
8	17	1000	32.4	16.8	59.9	1.39	50
15	15.5	1000	32.3	17.7	61.9	1.36	50
30	19.5	1000	41.5	17.1	48.2	1.87	50
50	29	1000	39.7	15.4	50.4	1.51	50

current-limited, 3- or 4-ampere supply to check the regulator and set the bias. The amplifier should have a reverse polarity protection diode rated at a voltage appropriate to the supply.

Make sure that you have a 50-ohm load on the amplifier output. The bias circuitry may not act correctly if there is no load. Slowly increase drive. Watch the output power and the current drain to make sure they are in line with efficiency. If the input match is bad, check C7 and C8. If the amplifier does not achieve its efficiency capability, RF that doesn't make it to the load may try to leave via the dc input line. This can cause C12 to explode off of the underside of the board. Also, under normal operation, the single unit Motorola specifies for C12 can barely handle the RF current that passes through it. I recommend paralleling 100-volt, 0.05- or 0.1-µF chip capacitors instead of using a single 0.1-µF chip.[5] This problem is difficult to analyze because C12 is on the underside of the board.

As mentioned earlier, the SSKW's bias circuitry may have to be tailored to the particular MRF154s used. One thing not mentioned on the schematics is the addition of an Arco 365 variable capacitor (C15 in Fig 1) across C14 to cancel some of the inductive reactance in the output transformer at 6 meters.

Performance

With a 50-volt power supply, the SSKW can produce over 1.1 kW from 2 to 54 MHz. At this power level, the MOSFETs' drain current runs at around 40 amperes, depending on the operating frequency. Table 1 shows amplifier performance data taken at six different frequencies. The SSKW can produce up to 1.5 kW below 30 MHz.

Especially below 30 MHz, the SSKW's harmonic output rises with output power. It may be possible to use only four low-pass filters to cover the 2 to 54 MHz range if a second harmonic of –40 dB can be tolerated. Since the transistors operate in push-pull, the second harmonic is usually not a problem.

Without its 5-dB input pad and operating with a 50-volt supply, the amplifier exhibits an input SWR of less than 2:1 (referred to 50 ohms) throughout its frequency range. Only below 5 MHz does its input SWR exceed 1.6:1. Inserting the 5-dB pad adds 10 dB of return loss and keeps the input SWR below 1.3:1 through the amplifier's operating range. I have not yet measured the amplifier's two-tone, thirdorder IMD performance.

Final Notes

For now, I can say that my goal of a compact, lightweight amplifier capable of at least 1 kW output is a reality. The SSKW has served me well in its intended DXpeditionary application. Nonetheless, the amplifier in its present form is still experimental. Before the SSKW could be acceptable for general use, output filtering and failureproof protection circuitry would have to be added. The next generation is on its way. The next project? Use the same 50-volt supply with power FETs capable of covering 144 to 432 MHz, and build it into a similar package!

Acknowledgment

I thank Helge Granberg for spending considerable time answering questions about this project.

Notes

[1] H. Granberg, "New MOSFETs Simplify High Power RF Amplifier Design," RF Design, Oct 1986, pp 43-48, 50, 52.

[2] H. Granberg, "A Compact 1-kW 2-50 MHz SolidState Amplifier," QEX, Jul 1990, pp 3-8.

[3] J. Paladino, "A Rapid Deployment Prototype Solid-State Kilowatt Linear Amplifier for 2 to 50 MHz Using MRF-154s," *1991 Central States VHF Conference Proceedings* (Newington: ARRL, 1991), pp 61-70. Available from the ARRL Bookshelf as #3614.

[4] H. Granberg, "MOSFET RF Power—An Update," Part 2, *QST*, Jan 1983, pp 30-33; also see Feedback, *QST*, Mar 1983, p 44. (Part 1 appeared in *QST*, Dec 1982, pp 13-16; also see Feedback, *QST*, Jan 1983, p 48.)

5American Technical Ceramics (ATC) has some new chip caps (the 900 series) that have higher voltage and current ratings than the capacitor Motorola specifies. I have samples but have not tried them yet.

By Jerry Pittenger, K8RA **From *QST*, September 1986**

An All-Band, 1500-Watt-Output 8877 Linear Amplifier

Part 1—This rock crusher, rated for continuous full-legal-limit output, can be built at home. It is, however, a major project requiring dedication and commitment.

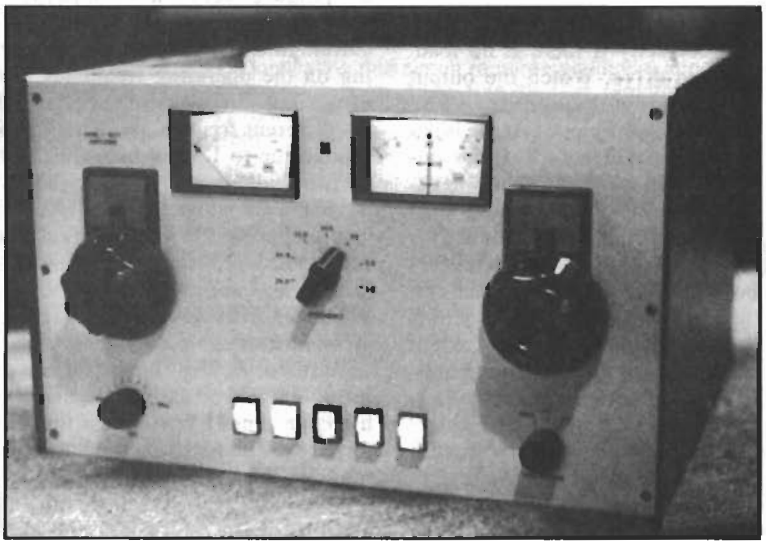

This article is the result of a 10-month project to build a legal-limit linear amplifier. The amplifier uses the popular EIMAC 8877 (3CX1500A7) high-mu power triode that can provide a continuous RF output of 1500 W to the antenna.

In recent years, I have built several different linear amplifiers, and I must admit that this previous experience was necessary to obtain the results achieved with this project.[1-4] I hope that by sharing this experience, others will benefit from it. Any amplifier design depends on the various components used and individual preferences. Therefore, you may not want—or be able—to duplicate this amplifier exactly.

The comment I receive most often from the amateur fraternity is about the high cost to build an amplifier like this. The criticism is valid. This amplifier is not inexpensive to build. Plan to spend from $1000 to $1200 for the RF deck, and another $500 to $600 on the power supply. If you really think about it, though, these costs are a bargain when you consider the performance and quality of the final product and the cost of an equivalent commercial unit.

This article is presented in two parts. In this part, I will describe the RF deck and power supply in general terms. Schematic diagrams and parts considerations are included. Part 2 gives detailed instructions for constructing the two units and considerations for the final testing and operation.

Preliminary Thoughts

Finding Parts

Finding parts can be a big task. Even the most difficult parts to find, such as the vacuum variable capacitors, vacuum relays and door-knob capacitors are available, however, and appear for sale in the ads (*QST* Ham-ads and the Yellow Sheets), or at hamfests and flea markets.[5] Probably the best source of parts is other hams who are actively building equipment. Go talk to these people and let them know what you are looking for. It's amazing how others will help, and even let you into their personal stores. There are people, like myself, who like to build amplifiers. Once you learn who these individuals are, keep in touch with them. They can help find the key parts.

Parts that are not available in the surplus market can be purchased new. This will be necessary for some parts, such as

Table 1
Recommended Tools

- Drill press or drill fixture (with set of highspeed bits)
- Band saw capable of cutting 1/16-inch-thick metal
- Chassis punches (5/8 inch to 1 inch)
- Fly cutter, 2-inch radius
- Vise
- Set of taps
- Common handtools (screwdrivers, pliers, soldering iron and gun)
- Volt-ohmmeter
- Variable power supply (5-26 V, 1 A)
- Dip oscillator

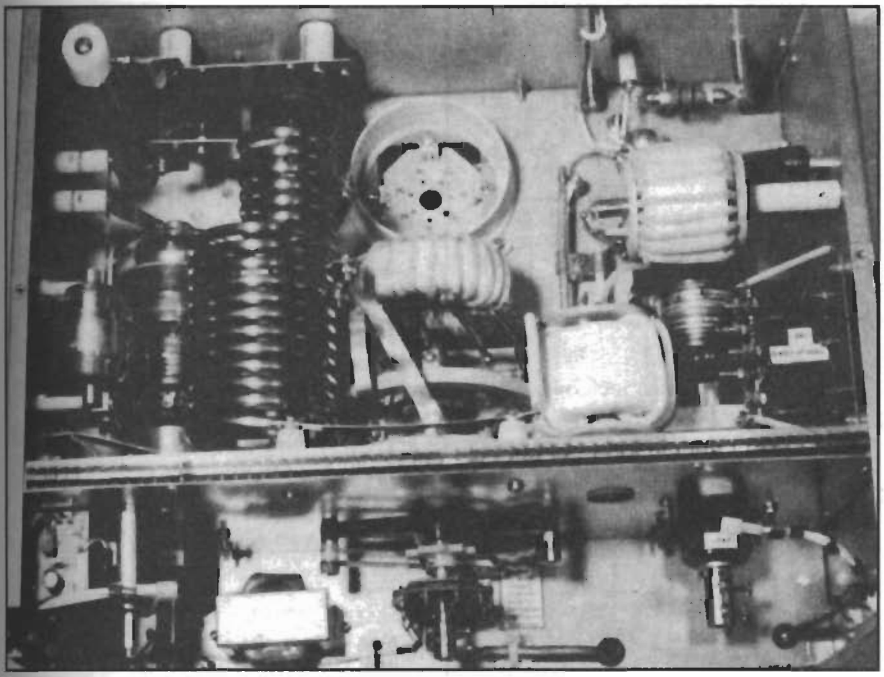

Fig 1—Top interior view of the 8877 linear amplifier RF deck.

cabinets, etc. Just remember that when you buy a new commercial amplifier, you pay the new price for every component.

Tools

A good assortment of hand tools, as well as some power tools, are necessary to complete this project. Table 1 shows the tools I recommend. In particular, I recommend that a drill press and band saw be available. You can do the job without all of the tools listed, but the job will be much more difficult.

Time

Time is probably the most valuable resource for most of us, and the one that may prove hardest to find. This project took well over 250 hours to complete. The key is to do each step right, and not hurry. Build the amplifier in a place where you can leave the project on the table and walk away. Plan each step and build in discrete modules. Work an hour or so whenever possible, and slowly, but surely, the modules will take shape. It is amazing how much you can accomplish using these small time segments. Also, great strides can be made on a Saturday or a Sunday. Commitment and consistency are the virtues required to finish the job.

RF Deck Circuit Description

The RF deck is designed to be a tabletop unit (see title photo). The power supply is remotely controlled and can be located almost anywhere. The amplifier design is based on proven circuitry. Included are all circuits required to provide a clean signal as well as adequate protection devices for the metal-ceramic 8877 tube.

Control Circuitry

Fig 2 shows the schematic diagram for the amplifier control circuitry and low-voltage power supply. The 117-V ac input from the high-voltage power supply enters the RF deck through a 5-conductor interconnecting control cable. Each control line is terminated in a pi-section filter as it enters the RF deck, to prevent RF from getting into the control cable and power supply. The pisection filters are constructed as an independent module.

The amplifier is powered up by the FIL ON/OFF switch, S1. Engaging S1 turns on the blower, filament power and 26-V dc power supply. The current inrush to the tube is limited by Ri, in series with the filament transformer primary. After approximately 1 second, K1 energizes and K1A shorts Ri thus providing full filament voltage to the tube. The K1 delay is controlled by R2 and C1 across the relay coil. R3, in series with the other leg of the filament transformer primary, is adjusted to provide the proper filament voltage (4.85 V ac) to the tube under load.

The 8877 requires a 3-minute warmup period to reach proper operating temperature. A solid-state timing circuit, formed by Q1 and Q2, locks the amplifier out of operation until the warmup period has elapsed. When the 26 V dc comes on, C2 charges through the 500-kilohm time-delay adjust and 1.2-megohm resistors. QI and Q2 form a high-impedance Darlington circuit, and the emitter of Q2 follows the voltage rise on C2. The high-impedance Darlington circuit is required to prevent the capacitor charge from draining through the transistors. After approximately three minutes, the potential at the emitter of Q2 reaches 18 V at which point the 4PDT relay, K2, engages. K2A applies 26 V dc to the K2 relay coil, removing the relay current load from Q2. The voltage also turns on the TIME pilot light located on the amplifier front panel to indicate that the warmup period is over. The same line also applies 26 V dc to S2B of the HV-ON push-button switch, which, when engaged, sends 26 V dc to the RF input/output relay circuits. K2B connects a 100-kilohm resistor across C2 to drain the charge from C2. This resets the 3-minute timer should the amplifier be turned off and immediately back on.

K2C and K2D are wired in parallel and apply 117 V ac to HV-ON switch S2A to energize the high-voltage power supply. The high-voltage power supply can't be turned on even if the HV-ON switch is engaged until after the 3-minute warmup period has ended. IN/OUT switch S3 allows the amplifier to be put in the standby mode with the amplifier turned on. Both HV-ON and IN/OUT front-panel push-button switches must be engaged to key the amplifier, thereby making it impossible to operate the amplifier without high voltage on the tube.

The amplifier is keyed by grounding the base of Q3 through the exciter TR-relay contact. A transistor is used to limit the current switched by the exciter VOX relay. This avoids a potential problem if the exciter VOX relay sparks on closure, which could damage the relay contacts. The "grid trip" break in the relay line causes the relays to drop out if the grid trip circuit actuates from too much grid current (approximately 120 mA). During normal operation, the grid trip break is shorted by a normally closed set of contacts on K3 (see Fig 3).

When the amplifier is keyed, the output RF relay must be closed before drive is applied to the tube—otherwise the tube will transmit for a brief period without a 50-ohm antenna load. This would not only be harmful to the tube, but also cause the grid-trip circuit to actuate. Therefore, a timing circuit, comprised of a 50-ohm resistor and i00--~iF capacitor, is included across the RF input relay K4 to allow vacuum relay KS time to close. The capacitor value depends on the relay used. Do not make the delay too long, since during the delay time, the exciter does not have a proper 50-ohm load. Check the time delay by placing a low voltage across the relay contacts and monitoring the contact closure on a dual-trace scope. I used a delay of about 20 ms.

RF Amplifier Circuit Design

The RF amplifier circuit is shown in Fig 3. The amplifier uses a tuned input network to minimize distortion products and provide a proper impedance match between the exciter and the tube. The input network is remotely switched, using small DPDT relays, to connect the correct pi-section for the selected band. A homemade switch deck is mounted on the band-switch shaft, in front of the subpanel, to ground the 12-V dc line for the proper input relay as selected by the main band switch. On 160 meters, the switch also controls a solenoid relay to add a 160-pF capacitance in parallel with the TUNE vacuum variable capacitor.

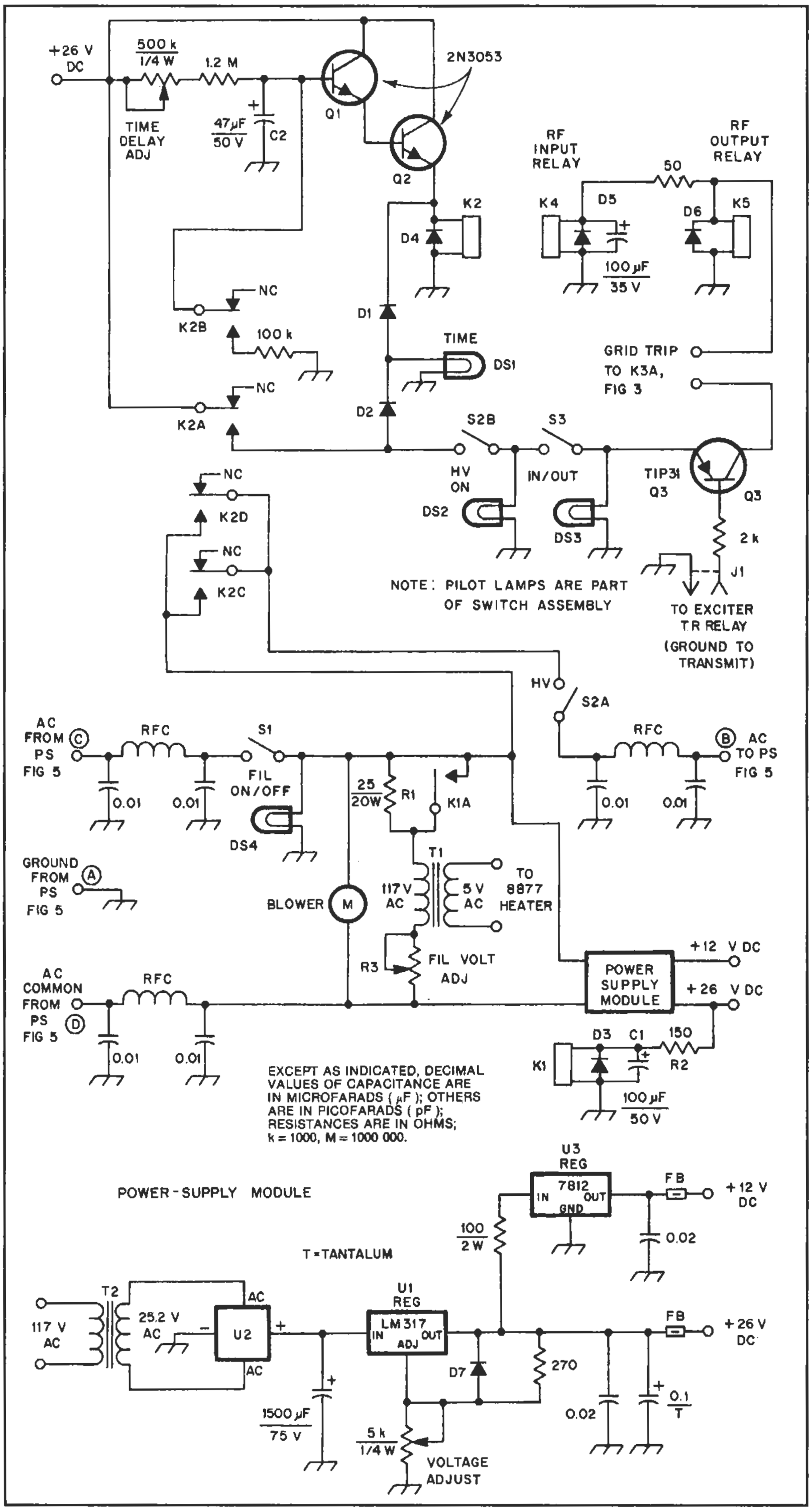
+26 V DC
500 k 1/4 W
1.2 M
TIME DELAY ADJ
47µF 50 V
C2
Q1
2N3053
Q2
K2
D4
D1
TIME
DS1
D2
K2B
NC
100 k
K2A
NC
RF INPUT RELAY
RF OUTPUT RELAY
50
K4
D5
D6
K5
100µF 35 V
GRID TRIP TO K3A, FIG 3
S2B
S3
HV ON
IN/OUT
DS2
DS3
TIP31 Q3
Q3
2 k
J1
TO EXCITER TR RELAY (GROUND TO TRANSMIT)
NOTE: PILOT LAMPS ARE PART OF SWITCH ASSEMBLY
K2D
K2C
NC
HV
S2A
AC FROM © PS FIG 5
RFC
S1
0.01
FIL ON/OFF
DS4
25 20W
R1
K1A
T1
117V AC
5 V AC
TO 8877 HEATER
BLOWER
M
R3
FIL VOLT ADJ
GROUND FROM Ⓐ PS FIG 5
AC COMMON FROM Ⓓ PS FIG 5
Ⓑ AC TO PS FIG 5
POWER SUPPLY MODULE
+12 V DC
+26 V DC
D3
C1
150
R2
K1
100µF 50 V
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF); RESISTANCES ARE IN OHMS; k = 1000, M = 1000 000.
POWER-SUPPLY MODULE
U3 REG
7812
IN OUT GND
FB
+12 V DC
0.02
100 2 W
T = TANTALUM
T2
117 V AC
25.2 V AC
U2
AC
U1 REG
LM317
IN OUT ADJ
D7
270
1500µF 75 V
5 k 1/4 W
VOLTAGE ADJUST
+26 V DC
0.02
0.1 T

Fig 2—Amplifier control circuit and low-voltage power-supply schematic diagram. Part numbers shown in parentheses are Radio Shack.

B—Blower, Dayton 4C004-i.
D1-D6—Diode, 1 kV, 2.5 A.
FB—Ferrite bead.
K1-K4—4PDT 24-V dc relay, Potter & Brumfield KHU17D11.
K5—SPDT vacuum relay, 26-V dc coil.
Q1,Q2—2N3053 NPN transistor.
Q3—TIP31 NPN transistor (276-2017).
R1—25 Ω, 20 W.
R2—150 Ω, 2 W.
R3—25 Ω, 25 W variable.
RFC—10 turns no. 14 enam wire on ¼-in-diam ferrite rod.
S1,S3—Alco 16TL5-11 SPST.
S2—Alco 16TL5-22 DPDT.
S4—Alco 16TZ pilot light.
Ti—Filament transformer, 5.0 V ac, 10 A, Peter Dahl Co.
T2—25.2 V ac, 1.0 A, Stancor P6469.
U1—50-V, 4-A bridge rectifier.

An effective ALC circuit, adjustable from a front-panel control, is included to avoid overdriving the tube. This feature is essential in this amplifier because the drive requirement is only about 80 W for 1500-W output. The ALC circuit samples the RF drive level through a 27-pF mica capacitor to generate a dc voltage that is fed back to the exciter for drive-power control.

The grid-trip-protection circuit shuts down the amplifier if grid current exceeds 120 mA. This protects the tube from tuning errors or other problems such as losing the antenna, or a tube flashover during operation. Although grid current flows through all paths from ground to the B – line, most of the grid current goes through R1. The current passing through R1 develops a voltage drop. For example, if 100 mA of grid current is drawn through Rl, 1 volt is developed (E = IR = 0.100 x 10). This voltage is used to turn on the transistor switch, Qi. When Qi turns on, the grid-trip relay, K3, energizes and opens the grid trip break in the RF relay control line to shut the amplifier down. R2 sets the current level at which Qi turns on. The front-panel GRID TRIP lamp goes out if the trip circuit is activated. The switch is reset by pressing S3.

The plate tank circuit uses a pi-L configuration because this design provides approximately 20-dB better harmonic suppression than the conventional pi design. The TUNE and LOAD capacitors are vacuum variable types to minimize space requirements and also optimize performance on 12 and 10 meters where small capacitance values are needed to achieve an acceptable tank-circuit Q. The 10- to 40-meter tank coil is homemade from ¼-inch copper tubing that is silver plated to minimize skin resistance. The 80-meter, 160-meter and L-coils are toroid designs to minimize space. Using a toroid for the L coil also helps isolate the L network from the rest of the tank circuit because of the toroid's self-shielding characteristics.

Metering circuits monitor plate and grid current, as well as filament voltage. Plate current is monitored by placing a meter in series with the B – line. Therefore, only a small dc voltage is across the meter. An additional position can be included on the FIL/GRID meter for plate voltage, but one is not shown in this design because a separate high-voltage meter is included in the power supply. It would be a good idea to include a high-voltage scale on the meter in case the RF deck is ever used with a different highvoltage supply. Grid current is monitored by measuring the voltage drop across R1. R3 is adjusted to give the cor-

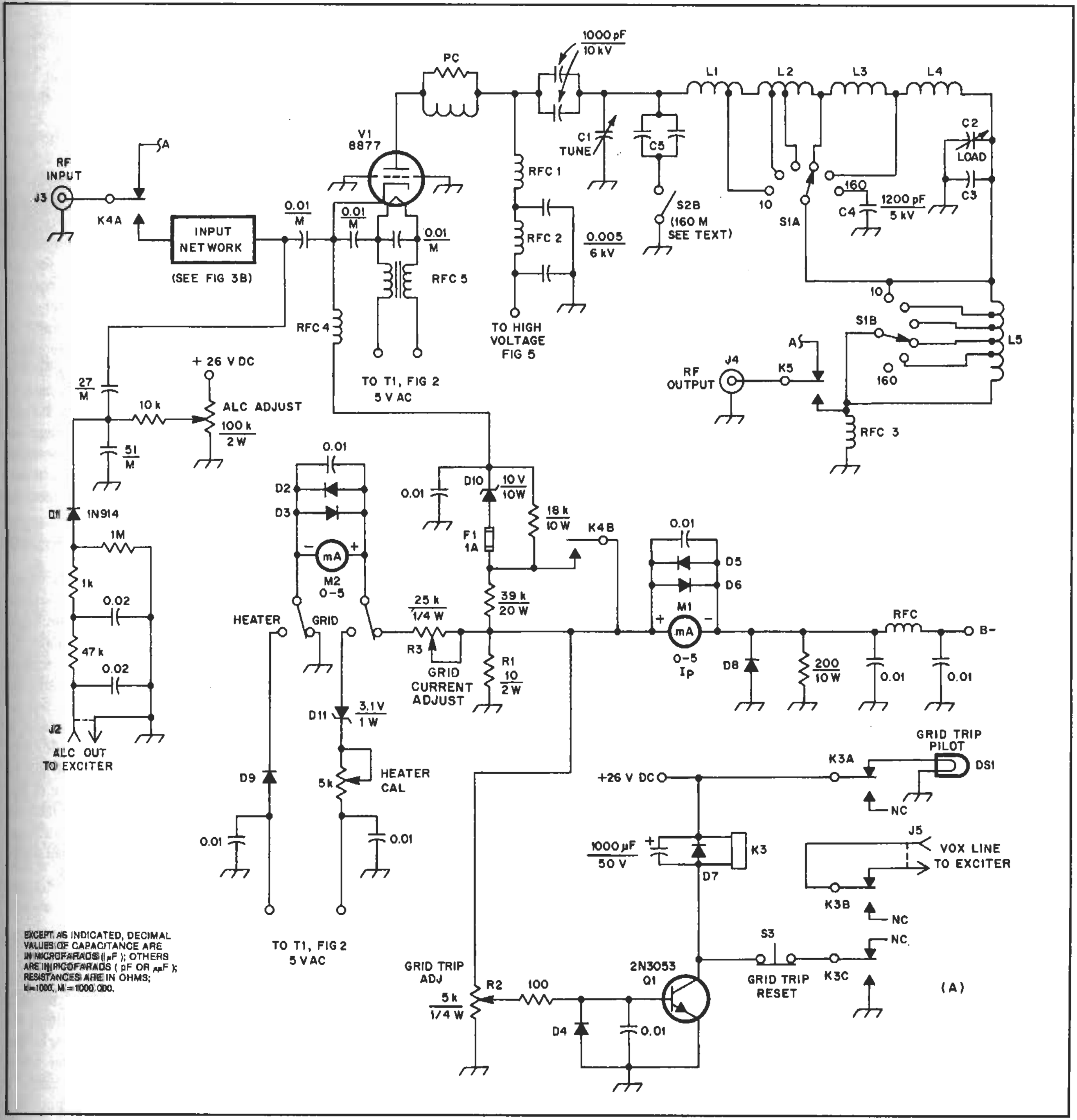
1000 pF
10 kV
PC
L1
L2
L3
L4
V1
8877
C1
TUNE
C5
C2
LOAD
C3
RF
INPUT
J3
A
K4A
INPUT
NETWORK
(SEE FIG 3B)
0.01
M
RFC 1
RFC 2
0.005
6 kV
S2B
(160 M
SEE TEXT)
10
S1A
160
C4
1200 pF
5 kV
RFC 5
RFC 4
TO HIGH
VOLTAGE
FIG 5
S1B
L5
RF
OUTPUT
J4
K5
RFC 3
TO T1, FIG 2
5 V AC
+ 26 V DC
27
M
10 k
ALC ADJUST
100 k
2 W
51
M
D1
1N914
1M
1k
0.02
47 k
J2
ALC OUT
TO EXCITER
D2
D3
M2
0-5
mA
HEATER
GRID
25 k
1/4 W
R3
GRID
CURRENT
ADJUST
D10
10 V
10W
F1
1A
18 k
10 W
39 k
20 W
R1
10
2 W
K4B
M1
D5
D6
0-5
Ip
D8
200
10 W
RFC
B-
D11
3.1 V
1 W
HEATER
CAL
5 k
D9
TO T1, FIG 2
5 VAC
GRID TRIP
PILOT
DS1
K3A
NC
+26 V DC
1000 µF
50 V
K3
D7
J5
VOX LINE
TO EXCITER
K3B
S3
GRID TRIP
RESET
K3C
2N3053
Q1
GRID TRIP
ADJ
R2
100
5 k
1/4 W
D4
(A)

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig 3—RF amplifier schematic diagram (see Part B on next page).

C1—Vacuum variable capacitor, 375 pF, 10 kV.
C2—Vacuum variable capacitor, 1000 pF, 10 kV.
C3—Mica transmitting capacitor, 100 pF, 5 kV.
C4—Mica transmitting capacitor, 3 × 400 pF, 5 kV.
C5—Fixed vacuum capacitor, 2 × 80 pF, 20 kV.
D1—Diode, 600 V, 1 A.
D2-D9—Diode, 1 kV, 2.5 A, HEP 170.
D10—Zener diode, 10 V, 1 W.
D11—Zener diode, 3.1 V, 1 W.
K1-K7—DPDT 12-V dc DIP relay (275-213).
K8—SPST 12-V dc relay (275-241).
RFC—10 turns no. 14 enam wire on 1/4-in-diam ferrite rod.
RFC1—Plate choke, 2 A, Peter Dahl Co.
RFC2—Air-wound coil, 15 turns, 1/2-in diam.
RFC3—Choke, 1 mH, 800 mA.
RFC4—110 turns no. 20 enam wire on 1/2-in diam fiber rod.
RFC5—Filament choke, 18 bifilar turns no. 14 enam wire on 1/2-in-diam ferrite rod, 6 inches long.
PC—Three 150-ohm, 2-W carbon resistors in parallel with 2-inch horseshoe loop of 1/2-inch silver-plated strap.
L1-L5—See Table 2.
M1,M2—Simpson Wide-Vue panel meter, 01253 bezel and 01165 lighting kit (See text).
S1—9-position, 2-pole switch, Radio Switch model 88, 13-kV, 30-A.
S2—Solenoid-controlled switch; see text.
S3—SPST normally closed momentary switch, Alco 16TL-11 with 6T-2 red lens.

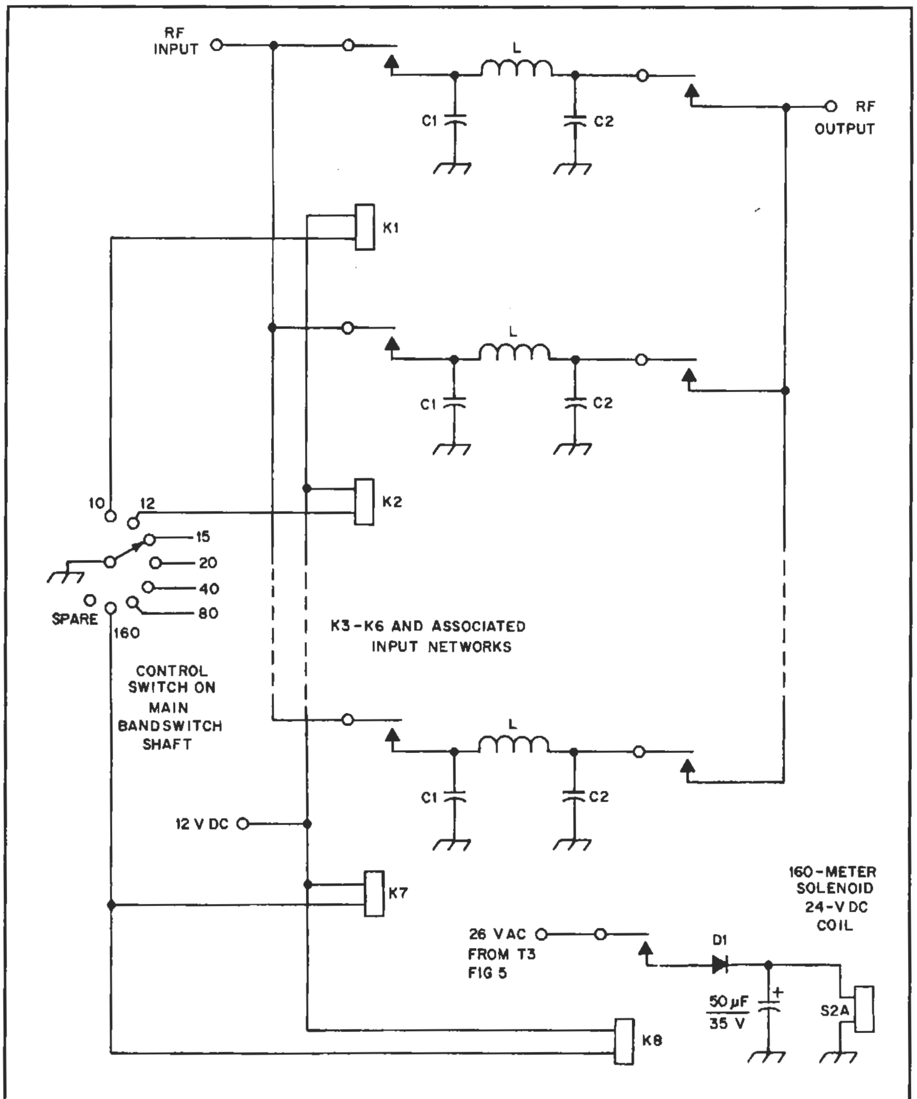
RF INPUT
L
C1
C2
RF OUTPUT
K1
K2
10
12
15
20
40
80
160
SPARE
CONTROL SWITCH ON MAIN BANDSWITCH SHAFT
K3-K6 AND ASSOCIATED INPUT NETWORKS
12 V DC
K7
26 VAC FROM T3 FIG 5
D1
160-METER SOLENOID 24-V DC COIL
50 µF 35 V
S2A
K8

BAND	C1 (pF)	L (µH)	C2 (pF)	Q	CORE	NO. OF TURNS	WIRE SIZE
160	4000	3.31	4000	2.35	T94-2	19	22
80	1240	2.06	1240	1.50	T94-2	15	22
40	680	1.03	680	1.67	T68-2	14	18
20	390	0.50	390	1.80	T68-2	9	18
15	270	0.32	270	1.90	T68-2	7	18
12	270	0.25	270	2.20	T68-2	6	18
10	160	0.27	160	1.50	T68-2	8	18

(B)

Fig 3B—see previous page.

rect gridcurrent meter reading. Filament voltage is measured by converting the ac voltage to dc and displaying the dc voltage on M2. The 3.1-V Zener diode expands the meter scale by not allowing conduction until the voltage reaches 3.1 V.

A vacuum relay is used for the amplifier output. The relay is small in size, quiet and capable of handling large RF currents.

RF Deck Parts Selection

Finding all the parts for the RF deck is a major task. If you are planning to build an amplifier, begin collecting parts as soon as possible. It is the first step because the physical layout of the amplifier will depend on the components available. Don't try to exactly duplicate the components I used. For example, vacuum variable capacitors come in many different shapes and sizes, with different mounting provisions. Actually, the parts you find may be better than the parts used in my RF deck. As an example, a 1500-pF vacuum variable LOAD capacitor would be much better than the 1000-pF unit I used.Therefore, use whatever resources you have available to acquire the parts— but a word of caution! Do not compromise too much when gathering components. If you cannot find what you need on the surplus market, buy the parts new. It may cost a little more, but if the project is not done right, you will never be happy with the final result.

Vacuum Variable Capacitors

Vacuum variable capacitors are often difficult to locate at reasnable prices. Plan to spend about $50 for the TUNE capacitor and $75 to $100 for the LOAD capacitor if vacuum capacitors are used. The TUNE capacitor should be at least 300 pF at 7 kV, and the LOAD capacitor should be at least 1000 pF at 3 kV.

An air variable capacitor can be used for the LOAD control, if desired. The minimum capacitance for the LOAD capacitor is 112 pF for 10 meters, which is not difficult to obtain with an air variable type. A rating of 1 kV, minimum, is recommended. However, it is a different story for the TUNE capacitor. The minimum required capacitance is 26 pF. The direct inter-electrode capacitance of the 8877 tube in grounded-grid service is 10 pF; therefore, the TUNE capacitor must have a minimum value of not more than 16 pF for 10 meters.

This is nearly impossible with a 300-pF air variable. In addition, the voltage requirements for the TUNE capacitor make any air variable rather large. For these reasons, a vacuum variable is recommended for the TUNE capacitor.

Meters

Good-quality meters with bezels are essential for good appearance. The Simpson Wide-Vue® meters I used were purchased at a hamfest. The bezels were ordered directly from Simpson because they seldom appear on the surplus market. Actually, almost any meter movement can be used, so don't pass up a good meter just because it reads 50 V or 100 mA on the scale. Any meter with a movement from 100 µA to 5 mA can be used. This allows use of approximately 90% of the meters available on the surplus market. I will give instructions later for calibrating any meter to read whatever current or voltage is required.

RF Band Switch

Good RF band switches are very difficult to locate. More problems are experienced with arcing band switches than with any other amplifier component. If the band switch selected has insufficient voltage insulation, it will arc to the wiper rotor on the high-impedance 10-meter position when operating on the lower-frequency bands. I obtained the band switch for my amplifier from Radio Switch Corp.[6] The model 88 switch is a 2-pole, 9-position unit with a 13-kV peak flashover/30-A contact rating. This switch will not arc! Its list price is currently $107, and it is well worth the money!

Miscellaneous Parts and Materials

Many of the small parts (capacitors, relays and resistors) can be purchased at Radio Shack. Their parts selection is good, and continues to increase. You can usually find a store around the corner in almost any city. Pioneer Electronics is also a good source for commercial-grade components.[7] Good-quality PC-board material can be found at almost any hamfest. Don't compromise here—use G10 glass-epoxy board. As for coils? Make them. Complete "how-to" instructions are given later.

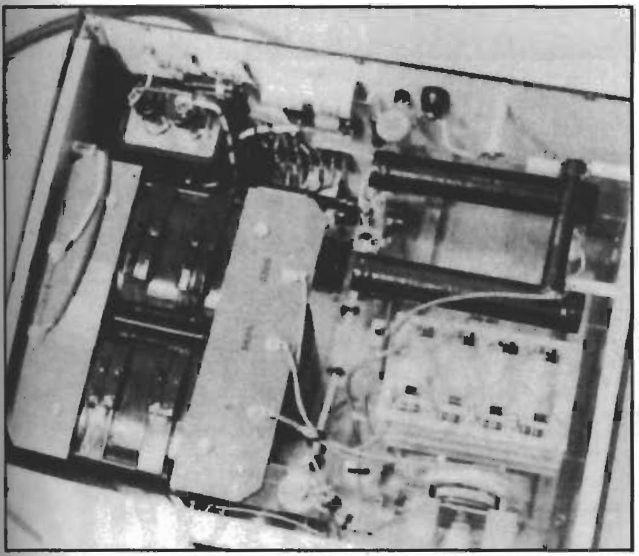

Fig 4—High-voltage power supply top interior view.

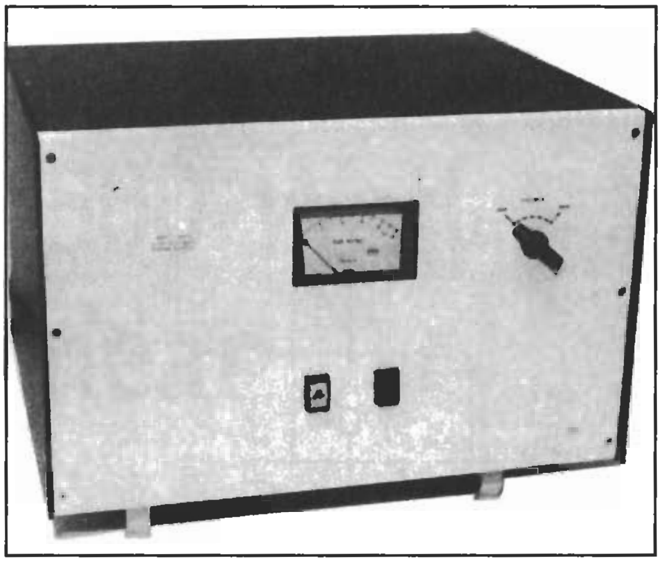

High-Voltage Power-Supply Circuit Description

The key to continuous duty in a high-power linear amplifier is the power supply. It must be able to deliver the required voltage and current on a continuous basis. Power supplies are usually the limiting factor in commercial linear amplifiers.

A WORD OF CAUTION IS IN ORDER. The power supply is a very dangerous piece of equipment! Give it proper respect. One mistake can be fatal. Use proper precautions in the construction and testing of this unit, and be careful to build a safe unit.

I recommend that the power supply be built first. The construction is not complex and it can serve as a training ground for amplifier building techniques, particularly for the first-time builder.

Power-Supply Design

The power supply is shown in Fig 4, and the schematic diagram is shown in Fig 5. The hypersil power transformer has a 234-V ac primary, and a 3300-V ac secondary that is tapped at 2600 V. This selection of two output voltages allows for a high- and lowpower capability. An alternative to this approach is to include a Variac® or Powerstat® autotransformer on the transformer primary.

The primary circuit of the power transformer includes a step-start circuit to protect the diode bank during the initial charge of C1, the 53-µF filter capacitor, when the power supply is turned on. Two 50-ohm, 25-W resistors, one in each leg of the primary, are shorted by time-delayed relays approximately 3 to 4 seconds after application of power. The more current drawn through the resistors at start up, the more voltage drop realized and this, in turn, protects the diode bank. The delay is provided by the time constant of the 500-ohm resistor and 100-µF capacitor. The relays must be dc types. Those I used have 90-V dc coils which allows power to be supplied from one 117-V leg of the primary. If 90-V relays can't be obtained, 24-V dc relays can be substituted. A 24-V dc power source must be provided if this is done.

The rectifier unit is a full-wave bridge with eight diodes in each leg. A 470-kilohm resistor and a 0.01-µF, 1-kV capacitor are wired in parallel with each diode to equalize the voltage and protect the diodes from voltage spikes.

The power supply is controlled remotely from the RF deck. A test switch has been incorporated to allow the supply to be energized without the RF deck. A shorted Cinch-Jones plug must be inserted into a socket in the rear of the supply for test switch S1 to operate.

Two pilot lights are mounted on the front panel. One pilot light is on whenever 234 V ac is present in the supply. The other lights when the power supply is activated.

A high-voltage meter is included on the front panel. The metering is done across a 25-ohm, 5-W resistor in series with the bleeder resistor. This voltage divider keeps the total high voltage off the meter. A 50-ohm, 50-W resistor in series with the high-voltage B+ circuit protects the tube and power supply from any current surge resulting from a tube flashover or other cause. In addition, a 0.6-ohm, 1-W resistor in series with the B+ line acts as a fuse resistor. A large current surge will cause the resistor to explode—an inexpensive protection device should a problem occur.

High-Voltage Power Supply Parts Selection

Transformer

It is important to find a good power transformer that can provide the proper operating voltages for the tube. Remember that some voltage drop will occur when current is drawn from the transformer. The voltage drop depends largely on the quality of the transformer (core and wire size), and can range from 200 V to over 1 kV. The transformer should have a 234-V ac primary. Transformers with 117-V primaries are usable only if two identical units can be wired in series to provide a 234-V primary. The secondaries can by wired in series or parallel, depending on the voltage requirements. Remember that the transformers must be identical.

The required transformer secondary voltage depends on the fmal voltage requirement of the tube and the power-supply circuitry. If a bridge rectifier is used, the power-supply high voltage will be about 1.4 times the secondary voltage. If a voltage doubler is used, the high voltage will be about 2.8 times the secondary voltage. A voltage doubler requires two filter capacitors, or more, so if a single oil-filled filter capacitor is to be used, the design can't be a voltage doubler. The *ARRL Handbook* contains circuits for both types of power supplies.[8]

The power-handling capability of a transformer can usually be estimated by its weight. As a rule, the heavier the transformer, the greater the power capability. The transformer for a 1500-W, continuous-duty amplifier will weigh 60-80 lb. The transformer used in this power supply was obtained from Peter Dahl Co.[9] The hypersil design provides a good ratio of power capability to size and weight. I have used several Peter Dahl transformer designs in the past and found them to be of excellent quality and reasonably priced.

Filter Capacitor

Enough filter capacitance is required to obtain good voltage regulation. What is enough? I have used as little as 18 µF and as much as 100 µF in power supplies. The required capacitance can be obtained with a single oil-filled capacitor or with a series string of computer-grade electrolytics. Either way, I recommend at least 25 µF be used, with at least a 10% voltage safety factor. The filter capacitor used in this power supply is a single oil-filled unit rated at 53 µF at 5kV dc. The capacitor was obtained

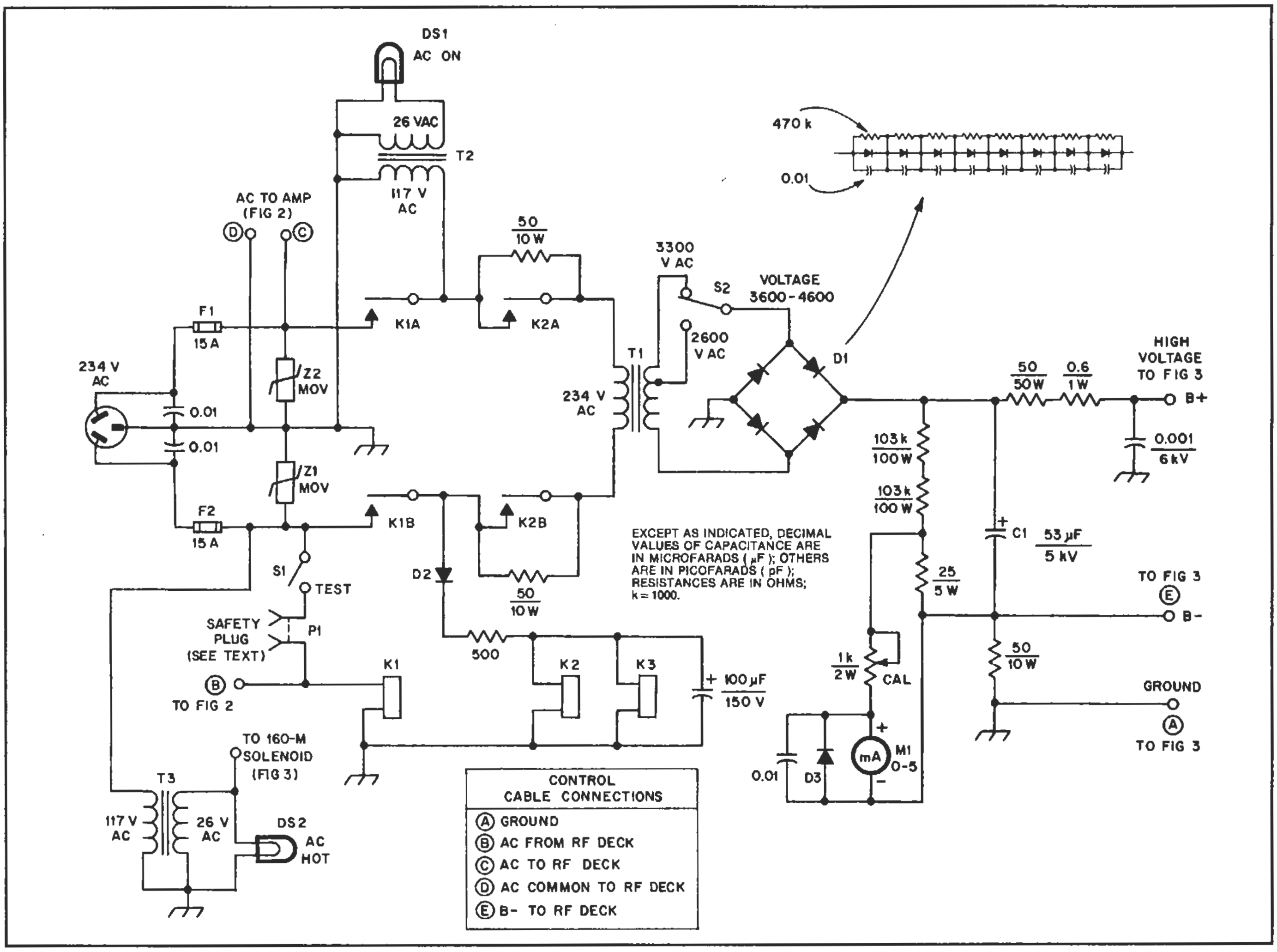
DS1
AC ON
26 VAC
T2
117 V
AC
AC TO AMP
(FIG 2)
Ⓓ
Ⓒ
50
10 W
470 k
0.01
3300
V AC
S2
VOLTAGE
3600-4600
2600
V AC
F1
15 A
K1A
K2A
T1
D1
234 V
AC
Z2
MOV
234 V
AC
0.01
0.01
Z1
MOV
F2
15 A
K1B
K2B
HIGH
VOLTAGE
TO FIG 3
50
50W
0.6
1 W
B+
0.001
6kV
103 k
100 W
103 k
100 W
C1
53 µF
5 kV
25
5 W
TO FIG 3
Ⓔ
B-
50
10 W
GROUND
Ⓐ
TO FIG 3
S1
TEST
D2
50
10 W
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF); RESISTANCES ARE IN OHMS; k = 1000.
SAFETY PLUG (SEE TEXT)
P1
Ⓑ
TO FIG 2
K1
500
K2
K3
100 µF
150 V
1 k
2 W
CAL
M1
0-5
mA
0.01
D3
TO 160-M SOLENOID (FIG 3)
T3
117 V
AC
26 V
AC
DS2
AC
HOT
CONTROL CABLE CONNECTIONS
Ⓐ GROUND
Ⓑ AC FROM RF DECK
Ⓒ AC TO RF DECK
Ⓓ AC COMMON TO RF DECK
Ⓔ B- TO RF DECK

Fig 5—High-voltage power supply schematic diagram. Part numbers in parentheses are RadioShack.

C1—Oil-filled capacitor, 53 µF, 5 kV, Peter Dahl Co.
D1—Diode bridge rectifier assembly, Peter Dahl Co, see text.
D2,D3—Diode, 1 kV, 2.5 A.
F1,F2—Fuse, 15 A.
K1—2PDT mercury plunger relay, Dayton 6X598-3.
K2,K3—SPDT relay, Potter Brumfield PRD1 DYO/9OVDC.
M1—High-voltage meter, 3½-inch Simpson Wide-Vue, 01253 bezel and 01165 lighting kit.
Z1,Z2—MOV transient suppressor, 117 V ac (276-568)
P1—Two-pin Cinch-Jones socket and plug (274-201 and 274-202)
P2—Eight-pin Cinch-Jones connector.
S1—SPST switch (275-690)
S2—Modified 6PST switch, Fair Radio Sales.
T1—Power transformer, 2600/3300-V ac sec, Peter Dahl Co.
T2,T3—Transformer, 26 V ac, 300 mA (273-1386)
Miscellaneous
Pilot lamp—Alco 16TZ, 6T-4 (yellow) and 6T-2 (red) lenses.
Cabinet—CTS model MCLS 10-17-14 black and white, SPP 10-14 black side panels.

from Peter Dahl Co, and is physically very small for the voltage and capacitance rating.

Diode Bridge Rectifier

The full-wave, diode-bridge rectifier is made up with 1000-PIV diodes rated at 3 A. The unit is a commercial module sold by Peter Dahl Co. Each diode string is built on a separate glass-epoxy board. The module is supplied with 1-inch angle brackets on each end, but because of space restraints, the angle was removed and the module was mounted in a vertical position using two Nylon bolts.

Should you decide to build the rectifier assembly, use good-quality diodes, such as HEP-170s or 1N5408s. Be sure to parallel each diode with a 470-kilohm resistor and a 0.01-µF, 1-kV capacitor.

High- Voltage Switch

The transformer has two taps on the secondary to provide a high- and low-voltage capability. The front-panel VOLTAGE 3600-4600 switch is fabricated from a 6-position, heavy-duty ceramic switch (Radio Switch Corp p/n 65). The switch detent and all but the second and fifth contacts are removed. New stops are fabricated from glass-epoxy board. Full high voltage appears across this switch, and therefore, it must be well insulated. The switch is mounted on two pieces of ¾-inch Plexiglas® to provide 2-inch spacing from any chassis or panel ground. A fiber shaft protrudes from the switch through the front panel. To protect the contacts, this switch must never be actuated when the power supply is on.

Construction Details

Next month, I will describe the unique construction details for building this high-power linear amplifier and power supply. In the meantime, should you be so inclined, get out there and find the parts! Remember that you should build the power supply first, so concentrate on those components.

Notes

[1] J. L. Pittenger, "Modular Linear Amplifier for the High-Frequency Amateur Bands," *Ham Radio*, Jan 1981.

[2] J. L. Pittenger, "A 3CX800A7 Linear Amplifier," *Ham Radio*, Aug 1984.

[3] J. L. Pittenger, '3CX1200A7 10 to 80 Meter Amplifier," *Ham Radio*, Aug 1985.

[4] W. I. Orr, ed, Radio Handbook (Indianapolis: H.W. Sams, 1975), 20th edition, Sect 22-4, "A Modern 3-1000Z Linear Amplifier for 80-10 Meters."

[5] Ham Trader Yellow Sheets, PO Box 2057, Glen Ellyn, IL 60138-2057.

[6] Radio Switch Corp, Rte 79, Marlboro, NJ 07746, tel 201-462-6100.

[7] Pioneer Standard Industrial Electronics, 1900 Troy St, Dayton, OH 45404, tel 513-236-9900.

[8] M. Wilson, ed, *The 1986 ARRL Handbook* (Newington: ARRL, 1985).

[9] Peter Dahl Transformer Co, 4007 Fort Blvd, El Paso, TX 79930, tel 915-566-5365.

By Jerry Pittenger, K8RA **From *QST*, October 1986**

An All-Band, 1500-Watt-Output 8877 Linear Amplifier

Part 2—Here's what you've been waiting for: detailed instructions for building a 1500-W RF deck and power supply.

Last month I described the circuitry and parts required for the 8877 linear amplifier and high-voltage power supply. This month I will cover the construction of both units. The power supply construction details are given first. Assuming that we have all the required parts in hand, we can determine the physical design of the power supply. To avoid costly mistakes, it is important to do adequate upfront planning before the first hole is drified.

Power Supply Construction

Meter Selection and Labeling

Any meter with a movement from 100 µA to 5 mA can be used for the high-voltage meter. This meter measures the voltage across the 25-ohm resistor at the bottom (B–) end of the bleeder string (see Fig 5 in Part 1). The more sensitive the meter, the higher the resistance setting of the variable 1-kilohm calibration resistor. The setting of the resistor is a simple Ohm's law problem. The maximum power-supply voltage on the meter scale causes a current equal to the meter-movement rating to flow through the meter, and thus give a full-scale reading. For example, the maximum scale on the meter in this supply is 5 kV dc. The meter has a 5-mA movement; therefore, if the supply is at 5 kV, 5 mA must flow through the calibration resistor and the meter for a fullscale reading. Looking at the complete bleeder string, the total resistance is the sum of the two 103-kΩ bleeder resistors plus the 25-ohm resistor, or 206,025 ohms. At 5 kY, approximately 24 mA (5000/206,025) flows through the string. The value of the calibration resistor, therefore, was selected to allow 5 mA through the meter and 19 mA through the 25-ohm resistor. Using Ohm's law, the value of the resistor should be approximately 95 ohms (19 mA x 25 ohms/5 mA). The required wattage rating of the resistor is 0.23 W (I^2R = 0.005 × 0.005 × 95), and a 1-kΩ, 2-W potentiometer was used. The meter used originally had a 0-50 scale. The scale was changed to read 0 to 5 kV.

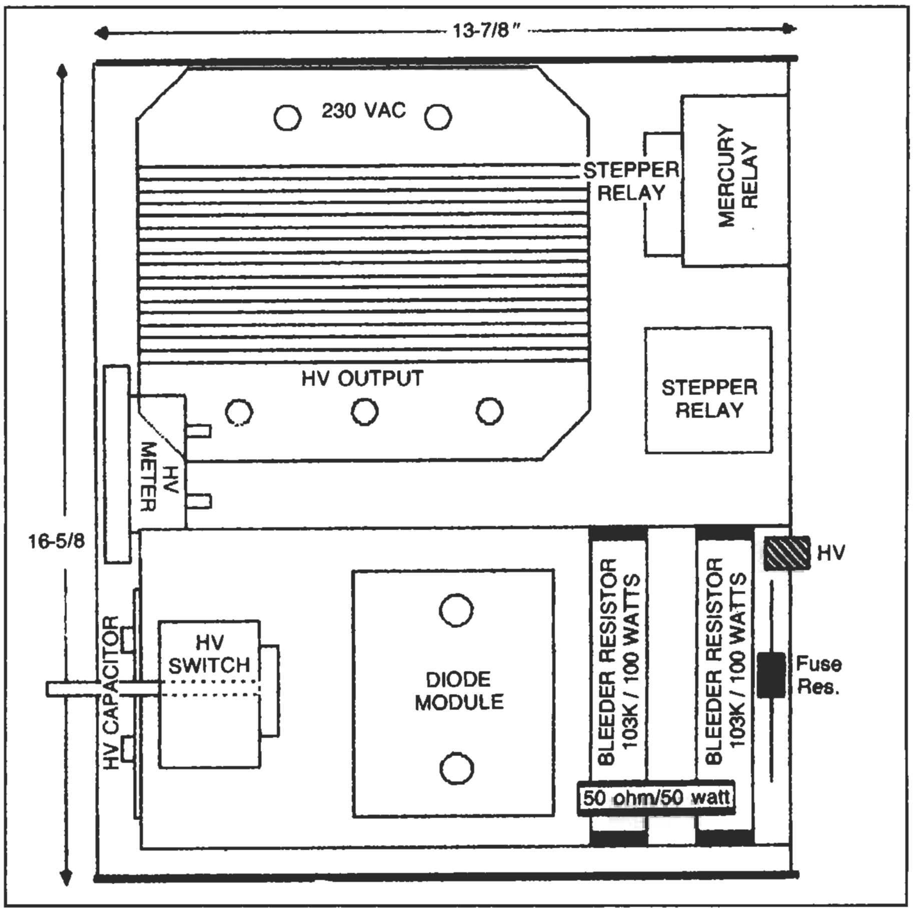

Fig 6—Typical computer-prepared layout drawing. Some component labels have been enlarged for legibility.

Panel/Chassis Layout

I do all my layouts with an Apple® Macintosh computer. An example of this layout is shown in Fig 6. A manual method may be used, instead. Cut out a piece of poster board the size of the front panel or chassis, as well as all the major components. Shuffle the pieces until you get an acceptable layout. This may seem like a lot of extra work, but "one picture's worth 1000 words." You will save time in the long run by going through this procedure, and not have to correct errors that otherwise are certain to occur. The panel layout should be symmetrical. Align switches and center meters. Make sure that components are properly spaced to accommodate the physical size of the parts behind the panel.

The computer-generated or manual layouts serve as a guide during construction, but are by no means sacred. Once you start putting the parts in place, you probably will make minor changes. Go ahead and make the changes, but always update the documentation.

Front-Panel Assembly

You probably will not be ready to cut metal until you are 3 to 4 months into the project. Do the front panel first, since the parts locations are fixed for symmetry. Parts behind the front panel can be moved to accommodate the front-panel design.

Cover the front panel with 3-inch-wide masking tape. The tape not only protects the panel from scratches, but also provides a way to lay out the panel with a pencil or pen. Before any holes are drilled, place all the major parts in their proper place in the cabinet to be sure that nothing obstructs the area behind the panel. Remember you have only one chance. Miss and it means a new panel, or cabinet!

Remove the front panel from the cabinet and center punch the panel where holes are to be drilled, as marked on the tape. Carefully drill a very small hole at each punch mark to serve as a guide, then cut the holes to final size. Holes up to approximately $^{7}/_{16}$ inch can be drilled with either a hand drill or a drill press. Holes larger than $^{7}/_{16}$ inch should be made with chassis punches. The meter hole is rectangular and is cut with a nibbler after drilling an access hole. Cut the meter hole about $^{1}/_{32}$ inch smaller than needed and finish with a large file to straighten the edges. Be careful using the file—it is very easy to let the file slip out of the hole and make a big scratch in the panel. The meter hole doesn't have to be perfect since a meter bezel is used.

After the front-panel metal work is complete, carefully remove the masking tape. The front panel should be labeled before mounting the components. Labeling is done with dry-transfer lettering available from art stores or Radio Shack. Apply the labeling by laying the letter or figure on the panel in the proper position and rubbing over it with a soft pencil. The character will be transferred to the panel. If a mistake is made, the character can be removed with masking or Scotch® tape. There is a mattefinish spray available to protect the lettering. I do not recommend using this spray. It will peel if bumped and does not work well.

The parts can now be mounted on the panel. The cabling to the front panel is connected through nylon multipin connectors (available from Radio Shack) to allow easy panel removal. The same technique is used in the RF deck.

Rear-panel Assembly

Placement of parts on the rear panel is not as critical as on the front panel. Attention should still be paid to symmetry, however. I have found that it is best to mount rear-panel parts after the major parts have been mounted inside the cabinet. The rear panel is drilled and labeled in a manner similar to the front panel. The black rear panel requires white lettering. White lettering kits are available at art supply stores and some electronics suppliers.

Fig 7 shows the inside rear panel of the power supply. The control and power cables enter the rear panel and are routed directly to barrier strips. Each terminal is labeled for clarity. The rear panel is accessible by removing the front panel and the transformer. It sounds like a big job, but it can be accomplished in about 10 minutes. It is necessary to provide good strain relief for the cables. Immediately inside the rear panel, a piece of $^{3}/_{4}$-inch aluminum angle stock is mounted, to which each cable is clamped. Large rubber grommets are used in the holes for cable protection.

Major Chassis Assembly

Power-supply components are extremely heavy and a good supporting structure is required. The $^{1}/_{16}$-inch-thick bottom cover supplied with the cabinet is replaced with a $^{1}/_{4}$-inch-thick base plate to provide an adequate foundation. The heavy plate is cut to size on a commercial metal shear. After the transformer, relays and filter capacitor are mounted, rubber-wheeled casters are bolted to the bottom so that the supply can be rolled, rather than carried from place to place.

The cabinet sides are formed by two removable panels. Parts can be mounted to the inner panel using countersunk screws and then covered with the $^{1}/_{8}$-inch painted cover plate to provide a professional appearance. A $^{1}/_{4}$-inch-thick sheet of Plexiglas® is mounted above the filter capacitor to support the high/low voltage switch, the diode bank and the bleeder resistors. The Plexiglas is supported by drilled and tapped holes for no. 6-32 countersunk screws in the rear and side panel. A post of $^{1}/_{2}$-inch aluminum bar stock supports the front-left corner of the Plexiglas.

The mercury-wetted power relay is mounted on the rear panel and must be positioned vertically. Mercury-wetted relays have a tendency to buzz if mounted on a solid surface. Use a rubber grommet to make a bushing in each mounting hole, or mount the relay on a rubber pad. The stepstart relays, K2 and K3, are also mounted on $^{1}/_{4}$-inch-thick rubber sheet to minimize noise.

Once the major components are mounted, wire them together, performing as much testing as possible along the way.

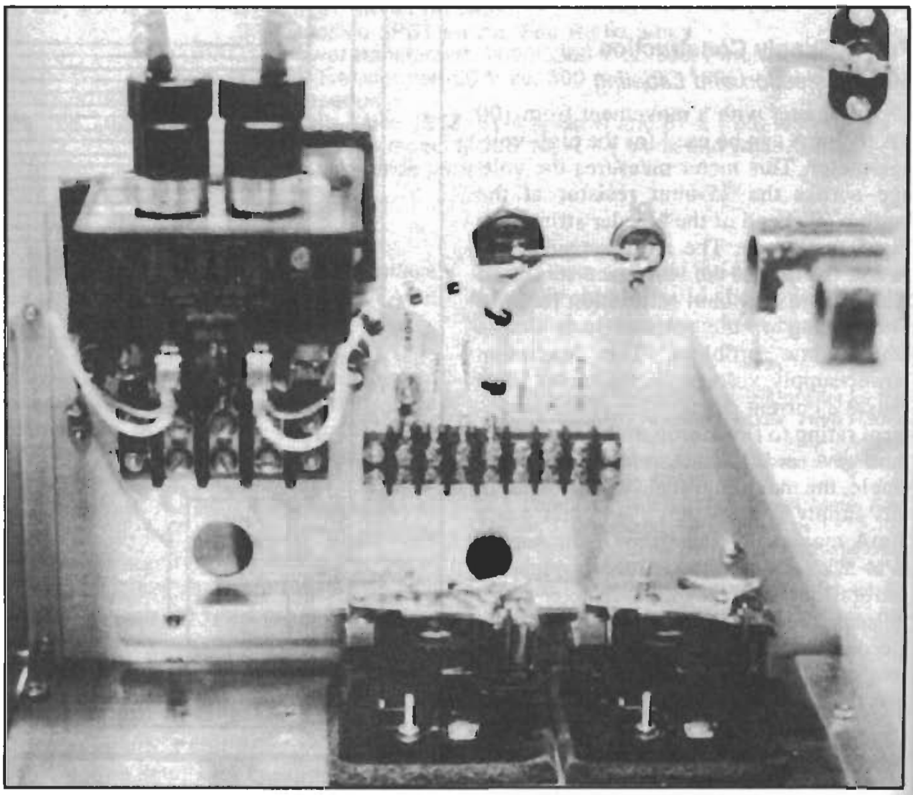

Fig 7—Interior view of rear panel. Note the mercury plunger relay mounting. Relays K2 and K3 are at the lower right.

Front- and rear-panel components are mounted and wired before those that are mounted to the cabinet sides and base plate.

Power-Supply Testing

When you are satisfied that the power supply has been correctly wired and carefully checked, go have a cup of coffee and come back later when your mind is fresh and check the wiring one more time. *Remember that this power supply can be a lethal device. One wrong move could be fatal.*

Final testing of the power supply is accomplished in three steps. Remove the wiring from the primary of the transformer and place a temporary line cord with a small variable autotransformer directly from a 117-V ac line to the primary. Turn the 117 V on and slowly run the variable autotransformer up. At 117 V, the power supply should be reading half-scale voltage. This test verifies that the diode bank and filter capacitor are correctly wired. Now is a good time to calibrate the front-panel meter. Using the variable auto-transformer, set the output voltage at a level that another VOM can measure accurately. For example, most VOMs can measure 1 kV. Adjust the calibration resistor so that the front-panel meter reads the same as the VOM.

The second test checks the primary 234-V ac circuitry. With the 234-V lines still disconnected from the primary, plug the power-supply line cord into a 234-V source. Remember that the shorted two-contact plug must be inserted into the rear-panel socket. Turn the power supply on with the test switch and listen for a 3- to 4-second time delay for the step-start relays. Check that 234 V appears across the two wires disconnected from the primary.

The final test is to try the entire power supply with 234 V applied to the primary. Check the high- and low-voltage indications on the front-panel meter.

Power-Supply Performance

The design results in a husky supply that delivers about 4.6 kV and 3.8 kV (no load) on the high- and low-voltage positions, respectively. In either position the power supply drops less than 400 V under full load. This is the performance you need to achieve full-power capability and good linearity.

RF Deck Construction

The physical design of the RF deck requires planning long before construction begins. Again, I used the Macintosh computer for my initial "paper design," but the same process can be done using paper and pencil. The important thing is to lay out the major components so that everything fits properly before starting to drill holes and cut metal. You must have all the major components in hand before doing the physical design, so you know what you have to work with.

Fig 8 shows the scaled Macintosh designs for the front, top and bottom views of

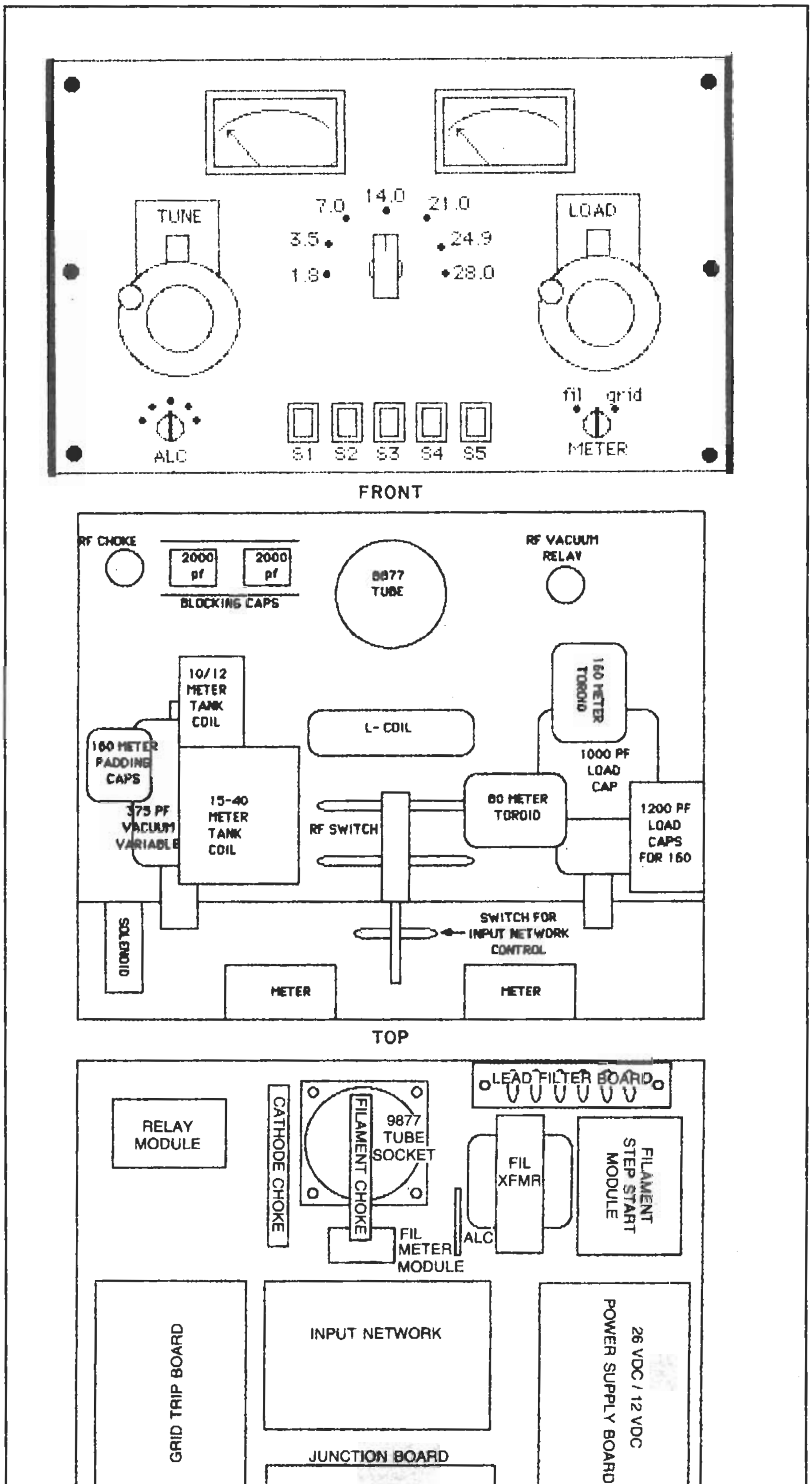
TUNE
7.0
14.0
21.0
3.5
24.9
1.8
28.0
LOAD
fil
grid
ALC
S1
S2
S3
S4
S5
METER
FRONT
RF CHOKE
2000 pf
2000 pf
BLOCKING CAPS
8877 TUBE
RF VACUUM RELAY
160 METER TOROID
10/12 METER TANK COIL
L-COIL
160 METER PADDING CAPS
375 PF VACUUM VARIABLE
15-40 METER TANK COIL
RF SWITCH
80 METER TOROID
1000 PF LOAD CAP
1200 PF LOAD CAPS FOR 160
SOLENOID
SWITCH FOR INPUT NETWORK CONTROL
METER
METER
TOP
RELAY MODULE
CATHODE CHOKE
FILAMENT CHOKE
9877 TUBE SOCKET
LEAD FILTER BOARD
FIL XFMR
FILAMENT STEP START MODULE
FIL METER MODULE
ALC
GRID TRIP BOARD
INPUT NETWORK
26 VDC / 12 VDC POWER SUPPLY BOARD
JUNCTION BOARD
SWITCH
SWITCH
SWITCH
SWITCH
SWITCH
BOTTOM

Fig 8—Scale layouts of the RF deck front, top and bottom views, as prepared on the Macintosh computer. Some component labels have been enlarged for legibility.

the amplifier. It is necessary to match the designs so that the front-panel controls end up at the right place on the panel. Everything must be drawn to scale to obtain the relative positions of the components. When designing the bottom and top layouts, only the major circuit boards and components like the grid trip, input network, low-voltage power supply, coils, capacitors and filament transformer are considered. The smaller components can be fitted in later. It is important that the unit be designed with maintainability in mind. Every component, large and small, must be accessible after the unit is completed. The amplifier was constructed in the following steps. Each major step is discussed in more detail later.

- Fabricate, build and test all printed circuit boards for the RF deck.
- Cut the holes in the front panel.
- Perform the major metal work on the subpanel and chassis plate.
- Mount the vacuum capacitors and band switch to the subpanel.
- Mount the PC boards, filament transformer and tube socket to the chassis plate.
- Wire and test the under-chassis control circuits.
- Fabricate and install the RF tank circuit.
- Label and calibrate meters.
- Complete the front panel and mate it to the chassis and cabinet.
- System test the amplifier with the power supply.

Printed-Circuit Board Fabrication

Making PC boards can be tedious, but with a little practice, good results can be obtained. Grouping interconnected circuits on the same PC board minimizes the cable harness between modules in the amplifier. In this design, nine PC board modules are required.

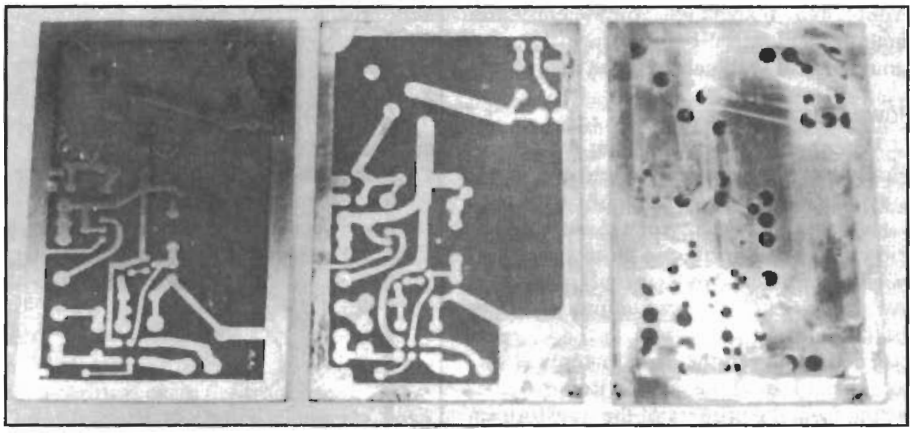

Fig 9—Things don't always work out. Three attempts were made to fabricate the grid-trip and bias PC board. The good board at left was subsequently damaged, requiring a fourth effort.

	Board Size
1) Input network	4" × 6"
2) Low-voltage power supply and timer circuit	4" × 6"
3) Grid-trip and bias circuit	4" × 6"
4) Switch bank harness interface to front panel	1" × 6"
5) Filament step-start circuit	2¾" × 3¾"
6) RF input/output relay timing circuit	2¾"
7) Line filters for control cable and blower	1¾" × 5¼"
8) Filament-voltage-meter circuit	1⅛" × 2"
9) ALC circuit	1¾" × 2¼"

All boards, except nos. 4, 7, 8 and 9, are mounted on ½-inch channel to the chassis plate. Therefore, the PC board layouts include a ground strip on the edges of each board to mount the boards to the channel.

The Macintosh computer, with a software package called "Draw," was used to lay out the circuit boards. The design of PC boards with the computer is beyond the scope of this article. PC boards can also be laid out using pencil and paper. Once the design is laid out to scale, the layout must be transferred to the board. First wash the board with a mild detergent to remove all grease and dirt. Using tape and special dry transfers (Radio Shack p/n 276-1577), copy the design onto the board. The transfers are not exact, and the hand drawing only serves as a guide. Submerge the board into an etching solution to remove the exposed copper. I use ferric-chloride etchant. A flood lamp over the etching tray warms the solution and speeds the etching process. Agitate the solution occasionally. The final result should be a nicely etched board ready for drilling the holes for mounting parts. Believe me, however, it doesn't always go as planned.

Fig 9 shows three tries at making the grid-trip and bias board. Actually the third board, on the left, was damaged, and a fourth was necessary. However, once you get the hang of it, it usually goes well.

After etching the board, polish it with

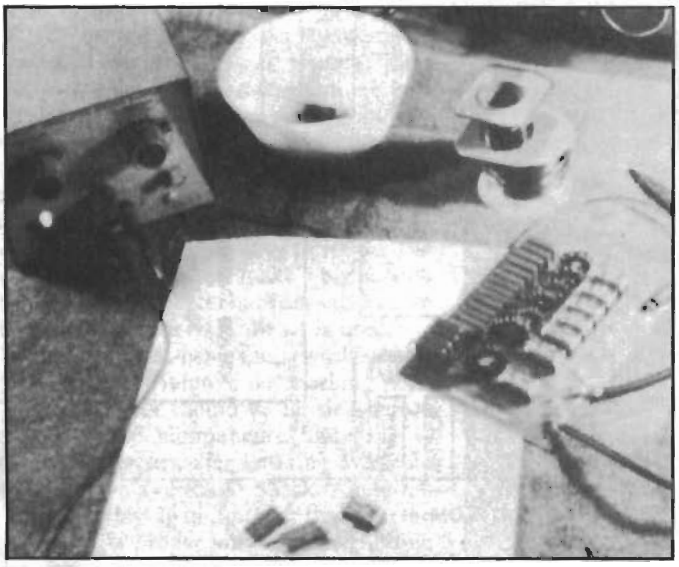

Fig 10—In-process testing of the input-network PC board. Each coil was trimmed to produce a 1:1 SWR into a 50-ohm load. Note the power supply required to actuate relays.

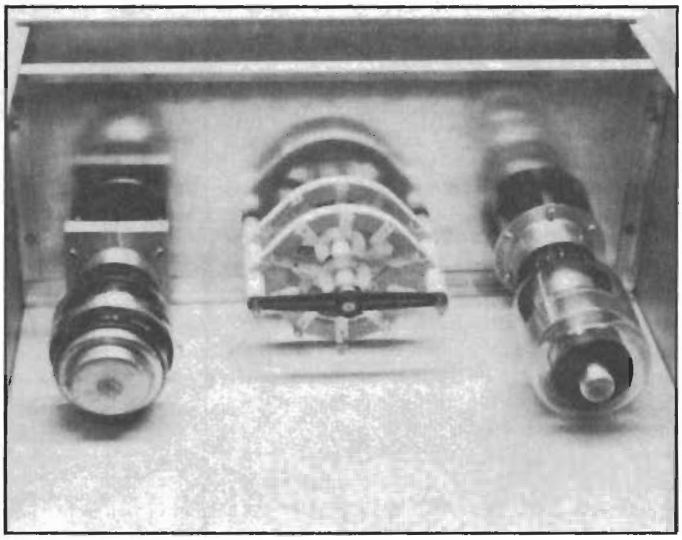

Fig 11—Subpanel installation in the cabinet. Note position of chassis plate, vacuum variable capacitors and band switch. Aluminum angle stockis used to mount the subpanel to the chassis plate and cabinet sides.

fine steel wool and drill the component mounting holes with a small drill. Use dry transfers to label the connections required to the board, before mounting parts. This will avoid wiring mistakes when the wiring harness is installed. Finally, mount and solder the components onto the board.

Test each board as much as possible before final assembly. For example, Fig 10 shows the input network being tested by hooking up coaxial cable and running 100 W through each section into a 50-ohm dummy load. The coils were adjusted at this time to give a flat 1:1 SWR. A power supply is needed to energize the relay for the pi-section being tuned.

Front-panel Fabrication

Cover the front panel with masking tape to protect it from scratches, and mark the mounting positions of the components. Take your time with this step since it is hard to recover if an error is made. Mark the front panel where the holes are to be cut and carefully drill a small pilot hole followed by the correct hole size. The holes for the band switch, capacitor control shafts, ALC control and multimeter switch are 3/8-inch diameter and can be made with a regular drill bit. The holes for the power switches are 5/8-inch diameter and require a chassis punch. The meter-mounting holes are the most difficult to make. The rectangular cutouts are marked about 1/32-inch smaller than required to protect against overcutting. Use a nibbler to cut the holes carefully, then file them to the exact size of the meter bezels. Be careful that the file doesn't slip and scratch the panel (disaster!). Using a large file will help avoid this tragedy.

I recommend doing all the front-panel metal work at one time to ensure proper component layout. When the metal work is complete, leave the masking tape on the panel for protection as it will be used as a template to locate the parts mounted behind the front panel.

Chassis and Subpanel Metal Work

This amplifier design has very little metal work that requires more than a hacksaw and a file. The chassis plate was purchased with the cabinet. The only other panel, except for front and rear panels, is the subpanel that mounts perpendicular to the chassis plate, and 3 inches behind the front panel. The subpanel shields the meter compartment from RF and serves as a mounting support for the band switch, vacuum capacitors and 80-meter toroid coil.

The chassis is mounted 3½ inches from the bottom of the cabinet to allow room under the chassis plate for the filament transformer. Therefore, the subpanel is cut to 6¾ × 16½ inches. Referring to Fig 11, ½-inch aluminum angle is attached with screws to each edge of the subpanel to provide a mounting flange to the chassis plate and cabinet side walls. The top piece of angle stock provides a mounting surface for a piece of gold-plated finger stock that seals the subpanel to the cabinet top plate.

Two large holes are required. A 3-inch-diameter hole is required in the chassis plate for the tube socket. A 5-inch-diameter hole is cut in the cabinet top plate and aligned directly above the tube socket to vent the air flowing from the tube and chimney. Use a large fly-cutter, available at most hardware stores. For safety, the fly-cutter should be used only on a drill press—never with a hand drill. Therefore, if you don't have a drill press, find a friend who has one. Before cutting the large holes, use 3-inch masking tape to cover the chassis plate and top panel to mark where the holes are to be drilled, and to protect the surfaces. For safety, clamp the panels onto a board and the base of the drill press before drilling.

Cut a piece of perforated aluminum stock to be slightly larger than the hole in the cabinet top panel. Clean the perforated metal well, and spray with paint to match the cabinet color. Fasten the perforated piece to the inside top panel with several small countersunk screws painted to match the cabinet.

Mounting Vacuum Variable Capacitors and Band Switch

The vacuum variable capacitors and band switch are mounted to the subpanel, but their control shafts must be aligned with the front panel design. Allow enough slack in the positioning of these components to perform precise alignment with the front-panel holes when the front panel is installed. To mark the hole positions on the subpanel, slide the subpanel against the rear of the front panel while both the front panel and chassis plate are bolted into place in the cabinet. Cover the subpanel with masking tape and mark the exact centers of the holes for the components. Drill the two holes in the subpanel for the vacuum variable capacitors with hole saws. Mount the vacuum variable capacitors and band switch on the subpanel.

Mounting Major Components to the Chassis Plate

The major components (PC boards, filament transformer and tube socket) are mounted to the chassis plate as shown in Fig 12. The PC boards are first mounted to ½-inch aluminum channel using sheet-metal screws for easy removal. Cover the bottom of the chassis plate with masking tape. Position and mark the PC boards, filament transformer and tube socket according to the planned physical layout. Drill the mounting holes in the chassis plate, then redrill the holes from the top with a countersink bit to allow flat-head countersunk screws to be used for mounting. This retains the flat surface on top of the chassis plate. With the ½-inch channels on the PC boards, mark the hole positions on the bottom of each channel with a pencil, using the predrilled chassis-plate mounting holes as a template. Remove the channels from the PC boards, mount the channels on the chassis plate, and remount the PC boards on the channels.

Under-Chassis Wiring

Complete the under-chassis wiring, according to the schematic diagram, using Teflon®-insulated wire. The nylon connectors near the front of the chassis connect to the front-panel power switches, the ALC potentiometer and the multimeter switch. This allows the front panel to be easily removed for rear-panel access. The small board in the front center of the chas-

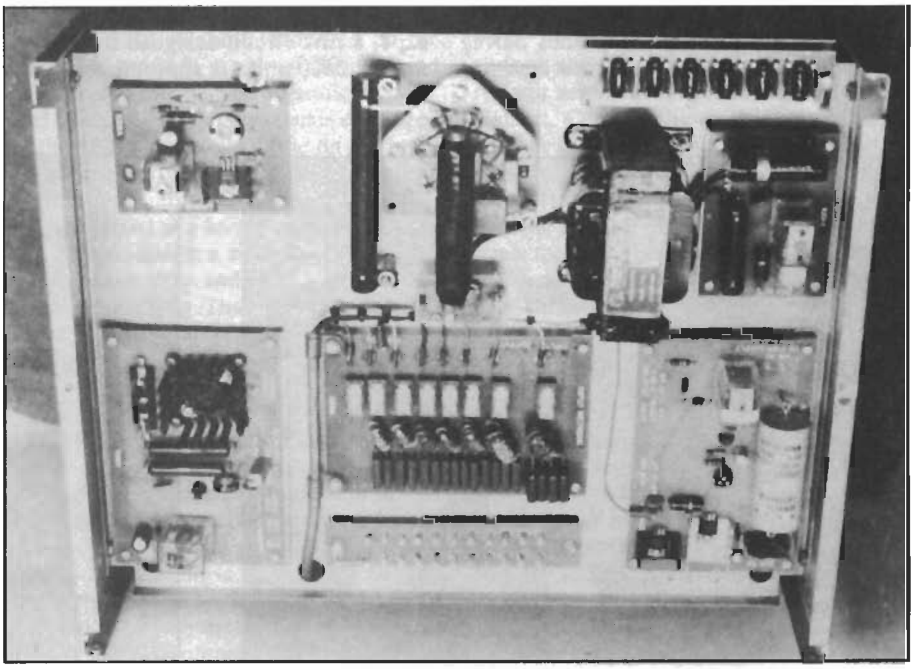

Fig 12—Bottom view of the RF deck showing placement of PC boards, transformer and tube socket.

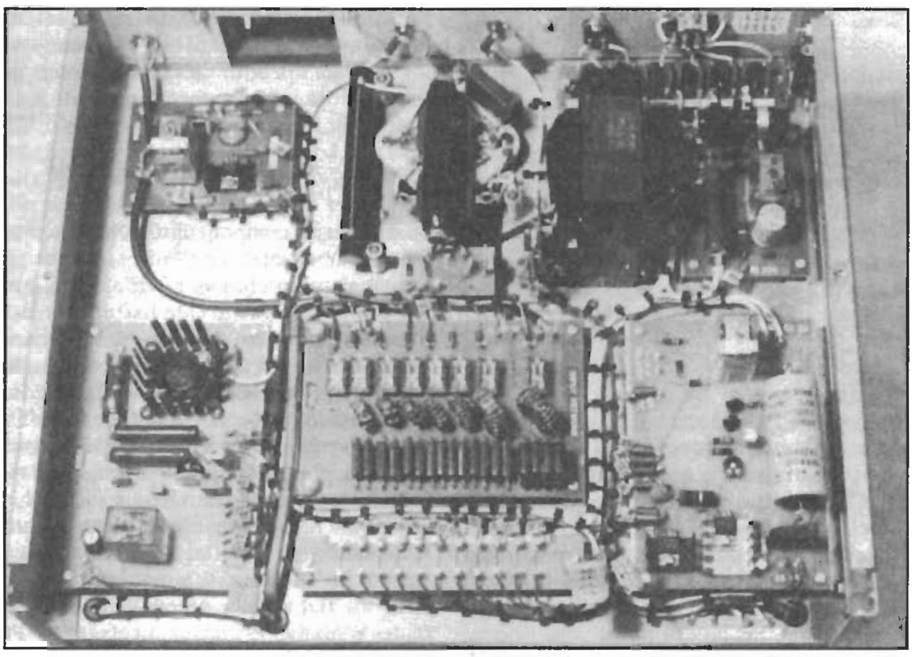

Fig 13—Bottom view of the RF deck showing harness wiring in place.

sis provides an easy way to mate the 12-pin nylon connector going to the power switches with the wiring harnesses under the chassis plate (see Fig 13). When making the PC boards, try to put all the connections to the board on one side. This allows access to the boards for maintenance without removing the wiring. Just unscrew the board from the channel and fold the board upward.

Each wire is labeled at each end with numbered tags, because the Teflon wire used is mostly the same color. A version of the schematic diagram was maintained with the wire numbers noted for easy wire tracing. The final wiring is cabled into harnesses with plastic cable ties. Don't be afraid to use plenty of ties, but at first only put a tie every inch or two. The ties will undoubtedly be cut several times during wiring to put in missing wires that are overlooked. When all wiring is complete and tested, put a cable tie every ½ inch on the major harnesses.

Amplifier Tank Circuit

The plate tank-coil set is shown in Fig 14. Before making the coils, detennine the tankcircuit parameters for the given tube plate impedance. The plate impedance can be determined from Eq 1.

Plate impedance = plate voltage/ (1.57 × plate current) (Eq 1)

Assuming the amplifier runs at 60% efficiency, the input power required for 1500-W output is approximately 2500 W. With 3200 V on the 8877 plate, plate current will be approximately 781 mA. From Eq 1, a tank circuit designed for approximately 2600 ohms is appropriate. A table of pi-L network component values is contained in Hoff's article, and the values used in this amplifier are summarized in Table 2.[10]

The 80- and 160-meter coils are wound on an assembly of three T225-2 toroid cores taped together with Scotch no. 27 glass-cloth tape. Use plenty of tape to provide good voltage insulation from the cores. The 80-meter coil is wound with 11 turns of no. 10 wire covered with Teflon sleeving. The 160-meter coil is wound with 19 turns of no. 12 wire, also covered with Teflon sleeving. The cores of both coils are mounted from 1 to 2 inches from the mounting wall on ceramic insulators. Each coil is sandwiched between two pieces of fiberglass material held together by a ceramic standoff running through the middle of each toroid (see Fig 15). Wind all coils in the same direction (clockwise or counterclockwise), to avoid a "bucking" action between coils.

The 10- through 40-meter coil is made from ¼-inch soft copper refrigerator tubing. Clean the tubing using fine steel wool (000) until the surface is smooth and bright. Use a piece of pipe with an OD equal to the desired coil ID and carefully wind the copper tubing around the pipe. After the smaller diameter part of the coil has been wound, change the pipe to a larger diameter, and wind the larger part of the coil. The inductance of the coil can be easily checked using a known fixed-mica capacitor with low internal inductance and a dip meter. After determining the resonant frequency, calculate the inductance using Eq 2.

Inductance = $1/(4 \pi f^2 C)$ (Eq 2)

Always wind the inductor with a few extra turns and then remove turns until the desired inductance is achieved. The induc-

---

[10] I.M. Hoff, "Pi Network Design for High Frequency Power Amplifiers," *Ham Radio*, Jun 1978.

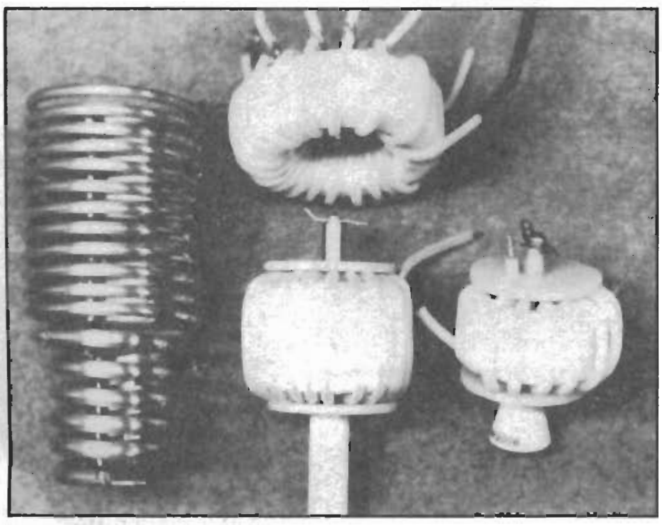

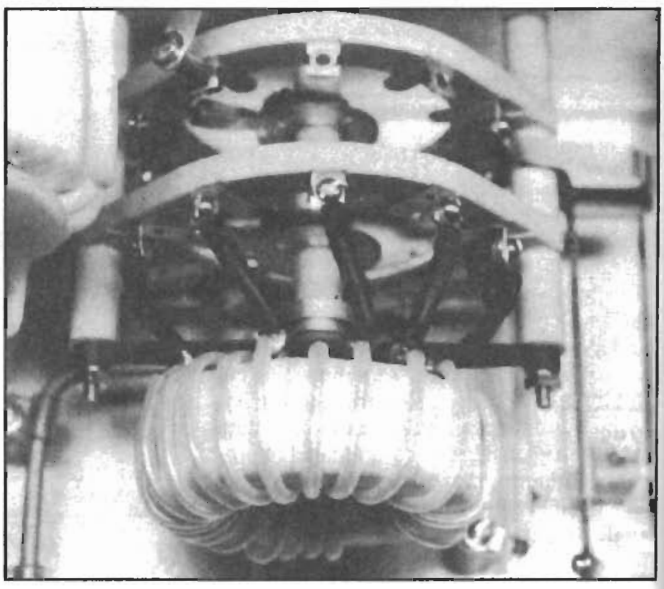

Fig 14—Plate tank coil set. Above: the various coils of the tank circuit. Right: the L toroid mounted on the rear of the band switch.

Table 2

Pi-L Network Design Parameters (See Fig 3, Part 1)

Freq (MHz)	*C1 (pF)*	*L1-L4 (µH)*	*C2 (pF)*	*L5 (µH)*	*Q*
1.8	462	22.02	2121	8.90	13.1
3.5	244	10.99	1132	4.45	13.4
7.0	113	6.03	503	2.44	12.4
14.0	55	3.08	245	1.24	12.2
21.0	37	2.05	164	0.83	12.2
24.9	30	1.70	130	0.70	12.0
29.7	26	1.48	112	0.60	12.0

L1, 10-12 meters—6 turns ¼-inch silver-plated copper tubing, 2-inch outside diameter, 2 inches long. 10-m tap at 3 turns.

L2, 15-40 meters—12 turns ¼-inch silver-plated copper tubing, 3¼-inch outside diameter, 4 inches long. 15-m tap at 2 turns, 20-m tap at 4 turns, 40-m tap at 12 turns.

L3, 80 meters—11 turns of no. 10 wire in Teflon sleeving over three T225A-2 iron-powder toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape.

L4, 160 meters—19 turns of no. 12 wire in Teflon sleeving over three T225A-2 toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape.

L5, L Coil—20 turns of no. 20 wire in Teflon sleeving over two T300-2 iron-powder toroidal cores taped together with Scotch no. 27 glass-cloth electrical tape. Tap as follows:

Band	*10-12*	*15*	*20*	*40*	*80*	*160*
Turn no.	2	3	4	6	10	20

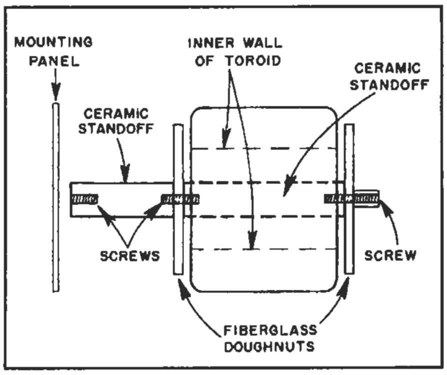

Fig 15—Mounting details of 80- and 160-meter toroids.

tance measurement is only an approximate value because there is always stray inductance introduced by the interconnecting leads and the band switch.

The copper tubing is silver plated only after the ends are configured to match the mounting lugs in the amplifier. Silver plating is simple with the right materials. Go to a photographic lab and get a couple of gallons of used fixer solution. Fixer solution is not usable when it becomes saturated with silver. The labs usually sell the spent fixer for silver recovery, so you may have to pay for it. Clean the coil by giving it another brushing with steel wool, wash it thoroughly in a mild detergent and rinse it well before plating. Put the fixer solution in a large plastic container and submerge the coil in it. A bright silver plate will form within seconds. The silver plate is not very thick, but it is sufficient to keep the copper from tarnishing. The coil taps are small wrap-around clamps made from silver-plated ½-inch-wide copper strap. Plate about 2 inches of copper strap, then drill a hole in one end for a no. 8-32 screw. Using a short piece of ¼-inch tubing as a jig, bend the end sharply around the tubing with a pair of duckbill pliers. Drill a second, matching hole, but squeeze the clamp in to provide a tight fit around the coil stock. Install the clamp with a flat washer on each side to provide a compression fit. The clamp should fit tightly. Don't solder the clamps until after the amplifier is tested. When the tap location is verified, heat the clamp and coil with a large solder gun and flow solder into the connection. After soldering the coil taps, spray the coil with a thin coat of clear plastic to retain the bright silver finish.

The L coil is wound on a pair of T300-2 toroid cores taped together with glass-cloth tape. The coil is mounted on the back of the band switch and supported by the connections to the coil. The leads are not soldered to the band switch, but are held in place with no. 6-32 screws.

Other major considerations in the RF tank circuit are the TUNE and LOAD capacitors. The TUNE capacitor is a 375-pF vacuum variable. Table 2 shows that though this capacitor will easily cover 10 through 80 meters, it is too small for 160 meters. One solution is to use a 500-pF vacuum variable, but units in this size class are not readily available in the surplus market, and you probably don't want to buy a new one at approximately $500. I solved the problem by switching in an additional 160-pF capacitor in parallel with the vacuum variable for 160 meters.

A 24-V dc solenoid, mounted on the subpanel, grounds a pair of 80-pF fixed vacuum capacitors. The solenoid is powered by a small 12-V relay that is energized by the 160-meter control line on the input network control switch. The grounding strap is made from a stiff piece of brass stock obtained at a hobby shop. A hinge is made by soldering a small piece of brass tubing to the strap and putting a solid rod through the tubing. The rod is supported on each end by a small piece of Plexiglas and allowed to pivot, thus forming a hinge. A piece of flexible braid is connected from the brass strap to ground for a good connection. The strap is held in the open position by a small spring pulling the strap back toward the subpanel. The contacts on the strap and the vacuum capacitors are from a 25-A power relay and soldered in place. A piece of ¼-inch brass bar stock is used to mount the contact on the vacuum capacitor side and to absorb the shock from the closure. A rubber grommet on the brass strap, where the rod from the solenoid hits the strap, also absorbs some of the shock. This scheme works smoothly and once thought out, took about one evening to install.

The LOAD capacitor is a 1000-pF, 3-kV unit with a fixed 100-pF, 5-kV fixed mica transmitting capacitor in parallel. This provides a total of 1100 pF for 80 meters. On 160 meters, the band switch adds an additional 1200 pF in parallel to provide a total of 2300 pF. The 1200-pF capacitor consists of three 400-pF mica transmitting capacitors in parallel.

Marking and Calibrating Meters

Any meter with a movement from 100 µA to 5 mA can be used for the front-panel meters. Use two identical meters, if possible, for esthetic reasons. In a very clean environment, begin by removing any internal meter shunts (on ammeters), or any series resistors inside the meter (on voltmeters) to obtain only the basic meter movement. Nothing else need be done for the multimeter, since the calibration resistors are a part of each metered circuit. A shunt resistor must be made for the plate-current meter. Wind about 2 feet of no. 22 enameled wire on a 2-W resistor of any value over 50 ohms, soldering the wire ends to the resistor terminals. Mount the resistor across the meter terminals to form a shunt. Connect the meter in series with another meter of known calibration and an adjustable power source. Trim the enameled wire 1 inch at a time until the meters read the same. If the meter being calibrated reads too low, the wire is too short. When the meters read the same, slip a piece of heat-shrink tubing over the resistor and seal the shunt.

Carefully remove the plastic face cover and the calibrated face plate from the meter. Do this in a clean environment because the magnet in the meter will attract metal shavings that could damage the movement. With a pencil eraser, rub off the original meter lettering, but not the analog scale tick marks. With small dry-transfer letters put the new numeric scale on the meter. Be careful not to bend the pointer when reassembling the meter.

Mounting the Front Panel

Remove the masking tape from the

panel, and carefully label the front panel with dry-transfer lettering before mounting any components. Mount the meters, potentiometer and switches in place, and complete the wiring. Wire nylon connectors to the power switches, ALC control, multimeter switch and meters to mate with the connectors in the RF deck. This will result in a totally removable panel.

When mounting the front panel to the cabinet, give careful attention to aligning the vacuum-capacitor and band-switch shafts. Loosen the component mounting screws and mate the shafts, then retighten the screws. Continue the alignment process until the controls operate smoothly.

Rear-Panel Assembly

Up to now, nothing has been done to the rear panel. Mount the panel and decide where the components should mount to obtain short connections. Lay the rear panel out in the same way as the front panel, and cut the necessary holes.

Mount the blower off the rear panel, located to allow good circulation of air up to the front of the under-chassis area and back to the tube socket. Positioning of the blower is not critical as long as air is not directly blown across the tube socket, which could cause backpressure.

Mount all components to the rear panel, and mount and wire the panel to the RF deck. The rear panel is not easily removable like the front panel.

Testing the Amplifier and Power Supply

Testing is the *big* moment and the climax of several months' work. First hook up the power-supply control cable and the ground cable. Leave the high-voltage cable off. Test the control circuits to ensure that the power supply can be turned on from the RF deck and the 3-minute time delay works. Make sure that the tube filaments are on. This isn't as easy as with a glass tube, which allows you to see the filaments glow. Let the tube run for about 10 minutes, then turn the power off. Immediately remove the tube and feel if the base is hot.

Test the amplifier first on 20 or 40 meters, since these bands use the midrange of the TUNE and LOAD capacitors. With high voltage applied to the tube, first key the amplifier with no input drive power. The resting idle current should be between 100 and 200 mA on the plate-current meter.

Now apply a little drive power to the tube. The plate current should rise. Move the TUNE and LOAD controls until power output is indicated on a wattmeter. Increase the drive while adjusting the TUNE and LOAD controls to achieve about 20 mA grid current and 650 to 700 mA of plate current, to realize 1500 W output. If the amplifier works properly on this band, proceed to the other bands and repeat the tests.

Efficiency should be at least 60% on all bands, except perhaps 10 and 12 meters, where efficiency may drop to 55%. If efficiency is poor, try moving the coil taps, but recognize that moving the taps also changes the Q of the coil. Decreasing the inductance and increasing the capacitance in the tank circuit will increase the Q. The amplifier may provide more power output over a larger frequency range with a lower tankcircuit Q, but harmonic suppression will also decrease and the amplifier could start generating interference or not meet FCC standards.

I must warn you one last time. *This device could kill you in one instant if you get tied into the high-voltage circuit. Put a good ground on both power supply and RF deck, and treat the equipment with proper respect!*

Conclusion

I have spent many enjoyable hours operating with this amplifier/power supply combination, with nothing but good signal reports. It was an exhausting task, but now that it's finished, I'm glad I did it. Try building one and you'll see what I mean.

By David Rutledge, KN6EK, et al. **From *QST*, May 1997**

High-Efficiency Class-E Power Amplifiers

Part 1—With 3 to 12 W of drive, you can push 300 to 500 W CW out of an $11 transistor! The trick is to use Class E.

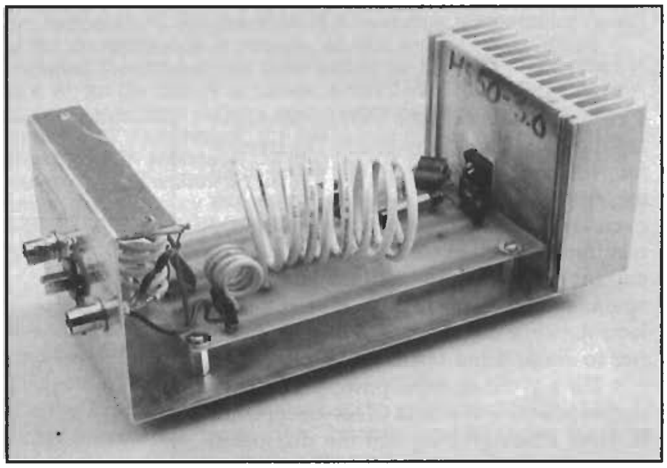

These 300 and 500-W 40-meter amplifiers evolved from a series of undergraduate student projects at Caltech. Our goal was to design an inexpensive amplifier that amateurs can easily duplicate.[1] The amplifiers use inexpensive, readily available power MOSFETs that can be driven by a QRP transceiver; in our case, the NorCal 40A. A block diagram of our setup is shown in Figure 1. The components of the station are shown in Figure 2.

Our 300-W amplifier uses an International Rectifier (IR) IRFP440; the 500-W amplifier employs an IRFP450. These transistors are widely used in switching power supplies, but we have not seen them previously reported for use as RF amplifiers. The MOSFETs have a maximum drain voltage of 500 V, with maximum RMS drain currents of 8.8 A for the IRFP440, and 14 A for the IRFP450. Both transistors are available from Digi-Key: The 440 costs $8; the 450 costs $11.

Class-E amplifiers are extremely efficient—about 90%. Because of the low loss, no cooling fan is required. No TR switch is needed; the received signal is piped through the amplifier itself. Even without an external filter, the amplifiers meet the FCC *requirements for spurious emissions.* They can be built and tuned up with an RF power meter, a multimeter, oscilloscope and a dummy load. Tune-up consists of adjusting an input coil for matching and an output coil to set the power level.

Operational Classes

In addition to the well-known Class A, B, and C operational modes, there are Class D and E.[2] In Class D and E, the devices operate as switches, half the time completely on, and the other half completely off. But transistors are not perfect switches. The MOSFETs have a resistance of about 1 Ω when on, and a capacitance of several hundred picofarads when off. Losses are greatly reduced in switching amplifiers, but there is a penalty: The output power no longer depends on the *drive power*, but rather on the *supply voltage*. This means that switching amplifiers are *not* linear amplifiers, and they are *not* suitable for SSB without additional limiting and modulating circuits. However, they are fine for CW, FSK and FM.

In a Class-D amplifier, a pair of transistors switch on and off, out of phase across an output transformer. Fred Raab, WA1WLW, recently developed a Class-D power amplifier that produces 250 W on 40 meters with an efficiency of 75%.[3] However, Class-D amplifiers are relatively complex. On the other hand, the Class-E circuit has significant advantages for the homebrewer because only one transistor, without gate bias or output transformer, is needed, and it can be driven by a low-power transceiver.

The Class-E Amplifier

The Class-E amplifier was invented and patented by Nathan Sokal, WA1HQC, and Alan Sokal, WA1HQB, in 1975.[4] It minimizes heat loss by having as little overlap as possible between voltage and current. Figure 3 shows an idealized Class-E circuit. As the switch opens and closes, the current alternately flows in the switch and in the load network. The switch voltage and current waveforms are shown in Figure 4. It may be easiest to understand the waveforms by starting at the beginning of the off-time interval. When the transistor turns off, the current flows into the resonant load *network, and there is a transient voltage* that rises and falls. With a properly de-

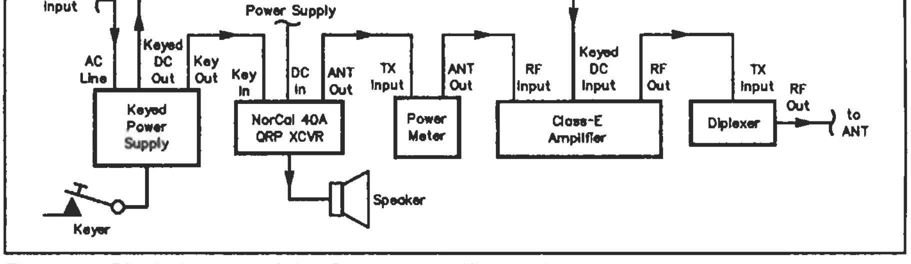

Figure 1—Block diagram of the Class-E amplifier station.

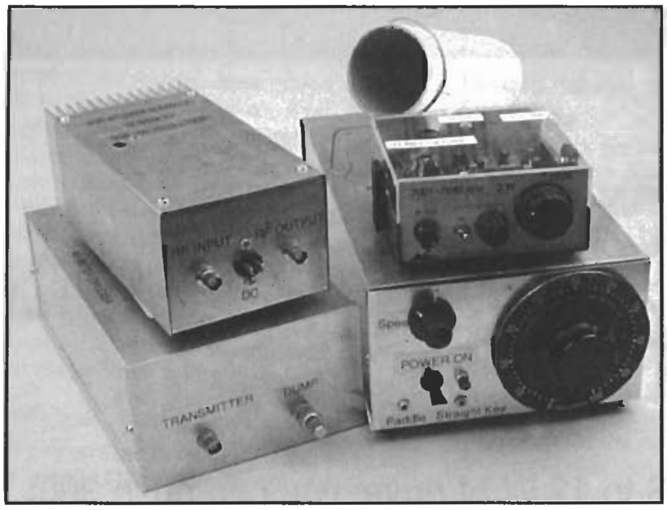

Figure 2—The components of the Class-E amplifier station. Clockwise from bottom left: the diplexer, amplifier, resonant speaker, NorCal 40A and keyed power supply. The latter houses the keyer and a pulse-stretching and shaping circuit.

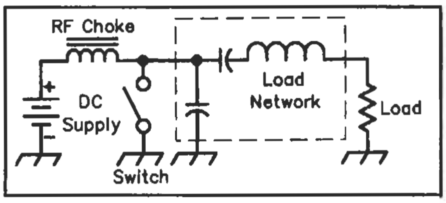

Figure 3—An idealized Class-E amplifier. The transistor is represented by a switch that opens and closes at RF.

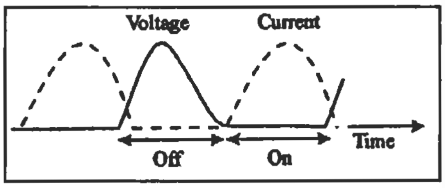

Figure 4—Class-E voltage and current waveforms. Class-E amplifiers reduce loss by keeping the overlap between voltage and current low.

signed load network, the voltage returns to zero smoothly with zero slope. The transistor switches on when both the voltage and the current are small, keeping losses low even if the switching is slow or slightly mistimed. Once the transistor turns on, the current rises smoothly until it switches off again, and the cycle repeats. The resonant load network does limit a Class-E amplifier to single-band operation.

For a given dc input power, the power output of a 90% efficient Class-E amplifier is three times greater, and the dissipated power is seven times smaller, than that of a 30% efficient Class-A amplifier. This means that for a given dissipated power, we can get *21 times* more power from a Class-E amplifier.

The Caltech Power Amplifiers

Both amplifiers have a common diagram (Figure 5) and PC board, but use components with different values. Several components are added to the basic Class-E circuit for matching and filtering. The MOSFET's gate impedance is rather low, and primarily capacitive, with a reactance of about 4 Ω. There is also a resistive component of about 2 Ω from parasitic series resistance in the gate itself, and the drain *on* resistance that is capacitively coupled to the gate. T1 reduces the 50-Ω impedance of the drive circuit to about 2 Ω to match the low resistance of the gate and sets the dc bias to 0 V. L1 is adjusted to cancel the gate capacitance; the input SWR is typically 1.5:1. C1 shunts L1 at high frequencies to reduce ringing in the VHF range. The 0-V gate bias ensures that the transistor is off when it is not driven, because this is far below the threshold voltage, which is about 4 V. We have never seen oscillations in these amplifiers.

C3 and L2 form a resonant network that produces the rising and falling voltage waveform needed for the Class-E ampli-

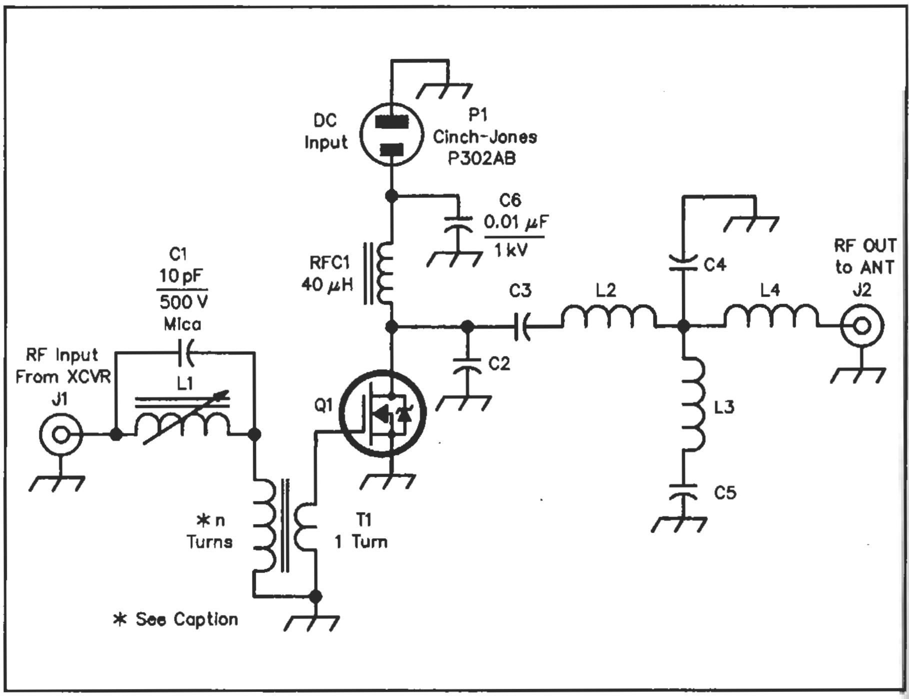

Figure 5—Circuit diagram and parts list for the amplifiers. Mica capacitors are available from Newark. L2, L3 and L4 are made with #10 THWN solid, insulated house wire sold in hardware stores. T1's core is an RF400-0, available from Communications Concepts (see Note 1 for supplier information). L1 is a Toko 10K (2.2 μH), available from Digi-Key. RFC1 is a J. W. Miller type 5240, available from Newark.

C1—10 pF, 500 V mica
C2—300 W, 270 pF; 500 W, 390 pF; (For C2 through C5, Cornell Dubilier mica capacitors, type CDV19, 1 kV, 5%
C3—300 W, 1500 pF; 500 W, 2000 pF
C4—300 and 500 W, 100 pF
C5—300 W, 680 pF; 500 W, 820 pF
C6—0.01 μF, 1 kV ceramic disc
J1, J2—BNC or SO-239 connectors
L1—2.2 μH, Toko 10K; available from Digi-Key.
L2—300 W, 9 turns; 500 W, 8 turns #10 THWN wound on 1¼-inch-OD plastic-pipe form; spread turns to fit holes in PC board.
L3—300 W, 4 turns; 500 W, 3 turns #14 THWN wound closely spaced on a ½-inch-diam drill bit
L4—300 W, 5 turns; 500 W, 5 turns #14 THWN wound on a ½-inch-diam drill bit form, turns spaced to occupy 1 inch
P1—Cinch-Jones P302AB
Q1—IRFP440 for 300-W amplifier; IRFP450 for 500-W amplifier; (use International Rectifier transistors only)
RFC1—40 μH, 3 A
T1—Pri: 300 W, 5 turns; 500 W 6 turns #26 stranded hook-up wire; wound on RF400-0 core, available from Communications Concepts
Misc: Berquist K10-104 insulating pad (thermal resistance of 0.2-K/W, 6-kV breakdown rating); available from Digi-Key.

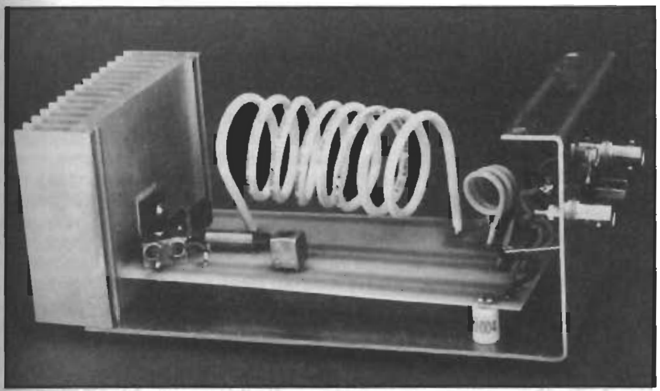

Figure 6—Inside the 500-W amplifier. The large coil is L2; L3 is the small coil. The circuit board is mounted on $^{1}/_{2}$-inch standoffs. BNC connectors are used for the RF input and output. No harm is done if the RF input and output are accidentally interchanged.

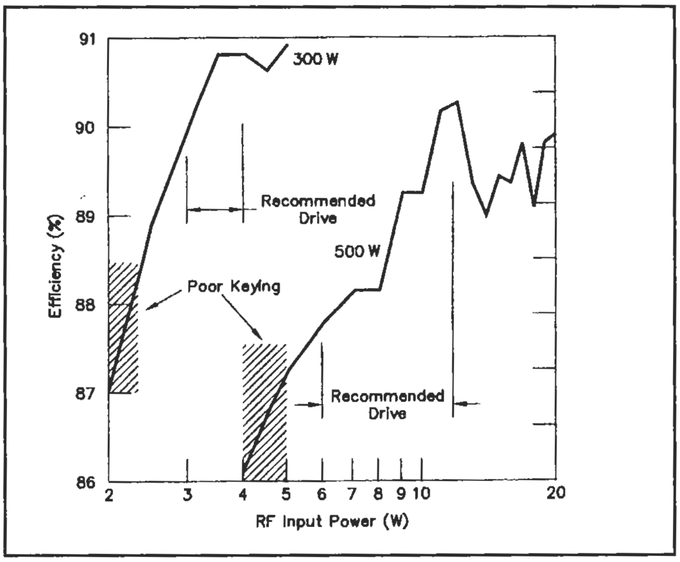

Figure 7—The measured efficiency plotted against RF input power. These measurements were taken with the heat sink at room temperature. In operation, the heat-sink temperature rises and the *on* resistance increases, so that the efficiency drops somewhat. Recommended drive power for the 300-W amplifier is 3 to 4 W; for the 500-W amplifier, 6 to 12 W. The dc input power was calculated from voltage and current measurements made with Fluke 87 multimeters; the RF output power was determined with a Bird 43P wattmeter using a 500-W element. Bird lists the accuracy of this wattmeter as ±25 W, but we were able to improve the accuracy of the measurement to an estimated 1% by thermal calibration. RF input power was measured with a Diamond SX-200 wattmeter.

fier. C5 and L3 act as a notch filter for the second harmonic. Without the notch filter, the second harmonic is typically between -25 and -30 dBc instead of the -40 dBc that the FCC requires on HF. In addition, C5 and L3 transform the 50-Ω antenna impedance to about 10 Ω, the appropriate load for a Class-E amplifier. RFC1 converts the 0 to 120-V dc input from the power supply to a current source, and C6 helps keep RF energy out of the power supply.

C4 and L4 form a low-pass filter to remove VHF harmonics. Without the filter, there are several harmonics at levels from -40 to -60 dBc in the frequency range from 130 to 210 MHz. With the filter, the VHF harmonics are reduced to the -70 to -80-dBc level.

Amplifier Construction

The 500-W amplifier is shown in Figure 6 with its cover removed. The transistor is mounted on a 3∞4$^{1}/_{8}$-inch heat sink with 1-inch fins (type HS50-3.0 from RF Parts, with a thermal resistance of 2 K/W with no fan) with a #6-32 bolt and nut. The transistor generates most of the heat (about 10% of the dc power) in the amplifier, so it must have good thermal contact to the heat sink. Because the transistor's case reaches high voltages, it must be electrically isolated from the heat sink. We use a Kapton pad manufactured by Berquist that has a thermal resistance of 0.2 K/W and a breakdown voltage of 6 kV. The heat-sink surface must be free of burrs, and the transistor should lie flat on the surface with minimal stress on the leads. If a torque screwdriver is available, International Rectifier recommends a mounting torque of 10 inch-pounds. The heat-sink baseplate is an aluminum L bracket bent from 0.050-inch-thick aluminum sheet. A U-shaped enclosure cover is made of 0.016-inch-thick aluminum sheet. A hole in the cover allows insertion of a plastic screwdriver for tuning L1.

Solder Q1 and C2 flush to the PC board to reduce VHF ringing on the gate signal. Good electrical contact is needed between Q1's source lead and the heat sink. Use a #6-32 bolt and nut, with washer spacers so that the source lead does not bend. Scope-probe pigtails soldered to Q1's gate and drain leads poke through holes in the baseplate. Rubber grommets in the holes prevent the pigtails from shorting to the chassis.

L4, L3 and L2 are made by winding solid insulated wire on pipe and drill-bit forms. Orient the coils at right angles to each other to reduce coupling between them. L2 has a 7-MHz Q of 350; L3 and L4, a Q of about 170. For C2, C3 and C5, use only 1-kV mica capacitors—even 500-V capacitors fail spectacularly with a burst of flame. If a 1-kV, 100-pF capacitor is not available for C4, substitute a series-connected pair of 500-V, 200-pF mica capacitors. For best filtering, mount C4 and L4 directly on the center pin and ground lug of J2. Mount L4 with its axis vertical to reduce coupling to the other coils.

The NorCal 40A Driver

For a driver, we use a NorCal 40A,[5] but its 2-W output is not enough to drive these amplifiers. Fortunately, the NorCal 40A can be modified to deliver greater power output.[6] We recommend 3 to 4 W drive for the 300-W amplifier, and 6 to 12 W for the 500-W amplifier. These drive levels give an efficiency in the 90% range (Figure 7). Drive levels lower than these give poor efficiency; higher drive levels increase the dissipated power without improving efficiency. Don't drive the 300-W amplifier with less than 2.3 W, and the 500-W amplifier with less than 5 W. At these low power levels, the transistor may not turn on fully at all supply voltages, subharmonic spurious components may be generated, and the amplifier may not key properly.

The recommended drive powers produce peak gate voltages of between 15 and 20 V. International Rectifier specifies a maximum peak gate voltage of 20 V to avoid rupturing the gate. Although the drive levels are close to this limit, experience shows them to be quite safe. The 20-V limit is more appropriate for the low frequencies used in power supplies than for RF voltages. In controlled tests, we've pushed the 300-W amplifier to 60-V gate-voltage peaks, *three times* the manufacturer's voltage limit, without damage.

Next month, we'll discuss the keyed power supply, keying waveform shaper and tune-up. Join us!

Notes

[1] A package of amplifier parts *only*, including PC board, components, connectors, heat sink and chassis is available at cost from Puff Distribution, Department of Electrical Engineering, MS 136-93, Caltech, Pasadena, CA 91125. Price: $50 for US orders, $60 for foreign orders. The price includes tax and shipping by surface mail. Make payment by check or money order only to "Caltech-Puff Distribution." Foreign checks must be drawn on a bank with a US branch office. Please provide your Amateur Radio call sign and specify which amplifier (the 300 or 500 W unit) you want. For more information, contact Dale Yee by e-mail at **yee@systems.caltech.edu**, fax 818-395-2137, or you can download an order form from: **http://www.systems.caltech.edu/EE/Faculty/rutledge/poweramp.html**. We do not offer power-supply components. Those parts are available from Communication Concepts, 508 Millstone Dr, Beavercreek, OH 45434-5840, tel 513-426-8600; Digi-Key, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539, **http://www.digikey. com**;

Newark Electronics, (many branches throughout the US; check your telephone book for a branch near you); main office: 4801 N Ravenswood Ave, Chicago, IL 06040-4496, tel 800-463-9275, 312-784-5100, fax 312-907-5217; RF Parts, 435 South Pacific St, San Marcos, CA 92069, tel 800-737-2787; Mouser Electronics, 2401 Hwy 287 N, Mansfield, TX 76062, tel 800-346-6873, 817-483-4422, fax: 817-483-0931 e-mail **sales@mouser.com**; **http://www.mouser.com**.

A template package is *not* available from the ARRL.

[2] Historically, amateurs have built high-power amplifiers with vacuum tubes rather than transistors to avoid complicated power-combining networks and many low-power transistors. See Dick Ehrhorn, W4ETO, "RF Power Amplifiers and Projects," *The 1996 ARRL Handbook*, Chapter 13. Exceptions to this are the elegant designs by Helge Granberg, K7ES/OH2ZE (SK). *RF Application Reports*, published by Motorola Inc, in 1995, contains over 20 Application Notes and Engineering Bulletins written by Helge Granberg on amplifiers with output powers of 20 to 1200 W. Communication Concepts sells the boards and components for these amplifiers. *Radio Frequency Transistors*, by Norm Dye and Helge Granberg, published by Butterworth-Heineman, Boston, 1993, is recommended reading. Joel Paladino, N6AMG, adapted one of Granberg's transistor amplifiers (see "An Experimental Solid-State Kilowatt Linear Amplifier for 2 to 54 MHz," *QST*, Sep 1992, pp 19-23). However, the power transistors alone cost $900! This led us to look for a less-expensive way to make transistor power amplifiers.

[3] Fred Raab, WA1WLW, "Simple and Inexpensive High-Efficiency Power Amplifier," *Communications Quarterly*, Winter 1996, pp 57-63.

[4] Nathan Sokal, WA1HQC, and Alan Sokal, WA1HQB, "Class-E, A New Class of High-Efficiency Tuned, Single-Ended Switching Power Amplifiers," *IEEE Journal of Solid-State Circuits*, Vol SC-10, Jun 1975, pp 168-175. This paper by father and son is a classic in the radio engineering literature, and is still the basic reference for the Class-E amplifier.

[5] The NorCal 40A designed by Wayne Burdick, N6KR, is available in kit form for $129 from Bob Dyer, KD6VIO, at Wilderness Radio, PO Box 734, Los Altos, CA 94023-0734, tel 415-494-3806, **http://www.fix.net/jparker/wild.html**. The NorCal 40A Web address is: **http://www.fix.net/~jparker/norcal.html**.

[6] A number of NorCal 40 modifications were published in *QRPp*, the magazine of the Northern California QRP club that first made the NorCal 40 kit available. For information on subscriptions to *QRPp* and back issues, contact Jim Cates, 3241 Eastwood Rd, Sacramento, CA 95821. Wayne Burdick, N6KR, suggests increasing the supply voltage for more power. If you do this, increase the voltage rating of Zener diode D12 to accommodate a larger peak collector voltage. For 3 W output, we use an MRF237 at Q7, the PA. To further increase the power to 7 W, we reduced L7 from 18 turns to 11 turns, and L8 from 18 turns to 14 turns. We replaced C45 and C47 with 560-pF, 300-V mica capacitors, and C46 with a 1500-pF, 100-V mica capacitor. For other approaches to raising the output power, see Dave Meacham, W6EMD, "5 Watts from your NorCal 40A," *QRPp*, Mar 1995, pp 6-7, and Ron Manabe, KN6VO, "Increasing the Output Power of the NorCal 40," *QRPp*, Jun, 1994 pp 42-45. Ron reports an output power of 7 W.

By David Rutledge, KN6EK, et al.

From *QST*, June 1997

High-Efficiency Class-E Power Amplifiers

Part 2—Class-E operation permits low-cost MOSFETs to develop considerable power.

Last month[7] we talked about Class-E amplifier fundamentals and began construction of a 40-meter unit. Now we'll tackle the power supply, keyed-waveform shaper and develop some power.

A Keyed Power Supply

Nonlinear Class-E operation sharpens the CW keying envelope, causing annoying key clicks. To prevent this, we key the power supply to shape the supply voltage. A separate 4×7×12-inch (HWD) enclosure houses the dc supplies, a stretcher circuit that delivers a stretched pulse to the driver and a shaper that produces the shaped pulse for the amplifier. Figure 8 shows how these circuits connect.

So that the RF drive does not end before the shaping pulse, the keying pulse to the NorCal 40A driver is stretched a few milliseconds. The stretcher (Figure 9) takes a keyer's CMOS logic signal and provides a buffered keying waveform to the shaper and a stretched keying waveform to the NorCal 40A driver. Dc supplies (Figure 10) provide 12 V dc to run the ICs and 0 to 120 V dc for the amplifier. A wave shaper (Figure 11) gives this 0 to-120 V dc supply voltage a controlled rise and fall time to avoid key clicks. Figure 12 shows the keyed power supply with its cover removed.

There are other advantages to using a keyed power supply to control the output power. The amplifier power dissipation is low at all supply voltage levels, so that loss is kept low throughout a keying pulse. Because the keyed power supply also acts as a solid-state TR switch, a relay is not needed. This is because the supply voltage is zero except during key down. (This feature works well with the NorCal 40A, because it, too, does not use a relay for switching.) At zero voltage, the drain-to-gate capacitance in a MOSFET is quite large, and the signal from the antenna is fed through the amplifier with a loss of only about 7 dB. The NorCal 40A receiver sensitivity is excellent,[8] and a 7 dB signal loss does not hurt reception at all. A 7-dB loss degrades the MDS to –130 dBm, still far below typical 40-meter antenna noise levels of –90 to –110 dBm. On the positive side, with 7 dB attenuation, the receiver is less susceptible to intermodulation distortion from other signals in the 40 meter band. In addition, the amplifier reduces AM broadcast signals by about 20 dB. The 7-dB loss does need to be made up at the audio end. For this, we mount a 2-inch-diameter speaker in a cardboard mailing tube cut to resonate at 650 Hz for CW reception. This gives a sound level that is quite adequate.[9]

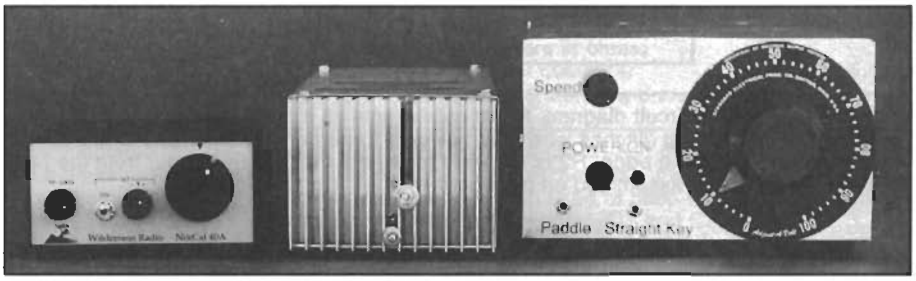

Transceiver, amplifier, and power supply for a 40-meter, 500-W station. The NorCal 40A transceiver driver is on the left, a 500-W amplifier is in the center and the power supply on the right. The amplifier's heat sink and transistor mounting screws are visible. The large dial controls the variable autotransformer, varying the RF output power.

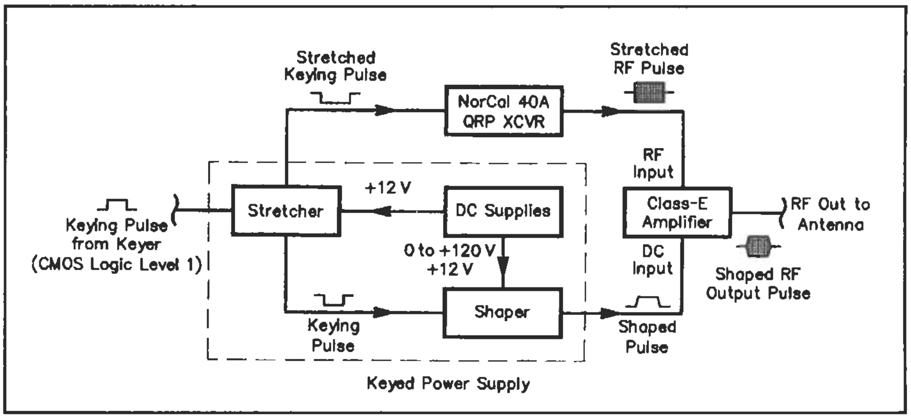

Figure 8—Block diagram showing the connections between the stretcher, dc supplies, shaper, NorCal 40A and the Class-E Amplifier. The stretcher, the dc supplies, and the shaper are in the keyed power supply. Power supply keying is done by a Curtis keyer IC (not shown) that provides a CMOS logic-level 1 during key down. The Curtis keyer IC and application note are available from MFJ Enterprises Inc, Box 494, Mississippi State, MS 39762, tel 800-647-1800, 601-323-5869, fax 601-323-6551; e-mail **mfg@mfjenterprises.com**; **http://www.mhjenterprises.com**.(Keyer circuits using Curtis ICs can be found in recent editions of *The ARRL Handbook.—Ed.*)

The Diplexer

For greater reduction of spurious emissions, we recommend following the amplifier with a band-pass diplexer (see Figure 13) to terminate out-of-band spurious components in a 50 Ω load.[10] Our diplexer (in a 3.5×6×10-inch [HWD] box) uses the equivalents of a 100-pF series capacitor, an 1800-pF shunt capacitor and air-wound inductors. Stretch or compress L2 to achieve minimum SWR. The measured loss of 40-meter signals was extremely low, only 4%. Our experience shows that a diplexer can reduce all spurious components to more than –55 dBc.

Tune-Up

Refer to Figure 3 in Part 1. There are two amplifier coil adjustments. First, with the cover on and the dc input off, L1 is set for minimum input SWR with full RF input. Typically, the SWR can be reduced to 1.5:1. If it cannot be brought below 2:1, try adding or subtracting a turn from L1.

Output power is peaked by stretching or squeezing L2. Note: For safety, the amplifier's cover should always be attached when the RF drive is applied. The cover also significantly lowers the inductance of L2. Be sure to turn off the supply voltage before you touch *any* amplifier parts! A high RF voltage will burn the skin. RF burns are deep and heal slowly. Having a keyed power supply helps here, because the amplifier supply voltage is zero except during key down.

Attach a dummy load and power meter to J2. With the RF input applied continuously, slowly increase the dc input voltage, while monitoring the gate and drain voltages using an oscilloscope with 10× high-impedance probes. (You should see waveforms similar to those in Figure 15, although the peak drain voltage should be about half of

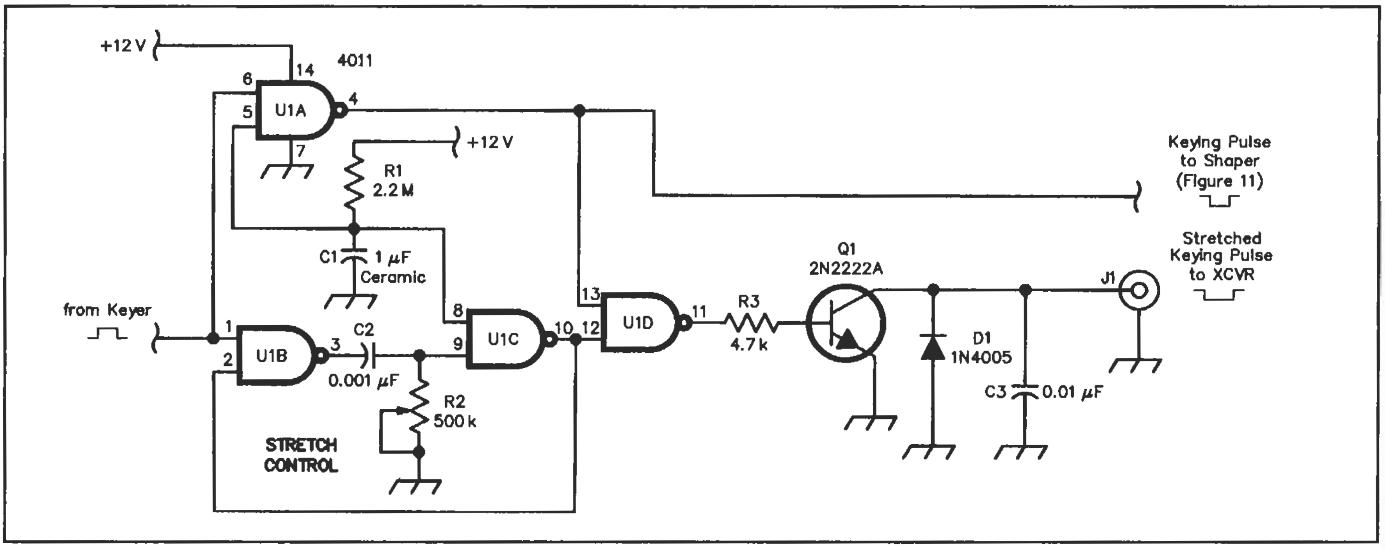
+12 V
4011
14
6
5
U1A
4
7
+12 V
R1
2.2 M
C1
1 µF
Ceramic
from Keyer
1
2
U1B
3
C2
0.001 µF
R2
500 k
STRETCH
CONTROL
8
9
U1C
10
12
13
U1D
11
R3
4.7 k
Q1
2N2222A
D1
1N4005
C3
0.01 µF
J1
Keying Pulse
to Shaper
(Figure 11)
Stretched
Keying Pulse
to XCVR

Figure 9—Stretcher circuit diagram. **NAND** gate A acts as a buffer and produces a keying pulse for the shaper. One input has an RC delay to prevent keying glitches when the power is turned on. Gates B and C are connected by an RC network that causes a pulse to be triggered on a falling edge. R2 is adjusted to ensure that the RF output from the NorCal 40A lasts longer than the shaping pulse. The components are assembled on perfboard.

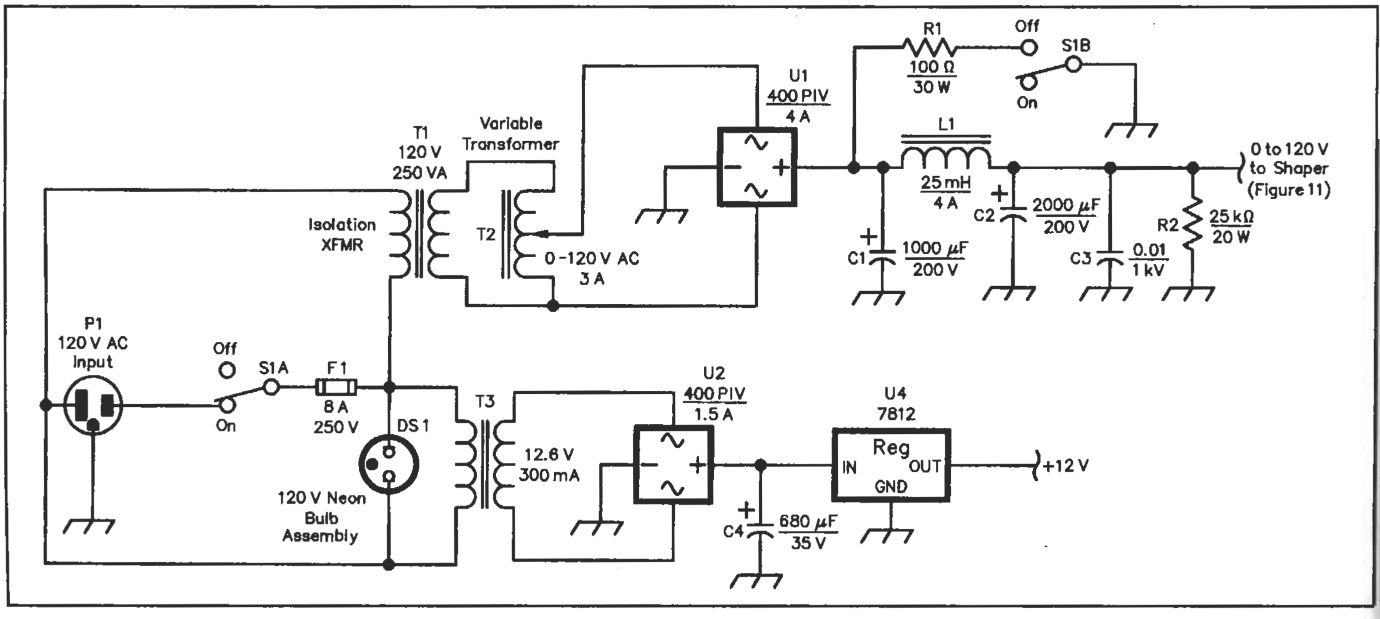
T1
120 V
250 VA
Variable
Transformer
Isolation
XFMR
T2
0-120 V AC
3 A
U1
400 PIV
4 A
R1
100 Ω
30 W
Off
S1B
On
L1
25 mH
4 A
C1
1000 µF
200 V
C2
2000 µF
200 V
C3
0.01
1 kV
R2
25 kΩ
20 W
0 to 120 V
to Shaper
(Figure 11)
P1
120 V AC
Input
Off
S1A
On
F1
8 A
250 V
DS1
120 V Neon
Bulb
Assembly
T3
12.6 V
300 mA
U2
400 PIV
1.5 A
C4
680 µF
35 V
U4
7812
Reg
IN
OUT
GND
+12 V

Figure 10—Dc supplies diagram. There are two isolated supplies: one with an unregulated 0 to 120-V output controlled by an autotransformer, and another with a +12-V regulated output. Two bleeder resistors help keep capacitor voltages at safe levels when the supply is turned off or its output reduced. The 4-A bridge rectifier is an RS404LR by Diodes, Inc, available from Digi-Key. The Panasonic 1000 and 2000-µF electrolytics in the filter are available from Digi-Key. The Stancor C-2686 25-mH choke is available from Newark. We scrounged the isolation transformer and variable autotransformer from old equipment. Look for them at swap meets, or purchase equivalents from Newark (Magnetek N55M 250-VA isolation transformer and the Staco 291 3-A variable autotransformer).

that in the figure.) The RF output should begin to rise. Increase the dc input until the forward RF output power is 25% of full power. Make sure the output SWR is 1.5:1 or less. Measure the dc input voltage, and adjust L2 to give 25% power at 60 V dc. Stretching L2 reduces its inductance and increases the output power, but usually lowers the amplifier's efficiency. Squeezing L2 reduces the output power and usually increases amplifier efficiency.

Increase the dc input voltage until the amplifier reaches maximum output power. The dc voltage should now be between 115 and 120 V, and the peak drain voltage of the RF-stage MOSFET should be between 380 and 420 V. Larger peak drain voltages run the risk of transistor failures. Lower voltages may indicate an excessive drain current, which can lead to a failure. If the voltage is too high, stretch the coils a bit more. If the voltage is too low, squeeze the coils. Check the RF drive and input SWR again. You may find that they have changed somewhat and that readjustment is needed. Measure the dc supply current and voltage, and calculate the amplifier efficiency to ensure that it is 85% or above. Use a 0.001-µF capacitor to bypass the voltmeter terminals because an RF voltage there can cause a significant measurement error. (In addition, you should realize that RF power meters often have error factors as high as 10%.)

Keyed-Waveform Shaping

The keying envelope is controlled by the three potentiometers in the shaper and stretcher circuits. R3 in the shaper circuit sets the rise time, R4 sets the fall time. R3 also helps control the power supply droop as the amplifier is keyed. R2 of Figure 9 determines the stretch in the pulse that keys the NorCal 40A, so that it does not stop transmitting before the end of the shaped

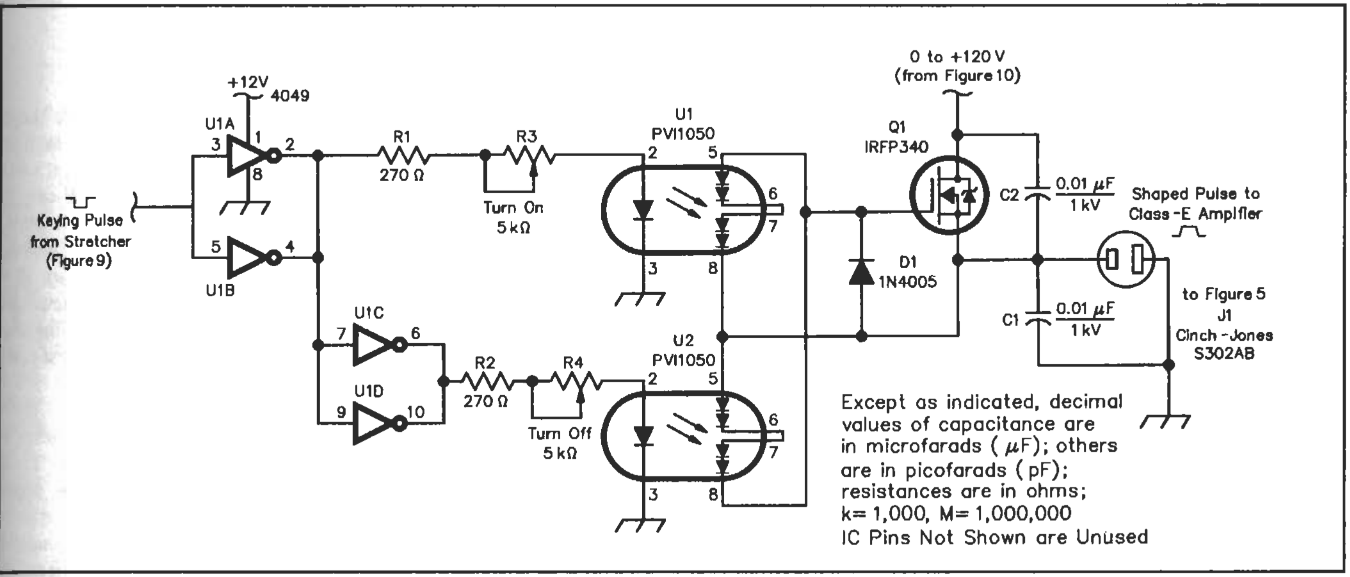

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k= 1,000, M= 1,000,000

IC Pins Not Shown are Unused

Figure 11—Shaper circuit diagram. A pair of optically isolated MOSFET drivers (type PVI1050) turn Q1 on and off to bring the supply voltage up and down. Keying waveform rise and fall times are adjusted by R3 and R4, which control the current to each MOSFET driver. The 270-Ω resistors limit the LED current to a safe level. The TURN-ON potentiometer is also used to keep the overshoot from the unregulated power supply to a reasonable level. D1 prevents Q1's gate voltage from drifting negative. C1 and C2 are RF bypasses. To keep the supply voltage from appearing across exposed contacts, a female socket is used at J1. All components are available from Digi-Key. Point-to-point wiring on perfboard is used.

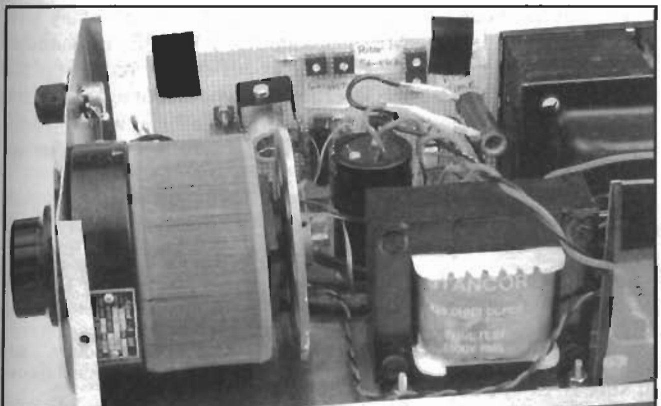

Figure 12—The keyed power supply with the cover removed, viewed from the side. T1 is mounted on the back plate, along with the fuse and connectors for the dc output, a keying line to the NorCal 40A and the ac line input. The **TUNE** switch keys the amplifier for testing. The variable autotransformer is on the left, mounted to the front panel. In the center foreground, mounted to the base plate, is L1, the 25-mH choke. Mylar sheets are taped to the perfboards for electrical isolation. The large components are attached directly to the box. The IRFP340 is mounted on the bottom of the box using a Berquist K10-104 Kapton insulating pad, #6-32 screw and a torque of 10 inch-pounds. The other shaping circuit components are mounted on perfboard. Layout is not critical.

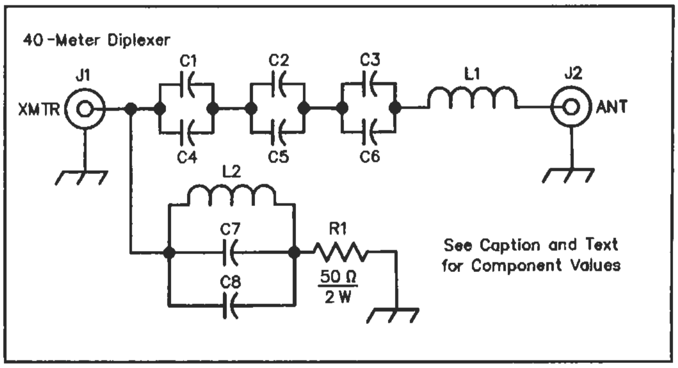

Figure 13—Schematic of our diplexer. It uses six 150 pF capacitors, two 910-pF capacitors and air-wound inductors made of $^3/_{16}$-inch-wide copper tape (see Note 1).

C1-C6—150 pF, 1 kV mica 5%-tolerance (Cornell Dubilier type CDV19; available from Newark Electronics, see Note 1)

C7, C8—910 pF (same type as above)

L1—24 turns of $^3/_{16}$-inch-wide copper tape, 1$^3/_8$ inch ID; see Note 1.

L2—4 turns of $^3/_{16}$-inch wide copper tape, 1 inch ID; see Note 1.

R1—50-Ω, 2-W (approx) termination, Jameco part number 71458; available from Jameco Electronics, 1355 Shoreway Rd, Belmont, CA 94002, tel 415-592-8097; fax 415-592-2503 and 415-595-2664)

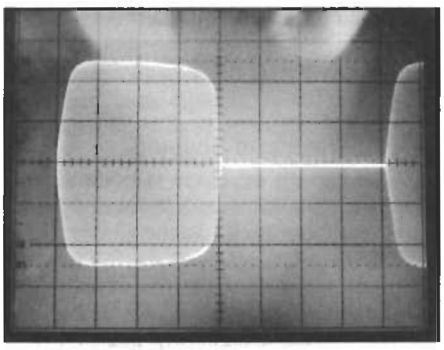

Figure 14—Keying waveform of the 500-W amplifier at 30 WPM. The horizontal axis is 10 ms per division. The rise and fall times are about 3 ms.

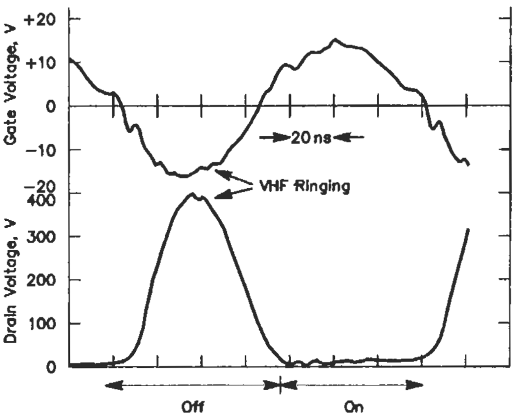

Figure 15—Drain waveform and ringing. Oscilloscope trace of the gate and drain voltages of the 300-W amplifier with 3 W drive. The dc supply voltage is 120 V; the input SWR is 1.6:1. Time scale is 20 ns per division; the transistor off and on times are shown. The transistor is off when the gate voltage is below the threshold, typically 4 V, and on when the gate voltage is several volts above the threshold. Peak gate voltage is 16 V, and the peak drain voltage is 400 V, safely within the manufacturer's ratings of 20 and 500 V, respectively.

pulse. We recommend using a keyer with adjustable weighting to offset the pulse stretching. The potentiometer settings interact somewhat, and there are variations at different sending speeds, so it is best to set them at the speed you commonly use. Adjust the controls for rise and fall times between 2 and 5 ms, and for a smooth keying envelope. Figure 14 shows keying at 30 WPM with rise and fall times of about 3 ms.

VHF Ringing

In many Class-E amplifiers, ringing in the VHF range can be seen on the gate and drain waveforms (Figure 15). This ringing can be quite pronounced, with bumps several volts high on the gate or drain or both. The bumps disappear when the RF input is removed, and that is why we refer to this as *ringing* rather than *oscillation*. We have compared measured spectral plots with *PSPICE* simulations and believe that the waves are driven by the sudden turn-on and turn-off of the transistor, acting rather like the gong of a bell.

We notice two distinct time periods and frequency ranges for the ringing. During the time the transistor is on, the ringing frequency is about 80 MHz. This appears to be a resonance of the external drain capacitor combined with the internal inductance of the capacitor and the transistor and may indicate a mismatched load. The on-ringing is usually small if the load is matched so that the drain voltage comes smoothly to zero before the transistor turns on.

The ringing while the transistor is off covers a broad range of frequencies—from 130 to 210 MHz. If the output low-pass filter is removed, the ringing can be seen easily on a spectrum analyzer at levels between −40 and −60 dBc. The off-ringing appears to be caused by a resonance of the external drain capacitor and its internal inductance, together with the transistor's internal drain capacitance and its inductance. The internal drain capacitance varies greatly with the drain voltage, so that the frequency is modulated as the drain voltage rises and falls. The low-pass filter reduces these harmonics to the −70 to −80 dBc range.

On the Air

The amplifiers meet the FCC requirements for spectral purity [confirmed in the ARRL Lab—*Ed.*]. NorCal 40A designer Wayne Burdick, N6KR, emphasizes that it is important to correctly tune the bandpass filter following the transmit mixer to minimize spurious emissions from the NorCal 40A.[11]

These amplifiers are excellent for chasing DX, schedules and "ragchews," particularly at this low point of the sunspot cycle. The amplifiers require no warm-up, no tune-up, and produce no fan or relay noise. We can vary the power from 1 W to full power via the variable autotransformer. The antenna SWR should be 1.5:1 or better because the amplifier is not protected against large mismatches. With a high SWR, the transistor will probably overheat. Check the dc voltage when the amplifier is delivering full power to the antenna to ensure that it remains between 115 V and 120 V. Readjust the coils if the voltage is too high or too low.

Our most common problem has been a poorly mounted PA transistor. If the transistor is not flat against the heat sink, heat transfer is poor and the transistor becomes quite hot; efficiency suffers (becoming usually less than 85%) and the amplifier may not reach full power. The output power may also drift downward, a sign that the transistor temperature is increasing and that the transistor is under stress.

Component temperature can be a good diagnostic tool. For the 500-W amplifier, our experience is that for CW QSOs longer than 30 minutes, the temperature of C3, C5 (Figure 5) and the heat sink rises to about 60× C. For the 300-W amplifier, the heat-sink temperature is about 50× C. This is hot to the touch, and it can be checked with a lab thermometer. The temperature will vary according to your operating style, and if the temperature is higher than you like, you can add a fan.

The Future

We see room for improvements: protection against antenna mismatches and employment of an inexpensive keyed *switching* power supply that is as lightweight as the amplifiers. Finally, it would be interesting to develop Class-E amplifiers for the other bands. We have built a 250-W amplifier for the 20 meter band that exhibits an efficiency of 88% with 10 W drive. We believe that Class-E amplifiers provide amateurs with good building challenges and operating fun at modest cost.

Acknowledgments

This work began as senior theses by Joyce Wong and Meng-Chen Yu at Caltech, and many of their ideas appear here. John Davis was supported by a scholarship from the James Irvine Foundation and the Army Research Office. Eileen Lau, Kai-Wai Chiu, and Jeff Qin received Caltech Summer Undergraduate Research Fellowships funded by the ETO Corporation and the Caltech Gates Grubstake Fund. We appreciate the help from all the people at the ETO Corporation, particularly Dick Ehrhorn, W0ID; Don Fowler, W1GRV; Tim Coutts; Chip Keen and Frank Myer. Mitsu Sakamoto, JA4FVE, of Kurashiki, Japan, sent us off-the-air recordings over the Internet. We would like to thank Wayne Burdick, N6KR; Wes Hayward, W7ZOI; and Bill Bridges, W6FA, for their advice.

Notes

[7] Eileen Lau, KE6VWU, et al, "High-Efficiency Class-E Amplifiers—*Part 1*," *QST*, May 1997, pp 39-42. The amplifier enclosures are 3$^1/_8$×4$^1/_8$×7$^1/_8$ HWD.

[8] Rick Lindquist, KX4V, "Low-Power Transceiver Kits You Can Build," Product Review, *QST*, Jun 1996, pp 45-50.

[9] See J. B. and R. V. Heaton, "An Electro-Acoustic CW Filter," *QST*, Apr 1983, pp 35-36; Wally Millard, K4JVT, "A Resonant Speaker for CW," Hints and Kinks, *QST*, Dec 1987, p 43; and Richard Clemens, KB8OAB, "More on Resonant Speakers," Hints and Kinks, *QST*, Jan 1989, p 37.—*Ed.*

[10] David Newkirk, WJ1Z (now W9VES), and Rick Karlquist, N6RK, "Mixers, Modulators, and Demodulators," *The 1996 ARRL Handbook*, Chapter 15, p 15.22.

[11] Telephone conversation with Wayne Burdick. This tuning is done by carefully adjusting C50 to peak the output power.

By William E. Sabin, W0IYH

From *QEX*, November 1999

A 100-W MOSFET HF Amplifier

We had the power supply in Mar/Apr[1] and the diplexer filters in Jul/Aug,[2] here's the main event: a reliable FET power amplifier that needs only 10 dBm of drive to produce a pristine 100 W output.

The two-stage amplifier described in this article and shown in Fig 1 is intended for SSB/CW/Data operation on all nine HF amateur bands. An input (R_{in}≈50 Ω) of about 10 mW (+10 dBm) is amplified to 100 W (PEP or average), continuous duty, with a gain of 40 1.0 dB from 1.8 to 29.7 MHz. Third-order, two-tone intermodulation distortion (IMD) products are 35 to 40 dB below 100 W, and higher-order products are also within the high-quality range for amateur SSB equipment, as shown in Fig 2. The main goal for this amplifier is to operate as a driver, with low adjacent-channel interference, for a legal-limit 1500-W linear amplifier. At this power level, adjacent-channel reduction is especially important. And, of course, it is used in the barefoot mode as well.

The power supply (40 V at 8 A) for the push-pull, class-AB MRF150MP (matched-pair) MOSFET output stage was described previously (see Note 1). Six diplexer filters (see Note 2) also known as invulnerable filters provide more than adequate harmonic attenuation for all nine HF bands. They present a broadband load impedance to the MOSFETs that helps to assure freedom from regeneration and oscillation, and good IMD performance. A resistive load impedance between 45 Ω and 55 Ω is recommended for best performance.

The MRF 150 was chosen because it is designed for linear, class-AB SSB operation and because it has high gain (gm) at the 30-MHz end of the HF spectrum. A desirable feature of the MOSFET power transistor is its ability to achieve low values of the higherorder IMD products.[3,4] As mentioned previously, these products contribute to adjacent channel SSB interference. The first stage uses high-gain, class-A push-pull MRF426 (matched-pair) BJTs that require 13.5 V at about 1.0 A from a separate supply. This supply is also the main supply for other system components. Matched pairs of both transistors are available for a small extra fee from at least two sources.[5,6] They are both listed in the current Motorola manual.[7]

One main idea for this amplifier is to operate it in a very low-stress manner that helps assure a low probability of failure for a very long time, which offsets the initial cost of the high-quality transistors. The MRF150 is a 50-V transistor operated at 40 V; the MRF426 is a 28-V transistor operated at 13.5 V. The required input level is low enough that most of the amplification at the signal frequency occurs in one gain block. Because of the good layout, circuit design and decoupling, the 40-dB gain value does not result in any stability problems.

The balanced amplifier greatly reduces even-order harmonics—especially the second—prior to any output filtering, as shown in Fig 3 for a 7.0-MHz signal. It should be 40 dB or more below a 100-W CW signal for each amateur band. This reduction has been found reliable, once achieved. The lowlevel signal source that drives this amplifier must have at least 50 dB of second-harmonic attenuation, since this amplifier will not suppress that harmonic. This is easy to accomplish, but must be considered during the equipment system design (see Fig 9) and while bench-testing the amplifier as shown in Fig 4.

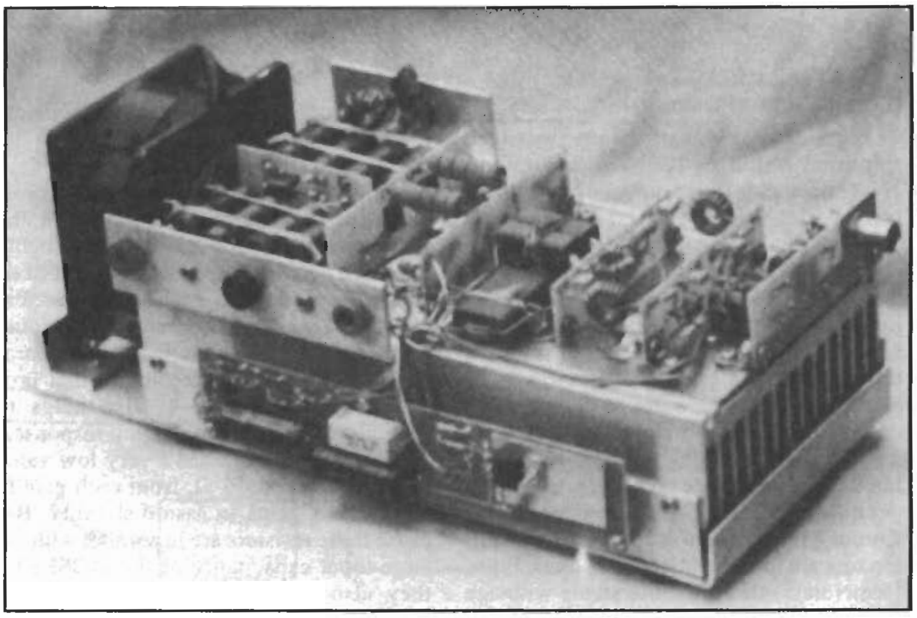

Fig 1—The 100-W broadband amplifier.

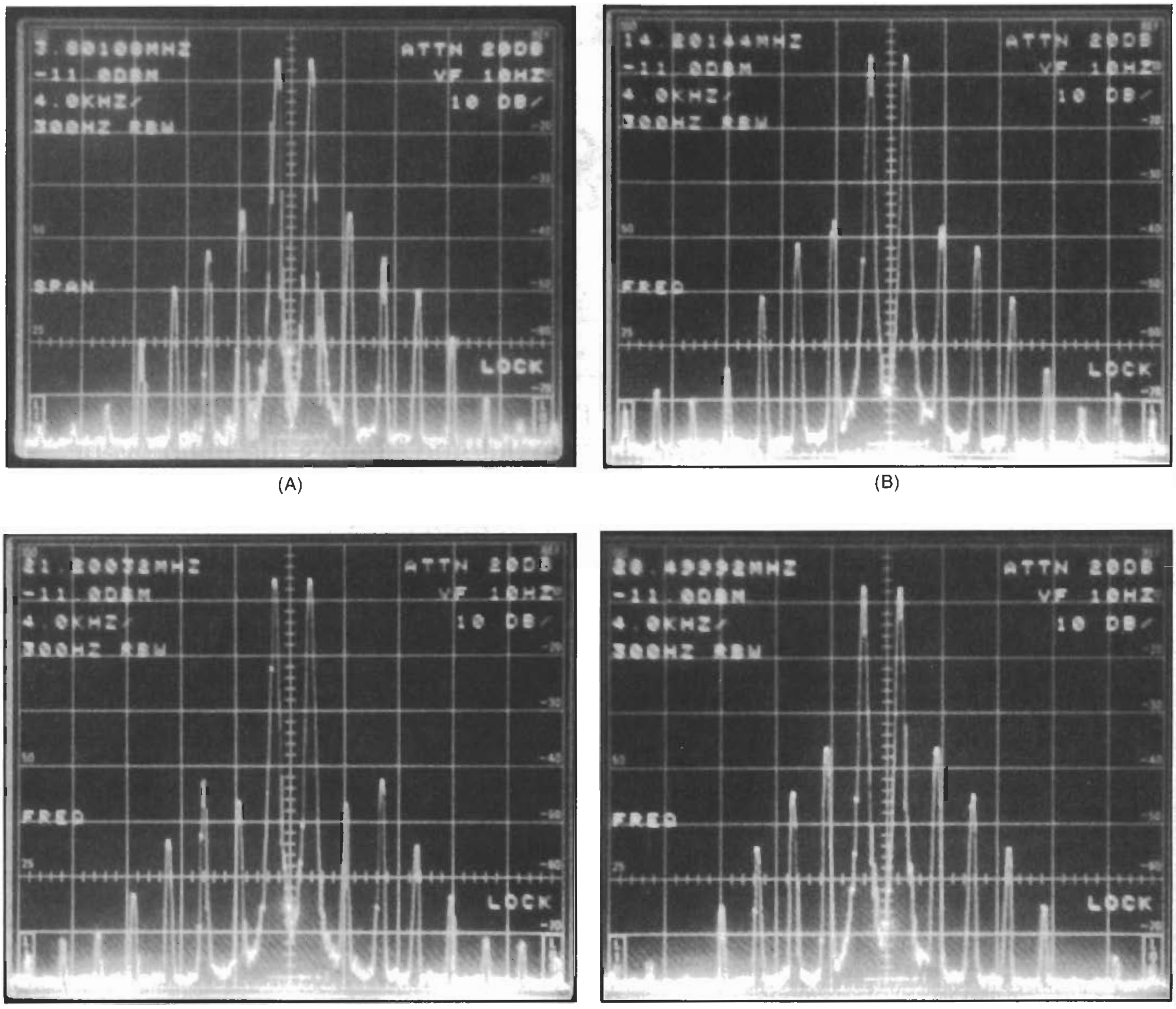

(A)

(B)

(C)

(D)

Fig 2—Two-tone IMD products: (A) 3.8 MHz, (B) 14.2 MHz, (C) 21.2 MHz, (D) 28.5 MHz.

Circuit General Discussion

Fig 5 is the schematic of the two-stage amplifier. It utilizes 1:1 choke baluns and 1:4 (impedance) step-up and step-down transmission line (Guanella) transformers. The choke baluns significantly improve the balance of the input and output stages. Notice also that T2, T3 and T4 have floating center taps rather than bypasses to ground. This is recommended to improve the even-harmonic balance[8] (verified). T4 and T5 run slightly warm as compared to conventional transformers that get quite hot. The 5-W feedback resistors get warm, but their large surface area limits their temperature rise. They also receive cooling air from the fan. All power resistors in the RF circuits are the excellent metal-oxide types (see Note 5) that are quite stable with age and temperature, and have very low reactance at 30 MHz (measured). Metal-film 1% resistors are used in several critical locations.

The first stage has resistance and inductance loading from collector to collector. A powerful free-running oscillation in the first stage at about 32.5 MHz was being triggered occasionally at the moment of dc supply turn-on while a fairly large 28.0 to 29.7 MHz input was present. The loading reduces stage gain and suppresses the parasitic collector circuit resonance that produced the oscillation. It also helps to flatten the amplifier frequency response.

The second stage has a very low value of RF resistance, 15 Ω, from each gate to ground that helps to assure stability. Because these resistors are in parallel with the large input capacitance of the MOSFETs, they also help to flatten the frequency response. Both stages have negative feedback networks that further assure stability and flatness of frequency response.

Two biasing networks are employed. The LM317 provides a highly regulated gate voltage for the second stage. The FET dissipation and linearity are very sensitive to this voltage; the bias voltage for SSB is fine-tuned for best signal purity using two-tone tests and a spectrum analyzer. Note the four 562-a resistors. If the LM317 fails short circuit, the voltage on the FET gates does not exceed 6.8 V, which will not damage the gates. The large drain current that results from this failure also does no damage because the 40-V power supply has current limiting and voltage fold-back that prevent harmful FET dissipation. The FETs are thus kept well within the safe-operating-area (SOAR) as defined in data sheets[9] (verified). It is important that each 56241 resistor from gate to ground be permanently attached directly between the gate and source tabs of the FETs themselves so that the gates are never floating. Wrap the resistor leads around the tabs so that they cannot come loose. This avoids accidental static charges that might ruin them. On the other hand, it is better to tack and not wrap the base-collector and gate-drain signal leads so that they can be easily disconnected. We want to be able to test individual segments of the circuitry easily. In my experience, the MRF150 has proven to be a rugged transistor, much more so when operated conservatively and with the power-supply safeguards that I mentioned.

The LM317 provides about 5.8 V for the particular pair of FETs that I used. Individual pairs of FETs will probably require an adjustment of this voltage for idling current and for IMD products that resemble Fig 2. The adjustment procedure is described later. This bias value is set to emphasize the reduction of the higher-order products rather than third- and fifth-order products. These lower-order products do not contribute as much to adjacent channel interference; in fact, an SSB speech processor will make them worse anyway. The higher-order products need to be reduced and the MRF150 has this capability.

The 2N3906 PNP transistor is a bias-current source. The value of this current is determined by the 4.7-Ω resistor and the base-to-ground voltage of the 2N3906. The two 562-Ω resistors force equal base currents into the MRF426s. This helps to assure equal performance. The 0.47-Ω resistor is part of a negative-feedback bias loop. If collector current increases, the voltage across this resistor increases and this reduces the base current. Thermal runaway is avoided by this strategy because the reduced base bias restricts the current increase to a small value. Because of the low dissipation of the MRF426s and the good heat sink, this method is very effective. These transistors are also well inside their SOAR requirements. This stage is very linear and contributes almost nothing to the overall IMD products. To verify this, it is necessary to turn off the drain voltage to the MRF150s and connect a spectrum analyzer across one of the 15-Ω resistors.

The diplexer filter method was decided uponafter a lot of experimentation with other low-pass filter methods as a solution that is free of problems caused by complex interactions between the filters and the output transistors. A special peculiar problem is discussed later. The MRF150 has high gain into the VHF region; the diplexers eliminated all problems associated with this fact. This approach is recommended as a simple way to assure correct operation for HF amateur-band operation of these high-frequency MOSFETs. I was able to get good enough operation with the more conventional low-pass filters, but this approach was by far the most satisfactory for an amateur-band amplifier, as confirmed by swept-frequency tests at all power levels into a 50-Ω load. More about complex loads later.

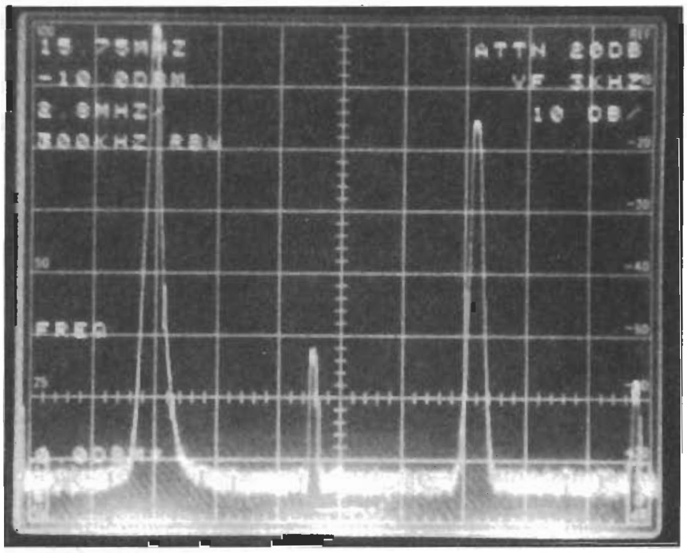

Fig 3—Wideband spectrum for a 7-MHz signal.

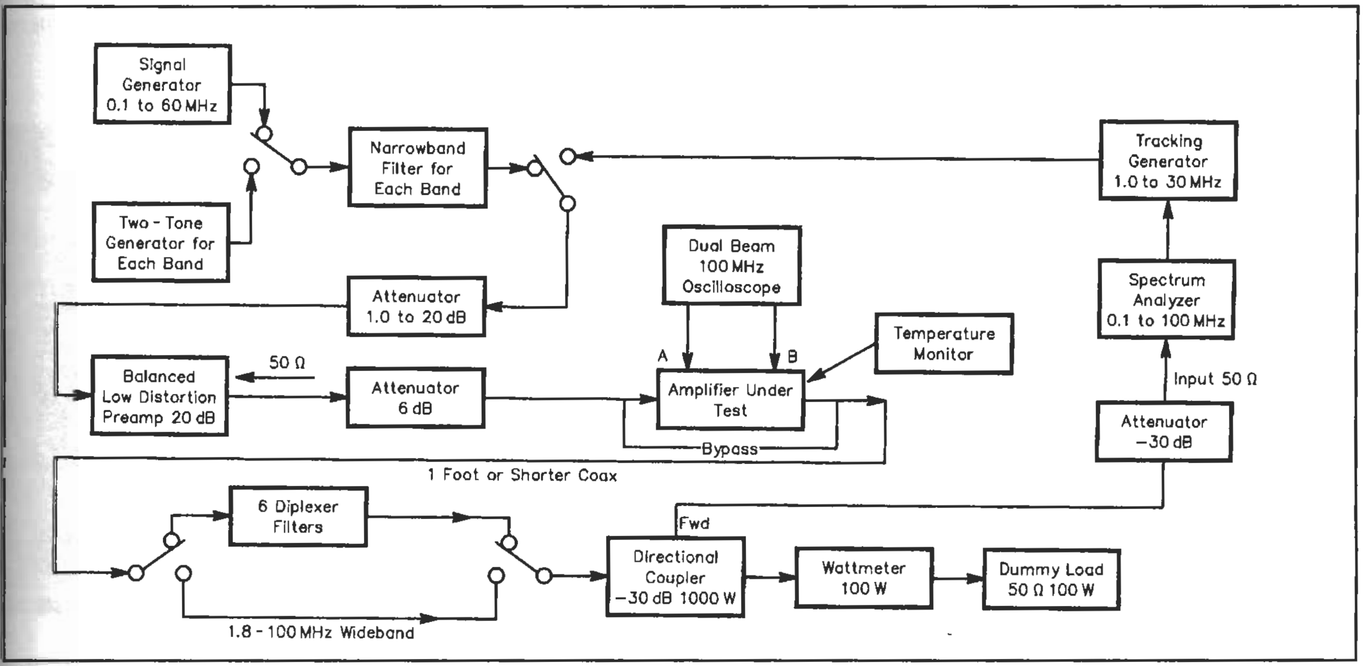
Signal Generator 0.1 to 60 MHz
Two-Tone Generator for Each Band
Narrowband Filter for Each Band
Tracking Generator 1.0 to 30 MHz
Dual Beam 100 MHz Oscilloscope
Spectrum Analyzer 0.1 to 100 MHz
Attenuator 1.0 to 20 dB
Temperature Monitor
A
B
Input 50 Ω
50 Ω
Balanced Low Distortion Preamp 20 dB
Attenuator 6 dB
Amplifier Under Test
Attenuator −30 dB
Bypass
1 Foot or Shorter Coax
6 Diplexer Filters
Fwd
Directional Coupler −30 dB 1000 W
Wattmeter 100 W
Dummy Load 50 Ω 100 W
1.8 - 100 MHz Wideband

Fig 4—Block diagram for equipment setup, design and testing of the MOSFET amplifier.

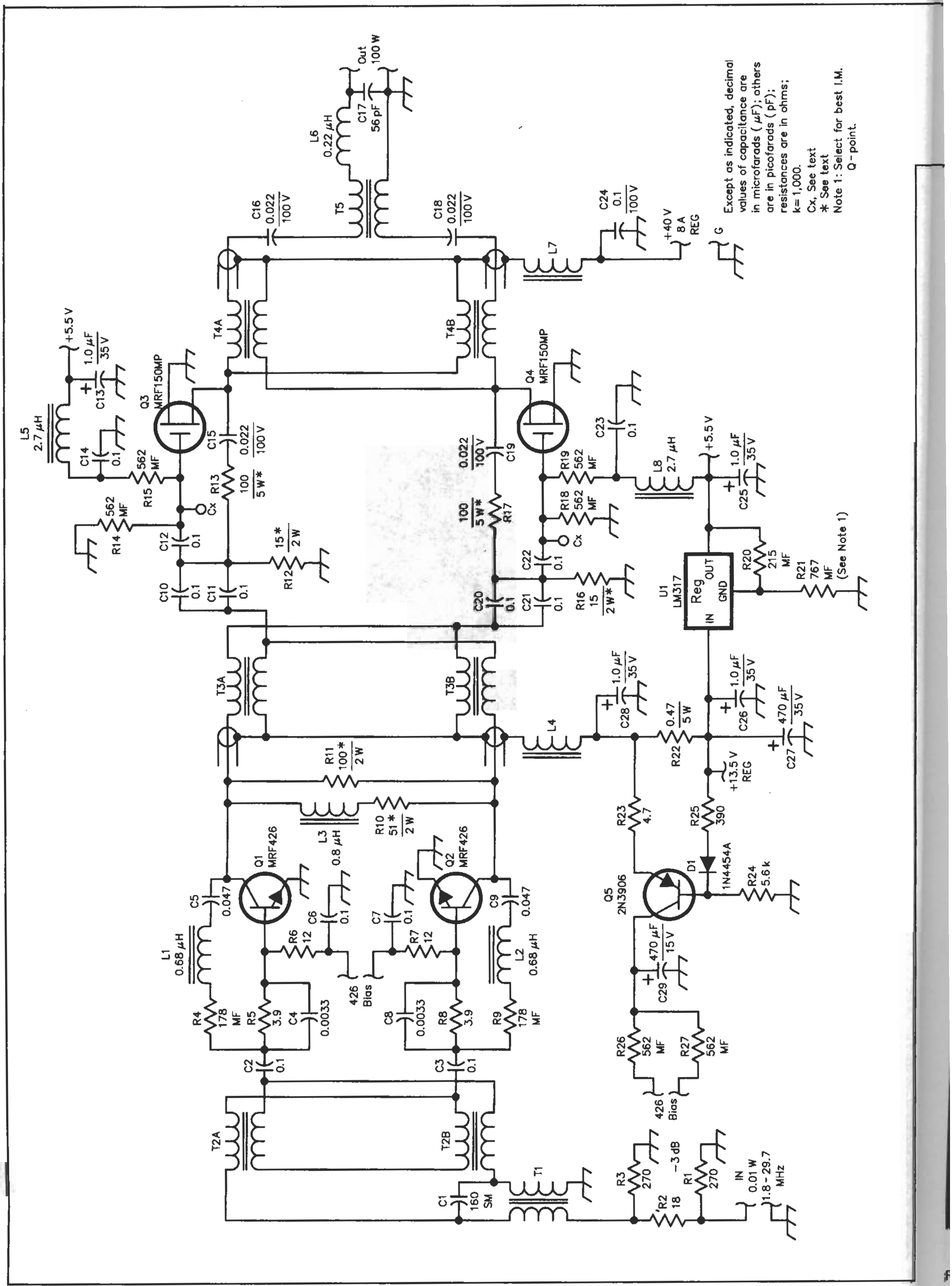

IN
0.01 W
1.8-29.7
MHz
R1
270
R2
18
R3
270
-3 dB
T1
C1
160
SM
T2A
T2B
C2
0.1
C3
0.1
R4
178
MF
R5
3.9
C4
0.0033
R6
12
C6
0.1
426
Bias
C7
0.1
R7
12
C8
0.0033
R8
3.9
R9
178
MF
L1
0.68 µH
C5
0.047
Q1
MRF426
Q2
MRF426
C9
0.047
L2
0.68 µH
L3
0.8 µH
R10
51*
2 W
R11
100*
2 W
T3A
T3B
L4
C28
1.0 µF
35 V
R22
0.47
5 W
C26
1.0 µF
35 V
C27
470 µF
35 V
+13.5 V
REG
R23
4.7
R25
390
D1
1N4454A
R24
5.6 k
Q5
2N3906
C29
470 µF
15 V
R26
562
MF
R27
562
MF
426
Bias
C10
0.1
C11
0.1
R12
15*
2 W
C12
0.1
R14
562
MF
R15
562
MF
Cx
R13
100
5 W*
C15
0.022
100 V
C14
0.1
L5
2.7 µH
C13
1.0 µF
35 V
+5.5 V
Q3
MRF150MP
C20
0.1
C21
0.1
R16
15
2 W*
C22
0.1
Cx
R18
562
MF
R19
562
MF
R17
100
5 W*
C19
0.022
100 V
Q4
MRF150MP
C23
0.1
L8
2.7 µH
+5.5 V
C25
1.0 µF
35 V
U1
LM317
Reg
IN
OUT
GND
R20
215
MF
R21
767
MF
(See Note 1)
T4A
T4B
C16
0.022
100 V
C18
0.022
100 V
T5
L6
0.22 µH
C17
56 pF
Out
100 W
L7
C24
0.1
100 V
+40 V
8 A
REG
G

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.
Cx, See text
* See text
Note 1: Select for best I.M. Q-point.

Fig 5—(left) Schematic diagram of a two-stage, 100-W amplifier. Unless otherwise specified, use ¼ W, 5%-tolerance carbon composition or film resistors. Resistors marked with an asterisk are metal-oxide units. MF indicates metal-film resistors. Both the MRF426 and MRF150 devices are matched pairs. All capacitors are 50 V, unless otherwise indicated.

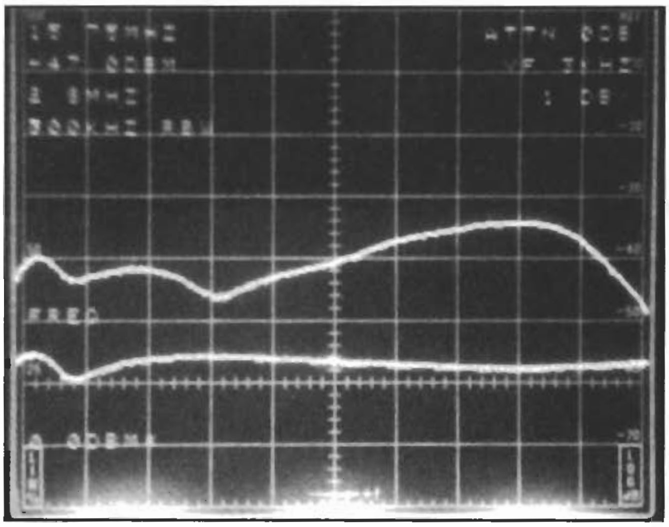

Fig 6—Gain variation, 1.8 MHz to 29.7 MHz, 1.0 dB per division vertical scale—The lower trace is a reference sweep that bypasses the 100-W amplifier—The response bump at low frequency is an artifact of the tracking generator.

Frequency Compensation

The bipolar transistors have gain values that decrease as frequency increases. The MOSFETS have large capacitances that also affect frequency-response roll-off. It was a major exercise to design networks that flatten the response from 1.8 to 29.7 MHz. Fig 5 shows the approach. The idea is to compensate smoothly from input to output in such a way that neither of the two stages is over-driven at any frequency. The criterion for this is to check IMD products and harmonics during the design process, which involves approximate analysis (see the MOSFET Stage Simulation sidebar) and negative feedback.[10] Beyond about 32 MHz, the gain falls off fairly rapidly (but not too rapidly). This is also desirable.

When testing the frequency response using the test setup of Fig 4, it is necessary to measure the frequency response of the signal path from tracking generator to spectrum analyzer while bypassing the 100-W amplifier. This reference response is then compared with the response with the amplifier in the path, as shown in Fig 6. Note the vertical scale: 1.0 dB per division. For the most credible results and ease of measurement, it is very desirable that the impedance looking back from the input be 50 Ω.

Signal Level Testing

We want to verify that the first stage is operating normally by measuring its RF voltages at 4.0 MHz. The class-A first-stage input impedance is close to 50 Ω, and the input level is +10 dBm. Temporarily disconnect the signal leads to the gates of the FETs. The voltage at each output of T3 is about 1.35 V. Reconnect the gate signal leads. The final output over the entire range into an unfiltered, wide-band 50-Ω load is then observed using the setup in Fig 4. A spectrum analyzer with tracking generator is very valuable for this. The accuracy of the wattmeter should be verified or calibrated by some means at both the 100-W and the 25-W levels (the power of each tone of a two-tone, 100-W PEP signal) in each amateur band.

If the second stage is working correctly, the output power is 100 W, or 71 V RMS across 50 Ω. These procedures assure that both stages are working properly and amplifying as intended. Because of the variations in the fabrication of the transistors, the total gain can vary a decibel up or down despite the use of negative feedback in each stage. I suggest the 3-dB attenuator at the input not be modified for simplicity reasons.

The drain-to-drain load impedance of the class-AB output stage is 12.5 Ω and the CW output power is 100 W, so the drain-to-drain ac voltage is 35.4 V. Fig 7 shows the dual-trace scope waveforms (chop mode) on each drain, superimposed on the 40-V supply, and also the drain-to-drain waveform that is confined to the linear region. The third-harmonic content is visible. It is interesting to note that although these waveforms show considerable non-linearity, the fundamental component is quite linear with respect to the gate signal level. I also found that a 50-V dc supply created more heat, but at 100W, did not improve linearity enough to make it worthwhile. The 35.4 V ac also appears across the two series-connected 100-Ω feedback resis-

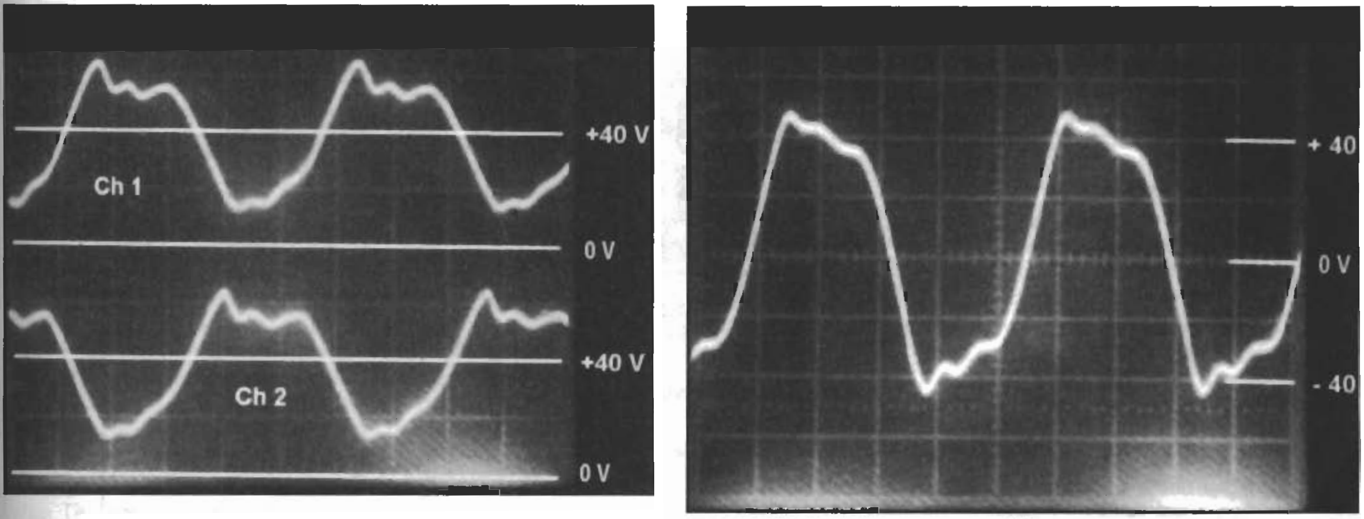

(A)

(B)

Fig 7—(A) drain and (B) drain-to-drain oscilloscope waveforms.

tors; they dissipate about (35.42)/200 = 6.3 W, or 3.2 W per resistor (64% of the 5 W rating).

A broadband, untuned power amplifier with flat frequency response and low distortion is not, by necessity, especially energy-efficient (ratio of RF output power to dc power). The output stage is less than 40% efficient for this reason. For SSB use, where the average power is not more that 25 W, even with speech processing, this is no problem at all. For continuous key-down at 100 W, the cooling fan is more than adequate.

Here is an important caution about using oscilloscope probes at the FET drains and the output connector: A 10:1 probe could be damaged (it happened to me) if used directly at this RF voltage level, especially at the upper end of the HF range. Use instead the homemade probe described previously (see Note 2) with a 50-Ω terminating resistor.

The quality of balance is checked by looking at the second harmonic on a spectrum analyzer using the setup in Fig 4 with the diplexer filters out of the signal path. At 100 W output, check the harmonics in each of the nine amateur bands. Capacitor C_x in Fig 5 is used (if needed) to improve the secondharmonic phase balance on the higher amateur bands. If the value is not well below 40 dBc, use C_x at one gate or the other to achieve 45 dBc. Extra pads are provided on the PC board for this capacitor. A value of between 22 pF and 56 pF should be adequate. Overkill is neither necessary nor desirable: The output filters will do the rest. For each decibel of power output below 100 W, the second harmonic normally drops about two decibels.

Construction Notes

Fig 1 shows how my version of the amplifier is constructed. A 0.125-inch (or 0.062-inch) double-clad PC board, 3.25 × 8.0 inches, is firmly attached to the heat sink. I suggest using this compact size for best reproducibility. The heat sink shown may not be presently available from RF Parts (see Note 5), but I also purchased a Model 99 sink from CCI (see Note 6) that is 6.5×12 inches. This can be easily tailored to the appropriate size with a band saw that has a metal-cutting blade. Pieces of angle and sheet aluminum can then be creatively fashioned to accommodate the fan (RadioShack #273-242) and the PC board that contains the bias circuitry. The transistors are bolted directly to the heat sink (through cutouts in the PC board) using heat-sink compound and tapped (and carefully deburred) #4-40 holes. Careful mounting of the transistors is essential. The main PC board has small sections of copper removed underneath the base and collector tabs of the MRF426s and underneath the gate and drain tabs of the MRF150s, so that accidental grounding is avoided. Use a hobby knife to define the areas and a hot soldering iron to peel off the copper.

A bottom cover is advised, as shown, so that air is funneled through the heat-sink fins efficiently. For a 100-W, continuous, single-tone output, the exhaust air temperature reaches 50°C. I used a RadioShack #22-174 multimeter with its temperature probe mounted inside the fins to monitor temperature during the design and testing phase.

The components are mounted on a set of seven small PC boards, six of which are mounted vertically (as shown) and bolted to the drilled-andtapped heat sink through the PC board, using #4-40 screws and small angle brackets. I used stiff, right angle #6 solder lugs that worked out very nicely. The seven PC boards are cut from a single 4×6-inch two-sided PC board, shown in Fig 8. The bias circuit board has a ground plane; the others do not. This set of boards is available from FAR Circuits.[11]

I chose this method because it is easy, makes the amplifier more compact, reduces stray L and C that can degrade the wideband frequency response and reduces stray couplings that can impair stability and harmonic balance. It also allows the ground plane to be one continuous surface, which is a plus factor. This approach worked out very well and I recommend it as a simple approach.

Temperature Rise

The cooling fan is important. This amplifier, as designed, should have the fan running, and I have found that a simple and reliable way to keep everything safe. It has been tested at 100 W continuously, for sev-

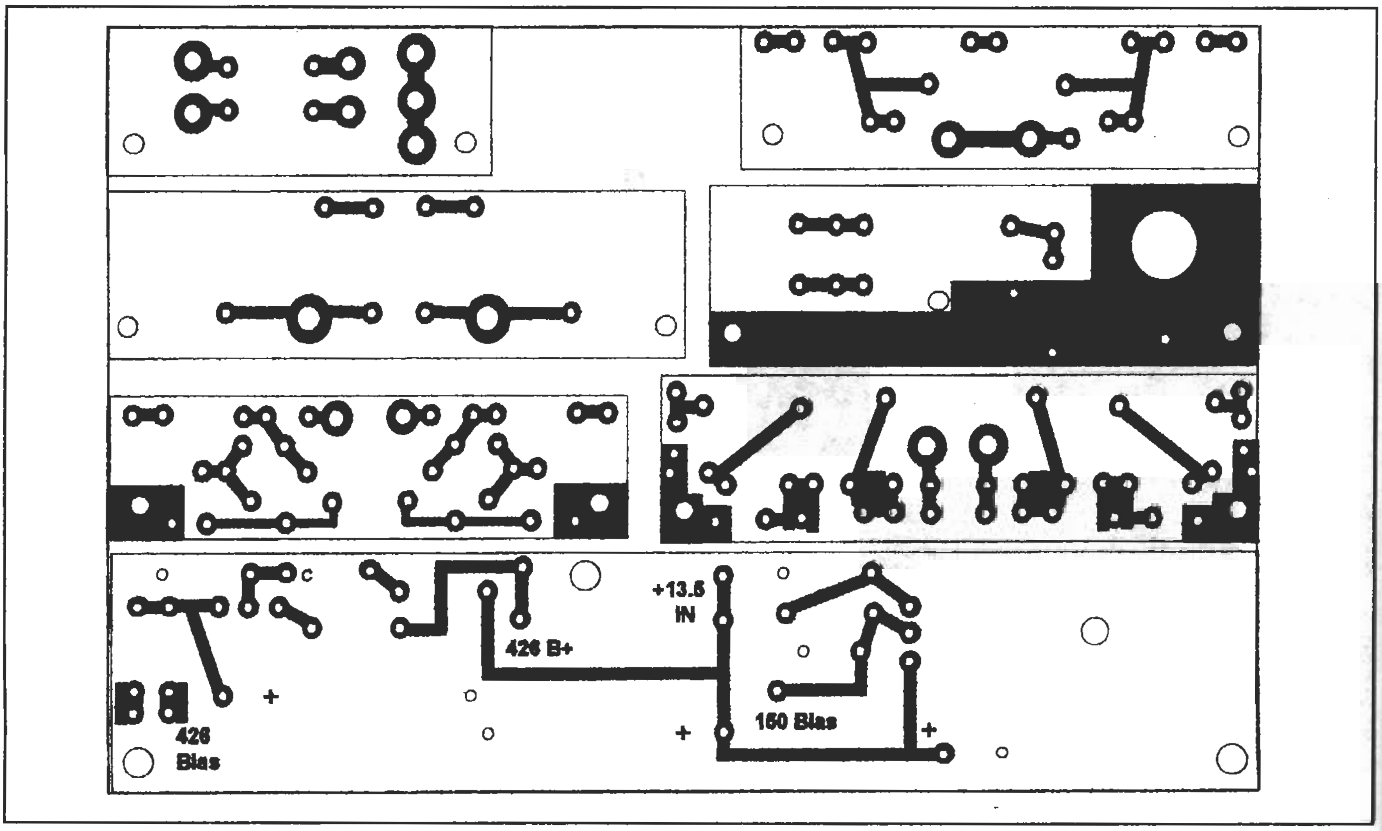

Fig 8—Etching pattern for a 4×6 inch (finished size) PC board that provides the seven individual boards.

MOSFET Stage Simulation

A simplified analysis of the second MOSFET stage is presented (Fig A) to illustrate the effect of the resistive negative feedback of the 100-Ω resistors and the 40-pF gateto-drain capacitance. The simulation diagram shows the voltagecontrolled current sources with a G_m of 6.0 S, which is assumed constant over frequency. The gate-to-source capacitance is 360 pF and the drain-to-source capacitance is 200 pF. These numbers are from the MRF150 data sheets.

The frequency plot (Fig B) obtained from the ARRL Radio Designer program shows the gain, MS21 (dB), with (lower trace) and without (middle trace) the 100-Ω feedback resistors. The gain variation is reduced from about 6 dB to about 1.5 dB. In this simplified model (good enough for this illustration), the gain drop is caused by the capacitors, especially the 40-pF capacitor whose influence is greatly magnified by the Miller effect. If this 40-pF capacitance is eliminated from the simulation, the frequency response variation—even without the feedback resistors—is less than 1.0 dB, as seen on the upper trace.

If the power output at 15 MHz is to be held constant, the turns ratio of the output transformer and the value of the feedback resistors can be manipulated to achieve that end. The simulation shows that a turns ratio of $1:\sqrt{2}$ and a resistor value of 120 Ω does this. The FET load impedance is now 25 Ω instead of 12.5 Ω. The increased voltage gain makes the Miller effect greater and increases the gain variation (lower trace) to 2.3 dB. The drain-to-drain voltage is now 50 V instead of 35 V, and this results in higher drain efficiency for the MOSFETs. However, the push-pull $1:\sqrt{2}$-turns-ratio transmission-line transformer is not as simple as the 1:2 that uses 25-Ω coax, and for me, that was the determining factor.

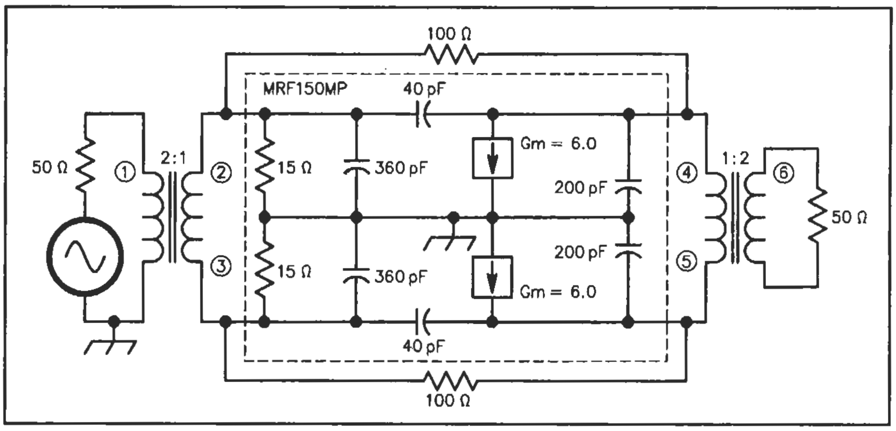

Fig A—A circuit used to analyze the PA output stage with ARD.

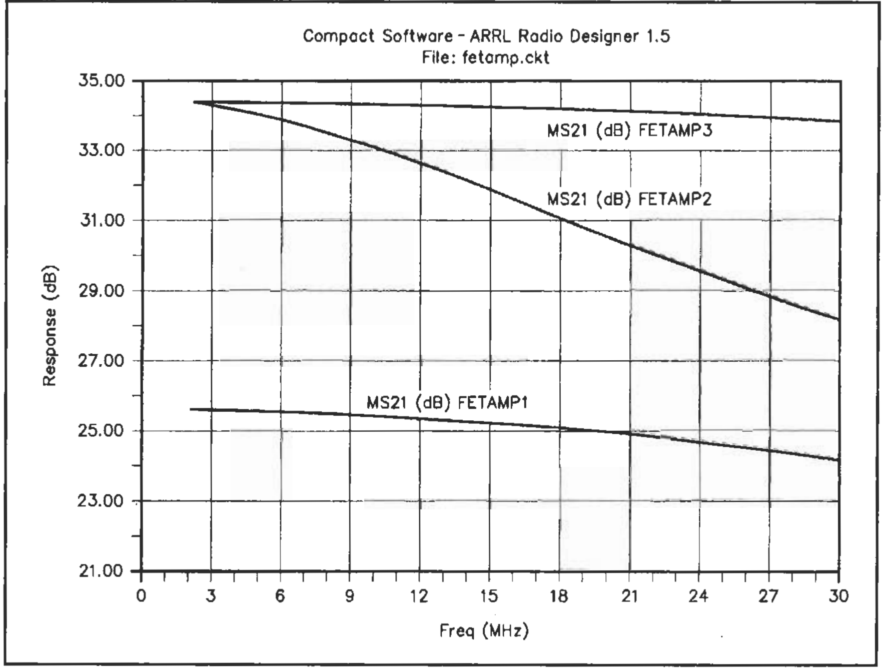

Fig B—The analysis results.

eral hours. The MOSFETswell separated on a good heat sink and with efficient airflow do not have a temperature-controlled gate-bias arrangement because I did not find it necessary at this power level. Because of the spread in threshold voltage of individual matched pairs, a one-time adjustment of gate bias is needed. The following simple procedure is used:

- Replace R21 with a resistor decade box set at 750 11.
- Place a 0-10 A meter in the +40-V line.
- With *no* signal input, switch on the amplifier and let the current reach its final value, which should reach about 1.5 A as the FETs warm up.
- Adjust the resistor value until DD is 1.5 A. After each adjustment, allow time for the FETs to reach a steady current value. This final value is not critical, but should be within 10%.
- Check the two-tone IMD at 100 W PEP (50 W average) and make further small changes, if needed, to resemble the IMD patterns in Fig 2.

In this amplifier, we are able to use the self-limiting feature of the MOSFET. This approach would not be appropriate in higher-power, higher-temperature amplifiers because of thermal runaway possibilities, where the decrease with temperature rise of the gate threshold voltage exceeds the decrease of the dc transconductance.[12,13] If fan reliability is a concern, add a thermal cutout switch to the +40-V line.

Load Impedance

Swept- and fixed-frequency tests at many signal levels and voltage valuesusing various values of parallel conductance and capacitive/inductive susceptancedid not reveal instability problems at SWR values less than 2:1. A tendency to oscillate in a peculiar manner occurred in the 20.0 to 29.7 MHz range, but only with input signal applied (a driven oscillation) and certain values of drive level, supply voltages less than 30 V, SWR of 3:1 or more, and a certain range of shunt capacitance at the output of the amplifier itself or at the output of the 12/10 meter diplexer. The instability shows up as a Christmas tree pattern in which discrete, uniformly-spaced 2- to 4-MHz sidebands appear on the main carrier. It is due to a parametric effect[14] involving the voltage-variable capacitances of the MOSFET transistors that is somehow enhanced by the shunt test capacitance. Shunt inductance did not produce the problem. The amplitude modulation of the carrier by the low-frequency oscillation is clearly visible on an oscilloscope. The low-frequency oscillation itself can be seen on a spectrum analyzer. At SWR values less than 2:1 and drain voltages greater than 35 V (40 V is recommended), the safety margin is large. The 0.22-µH, 56-pF network at the output greatly assisted with this problem. A coax cable from the PA to the diplexer assembly that is a foot or less duplicates my test setup. Under these conditions, there is no stability problem on any band caused by load-impedance values that my testing could identify. Actually, a resistive load between 45 Ω and 55 Ω is recommended for best SSB linearity, as mentioned previously. While driving a high-power linear amplifier to its rated two-tone SSB output, adjust its input impedance to 50 Ω resistive as closely as possible on each band, using a 50-Ω directional wattmeter. A broadband (untuned) solid-state driver amplifier such as this one requires such attention to load value for best results as compared to pi-network vacuum-tube PAs that can transform a fairly wide range of complex load impedance values to the correct tube plate load resistance at resonance.

It is difficult to achieve stability in a broadband, transistor power amplifier. Having done so, it is still wise to operate into the correct 50-Ω load as closely as possible. The monitoring of forward and reflected power in Fig 9 is helpful for this and assures clean, stable and reliable operation. This circuitry will cut back the drive level to a value that protects the output stage from excessive drain current or gate drive. Incidentally, the first stage goes into saturation far below the level that would damage the MOSFETs.

System Design

Fig 9 suggests a system implementation of the amplifier. For the flattest frequency response of the complete system from 1.8 to 29.7 MHz, the circuitry that drives this amplifier should also have a flat response and a 50-Ω output resistance—both easy to achieve. If it is not perfectly flat or not exactly 50 Ω, it may be necessary to slightly adjust the drive level on each band, which is not difficult. As mentioned before, the second-harmonic production of the circuitry preceding this amplifier must be of sufficiently low level (50 dBc) that balanced, push-pull operation is necessary. Narrow-band resonator filters are also very commonly used for amateur-band harmonic reduction in low-level exciter circuitry.[15,16]

The directional coupler detects excessive forward and reflected power. Both of these are displayed on a panel meter. The directional couplers forward port is also a

Parts List

C1—160 pF SM
C2, 3, 6, 7, 10-12, 14, 20-23—0.1 µF, 50-V CK05
C4, 8—0.0033 µF, 50-V CKO5
C5, 9—0.047 µF, 50-V CKO5
C13, 25, 26, 28—1.0 µF, 35-V tantalum
C15, 16, 18, 19—0.022 µF, 100-V CK05
C17—56 pF SM
C24—0.1 µF, 100-V CKO5
C27—470 µF, 35-V aluminum
C29—470 µF, 15-V aluminum
D1—1N4454A or equivalent
L1, 2—0.68 µH, T50-2 core, 10 turns #26 AWG
L3—0.80 µH, T50-2 core, 12 turns #26 AWG
L4—BN-43-3312, 4½ turns #22 hookup wire
L5, 8—2.7 µH molded
L6—0.22 µH, T50-2 core, 5 turns #26
L7—2 FB-43-5621 cores, 1½ turns #12 stranded
Q1, 2—MRF426 matched pair
Q3, 4—MRF150 matched pair
Q5—2N3906 PNP
R1, 3—270 Ω, ¼ W, 5% tolerance
R2—18 Ω, ¼ W, 5% tolerance
R4, 9—178 Ω, ⅛ W metal film, 1% tolerance
R5, 8—3.9 Ω, ¼ W, 5% tolerance
R6, 7—12 Ω, ¼ W, 5% tolerance
R10—51 Ω, 2 W metal oxide
R11—100 Ω, 2 W metal oxide
R12, 16—51 Ω, 2 W metal oxide
R13, 17—100 Ω, 5 W metal oxide
R14, 15, 18, 19, 26, 27—562 Ω, ⅛ W metal film, 1% tolerance
R20—215 Ω, ⅛ W metal film 1% tolerance
R21—767 Ω, ⅛ W metal film 1% tolerance, test selected
R22—0.47 Ω, 5W wire-wound
R23—4.7 Ω, ¼ W 5% tolerance
R24—5.6 kΩ, ¼ W, 5%
R25—390 Ω, ¼ W, 5%
T1—BN-43-202, 2½ turns #32 AWG, bifilar
T2A, B—BN-43-202, 2½ turns #32 AWG, bifilar
T3A, B—BN-43-3312, 2½ turns 25-Ω miniature coax*
T4A, B—(2) FB-43-5621, 1½ turns 25-Ω miniature coax*
T5—(2) FB-43-5621, 2½ turns 50-Ω miniature coax
U1—LM317 adjustable voltage regulator

Notes

*Microdot D260-4ii8-0000 available from Communication Concepts, Inc.
All cores are available from Amidon.
Closely matched transistor pairs are from RF Parts (see Note 5).

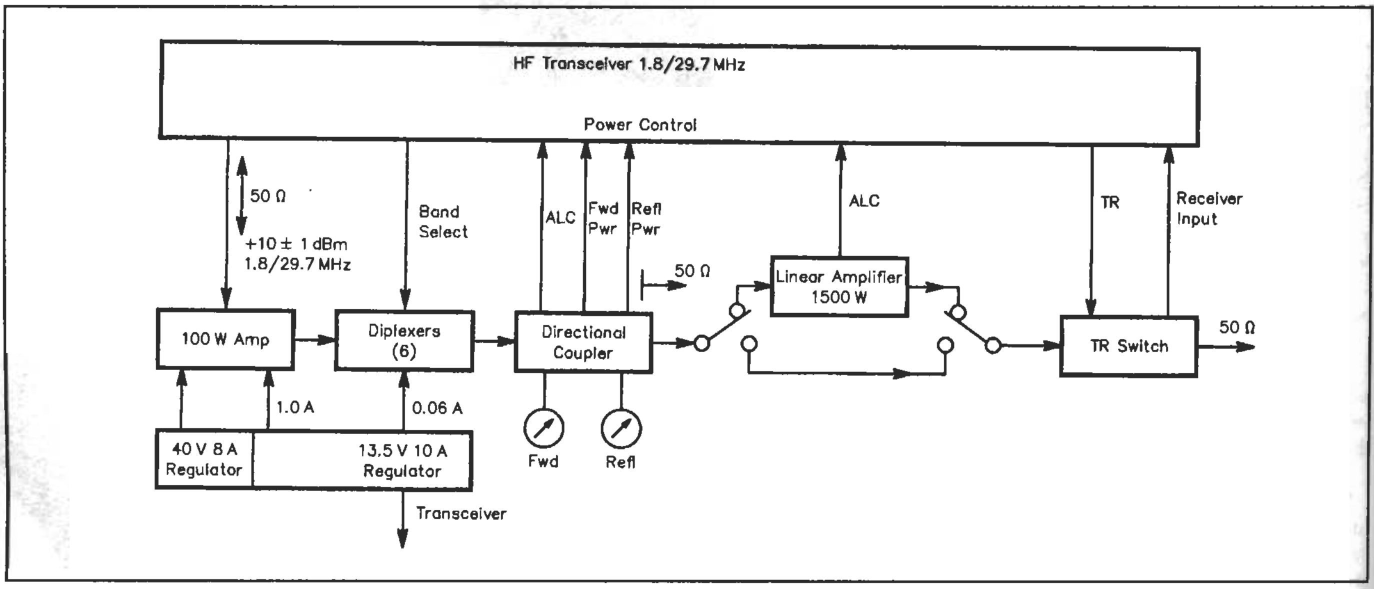

Fig 9—System implementation of the 100-W amplifier.

source of ALC voltage. These control voltages are fed back to the appropriate gain-controlled stages in the exciter. Gain control of the transistors in the 100-W amplifier is not recommended because changes in base or gate bias will degrade amplifier linearity, IMD products and possibly the flat frequency response. There are better ways to accomplish the gain-control task, such as preferably in a lowlevel IF amplifier. With respect to IMD, we are trying to control the odd-order curvature of the MOSFET transfer characteristic—especially for the higher-order products—by setting the gate-bias point. The bias value is found by looking at IMD on each amateur band and selecting the best compromise value. This was discussed previously in this article.

The diplexer filters are linked to the other band-switch circuitry so that the correct filter is always switched in. The TR relay is at the output of the diplexers. Do not hot-switch the diplexer filters because this might damage the inexpensive relays.

Conclusion

In conjunction with the power supply (see Note 1) and the diplexer filters (see Note 2), the amplifier described here is a basic module for homebrew equipment that should satisfy the requirements for a 100-W power level, a 1.8 to 29.7 MHz bandwidth and a high-quality signal. It should run trouble-free for a very long time. The initial cost of the four transistors is about $180, but they will last indefinitely if cared for properly. The ones that I use were abused considerably during the experimentation and continue to work perfectly. The approach that I suggest is best implemented with diplexer filters and a preamplifier that has very low second-harmonic output and low IMDboth easy to get at +10 dBm. The payoff for me is excellent performance and reliability, once the design was completed.

This project is suggested for the homebrew enthusiast who has at least some part-time access to lab-quality test equipment. Others who do not care to build may find the article, together with numerous other sources, interesting background information regarding MOSFET power-amplifier design and test methods.

Any attempt at building and testing this amplifier should also use the 40-V power supply in the referent of Note 1 or something similar. The automatic current limiting at 8 A, the automatic reduction of drain voltage, the short-circuit limiting to 4 A and the manual control down to 24 V help to protect the MOSFETs from mishaps that are bound to occur.

Notes

[1] W. Sabin, W0IYH, "Power Supply for a MOSFET Power Amplifier," *QEX*, Mar/Apr 1999, pp 50-54.

[2] W. Sabin, W0IYH, "Diplexer Filters for an HF MOSFET Power Amplifier," *QEX*, Jul/Aug 1999, pp 20-26.

[3] N. Dye and H. Granberg, *Radio Frequency Transistors* (New York: Butterworth-Heinemann, 1993), pp 42-43.

[4] W. Sabin and E. Schoenike, *HF Radio Systems and Circuits* (Tucker, Georgia: Noble Publishing Co, 1998), Chapter 12, Solid State Power Amplifiers, by R. Blocksome. This chapter is highly recommended regarding solid-state SSB power amplifiers. This book is also ARRL Order No. 7253. ARRL publications are available from your local ARRL dealer or directly from the ARRL. Mail orders to Pub Sales Dept, ARRL, 225 Main St, Newington, CT 06111-1494. You can call us toll-free at tel 888-277-5289; fax your order to 860-594- 0303; or send e-mail to **pubsales @arrl.org**. Check out the full ARRL publications line on the World Wide Web at **http://www.arrl.org/catalog**.

[5] RF Parts Co, 4355 Pacific St, San Marcos, CA 92069; tel 760-744-0700, 800-737-2787 (orders only), fax 760-744-1943; **rfp@rfparts.com**, **http://www.rfparts.com/**.

[6] Communication Concepts, Inc. (CCI), 508 Millstone Dr, Beavercreek, OH 45434; tel 937-426-8600, fax 937-429-3811; **cci.dayton@pobox.com**, **http://www. communication-concepts.com/**; ask for a catalog.

[7] *Motorola Wireless Semiconductor Solutions, Device Data*, Rev 9, Vol. II; Motorola #DL110/D. The data book is available from RF Parts (see Note 5). (A Motorola Selector Guide is available at **http://www.motsps.com/books/sguides/pdf/sg46rev18.pdf**. There is a MRF150 page at **http://www.mot-sps.com/products/rf_and_if/rf_transistors/high_power/power_mosfets/mrf150.html**. A MRF426 datasheet (MRF426/D)is available at **http://www.zettweb.com/CDROMs/cdrom013/pdf/mrf426rev0.pdf**, which is not a Motorola site—*Ed.*)

[8] Dye and Granberg, pp 113-115.

[9] *Motorola Wireless Semiconductor Solutions*; see pp 4.2-133 for the MRF151, similar to the MRF150. (See Note 7 for an MRF150 URL—*Ed.*)

[10] Dye and Granberg, Chapter 12.

[11] The set of boards costs $12 plus shipping and handling. FAR Circuits, 18N640 Field Ct, Dundee, IL 60118-9269; tel 847-836-9148 (Voice mail), fax 847-836-9148 (same as voice mail); **farcir@ais.net**, **http://www.cl.ais.net/farcir/**.

[12] Dye and Granberg, Chapter 4.

[13] H. Granberg, Wideband RF Power Amplifier, *RF design*, Feb 1988; also *Motorola RF Application Reports*, AR313, p 424.

[14] Dye and Granberg, similar top 124, Fig 7.7.

[15] *The ARRL Handbook* (Newington, CT: 1995-2000 editions), ARRL Order No. 1832; pp 17.60–17.66 and Fig 17.69. See Note 4 for purchasing information.

[16] W. Sabin, W0IYH, "Designing Narrow Band-Pass Filters with a BASIC Program," *QST*, May 1983, pp 23-29.

By Jim Valdes, WA1GPO **From *QST*, June 2003**

The FARA HF Project

Would you like an economical and relatively easy to build 30 W HF amplifier? For about $130 worth of parts you can give that QRP rig an extra 12 dB of muscle!

One of the larger and more active radio clubs in the Cape Cod, Massachusetts area is the Falmouth Amateur Radio Association (FARA). Some of its members are affectionately referred to as the "hackers"– those who enjoy the construction phase of Amateur Radio. The renewed interest in low power (QRP) radios in the 2 to 5 W output range resulted in the desire for a 10 to 12 dB gain RF linear amplifier producing 30 to 40 W of output power. Commercial amplifiers capable of being driven with less than 50 W below 30 MHz are prohibited by FCC regulations.[1] As a radio amateur, you are permitted to construct one amplifier per year for your personal use. The amplifier described here is relatively easy to fabricate, given a basic knowledge of electronics and some familiarity with hand tools. This article describes in detail the construction of a 12 dB, nominal 30 W output, 1.8 to 30 MHz RF power amplifier. It is intended for 12 to 14.7 V dc operation, making it ideal for mobile use.

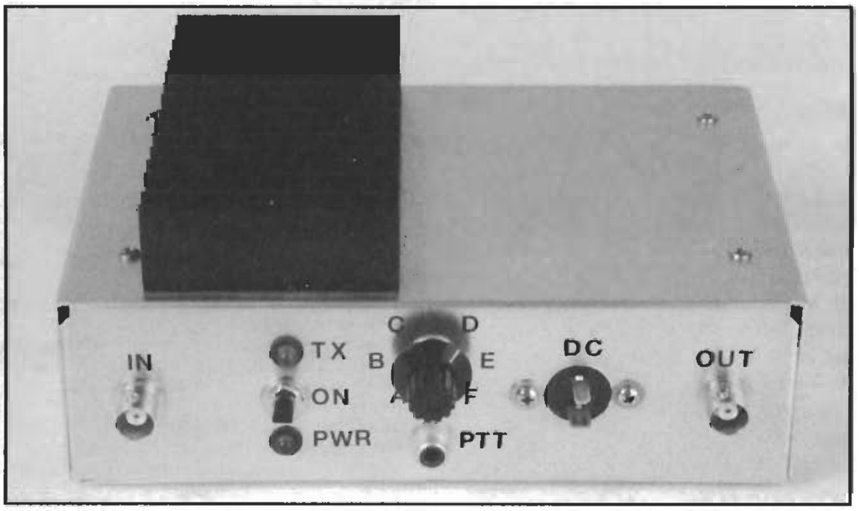

Figure 1—The completed 30 W amplifier. The band switch takes care of output low-pass filter switching.

Amplifier Description

The design of the amplifier is based upon common engineering practices. Ideas gathered from researching the handbooks published by the ARRL and The Radio Society of Great Britain (RSGB), as well as articles published in *QST* and other journals formed the basis of the design. This amplifier uses readily available components...many of the earlier designs were based on the Motorola MRF series of RF devices that are no longer available or are prohibitively expensive.

The amplifier is housed in a 5 × 7 × 2 inch aluminum box. It consists of two stacked circuit boards—an RF amplifier and a low-pass filter. The completed amplifier is shown in Figure 1. Figure 2 is the schematic diagram of the RF assembly together with the amplifier and low-pass filter parts list and Figure 3 is the low-pass filter schematic. The amplifier can be driven by 2 to 5 W at the RF input; an input attenuator consisting of R1, R2 and R3 must be selected (as noted in Figure 2) to ensure the proper drive level (2 W) for the push-pull (2SC2312C device types) class AB amplifier stage. T1 is wound on a small binocular core with a 4:1 ratio as detailed on the schematic. The low impedance secondary is a single turn center tapped; it carries the bias voltage to the output transistors. The bias voltage is derived from the LM317 regulator, which is operated as a switched current source. The LM317 is switched on when the internal PTT line is activated. Bias voltage is developed across the FES8J diode, which is in intimate thermal contact with the output transistors. The output transformer (T2) is also wound with a 4:1 ratio; it has a single-turn sense winding. The single-turn feedback winding provides some degree of negative feedback to flatten the gain and stabilize the input impedance over the HF frequency range. The RC network in the input base circuit establishes the overall gain. The 6.8 W series resistors determine the gain below 14 MHz, while the 4700 pF capacitors are effective above 14 MHz.

Figure 4 shows the output power versus frequency characteristics of the amplifier, including the second and third harmonic response. All measurements were made with an IFR1600S Communications Service Monitor. In all cases, harmonics were greater than 40 dB down referenced to the fundamental frequency (–40 dBc), and the amplifier meets current FCC requirements for spectral purity.[2] Note that the gain starts to decrease above 21 MHz, rolling off from a nominal 30 W to 20 W at 29 MHz. On 10 meters this still represents 10 dB of gain, a worthwhile improvement. The dc voltage to the output transistors is decoupled by the pi network at the center-tapped primary of the output transformer, T2. This winding carries substantial current and should be wound with #18 Teflon covered wire. Any IR drop in the dc circuit will severely degrade the amplifier performance. TR switching is provided by relay K1. The PTT line can be operated manually (pulled to ground) or RF activated switching circuitry can be installed to eliminate the need for external keying controls. The value of C12 (3.3 μF) determines the SSB time constant for the RF activated switch.

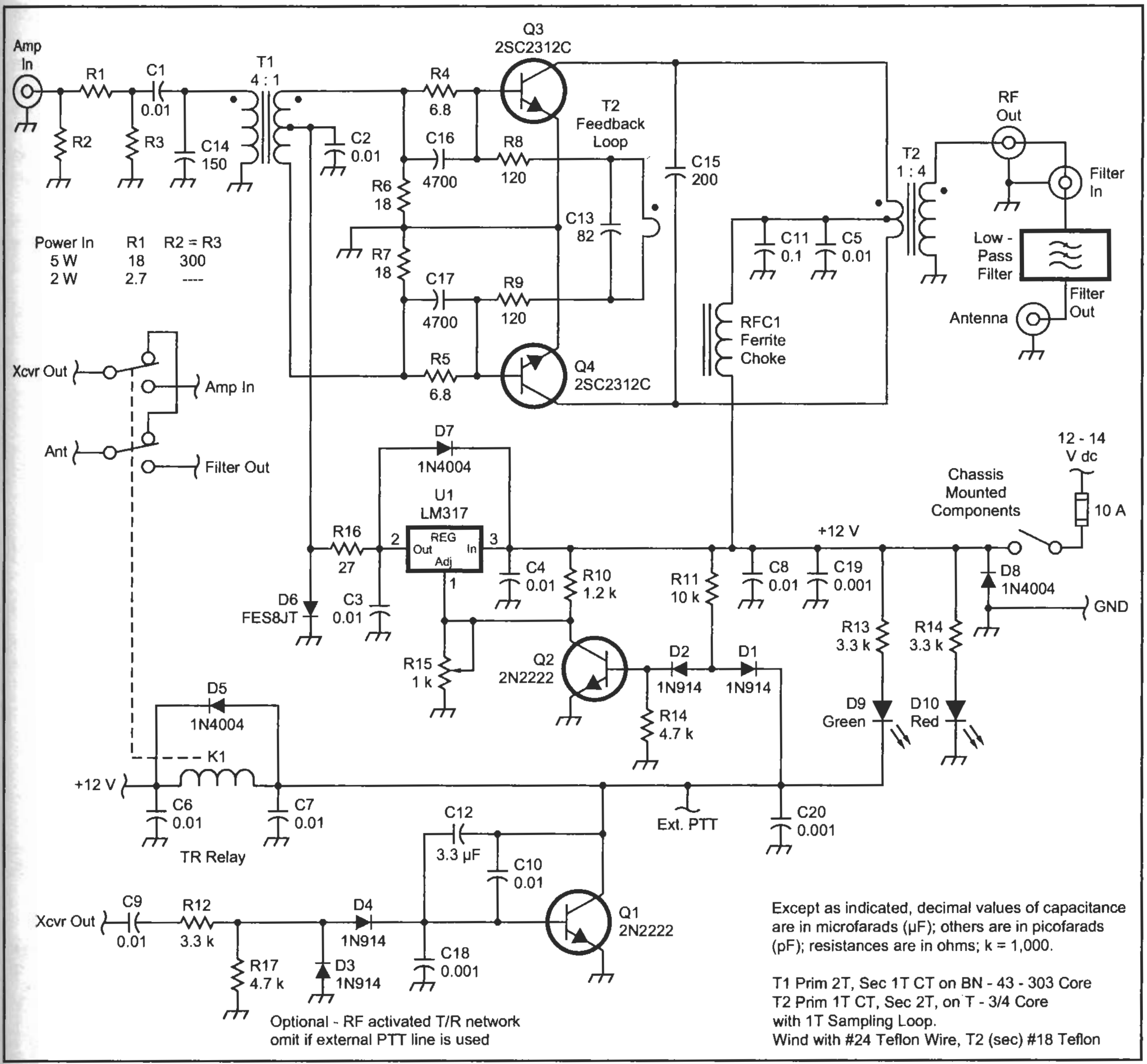
Amp In
R1
C1 0.01
R2
R3
C14 150
T1 4 : 1
R4 6.8
Q3 2SC2312C
C2 0.01
C16 4700
R8 120
R6 18
R7 18
T2 Feedback Loop
C13 82
C17 4700
R9 120
R5 6.8
Q4 2SC2312C
C15 200
C11 0.1
C5 0.01
RFC1 Ferrite Choke
T2 1 : 4
RF Out
Filter In
Low - Pass Filter
Filter Out
Antenna
Xcvr Out
Amp In
Ant
Filter Out
D7 1N4004
U1 LM317
REG Out In Adj
2
3
1
R16 27
D6 FES8JT
C3 0.01
C4 0.01
R10 1.2 k
R15 1 k
Q2 2N2222
D2 1N914
D1 1N914
R14 4.7 k
R11 10 k
C8 0.01
C19 0.001
+12 V
Chassis Mounted Components
12 - 14 V dc
10 A
D8 1N4004
GND
R13 3.3 k
R14 3.3 k
D9 Green
D10 Red
D5 1N4004
K1
+12 V
C6 0.01
C7 0.01
TR Relay
Ext. PTT
C20 0.001
C12 3.3 µF
C10 0.01
Q1 2N2222
Xcvr Out
C9 0.01
R12 3.3 k
D4 1N914
R17 4.7 k
D3 1N914
C18 0.001

Power In	R1	R2 = R3
5 W	18	300
2 W	2.7	----

Optional - RF activated T/R network
omit if external PTT line is used

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000.

T1 Prim 2T, Sec 1T CT on BN - 43 - 303 Core
T2 Prim 1T CT, Sec 2T, on T - 3/4 Core
with 1T Sampling Loop.
Wind with #24 Teflon Wire, T2 (sec) #18 Teflon

Figure 2—The RF amplifier schematic and parts list. (M) denotes Mouser Electronics, 1000 N Main St, Mansfield, TX 76063; tel 800-346-6873; **www.mouser.com**. (RF) denotes RF Parts Co, 435 S Pacific St, San Marcos, CA 92069; tel 800-737-2787; **www.rfparts.com**. (A) denotes Amidon, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; tel 800-898-1883; **www.amidon-inductive.com**. (F) denotes FAR Circuits, 18N640 Field Ct, Dundee, IL 60118; tel 847-836-9148; **www.farcircuits.net**.

RF Amplifier Board
C1-10—0.01 µF capacitor, (M) 140-100Z5-103Z.
C11—0.1 µF capacitor, (M) 80-CK06BX104K.
C12—3.3 µF capacitor, (M) 80-C340C335M5U.
C13—82 pF capacitor, (M) 5982-15-500V82.
C14—150 pF capacitor, (M) 5982-15-500V150.
C15—200 pF capacitor, (M) 5982-15-500V200.
C16, 17—4700 pF capacitor, (M) 140-50P5-472K-TB.
C18-20—0.001 µF capacitor, (M) 140-100Z5-102Z.
D1-D4—1N914 diode, (M) 610-1N914.
D5, D7, D8—1N4004 diode, (M) 583-1N4004.
D6—FES8JT diode, (M) 625-FES8JT.
D9—LED, green, with mount, (M) 512-HLMP4719.
D10—LED, red, with mount, (M) 512-HLMP4700.
K1—Relay, 12 V dc coil, DPDT, (M) 551-MR-12USR.
Q1, Q2—Transistor, switching, 2N2222, (M) 511-2N2222A.
Q3, Q4—Transistor, RF, 2SC2312C, (RF) 2SC2312C.
R1—300 Ω, 1 W, (M) 281-300.
R2, R3—18 Ω, 1 W, (M) 281-18.
R4, R5—6.8 Ω, ¼ W, (M) 30BJ250-6.8.
R6, R7—18 Ω, ¼ W, (M) 30BJ250-18.
R8, R9—120 Ω, ¼ W, (M) 30BY250-120.
R10—1.2 kΩ, ¼ W, (M) 30BJ250-1.2K.
R11—10 kΩ, ¼ W, (M) 30BJ250-10K.
R12-14—3.3 kΩ, ¼ W, (M) 30BJ250-3.3K.
R15—1 kΩ, potentiometer, (M) 531-PTC10H-1K.
R16—27 Ω, 1 W, (M) 281-27.
R17—4.7 kΩ, ¼ W, (M) 30BJ250-4.7K.
RFC1—RF choke, (RF) VK-200-3R.
T1—Transformer core, (A) BN-43-303.
T2—Transformer core, (RF) T-¾ core.
U1—IC, LM317T, (M) 512-LM317T.

Misc
2—TO-220 mounting kit, (M) 534-4724.
2—TO-220 thermal insulator pad, (M) 526-NTETP0006.
PC board, FARA RF amplifier, (F).

Low-Pass Filter Board
(see Figure 3 for component delineation)
2—100 pF capacitor, (M) 5982-15-500V100.
3—180 pF capacitor, (M) 5982-15-500V180.
3—330 pF capacitor, (M) 5982-15-500V330.
2—430 pF capacitor, (M) 5982-15-500V430.
1—560 pF capacitor, (M) 5982-15-500V560.
3—820 pF capacitor, (M) 5982-15-500V820.
3—1500 pF capacitor, (M) 5982-19-500V1500.
1—2700 pF capacitor, (M) 5982-19-500V2700.
12—0.01 µF capacitor, (M) 140-100Z5-102Z.

[continued on next page]

[continuation of Figure 2]

12—Relay, 12 V dc coil, DPDT, (M) 655-T7NS5D1-12.
1—Switch, 1 pole, 6 pos, (M) 10-YXX026.
HF Filter Kit, (A) HFFLT. Contains the cores and wire necessary to build the low-pass filters.
PC board, FARA LP Filter, (F).

Chassis Parts

Chassis box, Bud, (M) 563-AC-402.
Chassis cover, Bud, (M) 563-BPA-1589.
Heat sink, (M) 532-244609B02.
4—4-40 hex-type standoff, (M) 534-2201.
9—4-40 × ¼ pan head screw.
4—4-40 × ½ pan head screw.
6—4-40 flat washer.
9—4-40 lock washer.
2—4-40 nut.
2—8-32 × ¼ pan head screw.
2—8-32 lock washer.
2—BNC socket, chassis mount, (M) 161-9323.
RCA-type socket, female, chassis mount, (M) 161-1005.
Switch, toggle, SPDT, (M) 10TC320.
Jones-type socket, chassis mount, male, 2 pin, polarized, (M) 538-13023.

The low-pass filter assembly utilizes relays to select the proper filter network for the various frequency ranges. Relays were chosen to simplify the RF switching and to minimize cost. The six filters cover the nine amateur bands from 1.8 to 30 MHz; the frequency ranges and circuit constants are as noted on the LPF schematic diagram. The inductor cores and wire to wind the coils are available as a kit of parts from Amidon.[3] The L/C constants are the same as those recommended by WA2EBY for the MOSFET RF amplifier in *The ARRL Handbook*.[4] The filters are not used when the amplifier is off or when the PTT line is not activated—this permits multi-band listening and limited VHF use when a wide frequency transceiver is in use (like the Yaesu FT-817). There is no provision for ALC feedback, so caution must be exercised so as not to overdrive the amplifier.

Construction Hints

Although no step-by-step instructions are provided, a few hints will ease the assembly process. The circuit boards pictured are the prototype assemblies; they are not solder-plated. However, the available circuit boards (from FAR Circuits) are plated but do not have plated through-holes, so through-holes must be pinned and soldered.[5] Detailed drawings of the PC boards, the parts layout, coil-winding data and chassis templates can be found at **www.arrl.org/files/qst-binaries/faraamp.zip**. Saul, K1BI, the FARA Webmaster has also set up a site for the project. It can be found at **www.falara.org/tektalk/tektalkfs.html**. The circuit boards as they appear before wiring can be seen in Figure 5.

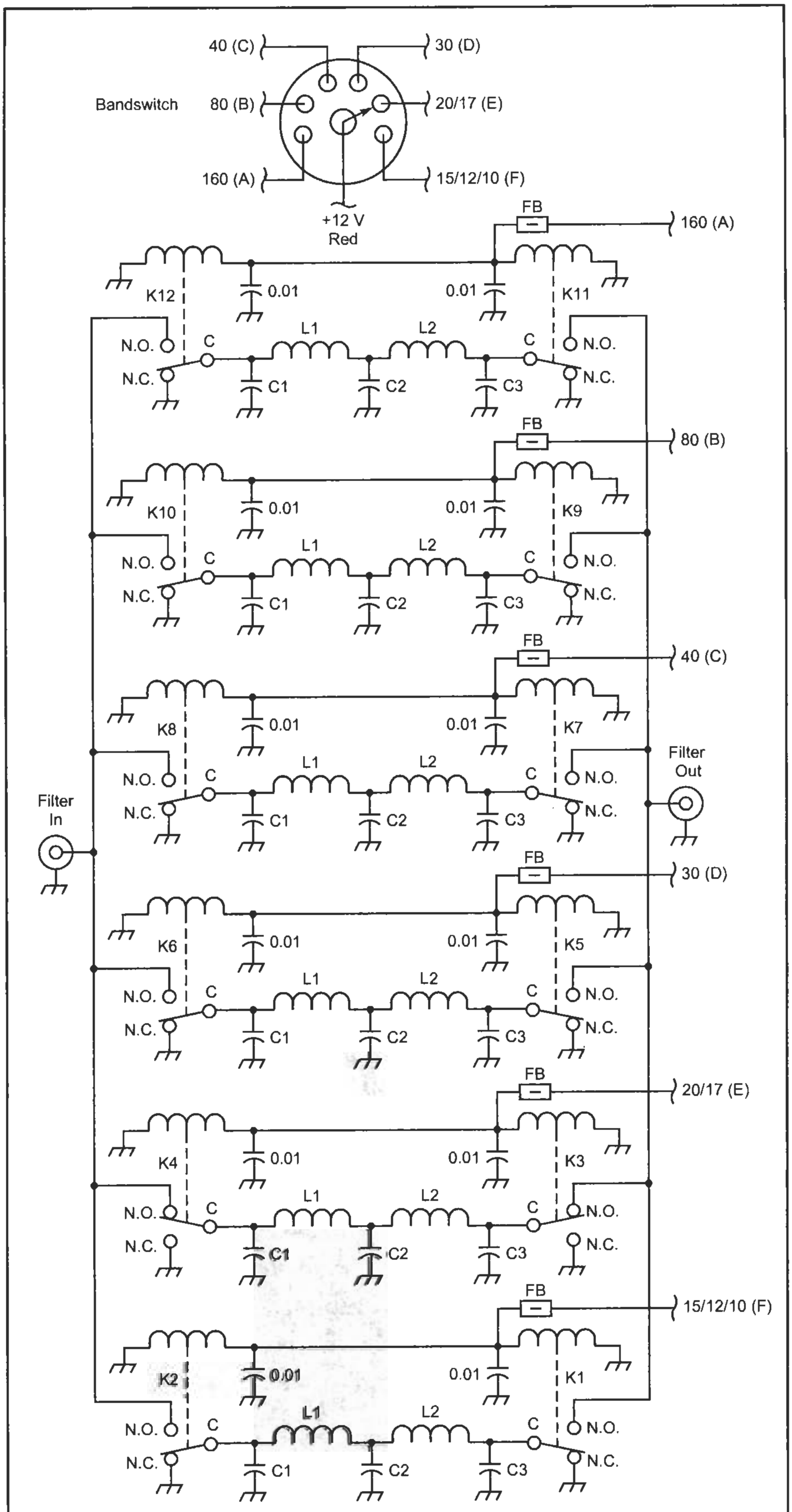

Cores T - 50 - X Amidon
L in number of turns (#22)
C in pF
Band (E) Shown Selected.
K3, K4 Energized.

Band	L1/2	C1/3	C2	Core
160 (A)	30	1500	2700	2
80 (B)	22	820	1500	2
40 (C)	16	430	820	2
30 (D)	14	330	560	2
20 - 17 (E)	11	180	330	6
15 - 10 (F)	8	100	180	6

Figure 3—The low pass filter (LPF) schematic.

Circuit Board Preparation

- Given the large ground plane area, the boards must be clean or you will experience difficulty when soldering to the foil. Plated boards are best; they are easier to solder. Some solder flux may improve the solderability of the board, but be sure to use *only* rosin core solder and a *non-corrosive* flux.
- Carefully inspect all soldered connections for cold solder joints. Good soldering technique is crucial to the performance of the amplifier.
- Periodically, the flux should be removed from the board during the construction phase with a suitable chemical cleaner.
- There are a number of holes to be drilled and pinned on each board; these are noted on the parts placements diagrams. Wires should be inserted through the board, bent into the shape of a "Z," formed flat against the board, soldered and cut.
- The four corner holes on each board must be sized as a clearance hole for the 4-40 mounting hardware.
- Components should be mounted flush to the board—fixed capacitors should be mounted as close to the foil as possible.
- The large rectangular blocks on the RF board must be trimmed in order to mount the RF output transistors and the bias diode.

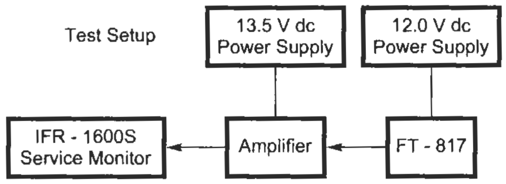

FT - 817, Power Output = 4.6 Watts ± .2 W (with −3 dB Pad)

FARA Amplifier Output Characteristics

Frequency (MHz)	RF Output (W)	2nd Harmonic (dB)	3rd Harmonic (dB)
1.8	23	-44	-43
3.5	33	-55	-63
7.2	33	-60	-50
10.1	28	-62	-45
14.2	31	-60	-50
18.1	31	-60	-52
21.2	30	-50	-47
24.9	25	-60	-53
28.4	21	-60	-56

Figure 4—Amplifier output data, including second and third harmonic response.

RF Amplifier Circuit Board

A view of the completed RF board can be seen in Figure 6. The following suggestions pertain to the amplifier board.

- Wind the secondary winding on T2 and mount it to the circuit board first.
- The emitters of the 2SC2312s (Q3, Q4) are intended to be grounded through a hole on the pad. The following modification is advised—bend a thin brass or copper strap into a "U" shape and solder it on both sides of the board. This lowers the impedance to ground.
- Next, mount the smaller fixed components, the resistors and capacitors.
- The semiconductors and the relay mount last.
- Do not mount D6, Q3 and Q4 until the assembly is fixed in the chassis and positioned relative to the heat sink.

Low-Pass Filter Circuit Board

This is a double-sided board; the reverse side is a ground plane with clearance etches for the various components. The LPF board can be seen in Figure 7. It may be necessary to form the capacitor leads slightly to conform to the hole spacing.

- Mount the filter components first, followed by the bypass capacitors and the jumper wires. Refer to the schematic for component values by frequency range.
- The relays mount last. Do not overheat their mounting pins when soldering.
- Do not mount the LPF assembly until the initial tune-up is completed. It is recommended that you pre-wire the band switch. It mounts between the two circuit boards when they are installed in the chas-

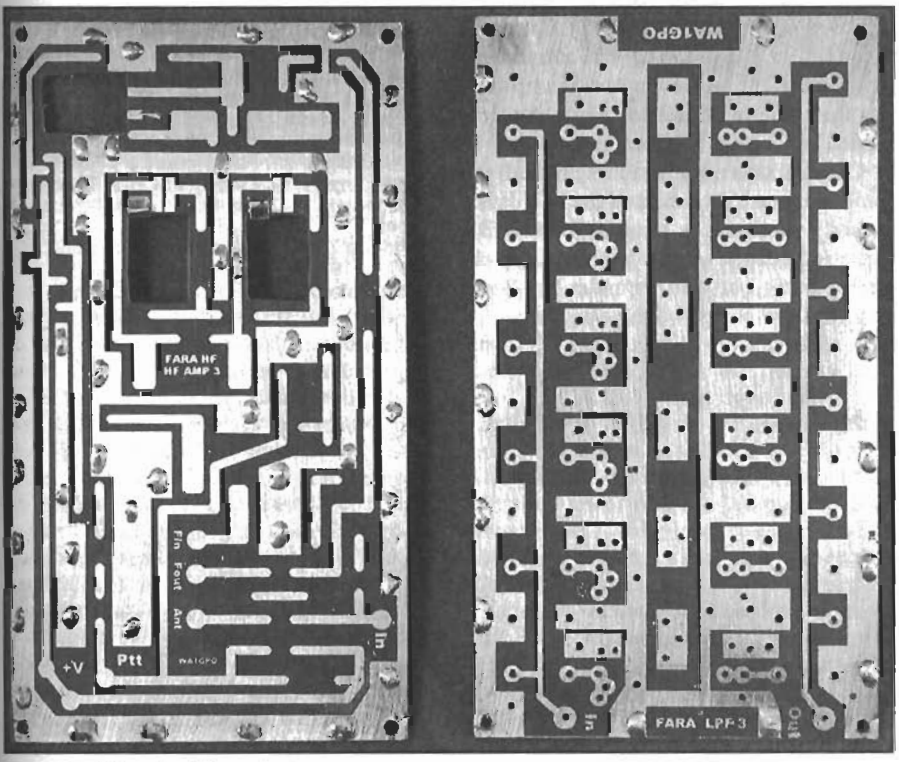

Figure 5—The circuit boards before wiring. The RF board is on the left and the LPF board is to the right. Note the pinned and soldered holes.

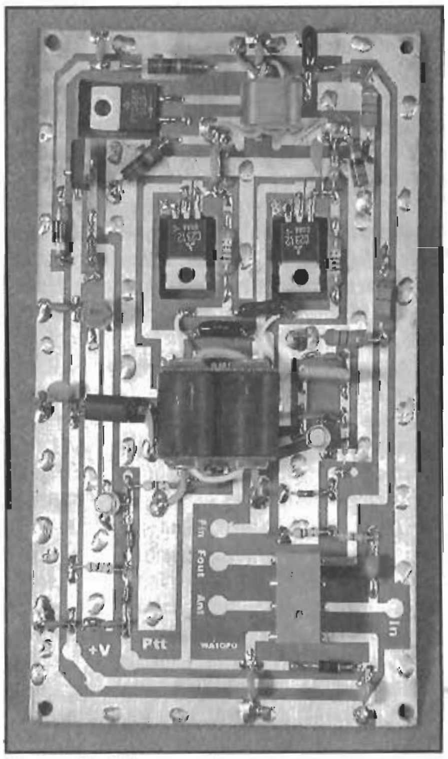

Figure 6—The completed RF assembly board.

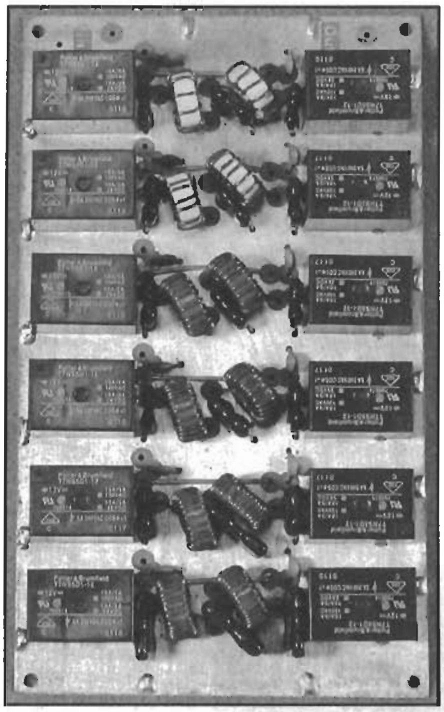

Figure 7—The completed LPF board. Note the relays that are used as filter switches–they are selected by the band switch.

sis and it is difficult to get to.

Chassis, Panel and Heat Sink

Full size templates for the chassis and panel can be found on the Web site. Trim the templates to size and fold and tape them to the aluminum box and heat sink. It is important to center punch the holes and drill a small pilot hole at each location. Enlarge the holes to size according to the dimensions. Letter the panel with dry transfer lettering available at office supply stores. Spray on several light coats of clear lacquer to protect the lettering. Mount the heat sink and the panel components, but do not mount the bandswitch at this time. Mount the RF amplifier board using a couple of 4-40 flat washers as spacers between the chassis and the circuit board at each corner. Install the 4-40 standoffs to hold the assembly in place. Mount D6, Q3 and Q4, using the TO-220 thermal pads and hardware to isolate the transistor mounting tab from the chassis. It is not necessary to isolate D6. Use RG-174 miniature coaxial cable for the internal RF connections.

Tune-up and Testing

As this is a broadband design there is no tune-up; only the bias adjustment needs to be set. A 12 to 14 V dc current limited variable supply is recommended for initial adjustment and testing. Use a dummy load at the amplifier output. Drive levels refer to the attenuator output, if used.

- Connect a temporary jumper between the F_{in} and F_{out} pads on the RF board.
- Preset R15 (1 kΩ potentiometer) so that the wiper is at ground.
- Apply 12 V dc and ground the PTT line (the relay should pull in).
- Briefly drive the input with 1 W at 14 MHz and note the power output.
- Again apply 1 W and increase bias (R15) until the output increases about 15%.
- Increase drive to 2 W and note the power output (about 25 W).
- Increase voltage to 14.7 V dc; note the output with 2 W of drive (about 35 W).

Remove power; remove the temporary jumper; mount the band switch; connect and install the LPF board. Verify that the power output over the range of 1.8 to 29 MHz is as expected. A view inside the bottom of the amplifier, with the LPF board visible, is shown in Figure 8.

Final Comments

The FCC has placed strict limitations on power amplifiers used below 30 MHz. Please review the appropriate FCC regulations before constructing this amplifier.[6] Kits offered for sale, even partial amplifier kits that require additional parts, are prohibited by current FCC regulations.[7]

Construction time, assuming that all the components are on hand and the boards are properly prepared, is about 4 hours. SSB operation results in the heat sink barely getting warm, and cooling under key-down conditions is more than adequate.

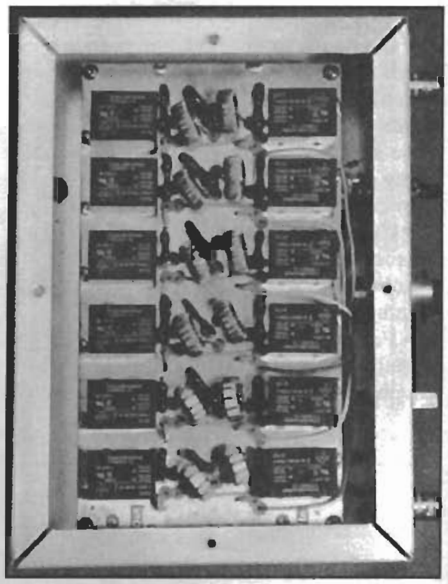

Figure 8—A bottom view inside the completed amplifier. The LPF board is mounted below the RF assembly.

A project of this scope is more fun when others participate. Harry, W2RKB, provided the necessary prodding to get it started. He also fabricated the circuit boards and the filter assemblies for the prototype amplifiers. Dave Hosom assisted with the photographs. A generous thank you is extended to them both. Give the FARA amplifier a try… it's a practical and rewarding project!

Notes

[1] Federal Communications Commission, Sec 97.315.

[2] See Note 1.

[3] Amidon Associates, Inc, 240 Briggs Ave, Costa Mesa, CA 92626; 800-898-1883; **www.amidon-inductive.com**.

[4] *The 2003 ARRL Handbook*, pp 17.91-17.97.

[5] FAR Circuits, 18N640 Field Court, Dundee, IL 60118; tel 847-836-9148; **www.farcircuits.net**.

[6] Federal Communications Commission, Sec 97.3 (19).

[7] See Note 6.

By Pat Bunn, N4LTA

From *QST*, October 1990

A Compact "Brick" for 6 Meters

Does your six-meter signal need some extra kick? This inexpensive amplifier bridges the gap from 10 to 100 watts.

Most transceivers available for the 6-meter band provide 10 W output. This power level provides satisfactory communications during good sporadic-E openings, but is not adequate for making long-haul contacts on meteor scatter or during F-layer openings. For the average station, an output level of 100 W seems to be a reasonable compromise between a 10-W exciter and a kilowatt amplifier. Unfortunately, 50-MHz amplifiers in this power class are not readily available.

I began operating on the 6-meter band a few years ago, near the lowest point in sunspot cycle 21. Sporadic E kept my interest high, and soon it became evident that I needed an amplifier. Being a builder at heart, I decided to "roll my own" and began to search the Amateur Radio literature for design examples. I found very few solid-state amplifiers, and only one in the 100-W class.[1]

I built the amplifier as specified, and I got good results. My 6-meter effectiveness made a quantum leap! One highlight was working Europe with a 12-foot-boom Yagi. The amplifier performed so well that several of my friends asked me to build them duplicates of it. With each revision, I tried to add design innovations, including on-board TR relays, a carrier-operated-relay (COR) circuit and transistor biasing, as well as low-pass filtering and more compact packaging. The product of these efforts is described in this article. At least 25 of these units have been constructed and are now in use. Several of these amplifiers have been used on DXpeditions, in beacon service, and as drivers for kilowatt-class amplifiers, and results have been excellent.

If you are looking for a low-cost 100-W amplifier, this proven design would make a nice addition to your 6-meter station. This circuit can also be adapted for use with an MRF140 FET, which produces over 160 W output with less than 10 W drive. See the sidebar for more details.

6 METER AMPLIFIER
DELAY
POWER

Circuit Highlights

The amplifier circuit, shown in Fig 1, is a single-stage, 10-dB-gain, class-B design using the popular Motorola MRF492 transistor. The RF circuitry is based on a test circuit in Motorola's RF Device Data book.[2] Input and output impedance matching is done with low-Q lumped circuits rather than with broadband transformers. The amplifier's bandwidth is about 1.5 MHz, which covers the DX portion of the 6-meter band with plenty to spare. And, the output network has a secondary dip at about 53 MHz, allowing CW/SSB work and FM operation without the need to retune the amplifier. The narrowband nature of the matching networks helps reduce harmonics. A half-wave low-pass filter in the output circuit further reduces harmonic content to meet current FCC specifications (see Fig 2).

The COR circuit is based on a high-gain operational amplifier that provides the sensitivity required for SSB operation. This circuit is much more effective than the Darlington-transistor COR circuits found in most commercial amplifiers. It works well with low-power SSB rigs that do not have easily accessible PTT lines, although a separate PTT input is also included.

The amplifier PC board[3] is mounted in a die-cast aluminum box (Hammond 1590R, Bud CU-247 or Jameco® part no. B5006) with an extruded-aluminum heat sink bolted to the bottom. (I used a 6-inch length of Wakefield no. 1527 heat sink. Any aluminum heat sink with similar dimensions should work fine.) The completed amplifier is operated inverted, with the removable lid serving as the bottom. The enclosure is painted flat black and labeled with white rub-on lettering. Stick-on rubber feet are attached at the corners of the bottom cover. The finished product

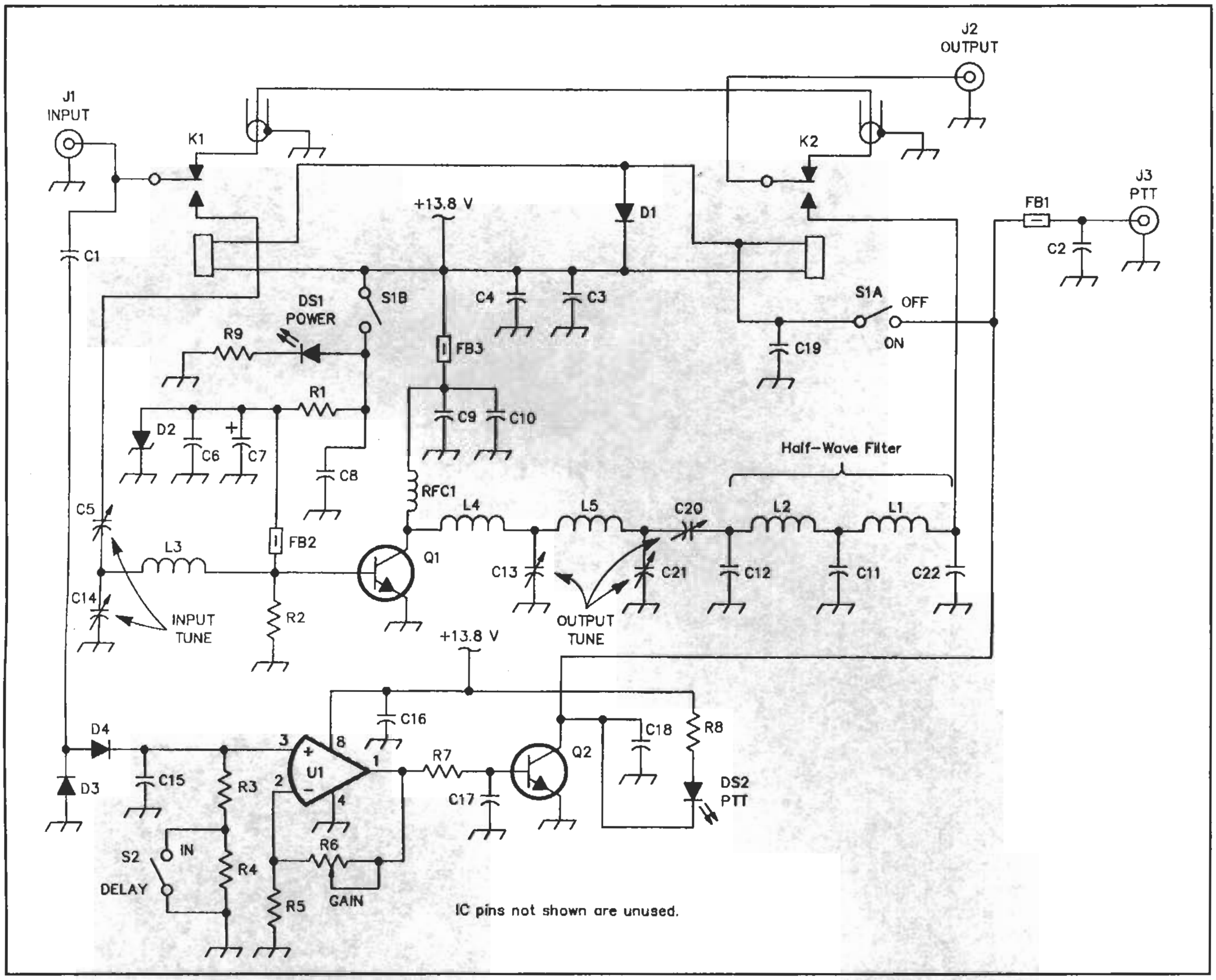
J2
OUTPUT
J1
INPUT
K1
K2
J3
PTT
FB1
C2
C1
+13.8 V
D1
S1B
C4
C3
S1A
OFF
ON
DS1
POWER
R9
FB3
C19
R1
D2
C6
C7
C9
C10
C8
Half-Wave Filter
RFC1
C5
L4
L5
C20
L2
L1
L3
FB2
Q1
C13
C21
C12
C11
C22
C14
INPUT
TUNE
R2
OUTPUT
TUNE
+13.8 V
C16
C18
R8
D4
3
8
U1
1
R7
Q2
D3
C15
R3
2
4
C17
DS2
PTT
R6
S2
IN
R4
DELAY
R5
GAIN
IC pins not shown are unused.

Fig 1—Schematic of the 100-W 6-meter amplifier. Liberal use of bypass capacitors and ferrite beads helps ensure stability and keeps RF out of the power supply.

C1—10 pF, ceramic disc.
C2—0.01 µF, polyester film.
C3, C10—0.001 µF, ceramic disc.
C4, C6, C8, C9, C15-C19—0.1 µF, polyester film.
C5—9- to 180-pF mica compression trimmer, Arco no. 463.
C7—1000 µF, 6.3 V, aluminum electrolytic.
C11—120 pF, 100 V, silver mica.
C12, C22—62 pF, 100 V, silver mica.
C13, C14, C20—50- to 380-pF mica compression trimmer, Arco no. 465.
C21—25- to 280-pF mica compression trimmer, Arco no. 464.
D1—1N4002.
D2—1N1200 stud-mount diode.
D3, D4—1N4148.
DS1—Green LED.
DS2—Red LED.
FB1, FB3—Ferrite bead, Amidon no. FB43-901 or similar.
FB2—VK-200 wide-band choke, 2½ turns no. 24 solid wire on Amidon no. FB43-5111 ferrite core.
J1, J2—Female RF connector (UHF, BNC, N, etc).
J3—Phono jack.
K1, K2—SPDT relay, 12-V dc coil, Omron no. G5L112P-PS-DC1 2. Available from Digi-Key.
L1, L2—4 turns no. 14 enam wire, 7/16 in. diam, 3/8 in. long.
L3—2 turns no. 14 enam wire, 3/8 in. diam, 1/2 in. long.
L4—no. 14 U-shaped wire loop, 3/8 in. diam, 9/16 in. finished length.
Q1—MRF492.
Q2—2N2222A.
R1—5-W bias resistor (see text).
R2—10 Ω, 1/2 W, carbon comp.
R3, R5—10 kΩ, 1/4 W, carbon comp.
R4—1 MΩ, 1/4 W, carbon comp.
R6—100-kΩ PC-board potentiometer.
R7-R9—1 kΩ, 1/4 W, carbon comp.
S1—DPDT miniature toggle.
S2—SPST miniature toggle.
U1—LM358.

Miscellaneous
Suitable die-cast-aluminum enclosure and heat sink.
Pair of binding posts or other suitable dc-supply connector.
Two 1- x 3-in, strips of double-sided foam tape.
Two LED holders.
Seven no. 4-40 × 1/4 in. machine screws.

looks professional, and also does a good job of cooling the power transistor and bias diode.

Construction

The amplifier is assembled on a double-sided printed circuit board. All RF components are surface-mounted on the component side of the PC board (see Fig 3). The COR components, including the relays, connect to traces on the other side of the board. The remaining non-component-side foil serves as a ground plane, and is electrically connected to the component-side ground plane via platedthrough holes. If you make your own PC board, you can replace these plated-through holes with no. 18 AWG wires soldered to both sides of the board, or, preferably, copper rivets such as those available from Frontier Microwave and Down East Microwave.[4] The relays also mount in plated-through holes. The PC-board patterns and part-placement overlay are available from the ARRL Technical Department Secretary.[5]

Drill the case for the switches, connectors and indicators you'll use. Begin PC-

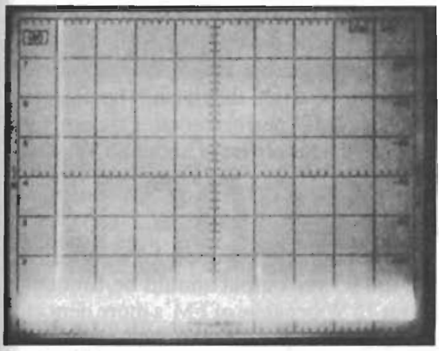

Fig 2—Worst-case spectral display of the 100-W 6-meter amplifier. Vertical divisions are each 10 dB; horizontal divisions are each 10 MHz. Output power was approximately 100 W at 50.2 MHz. All harmonics and spurious emissions are at least 60 dB below peak fundamental output when the amplifier is tuned as described in the text. The 100-W amplifier complies with current FCC specifications for spectral purity for equipment in this power-output class and frequency range. Better spectral purity can be obtained by using an external output filter (see the referent of note 1 for a suitable filter circuit).

board construction by mounting the parts that are soldered to traces on the other side of the board. This includes the COR parts and relays, the RG-174 jumper, and diode D1. Make sure that the diodes and Q2 are properly oriented.

The solder-side parts are surface-mounted. Keep lead lengths to a minimum (no more than 1/4 inch) on these parts. The bypass capacitors should have the shortest possible leads, and should be surface-mounted on the PC board.

Wind the coils using 1/4- and 3/8-inch drill bits as temporary forms. Cut the coil leads to 1/4 inch, bend them 90°, scrape the insulation off the ends and solder the coils to the board. Some of the coils may need to be compressed or spread during tune-up to get maximum output power, so allow for this when mounting them. Use Fig 3 as a guide. Mount the mica trimmers by bending the solder tabs parallel to the board and surface-mounting them. (Some trimmers have extra tabs that may short to the PC board; I cut these off with side cutters to eliminate this risk.) When surface-mounting the larger components, a high-wattage soldering gun is useful, but be careful: The traces may separate from the PC board if excessive heat is used.

When the PC board is complete, temporarily place the MRF492 in its mounting hole in the PC board, making sure to orient the transistor properly (the beveled lead is the collector). Place the PC board inside the box assembly. Carefully mark the transistor-mounting holes and the single PC-board mounting hole on the enclosure, through the PC board. Remove the PC board from the case, then drill and tap these holes for no. 4-40 hardware. Mark and drill the heat sink using the enclosure as a guide.

Drill and tap four more no. 4-40 holes, near the corners of the case, through the case and into the heat sink. Apply heat-sink compound between the die-cast box and the heat sink, then use no. 4-40 screws in these holes to secure the heat sink to the case.

Place the PC board into the case/heat-sink assembly again, and mark the location of the stud-mounted bias diode. Remove the board and drill and tap a 3/8-inch-deep no. 10-32 hole in the case/heat-sink assembly. The diode's threaded stud is more than 3/8 inch long, and should be carefully cut off (with a hacksaw) to 5/16 inch.

The rear of the PC board must be insulated from the box assembly, because many of the COR-circuit traces on the underside of the board are not at ground potential. I use 1/8-inch-thick double-sided foam tape to cover the solder-side traces and to support the board away from the case. Make sure the component leads are cut off closely at the rear of the PC board so that they don't punch through the foam tape and short to the box assembly. Mount the board inside the case, then screw the bias diode snugly into its mounting hole in the case/heat sink assembly.

Solder the MRF492 onto the board as the last step, *after* the transistor case has been mounted in the board and screwed into place. (This eliminates the possibility of cracking and ruining the $15 device!) Be sure to use a liberal amount of silicone heat-sink compound between the transistor and the box.

Mount the front-panel LEDs, switches and the RF connectors of your choice to the box. Any suitable RF connectors can be used. The connectors are wired to the PC board with no. 18 tinned bus-wire jumpers. Make these as short as possible. I used no. 24 insulated hookup wire to connect the switches and LEDs to the PC board. Finally, mount the dc-power binding posts to the rear panel. These are wired to the PC board with no. 18 bus wire run through FT-43-801 ferrite beads.

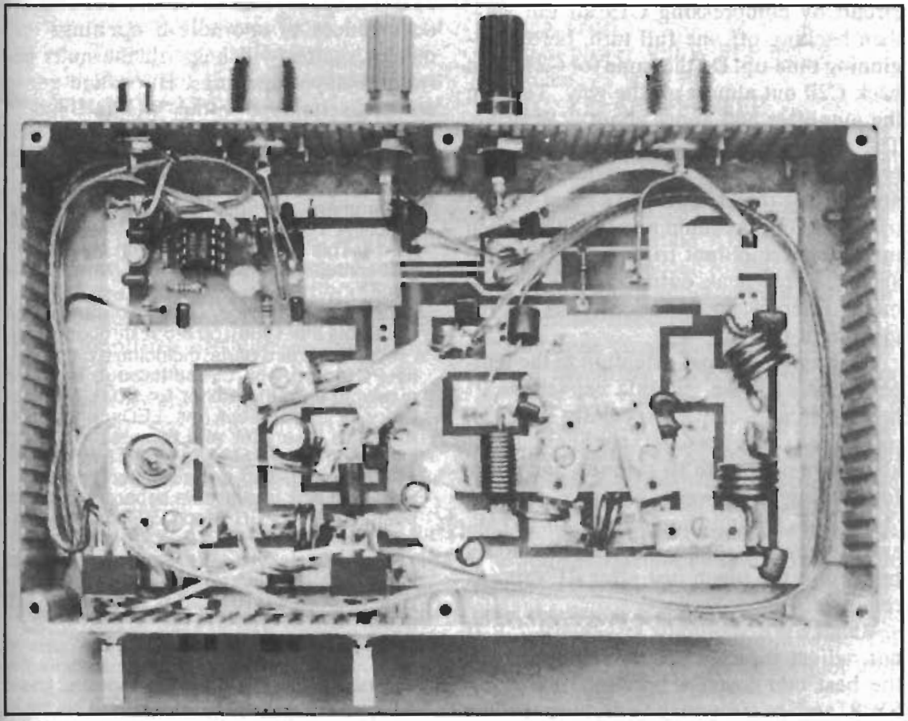

Fig 3—This inside view of the 100-W amplifier shows the circuit's simplicity. On the PC board, counterclockwise from upper left: the COR circuit; input-matching and bias networks, switches and indicators (lower left); MRF492 (lower center); collector choke and output network (lower right). The TA relays and connectors are at the upper center, and bias resistor R1 is just to the left of center.

Initial Testing

After the switches, LEDs and connectors have been wired, carefully check all connections. Make certain that the jumper from the PC board to bias diode D2 is connected! If all looks good, connect the amplifier to a 13.8-V power supply capable of sourcing 1 A. (If you have one, a small, current-limited power supply is best for testing.) If your power supply does not have reliable current metering, a meter capable of measuring 0-500 mA should be placed in series with the positive supply lead.

With the amplifier turned off, turn on the power supply. The meter should indicate 0. Turn on the amplifier and note the total indicated current. It should be between 50 and 300 mA. If it is much greater than 300 mA, quickly turn off the power supply and look for problems.

If the current is within these limits, you can go about setting the operating bias. This involves changing the value of R1 until the collector current is between 75 and 175 mA. Note that this is not total circuit current, which includes the diode

Using a MOSFET to Get More Power from the Basic MRF492 Amplifier Circuit

The MRF492 amplifier circuit described in the main text is easily modified for use with an MRF140 VMOS power FET, which has about 15 dB gain and over 160 W output—substantially more than the MRF492. Even the device cases are the same, so the changes required to operate this circuit with an MRF140 are quite straightforward. Specific details of the modifications required, including a schematic diagram, are available from the author for an SASE.

The main differences between the devices are that the MRF140, a MOSFET, needs different biasing and a higher supply potential (28 V) than the MRF492, and is more expensive ($88 versus $15 for the MRF492, at this writing). But if you have a 28-V supply and need more than 100 W or have a 5-W exciter, the change is probably worthwhile. The MRF140 is easily biased with a three-terminal regulator (an LM317T works well), and requires very few component changes other than those in the bias circuit. An additional capacitor and inductor are also required at the MRF140's gate.

Because the modified amplifier requires a 24- to 28-V supply, different TR relays must be used, or, alternatively, 12-V-coil relays can be wired in series. (The COR circuit needs no modification because the LM358 works fine at 28 V.)

Good heat sinking is more critical to amplifier performance at the higher power level, so I used a 3- × 5- inch slab of 1/8-inch-thick copper between the device and the case as a heat spreader. A larger heat sink and/or forced-air cooling are also desirable.—*N4LTA*

current and POWER LED current. The best way to measure collector current is to lift FB3 and insert a milliammeter in series with the collector-supply lead at that location. An alternative method is to use the following equation to approximate collector current:

$$I_C = I_{TOTAL} - [(V_{R1} \div R1) + ([V_{CC} - 1] \div 1000)]$$

where

I_C = collector current in milliamperes
V_{R1}= measured voltage across R1
V_{CC} = measured supply voltage

This equation lets you calculate collector current by subtracting the current through R1 ($V_{R1} \div R_1$) and the current through the **POWER** LED ($[V_{CC} — 1] \div 1000$) from the total supply current.

Using either method, substitute different resistors at R1 until the collector current falls into the 75- to 175-mA range. In almost every case, a 5-W resistor of 82, 91, 100, 110, 120 or 130 Ω will bring the current into that region. (I have never had to substitute more than three resistors to arrive in this bias-current range.) These 5-W resistors can be purchased for less than $0.40 each, so, although a 200-Ω, 5-W wire-wound potentiometer could be used for R1, all six resistors can be purchased for far less. And, these days, high-wattage potentiometers are hard to find.

Tune-Up

When the bias current has been set, the amplifier can be initially tuned. For this procedure, the amplifier should be connected to a power supply capable of delivering at least 15 A at 13.8 V. A power meter capable of indicating at least 100 W at 50 MHz should be connected at the amplifier output, and the meter's output to a 50-Ω load. It is also desirable to measure input power from the exciter, as well as amplifier-input SWR, but these can be measured after initial tune-up if two meters aren't available.

Set the input-circuit trimmers (CS and C14) near midrange. Because the output circuit can be tuned several ways for the proper impedance transformation, it's important to preliminarily adjust the output circuit by compressing C13 all the way, then backing off one full turn, before beginning tune-up. Do the same for C21, then back C20 out almost all the way. Turn on the amplifier and apply about 2 W drive. The COR should sense the input RF and switch immediately into transmit. Using a nonmetallic tool, tune the two input-circuit trimmers for maximum power output, then tune the three output trimmers (C13, C20 and C21) for peak output. These adjustments interact, so continue tweaking for maximum power.

During tuning, pay attention to the temperatures of these variable capacitors. If heating is noticeable (especially in C13), reset these capacitors and start again. Limit key-down periods to 10 seconds, and allow for cooling in between. Increase drive to about 9 W. Output should be 95 to 105 W, and the output-tuning capacitors should be cool to the touch just after RF drive is removed.

Input SWR should be less than 1.2:1. If not, adjust the input-circuit trimmers for the best compromise between low input SWR and maximum power output. The amplifier is now ready for operation. Some MRF492s seem to be "hotter" (exhibit more gain) than others; I have seen some devices produce 115 W easily and others strain to put out 95 W. On the air, the difference is so slight that it can't be detected. For best IMD performance, I recommend running the amplifier at about 85 W output on SSB.

FM Operation

This amplifier also works well on FM, but the power output should be limited to about 75 W because of FM's more demanding duty cycle. As mentioned earlier, the amplifier's bandwidth characteristic is double-humped; if it's tuned up at 50.1 MHz, the second hump falls in the 53-MHz area, allowing for SSB and FM operation without retuning.

If the amplifier is to be used for beacon, CW or FM operation exclusively, I suggest you configure it for class-C operation. Efficiency in class-C operation is in the 72% range, versus 60% for class B. To configure the amplifier for class-C use, connect the inductor using FB2 (the multi-hole ferrite bead) from Q1's base to ground, and eliminate all of the other bias components, including D2. If you like, you can add a panel switch to change between class-C and class-B operation.

The amplifier should never be operated class C on SSB. Doing so results in severe distortion.

Conclusion

I hope that you enjoy building this project, and will be motivated to try 50-MHz operation. As I write this, Japanese stations are being worked on the East Coast, and all continents are frequently heard. In the years to come, the regular occurrences of sporadic-E openings and meteor scatter will help fill the nulls between F-layer openings. Hurry and get in on the excitement on the "Magic Band"!

Notes

[1] T. Tammaru, "A Solid-State 6-Meter Linear Amplifier You Can Build," *QST*, May 1982, pp 11-14. This circuit was based on an MRF492 test circuit in Motorola's *RF Device Data* book.

[2] *Motorola RF Device Data*, Fourth edition (Motorola Inc: 1986), pp 3-552 through 3-555.

[3] A set of PC-board parts, including a high-quality PC board with plated-through holes, is available from the author for $100 (UPS delivery included). Switches, LEDs, enclosure, terminals, heat sink and connectors are not included. The PC board alone is available for $17 postpaid, and a black anodized-aluminum heat sink is available separately for $12 postpaid. (The ARRL and *QST* in no way warrant this offer.)

[4] Frontier Microwave, AD 1, Box 467, Ottsville, PA 18942, evening tel 215-795-2648; Down East Microwave, RR 1, Box 2310, Troy, ME 04987, tel 207-948-3741.

[5] Send a business-size SAE with return postage for 1 ounce with your request. Specify the information package for the 6-meter linear amplifier from October 1990 *QST*. This package includes full-size etching patterns for the front and rear sides of the amplifier PC board and a part-placement overlay.

By Richard Frey, K4XU From *QEX*, May 1999

A 300-W MOSFET Linear Amplifier for 50 MHz

Everyone is looking to go QRO on 6 meters. Build this "almost a brick" to boost your signals from 15 to 300 W.

In an earlier article, I described a 50 MHz, 125 V, 250 W, class-C amplifier using ARF448AJB highvoltage MOSFET devices.[1] This paper describes an improved version of an amplifier that is capable of class-AB linear operation. The design changes required and the procedures involved are explained and demonstrated. Complete descriptions of the amplifier and its construction are presented, as well as the measured performance.

High-voltage, high-power MOSFETs have been shown to be very capable RF power amplifiers.[2] The metal-gate architecture of the ARF series from Advanced Power Technology has raised the frequency limits for this type of device to 100 MHz. The APT448A/B is typical of the series. It has a 68,000square-mil die with a breakdown voltage rating, BV_{dss} of 450 V. The device is provided in the inexpensive TO-247 plastic package, and is available in common-source symmetric pairs. Like all MOSFETs, the gate threshold voltage, V_{th}, has a negative temperature coefficient. This makes operation as a linear amplifier difficult or impossible without compensation.

When forward biased with a constant gate voltage, the quiescent drain current rises as the temperature of the die increases. Operating at the typical drain voltage for these parts, about one third of the rated BV_{dss}, the power dissipation caused by the increasing ‘dq results in "hot spotting" and subsequent thermal runaway. This is an unstable system. The dissipation increases so rapidly that the outside surface of the case does not follow internal junction temperature well. As a result, a bias-compensation scheme that uses temperature sensing cannot keep up with the V_{th} shift, and the device is destroyed.

The power dissipation within the die is a direct function of the operating voltage. By lowering the operating voltage, the thermal loop gain can be reduced to a point where the gatethreshold shift can be compensated. Thermal stability can be achieved by sensing the case temperature. Linear operation thus becomes practical at 100 V and below. While this is less than 25% of the rated BV_{dss} and produces less gain, a very rugged and useful linear amplifier is the result.

Amplifier Description

The following were the design goals for the amplifier:

- Frequency range 50 to 51 MHz
- Input SWR < 1.5:1
- Gain> 13dB
- Output power 300 W PEP or CW
- Efficiency> 50%
- IMD3 > 25 dB below PEP

A push-pull topology was chosen for best output power and minimum harmonic content. The previously reported class-C design used $V_d = 125$ V. Since this is too high for reliable class-AB operation, 80 V was eventually chosen for this design. This is a compromise between gain, efficiency and thermal stability.

Since the gate input impedance is very low, it magnifies the effects of any stray inductance in the gate matching circuit. In a push-pull design, it is critical to maintain absolute symmetry between the two sides. This fact was demonstrated during the initial design work. One preliminary design had a slight asymmetry in the PCB artwork. The amplifier exhibited low efficiency, hot ferrite in the output transformer balun and poor distortion characteristics with asymmetrical IMD product amplitudes. This clearly demonstrates the benefit of symmetrically packaged devices.

A multiple-aperture ferrite bead was chosen for the input transformer. Brass tubing was used for the secondary, and the primary was wound inside the brass tubes to provide a very broadband balanced transformer design with minimum leakage reactance. Several cores and construction methods were evaluated, and this well-tried design proved best.

The typical HF push-pull amplifier employs a bifilar choke to decouple the drain-voltage feed. As frequency and power increase, this feed method becomes less practical because the need to reduce the number of windings to offset the stray capacity and the need to prevent core saturation conflict. In this design, a powdered-iron toroid was chosen for the feed choke core. This proved far superior to any of the ferrite cores evaluated. It is inexpensive and easy to reproduce.

Design

Three main areas must be addressed. The input matching must provide a balanced feed to a pair of low-impedance gates. The output must be matched to a suitable load impedance, and then transformed to a 50-Ω unbalanced load. The bias must be thermally compensated to track the negative temperature coefficient of the gate bias threshold.

The input is reasonably straightforward. Each gate input is 0.2 +*j*0.5, as estimated from the data sheet. The push-pull topology puts these two impedances in series, which makes the matching less difficult. A Smith Chart[3] program makes the actual design easy. The program used here is *WinSmith*.[4] The two gate impedances are added in series, and a network synthesized to transform the resulting impedance up to 50 Ω. The Smith Chart program only works with single-ended circuits, so the center tap was added later. See Fig 1.

The input-transformer design was chosen for its simplicity and relative ease of construction. Of several attempts using different material permeabilities, multiple beads and different conductor types, this proved to be the best and most consistent performer. The core is a Fair-Rite[5] "multi-aperture” core, part number 2843010402. The type-43 material has a μ_i of 850. At 50 MHz, type-61 material (μ_i of 125) would also be satisfactory. This transformer is essential to provide balanced drive to the gates of the MOSFETs. The secondary winding is $^3/_{16}$-inch brass tubing. Copper shim stock forms the connections to the brass tubing at each end of the transformer secondary. The two-turn primary is wound inside the tubing. This construction provides a very reproducible transformer with minimum leakage reactance and a very broad frequency response. It would be a suitable input transformer for a broadband amplifier covering 1 to 100 MHz.

The leakage reactance of the input transformer—referred to the secondary—is about 18 nH and is represented as Li on the simplified input schematic in Fig 1. The gate load is represented by the “Load R,” L2 and C3. Using all three parts of the gate impedance allows proper evaluation of the network bandwidth. A pi network consisting of C1, TL1 and C2 is used to step-up the gate load to the 12.5 Ω needed by the transformer and compensates for T1’s leakage reactance. Notice that the net stray inductance of the gate is almost enough to effect a match with a single shunt capacitor. This has actually been done, but it was not easy to fit all the parts in the available space; so it was judged unacceptable here. To transform the network into the required balanced configuration, the series TL1 is split into two equal parts; the shunt capacitors remain the same, and a neutral center tap is provided at the transformer secondary.

Because of the high currents circulating in the input network, it is imperative that C2 be a larger-sized, class-1 dielectric (C0G or NP0) capacitor. It must be a leadless, surfacemount chip type, or the value will need to be adjusted. The input tuning capacitor, a 900 pF mica compression trimmer, is mounted directly to the end of the input transformer.

The output network is also straightforward. The proper load impedance for class

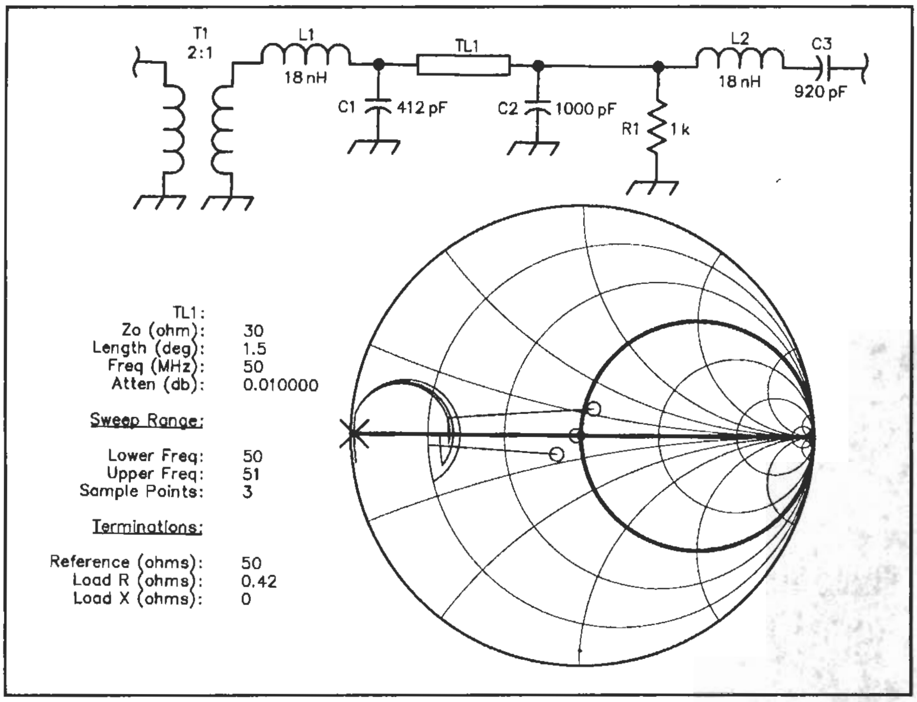
T1
2:1
L1
18 nH
TL1
L2
18 nH
C3
920 pF
C1
412 pF
C2
1000 pF
R1
1 k
TL1:
Zo (ohm): 30
Length (deg): 1.5
Freq (MHz): 50
Atten (db): 0.010000
Sweep Range:
Lower Freq: 50
Upper Freq: 51
Sample Points: 3
Terminations:
Reference (ohms): 50
Load R (ohms): 0.42
Load X (ohms): 0

Fig 1—Input matching network and calculation.

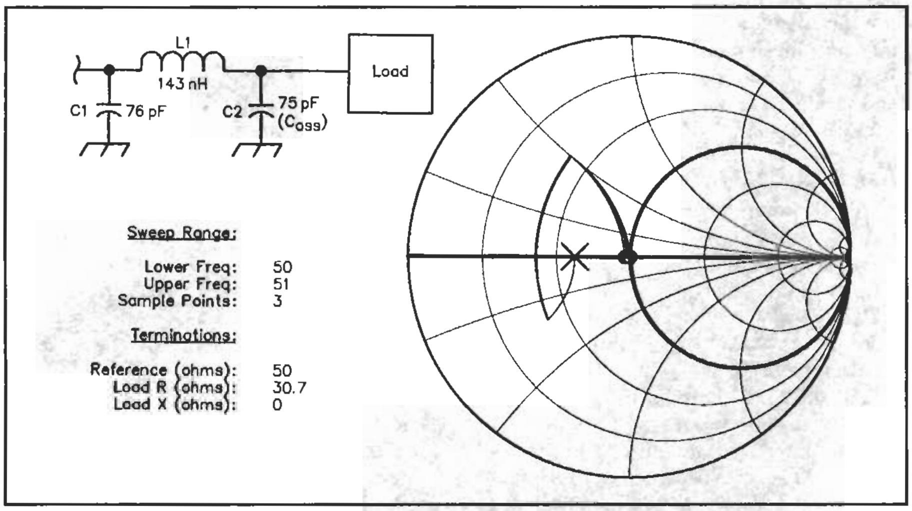
L1
143 nH
C1
76 pF
C2
75 pF
(Coss)
Load
Sweep Range:
Lower Freq: 50
Upper Freq: 51
Sample Points: 3
Terminations:
Reference (ohms): 50
Load R (ohms): 30.7
Load X (ohms): 0

Fig 2—Output matching network and calculation.

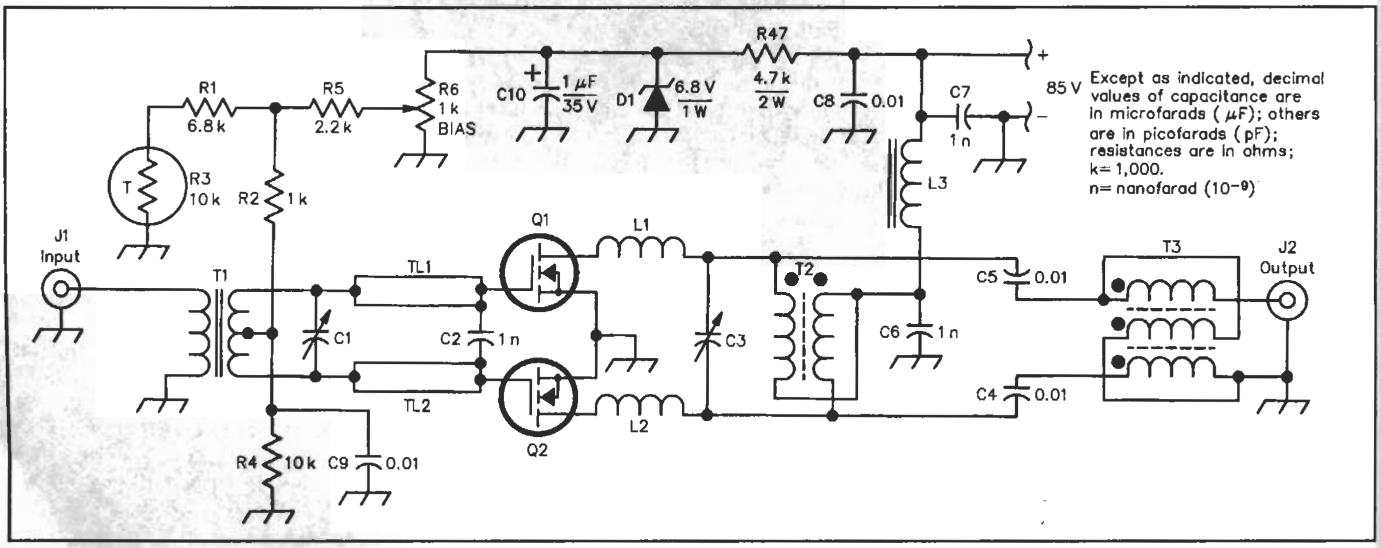
R47
4.7 k
2 W
+
85 V
−
R1
6.8 k
R5
2.2 k
R6
1 k
BIAS
C10
1 μF
35 V
D1
6.8 V
1 W
C8
0.01
C7
1 n
L3
T
R3
10 k
R2
1 k
J1
Input
T1
TL1
TL2
C1
C2
1 n
Q1
Q2
L1
L2
C3
T2
C5
0.01
C6
1 n
C4
0.01
T3
J2
Output
R4
10 k
C9
0.01
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000. n= nanofarad (10⁻⁹)

Fig 3—50 MHz amplifier schematic diagram.

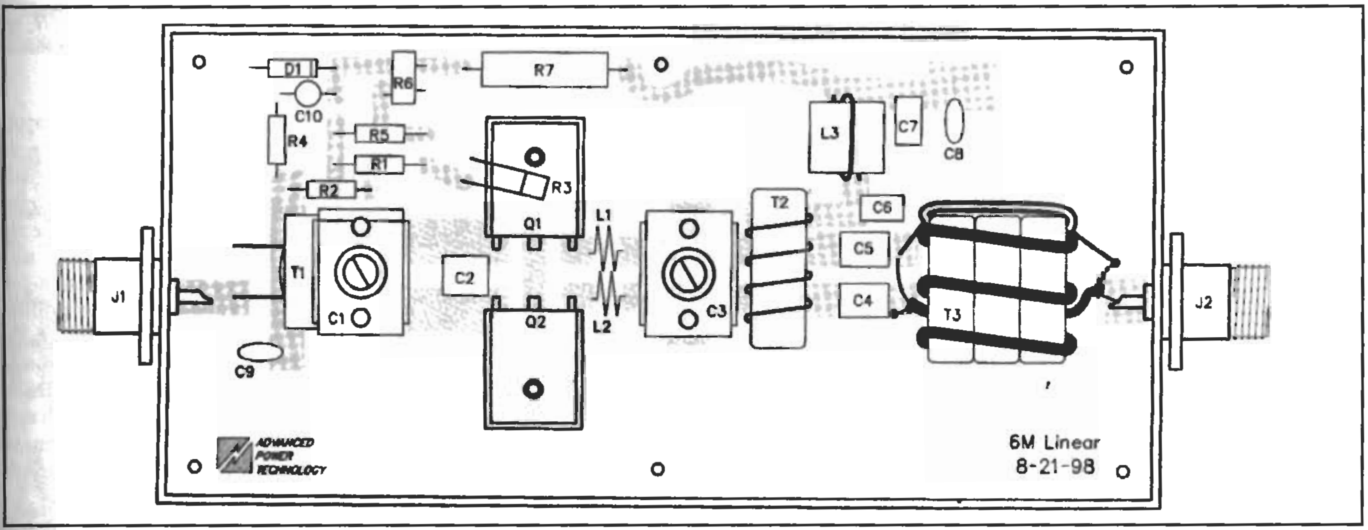

Fig 4—Amplifier parts layout.

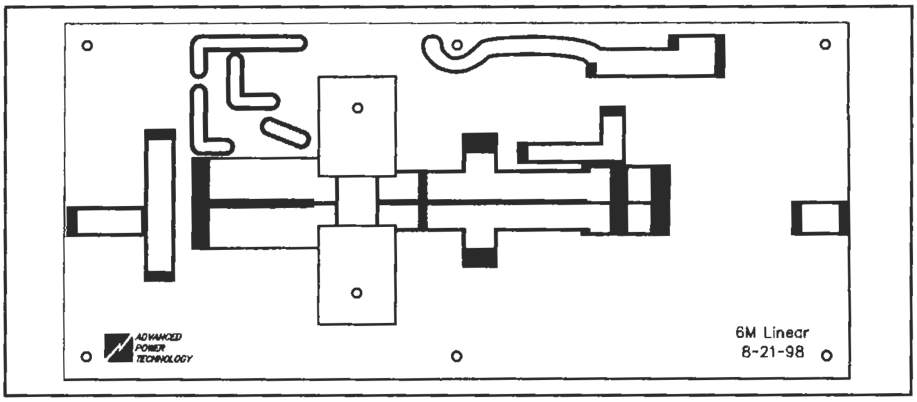

Fig 5—PCB artwork (not to scale, original size 3.35 (7.00 inches).

AB is calculated from the formula:

$$R_L = \frac{0.98\,V^2{}_{dd}}{2\,P_o} \qquad \text{(Eq 1)}$$

This is the load for each device, and P_O is one half of the total in a push-pull circuit. It is shunted by the output capacitance, C_{oss}. As was done for the gate circuit, both output impedances are series-connected to represent the total output impedance. The result for both devices in push-pull is 30.7 Ω in parallel with 75 pF, half the output capacitance of a single device. Though the design goal was 300 W PEP, the amplifier was actually designed for a 400 W load line. This gives a good compromise between efficiency and linearity.

In a classical design, a suitable transformer would be used to set the load impedance, and either the output power or the operating voltage would be adjusted to fit the available turns ratio. Normally, in a low-voltage HF design, the output capacitance is ignored because it is shunted by a much smaller load resistance. At 50 MHz, the effects of the output capacitance must be compensated, so a slightly different approach was taken in this circuit.

WinSMITH was used again to design the output matching. See Fig 2. The output impedance of 30 Ω is rotated south by the effect of the shunt output capacitance, C2. Two options present themselves for compensation. Some additional shunt capacitance could be added to further reduce the equivalent series real part to 12.5 Ω, a series L used to resonate the resulting series C, then a 4:1 transformer used to go to 50 Ω. However, building a reproducible low-loss 4:1 balanced transformer was very difficult, and compensating its leakage reactance further complicates the design. The second option was used: The equivalent series output capacitance was resonated first, then more series inductance was added to rotate the load all the way up to the $^1/_{50}$-mho conductance circle. Finally, a shunt capacitor was used to resonate the added X_L. The extra L and shunt C form an L-network, which transforms the 20-Ω equivalent series output impedance up to 50 Ω. This results in an easily duplicated design with a smooth, low-Q match.

The dc feed to the drains is provided through a shunt bifilar choke. At this frequency, most ferrite materials exhibit too much loss to be used at this impedance level. A powdered-iron core works famously here.

The balun-transformer function is provided by a simple coax and wire transformer. Two of the windings are provided by 50-Ω Teflon coax, and the third balancing winding by an additional single wire.

The bias network requires some explanation. Power MOSFETs have normal lot-to-lot variations in gate threshold voltage, V_{th}, forward transconductance, G_{fs}, and other parameters. A number of devices were checked for V_{th}, and they were all very close. They were all from the same die lot. The die lot number is marked on the package. For comparison, devices from another lot were checked and were uniformly a half a volt lower. Were this the case for the devices to be used in the amplifier, a dc block would need to be added to each side at the transformer, and the bias-feed network duplicated for each device. Since these devices were uniform, the additional complication of individual gate-bias adjustments was omitted in this design.

Because the gate-bias voltage required to maintain a particular value of idling drain current decreases as the temperature of the die increases, it is necessary to thermally compensate the gate-bias source, or the devices will "run away." A commonly available NTC resistor tracks the temperature of the case. (Refer to Fig 3.) This bias circuit has been in the literature for many years.[6] The ratio of R1 to R3 in part determines the degree of compensation. A smaller value of R1, or a larger value of R5, will increase the thermal sensitivity. A drop of thermally conductive glue keeps the thermistor in contact with the case. Proper operation is indicated when the set value of I_{dq} does not change after the heat sink gets hot from prolonged operation.

Construction

Refer to Fig 4 for the parts layout. A photo master of the artwork is shown in Fig 5. The original size of the artwork is 3.35x7 inches. The circuit board is 1-ounce-copper, double-sided $^1/_{16}$-inch G-10 PCB material. All four edges of the board and the three sides of the two rectangular cutouts for the transistors are wrapped with copper-foil tape that was soldered in place to

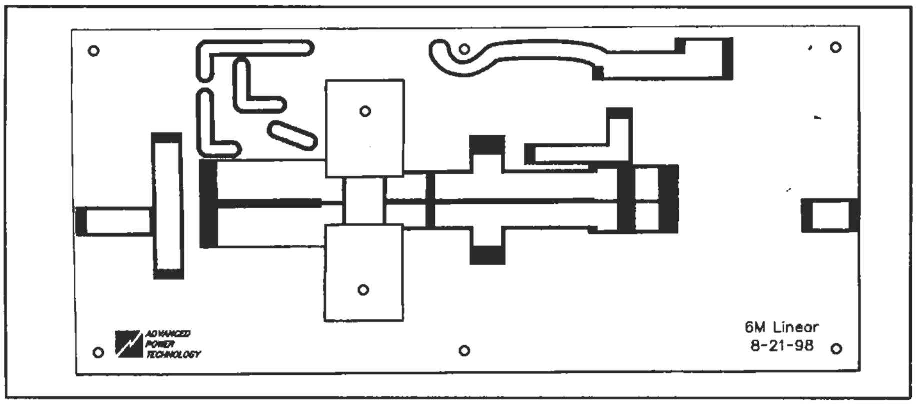

Fig 6—SSB IMD performance.

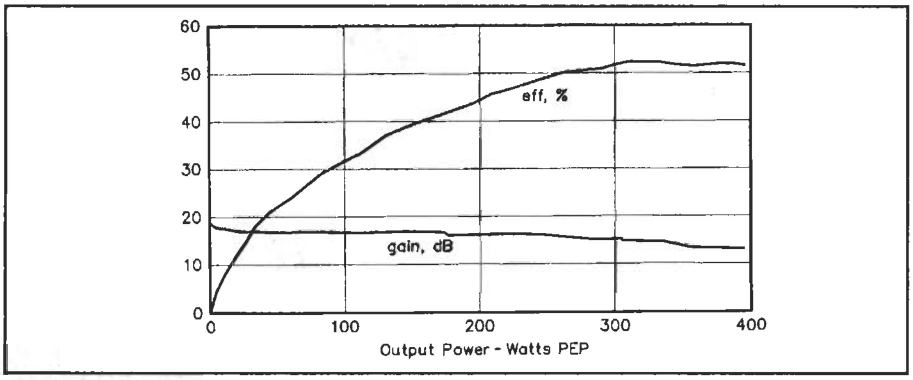

Fig 7—Amplifier efficiency and gain.

Table 1—Amplifier Parts List

C10—1µF, 35 V electrolytic capacitor
C1—215-790 pF Arco[8] #469 mica compression trimmer
C2, C6, C7—1000 pF, 500 V NPO chip cap, KD9 #2020N102J501P
C3—20-180 pF Arco #463 mica compression trimmer
C4-C5, C8-C9—0.01 µF, 500 V chip capacitor
D1—6.8 V, 1 W Zener diode
L1-L2—≈70 nH, 3t #18 AWG enameled wire 0.31-inch diameter 0.25-inch long
L3—2t #20 AWG on Fair-Rite #284301 0402 bead $\mu_i = 850$
Q1 —ARF448A
Q2—ARF448B
R3—10 kΩ NTC Fenwal[10] #140-103LAG-RB1
R6—1 kΩ, 0.5W 10-turn trimmer
T1—Primary 2t #20 PTFE, Secondary 3/16-inch brass tube on Fair-Rite #2843010402 balun core
T2—6t bifilar #20 PTFE on Amidon #T-94-2 toroid $\mu_i = 10$
T3—3t RG-316 coax, 3t #20 PTFE on three Fair-Rite #5961001801 toroids $\mu_i = 125$
TL1, TL2—30-Ω printed line, 0.6 inches long

provide a low-impedance continuous ground plane. The two cutouts for the transistors and the six mounting holes are the only holes in the board. All of the parts are surface mounted, which permits the board to be mounted directly to the heat sink.

This amplifier was built on a 7-inch length of AAVID #60765 heat-sink extrusion.[7] It is 3.5 inches wide, 1.5 inches deep and has nine fins. With 50 CFM of air blown across it, the devices will easily maintain thermal stability in a 30°C environment. The heat sink is not big enough for anything but very intermittent use without a fan to assure adequate airflow across the fins. The input and output connectors are each secured by two #4-40 screws and holes tapped into the base of the heat sink. A cover is recommended for safety; fairly high RF voltages are present.

Power Supply

Power for the amplifier needs to be fairly well regulated, since any ripple will show in the output signal as undesired 120-Hz AM. For on-air testing at the author's home, a very simple power supply was constructed using a 500 W, 120 to 240 V isolation transformer to drive a full-wave, centertapped rectifier circuit with 50,000 µF of filtering. Under SSB conditions, this is adequate. For CW, it needs some better regulation or the output power will sag over the length of a dash, and there will be some detectable hum. A regulated supply capable of providing 80 V at 6 A is needed.

Performance

This is the first-known, class-AB application of the ARF448 parts. Until now, the only other linear application is in a pulsed-mode linear amplifier for magnetic-resonance imaging. The SSB performance was encouraging because these devices were developed to serve the ISM plasma-generation market and no attention to linear performance was given in their design. The IMD performance with 200 mA of quiescent bias was better than expected. (See Fig 6.) The amplifier was tested with up to 0.5 A of I_{dq}. While the IMD performance improves somewhat at this level, the efficiency degrades significantly.

The gain and efficiency objectives have been met, as shown in Fig 7. The gain is 14.3 dB at 300 W PEP. The efficiency peaks at 51% at the same power. Under single-tone conditions, the drain efficiency is 61% at 250 W. The bandwidth of the amplifier is determined by the input network. The Smith-Chart plot of the input impedance shows the tracks for 50, 50.5 and 51 MHz. With the network adjusted for best match at 50.5 MHz, the SWR at the ± 0.5 MHz bandwidth points is 1.3:1. It would be difficult to increase this SWR bandwidth enough to cover the full 4 MHz of the amateur 6-m band without resorting to resistive loading, which would then reduce the available gain.

Conclusion

This paper has presented a 50 MHz, 300-W PEP linear amplifier using plastic-packaged, high-voltage MOSFET transistors. This is the first-known implementation of a full-duty-cycleclass-AB amplifier using these transistors. The design challenges, approaches to their solution and the resulting amplifier performance are shown. The parts, construction and mechanical layout all have been described in sufficient detail to permit duplication. The new line of plastic-packaged RF power transistors from APT offers designers a new cost-effective solution for efficient layout and performance.

Notes

[1] R. Frey, *A 50 MHz, 250 W Amplifier Using Push-Pull ARF448A/B*, APT9702, Advanced Power Technology Inc.

[2] R. Frey, "A Push-Pull 300 Watt Amplifier for 81.36 MHz," *Applied Microwaves and Wireless*, April 1998.

[3] Smith Chart is a trademark and property of Analog Instruments Co, New Providence, New Jersey.

[4] *WinSMITH*, copyright Eagleware Corp, 1995, available through Noble Publishing, Inc.

[5] Fair-Rite Products Corp, PO Box J, One Commercial Row, Walikill, NY 12589.

[6] H. Granberg, "Wideband RF Power Amplifier," *RF Design*, February 1988.

[7] AAVID Thermal Technologies Inc, Box 400, Laconia, NH 03247.

[8] Arco Electronics, 5310 Derry Aye, Agoura Hills, CA 91301.

[9] KD Components Inc, 2151 Challenger Way, Carson City, NV 89706.

[10] Fenwal Electronics, Inc, 450 Fortune Blvd, Milford, MA 01757.

By Paul Hewitt, WD7S From *QEX*, July 2000

A No-Bandswitch, Dual-Band VHF Desktop Amplifier

An old design idea together with new tubes yields legal-limit output on 6 and 2 meters from a small box.

Dual-band VHF amplifiers are by no means new in Amateur Radio. Several designs have appeared over the years, both commercial and homebrew.[1,2] The most recent commercial design was the Henry Tempo 6N2, now out of production for more than 20 years. There has never been a desktop capable of the new 1500-W PEP amateur power limit for both the 6- and 2-meter bands, until now. The amplifier I describe was designed primarily for meteor-scatter and weak-signal work, but the conservative design of the power supply and cooling system makes it well suited for EME (moonbounce) also. The amplifier uses a pair of Svetlana 3CX800A7s, operating in class AB2-push-pull on 2 meters and AB2-parallel on 6 meters. Features include: legal-limit output, no key-down time limit, no band switching, either-or operation, 50-60 W of drive for 1500 W PEP output on both bands, small size and light weight. In an effort to make this design easier to duplicate, off-the-shelf com-ponents and materials were used wherever possible. If you prefer not to "roll your own," Alpha Power should have a commercial version available soon.

Output Tank Circuit and Construction Details

Since this was a new tank design, I decided to use a box-in-a-box approach for the layout to make adjusting and debugging much easier. This should also make duplication of the tank circuits much easier for those of you who wish to use a separate, outboard power supply. The tank compartment (See Figs 1, 2 and 7) starts with a 7×7×2-inch Bud chassis with two surplus 11-pin ceramic sockets[3] mounted three inches apart (center-to-center) and four inches from the rear of the chassis. The sockets are mounted ¼ inch below the chassis in 2½-inch-square holes with each socket oriented at 45° to the hole (Fig 3). This is similar to an Eimac 2216 socket for the 8877. The Eimac #1906 chimneys were used in the normal manner. This is probably the upper frequency limit for this mounting scheme due to possible instability problems at UHF. One problem encountered while mounting the sockets this way is the small amount of tube exposed above the chimney to grasp for removal. This is easily solved by connecting several ¼-inch-wide hose clamps together end to end, then clamping them to the anode cooler to serve as a temporary handle.

The front, top and sides of the tank com-

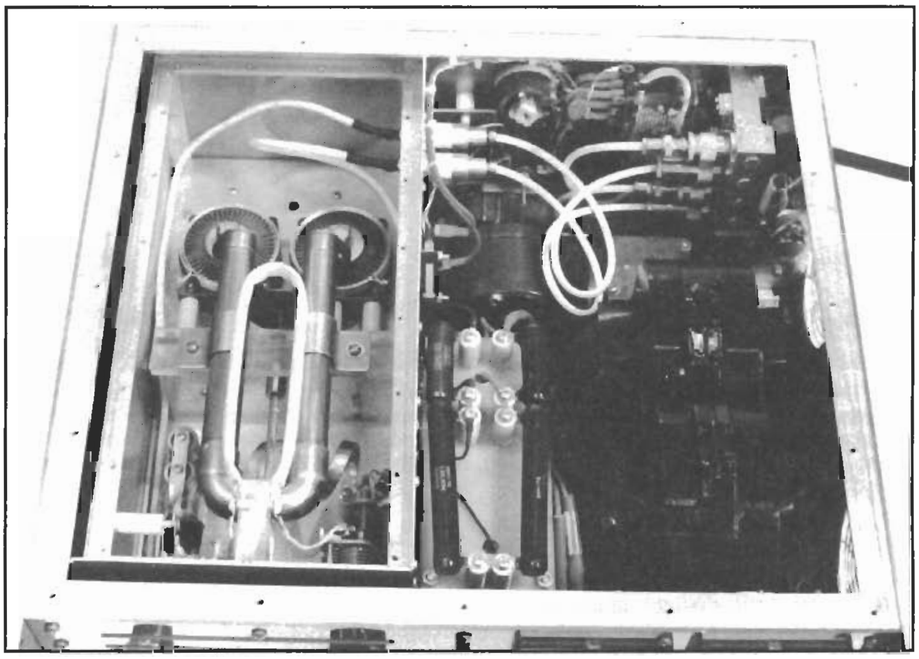

Fig 1—A top view of the amplifier. The plate transformer in the front right and high-voltage power assembly next to it occupy the majority of the control side with the blower, filament and control transformers sitting just behind. The tank compartment with the cover removed shows the plate line on its Plexiglas stand.

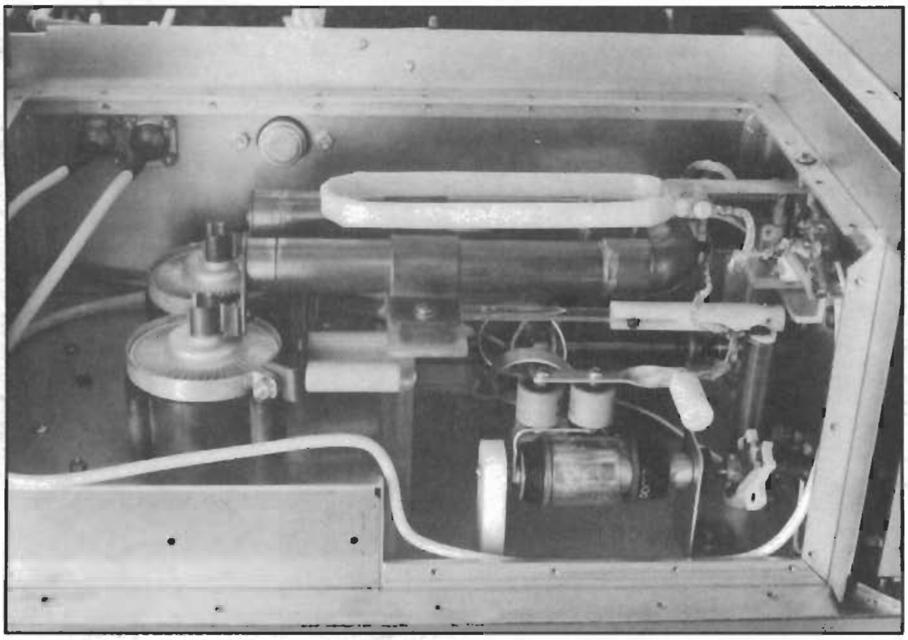

Fig 2—Side view of tank compartment. The 2-meter load capacitor on its ceramic standoffs mounted to the Plexiglas stand is in the center with the 6-meter tank coil just visible below the spline shaft assembly. The 6-meter tune capacitor on its **L** bracket is in the lower right front.

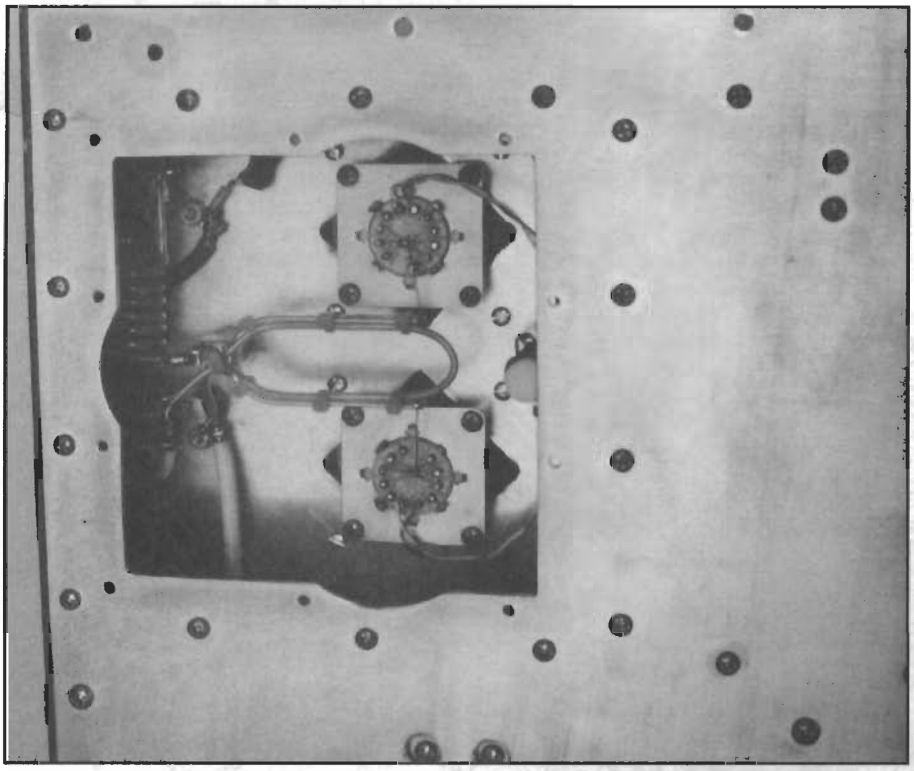

Fig 3—Bottom view of amplifier with cathode compartment cover removed The filament choke is to the right with the 2-meter input network visible centered between the tube sockets. Both input tuning capacitors are mounted on the rear panel and one of the 6-meter input coils is just visible on the left. Note the use of coax for exciter lines to keep feedback paths minimized.

partment are all made from 1/16-inch aluminum with 1/2×1/2×1/8-inch aluminum angle brackets at all corners. The completed compartment measures $7\times6\frac{1}{4}\times14\frac{1}{2}$ inches (WHD). The front and right-side compartment panels extend to the top of the main cabinet cover to keep inlet and outlet air separated. The compartment cover (not shown) has 176 0.2-inch holes above the tubes for cooling exhaust. When this box is attached to the floor of the main cabinet, it leaves a 4×7-inch hole at the back before the rear panel is attached. This opening serves as access for a dip meter when checking the 2-meter tank resonance and tuning range.

The 2-meter tank circuit consists of a shorted $\lambda/4$ balanced line section, tuned by a homebrew split-stator capacitor at the tube end of the line. Designing this type of tank circuit is made very pain-less with information and design exam-ples provided in references.[4] There are several advantages of using a push-pull tank circuit instead of a parallel arrangement at 144 MHz. In the push-pull arrangement, C_{out} and C_{in} of the tubes are in series, allowing a total C_{out} and C_{in} that is one quarter of that with the parallel arrangement. This allows lower values of loaded Q, resulting in higher efficiency and reduced compo-nent heating, minimizing thermal drift. The low value of C_{out} in this case allows the use of a $\lambda/4$ line in place of a $\lambda/2$ line, increasing bandwidth and decreasing size.

The line section, L2 of Fig 4, is made from 3/4-inch type-K copper pipe; the shorted end is made from one standard 90° copper elbow and one "street" 90° copper elbow. The open ends of the line are closed with brass plugs of the same outside diameter as the pipe, with a portion machined to fit inside the pipe. Standard copper pipe caps could be substituted for these plugs. The total length of the line section is $7\frac{3}{4}$ inches, from the ends to the inside of the shorted end. The plate-to-anode connections are made with 1/16-inch copper plate fastened between copper fuse clips on the anode connectors and threaded holes in the ends of the brass plugs.

The line section is supported by a Plexiglas stand and held in place by two copper clamps. The stators of the 2-meter **TUNE** capacitor, C3, are sup-ported by $1\frac{1}{2}$-inch ceramic standoffs attached to the Plexiglas stand (Fig 5). A 3/8×5/8-inch ear from each stator attaches to a 3/8-inch-wide strap around each of the tubes' anode coolers. The rotor of C3 is a two-inch disk of 1/16-inch copper, which is mounted to a piece of 1/4×28 brass all-thread rod. The threaded rod is held by a brass fitting (tapped for 1/4×28) that is attached to the Plexiglas stand. The threaded rod is coupled to a sliding spline shaft consisting of 1/4-inch Plexiglas rod inside a 1/2-inch piece of Teflon rod that has been drilled through the center. The Teflon tube is attached to a short piece of 1/4-inch stainless-steel rod that exits the tank compartment through a 1/4-inch panel bushing. This shaft is coupled to the turns counter with a synchronous gear belt (Fig 6). This shaft relocation via the gear belt serves only to improve the front-panel appearance; it may be omitted if desired.

The 2-meter tank circuit is coupled to the antenna by an adjustable reso-nant link. The link position is adjusted by a #10-24 screw through a threaded Plexiglas block, to which the link is attached (Fig 2). One end of the adjusting screw is supported by the top-cover angle bracket and the other end by a 2-inch-long piece of 1/2-inch angle mounted to the front panel of the compart-

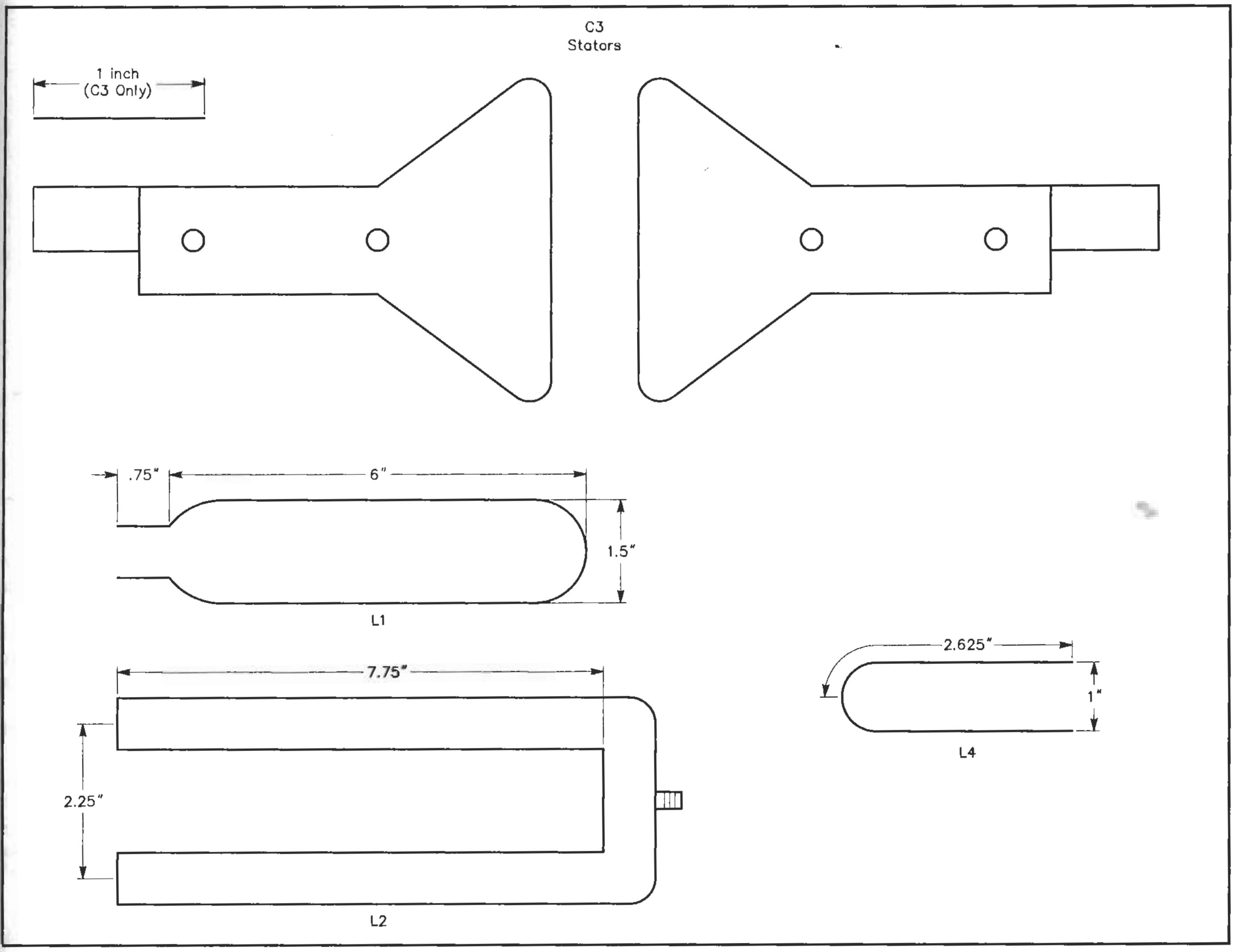

Fig 4—Details of 2-meter input and output tank indicators and the 2-meter TUNE capacitor stators.

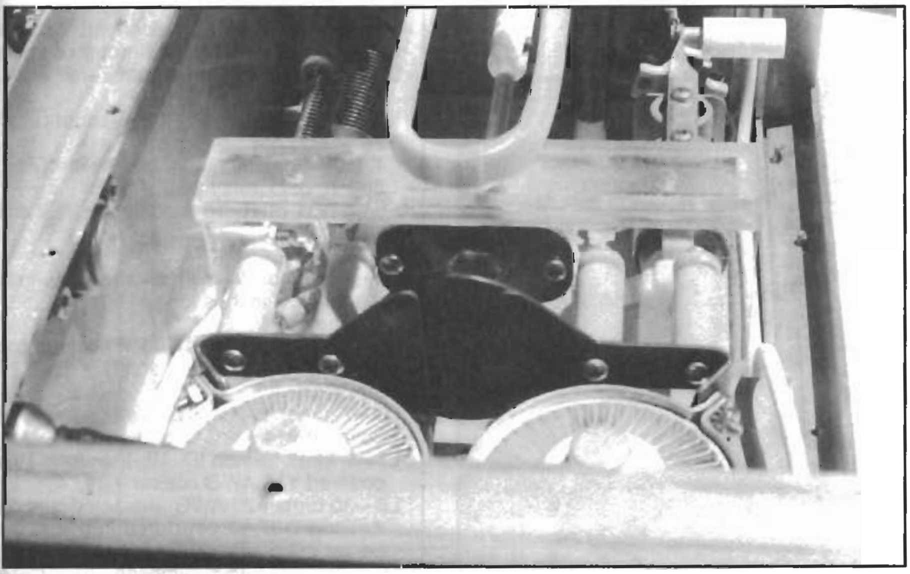

Fig 5—Rear view with the plate line removed showing the 2-meter TUNE capacitor.

ment. The link is made from $^3/_8 \times ^1/_{16}$-inch silver-plated copper strap, covered with a $^3/_8$-inch Teflon sleeve; it is attached to the Plexiglas block with four #10-32 nylon screws. Silver-plated coax braid attaches the link to the 2-meter **LOAD** capacitor, C2, and the isolation relay, K3. The link-adjustment screw is accessed through $^3/_8$-inch hole in the main cabinet top cover during initial setup, then the hole is covered with a plastic plug.

As with all push-pull circuitry, the key to successful operation is symmetry. All stray capacitance must be divided equally between both sides of the line and tubes. Also, try to keep ferrous materials out of the tank compartment to prevent imbalance.

The 6-meter tank is a normal π network designed for a plate-load resistance of 1150 W with a loaded Q of approximately 22. The design parameters are shown in Table 1. The input of the π network is connected to the center of the shorted end of the balanced line section where a #10-32 brass stud has been soldered in place. This is the low-impedance point of the line section, and anything but a dead short can be attached here without affecting the performance of the line.[5] To verify that the physical center of the line is also the RF center, couple a dip meter to the line and tune for a dip around 144 MHz. Then without moving the dip meter (DM), touch the tip of a lead pencil at points along the line until you find the spot that has the least effect on the DM. This is the attachment point for the plate choke and the 6-meter tank.

To keep the loaded Q of the π network as low as possible, full 51-54 MHz opera-

Table 1—Print out of operating parameters for the 6-meter input and output networks

```
The values for the PI matching network were calculated with       Conduction Angle     =    187.8     Degrees
Elmer (W5FD) Wingfield's new formulas found in the more recent    Peak Grid Voltage    =     36.8     V
ARRL handbooks.                                                   Zin @ Cathode        =     24.6     Ω
FREEWARE Courtesy of KD9JQ                                        Pin Drive (PEP)      =     41.0     W
Triode Amplifier Program  Version 2.0                             P0 @ Plate (PEP)     =   1600.0     W
For Grounded Grid Operation                                       P0 to Load (PEP)     =   1637.9     W
(2) USER Biased 3CX800A7 at 50.0 MHz Rated for FORCED AIR         DC Power Input       =   2405.5     W
DC Plate Volts        =   2150.0    V (2500 V Max)                Efficiency           =     66.5     %
Max Plate Voltage     =   2150.0    V                             Power Gain           =     16.0     dB
Peak Plate Swing      =   1900.0    V                             Cath to Grid Cap     =     52.00    pF
Min Plate Voltage     =    250.0    V                             Plate to Grid Cap    =     12.20    pF
Plate Current Peak    =      3.368  A
Plate Current DC      =      1.119  A                             T Match Input                   Pi Match Output
Grid Current DC       =      0.063  A                             RS   =   50.0 Ω                 RP  =  1154.8 Ω
Cath Current Peak     =      3.651  A                             L2   =    0.521 µH              CI  =    50.0 pF
Design Plate RL       =   1128.1    Ω                             CI   =   44.924 pF              L1  =     0.2377 µH
RL for Matching       =   1154.8    Ω                             L1   =    0.422 µH              C2  =   232.0 pF
Plate Dissipation     =    805.5    W (1600 W Max)                RFC  =    0.195 µH              RL  =    50.0 Ω
Grid Dissipation      =      3.2    W (8 W Max)                   Zin  =  22.6 –j 9.9 Ω           QL  =    22.0
Cathode Bias          =      8.2    V                             QL   =    5.0
```

tion was not attempted. The 6-meter **TUNE** capacitor, C1, operates very near its minimum capacitance because of strays and the C_{out} of both tubes. These strays account for a large portion of the 6-meter tune capacitance, which causes the majority of the tank circulating current to flow through the blocking caps, C7-C8. I used parallel Centralab 858 "door-knobs" for the blocking capacitors and I haven't experienced any problems. Capacitors with larger current ratings (such as HT-57s) would be a better choice for longer duty cycles.

The 3-30 pF vacuum-variable capacitor is mounted to the floor of the tank compartment with an L-bracket bent from 1/16-inch aluminum. Use the bearing retainer nut to attach the capacitor to the bracket, but be sure to add shims to the bearing to make up for space lost by the thickness of the bracket, or backlash will occur. The other end of the capacitor is supported by a #6-32 brass screw in the end of the capacitor that passes through a piece of Teflon bar stock mounted to the floor of the cabinet. A #6-32 hex nut holds the end of L3 and a piece of 3/8-inch strap (which supports C7 and C8) tight to the end of the vacuum cap. The other end of L3 is supported by a 1/2×1-inch Teflon standoff. Silver-plated braid ties this end to the 6-meter **LOAD** cap, which is mounted on the sidewall of the tank compartment.

Setting the Output Tanks

The 6-meter tank can be checked for tuning range with an SWR analyzer at the output connector. Connect a 1.1 kΩ, noninductive resistor from one of the tube anodes to ground to simulate the 1100-W plate-load resistance. You should be able to achieve a 1:1 match at both ends of the desired tuning range. Some stretching or compression of L3 may be required to bring the tuning range to 50-51 MHz. Be sure the top cover of the tank compartment is in place, as the stray capacitance it introduces accounts for a portion of the tuning capacitance for 6-meters.

The 2-meter tank was set up entirely

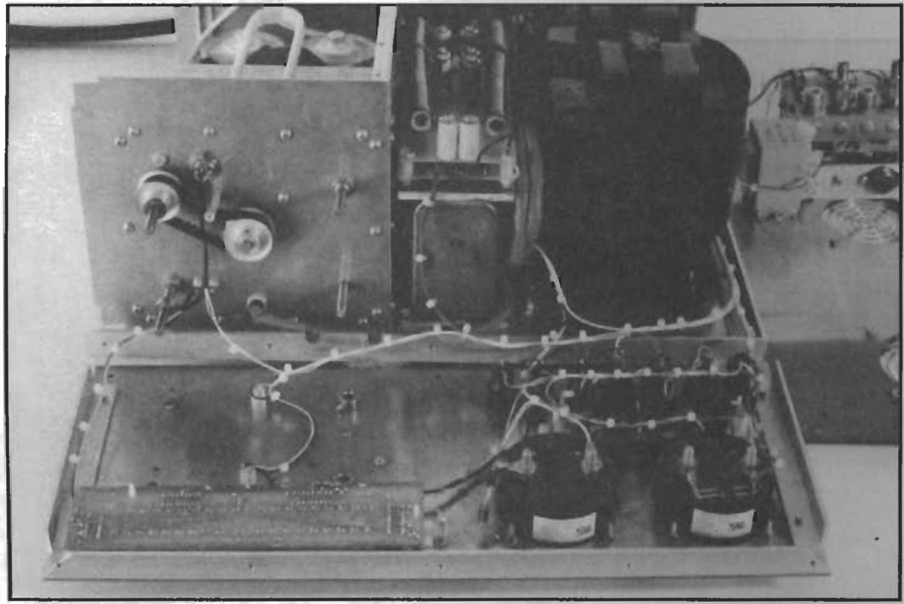

Fig 6—Front view of the amplifier with front panel folded down showing the location of high-voltage power-supply components at right and the gear-belt layout on the left. The bar-graph display PC board is visible in the lower left with the control board just beyond the two panel meters.

Fig 7—RF deck schematic. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

C1—3-30 pF vacuum-variable capacitor
C2—150 pF air-variable, 1400 V capacitor
C3—Split-stator air-variable (see text and Fig 4)
C4—325 pF air-variable, 1400 V capacitor
C5, C6—4.3-75 pF APC-style air-variable trimmer capacitor
C7, C8—1000 pF, 5 kV doorknob capacitor
C9—2500 pF, 2.5 kV feedthrough capacitor
D1—8.2 V, 50 W stud-mount Zener diode, 1N2806B
K3—Jennings RJ1A SPDT vacuum relay, 26.5 V dc coil
K4, K5—DowKey model 260B with "C" option, coax relay 26.5 V coil
K6—SPDT relay, 10 A contacts, 24 V dc coil
L1—See Fig 4
L2—See Fig 4
L3—2 turns of 3/8×1/16-inch strap, 1 1/2-inch diameter, 2 inches long
L4—See Fig 4
L5—8.5 inches of #14 Teflon covered copper wire with 6 inches tied tightly inside L4
L6—7 turns, 1/2-inch ID×1 1/8-inch long Teflon covered #14 AWG copper wire
L7—9 turns 1/2-inch ID×1 1/2-inch long Teflon covered #14 AWG copper wire
L8—30 turns #26 AWG enameled wire on an Amidon T-37-17 powdered-iron core over center conductor of coax
M1—0-2 A, 3.5-inch Simpson panel meter
M2—0-100 mA, 3.5-inch Simpson panel meter
RFC1—36 turns #18 AWG enameled wire, tight-wound on 1/2-inch-diameter Teflon rod
RFC2, RFC3—Ohmite Z-50
RFC4—Ohmite Z-144
V1, V2—Svetlana 3CX800A7

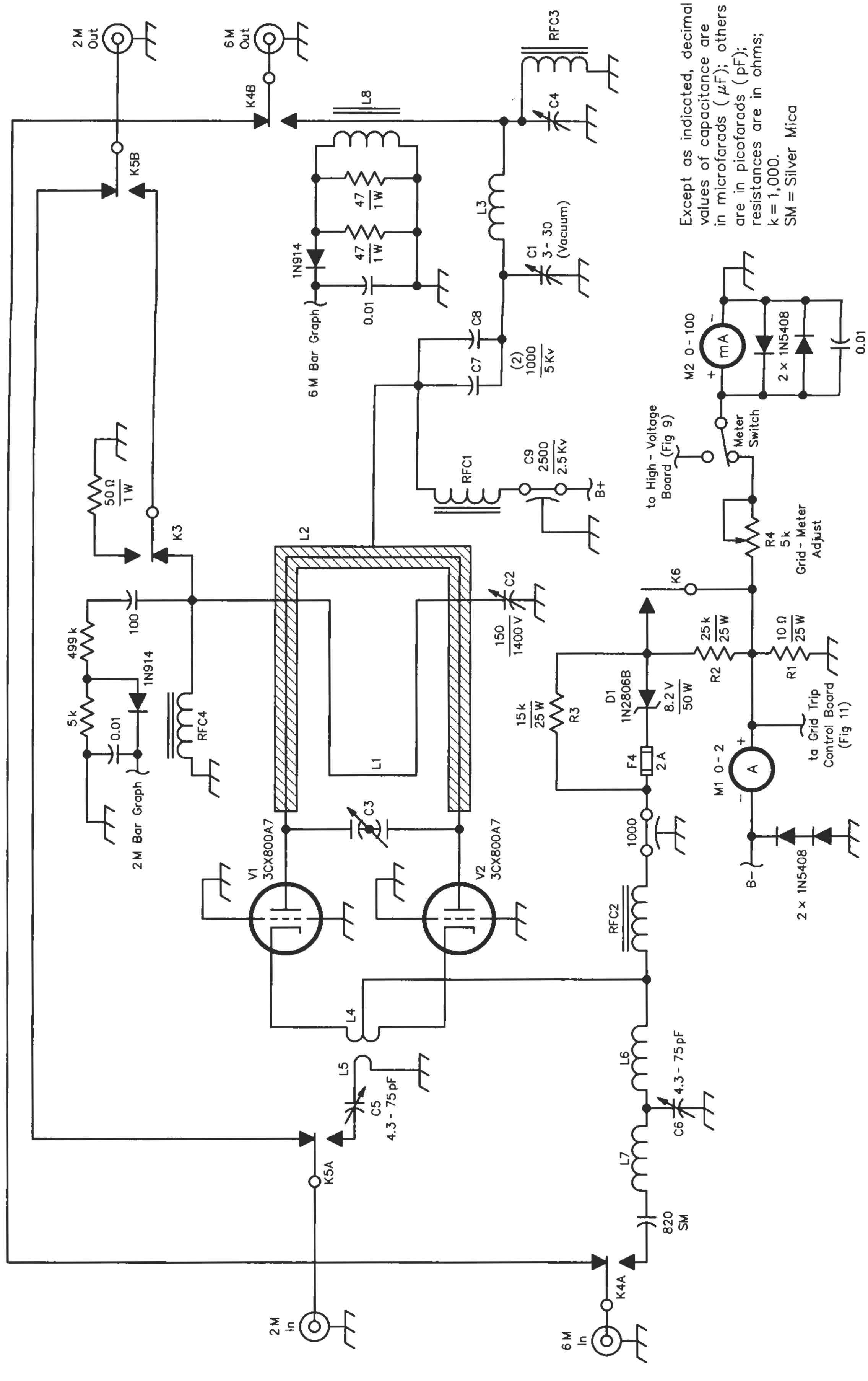
2 M Out
6 M Out
K4B
L8
RFC3
C4
K5B
47 1 W
47 1 W
1N914
0.01
6 M Bar Graph
L3
C1 3 - 30 (Vacuum)
C8
C7
(2) 1000 5 Kv
50 Ω 1 W
K3
RFC1
C9 2500 2.5 Kv
B+
to High - Voltage Board (Fig 9)
Meter Switch
M2 0 - 100
mA
2 × 1N5408
0.01 1 kV
L2
C2
150 1400 V
499 k
100
1N914
5 k
0.01
2 M Bar Graph
RFC4
L1
K6
R4 5 k Grid - Meter Adjust
25 k 25 W
R2
10 Ω 25 W
R1
15 k 25 W
R3
D1 1N2806B
8.2 V 50 W
F4 2 A
to Grid Trip Control Board (Fig 11)
M1 0 - 2
A
1000
C3
V1 3CX800A7
V2 3CX800A7
B−
2 × 1N5408
RFC2
L4
L5
C5 4.3 - 75 pF
L6
C6 4.3 - 75 pF
L7
820 SM
K5A
K4A
2 M In
6 M In

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms;
k = 1,000.
SM = Silver Mica

with a dip meter inserted through the opening in the back of the compart-ment. Set the output link about $^{3}/_{8}$-inch from the top plane of the line; this is close to the final position and adds stray capacitance. To keep loaded Q low, try to set the line length so it doesn't take very much tune capacitance to resonate the line at 144 MHz. Start with a line section a little longer than shown and carefully trim it to resonance. Remem-ber that 1 inch of line length equals approximately 10 MHz of tuning range! The tank should resonate at 144 MHz with about $^{1}/_{4}$ inch of space between the plates of C3. When the desired tuning range is found, install a $^{1}/_{4}$×28 lock nut on the threaded shaft of C3 to stop travel at the upper frequency limit. This will prevent accidental tuning of the tank to the third harmonic of 50 MHz, which could potentially damage the front end of a 2-meter exciter. Remember that the isolation of most coax relays is very poor at 144 MHz: It was only 35 dB with the DowKey 260B used here. For this reason, a Jennings RJ1A vacuum relay, K3, was added in series with the 2-meter transfer relay. The vacuum relay disconnects the 2-meter link from the coax and terminates the coax in a resistive load during 6-meter operation. Make sure the electrical length of the coax between the relays is more than $\lambda/10$ and less than $3\lambda/8$ at 144 MHz. With these safety measures in place, third-harmonic energy at the 2-meter input port was measured at −17 dBm during 1500-W 6-meter operation.

Input Networks

The 2-meter input network consists of a single tuned-link, air-core transformer (Figs 3 and 4). This coupling method requires slightly more drive than a link-

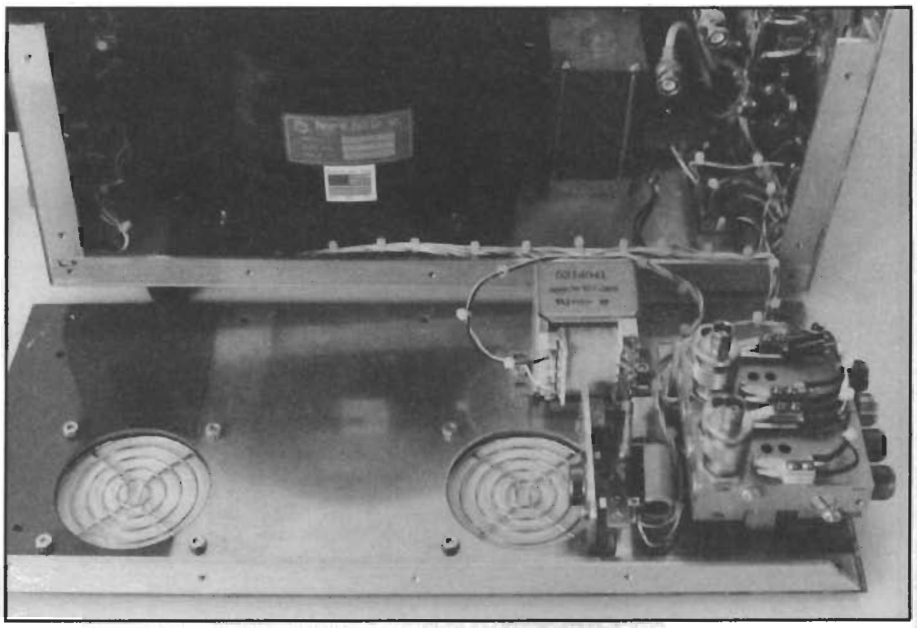

Fig 8—View of the right-hand side panel folded down showing the two transfer relays and the low-voltage power supply. The filament transformer is visible to the right of the plate transformer. Note the use of 80-mm fan covers for the cooling air inlets.

Table 2—Operating parameters of the amplifier

Plate voltage no load, 2350 V
Plate voltage @ 1 A, 2150 V
Zero signal plate current, 35 mA
6-Meter drive, single tone, 49 W
6-Meter power output,1420 W
6-Meter plate current, 980 mA
6-Meter grid current, 75 mA
Apparent efficiency, 67.4% (feed-through power not subtracted)
2-Meter drive, Single tone, 50 W
2-Meter power output, 1275 W
2-Meter plate current, 950 mA
2-Meter grid current, 68 mA
Apparent efficiency 62.4% (feed-through power not subtracted)

Table 3—Parts suppliers

Fair Radio Sales Co, Inc
1016 East Eureka St
PO Box 1105
Lima, OH 45804
tel 419-227-6573, 419-223-2196
fax 419-227-1313
e-mail **fairadio@wcoil.com**
URL **http://www.fairradio.com/**

Mouser Electronics
2401 Hwy 287 N
Mansfield, TX 76063
tel 800-346-6873
fax 817-483-0931
e-mail **sales@mouser.com**
URL **http://www.mouser.com/**

Newark Electronics
4801 N. Ravenswood Ave
Chicago, IL 60640-4496
tel 800-463-9275, 773-784-5100
URL **http://www.newark.com/**

Peter W. Dahl Co, Inc
5869 Waycross Ave
El Paso, TX 79924
tel 915-751-2300
fax 915-751-0768
e-mail **pwdco@pwdahl.com**
URL **http://www.pwdahl.com/**

RF Parts Co
435 S Pacific St
San Marcos, CA 92069
tel 760-744-0700, 800-737-2787
(orders only)
fax 760-744-1943
e-mail **rfp@rfparts.com**
URL **http://www.rfparts.com/**

Surplus Sales of Nebraska
1502 Jones St
Omaha, NE 68102-3112
tel 402-346-4750, 800-244-4567
(Orders only)
fax 402-346-2939
e-mail **grinnell@surplussales.com**
URL **http://www.surplussales.com/**

Svetlana Electron Devices
8200 S Memorial Pkwy
Huntsville, AL 35802
tel 256-882-1344, 800-239-6900
fax 256-880-8077
e-mail **sales@svetlana.com**
URL **http://www.svetlana.com/**

coupled, half-wavelength resonant line, but saves a lot of space. Two references[6, 7] give formulas and rules of thumb for designing this type of transformer, but it still required a lot of "cut and try" before suitable sizes for L4 and L5 were found. The loaded Q of the resonant input link is approximately 3.5 before the coupled-in resistance from the secondary modifies it. The degree of mutual coupling of air-core transformers is an elusive value that makes the final loaded Q difficult to calculate. The input tuning capacitor, C5, could be changed to a 25-pF unit since the one used here turned out to be much larger than needed. Both sides of C5 are above chassis ground; it requires a nonmetallic screwdriver to adjust.

Six-meter drive and B– are applied to the center tap of L4, providing parallel cathode drive for 50-MHz operation. A common **T** network, with a loaded Q of five, matches the 50-W line to the 24.6-Ω input impedance of the parallel tubes. All the cathode pins of the sockets are tied together with buss wire in a star pattern. One-inch-long bus-wire leads connect the ends of L4 to the center of the bus-wire stars on the sockets. Again, symmetry is all-important in balanced operation.

Both input networks were set up with an antenna analyzer and two 50-W carbon resistors to simulate the input load impedance. Tack-solder the resistors from cathode to ground on each socket with the shortest-possible leads. Short leads are very important on 2 meters because lead inductance and stray capacitance become quite significant. The resistors present

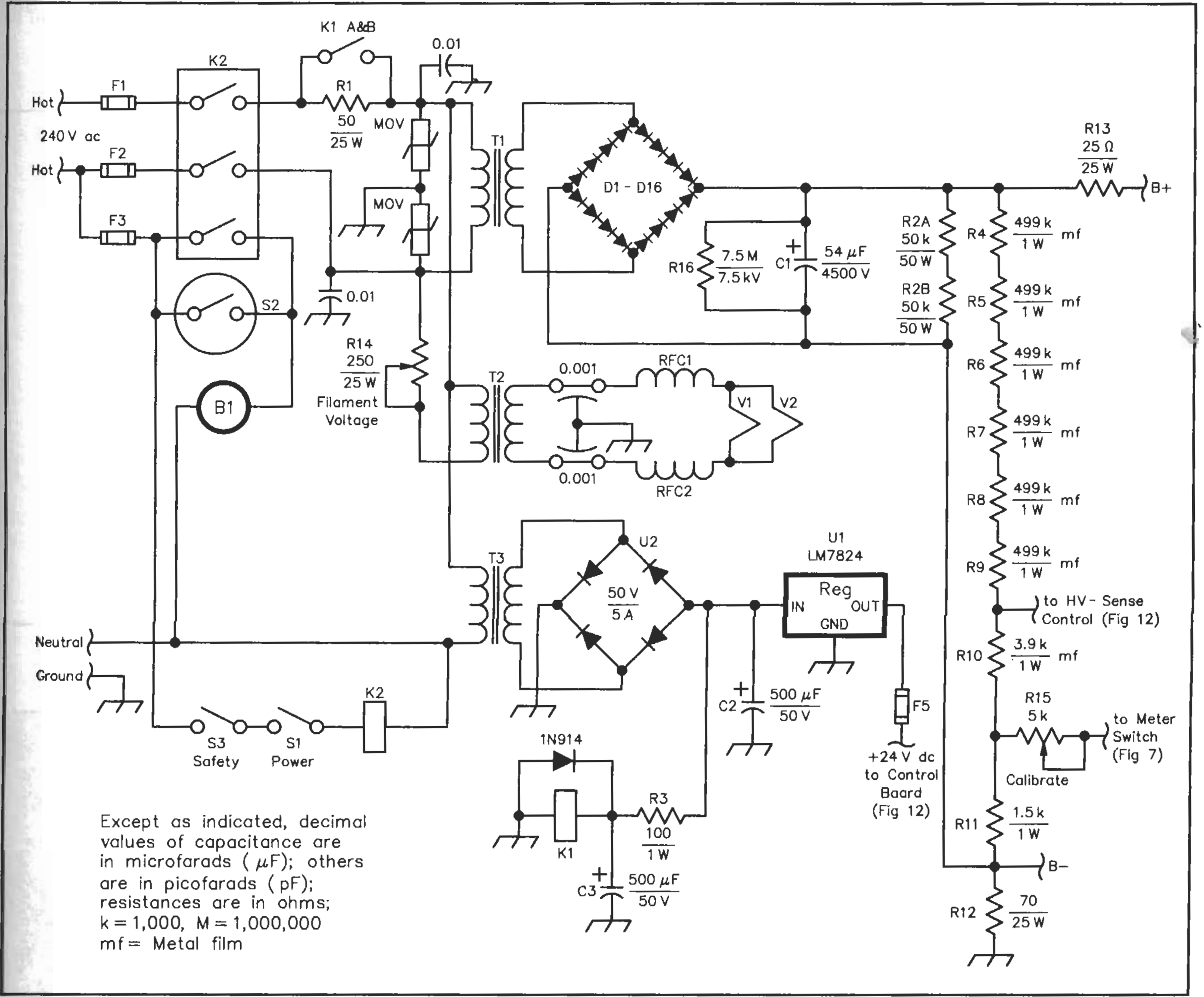

Hot
240 V ac
Hot
F1
F2
F3
K2
K1 A&B
R1
50
25 W
0.01
MOV
MOV
T1
D1 - D16
R16
7.5 M
7.5 kV
C1
54 µF
4500 V
R2A
50 k
50 W
R2B
50 k
50 W
R4
R5
R6
R7
R8
R9
499 k
1 W
mf
R13
25 Ω
25 W
B+
S2
0.01
R14
250
25 W
Filament Voltage
B1
T2
0.001
0.001
RFC1
RFC2
V1
V2
to HV - Sense Control (Fig 12)
R10
3.9 k
1 W
mf
T3
U2
50 V
5 A
U1
LM7824
Reg
IN
OUT
GND
Neutral
Ground
K2
S3
Safety
S1
Power
C2
500 µF
50 V
F5
+24 V dc to Control Board (Fig 12)
R15
5 k
Calibrate
to Meter Switch (Fig 7)
1N914
K1
R3
100
1 W
C3
500 µF
50 V
R11
1.5 k
1 W
B–
R12
70
25 W

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000
mf = Metal film

Fig 9—Power-supply schematic. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

B1—Dayton #4C761 squirrel-cage blower
C1—54 µF, 4500 V oil-filled electrolytic capacitor
C2, C3—500 µF, 50 V electrolytic capacitor
D1-D16—1N5409 1200 PIV, 3 A Silicon rectifiers, four diodes per string
F1, F2—20 A, 250 V ceramic fast-blow fuse
F3—3 A, 250 V fuse
F5—2 A slow-blow fuse
K1—DPST relay, 25 A contacts, 24 V dc coil
K2—3PST relay, 25 A contacts, 120 V ac coil
MOV—130 V metal oxide varistor, V130LA5
R1—150Ω 25 W wire-wound resistor
R2A, R2B—50 kΩ 50 W wire-wound resistor
R3—100 Ω, 1 W carbon resistor
R4-R9—499 kΩ, 1 W metal film resistor
R10—3.9 kΩ, 1 W metal film resistor
R11—1.5 kΩ, 1 W metal film resistor
R12—70 Ω, 25 W wire-wound resistor
R13—25 Ω, 25 W wire-wound resistor
R14—250 Ω, 25 W rheostat
R15—5 KΩ, 10-turn potentiometer
RFC1, RFC2—12 bifilar turns of #18 AWG enameled wire on 1/2-inch-diameter Teflon rod
S1—SPST 5 A lighted panel switch
S2—Temperature snap-disc control Grainger # 2E245
S3—SPDT 5 A microswitch
T1—Peter Dahl plate transformer: 240 V primary, 1800 V, 1 A CCS secondary
T2—240 V primary, 16 V, 5 A secondary
T3—120 V primary, 20 V, 2 A secondary
U1—LM7824 TO-3 case
U2—5 A, 50 V rectifier bridge

25 W to the parallel 6-meter network and 100 W to the 2-meter push-pull network. Connect the analyzer to the respective input ports and adjust each network for lowest SWR. Some adjustment of the 6-meter input coils (L6 and L7) may be required to achieve a 1:1 match. If the 2-meter network does not present a good match, some adjustment in length of L4 and L5 may be necessary. Again, the covers must be in place and tubes must be in their sockets when measuring the match. Both filament leads and the filament choke must be in their final positions because the cathode-to-filament capacitance affects input tuning at 144 MHz. Remember, this analyzer method only gets you close to the final tuning points; settings will be different under live conditions. Finally, don't forget to remove the temporary resistors from the cathode and tank circuits after the initial adjustments are done.

AC Mains and Low-Voltage Power Supplies

Several protective measures are built into the amplifier to protect the tubes and the operator. The ac mains are brought into the amplifier with four-conductor cord to keep the neutral and ground separated per the *NEC* (Fig 9). Three fuses are used to keep the blower's ac source separate from the fused plate transformer. This was done so the blower's cool-down delay still functions after a high-voltage fault, which removes ac from the entire amplifier in dual-fuse designs. A thermal snap switch in the tank compartment keeps the blower running after shutdown or a high-voltage fault if the exhaust air has been above 110°F and is not yet below 90°. This only happens after several minutes of continuous operation. You may want to use a switch with slightly higher ratings (120°F on, 110°F off) if your shack is often warmer than the 90° off point used here. The 3CX800A7's filaments dissipate only 20 W each during cutoff and natural convection is more than enough to cool the tubes when shut off.

AC voltage for K2, the main control relay, is also taken from F3, the blower fuse, to make sure blower voltage is available before the amplifier can be powered up. A safety switch that closes when the top cover is in place supplies ac to S1, the main power switch. Power-supply inrush protection is provided by K1, which closes approximately one second after K2 and effectively removes R1 from the transformer primaries. The delay period is set by R3 and C3. There are two MOVs across the ac lines after the contacts of K2 and K1. Arcing at the relay contacts can and will produce voltage spikes and spike protection on the fuse side of K2 will not always save the high-voltage diode strings and other components. Filament voltage is adjusted at the primary of T2 by R14, a 250-Ω, 25-W rheostat. Filament voltage is measured at the filament choke via two leads brought to the back panel of the amplifier. Purists may want to measure voltage at the filament pins, but the low filament-current demand of these tubes makes the voltage drop between the sockets and the choke negligible. The rheostat also provides passive filament-inrush protection.

The low-voltage supply—consisting of T3, U1 and U2—supplies regulated 24 V at 1 A. The bridge rectifier and 24-V regulator are mounted to a piece of $^3/_4$×1-inch aluminum angle that is placed directly in the cooling-air inlet path (Fig 8).

High-Voltage Power Supply

The high-voltage power supply consists of a Peter Dahl Hypersil plate transformer and a full-wave bridge rectifier with capacitor input filter. The supply is capable of producing 2150 V at 1 A CCS. The entire supply is assembled as one piece, then installed on the floor of the main cabinet next to the plate transformer (Fig 6). The oil-filled filter capacitor is sandwiched between two 3×6$^1/_2$-inch pieces of Plexiglas, in turn held together with four #10-24×5-inch-long flat-head screws. The rectifier strings and high-voltage-meter multiplier resistors are mounted to a 3×6$^1/_2$-inch PC board supported by $^1/_4$-inch spacers on the same four screws. A 3×6$^1/_2$-inch piece of fiberglass board holding the two bleeder resistors, B+ current limiting resistor and B– float resistor tops off the

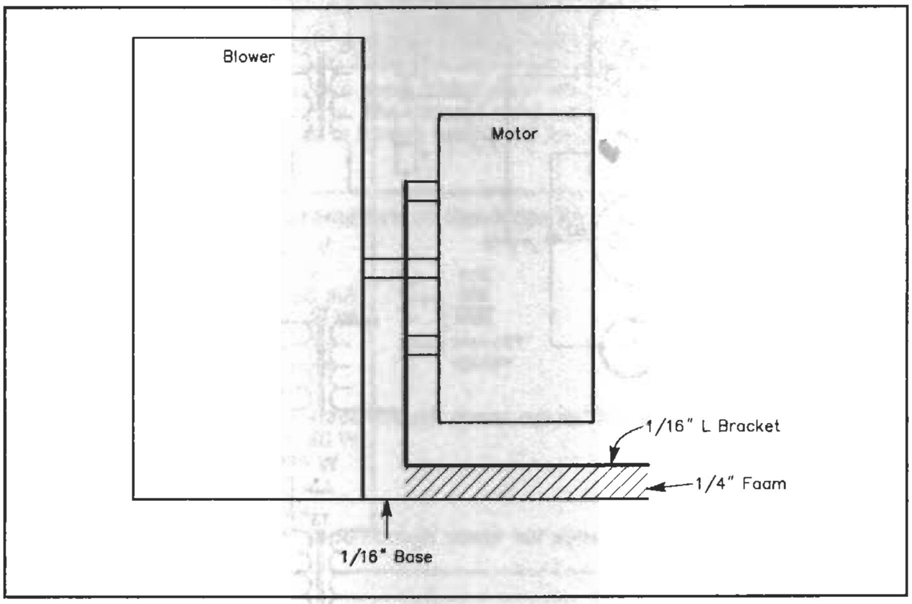

Fig 10—Layout of the blower modifications for sound reduction.

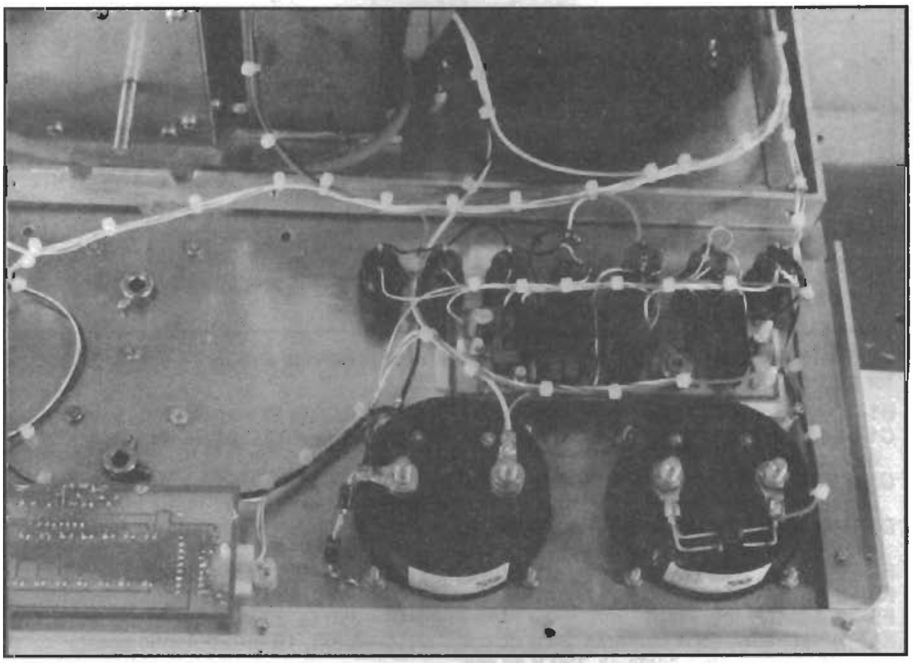

Fig 11—View of control board.

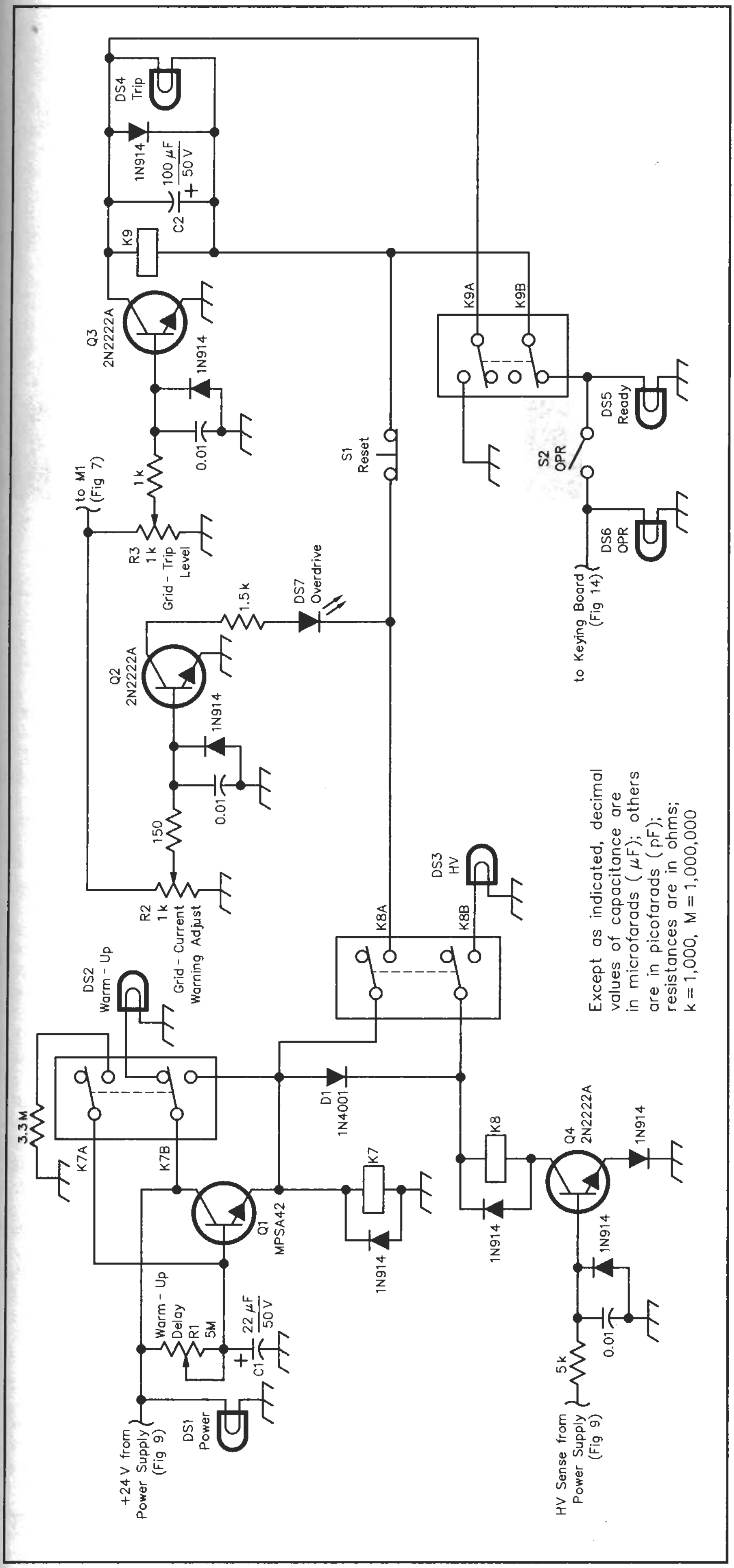
DS4 Trip
1N914
100 µF 50 V
C2
K9
Q3 2N2222A
1N914
0.01
1k
to M1 (Fig 7)
R3 1k Grid - Trip Level
K9A
K9B
DS5 Ready
S2 OPR
S1 Reset
DS6 OPR
to Keying Board (Fig 14)
DS7 Overdrive
1.5 k
Q2 2N2222A
1N914
0.01
150
R2 1k Grid - Current Warning Adjust
DS3 HV
K8A
K8B
DS2 Warm - Up
3.3M
K7A
K7B
D1 1N4001
K8
K7
Q4 2N2222A
1N914
Q1 MPSA42
1N914
1N914
1N914
Warm - Up Delay R1 5M
22 µF 50 V
C1
+24 V from Power Supply (Fig 9)
DS1 Power
5k
0.01
HV Sense from Power Supply (Fig 9)

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000

Fig 12—Control schematic. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

C1—22 µF, 50 V electrolytic capacitor
C2—100 µF, 50 V electrolytic capacitor
DS1—24 V lamp part of S1 (see Fig 9)
DS2, DS3, DS5—24 V panel light
DS4—24 V lamp part of S1
DS6—24 V lamp part of S2
K7, K8, K9—DPDT 5 A DIP relay, 24 V dc coil
R1—5 MΩ PC-mount potentiometer
R2, R3—1 kΩ PC-mount potentiometer
S1—Normally closed momentary-contact lighted panel switch
S2—SPST lighted panel switch

stack. All of the power resistors on the top board are mounted on one-inch ceramic standoffs. High voltage is routed to the tank compartment with test-lead wire and a high-voltage feed-through capacitor. The diode strings do not use equalizing resistors or capacitors as they are from the same batch and there is plenty of PIV headroom. Spike protection for the string is in the transformer primary, where it belongs. The high-voltage supply has three bleed-down paths for the filter capacitor: the bleeder resistors, the high-voltage meter multipliers and a 7.5 MΩ, 7.5-kV resistor at the terminals of the capacitor itself. Even with these redundant safety measures, *never* assume that they are working. *Always follow standard safety procedures when working with any high voltage, including the ac mains: It can kill you.*

Cooling

Whole-cabinet cooling is accomplished via a Dayton model 4C761 squirrel-cage blower. Cooling air is drawn into the cabinet at the right-hand side through two 2 3/4-inch holes, one on each side of the plate transformer. This removes heat from the bleeder resistors and other components before it enters the blower inlet. The Svetlana 3CX800A7 datasheet recommends airflow of 11 cfm at a back pressure of 0.2 inches (of water), for 600 W dissipation at sea level and 25°C inlet air temperature per tube. For two tubes, this is 22 cfm at 0.2 inches for 1200 W of dissipation. At a 1500-W output level with 60% efficiency, the tubes only dissipate 960 W.

The socket sub-mounting method I used resulted in a back pressure of 0.2 inches with this blower, as measured on my bench with a home-brew manometer. Since the blower is rated for 43 cfm at 0.2 inches (sea level assumed), there should be plenty of headroom for different elevations and inlet air temperatures. Remember that the above calculations are for continuous dissipation, while SSB and CW operation rarely approaches 50% of these values.

The mounting flange at the bottom of the blower was cut off so the outlet can align with the 2-inch-tall cathode chassis. The blower motor was taken off its squirrel cage and mounted to an aluminum **L** bracket. The squirrel cage was mounted to a 4×4-inch piece of 1/16-inch aluminum that supports the **L** bracket through a piece of 1/4-inch foam (Fig 10). This sound-isolation method was borrowed from Alpha Power Inc, and it works very well.

Control Board and Metering

The control PC board (Fig 11) is mounted on nylon standoffs just below the panel meters and above the row of switches and panel lights. As with most indirectly heated, oxide-coated cathodes, a warm-up period is required; for the 3CX800A7, a minimum of three minutes is recommended. When the amplifier is turned on, regulated 24 V is applied to the warm-up

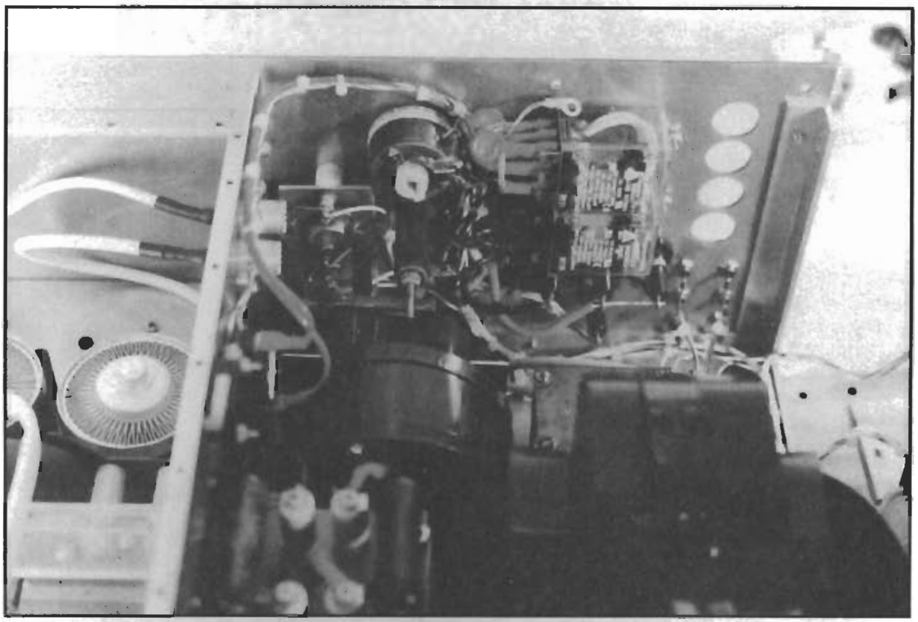

Fig 13—(right) View of panel showing K1 and K2. The bias heat sink and assembly is visible below the filament rheostat.

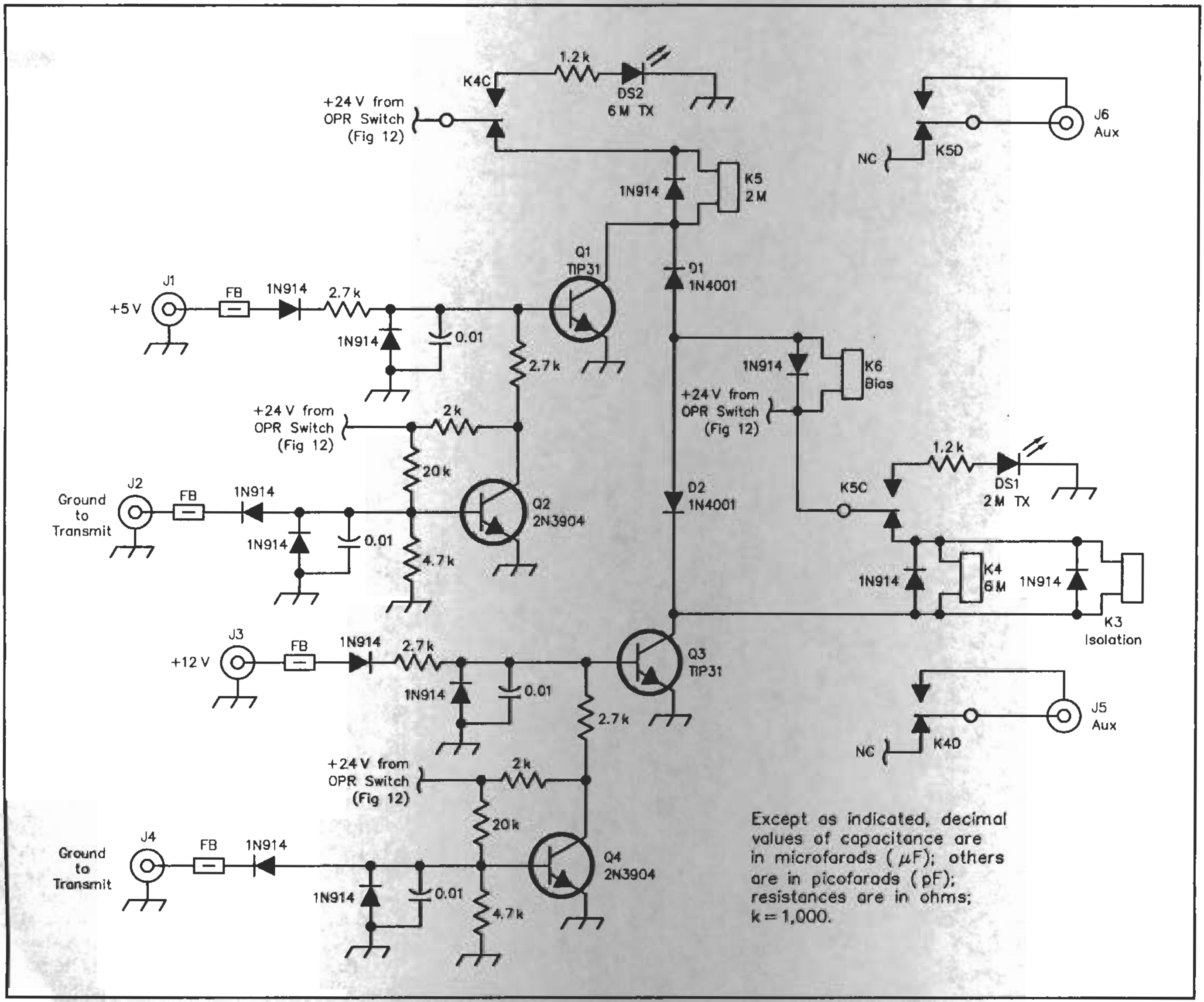

K4C
1.2 k
DS2
6 M TX
+24 V from OPR Switch (Fig 12)
1N914
K5 2M
K5D
NC
J6 Aux
Q1 TIP31
D1 1N4001
J1
+5 V
FB
1N914
2.7 k
0.01
2.7 k
1N914
K6 Bias
+24 V from OPR Switch (Fig 12)
2k
20k
J2
Ground to Transmit
Q2 2N3904
4.7k
D2 1N4001
K5C
DS1 2M TX
K4 6M
K3 Isolation
J3
+12 V
Q3 TIP31
J5 Aux
K4D
J4
Q4 2N3904

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000.

Fig 14—Schematic of the dual keying-circuit PC board. Unless otherwise specified, use 1/4 W, 5%-tolerance carbon composition or film resistors. See Table 3 for part-supplier contact information.

delay circuit consisting of Q1, R1 and C1 (Fig 12). When sufficient voltage appears at the base of Q1, relay K7 closes, with one set of contacts latching the relay on while the other set of contacts resets C1. If high-voltage is present, Q4 is switched on and K8 closes, lighting DS3 and supplying 24 V to the rest of the amp-lifier. Diode D1 was added between the field of K8 and the field of K7 to make sure K7 closes first. The proximity of the control board to the magnetic field of the plate transformer affects the timing of K7 slightly.

With the low grid-dissipation rating of these tubes, grid over-current protec-tion is necessary. Grid current is mea-sured as voltage drop across R1 of Fig 7, a 10-Ω, 25-W wire-wound resistor. R3 on the control board is set to fire K9 with Q3 at a grid current of 120 mA. One set of contacts latches the relay on; the other set interrupts the 24-V line to the **OPR** (operate) switch. Capacitor C2—across the field of K9—stops relay chatter on voice peaks. Lamp DS4 is part of the grid-reset switch S1, and it lights while K9 is energized. A second grid-current-sensing circuit, using Q2, is set to light DS7 on the front panel at 75 mA. This LED is used as a grid-current warning indicator and can be set at the grid-current level you prefer. The trip points of these two circuits and the grid-current meter reading are ad-justed while the amplifier is discon-nected from the ac supply.

To do so, connect a 24-V external supply to the 24-V bus of the amplifier after the timing circuit. Connect the positive lead of a variable-voltage dc supply to the ungrounded side of R1 of Fig 7 through an accurate dc milliammeter. Connect the external-supply negative lead to the chassis. With the grid-current panel meter disconnected, slowly increase the voltage of the bench supply until the milliam-meter reaches the set points of 120 mA for grid trip and 75 mA for the warning LED. Adjust R3 and R2 (Fig 12), respectively, to set the trip points. After the trip points are set, reconnect the grid-current meter and adjust R4 (Fig 7) for a full-scale reading of 100 mA using the same method as for the trip points. This "cold" method of setting grid trip points is a lot safer for the tubes than removing B+ and then applying drive to induce grid current.

The panel meter, M2, serves as the high-voltage meter as well as the grid-current meter. The calibration of the high-voltage portion was "roughed in" with the amplifier off and all voltages removed. A 1.5-V dc supply was connected to the multiplier side of R15 on the high-voltage rectifier board, then R15 is adjusted for a full-scale reading on M2. This corresponds to a full-scale reading of 3000 V. Meter calibration was then checked with a high-voltage probe and an accurate DMM, after the amplifier was turned on.

Plate current is measured directly in the B– line of the high-voltage supply, which is held slightly above ground by R12 on the power-supply assembly. Both meters are protected by 1N5408 diodes; two series-connected diodes on the B– bus keep it below approximately 1.4 V and two more are connected back-to-back from the multimeter to ground.

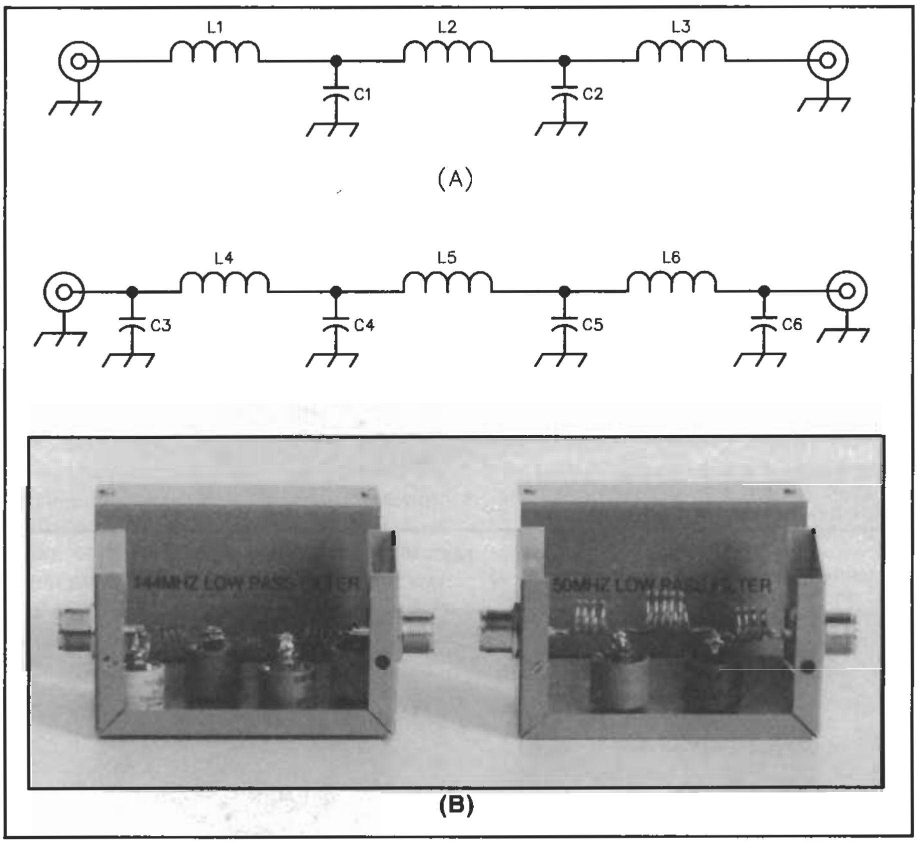

Fig 15—Details of the 2- and 6-meter low-pass filters. See Table 3 for part-supplier contact information.

50 MHz
C1, C2—50 pF doorknob capacitor
L1, L3—4 turns #14 AWG copper, $^5/_{16}$-inch ID×$^1/_2$-inch long
L2—5 turns #14 AWG copper, $^5/_{16}$-inch ID × $^5/_8$-inch long

144 MHz
C3, C6—25 pF doorknob capacitor
C4, C5—40 pF doorknob capacitor
L4-L6—3.5 turns #16 AWG silver-plated copper wire, $^1/_4$-inch ID×$^5/_{16}$-inch long

There are relative-output bar-graph displays for each band above the tune and load controls. These displays use two LM3914 bar-graph drivers per 20-segment display. A current-transformer pickup unit is used for 6 meters and a voltage-sensing pickup for 2 meters. The dual-bar-graph-display schematic can be found in Forrest Mims' book^8 and the 6- and 2-meter pickups are shown in Fig 7. The displays look appealing but turned out to be more trouble than they are worth, since I still use external wattmeters while I tune. The bar-graph displays' rapid response time would be better suited for plate- and grid-current meters.

Bias and Keying

Operating bias is developed across D1 of Fig 7, an 8.2-V, 50-W stud-mounted Zener diode. The Zener, along with K6, R2 and R3, is mounted on a 2×3-inch piece of 0.100-inch aluminum that serves as a heat sink (Fig 13). The entire assembly is mounted on two $^1/_2$-inch Teflon stand-offs on the back panel of the cabinet. The 2-A cathode fuse is mounted on the back panel alongside the ac-line fuses.

The keying PC board is mounted on the rear floor of the main cabinet (Fig 8). The PC board contains keying circuits for both 6 and 2 meters (Fig 14). Because they are identical, I'll only describe the operation of one.

At rest, Q2 is biased on, holding the base of Q1 low. Grounding J2 will take the base of Q2 low, turning it off. This allows voltage to appear at the base Q1, turning it on and energizing K5. Alternatively, applying +12 V to J1 will also turn Q1 on, keying the amplifier. Both keying inputs are logic-compatible. The bias relay is energized through D1 or D2, which comprise a two-input **OR** gate. Each transfer relay is a DowKey #260B with the "C" option, which is a pair of DPDT signaling switches. Each of the relays' supply voltages is routed through one of the opposite relay's normally closed signaling contacts. This mechanical **EXCLUSIVE-OR** gate keeps the amplifier from being simultaneously keyed on both bands. The normally open contacts are used to light transmit LEDs on the front panel; these are mounted between their respective tune and load controls to help eliminate confusion while tuning.

With this type of transfer relay, input and output relay sequencing is obviously

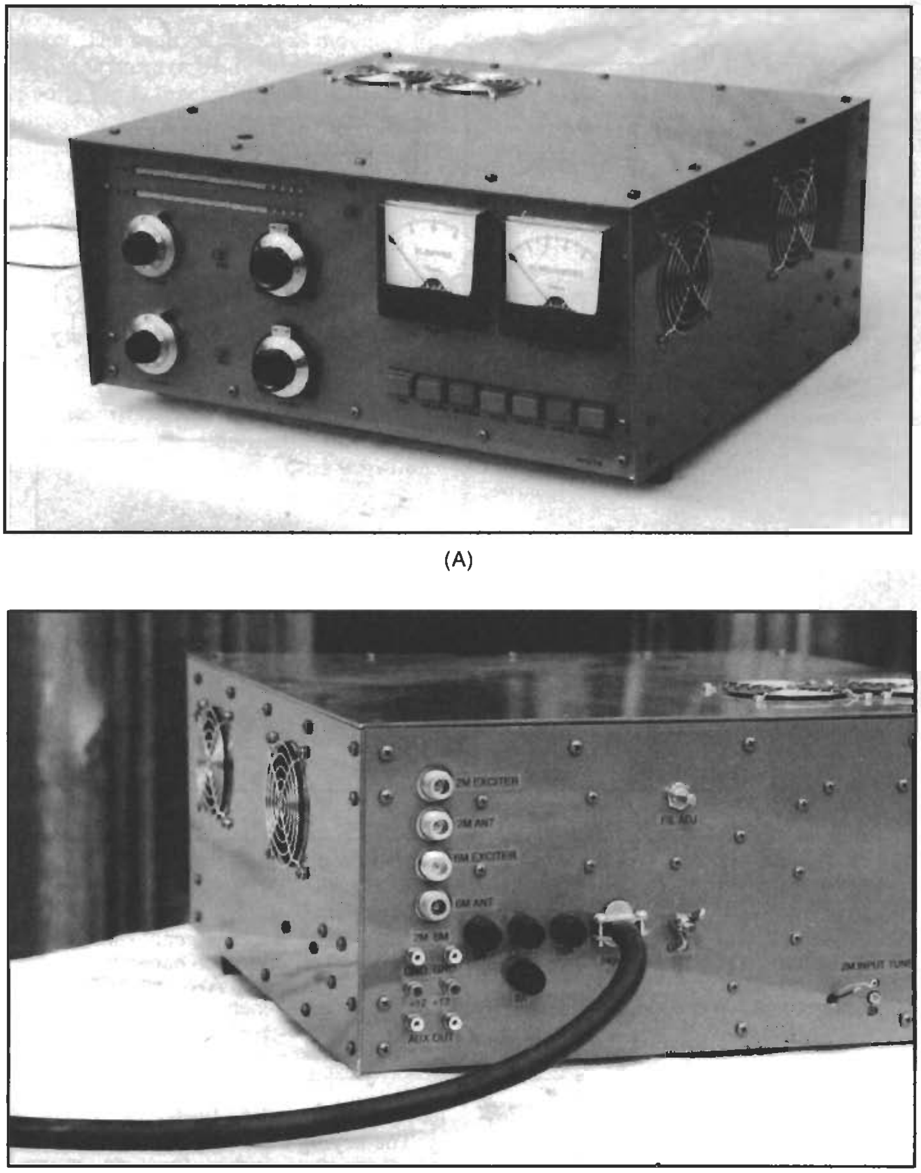

Fig 16—The completed amplifier ready for its ride to the mountain top. (A) Front view, (B) back panel.

not an option, and hot switching of the relays will result unless preven-tive measures are taken. One option is to key the amplifier and let it key the exciter. Each relay's spare set of signaling contacts was brought to the back panel for this possible use. Another option, which I employ, is to use an outboard keying sequencer. The sequencer takes care of my mast-mounted preamplifiers as well as the exciter-amplifier timing. A suitable sequencer can be found in the references.[9] When using separate exciters for each band, the spare set of signaling contacts can be used for audio muting of the unkeyed exciter while transmitting.

General Construction Notes

The entire amplifier cabinet was built using common hand and power tools. All aluminum was cut with carbide saw blades in a radial arm saw and table saw. Blades with a 5° negative hook angle seem to work best for aluminum. The main amplifier cabinet measures $18\times16\times7\frac{1}{2}$-inches (WDH). The plate transformer determined the cabinet height. The entire cabinet is built from 0.100-inch aluminum sheet and $\frac{3}{4}\times\frac{3}{4}\times\frac{1}{8}$-inch aluminum angle stock. All outside corners are secured with #8-32 pan-head screws tapped into the angle stock. The area under the plate transformer and high-voltage supply assembly was beefed up with a second layer of 0.100-inch aluminum on the cabinet floor. A square hole cut in the cabinet floor provides access the cathode compartment; it is normally covered with a piece of $\frac{1}{16}$-inch aluminum. All round holes were cut with chassis punches. The cabinet was painted with Dupont Chromaclear, which is a two-step (color coat/clear coat) automotive paint. After the base coat was applied over the primer, dry transfer lettering was applied to the front panel controls and Brothers P-touch labels were applied to the rear panel. Two coats of clear were applied over the base coat to finish the job.

All interior wiring was done with scrap Teflon-coated wire. The wiring harnesses were arranged so that all side panels can be removed and laid flat next to the chassis for ease of alignment and debugging. The intern-al coax runs were made with Teflon-dielectric coax to handle these high power levels, most other coax falls short, particularly at 144 MHz.

Adjustment and Operation

Use an ohmmeter to check the ac paths and RF deck for possible shorts and wiring mistakes. Be sure to blow out the entire cabinet with compressed air to remove any hidden debris. Set the filament-voltage rheostat to maximum resistance. If you haven't already done so, apply 24 V from an external supply to the 24-V bus and check the timing, control and keying circuitry for proper operation. Connect the amplifier to a suitable ac mains supply and connect an accurate DMM to the filament-voltage test points. Connect a 2-meter exciter and dummy load. Turn on the main power switch and quickly set the filament voltage to 13.5 V, or slightly less. *Never* operate the filaments below the 12.9-V recommended minimum. Be sure to check the filament voltage again after 5 to 10 minutes to detect any thermal drift in the rheostat. Verify that the plate is at approximately 2375 V.

Both tank circuits should be set to the tuning values found during initial setup. After the three-minute delay has elapsed, key the amplifier with no drive and check for a cathode idling current of approximately 30-50 mA. Apply a little 144-MHz drive and tune C5 in the input circuit for minimum reflected power. The 2-meter output link should be preset approximately $\frac{1}{2}$ inch above the top plane of the balanced line. Still with little drive, adjust the 2-meter tuning capacitor for maximum power output. Adjust the 2-meter load capacitor for peak output power and then leave it there until you are finished. Next, adjust link spacing as you would a load capacitor. Keep increasing the drive while increasing the link coupling for maximum power output, without exceeding grid-current limits and while also touching up the main tune capacitor, C3. Once the desired output power level is reached, check the link coupling by slowly decreasing the coupling until a slight decrease in output power is observed. Then increase the coupling slightly past the point where output power peaks and grid current is reduced to within operating limits. This should be very close to critical coupling, which will result in maximum efficiency. Go back and touch up the input network for minimum reflected power. The link is now set for operation and any further tuning adjustments can be made with C2 and C3.

Tune the 6-meter tank circuit the same way you would any other p net-work: by slowly increasing drive and tuning for maximum power output and best efficiency while touching up the input network for lowest reflected power. Then increase loading until a 2% decrease in power output is observed.

With values of loaded Q below 30 in the output tanks, there is no way to keep harmonics below the 60-dB-down figure required by the FCC without outboard filtering. Harmonic filters are a very small price to pay for the higher efficiency, lower drive require-ment and reduced thermal drift that the lower loaded Q provides. I've included descriptions of two suitable filters from the references;[10,11] they are very easy to build. The silver-mica capacitors originally in the filters have been replaced with surplus door-knob capacitors for better power-handling capability. With both filters installed all harmonics and spurious signals are more than 70 dB down from the fundamental.

Final Thoughts

This amplifier was completed just before the summer E-skip season and was immediately put on the air. On-the-air signal reports were all good, and they compared well to the 8877 amplifier I normally use. Close in IMD testing was done with the help of NN7DX who is located nine miles away, over flat ground. Results showed no excessive signal bandwidth while our antennas were positioned to bring signals down to S-9 levels during 1500-W output testing. The amplifier runs cool to the touch even after several hours.

The biggest problem encountered with the operation of the amplifier had to do with interfacing with certain exciters. Some of the new multiband exciters provide multiple antenna ports that can be configured for HF, 6- and 2-meter outputs, but provide only one buffered keying line. The ICOM-746 is built this way, but it does provide two unbuffered keying lines: one for HF and 6 meters and one for VHF. The ICOM-706-series radios also provide two unbuffered lines but only two antenna ports. Both of these radios require some sort of external switching for the keying line—and the antenna line in the case of the '706—if this amplifier and an HF amplifier are both used. ICOM does provide band logic as a variable voltage at the radio's **ACC** plug. I've designed a simple decoder/buffer that uses this logic and the radio's internal power supply to automatically switch the keying line between HF, 6- and 2-meter amplifiers. The buffer will sink up to 3 A of relay current. It also selects from three separate ALC input lines for those of you that employ sequencers using the ALC line for transmit inhibit. The decoder is described in "Automatic Amplifier Selection for the ICOM IC-746, -736 and -706MKII Transceivers," *QST*, May 2000, pp 33-36. With the decoder in place, amplifier selection with an IC-746 becomes totally automatic when changing bands between HF, 6 and 2 meters. For the IC-706 series, an additional coaxial relay can switch one of the antenna ports between HF and 6 meters to allow automatic selection. The decoder also works with the ICOM IC-736, choosing between HF and 6 meters. Although not verified, I've been told that the Yaesu FT-847 does have separate buffered keying lines for each of its four antenna ports, making interface to this amplifier easy. Performance figures for the amplifier are listed in Table 2 and the completed amplifier is shown in Fig 16. Enjoy!

Acknowledgements

Thanks to Marv Gonsier, W6FR, and Roy Scanlon, NN7DX, for help in reviewing the text.

Notes

[1] C. M. Maer Jr, W0IC, "The Perseids Powerhouse," *QST*, Oct 1959, p 32.

[2] R. M. Richardson, W4UCH, "A Kilowatt Amplifier for 6 and 2 Meters," *QST*, June 1973, p 16.

[3] The surplus sockets have a built-in 0.005 µf grid/screen bypass capacitor. RF Parts has a limited supply of these sockets. Substitutes include Johnson #124-0311-110 (with a grid collet) and Eimac SK1900.

[4] G. R. Jessop, G6JP, *VHF/UHF Manual*, 4th edition, (Hertfordshire, England: Radio Society of Great Britain, 1983) Chapter 3.

[5] E. P. Tilton, W1HDQ, *The Radio Amateur's VHF Manual*, (Newington: ARRL, 1972) p 77.

[6] R. D. Straw, Ed., *The 1999 ARRL Handbook* (Newington: ARRL, 1998), pp 6-46 and 6-47.

[7] R. Myers, W1FBY, *The 1975 ARRL Handbook* (Newington: ARRL, 1974), p 46.

[8] F. M. Mims III, *The Forrest Mims Engineer's Notebook*, (San Diego, California: Hightext Publications Inc, 1992) p 107.

[9] *The 1999 ARRL Handbook*, pp 22-53 through 22-56.

[10] I. White, G3SEK, Editor, *The VHF/UHF DX Book*, Vol 1 First Edition (Buckingham, England: DIR Publishing, 1992) p 12-35.

[11] R. Schetgen, KU7G, Ed. *The 1993 ARRL Handbook*, (Newington: ARRL, 1992) p 31-37, Fig 95.

By Bob Larkin, W7PUA **From *QST*, June 2000**

An 8-Watt, 2-Meter 'Brickette'

Put 20 mW into this little amplifier and get a 26-dB increase in transmitted power! Although the amplifier was designed initially for use with the DSP-10 transceiver, any 20-mW-output 2-meter exciter can enjoy the boost!

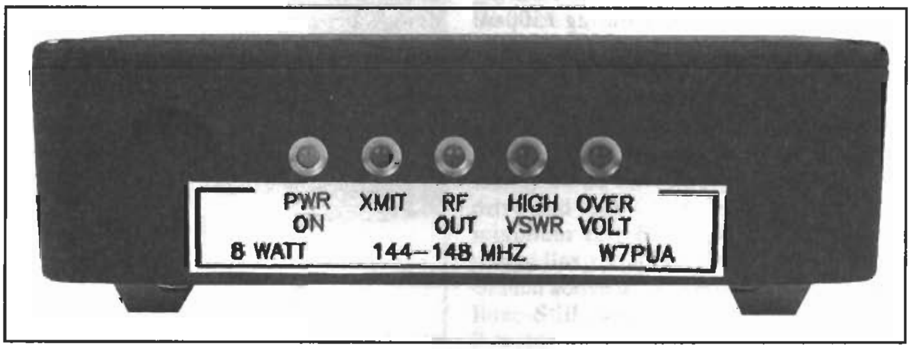

The five LEDs on the Brickette's front panel signal the amplifier's operating status. The power-output (**RF OUT**) and SWR (**HIGH VSWR**) LEDs vary in brightness depending on the forward and reflected power levels.

The DSP-10 transceiver[1] can be used with UHF/microwave transverters or as a very-QRP all-mode rig on 2-meters. The rig's 20-mW power output is sufficient for "barefoot" QSOs with locals, but that's hardly a big signal at any distance. Adding this little amplifier increases your fun quotient by raising the power level to 8 W, the high end of the QRP plateau. This brickette can also be used to drive an even higher-power, cascaded amplifier. Being a linear amplifier, it's suitable for use on all modes, including SSB, CW and FM.

The amplifier's front panel is quite simple, consisting of five monitoring LEDs. Knowing that everything is working correctly justifies the small amount of circuitry needed to operate these lights. The functions monitored include dc power on, transmit/receive status, power output, SWR and dc overvoltage. All amplifier control is done at the transceiver. Two RF cables connect to the transceiver output and receiver input, avoiding the need for an input relay in the amplifier.

This amplifier isn't limited to use with the DSP-10; any 2-meter transmitter that can deliver an output of 20 mW to drive the amplifier should work fine. In some cases, it might be desirable to use an input TR relay; there is sufficient room in the amplifier for adding one.

Circuitry

At the heart of this amplifier is an integrated power-amplifier module, manufactured by Mitsubishi. Such modules are used in many commercially manufactured transceivers and using one here makes amplifier construction and alignment simple. Within the module are two cascaded linear-amplifier transistors, along with their associated matching networks and biasing circuits for class AB operation. The module has 50-Ω input and output impedances. All we need to do is add a circuit to turn on the bias supply during transmit, install a low-pass filter for harmonic control and include an antenna-switching relay. To monitor amplifier operation, we tack on some simple circuitry.

Refer to the schematic in Figure 1. At the input pin of the RF module, U1, we apply 20 mW of drive. Blocking capacitors for the RF input and output ports are included within the module. Bypass capacitors on the three power leads are external, however. Ferrite beads, L6 to L8, prevent problems that might occur if RF gets on the power leads.

An L network consisting of L1 and C4 improves the impedance match to the module. Adding the network increases the output power by about 0.25 dB. Following the amplifier is a directional coupler (discussed later). Next is a five-pole low-pass filter. For simplicity, it is configured with the same coil and capacitor types as are used in the directional coupler. This filter attenuates the second and higher-order harmonics.

Separate connectors are available on the DSP-10 for the transmitter output and the receiver input. This simplifies adding the antenna relay for the amplifier since no switching is required at the amplifier input. A miniature relay is adequate at the amplifier output. Providing an isolation of about 30 dB, the relay, along with the PIN attenuator that is part of the DSP-10, provides plenty of protection for the receiver.

A lumped-element directional coupler, consisting of L2, L3 and C5 to C10, delivers power samples of the forward and reflected output signals. This directional coupler works quite well, providing a coupling of about −28 dB and a directivity[2] of 20 dB, but only over a narrow (12-MHz) bandwidth. For our application, this is adequate. Two diode detectors, built around D6 and D7, generate low-level dc signals that indicate the forward and reflected powers. These signals, in turn, are amplified by two sections of op amp U2. The resulting voltages are displayed on two LEDs, green for **RF OUT** (forward) power and red for **HIGH VSWR** (reflected) power. The LEDs serve as rough indicators of proper amplifier operation.

The **OVERVOLTAGE** indicator lights when the dc supply voltage exceeds 14.5 V, but does not automatically shut down the amplifier. The idea here is to supply a warning mechanism. When RF is not applied to the amplifier, it is quite resistant to supply overvoltage. As long as the supply overvoltage condition sets an alarm light, an observant operator will defer applying RF until the voltage is reduced.

Amplifier control is handled by the DSP-10. That transceiver has software-controlled relay sequencing, providing +5 V during transmit. Q2 and Q3 drive the antenna relay from this control. A ferrite bead, L11, and C24 keep RF at the antenna relay from getting back into the control circuits. D10 shunts the inductive kick from the deactivated relay coil. An LED, D2 (**XMIT**), across the relay displays the amplifier TR status.

Q4 provides an output that can control a follow-on amplifier. This output is an open-collector, ground-on-transmit type, compatible with many commercial and homebuilt amplifiers. For added driver-circuit protection, a reverse-voltage shunting diode, D11, is used.

Building the Amplifier

You could build the amplifier on a scrap of PC board. In fact, my first version is built that way (see the sidebar). For many builders, however, a PC board[3] is a more convenient way to assemble the project and that is shown in the photos.

This PC board is double-sided, with plated through holes. The backside is a solid ground plane allowing the board to be fastened directly to the aluminum enclosure. Such construction works very well for RF boards because a low-inductance ground path can be maintained throughout, reducing any interactions between the various circuits. However, this mounting method does not allow component leads to extend beyond the bottom surface of the board. I dealt with this primarily by using surface-mount parts[4] and by carefully bending the component leads so they behave like surface-mount parts.

All of the chip resistors, chip capacitors, diodes, surface-mount ICs and ferrite-bead chips are soldered to their board pads conventionally. Install the amplifier module *after* the board is mounted in the box. The antenna relay, fuse clips and the variable resistor all have their leads bent away from their host. Avoid making any bends close to the component body; make the bends at a point about 1/16 inch away from the component body. This approach eliminates mechanical stress on the lead attachment. These components end up about 1/16 inch above the board's top surface after they are soldered in place.

Figure 2 shows the construction of the four toroidal inductors, L2 through L5. Experience shows the inductor values are quite repeatable if the turns are always distributed in the same manner around the core. As the turns are pushed closer together, the inductance increases considerably. The matching-network coil, L1, is noncritical and it should be wound as shown in the part list and Figure 2; it should need little, if any, adjustment.

The LEDs all have long leads and are soldered to the board after bending the lead ends by about 1/8 inch. Be sure to keep track of the LEDs' longer (anode) leads: Those leads connect to the current-limiting resistors.

Putting the Board in a Box

Once the PC board is assembled, use it as a template for marking the hole locations in the enclosure, a standard Hammond die-cast box. Mount the board flat against the box bottom using #4-40 hardware. Because there will likely be some mold marks and box-identification letters where the board and RF module lie, make the enclosure's inner-bottom surface reasonably flat; you can do this with 60-grit sandpaper. Bend the leads of the five LEDs to apply a slight forward pressure on the lights as they slide into the holes in the box front. That holds the LEDs in alignment without needing adhesive.

Take care when tightening the PA module mounting screws to be sure that no pressure is applied to the ends of the module cover. The leads from the module need some trimming. They are above the board and need forming to get to the board level for soldering. Don't apply pressure at the edge of the module cover when doing this. You may need to hold each lead with needle-nose pliers to keep from damaging the case. It should be possible for the leads to reach the board surface within 1/8-inch of the cover.

Three short pieces of 50-Ω coax attach the board to J1, J2 and J3. Solder lugs under the jacks provide for ground connections at one end and PC-board pads take care of the other end of the coax.

Three leads run between connectors J4, J5 and J6 and the PC board. Each of these leads has a bypass capacitor (C28, C6 and C27, respectively) at the connector. Short, low-inductance leads are important on these capacitors. To help keep the RF signals inside the box, a small inductor wound on a ferrite core (L14, L15 and L16) is placed on each wire. Position these inductors close to their connectors. If #22 or #24 stranded hook-up wire with thin insulation is used, it's possible to wind the coils with the hook-up wire.

Turning on the Amplifier

Now, to see it work! First, connect a 50-Ω noninductive dummy load to the amplifier output.[5] If you have a variable-voltage power supply to use for initial testing, slowly raise the dc supply voltage from 0 V to 13.8 V. Otherwise, you'll need to rely on the power-supply fuse as protection from any serious construction errors when apply-

Did the Amplifier Always Look Finished?

No, it didn't. Some people may be able to put a finished amplifier on a PC board and have it work fine. I don't seem to be able to do that!

My first step in designing the amplifier was to search for suitable RF modules, mostly via the Internet. Once I selected the M57732L as having suitable power, gain and dc operating voltage, I drew a schematic in my notebook. Originally, the circuit had automatic shutdown for high SWR and full voltage regulation to deal with overvoltage. This looked too complex. It was necessary to cut off the drive to the amplifier, and the regulator had to be a low-dropout type (for these, the difference in voltage between the input and the output need be only a fraction of a volt). I couldn't find an integrated regulator that met my requirements and the thought of building one using op amps and transistors used more parts than I could justify.

I opted for simplification, putting alarm lights on for SWR and overvoltage and turned the automatic part over to the operator! A new schematic resulted, not unlike that of Figure 1. I went to the *ARRL Radio Designer*† and simulated the directional coupler and low-pass filter. A little playing around with the simulation showed that a single inductor value could be used for both subcircuits.

Next, I built the "breadboard" version using scraps of PC board. Testing showed that almost everything worked as expected. The **overvoltage** light did not have snap action, though. That was caused by using a resistor where the ferrite bead, L13, is now. I ran the input voltage temporarily up to 16 V and everything continued to function. Next, I left the amplifier running for an hour and nothing overheated. I checked the intermodulation products and harmonics and found them to be at satisfactory levels.

It was now time to layout the PC board. The experience of putting the breadboard together allowed a smarter final layout. The first version was on two boards and some of the connectors ended up on the front panel. This was all worked out for the final design.
—Bob Larkin, W7PUA

†ARRL order no. 6796. ARRL publications are available from your local dealer, or directly from the ARRL. See the ARRL Bookcase elsewhere in this issue, or check out the ARRL Web site at: **http://www.arrl.org/catalog/**.

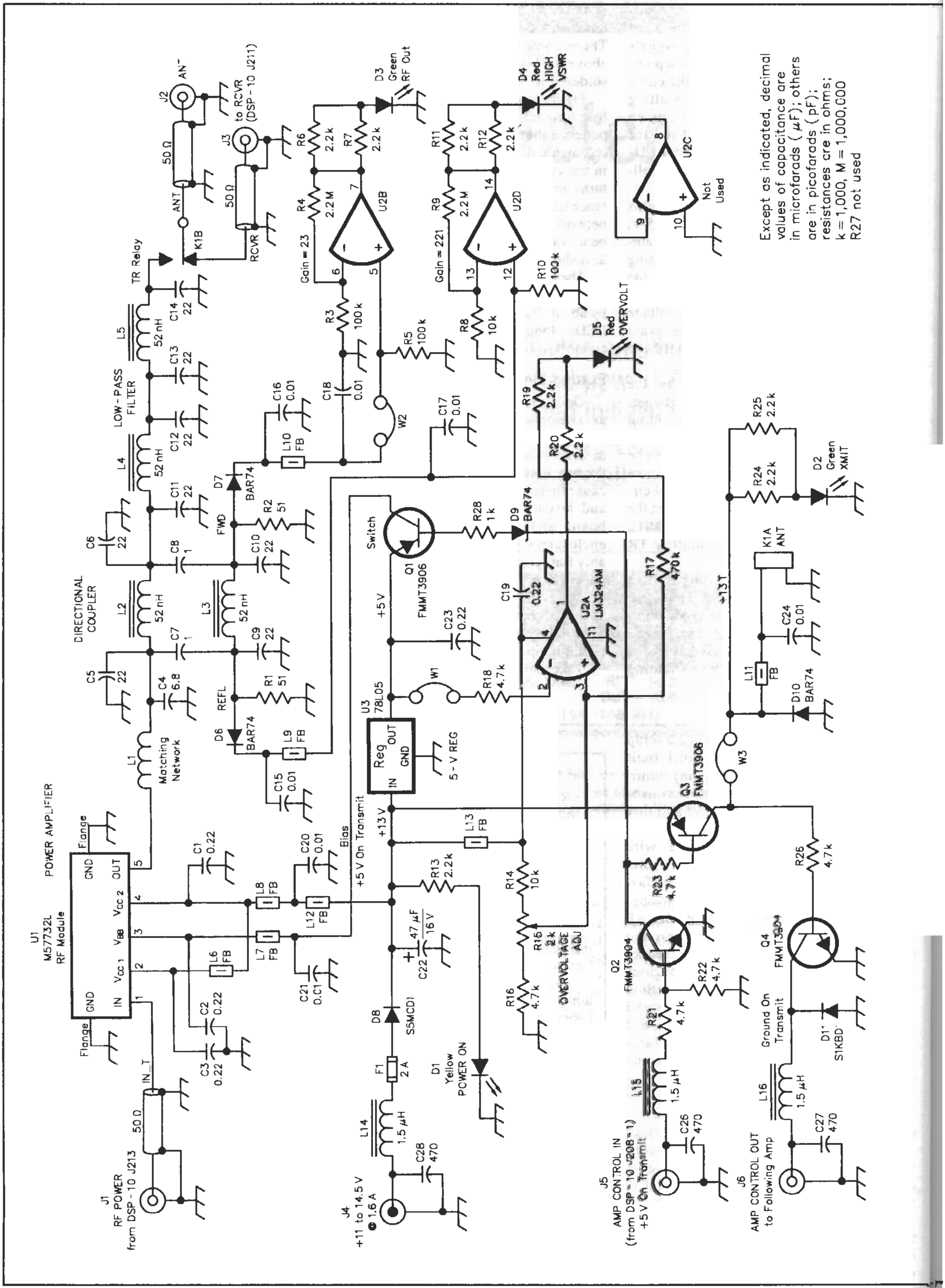
J1
RF POWER
from DSP-10 J213
50 Ω
IN_T
U1
M57732L
RF Module
POWER AMPLIFIER
GND
IN
Vcc 1
VBB
Vcc 2
OUT
GND
Flange
C1 0.22
C2 0.22
C3 0.22
L6 FB
L7 FB
L8 FB
L12 FB
C20 0.01
C21 0.01
Bias
+5 V On Transmit
L1
Matching Network
C4 6.8
C5 22
C6 22
C7 1
C8 1
L2 52 nH
L3 52 nH
DIRECTIONAL COUPLER
C9 22
C10 22
R1 51
R2 51
REFL
FWD
D6 BAR74
D7 BAR74
L9 FB
L10 FB
C15 0.01
C16 0.01
L4 52 nH
L5 52 nH
C11 22
C12 22
C13 22
C14 22
LOW-PASS FILTER
TR Relay
K1B
ANT
RCVR
50 Ω
J2
ANT
J3
to RCVR
(DSP-10 J211)
C18 0.01
R3 100 k
R5 100 k
W2
R4 2.2 M
R6 2.2 k
R7 2.2 k
Gain = 23
U2B
D3
Green
RF Out
R8 10 k
R9 2.2 M
R10 100 k
R11 2.2 k
R12 2.2 k
Gain = 221
U2D
D4
Red
HIGH
VSWR
U2C
Not Used
D5
Red
OVERVOLT
R19 2.2 k
R20 2.2 k
C17 0.01
Switch
Q1
FMMT3906
R28 1 k
D9 BAR74
+5 V
C23 0.22
C19 0.22
U2A
LM324AM
R17 470 k
U3
78L05
Reg
IN
OUT
GND
5-V REG
W1
R18 4.7 k
+13 V
L13 FB
R13 2.2 k
R14 10 k
R15 2k
OVERVOLTAGE ADJ
R16 4.7 k
C22 47 µF 16 V
D8 S5MCD1
F1 2 A
D1
Yellow
POWER ON
L14 1.5 µH
C28 470
J4
+11 to 14.5 V
@ 1.6 A
Q2
FMMT3904
R21 4.7 k
R22 4.7 k
R23 4.7 k
L15 1.5 µH
C26 470
J5
AMP CONTROL IN
(from DSP-10 J208-1)
+5 V On Transmit
Q3
FMMT3906
W3
D10 BAR74
L11 FB
C24 0.01
+13T
K1A
ANT
R24 2.2 k
R25 2.2 k
D2
Green
XMIT
R26 4.7 k
Q4
FMMT3904
Ground On Transmit
D11 S1KBD
L16 1.5 µH
C27 470
J6
AMP CONTROL OUT
to Following Amp
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k = 1,000, M = 1,000,000
R27 not used

Figure 1—Schematic of the 8-W amplifier. All resistors are 5% 1206 chips. These are available in small quantities from Mouser Electronics (Xicon). Unless otherwise noted all capacitors are either 1206 or 0805 chips (either size fits on the board). Capacitors less than 470 pF are NP0, while values from 470 pF and up are any general-purpose ceramic, such as X7R or Z5U. Component sources listed here are generally only a few of several that manufacture equivalent parts. If the amplifier is built without a pre-made PC board, change the chip components to leaded types. Component designations for the LEDs differ from *QST* style. Parts used in the amplifier are available from one or more of the following sources. Source abbreviations used in the parts list precede the company name: (DK) Digi-Key Corp, 701 Brooks Ave S, Thief River Falls, MN 56701; tel 800-344-4539, 218-681-6674; **http://www.digikey.com**; (ME) Mouser Electronics, 958 N Main, Mansfield, TX 76063; tel 800-346-6873, 817-483-4422; **http://www.mouser.com**; (RS) RadioShack—see your local distributor; **http://www.radioshack.com**; (RP) RF Parts, 435 South Pacific St, San Marcos, CA 92069; tel 800-737-2787, 760-744-0700; **http://www.rfparts.com**.

C22—47 µF, 16 V surface-mount electrolytic (DK PCE3033CT)
C26, C27, C28—470 pF, 50 V, leaded ceramic (ME 140-50P2-471K)
D1—Yellow T1 LED (DK160-1079)
D2, D3—Green T1 LED (DK 160-1080)
D4, D5—Red T1 LED (DK 160-1078)
D6, D7, D9, D10—BAR74 diode (DK BAR74ZXCT)
D8—5-A, 50-V power rectifier, SMC package (DK S5ACDICT)
D11—1-A, 100-V power rectifier, SMB package (DK S1BBDICT)
K1—SPDT 12-V dc miniature relay, Omron G5V-1-DC12 (DK Z774)
L1—2t, #20 or 22, 1/8-inch ID; see Figure 3.
L2-L5—52 nH; 5 turns, #26 enameled wire on a T-25-17 toroid; see Figure 3.
L6, L7, L9, L10, L11, L13—Ferrite SMT bead, 1206, 600 Ω at 100 MHz, Stewart HZ1206B601R (DK 240-1019-1)
L8, L12—Ferrite SMT bead, 3 A, 1206, 100 Ω at 100 MHz, Stewart HI1206N101R (DK 240-1008-1)
L14-L16—1.5 µH, 5 turns #22 or #24 enameled wire on an FT-23-43 core; see text.
F1—2-A, 5×20-mm fuse (DK F948) with two clips (DK F058)
J1, J5, J6—Phono jacks (RS 274-346)
J2, J3—BNC jack (RS 278-105)
J4—5.5-mm OD, 2.1-mm ID power connector (RS 274-1563)
Q1, Q3—FMMT3906 PNP transistor, SOT23 (DK FMMT3906CT)
Q2, Q4—FMMT3904 NPN transistor, SOT23 (DK FMMT3904CT)
R15—2 kΩ adjustable resistor, Bourns 3329H-1-202 (DK3329H-202)
U1—RF amplifier module, Mitsubishi M57732L, (RP M57732L)
U2—LM324AM dual op-amp (DK LM324AM)
U3—78L05 5-V positive regulator, SO-8 package (DK LM78L05ACM)
Misc: PC board, enclosure 3.7×4.7 ×1.3 inches, Hammond 1590BB (DK HM152), hardware.

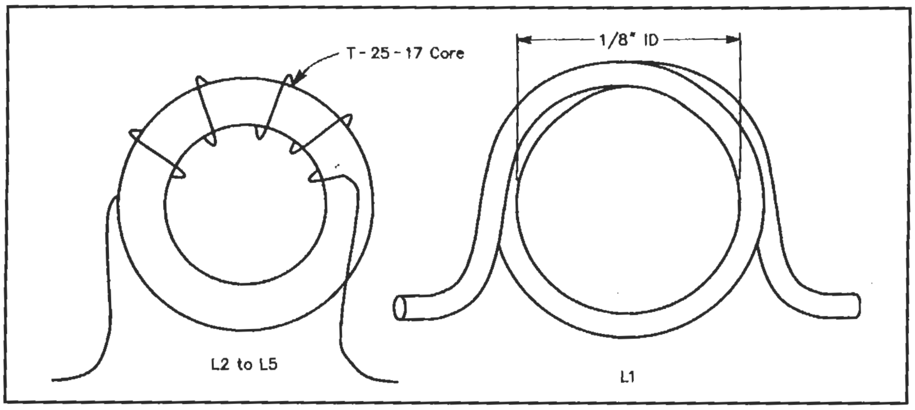

Figure 2—Coil details. Arrange the turns on the four toroids as shown here. L2 and L3 might need to have their turns spacing adjusted during tune-up.

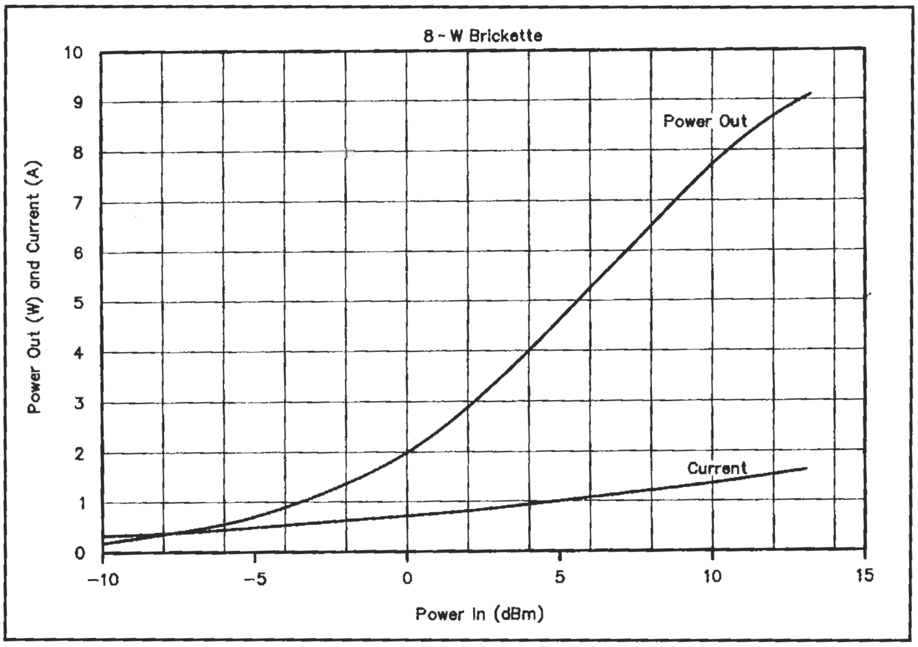

Figure 3—Measured power output and dc-supply current as a function of drive power. Notice that the output-power scale is not in decibels. This makes the output appear less compressed than it really is. The 1-dB compression point is at about 4 W.

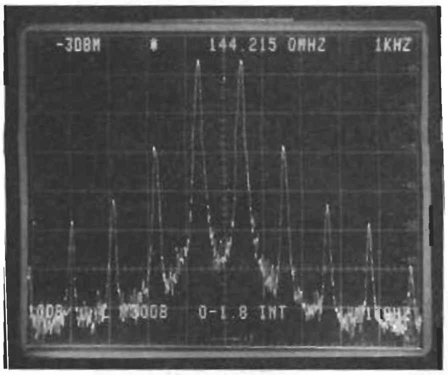

Figure 4—Measured intermodulation distortion with a PEP output of 8 W. Third-order products are down about 28 dB from peak power, fairly typical of this amplifier type. Perhaps more important is that the fifth and higher-order products continue to drop off as the order increases. These higher-order products are farther from the operating frequency and thus generally more disruptive to nearby stations. The current drawn during this two-tone test was 670 mA.

ing the full 13.8 V. At full voltage, the current drawn by the amplifier during receive should be about 10 mA. Next, set the **OVERVOLTAGE** alarm. Adjust the voltage at the wiper terminal of R15 to 4.68 V.[6] If you can vary the dc supply voltage, the **OVERVOLTAGE** LED should light at about 14.5 V.

When making the following adjustments, it is necessary to adjust the turns spacing on the toroidal inductors. For safety, *do this only when no RF drive is applied.*

Connect the TR control to the DSP-10 and connect J1, the amplifier's RF input, to the DSP-10 antenna connector. In CW mode, increase the DSP-10's RF power slowly until the **RF OUT** LED (D3) glows. Then adjust the turns spacing on L2 and L3 to extinguish the **HIGH VSWR** light (D4). This indicator is about 10 times more sensitive than the **RF OUT** LED, but it should be possible to extinguish the **HIGH VSWR** LED completely. Continue this adjustment process while increasing the drive power. You can measure the voltage at the output of the SWR op amp

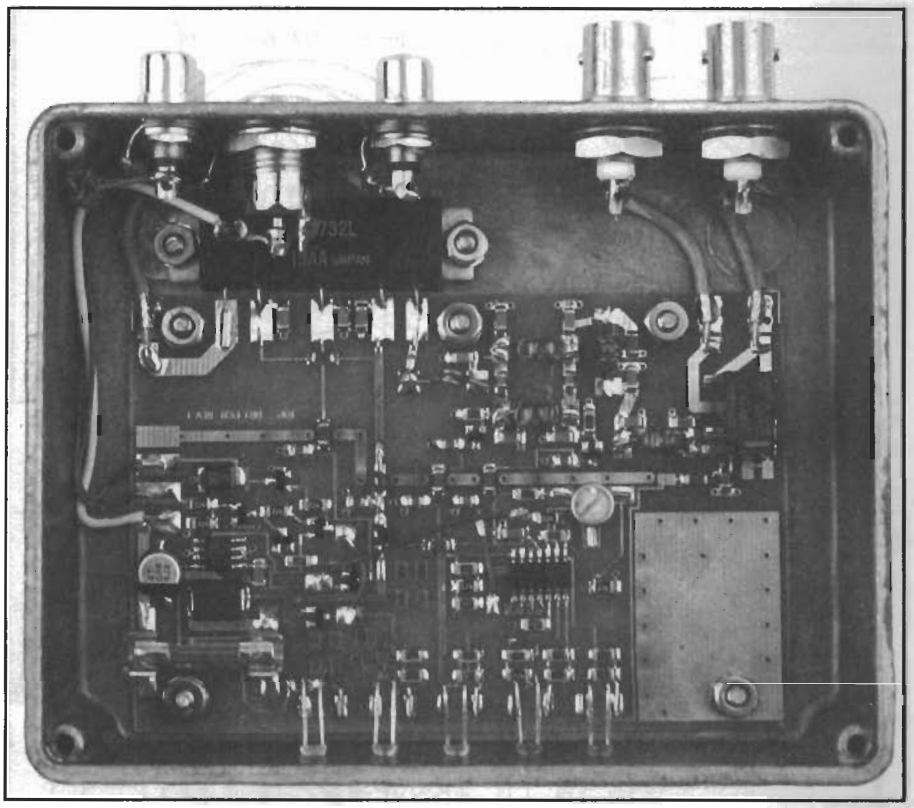

Pretty outside and pretty inside, the Brickette's construction reflects a caring hand. At the top-left rear the RF-amplifier module can be seen fastened to the base of the enclosure, which serves as a heat sink for the module. The relay is at the upper-right and a kludge area is visible at the lower right behind the front panel.

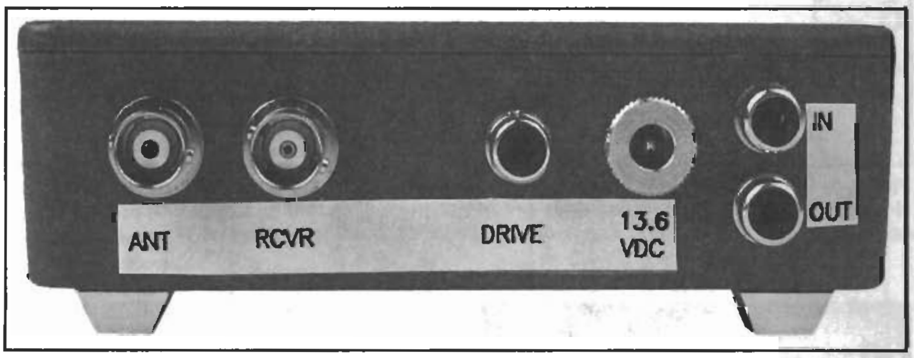

This rear-panel view of the 8-W Brickette shows compact but uncrowded I/O connectors clearly labeled.

(U2 pin 14) and continue to minimize this voltage, although the **VSWR** LED is not lit.

Optimize the amplifier matching by measuring the power output as indicated by the dc voltage at U2 pin 7 and adjusting L1's turn spacing (and perhaps changing the value of C4) for maximum output at high drive. Don't expect major changes with these adjustments as the amplifier is inherently quite well matched.

While in transmit, check that the bias voltage at pin 3 of U1 is about 4.9 V and that the idling current with no RF drive applied is about 140 mA.

Performance

Figure 3 shows the power output and current level for various drive levels. The CW output is over 8 W for the two devices that I tested. When tested with a two-tone input signal and an output of 8 W PEP, the IMD level is down 28 dB for third-order products and 43 dB for fifth-order products (see Figure 4). The strongest harmonic is the second (at about 292 MHz) and this is 65 dB down from peak output, more than enough to meet FCC 2002 requirements.[7]

Concluding Thoughts

Several areas on the PC board have either a ground plane section or pads for mounting transistors, resistors and other components. These areas are available to you to use for modifications or additions to the amplifier (perhaps you have an application that needs an input TR relay or you need a different control circuit).

Now you're ready to get on the air with a medium-sized signal. You can increase power even further by cascading another amplifier. Or, use some of the signal processing in the DSP-10 to work deeper into the noise. QRP power levels have pushed the idea of "working smarter instead of harder."

Notes

[1] Bob Larkin, W7PUA, "The DSP-10: An All-Mode 2-Meter Transceiver Using a DSP IF and PC-Controlled Front Panel," *QST*, —*Part 1*, Sept 1999, pp 33-41, —*Part 2*, Oct 1999, 34-40, —*Part 3*, Nov 1999, pp 42-45. Additional information on that project is available on the author's Web site, **http://www.proaxis.com/~boblark/dsp10.htm**. Any future information about this amplifier project will be placed at this Web site.

[2] The directivity of a coupler is the difference, in decibels, between the forward signal and the reflected signal, when the coupler is properly terminated in its designed load, in this case, 50 Ω. Couplers constructed from transmission lines have high directivity over a very wide frequency range. These lumped-element couplers are restricted in the frequency range where their isolation is high, but offer simplicity instead.

[3] Gerber files for making the PC board can be obtained from the author. Alternatively, unpopulated PC boards are available from Mashell Electric, PO Box 5, Eatonville, WA 98328. Price: $20 each in the US, $21 in Canada (air mail) and $22.50 (air mail) elsewhere. These boards have plated through holes, a solder mask on the component side and a silk-screened legend. Check their Web site **http://members.aol.com/w7slb/w7slb.htm** for details.

[4] See Sam Ulbing, N4UAU, "Surface Mount Technology—You Can Work with It!", *QST—Part 1*, Apr 1999, pp 33-39; —*Part 2*, May 1999, pp 48-50; —*Part 3*, Jun 1999, pp 34-36; —*Part 4*, pp 38-41. Additional comments from Avery Davis, WB4RTP, can be found in the *QST* Technical Correspondence, Feb 2000, p 70. Surface-mount techniques are used for the DSP-10 transceiver. Some people were initially concerned about dealing with the tiny components, but after constructing the board, most builders felt that it went well.

[5] A satisfactory dummy load for 144 MHz can be constructed from four 200-Ω, 3-W metal-oxide resistors. Arrange them in parallel around a coax connector. Short leads are important. *Do not use wire-wound resistors.*

[6] The **OVERVOLTAGE** alarm triggers when the dc supply voltage is greater than 14.5 V. If an adjustable dc supply is available, you can set this directly. Alternatively, set the voltage at the wiper of R15 to 0.360×(V – 0.6), where V is the dc supply voltage.

[7] Larry Price, W4RA, and Paul Rinaldo, W4RI, "WARC97, An Amateur Radio Perspective," *QST*, Feb 1998, pp 31-34. By these rules, an 8-W 2-meter transmitter must suppress the harmonics by at least 52 dB. Greater amounts of suppression are required for higher power levels.

By Dave Mascaro, WA3JUF **From *QST*, June 1990**

903-MHz Linear Amplifiers

Part 1—Looking for a gain block or two for your 903-MHz station? Here are *eleven* of them to cover just about any need you may have, from a simple receiving preamplifier to one with 23 dB gain and over 4 W output!

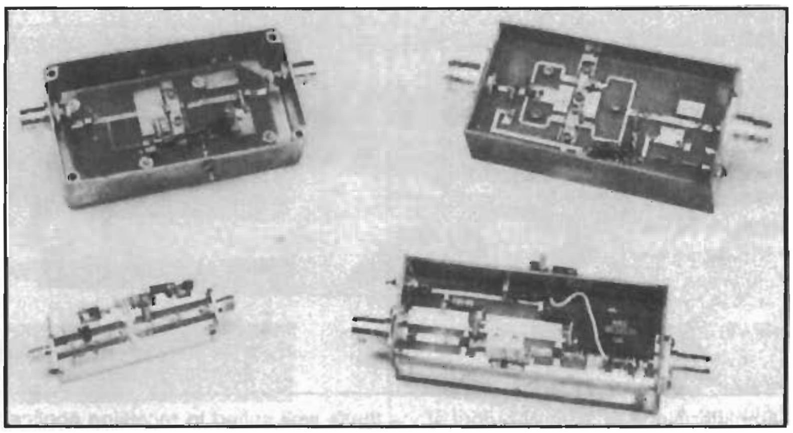

The 33-cm Amateur Radio band (902-928 MHz) is becoming well populated in many areas of the country. Propagation on 33 cm has traits similar to both 432 MHz and 1296 MHz, but has characteristics all its own at times, making it a very interesting band. It's pleasantly surprising what you can do with a few watts on 903 MHz with high-gain loop-Yagi antennas.

Articles have been published on transverters, amplifiers and receivers for 33 cm. I described a 759-MHz local oscillator[1] intended to be used with a 144-MHz IF for operation on 903 MHz. If you've built the transmit section of a transverter that gets you to the 10-mW level, or if you've purchased a commercial transverter such as Down East Microwave's low-power, no-tune unit,[2] you'll need to bring your transmitter power up to a usable level. Instead of showing a typical transmit-amplifier chain, many of which have been described in Amateur Radio literature, I will discuss eleven different gain blocks, for several different gains and power levels, in this two-part series.

I suggest that you build each amplifier in its own enclosure, rather than trying to build more than one amplifier in one box to eliminate connectors or reduce size. If you are building your own equipment, you probably aren't too concerned about size and compactness, and having separate enclosures makes for much easier tuning and troubleshooting. Many transistors used in 1296-MHz projects are suitable for use at 903 MHz; some even have more gain at 903.

Transistors with SD prefixes used in the following designs are manufactured by SGS-Thomson Microelectronics.[3] Some of these transistors don't operate at 12 V. Again, in building your own equipment, this shouldn't be a major problem. Gains listed are averages for several different devices tried. All of the amplifiers can be driven harder than indicated—for CW or FM operation only—but should be used at or below their rated power output for SSB work. All the designs can be used at lower power output; they are all linear amplifiers.

Construction

All of the amplifiers are built using similar techniques. Each design uses common $^1/_{16}$-inch-thick, G-10, double-sided, fiberglass-epoxy PC board for the microstrip circuitry. The ground-plane side of each PC board is unetched.

The microstrip boards can be made using what I call the "X-ACTO®-etch" method. This involves using a piece of clear tape as the resist. (Four-inch-wide clear tape is available at stationery stores.) After drawing the pattern on the board with a pencil or fine-tipped marker, cover it on both sides with clear tape and cut the pattern in the tape with an X-ACTO knife. Then, remove the tape from the areas to be etched and etch the boards in ferric chloride in a crock pot on low heat. Do this in a well-ventilated area! Etching takes about half an hour without agitation. The crock pot is a no-mess way to etch boards. Only an inch or so of ferric-chloride solution is needed. Of course, you can also use a photographic method to make boards, as I do for multiple or complicated boards.

After the PC board is etched, clean it with steel wool and drill holes at all dc- and RF-ground points using a no. 50 drill. No holes are needed for component leads, as all components are mounted on the microstrip side of each PC board. RF grounds must be located as close to the areas to be grounded as possible, to ensure low-inductance ground paths. There are at least two ways to do this. One is to install a rivet in each hole, flare it with an awl or ice pick, then flatten it by tapping with a small

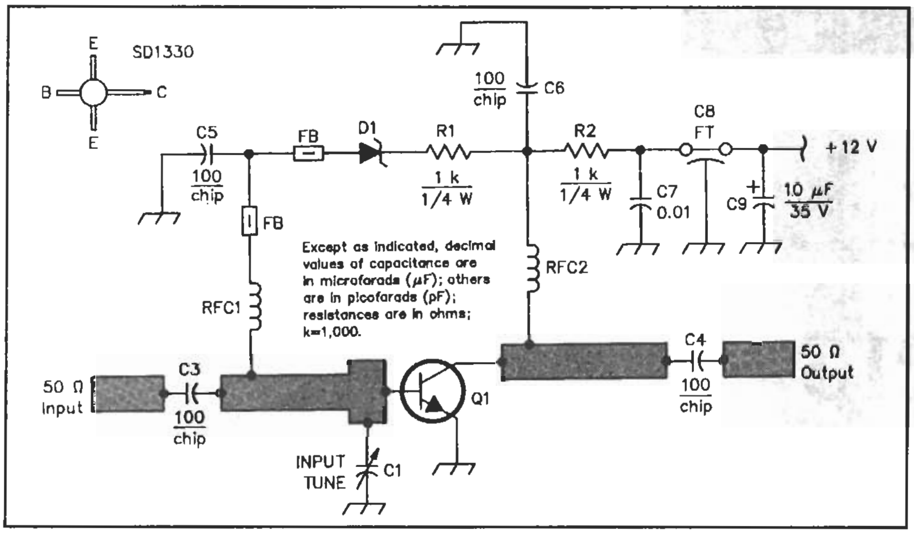
E
SD1330
B
C
E
C5
100
chip
FB
D1
R1
1 k
1/4 W
100
chip
C6
R2
1 k
1/4 W
C8
FT
+12 V
C7
0.01
C9
10 µF
35 V
FB
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.
RFC2
RFC1
C4
50 Ω
Output
100
chip
C3
50 Ω
Input
100
chip
Q1
INPUT
TUNE
C1

Fig 1—Schematic of the SD1330 receive preamplifier/low-level transmit stage. RFC1 and RFC2 consist of 8 turns of no. 24 enam wire closewound on a 0.1-inch-ID form (such as a 0.1-inch drill). D1 is a 9.1-V,½-W Zener. C1, a 12.5-pF trimmer, is available from Mouser (part no. 24AA071).

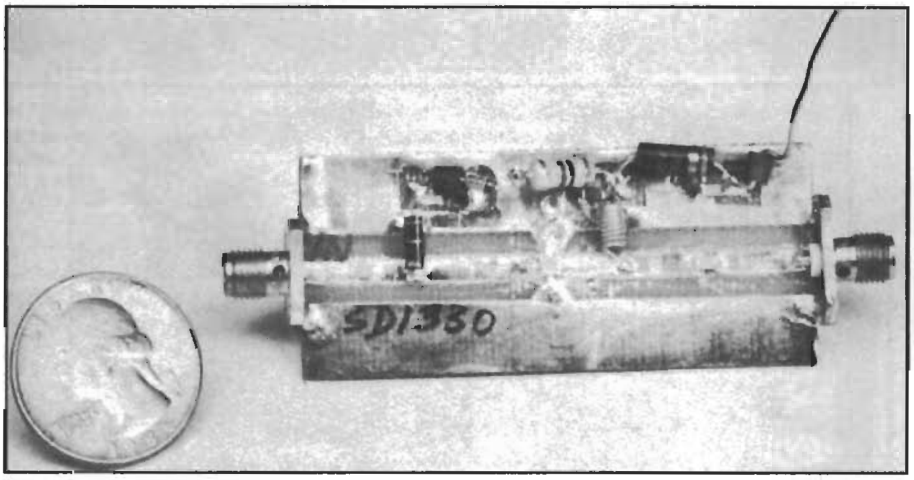

Fig 2—Amplifier No. 1, a 10-mW-output stage that's well suited to receiving applications, in prototype form.

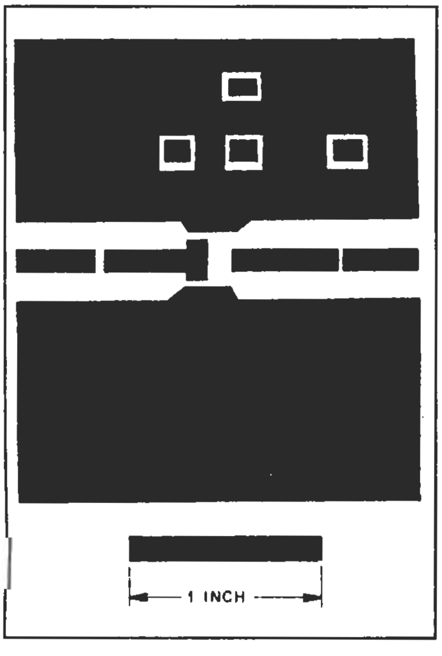

Fig 3—Full-scale PC-board artwork for the 10-mW SD1330 amplifier. Black areas represent unetched copperfoil. All parts are mounted on the foil side of the board. Crop the finished boards as necessary.

hammer on an anvil or other hard surface. Then, solder the rivets on both sides of the PC board. Alternatively, pieces of bus wire can be used for grounding.

Because parts placement is quite straightforward, I haven't provided parts-placement drawings for all of the designs. The schematics and PC-board artwork should provide all the guidance you'll need to assemble the boards. I can help with any questions that you might have when building any of the amplifiers.

All of the amplifiers described, except those operating under the 300-mW level, require some heat sinking. A small brass sheet is sufficient for the studded parts, but a finned heat sink an inch or two square is needed for the flanged-device amplifiers. Thermally conductive compound is required between the devices and their heat sinks. For studless devices, a small piece of brass shim stock soldered from the bottom of the device to the ground-plane side of the PC board should work fine.

An easy way to enclose an amplifier like those described is to make a housing using double-sided PC board or brass sheet for the four walls and bottom cover. The amplifier PC board makes the top cover. Input and output connectors (N, BNC or SMA) can be bolted or soldered to the end walls with their center pins soldered directly to the microstrip input/output lines in an end-launch configuration. When using PC-board material for housings, mount the side walls in such a way that the grounding is continuous from the connectors to the ground-plane side of the amplifier board.

Various die-cast aluminum boxes, such as Bud CU-123 and CU-124 and the Hammond 1590 series, also work great for enclosing these designs. Die-cast boxes usually provide sufficient heat sinking for amplifiers operating at less than 2 W output. Device studs or flanges should be attached directly to the boxes in these cases. All component leads must be kept as short as possible. This also applies to trimmer caps. Mount them flush to the PC board so that they act as capacitors—not as inductor/capacitor combinations. Use multiple rivets (or bus wires) to ground the trimmers.

Components

Use ceramic chip capacitors for the dc blocks and high-frequency bypassing. Some of the larger-value bypass capacitors, like 0.001 µF and 0.01 µF, are available in chip form also. These epoxy types are good for all the designs described here. I use inexpensive chip caps from Mouser Electronics.[4] Chip caps and Johanson piston trimmers are available from Microwave Components of Michigan.[5] Transistors are available from RF Parts[6] and RF Gain, Ltd.[7] Components, rivets and PC boards are available from Frontier Microwave.[8]

Tune-Up

In most of these amplifiers, the quiescent collector current I_{cq}) initially will have to be checked and set with no drive applied.

To check bias current, disconnect the cold end of the collector choke and insert a milliammeter in series with the choke. Adjust I_{cq} by changing the value of the collector-bias resistor or by changing the collector voltage. All amplifiers requiring 12 V can be powered directly from low-current, 12-V dc sources (a three-terminal, 12-V regulator, for example). Amplifiers running on 14-18 V dc or 21-23 V dc can be powered by an LM317T (or LM317K) adjustable regulator. Set the regulated voltage to a minimum, and adjust it upward from minimum while monitoring I_{cq}. Idling collector current can be adjusted to suit your gain and power-output requirements.

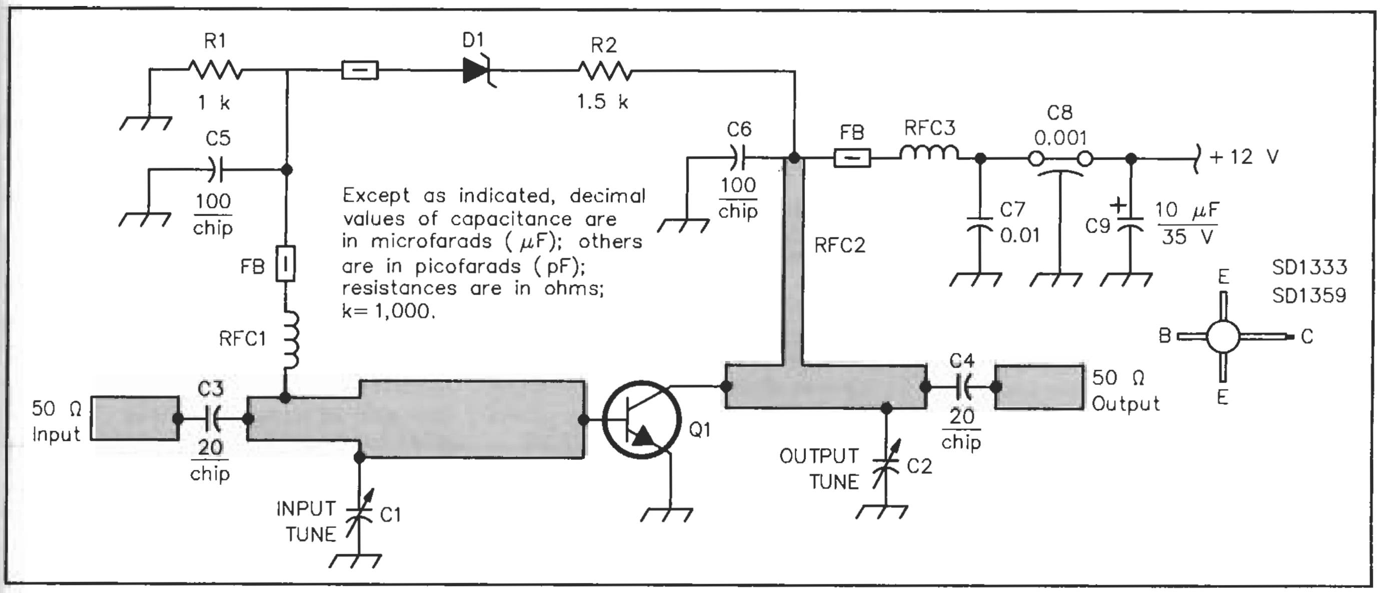

Fig 4—Schematic of the SD1333 and SD1359 amplifiers (Amplifiers No. 2 and No. 3). All components except Q1 are common to both designs (see text).

C1, C2—12.5-pF trimmer, Mouser no. 24AA071.
C7—0.01-µF disc ceramic.
D1—9.1-V, ½-W Zener.
RFC1, RFC3—8 turns of no. 24 enam wire, 0.1 inch ID, closewound.

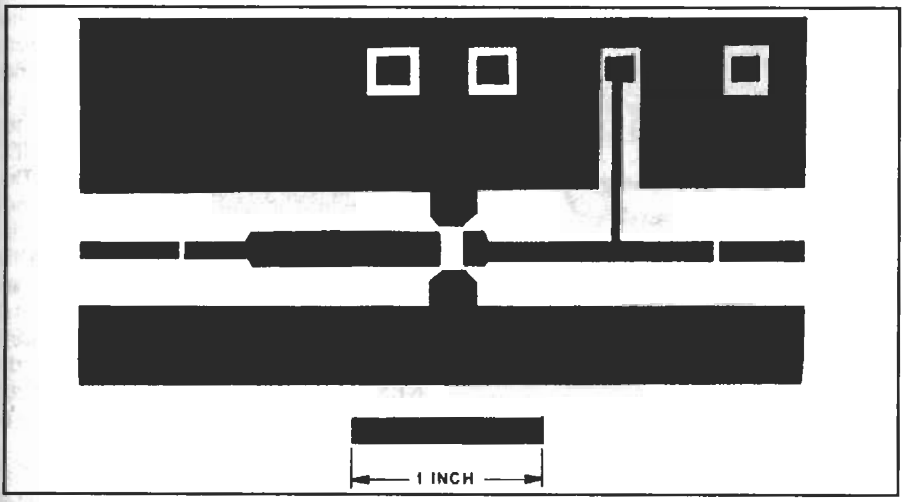

Fig 5—Full-scale PC-board network for the 100- and 250-mW amplifiers. Black areas represent unetched copper foil. All parts are mounted on the foil side of the board.

Amplifier No. 1: A Receive Preamplifier or 10-mW-Output, 13-dB-Gain Transmit Stage

The amplifier shown in Figs 1 and 2 is a low-power stage using an SD1330, Motorola MRF901 or NEC NE64535. This stage is best suited to the receiving side of a transverter. With a noise figure (NF) of 2 dB, it makes a fine front end by itself. With a good, low-noise GaASFET preamplifier in front of it, you'll have all the sensitivity you need in front of the receive mixer. For transmitting applications, where a low noise figure isn't important, an MMIC amplifier would be better because it is a lot easier to build for the same results.

The SD1330 amplifier is built on 0-10 double sided board; the artwork appears in Fig 3. After etching the board, install rivets at the RF and dc grounds. To mount the device, drill a hole in the PC board the size of Q1's macro-X case. This allows all four leads to be soldered to the microstrip without bending them.

The idling current for this amplifier should be set to 5 mA or less for good noise figure. The idling current is kept constant for different transistors and over different voltages and temperatures by current-limiting resistors Rl and R2 and Zener diode Dl. Tune the input trimmer for best NF or maximum gain if an NF-measurement setup isn't available.

Amplifier No. 2: 10 mW In, 100 mW Out

The SD1333 transistor, a macro-X plastic-packaged device used in this design, delivers 100-125 mW. This device is capable of a reasonable noise figure (2-3 dB) and good dynamic range, which allows it to be used as a second stage (following a GaAsFET preamplifier) in a receiving system. (Motorola's BFR96/MRF961/MRF962 also work well in this circuit; the BFR96 has slightly less gain.) Fig 4 shows the amplifier schematic and Fig 5 shows the full-scale artwork.

Zener-diode bias is used for simplicity and some temperature compensation. Set the quiescent current to 40-50 mA (not critical) by varying R1. A pot can be used initially, then the pot can be replaced by a fixed-value resistor. Or the collector voltage can be varied slightly.

After the rivets are installed in the PC board, drill a hole the diameter of Q1's molded package in the PC board. Mount the device in the board so its unbent leads are soldered directly to the microstrip board. Use the small variable capacitors to tune for maximum gain.

Amplifier No. 3: 9 dB Gain, 250 mW Out

This unit is a slightly higher-power version of the previous amplifier, using an SD1359 or Motorola TRF559 plastic-packaged device. The artwork and schematic are identical to that of Amplifier No. 2. Idling current, 40-50 mA, can be optimized by adjusting the collector supply between 10 and 13 V to set the stage gain.

After the PC-board rivets are installed, drill a hole in the PC board to accept Ql. Mount the device in the board so its unbent leads are soldered directly to the microstrip board. Tuning is as discussed with Amplifier No. 2.

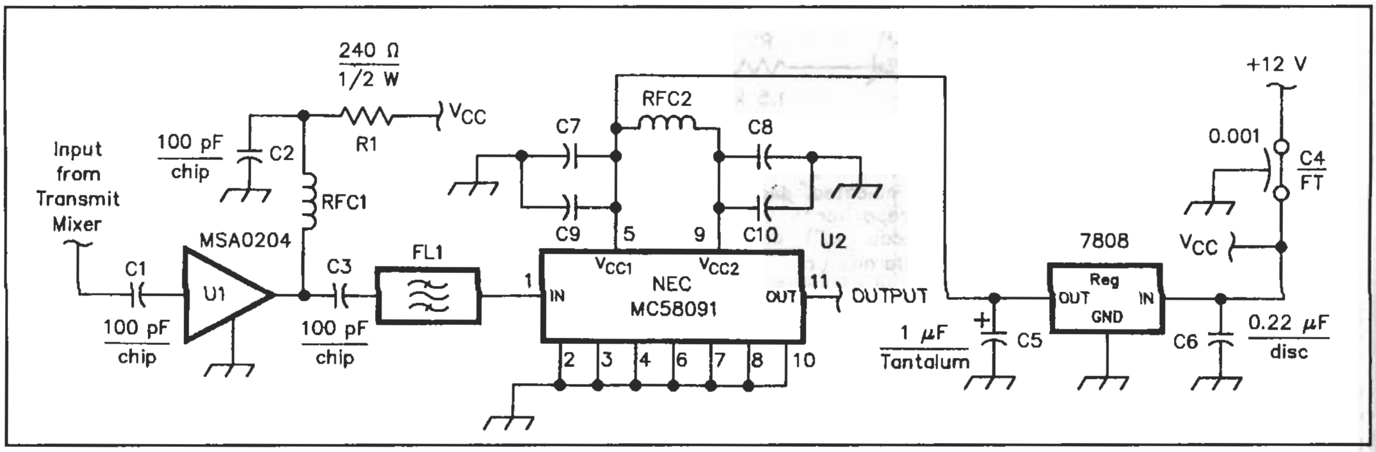

Fig 6—Schematic of the MSA 0204/MC5809L amplifier.

C7, C9—100-pF chip.
C8, C10—0.01-µF ceramic disc.
FL1—Two-stage Toko helical filter (Digi-Key no. TK23318).
RFC1—8 turns of no. 24 enam wire, 0.1 inch ID, closewound.
RFC2—1-µH molded choke.

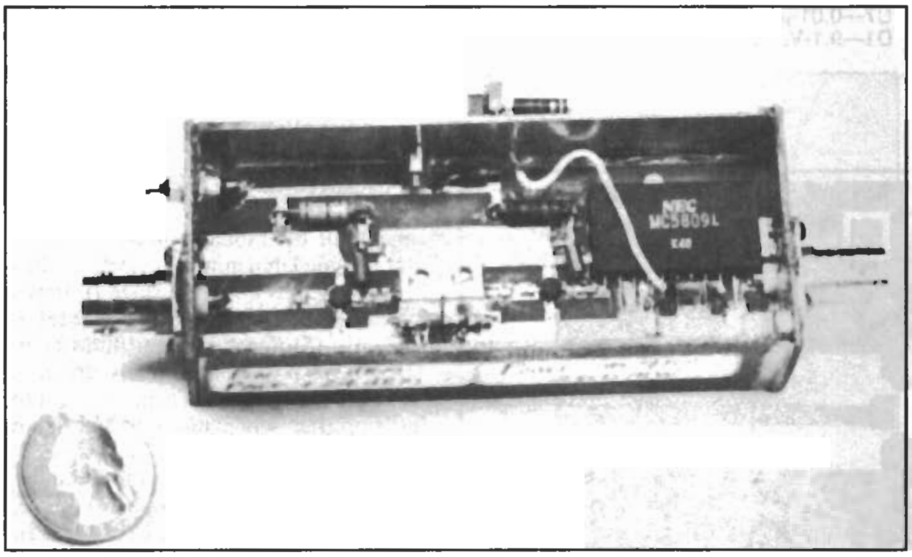

Fig 7—The 200-mW-output MSA 0204/MC5809L amplifier. In this prototype, I used MSA 0204 stages before and after the helical filter to provide addiional gain. The biasing components (the 7808 and associated parts) are mounted on the outside of the enclosure for convenience.

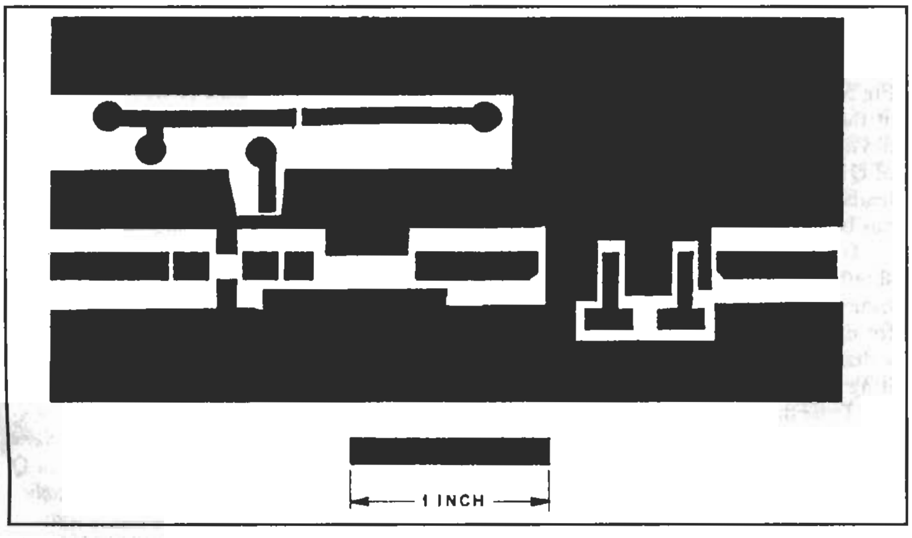

Fig 8—Full-scale PC-board artwork for the 200-mW amplifier. Black areas represent unetched copper foil.

Amplifier No.4:0.8 mW In, 200 mW Out

The amplifier shown in Figs 6 and 7 is a high-gain unit that can be used to follow the output of your 903-MHz transmit mixer. An Avantek MSA 0204 or Mini-Circuits MAR-2 MMIC is used for the first stage, followed by a Toko two-stage helical filter.[10] (In the prototype amplifier shown in Fig 7, I used two MMIC stages, in addition to the helical filter, before the final stage.) The output stage is an NEC MC5809L thick-film hybrid module. This NEC hybrid module is one of four such units intended for hand-held cellular telephones operating in the 800- to 960-MHz range. The MC5809-series amplifiers are rated at 150 mW minimum power output at 7.5 V dc. Using a more readily available power source (a 7808 regulator), the NE5809s I tried were linear at over 200 mW output. These hybrid amps are easy to use and are stable into any load.

The full-scale PC-board artwork is shown in Fig 8. The helical filter is used between stages to clean up the transmitter signal by filtering out the local-oscillator and image frequencies, but isn't required for applications where such filtering is done in other stages. A three-stage Toko filter[11] will also fit on the board, and could be used for better filtering. The filter leads are bent to the side to allow soldering to the PCboard traces. See Fig 9 for parts placement.

This amplifier is easy to build and get working. Because both active stages and the filter are designed for 50 Ω in and out, the amplifier uses 50-Ω microstrip throughout. No trimmer capacitors are needed and no bias adjustments are necessary—just apply the dc voltages and the drive signal, and peak the helical filter for maximum power output at 903 MHz!

Because more gain is available from this design than I needed, I tried mixing two signals at the input of the first stage. Using a T connection at the input of the first stage, I used the circuit to mix a 759-MHz local oscillator with a 144-MHz IF signal. The 903-MHz output signal was about 175 mW and was fairly clean. Using MMIC amplifiers for gain blocks, almost anything can be used as a mixer!

Amplifier No. 5: 35 mW In, 350 mW Out

The amplifier shown in Fig 10 is practically as simple as an MMIC amplifier. The transistor, an SD1598, produces more than 350 mW output when mounted in a 50-Ω line. The SD1598 (originally another part number) was designed by Bill Olson, W3HQT, years ago, when he was employed by Solid State Microwave (now SGS-Thomson Microelectronics). I've used this transistor, which I refer to as a "hot 2N3866," in amplifiers and frequency multipliers at frequencies from 144 MHz to 3.3 GHz. The SD1598 is in a studded package and the SD1598-1 is a studless package; either style is suitable for use at 903 MHz. I've built two-stage SD1598 amplifiers with 50-Ω lines that work from 400-1300 MHz. Such a two-stage amplifier fits nicely into a 1.5 × 3.6 × 1-inch Bud CU-123 the-cast box. Gain is 10 dB per stage at 903 MHz and 8-9 dB per stage at 1296 MHz.

PC-board artwork isn't necessary for this amplifier, as the 50-Ω line can be easily made on a piece of G-10 PC board using an X-ACTO knife and a straight edge. A 50-Ω line on 0-10 PC board is about 0.1-0.11 inches wide. After cutting the edges of the 50-Ω line, cut another line about a quarter inch from, and parallel to, the 50-Ω line on each side of the 50-Ω line. Using a soldering iron, heat the quarterinch-wide strips between the 50-Ω line and the outer sections and remove them with the knife or a needle-nose pliers. Leave some copper near the transistor for connecting Ql's emitter leads to the microstripline.

Drill a hole in the board for the transistor such that the leads of the transistor can be soldered flush to the microstripline and ground foils. This is important: Excessive lead lengths drastically reduce the gain of this amplifier. Attach a small heat sink to the stud of the device.

The SD1598 can be powered by a 12- to 18-V supply, depending on the gain and power output required. Use an LM317T voltage regulator in the power supply, and vary the voltage to set the gain of this lowlevel stage. This is an easy way to match the drive levels needed by the following stages.

Set Ql's quiescent current, I_{cq}, to 30-50 mA. Vary Rl to get the correct idling current. This bias current is not critical, and depends on the device and the collector voltage used. Tune up is simple: Adjust C8 for maximum power output.

Amplifier No. 6: 100 mW In, 1 W Out

A pair of SDl598s and a Wilkinson power divider/combiner combination are used in this design. The schematic for the combined amplifier is shown in Fig 11, and the PC-board artwork is shown in Fig 12. I didn't make an effort to terminate the 75-Ω Wilkinson divider/combiner; the amplifier works fine as is. Power sharing between the two devices is excellent, and a 3-dB improvement in power output is available over a single device. Saturated power output is well over 1 W.

The supply voltage and idling currents are the same as single-device Amplifier

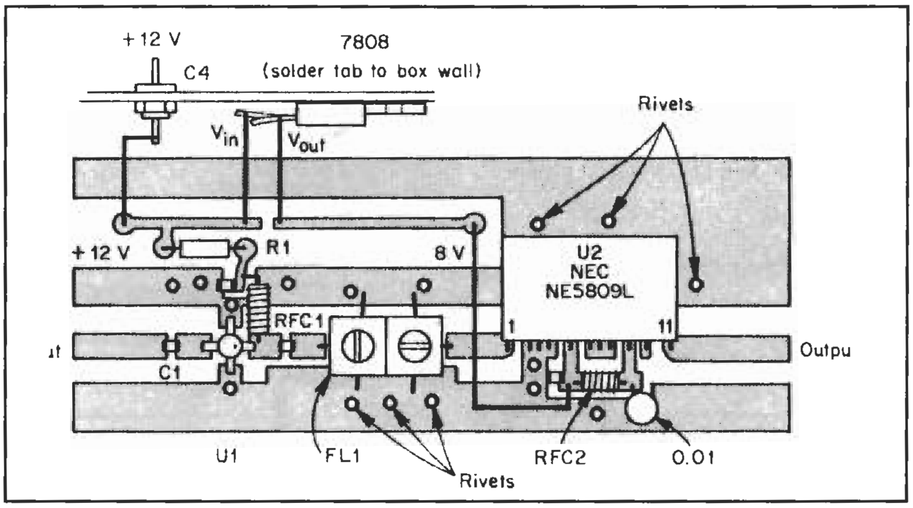

Fig 9—Parts-placement diagram for the 200-mW amplifier. All parts are mounted on the foil side of the board. Note the locations of rivets. At each V_{cc} lead on the MC5809L (pins 5 and 9), install a 100-pF chip capacitor and a 0.01-µF disc to ensure proper bypassing.

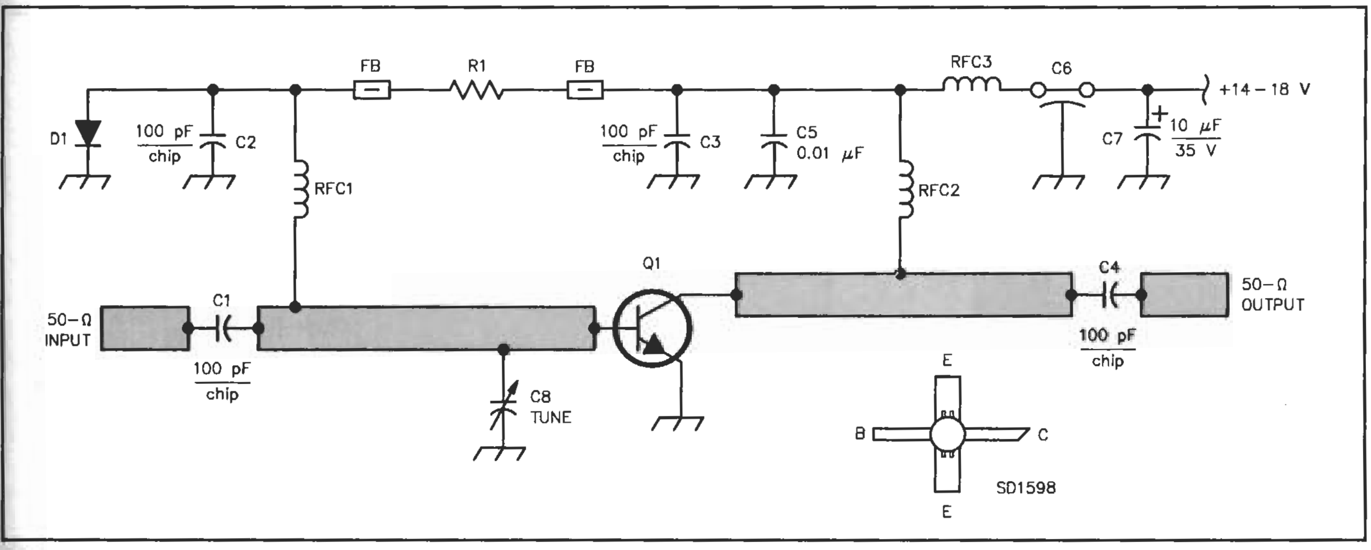

Fig 10—The 350-mW amplifier. Glue D1 to Q1's ceramic body for thermal composition.

C6—0.001-µF feedthrough.
C8—0.6- to 8-pF Johanson piston trimmer.
D1—1N4001.
Q1—SD1598.
R1—1.5-1.6 kΩ, ½ W.
RFC1, RFC2—8 turns of no. 26 enam wire, closewound, 0.1 inch ID.

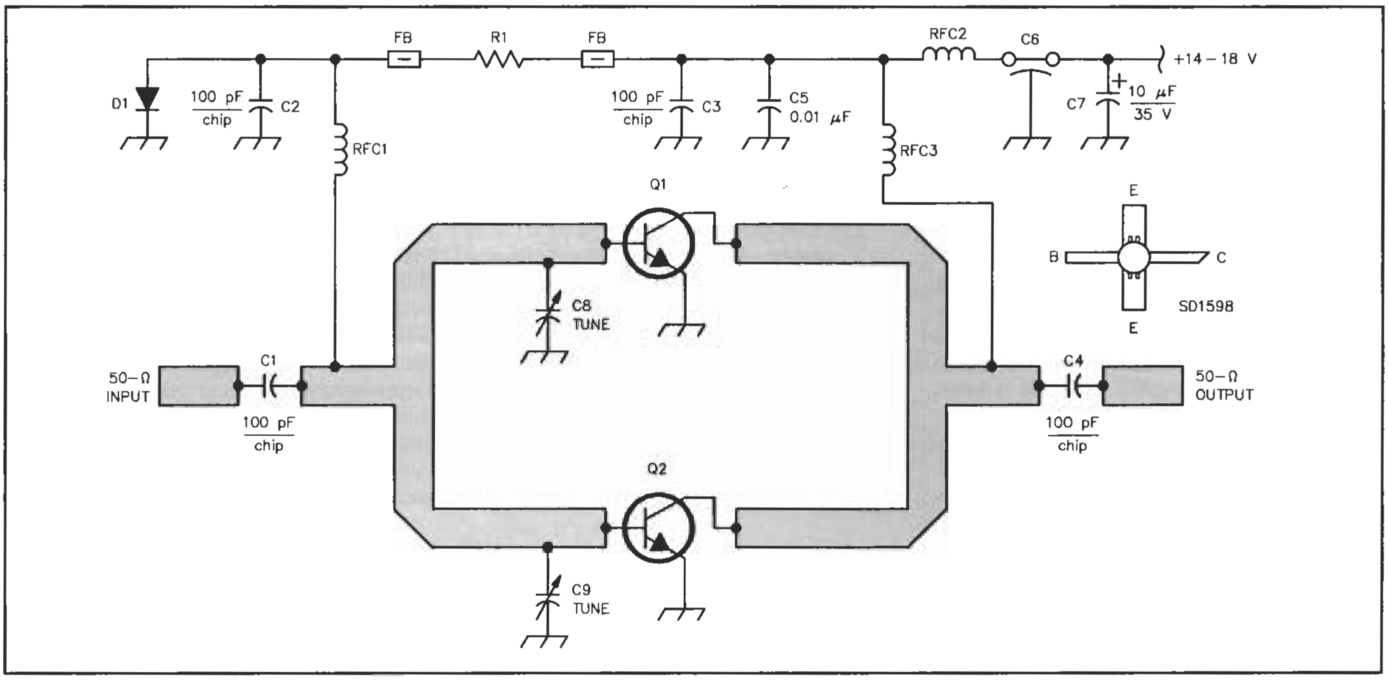
FB
R1
FB
RFC2
C6
+14 – 18 V
D1
100 pF
chip
C2
100 pF
chip
C3
C5
0.01 µF
C7
10 µF
35 V
RFC1
RFC3
Q1
E
B
C
SD1598
E
C8
TUNE
50-Ω
INPUT
C1
100 pF
chip
C4
50-Ω
OUTPUT
100 pF
chip
Q2
C9
TUNE

Fig 11—The 2x SD1598 amplifier. Glue D1 to Q2's ceramic body for thermal composition.

C8, C9—0.6- to 8-pF Johanson piston trimmer.
D1—1N4001.
Q1, Q2—SD1598.
R1—1.5-1.6 kΩ, ½ W
RFC1, RFC3—8 turns of no. 26 enam wire, closewound, 0.1 inch ID.

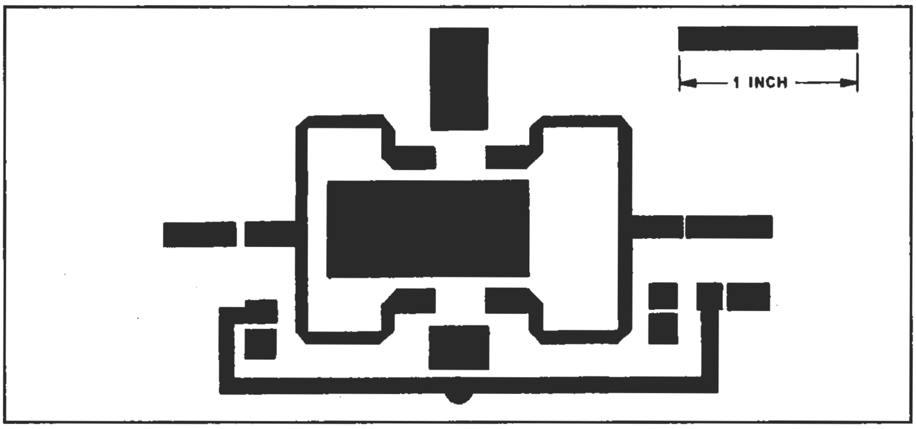

Fig 12—Full-scale PC-board artwork for the 2x SD1598 amplifier. Black areas represent unetched copper foil.

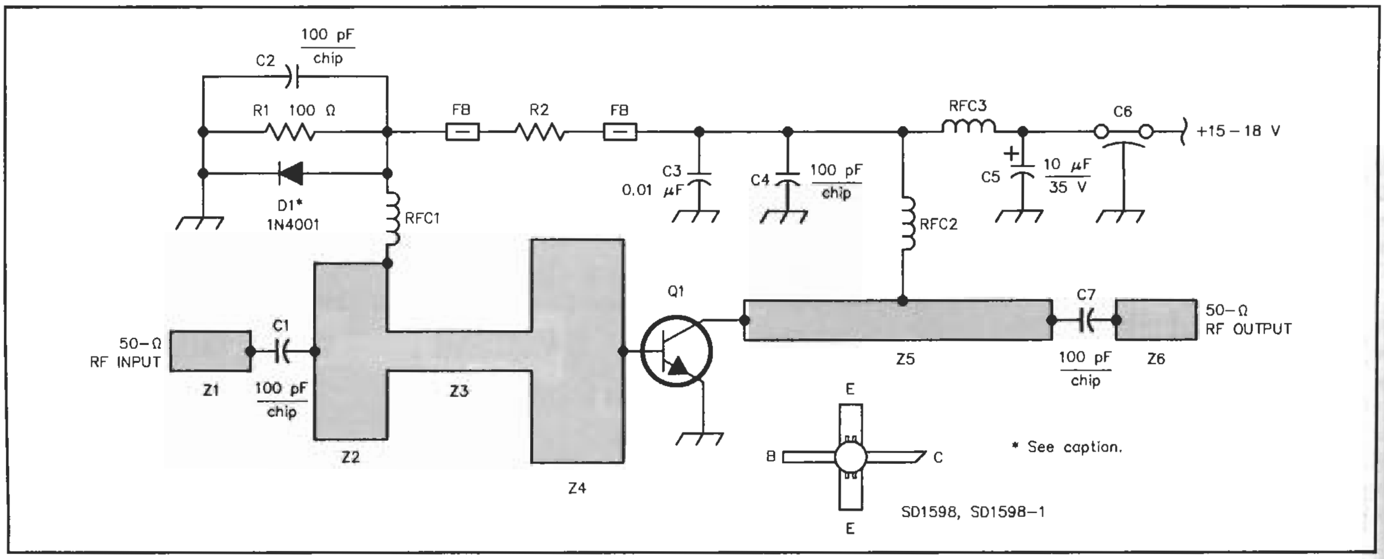
100 pF
chip
C2
R1 100 Ω
FB
R2
FB
RFC3
C6
+15 – 18 V
D1*
1N4001
RFC1
C3
0.01 µF
C4
100 pF
chip
C5
10 µF
35 V
RFC2
Q1
C7
50-Ω
RF OUTPUT
50-Ω
RF INPUT
C1
Z1
100 pF
chip
Z3
Z2
Z4
Z5
100 pF
chip
Z6
E
B
C
E
SD1598, SD1598–1
* See caption.

Fig 13—The 500-mW broadband amplifier. For power levels over 500 mW output, glue D1 to Q1's ceramic body for thermal compensation. RFC1 and RFC2 are implemented as PC-board traces.

Q1—SD1598 or SD1598-1.
R2—1.5-1.6 kΩ, ½ W
RFC3—8 turns of no. 28 enam wire, closewound, 0.1 inch ID.
Z1-Z5—Microstriplines. See text and Fig 14.

No. 5 (Fig 10). If this amplifier is mounted in a die-cast box, no additional heat sinking is required; If not, a small heat sink should be attached to the stud of each device. Adjust C8 and C9 for maximum power output.

Amplifier No.7: 12dB Gain, 500mW Out

The amplifier shown in Fig 13 is a broadband design that's been around for about 6 years. Bill Olson, W3HQT, designed it as a 1296-MHz amplifier. (The 1296-MHz version differs in that it has a small piston trimmer on the output circuit, about an inch from Qi's ceramic cap.)

This amplifier is a very versatile design; it works from 900-1300 MHz and has a power gain of 10-13 dB over this range. I built two test amplifiers for this design: one for the studded transistor and one for the flanged version. Over a dozen different devices were tried in these test circuits, both at 903 MHz and at 1296 MHz. Classes A, AB, C and pulsed-class-C operation were all tried, and all worked well.

The studded SD1598 is used in this design. Full-scale artwork appears in Fig 14; Fig 15 shows parts placement. As with the other designs, rivets are used at all dc and RF grounds. After the rivets are soldered in place, drill a hole in the board for the device. Trim Q1's leads with scissors, then solder them directly to the microstriplines and ground foils. If the amplifier is to be mounted in a die-cast box, the box will provide sufficient heat sinking. If you use another mounting method, attach a small heat sink to Q1's stud.

Power output for SSB is 500 mW, and

Fig 14—Full-scale PC-board pattern for the broadband 500-mW SD1598 amplifier. Black areas represent unetched copper foil.

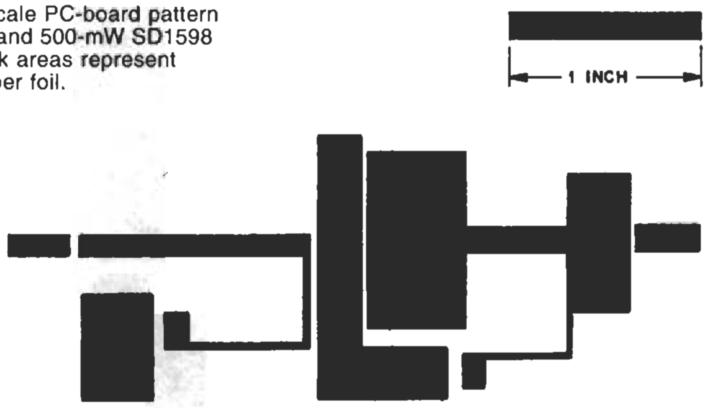

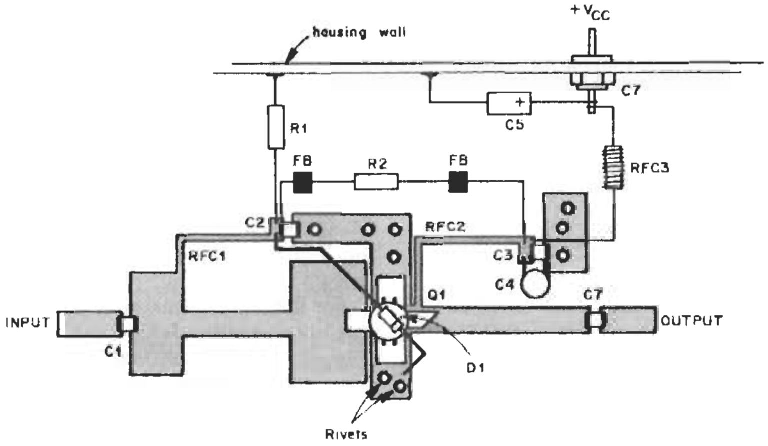
+V_CC
housing wall
C7
R1
C5
FB
R2
FB
RFC3
C2
RFC2
RFC1
C3
C4
Q1
C7
INPUT
OUTPUT
C1
D1
Rivets

Fig 15—Parts-placement diagram for the broadband 500-mW SD1598 amplifier. The PC-board edges are not shown. All components are mounted on the foil side of the PC board (except those mounted to the enclosure).

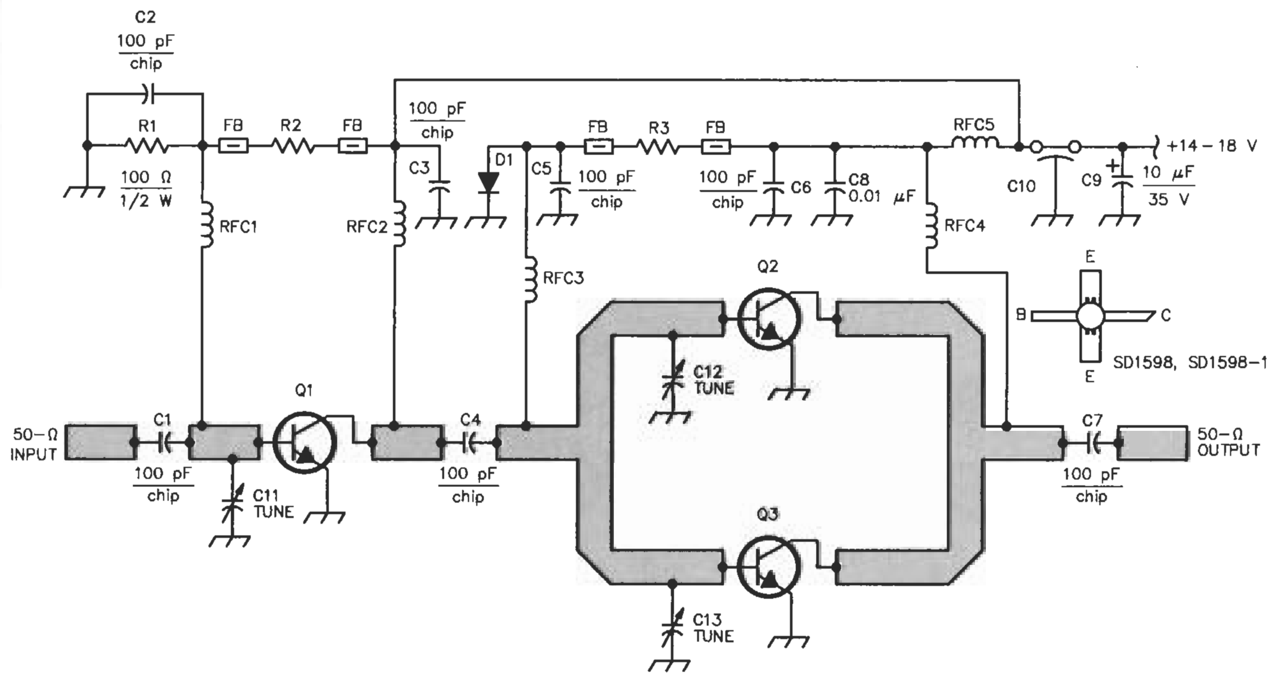
C2
100 pF
chip
R1
100 Ω
1/2 W
FB
R2
FB
RFC1
RFC2
100 pF
chip
C3
D1
C5
FB
R3
FB
100 pF
chip
100 pF
chip
C6
C8
0.01 µF
RFC5
+14 – 18 V
C10
C9
10 µF
35 V
RFC4
RFC3
Q2
E
B
C
E
SD1598, SD1598-1
C12
TUNE
Q1
50-Ω
INPUT
C1
100 pF
chip
C11
TUNE
C4
100 pF
chip
Q3
C13
TUNE
C7
100 pF
chip
50-Ω
OUTPUT

Fig 16—The two-stage, 1-W SD1598 amplifier. Glue D1 to Q2's ceramic body for thermal compensation.

C10—0.001-µF feedthrough.
C11, C13—0.6- to 8-pF Johanson piston trimmer.
Q1—SD1598 or SD-1598-1.
Q2, Q3—SD1598.
R2, R3—1.5-1.6 kΩ, ½ W.
RFC1-RFC5—8 turns of no. 24 enam wire, closewound, 0.1 inch ID.

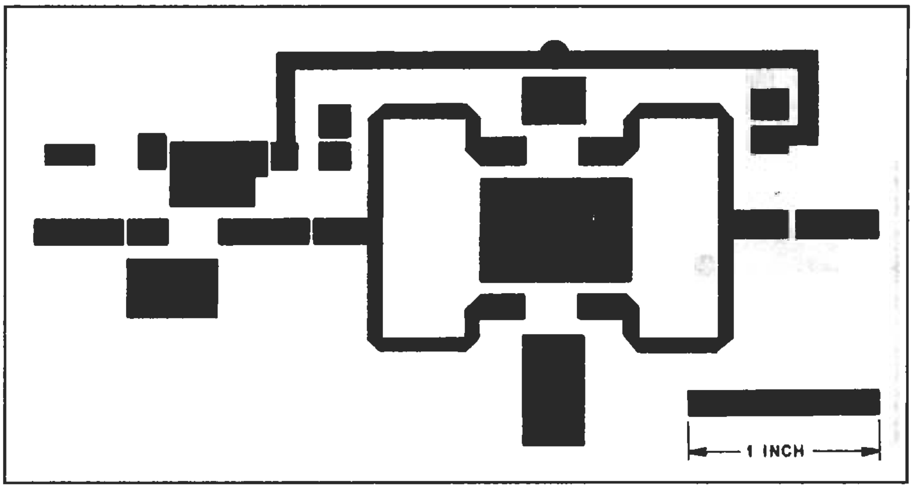

Fig 17—Full-scale PC-board pattern for the two-stage, 1-W SD1598 amplifier. Black areas represent unetched copper foil.

up to 1 W for CW and FM, depending on the device. In applications where the device will deliver over 500 mW, a IN4001 (DI) diode can be glued to Q1 for thermal stabilization, as shown pictorially in Fig 15. Power gain at 1 W output is 7 dB or so. By varying R2, set Q1's I_{cq}, for 30-50 mA for power levels up to 500 mW, and 15-20 mA for higher levels.

For low-level stages (up to 200mW out), you can use the SD1598-1 studless part in this circuit. Solder the studless device to the microstrip, then solder a piece of hobby brass to the device's gold-plated bottom area and to the ground plane. This provides enough heat sinking to dissipate a few hundred milliwatts.

At 903 MHz, no output trimmer capacitor is required. Some improvement in the output match can be achieved by trimming the width of the output line by a few thousandths of an inch with an X-ACTO knife.

Amplifier No. 8: 10 mW In, 1 W Out

A second 1-watt amplifier, this one using a pair of SD1598's driven by a single SD 1598, is shown in Fig 16. This design is similar to that of Amplifier No.6, with the addition of a driver stage. The artwork appears in Fig 17.

For the driver transistor, you can use an SD1598-1 studless device, as Q1 must provide only a hundred milliwatts or so. The driver is mounted in a hole in the PC board. A small piece of brass shim stock is soldered to the bottom of the device and to the ground-plane side of the PC board for heat sinking.

This amplifier runs on 14-18 V dc. The idling current for each device is around 30-50 mA. Again, the supply voltage can be provided by an LM317T regulator. Adjust R2 and R3 to individually set each device's idling current. Then, tune the amplifier by adjusting C11, C12 and C13 for maximum output. Vary the supply voltage to optimize the required power output or gain.

The 2.4 × 4.4 × 1-inch Bud CU-124, or an equivalent die-cast box, is a suitable enclosure for this amplifier. Mount the connectors on the ends of the box, and place the amplifier PC board inside the box body, instead of on the inside of the cover. The box provides sufficient heat sinking for this amplifier.

Next month, I'll describe three more 903-MHz linear amplifiers, including two 4-W units.

Notes

[1] D. Mascaro,"A 759-MHz Local Oscillator," *QEX*, May 1988, pp 12-15.

[2] No-tune transverters, as well as antennas and microwave components, are available from Down East Microwave, Box 2310, RR 1, Troy, ME 04987, tel 207-948-3741. Catalog available.

[3] The amplifiers in this article use transistors manufactured by SGS-Thomson Microelectronics, 211 Commerce Dr, Montgomeryville, PA 18936, tel 215-362-8500.

[4] Mouser Electronics, 2401 Hwy 287 N, Mansfield, TX 76063, tel 817-483.4422. Catalog available.

[5] Microwave Components of Michigan, 17141 Merriman, Romulus, MI 48174, evenlng tel 313-753-4581. Parts list available.

[6] RF Parts, 1320 Grand Ave, San Marcos, CA 92069, tel 619-744-0728.

[7] SGS-Thomson Microelectronics transistors are available through RF Gain Ltd, 100 Merrick Rd, Rockville Center, NY 11570, tel 800-645-2322. $50 minimum order.

[8] Frontier Microwave, RD 1, Box 467, Mink Rd, Ottsville, PA 18942, evening tel 215-795-2648.

[9] D. Mascaro,. "A Transverter Band-Switching Display and Universal Power Supply," *QEX*, Aug 1987, pp 8-14.

[10] Available from Digi-Key Corp, 701 Brooks Ave South, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539. Catalog available.

11Toko filters and other components are available from Steve Kostro, N2CEI, Box 341A, RD 1, Frenchtown, NJ 08825, evening tel 201-996-3584. Parts list available.

By Dave Mascaro, WA3JUF **From *QST*, July 1990**

903-MHz Linear Amplifiers

Part 2—Did you like the projects in Part 1? Here are three more amplifiers to suit your 903-MHz, 50-ohm gain-block needs.

In Part 1, I covered construction methods for building the 903-MHz linear amplifiers described in this twopart series. The amplifiers described this month cover a higher power range than those in Part 1, starting with a 2-W-output design, and finishing with a pair of 4-W output units (one with 13 dB gain, the other with 23 dB gain).

Amplifier No. 9: 100 mW In, 2 W Out

The transistor used in this amplifier, an SD1853, is a class-A device in a "strip-pac" flanged package. The PC-board artwork is shown in Fig 18, and the schematic appears in Fig 19. The parts-placement

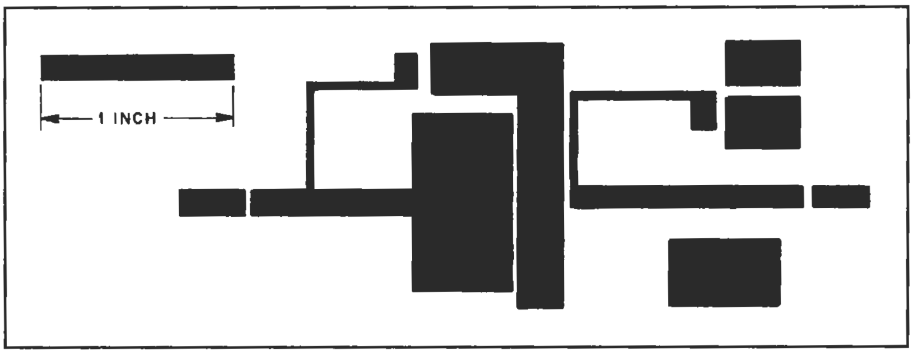

Fig 18—Full-scale PC-board pattern for the 2-W SD1853 amplifier. Black areas represent unetched copper foil.

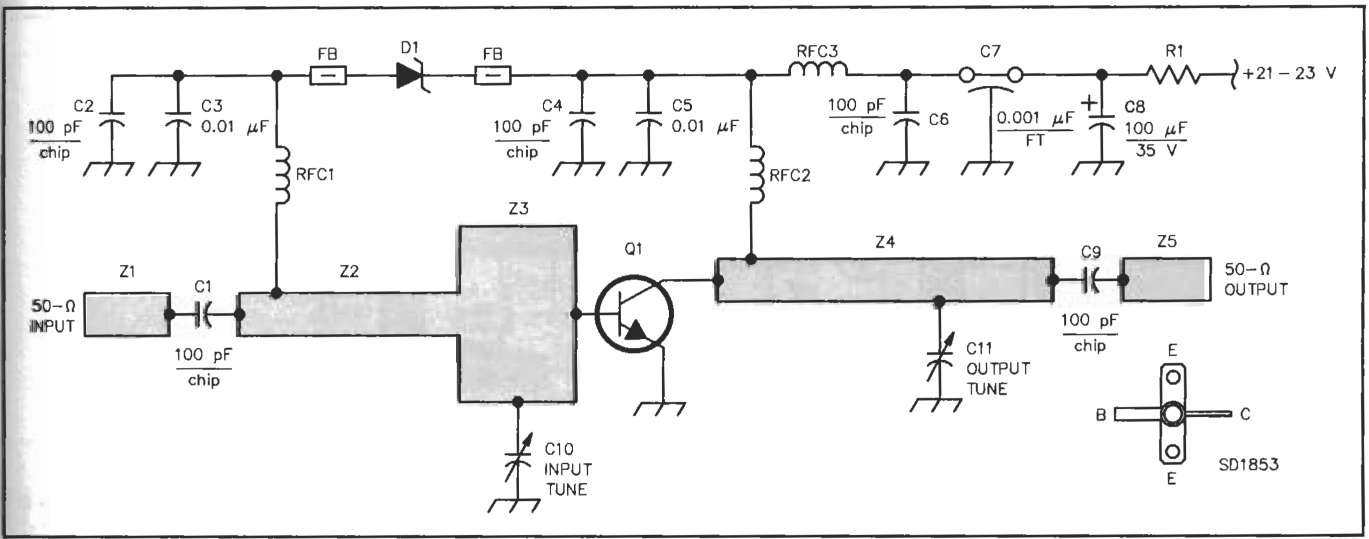
FB
D1
FB
RFC3
C7
R1
+21 – 23 V
C2
100 pF
chip
C3
0.01 µF
C4
100 pF
chip
C5
0.01 µF
100 pF
chip
C6
0.001 µF
FT
C8
100 µF
35 V
RFC1
RFC2
Z3
Q1
Z4
C9
Z5
Z1
C1
Z2
50-Ω
INPUT
50-Ω
OUTPUT
100 pF
chip
100 pF
chip
C11
OUTPUT
TUNE
C10
INPUT
TUNE
E
B
C
E
SD1853

Fig 19—The 2-W SD1853 amplifier. RFC1 and RFC2 are implemented as PC-board traces.

C10, C11—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1—SD1853.
R1—2-3 Ω, 1 W.
RFC3—8 turns of no. 26 enam wire, closewound, 0.1" ID.
Z1-Z5—Microstriplines. See text and Fig 18.

diagram is shown in Fig 20, and a photo of my prototype appears in Fig 21.

After the board is etched and rivets are installed, cut a hole and file its edges to accept Q1's flange. The SD1853 must be mounted to a heat sink. (A Bud CU-124 or equivalent die-cast box is adequate.) Use a piece of G-10 PC board between the amplifier board and the heat sink, to allow the device to be mounted flush to the microstrip. The device flange is the emitter, and a low-inductance ground connection is a must. Make this connection by using a piece of copper-foil tape soldered to the microstrip ground plane and placed between Q1's flange and the heat sink. Mount the transistor to the heat sink with no. 4-40 screws. Use heat-sink compound between Q1 and the heat sink.

An LM317T (or LM317K) adjustable regulator can be used to supply the voltage required by this amplifier. Use a 10-turn pot for the ADJUST control; smooth adjustment of the supply voltage to this amplifier is a must. Insert a milliammeter in series with current-limiting resistor R1, and slowly adjust the regulator output up from minimum until the idling current is 200-225 mA without RF drive applied. The exact supply voltage depends on the beta of

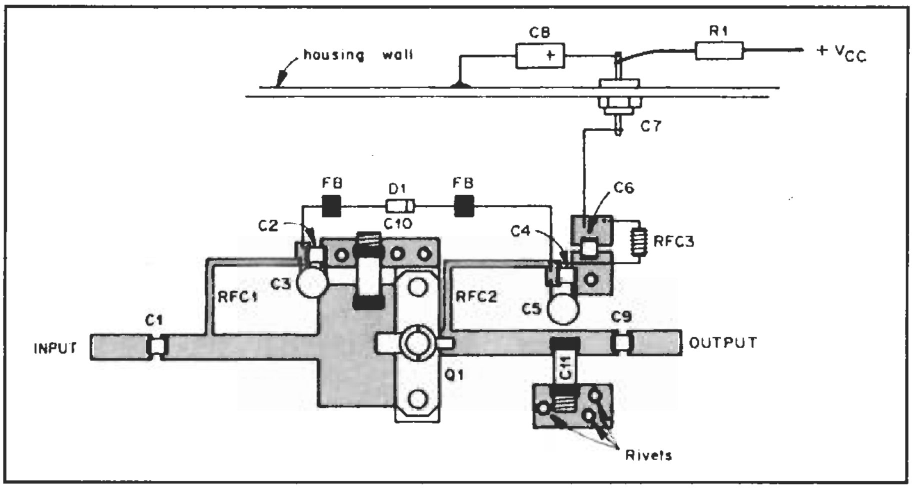
housing wall
C8
R1
+ V_{CC}
C7
FB
D1
FB
C6
C2
C10
C4
RFC3
RFC1
C3
RFC2
C5
C1
C9
INPUT
OUTPUT
C11
Q1
Rivets

Fig 20—Parts-placement diagram for the 2-W SD1853 amplifier. The PC-board edges are not shown. All components are mounted on the trace side of the PC board (except those mounted to the enclosure).

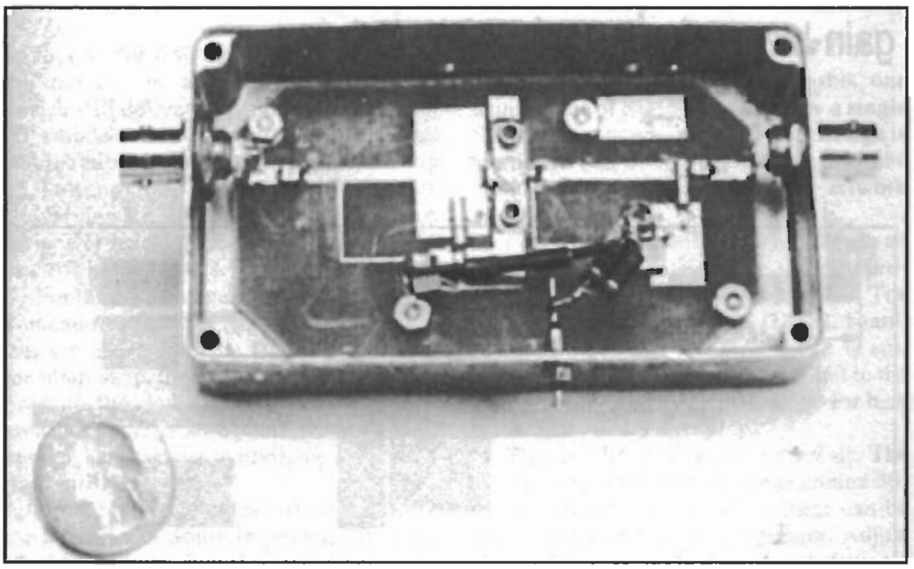

Fig 21—The prototype 2-W amplifier.

Fig 22 (below)—The 4-W, 2 x SD1853 amplifier.
C4-C6—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1, Q2—Sd1853.
R1—2-3 Ω, 1 W.
RFC1, RFC2—8 turns of no. 26 enam wire, closewound, 0.1" ID.
Z1-Z5—Microstriplines. See text and Fig 23.

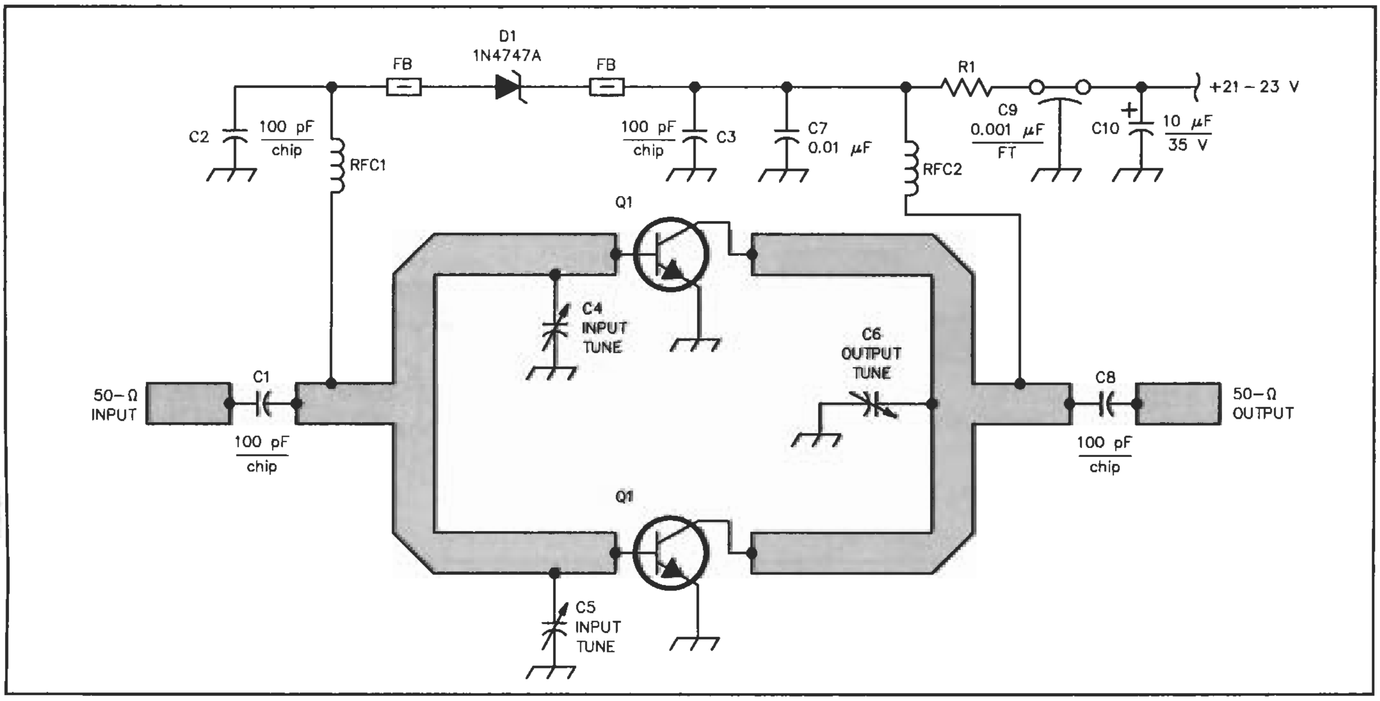
D1
1N4747A
FB
FB
R1
+21 – 23 V
C2
100 pF
chip
RFC1
100 pF
chip
C3
C7
0.01 µF
RFC2
C9
0.001 µF
FT
C10
10 µF
35 V
Q1
C4
INPUT
TUNE
C6
OUTPUT
TUNE
50–Ω
INPUT
C1
100 pF
chip
C8
100 pF
chip
50–Ω
OUTPUT
Q1
C5
INPUT
TUNE

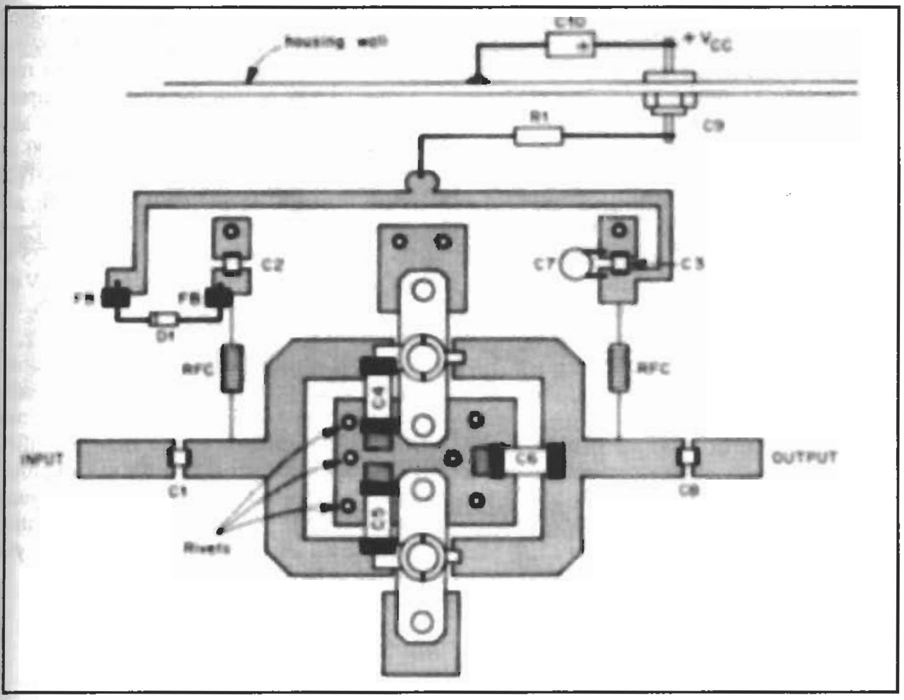
housing wall
C10
+Vcc
R1
C9
C2
C7
C3
FB
FB
D1
RFC
RFC
C4
C5
C6
INPUT
OUTPUT
C1
C8
Rivets

Fig 23—Parts-placement digram for the 2 x SD1853 amplifier. The PC-board edges are not shown. All components mount to the trace side of the PC board (except those mounted to the enclosure).

Q1, the value of R1, and D1's breakdown voltage. Idling current can be adjusted slightly to optimize gain and/or power output. The supply voltage must be removed during receive to minimize device heating.

Tune-up is simple. Apply drive and tune C10 and C11 alternately for maximum power output. If you have access to a return-loss bridge or network analyzer, you may want to trim the width of Z2 to improve the input return loss. Verify stability (as indicated by no output and no change in supply current) by tuning the trimmers through their ranges with no input signal applied.

Amplifier No. 10: 200 mW In, 4 W Out

This high-gain design consists of a pair of SD1853s combined in a Wilkinson power divider/combiner. As before, I made no effort to terminate the 75-Ω Wilkinson divider/combiner with balancing resistors. It worked just fine without them. Even the 1296-MHz version, which uses the same board layout, worked well in this configuration.

Fig 22 shows the schematic of this class-A amplifier, and Fig 23 shows parts placement. The PC-board pattern, the same

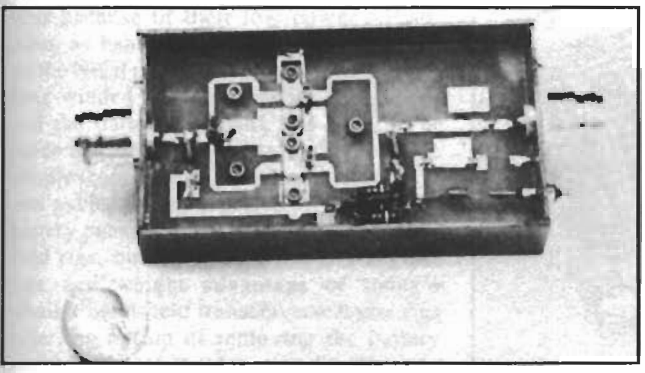

Fig 24—This prototype 4-W amplifier was built on a PC board with a spare driver stage that's bypassed with a brass strip.

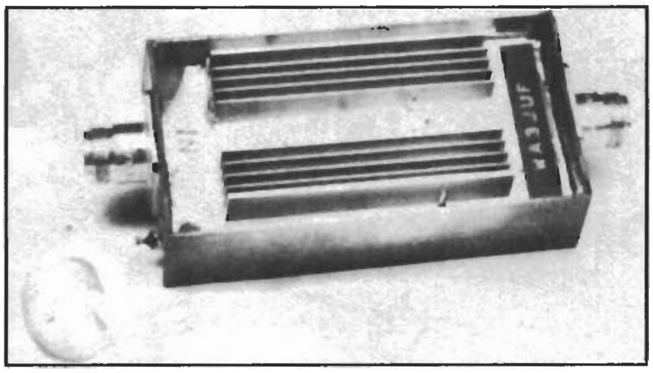

Fig 25—The 4-W amplifier's heat sink, bolted directly to the devices, must be positioned such that cooling air can flow over it during amplifier operation.

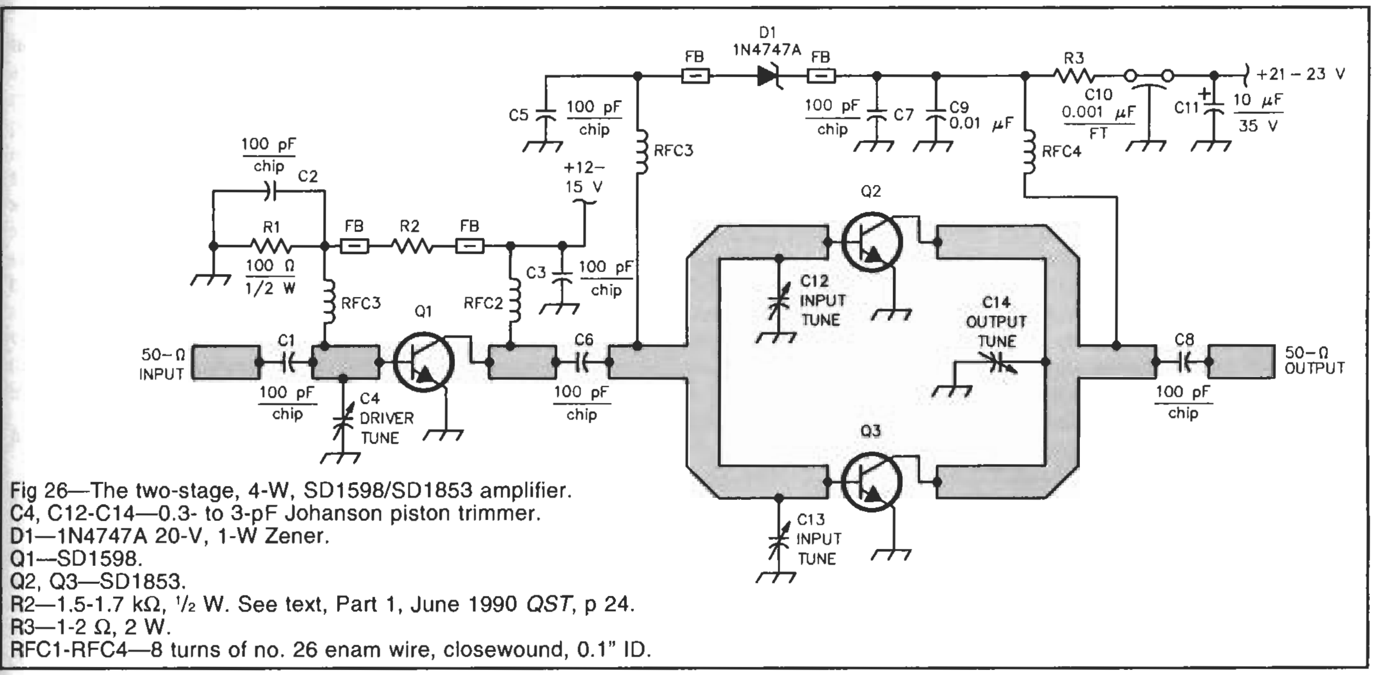
D1
1N4747A
FB
FB
R3
+21 – 23 V
C5
100 pF chip
RFC3
100 pF chip
C7
C9
0.01 μF
C10
0.001 μF
FT
C11
10 μF
35 V
RFC4
100 pF chip
C2
+12–15 V
Q2
R1
FB
R2
FB
100 Ω
1/2 W
C3
100 pF chip
RFC3
Q1
RFC2
C12
INPUT TUNE
C14
OUTPUT TUNE
50–Ω INPUT
C1
C6
C8
50–Ω OUTPUT
100 pF chip
C4
DRIVER TUNE
100 pF chip
100 pF chip
Q3
C13
INPUT TUNE

Fig 26—The two-stage, 4-W, SD1598/SD1853 amplifier.
C4, C12-C14—0.3- to 3-pF Johanson piston trimmer.
D1—1N4747A 20-V, 1-W Zener.
Q1—SD1598.
Q2, Q3—SD1853.
R2—1.5-1.7 kΩ, ½ W. See text, Part 1, June 1990 *QST*, p 24.
R3—1-2 Ω, 2 W.
RFC1-RFC4—8 turns of no. 26 enam wire, closewound, 0.1" ID.

as that for Amplifier No. 8, is shown in Fig 17 (shown in Part 1). The board's driver-stage traces are by-passed with a brass strip in this amplifier. Fig 24 shows my prototype mounted in a PC-board enclosure. A single base-bias source, using a current-limiting resistor and Zener diode, feeds both devices. Total idling current, measured in series with R1, is 400-450 mA.

The same mounting arrangement used with the single-SD1853 amplifier (Amplifier No. 9) applies to this unit. The device flanges must have low-inductance ground connections. Mount the transistors to a heat sink about 2 × 3 inches with ½ inch or taller fins, as shown in Fig 25. Again, the supply voltage is switched on only during transmit. To tune this amplifier, adjust C4, C5 and C6 alternately for maximum output.

Amplifier No. 11: 20 mW In, 4 W Out

A second 4-W amplifier, using a pair of SD1853s driven by a single SD1598, is shown in Fig 26. The final stage is Amplifier No. 10, and the driver is the same as that used in Amplifier No. 8. The PC-board artwork is shown in Fig 17; the parts placement can be done using Figs 23 and 16 as guides.

The power-supply and bias adjustments are the same as those for the previously described amplifiers using these devices. Transistor mounting is also the same as before, but extra heat sinking is required with this unit because of the class of operation and the power output.

Higher Power Output

Higher-power devices usable at 903 MHz are available from several manufacturers. NEC's NEL1320 (2SC3542) produces 18-20 W linear output at 903 MHz. These devices also work on the 1296-MHz band. These transistors, as well as single- and dual-stage amplifiers based on them, are available from Down East Microwave (DEM). (The DEM two-stage 3318PA runs around 18 W output for about 1 W input).

SGS-Thomson Microelectronics also makes high-power devices for the 800- to 960-MHz range. For example, the SD1423 is a 4-W-in, 30-W-out class AB device, and the SD1660 is a class-AB transistor that runs a whopping 120 W output for about 35 W drive. Both of these devices run on 28 V dc.

Summary

As I've shown in this two-part series, there's really nothing difficult about building your own 903-MHz equipment—from the local oscillator to the final amplifier. This ease of construction is a vast improvement over what it was just a few years ago—and the 903-MHz band's population has grown as a reflection of that. See you on the microwaves!

By Dave Mascaro, WA3JUF

From *QST*, August 1988

25-Watt Linear Amplifiers for 144 and 200 MHz

Do you need more punch from your hand-held or portable rig for 2 meters or 220 MHz? These little amplifiers can supply it!

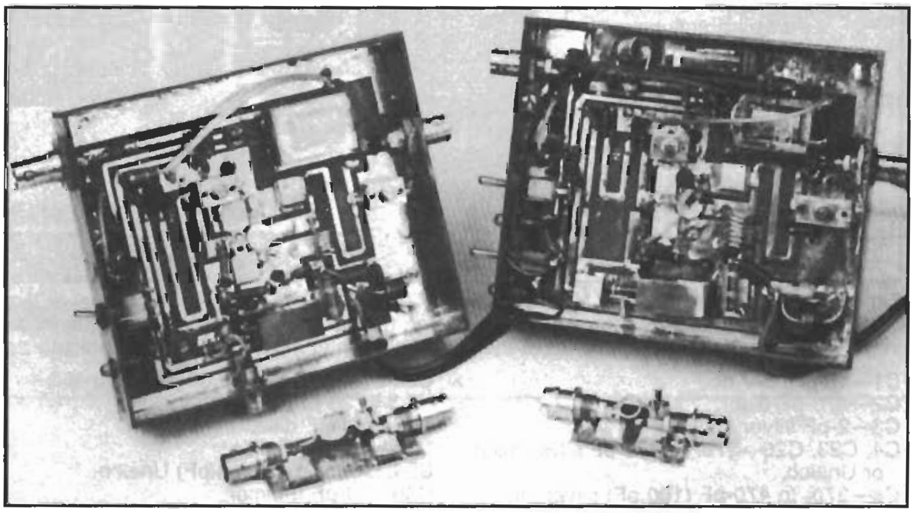

Many hams have 2-meter and 220-MHz hand-held rigs and low-power portable SSB rigs these days. Unfortunately, many of these radios have limited usefulness because of their low power output. Also, as hand-held radios get smaller, so do the NiCd packs that power them—a few long-winded transmissions on high power and the battery is dead! The low-power mode is usually good only for short-range simplex operation, or repeater use when you're close to the repeater. High-capacity battery packs are available for most hand-held rigs, but using them takes away the size and weight advantage of today's smaller hand-held transceivers. Some rigs offer the option of removing the battery pack and plugging the unit directly into 12-V dc, which gives you a small, lightweight rig, but you're still stuck with relatively low power.

The solution to this problem is the addition of a linear amplifier. An amplifier after the hand-held transceiver or portable rig that can give you 25 to 30 W of output power (depending on the output of the driving rig) without imposing large weight, cost and current-drain constraints is a great addition to your VHF station. An amplifier also allows you to use the same low-powered rig in the house and in the car with high power output. You can mount an amplifier under the seat or in the trunk of your car, and minimize the possibility of theft by taking the radio with you when you leave the car.

Amplifier Design

The amplifiers described in this article are capable of 25 to 30W of RF output for the drive power available from common handheld and portable rigs. The 2-meter and 220-MHz designs are so similar that we can describe them both in one discussion. The differences in the designs are covered in detail. Design, assembly, tuning and use of the amplifiers is virtually identical for both versions. Both versions even use the same PC board! (Note that two different boards are shown in the lead photo—they were prototypes of the final board design.)

Receiving preamplifiers are included in both units. They, too, are quite similar, differing only in some component values.

This power amplifier uses a single SD1274 bipolar transistor manufactured by Thomson Components/Mostek Corp.[1] The device is operated class AB for all-mode operation. Nominal dc power-supply voltage is 13.8. The amplifier will operate on any dc supply voltage between 12 and 14.5 (the typical automobile supply voltage range).

The main amplifier PC board is a microstripline design on standard $^1/_{16}$-inch-thick, double-sided, G-10 fiberglass-epoxy board. Input and output tuning capacitors are provided for maximizing gain and power output in a given band segment. If desired, the amplifiers can be tuned for broadband operation with only a slight reduction in power output across the operating frequency range.

The 2-meter and 220-MHz amplifiers are narrow-band designs. The 2-meter version gives 25 W output for 2 W of drive from 138 to 150 MHz, with a single tuning setting. When tuned for narrow-band operation, the amplifier gives about 20 W output for 1 W of drive. With the Kenwood TH-21AT on high power, my 2-meter amplifier puts out 25 W. On low power, I can adjust the amplifier for a maximum narrow-band gain, which gives about 7 to 8 W output. Efficiency is 50 to 60%, depending on tuning and power output.

The tuning range of the 220-MHz amplifier is 200 to 230 MHz. When tuned for narrow-band operation, the amplifier gives about 16 W output for 1 W of drive. Driving the amplifier with my ICOM IC-3AT on high power (about 1.5 W), the amplifier puts out 20 W. With the IC-3AT on low

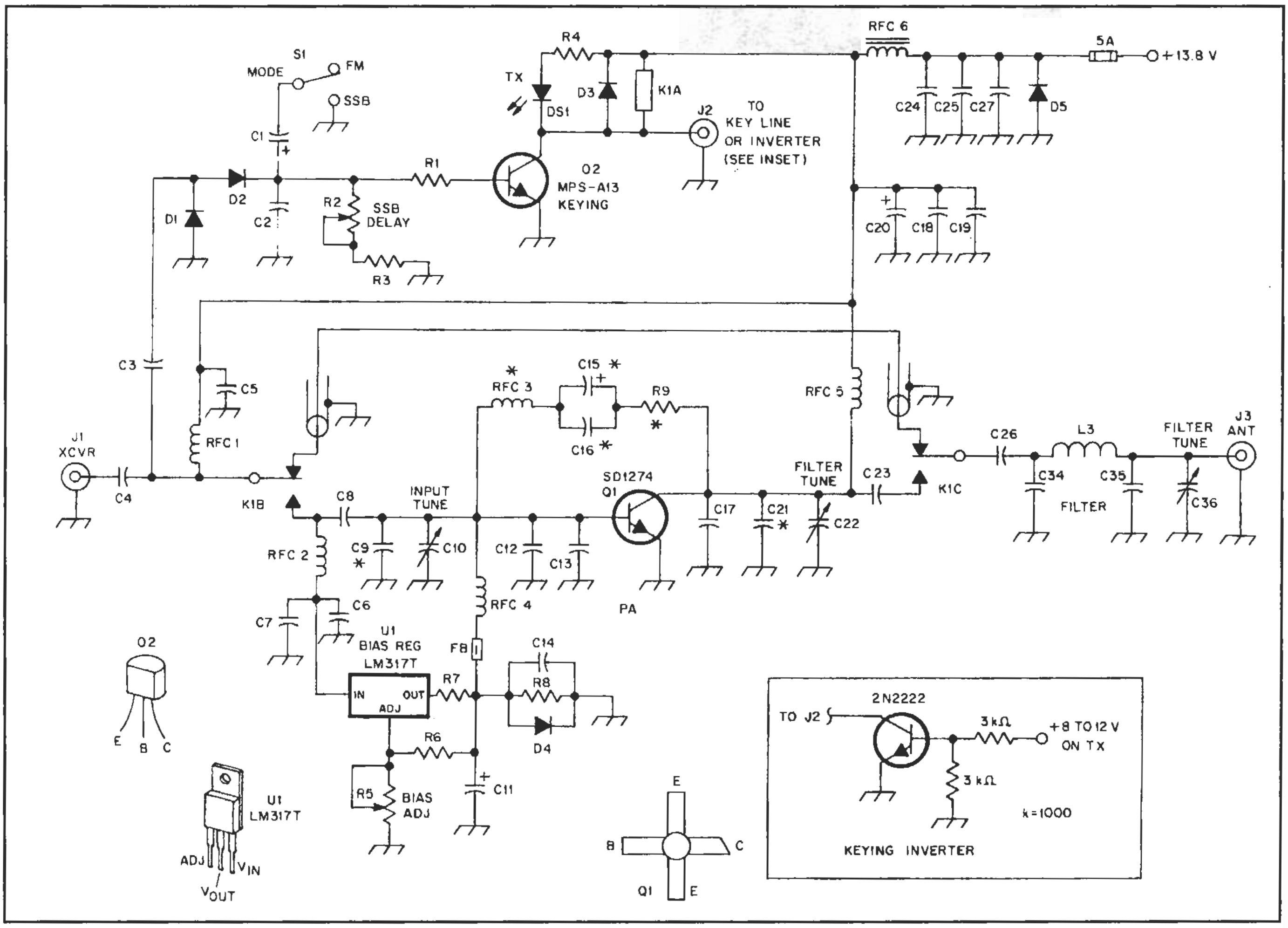
S1
MODE
FM
SSB
C1
R4
TX
DS1
D3
K1A
J2
TO KEY LINE OR INVERTER (SEE INSET)
RFC 6
5A
+13.8 V
C24
C25
C27
D5
R1
Q2 MPS-A13 KEYING
D2
D1
C2
R2
SSB DELAY
R3
C20
C18
C19
C3
C5
RFC 1
J1 XCVR
C4
K1B
C8
INPUT TUNE
RFC 3
C15
C16
R9
RFC 5
SD1274 Q1
FILTER TUNE
C23
K1C
C26
L3
FILTER TUNE
J3 ANT
C34
C35
FILTER
C36
RFC 2
C9
C10
C12
C13
C17
C21
C22
RFC 4
PA
C7
C6
U1 BIAS REG LM317T
IN
OUT
ADJ
FB
C14
R7
R8
R6
D4
R5
BIAS ADJ
C11
Q2
E B C
U1 LM317T
ADJ
VIN
VOUT
E
B
C
Q1
E
2N2222
TO J2
3 kΩ
+8 TO 12 V ON TX
3 kΩ
k=1000
KEYING INVERTER

Fig 1—Schematic of the 25-W VHF amplifiers. The 5-A fuse is included in the power lead (external to the amplifier). Be extremely careful when mounting the power transistor—see text for precautions. In the parts list below, values for the 220-MHz version are given in parentheses where they differ from those in the 2-meter version. Asterisks indicate parts not used in the 220-MHz version.

C1—15-µF, 6-V dc.
C2—470-pF disc.
C3—2-pF silver mica.
C4, C23, C26—270- to 470-pF silver mica or Unelco.
C8—270- to 470-pF (100 pF) silver mica or Unelco.
C5, C7, C14, C16, C19, C24—0.01-µF disc.
C6—0.22-µF disc.
C9—43-pF silver mica or Unelco.
C10—Arco 404—4 to 60 pF (Arco 403—4 to 40 pF).
C11, C15—1-µF, 15-V tantalum.
C12—180-pF (120-pF) Unelco.
C13, C18—220-pF Unelco.
C17—68-pF (51-pF) Unelco.
C20—10-µF, 35-V electrolytic.
C21*—33 pF silver mica or Unelco.
C22—Arco 404—4 to 60 pF (Arco 402—1.5 to 20 pF)
C25—0.1-µF disc.
C27—0.001-µF disc.
C34, C35—25-pF (18-pF) Unelco.
C36—10-pF trimmer.
D1, D2—1N4148 or 1N914.
D3, D4—1 N4001 or equiv.
D5—ECG 581 or equiv.
DS1—LED.
FB—Ferrite bead.
K1—Omron LZN2O3-UA-DC12 DPDT relay.
L3—3 turns (1 turn) no. 18, closewound, 0.2-in. (¼-in.) ID.
Q1—Thomson/Mostek SD1 274.
Q2—MPS-A13 NPN Darlington.
R1, R4—1.5-kΩ, ¼ W.
R2—500-kΩ, 10-turn potentiometer.
R3—10-kΩ, ¼ W.
R5—100-Ω miniature potentiometer.
R6—270-Ω, ¼ W.
R1, R8—10-Ω, ½ W.
R9*—15-Ω, ¼ W.
RFC1, RFC2, RFC4—0.47-µH molded choke.
RFC3*—0.15-µH molded choke.
RFC5—6 turns (5 turns) no. 16 enam, ¼-in. ID.
RFC6—VK200/4B ferrite choke.
S1—Miniature SPDT toggle.
U1—LM317T voltage regulator

Miscellaneous

2 BNC or N connectors (see text).
4 × 4 × 1½-in, heat sink.
5-A fuse and in-line holder.
2 × 1-in, scrap of thin sheet brass.

power (about 150 mW), I can tune the amplifier to give 4W output. Efficiency is about the same as the 2-meter version.

The saturated power output for both versions is more than 35 W in FM operation. The maximum FM power input is 4 W. Minimum power input for proper RF sensed keying operation is 100 mW.

The Power-Amplifier Circuit

See Fig 1. K1 is similar to relays used in VCR RF circuits. I have switched up to 40 W at 2 meters and 35 W at 220 MHz with these relays. They present a good match to 50 Ω. The loss through the amplifiers (when not in use) resulting from these relays is less than 1 dB. This is typical for 2-meter and 220-MHz amplifiers.

In addition to switching the transmitted and received signals, the relays also switch the dc supply voltage. In the transmit mode, K1 switches 13.8 V dc to the input of the bias regulator, U1. The regulator IC and its associated resistors (R7, R8) and bias diode (D5) supply a stiffer bias voltage to the base of Q1 than the more-common voltage-divider bias networks. This low-impedance bias source keeps the bias voltage constant over the range of RF drive levels. With a voltage-divider bias network, bias voltage can be upset by the base-emitter rectification (self bias) developed by the RF driving signal. In such a case, over-driving the amplifier causes the base bias to decrease, resulting in non-linear amplification. This effect is limited with the regulator-type bias

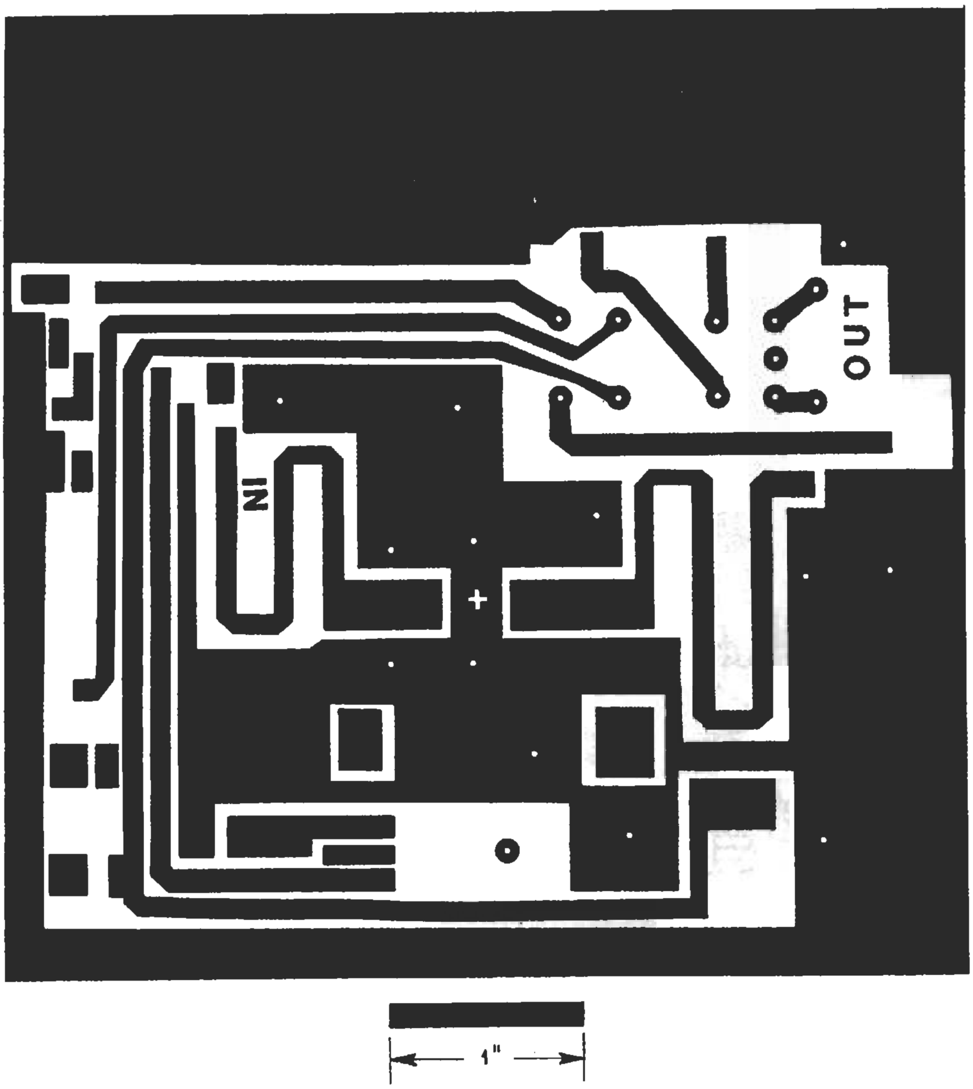

Fig 2—Main PC-board layout for the 2-meter and 220-MHz amplifiers. Artwork is full scale. Because this is a microstripline design, do not change the geometry or routing of board traces. The board must be drilled for rivets, mounting hardware and transistor mounting at the locations marked. The transistor-mounting hole is 0.38-in. diam and should be drilled at the + near the center of the PC board. All other holes (except for the relay-mounting holes near the output microstripline) are for rivets and no. 4-40 mounting hardware.

circuit used in these amplifiers.

TR switching is accomplished by an RF sensing circuit. A small amount of RF is sampled by C3 and rectified by Dl and D2, which turns on keying transistor Q2. Q2 pulls in K1, which switches the amplifier into the line.

When S1 is in the SSB position, a short drop-out delay is added in the RF-sensed-keying circuit to keep the amplifier keyed during brief pauses in speech. This delay is adjustable by varying R3. In the FM position, no delay is needed. S1 does not change the class of operation of Q1—it merely switches in the delay circuit.

The switching relays are wired so the amplifier can remain in-line at all times, with or without the supply voltage connected. Applying the supply voltage allows you to use the amplifier and preamp. Without the supply voltage connected, all transmitted and received signals pass through the amplifier.

The spectral output of these amplifiers is quite good, but to ensure clean signals, I've added filters to the output of each unit. Although there isn't much extra room for the filter components inside the amplifier cabinet, you can mount them on the back cover or in another small enclosure. The filters, shown in the title photo, are very simple to build and tune.

Construction

Each amplifier consists of two PC boards: the power-amplifier board and the preamp board (if used). After the PC boards are etched, holes are drilled with a no. 50 bit for installation of tinned grounding rivets at all RF and dc grounds (see Fig 2).[2] Install the rivets on the board as follows: After inserting the rivets, flare the inserted end with an awl. Next, flatten the rivet by tapping it lightly with a hammer, using an anvil or other solid surface under the PC board as a support. Solder the rivets on both sides of the PC board. Alternatively, pieces of wire can be soldered through the board—but because the wires will not be flush with the board, mounting the heat sink may be difficult.

Drill the holes for Q1 and the board-mounting screws as indicated in Fig 2. No holes are needed for component mounting because all parts are mounted on the trace side of the PC board. All the components except the preamp board are mounted next. (The preamp board is mounted after the amplifier has been tested, to prevent possible damage to the preamplifier.) Connect a piece of miniature 50-Ω Teflon® or RG-58 coaxial cable between K1 and the amplifier-input microstripline (see Fig 3).

After drilling or milling holes for Q1 in a suitable heat sink (see Fig 4), tap the mounting holes in the heat sink for no. 4-40 hardware, and mount the PC board to the heat sink. Trim the four leads of Q1 to about half their original length to make mounting easier.

Q1 can be mounted in one of several ways. The distance between the underside of the leads of Q1 and the heat-sink mounting area of Q1 is larger than the thickness of 1/16-in. G-10 board material. Because of this, the heat sink must be milled (see Fig 4), or small pieces of copper or brass must be soldered under each lead of Q1 to make up the difference in height. Alternatively, a second piece of G-10 material can be cut to the same size as the amplifier PC board

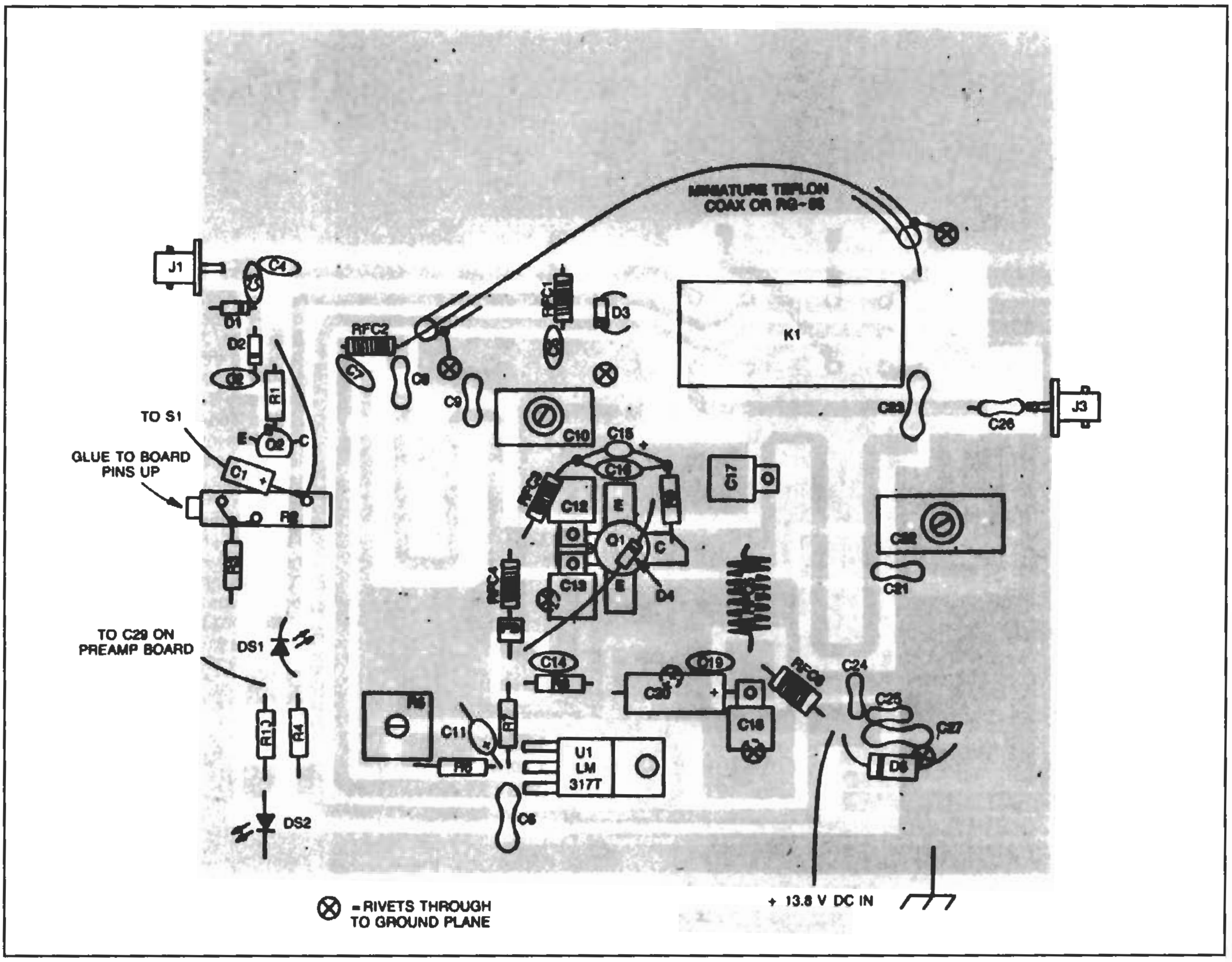

Fig 3—Parts-placement diagram for the VHF amplifiers. All parts are mounted *on the foil-trace side* of the PC board. The power-transistor mounting stud passes through the board and into the heat sink. D5 is epoxied to the ceramic cap of Q1. R2 (10-turn potentiometer) is epoxied to the PC board as indicated, with the pins facing up.0

and used as a spacer between the PC board and heat sink.

Be careful not to crack the body of Q1 where the leads of the device meet it—this can release beryllium-oxide (BeO) dust, *which is lethal.* The leads of the transistor can be bent down slightly to the PC board without affecting RF performance, but never force them in the opposite direction after the transistor has been tightened to the heat sink. Mount the transistor to the heat sink *first*, then solder its leads to the board, and *not vice versa.* Use thermally conductive compound between Q1's mounting surface and the heat sink.

Cut a piece of hobby-store sheet brass about as wide as the bias regulator IC and twice as long as the body of the device. This will be a heat sink for the bias regulator. When mounting the regulator IC, use thermally conductive compound between it and the brass sheet, and between the brass and the main PC board. After tightening the IC hardware to the PC board, bend the brass sheet in a U shape over the IC, leaving enough clearance for heat to escape between the IC and the brass heat sink. This heatsink arrangement can be seen near the bottom center of each amplifier in the lead photograph.

The Preamplifier Circuit

The high gain and low noise figure (NF) of a GaAsFET preamp are not necessary in amplifiers in this power class. I designed the preamp circuit (Fig 5) to use an inexpensive U310 FET. It has more than enough gain to overcome the losses of the amplifier switching circuits and the feed line to the amplifier. The preamp has more than 12 dB gain and noise figure of about 2 dB.

When S2 is closed, K2 closes, activating the preamplifier by applying the dc supply voltage to the drain circuit of Q3 and to the preamp ON LED, DS2. In the transmit mode, K1 switches off the supply voltage to K2, allowing K2 to drop out.

The preamp circuit is optional, and I didn't include it on the main amplifier PC board. If you're not going to build the preamp, solder foil-tape or wire jumpers between the two points on the main amplifier PC board where the preamp connects during assembly.

Building the Preamplifier

The schematic of the preamplifier circuit is shown in Fig 5. As mentioned earlier, although some component values differ between the 2-meter and 220-MHz versions, the two are essentially the same in all other respects. Components that have different values for the two versions are marked with asterisks on the schematic and the parts list. Unlike the main amplifier PC board, most preamp components are mounted through the board and soldered on the bottom side. The PC-board layout and parts-placement diagrams are shown in Figs 6 and 7, respectively.

As with most FET VHF preamplifiers, a shield is necessary between the input inductor and the active device. Solder a piece of scrap brass sheet to the ground foil between these circuit elements. (This preamp shield can be seen in the right-hand amplifier in the lead photo, just behind L1.)

After building the preamplifier, tune it

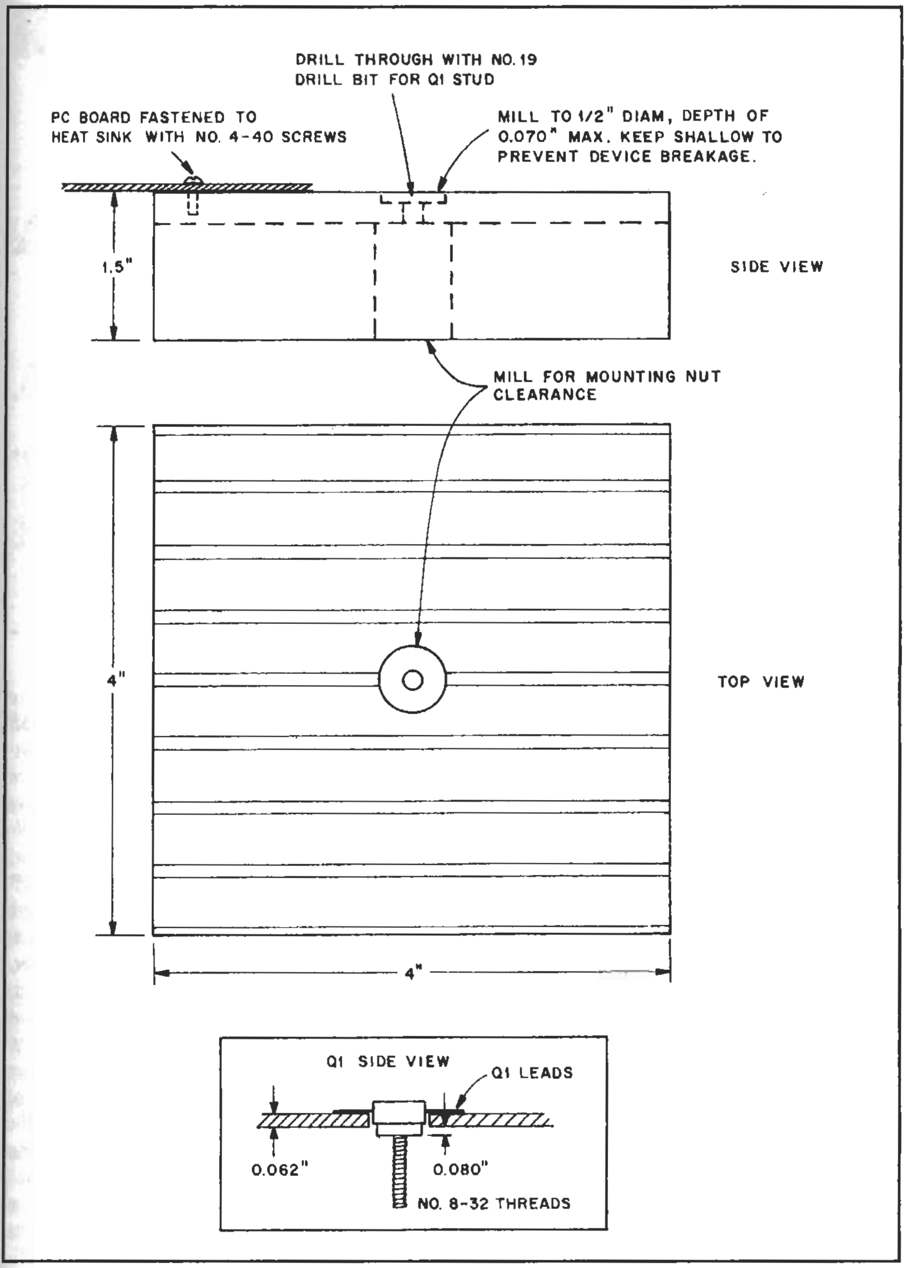

Fig 4—Heat-sink drilling dimensions for the VHF amplifiers. Overall heat-sink size isn't critical, but should be close to the dimensions shown. As an alternative to milling the heat sink to accept Q1, a piece of $^1/_{16}$-in. double-sided PC-board material can be used as a shim. Be careful not to stress the transistor lead-to-body connections during mounting.

for maximum gain with a signal generator or an on-the-air signal. If one is available to you, use a noise-figure meter and tune the preamp for best NF.

Building the Output Filters

I built the output filters for each amplifier on scraps of PC-board material. The construction is shown in the title photo. After building the filters, it's a good idea to use some silicone sealant on the coil (especially in the 2-meter filter) to hold the coil turns in place. Install the filter in a suitable enclosure, or in the amplifier cabinet.

To tune the filters, you'll need a receiver capable of receiving the second harmonic of the fundamental (288 MHz for the 2-meter filter, and 440 MHz for the 220-MHz filter), and a signal generator or an on-the-air signal. Connect the filter between the antenna (or signal generator) and the receiver, and tune C36 for minimum second-harmonic signal level.

Enclosures

I used the main amplifier PC board as the top cover for the unit (with the heat sink mounted on it). I made the sides and bottom cover of the enclosure from PC-board material. The side walls are soldered to the amplifier board, and the bottom cover is attached to the side walls with small brackets (also made of PC-board material) soldered to the bottom cover. Alternatively, the amplifier can be mounted in an aluminum enclosure.

Although I used BNC connectors for RF input and output, N connectors can also be used. Make the connections from the input connector to the PC board with a piece of no. 16 wire or a $^1/_8$-inch-wide strip of brass to the input microstripline. At the output, connect dc-blocking capacitor C26 between the microstripline and the antenna connector. Check the board for proper component placement and good solder joints, and get ready to tune up the amplifier!

Amplifier Tune-Up

As with the amplifier-circuit design and description, the tune-up procedure is very similar for both. The only difference is the exciter you use. Here's how to tune either version.

Disconnect RFC4 from the base of Q1. Connect a voltmeter to the free end of RFC4. Apply 13.8 V dc to the amplifier supply voltage leads. Using the voltmeter, verify that R5 is mounted so that output of U1 increases with clockwise rotation of R5. Turn R5 fully counterclockwise (minimum U1 output voltage). Reconnect RFC4.

Set the quiescent (no-drive) current to Q1 as follows: Disconnect one end of RFC6 and connect an ammeter in series with it. Apply 13.8 V dc to the amplifier through a 5-A fuse. Turn the preamp off. Do not apply RF drive during this adjustment. Using a clip lead, ground the collector of Q2. This should actuate K1. Check the TX LED, DS1, for operation. (If K1 actuates and the TX LED does not light, the LED may be installed in reverse.) Slowly adjust R5 for an idling current (through RFC6) of 75 to 100 mA. The amplifier should be stable; instabilities are indicated by erratic variation of Q1's quiescent current as R5 is adjusted.

Disconnect the ground lead of Q2. Q1's collector current should drop to zero. If the collector current does not drop to zero, the amplifier is unstable. If all is well, remove the ammeter and reconnect RFC6. If the amplifier is unstable, check all bypass capacitors and solder connections.

Apply about 100 mW of drive to the amplifier and check that the COR and delay circuits work properly. Adjust C10 and C22 for maximum power output. Increase drive power and retune for maximum output. After final assembly, tune the amplifier for the desired frequency and power level.

Final Assembly

The preamplifier board can now be installed on the main PC board. Mount the preamp on its edge (refer to the title photo and Fig 3) and solder the input and output microstriplines in place. Solder the ground foil of the preamp board to the ground foil of the main board in a few places to support the preamp board. Connect S2 and DS2 to the preamp board with hook-up wire (see Figs 5 and 6 for connection points). Apply power and check for proper operation of S2, K2 and DS2.

The enclosure can now be painted if you wish. Use masking tape to cover the connectors and heat sink. (Mount the LEDs and switches after painting.) You may want to

Fig 5—Schematic of the preamplifier circuit. K2 is identical to K1 on the main amplifier board. S2 and DS2 are mounted on the front panel of the amplifier. In the parts list below, values for 220-MHz version are given in parentheses where they differ from those of the 2-meter version.

C28, C29, C33—0.01-µF disc.
C30—2-pF silver mica.
C31, C32—20-pF trimmer.
DS2—LED.
D6—1 N4001.
L1, L2—5 turns (4 turns) no. 16 enam, 0.3-in. ID.
K2—Omron LZN2O3-UA-DC12 DPDT relay.
Q3—U310 FET.
R10—1.5-kΩ, ¼ W.
R11—91-Ω, ¼ W.
RFC7, RFC8—0.47-µH molded choke.
S2—Miniature SPDT toggle.

Miscellaneous
1 × ½-in. strip of thin sheet brass

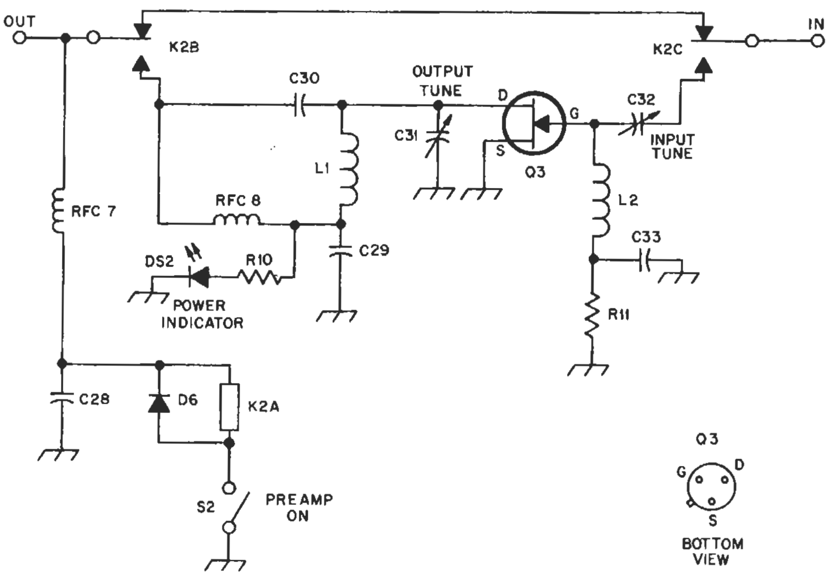

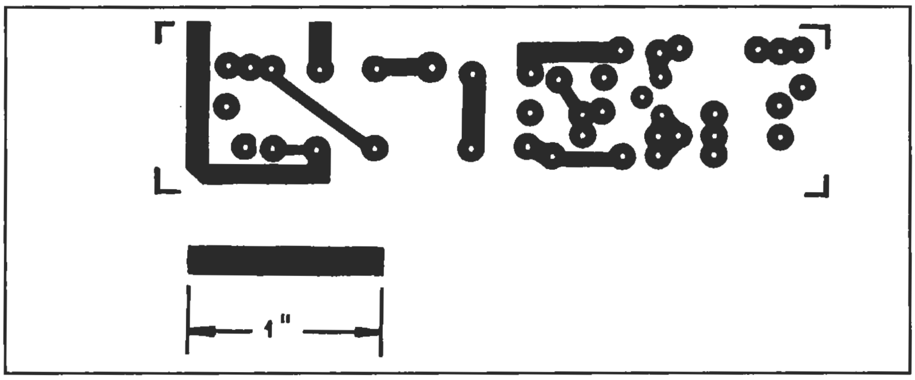

Fig 6—Full-size PC-board layout for the preamplifier circuit. Unlike the power-amplifier PC board, most of the preamplifier components are mounted on the non-trace side of the board.

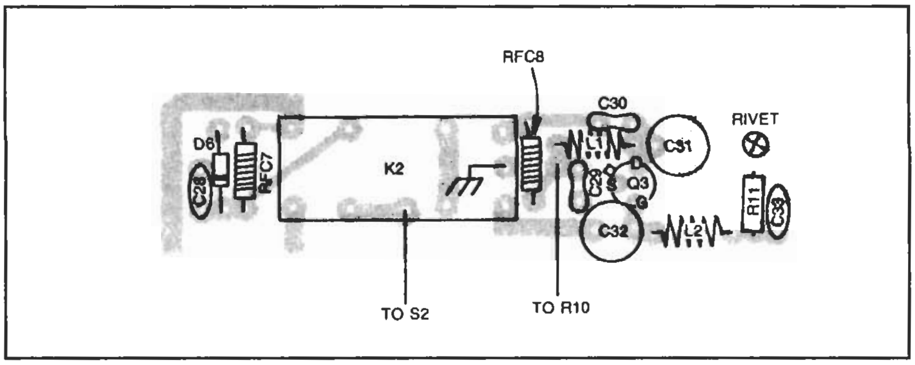

Fig 7—Parts-placement diagram for the VHF preamplifier. Mount all components on the unetched ground-plane side except for C30 and C31. Use a drill to clear the foil away from all mounting holes on the ground-plane side of the board (except for through-board ground connections). Solder all ground connections on both sides of the PC board (see text for rivet-installation instructions). Solder a brass-strip shield to the ground foil between L1 and Q3 on the component side of the PC board.

add stick-on rubber feet to the bottom cover to keep it from sliding around and being scratched.

Amplifier Operation

The 2-meter amplifier works very well with my Kenwood TH-21AT hand-held and ICOM IC-202S SSB/CW transceiver. I use one in my shack and one mobile—both at more than 25 W output. I drive the 220-MHz version of the amplifier with an ICOM IC-3AT (in its high-power mode).

Limit the amplifier power output to 25 W during linear operation. *All SSB amplifiers have a rated linear power output that should not be exceeded, even though the amplifier may be driven above that level. To keep your signal clean, do not overdrive any amplifier.* For exciters with fixed power outputs that would overdrive the amplifier, use the loss of a length of RG-58, or a discrete attenuator, between the radio and the amplifier to prevent overdrive. The gain of the preamplifier (if used) will overcome the attenuator or cable loss during receiving.

When using any amplifier or transverter in the shack, it is a good practice to "hard key" it. Hard keying is simply forcing the amplifier or transverter into the transmit mode with a switch closure or an applied voltage. To do this, run a keying line from the exciter to the amplifier. (Ground the collector of Q2 to hard key these amplifiers.) I added a phono jack to the 220-MHz amplifier to facilitate hard keying. If a positive voltage is used for keying, a 2N2222 transistor inverter can be used between the transceiver and the collector of Q2 (see the inset in Fig 1).

Summary

These amplifiers have served me well in the car and at home. After building one of these amplifiers, you'll probably find what I did: They're so handy to have around, and so easy to build, that you can't build just one!

Notes

[1] Thomson Components/Mostek Corp, Semiconductor Division, Commerce Dr, Montgomeryville, PA 18966. Thomson transistors are available through RF Gain, Ltd, 100 Merrick Rd, Rockville Center, NY 11570, tel 516-536-8868 or 800-645-2322.

[2] The following parts are available from Frontier Microwave, RD 1 Box 467, Ottsville, PA 18942: 100+ tinned rivets: $2; relays (K1, K2): 2 for $8. Prices include shipping. The ARRL and *QST* in no way warrant this offer.

By David D. Meacham, W6EMD **From *QST*, April 1984**

A High-Power 2-Meter Amplifier Using the New 3CX800A7

Tired of hearing exotic DX on 2 meters but not being able to work it? Build this amplifier, and you'll transform your pip-squeak signal into a real rockcrusher.

The 3CX800A7 triode recently announced by Varian EIMAC has a plate dissipation rating of 800 W and modest cooling requirements. Its oxide cathode gives high emission with only 20 W of heater power. With full ratings up to 350 MHz, this tube is a scaled-up version of the 8874 — a proven performer in amateur and commercial gear.

The amplifier described here is based on a design presented by Raymond F. Rinaudo, W6ZO, in January 1972 QST. Most of the changes in the new amplifier (Fig. 1) are designed to accommodate the larger size of the 3CX800A7 and its attendant higher capacitances and currents. A plate-current meter is not included because my station power supply is already well metered. However, there is room on the front panel of the amplifier for a plate-current meter if another builder wishes to add one. Other changes to Rinaudo's design are minor.

Construction

The amplifier chassis is mounted on a standard 19-inch wide, 5¼-inch-high aluminum rack panel (Bud No. SFA-1833).[1] A 5 × 13 × 3-inch aluminum chassis (Bud No. AC-422) is spaced 1¾ inches behind the panel by two aluminum end brackets. A 3½ × 4½ × 1-inch aluminum chassis (Bud No. AC-1402) houses the input circuitry and is mounted on the larger chassis between it and the front panel. The right-hand end bracket has a large lip on the rear for mounting the heater transformer and connectors. Fig. 2 shows mechanical details of the chassis and end brackets. The cabinet chosen is a lightweight aluminum unit made by TenTec (No. 19-0525).

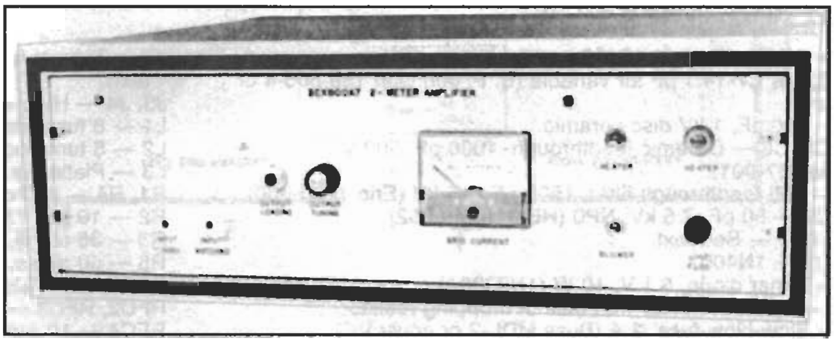

Input Circuit

In the cathode-driven configuration, the input impedance of the 3CX800A7 appears as a nominal capacitance of 26.5 pF in parallel with a resistive component that varies with operating conditions but is typically about 49 ohms. The computerdesigned input circuit of this amplifier operates at a loaded Q of about 2.1. It can be set anywhere in the 2-meter band for a VSWR of less than 1.3:1. C1 and C2 are predominantly matching and tuning controls, respectively; there is some interaction between them, however. When the input is tuned at 144.5 MHz, the input VSWR will he less than 1.4:1 from 144 to 145 MHz.

The tube socket is an EIMAC SK-1900 or a Johnson part No. 124-311-100. Its center is mounted 2 inches from the end of the input box, adjacent to the mounting bracket. The tube pins and input circuit are cooled by a small amount of air admitted to the input box from the pressurized output box. Three holes, made with a no. 50 drill, provide adequate flow. These holes are spaced in a close triangle and are located diagonally across from the variable capacitors. Air exhausts through the tuning holes. Fig. 3 shows details of the input circuitry.

The heater circuit includes two chokes, feedthrough capacitors, the filament transformer, a voltage-dropping resistor and a switch. The chokes are wound with a turn-to-turn spacing of about one-half the wire diameter. They are self-resonant (parallel resonant) just above the 2-meter band. Nominal heater voltage for the 3CX800A7 is 13.5 V. The closest available commercial transformer has a 14.0-V secondary, so R3 is used in the primary. Switching is set up so that the blower must be on before the heater can be activated. Conversely, this arrangement allows the blower to be left on after switching off the heater—a highly recommended practice.

Cathode bias is provided by a 5.1-V Zener diode, D3. Rl prevents the cathode voltage from soaring if the Zener fails. F2 will blow if excessive cathode current is drawn. R2 nearly cuts off plate current on receive. Because the grid is at dc ground, the negative supply lead must be kept above

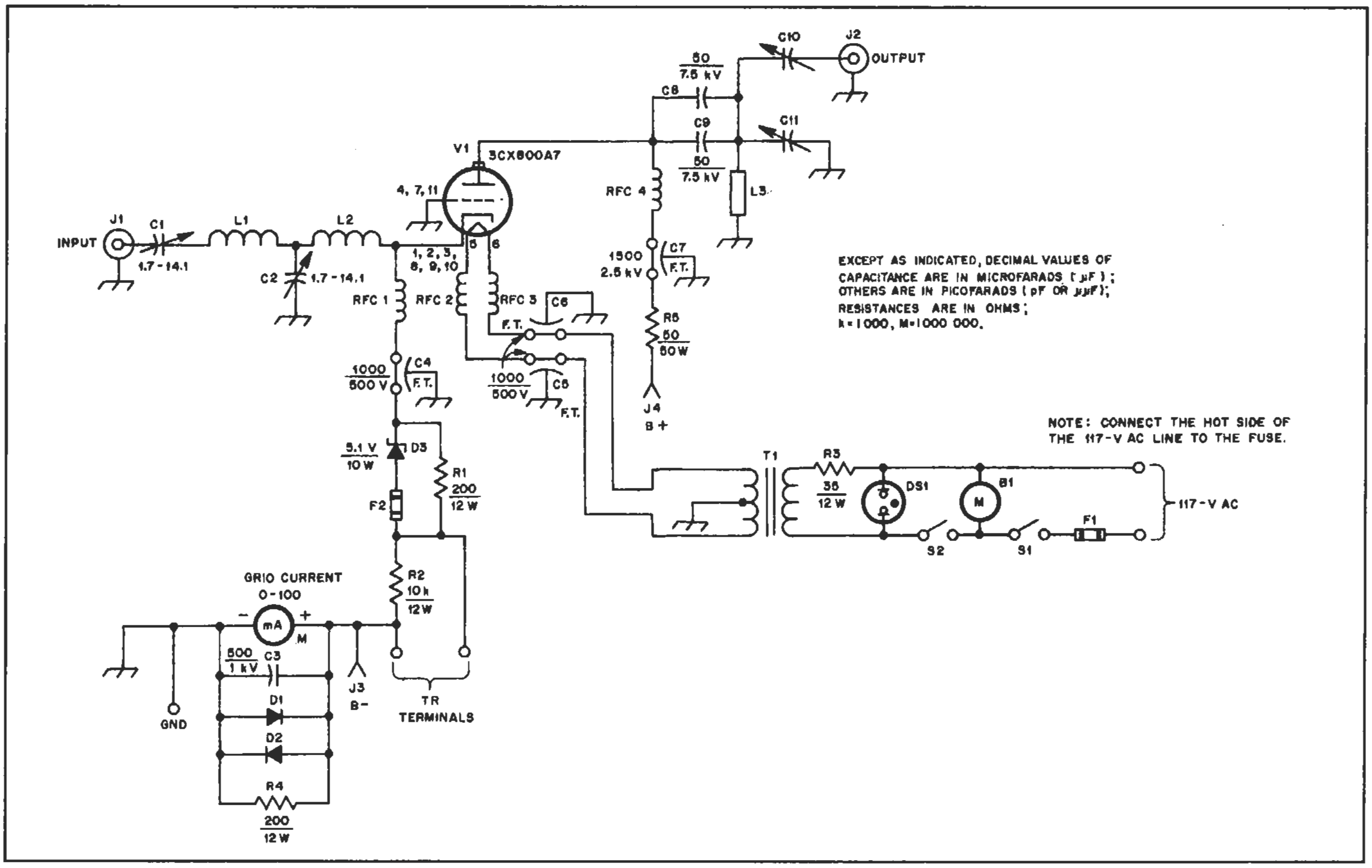
C10
J2
OUTPUT
50
7.5 kV
C8
C9
50
7.5 kV
C11
V1
3CX800A7
4, 7, 11
RFC 4
L3
J1
INPUT
C1
1.7-14.1
L1
L2
C2
1.7-14.1
1, 2, 3, 8, 9, 10
5
6
RFC 1
RFC 2
RFC 3
C6
F.T.
1500
2.5 kV
C7
F.T.
R5
50
50 W
J4
B+
1000
500 V
C4
F.T.
1000
500 V
C5
F.T.
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.
NOTE: CONNECT THE HOT SIDE OF THE 117-V AC LINE TO THE FUSE.
5.1 V
10 W
D3
R1
200
12 W
F2
T1
R3
35
12 W
DS1
B1
M
117-V AC
F1
S2
S1
GRID CURRENT
0-100
mA
M
R2
10 k
12 W
500
1 kV
C3
D1
D2
R4
200
12 W
J3
B-
TR
TERMINALS
GND

Fig. 1 — Schematic diagram of the 3CX800A7 amplifier.

B1 — Blower (Dayton 4C012 or equiv.); see text.
C1, C2 — 1.7-14.1 pF air variable (E. F. Johnson 189-505-4 or equiv.).
C3 — 500.pF, 1.kV disc ceramic.
C4, C5, C6 — Ceramic feedthrough, 1000 pF, 500 V (Erie 357-001).
C7 — EMI feedthrough filter, 1500 pF, 2.5 kV (Erie 1280-060).
C8, C9 — 50 pF, 7.5 kV, NP0 (HEC HT-50 852).
C10, C11 — See text.
D1, D2 — 1N4001.
D3 — Zener diode, 5.1 V, 10 W (1N3996A).
DS1 — NE51 in holder with built-in dropping resistor.
F1 — Slow-blow fuse, 2 A (Buss MDL-2 or equlv.).
F2 — 1.5 A fuse (Buss AGC 1$^{1}/_{2}$ or equiv.).
J1 — BNC bulkhead feedthrough connector (UG-492 A/U).
J2 — Type N female connector (part of C10 assembly — see text).
J3, J4 — High-voltage connector (Millen 37001 or equiv.).
L1 — 8 turns no. 16 tinned wire, $^{3}/_{8}$-in ID, $^{11}/_{16}$ in long.
L2 — 6 turns no. 16 tinned wire, $^{3}/_{8}$-in ID, $^{11}/_{16}$ in long.
L3 — Plate line, 1$^{5}/_{16}$ in wide, 8$^{3}/_{16}$ in long. See text and Fig. 4.
R1, R4 — 200 ohms, 12 W.
R2 — 10 kΩ, 12 W.
R3 — 35 ohms, 12 W.
R5 — 50 ohms, 50 W.
RFC1 — 1.0 µH choke, 300 mA (Miller 4602).
RFC2, RFC3 — 11 turns no. 20 HF wire on $^{3}/_{8}$-in Noryl rod.
RFC4 — 10 turns no. 16 tinned wire, $^{1}/_{2}$-in ID, 1$^{3}/_{16}$ in long.
T1 — Filament transformer, 14.0-V, 2-A secondary (Triad F-215X or equiv.).

ground for grid-current metering by M1. R4 keeps the negative side of the plate supply from rising if M1, D1 and D2 all open up. D1, D2 and C3 protect the meter from transients and RF voltages.

Output Circuit

The output tank circuit is a silver-plated quarter-wave strip line (Fig. 4) foreshortened by the tube, loading and tuning capacitances at, and near, its open end. It operates at a loaded Q of approximately 20. The silver-plated anode collet (Fig. 4) is made of 0.062-inch-thick brass sheet with Tech-Etch 134B finger stock soldered on the inside. It is supported by two Teflon® standoffs $^{1}/_{2}$-inch in diameter and 1 inch long. At the far end of the line, a silver-plated shorting block contacts the chassis and the strip line. This block (Fig. 4) is made from $^{7}/_{16}$-inch-thick brass 1$^{15}/_{16}$ inches wide and 1 inch high. EIMAC CF-800 finger stock is screwed and then soldered on both top and bottom. The chassis and stripline are both slotted to allow $^{5}/_{16}$ inch of shorting-block travel to set the tuning range of the capacitive-tuning paddle. A 1-inch-long Teflon standoff supports the center of the line.

The tuning (C11) and loading (C10) paddles are 1$^{3}/_{16}$-inch-diameter discs made from silver-plated, $^{1}/_{16}$-inch-thick brass. They are spaced 1$^{3}/_{8}$ inches center-to-center. The output loading paddle is nearest the tube, and its center is 1$^{13}/_{16}$ inches from the tube cooler surface. Spacing between the paddle and line during operation is about 0.135 inch. The output tuning boss, EIMAC part No. 720362, is tapped for a $^{1}/_{4}$-28 threaded rod. The output loading control and output connector are two separate EIMAC assemblies combined into one unit. They are available from EIMAC as Support Assembly No. 720361 and Sliding Probe Assembly No. 720407. For this amplifier, the outer conductor was lengthened to reach the front panel. The grid collet is also available from EIMAC as part No. 720359.

In addition to RFC4 and feedthrough capacitor C7, an energy-absorbing resistor (R5) is wired in series with the dc plate supply. This resistor will protect the tube and power supply in the event of a highvoltage arc. It provides necessary protection while dissipating only 12.5 W at 500-mA dc plate current.

Cooling

The blower specified provides a measured 25 CFM airflow through the output box and tube cooler at 0.42 inch of water-column static pressure. This amount of cooling is sufficient for 800 W of plate dissipation at sea level with inlet air temperatures up to 35° C. It is adequate for the same dissipation at 5000 feet of altitude with inlet-air temperatures up to 25° C.

Not evident from the photographs is the care taken to separate outgoing hot air from incoming cool air. This is accomplished by the addition of a simple dividing wall inside the cabinet running from the rear chassis cover of the output box to the rear of the cabinet. The material used is rigid fiberglass insulation — the kind you cut with a saw. It provides a measure of blower-noise suppression in addition to assuring cool inlet air.

Hot air leaving the anode cooler is directed to a homemade "honeycomb" RFI filter through a chimney made from rolled-up Teflon sheet. The filter is made of brass tubing and sheet with the honeycomb material soldered in place. See Fig. 5 for construction details. This filter acts as a waveguide-beyond-cutoff having high attenuation at 144 MHz.

The equation for calculating attenuation by this type of filter is

$$A_a = 32\ D/d \qquad \text{(Eq. 1)}$$

where A_a = aperture attenuation (dB), D = length of pipe, and d = inside diameter of pipe. In this case, the "pipe" is each cell of the honeycomb. The basic material for the honeycomb is cadmium-plated brass heat-radiator core.[2] Each hexagonal-shaped cell has a width between flat sides (diameter) of 0.100 inch. The material is 1/2-inch thick. Plugging these values into the equation yields an A_a of 160 dB—more than enough! From a practical standpoint, it is usually sufficient to make cell length at least three times the nominal cell diameter.

Tune-up

Initial work may be done with a dip meter, particularly on the output circuit. With the tube in place, but with no voltages applied, the shorting block on the end of the plate line should be set to give a paddle-tuning range that straddles the desired operating frequency. Bear in mind that when the tube is "hot," the resonant frequency will be somewhat lower than it is without electron flow.

The input-tuning capacitor can be dip-meter resonated with the matching capacitor set at one-half mesh for a start. Further work here must be done "hot" with a VSWR measuring device on the input.

Set the initial spacing between the output paddle and the plate line to 1/8-inch. Connect the output to a 50-ohm dummy load capable of handling at least 700W at 144 MHz through an accurate VHF wattmeter, such as a Bird model 43. Connect a driver capable of delivering about 20 W to the input through a VSWR-measuring device.

The heater of the 3CX800A7 should be run for at least three minutes before applying plate voltage. After the warm-up, short the TR terminals and apply about 1000 V to the plate. This should result in a small amount of idling plate current. Next, apply enough drive to produce a rise in plate current, and adjust the plate tuning for a peak in output power. Now tune and match the input circuit for minimum input VSWR. Next, adjust the output loading paddle for maximum power output while keeping the plate current dipped with the plate tuning paddle. Apply full plate voltage and higher drive power, and then repeak the output tuning and loading controls. Touch up the input tuning and matching, and the amplifier is ready for service.

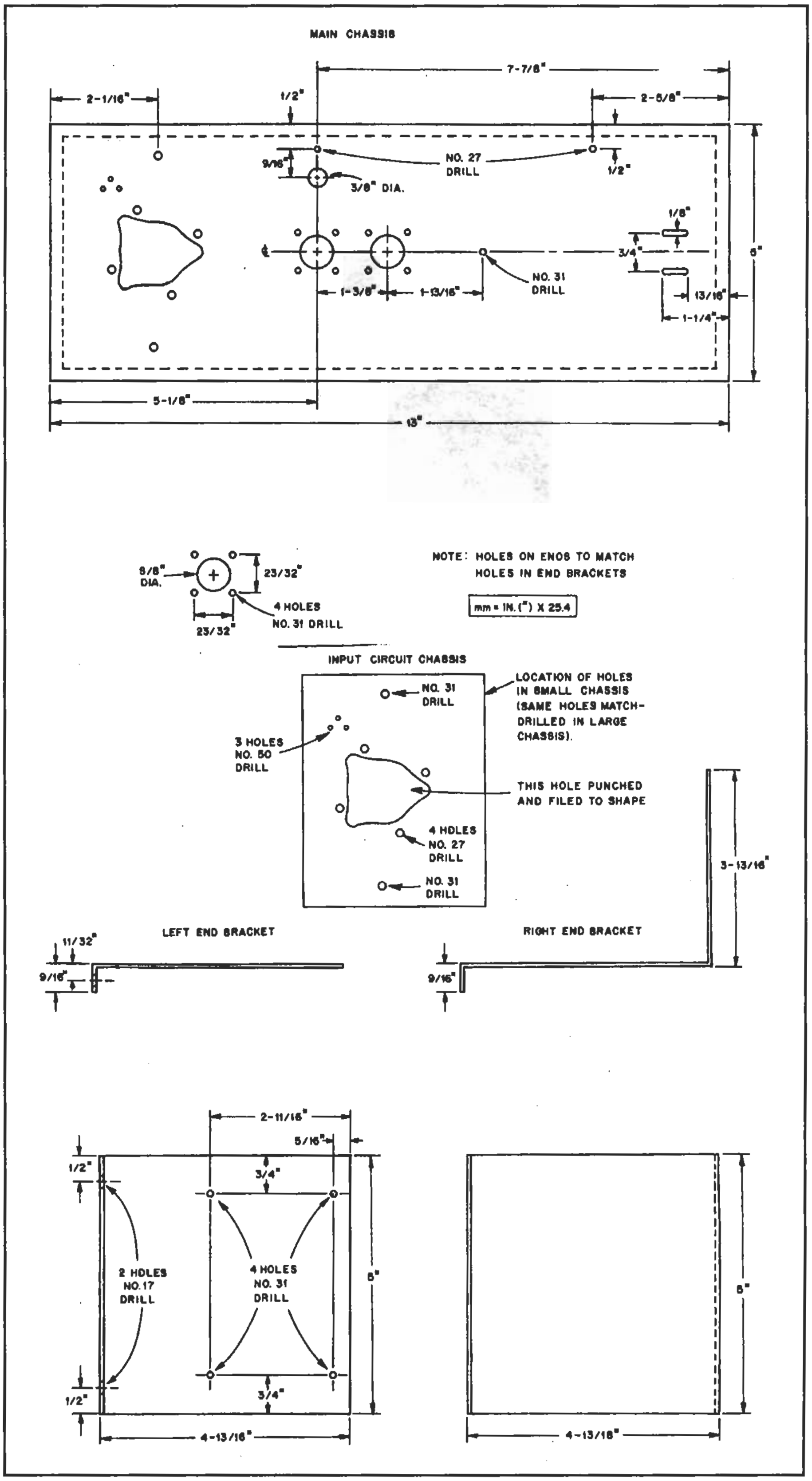
MAIN CHASSIS
7-7/8"
2-1/16"
1/2"
2-5/8"
9/16"
NO. 27 DRILL
1/2"
3/8" DIA.
1/8"
3/4"
NO. 31 DRILL
1-3/8"
1-13/16"
13/16"
1-1/4"
6"
5-1/8"
13"
5/8" DIA.
23/32"
4 HOLES NO. 31 DRILL
23/32"
NOTE: HOLES ON ENDS TO MATCH HOLES IN END BRACKETS
mm = IN. (") X 25.4
INPUT CIRCUIT CHASSIS
NO. 31 DRILL
LOCATION OF HOLES IN SMALL CHASSIS (SAME HOLES MATCH-DRILLED IN LARGE CHASSIS).
3 HOLES NO. 50 DRILL
THIS HOLE PUNCHED AND FILED TO SHAPE
4 HOLES NO. 27 DRILL
NO. 31 DRILL
3-13/16"
LEFT END BRACKET
RIGHT END BRACKET
11/32"
9/16"
9/16"
2-11/16"
5/16"
1/2"
3/4"
2 HOLES NO. 17 DRILL
4 HOLES NO. 31 DRILL
6"
3/4"
1/2"
4-13/16"
6"
4-13/16"

Fig 2—Mechanical details of the chassis and end brackets. The end brackets are made from 0.032-in aluminum, 5052 alloy or softer.

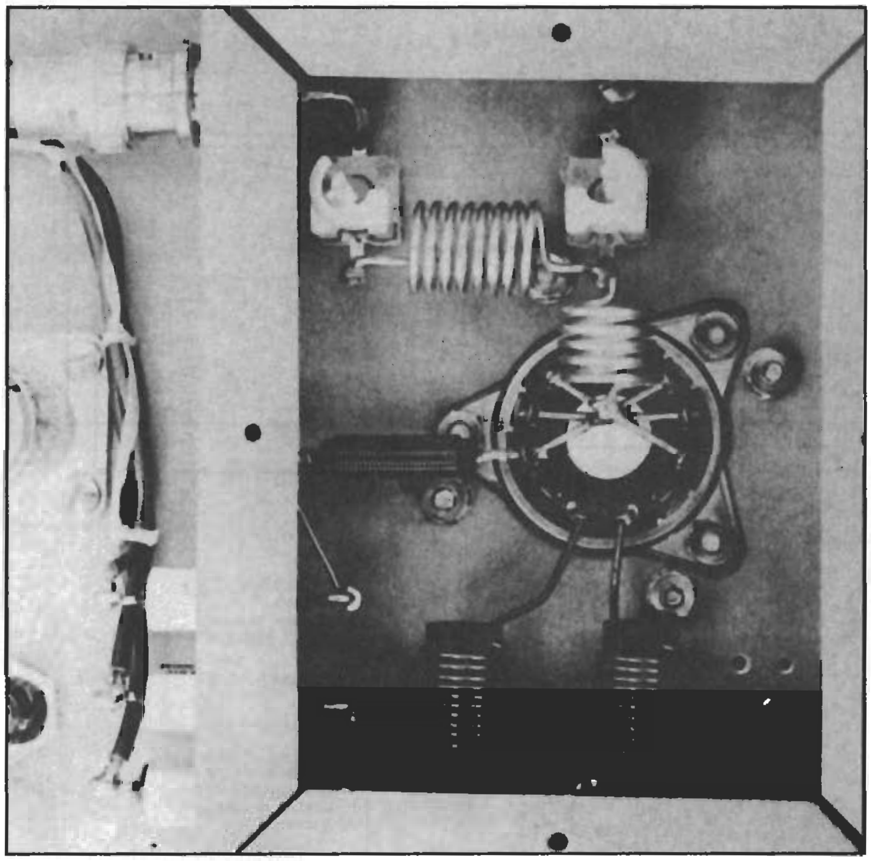

Fig 3—Close-up of the input circuitry and tube-socket wiring. Keep all leads as short as possible.

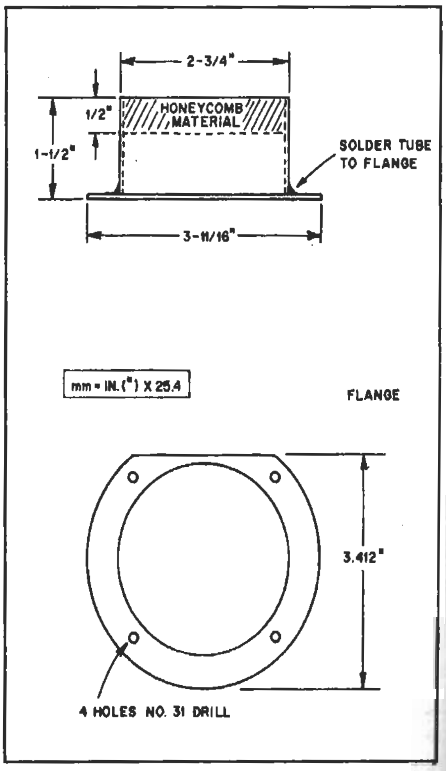
2-3/4"
1/2"
HONEYCOMB MATERIAL
1-1/2"
SOLDER TUBE TO FLANGE
3-11/16"
mm = IN.(") X 25.4
FLANGE
3.412"
4 HOLES NO. 31 DRILL

Fig 5—Mechanical details of the exhaust air RFI filter. The honeycomb material is soldered inside a 2¾-in OD, 0.065-in wall brass tube. The flange, which bolts to the chassis, is made from 0.125-in-thick brass sheet.

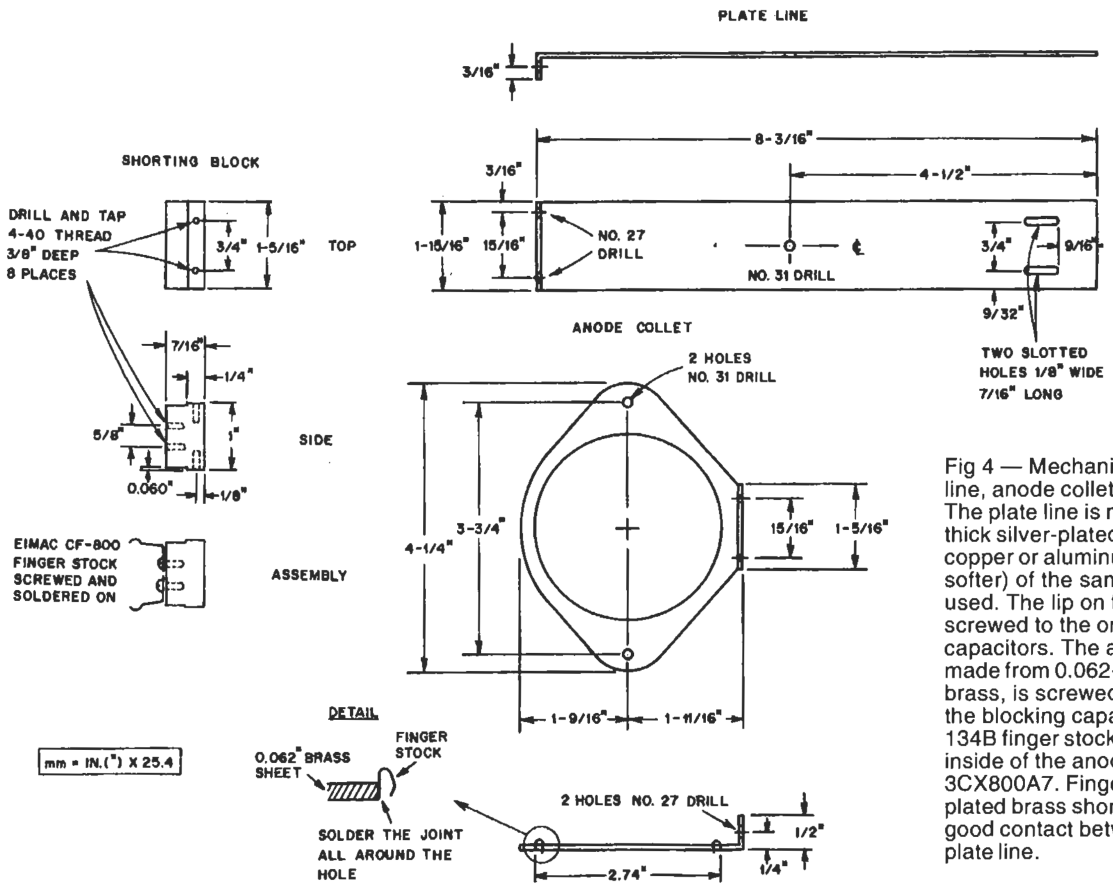
PLATE LINE
3/16"
8-3/16"
4-1/2"
3/16"
1-15/16"
15/16"
NO. 27 DRILL
NO. 31 DRILL
3/4"
9/16"
9/32"
TWO SLOTTED HOLES 1/8" WIDE 7/16" LONG
SHORTING BLOCK
DRILL AND TAP 4-40 THREAD 3/8" DEEP 8 PLACES
3/4"
1-5/16"
TOP
7/16"
1/4"
5/8"
1"
SIDE
0.060"
1/8"
EIMAC CF-800 FINGER STOCK SCREWED AND SOLDERED ON
ASSEMBLY
ANODE COLLET
2 HOLES NO. 31 DRILL
4-1/4"
3-3/4"
15/16"
1-5/16"
1-9/16"
1-11/16"
DETAIL
0.062" BRASS SHEET
FINGER STOCK
SOLDER THE JOINT ALL AROUND THE HOLE
2 HOLES NO. 27 DRILL
1/2"
1/4"
2.74"
mm = IN.(") X 25.4

Fig 4 — Mechanical details of the plate line, anode collet and shorting block. The plate line is made from 0.032-in-thick silver-plated brass, although copper or aluminum (5052 alloy or softer) of the same thickness may be used. The lip on the plate line is screwed to the one end of the blocking capacitors. The anode collet, which is made from 0.062-in-thick silver-plated brass, is screwed to the other ends of the blocking capacitors. Tech-Etch no. 134B finger stock soldered around the inside of the anode collet grips the 3CX800A7. Finger stock on the silver-plated brass shorting block ensures good contact between the chassis and plate line.

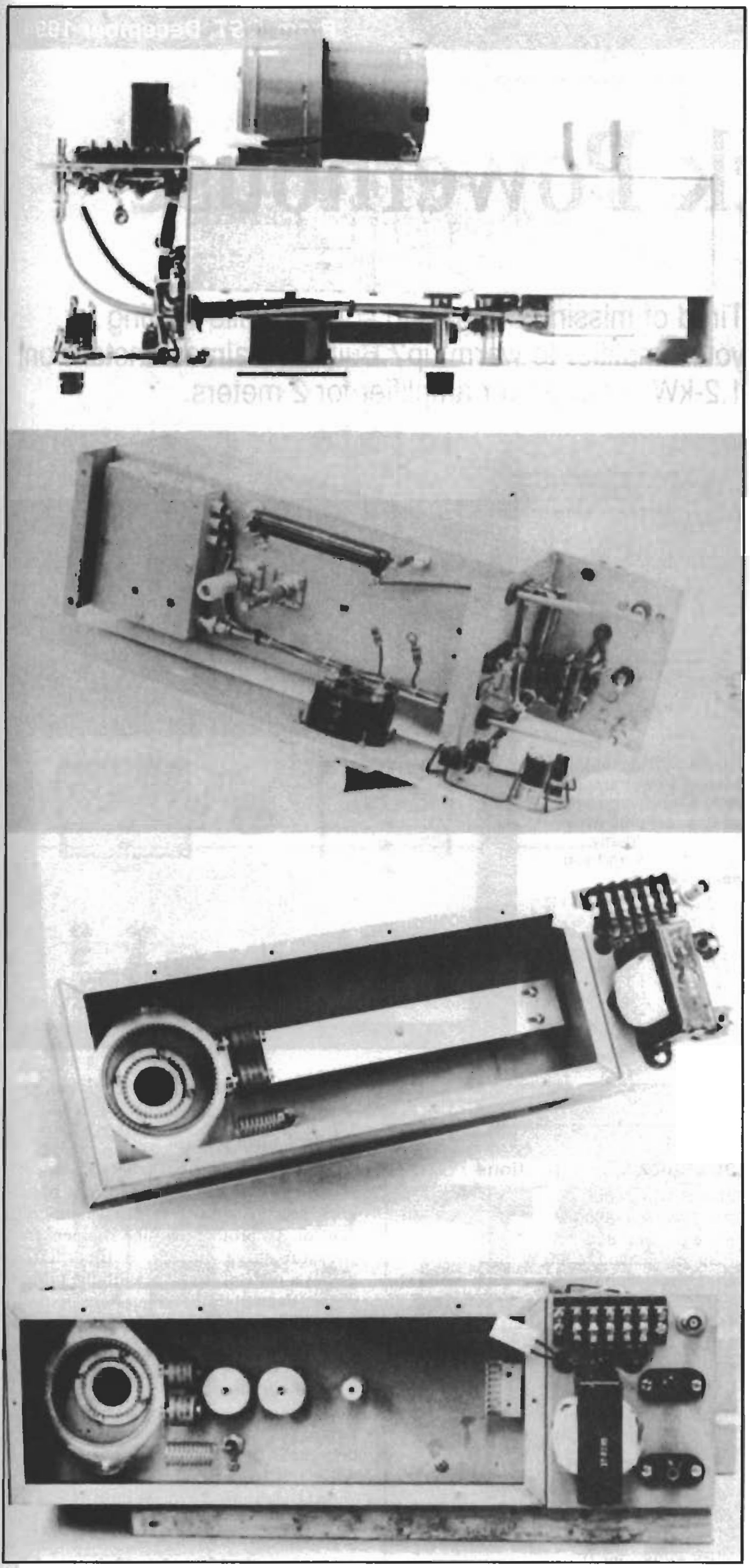

Fig 6—Various views of the completed 3CX800A7 amplifier.

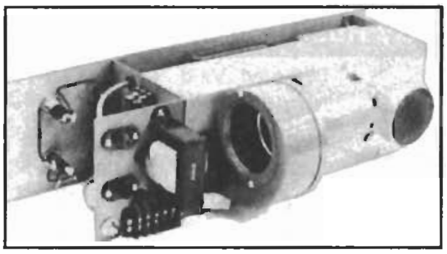

Fig 7—Rear view of the amplifier with the plate-compartment cover in place. Note the honeycomb RFI filter at the hot-air exhaust.

Table 1
Operating Conditions for the 3CX800A7 Amplifier

Plate supply voltage	2200 V
Zero-signal plate current*	65 mA
Single-tone plate current	500 mA
Grid bias (Zener bias)	–5.1 V
Single-tone grid current*	40 mA
Driving power	18.5 W
Output power	707 W
Gain	15.8dB
Efficiency**	64%

*Values may vary considerably from tube to tube.

**Actual tube efficiency is about one percent higher because of power loss in the 50-ohm series resistor in the plate lead (R5).

A few words of caution are in order. Remember that the heater voltage must never be applied without the blower running, and that the heater must warm up at least three minutes before applying plate voltage. Never exceed 60-mA dc grid current, even during tune-up. Also, because of the relatively low grid dissipation of the 3CX800A7, RF drive must never be applied unless plate voltage is applied to the tube and a suitable load is connected to the output. Following these simple rules will substantially increase tube life and amplifier reliability.

Typical Operating Conditions

With Zener-diode bias, the 3CX800A7 is best operated in Class AB2 for linear service. The data in Table 1 represent measured performance in linear service at 144 MHz. Complete data sheets are available from EIMAC. This amplifier is easily capable of conservative operation at 700-W output. Two of these tubes will run at the 1500-W output legal limit.

This amplifier uses straightforward construction techniques and is easily duplicated. Any builder will be rewarded with a great deal of satisfaction and a reliable amplifier.

Notes

[1] mm = in x 25.4; m = ft x 0.3048.

[2] One source for the honeycomb radiator core material is Technit West Division, 320 North Nopal St., Santa Barbara, CA 93103.

By Russ Miller, N7ART **From *QST*, December 1994**

The Quick Powerhouse

Tired of missing a new grid square while waiting for your amplifier to warm up? Build this almost-instant-on, 1.2-kW linear power amplifier for 2 meters.

It's that typical, early summer afternoon, when, out of the blue comes co-channel interference to the low-band (channels 2 through 6) TV channels. Time to turn on the 2-meter receiver. Whoops! There's a sporadic-E opening, and stations are booming in from over 1000 miles away. Quickly, I hit the switches to warm up my linear amplifier. Time marches on, and now the amp is ready—only took 3 or 4 minutes. But that juicy, far-away clatter via sporadic E is gone! Woe is me!

Having had this happen more than once, I was delighted to find that Eimac was producing a new tube with an almost-instant-warm-up filament that functions well at 2 meters—the 3CX1200Z7.[1] The Z7 is different from the 3CX1200A7 by virtue of its external grid ring, redesigned anode assembly and a 6.3-V ac filament. One advantage to the 3CX1200Z7 is the wide range of plate voltages that can be used, from 2000 to 5500. This amplifier looks much like the easily duplicated W6PO design,[2] except for the plate collet and the addition of some control circuitry. The plate collet is adapted from the W6PO 222-MHz design.[3] The RF deck is a compact unit, designed for table-top use (Figure 1). Table 1 gives some data on the 3CX1200Z7; Table 2 lists CW operating performance for this amplifier.

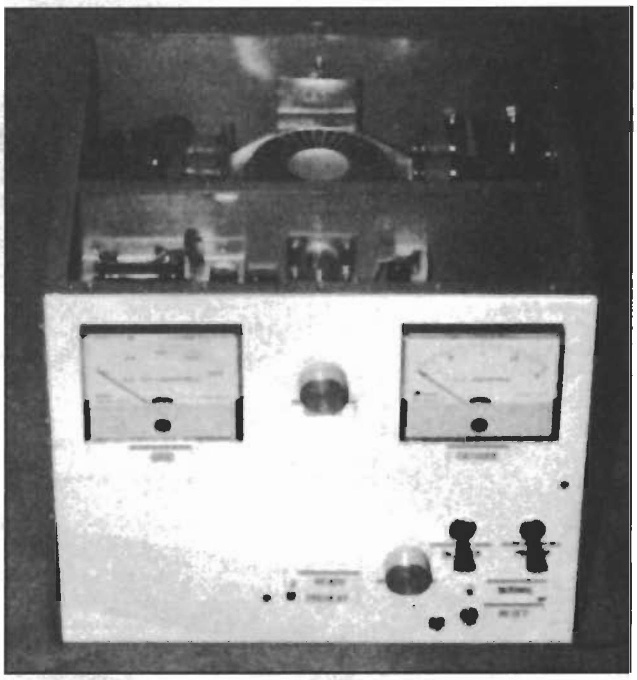

Figure 1—This table-top 2-meter power amplifier uses a quick-warm-up tube, a real plus when the band suddenly opens for DX and you want to join in.

Input Circuit

The tuned-filament T network matches the 50-Ω drive source to the filament input impedance, providing a very low input SWR. Tuning is easy and docile. Grid bias is provided by an 8.2-V, 50-W Zener diode. Cutoff bias is provided by a 10-kΩ, 25-W resistor. A relay on the control board shorts out the cutoff-bias resistor, to place the amplifier in the TRANSMIT mode.

I didn't use a tube socket. Instead, I bolted the tube directly to the top plate of the subchassis, using the four holes (drilled to clear a #6 screw) in the grid flange. Connections to the heater pins are via drilled and slotted brass rods. The input circuit is contained within a 3½×6×7¼-inch (HWD) subchassis (Figure 3).

Table 1
3CX1200Z7 Specifications

Maximum Ratings
Plate voltage: 5500 V
Plate current: 800 mA
Plate dissipation: 1200 W
Grid dissipation: 50 W

Table 2
CW Operating Data

Plate voltage:	3200 V
Plate current (operating):	750 mA
Plate current (idling):	150 mA
Grid current:	165 mA
DC Power input:	2400 W
RF Power output:	1200 W
Plate dissipation:	1200 W
Efficiency:	50%
Drive power:	85 W
Input reflected power:	1 W

Control Circuit

The control circuit (Figure 4) is a necessity. It provides grid overcurrent protection, keying control, and filament surge control. To protect the tube filament from stressful surge current, a timer circuit places a resistor in series with the primary of the filament transformer. After four seconds, the timer shorts the resistor, allowing full filament voltage to be applied. C2 and R4 establish the time delay.

Another timer inhibits keying for a total of 10 seconds, to give the internal tube temperatures a chance to stabilize. C1 and R3 determine the time constant of this timer. After 10 seconds, the amplifier can be keyed by grounding the keying line. When the amplifier is not keyed, it draws no plate current. When keyed, idle current is approximately 150 mA, and the amplifier only requires drive to produce output. A

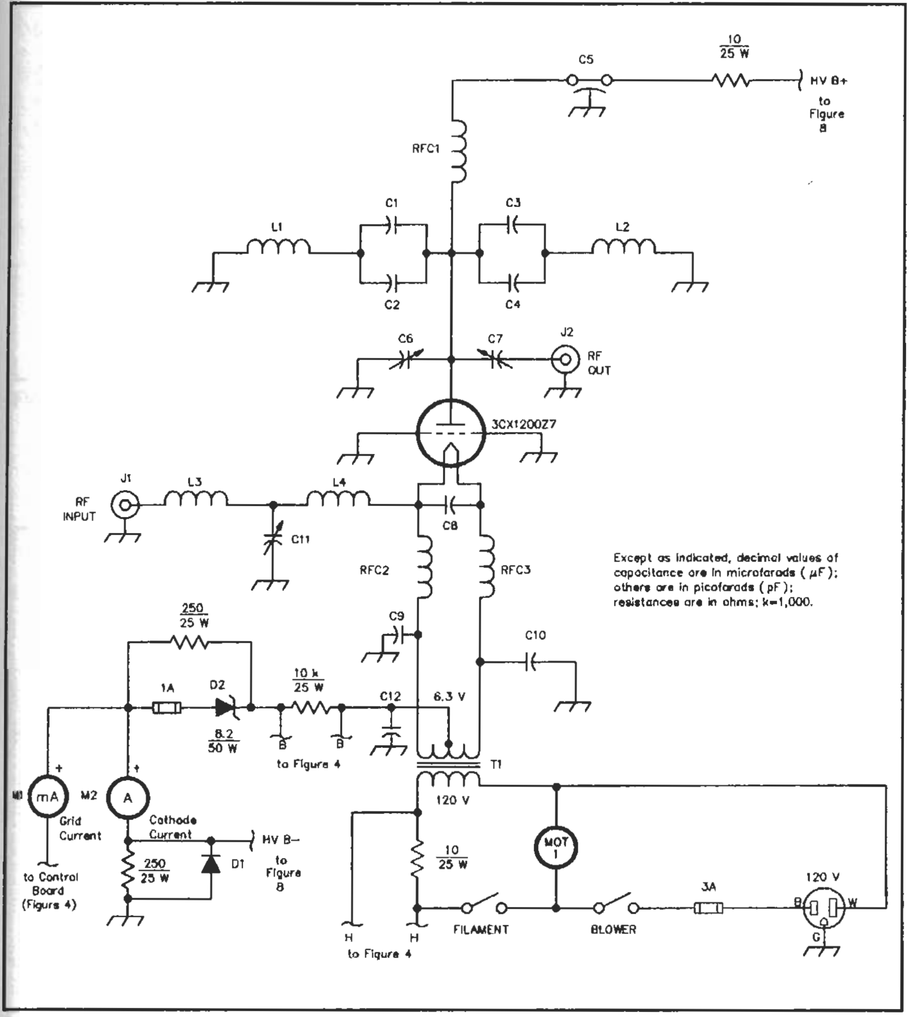
C5
10
25 W
HV B+
to Figure 8
RFC1
C1
C3
L1
L2
C2
C4
C6
C7
J2
RF OUT
3CX1200Z7
J1
RF INPUT
L3
L4
C11
C8
RFC2
RFC3
250
25 W
C9
C10
1A
D2
10 k
25 W
C12
6.3 V
8.2
50 W
B
B
to Figure 4
T1
M1
mA
M2
A
Grid Current
Cathode Current
HV B–
to Figure 8
250
25 W
D1
to Control Board (Figure 4)
120 V
10
25 W
MOT 1
3A
120 V
B
W
G
H
H
to Figure 4
FILAMENT
BLOWER

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000.

Figure 2—Schematic diagram of the 2-meter amplifier.

C1-C4—100 pF, 5 kV, type 850
C5—1000 pF, 5 kV
C6—Anode-tuning capacitor; see text and Figure 5 for details
C7—Output-loading capacitor; see text and Figure 7 for details
C8-C10—1000-pF silver mica, 500 V
C11—30-pF air variable
C12—0.01 µF, 1 kV
D1—1000 PIV, 3-A diode, 1N5408 or equiv
D2—8.2-V, 50-W Zener diode, ECG 5249A
J1—Chassis-mount BNC connector
J2—Type-N connector fitted to output coupling assembly (see Figure 7)
L1, L2—Plate lines; see text and Figure 6 for details
L3—5 t no. 14, 1/2-inch diameter, close wound
L4—3 t no. 14, 5/8-inch diameter, 1/4-inch spacing
RFC1—7 t no.14, 5/8-inch diameter, 1 3/8 inch long
RFC2, RFC3—10 t no. 12, 5/8-inch diameter, 2 inches long
T1—Filament transformer. Primary: 120 V; secondary: 6.3 V, 25 A, center tapped Available from Avatar Magnetics (Ronald C. Williams, W9JVF, 240 Tamara Trail, Indianapolis, IN 46217, 317-783-1211); part number AV-539
M1—Grid milliammeter, 200 mA dc full scale
M2—Cathode ammeter, 2 A dc full scale
MOT1 —140 free-air cfm, 120-V ac blower, Dayton 4C442 or equivalent.

Sources for some of the "hard to get parts" include:

Fair Radio Sales, 1016 E Eureka, Lima, OH 45802, tel 419-227-6573;
Surplus Sales of Nebraska, 1502 Jones Street, Omaha, NE 68102, tel 402-346-4750.

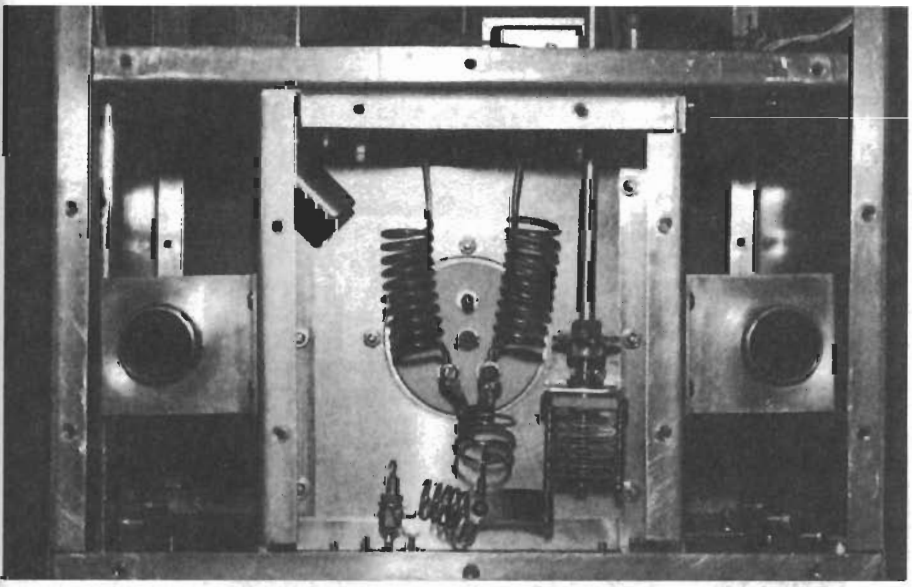

Figure 3—This view of the cathode-circuit compartment shows the input tuned circuit and filament chokes.

safety factor is built in: the keying circuit requires +12 V from the high-voltage supply. This feature ensures that high voltage is present before the amplifier is driven.

The grid overcurrent circuit should be set to trip if grid current reaches 200 mA. When it trips, the relay latches and the NORMAL LED extinguishes. Restoration requires you to press the RESET switch.

Plate Circuit

Figure 5 shows an interior view of the plate compartment.The anode collet is patterned after the one used in the W6PO 222-MHz amplifier. The differences are small. A 4×2 1/4-inch tuning capacitor plate and a 2×2-inch output coupling plate are centered on the collet. These parts are the same size and shape as those used on the 2-meter W6PO amplifier. The remaining difference is the diameter of the hole for the 3CX1200Z7 anode. Sufficient clearance must be left for the fingerstock. The hole diameter will be approximately 3 5/8 inches.

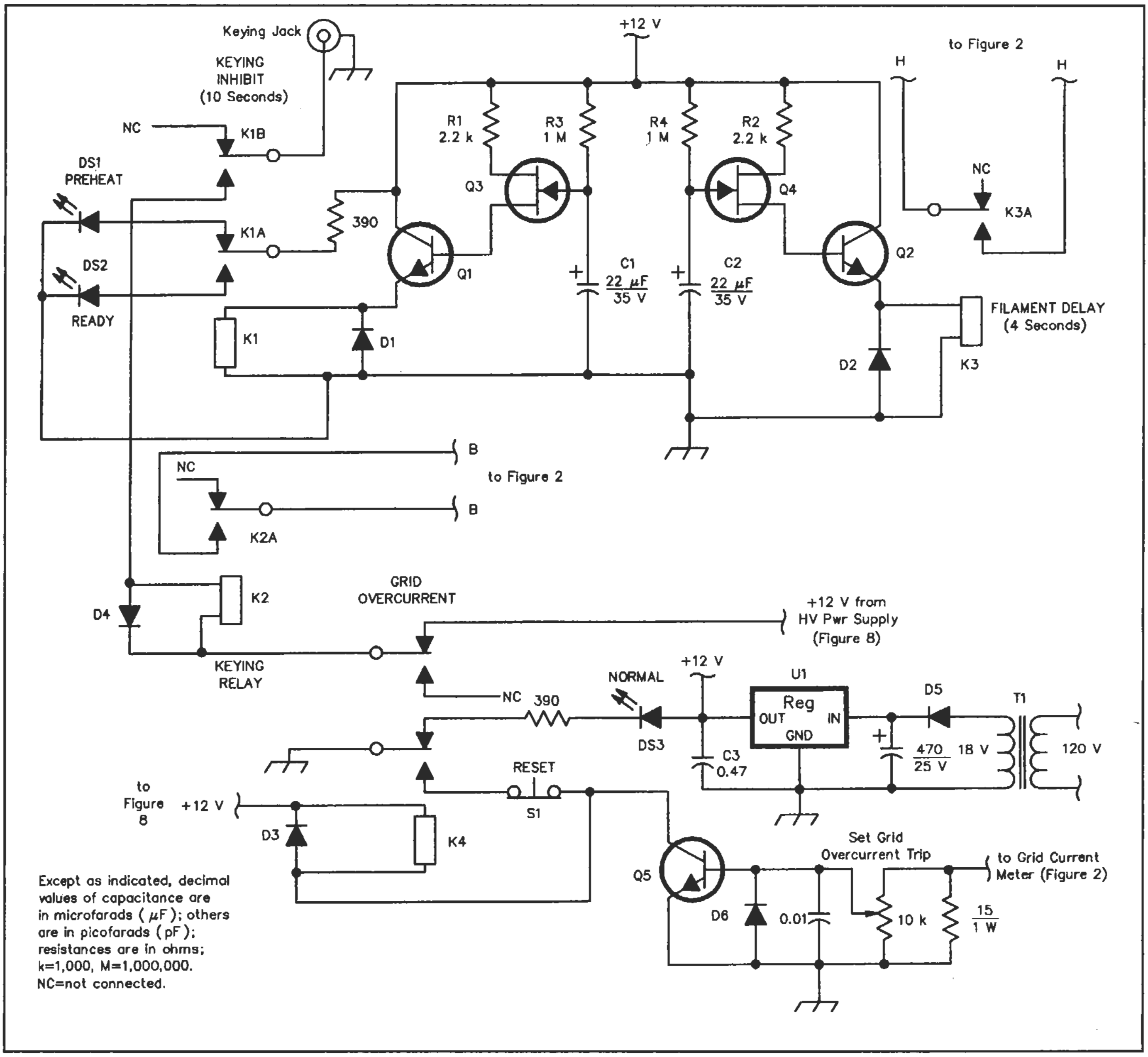
Keying Jack
KEYING INHIBIT (10 Seconds)
NC
K1B
DS1 PREHEAT
K1A
390
DS2
READY
K1
D1
Q1
R1 2.2 k
R3 1 M
Q3
+12 V
R4 1 M
R2 2.2 k
Q4
C1 22 µF 35 V
C2 22 µF 35 V
Q2
to Figure 2
H
H
NC
K3A
FILAMENT DELAY (4 Seconds)
K3
D2
B
to Figure 2
B
NC
K2A
K2
D4
KEYING RELAY
GRID OVERCURRENT
+12 V from HV Pwr Supply (Figure 8)
NC 390
NORMAL
DS3
+12 V
U1
Reg
OUT
IN
GND
C3 0.47
D5
T1
470 25 V
18 V
120 V
RESET
S1
to Figure 8
+12 V
D3
K4
Q5
D6
0.01
Set Grid Overcurrent Trip
10 k
15 1 W
to Grid Current Meter (Figure 2)

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000. NC=not connected.

Figure 4—Schematic diagram of the amplifier-control circuits.

C3—0.47-µF, 25-V tantalum capacitor
D1-D5—1N400i or equivalent
D6—1N4007 or equivalent
DS1—Yellow LED
DS2—Green LED
DS3—Red LED
K1—Keying-inhibit relay, DPDT, 12-V dc coil, 1-A contact rating (RadioShack 275-249 or equivalent)
K2—Amplifier keying relay, SPDT, 12-V dc coil, 2-A contact rating (RadioShack 275-248 or equivalent)
K3—Filament delay relay, SPST, 12-V dc coil, 2-A contact rating (RadioShack 275-248 or equivalent)
K4—Grid-overcurrent relay, DPDT, 12-V dc coil, 1-A contact rating (RadioShack 275-249 or equivalent)
Q1, Q2, Q5—2N2222A or equivalent
Q3—MPF1O2 or equivalent
Q4—2N38i9 or equivalent
S1—Normally closed, momentary pushbutton switch (RadioShack 275-1549 or equivalent)
T1—Power transformer, 120-V primary, 18-V, 1-A secondary
U1—+12 V regulator, 7812 or equivalent

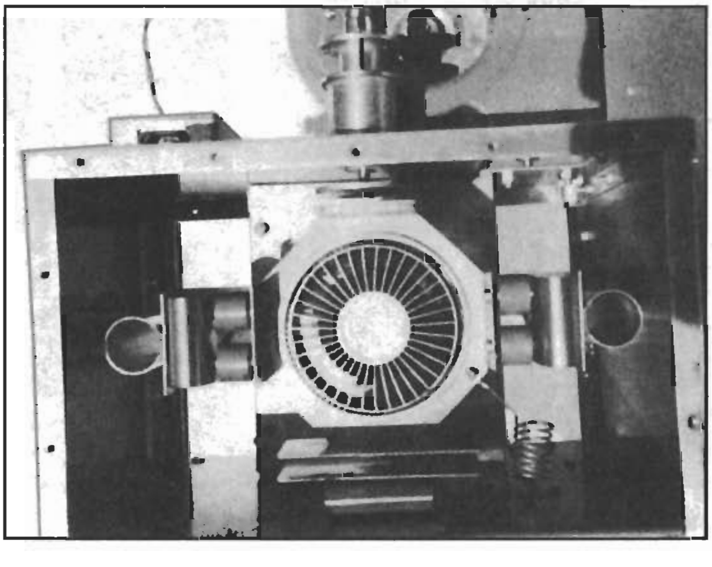

Figure 5—This top view of the plate compartment shows the plate-line arrangement, C1-C4 and the output coupling assembly.

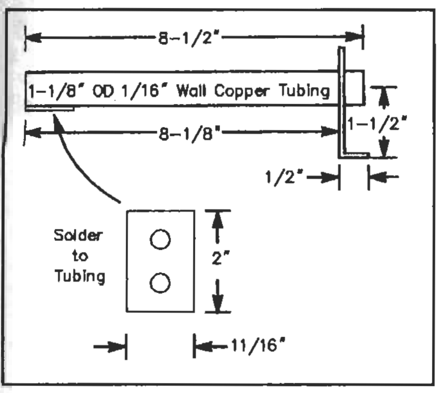
8-1/2"
1-1/8" OD 1/16" Wall Copper Tubing
8-1/8"
1-1/2"
1/2"
Solder to Tubing
2"
11/16"

Figure 6—Plate line details.

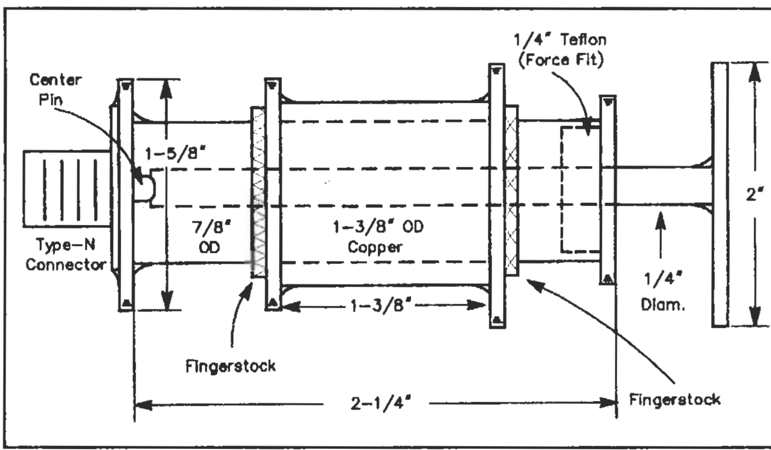
1/4" Teflon (Force Fit)
Center Pin
1-5/8"
Type-N Connector
7/8" OD
1-3/8" OD Copper
1-3/8"
2"
1/4" Diam.
Fingerstock
2-1/4"
Fingerstock

Figure 7—Details of the output coupling assembly.

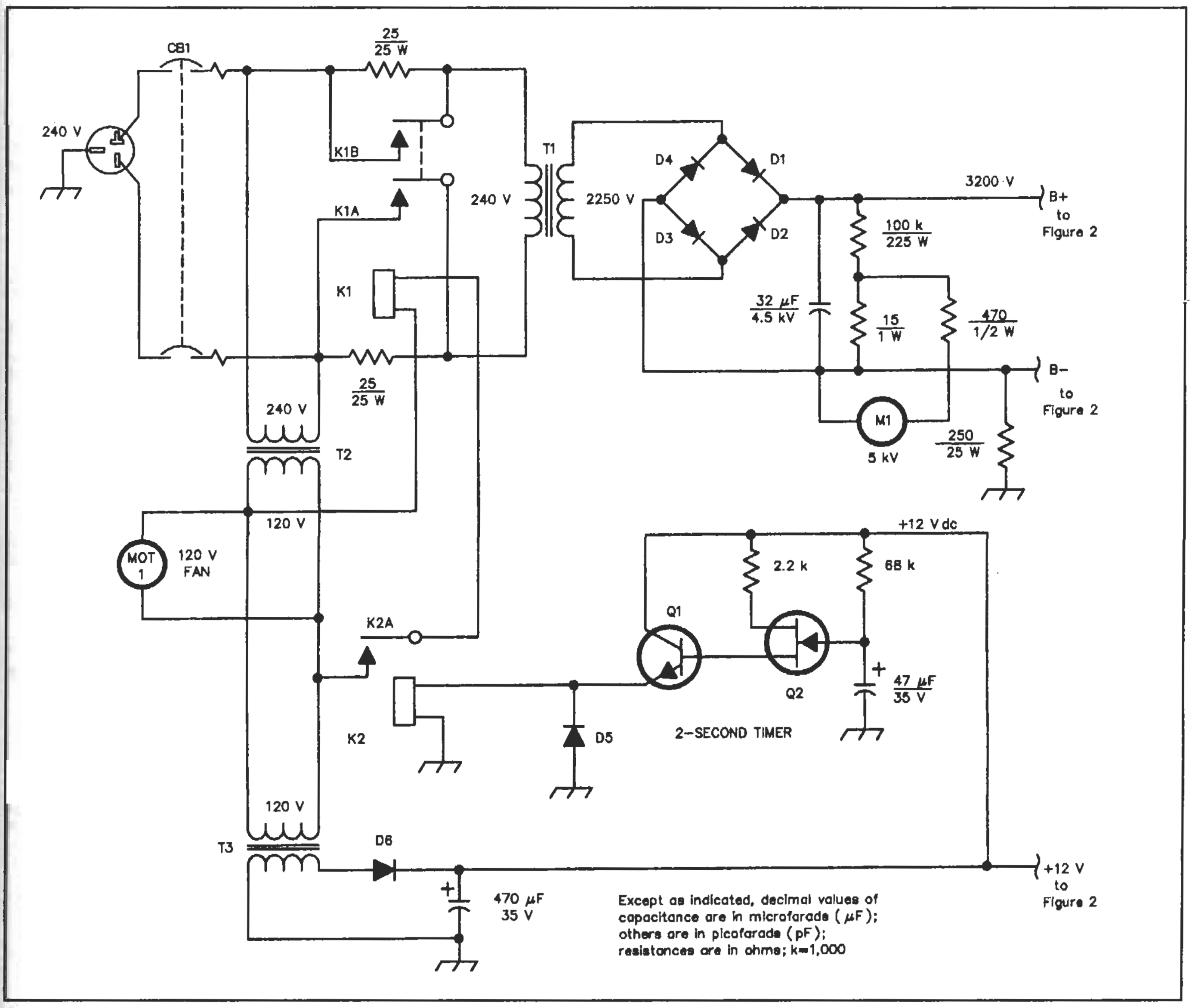
CB1
25
25 W
240 V
K1B
K1A
T1
240 V
2250 V
D4
D1
D3
D2
3200 V
B+ to Figure 2
100 k
225 W
32 µF
4.5 kV
15
1 W
470
1/2 W
K1
25
25 W
B- to Figure 2
240 V
T2
M1
5 kV
250
25 W
120 V
MOT 1
120 V FAN
+12 Vdc
2.2 k
68 k
Q1
K2A
Q2
47 µF
35 V
K2
D5
2-SECOND TIMER
120 V
T3
D6
+12 V to Figure 2
470 µF
35 V
Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000

Figure 8—Schematic diagram of the high-voltage power supply recommended for use with the power amplifier.

D1-D4—Strings of 4 each, 1000-PIV, 3-A diodes, 1N5408 or equivalent
K1—DPST relay, 120-V ac coil, 240-V-ac, 20-A contacts (Midland Ross 187-321200 or equivalent)
K2—SPDT miniature relay, 12-V dc coil (Radio Shack 275-248 or equivalent)
M1—High-voltage meter, 5 kV dc full scale (1-mA meter movement used with series resistors shown in drawing)
MOT1—Cooling fan, Torin TA-300 or equivalent
Q1—2N2222A or equivalent
Q2—MPF102 or equivalent
S1—20-A hydraulic/magnetic circuit breaker (Potter and Brumfield W68X2Q12-20 or equivalent)
T1—High-voltage power transformer, 240-V primary, 2250-V, 1.2-A secondary (Avatar AV-538 or equivalent)
T2—Stepdown transformer, Jameco 112125, 240-V to 120-V, 100 VA
T3—Power transformer, Jameco 104379, 120-V primary; 16.4 V, 1-A secondary (half used)

Figure 6 is a drawing of the plate line and Figure 7 is a drawing of the output coupling asssembly.

Cooling

The amplifier requires an air exhaust through the top cover, as the plate compartment is pressurized. You can fashion a chimney from a $3^1/_2$-inch waste-water coupling (black PVC) and a piece of $^1/_{32}$-inch-thick Teflon sheet. The PVC should extend down from the underside of the amplifier cover plate by $1^1/_8$ inches, with the Teflon sheet extending down $^3/_4$ inch from the bottom of the PVC.

The base of the 3CX1200Z7 is cooled by using bleed air from the plate compartment, which is directed at the tube base, through a $^7/_8$-inch tube set into the sub-chassis wall at a 45° angle.

The recommended blower will supply more tban enough air for any temperature zone. A smaller blower is not recommended, as it is doubtful that the base area will be cooled adequately. The 3CX1200Z7 filament draws 25 A at 6.3 V! It alone generates a great deal of heat around the tube base seals and pins, so good air flow is critical.

Construction

The amplifier is built into a 12×12×10-inch enclosure. A 12×10-inch partition is installed $7^1/_4$ inches from the rear panel. The area between the partition and the front panel contains the filament transformer, control board, meters, switches, Zener diode and miscellaneous small parts. Wiring between the front-panel area and the rear panel is through a $^1/_2$-inch brass tube, located near the shorted end of the right-hand plate line.

High voltage is routed from an MHV jack on the rear panel, through a piece of RG-59, just under the shorted end of the left-hand plate line. The cable then passes through the partition to a high-voltage standoff insulator made from nylon. This insulator is fastened to the partition near the high-voltage feed-through capacitor. A 10-Ω, 25-W resistor is connected between the insulator and the feed-through capacitor.

The plate lines are connected to the dc-blocking capacitors on the plate collet with $1^3/_4$×2-inch phosphor-bronze strips. The bottom of the plate lines are attached to the sides of the subchassis, with the edge of the L-shaped mounting bracket flush with the bottom of the subchassis.

When preparing the subchassis top plate for the 3CX1200Z7, cut a $2^{11}/_{16}$-inch hole in the center of the plate. This hole size allows clearance between the tube envelope and the top plate, without putting stress on the envelope in the vicinity of the grid flange seal.

Table 3
Power Supply Specifications
High voltage: 3200 V
Continuous current: 1.2 A
Intermittent current: 2 A
Step/Start delay: 2 seconds

Exercise care in placing the movable tuning plate and the movable output coupling disc, to ensure they cannot touch their fixed counterparts on the plate collet.

Operation

When the amplifier is first turned on, it cannot be keyed until:

- 10 seconds has elapsed
- High voltage is available, as confirmed by presence of +12 V to the keying circuit

Connect the amplifier to a dummy load through an accurate power meter capable of indicating 1500 W full scale. Key the amplifier and check the idling plate current. With 3200-V plate voltage, it should be in the vicinity of 150 mA. Now, apply a small amount of drive and adjust the input tuning for maximum grid current. Adjust the output tuning until you see an indication of RF output. Increase drive and adjust the output coupling and tuning for the desired output. Do not overcouple the output; once desired output is reached, do not increase loading.

When you shut down the amplifier, leave the blower running for at least three minutes after you turn off the filament voltage. I found the 3CX1200Z7 to be an excellent tube. I tried it with exessive drive, plate-current saturation, excessive plate dissipation—all the abuse it's likely to encounter in amateur applications. I had no problems, but I don't recommend you repeat these tests!

A Companion Power Supply

A good, solid-state high-voltage power supply is a necessity to ensure linearity in SSB operation. Specifications of the power supply I built are given in Table 3. Figure 8 is a schematic diagram of the supply. A power supply for a high-power linear amplifier should operate from a 240-V circuit, for best line regulation. I have specified a special, hydraulic/magnetic circuit breaker that doubles as the main power switch. I don't recommend you substitute a regular switch and fuses for this breaker, as fuses won't operate quickly enough to protect the amplifier in case of an operating abnormality. The bleeder resistor dissipates about 100 W, so I included a small fan to remove the excess heat.

Power Supply Construction

The power supply can be built into a 17×13×10-inch cabinet. The power transformer is quite heavy, so use $^1/_8$-inch aluminum for the cabinet bottom, and reinforce it with aluminum angle for extra strength. The diode bridge consists of four legs, each containing five diodes.

Power Supply Operation

When the front-panel breaker is turned on, the two, 25-Ω resistors in the primary circuit limit inrush current as the filter capacitor charges. After two seconds, K1 activates, shorting both resistors and allowing full line voltage to be applied to the transformer.

As with all high-voltage power supplies, you must be extremely careful! Before opening the cabinet, remove the ac-line plug from its receptacle, and confirm that the filter capacitor is discharged before working on the supply.

Conclusion

This amplifier is a reliable and cost-effective way to generate a big 2-meter signal—almost as quickly as a solid-state amplifier.

To ensure that the output of my amplifier meets current spectral purity requirements, I use a high-power version of the half-wave output filter that appears as Figure 16 on page 39-10 of the 1993 and 1994 editions of *The ARRL Handbook*. Although I did not make spectral measurements of the output, I can run full output while my wife Mary Lou watches TV in a nearby room of our home.

Another suitable filter is the one that appears in the 1990 *ARRL Handbook* (Figure 150, on page 31-72) as part of the description of "A Legal-Limit 2-Meter Tetrode Amplifier."

Notes

[1] Suggested retail price of the 3CX1200Z7 is $625. You can obtain it from: Henry Radio, 2050 S Bundy Dr, Los Angeles, CA 90025, tel 310-820-1234; Richardson Electronics, 40 W 267 Keflinger Rd, La Fox, IL 60147, tel 708-208-2200; RF Parts, 435 South Pacific St, San Marcos, CA 92069, tel 619-744-0700.

[2] W. Orr, Editor, *Radio Handbook*, 23rd ed. (Indianapolis: Howard W. Sams and Co., 1987), pp 18-2 through 18-7.

[3] This project is also described in the *Radio Handbook*, pp 18-11 through 18-15.

By David Munyon, W7DVB From *QST*, September 1984

A Cathode-Driven Tetrode for 6 Meters

Stability, power, low cost: Is this what you're looking for in a 6-meter do-it-yourself project? If so, look what's here!

The advantages of grounded-grid amplifiers are numerous and have been proven over the years. They include: simplicity of design, good thirdorder IMD characteristics, inherent stability (generally without the need for neutralization) and noncritical tuning. The only requirement is abundant driving power. Usually, this is a small price to pay, unless the driver you intend to use is one of the many popular, low-power (10- to 25-W-output) VHF transceivers. Also, with the tremendous increase in the use of 100%-duty-cycle modes such as RTTY and SSTV, many of us would rather not operate our exciters at half to full bore while driving an amplifier. If you're willing to sacrifice only one of the aforementioned advantages—simplicity—the remaining attributes of grounded-grid operation can be made available at a considerable reduction in driving power.

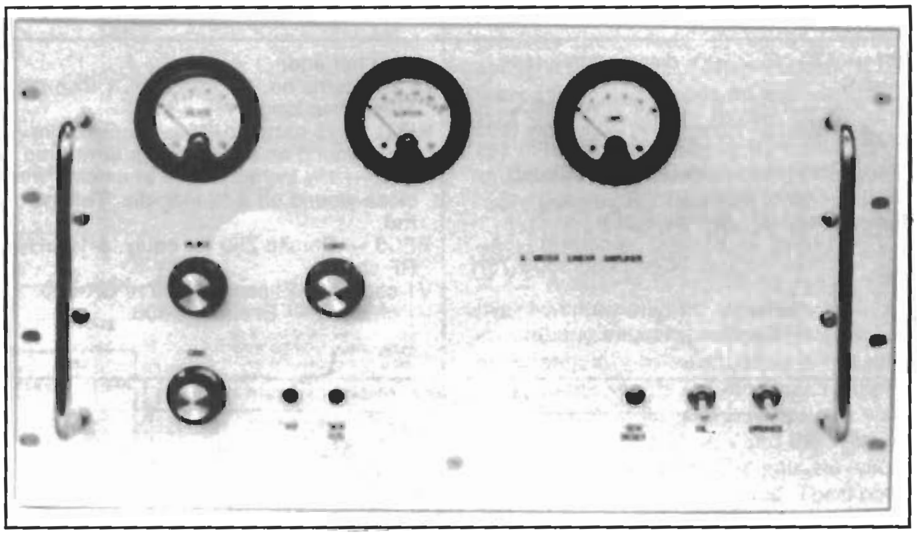

Enter the Tetrode

For many years, tetrodes (most often connected as triodes) have been used in grounded-grid circuits. But we have good reasons to consider the tetrode as a cathode-driven performer in its own right. First, the drive requirements are reduced. A 4-400A requires 40W of drive power for full output when triode connected. Only half that amount is needed to drive its class-AB1 counterpart. The second reason is cost. Tetrodes are still available as hamfest and surplus items at a fraction of their original cost. (Some of the newer triodes cost more than the exciters that drive them!)

But not all tetrodes can be triode connected. All of the external anode family of tubes, and a few others, have internal geometry that allows the control grid dissipation to be greatly exceeded if the tube is operated as a triode. Therefore, those tubes should be cathode driven only as tetrodes. The usual way to do this is to ground the control and screen grids to RF with low-inductance bypass capacitors, and then place operating voltages on them. The cathode can be driven in a normal fashion, and the tube operated in any class consistent with linear service. Class-AB1 has the lowest drive requirements and is the easiest with which to work.

Tetrode Vs. Triode

I concede that the addition of a screen element complicates the overall circuit, especially if the screen-circuit provisions are made properly. But, the circuit need be no more complex than that of its grid-driven equivalent. The cathode-driven tetrode requires grid bias and a well-regulated screen supply. Also, there is no correct way to load a tetrode without a meter to monitor the screen-grid current. This device is invaluable and should not be omitted for the sake of economy.

No discussion of the screen-grid circuit would be complete without some mention of an overcurrent-protection circuit. Several solutions have been offered in the past (including the use of no protection circuit!), and each has its disadvantages. Sensitive relays are expensive and their use can have an adverse effect on screen-voltage regulation. Current-limiting supplies do an excellent job of protecting the tube, but do not inform you when something is wrong. Since screen current is extremely sensitive to minor plate-voltage excursions and plateloading conditions, the builder has to be absolutely certain that these are correct, and that plate voltage is always present with screen voltage, if the screen-protection circuit is to be eliminated. The protection circuit is well worth the small cost involved. This amplifier incorporates an inexpensive and simple circuit that contains none of the disadvantages mentioned earlier.

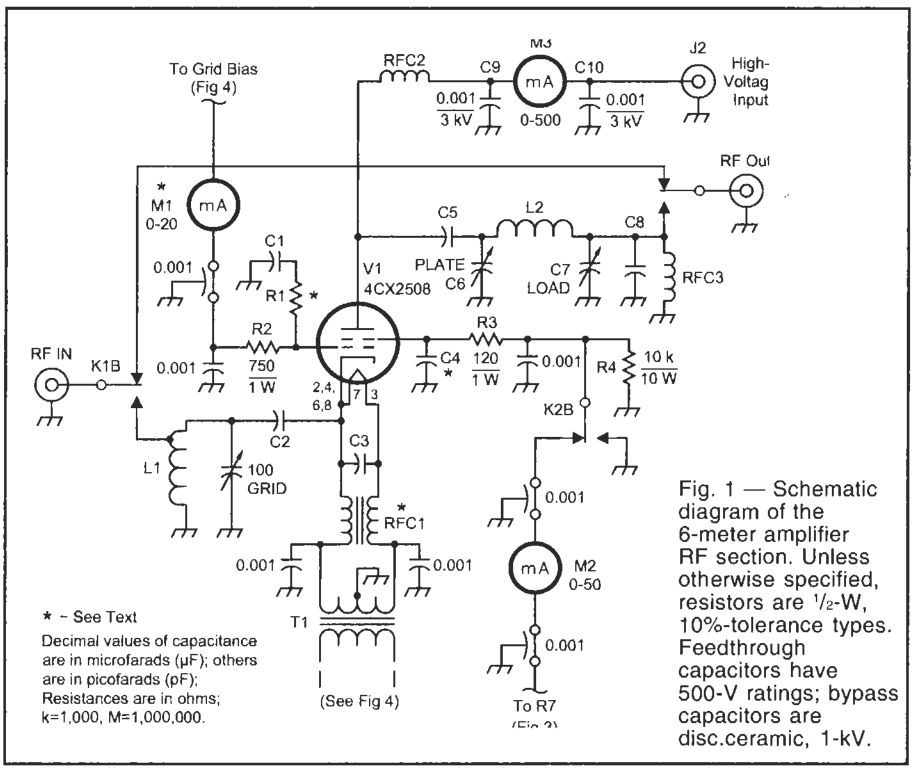
To Grid Bias
(Fig 4)
RFC2
C9
M3
mA
0-500
C10
J2
High-Voltag Input
0.001
3 kV
RF Out
M1
0-20
mA
C1
C5
L2
C8
0.001
R1
V1
4CX2508
PLATE
C6
C7
LOAD
RFC3
R2
750
1 W
R3
120
1 W
C4
0.001
R4
10 k
10 W
RF IN
K1B
0.001
2,4,
6,8
7
3
K2B
C2
C3
L1
100
GRID
RFC1
0.001
0.001
M2
0-50
mA
T1
(See Fig 4)
To R7
(Fig 3)
* - See Text
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.

Fig. 1 — Schematic diagram of the 6-meter amplifier RF section. Unless otherwise specified, resistors are 1/2-W, 10%-tolerance types. Feedthrough capacitors have 500-V ratings; bypass capacitors are disc.ceramic, 1-kV.

C1 — 470-pF, 2.5-kV disc-ceramic (see text).
C2, C3 — 0.001-µF, 500.V silver mica.
C4 — Part of tube socket assembly (see text).
C5 — 500-pF, 5-ky ceramic (Centralab 858S-500 or equlv.).
C6 — See text (approx. 6 pF).
C7 — 140-pF air variable, receiving type.
C8 — 75-pF ceramic (Centralab 850S-75N or equiv.).
L1 — 4 1/2 turns no. 14 bare wire, 1/2-inch-dia, 1 1/2 inches long, tapped one turn from hot end.
L2 — 6 turns no. 10 bare wire, 1-inch-dia, 1 1/2 inches long.
RFC1 — 15 turns no. 16 enameled wire closewound on a 1/2-inch-dia ferrite rod.
RFC2 — 1 5/8 inches no. 22 enameled wire close-wound on a 3/8-inch-dia. Teflon® rod.
RFC3 — Ohmite Z50 (or equiv. 5-10 µH RF choke).
V1 socket — Eimac SK-600 or SK-620.
V1 chimney — Elmac SK-606.

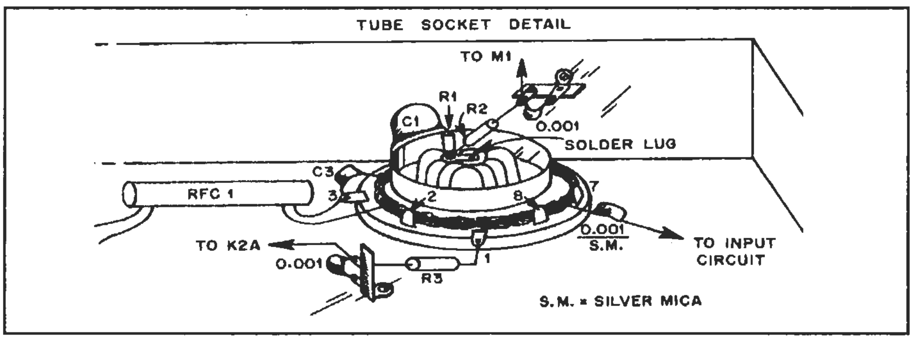
TUBE SOCKET DETAIL
TO M1
R1
R2
C1
0.001
SOLDER LUG
C3
RFC 1
3
2
8
7
0.001
S.M.
TO INPUT CIRCUIT
TO K2A
0.001
1
R3
S.M. = SILVER MICA

Fig 2—Tube socket detail. See text.

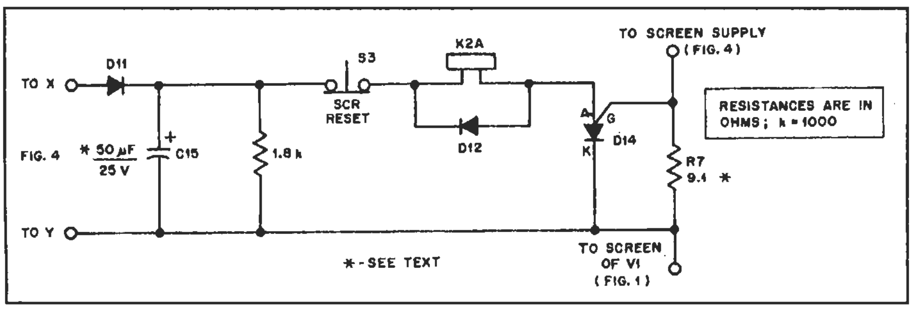
TO X
D11
FIG. 4
* 50 µF
25 V
C15
1.8 k
TO Y
S3
SCR RESET
K2A
D12
TO SCREEN SUPPLY
(FIG. 4)
A
G
K
D14
RESISTANCES ARE IN OHMS; k = 1000
R7
9.1 *
TO SCREEN OF V1
(FIG. 1)
* - SEE TEXT

Fig 3—Screen overcurrent-protection circuit. Resistors are 1/2-W, 10%-tolerance types.

D11, D12—1-kV, 2.5-A silicon diodes.
D14—4-A, 200-V SCR (Jameco C106B1, ECG 5455 or equiv.).
K2—DPDT 12-V dc relay, 1-A contacts (see text).

A Cathode-Driven 4CX250B

I became aware of the cathode-driven tetrode years ago while trying to get a 4-1000A to wotk on 50 MHz. The results were so good that, when it came time to build an amplifier for my new solid-state transverter, I decided that building a grid-driven amplifier would be a giant step backward.

The 4CX250B is an excellent tube for 6-meter use. This tube has a high plate-dissipation rating for its small physical size, has reasonable power-supply requirements, and requires approximately 10 W of drive in a cathode-driven, class-AB1 configuration. (If much more drive is available, an appropriate attenuator must precede the input circuit.) The only differences between this design and its grid-driven equivalent are the filament/cathode and control-grid circuits.

Filament/Cathode Circuit

In Fig. 1, notice that the cathode and one of the filament connections are tied together. Normally, the isolation created by the physical separation of the two elements is sufficient to ensure stability, and a filament choke would not be required. I felt that this would not be the case on 50 MHz, and a call to Eimac confirmed it. At Bill Orr's recommendation, I tied the cathode and filament together, and fed the filament through a choke. The cathode impedance is approximately 120 ohms. In the interest of best linearity and minimum, drive requirements, a tuned input circuit is used. The tank coil is tapped to present an input impedance of 50 ohms.

There are many sockets designed for use with this family of tubes. I recommend that you use one that incorporates a built-in screen bypass capacitor (1100 or 2700 pF); either one will do. A few of these types of sockets are so constructed that the four cathode pins are grounded internally; do not use one of these. An Eimac SK-600 or SK-620 socket is recommended. The four cathode pins (2,4, 6, 8) and one of the filament pins (7) of the socket must be wired together to form one low-inductance connection. The easiest way to do this is to carefully bend pins 1 and 3 parallel to the chassis, and then strap together pins 2, 4, 6, 7 and 8. Make certain that this strap forms a complete circle around the socket. For strap material, I use some stretched and flattened shield braid removed from RG-58 coaxial cable. All connections made to the cathode, or the filament (pin 7), can be made anywhere along this ring. See Fig. 2 for details.

Grid Circuit

The tube grid is accessed through the socket center connector and is bypassed to ground by C1. The capacitance of C1 is not as important as its physical size. C1 should be a high-voltage, disc-ceramic type (2 kV or more), whose body nearly spans the distance between the connection points. This provides for an absolute minimum lead

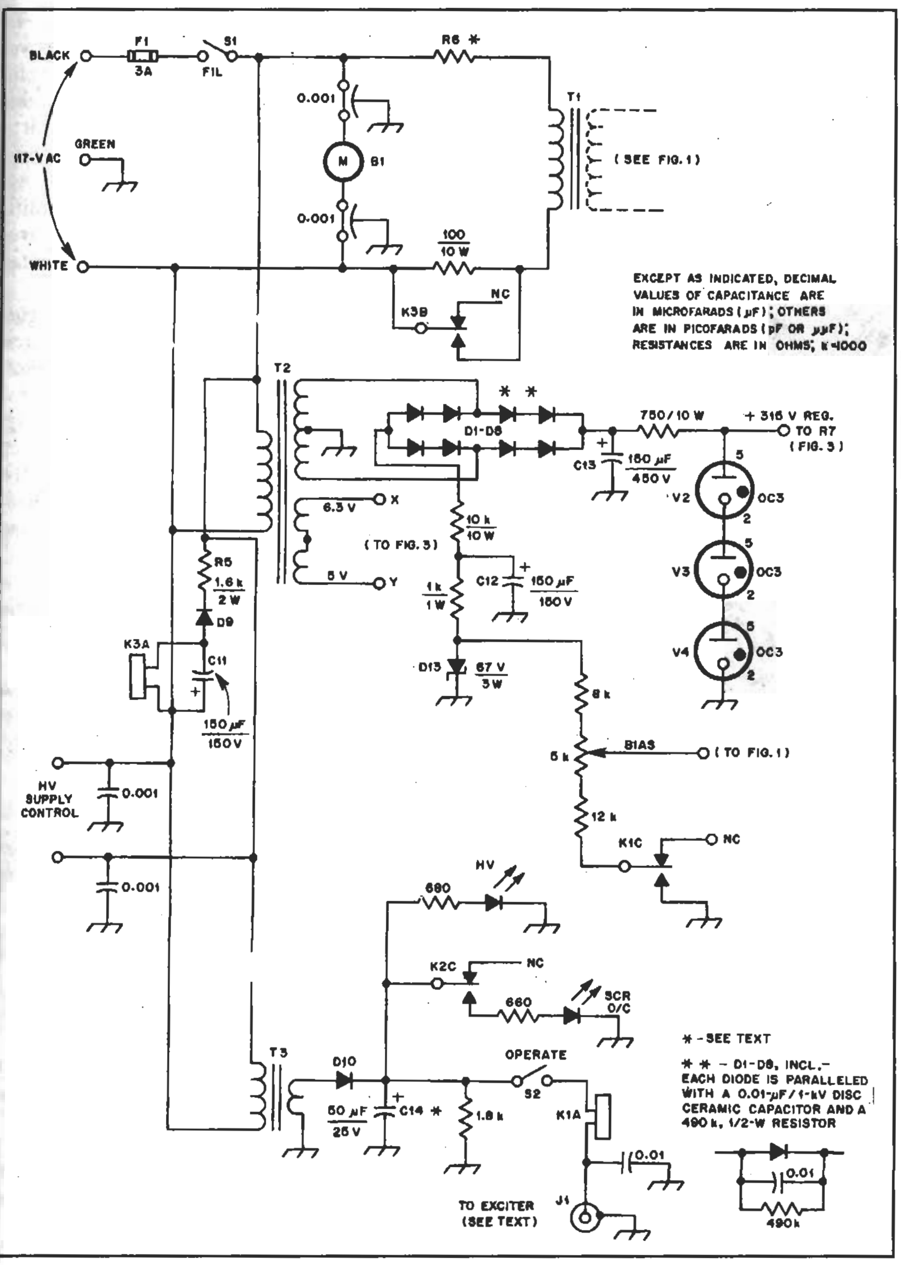
BLACK
F1
3A
S1
FIL
117-V AC
GREEN
WHITE
R6 *
0.001
M
B1
0.001
100
10 W
T1
(SEE FIG. 1)
K3B
NC
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000
T2
* *
D1-D8
750/10 W
+ 316 V REG.
TO R7
(FIG. 3)
C13
150 µF
450 V
V2
OC3
5
2
6.3 V
X
(TO FIG. 3)
5 V
Y
10 k
10 W
1 k
1 W
C12
150 µF
150 V
V3
OC3
V4
OC3
D13
87 V
5 W
R5
1.6 k
2 W
D9
K3A
C11
150 µF
150 V
HV
SUPPLY
CONTROL
0.001
0.001
8 k
5 k
BIAS
(TO FIG. 1)
12 k
K1C
NC
HV
680
K2C
NC
660
SCR
O/C
T3
D10
50 µF
25 V
C14 *
1.8 k
OPERATE
S2
K1A
0.01
TO EXCITER
(SEE TEXT)
J1
* - SEE TEXT
* * - D1-D8, INCL.- EACH DIODE IS PARALLELED WITH A 0.01-µF/1-kV DISC CERAMIC CAPACITOR AND A 490 k, 1/2-W RESISTOR
0.01
490k

Fig. 4 — Power-supply section of the 6-meter amplifier. Unless otherwise specified, resistors are ½-W, 10%-tolerance types; feedthrough and bypass capacitors have 500-V ratings.

B1 — Squirrel-cage blower, 56 CFM (minimum), 117-V ac motor.
D1-D10, incl. — 1-kV, 2.5-A silicon diodes.
D13 — 87-V, 5-W Zener diode.
K1 — 3PDT 12-V dc relay, 5-A contacts.
K3 — SPST 234-V ac relay, 3-A contacts (see text).
T1 — 117-V primary; 6-V. 2.6-A secondary (see text).
T2 — 117-V primary; 580-V c.t., 75-mA secondary with 6.3- and 5-V filament windings (see text).
T3 — 117-V primary; 12-V. 1-A secondary.

length. Use a solder lug on the grid pin, and solder the other capacitor lead directly to the inside of the air duct on the tube-socket bottom.

When my amplifier was first tested, I found that the plate current idled at about 40 mA, and was completely independent of grid voltage. This self-biasing was traced to a parasitic in the grid circuit, and was eliminated by the addition of a series-connected resistor, R1. If you wish to experiment, R1 should be a ½-W, carbon-composition resistor with a resistance value that is the minimum required for complete stability. R1 should be placed as shown in Fig. 2; I found that a 3.9-ohm resistor kept my amplifier unconditionally stable.

In the interest of stability, don't use chokes in place of R2 and R3 in the grid and screen circuits. M1, the grid-current meter, can have a range of from 0-1 mA to 0-20 mA or so, and is included only to ensure that grid current is never drawn, and the operation remains class-AB 1.

Screen Circuit

Relay contacts K2B are part of the screen overcurrent-protection circuit, the operation of which will be discussed later. Because the 4CX250B draws negative screen current on occasion, the screen-current meter should ideally be of the zero-center type. As these meters are expensive and not readily found, a standard meter can be pressed into service by using bleeder string, R4. Notice that 15 mA of bleeder current is drawn through the meter whenever K2 is open (normal). This offsets the actual zero reading of the meter to the 15-mA position and allows a negative current of 15 mA to be metered. This is more than adequate for the 4CX250B. If a zero-center meter is used, R4 may be eliminated unless the screen supply shunt regulator is replaced by a series regulator. In that case, the bleeder would be necessary to offset the effects of secondary emission.

For the overcurrent-protection circuit (Fig. 3), an SCR (D14) seems to be a natural. A very small gate voltage, developed across R7 by the screen current, turns on the SCR and causes K2 to close. This removes screen voltage from the tube and grounds the screen, preventing further abuse. A second set of relay contacts (K2C) is used to light a front-panel-mounted LED. D14 continues to lock out the screen voltage until the SCR is reset by removing its anode voltage. This is done by momentarily pushing the RESET switch, S3.

D11, C15 and the filament windings of T2 provide dc power for a 12-V relay at K2. A relay with a different voltage rating can be used if this supply is modified accordingly. The value for R7 was found experimentally by substituting a variable, low-voltage supply for the screen supply. A 0-100 mA meter and load resistor were placed in series with R7, and the voltage, slowly increased until the gate threshold was reached and the SCR fired. The threshold current was noted, and R7 varied until the desired current would trigger the SCR. With R7 equal to 9.1 ohms, the SCR fired consistently at 45 mA.

Since 15 mA of the current drawn through R7 will be bleeder current, this allows a maximum of 30 mA for screen current, which is well below the 12-W screen dissipation rating for the tube. Note that the entire overcurrent-protection circuit is hot with screen voltage, so proper precautions must be taken during construction. I built the protection circuit on a PC board, but it could have just as easily been chassis-mounted using terminal strips.

Construction

The amplifier is built on a 3 × 8 × 17-inch chassis; a 3 × 5 × 7-inch chassis houses the output network components. A 10½- × 19-inch rack panel is used for the amplifier front panel. The general physical layout should be followed, as it is well dictated by the flow of the circuit. All RF leads should be kept as short as possible, especially the bypass capacitor leads. Keep the dc leads cabled and

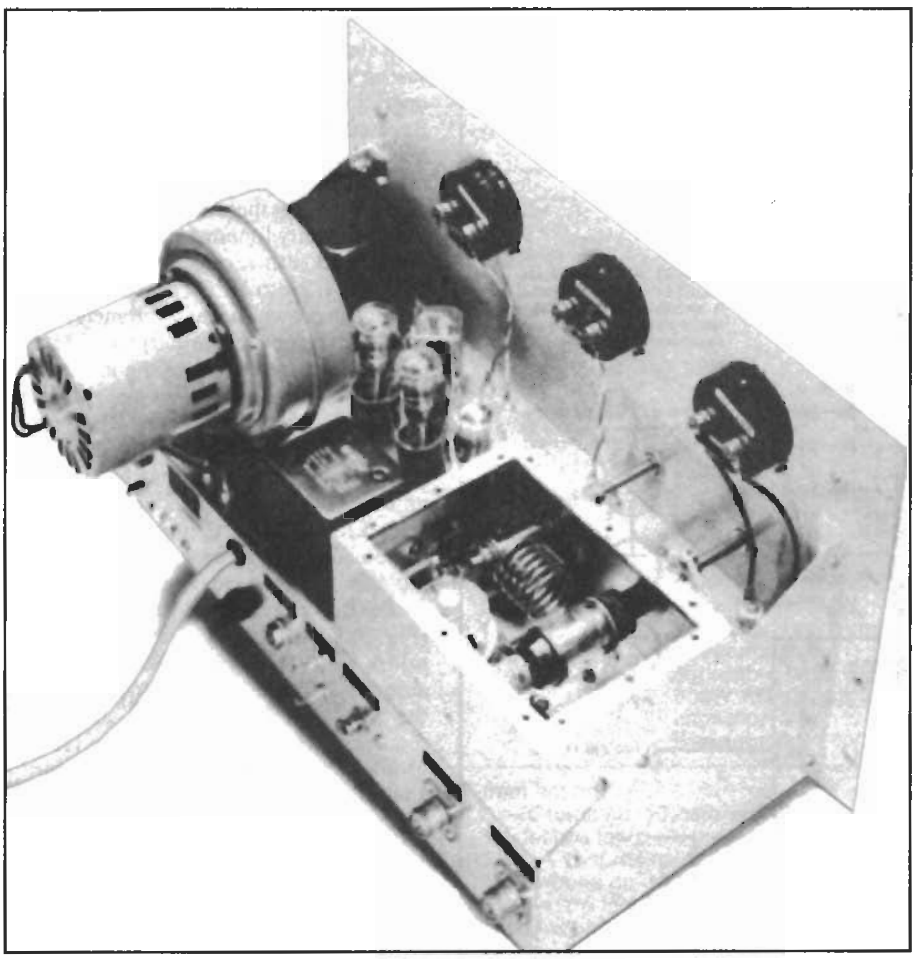

Fig 5—The final-amplifier-compartment cover has been removed for this photo.

as far away as possible from the RF leads. Most power-supply components are mounted on a fiberglass terminal board that is secured against the side of the chassis.

K3, D9, C11 and R5 (Fig. 4) constitute a time-delay circuit for filament inrush-current protection, and can be included if you feel that this is important. As it stands, the delay is approximately 1 second, which is more than ample. The 234-V ac relay is used on dc, and closes at about 70 V. The charging rate of C11 is controlled by its capacitance and the ohmic value of R5, and can be varied by changing their values.

Ideally, the input and output relays should be coaxial types; however, I used a single open-frame, 3PDT type and encountered no problems. The center pole is used for control (K1C), as that increases the isolation between the input and output circuits. In grid-driven amplifiers, I have never been able to use an open-frame relay for input and output switching. Because of the lower gain of this cathode-driven amplifier, I find no evidence of feedback or instability. Don't use plug-in type relays in their original form for K1. Remove the case, plug and connecting wires, and bolt the relay directly to the chassis. Use coaxial-cable braid to make the RF connections. The power supplies (Fig. 4) are of standard design. They may be replaced by any supplies yielding similar voltages, and need not reside on the same chassis as the amplifier. The high-voltage supply should produce 2-kV dc under load. Si turns on all supplies, including the external high-voltage supply. S2, the STANDBY/OPERATE switch, supplies power to K1, which allows the amplifier to be placed in operation; otherwise, the amplifier is simply bypassed.

If a 6.3-V ac filament transformer is used for T1, R6 will have to be.included to drop the filament voltage to 6 V. The ohmic value of R6 will have to be determined experimentally. Use a resistor with the highest wattage rating practical for good voltage stability. With an accurate voltmeter, measure the filament voltage at the tube socket with the filament choke in place, as there will be a slight voltage drop across the choke. Nothing will cause a 4CX250B to "go south" quicker than high filament voltage, and that means anything in excess of 6 V.

The filament choke (RFC1) consists of 15 bifilar turns of no. 16 enameled wire wound on half of an Amidon $^{1}/_{2}$-inch-diameter ferrite rod. The rod can be cut to a proper length by filing a small groove around its circumference, and then breaking it clean. Cover the rod with heat-shrink tubing or electrical tape before and after winding the choke. This will keep the rod and winding together, and make it self-supporting by its four leads.

The meters I used all have 0-1 mA movements and homemade shunts. The original meter-face calibration marks were carefully erased, and new ones applied using dry-transfer labels. Most such labels are ideal for panel marking, but are a bit too large for meter-face use. Your local stationery store should have transfers with smaller-sized numbers. During amplifier operation, care should be taken to avoid accidental contact with the plate-voltage meter, as it has high voltage on it.

Most VHF amplifiers have a higher-than-necessary Q in the output circuit because of the use of high minimum-capacitance air-variable capacitors. This is acceptable providing the rest of the output components are heavy-duty types and can handle the higher circulating currents. (The small vacuum variable I used is admittedly first class, and it may not be used by many other builders.) Rather than using an air-variable capacitor, a homemade two-disc system is a preferred choice. If you can find a neutralizing capacitor with plates of 2-inch diameter or so, that would be ideal. If you must use an air-variable capacitor, use one with the least minimum capacitance; remove all unnecessary plates leaving just enough to do the job.

Tune-Up

Before the tube is placed in the socket, close S1 and check that all of the operating voltages are correct. Ground the bottom end of the bias string at K1C, and check that the BIAS potentiometer range will supply –50 to –60 V; then set it at –55 V. J1 should be connected to the external-circuit control jack of your exciter so that it is grounded when your exciter is keyed. Open S1 and make certain all voltages have bled down, then install the tube. Close S1 and S2. After the tube warms up, with no drive applied, key the exciter and adjust the BIAS potentiometer so that the plate idling current is exactly 100 mA.

While the tube is idling, rotate all the tuning controls throughout their respective ranges. If there are no sudden plate-current increases, the amplifier is stable. With a small amount of drive applied, tune the GRID and PLATE controls for maximum output. Once you are satisfied that both of these controls will provide resonance, apply full drive and retune the plate to resonance. Increase loading slowly, retuning the PLATE control each time, until a screen current of 5 mA (the meter actually reads 20 mA) is indicated. Observe how sensitive the screen current is to the operation of the LOAD control, and that it is by far the best indicator of plate resonance. If grid current is indicated, you're overdriving the amplifier.

Adjust the drive level and the plate TUNE and LOAD controls so that the platecurrent meter indicates 250 mA and the screen-current meter indicates 5 mA at resonance. Under these conditions, this amplifier develops a power output of about 310 W. With modulation, voice peaks should not exceed 150 mA on the plate-

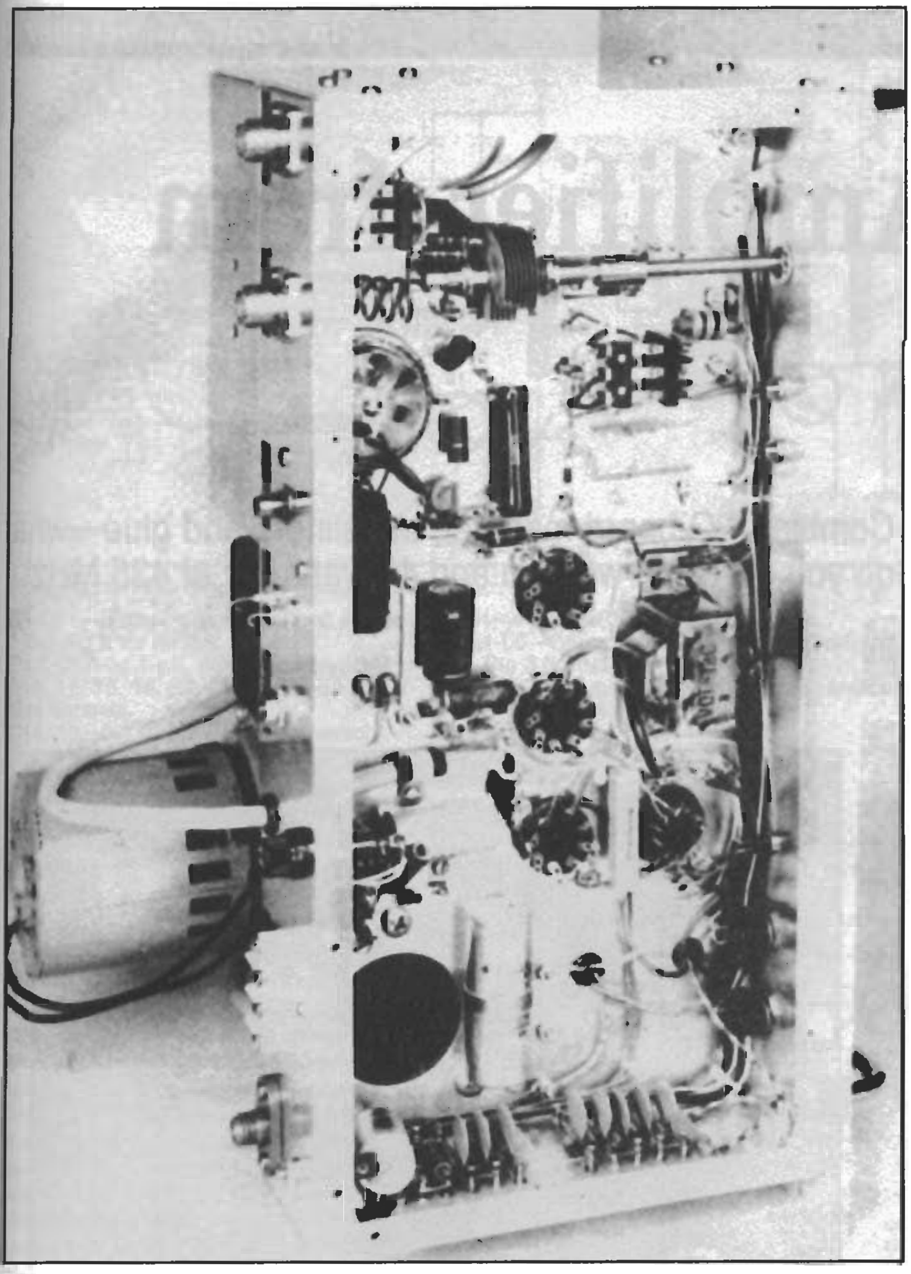

Fig 6—A bottom view of the 6-meter amplifier. The screen-protection circuit is mounted on the PC board to the right of the tube socket.

current meter, and a small amount of negative screen current is normal.

Summary

This amplifier has been in operation now for over a year, and all reports of its operation have been complimentary. I hope that this amplifier will be duplicated by many who had previously resigned themselves to grid-driven designs. Much of the material presented here is general enough to encourage interest in the cathode-driven tetrode as a viable alternative to the grounded-grid triode. I will be glad to answer all letters of inquiry regarding this amplifier; please include an s.a.s.e.

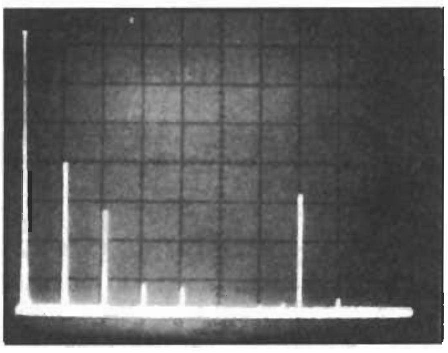

Fig 7—Spectral photo of the 6-meter amplifier without external filtering. Vertical divisions are each 10 dB; horizontal divisions are each 50 MHz. The fundamental (second pip from the left) has been notched approximately 40 dB by means of notch cavities to prevent analyzer overload. The seventh harmonic is approximately 48 dB below peak fundamental output.

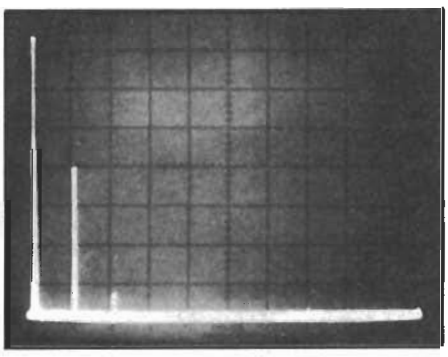

Fig 8—The addition of a simple filter (see Fig 9) at the amplifier output provides for excellent harmonic attenuation, and its use is recommended. All conditions are otherwise the same as those of Fig 7.

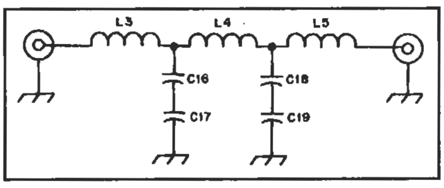

Fig 9—This simple harmonic-suppression filter, or one of similar characteristics, should be used at the output of the amplifier. The filter may be contained in a $2^1/_4 \times 2^1/_4 \times 5$-in (HWD) aluminum box.

C16-C19, incl.—110 pF, 1-kV silver-mica capacitors.

L3, L5—4 turns no. 14, $^5/_{16}$-in-ID, $^1/_2$ in long.

L4—5 turns no. 14, $^7/_{16}$-in-ID, $^5/_8$ in long.

By John Reed, W6IOJ **From *QST*, August 1987**

A UHF Amplifier–from Scratch

Combine PC-board material, transistors and glue—what do you get? Two watts in and 45 watts out at 435 MHz!

Most UHF RF equipment designs involve undefined reactance. For example, the self-inductance of a variable capacitor, or simply the method of attaching a 50-ohm cable to the circuit may constitute reactance that is critical to the operation of the circuit. The result is that complete success in duplicating a circuit is not likely unless the components and layout are exactly the same as in the original. This is nearly always a limiting factor for the do-it-yourselfer wanting to duplicate a project. This class-C UHF amplifier project uses homemade components that anyone can assemble. Tests show that their use has resulted in little or no compromise in the performance of the amplifier over a design using far more expensive commercial parts. My amplifier shows more than 45 W output with less than 2 W input at 435 MHz. The output stage efficiency is approximately 60%, and includes a double-tuned filter for improved spectral purity.

The key to this design is the glue-down stripline technique I have used in many other UHF projects.[1] This method permits easy modification of stripline parameters for optimizing circuitry. Striplines are cut from double-sided, glass-epoxy PC board having the same dimensions that you would choose using the etched-PC-board method. One side of the stripline is smeared with glue (Radio Shack all-purpose adhesive, 64-2307) and firmly pressed against the common base PC board. Parts can be soldered to the stripline immediately (without waiting for the glue to dry). No dc connection is required between the glueline foils, and changes can be made within minutes by lifting the glue-down stripline with a knife and replacing it with one having altered dimensions. This project gives another example of how stripline parameters can be varied to optimize a circuit: homemade variable capacitors.

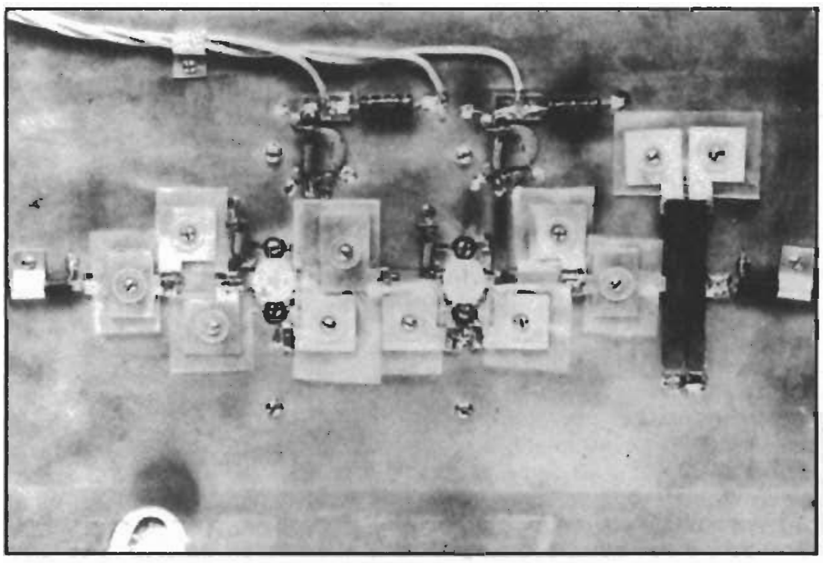

Circuit Details

See Fig 1. The two-stage amplifier uses an MRF641 (driver) and an MRF646 (final).[2] They were chosen primarily because of their availability in the surplus market. These are controlled-Q devices with a combined minimum gain of 12.6 dB. Although the output stage is rated at 45 W, typical saturated output with a 13.6-V collector supply is about 60 W (including drive power).

The home-brew compression-type variable capacitors, shown in Fig 2, dictated the overall design. I made them as small as possible (to minimize selfinductance), but kept them large enough to provide the necessary capacitance. The nominal $^5/_8$-inch-square compression plate results in a maximum capacitance of about 10 pF. The material used for the capacitor plates is 0.031-inch-thick Reynolds sheet aluminum, which is available at most hardware and home-supply stores. [Hobby store sheet brass works well in this application, is also commonly available, and is far easier to solder; it is a good alternative to aluminum—Ed.] The capacitor dielectric material is 2.7-mil polyethylene from a Dow Ziploc® heavy-duty freezer bag. Although I have tested this dielectric at a much higher voltage than it is subjected to in this application (at a high-impedance point in a vacuum-tube VHF amplifier), double thickness is used as a safeguard. The insulator between the adjustment screw and compression plate is made from plastic polymer. Most clear plastics used for miscellaneous household applications are of this type. I use plastic from a box used in packaging a tube of Grumbacher acrylic artists paint. Glass-epoxy insulation will work also, but at the higher-voltage

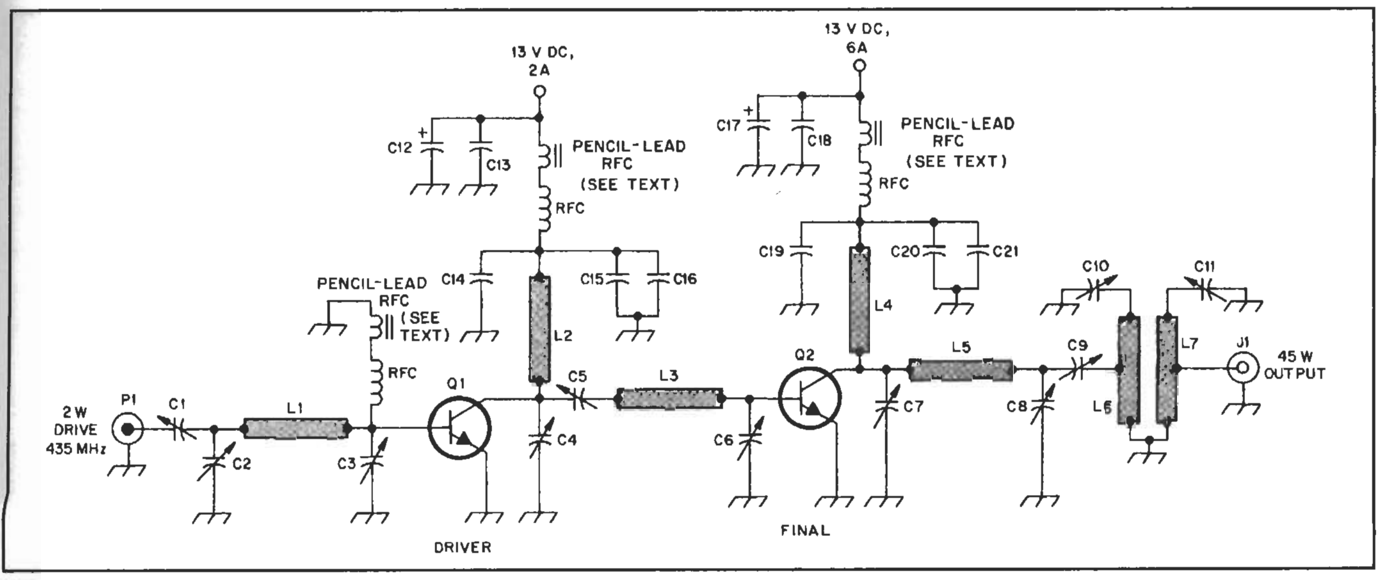
13 V DC,
2A
13 V DC,
6A
C12
C13
PENCIL-LEAD
RFC
(SEE TEXT)
RFC
C17
C18
PENCIL-LEAD
RFC
(SEE TEXT)
RFC
PENCIL-LEAD
RFC
(SEE
TEXT)
RFC
C14
C15
C16
C19
C20
C21
L2
L4
C10
C11
2 W
DRIVE
435 MHz
P1
C1
L1
C2
C3
Q1
C5
L3
C4
C6
Q2
L5
C7
C8
C9
L6
L7
J1
45 W
OUTPUT
DRIVER
FINAL

Fig 1—Schematic of the 435-MHz 45-W amplifier.

C1-C11—Homemade capacitors; see text and Fig 2 for details.
C12, C17—4.7-µF, 35-V electrolytic.
C13, 15, 16, 18, 20, 21—0.001-µF, 50-V disk ceramic.
C14, 19—0.1-µF, 50-V disk ceramic.
J1—Female BNC connector.
L1-L5—Glue-down striplines; see text and Fig 3 for details.
L6, L7—Air-gap striplines; see text and Fig 4 for details.
P1—Male BNC connector.
Q1—MRF641.
Q2—MRF646.
RFC—10 turns, no. 26 enameled copper, air-wound, $^3/_{16}$-in diameter.

applications (especially at C10 and C11) there will be some loss and component heating.

Once I settled on the capacitor configuration, the input and output stripline characteristics were determined experimentally for optimum matching to Q1 and Q2. The related glue-down stripline and pad details are shown in Fig 3. The L2 L4 $^1/_8$-inch-wide striplines act as RF chokes, making the inductance associated with the subsequent disc-ceramic capacitors uncritical. The pads include a number of the capacitor stators (C1, C4, C5, C6, C7 and C9). Foils which contact stator-adjustment screws are reamed out to approximately $^1/_8$-inch diameter to isolate the no. 2-56 screw from the stator. I did this with a large drill ($^3/_8$-inch). Screw contact with the bottom foil may also cause adjustment irregularities, so the foil was reamed appropriately. The stator plates must be completely deburred to prevent puncturing the plasticbag dielectric.

The amplifier output is coupled through C9 to a 1:1 inductively coupled, double-tuned filter. This 50-ohm filter consists of two striplines that are mounted parallel to each other and close together above the base PC board, as shown in Fig 4. Their close positioning establishes the necessary mutual coupling. Each line is tuned to resonance with a variable capacitor.

As indicated in Fig 1, dc-supply decoupling is identical for both stages. The 0.1-µF disk capacitors eliminate a bothersome low-frequency instability mode. RF chokes following these capacitors include two-turn sections with pencil-lead cores to minimize the possibility of parasitics. The 4.7-µF capacitors eliminate low-frequency power-supply instability problems.

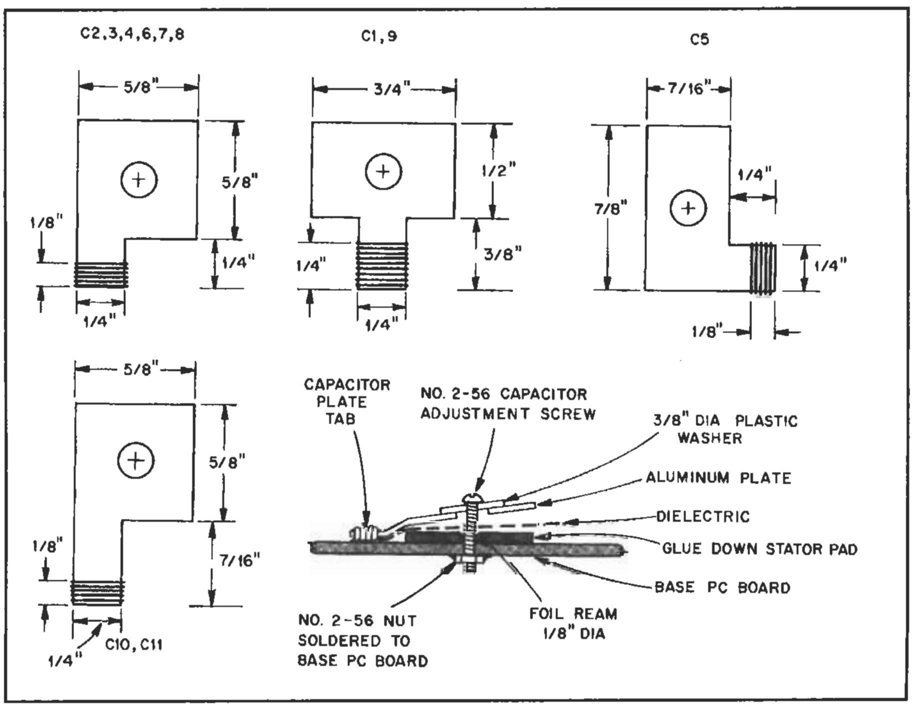
C2,3,4,6,7,8
5/8"
5/8"
1/8"
1/4"
1/4"
C1,9
3/4"
1/2"
3/8"
1/4"
1/4"
C5
7/16"
7/8"
1/4"
1/4"
1/8"
5/8"
5/8"
1/8"
7/16"
1/4"
C10, C11
CAPACITOR
PLATE
TAB
NO. 2-56 CAPACITOR
ADJUSTMENT SCREW
3/8" DIA PLASTIC
WASHER
ALUMINUM PLATE
DIELECTRIC
GLUE DOWN STATOR PAD
BASE PC BOARD
NO. 2-56 NUT
SOLDERED TO
BASE PC BOARD
FOIL REAM
1/8" DIA

Fig 2—Variable-capacitor construction details. The "rotor" plates are made from 0.031-inch Reynolds aluminum sheet stock and polished with 320-grit sandpaper to ensure that no burrs are left that may puncture the dielectric material. Plate center holes are $^3/_{16}$-inch diameter. Connections are made to the capacitors by means of no. 26 tinned copper wire, tightly wrapped (and crimped in a vise) around the tab on the capacitor and soldered to the appropriate circuit board connection. Make these connections carefully to ensure good electrical contact. See text for details on the dielectric and washer materials.

Assembly

The transistor-lead configurations are slightly modified to accommodate the glue-down striplines. Both the base and collector leads are bent up into an S shape so that the lead will overlap the stripline by approximately 1/16 inch when the transistor and stripline are mounted flush with the base PC board. The leads are only 0.005-inch thick and can be bent easily with long-nose pliers. Care must be taken to minimize stress at the lead-ceramic junc-

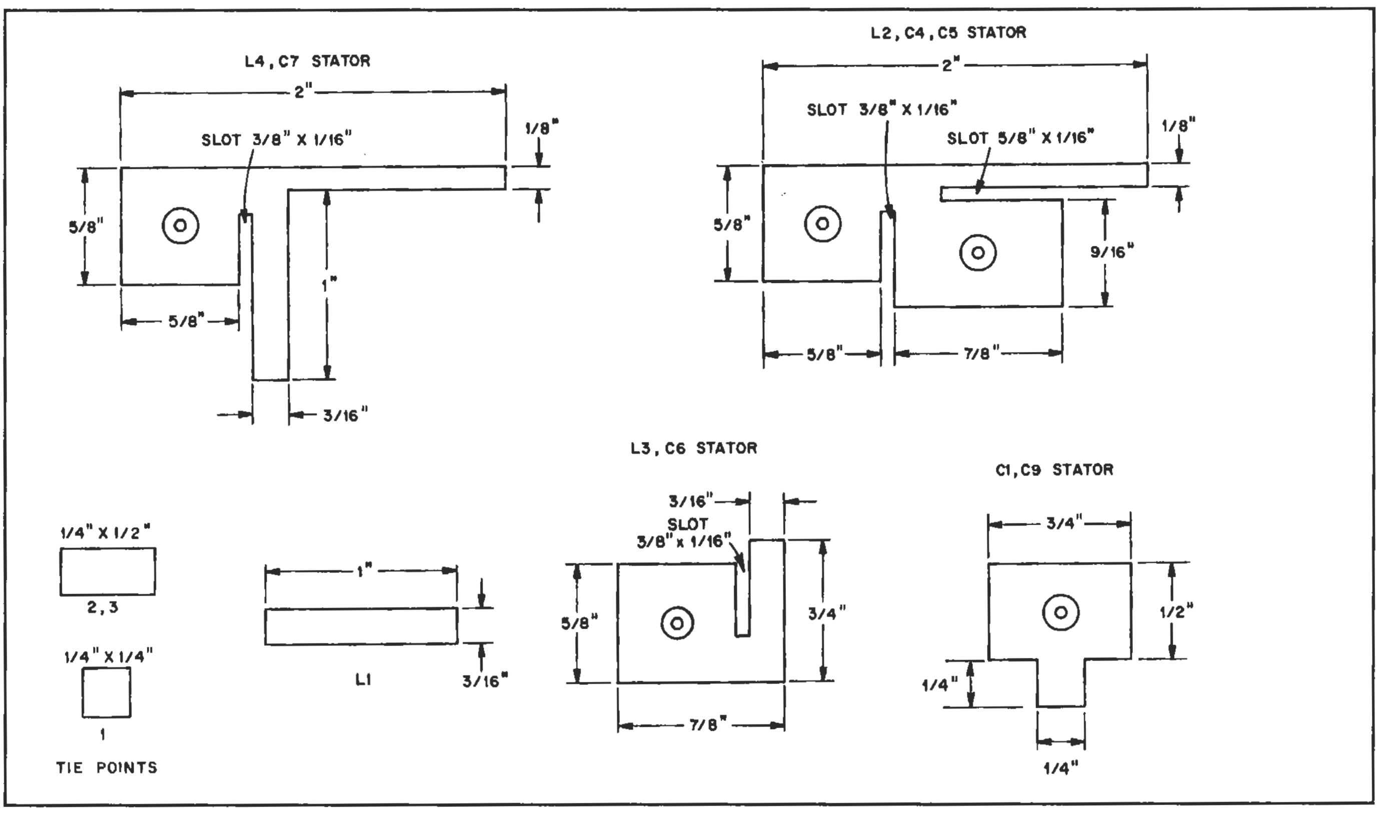
L4, C7 STATOR
2"
SLOT 3/8" X 1/16"
1/8"
5/8"
1"
5/8"
3/16"
L2, C4, C5 STATOR
2"
SLOT 3/8" X 1/16"
SLOT 5/8" X 1/16"
1/8"
5/8"
9/16"
5/8"
7/8"
L3, C6 STATOR
3/16"
SLOT
3/8" x 1/16"
5/8"
3/4"
7/8"
C1, C9 STATOR
3/4"
1/2"
1/4"
1/4"
1/4" X 1/2"
2, 3
1/4" X 1/4"
1
TIE POINTS
1"
L1
3/16"

Fig 3—Glue-down stator pad construction details. Slots are cut with a hacksaw and then polished with a pattern file. The material is 0.062-inch-thick double-sided glass-epoxy circuit board material.[6] After cutting and filing, the pads must be polished with fine steel wool to remove any burrs.

tions. Also, 3/32 inch of each emitter lead must be cut off with scissors to allow room for the glue-down pads.

Details of the mounting arrangement are shown in Fig 5. My transistor mounting method was dictated by the dimensions of an old 1/8-inch-thick aluminum panel (with many miscellaneous holes), which I pressed into service as a heat sink. First, I mounted each transistor on separate aluminum subbases. I then fastened the sub-bases to the heat sink. Thermally conductive compound is used between the transistors and the subbases and between the sub-bases and the heat sink. This arrangement permits removal of the heat sink without disturbing the transistor mountings when I need access to the bottom PC-board foil. Note that the base PC board cutout includes notches at the base and collector leads of the transistors to prevent shorting to the common foil.

Detailed assembly like that shown in Fig 6 is easy once the transistors are mounted. Simply smear glue on the pads and fit them into position. Start by positioning L1, followed by the C1 stator, then C1, C2 and C3. Drill holes in the base PC board for the no. 2-56 capacitor-adjustment screws as you go along. At the same time, solder the no. 2-56 nuts on the reverse side (using a screw to hold them in position). Bend the capacitor tabs to make certain the plates will be flush with the stator, inserting the capacitor dielectric as the final step. Make certain the dielectric has been pushed in between the capacitor plates as far as possible (maximum safety margin), and use an awl to puncture the dielectric at the adjustment-screw position. The adjustment screw can then be easily started through the dielectric.

Operation

UHF transistors are great devices, but they can be damaged permanently with just one voltage transient. For example, an accidental short by a probe during a troubleshooting exercise, or the discharge of a probe capacitor at the transistor's base lead can be disastrous. Another common problem is instability that causes selfoscillation and excessive collector voltage (no insta-

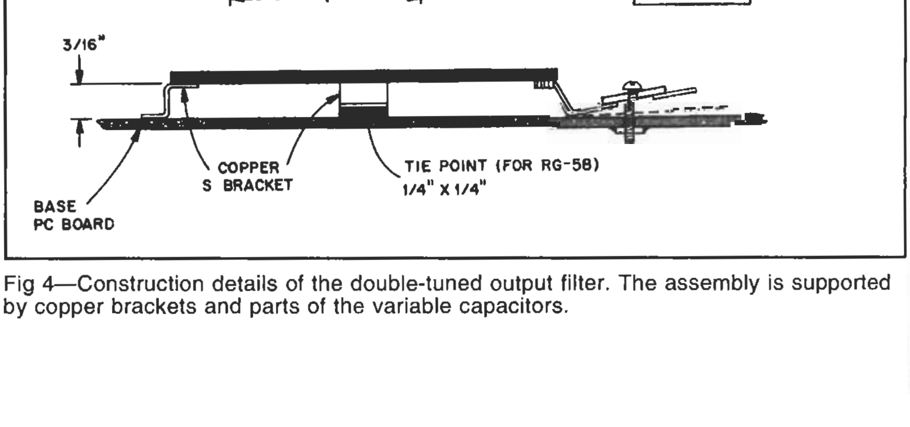
C9 STATOR
L6, L7
2 X 3/16 X 1/16-INCH
PC BOARD MATERIAL
COPPER
S BRACKET
C10, C11
SEE DETAIL
FIG 2
1/16"
1"
3/16"
COPPER
S BRACKET
TIE POINT (FOR RG-58)
1/4" X 1/4"
BASE
PC BOARD

Fig 4—Construction details of the double-tuned output filter. The assembly is supported by copper brackets and parts of the variable capacitors.

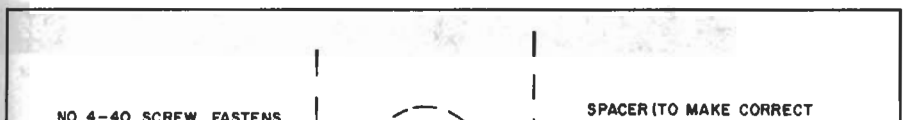
NO. 4-40 SCREW FASTENS BASE PC BOARD TO TRANSISTOR MOUNTING BASE AND HEAT SINK. (ONE ON EACH SIDE)
SPACER (TO MAKE CORRECT SPACING BETWEEN TRANSISTOR LEADS AND MOUNTING BASE)
Q1 OD CUT-OUT
EMITTER
BASE
COLLECTOR
BASE CUT-OUT NOTCH
COLLECTOR CUT-OUT NOTCH
GLUE-DOWN PAD L1
Q1
GLUE-DOWN PAD L2
BASE PC BOARD
Q1 MOUNTING SUB-BASE 5X2X1/8-INCH ALUMINUM
HEAT SINK 9X6X1/8-INCH ALUMINUM

Fig 5—Pictorial of transistor mounting method. The same scheme is used for both the driver and final transistors.

bility modes are evident in the amplifier described here). These possible failure modes can be avoided by limiting the collector supply to 5 V (or less) during tune-up, as practically all instability modes will be evident at this lower potential (particularly if the input drive is varied over a wide range). Following alignment at 5 V, it is surprising how little readjustment is required when the collector potential is increased to 13 V. (If you can't reduce your power supply voltage, you may want to use a 2N3055 as a pass transistor, driving it with an LM317T regulator [Radio Shack 276-2041 and 276-1778, respectively].) This low-voltage test procedure is mentioned in Motorola Application Note AN-548A.[3]

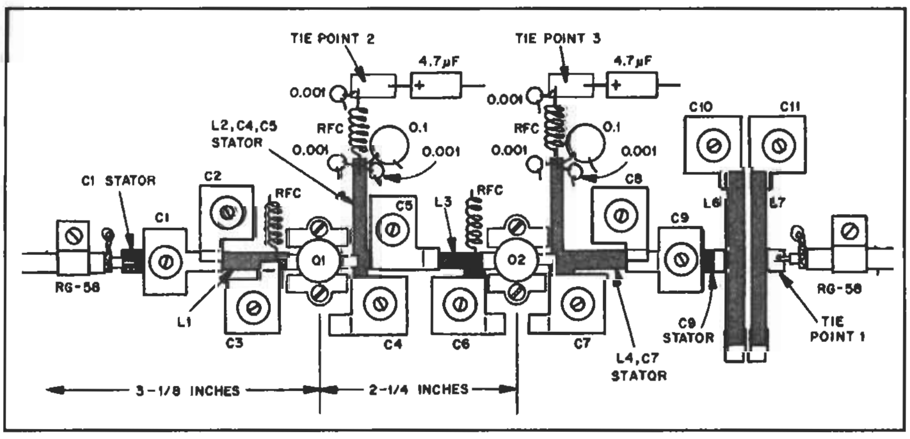
TIE POINT 2
4.7µF
TIE POINT 3
4.7µF
0.001
0.001
L2,C4,C5 STATOR
RFC
0.1
0.001
0.001
RFC
0.1
0.001
0.001
C10
C11
C1 STATOR
C2
RFC
C5
L3
RFC
C8
C1
Q1
Q2
C9
L6
L7
RG-58
RG-58
L1
C3
C4
C6
C7
C9 STATOR
TIE POINT 1
L4,C7 STATOR
3-1/8 INCHES
2-1/4 INCHES

Fig 6—Circuit board layout of the completed amplifier. Begin construction by measuring the center-to-center distance from Q1 to Q2 and then mount the transistors, working away from them as you mount the other components. Use great care when mounting and soldering the transistors into the circuit to avoid damaging them. The PC board measures 9½ in. x 6 in., and is made from double-sided, glass-epoxy material (Meshna PCB28—see note 1).

Initial Tune-Up

The dummy load/power-measuring assembly and narrow-band peak detector pick-off assembly used in tune-up are the same as the ones that I used with my 15-W transmitter.[4] The dummy load consists of ten 1/2-W resistors connected in parallel and arranged in a circle to minimize self-inductance. Power measurements were made by calibrating dummy-load temperature rise with a dc input (ambient air-temperature conditions, no forced-air cooling). This information is then compared to the temperature rise of the dummy load resulting from the RF output of the amplifier. The "thermistor" used for measuring the temperature rise is a 1N34A (reading the reverse resistance).

Under key-down conditions during alignment, I place a blower at the chassis end to assist the limited heat-sink capabilities of the amplifier. Also, the dummy load requires forced-air cooling, as it becomes very hot in about two minutes under keydown conditions.

Miscellaneous lengths of RG-8/M and RG-58 were used as attenuators to increase power measurement capability of this system.[5] In testing the 45-W amplifier, I used a total cable attenuation of 6.4 dB. This increases the power-measurement capability by 4.4 times.

With a 5-V collector supply and 2 W of drive, first tune the input capacitors for maximum Q1 collector current, then tune the remaining capacitors for maximum power output. The output should be about 8 W. Increase the collector supply to 13 V; minor readjustment will probably be required. My final measurements showed an output of 48 W.

Summary

This is a well-behaved amplifier. After completing an experimental version on the breadboard, I built a new assembly using the data in this article. Alignment of this final model was completed without any problems, and operation is flawless. The glue-down stripline technique is very reliable in this application, and is simple to use.

Notes

[1] J. Reed, "A Simple 435-MHz Transmitter," *QST*, May 1985, pp 14-18, 45.

[2] These transistors are available from RF Parts Company, 1320-16 Grand Aye, San Marcos, CA 92069, tel 619-744-0728.

[3] This note is reproduced in *Motorola RF Device Data* (fourth edition, first printing, 1986) on pages 6-61 through 6-65. Thus book is available from Motorola Literature Distribution, P0 Box 20912, Phoenix, AZ 85036, tel 602-994-6561. Cost is $4.75 (plus 15 percent of the total for shipping).

[4] See note 1.

[5] My 435-MHz measurements indicated that the RG-8/M I used has an attenuation of 9.0 dB per 100 feet (specified as 7.5 dB per 100 feet at 400 MHz); the RG-58 has 15.3 dB per 100 feet (specified as 12.0 dB per 100 feet at 400 MHz).

[6] Available from John J. Meshna, Jr. Inc, 19 Allerton St, Lynn, MA 01904, tel 617-595-2275.

By Tarmo Tammaru, WB2TMD **From *QST*, May 1982**

A Solid-State 6-Meter Linear Amplifier You Can Build

Does your 6-meter signal need a little more punch? This amplifier will solve your problem—without breaking the bank!

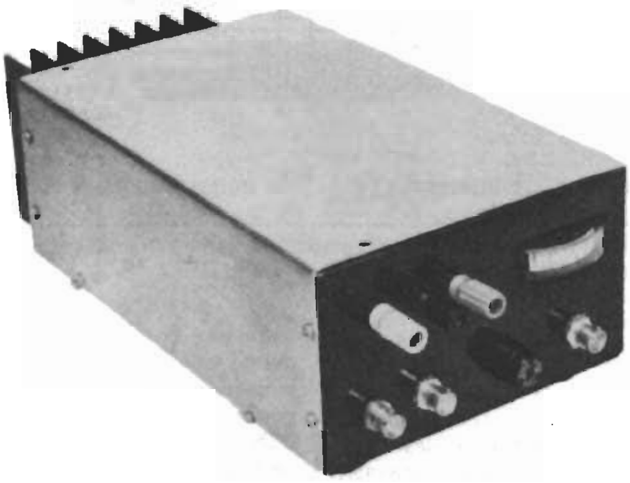

The new breed of 8-to 10-W, solid-state vhf transceivers offer compactness and relative freedom from TVI. However, there are times when 10 W will not do the job. This amplifier can correct that situation. I used it during the September vhf QSO party and was able to work every station I heard, except one.

The transistor used in the amplifier is a Motorola MRF-492, which is designed for use as an fm amplifier in the "vhf low" public-service band. In this service, the nominal output is rated at 75 W with a 12.5-V collector supply. The manufacturer's data sheet shows a maximum output of 120 W at 16 V. Hence, in ssb service, I felt it would be safe to operate at the 100-W PEP level from a 13.5-V supply. Although no numbers are given, the MRF492 is claimed to have "load mismatch capability at high line and rf overdrive."[1]

Circuit Description

The amplifier schematic diagram is shown in Fig. 1. Since I do not believe in "reinventing the wheel," the output matching network was taken directly from the Motorola data sheet. This network, incorporating L2 and L3 in the signal path, should provide excellent harmonic rejection. The network transforms the 50-Ω load impedance to the optimum transistor load of $0.6 + j1.0\ \Omega$. Note that this unit, like most rf amplifiers, does not have a 50-Ω output impedance. We simply adjust the output network to obtain the rated power into a 50-Ω load. Fig. 2A is an amplifier output spectral display. While the harmonic suppression is good, additional filtering is required before using the unit on the air. A simple, 5-pole, lowpass filter was placed at the amplifier output, and the resulting spectrum is shown in Fig. 2B. The low-pass filter circuit is given in Fig. 3.

Because the data sheet does not give values for input impedance in Class B service, the input circuit was derived basically by trial and error (the second version worked). The T matching network transforms the highly capacitive input impedance of approximately 1 Ω to a 50-Ω nonreactive load for the exciter.

In a Class B amplifier it is necessary to forward bias the base-emitter junction. There are basically two ways of doing this: the shunt diode method, and the emitter-follower method. The latter method undoubtedly results in lower inter-modulation products, but I have destroyed too many transistors using that method to try it with a $20 device. Forward bias is achieved by using the shunt regulator diode Dl. This diode must be connected to the same heat sink as the MRF492, Q1. The voltage across Dl is adjusted to give a Q1 collector current of 100 mA with no drive applied. Do not attempt to set the bias point by measuring the base voltage of Q1 — use a milliammeter in the collector supply lead. I used an unmarked, surplus, 15-A rectifier diode for Dl. Almost any diode will work, including power Zeners. I prefer a stud-mounted diode because it is easy to place in contact with the heat sink. Just be sure the polarity is such that the cathode is connected to the case. Several Motorola application notes specify a 1N4997 for this application. Unfortunately, this is a press-fit device. In any case, the heat sink must be grounded.

Parts Procurement

All parts, with the possible exception of an adequate heat sink, are readily available. The MRF492 has recently been advertised by Westcom[2] and Semiconductor Surplus.[3] RFC1 is a Nytronics shielded ferrite choke, designed to reduce coupling

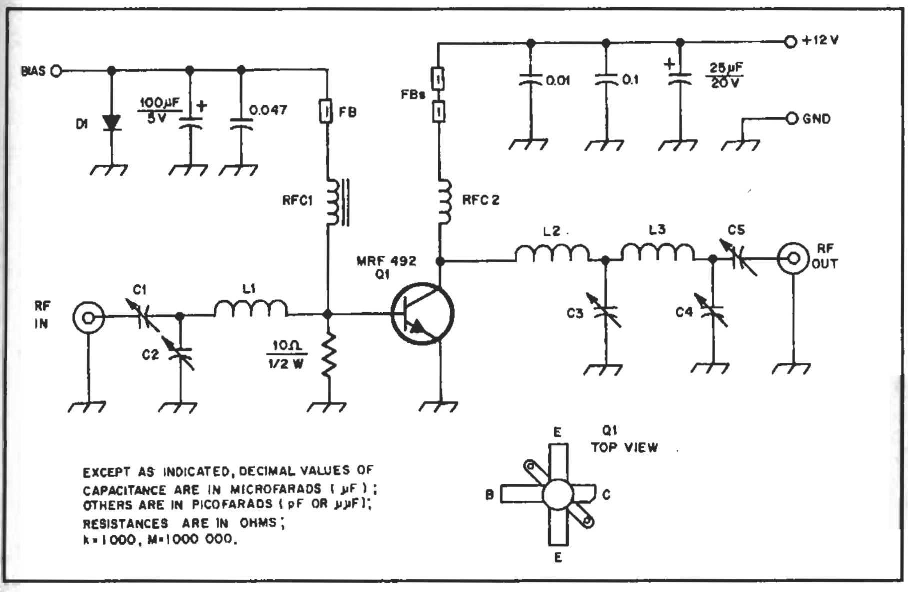
BIAS
D1
100µF 5V
+
0.047
FB
RFC1
FBs
0.01
0.1
+
25µF 20V
+12V
GND
RFC2
L2
L3
C5
RF OUT
MRF 492 Q1
C1
L1
RF IN
C2
10Ω 1/2 W
C3
C4
E
Q1 TOP VIEW
B
C
E
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig. 1 — The schematic diagram of the amplifier circuit board. Decimal value capacitors are disc ceramic and polarized capacitors are electrolytic. Resistors are 10% carbon type.

C1 — 9-to 180-pF mica compression trimmer, Arco 463 or equiv.
C2, C4, C5 — 50- to 380-pF mica compression trimmer, Arco 465 or equiv.
C3 — 80- to 480-pF mica compression trimmer, Arco 466 or equiv.
D1 — 15-A, 50-V rectifier diode (see text).
Fb — Ferrite bead, Radio Shack 273-1571 assortment or equiv.
L1 — 2 t. of no. 14 bare copper wire, $^{13}/_{32}$-inch ID, $^{3}/_{16}$-inch long.
L2— Half loop of no. 14 bare copper wire, $^{19}/_{32}$ inch high, $^{13}/_{32}$ inch long (see photograph).
L3— 2 t. of no. 14 bare copper wire, $^{13}/_{32}$-inch ID, $^{1}/_{4}$-inch long.
Q1 — Motorola rf power transistor, MRF492.
RFC1 — 6.8 µH, Nytronics SWD 6.8 or equiv. (see text).
RFC2 — No. 16 enameled wire, close wound over full length of a 330-Ω, 2-W carbon resistor.

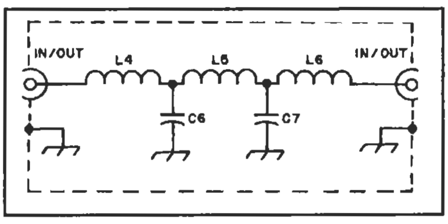

Fig. 3 — A simple low-pass filter for use with the 6-meter amplifier. The design information for this filter was taken from the 1982 ARRL *Radio Amateur's Handbook*. If the filter is mounted outside the amplifier cabinet, it should be enclosed in a metal box for shielding.

C6, C7 — 82-pF, 1000-V sliver-mica capacitor.
L4, L6 — 4 t. of no. 14 enameled wire. $^{3}/_{8}$-inch ID, $^{3}/_{4}$-inch long with $^{1}/_{8}$-inch leads (approximately 0.1 µH).
L5 — 6 t. of no. 14 enameled wire, $^{3}/_{8}$-inch ID, $^{1}/_{2}$-inch long with $^{1}/_{8}$-inch leads (approximately 0.2 µH).

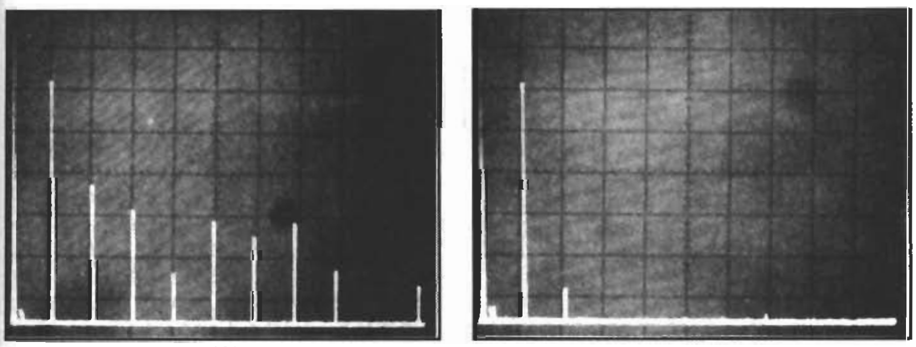
(A) (B)

Fig. 2 — The amplifier output matching network provides some harmonic suppression (A). To further reduce harmonics, a simple low-pass filter should be used (B). With the filter, all harmonics and spurious signals are more than 60 dB below the carrier, thus meeting current FCC spurious emission requirements for commercial equipment. In these displays, the carrier level has been reduced by means of a notch filter to avoid analyzer overload. Vertical divisions are each 10 dB and the horizontal divisions are each 50 MHz. These measurements were made in the ARRL lab.

between the base and collector circuit. The J. W. Miller 9250-682 should be as good. A toroidal inductor of 1 to 10 µH would also be suitable, since it is across an impedance of 1 Ω. Arco trimmers should be available from your local jobber. I used a surplus heat sink, measuring 3$^{1}/_{2}$ × 4$^{1}/_{2}$ × 1$^{3}/_{4}$ inches, which is quite adequate for ssb.[4] Anything much smaller will require a fan. Remember, it must be flat on one side. A replacement heat sink for any of the 80- to 100-W 2-meter amplifiers would be ideal.

The Circuit Board

All the components shown in Fig. 1 mount on, or through, the circuit board (Fig. 4). Use double-sided glass epoxy board and cut away narrow channels to separate the various land (foil) areas. To form the channels, use a steel ruler and a sharp utility knife to cut through the foil. Next, pull a hot soldering iron along the strip of foil to be removed. You should see it curl up and away from the board. Start at one end, and pry up with the knife if it gets stuck. You will have to drill two holes, slightly more than $^{1}/_{2}$ inch in diameter, through the board to provide clearance for Q1 and D1. Use a small round file to make the two clearance areas for the flanges of Q1. Connect the ground foils on the top and bottom of the board together by placing short lengths of wire through holes drilled in the board. Place the wires close to the transistor emitter tabs and near each corner of the board. Use pieces of the same type of wire you used to make the inductors, and drill the holes just large enough to pass the wire. Solder the wire on both sides.

Assembly

Make sure you have all the parts on hand before you solder anything to the board. This will ensure that everything fits properly. To keep the inductors in place, I drilled holes through the board and stuck the ends of the wire into the holes. You will have to remove the foil from around these holes on the bottom of the board to keep from shorting the inductors to ground. A large drill bit works fine for this. Center the board on the heat sink, and drill the mounting holes for Q1 and D1. Although you may be able to use screws and nuts to attach Q1 and D1 (depending on the heat sink design), a better method is to tap the holes for the desired screw thread. Mount the board to the heat sink near the corners, shimming it with washers to the same depth as the Q1 mounting flange. Drop Q1 in from the top of the board. The tabs of Q1 should rest just above the top of the board when the flange is in contact with the heat sink. Do not put any upward pressure on the transistor tabs. Never cut or file an rf power transistor body — they contain berillium oxide. It is highly toxic in powdered form and could be inhaled or absorbed through the skin.

Fasten Q1 to the heat sink before sol-

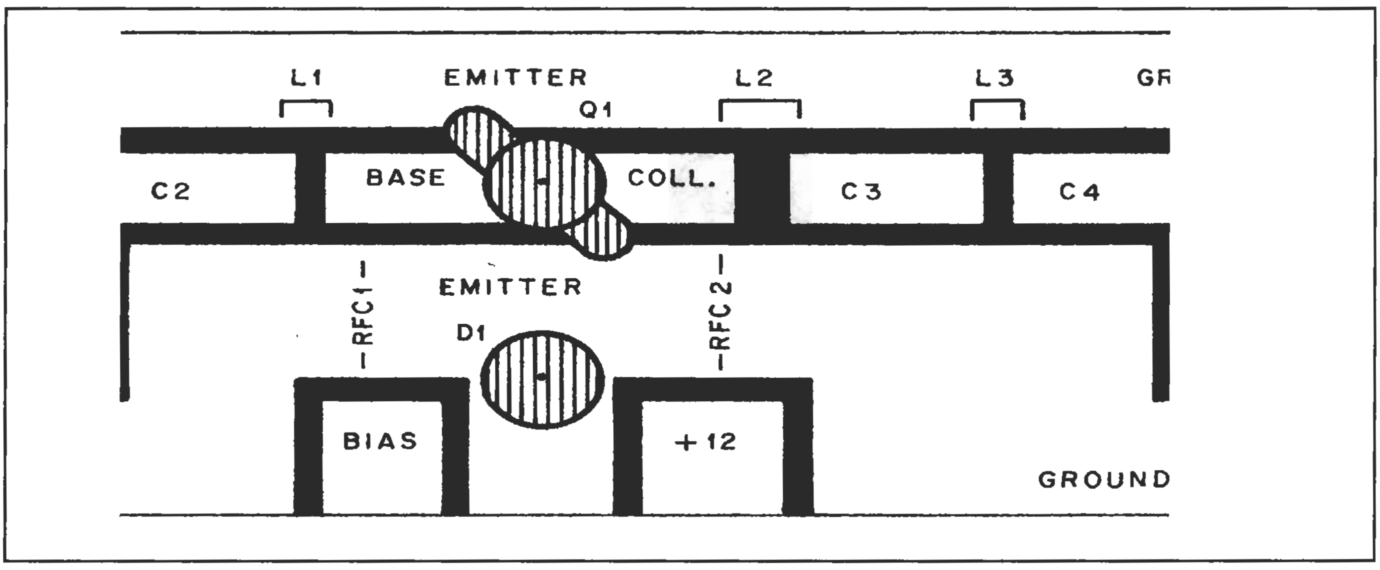

Fig 4—Full-scale circuit-board pattern and parts-placement guide for the 6-meter amplifier. Black represents those areas where copper has been removed by cutting or etching. The 10-ohm base resistor is connected between the base and ground foils, and is positioned next to RFC1.

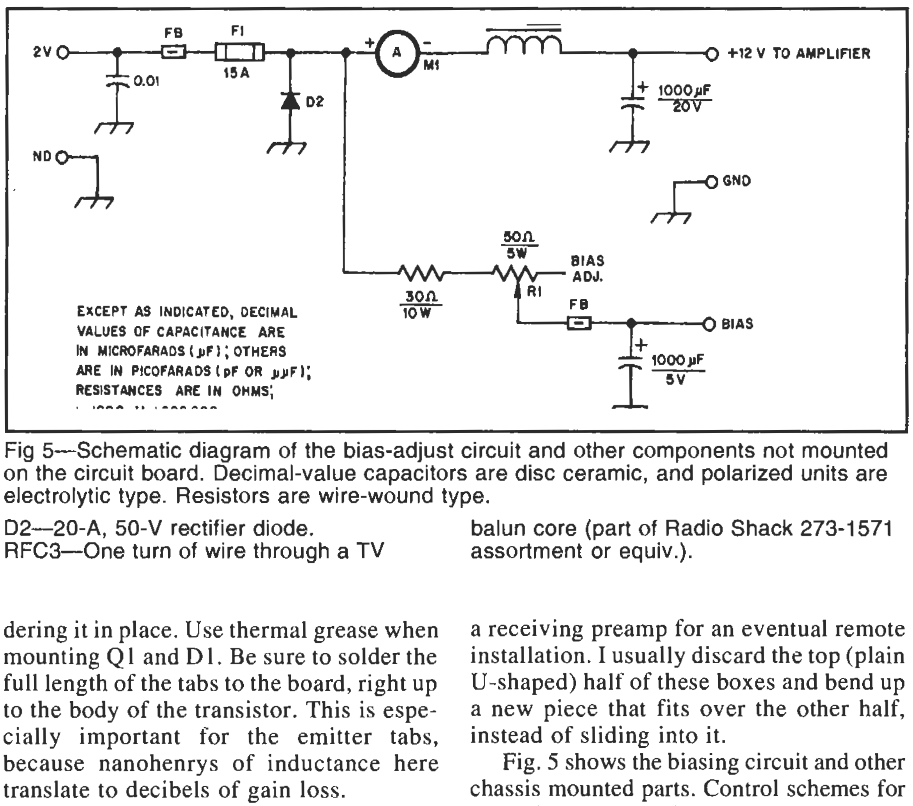
2V
FB
F1
15A
0.01
D2
A
M1
+12 V TO AMPLIFIER
1000μF 20V
ND
GND
50Ω 5W
BIAS ADJ.
R1
30Ω 10W
FB
BIAS
1000μF 5V
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (μF); OTHERS ARE IN PICOFARADS (pF OR μμF); RESISTANCES ARE IN OHMS;

Fig 5—Schematic diagram of the bias-adjust circuit and other components not mounted on the circuit board. Decimal-value capacitors are disc ceramic, and polarized units are electrolytic type. Resistors are wire-wound type.

D2—20-A, 50-V rectifier diode.

RFC3—One turn of wire through a TV balun core (part of Radio Shack 273-1571 assortment or equiv.).

dering it in place. Use thermal grease when mounting Q1 and D1. Be sure to solder the full length of the tabs to the board, right up to the body of the transistor. This is especially important for the emitter tabs, because nanohenrys of inductance here translate to decibels of gain loss.

Break off the little tabs on the sides of the mica compression trimmers. Bend the ends of the mounting terminals with pliers to aid in soldering the trimmers to the board. In doing this, be sure the adjusting screw will not touch the board when it is all the way down. C2, C3 and C4 should be oriented so that the screw is connected to the ground end of the capacitor. For C1 and C5 let the screw end be the 50-Ω side.

Putting It All Together

I installed the circuit board and heat sink on the back wall of a $3^1/_2 \times 6 \times 10$-inch Minibox. This is much larger than necessary, but does provide room for relays and a receiving preamp for an eventual remote installation. I usually discard the top (plain U-shaped) half of these boxes and bend up a new piece that fits over the other half, instead of sliding into it.

Fig. 5 shows the biasing circuit and other chassis mounted parts. Control schemes for use with a transceiver are described by Kapplin[5] and Ridpath[6] (also see the ARRL *Radio Amateur's Handbook*). D2 does not have to be heat sinked. It only has to last long enough to blow the fuse in case the wrong supply polarity is inadvertently applied. On the other hand, the 50-Ω bias adjust control (R1) gets quite warm, and should be mounted to the enclosure. In wiring the bias circuit, attach one wire from R1 to D1 and another wire from D1 to the bias point on the circuit board; all other connections are made to the circuit board. If R1 were connected to the circuit board, with a strap going to D1, you would almost certainly damage the transistor if the strap to D1 broke.

Alignment

It is necessary to set the bias before aligning the rf circuits. For an initial check, disconnect the wire between D1 and the circuit board (R1 is connected to D1). Apply power and measure the voltage across D1. It should be possible to lower it to 0.65 V, or less, by varying R1. Set R1 for the lowest voltage and connect the wire from D1 to the circuit board. Now reduce the resistance of R1 until the collector current is 100 mA. Be sure you are measuring collector current and not total supply current (you have about 300 mA flowing into D1).

For aligning the trimmers you will need, at least, an SWR indicator and a 50-Ω dummy load. You will, of course, not know how much power you are getting unless you also have a wattmeter. Connect the amplifier output to a dummy load and apply about 1 W of drive to the input through the SWR indicator. Adjust C1 and C2 for an SWR of 1:1. With CS snug, but not tight, adjust C3 and C4 for minimum collector current. With the SWR indicator or wattmeter connected between the output and the dummy load, adjust all the trimmers for maximum output. If more than one setting gives the same power, pick the one that corresponds to the lowest collector current. Slowly bring up the drive and keep readjusting all the trimmers. With 8 W of drive you should get an output of 100 W. The collector current will be approximately 15 A. Readjust the input for a 1:1 SWR at full power. This may, or may not coincide with maximum gain and you may have to compromise.

During testing you might consider using a harmonica to generate a multitone signal to reduce the average power dissipation (keep the speech processor off). If at any time the heat sink gets so hot you cannot hold your hand on it for five seconds, let it cool off. As a rough indication of power output, I found that at 100 W, a

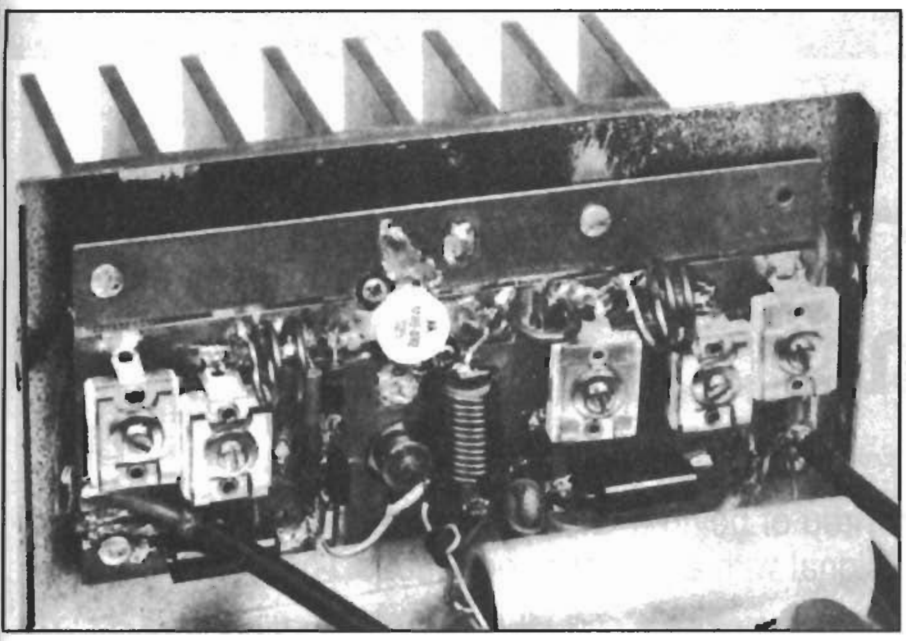
Interior view of the 6-meter amplifier showing the parts arrangement used by the author. The input matching network is at the right of the photo.

Drake DL 300 dummy load become too hot to hold after two minutes of key-down operation. To prevent rf burns, do not touch the dummy load while rf power is applied.

To maintain linearity, adjust the output circuits for 100 W, even if you are going to operate at 75 or 80 W. I evaluated the amplifier linearity by making two-tone IMD measurements with a spectrum analyzer. At a PEP output of 100 W, the third-order products were 25 dB below the PEP. At 80-W PEP they were down 30 dB. In over 100 contacts, I have not received any adverse comments on the signal quality.

Notes

[1] Motorola Inc., *Motorola Rf Data Manual*, 2nd ed. (Phoenix, AZ: Motorola Inc., 1980), p. 6-15.

[2] Westcom, 1320 Grand Ave., San Marcos, CA 92069.

[3] Semiconductor Surplus, 2822 N. 32nd St., No. 1, Phoenix, AZ 85008.

[4] mm = inches x 25.4.

[5] S. Kapplin, "Boots for QRP Rigs," *QST*, July 1981, pp. 15-20.

[6] I. Ridpath, "T-R Switching with PIN Diodes," *QST*, March 1981, pp. 19-21.

By Wilson Hoag, WA5OLT **From *QST*, April 1982**

Build A 6-Meter "Mini-Lini"

Tired of being a QRP station? Boost your low-level signal with a pair of sweep tubes.

Is low power enough for equality on 6 meters? Perhaps, but it has long been my contention that some stations are more equal than others! One of the great features of 6 meters is the ability to make contacts with just a few watts when the band is open. As solar activity begins to decline, band openings will be shorter and less frequent. This amplifier will add punch to your QRP signal.

The current offerings of solid-state 6-meter transceivers and transverters provide operating ease and good signal quality, but at the expense of rf power output. This, coupled with the paucity of linear amplifiers, resulting from the infamous amplifier ban, has caused an overall reduction in the average power output on 6 meters.

When I acquired my first solid-state rig (an IC-551) some time back, my tendency was to use its scan feature to locate a signal and then work the station with a 240-watt tube-type rig. A minor panic occurred when the tube rig "rolled over and died" midway through the '79-'80 F2 season. The Mini-Lini resulted from that panic.

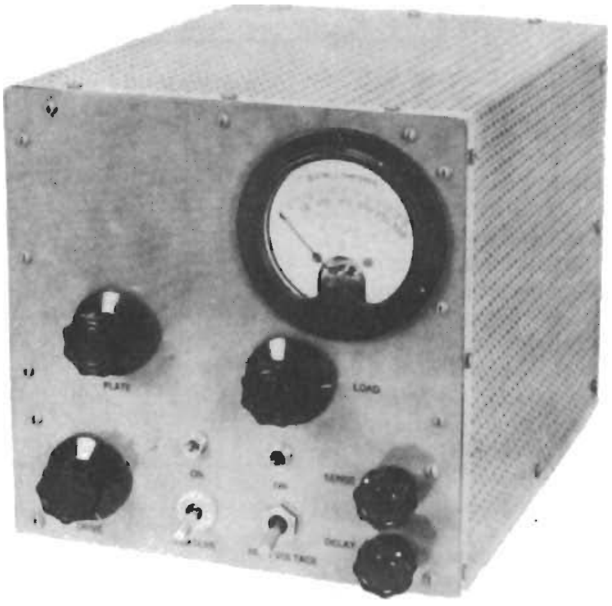

Design Approach

The amplifier was designed to be compatible with common power supplies, use a readily available tube, provide push-to-talk operation with the IC-551 and have approximately the same rf output as most tube-type transceivers. Fortunately, a design incorporating most of these features had been developed by Ed White, WA5RIA.[1] The major changes I made were in the methods of switching and bias adjustment. The tube used is the 6JB6, which is inexpensive, available and "happy" at 50 MHz. The biasing arrangement, an idea borrowed from Doug DeMaw,[2,3] allows the tubes to be matched for safe parallel operation. This amplifier provides about 50 watts of output when driven by a 10-watt exciter. More than one year of almost daily use has been trouble free and productive (42 states, including KH6 and KL7, plus JA and several Europeans).

Circuit Details

The circuit is conventional and typical of many low-band amplifiers. The original design used a capacitive input, but I thought that the parallel-tuned input method shown in Fig. 1 would provide more output when used with other QRP drivers. This was proven in practice: The amplifier delivers about 20 watts of output when driven by an IC-502. The potential problem of operating unmatched tubes in parallel has been avoided by biasing each tube individually. Component values shown will allow coverage of the entire 6-meter band.

The output network will not reject harmonic energy, and the second harmonic, as measured in the ARRL lab, was less than 40 dB below the fundamental (Fig. 2). This works out to a healthy 5 mW at 100 MHz. Most fm receivers in the neighborhood will detect this easily. FCC requirements for a commercial amplifier at this frequency are to have all spurious radiation at least 60 dB below the fundamental.

To meet this requirement (and to prevent a lot of RFI complaints) a 7-pole Chebyshev low-pass filter was built from data given in the 1982 *Radio Amateur's Handbook*, pages 6-11 and -12. Details of a filter with 0.01-dB of ripple and a 60-MHz cutoff frequency are given in Fig. 3. The coils can be wound on toroids, or be air-core types. Amidon T44-10 (or larger) cores should be adequate. The number of turns required will depend on core size or coil diameter, and can be calculated from data given on pages 2-12 and 2-30 of the *Handbook*. This filter reduced the second harmonic output to more than 65 dB below the fundamental (completely gone for all practical purposes). See Fig. 4.

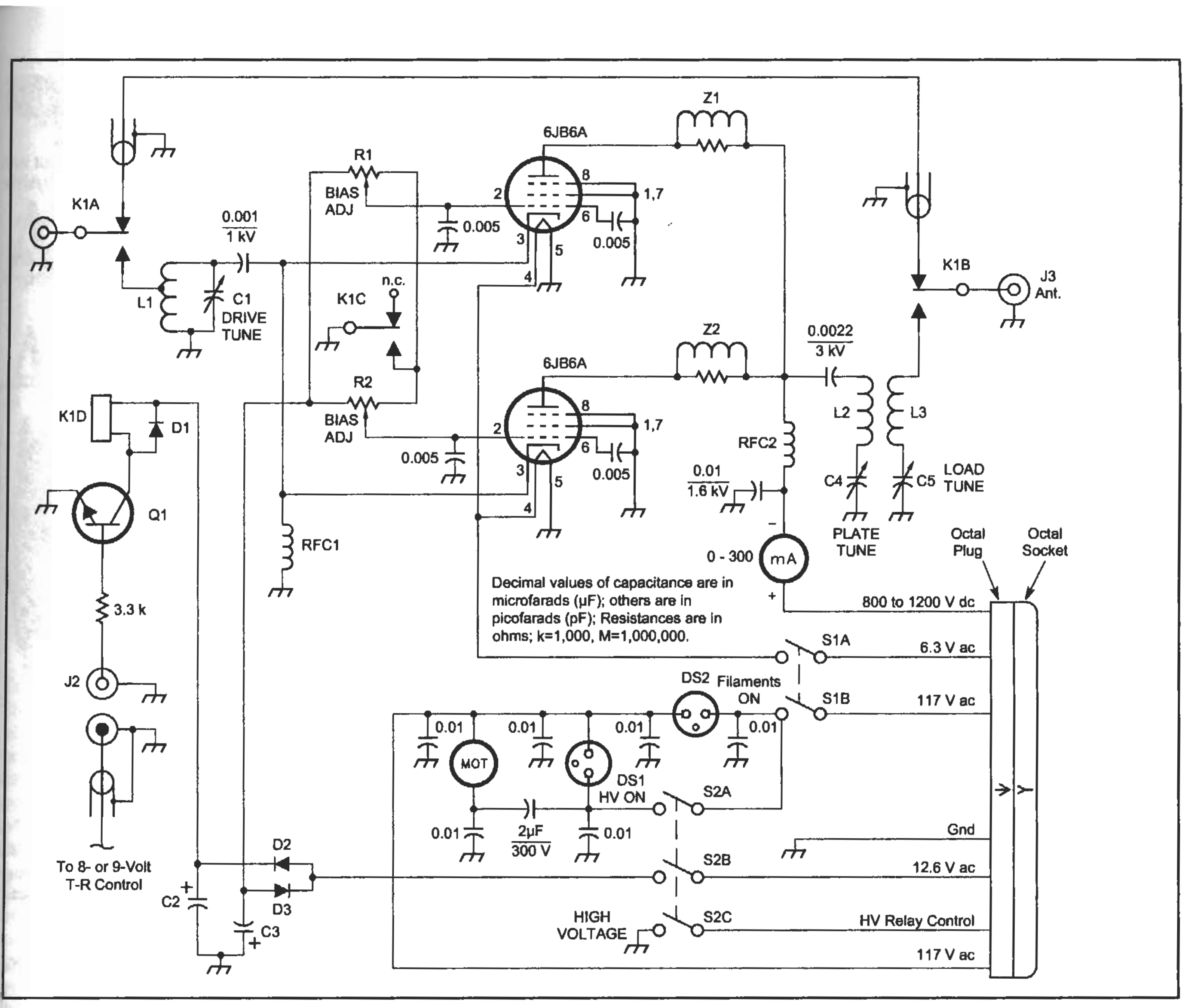
Z1
6JB6A
R1
BIAS ADJ
K1A
0.001
1 kV
2
8
1,7
6
0.005
3
5
0.005
4
K1B
J3
Ant.
L1
C1
DRIVE TUNE
K1C
n.c.
Z2
0.0022
3 kV
R2
BIAS ADJ
6JB6A
L2
L3
K1D
D1
0.005
2
8
1,7
6
3
5
0.005
4
RFC2
0.01
1.6 kV
C4
C5
LOAD TUNE
Q1
PLATE TUNE
Octal Plug
Octal Socket
0 - 300
mA
RFC1
3.3 k
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.
800 to 1200 V dc
S1A
6.3 V ac
J2
DS2 Filaments ON
S1B
117 V ac
0.01
0.01
0.01
0.01
MOT
DS1
HV ON
S2A
0.01
2µF
300 V
0.01
Gnd
D2
S2B
12.6 V ac
To 8- or 9-Volt T-R Control
C2
D3
C3
HIGH VOLTAGE
S2C
HV Relay Control
117 V ac

Fig. 1 — Schematic diagram of the 6-meter linear amplifier. The tube filaments can be wired In series, eliminating the need for a separate 6.3.V supply.

C1 — 50-pF miniature air variable, Cardwell 148-4 or equlv.
C2, C3 — 220-µF electrolytic, 50 V.
C4 — 25-pf air-variable, wide-spaced Hammarlund HF30X or equlv.
C5 — 100-pF air variable, Johnson 149-5 or equiv.
D1 — 1N914.
D2, D3 — 1 A, 50 V.
J1, J3 — SO-239.
J2 — Phono Jack.
K1 — 3pdt 12-V dc coil, rf type preferred, or KRP14DG.
L1 — 6 turns no. 14 enameled wire, 1/2-inch ID × 1-inch long, tapped at approximately 1 1/2 turns.
L2 — 5 turns no. 12 enameled wire, 1 1/2 inch ID × 2 1/2 inches long.
L3 — 2 turns no. 12 enameled wire, 1 1/2 inch ID × 1/2-inch long.
mA — 0-300 mA dc meter.
MOT — 117-V fan motor.
P1 — Chassis-mount octal plug.
Q1 — NPN power transistor, TIP31 or equivalent, (RS no. 276-2017).
R1, R2 — 10-kΩ, 2-W potentiometer.
RFC1 — 300-µH choke with ferrite bead on the ground lead.
RFC2 — 83 turns no. 28 enameled wire on a 1/2-inch diameter ceramic form. The coil is 2 inches long.
Z1, Z2 — 1 turn no. 16 enameled wire on 47-Ω, 2-W resistor.

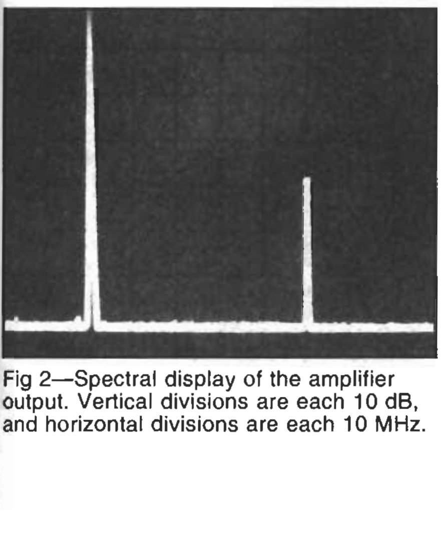

Fig 2—Spectral display of the amplifier output. Vertical divisions are each 10 dB, and horizontal divisions are each 10 MHz.

Construction

It should be possible to duplicate this amplifier for $40 or less, depending on the status of your junkbox. Oddly enough, the tube sockets proved to be the most difficult component to locate, since they are not a common catalog item. The 6JB6 uses a 9-pin NOVAR socket, and the best source turned out to be a shop that specialized in TV replacement parts. Tuning capacitor C4 is a junkbox item of questionable parentage. Any spacing greater than about 0.060 inch (mm = inches × 25.4) should be okay. RADIOKIT of Greenville, New Hampshire would be a good source for all of the variable capacitors.

My amplifier is constructed on an 8 × 10 × 2 1/2 inch chassis (BUD AC-1418) for a base, with 8 × 8-inch end panels. The panels were flanged with 1/2 × 1/2-inch aluminum channel for attachment of a cane-metal cover for safety and TVI protection. Component placement is not critical, but all component leads should be as short as possible. The resulting layout was determined largely by some mid-project design changes aimed at making the unit compatible with a newly acquired IC-502. Visible on the lower-right front panel are potentiometers that were reserved for control of a planned rf-operated T-R relay.[4]

The tube sockets are modified slightly to ground the necessary tube elements. Copper washers, cut from flashing copper, fit inside

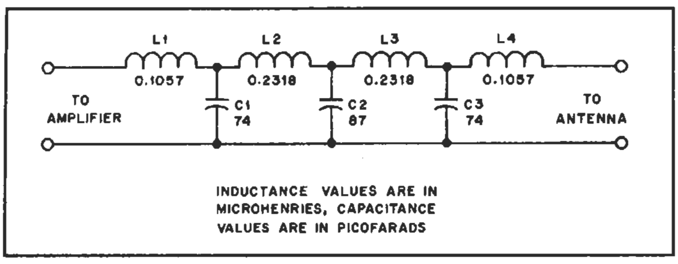
L1
0.1057
L2
0.2318
L3
0.2318
L4
0.1057
TO AMPLIFIER
C1 74
C2 87
C3 74
TO ANTENNA
INDUCTANCE VALUES ARE IN MICROHENRIES, CAPACITANCE VALUES ARE IN PICOFARADS

Fig 3—Schematic diagram of a 7-pole Chebyshev low-pass filter. Capacitors are silver-mica units, combined in parallel or series to obtain the design values. The text has information about winding the inductors.

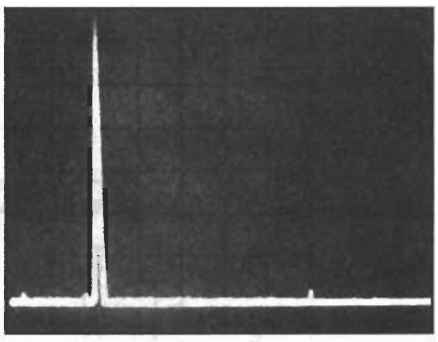

Fig 4—Spectral display of the amplifier output with a 7-pole Chebyshev filter. Vertical divisions are each 10 dB, and horizontal divisions are each 10 MHz.

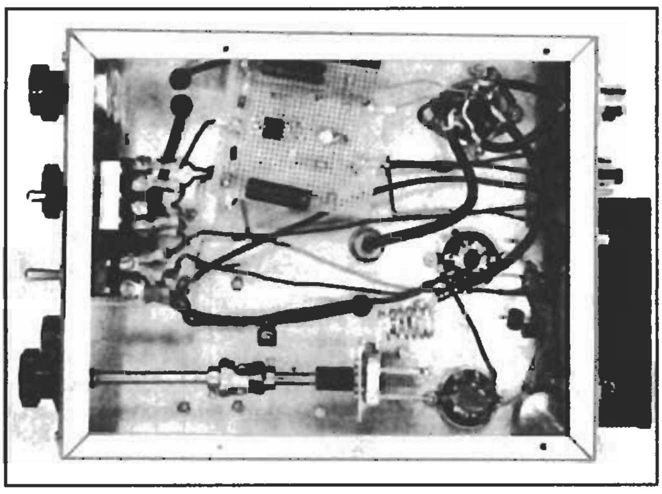

Fig 5—Bottom view of the amplifier chassis. Note the washers on the tube sockets, used to provide a ground connection for the appropriate pins.

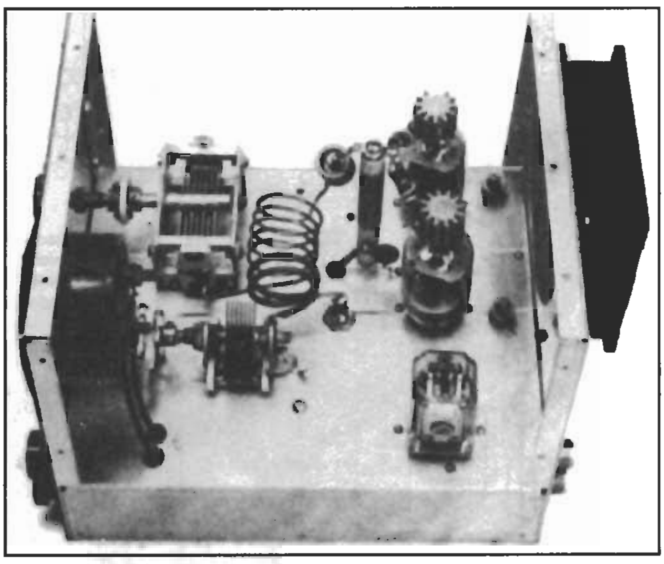

Fig 6—Top view of the Mini-Lini chassis. Large heat-dissipating plate caps and a fan help maintain cool operation of the tubes.

the pins on the underside of the sockets. Pins 1, 4, 7, 8 and 9 are bent over and soldered to the washer. The washers are grounded to the chassis with short pieces of no. 14 wire on opposite sides of the socket. Pins 2 and 6 of both tubes are bypassed to ground by connecting 0.005-µF disc-ceramic capacitors from these pins to the copper washers (Fig. 5). The control grids are raised above dc ground while providing a path for rf return. It also permits us to apply a dc bias to the control grid. This will establish the class of operation and cut the tubes off during receive, if desired.

Coils L1, L2 and L3 are wound with solid TW-insulated house wire from the local hardware store. I stripped the insulation from the wire before winding L1 and L2, but left it on L3 to prevent accidental contact with L2. The inside diameters shown are more the result of available cylindrical shapes than any electrical calculation. Coil spacing was adjusted, with the tubes in place, using a dip meter to ensure resonance at the proper frequency.

Heat-dissipating plate caps and a small fan provide cooling for the tubes. See Fig. 6. No thermal distress has been evident under any operating condition. The fan can be wired through S1 or S2, depending on whether cooling is desired during standby operation. Power connections to the fan should be isolated from the chassis.

A possible modification, shown on the schematic diagram, involves the use of a 3pdt relay for K1. The ground legs of R1 and R2 can be wired through one of the normally open contacts. This lifts the potentiometers above ground in the receive mode, applying full bias voltage and cutting off the tubes.

The bias voltage source and relay driver are "hard wired" on a small piece of perf board. The signal for T-R switching is applied to J2. This + 8-V signal is obtained from pin 6 of the IC-551 accessory socket. For use with other rigs, a 9-V battery, wired through a foot switch, works well. Current drain on this battery is low. An rf-operated T-R relay could be used in place of the directly keyed one described in this article. S2 disables the relay driver by removing the 12.6-V ac source in the standby position to permit straight-through operation.

Tune-Up and Operation

Initial tune-up is simple and ordinary. Connect a dummy load and the power supply to the appropriate jacks. Any high-voltage power supply that has an output of 750- to 1200-V dc should be satisfactory. Turn on S1 and allow the heaters to warm up for a minute or more. Switch S2 to turn on the other voltages. Actuate K1 to ground the bias resistors, R1 and R2. Adjust R1 and R2 to obtain 15 mA of idling current for each tube (30 mA total).

Connect the exciter to the amplifier through an SWR indicator. Actuate K1 and apply a small amount of drive. Adjust the position of the tap on L1 for minimum SWR. Be sure to remove all voltages each time you move the tap position!

Next, apply drive to the amplifier and adjust C1, C4 and C5 for maximum power output. For those lacking a wattmeter, an SWR indicator set in the FORWARD position can be used in the line between the amplifer and the dummy load.

Conclusion

The Mini-Lini should be ideal for those looking for more output from their solid-state rigs. Typical operating parameters are 50 watts out for 10 watts of drive, with 1050-V dc on the plates. Reports of signal quality have been complimentary. Enhance your "equality" and come join the fun on 6! I would be happy to answer any questions about the amplifier — enclose an s.a.s.e., please.

Notes

[1] The design was taken from personal correspondence with Ed White, WA5RIA.

[2] D. DeMaw, "Some Ground Rules for Sweep-Tube Linear-Amplifier Design," *QST*, July 1968, p. 30.

[3] D. DeMaw, "Some Thoughts About TV Sweep Tubes," *QST*, Feb. 1980, p. 11.

[4] D. DeMaw and J. Rusgrove, "An RF-Sensed Antenna Change-over Relay," *QST*, Aug. 1976, p. 21.

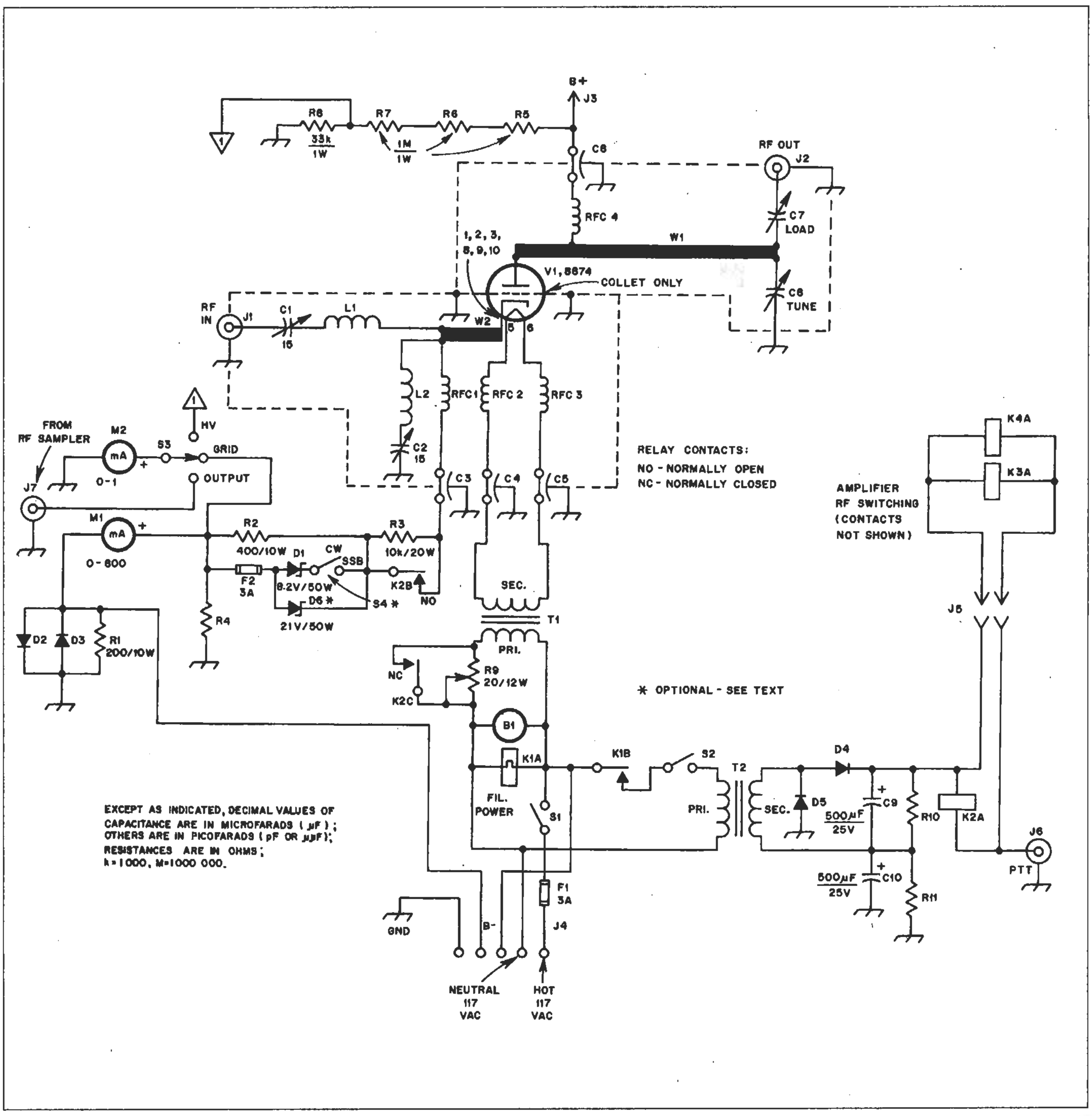
B+
J3
R8
33k
1W
R7
1M
1W
R6
R5
C8
RF OUT
J2
RFC 4
W1
C7
LOAD
1, 2, 3,
8, 9, 10
V1, 8874
COLLET ONLY
C6
TUNE
RF
IN
J1
C1
15
L1
W2
5
6
HV
L2
RFC1
RFC 2
RFC 3
C2
15
FROM
RF SAMPLER
M2
mA
0-1
S3
GRID
OUTPUT
J7
C3
C4
C5
RELAY CONTACTS:
NO - NORMALLY OPEN
NC - NORMALLY CLOSED
K4A
K3A
AMPLIFIER
RF SWITCHING
(CONTACTS
NOT SHOWN)
M1
mA
0-600
R2
400/10W
D1
CW
SSB
8.2V/50W
F2
3A
D6 *
S4 *
21V/50W
R3
10k/20W
K2B
NO
SEC.
T1
R4
D2
D3
R1
200/10W
PRI.
NC
K2C
R9
20/12W
* OPTIONAL - SEE TEXT
B1
K1A
FIL.
POWER
S1
K1B
S2
T2
PRI.
SEC.
D4
D5
C9
500µF
25V
C10
500µF
25V
R10
K2A
R11
J5
J6
PTT
F1
3A
J4
GND
B-
NEUTRAL
117
VAC
HOT
117
VAC

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig. 1 — Schematic diagram of the amplifier.

B1 — 50-ft^3 (1.4-m^3)/min blower, Ripley Sk2754-2A or equiv.
C1, C2 — Air-variable capacitor, 15 pF, E.F. Johnson 189-0565-001, 160-0107-001 or equiv.
C3-C5, incl. — Feedthrough capacitor, 500 pF, 300 V.
C6-C8, incl. — Homemade "flapper" capacitor. Details of construction in text and Fig. 3.
C9, C10 — Electrolytic capacitor, 500 µF, 25 V.
D1 — 50-watt, 8.2-volt Zener diode, IR Z-3307-C or equiv.
D2-D5, incl. — 1-A, 1000-PIV diode, 1N4007 or equiv.
D6 — 50-watt, 21-volt Zener diode (optional — see text).
F1, F2 — 3AG fuses.
J1 — Chassis mount BNC female connector, UG-1094/U.
J2 — Chassis mount N female connector, UG-58A/U.
J3 — High-voltage connector, Millen 37001.
J4, J5 — Power connectors, as available.
J6, J7 — RCA phono jacks.
K1 — Time-delay relay, 90 second, normally open contact, Amperite 115N090T.
K2 — Control relay, 28-volt coil, 1-A 4pdt contacts.
K3, K4 — Coaxial relays equipped with suitable connectors. K4 should have N connectors, K3 may be BNC or N.
L1 — 3½ turns no. 16 enam. wire, ¾-inch (19 mm) long, ¼-inch (6 mm) diameter.
L2 — 1½ turns no. 16 enam. wire, ⅝-inch (16 mm) long, ¼ inch (8 mm) diameter.
M1 — 1-mA meter movement with shunt to provide 600-mA full-scale deflection.
M2 — 1-mA meter movement with shunts to provide 90-mA (grid current) and 3-kV (plate voltage) full-scale deflection.
R4 — Grid-current shunt.
RFC1 — 10 turns no. 18 enam. wire, close wound, ¼ inch (6 mm) diameter.
RFC2, RFC3 — 10 turns no. 16 enam. wire, close wound, ¼ inch (6-mm) diameter.
RFC4 — 5 turns no. 16 wire, one inch (25 mm) long, ¼ inch (6 mm) diameter.
S1 — Toggle switch, spst.
S2 — Toggle switch, spst.
S3 — Rotary switch, single pole, three position.
S4 — Toggle switch, spst (optional, see text).
T1 — Filament transformer, 6.3-volt, 3-A, Stancor P-6466 or equiv.
T2 — Transformer, 12.6 volts, 1 A.

By Stephen J. Powlishen, K1FO From *QST*, October 1979

A Grounded-Grid Kilowatt Amplifier for 432 MHz

Stable, linear operation for tropo or moonbounce DX... that's the end result of this project.

In the last few years I've built several high-power 432-MHz amplifiers that used tubes from the 4CX250 family. While they worked well in Class C, their performance when biased for linear operation left something to be desired. My previous experience with grounded-grid triode amplifiers on 2 meters was so good that I decided to try the same approach on 70 cm. An Eimac 8874 high-mu triode was selected for this design and a crude prototype was built in a few evenings. After the design was verified, the amplifier described in this article was built. It is stable, compact and delivers over 500 watts output while requiring only a high-voltage supply and a source of about 25 watts of drive. The amplifier has been trouble free in over three years of heavy usage.

Circuit Description

A schematic diagram of the 432-MHz kilowatt is given in Fig. 1. W1 is a half-wavelength stripline which is tuned and loaded by C6 and C7 respectively. Plate choke RFC4 is connected at the approximate electrical center of the plate line. C8 functions as the plate-bypass capacitor. The half-wavelength cathode line is comprised of W2, L2 and C2. L1 and C1 serve to match the tube input impedance to the amplifier 50-ohm input. As the grid is grounded for dc as well as rf, D1 is used to develop operating bias at the cathode. R3 is switched in to supply near-cutoff bias during standby periods. M1 is used solely to monitor plate current in the high-voltage supply negative-return Lead. M2 is switched to read grid current, high voltage and relative output. The latter function is by means of an external line sampler.[1] With the exception of the multimeter functions the metering and bias circuits are similar to those in a 220-MHz amplifier.[2]

Separate coaxial relays attached to the

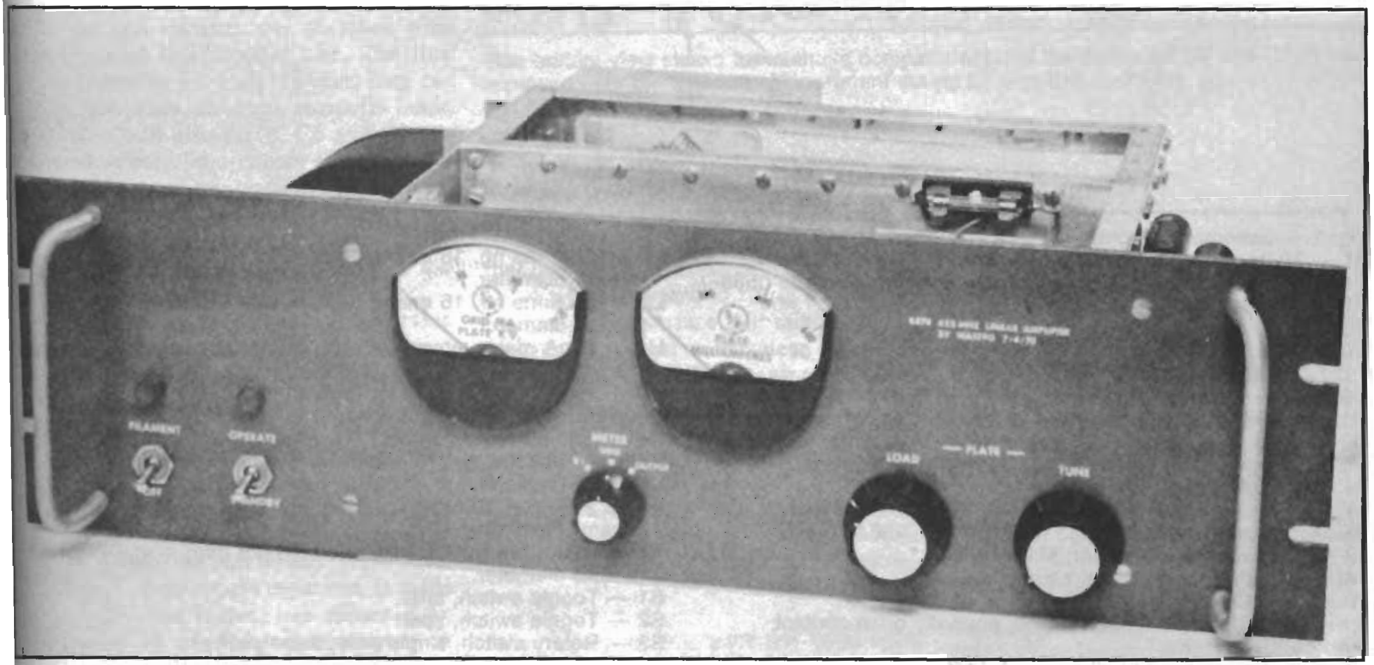

The high-power uhf amplifier. The toggle switches control filament power and standby/operate functions respectively. Multimeter function is selected with the switch located between the meters, while the plate tuning and loading controls are at the right. Modern knobs and homemade meter faces give the amplifier a commercial appearance.

input and output terminals allow the amplifier to be switched in and out of the line in a manner popular with hf amplifiers. Time-delay relay K1 prevents the amplifier from being switched into service for 90 seconds after the tube heater is energized, allowing the element to reach operating temperature. A normally closed contact of K2 applies full voltage to the heater during standby periods. The voltage is reduced during operation as recommended by the manufacturer.

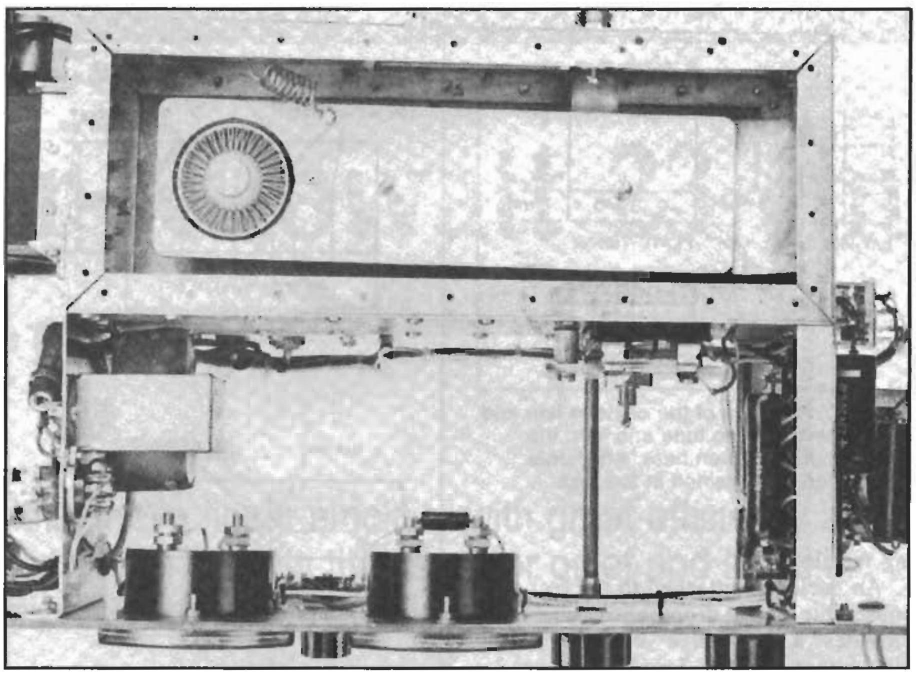
Top view of the amplifier, with the plate compartment cover removed. The tube, plate line (W2) and RFC4 may be seen at the top of the photo. Note the large number of holes drilled in the plate compartment to receive the cover hold-down screws. A tight seal is required to prevent rf and air leaks.

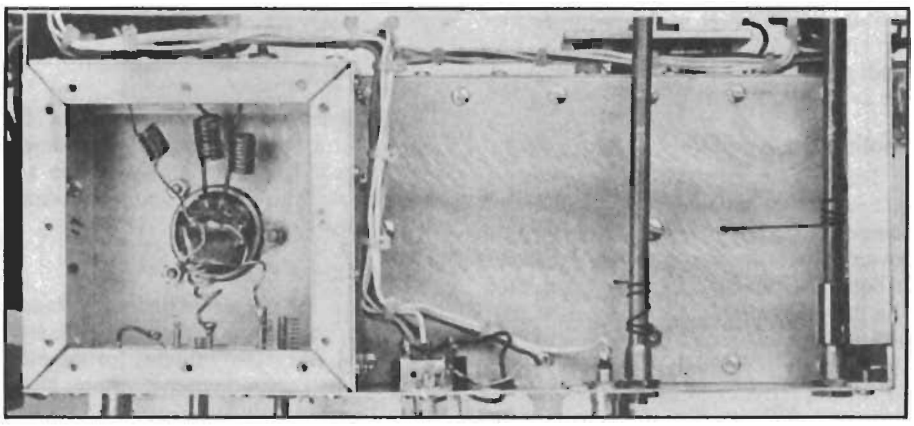
This bottom view shows the cathode compartment and the shafts for C6 and C7. A cover is placed over the cathode compartment during tuneup and operation.

Construction

Plate and cathode-compartment construction is from 0.032-inch (0.8-mm) thick aluminum sheet attached to $^1/_2$-inch (13-mm) aluminum angle stock. Some angle stock may be anodized, giving the surface a dull appearance. This material must be lightly sanded to remove the anodized metal, which is a poor conductor. Holes are drilled in the angle stock to allow attachment of the covers; these are tapped for no. 4-40 screws. Details of the 10.5 × 4 × 3-inch (267 × 102 × 76-mm) plate compartment may be seen in the top view photo. Construction of the cathode compartment is similar, and may be seen in the photo of the underside. It measures 4 × 4 × $1^3/_4$ inches (102 × 102 × 44 mm). The aluminum brackets holding the rf enclosures to the front panel also serve as end covers for the compartments. Compartment spacing from the panel is four inches (102 mm). A $5^1/_4$ × 19-inch (133 × 483-mm) rack panel is used.

The plate line was fabricated from a piece of $^1/_{16}$-inch (1.6-mm) thick brass. Fig. 2 gives detailed information for making the line. In addition to brass, lines were made from copper, both unplated and silver plated, with no discernible difference in efficiency. Double-sided G-10 printed circuit board would probably work as well. Best thermal stability was obtained with the unplated solid-copper line. The line is supported by 1.5-inch (38-mm) long ceramic insulators, although standoffs made of Teflon will also serve. C6 and C1 are made from beryllium-copper sheet. Details of their construction appear in Fig. 3. These "flappers" are moved with fishing line which is tied to $^1/_4$-inch (6.4-mm) fiber shafts. These shafts may be seen in the underside view.

The anode collet (Eimac no. 008294) is secured to the bottom of W1 with standard 60/40 solder. Use no. 4-40 screws and nuts to hold the collet in place during the soldering operation. The grid collet (Eimac no. 882931) is attached to the chassis with eight no. 4-40 machine screws and nuts. A poor ground connection for the grid will greatly increase the amplifier drive requirements or make the unit totally inoperative.

C8, the plate-bypass capacitor, is made from two brass plates, one mounted on either side of the plate compartment. A 0.005-inch (0.13-mm) thick piece of Teflon sheet is used for the dielectric material. While this Teflon thickness may seem inadequate, it is rated at 1000 volts per mil (0.03 mm) thickness. It is necessary to coat the dielectric with Dow Corning type DC-4 silicone grease to fill in any imperfections in the surface that might allow a leakage path and subsequent capacitor breakdown. This silicone grease has dielectric properties similar to Teflon* A no. 8 (4-mm) brass screw is used to hold the plates in place, and also acts as the high-voltage feedthrough terminal. A 3/8-inch (10-mm) diameter washer was sliced from a Teflon rod and used to center the screw in the hole. Fig. 4 gives details of the remain-

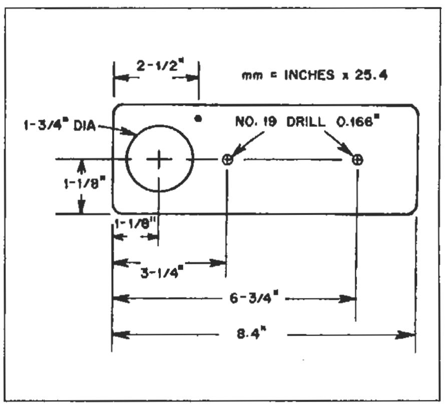

Fig 2—Dimensions of the plate line are given here. The line may be constructed from $^1/_{16}$-inch (1.6-mm) thick copper or brass. Corners of the line should be filed to give a $^3/_{16}$-inch (5-mm) radius.

*Alternatively, an Erie 2498-001-X5U0-102M 1000-pF 4-kV feedthrough capacitor may be used. This component is available from ARCOS, P.O. Box 546, East Greenbush, NY 12061.

By E.R. "Chip" Angle, N6CA **From *QST*, March 1985**

A Quarter-Kilowatt 23-cm Amplifier

Imagine, a linear amplifier with great efficiency and long-term stability that is super quiet and small in size. Sounds like HF? Not exactly...

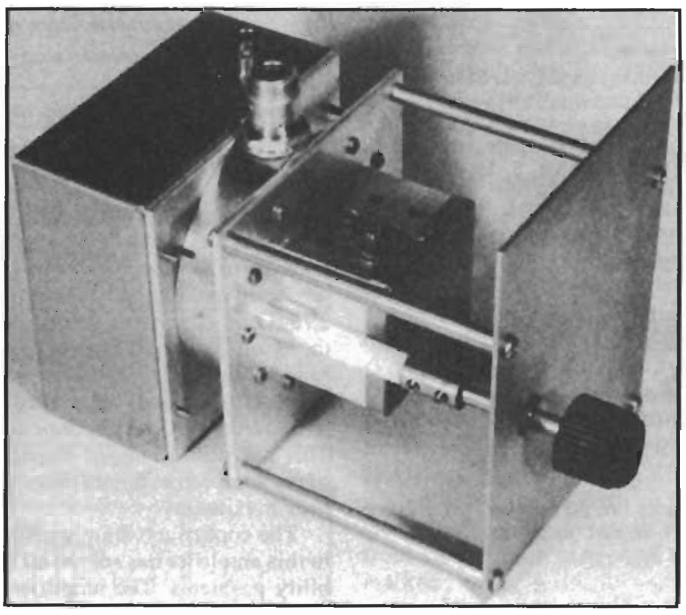

To me, there is nothing more frustrating than having to dig through a construction article to find out exactly what performance you can expect from the finished product. So here it is:

1) Grounded-grid 7289/2C39 cavity amplifier, single tube.

2) Linear operation (what you put in, you get out, only more of it).

3) Covers 1240 to 1300 MHz.

4) Power. gain ranges froth 12-20 dB depending on output power, input power, loading, anode voltage and grid bias voltage.

5) 50-ohm input and output — no stub tuner required.

6) Power output greater than 200W with about 12-W drive.

This is Part 1 of a two-part article. In this installment, I describe the design and construction of the RF deck. Part 2 describes power-supply construction, testing and operation.

This amplifier is a tried and proven design. Much development work has gone into this project. The amplifier works well, is reliable and can be duplicated. More than 50 of these amplifiers have been built to date. I have successfully worked many 1296-MHz EME (earth-moon-earth) stations with one of these amplifiers and a 384-element loop-Yagi array during the past year. Amplifiers of this design were used on both ends of the first California-to-Hawaii QSO on 1296 MHz. Another unit has logged more than 20,000 hours of continuous operation at the KH6HME beacon.

General Design Approach

A cavity amplifier is similar to a conventional amplifier designed for lower frequencies. The tube anode excites a resonant circuit, and power is in turn coupled into a load, usually 50 ohms. Instead of using coils and capacitors, as at lower frequencies, the cavity provides the resonant circuit necessary to tune the amplifier output.

The anode cavity of this amplifier is a squat cylinder. Cylinder height is set by mechanical tube requirements. The inside diameter of the cylinder sets the highest resonant frequency. Any capacitance added from the top to the bottom of the cavity will lower its resonant frequency, as will increasing the cavity diameter.

This amplifier uses $^{1}/_{8}$-inch-thick copper plates for the cavity top and bottom, and a thick-wall aluminum ring, cut from tubing, for the walls.[1] This heavy construction virtually eliminates all resonant-frequency variations caused by thermal and mechanical changes.

Fig. 1 is a schematic diagram of the cavity amplifier. The circuit is simple. Filament voltage and cathode bias enter the RF deck through feed-through capacitors (C4, C5) and RFC1 and 2. High voltage is fed to the anode through RFC3. C8, the anode bypass capacitor, is homemade from Teflon® dielectric sandwiched between a copper plate and the chassis.

The input pi network easily tunes the entire band at any power level. it is made from two Johanson piston trimmer capacitors and a "coil" made from copper wire. An input cavity is not necessary at 23 cm.

Output coupling is through a rotatable loop that serves as a variable loading control. This allows amplifier-tuning flexibility; it may be tuned for maximum gain or for maximum power. Light loading can produce stable power gains of up to 20 dB.

Amplifier tuning is accomplished with a homemade cylindrical coaxial capacitor with Teflon dielectric (C6). There are no moving metal parts to cause erratic performance. The Teflon rod/tube screws in and

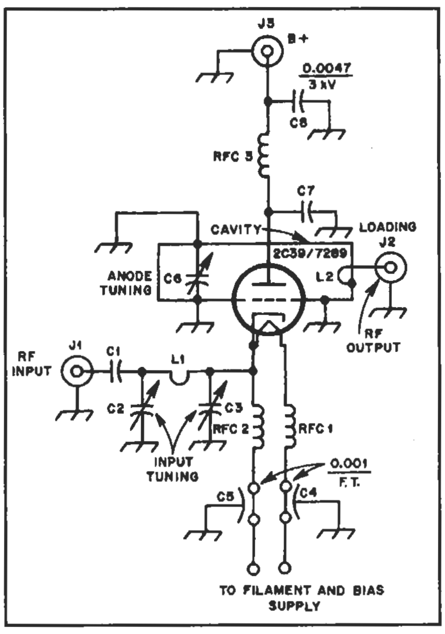

Fig. 1 — Schematic diagram of the 23-cm amplifier.

C1 — 3-pF dipped mica capacitor.
C2, C3 — 1- to 10-pF piston trimmer capacitor (Johanson no. 3957, 5201 or equiv.).
C6 — Anode-tuning capacitor. See text and Fig. 11.
C7 — Anode-bypass capacitor, 90 pF. Homemade from copper plate and Teflon sheet. See text and Figs. 5, 12 and 15.
C8 — Disc ceramic, 0.0047-µF, 3-kV capacitor.
J1 — 5-mm SMA connector, chassis mount, female.
J2 — Modified Type-N connector. See text and Fig. 7.
J3 — Female chassis-mount BNC connector.
Li — Loop of no. 18 bus wire soldered between C2 and C3. See Fig. 15.
L2 — Output-coupling loop. Part of output-connector assembly. See text and Fig. 7.
RFC1, RFC2 — 5 turns no. 20 bus wire, $^3/_{16}$-inch ID.
RFC3 — 3 turns no. 20 bus wire wound on a 20-ohm, 1-W carbon-composition resistor.

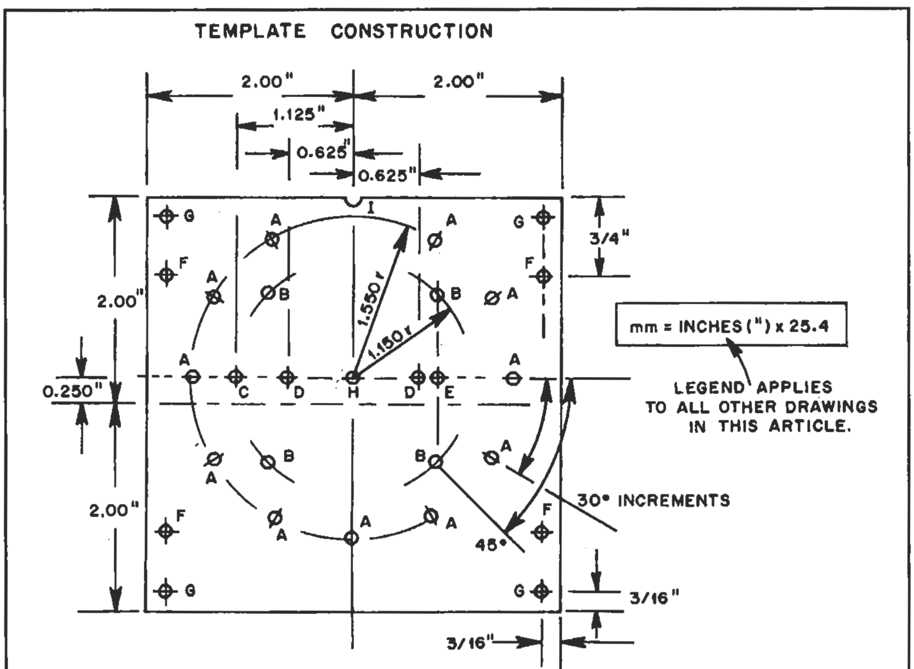

HOLES USED:
RING — A
ANODE PLATE — A, B, C, F, H
GRID PLATE — A, C, D, G, H
ANODE BYPASS-CAPACITOR PLATE — B, E, H
FRONT PANEL — C, G
I = RING ALIGNMENT NOTCH

- ALL HOLES 1/16" DIA. OR SMALLER
- MATERIAL: 1/16" ALUMINUM
- ALL DIMENSIONS IN INCHES
- ALL PLATES ARE VIEWED FROM FRONT PANEL
- HOLE DESIGNATIONS ARE MAINTAINED ON ALL DRAWINGS

Fig 2—Complete dimensions for the aluminum template.

out of the coaxial capacitor, increasing or decreasing the capacitance by changing the amount of Teflon dielectric inside the cylinder. With the rod all the way in, the dielectric is all Teflon; with the rod all the way out, the dielectric is all air.

Teflon has a relative dielectric constant (relative to air = 1) of 2.05, which means that the value of the capacitor with the rod all the way in is twice the value of the capacitor with the rod all the way out. Full capacitance will pull the resonant frequency of the amplifier down to 1240 MHz. Use of only one tuning adjustment means the amplifier will have more gain because cavity shunt capacitance has been minimized.

Thermal Considerations

The cavity walls are formed by a thick-wall aluminum ring, which is sandwiched between two thick copper plates. RF and thermal properties of these two metals are reasonably close, whereas brass is rather poor in both respects. The 7289/2C39 tube used in this amplifier is being run at 2-2½ times its normal dissipation rating; therefore it's important to have a cavity that remains thermally stable.

Most previously described amplifiers have used sheet brass in their construction. This has usually meant constant retuning of resonance to maintain output power at or near maximum.

The copper and aluminum construction in this amplifier has solved all thermal stability problems. The amplifier can easily be run key down for over an hour at 200-W output without retuning. This, of course, is obtained only with a good tube and water cooling. A practical water cooling system will be described in Part 2 of this article.

Water cooling keeps the internal structure of the tube thermally stable. When air cooling is used for output levels of 100 to 150 W, output power fluctuations are a direct result of internal tube changes. These changes vary from tube to tube and must be tested for. In some cases, otherwise perfectly good RF tubes have had poor thermal stability. Such tubes can make good drivers at lower power levels.

"Using Simple Hand Tools Will . . ."

Hand tools are great if you are skilled and patient. Most people want to hurry up and finish their new project. If that's you, then have a machine shop make all of the parts, leaving you only the final assembly. It should cost about $200. The parts are not difficult to fabricate, but the process is time consuming. If you have the time and patience to do it yourself, this amplifier can be very inexpensive.

Gathering the Materials

All of the materials used in this amplifier are fairly common and should be available from suppliers in most metropolitan areas. Some suppliers have "short sale" racks, where they sell odd pieces cut off standard lengths or sheets at reduced prices. The parts for this project are small enough to be fashioned from cutoff stock. Surplus-metal houses have some great buys, so start there if one is nearby.

The key to successfully completing this project is careful layout work before cutting or drilling any parts. Invest in a can of marking dye, a sharp scribe, an accurate rule, vernier calipers and several center punches. These tools are available at any machinists' supply shop. The marking dye will make cutting and filing lines much easier to see. Measure all dimensions as carefully as you can and then recheck them before cutting. Mark with a sharp scribe because the sharper the scribe, the finer the marked line, and the finer the marked line, the closer your cut will be to where it should be. Remember—the accuracy of your drilled holes is only as good as your center-punching ability, so use a fine punch for the first mark and then a bigger one to enlarge the mark enough for drilling.

Access to a drill press is a must. It's extremely difficult to drill holes accurately with a hand drill. Although they are not absolutely necessary, you should have access to a lathe or milling machine.

Other tools that will aid you with this project are a nibbling tool, a set of punches, a new set of files and some sharp drill bits. If you don't already have one, purchase a file card to clean metal shavings out of your files as you work. Clean, sharp files are faster and more accurate to work with. You'll also need an assortment of sandpaper for the final finish work.

The Template Approach

I highly recommend fabrication of a single template for marking and drilling the anode plate, anode bypass capacitor, cavity ring, grid plate and front panel. The template shown in Fig. 2 has all of the holes for these parts. If you use the template, you'll only have to make the careful measurements once — after that, it's simple to mark and drill the rest of the parts.

The template approach offers several other advantages. A template makes it much easier to maintain accuracy between the anode plate, cavity ring, grid plate and front panel; these parts will fit perfectly because they were all drilled from the same master. The template approach also makes it possible to set up a small production line if you decide to build more than one of these amplifiers and combine them for higher power, or if a friend wants to build an amplifier along with you.

See Fig. 2 for complete template dimensions. Start with a piece of 1/16-inch-thick aluminum stock that is larger than you need and degrease it with soap and water. Dry it off and spray it with marking dye. Scribe a 4-inch square on the stock and cut the template to size. A shear will make this job much easier, but it can be cut with hand tools and filed to size.

Carefully measure and scribe all holes. Note that holes A and B are on the circumference of circles. Use a compass to scribe the circles, and then locate the holes. After you have marked and checked all holes, centerpunch and drill them. The holes should be drilled with a 1/16-inch or smaller bit. Recheck all measurements. If you goof, start again. The time you spend making the template as perfect as you can will save you much time and aggravation when you make and assemble the other parts.

When you finish the template, mark the front side for future reference. All plates made from the template are marked and drilled from the front side (as viewed from the front panel).

Making the Copper Plates

Once you have completed the template, it will be easy to make the copper plates. The anode plate, grid plate and anode-bypasscapacitor plate are all made from 1/8-inch-thick copper. See Figs. 3, 4 and 5 for the dimensions of these pieces.

Measure and cut the three plates to the proper dimensions. Carefully break (deburr) all sharp edges to avoid small cuts to your fingers and hands.

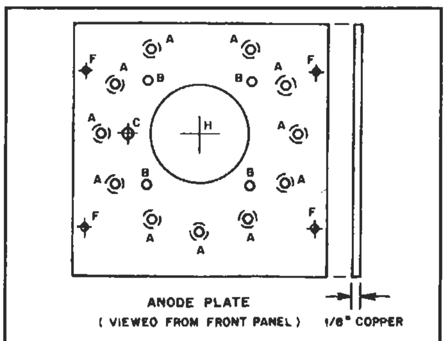

HOLE: A — 0.144-IN DIA. (NO. 27 DRILL). COUNTERSINK BACK SIDE TO 1/4-IN DIA. WITH A 100°C SINK.
B — 0.106-IN DIA. (NO. 36 DRILL), TAPPED NO. 6-32.
C — 5/32-IN DIA. (0.156). FOR ANODE TUNING POST, THIS SIDE.
F — 0.089-IN DIA. (NO. 43 DRILL). TAPPED NO. 4-40 FOR ANODE BOX MOUNTING.
H — 1.500-IN DIA.

Fig 3—Drilling details for the anode plate. See Fig 2 for additional information on hole location.

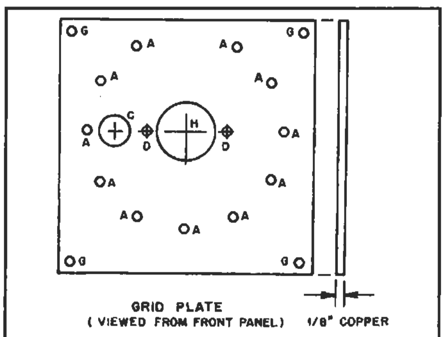

HOLE: A — 0.144-IN DIA. (NO. 27 DRILL).
C — 7/16-IN DIA. (0.438). ANODE TUNING COLLAR INSERTS FROM THIS SIDE.
D — 1/8-IN DIA. (0.125). FOR CATHODE-GRID MOUNTING.
G — 0.144-IN DIA. (NO. 27 DRILL). FOR FRONT PANEL MOUNTING.
H — 7/8-IN DIA. FOR GRID FINGER STOCK. FINGERS PROTRUDE OUT BACK SIDE.
NOTE: BRASS INPUT ENCLOSURE MOUNTS ON THIS SIDE.

Fig 4—Drilling details for the grid plate. See Fig 2 for additional information on hole location.

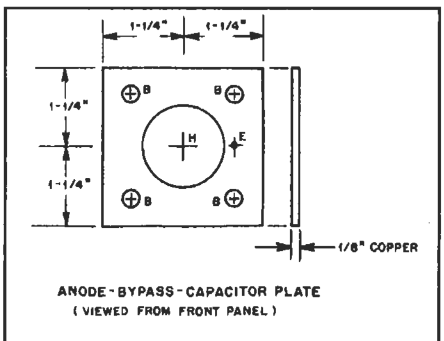

HOLES: B — 1/4-IN DIA.
E — 0.070-IN DIA. (NO. 50 DRILL), TAPPED FOR NO. 2-56 FOR HIGH-VOLTAGE LUG.
H — 1-1/4-IN DIA. FOR ANODE FINGER STOCK; FINGERS PROTRUDE OUT BACK SIDE.

Fig 5—Drilling details for the anode bypass capacitor plate. See Fig 2 for additional information on hole location.

Clean the plates with alcohol and spray them with marking dye. Clamp the aluminum template to each plate, and carefully scribe the correct holes. Remember that all plates do not have the same holes. The anode plate uses holes A, B, C, F and H; the grid plate uses holes A, C, D, G and H. The anode-bypass-capacitor plate uses holes B, E and H.

Use a small center punch to punch all holes lightly. If they then look accurate, enlarge them enough for drilling.

Copper isn't the easiest metal to work with. It's very stringy, and drilling it can be frustrating. You'll need the proper drill bits for best results. Special drills can be purchased, or you can use a grinder to carefully remove the sharp points on the outer edge of the cutting surface of each side of a standard drill bit. This will eliminate any tendency for the copper to grab. Practice on an old bit and be sure to grind it symmetrically. Modified drill bits can still be used on aluminum and other metals.

Always start with a smaller drill and work up to the final hole size. It's safer and more accurate. The larger holes can be cut with a flycutter, or you can drill a series of smaller holes around the inside of a larger hole and file to finish. Either way is fine. Use lots of cutting fluid to lubricate the drill bit, and wear safety glasses and an old shirt. Remember, some cutting fluids are not to be used on aluminum.

Start with a no.50 (0.070-inch) or smaller bit and drill pilot holes at each of your punched marks. The details for finishing each hole are listed in the drawings. Some holes are countersunk or tapped. Pay attention to the details, and take your time.

When you are through drilling, you must deburr each hole. Copper is soft, so it tends to rise up around the hole during drilling and deburring. Use a flat file for the initial cut, and then remove any remaining material with a countersink. File the copper plates flat again; a flush fit on both sides of the aluminum ring is important.

When all copper work is done, you should be able to stack the plates and see all pertinent holes align correctly. Enough tolerance is included in the dimensions to accommodate minor errors. After the holes are drilled, it can be difficult to tell which side of each plate is which, so mark the front side of each plate with a permanent marker.

Machining the Ring

The aluminum ring that forms the cavity wall is cut (sliced) from a length of 3½-inch OD tubing with a 3/8-inch wall thickness. See Fig. 6. The tubing ID is about 2¾ inches. The dimensions of the ring are the most critical in this amplifier. Tolerance of the ring thickness is ± 0.005 inch to maintain full band coverage.

The ring can be hacksawed or bandsawed out of the tubing, but take extreme care to be accurate. Cutting tubing straight isn't easy. Clamp the tubing to prevent rotating on the band saw. The final finish cut is best done on a lathe or milling machine, but careful filing will work.

Once the ring is the correct thickness, deburr the sharp edges and spray it with marking dye. Notice that the outside and inside diameters are not concentric. This is normal for large tubing. Lay the ring flat and find the thickest wall section. Scribe a

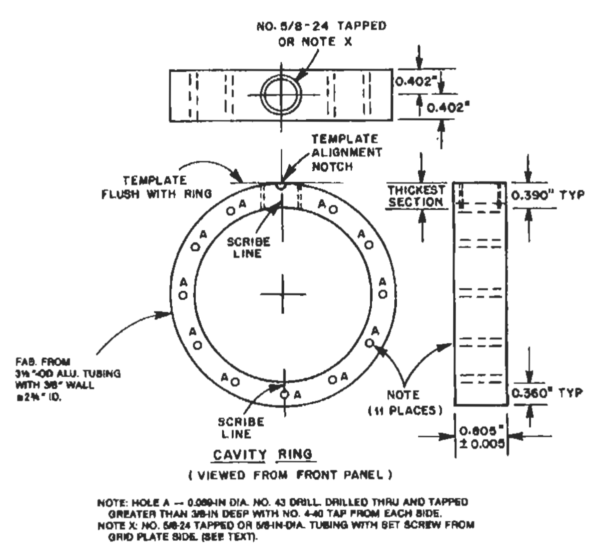
NO. 5/8-24 TAPPED
OR NOTE X
0.402"
0.402"
TEMPLATE
ALIGNMENT
NOTCH
TEMPLATE
FLUSH WITH RING
THICKEST
SECTION
0.390" TYP
A
SCRIBE
LINE
FAB. FROM
3½"-OD ALU. TUBING
WITH 3/8" WALL
≈2¾" ID.
NOTE
(11 PLACES)
0.360" TYP
SCRIBE
LINE
0.605"
± 0.005
CAVITY RING
(VIEWED FROM FRONT PANEL)
NOTE: HOLE A — 0.089-IN DIA. NO. 43 DRILL. DRILLED THRU AND TAPPED
GREATER THAN 3/8-IN DEEP WITH NO. 4-40 TAP FROM EACH SIDE.
NOTE X: NO. 5/8-24 TAPPED OR 5/8-IN-DIA. TUBING WITH SET SCREW FROM
GRID PLATE SIDE. (SEE TEXT).

Fig 6—Details of the cavity ring. See Fig 2 for additional information on hole location.

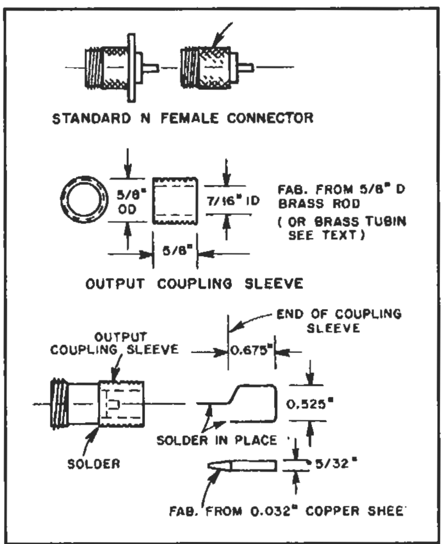
STANDARD N FEMALE CONNECTOR
5/8"
OD
7/16" ID
5/8"
FAB. FROM 5/8" D
BRASS ROD
(OR BRASS TUBIN
SEE TEXT)
OUTPUT COUPLING SLEEVE
END OF COUPLING
SLEEVE
OUTPUT
COUPLING SLEEVE
0.675"
0.525"
SOLDER IN PLACE
SOLDER
5/32"
FAB. FROM 0.032" COPPER SHEE

Fig 7—Output-probe/connector assembly details.

line across the wall at this point, across the center of the ring and across the wall on the other side. The scribed lines on each side of the ring will be used to align the template. The output connector will be placed at the thick wall section.

Carefully align notch I on the template with the line scribed on the thickest wall section on the ring. Clamp the template onto the ring. Mark each of the 11 holes labeled A on the template. After you mark the holes and remove the template, check alignment with the copper plates just in case. If everything lines up, center punch all eleven holes on one side of the ring only, and drill each hole completely through the ring. Use lots of cutting fluid. File the ring flat before and after deburring, taking care not to change the wall thickness. Tap each hole to accept no. 4-40 machine screws. Each hole will have to be tapped to a depth of at least $^3/_8$ inch from both sides because long taps don't exist. The inside of the ring doesn't need to be polished.

The hole for mounting the output connector can now be drilled. There are two ways to mount this connector, and either scheme works fme. Read ahead to the section on making the output connector for more information. The first method of mounting the connector involves tapping the ring with a no. $^5/_8$-24 tap and using a lathe to cut matching threads on the output connector coupling sleeve. Large taps are expensive, but a tap and die for Type-N connectors are handy if you do much building.

If you don't have access to a lathe or a large tap, the second method is easier. Make the output connector coupling sleeve from $^5/_8$-inch-OD brass or copper tubing, and drill the ring to just clear it. Then drill and tap the grid-plate side of the ring above the output connector to accept a setscrew. Also, drill a clearance hole in the grid plate for the setscrew. Use the setscrew to secure the output connector.

Output Connector

A standard Type-N chassis-mount female connector (silver plated) is used for the output probe/connector. See Fig. 7. First, remove the flange with a hacksaw and file flush with the connector body. Next, make the output-coupling sleeve that is right for your application (threaded orunthreaded, depending on how you fabricated the ring). The sleeve will be the same length in either case. The output-coupling loop is fashioned from a piece of 0.032-inch-thick copper sheet that is $^5/_{32}$ inches wide. Bend it to the dimensions shown in Fig. 7. We will solder the output connector together later.

Grid Compartment

The grid compartment measures 2 inches square by $1^1/_2$ inches high. See Fig. 8. It is made from brass and can be sawed out of square tubing or bent from sheet. The cover can be made from any material.

I use two small PC boards (Fig. 9) for

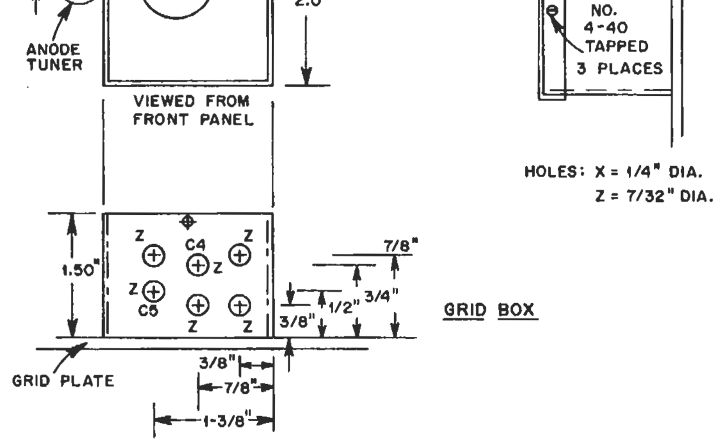
0.475"
1.05"
Z
C2
Z
X
C3
Z
GRID PLATE
INPUT
CONNECTOR
2.0"
0.800"
C
H
C
2.0"
ANODE
TUNER
VIEWED FROM
FRONT PANEL
COVER, FAB. FROM
0.020" ALU. OR BRASS
NO.
4-40
TAPPED
3 PLACES
HOLES: X = 1/4" DIA.
Z = 7/32" DIA.
1.50"
Z
C4
Z
Z
Z
C5
Z
Z
7/8"
3/4"
1/2"
3/8"
GRID BOX
GRID PLATE
3/8"
7/8"
1-3/8"

Fig 8—Input-compartment details.

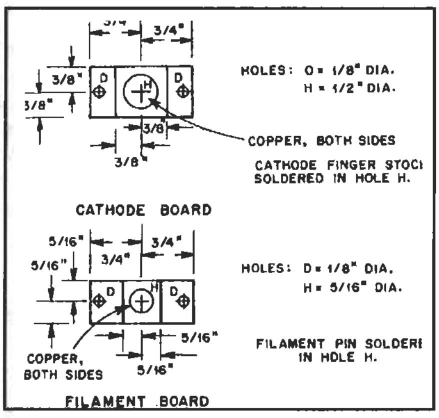
3/4"
3/4"
3/8"
3/8"
D
H
D
3/8
3/8"
HOLES: O = 1/8" DIA.
H = 1/2" DIA.
COPPER, BOTH SIDES
CATHODE FINGER STOC
SOLDERED IN HOLE H.
CATHODE BOARD
5/16"
3/4"
3/4"
5/16"
D
H
D
5/16"
5/16"
COPPER,
BOTH SIDES
HOLES: D = 1/8" DIA.
H = 5/16" DIA.
FILAMENT PIN SOLDER
IN HOLE H.
FILAMENT BOARD

Fig 9—Cathode and filament PC-board details.

holding the finger stock that makes contact with the filament pin and cathode ring on the 2C39 tube. These boards are cut from 1/16-inch-thick, double-sided G-10 glass-epoxy stock. The copper pattern is identical for both sides of each piece. Mark and drill or file the holes first, and then cut the boards to size. Small boards are difficult to hold while drilling them. Mark each side of each board and score the copper foil with a sharp knife.

The unwanted copper can be removed easily by heating the foil with a soldering iron and lifting it off. Use a flat file to deburr the boards. Do not use a countersink because the copper foil must be as close to the holes as possible to facilitate soldering the finger stock in place.

The input connector that I use is a 5-mm SMA type. This is an excellent RF connector, especially for low-power UHF applications. I highly recommend use of an SMA, but any small screw-on connector will do. If you really feel you have to use a BNC then do so, but it's a lousy connector at frequencies above 200 MHz. Remember to move the connector hole to accommodate its larger size.

The input connector must be as close as possible to the first input capacitor. The lead length of the input dc blocking capacitor must be as short as possible. The 3-pF capacitor is series resonant at 1200 MHz only with short (1/16-inch or less) leads.

Miscellaneous Bits and Pieces

There are still several small, but very important parts to fabricate. The front panel I use is shown in Fig. 10. It is made from a piece of 1/8-inch-thick aluminum sheet. Some builders may wish to mount the amplifier on a rack panel. Wash and dry your front-panel material and spray it with marking dye. Clamp it to the template and mark the holes. Check the hole alignment with the copper grid and anode plates. If all lines up correctly, center punch and drill the holes. The only front-panel control is for the anode tuning capacitor, which is adjusted by a 1/4-inch shaft protruding through a 3/8-inch panel bushing in hole C.

The anode tuning collar, shown in Fig. 11A, is made from a piece of 1/2-inch OD brass rod. This rod has a 3/8-inch hole drilled through its center, and it is turned down to 7/16-inch OD for half its length. The inside of the 1/2-inch-OD end is tapped to a depth of 1/4 inch to accept 3/8-24 threads. This collar will be inserted into hole C on the grid plate.

Fig. 11B also shows the anode tuning post. It is simply a length of 5/32-inch-OD brass rod that inserts into hole C on the copper anode plate. This rod will form one plate of the anode tuning capacitor.

The anode tuner (Fig. 11C) is machined from a piece of 3/8-inch-OD Teflon rod. One end of the rod is drilled out with a no. 21 drill. The outer wall of this end is threaded with a no. 3/8-24 tap. This is the end that will thread into the anode tuning collar and slip over the anode tuning post. The other end is turned down to fit inside a 1/4-inch shaft coupler.

Fig. 12 shows the remaining parts. The tuning shaft (A) is made from a piece of 1/4-inch brass rod. A coupler (B) to connect the tuning shaft to the anode tuner may be purchased or made. This also applies to the front-panel spacers (C). The Teflon dielectric for the anode bypass capacitor (D) is made from 0.010-inch-thick Teflon sheet. Use the template to locate holes B and H. Teflon washers and inserts (F) are used to insulate the mounting hardware for the anode bypass capacitor from the chassis. The inserts are made from 1/4-inch-OD Teflon rod. The washers are made from Teflon sheet. Sharpen a piece of 3/8-inch aluminum tubing and chuck it up in a drill press. This tool will cut neat, round washers from the sheet.

The box that encloses the anode compartment (Fig. 13) is fabricated from a Bud AU-1083 utility cabinet. Clean the chassis and spray it with marking dye. Secure the template to the side of the enclosure that contacts the anode plate and scribe the holes labeled F. Make sure that these holes line up with the holes on the copper anode plate. If they do, center punch and drill them to size. If air cooling is used, the blower will mount to this box.

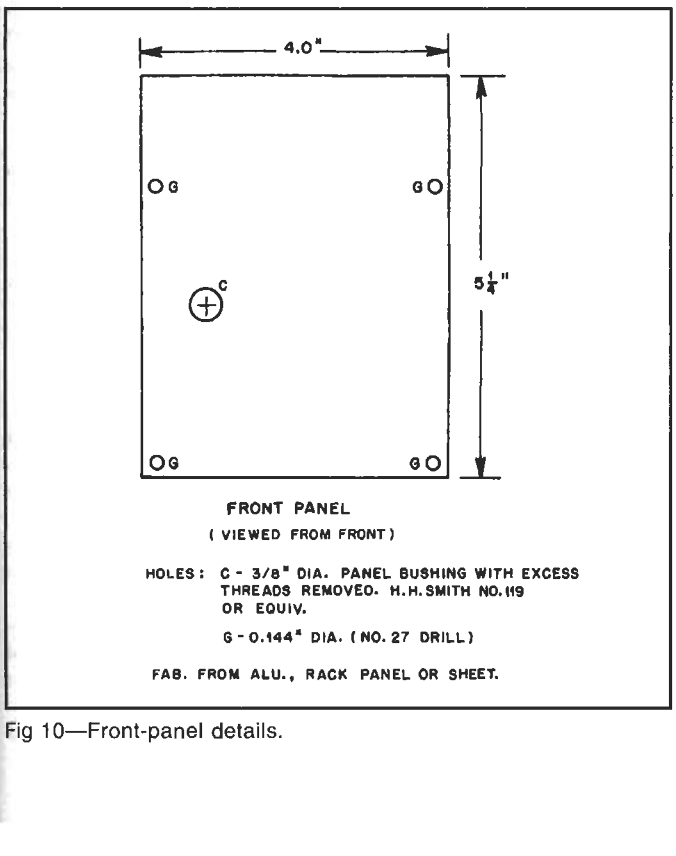
4.0"
5¼"
G
G
C
G
G
FRONT PANEL
(VIEWED FROM FRONT)
HOLES: C - 3/8" DIA. PANEL BUSHING WITH EXCESS THREADS REMOVED. H.H.SMITH NO.119 OR EQUIV.
G - 0.144" DIA. (NO. 27 DRILL)
FAB. FROM ALU., RACK PANEL OR SHEET.

Fig 10—Front-panel details.

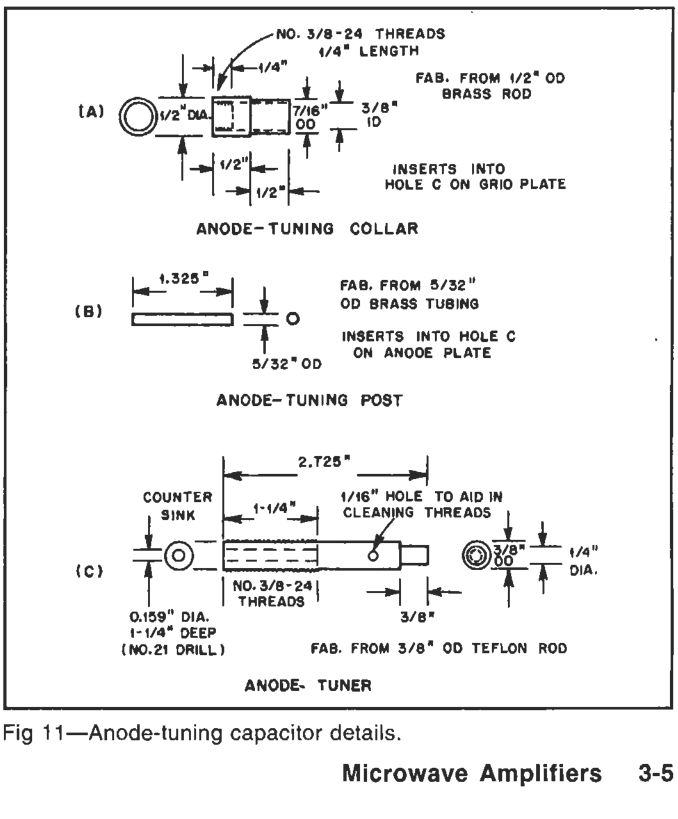
(A)
NO. 3/8-24 THREADS
1/4" LENGTH
1/4"
1/2" DIA.
7/16" OD
3/8" ID
1/2"
1/2"
FAB. FROM 1/2" OD BRASS ROD
INSERTS INTO HOLE C ON GRID PLATE
ANODE-TUNING COLLAR
(B)
1.325"
5/32" OD
FAB. FROM 5/32" OD BRASS TUBING
INSERTS INTO HOLE C ON ANODE PLATE
ANODE-TUNING POST
(C)
2.T25"
COUNTER SINK
1-1/4"
1/16" HOLE TO AID IN CLEANING THREADS
3/8" OD
1/4" DIA.
NO. 3/8-24 THREADS
3/8"
0.159" DIA.
1-1/4" DEEP
(NO.21 DRILL)
FAB. FROM 3/8" OD TEFLON ROD
ANODE- TUNER

Fig 11—Anode-tuning capacitor details.

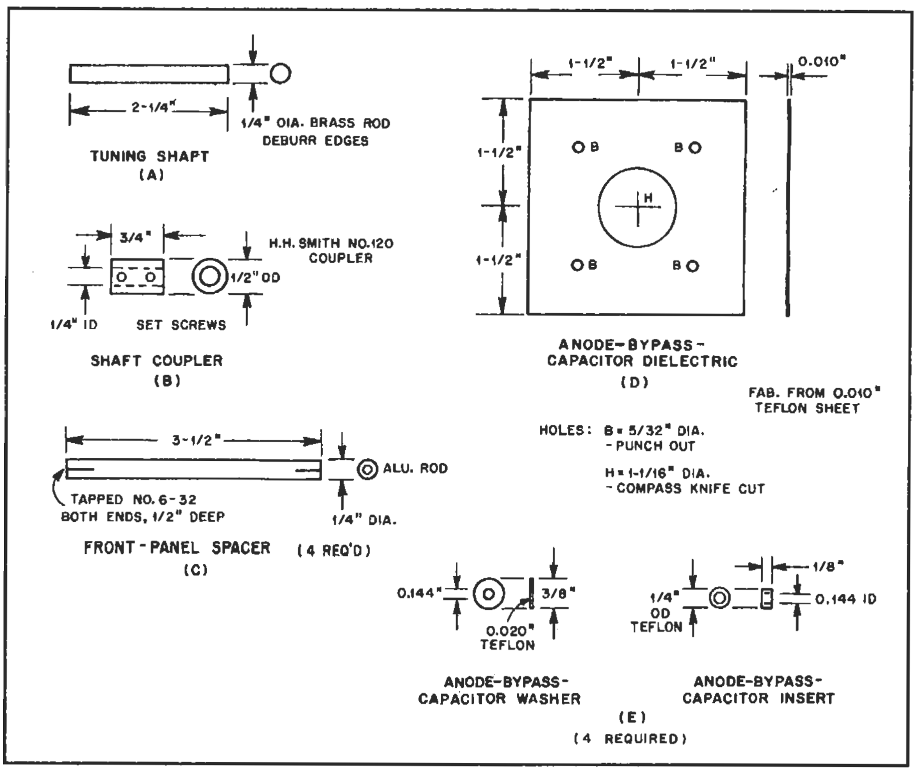
2-1/4"
1/4" DIA. BRASS ROD
DEBURR EDGES
TUNING SHAFT
(A)
3/4"
H.H. SMITH NO.120
COUPLER
1/2" OD
1/4" ID
SET SCREWS
SHAFT COUPLER
(B)
3-1/2"
ALU. ROD
TAPPED NO. 6-32
BOTH ENDS, 1/2" DEEP
1/4" DIA.
FRONT-PANEL SPACER
(C)
(4 REQ'D)
1-1/2"
1-1/2"
0.010"
1-1/2"
1-1/2"
B
B
H
B
B
ANODE-BYPASS-
CAPACITOR DIELECTRIC
(D)
FAB. FROM 0.010" TEFLON SHEET
HOLES: B = 5/32" DIA.
- PUNCH OUT
H = 1-1/16" DIA.
- COMPASS KNIFE CUT
0.144"
3/8"
0.020"
TEFLON
1/8"
1/4"
OD
TEFLON
0.144 ID
ANODE-BYPASS-
CAPACITOR WASHER
ANODE-BYPASS-
CAPACITOR INSERT
(E)
(4 REQUIRED)

Fig 12—Miscellaneous parts necessary to complete the amplifier.

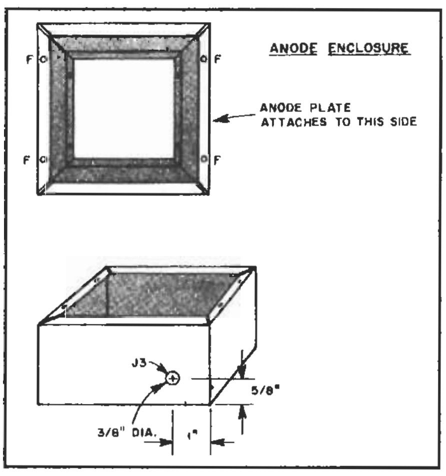
ANODE ENCLOSURE
F
F
F
F
ANODE PLATE
ATTACHES TO THIS SIDE
J3
5/8"
3/8" DIA.
1"

Fig 13—Anode-enclosure details.

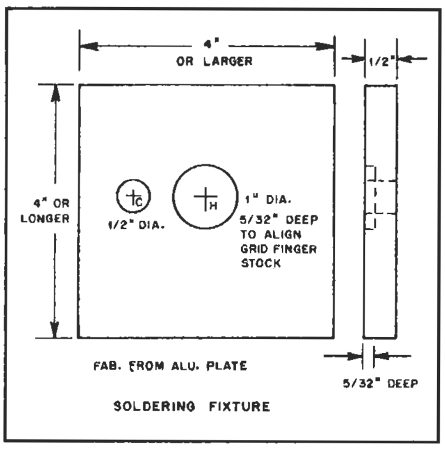
4"
OR LARGER
1/2"
4" OR
LONGER
C
1/2" DIA.
H
1" DIA.
5/32" DEEP
TO ALIGN
GRID FINGER
STOCK
FAB. FROM ALU. PLATE
5/32" DEEP
SOLDERING FIXTURE

Fig 14—Dimensions of the soldering fixture. See Fig 2 for more information on hole location.

Soldering the Subassemblies

Once all copper and brass parts are drilled and deburred, they should be cleaned with alcohol and Scotch-Brite®, a nonmetaffic pot cleaner, and washed in alcohol again. Set the pieces aside and avoid touching them. Fingerprints will inhibit soldering.

I have found that the best way to solder the heavy brass and copper parts is to first build the soldering fixture shown in Fig. 14. This soldering fixture, made from $^1/_2$-inch-thick aluminum plate, will evenly heat the entire assembly to be soldered. Even heating will allow you to do a much better soldering job than you could otherwise.

The soldering fixture should be preheated on a stove or hot plate until bits of solder placed on its surface just melt. At this point, reduce the heat slightly. Avoid excessive heat. If the copper parts placed on the fixture suddenly turn dark, it's too hot.

Solder the grid plate assembly first. You will need the copper grid plate, grid finger stock, anode tuning collar and brass input compartment.[2] Look at the drawings again to be sure that you know which parts go where. Insert the grid finger stock into hole H on the grid plate. As viewed from the front-panel side, the curved fingers will protrude out the back side, away from you. Apply liquid or paste flux and set the grid plate in the soldering fixture. The finger stock will fit in hole H in the fixture, allowing the grid plate to rest flush with the surface of the fixture. Next, apply flux to the anode tuning collar and insert it in hole C of the grid plate. Part of the tuning collar will slip into hole C in the soldering fixture. Make sure the collar seats flush with the grid plate. The flux should start to bubble.

Carefully apply solder directly to the joints of the installed parts. The solder should melt almost immediately and flow bright and smooth. Next, place the square brass input compartment in place and apply flux. In a few seconds, it can be soldered by running solder around the joints, inside and outside. If you have trouble getting it to flow on both sides, merely tap the brass box aside ($^1/_{16}$ inch) and return it to its original position.

Now comes the hard part — getting the soldered assembly away from the heat without disturbing the alignment. A pair of forceps is recommended, but long pliers will do. Carefully lift the assembly off the soldering fixture and set on a cooling rack. Do this without moving any part. The cooling rack can be any two pieces of metal that will allow clearance for the protruding parts. You can expedite cooling by using an ordinary hair dryer in the "cool" position to gently blow air across the assembly.

While the grid assembly is cooling, assemble the output connector. See Fig. 7. Place the modified Type-N female connector, threaded end down, on the soldering fixture. Apply flux to the top and install the output coupling sleeve. Allow both parts to heat before applying solder. Carefully remove the soldered output connector from thefixture. When it has cooled, solder one end of the loop to the center pin of the N connector and the other to the output coupling sleeve.

Now place the anode plate on the soldering fixture and allow to heat. Apply flux to hole C. Insert the anode tuning post ($^5/_{32}$-inch-OD brass tube) and allow to heat; apply solder. Remove the parts and cool. Next, solder the finger stock in hole H on the anode bypass capacitor plate.[3]

This completes the work with the soldering fixture. Be sure to let it cool off before handling! Save the fixture for future construction; you never know when you might want it again.

The anode plate and the anode-bypass-capacitor plate must be filed and then sanded flat on their butt surfaces to assure that there are no solder bumps or sharp points to puncture the Teflon dielectric. This must be done after soldering. The Teflon sheet is adequate insulation for many times the anode potential of this amplifier, but only if the surfaces it separates are smooth!

Next, clean the cathode and filament PC boards. Install the fmger stock in hole H of the cathode board. Apply flux to both sides of the board. Heat with a hot iron and apply solder around the circumference of hole H, soldering the finger stock on both sides of the board. Use the same technique to install the filament pin.[4]

After all parts have cooled, use a spray can of flux remover to clean them. Slight scrubbing with Scotch-Brite pot cleaner will finish them nicely. Congratulations: You have finished the pieces and are now

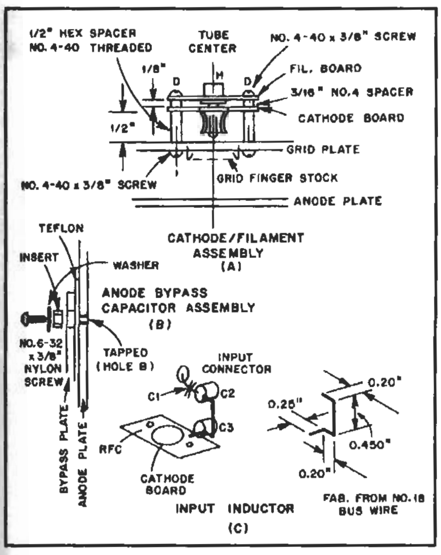
1/2" HEX SPACER
NO. 4-40 THREADED
TUBE CENTER
NO. 4-40 x 3/8" SCREW
1/8"
D
H
D
FIL. BOARD
3/16" NO.4 SPACER
CATHODE BOARD
1/2"
GRID PLATE
NO. 4-40 x 3/8" SCREW
GRID FINGER STOCK
ANODE PLATE
CATHODE/FILAMENT ASSEMBLY
(A)
TEFLON
INSERT
WASHER
ANODE BYPASS CAPACITOR ASSEMBLY
(B)
NO.6-32 x 3/8" NYLON SCREW
TAPPED (HOLE B)
BYPASS PLATE
ANODE PLATE
INPUT CONNECTOR
C1
C2
C3
RFC
CATHODE BOARD
INPUT INDUCTOR
(C)
0.20"
0.25"
0.450"
0.20"
FAB. FROM NO.18 BUS WIRE

Fig 15—Assembly details for the filament and cathode boards (A), the anode-bypass capacitor (B) and the input pi network (C).

ready to bolt the amplifier together.

Silver Plating

Over the years, many people have pushed silver plating as the only way to go. You may wish to silver plate the amplifier components before soldering them together, but I do not think it's necessary. I ran several tests to prove how much various types of plating affect performance of this amplifier. Remember that the RF skin conductivity of aluminum and copper is pretty good at 23 cm; they are much better than brass.

Four amplifiers were built for this test. They were plated as follows:

1) Nickel plated
2) Tin plated
3) Silver plated
4) Unplated

There was no difference in performance among the tin-plated, silver-plated and unplated versions. The nickel-plated amplifier exhibited 3-dB less gain.

In other words, it is not necessary to silver plate this amplifier; however, it does improve appearance by making the parts a similar color. Silver does tarnish, especially with fingerprints. The decision to plate or not to plate is up to you.

Assembly

After fabrication of all parts, assembly is simple. Figs. 15 through 17 show assembly details. Loosely fasten the grid and anode plates to the ring. Mount the input connector and capacitors on the input compartment. Loosely install the cathode and filament boards and their respective spacers. See Fig. 15A.

Now insert a 7289/2C39 tube. This will center up all finger stock. Place the Teflon anode tuner in its collar on the grid plate and screw it most of the way in. Now tighten all of the screws. The 7289/2C39 tube should slide in and out snugly, and the anode tuner should screw in and out smoothly.

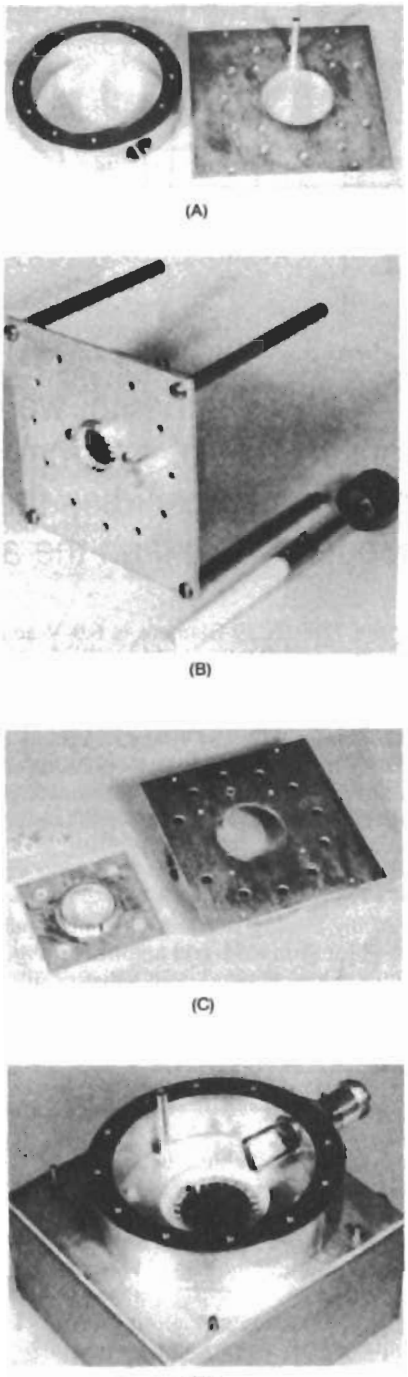
(A)

(B)

(C)

(D)

Fig. 16—The completed cavity ring and anode plate with anode tuning post soldered In place are shown at A. The photo at B shows the grid plate with finger stock, input compartment and anode tuning collar soldered in place. The completed anode tuner Is at the right. C shows the cavity ring attached to the anode plate. The anode-bypass capacitor is ready for installation. At D, the interior of the cavity as seen from the grid plate side is visible. The output probe/connector assembly is installed. The anode bypass capacitor and anode enclosure have been installed on the anode plate.

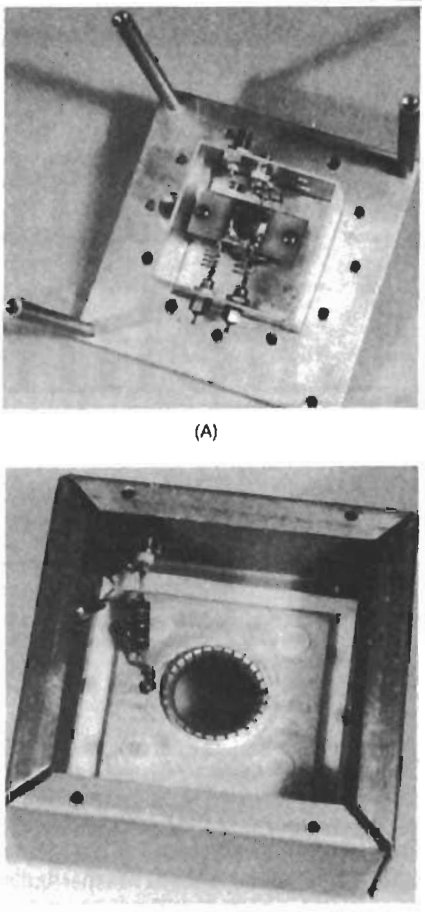
(A)

(B)

Fig 17—At A, the interior of the completed input compartment is visible. The photo at B shows the interior of the anode compartment with the anode bypass, RFC3, C8 and J3 installed.

The Teflon sheet and anode bypass capacitor plate can be installed now (Fig. 15B). Assemble the remaining input components, the filament feed-through capacitors and RFC5 (Fig. 15C). Screw the output probe into the cavity ring (or push in the probe and tighten the setscrew, depending on which method you chose). Install the high-voltage connector and other parts in the anode box. Mount the amplifier on the front panel and install the anode tuner shaft. This completes the assembly. Part 2 of this article will describe a complete power supply for the amplifier, a practical water cooling system, testing procedures, microwave radiation safety hazards, and amplifier tune-up and operation.

Notes

[1] mm = in x 25.4.

[2] The finger stock for this project is manufactured by Instrument Specialities, P.O. Box A, Delaware Water Gap, PA 18237. Contact them for the name of the closest distributor. The part numbers for this amplifier are: anode bypass capacitor plate, no. 97-70A; grid plate, no. 97-74A; cathode board, no. 97-420A; filament board, no. 97-280A.

[3] See note 2.

[4] See note 2.

By E.R. "Chip" Angle, N6CA **From *QST*, April 1985**

A Quarter-Kilowatt 23-cm Amplifier

Part 2—Last month, we described the design and construction of a 23-cm cavity amplifier. This installment describes the rest of the components needed to put it on the air.

After you complete Construction of the cavity amplifier described in March *QST*, you are ready to assemble the rest of the components needed to put it on the air. This month, I will discuss the filament, bias and high-voltage supplies; a whisper-quiet, high-efficiency water-cooling system; testing and hookup; and, finally, tune-up and operation.

Power Supplies

The filament and bias supplies for the cavity amplifier are shown schematically in Fig. 1. The manufacturer's specification for the 7289/2C39 filament is 6.0-V ac at 1 A. I have found that the use of a standard 6.3-V ac, 1-A transformer only slightly increases the tube emission without much loss of tube life. The filament should be allowed to warm up before operating the amplifier, so the filament, bias and high-voltage supplies incorporate separate primary switches.

Biasing

Many biasing schemes have been published for grounded-grid amplifiers. Fig. 1 shows a bias network that satisfies all of the following operating requirements:

1) external bias supply referenced to ground

2) low-power components

3) variable bias to accommodate tube-to-tube variations

4) TR switchable with relay contact or transistor to ground

5) bias-supply protection in case of a defective or shorted tube.

U2 provides a variable bias-voltage source, adjustable by R1. The output of U2 drives the base of Q1, which is used to increase the current-handling capability of

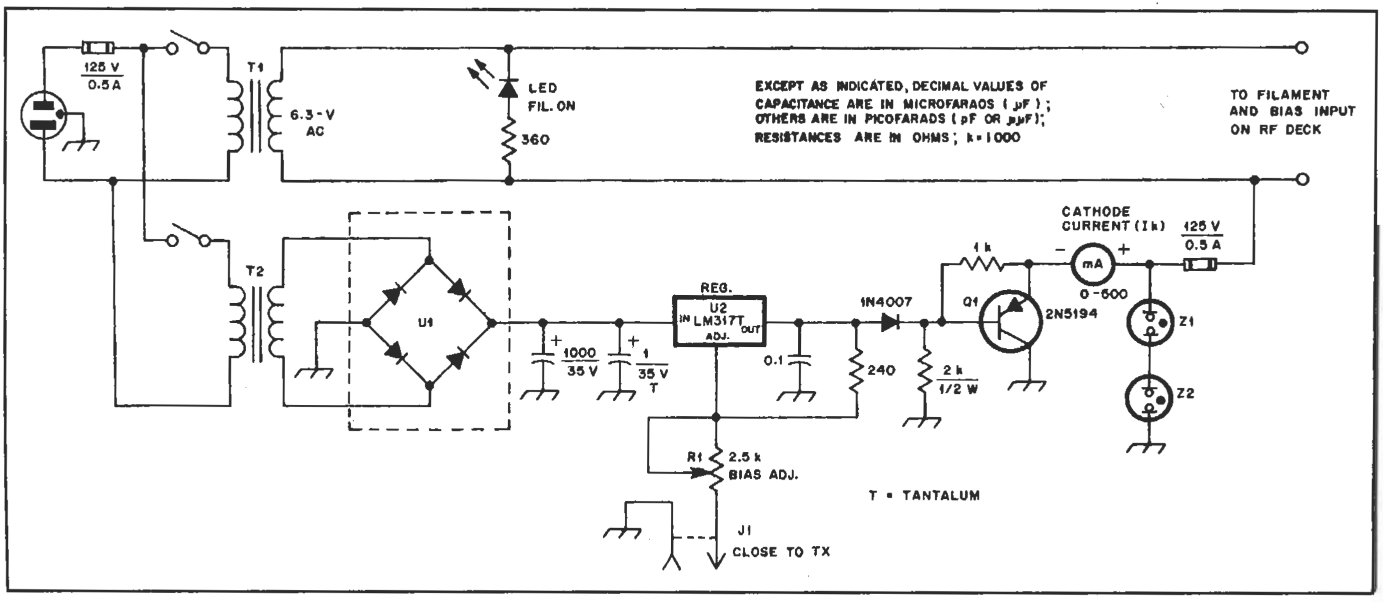
125 V
0.5 A
T1
6.3-V
AC
LED
FIL. ON
360
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k = 1000
TO FILAMENT AND BIAS INPUT ON RF DECK
T2
U1
REG.
U2
IN LM317T OUT
ADJ.
1000
35 V
1
35 V
T
0.1
240
1N4007
2 k
1/2 W
Q1
2N5194
1 k
CATHODE CURRENT (Ik)
mA
0-600
125 V
0.5 A
Z1
Z2
R1
2.5 k
BIAS ADJ.
T = TANTALUM
J1
CLOSE TO TX

Fig 1—Schematic diagram of the cavity-amplifier filament and bias supplies. All resistors are ¼-W carbon types unless otherwise noted.

J1—Female chassis-mount photo connector.
T1—Filament transformer. primary, 117 V; secondary, 6.3 V at 1 A.
T2—Power transformer. Primary, 117 V; secondary, 24 to 28 V at 50 mA or greater.
U1—Bridge rectifier, 50 PIV, 1 A.
U2—Adjustable 3-terminal regulator (LM317T or equiv.).
Z1, Z2—20-V unipolar metal-oxide varistor (General Semiconductor SA20 or equiv.) or two 20-V, 1-W Zener diodes.

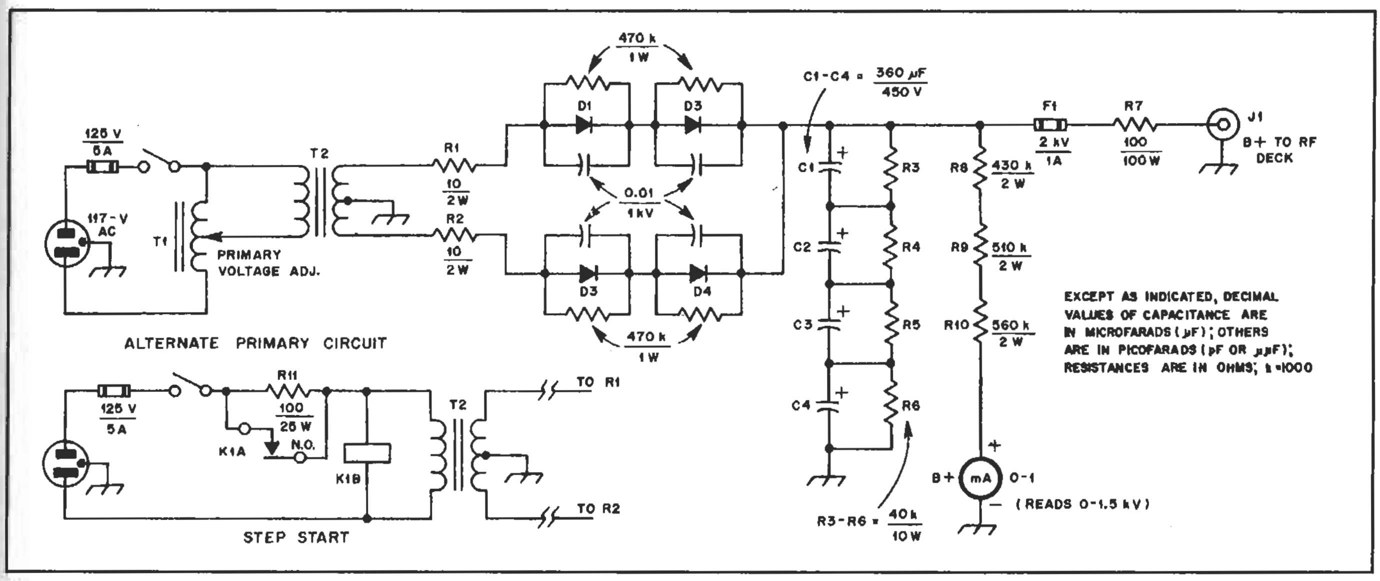

Fig 2—Schematic diagram of the amplifier high-voltage supply.

C1-C4—Electrolytic capacitor, 360 μF, 450 V.
D1-D4—Silicon rectifier, 1000 PIV, 3 A.
F1—High-voltage fuse, 2 kV, 1 A.
J1—Chassis-mount female BNC or MHV connector.
R3-R6—Wirewound resistor, 40 kΩ, 11 W.
T1—Variable autotransformer, 500 VA.
T2—High-voltage transformer. Primary, 117 V; secondary, 900 to 1050 V at 500 mA.

the bias supply. Q1 must be mounted on a heat sink. J1 is connected to the station TR switching system so that R1 is grounded on transmit and disconnected on receive. The approximate range of the bias supply is 6 to 20 V. Z1 and Z2 provide protection for Q1 in case of a shorted tube. The amplifier can be run without Z1 and Z2 if you keep the anode voltage below 1100 V.

High-Voltage Power Supply

A safe, reliable high-voltage power supply is described here. Of course, you can use any readily available HV supply; keep in mind, however, that the 7289/2C39 anode potential should never exceed 1400-V dc at full load and that the amplifier will withstand 1900-V dc at low cathode current and cutoff-bias conditions. For maximum power output, assuming adequate drive power is available, anode voltage under full load should be about 1200- to 1400-V dc.

Fig. 2 is a schematic diagram of the high-voltage supply. A power transformer (T2) that delivers 900- to 1050-V ac is ideal. The type of rectifier circuit used will depend on the type of transformer chosen. Each leg of the rectifier is made from two 1000-PIV, 3-A silicon diodes connected in series. Each diode is shunted with a 0.01-μF capacitor to suppress transient voltage spikes, and a 470-kΩ equalizing resistor.

Filtering is accomplished with a string of four 360-μF, 450-V electrolytic capacitors connected in series. R3-R6 equalize the voltage across each capacitor in the string and serve as bleeder resistors. Of course, a single oil-filled capacitor may be used here if available. Whatever type of filter you use, the total capacitance should be about 80 μF at a voltage rating of at least 1500-V dc. This value allows adequate "droop" of the anode voltage under high-current loads to protect the amplifier in case of RF overdrive or a defective tube.

Protective Circuitry

Some type of start-up protection should be incorporated in the primary. Fully discharged filter capacitors look like a dead short at supply turn-on. Initial surge current (until the capacitors charge) may be high enough to destroy the rectifiers. R1 and R2 provide some surge-current limiting, but either of the two primary configurations shown in Fig. 2 should be used. T1, a variable autotransformer (Variac and Powerstat are two common trade names), is ideal. In addition to allowing you to bring the primary up slowly (and charging the capacitors gradually), it also allows full control of amplifier output power by varying anode voltage.

The second method, a "step-start" system, uses a resistor in the T2 primary to limit the turn-on surge current. When the capacitors have charged, K1 is energized, shorting out R11 and applying full voltage to the T2 primary.

F1 and R7 protect against high-voltage arc-overs or short circuits. If sustained overcurrent is drawn, F1 will open and remove B+ from the RF deck. Use a high-voltage fuse here; standard fuses may arc when blown and not interrupt the B+ .R7 provides current limiting to protect the amplifier and power supply in case of a high-voltage arc.

Safety

An HV meter should always be used to monitor the status of the power supply. The values for R8-R10 shown in Fig. 2 will give a 1500-V dc full-scale reading on a 0-1 mA meter. RG-58 or -59 coaxial cable should be used for the high-voltage interconnection between the power supply and the RF deck. Ground the shield at both ends for safety and a good dc return.

Safety must be observed when working with all power supplies. These voltages are lethal! Always disconnect ac power and then discharge the filter capacitors before working on the power supply. Never guess or make assumptions about the status of a power supply. Assume it is hot.

Metering

Cathode-current monitoring is all that's really necessary for observing amplifier dc performance. Cathode current (I_K) is the sum of the plate (I_P) and grid (I_G) currents. Normally, when this amplifier is driven to 300- or 400-mA I_K, the grid current will be around 40 to 50 mA. The inclusion of a grid-current meter is not really necessary and only makes biasing and TR switching complicated.

Cooling

Desired output power and the level of drive power available will dictate what type of cooling to use. For intermittent duty (SSB, CW) at output levels less than 50 W, air cooling is satisfactory. Any small blower may be easily mounted to the aluminum box surrounding the tube anode. For high-duty-cycle modes and/or output levels greater than 50 W, water cooling is highly recommended. Greater than twice the normal air-cooled output power can be obtained from a water-cooled tube, and water cooling is quiet.

Tube Modification and Water Jacket

The first step is to remove the air radiator from the tube. The air radiator screws on, so it may simply be unscrewed without damage to the tube.

First, place a hose clamp around the tube anode. Secure the radiator fins in a vise and grip the hose clamp with a pair of large

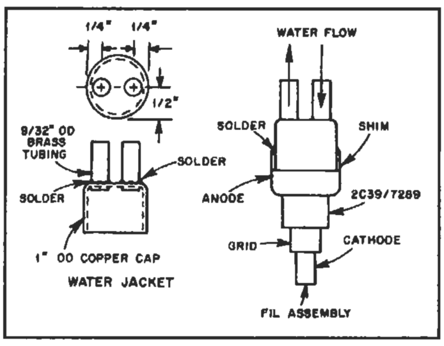
1/4"
1/4"
1/2"
9/32" OD BRASS TUBING
SOLDER
SOLDER
1" OD COPPER CAP
WATER JACKET
WATER FLOW
SOLDER
SHIM
ANODE
2C39/7289
GRID
CATHODE
FIL ASSEMBLY

Fig 3—Details of the solder-on water jacket.

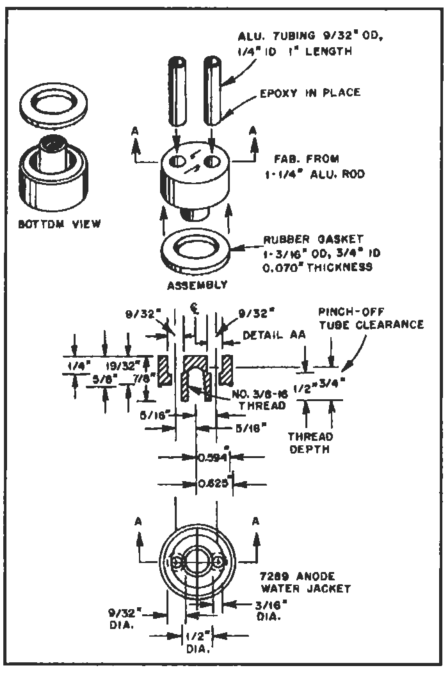
ALU. TUBING 9/32" OD, 1/4" ID 1" LENGTH
EPOXY IN PLACE
A
A
FAB. FROM 1-1/4" ALU. ROD
BOTTOM VIEW
RUBBER GASKET 1-3/16" OD, 3/4" ID 0.070" THICKNESS
ASSEMBLY
9/32"
℄
9/32"
PINCH-OFF TUBE CLEARANCE
DETAIL AA
1/4"
19/32"
5/8"
7/8"
NO. 3/8-16 THREAD
1/2"
3/4"
5/16"
5/16"
THREAD DEPTH
0.594"
0.625"
A
A
7289 ANODE WATER JACKET
9/32" DIA.
3/16" DIA.
1/2" DIA.

Fig 4—Details of the screw-on water jacket.

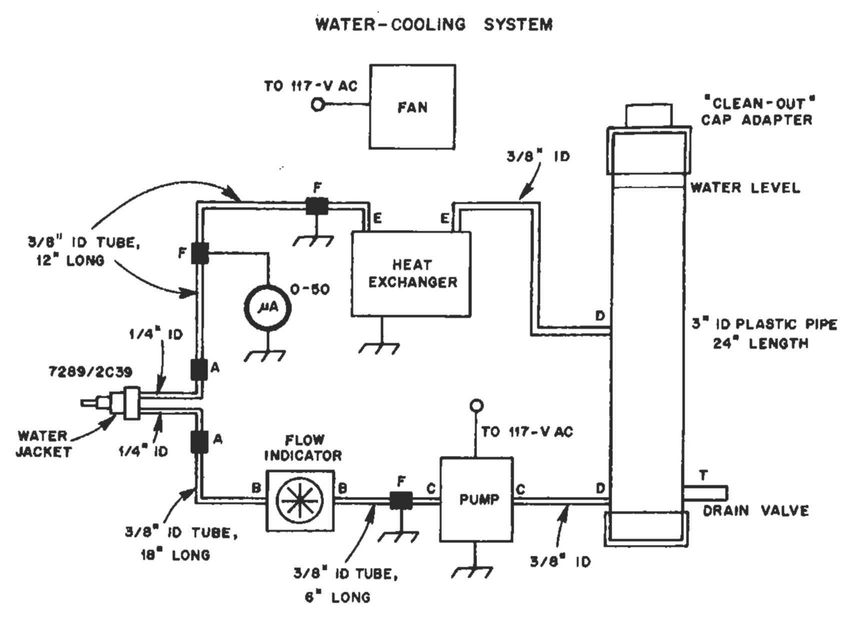
WATER-COOLING SYSTEM
TO 117-V AC
FAN
"CLEAN-OUT" CAP ADAPTER
3/8" ID
WATER LEVEL
F
E
E
3/8" ID TUBE, 12" LONG
F
HEAT EXCHANGER
0-50
μA
1/4" ID
D
3" ID PLASTIC PIPE 24" LENGTH
7289/2C39
A
WATER JACKET
1/4" ID
A
FLOW INDICATOR
TO 117-V AC
B
B
F
C
PUMP
C
D
T
DRAIN VALVE
3/8" ID TUBE, 18" LONG
3/8" ID TUBE, 6" LONG
3/8" ID

PART	DESCRIPTION		
A	1/4" TO 3/8" REDUCER.	D	3/8" OR 1/2" NPT TO 3/8" TUBE ADAPTER.
B	1/2" NPT TO 3/8" TUBE ADAPTER.	E	3/8" TUBE TO EXCHANGER ADAPTER.
C	3/8" TUBE TO PUMP ADAPTER.	F	3/8" OD COPPER TUBE W/WIRE SOLDERED TO IT. (SEE TEXT).

Fig 5—Details of the water-cooling system. Recommended pumps are: (1) Little Giant Pump Co. Model 1-42A or larger, available from most hardware stores; or (2) Calvert Engineering, Cal Pump Model 875S (160 gal/h), available from Calvert, 7051 Hayvenhurst Ave, Van Nuys, CA 91406, tel. 213-781-6029. The flow indicator (Model 15C; requires two 1/2-inch NPT adapters) is available from Proteus Industries, 240 Polaris Ave, Mountain View, CA 94043 tel. 415-964-4163.

pliers. Gently unscrew the tube from the radiator. If the hose clamp slips slightly, tighten it.

Some 7289/2C39 tubes use an air radiator that is attached with setscrews. To remove the radiator, simply remove the setscrews and pull the radiator off.

The air radiator will be replaced with a water jacket that allows water to be circulated past the tube anode and through a radiator, where it is cooled and circulated past the tube anode again. I have successfully used two different types of water jackets; both are described here.

The water jacket shown in Fig. 3 will work with any type of 7289/2C39. It is fabricated from a 1-inch-OD copper tubing cap and two short pieces of 9/32-inch-OD brass tubing. The copper tubing cap should be available from a local hardware store or plumbing supply house. Brass tubing is available from many hobby stores and metal supply houses.

Mark and drill the copper cap so that the brass tubing is a snug fit. Thoroughly clean the parts until they shine. Push the tubing into the holes in the end cap and degrease the assembly with alcohol. Using plenty of flux, solder the seam around each section of tubing. Allow the jacket assembly to cool.

Meanwhile, thoroughly clean the 7289/2C39 anode to a bright finish. Check the water jacket for fit. In some cases, you'll have to use a 0.005- to 0.010-inch-thick copper shim to fill the gap between the copper cap and the tube anode. This shim helps eliminate pin holes in the solder.

Using plenty of flux, solder the water jacket to the tube anode. Solder it quickly with a hot, high-wattage iron. Allow the tube to cool in the air after soldering to avoid thermal shock and possible breakage. After the tube has cooled, use plenty of alcohol to remove all traces of flux from the tube and water jacket.

The second type of water jacket is shown in Fig. 4. This jacket will work only with 7289/2C39 tubes that have a screw-on air radiator. It is designed to thread onto the tube anode just like the air radiator did. This jacket is machined from a piece of 1¼-inch aluminum rod. The water inlet and outlet tubes are made from 9/32-inch-OD, 1/4-inch-ID aluminum tubing that is epoxied in place. A rubber gasket seals the jacket against leaks.

If you have access to a lathe, you should have no trouble duplicating the jacket. You could have one made up at a local machine shop. Complete screw-on water jackets are also available from the author.

After you unscrew the air radiator from the 7289/2C39, check for and remove any burrs from the tube anode. The anode surface must be flat if the rubber gasket is to be effective. Screw the water jacket onto the tube. Tighten by hand only. Do not use any tools, or you could damage the tube or jacket! Do not use the water inlet and outlet tubes for leverage—they have thin walls and break easily.

Water System

Fig. 5 depicts the complete water-cooling system. Recommended pumps and accessories that have proven reliable and effective are listed in the caption.

Any small pump, such as a fountain pump, that can deliver 160 to 200 gallons per hour can be used here. Most inexpensive pumps are not self-priming, which means that they won't pump water if they have air in the rotor. Although water can be forced through the pump for the initial prime, my system uses gravity priming. The water reservoir is a 2-foot length of 3-inch-OD plastic pipe that is available from hardware or plumbing stores. The outlet is at the bottom, and the inlet about halfway up the column. The inlet is located here to eliminate aeration that ionizes the water and reduces its effectiveness. The outlet

Microwave Radiation Safety

Intense RF radiation concentrated on body tissues can produce heat damage; the extent and penetration will depend on the radio frequency in use and on exposure duration. You should be aware of the approximate intensity of RF radiation of the transmitting equipment and antennas you come in contact with.

RF intensity is commonly expressed in milliwatts per square centimeter (mW/cm^2), which is the power flowing away from a source through a unit sampling or interception area at some specified distance. Although the United States as yet has no federal RF protection standard, a useful interim guide is the 1982 standard of the American National Standards Institute (ANSI '82). The most stringent level in this standard is 1 mW/cm^2 for frequencies between 30 and 300 MHz. Above 300 MHz, the protection level rises until it reaches 5 mW/cm^2 at 1500 MHz. Beyond 1500 MHz, the recommended level remains at 5 mW/cm^2. These levels represent the average power density allowed over any six-minute period and are for the sum of all polarizations from a given source.

At 1296 MHz, where one wavelength (λ) equals 23 cm, a thick resonant dipole feeding a calibrated power meter with matched coaxial cable (itself free of pickup) may be used to obtain an indication of power density. A reasonably lossless resonant dipole has an effective aperture of $\lambda^2/8$; at 23 cm this is 66 cm^2. The power meter reading in milliwatts, divided by 66, is the indicated power density. For this to be a reliable indication, the dipole must be positioned far enough from the RF source to be in its far field. For a small source, the distance should be at least $\lambda/2$, and here that would be about 12 cm (45 inches). The dipole should be oriented for alignment with the dominant polarization. Note that the power meter must be capable of readings well below 1 mW.

This arrangement would be useful for checking leaks along the coaxial route that the high power (here 250 W) takes to a load, be it dummy load or antenna. Cable connectors may not be tightly secured, or they may be faulty. For equipment operating in the SHF region, waveguide flanges may not be clamped properly.

Direct measurement of electric field strength near an antenna (with a calibrated instrument, preferably one with the indicating meter shielded and possibly positioned at the center of the sampling dipole) is another way to check for adequate protection. A field strength of 60 V per meter (V/m) corresponds to 1 mW/cm^2; 134 V/m corresponds to 5 mW/cm^2. At a distance 60 cm (2 feet) from an isolated dipole fed with 26 watts, the field strength would be about 60 V/m. This is a far-field field strength for all frequencies where the half wavelength is less than 60 cm, or for frequencies above 250 MHz. For full 250 watts applied to the dipole, the 60 V/m level occurs at a distance of 1.8 meters (6 feet), and at this distance this holds for all frequencies above 80 MHz.

With SSB or CW keying, the fields during Amateur Radio operation are highly intermittent, and usually include considerable pauses or intervals for listening. These factors reduce the average power density over the six-minute averaging period.

Further information on RF safety and protection estimates can be found in Chapter 7 of *The Satellite Experimenter's Handbook,* published by the ARRL. The following rules of good practice for RF protection are recommended:

- Never operate an RF amplifier with equipment shielding removed.
- Never handle antennas with RF power applied.
- Never guess that RF levels are safe. Take the time to consult a reliable reference for an estimate, or measure levels carefully. Allow a "cushion" of about 6 dB (factor of four in power density). If possible, borrow an RF radiation monitor (after learning how to use it), or consult with a ham who is well informed on RF protection.
- Never look into an open end of a power waveguide; never point a powered directive antenna (a beam or a paraboloid, for example) toward people. Keep all VHF and UHF transmitting antennas as high as possible, distant from humans.
- Use good-quality, well-constructed coaxial cable and connectors to avoid RF leaks.
- Think RF and electrical safety first; test later!
- Watch *QST* for news on RF measurement techniques and progression, protection standards and proposed federal and state RE regulations.*—David Davidson, W1GKM*

directly feeds the pump. The pump and the reservoir outlet port should be mounted in the same plane. The pump should be oriented so that air bubbles will rise into the impeller output port and can be blown out once the pump starts running.

Flow Indicator and Heat Exchanger

Water cooling is best described as "super quiet." There is no noisy fan to reassure you that the tube is receiving adequate cooling. If water flow is reduced or cut off during amplifier operation, tube damage is virtually assured.

Flow interlocks and switches to shut down the amplifier if water flow is reduced are hard to find and expensive. Flow indicators, however, are inexpensive and reliable. A flow indicator has a spoked rotor that turns as water passes through the unit. If the wheel is turning, there is water flow; if not, you have a problem. Changes in flow rate can be observed by watching for speed changes in the rotor. A small lamp illuminates the flow indicator, making it easy to see rotation. The flow indicator should be mounted where it can be seen from the operating position and monitored during operation.

Heat exchangers, or radiators, remove the heat from water as it passes through. For this application, a small automobile transmission-oil cooler works great. Most auto- parts stores and speed shops have a good selection. Pick one that is similar in size and aspect ratio to a whisper fan (approximately 4 × 4 × 1 inches). Some come with mounting brackets. Look for a cooler with the input and output ports on the top so air bubbles will rise to the top and move on without becoming trapped. Trapped air degrades cooler performance.

If you use the amplifier for high-duty-cycle modes such as ATV or FM, or for long, slow-speed CW transmissions (EME, for example), you should use a small axial whisper fan to increase the effectiveness of the heat exchanger. A fan isn't necessary during normal operation, or even for sustained operation at moderate power levels, but I highly recommend one if you plan prolonged operation at maximum power. Locate the fan so the warm exhaust air won't heat up other equipment.

Hoses and Fittings

Most hardware stores carry a complete line of brass fittings and adapters that can be used for this project. Brass, however, will eventually corrode and pollute the water supply. Plastic fittings are cheaper and don't corrode, but they are harder to find. Recreational vehicle suppliers are my main source for these parts. They are used extensively in drinking water systems for mobile homes and travel trailers. Procure the fittings when you have the rest of the parts in hand, as there are many variables to consider.

You can use any relatively soft, thin-wall vinyl tubing for all water lines. The main runs are made from 3/8-inch-ID hose, while 1/4-inch-ID stock is used to connect to the 7289/2C39 water jacket. The 1/4-inch-ID tubing fits snugly over the 9/32-inch-OD inlet and outlet tubes on the water jacket, so no clamps are required. All other hose connections should be secured with stainless-steel clamps to prevent leaks. Any leaks mean air in the system and deterioration of cooling performance.

Safety

The tube anode, and hence the water jacket and water, are in direct contact with the high-voltage supply, so some safety precautions must be observed. Approximately 12 to 18 inches of tubing should run between the 7289/2C39 jacket and any other component in the cooling system. This will allow enough resistance in the water to provide adequate current limiting, should the water contact any components that are grounded.

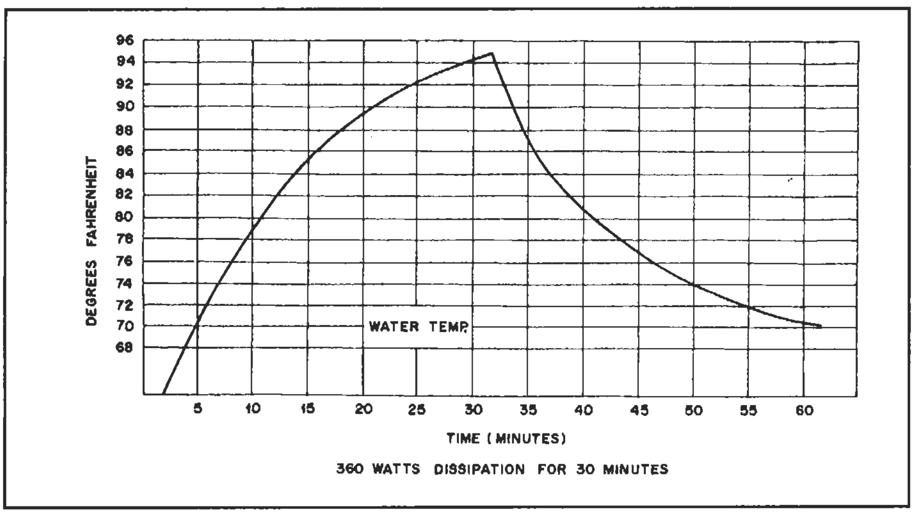

Fig 6—Performance graph of the water-cooling system.

It is best to ground the water supply at the pump. Do this by replacing a short section of the tubing that runs to the flow indicator with a piece of brass or copper tubing. Solder a wire to this metal tubing and connect the other end of the wire to your station ground. Use at least 24 inches of vinyl tubing between the anode cooling jacket and the ground point.

On the warm-water side of the 7289/2C39, run 12 inches of vinyl tubing to a small metal fitting or short section of metal tubing, and then another 12 inches of vinyl tubing to a grounded point (this can be at the heat exchanger). You can measure the water leakage current to ground by placing a microammeter between the metal fitting that connects the two vinyl hoses and ground. Leakage current should be less than 10 μA with clean water and an anode potential of 1 kV. As the water ages, the leakage current will rise; when this happens, replace the water.

Grocery stores carry distilled water for use in steam irons. It may be deionized and not truly distilled, but it works fine for about four to six months in this application. Filters can be purchased from scientific supply houses, but they're not really worth buying because deionized water is so cheap.

Do not use tap water under any circumstances! When you turn on the water system for the first time, run a gallon of water through it for half an hour to wash out fabrication impurities. Replace with clean water before using the system to cool the amplifier.

Water was chosen because it's inexpensive, nontoxic, nonflammable and easy to clean up if you have a leak. Better liquid coolants are available, but they are toxic. Don't use them!

Cooling Performance

I have used water-cooling systems for several years with no problems whatsoever. Fig. 6 is a graph of several transmit/receive cycles on a water-cooled, 500-W output, 23-cm power amplifier. For this test, I used two of the amplifiers described in this article coupled with a pair of hybrid combiners. This particular cooling system used 1 gallon of water. Experiments indicate that, during extended operation, the water temperature rises only 30° to 35°F above ambient room temperature. Typically, the tube anode and water average 10° to 15°F above ambient during casual operating.

Flow rates in this system are typically $^{1}/_{3}$ gallon per minute per tube, which is more than adequate. At this rate, more than 300 W of dissipation from a single inefficient 7289/2C39 were required to boil the water in the water jacket. The water should not be allowed to boil because this will heat the rubber gasket.

Tubes

It is not really necessary to buy a new 7289/2C39. Used tubes can be found surplus for around $1 to $5 and, in many cases, will perform as well as a new tube. Most used tubes have been sitting around for several years, so it's a good idea to run them through the dishwasher to clean them up and then run the filaments for about 24 hours. This will restore operation in many cases.

If you buy a new tube, you should be aware that the 7289/2C39 is being run far in excess of its ratings in this amplifier. The manufacturer's warranty will not cover tubes run in this application.

Contrary to popular opinion, glass tubes will work. Physically, they are not as rugged as the ceramic version, but the glass-to-metal seal seems to provide better shelf life than the ceramic seal. The glass tubes make great driver tubes and will work fine for power levels up to 100-W output. Pulse tubes (7815, 7211) are not recommended because of their poor thermal stability at high power levels. Also, they generally are 30 to 40 MHz lower in resonant frequency in this amplifier compared to the 7289/2C39. Some 7289 tubes can be as much as 30 MHz lower in frequency. Minor length adjustment of the anode-tuning post may be required to accommodate amplifier and tube differences.

Tube Insertion

Extreme care must be exercised when inserting the 7289/2C39 tube. Never force the tube in place, as damage (bending) of the cathode finger stock may result. Observe the layout of the finger stock to get an idea of how the tube inserts. Carefully position the tube so it is straight as you gently push. It should slide in snugly without any solid resistance.

Testing

After you have completed all of the parts for the amplifier, it's time to test everything before hooking it all together. Test the water-cooling system by turning it on and watching for steady water flow as indicated on the flow meter. The tube and water jacket can be removed from the cavity amplifier for this test.

Check all of the power-supply voltages first without connecting them to the RF deck. Then, without the tube in place, hook the bias and filament supplies to the cavity and check the voltages again at the tube finger-stock connections. Connect the high-voltage supply to the RF deck and bring the voltage up slowly with a variable autotransformer. Monitor the high voltage on the anode-bypass-capacitor plate, and look and listen for any possible arcing between the anode-bypass-capacitor plate and ground. Use extreme care when measuring and testing the high-voltage supply. If everything looks okay with the power supplies, shut them off and disconnect them.

You can make a safe, low-power test of the cavity resonance without applying any voltage. With the tube in place, insert a 2-inch-long coupling loop on the end of a piece of coaxial cable between the spring fingers of the anode down into the cavity. Connect the amplifier output probe/connector to a device capable of detecting low-level RF at 23 cm (for example, a spectrum analyzer or microwattmeter). Feed a signal from an L-band signal generator into cable attached to the wire coupling loop that you inserted into the cavity. Set the signal generator for various frequencies in the 23-cm band and tune the amplifier anode tuner. There will be sharp peak in output at cavity resonance.

This testing method can be used to determine cavity tuning range, anode-bypasscapacitor effectiveness and resonance of various tube types for use in this amplifier. Any cavity amplifier can be tested completely without ever applying high voltage. The better your test equipment, the easier the amplifier is to test. If all dimensions were followed strictly, the amplifier will tune as designed.

Amplifier Hookup

Installation and operation of this amplifier is relatively straightforward, but as with any amplifier, several precautions must be followed. If these are adhered to, the amplifier will provide years of reliable service.

The amplifier is designed to be operated in a 50-ohm system and should never be turned on without a good 50-ohm load connected to the output connector. Never operate it into an antenna that has not been tuned to 50 ohms!

Drive power to the amplifier should *never* exceed 15 W. *Never* apply drive power in excess of 1 W unless all operating voltages are present and the tube is biased on. Otherwise, the tube grid-dissipation rating will be exceeded and you will probably ruin it.

As in all TR-switched systems, some type of interlock or sequencing of transmit and receive functions should be incorporated. In most systems, the sequence for going into transmit is something like this: First, switch the antenna changeover relay from the receiver to the power amplifier. Next, bias the power amplifier on. Last, key the exciter and apply drive to the amplifier. To go to receive, unkey the exciter, remove operating bias from the amplifier and switch the antenna relay back to the receiver.

If the antenna relays are switched while the power amplifier is operating and putting out power, damage to the relay contacts and/or the amplifier is likely. If there is a momentary removal of the antenna while the power amplifier is biased on, oscillation may occur. This can damage the TR relay, the tube or even the receive preamplifier.

Tune-up and Operation

This is it — the big moment when you will see your project come to life! Connect an accurate UHF power meter and a 50-ohm antenna or load to the amplifier output connector. A Bird Model 43 wattmeter with a 100- or 250-W, 400-1000 MHz slug will give reasonable accuracy, depending on the purity of the drive signal. Apply filament power and tube cooling, and allow 3 to 5 minutes for the filaments to warm up. Turn on bias supply (the amplifier will draw maximum current if the anode voltage is applied without bias). Apply 300 to 400 V to the anode. There should be no current flowing in the tube as indicated on the cathode-current meter. Ground J1 on the bias supply to apply transmit bias and observe cathode current. As R1, the bias control, is turned clockwise, quiescent idling current should increase. Set for about 25 mA.

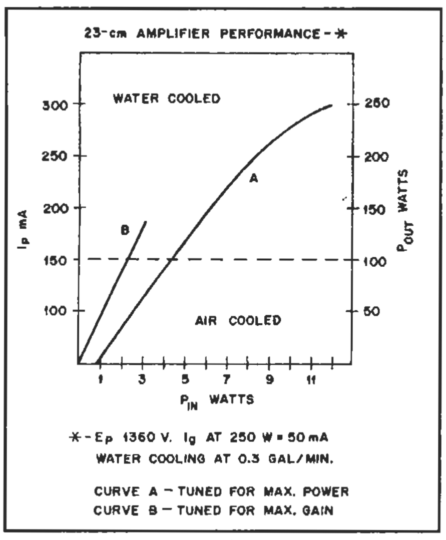

Fig 7—Performance of the cavity amplifier under different drive and plate-current conditions.

Apply 1 W of RF drive power. Turn the anode tuner while observing the RF output power meter and tune for maximum output. The output should go through a pronounced peak at'cavity resonance. Adjust C2 and C3 on the input tuning network for maximum amplifier output. If possible, use a directional wattmeter between the driver and the amplifier input to check that best input SWR and maximum amplifier output occur at roughly the same setting.

Depending on the amount of drive power available, you may want to tune the amplifier for maximum power output or maximum gain. Fig. 7 shows what you can expect from different drive levels.

Once the amplifier is tuned for best input SWR and maximum output with 1 W of drive, anode voltage and drive power can be increased. Increase both in steps; be sure to keep the anode tuner peaked for maximum output power. When you get to the 100-W output level, very carefully readjust the input circuit for maximum output. The input capacitor closest to the cathode is critical and should need to be rotated less than 90 degrees maximum. Maximum output power will be roughly coincident with best input SWR.

Increase the drive power and keep the anode tuner peaked for maximum output. Increase the drive until you reach the desired output level, but *do not exceed 400-mA I_K!* At 1400-V dc and 350-mA I_K, output power with a good tube should be about 230 to 250 W. At lower anode voltages, I_K will be higher for the same output power. Higher anode voltages result in higher gain, lower drive levels, lower grid current and lower plate current for a given output power.

The anode tuner's tuning rate is approximately 5 MHz per turn. Clockwise rotation of the tuner lowers the resonant frequency of the cavity. This control will require readjustment as you make large frequency excursions within the 23-cm band (for example, if you go from 1296 weak-signal work to the 1269-MHz satellite segment). You should also check the input SWR if you move more than 15 MHz. Generally, amplifier tuning does not change much after initial setup. You should be able to turn it on and use it without retuning as it heats up. Slight adjustments may be necessary, however, depending on cooling, inherent thermal differences from tube to tube and duty cycle of the operating mode. Always keep the anode tuner peaked for maximum output, and check it from time to time, especially while you are first learning how the amplifier operates.

The output loading control is the output connector and probe assembly. Loading is changed by minor rotational adjustment of the N connector. First loosen the jam-nut (or setscrew) slightly. While observing output power and keeping the anode tuner peaked, rotate the loading control ±30 degrees maximum for greatest output power. This should be done only once and should not need repeating unless another tube is installed. Even then it may not be required.

Conclusion

This cavity amplifier for the 23-cm band is capable of safe, reliable operation at output powers in excess of 200 W. More than 50 of these amplifiers are in operation, and you can build one, too. I would like to thank Mike Stahl, K6MYC, Bill Troetschel, K6UQH, William Jungwirth, AA6S, Lem Moeschler, W6KGS and Joseph Cadwallader, K6ZMW, for their help and encouragement during the development of this project.

References

Laakmann, "Cavity Amplifier for 1296 MHz," *QST*, Jan. 1968, pp. 17-22.
Laakmann, "Kilowatt Amplifier for 1296 MHz," *Ham Radio*, Aug. 1968, pp. 8-17.
"Power Amplifier for 23 cm," *VHF Communications*, April 1976, p. 222.
"Power Amplifier for 24 cm," *VHF Communications*, March 1978, p. 175.
Stefanskie, "1270 MHz ATV Power Amplifier," *Ham Radio*, June 1977, pp. 67-73.
Tech Report #6, Crawford Hill VHF Club, July 1971.
Tech Report #13, Crawford Hill VHF Club, Dec. 1972.

By Zack Lau, KH6CP/1 **From *QEX*, May 1994**

A 2-W, 13-cm Amplifier

This amplifier uses a resistively stabilized Hewlett Packard ATF-44101/AT-8140 power GaAs FET running class A to provide 14 dB of gain and a 2-W output. It wasn't too difficult to choose the device—there aren't that many linear devices that cover the 2.3 GHz band at this power level. Unfortunately, it is fairly pricey; the circa 1990 price was around $90. But since this transistor has been around a long time, it may be available surplus. Several hundred were sold at the bargain price of $3 each about a year and a half ago. I bought a few—I should have bought more.

Even if you do buy your FETs at bargain prices, you still will want to ensure that they are biased safely. As has been pointed out many times in the literature, having the negative gate bias supply fail while the drain supply is applied may result in the destruction of the device. Thus, designers have devised elaborate protection schemes to shut off the drain supply if the gate supply fails.

Fig 1 shows a simpler power-supply circuit. The drain supply is controlled by an ordinary 723 voltage regulator with current limiting. The current limiting protects the device—instead of the expensive FET getting hot and self-destructing, the cheap TIP 30 power transistor gets warm. Doing this eliminates the need for complex shutdown circuits to handle a failed bias supply. While there are more modern regulator chips than the 723, they aren't significantly better for this application. In fact, some are "improved" to the point where they are tougher to use. For example, chips with better current-sensing circuits often require special low-value resistors. While you could use the copper wire table in the *ARRL Handbook* to make your own low-value resistors from wire, it makes more sense to me to use a 723 with standard, readily available parts.

The gate bias supply is produced by an NE555 timer running as an oscillator and driving a voltage inverter. An LM337 adjustable 3-terminal voltage regulator controls the gate supply. Now, you have to be a little careful using adjustable voltage regulatorswith bias supplies. While I haven't experienced this myself, it is entirely possible that a noisy potentiometer could momentarily present an open circuit between the wiper and the resistance element. Conceivably, in some circuits, this could result in voltage spikes at the output of the voltage regulator as the voltage is adjusted. The cure for this is to wire the potentiometer so that it never presents an open circuit, even with an intermittent wiper, as I've done with R9 in Fig 1.

For some other GaAs FET circuits, it may be advantageous to use an active-bias supply which operates inside the feedback loop of the current limited supply. This allows the bias point to be unaffected by temperature variations. The current limiting acts as a failsafe, protecting the transistor if a more negative gate supply voltage is unable to turn off the transistor. I've successfully used this circuit when only a milliamp of bias current was needed. Due to the high current required by the gate circuit of the 2-W amplifier, I decided it was impractical to active-bias the FET in this application; the ICL7660 can't provide the needed current. But there are highercurrent versions of this chip, for those who don't mind the price.

RF Design

I found the AT-8140 rather difficult to get unconditionally stable. I ended up placing a 10041 chip resistor from gate to ground, which resulted in pretty good stability, except around 2.3 GHz. Unfortunately, it doesn't appear possible to dc decouple this resistor without adversely affecting stability, so quite a bit of bias current is needed. Because of this, it is conceivable that you may find a FET that needs too much gate voltage for a typical 10041 chip resistor to handle. Whether this is the case with a particular FET depends on both the threshold voltage and the transconductance of the device. The threshold voltage can vary quite a bit, particularly in older devices, though the manufacturers have tightened up this specification in recent data books. Worst case would be a large negative threshold voltage and a high transconductance. Stacking chip resistors in parallel might be a solution, though this might also invite unwanted parallel resonance effects, similar to what is seen with paralleled capacitors.

While in-band stability is not too critical, since the source and load impedance matching is usually pretty good at the operating frequency, I used a series resistor in the gate decoupling line to make the amplifier unconditionally stable, according to

Table 1—Measured Amplifier Performance

Input to MGF 1801 driver (dBm)	*Driver Output (dBm)*	*AT 8140 Output (dBm)*	*AT 8140 Gain (dB)*
–7.0	5.7	19.8	14.1
0.0	12.5	26.6	14.1
3.0	15.5	29.7	14.2
4.0	16.5	30.8	14.3
5.0	17.5	31.7	14.2
6.0	18.4	32.6	14.2
7.0	19.2	33.3	14.1
8.0	20.3	33.9	13.6
9.0	21.2	34.1	13.1
10.0	22.1	34.6	12.5
13.0	24.2	34.9	10.7

The 1-dB compression point is at 34.1/33.2 dBm output.
Comparing measurements made using an HP8563E spectrum analyzer against those made with an HP 453B/8481A power meter, a signal that measures 4.7 dBm on the spectrum analyzer reads out as 3.8 dBm on the power meter. The values listed are from the spectrum analyzer, taking into account a 30.2-dB, 25-W Bird attenuator inserted between the amplifier and the analyzer.

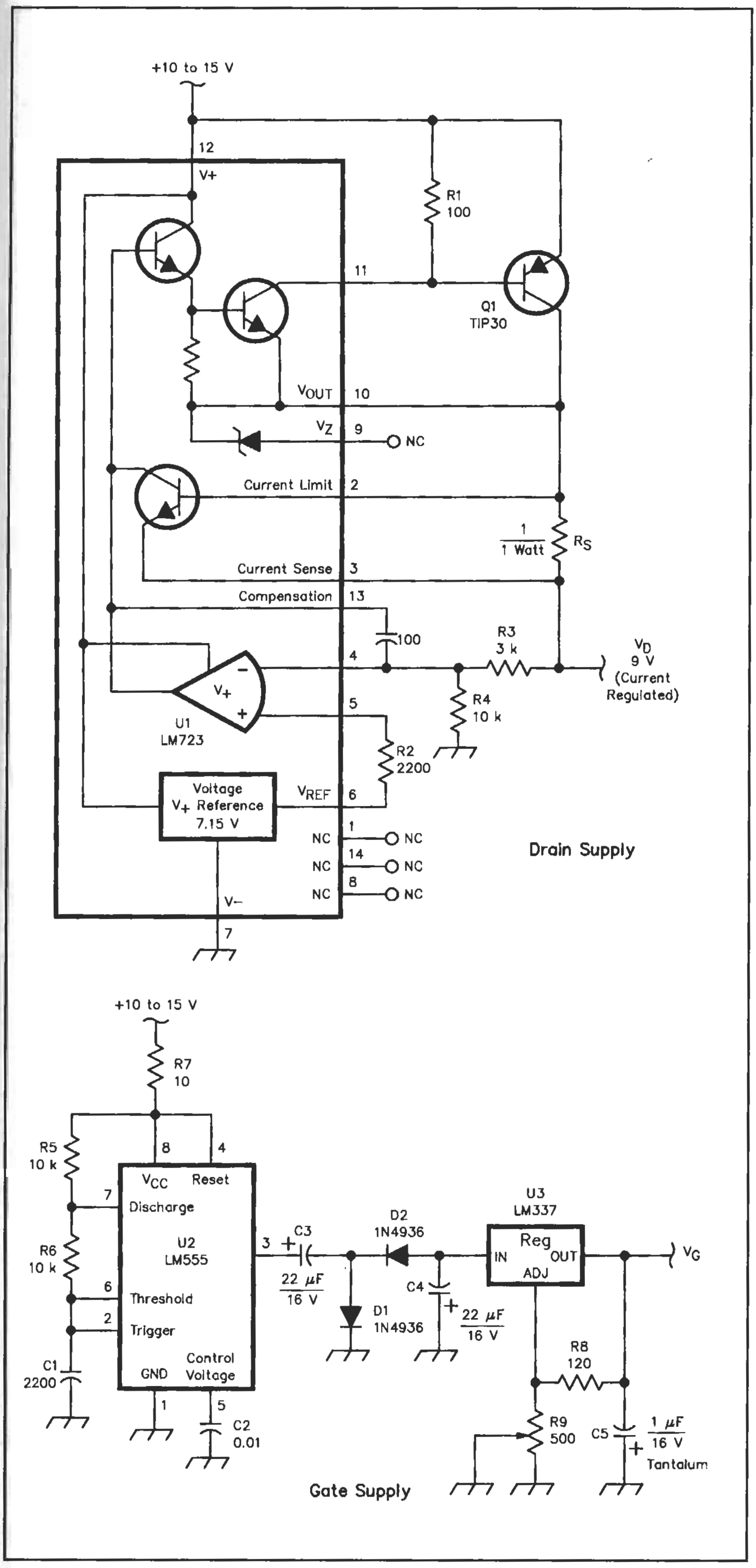
+10 to 15 V
12
V+
R1
100
11
Q1
TIP30
VOUT
10
VZ
9
NC
Current Limit
2
1
1 Watt
RS
Current Sense
3
Compensation
13
100
R3
3 k
4
VD
9 V
(Current Regulated)
V+
U1
LM723
5
R4
10 k
R2
2200
Voltage V+ Reference 7.15 V
VREF
6
1
NC
14
NC
8
NC
V–
7
Drain Supply
+10 to 15 V
R7
10
R5
10 k
8
4
VCC
Reset
7
Discharge
U2
LM555
R6
10 k
6
Threshold
2
Trigger
C1
2200
Control Voltage
GND
1
5
C2
0.01
3
C3
22 µF
16 V
D2
1N4936
D1
1N4936
C4
22 µF
16 V
U3
LM337
Reg
IN
OUT
ADJ
VG
R8
120
R9
500
C5
1 µF
16 V
Tantalum
Gate Supply

Fig 1—Power supply for the 2-W, 13-cm amplifier.

the computer model. This series resistor also improved the input match bandwidth, an important consideration if you don't want to tune the amplifier. Of course, improving the bandwidth with resistance also reduces the gain. The loss isn't too bad, however, seeing as this design is within a few dB of the maximum stable gain the data sheet specifies. The design was optimized with 150 Ω of series resistance, but I decided that such a high value was unwise. To accommodate such a high series resistance, it is entirely likely that the negative supply voltage would have to be more negative than the VGS limit. Then, if the 10041 resistor were to fail opencircuited, the supply could damage the gate of the transistor. For this reason, I chose to use a 5041 series resistor. I didn't reoptimize the design for 50 Ω of series resistance, since the likely improvement was small compared to the variations I've seen when optimizing the unit on the bench.

Finally, I discovered that additional gate circuit bypassing was required at audio frequencies to prevent the circuit from oscillating. This was done with a 1-µF capacitor that is not shown in the computer model. This frequency range wasn't covered by the model.

Several dB of additional saturated output power was obtained by modifying the output network with foil tabs. AT-8140s saturate at about 2.2 W output with around 200 mW of drive. An advantage to optimizing the circuit by modifying the 5041 striplines at the input and output of the amplifier is that the stability should not be adversely affected. But the board does have to be made larger to accommodate such tuning.

The original design was optimized using *Microwave Harmonica*, with the results shown in Table 2 and Fig 6. After I tweaked the output network to get maximum output power, I entered the changed circuit components into the computer and analyzed the circuit, resulting in the analysis shown in Table 3 and Fig 7. Table 1 shows the performance of the circuit in its final configuration.

Construction Notes

To ground the drain bypass capacitor and the gate resistor, I cut slots in the board and connected the pads to the ground plane with 1-mil copper foil. Copper foil was also used to connect the gate bypass capacitor, although in this case I trimmed the board so that it wasn't necessary to cut a slot.

For high-quality grounding without lots of tiny screws, I decided to try using a heat spreader made out of 1-mil copper foil. I cut a hole in the circuit board for the FET and then carefully soldered the foil across the hole in the circuit board, taking care not to put any solder where it would interfere with heat-sinking action. I didn't want any air pockets, as they are extremely poor conductors of heat. Then, after attaching the board to a piece of 0.25-inch sheet aluminum with suitably tapped holes, I used a scribe to punch holes for the 0-80 screws to

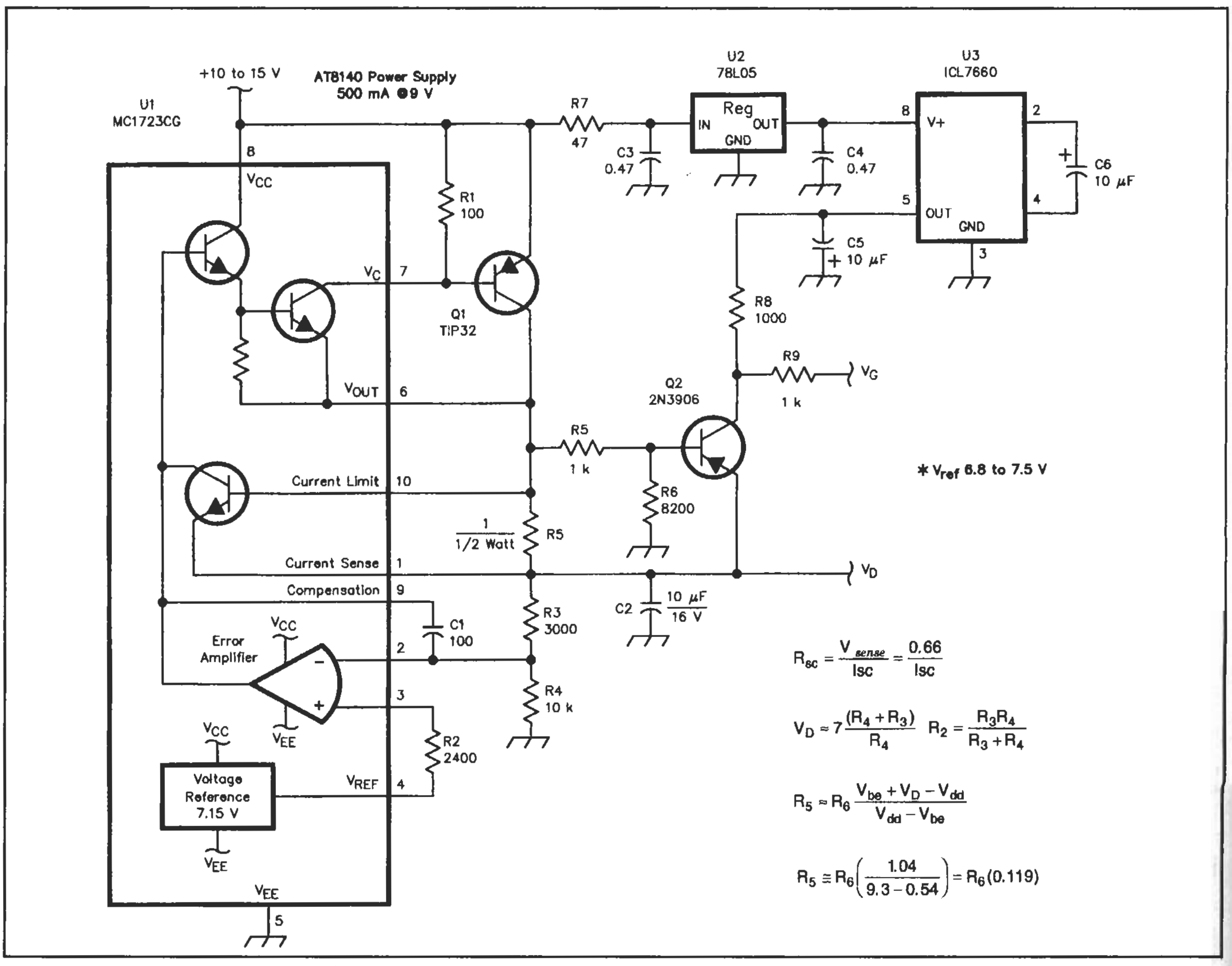
U1
MC1723CG
+10 to 15 V
AT8140 Power Supply
500 mA @9 V
8
V_{CC}
V_C
7
R1
100
Q1
TIP32
V_{OUT}
6
Current Limit
10
Current Sense
1
Compensation
9
Error Amplifier
V_{CC}
V_{EE}
2
3
C1
100
Voltage Reference 7.15 V
V_{CC}
V_{EE}
V_{REF}
4
V_{EE}
5
R2
2400
1/1/2 Watt
R5
R3
3000
R4
10 k
R7
47
C3
0.47
U2
78L05
Reg
IN
OUT
GND
C4
0.47
C5
10 µF
U3
ICL7660
8
V+
2
C6
10 µF
5
OUT
GND
4
3
R8
1000
R9
1 k
V_G
Q2
2N3906
R5
1 k
R6
8200
C2
10 µF
16 V
V_D
* V_{ref} 6.8 to 7.5 V

$$R_{sc} = \frac{V_{sense}}{Isc} \approx \frac{0.66}{Isc}$$

$$V_D \approx 7\frac{(R_4+R_3)}{R_4} \quad R_2 = \frac{R_3R_4}{R_3+R_4}$$

$$R_5 \approx R_6\frac{V_{be}+V_D-V_{dd}}{V_{dd}-V_{be}}$$

$$R_5 \cong R_6\left(\frac{1.04}{9.3-0.54}\right) \approx R_6(0.119)$$

Fig 2—An active biasing circuit suitable for use with amplifiers that draw minimal gate current. Do *not* use this with the 2-W amplifier.

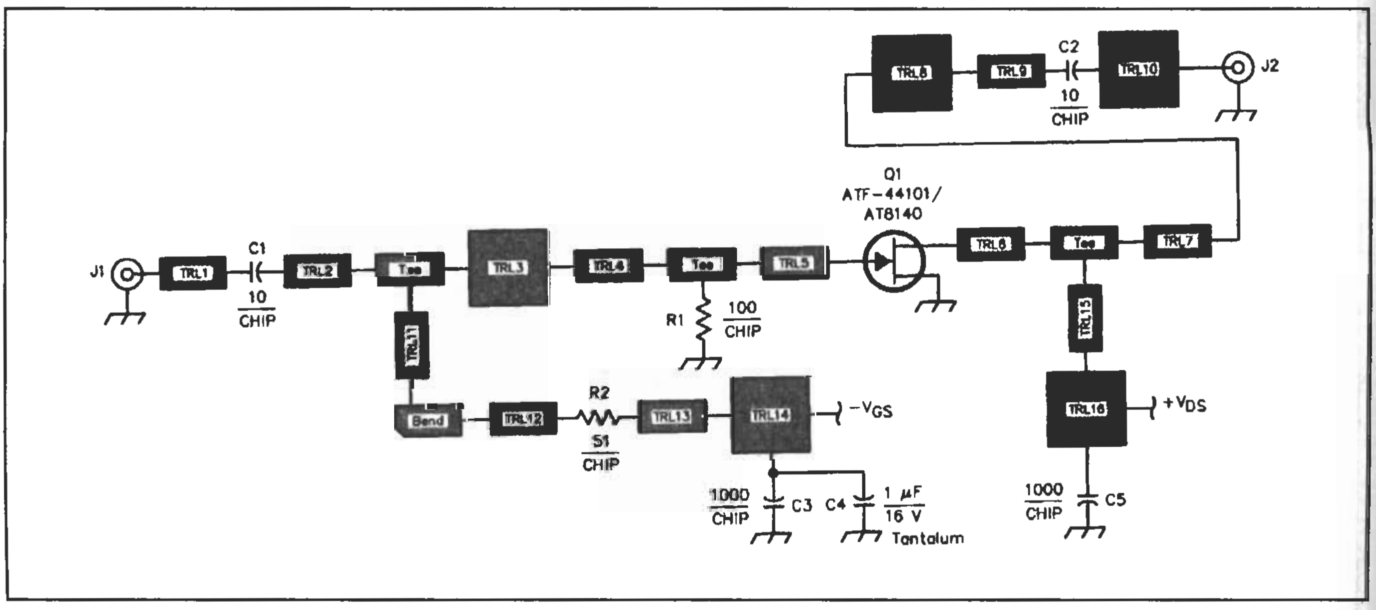
J1
TRL1
C1
10
CHIP
TRL2
Tee
TRL3
TRL4
Tee
TRL5
TRL11
Bend
TRL12
R2
51
CHIP
TRL13
TRL14
$-V_{GS}$
R1
100
CHIP
1000
CHIP
C3
C4
1 µF
16 V
Tantalum
Q1
ATF-44101/
AT8140
TRL6
Tee
TRL7
TRL15
TRL16
$+V_{DS}$
1000
CHIP
C5
TRL8
TRL9
C2
10
CHIP
TRL10
J2

Fig 3—Schematic diagram for the 2-W amplifier.

C1, C2—10-pF, high-quality porcelain chip capacitors. I used 55-mil capacitors, although 100-mil capacitors should work just fine.
C3, C5—1000-pF NPO chip capacitors
C4—1-µF, 16-V tantalum capacitor.
J1, J2—SMA jacks.
Q1—Hewlett Packard ATF-44101/AT-8410 medium power GaAs FET. This device must be properly heat-sinked. The absolute maximum channel temperature is 175°C. The case-to-junction thermal resistance is 23°C/W.
TRL1-16—Microstriplines etched on 0.031-inch-thick, $\varepsilon_r = 2.55$ Teflon circuit board.

Table 2—*Microwave Harmonica* Analysis (Original Circuit)

MICROWAVE HARMONICA PC V5.5 File: \mh\2304\qexamp.ckt 22-DEC-93 17:54:54

Freq GHz	MS11 dB AMP	PS11 deg AMP	MS21 mag AMP	PS21 deg AMP	MS12 mag AMP	PS12 deg AMP	MS22 dB AMP	PS22 deg AMP	MS21 dB AMP	K AMP
1.000	-0.70	-167.5	2.084	19.3	0.017	-53.7	-6.01	-128.7	6.38	1.64
1.100	-0.57	-177.2	1.988	0.1	0.018	-71.2	-5.35	-139.0	5.97	1.43
1.200	-0.47	173.3	1.941	-17.8	0.020	-87.3	-5.02	-148.2	5.76	1.27
1.300	-0.40	164.0	1.945	-35.0	0.022	-102.6	-4.96	-156.6	5.78	1.17
1.400	-0.36	154.7	1.999	-52.0	0.024	-117.7	-5.13	-164.3	6.01	1.09
1.500	-0.38	145.2	2.110	-69.2	0.028	-132.9	-5.59	-171.4	6.48	1.05
1.600	-0.48	135.2	2.288	-87.3	0.032	-148.9	-6.43	-177.6	7.19	1.02
1.700	-0.73	124.6	2.542	-106.8	0.039	-166.3	-7.84	178.8	8.10	1.01
1.800	-1.26	113.0	2.872	-128.6	0.047	174.0	-9.91	-177.3	9.16	1.01
1.900	-2.31	101.1	3.240	-153.5	0.056	151.3	-11.05	-156.6	10.21	1.01
2.000	-4.14	90.9	3.547	178.8	0.064	125.8	-8.31	-138.3	11.00	1.02
2.100	-6.75	87.5	3.682	149.7	0.070	98.9	-5.06	-141.0	11.32	1.04
2.200	-9.11	95.3	3.655	121.2	0.073	72.6	-3.15	-153.0	11.26	1.06
2.300	-10.01	108.4	3.587	93.9	0.075	47.6	-2.42	-167.3	11.09	1.07
2.304	-10.02	108.9	3.585	92.8	0.075	46.6	-2.41	-167.9	11.09	1.08
2.400	-10.08	120.9	3.583	66.5	0.078	22.4	-2.72	177.8	11.08	1.09
2.500	-9.15	137.4	3.638	36.0	0.082	-5.9	-4.55	162.9	11.22	1.11
3.000	-0.30	98.5	0.945	-114.5	0.025	-146.5	-1.26	-164.2	-0.49	1.17
4.000	-0.74	28.9	0.159	140.2	0.006	122.2	-0.11	150.2	-15.96	1.91
5.000	-2.01	-15.1	0.071	59.4	0.003	56.4	-0.08	123.6	-22.98	13.43
6.000	-0.78	1.8	0.044	120.4	0.003	125.4	-0.07	100.6	-27.20	9.68
7.000	-0.36	-60.0	0.158	68.8	0.014	86.8	-2.71	123.5	-16.02	8.12
8.000	-0.77	-109.6	0.086	-76.3	0.010	-50.3	-0.74	73.6	-21.26	14.53
9.000	-3.99	178.3	0.030	151.2	0.004	-174.8	-0.10	53.3	-30.56	50.93
10.000	-1.70	-138.8	0.016	5.4	0.003	45.4	-0.10	33.1	-36.06	75.19
11.000	-0.67	163.8	0.024	-91.2	0.006	-46.2	-0.47	12.8	-32.44	53.16
12.000	-0.95	116.5	0.031	106.0	0.010	152.0	-0.48	-0.6	-30.15	33.01

Table 3—*Microwave Harmonica* Analysis (Optimized for Output Power)

MICROWAVE HARMONICA PC V5.5 File: b:qexamp.ckt 22-DEC-93 18:04:28

Freq GHz	MS11 dB AMP	PS11 deg AMP	MS21 mag AMP	PS21 deg AMP	MS12 mag AMP	PS12 deg AMP	MS22 dB AMP	PS22 deg AMP	MS21 dB AMP	K AMP
1.000	-0.72	-168.6	2.386	39.7	0.020	-33.3	-17.28	-86.0	7.55	1.63
1.100	-0.67	-178.3	2.331	20.4	0.022	-50.9	-15.61	-117.3	7.35	1.42
1.200	-0.64	172.6	2.299	2.0	0.023	-67.5	-14.05	-142.9	7.23	1.27
1.300	-0.64	163.7	2.294	-15.7	0.026	-83.3	-12.69	-165.0	7.21	1.17
1.400	-0.66	155.0	2.314	-32.9	0.028	-98.6	-11.48	175.0	7.29	1.09
1.500	-0.71	146.5	2.365	-49.9	0.031	-113.6	-10.42	156.4	7.48	1.05
1.600	-0.79	137.8	2.453	-66.8	0.035	-128.4	-9.50	138.6	7.79	1.02
1.700	-0.91	128.9	2.588	-83.8	0.039	-143.3	-8.71	121.3	8.26	1.01
1.800	-1.10	119.4	2.786	-101.2	0.045	-158.5	-8.08	103.9	8.90	1.01
1.900	-1.42	108.8	3.071	-119.4	0.053	-174.6	-7.62	85.7	9.75	1.01
2.000	-1.98	96.3	3.482	-139.1	0.063	167.9	-7.46	65.7	10.84	1.02
2.100	-3.10	80.3	4.065	-161.7	0.078	147.5	-7.86	41.8	12.18	1.04
2.200	-5.77	57.1	4.815	170.7	0.096	122.2	-9.64	9.1	13.65	1.05
2.300	-14.58	10.6	5.418	135.8	0.113	89.5	-15.23	-52.9	14.68	1.07
2.304	-15.31	6.6	5.428	134.2	0.113	88.0	-15.55	-57.2	14.69	1.07
2.400	-10.67	-157.1	5.138	96.3	0.111	52.3	-13.32	174.5	14.22	1.09
2.500	-4.22	165.0	4.086	61.5	0.092	19.6	-7.41	119.7	12.23	1.11
3.000	-0.53	96.9	1.222	-35.2	0.032	-67.2	-2.33	21.6	1.74	1.17
4.000	-0.75	29.2	0.494	-155.6	0.019	-173.6	-1.06	-79.0	-6.13	1.69
5.000	-1.99	-15.1	0.261	93.6	0.013	90.6	-1.04	-169.9	-11.67	11.23
6.000	-0.78	1.9	0.137	130.1	0.010	135.1	-0.55	116.4	-17.28	7.89
7.000	-0.39	-60.0	0.101	160.7	0.009	178.7	-0.92	-3.1	-19.91	8.22
8.000	-0.75	-109.9	0.196	-1.0	0.023	25.0	-6.65	-66.0	-14.17	14.18
9.000	-4.01	178.4	0.126	-132.4	0.018	-98.4	-1.60	-147.6	-18.01	39.98
10.000	-1.70	-138.8	0.076	53.9	0.015	93.9	-2.18	130.1	-22.42	57.57
11.000	-0.66	163.8	0.071	-60.9	0.017	-15.9	-9.01	29.7	-22.92	50.28
12.000	-0.97	116.5	0.092	168.5	0.030	-145.5	-8.83	160.9	-20.71	31.18

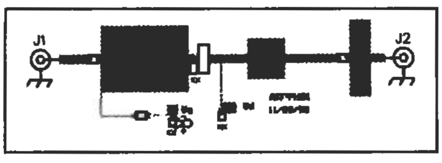

Fig 4—Parts placement diagram for the 2-W amplifier. Points labeled XX should be connected to the ground plane using copper foil.

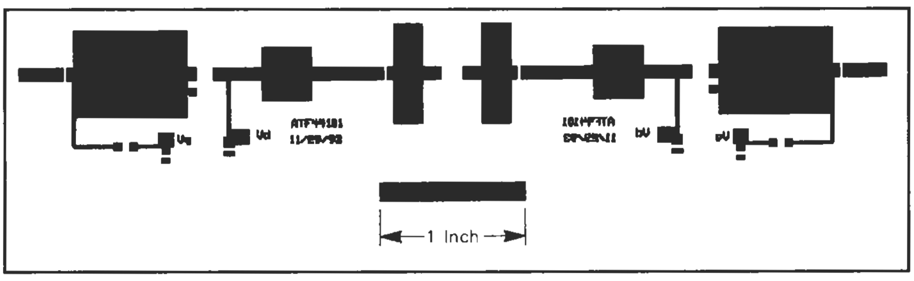

Fig 5—Etching pattern for the 2-W amplifier. Use 0.031-inch-thick, $\varepsilon_r = 2.55$ Teflon circuit board.

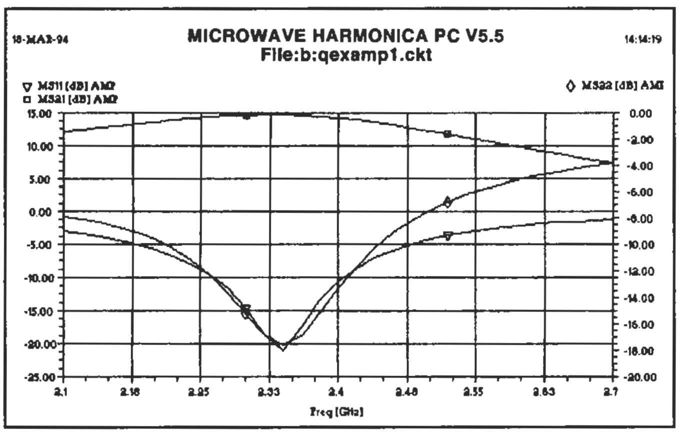

Fig 6—*Microwave Harmonica* analysis of the circuit's S parameters after computer optimization.

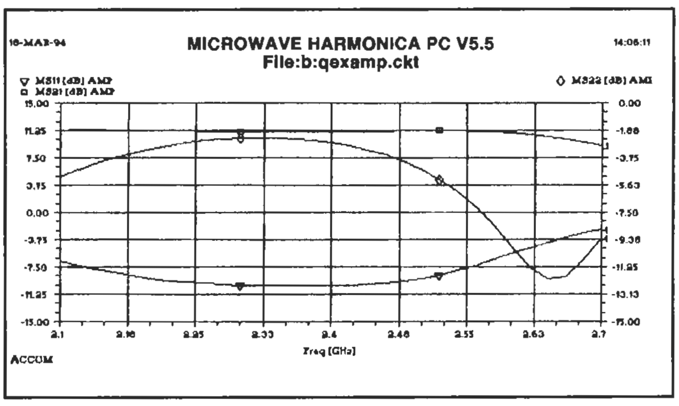

Fig 7—After the output network was optimized on the bench for maximum output power, *Microwave Harmonica* produced this analysis of the resulting circuit.

mount the FET. Ideally, clean holes would be punched into the copper, but I couldn't figure out a way to do that with the tools I have available. I used 0-80 screws, which are the largest screws I've found that will fit in the mounting holes. While I've used smaller screws in the past, I prefer using the largest screws possible since taps generally get tougher to use as they get smaller. I built one amplifier using 0.125-inch thick 6061 aluminum, but found that to be too thin to easily make the tapped holes for attaching the connectors and brass strips, although the finished mounting plate was stiff enough.

By Al Ward, WB5LUA **From *QST*, December 1985**

1296-MHz Solid-State Power Amplifiers

Explore the DX possibilities of the 23-cm band with these modern amplifiers.

With the ever-increasing number of ready-to-go 1296-MHz transverters available on the market today, there is a great demand for a simple and economical way to generate higher power than the typical 0.5 to 1 W output these transverters provide. If 1 W or less is used to drive a typical 2C39/7289 stripline or cavity amplifier, the low output power is often disappointing. A tube-type amplifier run with 1 kV on the plate typically offers a gain of 10 dB. This means that, at best, you can expect 5- to 10-W output from your 1-W or less input.

A popular way of generating higher power is to cascade two tube-type amplifiers for 50- to 100-W output. Here I will show you an alternative: two solidstate amplifiers that can replace the tubetype driver amplifier and provide 10- to 20-W output—enough to drive a two-tube amplifier to full output.

The NEC NEL1306 and the NEL1320 1300-MHz power transistors are an economical solid-state approach to generating moderate power levels (10-20W) at 1269 and 1296 MHz. These amplifiers can be used for terrestrial or sateffite work. When OSCAR 10 was designed, it was thought that 10 W into a modest gain antenna (20 dBi) would produce usable signals from the sateffite. Unfortunately, there were some problems, and the sensitivity of AO-10 was not as originally expected for the Mode-L uplink. These amplifiers can, however, be used as a driver for a higher-powered tube amplifier for Mode-L service. If all goes according to plan with the launch of AMSAT-OSCAR Phase IIIC, 10 to 20 W with a 20-dBi gain antenna will produce acceptable downlink signals.

The NEL1306 is rated for 6-W output at 1296 MHz at the 1-dB compression point; the NEL 1320 is rated at 20 W. These devices offer several advantages for amateur experimenters. They were designed for collector voltages of 12- to 13.6-V dc, making them ideal for portable and mobile operation. Although the price may seem high to someone familiar with HF parts, these devices are less expensive than most microwave power transistors. The NEL 1306 is in the $26 price range, while the NEL1320 costs about $42. California Eastern Laboratories makes these transistors available in single-lot quantities, so you don't have to be "in the business" to get your hands on them.[1]

The performance of the amplifiers I built and tested is shown in Table 1. The NEL1306 is a good buy. With 1.5-W input, 6- to 8-W output can be achieved. When the amplifier is tuned up at lower power levels, power gain can be as high as 10 dB. With 200-mW drive from my homemade transverter, an output power level of 2 W is attainable.[2]

Power gains as high as 17 dB are possible with a two-stage amplifier (an NEL1306 driving an NEL1320). With a mere 200 mW of drive, 10-W output is possible. When the pair of amplifiers is driven with 1 W and tuned for maximum power output, the 1-dB compression point of 18 W will be achieved.

Circuit Details

The basic design, shown schematically in Fig. 1, is an adaptation of a circuit described in the NEL1300 series data sheet. The design incorporates 30-ohm quarter-wavelength microstriplines on the input and output. C3, C4, C7 and C8, along with L1, form a pi network that matches the low input impedance of the device to 50 ohms. CS, C6, C9 and C10 and the 30-ohm transmission line (L2) form an output pi network that maximizes power transfer to 50 ohms. C10 is not always necessary, depending on variations among devices and circuit-board material.

I designed the amplifiers for 0.031-inch-thick, double-sided glass-epoxy circuit board. A 30-ohm line in this dielectric equates to a line width of 0.121 inch, which is equivalent to the width of the collector and base leads of the NEL1300 series

Table 1

Typical Operating Conditions for the 1296-MHz Solid-State Power Amplifiers

Device	*NE130681-12*	*NEL132081-12*
P_{out}(1-dB compression point)	7W	18W
Gain (1-dB compression point)	6 dB typ.	5 dB typ
Collector efficiency	40-50%	40-50%
Idling current	50 mA	150 mA
I_C @ 1-dB compression point	1.1 A	3.0 A
V_{CC}	13.5 V	13.5 V
Power input	14.9 W	40.5 W

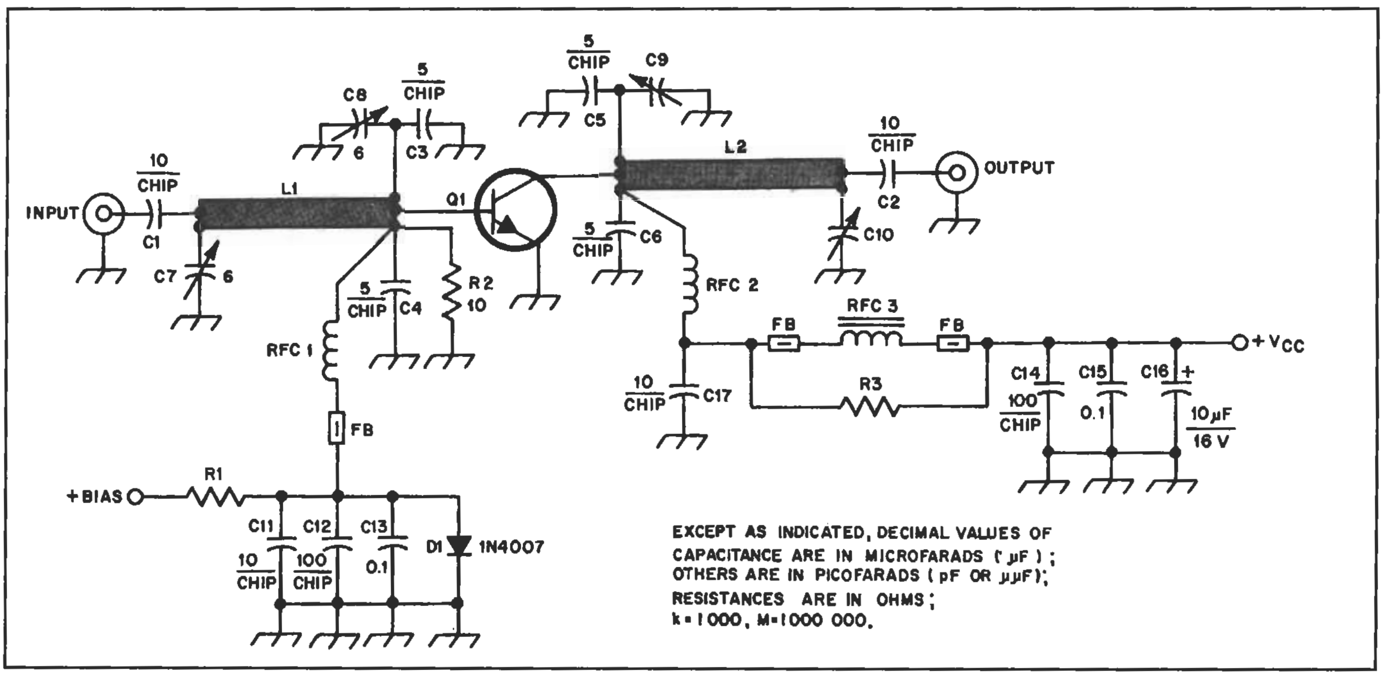
INPUT
10
CHIP
C1
C7
6
L1
C8
6
5
CHIP
C3
Q1
5
CHIP
C4
R2
10
RFC 1
FB
R1
+BIAS
C11
10
CHIP
C12
100
CHIP
C13
0.1
D1
1N4007
5
CHIP
C9
C5
L2
5
CHIP
C6
RFC 2
10
CHIP
C2
OUTPUT
C10
FB
RFC 3
FB
R3
10
CHIP
C17
C14
100
CHIP
C15
0.1
C16
10µF
16 V
+Vcc

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k=1000, M=1000 000.

Fig. 1—Schematic diagram of the NEL1306 and NEL1320 1296-MHz solid-state power amplifiers. The schematic is identical for both versions. Component values are the same except as noted.

C1, C2, C11, C17—10-pF chip capacitor.
C3, C4, C5, C6—3.6- to 5.0-pF chip capacitor.
C7, C8—1.8- to 6.0-pF miniature trimmer capacitor (Mouser 24AA070 or equiv. See text).
C9, C10—Same as C7 and C8 for the NEL1306 amplifier. For the NEL1320 version, 0.8- to 10-pF piston trimmers are used (Johanson 5200 series or equiv.).
C12, C14—100-pF chip capacitor.
C13, C15—0.1-µF disc ceramic capacitor.
C16—10-µF electrolytic capacitor.
D1—1N4007 diode.
L1, L2—30-ohm microstripline, ¼-wavelength long (see text).
Q1—NEC NEL130681-12 (6 W) or NEL132081-12 (18 W) *transistor.*
R1—82- to 100-Ω resistor, 2-W minimum. Vary for specified idling current.
R2—10-Ω, ¼-W carbon-composition resistor with "zero" lead length. See text.
R3—15-Ω, 1-W carbon-composition resistor.
RFC1—3t no. 24 wire, 0.125 inch ID, spaced 1 wire diam.
RFC2—1t no. 24 wire, 0.125 inch ID, spaced 1 wire diam.
RFC3—1-µH RF choke; 18t no. 24 enam. close-spaced on a T50-10 toroid core.

devices. This minimizes the discontinuity between L1, L2 and Q1.

Bias is provided by R1, R2 and D1. R1 can be optimized, if desired, to adjust the collector idling current.

I selected RFC1 and RFC2 by choosing the lowest possible reactance that will not affect power gain or output power. The RF chokes and the 10-pF bypass capacitors afford adequate decoupling at the frequency of operation. The values of RFC1 and RFC2 are purposely made different to avoid oscillations caused by bias-choke coupling.

After building several of these amplifiers, I noticed that the transistors sometimes generated low-frequency spurious signals. Although these signals were very low in amplitude and caused no problems, they were annoying. I found that I could eliminate them by keeping the high-frequency RF chokes in the collector circuit as small as possible and adding the parallel R3/RFC3 combination, as well as bypass capacitors C14, C15, C16. RFC3 must be capable of handling 1 A for the NEL1306 and 3 A for the NEL1320, so I made special RF chokes to withstand the current.

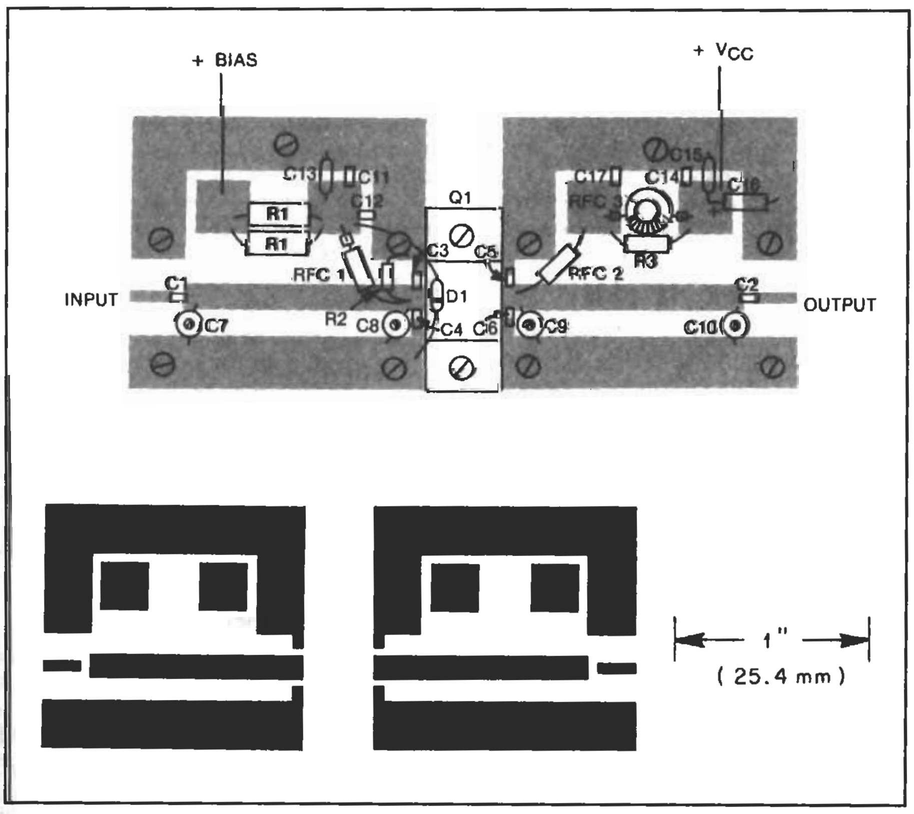
+ BIAS
+ Vcc
Q1
C13
C11
C12
R1
R1
RFC 1
R2
C8
C3
C5
D1
C4
C6
C9
C15
C17
C14
C16
RFC 2
R3
INPUT
C1
C7
C2
C10
OUTPUT
1"
(25.4 mm)

Fig 2—Parts-placement diagram and full-size etching pattern for the solid-state 1296-MHz power amplifiers. All components mount on the etched side of the board. The same PC boards are used for each version.

Construction

Identical construction techniques are used for both amplifiers. Most of the components are mounted to the PC board, and the board and transistor are mounted to an aluminum base plate made from 1/4-inch-thick stock. PC-board layout is shown in Fig. 2. Two separate boards are used—one for the input side and one for the output—and they are mirror images of each other. The copper is retained on the bottom side and serves as a ground plane. The grounded areas on the top side must have a good connection to the bottom ground plane for low-inductance grounding of the transistor emitter leads, matching capacitors and bias circuitry. Etched PC boards and partial parts kits for this project are available from A & A Engineering.[3]

Several effective methods of connecting the top and bottom ground planes are summarized here.

1) Plated-through holes at the critical grounding areas mentioned earlier.

2) Use of pins or screws that penetrate through the circuit board into the aluminum base plate at the critical areas.

3) Use of "wrap-around" foils on all edges of the ground plane.

Plated-through holes are often used in the commercial and military electronic marketplace but are not so easily reproduced in the average builder's circuitboard shop. I've found that the best technique for the home builder is a combination of methods 2 and 3. First, wrap thin copper or brass foil around the board edges and then solder the foil to the top and bottom. Sometimes called "shim stock," thin sheet metal is often available from hobby shops or metal suppliers. Next, drill holes through the board and use no. 4-40 screws to tie the circuit board to the base plate at the critical areas shown in Fig. 2.

The transistor must be mounted so that its leads lie flat against the PC board. This poses a slight problem, since the transistor leads protrude from the device 0.165 inch above the bottom of the flange, and the PC boards are only 0.031 inch high. There are a couple of ways around this problem. They work equally well, so choose the method that is easiest for you.

If you have access to a machine shop, you can mill out a 0.240-inch-wide by 0.090-inch-deep slot in the aluminum base plate to clear the transistor flange. This is a simple task on a milling machine, and you can probably find a local machine shop that will do small jobs when business is slow. The other method is to mount the transistor to the base plate and use a piece of 0.090-inch-thick aluminum sheet to bring the input and output PC boards up to the right height. The 0.090-inch dimension allows a wrap-around foil and solder buildup of 0.040 inch, maximum.

I used 1-inch-long screws to secure the transistor to holes drilled and tapped in the aluminum base plate. The extra screw length that protrudes from the bottom of the base plate allows the use of an external heat sink, which is suggested if continuous operation is desired.

The clearance holes in the NEL1300 devices are for no. 4-40 hardware. I drilled out the holes to accept no. 6-32 screws to make the assembly more rugged, although this may not be necessary. Be careful if you decide to drill out the holes to accept no. 6-32 hardware; the transistor flange is soft copper, and you could damage the device. Use a small vise to hold the transistor flange during the drilling operation.

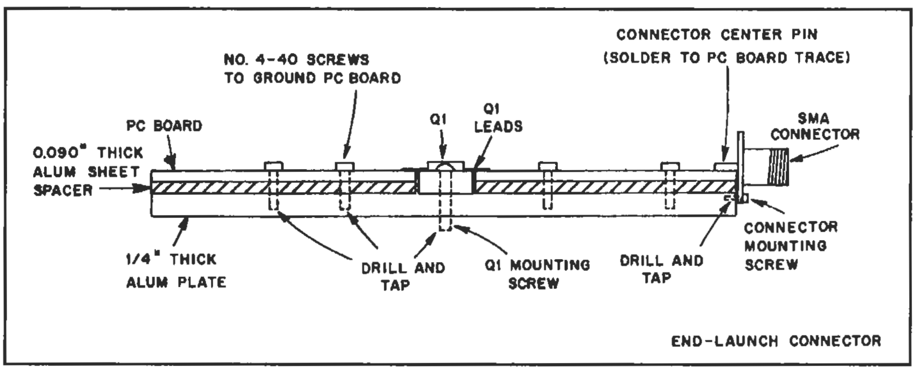

Fig 3—Construction details for the solid-state 1296-MHz power amplifiers. See text for additional information.

Solder the components to pads on the board using surface-mounting techniques. Silver solder (2%) is recommended for the chip capacitors, but 5N63 will work fine. Use a 15-W iron and solder quickly to avoid burning the metallization off the capacitors.

The transistor leads should be soldered in place only after the circuit boards and transistor have been firmly bolted down to the base plate. This is necessary to minimize any buildup of stress in the transistor leads. Chip capacitors C3, C4, C5 and C6 should be soldered directly onto the leads of Q1 to ensure the shortest possible lead length. Keep D1 close to Q1. Thermal compound will enhance heat transfer to D1 to ensure minimal drift in idling current with temperature changes.

In the original design of these amplifiers, Johanson piston-trimmer capacitors (5200 series or equiv.) were used for C7, C8, C9 and C10. These capacitors are fairly large, and in some instances the coupling between the bodies of C8 and C9 was enough to cause an in-band oscillation. Smaller variable capacitors should be used if at all possible. I've used a 1.8- to 6.0-pF miniature ceramic trimmer capacitor, Mouser Electronics part number 24AA070, in the input and output networks of the NEL1306 amplifier and in the input network of the NEL1320 amplifier with no performance degradation.[4] I did use the Johanson piston trimmers in the output network of the NEL1320 amplifier because of the high RF currents involved.

The PC board makes use of end-launch SMA-type connectors. Fig. 3 illustrates this technique. Four-hole, flange-mount SMA connectors can be mounted to the edge of the base plate using two of the four mounting holes. Drill and tap the base plate for no. 2-56 hardware. Be careful—it's easy to accidentally cross-thread or overtorque the no. 2-56 hardware.

An alternative approach is to mount the amplifier in an aluminum die-cast box (Bud CU-124B or Hammond 1590B) and run miniature 50-ohm coaxial cable such as RG-174 from the amplifier board to the connector. The amplifiers shown in Figs. 4 and 5 use standard SMA connectors mounted to the walls of the metal box. BNC or Type-N connectors should work equally as well. When preparing each end of the coaxial cable, try to keep the pigtail leads as short as possible (1/8 inch or less); otherwise the mismatch will be difficult to tune out.

I compared the performance of an amplifier with end-launch connectors to that of another that used the approach just described. I could measure no difference in gain or 1-dB compression point.

Considerable effort was put forth to make sure the amplifiers are stable. The devices have fairly high gain at the frequency of operation, so layout and good construction practices are very important. Here are some construction hints that can help ensure amplifier performance and stability.

1) Use the smallest (physical) size variable capacitors that will still handle the RF current.

2) Use wrap-around ground foils as noted. Grounding screws are required at the critical RF-ground areas near the shunt variable capacitors, shunt bypass capacitors and Q1 emitters.

3) Connect braids from the coaxial-cable jumpers to the same ground as the shunt variable capacitors.

4) Use as little lead length as possible on R2—less than 1/8 inch.

5) In some instances when the large piston trimmers are used, a shield approximately 3/4 inch high mounted on top of Q1 and grounded via the mounting screws can improve isolation between C8 and C9.

Tune-up and Operation

Measuring RF power at 1296 MHz can be difficult. I used a calibrated 20-dB directional coupler along with enough attenuator pads to allow power to be read with a Hewlett Packard HP430C power meter and an HP477B thermistor mount. Even better is the HP431 power meter with its associated HP478A thermistor mount (a newer version of the HP430C that does not

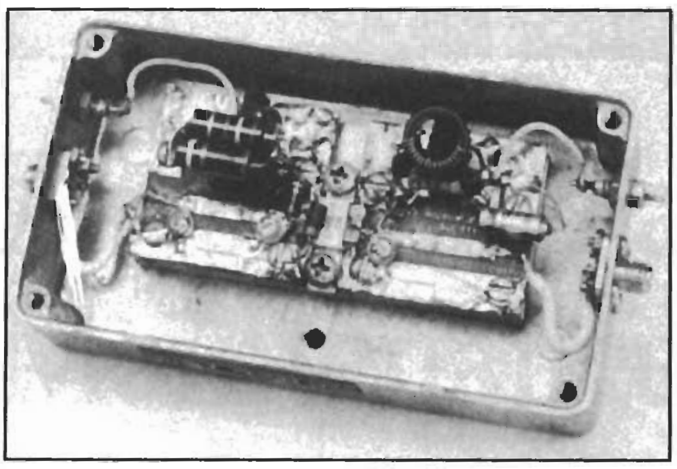

Fig 4—This NEL1306 amplifier was built inside a die-cast box. Miniature coaxial cable runs to the connectors.

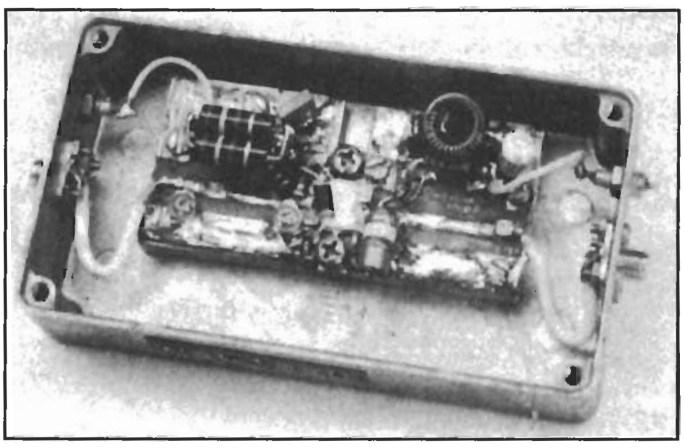

Fig 5—The NEL1320 amplifier is virtually identical to the NEL1306 version of Fig 4, except for the capacitors on the output strip line. C10 was not necessary on this version.

suffer from temperature-drift problems). Bird offers several low-power elements for the popular model 43 in this frequency range. Bird 400-1000 MHz elements are common, and they can be used with decreased accuracy.

Begin initial setup of each amplifier by terminating the input and output in good 50-ohm loads. I recommend that you use a fuse in the collector lead of Q1 to protect the device until you are sure everything is working normally. Start with all capacitors at minimum. Apply 12- to 13.5-V dc to the V_{CC} and bias terminals. The collector idling current should be as shown in Table 1. Vary the value of R1 for correct idling current.

For the NEL1306 amplifier, start out with 50 to 100 mW of drive. Adjust the output network for maximum power output and then peak the input network for maximum output power. Increase drive and repeak both matching networks for rated performance as shown in Table 1. Similarly, start out with approximately 1 W of drive for the NEL1320 and follow the same procedure. After a minute or two of operation at maximum power output, remove RF drive power and check to see that the collector idling current has not increased more than 25% over the initial setting. Keeping D1 in close contact with Q1 will minimize drift in idling current with temperature changes.

If you're going to use transistor switching to apply dc to the power amplifier stages during transmit, consider the following technique. Apply 13.5-V dc to the V_{CC} terminal during receive and transmit. Use a series transistor switch to apply 13.5-V dc to the bias terminal during transmit. A power transistor capable of carrying only a few hundred milliamperes of bias current, as opposed to several amperes of collector current, will be required. More important, the voltage drop across the transistor switch in the V_{CC} line will be eliminated. This will ensure maximum power output of the NEL1300 devices by keeping V_{CC} at 13.5-V dc.

Switching the bias port off during receive is important for another reason. Normally during receive periods, the amplifier is left unterminated. Sometimes the input port is also left open. Depending on the length of the unterminated 50-ohm cable on the amplifier ports, the unit may show signs of instability if it is drawing idling current.

Several amateurs in the Dallas area have duplicated these amplifiers with no problems. Other amplifiers of this design are in use in different parts of the country. The NEL1300 series amplifiers offer a simple and inexpensive means of generating medium power on 1296 MHz. You'll be amazed at what you can work with 18 W that you can't with 1 W.

I wish to thank everyone who offered technical advice, especially Wes Atchison, WA5TKU, for helping with the construction and evaluation of the prototype amplifiers.

Notes

[1] NEC transistors are available from California Eastern Laboratories, 3260 Jay St., Santa Clara, CA 95050, tel. 408-988-3500.

[2] Complete construction details for this transverter may be found in Chapter 32 of the *1986 ARRL Handbook*.

[3] A & A Engineering, 7970 Orchid Dr., Buena Park, CA 90620, tel. 714-521-4160.

[4] Mouser Electronics, 11433 Woodside Ave., Santee, CA 92071, tel. 619-449-2222.

By Ward Silver, N0AX From *QST*, September 2003

Amplifier Care and Maintenance

With few moving parts, your amplifier can be easy to overlook. Here are some ideas for taking care of amplifiers.

Amplifiers come in all shapes and sizes, large and small, light or heavy, tube or solid-state, VHF or HF. Regardless, they all need a little TLC from time to time. They can cost as much as top-of-the-line radios, so it's important that they get a little maintenance on a regular basis.

While this article focuses on amplifiers that use vacuum tubes, many of the ideas presented here can and should be applied to any amateur amplifier—HF or VHF/UHF. Solid-state amplifiers operate at lower voltages and generally have fewer points of failure, but they still need occasional maintenance.

Safety First

It is important to review good safety practices.[1] Tube amplifiers use power supply voltages well in excess of 1 kV and the RF output at full throttle can be hundreds of volts, as well. Almost every voltage in an amplifier can be lethal! Take care of yourself and use caution!

- *Power Control*—Know and control the state of both ac line voltage and dc power supplies. Physically disconnect line cords and other power cables when you are not working on live equipment. Use a lockout on circuit breakers. Double-check visually and with a meter to be absolutely sure power has been removed.
- *Interlocks*—Unless specifically instructed by the manufacturer's procedures to do so, never bypass or "rig" an interlock. This is rarely required except in troubleshooting and should only be done when absolutely necessary. Interlocks are there to protect you.
- *The One-Hand Rule*—Keep one hand in your pocket while making any measurements on live equipment. The hand in your pocket won't give current a chance to flow through you. It's also a good idea to wear shoes with insulating soles and work on dry surfaces. Current can be lethal even at milliampere levels—don't tempt the laws of physics.
- *Patience*—Repairing an amplifier isn't a race. Take your time. Don't work on equipment when you're tired or frustrated. Wait several minutes after turning the amplifier off to open the cabinet—capacitors can take several minutes to discharge through their bleeder resistors.
- *A Chicken Stick*—Make this simple safety accessory shown in Figure 1 and use it whenever you work on equipment in which hazardous voltages have been present. The ground wire should be heavy duty (12 gauge minimum) due to the high peak currents (hundreds of amperes) present when discharging a capacitor or tripping a circuit breaker. When equipment is opened, touch the tip of the stick to *every* exposed component and connection that you might come in contact with. Assume nothing—accidental shorts and component failures can put voltage in places it shouldn't be.
- *The Buddy System and CPR*—It's always a good idea to use the buddy system when working around any equipment that has the potential for causing serious injury. The buddy needn't be a ham, just anyone who can be nearby in case of trouble. Your buddy should know how to remove power and administer basic first aid. Since hams work around electrical equipment frequently, it would be a good idea to have your buddy or someone in the household know CPR, as well.[2]

Cleanliness

The first rule of taking good care of an amplifier is cleanliness. I realize that 90 percent of ham shacks have just failed the first rule. Amplifiers need not be kept sparkling new, but their worst enemy is heat. Excess heat accelerates component aging and stresses those expensive tubes and transformers. There are two areas to keep clean—the inside and the outside.

Outside the amplifier, you need to prevent dust and obstructions from blocking the paths by which heat is removed. This means keeping all ventilation holes free of the ever-present dust bunnies, pet hair and

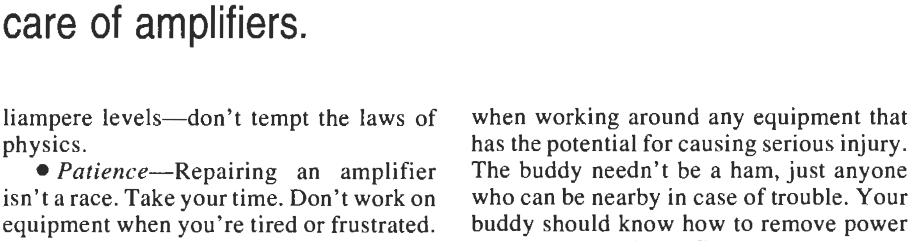
# 10 Eye - Bolt
Solder Lug and Nut
3/4" PVC Pipe, 12 inches Long
3/4" PVC Pipe Cap
2 - 3 Feet of Heavy, Flexible Wire (12 Gauge Minimum)
Heavy - Duty Alligator - Type Clip

Figure 1—The "chicken stick" is a great way to ensure that everything inside the amplifier that should be discharged actually is. It can be a life saver.

insects. Fan intakes are particularly susceptible to inhaling all sorts of "goop." Get out the vacuum cleaner and clean not only the amplifier, but the surrounding areas. Don't even think about letting liquids anywhere near the amplifier. One spilled cup of coffee can cause hundreds of dollars of damage.

Keep papers or magazines off the amplifier—even if the cover is solid metal. Paper acts as an insulator and keeps heat from being radiated through the cover. Amplifier heat sinks must have free air circulation to be effective. There should be at least a couple of inches of free space surrounding an amplifier on its sides and top. If the manufacturer recommends a certain clearance, mounting orientation or air flow, follow those recommendations.

Just as the outside needs to be kept clean, so does the inside. High voltage (HV) circuits attract dust like crazy. The dust slows heat dissipation and will eventually build up to the point where it arcs or carbonizes. Our friend the vacuum cleaner should make another appearance to remove any dust or dirt. If you find insects (or worse) inside the amp, try to determine how they got in and plug that hole. Window screening works fine to allow airflow while keeping out visitors. While you're cleaning the inside, this is a great time to perform a visual inspection as described in the next section.

Vacuuming works best with an attachment commonly known as a "crevice cleaner." Figure 2 shows a crevice cleaning attachment being used with a small paintbrush to dislodge and remove dust. Don't use the vacuum cleaner brush attachment; they're designed for floors, not electronics. Some vacuums also have a blower mechanism, but these rarely have enough punch to clean as thoroughly as a brush. Besides, that dust you blow all over is going to wind up in some other equipment, so it's best to take it out of circulation, so to speak. The brush will root dust out of tight places and off components without damaging them or pulling on connecting wires.

If you can't get a brush or attachment close enough, a spray can of compressed air will usually dislodge dust and dirt. If you use a rag or towel to wipe down panels or large components, be sure not to leave threads or lint behind. Never use a solvent or spray cleaner to wash down components unless the manufacturer advises doing so—you might leave behind a residue or damage the component.

Visual Inspection

Once the amplifier has been cleaned, it's time for a visual inspection. Remove any internal covers or access panels and...*stop!* Get out the chicken stick, clip its ground lead securely to the chassis and touch every exposed connection. Now, using a strong light and possibly a magnifier, look over the components and connections. Amplifiers have far fewer components than transceivers do, so it's quite feasible to look at every component and insulator. Look for cracks, signs of arcing, carbon traces (thin black lines), discoloration, loose connections, melting of plastic, and anything else that doesn't "look right." This is a great time to be sure that mounting and grounding screws are tight. While you're in there, does anything smell burnt? The nose can quickly detect the odors of toasted transformer, cooked capacitor or roasted resistor. Learn the smells of healthy and not-so-healthy components.

Make a note of what you find, repair, or replace—even if it's absolutely nothing. If you don't keep a shack notebook, start one. A simple spiral notebook with notes about maintenance, wiring, color coding, antenna behavior and so forth can be a big timesaver.

S. COHEN, N1SC; R. SCHETGEN, KU7G

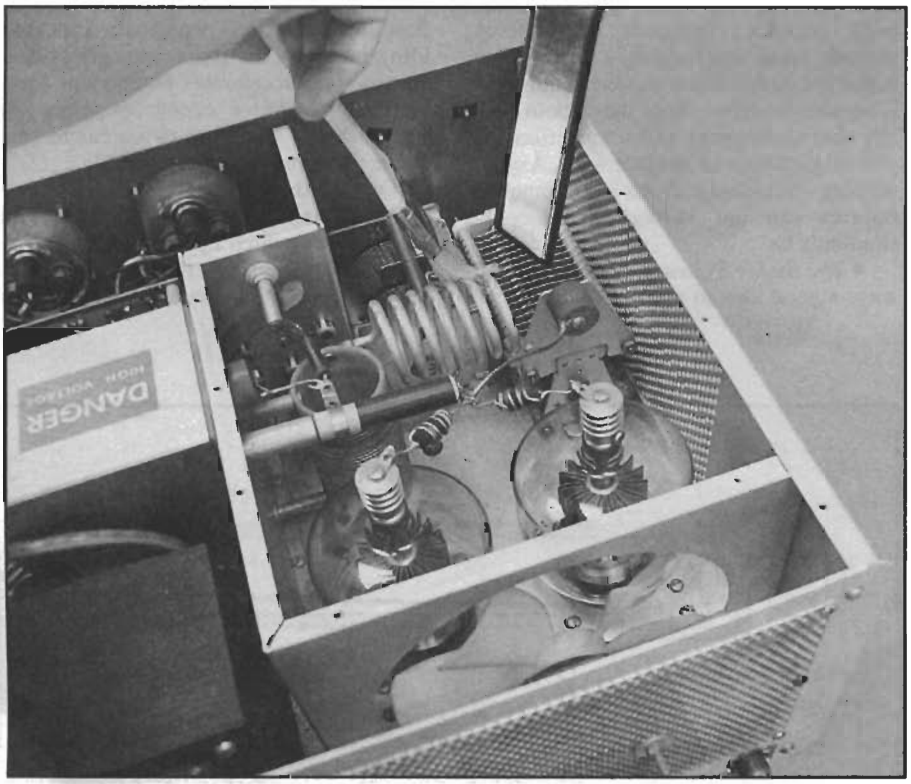

Figure 2—A small paintbrush and a vacuum cleaner crevice attachment make dust removal easy.

Electrical Components

An amplifier contains many heavy-duty HV and RF components. These can be expensive and hard to replace, so it's important that you take good care of them. Let's start with the power supply.

There are three basic parts to amplifier power supplies—the ac transformer and line devices, the rectifier/filter and the metering/regulation circuitry. Transformers need little maintenance except to be kept cool and be mounted securely. Line components such as switches, circuit breakers and fuses, if mechanically sound and adequately rated, are usually electrically okay, as well.

Rectifiers and the capacitors that filter the HV dc require occasional cleaning. Look for discoloration around components mounted on a printed circuit board (PCB) and make sure that all wire connections are secure. HV capacitors are generally electrolytic or oil and should show no signs of leakage, swelling or outgassing around terminals.

Located at the output of the filter, components that perform metering and regulation of voltage and current can be affected by heat or heavy dust. If there has been a failure of some other component in the amplifier—such as a tube—these circuits can be stressed severely. Resistors may survive substantial temporary overloads, but may show signs of overload, such as discoloration or swelling.

Amplifiers contain two types of relays—control and RF. Control relays switch ac and dc voltages and do not handle input or output RF energy. The usual problem encountered with control relays is oxidation or pitting of their contacts. A burnishing tool can be used to clean relay contacts. In a pinch a strip of ordinary paper can be pulled between contacts gently held closed. [Avoid the temptation to over-clean silver-plated relay and switch contacts, as the author points out later. It is easy to remove contact plating with excessive polishing and while silver-plated relay and switch contacts may appear to be dark in color, oxidized silver (black) is still a good conductor. Once the silver's gone, it's gone; contact erosion will then be pervasive.—*Ed.*] If visual inspection shows heavy pitting or discoloration or resistance measurements show the relay to have intermittent contact quality, it is best replaced.

RF relays are used to perform transmit-receive (TR) switching and routing of RF signals through or around the amplifier circuitry. Amplifiers designed for full break-in operation will usually use a high-speed vacuum TR relay. Vacuum relays are sealed and cannot be cleaned or maintained. When you replace RF relays, use a direct replacement part or one rated for RF service with the same characteristics as the original.

Cables and connectors are subjected to heavy heat and electrical loads in amplifiers. Plastics may become brittle and connections may oxidize. Cables should remain flexible and not be crimped or pinched under clamps or tie-downs. It's a good idea to gently wiggle cables while watching the connections at each end for looseness or bending. Connectors can be unplugged and reseated once or twice to clear oxide on contact surfaces. Carefully inspect any connector that seems loose. Be especially careful with connectors and cables in amplifiers that have RF decks that are in separate enclosures from those of their power supplies. Those interconnects are susceptible to both mechanical and electrical stress and you don't want an energized HV cable loose on the operating desk. Check both the soldered electrical integrity and the mechanical stability of those cables and make sure they are tightly fastened.

As with relays, switches found in amplifiers are either control function orientated or RF routers. Adequately rated control switches, if mechanically sound, are usually okay. Bandswitches are the most common RF switch—usually a rotary phenolic or ceramic type. A close visual inspection should show no pitting or oxidation on the wiper (the part of the switch that rotates between contacts) or the individual contacts. Arcing or overheating will quickly destroy rotary switches. Figure 3 is a photo of a heavy-duty band switch that has suffered severe damage from arcing. Slight oxidation is acceptable on silver-plated switches. Phosphor-bronze contacts can sometimes be cleaned with a light scrub from a pink pencil eraser, but plating can be easily removed, so use caution with this method and be sure to remove any eraser crumbs. Rotary switch contacts cannot be replaced easily although individual wafer sections may be replaced if an exact matching part can be obtained.

Amplifiers use all types of capacitors and resistors. When replacing them, be sure to use a part rated for the use to which it will be put. Voltage and power-handling ratings are particularly important, especially of those handling high RF currents. An RF tank capacitor replacement should be checked carefully for adequate RF voltage and current ratings, not just dc. HV resistors are generally long and thin to prevent arcing across their surfaces. Even if a smaller (and cheaper) resistor has an equivalent power rating, resist the temptation to substitute it. In a pinch, a series string of resistors of the appropriate combined value can be used to replace one HV unit. Don't use carbon resistors for metering circuits, use metal or carbon film types. The carbon composition types are too unstable.

If you are repairing or maintaining an old amplifier and manufacturer-specific parts are no longer available, the ham community has many sources for RF and HV components. Fair Radio Sales and Surplus Sales of Nebraska are familiar names.[3] Hamfests and Web sites such as **www.eham.net** or **www.k1dwu.net/hamtrader** often have amplifier components for sale. You might consider buying another amplifier of the same type in non-working condition for parts use.

Tubes

The single most expensive component in an amplifier is usually the vacuum tube that performs the amplification. Good maintenance of tubes starts with proper operation of the amplifier. Follow the manufacturer's instructions for input drive levels, duty cycles, tuning and output power level. Frequently check all metered voltages and current to be sure that the tubes are being operated properly and giving you maximum lifetime. Penta Labs has an excellent Web page on maintaining power tubes.[4]

The internal mechanical structures of tubes generally do not deal well with mechanical shock and vibration, so be gentle. The manufacturer may also specify how the amplifier is to be mounted, so read the operating manual.

Tubes generate a lot of heat, so it's important that whatever cooling mechanism employed is kept at peak efficiency. Airways should be clean, including between the fins on metal tubes. All seals and chimneys should fit securely and be kept clean. Wipe the envelope of glass tubes clean after handling them—fingerprints should be removed to prevent baking them into the surface.

On metal tubes that use finger-stock contacts, be sure the contacts are clean and make good contact all the way around the tube. Partial contact or dirty finger stock can cause asymmetric current and heating inside the tube, resulting in warping of internal grids and possibly causing harmonics or parasitics.

Plate cap connections and VHF parasitic suppressors should be secure and show no signs of heating. Overheated parasitic suppressors may indicate that the neutralization circuit is not adjusted properly. Inspect socket contacts and the tube pins to be sure all connections are secure, particularly high-current filament connections. Removing and inserting the tubes once or twice will clean the socket contacts.

Adjustments to the neutralizing network, which suppresses VHF oscillations by negative feedback from the plate to grid circuit, are rarely required except when you are replacing a tube or after you do major rewiring or repair of the RF components. The manufacturer will provide instructions on making these adjustments. If symptoms of VHF oscillations occur without changing a tube, then perhaps the tube characteristics or associated components have changed. Parasitic oscillations in high-

H. WARD SILVER, N0AX

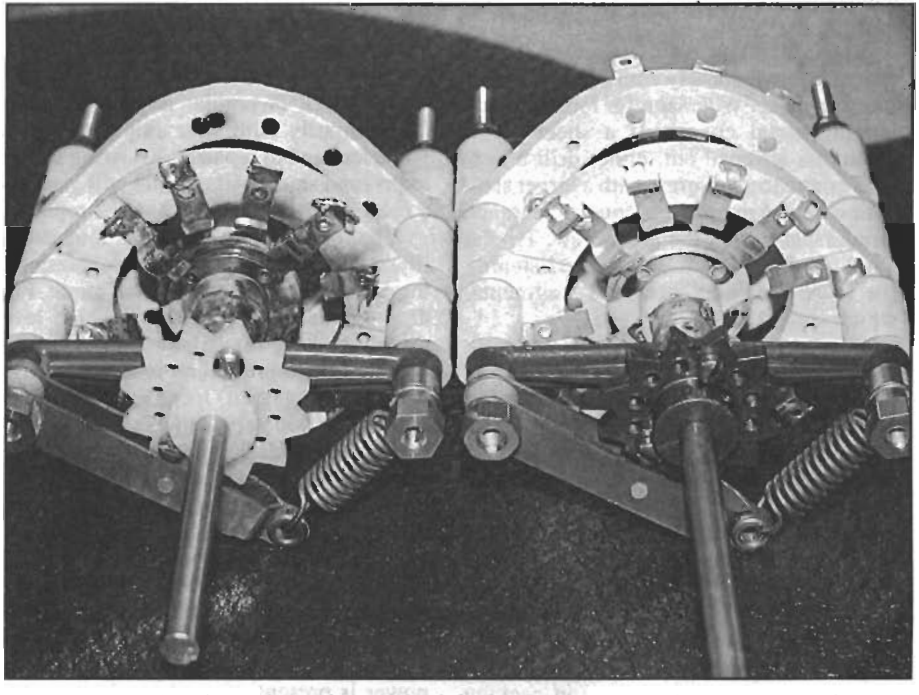

Figure 3—The band switch section on the left clearly shows the signs of destructive arcing.

power amplifiers can be strong enough to cause arcing damage. Perform a visual inspection prior to re-adjusting the neutralizing circuit.

Metering circuits rarely fail, but they play a key part in maintenance. By keeping a record of "normal" voltages and currents, you will have a valuable set of clues when things go wrong. This is perfect information for the shack notebook. Record tuning settings, drive levels, and tube voltages and currents on each band and with every antenna. When things change, you can refer back to the notebook instead of relying on memory.

Mechanical

While the amplifier is primarily an electronic beast, it has a significant number of mechanical parts that affect its well-being. Thermal cycling and heat-related stresses can result in mechanical connections loosening over time or material failures.

Switch shafts, shaft couplings and panel bearings all need to be checked for tightness and proper alignment. All mounting hardware needs to be tight, particularly if it supplies a grounding path. Examine all panel-mounted components, particularly RF connectors, and be sure they're attached securely. BNC and UHF connectors that are mounted with a single nut in a round panel hole are notorious for loosening with repeated connect/disconnections.

Rubber and plastic parts are particularly stressed by heat. If there are any belts, gears or pulleys, make sure they're clean and that dust and lint are kept out of their lubricant. Loose or slipping belts should be replaced. Check O-rings, grommets and sleeves to be sure they are not brittle or cracked. If insulation sleeves or sheets are used, check to be sure they are covering what they're supposed to. Never discard them or replace them with improperly sized or rated materials.

Enclosures and internal shields should all be fastened securely with every required screw in place. Watch out for loosely overlapping metal covers. If a sheet metal screw has stripped out, either drill a new hole or replace the screw with a larger size, taking care to maintain adequate clearance around and behind the new screw. Tip the amplifier from side to side while listening for loose hardware or metal fragments, which should all be retrieved.

A great time to clean the front and back panels and get gummy finger deposits off before they cause permanent finish damage is during maintenance. If the amplifier is missing a foot on the cabinet or an internal shock mount, replace it. A clean unit with a complete cabinet will have a significantly higher resale value than a dirty, grubby one, so it's in your interest to keep the equipment looking its best.

Shipping

When you are traveling with an amplifier or shipping it, some care in packing will prevent needless damage. Improper packing can also result in difficulty in collecting on an insurance claim, should damage occur. The original shipping cartons are a good method of protecting the amplifier for storage and sale, but they were not made to hold up to frequent shipping. If you travel frequently, it is best to get a sturdy shipping case made for electronic equipment.[5]

Some amplifiers require the power transformer to be removed before shipping. Check your owner's manual or contact the manufacturer to find out. Failure to remove it before shipping can cause major structural damage to the amplifier's chassis and case.

Tubes should also be removed from their sockets for shipment. It may not be necessary to ship them separately if they can be packed in the amplifier's enclosure with adequate plastic foam packing material. If the manufacturer of the tube or amplifier recommends separate shipment, however, do it!

Cleaning and Maintenance Plan

This discussion should have given you plenty to think about. It's easy to defer maintenance, but as with a vehicle, performance and lifetime are improved if a regular program is put into place. For amateur use, there is little need for maintenance more frequently than once per year. If there is a period of the year in which you are most active, put a note on the calendar about six weeks in advance to "open the hood," giving you time to obtain and replace any components.

Consider the maintenance requirements of your amplifier and what its manufacturer recommends. Sit down with your amplifier's manuals and make up a checklist of what major steps and tools are required. When maintenance time rolls around, you'll be prepared and be able to perform the job in the most efficient manner.

Troubleshooting

A benefit of regular maintenance will be familiarity with your amplifier should you ever need to repair it. Knowing what it looks (and smells) like inside will give you a head start on effecting a quick repair.

The following discussion is intended to illustrate the general flow of a troubleshooting effort, not be a step-by-step guide. Figure 4 shows a moderately-high-level troubleshooting flow chart. Before starting on your own amplifier, review the amplifier manual's "Theory of Operation" section and familiarize yourself with the schematic. If there is a troubleshooting procedure in the manual, follow it, of course.

You might be surprised how many "amplifier is dead" problems turn out to be simply a lack of ac power. Before even opening the cabinet of an unresponsive amplifier, be sure that ac is really present at the wall socket and that the fuse or circuit breaker is really closed. Assuming that ac power is present, trace through any internal fuses, interlocks and relays all the way through to the transformer primary terminals.

Hard failures in a high voltage power supply are rarely subtle, so it's usually clear if there is a problem and what components are involved. When you repair a power supply, take the opportunity to check all related components. If all defective components are not replaced, the failures may be repeated when the circuit is re-energized.

Rectifiers may fail open or shorted—test them using a DVM diode checker. An open rectifier will result in a drop in the HV output of 50 percent or more but will probably not overheat or destroy itself. A shorted rectifier failure is usually more dramatic and may cause additional rectifiers or filter capacitors to fail. If one rectifier in a string has failed, it may be a good idea to replace the entire string as the remaining rectifiers have been subjected to a higher-than-normal voltage.

High voltage filter capacitors usually fail shorted, although they will occasionally lose capacitance and show a rise in ESR (equivalent series resistance). Check the rectifiers and any metering components—they may have been damaged by the current surge caused by a shorted filter. Power transformer failures usually manifest themselves by insulation failures with consequent arcing of the windings. Either can result in the unmistakable aroma of overheated transformer. A failed transformer is generally not repairable.

Along with the HV plate supply, tetrode screen supplies occasionally fail, too. The usual cause is the regulation circuit that drops the voltage from the plate level. Operating without a screen supply can be damaging to a tube, so be sure to check the tube carefully after repairs.

If the power supply checks out okay and the tube's filaments are lit, check the resting or bias current. If it is excessive or very low, check all bias voltages and dc current paths to the tube, such as the plate choke, screen supply (for tetrodes) and grid or cathode circuits.

Having exhausted power supply and dc problems, you will then turn to the RF components or "RF deck." There is a natural tendency to forge ahead and swap in a known-good tube or tubes. Don't! Tubes are expensive and if the problem is elsewhere, you may damage the spares. Wait to swap tubes until you are sure that the tube is likely to be defective.

Check the input SWR to the amplifier. If it has changed (you did write down the normal SWR and drive levels, didn't you?) then you likely have a problem in the input circuitry or one or more tubes have failed. Perform a visual check of the input circuitry and the band switch, followed by an ohmmeter check of all input components.

If input SWR is normal and applying drive does not result in any change in plate current, you may have a defective tube, tube socket, or connection between the input circuits and the tube. Check the TR

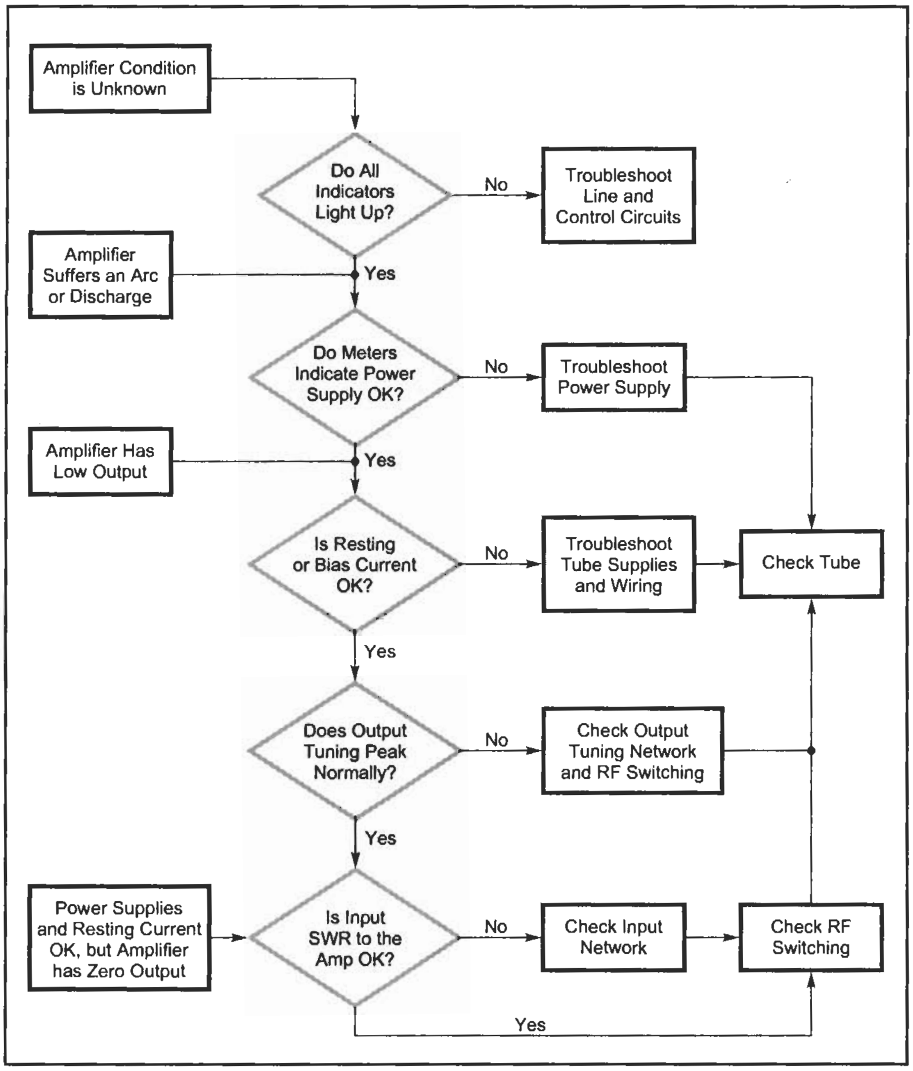
Amplifier Condition is Unknown
Do All Indicators Light Up?
No
Troubleshoot Line and Control Circuits
Amplifier Suffers an Arc or Discharge
Yes
Do Meters Indicate Power Supply OK?
No
Troubleshoot Power Supply
Amplifier Has Low Output
Yes
Is Resting or Bias Current OK?
No
Troubleshoot Tube Supplies and Wiring
Check Tube
Yes
Does Output Tuning Peak Normally?
No
Check Output Tuning Network and RF Switching
Yes
Power Supplies and Resting Current OK, but Amplifier has Zero Output
Is Input SWR to the Amp OK?
No
Check Input Network
Check RF Switching
Yes

Figure 4—This moderate-level flow chart is a good way to identify amplifier problems quickly in lieu of a misplaced or nonexistent manufacturer's troubleshooting procedure.

control circuits and relay. If plate current changes, but not as much as normal, try adjusting the output tuning circuitry. If this has little or no effect, the tube may be defective or a connection between the tube and output circuitry may have opened. If retuning has an effect, but at different settings than usual, the tube may be defective or there may be a problem in the tuning circuitry. A visual inspection and an ohmmeter check are in order.

The key to finding the trouble with your amplifier is to be careful and methodical, and to avoid jumping to false conclusions or making random tests. The manufac-turer's customer service department will likely be helpful if you are considerate and have taken careful notes detailing the trouble symptoms and any differences from normal operation. There may be helpful guidelines on the manufacturer's Web pages or from other Internet resources. Sometimes there is more than one problem—they work together to act like one very strange puzzle. Just remember that most problems are very simple and can be isolated by careful, step-by-step tests.

Summary

Amplifiers have been part of ham radio for many years. They are simple, reliable pieces of equipment that respond well to basic care and common sense. Take the time to know your amp—inside and out. If you take care of it, it will reward you with reliable service and maximum tube lifetime.

Notes

[1] Chapter 9 of the current *ARRL Handbook for Radio Communications* is an excellent source of safety information. Available from your local dealer or the ARRL Bookstore. Order no. 1921 (softcover), no. 1948 (hardcover). Telephone toll-free in the US 888-277-5289, or 860-594-0355, fax 860-594-0303; **www.arrl.org/shop/**; **pubsales@arrl.org**.

[2] Instructions for CPR can be found at: **about-the-web.com/spiritworks/web/Kaiser/html/adult.htm**.

[3] Two sources of HV and RF parts include Surplus Sales of Nebraska (**www.surplussales.com**) and Fair Radio Sales (**www.fairradio.com**). Others can be found at the ARRL Technical Information Service database (**www.arrl.org/tis/tisfind.html**).

[4] Penta Labs, "Tube Maintenance & Education" (**www.pentalaboratories.com/maintenance.asp**).

[5] Pelican (**www.pelican-shipping-cases.com**) and Anvil (**www.anvilsite.com**) make excellent shipping cases suitable for carrying amplifiers and radio equipment.

Build safe and reliable amplifiers. Produce loud and clean signals.

Planning to build an amplifier compares well to the work of a cook searching for a recipe and all the ingredients. Find the right design, and then locate the parts and components to begin building.

Shorten your discovery work with *ARRL's RF Amplifier Classics*! Use this book to find classic tube and solid-state designs at power levels from 5 W to 1.5 kW. Included are two-dozen amplifier projects and articles published in ARRL's popular journals between 1980 and 2003. There are amps for HF, MF, VHF and microwave. These are high quality works from respected authors such as Gary Breed, K9AY; Jerry Pittenger, K8RA; Bill Sabin, W0IYH; Al Ward, W5LUA; Dave Meacham, W6EMD and others.

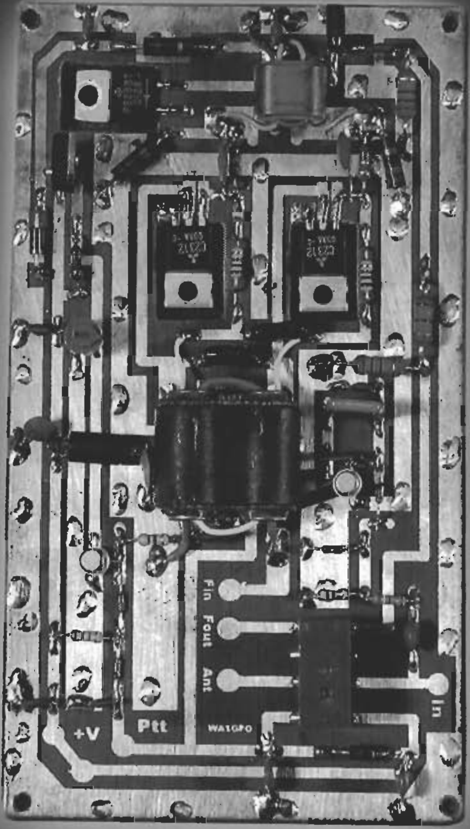

Turn dreams of constructing your first amp or next brick into reality!

ARRL *The national association for AMATEUR RADIO*
225 Main Street,
Newington, CT 06111-1494 USA
***ARRLWeb:* www.arrl.org/**

ISBN 0-87259-931-0
EAN
9 780872 599314
51995 >
ISBN 0-87259-931-0 ARRL Order No. 9310

TAB
ELECTRONICS

PRACTICAL ANTENNA HANDBOOK

Fourth Edition

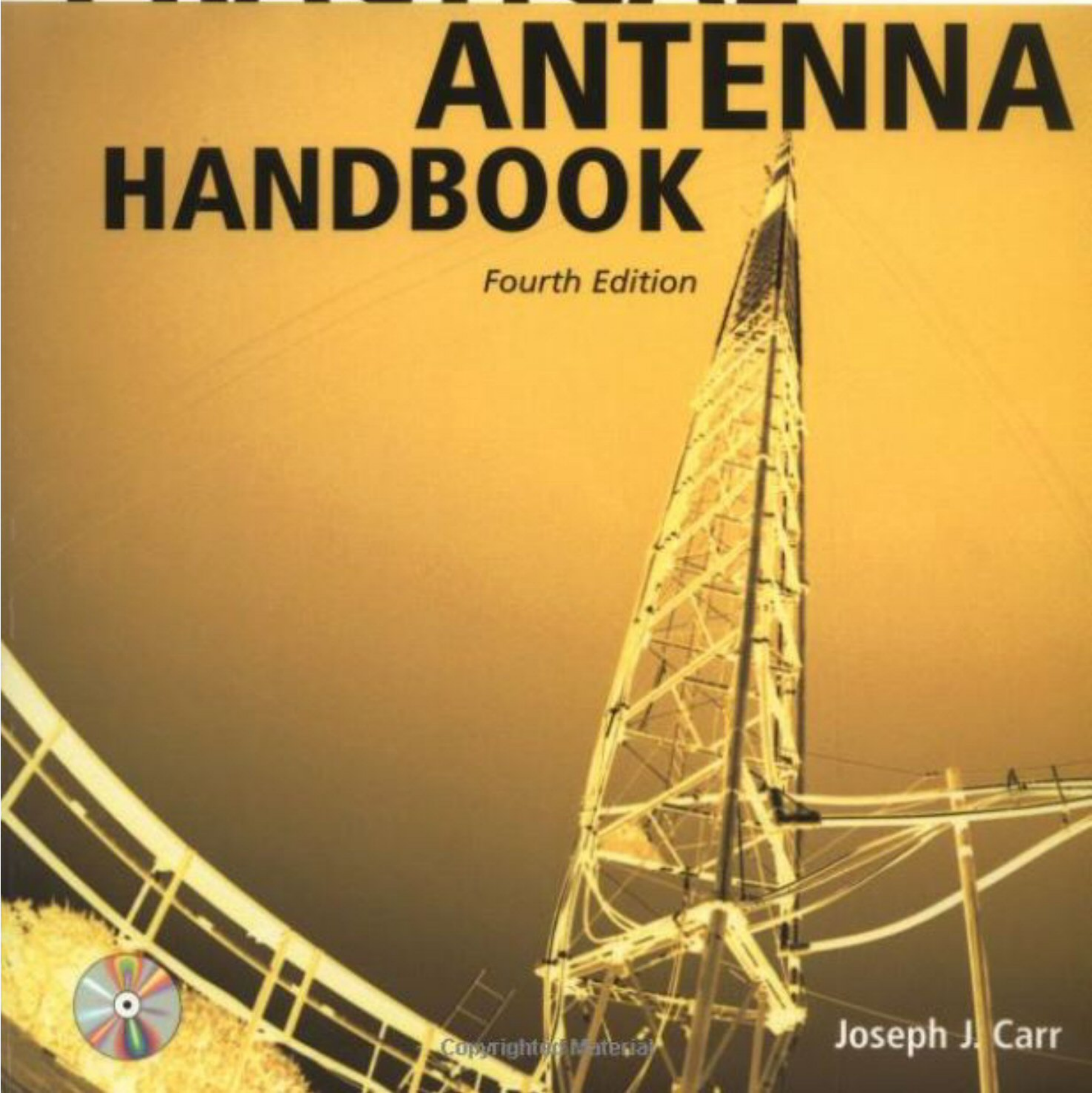

Joseph J. Carr

Copyrighted Material

Practical
Antenna
Handbook

Other Books by Joseph J. Carr

DC Power Supplies
Old Time Radios! Restoration and Repair
Secrets of RF Circuit Design

Practical Antenna Handbook

Fourth Edition

Joseph J. Carr

McGraw-Hill
New York Chicago San Francisco Lisbon London
Madrid Mexico City Milan New Delhi San Juan Seoul
Singapore Sydney Toronto

McGraw-Hill
A Division of The McGraw-Hill Companies

Copyright ©2001 by The McGraw-Hill Companies, Inc. All rights reserved. Manufactured in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

0-07-138931-8
The material in this eBook also appears in the print version of this title: 0-07-137435-3.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please contact George Hoare, Special Sales, at george_hoare@mcgraw-hill.com or (212) 904-4069.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. ("McGraw-Hill") and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hills prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS." McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

DOI: 10.1036/0071389318

Dedication

Dedicated in memoriam to Johnnie Harper Thorne, K4NFU:
a friend and colleague for nearly 30 years who is sorely missed.
Johnnie was a genius who knew some real smoke about antennas.
Killed by a drunk driver. . .

From the Publisher

The fourth edition of this book is published in memory of Joe Carr who passed away shortly after completing his manuscript. Joe was a valuable member of the McGraw-Hill family of authors and the electronics community as a whole. He will be missed by us all.

Scott L. Grillo
Editor-in-Chief

This page intentionally left blank

Contents

Introduction to the Fourth Edition *ix*

1 **Introduction to Radio Broadcasting and Communications** *1*

2 **Radio-wave Propagation** *5*

3 **Transmission Lines** *59*

4 **The Smith Chart** *95*

5 **Fundamentals of Radio Antennas** *123*

6 **High-Frequency Dipole and Other Doublet Antennas** *141*

7 **Vertically Polarized HF Antennas** *173*

8 **Multiband and Tunable-Wire Antennas** *203*

9 **Longwire Directional Antennas** *213*

10 **Hidden and Limited-Space Antennas** *231*

11 **Directional Phased Vertical Antennas** *245*

12 **Directional Beam Antennas** *255*

13 Antennas for Shortwave Reception ***271***

14 Large Wire Loop Antennas ***287***

15 Small Loop Receiving Antennas ***299***

16 Small Transmitting Loop Antennas ***319***

17 Antenna Modeling Software ***327***

18 VHF/UHF Transmitting and Receiving Antennas ***339***

19 Microwave Waveguides and Antennas ***369***

20 Antenna Noise Temperature ***417***

21 Antennas for Radio Astronomy ***421***

22 Adjusting, Installing, and Troubleshooting Antennas and Transmission Lines ***433***

23 Antennas for Radio Direction Finding (RDF) ***439***

24 Impedence Matching in Antenna Systems ***457***

25 Mobile, Emergency, Portable, and Marine Antennas ***479***

26 Antennas for Low-Frequency Operation ***501***

27 Measurement and Adjustment Techniques ***515***

28 General Antenna Mechanical Construction Techniques ***543***

29 Grounding the Antenna: What Is a Good Ground? ***573***

Index ***583***

McGraw-Hill Terms of Use

Introduction to the fourth edition

IT WAS WITH GREAT PLEASURE THAT I WROTE THIS FOURTH EDITION OF *PRACTICAL Antenna Handbook*. The original intent when the first edition was planned was to provide the reader with a practical, yet theoretical, book that could be used with only a minimal effort to actually design and install radio antennas. It was assumed that the readership would possess a wide range of levels of antenna sophistication, from the novice "newbie" to the professional engineer. That assumption proved to be correct as I have received letters from a wide variety of people.

The success of the three previous editions of *Practical Antenna Handbook* has been most gratifying. Clerks in radio stores told me that this book was outselling other titles 2:1. They confirmed that the types of people who buy the book meet the profile above. It was with surprise and delight that I noted that clerks in one radio store could cite the McGraw-Hill catalog number from memory, when they could not do that for other antenna books.

While the sales are deeply appreciated (they are an honor), there was one comment that stood out. The salesman at one radio store introduced me to an instructor from a U.S. Government communications school. He typically bought twenty copies of *Practical Antenna Handbook* at a time for use by his students in a training class. He told me that the reason why he selected my book over others was "...it's the *only book on the market* that people can give to a secretary, or clerk-typist, and expect them to be able to put up a working half-wavelength dipole two hours later." And, he stated, in his business that could literally happen any time.

Antennas have changed a lot over the years. Figure I-1 shows how antennas were between 1913 and 1940. This facility was the first U.S. Navy radio station (NAA) in Arlington, VA.[1] Those two 600- and 400-ft towers were taken down in 1940 to make room for National Airport (Washington, D.C.). The towers were reassembled in Annapolis, MD, where they lasted until recently.

Antenna technology has changed since 1940. In preparing this fourth edition of *Practical Antenna Handbook* all previous material was reviewed for accuracy and relevance. Additional material was added for the following topics:

- Wire antenna construction methods
- Antenna modeling software (miniNEC, WinNEC, and EZNEC for Windows)
- Antenna noise temperature
- Antennas for radio astronomy
- Antennas for Radio Direction Finding

I hope that you find this fourth edition as useful as the previous three editions. And thank you very much for honoring me by spending your hard-earned money on my book.

Joseph J. Carr, MSEE

---

[1] The NAA callsign has since been reassigned to the VLF station at Cutler, ME.

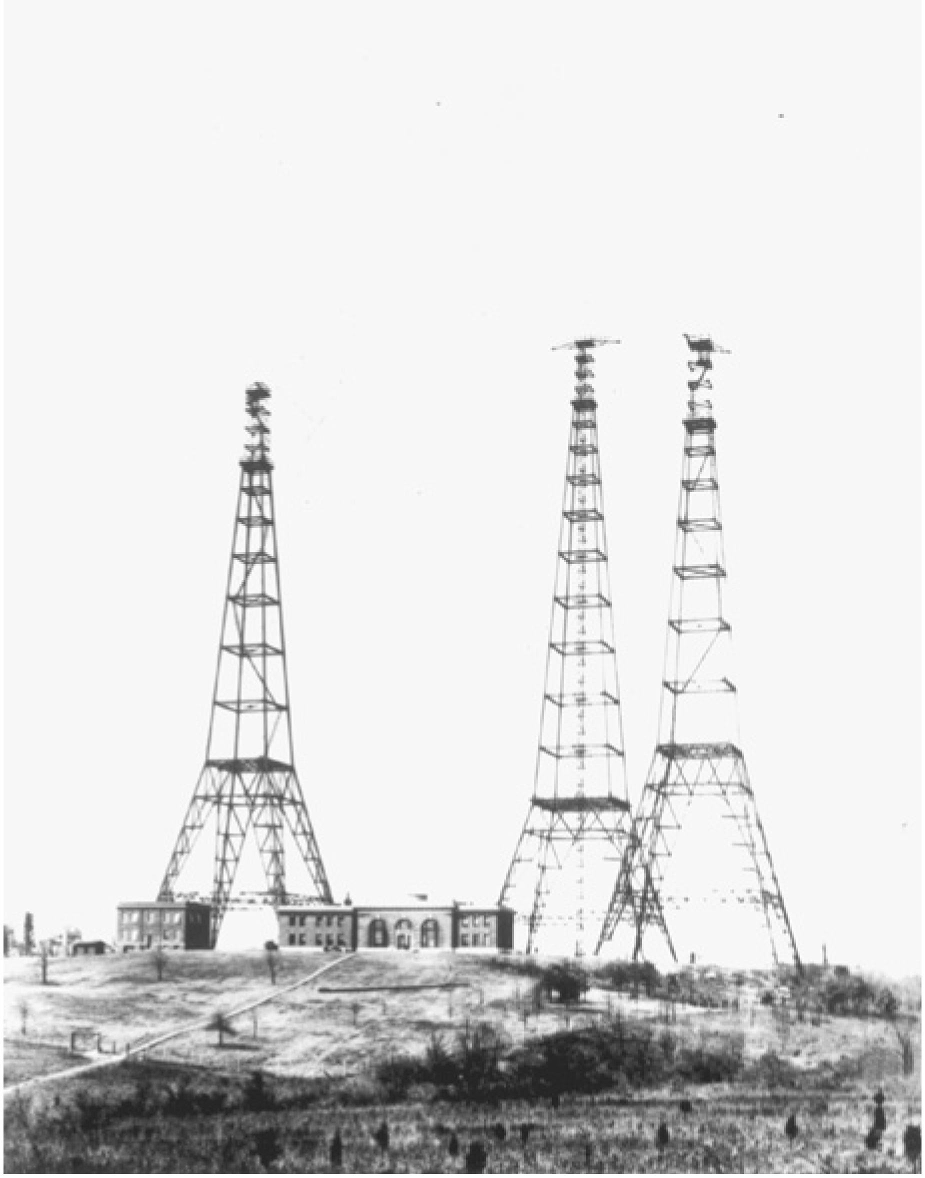

I-1 Massive antenna towers at Navy "Radio Arlington" (1913-1940). (Courtesy of the NAA.)

This page intentionally left blank

1
CHAPTER

Introduction to radio broadcasting and communications

RADIO BROADCASTING AND COMMUNICATIONS SEEMS TO HOLD A STRANGE KIND OF magical allure that attracts a wide variety of people and holds them for years. There is something fascinating about the ability to project yourself over vast intercontinental distances.

Radio communications have been with us now for the entire twentieth century and into the twenty-first. Experiments are on record as early as 1867, and by the turn of the century "wireless telegraphy" (as radio was called then) sparked the imaginations of countless people across the world. Radio communications began in earnest, however, when Guglielmo Marconi successfully demonstrated wireless telegraphy as a commercially viable entity. The "wireless" aspect to radio so radically changed communications that the word is still used to denote radio communications in many countries of the world. Marconi made a big leap to international fame on a cold December day in 1903, when he and a team of colleagues successfully demonstrated transatlantic wireless telegraphy. Until that time, wireless was a neighborhood—or crosstown at best—endeavor that was of limited usefulness. Of course, ships close to shore, or each other, could summon aid in times of emergency, but the ability to communicate over truly long distances was absent. All that changed on that fateful day in Newfoundland when Marconi heard the Morse letter "S" tickle his ears.

Wireless telegraphy was pressed into service by shipping companies because it immediately provided an element of safety that was missing in the prewireless days. Indeed, a ship that sank, leaving its crew and passengers afloat on a forbidding sea, was alone. Any survivors often succumbed to the elements before a chance encounter with a rescue vessel. Some early shipping companies advertised that their ships were safer because of wireless aboard. It was not until 1909, however, that wireless telegraphy proved its usefulness on the high seas. Two ships collided in the foggy Atlantic Ocean and were sinking. All passengers and crew members of both ships were in imminent danger of death in the icy waters. But radio operator Jack

Copyright 2001 - the McGraw-Hill Companies

Binns became the first man in history to send out a maritime distress call. There is some debate over which distress call Binns transmitted, but one thing is certain: It was not "SOS" (today's distress call), because SOS was not adopted until later. Binns probably transmitted either "CQD" or "CQE," both of which were recognized in those days before standardization. Regardless of which call was sent, however, it was received and relayed from ship to ship, allowing another vessel to come to the aid of the stricken pair of ships.

All radio prior to about 1916 was carried on via telegraphy (i.e., the on-off keying of a radio signal in the Morse code). But in 1916 some more magic happened. On a little hill in Arlington, Virginia, on a site that now overlooks the Pentagon and the U.S. Marine Corps base called *Henderson Hall*, there were (and still are) a pair of two-story brick buildings housing the naval radio station NAA (callsign since reassigned to the VLF station at Cutler, ME). On a night in 1916, radio operators and monitors up and down the Atlantic seaboard—from the midwest to the coast and out to sea for hundreds of miles—heard something that must have startled them out of their wits, for crackling out of the "ether," amidst the whining of Alexanderson alternators and "ZZZCHHT" of spark-gap transmitters, came a new sound—a human voice. Engineers and scientists at the Naval Research Laboratory had transmitted the first practical amplitude-modulated (AM) radio signal. Earlier attempts, prior to 1910, had been successful as scientific experiments, but they did not use commercially viable equipment.

Although radio activity in the early years was unregulated and chaotic, today it is quite heavily regulated. Order was brought to the bands (don't laugh, ye who tune the shortwaves) that was lacking before. Internationally radio is regulated by the International Telecommunications Union (ITU) in Geneva, Switzerland through the treaties arising from World Administrative Radio Conferences (WARC) held every 10 to 15 years. In the United States, radio communications are regulated by the Federal Communications Commission (FCC), headquartered in Washington, D.C.

Amateur radio has grown from a few thousand "hams" prior to World War I to more than 900,000 today, about one-third of them in the United States. Amateur operators were ordered off the air during World War I, and almost did not make a comeback after the war. There were, by that time, many powerful commercial interests that greedily coveted the frequencies used by amateurs, and they almost succeeded in keeping postwar amateurs off the air. But the amateurs won the dispute, at least partially. In those days, it was the frequencies with wavelengths longer than 200 m (i.e., 20 to 1500 kHz) that were valuable for communications.

The cynical attitude attributed to the commercial interests regarding amateurs was, "put 'em on 200 meters and below . . . they'll never get out of their backyards there!" But there was a surprise in store for those commercial operators, because the wavelengths shorter than 200 m are in the high-frequency region that we now call "shortwaves." Today, the shortwaves are well-known for their ability to communicate over transcontinental distances, but in 1919 that ability was not suspected.

I once heard an anecdote from an amateur operator "who was there." In the summer of 1921 this man owned a large, beautiful wire "flattop" antenna array for frequencies close to 200 m on his family's farm in southwestern Virginia. Using those frequencies he was used to communicating several hundred miles into eastern Ohio and down to the Carolinas. But, in September 1921 he went to college to study electrical engineering at the University of Virginia in Charlottesville. When he returned home for Thanksgiving he noticed that his younger brother had replaced the long flattop array with a short dipole antenna. He was furious, but managed through great effort to contain the anger until after dinner. Confronting his brother over the incredible sacrilege, he was told that they no longer used 150 to 200 m, but rather were using 40 m instead. Everyone "knew" that 40 m was useless for communications over more than a few blocks, so (undoubtedly fuming) the guy took a turn at the key. He sent out a "CQ" (general call) and was answered by a station with a callsign like "8XX." Thinking that the other station was in the 8th U.S. call district (WV, OH, MI) he asked him to relay a message to a college buddy in Cincinnati, OH. The other station replied in the affirmative, but suggested that ". . . you are in a better position to reach Cincinnati than me, I am *FRENCH* 8XX." (Callsigns in 1921 did not have national prefixes that are used today.) The age of international amateur communications had arrived! And with it came a new problem—the national identifiers in call signs became necessary (which is why American call signs begin with K, W, or N).

During the 1930s, radio communications and broadcasting spread like wildfire as technology and techniques improved. World War II became the first war to be fought with extensive electronics. Immediately prior to the war, the British developed a new weapon called RADAR (radio detection and ranging; now spelled radar). This tool allowed them to see and be forewarned of German aircraft streaming across the English Channel to strike targets in the United Kingdom. The German planes were guided by (then sophisticated) wireless highways in the sky, while British fighters defended the home island by radio vectoring from ground controllers. With night fighters equipped with the first "centimetric" (i.e., microwave) radar, the Royal Air Force was able to strike the invaders accurately—even at night. The first kill occurred one dark, foggy, moonless night when a *Beaufighter* closed on a spot in the sky where the radar in the belly of the plane said an enemy plane was flying. Briefly thinking he saw a form in the fog, the pilot cut loose a burst from his quad mount of 20-mm guns slung in the former bomb bay. Nothing. Thinking that the new toy had failed, the pilot returned to base—only to be told that ground observers had reported that a German *Heinkle* bomber fell from the overcast sky at the exact spot where the pilot had his ghostly encounter.

Radio, television, radar, and a wide variety of services, are available today under the general rubric "radio communications and broadcasting." Homeowners, and other nonprofessionals in radio, can own a receiver system in their backyard that picks up television and radio signals from satellites in geosynchronous orbit 23,000 mi out in space. Amateur operators are able to communicate worldwide on low power, and have even launched their own "OSCAR" satellites.

Some people had written off the HF radio spectrum in recent years, citing satellite technology as the reason. But the no-code license for amateur radio operators, which does not carry HF privileges, has proven to be a stepping stone to higher-class licenses, which do. Also, the shortwave broadcasting market received a tremendous boost during the Gulf War. When the troops of *Operation Desert Shield* and *Desert Storm* were assembling to take back Kuwait from the Iraqis, the sales of shortwave receivers jumped dramatically. And, following January 16, 1991, when the forces started pouring across the border into the actual fight, the sales skyrocketed out of sight. One dealer told me that he couldn't keep receivers in stock, and that he had sold out most models. That interest seems to have matured into long-term interest on the part of a significant number of listeners and new ham operators.

The antenna is arguably one of the most important parts of the receiving and/or transmitting station (Fig. 1-1). That is what this book is all about.

1-1 This AM/FM broadcast antenna tower bristles with two-way antennas.

2
CHAPTER

Radio-wave propagation

THE PROPAGATION OF RADIO SIGNALS IS NOT THE SIMPLE MATTER THAT IT SEEMS AT first glance. Intuitively, radio signal propagation seems similar to light propagation; after all, light and radio signals are both electromagnetic waves. But simple inverse square law predictions, based on the optics of visible light, fall down radically at radio frequencies because other factors come into play. In the microwave region of the spectrum, the differences are more profound because atmospheric pressure and water vapor content become more important than for light. For similar reasons, the properties of microwave propagation differ from lower VHF and HF propagation. In the HF region, solar ionization of the upper reaches of the atmosphere causes the kind of effects that lead to long-distance "skip" communications and intercontinental broadcasting. This chapter examines radio propagation phenomena so that you have a better understanding of what an antenna is used for and what parameters are important to ensure the propagation results that you desire.

Radio waves

Although today it is well recognized that radio signals travel in a wave-like manner, that fact was not always so clear. It was well known in the first half of the nineteenth century that wires carrying electrical currents produced an *induction field* surrounding the wire, which is capable of causing action over short distances. It was also known that this induction field is a magnetic field, and that knowledge formed the basis for electrical motors. In 1887, physicist Heinrich Hertz demonstrated that radio signals were *electromagnetic waves*, like light. Like the induction field, the electromagnetic wave is created by an electrical current moving in a conductor (e.g., a wire). Unlike the induction field, however, the *radiated field* leaves the conductor and propagates through space as an electromagnetic wave.

Copyright 2001 - the McGraw-Hill Companies

The *propagation* of waves is easily seen in the "water analogy." Although not a perfect match to radio waves, it serves to illustrate the point. Figure 2-1 shows a body of water into which a ball is dropped (Fig. 2-1A). When the ball hits the water (Fig. 2-1B), it displaces water at its point of impact, and pushes a leading wall of water away from itself. The ball continues to sink and the wave propagates away from it until the energy is dissipated. Although Fig. 2-1 shows the action in only one dimension (a side view), the actual waves propagate outward in all directions, forming concentric circles when viewed from above.

The wave produced by a dropped ball is not continuous, but rather is *damped* (i.e., it will reduce in amplitude on successive crests until the energy is dissipated and the wave ceases to exist). But to make the analogy to radio waves more realistic, the wave must exist in a continuous fashion. Figure 2-2 shows how this is done: a ball is dipped up and down in a rhythmic, or cyclic manner, successively rein-

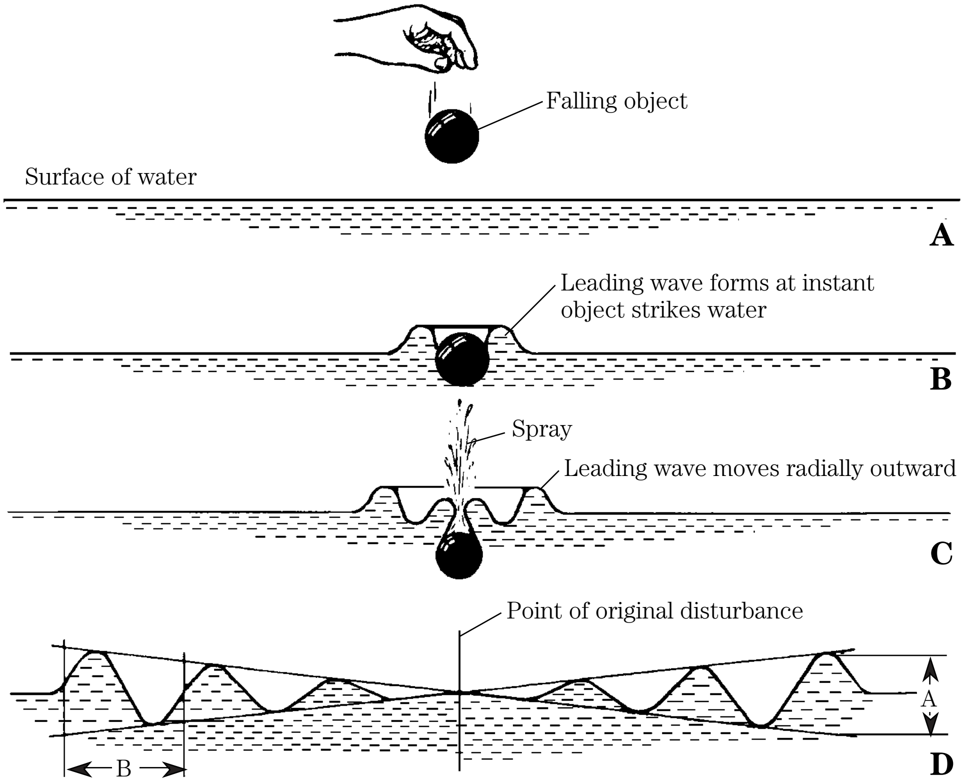
Falling object
Surface of water
A
Leading wave forms at instant object strikes water
B
Spray
Leading wave moves radially outward
C
Point of original disturbance
A
B
D

Notes: A Amplitude of leading wave
B Corresponds to 1 cycle of oscillation

2-1 A ball dropped into water generates a wavefront that spreads out from the point of original disturbances.

forcing new wave crests on each dip. The waves continue to radiate outward as long as the ball continues to oscillate up and down. The result is a *continuous wave train*.

There are two related properties of all waves that are important to radio waves as well: *frequency* (*f*) and *wavelength* (λ). The frequency is the number of oscillations (or cycles) per unit of time. In radio waves, the unit of time is the second, so frequency is an expression of the number of *cycles per second* (*cps*). If the period of time required for the leading wave to travel from point "A" to "B" is one second (1 s), and there are two complete wave cycles in that space, then the frequency of the wave created by the oscillating ball is 2 cps.

At one time, radio frequencies (along with the frequencies of other electrical and acoustical waves) were expressed in cps, but in honor of Heinrich Hertz, the unit was renamed the hertz (Hz) many years ago. Because the units are equal (1 Hz = 1 cps), the wave in Fig. 2-2 has a frequency of 2 Hz.

Because radio frequencies are so high, the frequency is usually expressed in kilohertz (kHz—1000s of Hz) and megahertz (MHz—1,000,000s of Hz). Thus, the frequency of a station operating in the middle of the AM broadcasting band can be properly expressed as 1,000,000 Hz, or 1000 kHz, or 1 MHz, all of which are equivalent to each other. Radio dials in North America are usually calibrated in kHz or MHz. In Europe and the rest of the world, on the other hand, it is not uncommon to find radio dials calibrated in meters, the unit of wavelength, as well as in frequency. In most equations used in radio antenna design, the proper units are hertz, kilohertz, and megahertz.

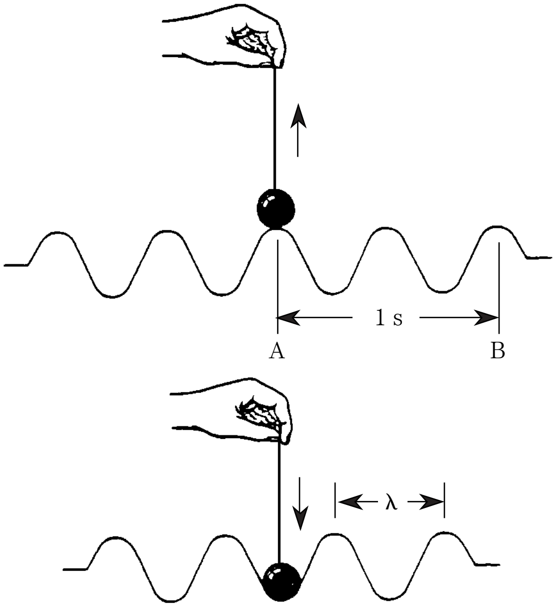

2-2 Bobbing ball on a string demonstrates continuous wave generation action.

The *wavelength* of any wave is the distance between like features on the waveform. In the case of Fig. 2-2, the wavelength (λ) is the distance between successive positive peaks. We could also measure the same distance between successive negative peaks, or between any two similar features on successive waves. In radio work, the wavelength of the signal is expressed in *meters or its subunits*.

The wavelength is proportional to the reciprocal of the frequency. The wavelength of any wave is related to the frequency so that $f\lambda = v$, where f is the frequency in Hz, λ is the wavelength in meters, and v is the *velocity of propagation* in meters per second (m/s). Because radio waves propagate at the speed of light (which is also an electromagnetic wave), approximately 300,000,000 m/s in both free space and the earth's atmosphere, the lowercase letter c is used to represent velocity (rather than v), so you can rewrite this expression in the form

$$f_{\text{Hz}} = \frac{c}{\lambda_{\text{meters}}} = \frac{300{,}000{,}000}{\lambda_{\text{meters}}} \qquad \textbf{[2.1]}$$

These equations are sometimes abbreviated for use with the units kHz and MHz:

$$F_{\text{kHz}} = \frac{300{,}000}{\lambda_{\text{meters}}} \qquad \textbf{[2.2]}$$

$$f_{\text{MHz}} = \frac{300}{\lambda_{\text{meters}}} \qquad \textbf{[2.3]}$$

You can get an idea of the order of length of these waves by solving Eq. 2.3 for several different frequencies: 100 kHz, 1 MHz (in the AM broadcast band), 10 MHz (in the shortwave bands), and 1000 MHz (microwave bands). If you work the equations, then you will find that these wavelengths are 3000 m (100 kHz), 300 m (1 MHz), 30 m (10 MHz), and 0.3 m, or 30 cm (1000 MHz). You can see from these numbers why 1 MHz is in what is called the *medium-wave band*, 10 MHz is in the *short wave band*, and 1,000 MHz is in the *microwave* ("very small" wave) band. At 100 kHz, which is 0.1 MHz, the wavelength is 3000 m, so this frequency is in the *long-wave band*.

The place where the water analogy falls down most profoundly is in the nature of the medium of propagation. Water waves move by moving water molecules; water is said to be the *medium* in which the wave propagates. At one time, scientists could not conceive of the "action at a distance" provided by radio waves, so they invented a hypothetical medium called *ether* (or *aether*) for propagating electromagnetic waves (such as radio waves and light). It was not until the late nineteenth century that American physicists Michaelson and Morley proved that the ether does not exist. Nonetheless, radio enthusiasts still refer to the "stuff" out of which radio waves arrive as the "ether." This terminology is merely an archaic, linguistic echo of the past.

The electromagnetic field: a brief review

A great deal of heavily mathematical material can be presented about electromagnetic waves. Indeed, developing Maxwell's equations is a complete field of study for specialists. In this section, you will not use this rigorous treatment because you can refer to engineering textbooks for that depth of information. The purpose here is to present a descriptive approach that is designed to present you with a basic understanding of the phenomena. The approach here is similar to the learning of a "conversational" foreign language, rather than undertaking a deep study of its grammar, syntax, and context. For those whose professional work routinely involves electromagnetic waves, this treatment is hopelessly simplistic. For that I make no apology, because it serves a greater audience. The goal here is to make you more comfortable when thinking about the propagation of electromagnetic fields in the radio portion of the electromagnetic spectrum.

Radio signals are *transverse electromagnetic* (TEM or EM) *waves* exactly like light, infrared (IR), and ultraviolet (UV), except for frequency. Radio waves have much lower frequencies than light, IR, or UV, hence they have much longer wavelengths. The TEM (EM) wave consists of two mutually perpendicular oscillating fields (see Fig. 2-3) traveling together in phase. One of the fields is an *electric field* and the other is a *magnetic field*.

Radio-wave intensity

The radio wave is *attenuated* (i.e. reduced in apparent power) as it propagates from the transmitter to the receiver. Although at some very high microwave frequencies, there is additional path loss as a result of the oxygen and water vapor content of the air. At other frequencies other losses exist. Radio waves at all frequencies suffer losses due to the *inverse square law*. Let's take a look at that phenomenon.

The electric field vector falls off in direct proportion to the distance traveled. The E field is measured in terms of volts per meter (V/m), or in the subunits millivolts per meter (mV/m) or microvolts per meter (μV/m). That is, if an E field of

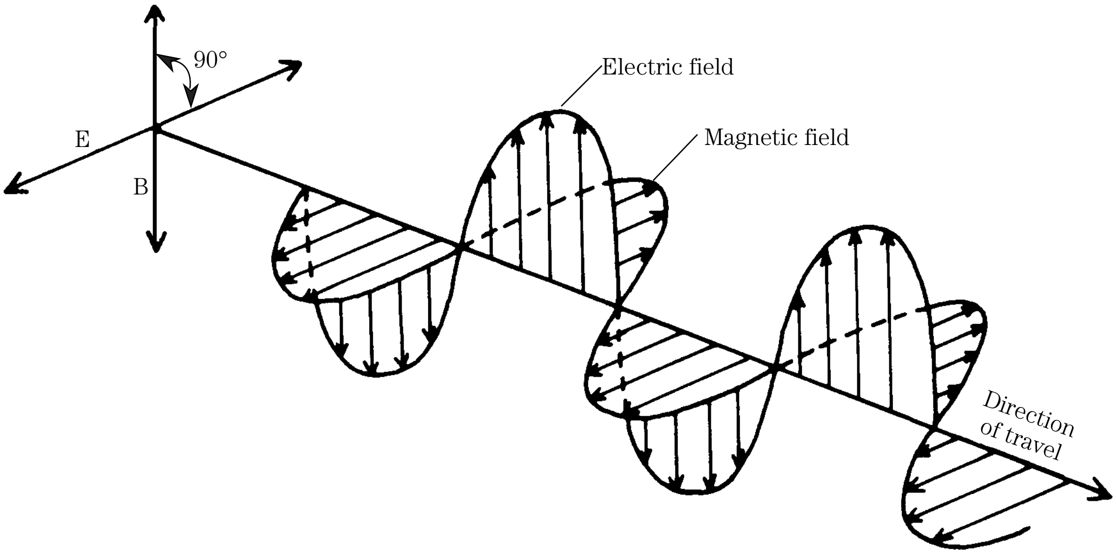

2-3 Electromagnetic wave consists of right angle electric and magnetic fields.

10 V/m crosses your body from head to toe, and you are about 2 m tall, then an electrical voltage is generated of (2 m)×(10 V/m), or 20 V. The reduction of the E field is linearly related to distance (i.e., if the distance doubles the E-field voltage vector halves). Thus, a 100 mV/m E field that is measured 1 mi from the transmitter will be 50 mV/m at 2 mi.

The power in any electrical system is related to the voltage by the relationship

$$P = \frac{E^2}{R} \tag{2.4}$$

where

P is the power in watts (W)
R is the resistance in ohms (Ω)
E is the electrical potential in volts (V)

In the case of a radio wave, the R term is replaced with the impedance (Z) of free space, which is on the order of 377 Ω. If the E field intensity is, for example, 10 V/m, then the *power density* of the signal is

$$P = \frac{(10\ \text{V/m})^2}{377\ \Omega} = 0.265\ \text{W/m}^2 = 26.5\ \text{mW/m}^2 \tag{2.5}$$

The power density, measured in watts per square meter (W/m^2), or the subunits (e.g., mW/cm^2), falls off according to the square of the distance. This phenomenon is shown graphically in Fig. 2-4. Here, you can see a lamp emitting a light beam that falls on a surface (A), at distance L, with a given intensity. At another surface (B), that is $2L$ from the source, the same amount of energy is distributed over an area (B) that is twice as large as area A. Thus, the power density falls off according to $1/d^2$, where d is the difference in distance. This is called the *inverse square law*.

Isotropic sources

In dealing with both antenna theory and radio wave propagation, a theoretical construct called an *isotropic source* is sometimes used for the sake of comparison, and for simpler arithmetic. You will see the isotropic model several places in this book. An isotropic source assumes that the radiator (i.e., an "antenna") is a very tiny spherical source that radiates energy equally well in all directions. The radiation pattern is thus a sphere with the isotropic antenna at the center. Because a spherical source is uniform in all directions, and its geometry is easily determined mathematically, the signal intensities at all points can be calculated from basic geometric principles.

For the isotropic case, you can calculate the average power in the extended sphere from

$$P_{\text{av}} = \frac{P_t}{4\pi d^2} \tag{2.6}$$

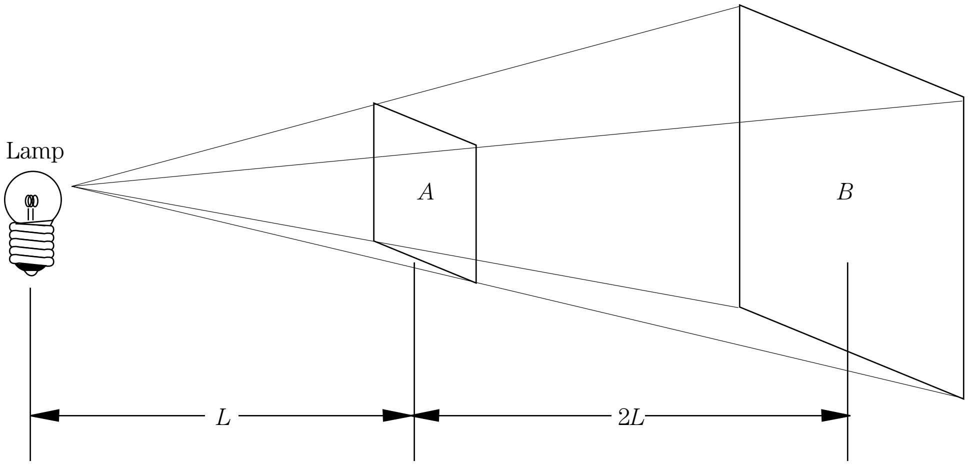

2-4 As a wave propagates, it spreads out according to the inverse square law, i.e., the area of B is 4 times that of A because it is twice as far from the source.

where

P_{av} is the average power per unit area
P_t is the total power
d is the radius of the sphere in meters (i.e., the distance from the radiator to the point in question)

The *effective aperture* (A_e) of the receiving antenna relates to its ability to collect power from the EM wave and deliver it to the load. Although typically smaller than the surface area in real antennas, for the theoretical isotropic case $A_e = \lambda^2/4\pi$. The power delivered to the load is

$$P_L = P_{\mathrm{av}} A_e \qquad \textbf{[2.7]}$$

By combining the two previous equations, the power delivered to a load at distance d is given by

$$P_L = \frac{P_t \lambda^2}{(4\pi)^2 d^2} \qquad \textbf{[2.8]}$$

where

P_L is the power to the load
λ is the wavelength (c/F) of the signal

From these expressions, there can then be derived an expression for ordinary path losses between an isotropic transmitter antenna and a receiver antenna:

$$L_{\mathrm{dB}} = 10 \log \left(\frac{P_t}{P_L} \right) \qquad \textbf{[2.9]}$$

or, by rearranging to account for individual terms:

$$L_{\text{dB}} = [20 \log d] + [20 \log F_{\text{MHz}}] + k \qquad \textbf{[2.10]}$$

where

L_{dB} is the path loss in decibels (dB)
d is the path length
F_{MHz} is frequency in megahertz (MHz)
k is a constant that depends on the units of d as follows:

k = 32.4 if d in kilometers
= 36.58 if d in statute miles
= 37.80 if d in nautical miles
= –37.87 if d in feet
= –27.55 if d in meters

The radiated sphere of energy gets ever larger as the wave propagates away from the isotropic source. If, at a great distance from the center, you take a look at a small slice of the advancing wavefront you can assume that it is essentially a flat plane, as in Fig. 2-5. This situation is analogous to the apparent flatness of the prairie, even though the surface of the earth is nearly spherical. You would be able to "see" the electric and magnetic field vectors at right angles to each other in the flat plane wavefront.

The *polarization* of an EM wave is, by definition, the *direction of the electric field*. Figure 2-6A shows vertical polarization, because the electric field is vertical with respect to the earth. If the fields were exchanged (as in Fig. 2-6B), then the EM wave would be horizontally polarized.

These designations are especially convenient because they also show the type of antenna used: vertical antennas produce vertically polarized signals, while hori-

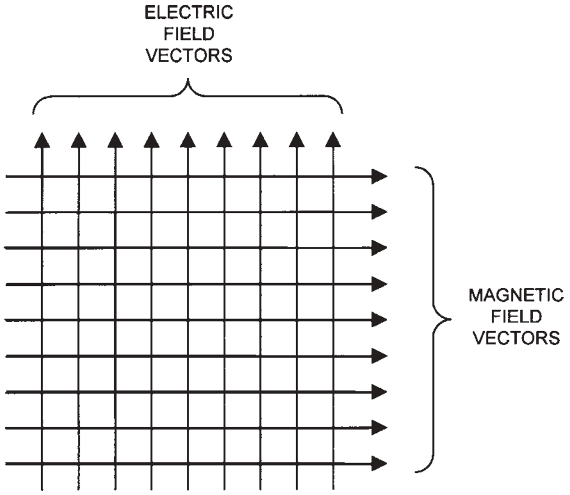

2-5 When viewed as a plane, the fields appear at right angles to each other.

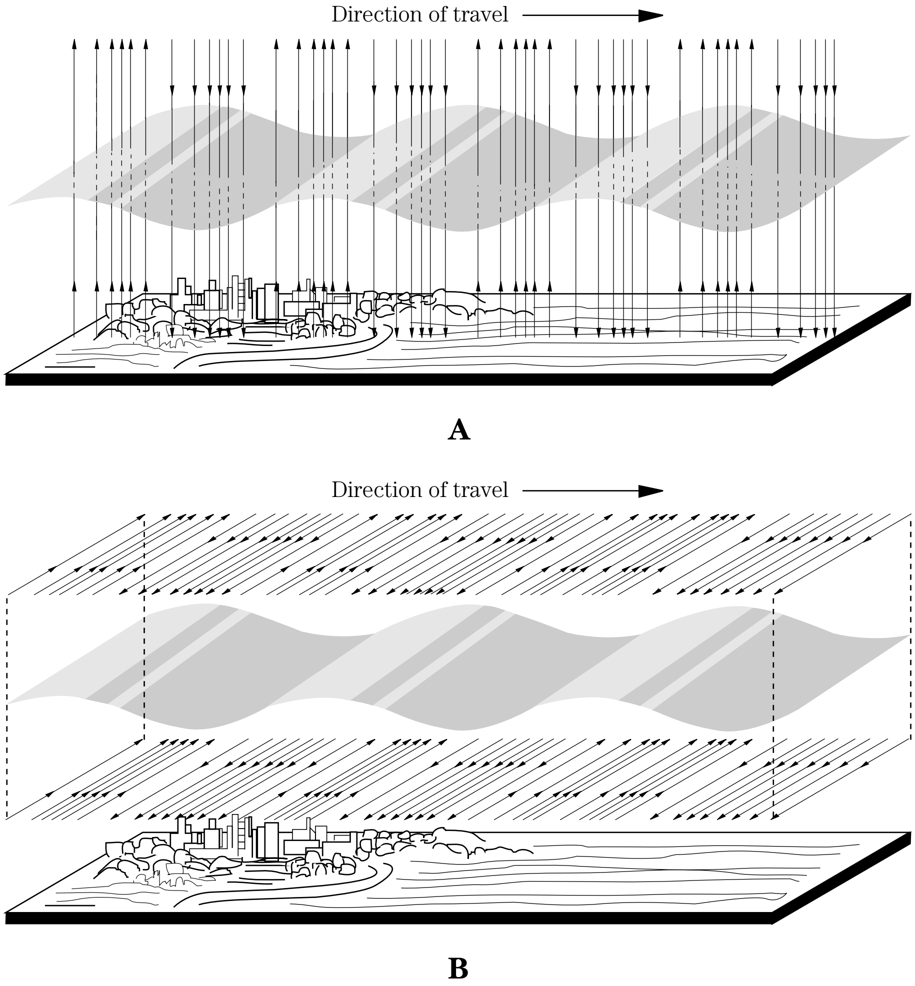

2-6 Wave polarization is determined by the direction of the electric field lines of force: (A) vertical polarized electromagnetic wave; (B) horizontally polarized wave.

zontal antennas produce horizontally polarized signals. At least one text erroneously states that antennas will not pick up signals of the opposite polarity. Such is not the case, especially in the high-frequency (HF) and lower very high-frequency (VHF) regions. At VHF, ultrahigh frequency (UHF), and microwave frequencies a loss of approximately 20 to 30 dB can be observed due to cross-polarization.

An EM wave travels at the speed of light, designated by the letter *c*, which is about 300,000,000 m/s (or 186,000 mi/s). To put this velocity in perspective, a radio signal originating on the sun's surface would reach earth in about 8 minutes. A terrestrial radio signal can travel around the earth seven times in one second. The velocity of the wave slows in dense media, but in air the speed is so close to the "free-space" value of *c*, that the same figures are used for both air and the near vacuum of outer space in practical problems. In pure water, which is much denser than air, the speed of radio signals is about one-ninth that of the free-space speed. This same phenomenon shows up in practical work in the form of the *velocity factor* (*V*) of transmission lines. In foam dielectric coaxial cable, for example, the value of *V* is 0.80, which means that the signal propagates along the line at a speed of 0.80*c*, or 80 percent of the speed of light.

This coverage of radio propagation considers the EM wave as a very narrow "ray" or "pencil beam" that does not diverge as it travels. That is, the ray remains the same width all along its path. This convention makes it easy to use ray tracing diagrams. Keep in mind, however, that the real situation, even when narrow-beamwidth microwave signals are used, is much more complicated. Real signals, after all, are sloppier than textbook examples: they are neither infinitesimally thin, nor nondivergent.

The earth's atmosphere

The electromagnetic waves do not need an atmosphere in order to propagate, as you will undoubtedly realize from the fact that space vehicles can transmit radio signals back to earth in a near vacuum. But when a radio wave does propagate in the earth's atmosphere, it interacts with the atmosphere, and its path of propagation is altered. A number of factors affect the interaction, but it is possible to break the atmosphere into several different categories according to their respective effects on radio signals.

The atmosphere, which consists largely of oxygen (O_2) and nitrogen (N) gases, is broken into three major zones: *troposphere*, *stratosphere*, and *ionosphere* (Fig. 2-7). The boundaries between these regions are not very well defined, and change both diurnally (over the course of a day) and seasonally.

The troposphere occupies the space between the earth's surface and an altitude of 6 to 11 km (4 to 7 mi). The temperature of the air in the troposphere varies with altitude, becoming considerably lower at greater altitude compared with ground temperature. For example, a +10°C surface temperature could reduce to –55°C at the upper edges of the troposphere.

The stratosphere begins at the upper boundary of the troposphere (6 to 11 km), and extends up to the ionosphere (≈50 km). The stratosphere is called an *isothermal region* because the temperature in this region is somewhat constant, despite altitude changes.

The ionosphere begins at an altitude of about 50 km (31 mi) and extends up to approximately 300 km (186 mi). The ionosphere is a region of very thin atmosphere. Cosmic rays, electromagnetic radiation of various types (including ultraviolet light from the sun), and atomic particle radiation from space (most of these from the sun also), have sufficient energy to strip electrons away from the gas molecules of the atmosphere. These freed electrons are called *negative ions*, while the O_2 and N molecules that lost electrons are called *positive ions*. Because the density of the air is so low at those altitudes, the ions can travel long distances before neutralizing each other by recombining. Radio propagation on some bands varies markedly between daytime and nighttime because the sun keeps the level of ionization high during

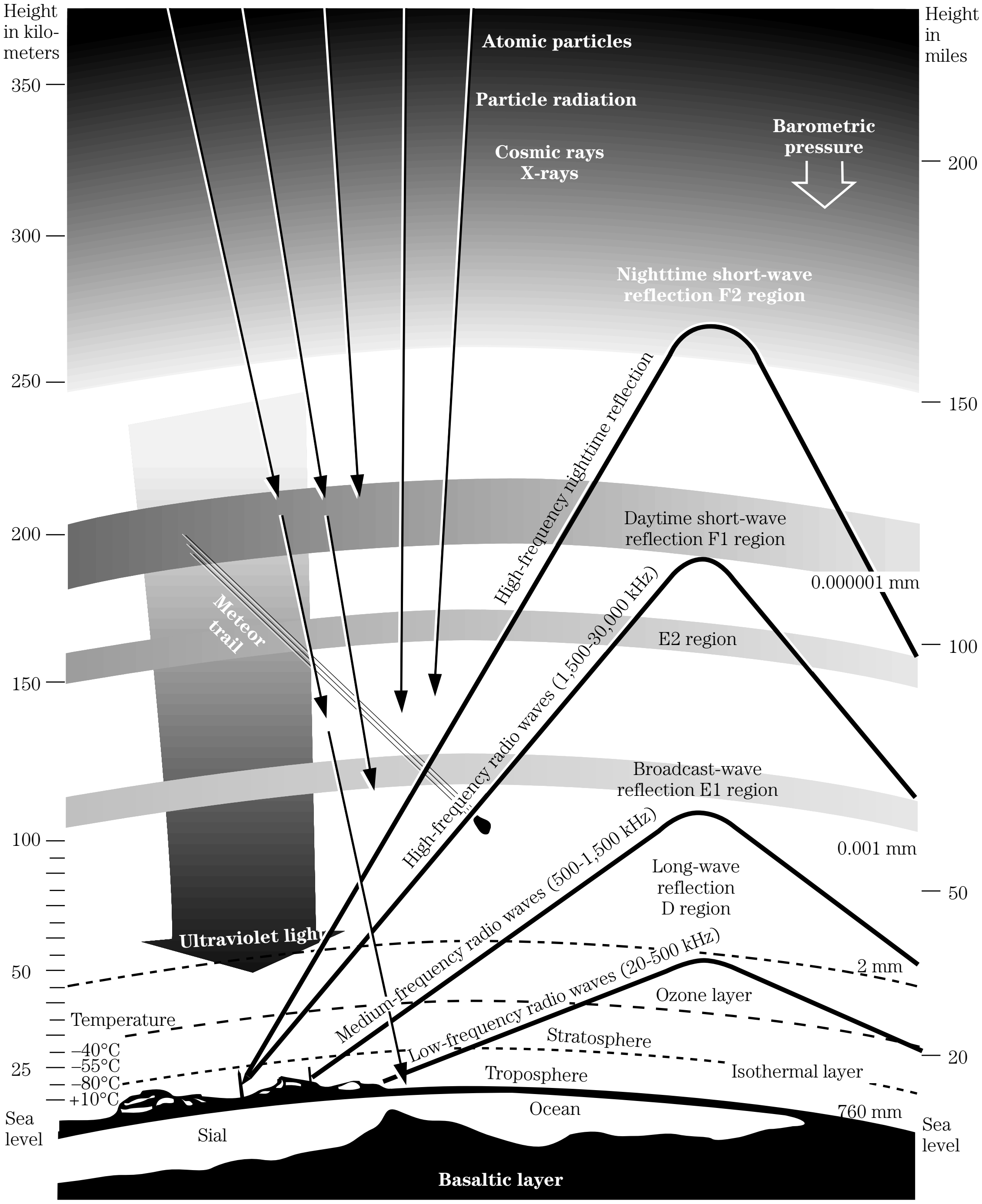
Height in kilometers
350
300
250
200
150
100
50
25
Sea level
Height in miles
200
150
100
50
20
Sea level
Atomic particles
Particle radiation
Cosmic rays
X-rays
Barometric pressure
Nighttime short-wave reflection F2 region
Daytime short-wave reflection F1 region
0.000001 mm
E2 region
Broadcast-wave reflection E1 region
0.001 mm
Long-wave reflection D region
2 mm
Ozone layer
Stratosphere
Isothermal layer
Troposphere
760 mm
Ocean
Sial
Basaltic layer
Meteor trail
Ultraviolet light
High-frequency nighttime reflection
High-frequency radio waves (1,500-30,000 kHz)
Medium-frequency radio waves (500-1,500 kHz)
Low-frequency radio waves (20-500 kHz)
Temperature
−40°C
−55°C
−80°C
+10°C

2-7 Radio propagation in the ionosphere is affected by a number of different physical factors: cosmic rays, atomic particles, solar radiation.

daylight hours, but the ionization begins to fall off rapidly after sunset, altering the radio propagation characteristics after dark. The ionization does not occur at lower altitudes because the air density is such that the positive and negative ions are numerous and close together, so recombination occurs rapidly.

EM wave propagation phenomena

Because EM waves are waves, they behave in a wave-like manner. Figure 2-8 illustrates some of the wave behavior phenomena associated with light and radio waves: *reflection*, *refraction*, and *diffraction*. All three play roles in radio propagation.

Reflection and refraction are shown in Fig. 2-8A. Reflection occurs when a wave strikes a denser reflective medium, as when a light wave strikes a glass mirror. The incident wave (shown as a single ray) strikes the interface between less dense and more dense mediums at a certain *angle of incidence* (a_i), and is reflected at exactly the same angle (now called the *angle of reflection* (a_r). Because these angles are equal, a reflected radio signal can often be traced back to its origin.

Refraction occurs when the incident wave enters a region of different density, and thereby undergoes both a velocity change and a directional change. The amount and direction of the change are determined by the ratio of the densities between the two media. If Zone B is much different from Zone A, then bending is great. In radio

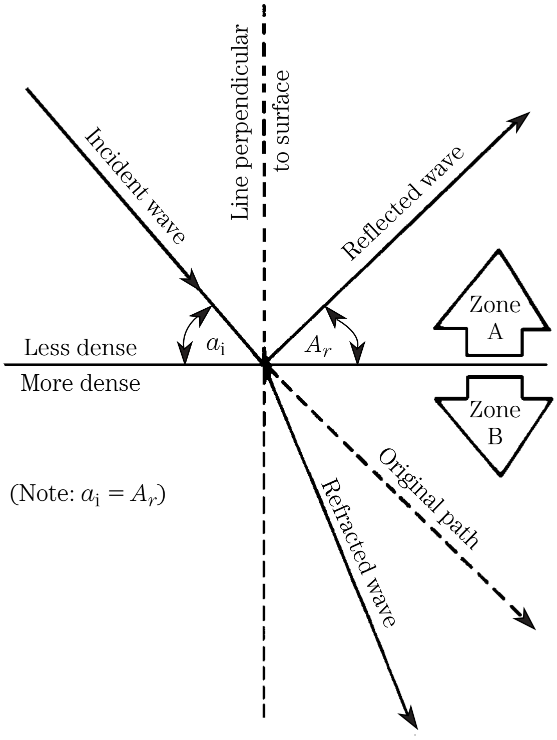

2-8A Reflection and refraction phenomena.

systems, the two media might be different layers of air with different densities. It is possible for both reflection and refraction to occur in the same system. Indeed, more than one form of refraction might be present. These topics will be covered in greater depth shortly.

Diffraction is shown in Fig. 2-8B. In this case, an advancing wavefront encounters an opaque object (e.g., a steel building). The shadow zone behind the building is not simply perpendicular to the wave, but takes on a cone shape as waves bend around the object. The "umbra region" (or diffraction zone) between the shadow zone ("cone of silence") and the direct propagation zone is a region of weak (but not zero) signal strength. In practical situations, signal strength in the cone of silence rarely reaches zero. A certain amount of reflected signals scattered from other sources will fill in the shadow a little bit. The degree of diffraction effect seen in any given case is a function of the wavelength of the signal, the size of the object, and its electromagnetic properties.

Propagation paths

There are four major propagation paths: *surface wave*, *space wave*, *tropospheric,* and *ionospheric*. The ionospheric path is important to medium-wave (MW) and HF propagation, but is not important to VHF, UHF, or microwave propagation. The space wave and surface wave are both *ground waves*, but they behave differently enough to warrant separate consideration. The surface wave travels in direct contact with the earth's surface and it suffers a severe frequency-dependent attenuation caused by absorption into the ground.

The space wave is also a ground-wave phenomenon, but it is radiated from an antenna many wavelengths above the surface. No part of the space wave normally travels in contact with the surface; VHF, UHF, and microwave signals are usually space waves. There are, however, two components of the space wave in many cases: *direct* and *reflected* (see Fig. 2-9).

The tropospheric wave is lumped with the direct space wave in some texts, but it has properties that actually make it different in practical situations. The troposphere

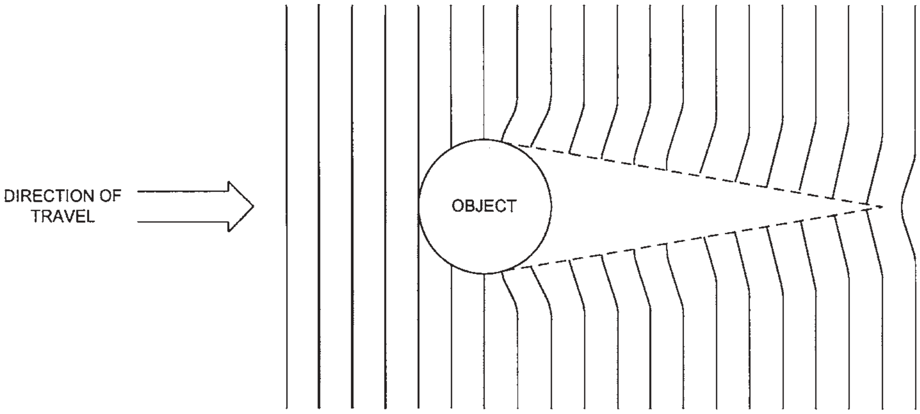

2-8B Diffraction phenomena.

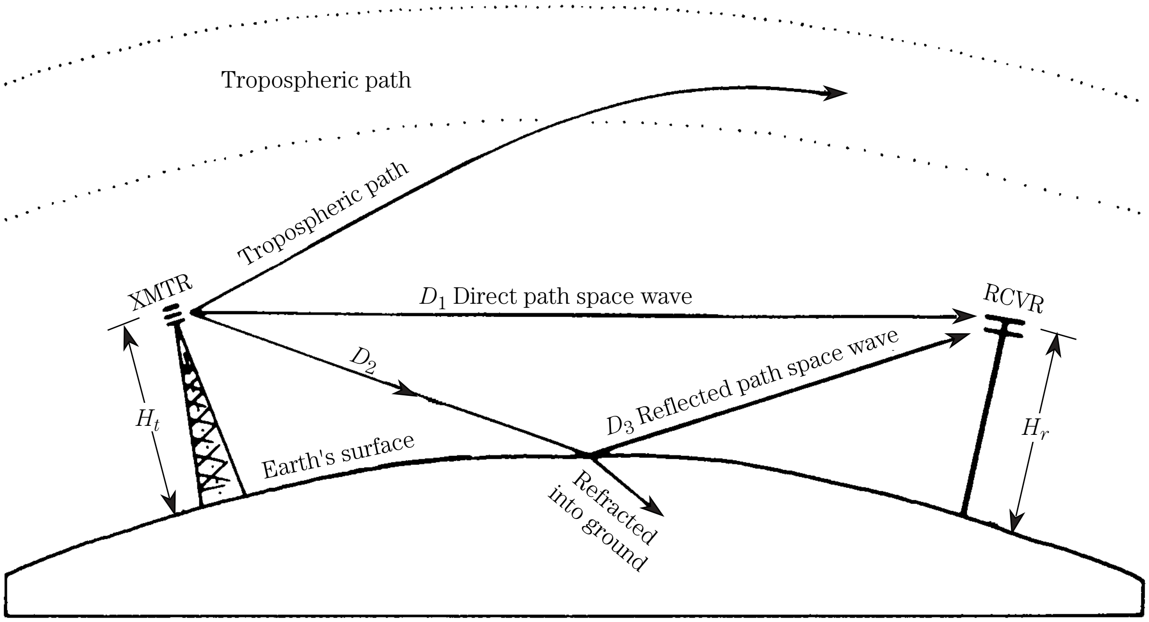

2-9 Space wave propagation.

is the region of earth's atmosphere between the surface and the stratosphere, or about 4 to 7 mi above the surface. Thus, most forms of ground wave propagate in the troposphere. But because certain propagation phenomena (caused mostly by weather conditions) only occur at higher altitudes, tropospheric propagation should be differentiated from other forms of ground wave.

The ionosphere is the region of earth's atmosphere that is above the stratosphere. The peculiar feature of the ionosphere is that molecules of air gas (O_2 and N) can be ionized by stripping away electrons under the influence of solar radiation and certain other sources of energy. In the ionosphere, the air density is so low that ions can travel relatively long distances before recombining with oppositely charged ions to form electrically neutral atoms. As a result, the ionosphere remains ionized for long periods of the day—even after sunset. At lower altitudes, however, air density is greater and recombination thus occurs rapidly. At those altitudes, solar ionization diminishes to nearly zero immediately after sunset, or never achieves any significant levels even at local noon.

Ionization and recombination phenomena in the ionosphere add to the noise level experienced at VHF, UHF, and microwave frequencies. The properties of the ionosphere are important to microwave technology because of the noise contribution. In satellite communications, there are some additional transionospheric effects.

Ground-wave propagation

The *ground wave*, naturally enough, travels along the ground, or at least in close proximity to it. There are three basic forms of ground wave: *space wave*, *surface* *wave*, and *tropospheric wave*. The space wave does not actually touch the ground. As a result, space wave attenuation with distance in clear weather is about the same as in free space (except above about 10 GHz, where H_2O and O_2 absorption increases dramatically). Of course, above the VHF region, weather conditions add attenuation not found in outer space.

The surface wave is subject to the same attenuation factors as the space wave, but in addition it also suffers ground losses. These losses are caused by ohmic resistive losses in the conductive earth. In other words, the signal heats up the ground. Surface wave attenuation is a function of frequency, and it increases rapidly as frequency increases. Both of these forms of ground-wave communications are affected by the following factors: wavelength, height of both the receiving and transmitting antennas, distance between the antennas, and the terrain and weather along the transmission path. Figure 2-10 is a nomograph that can be used to calculate the line of sight distances in miles from a knowledge of the receiver and transmitter antenna heights. Similarly, Figs. 2-11A and 2-11B show power attenuation with frequency and distance (Fig. 2-11A) and power attenuation in terms of field intensity (Fig. 2-11B).

Ground-wave communications also suffer another difficulty, especially at VHF, UHF, and microwave frequencies. The space wave is like a surface wave, but it is radiated many wavelengths above the surface. It is made up of two components (see Fig. 2-9 again): the *direct* and *reflected* waves. If both of these components arrive at the receiving antenna, they will add algebraically to either increase or decrease signal strength. There is always a phase shift between the two components because the two signal paths have different lengths (i.e., D_1 is less than $D_2 + D_3$). In addition, there may possibly be a 180° (π radians) phase reversal at the point of reflection (especially if the incident signal is horizontally polarized), as in Fig. 2-12. The following general rules apply in these situations:

- A phase shift of an odd number of half wavelengths causes the components to add, increasing signal strength (*constructive interference*).
- A phase shift of an even number of half wavelengths causes the components to subtract (Fig. 2-12), thus reducing signal strength (*destructive interference*).
- Phase shifts other than half wavelength add or subtract according to relative polarity and amplitude.

You can characterize the loss of signal over path D_1 with a parametric term n that is defined as follows:

$$n = \frac{S_r}{S_f} \qquad \textbf{[2.11]}$$

where

n is the signal loss coefficient
S_r is the signal level at the receiver in the presence of a ground reflection component
S_f is the free-space signal strength over path D_1 if no reflection took place.

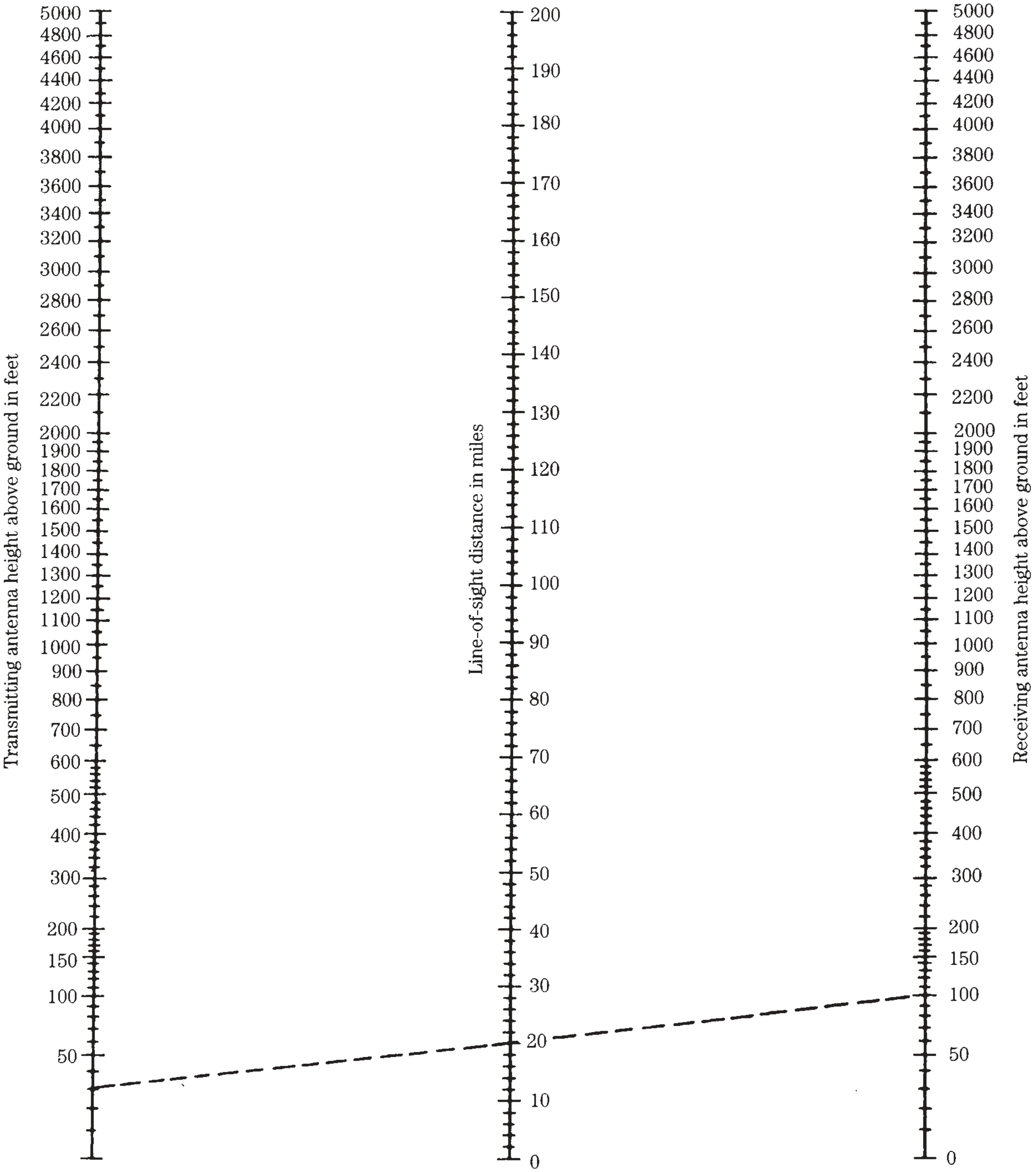

2-10 Nomograph showing the line-of-sight transmission distance as a function of receiving and transmitting antenna heights.

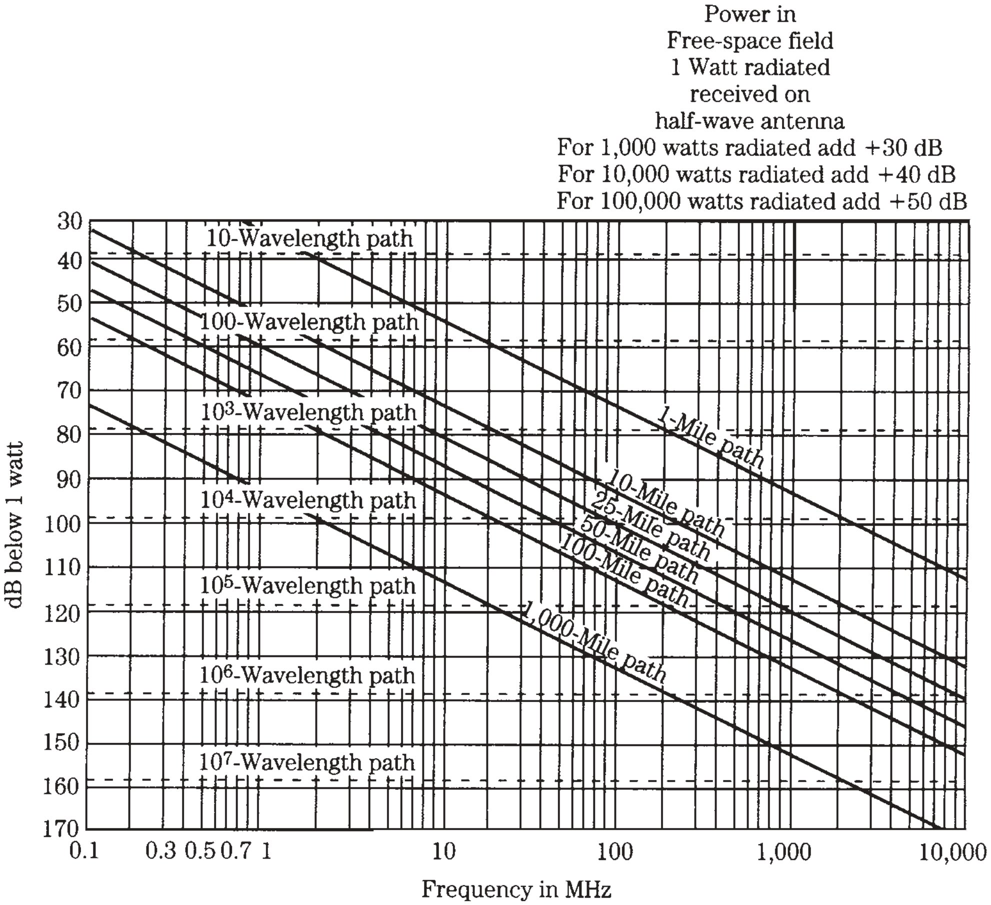

2-11A Power in a free-space field (normalized to 1 W).

You can calculate as follows:

$$n^2 = 4 \sin^2\left(\frac{2\pi h_t h_r}{\lambda D_1}\right) \qquad \textbf{[2.12]}$$

or

$$n = 2 \sin\left(\frac{2\pi h_t h_r}{\lambda D_1}\right) \qquad \textbf{[2.13]}$$

The reflected signal contains both amplitude change and phase change. The phase change is typically π radians (180°). The amplitude change is a function of frequency and the nature of the reflecting surface. The *reflection coefficient* can be characterized as

$$\gamma = pe^{j\phi} \qquad \textbf{[2.14]}$$

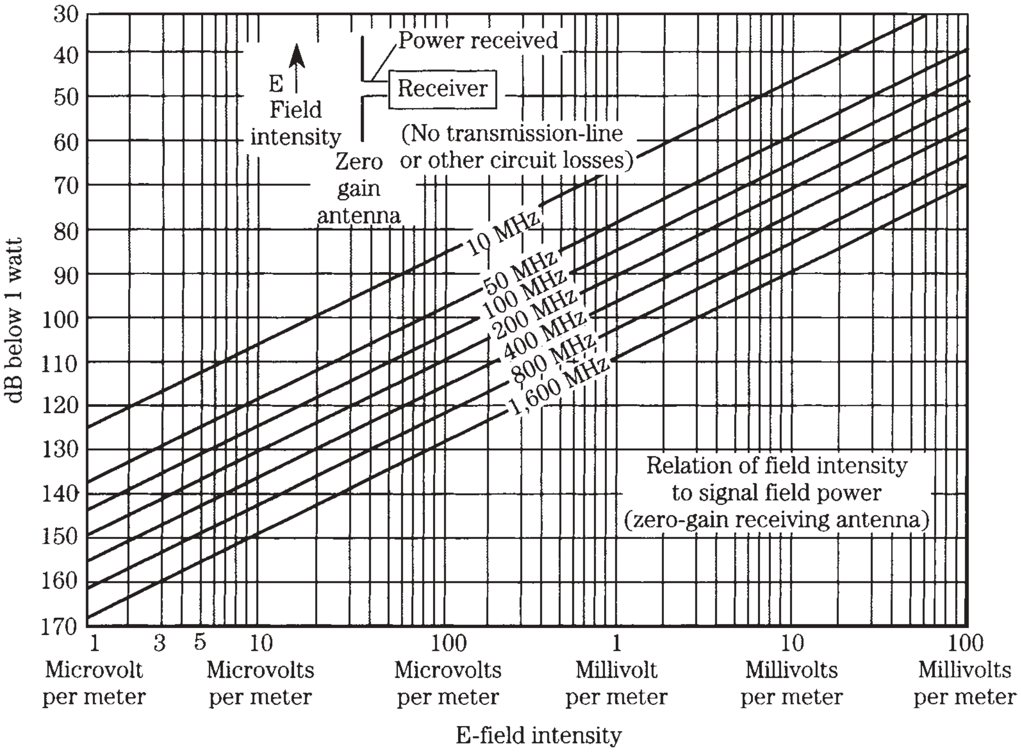

2-11B Relation of field strength to signal field power.

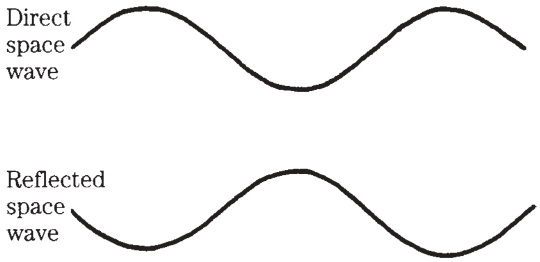

2-12 Out-of-phase waves cancel.

where

γ is the reflection coefficient
p is the amplitude change
ϕ is the phase change
j is the imaginary operator ($\sqrt{-1}$)

For smooth, high-reflectivity surfaces and a horizontally polarized microwave signal that has a shallow angle of incidence, the value of the reflection coefficient is close to –1.

The phase change of the reflected signal at the receiving antenna is at least π radians because of the reflection. Added to this change is an additional phase shift that is a function of the difference in path lengths. This phase shift can be expressed in terms of the two antenna heights and path length.

$$s = \pi + \left(\frac{4\pi h_t h_r}{\lambda D_1}\right) \qquad \textbf{[2.15]}$$

A category of reception problems called *multipath phenomena* exists because of interference between the direct and reflected components of the space wave. The form of multipath phenomena that is, perhaps, most familiar to many readers is ghosting in off-the-air television reception. Some multipath events are transitory in nature (as when an aircraft flies through the transmission path), while others are permanent (as when a large building, or hill, reflects the signal). In mobile communications, multipath phenomena are responsible for reception dead zones and "picket fencing." A *dead zone* exists when destructive interference between direct and reflected (or multiply reflected) waves drastically reduces signal strengths. This problem is most often noticed at VHF and above when the vehicle is stopped; and the solution is to move the antenna one-half wavelength (which at VHF and UHF is a matter of a few inches). *Picket fencing* occurs as a mobile unit moves through successive dead zones and signal enhancement (or normal) zones, and it sounds like a series of short noise bursts.

At VHF, UHF, and microwave frequencies, the space wave is limited to so-called line-of-sight distances. The horizon is theoretically the limit of communications distance, but the radio horizon is actually about 15 percent farther than the optical horizon (Fig. 2-13). This phenomenon is caused by refractive bending in the atmosphere around the curvature of the earth, and it makes the geometry of the situation look as if the earth's radius is $\frac{4}{3}$ the actual radius.

The refraction phenomenon occurs at VHF through microwave frequencies, but not in the visible light spectrum, because water and atmospheric pressure (which relates to the effects of atmospheric gases on microwave signals) become important contributors to the phenomenon. The *K factor* expresses the degree of curvature along any given path, while the *index of refraction n* measures the differential properties between adjacent zones of air.

The *K* factor, also called the *effective earth's radius factor*, is defined according to the relationship of two hypothetical spheres, both centered at the exact center of

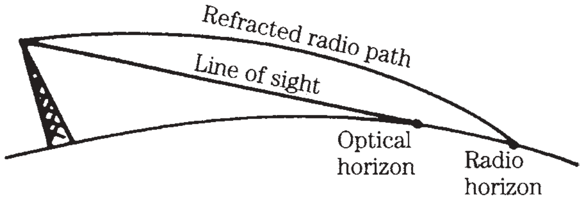

2-13 Phenomenon by which "greater than line of sight" communications occurs.

the earth. The first sphere is the earth's surface, which has a radius r_o (3440 nmi or 6370 km). The second sphere is larger than the first by the curvature of the signal "ray path," and has a radius r. The value of K is approximately

$$K = \frac{r}{r_o} \tag{2.16}$$

A value of $K = 1$ indicates a straight path (Fig. 2-14); a value of $K > 1$ indicates a positively curved path (refraction); and a value of $K < 1$ indicates a negatively curved path (subrefraction). The actual value of K varies with local weather conditions, so one can expect variation not only between locations, but also seasonally. In the arctic regions, K varies approximately over the range 1.2 to 1.34. In the "lower 48" states of the United States, K varies from 1.25 to 1.9 during the summer months (especially in the south and southeast), and from 1.25 to 1.45 in the winter months. The index of refraction n can be defined in two ways, depending on the situation. When a signal passes across boundaries between adjacent regions of distinctly different properties (as occurs in temperature inversions, etc.), the index of refraction is the ratio of the signal velocities in the two regions. In a homogeneous region n can be expressed as the ratio of the free-space velocity c to the actual velocity in the atmosphere V:

$$n = \frac{c}{V} \tag{2.17}$$

At the surface, near sea level, under standard temperature and pressure conditions, the value of n is approximately 1.0003, and in homogeneous atmospheres it will decrease by 4×10^{-8} per mile of altitude. The units of n are a bit cumbersome in equations, so the UHF/microwave communities tend to use a derivative parameter, N, called the *refractivity of the atmosphere*:

$$N = (n - 1) \times 10^6 \tag{2.18}$$

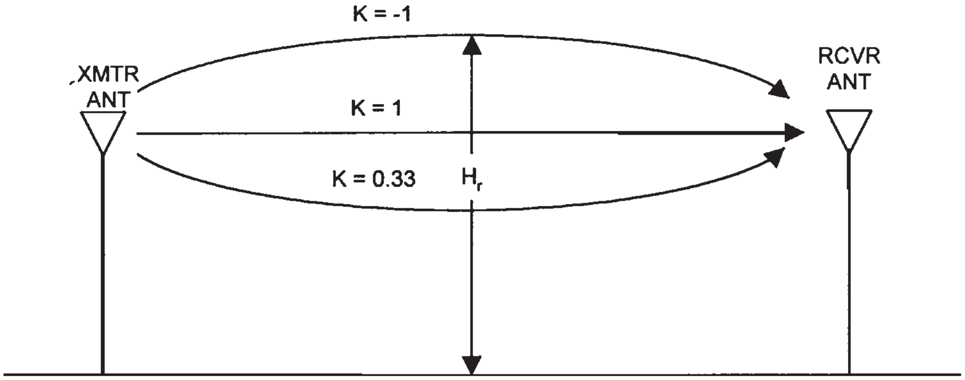

2-14 Refraction changes path length.

The value of N tends to vary from about 280 to 320, and both n and N vary with altitude. In nonhomogeneous atmospheres (the usual case), these parameters will vary approximately linearly for several tenths of a kilometer. All but a few microwave relay systems can assume an approximately linear reduction of n and N with increasing altitude, although airborne radios and radars cannot. There are two methods for calculating N:

$$N = \left(\frac{77.6}{T}\right)\left(P + \frac{4810 e_s H_{\text{rel}}}{T}\right) \qquad \textbf{[2.19]}$$

and

$$N = \frac{77.6}{T}\left(\frac{3.73 \times 10^5 e_s}{T^2}\right) \qquad \textbf{[2.20]}$$

where

P is the atmospheric pressure in millibars (1 torr = 1.3332 mbar)
T is temperature in kelvins
e_s is saturation vapor pressure of atmospheric water in millibars
H_{rel} is the relative humidity expressed as a decimal fraction (rather than a percentage)

Ray path curvature (K) can be expressed as a function of either n or N, provided that the assumption of a linear gradient d_n/d_h holds true:

$$K = \frac{1}{\left(1 + \dfrac{r_o d_n}{d_h}\right)} \qquad \textbf{[2.21]}$$

or

$$K = \frac{1}{\left(1 + \dfrac{d_N/d_h}{157}\right)} \qquad \textbf{[2.22]}$$

For the near-surface region, where d_n/d_h varies at about 3.9×10^{-8} m, the value of K is 1.33. For most terrestrial microwave paths, this value ($K = \frac{4}{3} = 1.33$) is called a *standard refraction*, and is used in calculations *in the absence of additional data*. For regions above the linear zone close to the surface, you can use another expression of refractivity:

$$N_a = N_s e^{-C_e(h_r - h_t)} \qquad \textbf{[2.23]}$$

where

N_1 is the refractivity at 1 km altitude
h_r is the height of the receiver antenna
h_t is the height of the transmit antenna

$C_e = L_n(N_s/N_1)$
N_a is the refractivity at altitude
N_s is the refractivity at the earth's surface

For simple models close to the surface, you can use the geometry shown in Fig. 2-15. Distance d is a curved path along the surface of the earth. But because the earth's radius r_o is about 4000 statute miles, and it is thus very much larger than a practical antenna height h, you can simplify the model in Fig. 2-15A to that in Fig. 2-15B. The underlying assumption is that the earth has a radio radius equal to about $4/3$ ($K = 1.33$) of the actual physical radius of its surface.

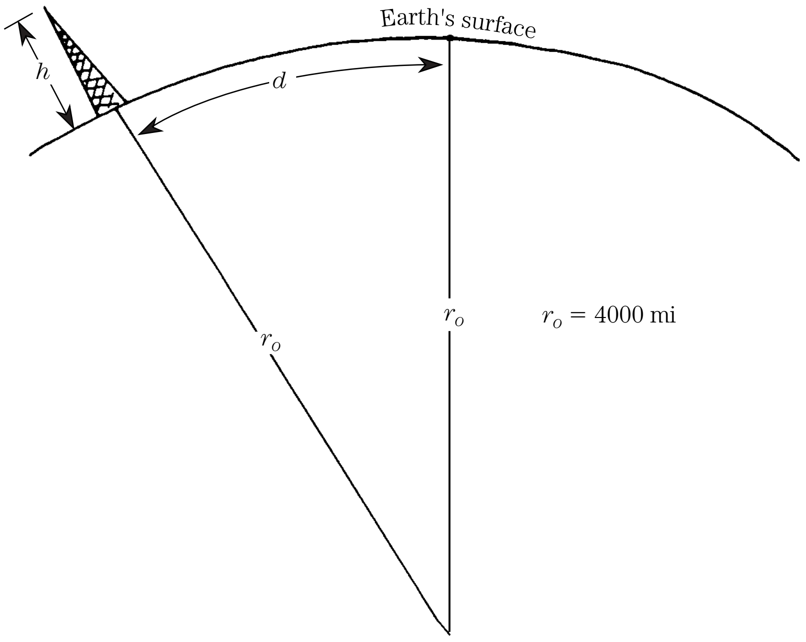

2-15A Geometry for calculating radio distances.

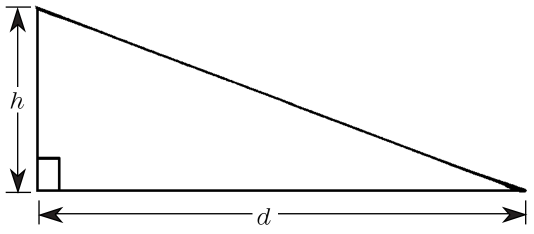

2-15B Simplified geometry.

The value of distance d is found from the expression

$$d = \sqrt{2r_c h_{\mathrm{ft}}} \qquad \textbf{[2.24]}$$

where

d is the distance to the radio horizon in statute miles
r_o is the normalized radius of the earth
h_{ft} is the antenna height in feet

Accounting for all constant factors, the expression reduces to

$$d = 1.414\sqrt{h_{\mathrm{ft}}} \qquad \textbf{[2.25]}$$

all factors being the same as defined above.

Example 2-1 A radio tower has a UHF radio antenna that is mounted 150 feet above the surface of the earth. Calculate the radio horizon (in statute miles) for this system.

Solution:

$$\begin{aligned} d &= 1.414\ (h_t)t \\ &= (1.414)(150\ \mathrm{ft})^{1/2} \\ &= (1.414)(12.25) \\ &= 17.32\ \mathrm{mi} \end{aligned}$$

For other units of measurement:

$$d_{\mathrm{nmi}} = 1.23\sqrt{h_{\mathrm{ft}}}\quad \mathrm{nmi} \qquad \textbf{[2.26]}$$

and

$$d_{\mathrm{km}} = 1.30\sqrt{h_{\mathrm{ft}}}\quad \mathrm{km} \qquad \textbf{[2.27]}$$

Surface-wave communications The surface wave travels in direct contact with the earth's surface, and it suffers a severe frequency-dependent attenuation because of absorption by the ground.

The surface wave extends to considerable heights above the ground level, although its intensity drops off rapidly at the upper end. The surface wave is subject to the same attenuation factors as the space wave, but in addition, it also suffers ground losses. These losses are caused by ohmic resistive losses in the conductive earth, and to the dielectric properties of the earth. In other words, the signal heats up the ground. Horizontally polarized waves are not often used for surface wave communications because the earth tends to short-circuit the E-field component. On vertically polarized waves, however, the earth offers electrical resistance to the E-field and returns currents to following waves (Fig. 2-16). The conductivity of the soil determines how much energy is returned. Table 2-1 shows the typical conductivity values for several different forms of soil.

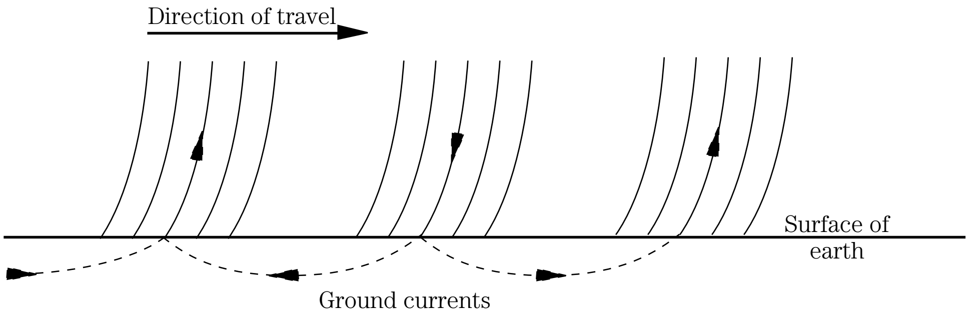

2-16 Distortion of vertically polarized electric field by lossy ground resistance.

Table 2-1. Sample soil conductivity values

Type of soil	**Dielectric constant**	**Conductivity (siemens/meter)**	**Relative quality**
Salt water	81	5	Best
Fresh water	80	0.001	Very poor
Pastoral hills	14–20	0.03–0.01	Very good
Marshy, wooded	12	0.0075	Average/poor
Rocky hills	12–14	10	Poor
Sandy	10	0.002	Poor
Cities	3–5	0.001	Very poor

The wavefront of a surface wave is tilted because of the losses in the ground that tend to retard the wavefront's bottom (also in Fig. 2-16). The tilt angle is a function of the frequency, as shown in Table 2-2.

Table 2-2. Tilt angle as a function of frequency

Frequency (kHz)	**Tilt angle ratio**	**Earth/seawater (degrees)**
20	2.07	4.3/2.08
200	104	13.4/0.13
2,000	64	32.3/0.5
20,000	25	35/1.38

Surface-wave attenuation is a function of frequency, and increases rapidly as frequency increases. For both forms of the ground wave, reception is affected by these factors: *wavelength, height* of both the receiving and transmitting antennas, *distance* between antennas, and both *terrain* and *weather* along the transmission path. In addition, the surface wave is affected by the ground losses in Table 2-1. Because of the ground loss effects, the surface wave is attenuated at a much faster rate than the inverse square law.

Ground-wave propagation frequency effects The frequency of a radio signal in large measure determines its surface-wave behavior. In the very low-frequency (VLF) band (<300 kHz), ground losses are small for vertically polarized signals, so medium-distance communications (up to several hundred miles) are possible. In the medium-wave band (300 to 3000 kHz, including the AM broadcast band), distances of 1000 mi are possible with regularity—especially at night. In the high-frequency (HF) band, ground losses are more considerable, so the surface-wave distance reduces drastically. It is possible, in the upper end of the HF band (3000 to 30,000 kHz) for surface-wave signals to die out within a few dozen miles. This phenomenon is often seen in the 15- and 10-m amateur radio bands, as well as the 11-m (27-MHz) Citizen's Band. Stations only 20 mi apart cannot communicate, but both can talk to a third station across the continent via ionospheric skip. Thus, the two stations close together must have a station more than 2000 mi away relay messages between them.

Tropospheric propagation The troposphere is the portion of the atmosphere between the surface of the earth and the stratosphere (or about 4 to 7 mi above the surface). Some older texts group tropospheric propagation with ground-wave propagation, but modern practice requires separate treatment. The older grouping overlooks certain common propagation phenomena that simply don't happen with space or surface waves.

Refraction is the mechanism for most tropospheric propagation phenomena. The dielectric properties of the air, which are set mostly by the moisture content (Fig. 2-17), are a primary factor in tropospheric refraction. Recall that refraction occurs in both light or radio-wave systems when the wave passes between mediums of differing density. Under that situation, the wave path will bend an amount proportional to the difference in density.

Two general situations are typically found—especially at UHF and microwave frequencies. First, because air density normally decreases with altitude, the top of a beam of radio waves typically travels slightly faster than the lower portion of the beam. As a result, those signals refract a small amount. Such propagation provides slightly longer surface distances than are normally expected from calculating the distance to the radio horizon. This phenomenon is called *simple refraction*, and was discussed in a preceding section.

A special case of refraction called *superrefraction* occurs in areas of the world where warmed land air goes out over a cooler sea (Fig. 2-18). Examples of such areas have deserts that are adjacent to a large body of water: the Gulf of Aden, the southern Mediterranean, and the Pacific Ocean off the coast of Baja California are examples. VHF/UHF/microwave communications up to 200 mi are reported in such areas.

The second form of refraction is weather-related. Called *ducting*, this form of propagation (Fig. 2-19) is actually a special case of superrefraction. Evaporation of sea water causes temperature inversion regions to form in the atmosphere—that is, layered air masses in which the air temperature is greater than in the layers below it. (*Note*: air temperature normally decreases with altitude, but at the boundary with an

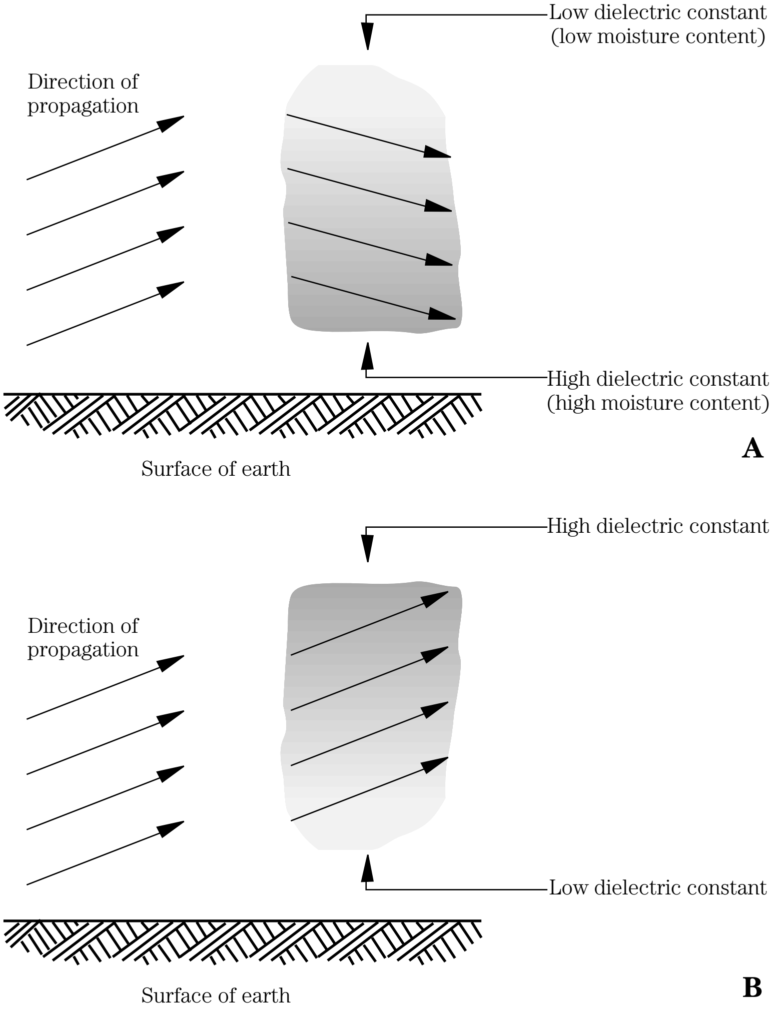
Low dielectric constant
(low moisture content)
Direction of propagation
High dielectric constant
(high moisture content)
Surface of earth
A
High dielectric constant
Direction of propagation
Low dielectric constant
Surface of earth
B

2-17 Refraction in the troposphere.

inversion region, it increases.) The inversion layer forms a “duct” that acts like a waveguide. In Fig. 2-19, the distance D_1 is the normal “radio horizon” distance, and D_2 is the distance over which duct communications can occur.

Ducting allows long-distance communications from lower VHF through microwave frequencies, with 50 MHz being a lower practical limit and 10 GHz being an ill-defined upper limit. Airborne operators of radio, radar, and other electronic equipment can sometimes note ducting at even higher microwave frequencies.

Antenna placement is critical for ducting propagation. Both the receiving and transmitting antennas must be either (1) inside the duct physically (as in airborne cases) or (2) able to propagate at an angle such that the signal gets trapped inside the duct. The latter is a function of antenna radiation angle. Distances up to 2500 mi

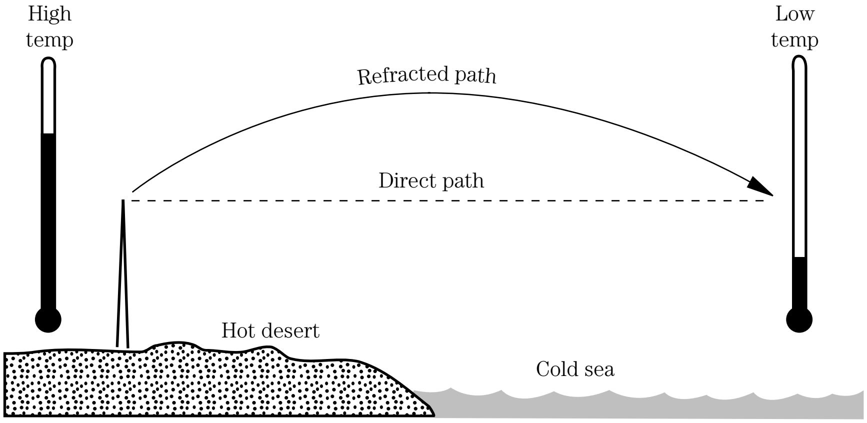
High temp
Low temp
Refracted path
Direct path
Hot desert
Cold sea

2-18 Superrefraction phenomena.

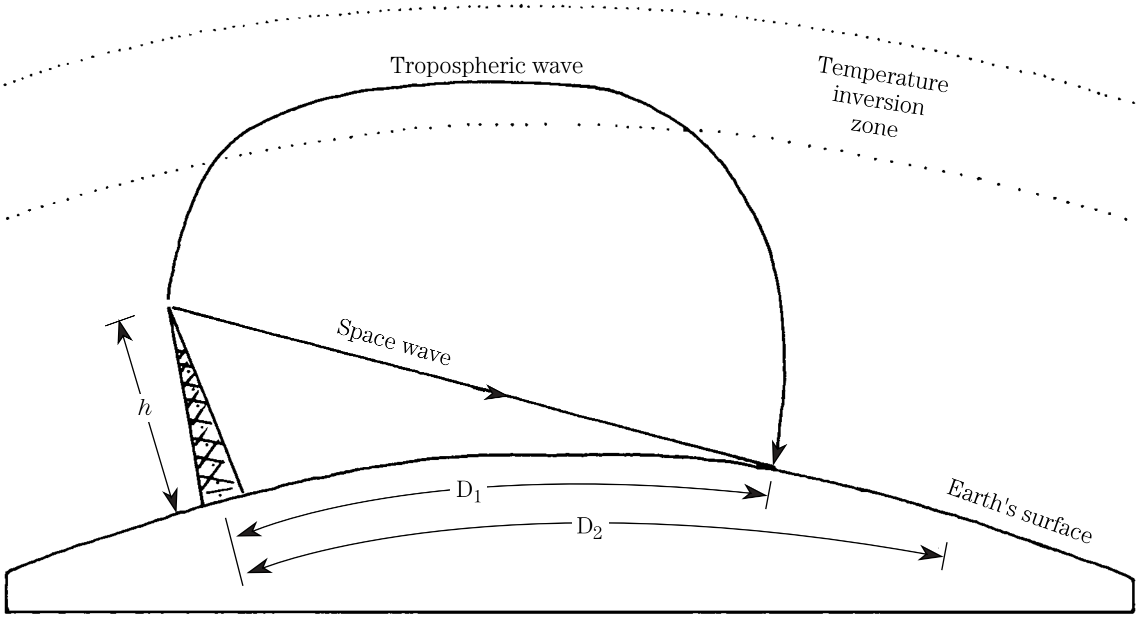
Tropospheric wave
Temperature inversion zone
Space wave
h
D1
D2
Earth's surface

2-19 Ducting phenomenon.

or so are possible through ducting. Certain paths where frequent ducting occurs have been identified: the Great Lakes to the Atlantic seaboard; Newfoundland to the Canary Islands; across the Gulf of Mexico from Florida to Texas; Newfoundland to the Carolinas; California to Hawaii; and Ascension Island to Brazil.

Another condition is noted in the polar regions, where colder air from the land mass flows out over warmer seas (Fig. 2-20). Called *subrefraction*, this phenomena bends EM waves away from the earth's surface—thereby *reducing* the radio horizon by about 30 to 40 percent.

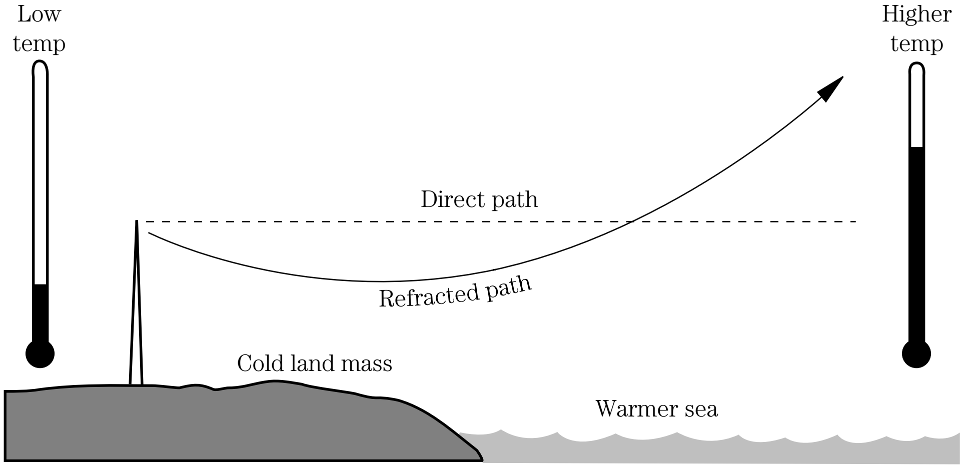

2-20 Subrefraction phenomena.

All tropospheric propagation that depends upon air-mass temperatures and humidity, shows diurnal (i.e., over the course of the day) variation caused by the local rising and setting of the sun. Distant signals may vary 20 dB in strength over a 24-hour period. These tropospheric phenomena explain how TV, FM broadcast, and other VHF signals can propagate great distances, especially along seacoast paths, at some times while being weak or nonexistent at others.

Diffraction phenomena Electromagnetic *waves* diffract when they encounter a radio-opaque object. The degree of diffraction, and the harm it causes, is frequency-related. Above 3 GHz, wavelengths are so small (approximately 10 cm) compared to object sizes that large attenuation of the signal occurs. In addition, beamwidths (a function of antenna size compared with wavelength) tend to be small enough above 3 GHz that blockage of propagation by obstacles is much greater.

Earlier in this chapter, large-scale diffraction around structures (such as buildings) was discussed. The view presented was from above, so it represented the horizontal plane. But there is also a diffraction phenomenon in the vertical plane. Terrain, or man-made objects, intervening in the path between UHF microwave stations (Fig. 2-21A) cause diffraction, and some signal attenuation. There is a minimum clearance required to prevent severe attenuation (up to 20 to 30 dB) from diffraction. Calculation of the required clearance comes from Huygens-Fresnel wave theory.

Consider Fig. 2-21B. A wave source A, which might be a transmitter antenna, transmits a wavefront to a destination C (receiver antenna). At any point along path A-C, you can look at the wavefront as a partial spherical surface (B_1-B_2) on which all wave rays have the same phase. This plane can be called an *isophase plane*. You can assume that the d_n/d_h refraction gradient over the height extent of the wavefront is small enough to be considered negligible.

Using ray tracing we see rays r_a incoming to plane [B_1-B_2], and rays r_b outgoing from plane [B_1-B_2]. The signal seen at C is the algebraic sum of all rays r_b. The signal pattern will have the form of an optical interference pattern with wave cancellation occurring between r_b waves that are a half-wavelength apart on [B_1-B_2]. The ray im-

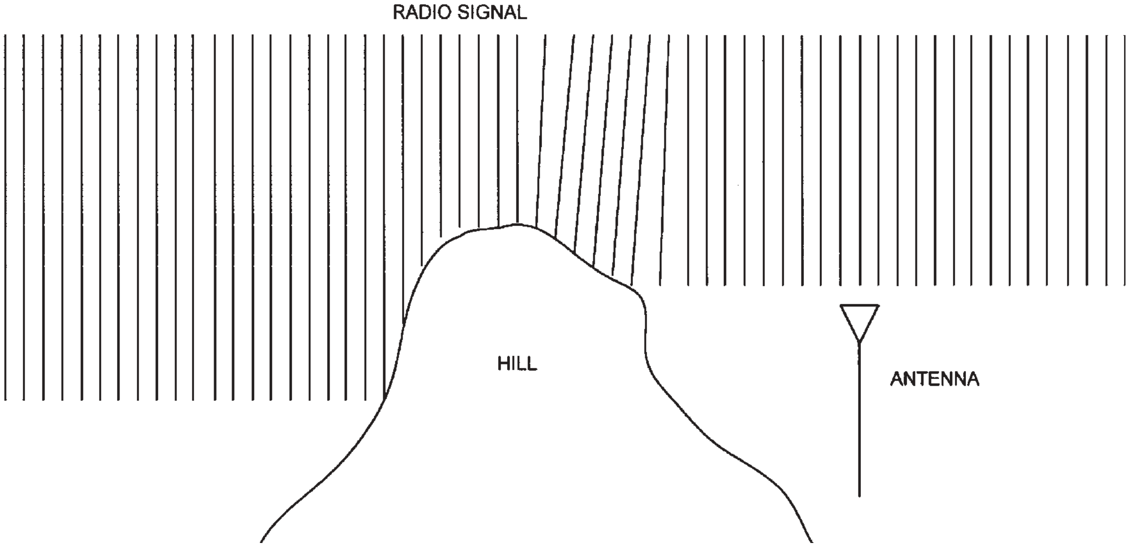

2-21A Terrain masking of VHF and higher-frequency signals.

pact points on plane [B_1-B_2] form radii R_n called *Fresnel zones*. The lengths of the radii are a function of the frequency and the ratio of the distances D_1 and D_2 (see Fig. 2-21B). The general expression is

$$R_n = M\sqrt{\frac{N}{F_{\mathrm{GHz}}}\left(\frac{D_1D_2}{D_1 + D_2}\right)} \qquad \textbf{[2.28]}$$

where

R_m is the radius of the nth Fresnel zone
F_{GHz} is the frequency in GHz
D_1 is the distance from the source to plane B_1-B_2
D_2 is the distance from destination to plane B_1-B_2
N is an integer (1, 2, 3, . . .)
M a constant of proportionality equal to

17.3 if R_m is in meters and D_1, D_2 are in kilometers and

72.1 if R_m is in feet and D_1, D_2 are in statute miles

If you first calculate the radius of the critical first Fresnel zone (R_1), then you can calculate the nth Fresnel zone from

$$R_n = R_1\sqrt{n} \qquad \textbf{[2.29]}$$

Example 2-2 Calculate the radius of the first Fresnel zone for a 2.5-GHz signal at a point that is 12 km from the source and 18 km from the destination.

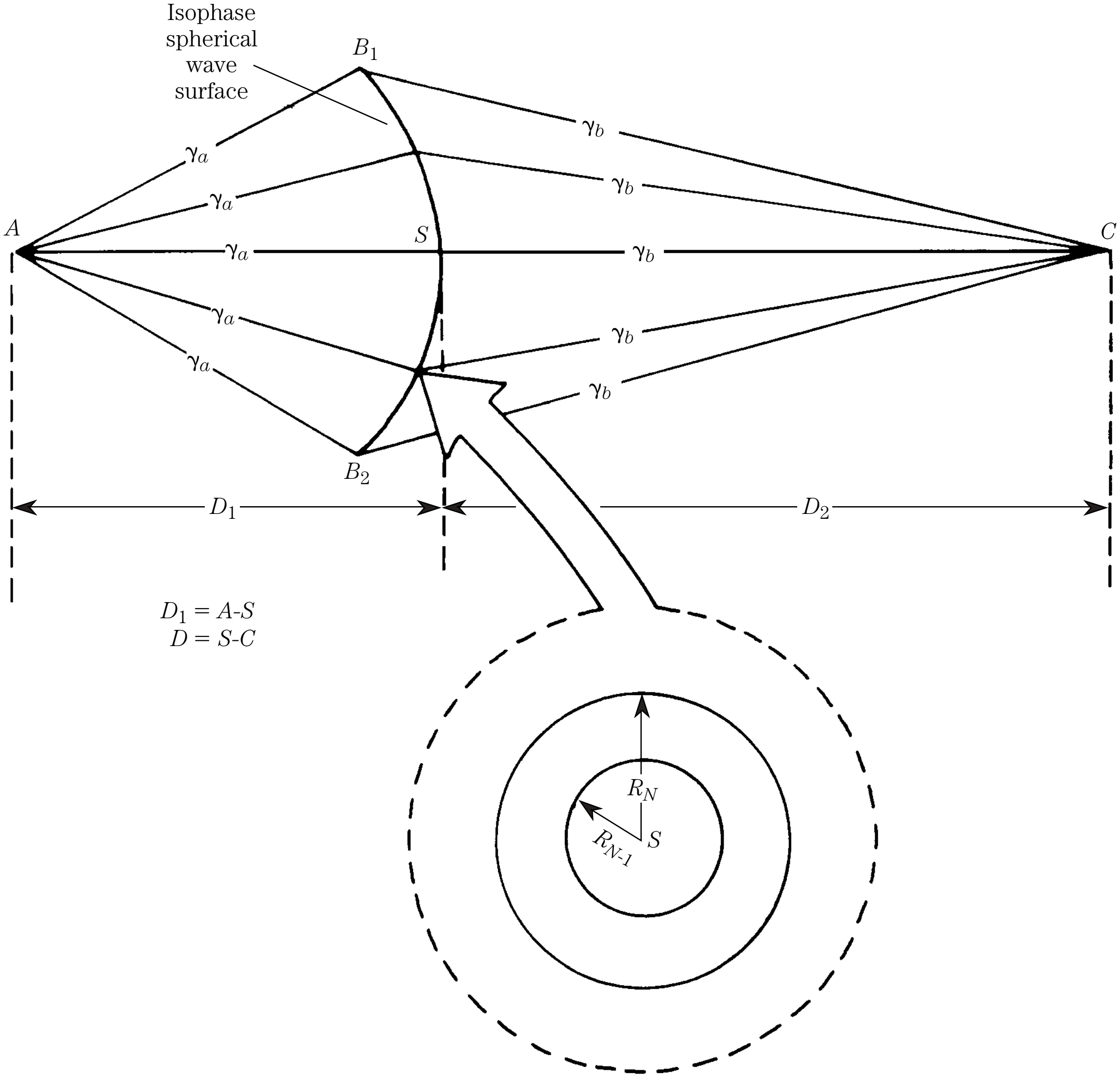
Isophase spherical wave surface
B_1
γ_a
γ_b
A
S
C
B_2
D_1
D_2
$D_1 = A\text{-}S$
$D = S\text{-}C$
R_N
$R_{N\text{-}1}$
S
(Rotated view)

2-21B Fresnel zone geometry.

Solution:

$$R_1 = M\sqrt{\frac{N}{F_{\text{GHz}}}\left(\frac{D_1 D_2}{D_1 + D_2}\right)}$$

$$R_1 = (17.3)\ \sqrt{\frac{1}{2.5}\left(\frac{12 \times 18}{12 + 18}\right)}$$

$$R_t = (17.3)\sqrt{(0.4)\ (7.2)}$$

$$R_1 = (17.3)\ (2.88) = 29.4\ \text{m}$$

For most terrestrial microwave systems an obstacle clearance of $0.6R_1$ is required to prevent diffraction attenuation under most normal conditions. However, there are conditions in which the clearance zone should be more than one Fresnel zone.

Fading mechanisms Fading is defined as a reduction in amplitude caused by reduced received signal power, changes in phase or polarization, wave cancellation, or other related problems that are not caused by a change in the output power level or other parameters associated with either the transmitter or its antenna system. You would not ordinarily think that line-of-sight radio relay links would experience fading, but that is not true. Fading does, in fact, occur, and it can reach levels of 30 dB in some cases (20 dB is relatively common). In addition, fading phenomena in the VHF-and-up range can last several hours, with some periods being reported of several days in duration (although very rare). There are several mechanisms of fading, and these will be dealt with in this section. HF fading caused by ionospheric mechanisms will be covered later.

Any or all of the mechanisms shown in Fig. 2-22 can occur in a given system. In all cases, two or more signals arrive at the receiver antenna (R_x). Ray A represents the direct path signal that is, ideally, the only signal to reach the destination. But it is also possible that a signal, in an elevated layer or other atmospheric anomaly, will cause refraction or subrefraction of the wave creating a second component, B. If this second signal arrives out of phase with A, then fading will occur (signal reinforcement—

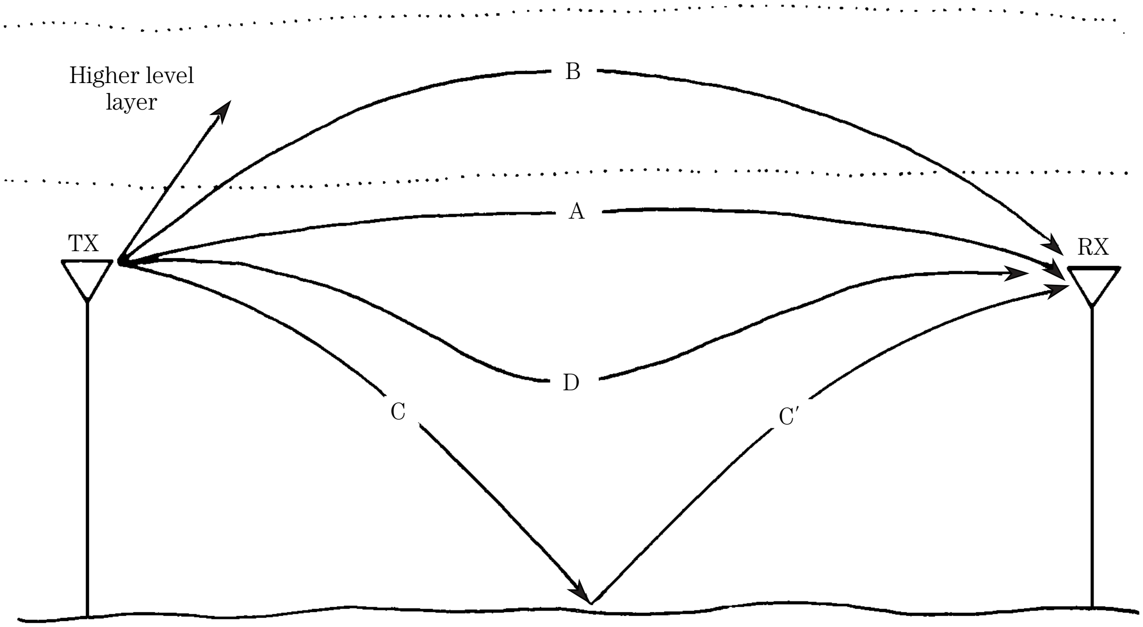

2-22 Multiple paths for signal to take between transmitter and receiver.

in phase—can also occur). It is also possible to see subrefraction fading, as in D. The classical multipath situation represented by ray C and its reflected component C' is also a source of fading.

These mechanisms are frequency-sensitive, so a possible countermeasure is to use *frequency diversity*. Hopping over a 5 percent frequency change will help eliminate fading in many cases. In cases where either system constraints, or local spectrum usage prevents a 5 percent delta (change), then try for at least 2 or 3 percent.

Over ocean areas or other large bodies of water there is a possibility of encountering fair weather surface ducting as a cause of fading. These ducts form in the mid latitudes, starting about 2 to 3 km from shore, up to heights of 10 to 20 mi; wind velocities are found in the 10 to 60 km/h range. The cause of the problem is a combination of power fading, due to the presence of the duct, and surface reflections (see Fig. 2-23). Power fading alone can occur when there is a superrefractive duct elevated above the surface. The duct has a tendency to act as a waveguide and focus the signal (Fig. 2-24). Although the duct shown is superrefractive, it is also possible to have a subrefractive duct.

Attenuation in weather Microwave communications above about 10 GHz suffer an increasingly severe attenuation because of water vapor and oxygen in the atmosphere. Figure 2-25 shows the standard attenuation in dB/km for microwave frequencies. Note that there are several strong peaks in an ever-increasing curve. Setting a system frequency in these regions will cause poor communications or will require a combination of more transmit power, better receiver sensitivity, and better antennas on receiving and transmitting locations. The curves shown in Fig. 2-25 assume certain standardizing conditions. Rain and other weather conditions can severely increase the attenuation of signals. In addition to attenuation, radar exhibits severe clutter problems when signals backscatter from rain cells.

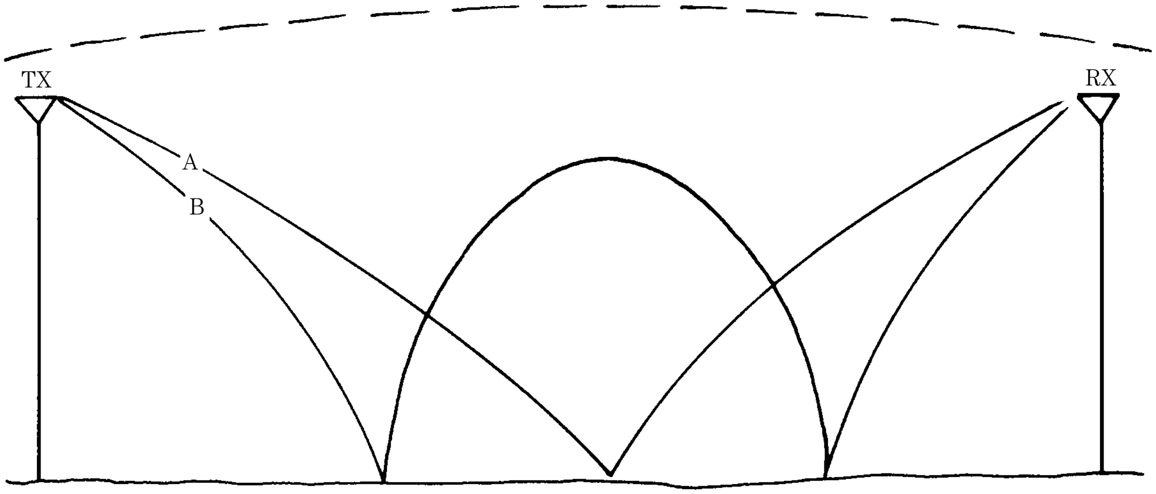

2-23 Multihop interference.

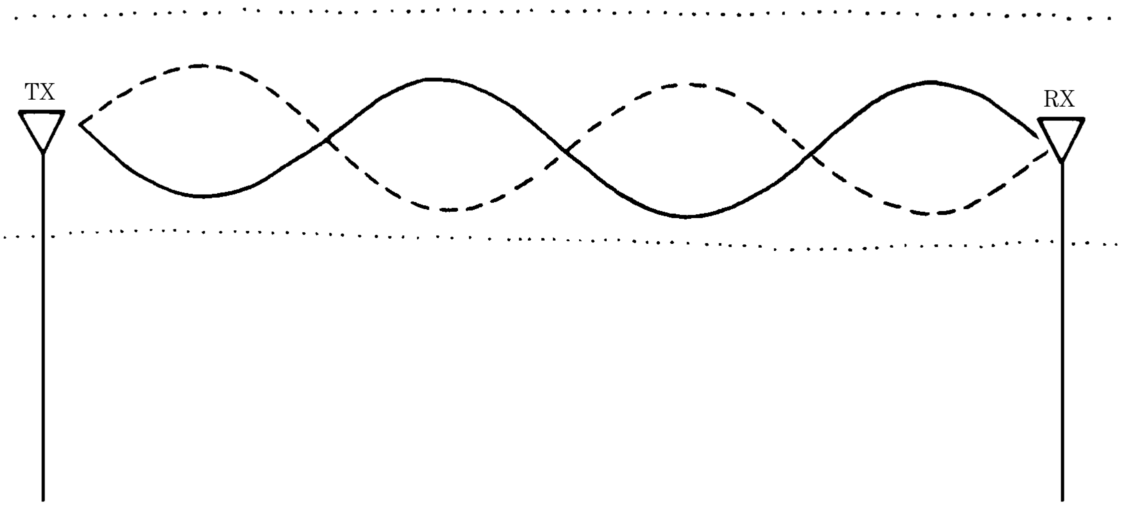

2-24 Wave interference.

Ionospheric propagation

Now let's turn your attention to the phenomenon of skip communications. Ionospheric propagation is responsible for the ability to do intercontinental broadcasting and communications. Long-distance radio transmission is carried out on the high-frequency (HF) bands (3 to 30 MHz), also called the *shortwave* bands. These frequencies are used because of the phenomenon called *skip*. Under this type of propagation, the earth's ionosphere acts as though it is a "radio mirror." Although the actual phenomenon is based on refraction (not reflection, as is frequently believed), the appearance to the casual observer is that shortwave and low-VHF radio signals are reflected from the ionosphere. The actual situation is a little different.

The key lies in the fact that this radio mirror is produced by ionization of the upper atmosphere. The upper portion of the atmosphere is called the *ionosphere* because it tends to be easily ionized by solar and cosmic radiation phenomena. The reason for the ease with which that region (30 to 300 mi above the surface) ionizes, is that the air density is very low. Energy from the sun strips away electrons from the outer shells of oxygen and nitrogen molecules. The electrons become negative ions, while the remaining portion of the atom forms positive ions. Because the air is so rarified at those altitudes, those ions can travel great distances before recombining to form electrically neutral atoms again. As a result, the average ionization level remains high in that region.

Several sources of energy will cause ionization of the upper atomosphere. Cosmic radiation from outer space causes some degree of ionization, but the majority of ionization is caused by solar energy.

The role of cosmic radiation was noticed during World War II, when British radar operators discovered that the distance at which their equipment could detect German aircraft was dependent upon whether or not the Milky Way was above the horizon. Intergalactic radiation raised the background microwave noise level, thereby adversely affecting the signal-to-noise ratio.

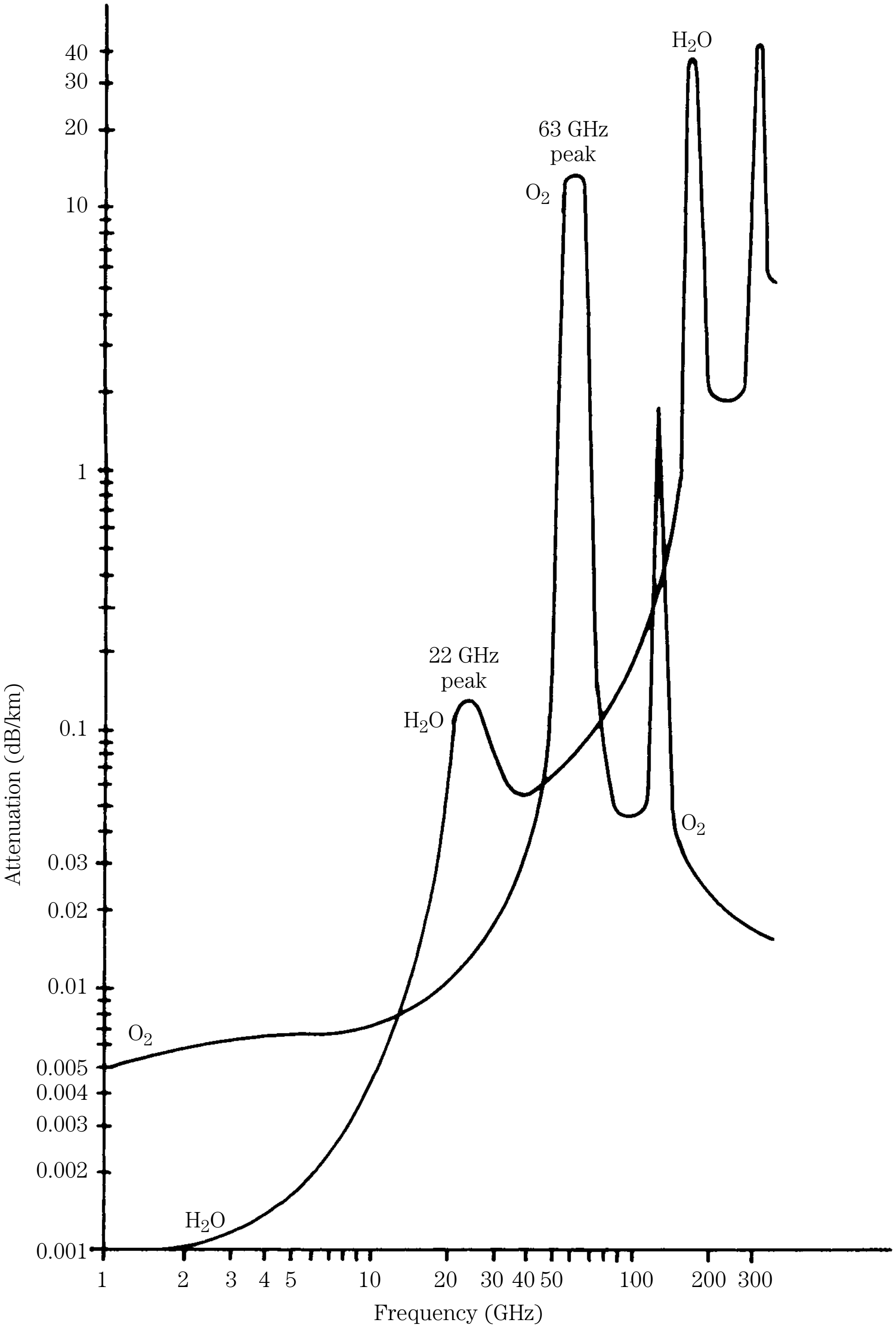

2-25 Atmospheric absorption of radio signals at microwave frequencies.

Events on the surface of the sun sometimes cause the radio mirror to seem to be almost perfect, and this situation makes spectacular propagation possible. At other times, however, solar disturbances (Fig. 2-26A) disrupt radio communications for days at a time.

There are two principal forms of solar energy that affect shortwave communications: *electromagnetic radiation* and *charged solar particles*. Most of the radiation is above the visible spectrum, in the ultraviolet and x-ray/gamma-ray region of the spectrum. Because electromagnetic radiation travels at the speed of light, solar events that release radiation cause changes to the ionosphere about 8 minutes later. Charged particles, on the other hand, having a finite mass must travel at a considerably slower velocity. They require two or three days to reach earth.

Various sources of both radiation and particles exist on the sun. Solar flares can release huge amounts of both radiation and particles. These events are unpredictable and sporadic. Solar radiation also varies over an approximately 27-day period, which is the rotational period of the sun. The same source of radiation will face the earth once every 27 days, and so events tend to be somewhat repetitive.

Solar and galactic noise affect the reception of weak signals. Solar noise can also affect radio propagation and act as a harbinger of changes in propagation patterns. Solar noise can be demonstrated by using an ordinary radio receiver and a directional antenna, preferably operating in the VHF/UHF regions of the spectrum (150 to 152 MHz frequently is used). Aim the antenna at the sun on the horizon at either sunset or sunrise. A dramatic change in background noise will be noted as the sun slides across the horizon.

2-26A Solar event that can affect radio propagation on earth.

Sunspots A principal source of solar radiation, especially the periodic forms, is sunspots (Fig. 2-26B). Sunspots can be as large as 70,000 to 80,000 miles in diameter, and generally occur in clusters. The number of sunspots varies over a period of approximately 11 years, although the actual periods since 1750 (when records were first kept) have varied from 9 to 14 years. The sunspot number is reported daily as the statistically massaged *Zurich Smoothed Sunspot Number*, or *Wolf Number*. The number of sunspots greatly affects radio propagation via the ionosphere. The low was in the range of 60 (in 1907), and the high was about 200 (1958).

Another indicator of ionospheric propagation potential is the *solar flux index* (SFI). This measure is taken in the microwave region (wavelength of 10.2 cm, or 2.8 GHz), at 1700 UTC at Ottawa, Canada. The SFI is reported by the National Institute for Standards and Technology (NIST) radio stations WWV (Fort Collins, CO) and WWVH (Maui, Hawaii).

The ionosphere offers different properties that affect radio propagation at different times. Variations occur not only over the 11 year sunspot cycle, but also diurnally and seasonally. Obviously, if the sun affects propagation in a significant way, then differences between nighttime and daytime, and between summer and winter, must cause variations in the propagation phenomena observed.

The ionosphere is divided, for purposes of radio propagation studies, into various layers that have somewhat different properties. These layers are only well defined

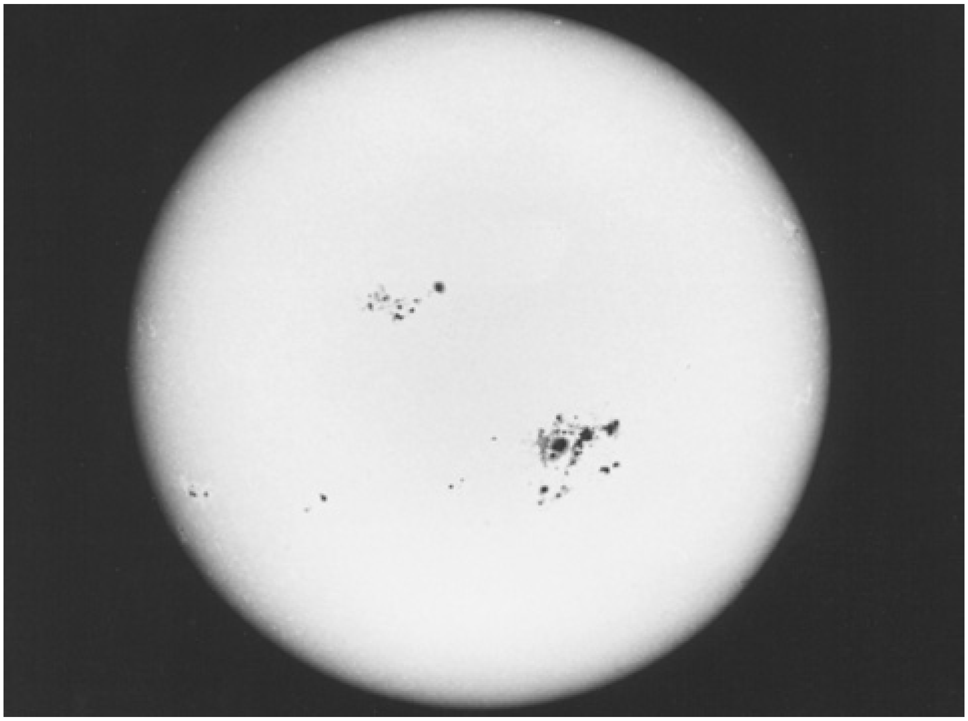

2-26B Sunspots.

in textbooks. However, even there you find a variation in the precise altitudes of the layers above the earth's surface. In addition, the real physical situation is such that layers don't have sharply defined boundaries, but rather fade one into another instead. Thus, the division into layers is somewhat arbitrary. These layers (Fig. 2-27) are designated D, E, and F (with F being further subdivided into F1 and F2 sublayers).

D layer The D layer is the lowest layer in the ionosphere, and exists from approximately 30 to 50 mi above the surface. This layer is not ionized as much as the higher layers, because all forms of solar energy that cause ionization are severely attenuated by the higher layers above the D layer. The reason for this is that the D layer is much more dense than the E and F layers, and that density of air molecules allows ions to recombine to form electroneutral atoms very quickly.

The extent of D layer ionization is proportional to the elevation of the sun, so it will achieve maximum intensity at midday. The D layer exists mostly during the warmer months of the year because of both the greater height of the sun above the horizon and the longer hours of daylight. The D layer almost completely disappears after local sunset. Some observers have reported sporadic incidents of D layer activity for a considerable time past sunset. The D layer exhibits a large amount of absorption of medium-wave and shortwave signals (to such an extent that signals below 4 to 6 MHz are completely absorbed by the D layer).

E layer The E layer exists from approximately 50 to 70 mi above the earth's surface, and it is considered the lowest region of the ionosphere that is important to radio communications. Like the D layer, this region is ionized only during the daylight hours, with ionization levels peaking at midday. The ionization level drops off sharply in the late afternoon, and almost completely disappears after local sunset.

During most of the year, the E layer is absorptive and it will not reflect radio signals. During the summer months, however, E layer propagation does occur. A phenomenon called "short skip" (i.e., less than 100 mi for medium wave and 1000 mi for shortwave signals) occurs in the E layer during the summer months, and in equatorial regions at other times.

A propagation phenomenon associated with the E layer is called *sporadic E propagation*. This phenomenon is caused by scattered zones of intense ionization in the E layer region of the ionosphere. The sporadic E phenomenon varies seasonally, and it is believed to be caused by the bombardment of solar particles. Sporadic E propagation affects the upper HF and lower VHF region. It is observed most frequently in the lower VHF spectrum (50 to 150 MHz), but it is also sometimes observed at higher frequencies. The VHF bands occasionally experience sporadic E propagation. Skip distances on VHF can reach 500 to 1500 miles on one hop—especially in the lower VHF region (including the 6-m band).

F layer The F layer of the ionosphere is the region that is the principal cause of long-distance shortwave communications. This layer is located from about 100 to 300 mi above the earth's surface. Unlike the lower layers, the air density in the F layer is low enough that ionization levels remain high all day, and decay slowly after local sunset. Minimum levels are reached just prior to local sunrise. Propagation in the F layer is capable of skip distances up to 2500 mi on a single hop. During the day there are actually two identifiable and distinct sublayers in the F layer region, and these are designated the Fl and F2 layers. The F1 layer is found approximately

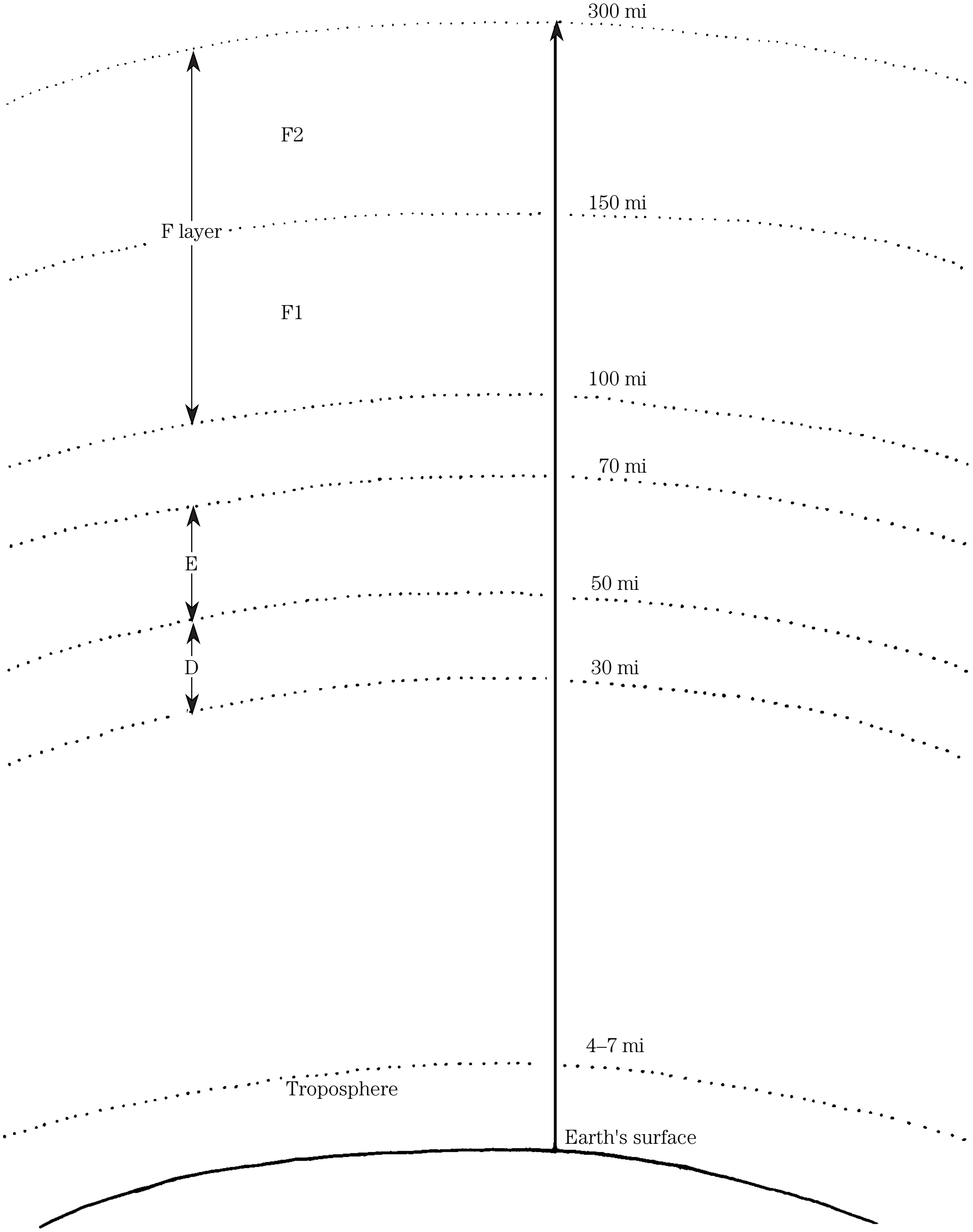
300 mi
F2
150 mi
F layer
F1
100 mi
70 mi
E
50 mi
D
30 mi
4–7 mi
Troposphere
Earth's surface

2-27 Classification of the earth’s atmosphere for radio propagation.

100 to 150 miles above the earth's surface, and the F2 layer is above the F1 extending up to the 270- to 300-mi limit. Beginning at local sundown, however, the lower regions of the F1 layer begin to deionize because of recombination of positive and negative ions. At some time after local sunset, the F1 and F2 layers have effectively merged to become a single reduced layer beginning at about 175 mi.

The height and degree of ionization of the F2 layer varies over the course of the day, with the season of the year, and with the 27-day sunspot cycle. The F2 layer begins to form shortly after local sunrise and reaches maximum shortly before noon. During the afternoon, the F2 layer ionization begins to decay in an exponential manner until, for purposes of radio propagation, it disappears sometime after local sunset. There is some evidence that ionization in the F layer does not completely disappear, but its importance to HF radio communication does disappear.

Measures of ionospheric propagation

There are several different measures by which the ionosphere is characterized at any given time. These measures are used in making predictions of radio activity and long-distance propagation.

The *critical frequency* and *maximum usable frequency* (MUF) are indices that tell us something of the state of ionization and communications ability. These frequencies increase rapidly after sunrise and international communications usually begin within 30 minutes.

Critical frequency F_c The *critical frequency*, designated by F_c, is the highest frequency that can be reflected when a signal strikes the ionosphere as a vertical (90° with respect to the surface) incident wave. The critical frequency is determined from an *ionogram*, which is a cathode-ray tube (CRT) oscilloscope display of the height of the ionosphere as a function of frequency. The ionogram is made by firing a pulse vertically (Fig. 2-28) at the ionosphere from the transmitting station. The critical frequency is that frequency that is just sufffficient to be reflected back to the transmitter site. Values of F_c can be as low as 3 MHz during the nighttime hours, and as high as 10 to 15 MHz during the day.

Virtual height Radio waves are refracted in the ionosphere, and those above a certain critical frequency are refracted so much that they return to earth. Such waves appear to have been reflected from an invisible radio "mirror." An observer on the earth's surface could easily assume the existence of such a mirror by noting the return of the "reflected" signal. The height of this apparent "mirror" is called the *virtual height* of the ionosphere. Figure 2-29 shows the refraction phenomenon by which a radio wave is bent sufficiently to return to earth. Virtual height is determined by measuring the time interval required for an ionosonde pulse (similar to that used to measure critical frequency) to travel between the transmitting station and a receiving station (Fig. 2-30). A radio signal travels at a velocity of 300,000,000 m/s (the speed of light). By observing the time between transmitting the pulse and receiving it, you can calculate the virtual height of the ionosphere.

Maximum usable frequency (MUF) The *maximum usable frequency* is the highest frequency at which communications can take place via the ionosphere over a given path. The MUF between a fixed transmitter site and two different, widely separated, receivers need not be the same. Generally, however, the MUF is approximately

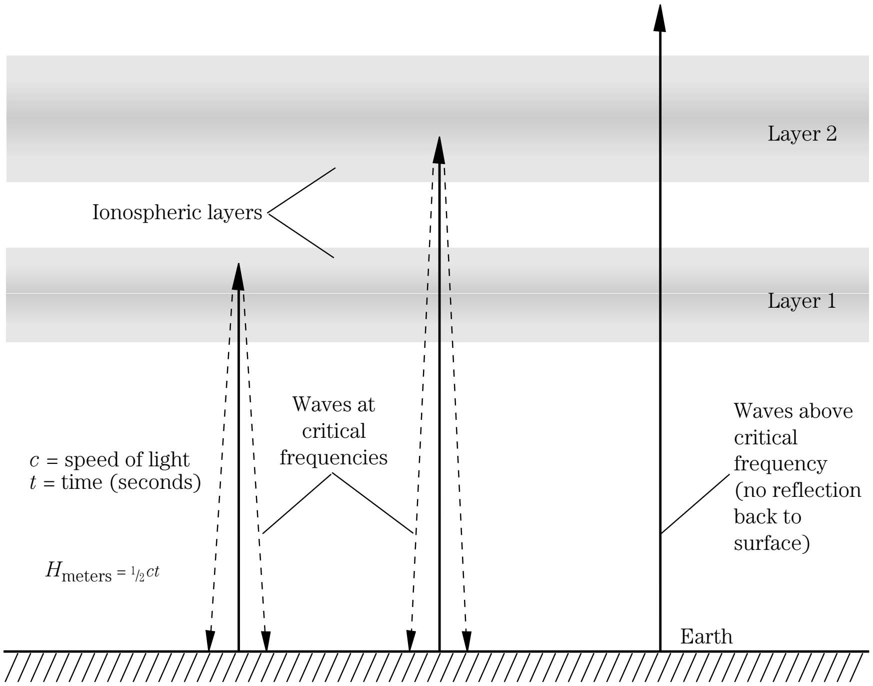
Layer 2
Ionospheric layers
Layer 1
Waves at
critical
frequencies
Waves above
critical
frequency
(no reflection
back to
surface)
c = speed of light
t = time (seconds)
$H_{\text{meters}} = {}^{1}/_{2}ct$
Earth

2-28 Finding critical height of the ionosphere.

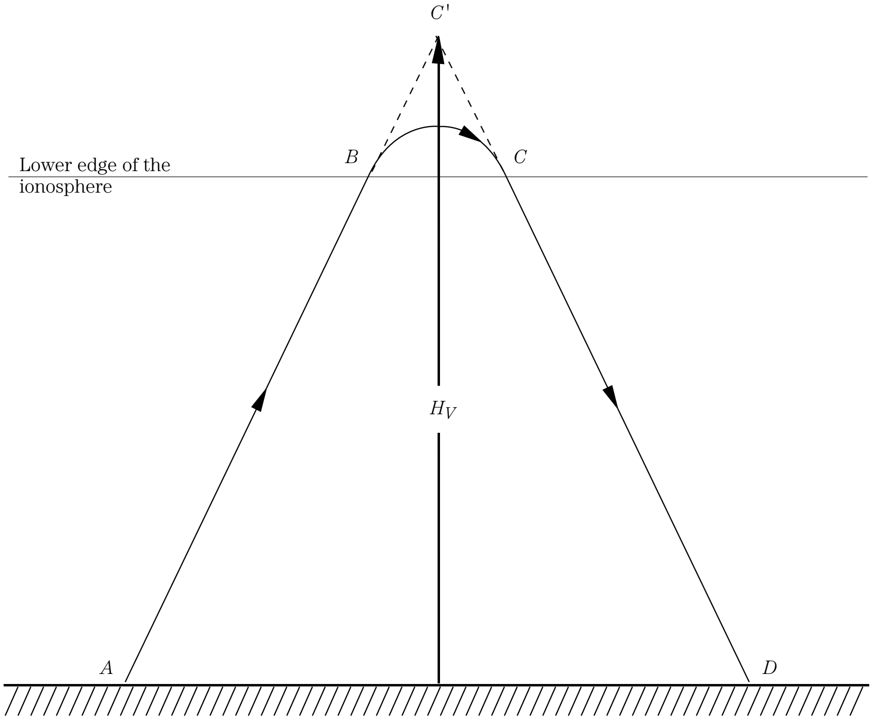
C'
B
C
Lower edge of the
ionosphere
H_V
A
D

2-29 Finding virtual height of the ionosphere.

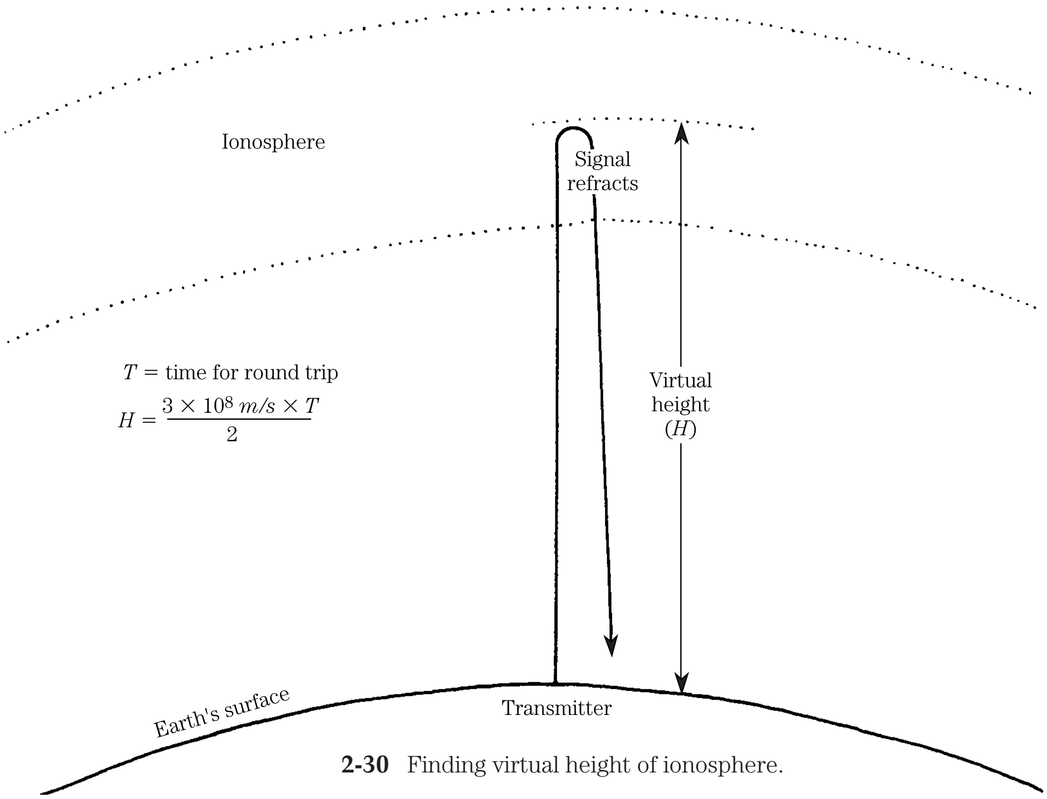

2-30 Finding virtual height of ionosphere.

3 times higher than the critical frequency. Both the MUF and the critical frequency vary geographically, and they become higher at latitudes close to the equator.

It is a general rule that the best propagation occurs at frequencies just below the MUF. In fact, there is a so-called *frequency of optimum traffic* (FOT) that is approximately 85 percent of the MUF. Both noise levels and signal strengths are improved at frequencies near the FOT.

Lowest usable frequency (LUF) At certain low frequencies, the combination of ionospheric absorption, atmospheric noise, miscellaneous static, and/or receiver signal-to-noise ratio requirements conspire to reduce radio communications. The lowest frequency that can be used for communications, despite these factors, is the *lowest usable frequency*.

Unlike the MUF, the LUF is not totally dependent on atmospheric physics. The LUF of a system can be varied by controlling the signal-to-noise ratio (SNR). Although certain factors that contribute to SNR are beyond our control, the effective radiated power (ERP) of the transmitter can be changed; a 2-MHz decrease in LUF is available for every 10-dB increase in the ERP of the transmitter.

Ionospheric variation and disturbances

The ionosphere is an extremely dynamic region of the atmosphere, especially from a radio operator's point of view, because it significantly alters radio propagation. The dynamics of the ionosphere are conveniently divided into two general classes: *regular variation* and *disturbances*. This section covers both types of ionospheric change.

Ionospheric variation There are several different forms of variation seen on a regular basis in the ionosphere: *diurnal*, *27-day* (monthly), *seasonal*, and *11-year cycles*.

Diurnal (daily) variation The sun rises and falls on a 24-hour cycle, and because it is the principal source of ionionization of the upper atmosphere, you can expect diurnal variation. During daylight hours the E and D levels exist, but these disappear at night. The height of the F2 layer increases until midday, and then it decreases until evening, when it disappears or merges with other layers. As a result of higher absorption in the E and D layers, lower frequencies are not useful during daylight hours. On the other hand, the F layers reflect higher frequencies during the day. In the 1- to 30-MHz region, the higher frequencies (>11 MHz) are used during daylight hours, and the lower frequencies (<11 MHz) at night.

27-day cycle Approximately monthly, this variation is caused by the rotational period of the sun. Sunspots are localized on the surface of the sun, so they will face the earth only during a portion of the month. As new sunspots are formed, they do not show up on the earthside face until their region of the sun rotates earthside.

Seasonal cycle The earth's tilt varies the exposure of the planet to the sun on a seasonal basis. In addition, the earth's yearly orbit is not circular; it is elliptical. As a result, the intensity of the sun's energy that ionizes the upper atmosphere varies with the seasons of the year. In general, the E, D, and F layers are affected—although the F2 layer is only minimally affected. Ion density in the F2 layer tends to be highest in winter, and less in summer. During the summer, the distinction between F1 and F2 layers is less obvious.

11-year cycle The number of sunspots, statistically averaged, varies on an approximately 11-year cycle. As a result, the ionospheric effects that affect radio propagation also vary on an 11-year cycle. Radio propagation, in the shortwave bands, is best when the average number of sunspots is at its highest.

Disturbances Disturbances in the ionosphere can have a profound effect on radio communications—and most of them (but not all) are bad. This section will briefly examine some of the more common forms.

Sporadic E layer A reflective cloud of ionization sometimes appears in the E layer of the ionosphere; this layer is sometimes called the E_s *layer*. It is believed that the E_s layer forms from the effects of wind shear between masses of air moving in opposite directions. This action appears to redistribute ions into a thin layer that is radio-reflective.

Sporadic E propagation is normally thought of as a VHF phenomenon, with most activity between 30 and 100 MHz, and decreasing activity up to about 200 MHz. However, about 25 to 50 percent of the time, sporadic E propagation is possible on frequencies down to 10 or 15 MHz. Reception over paths of 1400 to 2600 mi are possible in the 50-MHz region when sporadic E is present. In the northern hemisphere, the months of June and July are the most prevalent sporadic E months. On most days when sporadic E is present, it lasts only a few hours.

Sudden ionospheric disturbances (SIDs) The SID, or *Dellinger fade,* mechanism occurs suddenly, and rarely gives any warning. The SID can last from a few minutes to many hours. It is known that SIDs often occur in correlation with solar flares, or "bright solar eruptions," that produce an immense amount of ultraviolet radiation that impinges the upper atmosphere. The SID causes a tremendous increase in D layer ionization, which accounts for the radio propagation effects. The ionization is so intense that all receiver operators on the sunny side of the earth experience profound loss of signal strength above about 3 MHz. It is not uncommon for receiver owners to think their receivers are malfunctioning when this occurs. The sudden loss of signals on sunny side receivers is called *Dellinger fade*. The SID is often accompanied by variations in terrestrial electrical currents and magnetism levels.

Ionospheric storms The *ionospheric storm* appears to be produced by an abnormally large rain of atomic particles in the upper atmosphere, and is often preceded by SIDs 18 to 24 hours earlier. These storms tend to last from several hours, to a week or more, and are often preceded by two days or so by an abnormally large collection of sunspots crossing the solar disk. They occur, most frequently, and with greatest severity, in the higher latitudes, decreasing toward the equator. When the ionospheric storm commences, shortwave radio signals may begin to flutter rapidly and then drop out altogether. The upper ionosphere becomes chaotic; turbulence increases and the normal stratification into layers, or zones, diminishes.

Radio propagation may come and go over the course of the storm, but it is mostly dead. The ionospheric storm, unlike the SID, which affects the sunny side of the earth, is worldwide. It is noted that the MUF and critical frequency tend to reduce rapidly as the storm commences.

An ionospheric disturbance observed in November 1960 was preceded by about 30 minutes of extremely good, but abnormal, propagation. At 1500 hours EST, European stations were noted with S9+ signal strengths in the 7000- to 7300-kHz region of the spectrum, which is an extremely rare occurrence. After about 30 minutes, the bottom dropped out and even AM broadcast band skip (later that evening) was nonexistent. At the time, the National Bureau of Standards* radio station, WWV, was broadcasting a "W2" propagation prediction at 19 and 49 minutes after each hour (which is terrible!). It was difficult to hear even the 5-MHz WWV frequency in the early hours of the disturbance, and it disappeared altogether for the next 48 hours. This signal's fade-out occurred during the weekend of the annual ARRL "Sweepstakes" contest.

Ionospheric sky-wave propagation

Sky-wave propagation occurs because signals in the ionosphere are refracted so much that they are bent back toward the earth's surface. To observers on the surface, it looks like the signal was reflected from a radio mirror at the virtual height of the ionosphere. The *skip distance* is the surface distance between the transmitter point (*A* in Fig. 2-31) and the point where it returns to earth (point *C* in Fig. 2-31). The *ground-wave zone* is the distance from the transmitter site (*A* in Fig. 2-31) to where the ground wave fades to a low level, below usefulness (point *B* in Fig. 2-31). The *skip zone* is the distance from the outer edge of the ground-wave zone to the skip distance, or the distance from *B* to *C* in Fig. 2-31.

It is possible for the sky wave and the ground wave to interfere with each other at some frequencies, under some circumstances. When this happens, the sky wave has a relative phase that depends on its path length (among other

---

* Now the National Institute of Standards and Technology.

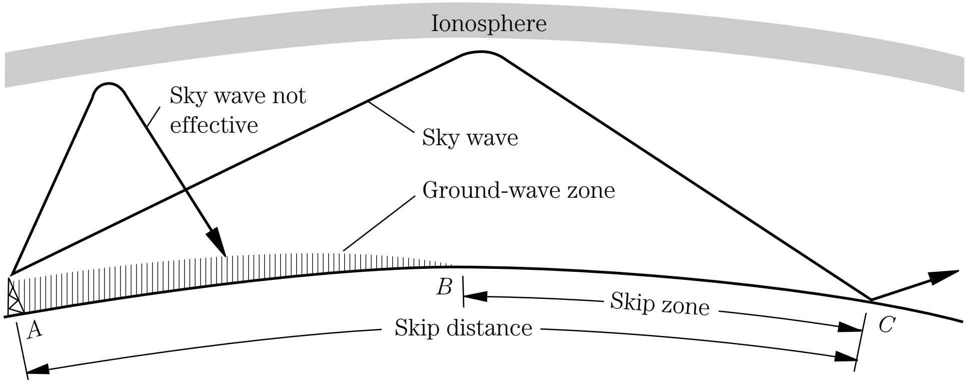

2-31 Relationship between ground wave and different cases of sky wave.

things), so it will arrive at some seemingly random phase relative to the ground wave. Thus, the sky wave can selectively strengthen or cancel the ground wave, giving rise to a type of fading.

Incident angle One of the factors that affects the length of the skip distance is the incident angle of the radio wave. This angle is partially a function of the frequency, and partially a function of the natural radiation angle of the antenna (α_r). The frequency effects are seen as a function of the ionization level and how much a given frequency is refracted.

The antenna radiation angle is the angle of the main vertical lobe with respect to the earth's surface; it is partially a function of its design, and partially of its installation configuration. For example, a ⅝-wavelength vertical antenna tends to have a lower angle of radiation than a ¼-wave vertical antenna. Similarly, the dipole's angle of radiation is a function of its height above ground.

Figure 2-32 shows how the angle of radiation affects skip distance. Low-angle-of-radiation signals tend to travel farther with respect to the earth's surface before refracting, so they produce the longest skip distances. Higher angles of radiation have shorter skip distances because they tend to return to earth more rapidly. In summertime, some high-frequency bands (e.g,. the 11-m—27 MHz—Citizens Band) offer high-angle "short skip" during the summer, and longer skip during the other months. At higher angles of radiation, there will be a *critical angle* wave and *escape angle* waves that are not returned to earth. These waves are not used for terrestrial communications or broadcasting. Figure 2-33 shows the difference between single-hop skip and multihop skip, as a function of incident angle. It is generally true that a multihop transmission is more subject to fading, and is weaker, than a single-hop transmission.

Using the ionosphere

The refraction of high-frequency and some medium-wave radio signals back to earth via the ionosphere gives rise to intercontinental HF radio communications. This phenomenon becomes possible during daylight hours and for a while after sunset when the ionosphere is ionized. Figure 2-34 reiterates the mechanism of long-distance skip communications. The transmitter is located at point *T*, while receiving stations

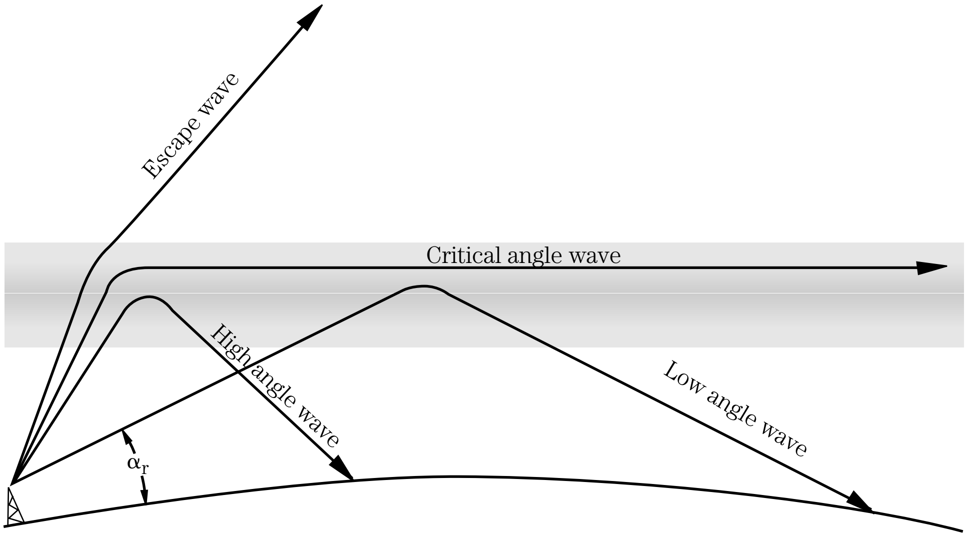
Escape wave
Critical angle wave
High angle wave
Low angle wave
α_r

2-32 Sky-wave propagation as a function of antenna radiation angle.

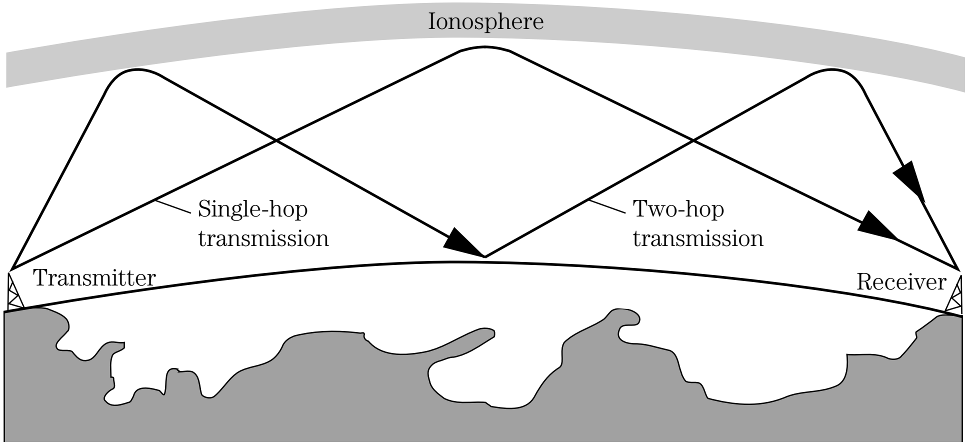
Ionosphere
Single-hop transmission
Two-hop transmission
Transmitter
Receiver

2-33 Single-skip and multihop-skip communications.

are located at sites R_1 and R_2. Signals 1 and 2 are not refracted sufficiently to be refracted back to earth, so they are lost in space. Signal 3, however, is refracted enough to return to earth, so it is heard at station R_1. The skip distance for signal 3 is the distance from T to R_1. At points between T and R_1, signal 3 is inaudible, except within ground-wave distance of the transmitter site (T). This is the reason why two stations 40 mi apart hear each other only weakly, or not at all, while both stations can communicate with a third station 2000 mi away. In amateur radio circles, it is common for South American stations to relay between two U.S. stations only a few miles apart. For an example of this problem, listen to the Inter-American and Halo Missionary

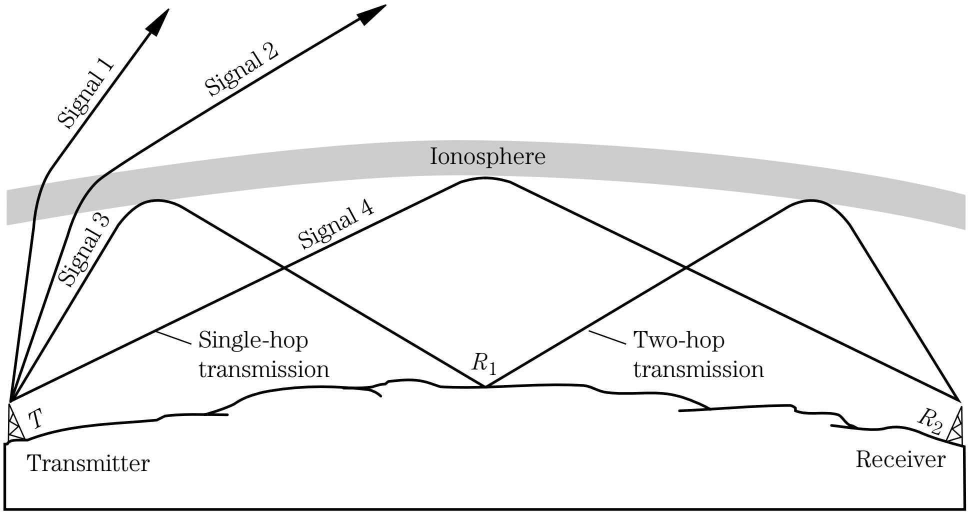

2-34 Effects of radiation angle on distance.

Nets on 21.290 MHz (15 m) daily from about 1700Z to 2100Z (ending time dependent upon traffic).

Multihop skip is responsible for the reception of the signal from transmitter T at site R_3. The signal *reflects* (not refracts) from the surface at R_1, and is retransmitted into the ionosphere, where it is again refracted back to earth.

Figure 2-35 shows a situation where skip signals are received at different distances depending upon the *angle of radiation* of the transmitting antenna. A high angle of radiation causes a shorter skip zone, but a lower angle of radiation results in a longer skip zone. Communication between any particular locations on any given frequency requires adjustment of the antenna radiation angle. Some international shortwave stations have multiple antennas with different radiation angles to ensure that the correct skip distances are available.

Great circle paths A *great circle* is a line between two points on the surface of a sphere, such that it lays on a plane through the earth's center and includes the two points. When translated to "radiospeak," a great circle is the shortest path on the surface of the earth between two points. Navigators and radio operators use the great circle for similar, but different, reasons. The navigator's reason is in order to get from here to there, and the radio operator's is to get a transmission path from here to there.

The heading of a directional antenna is normally aimed at the receiving station along its great circle path. Unfortunately, many people do not understand the concept well enough, for they typically aim the antenna in the wrong direction. For example, I live near Washington, D.C., which is on approximately the same latitude as Lisbon, Portugal. If I catch a lift on Superman's back, and he flies due East, we'll have dinner in Lisbon, right? Wrong. If you head due east from Washington, DC, across the Atlantic, the first landfall would be west Africa, somewhere near Zaire or Angola. Why? Because the great circle bearing 90° takes us far south. The *geometry of spheres*, not flat planes, governs the case.

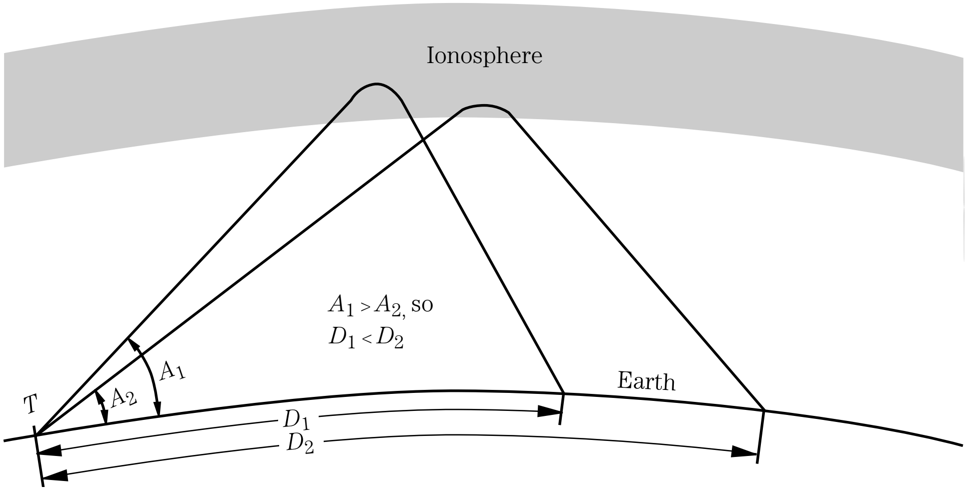

2-35 Skip propagation.

Figure 2-36 shows a great circle map centered on the Washington, D.C. area. These maps, or computer tabulations of the same data, can often be purchased for your own location by supplying your latitude and longitude to the service company that does the job. By drawing a line from your location at the center of the chart to the area you want to hear, and then extending it to the edge of the chart, you will obtain the beam heading required.

Long path versus short path The earth is a sphere (or more precisely, an oblate spheroid), so from any given point to any other point there are two great circle paths: the *long path* (major arc) and the *short path* (minor arc). In general, the best reception occurs along the short path. In addition, short-path propagation is more nearly "textbook," compared with long path reception. However, there are times when long path is better, or is the only path that will deliver a signal to a specific location from the geographic location in question.

Gray line propagation The *gray line* is the twilight zone between the nighttime and daytime halves of the earth. This zone is also called the *planetary terminator* (Fig. 2-37). It varies up to +23° either side of the north-south longitudinal lines, depending on the season of the year (it runs directly north-south only at the vernal and autumnal equinoxes). The D layer of the ionosphere absorbs signals in the HF region. This layer disappears almost completely at night, but it builds up during the day. Along the gray line, the D layer is rapidly decaying west of the line, and has not quite built up east of the line.

Brief periods of abnormal propagation occur along the gray line. Stations on either side of the line can be heard from regions, and at distances, that would otherwise be impossible on any given frequency. For this reason, radio operators often prefer to listen at dawn and dusk for this effect.

Scatter propagation modes

Ionospheric scatter propagation occurs when clouds of ions exist in the atmosphere. These clouds can exist in both the ionosphere and the troposphere, although

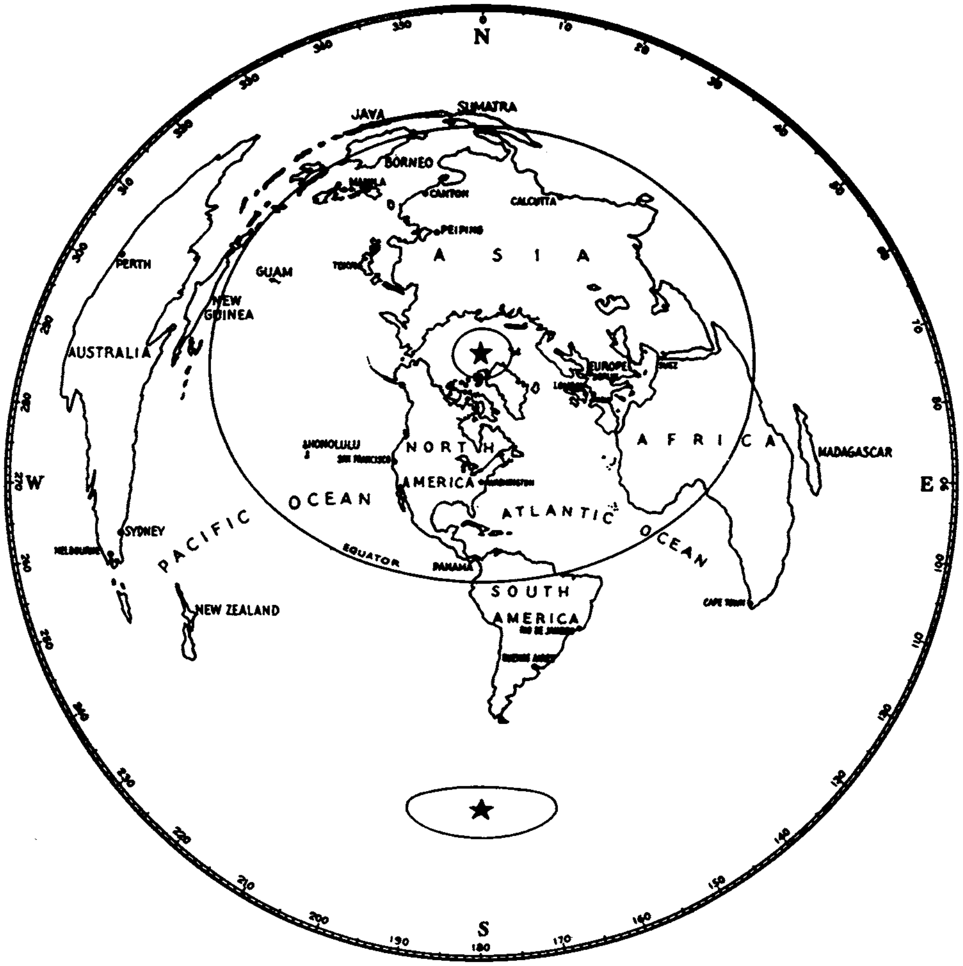
N
S
E
W
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
JAVA
SUMATRA
BORNEO
MANILA
CANTON
CALCUTTA
PEIPING
TOKIO
GUAM
A S I A
PERTH
NEW GUINEA
AUSTRALIA
EUROPE
BERLIN
LONDON
SUEZ
HONOLULU
SAN FRANCISCO
NORTH AMERICA
WASHINGTON
AFRICA
MADAGASCAR
PACIFIC OCEAN
ATLANTIC OCEAN
SYDNEY
MELBOURNE
EQUATOR
PANAMA
SOUTH AMERICA
RIO DE JANEIRO
BUENOS AIRES
NEW ZEALAND
CAPE TOWN

2-36 Azimuthal map centered on Washington, D.C. (*Courtesy of The ARRL Antenna Book.*)

the tropospheric model is more reliable for communications. Figure 2-38 shows the mechanism for scatter propagation. Radio signals from the transmitter are reflected from the cloud of ions to a receiver location that otherwise might not receive it. Scatter propagation occurs mostly in the VHF region, and it allows communications over extended paths that are not normally available.

There are at least three different modes of scatter from ionized clouds: *backscatter*, *side scatter*, and *forward scatter*. The backscatter mode is a bit like radar, in that the signal is returned back to the transmitter site, or to regions close to the transmitter. Forward scatter occurs when the reflected signal continues in the same azimuthal direction (with respect to the transmitter), but is redirected toward the earth's surface. Side scatter is similar to forward scatter, but the azimuthal direction might change.

Unfortunately, there are often multiple reflections from the ionized cloud, and these are shown as "multiple scatter" in Fig. 2-38. When these reflections are able to

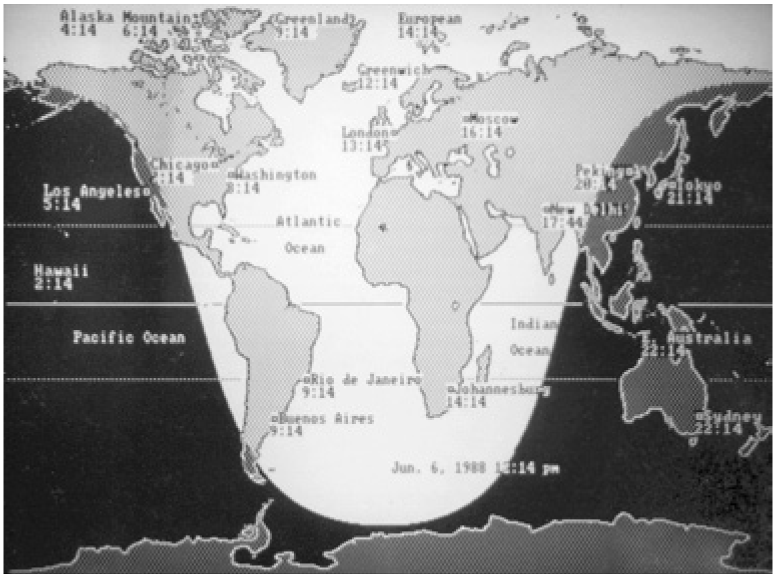
Alaska
4:14
Mountain
6:14
Greenland
9:14
European
14:14
Greenwich
12:14
London
13:14
Moscow
16:14
Chicago
:14
Washington
8:14
Los Angeles
5:14
Atlantic
Ocean
Peking
20:14
Tokyo
21:14
New Delhi
17:44
Hawaii
2:14
Pacific Ocean
Indian
Ocean
Australia
22:14
Rio de Janeiro
9:14
Johannesburg
14:14
Buenos Aires
9:14
Sydney
22:14
Jun. 6, 1988 12:14 pm

2-37 The planetary terminator ("gray line") provides some unusual propagation effects. (*Courtesy of MFJ Enterprises.*)

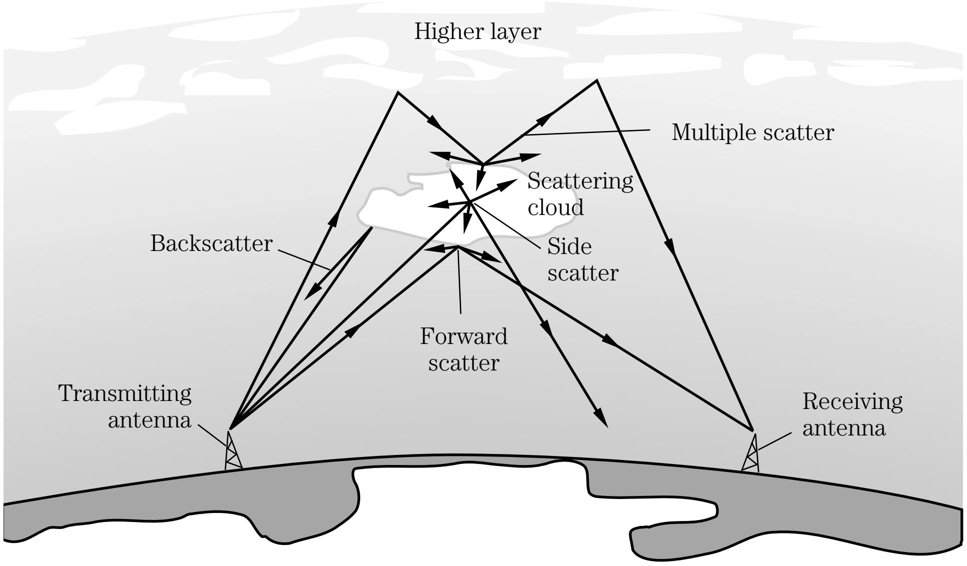
Higher layer
Multiple scatter
Scattering
cloud
Backscatter
Side
scatter
Forward
scatter
Transmitting
antenna
Receiving
antenna

2-38 Various modes of scatter propagation.

reach the receiving site, the result is a rapid, fluttery fading that can be of quite profound depths.

Auroral propagation

The *auroral effect* produces a luminescence in the upper atmosphere resulting from bursts of particles released from the sun 18 to 48 hours earlier. The light emitted is called the *northern lights* and the *southern lights*. The ionized regions of the atmosphere that create the lights form a radio reflection shield, especially at VHF and above, although 15 to 20 MHz effects are known. Auroral propagation effects are normally seen in the higher latitudes, although listeners in the southern tier of states in the United States are often treated to the reception of signals from the north being reflected from auroral clouds.

Meteor scatter propagation

When meteors enter the earth's atmosphere, they do more than simply heat up to the point of burning. The burning meteor leaves a wide, but very short duration, transient cloud of ionized particles in its path. These ions act as a radio mirror that permits short bursts of reception between sites correctly situated. Meteor scatter reception is not terribly reliable, although at least two companies offer meteor scatter communications services for users in the higher latitudes.

Other propagation anomalies

The ionosphere is a physically complex place, and even the extensive coverage in this chapter is not sufficient to do it justice. Indeed, entire books are available on the subject, and it is a valid engineering subspecialty. It is therefore not surprising that a number of propagation anomalies are known.

Nonreciprocal direction If you listen to an amateur band receiver on the East Coast of the United States, you will sometimes hear European stations—especially in the late afternoon. But when you try to work those stations there is no reply whatsoever. They simply don't hear you! This propagation anomaly causes the radio wave to travel different paths dependent on which direction it travels; i.e., an east→west signal is not necessarily the reciprocal of a west→east signal. This anomaly can occur when a radio signal travels through a heavily ionized medium in the presence of a magnetic field, which is exactly the situation when the signal travels through the ionosphere in the presence of the earth's magnetic field.

Another anomaly seen in the radio literature of the 1930s is the *Radio Luxembourg effect*. It is named after the radio station where it was first noticed. In a nonlinear ionosphere, it is sometimes noted that the modulation of superpower (i.e., > 500,000 W) shortwave broadcasters will be transferred to the carrier of a weaker signal in the same or nearby band. The interchange noted in the 1930s when this phenomenon was first discovered was between Radio Luxembourg and Britain's British Broadcasting Corporation (BBC) overseas outlets.

Propagation predictions

Propagation predictions for the VLF through low VHF bands are published each month in several magazines. Ham radio operators often use those in *QST* magazine (Fig. 2-39), while SWLs and others tend to prefer those published in magazines such as *Monitoring Times*. These charts relate the time of day in Universal Coordinated Time (UTC, formerly, GMT) and the frequency for transmission to different parts of the world.

Fading

Skip communications are not without problems. One phenomenon is fading (i.e., a variation in signal strength as perceived at the receiver site). This problem can sometimes be overcome by using one of several diversity reception systems. Three forms of diversity technique are used: *frequency diversity*, *spatial diversity*, and *polarity diversity*.

In the frequency diversity system (Fig. 2-40), the transmitter will send out two or more frequencies simultaneously with the same modulating information. Because the two frequencies will fade differentially, one will always be strong.

The spatial diversity system (Fig. 2-41) assumes that a single transmitter frequency is used. At the receiving site, two or more receiving antennas are used, spaced one-half wavelength apart. The theory is that the signal will fade at one antenna while it increases at the other. A three-antenna system is often used. Three separate, but identical, receivers, often tuned by the same master local oscillator, are connected to the three antennas. Audio mixing, based on the strongest signal, keeps the audio output constant while the radio-frequency (RF) signal fades.

Polarity diversity reception (Fig. 2-42) uses both vertical and horizontal polarization antennas to receive the signal. As in the space diversity system, the outputs of the vertical and horizontal receivers are combined to produce a constant level output.

Another form of fading, *selective fading*, derives from the fact that fading is a function of frequency. The carrier and upper and lower sidebands of an AM signal have slightly different frequencies, so they arrive out of phase with each other. Although this type of fading is lessened by using single-sideband (SSB) transmission, that does not help AM users. In those systems, some people use a filtering system that eliminates the carrier and one sideband; it then reconstitutes the AM signal with a product detector.

SSB receivers with stable local and product detector oscillators, and a sharp intermediate-frequency (IF) bandpass filter, can be used to reduce the effects of differential fading of AM signals because of the phasing of the lower sideband (LSB), upper sideband (USB), and carrier components. Carefully tune the receiver to only one sideband of the signal, and note when the heterodyne beatnote disappears. The correct point is characterized by the fact that you can then switch among USB, LSB, and continuous-wave (CW) modes without changing the received signal output.

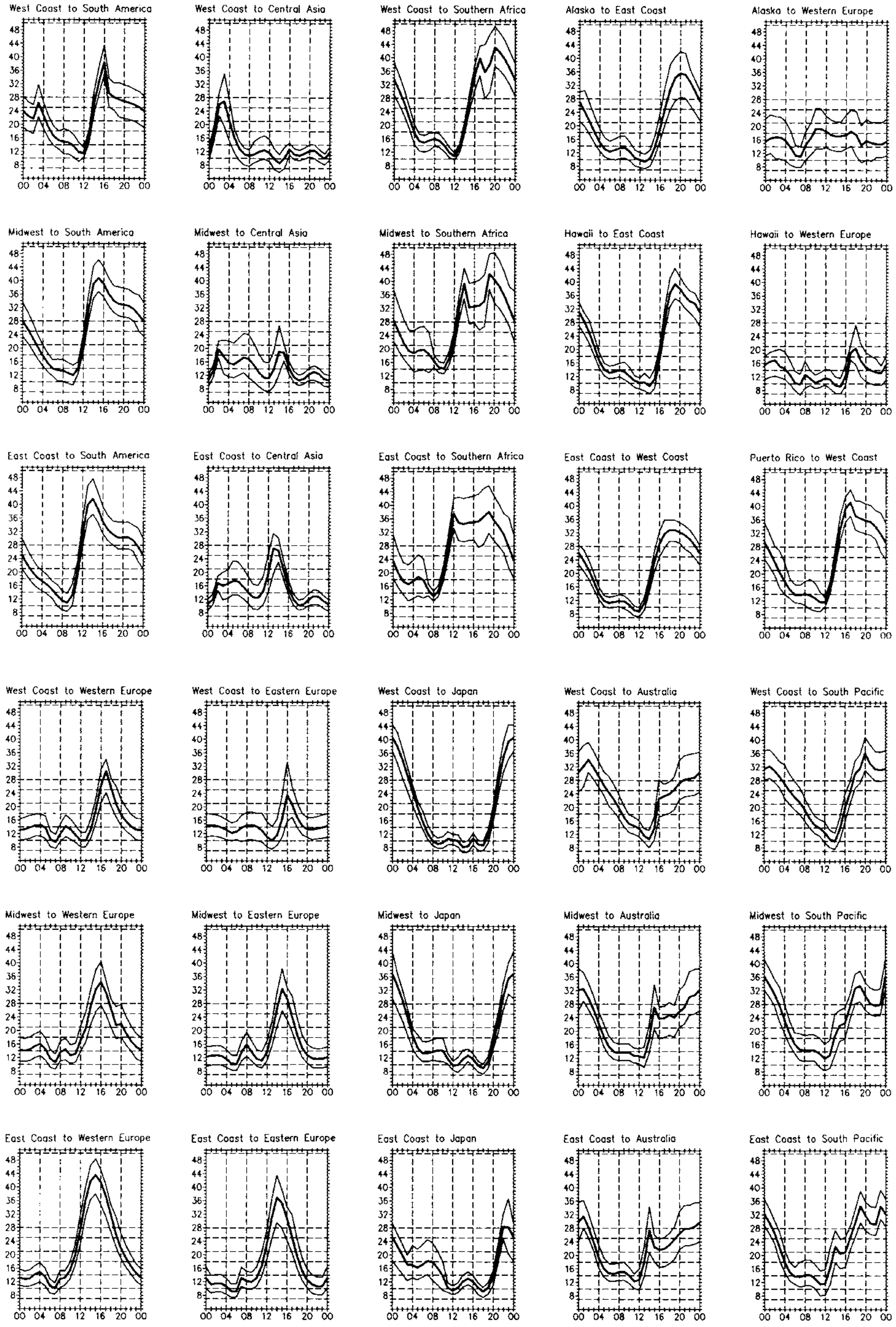

2-39 Propagation prediction charts appear in magazines such as *QST* every month. (*Courtesy of the American Radio Relay League.*)

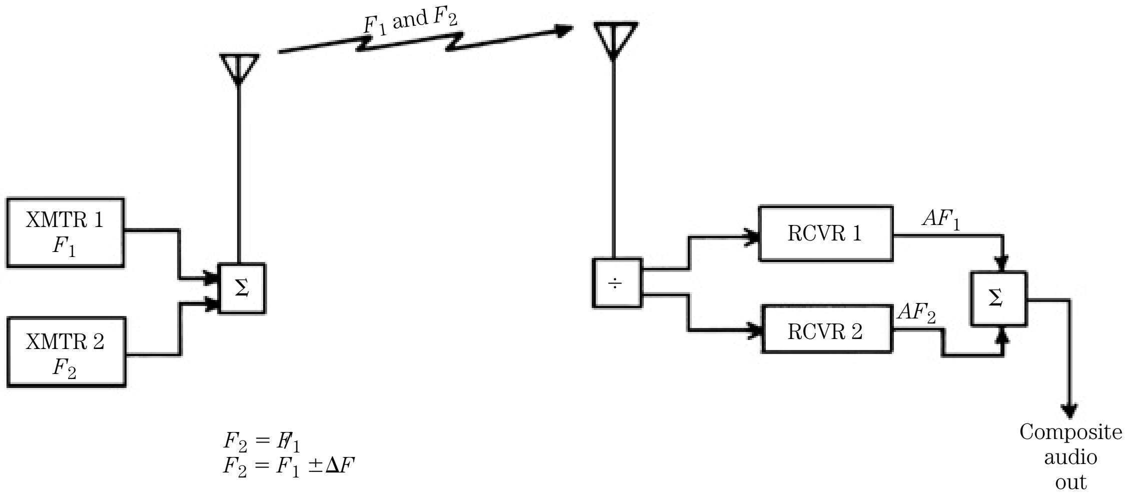
F_1 and F_2
XMTR 1 F_1
XMTR 2 F_2
Σ
÷
RCVR 1
RCVR 2
AF_1
AF_2
Σ
Composite audio out
$F_2 \neq F_1$
$F_2 = F_1 \pm \Delta F$

2-40 Frequency diversity reduces fading.

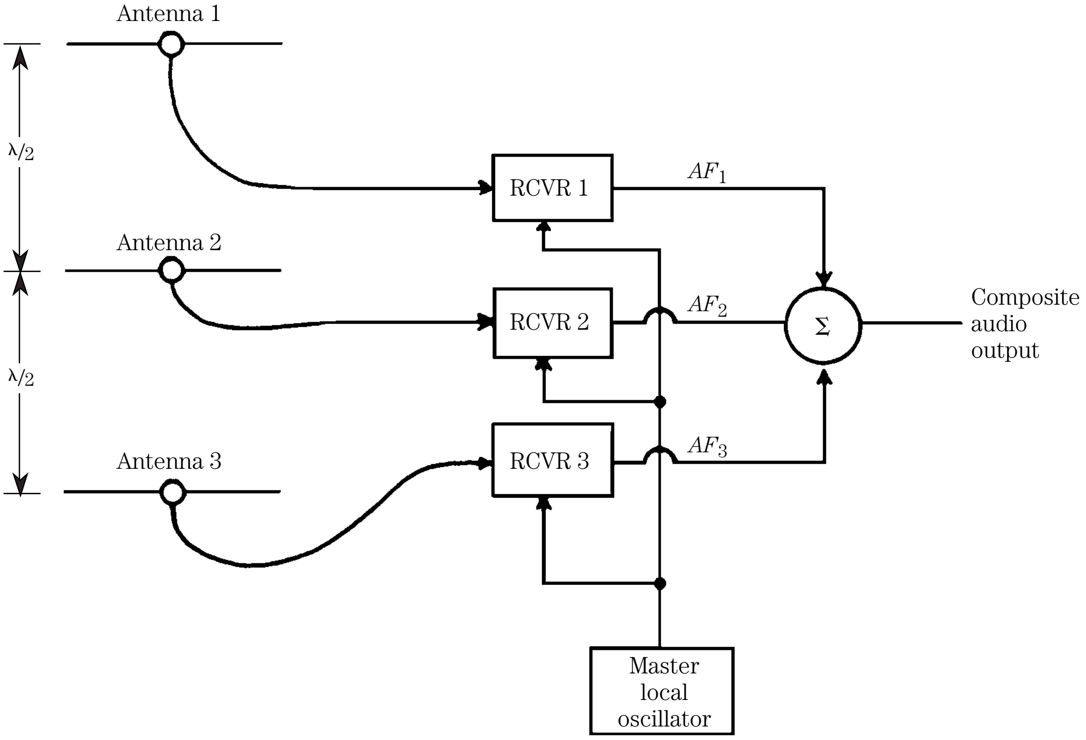
Antenna 1
Antenna 2
Antenna 3
λ/2
λ/2
RCVR 1
RCVR 2
RCVR 3
AF_1
AF_2
AF_3
Σ
Composite audio output
Master local oscillator

2-41 Spatial diversity.

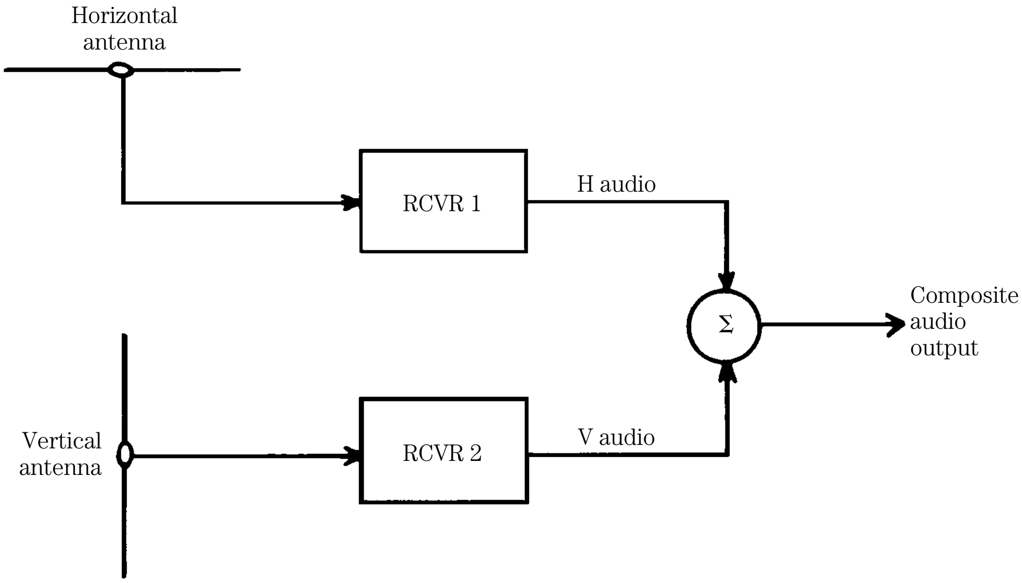
Horizontal antenna
RCVR 1
H audio
Σ
Composite audio output
Vertical antenna
RCVR 2
V audio

2-42 Polarization diversity.

3
CHAPTER

Transmission lines

TRANSMISSION LINES AND *WAVEGUIDES* ARE CONDUITS FOR TRANSPORTING RF SIGNALS between elements of a system. For example, transmission lines are used between an exciter output and transmitter input, and between the transmitter input and its output, and between the transmitter output and the antenna. Although often erroneously characterized as a "length of shielded wire," transmission lines are actually complex networks containing the equivalent of all the three basic electrical components: resistance, capacitance, and inductance. Because of this fact, transmission lines must be analyzed in terms of an *RLC* network.

Parallel and coaxial lines

This chapter will consider several types of transmission lines. Both step-function and sine-wave ac responses will be studied. Because the subject is both conceptual and analytical, both analogy and mathematical approaches to the theory of transmission lines will be used.

Figure 3-1 shows several basic types of transmission line. Perhaps the oldest and simplest form is the *parallel line* shown in Figs. 3-1A through 3-1D. Figure 3-1A shows an end view of the parallel conductor transmission line. The two conductors, of diameter d, are separated by a dielectric (which might be air) by a spacing S. These designations will be used in calculations later. Figure 3-1B shows a type of parallel line called *twin lead*. This is the old-fashioned television antenna transmission line. It consists of a pair of parallel conductors separated by a plastic dielectric. TV-type twin lead has a characteristic impedance of 300 Ω, while certain radio transmitting-antenna twin lead has an impedance of 450 Ω. Another form of twin lead is *open line*, shown in Fig. 3-1C. In this case, the wire conductors are separated by an air dielectric, with support provided by stiff (usually ceramic) insulators. A tie wire (only one shown) is used to fasten each insulator end to the main conductor. Some users of open line prefer the form of insulator or supporter shown in Fig. 3-1D. This

Copyright 2001 - the McGraw-Hill Companies

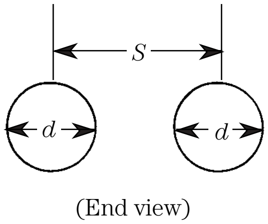
S
d
d
(End view)

3-1A Parallel line transmission line (end view).

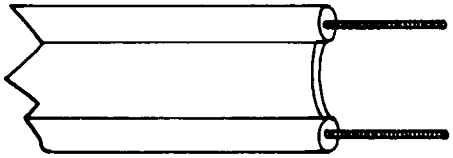

3-1B Twin-lead transmission line.

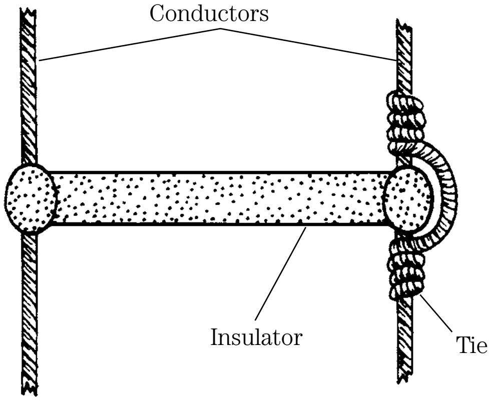
Conductors
Insulator
Tie

3-1C Parallel line construction details.

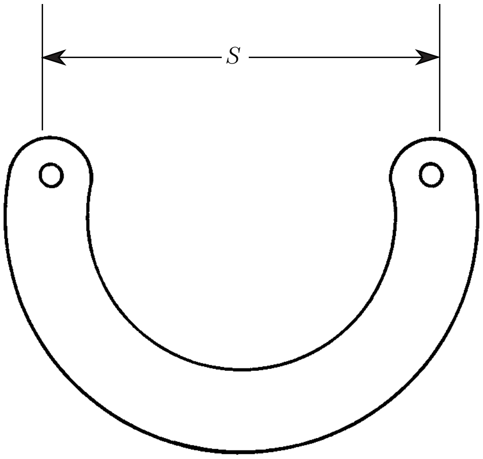

3-1D Horseshoe parallel line spreader.

form of insulator is made of either plastic or ceramic, and is in the form of a U. The purpose of this shape is to reduce losses, especially in rainy weather, by increasing the leakage currents path relative to spacing S.

Parallel lines have been used at VLF, MW, and HF frequencies for decades. Even antennas into the low VHF are often found using parallel lines. The higher impedance of these lines (relative to coaxial cable) yields lower loss in high-power applications. For years, the VHF, UHF, and microwave application of parallel lines was limited to educational laboratories, where they are well suited to performing experiments (to about 2 GHz) with simple, low-cost instruments. Today, however, printed circuit and hybrid semiconductor packaging has given parallel lines a new lease on life, if not an overwhelming market presence.

Figure 3-1E shows a form of parallel line called *shielded twin lead*. This type of line uses the same form of construction as TV-type twin lead, but it also has a braided shielding surrounding it. This feature makes it less susceptible to noise and other problems.

The second form of transmission line, which finds considerable application at microwave frequencies, is *coaxial cable* (Figs. 3-1F through 3-1L). This form of line consists of two cylindrical conductors sharing the same axis (hence "coaxial"), and separated by a dielectric (Fig. 3-1F). For low frequencies (in flexible cables) the dielectric may be polyethylene or polyethylene foam, but at higher frequencies *Teflon* and other materials are used. Also used, in some applications, are dry air and dry nitrogen.

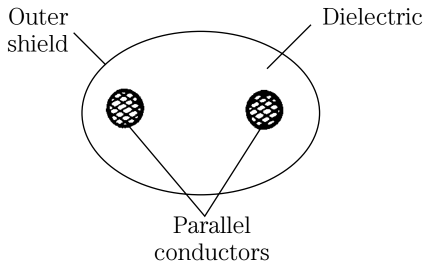

3-1E Shielded twin-lead transmission line.

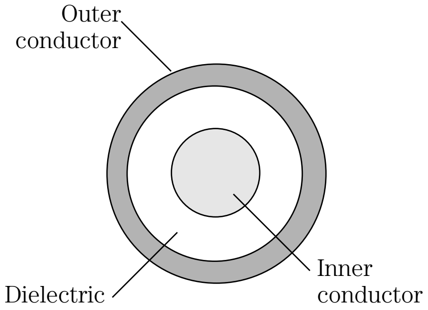

3-1F Coaxial cable (end view).

Several forms of coaxial line are available. Flexible coaxial cable is perhaps the most common form. The outer conductor in such cable is made of either braid or foil (Fig. 3-1G). Television broadcast receiver antennas provide an example of such cable from common experience. Another form of flexible or semiflexible coaxial line is *helical line* (Fig. 3-1H) in which the outer conductor is spiral wound. *Hardline* (Fig. 3-1I) is coaxial cable that uses a thin-wall pipe as the outer conductor. Some hardline coax used at microwave frequencies has a rigid outer conductor and a solid dielectric.

Gas-filled line is a special case of hardline that is hollow (Fig. 3-1J), the center conductor is supported by a series of thin ceramic or Teflon insulators. The dielectric is either anhydrous (i.e., dry) nitrogen or some other inert gas.

Some flexible microwave coaxial cable uses a solid "air-articulated" dielectric (Fig. 3-1K), in which the inner insulator is not continuous around the center conductor, but rather is ridged. Reduced dielectric losses increase the usefulness of the

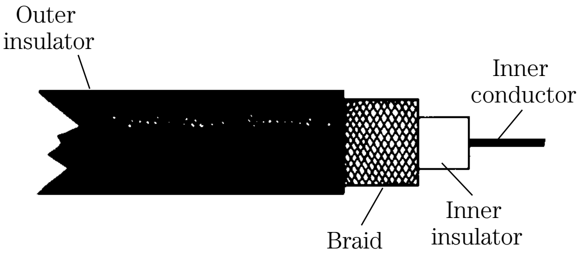
Outer insulator
Inner conductor
Braid
Inner insulator

3-1G Coaxial cable (side view).

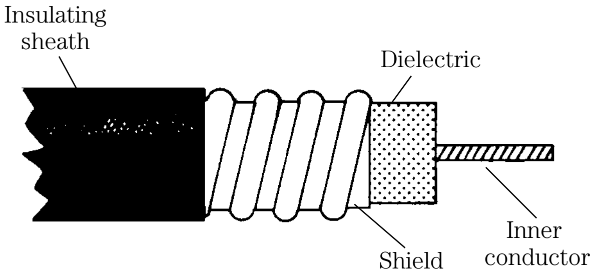
Insulating sheath
Dielectric
Shield
Inner conductor

3-1H Coaxial hardline cable.

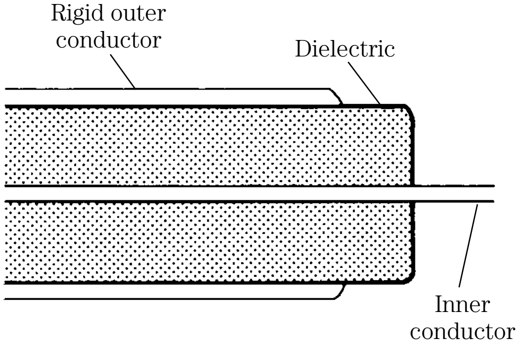
Rigid outer conductor
Dielectric
Inner conductor

3-1I Rigid coaxial line.

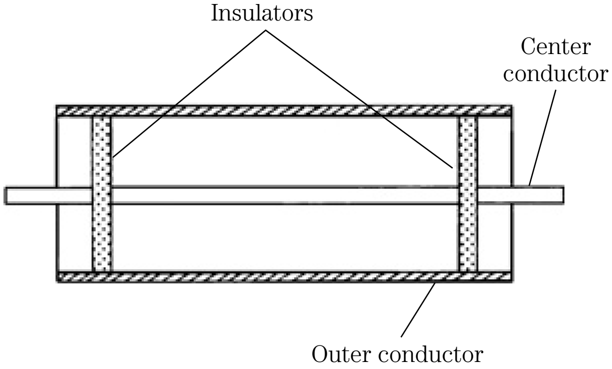
Insulators
Center conductor
Outer conductor

3-1J Gas-filled hollow coaxial line.

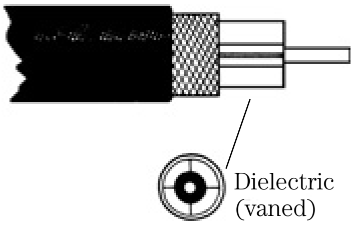
Dielectric (vaned)

3-1K Articulated coaxial line.

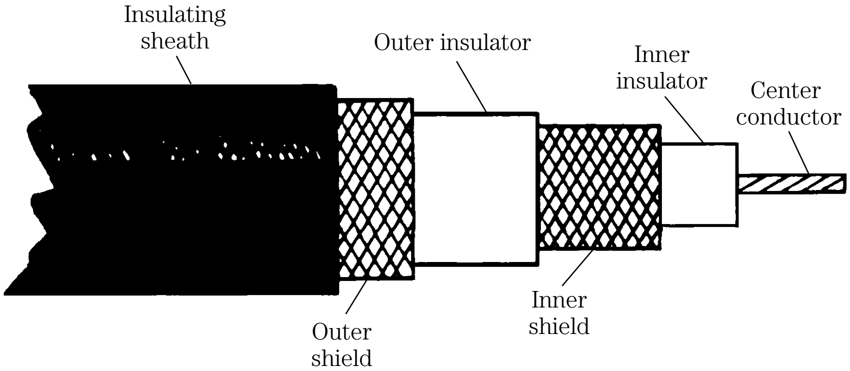
Insulating sheath
Outer insulator
Inner insulator
Center conductor
Inner shield
Outer shield

3-1L Double-shielded coaxial line.

cable at higher frequencies. Double-shielded coaxial cable (Fig. 3-1L) provides an extra measure of protection against radiation from the line, and EMI from outside sources, from getting into the system.

Stripline, also called *microstripline* (Fig. 3-1M), is a form of transmission line used at high UHF and microwave frequencies. The stripline consists of a critically sized conductor over a ground-plane conductor, and separated from it by a dielectric. Some striplines are sandwiched between two groundplanes and are separated from each by a dielectric.

Transmission line characteristic impedance (Z_o)

The transmission line is an *RLC* network (see Fig. 3-2), so it has a *characteristic impedance* Z_o, also sometimes called a *surge impedance*. Network analysis will show that Z_o is a function of the per unit of length parameters *resistance R*, *conductance G*, *inductance L*, and *capacitance C*, and is found from

$$Z_o = \sqrt{\frac{R + j\,\omega\,L}{G + j\,\omega\,C}} \qquad \textbf{[3.1]}$$

where

Z_o is the characteristic impedance, in ohms
R is the resistance per unit length, in ohms
G is the conductance per unit length, in mhos
L is the inductance per unit length, in henrys
C is the capacitance per unit length, in farads
ω is the angular frequency in radians per second ($2\pi F$)

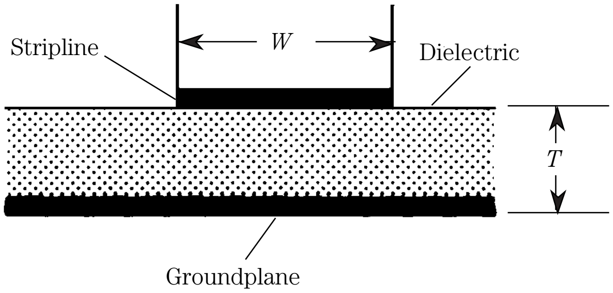

3-1M Stripline transmission line.

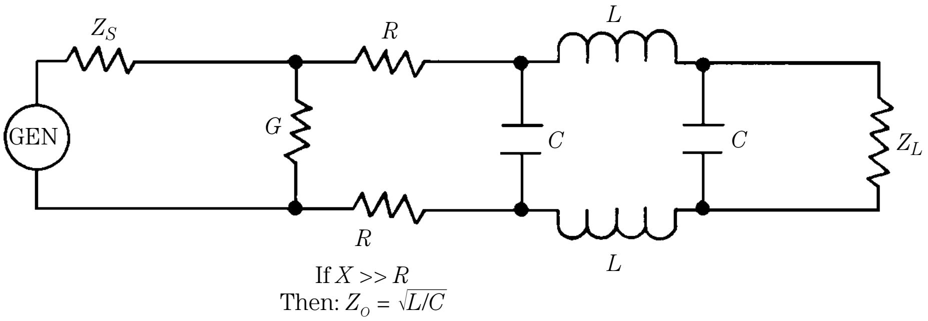

3-2 Equivalent circuit of transmission line.

In microwave systems the resistances are typically very low compared with the reactances, so Eq. 3.1 can be reduced to the simplified form:

$$Z_o = \sqrt{\frac{L}{C}} \qquad \textbf{[3.2]}$$

Example 3-1 A nearly lossless transmission line (R is very small) has a unit length inductance of 3.75 nH and a unit length capacitance of 1.5 pF. Find the characteristic impedance Z_o.

Solution:

$$Z_o = \sqrt{\frac{L}{C}}$$

$$= \sqrt{\frac{\left(3.75\ \text{nH} \times \frac{1\ \text{H}}{10^9\ \text{nH}}\right)}{\left(1.5\ \text{pF} \times \frac{1\ \text{F}}{10^{12}\ \text{pF}}\right)}}$$

$$= \sqrt{\frac{3.75 \times 10^{-9}\ \text{H}}{1.5 \times 10^{-12}\ \text{F}}}$$

$$= \sqrt{2.5 \times 10^3} = 50\ \Omega$$

The characteristic impedance for a specific type of line is a function of the conductor size, the conductor spacing, the conductor geometry (see again Fig. 3-1), and the dielectric constant of the insulating material used between the conductors. The dielectric constant e is equal to the reciprocal of the velocity (squared) of the wave when a specific medium is used:

$$e = \frac{1}{v^2} \qquad \textbf{[3.3]}$$

where

e is the dielectric constant (for a perfect vacuum $e = 1.000$)
v is the velocity of the wave in the medium

(a) Parallel line

$$Z_o = \frac{276}{\sqrt{e}} \log\left(\frac{2S}{d}\right) \qquad \textbf{[3.4]}$$

where

Z_o is the characteristic impedance, in ohms
e is the dielectric constant
S is the center-to-center spacing of the conductors
d is the diameter of the conductors

(*b*) *Coaxial line*

$$Z_o = \frac{138}{\sqrt{e}} \log\left(\frac{D}{d}\right) \tag{3.5}$$

where

D is the diameter of the outer conductor
d is the diameter of the inner conductor

(*c*) *Shielded parallel line*

$$Z_o = \frac{276}{\sqrt{e}} \log\left(2A \quad \frac{(1-B^2)}{(1+B^2)}\right) \tag{3.6}$$

where

$A = s/d$
$B = s/D$

(*d*) *Stripline*

$$Z_o = \frac{377}{\sqrt{e_t}} \left(\frac{T}{W}\right) \tag{3.7A}$$

where

e_t is the relative dielectric constant of the printed wiring board (PWB)
T is the thickness of the printed wiring board
W is the width of the stripline conductor

The relative dielectric constant e_t used above differs from the normal dielectric constant of the material used in the PWB. The relative and normal dielectric constants move closer together for larger values of the ratio W/T.

Example 3-2 A stripline transmission line is built on a 4-mm-thick printed wiring board that has a relative dielectric constant of 5.5. Calculate the characteristic impedance if the width of the strip is 2 mm.

Solution:

$$Z_o = \frac{377}{\sqrt{e_t}} \left(\frac{T}{W}\right)$$

$$= \frac{377}{\sqrt{5.5}} \left(\frac{4 \text{ mm}}{2 \text{ mm}}\right)$$

$$= \frac{377}{2.35} (2) = 321\ \Omega$$

In practical situations, we usually don't need to calculate the characteristic impedance of a stripline, but rather design the line to fit a specific system impedance (e.g., 50 Ω). We can make some choices of printed circuit material (hence dielectric constant) and thickness, but even these are usually limited in practice by the availability of standardized boards. Thus, stripline *width* is the variable parameter. Equation 3.2 can be arranged to the form:

$$W = \frac{377\ T}{Z_o \sqrt{e}} \tag{3.7B}$$

The impedance of 50 Ω is accepted as standard for RF systems, except in the cable TV industry. The reason for this diversity is that power handling ability and low loss operation don't occur at the same characteristic impedance. For example, the maximum power handling ability for coaxial cables occurs at 30 Ω, while the lowest loss occurs at 77 Ω; 50 Ω is therefore a reasonable tradeoff between the two points. In the cable TV industry, however, the RF power levels are minuscule, but lines are long. The tradeoff for TV is to use 75 Ω as the standard system impedance in order to take advantage of the reduced attenuation factor.

Transmission line characteristics

Velocity factor

In the section preceding this section, we discovered that the velocity of the wave (or signal) in the transmission line is less than the free-space velocity (i.e., less than the speed of light). Further, we discovered in Eq. 3.3 that velocity is related to the dielectric constant of the insulating material that separates the conductors in the transmission line. Velocity factor v is usually specified as a decimal fraction of c, the speed of light (3×10^8 m/s). For example, if the velocity factor of a transmission line is rated at "0.66," then the velocity of the wave is $0.66c$, or $(0.66)\ (3 \times 10^8 \text{ m/s}) = 1.98 \times 10^8$ m/s.

Velocity factor becomes important when designing things like transmission line transformers, or any other device in which the length of the line is important. In most cases, the transmission line length is specified in terms of *electrical length*, which can be either an angular measurement (e.g., 180° or π radians), or a relative measure keyed to wavelength (e.g., one-half wavelength, which is the same as 180°). The *physical length* of the line is longer than the equivalent electrical length. For example, let's consider a 1-GHz half-wavelength transmission line.

A rule of thumb tells us that the length of a wave (in meters) in free space is $0.30/F$, where frequency F is expressed in gigahertz; therefore, a half-wavelength line is $0.15/F$.

At 1 GHz, the line must be 0.15 m/1 GHz = 0.15 m. If the velocity factor is 0.80, then the *physical length* of the transmission line that will achieve the desired *electrical length* is [(0.15 m) (v)]/F = [(0.15 m) (0.80)]/1 GHz = 0.12 m. The derivation of the rule of thumb is “left as an exercise for the student.” (*Hint*: It comes from the relationship between wavelength, frequency, and velocity of propagation for any form of wave.)

There are certain practical considerations regarding velocity factor that result from the fact that the physical and electrical lengths are not equal. For example, in a certain type of phased-array antenna design, radiating elements are spaced a half-wavelength apart, and must be fed 180° (half-wave) out of phase with each other. The simplest interconnect is to use a half-wave transmission line between the 0° element and the 180° element. According to the standard wisdom, the transmission line will create the 180° phase delay required for the correct operation of the antenna. Unfortunately, because of the velocity factor, the physical length for a one-half electrical wavelength cable is shorter than the free-space half-wave distance between elements. In other words, the cable will be too short to reach between the radiating elements by the amount of the velocity factor!

Clearly, velocity factor is a topic that must be understood before transmission lines can be used in practical situations. Table 3-1 shows the velocity factors for several types of popular transmission line. Because these are nominal values, the actual velocity factor for any given line should be measured.

Table 3-1. Transmission line characteristics

Type of line	Z_o (ohms)	Velocity factor v
½-in. TV parallel line (air dielectric)	300	0.95
1-in. TV parallel line (air dielectric)	450	0.95
TV twin lead	300	0.82
UHF TV twin lead	300	0.80
Polyethylene coaxial cable	*	0.66
Polyethylene foam coaxial cable	*	0.79
Air-space polyethylene foam coaxial cable	*	0.86
Teflon	*	0.70

* Various impedances depending upon cable type.

Transmission line noise

Transmission lines are capable of generating noise and spurious voltages that are seen by the system as valid signals. Several such sources exist. One source is the coupling between noise currents flowing in the outer conductor and the inner conductor. Such currents are induced by nearby electromagnetic interference and other sources (e.g., connection to a noisy groundplane). Although coaxial design reduces noise pickup, compared with parallel line, the potential for EMI exists. Selection of high-grade line, with a high degree of shielding, reduces the problem.

Another source of noise is thermal noises in the resistances and conductances. This type of noise is proportional to resistance and temperature.

There is also noise created by mechanical movement of the cable. One species results from the movement of the dielectric against the two conductors. This form of noise is caused by electrostatic discharges in much the same manner as the spark created by rubbing a piece of plastic against woolen cloth.

A second species of mechanically generated noise is *piezoelectricity* in the dielectric. Although more common in cheap cables, one should be aware of it. Mechanical deformation of the dielectric causes electrical potentials to be generated.

Both species of mechanically generated noise can be reduced or eliminated by proper mounting of the cable. Although rarely a problem at lower frequencies, such noise can be significant at microwave frequencies when signals are low.

Coaxial cable capacitance

A coaxial transmission line possesses a certain capacitance per unit of length. This capacitance is defined by

$$C = \frac{24e}{\log\ (D/d)}\ \frac{\text{pF}}{\text{m}} \qquad \textbf{[3.8A]}$$

A long run of coaxial cable can build up a large capacitance. For example, a common type of coax is rated at 65 pF/m. A 150-m roll thus has a capacitance of 65 pF/m × (150 m), or 9750 pF. When charged with a high voltage, as is done in breakdown voltage tests at the factory, the cable *acts like a charged high-voltage capacitor*. Although rarely (if ever) lethal to humans, the stored voltage in new cable can deliver a nasty electrical shock and can irreparably damage electronic components.

Coaxial cable cutoff frequency F_o

The normal mode in which a coaxial cable propagates a signal is as a transverse electromagnetic (TEM) wave, but others are possible—and usually undesirable. There is a maximum frequency above which TEM propagation becomes a problem, and higher modes dominate. Coaxial cable should *not* be used above a frequency of

$$F = \frac{6.75}{(D + d)\sqrt{e}}\ \text{GHz} \qquad \textbf{[3.8B]}$$

where

F is the TEM-mode cutoff frequency
D is the diameter of the outer conductor, in inches
d is the diameter of the inner conductor, in inches
e is the dielectric constant

When maximum operating frequencies for cable are listed, it is the TEM mode that is cited. Beware of attenuation, however, when making selections for microwave frequencies. A particular cable may have a sufficiently high TEM-mode frequency, but still exhibit a high attenuation per unit length at X or Ku bands.

Transmission line responses

In order to understand the operation of transmission lines, we need to consider two cases: *step-function response* and the *steady-state ac response*. The step-function case involves a single event when a voltage at the input of the line snaps from zero (or a steady value) to a new (or nonzero) value, and remains there until all action dies out. This response tells us something of the behavior of pulses in the line, and in fact is used to describe the response to a single-pulse stimulus. The steady-state ac response tells us something of the behavior of the line under stimulation by a sinusoidal RF signal.

Step-function response of a transmission line

Figure 3-3 shows a parallel transmission line with characteristic impedance Z_o connected to a load impedance Z_L. The generator at the input of the line consists of a voltage source V in series with a source impedance Z_s and a switch S_1. Assume for the present that all impedances are pure resistances (i.e., $R + j0$). Also, assume that $Z_s = Z_o$.

When the switch is closed at time T_o (Fig. 3-4A), the voltage at the input of the line (V_{in}) jumps to $V/2$. In Fig. 3-2, you may have noticed that the LC circuit resembles a delay line circuit. As might be expected, therefore, the voltage wavefront propagates along the line at a velocity v of:

$$v = \frac{1}{\sqrt{LC}} \tag{3.9}$$

where

v is the velocity, in meters per second
L is the inductance, in henrys
C is the capacitance, in farads

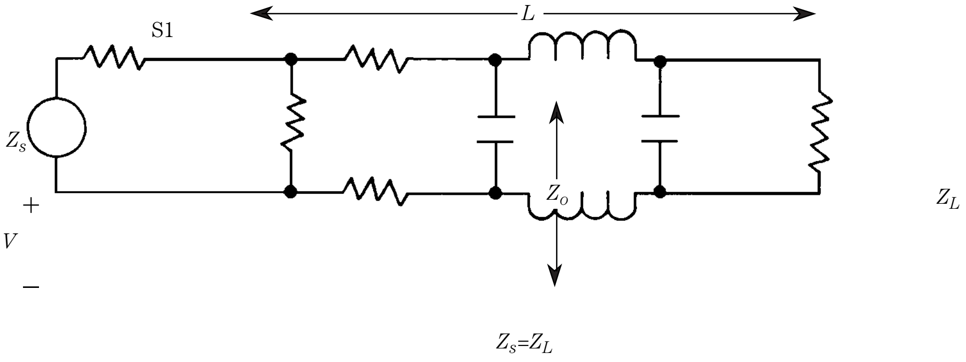

3-3 Schematic example of transmission line.

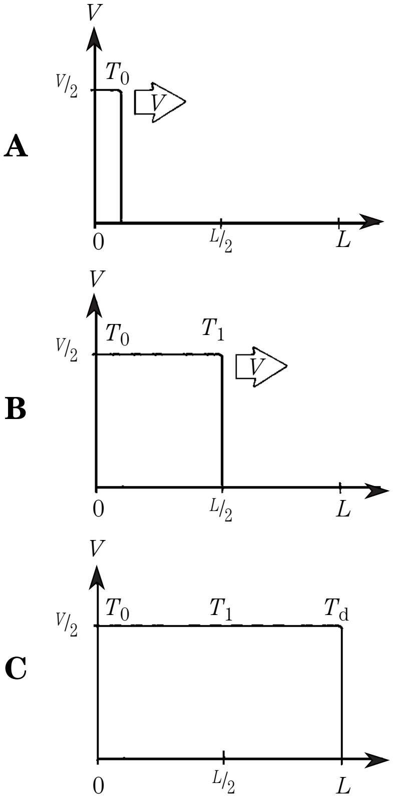

3-4 Step-function propagation along transmission line at three points.

At time T_1 (Fig. 3-4B), the wavefront has propagated one-half the distance L, and by T_d it has propagated the entire length of the cable (Fig. 3-4C).

If the load is perfectly matched (i.e., $Z_L = Z_o$), then the load absorbs the wave and no component is reflected. But in a mismatched system (Z_L is not equal to Z_o), a portion of the wave is reflected back down the line toward the generator.

Figure 3-5 shows the rope analogy for reflected pulses in a transmission line. A taut rope (Fig. 3-5A) is tied to a rigid wall that does not absorb any of the energy in the pulse propagated down the rope. When the free end of the rope is given a vertical displacement (Fig. 3-5B), a wave is propagated down the rope at velocity v (Fig. 3-5C). When the pulse hits the wall (Fig. 3-5D), it is reflected (Fig. 3-5E) and propagates back down the rope toward the free end (Fig. 3-5F).

If a second pulse is propagated down the line before the first pulse dies out, then there will be two pulses on the line at the same time (Fig. 3-6A). When the two pulses interfere, the resultant will be the algebraic sum of the two. In the event that a pulse train is applied to the line, the interference pattern will set up *standing waves*, an example of which is shown in Fig. 3-6B.

Reflection coefficient

The *reflection coefficient* Γ of a circuit containing a transmission line and load impedance is a measure of how well the system is matched. The absolute value of the re-

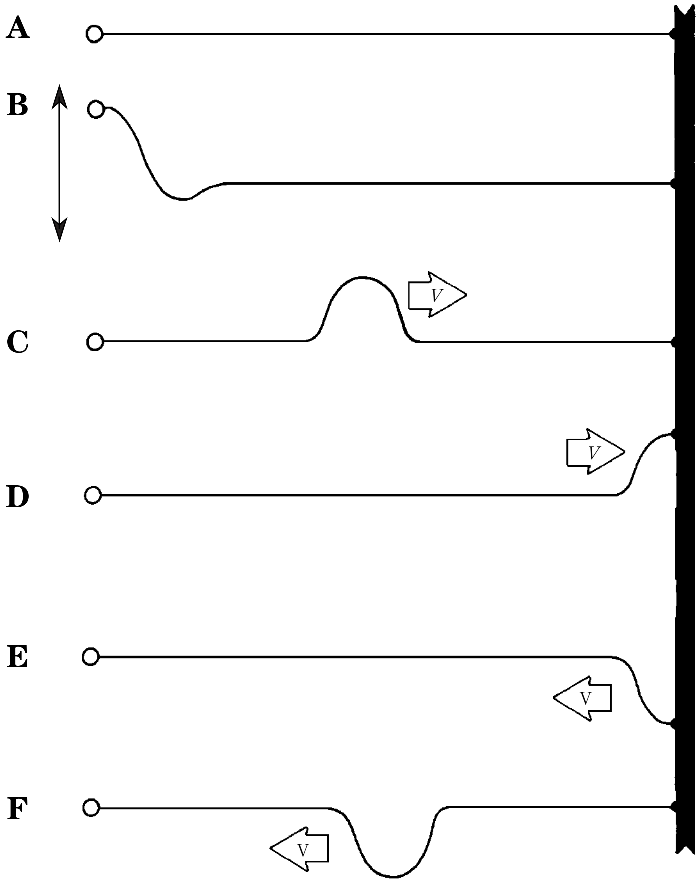
A
B
V
C
V
D
E
V
F
V

3-5 Rope analogy to transmission line.

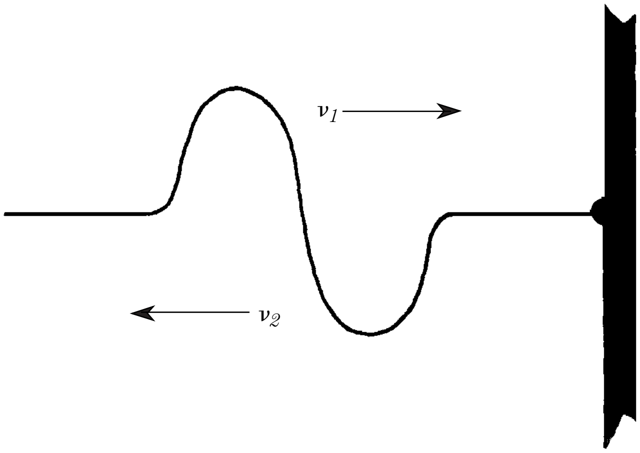

3-6A Interfering opposite waves.

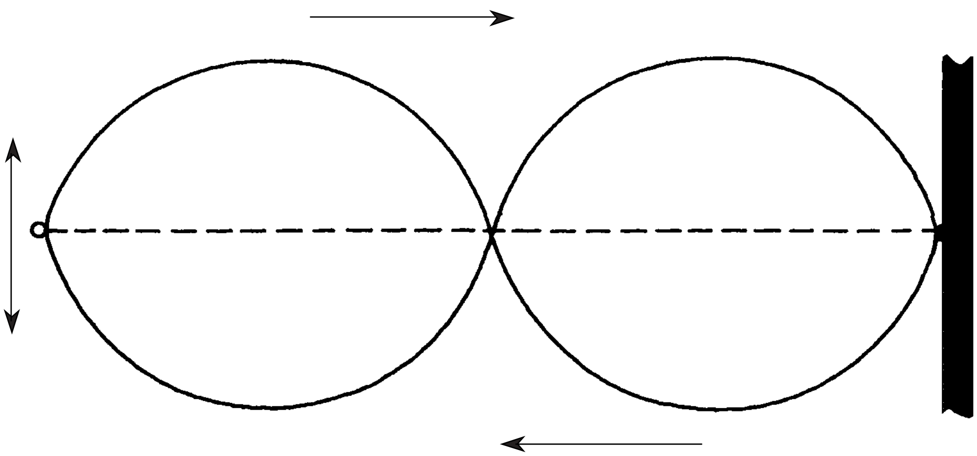

3-6B Standing waves.

flection coefficient varies from –1 to +1, depending upon the magnitude of reflection; Γ = 0 indicates a perfect match with no reflection, while –1 indicates a short-circuited load, and +1 indicates an open circuit. To understand the reflection coefficient, let's start with a basic definition of the resistive load impedance $Z = R + j0$:

$$Z_L = \frac{V}{I} \tag{3.10}$$

where

Z_L is the load impedance $R + j0$
V is the voltage across the load
I is the current flowing in the load

Because there are both reflected and incident waves, we find that V and I are actually the sum of incident and reflected voltages and currents, respectively. Therefore:

$$Z_L = \frac{V}{I} \tag{3.11A}$$

$$Z_L = \frac{V_{\text{inc}} + V_{\text{ref}}}{I_{\text{inc}} + I_{\text{ref}}} \tag{3.11B}$$

where

V_{inc} is the incident (i.e., forward) voltage
V_{ref} is the reflected voltage
I_{inc} is the incident current
I_{ref} is the reflected current

Because of Ohm's law, you can define the currents in terms of voltage, current, and the characteristic impedance of the line:

$$I_{\text{inc}} = \frac{V_{\text{inc}}}{Z_o} \qquad \textbf{[3.12]}$$

and

$$I_{\text{inc}} = \frac{-V_{\text{ref}}}{Z_o} \qquad \textbf{[3.13]}$$

(The minus sign in Eq. 3.13 indicates that a direction reversal has taken place.)

The two expressions for current (Eqs. 3.12 and 3.13) may be substituted into Eq. 3.11 to yield

$$Z_L = \frac{V_{\text{inc}} + V_{\text{ref}}}{\dfrac{V_{\text{inc}}}{Z_o} - \dfrac{V_{\text{ref}}}{Z_o}} \qquad \textbf{[3.14]}$$

The reflection coefficient Γ is defined as the ratio of reflected voltage to incident voltage:

$$\Gamma = \frac{V_{\text{ref}}}{V_{\text{inc}}} \qquad \textbf{[3.15]}$$

Using this ratio in Eq. 3.14 gives

$$\Gamma = \frac{Z_L - Z_o}{Z_L + Z_o} \qquad \textbf{[3.16]}$$

Example 3-3 A 50-Ω transmission line is connected to a 30-Ω resistive load. Calculate the reflection coefficient Γ.

Solution:

$$\Gamma = \frac{Z_L - Z_o}{Z_L + Z_o}$$

$$= \frac{(50\ \Omega) - (30\ \Omega)}{(50\ \Omega) + (30\ \Omega)}$$

$$= \frac{20}{80} = 0.25$$

Example 3-4 In Example 3-3, the incident voltage is 3 V rms. Calculate the reflected voltage.

Solution:

If

$$\Gamma = \frac{V_{\text{ref}}}{V_{\text{inc}}}$$

then

$$\begin{aligned} V_{\text{ref}} &= \Gamma\, V_{\text{inc}} \\ &= (0.25)\ (3\text{ V}) = 0.75\text{ V} \end{aligned}$$

The phase of the reflected signal is determined by the relationship of load impedance and transmission line characteristic impedance. For resistive loads ($Z = R + j0$): if the ratio Z_L/Z_o is 1.0, then there is no reflection; if Z_L/Z_o is less than 1.0, then the reflected signal is 180° out of phase with the incident signal; if the ratio Z_L/Z_o is greater than 1.0 then the reflected signal is in phase with the incident signal. In summary:

Ratio	**Angle of reflection**
$Z_L/Z_o = 1$	No reflection
$Z_L/Z_o < 1$	180°
$Z_L/Z_o > 1$	0°

The step-function (or pulse) response of the transmission line leads to a powerful means of analyzing the line, and its load, on an oscilloscope. Figure 3-7A shows (in schematic form) the test set-up for *time domain reflectometry* (TDR) measurements. An oscilloscope and a pulse (or square-wave) generator are connected in parallel across the input end of the transmission line. Figure 3-7B shows a pulse test jig built by the author for testing lines at HF. The small shielded box contains a TTL square-wave oscillator circuit. Although a crystal oscillator can be used, an *RC* timed circuit running close to 1000 kHz is sufficient. In Fig. 3-7B, you can see the test pulse generator box is connected in parallel with the cable under test and the input of the oscilloscope. A closer look is seen in Fig. 3-7C. A BNC “tee” connector and a double male BNC adapter are used to interconnect the box with the ’scope.

If a periodic waveform is supplied by the generator, then the display on the oscilloscope will represent the sum of reflected and incident pulses. The duration of the pulse (i.e., pulse width), or one-half the period of the square wave, is adjusted so that the returning reflected pulse arrives approximately in the center of the incident pulse.

Figure 3-8 shows a TDR display under several circumstances. Approximately 30 m of coaxial cable, with a velocity factor of 0.66, was used in a test setup similar to Fig. 3-7. The pulse width was approximately 0.9 microseconds (μs). The hori-

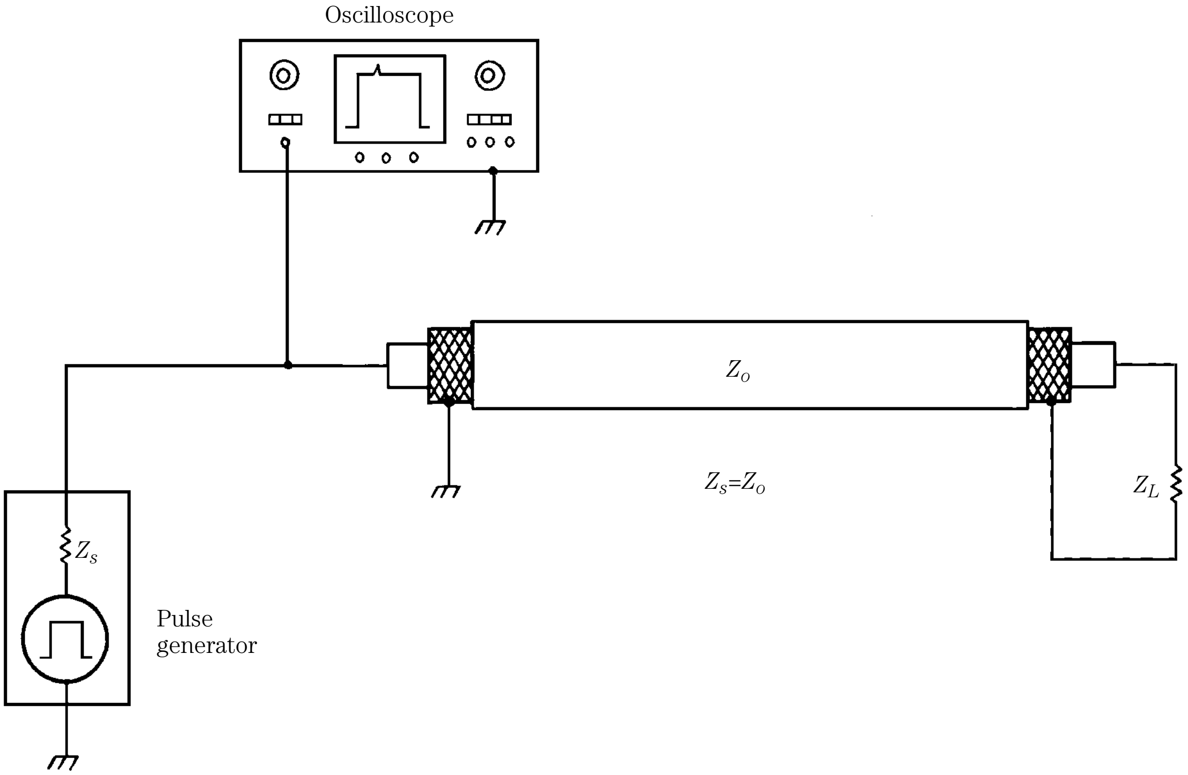
Oscilloscope
Z_o
Z_s=Z_o
Z_L
Z_s
Pulse generator

3-7A Time domain reflectometry setup.

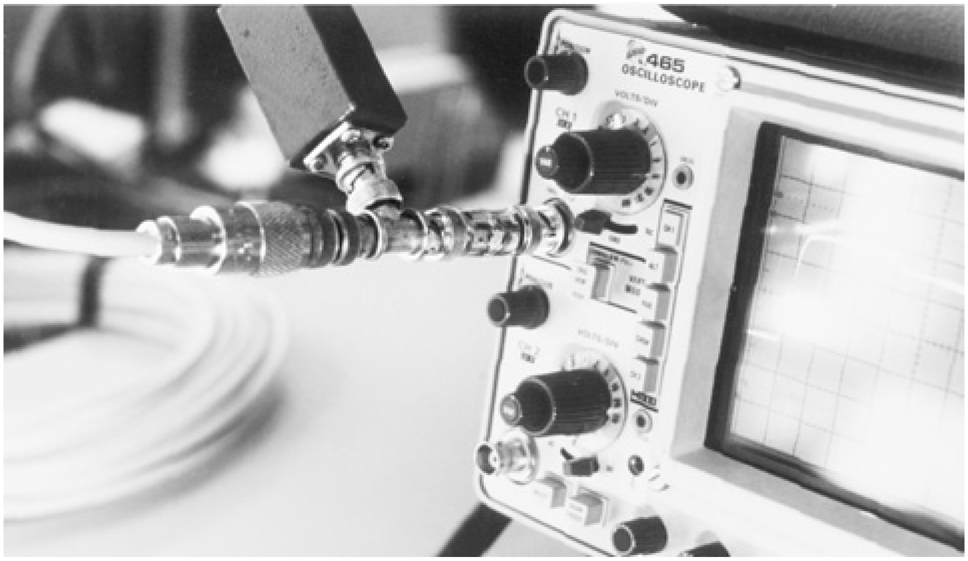
465
OSCILLOSCOPE

3-7B Test setup for impromptu time domain reflectometry.

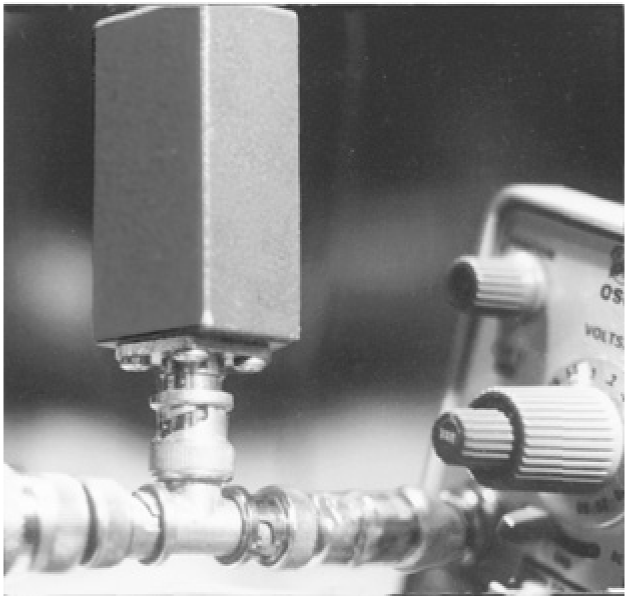

3-7C Close-up of RF connections.

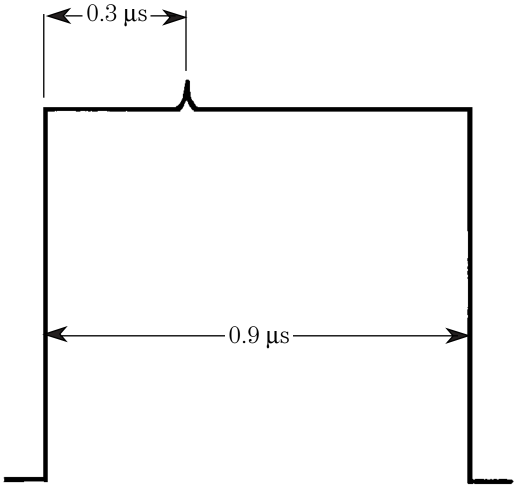

3-8A Idealized TDR pulse. Small "pip" on top is reflected signal interfering with forward pulse.

zontal sweep time on the 'scope was adjusted to show only one pulse—which, in this case, represented one-half of a 550-kHz square wave (Fig. 3-8B).

The displayed trace in Fig. 3-8B shows the pattern when the load is matched to the line ($Z_L = Z_o$). A slight discontinuity exists on the high side of the pulse, and this represents a small reflected wave. Even though the load and line were supposedly matched, the connectors at the end of the line presented a slight impedance discontinuity that shows up on the 'scope as a reflected wave. In general, any discontinuity in the line, any damage to the line, any too-sharp bend, or other anomaly, causes a slight impedance variation, and hence a reflection.

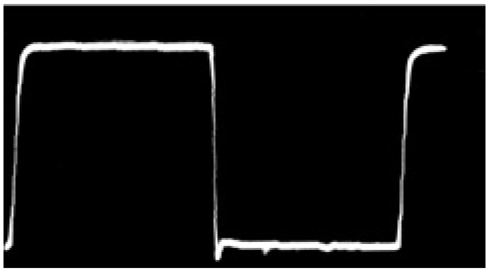

3-8B TDR pulse with no significant reflection.

Notice that the anomaly occurs approximately one-third of the 0.9-µs duration (or 0.3 µs) after the onset of the pulse. This fact tells us that the reflected wave arrives back at the source 0.3 µs after the incident wave leaves. Because this time period represents a round-trip, you can conclude that the wave required 0.3 µs/2, or 0.15 µs to propagate the length of the line. Knowing that the velocity factor is 0.66 for that type of line, you can calculate its approximate length:

$$\begin{aligned}\text{Length} &= cvT \qquad \textbf{[3.17]}\\ &= \frac{(3\times 10^8\ \text{m/s})}{\text{s}} \times (0.66) \times (1.5\times 10^{-7}\ \text{s})\\ &= 29.7\ \text{m}\end{aligned}$$

which agrees "within experimental accuracy" with the 30 m actual length prepared for the experiment ahead of time. Thus, the TDR setup (or a TDR instrument) can be used to measure the length of a transmission line. A general equation is

$$L_{\text{meters}} = \frac{cvT_d}{2} \qquad \textbf{[3.18]}$$

where

L is the length in meters
c is the velocity of light (3×10^8 m/s)
v is the velocity factor of the transmission line
T_d is the round-trip time between the onset of the pulse, and the first reflection

Figures 3-8C through 3-8H show the behavior of the transmission line to the stepfunction when the load impedance is mismatched to the transmission line (Z_L not equal to Z_o). Figure 3-8C shows what happens when the load impedance is less than the line impedance (in this case, $0.5Z_o$). The reflected wave is inverted, and

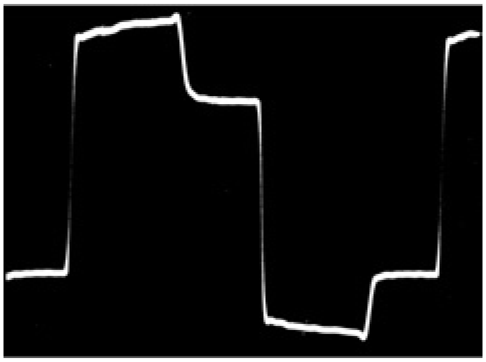

3-8C $Z_L < Z_o$.

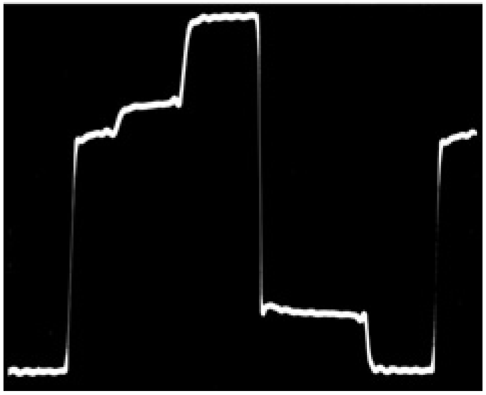

3-8D $Z_L > Z_o$.

sums with the incident wave along the top of the pulse. The reflection coefficient can be determined by examining the relative amplitudes of the two waves.

The opposite situation, in which Z_L is $2Z_o$, is shown in Fig. 3-8D. In this case, the reflected wave is in phase with the incident wave, so it adds to the incident wave as shown. The cases for a short-circuited load and an open-circuited load are shown in Figs. 3-8E and 3-8F, respectively. The cases of reactive loads are shown in Figs. 3-8G and 3-8H. The waveform in Fig. 3-8G resulted from a capacitance in series with a 50-Ω (matched) resistance; the waveform in Fig. 3-8H resulted from a 50-Ω resistance in series with an inductance.

The ac response of the transmission line

When a CW RF signal is applied to a transmission line, the excitation is sinusoidal (Fig. 3-9), so it becomes useful for us to investigate the steady-state ac response of the line. The term *steady-state* implies a sine wave of constant amplitude, phase, and

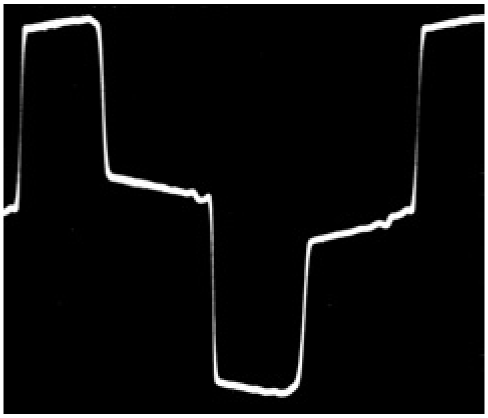

3-8E $Z_L = 0$.

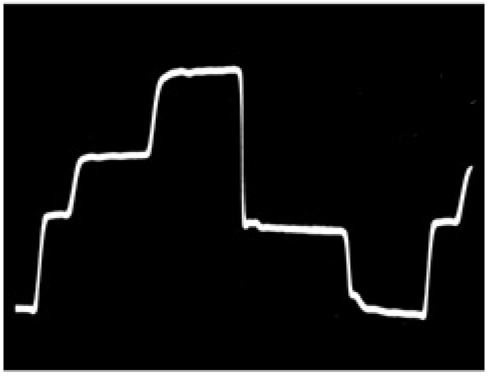

3-8F $Z_L = \infty$.

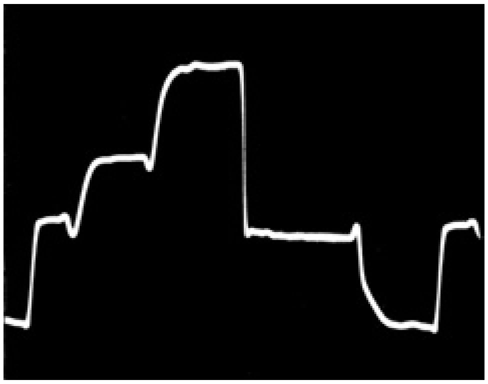

3-8G $Z_L = 50 - jX_C$.

3-8H $Z_L = 50 + jX_L$

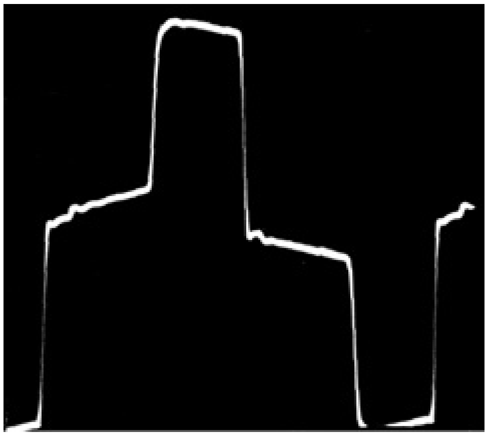

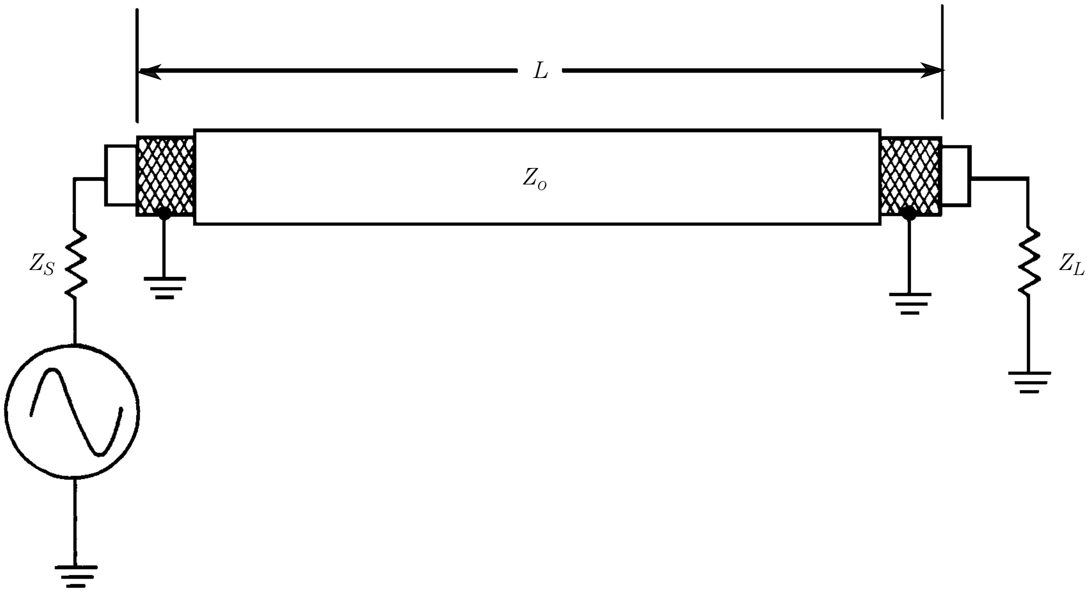

3-9 AC-excited transmission line.

frequency. When ac is applied to the input of the line, it propagates along the line at a given velocity. The ac signal amplitude and phase will decay exponentially:

$$V_R = Ve^{-yl} \quad \textbf{[3.19]}$$

where

V_R is the voltage received at the far end of the line
V is the applied voltage
l is the length of the line
y is the *propagation constant* of the line

The propagation constant y is defined in various equivalent ways, each of which serves to illustrate its nature. For example, the propagation constant is proportional to the product of impedance and admittance characteristics of the line:

$$y = \sqrt{ZY} \qquad \textbf{[3.20]}$$

or, since $Z = R + j\,\omega\,L$ and $Y = G + j\,\omega\,C$, we may write:

$$y = \sqrt{(R + jwl)\ (G + jwC)} \qquad \textbf{[3.21]}$$

You can also write an expression for the propagation constant in terms of the *line attenuation constant* a and *phase constant* B:

$$y = a + jB \qquad \textbf{[3.22]}$$

If you can assume that susceptance dominates conductance in the admittance term, and reactance dominates resistance in the impedance term (both usually true at microwave frequencies), then we may neglect the R and G terms altogether and write:

$$y = j\omega\sqrt{LC} \qquad \textbf{[3.23]}$$

We may also reduce the phase constant B to

$$B = \omega\sqrt{LC} \qquad \textbf{[3.24]}$$

or

$$B = \omega Z_o C \qquad \text{rad/m} \qquad \textbf{[3.25]}$$

and, of course, the characteristic impedance remains:

$$Z_o = \sqrt{LC} \qquad \textbf{[3.26]}$$

Special cases

The impedance "looking into" a transmission line (Z) is the impedance presented to the source by the combination of load impedance and transmission line characteristic impedance. Below are presented equations that define the looking-in impedance seen by a generator (or source) driving a transmission line.

The case where the load impedance, and line characteristic impedance, are matched is defined by

$$Z_L = R_L + \mathrm{j}0 = Z_o$$

In other words, the load impedance is resistive and equal to the characteristic impedance of the transmission line. In this case, the line and load are matched, and the impedance looking in will be a simple $Z = Z_L = Z_o$. In other cases, however, we find different situations where Z_L is not equal to Z_o.

1. Z_L is not equal to Z_o in a random-length lossy line:

$$Z = (Z_o)\left[\frac{Z_L + Z_o \tanh (yl)}{Z_o + Z_L \tanh (yl)}\right] \qquad \textbf{[3.27]}$$

where

Z is the impedance looking in, in ohms
Z_L is the load impedance, in ohms
Z_o is the line characteristic impedance, in ohms
l is the length of the line, in meters
y is the propagation constant

2. Z_L not equal to Z_o in a lossless, or very low loss, random-length line:

$$Z = (Z_o)\left[\frac{Z_L + jZ_o \tan (Bl)}{Z_o + jZ_L \tan (Bl)}\right] \qquad \textbf{[3.28]}$$

Equations 3.27 and 3.28 serve for lines of any random length. For lines that are either integer multiples of a half-wavelength, or odd-integer (i.e., 1, 3, 5, 7, . . ., etc.) multiples of a quarter-wavelength, special solutions for these equations are found—and some of these solutions are very useful in practical situations. For example, consider . . .

3. Half-wavelength lossy lines:

$$Z = (Z_o)\left[\frac{Z_L + Z_o \tanh (al)}{Z_o + Z_L \tanh (al)}\right] \qquad \textbf{[3.29]}$$

Example 3-5 A lossless 50-Ω (Z_o) transmission line is exactly one-half wavelength long and is terminated in a load impedance of $Z = 30 + j0$. Calculate the input impedance looking into the line. (*Note*: in a lossless line $a = 0$.)

Solution:

$$Z = (Z_o)\left[\frac{Z_L + Z_o \tanh (al)}{Z_o + Z_L \tanh (al)}\right]$$

$$Z = (50\ \Omega)\left\{\frac{30 + 50 \tanh[(0)(\pi)]}{50 + 30 \tanh[(0)(\pi)]}\right\}$$

$$= (50\ \Omega)\left[\frac{30 + 50 \tanh(0)}{50 + 30 \tanh(0)}\right]$$

$$= (50\ \Omega)\left(\frac{30 + 0}{50 + 0}\right)$$

$$= (50\ \Omega)\ (30/50) = 30\ \Omega$$

In Example 3-5 we discovered that the impedance looking into a lossless (or very low loss) half-wavelength transmission line is the load impedance:

$$Z = Z_L \qquad \textbf{[3.30]}$$

The fact that line input impedance equals load impedance, is very useful in certain practical situations. For example, a resistive impedance is not changed by the line length. Therefore, when an impedance is inaccessible for measurement purposes, the impedance can be measured through a transmission line that is an integer multiple of a half wavelength.

Our next special case involves a quarter-wavelength transmission line, and those that are *odd* integer multiples of quarter-wavelengths (of course, *even* integer multiples of a quarter wavelength obey the half-wavelength criteria).

4. Quarter-wavelength lossy lines:

$$Z = (Z_o)\left[\frac{Z_L + Z_o \coth(al)}{Z_o + Z_L \coth(al)}\right] \qquad \textbf{[3.31]}$$

and

5. Quarter-wavelength lossless or very low loss lines:

$$Z = \frac{(Z_o)^2}{Z_L} \qquad \textbf{[3.32]}$$

From Eq. 3.32, you can discover an interesting property of the quarter-wavelength transmission line. First, divide each side of the equation by Z_o:

$$\frac{Z}{Z_o} = \frac{[Z_o]^2}{Z_L Z_o} \qquad \textbf{[3.33]}$$

$$= \frac{Z_o}{Z_L} \qquad \textbf{[3.34]}$$

The ratio Z/Z_o shows an inversion of the load impedance ratio Z_L/Z_o, or, stated another way,

$$\frac{Z}{Z_o} = \frac{1}{Z_L/Z_o} \qquad \textbf{[3.35]}$$

Again, from Eq. 3.32, you can deduce another truth about quarter-wavelength transmission lines:
If

$$Z = \frac{\sqrt{Z_o}}{Z_L} \qquad \textbf{[3.36]}$$

then

$$ZZ_L = \sqrt{Z_o} \qquad \textbf{[3.37]}$$

which means

$$Z_o = \sqrt{ZZ_L} \qquad \textbf{[3.38]}$$

Equation 3.38 shows that a quarter-wavelength transmission line can be used as an *impedance matching network*. Called a *Q section*, the quarter-wavelength transmission line used for impedance matching requires a characteristic impedance Z_o—if Z is the source impedance and Z_L is the load impedance.

Example 3-6 A 50-Ω source must be matched to a load impedance of 36 Ω. Find the characteristic impedance required of a Q section matching network.

Solution:

$$\begin{aligned} Z &= \sqrt{ZZ_L} \\ &= \sqrt{(50\ \Omega)\ (36\ \Omega)} \\ &= \sqrt{1800\ \Omega^2} = 42\ \Omega \end{aligned}$$

6. Transmission line as a reactance: Reconsider Eq. 3.28, which related impedance looking in to load impedance and line length:

$$Z = (Z_o)\left[\frac{Z_L + jZ_o \tan(Bl)}{Z_o + jZ_L \tan(Bl)}\right] \qquad \textbf{[3.39]}$$

Now, for the case of a shorted line (i.e., $Z_L = 0$), the solution is

$$Z = (Z_o)\left[\frac{(0) + jZ_o \tan(B\,l)}{Z_o + j(0) \tan(B\,l)}\right] \qquad \textbf{[3.40]}$$

$$Z = (Z_o)\left[\frac{jZ_o \tan(Bl)}{Z_o}\right] \qquad \textbf{[3.41]}$$

$$Z = j\,Z_o \tan(Bl) \qquad \textbf{[3.42]}$$

Recall from Eq. 3.25 that

$$B = WZ_oC \qquad \textbf{[3.43]}$$

Substituting Eq. 3.43 into Eq. 3.42 produces

$$Z = j\,Z_o \tan(\omega Z_oCl) \qquad \textbf{[3.44]}$$

or

$$Z = j\,Z_o \tan(2\pi FZ_oCl) \qquad \textbf{[3.45]}$$

Because the solutions to Eqs. 3.44 and 3.45 are multiplied by the j operator, the impedance is actually a reactance ($Z = 0 + jX$). It is possible to achieve almost any possible reactance (within certain practical limitations) by adjusting the length of the transmission line and shorting the "load" end. This fact leads us to a practical method for impedance matching.

Figure 3-10A shows a circuit in which an unmatched load is connected to a transmission line with characteristic impedance Z_o. The load impedance Z_L is of the form $Z = R \pm jX$, and, in this case, it is equal to $50 - j20$.

A complex impedance load can be matched to its source by interposing the complex conjugate of the impedance. For example, in the case where $Z = 50 - j20$, the matching impedance network will require an impedance of $50 + j20\ \Omega$. The two impedances combine to produce a result of $50\ \Omega$. The situation of Fig. 3-10A shows a matching stub with a reactance equal in magnitude, but opposite in sign, with respect to the reactive component of the load impedance. In this case, the stub has a reactance of $+j20\ \Omega$ to cancel a reactance of $-j20\ \Omega$ in the load.

A *quarter-wavelength shorted stub* is a special case of the stub concept that finds particular application in microwave circuits. *Waveguides* (Chap. 19) are based on the properties of the quarter-wavelength shorted stub. Figure 3-10B shows a quarter-wave stub and its current distribution. The current is maximum across the short, but wave cancellation forces it to zero at the terminals. Because $Z = V/I$, when I goes to zero, the impedance becomes infinite. Thus, a quarter-wavelength stub has an infinite impedance at its resonant frequency, and redundant acts as an insulator. This concept may be hard to swallow, but the stub is a "metal insulator."

Standing wave ratio

The reflection phenomenon was noted earlier during the coverage of the step-function and single-pulse response of a transmission line; the same phenomenon also applies when the transmission line is excited with an ac signal. When a transmission

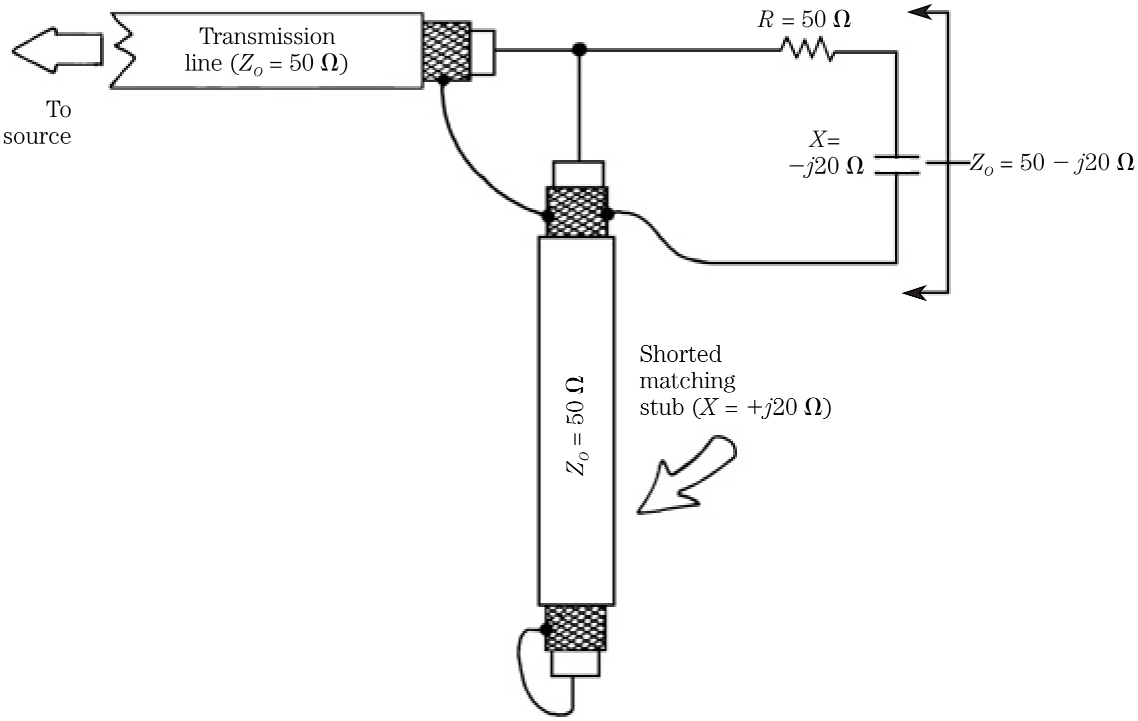

3-10A Stub matching system.

line is not matched to its load, some of the energy is absorbed by the load and some is reflected back down the line toward the source. The interference of incident (or "forward") and reflected (or "reverse") waves creates *standing waves* on the transmission line.

If the voltage or current is measured along the line, it will vary, depending on the load, according to Fig. 3-11. Figure 3-11A shows the voltage-versus-length curve for a matched line (i.e., where $Z_L = Z_o$). The line is said to be "flat" because the voltage (and current) is constant all along the line. But now consider Figs. 3-11B and 3-11C.

Figure 3-11B shows the voltage distribution over the length of the line when the load end of the line is *shorted* (i.e., $Z_L = 0$). Of course, at the load end the voltage is zero, which results from zero impedance. The same impedance and voltage situation is repeated every half-wavelength down the line from the load end toward the generator. Voltage *minima* are called *nodes*, and voltage *maxima* are called *antinodes*.

The pattern in Fig. 3-11C results when the line is unterminated (open) (i.e., $Z_L = \infty$). Note that the pattern is the same shape as Fig. 3-11B (shorted line), but the phase is shifted 90°. In both cases, the reflection is 100 percent, but the phase of the reflected wave is opposite.

Figure 3-11D shows the situation in which Z_L is not equal to Z_o, but is neither zero nor infinite. In this case, the nodes represent some finite voltage, $V_{\min}$, rather

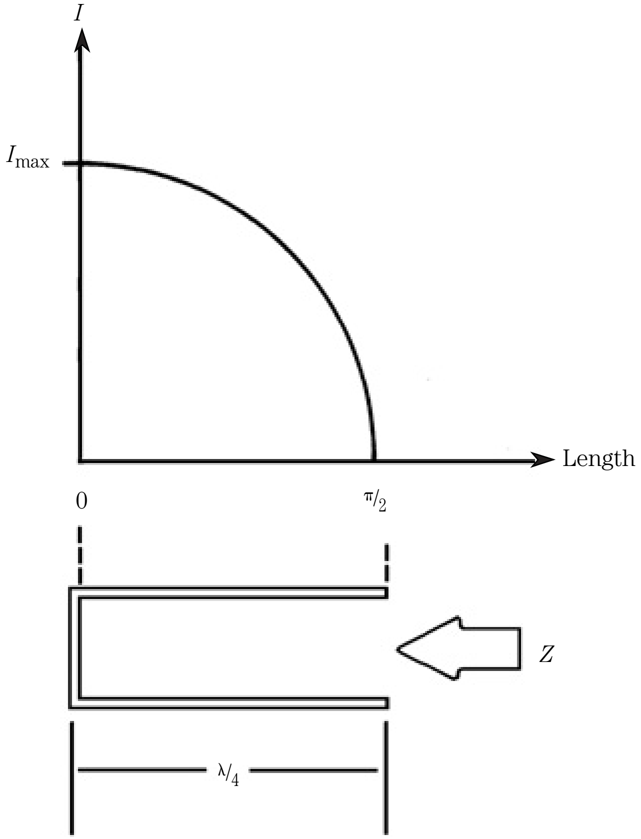

3-10B Quarter-wavelength stub.

than zero. The *standing wave ratio* (SWR) reveals the relationship between load and line.

If the current along the line is measured, the pattern will resemble the patterns of Fig. 3-11. The SWR is then called ISWR, to indicate the fact that it came from a current measurement. Similarly, if the SWR is derived from voltage measurements it is called VSWR. Perhaps because voltage is easier to measure, VSWR is the term most commonly used in most radio work.

VSWR can be specified in any of several equivalent ways:

1. *From incident voltage (V_i) and reflected voltage (V_i):*

$$\text{VSWR} = \frac{V_i + V_r}{V_i - V_r} \qquad \textbf{[3.46]}$$

2. *From transmission line voltage measurements (Fig. 3-11D):*

$$\text{VSWR} = \frac{V_{\text{max}}}{V_{\text{min}}} \qquad \textbf{[3.47]}$$

3. *From load and line characteristic impedances:*

$$(Z_L > Z_o)\ \text{VSWR} = Z_L/Z_o \qquad \textbf{[3.48]}$$

$$(Z_L < Z_o)\ \text{VSWR} = Z_o/Z_L \qquad \textbf{[3.49]}$$

4. *From incident (P_i) and reflected (P_r) power:*

$$\text{VSWR} = \frac{1 + \sqrt{P_r/P_i}}{1 - \sqrt{P_r/P_i}} \qquad \textbf{[3.50]}$$

5. *From reflection coefficient (Γ):*

$$\text{VSWR} = \frac{1 + \Gamma}{1 - \Gamma} \qquad \textbf{[3.51]}$$

It is also possible to determine the reflection coefficient Γ from a knowledge of VSWR:

$$\Gamma = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \qquad \textbf{[3.52]}$$

The relationship between reflection coefficient Γ and VSWR is shown in Fig. 3-11D.

VSWR is usually expressed as a *ratio*. For example, when Z_L is 100 Ω and Z_o is 50 Ω, the VSWR is $Z_L/Z_o = 100\ \Omega/50\ \Omega = 2$, which is usually expressed as VSWR = 2:1. VSWR can also be expressed in decibel form:

$$\text{VSWR} = 20 \log (\text{VSWR}) \qquad \textbf{[3.53]}$$

Example 3-7 A transmission line is connected to a mismatched load. Calculate both the VSWR and VSWR decibel equivalent if the reflection coefficient Γ is 0.25.

Solution:

$$(a)\ \text{VSWR} = \frac{1 + \Gamma}{1 - \Gamma}$$

$$= \frac{1 + 0.25}{1 - 0.25}$$

$$= \frac{1.25}{0.75} = 1.67{:}1$$

$$(b)\ \text{VSWR}_{\text{dB}} = 20 \log (\text{VSWR})$$

$$= (20)\ (\log 1.67)$$

$$= (20)\ (0.22) = 4.3\ \text{dB}$$

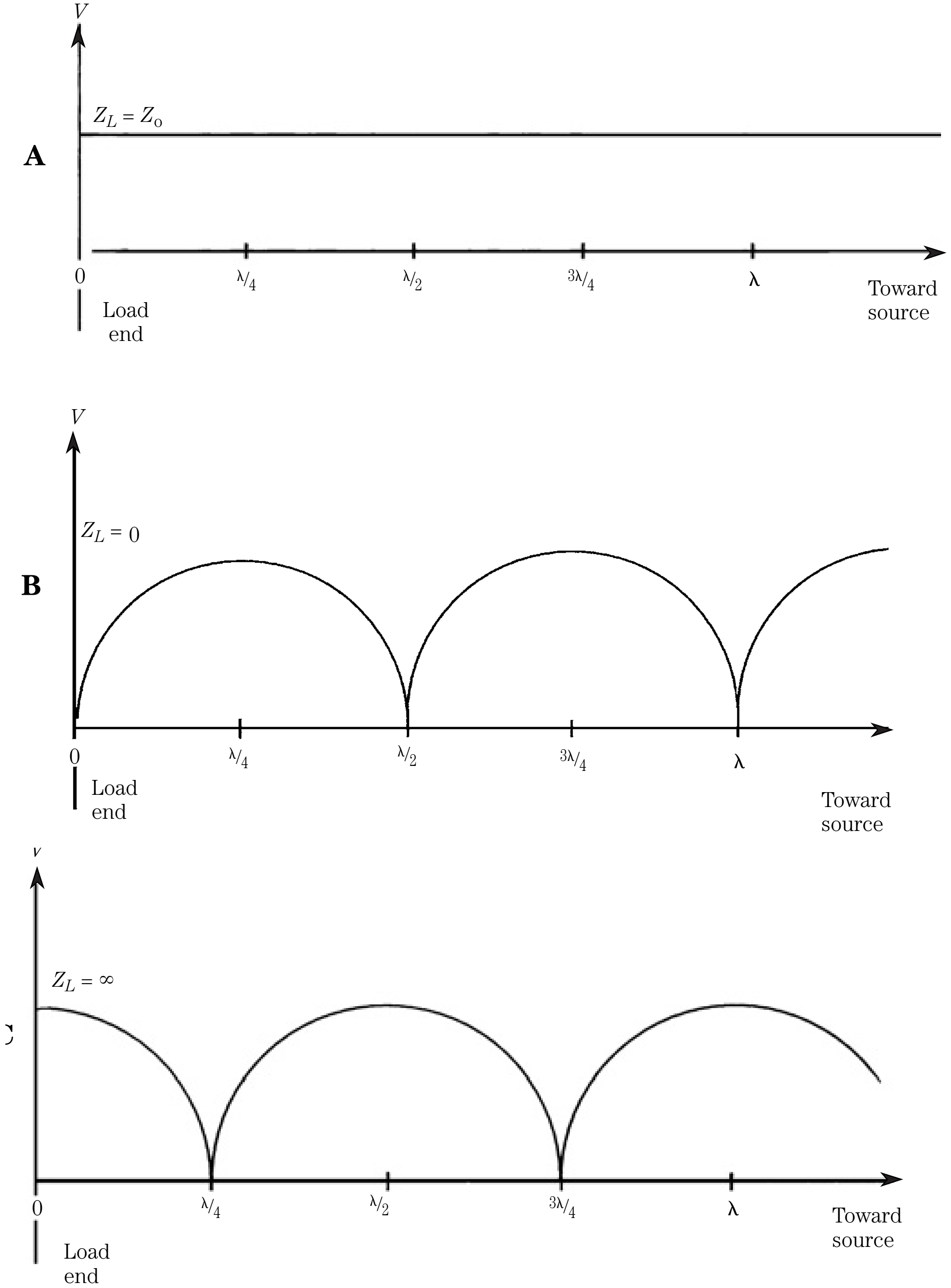

3-11 Voltage versus electrical length: (A) Matched impedances, (B) $Z_L = 0$, (C) Z_L = infinite.

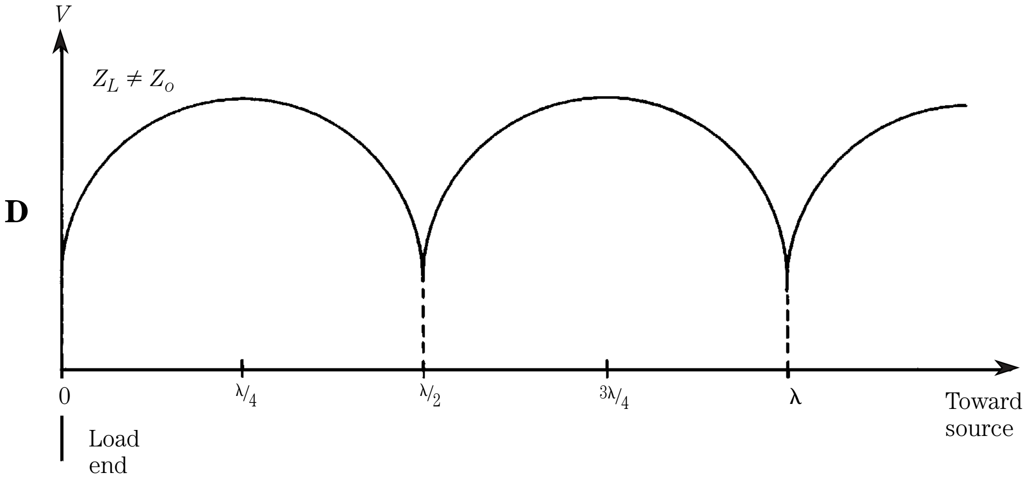

3-11 Continued: (D) Z_L not equal to Z_o.

The SWR is regarded as important in systems for several reasons. The base of these reasons is the fact that the reflected wave represents energy lost to the load. For example, in an antenna system, less power is radiated if some of its input power is reflected back down the transmission line, because the antenna feedpoint impedance does not match the transmission line characteristic impedance. The next section covers the problem of mismatch losses.

Mismatch (VSWR) losses

The power reflected from a mismatched load represents a loss, and will have implications that range from negligible to profound, depending on the situation. For example, one result might be a slight loss of signal strength at a distant point from an antenna. A more serious problem can result in the destruction of the output device in a transmitter. The latter problem so plagued early solid-state transmitters that designers opted to include shutdown circuitry to sense high VSWR, and turn down output power proportionally.

In microwave measurements, VSWR on the transmission lines (that interconnect devices under test, instruments, and signal sources) can cause erroneous readings—and invalid measurements.

Determination of VSWR losses must take into account *two* VSWR situations. Figure 3-9 shows a transmission line of impedance Z_o interconnecting a load impedance Z_L, and a source with an output impedance Z_s. There is a potential for impedance mismatch at *both ends* of the line.

In the case where one end of the line is matched (either Z_s or Z_L), the *mismatch loss* caused by SWR at the mismatched end is

$$\mathrm{ML} = -10 \log \left[1 - \left(\frac{\mathrm{SWR} - 1}{\mathrm{SWR} + 1} \right)^2 \right] \qquad \textbf{[3.54]}$$

which from Eq. 3.52 is

$$\mathrm{ML} = -10 \log (1 - \Gamma^2) \qquad \textbf{[3.55]}$$

Example 3-8 A coaxial transmission line with a characteristic impedance of 50 Ω is connected to the 50-Ω output (Z_o) of a signal generator, and also to a 20-W load impedance Z_L. Calculate the mismatch loss.

Solution:

(*a*) First find the VSWR:

$$\begin{aligned}\mathrm{VSWR} &= Z_o/Z_L \\ &= (50\ \Omega)/(20\ \Omega) = 2.5{:}1\end{aligned}$$

(*b*) Mismatch loss:

$$\begin{aligned}\mathrm{ML} &= -10 \log \left[1 - \left(\frac{\mathrm{SWR} - 1}{\mathrm{SWR} + 1} \right)^2 \right] \\ &= -10 \log \left[1 - \left(\frac{2.5 - 1}{2.5 + 1} \right)^2 \right] \\ &= -10 \log \left[1 - \left(\frac{1.5}{3.5} \right)^2 \right] \\ &= -10 \log [1 - (0.43)^2] \\ &= -10 \log [1 - 0.185] \\ &= -10 \log [0.815] \\ &= (-10)\ (-0.089) = 0.89\end{aligned}$$

When both ends of the line are mismatched, a different equation is required:

$$\mathrm{ML} = 20 \log [1 \pm (\Gamma_1 \times \Gamma_2)] \qquad \textbf{[3.56]}$$

where

Γ_1 is the reflection coefficient at the source end of the line, $(VSWR_1 - 1)/(VSWR_1 + 1)$

Γ_2 is the reflection coefficient at the load end of the line, $(VSWR_2 - 1)/(VSWR_2 + 1)$

Note that the solution to Eq. 3.56 has two values: $[1 + (\Gamma_1\Gamma_2)]$ and $[1 - (\Gamma_1\Gamma_2)]$.

The equations reflect the mismatch loss solution for low-loss or "lossless" transmission lines. This is a close approximation in some situations; however, it is insufficient when the line is lossy. Although not very important at low frequencies, loss becomes higher at microwave frequencies. Interference between incident and reflected waves produces increased current at certain antinodes—which increases ohmic losses—and increased voltage at certain antinodes—which increases dielectric losses. It is the latter that increases with frequency. Equation 3.57 relates reflection coefficient and line losses to determine total loss on a given line.

$$\text{Loss} = 10 \log \left(\frac{n^2 - \Gamma^2}{n - n\Gamma^2} \right) \qquad \textbf{[3.57]}$$

where

Loss is the total line loss in decibels
Γ is the reflection coefficient
n is the quantity $10^{(A/10)}$
A is the total attenuation presented by the line, in decibels, when the line is properly matched ($Z_L = Z_o$)

Example 3-9 A 50-Ω transmission line is terminated in a 30-Ω resistive impedance. The line is rated at a loss of 3 dB/100 ft at 1 GHz. Calculate (a) loss in 5 ft of line, (*b*) reflection coefficient, and (*c*) total loss in a 5-ft line mismatched per above.

Solution:

$$(a)\ A = \frac{3\text{ dB}}{100\text{ ft}} \times 5\text{ ft} = 0.15\text{ dB}$$

$$(b)\ \Gamma = \frac{Z_L - Z_o}{Z_L + Z_o}$$

$$= \frac{50 - 30}{50 + 30}$$

$$= 20/80 = 0.25$$

$$(c)\ n = 10^{(A/10)}$$

$$= 10^{(0.15/10)}$$

$$= 10^{(0.015)} = 1.04$$

$$\text{Loss} = 10 \log \left(\frac{n^2 - \Gamma^2}{n - n\Gamma^2} \right)$$

$$= 10 \log \left[\frac{(1.04)^2 - (0.25)^2}{1.04 - (1.04)\ (0.25)^2} \right]$$

$$= 10 \log \left[\frac{1.082 - 0.063}{1.04 - (1.04)\ (0.063)} \right]$$

$$= 10 \log \left(\frac{1.019}{1.04 - 0.066} \right)$$

$$= 10 \log \left(\frac{1.019}{0.974} \right)$$

$$= 10 \log (1.046)$$

$$= (10)\ (0.02) = 0.2 \text{ dB}$$

Compare the matched line loss ($A = 0.15$ dB) with the total loss (Loss = 0.2 dB), which includes mismatch loss and line loss. The difference (i.e., Loss – A) is only 0.05 dB. If the VSWR was considerably larger, however, the loss would rise.

4
CHAPTER

The Smith chart

THE MATHEMATICS OF TRANSMISSION LINES, AND CERTAIN OTHER DEVICES, BECOMES cumbersome at times, especially in dealing with complex impedances and "nonstandard" situations. In 1939, Phillip H. Smith published a graphical device for solving these problems, and an improved version of the chart followed in 1945. That graphic aid, somewhat modified over time, is still in constant use in microwave electronics, and other fields where complex impedances and transmission line problems are found. The *Smith chart* is indeed a powerful tool for the RF designer.

Smith chart components

The modern Smith chart is shown in Fig. 4-1. It consists of a series of overlapping orthogonal circles (i.e., circles that intersect each other at right angles). This chapter will dissect the Smith chart, so that the origin and use of these circles is apparent. The set of orthogonal circles makes up the basic structure of the Smith chart.

The normalized impedance line

A baseline is highlighted in Fig. 4-2, and bisects the Smith chart outer circle. This line is called the *pure resistance line*, and forms the reference for measurements made on the chart. Recall that a complex impedance contains both resistance and reactance, and is expressed in the mathematical form:

$$Z = R \pm jX \quad \textbf{[4.1]}$$

Copyright 2001 - the McGraw-Hill Companies

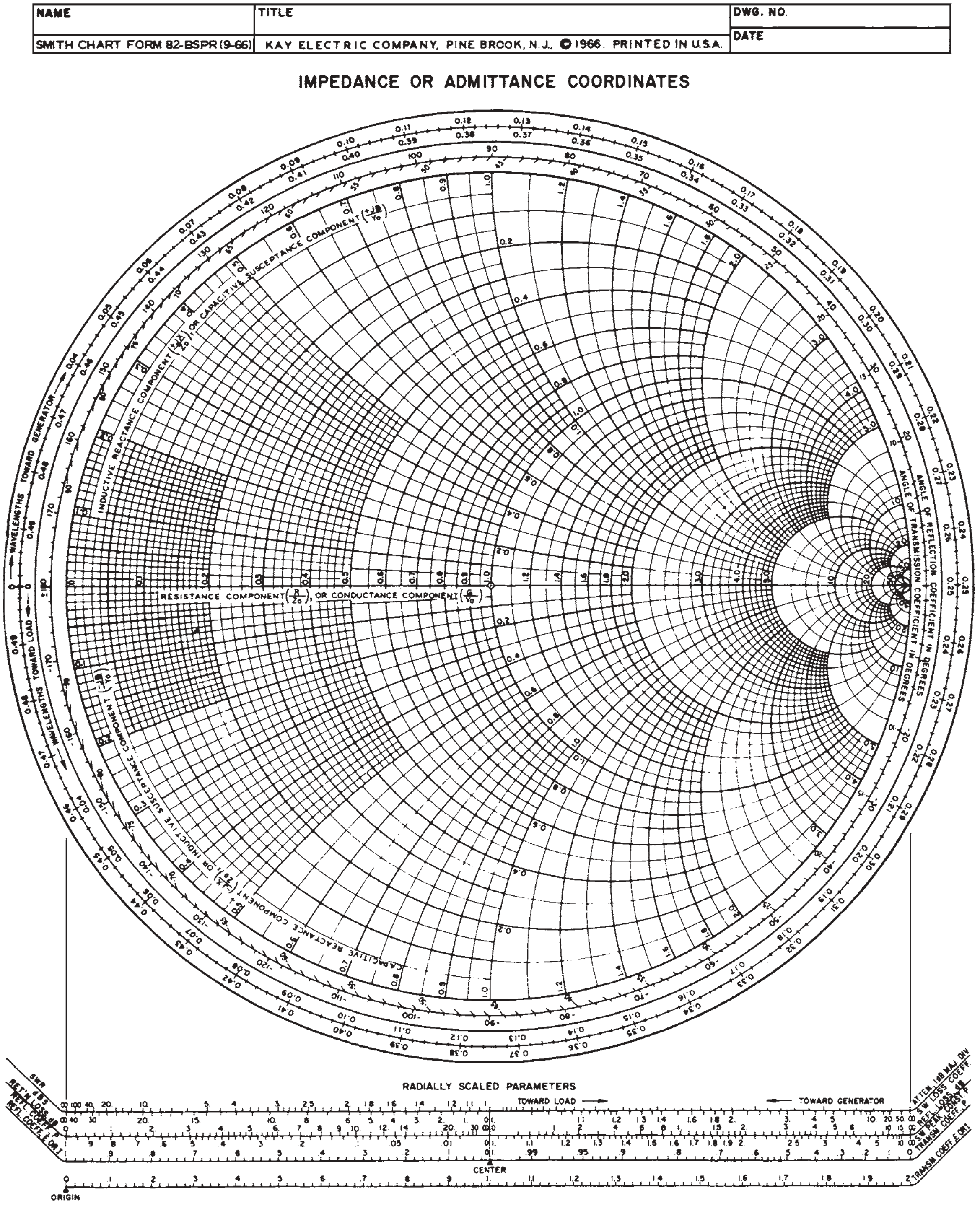

4-1 The Smith chart. (*Courtesy of Kay Elementrics*)

where

Z is the complex impedance
R is the resistive component of the impedance
X is the reactive component of the impedance*

The pure resistance line represents the situation where $X = 0$, and the impedance is therefore equal to the resistive component only. In order to make the Smith chart universal, the impedances along the pure resistance line are *normalized* with reference to system impedance (e.g., Z_o in transmission lines); for most microwave RF systems the system impedance is standardized at 50 Ω. In order to normalize the actual impedance, divide it by the system impedance. For example, if the load impedance of a transmission line is Z_L, and the characteristic impedance of the line is Z_o, then $Z = Z_L/Z_o$. In other words,

$$Z = \frac{R \pm jX}{Z_o} \qquad \textbf{[4.2]}$$

The pure resistance line is structured such that the system standard impedance is in the center of the chart, and has a normalized value of 1.0 (see point A in Fig. 4-2). This value derives from the fact that $Z_o/Z_o = 1.0$.

To the left of the 1.0 point are decimal fraction values used to denote impedances less than the system impedance. For example, in a 50-Ω transmission line system with a 25-Ω load impedance, the normalized value of impedance is 25 Ω/50 Ω or 0.50 ("B" in Fig. 4-2). Similarly, points to the right of 1.0 are greater than 1 and denote impedances that are higher than the system impedance. For example, in a 50-Ω system connected to a 100-Ω resistive load, the normalized impedance is 100 Ω/50 Ω, or 2.0; this value is shown as point C in Fig. 4-2. By using normalized impedances, you can use the Smith chart for almost any practical combination of system, and load and/or source, impedances, whether resistive, reactive, or complex.

Reconversion of the normalized impedance to actual impedance values is done by multiplying the normalized impedance by the system impedance. For example, if the resistive component of a normalized impedance is 0.45, then the actual impedance is

$$Z = (Z_{\text{normal}})\,(Z_o) \qquad \textbf{[4.3]}$$

$$= (0.45)\,(50\ \Omega) \qquad \textbf{[4.4]}$$

$$= 22.5\ \Omega \qquad \textbf{[4.5]}$$

The constant resistance circles

The *isoresistance circles*, also called the *constant resistance circles*, represent points of equal resistance. Several of these circles are shown highlighted in Fig. 4-3. These circles are all tangent to the point at the righthand extreme of the pure resis-

---

*According to the standard sign convention the inductive reactance (X_L) is positive (+) and the capacitive reactance (X_c) is negative (–). The term X in Eq. 4.1 above is the difference between the two reactances ($X = X_L - X_c$).

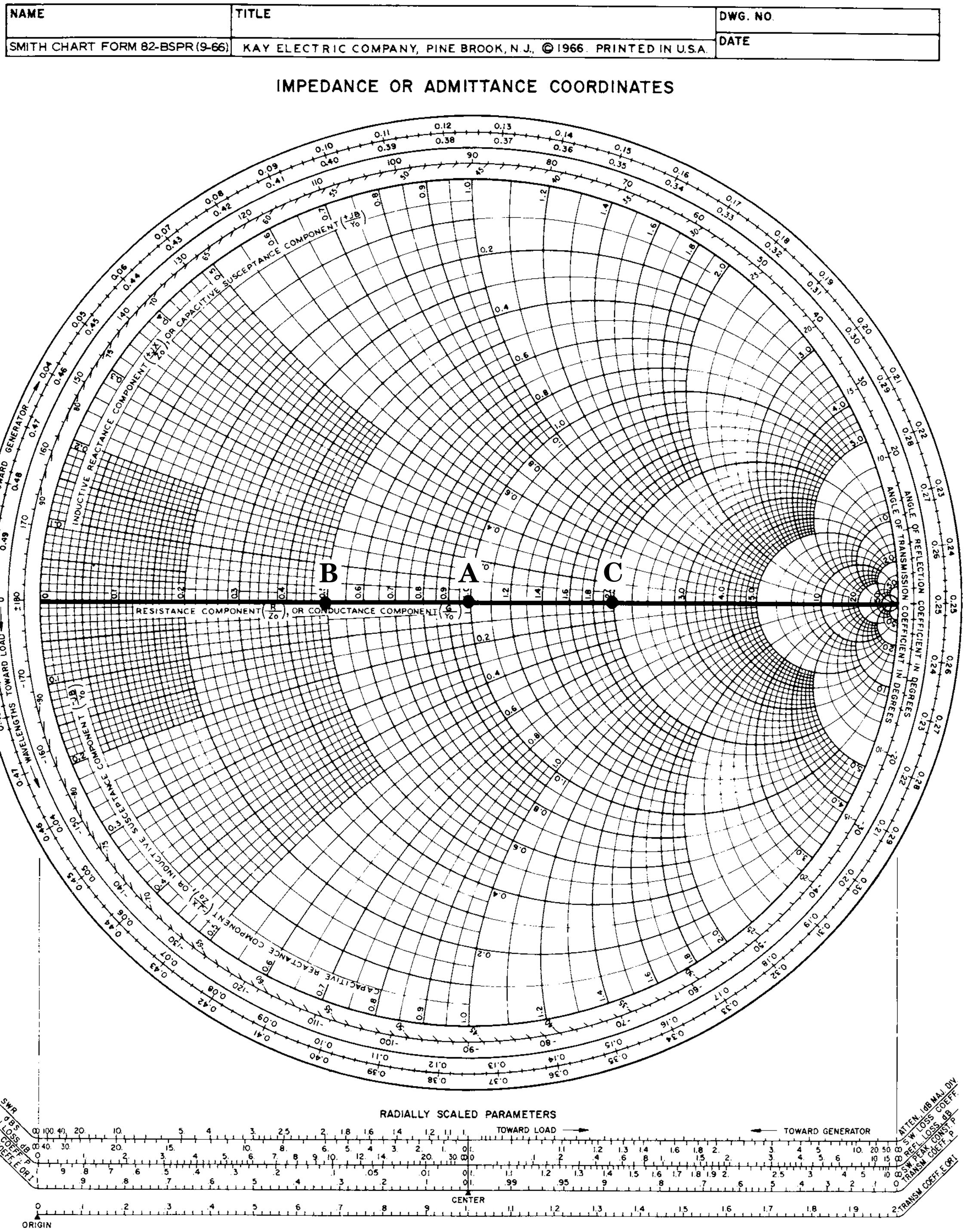
NAME
TITLE
DWG. NO.
SMITH CHART FORM 82-BSPR (9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J., © 1966. PRINTED IN U.S.A.
DATE
IMPEDANCE OR ADMITTANCE COORDINATES
A
B
C
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE COMPONENT (-jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
SWR
dBS
ATTEN. 1 dB MAJ. DIV.
S.W. LOSS COEFF.
REFL. LOSS dB
S.W. PEAK (CONST. P)
TRANSM. COEFF. P
CENTER
ORIGIN
TRANSM. COEFF. E OR I

4-2 Normalized impedance line. (*Courtesy of Kay Elementrics*)

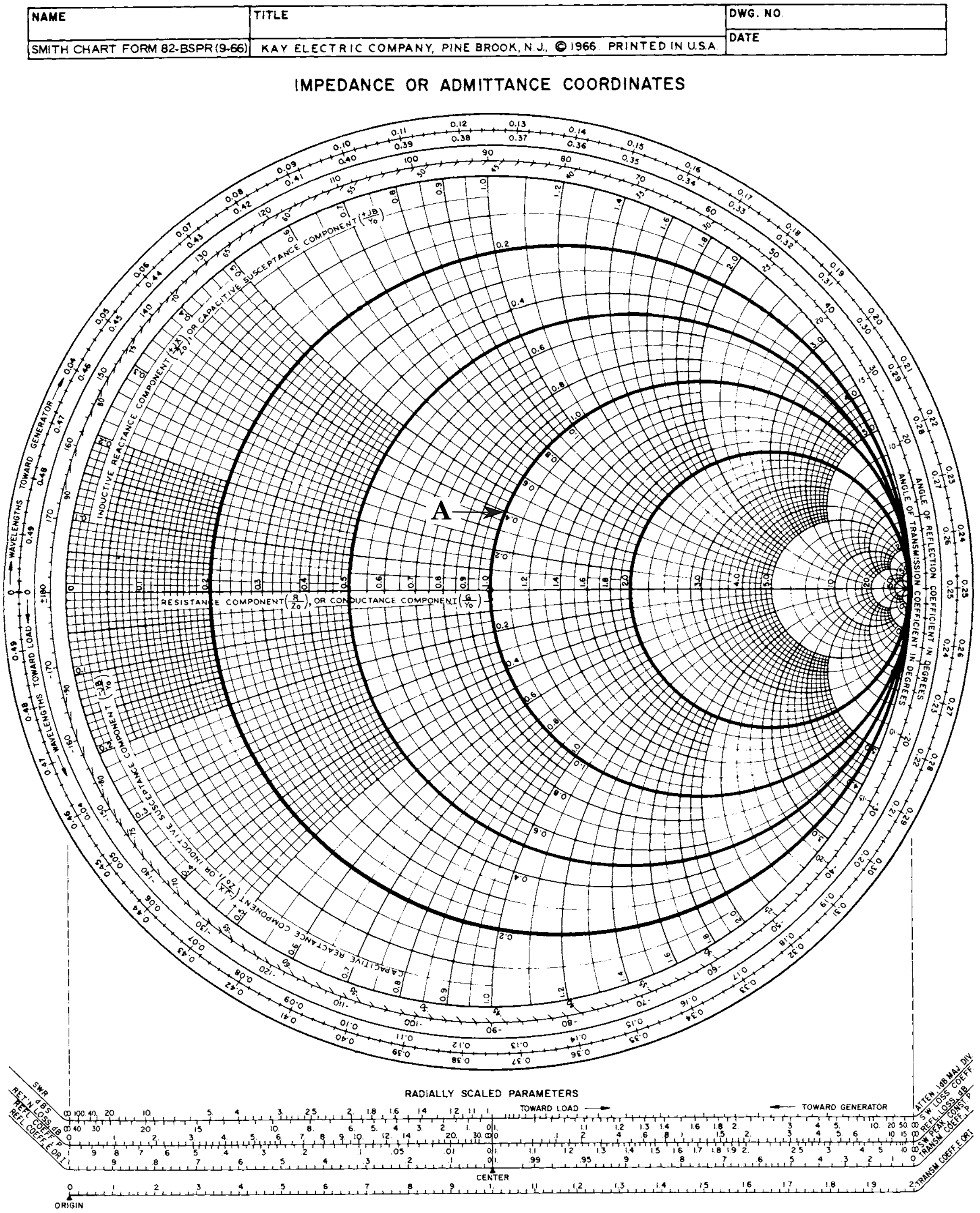
NAME
TITLE
DWG. NO.
DATE
SMITH CHART FORM 82-BSPR(9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J. © 1966 PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
A
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE COMPONENT (-jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
SWR
dBS
RET'N LOSS, dB
REFL. COEFF, P
REFL. COEFF, E OR I
ATTEN. 1dB MAJ. DIV.
S.W. LOSS COEFF
S.W. PEAK CONST. P
TRANSM. COEFF, E OR I
CENTER
ORIGIN

4-3 Constant resistance circles. (*Courtesy of Kay Elementrics*) tance line, and are bisected by that line. When you construct complex impedances (for which X = nonzero) on the Smith chart, the points on these circles will all have the same resistive component. Circle *A*, for example, passes through the center of the chart, so it has a normalized constant resistance of 1.0. Note that impedances that are pure resistances (i.e., $Z = R + j0$) will fall at the intersection of a constant resistance circle and the pure resistance line, and complex impedances (i.e., X not equal to zero) will appear at any other points on the circle. In Fig. 4-3, circle *A* passes through the center of the chart, so it represents all points on the chart with a normalized resistance of 1.0. This particular circle is sometimes called the *unity resistance circle*.

The constant reactance circles

Constant reactance circles are highlighted in Fig. 4-4. The circles (or circle segments) *above* the pure resistance line (Fig. 4-4A) represent the *inductive reactance* ($+X$), and those circles (or segments) *below* the pure resistance line (Fig. 4-4B) represent *capacitive reactance* ($-X$). In both cases, circle *A* represents a normalized reactance of 0.80.

One of the outer circles (i.e., circle *A* in Fig. 4-4C) is called the *pure reactance circle*. Points along circle *A* represent reactance only; in other words, an impedance of $Z = 0 \pm jX$ ($R = 0$).

Figure 4-4D shows how to plot impedance and admittance on the Smith chart. Consider an example in which system impedance Z_o is 50 Ω, and the load impedance is $Z_L = 95 + j55\ \Omega$. This load impedance is normalized to

$$Z = \frac{Z_L}{Z_o} \qquad \textbf{[4.6]}$$

$$= \frac{95 + j55\ \Omega}{50\ \Omega} \qquad \textbf{[4.7]}$$

$$= 1.9 + j1.1 \qquad \textbf{[4.8]}$$

An *impedance radius* is constructed by drawing a line from the point represented by the normalized load impedance, $1.9 + j1.1$, to the point represented by the normalized system impedance (1.0) in the center of the chart. A circle is constructed from this radius, and is called the *VSWR circle.*

Admittance is the reciprocal of impedance, so it is found from

$$Y = \frac{1}{Z} \qquad \textbf{[4.9]}$$

Because impedances in transmission lines are rarely pure resistive, but rather contain a reactive component also, impedances are expressed by complex notation:

$$Z = R \pm jX \qquad \textbf{[4.10]}$$

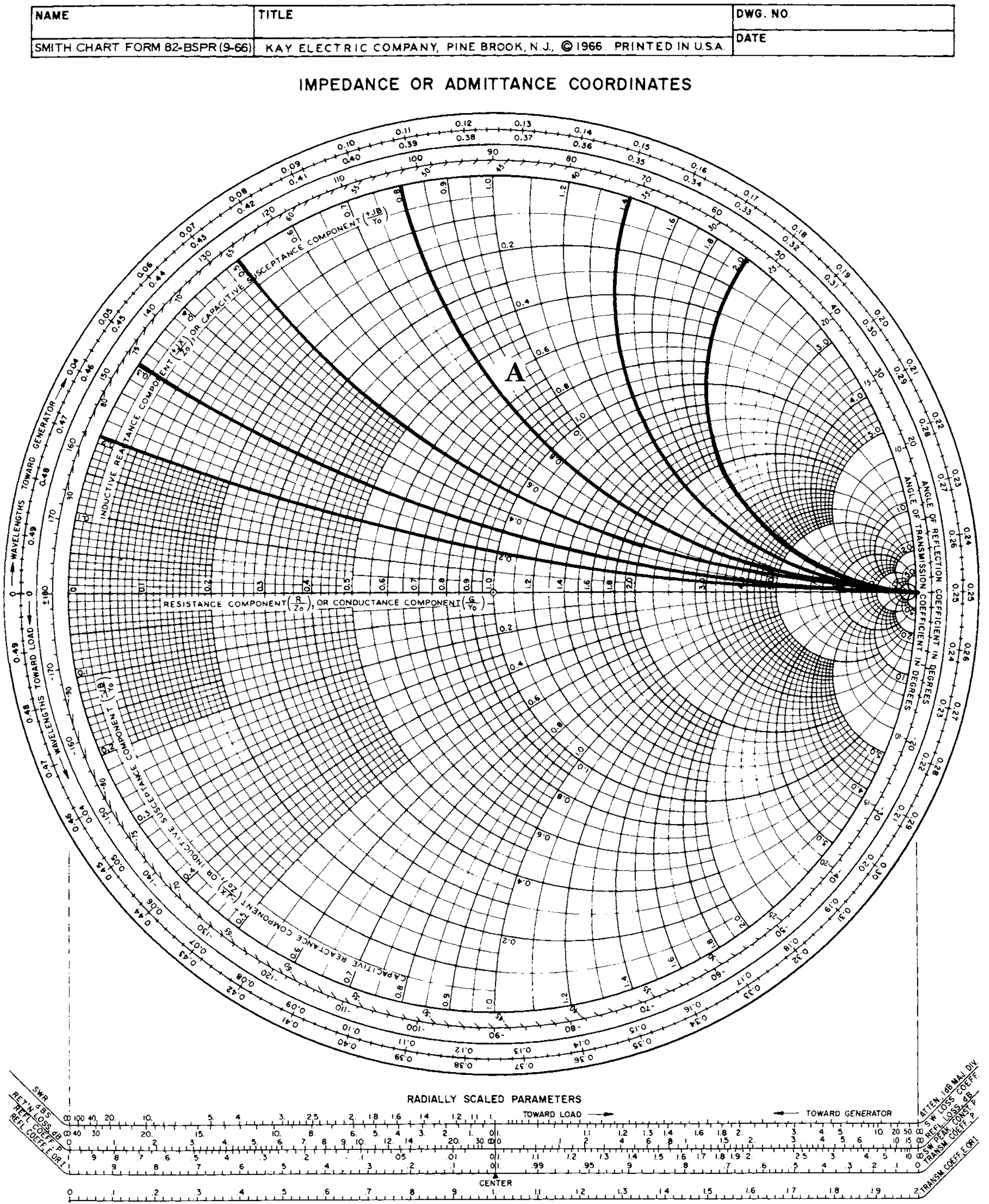
NAME
TITLE
DWG. NO.
DATE
SMITH CHART FORM 82-BSPR (9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J., © 1966 PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
A
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE COMPONENT (-jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
CENTER
SWR
RET'N dBS
RET'N LOSS, dB
REFL. COEFF. P
REFL. COEFF. E OR I
ATTEN. (dB) MAJ. DIV.
S.W. LOSS COEFF.
S.W. REFL. LOSS dB
S.W. PEAK (CONST. P)
TRANSM. COEFF. P
TRANSM. COEFF. E OR I

4-4A Constant inductive reactance lines. (*Courtesy of Kay Elementrics*)

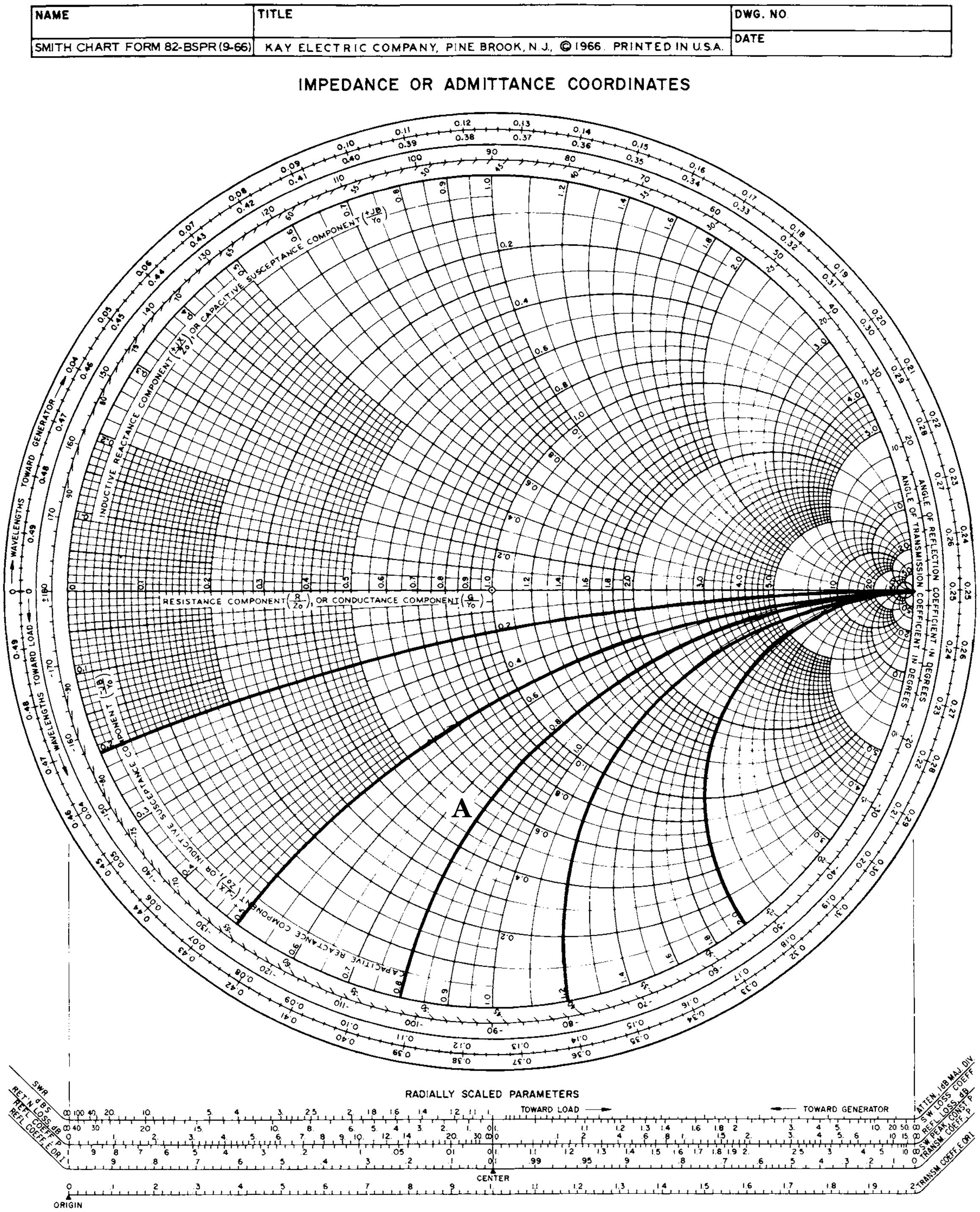
NAME
TITLE
DWG. NO
DATE
SMITH CHART FORM 82-BSPR(9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J. ©1966 PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT (+jX/Zo) OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo) OR INDUCTIVE SUSCEPTANCE COMPONENT (-jB/Yo)
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
A
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
SWR
dBS
RET'N LOSS, dB
REFL. COEFF. P
REFL. COEFF. E OR I
ATTEN. 1dB MAJ. DIV.
S.W. LOSS COEFF.
REFL. LOSS, dB
S.W. PEAK (CONST. P)
TRANSM. COEFF. P
TRANSM. COEFF. E OR I
CENTER
ORIGIN

4-4B Constant capacitive reactance lines. (*Courtesy of Kay Elementrics*)

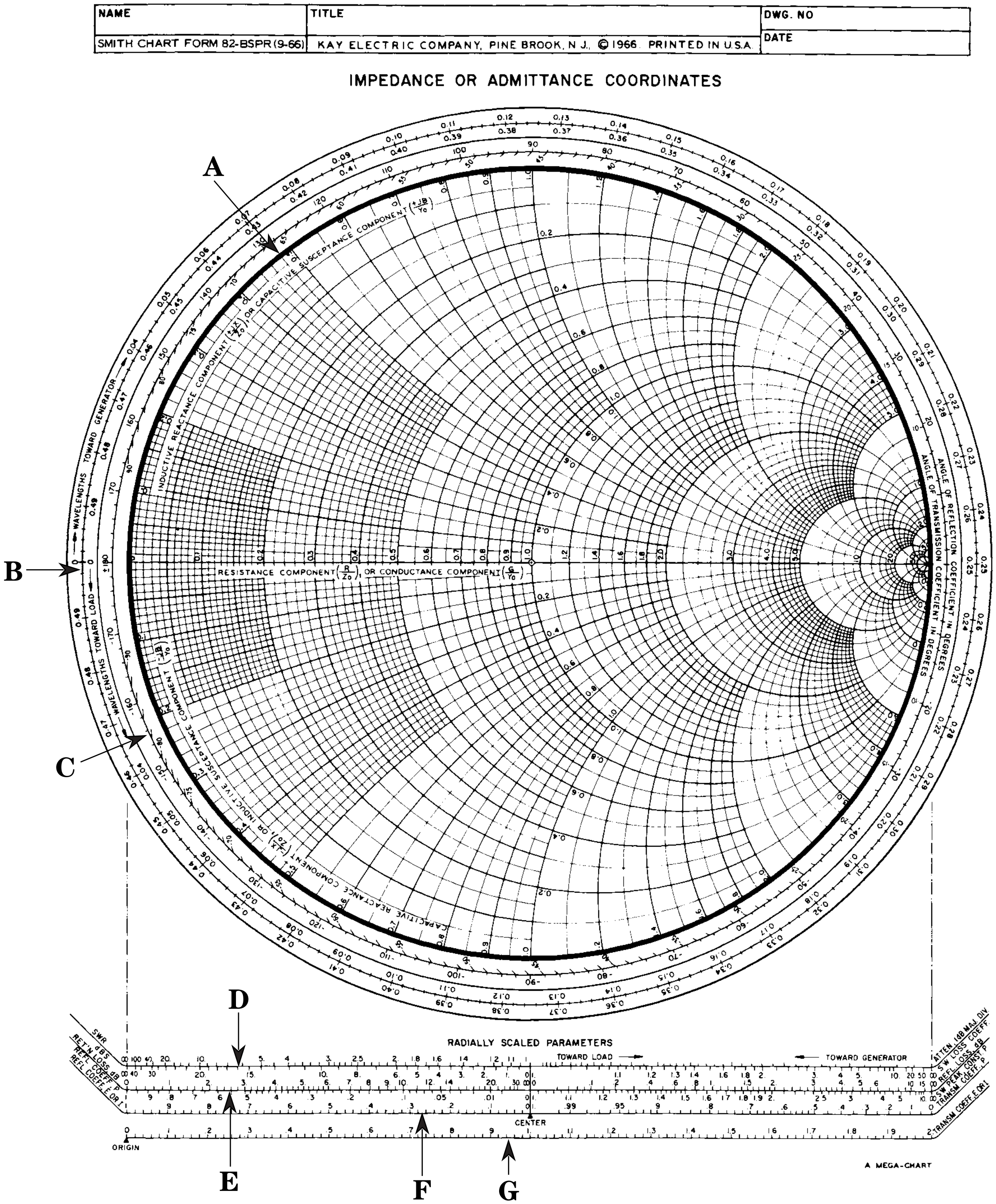
NAME
TITLE
DWG. NO
DATE
SMITH CHART FORM 82-BSPR(9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J. ©1966 PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
A
B
C
D
E
F
G
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo), OR INDUCTIVE SUSCEPTANCE COMPONENT (-jB/Yo)
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
SWR
RET'N LOSS dB
REFL. COEFF. P
REFL. COEFF. E OR I
ATTEN. 1dB MAJ. DIV.
S.W. LOSS COEFF.
REFL. LOSS dB
S.W. PEAK CONST. P
TRANSM. COEFF. E OR I
TRANSM. COEFF. P
CENTER
ORIGIN
A MEGA-CHART

4-4C Angle of transmission coefficient circle. (*Courtesy of Kay Elementrics*)

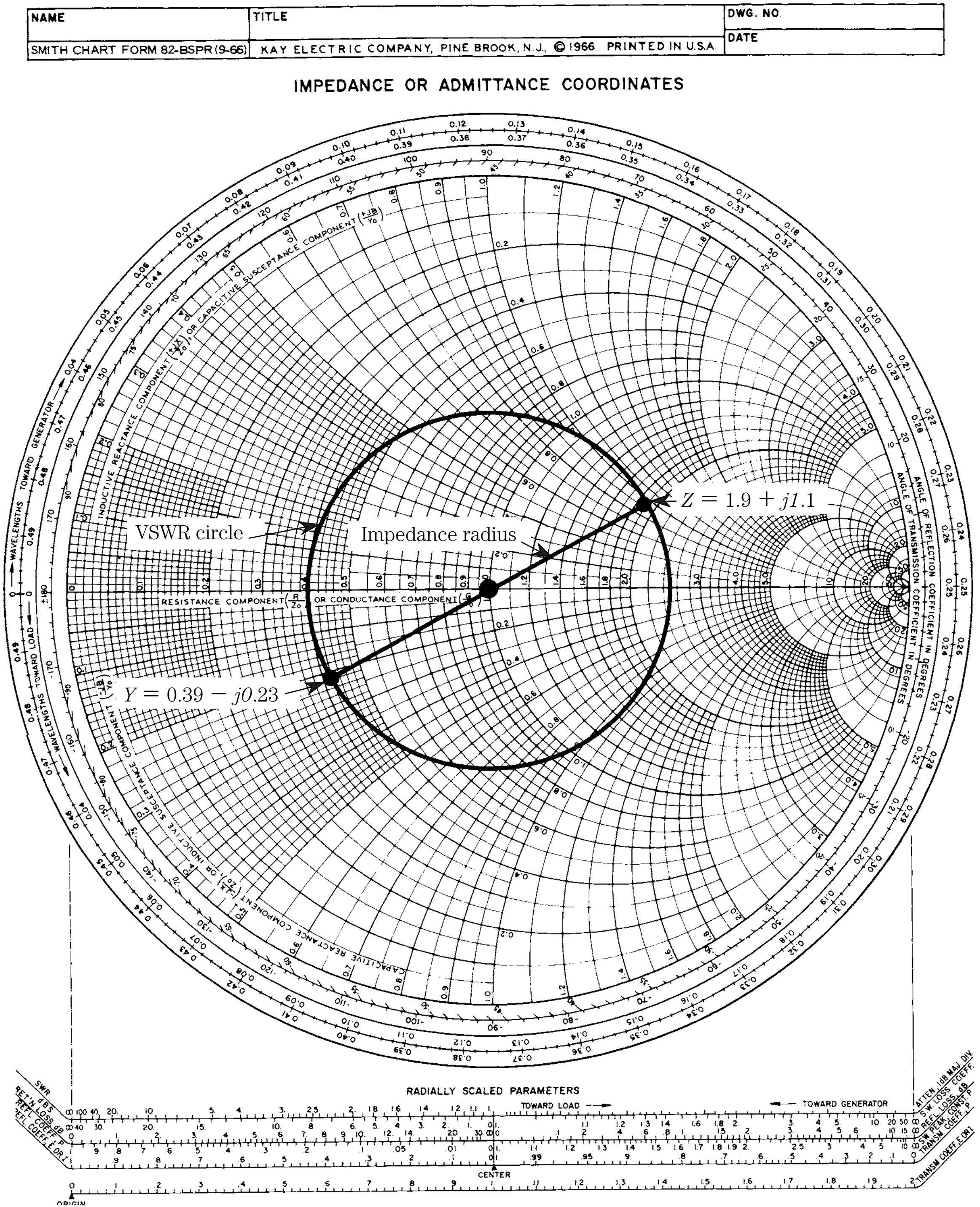

4-4D VSWR circles. (*Courtesy of Kay Elementrics*)

where

Z is the complex impedance
R is the resistive component
X is the reactive component

In order to find the *complex admittance*, take the reciprocal of the complex impedance by multiplying the simple reciprocal by the complex conjugate of the impedance. For example, when the normalized impedance is $1.9 + j1.1$, the normalized admittance will be:

$$Y = \frac{1}{Z} \qquad \textbf{[4.11]}$$

$$= \frac{1}{1.9 + j1.1} \times \frac{1.9 - j1.1}{1.9 - j1.1} \qquad \textbf{[4.12]}$$

$$= \frac{1.9 - j1.1}{3.6 + 1.2} \qquad \textbf{[4.13]}$$

$$= \frac{1.9 - j1.1}{4.8} = 0.39 - j0.23 \qquad \textbf{[4.14]}$$

One of the delights of the Smith chart is that this calculation is reduced to a quick graphical interpretation! Simply extend the impedance radius through the 1.0 center point until it intersects the VSWR circle again. This point of intersection represents the normalized admittance of the load.

Outer circle parameters

The standard Smith chart shown in Fig. 4-4C contains three concentric calibrated circles on the outer perimeter of the chart. Circle A has already been covered and it is the pure reactance circle. The other two circles define the wavelength distance (B) relative to either the load or generator end of the transmission line, and either the transmission or reflection coefficient angle in degrees (C).

There are two scales on the *wavelength* circle (B in Fig 4-4C), and both have their zero origin on the left-hand extreme of the pure resistance line. Both scales represent *one-half wavelength for one entire revolution*, and are calibrated from 0 through 0.50 such that these two points are identical with each other on the circle. In other words, starting at the zero point and travelling 360° around the circle brings one back to zero, which represents one-half wavelength, or $0.5\ \lambda$.

Although both wavelength scales are of the same magnitude (0 – 0.50), they are opposite in direction. The outer scale is calibrated clockwise and it represents *wavelengths toward the generator*; the inner scale is calibrated counterclockwise and represents *wavelengths toward the load*. These two scales are complementary at all points. Thus, 0.12 on the outer scale corresponds to (0.50 – 0.12) or 0.38 on the inner scale.

The *angle of transmission coefficient* and *angle of reflection coefficient* scales are shown in circle *C* in Fig. 4-4C. These scales are the relative phase angle between reflected and incident waves. Recall from transmission line theory (see Chap. 3), that a short circuit at the load end of the line reflects the signal back toward the generator 180° out of phase with the incident signal; an open line (i.e., infinite impedance) reflects the signal back to the generator in phase (i.e., 0°) with the incident signal. These facts are shown on the Smith chart by the fact that both scales start at 0° on the right-hand end of the pure resistance line, which corresponds to an infinite resistance, and it goes halfway around the circle to 180° at the 0 end of the pure resistance line. Note that the upper half-circle is calibrated 0 to +180°, and the bottom half-circle is calibrated 0 to –180°, reflecting indictive or capacitive reactance situations, respectively.

Radially scaled parameters

There are six scales laid out on five lines (*D* through *G* in Fig. 4-4C and in expanded form in Fig. 4-5) at the bottom of the Smith chart. These scales are called the *radially scaled parameters*—and they are both very important, and often overlooked. With these scales, we can determine such factors as VSWR (both as a ratio and in decibels), return loss in decibels, voltage or current reflection coefficient, and the power reflection coefficient.

The *reflection coefficient* Γ is defined as the *ratio of the reflected signal to the incident signal*. For voltage or current:

$$\Gamma = \frac{E_{\text{ref}}}{E_{\text{inc}}} \qquad \textbf{[4.15]}$$

and

$$\Gamma = \frac{I_{\text{ref}}}{I_{\text{inc}}} \qquad \textbf{[4.16]}$$

Power is proportional to the square of voltage or current, so:

$$P_{\text{pwr}} = \Gamma^2 \qquad \textbf{[4.17]}$$

or

$$\Gamma_{\text{pwr}} = \frac{P_{\text{ref}}}{P_{\text{inc}}} \qquad \textbf{[4.18]}$$

Example 10 W of microwave RF power is applied to a lossless transmission line, of which 2.8 W is reflected from the mismatched load. Calculate the reflection coefficient.

$$\Gamma_{\text{pwr}} = \frac{P_{\text{ref}}}{P_{\text{inc}}} \qquad \textbf{[4.19]}$$

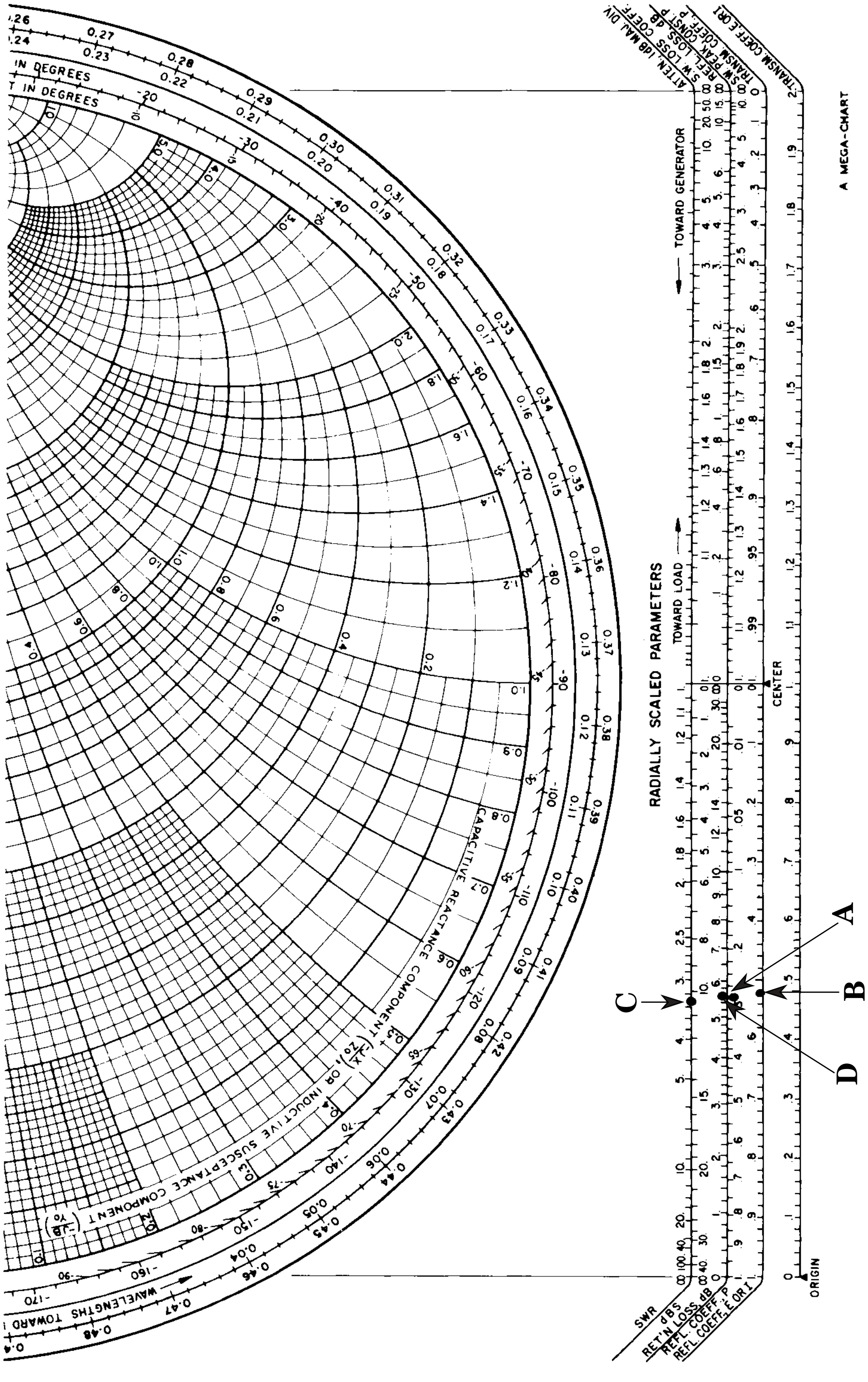
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
CENTER
ORIGIN
A MEGA-CHART
SWR
RET'N LOSS, dB
REFL. COEFF. P
REFL. COEFF. E OR I
ATTEN 1dB MAJ. DIV.
S.W. LOSS COEFF.
S.W. PEAK CONST. P
TRANSM. COEFF. P
TRANSM. COEFF. E OR I
CAPACITIVE REACTANCE COMPONENT
OR INDUCTIVE SUSCEPTANCE COMPONENT
WAVELENGTHS TOWARD
IN DEGREES
A
B
C
D

4-5 Radially scaled parameters.

$$\Gamma_{\text{pwr}} = \frac{2.8\ \text{W}}{10\ \text{W}} \qquad \textbf{[4.20]}$$

$$= 0.28 \qquad \textbf{[4.21]}$$

The voltage reflection coefficient Γ is found by taking the square root of the power reflection coefficient, so in this example it is equal to 0.529. These points are plotted at *A* and *B* in Fig. 4-5.

Standing wave ratio (SWR) can be defined in terms of reflection coefficient:

$$\text{VSWR} = \frac{1 + \Gamma}{1 - \Gamma} \qquad \textbf{[4.22]}$$

or

$$\text{VSWR} = \frac{1 + \sqrt{\Gamma_{\text{pwr}}}}{1 - \sqrt{\Gamma_{\text{pwr}}}} \qquad \textbf{[4.23]}$$

or, in our example,

$$\text{VSWR} = \frac{1 + \sqrt{0.28}}{1 - \sqrt{0.28}} \qquad \textbf{[4.24]}$$

$$= \frac{1 + 0.529}{1 - 0.529} \qquad \textbf{[4.25]}$$

$$= \frac{1.529}{0.471} = 3.25{:}1 \qquad \textbf{[4.26]}$$

or, in decibel form,

$$\text{VSWR}_{\text{dB}} = 20 \log (\text{VSWR}) \qquad \textbf{[4.27]}$$

$$= 20 \log (20) \qquad \textbf{[4.28]}$$

$$= (20)\ (0.510) = 10.2\ \text{dB} \qquad \textbf{[4.29]}$$

These points are plotted at *C* in Fig. 4-5. Shortly, you will work an example to show how these factors are calculated in a transmission line problem from a known complex load impedance.

Transmission loss is a measure of the one-way loss of power in a transmission line because of reflection from the load. *Return loss* represents the two-way loss, so it is exactly twice the transmission loss. Return loss is found from

$$\text{Loss}_{\text{ret}} = 10 \log (\Gamma_{\text{pwr}}) \qquad \textbf{[4.30]}$$

and, for our example in which $\Gamma_{\text{pwr}} = 0.28$,

$$\text{Loss}_{\text{ret}} = 10 \log (0.28) \qquad \textbf{[4.31]}$$

$$= (10)\ (-0.553) = -5.53\ \text{dB} \qquad \textbf{[4.32]}$$

This point is shown as *D* in Fig. 4-5.

The *transmission loss coefficient* can be calculated from

$$\text{TLC} = \frac{1 + \Gamma_{\text{pwr}}}{1 - \Gamma_{\text{pwr}}} \qquad \textbf{[4.33]}$$

or, for our example,

$$\text{TLC} = \frac{1 + (0.28)}{1 - (0.28)} \qquad \textbf{[4.34]}$$

$$= \frac{1.28}{0.72} = 1.78 \qquad \textbf{[4.35]}$$

The TLC is a correction factor that is used to calculate the attenuation caused by mismatched impedance in a lossy, as opposed to the ideal "lossless," line. The TLC is found from laying out the impedance radius on the *loss coefficient* scale on the radially scaled parameters at the bottom of the chart.

Smith chart applications

One of the best ways to demonstrate the usefulness of the Smith chart is by practical example. The following sections look at two general cases: *transmission line problems* and *stub matching systems*.

Transmission line problems

Figure 4-6 shows a 50-Ω transmission line connected to a complex load impedance Z_L of $36 + j40\ \Omega$. The transmission line has a velocity factor v of 0.80, which means that the wave propagates along the line at $\frac{8}{10}$ the speed of light ($c = 300{,}000{,}000$ m/s). The length of the transmission line is 28 cm. The generator V_{in} is operated at a frequency of 4.5 GHz and produces a power output of 1.5 W. See what you can glean from the Smith chart (Fig. 4-7).

First, normalize the load impedance. This is done by dividing the load impedance by the systems impedance (in this case, $Z_o = 50\ \Omega$):

$$Z = \frac{36 + j40\ \Omega}{50\ \Omega} \qquad \textbf{[4.36]}$$

$$= 0.72 + j0.8 \qquad \textbf{[4.37]}$$

The resistive component of impedance Z is located along the 0.72 pure resistance circle (see Fig. 4-7). Similarly, the reactive component of impedance Z is located by traversing the 0.72 constant resistance circle until the $+j0.8$ constant reactance circle is intersected. This point graphically represents the normalized load impedance $Z = 0.72 + j0.80$. A VSWR circle is constructed with an impedance radius equal to the line between 1.0 (in the center of the chart) and the $0.72 + j0.8$ point.

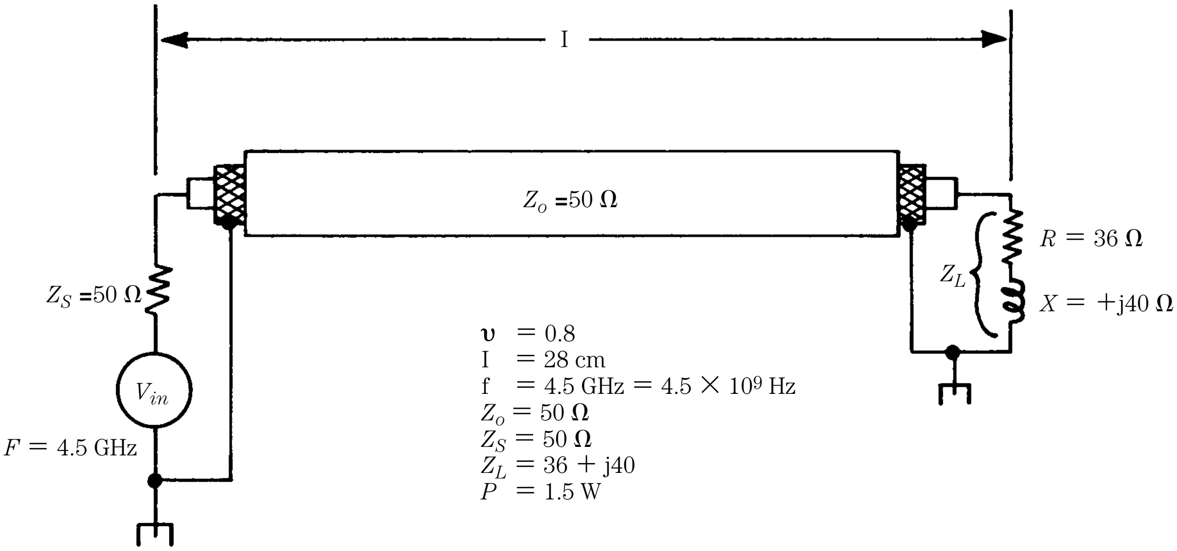

4-6 Transmission line and load circuit.

At a frequency of 4.5 GHz, the length of a wave propagating in the transmission line, assuming a velocity factor of 0.80, is

$$\lambda_{\text{line}} = \frac{c\ v}{F_{\text{Hz}}} \tag{4.38}$$

$$= \frac{(3 \times 10^8 \text{ m/s})\ (0.80)}{4.5 \times 10^9 \text{ Hz}} \tag{4.39}$$

$$= \frac{2.4 \times 10^8 \text{ m/s}}{4.5 \times 10^9 \text{ Hz}} \tag{4.40}$$

$$= 0.053 \text{ m} \times \frac{100 \text{ cm}}{\text{m}} = 5.3 \text{ cm} \tag{4.41}$$

One wavelength is 5.3 cm, so a half-wavelength is 5.3 cm/2, or 2.65 cm. The 28-cm line is 28 cm/5.3 cm, or 5.28 wavelengths long. A line drawn from the center (1.0) to the load impedance is extended to the outer circle, and it intersects the circle at 0.1325. Because one complete revolution around this circle represents one-half wavelength, 5.28 wavelengths from this point represents 10 revolutions plus 0.28 more. The residual 0.28 wavelengths is added to 0.1325 to form a value of (0.1325 + 0.28), or 0.413. The point 0.413 is located on the circle, and is marked. A line is then drawn from 0.413 to the center of the circle, and it intersects the VSWR circle at $0.49 - j0.49$, which represents the input impedance Z_{in} looking into the line.

To find the actual impedance represented by the normalized input impedance, you have to "denormalize" the Smith chart impedance by multiplying the result by Z_o:

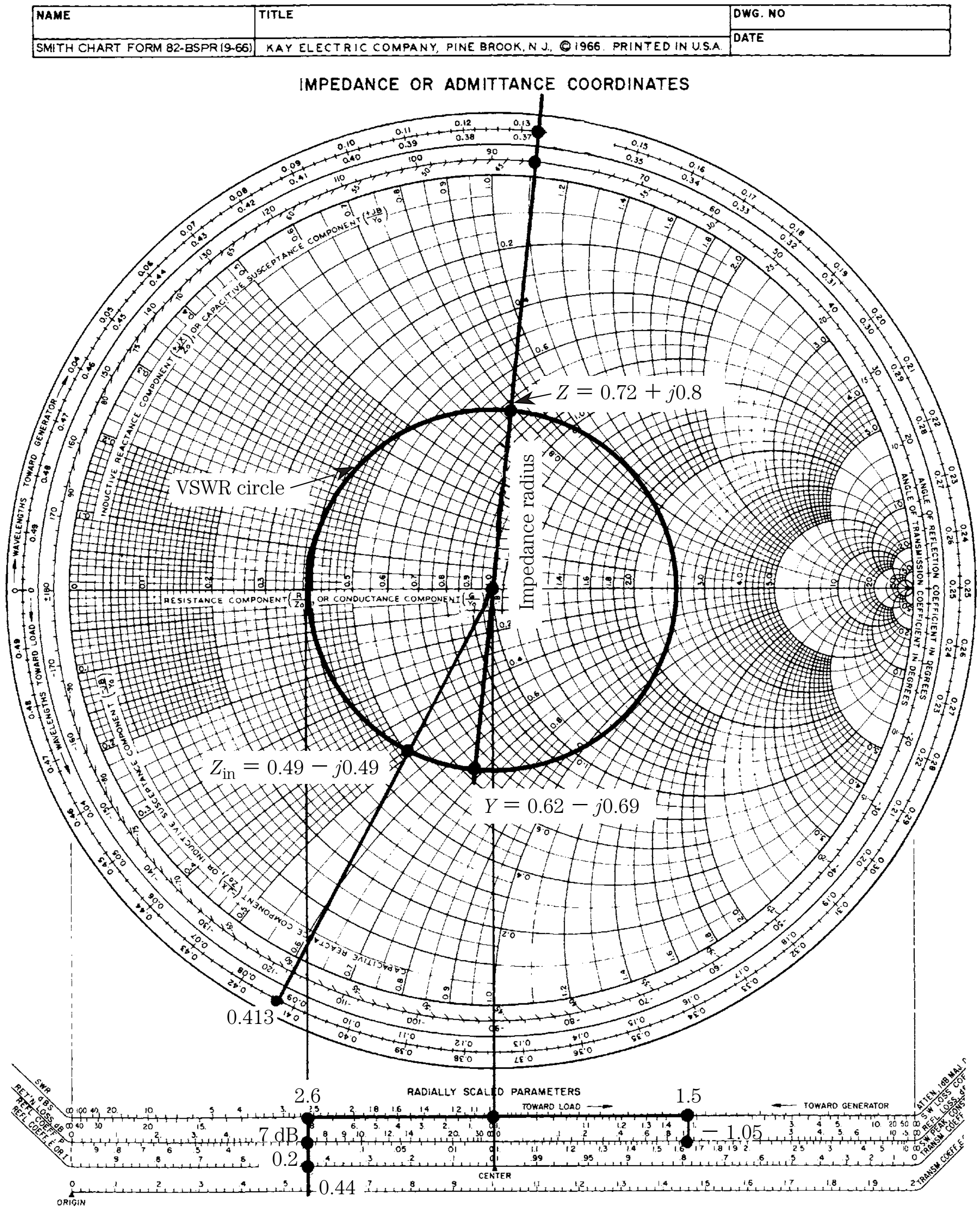

4-7 Solution to example. (*Courtesy of Kay Elementrics*)

$$Z_{\text{in}} = (0.49 - j0.49)\ (50\ \Omega) \qquad \textbf{[4.42]}$$

$$= 24.5 - j24.5\ \Omega \qquad \textbf{[4.43]}$$

It is this impedance that must be matched at the generator by a conjugate matching network.

The admittance represented by the load impedance is the reciprocal of the load impedance, and is found by extending the impedance radius through the center of the VSWR circle until it intersects the circle again. This point is found, and represents the admittance $Y = 0.62 - j0.69$. Confirming the solution mathematically:

$$Y = \frac{1}{Z} \qquad \textbf{[4.44]}$$

$$= \frac{1}{0.72 + j0.80} \times \frac{0.72 - j0.80}{0.72 - j0.80} \qquad \textbf{[4.45]}$$

$$= \frac{0.72 - j0.80}{1.16} = 0.62 - j0.69 \qquad \textbf{[4.46]}$$

The VSWR if found by transferring the “impedance radius” of the VSWR circle to the radial scales below. The radius (0.72 – 0.8) is laid out on the VSWR scale (topmost of the radially scaled parameters) with a pair of dividers from the center mark, and we find that the VSWR is approximately 2.6:1. The decibel form of VSWR is 8.3 dB (next scale down from VSWR), and this is confirmed by

$$\text{VSWR}_{\text{dB}} = 20 \log (\text{VSWR}) \qquad \textbf{[4.47]}$$

$$= (20) \log (2.7) \qquad \textbf{[4.48]}$$

$$= (20)\ (0.431) = 8.3\ \text{dB} \qquad \textbf{[4.49]}$$

The transmission loss coefficient is found in a manner similar to the VSWR, using the radially scaled parameter scales. In practice, once you have found the VSWR you need only drop a perpendicular line from the 2.6:1 VSWR line across the other scales. In this case, the line intersects the voltage reflection coefficient at 0.44. The power reflection coefficient Γ_{pwr} is found from the scale, and is equal to Γ^2. The perpendicular line intersects the power reflection coefficient line at 0.20.

The angle of reflection coefficient is found from the outer circles of the Smith chart. The line from the center to the load impedance ($Z = 0.72 + j0.8$) is extended to the *angle of reflection coefficient in degrees* circle, and intersects it at approximately 84°. The reflection coefficient is therefore 0.44/84°.

The transmission loss coefficient (TLC) is found from the radially scaled parameter scales also. In this case, the impedance radius is laid out on the *loss coefficient* scale, where it is found to be 1.5. This value is confirmed from:

$$\text{TLC} = \frac{1 + \Gamma_{\text{pwr}}}{1 - \Gamma_{\text{pwr}}} \qquad \textbf{[4.50]}$$

$$= \frac{1 + (0.20)}{1 - (0.21)} \qquad \textbf{[4.51]}$$

$$= \frac{1.20}{0.79} = 1.5 \qquad \textbf{[4.52]}$$

The *return loss* is also found by dropping the perpendicular from the VSWR point to the *RET'N LOSS, dB* line, and the value is found to be approximately 7 dB, which is confirmed by

$$\text{Loss}_{\text{ret}} = 10 \log (\Gamma_{\text{pwr}}) \text{ dB} \qquad \textbf{[4.53]}$$

$$= 10 \log (0.21) \text{ dB} \qquad \textbf{[4.54]}$$

$$= (10)\ (-0.677) \text{ dB} \qquad \textbf{[4.55]}$$

$$= 6.77 \text{ dB} = -6.9897 \text{ dB} \qquad \textbf{[4.56]}$$

The reflection loss is the amount of RF power reflected back down the transmission line from the load. The difference between incident power supplied by the generator (1.5 W in this example), $P_{\text{inc}} - P_{\text{ref}} = P_{\text{abs}}$ and the reflected power, is the *absorbed power* (P_a); or in the case of an antenna, the radiated power. The reflection loss is found graphically by dropping a perpendicular from the TLC point (or by laying out the impedance radius on the *REFL*. *LOSS*, *dB* scale), and in this example (Fig. 4-7) is –1.05 dB. You can check the calculations:

The return loss is –7 dB, so

$$-7 \text{ dB} = 10 \log \left(\frac{P_{\text{ref}}}{P_{\text{inc}}} \right) \qquad \textbf{[4.57]}$$

$$-7 = 10 \log \left(\frac{P_{\text{ref}}}{1.5 \text{ watts}} \right) \qquad \textbf{[4.58]}$$

$$\frac{-7}{10} = \log \left(\frac{P_{\text{ref}}}{1.5 \text{ watts}} \right) \qquad \textbf{[4.59]}$$

$$10^{(-7/10)} = \frac{P_{\text{ref}}}{1.5 \text{ W}} \qquad \textbf{[4.60]}$$

$$0.2 = \frac{P_{\text{ref}}}{1.5 \text{ W}} \qquad \textbf{[4.61]}$$

$$(0.2)(1.5 \text{ W}) = P_{\text{ref}} \qquad \textbf{[4.62]}$$

$$0.3 \text{ W} = P_{\text{ref}} \qquad \textbf{[4.63]}$$

The power absorbed by the load (P_a) is the difference between incident power P_{inc} and reflected power P_{ref}. If 0.3 W is reflected, then that means the absorbed power is (1.5 – 0.3), or 1.2 W.

The reflection loss is –1.05 dB, and can be checked from:

$$-1.05 \text{ dB} = 10 \log \left(\frac{P_a}{P_{\text{inc}}} \right) \qquad \textbf{[4.64]}$$

$$\frac{-1.05}{10} = \log \left(\frac{P_a}{1.5 \text{ W}} \right) \qquad \textbf{[4.65]}$$

$$10^{(-1.05/10)} = \frac{P_a}{1.5 \text{ W}} \qquad \textbf{[4.66]}$$

$$0.785 = \frac{P_a}{1.5 \text{ W}} \qquad \textbf{[4.67]}$$

$$(1.5 \text{ W}) \times (0.785) = P_a \qquad \textbf{[4.68]}$$

$$1.2 \text{ W} = P_a \qquad \textbf{[4.69]}$$

Now check what you have learned from the Smith chart. Recall that 1.5 W of 4.5-GHz microwave RF signal was input to a 50-Ω transmission line that was 28 cm long. The load connected to the transmission line has an impedance of 36 + j40. From the Smith chart:

Admittance (load)	0.62 – j0.69
VSWR	2.6:1
VSWR (dB):	8.3 dB
Reflection coefficient (E)	0.44
Reflection coefficient (P)	0.2
Reflection coefficient angle	84 degrees
Return loss	–7 dB
Reflection loss	–1.05 dB
Transmission loss coefficient	1.5

Note that in all cases the mathematical interpretation corresponds to the graphical interpretation of the problem, within the limits of accuracy of the graphical method.

Stub matching systems

A properly designed matching system will provide a conjugate match to a complex impedance. Some sort of matching system or network is needed any time the load impedance Z_L is not equal to the characteristic impedance Z_o of the transmission line. In a transmission line system, it is possible to use a *shorted stub* connected in parallel with the line, at a critical distance back from the mismatched load, in order to effect a match. The stub is merely a section of transmission line that is shorted at the end not connected to the main transmission line. The reactance (hence also susceptance) of a shorted line can vary from $-\lambda$ to $+\lambda$, depending upon length, so you can use a line of critical length L_2 to cancel the reactive component of the load impedance. Because the stub is connected in parallel with the line it is a bit easier to work with admittance parameters, rather than impedance.

Consider the example of Fig. 4-8, in which the load impedance is $Z = 100 + j60$, which is normalized to $2.0 + j1.2$. This impedance is plotted on the Smith chart in Fig. 4-9, and a VSWR circle is constructed. The admittance is found on the chart at point $Y = 0.37 - j0.22$.

In order to provide a properly designed matching stub, you need to find two lengths. L_1 is the length (relative to wavelength) from the load toward the generator (see L_1 in Fig. 4-8); L_2 is the length of the stub itself.

The first step in finding a solution to the problem is to find the points where the unit conductance line (1.0 at the chart center) intersects the VSWR circle; there are two such points shown in Fig. 4-9: $1.0 + j1.1$ and $1.0 - j1.1$. Select one of these (choose $1.0 + j1.1$) and extend a line from the center 1.0 point through the $1.0 + j1.1$ point to the outer circle (*WAVELENGTHS TOWARD GENERATOR*). Similarly, a line is drawn from the center through the admittance point 0.37 – 0.22 to the outer circle. These two lines intersect the outer circle at the points 0.165 and 0.461. The distance of the stub back toward the generator is found from:

$$L_1 = 0.165 + (0.500 - 0.461)\,\lambda \qquad \textbf{[4.70]}$$

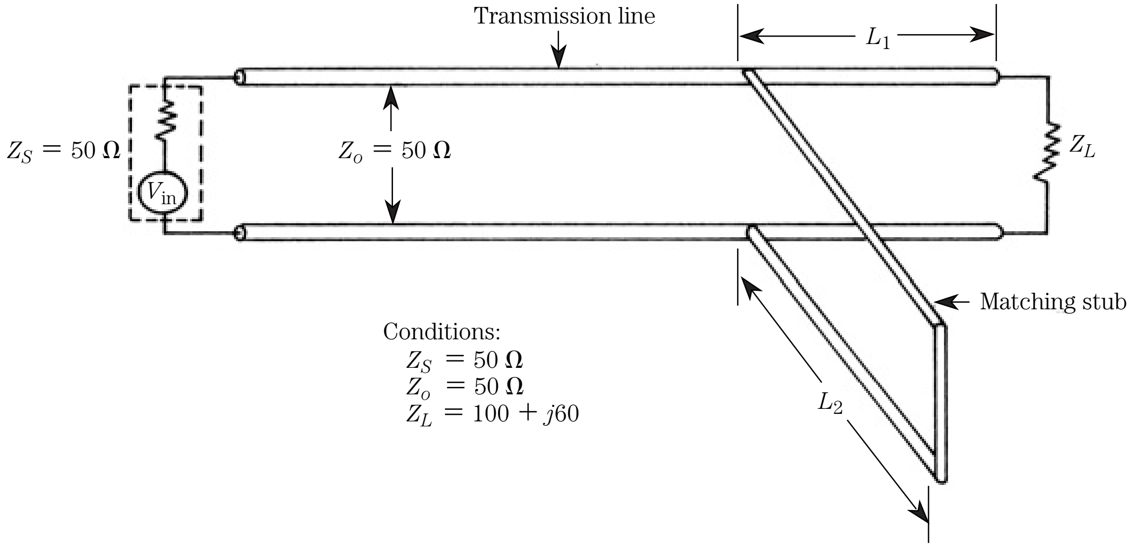

4-8 Matching stub length and position.

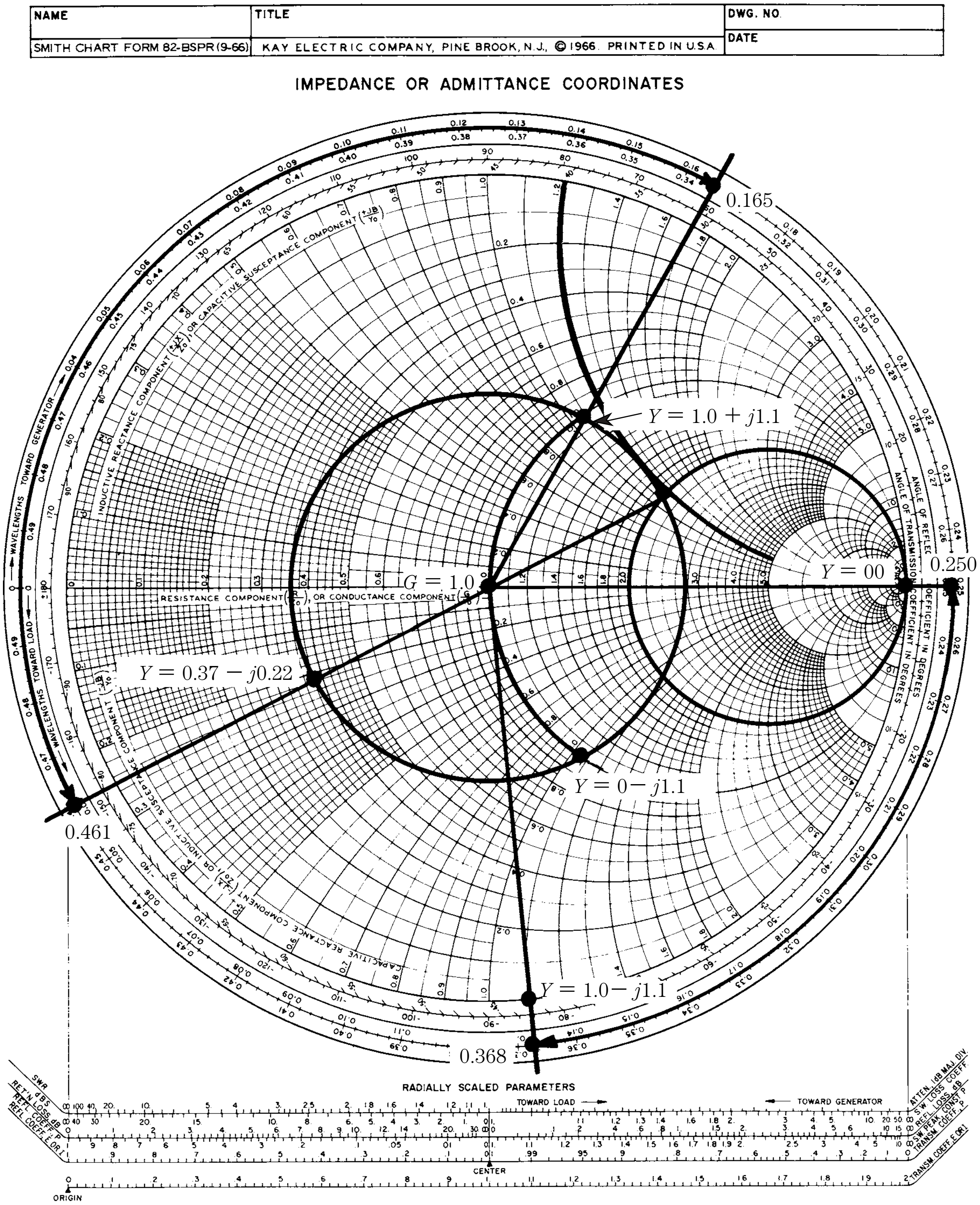
NAME
TITLE
DWG. NO.
DATE
SMITH CHART FORM 82-BSPR(9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J., © 1966 PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
0.165
Y = 1.0 + j1.1
Y = 00
0.250
G = 1.0
Y = 0.37 − j0.22
Y = 0 − j1.1
0.461
Y = 1.0 − j1.1
0.368
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo) OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (−jX/Zo) OR INDUCTIVE SUSCEPTANCE COMPONENT (−jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
CENTER
ORIGIN

4-9 Solution to problem. (*Courtesy of Kay Elementrics*)

$$= 0.165 + 0.039\,\lambda \qquad \textbf{[4.71]}$$

$$= 0.204\,\lambda \qquad \textbf{[4.72]}$$

The next step is to find the length of the stub required. This is done by finding two points on the Smith chart. First, locate the point where admittance is infinite (far right side of the pure conductance line); second, locate the point where the admittance is $0 - j1.1$. (Note that the susceptance portion is the same as that found where the unit conductance circle crossed the VSWR circle.) Because the conductance component of this new point is 0, the point will lay on the $-j1.1$ circle at the intersection with the outer circle. Now draw lines from the center of the chart through each of these points to the outer circle. These lines intersect the outer circle at 0.368 and 0.250. The length of the stub is found from

$$L_2 = (0.368 - 0.250)\lambda \qquad \textbf{[4.73]}$$

$$= 0.118\lambda \qquad \textbf{[4.74]}$$

From this analysis you can see that the impedance, $Z = 100 + j60$, can be matched by placing a stub of a length 0.118λ at a distance 0.204λ back from the load.

The Smith chart in lossy circuits

Thus far, you have dealt with situations in which loss is either zero (i.e., ideal transmission lines), or so small as to be negligible. In situations where there is appreciable loss in the circuit or line, however, you see a slightly modified situation. The VSWR circle, in that case, is actually a spiral, rather than a circle.

Figure 4-10 shows a typical situation. Assume that the transmission line is 0.60λ long, and is connected to a normalized load impedance of $Z = 1.2 + j1.2$. An "ideal" VSWR circle is constructed on the impedance radius represented by $1.2 + j1.2$. A line (A) is drawn, from the point where this circle intersects the pure resistance baseline (B), perpendicularly to the *ATTEN, 1 dB/MAJ. DIV.* line on the radially scaled parameters. A distance representing the loss (3 dB) is stepped off on this scale. A second perpendicular line is drawn, from the –3 dB point, back to the pure resistance line (C). The point where line C intersects the pure resistance line becomes the radius for a new circle that contains the actual input impedance of the line. The length of the line is 0.60λ, so you must step back (0.60 – 0.50)λ or 0.1λ. This point is located on the *WAVELENGTHS TOWARD GENERATOR* outer circle. A line is drawn from this point to the 1.0 center point. The point where this new line intersects the new circle is the actual input impedance (Z_{in}).The intersection occurs at $0.76 + j0.4$, which (when denormalized) represents an input impedance of $38 + j20\ \Omega$.

Frequency on the Smith chart

A complex network may contain resistance, inductive reactance, and capacitive reactance components. Because the reactance component of such impedances is a function of frequency, the network or component tends to also be frequency-sensitive.

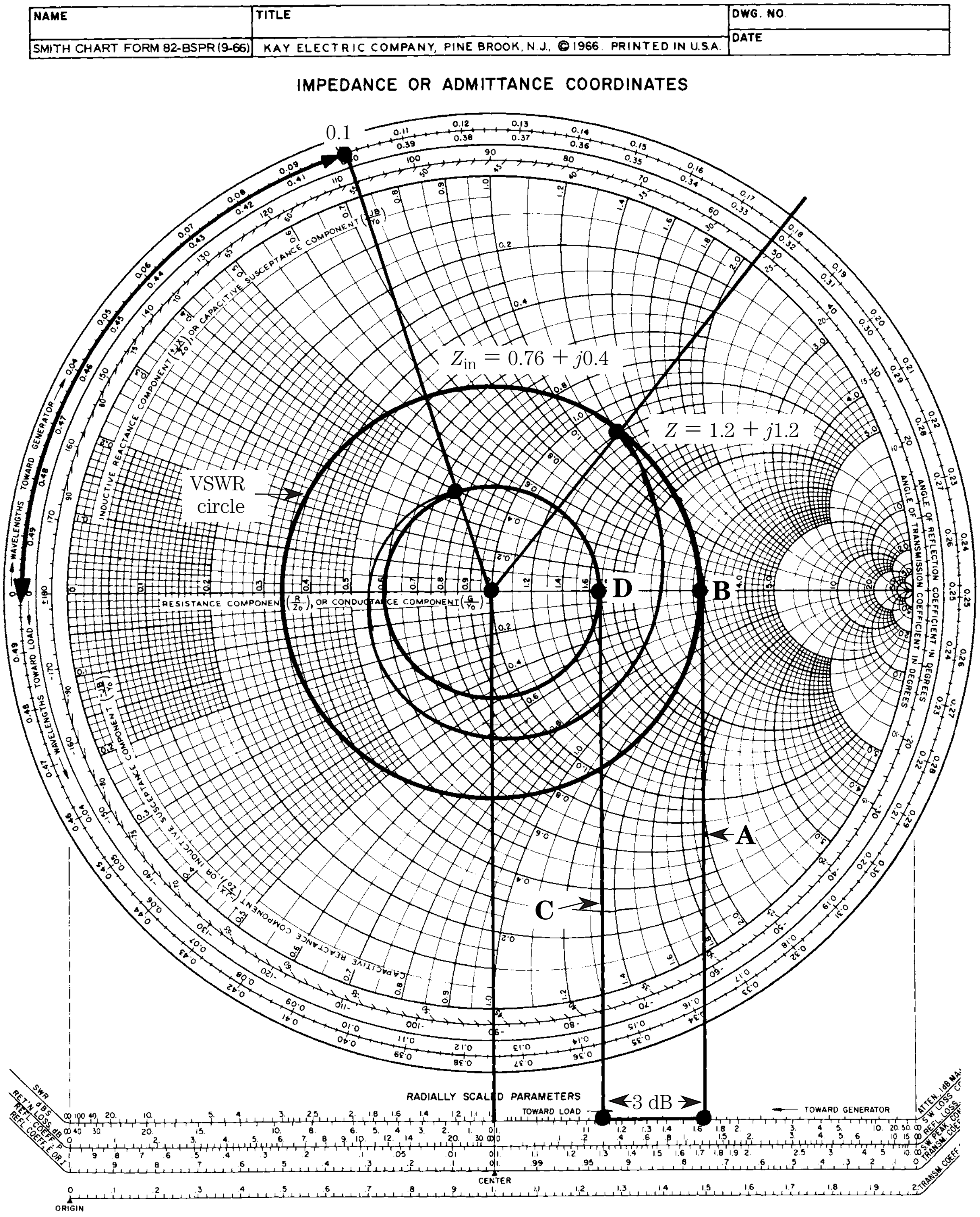
NAME
TITLE
DWG. NO.
DATE
SMITH CHART FORM 82-BSPR (9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J., © 1966. PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
0.1
$Z_{\text{in}} = 0.76 + j0.4$
$Z = 1.2 + j1.2$
VSWR circle
D
B
A
C
3 dB
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
CENTER
ORIGIN

4-10 Solution. (*Courtesy of Kay Elementrics*)

You can use the Smith chart to plot the performance of such a network with respect to various frequencies. Consider the load impedance connected to a 50-Ω transmission line in Fig. 4-11. In this case, the resistance is in series with a 2.2-pF capacitor, which will exhibit a different reactance at each frequency. The impedance of this network is

$$Z = R - j\left(\frac{1}{\omega C}\right) \tag{4.75}$$

or

$$Z = 50 - j\left(\frac{1}{2\,\pi\,F\,C}\right) \tag{4.76}$$

And, in normalized form,

$$Z' = 1.0 - \frac{j}{(2\,\pi\,F\,C) \times 50} \tag{4.77}$$

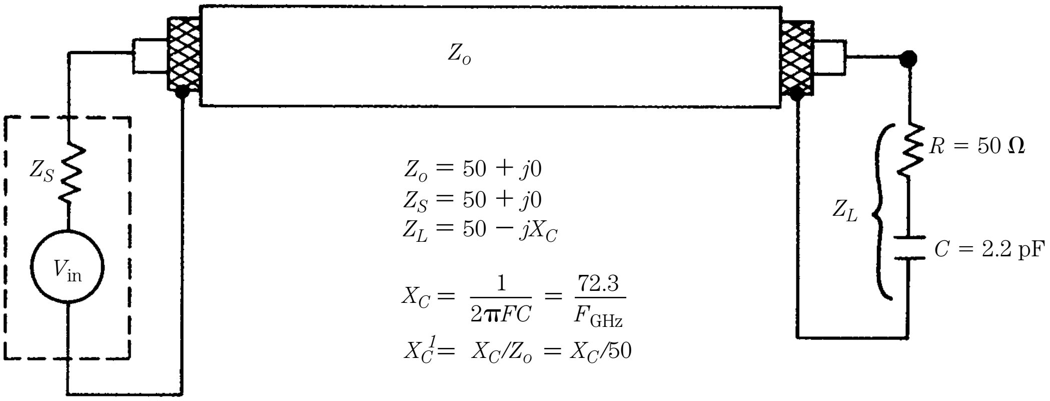

Z_o
Z_S
V_{in}
$Z_o = 50 + j0$
$Z_S = 50 + j0$
$Z_L = 50 - jX_C$
$X_C = \frac{1}{2\pi FC} = \frac{72.3}{F_{GHz}}$
$X_C^1 = X_C/Z_o = X_C/50$
Z_L
$R = 50\ \Omega$
$C = 2.2$ pF

Freq. (GHz)	X_C	
1	$-j72.3$	$-j1.45$
2	$-j36.2$	$-j0.72$
3	$-j24.1$	$-j0.48$
4	$-j18$	$-j0.36$
5	$-j14.5$	$-j0.29$
6	$-j12$	$-j0.24$

4-11 Load and source impedance transmission line circuit.

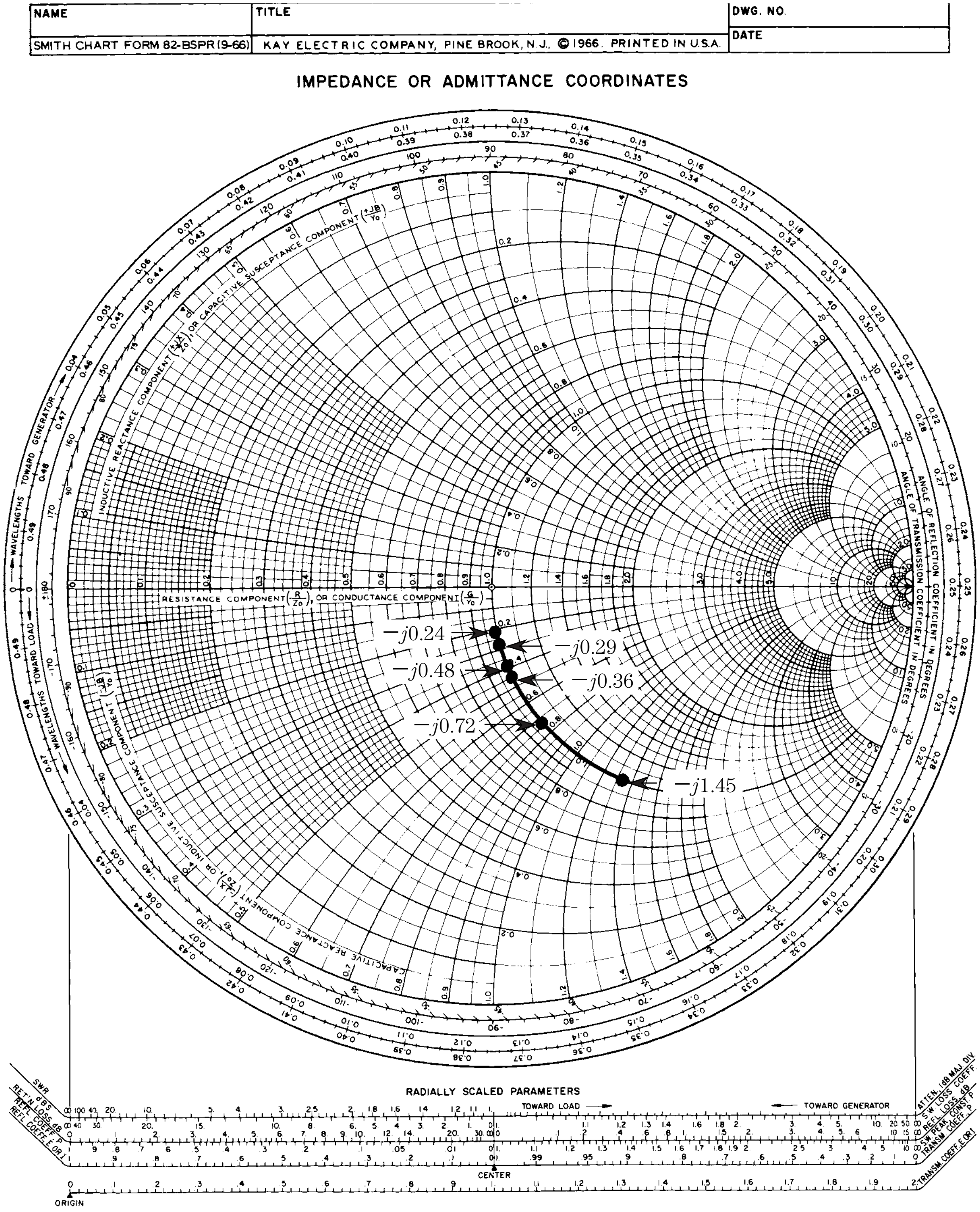
NAME
TITLE
DWG. NO.
DATE
SMITH CHART FORM 82-BSPR(9-66)
KAY ELECTRIC COMPANY, PINE BROOK, N.J. © 1966. PRINTED IN U.S.A.
IMPEDANCE OR ADMITTANCE COORDINATES
RESISTANCE COMPONENT (R/Zo), OR CONDUCTANCE COMPONENT (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo), OR CAPACITIVE SUSCEPTANCE COMPONENT (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (−jX/Zo), OR INDUCTIVE SUSCEPTANCE COMPONENT (−jB/Yo)
WAVELENGTHS TOWARD GENERATOR
WAVELENGTHS TOWARD LOAD
ANGLE OF REFLECTION COEFFICIENT IN DEGREES
ANGLE OF TRANSMISSION COEFFICIENT IN DEGREES
−j0.24
−j0.29
−j0.48
−j0.36
−j0.72
−j1.45
RADIALLY SCALED PARAMETERS
TOWARD LOAD
TOWARD GENERATOR
SWR
dBS
RET'N LOSS, dB
REFL. COEFF. P
REFL. COEFF. E OR I
ATTEN. 1dB MAJ. DIV.
S.W. LOSS COEFF.
S.W. PEAK (CONST. P)
TRANSM. COEFF. P
TRANSM. COEFF. E OR I
CENTER
ORIGIN

4-12 Solution. (*Courtesy of Kay Elementrics*)

$$= 1.0 - \frac{j}{6.9 \times 10^{-10} F} \tag*{\textbf{[4.78]}}$$

$$= 1.0 - \frac{j \times 7.23 \times 10^{10}}{F} \tag*{\textbf{[4.79]}}$$

Or, converted to gigahertz,

$$Z' = 1.0 - \frac{j72.3}{F_{\mathrm{GHz}}} \tag*{\textbf{[4.80]}}$$

The normalized impedances for the sweep of frequencies from 1 to 6 GHz are therefore

$$Z = 1.0 - j1.45 \tag*{\textbf{[4.81]}}$$

$$Z = 1.0 - j0.72 \tag*{\textbf{[4.82]}}$$

$$Z = 1.0 - j0.48 \tag*{\textbf{[4.83]}}$$

$$Z = 1.0 - j0.36 \tag*{\textbf{[4.84]}}$$

$$Z = 1.0 - j0.29 \tag*{\textbf{[4.85]}}$$

$$Z = 1.0 - j0.24 \tag*{\textbf{[4.86]}}$$

These points are plotted on the Smith chart in Fig. 4-12. For complex networks, in which both inductive and capacitive reactance exist, take the difference between the two reactances (i.e., $X = X_L - X_C$).

This page intentionally left blank

5
CHAPTER

Fundamentals of radio antennas

AN UNFORTUNATE OVERSIGHT IN MANY BOOKS ON RADIO ANTENNAS IS A LACK OF coverage on the most basic fundamentals of antenna theory. Most books, including the first draft of this one, start with a discussion of dipoles, but overlook that certain physical mechanisms are at work. An antenna is basically a *transducer* that converts electrical alternating current oscillations at a radio frequency to an electromagnetic wave of the same frequency. This chapter looks at the physics of how that job is accomplished.

The material in this chapter was adapted from a U.S. Army training manual on antennas and radio propagation. Although unfortunately no longer in print, the manual contained the best coverage of basics the author of this book could find. Given that U.S. government publications are not protected by copyright, this information can be brought to you in full.

Antenna fundamentals

The electric and magnetic fields radiated from an antenna form the electromagnetic field, and this field is responsible for the transmission and reception of electromagnetic energy through free space. An antenna, however, is also part of the electrical circuit of a transmitter (or a receiver); because of its distributed constants, it acts as a circuit containing inductance, capacitance, and resistance. Therefore, it can be expected to display definite voltage and current relationships in respect to a given input. A current through it produces a magnetic field, and a charge on it produces an electrostatic field. These two fields taken together form the induction field. To gain a better understanding of antenna theory, a review of the basic electrical concepts of voltage and electric field and of current and magnetic field is necessary.

Copyright 2001 - the McGraw-Hill Companies

Voltage and electric field

When a capacitor is connected across a source of voltage, such as a battery (Fig. 5-1), it is charged some amount, depending on the voltage and the value of capacitance. Because of the emf (electromotive force) of the battery, negative charges flow to the lower plate, leaving the upper plate positively charged. Accompanying the accumulation of charge is the building up of the electric field. The flux lines are directed from the positive to the negative charges and at right angles to the plates.

If the two plates of the capacitor are spread farther apart, the electric field must curve to meet the plates at right angles (Fig. 5-2). The straight lines in A become arcs in B, and approximately semicircles in C, where the plates are in a straight line. Instead of flat metal plates, as in the capacitor, the two elements can take the form of metal rods or wires. The three-dimensional view in Fig. 5-3 depicts the electric field more accurately. In A of Fig. 5-3, the wires are approximately 30° apart, and the flux lines are projected radially from the positively charged wire to the negatively charged wire. In B of Fig. 5-3, the two wires lie in a straight line, and the flux lines form a pattern similar to the lines of longitude around the earth. To bring out the picture more clearly, only the lines in one plane are given.

Assume that the sphere marked *E* in Fig. 5-3B is a transmitter supplying RF energy. The two wires then can serve as the antenna for the transmitter. RF energy is radiated from the antenna and charges move back and forth along the wires, alternately compressing and expanding the flux lines of the electric field. The reversals in polarity of the transmitter signal also reverse the direction of the electric field.

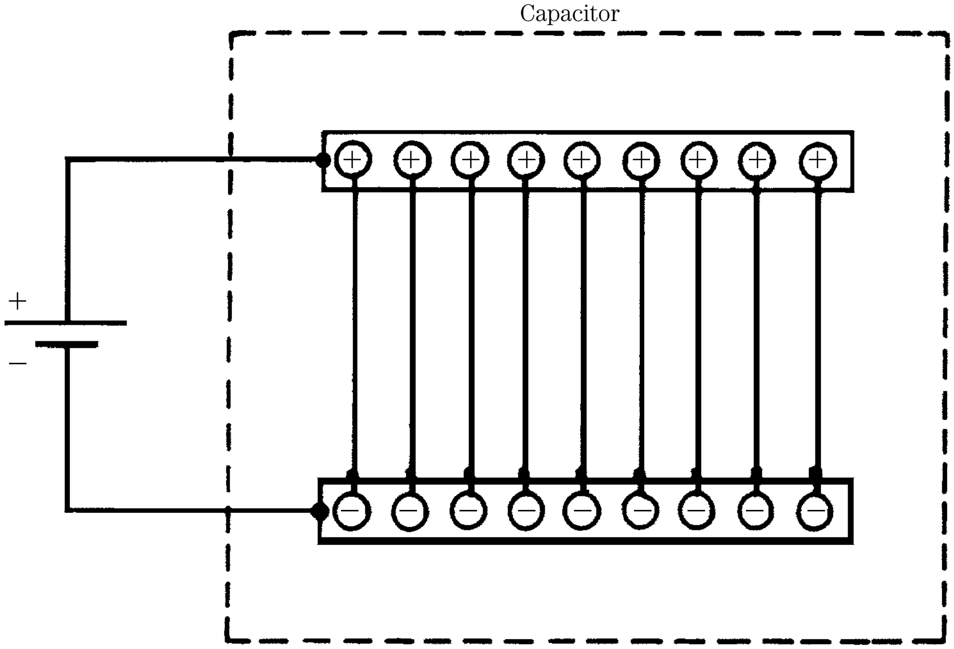

5-1 Charges on plates of a capacitor.

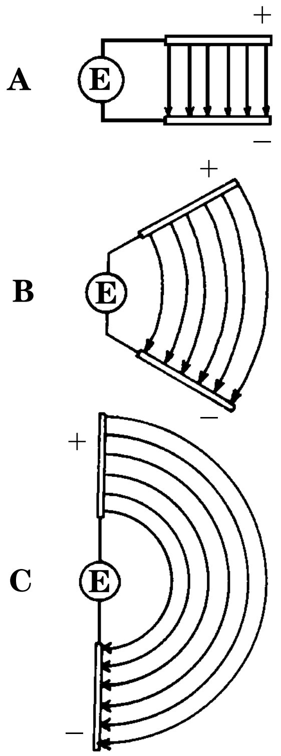

5-2 Electric field between plates at various angles.

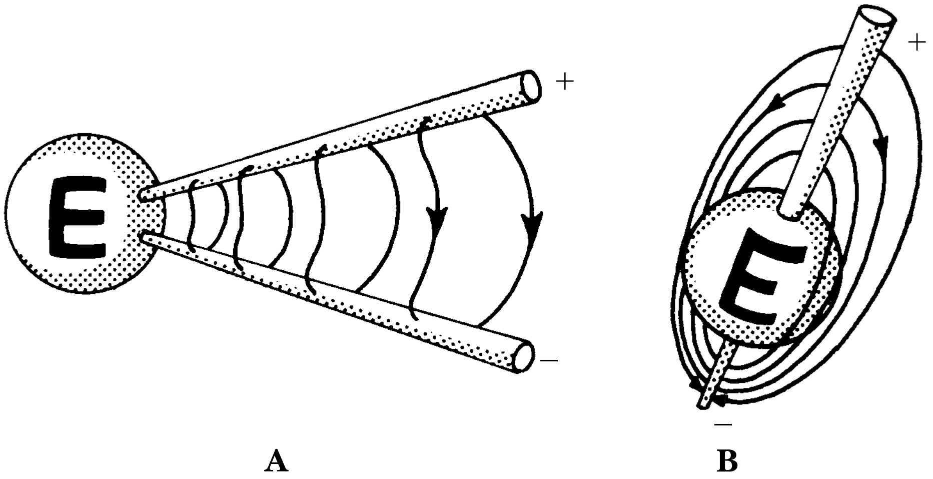

5-3 Electric field between wires at various angles.

When a charge is put on the plates of a capacitor by means of a battery, an electric field is set up between its plates. The flow of charge from source to capacitor ceases when the capacitor is fully charged, and the capacitor is said to be charged to a *voltage* equal, and of opposite polarity, to that of the source. The charged capacitor can be used as a source of emf since it stores energy in the form of an electric field. This is the same as saying that *an electric field indicates voltage*. The presence of an electric field about an antenna also indicates voltage. Since the polarity, and the amount of charge, depend on the nature of the transmitter output, the antenna voltage also depends on the energy source. For example, if a battery constitutes the source, the antenna charges to a voltage equal and opposite to that of the battery. If RF energy is supplied to a half-wave antenna, the voltage across the antenna lags the current by 90°. The half-wave antenna acts as if it was a capacitor, and it can be described as being *capacitive*.

Current and magnetic field

A *moving* charge along a conductor constitutes a current and produces a magnetic field around the conductor. Therefore, the flow of charge along an antenna also will be accompanied by a magnetic field. The intensity of this field is directly proportional to the flow of charge. When the antenna is uncharged, the current flow is maximum, since there is no opposing electric field. Because of this current flow, a charge accumulates on the antenna, and an electric field builds up in increasing opposition to the emf of the source. The current flow decreases and when the antenna is fully charged, the current no longer flows.

The magnetic field in the space around a current-carrying device has a specific configuration, with the magnetic flux lines drawn according to a definite rule (Fig. 5-4). Whereas, in the electric field, the electric lines are drawn from a positive charge to a negative charge, in the magnetic field, the flux lines are drawn according to the *left-hand rule*. The direction of current flow is upward along both halves of the antenna. The lines of magnetic flux form concentric loops that are perpendicular to the direction of current flow. The arrowheads on the loops indicate the direction of the field. If the thumb of the left hand is extended in the direction of current flow and the fingers are clenched, then the rough circles formed by the fingers indicate the direction of the magnetic field. This is the left-hand rule, or convention, which is used to determine the direction of the magnetic field.

Combined electric and magnetic fields

When RF energy from a transmitter is supplied to an antenna, the effects of charge, voltage, current, and the electric and magnetic fields are taking place simultaneously. These effects (Fig. 5-5) have definite time and space relationships to each other. If a half-wave antenna is used, the relations between charge and current flow can be predicted, because of the capacitive nature of this antenna. The voltage will lag the current by 90°, and the electric and magnetic fields will be 90° out of phase. With no electric field present (no charge), the current flow is unimpeded, and the magnetic field is maximum. As charge accumulates on the antenna, the electric field builds up in opposition to current flow and the magnetic field decreases in intensity. When the electric field reaches its maximum strength, the magnetic field has decayed to zero.

A reversal in polarity of the source reverses the direction of current flow as well as the polarity of the magnetic field, and the electric field aids the flow of current by

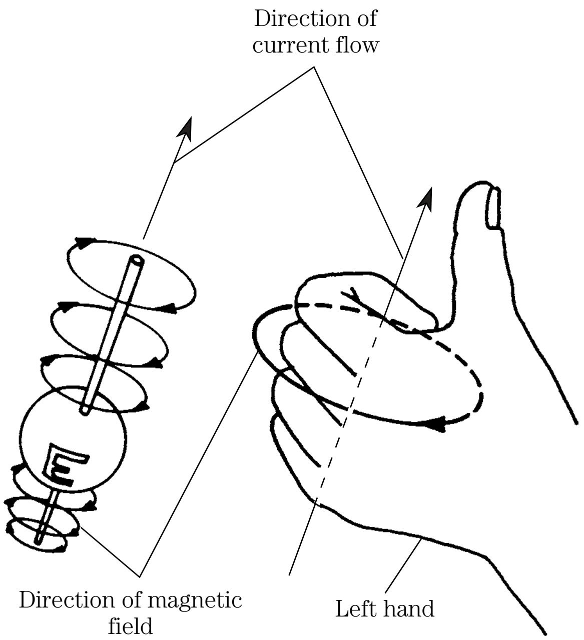

5-4 Magnetic field about a half-wave antenna (left-hand rule).

discharging. The magnetic field builds up to a maximum, and the electric field disappears as the charge is dissipated. The following half-cycle is a repetition of the first half-cycle, but in the reverse direction. This process continues as long as energy is supplied to the antenna. The fluctuating electric and magnetic fields combine to form the induction field, in which the electric and magnetic flux maximum intensities occur 90° apart in time, or in *time quadrature*. Physically, they occur at right angles to each other, or in *space quadrature*. To sum up, the electric and magnetic fields about the antenna are in space and time quadrature.

Standing waves

Assume that it is possible to have a wire conductor with one end extending infinitely, with an RF transmitter connected to this wire. When the transmitter is turned on, an RF current in the form of sine waves of RF energy moves down the wire. These waves of energy are called *traveling waves*. The resistance of the conductor gradually diminishes the amplitude of the waves, but they continue to travel so long as the line does not come to an end.

The antenna, however, has some finite length. Therefore, the traveling waves are halted when they reach the end of the conductor. Assume that the RF transmitter is turned on just long enough to allow one sine wave of energy to get on the line (Fig. 5-6A). This traveling wave is moving down the antenna toward the end. When the wave reaches the end of the conductor, the current path is broken abruptly. With the stoppage of current flow, the magnetic field collapses. A voltage is induced at the end of the conductor that causes current to flow *back toward the source*, as in Fig. 5-6B. The wave is reflected back to the source, and, if a continual succession of

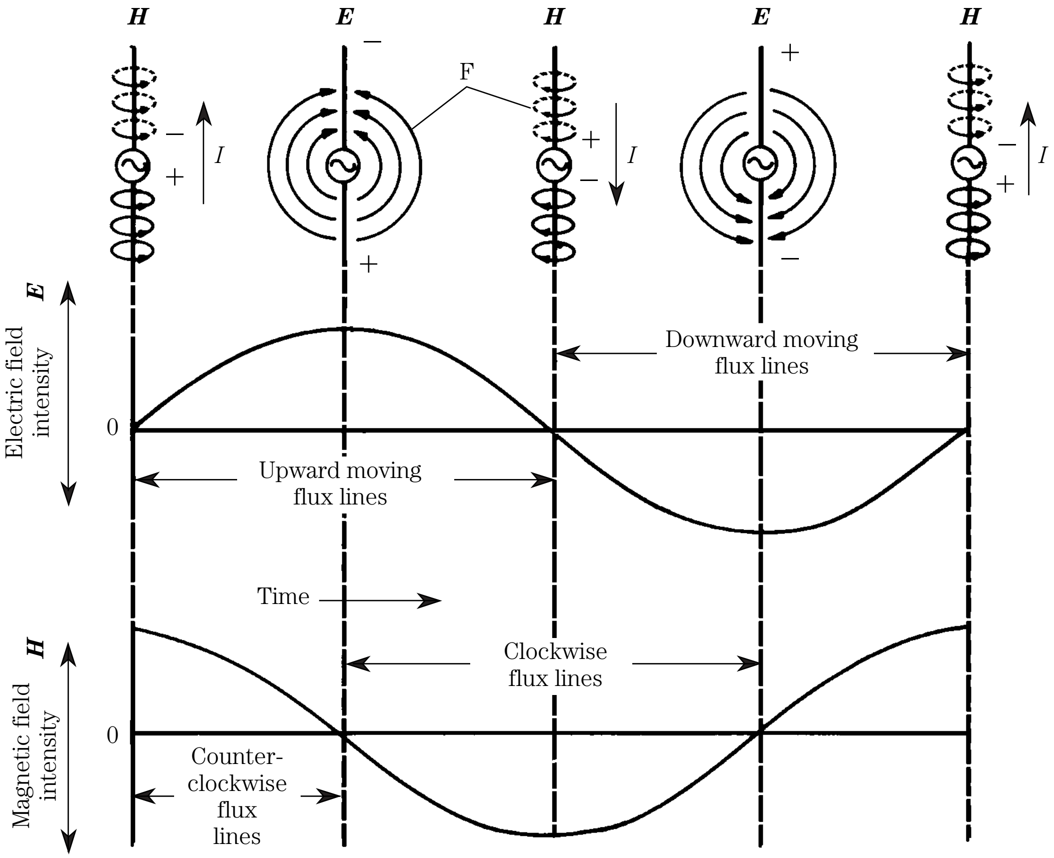
H
E
H
E
H
F
−
+
I
+
−
Electric field intensity
E
0
Downward moving flux lines
Upward moving flux lines
Time
Magnetic field intensity
H
0
Clockwise flux lines
Counter-clockwise flux lines

5-5 Electric and magnetic fields 90° out of phase.

waves is sent down the line, waves will be reflected in the same continual pattern. The wave moving from the transmitter toward the end is called the *incident* wave, and its reflection is called the *reflected* wave.

A continuous flow of incident waves results in a continuous flow of reflected waves. Because there is only one conductor, the two waves must pass each other. Electrically, the only current that actually flows is the *resultant* of both of these waves. The waves can reinforce or cancel each other as they move.

When they reinforce, the resultant wave is maximum; when they cancel, the resultant wave is minimum. In a conductor that has a finite length, such as an antenna, the points at which the maxima and minima of the resultant wave occur (Fig. 5-6C) are *stationary*. In other words, the maximum and minimum points stand still, although both the incident and reflected waves are moving. The resultant wave stands still on the line, only its amplitude being subject to change. Because of this effect, the resultant is referred to as a *standing wave*.

The development of the standing wave on an antenna by actual addition of the traveling waves is illustrated in Fig. 5-7. At the instant pictured in A, the incident and reflected waves just coincide. The result is a standing wave having twice the amplitude of either traveling wave. In B, the waves move apart in opposite directions, and

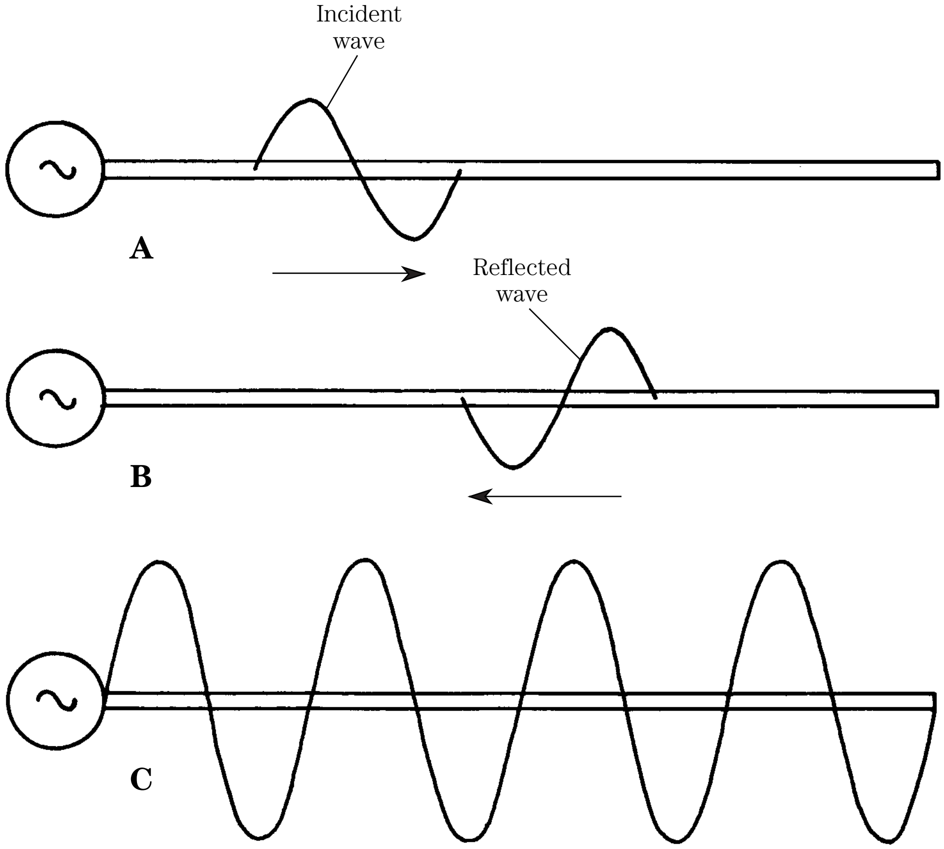

5-6 Traveling waves on an antenna and a typical resultant wave.

the amplitude of the resultant decreases, but the points of maximum and minimum do not move.

When the traveling waves have moved to a position of 180° phase difference, the resultant is zero along the entire length of the antenna, as shown in C. At this instant, there can be no current flow in the antenna. The continuing movement of the traveling waves, shown in D, builds up a resultant in a direction opposite to that in A. The in-phase condition of the traveling waves results in a standing wave, in E, equal in amplitude, but 180° out of phase with the standing wave in A.

If the progressive pictures of the standing wave are assembled on one set of axes, the result is as in Fig. 5-8. The net effect of the incident and reflected waves is apparent. The curves are lettered with reference to Fig. 5-7. As the traveling waves move past each other, the standing wave changes only its amplitude. The fixed minimum points are called *nodes*, and the curves representing the amplitude are called *loops*.

The concept of the standing wave can be applied to the half-wave antenna with reference to either current or voltage distribution at any instant. This application is possible because there are traveling waves of both voltage and current. Because voltage and current are out of phase on the half-wave antenna, the standing waves also are found to be out of phase.

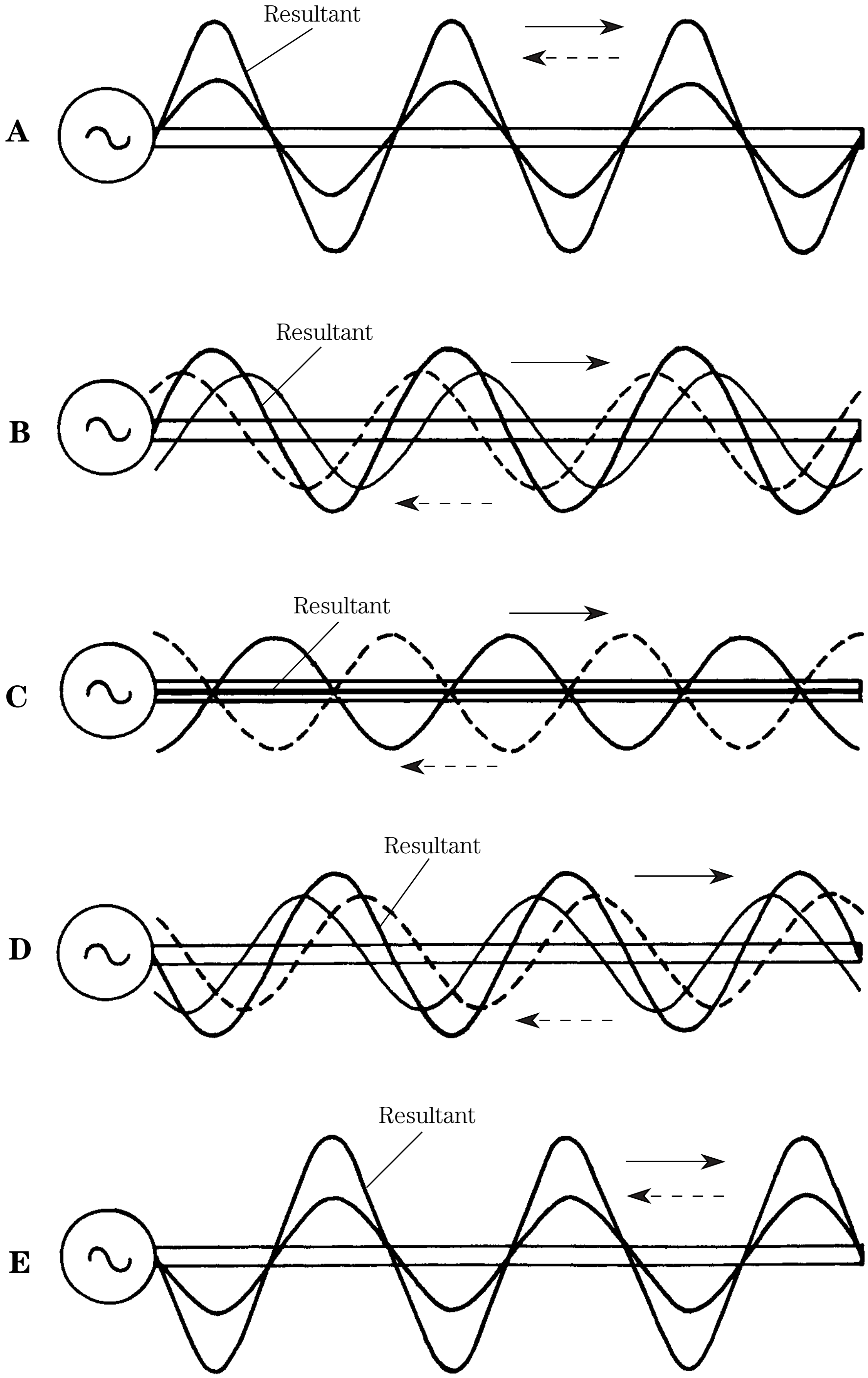
Resultant
A
Resultant
B
Resultant
C
Resultant
D
Resultant
E

5-7 Development of standing wave from traveling wave.

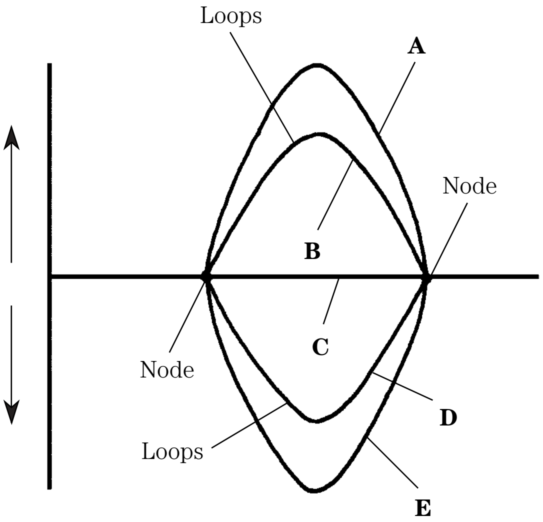

5-8 Standing waves.

Voltage and current distribution on half-wave antenna

When an RF transmitter is feeding a half-wave antenna, positive and negative charges move back and forth along the antenna (Figs. 5-9 and 5-10). The first picture shows the position of the charges at some arbitrary time, T_0. The RF charges being observed are at the ends of the antenna, and there is a maximum difference in potential between the ends, A and B. The remaining illustrations show the instantaneous positions of the charges at regular intervals of 22.5° throughout a complete cycle.

To the right of each instantaneous position of the charges are curves representing the current and voltage *at that particular time* for any point on the antenna. For example, at time T_0, the positive and negative charges are at points A and B on the antenna. The voltage between these points represents a maximum difference of potential. The current, being 90° out of phase in respect to the voltage, is everywhere zero. These distribution curves are *standing waves* derived in the same manner as those covered in the previous paragraph.

The next illustration shows the position of the charges at time T_1. The standing wave of current is a relative maximum at the center of the antenna. This current loop has nodes that remain at the ends of the antenna, and it is, therefore, 90° out of phase with the standing wave of voltage.

At T_2 and T_3, the charges move closer together, and the standing wave of voltage slowly decreases in amplitude. Conversely, the current loop increases in magnitude. When the charges meet after 90° of the RF cycle (T_4), the effect is that of having the positive and negative charges cancel. The voltage loop accordingly is zero everywhere on the antenna, and the current loop rises to its maximum value, unimpeded by any charge on the antenna.

At time T_5, the charges have passed each other, each charge having moved past the center point of the antenna. The polarity of the voltage loops is reversed, and they build up in the opposite direction, keeping the node always at the center point of the antenna. The reversal of polarity is shown in the charge positions at T_3, T_4, and

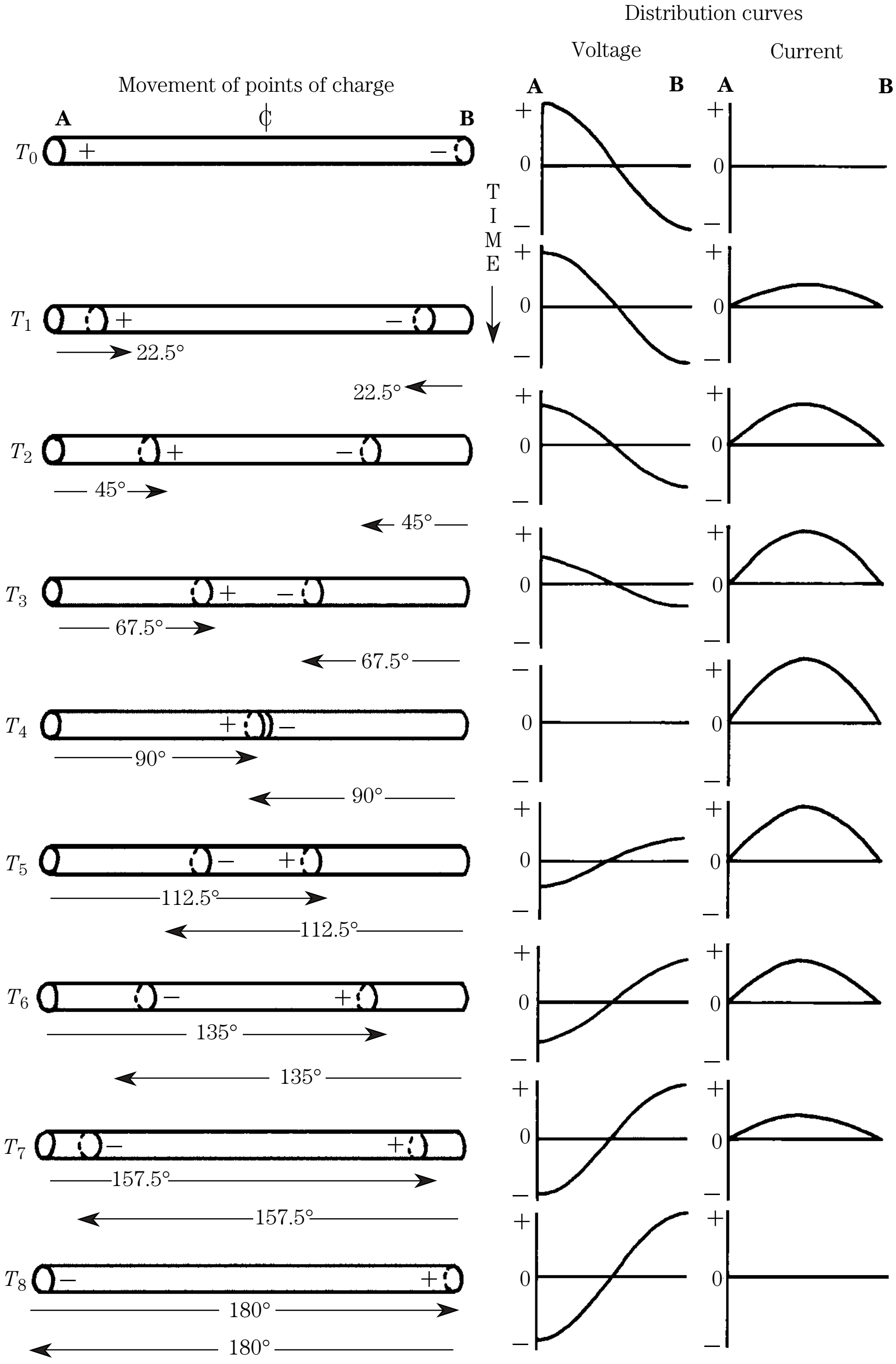
Distribution curves
Voltage
Current
Movement of points of charge
A
B
A
B
A
B
T_0
T_1
T_2
T_3
T_4
T_5
T_6
T_7
T_8
TIME
+
0
−
22.5°
45°
67.5°
90°
112.5°
135°
157.5°
180°

5-9 Voltage and current distribution in terms of positive and negative charges.

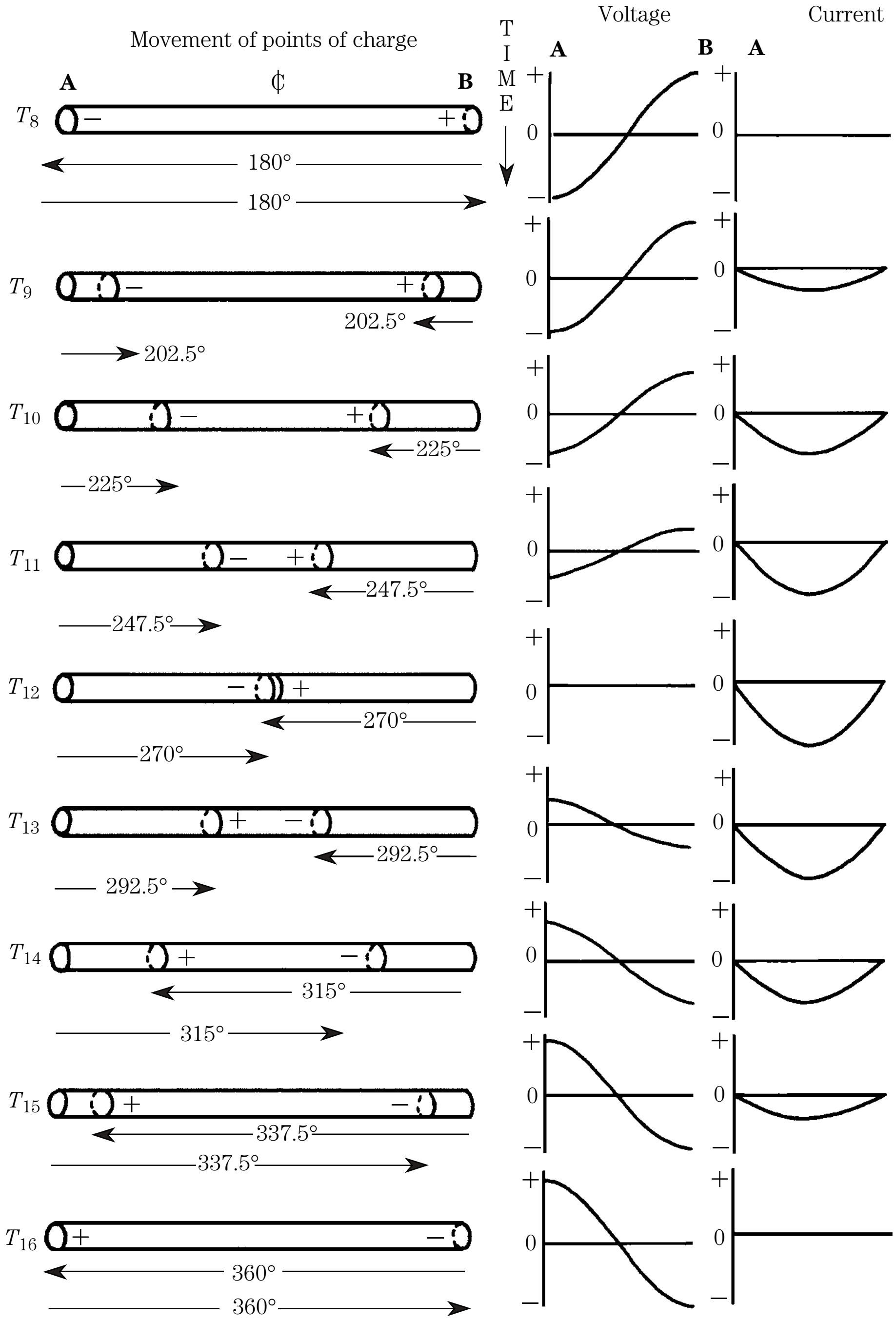
Movement of points of charge
Voltage
Current
TIME
A
B
A
₵
T8
T9
T10
T11
T12
T13
T14
T15
T16
180°
202.5°
225°
247.5°
270°
292.5°
315°
337.5°
360°
+
0
−

5-10 Voltage and current distribution in terms of positive and negative charges.

T_5. The separation of the charges also is accompanied by a decrease in the amplitude of the current loop.

From T_5 to T_8, the charges move out to the ends of the antenna. During this time, the voltage loops increase and the current loops decrease in amplitude. At time T_8, which occurs 180° after T_0 in the RF cycle, the charges have moved to opposite ends of the antenna. Compare the picture in T_0 to the picture in T_8. It is seen that the negative charge is now at point *A* and the positive charge at point *B*. Because the positions of the charges have been reversed from T_0 to T_8, the voltage loops in T_8 are 180° out of phase, compared with the loops in T_0.

From T_8 to T_{16} in Fig. 5-10, the movement of the charges is shown in the opposite direction, the current loop reaching a maximum at T_{12}. When the entire RF cycle is completed at time T_{16}, the charges have returned to the positions that they occupied at T_0. The distribution curves of voltage and current also are in their original conditions. The entire process then is repeated for each RF cycle.

Standing waves of voltage and current The distribution curves of the current and voltage are standing waves. This means that they are the resultants obtained by adding two traveling waves. The two traveling waves are associated with the positive and negative charges. The wave caused by the negative charge can be called the *incident wave* and the wave caused by the positive charge the *reflected wave*. This, however, is clearer when the concept of negative and positive charges is used.

The positive charge, taken at time T_0 in Fig. 5-9, produces a traveling wave of voltage, shown by the dashed line in Fig. 5-11A. The negative charge at the opposite end of the antenna produces an identical traveling wave (dash-dot curve). These two add together to produce the T_0 voltage distribution curve, which is the resultant wave of Fig. 5-9A. Both of these waveforms are identical, being the *standing wave of voltage* at time T_0. All of the following distribution curves of Fig. 5-11 are produced in the same manner. They are the standing wave resultants caused by the traveling waves accompanying the charges.

In Fig. 5-11B, each of the traveling waves has moved 45°, the positive traveling wave moving to the right and the negative traveling wave moving to the left. This time corresponds to T_2 in Fig. 5-9. The standing wave produced corresponds to the voltage distribution curve at T_2. The standing waves of current are produced in the same manner. The current curves at D, E, and F of Fig. 5-11 correspond to times T_0, T_2, and T_4 of Fig. 5-9.

Standing waves of voltage In Fig. 5-12A, voltage standing waves occurring at different times are brought together on one axis, *AB*, representing a half-wave antenna. Essentially, these are the same curves shown progressively in Figs. 5-9 and 5-10 as voltage distribution curves. They can be used to determine the voltage at any point on the antenna, at any instant of time. For example, if it is desired to know the variations of voltage occurring at point *Y* on the antenna over the RF cycle, the variations are graphed in respect to time, as shown in Fig. 5-12B. At T_0, the voltage at *Y* is maximum. From T_0 through T_3, the voltage decreases, passing through zero at T_4. The voltage builds up to a maximum in the opposite direction at T_8, returning through zero to its original position from T_8 to T_{16}.

Between T_0 and T_{16}, therefore, an entire sine-wave cycle, *Y*, is reproduced. This is also true of any other point on the antenna, with the exception of the node at *X*.

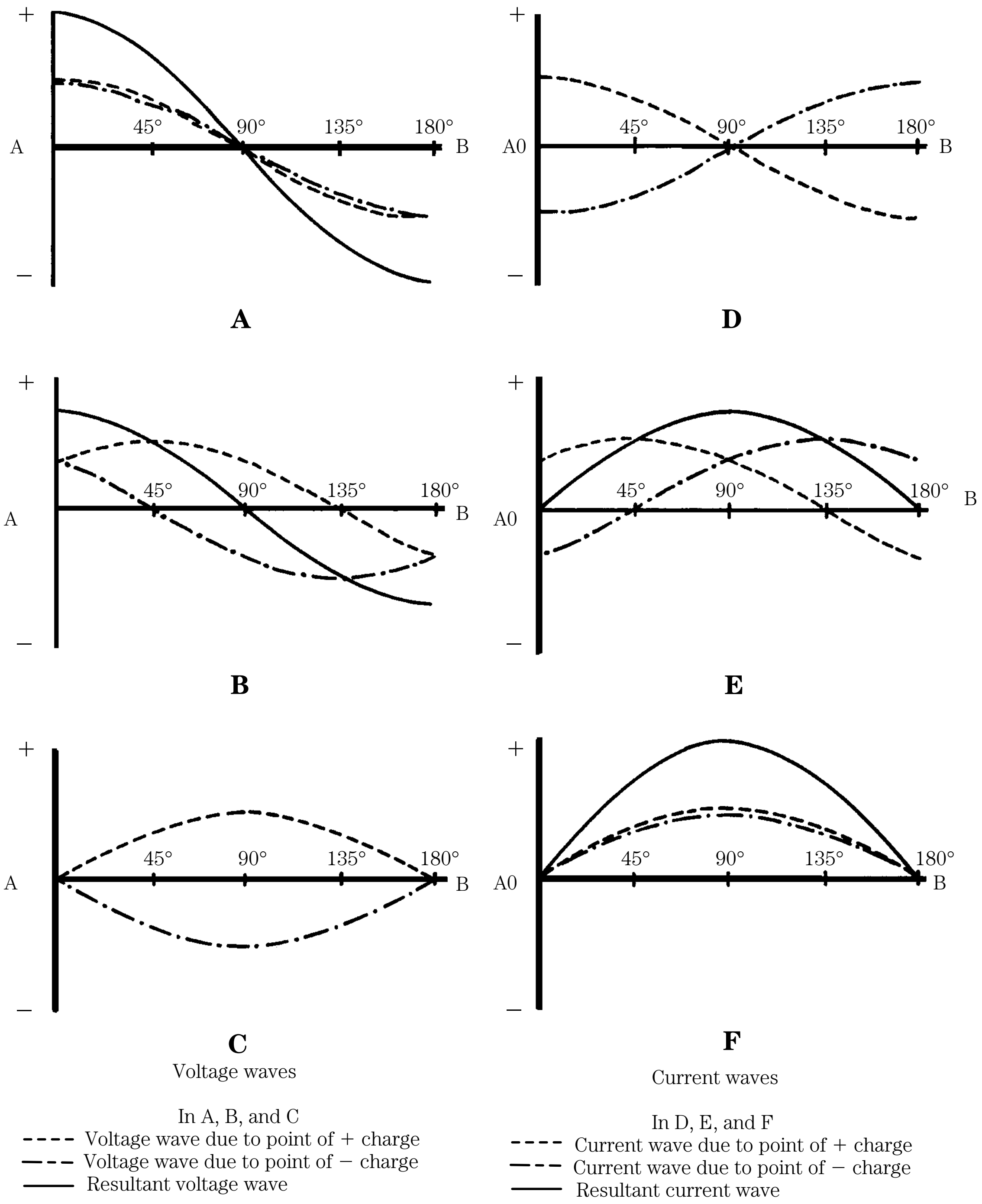
+
−
A
B
45°
90°
135°
180°
A0
A
B
C
D
E
F
Voltage waves
Current waves
In A, B, and C
Voltage wave due to point of + charge
Voltage wave due to point of − charge
Resultant voltage wave
In D, E, and F
Current wave due to point of + charge
Current wave due to point of − charge
Resultant current wave

5-11 Standing waves of voltage and current.

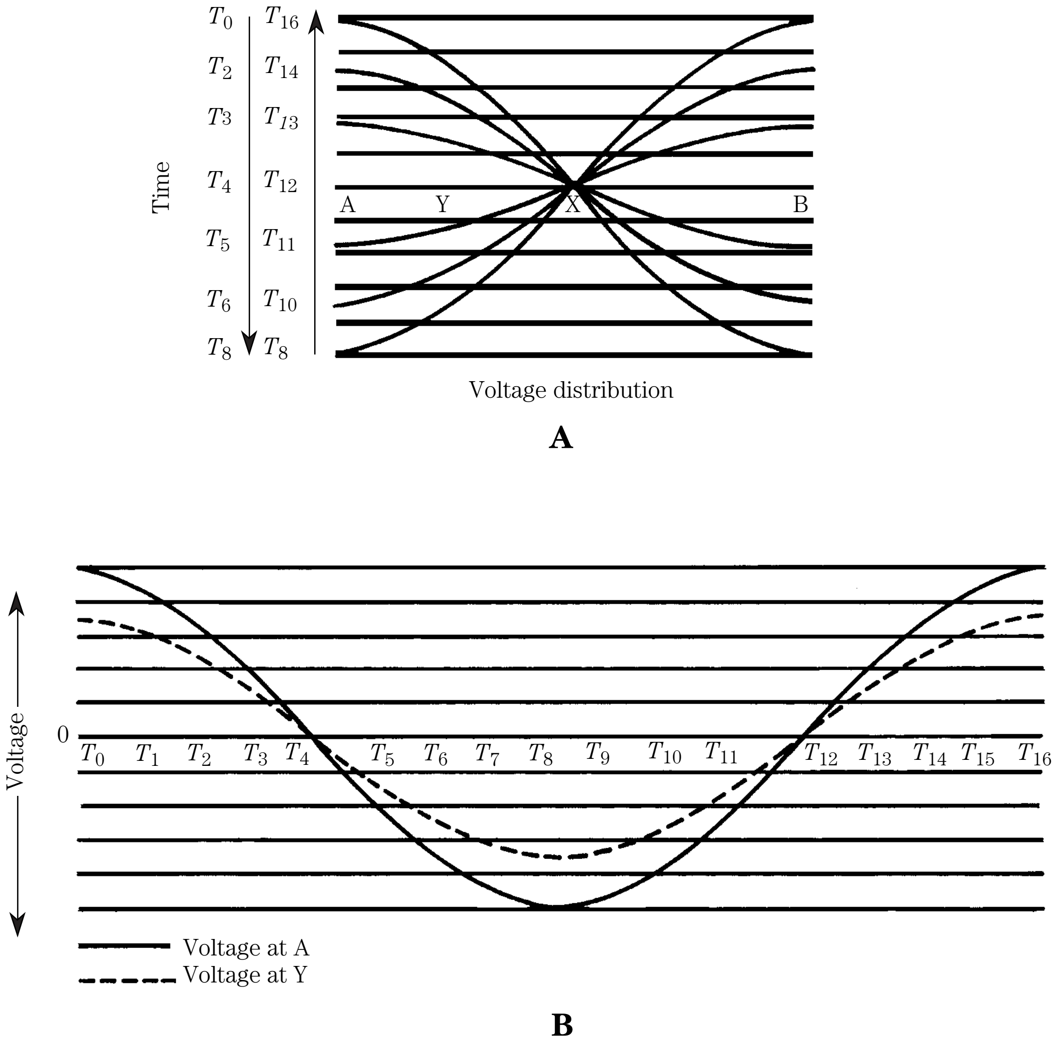

5-12 Standing waves of voltage at a point on the antenna.

The peak amplitude of the sine wave produced at any point depends on its position on the antenna. The nearer the point is to either end, the greater its peak amplitude.

Standing waves of current The standing waves of current occurring at various times through the RF cycle are assembled on a single axis in Fig. 5-13. This axis, AB, represents the half-wave antenna. If the current variations at point Y from T_0 to T_{16} are graphed in respect to time, the result is the sine wave in Fig. 5-13B. This is true for any point along the antenna with the exception of the nodes at the ends. The current has its greatest swing at *X*, the center of the antenna. Comparison of the voltage variation curve (Fig. 5-12A) with the current variation curve (Fig. 5-13A) shows the voltage curve leading the current curve by 90° at *Y*. This relation can be expected on any half-wave device.

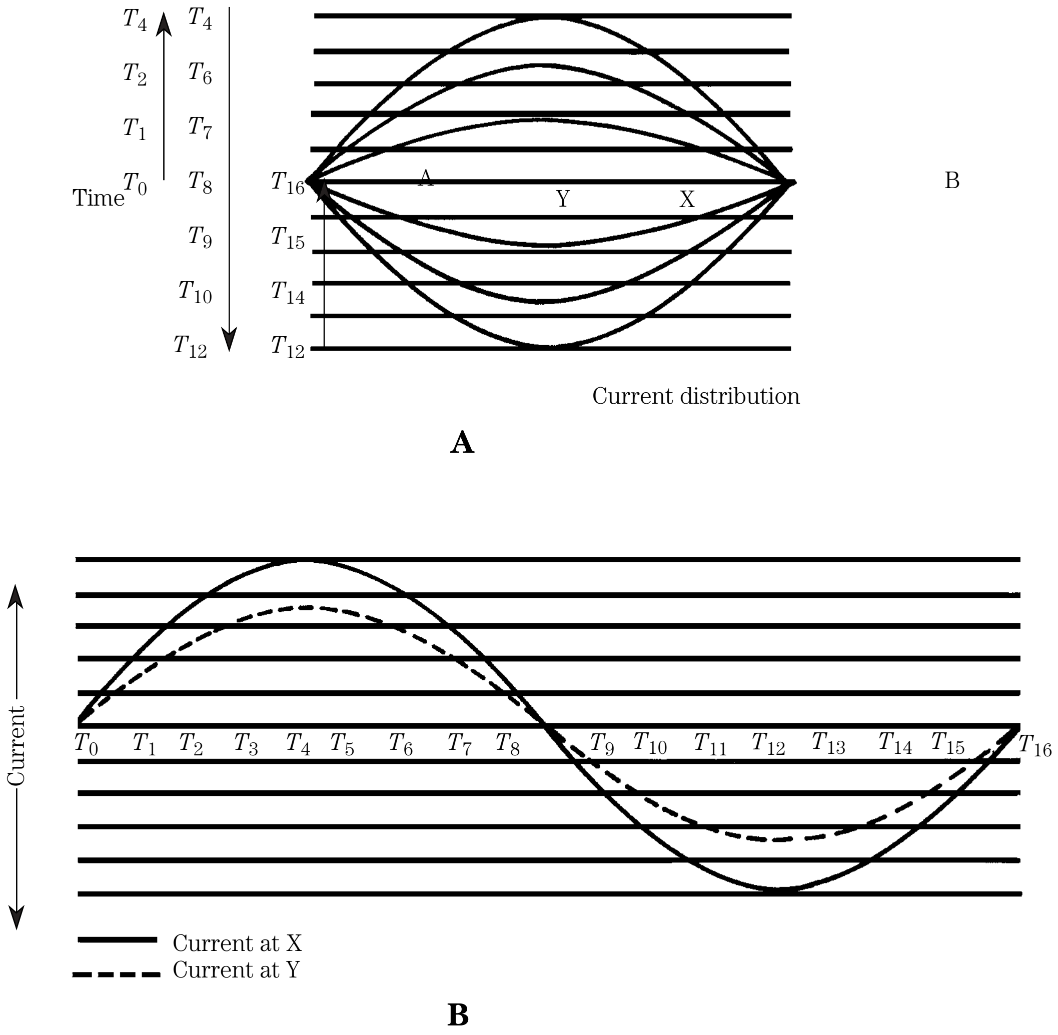
T4 T2 T1 T0 Time
T4 T6 T7 T8 T9 T10 T12
T16 T15 T14 T12
A
Y
X
B
Current distribution
A
Current
T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12 T13 T14 T15 T16
Current at X
Current at Y
B

5-13 Standing waves of current at a point on the antenna.

Measurement of standing waves In Fig. 5-14, the standing waves of voltage *E* and current *I* are indicated along the antenna. There are current nodes at *A* and *B* and a voltage node at *X*. These standing waves are found on any half-wave antenna. A meter that indicates the effective value (0.707 of peak) of the ac signal can be used to measure the standing waves present on the half-wave antenna.

Velocity of propagation and antenna length

In free space, electromagnetic waves travel at a constant velocity of 300,000 km (or approximately 186,000 mi) per second. The RF energy on an antenna, however, moves at a velocity considerably less than that of the radiated energy in free space

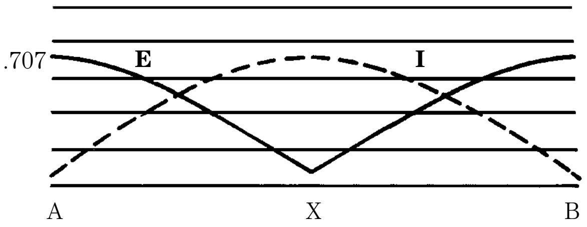

5-14 Standing waves measured with a meter.

because the antenna has a *dielectric constant* greater than that of free space. Because the dielectric constant of free space (air or vacuum) is approximately 1, a dielectric constant greater than 1 retards electromagnetic-wave travel.

Because of the difference in velocity between the wave in free space and the wave on the antenna, the *physical* length of the antenna no longer corresponds to its *electrical* length. The antenna is a half-wavelength electrically, but somewhat shorter than this physically. This is shown in the formula for the velocity of electromagnetic waves,

$$V = f\lambda \qquad \textbf{[5.1]}$$

where V is the velocity, f is the frequency, and λ is the wavelength. Since the frequency of the wave remains constant, a decrease in the velocity results in a decrease in the wavelength. Therefore, the wave traveling in an antenna has a shorter wavelength than the same wave traveling in free space, and the physical length of the antenna can be shorter.

The actual difference between the physical length and the electrical length of the antenna depends on several factors. A thin wire antenna, for example, has less effect on wave velocity than an antenna with a large cross section. As the circumference of the antenna increases, the wave velocity is lowered, as compared with its free-space velocity. The effect of antenna circumference on wave velocity is illustrated in the graph of Fig. 5-15.

Other factors are involved that lower wave velocity on the antenna. Stray capacitance, for example, increases the dielectric constant and lowers wave velocity. This capacitance can be caused by the line connecting the antenna to the transmitter, the insulators used to give physical support to the antenna, or nearby objects made of metallic or dielectric materials. The change in velocity resulting from stray capacitance is called *end effect* because the ends of the antenna are made farther apart electrically than they are physically. End effect is counteracted by making the physical length about 5 percent shorter than the electrical length, as expressed in the formula

$$\begin{aligned} L &= 0.95\left(\frac{492}{f}\right) \\ &= \frac{468}{f} \end{aligned} \qquad \textbf{[5.2]}$$

where L is the physical length in feet and f is the frequency in megahertz. This formula is accurate for all practical purposes in determining the physical length of a half-wavelength antenna at the operating frequency.

The capacitive end effect also slightly changes the standing waves of voltage and current. When the standing waves are measured, it is found that the nodes have some value and do not reach zero, because some current is necessary to charge the stray capacitance. The standing waves measured in Fig. 5-16 show the results of end effect.

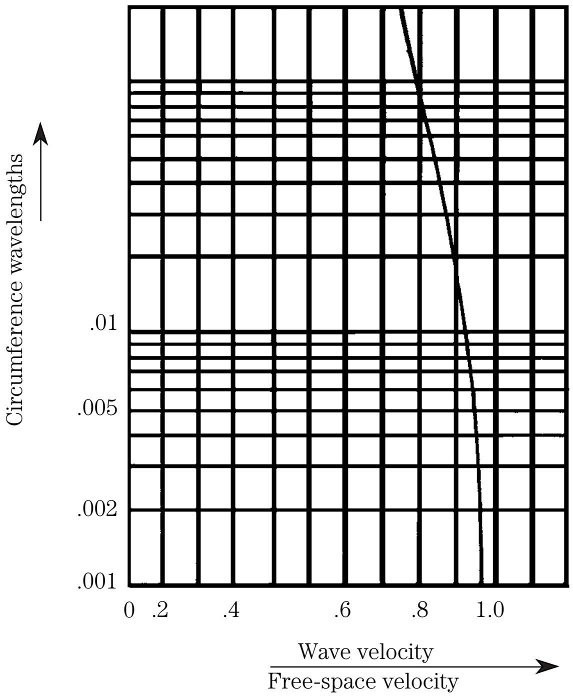

5-15 Effect of antenna circumference on wave velocity.

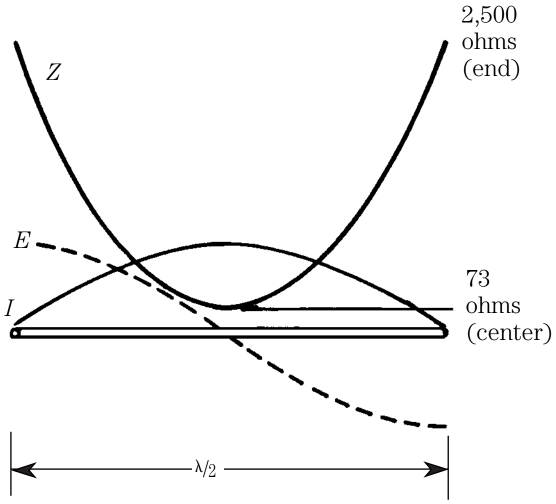

5-16 Impedance along half-wave antenna.

Resonance, resistance, and impedance

The antenna is a circuit element having distributed constants of inductance, capacitance, and resistance, which can be made to form a resonant circuit. The half-wave antenna is the shortest resonant length of antenna. However, antennas which are two or more half-wavelengths can also be resonant. Such antennas are said to operate on *harmonics*. If an antenna is four half-wavelengths at the transmitter frequency, it is being operated at the fourth harmonic of its lowest resonant frequency. In other words, this antenna is a half-wavelength at one-quarter of the frequency of operation. An antenna operating on the third harmonic is shown in Fig. 5-7.

Resistance A current flowing in the antenna must contend with three kinds of resistance. With the antenna considered as a radiator of energy, the power expended in the form of radiation can be thought of as an I^2R_t loss. R_t is called the *radiation resistance*. With the antenna considered as a conductor, a certain amount of energy is dissipated in the form of heat. In this I^2R_0 loss, R_0 is the *ohmic resistance*. There is also an I^2R loss because of the *leakage resistance* of dielectric elements, such as insulators. This R usually is included in the ohmic resistance.

The purpose of the antenna is to dissipate as much energy as possible in the form of radiation. The energy dissipated by the radiation resistance, therefore, is the useful part of the total power dissipated. Because the actual power loss depends on the ohmic resistance, this resistance should be kept as low as possible. In the half-wave antenna, the radiation resistance is large compared to the ohmic resistance, and most of the available energy is radiated. The half-wave antenna is, therefore, a very efficient radiator for most purposes.

For a half-wave antenna fed at the center point, the radiation resistance is equal to 73 Ω. The reference point is the center of the antenna at the time of peak current flow. Ohmic resistance is referred to this point. The total resistance is of importance in matching the antenna to a transmission line.

Impedance Because the half-wave antenna has different conditions of voltage and current at different points, and because impedance is equal to the voltage across a circuit divided by the current through it, the impedance will vary along the length of the antenna. If E is divided by I at each point of the voltage and current curves in Fig. 5-16, the result is the impedance curve, Z. The impedance is about 73 Ω at the center point and rises to a value of about 2500 Ω at the ends.

The impedance of the half-wave antenna usually is considered to be the impedance as *seen* by the transmitter at the input terminals. This impedance consists of both resistance and reactance. If the antenna is cut to a length of exact resonance, the reactance is zero and the impedance is purely resistive. However, if the antenna is longer or shorter than resonance, reactance is present. When the antenna is made shorter, capacitive reactance is present; when the antenna is made longer, inductive reactance is present.

The impedance at the antenna input terminals is important in terms of power efficiency. If the transmitter is feeding a nonresonant antenna, a power loss is caused by the reactive component of the antenna impedance. Conversely, if the frequency of the transmitter is changed, the electrical length of the antenna also changes. If the frequency is made somewhat higher, the electrical length is made greater, and inductive reactance is added to the impedance. If the frequency is lowered, the electrical length is shortened, and capacitive reactance is added to the impedance.

6
CHAPTER

High-frequency dipole and other doublet antennas

A MYTH AROSE IN RADIO COMMUNICATIONS CIRCLES SOME TIME AGO. PEOPLE CAME TO believe (especially in the ham and CB communities) that large antenna arrays are absolutely necessary for effective communications, especially over long distances. Overlooked, almost to the point of disdain, were effective (but simple) antennas that can be erected by inexperienced people and made to work well. The simple *dipole*, or *doublet*, is a case in point. This antenna is also sometimes called the *Hertz*, or *hertzian*, antenna because radio pioneer Heinrich Hertz reportedly used this form in his experiments.

The half-wavelength dipole is a balanced antenna consisting of two radiators (Fig. 6-1) that are each a quarter-wavelength, making a total of a half-wavelength. The antenna is usually installed horizontally with respect to the earth's surface, so it produces a horizontally polarized signal.

In its most common configuration (Fig. 6-1), the dipole is supported at each end by rope and end insulators. The rope supports are tied to trees, buildings, masts, or some combination of such structures.

The length of the antenna is a half-wavelength. Keep in mind that the physical length of the antenna, and the theoretical electrical length, are often different by about 5 percent. A free-space half-wavelength is found from

$$L = \frac{492}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[6.1]}$$

In a perfect antenna, that is self-supported many wavelengths away from any object, Eq. 6.1 will yield the physical length. But in real antennas, the length calculated

Copyright 2001 - the McGraw-Hill Companies

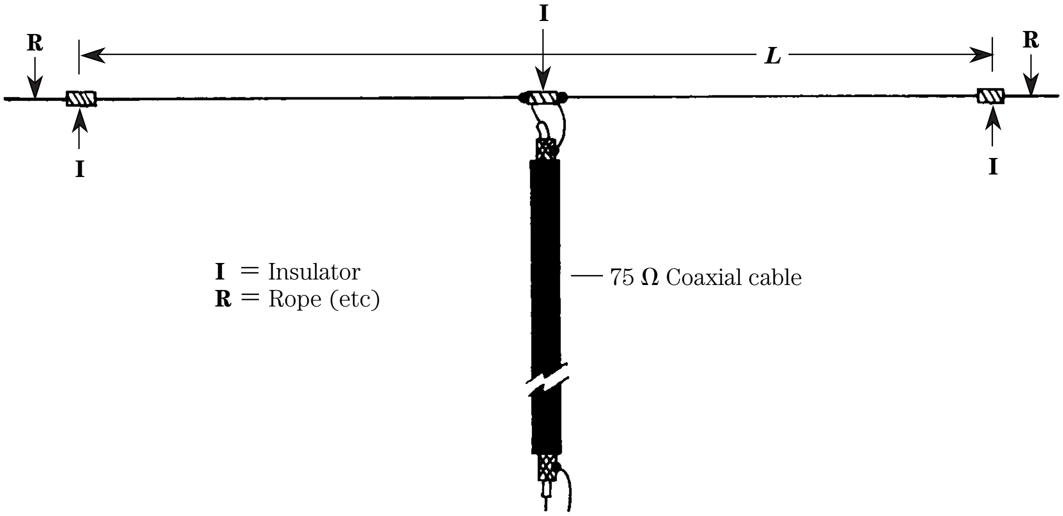

6-1 Simple half-wave dipole antenna.

above is too long. The average physical length is shortened by up to about 5 percent because of the velocity factor of the wire and capacitive effects of the end insulators. A more nearly correct *approximation* (remember that word, it's important) of a half-wavelength antenna is

$$L = \frac{468}{F_{\mathrm{MHz}}}\ \mathrm{ft} \tag{6.2}$$

where

L is the length of a half-wavelength radiator, in feet
F_{MHz} is the operating frequency, in megahertz

Example Calculate the approximate physical length for a half-wavelength dipole operating on a frequency of 7.25 MHz.

Solution:

$$\begin{aligned} L &= \frac{468}{F_{\mathrm{MHz}}}\ \mathrm{ft} \\ &= \frac{468}{7.25}\ \mathrm{ft} = 64.55\ \mathrm{ft} \end{aligned}$$

or, restated another way:

$$L = 64\ \mathrm{ft}\ 6.6\ \mathrm{in}$$

It is unfortunate that a lot of people accept Eq. 6.2 as a universal truth, a kind of immutable law of The Universe. Perhaps abetted by books and articles on antennas that fail to reveal the full story, too many people install dipoles without regard for reality. The issue is *resonance*. An antenna is a complex *RLC* network. At some frequency, it will appear like an inductive reactance ($X = +jX_L$), and at others it will appear like a capacitive reactance ($X = -jX_C$). At a specific frequency, the reactances are equal in magnitude, but opposite in sense, so they cancel each other out: $X_L - X_C = 0$. At this frequency, the impedance is purely resistive, and the antenna is said to be resonant.

The goal in erecting a dipole is to make the antenna resonant at a frequency that is inside the band of interest, and preferably in the portion of the band most often used by the particular station. Some of the implications of this goal are covered later on, but for the present, assume that the builder will have to custom-tailor the length of the antenna. Depending on several local factors (among them, nearby objects, the shape of the antenna conductor, and the length/diameter ratio of the conductor) it might prove necessary to add, or trim, the length a small amount to reach resonance.

The dipole feedpoint

The dipole is a half-wavelength antenna fed in the center. Figure 6-2 shows the voltage (V) and current (I) distributions along the length of the half-wavelength radiator element. The feedpoint is at a voltage minimum and a current maximum, so you can assume that the feedpoint is a current antinode.

At resonance, the impedance of the feedpoint is $R_o = V/I$. There are two resistances that make up R_o. The first is the ohmic losses that generate nothing but heat when the transmitter is turned on. These ohmic losses come from the fact that conductors have electrical resistance and electrical connections are not perfect (even when properly soldered). Fortunately, in a well-made dipole these losses are almost negligible. The second contributor is the *radiation resistance* R_r of the antenna. This resistance is a hypothetical concept that accounts for the fact that RF power is radiated by the antenna. The radiation resistance is the fictional resistance that would dissipate the amount of power that is radiated away from the antenna.

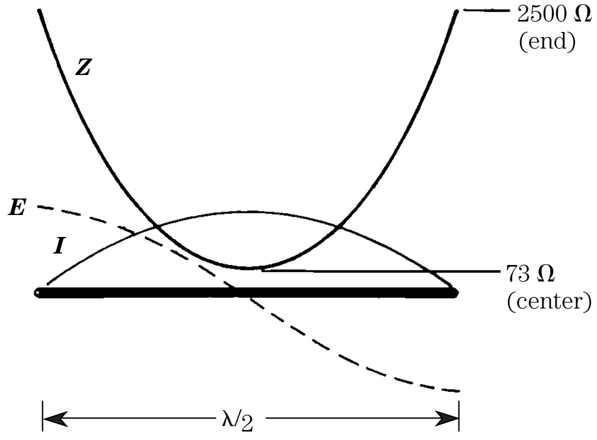

6-2 Graph of current, voltage, and impedance along half-wavelength dipole.

For example, suppose we have a large-diameter conductor used as an antenna, and it has negligible ohmic losses. If 1000 W of RF power is applied to the feedpoint, and a current of 3.7 A is measured, what is the radiation resistance?

$$R_r = \frac{P}{I^2} \qquad \textbf{[6.3]}$$

$$= \frac{1000\ \text{W}}{(3.7)^2} \qquad \textbf{[6.4]}$$

$$= 73\ \Omega \qquad \textbf{[6.5]}$$

It is always important to match the feedpoint impedance of an antenna to the transmission-line impedance. Maximum power transfer always occurs (in any system) when the source and load impedances are matched. In addition, if some applied power is not absorbed by the antenna (as happens in a mismatched system), then the unabsorbed portion is reflected back down the transmission line toward the transmitter. This fact gives rise to standing waves, and the so-called standing wave ratio (SWR or VSWR) discussed in Chap. 3. This is a problem to overcome.

Matching antenna feedpoint impedance seems to be simplicity itself because the free-space feedpoint impedance of a simple dipole is about 73 **Ω**, seemingly a good match to 75-**Ω** coaxial cable. Unfortunately, the 73-**Ω** feedpoint impedance is almost a myth. Figure 6-3 shows a plot of approximate radiation resistance (R_r) versus height above ground (as measured in wavelengths). As before, we deal in approximations in Fig. 6-3; in this case, the ambiguity is introduced by ground losses.

Despite the fact that Fig. 6-3 is based on approximations, you can see that radiation resistance varies from less than 10 **Ω**, to around 100 **Ω**, as a function of height. At heights of many wavelengths, this oscillation of the curve settles down to the free-space impedance (72 **Ω**). At the higher frequencies, it might be possible to install a dipole at a height of many wavelengths. In the 2-m amateur radio band (144 to 148 MHz), one wavelength is around 6.5 ft (i.e., 2 m × 3.28 ft/m), so "many wavelengths" is relatively easy to achieve at reasonably attainable heights. In the 80-m band (3.5 to 4.0 MHz), however, one wavelength is on the order of 262 ft, so "many wavelengths" is a practical impossibility.

There are three tactics that can be followed. First, ignore the problem altogether. In many installations, the height above ground will be such that the radiation resistance will be close enough to present only a slight impedance mismatch to a standard coaxial cable. The VSWR is calculated (among other ways) as the ratio:

1. $Z_o > R_r$:

$$\text{VSWR} = \frac{Z_o}{R_r} \qquad \textbf{[6.6]}$$

2. $Z_o < R_r$:

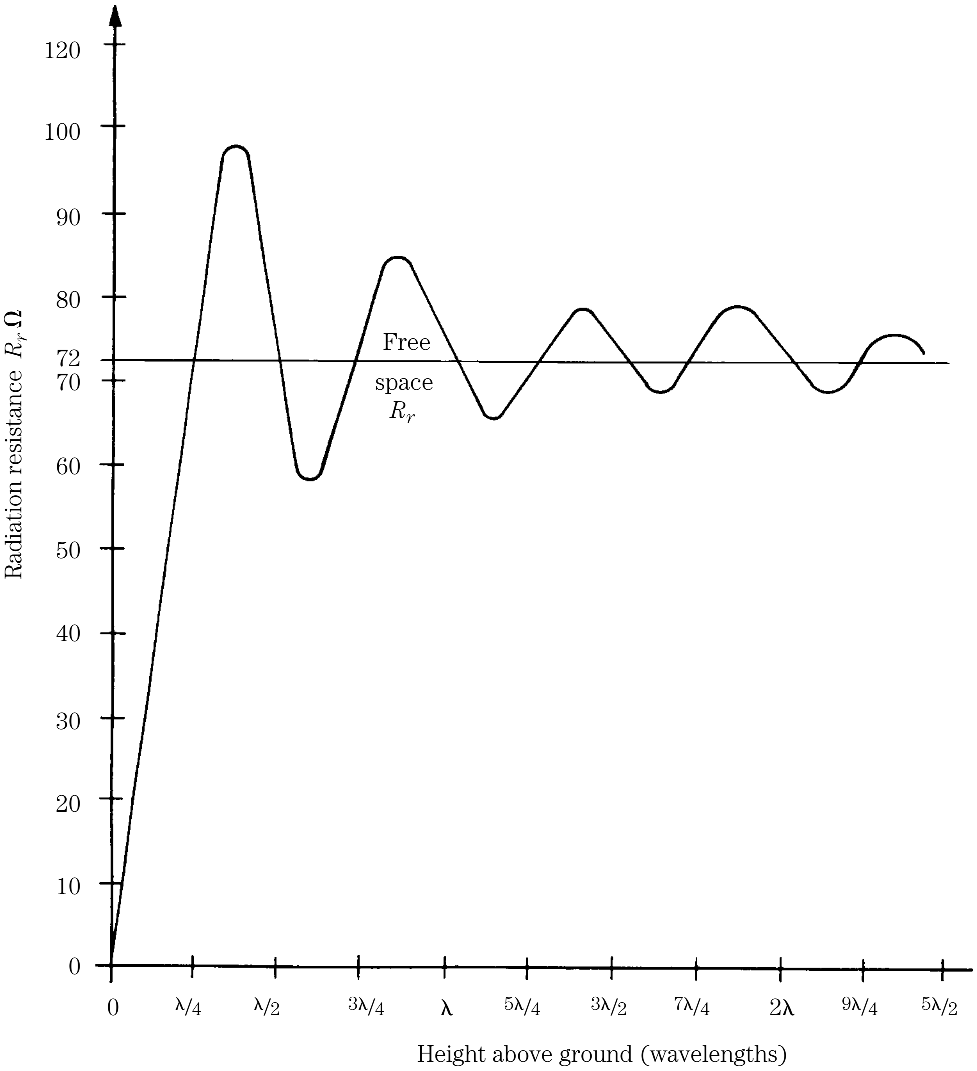

6-3 Radiation resistance versus height above ground.

$$\text{VSWR} = \frac{R_r}{Z_o} \qquad \textbf{[6.7]}$$

where

Z_o is the coaxial-cable characteristic impedance
R_r is the radiation resistance of the antenna

Consider an antenna mounted at a height somewhat less than a quarter-wavelength, such that the radiation resistance is 60 **Ω**. Although not recommended as good engineering practice (there are sometimes practical reasons) it is nonetheless necessary to install a dipole at less than optimum height. So, if that becomes necessary, what are the implications of feeding a 60-Ω antenna with either 52- or 75-Ω standard coaxial cable? Some calculations are revealing:

For 75-Ω coaxial cable:

$$\text{VSWR} = \frac{Z_o}{R_r} \qquad \textbf{[6.8]}$$

$$= \frac{75\ \Omega}{60\ \Omega} = 1.25{:}1 \qquad \textbf{[6.9]}$$

For 52-Ω coaxial cable:

$$\text{VSWR} = \frac{R_r}{Z_o} \qquad \textbf{[6.10]}$$

$$= \frac{60\ \Omega}{52\ \Omega} = 1.15{:}1 \qquad \textbf{[6.11]}$$

In neither case is the VSWR created by the mismatch too terribly upsetting.

The second approach is to mount the antenna at a convenient height, and use an impedance matching scheme to reduce the VSWR. In Chap. 23, you will find information on various suitable (relatively) broadbanded impedance matching methods including Q-sections, coaxial impedance transformers, and broadband RF transformers. “Homebrew” and commercially available transformers are available to cover most impedance transformation tasks.

The third approach is to mount the antenna at a height (Fig. 6-3) at which the expected radiation resistance crosses a standard coaxial cable characteristic impedance. The best candidate seems to be a height of a half-wavelength because the radiation resistance is close to the free-space value of 72 Ω, and is thus a good match for 75-Ω coaxial cable (such as RG-11/U or RG-59/U).

The dipole radiation pattern

The radiation patterns of various antennas are covered in this book. Some basic theory of patterns is repeated at each section, not to fill space, but rather to drive home a point and refresh reader memories. We keep harking back to the concepts *directivity* and *gain* (which are actually different expressions of the same thing).

Antenna theory recognizes a point of reference called the *isotropic radiator*. This device is a theoretical construct consisting of a spherical point source of RF radiation in all directions (see Chap. 2). It is truly *omnidirectional* (“all directions”) because it creates an ever-expanding sphere as the RF wavefront propagates outward. Antenna gain is a measure of how the antenna focuses available power away from a spherical wavefront to a limited number of directions (two, in the case of the dipole). Thus are related the concepts *directivity* and *gain*.

Always keep in mind that *directivity and gain are specified in three dimensions*. Too many times, people simplify the topic too much by publishing only part of the radiation pattern (azimuth aspect). In other words, the reader is given a pattern viewed from above that shows the directivity in the horizontal plane. But a signal does not propagate away from an antenna in an infinitely thin sheet, as such presentations seem to imply; rather, it has an elevation extent in addition to the azimuth extent. Thus, proper evaluation of an antenna takes into consideration both horizontal and vertical plane patterns.

Figure 6-4 shows the radiation pattern of a dipole antenna in free space "in the round." In the horizontal plane (6-4A), when viewed from above, the pattern is a "figure

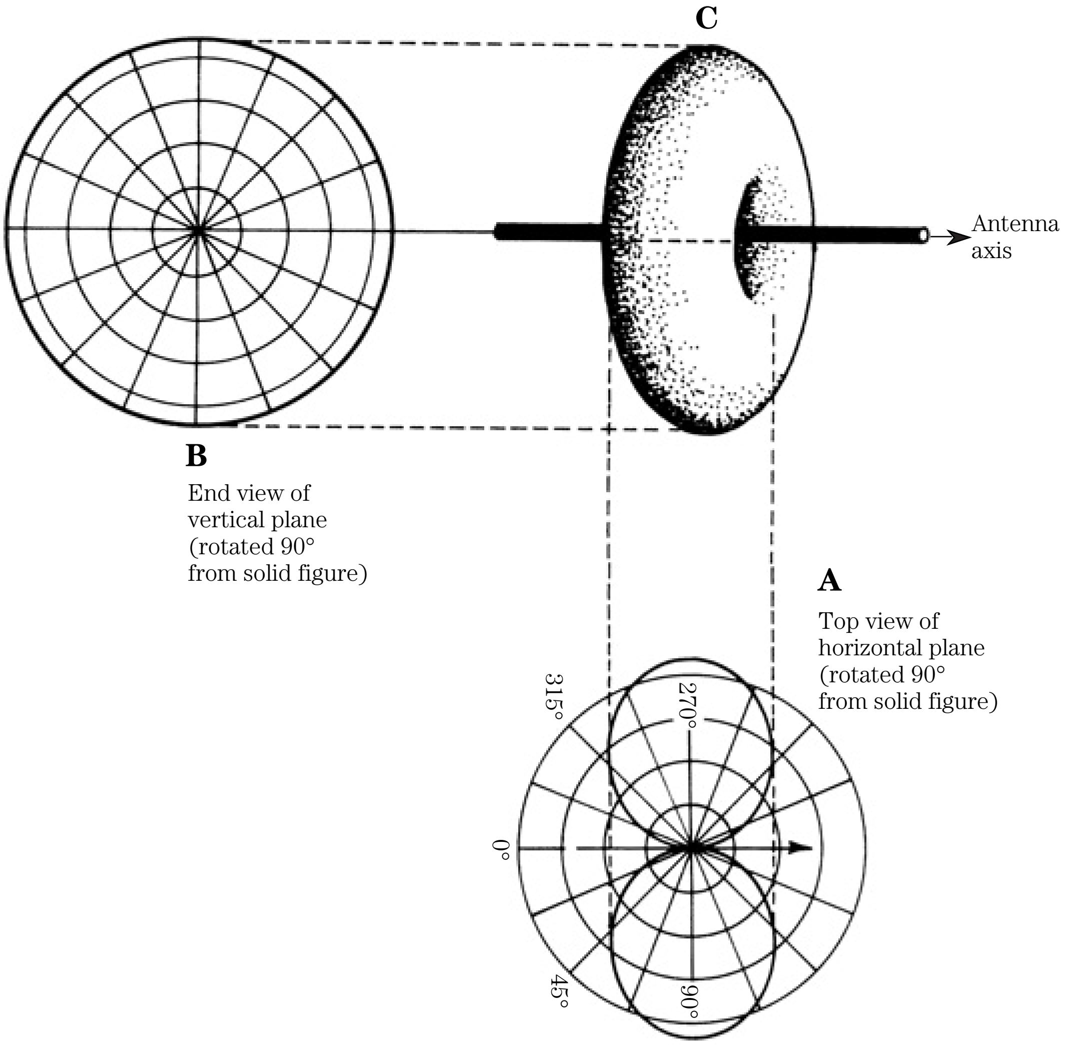
C
Antenna axis
B
End view of vertical plane (rotated 90° from solid figure)
A
Top view of horizontal plane (rotated 90° from solid figure)
315°
270°
0°
45°
90°

6-4 Idealized dipole radiation pattern.

8" that exhibits bidirectional radiation. Two main "lobes" contain the RF power from the transmitter, with sharp nulls of little or no power off the ends of the antenna axis. This pattern is the classical dipole pattern that is published in most antenna books.

Also shown, however, is the vertical plane pattern for a dipole antenna in free space. Note that when sliced in this aspect the radiation pattern is circular (Fig. 6-4B). When the two patterns are combined in the round, you can see the three-dimensional doughnut-shaped pattern (Fig. 6-4C) that most nearly approximates the true pattern of an unobstructed dipole in free space.

When a dipole antenna is installed close to the earth's surface, not in free space, as is the case at most stations, the pattern is distorted from that of Fig. 6-4. Two effects must be taken into consideration. First, and most important, is the fact that the signal from the antenna is reflected from the surface and bounces back into space. This signal will be phase-shifted both by the reflection and by the time required for the transit to occur. At points where the reflected wave combines in phase with the radiated signal, the signal is reinforced; in places where it combines out of phase, the signal is attenuated. Thus, the reflection of the signal from the ground alters the pattern from the antenna. The second factor is that the ground is lossy, so not all of the signal is reflected; some of it heats the ground underneath the antenna. Thus, the signal is attenuated at a greater rate than the inverse square law, so it further alters the expected pattern.

Figure 6-5 shows patterns typical of dipole antennas installed close to the earth's surface. The views in this illustration correspond to Fig. 6-4B, in that they are looking at the vertical plane from a line along the antenna axis. Therefore, the antenna is represented by *R* in each case shown. Figure 6-5A shows the case for a dipole

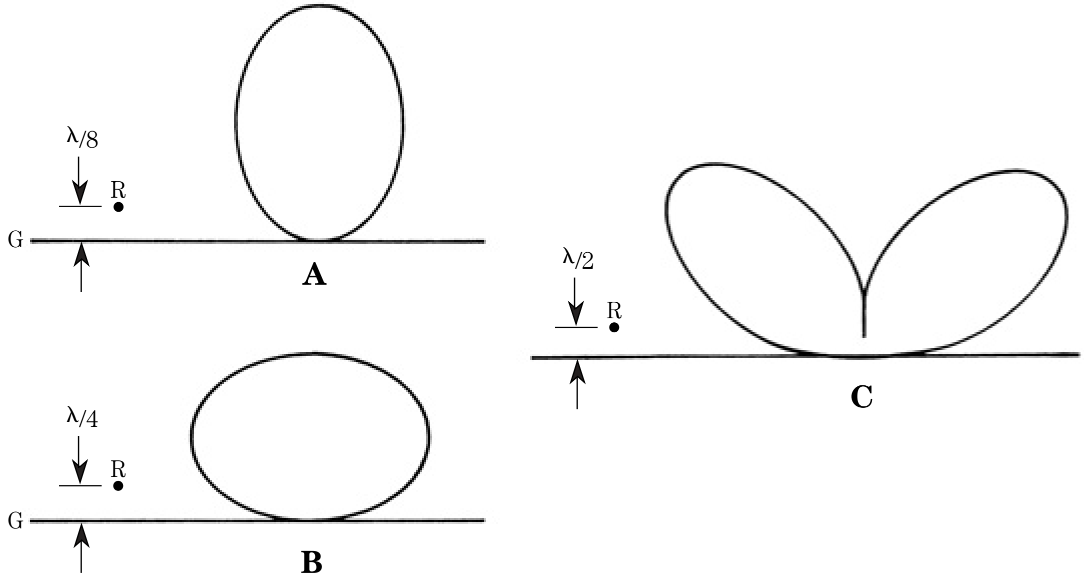

6-5 Vertical extent of dipole antenna at (A) ⅛ wavelength, (B) ¼ wavelength, and (C) ½ wavelength above ground.

installed at 1/8 wavelength above the surface. For this antenna, most of the RF energy is radiated almost straight up (not very useful). This type of antenna is basically limited to ground wave and very short skip (when available). The second case (Fig. 6-5B) shows the pattern when the antenna is a quarter wavelength above the earth's surface. Here the pattern is flattened, but it still shows considerable energy in the vertical direction (where it is useless). Finally, you can see the pattern obtained when the antenna is installed a half-wavelength above the surface. In this case, the pattern is best for long-distance work because energy is redirected away from straight-up into lobes at relatively shallow angles.

Tuning the dipole antenna

There are two issues to address when tuning an antenna (any antenna, not just the dipole): *resonance* and *impedance matching*. Although frequently treated in the literature as the same issue, they are not. This section deals mostly with the process of tuning the antenna to resonance. Although not all forms of antenna are resonant, the dipole *is* an example of a resonant antenna.

There is a lot of misinformation abroad concerning the tuning of antennas. Perhaps much of what is believed comes from the fact that VSWR is used as the indicator of both impedance matching and resonance. Quite a few people honestly, but erroneously, believe that the VSWR can be "tuned out" by adjusting the length of the feedline. That myth probably derives from the fact that voltage or current sensing instruments are used for VSWR measurement, and these are affected by transmission line length. But that fact is caused by a weakness in the instruments, not by radio physics.

There is only one proper way to tune a dipole antenna: *adjust the length of the antenna elements*, not the transmission line. It was in order to make these adjustments that we purposely did not initially tell you to solder the electrical connections at the center insulator.

Resonance The indicator of resonance is the minimum point in the VSWR curve. Figure 6-6 shows a graph of VSWR vs. frequency for several different cases. Curve *A* represents a disaster: a high VSWR all across the band. The actual value of VSWR can be anything from about 3.5:1 to 10:1, or thereabouts, but the cause is nonetheless the same: the antenna is either open or shorted; or it is so far off resonance as to appear to be open or shorted to the VSWR meter.

Curves *B* and *C* represent antennas that are resonant within the band of interest. Curve *B* represents a broadbanded antenna that is relatively flat all across the band and does not exhibit excessive VSWR until the frequency is outside of the band. Curve *C* is also resonant within the band, but this antenna has a much higher *Q* than curve *B.* In the naive sense, the broadbanded antenna is best, but that statement is true only if the broadness is not purchased at the expense of efficiency. Losses tend to broaden the antenna, but also reduce its effectiveness. So, if broad bandedness is purchased at the risk of increased loss, then it is less than desirable.

Curves *D* and *E* are resonant outside the band of interest. The curve marked *D* is resonant at a frequency on the low side of the band, so the dipole is too long. In this case, you need to shorten the antenna a bit to raise the resonant point inside

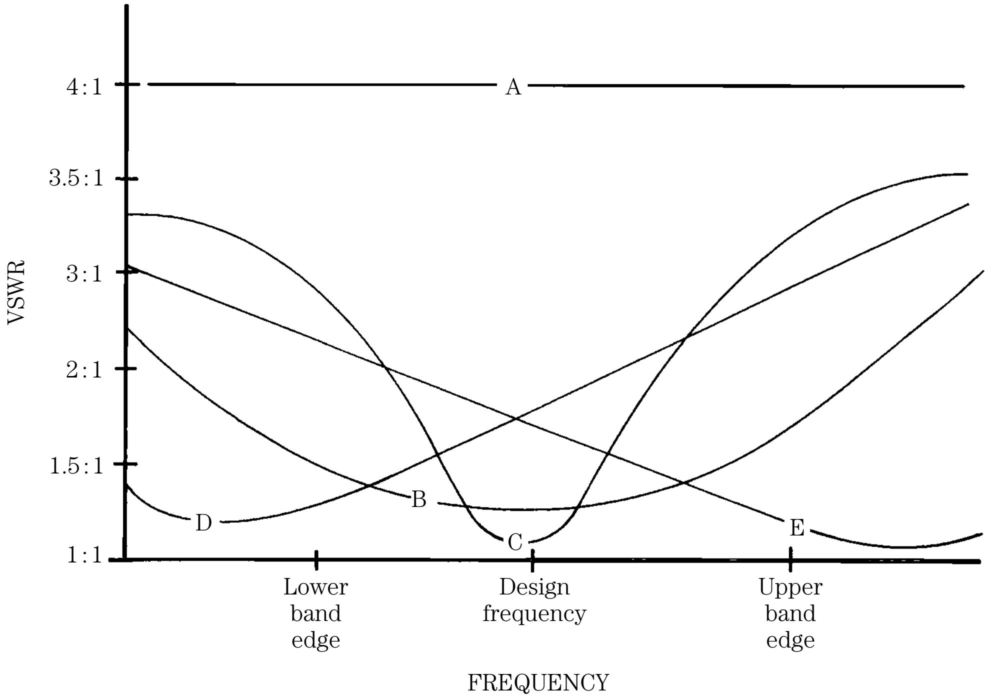

6-6 VSWR versus frequency for several cases.

the band. Curve *E* represents an antenna that is resonant outside the upper limit of the band, so this antenna is too short, and must be lengthened. Because this situation is possible, the antenna elements are typically made longer than needed when they are first cut.

How much to cut? That depends on two factors: how far from the desired frequency the resonant point is found, and which band is being used. The latter requirement comes from the fact that the "frequency per unit length" varies from one band to another. Look at an example of how to calculate this figure. The procedure is simple:

1. Calculate the length required for the upper end of the band.
2. Calculate the length required for the bottom end of the band.
3. Calculate the difference in lengths for the upper and lower ends of the band.
4. Calculate the width of the band in kilohertz by taking the difference between the upper frequency limit and the lower frequency limit.
5. Divide the length difference by the frequency difference; the result is in kilohertz per unit length.

Example 6-1 Calculate the frequency change per unit of length for 80 and for 15 m.

Solution For 80 m (3.5 to 4.0 MHz):

1. $L_{\text{ft}} = \dfrac{468}{4\text{ MHz}} = 117\text{ ft}$
2. $L_{\text{ft}} = \dfrac{468}{3.5\text{ MHz}} = 133.7\text{ ft}$
3. Difference in length: 133.7 ft – 117 ft = 16.7 ft
4. Frequency difference: 4000 kHz – 3500 kHz = 500 kHz
5. Calculate $\dfrac{\text{frequency}}{\text{unit length}} : \dfrac{500\text{ kHz}}{16.7\text{ ft}} = 30\text{ kHz/ft}$

For 15 m (21.0 – 21.45 MHz):

1. $L_{\text{ft}} = \dfrac{468}{21.45} = 21.82\text{ ft}$
2. $L_{\text{ft}} = \dfrac{468}{21} = 22.29\text{ ft}$
3. Difference in length: 22.29 ft – 21.82 ft = 0.47 ft
4. Convert to inches: 0.47 ft × 12 in/ft = 5.64 in
5. Frequency difference: 21,450 kHz – 21,000 kHz = 450 kHz
6. Calculate $\dfrac{\text{frequency}}{\text{unit length}} : \dfrac{450\text{ kHz}}{5.64\text{ in}} = 80\text{ kHz/in}$

At 80 m, the frequency change per foot is small, but at 15 m small changes can result in very large frequency shifts. You can calculate approximately how much to add (or subtract) from an antenna under construction from this kind of calculation. If, for example, you design an antenna for the so-called “net frequency” on 15 m (21,390 kHz), and find the actual resonant point is 21,150 kHz, the frequency shift required is 21,390 – 21,150, or 240 kHz. To determine how much to add or subtract (as a first guess):

1. The factor for 15 m is 80 kHz/in, which is the same as saying 1 in/80 kHz.
2. The required frequency shift is 240 kHz.
3. Therefore:

$$\text{Length change} = 240\text{ kHz} \times \frac{1\text{ in}}{80\text{ kHz}} \qquad \textbf{[6.12]}$$

$$= 3\text{ in} \qquad \textbf{[6.13]}$$

Each side of the antenna must be changed by half of the length calculated above, or 1.5 in. Because the first resonant frequency is less than the desired frequency, the length should be shortened 1.5 in. Once the length is correct, as proven by the VSWR curve, the connections at the center insulator are soldered and made permanent, and the antenna rehoisted to the operating level.

Impedance matching The difference between resonance and impedance matching is seen in the value of the VSWR minimum. While the minimum indicates the resonant point, its value is a measure of the relationship between the feedpoint impedance of the antenna and the characteristic impedance of the transmission line. Earlier in this chapter you learned that

1. $Z_o > R_r$:

$$\mathrm{VSWR} = \frac{Z_o}{R_r} \qquad \textbf{[6.14]}$$

2. $Z_o < R_r$:

$$\mathrm{VSWR} = \frac{R_r}{Z_o} \qquad \textbf{[6.15]}$$

where

Z_o is the coaxial-cable characteristic impedance
R_r is the radiation resistance of the antenna

Although knowledge of the VSWR will not show which situation is true, you can know that there is a high probability that *one* of them is true, and you can experiment to find which is the case. Of course, if the VSWR is less than about 1.5:1 or 2:1, then forget about it—the improvement is not generally worth the expense and cost. When coupled to a transmitter that is equipped with the tunable output network (most tube-type transmitters or final amplifiers), then it can accommodate a relatively wide range of reflected antenna impedances. But modern solid-state final amplifiers tend to be a little more picky about the load impedance. For these transmitters, a coax-to-coax *antenna tuning unit* (ATU) is needed.

Other dipoles

Thus far, the dipoles covered in this chapter have been the classic form, in which a half-wavelength single-conductor radiator element is connected to a coaxial transmission line. This antenna is typically installed horizontally at a half-wavelength above the earth's surface (or wherever convenient if that is impossible). This section looks at other forms of the dipole. Some of these dipoles are in every way the equal of the horizontal dipole, and others are basically compensation antennas in that they are used when a proper dipole is not practical.

Inverted-vee dipole

The *inverted-vee dipole* is a half-wavelength antenna fed in the center like a dipole. By the rigorous definition, the inverted-vee is merely a variation on the dipole theme. But in this form of antenna (Fig. 6-7), the center is elevated as high as possible from the earth's surface, but the ends droop to very close to the surface. Angle a can be almost anything convenient, provided that $a > 90$ degrees; typically, most inverted-vee antennas use an angle of about 120 degrees. Although essentially a compensation antenna for use when the dipole is not practical, many operators believe that it is essentially a better performer on 40 and 80 m in cases where the dipole cannot be mounted at a half-wavelength (64 ft or so).

By sloping the antenna elements down from the horizontal to an angle (as shown in Fig. 6-7), the resonant frequency is effectively lowered. Thus, the antenna will

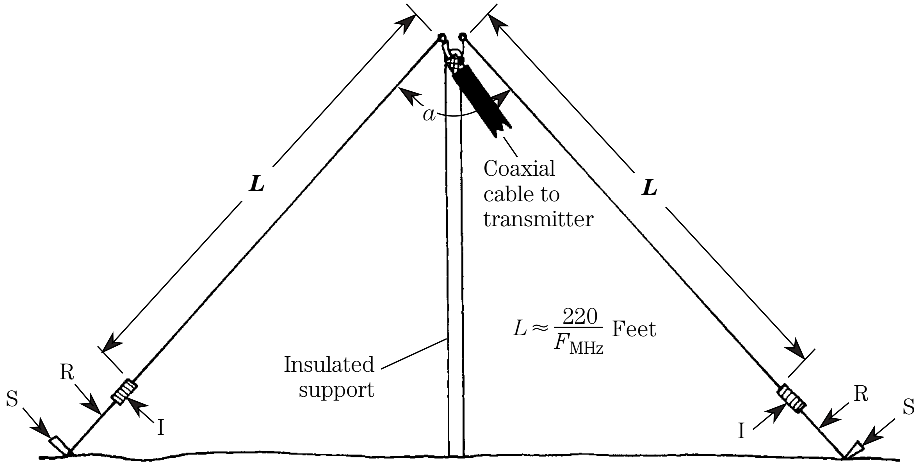
L
L
a
Coaxial cable to transmitter
$L \approx \frac{220}{F_{\text{MHz}}}$ Feet
Insulated support
R
S
I
I
R
S

I = Insulator
S = Support stake
R = Rope

6-7 Inverted-vee dipole.

need to be shorter for any given frequency than a dipole. There is no absolutely rigorous equation for calculation of the overall length of the antenna elements. Although the concept of "absolute" length does not hold for regular dipoles, it is even less viable for the inverted-vee. There is, however, a rule of thumb that can be followed for a starting point: Make the antenna about 6 percent shorter than a dipole for the same frequency. The initial cut of the antenna element lengths (each quarter wavelength) is

$$L = \frac{220}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[6.16]}$$

After this length is determined, the actual length is found from the same cut-and-try method used to tune the dipole in the previous section.

Bending the elements downward also changes the feedpoint impedance of the antenna and narrows its bandwidth. Thus, some adjustment in these departments is in order. You might want to use an impedance matching scheme at the feedpoint, or an antenna tuner at the transmitter.

Sloping dipole ("sloper" or "slipole")

The *sloping dipole* (Fig. 6-8) is popular with those operators who need a low angle of radiation, and are not overburdened with a large amount of land to install the antenna. This antenna is also called the *sloper* and the *slipole* in various texts. The author prefers the term "slipole," in order to distinguish this antenna from a sloping vertical of the same name. Whatever it is called, however, it is a half-wavelength dipole that is built with one end at the top of a support, and the other end close to the

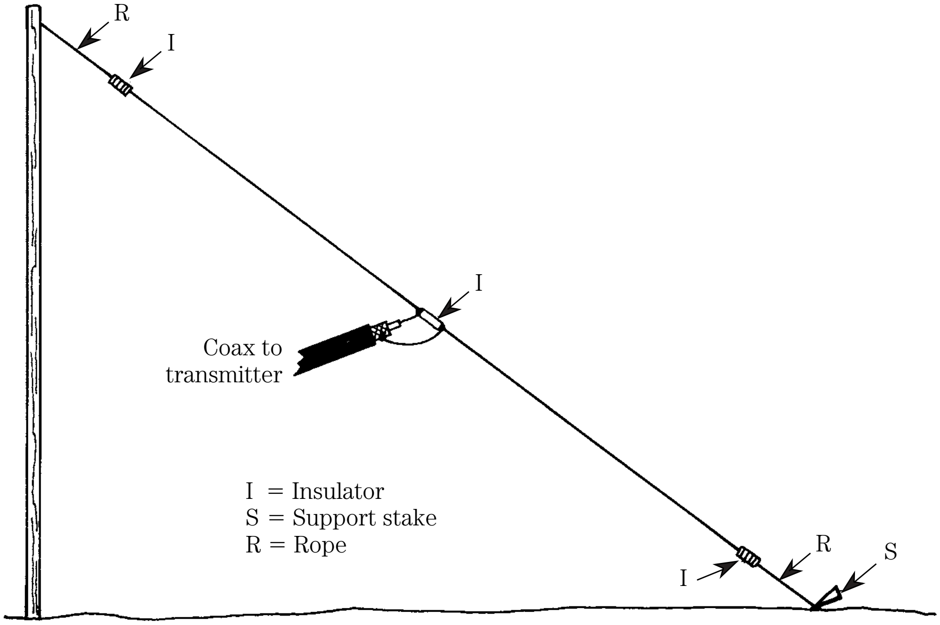

6-8 "Slipole" or "sloper."

ground, and being fed in the center by coaxial cable. Some of the same comments as obtained for the inverted-vee antenna also apply to the sloping dipole, so please see that section also.

Some operators like to arrange four sloping dipoles from the same mast such that they point in different directions around the compass (Fig. 6-9). A single four-position coaxial cable switch will allow switching a directional beam around the compass to favor various places in the world.

Broadbanded dipoles

One of the rarely discussed aspects of antenna construction is that the length/diameter ratio of the conductor used for the antenna element is a factor in determining the bandwidth of the antenna. In general, the rule of thumb states that large cross-sectional area makes the antenna more broadbanded. In some cases, this rule suggests the use of aluminum tubing instead of copper wire for the antenna radiator. On the higher-frequency bands that is a viable solution. Aluminum tubing can be purchased for relatively small amounts of money, and is both lightweight and easily worked with ordinary tools. But, as the frequency decreases, the weight becomes greater because the tubing is both longer and (for structural strength) must be of greater diameter. On 80 m, aluminum tubing is impractical, and at 40 m it is nearly so. Yet, 80 m is a significant problem, especially for older transmitters, because the band is 500 kHz wide, and the transmitters often lack the tuning range for the entire band.

Some other solution is needed. Here are three basic solutions to the problem of wide-bandwidth dipole antennas: *folded dipole*, *bowtie dipole*, and *cage dipole*.

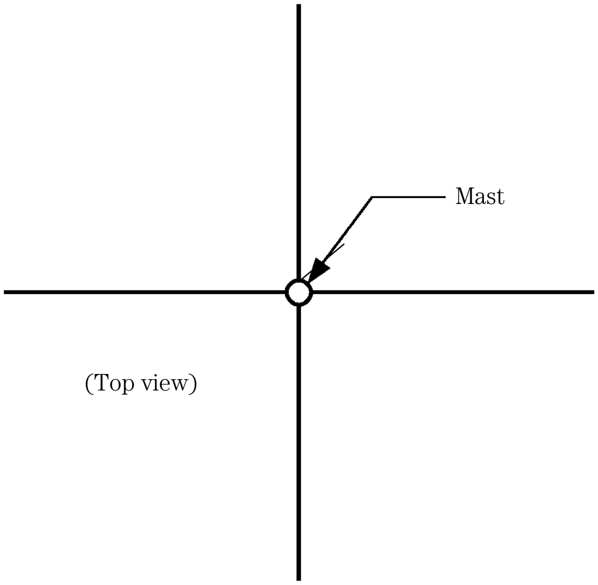

6-9 Directional antenna made of four slopers (top view).

Figure 6-10A shows the *folded dipole* antenna. This antenna basically consists of two half-wavelength conductors shorted together at the ends and fed in the middle of one of them. The folded dipole is most often built from 300-Ω television antenna twin-lead transmission line. Because the feedpoint impedance is nearly 300 Ω, the same type of twin lead can also be used for the transmission line. The folded dipole will exhibit excellent wide-bandwidth properties, especially on the lower bands.

A disadvantage of this form of antenna is that the transmitter has to match the 300-Ω balanced transmission line. Unfortunately, most modern radio transmitters are designed to feed coaxial-cable transmission line. Although an antenna tuner can be placed at the transmitter end of the feedline, it is also possible to use a 4:1 balun transformer at the feedpoint (Fig. 6-10B). This arrangement makes the folded dipole a reasonable match to 52- or 75-Ω coaxial-cable transmission line.

Another method for broadbanding the dipole is to use two identical dipoles fed from the same transmission line, and arranged to form a “bowtie” as shown in Fig. 6-11. The use of two identical dipole elements on each side of the transmission line has the effect of increasing the conductor cross sectional area so that the antenna has a slightly improved length/diameter ratio.

The *bowtie dipole* was popular in the 1930s and 1940s, and became the basis for the earliest television receiver antennas (TV signals are 3 to 5 MHz wide, so they require a broadbanded antenna). It was also popular during the 1950s as the so-called Wonder Bar antenna for 10 m. It still finds use, but it has faded somewhat in popularity.

The *cage diople* (Fig. 6-12) is similar in concept, if not construction, to the bowtie. Again, the idea is to connect several parallel dipoles together from the same transmission line in an effort to increase the apparent cross-sectional area. In the case of the cage dipole, however, spreader disk insulators are constructed to keep the wires separated. The insulators can be built from plexiglass, lucite, or ceramic.

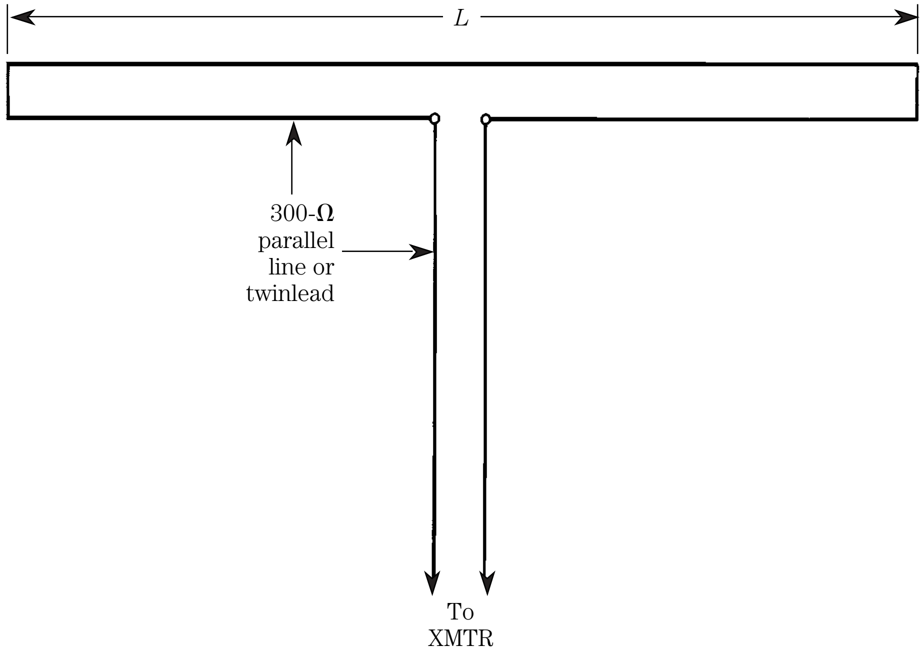
L
300-Ω
parallel
line or
twinlead
To
XMTR

6-10A Folded dipole fed with 300-Ω line.

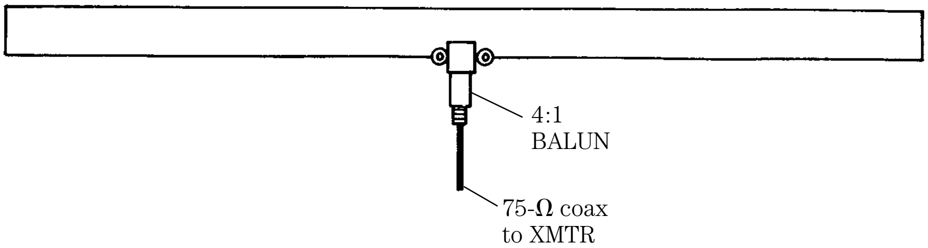
4:1
BALUN
75-Ω coax
to XMTR

6-10B Folded dipole fed with coaxial cable.

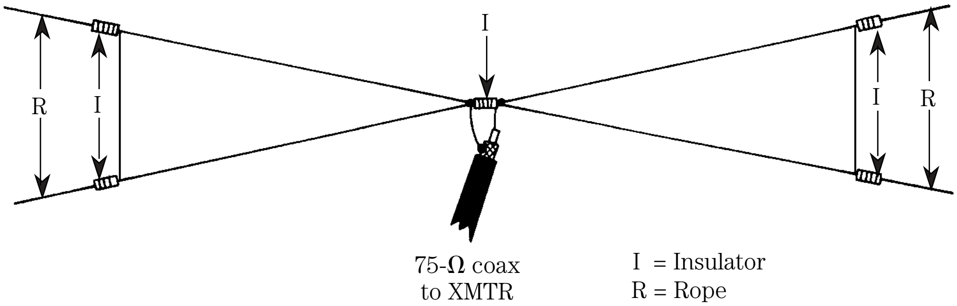
I
R
I
I
R
75-Ω coax
to XMTR
I = Insulator
R = Rope

6-11 Bowtie dipole.

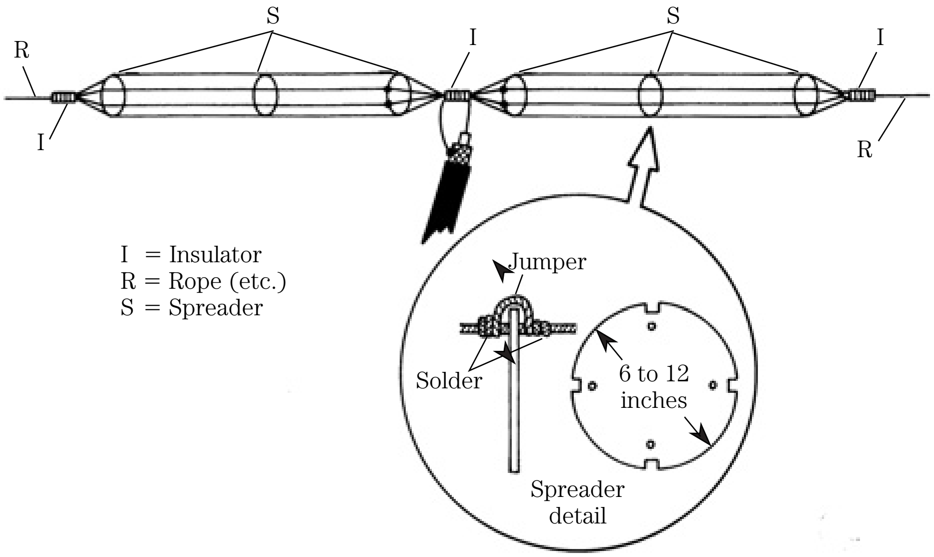

6-12 Cage dipole.

They can also be constructed of materials such as wood, if the wood is properly treated with varnish, polyurethene, or some other material that prevents it from becoming waterlogged. The spreader disks are held in place with wire jumpers (see inset to Fig. 6-12) that are soldered to the main element wires.

A tactic used by some builders of both bowtie and cage dipoles is to make the elements *slightly* different lengths. This “stagger tuning” method forces one dipole to favor the upper end of the band, and the other to favor the lower end of the band. The overall result is a slightly flatter frequency response characteristic across the entire band. On the cage dipole, with four half-wavelength elements, it should be possible to overlap even narrower sections of the band in order to create an even flatter characteristic.

Shortened coil-loaded dipoles

The half-wavelength dipole is too long for some applications where real estate is at a premium. The solution for many operators is to use a coil-loaded shortened dipole such as shown in Fig. 6-13. A shortened dipole (i.e., one which is less than a half-wavelength) is capacitive reactance. There is no reason why the loading coil cannot be any point along the radiator, but in Figs. 6-13A and 6-13B they are placed at 0 percent and 50 percent of the element length, respectively. The reason for this procedure is that it makes the calculation of coil inductances easier, and it also represents the most common practice.

Figure 6-13C shows a table of inductive reactances as a function of the percentage of a half-wavelength, represented by the shortened radiator. It is likely that the percentage figure will be imposed on you by the situation, but the general rule is to pick the largest figure consistent with the available space. For example, sup-

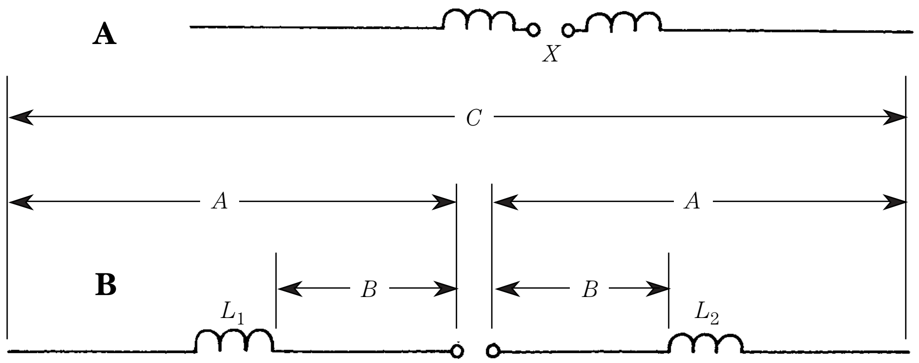

C

$L_1 = L_2 = L$

$C = 2A$

$A = \frac{1}{2}C$

$$C = \frac{468M}{F_{\mathrm{MHz}}}$$

$O \le M \le 1$

Percent of half-wavelength	Coils at feedpoint (Ω)	Coils at middle of radiators (Ω)
20	1800	2800
30	950	1800
40	700	1800
50	500	1300
60	360	950
70	260	700
80	160	500
90	75	160
95	38	80
98	15	30

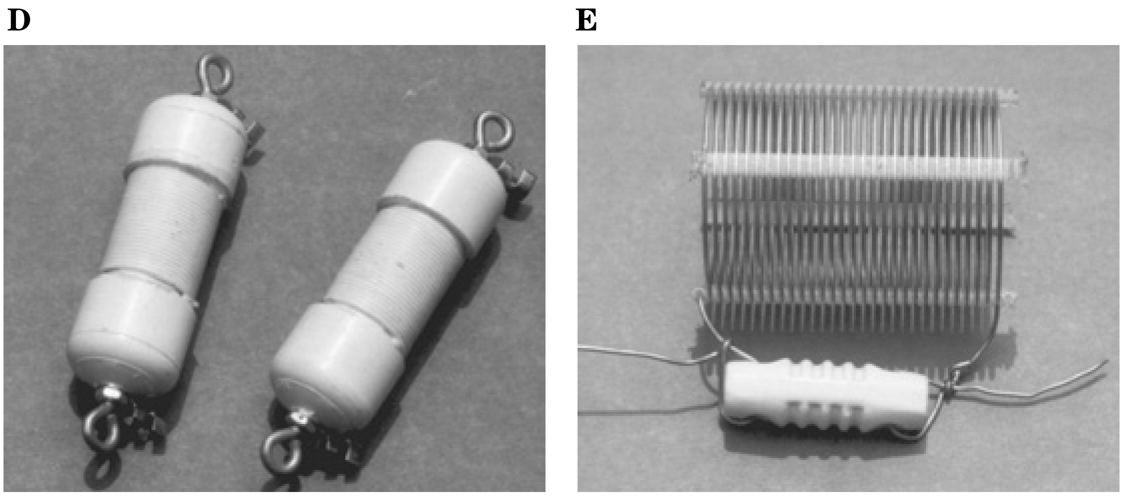

6-13 Shortened or loaded dipole. (A) inductors at feedpoint; (B) inductors midway along elements; (C) chart of reactances for coils; (D) commercially available coils; (E) homemade coil based on commercial coil stock.

pose you have about 40 ft available for a 40-m antenna that normally needs about 65 ft for a half-wavelength. Because 39 ft is 60 percent of 65 ft, you could use this value as the design point for this antenna. Looking on the chart, a 60 percent antenna with the loading coils at the midpoint of each radiator element wants to see an inductive reactance of 700 Ω. You can rearrange the standard inductive reactance equation ($X_L = 6.28\,FL$) to the form

$$L_{\mu H} = \frac{X_L \times 10^6}{6.28F} \quad \textbf{[6.17]}$$

where

$L_{\mu H}$ is the required inductance, in microhenrys
F is the frequency, in hertz (Hz)
X_L is the inductive reactance calculated from the table in Fig. 6-13C.

Example 6-2 Calculate the inductance required for a 60 percent antenna operating on 7.25 MHz. The table requires a reactance of 700 Ω for a loaded dipole with the coils in the center of each element (Fig. 6-15B).

Solution:

$$L_{\mu H} = \frac{X_L \times 10^6}{6.28F} \quad \textbf{[6.18]}$$

$$= \frac{(700)\ (10^6)}{(6.28)\ (7{,}250{,}000)} \quad \textbf{[6.19]}$$

$$= \frac{7 \times 10^8}{4.6 \times 10^7} \quad \textbf{[6.20]}$$

$$= 15.4\ \mu H \quad \textbf{[6.21]}$$

The inductance calculated above is approximate, and it might have to be altered by cut-and-try methods.

The loaded dipole antenna is a very sharply tuned antenna. Because of this fact, you must either confine operation to one segment of the band, or provide an antenna tuner to compensate for the sharpness of the bandwidth characteristic. However, efficiency drops, markedly, far from resonance even with a transmission line tuner. The function of the tuner is to overcome the bad effects on the transmitter, but it does not alter the basic problem. Only a variable inductor in the antenna will do that trick (at least one commercial loaded dipole once used a motor-driven inductor at the center feedpoint).

Figures 6-13D and E show two methods for making a coil-loaded dipole antenna. Figure 6-13D shows a pair of commercially available loading coils especially designed for this purpose. The ones shown here are for 40 m, but other models are also available. The inductor shown in Fig. 6-13E is a section of commercial coil stock connected to a standard end or center insulator. No structural stress is assumed by the coil—all forces are applied to the insulator, which is designed to take it.

Inductance values for other length antennas can be approximated from the graph in Fig. 6-14. This graph contains three curves for coil-loaded, shortened dipoles that are 10, 50, and 90 percent of the normal half-wavelength size. Find the proposed location of the coil, as a percentage of the wire element length, along the horizontal axis. Where the vertical line from that point intersects with one of the three curves, that intersection yields the inductive reactance required (see along vertical axis). Inductances for other overall lengths can be "rough-guessed" by interpolating between the three available curves, and then validated by cut and try.

Tunable dipoles

Dipoles are resonant antennas, so they naturally tend to prefer one frequency over the others. The VSWR will be quite low at the resonant point (assuming no feedline mismatch problems), and will rise at frequencies above and below resonance. If the antenna is a *high-Q* model, then the effect is quite profound, and it might render the antenna nearly useless at frequencies on the other end of the same band (especially where modern transmitters equipped with VSWR shutdown circuitry are used). Figure 6-15 shows a method for overcoming this problem. Here is how it works.

An antenna that is too long for the desired resonant frequency will act inductively (i.e., it will show a feedpoint impedance of the form $Z = R + jX_L$). To counteract the inductive reactance component $+jX_L$ it is necessary to add a bit of capacitive reactance $-jX_c$. This approach is taken in tuning certain antenna forms and it can be used with dipoles to make the antenna tunable over a range of about 15 percent of the frequency.

The tunable dipole of Fig. 5-15 is longer than other dipoles and it has an overall length of

$$L' = \frac{505}{F_{\text{MHz}}} \tag{6.22}$$

Note that the velocity constant is 505, rather than the 468 that is used with ordinary dipoles.

The required capacitive reactance, which is used to electrically shorten the antenna, is provided by C_1. In one effective design that I've tested, the capacitor was a 500-pF transmitting variable type. When the Heath Company was making kits, they offered an antenna tuner based on this idea. It was a motor-driven 500-pF variable capacitor inside of a weather-resistant metal-shielded case. The Heath tuner can be used on inverted vees and dipoles (if a center support is provided). A low-voltage (which is necessary for safety reasons) dc motor drives the shaft of the capacitor to tune the antenna to resonance.

Stacked dipoles

Figure 6-16 shows a *double dipole* (i.e., two half-wavelength dipole antennas spaced a half-wavelength apart). The transmission lines are connected in parallel at the receiver. This antenna provides about 3-dB gain over a single dipole, and it adds a bit of fade protection because two side-by-side antennas provide a bit of space diversity (see the end of Chap. 2 for information on diversity reception).

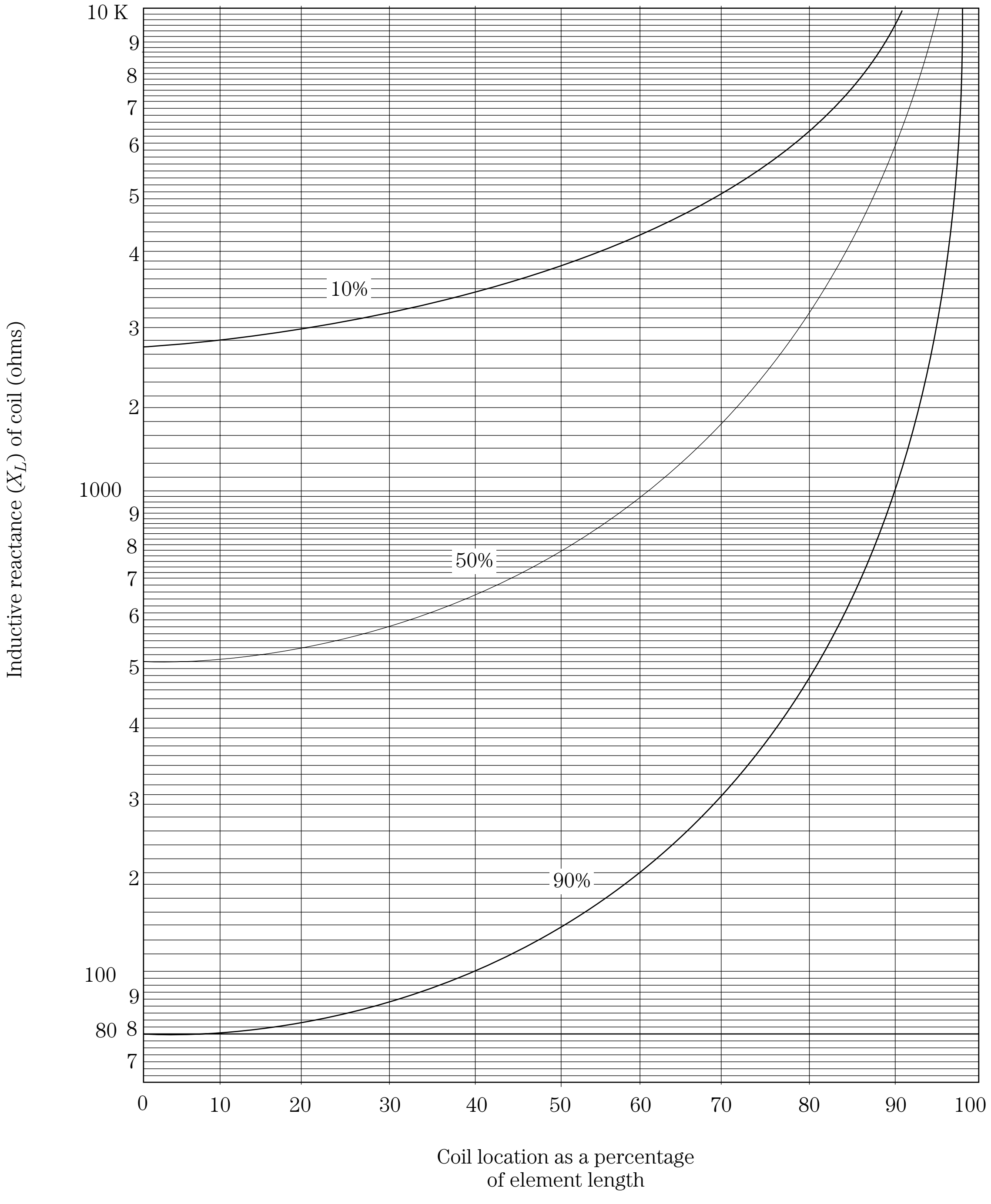

6-14 Inductive reactance versus coil location for shortened inductance-loaded dipole.

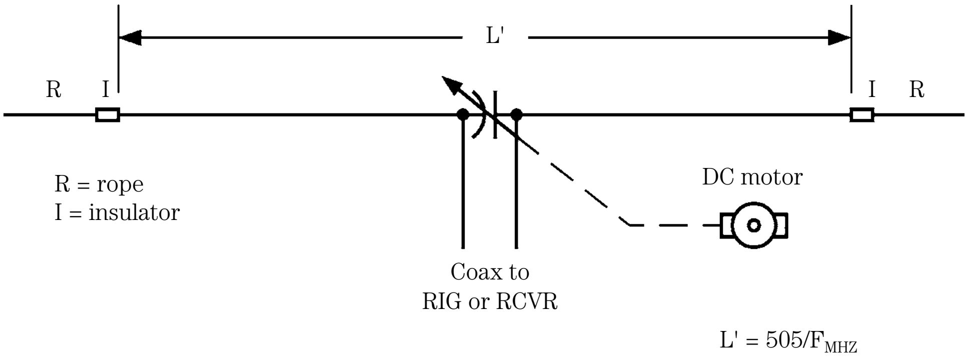
L'
R
I
I
R
R = rope
I = insulator
DC motor
Coax to
RIG or RCVR
$L' = 505/F_{MHZ}$

6-15 Electrically reducing the length of a dipole antenna by placing a capacitor at the feedpoint.

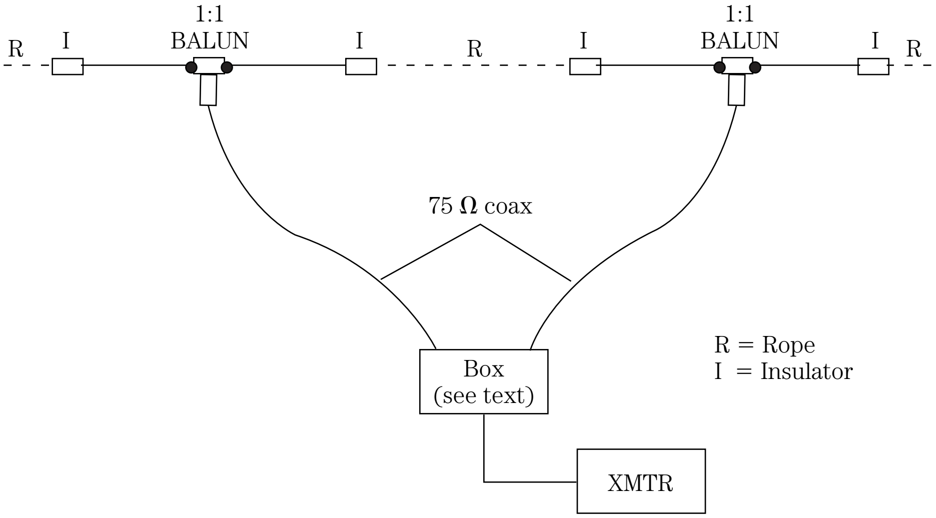
R
I
1:1
BALUN
I
R
I
1:1
BALUN
I
R
75 Ω coax
Box
(see text)
R = Rope
I = Insulator
XMTR

6-16 Phased dipoles provide a 3-dB gain over single dipole.

The two dipoles are supported by a common structure consisting of ropes (R) and end insulators (I) to support masts on the ends. In the center, a half-wavelength space is taken up by a rope so that the structure is maintained. The space is determined by

$$\text{Space}_{\text{feet}} = \frac{492}{F_{\text{MHz}}} \qquad \textbf{[6.23]}$$

and, for the dipoles,

$$\text{Length}_{\text{feet}} = \frac{468}{F_{\text{MHz}}} \qquad \textbf{[6.24]}$$

The stacked dipole gets a bit lengthy on low-frequency bands, but is easily achievable by most people on the upper HF bands.

The feedlines for the two dipoles can be connected directly in parallel and fed from the transmitter, provided that the antenna tuning unit will support one-half the normal expected impedance. Alternatively, a box can be provided that includes a matching transformer for 1:2 ratio. These can be built like a balun transformer on a toroidal core. A trifilar winding is used. Alternatively, a phasing box can be built that will allow altering the directionality of the antenna by 90°. This is done by using the switching circuit to reverse the sense of L_3.

Off-center-fed full-wave doublet (OCFD) antennas

An antenna that superficially resembles the Windom is the off-center-fed doublet (OCFD) antenna of Fig. 6-17. It is a single-band antenna, although at harmonics it will begin to act as a resonant, standing wave, longwire antenna. The overall length is one wavelength long:

$$L_{\text{feet}} = \frac{936}{F_{\text{MHz}}} \tag{6.25}$$

This antenna works best at heights of at least λ/2 above ground, so practical considerations limit it to frequencies above about 10 MHz (i.e., 30-m band). The feedpoint of the antenna is placed at a distance of λ/4 from one end, and is a good match for 75-Ω coaxial cable. A 1:1 balun transformer at the feedpoint is highly recommended. The pattern of a 1λ antenna is a four-lobe "cloverleaf," with the major lobes being about 53° from the wire. The gain is about 1 dB.

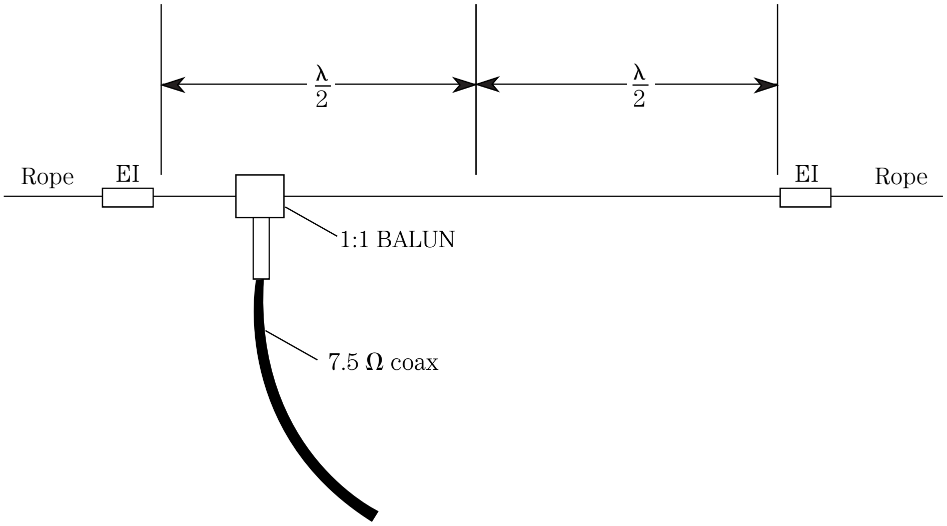

6-17 Off-center-fed 1-wire antenna.

Off-center-fed nonresonant sloper (OCFS)

Perhaps more viable for many people is the *nonresonant off-center-fed sloper* (OCFS) antenna of Fig. 6-18. This antenna consists of a wire radiator that must be longer than 3λ/2 at the lowest frequency of operation. The feedpoint is elevated at least λ/4 above ground at the lowest operating frequency. The antenna is fed with 75-Ω coaxial cable. The shield of the coax is connected to a λ/4 resonant radial (counterpoise ground). There should be at least one radial (more is better) per band of operation.

The far end of the radiator element is sloped to ground, where it is terminated in a 270-Ω noninductive resistor. The resistor should be able to dissipate up to one-third of the power level applied by the transmitter.

Double extended Zepp antenna

The *double extended Zepp* antenna (Fig. 6-19) provides a gain of about 2 dB over a dipole at right angles to the antenna wire plane. It consists of two sections of wire, each one of a length

$$L_{1\text{feet}} = \frac{600}{F_{\text{MHz}}} \qquad \textbf{[6.26]}$$

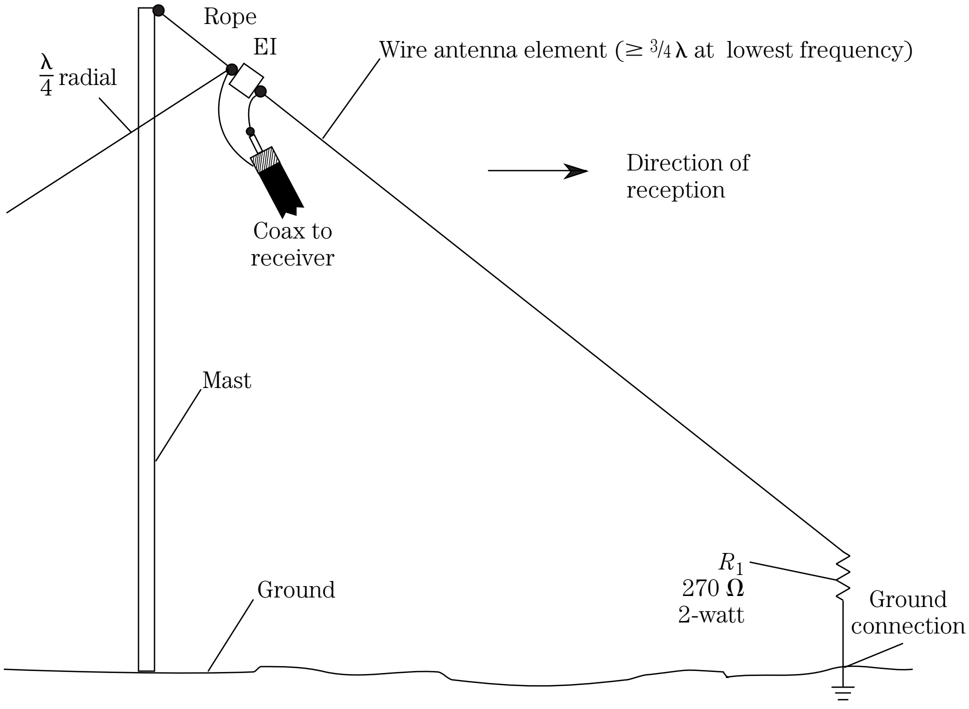
Rope
EI
$\frac{\lambda}{4}$ radial
Wire antenna element (≥ 3/4 λ at lowest frequency)
Direction of reception
Coax to receiver
Mast
R_1
270 Ω
2-watt
Ground
Ground connection

6-18 Terminated sloper antenna.

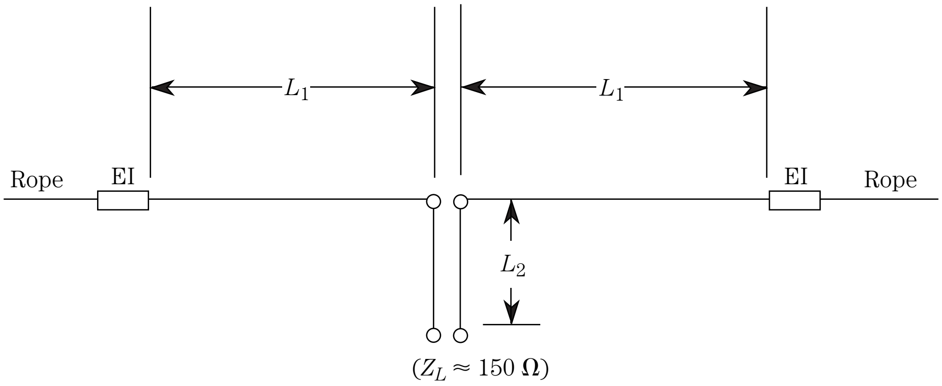

6-19 Double extended Zepp antenna.

Typical lengths are 20.7 ft on the 10-m band, 28 ft on the 15-m band, 42 ft on the 20-m band, and 84 ft on the 40-m band.

The double extended Zepp antenna can be fed directly with 450-Ω twin lead, especially if a balanced antenna tuner is available at the receiver. Alternatively, it can be fed from a quarter-wavelength matching section (made of 450-Ω twin lead, or equivalent open air parallel line), as shown, and a balun if coax is preferred. The length of the matching section should be

$$L_{2\text{feet}} = \frac{103}{F_{\text{MHz}}} \tag{6.27}$$

The double extended Zepp will work on several different bands. For example, a 20-m-band double extended Zepp will work as a Zepp on the design band, a dipole on frequencies below the design band, and as a four-lobed cloverleaf antenna on frequencies above the design band.

Collinear "Franklin" array antenna

Perhaps the cheapest approach to very serious antenna gain is the collinear Franklin array shown in Fig. 6-20. This antenna pushes the dipole and double extended Zepp concepts even farther. It consists of a half-wavelength dipole that is center-fed with a 4:1 balun and 75-Ω coaxial cable. At each end of the dipole, there is a quarter-wavelength *phase reversal stub* that end-feeds another half-wavelength element. Each element is a half-wavelength (λ/2) long, and its length can be calculated from

$$\frac{\lambda}{2} = \frac{492}{F_{\text{MHz}}} \tag{6.28}$$

The phase reversal stubs are a quarter-wavelength long, or one half the length calculated by Eq. 6.28.

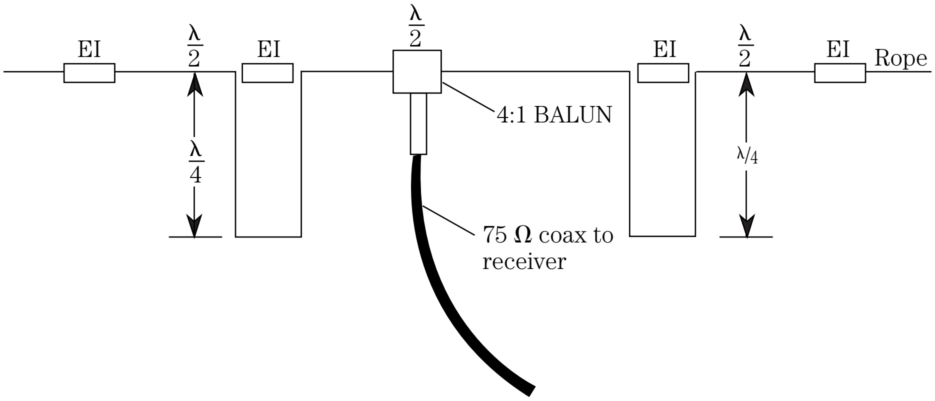

6-20 Wire collinear antenna.

The version of the "Collinear" shown in Fig. 6-20 has a gain of about 3 dB. There is no theoretical reason why you can't extend the design indefinitely, but there is a practical limit set by how much wire can be held by your supports, and how much real estate you own. A 4.5-dB version can be built by adding another half-wavelength section at each end, with an intervening quarter-wavelength phase reversal stub in between each new section, and the preceding section. Once you get longer than five half-wavelengths, which provides the 4.5-dB gain, the physical size becomes a bit of a bother for most folks.

The TCFTFD dipole

The *tilted, center-fed, terminated, folded dipole* (TCFTFD, also called the T^2FD or TTFD) is an answer to both the noise pickup and length problems that sometimes affect other antennas. For example, a random-length wire, even with antenna tuner, will pick up considerable amounts of noise. A dipole for 40 m is 66 ft long.

This antenna was first described publicly in 1949 by Navy Captain C. L. Countryman, although the U.S. Navy tested it for a long period in California during World War II. The TCFTFD can offer claimed gains of 4 to 6 dB over a dipole, depending on the frequency and design, although 1 to 3 dB is probably closer to the mark in practice, and less than 1 dB will be obtained at some frequencies within its range (especially where the resistor has to absorb a substantial portion of the RF power). The main attraction of the TCFTFD is not its gain, but rather its *broad bandedness*.

In addition, the TCFTFD can also be used at higher frequencies than its design frequency. Some sources claim that the TCFTFD can be used over a 5 or 6:1 frequency range, although my own observations are that 4:1 is more likely. Nonetheless, a 40-m antenna will work over a range of 7000 to 25,000 kHz, with at least some decent performance up into the 11-m Citizen's Band (27,000 kHz).

The basic TCFTFD (Fig. 6-21) resembles a folded dipole in that it has two parallel conductors of length *L*, spaced a distance *W* apart, and shorted together at the

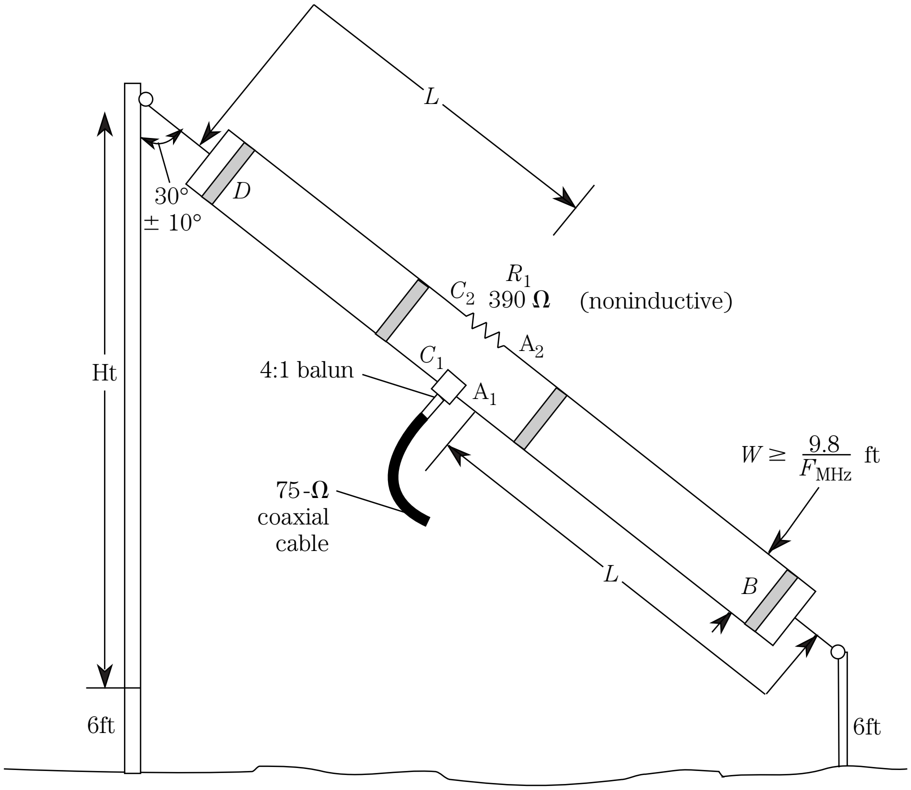

6-21 TTFD antenna.

ends. The feedpoint is the middle of one conductor, where a 4:1 balun coil and 75-Ω coaxial-cable transmission line to the transceiver are used. A noninductive, 390-Ω resistor is placed in the center of the other conductor. This resistor can be a carbon-composition (or metal-film) resistor, but it must not be a wirewound resistor or any other form that has appreciable inductance. The resistor must be able to dissipate about one-third of the applied RF power. The TCFTFD can be built from ordinary no.14 stranded antenna wire.

For a TCFTFD antenna covering 40 through 11 m, the spread between the conductors should be 19½ in, while the length *L* is 27 ft. Note that length *L* includes one-half of the 19-in spread because it is measured from the center of the antenna element to the center of the end supports.

The TCFTFD is a sloping antenna, with the lower support being about 6 ft off the ground. The height of the upper support depends on the overall length of the antenna. For a 40-m design, the height is on the order of 50 ft.

The parallel wires are kept apart by spreaders. At least one commercial TCFTFD antenna uses PVC spreaders, while others use ceramic. You can use wooden dowels of between 1-in and ⅝-in diameter; of course, a coating of varnish (or urethane spray) is recommended for weather protection. Drill two holes, of a size sufficient to pass the wire, that are the dimension *W* apart (19 in for 40 m). Once the spreaders are in place, take about a foot of spare antenna wire and make jumpers to hold the dowels in place. The jumper is wrapped around the antenna wire on either side of the dowel, and then soldered.

The two end supports can be made of 1 × 2 in wood treated with varnish or urethane spray. The wire is passed through screw eyes fastened to the supports. A support rope is passed through two holes on either end of the 1 × 2 and then tied off at an end insulator.

The TCFTFD antenna is noticeably quieter than the random-length wire antenna, and somewhat quieter than the half-wavelength dipole. When the tilt angle is around 30°, the pattern is close to omnidirectional. Although a little harder to build than dipoles, it offers some advantages that ought not to be overlooked. These dimensions will suffice when the "bottom end" frequency is the 40-m band, and it will work well on higher bands.

Vee-sloper antenna

The *vee-sloper antenna* is shown in Fig. 6-22. It is related to the vee beam (covered in Chap. 9), but it is built like a sloper (i.e., with the feed end of the antenna high above ground). The supporting mast height should be about half (to three-fourths) of the length of either antenna leg. The legs are sloped downward to terminating

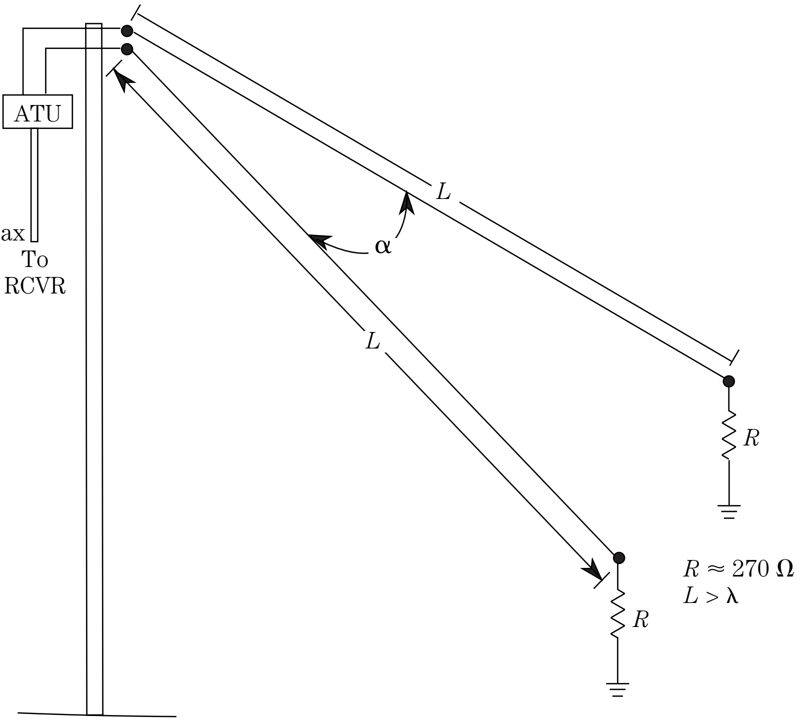

6-22 Vee-sloper antenna.

resistors at ground level. Each wire should be longer than 1λ at the lowest operating frequency. The terminating resistors should be on the order of 270 Ω (about one-half of the characteristic impedance of the antenna), with a power rating capable of dissipating one-third of the transmitter power. Like other terminating resistors, these should be noninductive (carbon composition or metal film).

The advantage of this form of antenna over the vee beam is that it is vertically polarized, and the resistors are close to the earth, so they are easily grounded.

Rhombic inverted-vee antenna

A variation on the theme is the vertically polarized rhombic of Fig. 6-23. Although sometimes called an *inverted vee*—not to be confused with the dipole variant of the same name—this antenna is half a rhombic, with the missing half being "mirrored" in the ground (similar to a vertical). The angle at the top of the mast (Φ) is typically ≥ 90°, and 120 to 145° is more common. Each leg (A) should be ≥λ, with the longer lengths being somewhat higher in gain, but harder to install for low frequencies. A requirement for this type of antenna is a very good ground connection. This is often accomplished by routing an underground wire between the terminating resistor ground and the feedpoint ground.

Multiband fan dipole

The basic half-wavelength dipole antenna is a very good performer, especially when cost is a factor. The dipole yields relatively good performance for practically no in-

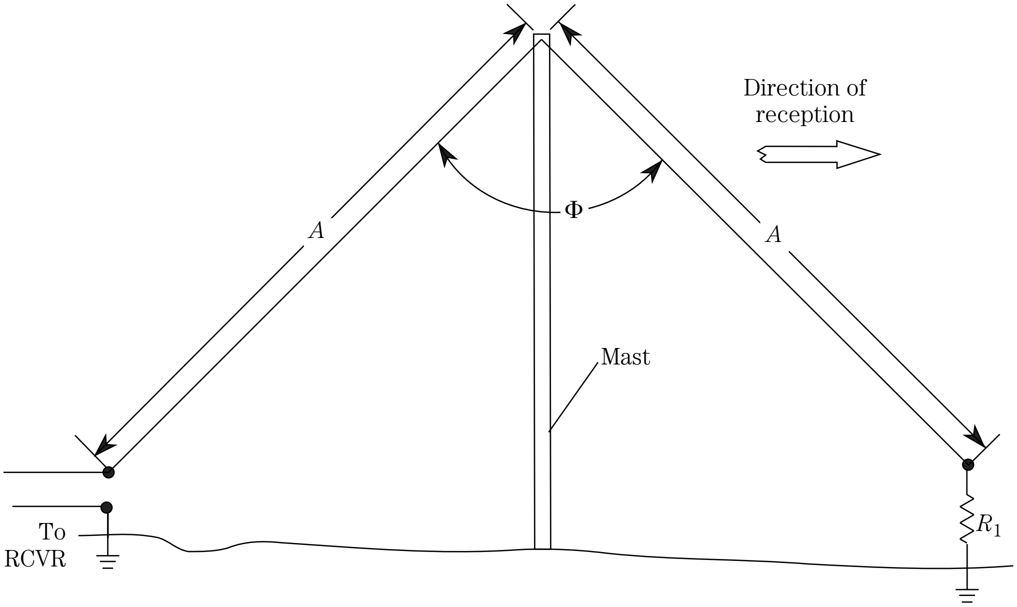

6-23 Inverted-vee antenna.

vestment. A standard half-wavelength dipole offers a bidirectional figure-8 pattern on its basic band (i.e., where the length is a half-wavelength), and a four-lobe cloverleaf pattern at frequencies for which the physical length is 3λ/2. Thus, a 40-m half-wavelength dipole produces a bidirectional pattern on 40 m, and a four-lobe cloverleaf pattern on 15 m.

The dipole is not easily multibanded without resorting to traps (covered in Chap. 8). One can, however, tie several dipoles to the same center insulator or balun transformer. Figure 6-24 shows three dipoles cut for different bands, operating from a common feedline and balun transformer: A1–A2, B1–B2, and C1–C2. Each of these antennas is a half-wavelength (i.e., $L_{\text{feet}} = 468/F_{\text{MHz}}$).

There are two points to keep in mind when building this antenna. First, try to keep the ends spread a bit apart, and second, make sure that none of the antennas is cut as a half-wavelength for a band for which another is 3λ/2. For example, if you make A1–A2 cut for 40 m, then don't cut any of the other three for 15 m. If you do, the feedpoint impedance and the radiation pattern will be affected.

The counterpoise longwire

The *longwire antenna* is an end-fed wire more than 2λ long. It provides considerable gain over a dipole, especially when a very long length can be accommodated. Although 75- to 80-m, or even 40-m longwires are a bit difficult to erect at most locations, they are well within reason at the upper end of the HF spectrum. Low-VHF

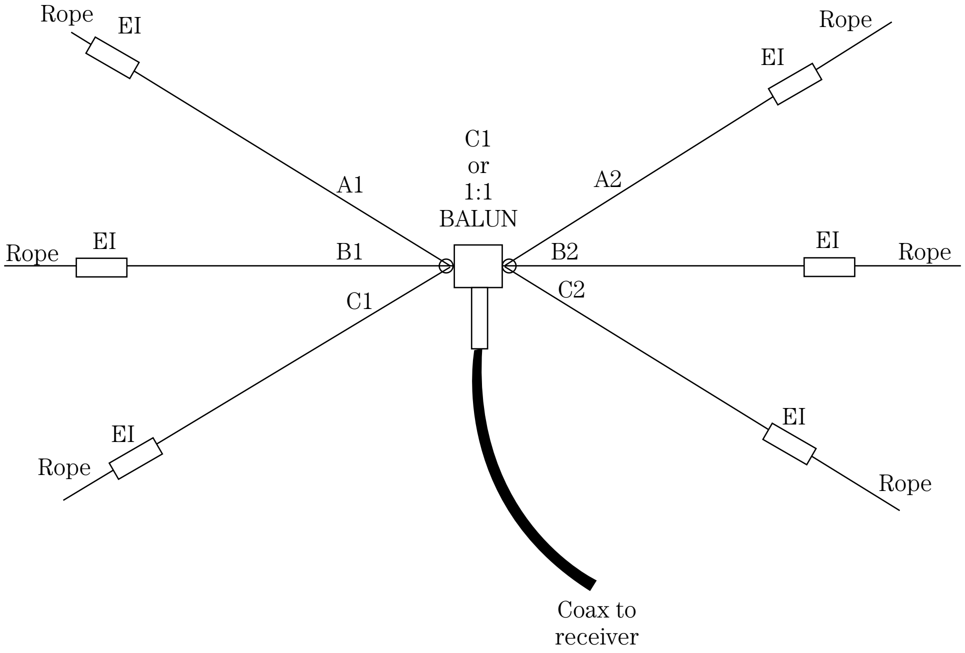

6-24 Multiband dipole antenna.

band operation is also practical. Indeed, I know one fellow who lived in far southwest Virginia as a teenager, and he was able to get his family television reception for very low cost by using a TV longwire (channel 6) on top of his mountain.

There are some problems with longwires that are not often mentioned. Two problems seem to insinuate themselves into the process. First, the Zepp feed is a bit cumbersome (not everyone is enamored of parallel transmission line). Second, how do you go about actually grounding that termination resistor? If it is above ground, then the wire to ground is long, and definitely not at ground potential for RF. If you want to avoid both the straight Zepp feed system employed by most such antennas, as well as the resistor-grounding problem, then you might want to consider the *counterpoise longwire antennas* shown in Fig. 6-25.

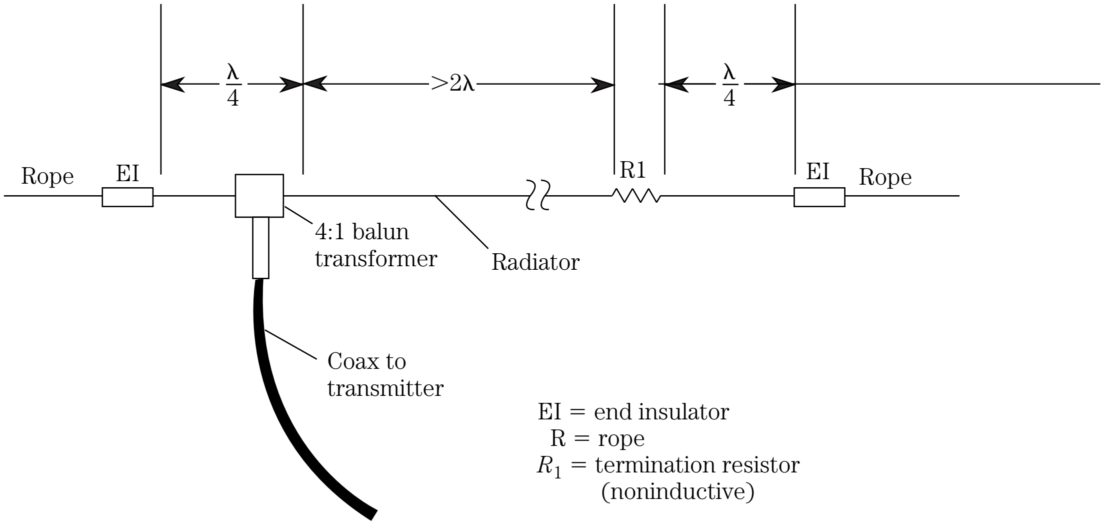
$\frac{\lambda}{4}$
>2λ
$\frac{\lambda}{4}$
Rope
EI
R1
EI
Rope
4:1 balun transformer
Radiator
Coax to transmitter
EI = end insulator
R = rope
R_1 = termination resistor (noninductive)

6-25A Longwire with λ/4 radials used as counterpoise ground at feedpoint and terminating resistor.

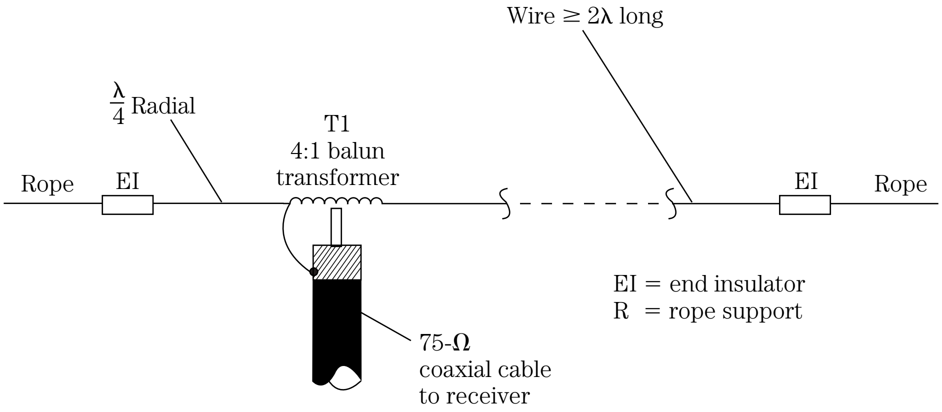
Wire ≥ 2λ long
$\frac{\lambda}{4}$ Radial
T1
4:1 balun
transformer
Rope
EI
EI
Rope
EI = end insulator
R = rope support
75-Ω
coaxial cable
to receiver

6-25B Balun transformer feed for longwire antenna.

A counterpoise ground is a structure that acts like a ground, but is actually electrically floating above real ground (and it is not connected to ground). A groundplane of radials is sometimes used as a counterpoise ground for vertical antennas that are mounted above actual earth ground. In fact, these antennas are often called *ground plane verticals*. In those antennas, the array of four (or more) radials from the shield of the coaxial cable are used as an artificial, or counterpoise, ground system.

In the counterpoise longwire of Fig. 6-25A, there are two counterpoise grounds (although, for one reason or another, you might elect to use either, but not both). One counterpoise is at the feedpoint, where it connects to the "cold" side of the transmission line. The parallel line is then routed to an antenna tuning unit (ATU), and from there to the transmitter. The other counterpoise is from the cold end of the termination resistor to the support insulator. This second counterpoise makes it possible to eliminate the earth ground connection, and all the problems that it might entail, especially in the higher end of the HF spectrum, where the wire to ground is of substantial length compared with 1λ of the operating frequency.

A slightly different scheme used to adapt the antenna to coaxial cable is shown in Fig. 6-25B. In this case, the longwire is a resonant type (nonterminated). Normally, one would expect to find this antenna fed with 450-Ω parallel transmission line. But with a λ/4 radial acting as a counterpoise, a 4:1 balun transformer can be used to effect a reasonable match to 75-Ω coaxial cable. The radial is connected to the side of the balun that is also connected to the coaxial cable shield, and the other side of the balun is connected to the radiator element.

7
CHAPTER

Vertically polarized HF antennas

IN PREVIOUS CHAPTERS, YOU HAVE FOUND THAT THE POLARITY OF AN ANTENNA IS THE direction of the electrical (E) field. Because the transmitted signal is an orthogonal electromagnetic wave, the magnetic field radiated from the antenna is at right angles to the electric field. The direction of the electric field, which sets the polarity of the antenna, is a function of the geometry of the radiator element. If the element is vertical, then the antenna polarity is also vertical. The signal propagates out from the radiator in all directions of azimuth, making this antenna an "omnidirectional" radiator.

Figure 7-1A shows the basic geometry of the vertical antenna: an RF generator (transmitter or transmission line from a transmitter) at the base of a radiator of length L. Although most commonly encountered verticals are quarter-wavelength ($L = \lambda/4$), that length is not the only permissible length. In fact, it may not even be the most desirable length. This chapter covers the standard quarter-wavelength vertical antenna (because it is so popular), and other-length verticals (both greater and less than quarter-wavelength).

The quarter-wavelength vertical antenna is basically half of a dipole placed vertically, with the "other half" of the dipole being the ground. Because of this fact, some texts show the vertical with a double-line ghost radiator, or image antenna, in the earth beneath the main antenna element. Figure 7-1B shows the current and voltage distribution for the quarter-wavelength vertical. Like the dipole, the quarter-wavelength vertical is fed at a current loop, so the feedpoint impedance is at a minimum (typically 2 to 37 Ω, depending upon nearby objects and design). As a result, the current is maximum and the voltage is minimum at the feedpoint. As you will see, however, not all vertical antennas are fed directly at the current loop. As a result, some designs require antenna tuning units to make them match the antenna impedance to the transmitter output impedance.

Figure 7-1C and 7-1D show the two basic configurations for the HF vertical antenna. Figure 7-1C shows the ground-mounted vertical antenna. The radiator

Copyright 2001 - the McGraw-Hill Companies

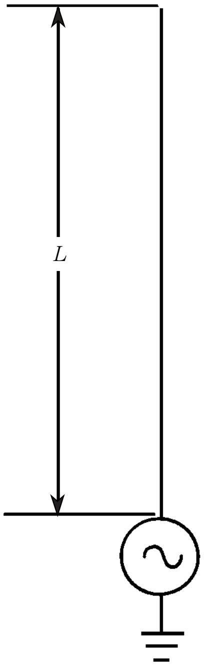

7-1A Basic elements of the vertical antenna.

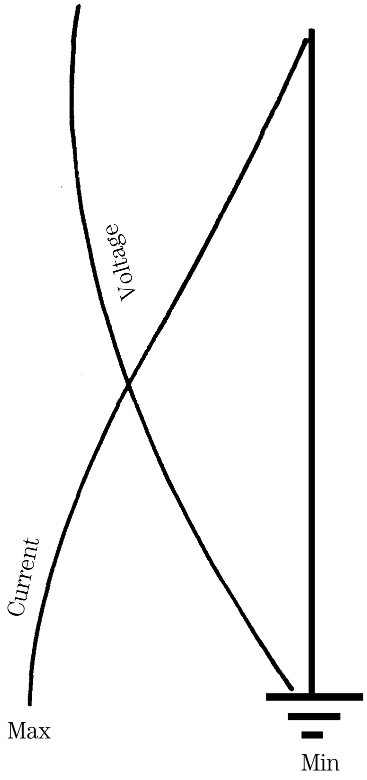
Voltage
Current
Max
Min

7-1B Current and voltage distribution along vertical.

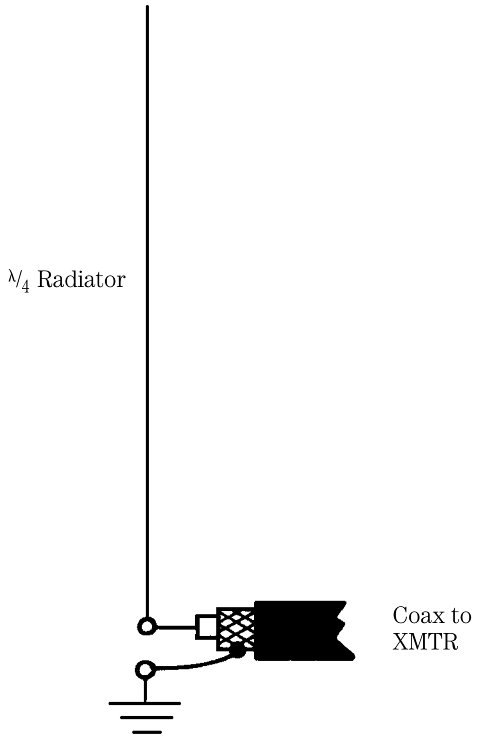
$\lambda/4$ Radiator
Coax to XMTR

7-1C Simple coaxial-fed vertical antenna. There is a slight mismatch, but it usually is within tolerable limits.

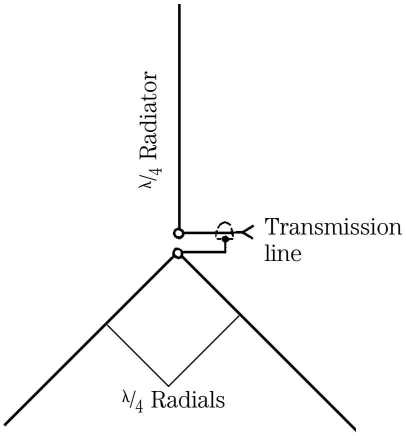
$\lambda/4$ Radiator
Transmission line
$\lambda/4$ Radials

7-1D Mast-mounted vertical uses radials as a counterpoise ground.

element is mounted at ground level, but it is insulated from ground. Because the antenna shown is a quarter-wavelength, it is fed at a current loop with 52-Ω coaxial cable. The inner conductor of the coaxial cable is connected to the radiator element, and the coaxial cable shield is connected to the ground. As you will see shortly, the ground system for the vertical antenna is critical to its performance. Normally, the feedpoint impedance is not 52 Ω, but rather is somewhat lower (37 Ω). As a result, without some matching there will be a slight VSWR, but in most cases, the VSWR is a tolerable tradeoff for simplicity. If the antenna has a feedpoint impedance of 37 Ω, which is the value usually quoted, then the VSWR will be 52 Ω/37 Ω, or 1.41:1.

A vertical mounted above the ground level is shown in Fig. 7-1D. This antenna is equally as popular as the ground mounted. Amateurs and CB operators find it easy to construct this form of antenna because the lightweight vertical can be mounted at reasonable heights (15 to 60 ft) using television antenna slip-up telescoping masts that are reasonably low in cost. A problem with the non-ground-level vertical antenna is that there is no easy way to connect it to ground. The solution to the problem is to create a counterpoise (artificial) ground with a system of quarter-wavelength radials.

In general, at least two radials are required for each band, and even that number is marginal. The standard wisdom holds that the greater the number of radials, the better the performance. Although that statement is true, there are both theoretical and practical limits to the number of radials. The theoretical limit is derived from the fact that more than 120 radials returns practically no increase in operational effectiveness, and at more than 16 radials, the returned added effectiveness per new radial is less than is the case for fewer radials. That is, going from 16 to 32 radials (doubling the number) creates less of an increase in received field strength at a distant point than going from 8 to 16 radials (both represent doubling the density of the radial system). The practical limit is 16 radials.

The radials of the off-ground-level vertical antenna can be at any angle. In Fig. 7-1D, they are "drooping radials" (i.e., the angle is greater than 90° relative to the vertical radiator element). Similarly, Fig. 7-1E shows a vertical antenna that is equipped with radials at exactly 90° (no common antenna has radials less than 90°). Both of these antennas are called *ground plane vertical antennas*.

The angle of the vertical's radials is said to affect the feedpoint impedance and the angle of radiation of the vertical antenna. Although those statements are undoubtedly true in some sense, there are other factors that also affect those parameters, and they are probably more important in most practical installations. Before digging further into the subject of vertical antennas, take a look at the subjects of angle of radiation and gain in vertical antennas.

Angle of radiation

Long-distance propagation in the HF region depends upon the ionospheric phenomena called "skip." (See Chap. 2 for a more extensive explanation.) In this type of propagation, the signal leaves the transmitting antenna at some angle a, called the *angle of radiation*, and enters the ionosphere where it is refracted back to earth at a distance D from the transmitting station. The signal in the zone between the outer edge of the antenna's ground-wave region and the distant skip point is weak or nonexistent.

The distance covered by the signal on each skip is a function of the angle of radiation. Figure 7-2 shows a plot of the angle of radiation of the antenna, and the distance to the first skip zone. The angle referred to along the vertical axis is the angle of radiation away from the antenna relative to the horizon. For example, an angle of

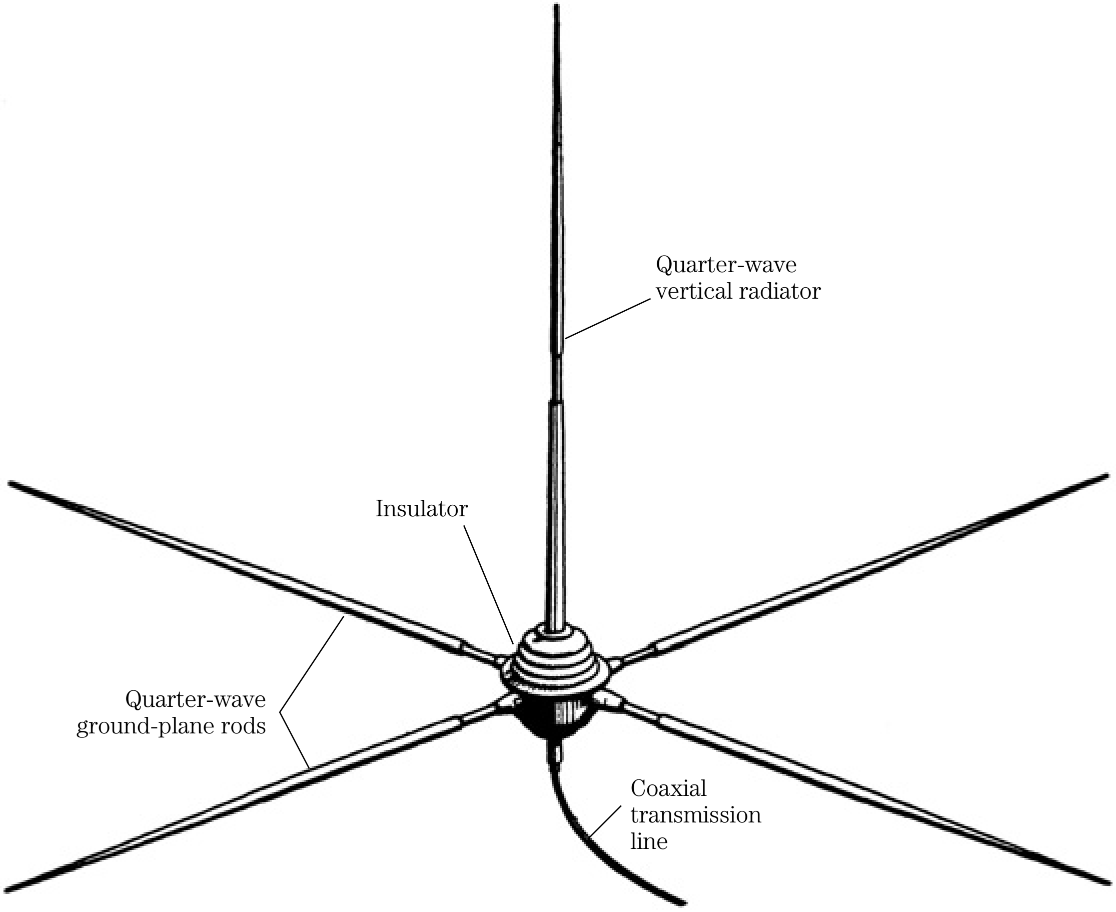
Quarter-wave
vertical radiator
Insulator
Quarter-wave
ground-plane rods
Coaxial
transmission
line

Ground-plane antenna

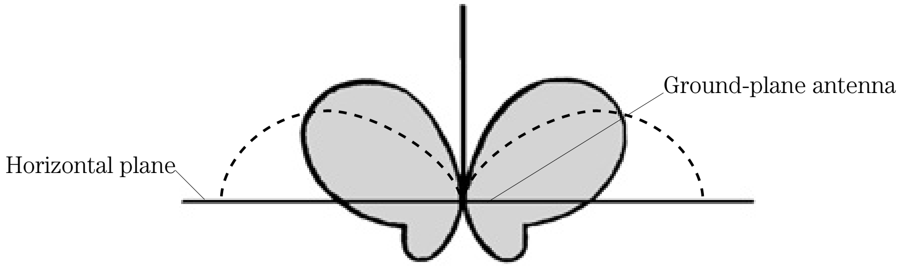
Ground-plane antenna
Horizontal plane

Radiation pattern

7-1E Ground-plane vertical antenna.

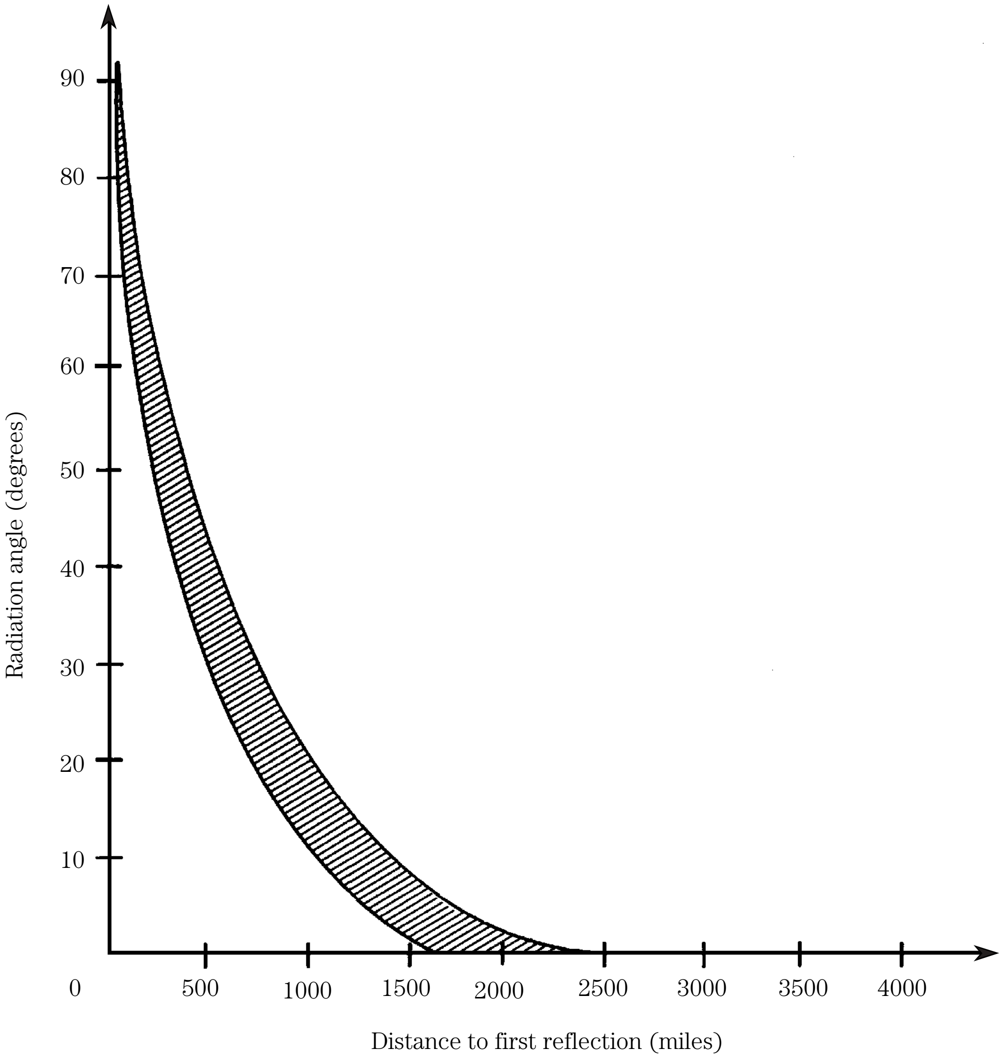

7-2 Effect of radiation angle on skip communications.

10° is elevated 10° above the horizon. Shorter distances are found when the angle of radiation is increased. At an angle of about 30°, for example, the distance per skip is only a few hundred miles.

Although you might expect on first blush to see a single line on the graph, there is actually a zone shown (shaded). This phenomenon exists because the ionosphere is found at different altitudes at different times of the day and different seasons of the year. Generally, however, in the absence of special event phenomena in the ionosphere, you can expect from 1500 to 2500 mi per bounce in the HF bands for low angles of radiation. Note, for example, that for a signal that is only a degree or two above the horizon the skip distance is maximum.

At distances greater than those shown in Fig. 7-2, the signal will make multiple hops. Given a situation where the skip distance is 2500 mi, covering a distance of 7500 mi requires three hops. Unfortunately, there is a signal strength loss on each hop of 3 to 6 dB, so you can expect the distant signal to be attenuated from making multiple hops between the earth's surface and the ionosphere. For maximizing distance, therefore, the angle of radiation needs to be minimized.

So what is the ideal angle of radiation? It is standard—but actually erroneous—wisdom among amateur radio operators (and even commercial operators, it turns out) that the lower the angle of radiation, the better the antenna. That statement is only true if long distance is wanted, so it reflects a strong bias toward the DX community. The correct answer to the question is: "It depends on where you want the signal to go." For example, I live in Virginia. If I want to communicate with stations in the Carolinas or New England, then it would behoove me to select a higher angle of radiation for radio conditions represented in Fig. 7-2 so that the signal will land in those regions. But if I wanted to work stations in Europe or Africa or South America, then a low angle of radiation is required. Because of the difference between performance of high and low angles of radiation, some stations have two antennas for each band: one each for high and low angles of radiation.

Figure 7-3 shows a signal from a hypothetical antenna located at point 0, in order to show what angle is meant by "angle of radiation." The beam from the antenna is elevated above the horizon (represented by the horizontal "tangent to horizon" line). The angle of radiation a is the angle between the tangent line and the center of the beam. This angle is not to be confused with the beamwidth, which is also an angle. In the case of beamwidth, we are talking about the thickness of the main lobe of the signal between points where the field strength is –3 dB down from the maximum signal (which occurs at point P); these points are represented by points X and Y in Fig. 7-3. Thus, angle b is the beamwidth, and angle a is the angle of radiation.

Gain in vertical antennas

Vertical antennas are considered *omnidirectional* because they radiate equally well in all directions. "Gain" in an antenna is not the creation of power, but rather a simple refocusing of energy from all directions to a specific direction; gain therefore infers directivity. According to the convention, then, the vertical antenna cannot have any gain because it radiates in all directions equally . . . gain infers directivity. Right? No, not really. Let's develop the theme more carefully.

Again consider the idea of an isotropic radiator (the word *isotropic* means equal power in all directions). Consider a spherical point source radiator located at point O in Fig. 7-4. Whatever the level of power available from the transmitter, it will be spread equally well over the entire surface of the sphere as it radiates out into space away from point O. If you measure the power distributed over some area A at a distance R from the source, then the power available will be a fraction of the total power:

$$P_{\text{avail}} = \frac{\text{total available power} \times A}{\text{total surface area of sphere}}$$

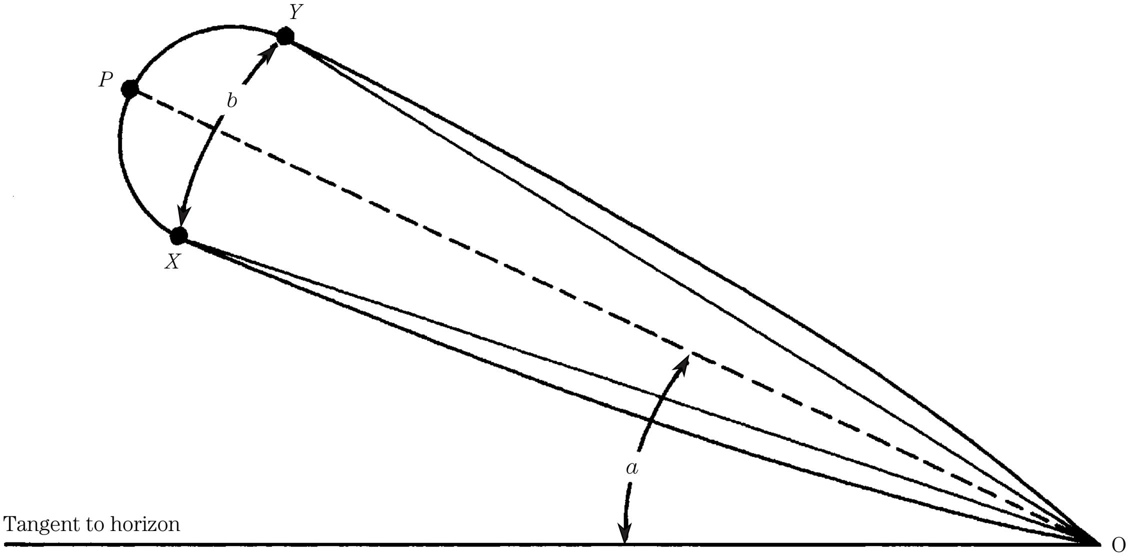

X, *Y*: −3 dB points
P: center beam
Angle *a*: elevation angle
Angle *b*: 3-dB beam width

7-3 Side view of vertical extent radiation pattern.

or, in math symbols:

$$P_a = \frac{P_s}{4\pi r^2} \qquad \textbf{[7.1]}$$

where

P_a is the power available per solid degree
P_s is the total radiated power, in watts
r is the radius of the sphere (i.e., the distance from O to P).

A practical rule of thumb for this problem is to calculate from the surface area of the sphere. If you perform the right calculations, you will find that there are approximately 41,253 square degrees on the surface of a sphere. By calculating the surface area of the beam front (also in square degrees), you can find the power within that region.

Now for the matter of gain in a vertical antenna. The vertical is not gainless because it does not, in fact, radiate equally well in all directions. In fact, the vertical is quite directional, except in the horizontal (azimuth) plane. Figure 7-5 shows the radiation pattern of the typical free-space vertical radiator. The pattern looks like a giant doughnut in free space (see solid pattern in Fig. 7-5). When sliced like a bagel, the pattern is the familiar circular "omnidirectional" pattern. When examined in the vertical plane, however, the plane looks like a sliced figure-8. The "gain" comes from the fact that energy is not spread over an entire sphere, but rather it is concentrated

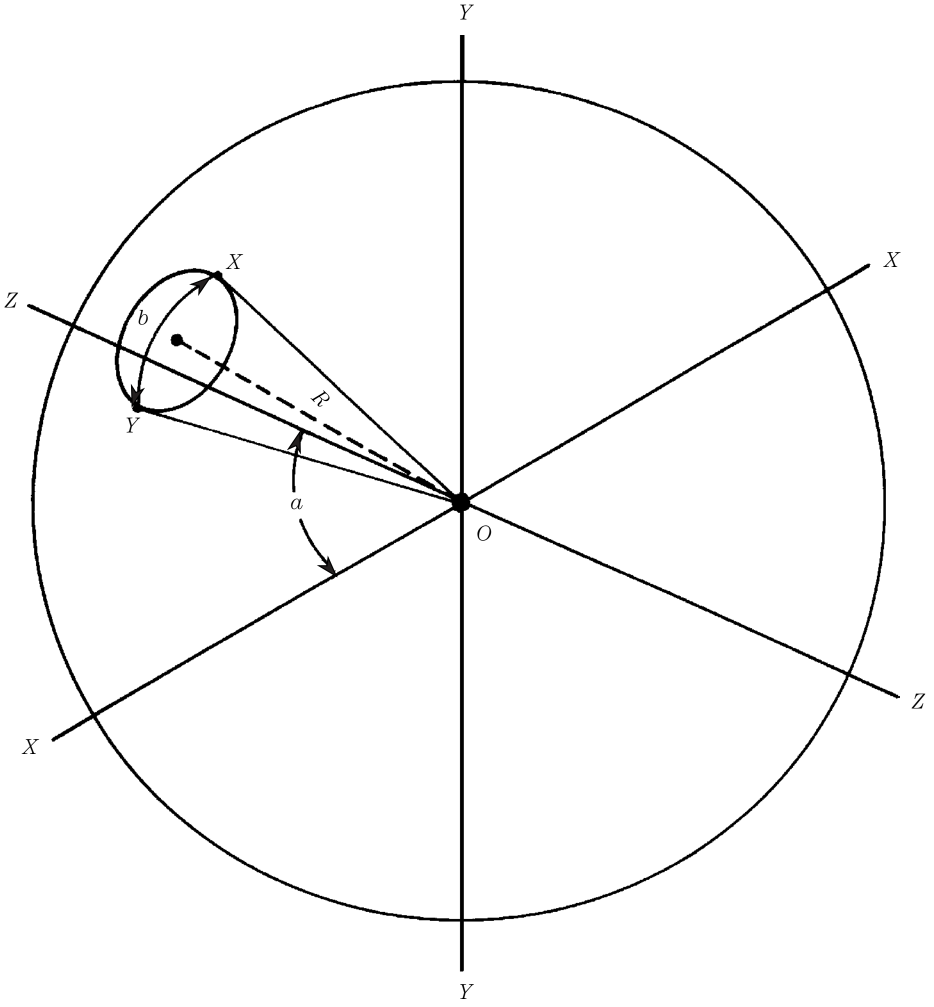

7-4 Spherical or "isotropic" radiation pattern.

to the toroidal doughnut-shaped region shown. Therefore, the power per unit area is greater than for the isotropic (truly omnidirectional) case.

Non-quarter-wavelength verticals

The angle of radiation for a vertical antenna, hence the shape of the hypothetical doughnut radiation pattern, is a function of the length of the antenna. (*Note:* Length in terms of vertical antennas is the same as height, and is sometimes expressed in degrees of wavelength, as well as feet and/or meters.) Figure 7-6A shows the approximate patterns for three different-length vertical antennas: quarter-wavelength, half-wavelength, and ⅝-wavelength. Note that the quarter-wavelength antenna has the highest angle of radiation, as well as the lowest gain of the three cases.

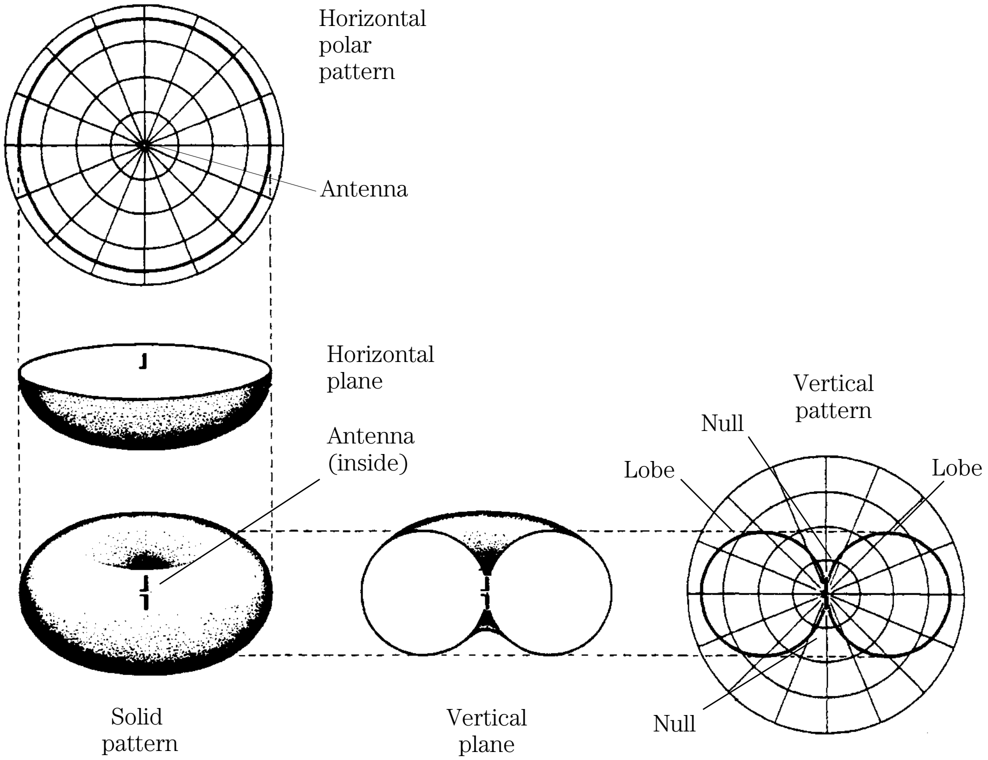
Horizontal
polar
pattern
Antenna
Horizontal
plane
Antenna
(inside)
Vertical
pattern
Null
Lobe
Lobe
Null
Solid
pattern
Vertical
plane

7-5 Vertical antenna radiation pattern.

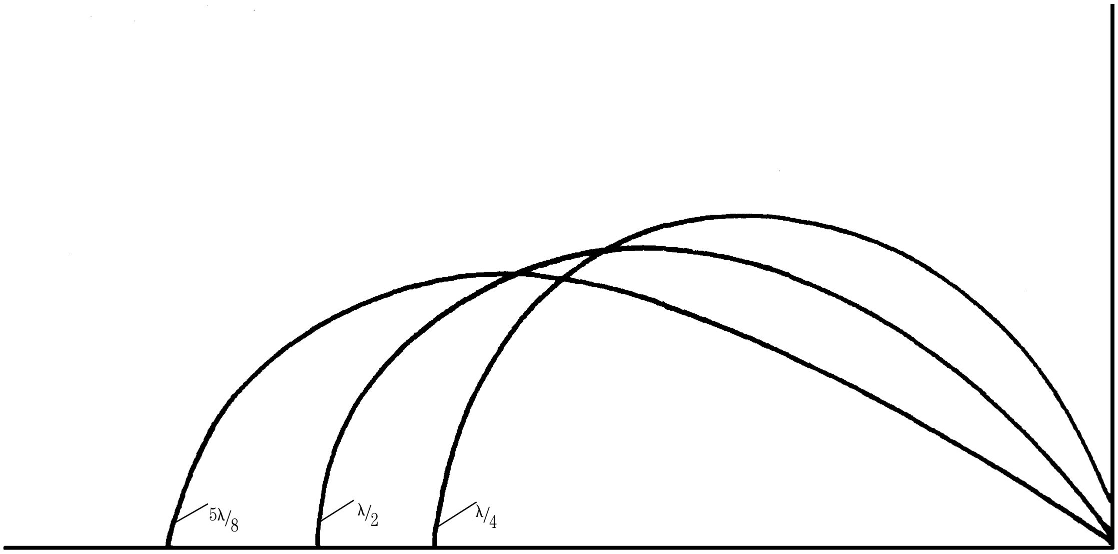

7-6A Vertical radiation pattern as a function of element length: Ideal.

The ⅝-wavelength antenna is both the lowest angle of radiation and the highest gain (compared with isotropic).

The patterns shown in Fig. 7-6A assume a perfectly conducting ground underneath the antenna. However, that is not a possible situation for practical antennas; all real grounds are lossy. The effect of ground losses is to pull in the pattern close to the ground (Fig. 7-6B). Although all of the patterns are elevated from those of Fig. 7-6A, the relationships still remain: the ⅝-wavelength radiator has the lowest angle of radiation and highest gain.

The feedpoint impedance of a vertical antenna is a function of the length of the radiator. For the standard quarter-wavelength antenna, the feedpoint radiation resistance is a maximum of 37 Ω, with only a very small reactance component. Figure 7-7 shows the approximate feedpoint impedances for antennas from nearly zero effective length to 120° of length.

Antenna length expressed in degrees derives from the fact that one wavelength is 360°. Thus, a quarter-wavelength antenna has a length of 360°/4 = 90°. To convert any specific length from degrees to wavelength, divide the length in degrees by 360. Thus, for a 90° antenna: 90°/360° = ¼ wavelength. The graph in Fig. 7-7A shows the antenna feedpoint impedance, both reactance and radiation resistance, for antennas from 60 to 120°; Fig. 7-7B shows the radiation resistance for antennas from near zero to 60°. Note that the radiation resistance for such short antennas is extremely small. For example, an antenna that is 30° long ($^{30}/_{360}$ = 0.083 wavelengths) has a resistance of approximately 3 Ω. It is generally the practice on such antennas to use a broadband impedance-matching transformer to raise the impedance of such antennas to a higher value (Fig. 7-8).

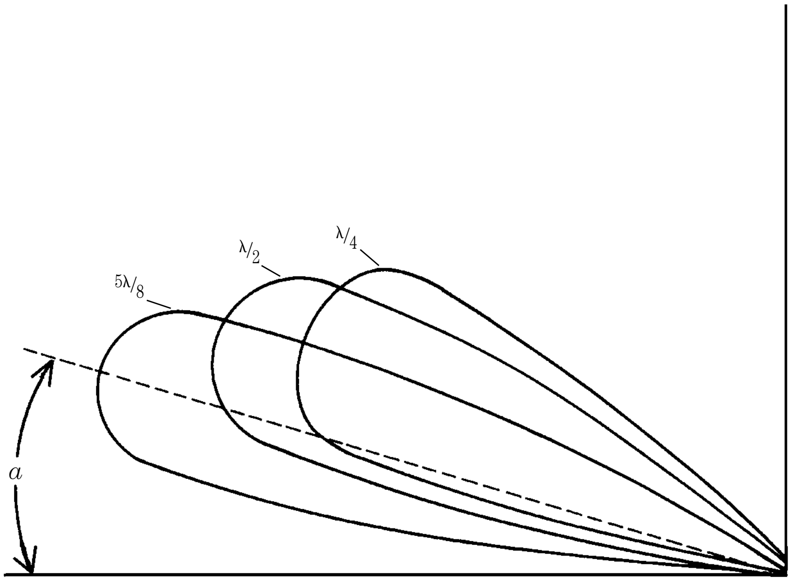

7-6B Accounting for ground losses.

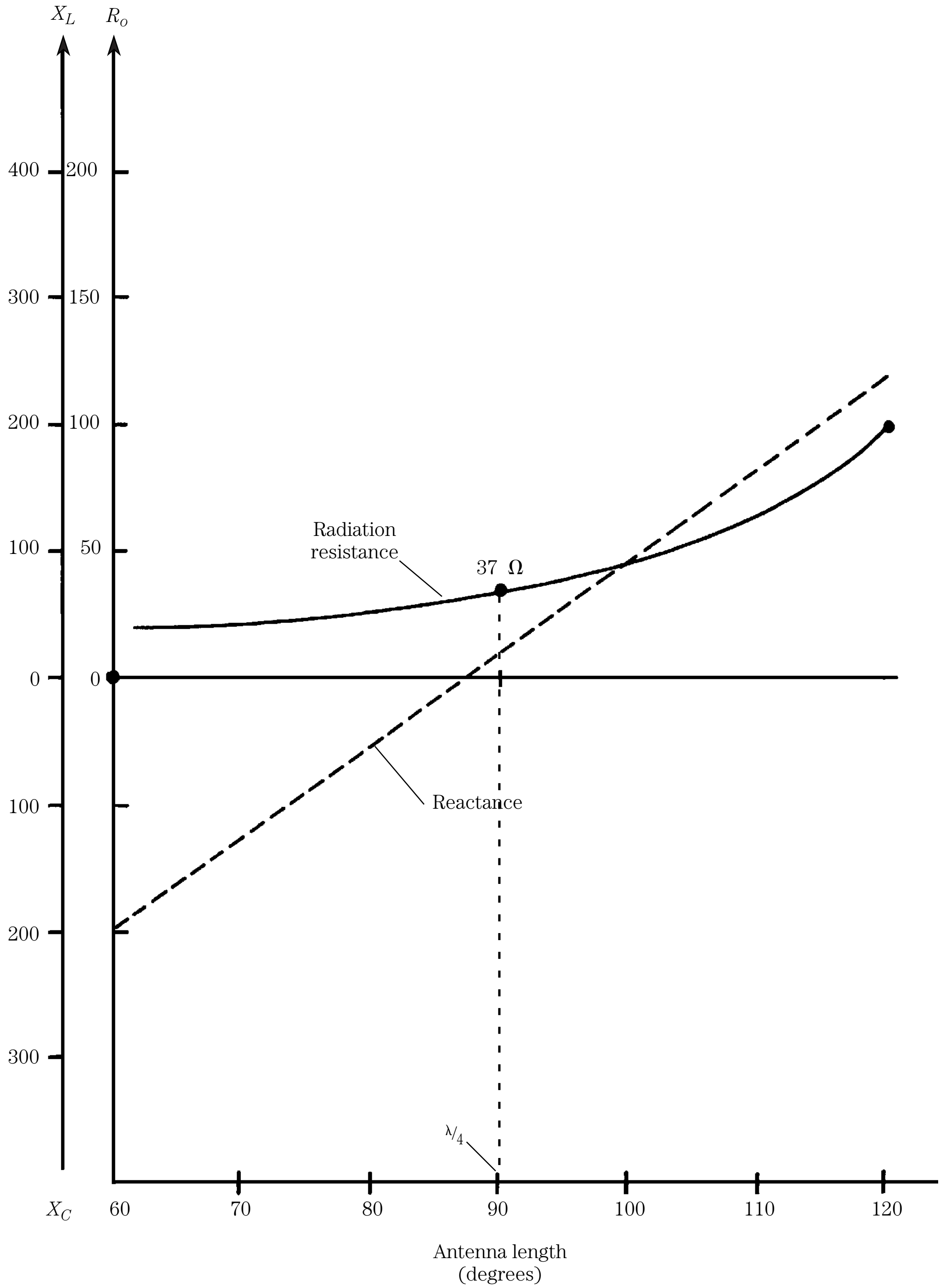

7-7A Antenna impedance as a function of antenna length.

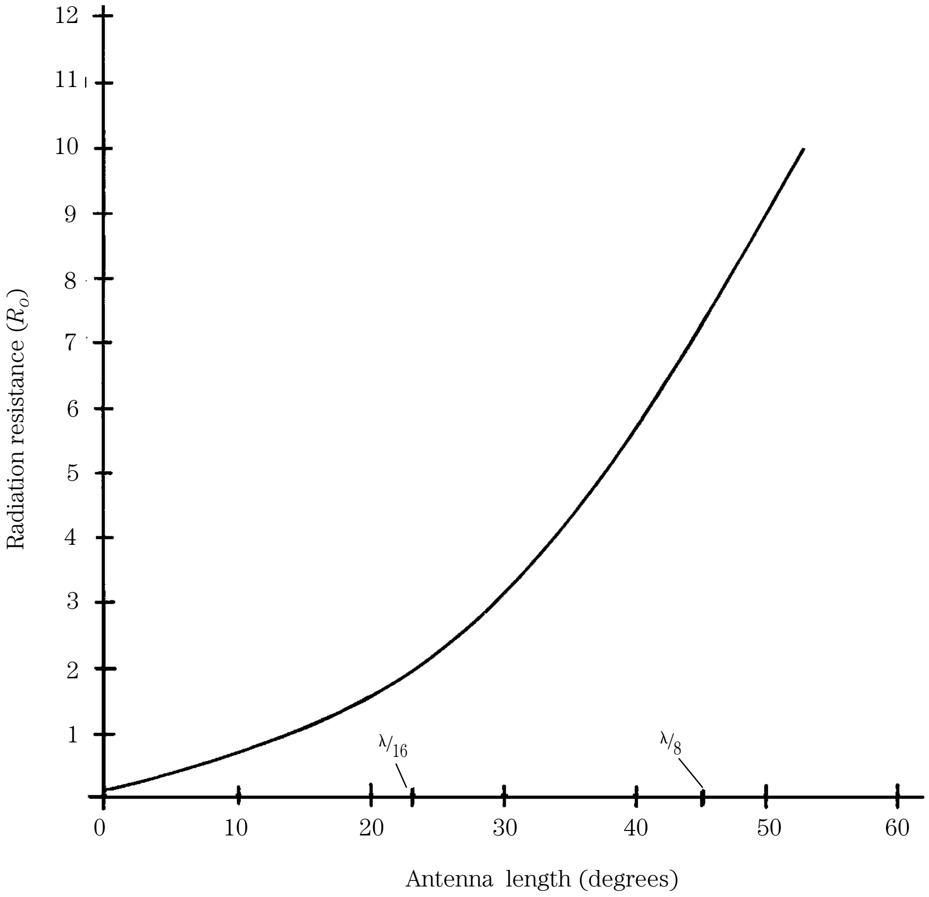

7-7B Radiation resistance as a function of antenna length.

The ground system for the vertical antenna

The vertical antenna works well only when placed over a good ground system. Chapter 28 gives details of proper ground systems for all radio antennas, verticals included, so only certain specifics are included here. The reader should also examine Chap. 28, however, in order to get a firmer grasp of the problem.

The usual way to provide a good ground for a vertical is to use a system of radials such as Fig. 7-9A. This case shows a view (from above) of 16 quarter-wavelength radials arranged to cover the full circle around the antenna. Each radial is a quarterwavelength, so it will have a length (in fact) of $246/F_{\mathrm{MHz}}$. All of the radials are connected together at the base of the antenna, and the ground side of the transmission line is connected to this system. The radials can be placed either on the surface or underground. One friend of the author built an extensive radial system on the bare dirt when his house was built, so when the sod was installed he had a very high-quality underground radial system.

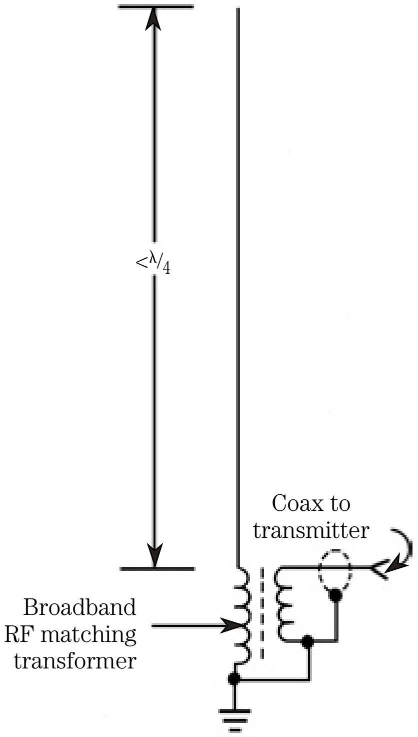

7-8 Impedance matching through broadband transformer.

If you decide to use an aboveground radial system, however, be sure to prevent people from tripping over it. There might be liability implications for people who trip and injure themselves, even when the person is an intruder or trespasser.

Some experts prefer to place a copper wire screen at the center of the radial system. The minimum size of this screen is about 2m square (6 × 6 ft). Connect it to the radials at the points shown using solder. Other experts will drive ground stakes into the ground at these points. Still another method is shown in Fig. 7-9B. This case shows a “spider web” of conductors shorting the radials at points a meter or two from the antenna. Again, some authorities recommend that ground rods be driven into the earth at the indicated points.

The exact number of radials to use depends in part on practical matters (how many can you physically install?). Use at least two radials per band, with four per band preferred for simple, low-cost systems. However, even four is considered a compromise case. The general rule is: *the more the better*. But it’s also true that there is a law of diminishing returns as the number of radials is increased. Figure 7-9C shows the approximate field intensity (mV/m) as a function of the number of radials. Notice that the field intensity does not increase as rapidly with the number of radials above 20 or so. Note that the Federal Communications Commission requires AM band (550 to 1700 kHz) stations to use 120 radials, but that number is not necessary for amateur stations. A practical upper limit of 16 radials is usually accepted for amateur radio work, and four radials work quite well.

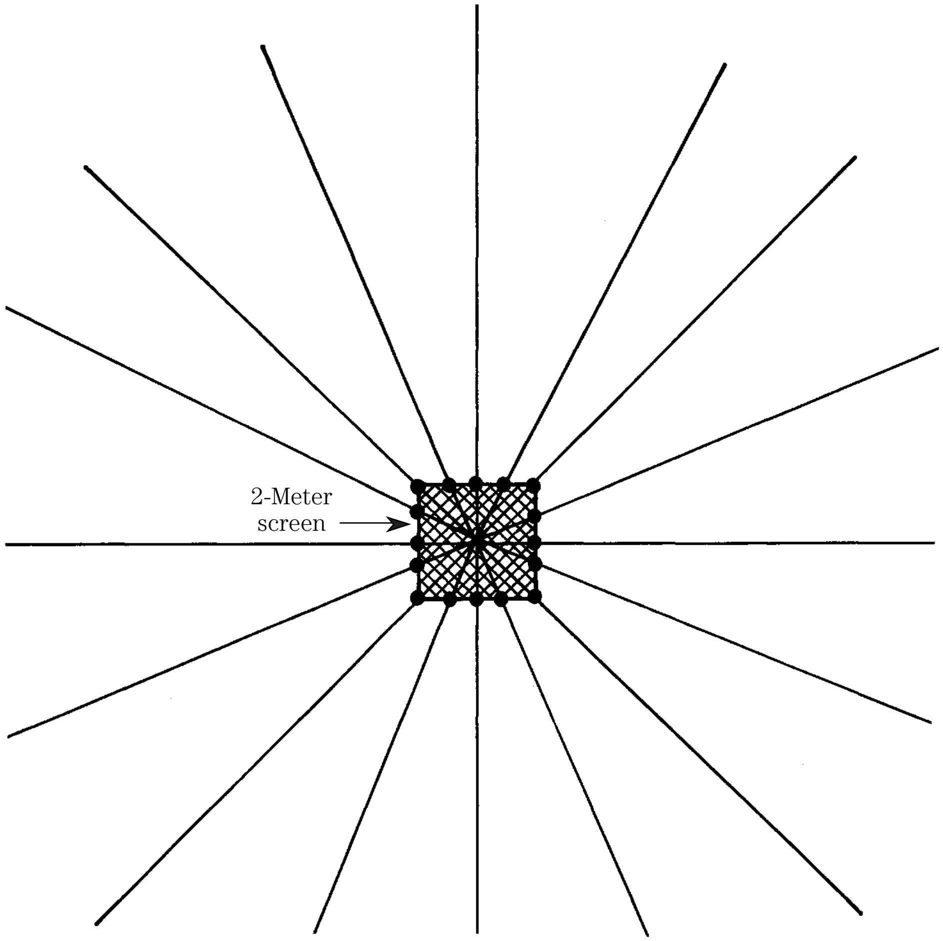

7-9A Comprehensive ground system for vertical antenna.

For vertical antennas mounted above ground, there is an optimum height above ground for the base of the antenna. This height is a quarter-wavelength above the actual ground plane. Unfortunately, that distance might not be the height above the surface. Depending upon ground conductivity and groundwater content, the height can be exactly a quarter-wavelength above the surface or slightly lower. The point is found from experimentation, and will, unfortunately, vary over the year if climatic changes are usual.

Variations on the vertical antenna theme

Thus far, the vertical antennas have been standard quarter- or ⅝-wavelength models. This section looks at several variations on the theme. Consider Fig. 7-10. This antenna is the vertical half-wavelength dipole. The vertical dipole is constructed in exactly the

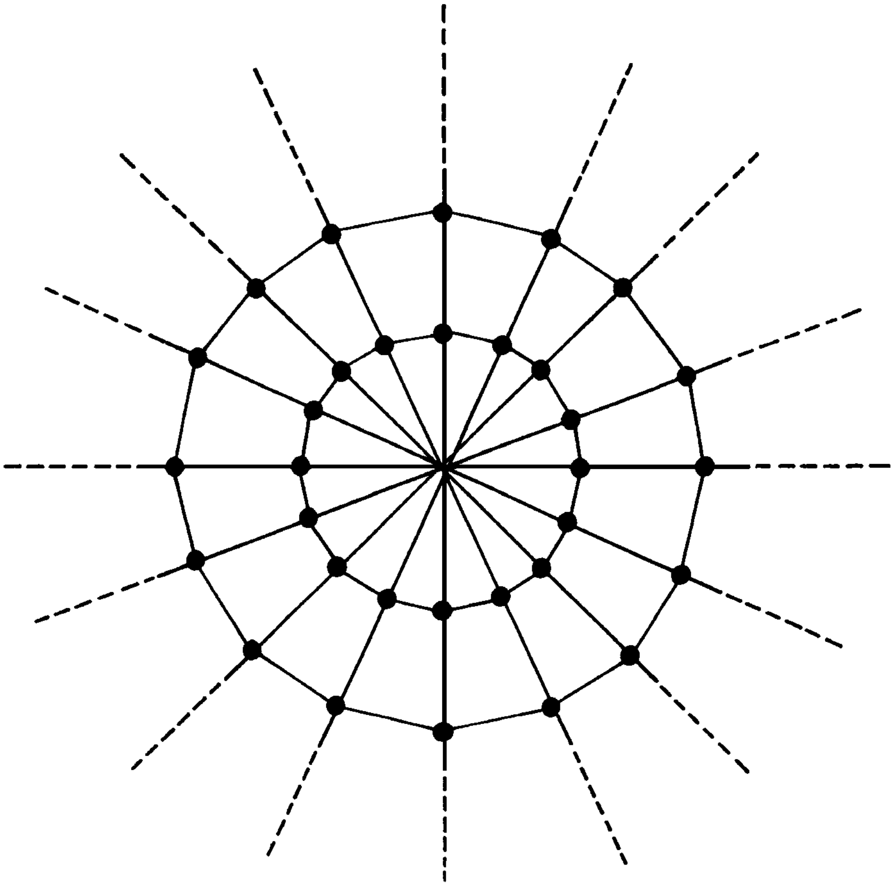

7-9B "Spider web" ground system.

same manner as the horizontal dipole, but is mounted in the vertical plane. In general, the section of the radiator that is closest to the ground should be connected to the shield end of the coaxial cable transmission line.

Like the horizontal dipole, the approximate length of the vertical dipole is calculated from

$$L_{\text{ft}} = \frac{468}{F_{\text{MHz}}} \qquad \textbf{[7.2]}$$

where

L_{ft} is the length, in feet
F_{MHz} is the operating frequency, in megahertz

Example Calculate the length of a half-wavelength vertical dipole for operation on a frequency of 14.250 MHz in the 20-m amateur radio band.

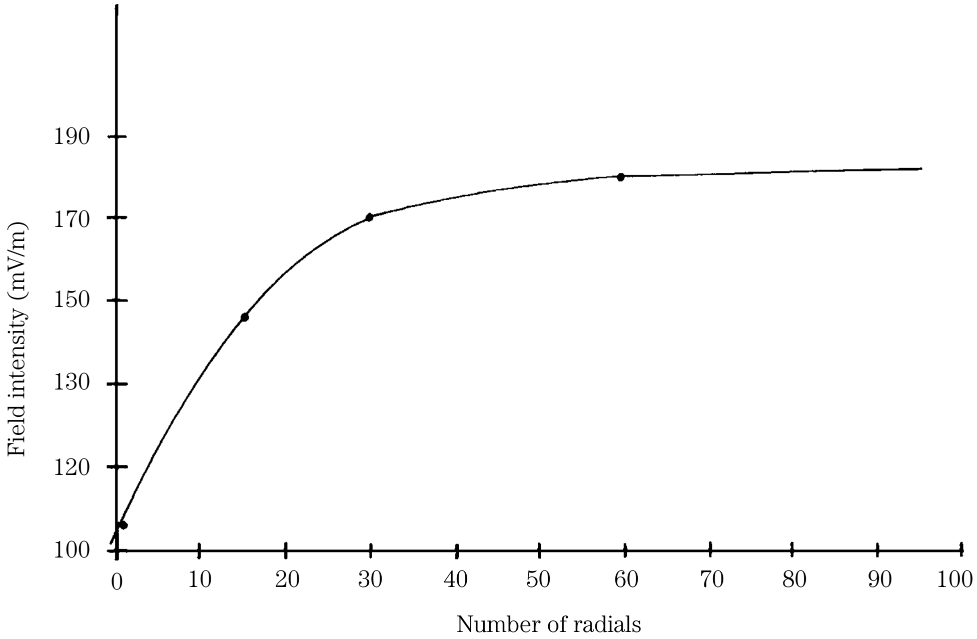

7-9C Effect of increasing the number of radials on field strength.

Solution:

$$L_{\text{ft}} = \frac{468}{F_{\text{MHz}}}$$

$$L_{\text{ft}} = \frac{468}{14.250} = 32.8 \text{ ft}$$

Note: The 0.8 ft part of this calculated length can be converted to inches by multiplying by 12: 0.8 × 12 = 9.6 in.

Each leg of the vertical dipole is one half of the calculated length, or

$$\frac{32.8 \text{ ft}}{2} = 16.4 \text{ ft}$$

The vertical dipole antenna is used in many locations where it is impossible to properly mount a horizontal dipole, or where a roof- or mast-mounted antenna is impossible to install because of logistics, a hostile landlord, and/or a homeowners' association. Some row house and town house dwellers, for example, have been successful with the vertical dipole. In the 1950s and 1960s, the vertical dipole was popular among European amateurs because of space restrictions found in many of those locations.

The construction of the vertical dipole is relatively straightforward. You must find or build a vertical support structure. In the case shown in Fig. 7-10, the support

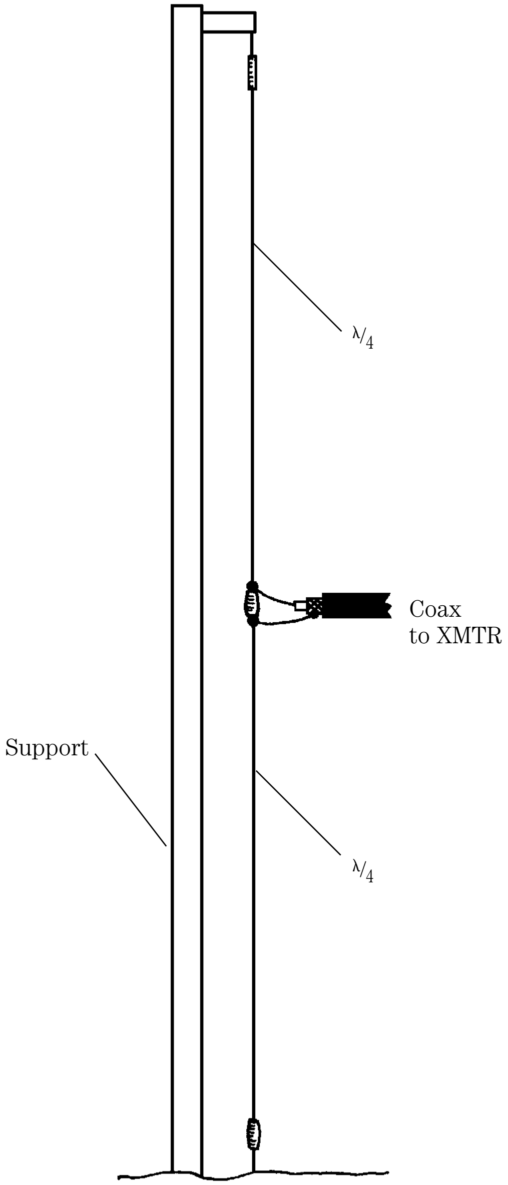

7-10 Vertical half-wavelength dipole antenna.

is a wooden or PVC mast erected for that purpose. Ropes and insulators at either end support the wire elements from the ends and keep the antenna taut. If the neighbors are a problem, then try to find some white PVC pipe that will make a fine flagpole and be patriotic—with a vertical dipole hidden inside of the pipe. In other cases, if your home is not metal-sided, and if it is high enough, then a support from the roof structure (or soffits) will make a proper mount.

One problem with the vertical dipole, and one that liability-conscious people need to consider, is that a high-impedance current node is found at the ends of a half- wavelength dipole. Anyone touching the antenna will likely receive a nasty RF burn (or shock) from this antenna.

A *coaxial vertical* is similar to the vertical dipole (and, in fact, it can be argued that it is a form of vertical dipole) in that it uses a pair of vertical radiator elements. In the case of the coaxial vertical antenna, however, the radiator that is closest to the ground is coaxial with the transmission line and the main radiator element. An example is shown in Fig. 7-11A. An insulator at the feedpoint separates the two halves of the radiator; in most cases, it is of smaller diameter than the coaxial sleeve (also called the *shield pipe* in some publications). The reasons for this arrangement are not entirely electrical, for the most part, but mechanical. The coaxial cable transmission line passes through the sleeve, and is itself coaxial to the sleeve.

The overall length of the coaxial vertical antenna is a half-wavelength, which consists of two quarter-wavelength sections. Both the radiator and the sleeve are quarter-wavelength. The length of each is found (approximately) from

$$L_{\text{ft}} = \frac{234}{F_{\text{MHz}}} \tag{7.3}$$

or

$$L_{\text{meters}} = \frac{72}{F_{\text{MHz}}} \tag{7.4}$$

You should recognize these as similar to the equation used previously to calculate half-wavelength antennas, but reduced by a factor of 2.

The coaxial vertical antenna was once popular with CB operators, and as such was called the *collinear antenna*. In some cases, you can find hardware from these antennas on the hamfest or surplus markets, and the pieces can be modified for amateur radio use. In the situation where a 10-m-band antenna is being built, it is a simple matter to cut the 11-m CB antenna for operation on a slightly higher frequency. In the case of the lower-frequency bands, however, it is a little more difficult and it is likely that only the insulator and mounting assembly are salvageable. Keep in mind, however, the fact that adjacent sizes of aluminum tubing are designed such that the inside diameter (ID) of the larger piece is a slip-fit for the outside diameter (OD) of the smaller piece. You can, therefore, connect adjacent sizes of aluminum tubing together without the need for special couplers, etc. With that in mind, salvaged insulator assemblies can be cut off with just 6 to 10 in of the former radiator and sleeve, and new radiators from “adjacent size” tubing installed.

The configuration shown in Fig. 7-11A is the manner of construction used by commercial antenna manufacturers for VHF and CB collinear vertical dipoles, but is a little difficult for amateurs (unless they happen to own a machine shop) to make the center insulator. For those people, some other method is indicated before this antenna is practical.

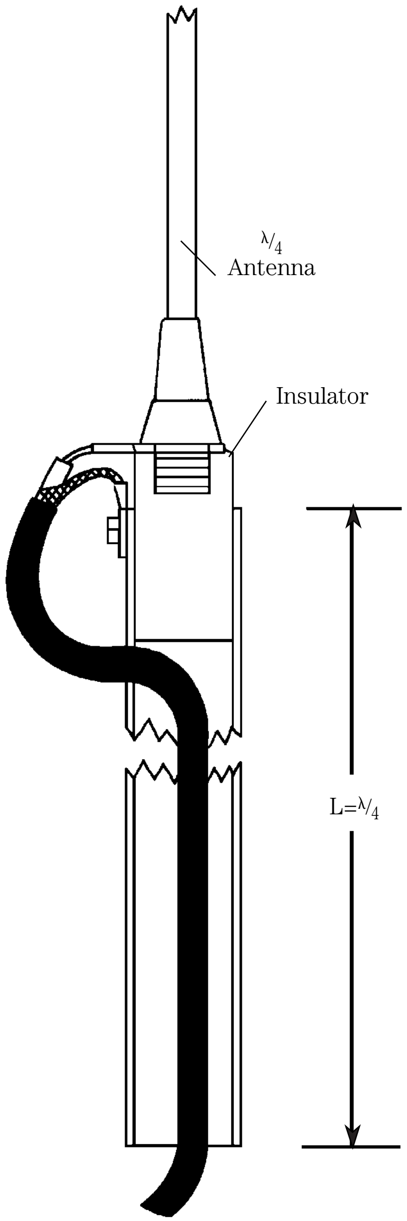

7-11A Construction of the vertical coaxial antenna.

Figure 7-11B shows a construction method that has been used by amateurs with good results. The radiator and shield pipe (sleeve) are joined together in an insulating piece of thick-wall PVC plumbing pipe, Lucite, or plexiglass tubing; 6 to 10 in of tubing are needed.

A gap of about 2 in is left between the bottom end of the radiator pipe and the top end of the shield pipe in order to keep them electrically insulated from each

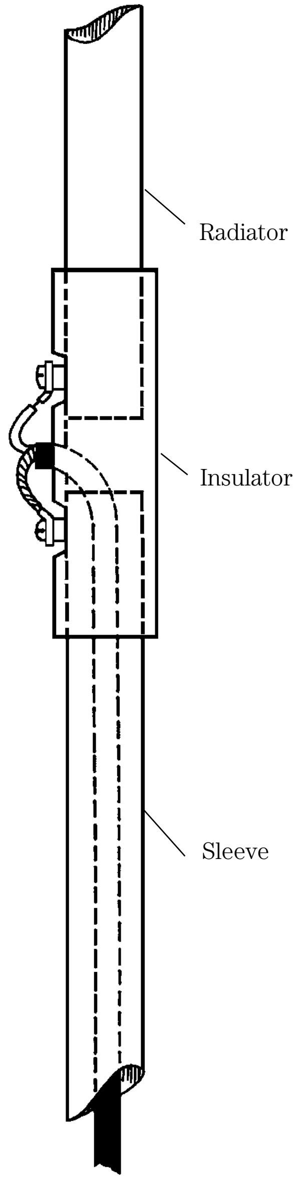

7-11B Vertical dipole made from tubing.

other, and to allow the coaxial cable to be passed through to the outside world. A hole in the insulator pipe is drilled for this purpose.

The aluminum tubing pieces for the radiator and the sleeve are fastened to the insulator with at least two heavy machine screws each. One of the machine screws on each can be used as the electrical connection between the coaxial cable and the pipes, provided that a larger hole is cut in the insulator at that point to admit the washer that provides the electrical pathway between the screw head and the aluminum pipe.

If you depend upon the machine screw touching the pipe at the edges of the hole cut for it, then there will probably be intermittent connection and all of the aggravation that ensues.

Mounting of the homebrew coaxial vertical antenna can be a “pain in the neck.” Normally, this antenna is mounted high in the air above ground, so some form of support is needed. Fortunately, you can use small-area metal supports connected to the sleeve for this purpose. Figure 7-11C shows one method for mounting that is popular. A pair of television antenna standoff mounting brackets are used to support the sleeve. Those brackets can be bought in sizes from 6 to 24 in. Note that a 2 × 4 piece of lumber is used between the building wall and the brackets. This wood serves as an

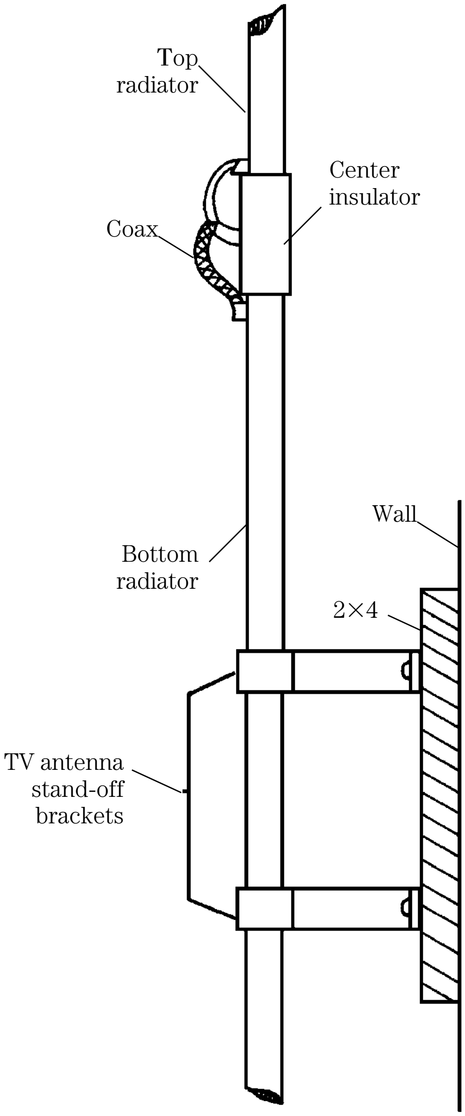

7-11C Mast mounting system for verticals.

insulator, so it should be varnished or painted. It is held to the wall with lag bolts, wing bolts, or some other effective method of anchoring. Keep in mind that the forces on the brackets increase tremendously during wind storms!

It is possible for the two vertical antennas shown previously to become a shock hazard to anyone who touches them. Both of these antennas are half-wavelength radiators and are of the dipole form of construction. The center point is used for feeding the antenna, so it forms the low-impedance point in the antenna. As a result, the ends of the antenna, one of which is close to the ground, are the high-impedance points—hence the voltages at those points within reach of prying hands playing in the yard can be high. It is wise to either mount the antennas so far above the ground that they cannot be reached, or build a small nonconductive fence around the bottom end of the antenna.

Vertical antenna construction

There are two general cases for installing vertical antennas: ground-level-mounted and non-ground-mounted. This section takes a brief look at both forms of mounting. We will concentrate on installation of homebrew verticals rather than commercial, because it is assumed that the vendors of such antennas will provide their own instructions.

The ground-level-mounted vertical is shown in Fig. 7-12. The typical vertical antenna is 8 to 40 ft high. Thus, although the actual weight of the antenna is small, the forces applied to the mounting structure (especially during windstorms) can be quite high. Don’t be fooled by the apparent light weight of the antenna in this respect.

The mounting structure for the vertical antenna can be a metal or wooden fence post buried in the ground. At least 2 ft of the fence post should be above ground. In the case of Fig. 7-12, a 4 × 4 wooden fence post is used as the mounting, but the principles are similar for all forms of post. A fence post hole is provided that is at least 2 ft deep. In some cases, it might be possible to use a 1-ft-deep gravel fill topped with back-filled dirt. In other cases, especially where a steel fence post is used, a concrete plug is placed at the bottom of the hole over a 4-in layer of gravel.

The antenna radiator element is installed onto the fence post using standoff insulators. Unfortunately, these insulators are difficult to find, so they might have to be omitted. Given that varnished or painted wood is not a terribly good conductor, it is not unreasonable to bolt the radiator directly to the 4 × 4 fence post. Use $\frac{5}{16}$-in (or larger) bolts, and make them long enough to fit completely through both the antenna element and the 4 × 4 post. Thus, 6-in-, 7-in-, or 8-in-long $\frac{5}{16}$-in bolts are the candidates for this job. Use at least two bolts, one at the bottom of the antenna radiator element, and one near the top of the fence post. A third bolt, halfway between the other two, would not be out of order.

If the antenna is quarter-wavelength, then no matching is generally necessary. Although the feedpoint impedance is not exactly 52 Ω, it is close enough (37 Ω) to form a reasonable match for 52-Ω coaxial cable (with VSWR = 1.4:1). The center conductor of the coaxial cable is connected to the radiator element, while the shield is connected to the ground system. In the example shown in Fig. 7-12, two ground methods are used. First, an 8-ft ground rod is driven into the earth at the base of the

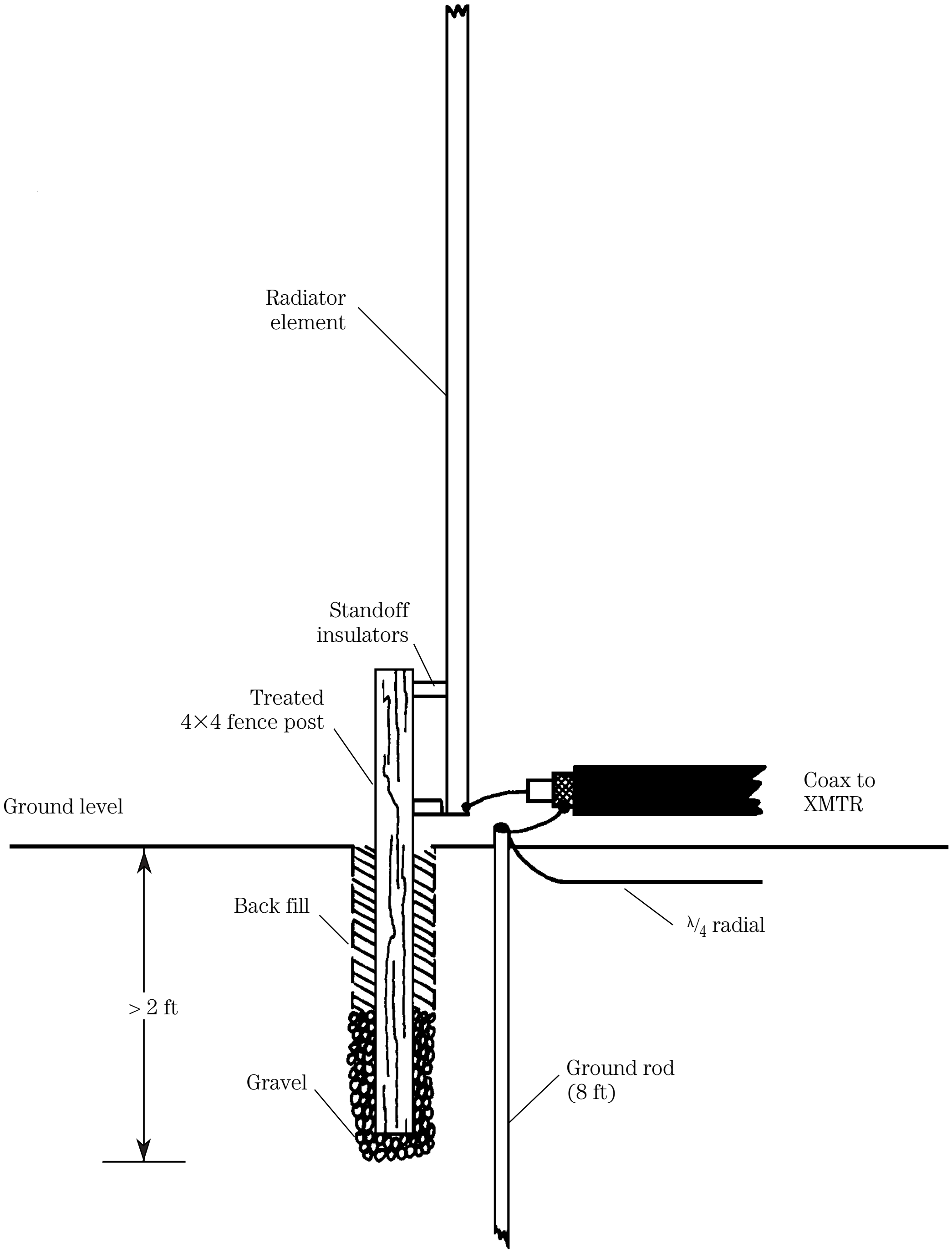
Radiator element
Standoff insulators
Treated 4×4 fence post
Coax to XMTR
Ground level
Back fill
$\lambda/_4$ radial
> 2 ft
Gravel
Ground rod (8 ft)

7-12 Ground-mounted vertical.

antenna; second, a system of quarter-wavelength radials is used. Remember that the ground system is absolutely essential.

A method for installing a vertical antenna above ground is shown in Fig. 7-13. In this case, a wooden support (2 × 4 or 4 × 4) is installed in a manner similar to Fig. 7-12, but with a deeper hole to counter the longer length. Alternatively, the wooden support is affixed to the side of a building wall, shed, or other preexisting structure. Once the support is arranged, however, the method of attachment of the radiator element is the same for the previous case, so that will not be repeated here.

Electrical connections to the antenna are also shown in Fig. 7-13. Because the antenna is above ground level, an electrical counterpoise ground consisting of a system of radials is absolutely essential; at least two radials per band must be provided. A small L bracket is used to support the radials and to provide an SO-239 coaxial connector for the coax. This connector is a chassis-mounted type with its center conductor connected to the radiator element. The shield of the connector is connected to the bracket, so it is also connected to the radial system.

In some installations, the antenna support structure will require guy wires to keep the structure stable. Do not use the radials as guy wires. The type of wire that normally works well for radials is too soft, and too easily stretched, for guy wire service. Use regular steel guy line, available where TV antenna supplies are sold, for this antenna. Make the lengths nonresonant, and break the guy lines up with egg insulators, if necessary, to achieve the nonresonance.

⅝-Wavelength verticals

Figure 7-14 shows the configuration for the ⅝-wavelength vertical antenna. Such an antenna generally gives a lower angle of radiation than the more common quarter-wavelength radiator, so presumably it works better for long distance.

The radiator of this antenna is made from 0.5-in to 1.5-in aluminum tubing. Again, remember that adjacent sizes fit together snugly to form longer sections. The physical length of the ⅝-wavelength radiator is found from

$$L_{\text{ft}} = \frac{585}{F_{\text{MHz}}} \tag{7.5}$$

or, in meters,

$$L_{\text{meters}} = \frac{180}{F_{\text{MHz}}} \tag{7.6}$$

The radials are the usual quarter-wavelength, and are made of no. 12 or no. 14 copper wire. These lengths are found from:

$$L_{\text{ft}} = \frac{246}{F_{\text{MHz}}} \tag{7.7}$$

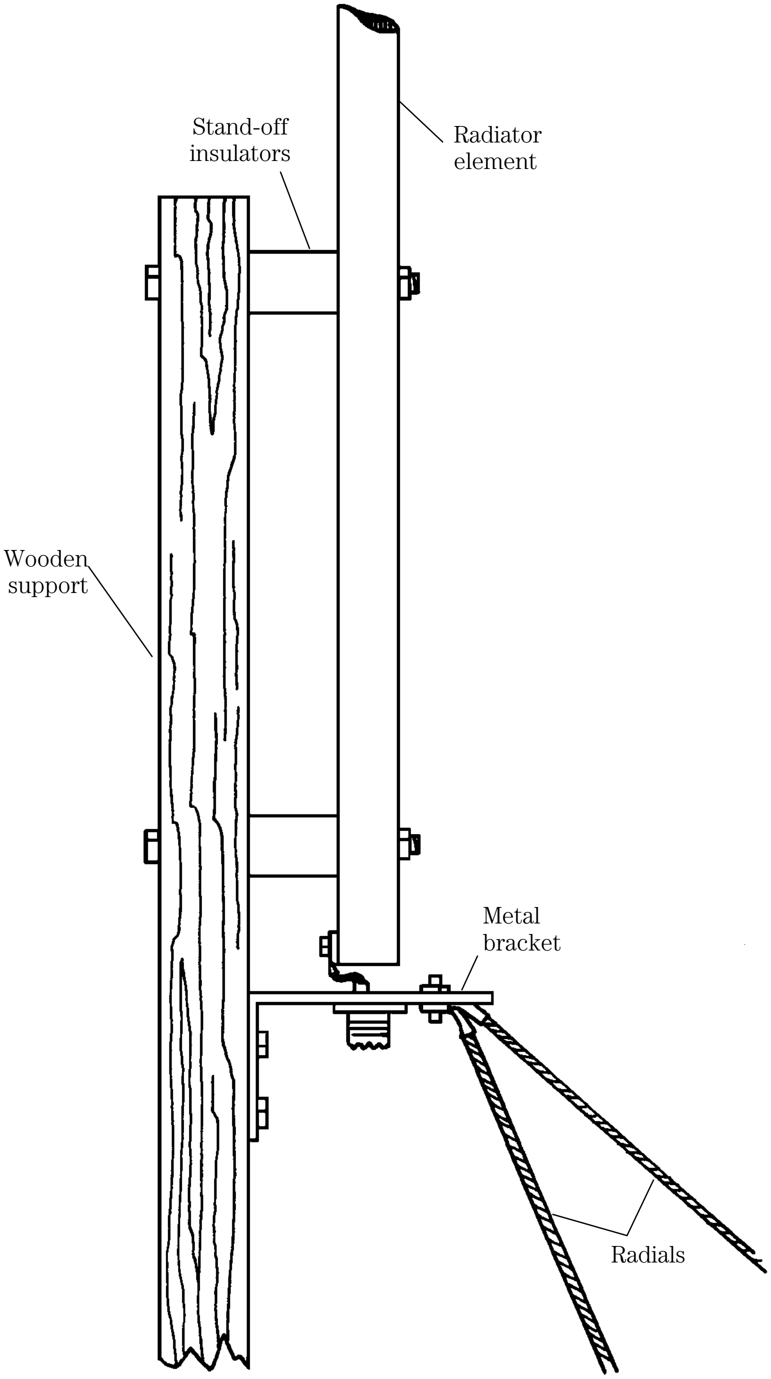
Stand-off
insulators
Radiator
element
Wooden
support
Metal
bracket
Radials

7-13 Details of feedpoint circuit on vertical.

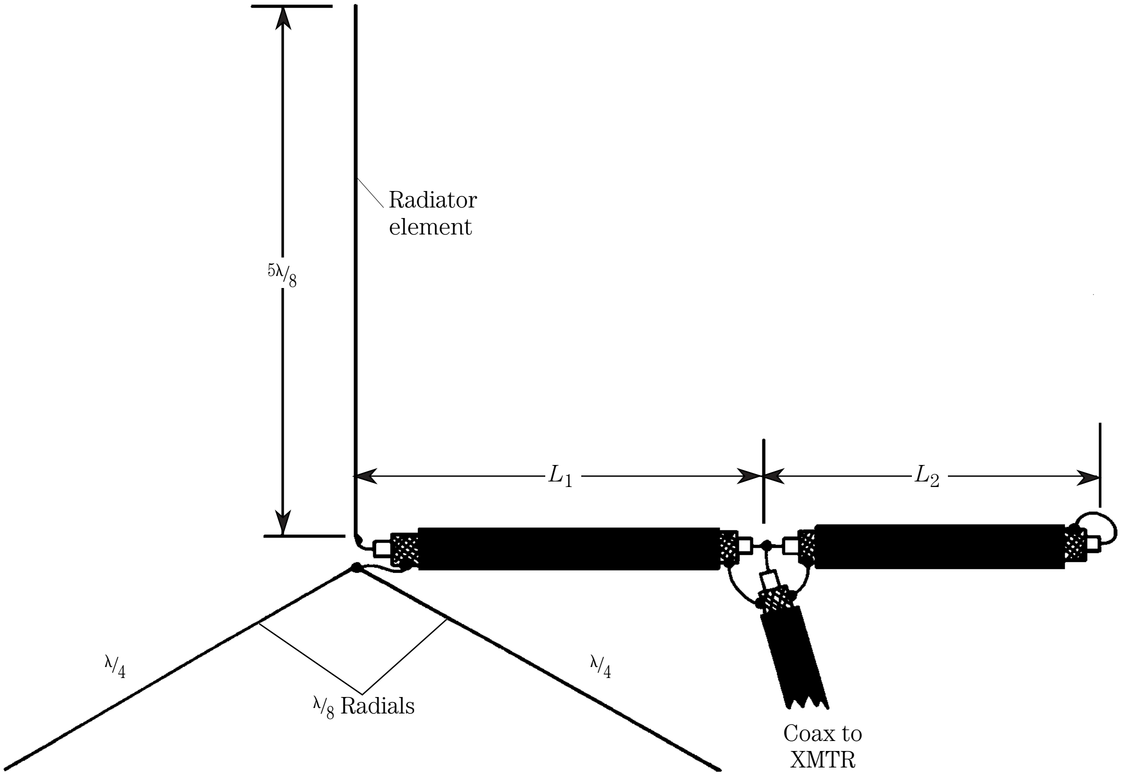

7-14 Using a Q section to match the feedpoint impedance of a vertical antenna.

or, in meters,

$$L_{\text{meters}} = \frac{75}{F_{\text{MHz}}} \qquad \textbf{[7.8]}$$

The feedpoint impedance of the ⅝-wavelength antenna is about 1600 **Ω**, not a good match for the ordinary coaxial cables that are routinely available on the amateur market. Some form of impedance matching is needed.

One option is to use a broadbanded RF transformer. These transformers will work throughout the HF spectrum, and match a wide variety of impedances to the 50-**Ω** standard system impedance.

Another option, especially for a single-band antenna, is to use a coaxial cable impedance transformer, such as shown in Fig. 7-14. The transformer consists of two sections of coaxial cable joined together, shown as L_1 and L_2 in Fig. 7-14. The lengths are found from

$$L_1 = \frac{122}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[7.9]}$$

or

$$L_1 = \frac{38}{F_{\text{MHz}}} \text{ m} \qquad [7.10]$$

and

$$L_2 = \frac{30}{F_{\text{MHz}}} \text{ ft} \qquad [7.11]$$

$$L_2 = \frac{9}{F_{\text{MHz}}} \text{ m} \qquad [7.12]$$

Grounded vertical antennas

The vertical antennas presented thus far in this chapter are called *series-fed verticals*, because the generator is essentially in series with the radiator element. Such an antenna must be insulated from ground. The other class of vertical is the *shunt-fed vertical*, which is grounded at one end (Fig. 7-15). There are three methods of shunt-feeding a grounded vertical antenna: *delta, gamma*, and *omega*. All three matching systems have exactly the same function: to form an impedance transformation between the antenna radiation resistance, at the feedpoint, and the coaxial cable characteristic impedance, as well as cancelling any reactance in the system.

The delta feed system is shown in Fig. 7-15A. In this case, a taut feed wire is connected between a point on the antenna, which represents a specific impedance on

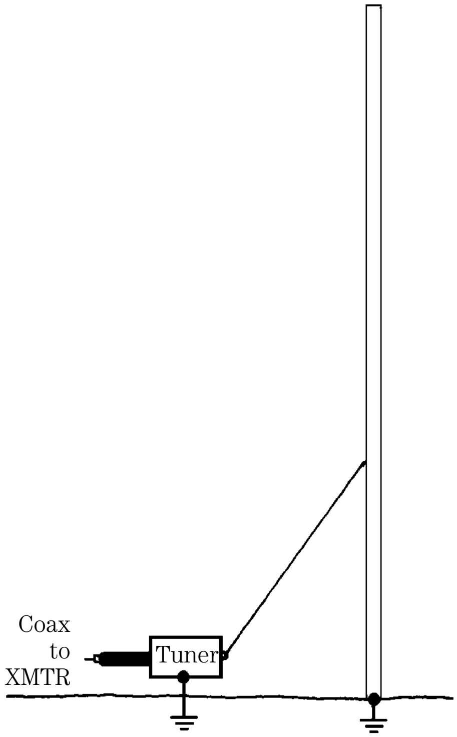

7-15A Delta-fed grounded vertical.

the antenna, and an antenna tuner. This method of feed is common on AM broadcast antennas (which are usually—perhaps always—verticals). Although you would think that the sloping feed wire would distort the pattern, that is not the case. The distortion of the pattern, if any, is minimal, hence it can be neglected.

The gamma feed system is shown in Fig. 7-15B. This method is commonly used by amateurs to feed Yagi beam antennas, so it is quite familiar in the amateur radio world. The feed system consists of a variable capacitor to tune the system, and a matching rod that parallels the antenna radiator element. It is important that the rod not be anywhere near a quarter-wavelength, or it would become a vertical antenna in its own right, and in fact would resemble the so-called *J-pole antenna*. A review of the gamma match is given in Chap. 12. The omega feed (shown in Fig. 7-15C) is similar to the gamma match except that a shunt capacitor is used.

Conclusion

The vertical antenna is a viable alternative for many situations, especially where real estate is at a premium. Contrary to popular opinion, the vertical antenna works well when installed properly and when due consideration has been given to matters such as the grounding and angle of radiation desired.

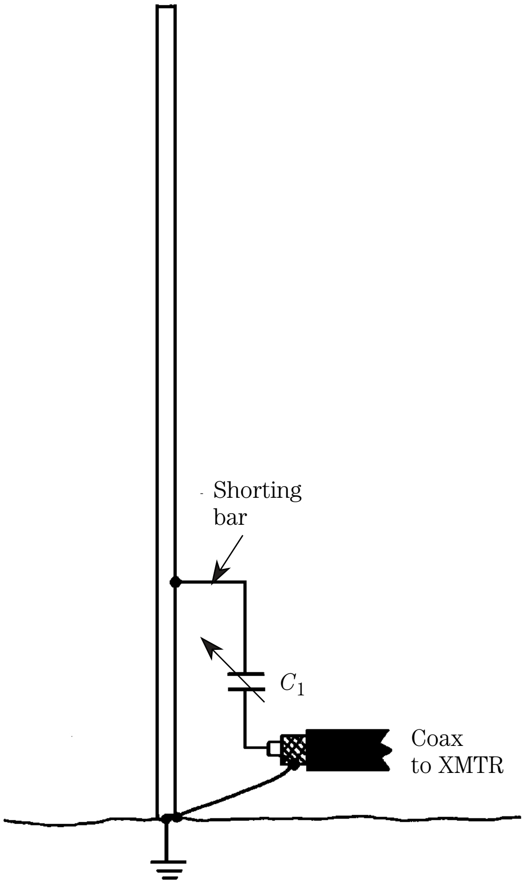

7-15B Gamma-fed grounded vertical.

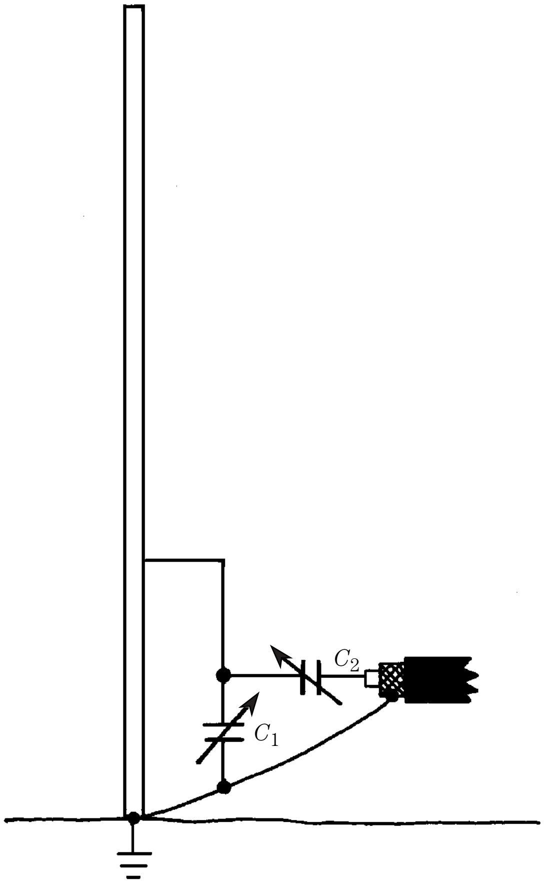

7-15C Omega-fed grounded vertical.

8
CHAPTER

Multiband and tunable-wire antennas

MOST COMMUNICATIONS OPERATORS REQUIRE MORE THAN ONE BAND, AND THAT MAKES the antenna problem exactly that—a problem to be solved. Amateur radio, commercial, and military operators are especially likely to need either multiple antennas for different bands, or a multiband antenna that operates on any number of different bands. This situation is especially likely on the high-frequency (HF) bands from 3.5 to 29.7 MHz.

Another problem regards the tunability of an antenna. Some amateur bands are very wide (several hundred kilohertz), and that causes any antenna to be highly variable from one end of the band to another. It is typical for amateurs to design an antenna for the portion of the band that they use most often, and then tolerate a high VSWR at the other frequencies. Unfortunately, when you see an antenna that seems to offer a low VSWR over such a wide range, it is almost certain that some problem exists that reduces the Q, and the antenna efficiency, to broaden the response. However, it is possible to tune an antenna for a wide band. It is also possible (now that amateurs have new HF bands) to use a single antenna between them, and then tune the difference out. For example, designing a single antenna for 21/24 MHz, 14/18 MHz, or 7/10 MHz should prove possible.

In this chapter we will take a look at both problems: the multiband and the tunable antenna.

Multiband antennas

Although a triband Yagi or quad beam antenna will undoubtedly work better than a wire antenna (when installed correctly!), the low-budget amateur operator need not lament any supposed inability to "get out" on wire antennas. To quote an old saying: "'Better' is the enemy of 'good enough.'" Or, to put it in terms of *Carr's law:* "If it's 'good enough' then don't waste a lot of energy fretting over making it 'better' unless you really want to make it *a lot better*."

Copyright 2001 - the McGraw-Hill Companies

Trap dipoles

Perhaps the most common form of multiband wire antenna is the *trap dipole* shown in Fig. 8-1A. In this type of antenna, one (or more) pairs of parallel resonant traps are placed in series with the quarter-wavelength elements of the dipole. The purpose of the traps is to block their own resonant frequency, while passing all other frequencies.

In the example of Fig. 8-1A, a 10-m trap isolates the first 8 ft or so (quarter-wavelength on 10 m) so that the antenna resonates on that band. A 40-m or 15-m signal, on the other hand, passes through the traps and uses the whole length of the antenna. (*Note*: A half-wavelength 40-m dipole works as a 3/2-wavelength antenna on 15 m.)

The overall length of the trap dipole will be a little less than the natural "non-trap" length for the lowest frequency of operation. At the low frequencies, the traps add a little inductance to the circuit so that the resonant point is lower than the natural resonant frequency. In general, most trap dipoles are just a few percent shorter than nontrap dipoles at the same band. The actual amount of shortening depends upon the values of the components in the traps, so consult the data for each trap purchased. Where more than one pair of traps is used in the antenna, make sure they are of the same brand and are intended to work together.

Another solution to this problem is shown in Fig. 8-1B. This type of antenna actually has two or more half-wavelength dipoles fed from the same transmission line. In this illustration, a total of three dipoles are fed from the same 75-Ω transmission line. There is no theoretical limit to how many dipoles can be accommodated, although there is certainly a practical limit. For one thing, there is a mechanical limit to how many wires are supportable (or desirable) hanging from any given support. There is also an electrical limit, although it is less defined. Having a lot of dipoles increases the possibility of radiating harmonics and other spurious emissions from your transmitter.

The 75-Ω coaxial cable is connected to the center feedpoint of the multidipole either directly or through a 1:1 balun transformer, as shown in Fig. 8-1B. Each an-

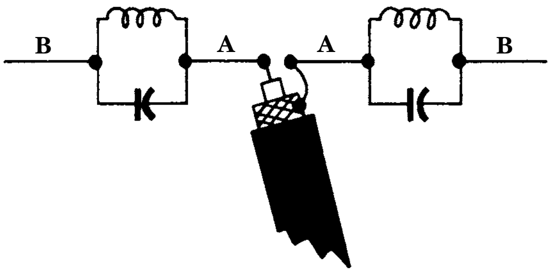

A=10-meter segment
A+B=15-meter/40-meter segment

8-1A Trap dipole for multiband operation. (*Courtesy of Hands-On Electronics and Popular Electronics*)

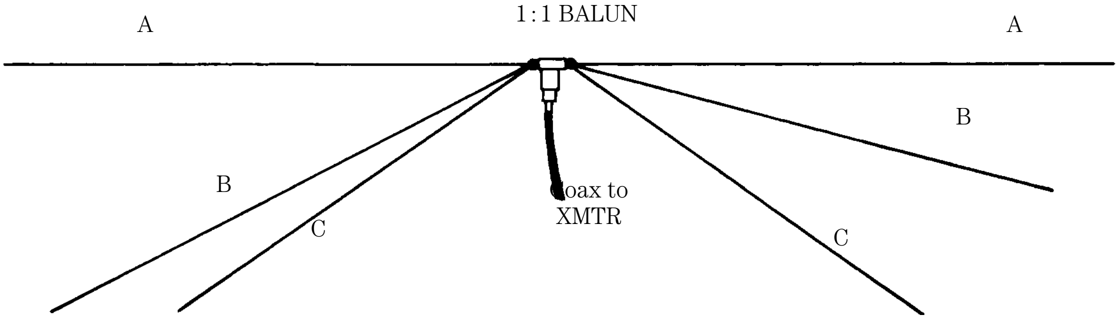

8-1B Multiband dipole consists of several dipoles fed from a common feedline.

tenna (*A-A*, *B-B*, or *C-C*) is a half-wavelength. Therefore, the overall length is found approximately from the standard dipole expressions:

Overall length ($A + A, B + B,$ or $C + C$):

$$L_{\mathrm{ft}} = \frac{468}{F_{\mathrm{MHz}}} \qquad \textbf{[8.1]}$$

or, for each element alone (*A*, *B*, or *C*):

$$L_{\mathrm{ft}} = \frac{234}{F_{\mathrm{MHz}}} \qquad \textbf{[8.2]}$$

As always, close to the earth's surface, these equations are approximations, and are not to be taken too literally. Some experimentation will probably be necessary to optimize resonance on each band. Also, be aware that the drooping dipoles (*B* and *C* in this case) may act more like an inverted-vee antenna (see Chap. 7) than a straight dipole, so the equation length will be just a few percent too short. In any event, a little spritzing with this antenna will yield results.

Some amateurs build the multiple dipole from four- or five-wire TV rotator lead. That type of wire is used to control antenna rotators, and has either four or five parallel wires in a flat arrangement similar to lamp cord. Cut each wire to the length required for a band, and strip off any unused portions.

Another possibility is the link-tuned dipole shown in Fig. 8-1C. In this situation, a single conductor is used for each half of the dipole, or actually inverted vee. The conductors are broken into segments A, B, and C (or more, if desired). Each segment is separated from the two adjacent sections by inline insulators (standard end insulators are suitable). Segment A is a quarter-wavelength on the highest frequency band of operation, $A + B$ is a quarter-wavelength on the next highest band of operation, and $A + B + C$ is a quarter-wavelength on the lowest frequency band of operation.

The antenna is "tuned" to a specific band by either connecting, or disconnecting, a similar wire (see inset) jumper across the insulator that breaks the connection between the segments. Either a switch or an alligator clip jumper will short out the insulator to effectively lengthen the antenna for a lower band. Some amateurs use single-pole 110-Vac power line switches to jumper the insulator. Although I have not tried this method, it should work.

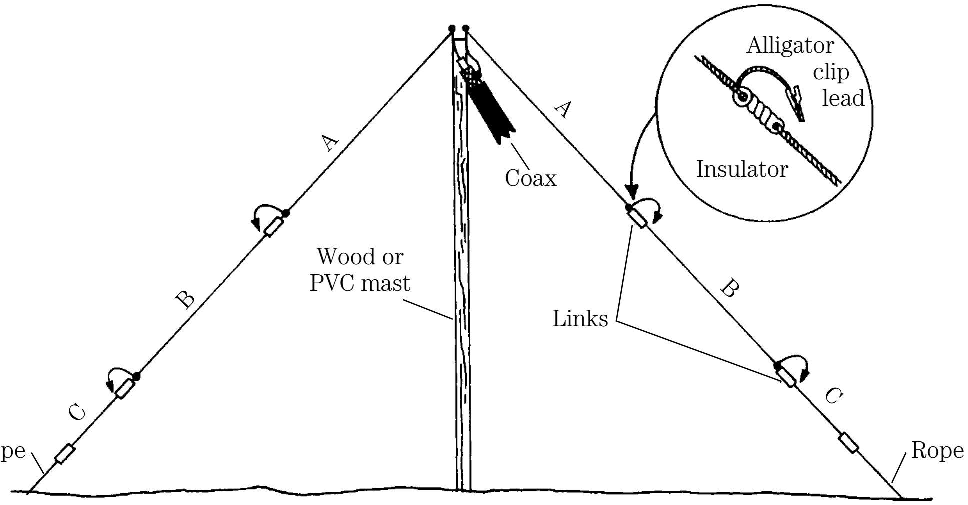

8-1C Multiband inverted vee uses shorting links to change bands.

A big disadvantage to this type of multiband antenna is that you must go out into the yard and manually switch the links to change bands, which probably explains why other antennas are a lot more popular, especially in northern latitudes.

Tuned feeder antennas

Figure 8-2A shows the tuned feeder type of antenna. This antenna can be used from 80 through 10 m, but it requires a special tuner and a length of parallel transmission line. There are two ways to get parallel transmission line: make it or buy it. Using no. 14 or 12 wire, and specially made insulators (also called *spreaders*), you can make 300-, 450-, or 600-Ω parallel transmission line. But that's a pain in the *ptusch* because you can also buy parallel line rather cheaply. I paid $16 for 100 ft of 450-Ω line recently.

One form of parallel line is ordinary TV-type twin lead, which has an impedance of 300 Ω. This line will take up to about 250 W, although some people use it at higher powers (not recommended!). The antenna of Fig. 8-2A uses 450-Ω parallel line. You can buy insulated 450-Ω twin lead (see Fig. 8-2B) that can be handled as easily as TV twin lead—and a lot more easily than open (uninsulated) parallel line.

The G5RV multiband dipole

Figure 8-3 shows the popular G5RV antenna. Although not without some problems, this antenna is very popular. It can be used either as a horizontal dipole, a sloper, or an inverted-vee antenna (which is how I used it). The dipole elements are each 51 ft long. The feedline can be either 300- or 450-Ω twin lead. For 300-Ω cases, use 29 ft of line, and for 450-Ω line, use 34 ft. One end of the parallel transmission line is connected to the antenna, and the other end is connected to a length of 50-Ω coaxial

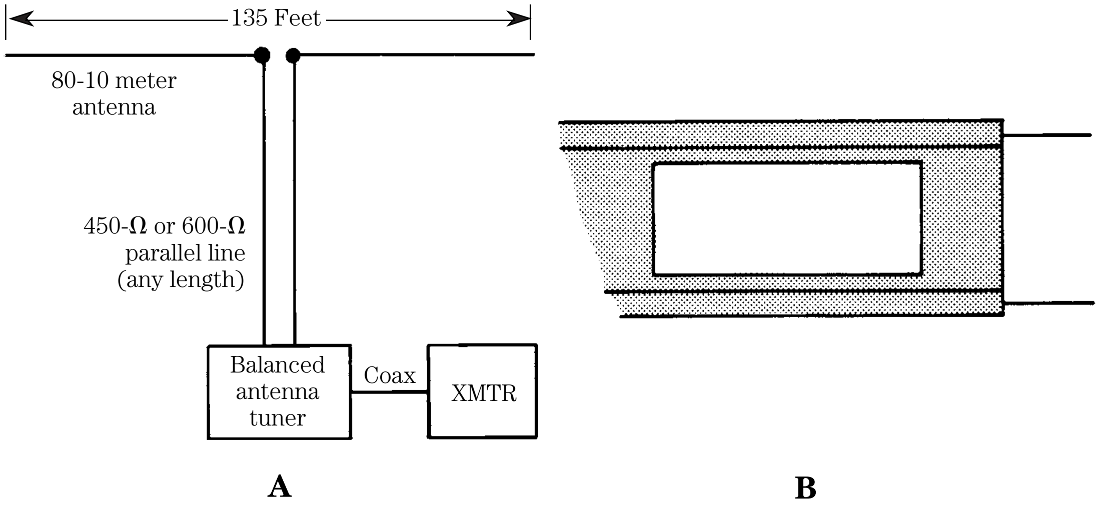
135 Feet
80-10 meter antenna
450-Ω or 600-Ω parallel line (any length)
Balanced antenna tuner
Coax
XMTR
A
B

8-2 Tuned feeder antenna can be used on several bands. (*Courtesy of Hands-On Electronics and Popular Electronics*)

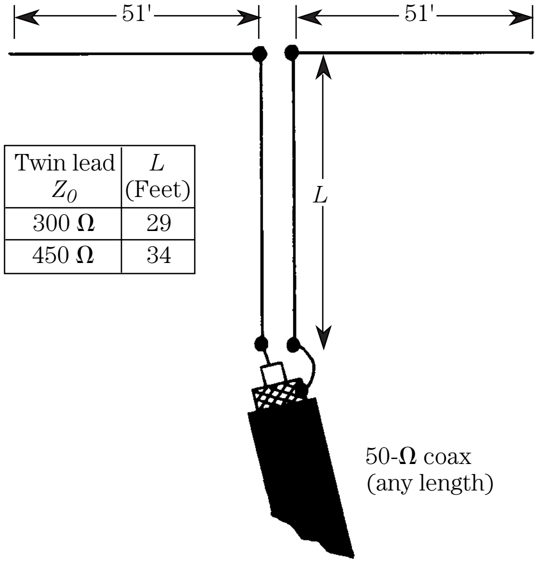
51'
51'
L
50-Ω coax (any length)

Twin lead Z_0	L (Feet)
300 Ω	29
450 Ω	34

8-3 G5RV antenna.
(*Courtesy of Hands-On Electronics and Popular Electronics*)

cable. Although most articles on the G5RV claim that any length of 50-Ω line will work, J. M. Haerle (*HF Antenna Systems: The Easy Way*) recommends that the 50-Ω segment should be at least 65 ft long.

Haerle is a little caustic in his comments on the G5RV, but his criticism is well taken. If you don't have a parallel transmatch, then the G5RV will work (especially if your rig can tolerate a 3:1 VSWR on some selected frequencies). Otherwise, use the antenna of Fig. 8-2A (or a transmatch with the G5RV).

Figure 8-4 shows the once-popular end-fed Zepp antenna. This antenna is a "monopole" in that it uses a half-wavelength radiator, but it is fed at a voltage loop rather than a current loop (i.e., the end of the antenna rather than the center). A 450- or 600-Ω parallel transmission line is used to feed the Zepp antenna. Although the line can theoretically be any length, practicality dictates a quarter-wavelength because of the fact that the antenna is fed at a high RF voltage point.

The Zepp (as shown) is a single-band antenna unless the transmission line is fed with a good, widerange transmatch, or other antenna tuner unit. The antenna tuner required is a balanced type, although a standard transmatch with a 4:1 balun transformer at the output will also work well. Using the Zepp on many bands is easy, but keep in mind that it is voltage-fed, and (at frequencies for which the parallel line is not an odd integer multiple of a quarter-wavelength) there will be a high-voltage node at the transmatch. That raises the possibility of "RF in the shack"—"hot" snippets of RF on the grounded chassis of station equipment (including microphones, where it is uncomfortable when touched to the lips).

Feeding parallel transmission line

Parallel transmission line is *balanced* with respect to ground, but coaxial cable is *unbalanced* to ground (i.e., one side of the coax is usually grounded). As a result, the standard amateur transmitter output will not drive parallel (balanced) feedline properly. You need to do one of two things: (1) buy (or build) an antenna tuner that is balanced on the output and unbalanced on the input or (2) convert a standard "coax-to-coax" transmatch or other form of antenna coupler to "coax-to-balanced" configuration. In some cases you can use a 4:1 balun transformer at the output of the coaxial cable tuner. Keep in mind, however, that some tuner manufacturers recommend against this practice. If you use the standard tee-network (or SPC) transmatch

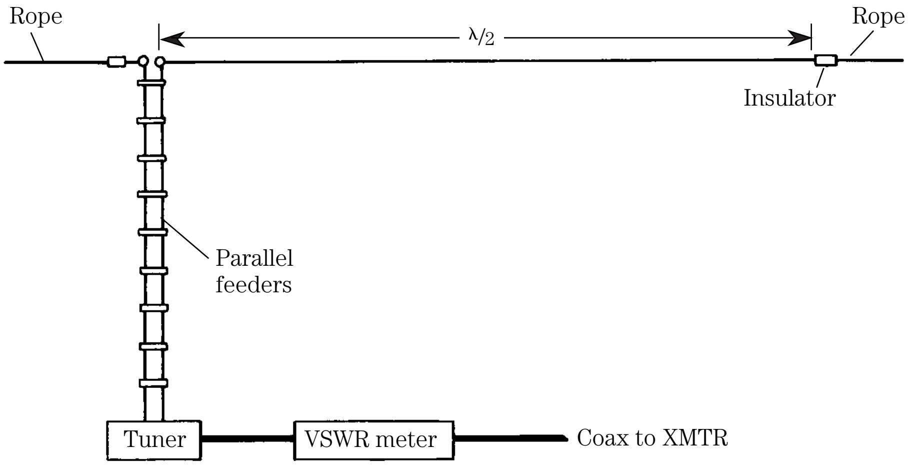

8-4 End-fed Zepp antenna.

then it is possible to make a balun for this purpose from some no.12 formvar- or enamel-covered wire and either a ferrite toroid or a short piece of PVC plumbing pipe. Figure 8-5 shows a balun coil construction project. Use 12 turns of no.12 formvar or enameled wire over a 1-in outside diameter piece of PVC pipe or tubing.

Next, we have the so-called longwire antenna. I use the term *so-called* because not all long wire antennas are truly longwires. A true longwire is longer than two wavelengths at the lowest frequency of operation. In Fig. 8-6 we see a longwire, or “random-length,” antenna fed from a tuning unit. If the antenna length *L* is greater than a quarter-wavelength, then the tuner consists of a single series capacitor (see below); and if it is shorter than a quarter-wavelength, the tuner is a series inductor. The standard tuner for this type of antenna, regardless of length, is a simple L-section coupler (also shown), which can be selected for L-section, series-L or series-C operation.

The Windom antenna (Fig. 8-7) has been popular since the 1920s. Although Loren Windom is credited with the design, there were actually a number of contributors. Coworkers with Windom at the University of Illinois were John Byrne, E. F. Brooke, and W. L. Everett, and they are properly cocredited. The designation of Windom as the inventor was probably due to the publication of the idea (credited to Windom) in the July 1926 issue of *QST* magazine. Additional (later) contributions were rendered by G2BI and GM1IAA (Jim MacIntosh). We will continue the tradition of crediting Loren Windom, with the understanding that others also contributed to this antenna design.

The Windom is a roughly half-wavelength antenna that will also work on even harmonics of the fundamental frequency. The basic premise is that the antenna radiation resistance varies from about 50 Ω to about 5000 Ω, depending upon the

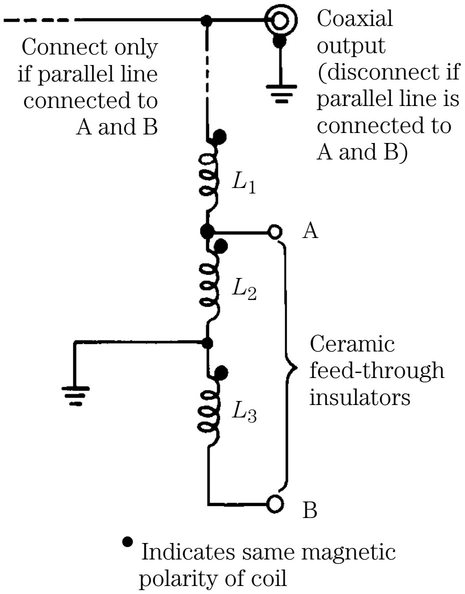

Connect only if parallel line connected to A and B
Coaxial output (disconnect if parallel line is connected to A and B)
L_1
A
L_2
Ceramic feed-through insulators
L_3
B
● Indicates same magnetic polarity of coil

8-5 Balun transformer converts coaxial line to parallel line feed system.

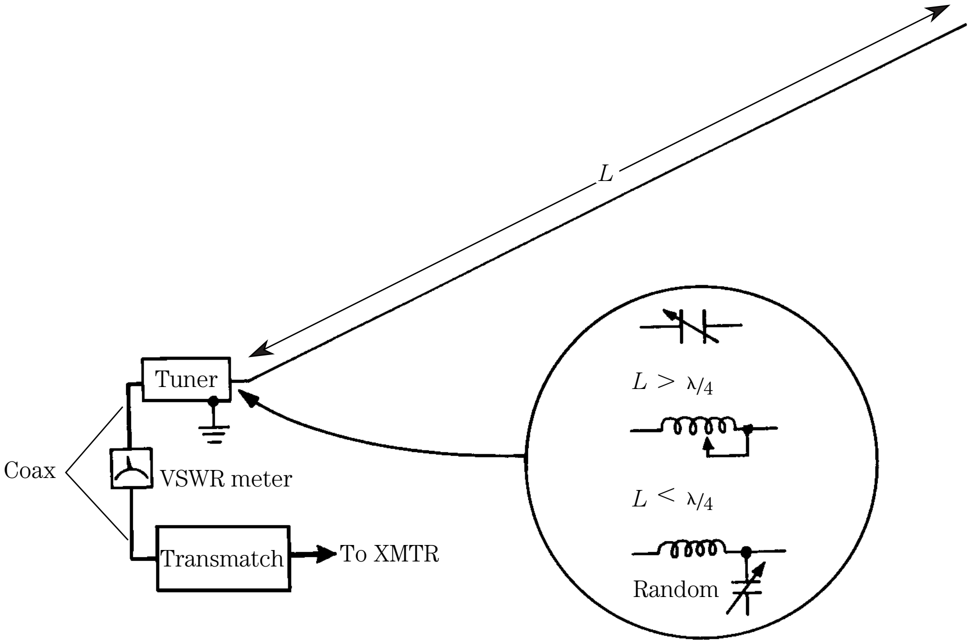

8-6 Random-length (a.k.a., if erroneously, "longwire") antenna.

selected feedpoint. When fed in the exact center, a current loop, the feedpoint impedance will be 50 Ω; similarly, end feeding the antenna finds a feedpoint impedance of about 5000 Ω. In Fig. 8-7A the feedpoint is tapped away from the center at a point that is about one-third the way from one end, at a point where the impedance is about 600 Ω.

The Windom antenna works well—but with some caveats. For example, the antenna has a tendency to put "RF in the shack" because of the fact that it is voltage-fed. Second, there is some radiation loss from the feedline. Finally, the antenna works poorly on odd harmonics of the fundamental frequency.

The antenna tuning unit can be either a parallel resonant, link-coupled *LC* tank circuit (see inset to Fig. 8-7A); or a reversed pi network. In the case of the Windom, the pi network is turned around backward from the usual configuration: C_1 is at the low-impedance end of the network, so it is larger than C_2. Design a pi network (see programs at end of book) to match 50 Ω on the transmitter end and 600 Ω on the antenna end.

A reasonable compromise Windom that reduces feedline radiation losses is shown in Fig. 8-7B. In this antenna a 4:1 balun transformer is placed at the feed point, and this in turn is connected to 75-Ω coaxial transmission line to the transmitter. A transmatch, or similar antenna tuner, is then connected between the transmitter and the transmission line.

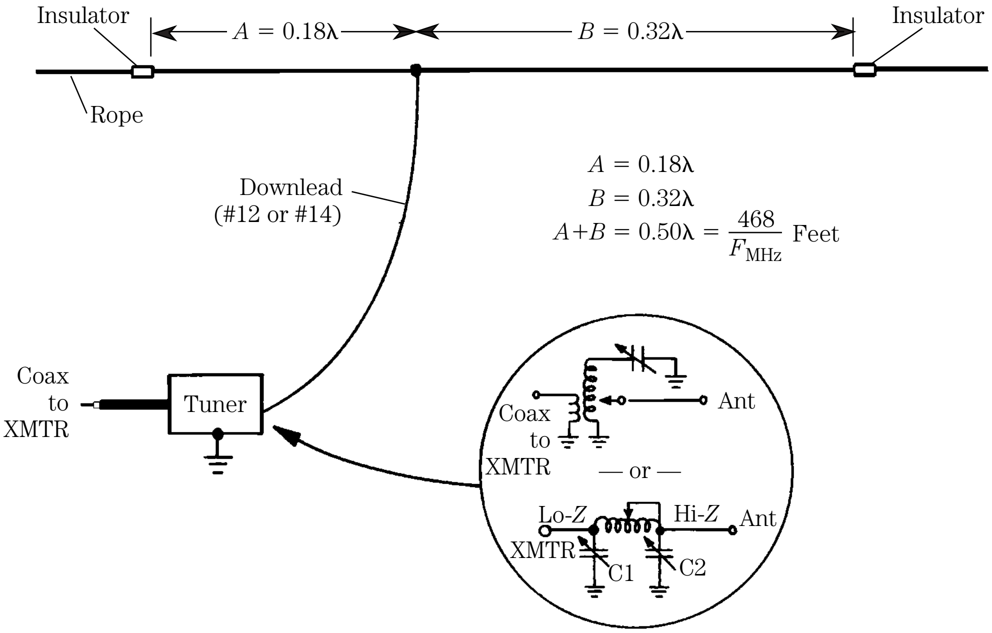
Insulator
A = 0.18λ
B = 0.32λ
Insulator
Rope
Downlead
(#12 or #14)
A = 0.18λ
B = 0.32λ
A+B = 0.50λ = 468/F_MHz Feet
Coax
to
XMTR
Tuner
Coax
to
XMTR
Ant
— or —
Lo-Z
Hi-Z
Ant
XMTR
C1
C2

8-7A Windom antenna.

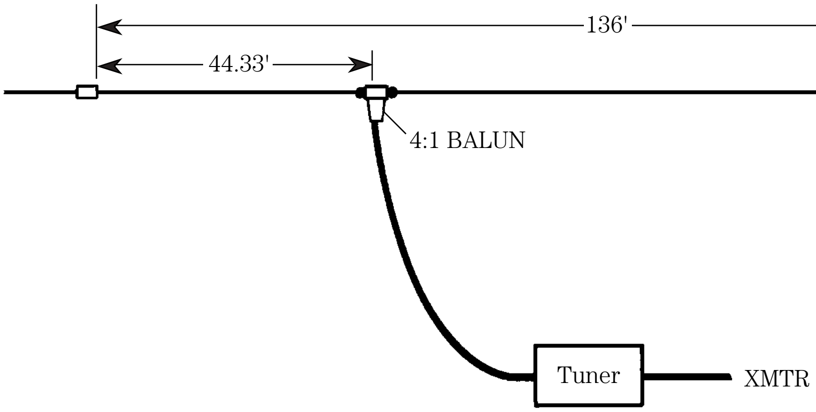
136'
44.33'
4:1 BALUN
Tuner
XMTR

8-7B Coaxial-fed Windom.

This page intentionally left blank

9
CHAPTER
Longwire directional antennas

YOU'VE HEARD IT DOZENS OF TIMES: THE RANDOM-LENGTH LONGWIRE ANTENNA IS THE "perfect" solution to awkward antenna problems. Whether it's a lack of real estate, cranky landlords, or a profound lack of dollars, the random-length longwire will do the job for you. Right? Well, now, that depends on who you ask, and what they did to make it work— or not work. One person says the longwire is not worth a plugged nickel; another is very ho-hum about it because his "kinda works"; still another is enthusiastic because hers is installed correctly and it works better than anything "since sliced pickle and liverwurst sandwiches." Over the years my various living arrangements have forced me to use longwires at many QTHs, as well as on Field Day. Why does the random-length longwire have such a varied reputation? Before we answer that question let's find out (for those who came in late) just what is a longwire antenna.

As an aside, let me point out that random-length wire antennas (less than 2λ) are not true longwire antennas. However, common (if erroneous) usage compels the inclusion of both types.

Longwire antennas

Longwire antennas are any of several types of resonant and nonresonant antennas. Any given longwire antenna may be both resonant and nonresonant, depending upon the operating frequencies used. In the "old days," when I was first starting in amateur radio, most resonant longwires were resonant over all HF bands because those bands were harmonically related to each other. But with the addition of the 10-, 18-, and 24-MHz band segments, that relationship no longer holds true for all cases.

Figure 9-1 shows the classic random length, nonresonant longwire antenna. It consists of a wire radiator that is at least a quarter-wavelength long, but is most often longer than a quarter-wavelength. The specific length is not critical, but it must be *greater than a quarter-wavelength at the lowest frequency of operation anticipated*. If you have a 90-ft wire, it will work on all HF bands above 3.5 MHz. In

Copyright 2001 - the McGraw-Hill Companies

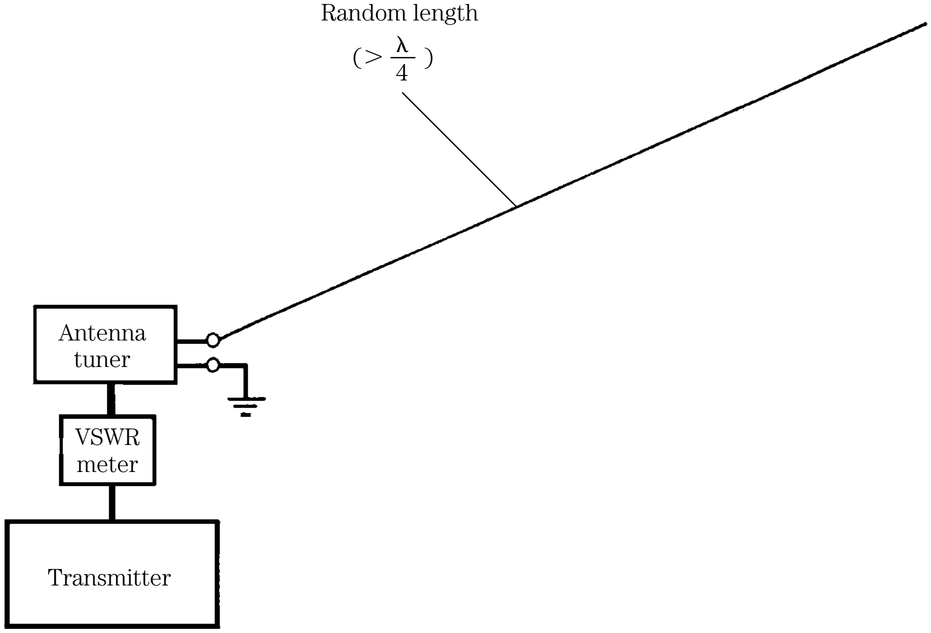

9-1 Random-length antenna.

most installations, the wire is no.12 and no.14 *copperweld*, or hard-drawn copper wire. I have successfully used both no.12 and no.14 house wire, but because it is solid (not stranded) it is not the best material. Stranded wire lasts longer in the wind, because solid wire fatigues and breaks quicker.

The longwire antenna is capable of providing gain over a dipole and a low angle of radiation (which is great for DX operators!). But these advantages are only found when the antenna is several wavelengths long, so it occurs only on typical HF antennas in the 21- through 29-MHz bands.

The longwire is end-fed, and therefore has a high impedance, except in those cases where the radiator happens to be quarter-wavelength. Because of this fact, it is necessary to use an antenna tuner between the low-impedance transmitter output (usually limited to 50- to 75-Ω impedances) and the antenna. You can buy any of several commercial antenna tuners, or make one yourself. Figure 9-2 shows a typical antenna tuner for longwires. It is an L-section coupler consisting of a series inductance and a shunt capacitance, both variable. The inductor can be homemade, or it can be made from a B&W miniductor (3029 is suitable). Alternatively, you can buy a used roller inductor from a hamfest. If you opt for the homemade or miniductor alternatives, then an alligator clip lead (short!) connected to one end of the coil can be used to short out unneeded turns when adjusting the inductance. I prefer the roller inductor method, because it allows the whole “shootin’ match” to be installed inside a shielded cabinet, helping the TVI/BCI (television interference/broadcast interference) situation.

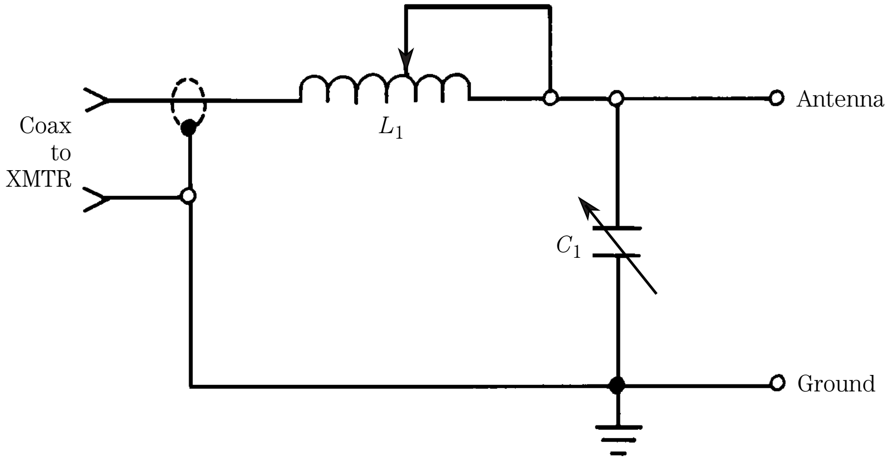

9-2 L-section coupler.

The capacitor (also available from the same sources) should be 150-pF to 250-pF maximum capacitance high-voltage variable. So-called transmitting variables are usually ok if the plate spacing is at least ⅛ in.

Tuning of the longwire is simplicity itself. If the tuner does not have a built-in VSWR meter, then install one in the line between the transmitter output and the tuner input. Adjust both L_1 and C_1 (they are a bit interactive so do it several times) for the lowest VSWR. If you use the alligator clip method, *turn off the darn transmitter before adjusting the clip position—RF burns are nasty!*

So what's the problem?

Ok, so we have decided to install a longwire. How do we make it work? First, make sure that it is long enough. The bare minimum lengths for HF bands are 70 ft for 3.5 MHz and up, 34 ft for 7 MHz and up, and so on. In general, the longer the better. The second problem is a *good ground*. The importance of a good ground cannot be underestimated, and it accounts for about 99 percent of the difference in reported performance of longwires.

A good ground consists of a short wire to either one very long ground rod or multiple ground rods spaced a couple feet apart. An important factor is the length of the ground wire. It must be considerably less than a quarter-wavelength. Use no.10 stranded wire (several parallel lengths) or braid.

When I was in college I lived on the second floor of a whacky student boarding house in Norfolk, VA. The ground was 24 ft away, and that made "short ground wire" a joke—and my longwire nearly inoperative. But I figured out a way around the problem. A quarter-wave radial was dropped out the window (see Fig. 9-3). Use more than one radial if possible. In one case, when the landlord was particularly cranky, I tacked the radial to the baseboard of my room (insulated on standoffs at the far end to protect against the high RF voltages present). It worked well!

A new product also comes to the rescue of those reader-hams who cannot put up a good longwire. The MFJ Enterprises (Box 494, Mississippi State, MS, 39762, 1-800-647-1800) model MFJ-931 artificial RF ground is installed in the ground line (see Fig. 9-4) and is used to tune the ground wire. Adjust the capacitance and inductance controls for maximum ground current as shown on the built-in meter. Be sure to follow the instructions, however, because high RF voltages can appear on a nonresonant ground wire. I wish I'd had one of these when I was in that whacky boarding house; discussing existentialist poetry can't hold a candle to DXing.

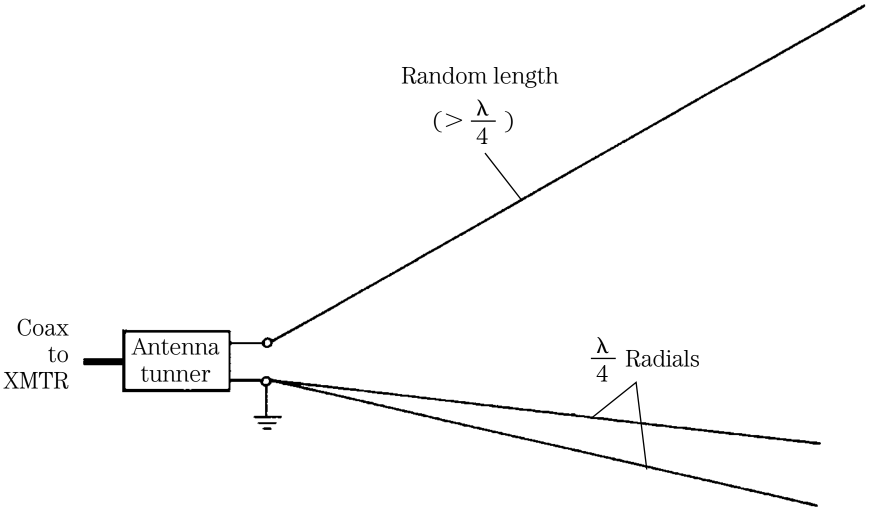
Random length
$(> \frac{\lambda}{4})$
Coax
to
XMTR
Antenna
tunner
$\frac{\lambda}{4}$ Radials

9-3 Radials improve the "ground" of random-length antenna.

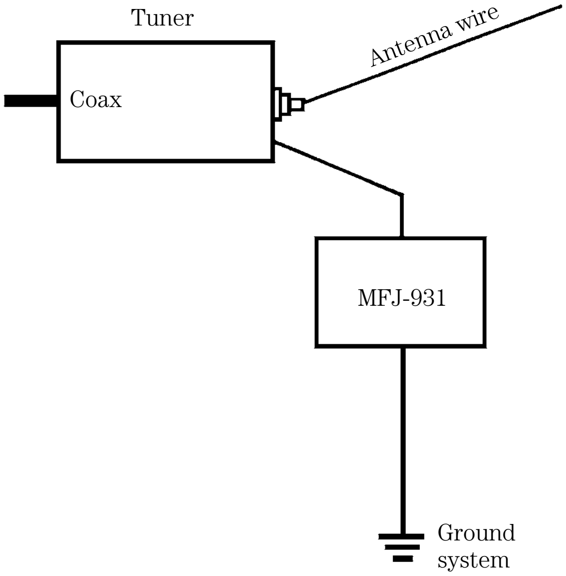
Tuner
Coax
Antenna wire
MFJ-931
Ground
system

9-4 MFJ ground line tuner installation.

When is a long wire a longwire?

For many years, the longwire has been a popular form of antenna. It is cheap, it is easy to construct, and—although reports vary—it has the potential to perform well. Properly constructed, it has a lot of utility. But what is a *real* longwire antenna? In past columns, I have used the term to mean an antenna such as Fig. 9-1. This form of antenna is popularly called a *longwire* if it is more than a quarter-wavelength. I and other authors have used the term longwire to mean this antenna, but that's not rigorously correct. A *true* longwire is an antenna that is many wavelengths long, or to be a little more rigorous, an antenna that is *more than two wavelengths long*. Although I will still use the term *longwire* for both forms of longwire antenna, proper rigor requires Fig. 9-1 to be called a "random-length antenna."

True longwire antennas

Figure 9-5 shows the true resonant longwire antenna. It is a horizontal antenna, and if properly installed, it is not simply attached to a convenient support (as is true with the random length antenna). Rather, the longwire is installed horizontally like a dipole. The ends are supported (dipole-like) from standard end insulators and rope.

The feedpoint of the longwire is one end, so we expect to see a voltage antinode where the feeder is attached. For this reason we do not use coaxial cable, but rather either parallel transmission line (also sometimes called *open-air* line or some such

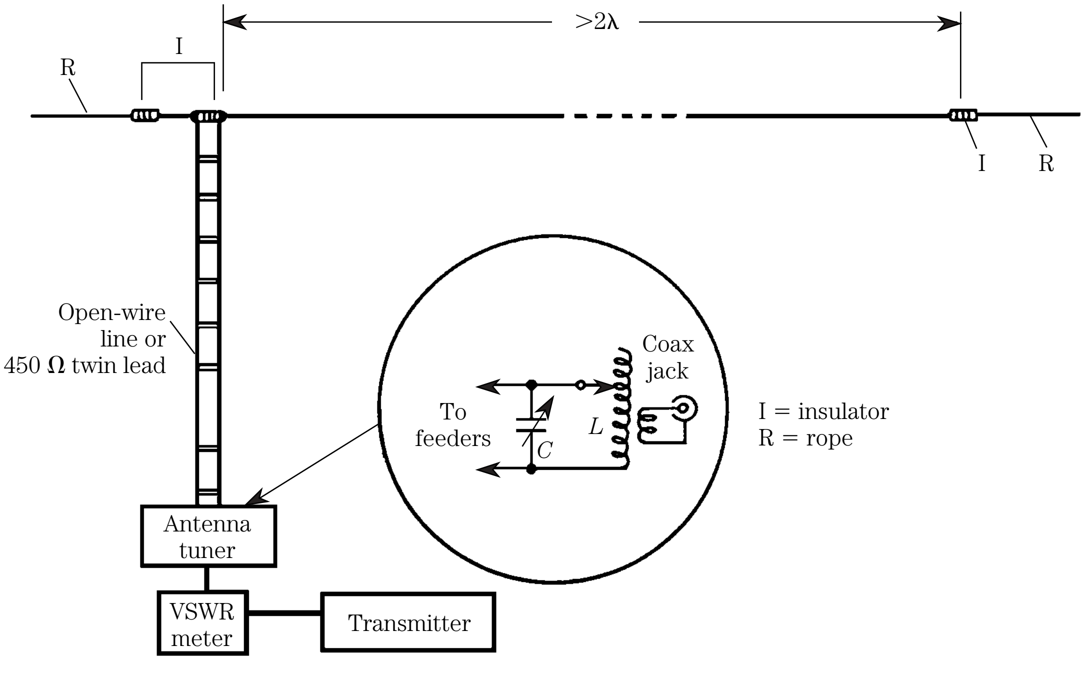

9-5 End-fed longwire with suitable tuner.

name), or 450-Ω twin lead. The transmission line is excited from any of several types of balanced antenna tuning unit (see Fig. 9-5). Alternatively, a standard antenna tuning unit (designed for coaxial cable) can be used if a 4:1 balun transformer is used between the output of the tuner and the input of the feedline.

What does "many wavelengths" mean? That depends upon just what you want the antenna to do. Figure 9-6 shows a fact about the longwire that excites many users of longwires: It has gain! Although a two-wavelength antenna only has a slight gain over a dipole; the longer the antenna, the greater the gain. In fact, it is possible to obtain gain figures greater than a three-element beam using a longwire, but only at nine or ten wavelengths.

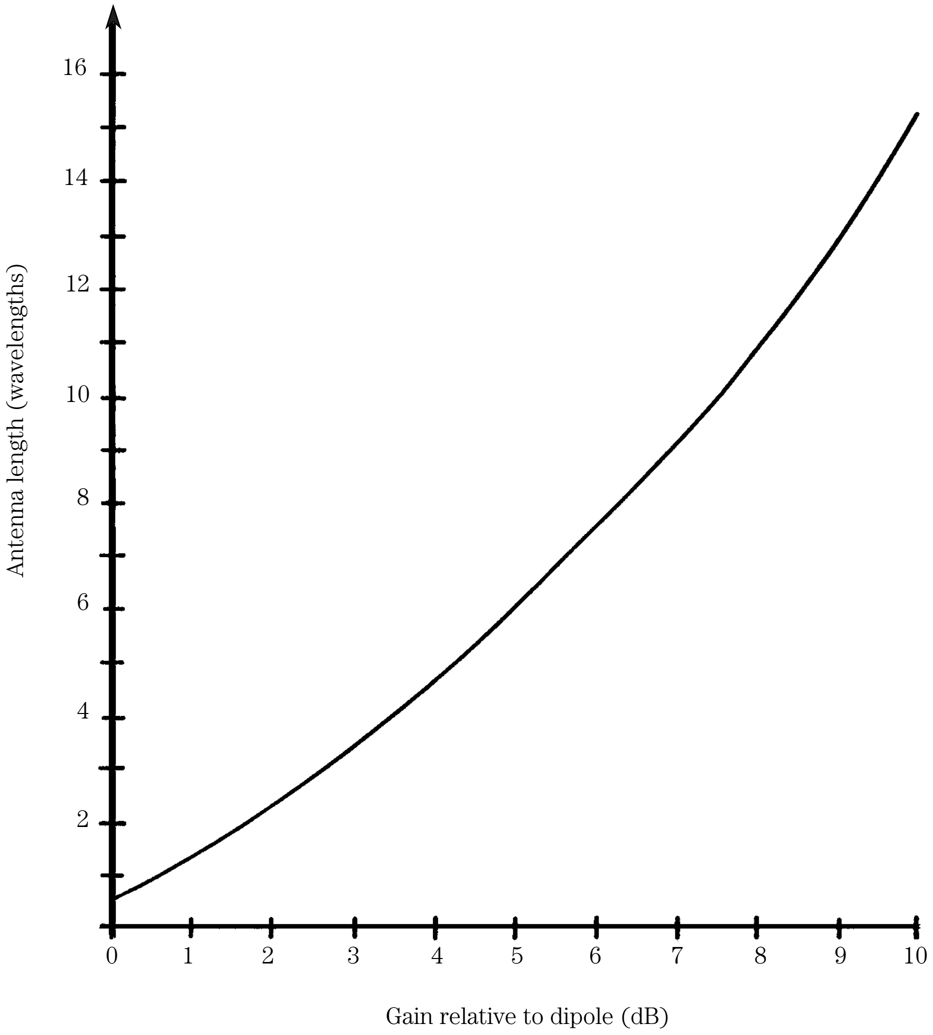

9-6 Antenna length versus gain over dipole.

What does this mean? One wavelength is $984/F_{\mathrm{MHz}}$ ft, so at 10 m (29 MHz) one wavelength is about 34 ft; at 75 m (3.8 MHz) one wavelength is 259 ft long. In order to meet the two-wavelength criterion a 10-m antenna need only be 68 ft long, while a 75-m antenna would be 518 ft long! For a ten-wavelength antenna, therefore, we would need 340 ft for 10 m; and for 75 m, we would need nearly 2,600 ft. Ah me, now you see why the longwire is not more popular. The physical length of a nonterminated resonant longwire is on the order of

$$L_{\mathrm{feet}} = \frac{492(N - 0.025)}{F_{\mathrm{MHz}}} \qquad \textbf{[9.1]}$$

(N is number of half-wavelengths in the radiator elements.)

Of course, there are always people like my buddy (now deceased) John Thorne, K4NFU. He lived near Austin, TX on a multiacre farmette that has a 1400-ft property line along one side. John installed a 1300 ft longwire and found it worked excitingly well. He fed the thing with homebrew 450-Ω parallel (open-air) line and a Matchbox antenna tuner. John's longwire had an extremely low angle of radiation, so he regularly (much to my chagrin on my small suburban lot) worked ZL, VK, and other Southeast Asia and Pacific basin DX, with only 100 W from a Kenwood transceiver.

Oddly enough, John also found a little bitty problem with the longwire that textbooks and articles rarely mention: electrostatic fields build up a high-voltage dc charge on longwire antennas! Thunderstorms as far as 20 mi away produce serious levels of electrostatic fields, and those fields can cause a buildup of electrical charge on the antenna conductor. The electric charge can cause damage to the receiver. John solved the problem by using a resistor at one end to ground. The "resistor" is composed of ten to twenty 10-MΩ resistors at 2 W each. This resistor bleeds off the charge, preventing damage to the receiver.

A common misconception about longwire antennas concerns the normal radiation pattern of these antennas. I have heard amateurs, on the air, claim that the maximum radiation for the longwire is

1. Broadside (i.e., 90°) with respect to the wire run or
2. In line with the wire run

Neither is correct, although ordinary intuition would seem to indicate one or the other. Figure 9-7 shows the approximate radiation pattern of a longwire when viewed from above. There are four main lobes of radiation from the longwire (A, B, C, and D). There are also two or more (in some cases many) minor lobes (E and F) in the antenna pattern. The radiation angle with respect to the wire run (G–H) is a function of the number of wavelengths found along the wire. Also, the number and extent of the minor lobes is also a function of the length of the wire.

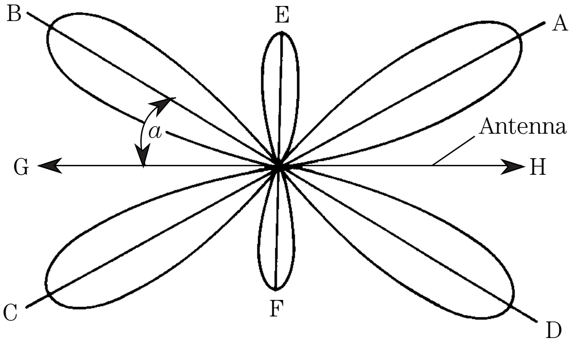

9-7 Radiation pattern of longwire antenna.

Nonresonant single-wire longwire antennas

The resonant longwire antenna is a *standing wave antenna*, because it is unterminated at the far end. A signal propagating from the feedpoint, toward the open end, will be reflected back toward the source when it hits the open end. The interference between the forward and reflected waves sets up stationary standing current and voltage waves along the wire.

A *nonresonant longwire* is terminated at the far end in a resistance equal to its characteristic impedance. Thus, the incident waves are absorbed by the resistor, rather than being reflected. Such an antenna is called a *traveling wave antenna*. Figure 9-8 shows a terminated longwire antenna. The transmitter end is like the feed system for other longwire antennas, but the far end is grounded through a terminating resistor R_1 that has a resistance R equal to the characteristic impedance Z_o of the antenna (i.e., $R = Z_o$). When the wire is 20 to 30 ft above the ground, Z_o is about 500 to 600 Ω.

The radiation pattern for the terminated longwire is a unidirectional version of the multilobed pattern found on the unterminated longwires. The angles of the lobes vary with frequency, even though the pattern remains unidirectional. The directivity of the antenna is partially specified by the angles of the main lobes. It is interesting to note that gain rises almost linearly with $n\lambda$, while the directivity function changes rapidly at shorter lengths (above three or four wavelengths the rate of change diminishes considerably). Thus, when an antenna is cut for a certain low frequency, it will work at higher frequencies, but the directivity characteristic will be different at each end of the spectrum of interest.

A two-wavelength (2λ) pattern is shown in Fig. 9-7. There are four major lobes positioned at angles of ±36° from the longwire. There are also four minor lobes—the strongest of which is –5 dB down from the major lobes—at angles of ±75° from the longwire. Between all of the lobes, there are sharp nulls in which little reception is possible. As the wire length is made longer, the angle of the main lobes pulls in tighter (i.e., toward the wire). As the lobes pull in closer to the wire, the number of minor lobes increases. At 5λ, there are still four main lobes, but they are at angles of ±22° from the wire. Also, the number of minor lobes increases to 16. The minor lobes are located at ±47°, ±62°, ±72°, and ±83° with respect to the wire. The minor lobes tend to be –5 to –10 dB below the major lobes. When the longwire gets very

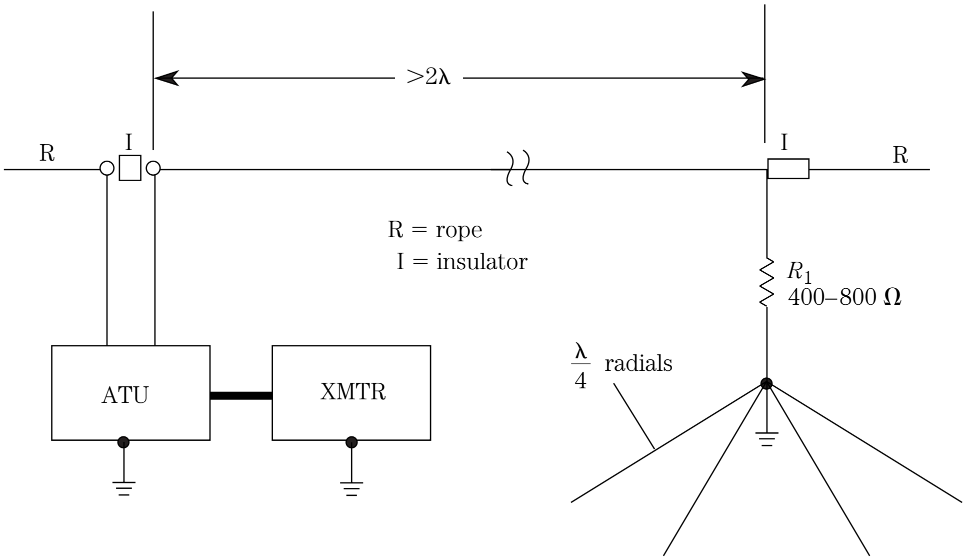

9-8 Longwire (greater than 2λ) antenna is terminated in a resistance (typically 400 to 800 Ω). Radials under the resistor improve the grounding of the antenna.

much longer than 5λ, the four main lobes begin to converge along the length of the wire, and the antenna becomes bidirectional. This effect occurs at physical lengths greater than about 20λ.

In general, the following rules apply to longwire antennas:

- On each side of the antenna, there is at least one lobe, minor or major, for each half-wavelength of the wire element. For the overall element, there is one lobe for every quarter-wavelength.
- If there are an even number of lobes on either side of the antenna wire, then half of the total number of lobes are tilted backward, and half are tilted forward; symmetry is maintained.
- If there are an odd number of lobes on either side of the wire, then one lobe on either side will be perpendicular to the wire, with the other lobes distributed either side of the perpendicular lobe.

Vee beams and rhombic beams

Longwire antennas can be combined in several ways to increase gain and sharpen directivity. Two of the most popular of these are the *vee beam* and the *rhombic* antennas. Both forms can be made in either resonant (unterminated) or nonresonant (terminated) versions.

Vee beams The *vee beam* (Fig. 9-9A) consists of two equal-length longwire elements (wire 1 and wire 2), fed 180° out of phase with each other, and spaced to produce an acute angle between them. The 180° phase difference is inherent in connecting the two wires of the vee to opposite conductors of the same parallel conductor feedline.

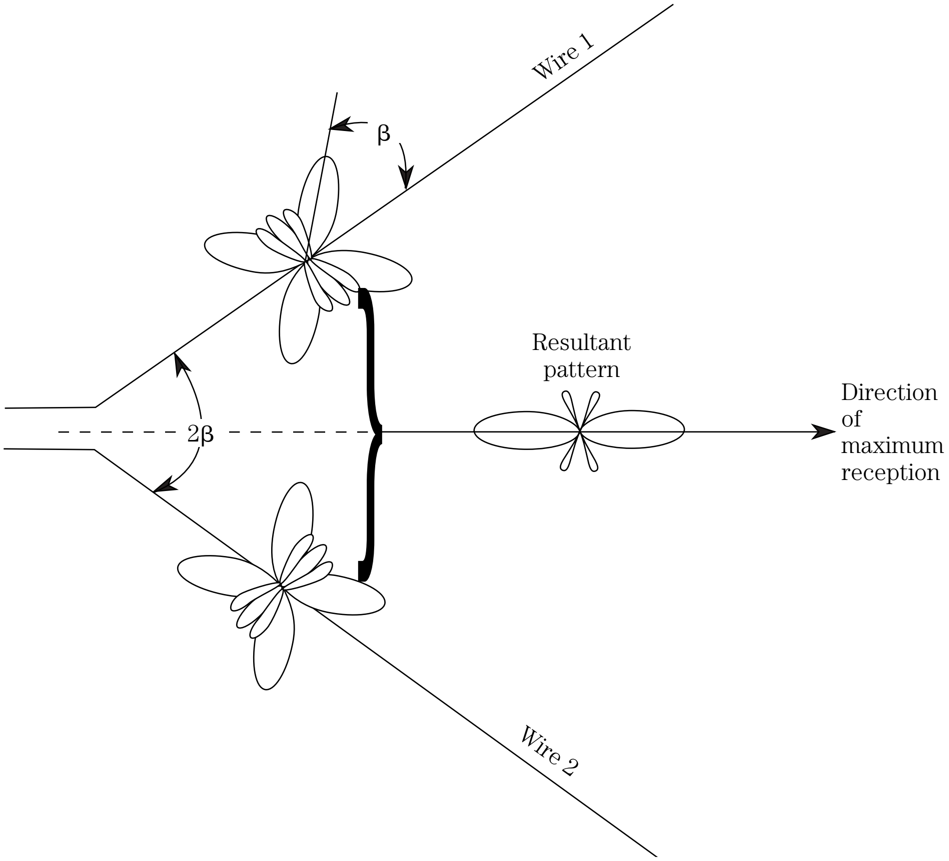

9-9A Radiation pattern of vee beam antenna consists of the algebraic sum of the two longwire patterns that make up the antenna.

The unterminated vee beam (Fig. 9-9A) has a bidirectional pattern that is created by summing together the patterns of the two individual wires. Proper alignment of the main lobes of the two wires requires an included angle, between the wires, of twice the radiation angle of each wire. If the radiation angle of the wire is β, then the appropriate included angle is 2β. To raise the pattern a few degrees, the 2β angle should be slightly less than these values. It is common practice to design a vee beam for a low frequency (e.g., 75-/80- or 40-m bands), and then to use it also on higher frequencies that are harmonics of the minimum design frequency. A typical vee beam works well over a very wide frequency range *only* if the included angle is adjusted to a reasonable compromise. It is common practice to use an included angle that is between 35° and 90°, depending on how many harmonic bands are required.

Vee beam patterns are based on an antenna height that is greater than a half-wavelength from the ground. At low frequencies, such heights will not be practical, and you must expect a certain distortion of the pattern because of ground reflection effects.

Gain on a vee beam antenna is about 3 dB higher than the gain of the single-wire longwire antenna of the same size, and it is considerably higher than the gain of a dipole (see Fig. 9-9B). At three wavelengths, for example, the gain is 7 dB over a dipole. In addition, there may be some extra gain because of mutual impedance effects—which can be about 1 dB at 5λ and 2 dB at 8λ.

Nonresonant vee beams Like the single-wire longwire antennas, the vee beam can be made nonresonant by terminating each wire in a resistance that is equal to the antenna's characteristic impedance (Fig. 9-10). Although the regular vee is a standing wave antenna, the terminated version is a traveling wave antenna and is thus unidirectional. Traveling wave antennas achieve unidirectionality because the terminating resistor absorbs the incident wave after it has propagated to the end of the wire. In a standing-wave antenna, that energy is reflected backwards toward the source, so it can radiate oppositely from the incident wave.

Rhombic beams The *rhombic beam antenna*, also called the *double vee*, consists of two vee beams positioned end-to-end with the tips connected. The unidirectional, nonresonant (terminated) rhombic is shown in Fig. 9-11. The unterminated resonant form gives approximately the same gain and directivity as a vee beam of the same size. The nonresonant rhombic has a gain of about 3 dB over a vee beam of the same size (see Fig. 9-9 again).

Two angles are present on the rhombic antenna. One-half of the included angle of the two legs of one wire is the *tilt angle* (ϕ), while the angle between the two wires is the *apex angle* (θ). A common rhombic design uses a tilt angle of 70°, a length of 6λ for each leg (two legs per side), and a height above the ground of 1.1λ.

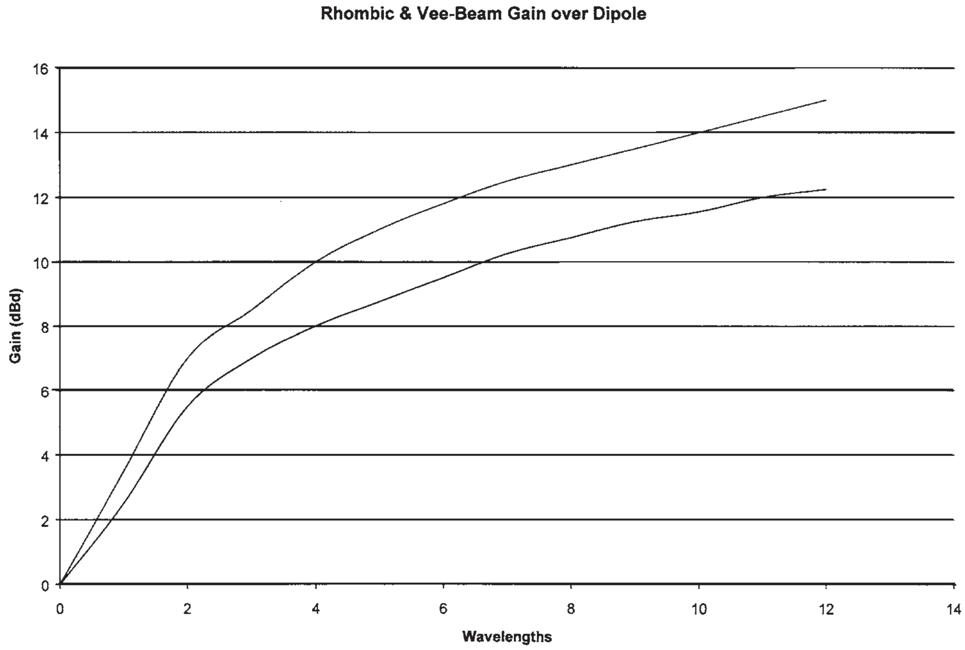

9-9B Gain versus length of vee beam and rhombic beam antennas.

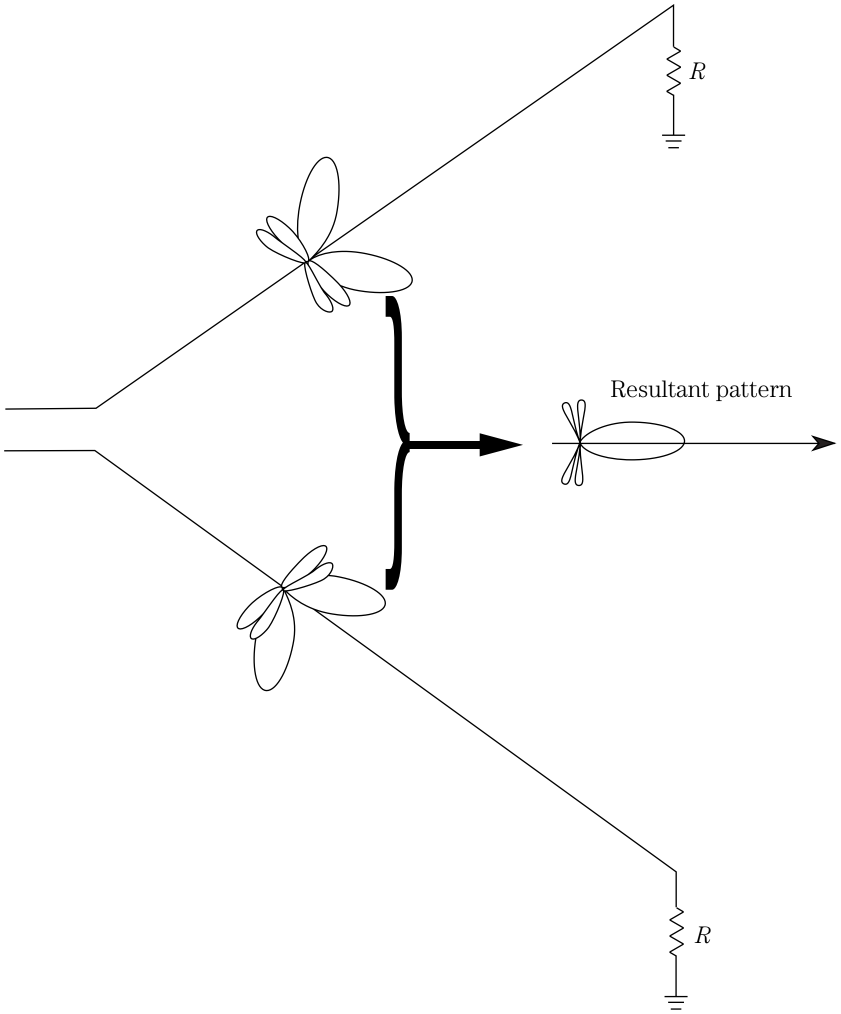

9-10 Nonresonant vee beam antenna.

The termination resistance for the nonresonant rhombic is 600 to 800 **Ω**, and must be noninductive. For transmitting rhombics, the resistor should be capable of dissipating at least one-third of the average power of the transmitter. For receive-only rhombics, the termination resistor can be a 2-W carbon composition or metal-film type. Such an antenna works nicely over an octave (2:1) frequency range.

Beverage or "wave" antennas

The *Beverage* or *wave antenna* is considered by many people to be the best receive antenna available for very low frequency (VLF), AM broadcast band (BCB), medium

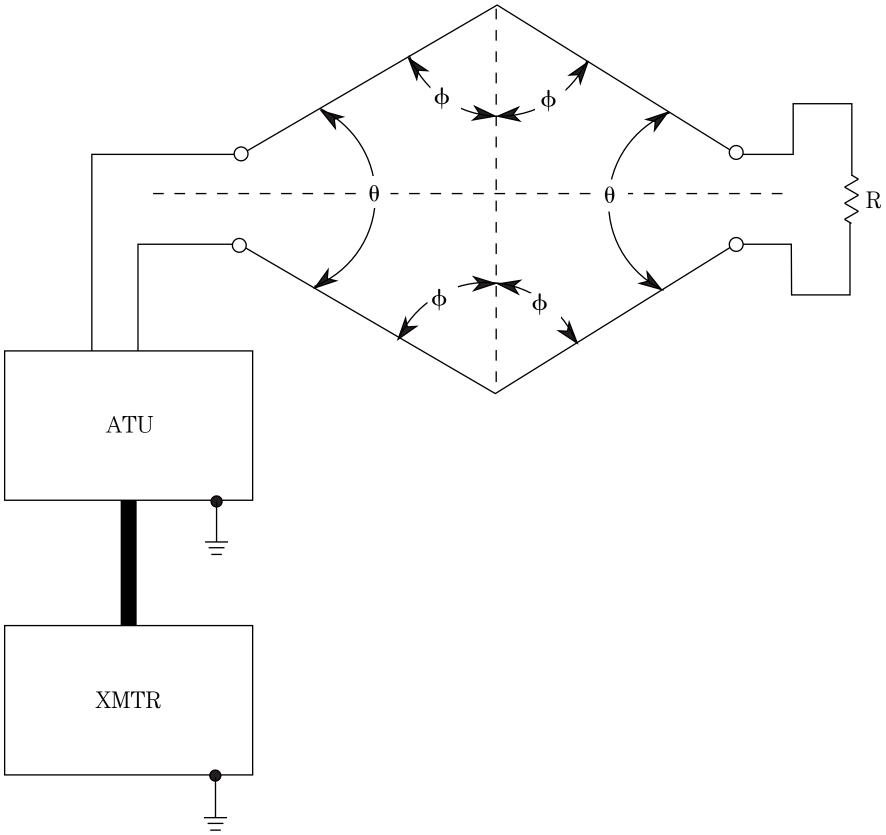

9-11 Terminated rhombic antenna.

wave (MW), or tropical band (low HF region) DXing. The Beverage was used by RCA at its Riverhead, Long Island (NY) station in 1922, and a technical description by Dr. H. H. Beverage (for whom it is named) appeared in *QST* magazine for November 1922, in an article titled “The Wave Antenna for 200-Meter Reception.” In 1984, an edited and updated version of the 1922 article appeared in the same magazine. In 1921, Paul Godley, under sponsorship of the American Radio Relay League, journeyed to Scotland to erect a receiving station at Androssan. His mission was to listen for amateur radio signals from North America. As a result of politicking in the post–World War I era, hams were consigned to the supposedly useless shortwave ($\lambda < 200$ m), and it was not clear that reliable international communications were possible. Godley went to Scotland to see if that could happen; he reportedly used a wave antenna for the task (today, called the *Beverage*).

The Beverage antenna is a longwire of special design, more than one wavelength (1λ) long (Fig. 9-12), although some authorities maintain that $>0.5\lambda$ is minimally sufficient. The Beverage provides good directivity and good gain, but is not very efficient.

As a result, it is preferred for receiving, and is less useful for transmitting. This is an example of how different attributes of various antennas make the Law of Reciprocity an unreliable sole guide to antenna selection. Unlike the regular longwire, which is of a different design, the Beverage is intended to be mounted close to the earth's surface (typically $< 0.1\lambda$); heights of 8 to 10 feet is the usual prescription.

Figure 9-12 shows the basic single-wire Beverage antenna. It consists of a single conductor (no.16 to no.8 wire, with no.14 being most common) erected about 8 to 10 ft above ground. Some Beverages are unterminated (and bidirectional), but most of them are terminated at the far end in a resistance R equal to the antenna's characteristic impedance Z_o. The receiver end is also terminated in its characteristic impedance, but generally requires an impedance matching transformer to reduce the antenna impedance to the 50-Ω standard impedance used by most modern transmitters.

The Beverage works best in the low-frequency bands (VLF through MW), although at least some results are reportedly relatively easy to obtain up to 25 m (11.5 MHz). Some questionably successful attempts have been made at making Beverage antennas work as high as the 11-m Citizens Band or the 10-m ham band (29.7 MHz).

The Beverage antenna works on vertically polarized waves arriving at low angles of incidence. These conditions are normal in the AM BCB, where nearly all transmitting antennas are vertically polarized. In addition, the ground- and sky-wave propagations found in these bands (VLF, BCB, and low HF) are relatively consistent. As the frequency increases, however, two factors become increasingly dominant. First, the likelihood of horizontal polarization increases because of the size of a wavelength at those frequencies. The polarization of the received signal not only changes in those bands, but does so constantly when conditions are unsettled. It is the strong dependence of the

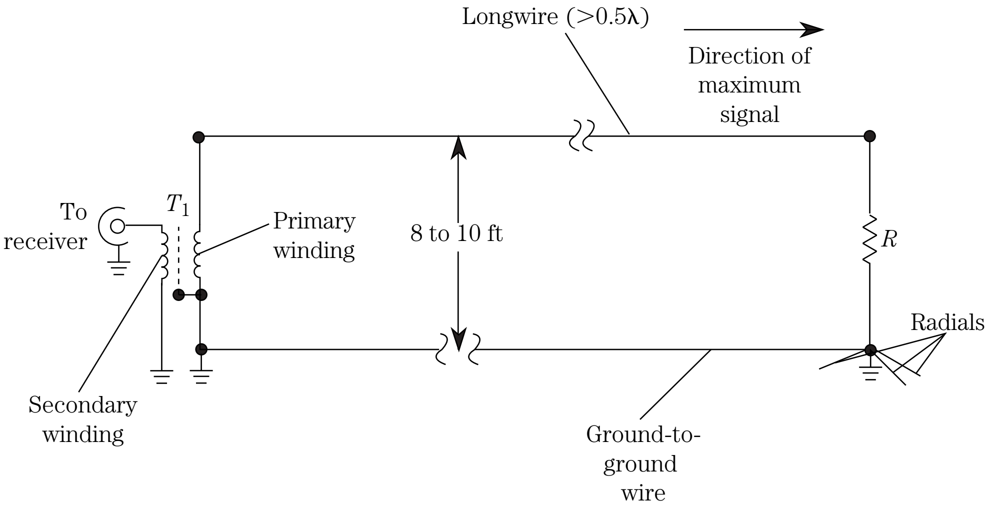

9-12 Beverage antenna.

Beverage on relatively constant vertical polarization that makes me suspect the claims of Beverage-like performance above the 25- or even 31-m bands.

The Beverage depends on being erected over poorly conductive soil, even though the terminating resistor needs a good ground. Thus, one source claimed that sand beaches adjacent to salty marshes make the best Beverage sites (a bit of an overstatement). Figure 9-13 shows why poorly conductive soil is needed. The E-field vectors are launched from the transmitting antenna perpendicular to the earth's surface. Over perfectly conducting soil, the vertical waves would remain vertical. But over imperfectly conducting soil, the field lines tend to bend close to the point of contact with the ground. As shown in the inset of Fig. 9-13, the bending of the wave provides a horizontal component of the E-field vector, and this provides the means of generating an RF current in the conductor wire.

A debate among Beverage antenna fans regards the best length for the antenna. Some sources state that the length can be anything greater than or equal to 0.5λ, yet others say greater than or equal to1λ is the minimum size. One camp says that the length should be as long as possible, while others say it should be close to a factor called the *maximum effective length* (MEL), which is

$$\mathrm{MEL} = \frac{\lambda}{4\left(\frac{100}{K} - 1\right)} \qquad \textbf{[9.2]}$$

where

MEL is the maximum effective length, in meters
λ is the wavelength, in meters
K is the velocity factor, expressed as a percent

Misek, who may well be the leading exponent of the Beverage antenna, uses numbers like 1.6λ to 1.7λ over the 1.8- to 7.3-MHz region, and 0.53λ to 0.56λ on frequencies lower than 1.8 MHz. Doctor Beverage was once quoted as saying that the optimum length is 1λ.

Like the longwire antenna, the Beverage needs a termination resistor that is connected to a good ground. This requirement might be harder to meet on Beverage antennas because they work best over lossy ground, which doesn't make a very good ground connection. As in the longwire case, insulated or bare wires, a quarter-wavelength long, make the best radials. However, a substantial improvement in the ground is possible using just bare wires measuring 15 to 20 ft long (which is much less than λ/4), buried in the soil just below the surface (far enough to prevent erosion from bringing it to the surface). Many articles and books on Beverages show ground rods of 2 or 3 ft long, which borders on the ridiculous. Poor soil requires longer ground rods, on the order of 6 to 8 ft. Copper-clad steel makes the best rods.

In addition to the radials and ground rod, Misek also recommends using a wire connection between the ground connection at the termination resistor, and the

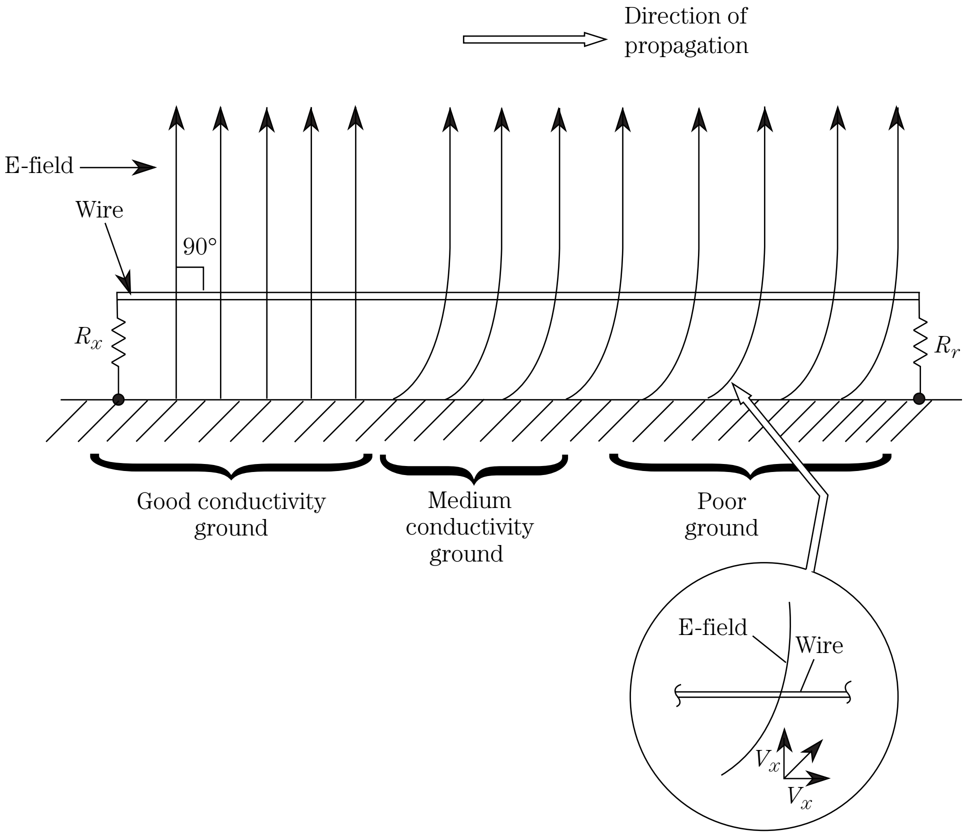

9-13 Performance of Beverage antenna over different values of ground conductance.

ground connection at the receiver transformer (see again Fig. 9-12). According to Misek, this wire helps to stabilize the impedance variations at higher frequencies.

Installation of the Beverage antenna is not overly critical if certain rules are followed. The antenna should be installed at a height of 6 to 10 ft off the ground, and it should be level with the ground over its entire length. If the ground is not flat enough to make a level installation possible, then try to use a height that is 6 to 10 ft above the average terrain elevation along its run. A popular installation method is to erect 16-ft 4 × 4 lumber, such that 3 to 4 ft are buried in a concrete-filled posthole. Use lumber that is treated for outdoor use (i.e., pressure-treated lumber sold for decks and porches. The wire can be fastened to the 4 × 4 posts using either ceramic standoff ("beehive") insulators or electric fence insulators (which some people deem preferable). Try to use one contiguous length of wire for the antenna, if possible, in order to avoid soldered splices and joints.

One of the Beverage installation difficulties shared with the longwire is the need to slope down to a point where a termination resistor can be easily installed close to the ground. While the longwire can be sloped over a large portion of its length, the Beverage should be sloped downward only over the last 60 ft or so.

Steerable notch Beverage antennas

A Beverage erected with two wires—parallel to each other, at the same height, spaced about 12 in apart (Fig. 9-14), with a length that is a multiple of a half-wavelength—is capable of *null steering*. That is, the rear null in the pattern can be steered over a range of 40° to 60°. This feature allows strong, off-axis signals to be reduced in amplitude so that weaker signals in the main lobe of the pattern can be received. There are at least two varieties of the *steerable wave Beverage* (SWB).

If null steering behavior is desired, then a phase control circuit (PCC) will be required—consisting of a potentiometer, an inductance, and a variable capacitor in series with each other. Varying both the "pot" and the capacitor will steer the null. You can select the direction of reception, and the direction of the null, by using a switch to swap the receiver and the PCC between port A and port B.

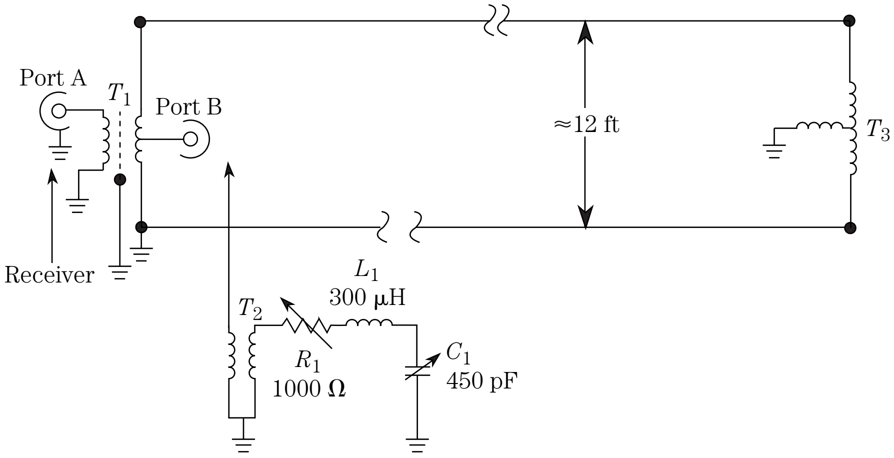

9-14 Beverage antenna with a steerable null.

This page intentionally left blank

10
CHAPTER

Hidden and limited-space antennas

ONE OF THE MOST SIGNIFICANT IMPEDIMENTS TO AMATEUR RADIO OPERATORS, CB operators, and shortwave listeners is the space available for their antennas. In many thousands of other cases, the limitation is less one of space, but of regulators. More and more subdivisions are built with covenants on the deed that prohibit the buyer from installing outdoor antennas. Once limited to townhouse developments, where that breed of contemptible vermin (called the Homeowners Committee) routinely intruded on the affairs of people who mistakenly think they "own" their townhouse (ownership implies right of use, which is limited by the covenants). These onerous covenants are now routinely placed on single-family dwellings as well. In fact, it is the single most serious threat to amateur communications people in the country today. Other homeowners are no longer angered by the restriction on antennas because their television reception is now carried to them via cable systems in most parts of the country. Even where cable is not available, most users can install a moderately sized television antenna in their attic, or use rabbit ears. In other cases, the townhouse community will install a single master TV antenna and then distribute signals to each unit. The result is that the amateur radio operator, CB operator, and shortwave listener are left to fend for themselves without assistance from neighbors.

In this chapter we will examine some of the alternatives available to those readers who have either a limited space situation (such as a small city lot), or are unable to move out of a subdivision where there are stupid rules against outdoor antennas. The suggestions contained in this chapter are not universal, and indeed the author recommends that you adapt, as well as adopt, these recommendations, and come up with some of your own. Creativity within the constraints of the laws of physics governing radio antennas is encouraged.

Copyright 2001 - the McGraw-Hill Companies

Hidden antennas

A hidden antenna is one that is either completely shielded from view or disguised as something else. Alternatively, you could also include in this category antennas that are in semiopen view, but which are not too obvious (except to the trained and diligent eye). Some people have opted for "hidden" longwires made of very fine wire (no. 26 enameled wire is popular). The user will install the wire in the open, high off of the ground (as in an apartment installation), and operate without anyone knowing the difference. One chap used no. 22 wire suspended between two 16-story apartment house buildings that were 100 yards apart. He had one of the best working longwires in town until a windy day when the whole thing came down. No one was injured; but had someone been hurt, there might have been a lawsuit. Hidden antennas must be designed with an eye toward causing others no harm. It is neither ethical nor smart to place others at risk in enjoying our hobbies.

The dipole is a popular antenna with both shortwave listeners and amateur radio operators. Indeed, for the CB operator who wants to get on the air from an apartment or townhouse (or restricted single-family home), the dipole can represent a respectable alternative.

As you learned in Chap. 6, the dipole is a horizontal wire (or pipe) antenna that is a half-wavelength long and fed in the center (ideally) with 75-Ω coaxial cable. In the "townhouse" dipole, it is possible to build the antenna entirely inside the attic of the building. The length of the dipole is given approximately by

$$L_{\mathrm{ft}} = \frac{468}{F_{\mathrm{MHz}}} \qquad \textbf{[10.1]}$$

If you do some quick calculations you will find that antennas for the 10-, 13-, 15-, and (possibly) even 18- and 20-m bands, will fit entirely inside the typical townhouse attic. This statement is also true of the 11-m citizen's band antenna: It will fit inside the standard townhouse attic. But what about the lower frequency bands?

Figure 10-1 shows a possible solution to the use of the lower frequencies in the townhouse situation. The two quarter-wavelength arms of the dipole are ideally installed in line with each other, as was shown in the chapter on horizontal antennas. But in a sticky situation we can also install the dipole with the arms bent to accommodate the space available. In this example, only one of many possible methods for accomplishing this job is shown. Here each quarter-wavelength section is composed of two legs, *AB* and *CD*, respectively. Ideally, segments *B* and *C* are the longest dimensions. Also, if possible, make segments *A* and *D* equal lengths. Another method is to reverse the direction of one end leg, say for example *D*, and run it to the other corner of the building over the peak of the roof.

The author is almost hesitant to offer only one drawing, despite space constraints, in the fear that readers will take the offered pattern as the only authorized version, or solution. In reality, you might not be able (for a variety of reasons) to use the exact pattern shown (so ad lib a little bit).

How about performance? Will the constrained dipole of Fig. 10-1 work as well as a regular dipole installed a wavelength or two off the ground and away from objects?

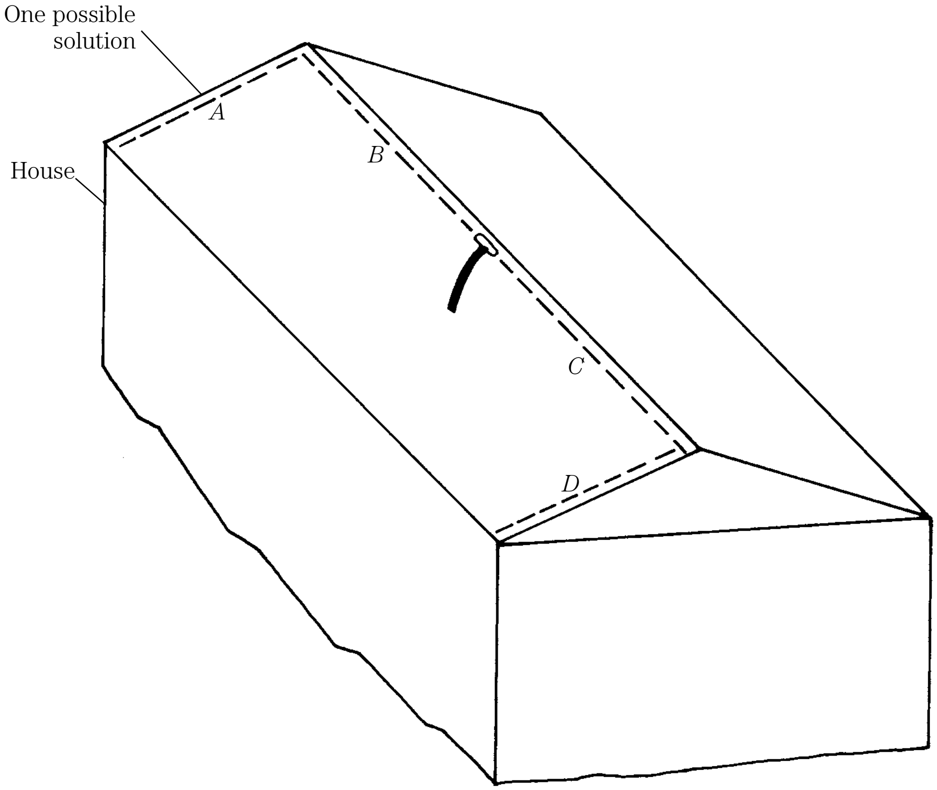

10-1 Installing dipole in attic.

In a word: no. But that is not the problem being solved; getting on the air at all is the present problem. You will find that the pattern of the constrained dipole is distorted compared with that of the regular dipole. In addition, the feedpoint impedance is not going to be 73 Ω (except if by fluke), so you will be required to use an antenna tuner of some kind. You are well advised to read the chapter on antenna instruments and measurements so that you can be prepared to figure out any feed problems that crop up—as well they might.

The wire used in the constrained dipole (or other forms of attic antenna) should be mounted on TV-type screw-in standoff insulators. Almost any outlet that sells TV antennas, or installation parts, will have them. These standoff insulators are also available in many hardware stores, and department stores that sell TV antennas or "Harry and Harriet Homeowner" supplies, in addition to electronics parts suppliers. Do not simply tape the wire to the wooden underside of your roof. The reason is simple: in a poorly tuned antenna, voltages can get high enough at the ends to produce *corona* effects; arcing could be a fire hazard. Also, use insulated wire to avoid accidental contact in the event that someone is in the attic when you are operating. Be aware that the insulation might affect the propagation velocity of signals in the wire, so it will slightly alter the required length.

Another alternative for the attic antenna, and one that avoids some of the problems of corona effects, is the nonresonant loop shown in Fig. 10-2. Although specified as a top view, the loop can be installed in the configuration that best uses the space. In fact, the best performance will be bidirectional when the loop is installed vertically. In this case, a large loop (as large as can be accommodated in the space available) is installed in the attic. Again, use standoff insulators and insulated wire for the installation.

The giant loop is fed with parallel line, and is tuned with a balanced antenna tuning unit. As was true with the constrained dipole, the performance is not to be equated with the performance of more regular antennas; but, with the prospect of not being on the air at all . . . ! Once again we have a compromise antenna for a compromising situation.

Another ploy is the old “flagpole trick” shown in Fig. 10-3. Some developments allow homeowners to vent their patriotism by installing flagpoles—and flagpoles can be disguised antennas. In the most obvious case, you can install a brass or aluminum flagpole and feed it directly from an antenna tuning unit. For single-band operation, especially on the higher frequencies, you can delta-feed the “flagpole” unobtrusively, and call your flagpole a vertical antenna. But that is not always the best solution.

Figure 10-3 shows two methods for creating a flagpole antenna, and both depend on using white PVC plumbing pipe as the pole. The heavier grades of PVC pipe

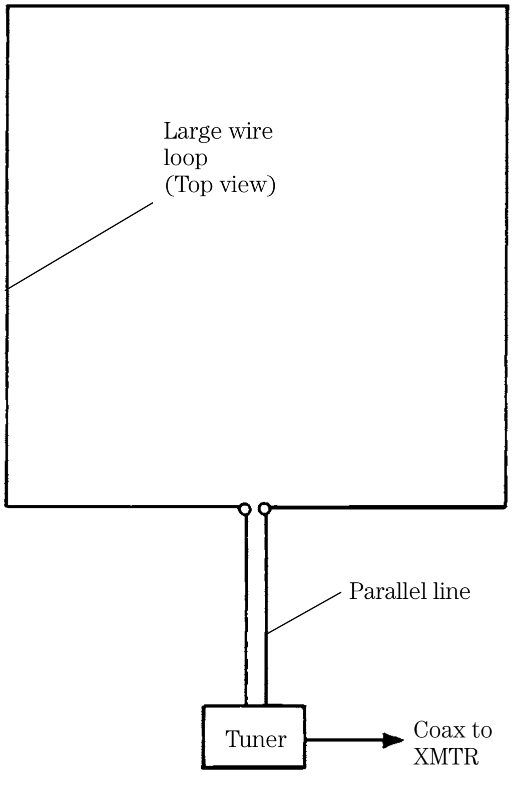

10-2 Large loop antenna.

are self-supporting to heights of 16 to 20 ft, although lighter grades are not self-supporting at all (hence are not usable).

Figure 10-3A shows the use of a PVC flagpole in which a no. 12 (or no. 14) wire is hidden inside. This wire is the antenna radiator. For some frequencies, the wire will be resonant, and for others, it will surely be nonresonant. Because of this problem an antenna tuning unit is used either at the base of the antenna or inside at the transmitter.

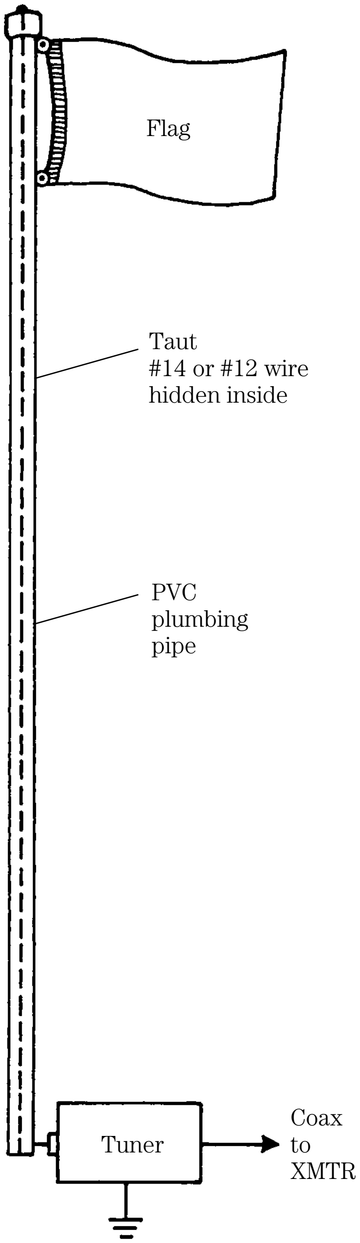
Flag
Taut
#14 or #12 wire
hidden inside
PVC
plumbing
pipe
Tuner
Coax
to
XMTR

10-3A Flagpole antenna made from thick-walled PVC pipe.

If the wire is too long for resonance (as might happen in the higher bands), then place a capacitor in series with the wire. Various settings may be required, so use a multisection transmitting variable that has a total capacitance selectable to more than 1000 pF. Alternatively, use a vacuum variable capacitor of the same range.

In cases where the antenna is too short for resonance, as will occur in the lower bands, insert an inductance in series with the line to “lengthen” it. Another alternative is to use an L-section tuner at the feedpoint.

A good compromise situation is the use of a 16-ft length of “flagpole” pipe with a 16-foot wire embedded inside. The 16-foot wire is resonant at 20 m, so it will perform like a vertical antenna at those frequencies. The tuner will then accommodate frequencies above and below 20 m.

Another alternative is the version shown in Fig. 10-3B. In this case the wire radiator is replaced with a section of aluminum tubing. A wooden or plastic insert is fashioned with a drill and file to support the aluminum tubing inside the PVC tubing. One way to make the support is to use a core bit in an electric drill to cut out a disk that fits snugly inside the PVC tubing. Rat out the center hole, left by the core bit pilot, to the outside diameter of the aluminum tubing. The support can be held in place with screws from the outside, or simply glued in place.

The problem of operating with a hidden antenna is a serious challenge. But with some of these guidelines and a little creativity, you can get on the air and enjoy your amateur radio hobby.

Limited-space antennas

Many people live in situations where it is permissible to install an outdoor antenna, but it is not practical to install a full-size antenna. I once lived in a house that was 16 ft wide and 37 ft long, on a 33 × 100-ft lot. Very few full-sized antennas could be installed on that lot because of the space constraints. Beam antennas were out because county laws required that the antenna not hang over the property line. Although vertical antennas were possible, there was a period of time when I used other antennas that were easier to install. In this chapter, you will examine some of the options open to those with limited space for amateur radio, CB, or SWL antennas.

Once again we return the simple dipole as the basis for our discussion. In Fig. 10-4, you see several alternatives for installing an outdoor antenna in a limited space. In Fig. 10-4A, the slanted dipole (or slipole) antenna uses the standard dipole configuration, but one end is connected to the high point of the building, while the other is anchored near the ground. The coaxial cable is connected to the midpoint of the antenna in the usual manner for regular dipoles. If the end of the dipole is within reach of people on the ground, then they may get a nasty RF burn if the antenna is touched while you are operating. Take precautions to keep people and pets away from this lower end.

Another method is the vee dipole shown in Fig. 10-4B. In a regular dipole installation, the ends of the antenna are along the same axis (in other words, at an angle of 180°). In the example of Fig. 10-4B, however, the angle between the elements is less than 180°, but greater than 90°. In some cases we might want to bend the elements, rather than install them in a vee shape. Figure 10-4C shows an angled dipole with four segments. For the best performance (but not as good as that

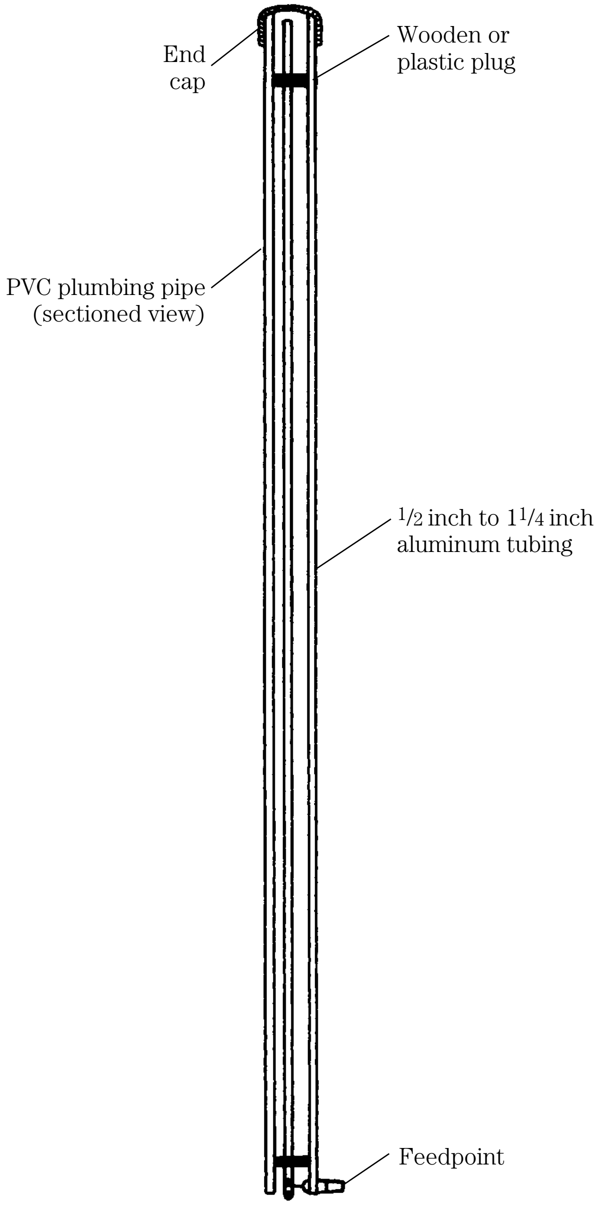
End cap
Wooden or plastic plug
PVC plumbing pipe (sectioned view)
1/2 inch to 1 1/4 inch aluminum tubing
Feedpoint

10-3B PVC flagpole hiding an aluminum tubing vertical.

of a regular dipole), make $A = D$ and $B = C$. In all three examples, Figs. 10-4A through 10-4C, you can expect to find the length needed for resonance varies somewhat from the standard $468/F_{\mathrm{MHz}}$ value, and that the feedpoint impedance is other than 73 Ω. Also, the pattern will be distorted with respect to the regular dipole. Although these antennas do not work as well as a properly installed dipole, the performance is sufficient to allow successful operation.

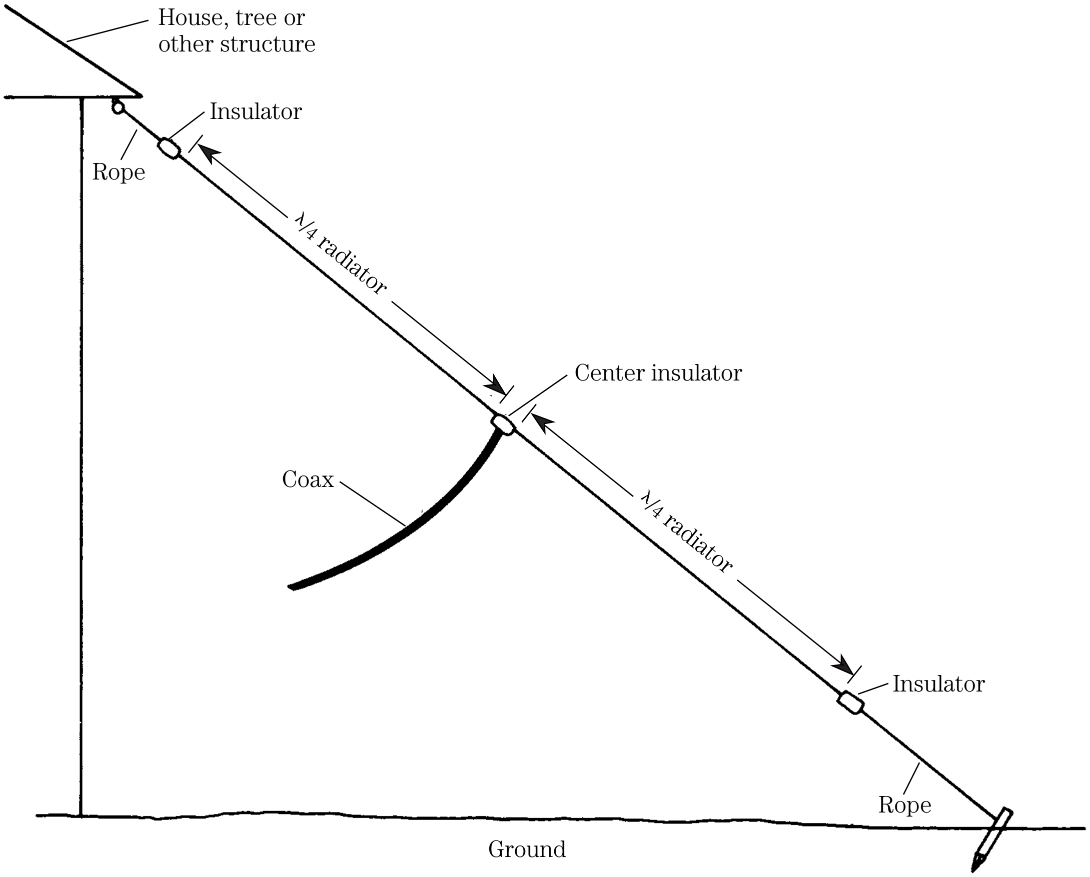

10-4A Sloper dipole.

Another limited-space wire antenna is the (so-called) half-slope shown in Fig. 10-4D. Although single-band versions are often seen, the example in Fig. 10-4D is a multiband version. Resonant traps separate the different band segments. This antenna operates like the vertical, but it is not omnidirectional. Also, the feedpoint impedance will be different from the regular vertical situation. If the impedance varies too much for comfort, insert an antenna tuner, such as the Transmatch, in the coaxial line between the transmitter and the antenna.

Figure 10-5 shows another antenna that's useful for limited-space situations. Although it is easily constructed from low-cost materials, the antenna is also sold by several companies under various rubrics including “cliff-dweller,” “apartment house,” “townhouse,” or “traveler's” antennas. The antenna consists of a 4- to 16-ft section of aluminum or copper tubing. Some of the commercial antennas use a telescoping tubing that can be carried easily in luggage. As was true with the longwire, the window-sill antenna is tuned to resonance with an L-section coupler.

The L-section coupler must be tuned to produce the lowest possible VSWR, so either an RF power meter or a VSWR meter must be installed in the coaxial cable between the transmitter and the tuner. Again, a good ground (or radial system) will

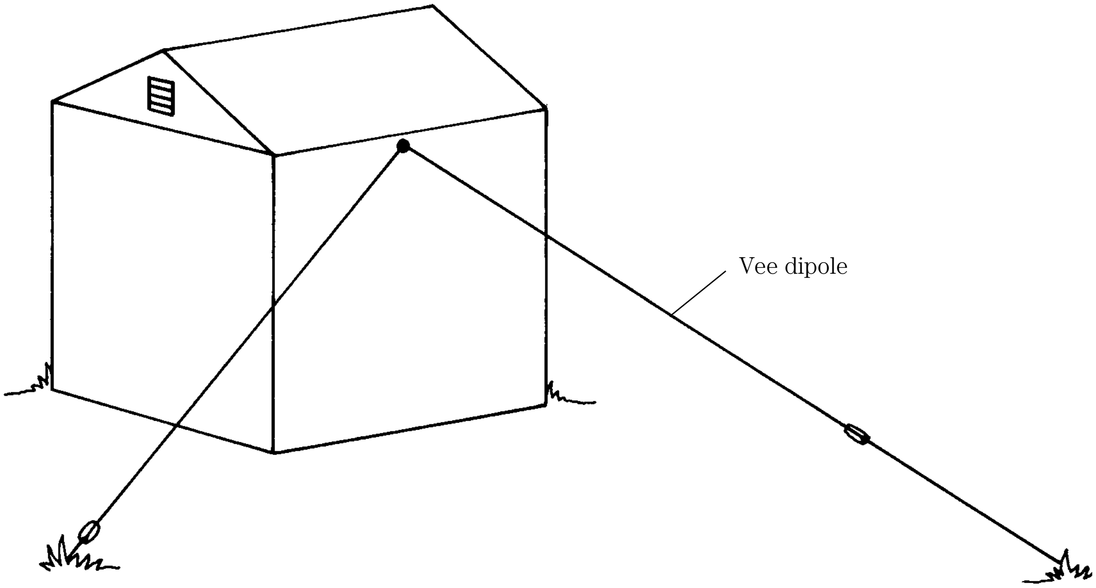

10-4B Vee dipole.

greatly improve the operation of this antenna. The performance should be considerably less than that of a good longwire, but it will work DX for you even on lower frequencies. Once again it is not the best antenna made, but it will get you on the air.

Figure 10-6 shows the use of a mobile antenna for a fixed or portable location. I used this type of antenna at one QTH to good effect. Mount a mobile antenna, such as the Hustler, on the window sill, or other convenient mount (at least one amateur radio operator uses one mounted on the roof of the house). Grounding is essential for this antenna, as was also true for the longwire and window-sill antenna (Fig. 10-5). The operation of this antenna is improved by installing at least two radials per band as a counterpoise ground. My installation worked well, even though only two radials were present.

A directional *rotatable* dipole is shown in Fig. 10-7. This antenna is made from a pair of mobile antennas connected “back-to-back” on a horizontal length of 1 × 2 lumber and fed in the center with a coaxial cable. The end pieces of the mobile antenna set the resonance, and you must recognize that adjusting one requires a countervailing adjustment of the other as well.

Two examples of *helically* wound antennas are shown in Fig. 10-8. In this type of antenna, an insulating mast is wound with a half-wavelength of antenna wire. The overall length of the antenna is considerably less than a half-wavelength, except at the highest frequencies. In order to dissipate the high voltages that tend to build up at the ends of these antennas, a capacitance hat is used. These “hats” are either disks (pie tins work well) or rods of conductor about 16 to 24 in long. The version shown in Fig. 10-8A is a vertical antenna, and like other verticals, it must be installed over either a good ground or a counterpoise ground. The version shown in Fig. 10-8B is horizontally polarized.

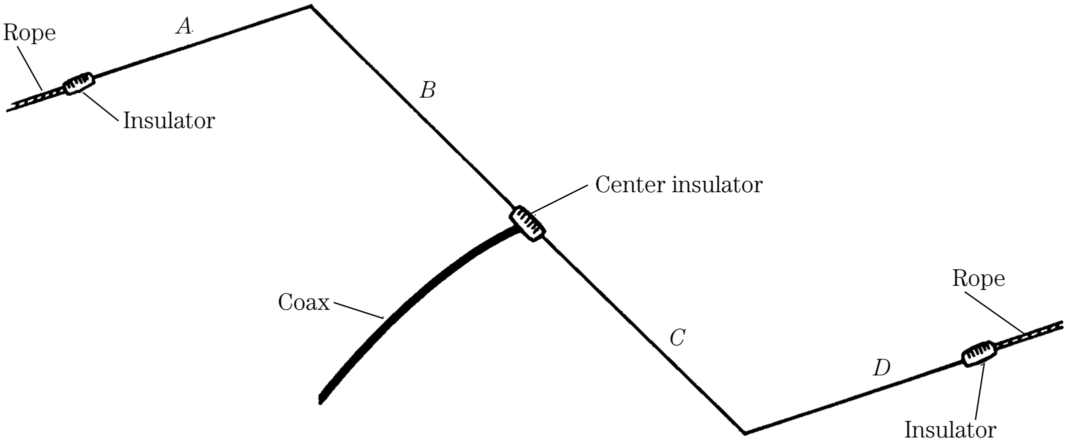
Rope
A
B
Insulator
Center insulator
Coax
Rope
C
D
Insulator

10-4C Bent dipole.

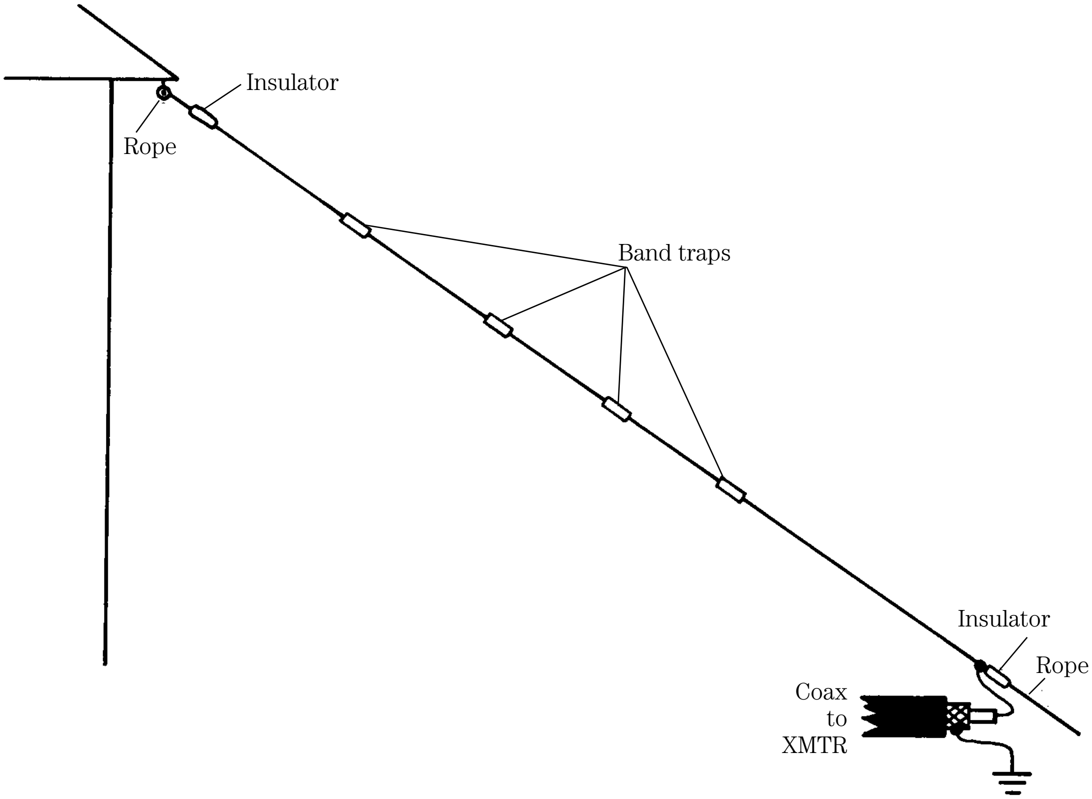
Insulator
Rope
Band traps
Insulator
Rope
Coax
to
XMTR

10-4D Sloper trap vertical.

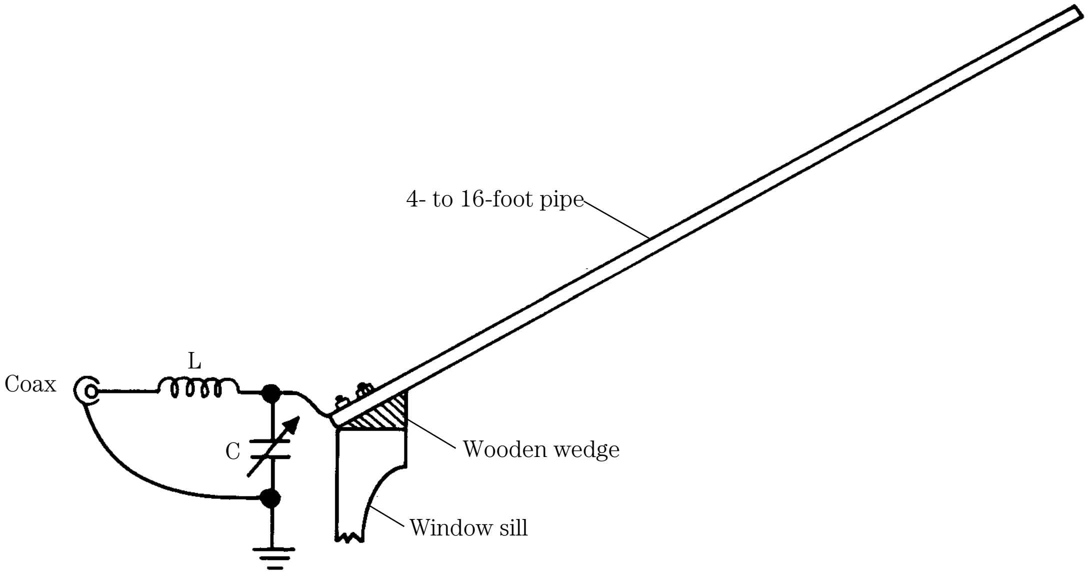
4- to 16-foot pipe
L
Coax
C
Wooden wedge
Window sill

10-5 Window sill antenna.

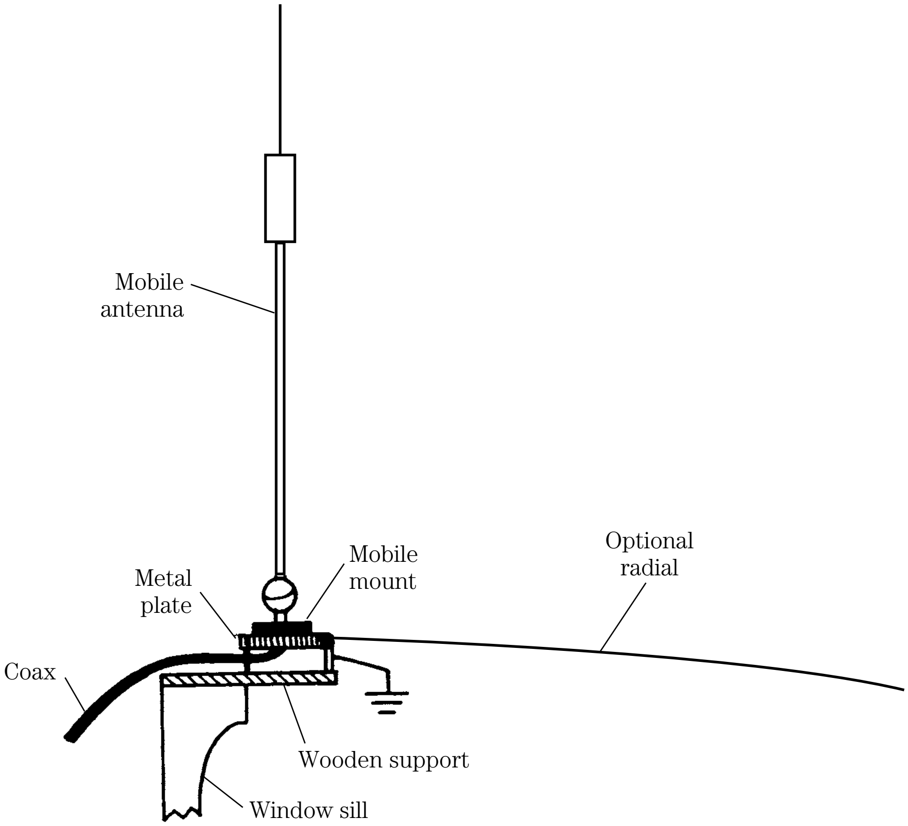
Mobile
antenna
Metal
plate
Mobile
mount
Optional
radial
Coax
Wooden support
Window sill

10-6 Use of mobile antenna as a window sill antenna.

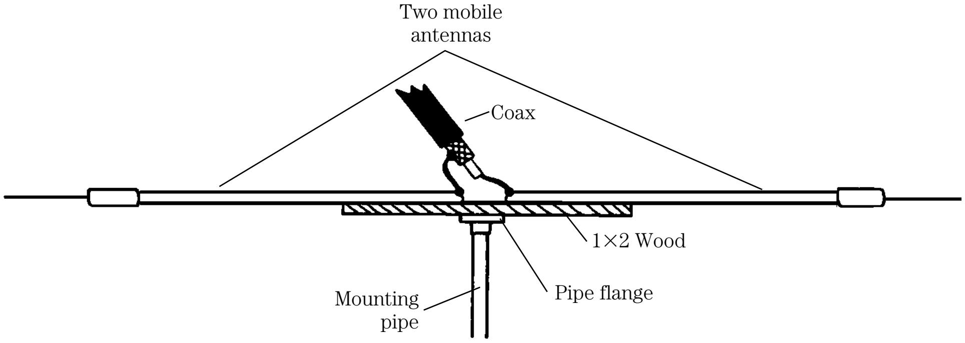
Two mobile antennas
Coax
1×2 Wood
Pipe flange
Mounting pipe

10-7 Use of two mobile antennas as a rotatable dipole.

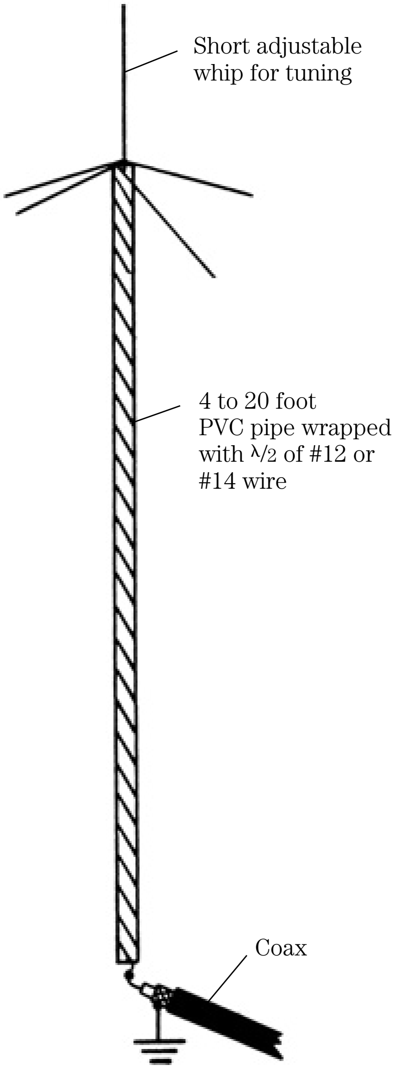
Short adjustable whip for tuning
4 to 20 foot PVC pipe wrapped with λ/2 of #12 or #14 wire
Coax

10-8A Helically vertical antenna.

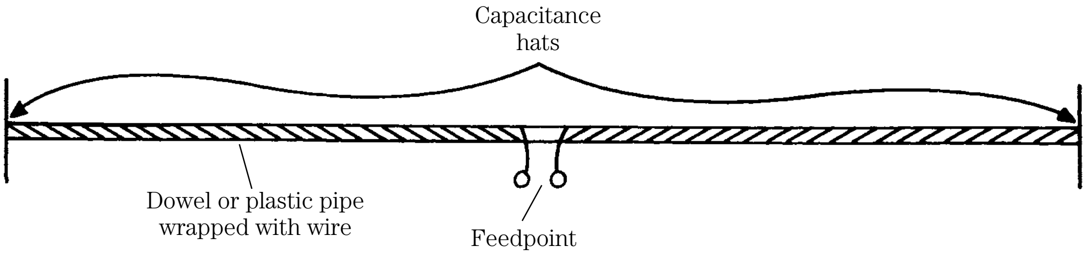
Capacitance hats
Dowel or plastic pipe wrapped with wire
Feedpoint

10-8B Helically wound dipole antenna.

This page intentionally left blank

11
CHAPTER

Directional phased vertical antennas

THE VERTICAL ANTENNA IS A PERENNIAL FAVORITE WITH RADIO COMMUNICATIONS users. The vertical is either praised, or cursed, depending upon the luck of the owner. "DXability" is usually the criterion for judging the antenna's quality. Some amateurs can't get out of their backyards with a vertical, and they let everyone within earshot know that such and such a brand is no good. Yet, another person routinely works New Zealand or Australia on 15 m using exactly the same brand of vertical. The proper installation of vertical antennas is dealt with in another chapter, so, for the present, let's look at another problem attributed to vertical antennas.

That problem is that vertical antennas are *omnidirectional* in the azimuth aspect; that is, they send out and receive equally well from all directions. Some people moan that this pattern dissipates their power, and gives them a weaker signal "out where it counts" (true). However, the main disadvantage of the omnidirectional pattern is noise (QRN and QRM). "QRN" is natural noise from thunderstorms and other sources. "QRM" is man-made noise, and can consist of other stations or the many assorted forms of electrical filth that pollute the airwaves. All forms of noise, however, have one thing in common: they are directional with respect to the station. In other words, if you could null signals coming from the direction of the noise source (or undesired station), you would be able to hear desired stations much better. A directional antenna performs this task, so let's look at some vertically polarized directional antennas.

Although most amateurs seem to think that the effective radiated power (ERP) increase that the directional antenna gives them is the real reason to own one, the main benefit is actually on receive. Think about it for a moment. With anywhere from 100 to 1500 W available, the increase or decrease in signal strength (due to the directivity of the antenna) results in a minimal difference on the receive end, especially during good DX conditions. If we rotate the directional pattern, to null out interference, then we usually find that the change in our signal strength perceived by the other guy is small; the S meter reading of the desired station is minimally affected; but the amplitude of the interference source is greatly attenuated! The overall effect is an apparent increase

Copyright 2001 - the McGraw-Hill Companies

in the other guy's signal, even though the S meter tells a slightly different story. The improvement of signal-to-noise ratio (SNR) is tremendously improved.

Directivity and phasing

So, how does a vertical antenna owner get the benefit of directivity without the kilobuck investment that a beam or quad costs? The usual solution is to use phased verticals. AM broadcast stations, with more than one tower, are using this type of system (although for different reasons than hams). The idea is to place two or more antennas in close proximity and feed them at specific phase angles to produce a desired radiation pattern. A lot of material is available in the literature on phased vertical antenna systems, and it is far too much to be reproduced here. There are "standard patterns" dating from before World War II that are created with different spacings and different phase angles of feed current. In this chapter, we will consider only one system.

Figure 11-1 shows the patterns for a pair of quarter-wavelength vertical antennas spaced a half-wavelength (180°) apart. Without getting into complex phase shifting networks, there are basically two phasings that are easily obtained: 0° (antennas in phase) and 180° (antennas out of phase with each other).

When the two antennas (*A* and *B*) are fed in phase with equal currents (Fig. 11-1A), the radiation pattern (shown somewhat idealized here) is a bidirectional figure 8 that is directionally perpendicular to the line of centers between the two antennas; this pattern is called a *broadside* pattern. A sharp null exists along the line of centers (*A*-*B*).

When the antennas are fed out of phase with each other by 180° (Fig. 11-1B), the pattern rotates 90° (a quarter way around the compass) and now exhibits directivity along the line of the centers (*A*-*B*); this is the "end fire" pattern. The interference cancelling null is now perpendicular to line *A*-*B*.

It should be apparent that you can select your directivity by selecting the phase angle of the feed currents in the two antennas. Figure 11-2 shows the two feeding

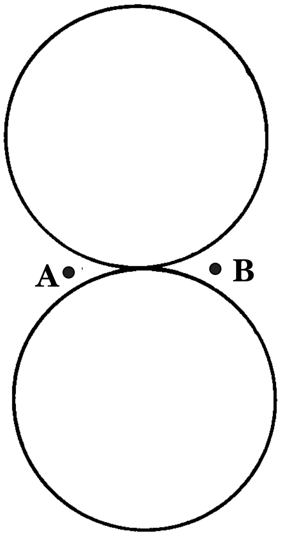

11-1A Pattern of two radiators in-phase, spaced a half-wavelength apart.

11-1B Pattern of two radiators out of phase, spaced a half-wavelength apart.

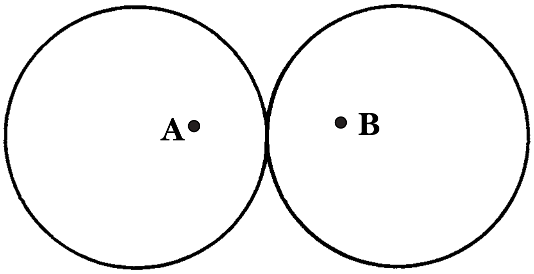

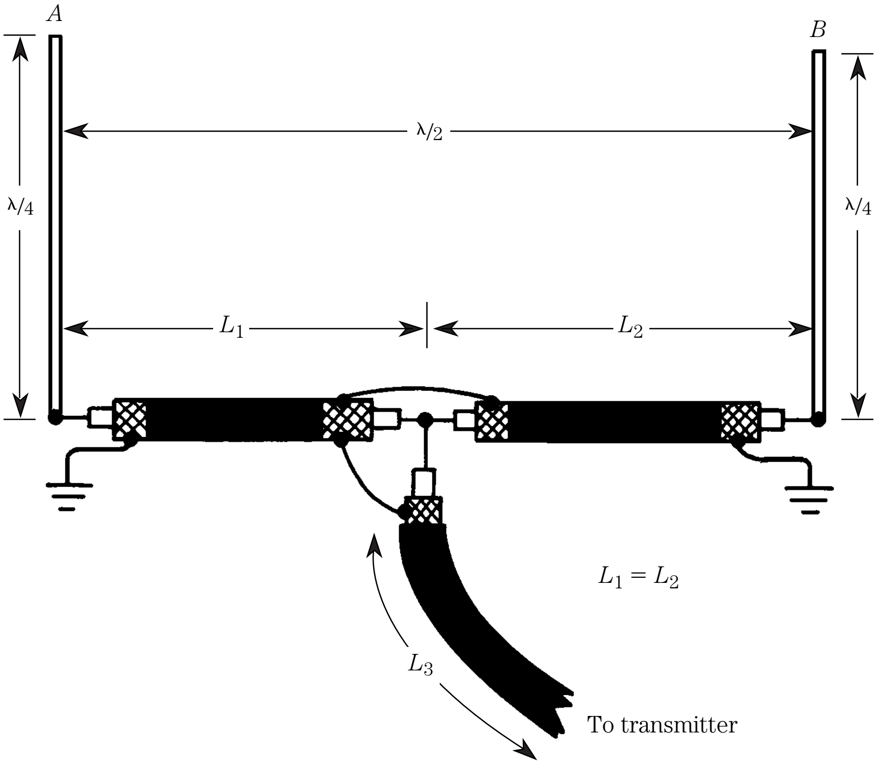

11-2A Feeding a phased array antenna in phase.

systems usually cited for in-phase (Fig. 11-2A) and out-of-phase (Fig. 11-2B) systems. Figure 11-2A shows the coax from the transmitter coming to a coax tee connector. From the connector to the antenna feedpoints are two lengths of coax (L_1 and L_2) that are equal to each other, and identical. Given the variation between coaxial cables, I suspect that it would work better if the two cables were not merely the same length ($L_1 = L_2$), but also that they came from the same roll.

The second variation, shown in Fig. 11-2B, supposedly produces a 180° phase shift between antenna *A* and antenna *B*, when length L_3 is an electrical half-wavelength. According to a much-publicized theory, the system of Fig. 11-2B ought to

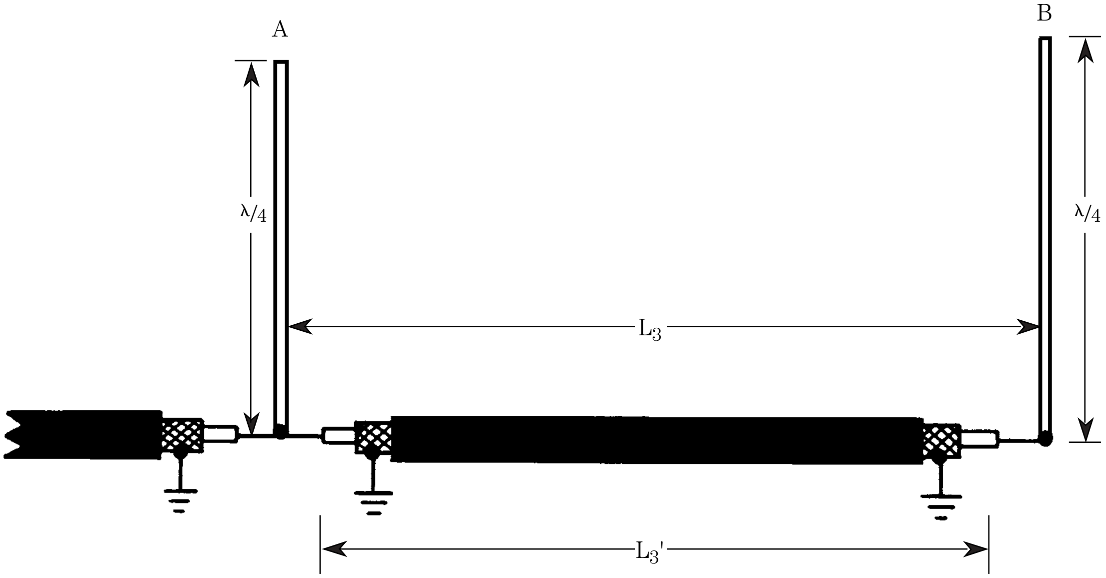

11-2B Feeding a phased array antenna out of phase.

produce the pattern of Fig. 11-1B—yet experience shows this claim is false. It seems that there are several problems with the system in Fig. 11-2B.

First, coax has a property called *velocity factor* (V_F), which is the fraction of the speed of light at which signals in the cable propagate. The V_F is a decimal fraction on the order of 0.66 to 0.90, depending upon the type of coax used. Unfortunately, the physical spacing between *A* and *B* is a real half-wavelength ($L_3 = 492/F$), but the cable length is shorter by the velocity factor [$L_3' = (V_F \times 492)/F$].

Consider an example. A 15-m phased vertical antenna system will have two 11-ft radiators, spaced 22 ft apart (approximately, depending upon exact frequency). If we use foam coax, with $V_F = 0.80$, the cable length is 0.8 × 22 ft, or 17.6 ft. In other words, despite lots of publicity, the cable won't fit between the towers!

Second, the patterns shown in Fig. 11-1 are dependent upon one condition: the *antenna currents are equal*. If both of them are the same impedance, and are fed from the same transmitter, then it is reasonable to assume that the currents are equal—right? No, wrong! What about coax loss? Because of normal coax loss, which increases at higher frequencies, the power available to antenna *B* in Fig. 11-1B is less than the power available to antenna *A*. Thus, the pattern will be somewhat distorted, because the current produced in *B* is less than the current in *A*, when they should be equal.

The first problem is sometimes fixed by using unequal lengths for cables L_1 and L_2 (Fig. 11-2A), and using it for the out-of-phase case. For example, if we make L_1 one-quarter wavelength and L_2 three-quarter wavelength (Fig. 11-2C), antenna *A* is fed with a 90° phase lag (relative to the tee connector signal), while antenna *B* is fed with a 270° phase shift. The result is still a 180° phase difference. Unfortunately, we

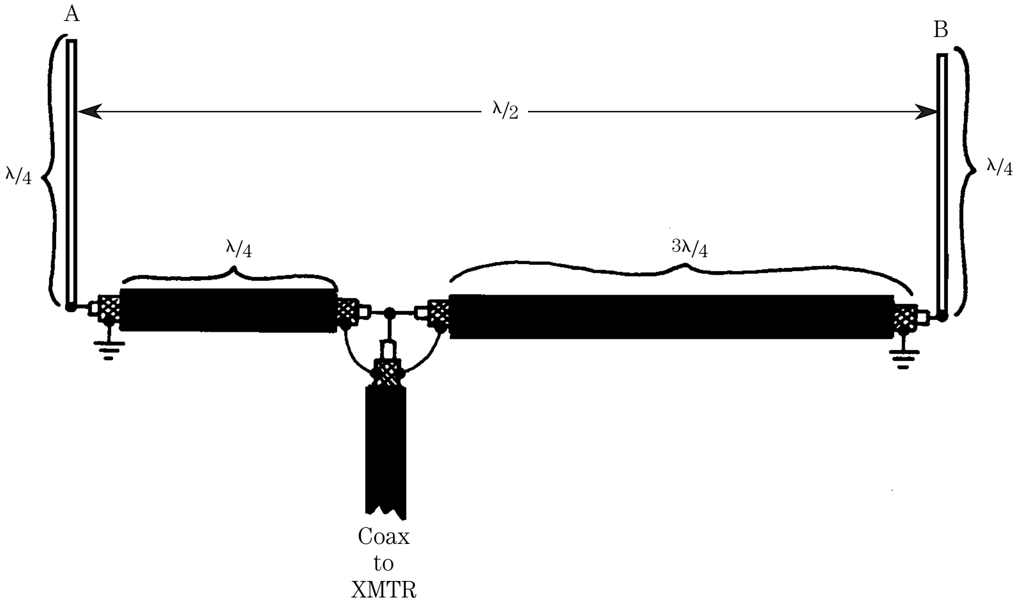

11-2C Corrected feed for a phased array antenna.

have not solved the current level problem, and may have actually made it worse by adding still more lossy cable to the system.

There is still another problem that is generic to the whole class of phased verticals. Once installed, the pattern is fixed. This problem doesn't bother most point-to-point commercial stations, or broadcasters, because they tend to transmit in only one direction. But amateurs are likely to need a rotatable pattern. Neither the antennas in Fig. 11-1A nor that in Fig. 11-1B is rotatable without a lot of effort—like changing the coax feeds, or physically digging up the verticals and repositioning them.

Fortunately, there is a single solution to all three problems. Figure 11-3 shows a two-port phasing transformer made from a toroidal balun kit. Use the kind of kit that makes a 1:1 balun transformer. Although we are not making a balun, we will need enough wire to make three windings, and that is the normal case for 1:1 baluns. Amidon Associates and others make toroidal balun kits.

Wind the three coils in trifilar style, according to the kit instructions. The dots in Fig. 11-3 show the "sense" of the coils, and they are important for correct phasing; call one end the "dot end" and the other end the "plain end" to keep them separate. If the dot end of the first coil is connected to J_3 (and the transmitter), then connect the dot end of the second coil to the 0° output (J_1, which goes to antenna A). The third coil is connected to a DPDT RF relay or switch. In the position shown, S_1 causes the antennas to be 180° out of phase. In the other position, the "sense" of the third coil is reversed, so the antennas are in phase.

Another phasing method is shown in Fig. 11-4. In this scheme, two convenient, but equal, lengths of coaxial cable (L_1 and L_2) are used to carry RF power to the antennas. One segment (L_1) is fed directly from the transmitter's coaxial cable (L_3), while the other is fed from a phasing switch. The phasing switch is used to either by-

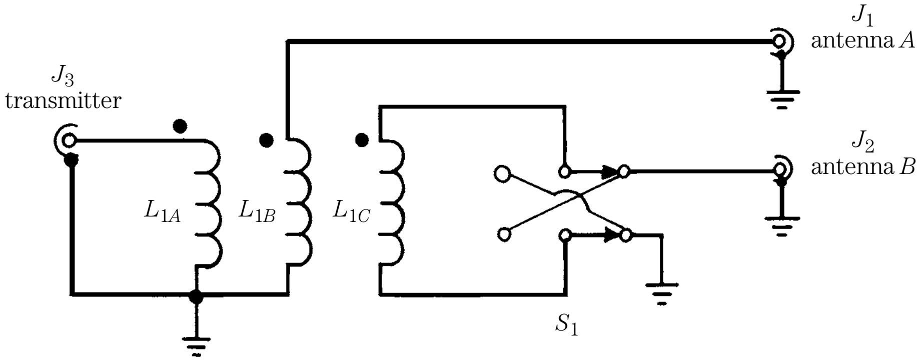

11-3A Phasing transformer circuit.

pass or insert a phase-shifting length of coaxial cable (L_4). For 180° phasing use the following equation to find the length (L_4):

$$L = \frac{492\ V_F}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[11.1]}$$

where

L is the length of $L_{4,}$ in feet
V_F is the velocity factor (a decimal fraction)
F_{MHz} is the operating frequency, in megahertz

Some people use a series of switches to select varying amounts of phasing shift from 45° to 270°. Such a switch allows them to select any number of other patterns for special situations.

360° directional array

The phased vertical antenna concept can be used to provide round-the-compass control of the antenna pattern. Figure 11-5A shows how three quarter-wavelength verticals (arranged in a triangle that is a half-wavelength on each side) can be used to provide either end-fire or broadside patterns from any pair (*A*-*B*, *A*-*C*, or *B*-*C*). Any given antenna (*A*, *B*, or *C*) will be grounded, fed at 0°, or fed with 180°. The table in Fig. 11-5B shows the relative phasing for each direction that was labelled in Fig. 11-5A. Either manual phase changing or switch-operated phase changing can be used, although the latter is preferred for convenience. Some international showcase broadcasters use antenna arrays formed into two or more concentric circles of vertical elements, with one element at the center. Selection of elements and phasing determines directivity and gain.

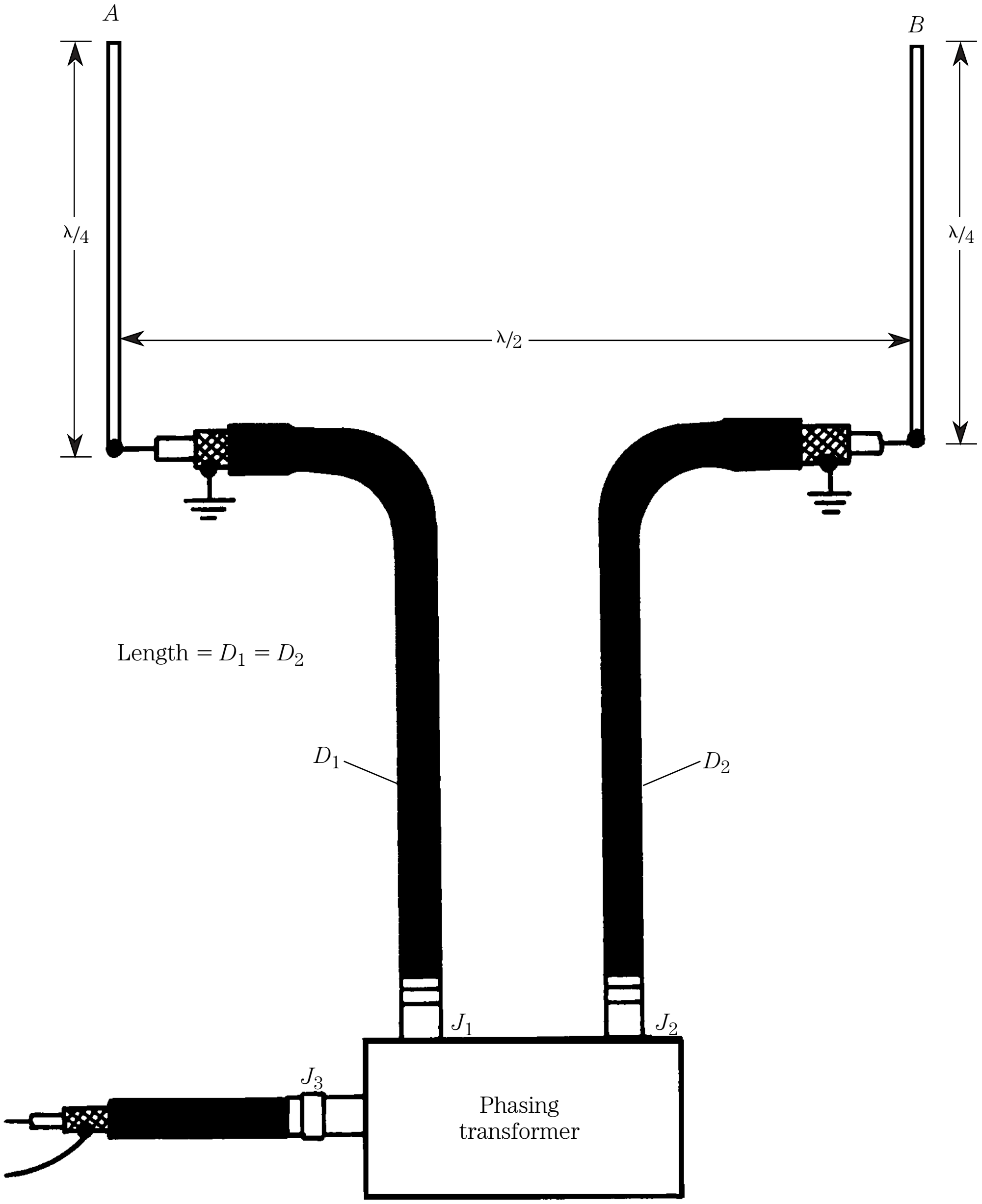
A
B
λ/4
λ/4
λ/2
Length = $D_1 = D_2$
D_1
D_2
J_1
J_2
J_3
Phasing transformer

11-3B Connection to antennas.

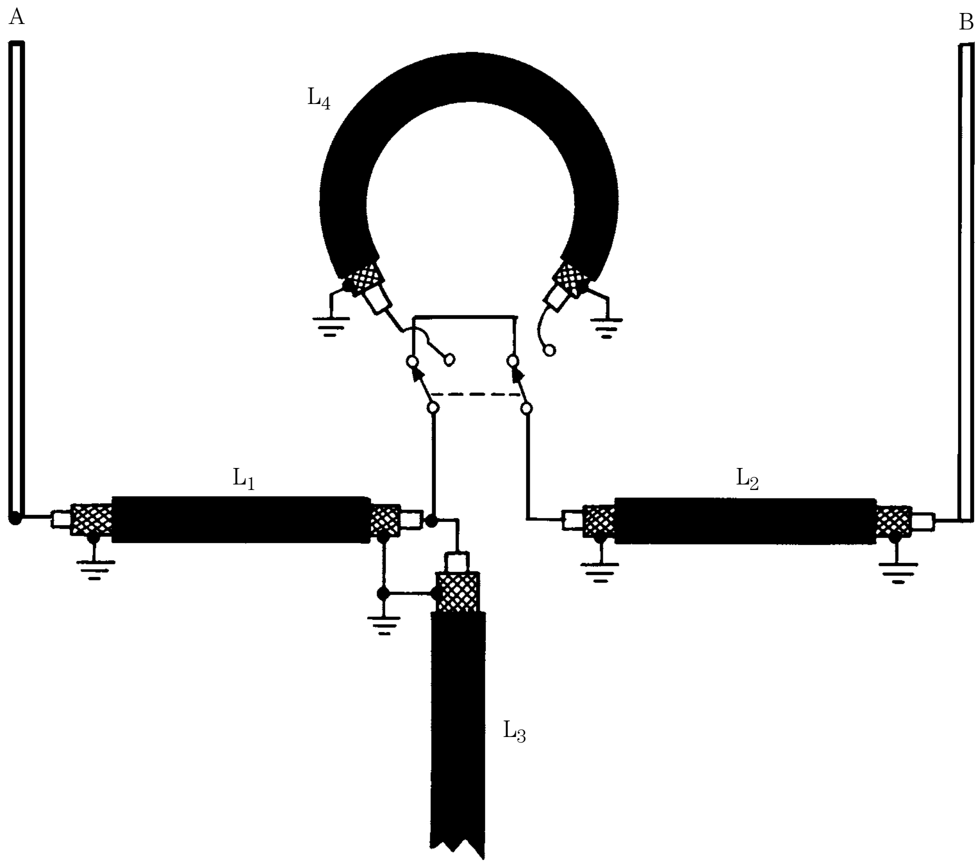

11-4 Phase-shifting antenna circuit.

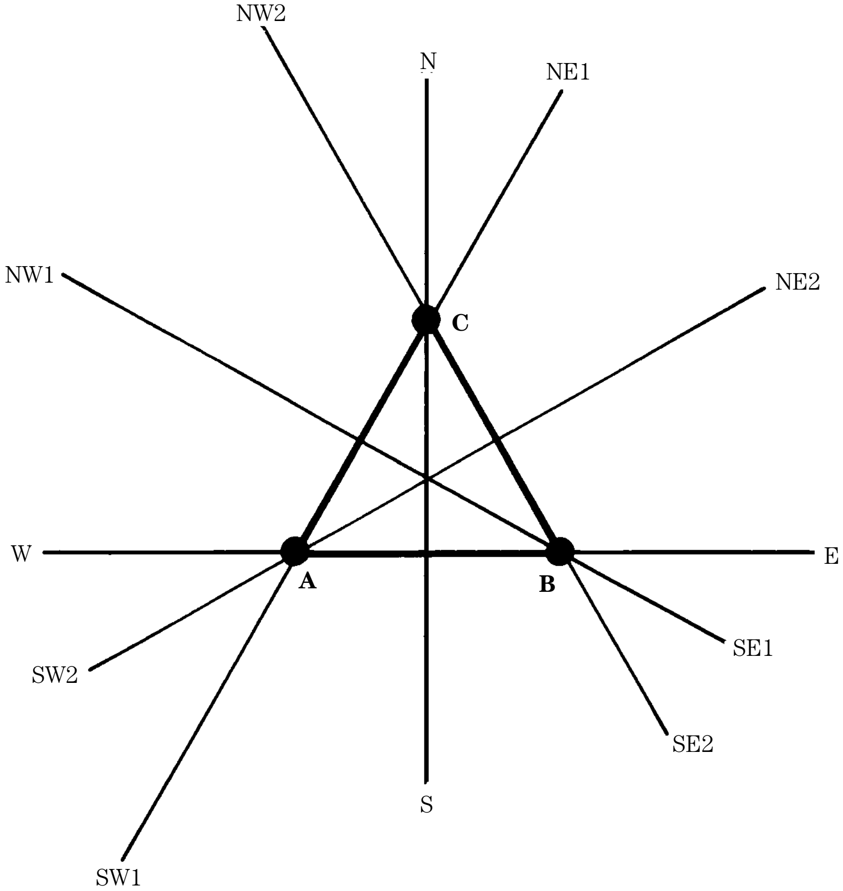
NW2
N
NE1
NW1
NE2
C
W
E
A
B
SE1
SW2
SE2
S
SW1

11-5A Three-element phased array.

Direction	Ant A	Ant B	Ant C
N-S	0°	0°	GND
NE1-SW1	0°	GND	180°
NE2-SW2	GND	0°	0°
E-W	0°	180°	GND
SE1-NW1	0°	GND	0°
SE2-NW2	GND	0°	180°

11-5B Table of feed phasing for the three-element array. Ant = antenna. GND = ground

12
CHAPTER

Directional beam antennas

THE DIRECTIONAL BEAM ANTENNA DOES SEVERAL JOBS. FIRST, IT PROVIDES AN APPARENT increase in radiated power, because it focuses available transmitter power into a single (or at worst limited) direction. For this reason, a bidirectional dipole has a gain of approximately 2 dB over an isotropic radiator. Add one or more additional elements, and the focusing becomes nearly unidirectional, which increases the *effective radiated power* (ERP) even more. Second, the beam increases the received signal available at the inputs of the receiver. Antennas are generally reciprocal, so they will work for receiving as they do for transmitting. Finally, the directivity of the beam antenna allows the operator to null interfering stations. In fact, it is the last attribute of the beam that is most useful on today's crowded bands. All in all, if your funds are too little to provide both increased RF power and a good antenna system, then spend what is available on the antenna—not on the power.

In this chapter we will focus on directional antennas that can be built relatively easily. It is assumed that most readers who want a triband multielement Yagi will prefer to buy a commercial product, rather than build a homebrew model. The material herein concentrates on homebrew projects that are within the reach and capabilities of most readers. The first of these is not a beam antenna at all, but rather a *rotatable dipole.*

Rotatable dipole

The *dipole* is a bidirectional antenna with a figure-8 pattern (when viewed from above). The dipole is a half-wavelength and is usually installed horizontally, although vertical half-wavelength dipoles are known. Although the length of the dipole is too great for rotatability at the lower bands, it is within reason for the higher band. For example, the size of the halfwave dipole is approximately 16 ft on 10 m and 22 ft on 15 m. Even the 33-ft length on 20 m is not unreasonable for amateur constructors. The length of the dipole is found from

Copyright 2001 - the McGraw-Hill Companies

$$L = \frac{468}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[12.1]}$$

This length is approximate because of end effects and other phenomena, so some "cut and try" is required.

Example 12-1 Find the length of a dipole antenna for a frequency of 24.930 MHz in the 12-m amateur radio band.

Solution:

:

$$L = \frac{468}{F_{\text{MHz}}}\ \text{ft}$$

$$= \frac{468}{24.930\ \text{MHz}} = 18.77\ \text{ft}$$

The half-wave dipole is fed in the center by coaxial cable. Each element of the dipole is one-half of the overall length (or, in the example given, about 9.4 ft).

Figure 12-1 shows a rotatable dipole that can be designed for use on 15, 12, and 10 m. The radiator elements are made from 10-ft lengths of $\frac{3}{4}$-in aluminum tubing. The tubing is mounted on "beehive" standoff insulators, which in turn are mounted on a 4-ft length of 2 × 2 lumber. The lumber should be varnished against weathering. In a real pinch, the elements can be mounted directly to the lumber without the insulators, but this is not the recommended practice.

The mast is attached to the 2 × 2 lumber through any of several means. The preferred method is the use of a 1-in pipe flange. These devices are available at hardware stores under the names *floor flange* and *right-angle flange.*

The 10-ft lengths of pipe are the standard lengths available in hardware stores, so it was selected as being closest to the required 22 ft for 15 m. A 0.14-µH loading coil is used at the center, between the elements, in order to make up for the short length. The dimensions of the coil are 4 to 5 turns, 0.5-in diameter, 4-in length. For low power levels, the coil can be made of no. 10 (or no. 12) solid wire—and, for higher levels, $\frac{1}{8}$-in copper tubing.

There are two basic ways to feed the antenna, and these are shown in details A and B in Fig. 12-1. The traditional method is to connect the coaxial cable (in parallel) across the inductor. This method is shown in Fig. 12-1, detail A. A second method is to link couple the coil to the line through a one- to three-turn loop (as needed for impedance matching). This is the method that would be used for a toroidal inductor.

Lower frequencies can be accommodated by changing the dimensions of the coil. The coil cannot be scaled, simply because the relative length of the antenna changes as the frequency changes. But it is possible to cut and try by adding turns to the coil, one turn at a time, and remeasuring the resonant frequency. Adding inductance to the coil will make the antenna usable on 17 m and 20 m, as well as on 15 m.

Another method for building a rotatable dipole for lower frequencies is to increase the element lengths. On 17 m, the overall length is approximately 27.4 ft, so each element length is 13.7 ft long. This length can be achieved by either of two methods. First, adjacent sizes of aluminum tubing are designed so that the smaller will be a slip-fit inside of the larger. What constitutes "adjacent sizes" depends on the wall thickness, but for one common brand, the $\frac{7}{8}$-in is adjacent to the $\frac{3}{4}$-in size. You

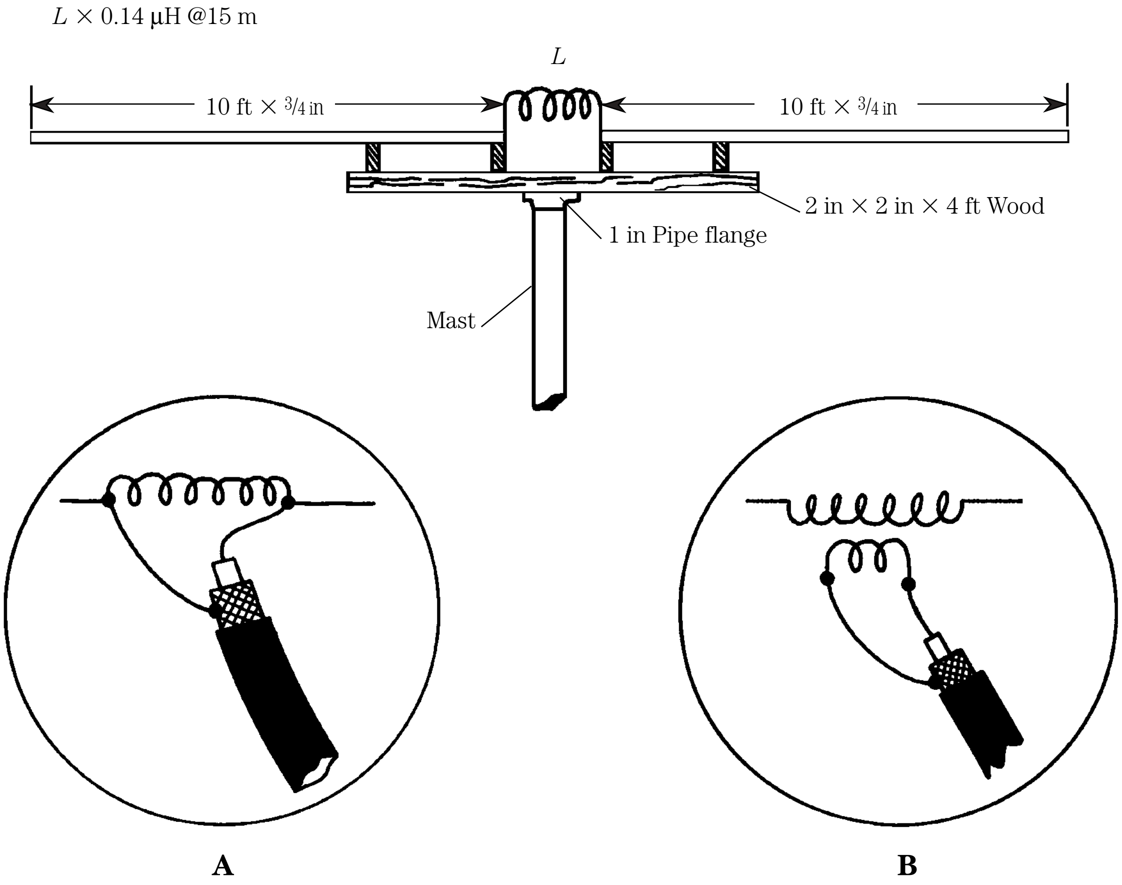

12-1 Rotatable dipole antenna. Inset A shows conventional feed; inset B shows transformer feed.

can use two smaller lengths to make the larger lengths of pipe, and cut it to size. This method is only available to those readers who have a commercial or industrial metals distributor nearby, because the 16-ft lengths are not generally available from hardware stores.

Bands higher than 15 m (i.e., 12 and 10 m) can be accommodated by using the 10-ft lengths of tubing, but without the inductor. The tubing is cut to the desired half-wavelength size and used directly.

Yagi beam antennas

A *Yagi* antenna is one member of a class of directional beam antennas that are popular in the HF bands. Figure 12-2 shows the pattern (viewed from above) typical of the beam antenna. The antenna is located at point P and fires signals in the direction shown by the arrow. The beamwidth of the antenna is the angle a between the points on the main lobe that are −3 dB down from the center point C.

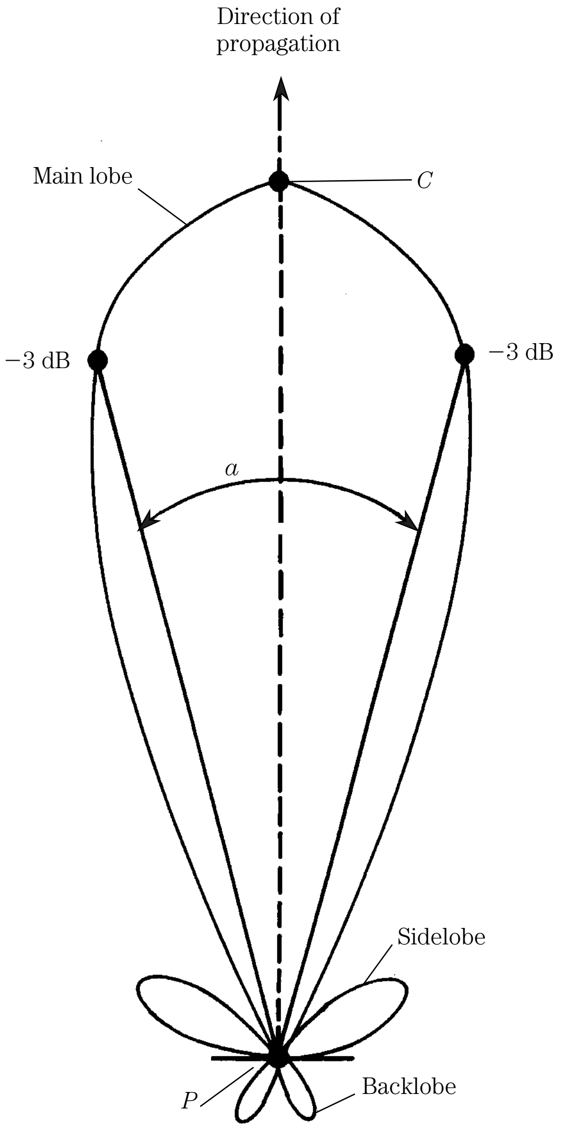

12-2 Pattern of beam antenna.

A perfect beam antenna will have only the main lobe, but that situation occurs only in dreams. All real antennas have both sidelobes and backlobes, also shown in Fig. 12-2. These lobes represent wasted power transmitted in the wrong direction during transmission and interference opportunities while receiving. The goal of the antenna designer is to increase the main lobe while decreasing the sidelobes and backlobes.

Figure 12-3 shows, schematically, the basic *Yagi-Uda* antenna (usually called simply *Yagi*). The driven element is a simple half-wavelength dipole fed in the center. There are two additional elements: *reflectors* and *directors*. These are called *parasitic elements* because they are not directly excited by RF, but rather, they receive energy radiated from the driven element and then reradiate it. The reflector is placed behind the driven element, and is typically about 4 percent longer than the driven element. The director is placed in front of the driven element (relative to the direction of propagation). The director is typically about 4 percent shorter than the driven element. Although there is no fixed rule regarding the number of either reflectors or directors, it is common practice to use a single director and a driven element for two-element beams, and a single reflector and a single director, in addition to the driven element, for three-element beams. Again, additional reflectors can be added for four- and more element beams, but standard practice calls for addition of directors instead.

The length of the elements is given by

$$L = \frac{K}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[12.2]}$$

where

L is the length, in feet
F_{MHz} is the frequency, in megahertz
K is a constant

The spacing of the elements is typically from 0.15 to 0.308 wavelengths, although 0.2 and 0.25 are the most common values.

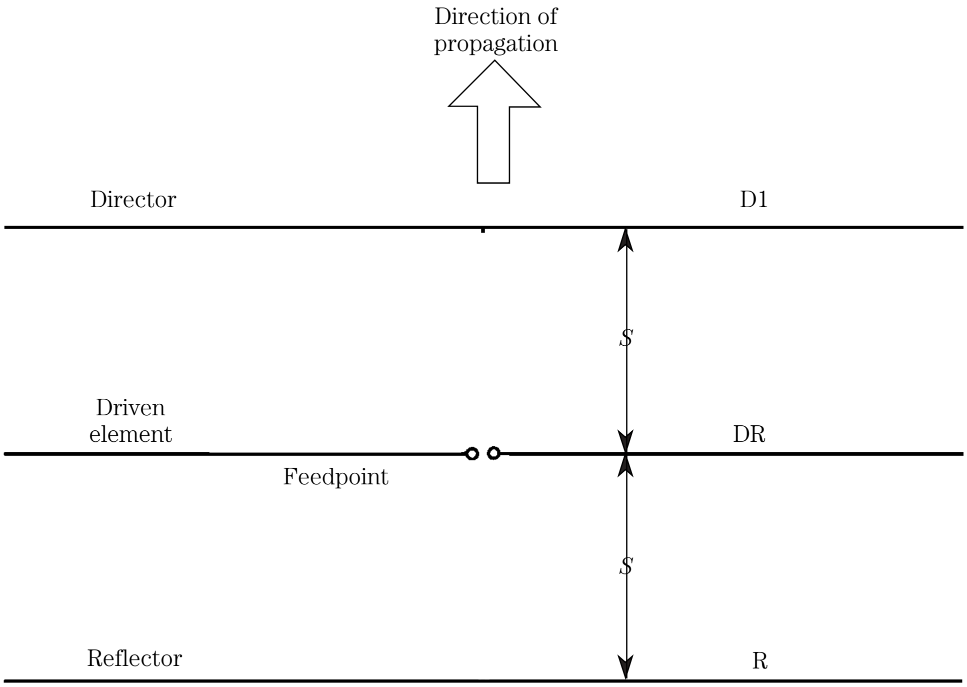

12-3 Basic Yagi-Uda antenna.

Example Calculate the approximate element lengths for a three-element 15-m beam designed to operate on a frequency of 21.39 MHz.

Solution:

1. Driven element ($K = 478$):

$$L = \frac{K}{F_{\text{MHz}}}\text{ ft} = \frac{478}{21.39\text{ MHz}}\text{ ft} = 22.34\text{ ft}$$

2. Reflector ($K = 492$):

$$L = \frac{K}{F_{\text{MHz}}}\text{ ft} = \frac{492}{21.39\text{ MHz}}\text{ ft} = 23\text{ ft}$$

3. Director ($K = 461.5$):

$$L = \frac{K}{F_{\text{MHz}}}\text{ ft} = \frac{461.5}{21.39\text{ MHz}}\text{ ft} = 21.58\text{ ft}$$

4. Element Spacing (K = 142):

$$L = \frac{K}{F_{\text{MHz}}}\text{ ft} = \frac{142}{21.39\text{ MHz}}\text{ ft} = 6.64\text{ ft}$$

The elements of a rotatable beam antenna can be built in a manner similar to the rotatable dipole described earlier. In the beam antenna, however, a boom is needed between the elements to support them. The boom can be made of metal or wood. In the case of a metal boom, the driven element must be insulated from the boom, even though the parasitic elements can be mounted directly to it. In general, it is usually better to use wood as a matter of convenience. Metal boom antennas can be obtained from commercial sources. The wood boom is easy to build and maintain, even though a little less durable than a metal boom.

The feedpoint impedance of a dipole is on the order of 72 Ω in free space, although the actual impedance will vary above and below that figure for antennas close to the earth's surface. In addition, adding parasitic elements reduces the impedance even more. The feedpoint impedance of the antenna is too low to be directly fed with coaxial cable, so some means of impedance matching is needed. Some people feed the antenna through an impedance matching balun transformer. Figure 12-4 shows the *gamma match* system. The driven element of the Yagi is not broken in the center, as in the case of the simple dipole. The outer conductor, or shield, of the coaxial cable is connected to the center point of the driven element. The center conductor is connected to the gamma match element. The dimensions of the gamma match are

1. Gamma match length: $L/10$
2. Gamma match director: $D/3$
3. Spacing of gamma match from driven element: $L/70$

where

L is the length of the driven element
D is the diameter of the driven element

The capacitor in series with the center conductor of the coaxial cable has a value of approximately 8 pF per meter of wavelength at the lowest frequency in the band of operation, or approximately,

$$C = \frac{2400}{F_{\text{MHz}}} \text{ pF} \qquad \textbf{[12.3]}$$

The capacitor must be a high-voltage transmitting variable type. In general, the gamma match capacitors are either air or vacuum variables.

There are three aspects to the adjustment of the Yagi antenna. Resonance is determined by the length of the element. The length is increased or decreased in order to find the resonant point. This point can be determined by the use of a noise bridge, VSWR meter or other means. The capacitor, and the shorting bar/clamp, are adjusted to match the impedance of the antenna to the transmission line impedance. For the dimensions shown, the coaxial cable should be 52 Ω (RG-58 or RG-8).

It is not necessary to use tubing or pipes for the antenna elements in order to obtain the benefits of the Yagi beam antenna. An example of a wire beam is shown in Fig. 12-5. The wire beam is made as if it were two half-wavelength dipoles, installed parallel to and about 0.2 to 0.25 wavelengths apart from each other. Although multielement wire beams are possible, the two-element version is the most common. Perhaps the most frequent

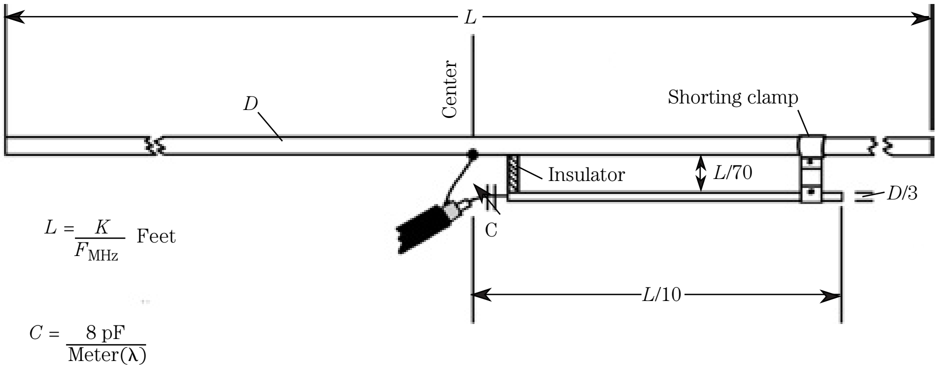

12-4 Gamma feed provides impedance matching.

use of the wire beam antenna is on the lower bands (e.g., 40 m and 75/80 m) where rotatable beams are more difficult to build.

20-m ZL-special beam

The antenna shown in Fig. 12-6 is a close relative of the Yagi beam. It consists of a pair of folded dipoles, mounted approximately 0.12 wavelengths apart. The elements are 30.5 ft in length, and the spacing is 7.1 ft. The elements can be built from aluminum tubing if the antenna is to be rotatable. Alternatively, for a fixed antenna, the elements can be made of 300-Ω television-type twin lead. If the tubing type of construction is selected, then make the size of the tubing, and its spacing, sufficient for 300-Ω parallel transmission lines, according to

$$Z_o = 276 \log \frac{2S}{d} \qquad \textbf{[12.4]}$$

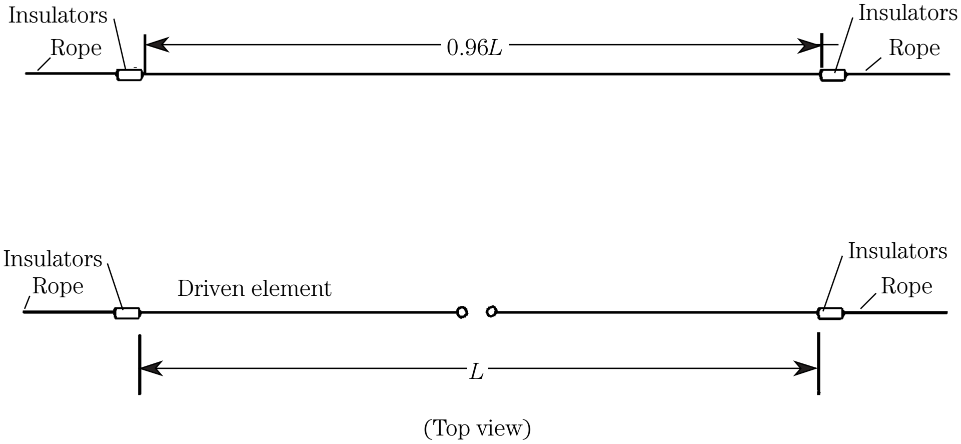
Insulators
Rope
0.96L
Insulators
Rope
Insulators
Rope
Driven element
Insulators
Rope
L
(Top view)

12-5 Wire beam useful for low frequencies.

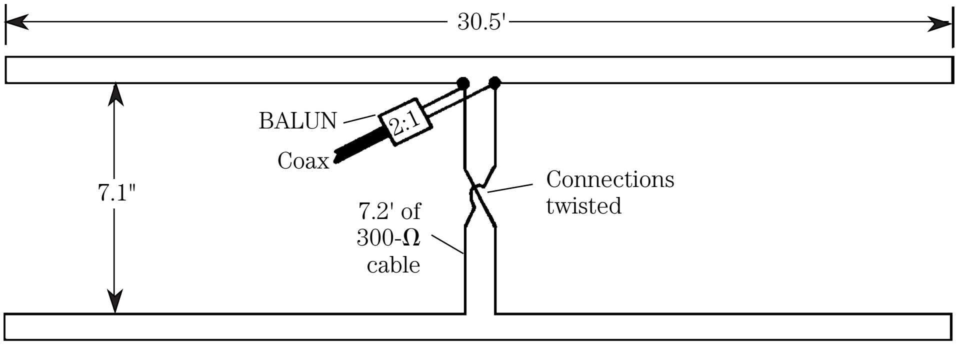
30.5'
BALUN
2:1
Coax
7.1"
7.2' of
300-Ω
cable
Connections
twisted

12-6 ZL-special beam antenna.

where

Z_o is the impedance of the line (300 Ω)
S is the center-to-center spacing of the parallel conductors on the elements
d is the diameter of the conductors

The two half-wavelength elements of the ZL-special are fed 135° out of phase with each other. The feedline is connected to one of the dipoles directly; and then to the other through a length of 300-Ω twin lead that has an electrical length of about 45° (λ/8).

The feedpoint impedance is on the order of 100 to 150 Ω, so it will make a good match to either 52- or 75-Ω coaxial cable if a 2:1 impedance matching transformer is used.

Cubical quad beam antenna

The *cubical quad* antenna is a one-wavelength square wire loop. It was designed in the mid-1940s at radio station HCJB in Quito, Ecuador. HCJB is a Protestant missionary shortwave radio station with worldwide coverage. The location of the station is at a high altitude. This fact makes the Yagi antenna less useful than it is at lower altitudes. According to the story, HCJB originally used Yagi antennas. These antennas are fed in the center at a current loop, so the ends are high-voltage loops. In the thin air of Quito, the high voltage at the ends caused corona arcing, and that arcing periodically destroyed the tips of the Yagi elements. Station engineer Clarence Moore designed the cubical quad antenna (Fig. 12-7) to solve this problem. Because it is a full-wavelength antenna, each side being a quarter wavelength, and fed at a current loop in the center of one side, the voltage loops occur in the middle of the adjacent sides—and that reduces or eliminates the arcing. The elements can be fed in the center of a horizontal side (Figs. 12-7A and 12-8A), in the center of a vertical side (Fig. 12-8B), or at the corner (Fig. 12-8C).

The antenna shown in Fig. 12-7A is actually a quad loop rather than a cubical quad. Two or more quad loops, only one of which needs to be fed by the coax, are used to make a cubical quad antenna. If only this one element is used, then the antenna will have a figure-8 azimuthal radiation pattern (similar to a dipole). The quad loop antenna is preferred by many people over a dipole for two reasons. First, the quad loop has a smaller "footprint" because it is only a quarter-wavelength on each side (A in Fig. 12-7A). Second, the loop form makes it somewhat less susceptible to local electromagnetic interference (EMI).

The quad loop antenna (and the elements of a cubical quad beam) is mounted to spreaders connected to a square gusset plate. At one time, carpets were wrapped around bamboo stalks, and those could be used for quad antennas. Those days are gone, however, and today it is necessary to buy fiberglass quad spreaders. A number of kits are advertised in ham radio magazines.

The details for the gusset plate are shown in Fig. 12-7B. The gusset plate is made of a strong insulating material such as fiberglass or ¾-in marine-grade plywood. It is mounted to a support mast using two or three large U bolts (stainless steel to prevent corrosion). The spreaders are mounted to the gusset plate using somewhat smaller U bolts (again, use stainless steel U bolts to prevent corrosion damage).

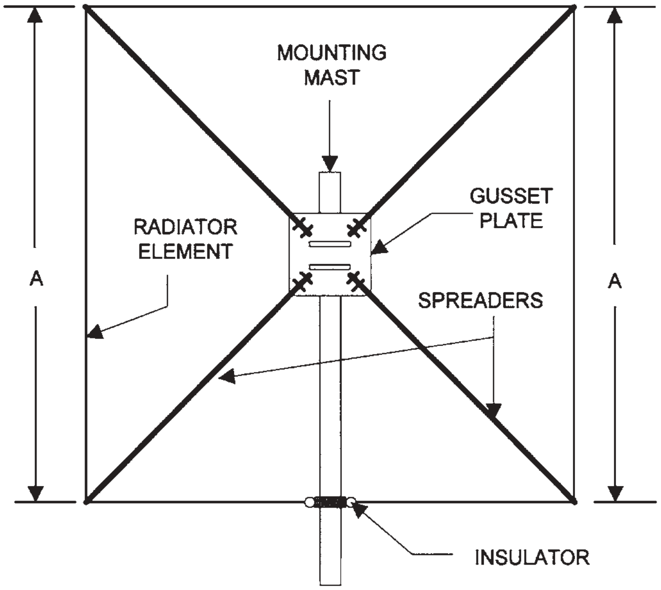
MOUNTING
MAST
GUSSET
PLATE
RADIATOR
ELEMENT
A
A
SPREADERS
INSULATOR

12-7A Quad loop antenna.

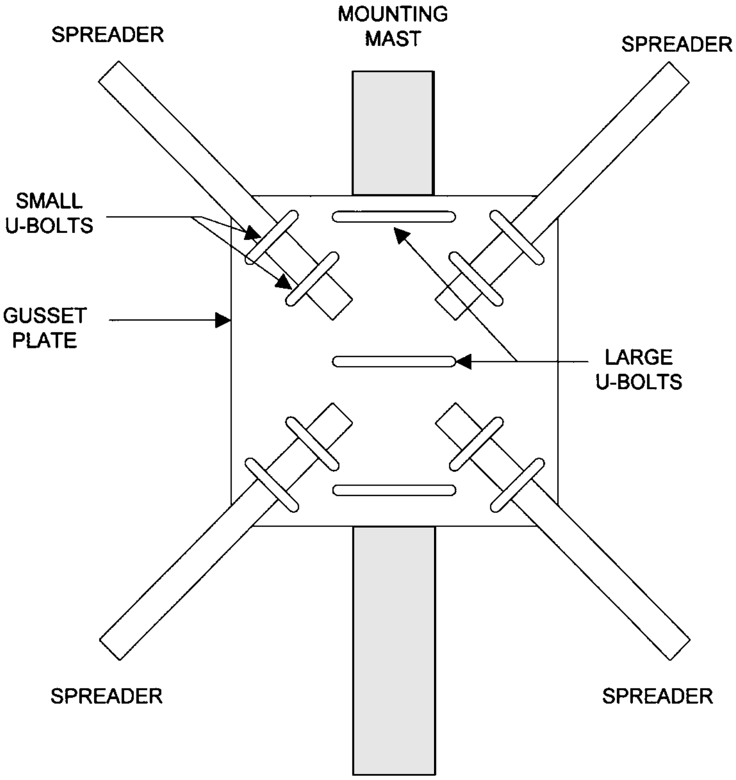
SPREADER
MOUNTING
MAST
SPREADER
SMALL
U-BOLTS
GUSSET
PLATE
LARGE
U-BOLTS
SPREADER
SPREADER

12-7B Quad loop antenna.

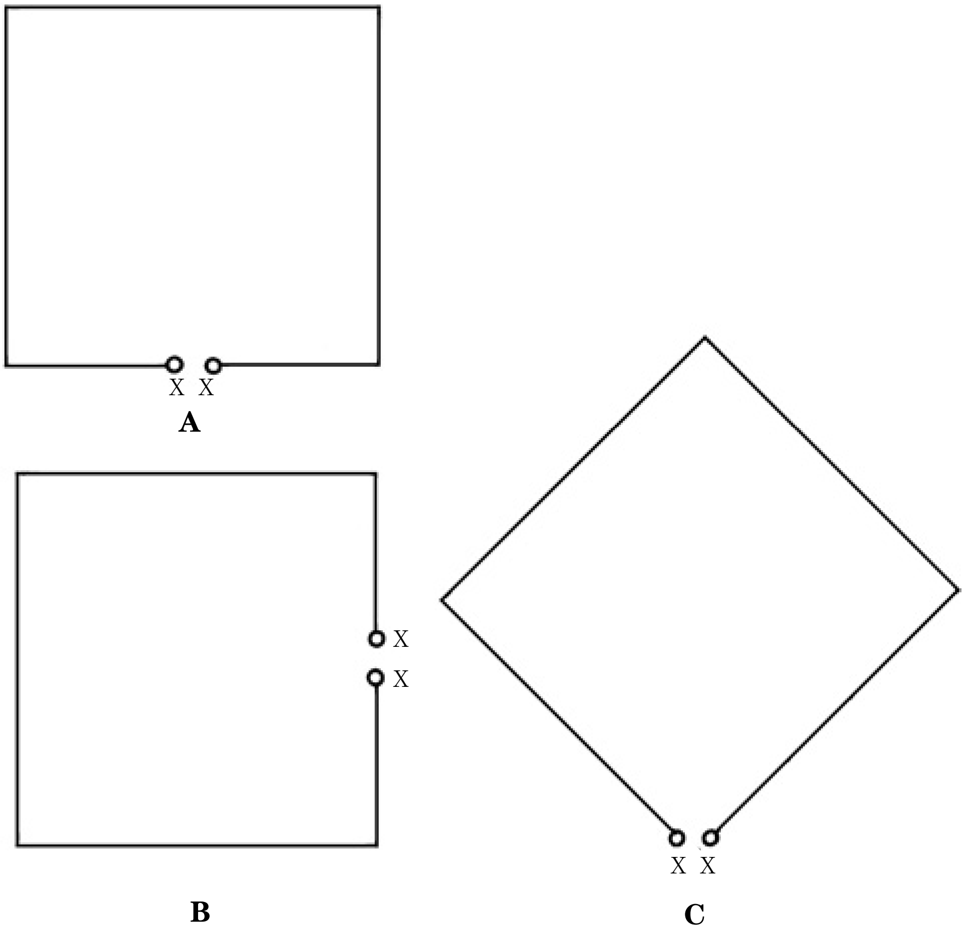

12-8 Feed options for the quad.

There is a running controversy regarding how the antenna compares with other beam antennas, particularly the Yagi. Some experts claim that the cubical quad has a gain of about 1.5 to 2 dB higher than a Yagi (with a comparable boom length between the two elements). In addition, some experts claim that the quad has a lower angle of radiation. Most experts agree that the quad seems to work better at low heights above the earth's surface, but the difference disappears at heights greater than a half-wavelength.

The quad can be used as either a single-element antenna or in the form of a beam. Figure 12-9 shows a pair of elements spaced 0.13 to 0.22 wavelengths apart. One element is the driven element, and it is connected to the coaxial-cable feedline directly. The other element is a reflector, so it is a bit longer than the driven element. A tuning stub is used to adjust the reflector loop to resonance.

Because the wire is arranged into a square loop, one wavelength long, the actual length varies from the naturally resonant length by about 3 percent. The driven

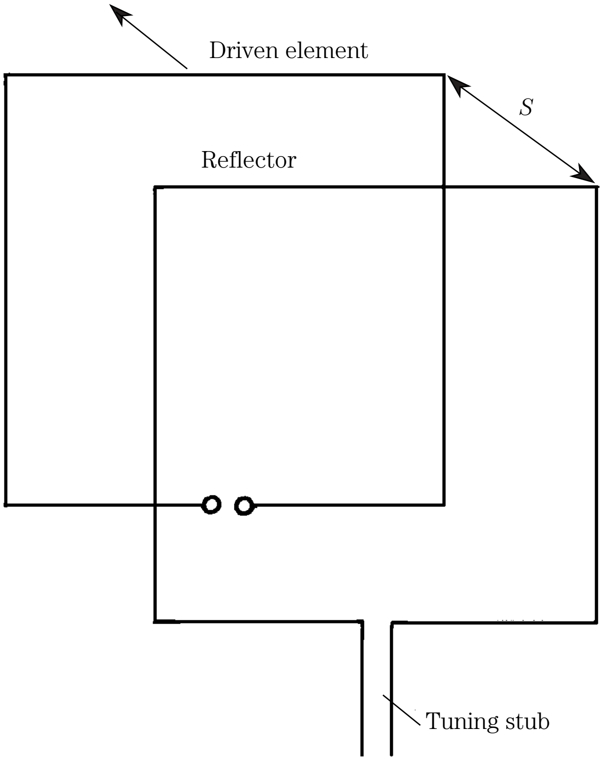

$0.13\lambda \leq S \leq 0.22\lambda$

12-9 Quad beam antenna.

element is about 3 percent longer than the natural resonant point. The overall lengths of the wire elements are

1. Driven element: $$L = \frac{1005}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[12.5]}$$

2. Reflector: $$L = \frac{1030}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[12.6]}$$

3. Director: $$L = \frac{975}{F_{\text{MHz}}} \text{ ft} \qquad \textbf{[12.7]}$$

One method for the construction of the quad beam antenna is shown in Fig. 12-10. This particular scheme uses a 12 × 12-in wooden plate at the center, bamboo (or fiberglass) spreaders, and a wooden (or metal) boom. The construction must be heavy-duty in order to survive wind loads. For this reason, it is probably a better solution to buy a quad kit consisting of the spreaders and the center structural element.

More than one band can be installed on a single set of spreaders. The size of the spreaders is set by the lowest band of operation, so higher frequency bands can be accommodated with shorter loops on the same set of spreaders.

This quad antenna is an example of a multielement, large loop antenna. Additional information on large loops, but not in a beam antenna array, is found in Chap. 14.

Inverted bobtail curtain (Thorne array)

The *bobtail curtain* antenna is a fixed array consisting of three individual quarter-wavelength elements spaced a half-wavelength apart, and fed from the top by a

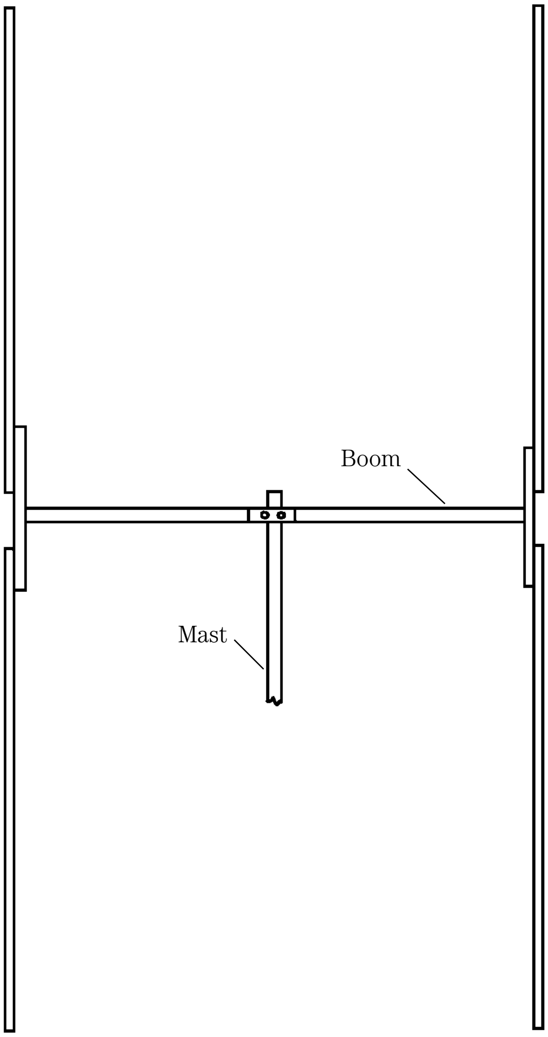

12-10 Quad construction.

shorting element or wire. The *inverted bobtail curtain,* or *Thorne array,* consists of an *upside down* bobtail curtain as shown in Fig. 12-11. The radiator elements are each a quarter-wavelength long. Their lengths are found from

$$L = \frac{246}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[12.8]}$$

The lengths of spacing between the elements are exactly twice above the value, or

$$L = \frac{492}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[12.9]}$$

The antenna is fed at the base of the center element, through a parallel resonant tuner. The capacitor is a 100- to 200-pF transmitting variable, while the inductor is set to resonate at the band desired (with the capacitor at half to three-quarters full capacitance). A loop or link coupling scheme connects the tuner to the transmission line.

An alternate feed method (Fig. 12-12) worked out by the late J. H. Thorne (K4NFU/5), feeds the end elements from the shield of the coaxial cable, and the center element of the array is fed from the center conductor of the coaxial cable. A coaxial impedance-matching section is used between the cable transmitter and the antenna feedpoint.

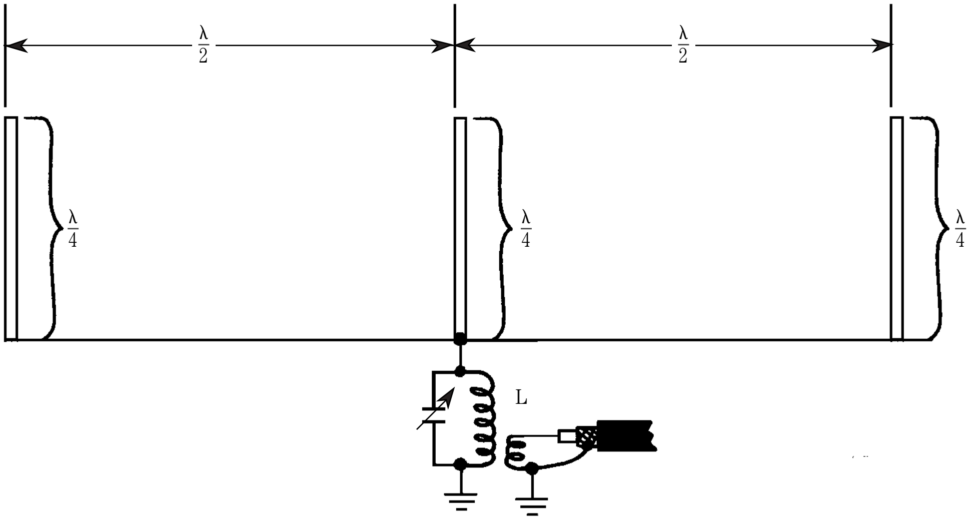

12-11 Inverted bobtail curtain.

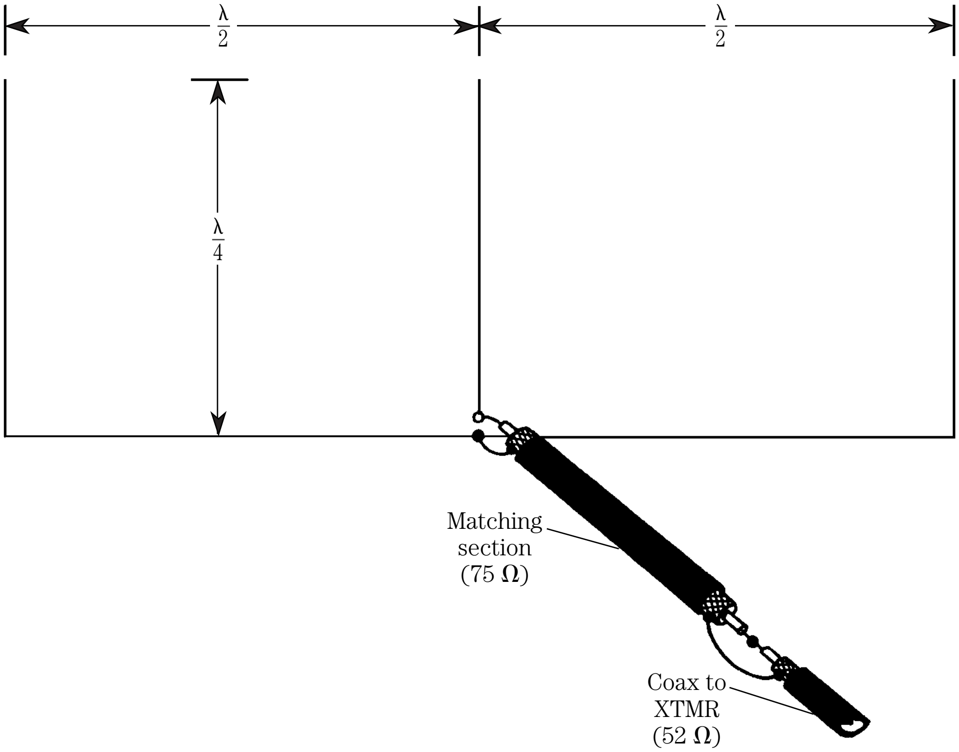
λ/2
λ/2
λ/4
Matching section (75 Ω)
Coax to XTMR (52 Ω)

12-12 Thorne array.

This page intentionally left blank

13
CHAPTER
Antennas for shortwave reception

ALTHOUGH MOST OF THIS BOOK ADDRESSES ANTENNAS FOR TRANSMITTING, THERE IS A certain body of material that pertains purely to receiving antennas. This material also needs to be addressed, and that is the function of this chapter. There are two readers in mind for this chapter. First, and foremost, is the shortwave listener (SWL). Second, however, is the amateur radio operator who wants to either use a separate receiving antenna on the main station receiver or use an ancillary receiver (common among DXers).

The law of reciprocity

Antennas possess a property called *reciprocity.* That is a fancy way of saying that an antenna works on reception the same way it does for transmission. Although articles occasionally appear in the literature claiming an HF or VHF design which violates the *law of reciprocity,* to date all have depended on either variable definitions, false premises, or faulty measurements. There is even a school of thought that falsely argues against antenna reciprocity based on an ionospheric anomaly in which propagation depends on the *direction* traveled (see the end of Chap. 2). The bottom line of reciprocity, for the SWL, is that every antenna described in this book can be also used with equal results on receive. For example, a half-wavelength dipole works equally well as a receiver antenna, or as a transmitter antenna.

Which properties are important?

Selecting a receiver antenna is a function of several factors. Assuming that you want more than a simple longwire (which we will deal with shortly), you will want to home in on the properties desired for your particular monitoring application. Is the antenna to be fixed or rotatable? Do you want omnidirectional or directional reception? In which plane? What about gain?

Copyright 2001 - the McGraw-Hill Companies

What about gain? What is gain, for that matter? The concept of antenna gain derives from the fact that directional antennas focus energy. Two kinds of gain figure are often quoted: gain referenced to a dipole antenna on the same frequency, and gain relative to a theoretical construct called an *isotropic radiator.* In the isotropic case, the reference radiator to which the antenna is compared is a spherical point source that radiates equally well in all directions. The dipole exhibits a figure-8 radiation and reception pattern.

The gain of an antenna is merely the ratio (usually expressed in decibels) of the power radiated in a given direction by two antennas (i.e., the reference antenna and the test antenna). If an antenna gain is listed as “8-dB gain over isotropic,” this means that, in the direction specified, the power radiated is 8 dB higher than the same total power applied to an isotropic radiator.

So, of what use is antenna gain? Two answers immediately present themselves. First, by accumulating more signal, the antenna essentially makes your receiver more sensitive. Note that the gain of the antenna does not create a higher powered signal, it merely increases the apparent signal power by focusing energy from a given direction. And, note well that *gain* implies *directivity.* Any antenna that claims to provide gain, but that is truly omnidirectional in all planes, is a fraud. The fundamental assumption is that *gain implies directivity.*

The concept of directivity (hence also of gain) is often taken to mean *horizontal directivity,* which is the case of a dipole antenna. But all forms of antenna radiate in three-dimensional space. Azimuth angle of radiation, and elevation angle of radiation, are both important. Certain 2-m vertical antennas are listed as “gain antennas,” yet the pattern in the horizontal direction is 360°—implying omnidirectional behavior. In the vertical plane, however, lost energy is compressed into a smaller range of elevation angles, so gain occurs by refocusing energy that would have been radiated at a higher than useful angle.

The second application of directivity is in suppressing interfering signals. On the regular AM and FM broadcast bands, channelization permits receiver selectivity to overcome adjacent channel interference in most cases. But in the HF amateur radio and international broadcast bands, channelization is either nonexistent, poorly defined, or ignored altogether. In these cases, interfering “adjacent” channel signals can wipe out a weaker, desired station. Similarly, with cochannel interference (i.e., when both stations are on the same frequency), two or more signals compete in a “dog fight” that neither will ever totally win. Consider Fig. 13-1A. Assume that two 9540-kHz signals, S_1 and S_2, arrive at the same omnidirectional vertical antenna. Either both signals will be heard, or the stronger signal will drown out the weaker signal.

Now, consider Fig. 13-1B. Here, a dipole is used as the receiving antenna, so a little directivity is obtained. The main lobes of the dipole are wide enough to provide decent reception of signal S_1, even though the antenna is positioned such that S_1 is not along the maximum line (dotted). But the positioning shown places the interfering cochannel signal (S_2) in the null off the ends of the dipole, so it weakens it considerably. The result will be enhanced reception of S_1. In Fig. 13-1B, the idea is not to exploit the ability of the gain antenna to increase the level of S_1. Indeed, by placing the antenna as shown, we are not getting S_1 levels as high as might otherwise be

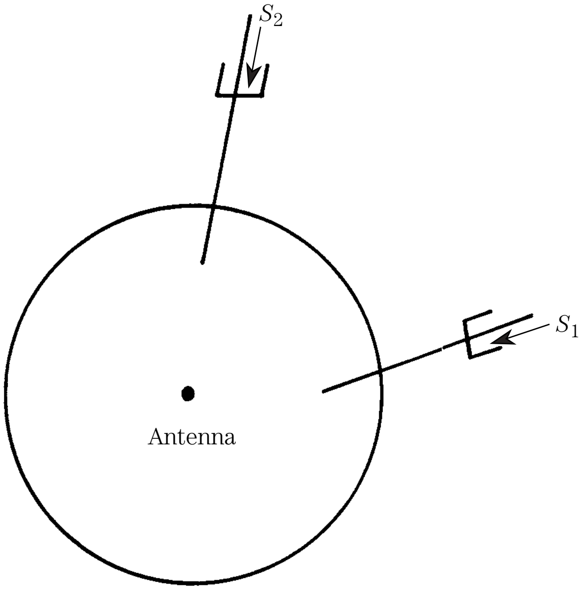

13-1A Omnidirectional antenna picks up cochannel interfering signals equally well.

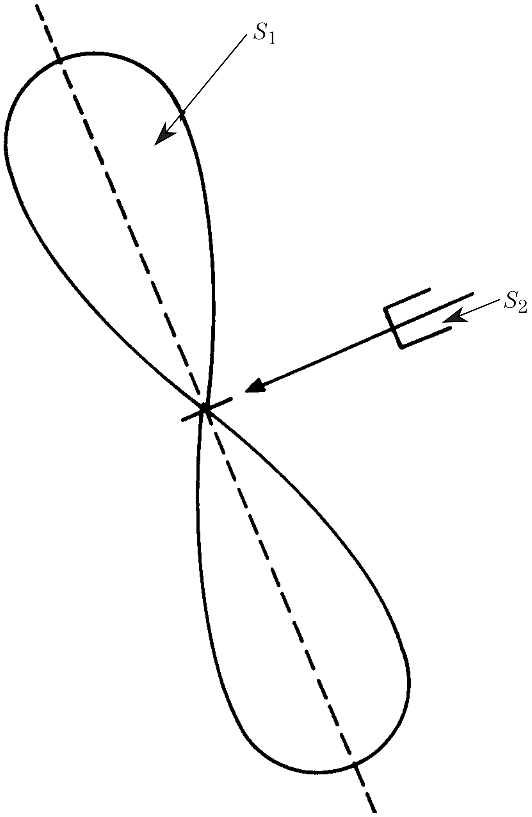

13-1B Directional antenna can discriminate against unwanted signal.

possible. The idea here is to place the unwanted signal (S_2) into the "notch" in order to make it considerably weaker. Note that the notch is sharper than the peak of the main lobe. If the dipole is placed on a mast, with an antenna rotator, this ability is increased even more.

Another antenna parameter, of considerable interest, is *angle of radiation a*, which (by reciprocity) also means angle of reception. Because HF propagation over long distances is created by skip phenomena, the angle at which the signal hits the ionosphere becomes extremely important. Figure 13-2 shows two situations from the same station. Signal S_1 has a high angle of radiation (a_1), so its skip distance (D_1) is relatively short. On signal S_2, however, the angle of radiation (a_2) is low, so the skip distance (D_2) is much longer than D_1.

So which situation do you want in your antenna? The impulsive answer would be the long distance angle of radiation (a_2), but that is often wrong. The correct answer is: "It depends!" The desired angle of radiation is a function of whether you want to receive a station from point A or point B.

The angle of radiation of the antenna is fixed by its design, that is, by antenna physics. The desired angle is a function of the ionospheric properties at the time of interest, and the operating frequency. For this reason, some well-equipped radio hobbyists have several antennas, of differing properties, to enhance their listening.

Connection to the receiver

It's simply too naive to state, I suppose, but let's do it anyway: An antenna must be properly connected to the receiver before it can be effective. If your antenna uses coaxial cable, and the receiver accepts coax, then no discussion is needed: Attach

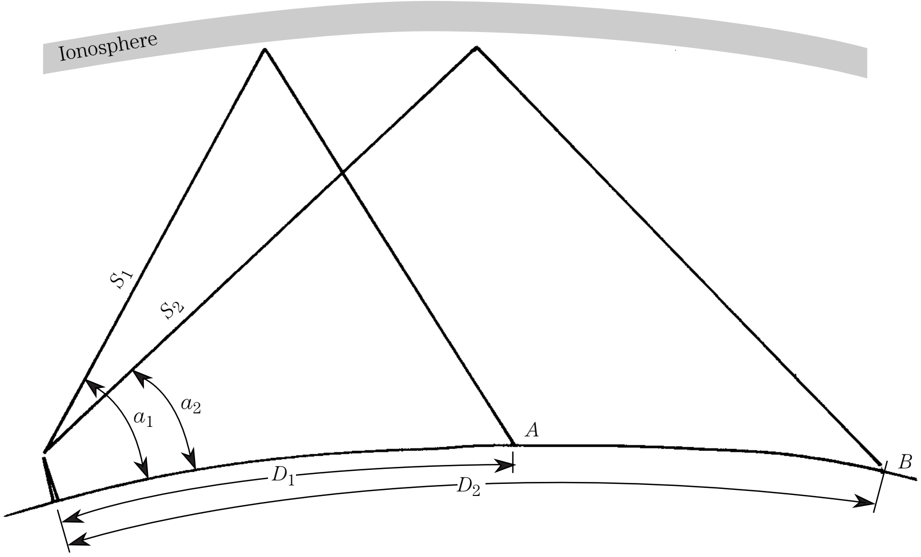

13-2 Skip phenomena dependence on angle of radiation.

the proper coax connector and plug in. But in other cases, non-coaxial-cable antennas are used.

There are two major forms of antenna input connector used on shortwave receivers. One form uses two (or three) screws intended for either wrapped wire leads or spade lugs, while the other is one or more varieties of coaxial connector. This section covers how each type is connected to a single-wire antenna lead-in.

Consider first the screw-type connector (Fig. 13-3A). Depending upon the design, there will be either two or three screws. If only two screws are found, then one is for the antenna wire and the other is for the ground wire. These screws will be marked something like "A/G" or "ANT/GND," or with the schematic symbols for antenna and ground.

Three-screw designs are intended to accommodate balanced transmission lines such as twin lead, or parallel ladder line. Shortwave listeners can sometimes use ordinary ac line cord (called *zipcord*) as an antenna transmission line. Zipcord has an impedance that approximates the 75-Ω impedance of a dipole. When parallel lines of any type are used, connect one lead to A1 and the other to A2. Of course, the ground terminal (G) is connected to the earth ground.

For single-lead antenna lines connect a jumper wire or bar (i.e., a short piece of bare no. 22 solid hookup wire) between A2 and G. This jumper converts the balanced input line to unbalanced. The A2/G terminal is connected to earth ground, while A1 is connected to the single-lead antenna wire.

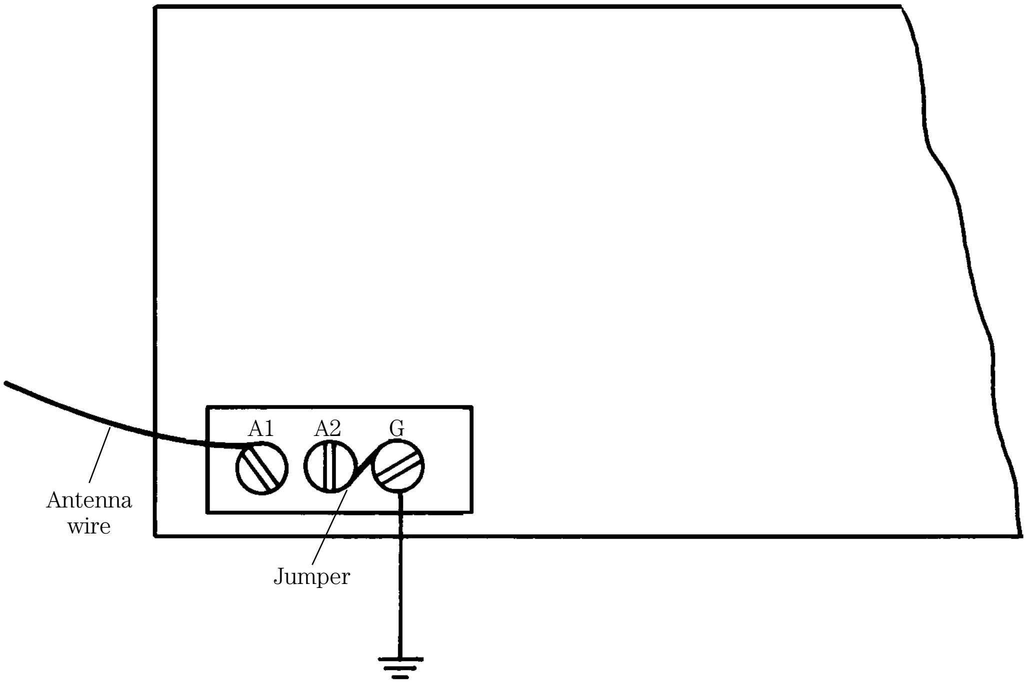

13-3A Connection of wire antenna to balanced antenna terminals on receiver.

On receivers that use an SO-239 coaxial connector, we can use either of two techniques to connect a single-lead wire. First, we can obtain the mating PL-259 plug, and solder the antenna lead to the center conductor pin. The PL-259 connector is then screwed into the mating SO-239 chassis connector. Regardless of the type of coaxial connector, however, the mate can be used for the antenna lead wire. But for SO-239 connectors another alternative is also available. Figure 13-3B shows a "banana plug" attached to the lead wire and inserted into the receptacle of the SO-239.

DANGER!

Certain low-cost receivers, especially older vacuum-tube models, have a so-called ac/dc or transformerless internal dc power supply. On most receivers, the dc common is the chassis, which also serves as the RF signal common. But on ac/dc models the neutral wire of the ac power line serves as the dc common, and it is kept floating as a "counterpoise" ground above the chassis ground used by the RF signals. A capacitor (C_1 in Fig. 13-4) sets the chassis and counterpoise ground at 0-V RF potential, while keeping the counterpoise isolated for dc and 60 Hz ac. A danger exists if either the ac plug is installed backwards or someone plugs the socket in the wall incorrectly (often happens!). Even if C_1 is intact, *a nasty shock can be felt by touching the antenna ground* (G or GND) *terminal.* The capacitive reactance of C_1 is about 2.7 MΩ for 60-Hz ac, so at least a "bite" is going to happen. But if that capacitor is shorted, which is likely on older receivers, then the bite is considerably worse, and might even prove *fatal.* The problem, in that case, is that reversed ac line polarity will set the hot line from the ac socket on the ground lead. The least to expect is massive fireworks and a possible fire hazard; the most to expect is a fire, and your possible electrocution.

The usual advice given to owners of such radios is to make sure that C_1 is intact before using the radio. I prefer a better solution: buy, install, and use a 120:120 Vac isolation transformer to isolate your receiver from the ac power lines. Such a transformer is standard practice in repair shops, and it should also be standard practice in your house.

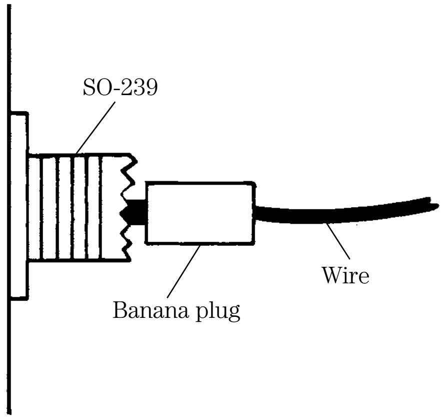

13-3B Connection of wire antenna to coaxial input receiver.

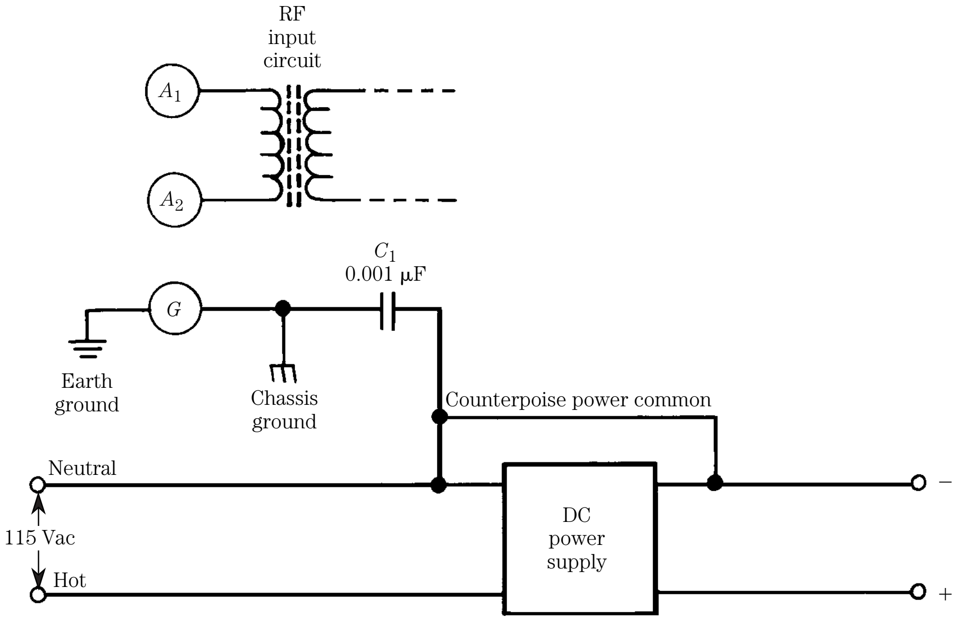

13-4 Dangerous form of antenna/receiver circuitry.

Wire antennas

This section reviews simple wire antennas that are suitable for the reception of shortwave signals, although not necessarily for transmitting. Once again, you are reminded that the law of reciprocity permits you to use any transmitting antenna found in other chapters for receiving also.

Figure 13-5 shows the common receiving longwire. The antenna element should be 30 to 150 ft in length. Although most texts show it horizontal to the ground (and indeed, a case can be made that performance is better that way), it is not strictly necessary. If you must slope the wire, then it is doubtful that you will notice any reception problems.

The far end of the wire is attached to a supporting structure through an insulator and a rope. The support structure can be another building, a tree, or a mast installed especially for this purpose. Chapter 28 deals with antenna construction practices.

Wind will cause motion in the antenna wire and its supporting structure. Over time, the wind movement will fatigue the antenna wire and cause it to break. Also, if a big enough gust of wind (or a sustained storm) comes along, then even a new antenna can either sag badly or break altogether. You can do either of two things to reduce the problem. First, as shown in Fig. 13-5, a door spring can be used to provide a little variable slack in the wire. The spring tension is selected to be only partially expanded under normal conditions, so wind will increase the tension, and stretch the spring. Make sure that the spring is not too strong to be stretched by the action of wind on the antenna, or no good is accomplished.

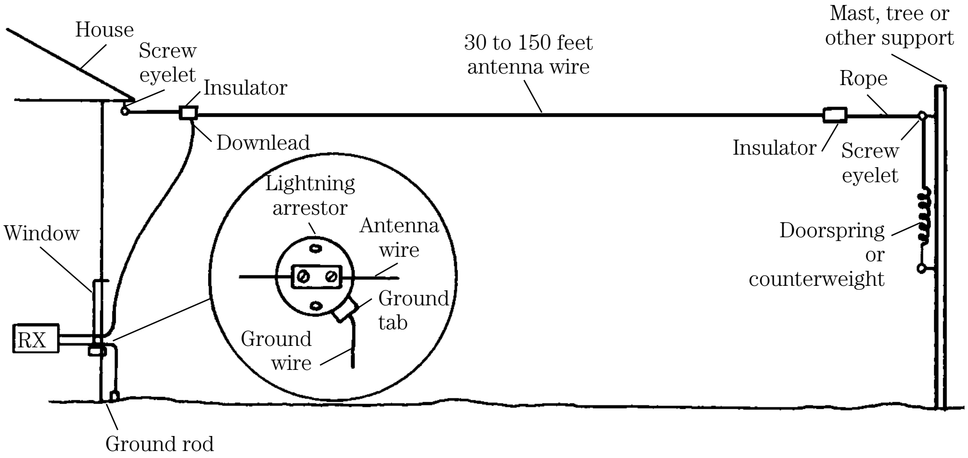
House
Screw eyelet
Insulator
30 to 150 feet antenna wire
Mast, tree or other support
Rope
Downlead
Insulator
Screw eyelet
Lightning arrestor
Antenna wire
Window
Doorspring or counterweight
Ground tab
Ground wire
RX
Ground rod

13-5 Longwire SWL antenna.

Another tactic is to replace the spring with a counterweight that is heavy enough to keep the antenna nearly taut under normal conditions, but not so heavy that it fails to move under wind conditions. In other words, the antenna tension should exactly balance the counterweight under normal conditions, and not be too great that it stretches the antenna wire excessively.

The antenna wire should be either no. 12 or no. 14 hard-drawn copper, or *Copperweld* stranded wire. The latter is actually steel-core wire, but has a copper coating on the outside. Because of skin effect, RF signals flow only in the outer copper coating. Soft-drawn copper wire will stretch and break prematurely, so it should be avoided.

The downlead of the antenna must be insulated, and it should also be stranded wire (which breaks less easily than solid wire). Again, no. 12 or no. 14 wire should be used, although no. 16 would be permissible. The point where the downlead and antenna are joined should be soldered to prevent corrosion of the joint. Mechanical strength is provided by proper splicing technique (see Fig. 13-6). Do not depend on the solder for mechanical strength, for it has none.

There are several ways to bring a downlead into the building. First, if you can tolerate a slight crack in the junction of the sash and sill, then run the wire underneath the sash and close the window. Alternatively, you can buy a flat strap connector to pass under the window. This method is electrically the same as running the lead, but is mechanically nicer. Chapter 28 deals with several methods, and should be consulted.

Grounding

The ground lead should be a heavy conductor, such as heavy wire or braid. The shield stripped from RG-8 or RG-11 coaxial cable is suitable for most applications. For reception purposes only, the ground may be a cold-water pipe inside the house. Do not use the hot-water pipes (which are not well grounded) or gas pipes (which

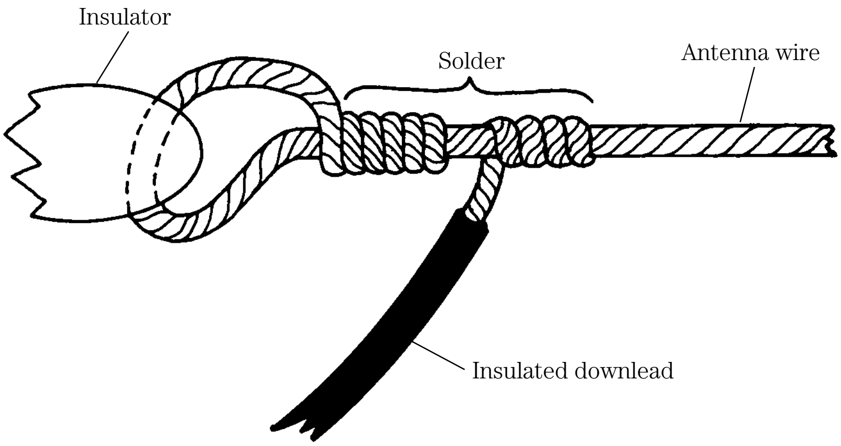

13-6 Construction details.

are dangerous to use). Also, be aware that residential air conditioner liquid lines look like copper cold-water pipes in some cases. Don't use them. Chapter 29 deals with grounding.

The lightning arrester is a safety precaution, and it *must* be used! Its purpose is to supply an alternative path to ground in the event of a lightning strike. Although at least one text calls the arrester optional, it is not. Besides the obvious safety reasons (which are reason enough), there are also legal and economic reasons for using the arrester. Your local government building and/or fire codes might require a lightning arrester for outdoor antennas. Also, your insurance company might not honor your homeowner's policy if the lightning arrester (required by local code) is not used. The antenna lightning arrester is *not optional,* so use it.

WARNING! DO NOT EVER ATTEMPT TO INSTALL AN ANTENNA BY CROSSING A POWER LINE! **EVER!** NO MATTER WHAT YOU BELIEVE OR WHAT YOUR FRIENDS TELL YOU, IT'S NEVER SAFE AND IT COULD KILL YOU.

The rest of this discussion touches on antennas, other than the receiver longwire. Because construction details are similar, we will not repeat them. You are, by the way, encouraged to also read Chap. 28 to glean more details on antenna construction methods.

The flattop antenna

The "flattop" antenna is shown in Fig. 13-7. This antenna is a relative of the longwire, with the exception that the downline is at the approximate center of the antenna section. The flattop should be at least a half-wavelength ($492/F_{\mathrm{MHz}}$) at the lowest frequency of operation. The advantage of the flattop is that it allows the maximum use of space in the configuration shown.

Vertical antennas

It is also possible to build shortwave listener antennas in vertical polarization. Figure 13-8 shows one version of this type of antenna. The support (which could

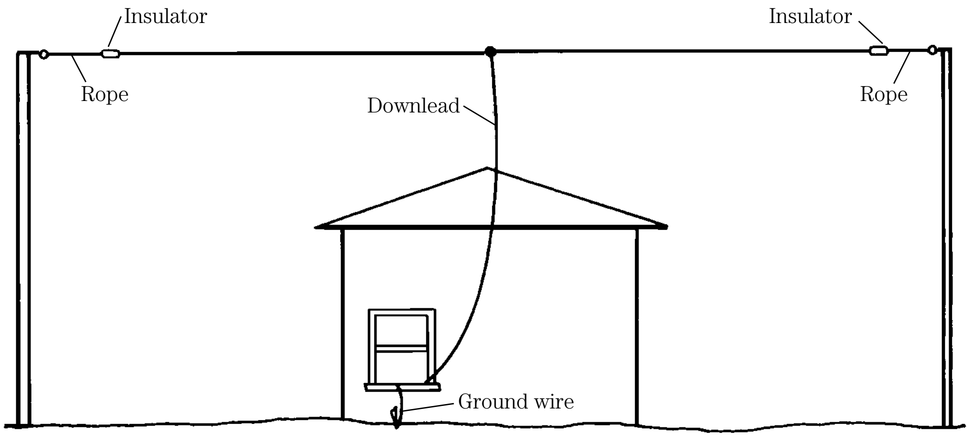
Insulator
Insulator
Rope
Downlead
Rope
Ground wire

13-7 Flattop SWL antenna.

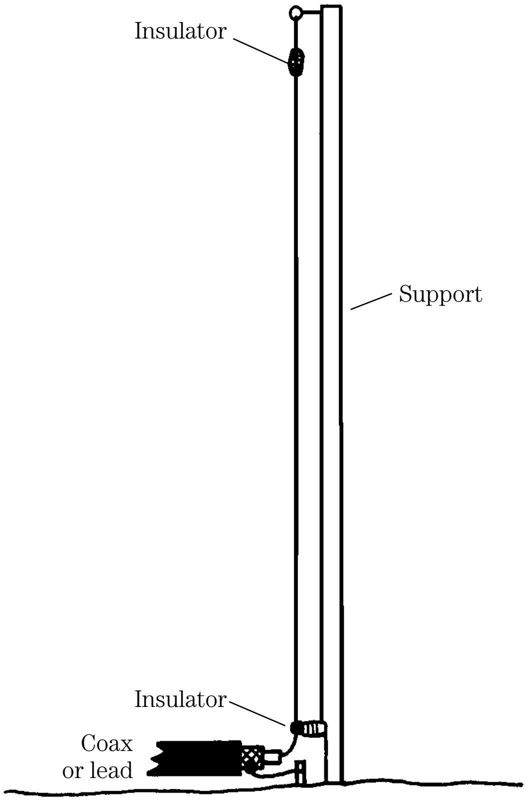
Insulator
Support
Insulator
Coax
or lead

13-8 Wire vertical SWL antenna—works for hams too.

be a tree or building) should have enough height to be at least a quarter-wavelength on the lowest frequency of operation. The antenna is fed at the base with coaxial cable. The center conductor of the coax is connected to the antenna element, and the shield is connected to the ground rod at the base of the structure. You are encouraged to see Chap. 10, which deals with limited space and hidden antennas, for a different version of this antenna. It is possible to install the wire (or multiple wires of different lengths) inside of a length of PVC plumbing pipe. The pipe serves as the support structure, and the conductors are placed inside.

Different lengths of conductor are required for different bands of operation. You can calculate the length (in feet) required for quarter-wavelength vertical antennas from $L_{\mathrm{ft}}=246/F_{\mathrm{MHz}}$. Figure 13-9A shows how several bands are accommodated from the same feedline on the same support structure. In this particular case eight different antenna elements are supported from the same tee bar. Be sure to insulate them from each other, and from the support structure. Again, PVC piping can be used for the support. Another method for accommodating several different bands is to tie the upper ends of the wires to a sloping rope (as in Fig. 13-9B).

Wire directional antenna

A directional antenna has the ability to enhance reception of desired signals, while rejecting undesired signals arriving from slightly different directions. Although directivity normally means a beam antenna, or at least a rotatable dipole, there are certain types of antenna that allow fixed antennas to be both directive and variable. See Chap. 7 for fixed but variable directional antennas and Chap. 11 for fixed and non-variable directive arrays. Those antennas are transmitting antennas, but they work equally well for reception. This section shows a crude, but often effective, directional antenna that allows one to select the direction of reception with pin plugs or switches.

Consider Fig. 13-10. In this case, a number of quarter-wavelength radiators are fanned out from a common feedpoint at various angles from the building. At the near end of each element is a female banana jack. A pair of balanced feedlines from the receiver (300-Ω twin lead, or similar) are brought to the area where the antenna elements terminate. Each wire in the twin lead has a banana plug attached. By selecting which banana jack is plugged into which banana plug, you can select the directional pattern of the antenna. If the receiver is equipped with a balanced antenna input, then simply connect the other end of the twin lead direction to the receiver. Otherwise, use one of the couplers shown in Fig. 13-11.

Figure 13-11A shows a balanced antenna coupler that is tuned to the frequency of reception. The coil is tuned to resonance by the interaction of the inductor and the capacitor. Antenna impedance is matched by selecting the taps on the inductor to which the feedline is attached. A simple RF broadband coupler is shown in Fig. 13-11B. This transformer is wound over a ferrite core, and consists of 12 to 24 turns of no. 26 enameled wire, with more turns being used for lower frequencies, and fewer for higher frequencies. Experiment with the number of turns in order to determine the correct value. Alternatively, use a 1:1 balun transformer instead of Fig. 13-11B; the type intended for amateur radio antennas is overkill powerwise, but it will work nicely.

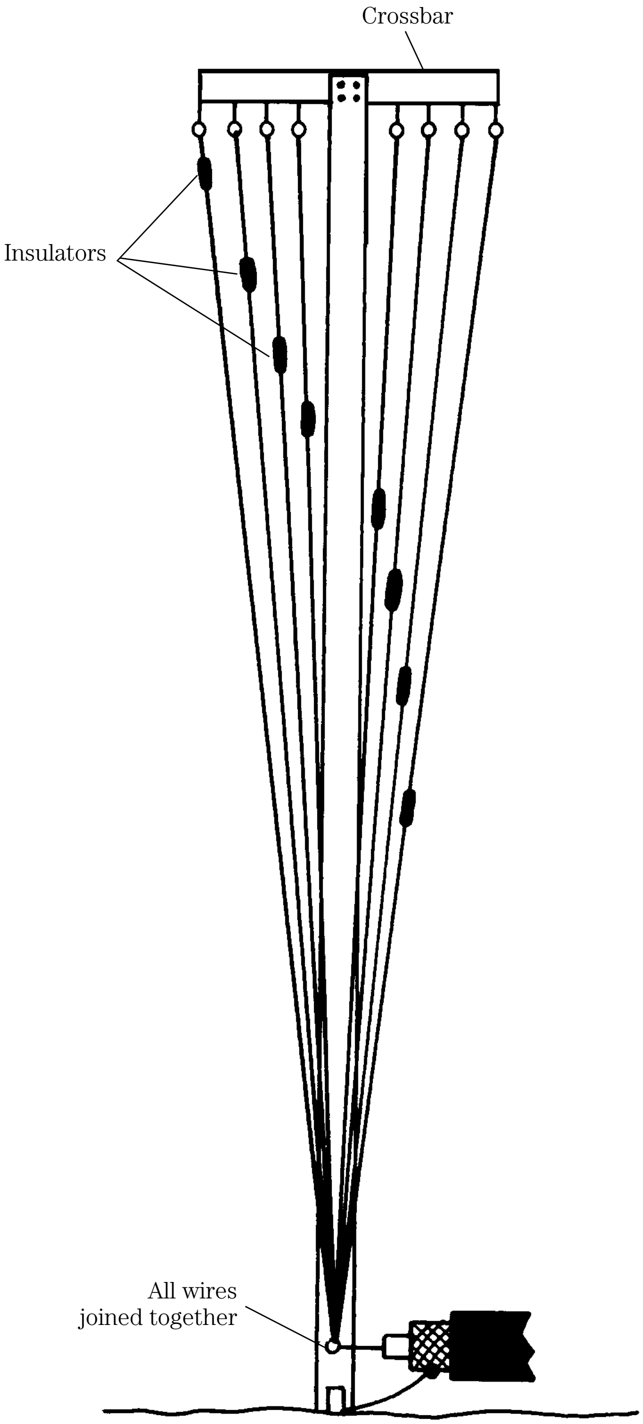
Crossbar
Insulators
All wires
joined together

13-9A Multiband wire vertical SWL antenna: tee bar style.

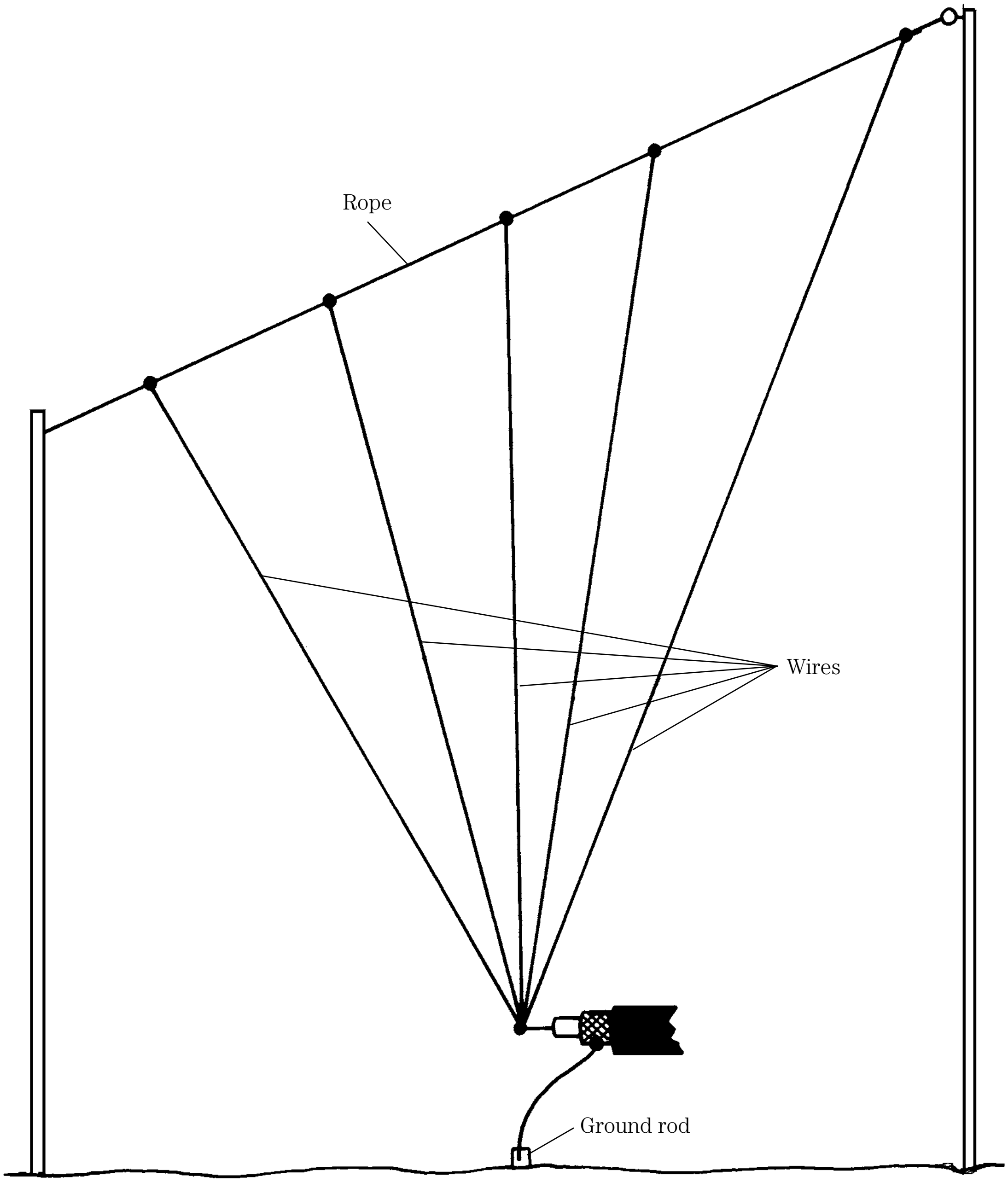
Rope
Wires
Ground rod

13-9B Multiband wire vertical SWL antenna: fan style.

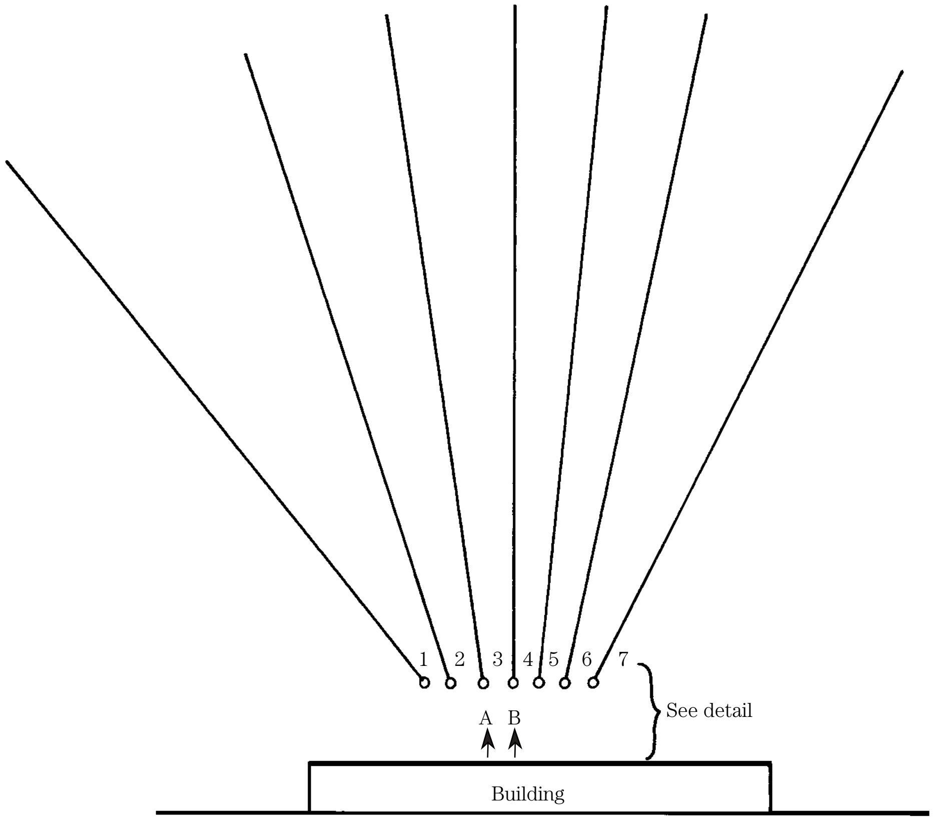

13-10 Directional SWL antenna (view from above).

The antenna of Fig. 13-10 works by phasing the elements so as to null, or enhance (as needed), certain directions. This operation becomes a little more flexible if you build a phasing transformer, as shown in Fig. 13-11C and 13-11D. Windings L_1, L_2, and L_3 are wound “trifilar” style onto a ferrite core. Use 14 turns of no. 26 enameled wire for each winding. The idea in this circuit is to feed one element from coil L_2 in the same way all of the time. This port becomes the 0° phase reference. The other port, B, is fed from a reversible winding, so it can either be in phase or 180° out of phase with port A. Adjust the DPDT switch and the banana plugs of Fig. 13-10 for the best reception.

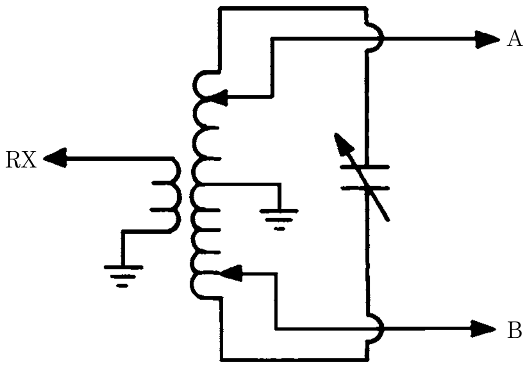

13-11A Antenna matching tuner.

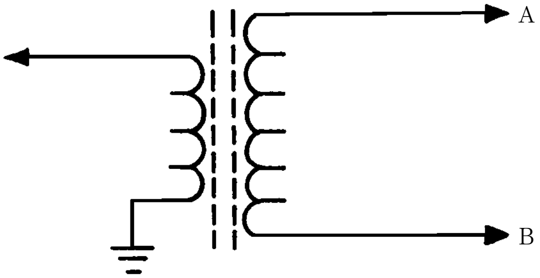

13-11B Antenna matching transformer.

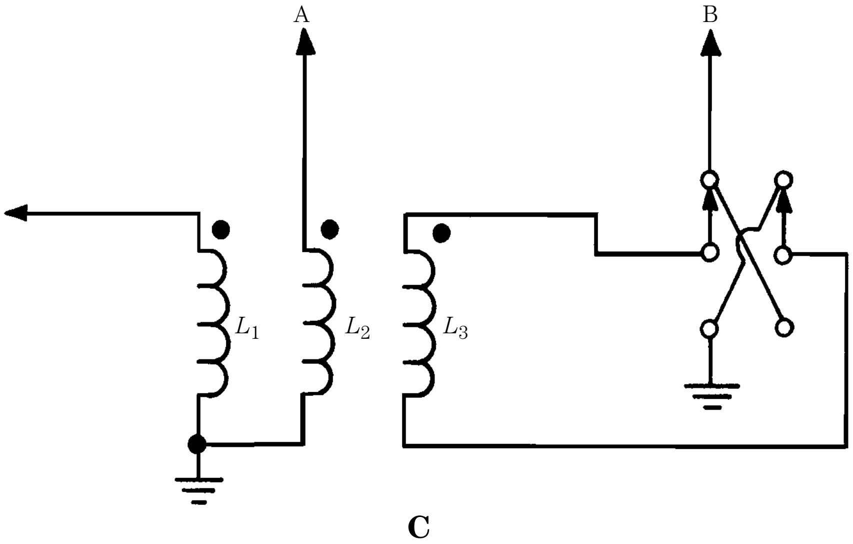

13-11C Phase-switching antenna matcher.

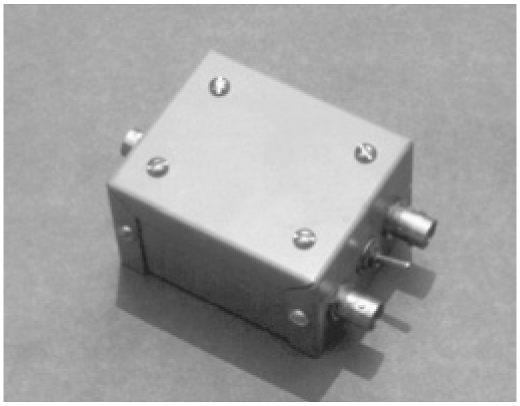

13-11D Phasing box permits phasing antennas 0° to 180°.

14
CHAPTER
Large wire loop antennas

THERE ARE TWO BASIC FORMS OF LOOP ANTENNAS: SMALL AND LARGE. THESE TWO TYPES have different characteristics, work according to different principles, and have different purposes. Small loops are those in which the current flowing in the wire has the same phase and amplitude at every point in the loop (which fact implies a very short wire length, i.e., less than 0.2λ). Such loops respond to the magnetic field component of the electromagnetic radio wave. A large loop antenna has a wire length greater than 0.2λ, with most being either λ/2, 1λ, or 2λ. The current in a large loop varies along the length of the wire in a manner similar to other wire antennas.

λ/2 large loops

The performance of large wire loop antennas depends in part on their size. Figure 14-1 shows a half-wavelength loop (i.e., one in which the four sides are each λ/8 long). There are two basic configurations for this antenna: *continuous* (S_1 closed) and *open* (S_1 open). In both cases, the feedpoint is at the midpoint of the side opposite the switch.

The direction of the main reception, or radiation lobe (i.e., the direction of maximum reception), depends on whether S_1 is open or closed. With S_1 closed, the main lobe is to the right (solid arrow); and with S_1 open, it is to the left (broken arrow). Direction reversal can be achieved by using a switch (or relay) at S_1, although some people opt for unidirectional operation by eliminating S_1, and leaving the loop either open or closed.

The feedpoint impedance is considerably different in the two configurations. In the closed-loop situation (i.e., S_1 closed), the antenna can be modeled as if it were a half-wavelength dipole bent into a square and fed at the ends. The feedpoint (X_1-X_2) impedance is on the order of 3 kΩ because it occurs at a voltage antinode (current node). The current antinode (i.e., I_{max}) is at S_1, on the side opposite the feedpoint. An antenna tuning unit (ATU), or RF impedance transformer, must be used to match the lower impedance of the transmission lines needed to connect to receivers.

The feedpoint impedance of the open-loop configuration (S_1 open) is low because the current antinode occurs at X_1-X_2. Some texts list the impedance as "about 50 Ω," but my own measurements on several test loops were somewhat higher (about 70 Ω). In either case, the open loop is a reasonable match for either 52- or 75-Ω coaxial cable.

Copyright 2001 - the McGraw-Hill Companies

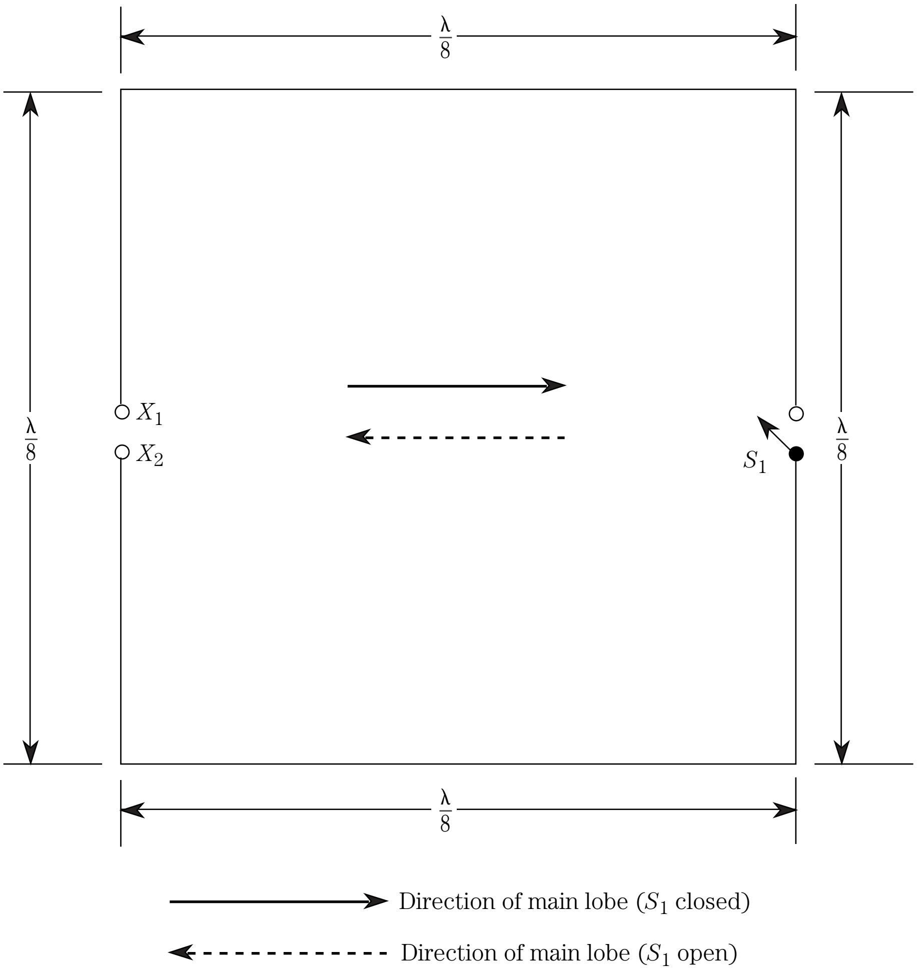
$\frac{\lambda}{8}$
$\frac{\lambda}{8}$
$\frac{\lambda}{8}$
$\frac{\lambda}{8}$
X_1
X_2
S_1
Direction of main lobe (S_1 closed)
Direction of main lobe (S_1 open)

14-1 Half-wave square loop antenna with reversal switch.

Neither λ/2 loop configuration shows gain over a dipole. The figure usually quoted is −1-dB forward gain (i.e., a loss compared with a dipole), and about 6-dB front-to-back ratio (FBR). Such low values of FBR indicate that there is no deep notch ("null") in the pattern.

A lossy antenna with a low FBR seems like a born loser, and in most cases it is. But the λ/2 loop finds a niche where size must be constrained, for one reason or another. In those cases, the λ/2 loop can be an alternative. These antennas can be considered limited-space designs, and can be mounted in an attic, or other limited-access place, as appropriate.

A simple trick will change the gain, as well as the direction of radiation, of the closed version of the λ/2 loop. In Fig. 14-2, a pair of inductors (L_1 and L_2) are inserted into the circuit at the midpoints of the sides adjacent to the side containing the feedpoints. These inductors should have an inductive reactance X_L of about 360 Ω in the center of the band of operation. The inductance of the coil is

$$L_{\mu H} = \frac{3.6 \times 10^8}{2\pi F_{Hz}} \qquad [14.1]$$

where

$L_{\mu H}$ is the coil inductance, in microhenrys (µH)

F_{Hz} is the midband frequency, in hertz (Hz)

Example Find the inductance for the coils in a loaded half-wavelength closed-loop antenna that must operate in a band centered on 10.125 MHz.

Solution:

[*Note:* 10.125 MHz = 10,125,000 Hz]

$$L_{\mu H} = \frac{3.6 \times 10^8}{2\pi\ (10{,}125{,}000\ \text{Hz})} = 5.7\ \mu\text{H}$$

The coils force the current antinodes toward the feedpoint, reversing the direction of the main lobe, and creating a gain of about +1 dB over a half-wavelength dipole.

The currents flowing in the antenna can be quite high, so when making the coils, be sure to use a size that is sufficient for the power and current levels anticipated. The 2- to 3-in B&W Air-Dux style coils are sufficient for most amateur radio use. Smaller coils are available on the market, but their use is limited to low-power situations.

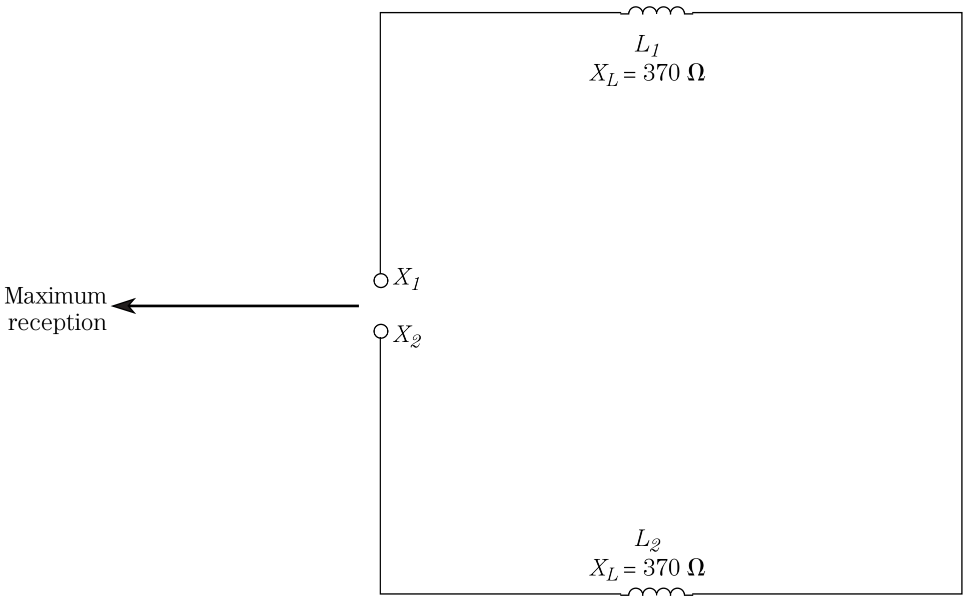

14-2 Use of inductive loading to reduce the size of antenna, and make the pattern more uniform.

1λ large loops

If size is not forcing you to a λ/2 loop, then a 1λ loop might be just the ticket. It produces a gain of about +2 dB over a dipole in the directions that are perpendicular to the plane of the loop. The azimuth patterns formed by these antennas are similar to the figure-8 pattern of the dipole. Three versions are shown: the square loop (Fig. 14-3), the diamond loop (Fig. 14-4), and the delta loop (a.k.a. D-loop and triangle—Fig. 14-5). The square and diamond loops are built with λ/4 on each side, and the delta loop is λ/3 on each side. The overall length of wire needed to build these antennas is

$$L_{\text{feet}} = \frac{1005}{F_{\text{MHz}}} \tag{14.2}$$

The polarization of the three loop antennas is horizontal, because of the location of the feedpoints. On the square loop, moving the feedpoint to the middle of either vertical side will provide vertical polarization. Similarly, on the diamond loop vertical polarization is realized by moving the feedpoint to either of the two adjacent apexes. On the delta loop, placing the feedpoint at either of the two other apexes produces a diagonal polarization that offers approximately equal vertical and horizontal polarization components.

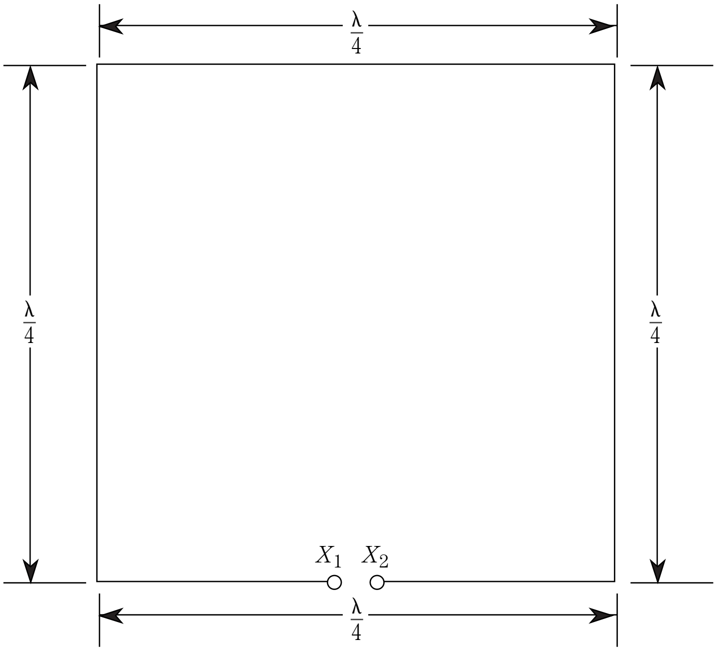

14-3 Quarter-wavelength square loop (single-element quad).

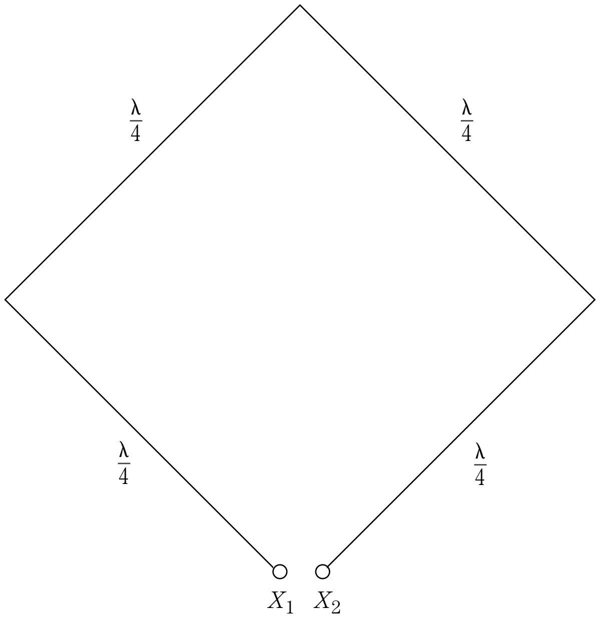
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
$\frac{\lambda}{4}$
X_1
X_2

14-4 Bottom-fed loop.

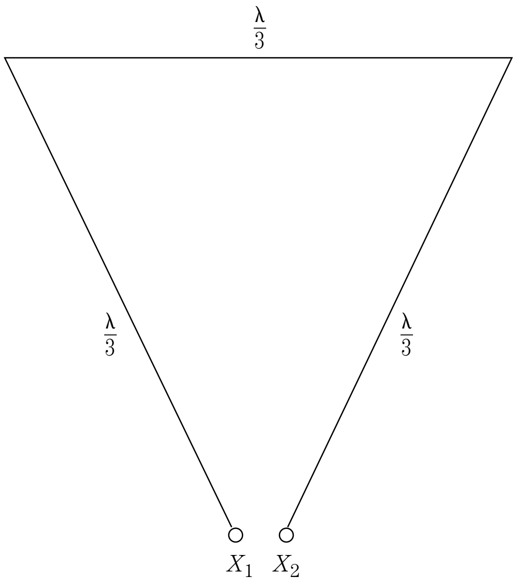
$\frac{\lambda}{3}$
$\frac{\lambda}{3}$
$\frac{\lambda}{3}$
X_1
X_2

14-5 Delta loop antenna.

The feedpoint impedance of the 1λ loop is around 100 Ω, so it provides a slight mismatch to 75-Ω coax and a 2:1 mismatch to 52-Ω coax. A very good match to 52-Ω coax can be produced using the scheme of Fig. 14-6. Here, a quarter-wavelength coaxial cable matching section is made of 75-Ω coaxial cable. The length of this cable should be

$$L_{\text{feet}} = \frac{246V}{F_{\text{MHz}}} \qquad \textbf{[14.3]}$$

where

L_{feet} is the length, in feet (ft)

V is the velocity factor of the coax

F_{MHz} is the frequency, in megahertz (MHz)

The impedance Z_o of the cable used for the matching section should be

$$Z_o = \sqrt{Z_L Z_s} \qquad \textbf{[14.4]}$$

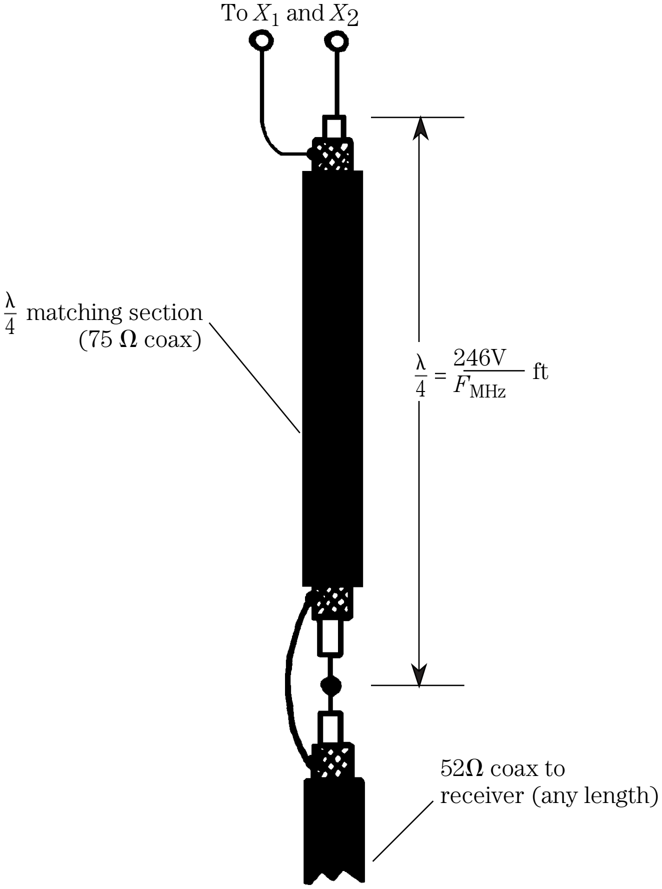
To X_1 and X_2
$\frac{\lambda}{4}$ matching section (75 Ω coax)
$\frac{\lambda}{4} = \frac{246V}{F_{\text{MHz}}}$ ft
52Ω coax to receiver (any length)

14-6 Quarter-wavelength coaxial matching section.

where

Z_o is the characteristic impedance of the coax used in the matching section, in ohms
Z_L is the feedpoint impedance of the antenna, in ohms
Z_s is the source impedance (i.e., the 52-Ω characteristic impedance of the line to the receiver in standard systems)

When Eq. 14.4 is applied to this system, where $Z_s = 52\ \Omega$, $Z_L = 100\ \Omega$:

$$Z_o = \sqrt{(100\ \Omega)(52\ \Omega)} = 72\ \Omega \qquad \textbf{[14.5]}$$

This is a very good match to 75-Ω coaxial cable.

Demiquad loop antenna

The *demiquad* is a single-element 1λ quad antenna. The length of the antenna is, like the cubical quad beam antenna (see Chap. 12), one wavelength. Figure 14-7 shows a type of demi-quad based on the tee-cross type of mast.

The impedance-matching section is a quarter-wavelength piece of 75-Ω coaxial cable (RG-58/U or RF-11/U). The length of the matching section is determined from:

$$L = \frac{246V}{F_{\text{MHz}}} \qquad \textbf{[14.6]}$$

where

L is the overall length, in feet
F_{MHz} is the frequency, in megahertz
V is the velocity factor of the coaxial cable (typically 0.66, 0.70, or 0.80)

Delta loop

The *delta loop antenna,* like the Greek uppercase letter “delta” (Δ) from which it draws its name, is triangle-shaped (Fig. 14-8). The delta loop is a full wavelength, with elements approximately 2 percent longer than the natural wavelength (like the quad). The actual length will be a function of the proximity and nature of the underlying ground, so some experimentation is necessary. The approximate preadjustment lengths of the sides are found from:

$$L_1 = \frac{437}{F_{\text{MHz}}}\ \text{ft} \qquad \textbf{[14.7]}$$

$$L_2 = L_3 = \frac{296}{F_{\text{MHz}}} \qquad \textbf{[14.8]}$$

The delta loop antenna is fed from 52-Ω coaxial cable through a 4:1 balun transformer. The delta loop can be built in a fixed location, and will offer a bidirectional pattern.

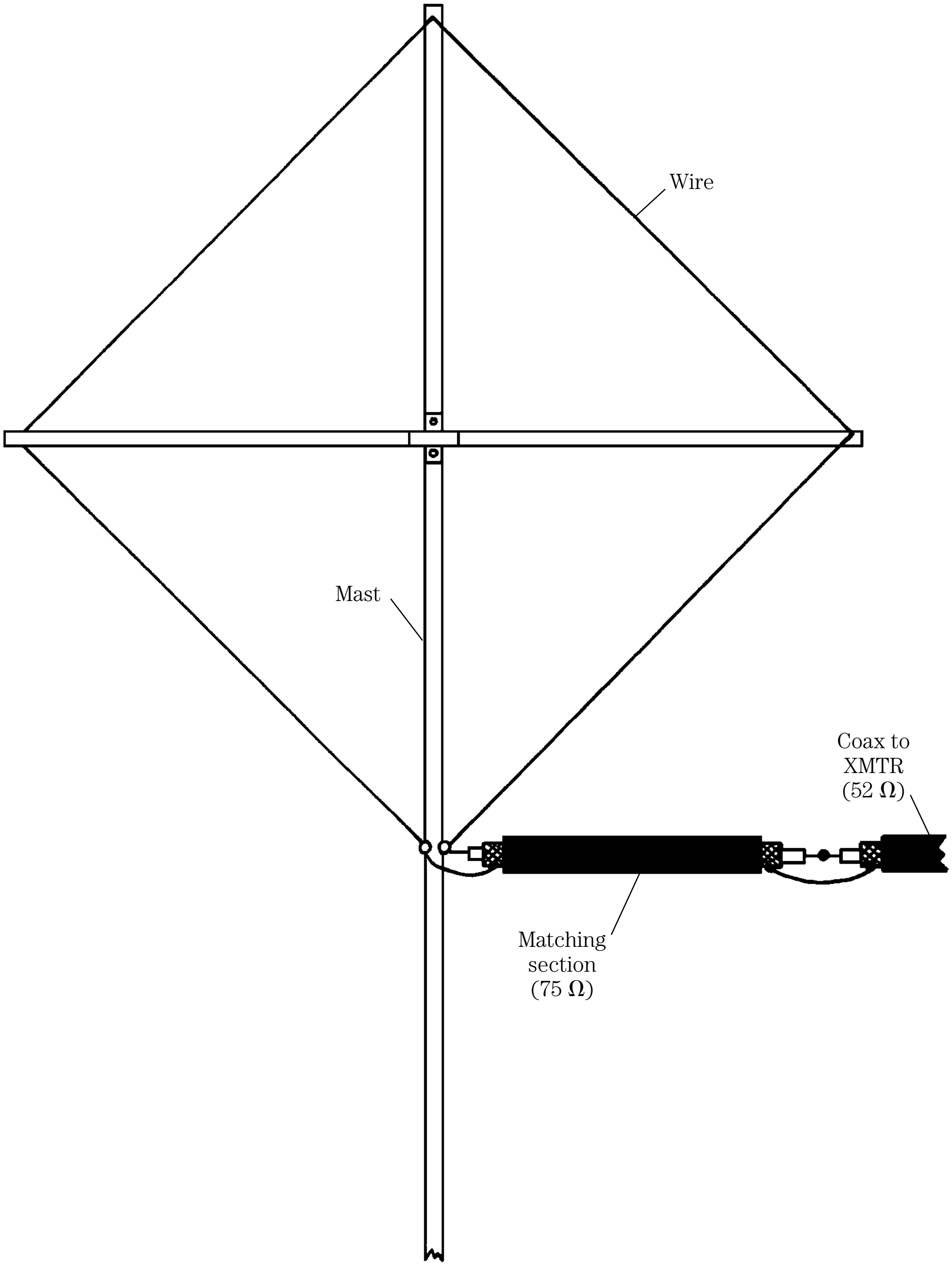
Wire
Mast
Coax to
XMTR
(52 Ω)
Matching
section
(75 Ω)

14-7 Demiquad antenna.

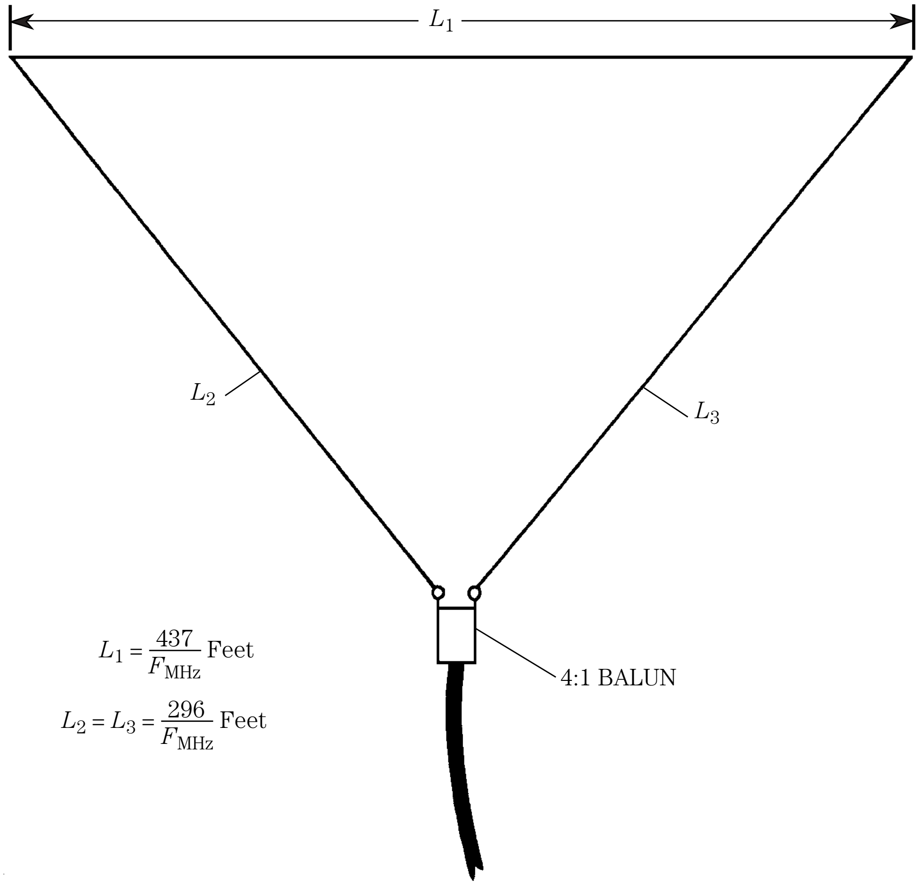
L_1
L_2
L_3
$L_1 = \frac{437}{F_{\text{MHz}}}$ Feet
$L_2 = L_3 = \frac{296}{F_{\text{MHz}}}$ Feet
4:1 BALUN

14-8 Delta loop antenna.

Half-delta sloper (HDS)

The *half-delta sloper* (HDS) antenna (Fig. 14-9) is similar to the full delta loop, except that (like the quarter-wavelength vertical) half of the antenna is in the form of an “image” in the ground. Gains of 1.5 to 2 dB are achievable. The HDS antenna consists of two elements: a λ/3-wavelength sloping wire and a λ/6 vertical wire (on an insulated mast), or a λ/6 metal mast. Because the ground currents are very important, much like the vertical antenna, either an extensive radial system at both ends is needed, or a base ground return wire (buried) must be provided.

The HDS will work on its design frequency, plus harmonics of the design frequency. For a fundamental frequency of 5 MHz, a vertical segment of 33 ft and a sloping section of 66 ft is needed. The lengths for any frequency are found from

$$\frac{\lambda}{3} = \frac{328}{F_{\text{MHz}}} \tag{14.9}$$

and

$$\frac{\lambda}{6} = \frac{164}{F_{\mathrm{MHz}}} \qquad \textbf{[14.10]}$$

The HDS is fed at one corner, close to the ground. If only the fundamental frequency is desired, then you can feed it with 52-Ω coaxial cable. But at harmonics, the feedpoint impedance changes to as high as 1000 Ω. If harmonic operation is intended, then an antenna tuning unit (ATU) is needed at point *A* to match these impedances.

Bisquare loop antenna

The *bisquare antenna,* shown in Fig. 14-10, is similar to the other large loops, except that it is λ/2 on each side, making a total wire length of two wavelengths. This antenna is built like the diamond loop shown earlier (i.e., it is a large square loop fed at an apex that is set at the bottom of the assembly). In this case, the loop is fed either with an antenna tuning unit (to match a 1000-Ω impedance) or a quarter-wavelength matching section made of 300-Ω or 450-Ω twin-lead transmission line. A 1:1 balun transformer connects the 75-Ω coaxial cable to the matching section.

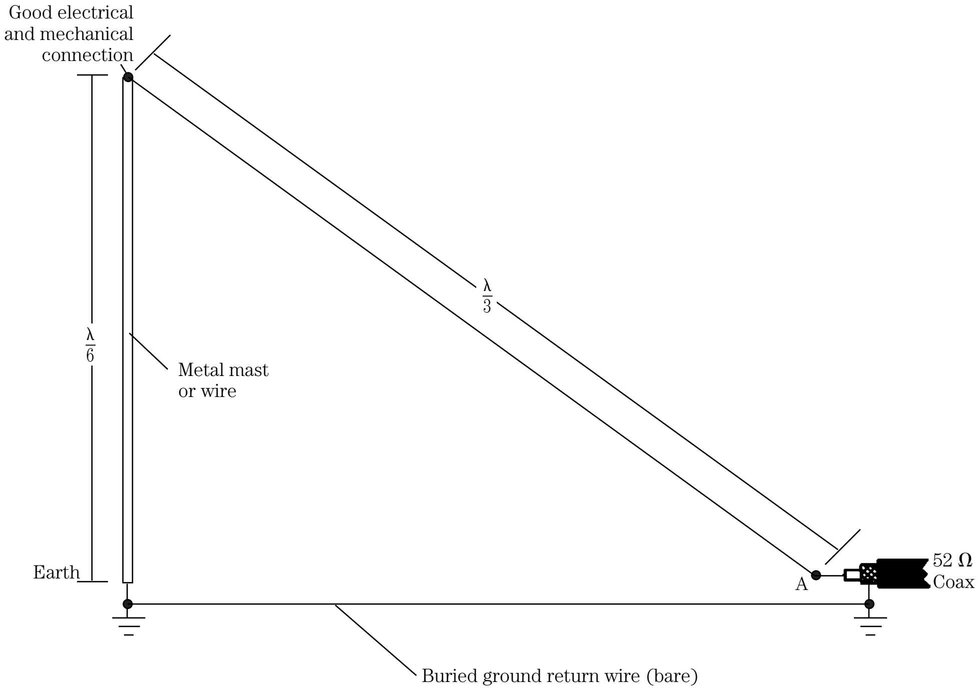
Good electrical
and mechanical
connection
λ/3
λ/6
Metal mast
or wire
Earth
A
52 Ω
Coax
Buried ground return wire (bare)

14-9 Half-delta loop antenna.

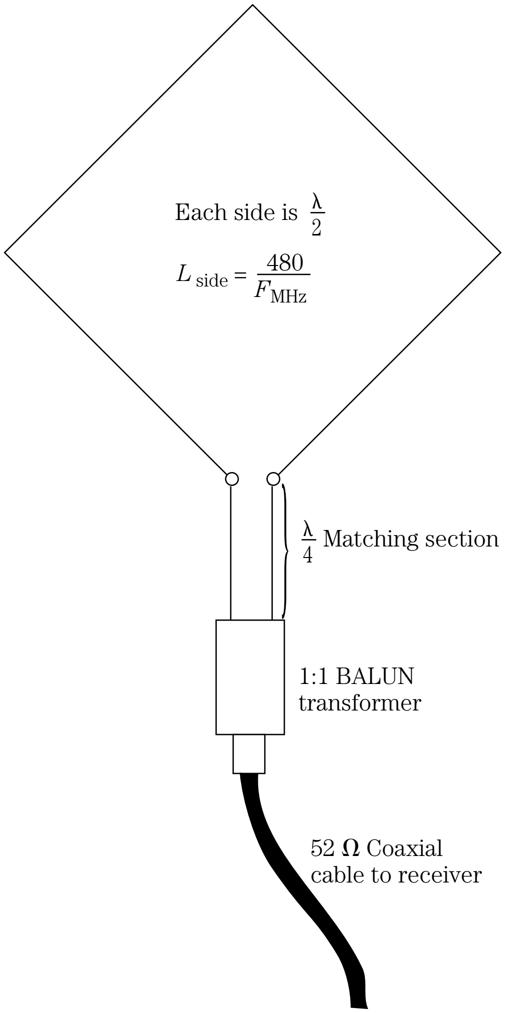
Each side is $\frac{\lambda}{2}$
$L_{\text{side}} = \frac{480}{F_{\text{MHz}}}$
$\frac{\lambda}{4}$ Matching section
1:1 BALUN transformer
52 Ω Coaxial cable to receiver

14-10 Bisquare 2λ square loop antenna.

The bisquare antenna offers as much as 4-dB gain broadside to the plane of the antenna (i.e., in and out of the book page), in a figure-8 pattern, on the design frequency. It is horizontally polarized. When the frequency drops to one-half of the design frequency, the gain drops to about 2 dB, and the antenna works like the diamond loop covered previously.

This page intentionally left blank

15
Chapter

Small loop receiving antennas

Radio direction finders and people who listen to the AM broadcasting bands, VLF, medium-wave, or the so-called low-frequency tropical bands are all candidates for a small loop antenna. These antennas are fundamentally different from large loops and other sorts of antennas used in these bands. Large loop antennas have a length of at least 0.5λ, and most are quite a bit larger than 0.5λ. Small loop antennas, on the other hand, have an overall length that is less than 0.22λ, with most being less than 0.10λ.

The small loop antenna responds to the magnetic field component of the electromagnetic wave instead of the electrical field component. One principal difference between the large loop and the small loop is found when examining the RF currents induced in a loop when a signal intercepts it. In a large loop, the current will vary from one point in the conductor to another, with voltage varying out of phase with the current. In the small loop antenna, the current is the same throughout the entire loop.

The differences between small loops and large loops show up in some interesting ways, but perhaps the most striking is the directions of maximum response—the main lobes—and the directions of the nulls. Both types of loops produce figure-8 patterns but in directions at right angles with respect to each other. The large loop antenna produces main lobes *orthogonal,* at right angles or "broadside," to the plane of the loop. Nulls are off the sides of the loop. The small loop, however, is exactly the opposite: The main lobes are off the sides of the loop (in the direction of the loop plane), and the nulls are broadside to the loop plane (Fig. 15-1A). Do not confuse small loop behavior with the behavior of the loopstick antenna. Loopstick antennas are made of coils of wire wound on a ferrite or powdered-iron rod. The direction of maximum response for the loopstick antenna is broadside to the rod, with deep nulls off the ends (Fig. 15-1B). Both loopsticks and small wire loops are used for radio direction-finding and for shortwave, low-frequency medium-wave, AM broadcast band, and VLF listening.

The nulls of a loop antenna are very sharp and very deep. Small changes of pointing direction can make a profound difference in the response of the antenna. If you point a loop antenna so that its null is aimed at a strong station, the signal strength of

Copyright 2001 - the McGraw-Hill Companies

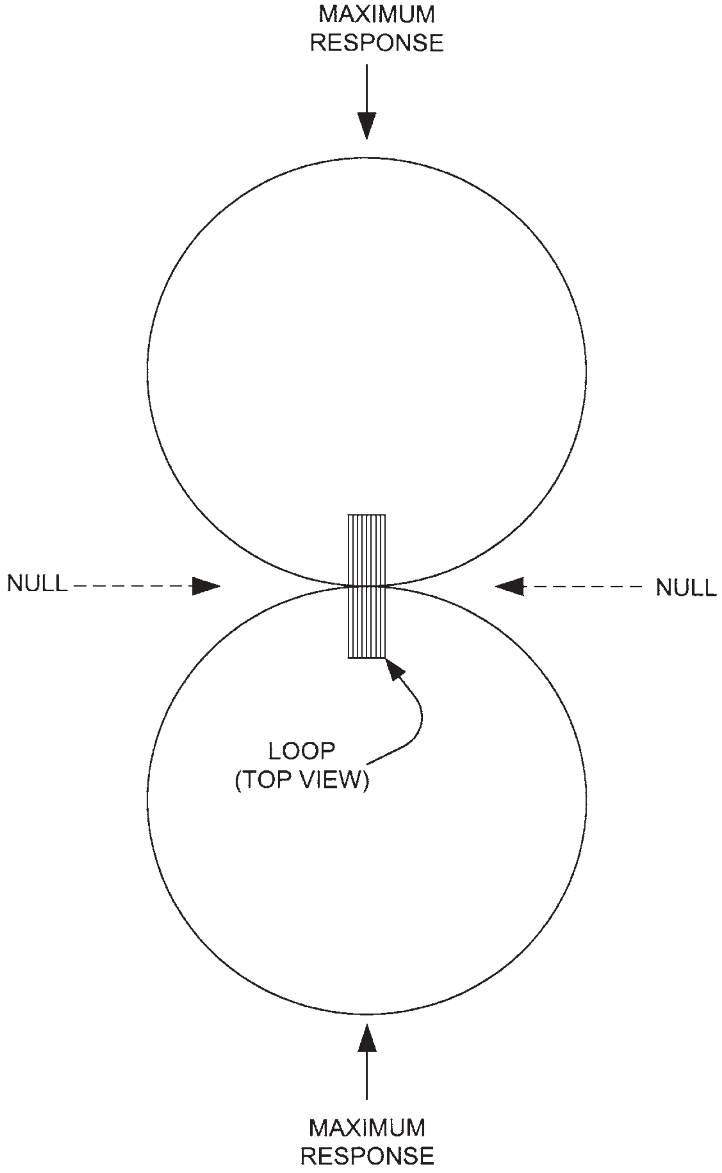

15-1A Large loop antenna.

the station appears to drop dramatically at the center of the notch. Turn the antenna only a few degrees one way or the other, however, and the signal strength increases sharply. The depth of the null can reach 10 to 15 dB on sloppy loops and 30 to 40 dB on well-built loops (30 dB is a very common value). I have seen claims of 60-dB nulls for some commercially available loop antennas. The construction and uniformity of the loop are primary factors in the sharpness and depth of the null.

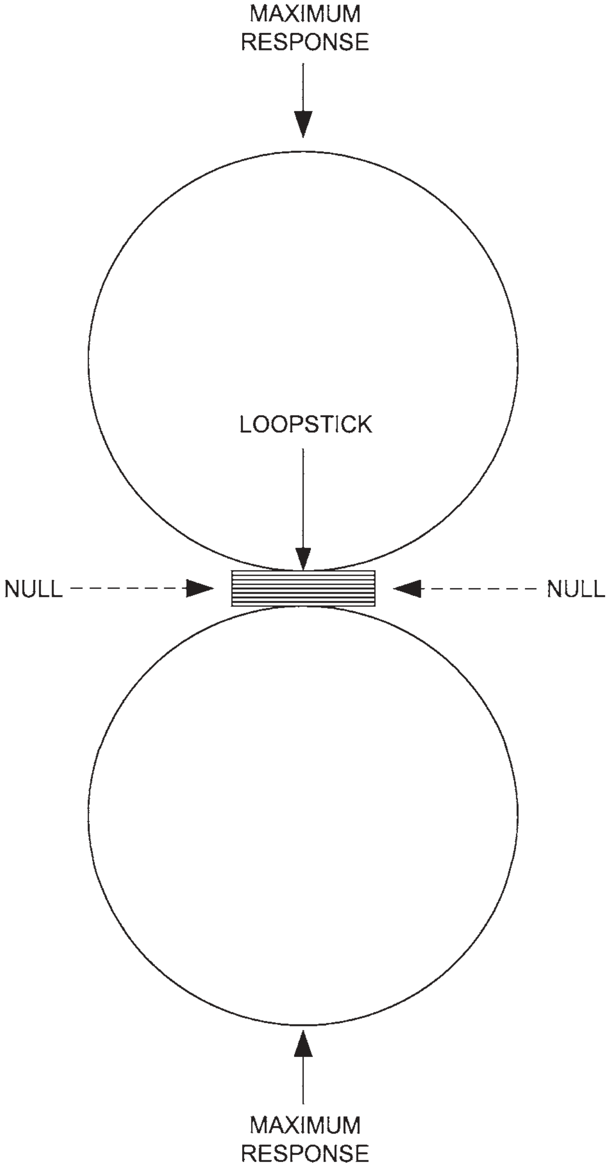

15-1B Loopstick antenna.

At one time, the principal use of the small loop antenna was radio direction-finding, especially in the lower-frequency bands. The RDF loop is mounted with a compass rose to allow the operator to find the direction of minimum response. The null was used, rather than the peak response point, because it is far narrower than the peak. As a result, precise determination of direction is possible. Because the null is bidirectional, ambiguity exists as to which of the two directions is the correct direction. What the direction-finder "finds" is a line along which the station exists. If the line is found from two reasonably separated locations and the lines of direction are plotted on a map, then the two lines will cross in the area of the station. Three or more lines of direction (a process called *triangulation*) yields a pretty precise knowledge of the station's actual location.

Today, these small loops are still used for radio direction-finding, but their use has been extended into the general receiving arena, especially on the low frequencies. One of the characteristics of these bands is the possibility of strong local interference smothering weaker ground-wave and sky-wave stations. As a result, you cannot hear cochannel signals when one of them is very strong and the other is weak. Similarly, if a cochannel station has a signal strength that is an appreciable fraction of the desired signal and is slightly different in frequency, then the two signals will heterodyne together and form a whistling sound in the receiver output. The frequency of the whistle is an audio tone equal to the difference in frequency between the two signals. This is often the case when trying to hear foreign BCB signals on frequencies (called *split frequencies*) between the standard spacing. The directional characteristics of the loop can help if the loop null is placed in the direction of the undesired signal.

Loops are used mainly in the low-frequency bands even though such loops are either physically larger than high-frequency loops or require more turns of wire. Loops have been used as high as VHF and are commonly used in the 10-m ham band for such activities as hidden transmitter hunts. The reason why low frequencies are the general preserve of loops is that these frequencies are more likely to have substantial ground-wave signals. Sky-wave signals lose some of their apparent directivity because of multiple reflections. Similarly, VHF and UHF waves are likely to reflect from buildings and hillsides and so will arrive at angles other than the direction of the transmitter. As a result, the loop is less useful for the purpose of radio direction-finding. If your goal is not RDF but listening to the station, this is hardly a problem. A small loop can be used in the upper shortwave bands to null a strong local ground-wave station in order to hear a weaker sky-wave station. Finally, loops can be useful in rejecting noise from local sources, such as a "leaky" electric power line or a neighbor's outdoor light dimmer.

Let's examine the basic theory of small loop antennas and then take a look at some practical construction methods.

Grover's equation

Grover's equation (Grover, 1946) seems closer to the actual inductance measured in empirical tests than certain other equations that are in use. This equation is

$$L_{\mu\mathrm{H}} = (K_1 N^2 a) \times \left\{ \ln \left[\frac{K_2 a N}{(N+1)\, b} \right] + K_3 + \left[\frac{K_4 \, (N+1)\, b}{aN} \right] \right\}$$

where $L_{\mu\mathrm{H}}$ is the inductance, in microhenrys (μH)
a is the length of a loop side, in centimeters (cm)
b is the loop width, in centimeters (cm)

n is the number of turns in the loop
K_1 through K_4 are shape constants and are given in Table 15-1
ln is the natural log of this portion of the equation

Air core frame loops ("box" loops)

A wire loop antenna is made by winding a large coil of wire, consisting of one or more turns, on some sort of frame. The shape of the loop can be circular, square, triangular, hexagonal, or octagonal. For practical reasons, the square loop seems to be most popular. With one exception, the loops considered in this section will be square, so you can easily duplicate them.

The basic form of the simplest loop is shown in Fig. 15-2. This loop is square, with sides the same length A all around. The width of the loop (B) is the distance from the first turn to the last turn in the loop, or the diameter of the wire if only one turn is used. The turns of the loop in Fig. 15-2 are *depth wound,* meaning that each turn of the loop is spaced in a slightly different parallel plane. The turns are spaced evenly across distance B. Alternatively, the loop can be wound such that the turns are in the same plane (this is called *planar winding*). In either case, the sides of the loop (A) should be not less than five times the width (B). There seems to be little difference between depth- and planar-wound loops. The far-field patterns of the different shape loops are nearly the same if the respective cross-sectional areas (πr^2 for circular loops and A^2 for square loops) are less than $\lambda^2/100$.

The reason why a small loop has a null when its broadest aspect is facing the signal is simple, even though it seems counterintuitive at first blush. Take a look at Fig. 15-3. Here, we have two identical small loop antennas at right angles to each other. Antenna A is in line with the advancing radio wave, whereas antenna B is broadside to the wave. Each line in the wave represents a line where the signal strength is the same, i.e., an "isopotential line." When the loop is in line with the signal (antenna A), there is a difference of potential from one end of the loop to the other, so current can be induced in the wires. When the loop is turned broadside, however, all points on the loop are on the same potential line, so there is no difference of potential between segments of the conductor. Thus little signal is picked up (and the antenna therefore sees a null).

The actual voltage across the output terminals of an untuned loop is a function of the angle of arrival of the signal α (Fig. 15-4), as well as the strength of the signal and the design of the loop. The voltage V_o is given by

Table 15-1. Shape constants

Shape	K_1	K_2	K_3	K_4
Triangle	0.006	1.155	0.655	0.135
Square	0.008	1.414	0.379	0.33
Hexagon	0.012	2.00	0.655	0.135
Octagon	0.016	2.613	0.7514	0.0715

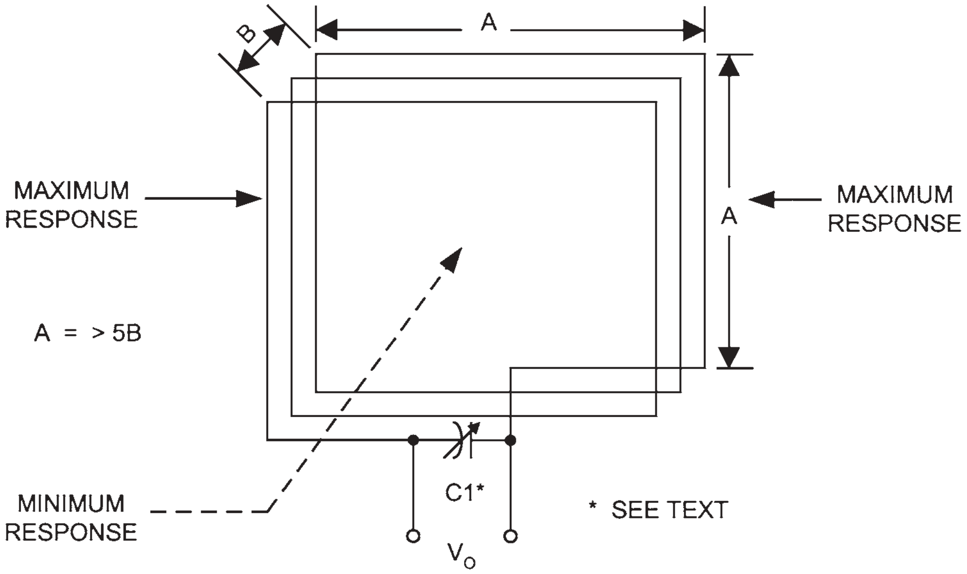

15-2A A simple loop antenna.

$$V_o = \frac{2\pi ANE_f \cos(\alpha)}{\lambda}$$

where V_o is the output voltage of the loop
A is the area of the loop, in square meters (m^2)
N is the number of turns of wire in the loop
E_f is the strength of the signal, in volts per meter (V/m)
a is the angle of arrival of the signal
λ is the wavelength of the arriving signal

Loops are sometimes specified in terms of the *effective height* of the antenna. This number is a theoretical construct that compares the output voltage of a small loop with a vertical piece of the same kind of wire that has a height of

$$H_{eff} = \frac{2\pi NA}{\lambda}$$

If a capacitor (such as C_1 in Fig. 15-2) is used to tune the loop, then the output voltage V_o will rise substantially. The output voltage found using the first equation is multiplied by the loaded Q of the tuned circuit, which can be from 50 to 100:

$$V_o = \frac{2\pi ANE_f Q \cos(\alpha)}{\lambda}$$

Even though the output signal voltage of tuned loops is higher than that of untuned loops, it is nonetheless low compared with other forms of antenna. As a result, a loop preamplifier usually is needed for best performance.

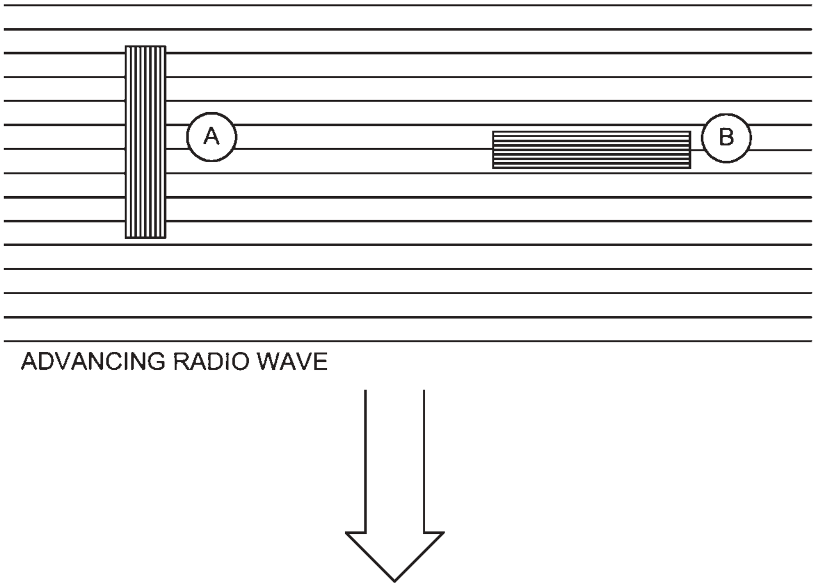
A
B
ADVANCING RADIO WAVE

15-3 Two small loop antennas at right angles to each other.

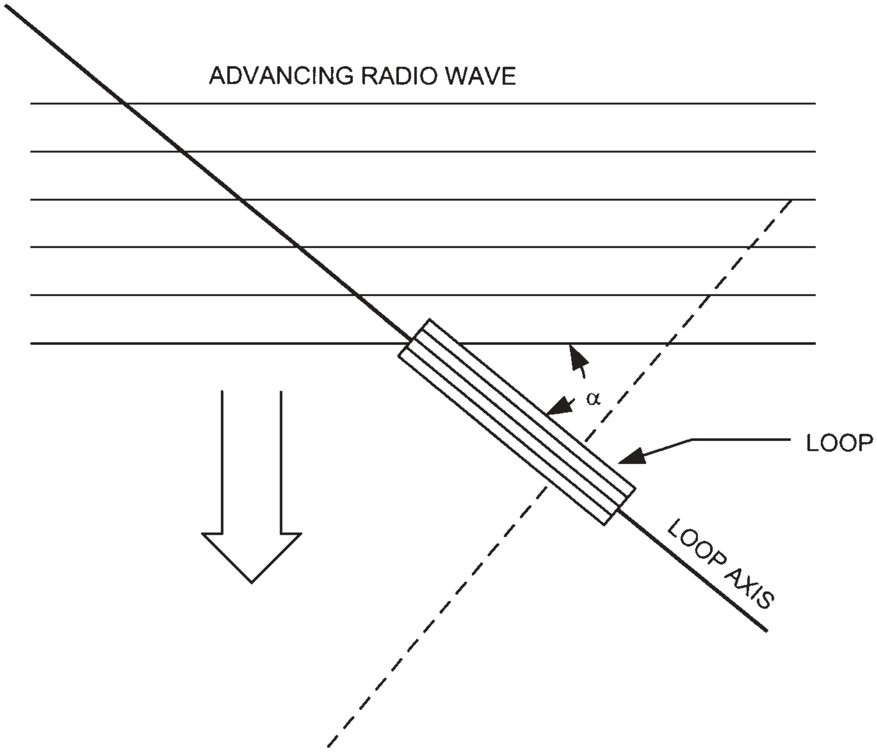
ADVANCING RADIO WAVE
α
LOOP
LOOP AXIS

15-4 An untuned loop antenna.

Transformer loops

It is common practice to make a small loop antenna with two loops rather than just one. Figure 15-5 shows such a transformer loop antenna. The main loop is built exactly as discussed above: several turns of wire on a large frame, with a tuning capacitor to resonate it to the frequency of choice. The other loop is a one- or two-turn *coupling loop.* This loop is installed in very close proximity to the main loop, usually (but not necessarily) on the inside edge not more than a couple of centimeters away. The purpose of this loop is to couple signal induced from the main loop to the receiver at a more reasonable impedance match.

The coupling loop is usually untuned, but in some designs a tuning capacitor (C_2) is placed in series with the coupling loop. Because there are many fewer turns on the coupling loop than on the main loop, its inductance is considerably smaller. As a result, the capacitance to resonate is usually much larger. In several loop antennas constructed for purposes of researching this chapter, I found that a 15-turn main loop resonated in the AM BCB with a standard 365-pF capacitor, but the two-turn

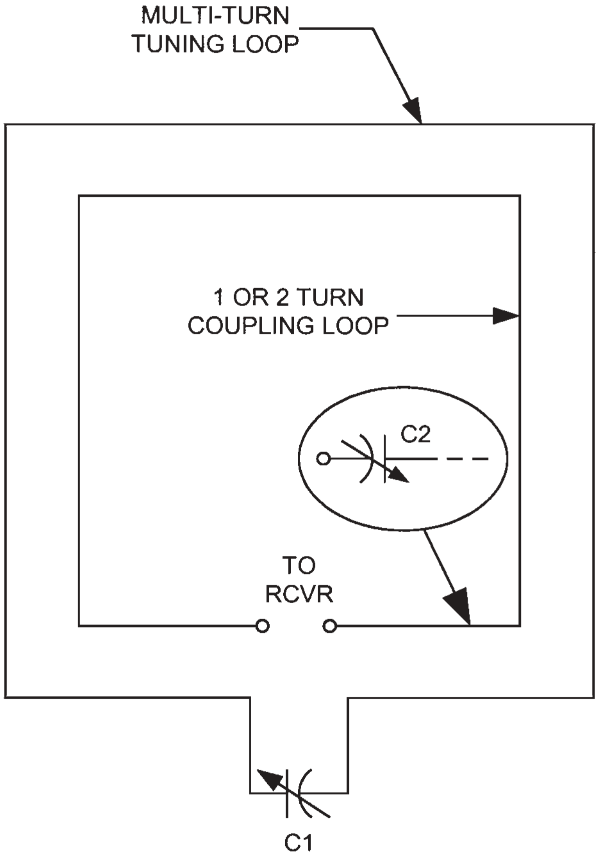

15-5 A transformer loop antenna.

coupling loop required three sections of a ganged 3 × 365-pF capacitor connected in parallel to resonate at the same frequencies.

In several experiments, I used computer ribbon cable to make the loop turns. This type of cable consists of anywhere from 8 to 64 parallel insulated conductors arranged in a flat ribbon shape. Properly interconnected, the conductors of the ribbon cable form a continuous loop. It is no problem to take the outermost one or two conductors on one side of the wire array and use them for a coupling loop.

Tuning schemes for loop antennas

Loop performance is greatly enhanced by tuning the inductance of the loop to the desired frequency. The bandwidth of the loop is reduced, which reduces front-end overload. Tuning also increases the signal level available to the receiver by a factor of 20 to 100 times. Although tuning can be a bother if the loop is installed remotely from the receiver, the benefits are well worth it in most cases.

There are several different schemes available for tuning, and these are detailed in Fig. 15-6. The parallel tuning scheme, which is by far the most popular, is shown in Fig. 15-6A. In this type of circuit, the capacitor (C_1) is connected in parallel with the inductor, which in this case is the loop. Parallel resonant circuits have a very high impedance to signals on their resonant frequency and a very low impedance to other frequencies. As a result, the voltage level of resonant signals is very much larger than the voltage level of off-frequency signals.

The series resonant scheme is shown in Fig. 15-6B. In this circuit, the loop is connected in series with the capacitor. A property of series resonant circuits is that they offer a high impedance to all frequencies except the resonant frequency (exactly the opposite of the case of parallel resonant circuits). As a result, current from the signal will pass through the series resonant circuit at the resonant frequency, but off-frequency signals are blocked by the high impedance.

There is a wide margin for error in the inductance of loop antennas, and even the precise-looking equations to determine the required values of capacitance and inductance for proper tuning are actually only estimations. The exact geometry of the loop "as built" determines the actual inductance in each particular case. As a result, it is often the case that the tuning provided by the capacitor is not as exact as desired, so some form of compensation is needed. In some cases, the capacitance required for resonance is not easily available in a standard variable capacitor, and some means must be provided for changing the capacitance. Figure 15-6C shows how this is done. The main tuning capacitor can be connected in either series or parallel with other capacitors to change the value. If the capacitors are connected in parallel, then the total capacitance is increased (all capacitances are added together). If the extra capacitor is connected in series, however, then the total capacitance is reduced. The extra capacitors can be switched in and out of a circuit to change frequency bands.

Tuning of a remote loop can be a bother if it is done by hand, so some means must be found to do it from the receiver location (unless you enjoy climbing into the attic or onto the roof). Traditional means of tuning called for using a low-rpm dc motor, or stepper motor, to turn the tuning capacitor. A very popular combination was the little 1- to 12-rpm motors used to drive rotating displays in retail store show

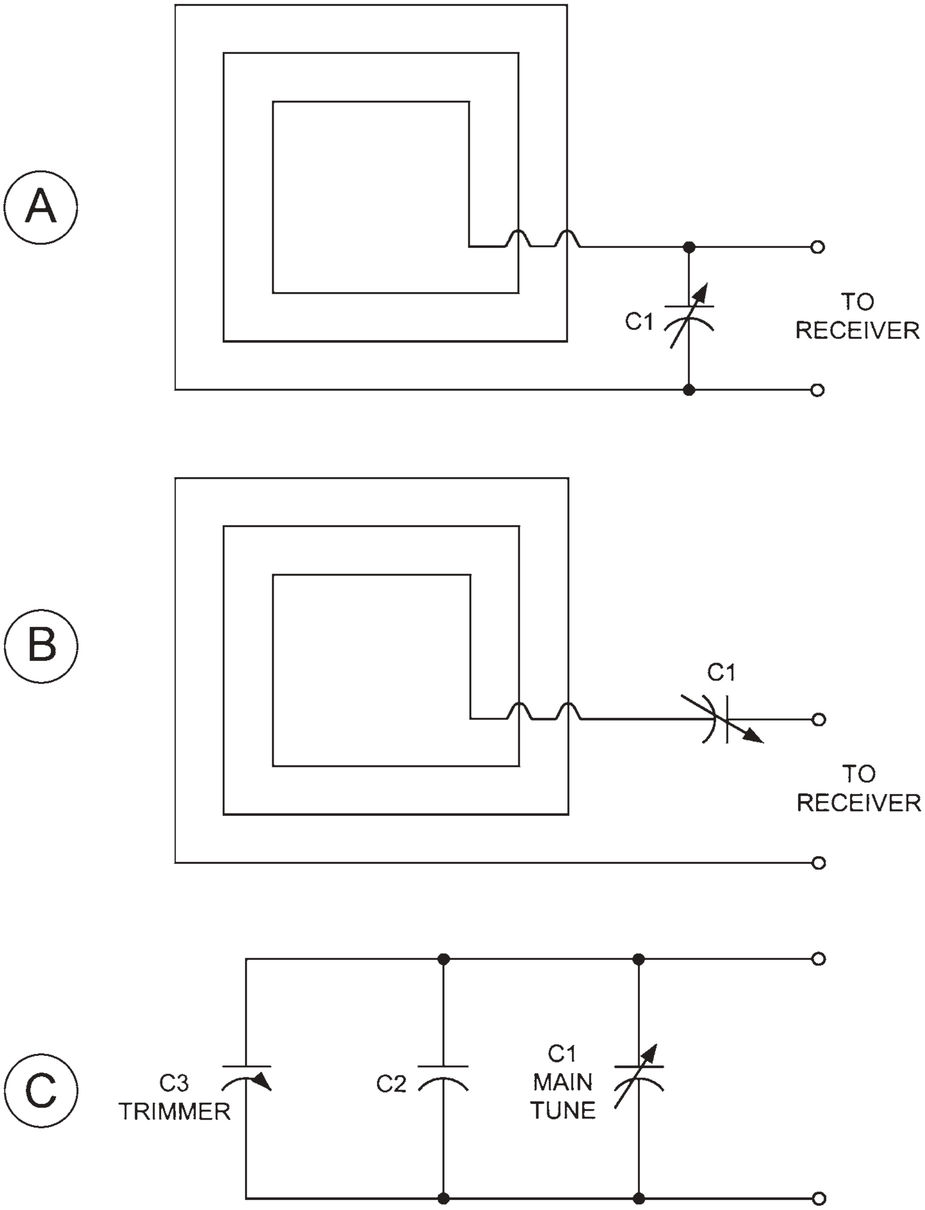
A
C1
TO
RECEIVER
B
C1
TO
RECEIVER
C
C3
TRIMMER
C2
C1
MAIN
TUNE

15-6 Various tuning schemes: (A) parallel, (B) series resonant, (C) series or parallel.

windows. But this approach is not really needed today. We can use varactor voltage-variable capacitance diodes to tune the circuit.

A varactor works because the junction capacitance of the diode is a function of the applied reverse-bias voltage. A high voltage (such as 30 V) drops the capacitance, whereas a low voltage increases it. Varactors are available with maximum capacitances of 22, 33, 60, 100, and 400 pF. The latter are of most interest to us because they have the same range as the tuning capacitors normally used with loops.

Figure 15-7 shows how a remote tuning scheme can work with loop antennas. The tuning capacitor is a combination of a varactor diode and two optional capacitors: a fixed capacitor (C_1) and a trimmer (C_2). The dc tuning voltage (V_t) is provided from the receiver end from a fixed dc power supply (V+). A potentiometer (R_1) is used to set the voltage to the varactor, hence also to tune the loop. A dc blocking capacitor (C_3) keeps the dc tuning voltage from being shorted out by the receiver input circuitry.

The sports fan's loop

OK, sports fans, what do you do when the best game of the week is broadcast only on a low-powered AM station and you live at the outer edge of their service area where the signal strength leaves much to be desired? You use the sports fan's loop antenna, that's what! I first learned of this antenna from a friend of mine, a professional broadcast engineer, who worked at a religious radio station that had a pipsqueek signal but lots of fans. It really works—*one might say it's a miracle.*

The basic idea is to build a 16-turn, 60-cm^2 tuned loop and then place the AM portable radio at the center so that its loopstick is aimed such that its null end is broadside of the loop. When you do so, the nulls of both the loop and the loopstick are in the same direction. The signal will be picked up by the loop and then coupled to the radio's loopstick antenna. Sixteen-conductor ribbon cable can be used for making the loop. For an extra touch of class, place the antenna and radio assembly on a dining room table lazy Susan to make rotation easier. A 365-pF tuning capacitor is used to resonate the loop. If you listen to only one station, then this capacitor can be a trimmer type.

Shielded loop antennas

The loop antennas discussed thus far in this chapter have all been unshielded types. Unshielded loops work well under most circumstances, but in some cases their pattern is distorted by interaction with the ground and nearby structures (trees, buildings, etc.). In my own tests, trips to a nearby field proved necessary to measure the depth of the null because of interaction with the aluminum siding on my house. Figure 15-8 shows two situations. In Fig. 15-8A we see the pattern of the normal "free space" loop, i.e., a perfect figure-8 pattern. When the loop interacts with the nearby environment, however, the pattern distorts. In Fig. 15-8B we see some filling of the notch for a moderately distorted pattern. Some interactions are so severe that the pattern is distorted beyond all recognition.

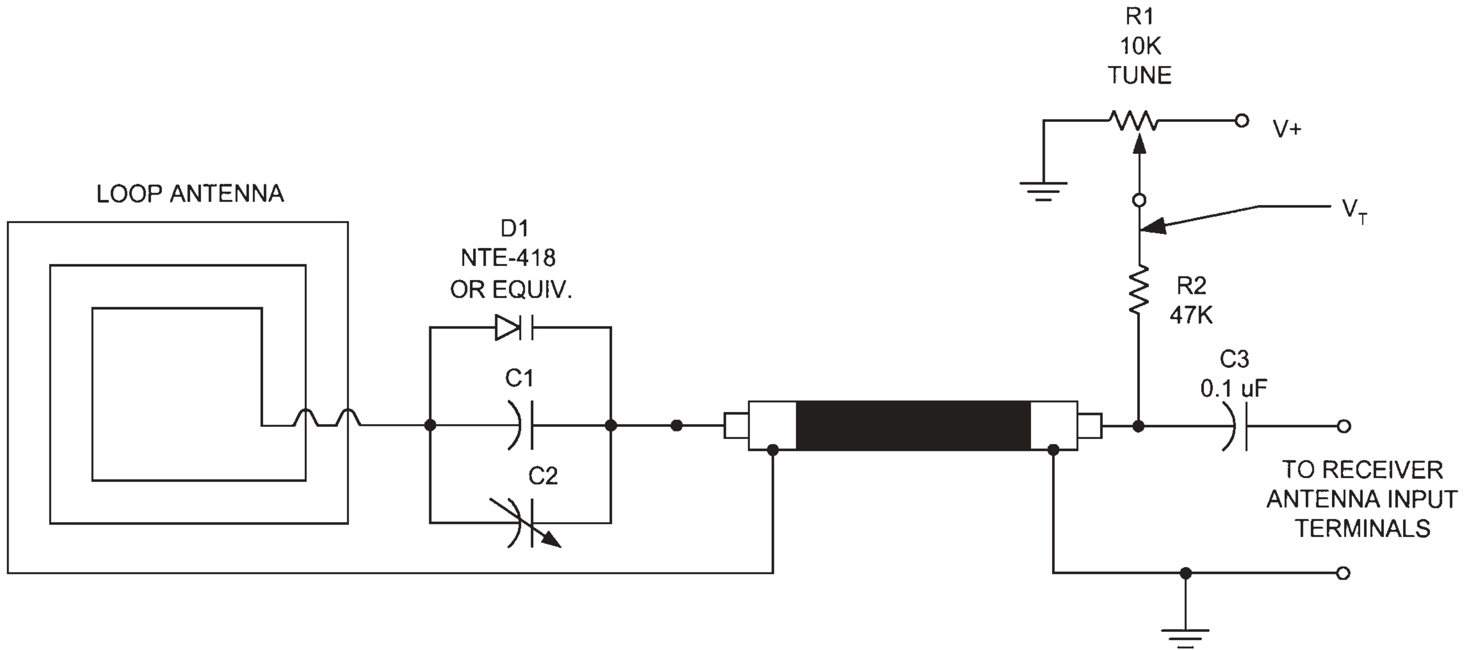
R1
10K
TUNE
V+
V_T
LOOP ANTENNA
D1
NTE-418
OR EQUIV.
C1
C2
R2
47K
C3
0.1 uF
TO RECEIVER
ANTENNA INPUT
TERMINALS

15-7 Remote tuning scheme.

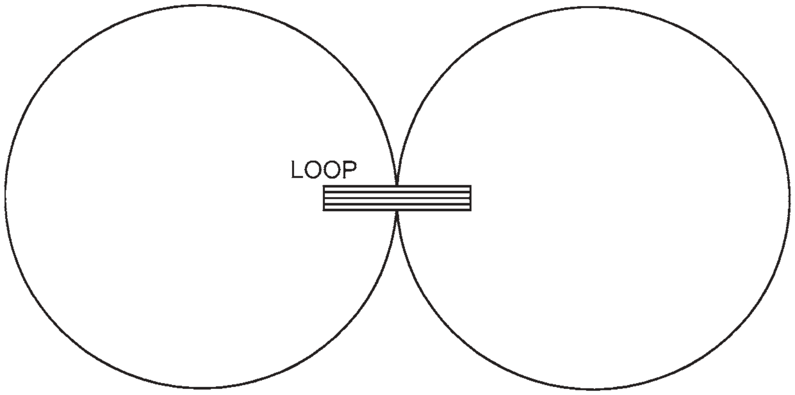

15-8A A normal "free space" loop.

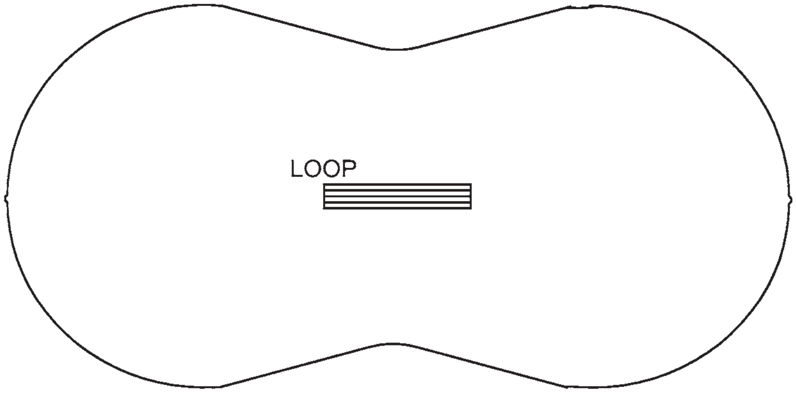

15-8B Filling of the notch.

The solution to the problem is to reduce interaction by shielding the loop, as in Fig. 15-9. Loop antennas operate on the magnetic component of the electromagnetic wave, so the loop can be shielded against voltage signals and electrostatic interactions. In order to prevent harming the ability to pick up the magnetic field, a gap is left in the shield at one point.

There are several ways to shield a loop. You can, for example, wrap the loop in adhesive-backed copper-foil tape. Alternatively, you can wrap the loop in aluminum foil and hold it together with tape. Another method is to insert the loop inside a copper or aluminum tubing frame. Or—the list seems endless.

Using a loop antenna

Most readers will use a loop for DXing rather than hidden transmitter hunting, navigation, or other RDF purposes. For the DXer, there are actually two uses for the

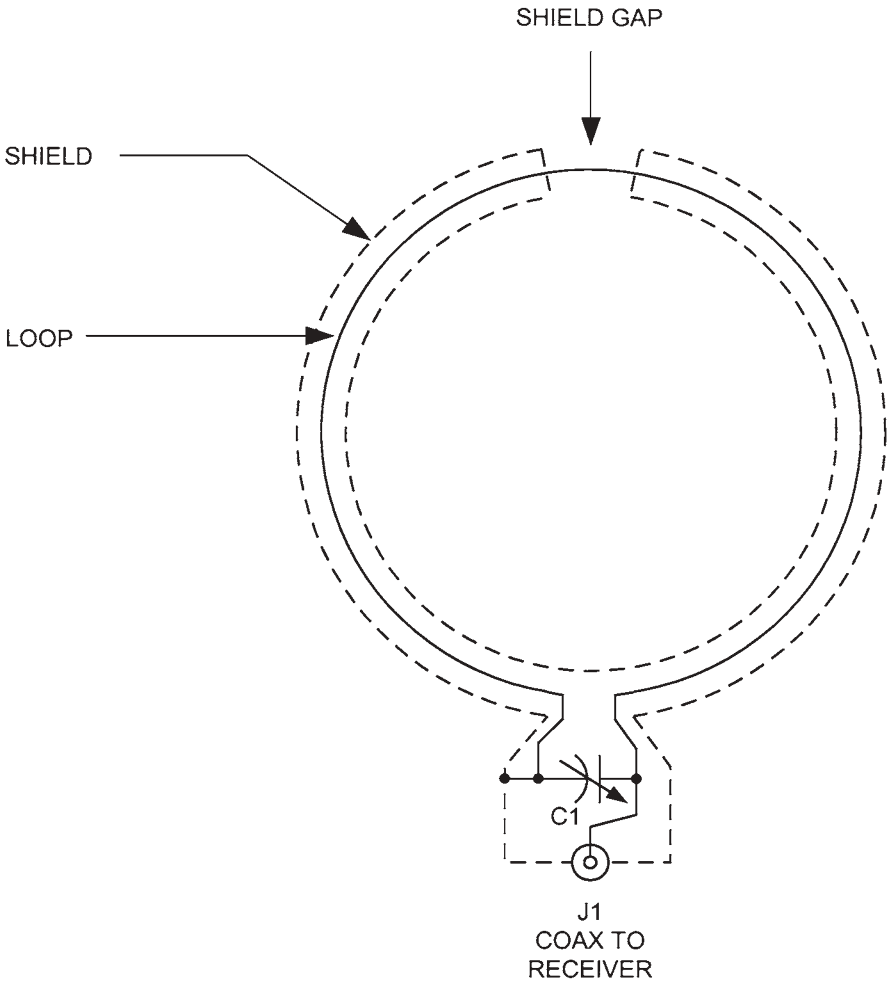

15-9 Shielding the loop.

loop. One is when you are a renter or live in a community that has routine covenants against outdoor antennas. In this situation, the loop will serve as an active antenna for receiving AM BCB and other low-frequency signals without the neighbors or landlord becoming PFJs ("purple-faced jerks").

The other use is illustrated by the case of a friend of mine. He regularly tunes in clear channel WSM (650 kHz, Nashville) in the wee hours between Saturday evening ("Grand Ole Opry" time) and dawn. However, this "clear" channel of WSM is not really so clear, especially without a narrow filter in the receiver. He uses a loop antenna to null out a nearby 630-kHz signal that made listening a bit dicey and can now tape his 1940s–1950s vintage country music.

It is not necessary to place the desired station directly in the main lobes off the ends of the antenna but rather to place the nulls (broadside) in the direction of the offending station that you want to eliminate. So what happens if the offending station and the desired station are in a direct line with each other and your receiving location is in the middle between them? Both nulls and lobes on a loop antenna are bidirectional, so a null on the offending station also will null the desired station in the opposite direction.

One method is to use a sense antenna to spoil the pattern of the loop to a cardioid shape. Another method is to use a spoiler loop to null the undesired signal. The spoiler loop is a large box loop placed 1 to 3 ft (found experimentally) behind the reception loop in the direction of the offending signal. This method was first described by Levintow and is detailed in Fig. 15-10. The small loopstick may be the antenna inside the receiver, whereas the large loop is a box loop such as the sports fan's loop. The large box loop is placed about 33 to 100 cm behind the loopstick and in the direction of the offending station. The angle with respect to the line of centers should be 60° to 90°, which also was found experimentally. It is also possible to use two air core loops to produce an asymmetrical receiving pattern.

Sharpening the loop

Many years ago, the *Q-multiplier* was a popular add-on accessory for a communications receiver. These devices were sold as Heathkits, and many construction projects could be found in magazines and amateur radio books. The Q-multiplier has the effect of seeming to greatly increase the sensitivity of a receiver, as well as greatly reducing the bandwidth of the front end. Thus it allows better reception of some stations because of increased sensitivity and narrowed bandwidth.

A Q-multiplier is an active electronic circuit placed at the antenna input of a receiver. It is essentially an Armstrong oscillator, as shown in Fig. 15-11, that does not quite oscillate. These circuits have a tuned circuit (L_1/C_1) at the input of an amplifier stage and a feedback coupling loop (L_3). The degree of feedback is controlled by the coupling between L_1 and L_3. The coupling is varied by varying both how close the two coils are and their relative orientation with respect to each other. Certain other circuits use a series potentiometer in the L_3 side that controls the amount of feedback.

The Q-multiplier is adjusted to the point that the circuit is just on the verge of oscillating, but not quite. As the feedback is backed away from the threshold of oscillation, but not too far, the narrowing of bandwidth occurs, as does the increase in sensitivity. It takes some skill to operate a Q-multiplier, but it is easy to use once you get the hang of it and is a terrific accessory for any loop antenna.

Loop amplifier

Figure 15-12 shows the circuit for a practical loop amplifier that can be used with either shielded or unshielded loop antennas. It is based on junction field effect transistors (JFET) connected in cascade. The standard common-drain configuration is used for each transistor, so the signals are taken from the source terminals. The drain terminals are connected together and powered from the +12-V dc power

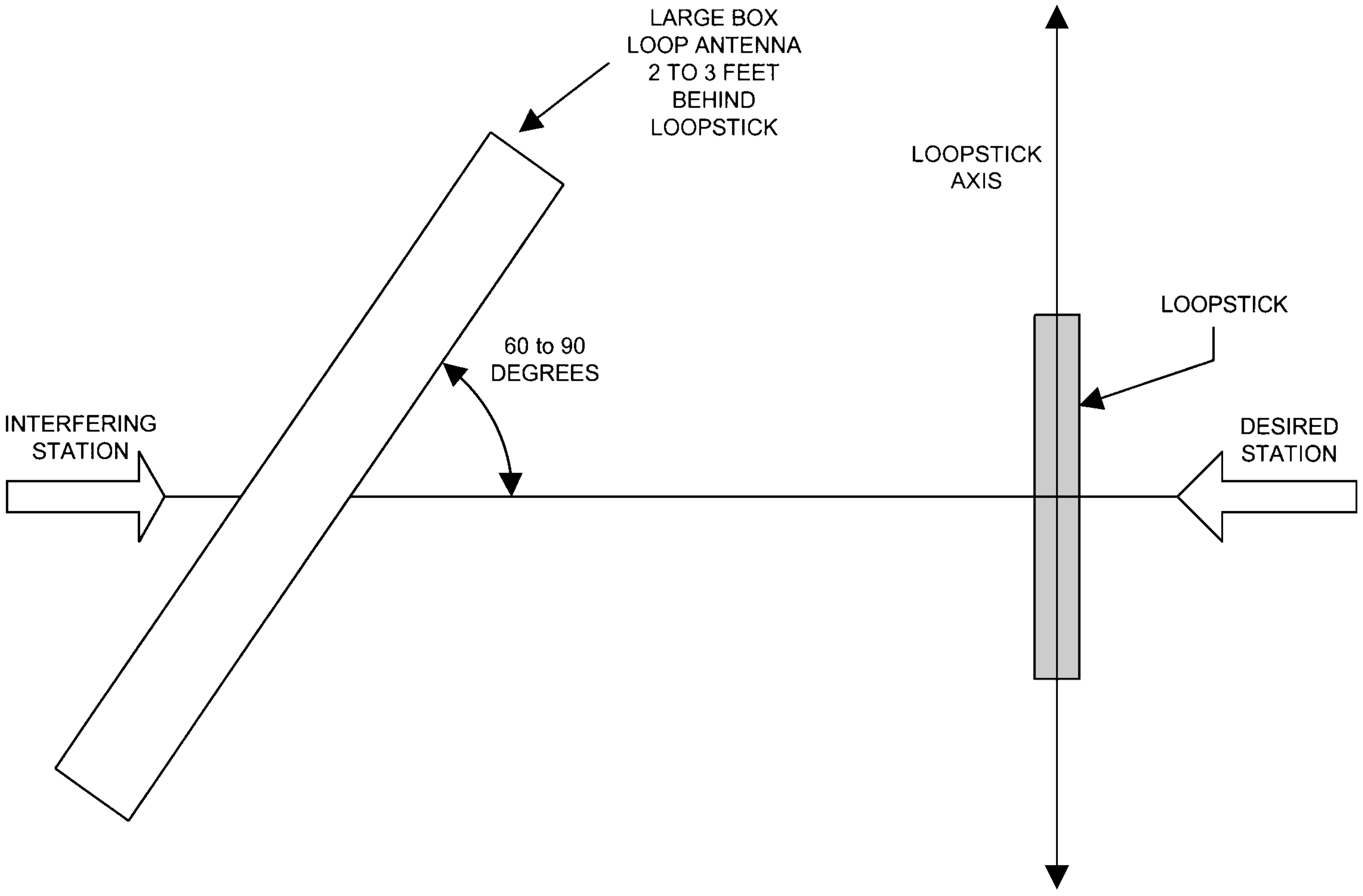
LARGE BOX LOOP ANTENNA 2 TO 3 FEET BEHIND LOOPSTICK
LOOPSTICK AXIS
LOOPSTICK
60 to 90 DEGREES
INTERFERING STATION
DESIRED STATION

15-10 Use of a spoiler loop to null an undesired signal.

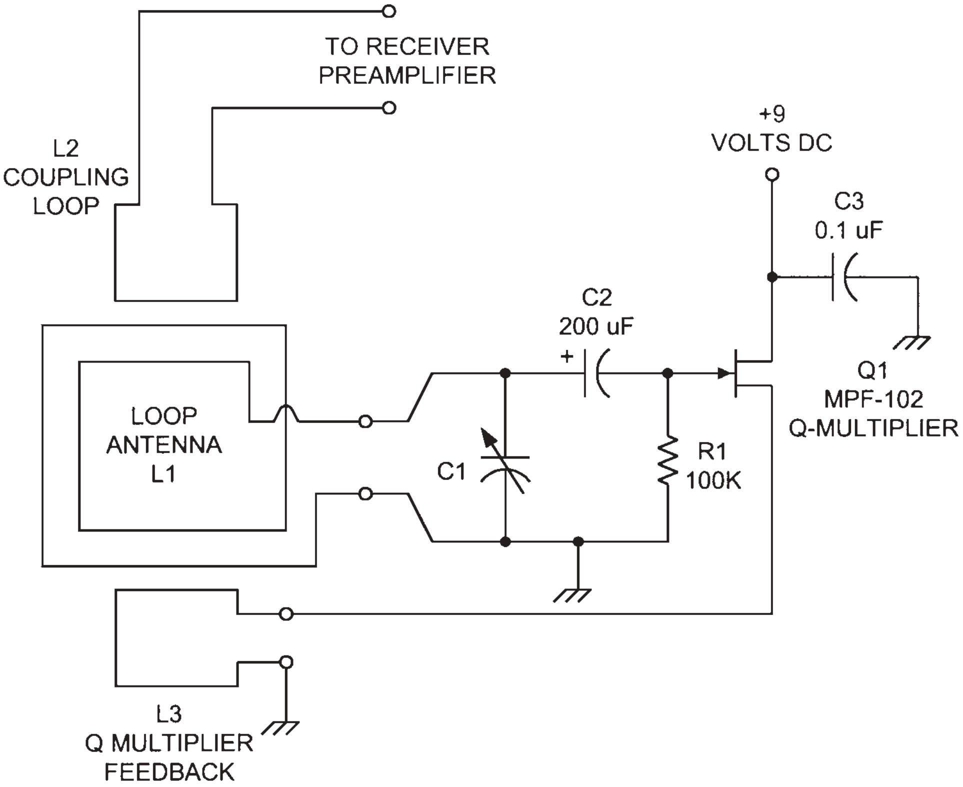
TO RECEIVER
PREAMPLIFIER
L2
COUPLING
LOOP
+9
VOLTS DC
C3
0.1 uF
C2
200 uF
+
Q1
MPF-102
Q-MULTIPLIER
LOOP
ANTENNA
L1
C1
R1
100K
L3
Q MULTIPLIER
FEEDBACK

15-11 A Q-multiplier.

supply. A 2.2-μF bypass capacitor is used to put the drain terminals of Q_1 and Q_2 at ground potential for ac signals while keeping the dc voltage from being shorted out.

The two output signals are applied to the primary of a center-tapped transformer, the center tap of which is grounded. To keep the dc on the source terminals from being shorted through the transformer winding, a pair of blocking capacitors (C_4, C_5) is used.

The input signals are applied to the gate terminals of Q_1 and Q_2 through dc blocking capacitors C_2 and C_3. A pair of diodes (D_1, D_2) is used to keep high-amplitude noise transients from affecting the operation of the amplifier. These diodes are connected back to back in order to snub out both polarities of signal.

Tuning capacitor C_1 is used in lieu of the capacitor in the loop and is used to resonate the loop to a specific frequency. Its value can be found from the equation given earlier.

The transistors used for the push-pull amplifier (Q_1, Q_2) can be nearly any general-purpose JFET device (MPF-102, MPF-104, etc.). A practical approach for many people is to use transistors from service replacement lines, such as the NTE-312 and NTE-316 devices.

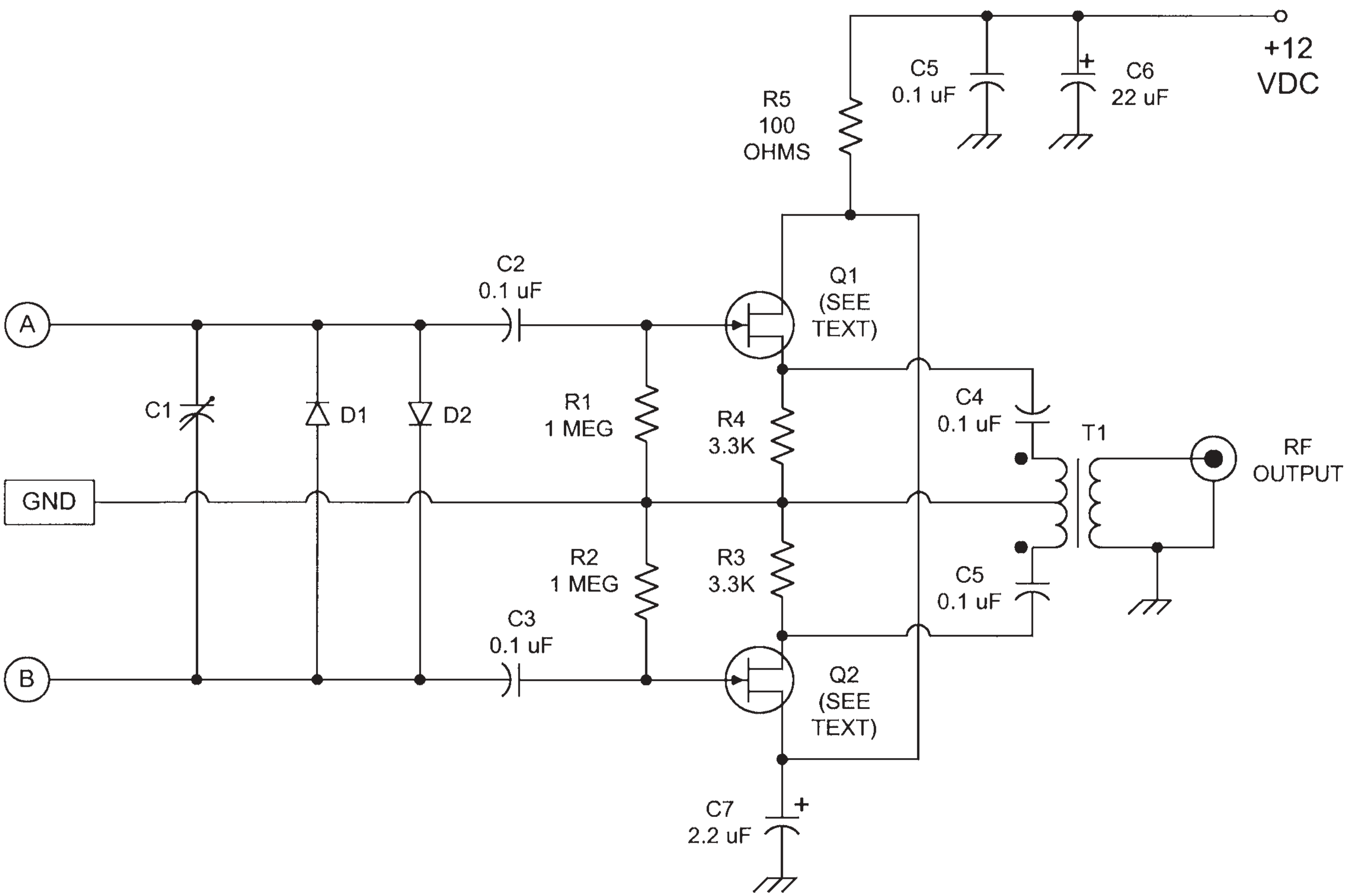
+12
VDC
R5
100
OHMS
C5
0.1 uF
C6
22 uF
A
GND
B
C1
D1
D2
C2
0.1 uF
C3
0.1 uF
R1
1 MEG
R2
1 MEG
R4
3.3K
R3
3.3K
Q1
(SEE
TEXT)
Q2
(SEE
TEXT)
C4
0.1 uF
C5
0.1 uF
T1
RF
OUTPUT
C7
2.2 uF

15-12 A practical loop amplifier.

Special problem for VLF/LF loops

A capacitance is formed whenever two conductors are side by side. A coil produces capacitance as well as inductance because the turns are side by side. Unfortunately, with large multiturn loops, this capacitance can be quite large. The "distributed capacitance" of the loop causes a self-resonance with the inductance. The loop does not work well at frequencies above the self-resonant point, so it is sometimes important to raise the self-resonance to a point where it does not affect operation at the desired frequencies.

Figure 15-13 shows a solution that raises the self-resonant point. The turns are broken into two or more groups and separated by a space. This method reduces the effective capacitance by placing the capacitances of each group of wires in series with the others.

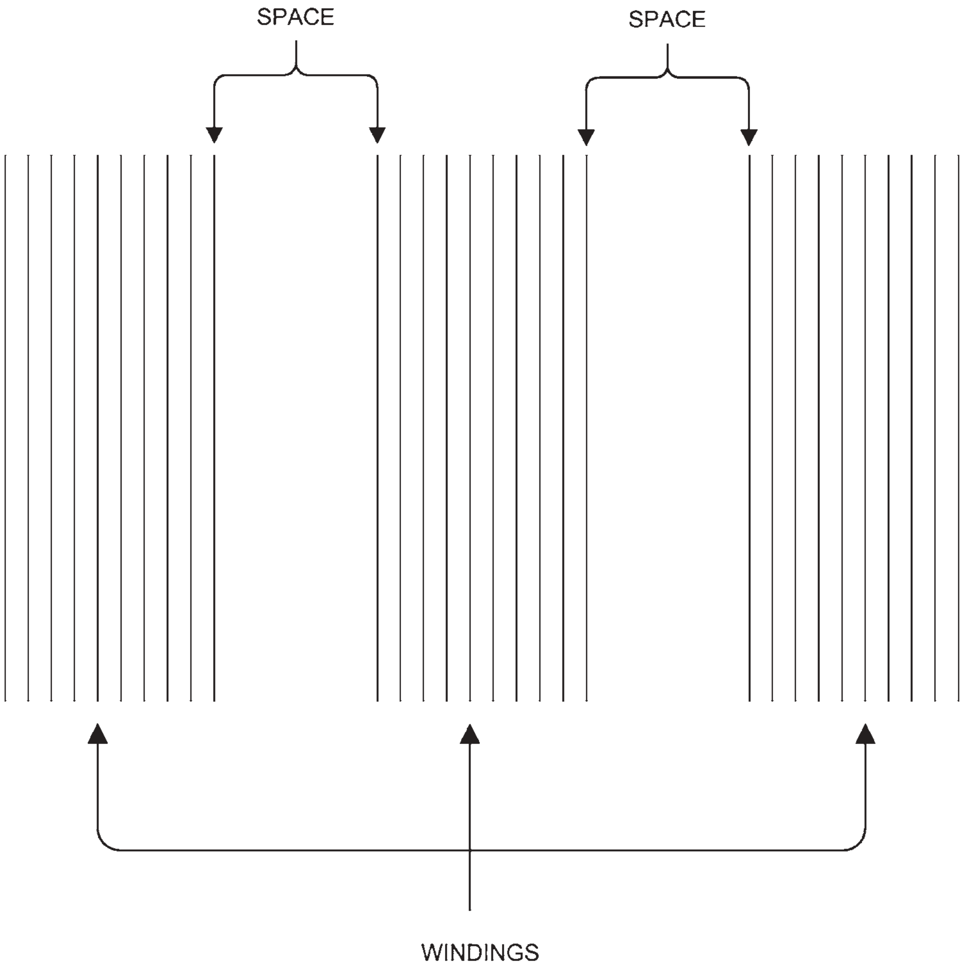

15-13 Raising the self-resonant point.

Coaxial-cable loop antennas

One of the more effective ways to make a shielded loop is to use coaxial cable. Figure 15-14 shows the circuit of such a loop. Although only a single-turn loop is shown, there can be any number of turns. One reader made a 100-kHz LORAN (a navigation system) loop using eight turns of RG-59/U coaxial cable on an 8-ft diameter.

Note the special way that the coaxial cable is connected. This method is called the *Faraday connection* after the fact that the shield of the coax forms a *Faraday shield.* At the output end, the center conductor of the coaxial cable is connected to the center conductor of the coaxial connector. The coax shield is connected to the connector ground/shield terminal. At the other end of the loop, the shield is left floating, but the center conductor is connected to the shield. Note very carefully that the center conductor at the far end is connected to the shield *at the connector,* not at just any convenient point.

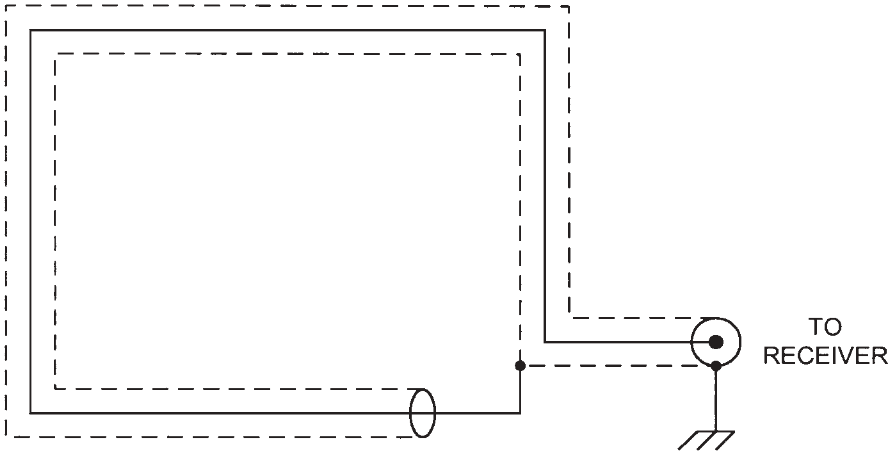

15-14 A coaxial-cable shielded loop.

16
Chapter

Small transmitting loop antennas

Small transmitting loops are those which are less than one-third wavelength (λ/3) according to Hart (1986) or between 0.04λ and 0.10λ according to Belrose (1993). Other sources claim that loops up to about 0.20λ are small. The key characteristic is that the current is equal in all points of the loop rather than varying with loop length the way the current in a large loop (or other antenna) does.

One of the parameters of the small loop antenna is a very low radiation resistance (it is a function of the area of the loop). The radiation resistance of the small loop antenna is less than 1 Ω and often considerably less than this figure (e.g., 0.05 Ω). This means that the copper or aluminum losses of the antenna must be kept very, very low. For such reasons, it is common to assemble the small transmitting loop antenna from 1- to 4-in copper or aluminum pipe. For this reason, the square (Fig. 16-1A) or octagon (Fig. 16-1B) shapes are preferred. The reason this is true is purely mechanical: The square loop can be built with 90° elbow joints, and the octagon can be built with 45° elbow joints. For antennas whose circumference is greater than λ/8, standard ¾-in copper pipe can be used in construction, but for smaller-sized loops, larger-sized copper piping is used.

The small transmitting loop tends to be highly inductive in its reactance. Values of inductive reactance to 1000 Ω are not unusual, although closer to 100 Ω is the norm. Because of this inductive reactance, the small transmitting loop antenna is tuned by one or more series capacitors.

A sample loop was modeled by Belrose (1993). It was 1 m in diameter and was made of 2.54-cm (1-in) copper pipe. For this loop, the gains were

10 MHz	2.88 dBi
14 MHz	0.22 dBi
21 MHz	1.14 dBi
30 MHz	1.42 dBi

Note that the gains were all less than that of a dipole (12.15 dBi) but in the upper regions of the frequency spectrum were quite decent. Even at 10 MHz, the 1-m

Copyright 2001 - the McGraw-Hill Companies

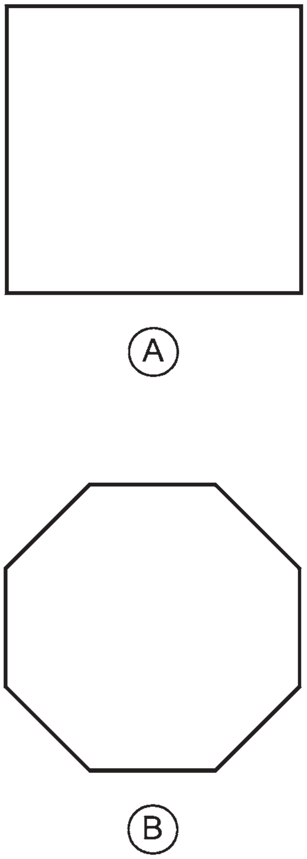

16-1 Small transmitting loop antennas: (A) square, (B) octagon.

small loop antenna was only −5.03 dBi less than the dipole. The loop probably would show similar gains up to the point of self-resonance.

Calculations of the current and voltage within the loop are substantial. The loop impedance at 10 MHz is $0.088\ 1\ j161\ \Omega$. This means that a 150-W transmitter, certainly modest for these frequencies, would produce 41 A circulating inside the loop and a voltage across the tuning capacitor of 6600 V. As Belrose (1993) warns: "Don't touch your loop when transmitting!"

The tuning capacitor in Belrose's loop should be rated for at least 10,000 V to guard against anyone using more than 150 W. Keep in mind that a power of 500 W will produce potentials across the capacitor of 30,000 V, so you will need a capacitor with at least $^1/_2$-in spacing. A 1-in spacing is good for about 75,000 V. It is probably best to use a vacuum-variable capacitor for the tuning capacitor of a small loop antenna if you can afford one. Regardless, the loop power should be kept under control!

The Q of the antenna at any given resonant frequency is very, very high. For the 1-m loop antenna described earlier, the Q at 10 MHz was 1824, which resulted in a bandwidth of 5.5 kHz (although in practice it is twice this amount before VSWR tolerances on the transmitter are exceeded). Regardless, the bandwidth of the resonated loop is tiny, so some means must be provided for retuning the loop as you shift frequency. If you do not shift frequency, then the loop can be tuned and forgotten, but for the rest of the radio fraternity, perhaps a remotely tuned capacitor is in order. Getting a large capacity variable in the voltages that are required for operation with a good power level is a bit tricky.

Design equations

The design equations developed by Hart (1986) are as follows:

Radiation resistance (R_R):

$$R_R = 3.38 \times 10^{-8} \, (F^2 A)^2 \qquad \textbf{[16.1]}$$

Loss resistance (R_L):

$$R_L = \frac{9.96 \times 10^{-4} S \sqrt{F}}{D} \qquad \textbf{[16.2]}$$

Efficiency (η):

$$\eta = \frac{R_R}{R_R + R_L} \qquad \textbf{[16.3]}$$

Inductance (L):

$$\mathrm{L} = 1.9 \times 10^8 \mathrm{S} \left(7.353 \log_{10} \frac{96S}{\pi D} - 6.386\right) \qquad \textbf{[16.4]}$$

Inductive reactance (X_L):

$$X_L = 2\pi F L \times 10^6 \qquad \textbf{[16.5]}$$

Tuning capacitance (C):

$$C = \frac{1}{2\pi F X_L \times 10^6} \qquad \textbf{[16.6]}$$

Quality factor (Q):

$$Q = \frac{F}{\Delta F} = \frac{X_L}{2\,(R_R + R_L)} \qquad \textbf{[16.7]}$$

Bandwidth (ΔF):

$$\Delta F = \frac{F}{Q} \qquad \textbf{[16.8]}$$

Distributed capacitance (C_D):

$$C_D = 0.82S \qquad \textbf{[16.9]}$$

Capacitor voltage (V_C):

$$V_C = \sqrt{P X_L Q} \qquad \textbf{[16.10]}$$

where

A is the area of the loop, in square feet (ft2)
S is the length of the loop conductor, in feet (ft)
D is the diameter of the conductor, in inches (in)
P is the applied transmitter output power, in watts (W)

Some sample loop antennas

The Patterson (1967) loop antenna is shown in Fig. 16-2. It is made from 1.5-in copper tubing. Segments are cut and are joined together by eight 45° elbow joints, giving the octagon shape. Each segment is 0.5 ft long. The tuning is accomplished by three capacitors, two of which are a split-stator unit (two capacitors on the same shaft). The tuning control is the split-stator capacitor, whereas the loading control is a single capacitor.

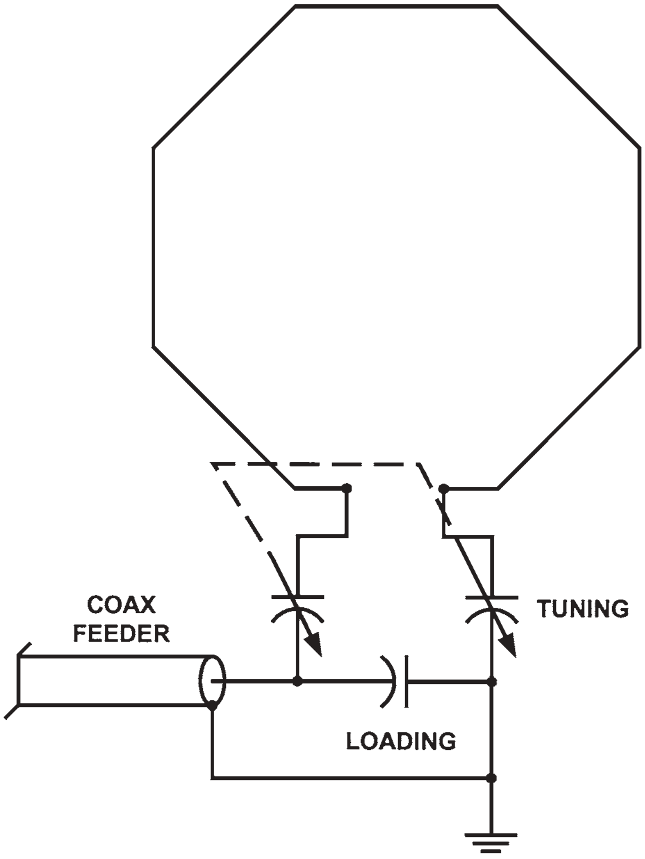

16-2 A Patterson loop antenna.

Another loop antenna is shown in Fig. 16-3. This antenna has a Faraday feed system rather than a capacitor-coupled feed system, as did the Patterson loop. The tuning is accomplished by C_A, the series tuning capacitor. The requirement must be met that $D_1/D_2 = 5$.

Still another loop is shown in Fig. 16-4. This loop relies on magnetic coupling to perform the coupling of the transmitter. In this loop antenna, the coupling is via a small coupling loop and 50-Ω coaxial cable to the transmitter. Capacitor C_1 is used to resonate the loop, whereas capacitors C_2 and C_3 serve the purposes of loading and resonating the coupling loop. According to Mozzochi (1993), the voltages and currents with respect to the capacitors are

$$V_{C1} = 1800\sqrt{P} \quad \textbf{[16.11]}$$

$$V_{C2} = 18\sqrt{\mathrm{P}} \quad \textbf{[16.12]}$$

$$V_{C3} = 17\sqrt{\mathrm{P}} \quad \textbf{[16.13]}$$

$$I_{C1} = 2.2\sqrt{\mathrm{P}} \quad \textbf{[16.14]}$$

$$I_{C2} = 0.21\sqrt{P} \quad \textbf{[16.15]}$$

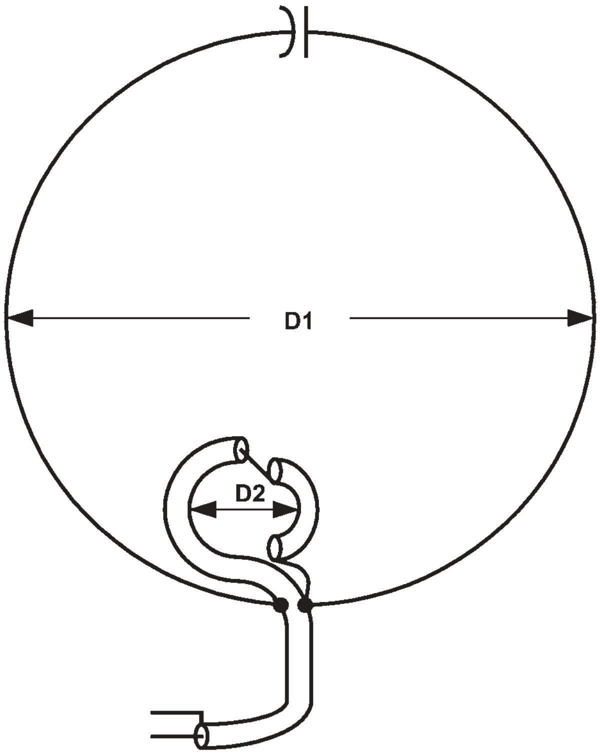

16-3 A loop antenna with a Faraday feed system.

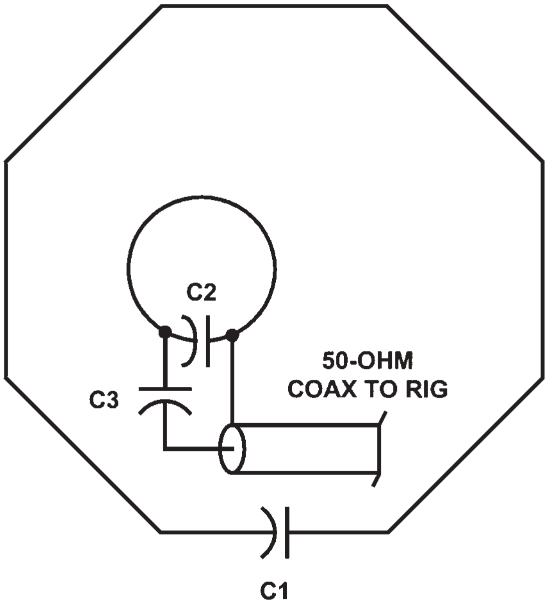

16-4 A loop antenna with magnetic coupling.

$$I_{C3} = 0.15\sqrt{P} \qquad \textbf{[16.16]}$$

The radiation patterns for the loop antenna are shown in Fig. 16-5. Note that four elevations are given (0°, 20°, 45°, and 60°).

The bottom line is that small transmitting loop antennas are not very good for those who can afford to put up a better antenna, but for those whose tight quarters permit only a small transmitting loop, they are quite viable.

References

Belrose, John S. (1993), "An Update on Compact Transmitting Loops," *QST,* November.

Erwin, David (1991), "Very Small Transmitting and Receiving Only Antennas," in *HF Antenna Collection,* Radio Society of Great Britain, Potters Bar, Herts., UK, Chap. 5.

Grover, F. W. (1946), *Inductance Calculation-Working Formulas and Tables,* Van Nostrand, New York.

Hart, Ted (1986), "Small, High-Efficiency Loop Antennas: An Alternative Antenna for Small Spaces," *QST,* June.

Koontz, Floyd (1993), "A High-Directivity Receiving Antenna for 3.8 MHz," *QST,* August, pp. 31–34.

McCoy, L. (1968), "The Army Loop in Ham Communications," *QST,* March, pp. 17–18, 150; see also "Technical Correspondence" column, *QST,* May, pp. 49–51, and November, pp. 46–47.

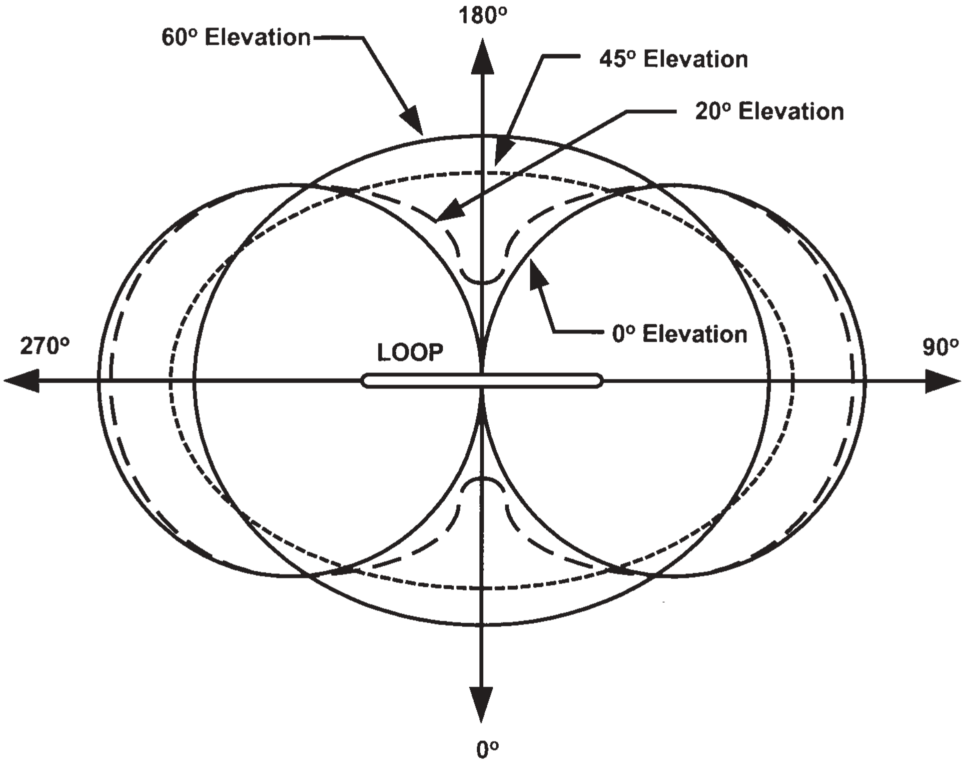
60° Elevation
180°
45° Elevation
20° Elevation
0° Elevation
270°
LOOP
90°
0°

16-5 Radiation patterns.

Mozzochi, Charles J. (1993), "A Small Loop Antenna for 160 Meters," *QST,* pp. 32–34.

Orr, William (1994), "Feedback on Hart's article," *QST,* July.

Patterson, K. (1967), "Down-to-Earth Army Antennas," *Electronics,* August 21, pp. 111–114.

Somerfield, A. (1952), *Electrodynamics,* Academic Press, New York.

This page intentionally left blank

17
CHAPTER
Antenna modeling software

One of the significant contributions of computer technology to antenna design is the improvement of modeling and simulation. Modeling and simulation are used in a wide variety of applications, including management, science, and engineering. We can now model just about any process, any device, or any circuit that can be reduced mathematically. The purpose of modeling, at least in engineering, is to do the design cheaply on the computer before "bending metal." The old "cut and try" method works, to be sure, but it is costly in time and money (two things perpetually in short supply). If performance issues and problems can be solved on a computer, then we are time and money ahead of the game. Also, modeling and simulation make it possible to look at more alternatives and to gauge the effect of a change in an antenna design before the change is made.

Modeling is used not just in engineering but also in other processes. Organizations and management consultants use modeling and simulation to study their operations and find the most efficient way of doing things. I saw an example where an inefficient truck terminal and warehouse operation was improved significantly by modeling on a Windows desktop computer. The operators modeled the existing system and several alternative systems, and found that one particular rule (involving bulk deliver trucks such as tankers and dry bulk carriers) was jamming up the normal package deliveries at the warehouse dock. They could have tried different sets of rules over a couple of years, collected lots of data, presented scores of viewgraphs to management committees, and still have overlooked the actual cause of the problem. The model showed the problem in only a few minutes, however.

What is a model?

Models are no big mystery. If you have fingers and can count, then you can make models. In fact, you probably do make models mentally, e.g., when you look at how much cash you have in your pocket when confronted with several different purchase choices but cannot afford all of them. The "model" is the "what if" considerations that lead to your selection.

Copyright 2001 - the McGraw-Hill Companies

Youngsters in the schools are learning that the equations and formulas they learn in algebra and trigonometry are "models." This designation was not used when I was in school, but it is an apt description in most cases.

One of the earliest modeling methods for the masses was the computer-based spreadsheet program. Software such as Microsoft Excel can be used to test various scenarios by entering different values into the cells. I have seen spreadsheets used for the originally intended business applications, as well as engineering applications (including antenna design).

Antenna modeling software tends to use the "method of moments," in which each wire in the system is broken into 1 to N segments, and the current in each segment is calculated.

Some common programs

Some years ago an antenna modeling program came on the scene. Developed by the Navy, this program used numerical methods to perform the complex calculations needed to model an antenna. Called the *Numerical Electromagnetic Computation* (NEC) *program,* a smaller version soon became available known as miniNEC (which runs on PCs). You can download miniNEC-3 from several sources in the amateur radio newsgroups in the United States or United Kingdom, or from the Navy Postgraduate School Web site. A miniNEC-4 version is also available, but I understand that it requires a license to use and cannot be exported without a permit from the U.S. government.

There are also a number of private vendors who offer the miniNEC at a price. Some of the better vendors offer their own versions, based on the miniNEC concept, but with some value added to make it worthwhile to buy theirs instead of using the "freebie" version.

The miniNEC programs model wire antennas using the coordinate system shown in Fig. 17-1. This is a standard three-dimensional Cartesian coordinate system in which the y axis is horizontal, the z axis is vertical, and the x axis is orthogonal to both the y and z axes. The idea is to specify the starting and ending points of the wires making up the antenna in terms of the xyz coordinate system. Figure 17-2 shows a single-wire antenna laid out in the x axis. This axis is in the horizontal plane but orthogonal to the y axis. The antenna designer will specify X_1 and X_2, which are the starting and ending points of the wire.

The antenna can consist of a large number of wires, although the larger the number, the longer is the processing time. Each wire is broken into a number of different segments for calculation purposes.

NEC-4

The NEC-4 software is the latest in the NEC series and is proprietary to the University of California, from whom a license must be obtained to use it. Because of national defense considerations, an export license is required to take NEC-4 overseas. The NEC-4 software will produce modeling of underground radials, elements of varying diameter, and carefully constructed close-spaced parallel wires.

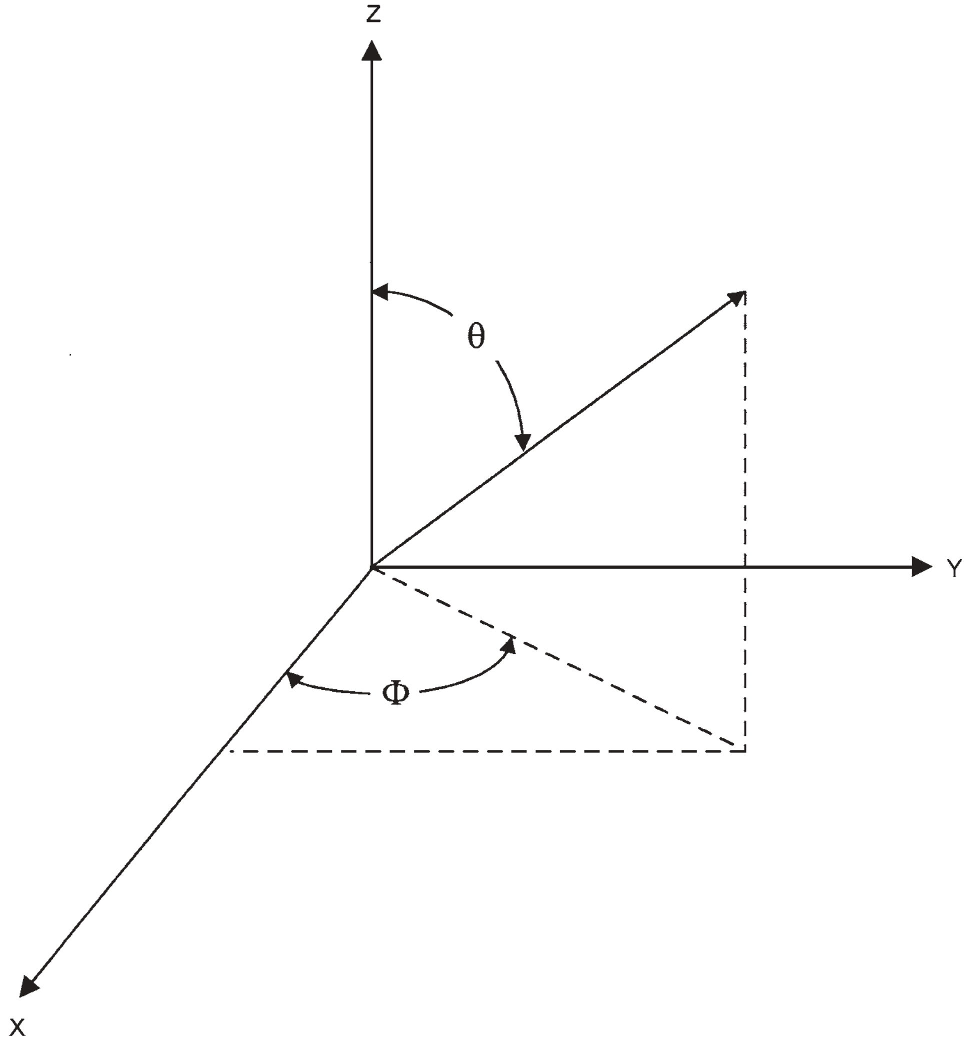

17-1 *XYZ* coordinate system for NecWin modeling software.

EZNEC Professional

Roy Lewallen, W7EL, provides the EZNEC Pro software. This software has an option for NEC-4, provided that the license is obtained from the University of California (again, export restrictions apply).

GNEC

Nittany-Scientific makes a version of NEC-4 known as GNEC. It permits the use of catenary wires, helices, networks, and other possibilities. Outputs include 3-D, polar plots, and rectangular plot graphs, as well as a large array of tabular reports.

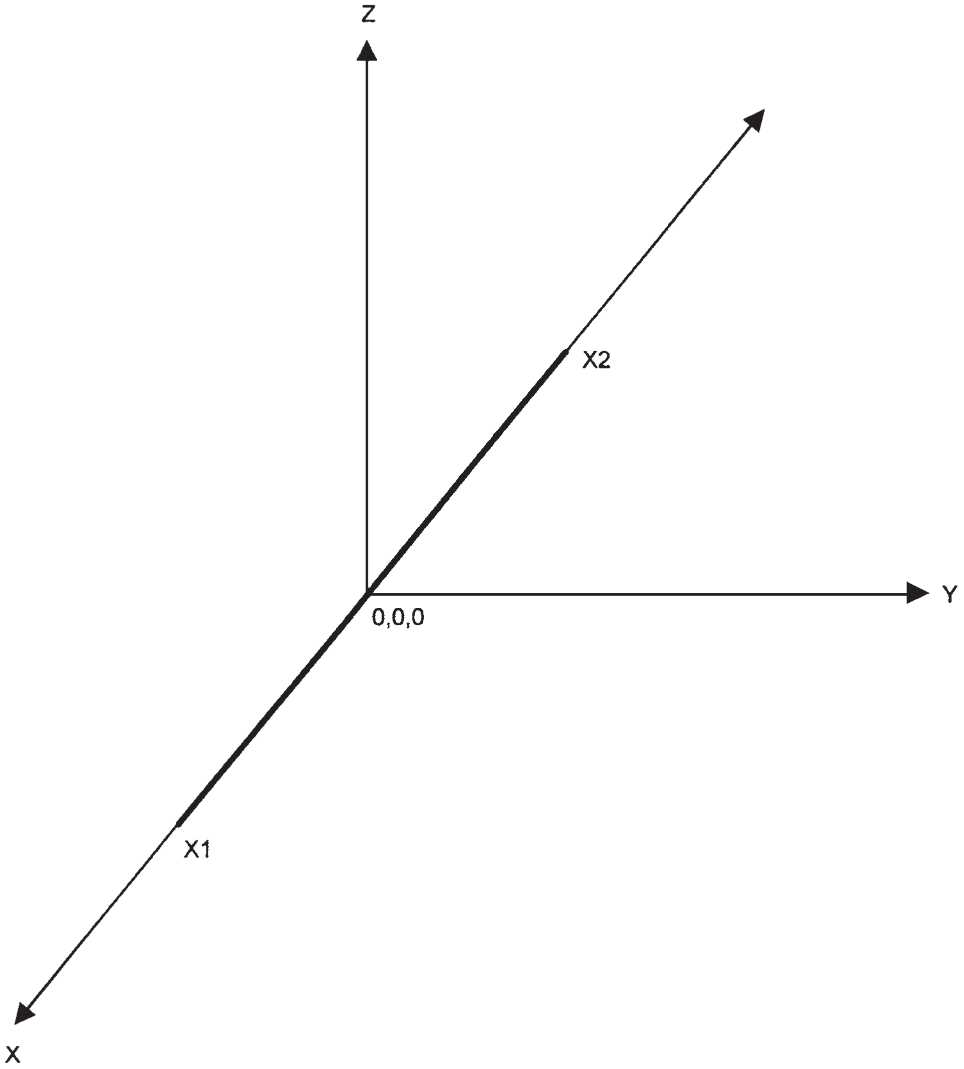

17-2 Example of a single wire laid out on *XYZ* coordinate system.

NEC-2

This is a high-capability version of the NEC software, and it is in the public domain. It is limited to antenna elements of constant diameter, although some software providers who base their wares on NEC-2 provide corrections for multidiameters. It is fitted with the Sommerfield-Norton ground model for modeling horizontal wires close to the earth's surface.

NECWin Plus

This software is provided by Nittany-Scientific. It features true spreadsheet geometry. The software also offers stepped-diameter corrections, a gain-averaging test, a

CAD file (*.dxf), 2-D and 3-D plots, 2-D and 3-D antenna views, and certain graphic outputs. There are two versions, NECWin Pro and NECWin Basic (NECWin Basic is reviewed in this chapter).

EZNEC for Windows

Roy Lewallen's EZNEC for Windows is available in both NEC-2 and NEC-4 versions, although the NEC-4 version requires a license from the University of California to operate. It offers 2-D and 3-D plots, 3-D plots with 2-D slicing, ground-wave output, stepped-diameter correction, and various shortcuts to antenna modeling. The standard EZNEC is limited to 500 segments (the Pro version handles a much larger number of segments).

Expert MiniNEC

This new version of MiniNEC can handle sharp angles in antenna geometry directly and handles antennas close to the ground much more accurately than MiniNEC. Several versions of Expert MiniNEC are available from EM Scientific. These are MiniNEC for Windows and MiniNEC Broadcast Professional.

MiniNEC

MiniNEC Version 3.13 is the latest of the public-domain MiniNEC code. It runs on a PC in the DOS mode. It handles element lengths of changing diameter. Horizontal antennas must be at least 0.2λ above ground for accurate results to occur.

NEC4Win95

Orion (Canada) offers a Windows 95/98 version of MiniNEC known as NEC4Win95. It offers spreadsheet antenna geometry input, pull-down boxes for all other antenna parameters, and a pattern-plotting routine that includes lobe identification and bandwidth.

NECWin Basic

Many amateur and professional users have been increasingly dissatisfied with the commercial miniNEC-based antenna modeling programs on the market (mostly because they only run in the DOS environment and are clumsy to use). A low-cost Windows-based modeling program is NecWin Basic for Windows. This program appears to be a DOS-based NEC engine run with a Windows graphic user interface written in Visual Basic 3.0. The vendor offering NecWin Basic for Windows is Nittany Scientific, Inc. [Airline Highway, Suite 361, Hollister, CA, 95023-5621; phone: (408) 634-0573; Web site: *www.nittany-scientific.com*]. The basic version (which I test drove and now use) costs less than $100, and there is a professional version for a higher price.

NecWin Basic for Windows allows you to specify the geometry of the antenna in the *xyz* coordinate system. This process is made a lot easier by the fact that it uses a "spreadsheet-like" grid (Fig. 17-3). Once the geometry is entered, you can select from several different ground situations (including "none" for free-space calculations), the operating frequency, and units of measure used for the antenna wires (meters is default). You also can customize the calculations using the "Output" button. Once the geometry and parameters are entered, you can visualize the geometry in "stick form" using the "Eye" button on the toolbar. The "Traffic Light" button performs the NEC calculations, whereas the “Polar Coordinates” button creates the elevation and azimuth pattern graphics. When you “Save Screen” in the “File” menu, you will produce a bitmap (*.bmp) file of the pattern.

The particular geometry entered in Fig. 17-3 is for a long-wire antenna (Fig. 17-4) that is two wavelengths long and has a quarter-wavelength radial used as a counterpoise ground. There are no wires in the y and z planes, so the values of Y_1, Y_2, Z_1, and Z_2 are zero. The antenna is described with the coordinates X_1 = 0.000 and X_2 = 14.56 m. The operating frequency is set to 25-MHz on another screen. When the NEC calculations are made and the patterns displayed, the azimuth (Fig. 17-5) and elevation (Fig. 17-6) patterns can be printed out.

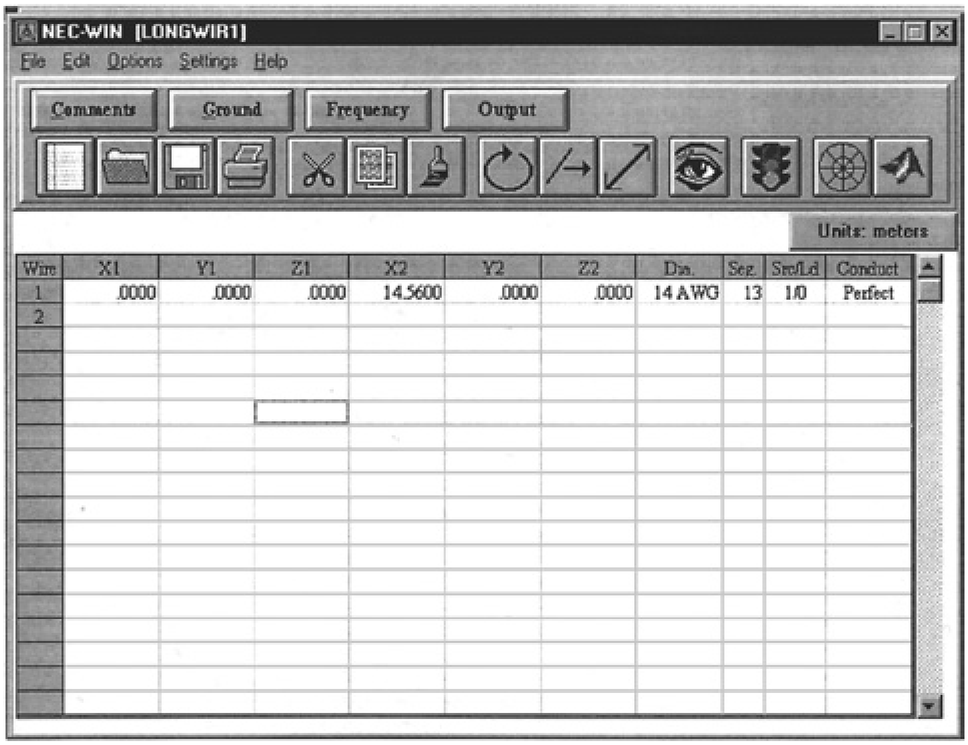
NEC-WIN [LONGWIR1]
File Edit Options Settings Help
Comments
Ground
Frequency
Output
Units: meters

Wire	X1	Y1	Z1	X2	Y2	Z2	Dia.	Seg.	Src/Ld	Conduct
1	.0000	.0000	.0000	14.5600	.0000	.0000	14 AWG	13	1/0	Perfect
2										

17-3 NecWin screen.

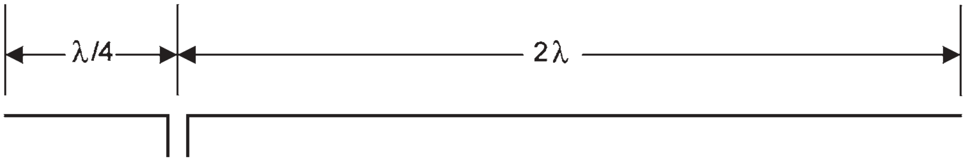

17-4 Off-center fed doublet antenna.

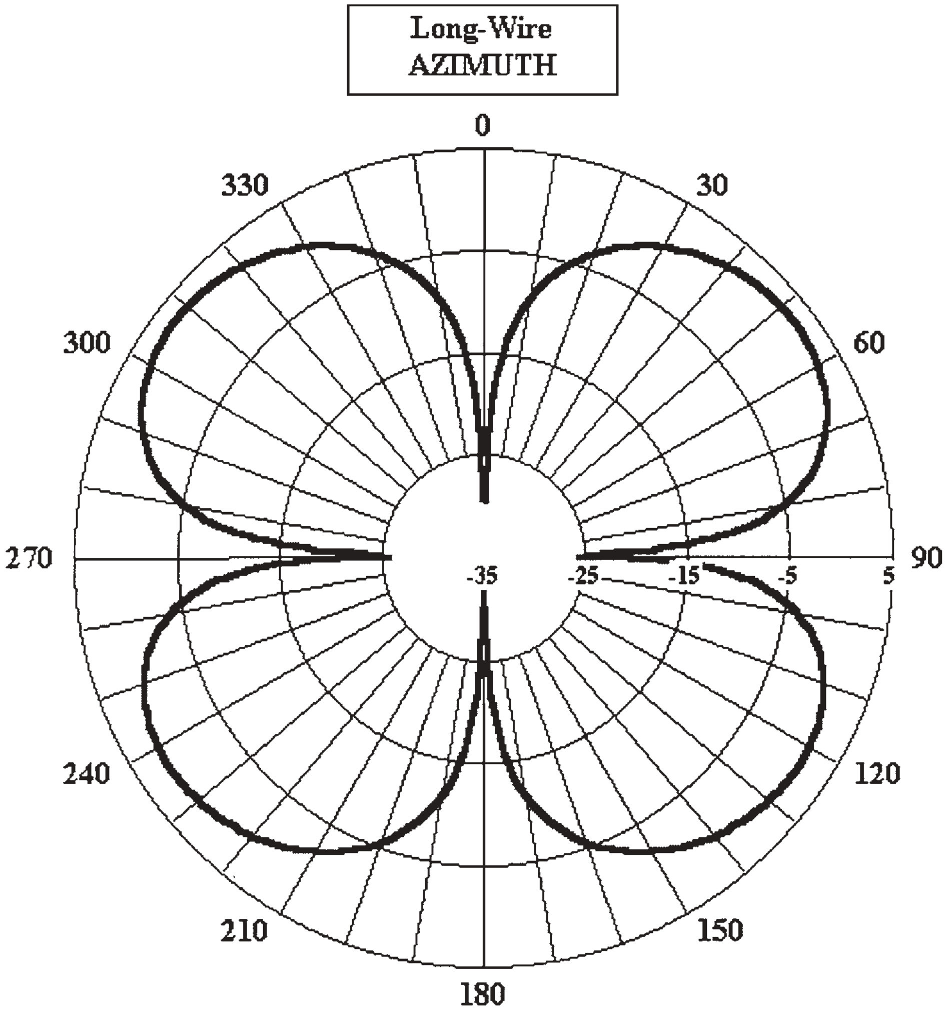

17-5 Modeled horizontal pattern.

NecWin Basic for Windows also will print out a data table of the frequency, wavelength, and VSWR. You can set the starting frequency, number of calculations, and step frequency.

If there is any criticism of the NecWin Basic for Windows program, it is that it does not graph the VSWR versus frequency curve. This is a little harder in Visual Basic 3.0 than in later versions, but users who also have Microsoft Excel on their computers can hand enter the numbers from the NecWin Basic for Windows output report and get a chart similar to Fig. 17-7.

The low price of NecWin Basic for Windows makes it useful to amateur radio operators, shortwave and scanner enthusiasts, and other amateur antenna builders. If

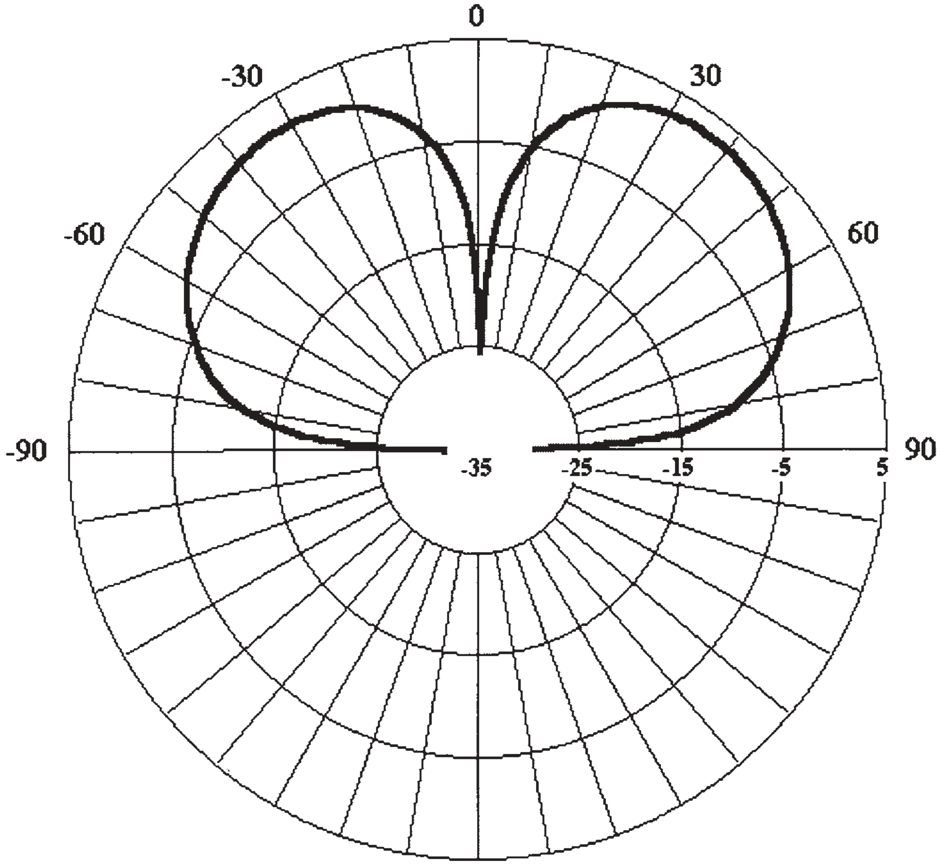

17-6 Modeled vertical pattern.

you are serious about antennas, then you need to obtain this software. It will perform well for many professional users, as well as student users in schools. However, professional users might want to consider buying the higher-priced NecWin Professional for Windows version.

EZNEC for Windows

Roy Lewallen, W7EL, produces the EZNEC for Windows software. The software extends Roy's earlier work on DOS versions of EZNEC but is still based on the NEC-2 engine (a NEC-4 professional version is available for those with the required license from the University of California).

Figure 17-8 shows the basic window that opens when you open an antenna project (in this case, a 10-MHz dipole). Given in the active window are file name, frequency, wavelength (in meters), number of wires and segments, number of sources, number of loads, transmission lines, ground type, wire loss, the units being calculated

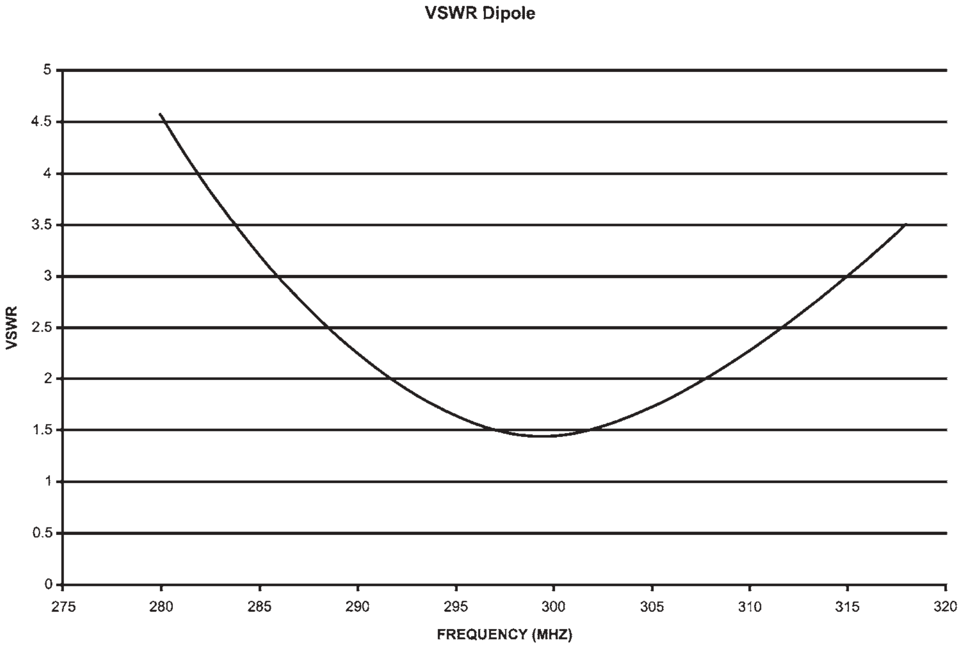

17-7 VSWR curve of the modeled antenna.

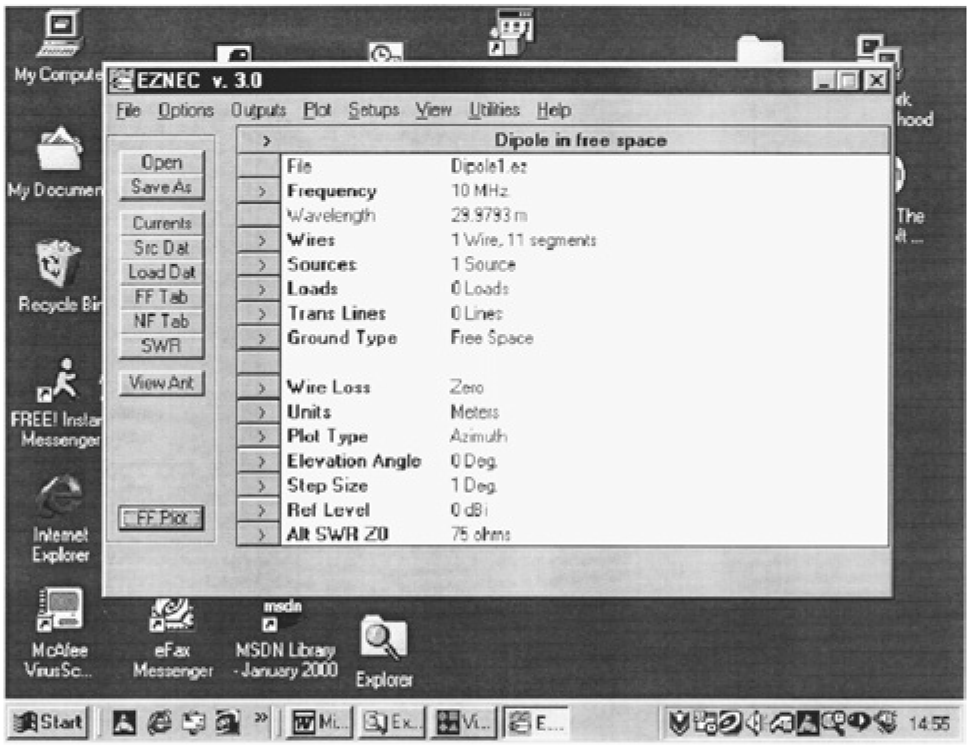
EZNEC v. 3.0
File Options Outputs Plot Setups View Utilities Help
Open
Save As
Currents
Src Dat
Load Dat
FF Tab
NF Tab
SWR
View Ant
FF Plot
My Computer
My Documents
Recycle Bin
FREE! Instant Messenger
Internet Explorer
McAfee VirusSc...
eFax Messenger
MSDN Library - January 2000
Explorer
Start
14:55

Dipole in free space

File Dipole1.ez
Frequency 10 MHz
Wavelength 29.9793 m
Wires 1 Wire, 11 segments
Sources 1 Source
Loads 0 Loads
Trans Lines 0 Lines
Ground Type Free Space

Wire Loss Zero
Units Meters
Plot Type Azimuth
Elevation Angle 0 Deg.
Step Size 1 Deg.
Ref Level 0 dBi
Alt SWR Z0 75 ohms

17-8 Basic window of EZNEC for Windows.

(meters, feet, etc.), plot type (azimuth, elevation), elevation angle, step size, a reference level, and an alternate SWR Z_o.

There are command buttons to the left of the main window. These include Open, Save As, Currents, Sto Dat, Load Dat, FF Tab, NF Tab, SWR, View Antenna, and FF Plot. The "View Antenna" button provides a view of the *xyz* coordinate system along with the wires (Fig. 17-9). The "FF Plot" button plots the far field of the antenna, and the "SWR" button plots the SWR curve based on parameters you give it (starting frequency, stopping frequency, frequency step).

The "View Antenna" window is shown in Fig. 17-9. This antenna view provides the *xyz* coordinate system, along with the wires. There are scroll bars to the left of the main window where the antenna appears that are used to vary the image to better see it. A series of three "Reset" buttons resets the image to the initial setting.

If you wish to plot the radiation pattern (azimuth or elevation) of the antenna, then press "FF Plot" in the main window. The result will be as shown in Fig. 17-10. Some data appears in the lower window on the plot. The data in this case are elevation angle, outer ring gain, slice gain maximum, front-to-side ratio, beamwidth, side-

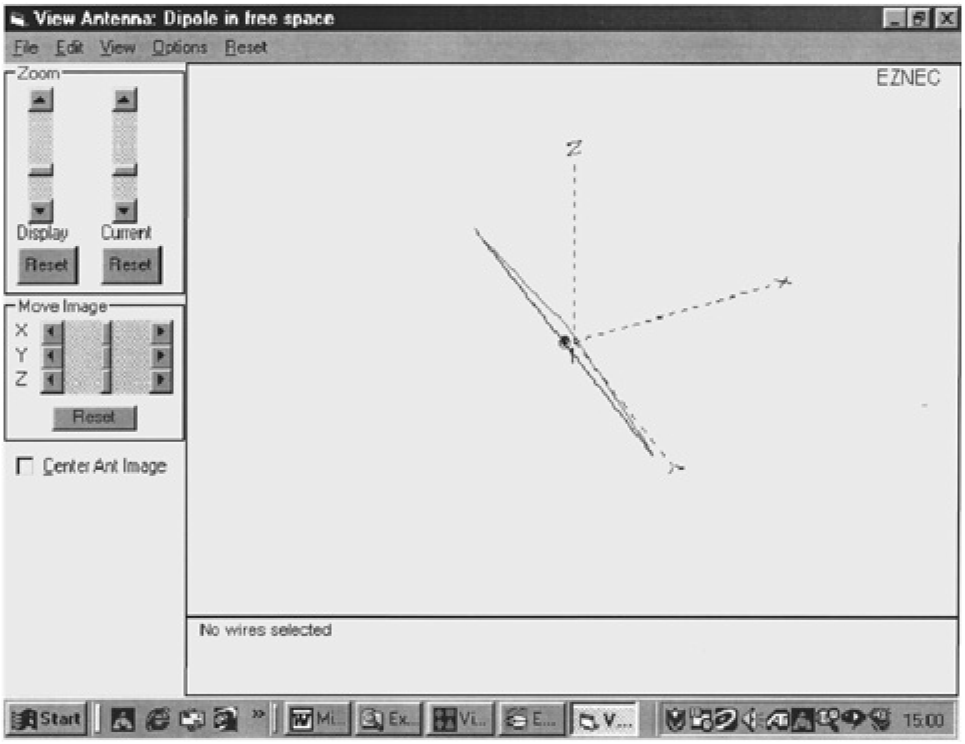
View Antenna: Dipole in free space
File
Edit
View
Options
Reset
Zoom
Display
Current
Reset
Reset
Move Image
X
Y
Z
Reset
☐ Center Ant Image
EZNEC
Z
X
Y
Start
15:00

No wires selected

17-9 View Antenna window of EZNEC for Windows.

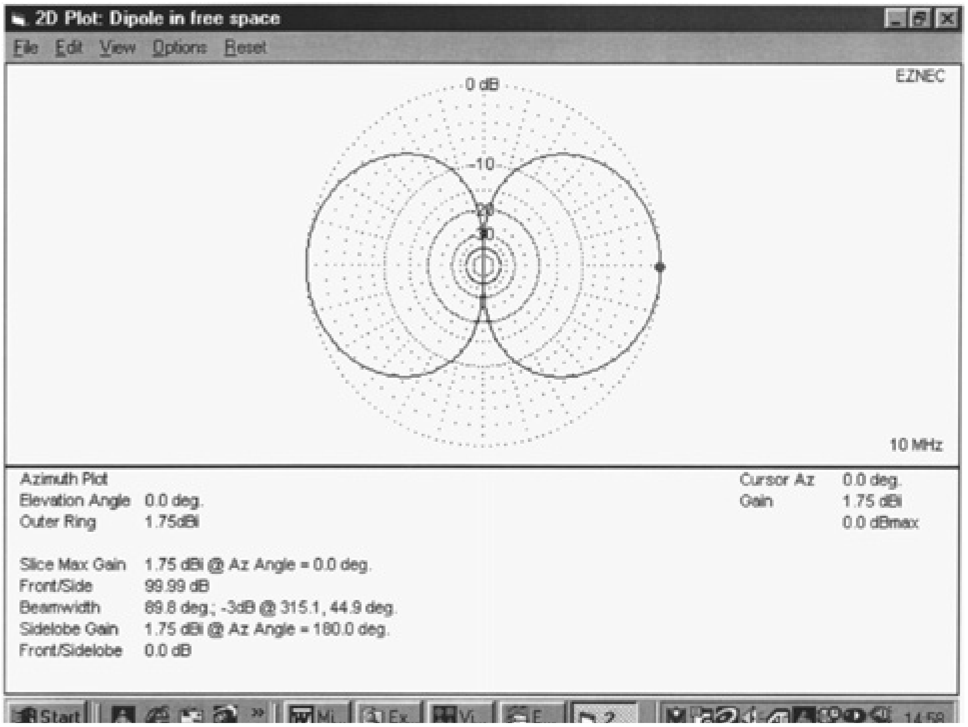

Azimuth Plot
Elevation Angle 0.0 deg.
Outer Ring 1.75dBi

Slice Max Gain 1.75 dBi @ Az Angle = 0.0 deg.
Front/Side 99.99 dB
Beamwidth 89.8 deg.; -3dB @ 315.1, 44.9 deg.
Sidelobe Gain 1.75 dBi @ Az Angle = 180.0 deg.
Front/Sidelobe 0.0 dB

Cursor Az 0.0 deg.
Gain 1.75 dBi
0.0 dBmax

17-10 Result of pressing "FF Plot" button in main window of EZNEC for Windows.

lobe gain, front/sidelobe, cursor azimuth, and gain. You can print out the plot of the radiation pattern with either the data present or the data hidden.

One of the really nice things about EZNEC for Windows is the availability of VSWR plots without going to the professional edition. Figure 17-11 shows the plot for the dipole under consideration. Note that the data given include the starting frequency, the SWR at the starting frequency, the impedance at the starting frequency, the reflection coefficient at the starting frequency, and the source impedance.

References

Burke, G. J., and A. J. Poggio, "Numerical Electromagnetics Code (NEC)—Method of Moments," Naval Ocean Systems Center Technical Document 116, January 1981.

King, R. W. P., *Tables of Antenna Characteristics,* IFI/Plenum Data Corporation, New York, 1971.

Prasad, S., and R. W. P. King, "Experimental Study of L-, T- and Related Transmission Line Antennas," *Journal of Research of the National Bureau of Standards—D Radio Propagation,* Vol. 65D(5): 1961.

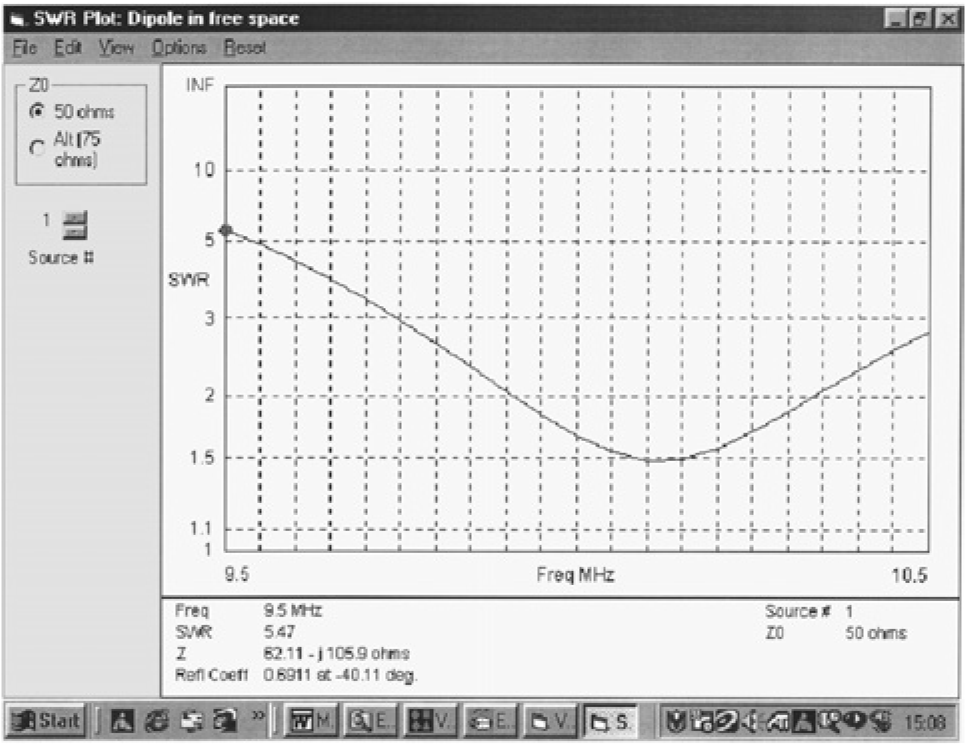
SWR Plot: Dipole in free space
File Edit View Options Reset
Z0
50 ohms
Alt (75 ohms)
1
Source #
INF
10
5
SWR
3
2
1.5
1.1
1
9.5
Freq MHz
10.5
Start
15:08

Freq 9.5 MHz
SWR 5.47
Z 62.11 - j 105.9 ohms
Refl Coeff 0.6911 at -40.11 deg.
Source # 1
Z0 50 ohms

17-11 VSWR plot of a dipole with EZNEC for Windows.

Rockway, J. W., and J. C. Logan, MiniNEC Professional for Windows, EM Scientific, Inc., Carson City, NV, 1995.

Rockway, J. W., and J. C. Logan, MiniNEC Broadcast Professional for Windows, EM Scientific, Inc., Carson City, NV, 1996.

Rockway, J. W., and J. C. Logan, "Advances in MiniNEC," *IEEE Antennas and Propagation Magazine,* 37(4): 1995.

18
CHAPTER

VHF/UHF transmitting and receiving antennas

THE VHF/UHF SPECTRUM IS COMMONLY ACCEPTED TO RANGE FROM 30 MHZ TO 900 MHZ, although the upper breakpoint is open to some differences of opinion. The VHF spectrum is 30 MHz to 300 MHz, and the UHF spectrum is 300 MHz to 900 MHz. Above 900 MHz is the microwave spectrum. These bands are used principally for local “line-of-sight” communications, according to the standard wisdom. However, with the advent of OSCAR satellites, the possibility of long-distance direct communications is a reality for VHF/UHF operators. In addition, packet radio is becoming common; this means indirect long-distance possibilities through networking. For the low end of the VHF spectrum (e.g., 6-m amateur band), long-distance communications are a relatively common occurrence.

In many respects, the low-VHF region is much like the 10-m amateur band and 11-m Citizens Band: skip is not an infrequent occurrence. Many years ago, I recall an event where such skip caused many a local police officer to skip a heart beat. In those days, our police department operated on 38.17 MHz, which is between the 6-m and 10-m amateur bands. They received an emergency broadcast concerning a bank robbery at a certain Wilson Boulevard address. After a race to the county line, they discovered that the reported address would be outside of the county, and in fact did not exist even in the neighboring county (a number was skipped). The problem was traced to a police department in a southwest city that also had a Wilson Boulevard, and for them the alarm was real.

The principal difference between the lower frequencies and the VHF/UHF spectrum is that the wavelengths are shorter in the VHF/UHF region. Consider the fact that the wavelengths for these bands range from 10 m to 1 m for the VHF region, and from 1 m to 33 cm for the UHF region. Most antenna designs are based on wavelength, so that fact has some implications for VHF/UHF antenna design. For example, because bandwidth is a function of length/diameter ratio for many classes of antenna, broadbanding an antenna in the VHF/UHF region is relatively easy. If, say, 25-mm (i.e., 1-in) aluminum tubing is used to make a quarter-wavelength vertical, then the approximate *L/D* ratio is 790 in the 8-m band and 20 in the 2-m band. This feature is fortunate, because the VHF/UHF bands tend to be wider than the HF bands.

Copyright 2001 - the McGraw-Hill Companies

Another point to make is that many of the mechanical chores of antenna design and construction become easier for VHF/UHF antennas. One good example is the delta impedance-matching scheme. At 80 m, the delta-match dimensions are approximately 36 × 43 ft, and at 2 m they are 9.5 × 12 *inches.* Clearly, delta matching is a bit more practical for most users at VHF than at HF.

Types of antennas usable for VHF/UHF

The concept "VHF/UHF antenna" is only partially valid because virtually all forms of antenna can be used at HF, MW, and VHF/UHF. The main limitations that distinguish supposedly VHF/UHF designs from others are mechanical: there are some things that are simply much easier to accomplish with small antennas. Besides the delta match mentioned previously, there is the ease of construction for multielement antennas. A 14-element 20-m beam would be a wonderful thing to have in a QRM-laden DX pile-up, but is simply too impractical for all but a few users because of its size. If you look on embassy rooftops around the world you will see many-element Yagi and log periodic HF antennas supported on massive towers…and some of them use a standard Size 25 tower (common for amateur use) as the *antenna boom*! A 14-element 80/75-m Yagi approaches impossibility. But at 2 m, a 14-element Yagi beam antenna can be carried by one person, in one hand, unless the wind is acting up.

Safety note Large array beams, even at VHF/UHF, have a relatively high "windsail area," and even relatively modest winds can apply a lot of force to them. I once witnessed a large, strong technician blown off a ladder by wind acting on a modest, "suburban" sized, TV antenna. *It can happen to you, too. So always install antennas with a helper, and use hoists and other tools to actually handle the array.*

Lower band antennas on VHF/UHF

Between 1958 and 1962, a friend and I had access to a radio club amateur radio station in a Red Cross chapter house in Virginia. The "antenna farm" consisted of a 14-element 2-m beam, a three-element triband HF beam (10, 15, and 20 m), and a five-band (80 to 10 m) trap dipole. All of the coaxial cables came into the station through a wall; they were kept disconnected and shorted out when not in use because of the senior Red Cross official's concern over lightning.

One night, attempting to connect the 2-m beam to the Gonset "gooney box" 2-m AM transceiver, my friend accidentally used the cable from the five-band trap dipole instead. We worked a lot of stations that contest weekend, and scored lots of points. Later, we discovered the error, and asked a more technically competent adult (we were teenagers), "Why the good reports?" He then gave us a lesson in longwire antenna theory. A good longwire is many wavelengths long. Consider that a half-wave antenna on 2 m is 80 m/2 m or 40 wavelengths shorter than an 80-m half-wave antenna. Thus, the 80-m antenna, counting foreshortening of physical lengths because of the traps, was on the order of 35 to 38 wavelengths long on 2 m. We had a highly directional, but multilobed, pattern.

Similarly, 40- to 10-m and 80- to 10-m trap verticals are often usable on VHF/UHF frequencies without any adjustments. Similarly, Citizens Band 11-m antennas, many of which are $\frac{5}{8}$-wavelength (18 ft high), will sometimes work on VHF frequencies. Check the VSWR of an HF antenna on 2 m with a reliable VHF/UHF VSWR meter (or RF wattmeter) to discover the truth about any particular antenna. Always use the low-power setting on the transmitter to limit damage in cases where the specific antenna is not usable on a specific frequency.

The lesson to be learned is that antennas are often usable on frequencies much higher than the design frequency, even though useless on nearby bands. Care must be exercised when initially checking out the antenna, but that is not an inordinate difficulty.

VHF/UHF antenna impedance matching

The VHF/UHF antenna is no more or less immune from the need for impedance matching than lower-frequency antennas. However, some methods are easier (coax baluns, delta match, etc.) and others become either difficult or impossible. An example of the latter case is the tuned *LC* impedance-matching network. At 6 m, and even to some limited extent 2 m, inductor and capacitor *LC* networks can be used. But above 2 m other methods are more reasonable. We can, however, mimic the *LC* tuner by using stripline components, but that approach is not always suited to amateur needs.

The balun transformer makes an impedance transformation between *bal*anced and *un*balanced impedances. Although both 1:1 and 4:1 impedance ratios are possible, the 4:1 ratio is most commonly used for VHF/UHF antenna work. At lower frequencies it is easy to build broadband transformer baluns, but these become more of a problem at VHF and above.

For the VHF/UHF frequencies, a 4:1 impedance ratio coaxial balun (Fig. 18-1A) is normally used. Two sections of identical coaxial cable are needed. One section (*A*) has a convenient length to reach between the antenna and the transmitter. Its characteristic impedance is Z_o. The other section (*B*) is a half-wavelength long at the center of the frequency range of interest. The physical length is found from

$$L = \frac{5904\,V}{F_{\mathrm{MHz}}} \qquad \mathbf{[18.1]}$$

where

L is the cable length, in inches
F_{MHz} is the operating frequency, in megahertz
V is the velocity factor of the coaxial cable

The velocity factors of common coaxial cables are shown in the following table.

Coaxial cable velocity factors

Regular polyethylene	0.66
Polyethylene foam	0.80
Teflon	0.72

Chap18_Carr 4/10/01 3:44 PM Page 342

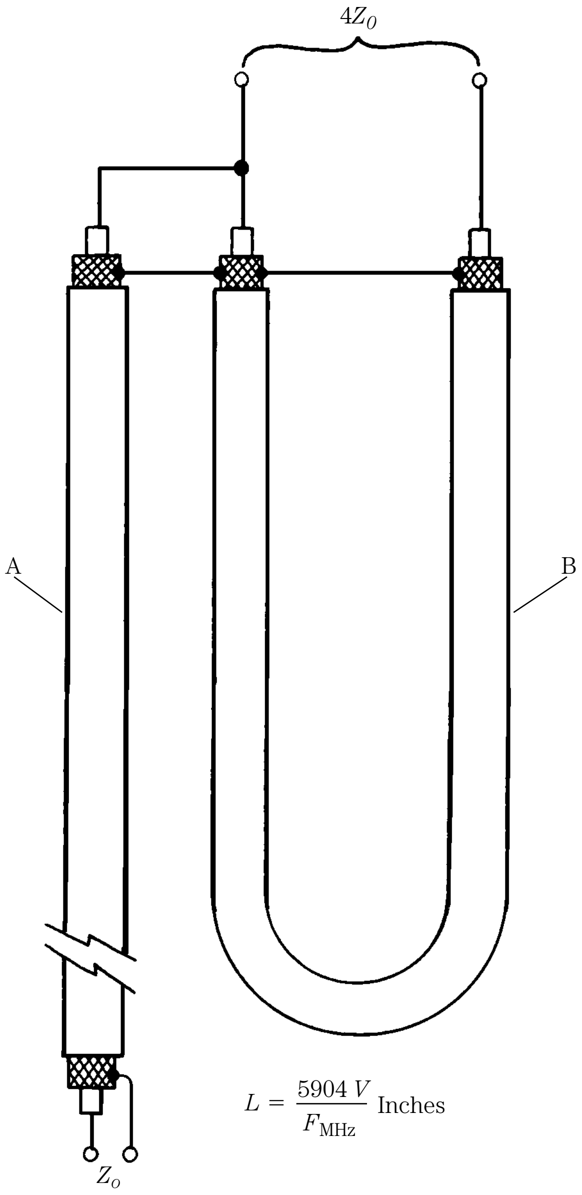
$4Z_0$
A
B
$L = \frac{5904\,V}{F_{\text{MHz}}}$ Inches
Z_o

18-1A Coaxial 4:1 balun transformer.

Example 18-1 Calculate the physical length required of a 146-MHz 4:1 balun made of polyethylene foam coaxial cable.

Solution:

$$L = \frac{5904\,V}{F_{\text{MHz}}}$$

$$= \frac{(5904)\ (0.80)}{146 \text{ MHz}}$$

$$= \frac{4723.2}{146 \text{ MHz}} = 32.4 \text{ in}$$

A mechanical method of joining the coaxial cables is shown in Fig. 18-1B. In this example, three SO-239 coaxial receptacles are mounted on a metal plate. This arrangement has the effect of shorting together the shields of the three ends of coaxial cable. The center conductors are connected in the manner shown. This method is used especially where a mounting bracket is available on the antenna. The lengths of coaxial cable need PL-259 coax connectors installed in order to use this method.

The "delta match" gets its name from the fact that the structure of the matching element has the shape of the Greek letter delta, or a triangle. Figure 18-2A shows the basic delta match scheme. The matching element is attached to the driven element of the antenna (symmetrically, about the center point of the antenna). The width (A) of the delta match is given by

$$L = \frac{1416}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.2]}$$

While the height of the match (B) is

$$L = \frac{1776}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.3]}$$

The transmission line feeding the delta match is balanced line, such as parallel transmission line or twin lead. The exact impedance is not terribly critical because the dimensions (especially A) can be adjusted to accommodate differences. In general, however, either 450- or 600-Ω line is used, although 300-Ω line can also be used. Figure 18-2B

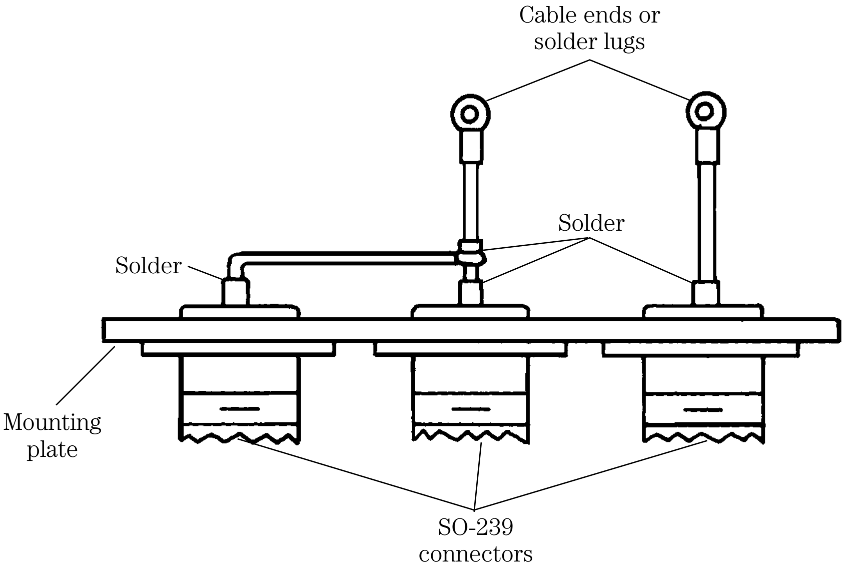

18-1B Practical implementation of 4:1 balun using connectors.

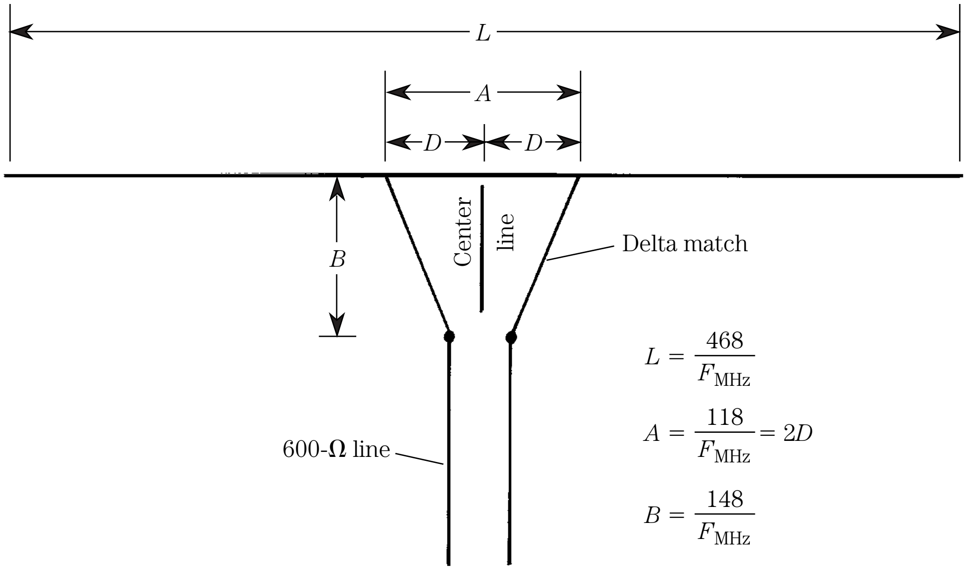
L
A
D
D
Center line
B
Delta match
600-Ω line
$L = \frac{468}{F_{\mathrm{MHz}}}$
$A = \frac{118}{F_{\mathrm{MHz}}} = 2D$
$B = \frac{148}{F_{\mathrm{MHz}}}$

18-2A Delta feed matching system.

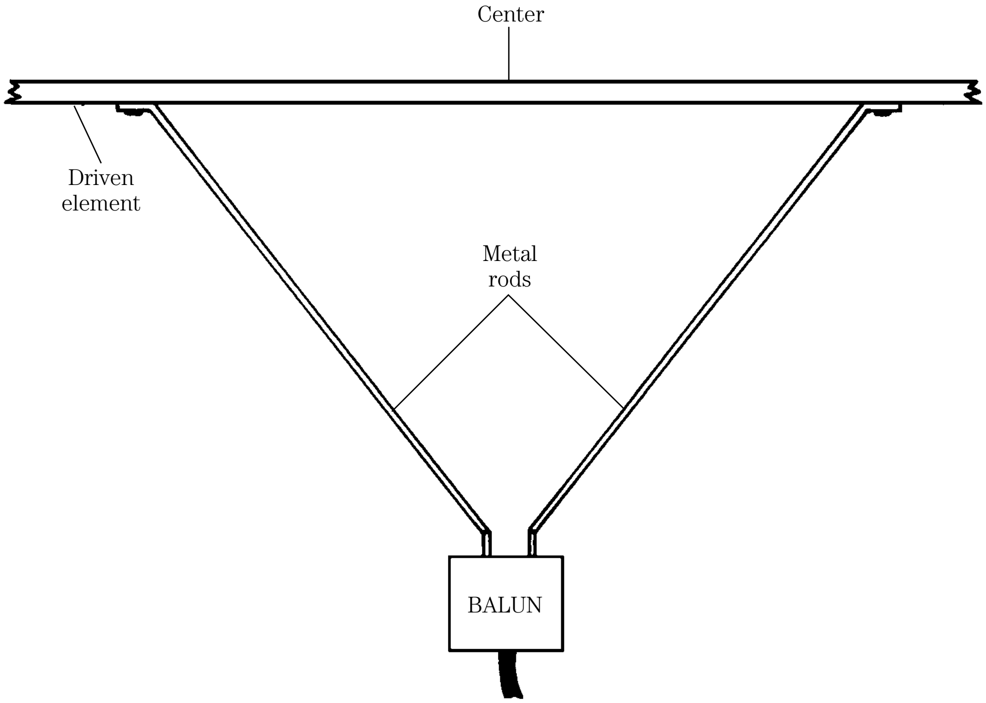
Center
Driven element
Metal rods
BALUN

18-2B Practical VHF delta match.

shows a method for using coaxial cable with the delta match. The impedance is transformed in a 4:1 balun transformer (see Fig. 18-1A). The elements of the delta match can be made from brass, copper, or aluminum tubing, or a bronze brazing rod bolted to the main radiator element.

A stub-matching system is shown in Fig. 18-3. In this case, the impedance transformation is accomplished despite the half-wavelength shorted stub of transmission line. The exact impedance of the line is not very critical, and is found from

$$Z_o = 276 \log_{10} \frac{2S}{d} \qquad \textbf{[18.4]}$$

The matching stub section is made from metal elements such as tubing, wire, or rods (all three are practical at VHF/UHF frequencies). For a $^3\!/_{16}$-in rod, the spacing is approximately 2.56 in to make a 450-Ω transmission line. A sliding short circuit is used to set the electrical length of the half-wave stub. The stub is tapped at a distance

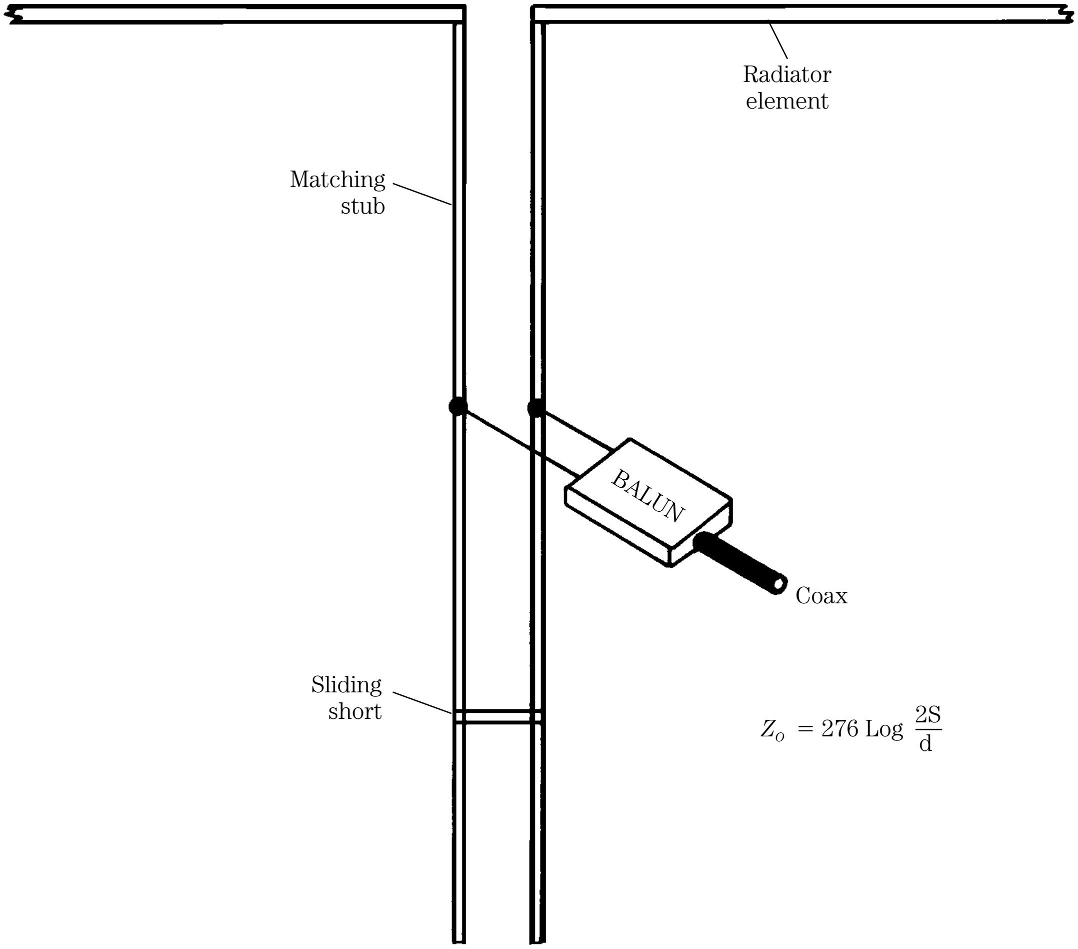

18-3 Stub matching.

from the antenna feedpoint that matches the impedance of the transmission line. In the example shown, the transmission line is coaxial cable, so a 4:1 balun transformer is used between the stub and the transmission line. The two adjustments to make in this system are: (1) the distance of the short from the feedpoint and (2) the distance of the transmission line tap point from the feedpoint. Both are adjusted for minimum VSWR.

The gamma match is basically a half delta match, and operates according to similar principles (Fig. 18-4). The shield (outer conductor) of the coaxial cable is connected to the center point of the radiator element. The center conductor of the coaxial cable series feeds the gamma element through a variable capacitor.

VHF/UHF antenna examples

Although it is probably not necessary to reiterate the point, VHF/UHF antennas are not substantially different from HF antennas, especially those for the higher bands. However, for various practical reasons there are several forms that are specially suited, or at least popular, in the VHF/UHF bands. In this section we will take a look at some of them.

Coaxial vertical

The coaxial vertical is a quarter-wavelength vertically polarized antenna that is popular on VHF/UHF. There are two varieties. In Fig. 18-5A we see the coaxial antenna made with coaxial cable. Although not terribly practical for long-term installation, the coax-coax antenna is very useful for short-term, portable, or emergency applications. For example, a boater found himself adrift, and in dire trouble, after a storm damaged the boat. The mast-top VHF antenna was washed away, leaving only the end of the coaxial cable dangling loose. Fortunately, the boat operator was a two-way radio technician, and he knew how to strip back the coaxial cable to make an impromptu coaxial vertical.

The coax-coax antenna shown in Fig. 18-5A uses a quarter-wavelength radiator and a quarter-wavelength sleeve. The sleeve consists of the coax braid stripped back and folded down the length of the coax cable. The maximum length is found from the equation below (actual length is trimmed from this maximum):

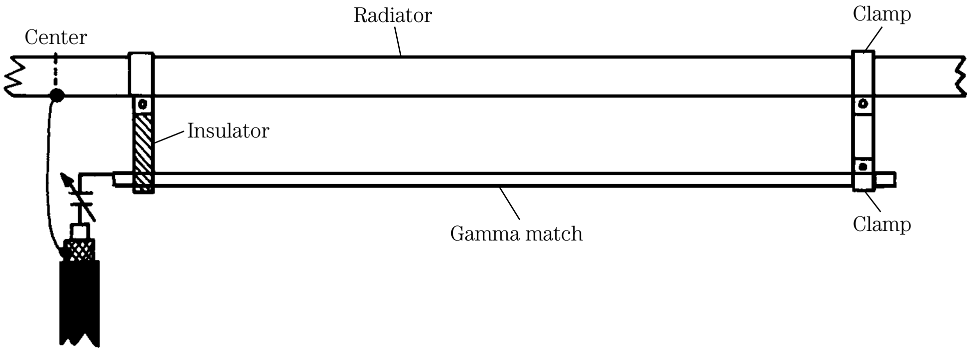

18-4 Gamma matching.

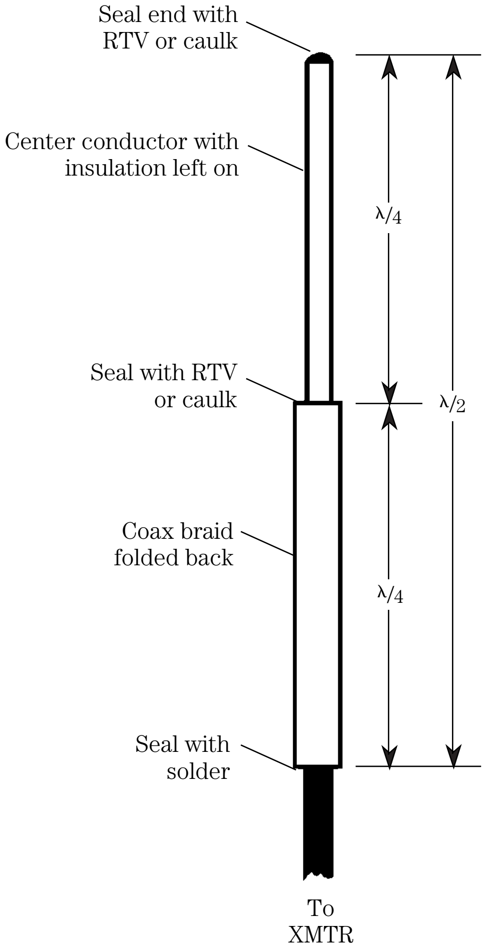
Seal end with RTV or caulk
Center conductor with insulation left on
Seal with RTV or caulk
Coax braid folded back
Seal with solder
To XMTR
λ/4
λ/2
λ/4

18-5A Coaxial vertical based on coaxial cable.

$$L = \frac{2952\ V}{F_{\mathrm{MHz}}}\ \text{in} \qquad \textbf{[18.5]}$$

The antenna is mounted by suspending it from above by a short piece of string, twine, or fishing line. From a practical point of view, the only problem with this form of antenna is that it tends to deteriorate after a few rainstorms. This effect can be reduced by sealing the end, and the break between the sleeve and the radiator, with either silicone *RTV* or bathtub caulk.

A more permanent method of construction is shown in Fig. 18-5B. The sleeve is a piece of copper or brass tubing (pipe) about 1 in in diameter. An end cap is fitted over the end and sweat-soldered into place. The solder is not intended to add mechanical strength, but rather to prevent weathering from destroying the electrical contact between the two pieces. An SO-239 coaxial connector is mounted on the end cap. The coax is connected to the SO-239 inside the pipe, which means making the connection before mounting the end cap.

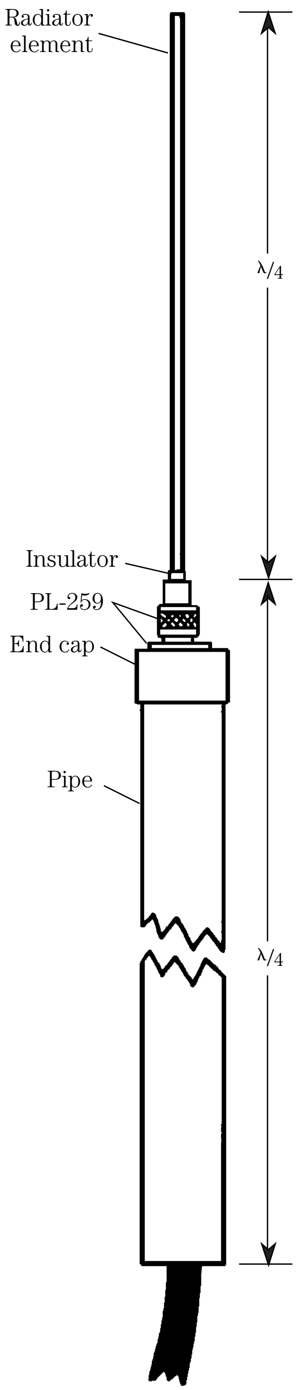
Radiator element
Insulator
PL-259
End cap
Pipe
λ/4
λ/4

18-5B Tubing coaxial vertical.

The radiator element is a small piece of tubing (or brazing rod) soldered to the center conductor of a PL-259 coaxial connector. An insulator is used to prevent the rod from shorting to the outer shell of the PL-259. (Note: an insulator salvaged from the smaller variety of banana plug can be shaved a small amount with a fine file and made to fit inside the PL-259. It allows enough center clearance for $^1\!/_8$-inch or $^3\!/_{16}$-inch brass tubing.)

Alternatively, the radiator element can be soldered to a banana plug. The normal-size banana plug happens to fit into the female center conductor of the SO-239.

Collinear vertical

Gain in antennas is provided by *directivity.* In other words, by taking the power radiated by the antenna, and projecting it into a limited direction, we obtain the appearance of higher radiated power. In fact, the *effective radiated power* (ERP) of the antenna is merely its feedpoint power multiplied by its gain. Although most antenna patterns are shown in the horizontal dimension (as viewed from above), it is also possible to obtain gain by compressing the vertical aspect. In this manner it is possible to have a vertical antenna that produces gain. Figure 18-6 shows a collinear gain antenna, with vertical polarization and a horizontally omnidirectional pattern. Incidentally, when mounted horizontally the pattern becomes bidirectional.

The collinear antenna shown in Fig. 18-6 is basically a pair of stacked collinear arrays. Each array consists of a quarter-wavelength section *A* and a half-wavelength section *C* separated by a quarter-wavelength phase reversing stub *B.* The phase reversal stub preserves in-phase excitation for the outer element (referenced to the inner element).

The feedpoint is between the two elements of the array (i.e., between the *A* sections). The coaxial-cable impedance is transformed by a 4:1 balun transformer (see Fig. 18-1A). Alternatively, 300-Ω twin lead can be used for the transmission line. If this alternative is used, then the use of UHF shielded twin lead is highly recommended. If the transmitter lacks the balanced output needed to feed twin lead, then use a balun at the input end of the twin lead (i.e., right at the transmitter).

Yagi antennas

The Yagi beam antenna is a highly directional gain antenna, and is used both in HF and VHF/UHF systems. The antenna is relatively easy to build at VHF/UHF. In fact, it is easier than for HF systems. The basic Yagi was covered in Chap. 12, so we will only show examples of practical VHF devices. A 6-m Yagi antenna is shown in Fig. 18-7. This particular antenna is a four-element model. The reflector and directors can be mounted directly to a metallic boom, because they are merely parasitic. The driven element, however, must be insulated from the metal boom.

The driven element shown in Fig. 18-7 is a folded dipole. While this is common practice at VHF, because it tends to broadband the antenna, it is not strictly necessary. The dimensions of the driven element are found from Eq. 18.4. Set the equation equal to 300 Ω, select the diameter of the tubing from commercially available sources, and then calculate the spacing.

Example 18-2 Calculate the spacing of a 300-Ω folded dipole when $^3\!/_4$-in tubing is used in its construction.

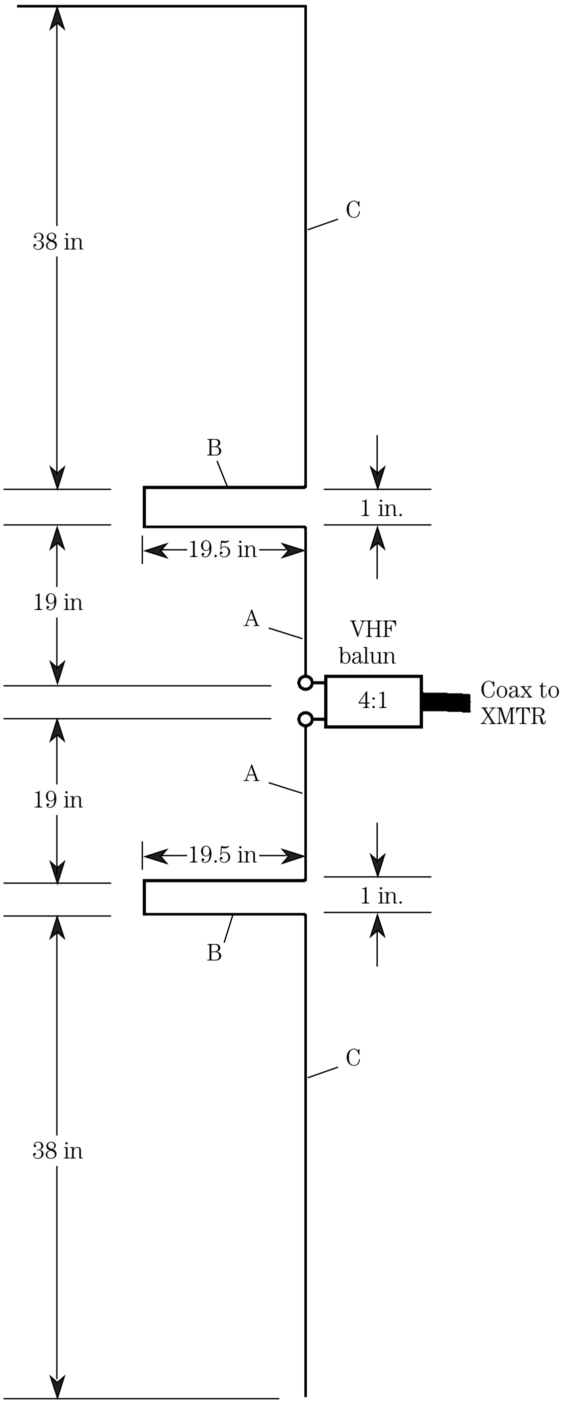
C
38 in
B
1 in.
19.5 in
19 in
A
VHF balun
4:1
Coax to XMTR
A
19 in
19.5 in
1 in.
B
C
38 in

18-6 Vertical collinear antenna.

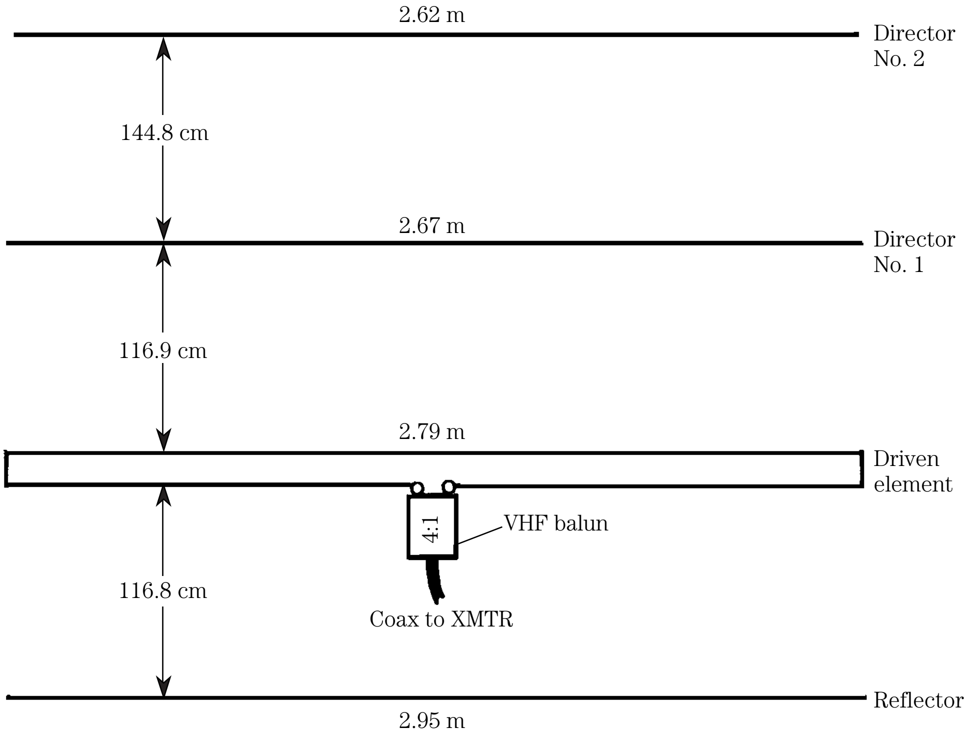
2.62 m
Director No. 2
144.8 cm
2.67 m
Director No. 1
116.9 cm
2.79 m
Driven element
4:1
VHF balun
116.8 cm
Coax to XMTR
Reflector
2.95 m

18-7 Six-meter beam antenna.

Solution:

$$Z_o = 276 \log_{10} \frac{2S}{d}$$

$$300\ \Omega = 276 \log_{10} \frac{2S}{d}$$

$$\frac{300}{276} = \log_{10} \frac{2S}{d}$$

$$1.08 = \log_{10} \frac{2S}{d}$$

$$10^{1.08} = \frac{2S}{d}$$

Using 0.75-in pipe results in

$$10^{1.08} = \frac{2S}{0.75}$$

$$12.02 = 2.67S$$

so

$$S = 4.5 \text{ in}$$

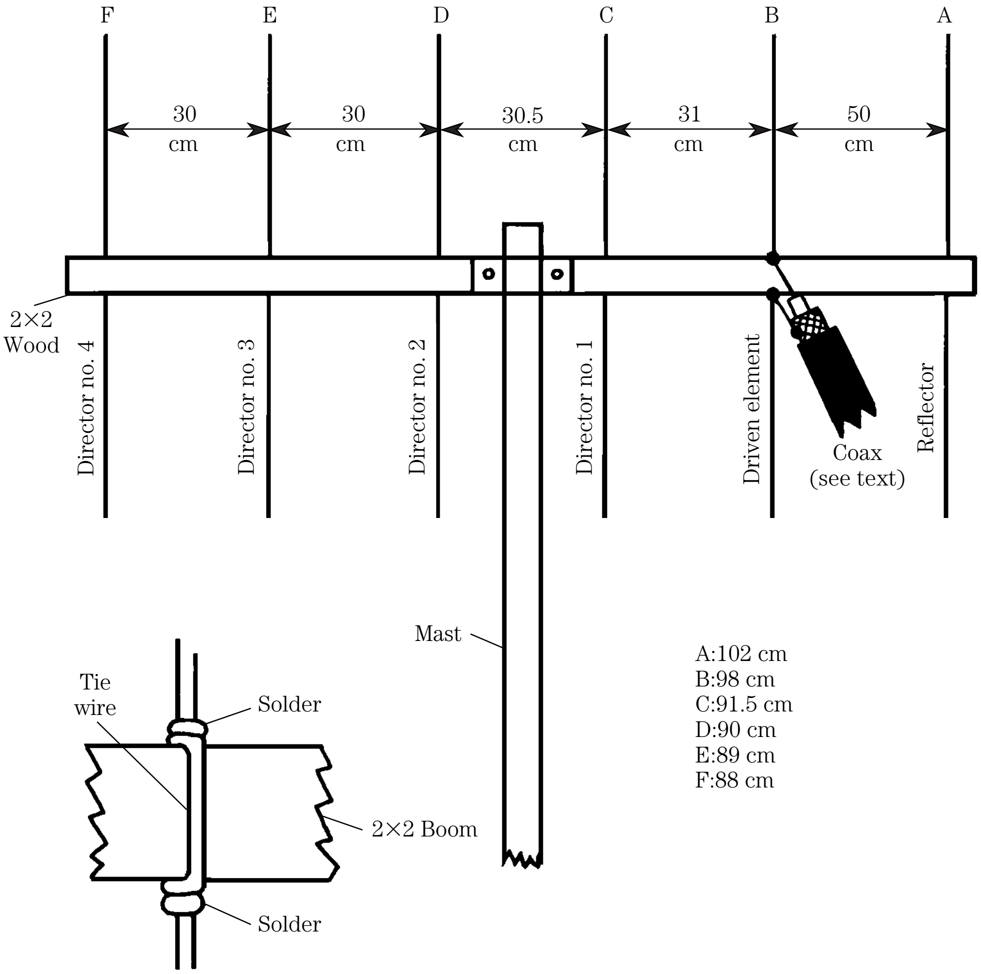
F
E
D
C
B
A
30 cm
30 cm
30.5 cm
31 cm
50 cm
2×2 Wood
Director no. 4
Director no. 3
Director no. 2
Director no. 1
Driven element
Coax (see text)
Reflector
Mast
A:102 cm
B:98 cm
C:91.5 cm
D:90 cm
E:89 cm
F:88 cm
Tie wire
Solder
2×2 Boom
Solder

18-8 Two-meter vertical beam.

Two-meter yagi

Figure 18-8 shows the construction details for a six-element 2-meter Yagi beam antenna. This antenna is built using a 2 × 2-in wooden boom and elements made of either brass or copper rod. Threaded brass rod is particularly useful, but not strictly necessary. The job of securing the elements (other than the driven element) is easier when threaded rod is used, because it allows a pair of hex nuts, one on either side of the 2 × 2-in boom, to be used to secure the element. Nonthreaded elements can be secured with RTV sealing a press-fit. Alternatively, tie wires (see inset to Fig. 18-8) can be used to secure the rods. A hole is drilled through the 2 × 2 to admit the rod or tubing. The element is secured by wrapping a tie wire around the rod on either side of the 2 × 2, and then soldering it in place. The tie wire is no. 14 to no. 10 solid wire.

Mounting of the antenna is accomplished by using a mast secured to the boom with an appropriate clamp. One alternative is to use an end-flange clamp, such as is sometimes used to support pole lamps, etc. The mast should be attached to the boom at the center of gravity, which is also known as the *balance point.* If you try to balance the antenna in one hand unsupported, there is one (and only one) point at which it is balanced (and won't fall). Attach the mast hardware at, or near, this point in order to prevent normal gravitational torques from tearing the mounting apart.

The antenna is fed with coaxial cable at the center of the driven element. Ordinarily, either a matching section of coax, or a gamma match, will be needed because the effect of parasitic elements on the driven element feedpoint impedance is to reduce it.

⅝-wavelength 2-m antenna

The ⅝-wavelength antenna (Fig. 18-9) is popular on 2 m for mobile operation because it is easy to construct, and it provides a small amount of gain relative to a dipole. The radiator element is ⅝-wavelength, so its physical length is found from:

$$L = \frac{7380}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.6]}$$

The ⅝-wavelength antenna is not a good match to any of the common forms of coaxial cables. Either a matching section of cable, or an inductor match, is normally used. In Fig. 18-9 an inductor match is used. The matching coil consists of 2 to 3 turns of no. 12 wire, wound over a ½-in OD form, ½-in long. The radiator element can be tubing, brazing rod, or a length of heavy "piano wire." Alternatively, for low-power systems, it can be a telescoping antenna that is bought as a replacement for portable radios or televisions. These antennas have the advantage of being adjustable to resonance without the need for cutting.

J-pole antennas

The *J-pole antenna* is another popular form of vertical on the VHF bands. It can be used at almost any frequency, although the example shown in Fig. 18-10 is for 2 m. The antenna radiator is ¾-wavelength long, so its dimension is found from

$$L = \frac{8838}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.7]}$$

and the quarter-wavelength matching section length from

$$L = \frac{2952}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.8]}$$

Taken together the matching section and the radiator form a parallel transmission line with a characteristic impedance that is 4 times the coaxial cable impedance. If 50-Ω coax is used, and the elements are made from 0.5 in OD pipe, then a spacing of 1.5 in will yield an impedance of about 200 Ω. Impedance matching is accomplished by a gamma match consisting of a 25-pF variable capacitor, connected by a clamp to the radiator, about 6 in (experiment with placement) above the base.

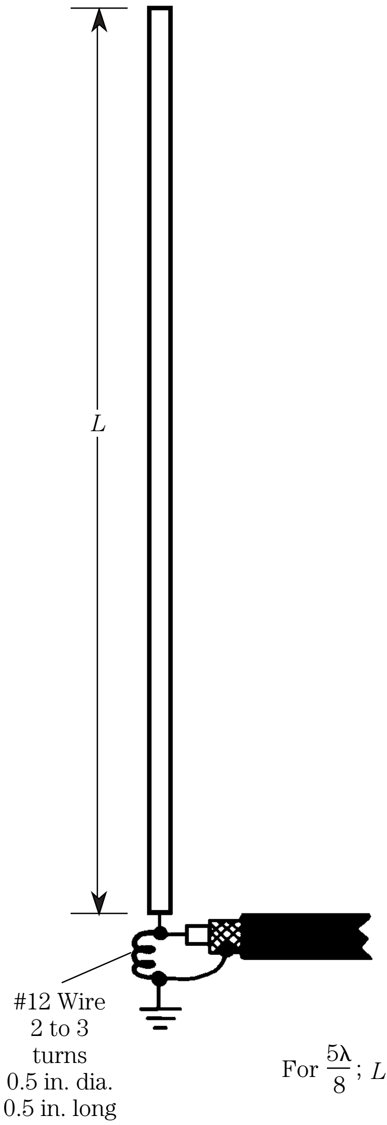
L
#12 Wire
2 to 3
turns
0.5 in. dia.
0.5 in. long

$$\text{For } \frac{5\lambda}{8};\ L_{FT} = \frac{615}{F_{MHz}}$$

$$L_{IN} = \frac{7380}{F_{MHz}}$$

18-9 $\frac{5}{8}$-wavelength 2-m antenna.

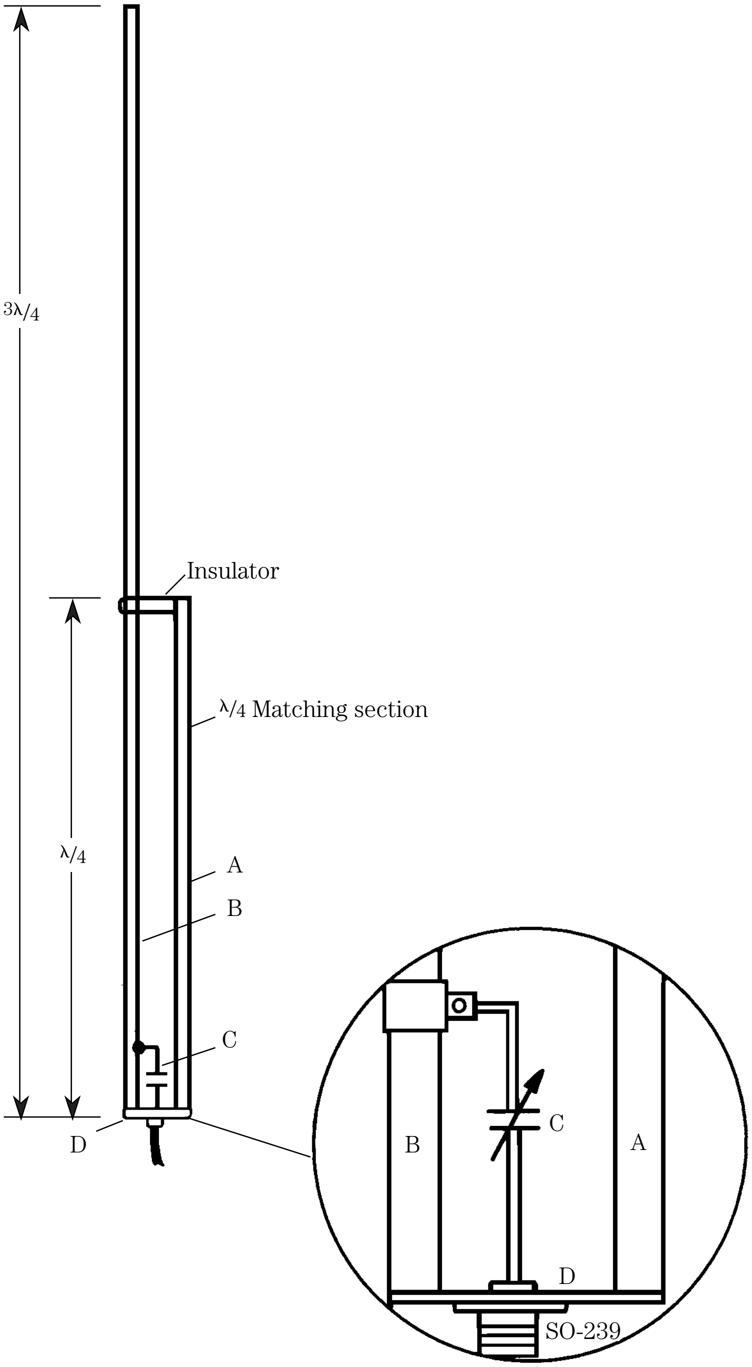
3λ/4
Insulator
λ/4 Matching section
λ/4
A
B
C
D
B
C
A
D
SO-239

18-10 J-pole antenna.

Groundplane

The *groundplane antenna* is a vertical radiator situated above an artificial RF ground consisting of quarter-wavelength radiators. Groundplane antennas can be either $\frac{1}{4}$-wavelength or $\frac{5}{8}$-wavelength (although for the latter case impedance matching is needed—see the previous example).

Figure 18-11 shows how to construct an extremely simple groundplane antenna for 2 m and above. The construction is too lightweight for 6-m antennas (in general), because the element lengths on 6-m antennas are long enough to make their weight too great for this type of construction. The base of the antenna is a single SO-239 chassis-type coaxial connector. Be sure to use the type that requires four small machine screws to hold it to the chassis, and not the single-nut variety.

The radiator element is a piece of $\frac{3}{16}$-in or 4-mm brass tubing. This tubing can be bought at hobby stores that sell airplanes and other models. The sizes quoted just happen to fit over the center pin of an SO-239 with only a slight tap from a lightweight hammer—and I do mean *slight* tap. If the inside of the tubing and the connector pin are pretinned with solder, then sweat soldering the joint will make a good electrical connection that is resistant to weathering. Cover the joint with clear lacquer spray for added protection.

The radials are also made of tubing. Alternatively, rods can also be used for this purpose. At least four radials are needed for a proper antenna (only one is shown in Fig. 18-11). This number is optimum because they are attached to the SO-239 mounting holes, and there are only four holes. Flatten one end of the radial, and drill a small hole in the center of the flattened area. Mount the radial to the SO-239 using small hardware (4-40, etc.).

The SO-239 can be attached to a metal L bracket. While it is easy to fabricate such a bracket, it is also possible to buy suitable brackets in any well-equipped hardware store. While shopping at one do-it-yourself type of store, I found several reasonable candidate brackets. The bracket is attached to a length of 2 × 2-in lumber that serves as the mast.

Halo antennas

One of the more saintly antennas used on the VHF boards is the halo (Fig. 18-12). This antenna basically takes a half-wavelength dipole and bends it into a circle. The ends of the dipole are separated by a capacitor. In some cases, a transmitting-type mica “button” capacitor is used, but in others (and perhaps more commonly), the halo capacitor consists of two 3-in disks separated by a plastic dielectric. While air also serves as a good (and perhaps better) dielectric, the use of plastic allows mechanical rigidity to the system.

Quad beam antennas

The quad antenna was introduced in the chapter on beams. It is, nonetheless, also emerging as a very good VHF/UHF antenna. It should go without saying that the antenna is a lot easier to construct at VHF/UHF frequencies than it is at HF frequencies! Figure 18-13 shows a modest example.

There are several methods for building the quad antenna, and Fig. 18-13 represents only one of them. The radiator element can be any of several materials,

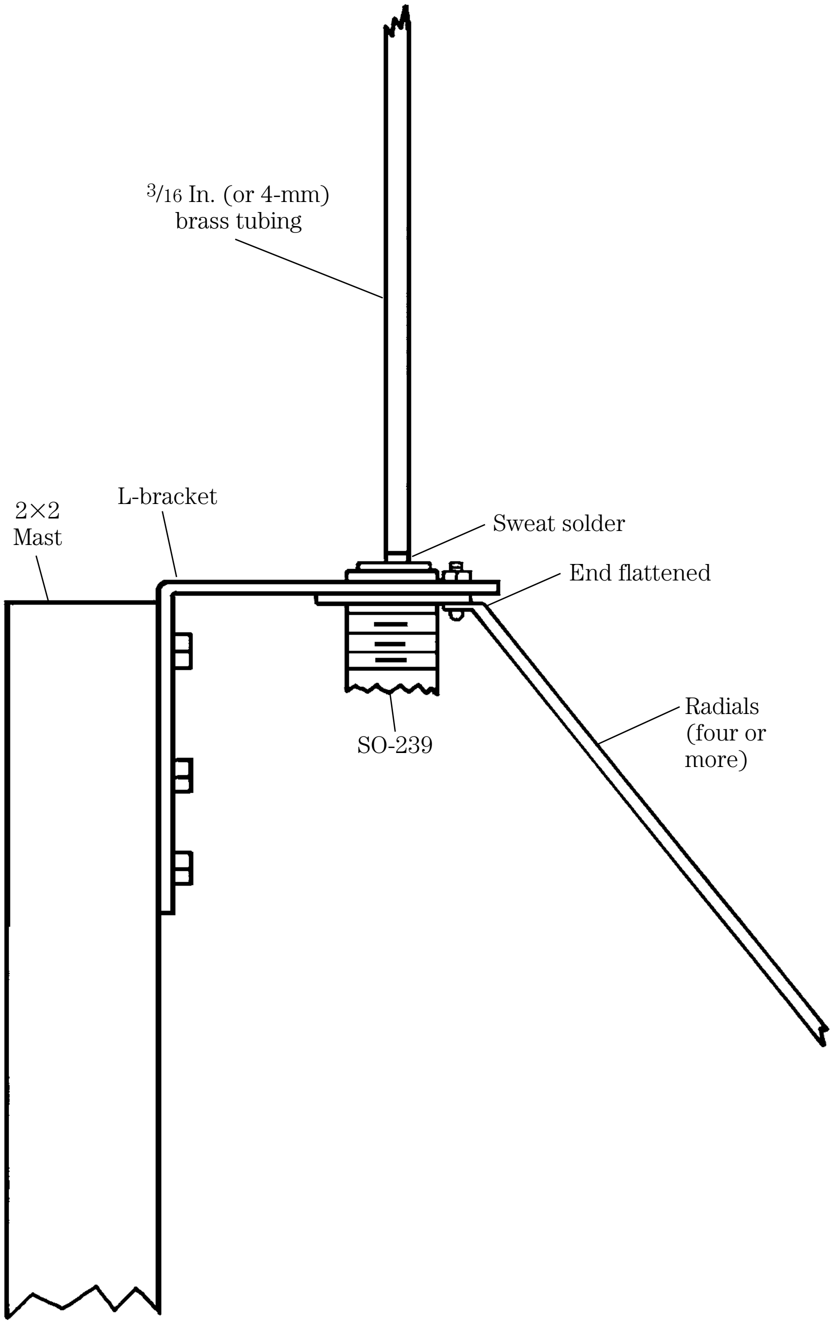
3/16 In. (or 4-mm)
brass tubing
L-bracket
2×2
Mast
Sweat solder
End flattened
Radials
(four or
more)
SO-239

18-11 Small VHF groundplane construction.

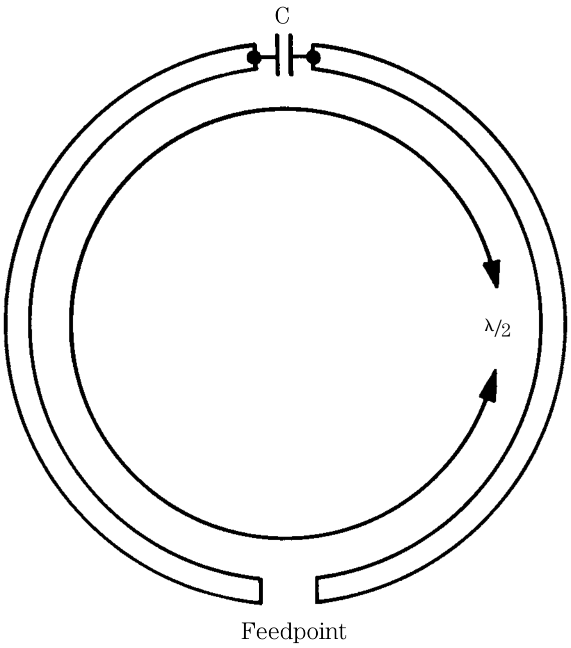

18-12 Halo antenna.

including heavy solid wire (no. 8 to no. 12), tubing, or metal rods. The overall lengths of the elements are given by:

Driven element:

$$L = \frac{11{,}826}{F_{\mathrm{MHz}}}\ \text{in} \tag{18.9}$$

Reflector:

$$L = \frac{12{,}562}{F_{\mathrm{MHz}}}\ \text{in} \tag{18.10}$$

Director:

$$L = \frac{11{,}248}{F_{\mathrm{MHz}}}\ \text{in} \tag{18.11}$$

There are several alternatives for making the supports for the radiator. Because of the lightweight construction, almost any method can be adapted for this purpose. In the case shown in Fig. 18-13, the spreaders are made from either 1-in furring strips, trim strips, or (at above 2 m) even wooden paint stirring sticks. The sticks are cut to length, and then half-notched in the center (Fig. 18-13, detail B). The two spreaders for each element are joined together at right angles and glued (Fig. 18-13, detail C). The spreaders can be fastened to the wooden boom at points S in detail C. The usual rules regarding element spacing (0.15 to 0.31 wavelength) are followed. See the information on quad antennas in Chap. 12 for further details. Quads have been successfully built for all amateur bands up to 1296 MHz.

Overall lengths:

$$\text{Reflector: } L = \frac{12{,}562}{F_{\text{MHz}}} \text{ in}$$

$$\text{Director: } L = \frac{11{,}248}{F_{\text{MHz}}} \text{ in}$$

$$\text{Driven element: } L = \frac{11{,}826}{F_{\text{MHz}}} \text{ in}$$

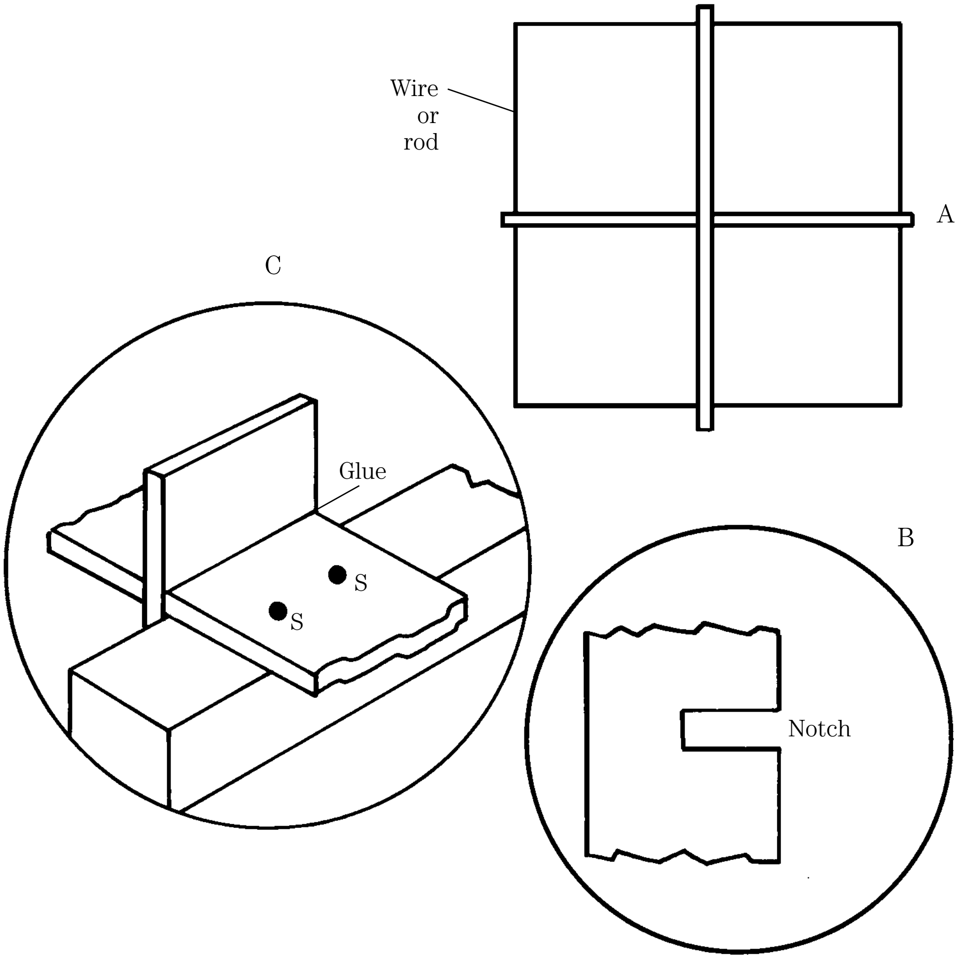

18-13 Quad loop construction on VHF.

VHF/UHF scanner band antennas

The hobby of shortwave listening has always had a subset of adherents who listen exclusively to the VHF/UHF bands. In those bands are found a rich variety of services from commercial two-way radios, police and fire systems, and others. Many of these people are serious DXers, but others have a little of the voyeur in them (they like to listen to what's going on in the community by monitoring the police frequencies). As long as they neither respond to the scene of a police or fire action, nor use the information gained from monitoring in an illegal manner, they are perfectly free to listen in.

A few people have an unusually practical element to their VHF/UHF listening. At least one person known to the author routinely tunes in the local taxicab company's frequency as soon as she orders a cab. She then listens for her own address. She knows from that approximately when to expect the cab.

In the early 1960s, when all VHF monitor receivers were crystal-controlled (and relatively expensive), the listener had a very limited selection of frequencies to choose from. Having worked in various shops that sold those receivers, I can recall that most of our customers tended to be police officers, fire fighters (or volunteers), or journalists covering the local crime news. However, modern scanners operate on both major VHF bands (low and high), plus UHF bands and others.

Scanner-vision antennas

The antennas used by scanner listeners are widely varied, and (in some cases) overpriced. Although it is arguable that a total coverage VHF/UHF antenna is worth the money, and it probably is, there are other possibilities that should be considered.

First, don't overlook the use of television antennas for scanner monitoring! The television bands (about 80 channels from 54 MHz to around 800 MHz) encompass most of the ordinarily used scanner frequencies. Although antenna performance is not optimized for the scanner frequencies, it is also not "zero" on those frequencies. If you already have an "all channel" TV antenna installed, then it is a simple matter to connect the antenna to the scanner receiver. (Note: if the antenna uses 300-Ω twin lead, then install a 4:1 balun transformer that accepts 300 Ω in and produces 75 Ω out. These transformers are usually available at TV shops, video shops—including videotape rental places—and Radio Shack stores.)

The directional characteristic of the TV antenna makes it both an advantage and a disadvantage to the scanner user. If the antenna has a rotator, then there is no problem. Just rotate the antenna to the direction of interest. However, if the antenna is fixed, and the station of interest is elsewhere than where the antenna is pointed, then there is a bit of a problem. Nevertheless, some reception is possible for at least three reasons. One is that the main beam of the antenna is not infinitely thin, so the station of interest might well be within the beam—even if off the point a little bit. Second, there are always sidelobes on an antenna. These are areas outside of the main beam that offer reduced, but nonzero, reception characteristics. Finally, the sidelobes and main lobe of the antenna are optimized for the TV bands, and may not obtain the same directivity on certain scanner frequencies. It is, therefore, possible that a TV antenna will have an unusual lobe in the direction of interest for the scanner channel.

The TV antenna also offers other possibilities. The hardware and electronics, normally found useful on TV antenna systems, are also useful for scanner antennas. The use of the mounting hardware is obvious, but less obvious is the use of components such as multiset couplers, impedance transformers, and wideband amplifiers. For example, Fig. 18-14 shows two scanner antennas joined together into a single

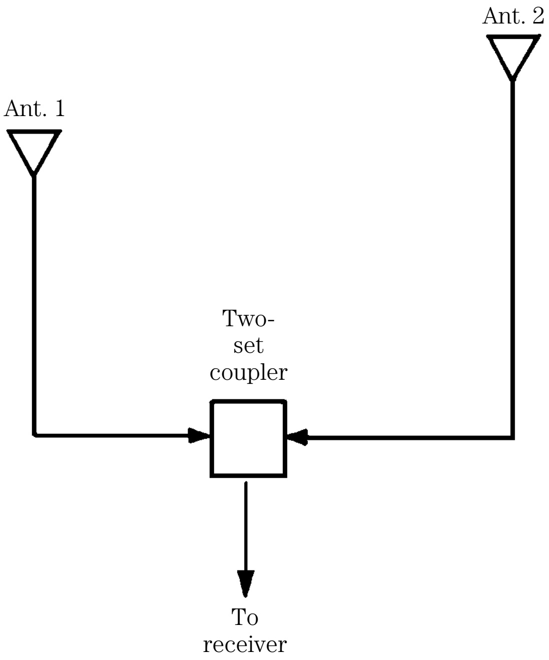

18-14 Combining two antennas with TV hardware.

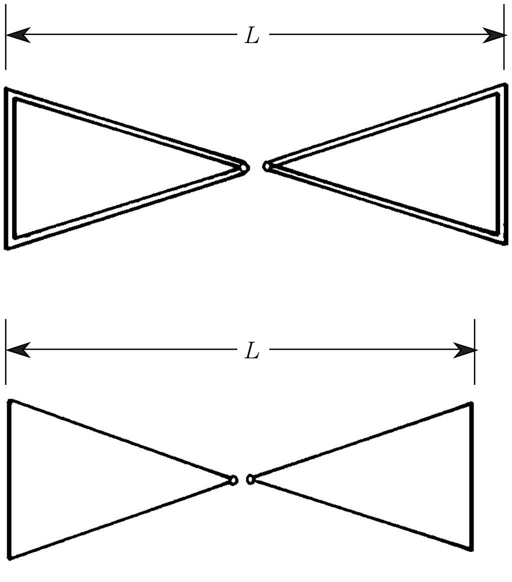

18-15 VHF/UHF "bowtie" dipoles for broadband reception**.**

transmission line using a two-set coupler. Although intended to allow two television sets to receive a signal from the same antenna, the device also works to combine two antennas into a single transmission line. Two popular TV uses include two antennas for different directions and VHF and UHF antennas that share the same downlead. Be sure to buy a weatherproof model if you intend to mount the coupler on the antenna mast (the correct place for it) (Fig. 18-15).

TV antenna amplifiers are used to increase the weak signal from a distant station to a level compatible with the TV receiver. Because these amplifiers are wideband designs, they will also work well for the scanner bands that are inclusive within the 54- to 800-MHz TV bands. Some models are intended to be part of the antenna, while others are mounted indoors, some distance from the antenna. The basic rule of thumb is to place the amplifier as close as possible to the "head end" of the system where the antenna is located. This situation will permit signals to be built up to a level considerably above the noise level before being attenuated in the transmission line losses.

TV/FM receiver folded dipoles made of twin-lead transmission line also work well at other VHF/UHF frequencies. Figures 18-16A and 18-16B show the details of their construction.

Scanner skyhooks

Other antennas not to be overlooked are the CB, or amateur radio, high-frequency (shortwave) antennas. Many amateur radio operators use their 80-m dipole antenna

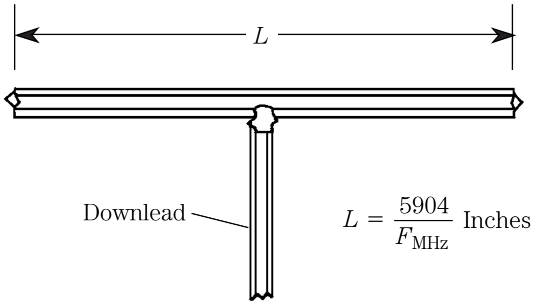

18-16A Receiver folded dipole antenna made of 300 Ω twin lead.

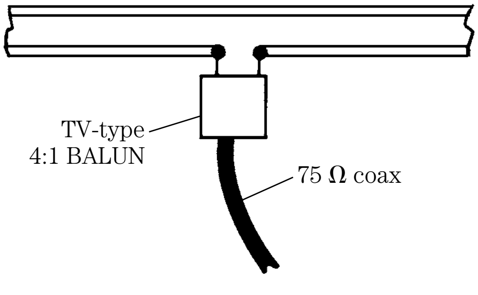

18-16B Poles made of twin-lead transmission VHF frequencies.

on 2 m (144 to 148 MHz) and find that it has the gain and directivity characteristics of a longwire antenna. Similarly, with vertical "ham" and CB antennas, they have apparent gain on some VHF frequencies and will at least perform to some extent on others.

Even the lowly "random length" wire antenna used for your shortwave receiver ought to turn in decent performance as a VHF/UHF longwire antenna. These antennas are simply a 30- to 150-ft length of no. 14 wire attached to a distant support.

Figure 18-17 shows two variants to an antenna that is popular with VHF/UHF receiver operators: the *drooping dipole.* These antennas are similar to the inverted-vee dipole, in that they consist of two quarter-wavelength radiator elements. Unlike the ordinary dipole, the ends of the radiators, in the single-band drooping dipole of Fig. 18-17A, are lowered. The drooping dipole can be made using an SO-239 UHF coaxial connector as the support for either stiff wire, brass hobbyist tubing, or brazing rod elements cut to λ/4 each. The antenna can be made into a multiband, or very wideband, antenna by connecting several pairs of drooping dipoles in parallel, as in Fig. 18-17B.

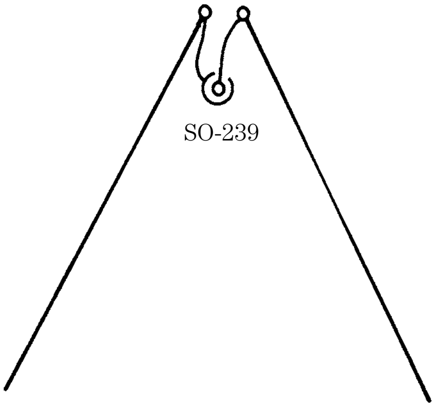

18-17A Drooping dipole.

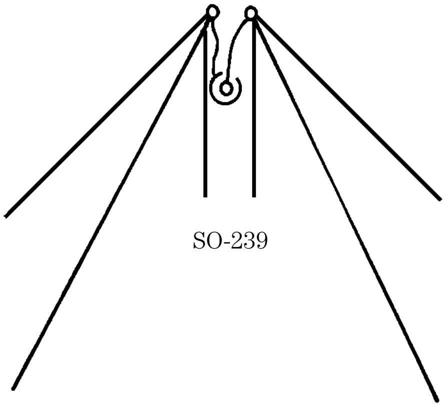

18-17B Multiband drooping dipole.

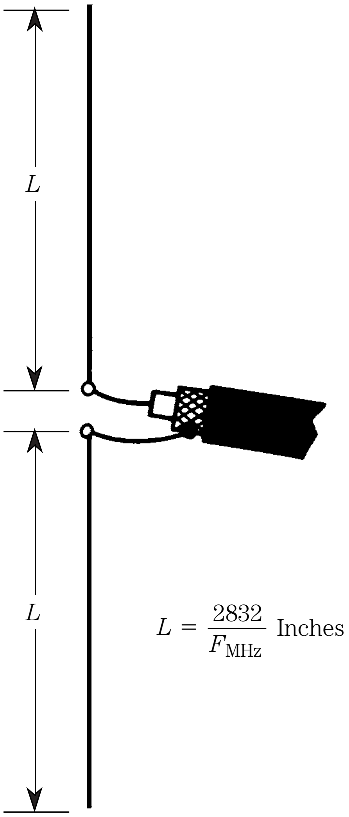

18-18A Vertical dipole.

The vertical dipole (Fig. 18-18A) also finds use in the VHF/UHF region. Each element is λ/4 long, and its physical length can be found from

$$L = \frac{2832}{F_{\text{MHz}}} \text{ in} \qquad \textbf{[18.12]}$$

This antenna can be built with the same types of materials as the drooping dipole. Although several different construction techniques are popular, the method of Fig. 18-18B is probably the most popular with homebrew builders.

A groundplane antenna for VHF/UHF frequencies is shown in Fig. 18-19. The antenna sits on an L bracket mounted to a 2 × 4-in wooden mast or support. The L bracket can be manufactured for the purpose, or purchased at Happy Harry's hardware emporium. The base of the antenna is a chassis-mount SO-239 UHF coaxial connector. A quarter-wave radiator element is made from $\frac{1}{8}$-in brass tubing, or brazing rod, soldered to the center conductor pin of the SO-239. Similarly, radials can be made from the same type of material, but soldered directly to the body of the SO-239 (this action connects it to the coaxial cable shield). The four mounting holes of the SO-239 make reasonable anchors for the radials. Solder them for electrical integrity.

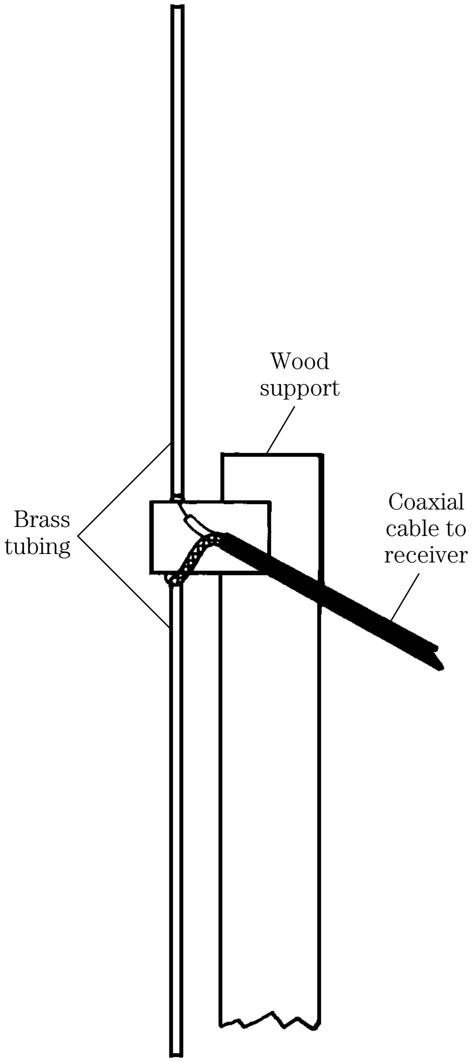

18-18B Construction of VHF vertical dipole.

A coaxial vertical can be made from 1-in brass tubing or copper plumbing pipe, and either brazing rod or brass tubing (Fig. 18-20). The sleeve is fitted with an end cap that has an SO-239 mounted on it (see inset). The radiator element can be made with a mate to the SO-239, namely the PL-259 coaxial plug. The radiator element is soldered to the center conductor, and it is insulated from the shield. You can buy coaxial whips with this type of construction, so you may not have to actually build the radiator element.

Additional gain, about +3 dB, can be achieved by stacking VHF/UHF antennas together. Figure 18-21 shows a typical arrangement in which two half-wavelength dipole antennas are connected together through a quarter-wavelength harness of RG-59/U coaxial cable. This harness is shorter than a quarter wavelength by the velocity factor of the coaxial cable:

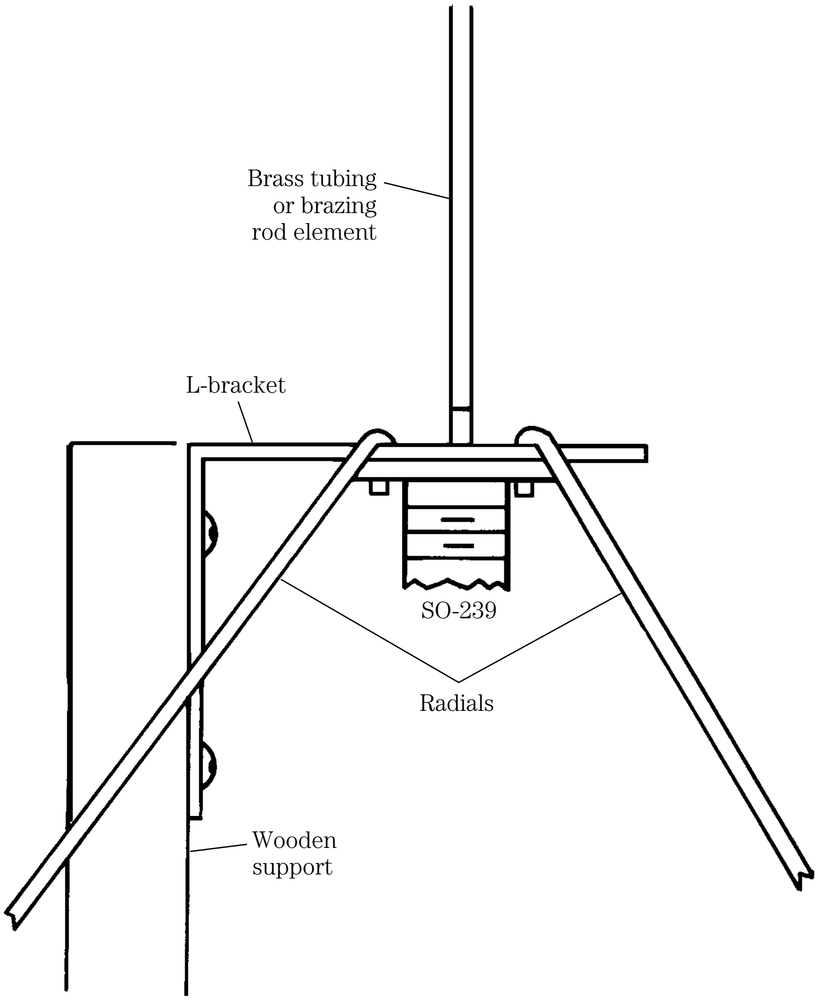
Brass tubing
or brazing
rod element
L-bracket
SO-239
Radials
Wooden
support

18-19 Construction of VHF groundplane for receivers.

$$L = \frac{2832V}{F_{\text{MHz}}} \qquad \textbf{[18.13]}$$

where

L is the length, in inches
V is the velocity factor (typically 0.66 or 0.80 for common coax)
F_{MHz} is the frequency, in megahertz

The antennas can be oriented in the same direction to increase gain, or orthogonally (as shown in Fig. 18-21) to make it more omnidirectional.

Because the impedance of two identical dipoles, fed in parallel, is one-half that of a single dipole, it is necessary to have an impedance-matching section

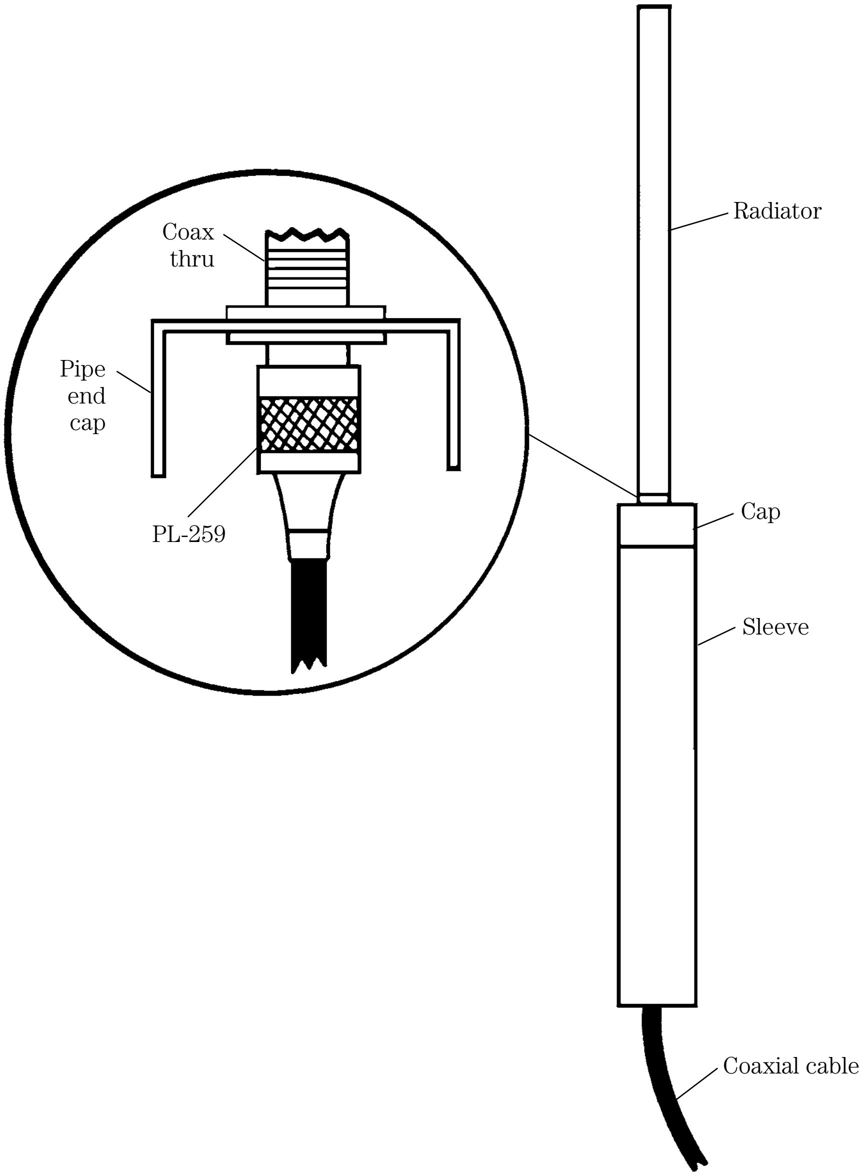
Coax thru
Pipe end cap
PL-259
Radiator
Cap
Sleeve
Coaxial cable

18-20 Coaxial dipole vertical.

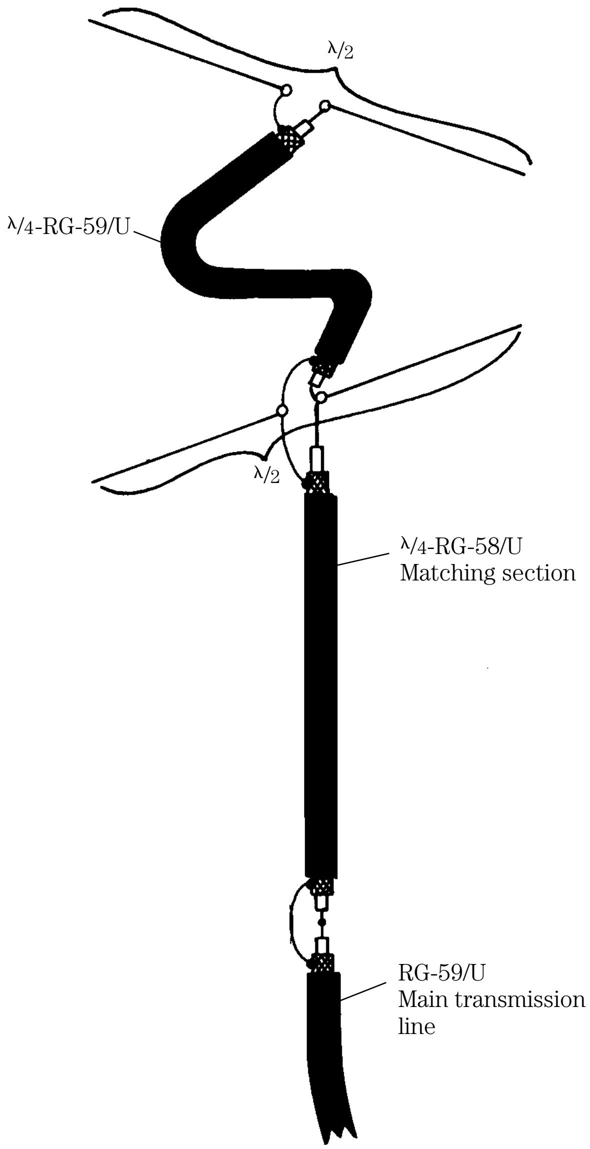
λ/2
λ/4-RG-59/U
λ/2
λ/4-RG-58/U
Matching section
RG-59/U
Main transmission
line

18-21 Stacking VHF antennas.

made of RG-58/U coaxial cable. This cable is then fed with RG-59/U coax from the receiver.

There is nothing magical about scanner receivers that require any form of antenna that is significantly different from other VHF/UHF antennas. Although the designs might be optimized for VHF or UHF, these antennas are basically the same as others shown in this book. As a matter of fact, almost any antennas, from any chapter, can be used by at least some scanner operators.

19
CHAPTER

Microwave waveguides and antennas

THE MICROWAVE PORTION OF THE RADIO SPECTRUM COVERS FREQUENCIES FROM ABOUT 900 MHz to 300 GHz, with wavelengths in free-space ranging from 33 cm down to 1 mm. Transmission lines can be used at frequencies from dc to about 50 or 60 GHz, although, above 5 GHz, only short runs are practical, because attenuation increases dramatically as frequency increases. There are three types of losses in conventional transmission lines: *ohmic, dielectric,* and *radiation.* The ohmic losses are caused by the current flowing in the resistance of the conductors making up the transmission lines. Because of the skin effect, which increases resistance at higher frequencies, these losses tend to increase in the microwave region. Dielectric losses are caused by the electric field acting on the molecules of the insulator and thereby causing heating through molecular agitation. Radiation losses represent loss of energy as an electromagnetic wave propagates away from the surface of the transmission line conductor.

Losses on long runs of coaxial transmission line (the type most commonly used) cause concern even as low as the 400-MHz region. Also, because of the increased losses, power handling capability decreases at high frequencies. Therefore, at higher microwave frequencies, or where long runs make coax attenuation losses unacceptable, or where high power levels would overheat the coax, *waveguides* are used in lieu of transmission lines.

What is a waveguide? Consider the light pipe analogy depicted in Fig. 19-1. A flashlight serves as our "RF source," which (given that light is also an electromagnetic wave) is not altogether unreasonable. In Fig. 19-1A the source radiates into free space and spreads out as a function of distance. The intensity per unit area, at the destination (a wall), falls off as a function of distance (D) according to the *inverse square law* ($1/D^2$).

But now consider the transmission scheme in Fig. 19-1B. The light wave still propagates over distance D, but is now confined to the interior of a mirrored pipe. Almost all of the energy (less small losses) coupled to the input end is delivered to the output end, where the intensity is practically undiminished. Although not

Copyright 2001 - the McGraw-Hill Companies

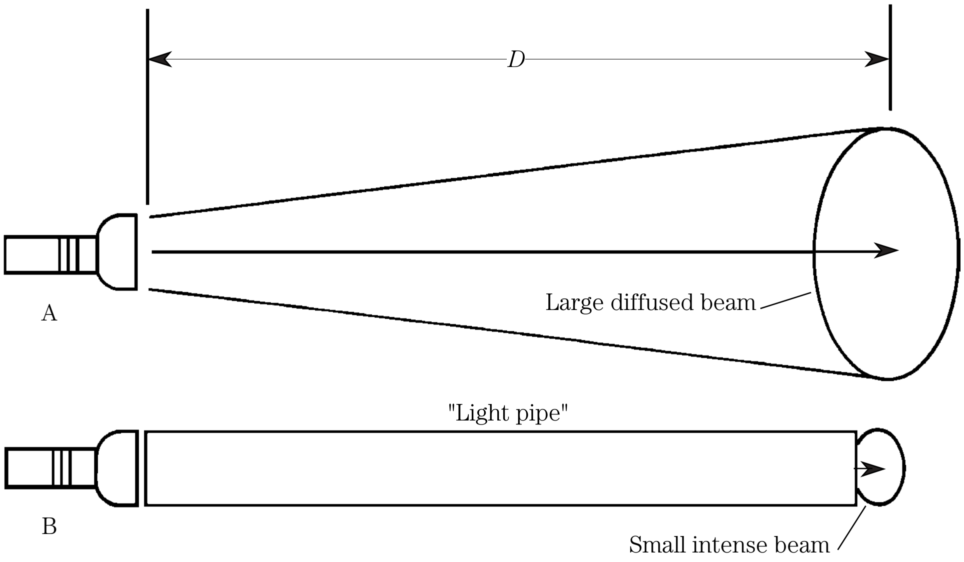

19-1 Waveguide analogy to light pipe.

perfect, the light pipe analogy neatly summarizes, on a simple level, the operation of microwave waveguides.

Thus, we can consider the waveguide as an "RF pipe" without seeming too serenely detached from reality. Similarly, fiber-optic technology is waveguidelike at optical (IR and visible) wavelengths. In fact, the analogy between fiber optics and waveguide can withstand more rigorous comparison than the simplistic light pipe analogy.

The internal walls of the waveguide are not mirrored surfaces, as in our optical analogy, but are, rather, electrical conductors. Most waveguides are made of aluminum, brass, or copper. In order to reduce ohmic losses, some waveguides have their internal surfaces electroplated with either gold or silver, both of which have lower resistivities than the other metals mentioned above.

Waveguides are hollow metal pipes, and can have either circular or rectangular cross sections (although the rectangular are, by far, the most common). Figure 19-2 shows an end view of the rectangular waveguide. The dimension a is the wider dimension, and b is the narrower. These letters are considered the standard form of notation for waveguide dimensions, and will be used in the equations developed in this chapter.

Development of the rectangular waveguide from parallel transmission lines

One way of visualizing how a waveguide works is to develop the theory of waveguides from the theory of elementary parallel transmission lines (see Chap. 3). Figure 19-3A shows the basic parallel transmission line which was introduced in Chap. 3. The line consists of two parallel conductors separated by an air dielectric. Because air won't support the conductors, ceramic or other material insulators are used as supports.

There are several reasons why the parallel transmission line *per se* is not used at microwave frequencies. Skin effect increases ohmic losses to a point that is unacceptable. Also, the insulators supporting the two conductors are significantly more

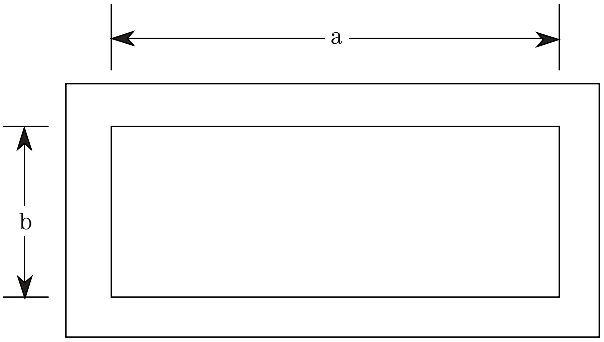

19-2 Rectangular waveguide (end view).

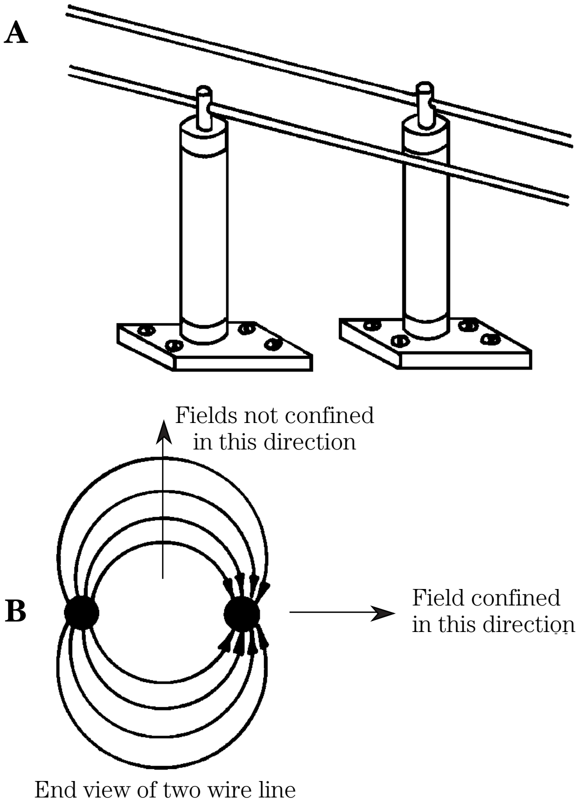
A
Fields not confined
in this direction
B
Field confined
in this direction
End view of two wire line

19-3 Development of waveguide from parallel line.

lossy at microwave frequencies than at lower frequencies. Finally, radiation losses increase dramatically. Figure 19-3B shows the electric fields surrounding the conductors. The fields add algebraically (either constructively or destructively), resulting in pinching of the resultant field along one axis, and bulging along the other. This geometry increases radiation losses at microwave frequencies.

Now let's consider the quarter-wavelength shorted stub. The "looking-in" impedance of such a stub is infinite. When placed in parallel across a transmission line (Fig. 19-4A) the stub acts like an insulator. In other words, at its resonant frequency, the stub is a *metallic insulator,* and can be used to physically support the transmission line.

Again, because the impedance is infinite, we can connect two quarter-wavelength stubs in parallel with each other across the same points on the transmission line (Fig. 19-4B) without loading down the line impedance. This arrangement effectively forms a half-wavelength pair. The impedance is still infinite, so no harm is done. Likewise, we can parallel a large number of center-fed half-wavelength pairs along the line, as might be the case when a long line is supported at multiple points. The waveguide is analogous to an *infinite number of*

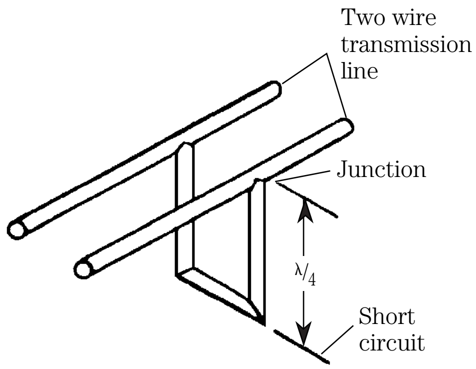

19-4A
Quarter-wave stub analogy.

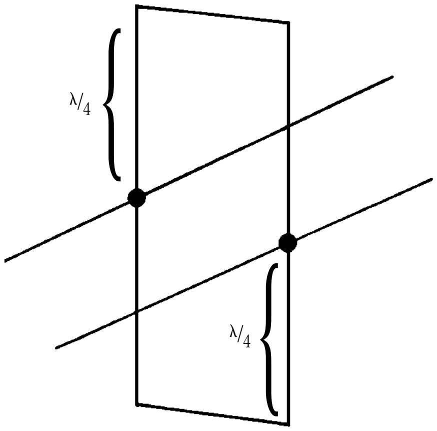

19-4B
Quarter-wave stub analogy extended.

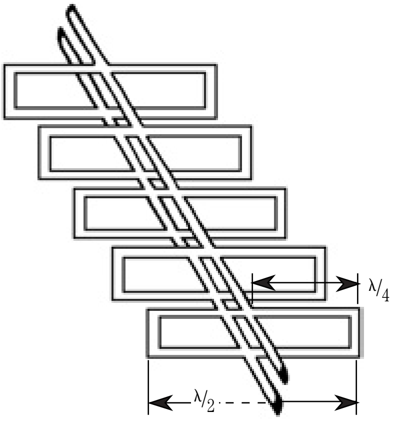

19-4C
Quarter-wave stub analogy extended even further.

center-fed "half-wave pairs" of quarter-wave shorted stubs connected across the line. The result is the continuous metal pipe structure of the common rectangular waveguide (Fig. 19-4C).

On first glance, relating rectangular waveguide to quarter-wavelength shorted stubs seems to fall down, except at the exact resonant frequency. It turns out, however, that the analogy also holds up at other frequencies, so long as the frequency is higher than a certain minimum cutoff frequency. The waveguide thus acts like a high-pass filter. There is also a practical upper frequency limit. In general, waveguides support a bandwidth of 30 to 40 percent of cutoff frequency. As shown in Fig. 19-5, the center line of the waveguide (which represents the points where the conductors are in the parallel line analogy) becomes a "shorting bar" between segments, and that "bar" widens or narrows according to operating frequency. Thus, the active region is still a quarter-wavelength shorted stub.

Below the cutoff frequency, the structure disappears entirely, and the waveguide acts like a parallel transmission line with a low-impedance inductive reactance shorted across the conductors. When modeled as a pair of quarter-wavelength stubs, the a dimension of the waveguide is a half-wavelength long. The cutoff frequency is defined as the frequency at which the a dimension is less than a half-wavelength.

Propagation modes in waveguides

The signal in a microwave waveguide propagates as an electromagnetic wave, not as a current. Even in a transmission line, the signal propagates as a wave because the current in motion down the line gives rise to electric and magnetic fields, which behave as an electromagnetic field. The specific type of field found in transmission lines, however, is a *transverse electromagnetic* (*TEM*) field. The term *transverse* implies things at right angles to each other, so the electric and magnetic fields are perpendicular to the direction of travel. In addition to the word "transverse," these right-angle waves are said to be "normal" or "orthogonal" to the direction of travel—three different ways of saying the same thing: right-angledness.

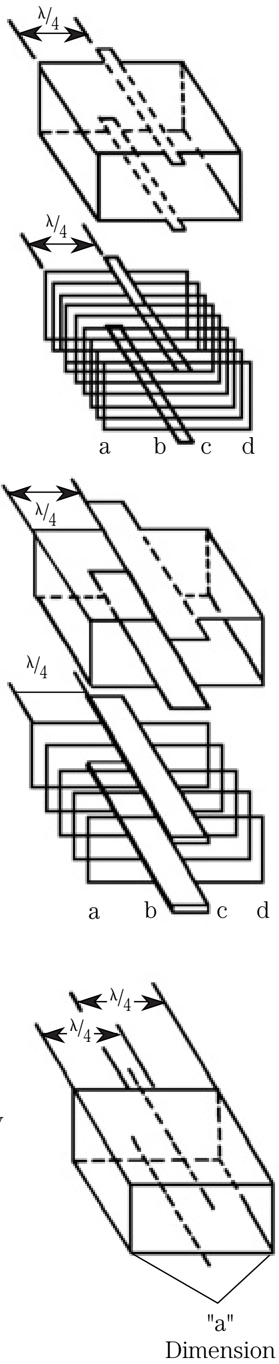
λ/4
λ/4
a
b
c
d
λ/4
λ/4
a
b
c
d
λ/4
λ/4
"a"
Dimension

A. Operating frequency

B. Increasing frequency

C. Decreasing frequency

19-5
Changing frequency does not affect the analogy.

Boundary conditions

The TEM wave will not propagate in a waveguide because certain *boundary conditions* apply. Although the wave in the waveguide propagates through the air (or inert gas dielectric) in a manner similar to free-space propagation, the phenomenon is bounded by the walls of the waveguide, and that implies certain conditions must be met. The boundary conditions for waveguides are

1. The electric field must be orthogonal to the conductor in order to exist at the surface of that conductor.
2. The magnetic field must not be orthogonal to the surface of the waveguide.

In order to satisfy these boundary conditions the waveguide gives rise to two types of propagation modes: *transverse electric mode* (TE mode), and *transverse magnetic mode* (TM mode). The TEM mode violates the boundary conditions because the magnetic field is not parallel to the surface, and so does not occur in waveguides.

The transverse electric field requirement means that the E field must be perpendicular to the conductor wall of the waveguide. This requirement is met by use of a proper coupling scheme at the input end of the waveguide. A vertically polarized coupling radiator will provide the necessary transverse field.

One boundary condition requires that the magnetic (H) field must not be orthogonal to the conductor surface. Because it is at right angles to the E field, it will meet this requirement (see Fig. 19-6). The planes formed by the magnetic field are parallel to both the direction of propagation and the wide dimension surface.

As the wave propagates away from the input radiator, it resolves into two components that are not along the axis of propagation, and are not orthogonal to the walls. The component along the waveguide axis violates the boundary conditions, so it is rapidly attenuated. For the sake of simplicity, only one component is shown in Fig. 19-7. Three cases are shown in Fig. 19-7: high, medium, and low frequency. Note that the angle of incidence with the waveguide wall increases as frequency drops. The angle rises toward 90° as the cutoff frequency is approached from above. Below the cutoff frequency the angle is 90°, so the wave bounces back and forth between the walls without propagating.

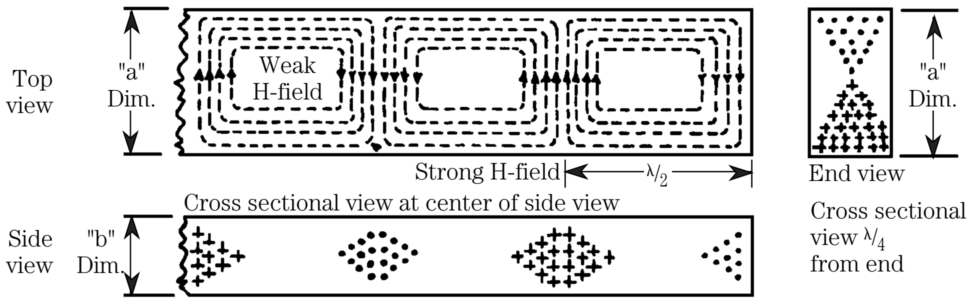

19-6 Magnetic fields in waveguide.

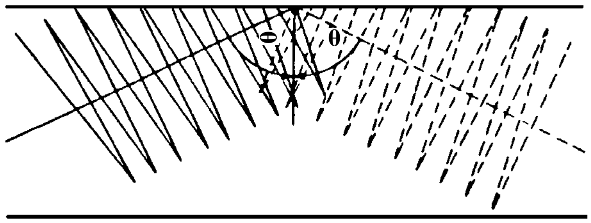

A. High frequency

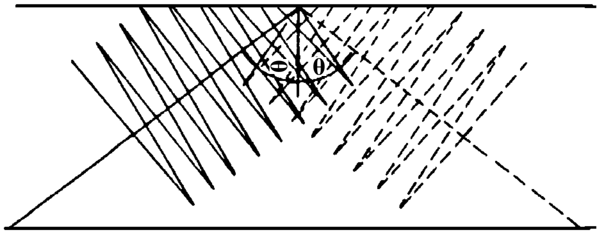

B. Medium frequency

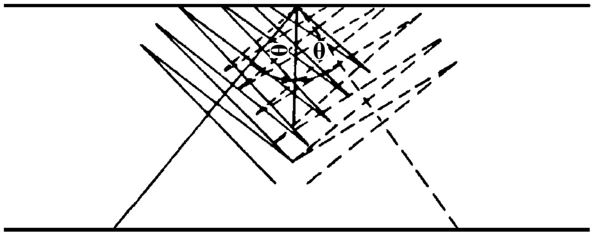

C. Low frequency
(approaching cutoff)

θ = ∠ of incidence
θ = ∠ of reflection

19-7
Frequency effect on propagating wave.

Coordinate system and dominant mode in waveguides

Figure 19-8 shows the coordinate system used to denote dimensions and directions in microwave discussions. The a and b dimensions of the waveguide correspond to the x and y axes of a cartesian coordinate system, and the z axis is the direction of wave propagation.

In describing the various modes of propagation, use a shorthand notation as follows:

$$Tx_{m,n}$$

where

x is E for transverse electric mode, and M for transverse magnetic mode
m is the number of half-wavelengths along the x axis (i.e., a dimension)
n is the number of half-wavelengths along the y axis (i.e., b dimension)

The TE_{10} mode is called the *dominant mode, and is the best mode for low attenuation propagation in the z axis*. The nomenclature TE_{10} indicates that there is one half-wavelength in the *a* dimension and zero half-wavelengths in the *b* dimension. The dominant mode exists at the lowest frequency at which the waveguide is a half-wavelength.

Velocity and wavelength in waveguides

Figures 19-9A and 19-9B show the geometry for two wave components, simplified for the sake of illustration. There are three different wave velocities to consider with respect to waveguides: *free-space velocity c, group velocity V_g,* and *phase velocity V_p.* The free-space velocity is the velocity of propagation in unbounded free space (i.e., the speed of light $c = 3 \times 10^8$ m/s).

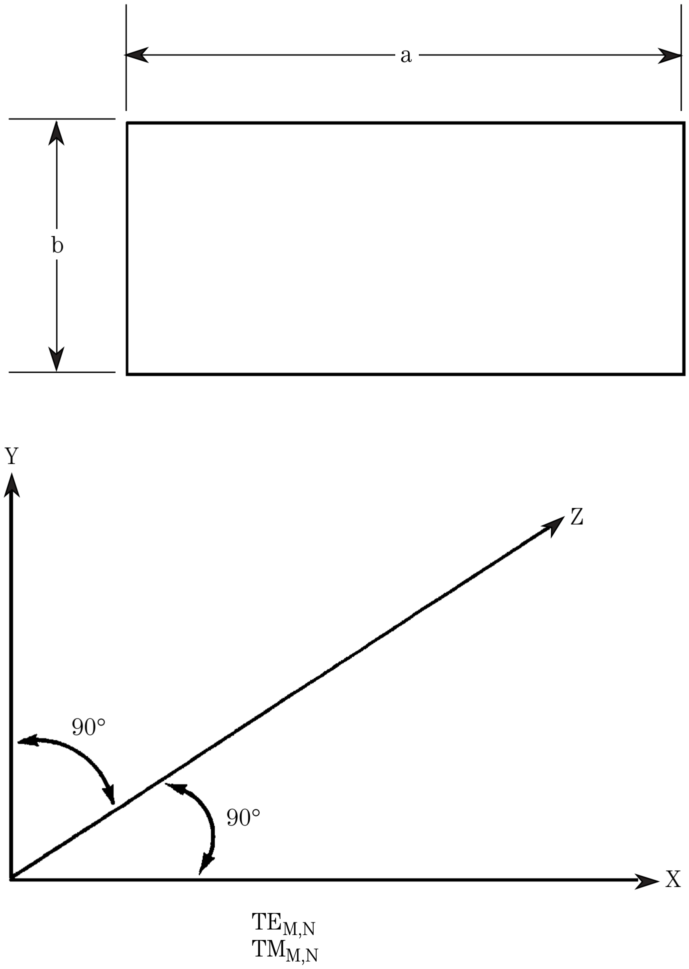

19-8 Rectangular waveguide coordinate system.

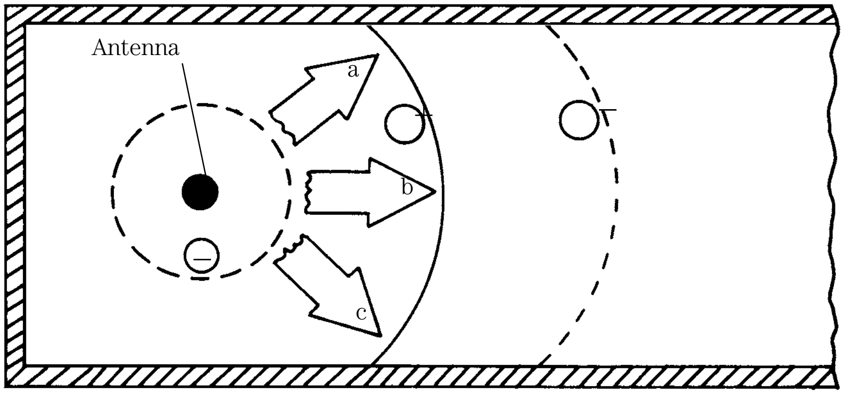

19-9A Antenna radiator in capped waveguide.

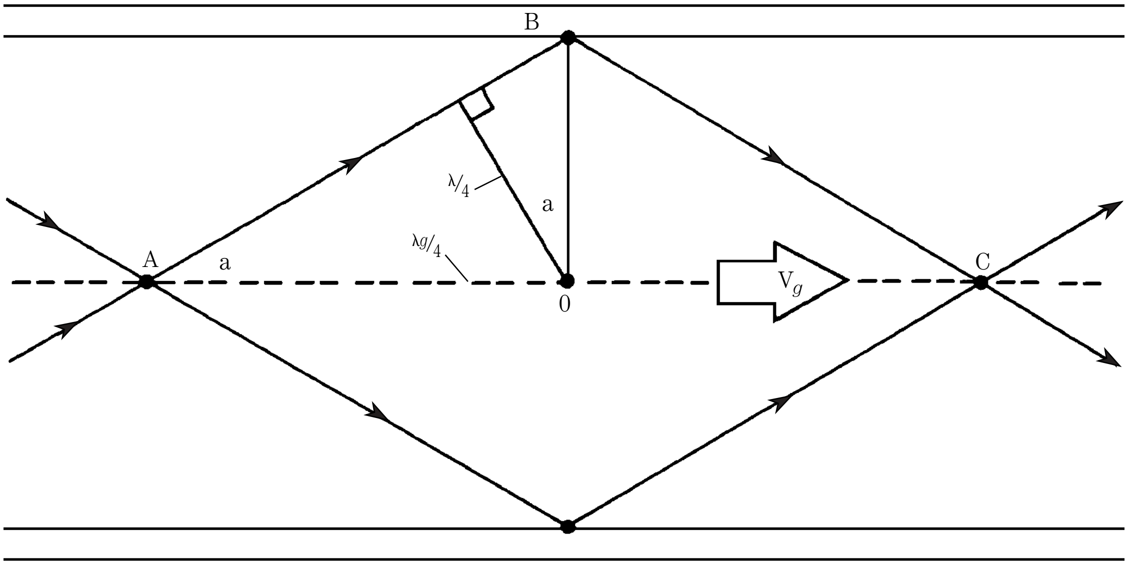

19-9B Wave propagation in waveguide.

The *group velocity* is the straight line velocity of propagation of the wave down the center line *(z* axis) of the waveguides. The value of V_g is always less than c, because the actual path length taken, as the wave bounces back and forth, is longer than the straight line path (i.e., path *ABC* is longer than path *AC*). The relationship between c and V_g is

$$V_g = c \sin a \qquad \textbf{[19.1]}$$

where

V_g is the group velocity, in meters per second
c is the free-space velocity (3×10^8 m/s)
a is the angle of incidence in the waveguide

The *phase velocity* is the velocity of propagation of the spot on the waveguide wall where the wave impinges (e.g., point B in Fig. 19-9B). This velocity, depending upon the angle of incidence, can actually be faster than both the group velocity and the speed of light. The relationship between phase and group velocities can be seen in the beach analogy. Consider an ocean beach, on which the waves arrive from offshore at an angle other than 90°. In other words, the arriving wavefronts are not parallel to the shore. The arriving waves have a group velocity V_g. But as a wave hits the shore, it will strike a point down the beach first, and the "point of strike" races up the beach at a much faster phase velocity V_p, that is even faster than the group velocity. In a microwave waveguide, the phase velocity can be greater than c, as can be seen from Eq. 19.2:

$$V_p = \frac{c}{\sin a} \qquad \textbf{[19.2]}$$

Example 19-1 Calculate the group and phase velocities for an angle of incidence of 33°.

Solution:

(*a*) *Group velocity*

$$\begin{aligned} V_g &= c \sin a \\ &= (3 \times 10^8)(\sin 33°) \\ &= (3 \times 10^8)(0.5446) = 1.6 \times 10^8 \text{ m/s} \end{aligned}$$

(*b*) *Phase velocity*

$$\begin{aligned} V_p &= c/\sin a \\ &= (3 \times 10^8 \text{ m/s})/\sin 33° \\ &= (3 \times 10^8 \text{ m/s})/(0.5446) \\ &= 5.51 \times 10^8 \text{ m/s} \end{aligned}$$

For this problem the solutions are

$$\begin{aligned} c &= 3 \times 10^8 \text{ m/s} \\ V_p &= 5.51 \times 10^8 \text{ m/s} \\ V_g &= 1.6 \times 10^8 \text{ m/s} \end{aligned}$$

We can also write a relationship between all three velocities by combining Eqs. 19.1 and 19.2, resulting in

$$c = \sqrt{V_p V_g} \tag{19.3}$$

In any wave phenomenon the product of *frequency* and *wavelength* is the *velocity.* Thus, for a TEM wave in unbounded free space we know that:

$$c = F\lambda_o \tag{19.4}$$

Because the frequency F is fixed by the generator, only the wavelength can change when the velocity changes. In a microwave waveguide we can relate phase velocity to wavelength as the wave is propagated in the waveguide:

$$V_p = \frac{\lambda \times c}{\lambda_o} \tag{19.5}$$

where:

V_p is the phase velocity, in meters per second
c is the free-space velocity (3×10^8 m/s)
λ is the wavelength in the waveguide, in meters
λ_o is the wavelength in free space *(c/F)*, in meters (see Eq. 19.4)

Equation 19.5 can be rearranged to find the wavelength in the waveguide:

$$\lambda = \frac{V_p \lambda_o}{c} \tag{19.6}$$

Example 19-2 A 5.6-GHz microwave signal is propagated in a waveguide. Assume that the internal angle of incidence to the waveguide surfaces is 42 degrees. Calculate (*a*) phase velocity, (*b*) wavelength in unbounded free space, and (*c*) wavelength of the signal in the waveguide.

Solution:

(*a*) *Phase velocity*

$$\begin{aligned} V_p &= \frac{c}{\sin a} \\ &= \frac{3 \times 10^8 \text{ m/s}}{\sin 42°} \\ &= \frac{3 \times 10^8 \text{ m/s}}{0.6991} = 4.5 \times 10^8 \text{ m/s} \end{aligned}$$

(*b*) *Wavelength in free space*

$$\begin{aligned} \lambda_o &= c/F \\ &= (3 \times 10^8 \text{ m/s})/(5.6 \times 10^9 \text{ Hz}) \\ &= 0.054 \text{ m} \end{aligned}$$

(*c*) *Wavelength in waveguide*

$$\lambda = \frac{V_p \lambda_o}{c}$$

$$= \frac{(4.5 \times 10^8 \text{ m/s})(0.054 \text{ m})}{3 \times 10^8 \text{ m/s}} = 0.08 \text{ m}$$

Comparing, we find that the free-space wavelength is 0.054 m, and the wavelength inside of the waveguide increases to 0.08 m.

Cutoff frequency (F_c)

The propagation of signals in a waveguide depends, in part, upon the operating frequency of the applied signal. As covered earlier, the angle of incidence made by the plane wave to the waveguide wall is a function of frequency. As the frequency drops, the angle of incidence increases toward 90°.

The propagation of waves depends on the angle of incidence and the associated reflection phenomena. Indeed, both phase and group velocities are functions of the angle of incidence. When the frequency drops to a point where the angle of incidence is 90°, then group velocity is meaningless.

We can define a general mode equation based on our system of notation:

$$\frac{1}{(\lambda_c)^2} = \left(\frac{m}{2a}\right)^2 + \left(\frac{n}{2b}\right)^2 \qquad \textbf{[19.7]}$$

where

λ_c is the longest wavelength that will propagate
a, b are the waveguide dimensions (see Fig. 19-2)
m, n are integers that define the number of half-wavelengths that will fit in the a and b dimensions, respectively

Evaluating Eq. 19.7 reveals that the longest TE-mode signal that will propagate in the dominant mode (TE_{10}) is given by

$$\lambda_c = 2a \qquad \textbf{[19.8]}$$

from which we can write an expression for the cutoff frequency:

$$F_c = \frac{c}{2a} \qquad \textbf{[19.9]}$$

where

F_c is the lowest frequency that will propagate, in hertz
c is the speed of light (3×10^8 m/s)
a is the wide waveguide dimension

Example 19-3 A rectangular waveguide has dimensions of 3 × 5 cm. Calculate the TE_{10} mode cutoff frequency.

Solution:

$$F_c = \frac{c}{2a}$$

$$= \frac{(3 \times 108\ \text{m/s}}{(2)\left(5\ \text{cm} \times \dfrac{1\ \text{m}}{100\ \text{cm}}\right)}$$

$$= \frac{3 \times 10^8\ \text{m/s}}{(2)(0.05\ \text{m})} = 3\ \text{GHz}$$

Equation 19.7 assumes that the dielectric inside the waveguide is air. A more generalized form, which can accommodate other dielectrics, is

$$F_c = \frac{1}{2\sqrt{ue}} \sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2} \qquad \textbf{[19.10]}$$

where

e is the dielectric constant
u is the permeability constant

For air dielectrics, $u = u_o$ and $e = e_o$, from which

$$c = \frac{1}{\sqrt{u_o e_o}} \qquad \textbf{[19.11]}$$

To determine the cutoff wavelength, we can rearrange Eq. 19.10 to the form:

$$\lambda_c = \frac{2}{\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}} \qquad \textbf{[19.12]}$$

One further expression for air-filled waveguide calculates the actual wavelength in the waveguide from a knowledge of the free-space wavelength and actual operating frequency:

$$\lambda_g = \frac{\lambda_o}{\sqrt{1 - \left(\frac{F_c}{F}\right)^2}} \qquad \textbf{[19.13]}$$

where

λ_g is the wavelength in the waveguide
λ_o is the wavelength in free space
F_c is the waveguide cutoff frequency
F is the operating frequency

Example 19-4 A waveguide with a 4.5-GHz cutoff frequency is excited with a 6.7-GHz signal. Find (*a*) the wavelength in free space and (*b*) the wavelength in the waveguide.

Solution:

(*a*)

$$\begin{aligned}\lambda_o &= c/F \\ &= \frac{3 \times 10^8 \text{ m/s}}{6.7 \text{ GHz} \times \dfrac{10^9 \text{ Hz}}{1 \text{ GHz}}} \\ &= \frac{3 \times 10^8 \text{ m/s}}{6.7 \times 10^9 \text{ Hz}} = 0.0448 \text{ m}\end{aligned}$$

(*b*)

$$\begin{aligned}\lambda_g &= \frac{\lambda_o}{\sqrt{1 - \left(\dfrac{F_c}{F}\right)^2}} \\ &= \frac{0.0448 \text{ m}}{\sqrt{1 - \left(\dfrac{4.5 \text{ GHz}}{6.7 \text{ GHz}}\right)^2}} \\ &= \frac{0.0448 \text{ m}}{1 - 0.67} \\ &= \frac{0.0448}{0.33} = 0.136 \text{ m}\end{aligned}$$

Transverse magnetic modes also propagate in waveguides, but the base TM_{10} mode is excluded by the boundary conditions. Thus, the TM_{11} mode is the lowest magnetic mode that will propagate.

Waveguide impedance

All forms of transmission line, including the waveguide, exhibit a characteristic impedance, although in the case of waveguide it is a little difficult to pin down conceptually. This concept was developed for ordinary transmission lines in Chap. 3. For a waveguide, the characteristic impedance is approximately equal to the ratio of the electric and magnetic fields (*E/H*), and converges (as a function of frequency) to the intrinsic impedance of the dielectric (Fig. 19-10). The impedance of the waveguide is a function of waveguide characteristic impedance (Z_o) and the wavelength in the waveguide:

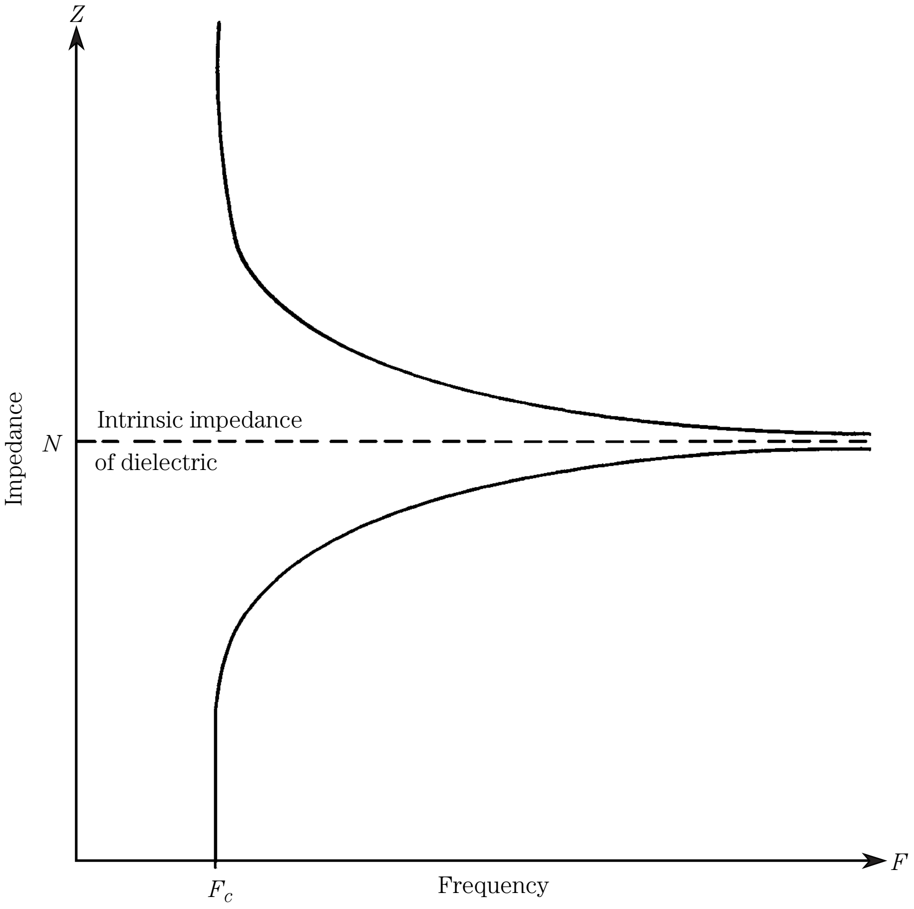

19-10 Impedance versus frequency.

$$Z = \frac{Z_o \lambda_g}{\lambda_o} \qquad \textbf{[19.14]}$$

Or, for rectangular waveguide, with constants taken into consideration:

$$Z = \frac{120\,\pi\,\lambda_g}{\lambda_o} \qquad \textbf{[19.15]}$$

The *propagation constant B* for rectangular waveguide is a function of both cutoff frequency and operating frequency:

$$B = W\sqrt{eu}\,\sqrt{1-\left(\frac{F_c}{F}\right)^2} \qquad \textbf{[19.16]}$$

from which we can express the TE-mode impedance:

$$Z_{\mathrm{TE}} = \frac{\sqrt{ue}}{\sqrt{1-\left(\frac{F_c}{F}\right)^2}} \qquad \textbf{[19.17]}$$

and the TM-mode impedance:

$$Z_{\mathrm{TM}} = 377\sqrt{1-\left(\frac{F_c}{F}\right)^2} \qquad \textbf{[19.18]}$$

Waveguide terminations

When an electromagnetic wave propagates down a waveguide, it must eventually reach the end of the guide. If the end is open, then the wave will propagate into free-space. The horn radiator is an example of an unterminated waveguide. If the waveguide terminates in a metallic wall, then the wave reflects back down the waveguide, from whence it came. The interference between incident and reflected waves forms standing waves (see Chap. 3). Such waves are stationary in space, but vary in the time domain.

In order to prevent standing waves, or more properly, the reflections that give rise to standing waves, the waveguide must be *terminated* in a matching impedance. When a properly designed antenna is used to terminate the waveguide, it forms the matched load required to prevent reflections. Otherwise, a *dummy load* must be provided. Figure 19-11 shows several types of dummy load.

The classic termination is shown in Fig. 19-11A. The “resistor” making up the dummy load is a mixture of sand and graphite. When the fields of the propagated wave enter the load, they cause currents to flow, which in turn cause heating. Thus, the RF power dissipates in the sand-graphite rather than being reflected back down the waveguide.

A second dummy load is shown in Fig. 19-11B. The resistor element is a carbonized rod critically placed at the center of the electric field. The E field causes currents to flow, resulting in I^2R losses that dissipate the power.

Bulk loads, similar to the graphite-sand chamber, are shown in Fig. 19-11C, D, and E. Using bulk material such as graphite or a carbonized synthetic material, these loads are used in much the same way as the sand load (i.e., currents set up, and I^2R losses dissipate the power).

The resistive vane load is shown in Fig. 19-11F. The plane of the element is orthogonal to the magnetic lines of force. When the magnetic lines cut across the vane, currents are induced, which gives rise to the I^2R losses. Very little RF energy reaches the metallic end of the waveguide, so there is little reflected energy and a low VSWR.

There are situations where it isn’t desirable to terminate the waveguide in a dummy load. Several reflective terminations are shown in Fig. 19-12. Perhaps the simplest form is the permanent end plate shown in Fig. 19-12A. The metal cover must be welded or otherwise affixed through a very-low-resistance joint. At the substantial power levels typically handled in transmitter waveguides, even small resistances can be important.

The end plate (shown in Fig. 19-12B) uses a quarter-wavelength cup to reduce the effect of joint resistances. The cup places the contact joint at a point that is a

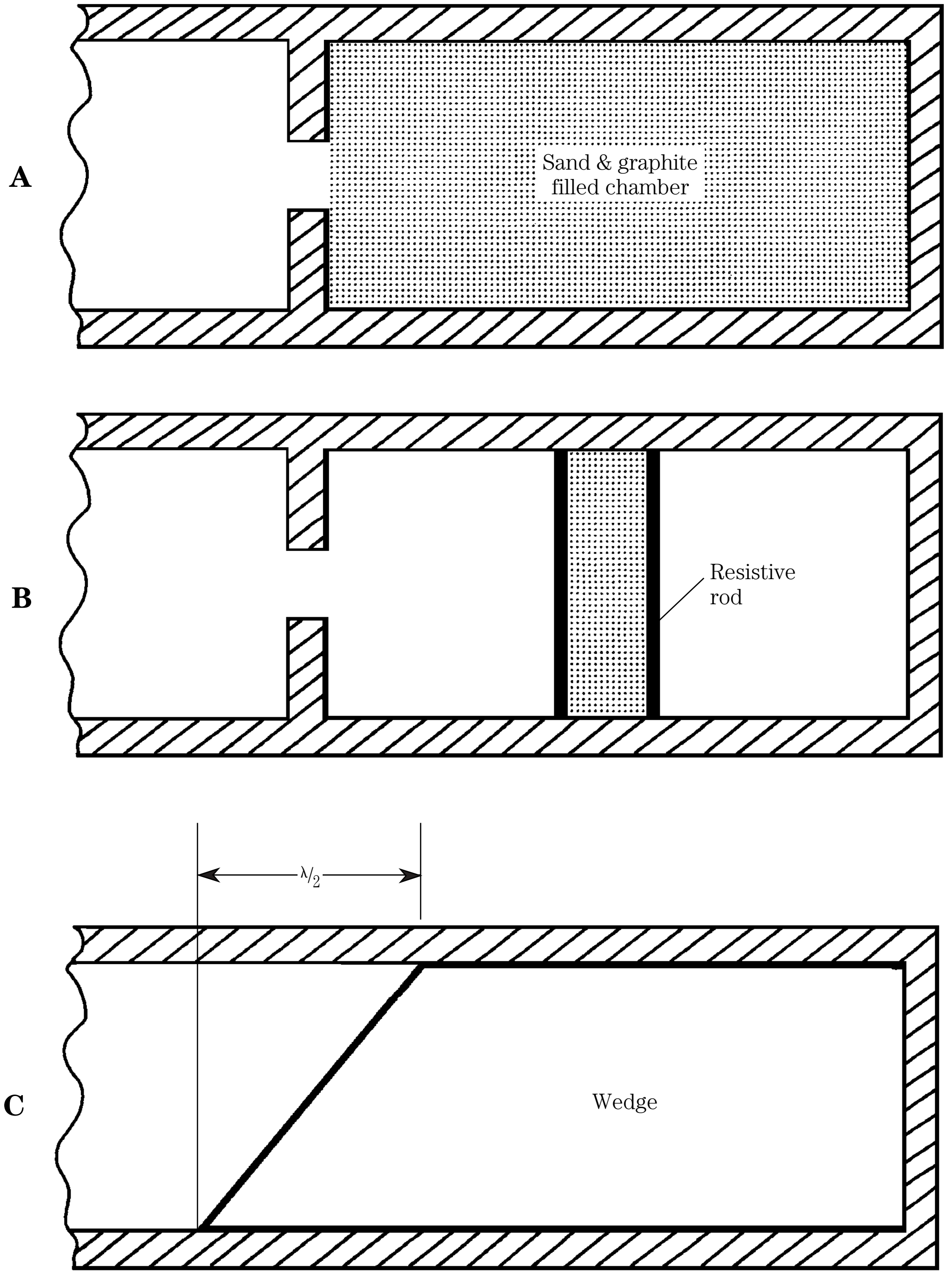
A
Sand & graphite filled chamber
B
Resistive rod
C
$\lambda/2$
Wedge

19-11 Dummy loads: (A) sand-graphite chamber; (B) resistive rod; (C) wedge.

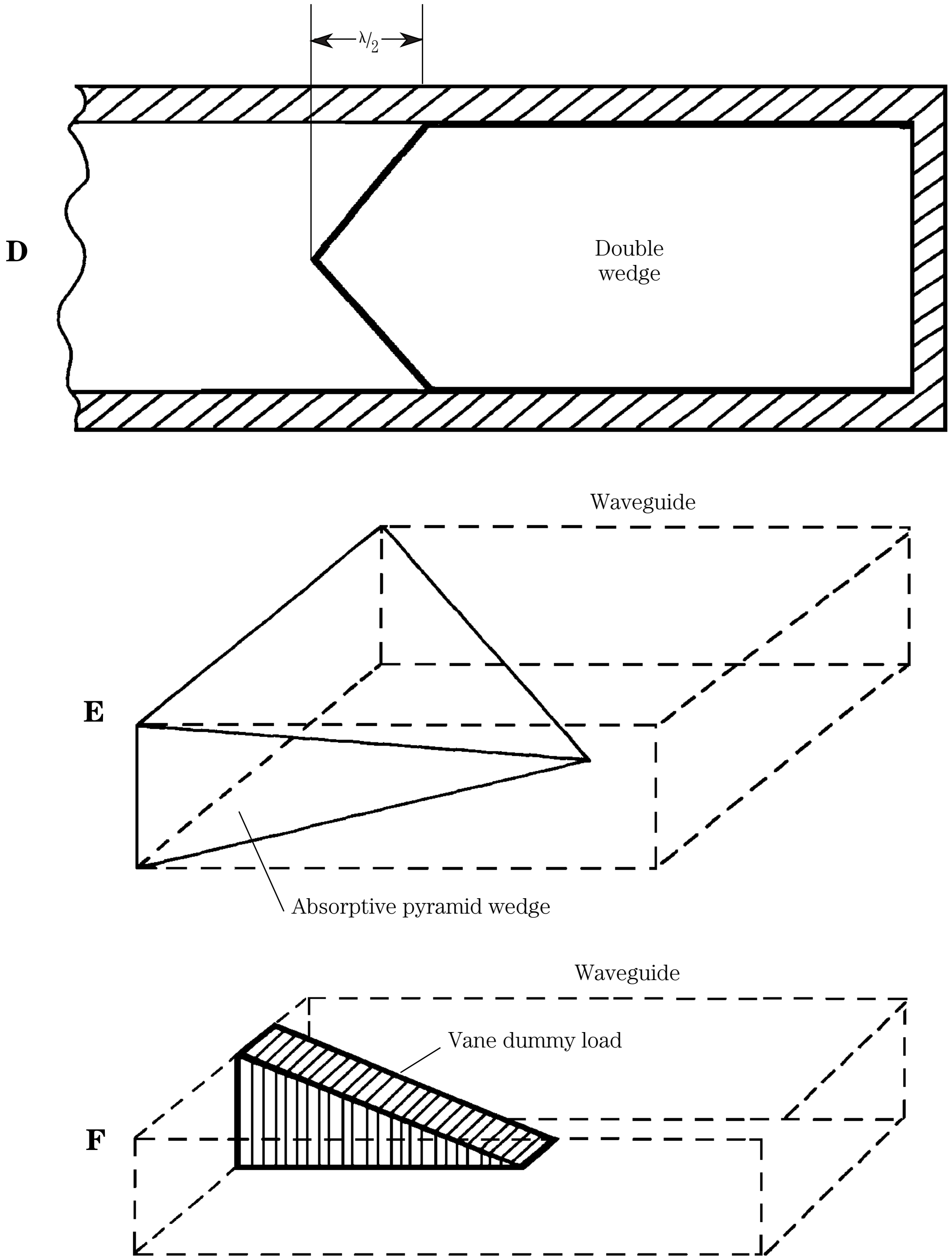
λ/2
D
Double wedge
Waveguide
E
Absorptive pyramid wedge
Waveguide
Vane dummy load
F

19-11 Dummy loads: (D) double wedge; (E) pyramid wedge; (F) vane dummy load.

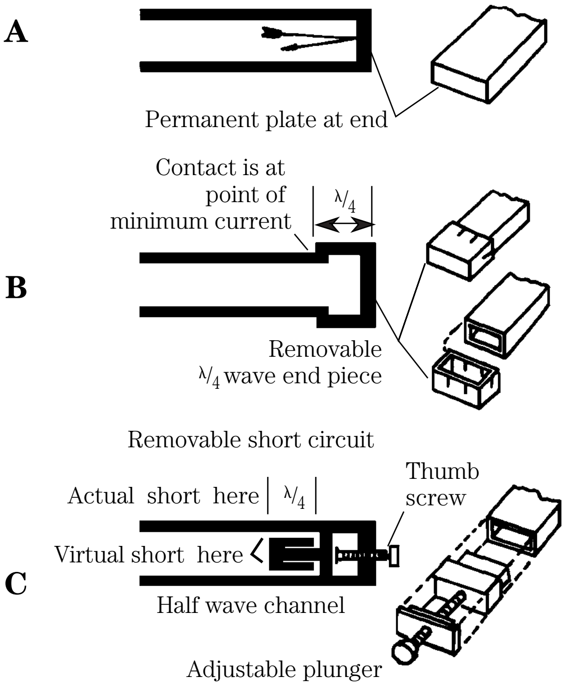
A
Permanent plate at end
Contact is at point of minimum current
λ/4
B
Removable λ/4 wave end piece
Removable short circuit
Actual short here
λ/4
Thumb screw
Virtual short here
C
Half wave channel
Adjustable plunger

19-12
End terminations.

quarter-wavelength from the end. This point is a minimum-current node, so I^2R losses in the contact resistance become less important.

The adjustable short circuit is shown in Fig. 19-12C. The walls of the waveguide and the surface of the plunger form a half-wavelength channel. Because the metallic end of the channel is a short circuit, the impedance reflected back to the front of the plunger is zero ohms, or nearly so. Thus, a virtual short exists at the points shown. By this means, the contact (or joint) resistance problem is overcome.

Waveguide joints and bends

Joints and bends in any form of transmission line or waveguide are seen as impedance discontinuities, and so are points at which disruptions occur. Thus, improperly formed bends and joints are substantial contributors to a poor VSWR. In general, bends, twists, joints, or abrupt changes in waveguide dimension can deteriorate the VSWR by giving rise to reflections.

Extensive runs of waveguide are sometimes difficult to make in a straight line. Although some installations do permit a straight waveguide, many others require directional change. This possibility is especially likely on shipboard installations. Figure 19-13A shows the proper ways to bend a waveguide around a corner. In each case, the radius of the bend must be at least two wavelengths at the lowest frequency that will be propagated in the system.

The “twist” shown in Fig. 19-13B is used to rotate the polarity of the E and H fields by 90°. This type of section is sometimes used in antenna arrays for phasing the elements. As in the case of the bend, the twist must be made over a distance of at least two wavelengths.

When an abrupt 90° transition is needed, it is better to use two successive 45° bends spaced one-quarter wavelength apart (see Fig. 19-13C). The theory (behind this kind of bend) is to cause the interference of the direct reflection of one bend, with the inverted reflection of the other. The resultant relationship between the fields is reconstructed as if no reflections had taken place.

Joints are necessary in practical waveguides because it simply isn't possible to construct a single length of practical size for all situations. Three types of common joints are used: *permanent, semipermanent,* and *rotating.*

To make a permanent joint the two waveguide ends must be machined extremely flat so that they can be butt-fitted together. A welded or brazed seam bonds the two sections together. Because such a surface represents a tremendous discontinuity, reflections and VSWR will result unless the interior surfaces are milled flat and then polished to a mirror-like finish.

A *semipermanent joint* allows the joint to be disassembled for repair and maintenance, as well as allowing easier on-site assembly. The most common example of this class is the *choke joint* shown in Fig. 19-14.

One surface of the choke joint is machined flat and is a simple butt-end planar flange. The other surface is the mate to the planar flange, but it has a quarter-wavelength circular slot cut at a distance of one-quarter wavelength from the waveguide aperture. The two flanges are shown in side view in Fig. 19-14A, and the slotted end view is shown in Fig. 19-14B. The method for fitting the two ends together is shown in the oblique view in Fig. 19-14C.

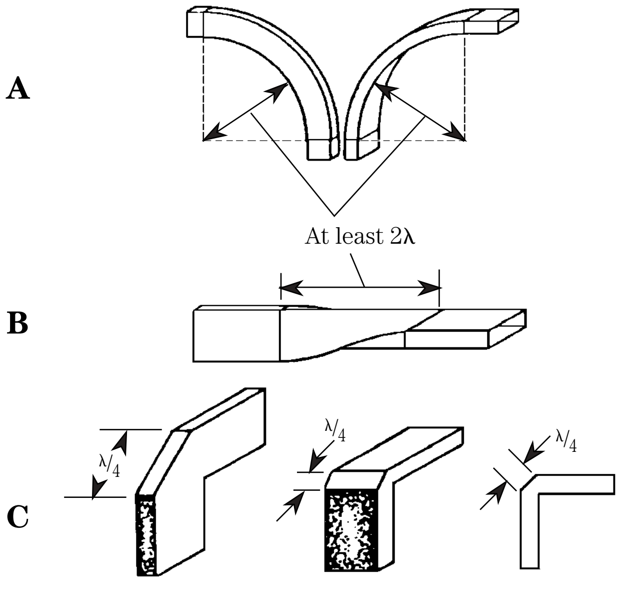

Narrow dimension

Wide dimension

Side view

19-13 Bends in waveguide must be gentle.

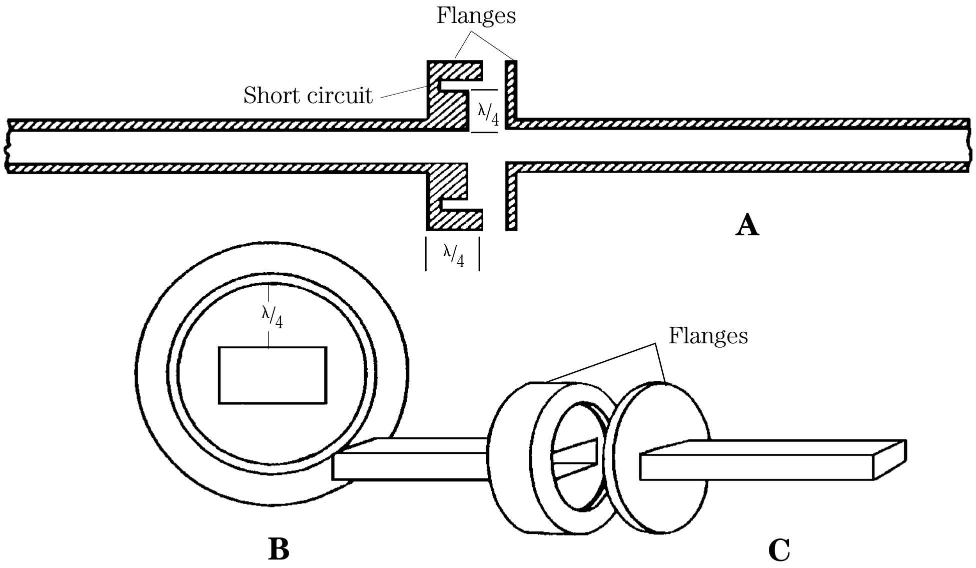

19-14 Choke joint.

Rotating joints are used in cases where the antenna has to point in different directions at different times. Perhaps the most common example of such an application is the radar antenna.

The simplest form of rotating joint is shown in Fig. 19-15. The key to its operation is that the selected mode is symmetrical about the rotating axis. For this reason, a circular waveguide operating in the TM_{01} mode is selected. In this rotating choke joint, the actual waveguide rotates but the internal fields do not (thereby minimizing reflections). Because most waveguide is rectangular, however, a somewhat more complex system is needed. Figure 19-16 shows a rotating joint consisting of two circular waveguide sections inserted between segments of rectangular waveguide. On each end of the joint, there is a rectangular-to-circular transition section.

In Fig. 19-16, the rectangular input waveguide operates in the TE_{10} mode that is most efficient for rectangular waveguide. The E-field lines of force couple with the circular segment, thereby setting up a TM_{01} mode wave. The TM_{01} mode has the required symmetry to permit coupling across the junction, where it meets another transition zone and is reconverted to TE_{10} mode.

Waveguide coupling methods

Except possibly for the case where an oscillator exists inside a waveguide, it is necessary to have some form of input or output coupling in a waveguide system. There are three basic types of coupling used in a microwave waveguide: *capacitive* (or *probe*), *inductive* (or *loop*), and *aperture* (or *slot*).

Capacitive coupling is shown in Fig. 19-17. This type of coupling uses a vertical radiator inserted into one end of the waveguide. Typically, the probe is a quarter-wavelength in a fixed-frequency system. The probe is analogous to the vertical antennas used at lower frequencies. A characteristic of this type of radiator is that the E field is parallel to the waveguide top and bottom surfaces. This arrangement satisfies the first boundary condition for the dominant TE_{10} mode.

The radiator is placed at a point that is a quarter-wavelength from the rear wall (Fig. 19-17B). By traversing the quarter-wave distance (90° phase shift), being reflected from the rear wall (180° phase shift), and then retraversing the

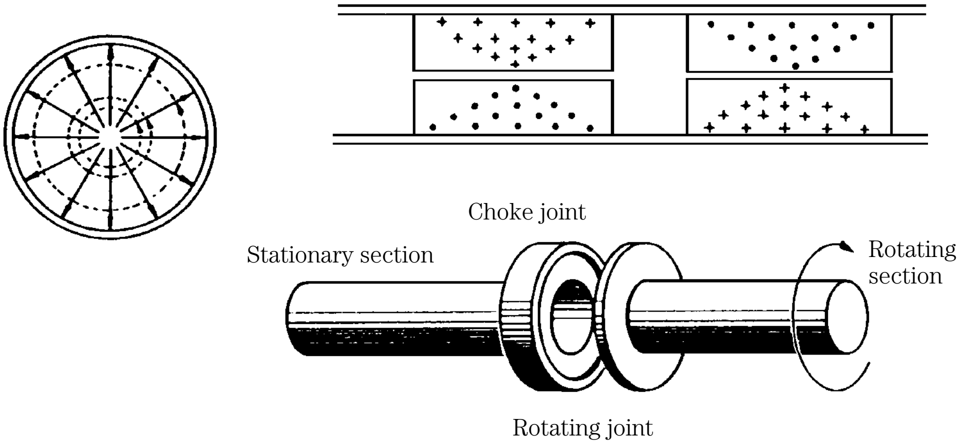
Choke joint
Stationary section
Rotating section
Rotating joint

19-15 Basic rotating joint.

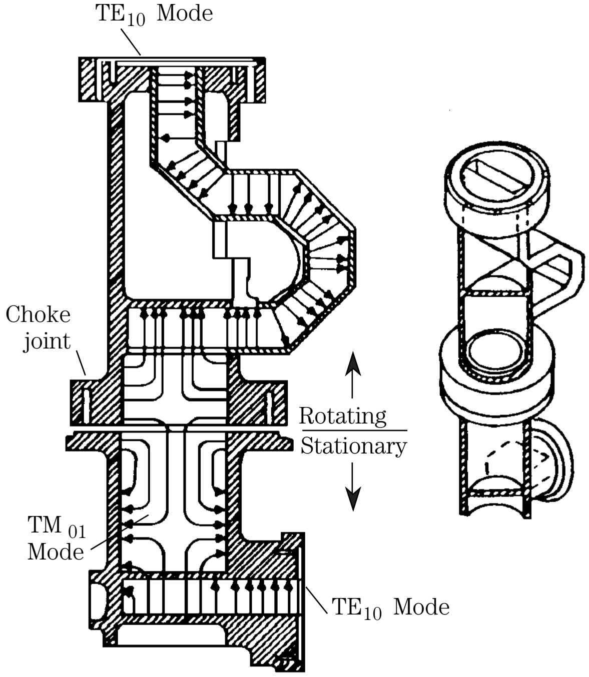
TE10 Mode
Choke joint
Rotating
Stationary
TM 01 Mode
TE10 Mode

19-16
Representative practical rotating joint.

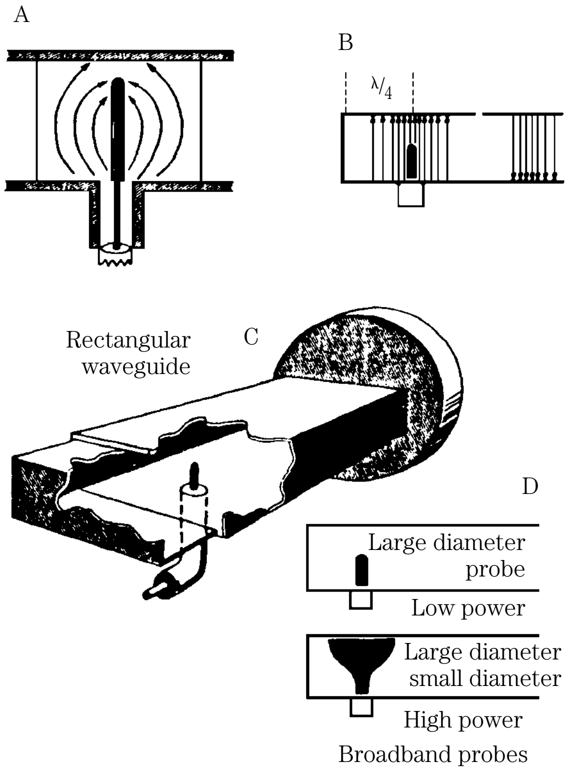

19-17
Probe (capacitive) coupling.

quarter-wavelength distance (another 90° phase shift), the wave undergoes a total phase shift of one complete cycle, or 360°. Thus, the reflected wave arrives back at the radiator in phase to reinforce the outgoing wave. Hence, none of the excitation energy is lost.

Some waveguides have an adjustable end cap (Fig. 19-17C) in order to accommodate multiple frequencies. The end cap position is varied to accommodate the different wavelength signals.

Figure 19-17D shows high- and low-power broadband probes that are typically not a quarter-wavelength except at one particular frequency. Broadbanding is accomplished by attention to the diameter-to-length ratio. The *degree of coupling* can be varied in any of several ways: the *length* of the probe can be varied; the *position* of the probe in the E field can be changed; or *shielding* can be used to partially shade the radiator element.

Inductive, or loop coupling, is shown in Fig. 19-18. A small loop of wire (or other conductor) is placed such that the number of magnetic flux lines is maximized. This form of coupling is popular on microwave receiver antennas, in order to make a waveguide-to-coaxial cable transition. In some cases, the loop is formed by the pigtail lead of a detector diode that, when combined with a local oscillator, downconverts the microwave signal to an IF frequency in the 30- to 300-MHz region.

Aperture, or slot coupling, is shown in Fig. 19-19. This type of coupling is used to couple together two sections of waveguide, as on an antenna feed system. Slots can be designed to couple either electric, magnetic, or electromagnetic fields. In Fig. 19-19, slot *A* is placed at a point where the E field peaks, so it allows electrical field coupling. Similarly, slot *B* is at a point where the H field peaks, so it allows magnetic field coupling. Finally, we see slot *C,* which allows electromagnetic field coupling.

Slots can also be characterized according to whether they are *radiating* or *nonradiating.* A nonradiating slot is cut at a point that does not interrupt the flow of currents in the waveguide walls. The radiating slot, on the other hand, does interrupt currents flowing in the walls. A radiating slot is the basis for several forms of antenna, which are discussed at the end of this chapter.

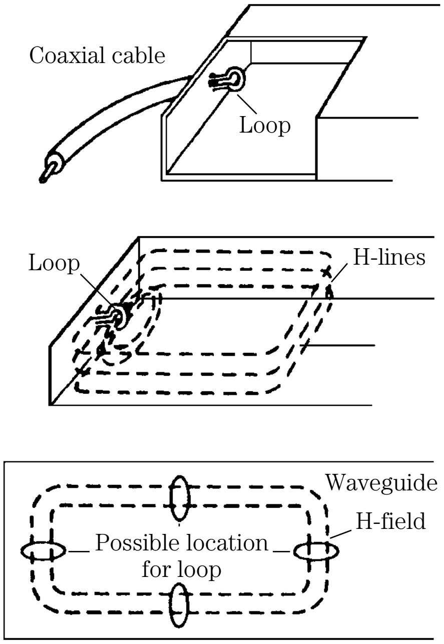

19-18
Loop (inductive) coupling.

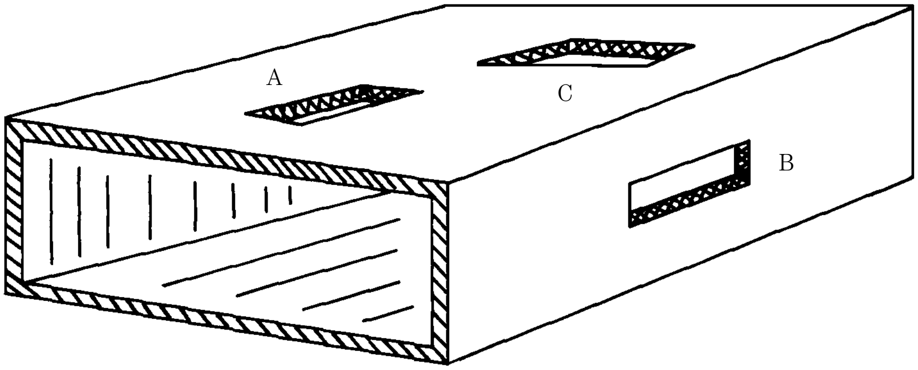

19-19 Slot coupling.

Microwave antennas

Antennas are used in communications and radar systems at frequencies from the very lowest to the very highest. In both theory and practice, antennas are used until frequencies reach infrared and visible light, at which point optics becomes more important. Microwaves are a transition region between ordinary "radio waves" and "optical waves," so (as might be expected) microwave technology makes use of techniques from both worlds. For example, both dipoles and parabolic reflectors are used in microwave systems.

The purpose of an antenna is to act as a *transducer* between either electrical oscillations or propagated "guided waves" (i.e., in transmission lines or waveguides) and a propagating electromagnetic wave in free space. A principal function of the antenna is to act as an *impedance matcher* between the waveguide, or transmission line, impedance and the impedance of free space.

Antennas can be used equally well for both receiving and transmitting signals because they obey the *law of reciprocity.* That is, the same antenna can be used to both receive and transmit with equal success. Although there might be practical or mechanical reasons to prefer specific antennas for one or the other mode, electrically they are the same.

In the transmit mode, the antenna must radiate electromagnetic energy. For this job, the important property is *gain* G. In the receive mode, the job of the antenna is to gather energy from impinging electromagnetic waves in free space. The important property for receiver antennas is the *effective aperture* A_e, which is a function of the antenna's physical area. Because of reciprocity, a large gain usually infers a large effective aperture and vice versa. *Effective aperture* is defined as the area of the impinging radio wavefront that contains the same power as is delivered to a matched resistive load across the feedpoint terminals.

The isotropic "antenna"

Antenna definitions and specifications can become useless unless a means is provided for putting everything on a common footing. Although a variety of systems exist for describing antenna behavior, the most common system compares a specific antenna with a theoretical construct, called the *isotropic radiator.*

An isotropic radiator is a spherical point source that radiates equally well in all directions. By definition, the directivity of the isotropic antenna is unity (1), and all antenna gains are measured against this standard. Because the geometry of the sphere and the physics of radiation are well known, we can calculate field strength and power density at any point. These figures can then be compared with the actual values from an antenna being tested. From spherical geometry, we can calculate isotropic power density at any distance R from the point source:

$$P_d = \frac{P}{4\pi R^2} \qquad \textbf{[19.19]}$$

where

P_d is the power density, in watts per square meter
P is the power in watts input to the isotropic radiator
R is the radius in meters at which point power density is measured

Example 19-5 Calculate the power density at a distance of 1 km (1000 m) from a 1000-W isotropic source.

Solution:

$$P_d = \frac{P}{4\pi R^2}$$

$$= \frac{(1000 \text{ W})}{4\pi\ (1000 \text{ m})^2}$$

$$= 7.95 \times 10^{-5} \text{ W/m}^2$$

The rest of this chapter covers antenna gains and directivities that are relative to isotropic radiators.

Near field and far field

Antennas are defined in terms of *gain* and *directivity,* both of which are measured by looking at the radiated field of the antenna. There are two fields to consider: *near field* and *far field.* The patterns published for an antenna tend to reflect far-field performance. The far field for most antennas falls off according to the *inverse square law.* That is, the intensity falls off according to the square of the distance ($1/R^2$), as in Eq. 19.19.

The near field of the antenna contains more energy than the far field because of the electric and magnetic fields close to the antenna radiator element. The near field tends to diminish rapidly according to a $1/R^4$ function. The minimum distance to the edge of the near field is a function of both the wavelength of the radiated signals and the antenna dimensions:

$$r_{\min} = \frac{2\,d^2}{\lambda} \qquad \textbf{[19.20]}$$

where

$r_{\min}$ is the near-field distance
d is the largest antenna dimension
λ is the wavelength of the radiated signal (all factors in the same units)

Example 19-6 An antenna with a length of 6 cm radiates a 12-cm wavelength signal. Calculate the near-field distance.

Solution:

$$r_{\min} = \frac{2d^2}{\lambda}$$

$$= \frac{(2)\,(6\text{ cm})^2}{12\text{ cm}}$$

$$= \frac{72}{12} = 6\text{ cm}$$

Antenna impedance

Impedance represents the total opposition to the flow of alternating current (e.g., RF), and includes both resistive and reactive components. The reactive components can be either capacitive or inductive, or a combination of both. Impedance can be expressed in either of two notations:

$$Z = \sqrt{R^2 + (X_L - X_c)^2} \qquad \textbf{[19.21]}$$

or

$$Z = R \pm jX \qquad \textbf{[19.22]}$$

Of these, Eq. 19.22 is perhaps the more commonly used in RF applications. The reactive part of antenna impedance results from the magnetic and electrical fields (close to the radiator) returning energy to the antenna radiator during every cycle. The resistive part of impedance consists of two elements: *ohmic losses* R_o and *radiation resistance* R_r. The ohmic losses are due to heating of the antenna conductor elements by RF current passing through, as when current passes through any conductor.

The radiation resistance relates to the radiated energy. An efficiency factor k compares the loss and radiation resistances:

$$k = \frac{R_r}{R_r + R_o} \qquad \textbf{[19.23]}$$

The goal of the antenna designer is to reduce R_o to a minimum. The value of R_r is set by the antenna design and installation, and is defined as the quotient of the voltage over the current at the feedpoint, less losses.

Dipole antenna elements

The *dipole* is a two-pole antenna (Fig. 19-20) that can be modeled as either a single radiator fed at the center (Fig. 19-20A) or a pair of radiators fed back to back (Fig. 19-20B). RF current from the source oscillates back and forth in the radiator element, causing an electromagnetic wave to propagate in a direction *perpendicular* to the radiator element. The polarity of any electromagnetic field is the direction of the electrical field vector (see Chap. 2). In the dipole, the polarization is parallel to the radiator element: a horizontal element produces a horizontally polarized signal, while a vertical element produces a vertically polarized signal.

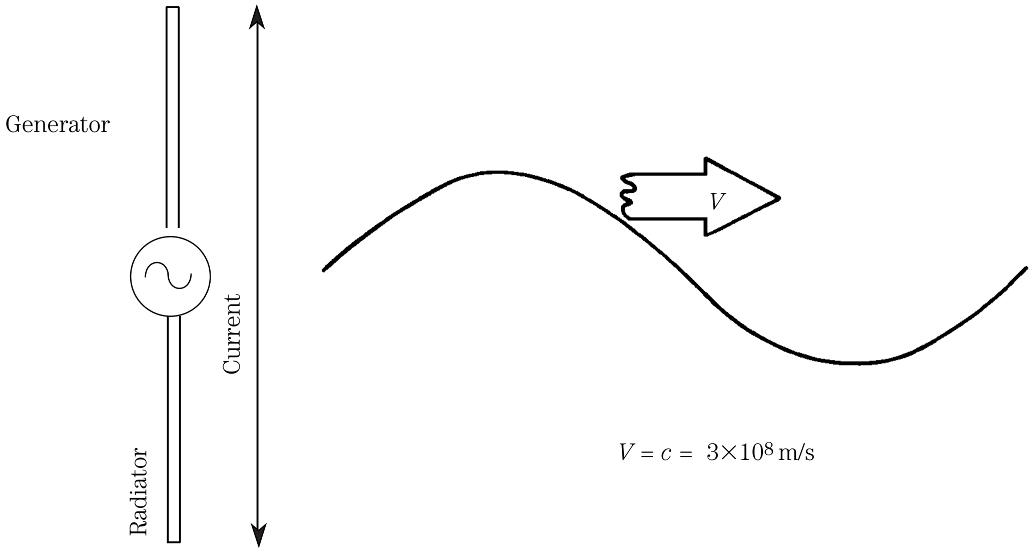

19-20A Basic dipole antenna showing propagation.

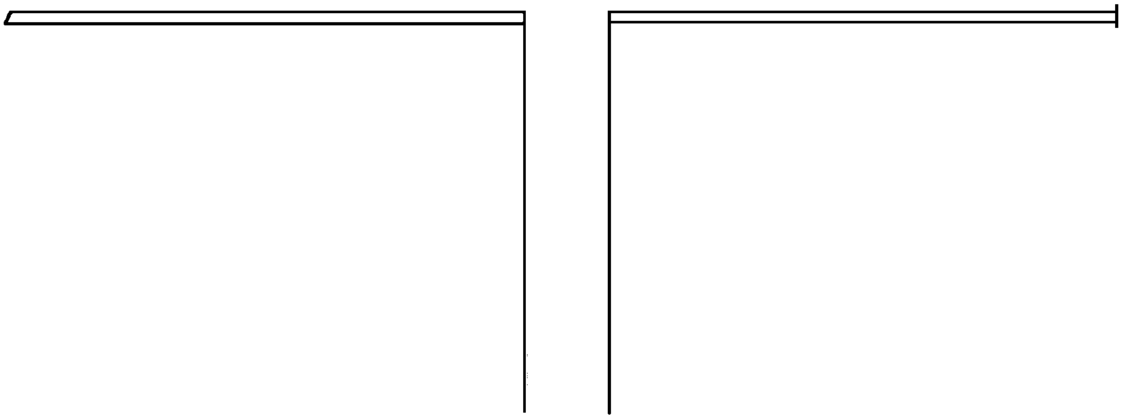

19-20B Basic dipole antenna.

Figure 19-21 shows the radiator patterns for the dipole viewed from two perspectives. Figure 19-21A shows the pattern of a horizontal half-wavelength dipole as viewed from above. This plot shows the *directivity* of the dipole: maximum radiation is found in two lobes perpendicular to the radiator length. The plot in Fig. 19-21B shows the end-on pattern of the dipole. This omnidirectional pattern serves for a vertically polarized dipole viewed from above. The end-on pattern of a horizontal dipole would be similar, except that it is distorted by ground effects unless the antenna is a very large number of wavelengths above the ground.

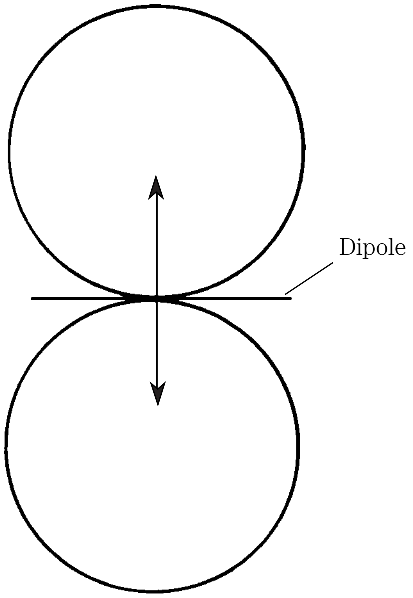

19-21A
Dipole pattern (horizontal).

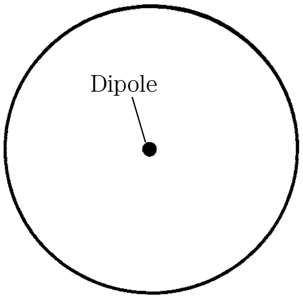

19-21B
Dipole pattern (vertical).

A microwave dipole is shown in Fig. 19-22. The antenna radiator element consists of a short conductor at the end of a section of waveguide. Although most low-frequency dipoles are a half-wavelength, microwave dipoles might be either a half-wavelength, less than a half-wavelength, or greater than a half-wavelength, depending upon application. For example, because most microwave dipoles are used to illuminate a reflector antenna of some sort, the length of the dipole depends upon the exact illumination function required for proper operation of the reflector. Most, however, will be a half-wavelength.

Antenna directivity and gain

The dipole discussed illustrated a fundamental property of the type of antenna generally used at microwave frequencies: *directivity* and *gain.* These two concepts are different but so interrelated that they are usually discussed at the same time. Because of the directivity, the antenna focuses energy in only two directions, which means that all of the energy is found in those directions (Fig. 19-21A), rather than being distributed over a spherical surface. Thus, the dipole has a gain approximately 2.1 dB greater than isotropic. In other words, the measured power density at any point will be 2.1 dB higher than the calculated isotropic power density for the same RF input power to the antenna.

Directivity The directivity of an antenna is a measure of its ability to direct RF energy in a limited direction, rather than in all (spherical) directions equally. As shown in Fig. 19-21A, the horizontal directivity of the dipole forms a bidirectional figure-8 pattern. Two methods for showing unidirectional antenna patterns are shown in Fig. 19-23. The method of Fig. 19-23A is a polar plot viewed from above. The main lobe is centered on 0°. The plot of Fig. 19-23B is a rectangular method for displaying the same information. This pattern follows a (sin x)/x function or, for power, $[(\sin x)/x]^2$.

Directivity D is a measure of relative power densities:

$$D = \frac{P_{\text{max}}}{P_{\text{av}}} \tag{19.24}$$

Or, referenced to isotropic,

$$D = \frac{4\pi}{\Phi} \tag{19.25}$$

where

D is the directivity
P_{max} is the maximum power
P_{av} is the average power
Φ is the solid angle subtended by the main lobe

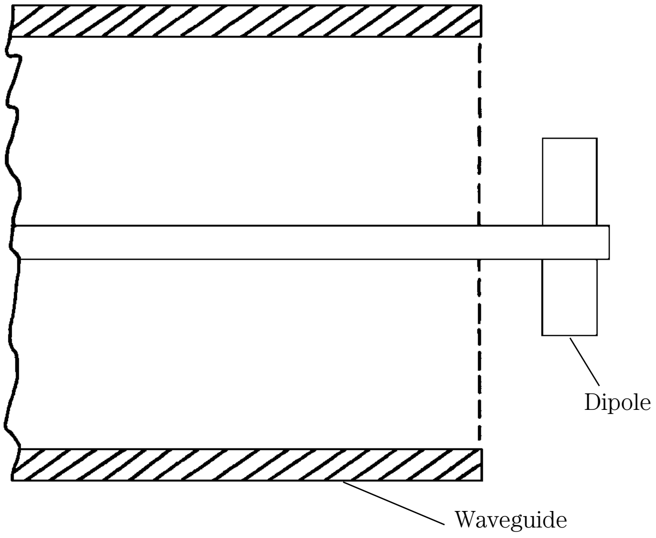

19-22
Microwave dipole radiator.

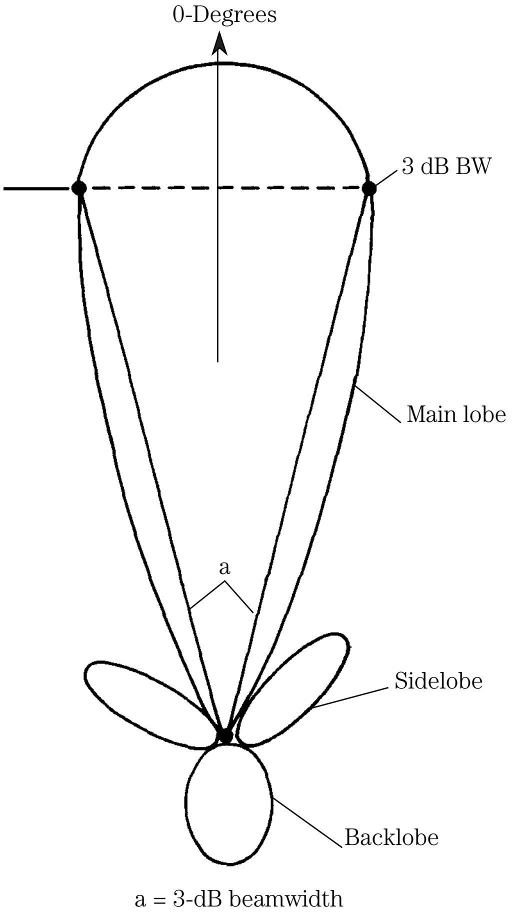

19-23A
Directional antenna pattern (top view).

The term Φ is a solid angle, which emphasizes the fact that antenna patterns must be examined in at least two extents: horizontal and vertical.

A common method for specifying antenna directivity is *beamwidth* (BW). The definition of BW is the angular displacement between points on the main lobe (see Figs. 19-23A and 19-23B), where the power density drops to one-half (−3 dB) of its maximum main lobe power density. This angle is shown in Fig. 19-23A as *a*.

In an ideal antenna system, 100 percent of the radiated power is in the main lobe, and there are no other lobes. But in real antennas certain design and installation anomalies cause additional minor lobes, such as the *sidelobes* and *backlobe* shown in Fig. 19-23A. Several problems derive from the minor lobes. First is the loss of usable power. For a given power density required at a distant receiver site, the transmitter must supply whatever additional power is needed to make up for the minor lobe losses.

The second problem is intersystem interference. A major application of directional antennas is the prevention of mutual interference between nearby cochannel stations. In radar systems, high sidelobes translate to errors in detected targets. If, for example, a sidelobe is large enough to detect a target, then the radar display will show this off-axis target as if it was in the main lobe of the antenna. The result is an azimuth error that could be important in terms of marine and aeronautical navigation.

Gain Antenna gain derives from the fact that energy is squeezed into a limited space instead of being distributed over a spherical surface. The term *gain* implies that the antenna creates a higher power when, in fact, it merely concentrates the power into a single direction that would otherwise be spread out over a larger area. Even so, it is possible to speak of an apparent increase in power. Antenna-transmitter systems are often rated in terms of *effective radiated power* (ERP). The ERP is the product of the transmitter power and the antenna gain. For example, if an antenna has a gain of +3 dB, the ERP will be twice the transmitter output power. In other words, a 100-W output transmitter connected to a +3-dB antenna will produce a power density at a distant receiver equal to a 200-W transmitter feeding an isotropic radiator. There are two interrelated gains to be considered: *directivity gain* G_d *and power gain* G_p.

The directivity gain is defined as the quotient of the maximum radiation intensity over the average radiation intensity (note the similarity to the directivity definition). This measure of gain is based on the shape of the antenna radiation pattern, and can be calculated with respect to an isotropic radiator ($D=1$) from:

$$G_d = \frac{4\pi P_a}{P_r} \qquad \textbf{[19.26]}$$

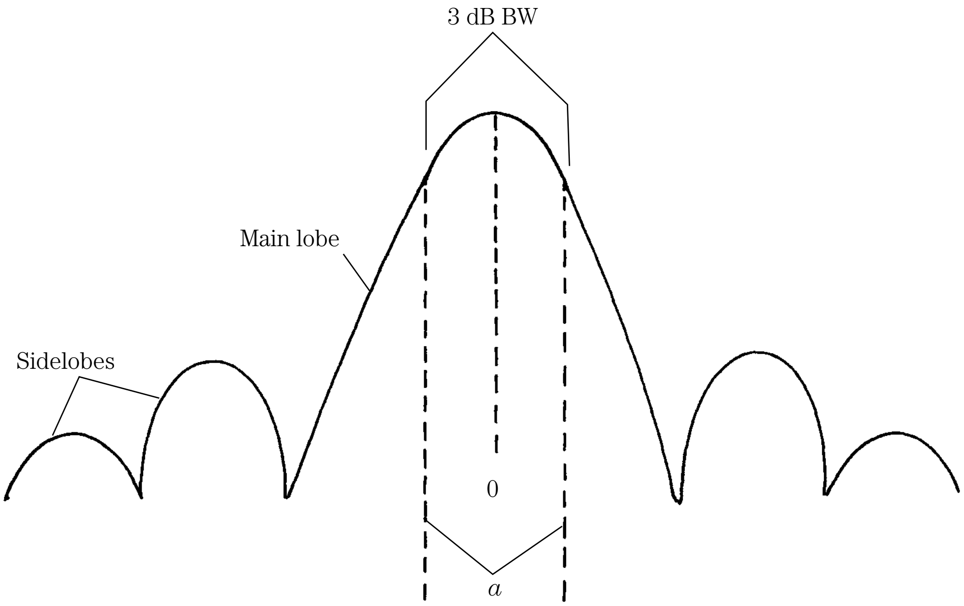

19-23B Graphically presented pattern.

where

G_d is the directivity gain
P_a is the maximum power radiated per unit of solid angle
P_r is the total power radiated by the antenna

The power gain is similar in nature, but slightly different from directivity gain; it includes dissipative losses in the antenna. Not included in the power gain are losses caused by cross-polarization or impedance mismatch between the waveguide (or transmission line) and the antenna. There are two commonly used means for determining power gain:

$$G_p = \frac{4\pi P_a}{P_n} \qquad \textbf{[19.27]}$$

and

$$G_p = \frac{P_{\text{ai}}}{P_i} \qquad \textbf{[19.28]}$$

where

P_a is the maximum radiated per unit solid angle
P_n is the net power accepted by the antenna (i.e., less mismatch losses)
P_{ai} is the average intensity at a distant point
P_i is the intensity at the same point from an isotropic radiator fed the same RF power level as the antenna

(Equations assume equal power to antenna and comparison isotropic source.)

Provided that ohmic losses are kept negligible, the relationship between directivity gain and power gain is given by:

$$G_p = \frac{P_r G_d}{P_n} \qquad \textbf{[19.29]}$$

(All terms as previously defined.)

Relationship of gain and aperture Antennas obey the law of reciprocity, which means that any given antenna will work as well to receive as to transmit. The function of the receiver antenna is to gather energy from the electromagnetic field radiated by the transmitter antenna. The aperture is related to, and often closely approximates, the physical area of the antenna. But in some designs the effective aperture A_e is less than the physical area A, so there is an effectiveness factor n that must be applied. In general, however, a high-gain transmitter antenna also exhibits a high receiving aperture, and the relationship can be expressed as

$$G = \frac{4\,\pi A_e n}{\lambda^2} \qquad \textbf{[19.30]}$$

where

A_e is the effective aperture
n is the aperture effectiveness ($n=1$ for a perfect, lossless antenna)
λ is the wavelength of the signal

Horn antenna radiators

The horn radiator is a tapered termination of a length of waveguide (see Fig. 19-24 (A-C) that provides the impedance transformation between the waveguide impedance and the free-space impedance. Horn radiators are used both as antennas in their own right, and as illuminators for reflector antennas. Horn antennas are not a perfect match to the waveguide, although standing wave ratios of 1.5:1 or less are achievable. The gain of a horn radiator is proportional to the area A of the flared open flange ($A = ab$ in Fig. 19-24B), and inversely proportional to the square of the wavelength:

$$G=\frac{10A}{\lambda^2} \qquad \textbf{[19.31]}$$

where

A is the flange area
λ is the wavelength (both in same units)

The −3-dB beamwidth for vertical and horizontal extents can be approximated from:

Vertical

$$\Phi_v = \frac{51\lambda}{b} \quad \text{degrees} \qquad \textbf{[19.32]}$$

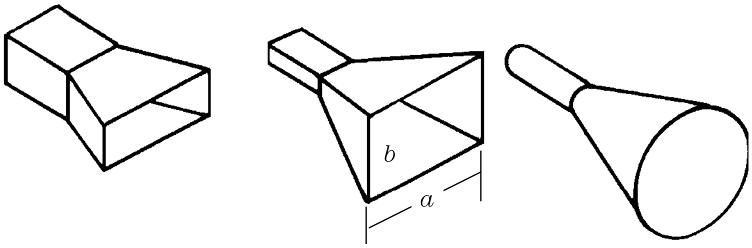

Rectangular Pyramidal Conical

A **B** **C**

19-24 Horn radiators.

Horizontal

$$\Phi_h = \frac{70\lambda}{a} \qquad \text{degrees} \tag{19.33}$$

where

Φ_v is the vertical beamwidth, in degrees
Φ_h is the horizontal beamwidth, in degrees
a, b are dimensions of the flared flange
λ is the wavelength

A form of antenna, related to the horn, is the cavity antenna of Fig. 19-25. In this type of antenna, a quarter-wavelength radiating element extends from the waveguide (or transmission line connector) into a resonant cavity. The radiator element is placed a quarter-wavelength into a resonant cavity; and is spaced a quarter-wavelength from the rear wall of the cavity. A tuning disk is used to alter cavity dimensions in order to provide a limited tuning range for the antenna. Gains to about 6 dB are possible with this arrangement.

Reflector antennas

At microwave frequencies, it becomes possible to use *reflector antennas* because of the short wavelengths involved. Reflectors are theoretically possible at lower frequencies, but because of the longer wavelengths, the antennas would be so large that they become impractical. Several forms of reflector are used (Figs. 19-26 and 19-27). In Fig. 19-26 we see the *corner reflector* antenna, which is used primarily in the high-UHF and low-microwave region. A dipole element is placed at the "focal point" of the corner reflector, so it receives (in phase) the reflected wavefronts from the surface. Either solid metallic reflector surfaces or wire mesh may be used. When mesh is used, however, the holes in the mesh must be $\frac{1}{12}$-wavelength or smaller.

Figure 19-27 shows several other forms of reflector surface shape, most of which are used in assorted radar applications.

Parabolic "dish" antennas

The parabolic reflector antenna is one of the most widespread of all the microwave antennas, and is the type that normally comes to mind when thinking of microwave systems. This type of antenna derives its operation from physics similar to optics, and is possible because microwaves are in a transition region between ordinary radio waves and infrared/visible light.

The dish antenna has a paraboloid shape as defined by Fig. 19-28. In this figure, the dish surface is positioned such that the center is at the origin (0,0) of an x-y coordinate system. For purposes of defining the surface, we place a second vertical axis called the *directrix* (y') a distance behind the surface equal to the focal length (u). The paraboloid surface follows the function $y^2 = 4ux$, and has the property that a line from the focal point F to any point on the surface is the same length as a line from that same point to the directrix (in other words, MN = MF).

If a radiator element is placed at the focal point F, then it will illuminate the reflector surface, causing wavefronts to be propagated away from the surface in

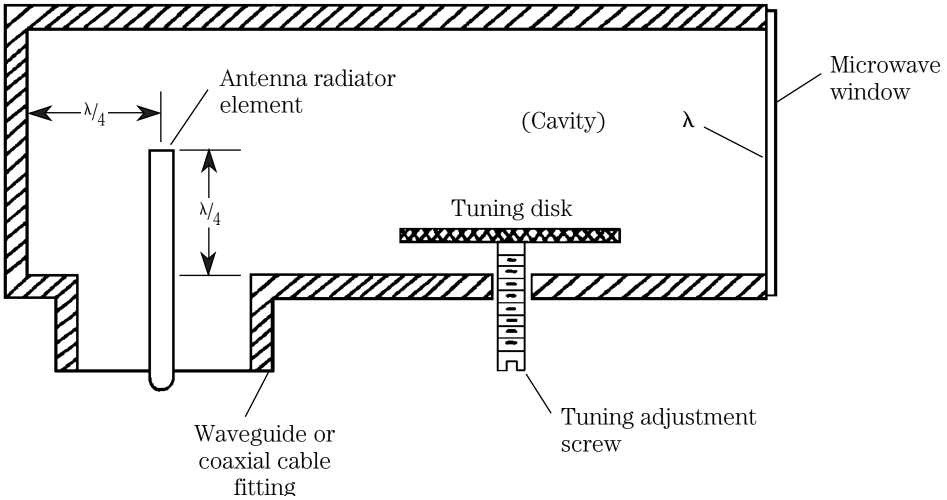
Antenna radiator element
$\lambda/4$
$\lambda/4$
(Cavity)
λ
Microwave window
Tuning disk
Waveguide or coaxial cable fitting
Tuning adjustment screw

19-25 Cavity antenna.

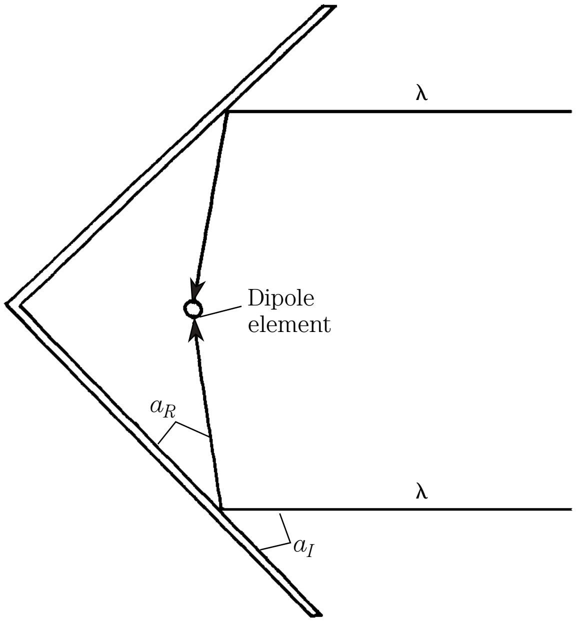
λ
Dipole element
a_R
λ
a_I

19-26 Corner reflector.

phase. Similarly, wavefronts, intercepted by the reflector surface, are reflected to the focal point.

Gain The gain of a parabolic antenna is a function of several factors: dish diameter, feed illumination, and surface accuracy. The dish diameter D should be large compared with its depth. Surface accuracy refers to the degree of surface irregularities. For commercial antennas, $\frac{1}{8}$-wavelength surface accuracy is usually sufficient,

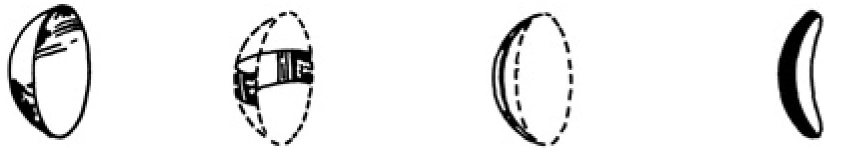

A. Paraboloid

B. Truncated paraboloid (surface search)

C. Truncated paraboloid (height finding)

D. Orange-peel paraboloid

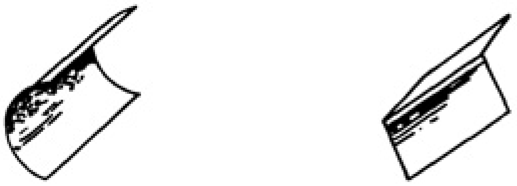

E. Cylindrical paraboloid

F. Corner reflector

19-27 Reflector antennas.

although on certain radar antennas the surface accuracy specification must be tighter.

The feed illumination refers to how evenly the feed element radiates to the reflector surface. For circular parabolic dishes, a circular waveguide feed produces optimum illumination, and rectangular waveguides are less than optimum. The TE_{11} mode is desired. For best performance, the illumination should drop off evenly from the center to the edge, with the edge being 210 dB down from the center. The diameter, length, and beamwidth of the radiator element (or horn) must be optimized for the specific *F/d* ratio of the dish. The cutoff frequency is approximated from

$$f_{\text{cutoff}} = \frac{175{,}698}{d_{\text{mm}}} \quad \textbf{[19.34]}$$

where

f_{cutoff} is the cutoff frequency
d is the inside diameter of the circular feedhorn

The gain of the parabolic dish antenna is found from

$$G = \frac{k\,(\pi D)^2}{\lambda^2} \quad \textbf{[19.35]}$$

where

G is the gain over isotropic
D is the diameter
λ is the wavelength (same units as D)
k is the reflection efficiency (0.4 to 0.7, with 0.55 being most common)

The −3-dB beamwidth of the parabolic dish antenna is approximated by

$$\mathrm{BW} = \frac{70\,\lambda}{D} \qquad \textbf{[19.36]}$$

and the focal length by

$$F = \frac{D^2}{16d} \qquad \textbf{[19.37]}$$

For receiving applications, the effective aperture is the relevant specification and is found from

$$A_e = k\pi(D/2)^2 \qquad \textbf{[19.38]}$$

The antenna pattern radiated by the antenna is similar to Fig. 19-23B. With horn illumination, the sidelobes tend to be 23 to 28 dB below the main lobe, or 10 to 15 dB below isotropic. It is found that 50 percent of the energy radiated by the parabolic dish is within the −3-dB beamwidth, and 90 percent is between the first nulls on either side of the main lobe.

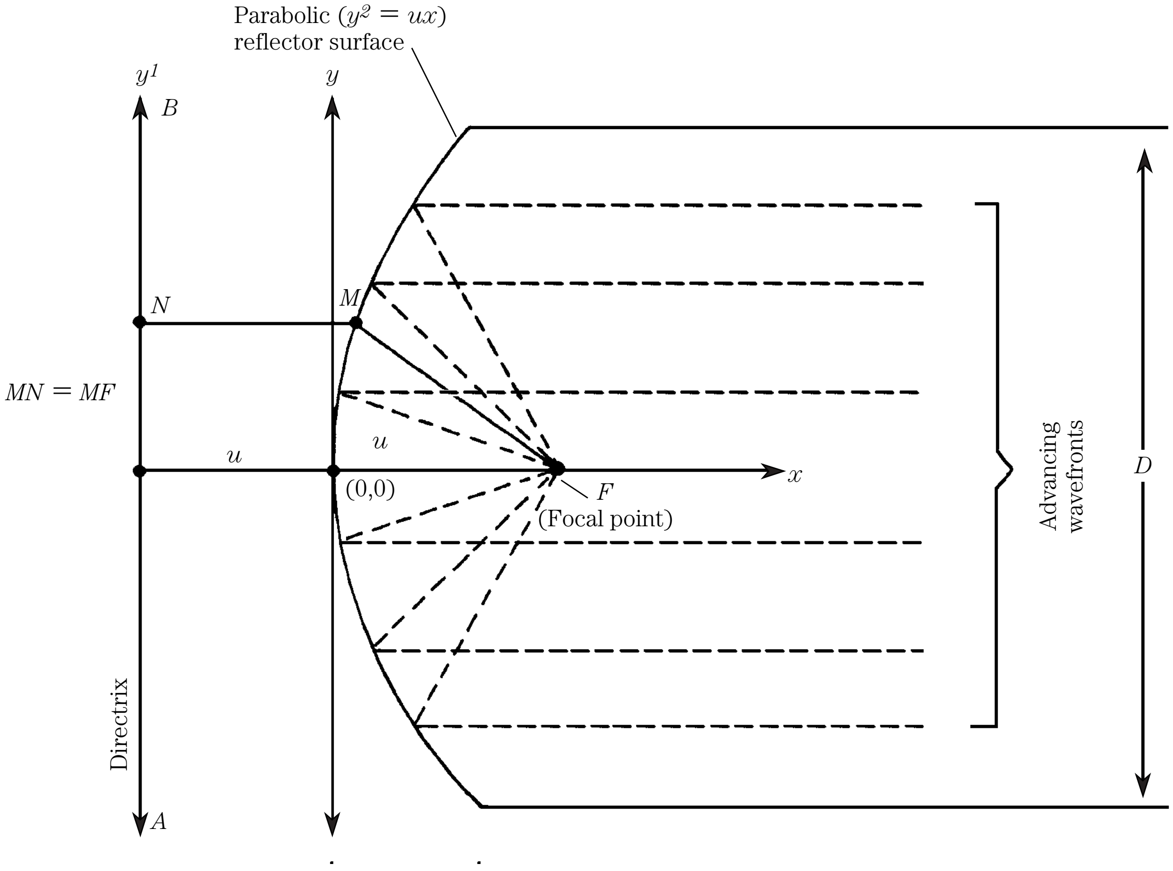
Parabolic ($y^2 = ux$)
reflector surface
y^1
y
B
N
M
$MN = MF$
u
u
(0,0)
x
F
(Focal point)
Advancing wavefronts
D
Directrix
A

19-28 Ray tracing shows operation of parabolic antenna.

If a dipole element is used for the feed device, then a *splash plate* is placed ¼-wavelength behind the dipole in order to improve illumination. The splash plate must be several wavelengths in diameter, and is used to reflect the backlobe back toward the reflector surface. When added to the half-wave phase reversal inherent in the reflection process, the two-way quarter-wavelength adds another half-wavelength and thereby permits the backwave to move out in phase with the front lobe wave.

Parabolic dish feed geometries Figure 19-29 shows two methods for feeding parabolic dish antennas, regardless of which form of radiator (horn, dipole, etc.) is used. In Fig. 19-29A we see the method in which the radiator element is placed at the focal point, and a waveguide (or transmission line) is routed to it. This method is used in low-cost installations such as home satellite TV receive-only (TVRO) antennas.

Figure 19-29B shows the *Cassegrain feed* system. This system is modeled after the Cassegrain optical telescope. The radiator element is placed at an opening at the center of the dish. A hyperbolic subreflector is placed at the focal point, and it is used to reflect the wavefronts to the radiator element. The Cassegrain system results in lower-noise operation because of several factors: less transmission line length, lower sidelobes, and the fact that the open horn sees sky instead of earth

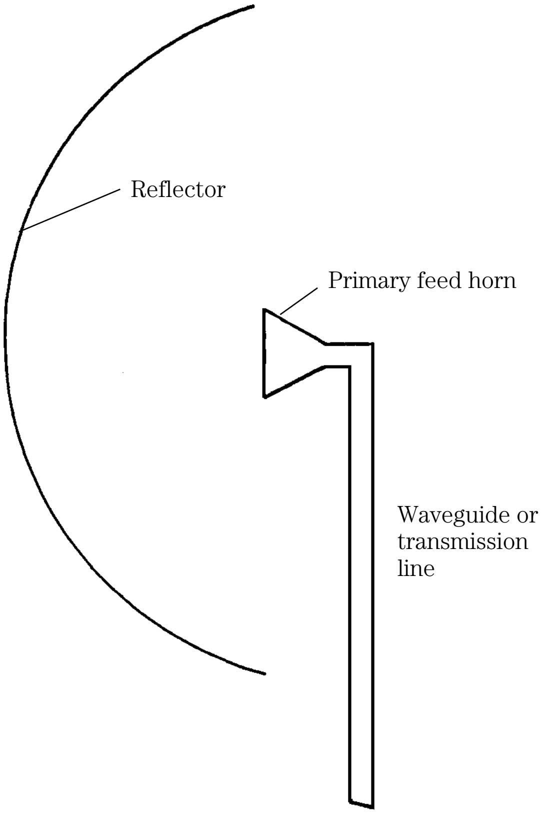

19-29A
Parabolic antenna conventional feed.

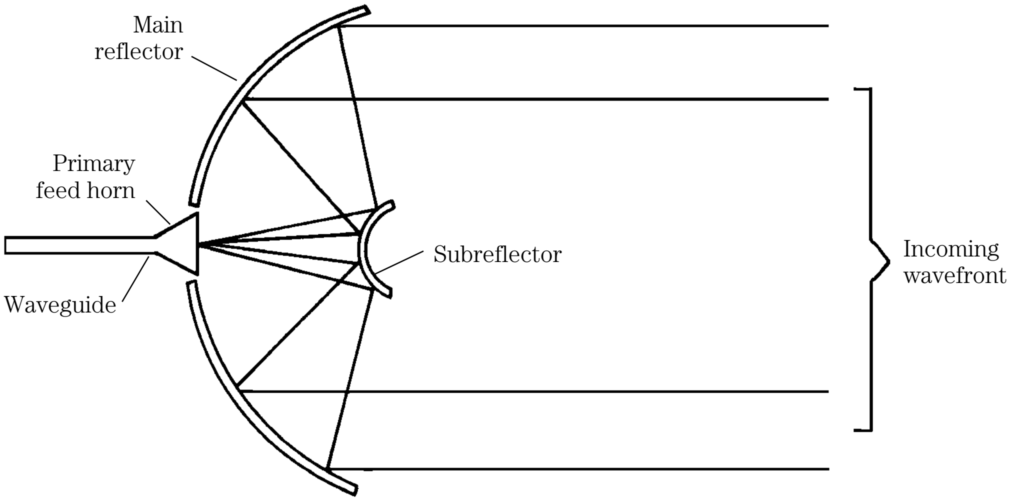

19-29B Parabolic antenna Cassegrain feed.

(which has lower temperature); on the negative side, galactic and solar noise might be slightly higher on a Cassegrain dish.

Figure 19-30A shows the *monopulse* feed geometry. In this system, a pair of radiator elements are placed at the focal point, and fed to a power splitter network that outputs both *sum* and *difference* signals (Fig. 19-30B). When these are combined, the resultant beam shape has improved −3-dB beamwidth as a result of the algebraic summation of the two.

Array antennas

When antenna radiators are arranged in a precision array, an increase in gain occurs. An array might be a series of dipole elements, as in the broadside array of Fig. 19-31 (which is used in the UHF region), or a series of slots, horns, or other radiators. The overall gain of an array antenna is proportional to the number of elements, as well as the details of their spacing. In this, and other antennas, a method of *phase shifting* is needed. In Fig. 19-31, the phase shifting is caused by the crossed feeding of the elements, but in more modern arrays, other forms of phase shifter are used.

Two methods of feeding an array are shown in Fig. 19-32. The *corporate feed* method connects all elements, and their phase shifters, in parallel with the source. The *branch feed* method breaks the waveguide network into two (or more) separate paths.

Solid-state array antennas Some modern radar sets use solid-state array antennas consisting of a large number of elements, each of which is capable of shifting the phase of a microwave input signal. Two forms are known: *passive* (Fig. 19-33A) and *active* (Fig. 19-33B). In the passive type of element, a ferrite (or PIN diode) phase shifter is placed in the transmission path between the RF input and the radiator element (usually a slot). By changing the phase of the RF signal selectively, it is possible to form and steer the beam at will. A 3-bit (i.e., three discrete state) phase

Chap19_Carr 4/10/01 3:40 PM Page 410

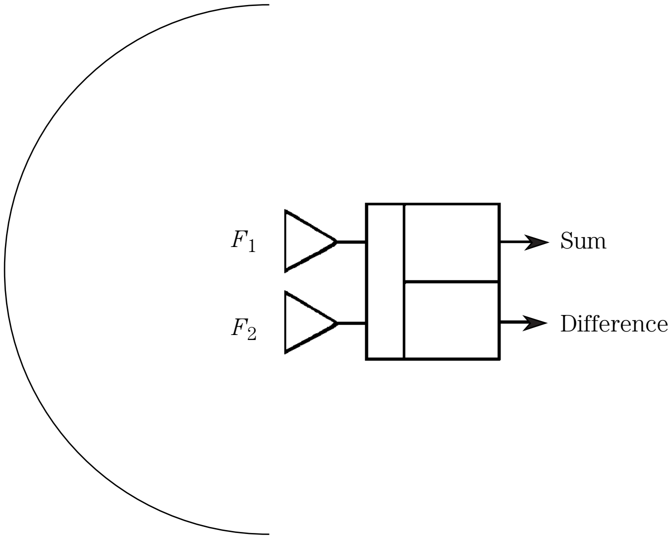
F_1
F_2
Sum
Difference

19-30A
Monopulse feed.

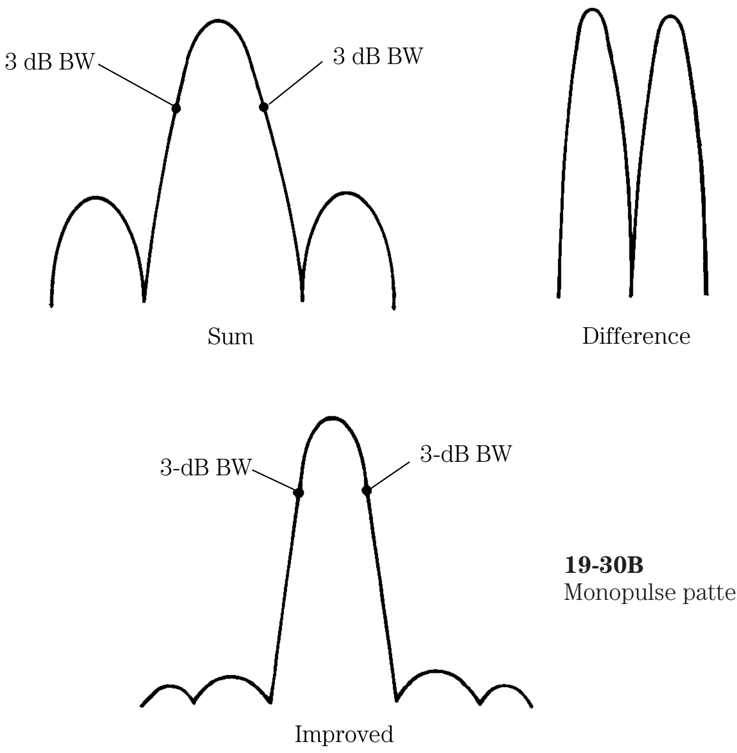

19-30B
Monopulse patterns.

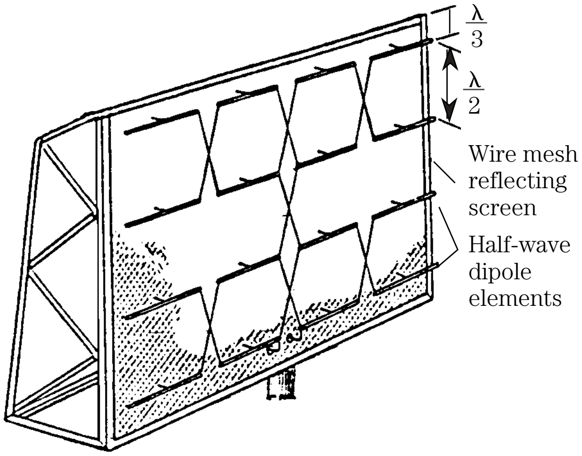

19-31
Reflector array antenna.

shifter allows the phase to shift in 45° increments, while a 4-bit phase shifter allows 22.5° increments of phase shift.

The active element contains, in addition to a phase shifter, a transmit power amplifier (1 or 2 W) and a low-noise amplifier (LNA) for receiving. A pair of transmit/receive (T/R) switches select the path to which the RF signal is directed. The total output power of this antenna is the sum of all output powers from all elements in the array. For example, an array of one thousand 2-W elements makes a 2000-W system.

Slot array antennas

A resonant slot (cut into a wall of a section of waveguide) is somewhat analogous, if not identical, to a dipole. By cutting several slots in the waveguide, we obtain the advantages of an array antenna in which the elements are several slot radiators. Slot array antennas are used for marine navigation radars, telemetry systems, and the reception of microwave television signals in the Multipoint Distribution Service (MDS) on 2.145 GHz.

Figure 19-34 shows a simple slot antenna used in telemetry applications. A slotted section of rectangular waveguide is mounted to a right-angle waveguide flange. An internal wedge (not shown) is placed at the top of the waveguide and serves as a matching-impedance termination to prevent internal reflected waves. Directivity is enhanced by attaching flanges to the slotted section of waveguide parallel to the direction of propagation (see end view of Fig. 19-34).

Figure 19-35 shows two forms of *flatplate array* antennas constructed from slotted waveguide radiator elements (shown as insets). Figure 19-35A shows the rectangular array, and Fig. 19-35B shows the circular array. These flatplate arrays are used extensively in microwave communications and radar applications.

The feed structure for a flatplate array is shown in Fig. 19-35C. The antenna element is the same sort as shown in Figs. 19-35A and 19-35B. A distribution waveguide is physically mated with the element, and a coupling slot is provided between the two waveguides. Energy propagating in the distribution system waveguide is

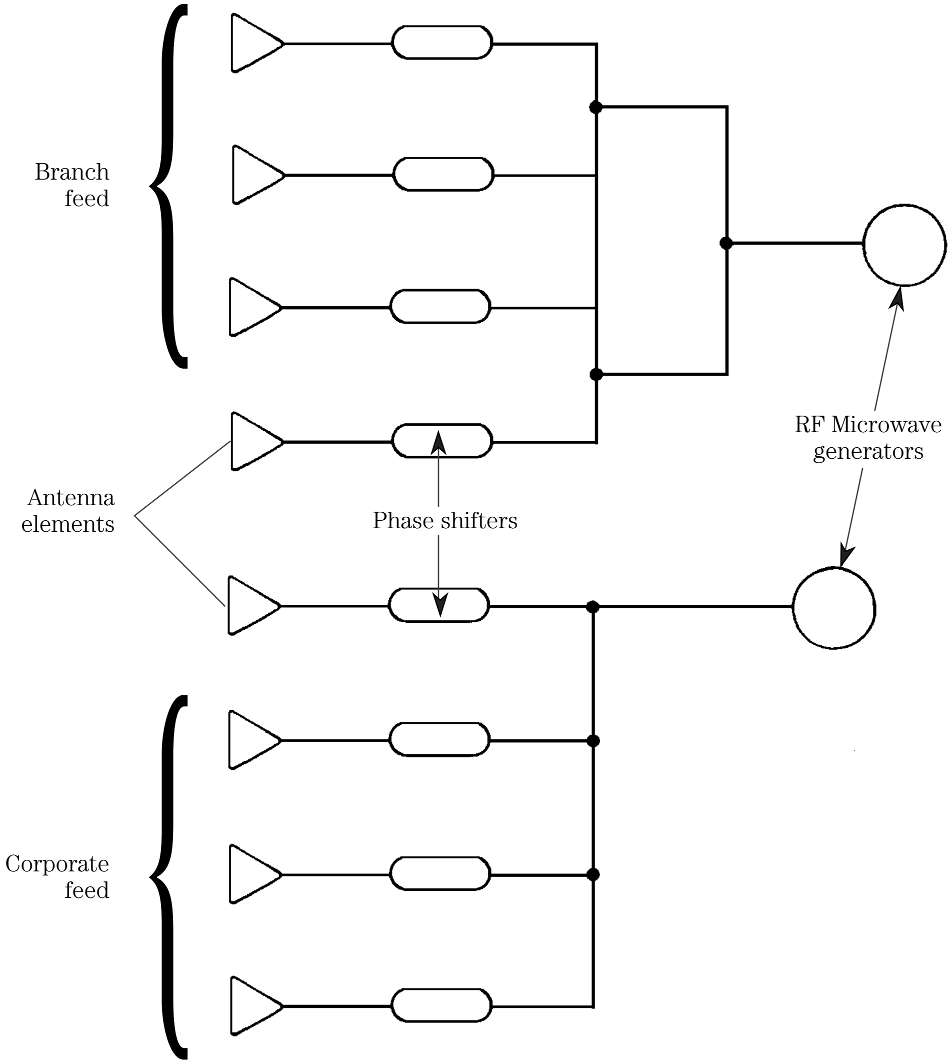
Branch
feed
Antenna
elements
Phase shifters
RF Microwave
generators
Corporate
feed

19-32 Branch feed and corporate feed.

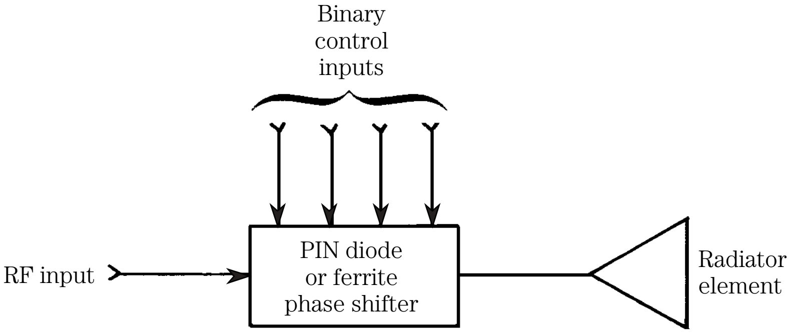
Binary
control
inputs
RF input
PIN diode
or ferrite
phase shifter
Radiator
element

19-33A Phase shifter.

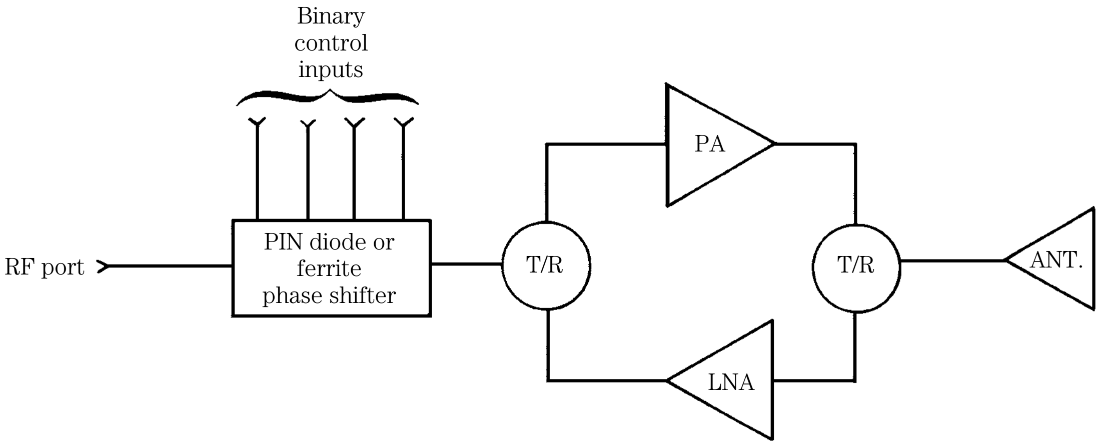
Binary
control
inputs
PA
RF port
PIN diode or
ferrite
phase shifter
T/R
T/R
ANT.
LNA

19-33B Phase-shift T/R element.

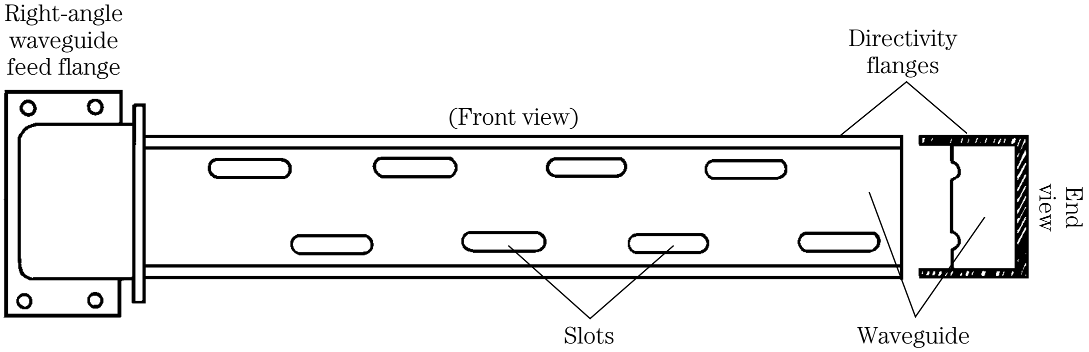
Right-angle
waveguide
feed flange
Directivity
flanges
(Front view)
End
view
Slots
Waveguide

19-34 Slot antenna (front view).

coupled into the antenna radiator element through this slot. In some cases metallic or dielectric phase-shifting stubs are also used in order to fine-tune the antenna radiation pattern.

Microwave antenna safety note

Microwave RF energy is dangerous to your health. Anything that can cook a roast beef can also cook you! The U.S. government sets a safety limit for microwave exposure of 10 mW/cm^2 averaged over 6 minutes; some other countries use a level one-tenth of the U.S. standard. The principal problem is tissue heating, and eyes seem especially sensitive to microwave energy. Some authorities believe that cataracts form from prolonged exposure. Some authorities also believe that genetic damage to offspring is possible as well as other long-term effects as a result of cumulative exposure.

Because of their relatively high gain, microwave antennas can produce hazardous field strengths in close proximity—even at *relatively low RF input power levels.* At

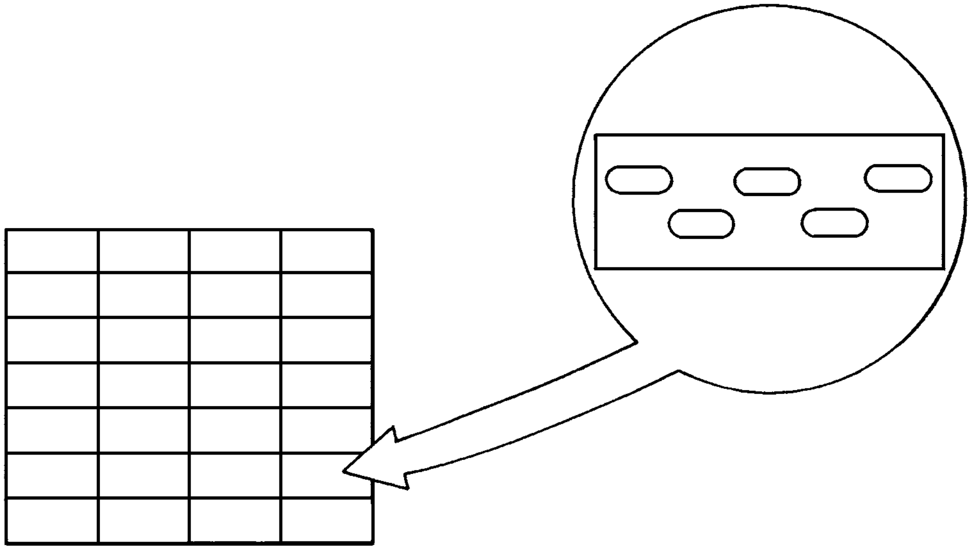

19-35A Flatplate slot array.

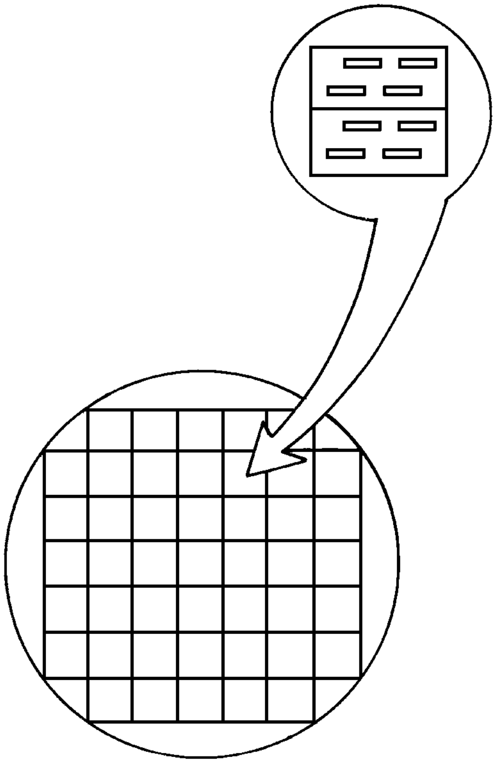

19-35B
Flatplate antenna.

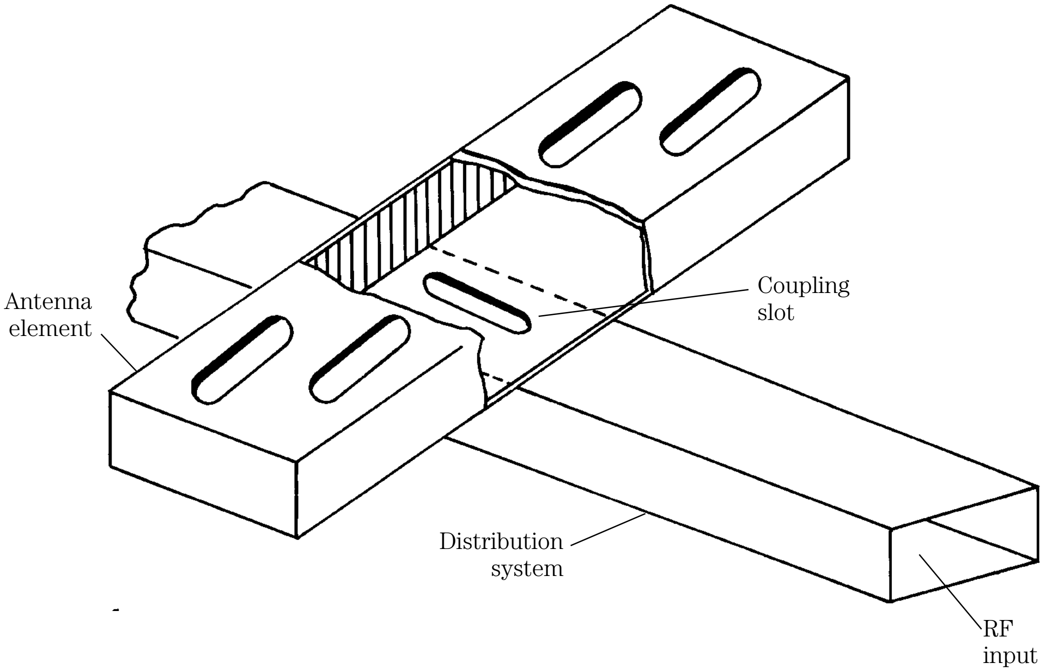

19-35C Flatplate antenna feed coupling.

least one technician in a TV satellite earth station suffered abdominal adhesions, solid matter in the urine, and genital dysfunction after servicing a 45-m-diameter 3.5-GHz antenna with RF power applied.

Be very careful around microwave antennas. Do not service a radiating antenna. When servicing nonradiating antennas, be sure to stow them in a position that prevents the inadvertent exposure of humans, should power accidentally be applied. A *Radiation Hazard* sign should be prominently displayed on the antenna. Good design practice requires an interlock system that prevents radiation in such situations. “Hot” transmitter service should be performed with a shielded dummy load replacing the antenna.

This page intentionally left blank

20
CHAPTER

Antenna noise temperature

RADIO RECEPTION IS ESSENTIALLY A MATTER OF SIGNAL-TO-NOISE RATIO (SNR). SIGNALS must be some amplitude above the noise floor of the system in order to be received properly. All electronic systems (receivers and antennas included) have inherent noise, even if there is no power flowing in them. One of the goals of the antenna designer is to minimize the noise so that weak signals are not obscured. One of the basic forms of noise seen in systems is the *thermal noise.* Even if the amplifiers in the receiver add no additional noise (they will!), there will be thermal noise at the input.

If you replace the antenna with a resistor matched to the system impedance that is totally shielded, there will still be noise present. The noise is produced by the random motion of electrons inside the resistor. At all temperatures above absolute zero (about −273.16°C) the electrons in the resistor material are in random motion. At any given instant there will be a huge number of electrons in motion in all directions. The reason why there is no discernible current flow in one direction is that the motions cancel each other out even over short time periods. The noise power present in a resistor is:

$$P_N = KTBR \qquad \textbf{[20.1]}$$

where

P_N is the noise power, in watts
T is the temperature, in kelvins (K)
K is Boltzmann's constant (1.38×10^{-23} K)
B is the bandwidth, in hertz
R is the resistance, in ohms

Note: By international agreement T is set to 290 K.

Consider a receiver with a 1-MHz bandwidth and an input resistance of 50 Ω. The noise power is $(1.38 \times 10^{-23}\ \text{K}) \times (290\ \text{K}) \times (1{,}000{,}000\ \text{Hz}) \times (50\ \Omega) = 2 \times 10^{-13}$ W. This noise is called *thermal noise, thermal agitation noise,* or *Johnson noise.*

A resonant antenna can be modeled as an impedance consisting solely of a resistor with a value equal to the feedpoint impedance. If an antenna has a 50-Ω feedpoint

Copyright 2001 - the McGraw-Hill Companies

impedance, then it will generate exactly the same amount of thermal noise as a resistor of the same value.

Noise factor, noise figure, and noise temperature

The noise performance of a receiving system can be defined in three different, but related, ways: *noise factor* F_n, *noise figure (NF),* and *equivalent noise temperature* T_e; these properties are definable as a simple ratio, decibel ratio, or kelvin temperature, respectively.

Noise factor (F_n)

For components such as resistors, the noise factor is the ratio of the noise produced by a real resistor to the simple thermal noise of an ideal resistor. The noise factor of a radio receiver (or any system) is the ratio of output noise power P_{no} to input noise power P_{ni}:

$$F_n = \left.\frac{P_{\text{no}}}{P_{\text{ni}}}\right|_{T=290\text{ K}} \quad \textbf{[20.2]}$$

In order to make comparisons easier, the noise factor is usually measured at the standard temperature (T_o) of 290 K (standardized room temperature), although in some countries 299 K or 300 K is commonly used (the differences are negligible).

It is also possible to define noise factor F_n in terms of output and input signal-to-noise ratios:

$$F_n = \frac{S_{\text{ni}}}{S_{\text{no}}} \quad \textbf{[20.3]}$$

where

S_{ni} is the input signal-to-noise ratio
S_{no} is the output signal-to-noise ratio

Noise figure (NF)

The noise figure is a frequently used measure of a receiver's "goodness," or its departure from "idealness." Thus, it is a *figure of merit.* The noise figure is the noise factor converted to decibel notation:

$$\text{NF} = 10 \log F_n \quad \textbf{[20.4]}$$

where

NF is the noise figure, in decibels
F_n is the noise factor
LOG refers to the system of base 10 logarithms

Noise temperature (T_e)

The noise "temperature" is a means for specifying noise in terms of an equivalent temperature. Evaluating the noise equations shows that the noise power is directly proportional to temperature in kelvins, and also that noise power collapses to zero at the temperature of absolute zero (0 K).

Note that the equivalent noise temperature T_e is *not* the physical temperature of the amplifier, but rather a theoretical construct that is an *equivalent* temperature that produces that amount of noise power. The noise temperature is related to the noise factor by

$$T_e = (F_n - 1)\,T_o \qquad \textbf{[20.5]}$$

and to noise figure by

$$T_e = KT_o \log^{-1}\left(\frac{\mathrm{NF}}{10}\right) - 1 \qquad \textbf{[20.6]}$$

Noise temperature is often specified for receivers and amplifiers in combination with, or in lieu of, the noise figure. The noise temperature concept is also applied to antennas where it is related to the amount of thermal noise generated by the resistive component of the antenna feedpoint impedance.

The antenna-receiver system will be afflicted by three different noise sources external to the receiver. The first is the thermal noise temperature of the feedpoint impedance (T_R). The sky exhibits a noise temperature that depends on where the antenna main lobe is pointed. Similarly, the ground has a noise temperature that consists of components reflected from the sky as well as components of its own caused by whatever thermal agitation exists. In a typical system (Fig. 20-1) the

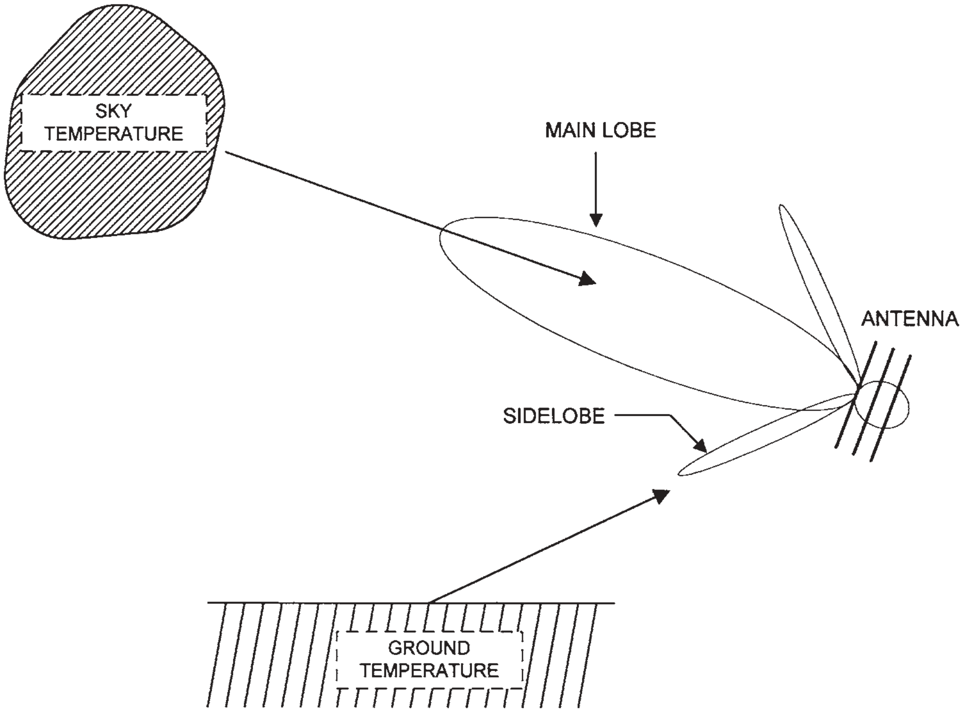

20-1 Contributors to antenna noise temperature.

main lobe will be pointed toward the sky noise source, while the sidelobes will pick up noise from the ground. The total noise temperature of the antenna is

$$T_{\text{ANT}} = (M \times T_{\text{SKY}}) + \xi\,(1 - M)T_{\text{GND}} + T_R \qquad \textbf{[20.7]}$$

where

T_{ANT} is the equivalent noise temperature of the antenna
T_{SKY} is the noise temperature of the sky
T_{GND} is the noise temperature of the ground
T_R is the feedpoint resistance noise temperature
M is the fraction of the total energy that enters the main lobe
ξ is the fraction of sidelobes that are viewing the ground (only one of several sidelobes is shown in Fig. 20-1)

21
CHAPTER

Antennas for radio astronomy

FOR CENTURIES ASTRONOMERS HAVE SCANNED THE HEAVENS WITH OPTICAL TELESCOPES. But today, astronomers have many more tools in their bag, and one of them is *radio* astronomy. The field of radio astronomy emerged in the 1930s and 1940s through the work of Grote Reber and Carl Jansky. Even during World War II, progress was made as many tens of thousands of operators were listening to frequencies from dc to near daylight (well, actually, the low-end microwave bands). British radar operators noted during the Battle of Britain that the distance at which they could detect German aircraft dropped when the Milky Way was above the horizon.

Although there is a lot of amateur radio astronomy being done, most of it requires microwave equipment with low-noise front ends. However, there are several things that almost anyone can do.

The topic of antennas for radio astronomy can include nearly all forms of directional gain antenna. It is common to see Yagis, ring Yagis, cubical quads, and other antennas for lower-frequency use (18 to 1200 MHz). Microwave gain antennas can be used for higher frequencies. Indeed, many amateur radio astronomers appropriate TV receive-only (TVRO) satellite dish antennas for astronomy work. In this chapter we will cover some antennas that are not found in other chapters, at least not in this present context.

Jupiter reception

Jupiter is a strong radio source (next to the sun, it's the strongest in the sky). It produces noiselike signals over the spectrum 5 to 40 MHz, with peaks between 18 and 24 MHz. One source claims that the radio signals come from massive storms on the largest planet's surface, apparently triggered by the transit of the jovian moons through the planet's magnetic field. The signals are plainly audible on the HF band any time Jupiter is above the horizon, day or night. However, in order to eliminate the possibility of both local and terrestrial skip signals from interfering, Jupiter DXers prefer to listen during the hours after 2100 [or whenever the maximum usable frequency (MUF) drops significantly below 18 MHz] and local sunup. Listen to the

Copyright 2001 - the McGraw-Hill Companies

amateur 17- or 15-m bands. If you hear zilch activity, then it's a good bet that the MUF has dropped enough to make listening worthwhile. Even during the day, however, it is possible to hear jovian signals, but differentiating them from other signals or solar noise is difficult.

So what do you need to DX Jupiter? It would help to have a decent radio receiver that works well over the range 18 to 24 MHz. Some cheaper radio receivers are not desirable, but most modern communications receivers are fine. The radio signals are rising and falling "swooshing" noises. The chances of receiving a signal from Jupiter are about 1 in 6, according to several radio astronomers.

The antenna can be a simple dipole cut for the middle of the 18- to 24-MHz band, which happens to be a 15-m amateur radio band antenna. The antenna should be installed in the normal manner for any dipole, except that the wire must run east-west in order to pick up the southerly rising planet.

Figure 21-1 shows a broadband dipole that covers the entire frequency region of interest (18 to 24 MHz) by paralleling three different dipoles: one cut for 18 MHz, one cut for 21 MHz, and one cut for 24 MHz. The dimensions are

A	19.5 ft	24 MHz
B	22.3 ft	21 MHz
C	26 ft	18 MHz

There are several approaches to making this type of antenna. One is to use three-conductor wire, and cut the wires to the lengths indicated above. Another is to use a homemade spacer to spread the wires apart.

Ring antenna

Another popular jovian radio antenna is the ring radiator, two versions of which are shown in Fig. 21-2. This antenna is made of a 5-ft-diameter loop of $\frac{1}{2}$-in-diameter soft-drawn copper plumbing pipe. The single-ended version is shown in Fig. 21-2A. In this antenna the loop is open-ended. The center conductor of the coaxial cable feedline is connected to one end of the ring radiator, while the coax shield is connected to the chicken wire ground plane. The balanced version (Fig. 21-2B) has an RF transformer (T_1) at the feedpoint.

The ring radiator antenna should have a bandpass preamplifier. The preamp is needed because of the low pickup of this kind of antenna, and that preamp should be mounted as close as possible to the antenna. The bandpass characteristic is obtained by filtering. The idea is to get rid of terrestrial signals in the adjacent bands. The typical preamplifier is broadbanded (which is what we want), so will pick up not only the desired signals but others as well (which is what we don't want). Even a 5-W CB transmitter a few blocks away can drive the preamplifier into saturation, so it's wise to eliminate the undesired signals before they get into the preamplifier. In the case of the single-ended amplifier, a single-ended preamplifier is used. But for the balanced version (Fig. 21-2B) a differential preamplifier is used.

The loop (both versions) is mounted about 7 or 8 in above a groundplane made of chicken wire, metal window screen, copper sheeting, or copper foil. The copper sheeting

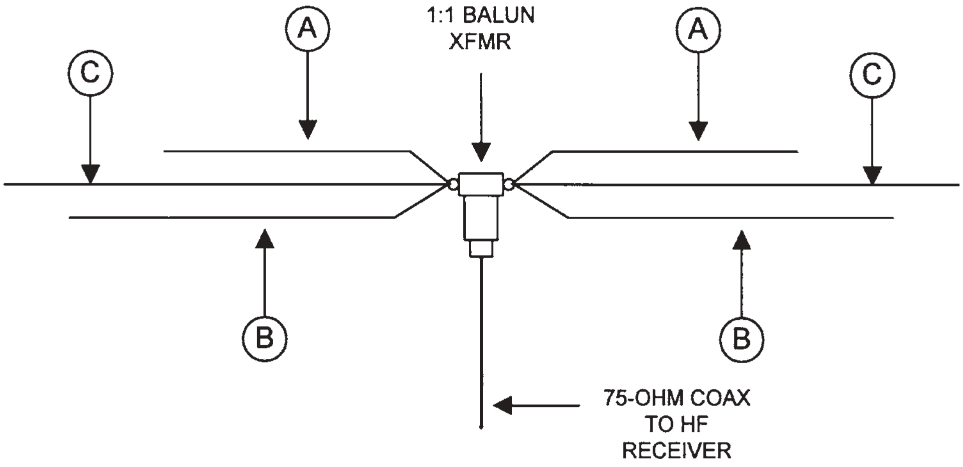
1:1 BALUN
XFMR
A
A
C
C
B
B
75-OHM COAX
TO HF
RECEIVER

21-1 Wideband dipole such as might be used for Jupiter reception.

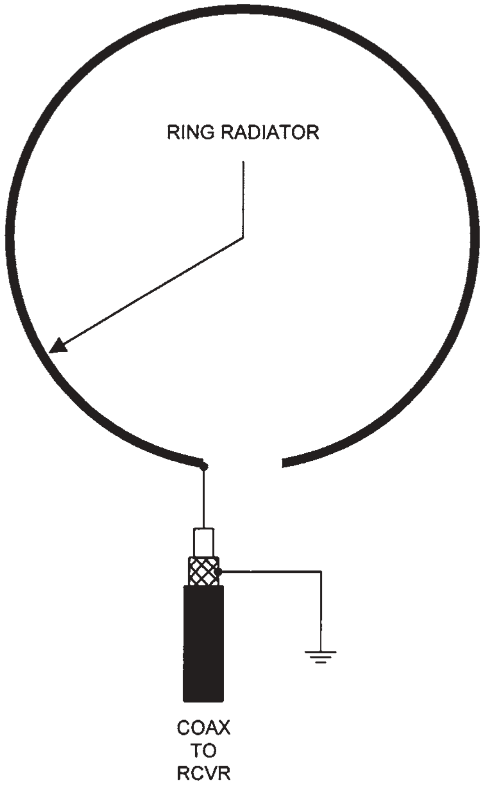
RING RADIATOR
COAX
TO
RCVR

21-2A Ring radiator: single-ended.

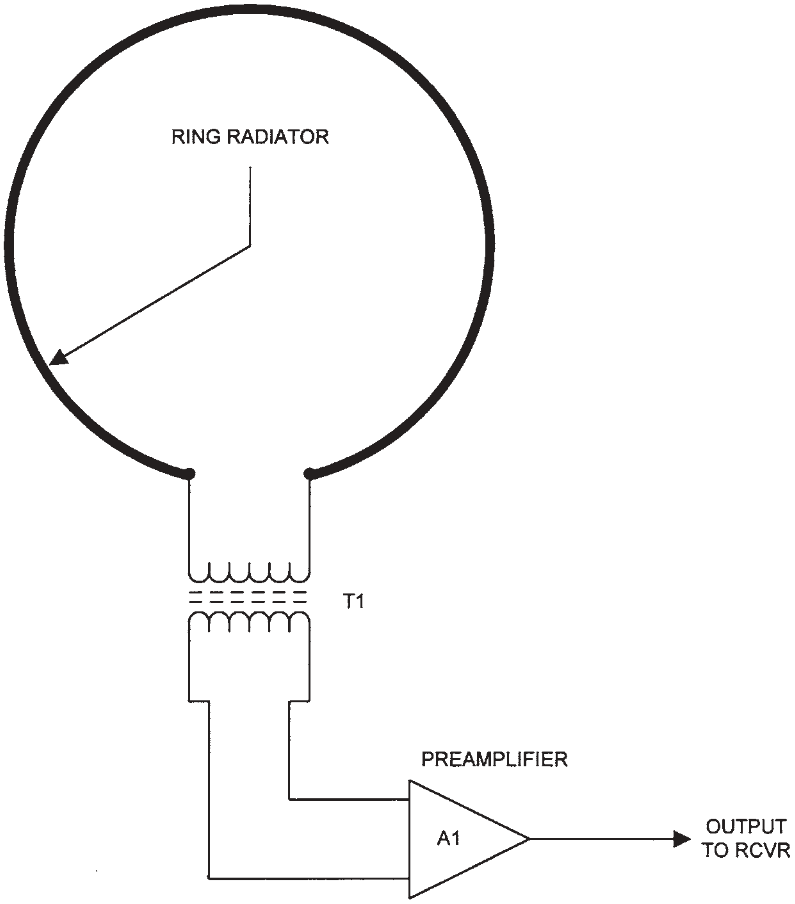

21-2B Ring radiator: balanced.

or foil is best, but costs a lot of money, and turns ugly green after a couple weeks in the elements. If you use screen, make sure that it is a metallic screen. Some window and porch screening material is made of synthetic materials that are insulators.

Figure 21-3A is a mechanical side view of the ring radiator antenna, while Fig. 21-3B is a side view. The antenna is mounted above the screen with insulators. These can be made of wood, plastic, or any other material. The frame holding the groundplane screen (Fig. 21-3B) can be made from 1 × 2-in lumber. Note that the frame has interior crosspieces to support the antenna, as well as the outer perimeter. The larger outer perimeter is needed because the screen groundplane should extend beyond the diameter of the radiator element by about 10 to 15 percent.

DDRR

The directional discontinuity ring radiator (DDRR) antenna is shown in Fig. 21-4A, while a side view showing the mounting scheme is shown in Fig. 21-4B. It is typically

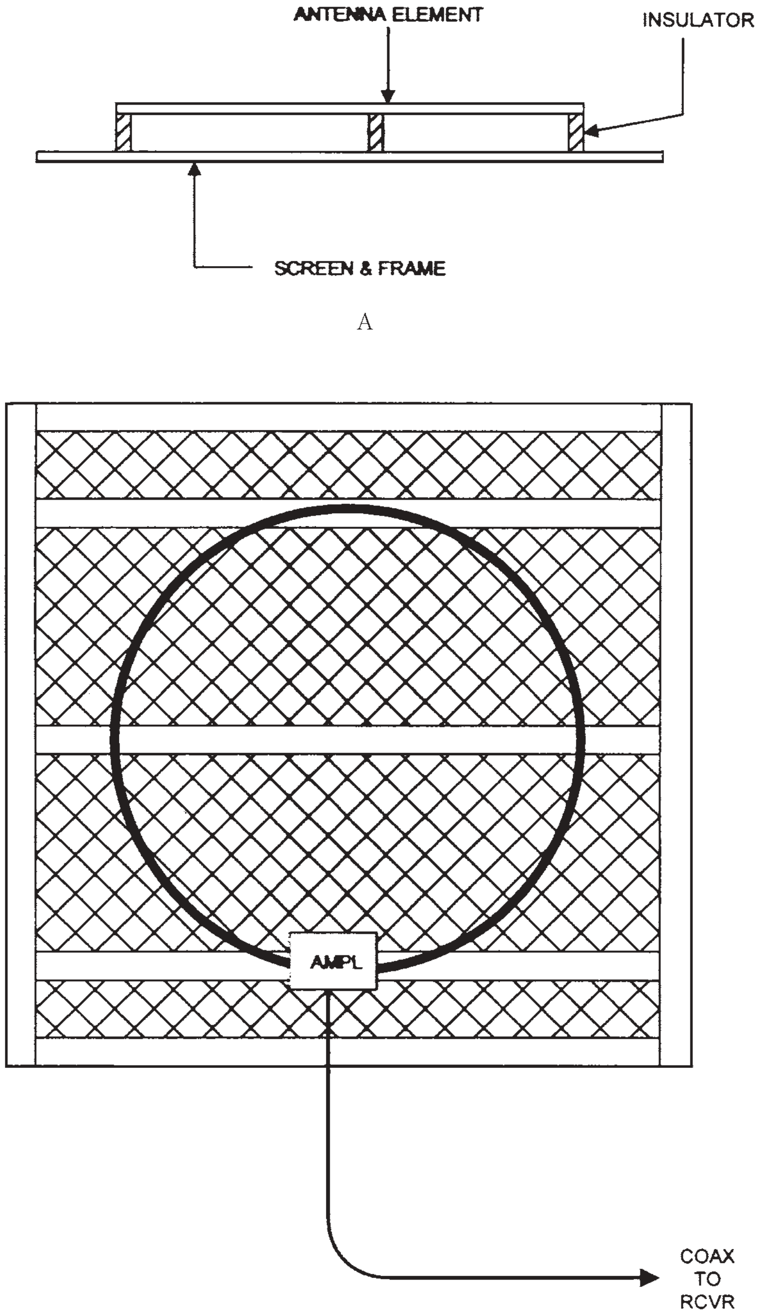
ANTENNA ELEMENT
INSULATOR
SCREEN & FRAME
AMPL
COAX
TO
RCVR

A

B

21-3 Ring radiator mounted over a ground screen.

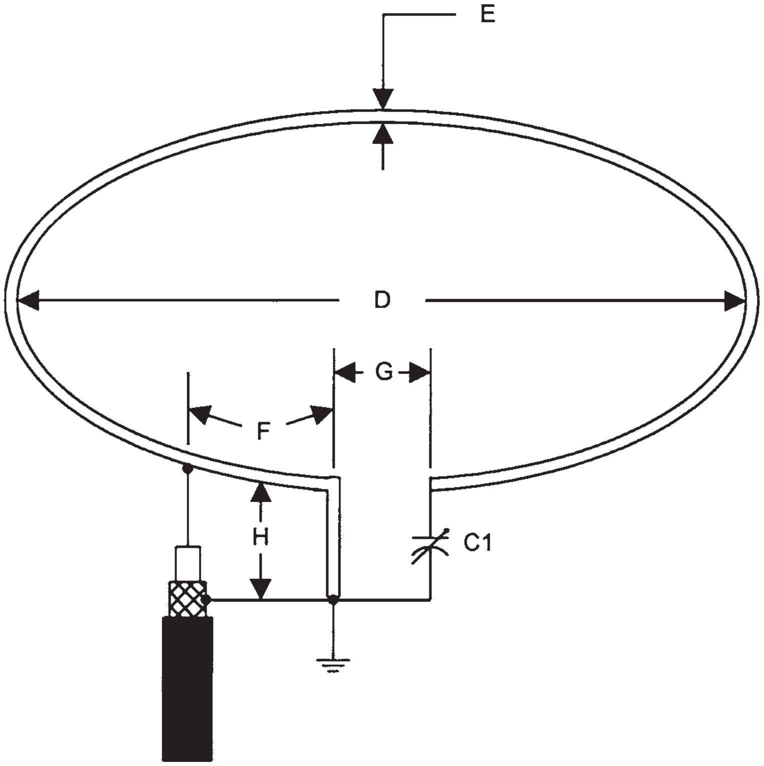

21-4A DDRR antenna.

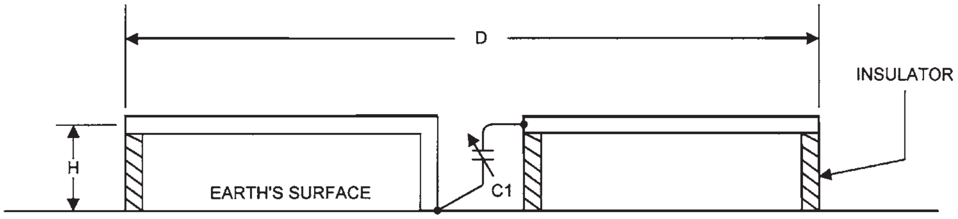

21-4B DDRR antenna: side view.

mounted about 1 ft off the ground ($H=12$ in). The DDRR consists of two sections, one vertical and one horizontal. The short vertical section has a length equal to the height H of the antenna above the ground. One end of the vertical segment is grounded. The horizontal section is a loop with a diameter D.

The conductor diameter E is at least 0.5 in at 28 MHz, and 4 in at 4 MHz. Because of the loop, some people call this the hula hoop antenna. One author recommends using a 2-in automobile exhaust pipe bent into the correct shape by an auto muffler dealer. The far end of the loop is connected to ground through a small-value tuning capacitor C_1. The actual value of C_1 is found experimentally and is used to resonate the antenna to a particular operating frequency.

The feedline of the DDRR antenna is coaxial cable connected such that the shield is grounded at the bottom end of the vertical section. The center conductor is connected to the ring radiator a distance F from the vertical section. The length F is determined by the impedance that must be matched. The radiation resistance is approximated by

$$R_T = \frac{2620H^2}{\lambda_2} \tag{21.1}$$

where

R_T is the radiation resistance, in ohms
H is the height of the antenna off the ground
λ is the wavelength

(Both H and λ are in the same units.)

The approximate values for the various dimensions of the DDRR are given below in general terms, with examples in Table 21-1:

D	0.078λ
H	$0.11D$
F	$0.25H$
E	0.5 to 4 in
G	See Table 21-1

Table 21-1. Examples of dimensions for DDRR

	Band (MHz)							
Dimension	**1.8**	**4**	**7.5**	**15**	**22**	**30**	**50**	**150**
G (in)	16	7	5	3	2.5	2	1.5	1
C_1 (pF)	150	100	75	35	15	12	10	6
F (in)	12	6	6	1.5	1.5	1	1	0.5
H (in)	48	24	11	6	4.75	3	1.5	1
D (ft)	36	18	9	4.5	3.33	2.33	1.4	6
E (in)	5	4	2	1	0.75	0.75	0.5	0.25

The construction details of the DDRR are so similar to those of the ring radiator that the same diagram can be used (see Fig. 21-3 again).

The normal attitude of the DDRR for communications is horizontal. However, for Jupiter radio reception, the antenna groundplane screen can be tilted to face Jupiter's point in the sky.

Helical antennas

The helical antenna (Fig. 21-5) provides moderately wide bandwidth and circular polarization. Because of the polarization some people find the helical antenna to be particularly well suited to radio astronomy reception. The antenna (of diameter D) will have a circumference C of 0.75λ to 1.3λ. The pitch of the helix (S) is the axial length of one turn, while the overall length $L=NS$ (where N is the number of turns). The ratio S/C should be 0.22λ to 0.28λ. At least three turns are needed to produce axial-mode main lobe maxima.

The diameter or edge of the groundplane G should be on the order of 0.8λ to 1.1λ if circular and/or square, respectively. The offset between the groundplane and the first turn of the helix is 0.12λ.

The approximate gain of the helical antenna is found from

$$\text{Gain} = 11.8+10\log(C^2NS) \qquad \text{dBi} \tag{21.2}$$

The pitch angle ϕ and turn length γ for the helical antenna are given by

$$\phi = \tan^{-1}\left(\frac{\pi D}{S}\right) \tag{21.3}$$

and

$$\gamma = \sqrt{(\pi D)^2+S^2} \tag{21.4}$$

The beamwidth of the helical antenna is

$$\Theta = \frac{\mathrm{K}}{\sqrt{\mathrm{NS}}} \tag{21.5}$$

where

Θ is the beamwidth, in degrees
N is the number of turns
S is the pitch, in wavelengths
C is the circumference, in wavelengths
K is 52 for the −3-dB beamwidth and 115 for the beamwidth to the first null in the pattern

The short section between the helix and the groundplane is terminated in a coaxial connector, allowing the antenna to be fed from the rear of the groundplane. The feedpoint impedance is approximately 140 Ω.

Multiple helical antennas

Stacking helical antennas allows a radiation pattern that is much cleaner than the normal one-antenna radiation pattern. It also provides a good way to obtain high gain with only a few turns in each helix. If two helixes are stacked, then the gain will be the same as for an antenna that is twice the length of each element, while for four stacked antennas the gain is the same as for a single antenna 4 times as long. Figure 21-6 shows a side view of the stacked helixes.

The feed system for stacked helixes is a little more complex than for a single helix. Figure 21-7A shows an end view of a set of four stacked helical antennas. Tapered lines (TL) are used to carry signal from each element and the coaxial connector (B). In this case, the coaxial connector is a feed-through “barrel” SO-239 device at the center of the groundplane (B). A side view of the tapered line system

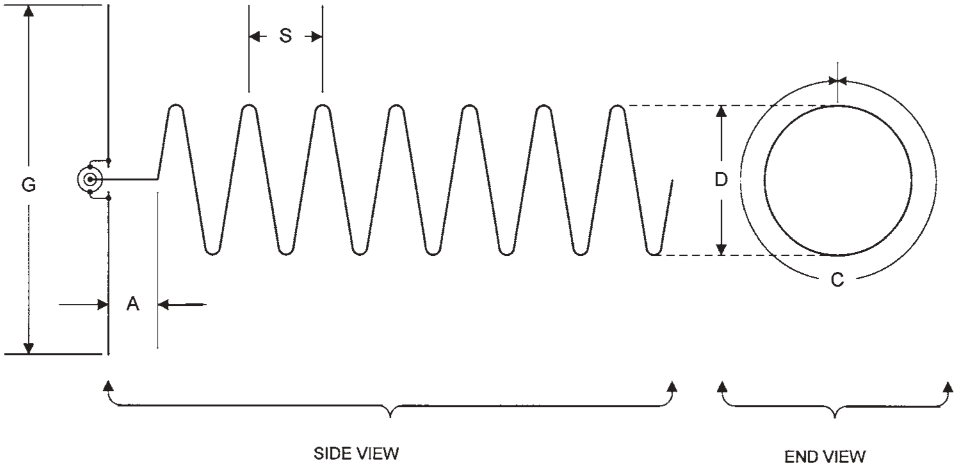
S
G
D
C
A
SIDE VIEW
END VIEW

21-5 Helical antenna.

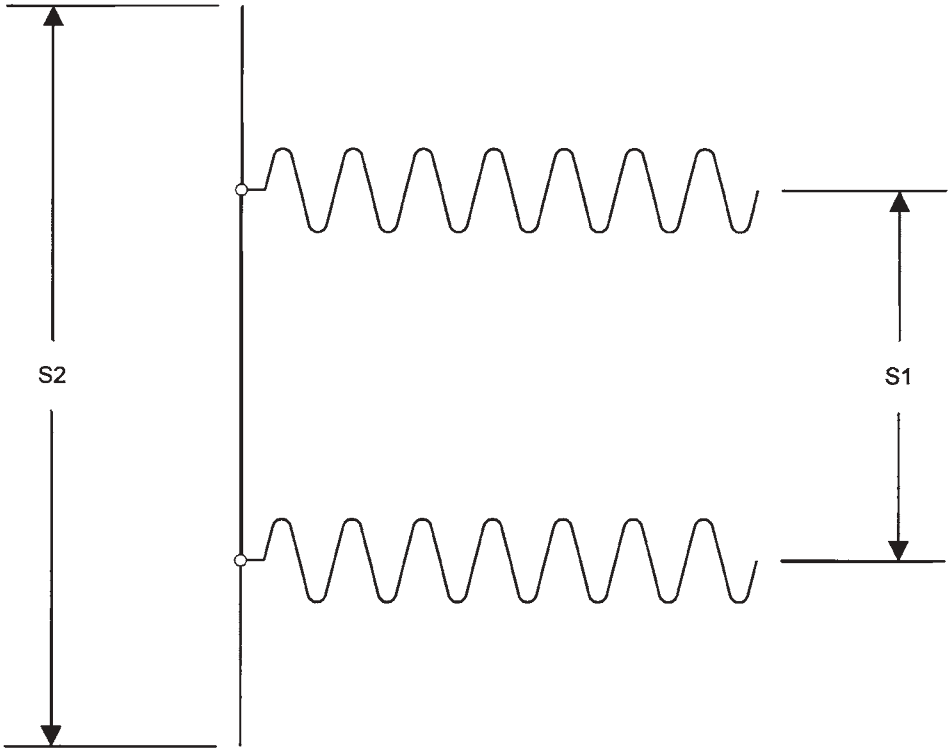

21-6 Dual/quad helical antenna.

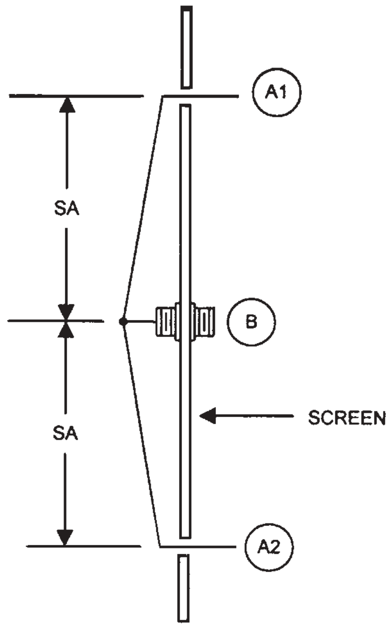
A1
SA
B
SA
SCREEN
A2

B

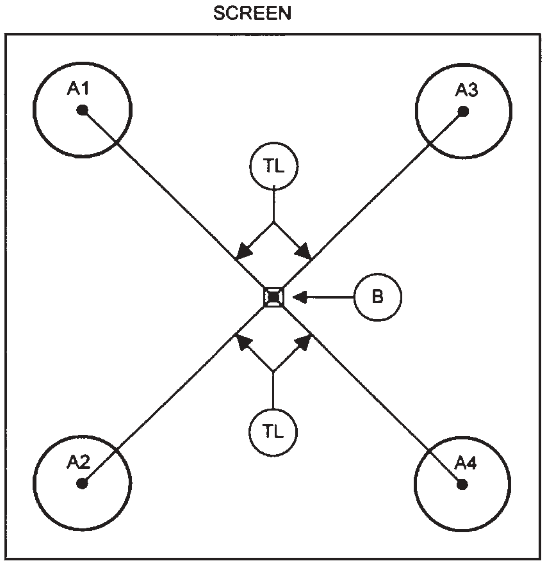
SCREEN
A1
A3
TL
B
TL
A2
A4

A

21-7 Frontal view of quad helical antenna.

is shown in Fig. 21-7B. The length of the tapered lines is 1.06λ, while the center-to-center spacing between the helical elements is 1.5λ. The length of each side of the groundplane is 2.5λ. In the case of Fig. 21-7, the antenna is fed from the front of the groundplane.

Interferometer antennas

The resolution of an antenna is set by its dimensions a relative to the wavelength λ of the received signal. Better resolution can be achieved by increasing the size of the antenna, but that is not always the best solution. Figure 21-8 shows a *summation interferometer array.* Two antennas with aperture a are spaced S wavelengths apart. The radiation pattern is a fringe pattern (Fig. 21-9). This pattern consists of a series of maxima and nulls. The resolution angle ϕ to the first null is

$$\phi = \frac{S+a}{57.3} \qquad \textbf{[21.6]}$$

The interferometer can be improved with additional antennas in the array. Professional radio astronomers use very wide baseline antennas. With modern communications it is possible to link radio telescopes on different continents to make the widest possible baseline.

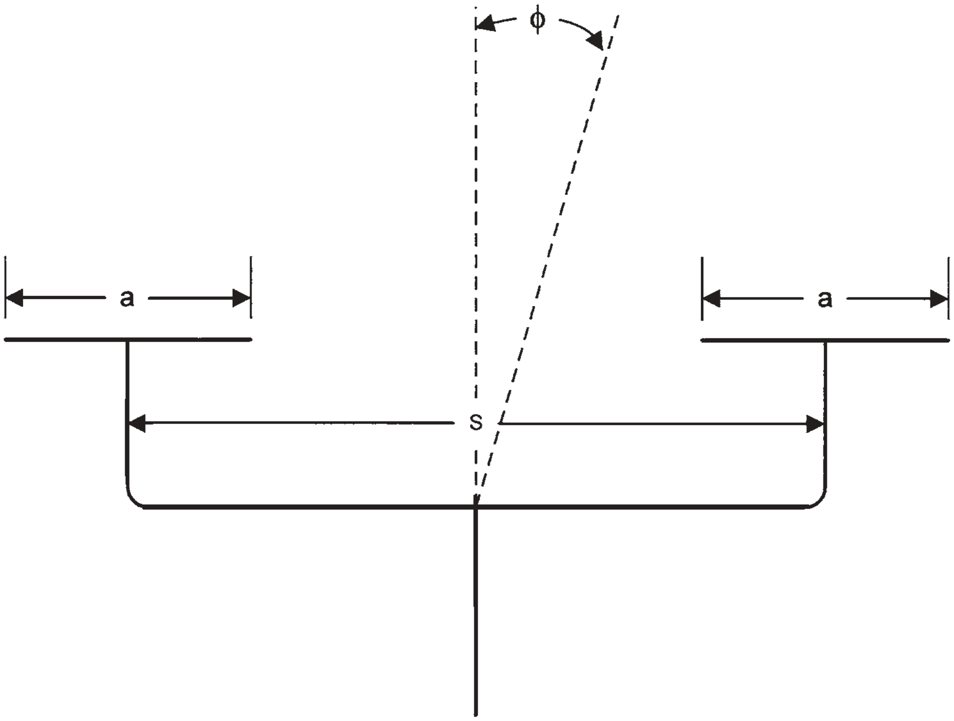

21-8 Interferometer array.

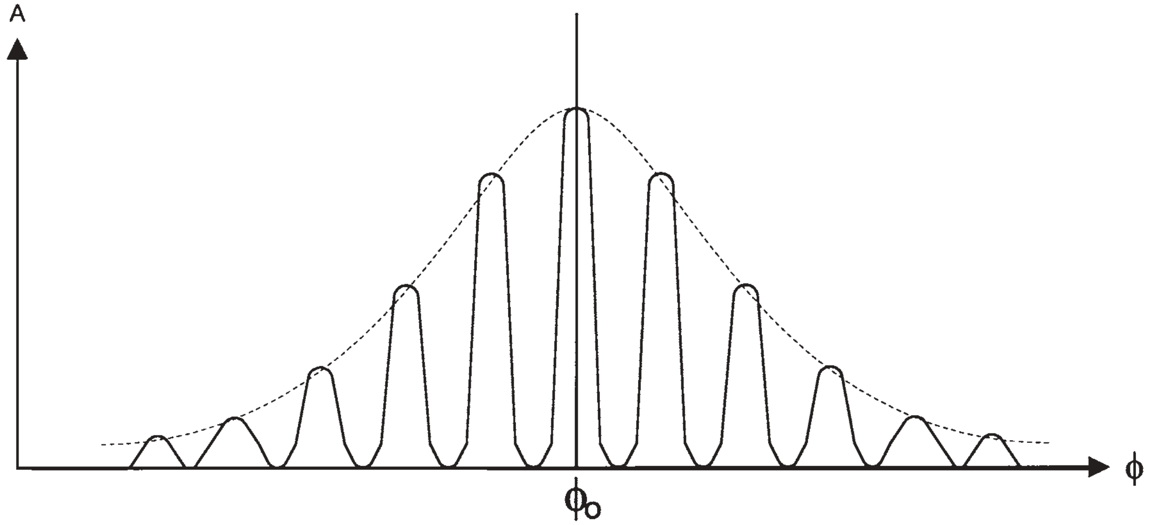

21-9 Interferometer pattern.

22
CHAPTER

Adjusting, installing, and troubleshooting antennas and transmission lines

ANTENNA AND TRANSMISSION LINE MEASUREMENTS SHOULD BE MADE WHEN THE ANTENNA is first installed and periodically thereafter. If some difference in operation is noted, the same measurements should be repeated. Many antenna measurements are difficult to make with any degree of accuracy. There are some things about antennas that can and should be measured, however, regardless of the difficulty. For example, VSWR and the resonant frequency of the antenna are readily accessible. It's also possible to measure the impedance of the antenna feedpoint. You can measure the VSWR either with a special VSWR meter (often built into transmitters or antenna tuning units), or by using an RF wattmeter.

By frequency stepping through the band and testing the VSWR at various frequencies, one can draw a VSWR curve (Fig. 22-1) that shows how the antenna performs across the band. The resonant frequency is the point where the VSWR dips to a minimum (which may or may not be the much sought-after 1:1).

You can use the resonant frequency to figure out whether the antenna is too long (resonant frequency lower than the hoped-for design frequency), or too short (resonant frequency above the design frequency). Figure 22-2 shows all three situations. Curve A represents the desired or ideal curve that is centered on the desired frequency F_D. If the antenna is too long, however, the resonant frequency will be shifted downward towards F_1, and curve B will be observed. Similarly, if the antenna is too short, curve C will be found as the resonant frequency shifts upband to F_2.

But resonant frequency and VSWR curves are not the entire story because they don't tell us anything about the impedance presented by the antenna. One cannot get the VSWR to be 1:1 unless the antenna impedance and transmission line impedance are the same. For example, a dipole has a nominal textbook impedance of 73 Ω,

Copyright 2001 - the McGraw-Hill Companies

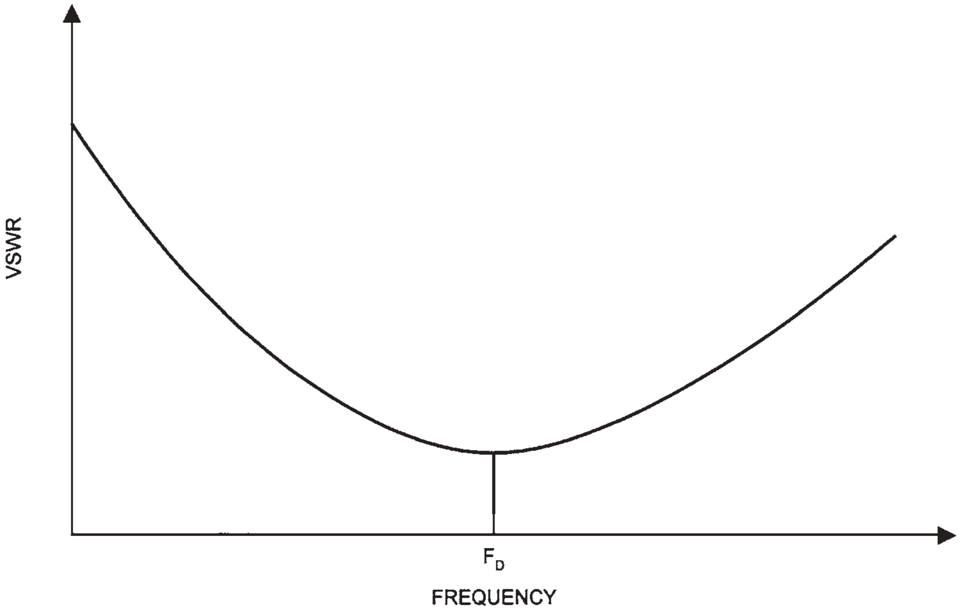

22-1 VSWR-vs-frequency plot.

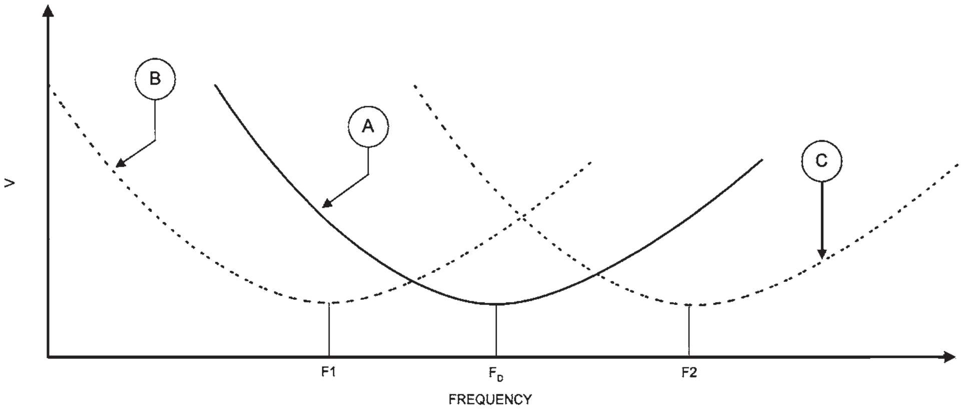

22-2 VSWR-vs-frequency plot for three frequencies.

so it makes a very good match to 75-Ω coaxial cable. But the actual impedance of a real dipole may vary from a few ohms to more than 100 Ω. For example, if your antenna exhibits a feedpoint impedance of (say) 25 Ω, using 75-Ω coaxial cable to feed it produces a VSWR=75/25=3:1. Not too great. Measuring the feedpoint impedance is therefore quite important to making the antenna work properly.

There are any number of instruments on the market that will aid in making antenna measurements. Some of them are quite reasonably priced (or can be built), while others are beyond the reach of all but the most ardent and well-endowed enthusiasts. In this chapter we will look at a newer breed of more universal instrument called the *SWR analyzer.*

VSWR analyzers

One basic premise in this chapter is that the instruments used for measuring VSWR must be accessible to people who don't have a commercial or ham operator's license, as well as those who do (we must remember our SWL and scanner friends). Some of the instruments used by hams meet that requirement, but a relatively new breed of instrument called the SWR analyzer provides a lot of capability to the SWL, scanner operator, and ham radio operator alike. It uses a low-power RF signal generator and some clever circuitry to measure the VSWR of the antenna. One model also measures the feedpoint resistance.

The MFJ-259 VSWR analyzer (MFJ Enterprises, Inc., P.O. Box 494, Mississippi State, MS, 39762) is shown in Fig. 22-3. This instrument combines a VSWR analyzer with a digital frequency counter, and operates over the range 1.8 to 170 MHz. A band switch is set to the desired band, and then the *tune* control is set to the desired frequency. The meter will then read the VSWR at the design frequency. Alternatively, you can adjust the tune control until the minimum VSWR is found. This frequency is the actual, versus the desired, resonant frequency of the antenna. The front panel of the MFJ-259 has two meters, *SWR* and *RESISTANCE.* The SWR meter is calibrated up to 3:1, with a little uncalibrated scale to indicate higher SWRs.

The resistance meter is calibrated from 0 to 500 Ω, which is consistent with the SWR range. Two controls on the front panel are *tune* and *frequency* (a band switch). The MFJ-259 has a digital frequency meter to measure the operating frequency of the internal oscillator. This frequency counter can also be used to measure the frequency of external signal sources. (*Do not* connect the counter to the output of a transmitter; the instrument will be destroyed.) The top end of the MFJ-259 has a number of controls and connectors. An SO-239 UHF-style coaxial connector is provided for the antenna connection. A BNC coaxial connector is provided to apply external signal to the frequency counter, while a push button *input* switch is available to switch the counter from internal to external signal sources. Another push button switch is used to set the gate timing of the counter (a red LED on the front panel blinks every time the gate is triggered). The tuning is from 1.8 MHz to 174 MHz, while the counter will measure up to 200 MHz.

The MFJ-259 will work from an external 12-Vdc source, or from an internal battery pack consisting of eight size-AA standard cells. MFJ recommends that either alkaline or rechargeable batteries, rather than ordinary zinc-carbon cells, be used in order to reduce the possibility of leakage that can damage the instrument (this is good practice in all battery-powered instruments).

Unlike many lesser SWR meters, this instrument is not fooled by antennas that have impedances consisting of both resistance and reactance elements. An example in the manual demonstrates an impedance of $25+j25\ \Omega$ (i.e., R is 25 Ω and reactance X is also 25 Ω). When connected to a 50-Ω load one might be tempted to think the VSWR is 1:1, and some cheaper meters will so indicate. But the actual SWR is 2.6:1, which is what the MFJ-259 will read.

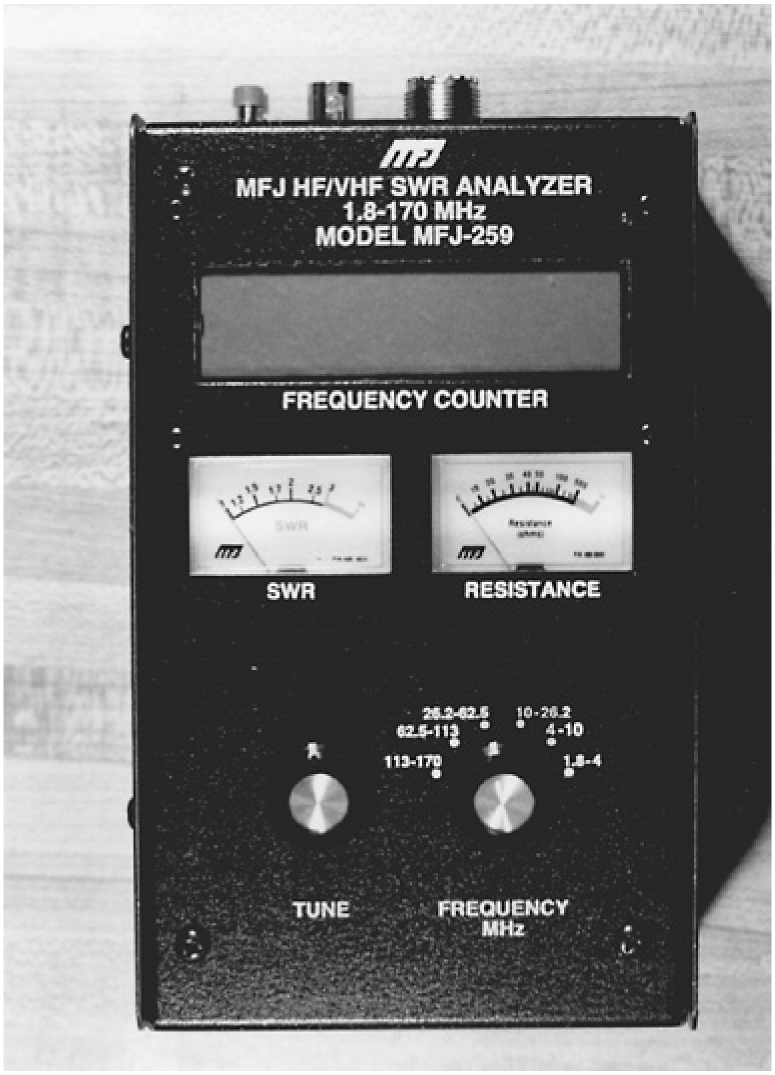
MFJ
MFJ HF/VHF SWR ANALYZER
1.8-170 MHz
MODEL MFJ-259
FREQUENCY COUNTER
SWR
RESISTANCE
26.2-62.5
10-26.2
62.5-113
4-10
113-170
1.8-4
TUNE
FREQUENCY
MHz

22-3 MFJ-259 VSWR analyzer.

The resistance measurement assumes a resistive load (i.e., the measurement is made at the antenna's resonant frequency), and is referenced to 50 Ω. The VSWR and resistance measurements should be consistent with each other. If the VSWR is 2:1, then the resistance should be either 100 Ω (100/50 = 2:1) or 25 Ω (50/25 = 2:1). If the resistance is not consistent with the VSWR reading, then you should assume that the impedance has a significant reactive component and take steps to tune it out.

In addition to antenna measurements, the MFJ-259 is equipped to measure a wide variety of other things. It will measure the velocity factor of transmission line, help in tuning or adjusting matching stubs or matching networks, measure capacitance or inductance, and determine resonant frequency of *LC* networks.

Doping out coaxial cable

When you install an antenna, or do a bit of preventive maintenance, or find the antenna is not working properly, one thing to check is the transmission line. Two basic measurements are popular. Figure 22-4 shows how to make ohmic checks. The cable consists of an inner conductor and an outer conductor (shield). With terminals A and B open there should not be any resistive path across the input terminals (as shown). If a high resistance is seen, then there might be some contamination in the system, or the insulation has failed, permitting a current path. A low resistance indicates a short circuit. If the cable has been cut, or an object passed through, or the connector is messed up, then a short can result.

If terminals A-B are shorted together, a low resistance should be noted. If not, then it is likely that the center conductor is open. Of course, if connectors are on the line, either the shield or center conductor could be at fault.

The other issue is loss of the transmission line. Cable losses get higher as cables age, so low loss on installation does not guarantee low loss later on. Losses in new cables run from 0.2 dB/100 ft at 4 MHz. At 30 MHz the losses are 0.7 dB/100 ft to 2.6 dB/100 ft. At VHF/UHF frequencies the losses rise considerably.

Figure 22-5 shows a test setup for coaxial cable. It is necessary to know the length of the piece of coax under test. For ease of calculations a 100-ft section

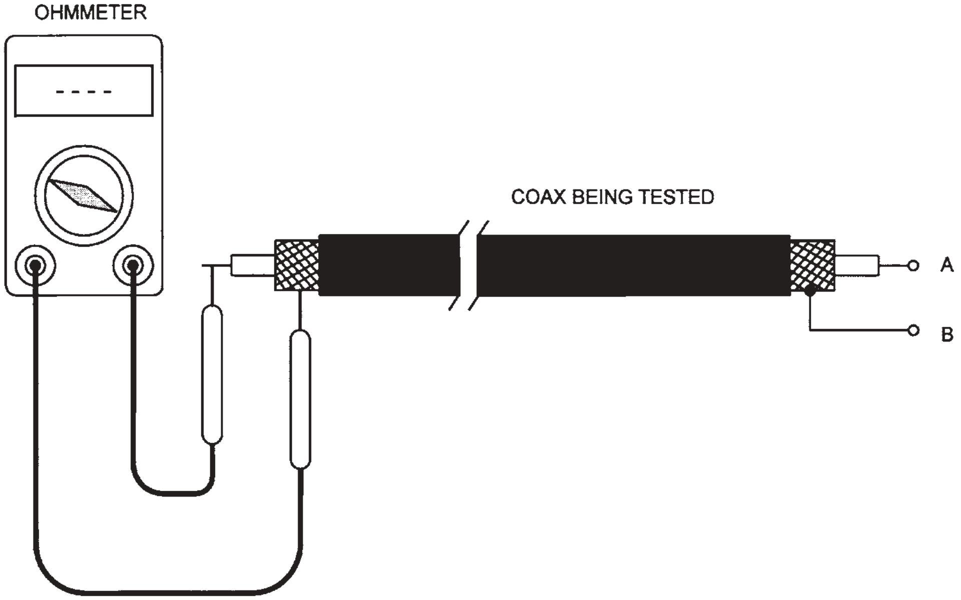

22-4 Ohmmeter used to test coax.

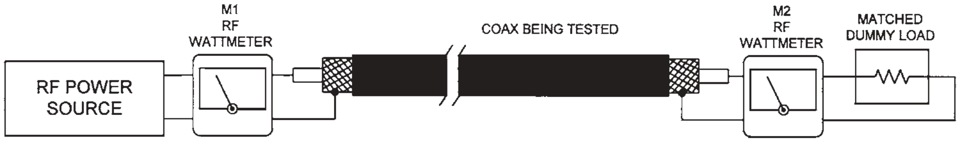
M1
RF
WATTMETER
COAX BEING TESTED
M2
RF
WATTMETER
MATCHED
DUMMY LOAD
RF POWER
SOURCE

22-5 Attenuation test for coax.

should be used. Two RF power meters (M_1 and M_2) are used to measure the input power to the line (M_1) and the delivered power (M_2). The difference in power determines the loss. The loss is

$$\text{Loss}_{\text{dB}} = 10 \log \left(\frac{P_{M2}}{P_{M1}} \right) \tag{22.1}$$

where

Loss is the loss of the cable, in dB
P_{M1} is the power reading on M_1
P_{M2} is the power reading on M_2

If the cable length is 100 ft, then you already have the loss in dB/100 ft. But if length L is anything other than 100 ft, then you need to make the following calculation:

$$\text{Loss}_{\text{dB/100ft}} = \frac{\text{Loss}_{\text{dB}}}{\text{L}_{\text{ft}}} \times 100 \tag{22.2}$$

23
CHAPTER

Antennas for radio direction finding (RDF)

RADIO DIRECTION FINDING IS THE ART AND PRACTICE OF EITHER LOCATING YOURSELF OR A radio station by using a directional radio antenna and receiver. When the FCC wants to locate an illegal station that is transmitting, they will use radio direction finders to triangulate the position. If they find the bearing from two stations they will locate the station at the intersection. However, there is a fair degree of ambiguity in the measurement. As a result, radio direction finders typically use three or more (hence "*tri*angulate") sites. Each receiving site that can find the bearing to the station reduces the overall error.

At one time aviators and seafarers relied on radio direction finding. It is said that the Japanese air fleet that attacked Pearl Harbor, Hawaii, on December 7, 1941 homed in on a Honolulu AM radio station. During the 1950s and early 1960s AM radios came with two little circled triangle marks at the 640-kHz and 1040-kHz points on the dial. These were the CONALRAD frequencies that you could tune to in case of a nuclear attack (right before you kissed your butt good-bye, I suspect). All other radio stations were off the air except the CONALRAD stations. The enemy was prevented from using these frequencies for RDF because the system used several stations that transmitted in a rapidly rotating pattern. No one station was on the air long enough to allow a "fix." The result was a wavering sound to the CONALRAD station (which we heard during tests) that would confuse any enemy who tried to DX his way into our cities with nuclear weapons.

Radio direction finders based on the AM broadcast band (BCB) looked a bit like Fig. 23-1. A receiver with an S meter (which measures signal strength) is equipped with a rotatable ferrite loopstick antenna to form the RDF unit. A degree scale around the perimeter of the antenna base could be oriented toward north so that the bearing could be read.

Loopstick antennas have a figure-8 reception pattern (Fig. 23-2A) with the maxima parallel to the loopstick rod and the minima off the ends of the rod. When the antenna is pointed at the signal, maximum reception strength is achieved. Unfortunately, the maxima are so broad that it is virtually impossible to find the true point on the compass dial where the signal peaks. The peak is too shallow for that

Copyright 2001 - the McGraw-Hill Companies

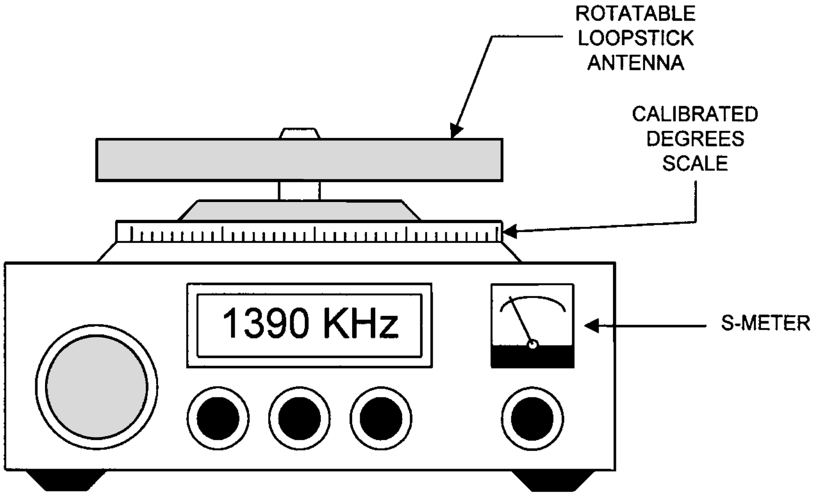
ROTATABLE LOOPSTICK ANTENNA
CALIBRATED DEGREES SCALE
1390 KHz
S-METER

23-1 Radio direction finding (RDF) receiver using a loopstick antenna.

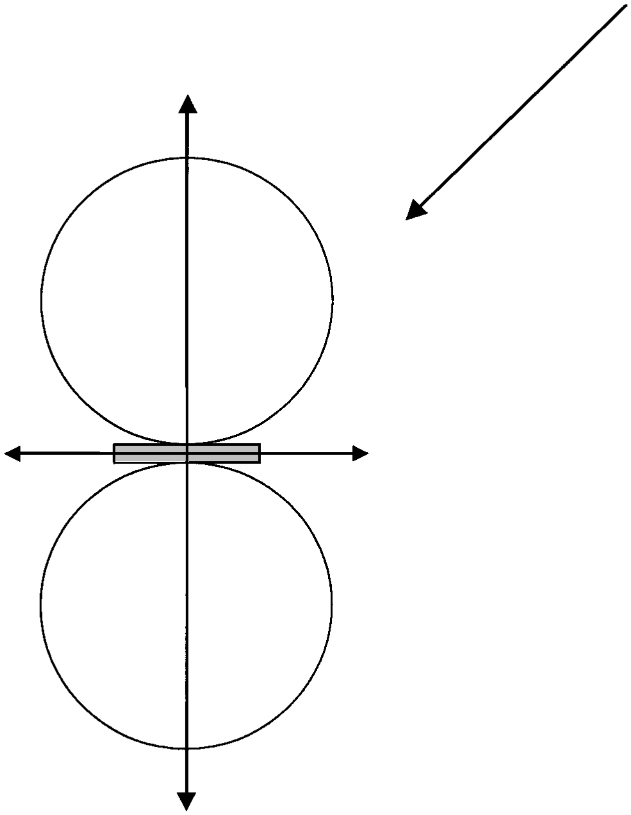

23-2A Pattern of loopstick antenna: not oriented towards signal.

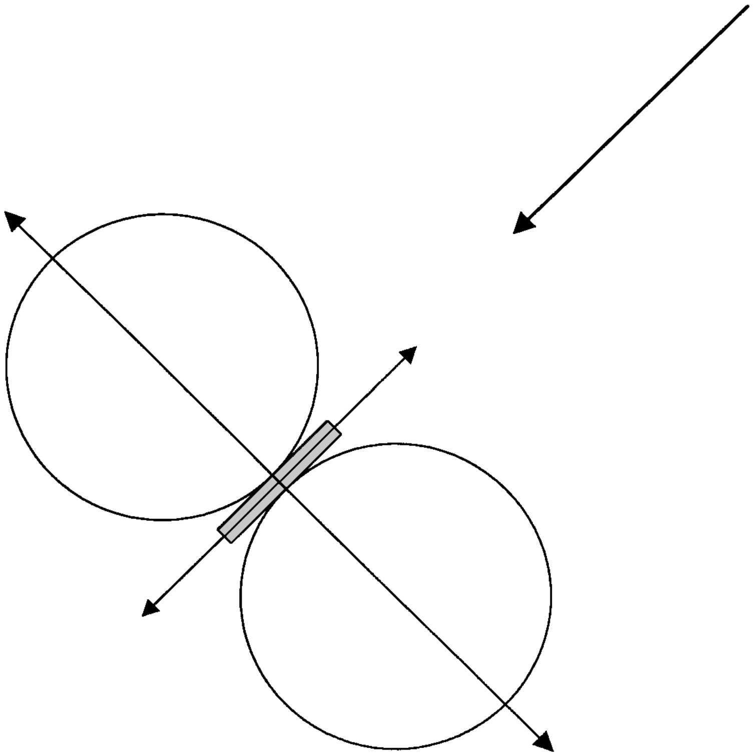

23-2B Pattern of loopstick antenna: oriented towards signal.

purpose. Fortunately, the minima are very sharp. You can get a good fix on the direction of the signal by pointing the minima toward the station. This point is found by rotating the antenna until the audio goes to zip or the S meter dips to a minimum (Fig. 23-2B).

The loopstick is a really neat way to do RDF—except for one little problem: The darn thing is bidirectional. There are two minima because, after all, the pattern is a figure 8. You will get exactly the same response from placing either minima in the direction of the station. As a result, the unassisted loopstick can only show you a line along which the radio station is located, but can't tell you which direction it is. Sometimes this doesn't matter. If you know the station is in a certain city, and that you are generally south of the city, and can distinguish the general direction from other clues, then the line of minima of the loopstick will refine that information. A compass helps, of course. Shortly we will take a look at an impromptu radio direction finder using a portable radio.

The solution to the ambiguity problem is to add a sense antenna to the loopstick (Fig. 23-3). The sense antenna is an omnidirectional vertical whip, and its signal is combined with that of the loopstick in an *RC* phasing circuit. When the two patterns are combined, the resultant pattern will resemble Fig. 23-4. This pattern is called a *cardioid* because of the heart shape it exhibits. This pattern has only one null, so it resolves the ambiguity of the loopstick used alone.

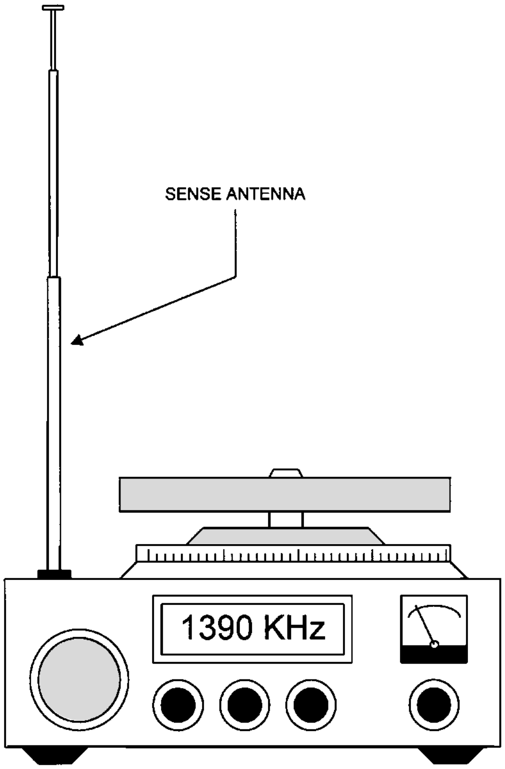

23-3 Addition of sense antenna overcomes directional ambiguity of the loopstick.

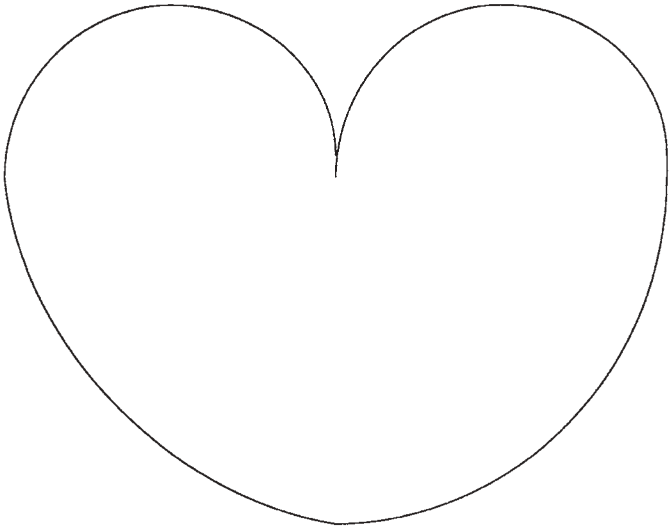

23-4 Cardioid pattern of sense-plus-loopstick antennas.

Field improvisation

Let's suppose you are out in the woods trekking around the habitat of lions, tigers, and bears (plus a rattlesnake or two for good measure). Normally you find your way with a compass, a Geological Survey 7.5-min topological map, and a Global Positioning System (GPS) receiver. Those little GPS marvels can give you real good latitude and longitude indication. But what happens if it breaks or a bear eats it? The answer to your direction finding problem might be the little portable AM BCB radio (Fig. 23-5) that you brought along for company.

Open the back of the radio and find the loopstick antenna. You will need to know which axis it lies along. In the radio shown in Fig. 23-5 the loopstick is along the top of the radio, from left to right. In other radios it is vertical, from top to bottom. Once you know the direction, you can tune in a known AM station and orient the radio until you find a null. Your compass can give you the bearing. If you know the approximate location of the station, then you can reverse the compass direction from it and mark the line on the topo map. Of course, it's still a bidirectional indication, so all you know is the line along which you are lost.

But then you tune into a different station in a different city (or at least wide enough from the line to the other station to make a difference) and take another reading. Your approximate location is where the two lines cross. Take a third, fourth, and fifth reading and you will home in pretty tight. If you were smart enough to plan ahead, you will have selected candidate stations in advance and located their latitude and longitude. Alternatively, you would have bought the topo map that covers their location as well as where you want to hike, and from those maps you can find the latitudes and longitudes of the distant stations.

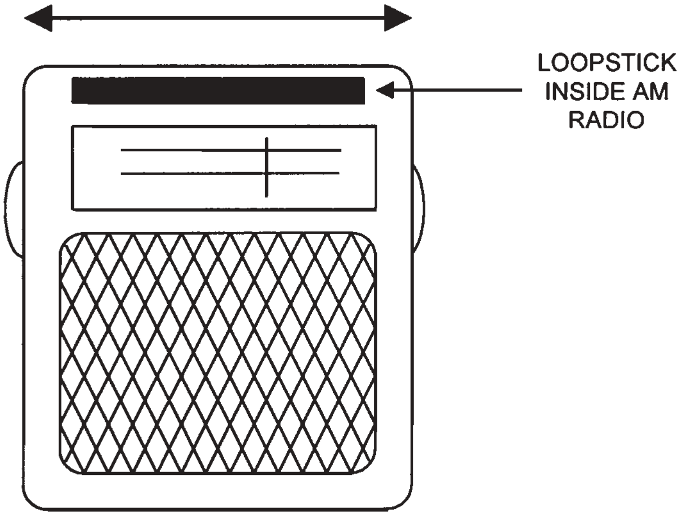

23-5 AM portable radio. Dark bar represents position and direction of internal loopstick.

Regular loop antennas

Regular wire loop antennas (Fig. 23-6) are also used for radio direction finding. In fact, in some cases the regular loop is preferred over the loopstick. The regular loop antenna may be square (as shown), circular, or any other regular "*n*-gon" (e.g., hexagon), although for practical reasons the square is easier to build. The loop has pretty decent inductance even with only a few turns. One loop I built was 24 in square (A in Fig. 23-6) and had, if I recall correctly, about 10 turns of wire spaced over a 1-in width (B in Fig. 23-6). It resonated to the AM BCB with a standard 365-pF "broadcast" variable capacitor.

When you use a regular loop antenna be aware that the antenna has a figure-8 pattern like the loopstick, but it is oriented 90° out of phase with the loopstick antenna. In the regular loop the minima (nulls) are perpendicular to the plane of the loop, while the maxima are off the sides. In Fig. 23-6 the minima are in and out of the page, while the maxima are left and right (or top and bottom).

Fox hunting

An activity popular with ham radio operators in the 1960s was "fox hunting." A fiendishly clever ham (the "fox") would go hide with a mobile or portable transmitter (usually on either 10 or 75 m). The "hunters" would then RDF the fox's brief transmissions and try to locate the transmitter. If you could locate the antenna, you located the station (according to the most common set of rules). Sometimes the hunt got a little wild as hunters raced each other (a no-no) in the final stretch.

One friend of mine had an interesting experience. He drove a 1949 Plymouth that was painted hideous dark green with fire-engine-red hub caps. It looked just plain awful (especially since the paint looked like it was put on with a whisk broom!). He used a Gonset converter linked to the AM BCB receiver installed in the car. The antenna was a wire loop on the end of a broom handle, similar to Fig. 23-7. The antenna was mounted about where you would expect the left-side rear view mirror to be located. He could reach his hand out the window and rotate the antenna while listening for the minima on the receiver. Worked rather well, and he won some fox hunts.

The fox hunts in his area were usually held on Saturday or Sunday morning, and after it was all over the whole crew would get to a restaurant for brunch or beer or something. One morning, however, my buddy with the hideous green car was racing down a residential street to be the first to the transmitter. Others were in the area, and the end game was in motion. Unfortunately, a little old lady called the police complaining about a "…nut racing up and down the street in a strange car waving a cross out the window." Breakfast had to wait that morning, I bet.

Shortwave and AM BCB "skip" RDF

Radio direction finding is most accurate over relatively short distances. If you can use the ground wave, then all the better (which is what you use during daylight hours for nearly all AM BCB stations). Skip rolls in on the AM BCB after local

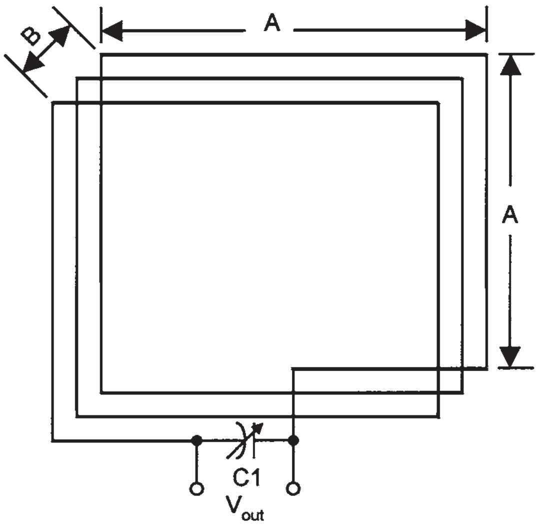

23-6 Wire loop antenna.

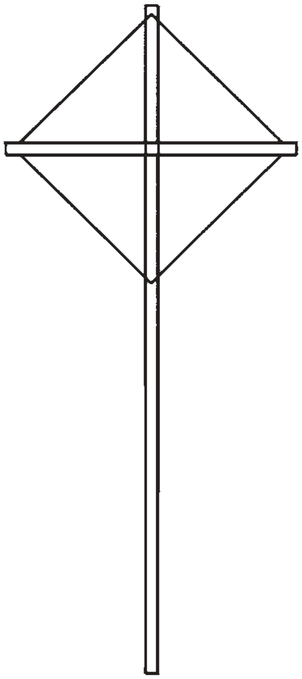

23-7 "…waving a cross out the window" antenna.

sundown, so you can hear all manner of stations all up and down the dial. You can RDF distant stations. Some rather interesting sites have turned up when SWL DXers RDF'ed some of the infamous "numbers stations." Those stations transmit either CW or voice numbers groups, and are believed to be sending messages to spies around the world. RDFing by DXers has located some of these transmitters at, well, "interesting" sites.

Unfortunately, there are some problems with skip RDFing. When we look at propagation drawings of skip in textbooks we usually see a one plane view. The curvature of the earth shows, as does the transmitter and receiver site. A "pencil beam" radio signal travels at some angle up the ionosphere, where it is "reflected" (actually, it's a refraction phenomenon but looks like reflection to an observer on the surface) back to earth. We can tell from the drawing that the angle of incidence equals the angle of reflection, just like they told us in high school science classes. Oops! The real world is not so neat and crisp, however.

In the real world the wave might encounter a different ionization density along its path of travel and therefore be deflected from its original path. It might return to earth at a location offset from the direction it appears. If we look at the true bearing from the receiver site to the transmitter site, and then note the azimuthal angle of arrival of the signal, we note that something is amiss.

Actual reflection also causes some problems, especially when RDFing a station in the high end of the HF band or the VHF/UHF bands. Radio waves will reflect from geological features such as mountains, as well as man-made structures (e.g., buildings). If the reflection is strong enough it might appear to be the real signal, and cause a severe error in RDFing. Be wary of RDF results when the "skip is in."

Sense antenna circuit

Figure 23-8 shows a method for summing together the signals from an RDF antenna (such as a loop) and a sense antenna. The two terminals of the loop are connected to the primary of an RF transformer. This primary (L_{1A}) is center-tapped and the center tap is grounded. The secondary of the transformer (L_{1B}) is resonated by a variable capacitor C_1. The dots on the transformer coils indicate the 90° phase points.

The top of L_{1B} is connected to the sense circuit, and to the receiver antenna input. The phasing control is a potentiometer (R_1). The value of this pot is usually 10 to 100 kΩ, with 25 kΩ being a commonly seen value. Switch S_1 is used to take the sense antenna out of the circuit. The reason for this switch is that the nulls of the loop or loopstick are typically a lot deeper than the null on the cardioid pattern. The null is first located with the switch open. When the switch is closed you can tell by the receiver S meter whether or not the correct null was used. If not, then reverse the direction of the antenna and try again.

Adcock antennas

The Adcock antenna has been around since 1919 when it was patented by F. Adcock. Figure 23-9 shows the basic Adcock RDF array. This antenna consists of two center-fed nonresonant (but identical) vertical radiators. Each side of each element is at least 0.1λ long, but need not be resonant (which means the antenna can be used

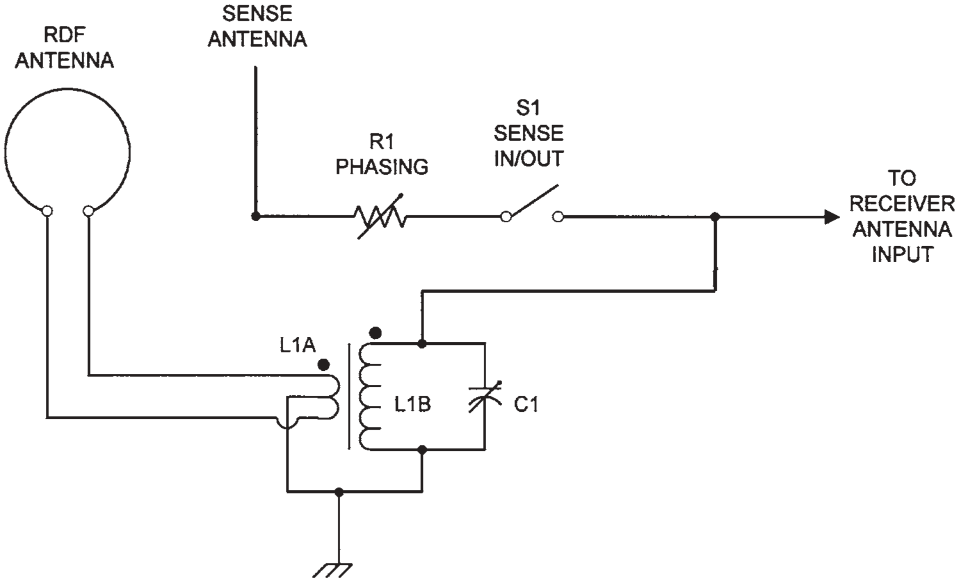

23-8 Sense circuit for a loop/sense antenna RDF.

over a wide band). The elements are spaced from 0.1λ to 0.75λ, although the example shown here is spaced 0.125λ.

The Adcock antenna is vertically polarized, so it will respond to the vertically polarized wave very much like loop or phased array antennas of similar size. The horizontally polarized wave, however, affects all elements the same, so the currents are essentially cancelled, resulting in no pattern. This characteristic makes the Adcock antenna suitable for high-frequency shortwave RDFing.

The pattern for an Adcock antenna is shown in Fig. 23-10. This pattern was generated using the NecWin Basic for Windows program (Chap. 17). The example antenna is a 10-MHz (30-m band) Adcock that uses 1.455-m elements (total 2.91 m on each side), spaced 4 m apart. The pattern is a traditional figure-8 with deep nulls at 0° and 180°. The antenna can be rotated to find a null in the same manner as a loop.

Watson-Watt Adcock array

Figure 23-11 shows the Watson-Watt Adcock RDF array. It consists of two Adcock arrays arranged orthogonally to each other. It is common practice to arrange one Adcock in the east-west direction and the other in the north-south direction. These are fed to identical receivers that are controlled by a common local oscillator (LO). The outputs of the receivers are balanced, and are used to drive the vertical and horizontal plates of a cathode-ray oscilloscope (CRO). Figure 23-12 shows the patterns achieved by signals of various phases arriving at the Watson-Watt array. The patterns of Figs. 23-12A and 23-12B are made from signals 180° out of phase, while the signal of Fig. 23-12C has a 90° phase difference.

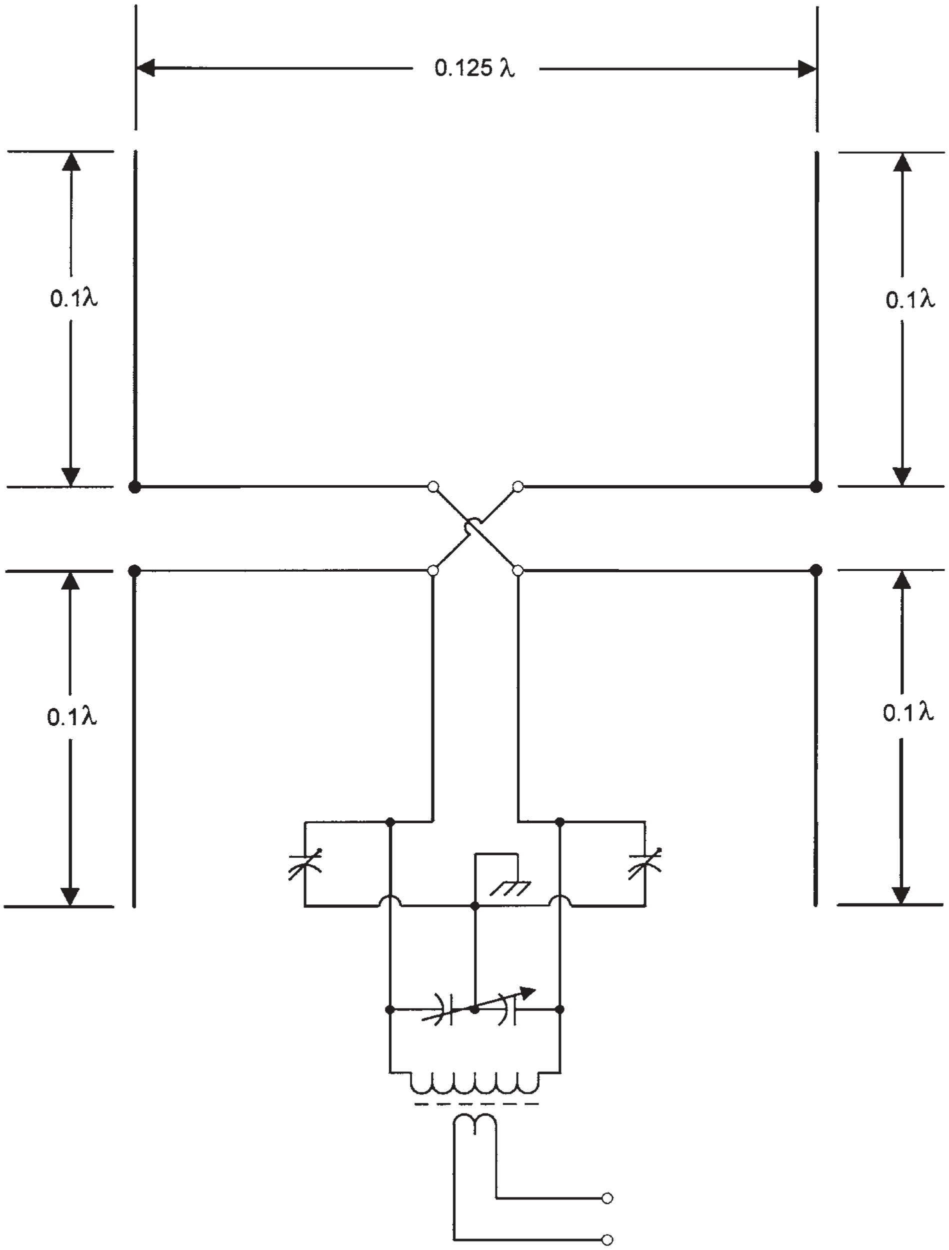
0.125 λ
0.1λ
0.1λ
0.1λ
0.1λ

23-9 Adcock array RDF antenna.

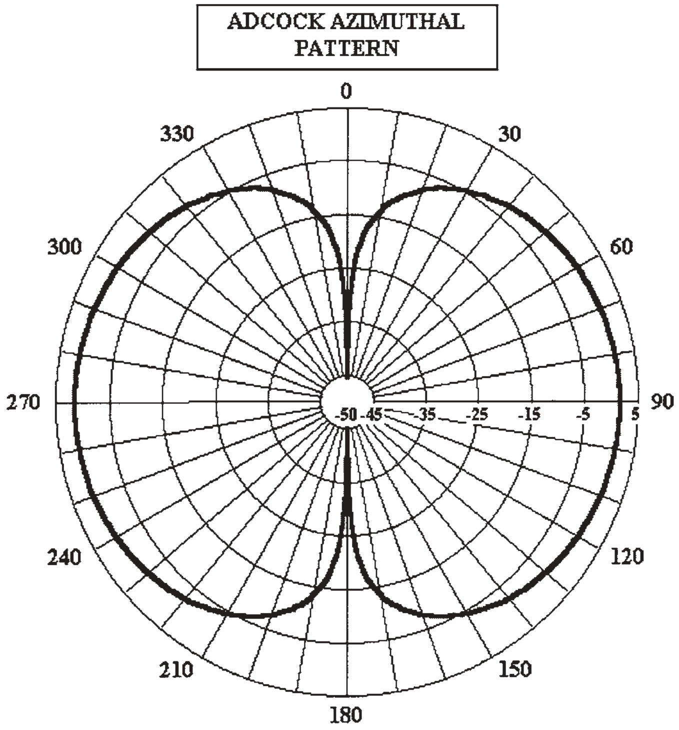

23-10 Adcock pattern.

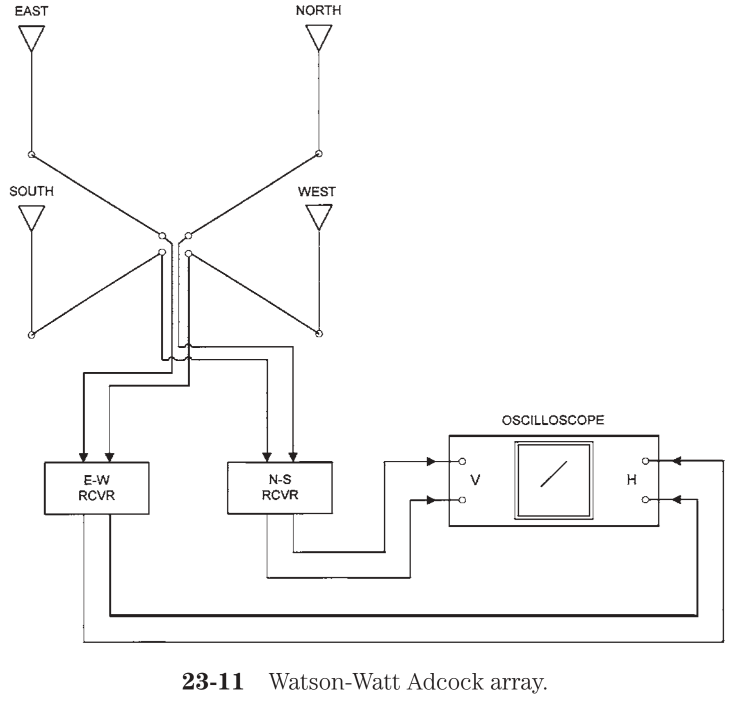
EAST
NORTH
SOUTH
WEST
OSCILLOSCOPE
E-W RCVR
N-S RCVR
V
H

23-11 Watson-Watt Adcock array.

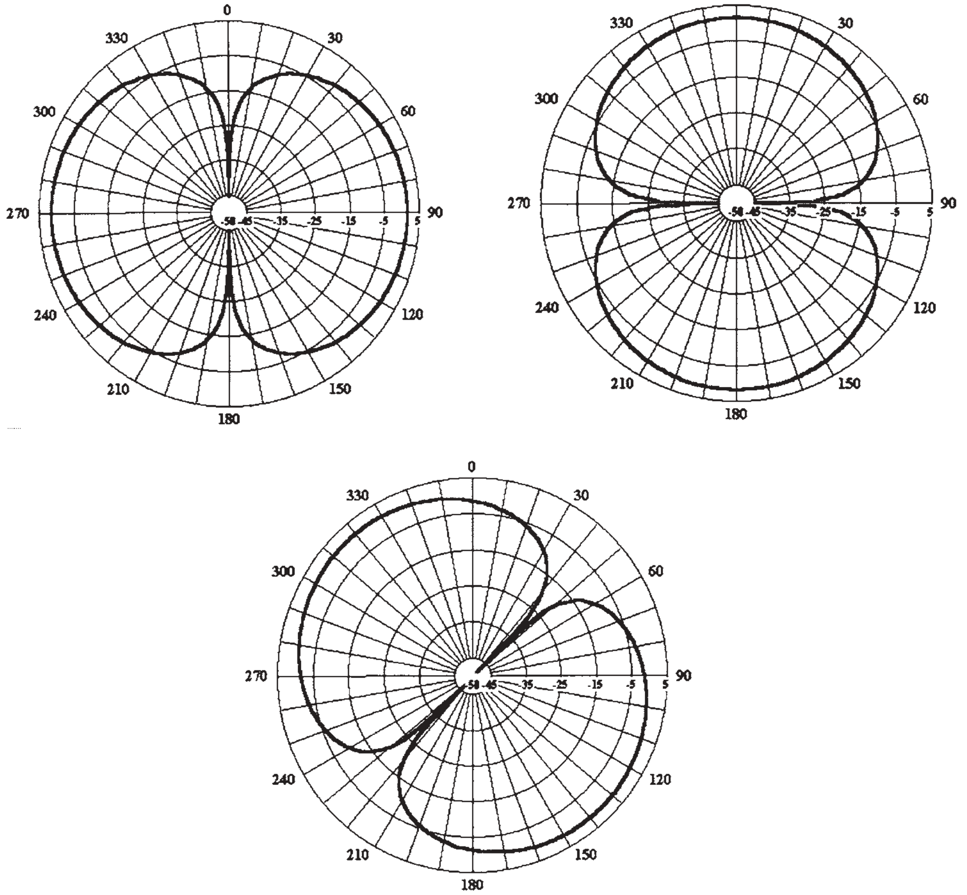

23-12 Patterns for different signals on the Watson-Watt array.

Doppler RDF antennas

Figure 23-13 shows the basic concept of a Doppler RDF antenna. The Doppler effect was discovered in the nineteenth century. A practical example of the Doppler effect is seen when a wailing ambulance siren is approaching you, and then passes you and heads away from you. The wailing pitch will rise as the sound source approaches, and then fall as the sound source recedes in the opposite direction.

In a radio system, when the antenna and signal source move with respect to each other, a Doppler shift is generated. This shift is proportional to the relative speed difference. The Doppler RDF antenna of Fig. 23-13 uses a rotating antenna. The signal approaches from a single direction, so there will be a predictable Doppler shift at any point on the circular path of the antenna. The magnitude of the phase shift is

$$S = \frac{R\omega F_c}{c} \qquad \textbf{[23.1]}$$

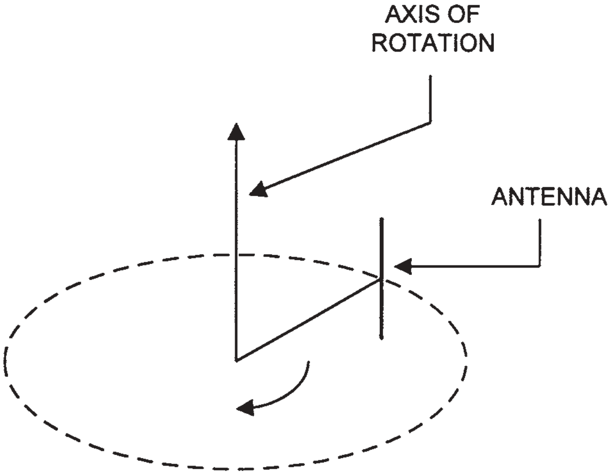

23-13 Doppler antenna.

where

S is the Doppler shift, in hertz
R is the radius of rotation, in meters
ω is the angular velocity of the antenna, in radians per second
F_c is the carrier frequency of the incoming signal, in hertz
c is the velocity of light (3×10^8 m/s)

In theory this antenna works nicely, but in practice there are problems. One of the big problems is getting a large enough Doppler shift to easily measure. Unfortunately, the rotational speed required of the antenna is very high—too high for practical use. However, the effect can be simulated by using a number of antennas, arranged in a circle, that are sequentially scanned. The result is a piecewise approximation of the effect seen when the antenna is rotated at high speed.

Wullenweber array

One of the problems associated with small RDF antennas is that they have such a small aperture that relatively large distortions of their pattern result from even small anomalies. Follow that? What it means is that the pattern is all messed up by small defects. If you build a wide-aperture direction finder (WADF), however, you can average the signals from a large number of antenna elements distributed over a large-circumference circle. The Wullenweber array (Fig. 23-14) is such an antenna. It consists of a circle of vertical elements. In the HF band the circle can be 500 to 2000 ft in diameter.

A goniometer rotor spins inside the ring to produce an output that will indicate the direction of arrival of the signal as a function of the position of the goniometer. The theoretical resolution of the Wullenweber array is on the order of 0.1°, although practical resolutions of about 2.8° are commonly seen.

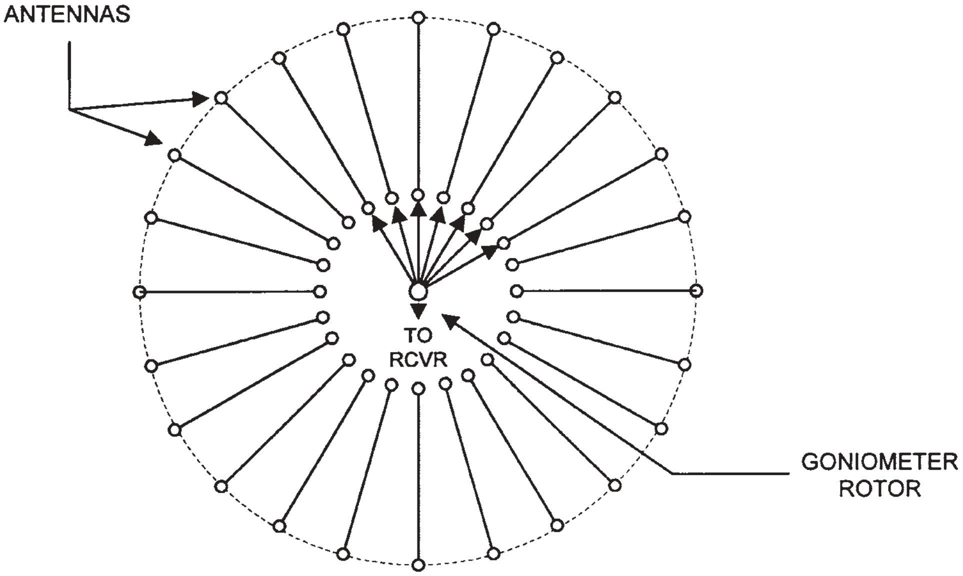

23-14 Wullenweber RDF array.

Time difference of arrival (TDOA) array

If you erect two antennas with distance *d* apart, then arriving signals can be detected by examining the time-of-arrival difference. Figure 23-15A shows an example signal. If the advancing wavefront is parallel to the line between the antennas, then it will arrive at both antennas at the same time. The TDOA is zero in that case. But if the signal arrives at an angle (as in Fig. 23-15A), it will arrive at one antenna first. From the difference between the time of arrival at the two antennas we can discern the direction of arrival.

There is an ambiguity in the basic TDOA array in that the combined output will be the same for conjugate angles, i.e. at the same angle from opposite directions. This problem can be resolved by the system shown in Fig. 23-15B. The signals from ANT1 and ANT2 are designated V_1 and V_2, respectively. These signals are detected by receivers (RCVR1 and RCVR2), and are then threshold detected in order to prevent signal-to-noise problems from interfering with the operation. The outputs of the threshold detectors are used to trigger a sawtooth generator that controls the horizontal sweep on an oscilloscope.

The two signals are then delayed, and one is inverted. The reason for inverting one signal is to allow the operator to distinguish them on the CRT screen. The reason why this is necessary is the ambiguity. If the antennas in Fig. 23-15A are arrayed east to west, then the line perpendicular to the line between them is north to south.

If we designate north as 0°, then the signal shown arrives at an angle of 330°. That means it will arrive at ANT1 before it arrives at ANT2. A signal arriving from a bearing of 30° will produce the same output signal, even though it arrives at ANT2 before ANT1. All signals of bearing $0 \leq x < 180°$ will arrive at ANT2 first, while all

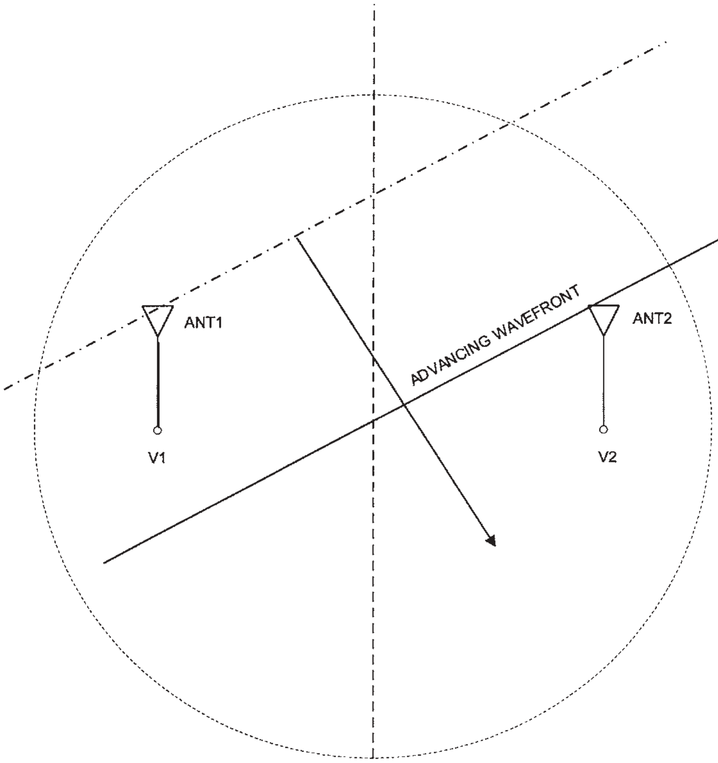
ANT1
ANT2
ADVANCING WAVEFRONT
V1
V2

23-15A Time difference of arrival (TDOA) RDF array.

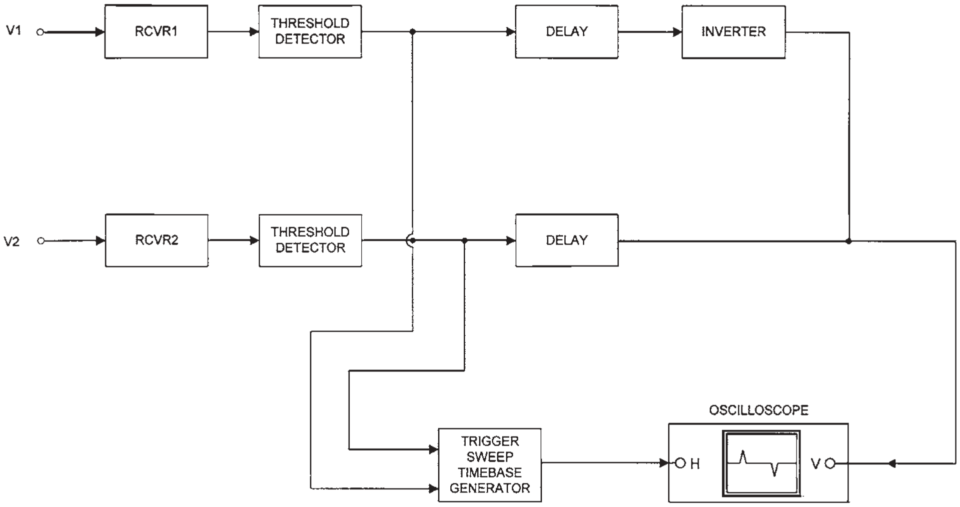
V1
RCVR1
THRESHOLD DETECTOR
DELAY
INVERTER
V2
RCVR2
THRESHOLD DETECTOR
DELAY
OSCILLOSCOPE
TRIGGER SWEEP TIMEBASE GENERATOR
H
V

23-15B Block diagram of typical TDOA circuit.

signals $180 \leq x < 360°$ will arrive at ANT1 first. Yet both will produce the same blip on the oscilloscope screen. The solution to discerning which of the two conjugate angles is intended is to invert the ANT1 signal. When this is done, the ANT1 signal falls below the baseline on the CRT screen, while the ANT2 signal is above the baseline. By noting the time difference between the pulses, and their relative position, we can determine the bearing of the arriving signal.

Switched-pattern RDF antennas

Suppose we have a unidirectional pattern such as the cardioid shown in Fig. 23-16. If we can rapidly switch the pattern back and forth between two directions that are 180° apart, then we can not only discern direction, but also we can tell whether an off-axis signal is left or right. This feature is especially useful for mobile and portable direction finding.

In Fig. 23-16 we have three different positions for a signal source. When the signal source is at point *A,* it will affect the pattern to the left more than the pattern to the right, so the meter will read left. If the signal source is at point *B,* on the other hand, the signal affects both pattern positions equally, so the meter reads zero. Finally, if the signal arrives from point *C,* it affects the right-hand pattern more than the left-hand pattern, so the meter reads right.

Figure 23-17 shows how such a system can be constructed. This system has been used by amateur radio operators with "rubber ducky" VHF antennas and a sin-

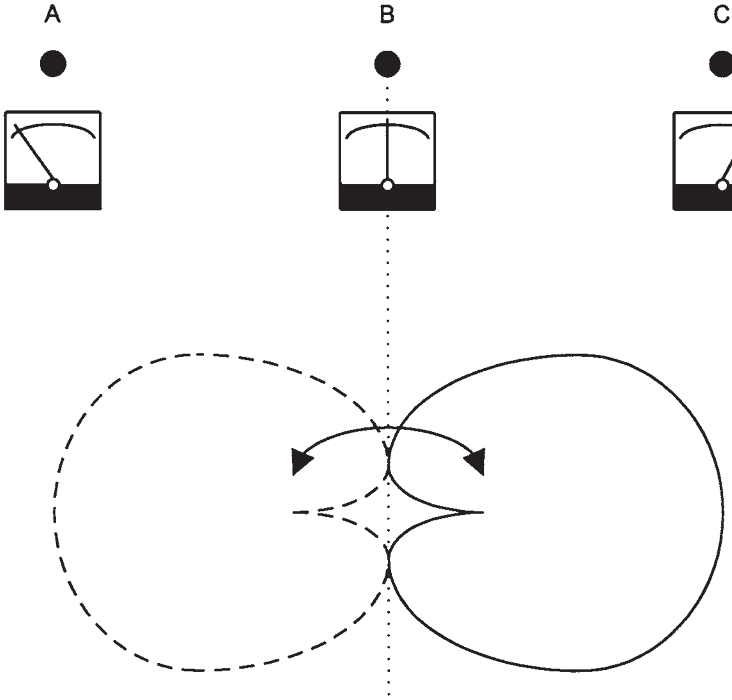

23-16 Switched-pattern RDF system.

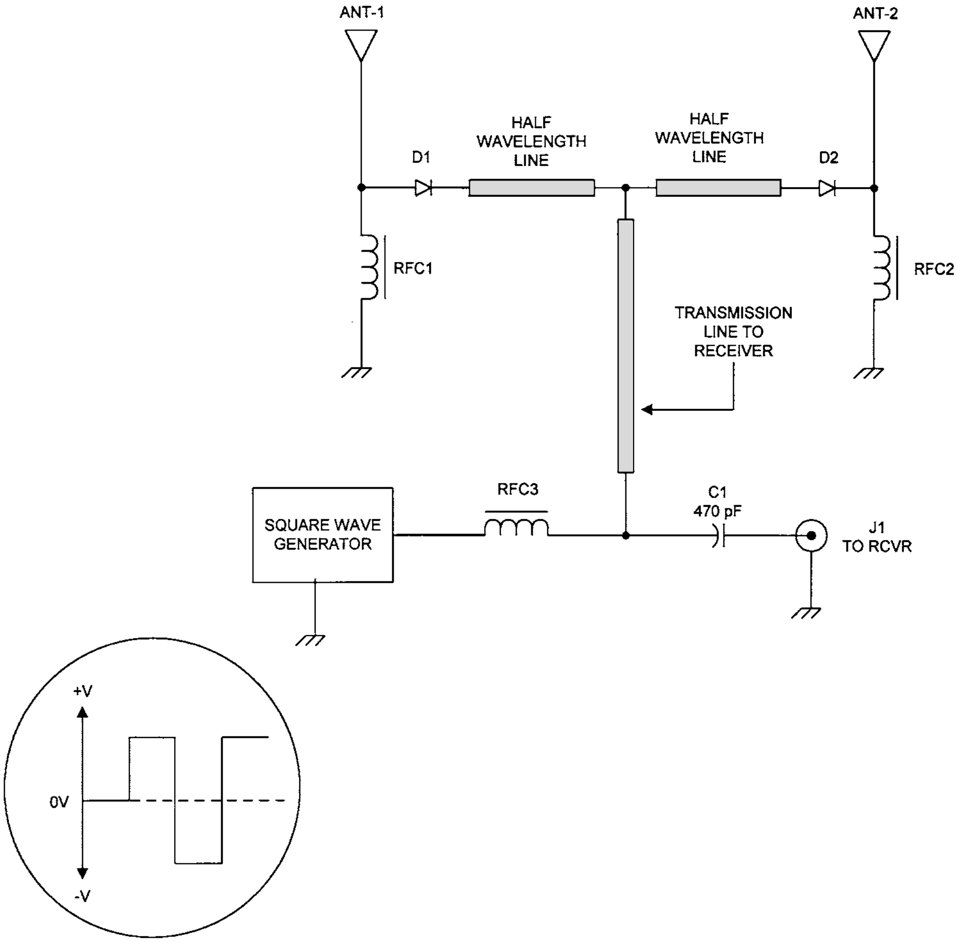
ANT-1
ANT-2
HALF WAVELENGTH LINE
HALF WAVELENGTH LINE
D1
D2
RFC1
RFC2
TRANSMISSION LINE TO RECEIVER
RFC3
C1
470 pF
SQUARE WAVE GENERATOR
J1
TO RCVR
+V
0V
-V

23-17 Double-ducky RDF system based on the switched-pattern system.

gle receiver, where it is commonly called the *double-ducky direction finder* (DDDF). The antennas are spaced from 0.25 λ to 1λ apart over a good groundplane (such as the roof of a car or truck). If no groundplane exists, then a sheet metal groundplane should be provided.

In Fig. 23-17 we see the antennas are fed from a common transmission to the receiver. In order to keep them electrically the same distance apart, a pair of identical half-wavelength sections of transmission line are used to couple to the antennas.

Switching is accomplished by using a bipolar square wave and PIN diodes. The bipolar square wave (see inset to Fig. 23-17) has a positive peak voltage and a negative peak voltage on opposite halves of the cycle. The PIN diodes (D_1 and D_2) are connected in opposite polarity to each other. Diode D_1 will conduct on negative excursions of the square wave, while D_2 conducts on positive excursions. The antenna connected to the conducting diode is the one that is connected to the receiver, while the other one is parasitic. The active antenna therefore switches back and forth between ANT1 and ANT2.

This antenna is coupled to the receiver through a small-value capacitor (C_1) so that the square wave does not enter the receiver. This allows us to use the transmission line for both the RF signal and the switching signal. An RF choke (RFC_3) is used to keep RF from the antenna from entering the square-wave generator.

The DDDF antenna produces a phase modulation of the incoming signal that has the same frequency as the square wave. This signal can be heard in the receiver output. When the signal's direction of arrival is perpendicular to the line between the two antennas, the phase difference is zero, so the audio tone disappears.

The pattern of the DDDF antenna is bidirectional, so there is the same ambiguity problem as exists with loop antennas. The ambiguity can be resolved by either of two methods. First, a reflector can be placed $\lambda/4$ behind the antennas. This is attractive, but it tends to distort the antenna pattern a little bit. The other method is to rotate the antenna through 90° (or walk an L-shaped path).

Conclusion

Radio direction finding can be very useful for locating RF noise sources, illegal stations, and other RF sources. It can also be used to locate yourself if you can get bearings on at least two stations. The lines of bearing will cross at a location approximately where you are standing—try it, you'll like it.

24

CHAPTER

Impedance matching in antenna systems

ONE OF THE FIRST THINGS THAT YOU LEARN IN RADIO COMMUNICATIONS AND broadcasting is that antenna impedance must be matched to the transmission line impedance, and that the transmission line impedance must be matched to the output impedance of the transmitter. The reason for this requirement is that maximum power transfer, between a source and a load, always occurs when the system impedances are matched. In other words, more power is transmitted from the system when the load impedance (the antenna), the transmission line impedance, and the transmitter output impedance are all matched to each other.

Of course, the trivial case is where all three sections of our system have the same impedance. For example, we could have an antenna with a simple 75-Ω resistive feedpoint impedance (typical of a half-wave dipole in free space), and a transmitter with an output impedance that will match 75 Ω. In that case, we need only connect a standard-impedance 75-Ω length of coaxial cable between the transmitter and the antenna. Job done! Or so it seems.

But there are other cases where the job is not so simple. In the case of our "standard" antenna, for example, the feedpoint impedance is rarely what the books say it should be. That ubiquitous dipole, for example, is nominally rated at 73 Ω, but even the simplest antenna books tell us that value is an approximation of the theoretical free-space impedance. At locations closer to the earth's surface, the impedance could vary over the approximate range of 30 to 130 Ω and it might have a substantial reactive component.

But there is a way out of this situation. We can construct an impedance-matching system that will marry the source impedance to the load impedance. This chapter examines several matching systems that might prove useful in a number of antenna situations.

Copyright 2001 - the McGraw-Hill Companies

Impedance-matching approaches

Antenna impedance can contain both reactive and resistive components. In most practical applications, we are searching for a purely resistive impedance ($Z = R$), but that ideal is rarely achieved. A dipole antenna, for example, has a theoretical free-space impedance of 73 Ω at resonance. As the frequency applied to the dipole is varied away from resonance, however, a reactive component appears. When the frequency is greater than resonance, then the antenna tends to look like an inductive reactance, so the impedance is $Z = R + jX$. Similarly, when the frequency is less than the resonance frequency, the antenna looks like a capacitive reactance, so the impedance is $Z = R - jX$. Also, at distances closer to the earth's surface the resistive component may not be exactly 73 Ω, but may vary from about 30 to 130 Ω. Clearly, whatever impedance coaxial cable is selected to feed the dipole, it stands a good chance of being wrong.

The method used for matching a complex load impedance (such as an antenna) to a resistive source (the most frequently encountered situation in practical radio work) is to interpose a matching network between the load and the source (Fig. 24-1). The matching network must have an impedance that is the *complex conjugate* of the complex load impedance. For example, if the load impedance is $R + jX$, then the matching network must have an impedance of $R - jX$; similarly, if the load is $R - jX$ then the matching network must be $R + jX$. The sections to follow look at some of the more popular networks that accomplish this job.

L-section network

The L-section network is one of the most used, or at least the most published, antenna matching networks in existence: it rivals even the pi network. A circuit for the L-section network is shown in Fig. 24-2A. The two resistors represent the source (R_1) and load (R_2) impedances. The elementary assumption of this network is that $R_1 < R_2$. The design equations are

$$R_1 < R_2 \text{ and } 1 < Q < 5 \qquad \textbf{[24.1]}$$

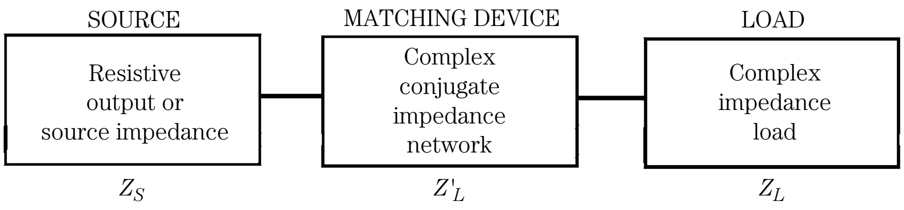
SOURCE
MATCHING DEVICE
LOAD
Resistive output or source impedance
Complex conjugate impedance network
Complex impedance load
Z_S
Z'_L
Z_L

$Z_S = R$

If:
$Z_L = R + jX$, then $Z'_L = R - jX$
$Z_L = R - jX$, then $Z'_L = R + jX$

24-1 Transmitter, matcher, and antenna system.

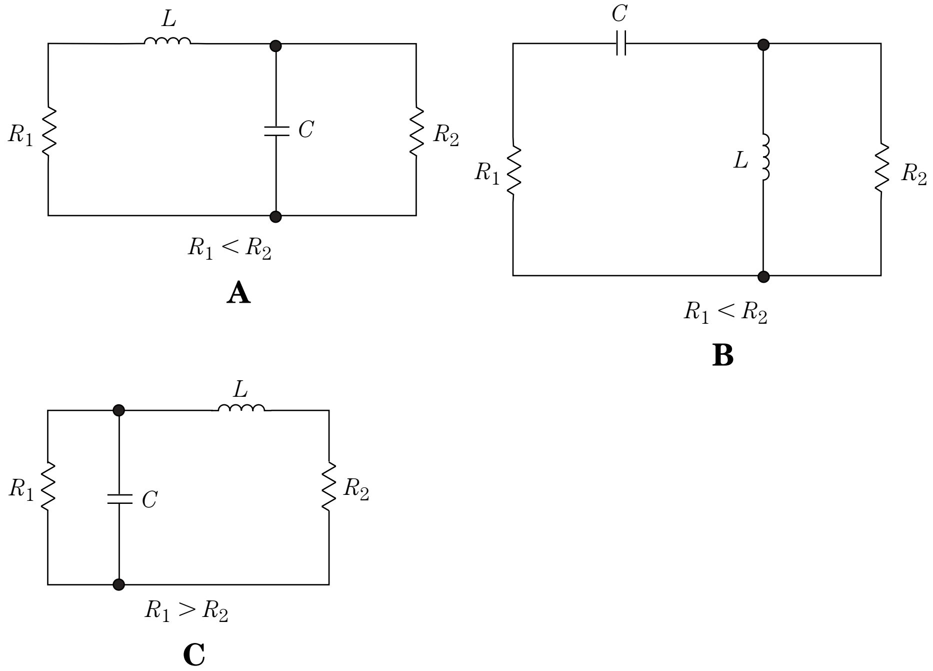

24-2 (A) L-section network; (B) reverse L section; (C) inverted L-section network.

$$X_L = 6.28FL = Q \times R_1 \qquad \textbf{[24.2]}$$

$$X_C = \frac{1}{6.28FC} \qquad \textbf{[24.3]}$$

$$Q = \sqrt{\frac{R_2}{R_1} - 1} \qquad \textbf{[24.4]}$$

also,

$$Q = \frac{X_\mathrm{L}}{R_1} = \frac{R_2}{X_\mathrm{C}} \qquad \textbf{[24.5]}$$

You will most often see this network published in conjunction with less-than-quarter-wavelength longwire antennas. One common fault of those books and articles is that they typically call for a “good ground” in order to make the antennas work properly. But they don’t tell you (1) what a “good ground” is and (2) how you can obtain it. Unfortunately, at most locations a good ground means burying a lot of copper conductor—something that most of us cannot afford. In addition, it is often the case that the person who is forced to use a longwire, instead of a better antenna, cannot construct a “good ground” under any circumstances because of landlords (and/or logistics) problems. The very factors that prompt the use of a longwire antenna in the first place also prohibit any form of practically obtainable "good ground." But there is a way out: *radials*. A good ground can be simulated with a counterpoise ground constructed of quarter-wavelength radials. These radials have a length in feet equal to $246/F_{\mathrm{MHz}}$, and as few as two of them will work wonders. Another form of L-section network is shown in Fig. 24-2B. This circuit differs from the previous circuit in that the roles of the L and C components are reversed. As you might suspect, this role reversal brings about a reversal of the impedance relationships: in this circuit the assumption is that the driving source impedance R_1 is larger than the load impedance R_2 (i.e., $R_1 > R_2$). The equations are shown below:

$$R_2 > R_1 \tag{24.6}$$

$$X_L = R_2 \sqrt{\frac{R_1}{(R_2 - R_1)}} \tag{24.7}$$

$$X_C = \frac{(R_1 R_2)}{X_L} \tag{24.8}$$

Still another form of L-section network is shown in Fig. 24-2C. Again, we are assuming that driving source impedance R_1 is larger than load impedance R_2 (i.e., $R_1 > R_2$). In this circuit, the elements are arranged similar to Fig. 24-2A, with the exception that the capacitor is at the input rather than the output of the network. The equations governing this network are:

$$R_1 > R_2 \text{ and } 1 < Q < 5 \tag{24.9}$$

$$X_L = 6.28FL = \sqrt{(R_1 R_2) - (R_2)^2} \tag{24.10}$$

$$X_C = \frac{1}{6.28FC} = \frac{R_1 R_2}{X_L} \tag{24.11}$$

$$C = \frac{1}{6.28FX_C} \tag{24.12}$$

$$L = \frac{X_L}{6.28F} \tag{24.13}$$

Thus far, we have considered only matching networks that are based on inductor and capacitor circuits. But there is also a possibility of using transmission line segments as impedance-matching devices. Two basic forms are available to us: quarter-wave sections and the series matching section.

Pi networks

The pi network shown in Fig. 24-3 is used to match a high source impedance to a low load impedance. These circuits are typically used in vacuum tube RF power amplifiers that need to match low antenna impedances. The name of the circuit comes from its resemblance to the Greek letter pi. The equations for the pi network are:

$$R_1 > R_2 \text{ and } 5 < Q < 15 \tag{24.14}$$

$$Q > \sqrt{\frac{R_1}{R_2} - 1} \tag{24.15}$$

$$X_{C2} = \frac{R_2}{\sqrt{\left(\frac{R_2}{R_1}\right)(1 + Q^2) - 1}} \tag{24.16}$$

$$X_{C1} = \frac{R_1}{Q}$$

$$X_L = \frac{R_1 \{(Q + (R_2/X_{C2})\}}{Q^2 + 1} \tag{24.17}$$

Split-capacitor network

The split-capacitor network shown in Fig. 24-4 is used to transform a source impedance that is less than the load impedance. In addition to matching antennas, this circuit is also used for interstage impedance matching inside communications equipment. The equations for design are

$$R_1 < R_2 \tag{24.18}$$

$$Q > \sqrt{\frac{R_1}{R_2} - 1} \tag{24.19}$$

$$X_L = \frac{R_2}{Q} \tag{24.20}$$

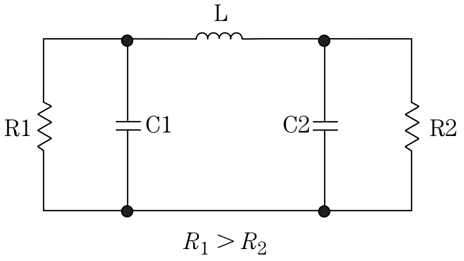

24-3 Pi network.

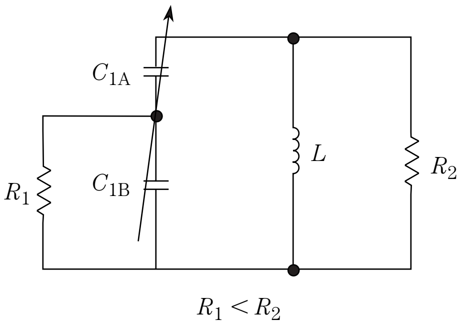

24-4 Split-capacitor network.

$$X_{C1B} = \sqrt{\frac{R_1 (Q_2 + 1)^2}{R_2}} - 1 \qquad \textbf{[24.21]}$$

$$X_{C1A} = \frac{R_2 Q}{Q^2 + 1} \left(1 - \frac{R_1}{Q X_{C1A}} \right) \qquad \textbf{[24.22]}$$

Transmatch circuit

One version of the *transmatch* is shown in Fig. 24-5. This circuit is basically a combination of the split-capacitor network and an output tuning capacitor (C_2). For the HF bands, the capacitors are on the order of 150 pF per section for C_1, and 250 pF for C_2. The collar inductor should be 28 μH. The transmatch is essentially a coax-to-coax impedance matcher, and is used to trim the mismatch from a line before it affects the transmitter.

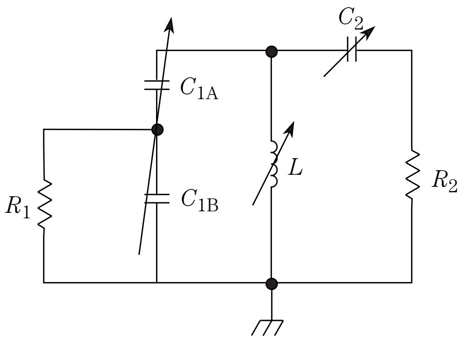

24-5 Split-capacitor transmatch network.

Perhaps the most common form of transmatch circuit is the tee network shown in Fig. 24-6. This network is lower in cost than some of the others, but suffers a problem. Although it does, in fact, match impedance (and thereby, in a naive sense, “tunes out” VSWR on coaxial lines), it also suffers a high pass characteristic. The network, therefore, does not reduce the harmonic output of the transmitter. The simple tee network does not serve one of the main purposes of the antenna tuner: harmonic reduction.

An alternative network, called the *SPC transmatch*, is shown in Fig. 24-7. This version of the circuit offers harmonic attenuation, as well as matching impedance.

Figure 24-8 shows commercially available antenna tuners based on this transmatch design. The unit shown in Fig. 24-8A is manufactured by MFJ Electronics, Inc. It contains the usual three tuning controls, here labeled *transmitter, antenna, and inductor*. Included in this instrument is an antenna selector switch that allows the operator to select a coax antenna through the tuner, to connect input to output (coax) without regard to the tuner, select a balanced antenna, or connect an internal dummy load. The instrument also contains a multifunction meter that can measure 200 W or 2000 W (full scale), in either forward or reverse directions. In addition, the meter operates as a VSWR meter.

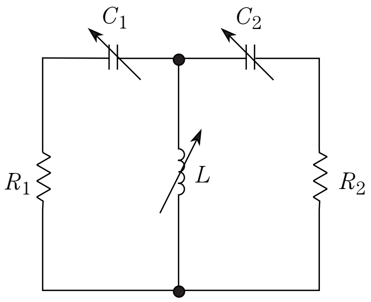

24-6 Tee-network transmatch.

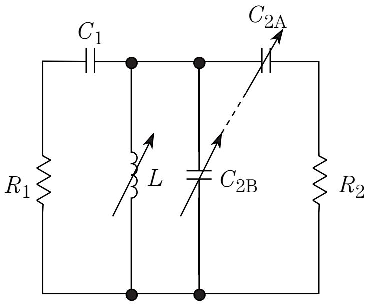

24-7 Improved transmatch offers harmonic attenuation.

Figures 24-8B through 24-8D show an imported tuner from the United Kingdom. This instrument, called the *Nevada* model, is a low-cost model, but contains the three basic controls. For proper operation, an external RF power meter, or VSWR meter, is required. The tuner is shown in Fig. 24-8B; also shown is a Heathkit transmatch antenna tuner. The rear panel of the Nevada instrument is shown in Fig. 24-8C. There are SO-239 coaxial connectors for input and unbalanced output, along with a pair of posts for the parallel line output. A three-post panel is used to select which antenna the RF goes to: unbalanced (coax) or parallel. The internal circuitry of the *Nevada* is shown in Fig. 24-8D. The roller inductor is in the center and allows the user to set the tuner to a wide range of impedances over the entire 3- to 30-MHz HF band.

MFJ VERSA TUNER V
TRANSMITTER
ANTENNA SELECTOR
COAX
ANTENNA
INDUCTOR
DUMMY LOAD
WIRE BAL LINE
COAX DIRECT
SWR SET
METER
MFJ

24-8A MFJ antenna tuner.

24-8B Author's Heathkit antenna tuner and the Nevada tuner from the United Kingdom.

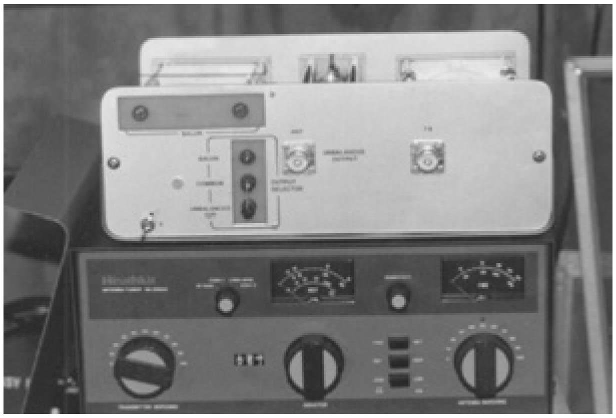

24-8C Rear of Nevada tuner.

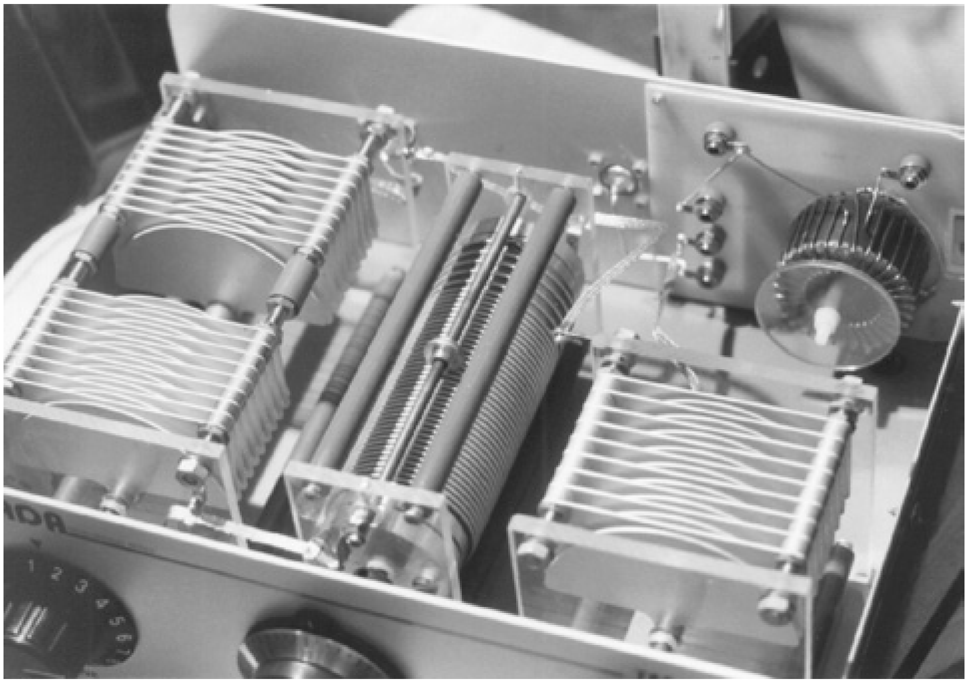

24-8D Inside of Nevada tuner.

Coaxial cable BALUNS

A balun is a transformer that matches an unbalanced resistive source impedance (such as a coaxial cable), to a "balanced" load (such as a dipole antenna). With the circuit of Fig. 24-9, we can make a balun that will transform impedance at a 4:1 ratio, with $R_2 = 4 \times R_1$. The length of the balun section of coaxial cable is

$$L_{\text{ft}} = \frac{492V}{F_{\text{MHz}}} \qquad \textbf{[24.23]}$$

where:

L_{ft} is the length, in feet
V is the velocity factor of the coaxial cable (a decimal fraction)
F_{MHz} is the operating frequency, in megahertz

Matching stubs

A shorted stub can be built to produce almost any value of reactance. This fact can be used to make an impedance-matching device that cancels the reactive portion of a complex impedance. If we have an impedance of $Z = R + j30\ \Omega$ we need to make a stub with a reactance of $-j30\ \Omega$ to match it. Two forms of matching stub are shown in Figs. 24-10A and 24-10B. These stubs are connected exactly at the feedpoint of the complex load impedance, although they are sometimes placed farther back on the line at a (perhaps) more convenient point. In that case, however, the reactance required will be transformed by the transmission line between the load and the stub.

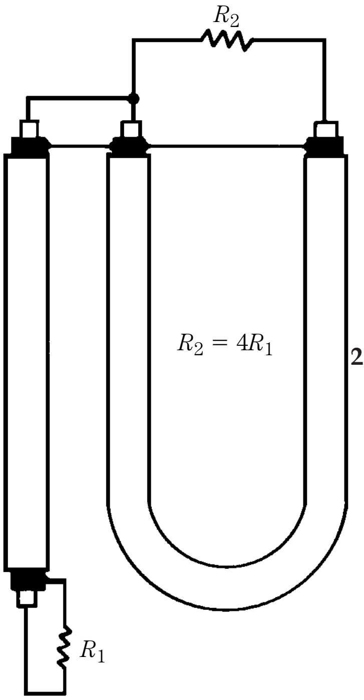

24-9 Coaxial balun transformer.

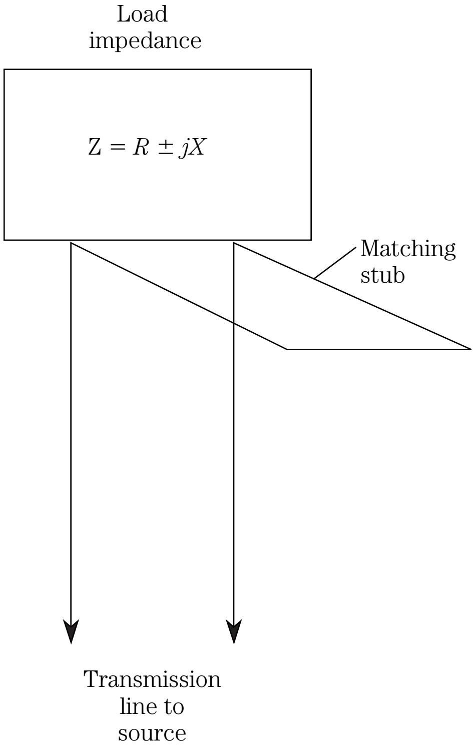
Load impedance
$Z = R \pm jX$
Matching stub
Transmission line to source

24-10A Stub match scheme.

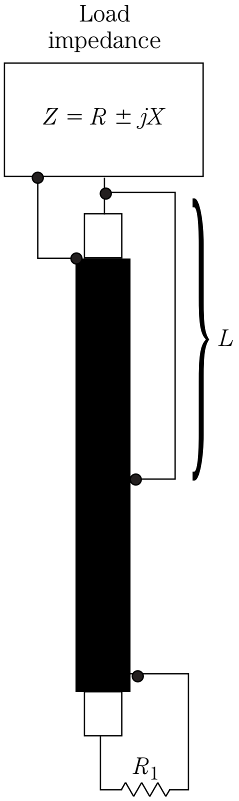

24-10B Coaxial stub matching.

Quarter-wave matching sections

Figure 24-11 shows the elementary quarter-wavelength transformer section connected between the transmission line and the antenna load. This transformer is also sometimes called a *Q section*. When designed correctly, this transmission line transformer is capable of matching the normal feedline impedance Z_s to the antenna feedpoint impedance Z_R. The key factor is to have available a piece of transmission line that has an impedance Z_o of

$$Z_o = \sqrt{Z_s Z_R} \qquad \textbf{[24.24]}$$

Most texts show this circuit for use with coaxial cable. Although it is certainly possible, and even practical in some cases, for the most part there is a serious flaw in using coax for this project. It seems that the normal range of antenna feedpoint impedances, coupled with the rigidly fixed values of coaxial-cable surge impedance available on the market, combines to yield unavailable values of Z_o. Although there are certainly situations that yield to this requirement, many times the quarter-wave section is not usable on coaxial-cable antenna systems having standard impedance values.

On parallel transmission line systems, on the other hand, it is quite easy to achieve the correct impedance for the matching section. We use the equation above to find a value for Z_o, and then calculate the dimensions of the parallel feeders. Be-

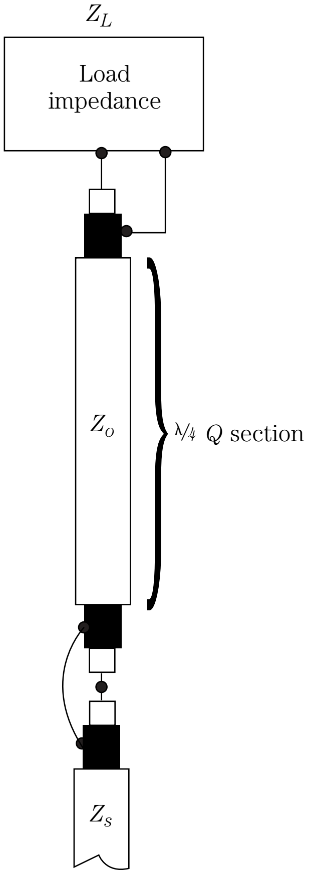

24-11 Quarter-wavelength Q section.

cause we know the impedance, and can more often than not select the conductor diameter from available wire supplies, you can use the equation below to calculate conductor spacing:

$$S = D \times 10^{(Z/276)} \qquad \textbf{[24.25]}$$

where

S is the spacing
D is the conductor diameter
Z is the desired surge impedance.
From there you can calculate the length of the quarter-wave section from the familiar $246/F_{\mathrm{MHz}}$. *D and S* are in the same units.

Series matching section

The quarter-wavelength section, covered in the preceding section, suffers from drawbacks: it must be a quarter-wavelength and it must use a specified (often nonstandard) value of impedance. The series matching section is a generalized case of the same idea, and it permits us to build an impedance transformer that overcomes these faults. According to *The ARRL Antenna Book*, this form of transformer is capable of matching any load resistance between about 5 and 1200 **Ω**. In addition, the transformer section is not located at the antenna feedpoint.

Figure 24-12 shows the basic form of the series matching section. There are three lengths of coaxial cable: L_1, L_2, and the line to the transmitter. Length L_1 and the line to the transmitter (which is any convenient length) have the same characteristic impedance, usually 75 **Ω**. Section L_2 has a different impedance from L_1 and the line to the transmitter. Note that only standard, easily obtainable values of impedance are used here.

The design of this transformer consists of finding the correct lengths for L_1 and L_2. You must know the characteristic impedance of the two lines (50 **Ω** and 75 **Ω** given as examples), and the complex antenna impedance. In the case where the antenna is non resonant, this impedance is of the form $Z = R \pm jX$, where R is the resistive portion, X is the reactive portion (inductive or capacitive), and j is the so-called imaginary operator (i.e., square root of minus one). If the antenna is resonant, then $X = 0$, and the impedance is simply R.

The first chore in designing the transformer is to *normalize* the impedances:

$$N = \frac{Z_{L1}}{Z_o} \qquad \textbf{[24.26]}$$

$$R = \frac{R_L}{Z_o} \qquad \textbf{[24.27]}$$

$$X = \frac{X_L}{Z_o} \qquad \textbf{[24.28]}$$

The lengths are determined in electrical degrees, and from that determination we can find length in feet or meters. If we adopt ARRL notation and define $A = \tan L_1$, and $B = \tan L_2$, then the following equations can be written:

If: $Z_L = R \pm jX_L$

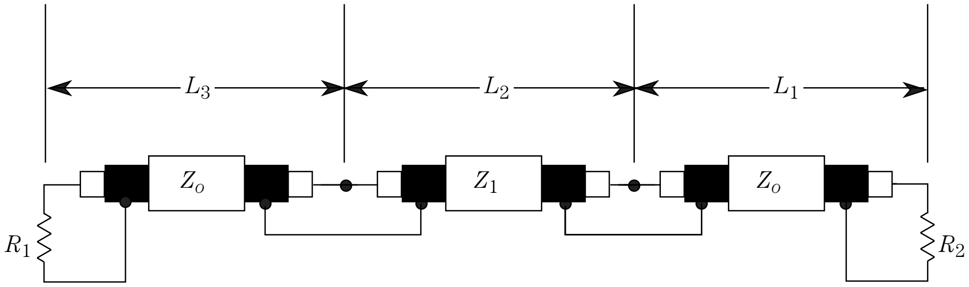

24-12 Series matching section.

$$\tan L_2 = B = \frac{(r-1)^2 + X^2}{r\{N - (1/N)^2\} - (r-1)^2 - X^2} \qquad \textbf{[24.29]}$$

$$\tan L_1 = A = \frac{\{N - (r/R)\} \times B + X}{r + (XNB) - 1} \qquad \textbf{[24.30]}$$

where:

$$N = \frac{Z_1}{Z_o}$$

$$r = \frac{R_L}{Z_o}$$

$$Z = \frac{X_L}{Z_o}$$

Constraints:

$$Z_1 > Z_o \sqrt{\mathrm{VSWR}} \qquad \textbf{[24.31]}$$

$$Z_1 < Z_o \sqrt{\mathrm{VSWR}} \qquad \textbf{[24.32]}$$

If $L_1 < 0$, then add 180°
If $B < 0$, then Z_1 is too close to Z_o
Z_1 not equal to Z_o
$Z_o \sqrt{\mathrm{VSWR}} < Z_1 < Z_o \sqrt{\mathrm{VSWR}}$

Physical length in feet:

$$L_1 = \frac{L_1 \lambda}{360} \qquad \textbf{[24.33]}$$

$$L_2 = \frac{L_2 \lambda}{360} \qquad \textbf{[24.34]}$$

where

$$\lambda = \frac{984 \times \text{velocity factor}}{\text{frequency in megahertz}} \qquad \textbf{[24.35]}$$

The physical length is determined from arctan *A* and arctan *B*, divided by 360 and multiplied by the wavelength along the line and the velocity factor.

Although the sign of *B* can be selected as either – or +, the use of + is preferred because a shorter section is obtained. In the event that the sign of *A* turns out negative, add 180° to the result.

There are constraints on the design of this transformer. For one thing, the impedance of the two sections L_1 and L_2 cannot be too close together. In general, the following relationships must be observed:

Either

$$Z_{L1} > Z_o \times \text{SWR} \qquad \textbf{[24.36]}$$

or

$$\frac{Z_{L1} < Z_o}{\text{SWR}} \qquad \textbf{[24.37]}$$

Ferrite core inductors

The word *ferrite* refers to any of several examples of a class of materials that behave similarly to powdered iron compounds, and are used in radio equipment in the form of inductors and transformers. Although the original materials were of powdered iron, and indeed the name *ferr*ite still implies iron, many modern materials are composed of other compounds. According to the Amidon Associates literature, ferrites with a permeability u of 800 to 5000 are generally of the manganese-zinc type of material, and cores with permeabilities of 20 to 800 are of nickel-zinc. The latter are useful in the 0.5- to 100-MHz range.

Toroid cores

A *toroid* is a doughnut-shaped object, so one can reasonably expect a toroidal core to be an inductor (or transformer) form, made of a ferrite material, in the general shape of a doughnut. The type of core must be known for application purposes, and is given by the type number. The number will be of the form: FT – *xx* – *nn*, where FT means ferrite toroid and describes the shape, *xx* indicates the size, and *nn* indicates the material type. The F in FT is sometimes deleted in parts lists, and the core defined as a "T-*xx*-*nn*."

A chart is provided by Amidon that gives the dimensions, and a description of the properties of the different types of material, along with a lot of other physical data. Some of these data are also available in the annual ARRL publication *The ARRL Handbook for the Radio Amateur* (the same material has appeared in earlier editions also).

Tables are derived in part from both Amidon and ARRL sources. From these tables, you can see the sizes and properties of various popular toroids. These tables, incidentally, are not exhaustive of either the variety of toroids available, or all of the properties of the toroids mentioned. Using the nomenclature mentioned, a T-50-2 core refers to a core that is useful from 1 to 30 MHz, has a permeability of 10, is painted red, and has the following dimensions: OD = 0.500 in (1.27 cm), ID = 0.281 in (0.714 cm) and a height (i.e., thickness) of 0.188 in (0.448 cm).

Toroidal transformers

A magazine reader once asked a question of this author concerning the winding protocol for toroidal transformers as seen in textbooks and magazine articles. My correspondent included a partial circuit (Fig. 24-13A) as typical of the dilemma. The question was "How do you wind it?" and a couple of alternative methods were proposed. At first I thought it was a silly question because the answer was "obvious,"

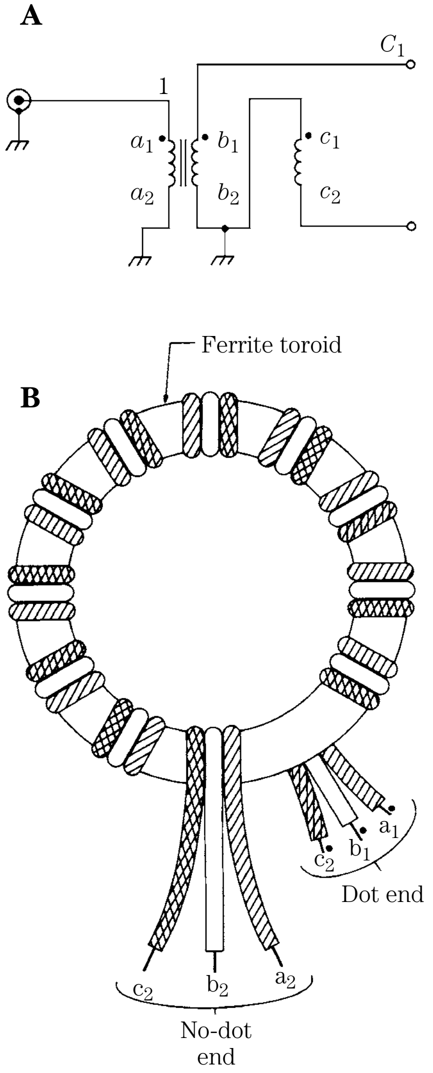

24-13 Broadband RF transformer.

and then I realized that perhaps I was wrong, and to many people the answer was not at all that obvious.

The answer to the question is that all windings are wound together in a “multifilar” manner. Because there are three windings, in this case we are talking about “trifilar” windings. Figure 24-13B shows the trifilar winding method. For the sake of clarity, I have patterned all three wires differently so that you can follow them. This practice is also a good idea for practical situations. Since most small construction projects use no.26, no.28 or no.30 enameled wire to wind coils, I keep three colors of each size on hand, and wind each winding with a different color [Note: for transmitting antenna transformers use no.16, no.14, no.12, or no.10 wire]. Otherwise, label the ends with adhesive labels.

The dots in the schematic, and on the pictorial, are provided to identify one end of the coil windings. Thus, the dot and no-dot ends are different from each other, and it usually makes a difference to circuit operation which way the ends are connected into the circuit (the issue is signal phasing).

Figure 24-14 shows two accepted methods for winding a multifilar coil on a toroidal core. Figure 24-14A is the same method as in Fig. 24-13B, but on an actual toroid instead of a pictorial representation. The wires are laid down parallel to each other as shown previously. The method in Fig. 24-14B uses twisted wires. The three wires are "chucked up" in a drill and twisted together before being wound on the core. With one end of the three wires secured in the drill chuck, anchor the other end of the three wires in something that will hold it taut. Some people use a bench vise for this purpose. Turn on the drill at slow speed and allow the wires to twist together until the desired pitch is achieved.

Be very careful when performing this operation. If you don't have a variable speed electric drill (so that it can be run at very low speed), then use an old-fashioned manual hand drill. If you use an electric drill, then wear eye protection. If the wire breaks, or gets loose from its mooring at the end opposite the drill, it will whip around wildly until the drill stops. That whipping wire will cause painful welts on the skin, and can easily damage eyes permanently.

Of the two methods for winding toroids, the method of Figs. 24-13B and 24-14A is preferred. When winding toroids, at least those of relatively few windings, pass the wire through the "doughnut hole" until the toroid is about in the middle of the length of wire. Then, loop the wire over the outside surface of the toroid, and pass it through the hole again. Repeat this process until the correct number of turns is wound onto the core. Be sure to press the wire against the toroid form, and keep it taut as you wind the coils.

Enameled wire is usually used for toroid transformers and inductors and that type of wire can lead to a problem. The enamel can chip and cause the copper conductor to contact the core. On larger cores, such as those for antenna-matching tranformers (and baluns used at kilowatt power levels), the practical solution is to wrap the bare toroid core in a layer of fiberglass packing tape. Wrap the tape exactly as if it was wire, but overlap the turns slightly to ensure covering the entire circumference of the core.

On some projects, especially those in which the coils and transformers use very fine wire (e.g., no.30), you may experience a tendency for the wire windings to unravel after the winding is completed. This problem is also easily curable. At the ends of the windings, place a tiny dab of rubber cement or RTV silicone sealer.

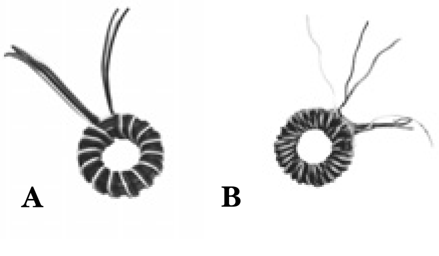

24-14 Winding a toroidal transformer
A) parallel wound,
B) twist wound.

Mounting toroid cores Now that you have a properly wound toroidal inductor or transformer, it is time to actually mount it in a circuit being constructed. There are three easy ways to do this job. First, you might be able to ignore it. If the wire is heavy enough, then just use the wire connections to the circuit board or terminal strip to support the component. But if this is not satisfactory, and in mobile equipment (or wherever else vibration is a factor) it won't be, then try laying the toroid flat on the board and cementing it in place with silicone seal or rubber cement. The third method is to drill a hole in the wiring board and use a screw and nut to secure the toroid. Do not use metallic hardware for mounting the toroid! Metallic fasteners will alter the inductance of the component and possibly render it unusable. Use nylon hardware for mounting the inductor, or transformer.

How many turns? Three factors must be taken into consideration when making toroid transformers or inductors: *toroid size, core material,* and *number of turns of wire.* The toroid size is selected as a function of power-handling capability, or for convenience of handling. The core material is selected according to the frequency range of the circuit. The only thing left to vary is the number of turns. The size and core material yields a figure called the A_L factor. The required value of inductance and the A_L factor are used in the following equation:

$$N = 100\sqrt{\frac{L_{\mu\mathrm{H}}}{A_L}} \qquad \textbf{[24.38]}$$

where:

N is the number of turns
$L_{\mu\mathrm{H}}$ is the inductance in microhenrys
A_L is the core factor, in microhenrys per 100 turns

Example Calculate the number of turns required to make a 5-μH inductor on a T-50-6 core. The A_L factor is 40.

Solution:

$$\begin{aligned} N &= 100\sqrt{\frac{L_{\mu\mathrm{H}}}{A_L}} \\ &= 100\sqrt{\frac{5_{\mu\mathrm{H}}}{40}} \\ &= 100\sqrt{0.125} \\ &= (100)\,(0.35) = 35 \end{aligned}$$

Don't take the equation value too seriously, however, because it is my experience that a wide tolerance exists on amateur-grade ferrite cores. Although it isn't too much of a problem when building transformers, it can be critical when making inductors for a tuned circuit. When you find that the tuned circuit takes considerably more (or less) capacitance than called for in the standard equation, and all of the stray capacitance is properly taken into consideration, then it may be that the actual A_L value of your particular core is different from the table value.

Ferrite rods

Another form of ferrite core available on the market is the rod, shown in Fig. 24-15A. This type of core is used to make RF chokes, such as the RFC used in the vacuum tube filament lines of a linear amplifier power tube (Fig. 24-15B). They are also used for baluns by some people. The two windings are wound in a bifilar manner over the ferrite rod. Of course, the wires used must be heavy enough to carry the filament current of the tube. As was true in the toroidal transformer, I use two different wire colors in order to make it easy to identify windings.

Ferrite rods are also used in receiving antennas. Although the amateur use is not extensive, there are places where a ferrite rod antenna (or “loopstick”) is used. For example, in radio direction finding antennas, it is common to see the ferrite loop. Also, some amateurs report that they use a loopstick receiving antenna when operating on crowded bands, such as 40 and 75 m. The small loopstick has an extremely directional characteristic, so it is capable of nulling out interfering signals. Of course, you would not want to use the loopstick for transmitting, so some means must be found for transferring the antennas over between the transmit and receive functions.

Mounting ferrite rods Ferrite rods can be mounted in several ways, two of which are analagous to the methods used on toroids. We can, for example, mount the rod using either its own wires for support or by using a dab of cement or silicone sealer to fasten it to the board. Although we cannot use simple nylon screws the way we can on toroids, we can use insulating cable clamps to secure the ends of the rod to the board.

Toroid broadbanded impedance matching transformers

The toroidal transformer forms a broadbanded means for matching antenna impedance to the transmission line, or matching the transmission line to the transmitter. The other matching methods (shown thus far) are frequency-sensitive, and must be readjusted whenever the operating frequency is changed even a small amount. Although this problem is of no great concern to fixed-frequency radio stations, it is of critical importance to stations that operate on a variety of frequencies.

Figure 24-16A shows a trifilar transformer that provides a 1:1 impedance ratio, but it will transform an unbalanced transmission line (e.g., coaxial cable) to a

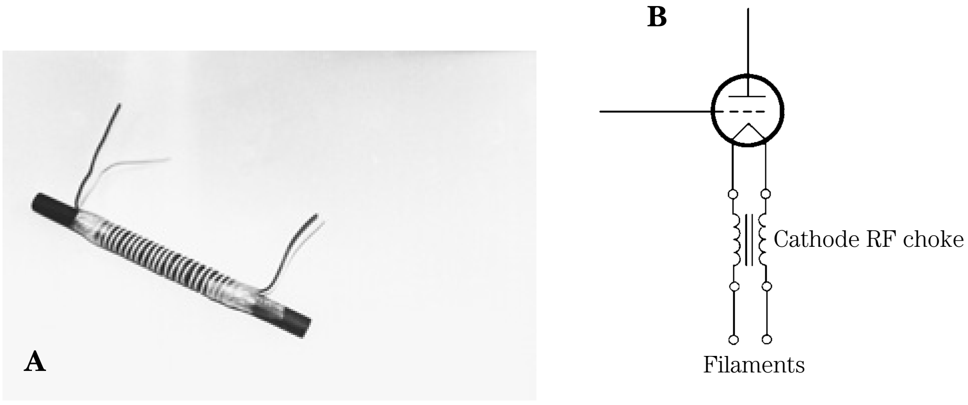

24-15 Ferrite rod inductor (A) construction, (B) use in amplifier.

balanced signal required to feed a dipole antenna. Although it *provides no impedance transformation*, it does tend to balance the feed currents in the two halves of the antenna. This fact makes it possible to obtain a more accurate figure-8 dipole radiation pattern in the horizontal plane. Many station owners make it standard practice to use a balun at the antenna feedpoint.

The balun shown in Fig. 24-16B is designed to provide the unbalanced to balanced transformation, while also providing a 4:1 impedance ratio. Thus, a 300 Ω folded dipole feedpoint impedance will be transformed to 75 Ω unbalanced. This type of balun is often included inside antenna tuners, including all three models shown in Figs. 24-8A and 24-8B. A variable (or at least settable) broadbanded transformer is shown in Fig. 24-16C. In this case the output winding is tapped, and the operator selects the correct tap needed to provide the desired impedance ratio. The usual turns ratio criterion applies.

Another multiple-impedance transformer is shown in Fig. 24-16D. In this case, the operator can select impedance transformation ratios of 1.5:1, 4:1, 9:1 or 16:1. A commercial version of this type of transformer is shown in Fig. 24-16E. It is manufactured by Palomar Engineers, and is intended for vertical antenna feeding in the HF bands. It will, however, work well on antennas other than simple verticals. A top view of the vertical antenna feeding transformer is in Fig. 24-16F.

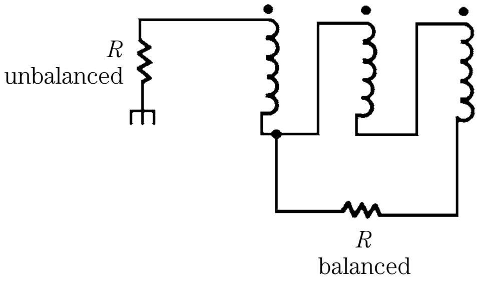

24-16A BALUN 1:1.

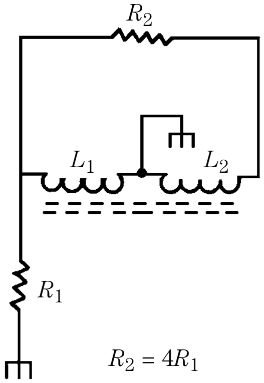

24-16B Balun 1:4.

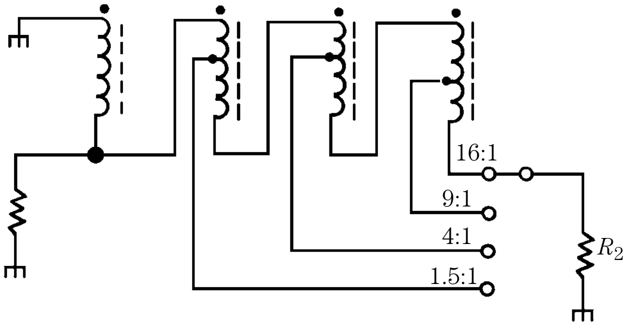
16:1
9:1
4:1
1.5:1
R_1
R_2

24-16C Tapped BALUN.

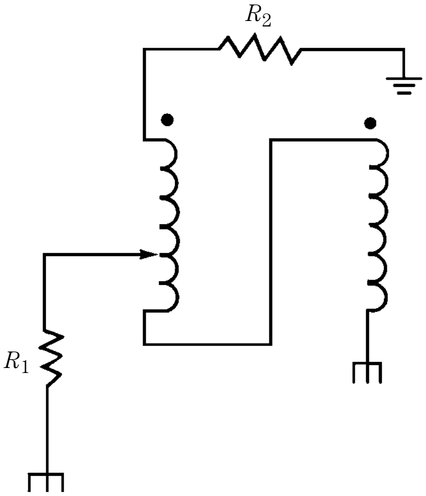

24-16D Multi-impedance BALUN.

24-16E Vertical antenna feeding transformer.

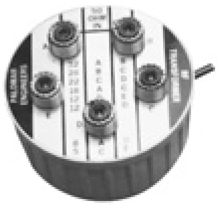

24-16F Top view of Fig. 24-16E.

This page intentionally left blank

25
CHAPTER

Mobile, emergency, portable, and marine antennas

MOBILE OPERATION OF RADIO COMMUNICATIONS EQUIPMENT DATES BACK TO ONLY A little later than "base station" operation. From its earliest times, radio buffs have attempted to place radio communications equipment in vehicles. Unfortunately, two-way radio was not terribly practical until the 1930s, when the earliest applications were amateur radio and police radio (which used frequencies in the 1.7- to 2.0-MHz region). Over the years, the land mobile and amateur radio mobile operation has moved progressively higher in frequency because of certain practical considerations. The higher the frequency, for example, the shorter the wavelength; and, therefore, the shorter a full-size antenna. On the 11-m Citizens Band, for example, a quarter-wavelength whip antenna is 102 in long, and for the 10-m amateur band only 96 in long. At VHF frequencies, antennas become even shorter. As a result, much mobile activity takes place in the VHF and UHF region.

The amateur 144-, 220-, and 440-MHz bands are popular because of several factors, not the least of which is the ease of making ¼- and ⅝-wavelength antennas. Because low-cost commercial antennas are available for these frequencies, however, we will not examine such antennas in this chapter. Rather, we will concentrate on high-frequency antennas.

Mobile HF antennas

High-frequency (HF) mobile operation requires substantially different antennas than VHF or UHF. Quarter-wavelength antennas are feasible only on the 11- and 10-m bands, with some argument in favor of 13 m as well. But by the time the frequency drops to the 21-MHz (15-m) band, the antenna size must be approximately 11 ft long, and that is too long for practical mobile operation. Because of practical considerations,

Copyright 2001 - the McGraw-Hill Companies

you must limit antenna size to 8 or 9 ft; in the amateur radio bands the 8-ft antenna is most popular because it is resonant on 10 m. At all lower frequencies, the 8-ft whip becomes capacitive, and therefore requires an equal inductive reactance to cancel the capacitive reactance of the antenna.

Figure 25-1 shows three basic configurations of coil-loaded HF antennas for frequencies lower than the natural resonant frequency of the antenna. In each case, the antennas are series fed with coaxial cable from the base; point *A* is connected to the coaxial cable center conductor, and point *B* is connected to the shield, and the car body (which serves as ground). The system shown in Fig. 25-1A is base-loaded. Although convenient, there is some evidence that the current distribution is less than optimum. The version shown in Fig. 25-1B is center-loaded, and results in an improved current distribution. This configuration is probably the most common among commercially available HF mobile antennas, although the coil is often located slightly above the center point. Finally, we see the top-loaded coil system in Fig. 25-1B. In all three cases, the point is to use the inductance of the coil to cancel the capacitive reactance of the antenna.

A modified version of the coil-loaded HF mobile antenna is shown in Fig. 25-2A. In this configuration the loading inductance is divided between two individual coils, L_1 and L_2. Coil L_1 is adjustable with respect to the antenna, while L_2 is fixed; coil L_2 is tapped, however, in order to match the impedance of the antenna to the charac-

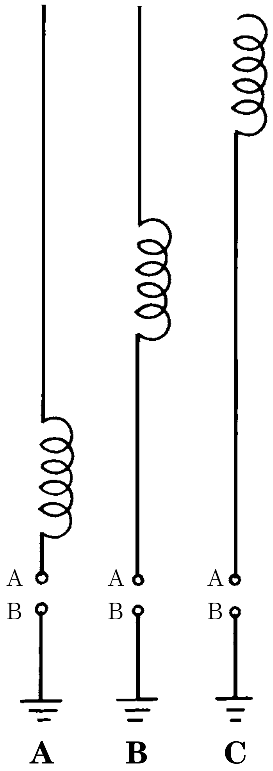

25-1 Loaded mobile antennas.

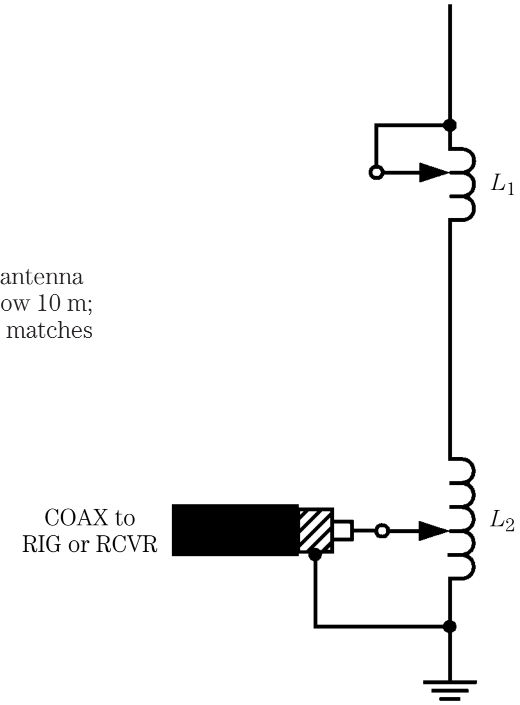

25-2A Practical mobile antenna for HF bands below 10 m; L_1 resonates, L_2 matches impedance.

teristic impedance of the coaxial-cable transmission line. When tuning this antenna, two instruments are needed: a field strength meter and a VSWR meter. The field strength meter gives a relative indication of the amount of power radiated from the antenna, while the VSWR meter helps determine the state of the impedance match. A pair of relative field strength meter projects are discussed later in this chapter.

Variations on the theme are shown in Figs. 25-2B and 25-2C. A helical whip is shown in Fig. 25-2B. In this configuration, the inductor is distributed along the length of a fiberglass antenna rod. The conductor is a wire that is helically wound on the fiberglass shaft. An adjustable tip sets the antenna to resonance. The length of this tip can be set by a local field strength meter. Fixed versions of this type of antenna are very popular on the Citizens Band where they may be as short as 30 in, or as long as 48 in. There are also some amateur radio commercial antennas based on this concept.

The other variant is shown in Fig. 25-2C. In this case, the lower end of the radiator consists of a metal tube topped with a loading coil. An adjustable shaft at the top end is used to tune the antenna to resonance. This form of antenna is popular among commercial makers of HF mobile antennas, such as the Hustler. The fixed shaft tends to be universal for all bands, while the coil and adjustable shaft form a separate "resonator" for each band. Multiband operation of this form of antenna can be accomplished by using a bracket such as shown in Fig. 25-2D. Although early versions of this scheme were homebrewed, several manufacturers currently make factory-built versions. The idea in Fig. 25-2D is to mount two or more resonators to a common fixed shaft.

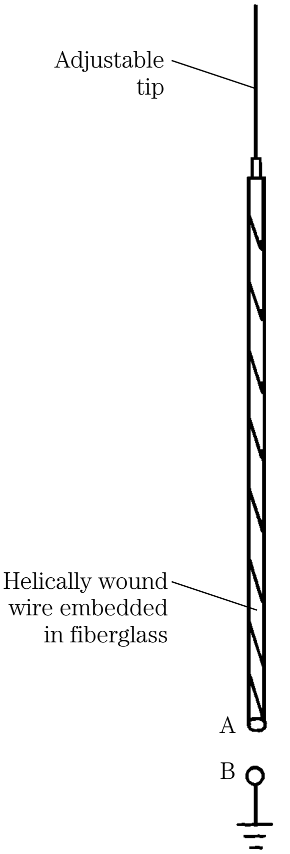
Adjustable tip
Helically wound wire embedded in fiberglass
A
B

25-2B Helically wound mobile antenna.

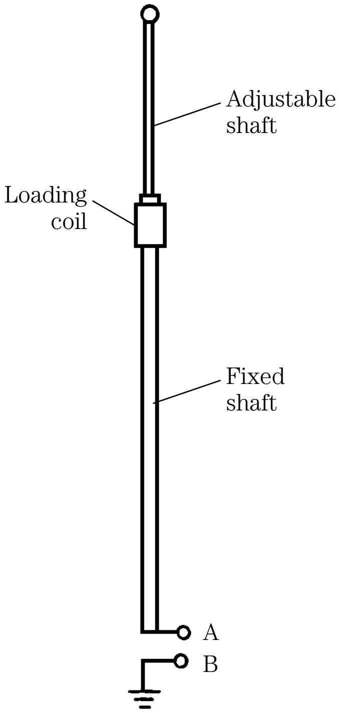
Adjustable shaft
Loading coil
Fixed shaft
A
B

25-2C Top-loaded mobile antenna common in amateur radio.

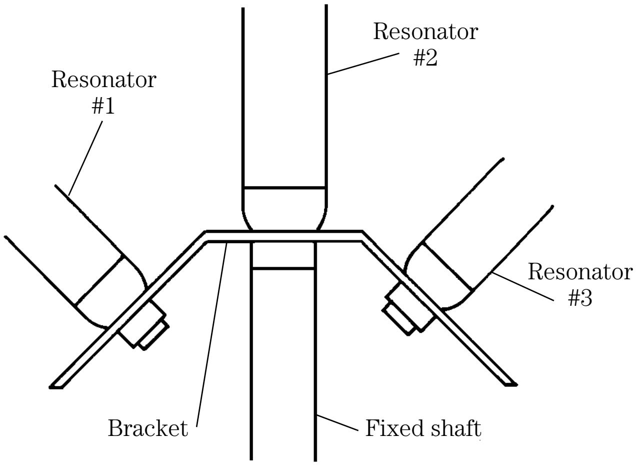
Resonator #2
Resonator #1
Resonator #3
Bracket
Fixed shaft

25-2D Multiband top-loaded antenna.

A common problem with all coil-loaded mobile antennas is that they tend to be very high *Q* antennas. In other words, they are very sharply tuned. The VSWR tends to rise rapidly as the operating frequency departs from the frequency to which the antenna is tuned; as little as 25-kHz change of operating frequency will detune the antenna significantly. Although an antenna tuner at the output of the transmitter will reduce the VSWR to a point that allows the transmitter to operate, that type of tuner is merely a line flattener that does not fully address the problem. The actual problem is that the antenna is not resonant. The efficiency of the antenna drops off rapidly as the frequency changes. The only cure for this problem is to readjust the resonator's adjustable shaft as the band segment is changed (not merely the band, but the *band segment*). Unfortunately, this solution requires tools, and the operator must get out of the vehicle in order to do a good job of retuning. Another solution is to use a motor-driven variable inductor for the loading coil. Several manufacturers offer both base-loaded and center-loaded coils that are either motor-driven or relay-selectable to permit frequency changing from the operator's seat.

Tuning HF mobile antennas

Although the procedures for tuning certain specific antennas might be different from those given below, the basic principles are the same, and can easily be adapted to any given situation. There are two situations to consider. First is the antenna in which there is either a single adjustable (or tap-selectable) loading coil, or a fixed loading coil and an adjustable shaft resonator. Second is the case (Fig. 25-2A) where there is an adjustable or tapped coil in series with a tapped impedance-matching coil at the base.

Case no. 1. In this case, we assume that the antenna has a single adjustment, either a shaft resonator or adjustable coil. In this case we need to use a field strength meter (FSM) to measure the relative field strength of the radiated signal. The antenna resonator is adjusted until the radiated field strength is maximum. Of course, this procedure must be done in steps, keying the transmitter after each adjustment to see what happened. Alternatively, a VSWR meter can be used to set the resonator to a minimum VSWR.

Case no. 2. Here, we have a coil at the base that is fixed with respect to the antenna radiator element, but is tapped with respect to the coaxial cable from the transmitter. A second coil is also used in the antenna (see Fig. 25-2A), which can be either in the center or at the top. This coil is adjustable for setting resonance. Alternatively, this coil may be fixed, and an adjustable resonator shaft may be used to set operating frequency. In this type of antenna the upper coil, or resonator shaft, is adjusted to resonance by using a field strength meter. The lower coil is adjusted for minimum VSWR. Both adjustments are needed to complete the job.

Field strength meters

A *field strength meter* (FSM) is an instrument that measures the radiated field from an antenna. Commercial engineering-grade instruments are calibrated in terms of either watts/square centimeter or volts/meter, and are used for jobs such as broadcast station "proof-of-performance" tests and other professional jobs. For adjusting antennas, however, a considerably different instrument is sufficient. This section describes two simple, passive (which means no dc power is required) field strength meters usable for adjusting HF radio antennas, including both CB and amateur radio antennas.

Two forms are shown in Fig. 25-3; both are basically variations on the old-fashioned "crystal set" theme. Figure 25-3A shows the simplest form of untuned FSM. In this circuit a small whip antenna (used for signal pickup) is connected to one end of a grounded RF choke (RFC). The RF voltage developed across the RFC is applied to a germanium diode detector; either IN34 or IN60 can be used. Silicon diodes are normally preferred in signal applications, but in this case we need the lower contact potential of germanium diodes in order to improve sensitivity (V_g is 0.2 to 0.3 V for Ge and 0.6 to 0.7. V for Si). A potentiometer is used both as the load for the diode and as a sensitivity control to set the meter reading to a convenient level.

The untuned version of the FSM is usable even at low powers, but suffers from a lack of sensitivity. Only a certain amount of signal can be developed across the RFC, so this limits the sensitivity. Also, the RFC does not make a good impedance match to the detector diode D_1. An improvement is possible by adding a tuned circuit and an impedance-matching scheme, as shown in Fig. 25-3B. In this case a variable capacitor C_3 is used as a tune control. The tuning capacitor is parallel resonant with inductor L_1. A tapped capacitive voltage divider (C_1/C_2) is used to provide impedance matching to the diode. Figure 25-3C gives values of capacitance and inductance for various bands.

Operation of the tunable FSM is simple and straightforward. Set the sensitivity control to approximately half-scale, and then key the transmitter; adjust the tuning control (C_3) for maximum deflection of the meter pointer (readjustment of the sensitivity control may be needed). After these adjustments are made, the tunable FSM works just like any other FSM.

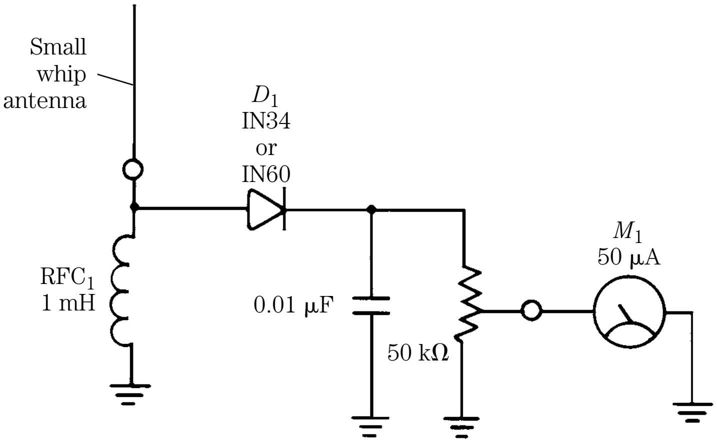

25-3A Simple field strength meter.

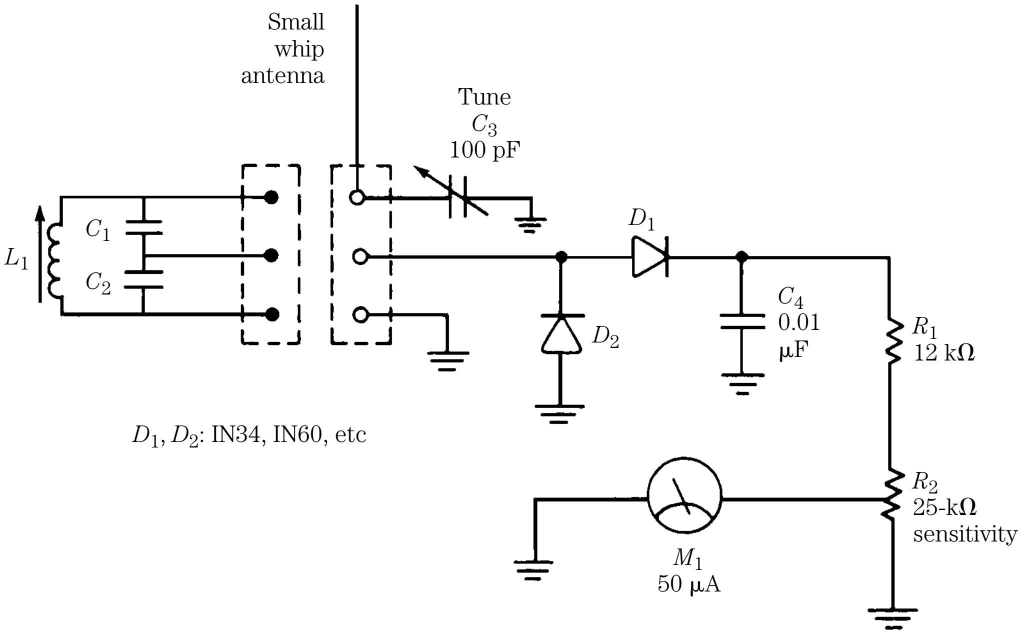

25-3B Tuned field strength meters.

25-3C Values chart for field strength meters.

Band (meters)	L_1 (µH)	C_1 (pF)	C_2 (pF)
75/80	27	25	100
40	10	15	68
30	4.4	15	68
20	2.25	10	68
17	1.7	10	56
15	1.3	10	47
13	0.8	10	39
10	0.5	10	39

Antennas for emergency operations

Some time ago, I met an interesting character at a convention. As a medical doctor, he was a medical missionary working at a relief station in Sudan. Because of his unique business address, he was able to discuss mobile and portable antennas for communications from the boondocks. His *bona fides* for this knowledge includes the fact that he is licensed to operate on both the amateur radio bands, and as a landmobile (or point-to-point) station in the 6.2-MHz band. The desert where he travels is among the worst in the world. The path they euphemistically call a "road" is occasionally littered with camel corpses because of the harsh conditions. The doctor's organization requires him to check in twice daily on either 6.2 MHz or

3.885 MHz (which some missionary hams in Africa use as an unofficial calling frequency). If he misses two check-ins in a row, then the search and rescue planes are sent up. As a result of his unique "house calls," he does a lot of mobile and portable operating in the lower-HF region of the spectrum. His problem is this: How do you reliably get through the QRM and tropical QRN with only 200 W PEP and a standard loaded mobile antenna?

Another fellow I once met works in Alaska for a government agency. He faces many of the same problems as the doctor in Sudan, but at close to 100° colder. He frequently takes his 100-W mobile rig into the boondocks with him in a four-wheel drive vehicle. Again, with only 100 W into a poor-efficiency loaded mobile antenna, how does one reliably cut through the interference to be heard back at the homestead?

An earthquake, or hurricane, strikes your community. Antenna towers collapse, tribanders become tangled masses of aluminum tubing, dipoles are snarled globs of no.14 Copperweld, and the rig and linear amplifier are smashed under the rubble of one corner of your house. All that remains is the 100-W HF rig in your car. How do you reliably establish communications in "kilowatt alley" with a 100-W mobile driving a 75-W loaded whip? Of course, you always got through one way or another before, but now communications are not for fun—they are deadly serious. Somehow, the distant problems of a Sudanese missionary doctor and a KL7 government forester don't seem too very far away.

For these operators, communications often means life or death for someone, perhaps themselves. Given the inefficiency of the loaded whips typically used as mobile antennas in the low-HF region, the generally low power levels used in available mobile rigs, and the crowded band conditions on the 80-, 75-, and 40-m bands, it becomes a matter of more than academic interest how you might increase the signal strength from your portable (or mobile) emergency station. Anything we can do, easily and cheaply, to improve the signal is like having money in the bank. Fortunately, there are several tricks of the trade that will help us out in a pinch.

Figure 25-4 shows a typical mobile antenna for a low-HF band. Because quarter-wavelength antennas on these frequencies are 30 to 70 ft high, full-size vertical whip antennas are not practical. In fact, at frequencies below 10 m, full-size whips are not generally used. Short antennas exhibit capacitive reactance, so we add a loading coil (inductor L in Fig. 25-4) to the radiator to make up the difference (its inductive reactance cancels the capacitive reactance of the antenna). The inductor in such a loaded antenna can be placed almost anywhere along the radiator, although base-, center-, and top-loaded designs predominate. The actual inductance needed varies somewhat with coil placement, as does antenna performance. The "resonators" used on commercial low-HF mobile antennas are loading coils encapsulated in a weathertight housing.

The mobile configuration is inefficient by its nature, and little can be done to improve matters. Of course, an antenna-matching device or tuner will help in optimizing the power transfer to the antenna, and should always be used in any event (especially with solid-state final amplifiers, which don't tolerate VSWR as easily as do tube finals). In the portable configuration, however, we can both improve the performance of "mobile" antennas and look into antenna options not open to the mobile

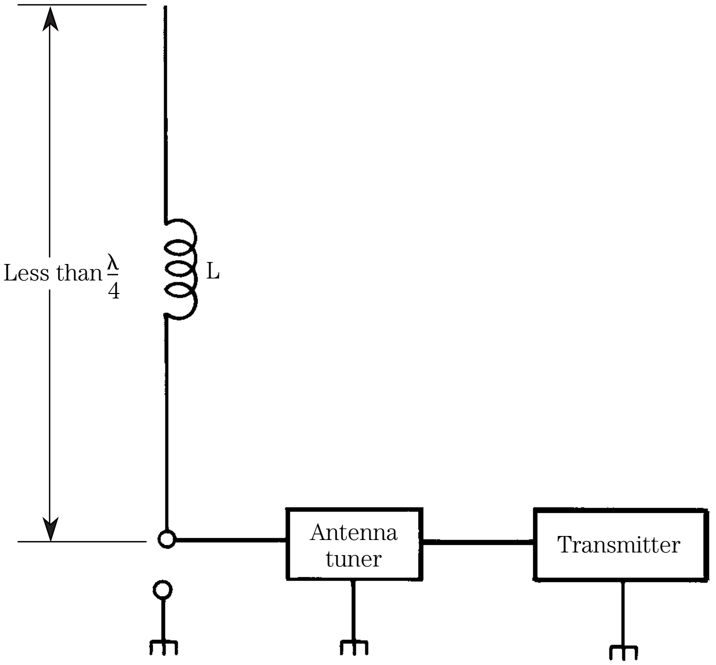

25-4 Basic mobile antenna system often proves unsatisfactory in marginal cases.

operator. We can make a basic assumption for these improvements: The operator needing emergency or urgent communications is not driving down the interstate somewhere, but is stopped in the back country and needs to call out.

In cases of emergencies on most highways, we are in range of some repeater, so we would use a VHF band (probably 2 m) to contact police or other emergency services through a repeater autopatch. In fact, with the wide availability of repeaters around the country it behooves any amateur, backpacking or four-wheeling into remote areas, to be familiar with repeater locations and frequencies. From hilltops, especially, it is often possible to hit repeaters from a surprisingly long distance. I have seen hand-held 2-m rigs trigger mountain top repeaters from the mountains of southwestern Virginia, and presume that other areas of the country are as well off repeaterwise. This chapter deals with HF rigs, especially those operating in the lower end of the HF spectrum, in situations where a temporary antenna must be erected.

Most amateurs, who have low-HF mobile rigs, will testify that getting out is a pain in the mike button. Antenna efficiencies are simply too poor. One factor in this dismal equation is that the vehicle body makes a very poor antenna ground system. One could, I suppose, carry a supply of 6- or 8-ft copper-clad steel ground rods, and a 9-lb sledge hammer to drive them into the earth wherever a portable station is needed. Somehow, that solution doesn't seem too appealing at this point! Besides, have you ever tried to remove an effective 8-ft ground rod from the ground? Forget it, it's too much pain for too little gain. A better solution is to provide a counterpoise groundplane, as shown in Fig. 25-5. In Fig. 25-5A we see the electrical situation and connection, while in Fig. 25-5B we see the mechanical scheme for a specific situation: A groundplane consists of two or more (even *one* helps) quarter-wavelength radials connected to the antenna ground point (i.e., where the coaxial cable shield connects to the vehicle body). The radials are made of no. 14 wire, so are relatively easy to stow.

Figure 25-5B shows a workable system that will improve the performance of a mobile rig in stationary situations. The mobile antenna uses the normal base mount attached to the rear quarter panel of the car adjacent to the trunk lid. An all-metal grounding-type binding post is installed through an extra hole drilled in the base

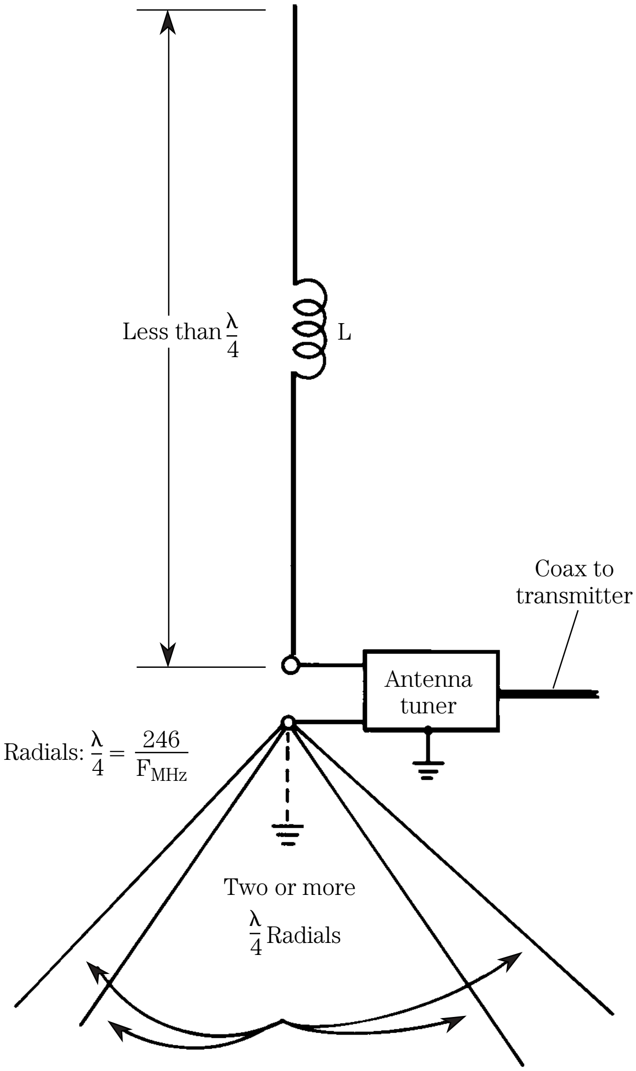
Less than $\frac{\lambda}{4}$
L
Coax to transmitter
Antenna tuner
Radials: $\frac{\lambda}{4} = \frac{246}{F_{MHz}}$
Two or more
$\frac{\lambda}{4}$ Radials

25-5A Two or more radials improve efficiency of mobile antenna used in a portable communications setup.

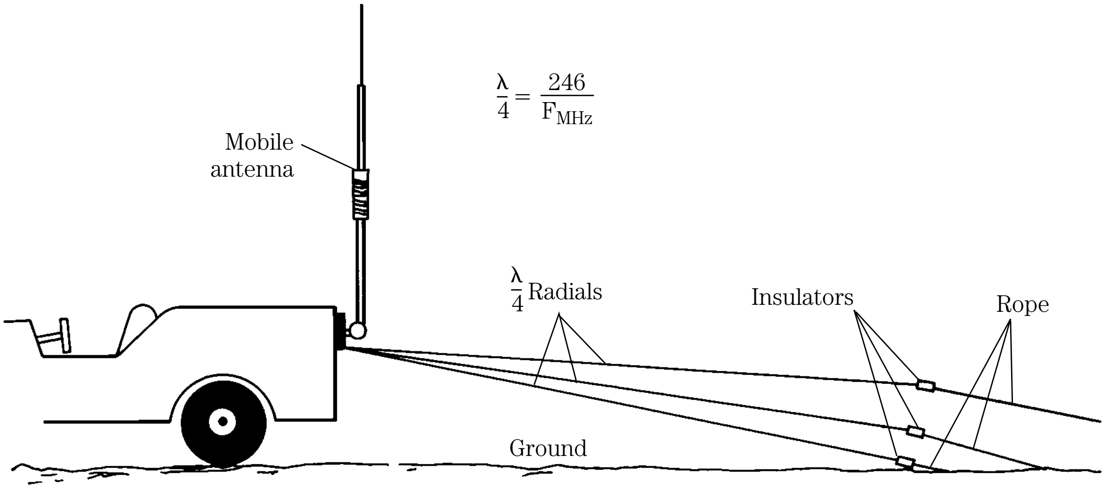
$\frac{\lambda}{4} = \frac{246}{F_{MHz}}$
Mobile antenna
$\frac{\lambda}{4}$ Radials
Insulators
Rope
Ground

25-5B Multiple radials are even better.

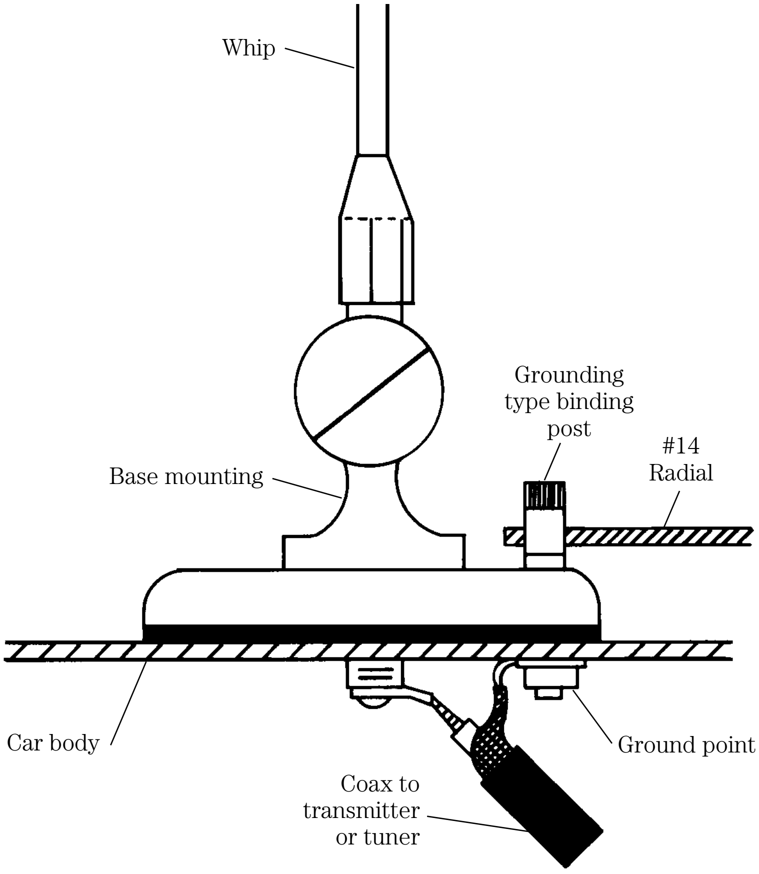
Whip
Grounding type binding post
#14 Radial
Base mounting
Car body
Ground point
Coax to transmitter or tuner

25-5C Connection of radials to mobile antenna.

insulator (see Fig. 25-5C). Radials for portable operation are attached at this point. Although the binding post is small, it easily accommodates two no. 14 radials. The owner used the vehicle for camping, and the radials were used to improve antenna efficiency. I don't believe the dramatic accounts enthusiastically reported by the owner of the car, but my own Field Day experience (and the testimony of the Sudanese missionary doctor) lead me to expect considerable improvement over the unaided loaded whip.

A potential solution to the inefficiency problem is to replace the mobile antenna with a more efficient, but stowable, antenna that can be brought out and erected when needed. One candidate is a surplus military HF whip antenna. Intended for jeeps and communications trucks, these antenna/tuner combinations are collapsible and are as efficient as any on the market. Although my Alaskan friend could use such a surplus find, my Sudanese friend could not. He told me that Americans in the boonies, even (perhaps especially) missionaries, are always suspected by pinheaded local authorities of being agents of the dreaded "Cuban Invasion Authority." Showing up "out there" with obviously military radio equipment serves only to seal that belief!

The common dipole is often looked down upon by the owners of massive array antennas, but those old-fashioned antennas are capable of turning in some impressive results. The dipole is made by connecting two quarter-wavelength pieces of wire to a coaxial-cable transmission line. One length is connected to the center conductor, and the other is connected to the shield of the coax. In a pinch, zip (i.e., lamp) cord and twisted pairs of hookup wire will do for a transmission line. Figure 25-6 shows the common dipole and the normal equation for determining *approximate* length. The actual length is found by trimming length until the VSWR drops to its lowest point, a nicety that might not seem altogether important in an emergency. The ends of the dipole must be supported on trees, masts, or some other elevated structure. Unfortunately, mounting points aren't easily found. My Sudanese friend doesn't see many trees!

Figure 25-7 shows another alternative antenna that works well for portable operation. Based on the dipole, this inverted-vee dipole does not require two end supports but rather a single center support. The length of each leg is 6 percent longer than for a nominal dipole. Because the application is both emergency-related and temporary in nature, we can get away with construction methods that would be unthinkable in more permanent installations.

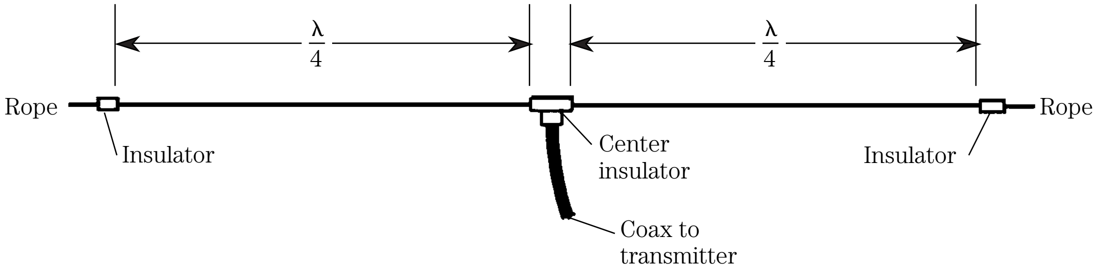

25-6 Basic dipole.

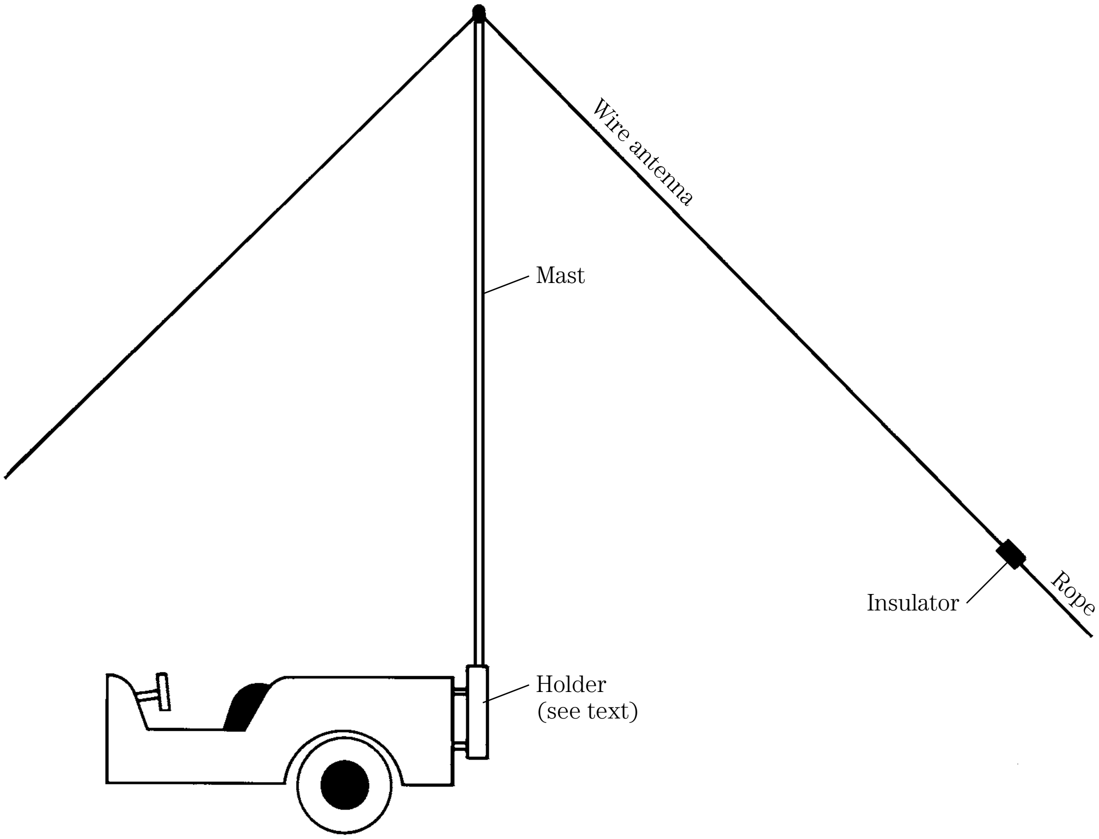

25-7 Portable inverted-vee antenna.

Three problems present themselves. First, the antenna must be portable for backpackers, or stowable in the case of those people who motor into the back country. Size and weight are major constraints in both cases, even though the vehicular case is a lot easier to work with than the backpacker. Second, what materials and means of construction are needed for the mast? Third, how is the mast supported? Because the answer to the third question depends in part on the answer selected for the second, we will deal with it first. The first question is avoided at this point because I have selected the vehicular case.

One alternative is to use a telescoping TV antenna mast to support your antenna. We can either mount a mobile whip and its associated radials at the top, or install an inverted-vee per Fig. 25-7. These masts collapse to 6 or 8 ft in length, but can be slipped up to heights of 18, 25, 30, 40, or 50 ft, depending upon the type selected. Keep in mind, when shopping for these masts, that the larger models are considerably heavier than shorter models, and they require two or more people to install them. Erecting a 40- or 50-ft telescoping mast is not a single-person job, even though you might know someone who has done it. I've done it, and won't do it again! Even the 30-ft models are a bit hairy to install alone.

Another mast that can be used for the inverted vee is made from PVC plastic plumbing pipe. If it can be carried on or in the vehicle, then lengths up to 10 ft are available. Longer lengths are made at the site of erection by joining together 10-ft or shorter sections with couplings (also available at plumbing supply outlets). Be careful of using PVC pipe that is too small, however. PVC pipe is relatively thin walled, and it is therefore flexible. Sizes below 1.5-in diameter will not easily stand alone without guying. While a single 10-ft section might be self-supporting, two or more sections together will not support themselves and the weight of the antenna. Guying can be accomplished with ropes, or on a temporary basis, heavy twine.

Still another alternative is to carry steel TV antenna masts. Available in 5- and 10-ft lengths, these masts are flared on one end and crimped on the other, so that they can be joined together to form longer lengths. In addition, the same sources that sell these masts also sell the guy wire rings that help support the mast when installed. There is also a variety of rooftop mounting devices that aid also in ground mounting the antenna.

I used the TV mast solution on Field Day expeditions for many years. The inverted-vee antenna is, of course, a natural for this solution. Once, however, we used a Hustler mobile antenna mounted at the top of a 20-ft mast, with four radials, and it worked surprisingly well with the 30-W limit for extra points credit then available for Field Day contestants.

There are, then, several alternatives for masts that slip up, snap together, or are otherwise easily unstowed and installed. Let's now turn our attention to base mounting schemes that can be used to support the mast. Of course, in a life or death pinch, you could always just pile rocks around the base, or hold the darn thing up as you transmit. But a little forethought could eliminate those problems easily.

One of the first solutions I'll deal with was seen on the Outer Banks of North Carolina, and was provided to me by a CB operator. Surf fishers on the Outer Banks use four-wheel drive vehicles to get out on the beach (to the surf, where the big sea bass lurk). Welded to either the front or rear bumper attachments are steel tubes (see Fig. 25-7) used for mounting the very, very long surf casting rods they use to catch fish. The CB operator had a 20-ft mast consisting of two 10-ft TV mast sections mounted in one of the rod holders. At the upper end of the mast was his 11-m groundplane antenna. The same method of mounting would also support similar amateur antennas, inverted-vee dipoles, or VHF/UHF antennas.

Given that the antenna installation will be temporary, lasting only a few hours or a couple of days in the worst case, we need not worry about long-term integrity, or the practicality of the installation. Mounting the mast to the back of a four-wheeler or pickup truck with a pair of U bolts is not terribly practical if you must move the vehicle, but works nicely if you plan to camp (or are stranded) for a few days.

For lightweight masts, up to about 25 ft, the base support could be an X-shaped base made of 2 × 4-in lumber (Fig. 25-8A), or even a Christmas tree holder. Alternatively, a TV antenna rooftop tripod mount (Fig. 25-8B) is easily adopted for use on the ground. None of these three alternatives can be depended upon for self-supporting installations, and must be guyed even if used for only a short period. Again, because of the temporary nature of the installation, aluminum or wooden tent pegs can be used to anchor the guy wires. Although they are insufficient for long-term installations, they work fine for the short run.

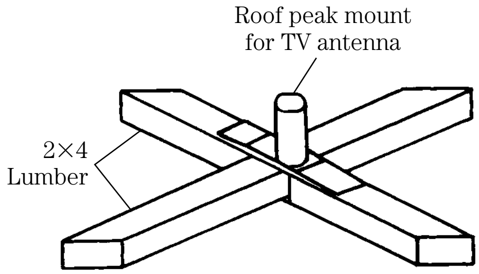

25-8A Wooden antenna stand.

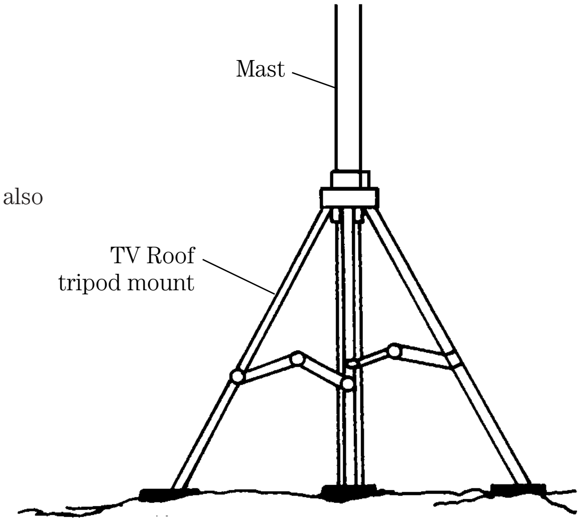

25-8B TV-antenna tripod also works in portable situation for HF antenna masts.

Another problem

Field Day is a reasonably good training ground for people who anticipate operating radio equipment under primitive conditions. If you pay attention to the sometimes humorous foul-ups seen while getting on the air from atop High Peak Mountain in Goose Bump National Park, then you might avoid some of the same mistakes when foul-ups mean more than a quick drive to the nearest town. You will learn, for example, which simple tools are a must for constructing and repairing simple antennas (including the 12-Vdc soldering iron that runs off the vehicle battery). You will also learn a bit about electricity needed to run the rig.

Two alternatives present themselves. First, you could use 12 Vdc from the vehicle electrical system, which makes sense with the current crop of 12 Vdc HF rigs on the market. Or, you could operate from 120 Vac generated by a light plant generator.

When boondocking in a four-wheeler (or other vehicle), it is wise to use a dual battery system such as shown in Fig. 25-9. Two separate 12-Vdc auto batteries (preferably high ampere-hour capacity) are connected essentially in parallel with the alternative charging system. Diodes D_1 and D_2 are rated at 100 A, 50 V peak inverse voltage (PVI), and are used to isolate the two batteries from each other. Diode assemblies such as this can be built, or purchased from van conversion and recreational

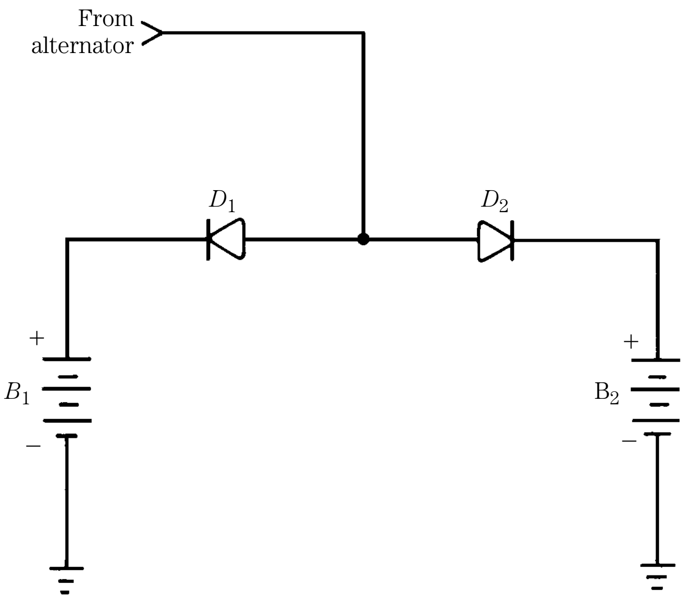

25-9 Diode pack allows charging two batteries from system charger.

vehicle shops. Van and RV owners' use is the same as yours: It's a darn shame to run down the battery needed to start the vehicle just by operating your rig. You not only cannot start the vehicle in such a case, but cannot muster enough power to call for the wilderness equivalent of "Triple A."

One experienced boondocker I met in southern Arizona totes either a 450-W Kawasaki or 500-W Honda mini light plant generator in the back of his Bronco. These models are surprisingly lightweight and quiet operating. One of the generators has a 12-Vdc, 8-A outlet that can be used to charge a dead battery. The other has only the 110-Vac outlet, but can be used for battery charging if a small charger is provided. Simple 10-A battery chargers can usually be bought at auto parts or accessories stores, for about $60.

Operating radio communications equipment successfully under primitive conditions depends upon two major factors: available electrical power and a proper, efficient, antenna system. Although dealing in detail with the means of obtaining power in remote locations is beyond the scope of this book, we have provided some ideas from which you can start planning your own "survival" radio system.

Marine radio antennas

Radio communications considerably lessened the dangers inherent in sea travel. So much so, in fact, that the maritime industry took to the new "wireless telegraphy" earlier than any other segment of society. In the early days of radio, a number of exciting rescues occurred because of wireless. Even the infamous *S.S. Titanic* sinking might have cost less human life if the wireless operator aboard a nearby ship had been on duty. Today, shore stations and ships maintain 24-hour surveillance, and vessels can be equipped with autoalarm devices to wake up those who are not on duty. Even small pleasure boats are now equipped with radio communications, and these are monitored around the clock by the U.S. Coast Guard.

The small shipboard operator can have a selection of either HF single-sideband communications or VHF-FM communications. The general rule is that a station must have the VHF-FM, and can be licensed for the HF SSB mode only if VHF-FM is also aboard. The VHF-FM radio is used in coastal waters, inland waters, and harbors. These radios are equipped with a high/low power switch that permits the use of low power (1 W typically) in the harbor, but higher power when under way. The HF SSB radios are more powerful (100 W typically), and are used for offshore long-distance communications beyond the line-of-radio-sight capability of the VHF-FM band.

The problems of antennas on boats are the same as for shore installations, but are aggravated by certain factors. Space, for one thing, is less on a boat so most HF antennas must be compensation types. Also, grounds are harder to come by on a fiberglass or wooden boat, so external grounds must be provided.

A typical powerboat example is shown in Fig. 25-10. The radio is connected to a whip antenna through a transmission line (and a tuner on HF), while also being grounded to an externally provided groundplate. Over the years, radio grounds have taken a lot of different forms. For example, the ground might be copper or aluminum foil cemented to the boat hull. Alternatively, it might be a bronze plate or hollow bronze tube along the center line of the boat, or along the sides just below the water line. The ground will also be connected to the engine. Careful attention must be paid

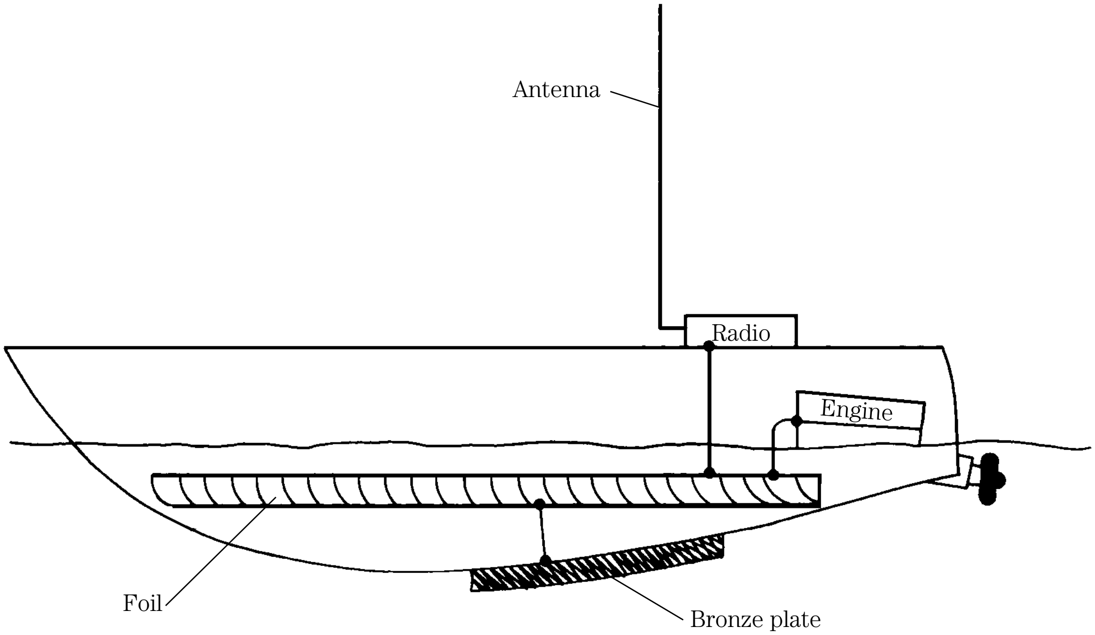

25-10 Motorboat grounding system.

to the electrical system of the boat when creating external grounds to prevent electrolytic corrosion from inadvertent current flows.

On sailboats the whip antenna (especially VHF-FM) might be mast-mounted as shown in Fig. 25-11. The same grounding scheme applies with the addition of a metal keel or metal foil over a nonmetallic keel.

The whips used for boat radios tend to be longer than land mobile antennas for the same frequency. The VHF-FM whip (Fig. 25-12) can be several quarter-wavelengths and take advantage of gain characteristics thereby obtained. Whips for the HF bands tend to be 10 to 30 ft in length, and often look like trolling rods on power boats. You will also see Citizens Band radios on board boats, and the CB antennas are also whips. It is not a good idea to rely solely on CB for boat communications, because it is a lot less

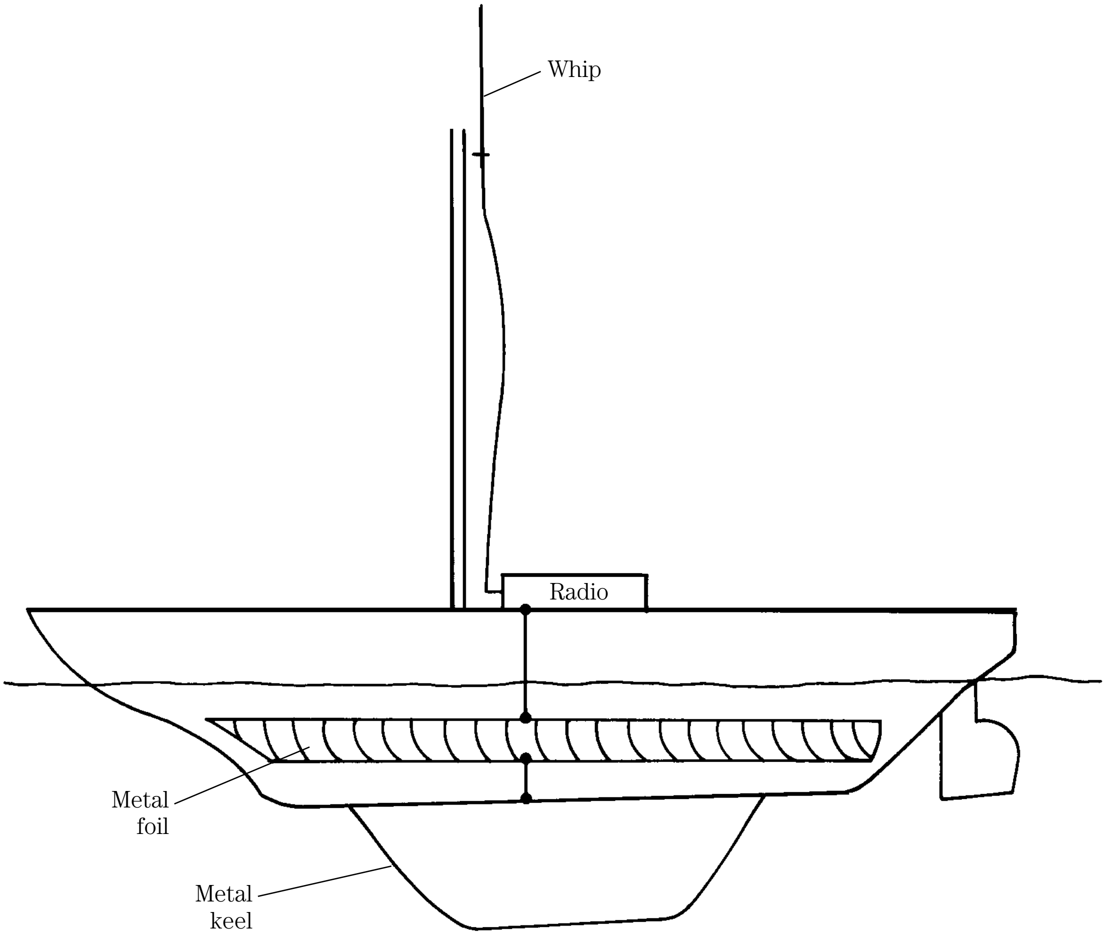

25-11 Sailboat grounding.

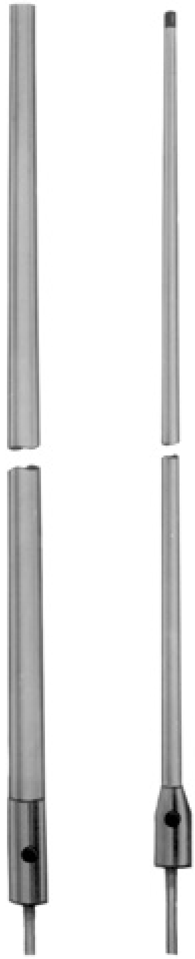

25-12 VHF-FM boat antenna.
Courtesy of Antenna Specialist Company

likely to be useful in emergencies. Also, some boaters use amateur radio sets, often illegally, for emergency communications in lieu of a proper HF-SSB unit.

Longwire antennas also find use in marine service. Figure 25-13 shows two installations. The antenna in Fig. 25-13A shows a wire stretched between the stern and bow by way of the mast. The antenna is end-fed from an antenna tuner or “line flattener.” The longwire shown in Fig. 25-13B is similar in concept, but runs from the bottom to the top of the mast. Again, a tuner is needed to match the antenna to the radio transmission line. Notice that the antennas used in this manner are actually not “longwires” in the truly rigorous sense of the term, but rather “random length” antennas.

The format of the tuner can be any of several designs, shown in Fig. 25-14. The reversed L-section coupler shown in Fig. 25-14A is used when the antenna radiator element is less than a quarter-wavelength. Similarly, when the antenna is greater than a quarter-wavelength, the circuit of Fig. 25-14B is the tuner of choice. This circuit is a modified L-section coupler that uses two variable capacitors and the inductor.

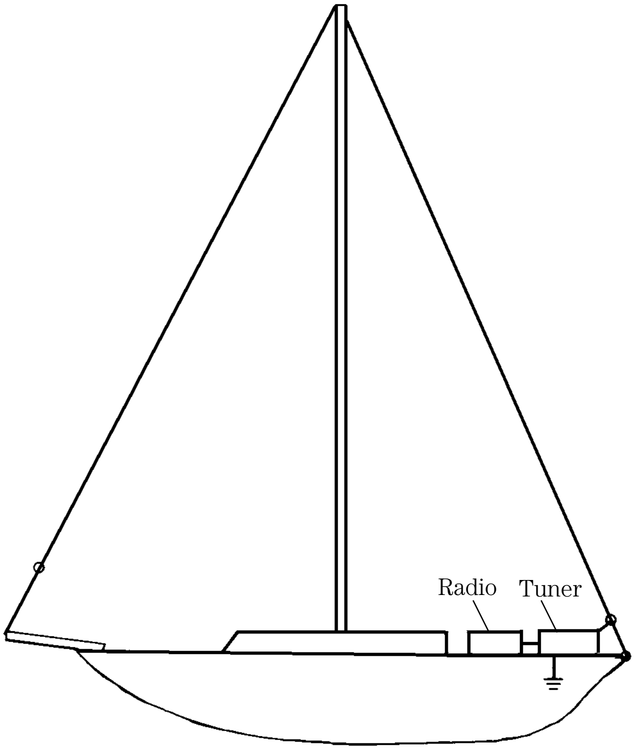
Radio
Tuner

25-13A Sailboat HF antenna.

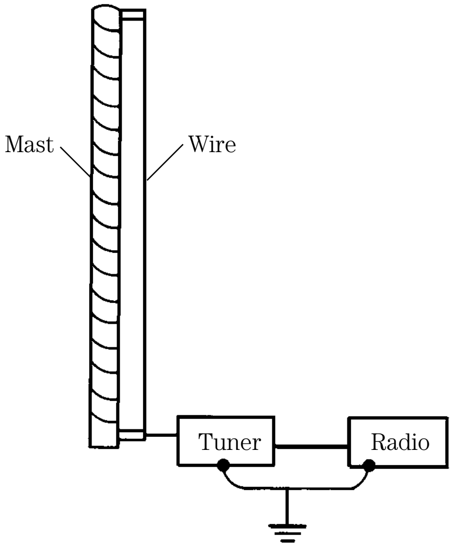
Mast
Wire
Tuner
Radio

25-13B Mast antenna.

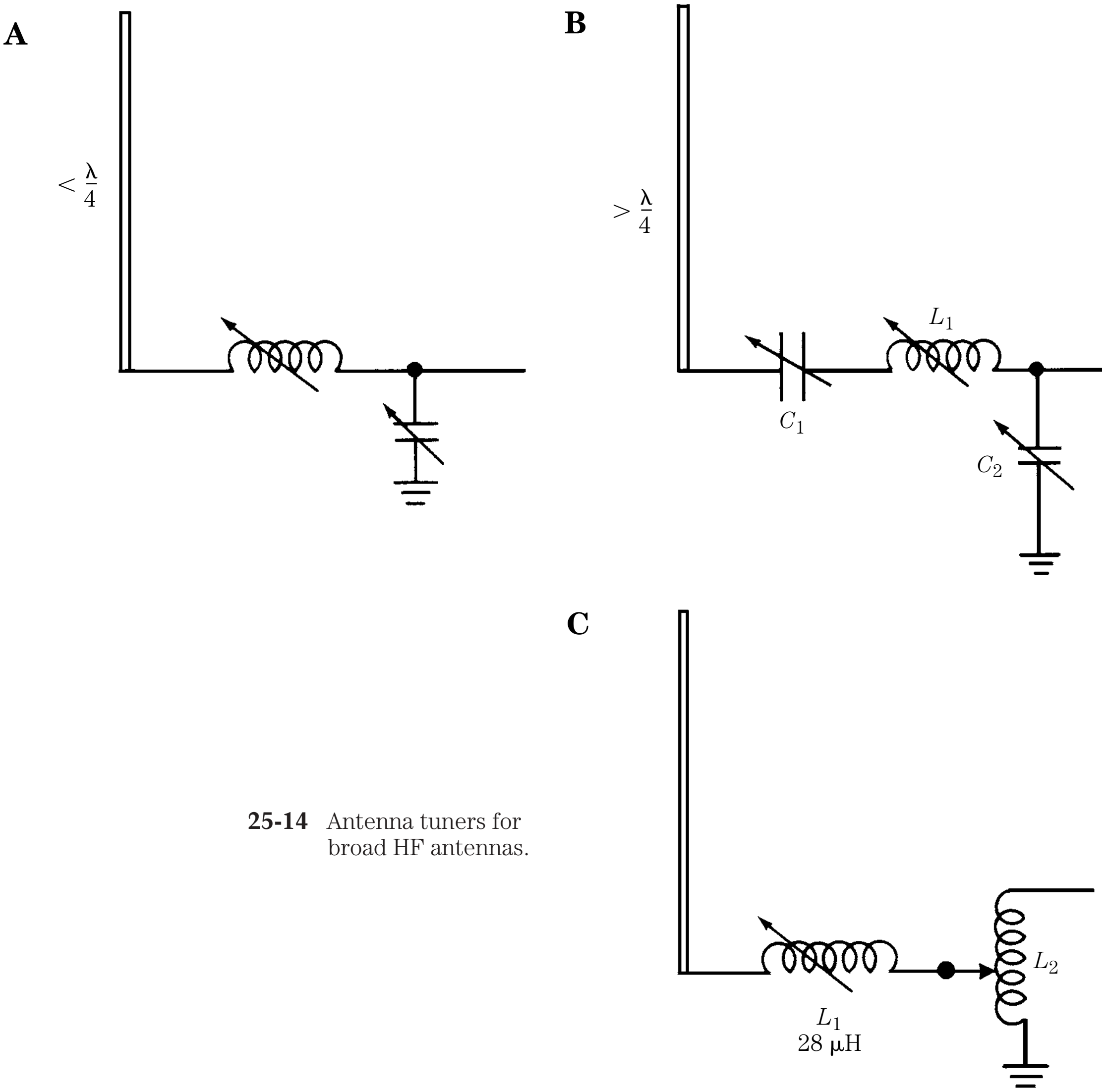

25-14 Antenna tuners for broad HF antennas.

Finally, we see in Fig. 25-14C the coupler used on many radios for random-length antennas. Two variable inductors are used; L_1 is used to resonate the antenna, and L_2 is used to match the impedance looking back to the transmitter to the system impedance. In some designs, the inductors are not actually variable, but rather use switch-selected taps on the coils. The correct coil taps are selected when the operator selects a channel. This approach is less frequently encountered today, when frequency synthesizers give the owner a selection of channels to use. In those cases, either the antenna must be tuned every time the frequency is changed, or an automatic (or motor-driven) preselected tuner is used.

The line flattener (Fig. 25-15) is a standard transmatch antenna tuner that flattens coaxial-cable transmission line for VSWR. This type of tuner is especially useful for transmitters with solid-state finals that are not “happy” with high VSWR. Some of those designs incorporate shutdown circuits that reduce (and then cut off) power as the VSWR increases. The line flattener basically tunes out the VSWR at the transmitter. It does nothing to tune the antenna, but only makes the transmitter operable.

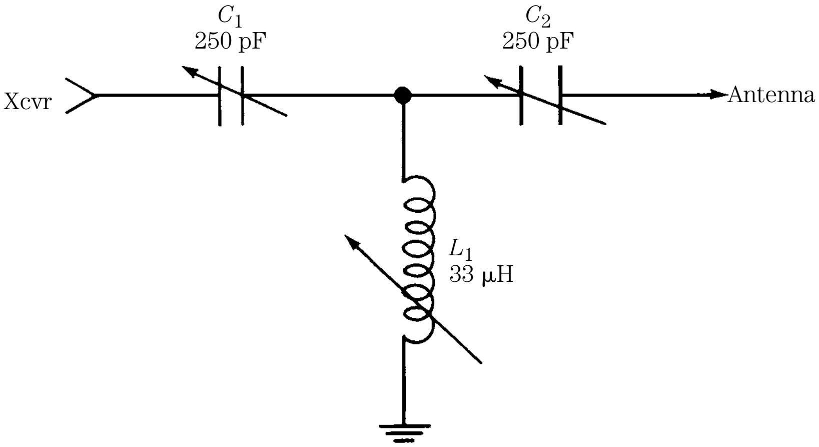

25-15 Line flattener tuner.

26
CHAPTER

Antennas for low-frequency operation

LOW-FREQUENCY OPERATION (E.G., 160- AND 75/80-METER BAND) POSES CERTAIN difficulties for the antenna. Of course, the first thing that springs to mind is the large size of those antennas. A half-wavelength dipole is between 117 and 133 ft long on 75/80 m, and on 160 m it is about twice that length. On my own suburban lot, I cannot erect a half-wave 75/80-m band antenna and stay within the property lines.

A similar situation is seen with vertical antennas. Although a 40-m vertical (33 ft high) is not an unreasonable mechanical job, the 66 ft 75/80-m vertical (never mind the 120-ft 160-m vertical) is a nightmare. In addition, the local authorities might not require any special inspections or permits (check!) on the 33-ft antenna, yet impose rigid and very exacting requirements on the higher structure. On suburban or urban lots, a typical 40-m antenna might well be able to fall over and still not cross the property line—or come in close proximity to power lines. A longer antenna, however, almost inevitably suffers one—or the other—problem when it falls.

Grounding

Still another problem involves grounding. The ground system of a higher-frequency vertical antenna can be installed using $\lambda/4$ radials buried a few inches below the surface. But that same idea becomes problematic when the radials are 66 or 125 ft long. Burying half a dozen 125-ft radials might remind your neighbors of the Galloping Gung-Ho Gopher—and not to their amusement!

There are, however, some solutions to the problem—without having you buy a farm in the flatter regions of the midwest. For example, the length problem can be solved by using one of the bent dipoles covered in Chap. 10. No, they are not

Copyright 2001 - the McGraw-Hill Companies

as effective as an antenna that is correctly installed. But they will serve to get you on the air. Some of them work surprisingly well.

The ground radial problem can be solved by bending the radial system around your property (Fig. 26-1). You will never have to cross your neighbor's line. And, as for the gopher track appearance, it is not necessary to work sloppily, and you can install radials so that nary an eye, practiced or otherwise, can see their location.

Shortened vertical antennas

The biggest problem for most low-frequency DXers, as you have seen, is the excessive size of antennas for those frequencies; it is "not for nothing" that those AM broadcast band (<1.6 MHz) towers are usually hundreds of feet tall. But there are ways to shorten an antenna—not for free, because the TANSTAAFL[*] principle still

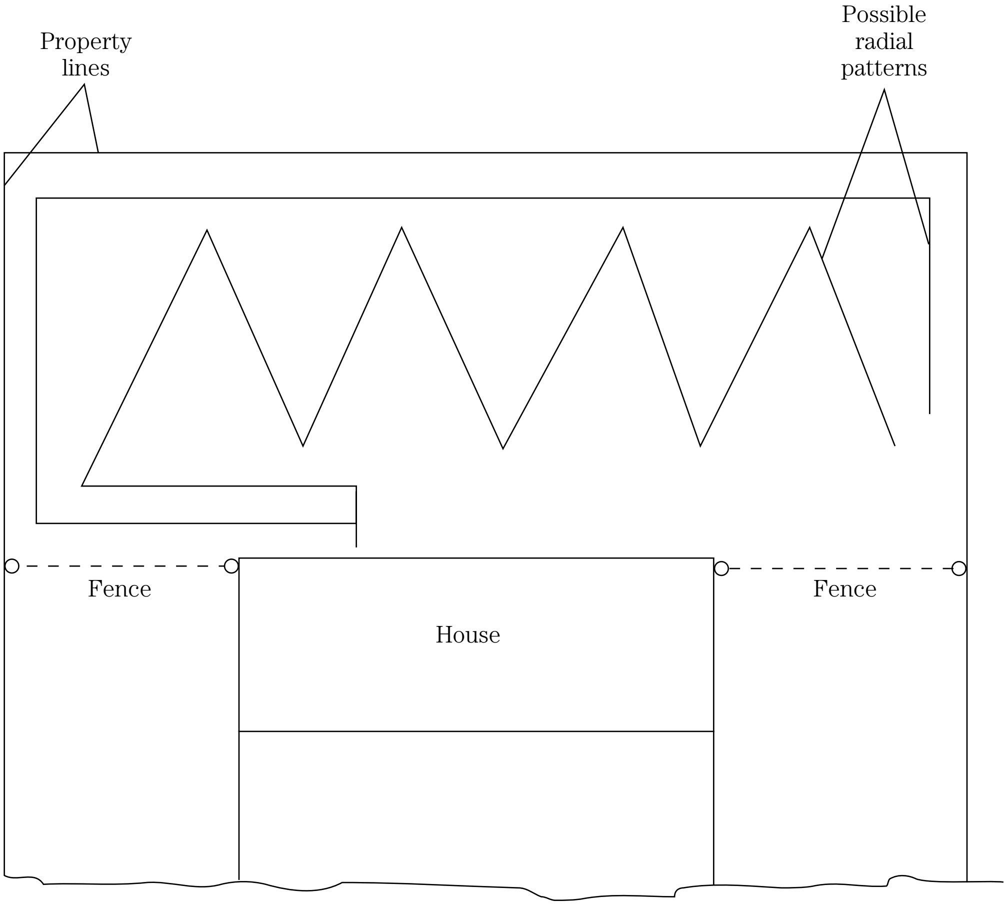

26-1 Methods for laying out low-frequency radials in a cramped space.

---

[*] TANSTAAFL = There ain't no such thing as a free lunch!

applies—to a point where it becomes mechanically possible. Let me reiterate once again that these compensation antennas will not work as well as a properly installed full-sized antenna, but they will serve to get you on the air on frequencies where it is otherwise utterly impossible. Several different compensation configurations are popular, and these are shown in Fig. 26-2.

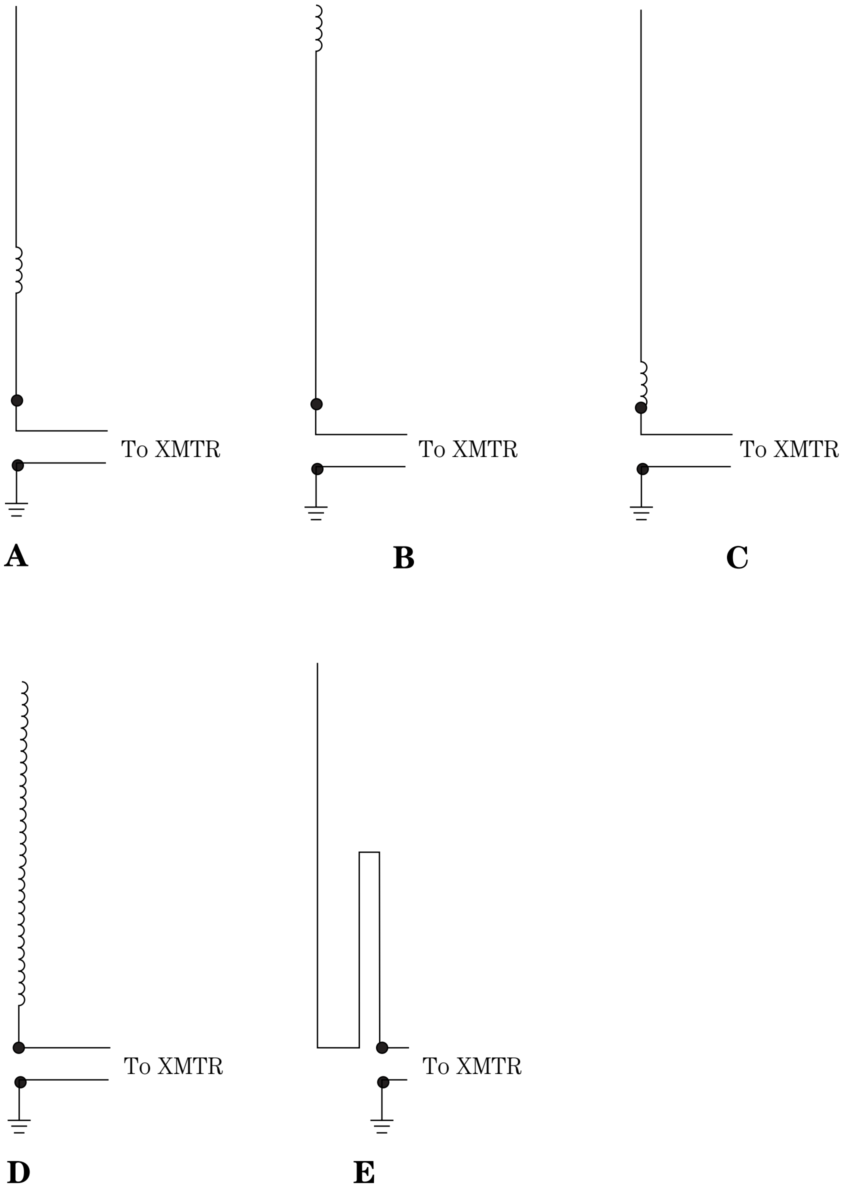

26-2 Inductance loading of vertical antennas: (A) center; (B) top; (C) base; (D) linear or length-loaded; (E) linear hairpin-loaded.

The basic foundation for these antennas is a very short vertical antenna. The "standard" vertical antenna is a quarter-wavelength (λ/4) (i.e., 90° electrical length), and it is unbalanced with respect to ground.

Recall that a vertical antenna that is too short for its operating frequency (i.e., less than λ/4) will exhibit capacitive reactance. In order to resonate that antenna, it is necessary to cancel the capacitive reactance with an equivalent inductive reactance, such as $|X_L| = |X_C|$. By placing an inductance in series with the antenna radiator element, therefore, we can effectively "lengthen" it electrically. Of course, what is really happening is that the effects of the lower operating frequency are being accommodated (i.e., "cancelled out") of a too-short antenna.

An antenna that is reactance-compensated for a different frequency is said to be *loaded*, or in the case of a very low frequency antenna, *inductively loaded*. Three basic forms of loading are popular: *discrete loading, continuous loading,* and *linear loading*.

Discrete loading means that there is a discrete, or lumped, inductance in series with the antenna radiator (Figs. 26-2A through 26-2C). These antennas are so constructed that a loading coil is placed at the center (Fig. 26-2A), top (Fig. 26-2B), or bottom (Fig. 26-2C) of the radiator element.

You will recognize these configurations as being the same as those found on mobile antennas. Indeed, low-band mobile antennas can be used in both mobile and fixed installations. Note, however, that although it is convenient to use mobile antennas for fixed locations (because they are easily available in "store bought" form); they are less efficient than other versions of the same concept. The reason is that the mobile antenna, for low frequencies, tends to be based on the standard 96- to 102-in whip antenna used by amateur operators on 10 m or Citizens Band operators on 11 m. In fixed locations, on the other hand, longer radiator elements (which are more efficient) are more easily handled. For example, a 16- to 30-ft high aluminum radiator element can easily be constructed of readily available materials. The 16-ft element can be bought in the form of one or two lengths of 1- to 1.5-in aluminum tubing at do-it-yourself outlets.

A problem seen with these antennas is that they tend to be rather high Q, so the bandwidth is necessarily narrow. An antenna might work in the center of a band, but present a high VSWR at the ends of the band, and thus be unusable. This problem is solved by making the inductor variable so that slightly different inductance values can be selected at different frequencies across the band. For the base-loaded version (Fig. 26-2C), this is particularly easy: a rotary inductor (perhaps motor driven for remote operation) can be used. For the other configurations, a tapped fixed inductor can be used instead. Each tap represents a different inductance value. Either clip connectors, or relay connections, can be used to select which tap is used.

A continuously loaded antenna has the inductance distributed along the entire length of the radiator (Fig. 26-2D). Typical of these antennas is the helically wound verticals in which about a half-wavelength of insulated wire is wound over an insulating form (such as a length of PVC pipe or a wooden dowel); the turns of the coil are spread out over the entire length of the insulated support.

Linear loading (Fig. 26-2E) is an arrangement whereby a section of the antenna is folded back on itself like a stub. Antennas of this sort have been successfully built from the same type of aluminum tubing as regular verticals. For 75/80-m (3.75-MHz) operation a length of 30 ft for the radiator represents 41°, while a normal λ/4 vertical is 90°. The difference between 90° and 41° of electrical length is made up by the "hairpin" structure at the base.

Shortened horizontal antennas

The same strategies that worked for vertical antennas also work for horizontal antennas, although in the horizontal case we are simulating a half wavelength (180°) balanced antenna—rather than a λ/4 unbalanced antenna. Figure 26-3 shows several different shortened, low frequency antennas that are based on the same methods as the verticals shown previously.

Figures 26-3A through 26-3C show discretely loaded dipole antennas. Figures 26-3A and 26-3B are center-loaded versions, and Fig. 26-3C is a center (of each element) -loaded version. In each case, it is assumed that the radiator elements are the same physical length. In the example of Fig. 26-3A, the coil is tapped to provide a match to either 52- or 75-Ω transmission line. In some variants, the shield of the coaxial-cable transmission line is connected directly to the junction of the coil and one radiator element, and the coax center conductor is connected to the tap that best matches the impedance of the line.

The impedance-matching problem is solved a little differently in Fig. 26-3B. In this type of antenna, the coil is used to center-load the dipole, but the transmission line is connected to a link wound on the same form as the loading inductor. The turns ratio between the loading inductor and the coupling link determines the impedance match.

A "hairpin" linear loading scheme is shown in Fig. 26-3D. This design is basically the same as for the vertical case, but it is balanced out of respect for the design of the dipole. This type of design is used for both the driven and parasitic elements on some commercial 40-m beam antennas. In addition, some "add-on" 40-m dipoles (designed for beam antennas that were intended for operation at frequencies in the 20-m and higher bands) use this method. Such antennas are essentially rotatable dipoles.

The continuously loaded dipole of Fig. 26-3E is constructed like the helical vertical (i.e., about a half-wavelength of insulated wire is wound over the entire length of an insulated rod or pipe of some sort). The winding can be either broken at its center point to accommodate the feedline, or link-coupled to the transmitter, or receiver, as shown in Fig. 26-3F. Of the two feed methods, the most popular appears to be the type that breaks the winding into two pieces.

The form of continuously loaded dipole in Fig. 26-3E has a combination of the two methods of feed. The distributed loading coil is broken into two sections, as is often done on continuously loaded dipoles. But in this case, a portion of the overall inductance is made by using a discrete inductor that is part of a toroidal transformer T_1. The length of the wire used for the continuous loading coil is somewhat shorter (find by experimentation) than otherwise would be the case because of the inductance of the transformer secondary. Transformer T_1 can be wound using the same sort of toroidal ferrite or powdered iron core used for Balun transformers.

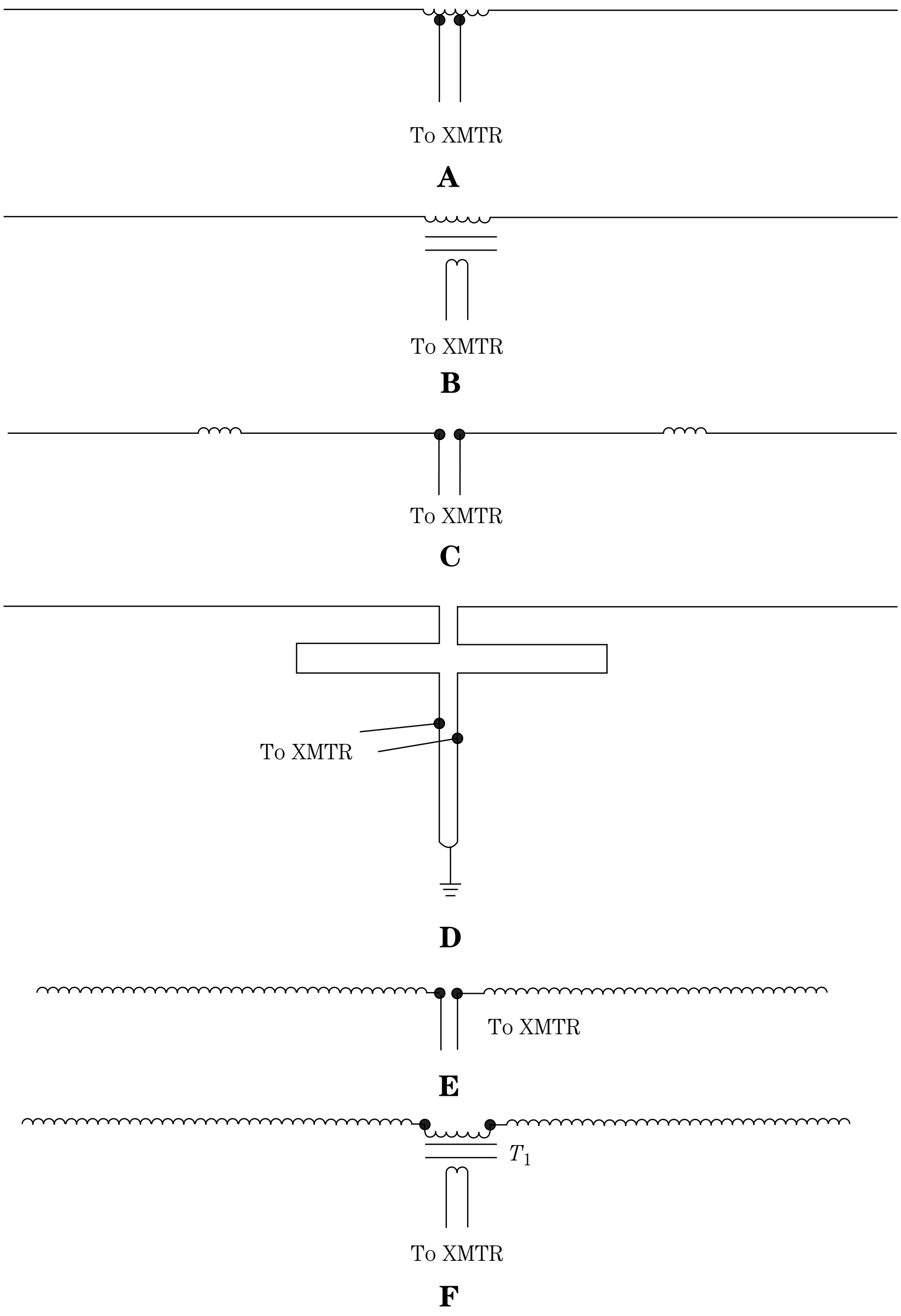
To XMTR
A
To XMTR
B
To XMTR
C
To XMTR
D
To XMTR
E
T_1
To XMTR
F

26-3 Inductance loading of a dipole: (A) center loading with impedance tapped coil; (B) center loading with transformer coupling; (C) loading at center of elements; (D) linear loading (hairpin); (E) linear or length loading; (F) transformer coupling.

Loaded tower designs

Many amateur radio stations and other services use a rotatable beam antenna on top of a tower. Typical towers are from 30 to 120 ft in height. In my locality, towers to 30 ft, or less than 3 ft above the roof line, whichever is taller, can be erected without a permit as long as they are attached to the house and will not fall onto a power line, or across the property line, in a catastrophic failure. Higher towers must be erected under a mechanical permit, and properly inspected by the county. These towers typically support a two-element (or more, three is popular) Yagi /Uda beam antenna, or a quad, or some other highly directional antenna. But they can also be treated as a vertical antenna under the right circumstances.

If the tower is close to 66 ft high (which is a popular height for amateur radio towers), then the tower (already $\lambda/4$) can be used as a resonant vertical on 75/80 m. The same tower can also be used on lower frequencies if proper antenna tuning unit (ATU) components are provided.

Figure 26-4 shows a situation in which an 80- to 110-ft tower is insulated from ground, and is considered a random-length vertically polarized Marconi antenna. Like other such antennas, it is fed by a simple L-section, or reverse L-section, ATU. Note the approximate values of the inductors and capacitors in each configuration, and also their relationships. On the 40- and 75/80-m versions, the tower is too long, so a series capacitance and shunt inductance are used for the ATU (i.e., Reverse L-section coupler). On the other hand, the tower is too short for 160 m, so the inductor and capacitor are reversed.

There is only one small problem with the design of Fig. 26-4: Compared with grounded towers, insulated towers are expensive to buy and install. In addition, lightning protection is probably better in the grounded tower. So how do we work a grounded tower? See Fig. 26-5. In this antenna, the tower is grounded at its base. If the tower is mounted to a concrete pedestal (the usual arrangement), then a separate ground rod and ground wire adjacent to the pedestal must be provided. Concrete is not a good insulator, but it is also not a good conductor. A *delta feed* system is used with this antenna. That is, a single wire from the ATU is connected to a point on the antenna where an impedance match can be achieved. It is important that the ATU be spaced away from the tower base (as shown), and the wire must be run straight to the feedpoint on the antenna. This system is an example of several possible *shunt feeding* systems.

Random-length Marconi

One of the old standbys for all bands is the random-length Marconi antenna. Such antennas consist of a length of wire, typically (but not always) less than $\lambda/4$, and fed at one end with coaxial cable and an L-section coupler (Fig. 26-6). The antenna radiator element can be angled in any direction as needed, but it works best if the radiator is either as horizontal, or vertical, as possible (for pure polarization). The usual situation, however, is to run the wire at an inclined angle, or with about equal portions horizontal and vertical (see the inverted L) which yields complex polarization.

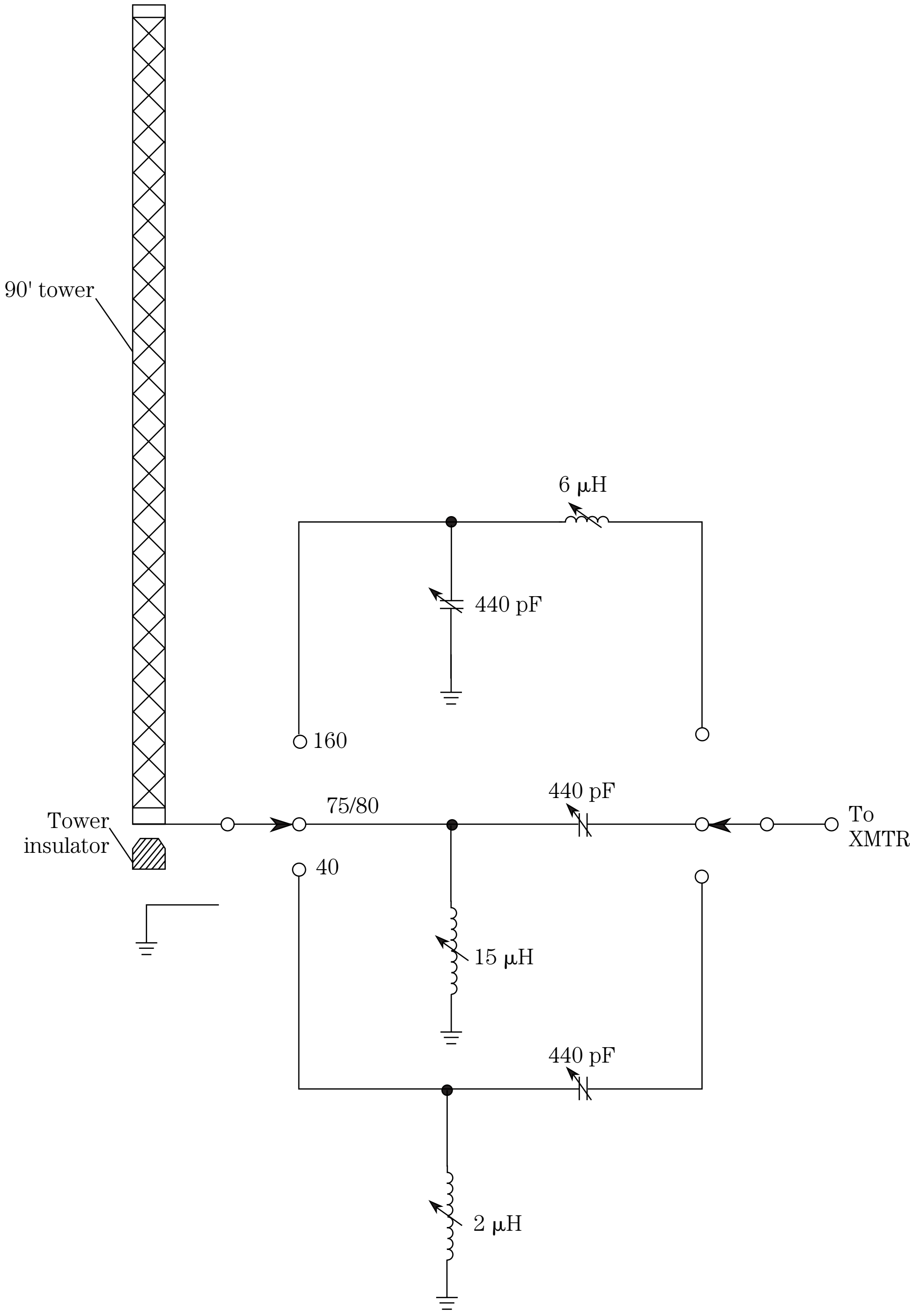
90' tower
6 µH
440 pF
160
75/80
440 pF
To
XMTR
Tower
insulator
40
15 µH
440 pF
2 µH

26-4 Method for impedance matching a tower antenna.

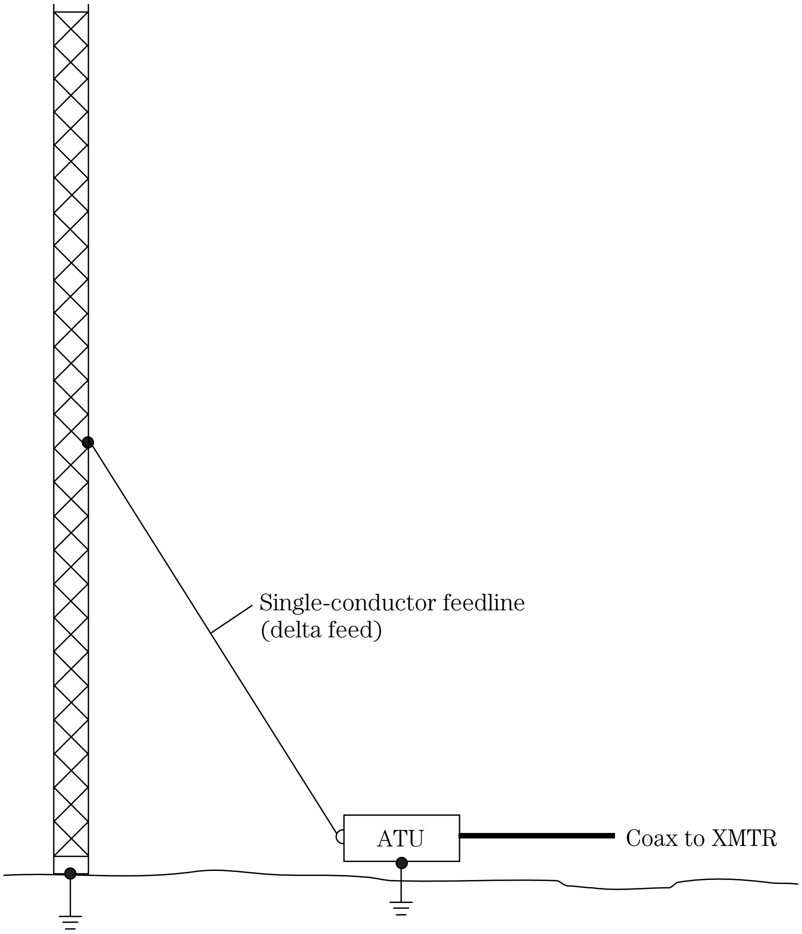
Single-conductor feedline
(delta feed)
ATU
Coax to XMTR

26-5 Delta matching a grounded vertical tower antenna.

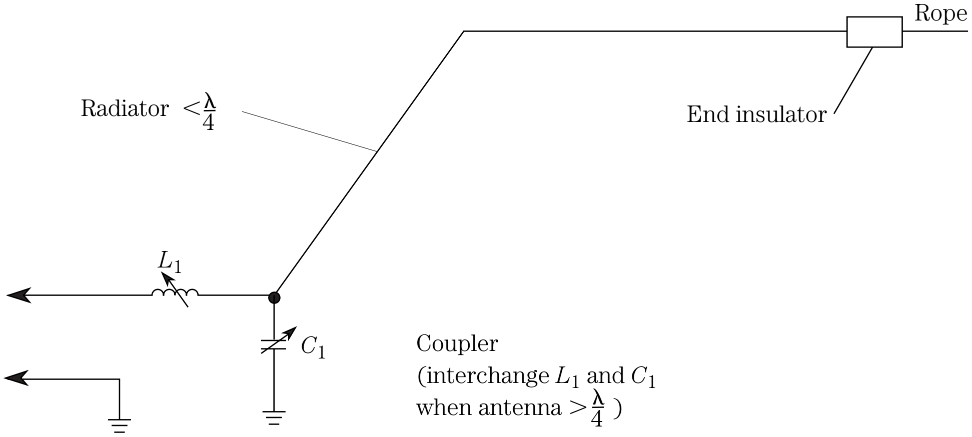
Rope
Radiator $<\frac{\lambda}{4}$
End insulator
L_1
C_1
Coupler
(interchange L_1 and C_1
when antenna $>\frac{\lambda}{4}$)

26-6 Random-length wire antenna.

The L-section coupler shown in Fig. 26-6 is set up for the case where the antenna radiator element is less than λ/4 (i.e., a series inductor and a shunt capacitor). If the antenna is longer than λ/4 on some frequency, then reverse the positions of the capacitor and inductor.

Inverted-L antennas

Another popular antenna for low frequencies is the *λ/4 inverted-L* (Fig. 26-7). In this type of antenna, two sections are erected at a 90° angle with respect to each other; one vertical and the other horizontal. One way to think of this antenna is bent λ/4 vertical, although some people liken it to a top-loaded vertical. The feedline can be 52- or 75-Ω coaxial cable.

It is generally the case that the sections of the inverted-L are equal in length (λ/8 each), but that is not strictly necessary. As the vertical section becomes longer (with overall length remaining at λ/4), the angle of radiation depresses.

One popular method of construction is to use a tower for the vertical section, and a run of wire for the horizontal section. If you already have a 60-ft tower to accommodate the beam antenna used on higher frequencies, then it is relatively easy to build an inverted-L antenna for 160 m.

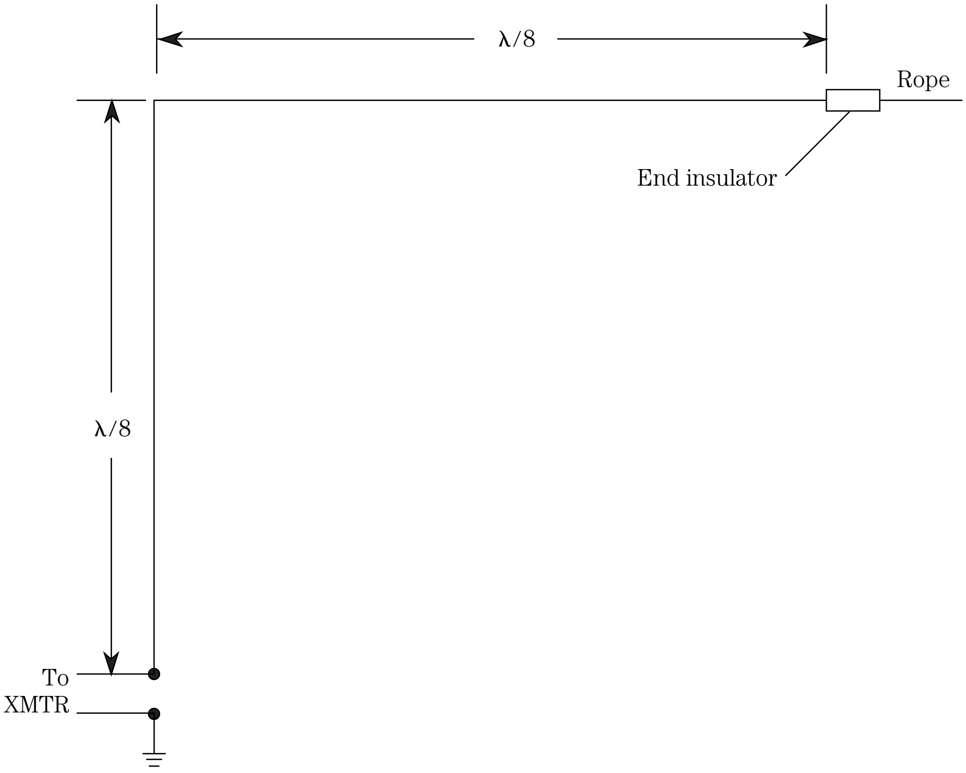

26-7 L-section quarter-wavelength antenna.

The linearly loaded "tee" antenna

A "tee" antenna consists of a horizontal radiator element fed at the center with a single conductor wire (not coax) that is a quarter-wavelength long, and goes away from the radiator at a right angle. Such antennas were popular prior to World War II in the United States, although popularity fell off in the late 1930s as other types (and reasonably priced transmission lines) became available. The tee is possibly one of the oldest forms of radio transmitting antennas. Like other antennas at low frequencies, however, it is far too long for easy use in most locations.

We can, however, linearly load the tee by folding the ends of the radiator back on themselves to form the zigzag pattern of Fig. 26-8. Popular in Europe for some time, this antenna can provide reasonable performance on the lower frequencies without using too much horizontal space.

The radiator element consists of three sections that are each about $\lambda/6$ long ($L_1 = 164/F_{\mathrm{MHz}}$), and spaced 8 to 12 in apart; they are parallel to each other. The quarter-wavelength transmission line is connected to the center point of the middle section of the radiator element.

Figure 26-9 shows how the linearly loaded tee antenna could be built using wire for the radiator elements. Spreaders made of 24-in-long 1 × 2-in lumber (treated against the weather), plastic, or some other synthetic material can be used. Each spreader has three holes drilled in it spaced about 10 in apart. Spreaders A and E are end spreaders, and are identical—except for being upside down with respect to each other. In each case, one of the radiator element conductor sections is terminated, while the other passes through to the back of the spreader to join the center radiator element. The other spreaders are used either for center support, as in C (note the transmission line attachment), or for interim support between the center insulator and the end insulators.

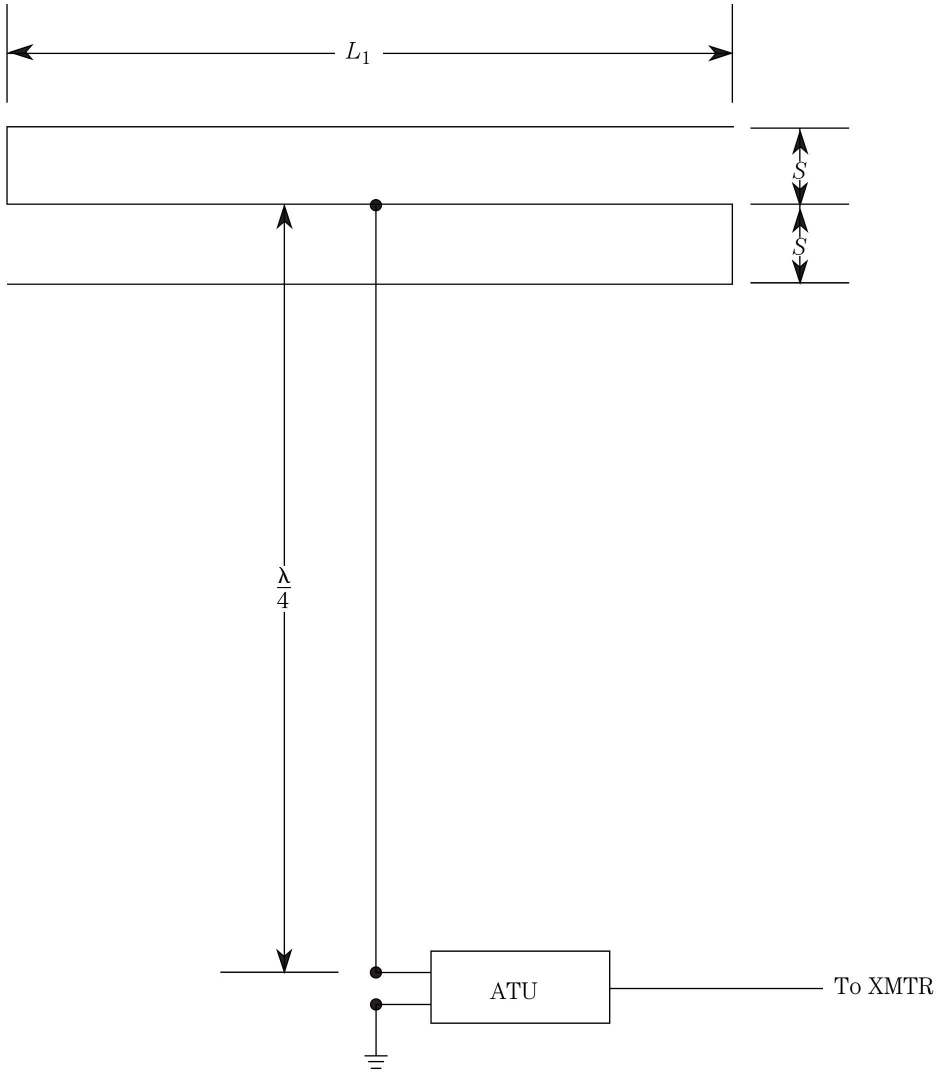
L_1
S
S
$\frac{\lambda}{4}$
ATU
To XMTR

$$L_1 = \frac{164}{F_{\mathrm{MHz}}}$$

$$\frac{\lambda}{4} = \frac{246}{F_{\mathrm{MHz}}}$$

26-8 Tee-loaded antenna.

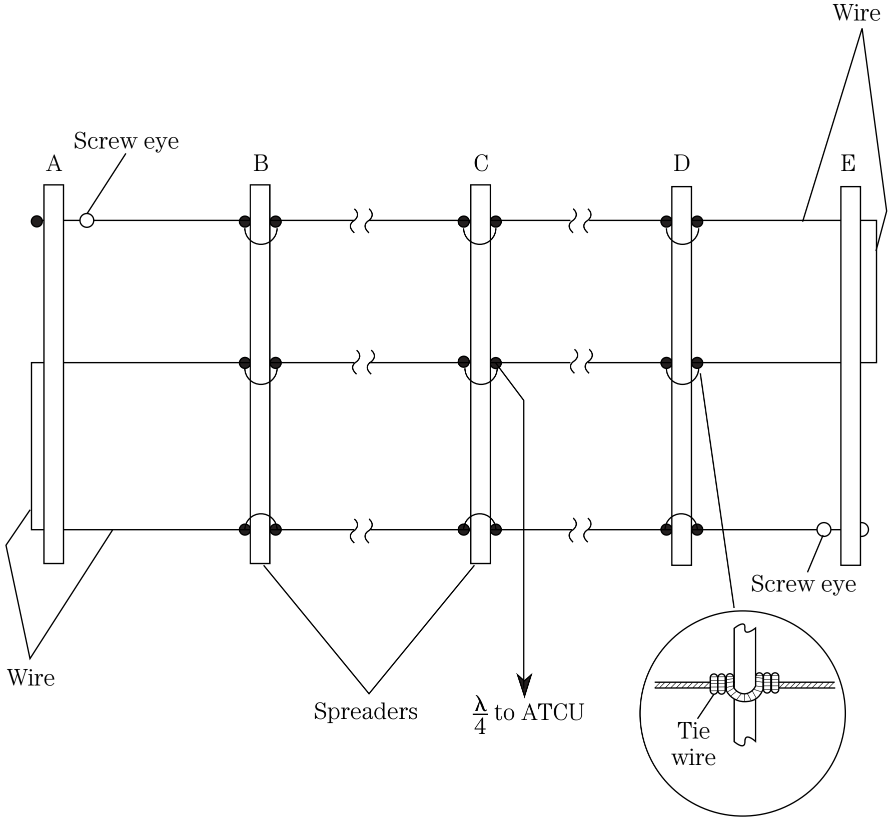
Wire
Screw eye
A
B
C
D
E
Screw eye
Wire
Spreaders
$\frac{\lambda}{4}$ to ATCU
Tie wire

26-9 Shortened tee antenna.

This page intentionally left blank

27
CHAPTER

Measurements and adjustment techniques

THIS CHAPTER EXAMINES SOME OF THE INSTRUMENTS AND TECHNIQUES FOR TESTING antenna systems, whether brand new installations, or in troubleshooting situations on older antenna. The basic radio system is shown in Fig. 27-1. We have several elements in the system: transmitter, low-pass filter, impedance-matching unit, a coaxial relay if the receiver is separate from the transmitter, and the antenna. Connecting these elements are lengths of transmission line. In most modern radio systems below the microwaves, the transmission line is coaxial cable.

The low-pass filter and matching unit may be considered optional by some, but some designers regard them as essentially standard equipment, especially for HF and low-band VHF systems. The low-pass filter has the job of removing harmonics from the output signal that could interfere with other radio systems or radio and TV broadcast reception. It will pass only those frequencies below a certain cutoff point. The impedance matching unit is used to "tune out" impedance mismatches that cause the transmitter to see an excessive VSWR. These units are common in VLF, MW, and HF communications stations, as well as AM broadcasting stations. The matching unit also provides additional attenuation of the harmonics, so it makes the output even cleaner than is possible with the low-pass filter alone.

Another reason to use the matching unit is to allow the radio transmitter to put out the maximum allowable RF power. Modern solid-state final amplifiers are not tolerant of VSWR. In addition, these units use fixed-tuned low-pass filters for each band rather than the wide-range pi networks common on vacuum tube transmitters. Solid-state transmitters usually include a feedback *automatic load control* (ALC) that reduces output power when a high VSWR is sensed. The VSWR cut-in knee begins around 1.5:1 and completely shuts off the transmitter when the VSWR gets high (typically above 2.5:1 or 3:1). The external T/R relay is not used on most modern systems because the receiver and transmitter are typically housed inside of the same box. The coaxial relay was used in days past, when

Copyright 2001 - the McGraw-Hill Companies

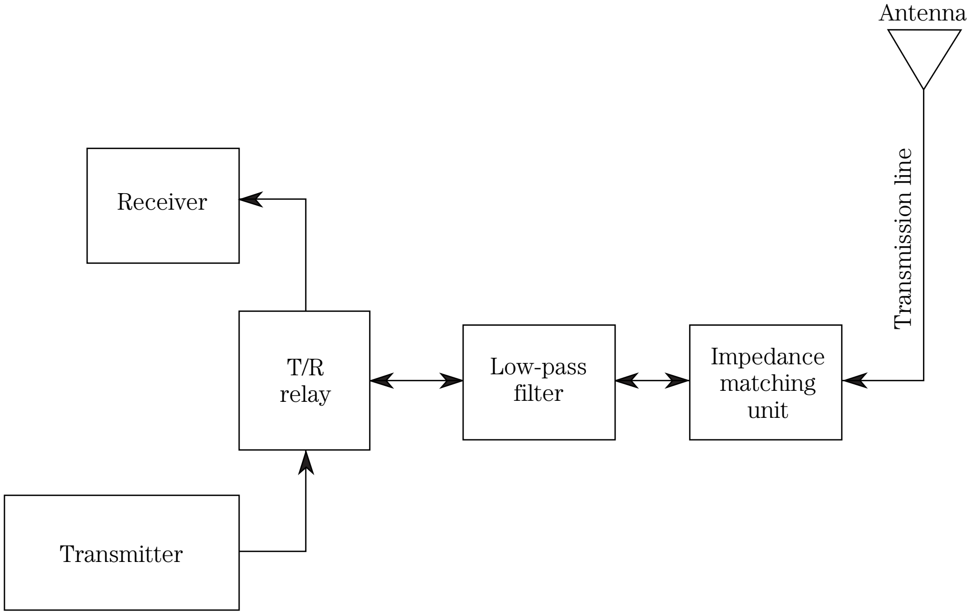

27-1 Basic communications radio station setup.

a separate receiver needed to use the same antenna. In addition, many modern solid-state equipments use PIN diode T/R switching.

Transmission lines

The transmission line is not merely a wire that carries RF power to the antenna. It is actually a complex circuit that simulates an infinite *LC* network. There is a characteristic impedance Z_o, also called *surge impedance*, which describes each transmission line. This impedance is the square root of the ratio of the capacitance and inductance per unit of length. When a load having a "resistive-only" impedance equal to the surge impedance of the transmission line is connected, then we will see maximum transfer of power between the line and the antenna.

We cannot deal extensively with transmission-line theory here, and refer the reader instead to previous chapters. We must, however, have at least some idea of what the circuit looks like. Figure 27-2 shows a model of a transmission line in which Z_o is the surge impedance of the line, R_1 is the load impedance of the antenna, and R_2 is the output impedance of the transmitter. In a properly designed system all three impedances either will be equal ($Z_o = R_1 = R_2$) or a matching network will make them equal.

We must consider the electrical situation along the transmission line in order to understand the readings that we see on our instruments. Figure 27-3 shows several possible situations. These graphs are of the RF voltage along the line, with voltage on the vertical axis and transmission line length (expressed in wavelengths of the RF signal) along the horizontal axis. When the system is matched ($Z_o = R_1$), the voltage

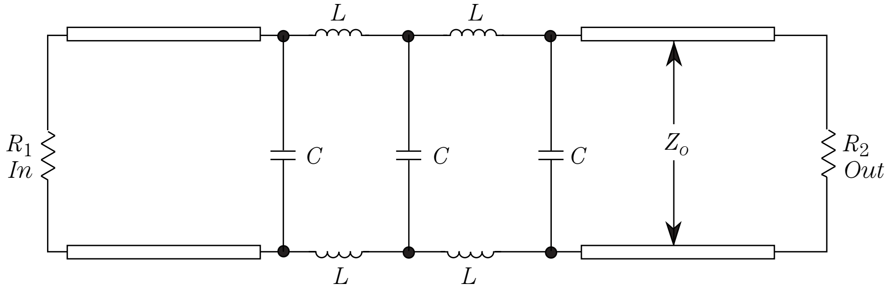

L = Inductance per unit length
C = Capacitance per unit length

27-2 Equivalent circuit of transmission line.

is the same everywhere along the line (Fig. 27-3A). This line is said to be "flat." But when Z_o and R_2 are not equal, then the voltage varies along the line with wavelength. In mismatched systems, not all of the power is radiated by the antenna, but rather is reflected back to the transmitter. The forward and reflected waves combine algebraically at each point along the line to form standing waves (Fig. 27-3B). We can plot the voltage maxima ($V_{\max}$) and minima ($V_{\min}$). Keep this graph in mind for a few minutes because you will refer again to it when we deal with VSWR.

Two special situations occur in transmission line and antenna systems that yield similar results. The entire forward power is reflected back to the transmitter (none radiated) if the load (i.e., antenna) end of the transmission line is either open or shorted. The voltage plot for an open transmission line (R_2 is infinite) is shown in Fig. 27-3C, and that for the shorted line is shown in Fig. 27-3D. Note that they are very similar to each other except for where the minima ($V_{\min} = 0$) occur. The minima are offset from each other by 90° (i.e., quarter-wavelength).

Calculating standing wave ratio

The VSWR can be calculated from any of several bits of knowledge. Even if you don't have a VSWR meter, therefore, it is possible to determine VSWR. If the antenna load impedance (R_2) is not equal to Z_o, then we can calculate VSWR from one of the following:

If Z_o is greater than R_2:

$$\text{VSWR} = \frac{Z_o}{R_2} \quad \textbf{[27.1]}$$

If Z_o is less than R_2:

$$\text{VSWR} = \frac{R_2}{Z_o} \quad \textbf{[27.2]}$$

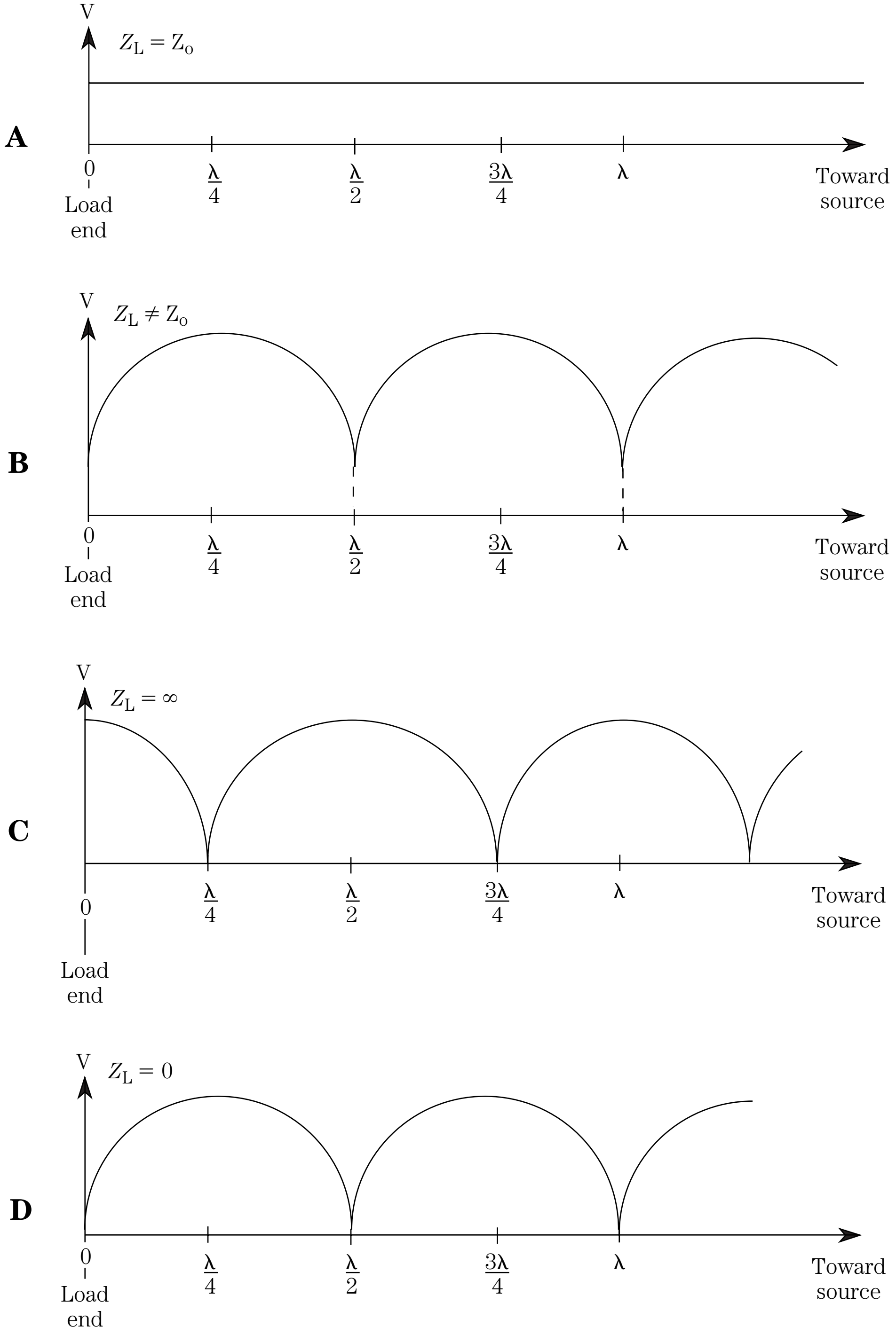

27-3 Voltage along line versus line electrical length: (A) flat line with matched impedance; (B) unequal impedances; (C) open load; (D) shorted load.

We can also measure the forward and reflected power, and calculate the VSWR from those readings:

$$\text{VSWR} = \frac{1 + \sqrt{(P_r/P_f)}}{1 - \sqrt{(P_r/P_f)}} \qquad \textbf{[27.3]}$$

where

VSWR is the voltage standing wave ratio
P_r is the reflected power
P_f is the forward power

If we can measure either the voltage maximum and minimum, or the current maximum and minimum, then we can calculate SWR:

$$\text{VSWR} = \frac{V_{\text{max}}}{V_{\text{min}}} \qquad \textbf{[27.4]}$$

Finally, if the forward and reflected voltage components at any given point on the transmission line can be measured, then we can calculate the VSWR from:

$$\text{VSWR} = \frac{V_f + V_r}{V_f - V_r} \qquad \textbf{[27.5]}$$

where

V_f is the forward voltage component
V_r is the reflected voltage

The last equation, based on the forward and reflected voltages, is the basis for many modern VSWR and RF power meters.

Impedance bridges

We can make antenna impedance measurements using a variant of the old-fashioned Wheatstone bridge. Figure 27-4A shows the basic form of the bridge in its most generalized form. The current flowing in the meter will be zero when $(Z_1/Z_2) = (Z_3/Z_4)$. If one arm of the bridge is the antenna impedance, then we can adjust the others to make the bridge null to make the measurement. A typical example is shown in Fig. 27-4B. The antenna connected to J_2 is one arm of the bridge, while R_2 is a second. The value of R_2 should be 50 Ω or 75 Ω, depending upon the value of the expected antenna impedance. The choice of 68 Ω is a good compromise for meters to operate on both types of antennas. The other two arms of the bridge are the reactances of C_{1A} and C_{1B}, which is a single differential capacitor. Tune C_1 until the meter is nulled, and then read the antenna from the dial. At least one instrument allows the technician to plug in a resistor element equal to system impedance.

Calibrating the instrument is simple. A series of noninductive carbon composition resistors having standard values from 10 to 1000 Ω are connected across J_2. The meter is then nulled, and the value of the load resistor is inscribed on the dial at the point.

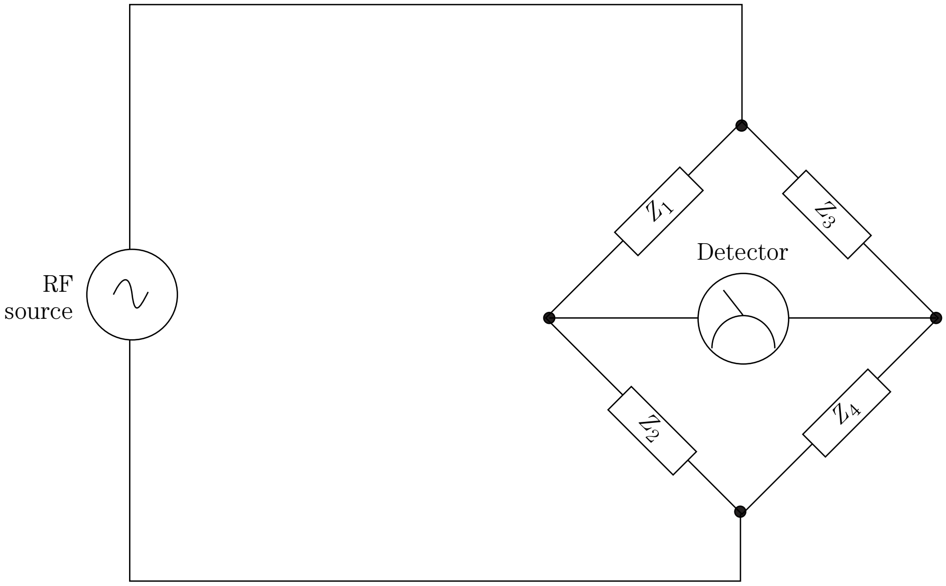

RF
source
Detector
Z_1
Z_2
Z_3
Z_4

27-4A Basic Wheatstone bridge.

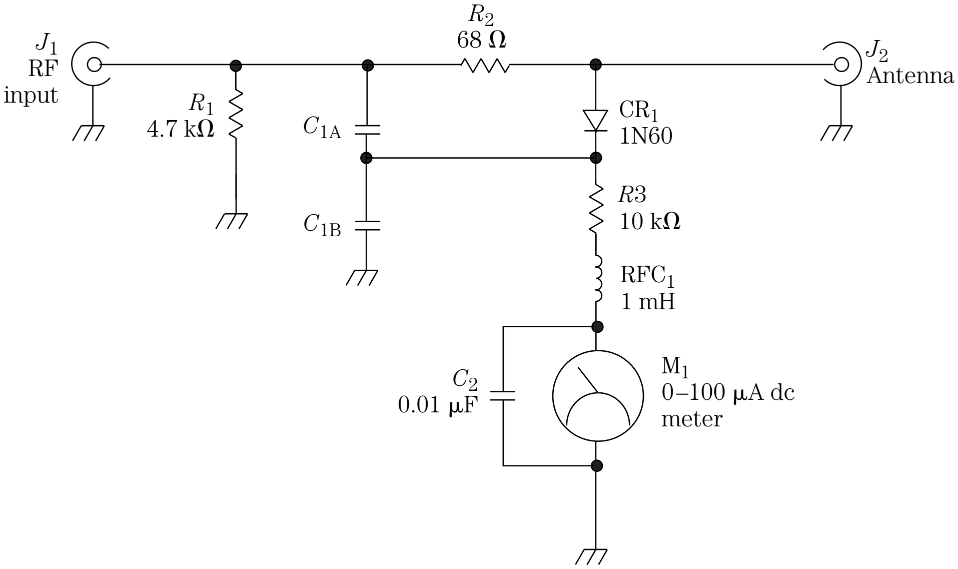

J_1
RF
input
R_1
4.7 kΩ
C_{1A}
C_{1B}
R_2
68 Ω
CR_1
1N60
$R3$
10 kΩ
RFC_1
1 mH
C_2
0.01 µF
M_1
0–100 µA dc
meter
J_2
Antenna

27-4B SWR bridge.

The basic circuit of Fig. 27-4B is useful only to measure the resistive component of impedance. We can modify the circuit as shown in Fig. 27-4C to account for the reactive component. An example of a commercial hand-held radio antenna impedance bridge is shown in Fig. 27-5.

The RF noise bridge

This section explores a device that was once associated only with engineering laboratories, but turns out to have applications in general communications servicing as well: the *RF noise bridge*. It is one of the most useful, low-cost, and often overlooked test instruments in our armamentarium.

Several companies have produced low-cost noise bridges: Omega-T, Palomar Engineers, The Heath Company, and, most recently, MFJ. The Omega-T and the Palomar Engineers models are shown in Fig. 27-5. The Omega-T device is a small cube with minimal dials, and a pair of BNC coax connectors (*antenna* and *receiver*). The dial is calibrated in ohms, and measures only the resistive component of impedance. The Palomar Engineers device is a little less eye appealing, but does everything the Omega-T does; in addition, it allows you to make a rough measurement of the reactive component of impedance. The Heath Company added their Model HD-1422 to the lineup. And the latest entry is the MFJ-204B antenna bridge.

Over the years, some people have found the noise bridge very useful for a variety of test and measurement applications, especially in the HF and low-VHF regions.

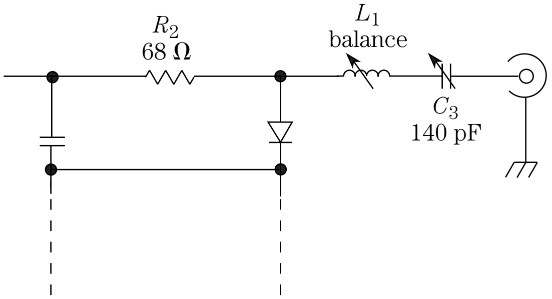

27-4C Circuit with balance control added.

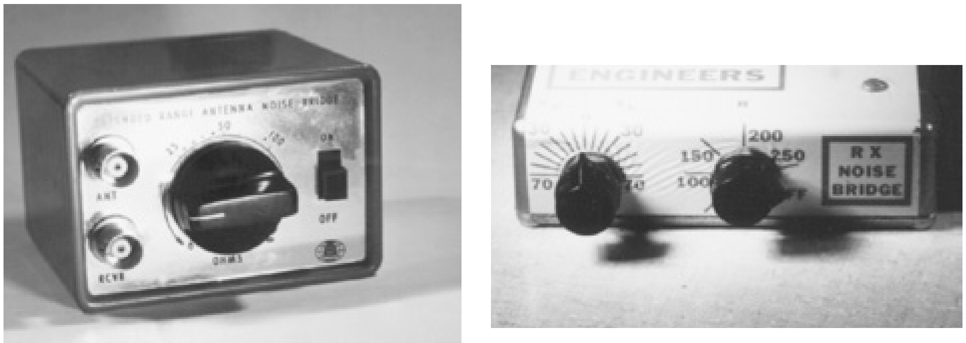

27-5 Commercially available amateur noise bridges.

Those applications are not limited to the testing of antennas (which is the main job of the noise bridge). In fact, although the two-way technician (including CB) or amateur radio operator, will measure antennas, tuned circuits, and resonant cavities with the device, consumer electronics technicians will find other applications.

Figure 27-6 shows the circuit of a noise bridge instrument. The bridge consists of four arms. The inductive arms (L_{1b} and L_{1c}) form a trifilar wound transformer over a ferrite core with L_{1a}, so a signal applied to L_{1a} is injected into the bridge circuit. The measurement consists of a series circuit of a 200-**Ω** potentiometer and a 120-pF variable capacitor. The potentiometer sets the range (0 to 200 **Ω**) of the resistive component of measured impedance, while the capacitor sets the reactance component. Capacitor C_2 in the unknown arm of the bridge is used to balance the measurement capacitor. With C_2 in the circuit, the bridge is balanced when C is approximately in the center of its range. This arrangement accommodates both inductive and capacitive reactances, which appear on either side of the "zero" point (i.e., the midrange capacitance of C). When the bridge is in balance, the settings of R and C reveal the impedance across the unknown terminal.

A reverse-biased zener diode (zeners normally operate in the reverse-bias mode) produces a large amount of noise because of the avalanche process inherent in zener operation. Although this noise is a problem in many other applications, in a noise bridge it is highly desirable: the richer the noise spectrum the better the performance. The spec-

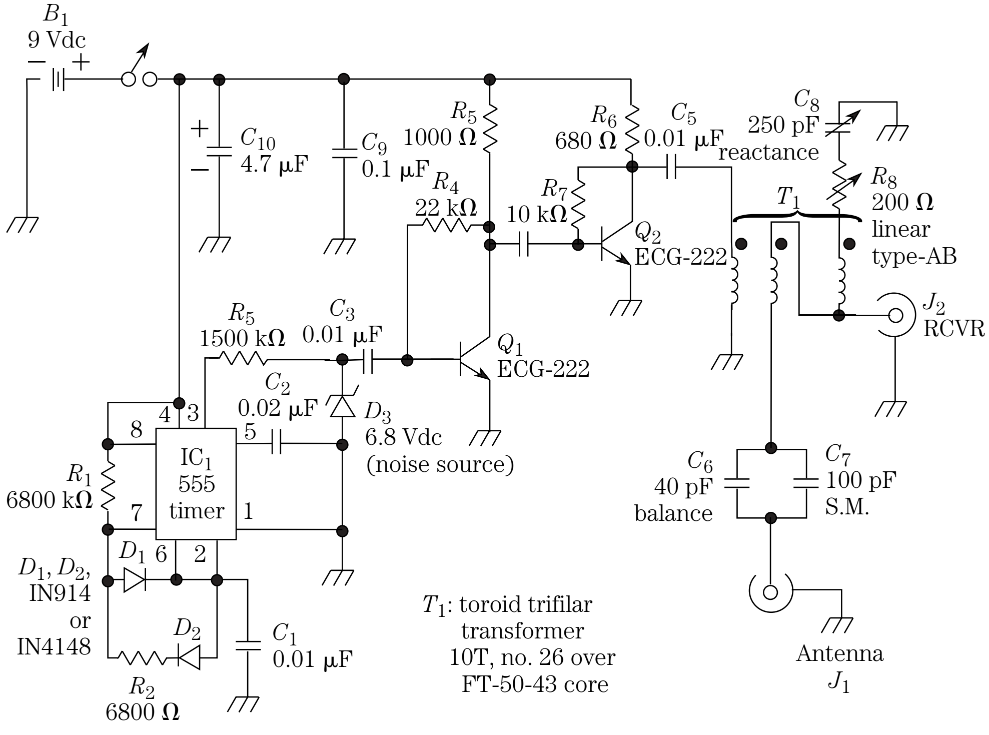
B_1
9 Vdc
− +
+
−
C_{10}
4.7 μF
C_9
0.1 μF
R_5
1000 Ω
R_4
22 kΩ
R_6
680 Ω
R_7
10 kΩ
C_5
0.01 μF
Q_2
ECG-222
C_8
250 pF
reactance
R_8
200 Ω
linear
type-AB
T_1
J_2
RCVR
R_5
1500 kΩ
C_3
0.01 μF
Q_1
ECG-222
C_2
0.02 μF
D_3
6.8 Vdc
(noise source)
8
4
3
5
IC$_1$
555
timer
7
1
6
2
R_1
6800 kΩ
D_1
D_1, D_2,
IN914
or
IN4148
D_2
R_2
6800 Ω
C_1
0.01 μF
C_6
40 pF
balance
C_7
100 pF
S.M.
T_1: toroid trifilar
transformer
10T, no. 26 over
FT-50-43 core
Antenna
J_1

27-6 Noise bridge circuit.

trum is enhanced because of the 1-kHz square-wave modulator that chops the noise signal. An amplifier boosts the noise signal to the level needed in the bridge circuit.

The detector used in the noise bridge is a tunable receiver covering the frequencies of interest. The preferable receiver uses an AM demodulator, but both CW (morse code) and SSB receivers will do in a pinch. The quality of the receiver depends entirely on the precision with which you need to know the operating frequency of the device under test.

Adjusting antennas

Perhaps the most common use for the antenna noise bridge is finding the impedance and resonant points of an HF antenna. Connect the *receiver* terminal of the bridge to the *antenna* input of the HF receiver through a short length of coaxial cable as shown in Fig. 27-7. The length should be as short as possible, and the characteristic impedance should match that of the antenna feedline. Next, connect the coaxial feedline from the antenna to the *antenna* terminals on the bridge. You are now ready to test the antenna.

Set the noise bridge resistance control to the antenna feedline impedance (usually 50 to 75 Ω for most common antennas). Set the reactance control to midrange (zero). Next, tune the receiver to the *expected* resonant frequency (F_{exp}) of the antenna. Turn the noise bridge *on*, look for a noise signal of about S9 (will vary on different receivers), and if—in the unlikely event that the antenna is resonant on the expected frequency—you will find yourself right in the middle of the null.

Adjust the *resistance* control R on the bridge for a null, i.e., minimum noise as indicated by the S meter. Next, adjust the *resistance* control C for a null. Repeat the adjustments of the R and C controls for the deepest possible null, as indicated by the lowest noise output on the S meter (there is some interaction between the two controls).

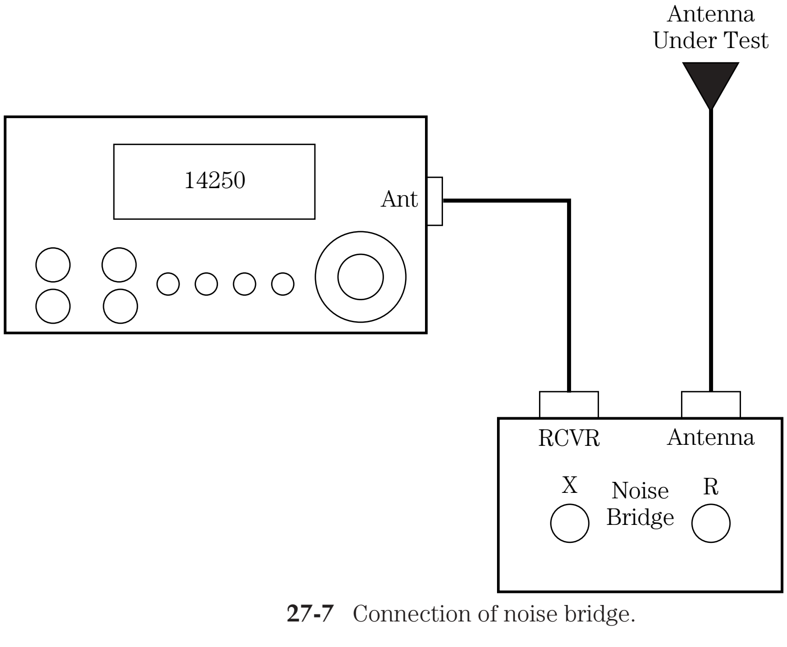

27-7 Connection of noise bridge.

A perfectly resonant antenna will have a reactance reading of zero ohms, and a resistance of 50 to 75 Ω. Real antennas might have some reactance (the less the better), and a resistance that is somewhat different from 50 to 75 Ω. Impedance-matching methods can be used to transform the actual resistive component to the 50- or 75-Ω characteristic impedance of the transmission line. The results to expect:

1. If the resistance is close to zero, then suspect that there is a short circuit on the transmission line. Suspect an open circuit if the resistance is close to 200 Ω.
2. A reactance reading on the X_L side of zero indicates that the antenna is too long, while a reading on the X_C side of zero indicates an antenna that is too short.

An antenna that is too long or too short should be adjusted to the correct length. To determine the correct length, we must find the actual resonant frequency F_r. To do this, reset the *reactance* control to zero, and then *slowly* tune the receiver in the proper direction—downband for too long and upband for too short—until the null is found. On a high-Q antenna, the null is easy to miss if you tune too fast. Don't be surprised if that null is out of band by quite a bit. The percentage of change is given by dividing the expected resonant frequency F_{exp} by the actual resonant frequency F_r, and multiply by 100:

$$\text{Change} = \frac{F_{\text{exp}} \times 100\%}{F_r} \qquad \textbf{[27.6]}$$

Connect the antenna, noise bridge, and receiver in the same manner as above. Set the receiver to the expected resonant frequency (i.e., approximately 468/*F* for half-wavelength types and 234/*F* for quarter-wavelength types). Set the resistance control to 50 Ω or 75 Ω, as appropriate for the normal antenna impedance and the transmission line impedance. Set the reactance control to zero. Turn the bridge on and listen for the noise signal.

Slowly rock the reactance control back and forth to find on which side of zero the null appears. Once the direction of the null is determined, set the reactance control to zero, and then tune the receiver toward the null direction (downband if null is on X_L side and upband if it is on the X_C side of zero).

A less-than-ideal antenna will not have exactly 50- or 75-Ω impedance, so some adjustment of R and C to find the deepest null is in order. You will be surprised how far off some dipoles and other forms of antennas can be if they are not in "free space," (i.e., if they are close to the earth's surface).

Nonresonant antenna adjustment

We can operate antennas on frequencies other than their resonant frequency if we know the impedance:

$$X_C = X = \frac{55}{68 - C} - 2340 \qquad \textbf{[27.7]}$$

or

$$X_L = X = 2340 - \frac{159{,}155}{68 + C} \qquad \textbf{[27.8]}$$

Now, plug *"X"* calculated from Eq. 27.7 or 27.8 into $X_f = X/F$ where F is the desired frequency in MHz.

Other jobs

The noise bridge can be used for a variety of jobs. We can find the values of capacitors and inductors, determine the characteristics of series- and parallel-tuned resonant circuits, and adjust transmission lines.

Transmission line measurements Some antennas and non-noise measurements require antenna feedlines that are either a quarter-wavelength or half-wavelength at some specific frequency. In other cases, a piece of coaxial cable of specified length is required for other purposes: for instance, the dummy load used to service depth sounders is nothing but a long piece of shorted coax that returns the echo at a time interval that corresponds to a specific depth. We can use the bridge to find these lengths as follows:

1. Connect a short circuit across the *unknown* terminals and adjust R and X for the best null at the frequency of interest. (*Note*: both will be near zero.)
2. Remove the short circuit.
3. Connect the length of transmission line to the *unknown* terminal—it should be longer than the expected length.
4. For quarter-wavelength lines, shorten the line until the null is very close to the desired frequency. For half-wavelength lines, do the same thing, except that the line must be shorted at the far end for each trial length.

The velocity factor of a transmission line (usually designated by the letter V in equations) is a decimal fraction that tells us how fast the radio wave propagates along the line relative to the speed of light in free space. For example, foam dielectric coaxial cable is said to have a velocity factor of $V = 0.80$. This number means that the signals in the line travel at a speed 0.80 (or 80 percent) of the speed of light.

Because all radio wavelength formulas are based on the velocity of light, you need the V value to calculate the physical length needed to equal any given electrical length. For example, a half-wavelength piece of coax has a physical length of $(492 \times V)/F_{\mathrm{MHz}}$ feet. Unfortunately, the real value of V is often a bit different from the published value. You can use the noise bridge to find the actual value of V for any sample of coaxial cable as follows:

1. Select a convenient length of the coax more than 12 ft in length and install a PL-259 RF connector (or other connector compatible with your instrument) on one end, and short-circuit the other end.
2. Accurately measure the physical length of the coax in feet; convert the "remainder" inches to a decimal fraction of a foot by dividing by 12 (e.g., 32 ft 8 in= 32.67 ft because 8 in/12 in = 0.67). Alternatively, cut off the cable to the nearest foot and reconnect the short circuit.
3. Set the bridge *resistance* and *reactance* controls to zero.
4. Adjust the monitor receiver for deepest null. Use the null frequency to find the velocity factor $V = FL/492$, where V is the velocity factor (a decimal fraction); F is the frequency in megahertz; and L is the cable length in feet.

Tuned circuit measurements An inductor/capacitor (*LC*) tuned "tank" circuit is the circuit equivalent of a resonant antenna, so there is some similarity between the two measurements. You can measure resonant frequency with the noise bridge to within ±20 percent (or better if care is taken). This accuracy might seem poor, but it is better than you can usually get with low-cost signal generators, dip meters, absorption wavemeters, and the like.

A *series tuned circuit* exhibits a low impedance at the resonant frequency and a high impedance at all other frequencies. Start the measurement by connecting the series tuned circuit under test across the *unknown* terminals of the dip meter. Set the *resistance* control to a low resistance value, close to zero ohms. Set the *reactance* control at midscale (zero mark). Next, tune the receiver to the expected null frequency, and then tune for the null. Make sure that the null is at its deepest point by rocking the R and X controls for best null. At this point, the receiver frequency is the resonant frequency of the tank circuit.

A *parallel resonant circuit* exhibits a high impedance at resonance and a low impedance at all other frequencies. The measurement is made in exactly the same manner as for the series resonant circuits, except that the connection is different. Figure 27-8 shows a two-turn link coupling that is needed to inject the noise signal into the parallel resonant tank circuit. If the inductor is the toroidal type, then the link must go through the hole in the doughnut-shaped core and then connect to the *unknown* terminals on the bridge. After this, proceed exactly as you would for the series tuned tank measurement.

Capacitance and inductance measurements The bridge requires a 100-pF silver mica test capacitor and a 4.7-μH test inductor, which are used to measure inductance and capacitance, respectively. The idea is to use the test components to form a series-tuned resonant circuit with an unknown component. If you find a resonant frequency, then you can calculate the unknown value. In both cases, the series tuned circuit is connected across the *unknown* terminals of the dip meter, and the series tuned procedure is followed.

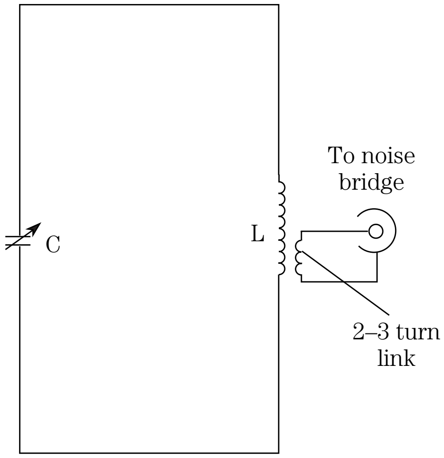

27-8 Noise bridge connection to *LC* tank circuits.

To measure inductance, connect the 100-pF capacitor in series with the unknown coil across the *unknown* terminals of the dip meter. When the null frequency is found, find the inductance from

$$L = \frac{253}{F^2} \qquad \textbf{[27.9]}$$

where

L is the inductance, in microhenries
F is the frequency, in megahertz

Connect the test inductor across the *unknown* terminals in series with the unknown capacitance. Set the *resistance* control to zero, tune the receiver to 2 MHz, and readjust the *reactance* control for null. Without readjusting the noise bridge control, connect the test inductor in series with the unknown capacitance and retune the receiver for a null. Capacitance can now be calculated from:

$$C = \frac{5389}{F^2} \qquad \textbf{[27.10]}$$

where:

C is in picofarads
F is in megahertz

Dip oscillators

One of the most common instruments for determining the resonant frequency of an antenna is the so-called dip oscillator or dip meter. Originally called the *grid dip meter*, the basis for this instrument is the fact that its output energy can be absorbed by a nearby resonant circuit (or antenna, which is an electrically resonant LC tank circuit). When the inductor of the dip oscillator (see Fig. 27-9) is brought into

SOLID STATE
DIP METER
HEATHKIT

27-9 Dip meter.

close proximity to a resonant tank circuit, and the oscillator is operating on the resonant frequency, then a small amount of energy is transferred. This energy loss shows up on the meter pointer as a slight "dipping" action. The dip is extremely sharp, and is easily missed if the meter frequency dial is tuned too rapidly.

Antennas are resonant circuits, and can be treated in a manner similar to *LC* tank circuits. Figure 27-10A shows one way to couple the dip oscillator to a vertical antenna radiator. The inductor of the dipper is brought into close proximity to the base of the radiator. Figure 27-10B shows the means for coupling the dip oscillator

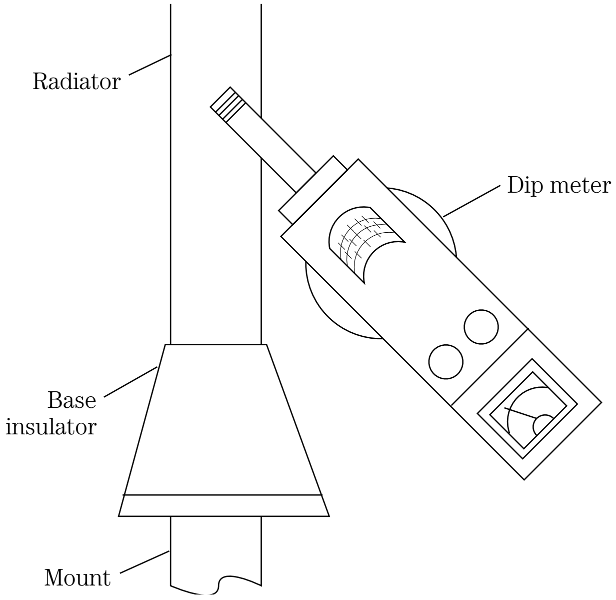
Radiator
Dip meter
Base insulator
Mount

27-10A Direct coupling the dip meter to antenna.

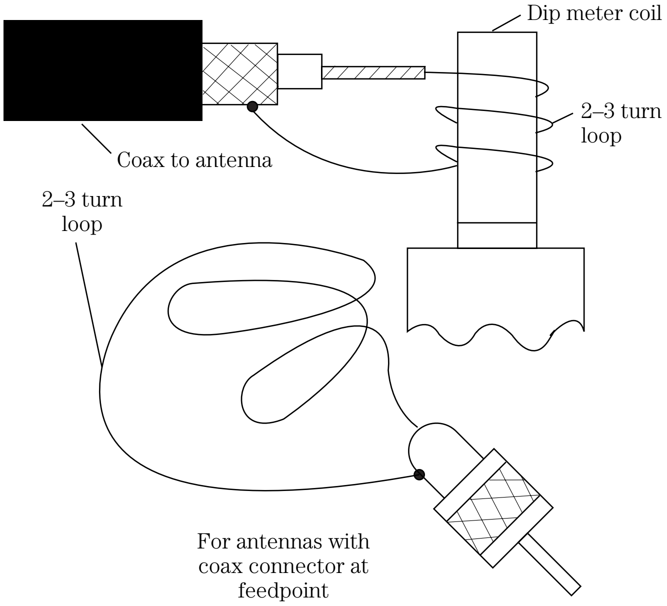
Dip meter coil
2–3 turn loop
Coax to antenna
2–3 turn loop
For antennas with coax connector at feedpoint

27-10B Coupling the dip meter to coaxial cable.

to systems where the radiator is not easily accessed (as when the antenna is still erected). We connect a small two- or three-turn loop to the transmitter end of the transmission line, and then bring the inductor of the dipper close to it. A better way is to connect the loop directly to the antenna feedpoint.

There are two problems with dip meters that must be recognized in order to best use the instrument. First, the dip is very sharp. It is easy to tune past the dip and not even see it. To make matters worse, it is normal for the meter reading to drop off gradually from one end of the tuning range to the other. But if you tune very slowly, then you will notice a very sharp dip when the resonant point is reached.

The second problem is the dial calibration. The dial gradations of inexpensive dip meters are too close together and are often erroneous. It is better to monitor the output of the dip oscillator on a receiver and depend upon the calibration of the receiver for data.

Selecting and using RF wattmeters and antenna VSWR meters

A key instrument required in checking the performance of, or troubleshooting, radio transmitters is the RF power meter (or “wattmeter”). These instruments measure the output power of the transmitter, and display the result in watts, or some related unit. Closely related to RF wattmeters is the antenna VSWR meter. This instrument also examines the output of the transmitter and gives a relative indication of output power. It can be calibrated to display the dimensionless units of *voltage standing wave ratio.* Many modern instruments, a couple of which will be covered as examples in this chapter, combine both RF power and VSWR measurement capabilities.

Measuring RF power

Measuring RF power has traditionally been notoriously difficult, except perhaps in the singular case of continuous-wave (CW) sources that produce nice, well-behaved sine waves. Even in that limited case, however, some measurement methods are distinctly better than others.

The peak voltage of a waveform is 100 V (i.e., peak-to-peak 200 V). Given that the CW waveform is sinusoidal, we know that the root mean square (RMS) voltage is 0.707 V. The output power is related to the RMS voltage across the load by

$$P = \frac{(V_{\text{rms}})^2}{Z_o} \qquad \textbf{[27.11]}$$

where:

P is the power, in watts
V_{rms} is the RMS potential, in volts
Z_o is the load impedance, in ohms

If we assume a load impedance of 50 Ω, then we can state that the power in our hypothetical illustration waveform is 100 W.

We can measure power on unmodulated sinusoidal waveforms by measuring either the RMS or peak values of either voltage or current, assuming that a constant-value resistance load is present. But the problem becomes more complex on modulated signals. The various power readings on a Bird model 4311 peak power meter, the peak (PEP) and average powers, vary markedly with modulation type.

One of the earliest forms of practical RF power measurement was the thermocouple RF ammeter (see Fig. 27-11). This instrument works by dissipating a small amount of power in a small resistance inside the meter, and then measuring the heat generated with a thermocouple. A dc current meter monitors the output of the thermocouple device, and indicates the level of current flowing in the heating element. Because it works on the basis of the power dissipated heating a resistance, *a thermocouple RF ammeter is inherently an RMS-reading device*. Because of this feature it is very useful for making average power measurements. If we know the RMS current and the resistive component of the load impedance, and if the reactive component is zero or very low, then we can determine RF power from the familiar expression:

$$P = I^2 \times R_i \tag{27.12}$$

There is, however, a significant problem that keeps thermocouple RF ammeters from being universally used in RF power measurement: those instruments are highly frequency-dependent. Even at low frequencies, it is recommended that the meters be mounted on insulating material with at least ⅜-in spacing between the meter and its metal cabinet. Even with that precaution, however, there is a strong frequency dependence that renders the meter less useful at higher frequencies. Some meters are advertised to operate into the low-VHF region, but a note of caution is necessary. That recommendation requires a copy of the calibrated frequency response curve for that specific meter, so that a correction factor can be added (or subtracted) from

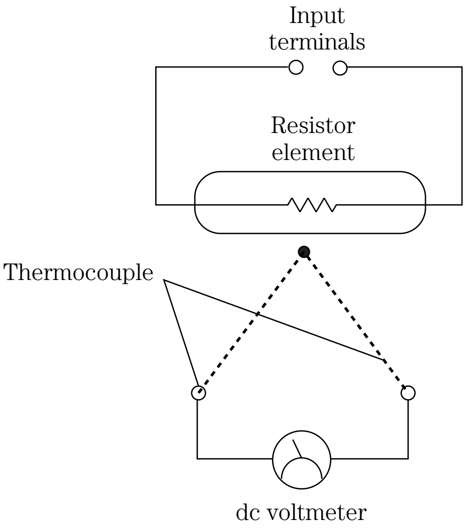

27-11 Circuit for thermocouple RF ammeter.

the reading. At 10 MHz and higher, the readings of the thermocouple RF ammeter must be taken with a certain amount of skepticism unless the original calibration chart is available.

We can also measure RF power by measuring the voltage across the load resistance (see Fig. 27-12). In the circuit of Fig. 27-12, the RF voltage appearing across the load is scaled downward to a level compatible with the voltmeter by the resistor voltage divider (R_2/R_3). The output of this divider is rectified by CR_1 and filtered to dc by the action of capacitor C_2.

The method of measuring the voltage in a simple diode voltmeter is valid only if the RF signal is unmodulated and has a sinusoidal waveshape. While these criteria are met in many transmitters, they are not universal. If the voltmeter circuit is peak reading, as in Fig. 27-12, then the peak power is

$$P = \frac{V_o^2}{R_1} \tag{27.13}$$

The average power is then found by multiplying the peak power by 0.707. Some meter circuits include voltage dividers that precede the meter and thereby convert the reading to RMS, and thus convert the power to average power. Again, it must be stressed that terms like RMS, average, and peak have meaning only when the input

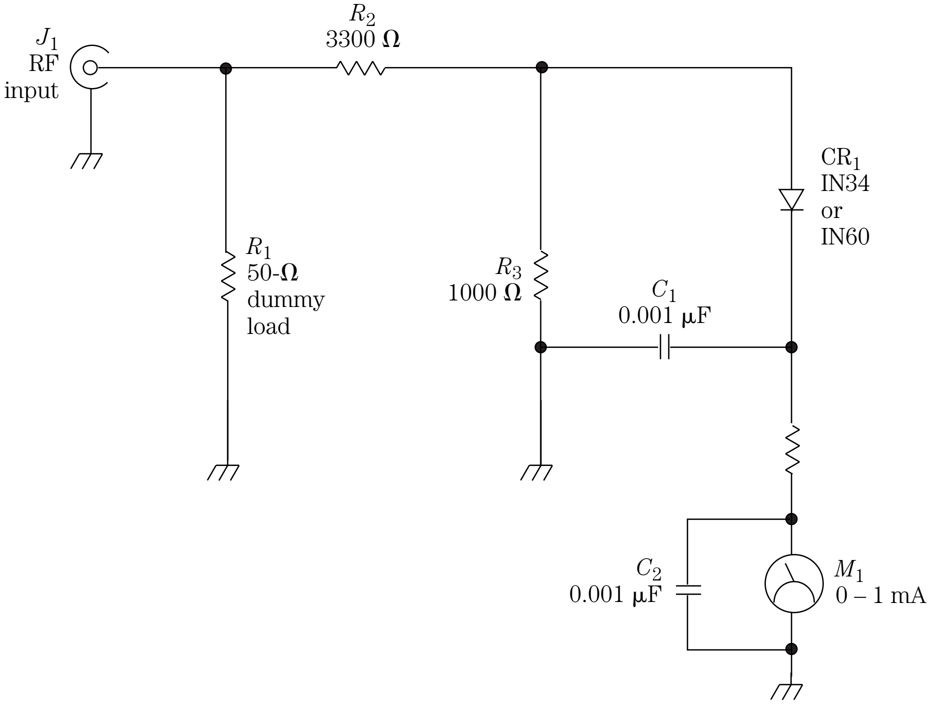

27-12 RF voltmeter "wattmeter."

RF signal is both unmodulated and sinusoidal. Otherwise, the readings are meaningless unless calibrated against some other source.

It is also possible to use various bridge methods for measurement of RF power. Figure 27-13 shows a bridge set up to measure both forward and reverse power. This circuit was once popular for VSWR meters. There are four elements in this quasi-Wheatstone bridge circuit: R_1, R_2, R_3, and the antenna impedance (connected to the bridge at J_2). If R_{ant} is the antenna resistance, then we know that the bridge is in balance (i.e., the null condition) when the ratios R_1/R_2 and R_3/R_{ant} are equal. In an ideal situation, resistor R_3 will have a resistance equal to R_{ant} , but that might overly limit the usefulness of the bridge. In some cases, therefore, the bridge will use a compromise value such as 67 Ω for R_3. Such a resistor will be usable on both 50- and 75-Ω antenna systems with only small errors. Typically, these meters are designed to read relative power level, rather than the actual power.

An advantage of this type of meter is that we can get an accurate measurement of VSWR by proper calibration. With the switch in the *forward* position, and RF power applied to J_1 (XMTR), potentiometer R_6 is adjusted to produce a full-scale deflection on meter M_1. When the switch is then set to the *reverse* position, the meter will read reverse power relative to the VSWR. An appropriate VSWR scale is provided.

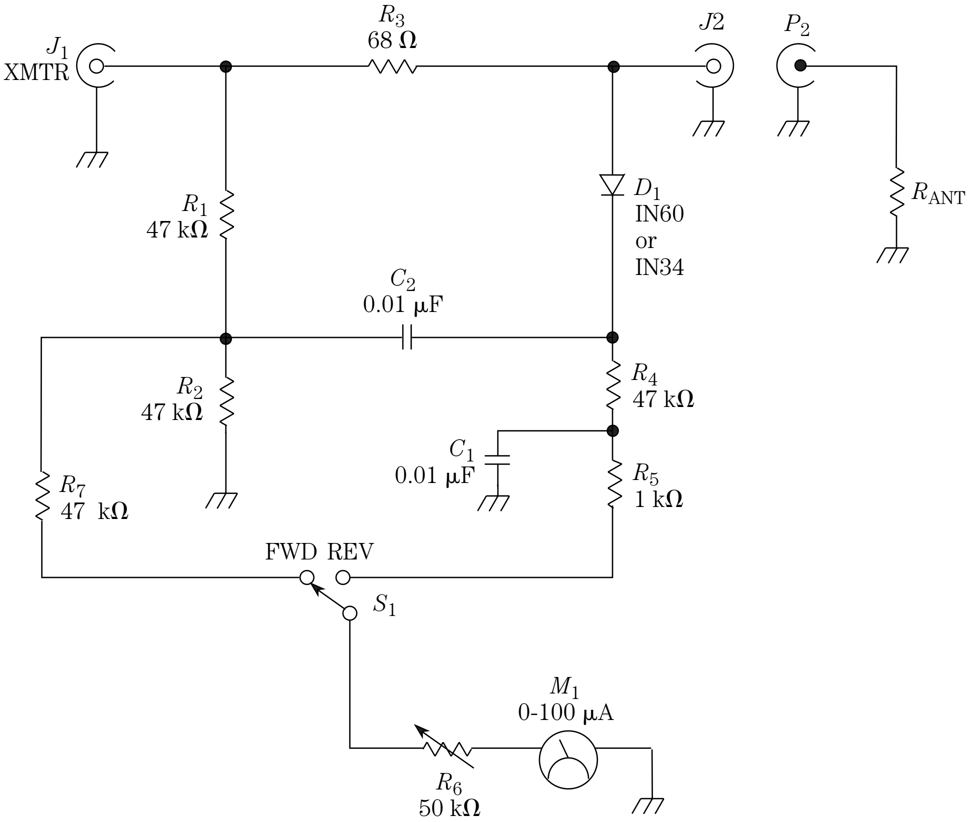

27-13 A bridge arrangement to measure both forward and reverse power.

A significant problem with the bridge of Fig. 27-13 is that it cannot be left in the circuit while transmitting because it dissipates a considerable amount of RF power in the internal resistances. These meters, during the time when they were popular, were provided with switches that bypassed the bridge when transmitting. The bridge was only in the circuit when making a measurement.

An improved bridge circuit is the capacitor/resistor bridge in Fig. 27-14; this circuit is called the micromatch bridge. Immediately, we see that the micromatch is improved over the conventional bridge because it uses only 1 Ω in series with the line (R_i). This resistor dissipates considerably less power than the resistance used in the previous example. Because of this low-value resistance, we can leave the micromatch in the line while transmitting. Recall that the ratios of the bridge arms must be equal for the null condition to occur. In this case, the capacitive reactance ratio of C_1/C_2 must match the resistance ratio R_1/R_{ant}. For a 50-Ω antenna, the ratio is 1⁄50, and for 75-Ω antennas it is 1⁄75 (or, for the compromise situation, 1⁄68). The small-value trimmer capacitor (C_2) must be adjusted for a reactance ratio with C_1 of 1⁄50, 1⁄75, or 1⁄68, depending upon how the bridge is set up.

The sensitivity control can be used to calibrate the meter. In one version of the micromatch, there are three power ranges (10, 100, and 1000 W). Each range has its own sensitivity control, and these are switched in and out of the circuit as needed.

The monomatch bridge circuit in Fig. 27-15 is the instrument of choice for HF and low-VHF applications. In the monomatch design, the transmission line is segment

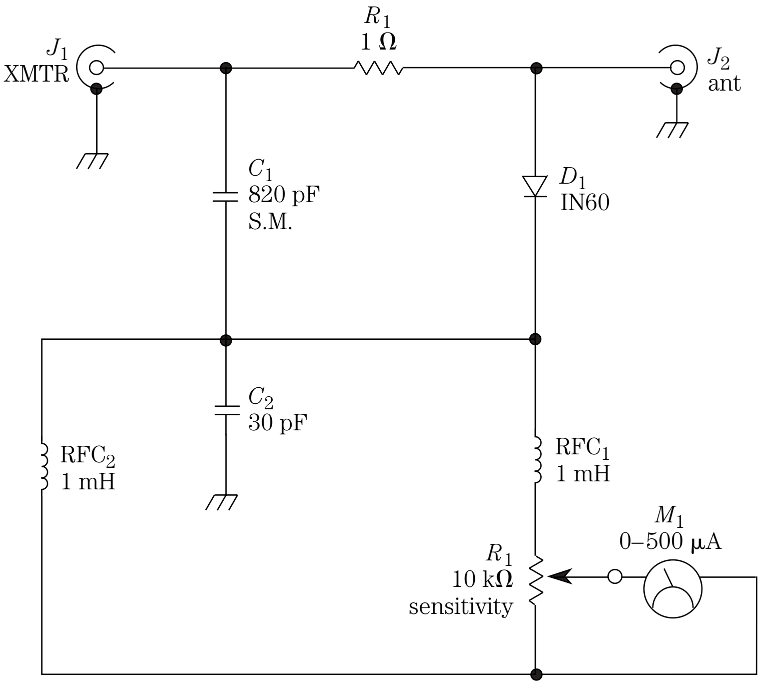

27-14 Micromatch wattmeter.

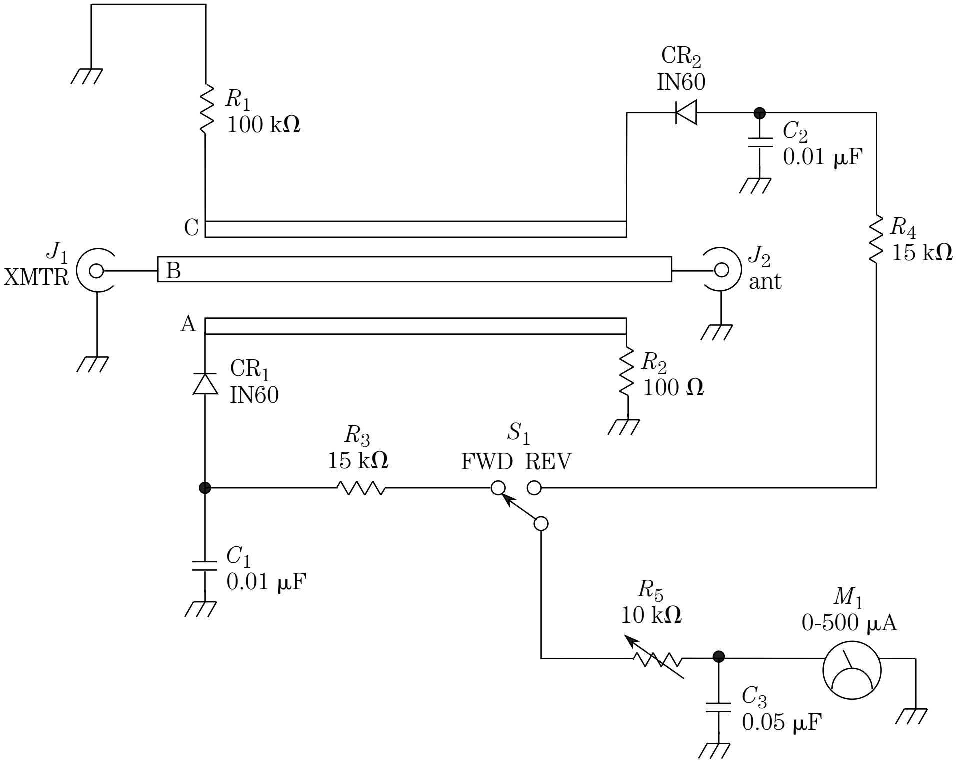

27-15 Monomatch wattmeter.

B, while RF sampling elements are formed by segments A and C. Although the original designs were based on a coaxial-cable sensor, later versions used either printed circuit foil transmission line segments or parallel brass rods for A, B, and C.

The sensor unit is basically a directional coupler with a detector element for both forward and reverse directions. For best accuracy, diodes CR$_1$ and CR$_2$ should be matched, as should R_1 and R_2. The resistance of R_1 and R_2 should match the transmission line surge impedance, although in many instruments a 68-Ω compromise resistance is used.

The particular circuit shown in Fig. 27-15 uses a single dc meter movement to monitor the output power. Many modern designs use two meters (one each for forward and reverse power).

One of the latest designs in VSWR meter sensors is the current transformer assembly shown in Fig. 27-16. In this instrument, a single-turn ferrite toroid transformer is used as the directional sensor. The transmission line passing through the hole in the toroid “doughnut” forms the primary winding of a broadband RF transformer. The secondary, which consists of 10 to 40 turns of small enamel wire, is connected to a measurement bridge circuit ($C_1 + C_2$ + load) with a rectified dc output.

Figures 27-17 and 27-18 show instruments based on the current transformer technique. Shown in Fig. 27-17 is the Heath model HM-102 high-frequency VSWR/power meter. The sensor is a variant on the current transformer method. This

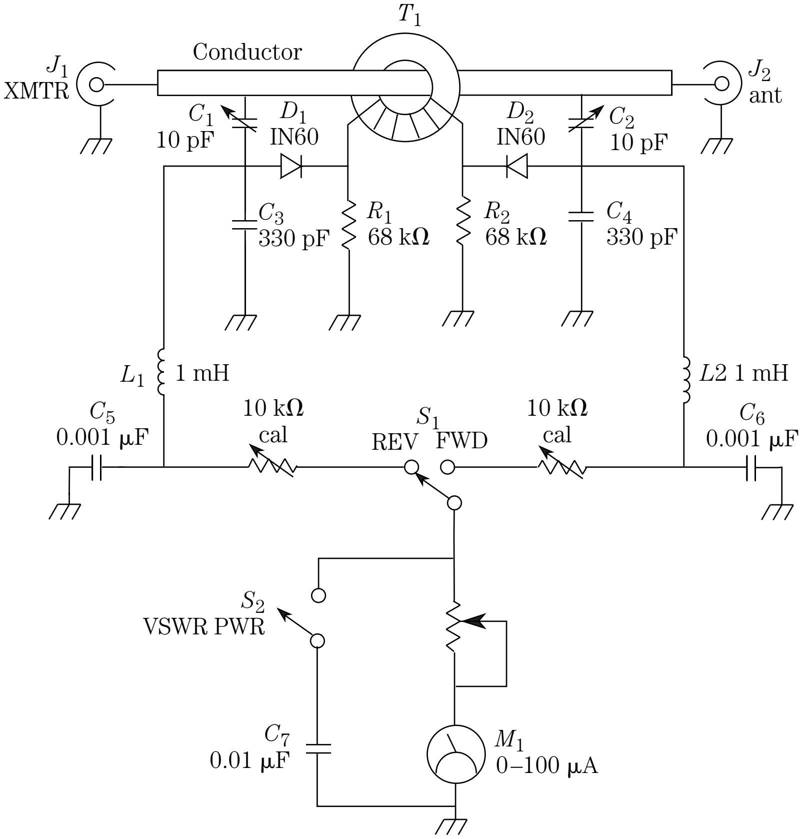
T_1
Conductor
J_1
XMTR
J_2
ant
C_1
10 pF
D_1
IN60
D_2
IN60
C_2
10 pF
C_3
330 pF
R_1
68 kΩ
R_2
68 kΩ
C_4
330 pF
L_1
1 mH
$L2$ 1 mH
C_5
0.001 μF
10 kΩ
cal
S_1
REV
FWD
10 kΩ
cal
C_6
0.001 μF
S_2
VSWR PWR
C_7
0.01 μF
M_1
0–100 μA

27-16 Current transformer wattmeter.

27-17 Amateur radio RF wattmeter.

instrument measures both forward and reflected power, and it can be calibrated to measure VSWR.

The Bird model 43 Thruline RF wattmeter shown in Fig. 27-18 has for years been one of the industry standards in communications service work. Although it is slightly more expensive than lesser instruments, it is also versatile, and it is accurate and rugged. The Thruline meter can be inserted into the transmission line of an

27-18 Bird model 43 RF wattmeter.

antenna system with so little loss that it may be left permanently in the line during normal operations. The Model 43 Thruline is popular with land mobile and marine radio technicians.

The heart of the Thruline meter is the directional coupler transmission line assembly shown in Fig. 27-19A; it is connected in series with the antenna, or dummy load, transmission line. The plug-in directional element can be rotated 180° to measure both forward and reverse power levels. A sampling loop and diode detector are contained within each plug-in element. The main RF barrel is actually a special coaxial line segment with a 50-Ω characteristic impedance. The Thruline sensor works due to the mutual inductance between the sample loop and center conductor of the coaxial element. Figure 27-19B shows an equivalent circuit. The output voltage from the sampler (e) is the sum of two voltages, e_r and e_m. Voltage e_r is created by the voltage divider action of R and C on transmission line voltage E. If R is much less then X_C, then we may write the expression for e_r as

$$e_r = \frac{RE}{X_C} = RE(j\omega \mathrm{C}) \qquad \textbf{[27.14]}$$

Voltage e_m, on the other hand, is due to mutual inductance, and is expressed by

$$e_m = I(j\ \omega) \pm M \qquad \textbf{[27.15]}$$

We now have the expression for both factors that contribute to the total voltage e. We know that:

$$e = e_r + e_m \qquad \textbf{[27.16]}$$

so, by substitution,

$$e = j\omega M\left[\left(\frac{E}{Z_o}\right) \pm I\right] \qquad \textbf{[27.17]}$$

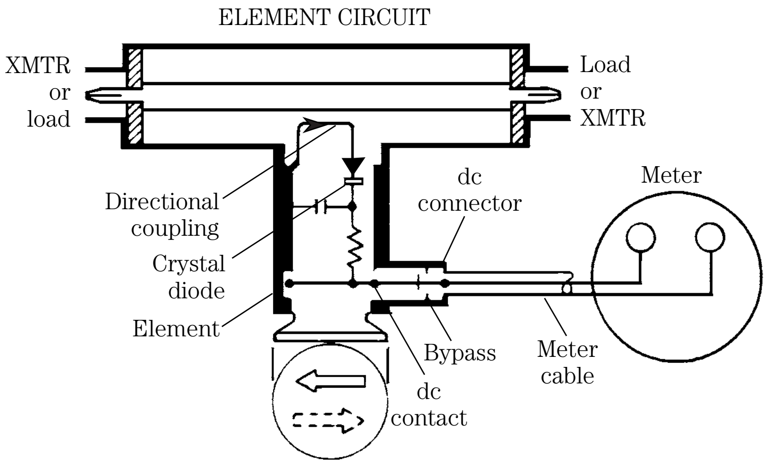
ELEMENT CIRCUIT
XMTR or load
Load or XMTR
Directional coupling
Crystal diode
Element
dc connector
Meter
Bypass
Meter cable
dc contact

27-19A Thruline sensor circuit.

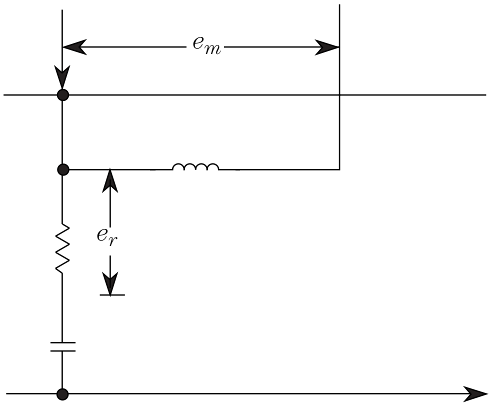

27-19B Circuit of pick-up element in a Model 43 wattmeter.

At any given point in a transmission line, E is the sum of the forward (E_f) and reflected (E_r) voltages, and the line current is equal to

$$I = \frac{E_f}{Z_o} - \frac{E_r}{Z_o} \qquad \textbf{[27.18]}$$

where Z_o is the transmission line impedance.

We may specify e in the forms

$$e = \frac{j\,\omega\,M(2E_f)}{Z_o} \qquad \textbf{[27.19]}$$

and

$$e = \frac{j\,\omega\,M(2E_r)}{Z_o} \qquad \textbf{[27.20]}$$

The output voltage e of the coupler, then, is proportional to the mutual inductance and frequency (by virtue of $j\omega M$). But the manufacturer terminates R in a capacitive reactance, so the frequency dependence is lessened (see Fig. 27-19C). Each element is custom-calibrated, therefore, for a specific frequency and power range. Beyond the specified range for any given element, however, performance is not guaranteed. There are a large number of elements available that cover most commercial applications. The Thruline meter is not a VSWR meter, but rather a power meter. VSWR can be determined from the formula, or by using the nomographs in Fig. 27-20. Some of the Thruline series intended for very high power (Fig. 27-19D) applications use an in-line coaxial cable coupler (for broadcast-style hardline) and a remote indicator.

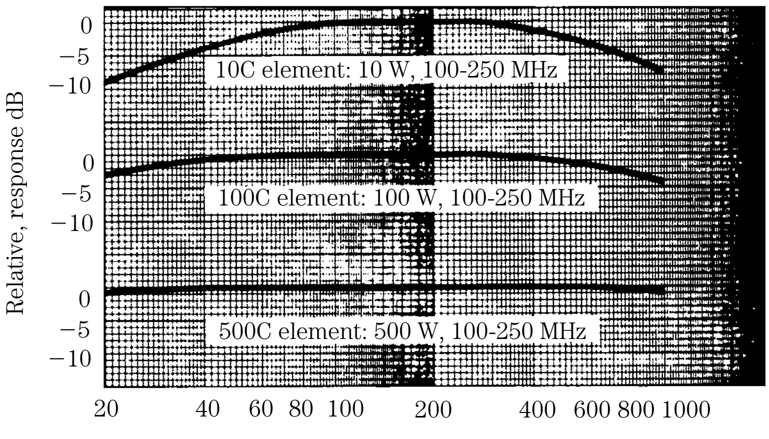

27-19C Power-frequency calibration.

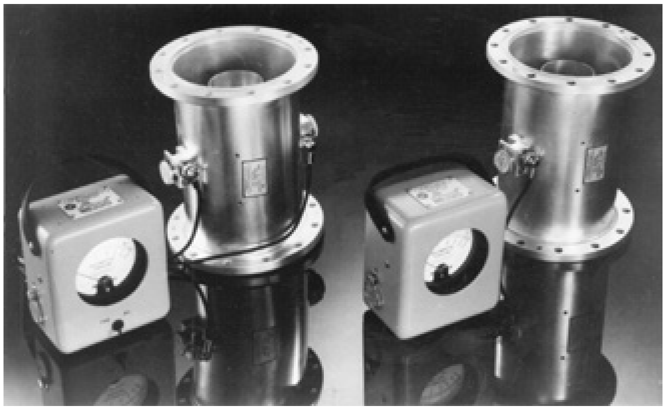

27-19D High-power RF wattmeters.

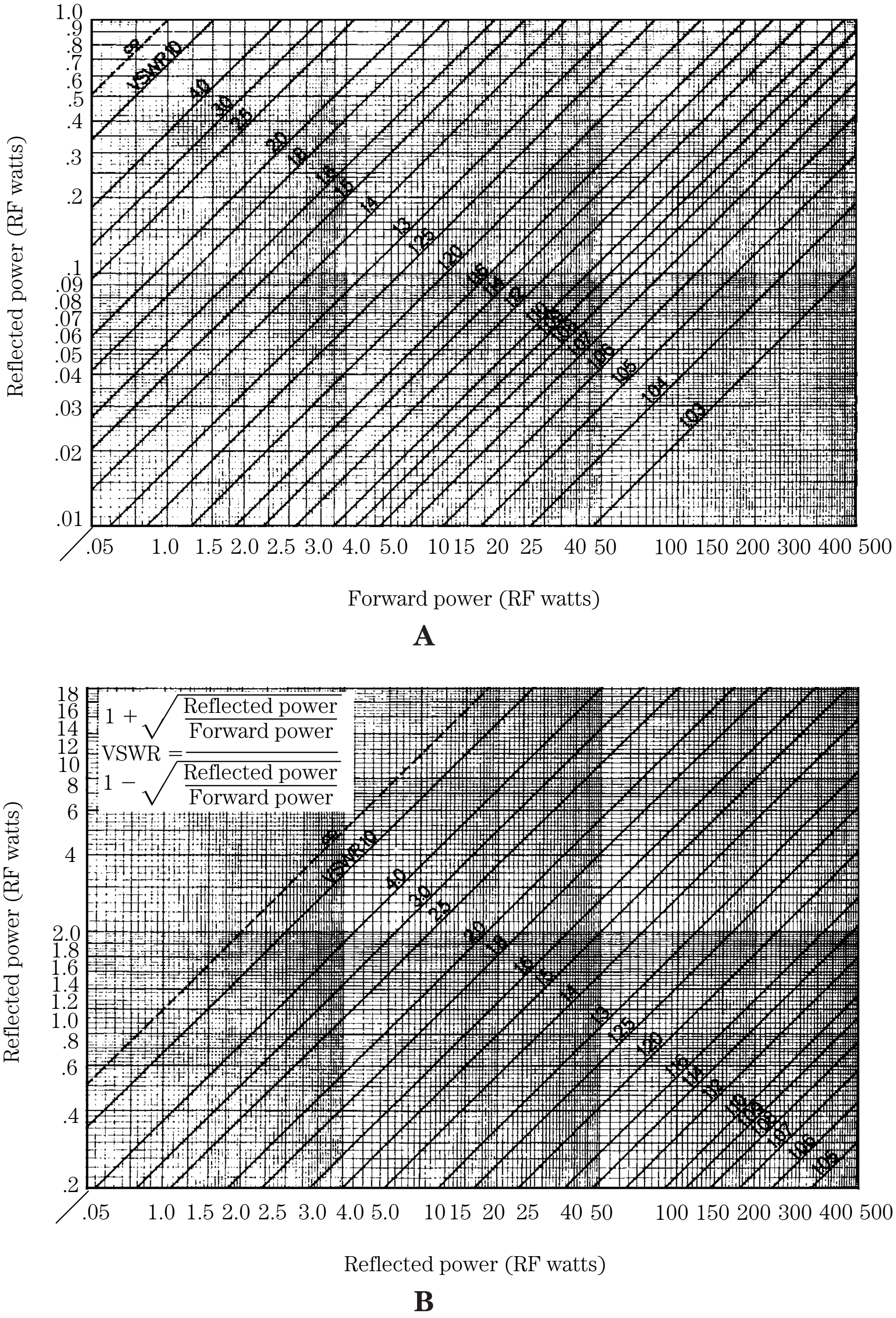

27-20 Forward versus reflected power graphs.

Dummy loads

A *dummy load* is a substitute antenna for making measurements and tests. In fact, British radio engineers often refer to dummy loads as “artificial aerials.” There are several uses for these devices. Radio operators should use dummy loads to tune up on crowded channels and (only then) transfer to the live antenna.

The other use is in troubleshooting antenna systems. Suppose we have a system in which the VSWR is high enough to affect the operation of the transmitter. You can disconnect each successive element and connect the dummy load to its output. If the VSWR goes down to the normal range, then the difficulty is downstream (i.e., toward the antenna). You will eventually find the bad element (which is usually the antenna itself).

Figure 27-21 shows the most elementary form of dummy load, which consists of one or more resistors connected in parallel, series, or series-parallel so that the total resistance is equal to the desired load impedance. The power dissipation is the sum of the individual power dissipations. *It is essential that noninductive resistors be used for this application.* For this reason, carbon-composition or metal-film resistors are used. For very low frequency work, it is permissible to use special counter wound wire low-inductance resistors. These resistors, however, cannot be used over a few hundred kilohertz.

Several commercial dummy loads are shown in Figs. 27-22 through 27-24. The load in Fig. 27-22A is a 5-W model, and is typically used in Citizens Band servicing. The resistor is mounted directly on a PL-259 coaxial connector. These loads typically work to about 300 MHz, although many are not really useful over about 150 MHz. A higher-power version of the same type is shown in Fig. 27-22B. This device works to the low VHF region, and dissipates up to 50 W. I have used this dummy load for servicing high-VHF land mobile rigs, VHF-FM marine rigs, and low-VHF land mobile rigs. Very high power loads are shown in Fig. 27-23. These devices are *Bird Electronics* “coaxial resistors,” and operate to power levels up to 10, 30, and 40 kW. These high-power loads are cooled by flowing water through the body of the resistor and then exhausting the heat in an air-cooled radiator.

Our final dummy load resistor is shown in Fig. 27-24. The actual resistor is shown in Fig. 27-24A, and a schematic view is shown in Fig. 27-24B. The long, high-power noninductive resistor element is rated at 50 **Ω**, and can dissipate 1000 W for

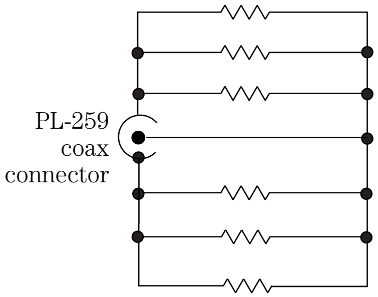

27-21 A very basic dummy load.

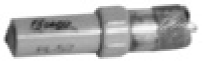

27-22A CB-style 5-W dummy load.

27-22B 15-W dummy load.

WATTS
X1
X10
BLACK SCALE
RED SCALE

27-23 Termaline RF wattmeter.

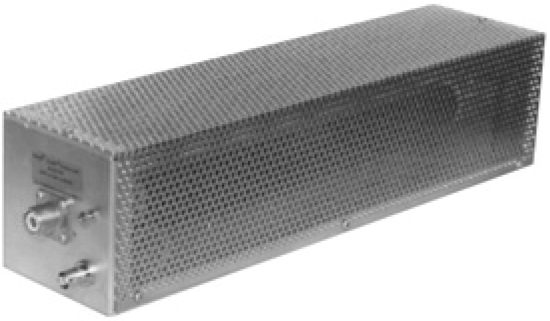

27-24A Drake DL-1000 dummy load.

several minutes. If longer times, or higher powers, are anticipated, then forced air cooling is applied by adding a blower to one end of the cage. The device is modified by adding the BNC sampling jack. This jack is connected internally to either a two-turn loop, made of no.22 insulated hookup wire, or brass rods that are positioned alongside the resistor element. It will therefore pick up a sample of the signal so that it can be viewed on an oscilloscope, or used for other instrumentation purposes.

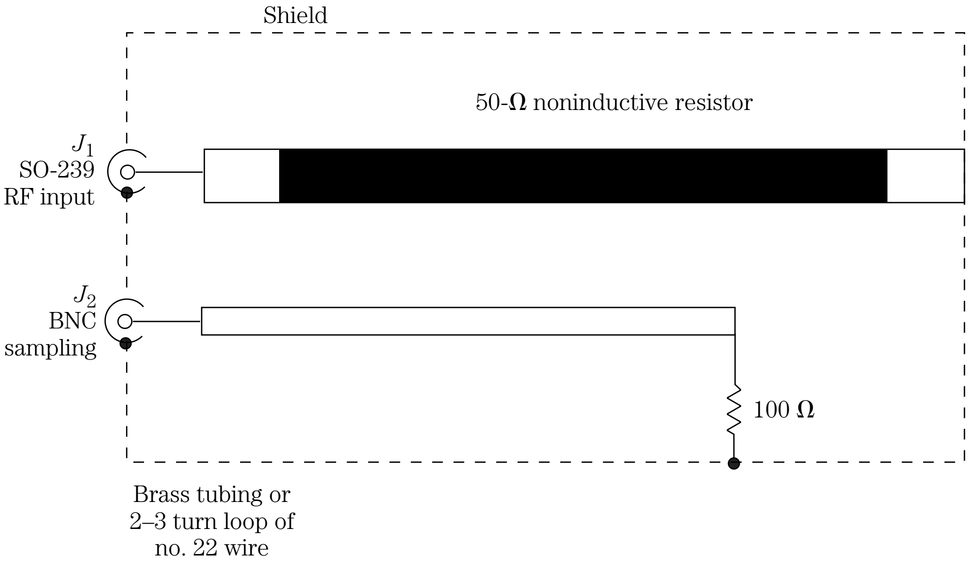
Shield
50-Ω noninductive resistor
J_1
SO-239
RF input
J_2
BNC
sampling
100 Ω
Brass tubing or
2–3 turn loop of
no. 22 wire

27-24B HF dummy load with pickup loop.

28
CHAPTER

General antenna mechanical construction techniques

ANTENNA INSTALLATION CHORES VARY IN COMPLEXITY FROM THE SIMPLE, WHICH A teenager, with no assistance, can execute properly and safely, all the way up to large-scale projects that are best left to professional antenna riggers. If your only antenna installation chore is a dipole handing between the peak of your house roof and a nearby tree, then little need be said besides "Be careful."

The purpose of this chapter is to cover both wire and intermediate installation jobs that can be done by amateurs if instructed properly. Keep in mind when planning an installation, however, that the information given herein is merely informal guidance. Always have your plans reviewed by a professional mechanical or structural engineer licensed to practice in your state. In many locales, there are good insurance reasons for this step. Also, learn and abide by local mechanical, electrical, and building codes. It is simply not worth the pain, and risk, to skirt around local laws.

Antenna erection safety

Before dealing with the radio and performance issues, let's deal first with a few antenna safety matters. You do not want to be hurt either during installation or during the next wind storm. Two problems present themselves: *reliable mechanical installation* and *electrical safety.*

Radio antennas are inherently dangerous to erect if certain precautions are not followed. It is not possible to foresee all the situations you might face in erecting an antenna. I would like to give you all possible warnings, but this is not even possible. You are on your own and must take responsibility for installing your own antenna. I can, however, give you some general safety guidelines. Knowledge of what you face and some hard-nosed, sound judgment, modulated by a little bit of common sense are the best tools on any antenna job.

Copyright 2001 - the McGraw-Hill Companies

One rule that is an absolute is that no antenna should *ever* be erected where the antenna, the feedline, or any part thereof crosses over or under a power line or pole transformer—*ever*! This is a "no kidder"—don't do it. Every year we read sad news in magazines or newspapers about a colleague being electrocuted while installing or working on an antenna. In all these tragic cases, the antenna somehow came into contact with an ac power line. Keep in mind one dictum and make it an absolute:

There is never a time or situation when it is safe to let an antenna contact the electrical power lines!

This advice includes dipoles and other long wire antennas "thrown over" supposedly insulated lines, as well as antennas built from aluminum tubing. The excuse that the power lines are insulated is hogwash. Old insulation crumbles on contact with even a thin wire antenna. *Don't do it*! The operant word is *never*!

Power lines look insulated, but there are often small breaks or weakened spots (especially more than a couple days after installation) that can bring the antenna into contact—lethal contact—with a "hot" power line. Every year or so we hear about an SWL, scanner/monitor buff, or ham operator being *killed* by tossing an antenna wire over a power line. Avoid making yourself into a high-powered resistor—it might blow your "life fuse."

And the same rule applies to situations where the antenna can fall onto a power line if it fails mechanically. You have to examine the situation to see if there is any possible way for that antenna or its support structure to fall onto a power line if it breaks in any way whatsoever. Figure 28-1 is a diagram of my lot in Virginia. A 23-ft mast is erected at the back of the house, and its 23-ft "fall radius" is plotted. It should not intersect either the power lines or the cable TV line, even when it fails and is wind-whipped.

Note that the antenna could fall over the property line onto the next lot to the rear. After consultation with the other landowner and the county mechanical inspector, it was determined that my rear fence is sufficient protection for the other homeowner. Because it is technically on my land, the fence damage is my responsibility. However, should the other homeowner change his mind or a new homeowner buys the place, then my antenna may have to be moved.

Another caution is that you must be physically fit to do the work. While the on-the-ground portion of the work is usually not too strenuous, any climbing at all, even on ladders, can be taxing. Antenna materials are deceptively lightweight on the ground, but when you get up on even a small ladder, they are remarkably difficult to handle. Attempting to manhandle a 22-ft vertical once wiped my back out to the extent that I needed 800-mg of ibuprofen three times a day for about a week—and I consider myself lucky that the pain hit me after I had dismounted the ladder. Besides, if you could see me, you would wonder why a man my size was on any ladder in the first place!

Before using a ladder, learn how to use a ladder. A lot of homeowners, whether putting up antennas or painting the upstairs windows, fall off ladders that were being used incorrectly. There is a right way and a wrong way to use each of the different types of ladder; learn to use them correctly.

If the wind blows even a few kilometers per hour, the danger is magnified considerably. I recall a friend, who is a large, strong "bear of a man," attempting to install

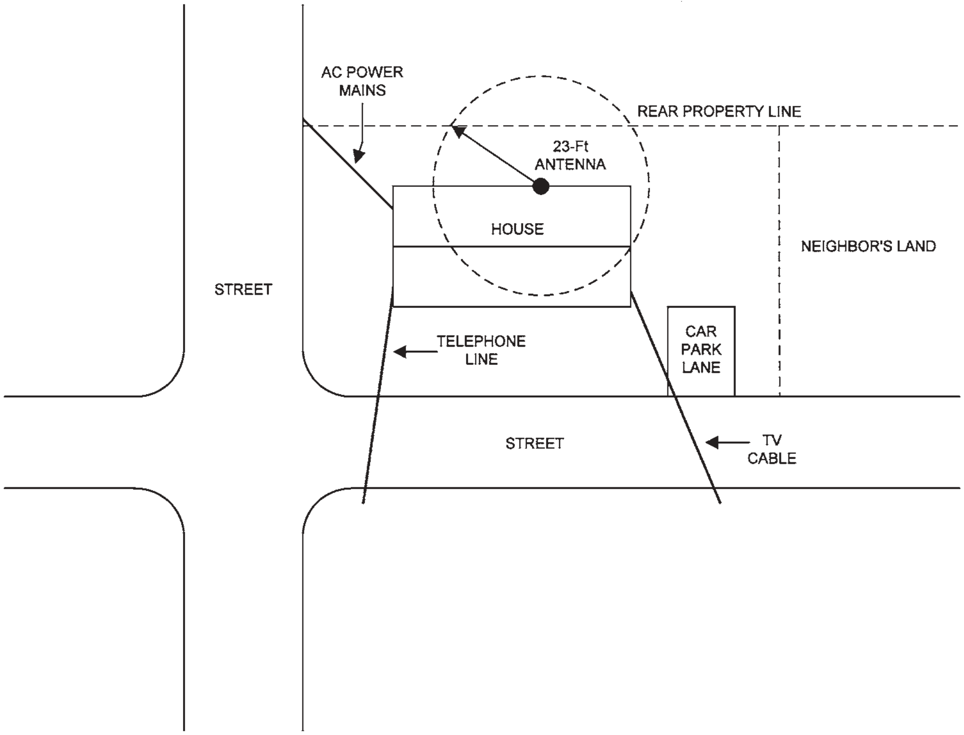

28-1 Antenna installation on suburban lot.

a 26-element television "all channel" rotatable antenna on the roof of his two-story beach house. He wanted to be able to pick up Richmond, VA; Norfolk, VA; Washington, DC; and Baltimore, MD, on television, plus scanner bands, from his beach house on the Chesapeake Bay. The antenna was handled easily with one hand on the ground and with no wind blowing. But up on the roof, it was a different story! "Big Al" was on the peak of the roof when an off-shore gust of wind came up suddenly and caught the antenna. It acted like a hang glider and pulled Al off the roof, plunging him down two stories to the patio below. He fractured his pelvis and busted a leg. Two months of orthopedic casts and a year of physical therapy followed—not to mention lost wages. Expensive TV antenna, I reckon. *Be careful*!

One good rule is to always work under the "buddy system." Ask as many friends as are needed to do the job safely, and always have at least one assistant, even when you think you can do it alone.

Always use quality materials and good work practices. Antennas, being potentially dangerous, always should have the best of both materials and workmanship in order to keep quality high. It is not just the electrical or radio reception workings that are important but also the ability to stay up in the air and be safe.

When planning an antenna job, keep in mind that pedestrian traffic in your yard possibly could affect the antenna system. Wires are difficult to see, and if an antenna wire is low enough to intersect someone's body, then it is possible to cause very serious injuries to passersby. In the United States, even when the person is a trespasser (or even a burglar!), the courts may hold a homeowner liable for injuries caused by an inappropriately designed and installed antenna. Take care for safety not only of yourself but also of others.

Consider a typical scenario involving a four- or five-band trap vertical antenna (Fig. 28-2). It will be 18 to 26 ft tall, judging from the advertisements in magazines, and will be mounted on a roof or mast 12 to 30 ft off the ground. Consider the case of a 25-ft-tall trap vertical installed atop a 15-ft-high telescoping TV antenna mast. The total height above ground will be the sum of the two heights: 15 ft + 25 ft = 40 ft. The tip of the vertical is 40-ft above ground level. You must select a location at which a 40-ft aluminum pole can fit—and fall. This requirement limits the selection of locations for the antenna. If this is done, then a wind storm cannot result in a shorted or downed power line if the antenna falls over.

When installing a vertical antenna, especially one that is not ground-mounted, make sure that you have help. It takes at least two people to safely install a standard HF vertical antenna, and more may be needed for especially large models. If you are alone, then go find some friends to lend an arm or two. Wrenched backs, smashed antennas, crunched house parts, and other calamities simply do not happen as often on a well-organized work party that has sufficient hands to do the job safely.

The second issue pertaining to installing antennas is old-fashioned mechanical integrity. Two problems are often seen. First, you must comply with local building

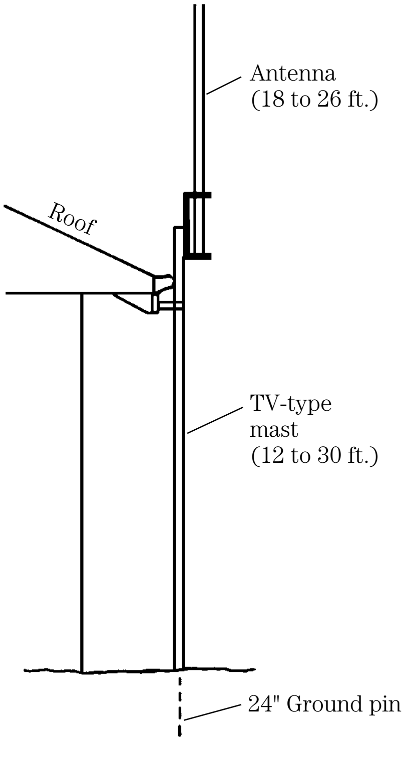

28-2 Mast-mounted vertical antenna.

codes, regulations, and inspections. The local government has a reasonable interest and absolute right to impose reasonable engineering standards on the electrical and mechanical installation of radio antennas. The second issue is that it is in your own best interest to make the installation as good as possible. Consider the local regulations as the minimum acceptable standard, not the maximum.

Both these mechanical integrity issues become extremely important if a problem develops. For example, suppose that a wind or snow storm wrecks the antenna, plus part of your house. Your insurance company may not pay off if your local government requires an inspection and you failed to get one. Make sure that the mechanical and/or electrical inspector (as required by the law) leaves a certificate or receipt proving that the final inspection was done. It could come in handy when disputing the insurance company over damage.

A quality installation starts with the selection of good hardware for the installation. Any radio-TV parts distributor who sells TV antenna hardware will have what you need in most cases. I used Radio Shack standoff brackets, ground pin, and a 19-ft telescoping mast in one installation. Wherever you buy, select the best-quality, strongest materials that you can find. Opt for steel masts and brackets over aluminum, no matter what the salesperson behind the counter tells you. Keep in mind that although salespeople can be knowledgeable and even helpful, you, not they, are responsible for the integrity of the installation.

Local government One necessary reminder is that your local government might have some interesting ideas—legal requirements, actually—concerning your antenna installation. The electrical, mechanical, and zoning codes must be observed. There is a great deal of similarity between local codes because most of them are adaptations (even adoptions) from certain national standards. But there are enough differences that one needs to consult local authorities. Indeed, you may need a license or building permit to install an antenna in the first place.

One problem that SWLs and scanner/monitor buffs face in the United States is that their antennas are not protected by the Federal Communications Commission, as are ham antennas. Local governments have limited rights to regulate ham antennas—only “reasonable” mechanical and electrical standards can be imposed. It may be illegal for you to install any type of antenna.

Save all paperwork regarding your building permit, including the inspection certificates, decals, letters or papers, and the original drawings (with the local building inspector’s stamps). If a casualty occurs, then your insurance company may elect not to pay off if you have violated an electrical, mechanical, building, or zoning code. This advice may be overlooked by an enthusiastic antenna builder, but it could prove to be a costly oversight if something bad happens.

In some jurisdictions, there might be another limitation on antenna location. Some local governments have a requirement that the antenna be able to fall over and land entirely on your own property. About 30 years ago, a Maryland county required that antennas be installed double their own length plus 50 ft from the nearest property line—until a local ham radio club sued on grounds that about 99.99 percent of the rooftop TV antennas in the county were noncompliant. In one humorous scene, a group of amateur radio operators went down to the county offices and asked for about 50,000 complaint forms. They filled out several thousand against homeowners with TV antennas before they were stopped by the county officer. Before installing any antenna, check with the local building codes!

Wire antenna construction

Volumes have been written on wire antennas. Indeed, many of my antenna books spend a lot of space on the topic. What seems to be missing in many books is details on the mechanical aspects of wire antenna construction and installation. Although some books cover the mechanical details, much is often missing. Let's take a look at some of the basics.

The basic wire antenna

Wire antennas come in a bewildering variety of forms, but for our purposes, the "standard model" is our old friend the half-wavelength horizontal dipole (Fig. 28-3). The overall length (A) in meters is found from $144/F_{\mathrm{MHz}}$. Each element (B) is one-half the overall length, or $72/F_{\mathrm{MHz}}$. The antenna's nominal feedpoint impedance at resonance is 73 Ω, so it is a good match to standard 75-Ω coaxial cable. It is common practice to use a 1:1 balun transformer at the feedpoint in order to balance currents and keep the pattern according to textbook predictions.

Center insulators

It was once common (but poor) practice to strip a few centimeters of coaxial cable, part the braid and center conductor, and connect them to each side of the dipole. We could then wonder why the darn thing fell down in the next wind storm. Even if you elect to not use the balun transformer, a ready-made center insulator is a must. Figure 28-4A shows a common form of center insulator for use with dipoles and other wire antennas. There is an SO-239 UHF coaxial connector on the bottom to accept the feedline and two eyebolts mounted with solder connectors for the antenna elements on the sidewalls. At the top is a small eyebolt that is used to strain-relieve the system, or to hang it from a center support (as when inverted-V dipoles are built).

A larger form of center insulator is shown in Fig. 28-4B. The enlarged down section indicates that this device contains either a 4:1 or 1:1 impedance ratio balun

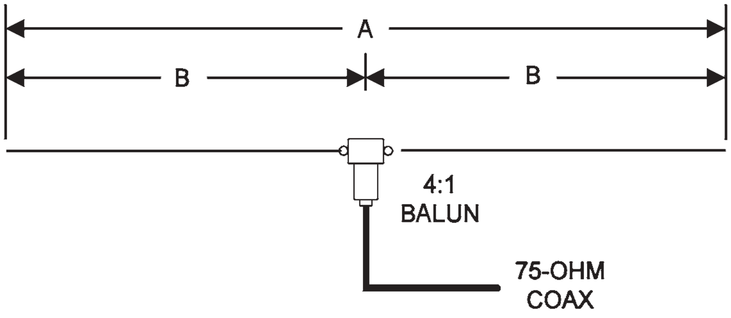

28-3 Dipole antenna.

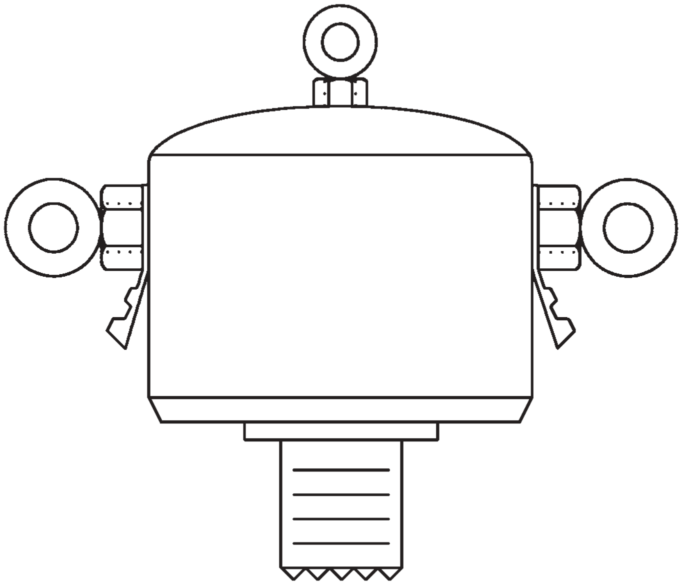

28-4A Center insulator.

transformer. The 1:1 size is recommended for ordinary dipoles, whereas a 4:1 unit is used for folded dipoles and certain other wire antennas.

Connections to the device, whether straight center insulator or balun form, are shown in Fig. 28-5. For simplicity, only one side is shown (the other side is identical). The elements (*B* in Fig. 28-3) typically are made of no. 12 or no. 14 antenna wire. The best wire is stranded and is copper-clad steel core wire (Copperweld).

The end of the antenna wire is passed through the eyebolt and then doubled back and twisted onto itself 7 to 10 times. The remainder of the wire is then soldered to the terminal. Some people prefer to use a separate "pigtail" wire to connect the antenna element to the solder terminal. It is fastened to the antenna element wire between the 7- to 10-turn twist and the eyebolt.

When preparing the center connector/balun, it is good practice to use steel wool or sandpaper on the solder lug to remove oxidation. Oxidation makes it hard to solder and forces you to use too much heat to make a good connection. Also, be sure to have a large pair of pliers to crimp the wings of the solder lugs onto the wire.

Solder assists in making a good connection at the solder lugs, but the crimp should be good enough to do the job. The solder then provides corrosion protection. Also, tin the 7- to 10-turn twist section with solder. This does not increase the strength or provide better electrical connection—it is protection against the corrosion that will occur with copper wires.

Note Use only resin-core solder for antennas! Typical solder is 50/50 or 60/40 lead/tin ratio and may be labeled "radio-TV" or "electronic" or something similar. Plumber's solder is acid core and will destroy your antenna within a short time after it is erected.

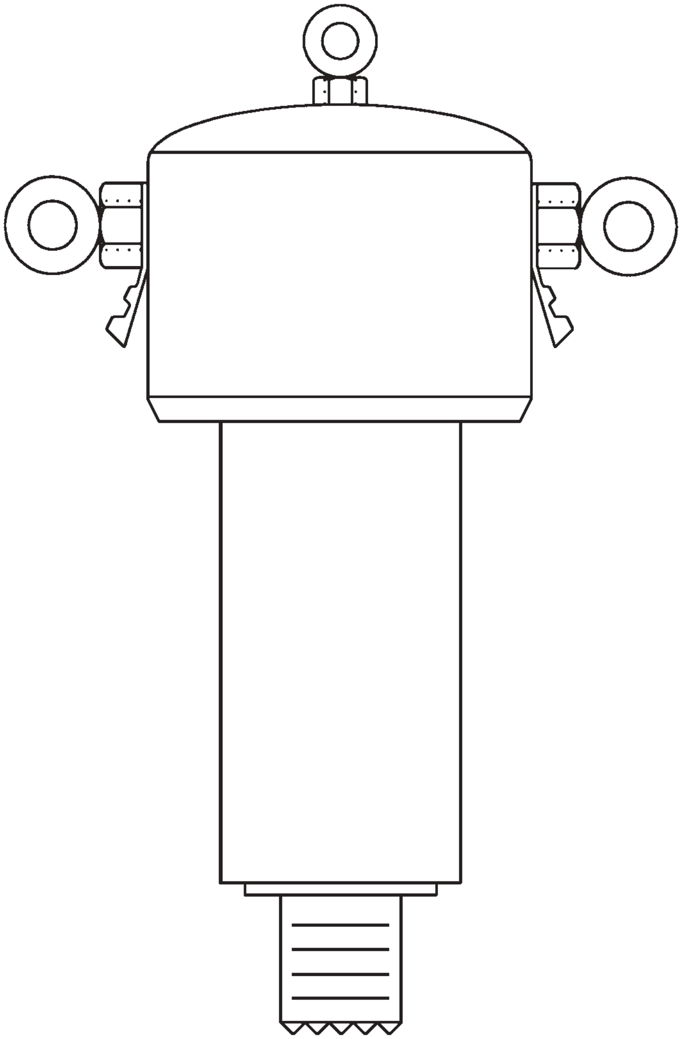

28-4B Balun/center insulator.

Figure 28-6 shows the dipole installed. The ends of the two wire elements are connected to end insulators, which are in turn supported by rope to a vertical support (e.g., mast, tree, roof line of the house, etc.). Although shown nice and straight here, the actual installation will droop a bit in the center due to the weights of the center insulator or balun, the coaxial cable, and the wire itself. Gravity!

Note that the coaxial cable is formed into a 15- to 25-cm loop just below the coaxial connector to the center insulator or balun. This forms a species of RF choke to keep currents from flowing in the outer shield and improves the pattern. The coaxial cable loop is bound together with black electrician's tape or some similar adhesive medium and then fastened to the top eyelet with stout string or fishing line for strain relief.

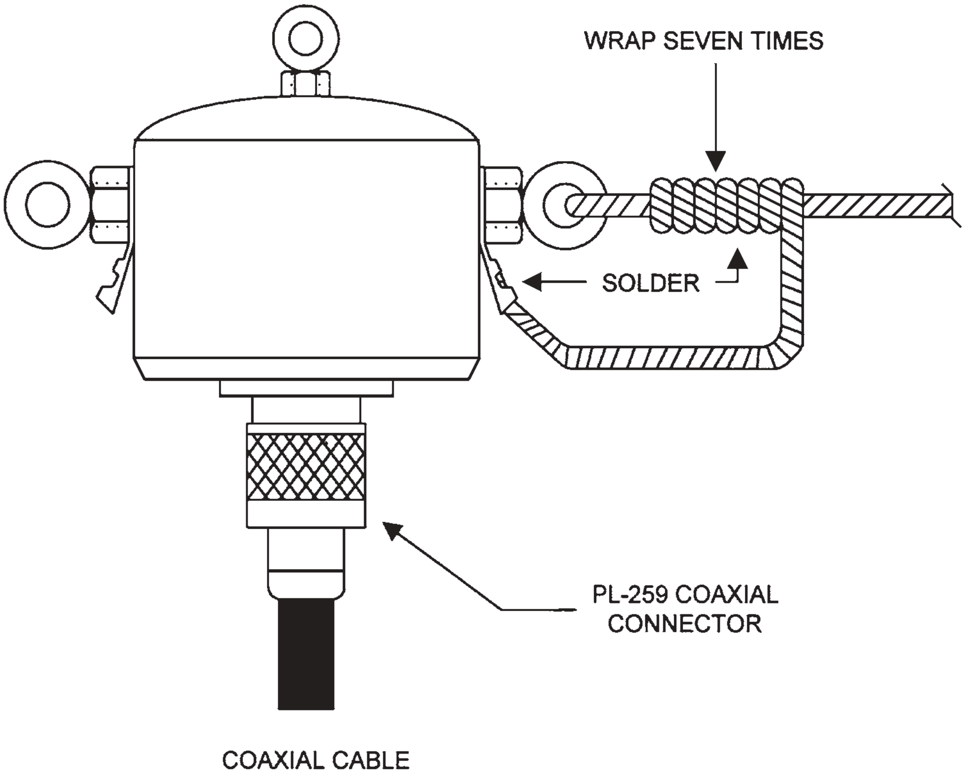

28-5 Center insulator/balun connections.

Support rope

The support rope selected should be large enough to provide support but small enough to fit through the holes of the end insulators. Cotton rope is singularly unsuited to this task because it weathers poorly and will rot through in short order. Nylon rope, sail rope, and parachute rope are perhaps the best types, especially if a particular lot has been tested for high tensile strength. Remember, the wind loading of your antenna, even a short dipole, can be tremendous!

End insulators

End insulators come in a large variety of shapes and sizes (I have even seen one chap use old toothbrushes, suitably drilled, of course). The main categories, however, are represented in Fig. 28-7. The version in Fig. 28-7A is the standard form, whereas that in Fig. 28-7B is an “egg” insulator. The egg insulator has two sets of wire grooves and through-holes that are orthogonal to each other (only one set is shown). Wire passes through one hold and set of grooves, and the supporting rope passes through the other hole/groove set.

The standard form shown in Fig. 28-7A has a number of grooves or ridges cut along the body. These grooves are intended to lengthen the electrical path between

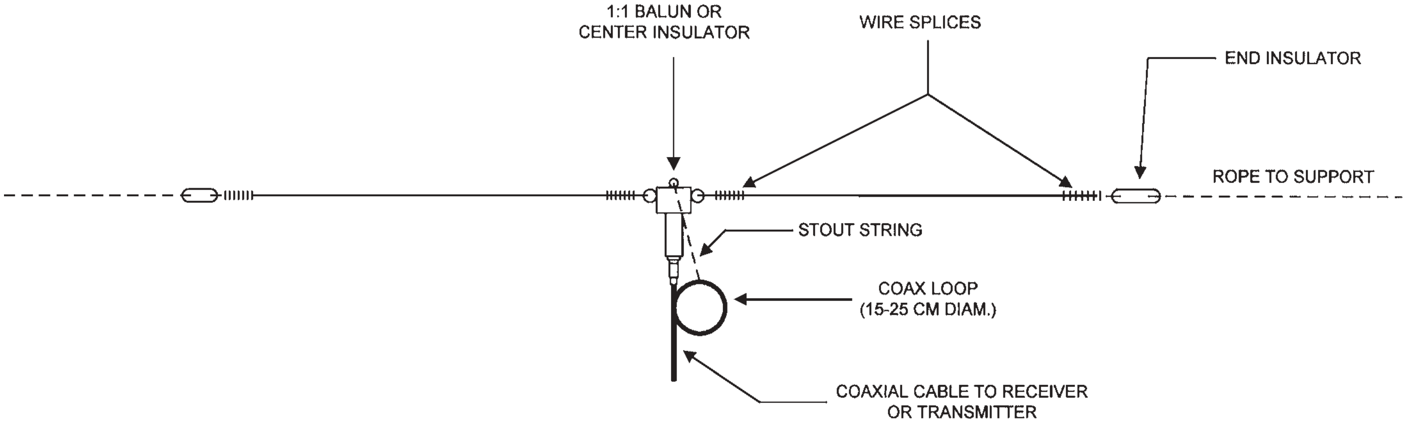
1:1 BALUN OR
CENTER INSULATOR
WIRE SPLICES
END INSULATOR
ROPE TO SUPPORT
STOUT STRING
COAX LOOP
(15-25 CM DIAM.)
COAXIAL CABLE TO RECEIVER
OR TRANSMITTER

28-6 Dipole and wire antenna installation.

ends of the insulator and therefore serve to reduce losses in the insulator. Typical materials used to construct insulators include ceramic, glass, nylon, and plastics.

Figure 28-8 shows the connections to the end insulator. The electrical connections to the antenna wire are the same as for the center insulator. The only proviso here is that for transmitting antennas, it is wise not to leave any sharp points sticking up. The end of the dipole is at high RF voltages, so sharp points tend to cause corona sparking. Cut off any extraneous leads that form sharp points, and then smooth the whole affair down with tinning (i.e., solder).

The rope is treated in a similar manner as the wire, but it must be knotted in some manner that tends to pull together as the rope is pulled into tension. Some people use the "hangman's knot," but this seems a bit excessive.

Strain relief of the antenna

Figure 28-9 shows a method of strain relief for the center insulator or balun. This point is at the highest tension and probably is the weakest because of the number of

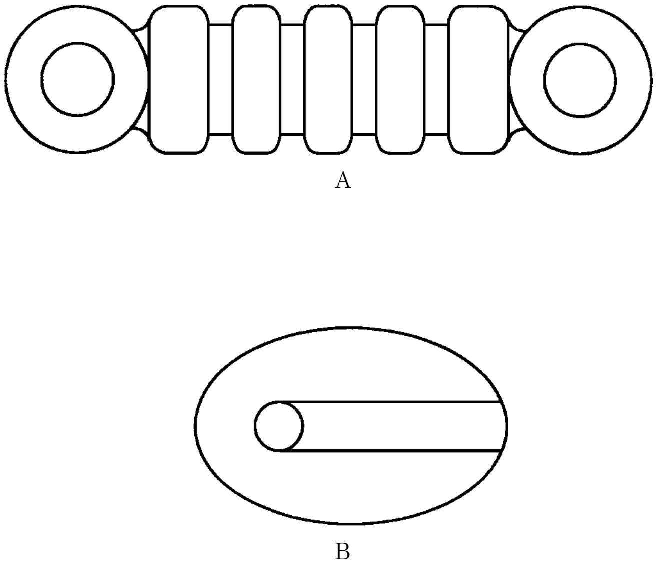

28-7 End insulators.

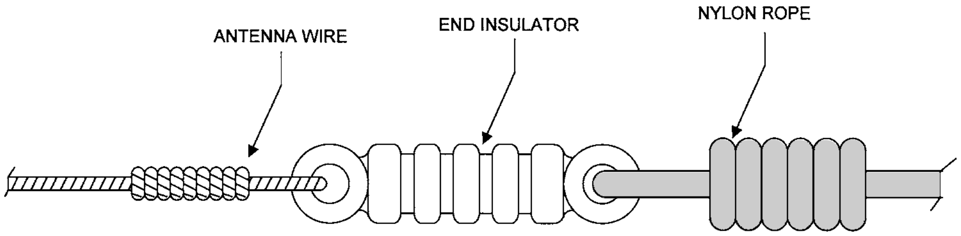

28-8 Connections to end insulators.

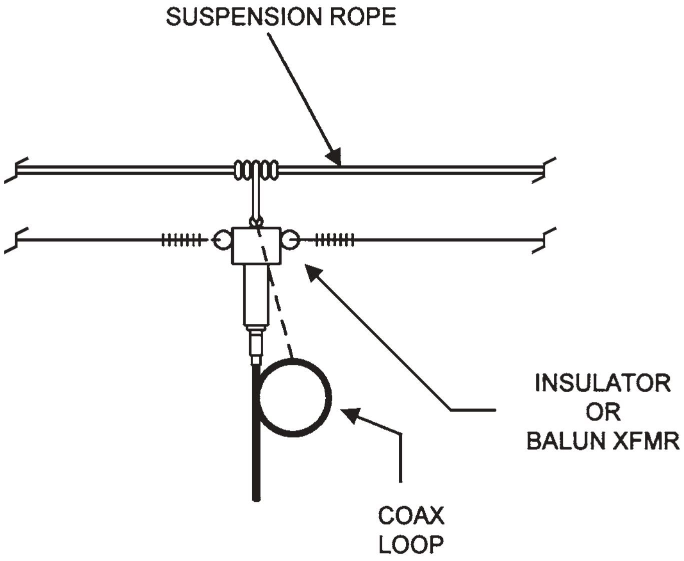

28-9 Strain relieving the center insulator or balun.

connections and things that can break. A suspension rope is run between the antenna's vertical supports, parallel to the antenna wire. It is tied to the top eyelet on the center insulator or balun by a short bit of rope about 15 to 20 cm long.

Special case of the folded dipole

The folded dipole (Fig. 28-10) is a special case of the half-wavelength horizontal dipole. It consists of two close-spaced conductors shorted together on the extreme ends. One of the conductors is split in the center and used as the feedpoint. The feedpoint impedance at resonance is around 280 Ω, so it makes a good match to 300-Ω television antenna twin lead. The usual practice is to build the entire antenna and feedline out of twin lead. An alternative (as shown in Fig. 28-10) is to provide a 4:1 balun transformer at the feedpoint to match 75-Ω coaxial cable. Despite the fact that 300-Ω TV twin lead is disappearing from the market, the folded dipole maintains a certain following among SWLs and hams.

The problem is how to install the antenna such that the terribly weak 300-Ω twin lead wires do not break at the slightest provocation. Figure 28-11 shows a solution used by one fellow (admittedly a fine worker in plastics and other materials). He fashioned a center insulator and the end insulators from a piece of strong Lucite material.

The 300-Ω TV twin lead is prepared as shown in Fig. 28-11A. The antenna element is notched in the center (break along the bottom) for about 1 to 1.5 cm. A hand punch is used to place two or three holes in the center insulation on either side of the break. Either a large-size leather punch or a paper punch can be used to make

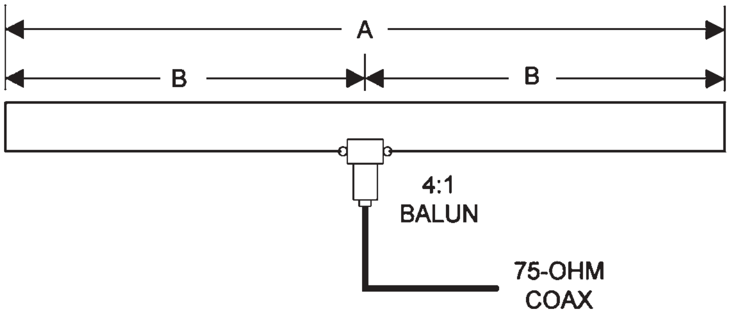
A
B
B
4:1
BALUN
75-OHM
COAX

28-10 Folded dipole.

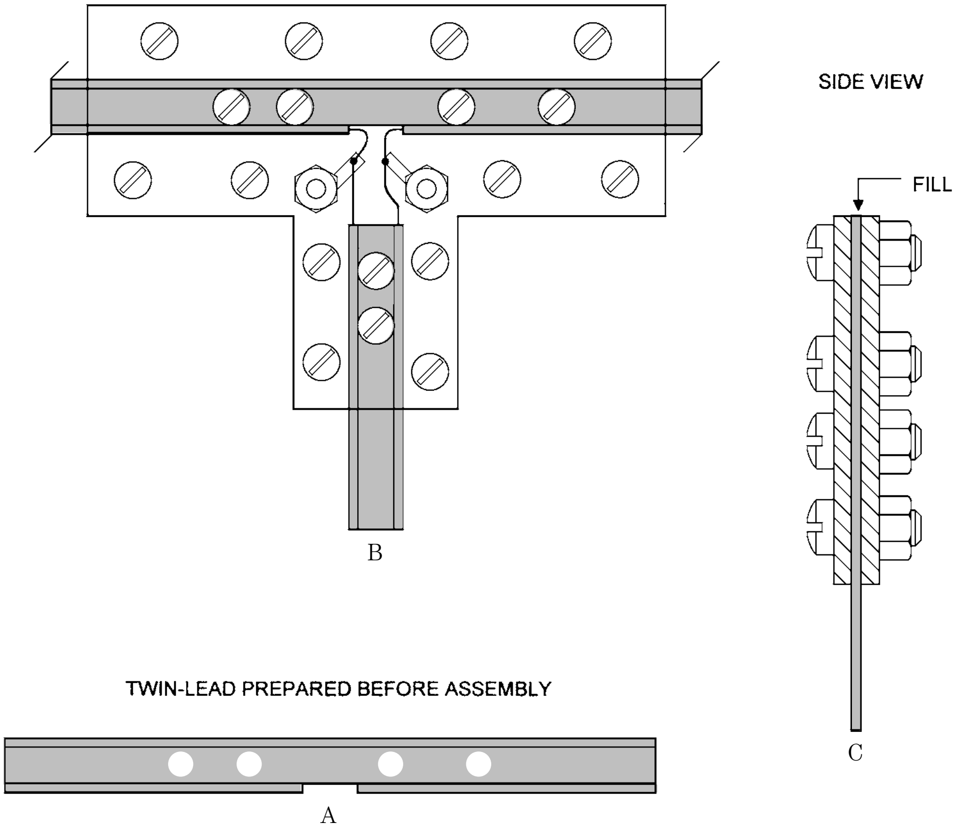
SIDE VIEW
FILL
B
C
TWIN-LEAD PREPARED BEFORE ASSEMBLY
A

28-11 Center insulator for ribbon cable folded dipole: (A) cable preparation, (B) front/back view, (C) edge view.

the holes. In a test, I was able to make the required holes using a hand-operated paper punch, but it required a large force (and turned my hand temporarily purple!).

The center insulator is shown in Fig. 28-11B. It is made from a piece of strong plastic, Lucite, or other insulating material. Two identical sections, front and back, are needed. A number of 5-mm holes are drilled into both pieces at the points shown to clamp the twin lead. A pair of solder lugs is provided to make connections between the antenna element and the transmission line (either twin lead or leads from a balun). The screws and nuts used to fasten the two halves of the insulator are made of nylon to prevent interaction with the antenna.

A side view is shown in Fig. 28-11C. Note that the twin lead causes a gap that can catch water. It also makes it possible to break one or both insulators by overtightening the nuts. In order to prevent this, a gasket of similar material is glued into the space as filler. The gasket material should be about as thick as the twin lead.

If you opt for a balun transformer, then it can be wired to the downleg of the tee insulator using the holes that normally would clamp the twin-lead transmission line.

An end insulator for twin-lead antennas is shown in Fig. 28-12. It is constructed in a similar manner to the center insulator. Two scenarios are possible. The first, and that used in a test antenna, is to make the clamping fixture from metal such as 3- to 6-mm brass or copper stock. The ends of the antenna wires are shorted, so this is not a great problem.

An end insulator is then used with a rope in the normal manner. A chain-link section is used to fasten the clamping fixture and the end insulator. I attempted this with a nylon chain link, about 1.5 × 2.5 in, from a do-it-yourself hardware store. It was only with a bit of difficulty that the ends were pulled apart, slipped into both the end insulator and the clamping fixture, and then reclosed.

The second approach is to make the clamping fixture out of the same material as the center insulator. It then becomes the end insulator, so the other end insulator and that cursed chain link can be ignored. If you opt for this approach, it is a good idea to smooth and polish the rope hole on the clamping fixture to prevent chaffing of the rope.

Masts and other supports for wire antennas

The end supports of the antenna can be anything that provides height: a mast, a tree, or the roof line of a building. Figure 28-13 shows the use of a mast, but either of the other forms of support could be used as well. The support rope is not tied off at the

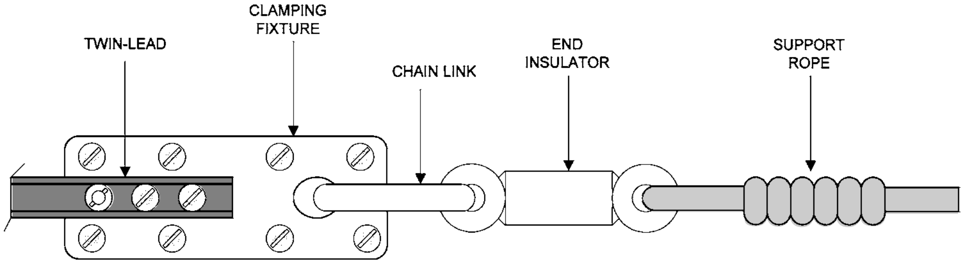

28-12 Folded dipole end insulator.

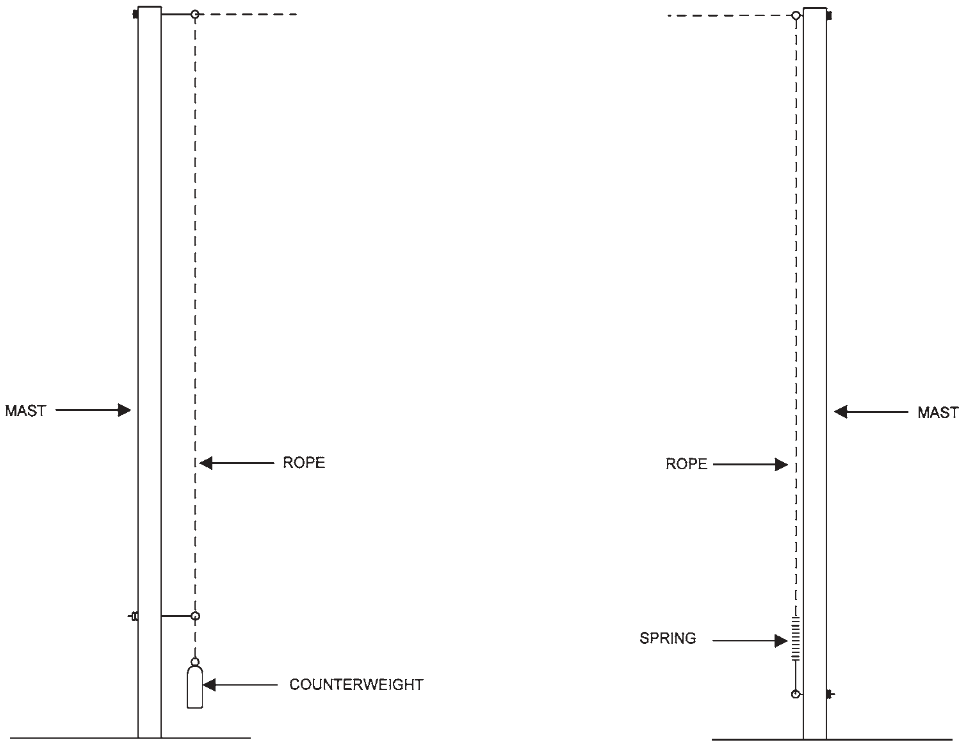

28-13 Mounting mast schemes.

top, as is true in all too many installations, but rather is brought down to ground level. This approach facilitates raising and lowering the antenna for maintenance and tuning. Be sure to provide enough “dead slack” to make lowering the antenna feasible.

The antenna support rope is connected to a spring on one end and a counterweight on the other end. The spring should be stout enough so that it is only slightly extended when the full weight of the antenna places it in tension. I have used doorsprings for this task, but only the stoutest varieties.

The counterweight should be just enough to balance the weight of the antenna and keep it not quite taut. As wind moves the antenna up and down, the counterweight raises and lowers, thereby reducing the chances of straining the antenna wires to their limit. Any number of counterweights can be used. I have seen a number of types: drapery cord weights, a small bucket of rocks, a gaggle of fishing weights, and (in one case) a burned-out automobile starter motor (the mounting hole on the front boss of the motor was ideal for accepting the rope).

Figures 28-14A and 28-14B show two methods for making the connections at the top of the mast. Although egg insulators are shown here for the sake of fairness (the alternate form was shown earlier), either form can be used. A pulley is mounted on the top of the mast with a link section and a stout eyebolt that passes through the mount. In the case of a tree, I have used a steel band of the type used to mount TV antennas to house chimneys. By making the down rope a closed loop, you gain the

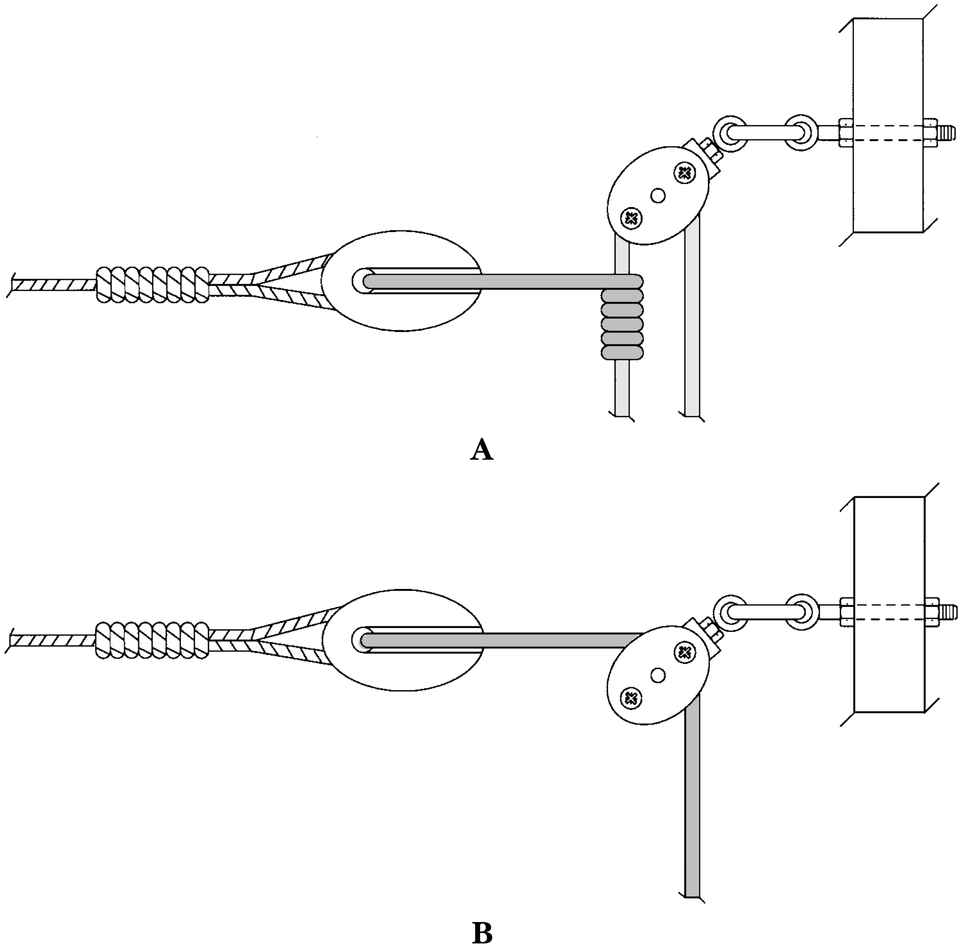

28-14 Two alternate pulley systems for wire antenna mounting.

ability to raise and lower the antenna. But do not let a loose end slip, or you will be chasing the rope up the mast as it slips through the pulley.

Pulleys work nicely, but they are mechanical contraptions that tend to fail occasionally. When they fail, the rope gets stuck. Pulleys also can corrode rather badly, with the same end result. All in all, a pulley is not the best solution for most people.

Figure 28-15 shows a better way. In this approach, a U-bolt is fastened to the top of the support mast. If the U-bolt is made of brass, then all the better, for it will not corrode. The rope can be passed through the U-bolt as in the pulley system but without the mechanical failure problem. Be sure to install the U-bolt all the way up to the threads so that the rope will not chafe against them.

The method shown in Fig. 28-16 can only be used by a horse lover! This method uses a brass stirrup in place of the U-bolt. If you have switched to cars or trucks and no longer need your horse stirrups, then perhaps this is a viable means of attaching the support rope to the top of the mast.

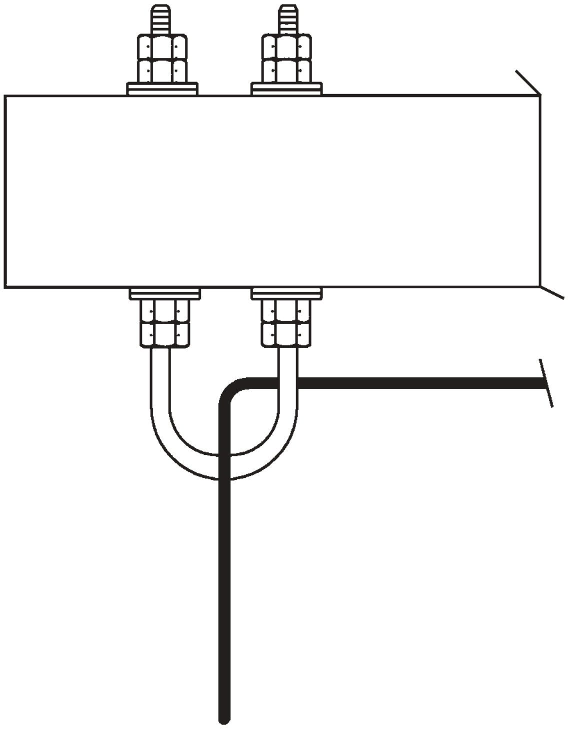

28-15 U-bolt end mounting.

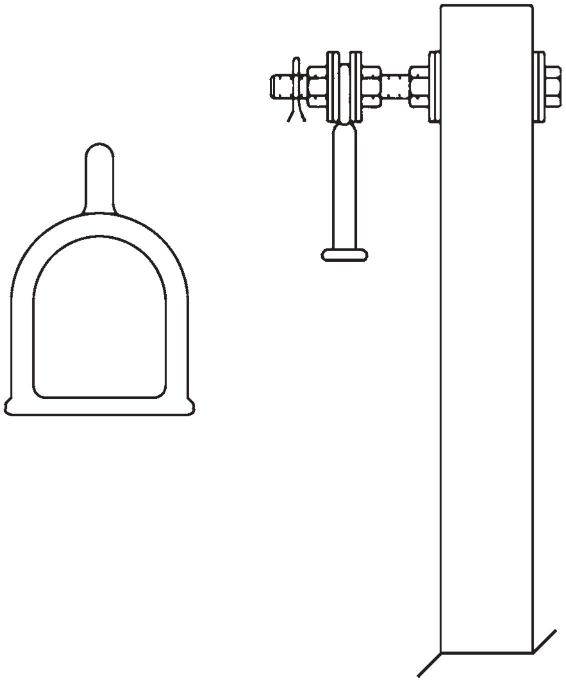

28-16 Stirrup end mounting.

Installing vertical antennas

Despite continuous (and mostly untrue) stories of very poor performance, the four- or five-band commercially manufactured trap vertical antenna remains one of the most popular amateur radio “antlers” in the HF bands. They are economical compared with directional rotating beam antennas, easy to erect, and do not have to occupy a lot of real estate (the “footprint” of the vertical can be very small, especially if you bury the radials). Unfortunately, a misinstalled vertical is both dangerous and a very poor performer. This section covers both problems in order to give you a good chance of success.

Electrical installation

Figure 28-17 shows the usual form of multiband trap vertical antenna. Each trap (*TR1*–*TR3*) is a parallel resonant LC tank circuit that blocks a certain frequency but passes all others. In Fig. 28-17, *TR1* is the 10-m trap, *TR2* is the 15-m trap, and *TR3* is the 20-m trap. No 40-m trap is needed because the antenna resonates the entire length of the tubing on 40 m. Each section (except perhaps the 10-m section) is actually a little shorter than might be expected from the standard quarter-wavelength formulas. This is because the traps tend to act inductively and so lessen the length required to resonate on any given band.

The vertical antenna manufacturer might give suggested lengths for the various segments between traps. Do not make the mistake of assuming that these are absolute numbers. They are only recommended starting points, even though the literature packed with the antenna may suggest otherwise. Loosely (meaning do not tighten the clamps too much) but safely install the antenna, and then adjust each segment for resonance. Start with the 10-m band, and then work each lower-frequency band in succession: 10-15-20-40, etc. After each lower band is adjusted, recheck the higher bands to make sure that nothing has shifted, because there might be a little interaction between bands. Once the antenna is properly resonant, tighten the clamps, and make the final installation. I know this is a pain in the neck, and means putting the antenna up and taking it down a couple of times, but it pays dividends in the end. I failed to do this once and found that the 15-m band was useless; it resonated at 19.2 MHz.

Radials make or break a vertical antenna; they form the ground plane for the antenna. AM broadcast stations typically use vertical antennas and must have 120 radials for the ground plane. Antenna reference books usually contain a graph plotted to show the effectiveness of radials and demonstrate a decreasing return on investment after about 32 radials. For amateur work, I recommend not less than 2 radials per band, and preferably 4, arranged so that they are equally spaced around the antenna. If you cannot space them correctly, never fear; they will work anyway. On a four-band antenna, this means 16 radials, which really is not a lot. Of course, the general rule is that the more radials, the better is the antenna, at least up to the point where diminishing returns are realized.

The radials are made of no. 14 wire and must be a quarter wavelength ($246/F_{\text{MHz}}$). Figure 28-18 shows how to mount the radials on a mast-mounted (or roof-mounted) installation, and Fig. 28-19 shows radials in a ground-mounted situa-

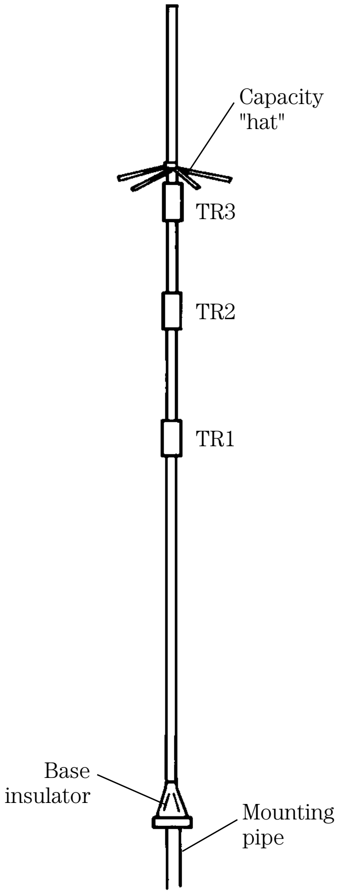

28-17 Mounting commercial trap vertical.

tion. I do not recommend radials on the surface for ground-mounted antennas. It is too easy for a guest or even a trespasser to trip over the radials, and this could land you in court. Wire radials can be buried using a spade tip to cut a 3- to 4-in slit, and they even make for a better ground.

Ground mounting verticals and masts

In the preceding section, the essential assumption was that the antenna is mounted above ground, on a mast or other support. In this section you will take a look at the methods of ground mounting the antenna or its mast. The same mast ideas are also useful for erecting supports for dipoles and other wire antennas.

Figure 28-20 shows a standard chain-link fence post used as a support for either a vertical antenna or a mast (which could be metal, PVC, or wood). The typical fence post is 1.25 to 2.00 in outside diameter and made of thick-wall galvanized steel. Thus such a pipe makes a tough installation. Some authorities believe that the pipe will last longer if painted its entire length (especially the underground portion) with a rust-inhibiting paint such as Rustoleum.

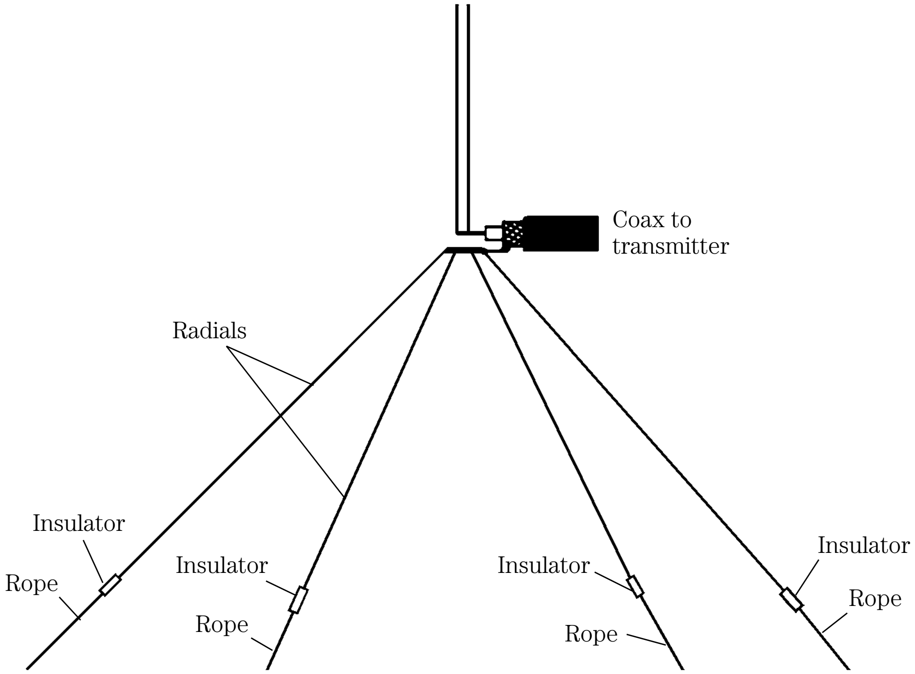
Coax to
transmitter
Radials
Insulator
Rope
Insulator
Rope
Insulator
Rope
Insulator
Rope

28-18 Feeding the mast-mounted vertical antenna.

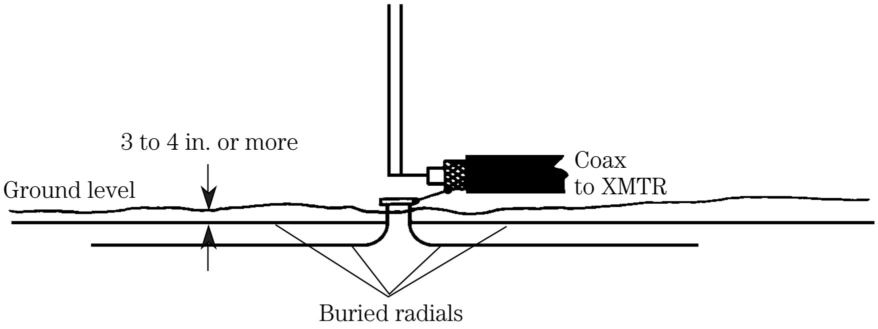
3 to 4 in. or more
Ground level
Coax
to XMTR
Buried radials

28-19 Feeding the ground-mounted vertical antenna.

The beauty of the fence post as an antenna support is that it is relatively easy to install, and supplies are obtained easily from hardware stores. The post is mounted in a concrete plug set at depth *D* in the ground (see again Fig. 28-20). The depth *D* is a function of local climate and local building codes. In the absence of local regulations to the contrary, make the hole at least 24 in deep (or more). The depth is determined by the 100-year depth of the frost line in your area, a figure well familiar to local building experts. Keep in mind, however, that shallower depths may be legal for

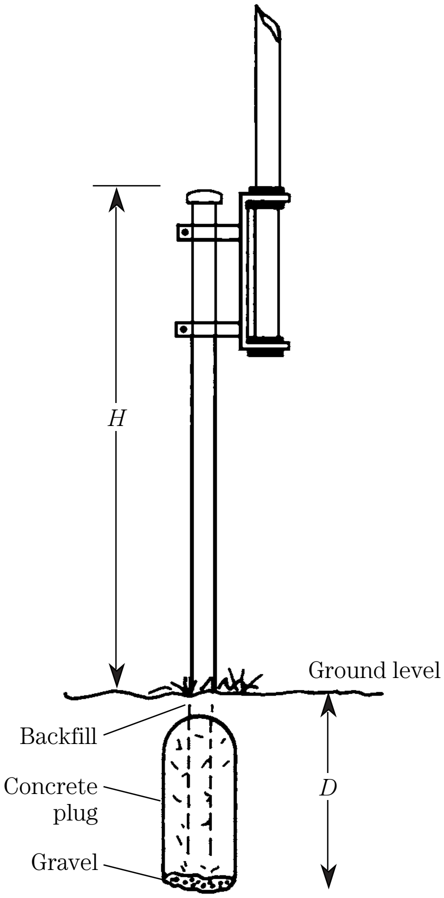

28-20 Fence post mounting.

fence posts, but installation of a vertical antenna (or mast) on top of the fence post changes the mechanical situation considerably.

The hole can be dug with a post-hole digger tool or an earth-auger-bit tool. The latter often can be rented in either gasoline engine, electrical motor, or manual versions from tool rental stores.

Once the hole is dug, place about 4 in (or local requirement) of gravel at the bottom of the hole. This gravel can be bought in bags from hardware stores for small jobs. The post is installed at this point and made plumb (i.e., it is made to sit vertical in the hole). Gravel can be used to force the post to remain upright as the job is finished. On top of the gravel is a concrete plug that fills the hole to at least 4 in below the surface. The rest of the hole is backfilled with dirt, unless you want the plug to extend all the way to the surface. Do not put anything on top of the post for at least 4 days or whatever period of time the concrete manufacturer recommends. Concrete needs time to cure, and 4 to 7 days seems to be the recommended period of time. It is probably a good idea to install the post one weekend and then complete the antenna installation the following weekend.

Chain-link fence posts tend to be about 4 ft high, and this is sufficient for installation of masts for inverted-V and dipole support. However, for vertical installation, make the height H of the post about 18 in (otherwise, the antenna tries to act differently from the instructions).

If the antenna is a standard vertical with the usual form of offset base (as shown in Fig. 28-20), then use either the mounting brackets that come with the antenna or U-bolts fastened around the fence post. Never use a single mounting point; always use at least two-point mounting to prevent the antenna from pivoting and shearing off the mounting hardware.

Figure 28-21 shows how to mount a 2 × 4 mast to a fence post. Good-quality 2 × 4 lumber can be purchased in lengths up to 20 ft, although, 8-, 10-, 12-, and 16-ft lengths are the most commonly available (in my area, 16- and 20-ft lengths have to be bought at professional "contractor" lumberyards rather than those which serve homeowners). The 2 × 4 selected should be kiln-dried and pressure-treated. It also should be painted in order to minimize the effects of weathering on the mechanical integrity of the installation. Good-quality 2 × 4 lumber is relatively inexpensive, but

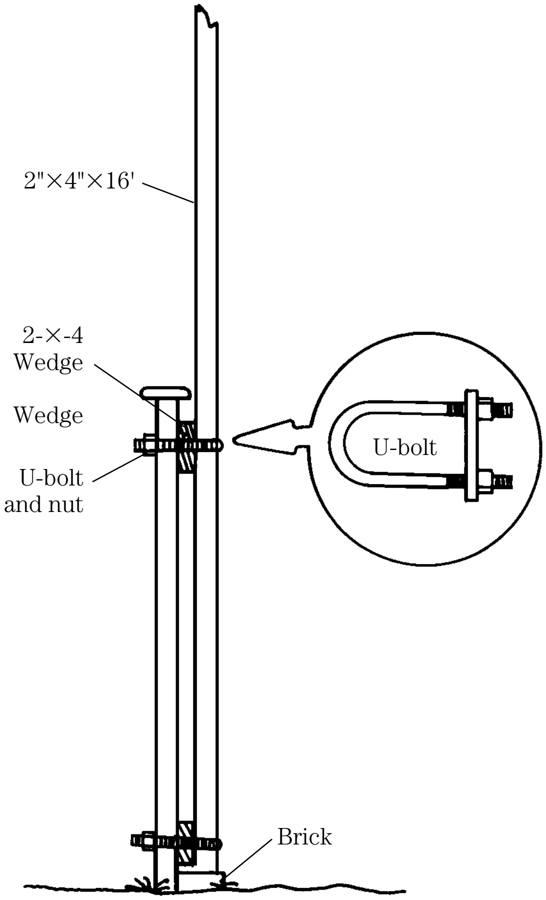

28-21 Making a mast mounted to a fence post.

the less costly material (which is only marginally less costly than the best) will warp and ruin the installation within 1 year.

The 2 × 4 is attached to the fence post using U-bolts. A 2 × 4 scrap is used as a wedge to take up the difference between the post and the mast. In some installations the U-bolt will go around the perimeter of the 2 × 4, whereas in others a pair of holes will be drilled in the 2 × 4 to admit the U-bolt arms. The U-bolt must be at least $^1/_4$-in no. 20 thread, and either a $^5/_{16}$-in bolt or a $^3/_8$-in bolt is highly recommend if the mast supports any significant weight. Some installers like to place a brick at the base of the 2 × 4. This support bears the static and dynamic loads (i.e., gravity and wind) that eventually would cause the mast to list because of the eccentric loading on the footer. The auxiliary purpose is to keep the butt end of the 2 × 4 off the ground and thereby prevent rot. Some additional mechanical stability might be afforded by replacing the brick with a cinder block that is filled with concrete. Another approach is to install a pressure-treated 4 × 4 of the sort used to support decks (Fig. 28-22). Hardware stores and lumber yards sell the specially treated 4 × 4 posts and the concrete for the footers. These footers are dropped onto a gravel bed at the bottom of a hole (see the preceding section for a discussion of how deep the hole must be). A typical deck footer has a plastic or heavy-duty metal mount to support the 4 × 4. After the footer and mount are installed, the hole is backfilled with soil. Some people prefer to backfill, at least to a point above the mount, with builder's gravel (of the same sort as used beneath the concrete footer), or concrete can be used for maximum stability.

Installing towers

Antenna towers are used extensively in amateur and commercial radio communications systems for a very obvious reason: Antennas tend to work better "at altitude" than on the ground. Amateur operators tend to install 30- to 100-ft towers, although the 40- and 50-ft models probably predominate for practical (and in some areas

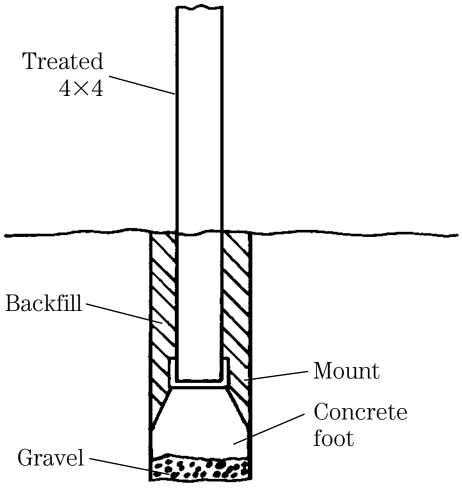

28-22 A 4 × 4 antenna support.

legal) reasons. Local laws vary considerably, from one area to another. Be sure to consult the bureaucracy before erecting a tower.

In my own area, a 30-ft tower that is physically attached to a two-story house does not need a permit or inspection (but must be constructed according to certain regulations), although anything mounted away from the house (or at a height greater than 30 ft, if attached to the house) requires a permit and the usual mechanical inspection. The law requires that an inspection be performed when the hole for the footing is completed and 4 in of gravel is on the bottom. The inspector will take a yardstick and measure both the depth of the hole and the depth of the gravel. A final inspection is performed when the antenna is installed. Prior to all this mess, a permit must be obtained (which means drawings submitted).

Several different types of towers are used by amateurs. Some are square-base designs, and others are triangular-base designs. At least one commercial model is a round pipe that has a rotator at the base. In this model, the entire tower turns when the antenna is rotated. Several tower configurations are seen commonly. Figure 28-23 shows the fixed form of tower. This type of tower costs the least, and a case can be

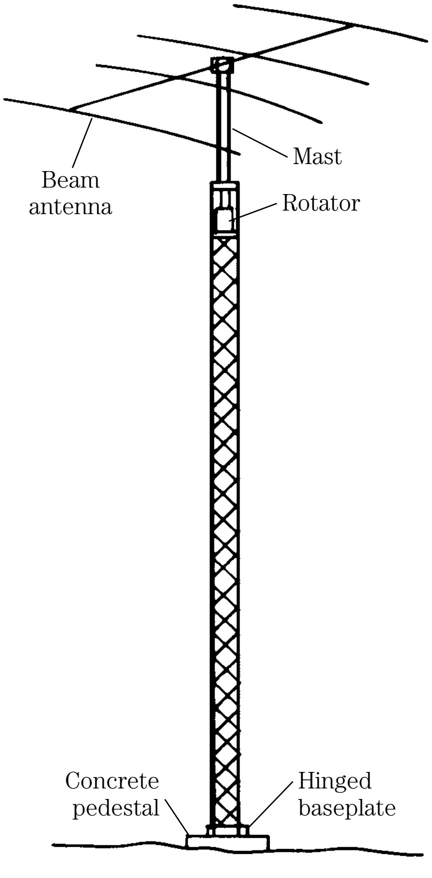

28-23 Antenna tower on hinged baseplate.

made for it also being the most stable. It is made of 10-ft sections and an 8- to 12-ft top section that contains the thrust bearing for the antenna mast and rotator mounting platform. Some fixed towers are self-supporting up to a point, but the use of guy wires is always recommended unless a firm mounting point (against the side of a building, for example) is provided about 18 ft above the surface. The fixed tower can be made to tilt over at the base by using a hinged baseplate on the mounting pedestal.

The tilt-over tower is shown in Fig. 28-24. This type of tower uses the same sort of construction as the fixed model but is hinged at a point between one-third and one-half the total height. A winch at the base drives a steel wire that is used to raise or lower the tower between fully extended and fold-over positions.

Still another configuration is shown in Fig. 28-25. This type of tower is the slip-up variety. A wide base section supports a smaller movable upper section. A winch is used to lift the upper section into its fully extended position. Typical examples will raise from a low position of 20 ft, to a height of 40 ft or so when fully extended (or 20/50 ft). When the tower is fully extended, it is locked into place at the bottom of the upper section with steel bolts. A lot of shear force is applied to these bolts, so it is wise to use several very hard stainless steel bolts for the lockdown.

There is a serious danger inherent in the design of the slip-up tower: the guillotine effect. If the upper section comes loose while you are working on it, it will come plummeting down the shaft formed by the lower section and shear off any arms, legs, or other body parts that get in the way. Whenever you work on this form of tower, use steel fence posts (or similar pieces of metal) as a safety measure (Fig. 28-26); at least two should be used, and both should be attached securely at both ends with

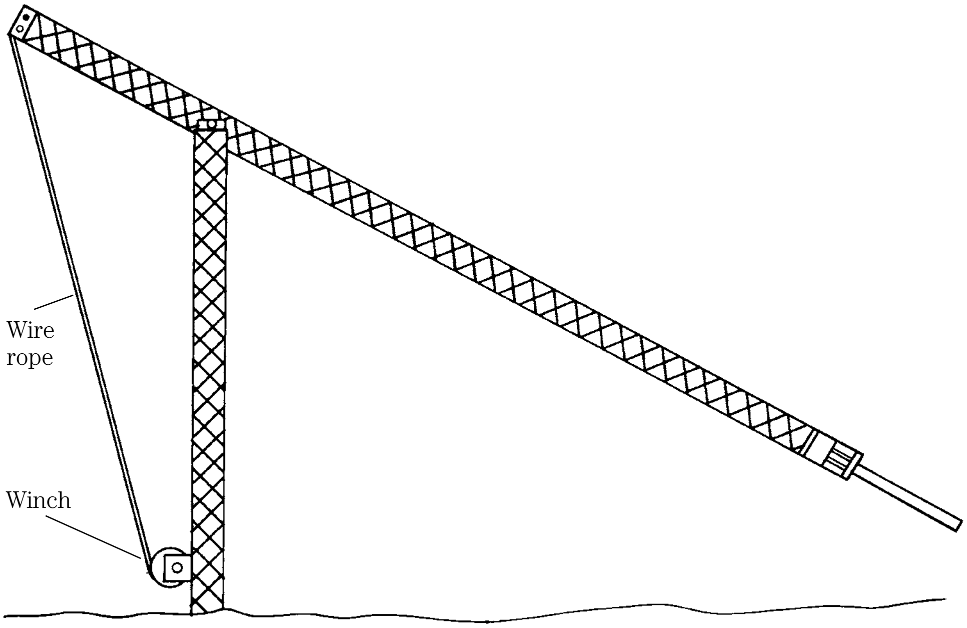

28-24 Fold-over tower.

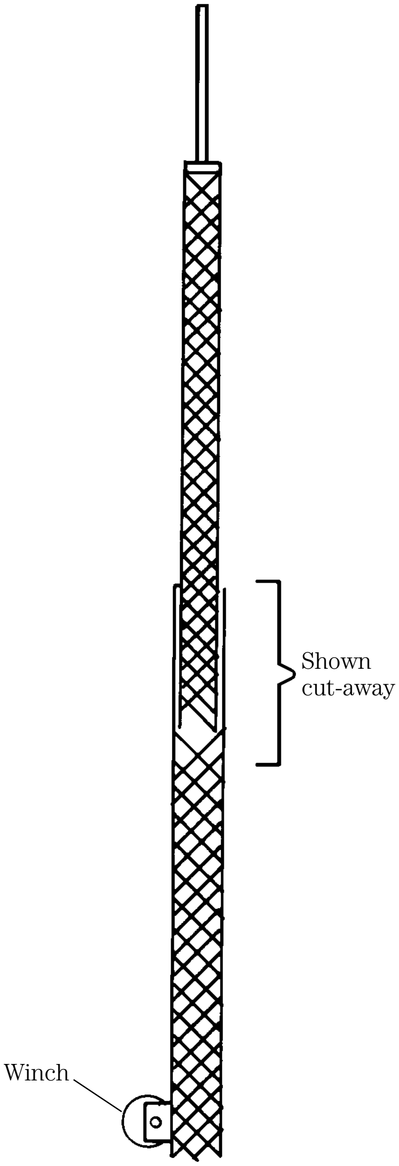

28-25 Slip-up tower.

rope. The purpose of the rope is to prevent the safety pipes from working loose and falling out. These pipes are used in addition to, not instead of, the normal bolt fasteners that keep the antenna tower erect.

The antenna tower, regardless of its configuration, is typically installed on a buried pedestal made of concrete (Fig. 28-27). The construction must follow local laws, but there are some general considerations to keep in mind. The surface area of the pedestal is 242 to 362 in. The top surface is fitted with either a fixed or hinged baseplate (which must be installed before the concrete sets). The depth of the

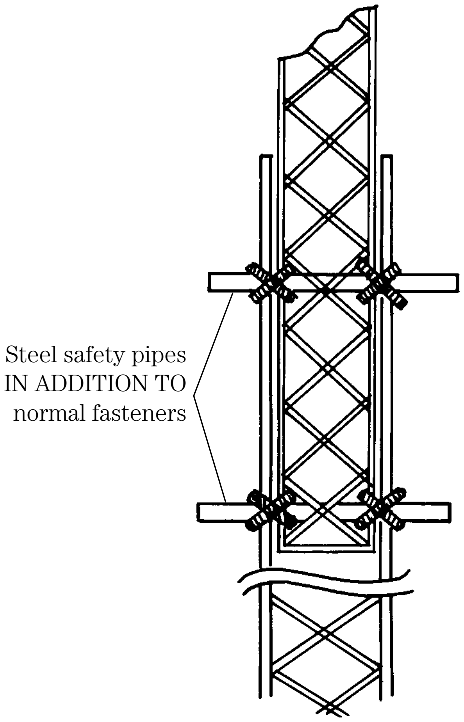
Steel safety pipes
IN ADDITION TO
normal fasteners

28-26 Safety precaution.

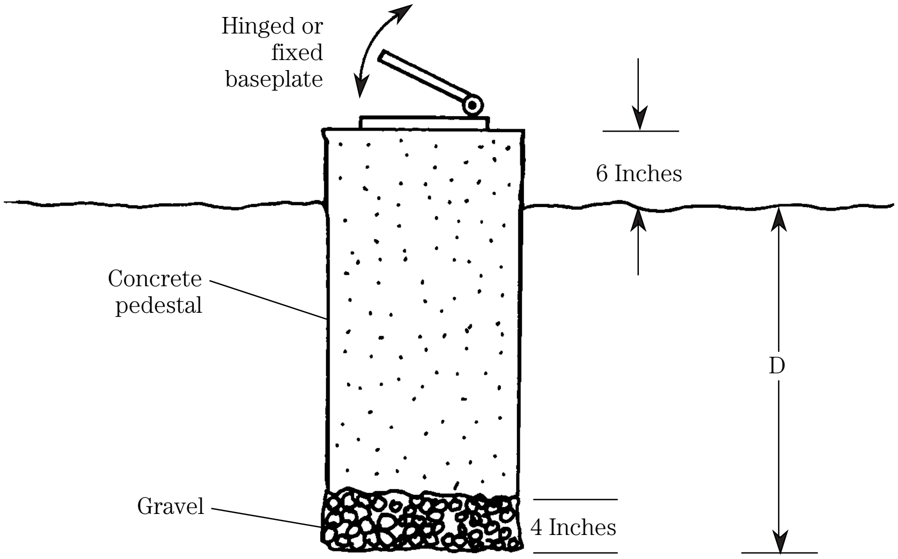
Hinged or
fixed
baseplate
6 Inches
Concrete
pedestal
D
Gravel
4 Inches

28-27 Ground pedestal.

pedestal and the depth of the gravel base tend to reflect the local frost line. Regardless of how general local law is, good advice is to make the pedestal at least 28 in deep—even if local law permits less.

Raising the tower can be accomplished by either of two methods, although, in my opinion, only one of them is safe. Figure 28-28 shows a method by which the tower is laid out on the ground and then raised with a heavy rope over a high support (such as the peak of the house roof). This method is dangerous unless certain precautions are taken. For example, a support must be placed beneath the antenna as it is raised. This requirement means that a constantly increasing height of the tower demands a constantly increasing lower support.

The end-over-end method of raising a tower is shown in Fig. 28-29. A "gin pole" is required for this job. The gin pole is a length of pipe fitted with a pulley at the top and a pair of clamps at the bottom. The gin pole is clamped onto the top end of the highest section so that it can be used to raise the next 10-ft section into place, where it can be bolted down. The gin pole is then raised to the top of the newly installed section, where the operation is repeated again. The end-over-end method is preferred by tower manufacturers and is the method recommended by most of them.

The antenna tower will be most stable if guy wires are installed at appropriate places determined by the tower manufacturer. Although many people are inclined to install towers "free standing" and some towers are advertised as such, it is always good practice to use guy wires attached to at least one point (or more, for tall towers). Use at least four guy wires on a square tower or three on a triangle tower (one in each direction). Figure 28-30 shows two methods for securing the guy wires to the ground. One approach is to bury a heavy concrete or metal "deadman" weight underground. Alternatively, use at least one 4- to 6-ft pipe (1 to 1.5-in in diameter) buried up to a length of 12 to 24 in (18 in is shown in Fig. 28-30).

Note It is absolutely essential to use safety belts when climbing a tower. Use two belts that are completely independent of each other so that no "single-point failure" can eliminate the belt's function. Always keep one belt fastened to the tower; when climbing, unhook one belt and move it up on the tower, and then disconnect the other only when the first is reattached. Be sure to inspect the belts prior to use. Leather and metal wear out and break, and you do not want to find that failure at 50 ft up. Also, do not buy cheap belts.

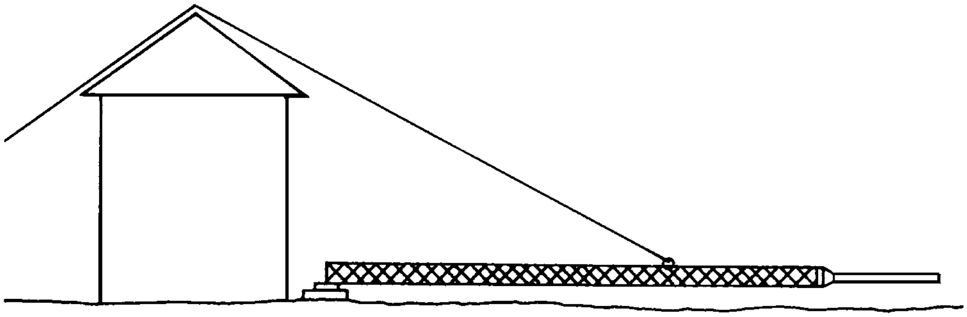

28-28 Dangerous method for erecting tower.

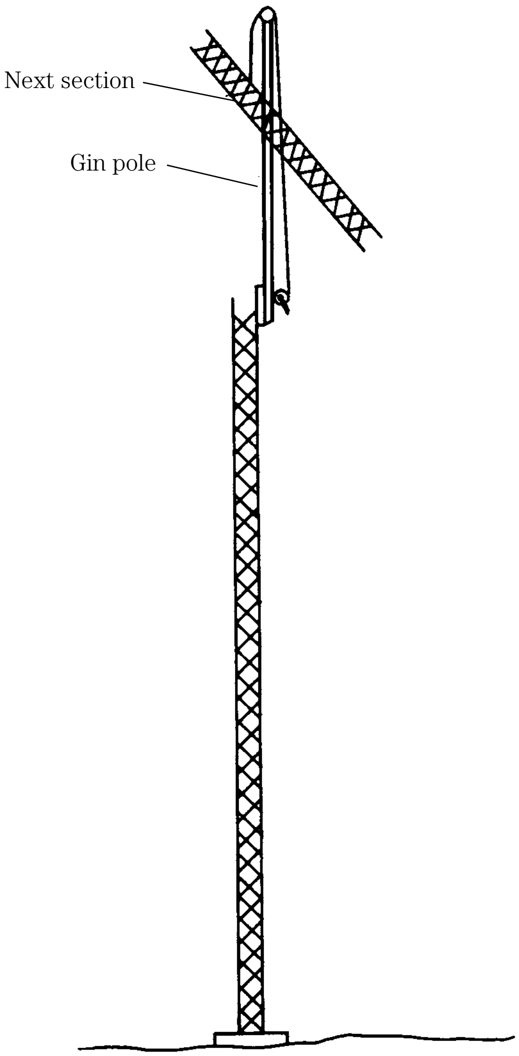
Next section
Gin pole

28-29 Proper tower installation using gin pole.

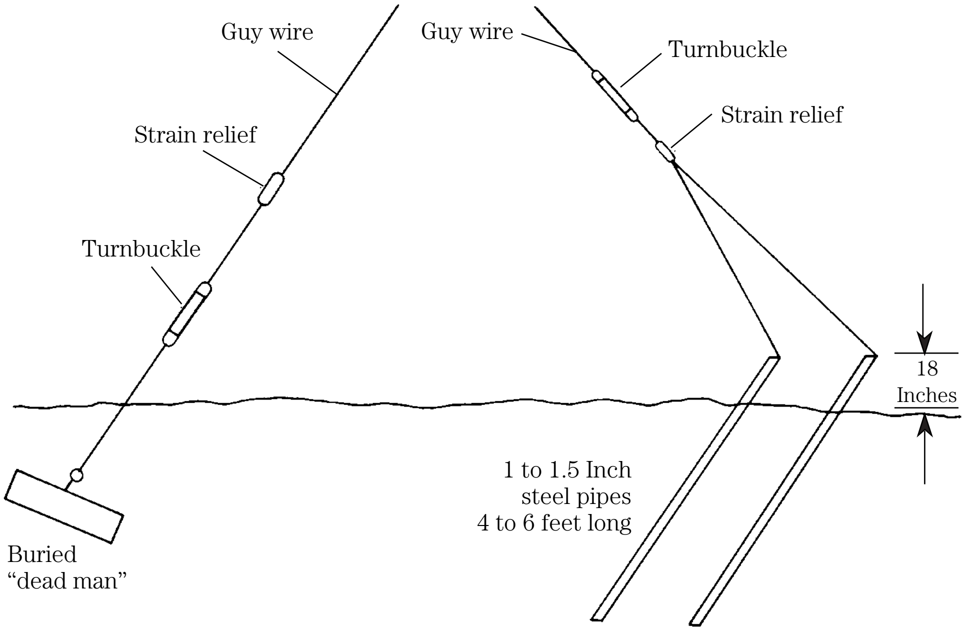
Guy wire
Strain relief
Turnbuckle
Buried
"dead man"
Guy wire
Turnbuckle
Strain relief
18
Inches
1 to 1.5 Inch
steel pipes
4 to 6 feet long

28-30 Ground stakes for guy wires.

29
CHAPTER

Grounding the antenna: What is a good ground?

THE SUCCESS OR FAILURE OF A RADIO ANTENNA SYSTEM OFTEN (PERHAPS USUALLY) hangs on whether or not it has a good RF ground. Poor grounds cause most antennas to operate at less than best efficiency. In fact, it is possible to burn up between 50 and 90 percent of your RF power heating the ground losses under the antenna, instead of propagating into the air. Ground resistances can vary from very low values of 5 Ω, up to more than 100 Ω (5 to 30 Ω is a frequently quoted range). RF power is dissipated in the ground resistance. The factors that affect the ground resistance include the conductivity of the ground, its composition, and the water content. The ideal ground depth is rarely right on the surface, and depending on local water table level might be a couple meters or so below the surface.

It is common practice among some amateur radio operators to use the building electrical ground wiring for the RF antenna ground of their station. Neglecting to install an outdoor ground that will properly do the job, they opt instead for a single connection to the grounded “third wire” in a nearby electrical outlet. Besides being dangerous to work with, unless you know what you are doing, this is also a very poor RF ground. It is too long for even the low RF bands, and radiates RF around the house in large quantity. Stations that use the household electrical wiring as the radio ground tend to cause TV interference, broadcast interference, and other electromagnetic interference both in their own building and in nearby buildings.

Fortunately, there are some fixes that will help your situation. We can reduce the ground resistance by either altering the composition of the earth surrounding the ground point, or by using a large surface area conductor as the ground point.

Figure 29-1 shows the traditional ground rod used on small radio stations, including amateur stations. Use either a copper (or copper-clad steel) rod at least 6 ft long (8 ft preferred). Electrical supply houses, as well as amateur radio and communications equipment suppliers, also sell these ground rods. Do not use the nonclad

Copyright 2001 - the McGraw-Hill Companies

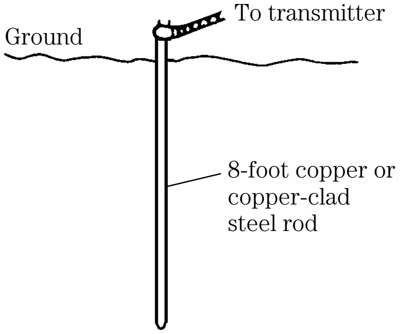

29-1 Single ground rod works—almost.

steel types sold by some electrical supply houses. They are usable by electricians when making a service entrance ground on your home or workplace, but RF applications require the low skin resistance of the copper-clad variety. The rod need not be all copper, because of skin effect forcing the RF current to flow only on the outer surface of the rod. Try to use an 8-ft rod if at all possible, because it will work better than the shorter kind. Do not bother with the small TV-antenna 4-ft ground rods; they are next to useless for HF radio stations. Drive the ground rod into the earth until only 6 in or so remains above the surface. Connect a ground wire from your station to the ground rod.

The ground wire should be as short as possible. Furthermore, it should be a low-inductance conductor. Use either heavy braid (or the outer conductor stripped from RG-8 or RG-11 coaxial cable) or sheet copper. You can buy rolls of sheet copper from metal distributors in widths from 4 in up to about 18 in. Some amateurs prefer to use 7-in-wide foil that is rated at a weight of 1 lb/linear ft. Sweat-solder the ground wire to the rod. You can get away with using mechanical connections like the electricians use, but you will eventually have to service the installation when corrosion takes its toll. I prefer to use soldered connections, and then cover the joint with either petroleum jelly or acrylic spray lacquer.

Another alternative is to use a copper plumbing pipe as the ground rod. The pipe can be purchased in 8-ft through 16-ft lengths from plumbing supply shops or hardware stores. The pipe selected should be ¾ in or larger. Some people report using up to 2-in pipe for this application. The surface area of the hollow pipe is greater than that of a solid rod of the same diameter. Because of certain current flow geometries in the system, however, the ground resistance is not half the resistance of a rod of the same diameter, but is nonetheless lower.

Driving a long pipe into the ground is not easily done. Unlike the copper-clad steel rod, the pipe has no compression strength and will deform when you hit it with a hammer or other driving tool. To overcome this problem (see Fig. 29-2) you can use a garden hose as a water drill. Sweat solder a tee joint on the top end of the pipe, and then sweat-solder a faucet fitting that matches the garden hose end on one end of the tee joint. Cap off the other port of the tee joint. Use the tee

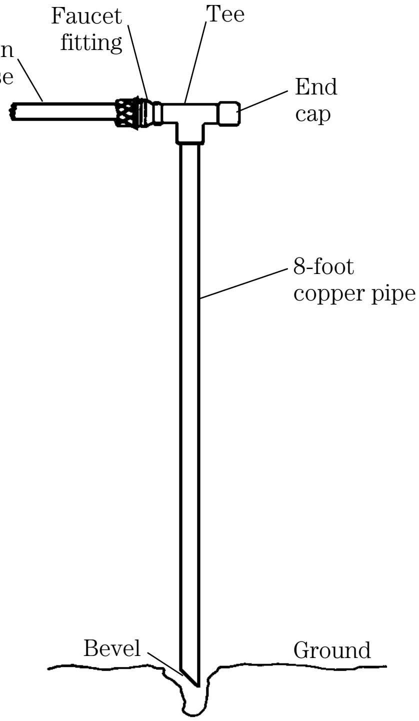

29-2 Hose method of installing ground pipe.

joint as a handle to drive the pipe into the ground. When water pressure is applied, the pipe will sink into the ground as you apply a downward pressure on the tee handle. In some cases, the pipe will slip into the ground easily, requiring only a few minutes. In other cases, where the soil is hard or has a heavy clay content, it will take considerably greater effort and more time. When you finish the task, turn off the water and remove the garden hose. Some people also recommend unsoldering and removing the tee joint.

Altering soil conductivity

The conductivity of the soil determines how well, or how poorly, it conducts electrical current (Table 29-1). Moist soil over a brackish water dome conducts best (southern swamps make better radio station locations), and the sand of the western deserts makes the worst conductor. Figure 29-3 shows a method for reducing the soil electrical resistance by treating with one of two chemicals: copper sulfate or common rock salt. The rock salt is one of several salt materials used for snow and ice melting in snow-prone areas of the country. If you cannot locate rock salt in a hardware store, then look for a store that sells ice-cream-making supplies. Rock salt is a principal ingredient in the process (but not the product).

Table 29-1. Sample soil conductivity values

Type of soil	Dielectric constant	Conductivity (siemens /meter)	Relative quality
Salt water	81	5	Best
Fresh water	80	0.001	Very poor
Pastoral hills	14–20	0.03–0.01	Very good
Marshy, wooded	12	0.0075	Average/poor
Sandy	10	0.002	Poor
Cities	3–5	0.001	Very poor

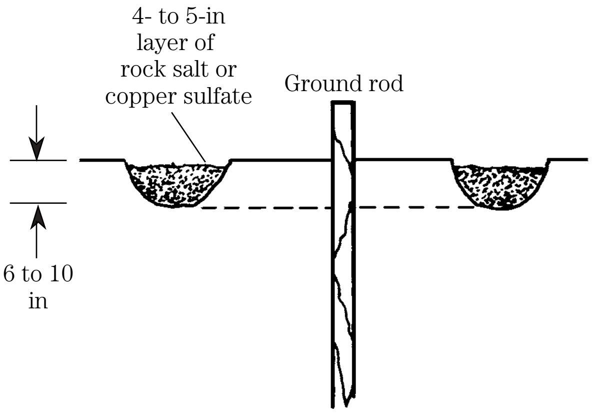

29-3A Ground rod with "improved" ground.

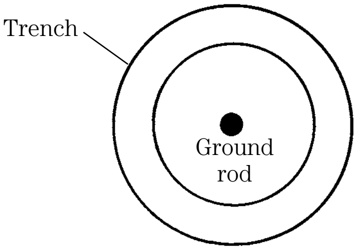

29-3B Trenching pattern.

Figures 29-3A and 29-3B show side and top views (respectively) of a slit trench method of applying the chemical treatment. Dig a 6- to 10-in-deep trench, about 12 to 24 in from the ground rod. Fill that trench with a 4- or 5-in layer of rock salt or copper sulfate. Cover the remaining depth with soil removed when you dug the trench. Water the trench well for about 15 minutes. The treatment will need to be repeated every 12 to 36 months, depending upon local rainfall and soil composition.

Figure 29-3C shows an alternative method: the salt pipe. Use either copper or PVC plumbing pipe, up to 4 in diameter (although 2 to 3 in is easier to work). The overall length of the pipe should be at least 18 to 24 in, although longer is useful.

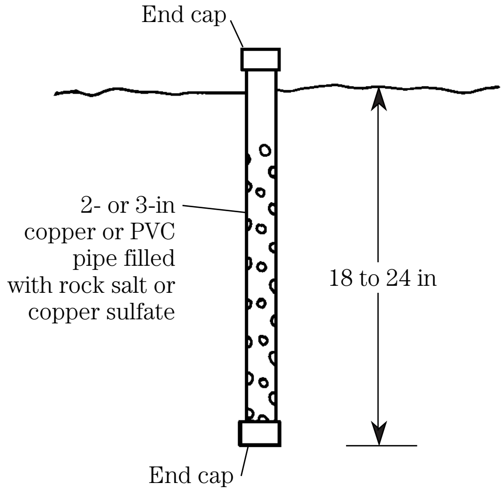

29-3C Slow release ground treatment pipe.

Drill a large number of small holes in the pipe (no hole over ⅜ in), sparing only the end that will be above the surface. Fill with rock salt or copper sulfate and cap off both ends.

Install several salt pipes around the ground rod. Installation is best accomplished with a fence posthole digger. Drop the salt pipe into the hole and backfill with water. Remove the top endcap and hose down for about 15 minutes. Refill the pipe occasionally to account for the salt leaching out.

Multiple ground rods

The key to a low-resistance ground is the surface area in contact with soil. One means for gaining surface area, and thereby reducing ground resistance, is to use multiple ground rods. Although multiple 4-ft rods are better than a single 4-ft ground rod, the use of 6- or 8-ft rods is recommended even in cases where multiple rods are used.

Figure 29-4A shows the use of three ground rods in the same system. The 8-ft rods are placed 12 to 18 in apart for low- and medium-power levels, and perhaps 30 in for higher amateur power levels. An electrical connection is made between the rods on the surface using either copper stripping or copper braid. The connections are sweat-soldered in the manner described above, with the feedpoint at the center rod.

Figure 29-4B shows a more complex system with a larger number of rods laid out in an array. The rods are arrayed about 24 in apart. As in the previous case, the rods are electrically connected (by braid or foil) on the surface. Again, connection to the system is made from the rod at the geometric center of the array.

Radials/counterpoise grounds

The effectiveness of the ground system is enhanced substantially by the use of radials either above ground or buried under the surface. Figure 29-5 shows a vertical

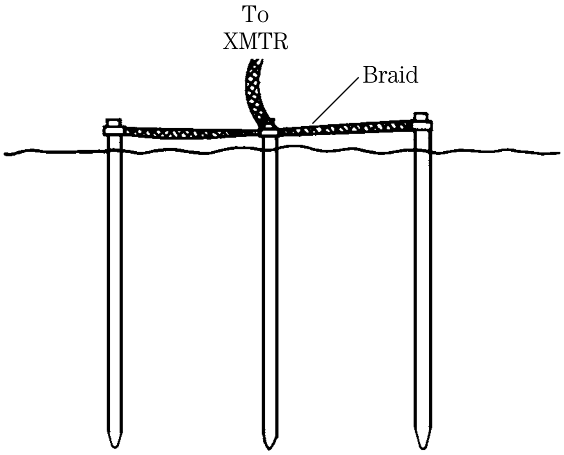

29-4A Multiple ground rod system.

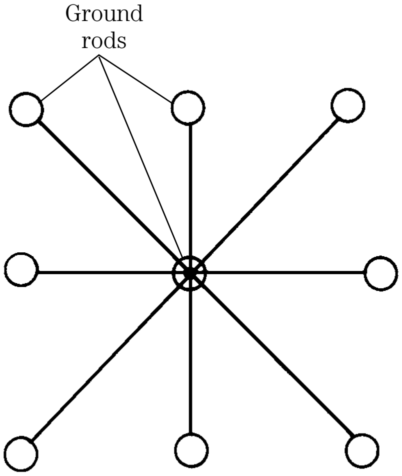

29-4B Connection scheme.

antenna with three different forms of radials: aboveground, subsurface, and ground rod. It is not unreasonable to use both radials and a ground rod. Note (from Chap. 7) that vertical antennas are relatively ineffective unless provided with a good ground system, and for most installations that requirement is best met through a system of ground radials.

An effective system of radials requires a large number of radials. Although as few as two quarter-wavelength resonant radials will provide an improvement, the best performance is to use more. Broadcasters in the AM band (550 to 1640 kHz) are advised to use 120 half-wavelength radials. Installing more than 120 radials is both expensive and time-consuming, but does not provide any substantial improvement. For amateur and small commercial stations, use a minimum of 16 quarter-wavelength radials.

Above ground, the use of insulated wire is recommended, but not required. Below ground, noninsulated wire is preferred. Although some sources claim that any

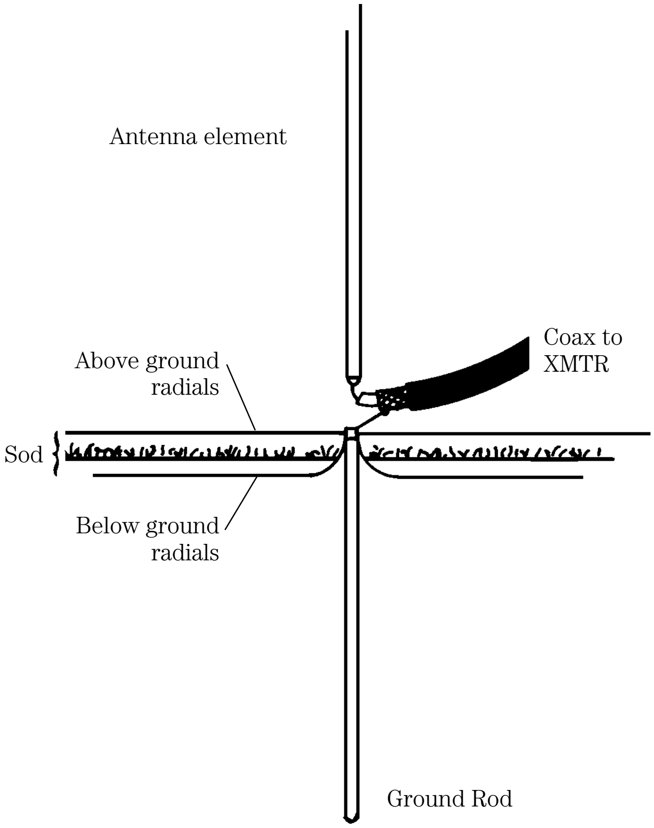

29-5 Combining ground rod and buried radials.

size wire from no. 26 up to no. 10 can be used, it is best to use larger sizes in that range (i.e., no. 14 through no. 10). Either solid or stranded wire can be used.

The layout for a system of radials in a vertical antenna system is depicted in a view from above in Fig. 29-6. Here, the radials are laid out in a uniform pattern around the antenna element. This coverage provides both the lowest resistance and the best radiation pattern for the antenna. Solder all radials together at a common point, which might be the ground or mounting rod used to support the vertical antenna.

Station grounding

It does no good to provide a topflight ground system, such as those shown earlier in this chapter, if the connection between the station equipment and the ground system is substandard. Figure 29-7 shows a method used by the author to good effect. On the back of the operating position is a sheet of copper, 7 in wide, running the length of the equipment platform. This form of copper, in the 1-lb/sq ft weight, is used on older houses for roofing flashing. Each piece of equipment is connected to the ground sheet through a short length of braid harvested from RG-8 or RG-11 coaxial cable shield. RF accessories, such as the low-pass TV interference filter (if used—it should be) are mounted directly to the copper sheet. In one installation, the

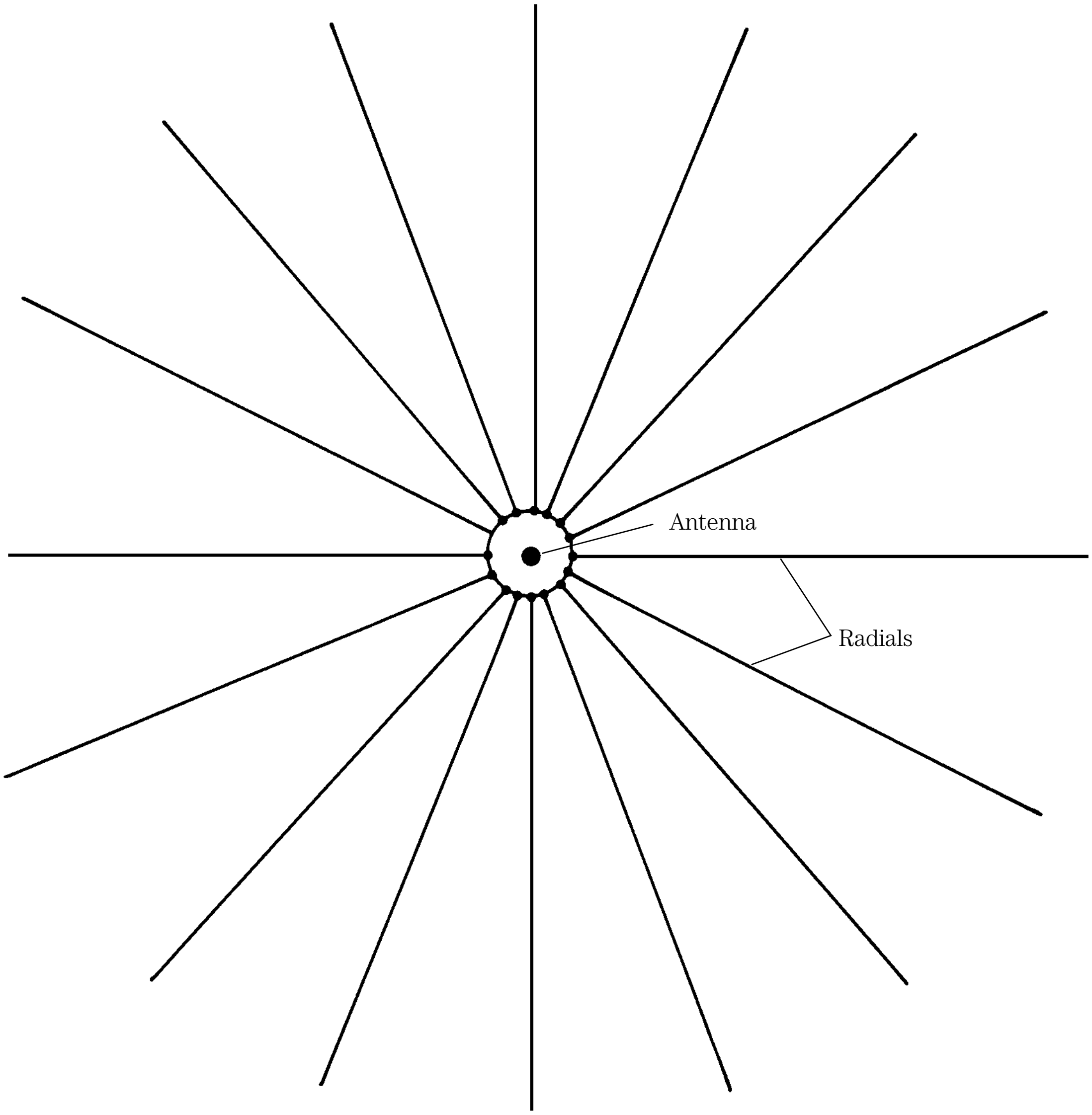

29-6 Radial pattern.

author was able to drop the copper sheet down from the table to connect directly to the ground system outside the building. The run was less than 40 in. But in other cases, a short length of braid wire will be more practical.

Tuning the ground wire

An alternative that some operators use is the ground wire tuner. These instruments insert an inductor or capacitor (or a *LC* network) in series with the ground line. You

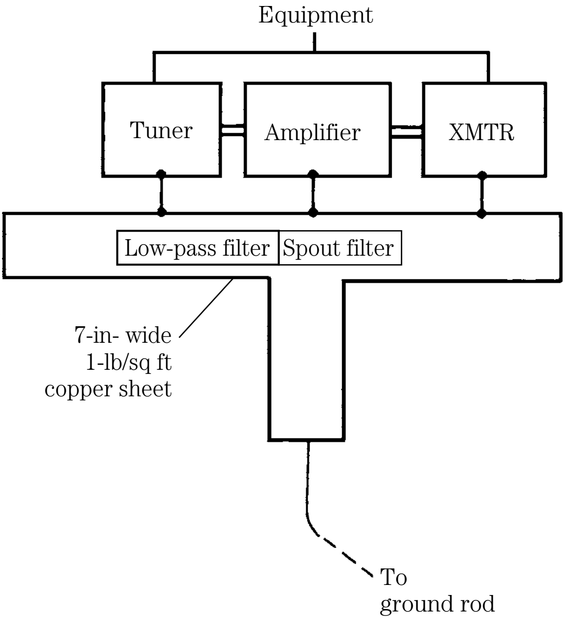
Equipment
Tuner
Amplifier
XMTR
Low-pass filter
Spout filter
7-in- wide
1-lb/sq ft
copper sheet
To
ground rod

29-7 Ground system inside shack.

adjust the ground line tuner for maximum ground current at the operating frequency. MFJ Electronics, Inc., Mississippi State, MS, makes one of these devices.

Conclusion

A high-quality, low-resistance ground might seem costly to install, but in reality it pays rich dividends in the functioning of your antenna. Don't overlook the quality of the ground, or you might be in the position of being penny wise and pound foolish.

This page intentionally left blank

Index

Copyright 2001 - the McGraw-Hill Companies

This page intentionally left blank

Note: boldface numbers indicate illustrations; italic *t* indicates a table.

A

absorbed power (P_a), transmission lines, 113–114
absorption of microwaves, **38**
Adcock antennas, 446–447, **448, 449**
Adcock, F., 446
adjustable plunger or short-circuit termination, microwave waveguides and antennas, 388, **388**
adjustment techniques (*See* measurement and adjustment techniques)
admittance, 100, 105, 112–114
air core frame loops (“box” loops), 303–305, **304**
air-articulated line, 62–63, **63**
airborne radio, 30
Alexanderson alternators, 2
aluminum vs. copper radiators, in high–frequency dipoles and doublets, 154
AM radio, 47, 55
  Beverage or “wave” antenna in, 224–228, **226, 228, 229**
  vertically polarized HF antennas in, 186
  radio direction finding (RDF) antennas in, 439
AM/FM broadcast antenna tower, **4**
amateur radio, 2, 3, 29, 49, 55
  loop antennas, small receiver-type in, 302
  multiband and tunable-wire antennas in, 203
  vertically polarized HF antennas in, 176, 179, 186
ammeter, 530, **530**
amplifier, for loop antennas, small receiver-type, 313–316, **316**
amplitude modulation (AM), 2, 8, 29
angle of incidence, 16, **16**, 22
angle of radiation, 50, 214, 274, **274**
angle of reflection, 16, **16**, 75
angle of reflection coefficient, **103**, 106, 112
angle of transmission coefficient circle, Smith charts, **103**, 106
angled or bent dipole antenna, 236–237, **240**
antenna matching tuner, 281, **285**
antenna noise temperature, 417–420
antenna tuning unit (ATU) (*See also* tuning), 463–464, **464, 465**
  in (λ/2 large loop antennas, 287
  in high–frequency dipoles and doublets, 152, 172
  in longwires, 214
  in multiband and tunable-wire antennas, 210
antenna tuning unit (ATU) (*Cont.*):
  in vertically polarized HF antennas, 173
antinodes, 87
aperture and aperture (slot) coupling, waveguides and antennas, 390–393, **393**, 402–403
apex angle, rhombic beam longwire, 223
approximation, in high–frequency dipoles and doublets, 142–143
arcing, 233
Armstrong oscillator circuit, 313, **315**
array antennas:
  branch feed systems for, 409, **412**
  corporate feed systems for, 409, **412**
  flatplate array antennas, 411, 413, **414, 415**
  phase shifting in, 409, **412, 413**
  slot array antennas, 411–413, **413**
  solid-state array antennas in, 408–411
  Watson-Watt Adcock array antennas, 447–450, **449, 450**
astronomy (*See* radio astronomy antennas)
atmospheric layers and radio propagation, 14–16, **15**, 18, 40–43,**42**
atmospheric pressure vs. radio propagation, 25

attenuation, 9–10, 28, 32, 36, 82, 377
attic-installed antennas, 232–234, **233**
auroral propagation, 54
automatic load control (ALC), 515
azimuth aspect, high–frequency dipoles and doublets, 147

B

backlobes in microwave waveguides and antennas, 400, **400, 401**
backscatter, 52
balanced transmission line, in multiband and tunable-wire antennas, 208–209, **209**
  (*See also* baluns)
baluns:
  coaxial cable, 465, **466**
  in directional phased vertical antennas, 249
  in high–frequency dipoles and doublets, 155, 171, **171**
  in longwires, 171, **171**
  multi-impedance, 476, **477**
  in multiband and tunable-wire antennas, 208–209, **209**, 210
  strain relief for, 553–554, **554**
  tapped, 476, **477**
  in VHF/UHF transmitting/receiving antennas, 341–346, **342, 343**
bandwidth, 2
  in loop antennas, small transmitter-type, 320–321, 320
  in VHF/UHF transmitting/receiving antennas, 339
basic radio system, 515, **516**
battery chargers/battery packs, 126, 493–494, **494**
beam antennas:
  cubical quad beam antennas, 263–267, **264, 266, 267**
  inverted bobtail curtain (Thorne array) antenna, 267–269, **268, 269**
  quad beam VHF/UHF antenna, 356, 358, **359**
  VHF/UHF transmitting/receiving antennas, 349–352, **351, 352**
  Yagi antennas, 257–262, **259**, 265, 349–352, **351, 352**
beam antennas (*Cont.*):
  ZL-special beam antenna, 20-m, 262–263,**262**
beamwidth:
  microwave waveguides and antennas, 400, **400, 401**
  parabolic “dish” antennas, 407
bends in microwave waveguides and antennas, 388–390, **389, 390**
bent dipole antenna, 236–237, **240**
Beverage or “wave” antenna, 224–228, **226, 228, 229**
  grounding of, 227–228, **228**
  installation of, 228
  insulators in, 228
  maximum effective length (MEL) in, 227
  steerable notch Beverage antennas, 229, **229**
Beverage, H.H., 225
Binns, Jack, first maritime radio distress call, 1–2
bidirectionality of loopsticks in RDF antennas, 441
bisquare loop large antenna, 296–297, **297**
boundary conditions in microwave waveguides and antennas, 375–376
“bowtie” dipoles, 154, 155, **156, 361**, 362
“box” loops, 303–305
branch feed systems for array antennas, 409, **412**
bridges:
  impedance bridges, 519–521, **520, 521**
  noise bridge, 521–522, **521, 522, 523**
  RF noise bridge, 522–527, **521**
  RF power measurement using, 532, **532**
British Broadcasting Corporation (BBC), 54
broadband impedance matching, in high–frequency dipoles and doublets, 146
broadbanded dipoles, 154–157
broadside radiation pattern in directional phased vertical antennas, 246
Brooke, E.F., 209
burn or shock hazard through antennas, 191, 195, 215, 233, 276, 279
Byrne, John, 209

C

cable TV, 67
cage dipoles, 154, 155, 157, **157**
call signs, 3
capacitance, 126, 138, 526–527
  coaxial cable, 69
  loop antennas, small receiver-type, 317
  loop antennas, small transmitter-type, 321
  transmission lines, 64
capacitance tuning of loop antennas, small receiver-type, 307–309
capacitive (probe) coupling in microwave waveguides and antennas, 390–393, **392**
capacitive antennas, 126
capacitive reactance, 140
  in high–frequency dipoles and doublets, 143
  impedance matching and, 458
  using Smith charts, 100, **102**, 117
capacitors, 124–126, **124**
  in longwires, 215
  in loop antennas, small receiver-type, 304, 307–309
  in loop antennas, small transmitter-type, 320, 321
  in Yagi antennas, 261
cardioid radiation pattern of sense antenna in RDF antennas, 441, **442**
Cartesian coordinate system used in modeling software, 328, **329, 330**
Cassegrain feed for parabolic “dish” antennas, 408–409, **409**
cavity antennas, 404, **405**
center insulators, 548–550, **549, 550, 551**
characteristic impedance (Z_o), 516
  of Beverage antenna, 226
  of longwires, 223
  of microwave waveguides and antennas, 383–384
  of transmission line, 64–67, 89

charge, 124–126, **124**, 127, 131–137, **132, 133**
charged solar particles vs. propagation, 39
choke joints in microwave waveguides and antennas, 389, **390**
circular loop transmitter antenna, 323, **323**
Citizens' Band (CB) radio, 29, 48, 166, 176, 191, 341, 496
coaxial cable, 59, 61–62, **61, 62**
  baluns, 465, **466**
  capacitance in, 69
  characteristic impedance of, 66
  coaxial vertical scanner antenna using, 365–368, **367, 368**
  cutoff frequency (F_o) of, 69
  doping out, 437–438
  Faraday connection in, 318
  Faraday shield in, 318
  length of, calculating, 438
  loop antennas made using, 318, **318**
  losses in, 369, 438
  matching section of, quarter-wavelength coax, 292
  multiband and tunable-wire antennas and, 208
  testing, 437–438, **437, 438**
  velocity factor (V_F) of, 248
  vertical coax VHF/UHF antenna, 346–349, **347, 348**
coaxial vertical antennas, 191–194, **192, 193**
coaxial vertical scanner antenna, 365–368, **367, 368**
cochannel interference in loop antennas, small receiver-type, 302
coil-loaded mobile antenna, 480–481, **480**
collinear "Franklin" array antenna, 165–166, **166**
collinear vertical VHF/UHF antenna, 349, **350**
commercial operators, multiband and tunable-wire antennas, 203
complex admittance, Smith charts, 105
complex conjugates in impedance matching, 458
CONALRAD station frequencies, 439
conductance of transmission lines, 64
conductivity of soil, 27, 28*t*
cone of silence, 17
constant reactance circles, Smith charts, 100–105
constant resistance circles, Smith charts, 97–100, **99**
construction techniques (*See* installation and construction techniques)
constructive interference, 19
continuous loading in low-frequency operation antennas, **503**, 504–505, **506**
continuous wave (CW), 55
continuous wave trains, 7, **7**
coordinate system to denote dimensions in microwave waveguides and antennas, 376–377, **377**
coordinate system used in modeling software, 328, **329, 330**
copper plumbing pipes as ground rod, 574
copper sulfate to alter soil conductivity, 575–576
copperweld wire, 214
core material of transformers, 474
corner reflector antennas, 404, **405, 406**
corona effects, 233
corporate feed systems for array antennas, 409, **412**
cosmic radiation noise, 37
counterpoise grounds, 577–579, **579**
  in emergency operations antennas, 487–488, **488**
  in longwire, 170–172, **171**
coupling loop antennas, 306–307, **306**
coupling methods for microwave waveguides and antennas, 390–393
CQD/CQE distress calls, 2
critical angle, 48
critical frequency (F_c), 43, **44**
cross–polarization, 13
cubical quad beam antennas, 263–267, **264, 266, 267**, 421
current, 123, 126, 128, 129, 131–137, **132, 133**
current (*Cont.*):
  in directional phased vertical antennas, 248
  in (λ/2 large loop antennas, 289
  in high–frequency dipoles and doublets, 143–146, **143**
  in loop antennas, large designs, 299
  in loop antennas, small receiver-type, 299
  in loop antennas, small transmitter-type, 320
  standing waves of, 134, **135**, 136, **137**, 139
  in vertically polarized HF antennas, 173, **174**
  in loop antennas, small transmitter-type, 323
current standing wave ratio (ISWR), 88
current transformer wattmeter, 534–538, **535, 536, 537**
cutoff frequency (F_c):
  in microwave waveguides and antennas, 381–383
  in parabolic "dish" antennas, 406
cycles per second, 7

D

D layer of ionosphere, 41, **42**, 46, 51
D loop (1λ) large antenna, 290–293, **291**
damping, 6
dead zones, 23
Dellinger fade, 46–47
delta feed match systems:
  in low-frequency operation antennas, 507, **509**
  in VHF/UHF transmitting/receiving antennas, 343, **344**, 345
delta loop (1λ) large antenna, 290–293, **291**, **295**
delta-fed grounded vertical antenna, 200–201, **200**
demiquad loop large antenna, 293, **294**
depth wound loop antennas, small receiver-type, 303
design equations for loop antennas, small transmitter-type, 321–322
destructive interference, 19

diamond loop (1λ) large antenna, 290–293, **291**
dielectric constants, 138
dielectric loss, 369
dielectric properties of air, 29–30, **30**
diffraction, 16–17, **17**, 32–35
dip oscillator/dip meter, 527–529, **527, 528**
dipoles and doublets, 141–172, **142**
  aluminum vs. copper radiators in, 154
  angled or bent dipole, 236–237, **240**
  antenna tuning unit (ATU) for, 152, 172
  approximation of, 142–143
  attic installation of, 232–234, **233**
  azimuth aspect of, 147
  baluns in, 155, 171, **171**
  “bowtie” dipoles in, 154, 155, **156, 361**, 362
  broadband impedance matching in, 146
  broadbanded dipoles in, 154–157
  cage dipoles in, 154, 155, 157, **157**
  capacitive reactance in, 143
  center insulators, 548–550, **549, 550, 551**
  collinear “Franklin” array antenna in, 165–166, **166**
  counterpoise longwire in, 170–172, **171**
  current in, 143–146, **143**
  directivity in, 146–247
  double dipoles in, 160
  double extended Zepp antenna in, 164–165, **165**
  drooping dipole scanner antenna in, 363, **363**
  electrical vs. physical length of, 141, 142–143, 149–151, 160, **162**
  emergency operations antennas, 490, **490**
  end or egg insulators, 551–553, **553**
  feedpoint in, 143–146, 256–257, **257**
  floor flanges mounting hardware in, 256
  folded dipole, 154–155, **156**, 362, **362**, 554–556, **555, 556**
dipoles and doublets (*Cont*.):
  gain in, 146–147
  ground plane verticals in, 172
  height of antenna vs. radiation patterns in, 148–149, **148**
  height of antenna vs. resistance in, 145–146, **145**
  hidden and limited-space antennas, 232–234, **233**
  high-Q models of, 160
  high–frequency dipoles and doublets, 141–172, **142**
  impedance and impedance matching in, 143–146, 149, 151–152, 155, 237
  inductive reactance in, 143, 157, 158*t*, 159–160, **161**
  installation plan for, **552**
  insulators for, 141, 256, 548–550, **549, 550, 551**, 551–553, **553**
  inverted-vee dipole in, 152–153, **153**, 490–491, **491**
  isotropic radiation in, 146
  length of, 141, 142–143, 149–151, 160, 232, 255–257
  limited-space designs for, 236–244
  loading coils in, 157–160, **157**
  losses in, 143–146
  microwave waveguides and antennas and, elements of, 396–398, **397, 399**
  mounting of, 141, 233, 256
  multiband fan dipole in, 169–170, **170**
  off-center-fed full-wave doublet (OCFD) in, 163, **163**
  off-center-fed nonresonant sloper (OCFS), 164, **164**
  ohmic loss in, 143–146
  ominidirectionality of, 146
  phase reversal stubs in, 165
  Q-factor in, 160
  Q-sections in, 146
  radiation patterns of, 146–152, **147**
  radiation resistance in, 143–146
  radiators of, 141
  reactance in, 143
  resistance in, 143–146, **145**
  resonance in, 143, 149–151, **150**, 237
  rhombic inverted-vee antenna in, 169, **169**
dipoles and doublets (*Cont*.):
  right-angle flange mounting hardware in, 256
  RLC networks and, 143
  rotatable dipole antennas, 239, **242**, 255–257, **257**
  shortened coil-loaded dipoles in, 157–160, **158**
  shortwave reception antennas, 272, **273**, 274
  sloping (“sloper” or “slipole”) dipole in, 153–154, **154, 155**, 236, **238**
  space diversity in, 160
  stacked dipoles in, 160–163, **162**
  stagger tuning of, 157
  standing waves and standing wave ratio (SWR) in, 144
  strain relief in, 553–554, **554**
  support ropes for, 551
  three-dimensional (ideal) radiation pattern of, 147–148, **147**
  tilted, center-fed, terminated, folded dipole (TCFTFD or T + 2 + FD, TTFD) in, 166–168
  transformers for, 146
  tunable dipoles in, 160
  tuning of, 149–152
  vee dipoles, 236, **239**
  vee-sloper antenna in, 168–169, **168**
  vertical dipole scanner antenna in, 364, **364, 365**
  vertical plane radiation pattern of, 148–149, **148**
  voltage in, 143–146, **143**
  voltage standing wave ratio (VSWR) in, 144, 146, 149–151, **150**, 160
  wavelength measurement of, 141–142
  wire antenna, basic design, 548, **548**
  Zepp antenna in, 164–165, **165**, 171
direct vs. reflected space waves, 17–19, **18**
direction of radio waves, 54
directional array, 360-degree phased, 250–254, **253**, 254*t*
directional beam antennas, 255–269

directional beam antennas (*Cont.*):
- cubical quad beam antennas, 263–267, **264, 266, 267**
- directivity of, 255
- effective radiated power (ERP) in, 255
- electromagnetic interference (EMI), 263
- feedpoints, 256–257, **257** 260–261, **261**, 263, **265**
- floor flanges mounting hardware in, 256
- gamma match feed system for, 260–261, **261**
- insulators, 256
- inverted bobtail curtain (Thorne array) antenna, 267–269, **268, 269**
- length of, 255–257, 259–260
- mounting of, 256
- quad loop antennas in, 263–267, **264, 266, 267**
- right-angle flange mounting hardware in, 256
- rotatable dipole as, 255–257, **257**
- Yagi antennas as, 257–262, **259**, 265
- Yagi-Uda antenna as, 258
- ZL-special beam antenna, 20-m, 262–263,**262**

directional discontinuity ring radiator (DDRR) antenna, 424–427, **426**, 427*t*

directional phased vertical antennas, 245–254
- broadside radiation pattern in, 246
- current in, 248
- directivity of, 246–250
- effective radiated power (ERP) in, 245–246
- feed phasing table for, 254*t*
- feedpoint in, 247–248, **247, 248**
- length of, variable lengths in, 248–249, **249**
- noise in, 245
- omnidirectionality of, 245
- phase-shifting circuit for, **252**
- phasing in, 246–250
- phasing switch in, 249–250, **252**
- radiation patterns in, 246–250, **246, 247**
- signal-to-noise ratio (SNR) in, 246

directional phased vertical antennas (*Cont.*):
- 360-degree directional array, 250, **253**, 254*t*
- three-element, 250, **253**
- transformers in, 249–250, **250, 251**
- two-port phasing transformer for, 249–250, **250, 251**
- velocity factor (V_F) of coax in, 248

directional wire antenna, for shortwave reception, 281–286, **284**

directivity:
- in directional beam antennas, 255
- in directional phased vertical antennas, 246–250
- in high–frequency dipoles and doublets, 146–247
- in microwave waveguides and antennas, 395, 397, 398–403, **400**
- in shortwave reception antennas, 272
- in VHF/UHF transmitting/receiving antennas, 349

directivity gain (G_d) in microwave waveguides and antennas, 401–402

directors, Yagi antennas, 258–259

directrix, parabolic "dish" antennas, 404

discrete loading in low-frequency operation antennas, **503**, 504–505

dish antennas (*See* parabolic "dish" antennas)

distance between antennas, 28

distance calculation, for radio, 26–27, **26**

distress calls, 2

distributed capacitance, in loop antennas, small transmitter-type, 321

disturbance, ionospheric, 45–47

diurnal (*See* time–of–day variations)

diversity propagation/reception, 55–58, **57, 58**, 160

dominant mode in microwave waveguides and antennas, 376–377

Doppler antennas, 450–451, **451**

double dipoles, 160

double ducky direction finders (DDDF), 455–456, **455**

double extended Zepp antenna, 164–165, **165**

double vee longwire antennas, 223

double-shielded coaxial, 63, **63**

doublet antennas (*See* high–frequency dipoles and doublets)

drooping dipole scanner antenna, 363, **363**

drooping dipole multiband antennas, 205

"drooping radials," in vertically polarized HF antennas, 176, **177**

ducting, 29–31, **31**, 36

dummy loads, 385–388, **386**, 540–542, **540, 541, 542**

DXing using loop antennas, small receiver-type, 311–313

E

E layer of ionosphere, 41, **42**, 46

effective aperture (A_e), 11, 394, 407

effective Earth's radius factor (K factor), 23–27

effective height of antennas, 304

effective radiated power (ERP), 45
- in directional beam antennas, 255
- in directional phased vertical antennas, 245–246
- in microwave waveguides and antennas, 401
- in VHF/UHF transmitting/receiving antennas, 349

efficiency, in loop antennas, small transmitter-type, 321

E–field components, 27

electric field, 123, 124–126, **125**, 127, **128**
- in longwires, 219
- in microwave waveguides and antennas, 383, 395–396
- in vertically polarized HF antennas, 173

electrical installation, vertical antennas, 560–561

electrical vs. physical length, 137–140, **139**
- of high–frequency dipoles and doublets, 141–143, 149–151, 160, **162**

electrical vs. physical length (*Cont.*):
  impedance matching and, 470–471
  of longwires, 213–214, 217–219, 221
  of transmission line, 67–68
electromagnetic fields, 8–16, **9**, 123, 125–127
electromagnetic interference (EMI), 263
electromagnetic radiation vs. propagation, 39
electromagnetic waves, 5
electrostatic fields/discharge, in longwires, 219
eleven–year cycle of radio propagation, 46
EM Scientific, 331
emergency operations antennas, 485–493
  battery packs for, 493–494, **494**
  counterpoise ground plane for, 487–488, **488**
  dipoles in, 490, **490**
  grounding and ground systems for, 487–488
  increasing signal strength from portable/mobile stations for, 486
  inverted-vee dipole in, 490–491, **491**
  mobile antennas and, 486–487, **487**
  mounting of, on vehicles, 491–492
  multiple radials from vehicle-mounted antenna for, 488, **489**
  radial-to-mobile-antenna connection detail in, 488, **489**
  repeaters and, 487
  telescoping antennas for, 491
  temporary antenna construction for, 487–488, **488**
  temporary mast mount for, 492, **493**
  VHF band for, 487
end effect, 138–139
end or egg insulators, 551–553, **553**
end plates/end terminators, in microwave waveguides and antennas, 385, 388, **388**
end-fed longwire, 217–219, **217**
end-fed Zepp antenna, 208, **208**
environmental/natural noise (QRN), 245
erecting antennas safely, 543–548
escape angle, 48
ether, 8
Everett, W.L., 209
Expert MiniNEC modeling software, 331
EZNEC for Windows modeling software, 331, 334–337, **335, 336, 337**
EZNEC Professional modeling software, 329

F

F layer of ionosphere, 41–43, **42**, 46
fading, 35–36, 35, 55–58, 55
  Dellinger fade, 46–47
  selective fading, 55
fan dipole, multiband, 169–170, **170**
far field calculations in microwave waveguides and antennas, 395–396
Faraday connection, 318
Faraday shield, 318
Federal Communications Commission (FCC), 2
feed geometries for parabolic "dish" antennas, 408–409, **408, 409**
feed illumination in parabolic "dish" antennas, 406
feeding transformer, 476, **477**
feedpoints:
  λ/2 large loop antennas, 287–288
  dipoles, 256–257, **257**
  directional beam antennas, 256–257, **257**, 260–261, **261**
  directional phased vertical antennas, 247–248, **247, 248**
  high–frequency dipoles and doublets, 143–146
  longwires, 217–218
  multiband and tunable-wire antennas, 204, 210
  one/1λ large loop antennas, 292
  quad beam antennas, 263, **265**
  vertically polarized HF antennas, 197, **198**
  Yagi antennas, 260–261, **261**
fence posts as supports for vertical antennas, 561–565, **563, 564**
ferrite core inductors, 471–477
ferrite rods, 475, **475**
fiber optics, 370
field strength meters (FSM), for mobile HF antennas, 483–485, **484, 485**, 485*t*
field strength, 21, **22**
figure of merit, 418
filters, 55, 515
five-eighths-wavelength antennas, 197–200
  2-meter VHF/UHF antenna, 353, **354**
  vertical antenna, 181, **182**
flagpole-installed antennas, 234–236, **235, 237**
flatplate array antennas, 411, 413, **414, 415**
flattop antennas, shortwave reception, 279, **280**
floor flanges mounting hardware, 256
flux, magnetic, 126, 127
FM, marine radio antennas, 495
fold- or tilt-over tower, 567, **567**
folded dipole, 154–155, **156**
  installation of, 554–556, **555, 556**
  tilted, center-fed, terminated, folded dipole (TCFTFD or T^2FD, TTFD) in, 166–168, **167**
  VHF/UHF antennas, 362, **362**
forward power, 538, **539**
forward scatter, 52
"fox hunting" antenna, 444, **445**
"Franklin" array antenna, 165–166, **166**
free space, 14, 19, **21**, 67–68, 137–138
free-space velocity (C) in microwave waveguides and antennas, 377, 380
frequency, 2, 7, 9
  in microwave waveguides and antennas, 373, **374**, 375, **376**, 380, 383–384, **384**
  using Smith charts, 117–121, **119, 120**
  tilt angle vs., 28, 28*t*
  VHF/UHF transmitting/receiving antennas, 339, 341
  VSWR vs., 434–437, **435**
frequency diversity, 36, 55, **57**
frequency modulation (FM), 32

frequency of optimum traffic (FOT), 45
Fresnel zones, 33–34, **34**, 35
fundamentals of radio antennas, 123–140

G

G5RV multiband dipole antenna, 206–208, **207**
gain:
  in $\lambda/2$ large loop antennas, 288–289, **289**
  in high–frequency dipoles and doublets, 146–147
  in longwires, 214, 218, **218**, 223
  in loop antennas, small transmitter-type, 319–320
  in microwave waveguides and antennas, 394–395, 398–403
  in microwave waveguides and antennas, vs. aperture in, 402–403
  in parabolic “dish” antennas, 405–408
  in shortwave reception antennas, 272
  in vertically polarized HF antennas, 179–181
  in VHF/UHF transmitting/receiving antennas, 349
galactic noise, 39, 421
gamma match feed system:
  in Yagi antennas, 260–261, **261**
  in VHF/UHF transmitting/receiving antennas, 346, **346**
gamma rays, 39
gamma-fed grounded vertical antenna, 200–201, **201**
gas-filled line, 62, **63**
ghost radiator, vertically polarized HF antennas, 173
gin pole to raise towers, 570, **571**
Global Positioning System (GPS), 443
GNEC modeling software, 329–330
graphite-sand dummy load in microwave waveguides and antennas, 385
gray line transmission, 51, **53**
great circle paths, 50–51, **52**
grid dip meter, 527–529, **528**
ground losses, vertically polarized HF antennas, 183
ground plane vertical antennas, 172, 176, **177**
ground rods, 573–574, **574**, 575, **575**, 577, **578**
ground stakes and guy wires, 570, **572**
ground wave, 17–29, **18**, 47–48, **48**, 302
ground-mounted vertical antennas, 561–565, **562**
grounded vertical antennas, 200–201
grounding and ground systems, 573–581
  in Beverage antenna, 227–228, **228**
  copper plumbing pipes as ground rod, 574
  copper sulfate to alter soil conductivity in, 575–576
  counterpoise grounds, 577–579, **579**
  in emergency operations antennas, 487–488
  ground rods for, 573–574, **574**, 574–575, **575**, 577, **578**
  in hidden and limited-space antennas, 238–239
  impedance matching and, 459–460
  interference and, 573
  length of ground wires in, 574
  in longwires, 215–216, **216**
  in low-frequency operation antennas, 501–502, **502**
  in marine radio antennas, 495–496, **495, 496**
  radials for, 577–579, **579, 580**
  “salt pipe” to apply chemical soil-conductivity improvers, 576–577, **577**
  salt to alter soil conductivity in, 575–576
  in shortwave reception antennas, 276, **277**, 278–279, **279**
  soil conductivity and, altering, 575–577, 576*t*, **576, 577**
  station grounding, 579–580, **581**
  trenching to apply chemical soil-conductivity improvers, 576, **576**
grounding and ground systems (*Cont.*):
  tuning ground wire, 580–581
  in vertically polarized HF antennas, 176, 185–187, 195–197, **196**
groundplane VHF/UHF antenna, 356, **357**
groundplane scanner antenna, 364, **366**
group velocity (V_g) in microwave waveguides and antennas, 377–379
Grover’s equation, for loop antennas, small receiver-type, 302–303, 303*t*
guy wires and ground stakes, 570, **572**

H

“hairpin” loading scheme, 505, **506**
half-delta sloper (HDS), 295–296, **296**
half-slope vertical antenna, 238, **240**
half-wave antennas, 126, 181, **182**
  current distribution in, 131–137, **132, 133**
  electric vs. physical length of antenna, 137–139, **139**
  impedance matching in, 139–140, **139**
  standing waves and, 129
  voltage distribution in,131–137, **132, 133**
half-wavelength line, 83
halo VHF/UHF antenna, 356, **358**
hardline, 62, **62**
harmonics, 140, 209, 210, 515
height of antenna, 19, **20**, 25, 28
  distance calculation for, 26–27, **26**
  in high–frequency dipoles and doublets, vs. radiation patterns in, 148–149, **148**
  in high–frequency dipoles and doublets, vs. resistance in, 145–146, **145**
  in loop antennas, small receiver-type, 304
  in vertically polarized HF antennas, 181, 187
helical antennas, 427–428, **429**
helical line, 62

helical whip mobile antenna, 481, **482**
helically wound antennas, **239**, 242, **243**
hertz as unit of frequency, 7
Hertz or hertzian antenna (*See* high–frequency dipoles and doublets)
Hertz, Heinrich, 5, 7
heterodyning, 55
hidden and limited-space antennas, 231–244
  angled or bent dipole, 236–237, **240**
  attic-installed, 232–234, **233**
  dipoles as, 232–234, **233**, 236
  flagpole installation for, 234–236, **235, 237**
  ground system for, 238–239
  half-slope vertical antenna as, 238, **240**
  helically wound antennas as, **239**, 242, **243**
  L-section couplers in, 238–239
  limited-space designs for, 236–244
  longwires as, 232
  mobile antenna used as windowsill-mounted, 239, **241**
  mounting of, 233
  nonresonant loop antennas as, 234–235, **234**
  rotatable dipole antennas as, 239, **242**
  safety issues for, 233
  sloping ("sloper" or "slipole") dipole as, 236, **238**
  vee dipoles as, 236, **239**
  voltage standing wave ratio (VSWR) in, 238
  window-sill antennas as, 238, **241**
high frequency (HF), 3, 13, 17, 29, 35, 37, 41, 48, 61, 170–171, 203, 479–485
high-frequency dipoles and doublets (*See* dipoles and doublets)
high-power RF wattmeters, 538, **538**
high-Q dipoles and doublets, 160
hinged baseplate tower, 566–567, **538**
history of radio broadcasting and communications, 1–4
hopping, 36
horizontal low-frequency operation antennas, 505
horizontal polarization, 12–13, **13**, 27, 55, 173
  Beverage antenna and, 226–227
  helically wound antennas and, **239**, 242, **243**
horn antenna radiators, 403–404, **403**
Huygens–Fresnel wave theory, 32

I

I^2R losses in microwave waveguides and antennas, 385
image antenna, vertically polarized HF antennas, 173
imaginary operator (j), 22
impedance and impedance matching, 140, 433, 457–477, 525
  antenna tuners for, 463–464, **464, 465**
  in antennas, 139, **139**
  approaches to, 458–462, **458**
  in Beverage antenna, 226
  capacitive reactance in, 458
  characteristic impedance (Z_o) of transmission line and, 64–67
  in coaxial cable baluns, 465, **466**
  complex conjugates in, 458
  in dipoles, 237
  electrical vs. physical length calculations in, 470–471
  feeding transformer for, 476, **477**
  ferrite core inductors in, 471–477
  ferrite rods in, 475, **475**
  grounding and, 459–460
  in high–frequency dipoles and doublets, 143–146, 149, 151–152, 155
  impedance bridges in, 519–521, **520, 521**
  impedance matching network (Q sections) in, 85
  inductive reactance in, 458
  L-section network for, 458–460, **459**
  in λ/2 large loop antennas, 287–288
  in longwires, 214, 223
  in low-frequency operation antennas, 505
impedance and impedance matching (*Cont.*):
  matching stub systems for, 465–466, **466, 467**
  matching units for, 515
  in microwave waveguides and antennas, 383–384, **384**, 394, 396
  multi-impedance balun for, 476, **477**
  in multiband and tunable-wire antennas, 210
  noise and, 417–418
  normalization of impedance in, 469–470
  in 1λ large loop antennas, 292–293, **292**
  pi networks for, 461, **461**
  quarter-wave matching (Q section) sections for, 467–468, **468, 469**
  radials in, 460
  series matching section for, 468–471
  using Smith charts, 110–112, **111**, 115, 119–121, **119**
  using Smith charts, impedance radius in, 100
  using Smith charts, normalized impedance line in, 95–97, **98**
  SPC transmatch network in, 463, **463**
  split-capacitor networks for, 461–462, **462**
  stub matching systems for, 115
  tapped balun for, 476, **477**
  tee-network transmatch for, 463, **463**
  toroid broadbanded transformers for, 475–477, **476**
  toroid cores and, 471, **472**
  toroidal transformers and, 471–475, **472**
  in towers, 507, **508**
  transmatch circuits for, 462–471, **462**
  in transmission lines, 71, 72, 82–86, 109, 110–112, **111**, 112
  in vertically polarized HF antennas, 173, 175–176, 183, **184, 186**, 195–197, 199–200, **199**
  in VHF/UHF transmitting/receiving antennas, 341–346, **342**

- impedance and impedance matching (*Cont.*):
  - Wheatstone bridge in, 519–521, **520**
  - in Yagi antennas, 260–261, **261**
- impedance bridges, 519–521, **520, 521**
- incident angle, 48
- incident power (P_I) of transmission lines, 89
- incident waves, 128, 134
- index of refraction, 23–24
- inductance, 526–527
  - ferrite core inductors in, 471–477
  - Grover's equation in, 302–303, 303*t*
  - in longwires, 214
  - in loop antennas, small receiver-type, 307
  - in loop antennas, small transmitter-type, 321
  - in transmission lines, 64
- induction fields, 5, 123, 127
- inductive (loop) coupling in microwave waveguides and antennas, 390–393, **393**
- inductive loading in low-frequency operation antennas, **503**, 504–505
- inductive reactance, 140
  - in high–frequency dipoles and doublets, 143, 157, 158*t*, 159–160, **161**
  - impedance matching and, 458
  - in $\lambda/2$ large loop antennas, 288–289, **289**
  - in loop antennas, small transmitter-type, 319, 321
  - using Smith charts, 100, **101**, 117
- infrared (IR) radiation, 9
- installation and construction techniques, 433–438, 543–572
  - center insulators, 548–550, **549, 550, 551**
  - electrical installation, vertical antennas, 560–561
  - end or egg insulators, 551–553, **553**
  - fence posts as supports for vertical antennas, 561–565, **563, 564**
  - fold- or tilt-over tower, 567, **567**
- installation and construction techniques (*Cont.*):
  - folded dipole installation, 554–556, **555, 556**
  - four/4x4 support posts for, 565, **565**
  - gin pole to raise towers, 570, **571**
  - ground-mounted vertical antennas, 561–565, **562**
  - guy wires and ground stakes, 570, **572**
  - hinged baseplate tower, 566–567, **538**
  - insulators, 548–550, **549, 550,** 551–553, **551, 553**
  - mast-mounted vertical antenna, 546, 560–561, **546**, **561, 562**
  - masts and supports, 556–559, **557, 558**
  - pedestals for towers, 568–570, **569**
  - permits, zoning, and regulations, 547–548
  - pulley system for mounting, 557–558, **558**
  - raising towers, 570, **570, 571**
  - safety issues, 543–548
  - slip-up tower, 567–568, **568, 569**
  - stirrup-end mounting hardware, 558, **559**
  - strain relief, 553–554, **554**
  - support ropes, 551
  - suburban lot installation plan, **545**
  - tower installation, 565–572, **538**
  - U-bolt mounting hardware, 558, **559**
  - vertical antenna installation, 560–572
  - wire antenna, basic design, 548, **548**
- insulators (or spreaders), 548–550, **549, 550, 551**, 551–553, **553**
  - in Beverage antenna, 228
  - in dipoles, 256
  - in directional beam antennas, 256
  - in high–frequency dipoles and doublets, 141
  - in microwave waveguides and antennas, 371–372
  - in multiband and tunable-wire antennas, 206
  - in vertically polarized HF antennas, 175, 195
- intensity of radio waves, 9–10
- interference, 19, 36, **36, 37**, 573
  - electromagnetic interference (EMI), 263
  - in longwires, 214
  - in loop antennas, small receiver-type, 302
  - in microwave waveguides and antennas, 400–401
  - in shortwave reception antennas, 272, **273**, 274
- interferometer antennas, 431, **431, 432**
- intermediate frequency (IF), 55
- International Telecommunications Union (ITU), 2
- inverse square law ($1/D^2$), 9–10, **11**, 369–370
- inverted bobtail curtain (Thorne array) antenna, 267–269, **268, 269**
- inverted-L antenna, 510, **510**
- inverted-vee antenna, rhombic, 169, **169**
- inverted-vee dipole, 152–153, **153**
  - emergency operations antennas, 490–491, **491**
  - multiband and tunable-wire antennas, 205, **206**
- ionization, 5, 14–16, **15**, 18, 37, 41, 46, 47, 51, 52, 54
- ionograms, 43
- ionosphere, 14–16, **15**, 18, 37, 40–43, **42**, 176
- ionospheric disturbances, 45–47
- ionospheric propagation, 17–18, **18**, 37–43, 176
- ionospheric scatter propagation, 51–54, **53**
- ionospheric storms, 47
- ionospheric variation, 45–47
- iosthermal region of atmosphere, 14, **15**
- isophase plane, 32
- isopotential line, loop antennas, small receiver-type, 303
- isoresistance circles, using Smith charts, 97–100, **99**
- isotropic antenna waveguide, 394–403
  - backlobes in, 400, **400, 401**
  - beamwidth in, 400, **400, 401**

isotropic antenna waveguide (*Cont.*):
  dipole antenna elements and, 396–398, **397, 399**
  directivity in, 397, 398–403, **400**
  gain in, 398–403
  gain vs. aperture in, 402–403
  impedance in, 396
  interference and, 400–401
  near field and far field calculations in, 395–396
  ohmic loss in, 396
  radiation patterns and, 397, **398**
  radiation resistance in, 396
  sidelobes in, 400, **400, 401**
isotropic radiation:
  in high–frequency dipoles and doublets, 146
  in microwave waveguides and antennas, 394
  in shortwave reception antennas, 272
  in vertically polarized HF antennas, 179–181, **181**
isotropic sources, 10–14
ISWR (*See* current standing wave ratio)

J

J-pole antennas, 201, 353–355, **355**
Jansky, Carl, 421
JFET amplifier for loop antennas, small receiver-type, 313–316, **316**
Johnson noise, 417
joints and bends in microwave waveguides and antennas, 388–390, **389, 390**
Jupiter reception antenna, 421–422, **423**

K

K factor, 23–27
kilohertz as unit of frequency, 7
Ku bands, 69

L

L-section coupler, 214, **215**, 238–239
L-section network for impedance matching, 458–460, **459**
lag, voltage, 126
λ/2 large loop antennas, 287–289, **288, 289**
λ/4 inverted-L antenna, 510, **510**
Law of Reciprocity, 226, 271, 394
LC tank circuit, 526, **526**
leakage resistance, 140
left-hand rule, 126, **127**
length of antenna elements:
  in coaxial cable, 438
  in dipoles, 141, 142–143, 149–151, 160, 232, 255–256
  in directional beam antennas, 255–257, 259–260
  in directional phased vertical antennas, variable lengths in, 248–249, **249**
  in high–frequency dipoles and doublets, 141, 142–143, 149–151, 160, 232, 255–256
  in longwires, 213–215, 218–219, **218**, 221, 223
  in low-frequency operation antennas, 504
  in microwave waveguides and antennas, 392
  in multiband and tunable-wire antennas, 204
  in vertical antennas, 504
  in vertically polarized HF antennas, 181, 183, **184**
  in VHF/UHF transmitting/receiving antennas, 339
  in Yagi antennas, 259–260
Lewallen, Roy, 329, 331, 334
licensing, 3
light, speed, 8, 9, 13–14, 39, 525
limited-space antennas, 231–244
line attenuation constant, 82
line flattener tuner for marine radio antennas, 500, **500**
line-of-sight transmission, 19, **20**, 23–24, **23**, 35, 339
linear loading in low-frequency operation antennas, **504–505**
linearly loaded tee antenna, 511, **512**
link-tuned dipole, multiband and tunable-wire antennas, 205
loaded tower designs, 507
loading, **503**, 504–505, **506**
  automatic load control (ALC), 515
  in low-frequency operation antennas, **503**, 504–505, **506**
loading (*Cont.*):
  in low-frequency operation antennas, loaded tower designs for, 507
loading coils, dipoles and doublets in, 157–160, **157**
long path transmission, 51
long wave band, 8
longwires, 213–229
  angle of radiation in, 214
  antenna tuning unit (ATU) in, 172, 214
  apex angle in, 223
  baluns in, 171, **171**
  Beverage or “wave” antenna in, 224–228, **226, 228, 229**
  capacitors for, 215
  characteristic impedance in, 223, 226
  counterpoise, 170–172, **171**
  double vee antennas, 223
  electrical vs. physical length of, 213–214, 217, 218–219, 221
  electrostatic fields and electric charge in, 219
  end-fed longwire, 217–219, **217**
  feedpoint in, 217–218
  flagpole installation of, 234–236, **235, 237**
  gain in, 214, 218, **218**, 223
  ground plane verticals in, 172
  grounding of, 215–216, **216**
  as hidden and limited-space antennas, 232
  impedance and impedance matching in, 214, 223, 226
  inductors for, 214
  interference and, 214
  L-section coupler for, 214, **215**
  length of, 213–215, 218–219, **218**, 221, 223
  as marine radio antennas, 497, **498**
  maximum effective length (MEL) in, 227
  as multiband and tunable-wire antennas, 209, **210**
  non-resonant single-wire longwire antennas in, 220
  nonresonant vee beams, 223, **224**
  quarter-wavelength, 214, 217
  radials for, 215

longwires (*Cont.*):
  radiation patterns in, 219, **220**, 220–221, **222**
  random-length type, 213–214, **214**
  resonant vs. non-resonant types in, 213
  rhombic beam, 221, 223–224, **225**
  safety issues in, 215
  as shortwave reception antennas, 277–278, **278**
  as standing wave antennas, 220
  steerable notch Beverage antennas and, 229, **229**
  terminated antennas as, 220, **221**
  tilt angle in, 223
  transmission line for, 217–218
  traveling wave antennas in, 220, 223
  true longwire antenna, 217–219, **217**
  tuning of, 214, 215, **215**
  vee beam, 221–223, **222**
  VHF/UHF transmitting/receiving antennas, 340–341
  voltage standing wave ratio (VSWR) in, 215
  wavelength-measurement in, 217, 218–219
  window-sill antennas, 238, **241**
  wire used in, 214
  Zepp antenna in, 164–165, **165**
"looking in" impedance, 82–86, 372
loop antennas, 129, 131
loop antennas, large design, 287–297, 299
  (*See also* loop antennas, small receiver-type; loop antennas, small transmitter-type)
  antenna tuning unit (ATU) in, 287
  bisquare loop large antenna, 296–297, **297**
  current in, 289
  delta loop (1λ) large antenna, 290–293, **291**
  delta loop large antenna, 293, **295**
  demiquad loop antenna, 293, **294**
  diamond loop (1λ) large antenna, 290–293, **291**
  feedpoints in, 287–288, 292
  gain in, 288–289, **289**
  half-delta sloper (HDS), 295–296, **296**
loop antennas, large design (*Cont.*):
  impedance and impedance matching in, 292–293, **292**
  impedance in, 287–288
  inductive reactance in, 288–289, **289**
  λ/2 large loop antennas in, 287–289, **288, 289**
  loss in, 288–289, **289**
  nulls in, 299–300, **300, 301**
  (1λ) large loop antennas, 290–293, **290, 291**
  orthogonal lobes in, 299–300, **300, 301**
  radiation patterns in, 287, 299–300, **300, 301**
  square loop (1λ) large antenna, 290–293, **290**
loop antennas, small receiver-type, 299–318
  (*See also* loop antennas, large design; loop antennas, small transmitter-type)
  air core frame loops ("box" loops) in, 303–305, **304**
  amateur radio, 302
  amplifier for, 313–316, **316**
  angle of arrival of signal in, voltage vs., 303, **305**
  Armstrong oscillator circuit for, 313, **315**
  capacitance and capacitors in, 304, 307–309, 317
  capacitance tuning of, 307–309
  coaxial cable loop antennas, 318, **318**
  cochannel interference, 302
  coupling loops in, 306–307, **306**
  current in, 299
  depth wound loops in, 303
  DXing using, 311–313
  effective height of, 304
  Faraday connection and Faraday shield in, 318
  Grover's equation in, 302–303, 303*t*
  height of, 304
  inductance in, 307
  isopotential line in, 303
  low frequency (LF), 302
  nulls in, 299–300, **300, 301**, 303, **305**
  parallel tuning of, 307–309, **308**
  planar windings in, 303
loop antennas, small receiver-type (*Cont.*):
  preamplifiers for, 304
  Q–multipliers for, 313, **315**
  radiation patterns in, 299–300, **300, 301**, 303, 313
  radio direction finding (RDF) using, 301
  remote tuning in, 307–309, **310**
  self-resonant point in, 317, **317**
  series resonant tuning in, 307–309, **308**
  shape constants in, 303, 303*t*
  sharpening of, 313
  shielded loop antennas as, 309–311, **311, 312**
  split frequencies in, 302
  spoiler loop for, 313, **314**
  sports fan's loop as, 309
  transformer loops in, 306–307, **306**
  triangulation of radio signal using, 302
  tuning of, 304, 307–309, **308**, 312–313
  UHF, 302
  use of, 311–313
  varactor tuning in, 309, **310**
  VHF, 302
  VLF/LF loops, special problems in, 317
  voltage in, 303, 304
loop antennas, small transmitter-type, 319–325
  (*See also* loop antennas, large designs ; loop antennas, small receiver-type)
  bandwidth of, 320–321
  capacitance and capacitors in, 320–321
  capacitor voltage in, 321
  circular loop transmitter antenna as, 323, **323**
  current in, 320, 323
  design equations for, 321–322
  distributed capacitance in, 321
  efficiency in, 321
  gain in, 319–320
  inductance in, 321
  inductive reactance in, 319, 321
  loss resistance in, 321
  octagonal loop transmitter antenna in, 319, **320**, 322, **322**, 323, **324**

loop antennas, small transmitter-type (*Cont.*):
  Patterson octagonal loop transmitter antenna as, 322, **322**
  quality factor (Q) in, 320–321
  radiation patterns in, 324, **325**
  radiation resistance in, 319, 321
  resistance in, 319, 321
  square loop transmitter antenna in, 319, **320**
  tuning capacitance in, 320–321
  voltage in, 320, 323
  voltage standing wave ratio (VSWR) in, 320
loss, 11–12, 19, 27, 140
  in coaxial cable, 438
  in high–frequency dipoles and doublets, 143–146
  in λ/2 large loop antennas, 288–289, **289**
  loss coefficient scale in, 109
  in microwave waveguides and antennas, 396
  in multiband and tunable-wire antennas, 210
  in Smith charts, 108–109
  in transmission lines, 83–84, 91–94, 369
  in vertically polarized HF antennas, 183
loss coefficient scale, 109, 112–113
loss resistance, in loop antennas, small transmitter-type, 321
lossy circuits, Smith charts, 117, **118**
low frequency (LF), 421
  loop antennas, small receiver-type, 302
  low-frequency operation antennas, 501–513
  VHF/UHF transmitting/receiving antennas vs., 339
low-frequency operation antennas, 501–513
  continuous loading in, **503**, 504–505, **506**
  delta feed systems in, 507, **509**
  discrete loading in, **503**, 504–505
  grounding and ground systems for, 501–502, **502**
  "hairpin" loading scheme in, 505, **506**
  impedance matching in, 505
low-frequency operation antennas (*Cont.*):
  impedance matching in, of towers, 507, **508**
  inductance loading of vertical antennas in, **503**
  inductive loading in, **503**, 504–505
  inverted-L antenna for, 510, **510**
  λ/4 inverted-L antenna for, 510, **510**
  length of vertical antenna in, 504
  linear loading in, **503**, 504–505
  linearly loaded tee antenna for, 511, **512**
  loaded tower designs for, 507
  loading in, **503**, 504–505, **506**
  radial layout for, 502, **502**
  random-length Marconi antenna for, 507–510, **509**
  shortened horizontal antennas for, 505
  shortened tee antenna, 511, **513**
  shortened vertical antennas for, 502–505, **503**
  shunt feed systems in, 507
lower sideband (LSB), 55
lowest usable frequency (LUF), 45

M

magnetic fields, 123, 126, 127, **128**,
  in microwave waveguides and antennas, 375–376, **375**, 383, 395–396
magnetic flux, 126, 127
man-made noise (QRM), 245
Marconi, Guglielmo, 1
marine radio antennas, 1–2, 494–497, **497**
  Citizens Band radios and, 496
  grounding and ground systems for, 495–496, **495, 496**
  HF SSB mode in, 495
  line flattener tuner for, 500, **500**
  longwire antennas, 497, **498**
  tuners, 497–500, **499, 500**
  VHF-FM use in, 495
  whip antennas for, 496–497, **497**
mast-mounted vertical antenna, 546, **546**, 560–561, **561, 562**
masts and supports, 556–559, **557, 558**
matching section, quarter-wavelength coax, 292
matching stub systems, 465–466, **466, 467**
maximum effective length (MEL), 227
maximum usable frequency (MUF), 43–45, 47
Maxwell's equations, 8–9
measurement and adjustment techniques, 433–438, 515–542
  antenna adjustment in, 523–524
  automatic load control (ALC) in, 515
  basic radio system in, 515, **516**
  bridge to measure RF power in, 532, **532**
  capacitance measurement in, 526–527
  current transformer wattmeter in, 534–538, **535, 536, 537**
  dip oscillator/dip meter in, 527–529, **527, 528**
  dummy loads in, 540–542, **540, 541, 542**
  filters in, 515
  forward power in, 538, **539**
  grid dip meter in, 527–529, **528**
  harmonics in, 515
  high-power RF wattmeters for, 538, **538**
  impedance bridges in, 519–521, **520, 521**
  impedance in, 525
  inductance measurement in, 526–527
  LC tank circuit in, 526, **526**
  matching unit in, 515
  micromatch wattmeter in, 533, **533**
  monomatch wattmeter in, 533–534, **534**
  nonresonant antenna adjustment in, 524–525
  parallel resonant circuits in, 526
  power in, 515, 531–532
  power-frequency calibration in, 538, **538**
  reactance testing/adjustment in, 524
  reflected power in, 538, **539**
  resistance testing/adjustment in, 523–524
  resonance testing/adjustment in, 524

measurement and adjustment techniques (*Cont.*):
  RF noise bridge in, 521–522, **521**
  RF power measurements in, 529–539
  RF wattmeters for, 529–539, **531**
  series tuned circuits in, 526
  surge impedance (Z_o) in, 516
  thermocouple for RF ammeter in, 530, **530**
  transmission lines in, 516–517, 525
  tuned circuit measurements in, 526
  tuning antennas in, 524
  voltage measurement in, 531
  voltage standing wave ratio (VSWR) in, 515, 517–519, 529, 532
  Wheatstone bridge in, 519–521, **520**
medium wave (MW), 8, 17, 29, 48, 61, 224–228, **226, 228, 229**
meteor scatter propagation, 54
micromatch wattmeter, 533, **533**
microwave transmissions, 3, 5, 8, 13, 17, 18, 19, 22, 23, 29, 30, 32, 37, 40, 61, 69, 369
microwave waveguides and antennas , 369–415, 421
  (*See also* parabolic "dish" antennas)
  adjustable plunger or short-circuit termination in, 388, **388**
  antenna design using, 394
  aperture (slot) coupling in, 390–393, **393**
  aperture in, 402–403
  array antennas in, 409–411
  attenuation in, 377
  backlobes in, 400, **400, 401**
  beamwidth in, 400, **400, 401**
  boundary conditions in, 375–376
  branch feed systems for, 409, **412**
  capacitive (probe) coupling in, 390–393, **392**
  cavity antennas, 404, **405**
  characteristic impedance in, 383–384
  choke joints in, 389, **390**
  coordinate system to denote dimensions in, 376–377, **377**
  corner reflector antennas, 404, **405, 406**
microwave waveguides and antennas (*Cont.*):
  corporate feed systems for, 409, **412**
  coupling methods for, 390–393
  cutoff frequency (F_c) in, 381–383
  dielectric loss, 369
  dipole antenna elements and, 396–398, **397, 399**
  directivity gain (G_d) in, 401–402
  directivity in, 395, 397, 398–403, **400**
  dominant mode in, 376–377
  dummy loads in, 385–388, **386**
  effective aperture (A_e) in, 394
  effective radiated power (ERP) in, 401
  electric fields in, 383, 395–396
  end plate in, 385, 388, **388**
  flatplate array antennas, 411, 413, **414, 415**
  free-space velocity (C) in, 377, 380
  frequency in, 373, **374**, 375, **376**, 380, 383–384, **384**
  gain in, 394, 395, 398–403
  gain vs. aperture in, 402–403
  graphite-sand dummy load in, 385
  group velocity (V_g) in, 377, 378–379
  horn antenna radiators in, 403–404, **403**
  I^2R losses in, 385
  impedance and impedance matching in, 383–384, 394, 396
  inductive (loop) coupling in, 390–393, **393**
  insulators in, 371–372
  interference and, 400–401
  inverse square law ($1/D^2$), 369–370
  isotropic antenna waveguide, 394–403
  isotropic radiation in, 394
  joints and bends in, 388–390, **389, 390**
  law of reciprocity and, 394
  length in, 392
  losses in, 396
  magnetic fields in, 375–376, **375**, 383, 395–396
  materials used in, 370
  near field and far field calculations in, 395–396
microwave waveguides and antennas (*Cont.*):
  ohmic loss in, 369, 371, 396
  parabolic "dish" antennas as, 404–409, **407, 408, 409**
  phase shifting in, 409, **412, 413**
  phase velocity (V_p) in, 377, 379, 380
  pipe analogy of, 369–370, **370**
  power gain (G_p) in, 401–402
  propagation constant (B) in, 384
  propagation modes in, 373–381
  quarter-wavelength shorted stub vs., 372–373, **372, 373, 374**
  radiating vs. nonradiating slots in, 393
  radiation loss in, 369, 372
  radiation patterns and, 397, **398**
  radiation resistance in, 396
  rectangular waveguide using parallel transmission lines and, 370–373, **371**
  reflector antennas, 404, **405, 406**
  resistive vane dummy load in, 385
  rotating joints in, 390, **391**
  safety issues, 413–415
  semipermanent joints in, 389
  shielding in, 392
  sidelobes in, 400, **400, 401**
  skin effect in, 369, 371
  slot array antennas in, 411–413, **413**
  solid-state array antennas in, 408–411
  terminations for, 385–388
  transducer action of, 394
  transverse electric mode (TE mode) in, 375
  transverse electromagnetic (TEM) fields in, 373
  transverse magnetic mode (TM mode) in, 375
  velocities in, 377–381, **378**
  voltage standing wave ratio (VSWR) in, 388
  wavelength in, 377–381, **378**
  wedge dummy loads in, 385, **386, 387**
military radio, multiband and tunable-wire antennas, 203
MiniNEC modeling software, 331
miniNEC-3/miniNEC-4 for modeling software, 328–329

- "mirroring" of radio waves, 43, 54
- mismatch (VSWR) losses, transmission lines, 91–94
- mixing, 55
- mobile HF antennas, 479–485
  - coil-loaded antenna as, 480–481, **480, 351**
  - field strength meters (FSM) for, 483–485, **484, 485**, 485*t*
  - helical whip antenna as, 481, **482**
  - multiband top-loaded antenna as, 481, **482**
  - Q factor in, 483
  - radial-to-mobile-antenna connection detail in, 488, **489**
  - resonance in, 483
  - top-loaded antenna as, 481, **482**
  - tuning of, 483–485
  - voltage standing wave ratio (VSWR) in, 483
  - used as windowsill-mounted, 239, **241**
- modeling software for antennas, 327–338
  - coordinate system used in, 328, **329, 330**
  - Expert MiniNEC, 331
  - EZNEC for Windows, 331, 334–337, **335, 336, 337**
  - EZNEC Professional, 329
  - GNEC, 329–330
  - MiniNEC, 331
  - miniNEC-3/miniNEC-4, 328–329
  - NEC-2, 330
  - NEC4Win95, 331
  - NECWin Basic, 331–334, **332**
  - NECWin Plus, 330–331
  - Numerical Electromagnetic Computation (NEC) programs in, 328
  - radiation patterns displayed in EZNEC for Windows, 336–337, **337**
  - radiation patterns displayed in NECWin Basic, 332, **333, 334**
  - voltage standing wave ratio (VSWR) modeling in, 333, **335**
  - voltage standing wave ratio (VSWR) modeling in, 337, **338**
- monomatch wattmeter, 533–534, **534**
- monopulse feeds, parabolic "dish" antennas, 409, **410**
- monthly variations in radio propagation, 46
- Morse code, 1, 2
- mounting (*See also* construction and installation techniques)
  - Beverage antenna, 228
  - dipole, 233, 256
  - directional beam antenna, 256
  - emergency operations antenna, on vehicles, 491–492
  - fence posts as supports for vertical antenna, 561–565, **563, 564**
  - floor flanges in, 256
  - hidden and limited-space antenna, 233
  - high–frequency dipole and doublet, 141
  - mast-mounted vertical antenna, **546**
  - pulley system for, 557–558, **558**
  - right-angle flange hardware in, 256
  - shortwave reception antenna, 277–278, **278**
  - stirrup-end hardware in, 558, **559**
  - temporary mast mount for emergency antenna, 492, **493**
  - U-bolt hardware, 558, **559**
  - vehicle-mounted emergency antenna, 491–492
  - vertically polarized HF antenna, 186, 189–190, 194–195, **194**, 195–197, **196**, 560–565, **561**
  - VHF/UHF transmitting/receiving antenna, 347
- multi-impedance balun, 476, **477**
- multiband and tunable-wire antennas, 203–212
  - antenna tuning unit (ATU) for, 210
  - balanced transmission line and, 208
  - balancing line, 208–209, **209**
  - baluns for, 208–209, **209**, 210
  - coaxial cable in, 208
  - dipoles in, 204–205, **205**
  - end-fed Zepp antenna in, 208, **208**
  - feeding parallel transmission line for, 208–212
  - feedpoint in, 204, 210
  - G5RV multiband dipole antenna in, 206–208, **207**
- multiband and tunable-wire antennas (*Cont.*):
  - harmonics in, 210
  - impedance in, 210
  - insulators (spreaders) in, 206
  - inverted-vee dipole for, 205, **206**
  - length of, 204
  - link-tuned dipole for, 205
  - longwires for, 209, **210**
  - losses in, 210
  - quad beam antennas for, 203
  - random-length longwire antennas in, 209, **210**
  - transformers for, 210, **211**
  - transmatches in, 208–209, **209**
  - trap dipoles in, 204–206, **204, 205**
  - tunability of, 203
  - tuned feeder antennas in, 206, **207**
  - tuning of, 205
  - voltage standing wave ratio (VSWR) and, 203
  - Windom antenna as, 209–210, **211**
  - Yagi antennas as, 203
- multiband fan dipole, 169–170, **170**
- multiband top-loaded mobile antenna, 481, **482**
- multifilar windings in transformers, 472
- multihop communication, 23, **49**, 49–50
- multihop interference, 36, **36**
- multiple helical antennas, 428–431, **429, 430**
- multiple radials from vehicle-mounted antenna, 488, **489**

N

- NAA radio station, 2
- National Bureau of Standards (*See* National Institute of Standards and Technology)
- national identifiers in call signs, 3
- National Institute for Standards and Technology (NIST), 40, 47
- Naval Research Laboratory, 2
- near field and far field calculations in microwave waveguides and antennas, 395–396
- NEC-2 modeling software, 330
- NEC4Win95 modeling software, 331
- NECWin Basic modeling software, 331–334, **332**

- NECWin Plus modeling software, 330–331
- Nittany-Scientific, 329, 330, 331
- nodes, 87, 129
- noise, 37, 68–69
  - antenna noise temperature as, 417–420
  - contributors to antenna noise temperature and, 419–420, **419**
  - environmental/natural noise (QRN) in, 245
  - figure of merit and, 418
  - impedance and, 417–418
  - Johnson noise in, 417
  - man-made noise (QRM) in, 245
  - noise bridge in, 521–522, **521, 522, 523**
  - noise factor (F_n) in, 418
  - noise figure (NF) in, 418
  - noise temperature (T_e) in, 418–420
  - omnidirectionality vs., 245
  - piezoelectric, 69
  - RF noise bridge in, 521–522, **521**
  - signal-to-noise ratio (SNR) in, 246
  - sky noise sources, 419–420
  - thermal agitation noise, 417
  - thermal noise in, 417–420
  - in transmission lines, 68–69
  - in Yagi antennas, 261
- noise bridge, 521–522, **521, 522, 523**
- noise factor (F_n), 418
- noise figure (NF), 418
- noise temperature (T_e), 418–420
- nonreciprocal direction propagation, 54
- nonresonant antenna adjustment, 524–525
- nonresonant loop antennas, 234–235, **234**
- nonresonant single-wire longwire antennas, 220–220
- nonresonant vee beam longwire antenna, 223, **224**
- normalization of impedance, 95–97, **98**, 469–470
- northern lights, 54
- null steering, 229
- nulls, in loop antennas, 299–300, **300, 301**, 303, **305**
- Numerical Electromagnetic Computation (NEC) programs, 328

O

- 1λ large loop antennas, 290–293, **290, 291**
- octagonal loop transmitter antenna, 319, **320**, 322, **322**, 323, **324**
- off-center-fed full-wave doublet (OCFD), 163, **163**
- off-center-fed nonresonant sloper (OCFS), 164, **164**
- ohmic loss, 143–146, 369, 371, 396
- ohmic resistance, 19, 27, 140
- ohmmeter, coaxial cable test, 437–438, **437**
- Ohm's law, transmission lines, 74
- Ω-fed grounded vertical antenna, 200–201, **202**
- ominidirectionality:
  - in high–frequency dipoles and doublets, 146
  - noise vs., 245
  - in shortwave reception antennas, 272, **273**
  - in vertically polarized HF antennas, 173, 179–181
- open line, 59
- Orion, 331
- orthogonal lobes, in loop antennas, 299–300, **300, 301**
- OSCAR satellite communications, 3, 339
- out–of–phase waves, **22**

P

- parabolic "dish" antennas, 404–409, **407, 408, 409**, 421
  - beamwidth in, 407
  - Cassegrain feed in, 408–409, **409**
  - cutoff frequency in, 406
  - directrix in, 404
  - effective aperture in, 407
  - feed geometries for, 408–409, **408, 409**
  - feed illumination in, 406
  - gain in, 405–408
  - monopulse feeds, 409, **410**
  - ray tracing showing operation of, **407**
  - splash plates in, 408
  - x-y coordinates in, 404
- parallel lines, 59–63, **60**, 65–66
- parallel resonant circuits, 526
- parallel tuning in loop antennas, 307–309, **308**
- parasitic elements in Yagi antennas, 258–259
- paths, 17–18, **18**
- Patterson octagonal loop transmitter antenna, 322, **322**
- pedestals for towers, 568–570, **569**
- permits, zoning, and regulations, 547–548
- phase, 126, **128**
- phase constant, 82
- phase control circuit (PCC), 229
- phase reversal stubs, collinear "Franklin" array antenna in, 165–166, **166**
- phase shift, 19, 21–23, 129, 409, **412, 413**, 450–451, **451**
- phase velocity (V_p) in microwave waveguides and antennas, 377, 379, 380
- phase-shifting circuit, directional phased vertical antennas, 252, **252**
- phase-switching antenna matcher, shortwave reception antennas, 284, **285**
- phasing box, shortwave reception antennas, 284, **286**
- phasing switch, directional phased vertical antennas, 249–250, **252**
- phasing transformer, 249–250, **250, 251**
- phasing, phased radials, directional phased vertical antennas, 246–250
- pi networks, 461, **461**
- "picket fencing," 23
- piezoelectricity, 69
- planar windings in loop antennas, 303
- planetary terminator, 51, **53**
- plunger or short-circuit termination, microwave waveguides and antennas, 388, **388**
- polarity, 126-127, 131, 134
  - reversal of, 126–127, 131, 134
- polarization of radio waves, 12–13, **13**, 27, **28**, 29, 55, 173
  - in Beverage antenna, 226–227
  - in directional phased vertical antennas (*See* directional phased vertical antennas)

- polarization of radio waves (*Cont.*):
  - in vertically polarized HF antennas, 173–202
- polarization diversity, 55, **58**
- power, 11, **21**, 140, 515, 531–532
  - in RF power measurements, 529–539
  - in vertically polarized HF antennas, 180
- power density, 10
- power fading, 36
- power gain (G_p) in microwave waveguides and antennas, 401–402
- power-frequency calibration in, 538, **538**
- preamplifiers for loop antennas, 304
- predictions of propagation, 55, **56**
- propagation constant (B), 82, 384
- propagation modes in microwave waveguides and antennas, 373–381
- propagation of radio waves, 5–58
- propagation predictions, 55, **56**
- pulley system for mounting, 557–558, **558**
- pure reactance circle, Smith charts, 100, **103**
- pure resistance line, Smith charts, 95–97, **98**

Q

- Q–multipliers for loop antennas, small receiver-type, 313, **315**
- QRM noise, 245
- QRN noise, 245
- quad beam VHF/UHF antenna, 356, 358, **359**
- quad beam antennas, 203
- quad loop antennas, 263–267, **264, 266, 267**
- quality factor (Q), 85, 160, 320–321, 483
- quarter-wave matching (Q section) sections, 85, 146, 199–200, **199**, 467–468, **468, 469**
- quarter-wave vertical polarized HF antenna, 173
- quarter-wavelength longwires, 214, 217
- quarter-wavelength lines, 84–85
- quarter-wavelength shorted stub
  - "looking in" impedance in, 372
  - microwave waveguides and antennas vs., 372–373, **372, 373, 374**
  - transmission lines and, 86, **87, 88**
- quarter-wavelength vertical antenna, 181, **182**, 183, 187

R

- radar, 3, 30, 52
- radial-to-mobile-antenna connection detail, 488, **489**
- radially scaled parameters, Smith charts, 106–109, **107**
- radials:
  - in grounding and ground systems, 577–579, **579, 580**
  - in impedance matching, 460
  - in longwires, 215
  - in low-frequency operation antennas, 502, **502**
  - multiple, from vehicle-mounted antenna for, 488, **489**
  - radial-to-mobile-antenna connection detail in, 488, **489**
  - in vertically polarized HF antennas, 176, 186, **189**, 560–561
- radiated fields, 5
- radiating vs. nonradiating slots, microwave waveguides and antennas, 393
- radiation angle, 30, 48–50, **50**
- radiation loss, 369, 372
- radiation patterns:
  - in Adcock antenna, **449**
  - in beam antennas, 257–258, **258**
  - cardioid, 441, **442**
  - in directional phased vertical antennas, 246–250, **246, 247**
  - EZNEC for Windows modeling of, 336–337, **337**
  - in high–frequency dipoles and doublets, 146–152, **147**
  - in isotropic antenna waveguide, 397, **398**
  - in λ/2 large loop antennas, 287
  - in longwires, 219, **220**, 220–221, **222**
  - in loop antennas, large designs, 299–300, **300, 301**
  - in loop antennas, small receiver-type, 299–300, **300, 301**, 303, 313
- radiation patterns (*Cont.*):
  - in loop antennas, small transmitter-type, 324, **325**
  - in loopstick antennas, 439, **440, 441**
  - NECWin Basic modeling of, 332, **333, 334**
  - in sense antennas, 441, **442**
  - in switched-pattern RDF antenna, 454–456, **454**
  - in vertically polarized HF antennas, 178–181, **181, 182**
  - in Watson-Watt Adcock array antennas, **450**
  - in Yagi antennas, 257–258, **258**
- radiation resistance, 140
  - in high–frequency dipoles and doublets, 143–146
  - in loop antennas, small transmitter-type, 319, 321
  - in microwave waveguides and antennas, 396
  - in vertically polarized HF antennas, 183, **184, 185**
- radiators, high–frequency dipoles and doublets, 141
- radio astronomy antennas, 421–432
  - directional discontinuity ring radiator (DDRR) antenna in, 424–427, **426**, 427*t*
  - helical antennas in, 427–428, **429**
  - interferometer antennas in, 431, **431, 432**
  - Jupiter reception antenna, 421–422, **423**
  - multiple helical antennas in, 428–431, **429, 430**
  - ring antenna, 422–424, **423, 424, 425**
  - summation interferometer array antenna, 431, **431**
- radio direction finding (RDF), 439–456
  - Adcock antennas for, 446–447, **448, 449**
  - AM broadcast band (BCB) in, 439
  - birdirectionality of loopsticks in, 441
  - cardioid radiation pattern of sense antenna in, 441, **442**
  - CONALRAD station frequencies and, 439
  - Doppler antennas, 450–451, **451**

radio direction finding (RDF) (*Cont.*):
  double ducky direction finders (DDDF) in, 455–456, **455**
  “fox hunting” antenna in, 444, **445**
  Global Positioning System (GPS) and, 443
  improvisation of, using portable AM radio receiver, 443, **443**
  loop antennas in, small receiver-type, 301
  loopstick antennas in, 439–440, **440**
  measurements in, 433
  radiation patterns of loopsticks in, 439, **440, 441**
  regular wire loop antennas for, 444, **445**
  “rubber ducky” VHF antennas in, 454–456, **455**
  S meters in, 439
  sense antenna for, 441, **442**, 446, **447**
  shortwave and AM BCB “skip” in, 444–446
  switched-pattern RDF antenna in, 454–456, **454**
  time difference of arrival (TDOA) array antenna in, 452–454, **453**
  triangulation of radio signal in, 439
  Watson-Watt Adcock array antennas in, 447–450, **449, 450**
  wide-aperture direction finder (WADF) in, 451
  Wullenweber array antennas in, 451, **452**
Radio Luxembourg effect, 54
radio system, 515, **516**
radio vectoring, 3
radio wave configuration, 5–6
random-length longwires, 209, **210**, 213–214, **214**, 507–510, **509**
ray tracing, 32–33, **407**
reactance, 140, 524
  (*See also* capacitive reactance; inductive reactance)
  in high–frequency dipoles and doublets, 143
  Smith charts, 109, 117
  Smith charts, capacitive reactance in, 100, **102**
reactance (*Cont.*):
  Smith charts, constant reactance circles in, 100–105
  Smith charts, inductive reactance in, 100, **101**
  Smith charts, pure reactance circle in, 100, **103**
  in transmission lines, 85–86, 109, 525
  in vertically polarized HF antennas, 183, **184**
Reber, Grote, 421
receivers, 123
reciprocity, law, 226, 271, 394
recombination phenomena, 18
rectangular waveguide using parallel transmission lines, 370–373, **371**
reflected power (P_r), 89, 538, **539**
reflected waves, 128, 134
reflection, 16–17, 21, 22, 50, 52, 161
reflection coefficient, 21, 22
  Smith charts, **103, 106**, 106–109, 112, 114
  transmission lines, 71–79, 89, 112, 114
reflection loss, 113–114
reflector antennas, 404, **405, 406**
reflectors, Yagi antennas, 258–259
refraction, 16–17, 16, 23–24, **24**, 29–30, **30**, 48
refractivity of atmosphere, 24–27
regular wire loop antennas, 444, **445**
remote tuning, loop antennas, small receiver-type, 307–309, **310**
repeaters for emergency operations antennas, 487
resistance, 27, 140, 369, 434–437, 523–524
  in high–frequency dipoles and doublets, 143–146, **145**
  in loop antennas, small transmitter-type, 319, 321
  in Smith charts, constant or isoresistance circles in, 97–100, **99**
  in Smith charts, pure resistance line in, 95–97, **98**
  in Smith charts, unity resistance circle in, 100
  in transmission lines, 64, 525
resistance (*Cont.*):
  in vertically polarized HF antennas, 183
resistive vane dummy load in microwave waveguides and antennas, 385
resonance, 140, 524
  in dipoles, 237
  in high–frequency dipoles and doublets, 143, 149–151, **150**
  in mobile HF antennas, 483
resonant frequency, 433
resonant vs. non-resonant longwires, 213
return loss, 108–109, 113, 114
RF ammeter, 530, **530**
RF noise bridge, 521–522, **521**
RF power measurements, 529–539
RF wattmeters, 529–539, **531**
rhombic beam longwire antenna, 221, 223–224, **225**
rhombic inverted-vee antenna, 169, **169**
right-angle flange mounting hardware, 256
ring antenna, 422–424, **423, 424, 425**
RLC networks, 59, 143
rotatable dipole antennas, 239, **242**, 255–257, **257**
rotating joints in microwave waveguides and antennas, 390, **391**
“rubber ducky” VHF antennas, 454–456, **455**

S

S meters in RDF antennas, 439
safety issues, 543–548
  burns or shocks from RF through vertical antennas, 191, 195
  hidden and limited-space antennas, 233
  longwires, 215
  microwave waveguides and antennas, 413–415
  shortwave reception antennas, 276, **277**, 279
  VHF/UHF transmitting/receiving antennas, 340
“salt pipe” to apply chemical soil-conductivity improvers, 576–577, **577**

salt to alter soil conductivity, 575–576
satellite communications, 3, 339, 421
scanner band VHF/UHF antennas, 360–368
scanner skyhook VHF/UHF antennas, 362–368, **363**
scanner-vision antennas, 360–362, **361**
scatter propagation, 51–54, **53**
seasonal variations in radio propagation, 24, 40, 41, 46, 51, 178
selective fading, 55
self-resonant point, loop antennas, small receiver-type, 317, **317**
semipermanent joints in microwave waveguides and antennas, 389
sense antenna in RDF, 441, **442**, 446, **447**
series matching section, 468–471
series resonant tuning, loop antennas, small receiver-type, 307–309, **308**
series tuned circuits, 526
series-fed vertical antennas, 200
shape constants, loop antennas, small receiver-type, 303, 303*t*
shielded loop antennas, 309–311, **311, 312**
shielded parallel line, characteristic impedance of, 66
shielded twin lead, 61, **61**
shielding in microwave waveguides and antennas, 392
shocks hazards from RF through antennas, 191, 195, 215, 233, 276, 279
short path transmission, 51
short skip, 41
short-circuit termination, in microwave waveguides and antennas, 388, **388**
shorted lines, 87
shorted stubs, transmission lines, 86, **87**
shorted stubs (*See* stub matching systems)
shortened coil-loaded dipoles, 157–160, **158**
shortened tee antenna, 511, **513**
shortwave and shortwave reception antennas, 2–4, 8, 37, 41, 54–55, 271–286
shortwave and shortwave reception antennas (*Cont.*):
  angle of radiation in, 274, **274**
  antenna matching transformer in, 281, **285**
  antenna matching tuner in, 281, **285**
  connection to receiver in, 274–276, **275, 276**
  dipole as, 272, **273**, 274
  directional wire antenna in, 281–286, **284**
  directivity of, 272
  flattop antennas for, 279, **280**
  gain in, 272
  grounding for, 276, **277**, 278–279, **279**
  interference and, 272, **273**, 274
  isotropic radiation in, 272
  law of reciprocity in, 271
  longwire antennas for, 277–278, **278**
  mounting for, 277–278, **278**
  multiband wire vertical antennas for, 281, **282, 283**
  omnidirectionality in, 272, **273**
  phase-switching antenna matcher in, 284, **285**
  phasing box for, 284, **286**
  properties of importance in, 271–274
  radio direction finding (RDF) antennas for, 444–446
  safety issues in, 276, **277**, 279
  skip propagation and, 274
  transformers in, 281
  vertical antennas for, 279–281, **280, 282, 283**
shunt feed systems, 200, 507
side scatter, 52
sidelobes in microwave waveguides and antennas, 400, **400, 401**
signal-to-noise ratio (SNR), 37, 45, 246
simple refraction, 29
single-sideband (SSB), 55, 495
skin effect, 369, 371
skip distance, 47–48, **48**, 47
skip propagation, 5, 29, 37, 41, 47–48, **48, 49, 51**, 55
  radio direction finding (RDF) antennas and, 444–446
skip propagation (*Cont.*):
  shortwave reception antennas and, 274
  vertically polarized HF antennas and, 176–179, **178**
  VHF/UHF transmitting/receiving antennas and, 339
skip zone, 47–48, **48**
sky noise sources, 419–420
sky wave, 47–48, **48, 49**, 302
skyhook VHF/UHF antennas, 362–368, **363**
slip-up tower, 567–568, **568, 569**
sloping (“sloper” or “slipole”) dipole, 153–154, **154, 155**, 236, **238**
  half-delta sloper (HDS), 295–296, **296**
  off-center-fed nonresonant sloper (OCFS), 164, **164**
  vee-sloper antenna in, 168–169, **168**
slot array antennas, 411–413, **413**
Smith charts, 95–121, **96**
  absorbed power (P_a) in, 113–114
  admittance in, 100, 112, 114
  angle of reflection coefficient circle in, **103**, 106, 112
  angle of transmission coefficient circle in, **103**, 106
  applications for, 109–121
  capacitive reactance in, 100, **102**, 117
  complex admittance in, 105
  components of, 95–109
  constant reactance circles in, 100–105
  constant resistance circles in, 97–100, **99**
  frequency in, 117–121, **119, 120**
  impedance and impedance matching, 110–112, **111**, 115, 119–121, **119**
  impedance radius in, 100
  inductive reactance in, 100, **101**, 117
  isoresistance circles in, 97–100, **99**
  loss coefficient scale in, 109, 112–113
  lossy circuits and, 117, **118**
  normalized impedance line in, 95–97, **98**

Smith charts (*Cont.*):
  pure reactance circle in, 100, **103**
  pure resistance line in, 95–97, **98**
  radially scaled parameters in, 106–109, **107**
  reactance in, 109, 117
  reflection coefficient in, 106–109, 114
  reflection loss in, 113–114
  return loss in, 108–109, 113, 114
  standing wave ratio (SWR) in, 108
  stub matching systems problems solved using, 109, 114–117, **115, 116**
  susceptance in, 117
  transmission line problems solved using, 109–114
  transmission loss coefficient (TLC) in, 112, 114
  transmission loss in, 108–109
  unity resistance circle in, 100
  voltage standing wave ratio (VSWR) circle in, 100, **104**, 112, 114, 115, 117
  wavelength calculation in, 110, 115
  wavelength circles in, **103**, 105
Smith, Phillip H., 95
soil conductivity, 27, 28*t*, 575–577, 576*t*, **576**, 577
solar flares vs. radio propagation, 37, 39, 46–47
solar flux index (SFI), 40
solar storms, 47
solid-state array antennas, 408–411
SOS, 2
southern lights, 54
space diversity, high–frequency dipoles and doublets, 160
space quadrature fields, 127
space waves, 17–19, **18**
spark–gap transmitters, 2
spatial diversity, 55, **57**
SPC transmatch network, 463, **463**
"spider web" ground system, vertically polarized HF antennas, 185, **188**
splash plates, parabolic "dish" antennas, 408
split frequencies, loop antennas, small receiver-type, 302
split-capacitor networks, 461–462, **462**
spoiler loop, loop antennas, small receiver-type, 313, **314**
sporadic E propagation, 41, 46
sports fan's loop antennas, small receiver-type, 309
spreaders, multiband and tunable-wire antennas, 206
square loop (1λ) large antenna, 290–293, **290**
square loop transmitter antenna, 319, **320**
stacked dipoles, 160–163, **162**
stagger tuning in high–frequency dipoles and doublets, 157
standard refraction, 25
standing wave antennas, 220
standing wave ratio (SWR), 434
  in high–frequency dipoles and doublets, 144
  in Smith charts, 108
  in transmission lines, 86–91, **90**
standing waves, 87, 127–131, **130, 131**
  current, 134, **135**, 136, **137**, 139
  in high–frequency dipoles and doublets, 144
  measurement of, 137, 138
  in transmission lines, 71, **72, 73**
  voltage, 134–136, **135, 136**, 139
station grounding, 579–580, **581**
stationary waves, 128
steady-state ac response, transmission lines, 70, 79–82, **81**
steerable notch Beverage antennas, 229, **229**
step–function response, transmission lines, 70–71, **70, 71**
stirrup-end mounting hardware, 558, **559**
storms, ionospheric, 47
strain relief, 553–554, **554**
stratosphere, 14, **15**, 18
stray capacitance, 138
stripline or microstripline, 63, **64**, 66–67
stub matching systems, 114–117, **115**, 465–466, **466, 467**
  impedance matching in, 115
  length of stub calculation in, 117
  shorted stubs, 114
  Smith charts for, 109, 114–117, **115, 116**
stub matching systems (*Cont.*):
  in VHF/UHF transmitting/receiving antennas, 345–346, **345**
  voltage standing wave ratio (VSWR) in, 115, 117
  wavelength calculation in, 115
subrefraction, 31, **32**, 36
sudden ionospheric disturbances (SIDs), 46–47
summation interferometer array antenna, 431, **431**
sunspots vs. radio propagation, 40–41, **40**, 43, 46, 47
superrefraction, 29, **31**, 26
support posts, 565, **565**
support ropes, 551
suburban lot installation plan, **545**
surface waves, 17–19, **18**, 27–29
surge impedance (*See* characteristic impedance)
susceptance, Smith charts, 117
switched-pattern RDF antenna, 454–456, **454**
SWR analyzer, 434

T

360-degree directional array, 250, **253**, 254*t*
tapped balun, 476, **477**
tee antennas, 511, **512, 513**
tee-network transmatch, 463, **463**
Teflon, 61, 62
telegraph, 1, 2
telescoping antennas, emergency operations, 491
television, 3, 32
television interference/broadcast interference (TVI/BCI), 214
television-scanner VHF/UHF antennas, 360–362, **361**
temperature effect on propagation, 29–32, **32**
temporary antenna construction, emergency operations, 487–488, **488**
temporary mast mount for emergency antennas, 492, **493**
terminated longwire antennas, 220, **221**
terminations for microwave waveguides and antennas, 385–388
terrain effects, 28, 32, **33**
thermal agitation noise, 417

thermal noise, 68, 417–420
thermocouple for RF ammeter, 530, **530**
Thorne array antenna, 267–269, **268, 269**
three-dimensional (ideal) radiation pattern, dipoles and doublets, 147–148, **147**
tilt angle of surface wave, 28, 28*t*
tilt angle, rhombic beam longwire, 223
tilted, center-fed, terminated, folded dipole (TCFTFD or T^2FD, TTFD), 166–168, **167**
time difference of arrival (TDOA) array antenna, 452–454, **453**
time domain reflectometry (TDR), 75–79, **76, 77, 78, 79, 80, 81**
time quadrature fields, 127
time-of-day variations in radio propagation, 14, 16, 32, 39–41, 43, 46, 51, 178
top-loaded mobile antenna, 481, **482**
toroid broadbanded transformers, 475–477, **476**
toroid cores, 471, **472**, 474
toroidal transformers, 471–475, **472**
towers, **538**, 565–572
  fold- or tilt-over, 567, **567**
  gin pole to raise, 570, **571**
  guy wires and ground stakes for, 570, **572**
  hinged baseplate, **538**, 566–567
  impedance matching in, 507, **508**
  low-frequency operation antennas, loaded tower designs for, 507
  masts and supports for, 556–559, **557, 558**
  pedestals for, 568–570, **569**
  raising of, 570, **570, 571**
  safety issues in, 543–548
  slip-up, 567–568, **568, 569**
transducer action, microwave waveguides and antennas, 394
transformer loop antennas, 306–307, **306**
transformers, 199–200
  core material in, 474
  current transformer wattmeter in, 534–538, **535, 536, 537**
transformers (*Cont.*):
  in directional phased vertical antennas, 249–250, **250, 251**
  feeding type, 476, **477**
  ferrite rods in, 475, **475**
  in high–frequency dipoles and doublets, 146
  multi-impedance balun and, 476, **477**
  in multiband and tunable-wire antennas, 210, **211**
  multifilar windings in, 472
  in shortwave reception antennas, 281
  tapped balun and, 476, **477**
  toroid core mounting in, 471, **472**, 474
  toroid broadbanded, for impedance matching, 475–477, **476**
  toroidal type, 471–475, **472**
  in transformer loop antennas, 306–307, **306**
  trifilar windings in, 472
  turns in, 474
  in VHF/UHF transmitting/receiving antennas, 341–346, **342, 343**
  windings in, 472–473, **473**
transmatch circuits, 208–209, **209**, 462–471, **462**
transmission lines, 59–94, 516–517
  absorbed power (P_a) in, 113–114
  adjustments to, 433–438
  admittance in, 112, 114
  air-articulated line as, 62–63, **63**
  angle of reflection in, 75
  antinodes in, 87
  attenuation constant of, 82
  balanced transmission line and, 208
  cable TV, 67
  capacitance in, 64, 69
  characteristic impedance (Z_o), 64–67, 89
  coaxial cable, 59, 61–62, **61, 62**, 66, 69, 208
  (*See also* coaxial cable)
  conductance in, 64
  current standing wave ratio (ISWR) in, 88
  cutoff frequency (F_o) in, 69
  dielectric loss in, 369
  double-shielded coaxial as, 63, **63**
transmission lines (*Cont.*):
  electrical vs. physical length of, 67–68
  equivalent circuit for, **64**, 516, **517**
  gas-filled line as, 62, **63**
  half-wavelength line as, 83
  hardline as, 62, **62**
  helical line as, 62
  impedance and impedance matching, 71, 72, 82–86, 109–112, **111**
  impedance matching network (Q section) in, 85
  incident power (P_I) in, 89
  inductance in, 64
  installation tips for, 433–438
  insulators in, 61
  load impedance in, 109–110
  longwires and, 217–218
  “looking-in” impedance of, 82–86
  loss coefficient scale in, 112–113
  loss in, 83–84, 91–94, 369
  measurement and adjustment techniques for, 516–517, 525
  mismatch (VSWR) losses in, 91–94
  multiband and tunable-wire antennas and, feeding, 208–212
  nodes in, 87
  noise and, 68–69
  ohmic loss in, 369
  Ohm’s law and, 74
  open line as, 59
  parallel lines as, 59–63, **60**, 65–66
  phase constant of, 82
  propagation constant of, 82
  Q section in, 85
  quarter-wavelength lines as, 84–85
  quarter-wavelength shorted stubs in, 86, **87, 88**
  radiation loss in, 369
  reactance in, 85–86, 109, 525
  rectangular waveguide using parallel type, 370–373, **371**
  reflected power (P_r) in, 89
  reflection coefficient in, 71–79, 89, 112, 114
  reflection loss in, 113–114
  resistance in, 64, 525
  responses in, 70–94
  return loss in, 113, 114
  RLC networks and, 59

transmission lines (*Cont.*):
  shielded parallel line as, 66
  shielded twin lead as, 61, **61**
  shorted lines and, 87
  shorted stubs, 86, **87**, 114
  (*See also* stub matching systems)
  skin effect in, 369
  Smith charts for, 109–114
  standing wave ratio (SWR) in, 86–91, **90**
  standing waves in, 71, **72, 73**, 87
  steady-state ac response in, 70, 79–82, **81**
  step–function response in, 70–71, **70, 71**
  stripline or microstripline as, 63, **64**, 66–67
  stub matching systems (*See* stub matching systems)
  surge impedance (Z_o) in, 516
  time domain reflectometry (TDR) measurements in, 75–79, **76, 77, 78, 79, 80, 81**
  transmission loss coefficient (TLC) in, 112, 114
  troubleshooting tips for, 433–438
  twin lead as, 59, **60**
  velocity factor of, 67–68, 68*t*, 525
  voltage minima and maxima in, 87
  voltage standing wave ratio (VSWR) in, 88–94, **90**, 112, 114, 517, **518**
  waveguides vs., 86
  wavelength calculation in, 110
transmission loss, using Smith charts, 108–109
transmission loss coefficient (TLC), 112, 114
transmitters, 123
transmitting antennas, small loops (*See* loop antennas, small transmitter-type)
transverse electric mode (TE mode) in microwave waveguides and antennas, 375
transverse electromagnetic (TEM) fields, 9, 69, 373–375
transverse magnetic mode (TM mode) in microwave waveguides and antennas, 375
trap dipoles, 204–206, **204, 205**
traveling wave antennas, 220, 223
traveling waves, 127, 128–129, **129**, 134
trenching to apply chemical soil-conductivity improvers, 576, **576**
triangle loop (1λ) large antenna, 290–293, **291**
triangulation of radio signal, 302, 439
trifilar windings in transformers, 472
tropical band, Beverage or "wave" antenna in, 224–228, **226, 228, 229**
troposphere, 14, **15**, 29, 51
tropospheric propagation, 17–19, **18**, 29–32
troubleshooting tips, 433–438
true longwire antenna, 217–219, **217**
tubing construction of vertical dipole,192–194, **193**
tunability, multiband and tunable-wire antennas, 203
tunable dipoles, 160
tunable-wire antennas, 203–212
tuned circuit measurements, 526
tuned feeder antennas, 206, **207**
tuning, 524
  antenna matching tuner in, 281, **285**
  of ground wires, 580–581
  of high–frequency dipoles and doublets, 149–152
  of longwires, 214, **215**, 215
  of loop antennas, small receiver-type, 304, 307–309, **308**, 312–313
  of marine radio antennas, 497–500, **499, 500**
  of mobile HF antennas, 483–485
  of multiband and tunable-wire antennas, 203–205
  of shortwave reception antennas, 281
tuning antennas, 203
tuning capacitance and capacitors, loop antennas, small transmitter-type, 320–321
turns in transformers, 474
TV receive-only (TVRO) satellite dishes, 421
twenty–seven day cycle of radio propagation, 46
twin lead, 59, **60**
two-meter Yagi VHF/UHF antenna, 352–353, **352**
two-port phasing transformer for directional phased vertical antennas, 249–250, **250, 251**

U

U-bolt mounting hardware, 558, **559**
ultrahigh frequency (UHF), 13, 17–19, 23, 29, 32, 39, 61
  in loop antennas, small receiver-type, 302
  in VHF/UHF transmitting/receiving antennas, 339–368
ultraviolet (UV) radiation, 9, 39
umbra region, 17
unity resistance circle, Smith charts, 100
Universal Coordinated Time (UTC), 55
upper sideband (USB), 55

V

varactor tuning, in loop antennas, small receiver-type, 309, **310**
variation, ionospheric, 45–47
vee beam longwire antennas, 221–223, **222**
vee dipole, 236, **239**
vee-sloper antenna, 168–169, **168**
velocity factor (V_F), 14, 525
  of coax, 248
  of microwave waveguides and antennas, 377–381, **378**
  of transmission lines, 67–68, 68*t*
velocity of propagation, 8, 13–14, 137–139, **139**
vertical antennas (*See* vertically polarized HF antennas)
vertical coax VHF/UHF antenna, 346–349, **347, 348**
vertical dipole scanner antenna, 364, **364, 365**
vertical extent radiation pattern, vertically polarized HF antennas, 179, **180**
vertical plane radiation pattern, high–frequency dipoles and doublets, 148–149, **148**
vertical polarization, 12–13, **13**, 27, **28**, 55, 173

vertical polarization (*Cont.*):
  Beverage antenna, 226–227
  helically wound antennas, **239**, 242, **243**
  vertically polarized HF antennas, 173–202
vertically polarized HF antennas, 173–202, **174**
  above-ground mounted, 197
  angle of radiation in, 176–179, **178**
  antenna tuning units (ATU) in, 173
  coaxial vertical antennas in, 191–194, **192, 193**
  coaxial vertical scanner antenna in, 365–368, **367, 368**
  coaxial-fed, **175**
  collinear vertical VHF/UHF antenna, 349, **350**
  collinear vertical antennas in, 191–194, **192**
  construction of, 195–197
  current in, 173, **174**
  delta-fed grounded vertical antenna in, 200–201, **200**
  directionally phased (*See* directional phased vertical antennas)
  "drooping radials" in, 176, **177**
  electric fields in, 173
  electrical connections for,197, **198**, 560–561
  feedpoint circuits for, 197, **198**
  fence posts as supports for vertical antennas, 561–565, **563, 564**
  five-eighths-wavelength antenna, 181, **182**, 183, 197–200
  four/4x4 support posts for, 565, **565**
  gain in, 179–181
  gamma-fed grounded vertical antenna in, 200–201, **201**
  ghost radiator in, 173
  ground losses in, 183
  ground plane vertical antennas in, 176, **177**
  ground-mounted, 176, 195–197, **196**, 561–565, **562**
  grounding and ground system for, 176, 185–187, **188**, 195–197, **196**, 200–201
  half-slope vertical antenna, 238, **240**
vertically polarized HF antennas (*Cont.*):
  half-wave vertical dipole in, 187–191, **190**
  half-wavelength antenna, 181, **182**
  height of antenna in, 181, 187
  helically wound antennas, **239**, 242, **243**
  image antenna in, 173
  impedance and impedance matching, 173, 175–176, 183, **184**, **186**, 195–197, 199–200, **199**
  installation of, 560–572
  insulators in, 175, 195
  isotropic radiation in, 179–181, **181**
  J-pole antennas in, 201
  length of, 181, 183, **184**
  limited-space designs, 238, **240**
  losses in, 183
  low-frequency operation antennas, 502–505, **503**
  mast-mounted, **175**, 194–195, **194, 546**, 560–561, **561, 562**
  mounting, 186, 189–190, 194–197, **194, 196**, 560–565, **561**
  non-ground-mounted, 195, 197
  non-quarter-wavelength verticals in, 181–185
  omega-fed grounded vertical antenna in, 200–201, **202**
  omnidirectionality in, 173, 179–181
  power calculation in, 180
  Q-sections in, 199–200, **199**
  quarter-wavelength antenna, 173, 181, **182**, 183, 187
  radials used in, 176, 186, **189**, 560–561
  radiation patterns in, 178–181, **181, 182**
  radiation resistance in, 183, **184, 185**
  reactance in, 183, **184**
  resistance in, 183
  safety issues, burns or shocks from RF through vertical antennas, 191, 195
  series-fed vertical antennas in, 200
  shortwave reception antennas, 279–281, **280, 282, 283**
vertically polarized HF antennas (*Cont.*):
  shunt-fed vertical antennas in, 200
  skip propagation and, 176–179, **178**
  "spider web" ground system for, 185, **188**
  tower installation, **538**, 565–572
  transformers in, 199–200
  tubing construction of vertical dipole,192–194, **193**
  vertical coax VHF/UHF antenna, 346–349, **347, 348**
  vertical dipole scanner antenna in, 364, **364, 365**
  vertical extent radiation pattern in, 179, 180
  VHF/UHF transmitting/receiving antennas, 340–341
  voltage in, 173, **174**
  voltage standing wave ratio (VSWR) in, 176, 195
  window-sill antennas, 238, **241**
  Yagi antennas and, 201
very high frequency (VHF), 13, 17–19, 23, 29, 30, 32, 39, 41, 46, 54, 55, 170–171
  emergency operations antennas and, 487
  loop antennas, small receiver-type and, 302
  marine radio antennas and, 495
  vertically polarized HF antennas and, 191
  VHF/UHF transmitting/receiving antennas and, 339–368
very low frequency (VLF), 29, 55, 61
  Beverage or "wave" antenna in, 224–228, **226, 228, 229**
VHF/UHF transmitting/receiving antennas, 339–368
  baluns in, 341–346, **342, 343**
  bandwidth of, 339
  beam antennas in, 349–352, **351, 352**
  "bowtie" dipoles in, **361**, 362
  citizens' band antennas as, 341
  coaxial vertical antenna for, 346–349, **347, 348**
  coaxial vertical scanner antenna in, 365–368, **367, 368**
  collinear vertical antenna, 349, **350**

VHF/UHF transmitting/receiving antennas (*Cont.*):
  delta feed match for, 343, **344**, 345
  design limitations in, 340
  directivity in, 349
  drooping dipole scanner antenna in, 363, **363**
  effective radiated power (ERP) in, 349
  five-eighths-wavelength 2-meter antenna for, 353, **354**
  folded dipoles in, 362, **362**
  frequency spectrum of, 339, 341
  gain in, 349
  gamma match system for, 346, **346**
  groundplane antenna for, 356, **357**
  groundplane scanner antenna in, 364, **366**
  halo antenna for, 356, **358**
  impedance matching in, 341–346, **342**
  J-pole antenna for, 353–355, **355**
  length of, 339
  line-of-sight communications and, 339
  longwires as, 340–341
  low frequency (LF) vs., 339
  lower band antennas on VHF/UHF, 340–341
  mounting, 347
  OSCAR satellite communications and, 339
  quad beam antenna for, 356, 358, **359**
  safety issues in, 340
  satellite communications and, 339
  scanner band antennas, 360–368
  scanner skyhook antennas in, 362–368, **363**
  scanner-vision antennas in, 360–362, **361**
  size of, as limiting factor, 340
  skip propagation in, 339
  stub matching system for, 345–346, **345**
  transformers in, 341–346, **342, 343**
  two-meter Yagi antenna, 352–353, **352**
  vertical antennas as, 340–341
  vertical dipole scanner antenna in, 364, **364, 365**
  voltage standing wave ratio (VSWR) in, 341
  wavelength in, 339
  Yagi antennas in, 349–352, **351, 352**
virtual height, 43, **44, 45**
voltage, 123–120, 131–137, **132, 133**, 531
  (*See also* voltage standing wave ratio)
  in high–frequency dipoles and doublets, 143–146, **143**
  lag, 126
  in loop antennas, small receiver-type, 303, 304
  in loop antennas, small transmitter-type, 320, 323
  minima and maxima of (*See* antinodes; nodes)
  standing waves of, 134–136, **135, 136**, 139
  in vertically polarized HF antennas in, 173, **174**
  in vertically polarized HF antennas, 195
  in VHF/UHF transmitting/receiving antennas, 341
  in Yagi antennas, 261
voltage standing wave ratio (VSWR), 88–91, **90**, 515, 517–519, 529, 532
  calculation of, 517–519
  EZNEC for Windows modeling of, 337, **338**
  frequency vs., 434–437, **435**
  in hidden and limited-space antennas, 238
  in high–frequency dipoles and doublets in, 144, 146, 149–151, **150**, 160
  in longwires, 215
  in loop antennas, small transmitter-type, 320
  measurement of, 433, 435–437, **436**, 529, 532
  in microwave waveguides and antennas, 388
  in mobile HF antennas, 483
  in multiband and tunable-wire antennas, 203
  NECWin Basic modeling of, 333, **335**
  using Smith charts circle, 100, **104**, 112, 114, 115, 117
  in stub matching systems, 115, 117
  in transmission lines, 91–94, 112, 114, 517, **518**

W

water analogy of radio wave propagation, 6–7
Watson-Watt Adcock array antennas, 447–450, **449, 450**
wattmeters, 529–539, **531, 533, 534, 535, 536, 537, 538**
waveguides, 86, 369–415
  (*See also* microwave waveguides and antennas)
wavelength calculation, using Smith charts, **103**, 105, 110, 115
wavelength, 7–8, 28
  high–frequency dipoles and doublets, 141–142
  longwires, 217, 218–219
  microwave waveguides and antennas, 377–381, **378**
  VHF/UHF transmitting/receiving antennas, 339
"wave" antenna, 224–228, **226, 228, 229**
weather–related propagation phenomena, 28, 36
wedge dummy loads, in microwave waveguides and antennas, 385, **386, 387**
Wheatstone bridge, 519–521, **520**
whip antennas for marine radio, 496–497, **497**
wide-aperture direction finder (WADF), 451
windings in transformers, 472–473, **473**
Windom antenna, 209–210, **211**
Windom, Loren, 209
window-sill antennas, 238, **241**
wire antenna, basic design, 548, **548**
wire used in longwires, 214
wireless telegraphy, 1–2
Wolf Number, 40
World Administrative Radio Conferences (WARC), 2

Wullenweber array antennas, 451, **452**
WWV radio station, 40, 47
WWVH radio station, 40

X
X bands, 69
x-ray radiation, 39

Y
Yagi antennas, 203, 257–262, **259**, 421
  capacitors in, 261
  directors in, 258–259
  feedpoints, 260–261, **261**
  gamma match feed system for, 260–261, **261**
  impedance matching in, 260–261, **261**
  length of, 259–260
  noise and VSWR in, 261
  parasitic elements in, 258–259
  quad beam antennas vs., 265
  radiation pattern of, 257–258, **258**
  reflectors in, 258–259
  two-meter Yagi VHF/UHF antenna, 352–353, **352**
  vertically polarized HF antennas, 201
  VHF/UHF transmitting/receiving antennas, 349–352, **351, 352**
  Yagi-Uda antenna, 258

Z
Zepp antenna, 164–165, **165**, 171
  end-fed multiband, 208, **208**
zipcord, 275
ZL-special beam antenna, 20-m, 262–263, **262**
Zurich Smoothed Sunspot Number, 40

About the Author

Joe Carr was a military electronics technician and author. Along with the best-selling *Practical Antenna Handbook*, his other books with McGraw-Hill include *Secrets of RF Circuit Design*. Carr also contributed a monthly column to *Nuts & Volts* magazine entitled “Open Channel.”

This page intentionally left blank

This page intentionally left blank

SOFTWARE AND INFORMATION LICENSE

The software and information on this diskette (collectively referred to as the “Product”) are the property of The McGraw-Hill Companies, Inc. (“McGraw-Hill”) and are protected by both United States copyright law and international copyright treaty provision. You must treat this Product just like a book, except that you may copy it into a computer to be used and you may make archival copies of the Products for the sole purpose of backing up our software and protecting your investment from loss.

By saying “just like a book,” McGraw-Hill means, for example, that the Product may be used by any number of people and may be freely moved from one computer location to another, so long as there is no possibility of the Product (or any part of the Product) being used at one location or on one computer while it is being used at another. Just as a book cannot be read by two different people in two different places at the same time, neither can the Product be used by two different people in two different places at the same time (unless, of course, McGraw-Hill's rights are being violated).

McGraw-Hill reserves the right to alter or modify the contents of the Product at any time.

This agreement is effective until terminated. The Agreement will terminate automatically without notice if you fail to comply with any provisions of this Agreement. In the event of termination by reason of your breach, you will destroy or erase all copies of the Product installed on any computer system or made for backup purposes and shall expunge the Product from your data storage facilities.

LIMITED WARRANTY

McGraw-Hill warrants the physical diskette(s) enclosed herein to be free of defects in materials and workmanship for a period of sixty days from the purchase date. If McGraw-Hill receives written notification within the warranty period of defects in materials or workmanship, and such notification is determined by McGraw-Hill to be correct, McGraw-Hill will replace the defective diskette(s). Send request to:

Customer Service
McGraw-Hill
Gahanna Industrial Park
860 Taylor Station Road
Blacklick, OH 43004-9615

The entire and exclusive liability and remedy for breach of this Limited Warranty shall be limited to replacement of defective diskette(s) and shall not include or extend to any claim for or right to cover any other damages, including but not limited to, loss of profit, data, or use of the software, or special, incidental, or consequential damages or other similar claims, even if McGraw-Hill has been specifically advised as to the possibility of such damages. In no event will McGraw-Hill's liability for any damages to you or any other person ever exceed the lower of suggested list price or actual price paid for the license to use the Product, regardless of any form of the claim.

THE McGRAW-HILL COMPANIES, INC. SPECIFICALLY DISCLAIMS ALL OTHER WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, ANY IMPLIED WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Specifically, McGraw-Hill makes no representation or warranty that the Product is fit for any particular purpose and any implied warranty of merchantability is limited to the sixty day duration of the Limited Warranty covering the physical diskette(s) only (and not the software or in-formation) and is otherwise expressly and specifically disclaimed.

This Limited Warranty gives you specific legal rights; you may have others which may vary from state to state. Some states do not allow the exclusion of incidental or consequential damages, or the limitation on how long an implied warranty lasts, so some of the above may not apply to you.

This Agreement constitutes the entire agreement between the parties relating to use of the Product. The terms of any purchase order shall have no effect on the terms of this Agreement. Failure of McGraw-Hill to insist at any time on strict compliance with this Agreement shall not constitute a waiver of any rights under this Agreement. This Agreement shall be construed and governed in accordance with the laws of New York. If any provision of this Agreement is held to be contrary to law, that provision will be enforced to the maximum extent permissible and the remaining provisions will remain in force and effect.

$14

More Wire Antenna Classics

Volume 2

More **of the best articles from ARRL publications**

More Wire Antenna Classics Volume 2

More Wire Antenna Classics

Volume 2

Compiled by Chuck Hutchinson, K8CH

Production:
Paul Lappen

Cover Design:
Sue Fagan

Published by:
ARRL *The national association for AMATEUR RADIO*
225 Main Street
Newington, CT 06111-1494

Contents

1 Dipoles

2 Off-Center-Fed Dipoles

3 Multiband Antennas

4 Loop Antennas

5 Collinear Antennas

6 Wire Beams

7 Vertically Polarized

8 Receiving Antennas

9 Miscellanea

CHAPTER ONE

DIPOLES

1-1 **An Inverted V for Field Day**
From June 1968 *QST*

1-4 **E-Z-Up Antenna for 75 and 40**
From October 1961 *QST*

1-6 **A Zip-Cord Special Antenna**
From May 1972 *QST* (Hints & Kinks)

1-7 **Zip-Cord Special Antennas–Do They Work?**
From March 1979 *QST*

1-9 **Piggyback Antenna for the Ten-Meter Band**
From April 1973 *QST* (Hints & Kinks)

1-10 **RFD-1 and RFD-2: Resonant Feed Line Dipoles**
From August 1991 *QST*

1-14 **Off-Center-Loaded Dipole Antennas**
From September 1974 *QST*

1-20 **A Simple Broadband Dipole for 80 Meters**
From September 1993 *QST*

1-25 **Once More With the 80-Meter Broadband Dipole**
From June 1985 *QST* (Technical Correspondence)

1-26 **Fat Dipoles**
From *The ARRL Antenna Compendium, Vol 2*

By Frank Gue, VE3DPC, ex-VE6BH **From *QST*, June 1968**

An 80-Meter Inverted V for Field Day

It's that time of year again when portable operation is a delight. Here is a full-sized low-frequency antenna that you can put up in the field or at home without any help.

A Field Day antenna for 80 meters should be light, portable, inexpensive, capable of easy erection without any special equipment, and it should not contain any fragile parts.

Of course, the antenna must perform adequately.

The popular inverted V suggests itself as an answer to these requirements. It can be constructed so that it guys itself, minimizing the inevitable exasperating tangle of rope and wire, and it requires only one mast. Like all dipoles the inverted V can be fed at the center with low impedance line, and elements for several bands can be connected in parallel at the feed point of the antenna.

The inverted V shown in the photographs and sketches is simple, easily managed (20 minutes to erect, 10 minutes to dismantle), and costs under $15 for all components. Not only is it a convenient antenna for portable operation, but it can serve as a fine permanent antenna for the home station as well. In fact, that's what has happened to the inverted V pictured; it's been in use in the back yard of VE3DPC ever since Field Day, 1967.

Assembly

As shown in Figure 1, the antenna support consists of three 10-foot sections of 1¼-inch TV mast. One end of each of the two lower sections is swaged, permitting all three sections to be force-fitted together. Hooks of music wire were installed in the top section, so that the element wire and feed line could be easily coiled for storage.

The center insulator (Figure 2) is made of maple; however, if this wood cannot be found, birch or fir can be used instead.

Figure 2 shows how the center insulator was drilled and Figure 3 shows how the feed line and antenna element were attached. Dimensions are obviously not critical, and the sketches are intended to show in a general way how the insulator was made, not to give precise detail. Just be careful where you drive the screws!

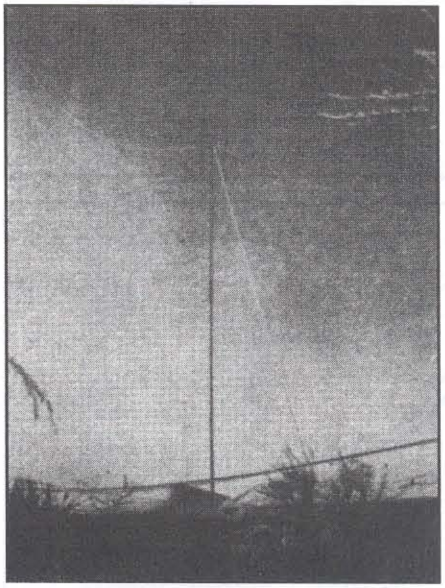

The inverted V as it looked this past winter in the author's backyard.

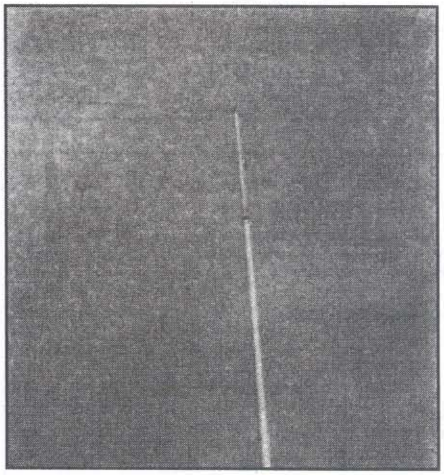

Upper portion of the 80-meter inverted V as it appeared at the Field Day site.

Softwood (pine) was picked for the end insulators (Figure 4) because it is adequate for the application and easier to work with than maple. Fragile ceramics should not be used in place of the wood insulators.

The 80-meter element (Figures 3 and 5) is made from small-gauge, Teflon-insulated, stranded Copperweld that was bought at a surplus outfit. This wire is worth looking for; it's tough, durable, light, and easy to handle. If operation on both 80 and 40 is desired, the builder can connect a 40 meter element to points A and B in Figure 3.[1]

The feed line can be 50-ohm coax, 72-ohm coax or 72-ohm Twin-Lead. I chose the latter because it is low cost, lightweight, very flexible, easy to store, and because it lends itself to such outrageous techniques as being tied in knots; however, since most transmitters have coax output, either a balun or an antenna coupler must be used between the feed line and the transmitter to maintain balance.[2] If elements for more than one band are to be used, the antenna coupler is preferred because it will reject harmonics of the transmitting frequency that might fall in the range of one or more of the paralleled antennas.

Raising and Lowering

Transporting the antenna to and from the Field Day site is easy. Tape the three mast sections together, wrap them with cloth to protect the car's finish, and tie the works between one of the door handles and the rear bumper. If the antenna extends past the back of the car, don't forget to tie a red flag on the end of the bundle to alert other drivers.

Upon arrival at the Field Day site, unpack the antenna and join the TV masts together. The three-section support is light enough to be "walked up" by one man, if the wind is under 10 miles per hour; however, prior to the antenna raising, you must put stakes into the ground, determine the

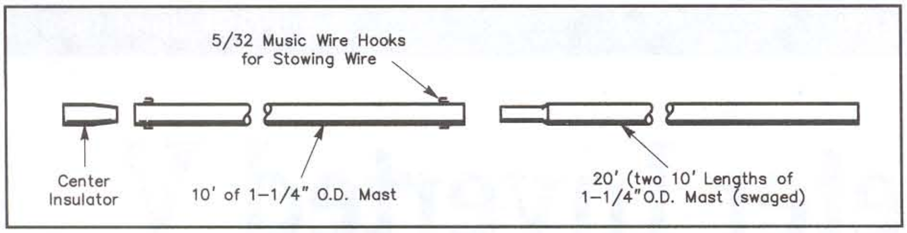
5/32 Music Wire Hooks
for Stowing Wire
Center
Insulator
10' of 1-1/4" O.D. Mast
20' (two 10' Lengths of
1-1/4"O.D. Mast (swaged)

Figure 1—The makings of the inverted-V center support.

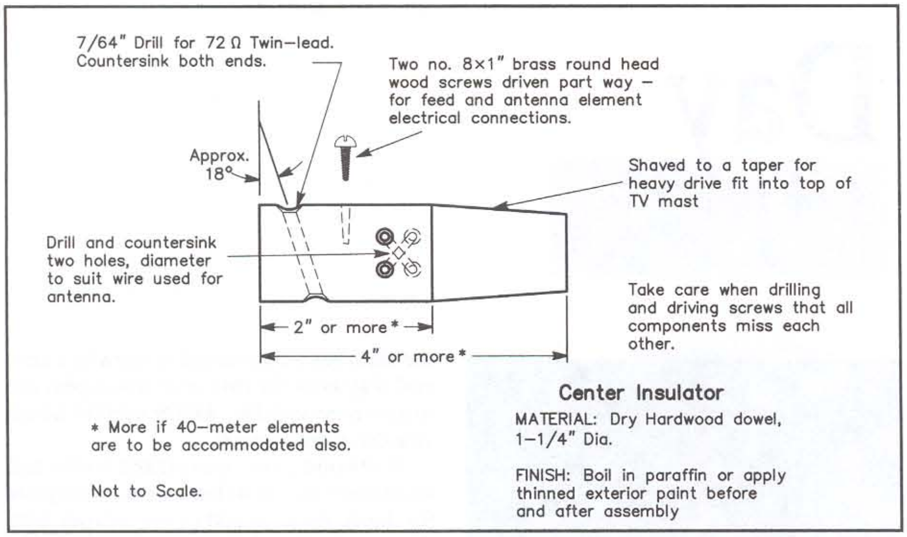
7/64" Drill for 72 Ω Twin-lead.
Countersink both ends.
Two no. 8×1" brass round head
wood screws driven part way -
for feed and antenna element
electrical connections.
Approx.
18°
Shaved to a taper for
heavy drive fit into top of
TV mast
Drill and countersink
two holes, diameter
to suit wire used for
antenna.
Take care when drilling
and driving screws that all
components miss each
other.
2" or more*
4" or more*
Center Insulator
MATERIAL: Dry Hardwood dowel,
1-1/4" Dia.
FINISH: Boil in paraffin or apply
thinned exterior paint before
and after assembly
* More if 40-meter elements
are to be accommodated also.
Not to Scale.

Figure 2—An X-ray view of the center insulator.

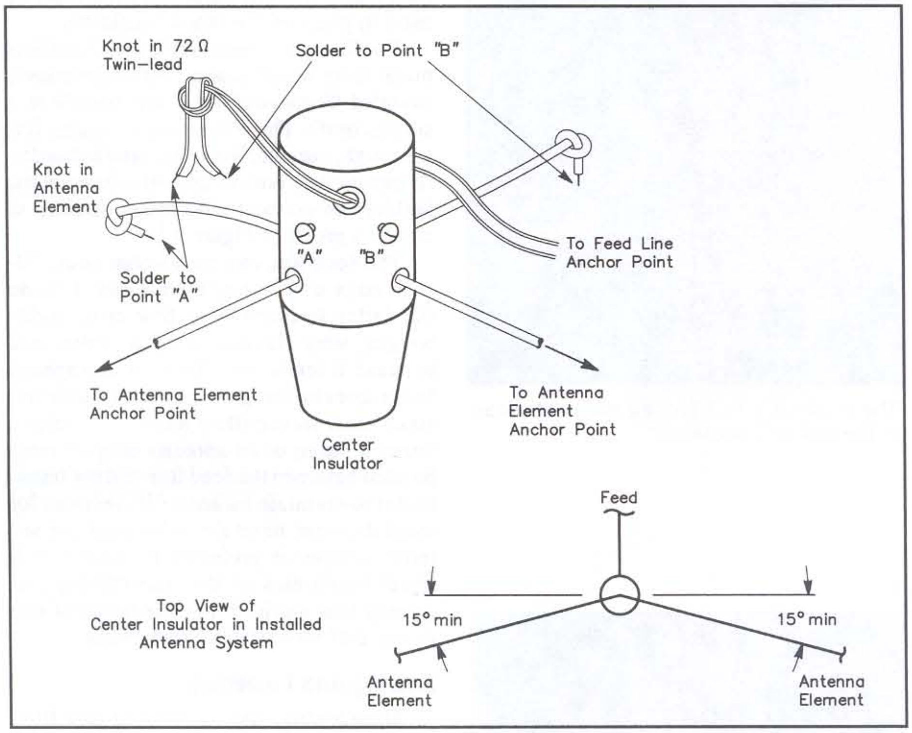
Knot in 72 Ω
Twin-lead
Solder to Point "B"
Knot in
Antenna
Element
"A"
"B"
To Feed Line
Anchor Point
Solder to
Point "A"
To Antenna Element
Anchor Point
To Antenna
Element
Anchor Point
Center
Insulator
Feed
Top View of
Center Insulator in Installed
Antenna System
15° min
15° min
Antenna
Element
Antenna
Element

Figure 3—Details of how the feed line and element wires should be attached to the center insulator, and a top view of the installed insulator.

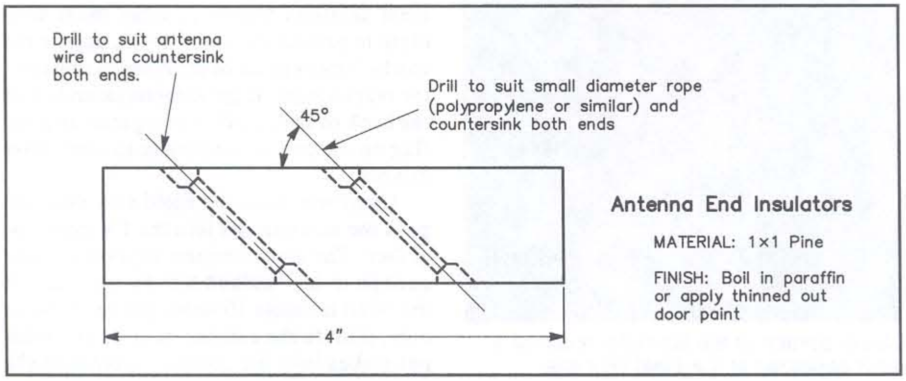
Drill to suit antenna
wire and countersink
both ends.
45°
Drill to suit small diameter rope
(polypropylene or similar) and
countersink both ends
Antenna End Insulators
MATERIAL: 1×1 Pine
FINISH: Boil in paraffin
or apply thinned out
door paint
4"

Figure 4—A sketch of one of the wooden end insulators. Two insulators are required for each dipole used.

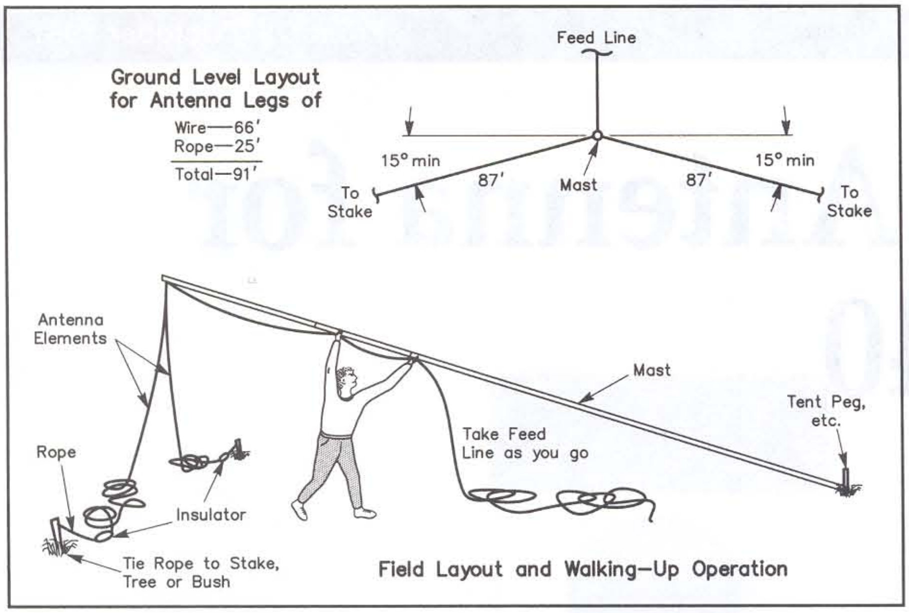

Figure 5—By using the dimensions and layout shown in the upper right corner of the sketch—so that ground stakes could be installed and proper length guys attached between them and the mast—the author was able to raise his 80-meter inverted V without any help. For other element lengths than shown, one man can easily put up the antenna by using simple geometry to determine the amount of rope required and the locations of the stakes.

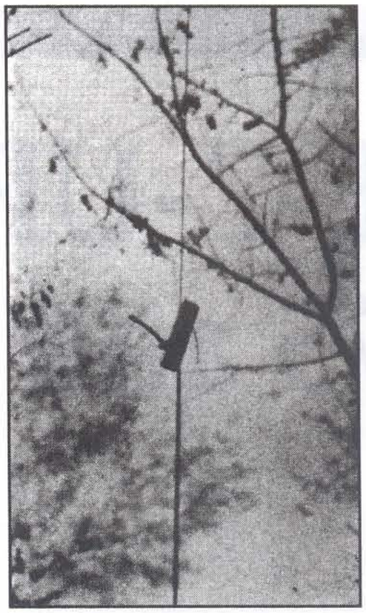

A sideview of one of the wooden end insulators.

exact length the guys will be when the antenna is in its operating position, and attach the guys to the stakes (Figure 5).

If help is available, it won't be necessary to install stakes or measure guys prior to putting up the antenna. Have the helpers pay out each half of the antenna element and keep slack out of the wire as the mast goes up. Station the men about 45 degrees off the line of the mast as it lies on the ground, and then, once the mast is up, have the helpers walk around to the final positions shown in Figure 5. The last few degrees of "walking up" is where your control is poorest.

With the rear guy (feed line) in your hand, move quickly and smoothly past top dead center and promptly and firmly apply back tension on the antenna element. Once the feed line is snubbed, you can easily correct the mast for lean by lifting the base and putting it where it suits you. If you are fortunate enough to have three helpers, there is no problem at all. Give the third man the feed line, station him as far away as possible from the mast and antenna element, and have him take up the slack as the mast goes up.

Base support can be little or nothing. The mast, which is now permanently erected in my back yard, rests directly on the grass. It was moved a couple of inches during each mowing operation, and no apparent harm came to the lawn.

Lowering the antenna is brutal but quick. Release the back guy (feed line) and let the whole system fall. If the ground is reasonably flat, this will be OK. But if the ground is very rough, the pipe may take a set as it strikes the ground. Under such conditions, it would be better to use helpers to reverse the erecting procedure.

The antenna has been left up for a period of several months, during which there have been many autumn storms and a severe ice storm. It has shown no signs of distress, and it has performed normally when wet or loaded with ice. The small size of the inverted V's parts has led to little wind and ice loading and is probably the main reason for the antenna's durability.

---

[1] Although dimensions for an 80-meter inverted V are given in Figure 5, these aren't necessarily the optimum dimensions for every installation: there are too many variables involved. If the antenna is not to be used with a transmatch and open-wire line, it is best to experimentally determine the length of the inverted V. This can be done by starting with an overall length equal to about 525 divided by the frequency in megahertz. The ends of the inverted V should then be trimmed until the SWR is minimum at the operating frequency. A simpler method is described in footnote 2.

[2] An easier arrangement, especially for multiband operation, can be achieved by using a transmatch and either 300-ohm Twin-Lead or 450-ohm Ladder-Line. Only one dipole—about 100 feet long (length not critical)—is all that is necessary to cover all the amateur bands between 3.5 and 30 MHz.

By John C. Allred, W5LST **From *QST*, October 1961**

E-Z-UP Antenna for 75 and 40

Full-size dipoles for the 75- and 40-meter bands occupy more space than is conveniently available on the 75 by 113-foot lot at W5LST. The increasing popularity of the "drooping," or inverted-vee, dipole antenna among amateurs led us to investigate it for our somewhat crowded conditions. Based on the electrical design of Glanzer,[1] this system has performed meritoriously at W5LST. Requiring only one support, it was surprisingly easy to erect, gives a satisfactory s.w.r. over the phone bands and, importantly, the cost was less than thirty dollars complete. A plan-view sketch is shown in Figure 1.

The Mast

The mast is a telescoping Channel-master, capable of 50-foot height, but extended only to 35 feet. Extending the upper sections to less than their full lengths gives rigidity to the mast, and has apparently eliminated the need for guys on each section, as recommended by the manufacturer. To date this mast has withstood gusts of 50 miles per hour without a shudder; it is yet to be tested in a real gale, however.

As shown in Figure 2, a 9-foot length of 3-inch pipe is cast in concrete with 6 feet of its length extending above ground. Three pairs of clamps, such as those used on chain-link fence, secure the mast to the upright standard. During erection, these clamps are loosened and all three lie at the base of the standard, so that the mast need only be lifted about 6 inches to be put in place. When the mast is in place, the clamps are raised and tightened.

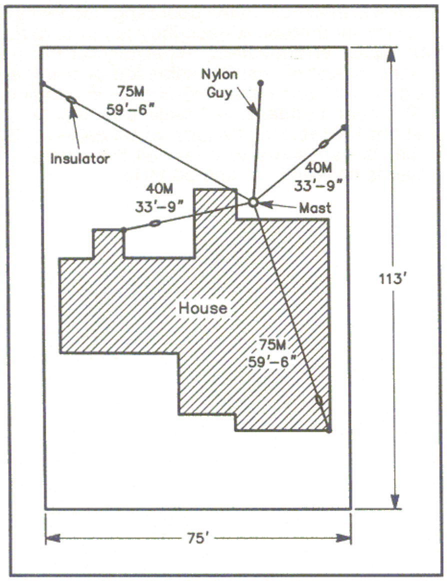

Figure 1—W5LST's layout for effective 40- and 80-meter antennas on a small lot. The two "drooping" dipoles are fed in parallel with a single coax line. Nylon-line extensions are used to reach convenient anchorages.

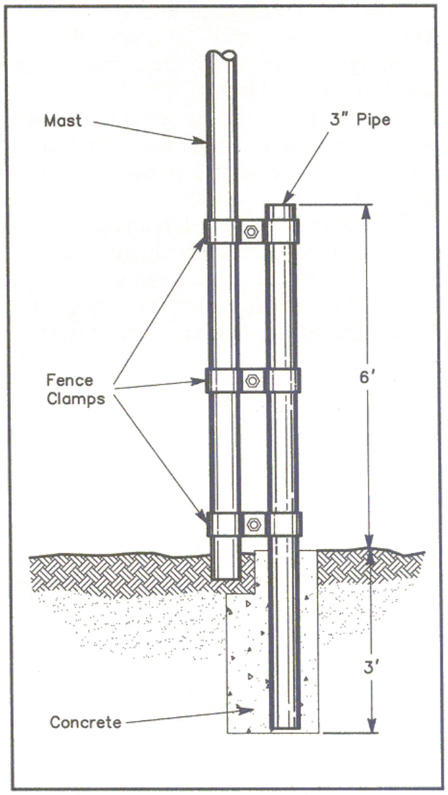

Figure 2—The mounting for the antenna mast. The 9-foot pipe is guyed temporarily while the concrete is poured.

Rigging

Except for the antenna conductors, all rigging is of nylon line of 500-pound test. A halyard is reeved through a pulley of suitable size which is wired securely to the top of the mast. The two ends of the halyard are made fast to a harness snap which, in turn, supports the center of the antennas. Provision of this halyard has proved to be a great convenience in permitting inspection of the antenna connections and the adjustment of tension in the wires without the necessity for lowering the mast.

Nylon line has a tendency to ravel at its ends but this problem is easily solved. Most fastenings were made with two halfhitches, followed by sewing the end of the line to itself with thread, as shown in Figure 3A, and doping with one of the quick-drying model-airplane cements.

Some weeks after the initial installation, it became apparent that some additional stabilization of the mast against occasional strong northerly winds would be desirable. Accordingly a nylon line was run from the harness snap at the top of the mast to a convenient anchor in the back yard, which happened to be the top of the children's swing set. Experience seems to show that the antenna wires, together with the additional nylon line, stabilize the mast against aerodynamically-excited vibration, without any appreciable strain on the antennas.

The Antennas

As shown in Figure 3, the two antennas are connected in parallel at the top of the mast. The lower ends are connected to convenient tie points so that the two legs of a given antenna are more or less in a straight line. To our great surprise, very little effect is produced by moving the ends of the an-

---

[1] Glanzer, "The Inverted V-Shaped Dipole," *QST*, August, 1960.

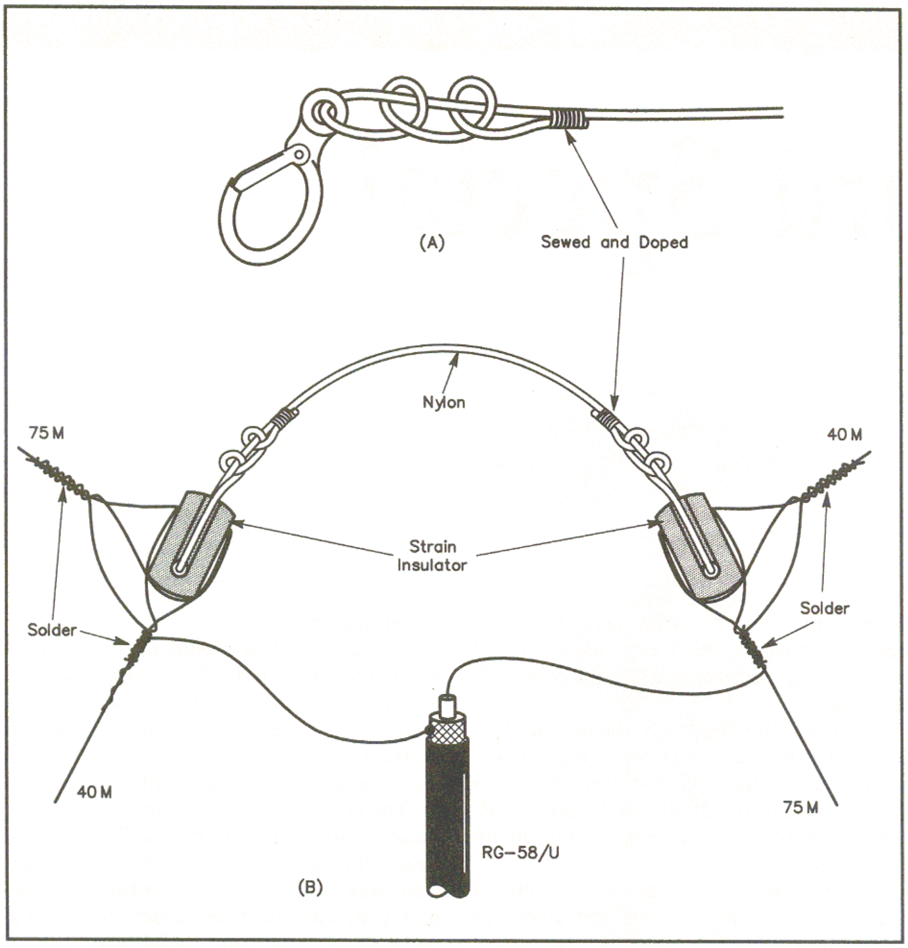

Figure 3—The halyard and feeder arrangement. The coaxial cable is sealed securely with polyethylene tape after the connections are made. The harness snap of A at the end of the hoisting halyard engages the bridle between the two insulators in B.

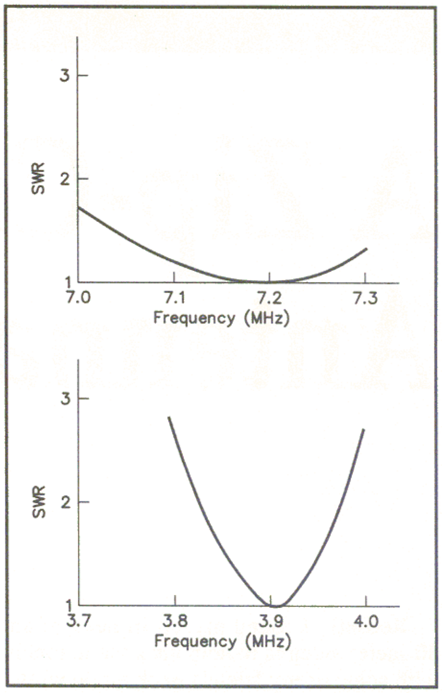

Figure 4—SWR curves as indicated by a "Monimatch"-type SWR indicator.

tennas either horizontally or vertically. There is apparently negligible electrical interaction between them as indicated by the s.w.r. bridge.

Pruning of the antennas is, as always, desirable. In our case, the optimum lengths of each leg turned out to be 33 feet 9 inches for 40 meters and 59 feet 6 inches for 75 meters. Figure 4 shows the performance of the antennas on 40 and 75 as measured by a Heath s.w.r. bridge. Although the antennas tune sharply, they are usable over the entire phone band in each case, and it would be difficult to imagine a better performer on the lower-frequency bands than this simple antenna system.

A Zip-Cord Special Antenna

Recently I found myself in need of an 80-meter antenna to help me keep in touch with some of my friends back home while operating portable in Arizona. The antenna and feed line were not readily available, but I remembered that one of my friends had recommended zipcord as a substitute feed line for a dipole, provided low power was used. Going one step further, I felt the entire antenna could be made of this low-priced material. I purchased 80 feet of zip cord at a local hardware store (at 4 cents a foot) and peeled it apart for 60 feet. This gave me the required 120 feet for a dipole, plus a 20-foot feed line which was taped to prevent further splitting. The feeder was fastened to a 2×4 which was used as a center support. The ends were extended as far as possible, then tied with heavy cord to a fence that was in the trailer park where I was staying. This made a slight inverted-V antenna.

To reach the desired frequency of 3935 kHz it was necessary to shorten the antenna somewhat. I did this by folding three feet on each end of the dipole back upon itself. This enabled me to keep skeds with many "7s" at home.

Having met with success on 80-meters I immediately began to eye the 40-meter band for possible use. I then measured out 32 feet from the center and cut the wire there, then tied a knot in the two pieces, leaving a couple of inches of bare wire dangling. Now for forty meters, I merely disconnect the wires and reconnect them for 80 meters.

As the *ARRL Antenna Book* states, a 40-meter dipole will often present a good match at 15-meters, and this one is no exception; I have worked many stations across the country on 15 meters with low power while using this simple antenna.

The *Handbook* or *Antenna Book* has not recommended this type of feed line in many years, though using a twisted pair at one time was usual practice in feeding a doublet. But, when the need is there for a quick and easy antenna, I don't think you can find an easier one as long as you remember to use low-power transmitters (100 watts or less) and trim the wire for a low value of SWR. — *Frank S. Wise, W7QYC*

By Jerry Hall, K1TD From *QST*, March 1979

Zip-Cord Antennas– Do They Work?

Basic Amateur Radio: Parallel power cord is readily available and is easy to work with. How efficient is it when used at radio frequencies? Well, that depends.

Zip cord is readily available at hardware and department stores, and it's not expensive. The nickname, zip cord, refers to that parallel-wire electrical cord with brown or white insulation used for lamps and many small appliances. The current price is about 6 cents a foot in hundred-foot rolls in most parts of the U.S. The conductors are usually no. 18 stranded copper wire, although larger sizes may also be found. Zip cord is light in weight and easy to work with.

For these reasons, zip cord is frequently used as both the transmission line and the radiator section for an emergency dipole antenna system. The radiator section is obtained simply by "unzipping" or pulling the two conductors apart for the length needed to establish resonance for the operating frequency band. The initial dipole length can be determined from the equation $\ell = 468/f$, where ℓ is the length in feet and f is the frequency in megahertz. (It would be necessary to unzip only half the length found from the formula, since each of the two wires becomes half of the dipole.) The insulation left on the wire will have some loading effect, so a bit of length trimming may be needed for exact resonance at the desired frequency.

For installation, many amateurs like to use the electrician's knot shown in Figure 1 at the dipole feed point to keep the transmission-line part of the system from unzipping itself under the tension of dipole suspension. This way, if zip cord of sufficient length for both the radiator and the feed line is obtained, a solder-free installation can be made right down to the input end of the line. Granny knots (or any other variety) can be used at the ends with cotton cord to suspend the system. You end up with a lightweight low-cost antenna system that can serve for portable or emergency use.

But just how efficient is a zip-cord antenna system? Since it is easy to locate the materials and simple to install, how about using such for a more permanent installation? Upon casual examination, zip cord looks about like high-quality 72-ohm balanced feed line. Does it work as well? Ask several amateurs these questions and you're likely to get answers ranging all the way from, "Yes, it's a very good antenna system!" to "Don't waste your time and money – it's not worth it!" Myths and hearsay seem to prevail, with little factual data. The intent of this article is to rectify that situation.

Zip Cord as a Transmission Line

In order to determine the electrical characteristics of zip cord as a radiofrequency transmission line, we purchased a 100-foot roll and subjected it to tests in the ARRL laboratory with an rf impedance bridge. Zip cord is properly called parallel power cord. The variety purchased was manufactured for GC Electronics, Rockford, IL, being 18-gauge, brown, plastic-insulated type SPT-1, GC cat. no. 14-118-2G42. Undoubtedly, minor variations in the electrical characteristics will occur among similar cords from different manufacturers, but the results presented here are probably typical.

As the first step, we checked the physical length of the wire. GC generously provided nearly 4 feet over the specified 100-ft length, and the extra length was promptly lopped off.

We wanted to avoid measurement errors that might arise from coiling the wire on the supply spool (inductance between turns) or laying it out on the vinyl-tiled cement-slab floor (capacitance to a large surface), so the second step was to suspend the wire in a long hallway, about a foot or so below the false ceiling. Cotton twine was used for the supporting material. Undaunted by snide comments from fellow staff members about cheating the telephone company out of installation work and facetious suggestions

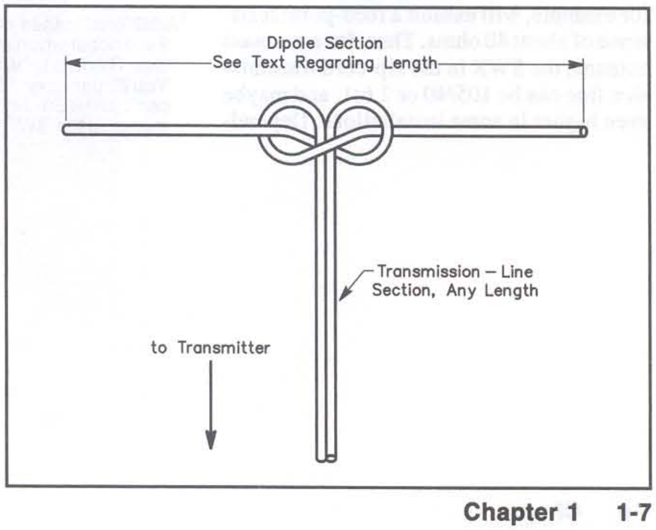
Dipole Section
See Text Regarding Length
Transmission – Line Section, Any Length
to Transmitter

Figure 1—This electrician's knot, often used inside lamp bases and appliances in lieu of a plastic grip, can also serve to prevent the transmission-line section of a zip-cord antenna from unzipping itself under the tension of dipole suspension. To tie the knot, first use the right-hand conductor to form a loop, passing the wire behind the unseparated zip cord and off to the left. Then pass the left-hand wire of the pair behind the wire extending off to the left, in front of the unseparated pair, and thread it through the loop already formed. Adjust the knot for symmetry while pulling on the two dipole wires.

that we use two paper cups and a string for our intercom, we continued with preparation for the tests.

With a General Radio 1606A rf impedance-measuring bridge, we made measurements at the input end of the line on 10, 15 and 29 MHz, while the far end of the line was terminated first in a short and then in an open circuit. We did notice during measurements that our readings changed somewhat as people walked in the hallway, especially during the 29-MHz tests. After a seemingly endless period of shooing away would-be helpers who disrupted measurements by grabbing the line, hanging objects on it, and so forth, we finally arrived at a set of readings considered to be satisfactory for our purposes. From there, a bit of work with an electronic calculator and a Smith Chart told us what we wanted to know — the electrical characteristics at radio frequencies.

The results? Well, as the saying goes, I have some good news and some bad news.

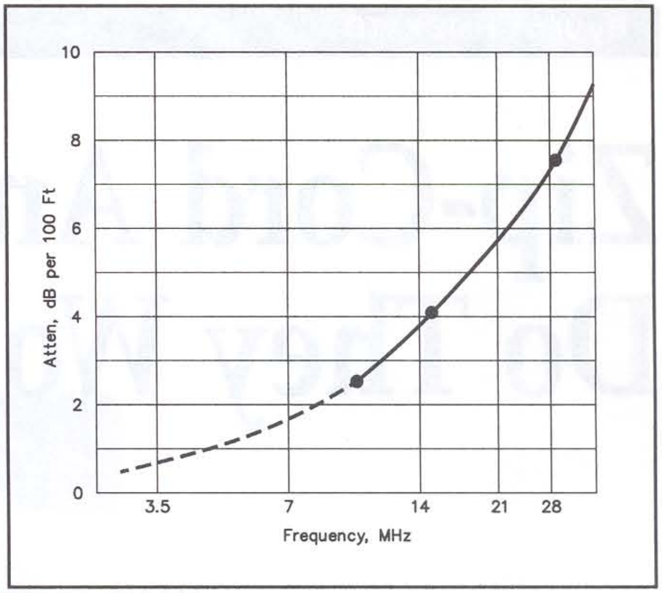

Figure 2—Attenuation of zip cord in decibels per hundred feet when used as a transmission line at radio frequencies. Measurements were made only at the three frequencies where plot points are shown, but the curve has been extrapolated to cover all high-frequency amateur bands.

First, the Good News

If ever you need a balanced 105-ohm transmission line, then zip cord is the stuff for you! Its characteristic impedance was determined to be 107 ohms at 10 MHz, dropping in value to 105 ohms at 15 MHz and to a slightly lower value at 29 MHz. The nominal value is 105 ohms at hf. The velocity factor of the line was determined to be 69.5 percent.

This reported change in impedance with frequency may raise a few eyebrows, but it is a fact that most lines do not exhibit a constant characteristic impedance across a range of frequencies. However, this writer has not previously encountered any line which was not "flat" between 3 and about 50 MHz. Zip cord may be exceptional from this standpoint, especially considering that the insulating material was not chosen for its qualities at rf.

And Now the Bad News

Who needs a 105-ohm line, especially to feed a dipole? Most of us know that a dipole in free space exhibits a feed-point resistance of 73 ohms, and at heights above ground of less than 1/4 wavelength the resistance is even lower. An 80-meter dipole at 35 feet, for example, will exhibit a feed-point resistance of about 40 ohms. Thus, for a resonant antenna, the SWR in the zip-cord transmission line can be 105/40 or 2.6:1, and maybe even higher in some installations. Depending on the type of transmitter in use, the rig may not like working into the load presented by the zip-cord antenna system.

But the really bad news is still to come–line loss! Figure 2 is a plot of line attenuation in decibels per hundred feet of line versus frequency. Chart values are based on the assumption that the line is perfectly matched (sees a 105-ohm load as its terminating impedance). For lengths other than 100 feet, multiply the figure by the actual length in feet and then divide by 100.

In a feed-line, losses up to about one decibel or so can be tolerated, because at the receiver a 1-dB difference in signal strength is just barely detectable. But for losses above about 1-dB, beware. Remember that if the total losses are 3 dB, half of your power will be used up just to heat the transmission line.[1]

Based on this information, we can see that a hundred feet or so of zip-cord transmission line on 80 meters might be acceptable, as might 50 feet on 40 meters. But for longer lengths and higher frequencies, look out! The losses become appreciable.

What About Zip Cord Wire as the Radiator?

For years, amateurs have been using ordinary copper house wire as the radiator section of an antenna, erecting it without bothering to strip the plastic insulation. Other than the loading effects of the insulation mentioned earlier, no noticeable change in performance has been noted with the insulation present. And the insulation does offer a measure of protection against the weather. These same statements can be applied to single conductors of zip cord.

The situation in a radiating wire covered with insulation is not quite the same as in two parallel conductors, where there may be a leaky dielectric path between the two conductors. In the parallel line, it is this current leakage which contributes to line losses. The current flowing through the insulation on a single radiating wire is quite small by comparison, and so as a radiator the efficiency is high.

Now back to the original question: How efficient *is* a zip-cord antenna system? Well, that *does* depend, on the length of the wire used for the feed-line section and on the frequency. In a pinch on 160, 80, 40 and perhaps 20 meters, communications can certainly be established with this kind of antenna. For higher frequencies, especially with long line lengths for the feeder, you're on your own.

---

[1] Additional losses over those charted in Figure 2 will occur when standing waves are present. See Gibilisco, "What Does Your SWR Cost You?" January 1979 *QST*. Trouble is, you can't use a 50- or 75-ohm SWR instrument to measure the SWR in zip-cord line accucately.

From *QST*, April 1973 (Hints & Kinks)

Piggyback Antenna For The Ten-Meter Band

Most amateurs from Novices on up use dipole antennas for the eighty- and forty-meter bands. In many cases, these two antennas were all the Novice needed since the fifteen-meter band could also be covered with the forty-meter dipole.

However, with the advent of the new ten-meter subband for Novices, a simple and inexpensive method of getting the needed coverage is to use a "piggyback" antenna. It can be easily added to an existing dipole, eliminating the need for additional supports or feed lines. Also, it will have no effect on either an eighty- or forty-meter dipole.

Of course, performance will not be as good as that obtained with a beam, but it should not be sold short either. A KH6 was worked during a recent contest with a power output of 1.8 watts. — *Robert M. May, II, WA4DEG*

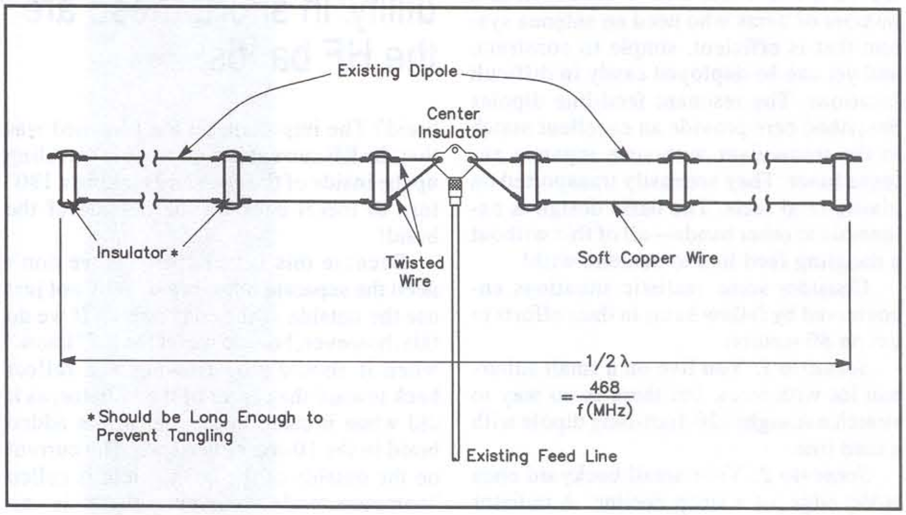
Existing Dipole
Center Insulator
Insulator *
Twisted Wire
Soft Copper Wire
1/2 λ
$= \frac{468}{f(\text{MHz})}$
* Should be Long Enough to Prevent Tangling
Existing Feed Line

By James E. Taylor, W2OZH

From *QST*, August 1991

RFD-1 and RFD-2: Resonant Feed-Line Dipoles

This unique design offers simplicity and general utility. In short, these are "reel" great antennas for the HF bands.

During several decades of operation on the 80-meter band, I have encountered a number of hams who need an antenna system that is efficient, simple to construct, and yet can be deployed easily in difficult locations. The resonant feed-line dipoles described here provide an excellent match to the transceiver without a separate antenna tuner. They are easily transported on plastic cord reels. The basic design is extensible to other bands—all of this without a dangling feed line to contend with!

Consider some realistic situations encountered by fellow hams in their efforts to get on 80 meters:

Scenario 1: You live on a small suburban lot with trees, but there is no way to stretch a straight 120-foot-long dipole with a feed line.

Scenario 2: Your small backyard ends at the edge of a steep decline. A radiator wire could be snaked through the trees, but running a separate feed line is out of the question.

Scenario 3: You live in a high-rise building and could secretly stretch a wire upward or downward, but there is no way to install a feed line.

Scenario 4: You're going on vacation to the mountains and you know that your cottage is nestled among hundred-foot-high pines You can take along a bow and arrow and cord, but then what would you do?

In response to these problems, I began to search for a universal solution under the following ground rules. (1) No separate antenna tuner would be required, (2) the antenna could be deployed with no more difficulty than stringing up a length of coax and (3) the antenna could be easily stored and unwound from a cord reel without a tangle of cable and wire.

Several years ago I used a 10-meter vertical linear coaxial sleeve antenna, as shown in Figure 1. A vertical dipole is constructed from a quarter-wave whip and a quarter wavelength of shielding braid. Its feed line passes through the braid, yielding a simplified geometry. Although this concept could be adapted to 80 meters, who wants to deal with 60 feet of shielding braid? The important lesson I learned was that the RF current has no trouble traveling up the inside of the coax and making a 180° turn to travel back on the outside of the braid!

Because this is true, perhaps we don't need the separate outer braid. Why not just use the outside of the coax itself? If we do this, however, how do we let the RF "know" when it should stop flowing and reflect back toward the center of the radiator, as it did when it came to the end of the added braid in the 10-meter vertical? The current on the outside of the coax shield is called "common-mode" current—there is no counteracting equal and opposite current as there is inside the coax. One design approach utilizes the primary function of a balun transformer: to place in unbalanced reactance in the path of this common-mode current without affecting the desired balanced transmission line currents. The development of this concept is discussed in the Appendix.

The common-mode current on the coax shield is transformed by a quarter-wavelength stub to maximum near the transceiver. It was not surprising to note experimentally that a coil of a few turns placed at the fed end of the dipole decreases the resonant frequency of the system substantially (because of the added inductive reactance). However, an unexpected dividend of this is that, at the resonant frequency, an almost perfect impedance match to the 50-ohm source is realized! Now we have the design for our simplified antenna system! We can increase the resonant frequency by moving the coil along the cable away from the current maximum while retaining the perfect match. And at the same time we can decrease the common-mode current on this part of the line, because of the coil's inductive reactance. For the values chosen, the coil is near self-resonance from the distributed capacitance of the coil windings. The equivalent parallel-resonant circuit serves to increase the reactance at this point in the antenna, which assures a reflection when the RF reaches this virtual end of the dipole. To my knowledge, this configuration and method of resonating is unique and novel, and I refer to it as a T choke.

Constructon, Installation and Adjustment

The simple arrangement of the resonant coaxial linear dipole is shown in Figure 2. The dimensions and test results are for a nominal

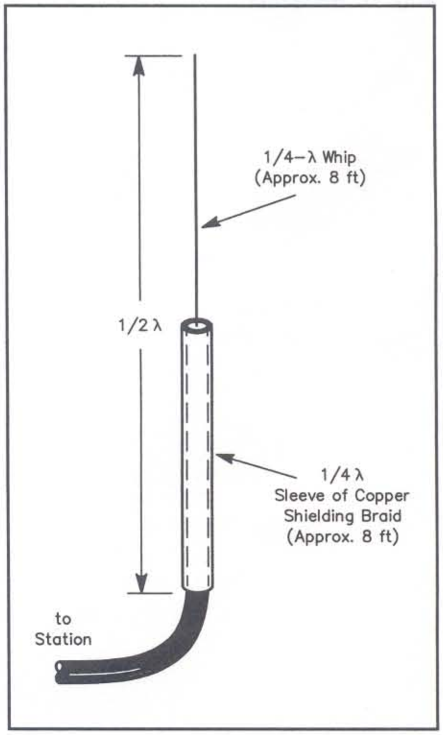

Figure 1—A vertical linear coaxial sleeve antenna for 10 meters.

frequency of 3.95 MHz. To dramatize the simplicity of the antenna, I list all required parts in Table 1.

The coaxial cable connector is assembled at the input end of the coax for connecting to the transceiver. The center conductor of the far end of the cable is connected to the antenna wire to form a hook and eye, and is securely soldered. This junction is potted in the center of a short length of PVC pipe to form a robust center insulator assembly. At the outset, 13 turns of cable can be wound on the cord reel at a point approximately 59 feet plus 25 inches from the center of the dipole. The 13 turns can then be taped in place on the reel with duct tape or equivalent. The remaining coax and wire can now be wound on the reel for ease of transporting. To avoid kinks when unwinding, secure the end of the wire and rotate the reel, keeping the wire and coax taut.

Ground losses are a great enemy of an HF antenna, especially at the lower frequencies, so place the dipole for 80 meters as far above the ground as possible! After installation, check the resonant frequency with a noise bridge or an SWR meter. If these construction details are followed, the resonant frequency should be approximately 3.95 MHz. If the resonance indication is indefinite or if the resistance is not close to 50 ohms, adjust the self-resonance of the coil by moving the turns slightly on the reel. This alters the inter-turn capacitance, permitting adjustment of the reflection of RF at the end of the dipole. During this adjustment, remember that the greatest effect is between the input and output turns, where the voltage difference is greatest. If you want to lower the resonant frequency of the antenna, remove the tape and rotate the reel to increase the 25-inch distance, thereby increasing the length of the dipole. This is done in a manner to retain the 13-turn coil—you are winding and unwinding equal lengths of cable. When the desired resonant frequency is attained, you are ready to operate! This is a broadband antenna, and not much adjustment is required.

The 13 turns is mentioned only as a nominal coil size. I have used both 11-turn and 13-turn coils. Self-resonance can be adjusted as mentioned earlier. As the number of turns is increased, the initial 25-inch distance will be decreased. For example, in my test installation a change from 11 turns to 13 turns altered this distance required for resonance from 68 inches to 25 inches. A greater impedance at this point serves to improve the isolation of the dipole from ground.

The RFD-2 for Two-Band Operation

The preceding construction information is for an antenna dedicated to a single amateur band, the RFD-1. With a slightly different configuration, referred to as the RFD-2, the RFD-1 design can be extended to cover the 40-meter band without severing the coaxial cable. Conversion from 80 to 40 meters and back requires only a few minutes. The RFD-2 design allows you to change operation to 40 meters for a few days without permanently altering the 80-meter lengths. In other words, the total dimensions of the 80-meter coaxial feedline dipole are retained but operation is adapted to 40 meters. Tbis is done simply by winding coaxial cable on reels. This is readily achieved by altering the winding of the coax on the original reel and then adding a second coil near the far end of the dipole. Once the values have been established, this band change is accomplished simply.

Table 1
RFD-1 Parts List

- 1–118-foot length RG-8X (minifoam) coaxial cable
- 1–59-foot length no. 12 stranded or Copperweld wire
- 1–PL-259 male coaxial cable connector
- 1–10-inch-diameter cord reel (Doscocil model no. 32500 or equiv.)
- 1–3-inch length of 1/2-inch OD PVC pipe (for center insulator)
- 1–Epoxy potting compound (sufficient to fill pipe)

The RFD-2, as it has evolved at W2OZH, is shown in Figure 3. Total dimensions are given in Figure 3A. The length of the feed line has been increased from that of the RFD-1 to 143 feet to make allowance for the cable used in the 13-turn coil. This assures an isolating stub which is approximately 1/4 λ on 80 meters, and at the same time permits a stub length that is close to 1/4 λ on 40 meters, thereby achieving the desired high impedance for isolation on both bands. Also, the no. 12 terminating antenna wire has been replaced by a length of coaxial cable—the outer jacket provides insulation so the coax can now be coiled to provide the desired dipole length on 40 meters.

As pointed out earlier, a self-resonant coil is used to assure a high coefficient of reflection at the fed end of the 80-meter dipole. For other bands it is necessary to calculate the approximate number of turns required to approach self-resonance. A suitable winding for 80 meters was determined to be 13 turns on a 6-inch diameter reel, so the numbers for other bands can be obtained by a simple scaling calculation. Details are given in the Appendix. For guidance in adapting the design for the other HF bands, the approximate number of turns for the terminating coils is shown in Table 2.

For wavelengths shorter than 40 meters it may be desirable to use a radiating section longer then 1/2 λ. The length either side of center can be any odd number of quarter wavelengths. For example, on 20 meters the radiator could be 3/2 λ (three

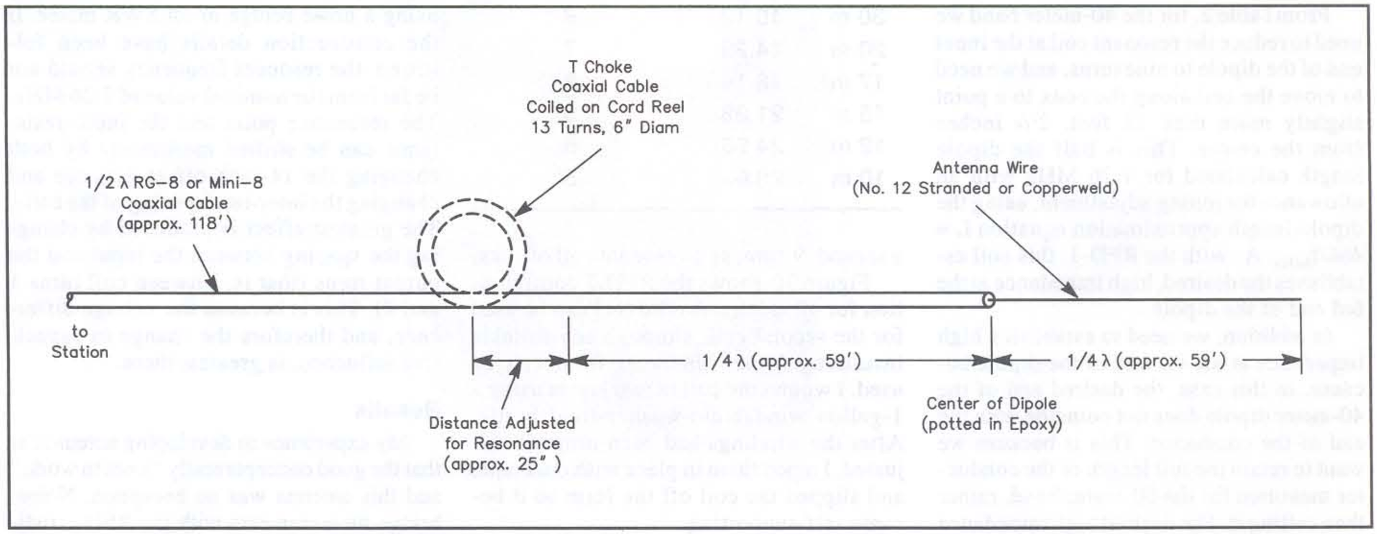

Figure 2—The RFD-1 (resonant feed-line dipole) antenna for 80 meters.

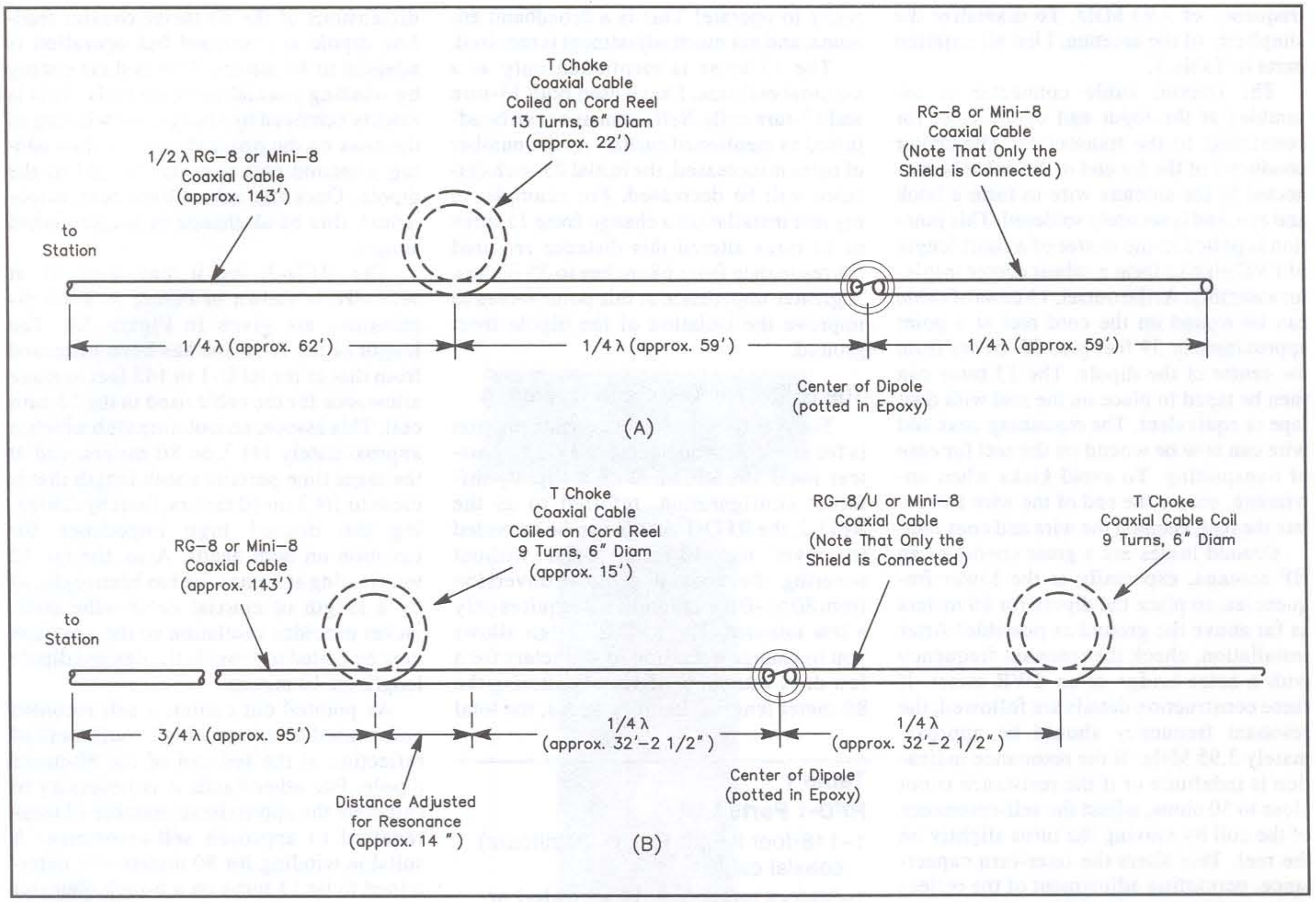
T Choke
Coaxial Cable
Coiled on Cord Reel
13 Turns, 6" Diam
(approx. 22')
1/2 λ RG-8 or Mini-8
Coaxial Cable
(approx. 143')
to
Station
RG-8 or Mini-8
Coaxial Cable
(Note That Only the
Shield is Connected)
1/4 λ (approx. 62')
1/4 λ (approx. 59')
1/4 λ (approx. 59')
Center of Dipole
(potted in Epoxy)
(A)
T Choke
Coaxial Cable
Coiled on Cord Reel
9 Turns, 6" Diam
(approx. 15')
RG-8 or Mini-8
Coaxial Cable
(approx. 143')
RG-8/U or Mini-8
Coaxial Cable
(Note That Only the
Shield is Connected)
T Choke
Coaxial Cable Coil
9 Turns, 6" Diam
to
Station
3/4 λ (approx. 95')
1/4 λ
(approx. 32'-2 1/2")
1/4 λ
(approx. 32'-2 1/2")
Center of Dipole
(potted in Epoxy)
Distance Adjusted
for Resonance
(approx. 14")
(B)

Figure 3—The RFD-2 antenna can be configured for either of two bands. Shown at A is the setup for 80 meters, and at B for 40 meters.

quarter waves either side of center). Depending upon the antenna orientation, the additional radiation lobes may be advantageous. Also, the unused conductor lengths at the ends are less apt to cause trouble because of parasitic excitation from the main radiator.

Arrangement and Construction

From Table 2, for the 40-meter band we need to reduce the resonant coil at the input end of the dipole to nine turns, and we need to move the coil along the coax to a point slightly more than 32 feet, 2¾ inches from the center. This is half the dipole length calculated for 7.26 MHz with an allowance for tuning adjustment, using the dipole-length approximation equation $L = 468/f_{MHz}$. As with the RFD-1, this coil establishes the desired, high impedance at the fed end of the dipole.

In addition, we need to establish a high impedance at the far end of the dipole because, in this case, the desired end of the 40-meter dipole does not coincide with the end of the conductor. This is because we want to retain the full length of the conductor measured for the 80-meter band, rather than cutting it. The desired high impedance reflecting termination is achieved by use of a second, 9-turn, self-resonant coil of coax.

Table 2
Scaling of RFD-1 to Other Bands

Band	*Frequency*	*Turns*
160 m	1.90	19
80 m	3.95	13
40 m	7.26	9
30 m	10.12	8
20 m	14.29	7
17 m	18.14	6
15 m	21.38	6
12 m	24.96	5
10 m	28.65	5

Figure 3B shows the RFD-2 configuration for 40 meters. A cord reel can be used for the second coil, although any suitable insulating 6-inch diameter form can be used. I wound the coil in two layers using a 1-gallon windshield-washer-fluid bottle. After the windings had been properly adjusted, I taped them in place with duct tape, and slipped the coil off the form so it became self-supporting.

As with the RFD-1, it is desirable to place the antenna as high above ground as possible to reduce ground losses and to belp achieve low-angle radiation, if that is desired. The dipole does not have to be installed in a straight line, and if you can gain appreciably greater height by bending it, this may be a desirable compromise.

After the antenna with its two coils is erected, measure the resonant frequency using a noise bridge or an SWR meter. If the construction details have been followed, the resonant frequency should not be far from the nominal value of 7.26 MHz. The resonance point and the input resistance can be shifted moderately by both changing the 14-inch offset distance and changing the inter-turn spacing of the coils. The greatest effect is produced by changing the spacing between the input and the output turns (that is, between coil turns 1 and 9). This is because the voltage difference, and therefore the change in capacitive influence, is greatest there.

Results

My experience in developing antennas is that the good concepts really "want to work," and this antenna was no exception. Noise-bridge measurements with the RFD-1 indicated an input resistance at resonance of

49 ohms, and from a practical standpoint the SWR is 1:1. An H-field antenna probe was used to evaluate the power radiated at the center of the dipole compared with that at the current loop near the feed-point. The current ratio was 5.5 to 1, which corresponds to a power ratio of 30 to 1, or 15 decibels. This indicates that the coil is very effective in attenuating the common-mode current flowing back toward the feed point.

For the RFD-2 with the 40-meter dimensions shown in Figure 3B, the initial resonance point was within the 40-meter phone band and the SWR was essentially 1:1. Slight adjustment brought the resonant frequency up to the desired value, and the input resistance was very close to 50 ohms. A salient characteristic of the RFD antennas is the ease with which an impedance match is attained. It appears always to be easy to get a reflected power indication of zero. This is probably because the common-mode current on the shield of the coax is indistinguishable from the desired radiating current—in other words, the common mode is used rather than avoided.

Appendix

When the separate outer braid of the antenna shown in Figure 1 is removed, there must be some way to let the RF "know" when it should stop flowing on the shield of the coax and reflect back toward the center of the radiator. A balun box that I made up for another purpose is one approach: 30 turns of bifilar winding on an Amidon T-200-2 iron-powder toroidal core. The turns formula for such a balun transformer is

$$T=\sqrt{\frac{\text{desired L}(\mu\text{H})}{A_L}}$$

where

T = no. of turns

A_L = inductance index (microhenries per 100 turns)

From this, the inductive reactance for unbalanced current is

$$X_L=2\pi fL=\frac{2\pi fT^2A_L}{10^4}$$

At a frequency of 4 MHz, this 30-turn coil has an unbalanced reactance of only 270 ohms. We would need about ten of these in series to support the RF field at the end of a dipole antenna!

An alternative method involves placing a toroidal isolation transformer in the coaxial line at the fed end of the linear dipole. I actually wound such a transformer on two stacked T-200-2 cores, to provide a 1-kW power capability. This configuration worked, after a fashion, but the impedance match was less than desired, probably because of excessive capacitance between secondary and primary windings. However, before I took steps to control this, a simpler approach came to mind.

We need to isolate the transceiver from the fed end of the dipole, so why not cut the coaxial feed line to be a quarter wavelength long? (This length is measured in free-space, not in the line.) This serves to tranform the high impedance of the fed end of the dipole to the low impedance at the grounded transceiver. Any unused portion of this approximate 60-foot length of coax can be wound in a coil and used for further isolation and, as it turns out, for tuning the system to resonance.

Scaling for a Second Band

For coverage of a second band, we would like to have the same inductive reactance that we had for the 80-meter coil.

From handbooks, the inductance of a coil of assumed dimensions is

$$L = AN^2$$

where

A = a constant determined by the coil geometry

N = number of turns

Thus, we have for the reactance

$$X_L = 2\pi fL = 2\pi fAN^2$$

For equal reactances we can calculate N_2, the new number of turns, at frequency f_2, from N_1, the known number of turns at frequency f_1, by the equation

$$N_2=N_1\sqrt{\frac{f_1}{f_2}}$$

By Jerry Hall, K1PLP **From *QST*, September 1974**

Off-Center-Loaded Dipole Antennas

In these times when much of our amateur population lives in urban areas, the subject of shortened antennas for the lower frequency amateur bands is a very popular one. Physically short ground-mounted vertical antennas with lumped-constant loading to make them resonant can be quite efficient radiators, if a good radial system has been installed. This has certainly been evidenced in Sevick's series of recent *QST* articles.[1] To many amateurs, however, the "hitch" in constructing such a system is the installation of a *good* radial system. It must be admitted that for the "top" amateur bands, 160 and 80/75 meters, an efficient system of buried radials requires a sizable amount of real estate, even for a physically short radiator. On the average city-size lot, 50 or 75 by 120 to 150 feet, it's almost impossible to install a highly efficient radial system for 80/75 meters, much less for 160 meters, when structures like a house and perhaps a separate garage exist. Or to some amateurs, just the thought of burying hundreds or maybe thousands of feet of wire is enough to turn off any enthusiasm for the project. What's the alternative? A dipole type of antenna with lumped constant loading. At modest heights, 30 or 40 feet, such an antenna will prove to be quite satisfactory if it is physically longer than about 0.2 wavelength. Shorter lengths may also be used, at reduced efficiency. Such an antenna can be fed directly with 50-Ω coaxial line, and it can be operated with no earth ground. (Of course the chassis of the transmitter and/or receiver should be grounded adequately for protection against shock hazard.)

Nearly all of us are familiar with the concept behind the use of inductive loading. A vertical antenna which is shorter than a quarter wave (or a dipole antenna which is shorter than a half wave) will exhibit capacitive reactance at its base (or center) feed point. To cancel such capacitive reactance, a coil having the proper inductive reactance may be connected in series with the base feed point of the vertical. The same result will be obtained through the use of two such coils for a dipole, one coil connected in series with each half. It is not necessary for the inductor to be installed at the feed point, however. In fact, greater radiating efficiency results through improved current distribution if the inductor is located along the radiator some distance away from the low-impedance feed point, *viz*, in the manner of a center-loaded mobile whip antenna. Figure 1 shows this concept extended to a dipole element, with off-center loading. The inductors resonate the antenna to the operating frequency, but do little actual radiating themselves. (This is in contrast to helically wound or continuously loaded elements, where a long thin inductor is the radiator as well as the loading element.)

In the antenna represented by Figure 1, there are many variable factors to be considered when a practical antenna for a given frequency is being constructed. Of primary consideration from an efficiency standpoint is the overall length, shown as dimension A. Another consideration for efficiency is the distance of the coils from center, dimension B. The longer the overall length (A), up to a half wave, and the farther the loading coils are placed from the center (B), the greater is the efficiency of the antenna. However, the greater is distance B (for a fixed overall antenna size), the larger the inductors must be to maintain resonance. Theoretically, if the coils were placed at the outer ends of the dipole, they should be infinite in value to maintain resonance. Capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce this requirement to a more practical value.

What Inductance Values?

As a matter of personal interest, this writer has been doing experimental work for a number of years with off-center-loaded antennas. One big drawback to such experimentation was the ever-present need for a large amount of cut-and-try work to arrive at resonance whenever a new set of dimensions was to be used. Probably the number of pruned-off turns from coil stock from such experiments, if straightened out and soldered end to end, would make up several full-sized half-wave antennas for the 160-meter band. Therefore, most of the writer's work of late in this area has been in going through paperwork exercises, looking for a way whereby at least "ball-park" values of inductance needed for a particular system could be calculated.

The equation contained in the Mobile chapter of *The ARRL Antenna Book* for determining the capacitance of a vertical antenna shorter than a quarter wavelength looked promising in early computations,

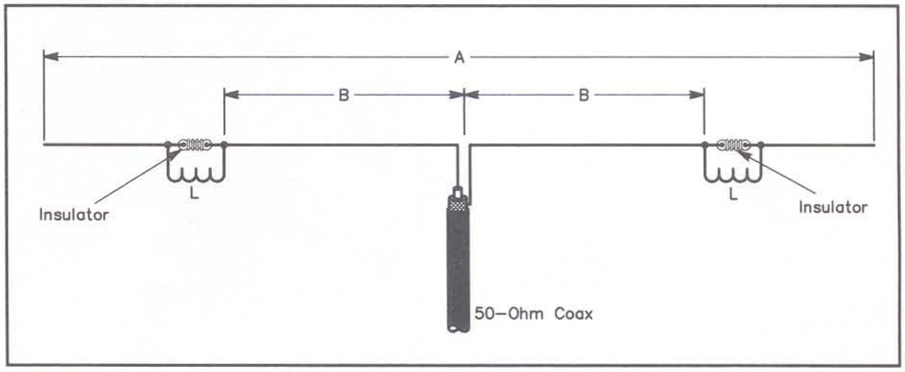

Figure 1—A dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance.

and, indeed, it became the basis for the calculation procedure which finally resulted.

This procedure has been found to produce results much closer than mere "ball-park" values for the necessary inductance – for wire antennas "in the clear" at moderate heights, the final inductance values found by cut-and-try pruning for lowest SWR at the desired frequency have been so close to the value from calculations that a laboratory bridge was necessary to measure the difference. The results are equally good for elements using tubing. Once the needed inductance value is determined by calculations, it is generally found sufficient to obtain coil dimensions from an ARRL L/C/F Calculator (see LaPlaca*) or by equation. Any significant pruning which has been found necessary could always be attributed to objects in proximity to the ends of the antenna.

The complete set of calculations is expressed in the mathematical relationship below as Eq. 1, presented here primarily for mathematics buffs or those having access to electronic computers.

This equation yields the inductance required, in microhenrys, for single-band resonance of a shortened antenna of a particular physical size at a given frequency, for a specific position of the loading coils from the center of the antenna. To spare the reader the task of performing some rather tedious calculations, Figure 2 has been prepared from Eq. 1. The curves of the chart have been normalized, and may be used for any frequency of resonance. The chart is based on a halfwavelength/diameter ratio of the radiator of approximately 24,000. (This corresponds to No. 14 wire on 80 meters or No. 8 wire on 160 meters.) For "thinner" conductors, the required inductance will be somewhat greater than that determined from Figure 2, and less inductance will be required for "thicker" conductors.

The use of the chart is as follows: At the intersection of the appropriate curve from the body for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimensions A and B are shown in Figure 1, and for use with the chart are expressed as percentages. Dimension A is taken as percent length of the shortened antenna with respect to the length of a resonant half-wave dipole of the same conductor material. Dimension B is taken as the percent of coil distance from the feed point to the end of the shortened antenna. For example, resonating an antenna which is 50% or half the size of a half-wave dipole (one-quarter wavelength overall), with loading coils positioned midway between the feed point and each end (50% out), would require loading coils having an inductive reactance of approximately 950 ohms at the operating frequency. If the antenna is hung "in the clear," and if the length/diameter ratio of the conductor is near 24,000, inductance values as determined from the chart will be very close to actual values required. (Eq. 1 above takes the diameter of the radiator into account, and thus may be used for any length/diameter ratio.) For practical purposes, dimension B may be taken as that distance from the center of the feed-point insulator to the inside eye of the loading-coil insulator, and dimension A as the eye-to-eye distance inside the end insulators (which are not drawn in Figure 1).

Proximity of surrounding objects in individual installations may require some pruning of the coils, and the exact amount of final inductance required should be determined experimentally. If the antenna is hung in inverted-V style, with the ends brought near the earth, the required inductance will almost always be somewhat less than that determined from the chart or equation. A grid-dip meter, Macromatcher, (see Hall and Kaufmann*) or SWR indicator may be used, during the final adjustment procedure.

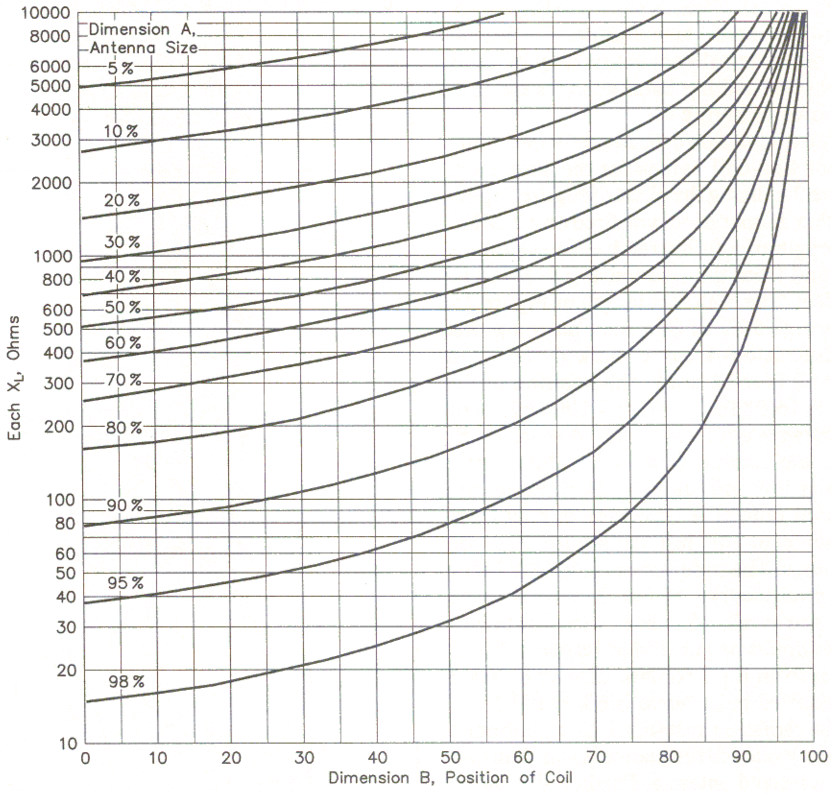

Figure 2—Chart for determining approximate inductance values for off-center-loaded dipoles. At the intersection of the appropriate curve from the body of the chart for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at left. See Figure 1 regarding dimensions A and B.

(Eq. 1):

$$L_{\mu H} = \frac{10^6}{68\pi^2 f^2}\left\{ \frac{\left[ln\frac{24\left(\frac{234}{f}-B\right)}{D} -1\right]\left[\left(1-\frac{fB}{234}\right)^2-1\right]}{\frac{234}{f}-B} - \frac{\left[ln\frac{24\left(\frac{A}{2}-B\right)}{D}-1\right]\left[\left(\frac{\frac{fA}{2}-fB}{234}\right)^2-1\right]}{\frac{A}{2}-B}\right\}$$

where
$L_{\mu H}$ = inductance required for resonance
ln = natural log
f = frequency, megahertz
A = overall antenna length, feet
B = distance from center to each loading coil, feet
D = diameter of radiator, inches

Practical Antennas

Although one might erect an inductively loaded antenna that is cut for a single amateur band, it is possible to use the antenna itself for two, three, or more bands of operation, if provision is made to lower the antenna for band changes. A simple rope halyard and pulley arrangement at one of the supports will do the trick. Figure 3A shows a 3-band antenna of this nature, for 160, 80, and 20 meters. If the insulators shown are left open, with nothing bridging them, the antenna is a simple half-wave dipole cut for 14.18 MHz. (The 48.5-foot lengths act merely as support wires, and have negligible effect on operation of the antenna.) If the insulators are bridged with short lengths of antenna wire, the antenna becomes a center-fed 80-meter dipole, resonant at about 3.6 MHz. For 160-m operation the 20-meter insulators may be bridged with loading coils to resonate the antenna at 1.8 MHz, as shown in Figure 3A. Burndy or other manufacturers' "Servit" type of electrical connectors may be used, for ease in making band changes quickly, as shown in Figure 4.

The calculation procedure for determining loading-coil values for the antenna of Figure 3A, using the chart of Figure 2, goes like this. If operation is desired on 1.8 MHz, the length of a full-sized half-wave dipole is found from the relationship 468/*f* to be 260 feet. The 130-foot length of Figure 3A represents 50% of this size, meaning that the dimension-A curve marked "50%" in Figure 2 is to be used. The position of the coils is 16.5/(16.5 + 48.5) × 100 or 25% of the distance out from center, dimension B. From the intersection of 25 (horizontal scale at bottom), and the 50% curve, the required inductive reactance is read from the scale at the left of Figure 2 to be 650 ohms. The inductance, *L*, is $650/2\pi f$ or 57.5 microhenrys, if No. 8 wire is to be used. For smaller diameter wire, the inductance should be somewhat larger. (Calculations from Eq. 1 for No. 12 wire indicate the required inductance is 60.99 µH.)

The radiation resistance of a shortened antenna loaded to resonance is less than that of a full-sized antenna. Further, the shortened antenna is "sharper," meaning that the change in reactance versus frequency is greater. In other words, the shortened antenna acts as a tuned circuit having a higher *Q* than a full-sized antenna. To check these characteristics, the line input impedances for the antenna of Figure 3A were measured with a laboratory bridge, and the electrical line length at the measurement frequency was then taken into account to determine the impedance at the antenna feed point. The antenna was constructed of No. 12 wire and hung at a height of 50 feet as a "flat-top" radiator.

The solid curve of Figure 3B is a plot of the feed-point impedance versus frequency for this antenna. The plot on Smith Chart coordinates is more meaningful than a simple SWR-vs.-frequency curve because the magnitudes of the resistive and reactive components are shown, as well as the sign of the reactance. (Capacitive reactance is negative, plotted to the left of the vertical center line, and inductive reactance is positive, plotted to the right.) In this presentation, a 50-ohm nonreactive impedance will appear at the exact center of the chart. The SWR in 50-ohm line for a given frequency may be determined by first noting the distance from the center of the chart to the particular impedance plot on the curve, and next measuring this same distance down the vertical center line from chart center (a drawing compass is helpful for this task), and finally dividing 50 into the value read at that point on the center line. For example, the SWR at 1.8 MHz equals 120/50 or 2.4, as indicated by the segment of the 2.4 SWR circle in Figure 3B. It may be seen that resonance (zero reactance) occurs at approximately 1810 kHz, where the resistance is about 22 ohms. The SWR at resonance is 2.33:1, and climbs to 3:1 at 1825 kHz. At 1850 kHz, the SWR is 10:1. Without any matching provisions the antenna is relatively sharp, as mentioned earlier. If one sets the usable bandwidth as the frequency range where the SWR is 3:1 or less, it is approximately 35 kHz, or 1.9% of the resonant frequency. As far as efficiency is concerned, ohmic losses are low, and the antenna is a good performer on 160 meters. Because of its horizontal polarization, it has proved to be most effective at night, and stations several hundred miles away have been worked with S-9 reports received for the 50-watt signal.

For a comparison of impedances, the broken curve of Figure 3B is a plot of measured impedances of a full-size half-wave dipole, 260 feet long overall, hung in place of the shortened antenna. From this curve it may be seen that resonance occurs at 1810 kHz, where the resistance is 59 ohms. The 3:1-SWR bandwidth for the half-wave antenna is in the order of 60 kHz, or 3.3% of the reso-

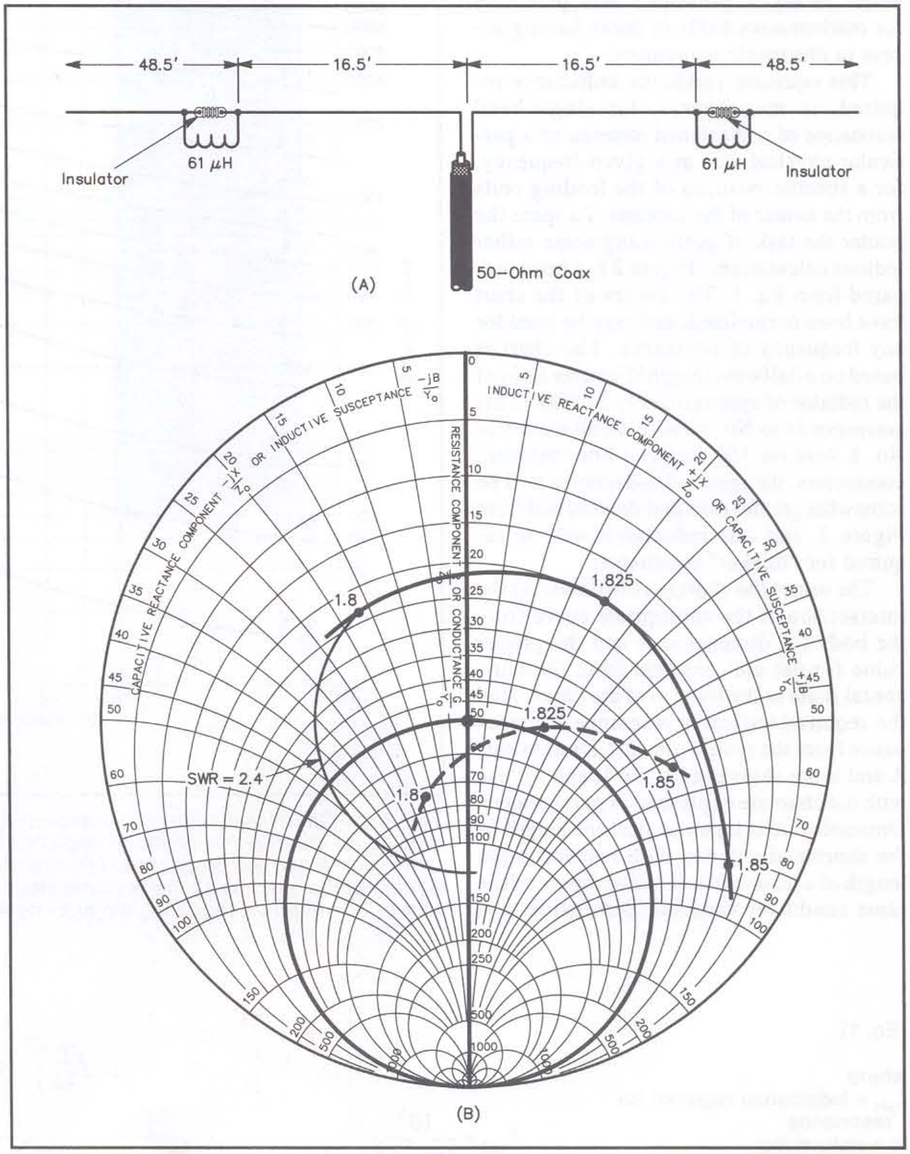

Figure 3—At A, an 80-meter dipole loaded for 160-meter operation. The inductors are 37 turns of coil stock having a 3-inch diameter and 10 turns per inch (B&W 3035). At B, the impedance plot of this antenna installed at a height of 50 feet (solid curve), and the plot of a 160-m half-wave dipole (broken curve).

nant frequency. It is interesting to note on this curve that the SWR at resonance is 1.18:1, and that it is a somewhat lower value, 1.15:1, at a frequency a few kilohertz above resonance. (Measurements were made every 5 kHz across this band, but plot points are shown only for 25-kHz increments to avoid crowding of the data.) This evidence refutes the oft-heard statement that the SWR-vs.-frequency curve is *always* lowest at antenna resonance. Points to remember are that the SWR in a transmission line is completely dependent upon the characteristic-impedance value of the line in use. Using a line of different impedance may shift the position of the SWR curve along the frequency axis in a simple SWR-vs.-frequency plot. This is definitely true in this case — if the 160-meter half-wave dipole were to be fed with 75-ohm line, the lowest SWR would occur at a frequency about 5 kHz *below* antenna resonance, whereas with 50-ohm line the lowest SWR is at a frequency slightly *above* resonance. The reason this happens is that the resistive component of the impedance, which consists of the radiation resistance plus any loss resistance, is not constant with frequency, even over a rather narrow frequency range. It must be acknowledged that the differences here are very slight, however, and for practical purposes the frequency of lowest SWR is (within a few kilohertz) the resonant frequency of the antenna.

Another point concerning the SWR values bears noting. The values as determined from the plots in the manner described above are quite accurate, having been determined by measurements with laboratory equipment. In contrast, measurements with simple SWR indicators usually cannot be relied upon for anywhere near the equivalent accuracy.

For example, the author owns a commercially manufactured SWR indicator of the Monimatch type (see McCoy*) which, under a particular set of conditions, indicates a 2.5:1 SWR in a line where laboratory measuring equipment shows the *true SWR to be 4:1*. A significant difference! Herein lies another reason why impedance plots on Smith Chart coordinates are more meaningful than a simple SWR-vs.-frequency curve — greater accuracy may generally be expected.

A Half-Size 80-Meter Antenna

Figure 5A shows the 3-band concept described earlier as it can be applied to 80, 40, and 20 meters. Its overall length is 66 feet, not a difficult length to use on a small lot. This antenna was constructed for 80-m operation with a design-center frequency of 3.55 MHz, using No. 12 antenna wire and 40-μH loading coils—27 turns of stock having a diameter of 3 inches and a pitch of 10 turns per inch (tpi). Feed-point impedances versus 80-m frequency for the antenna, hung at a height of 50 feet, are shown by the solid curve at B of Figure 5. Actual resonance occurred at 3.54 MHz, where the resistance was about 26 ohms. The bandwidth within which the SWR is 3:1 is

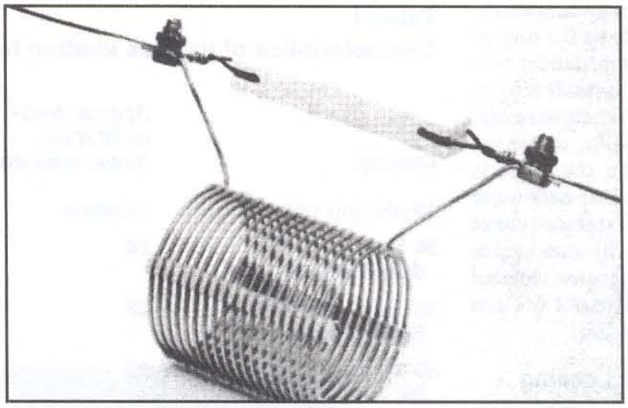

Figure 4—Copper electrical service connectors, sold under one trade name of Servit, provide a simple means of installing the loading coils. The antenna wire and the ends of the coil wires should be tinned to prevent corrosion. In addition, a protective coating of acrylic spray may be used at each connection.

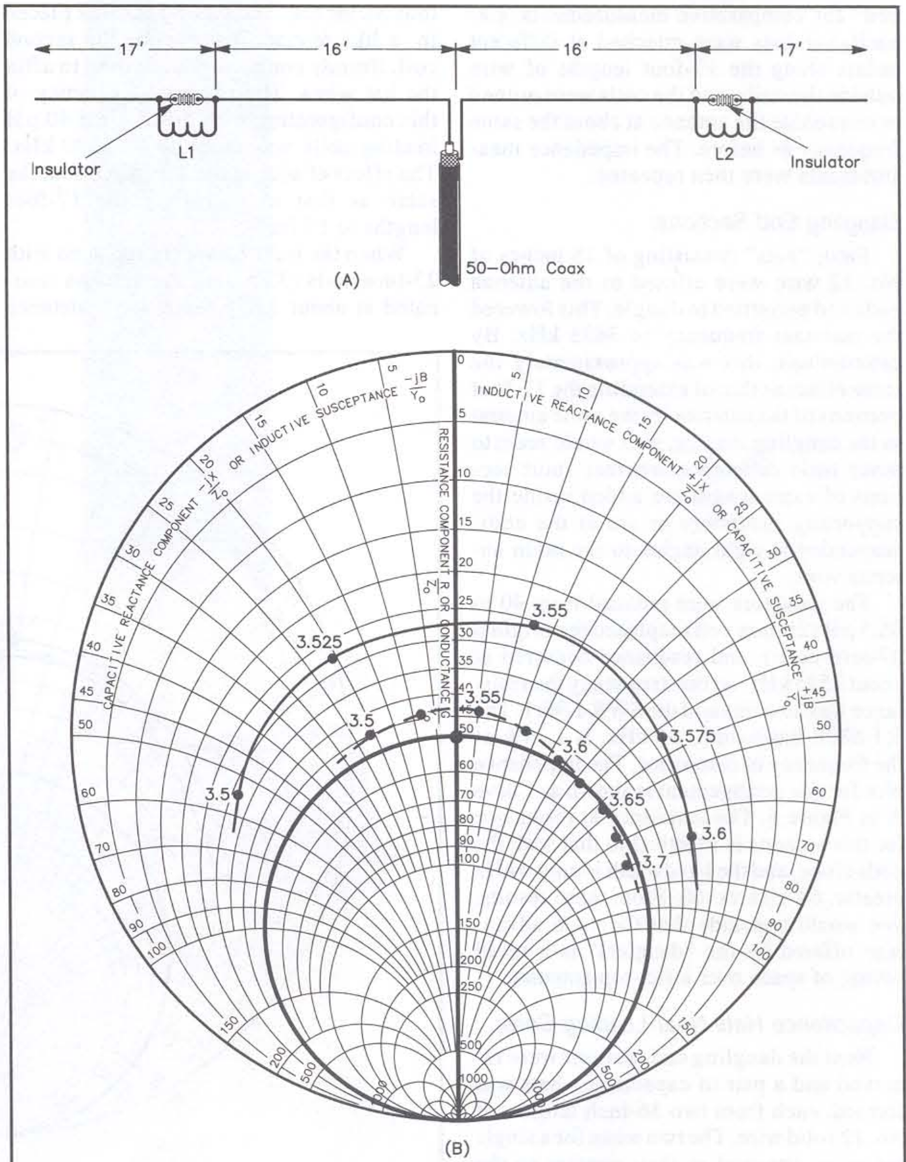
17'
16'
16'
17'
L1
L2
Insulator
Insulator
(A)
50–Ohm Coax
INDUCTIVE SUSCEPTANCE $\frac{-jB}{Y_o}$
OR INDUCTIVE REACTANCE COMPONENT $\frac{+jX}{Z_o}$
CAPACITIVE REACTANCE COMPONENT $\frac{-jX}{Z_o}$
OR CAPACITIVE SUSCEPTANCE $\frac{+jB}{Y_o}$
RESISTANCE COMPONENT R OR CONDUCTANCE G $\frac{R}{Z_o}$
3.5
3.525
3.55
3.575
3.6
3.65
3.7
(B)

Figure 5—At A, a 40-meter dipole loaded for 80-meter operation. For resonance at 3.55 MHz the coils should be approximately 40 μH (27 turns of stock); 3.75 MHz, 35 μH (24 turns); 3.9 MHz, 31 μH (22 turns); and 4.0 MHz, 29 μH (21 turns). These are circulated inductance values for No. 12 antenna wire. Coil stock referenced above is 3-inch diameter, 10 turns per inch (B&W 3035). At B, the impedance plot of the 3.55-MHz version (solid curve) and of an 80-m half-wave dipole (broken curve).

60 kHz, or 1.69% of the resonant frequency.

Also shown in Figure 5B, by the broken curve, are the feed-point impedances of a half-wave dipole, 132 feet overall length, hung in place of the shortened antenna. Resonance occurs at 3.54 MHz, where the resistance is 43.5 ohms and the SWR is 1.15:1. The broader nature of the half-wave antenna is exhibited by the "tighter" curve which swings closer to the 50-ohm center point of the chart than the shorter, loaded antenna. The SWR at 3.5 MHz is 1.6:1 and remains below 3:1 to 3.67 MHz.

Table 1
Characteristics of various loading techniques, 66-foot 80-m dipole.

Loading	*Approx. feed-point resistance, resonance*	*SWR at resonance*	*3:1-SWR bandwidth, % of resonant freq.*
40-µH coils only	26 ohms	1.92:1	1.69
36.5-µH coils, 18″ dangling ends	26	1.90:1	1.79
36″ hats outside 32.5-µH coils	23	2.15:1	1.68
30-µH coils, 36″ hats at ends	25	1.98:1	2.05
None (λ/2 dipole)	43.5	1.15	Greater than 3.6

Coil positions for each loaded antenna were 16 feet from antenna center. All antennas were constructed of No. 12 wire and installed at a height of 50 feet.

Capacitive and Inductive Loading

One would assume that a combination of capacitive and inductive loading might provide a different feed-point impedance than would inductive loading alone, because of different current distributions in the radiators. To check out this assumption, the antenna of Figure 5A was used as a "test bed" for comparative measurements. Capacitance hats were attached at different points along the 17-foot lengths of wire outside the coils, and the coils were pruned to reresonate the antenna at about the same frequency as before. The impedance measurements were then repeated.

Dangling End Sections:

First, "hats" consisting of 18 inches of No. 12 wire were affixed to the antenna ends and permitted to dangle. This lowered the resonant frequency to 3435 kHz. By calculations, this was approximately the same effect as that of extending the 17-foot portions of the antenna by the same amount as the dangling lengths, so it would seem to make little difference whether short sections of extra length are added inside the supporting insulators or are at the ends, suspended at right angles to the main antenna wire.

The inductors were reduced from 40 to 36.5 µH (25-turn coils replaced the original 27-turn coils), and resonance occurred at about 3575 kHz. At this frequency the resistance was 26 ohms and the SWR 1.90:1. The 3:1-SWR bandwidth, 64 kHz, is 1.79% of the frequency of resonance. The impedance plot for this arrangement is shown as Curve A in Figure 6. The resistance at resonance for this antenna is identical to that with the coils alone, and the bandwidth is only 4 kHz greater, 64 kHz vs. 60. From these results, one would conclude that the main advantage offered by the "danglers" is a small saving of space over a flat-top antenna.

Capacitance Hats Near Loading Coils:

Next the dangling end sections were removed and a pair of capacitance hats was formed, each from two 36-inch lengths of No. 12 solid wire. The two wires for a single hat were attached at their centers to the antenna wire at a point just outside one of the loading coils. The hat wires were then bent radially to form an X at right angles to the antenna wire, like four spokes of a wheel with the main antenna wire at the hub. The diameter of the X-shaped hat was thus 36 inches. The second hat was placed in a like manner just outside the second coil. Burndy connectors were used to affix the hat wires. The resonant frequency of this configuration with the original 40-µH loading coils was found to be 3290 kHz. The effect of adding the hats was about the same as that of extending the 17-foot lengths to 19 feet.

When the inductors were replaced with 23-turn coils (32.5 µH), the antenna resonated at about 3.575 MHz, the resistance being 23 ohms. The SWR at resonance is 2.15:1, and the 3:1-SWR bandwidth for this configuration is 60 kHz, 1.68% of the resonant frequency. The impedance of this arrangement versus frequency is shown by Curve B of Figure 6.

It is surprising to note that, by the standards of most amateurs, the characteristics of this antenna are not as good as those of the same length antenna with loading coils alone. The SWR at resonance for the antenna with combination capacitive and in-

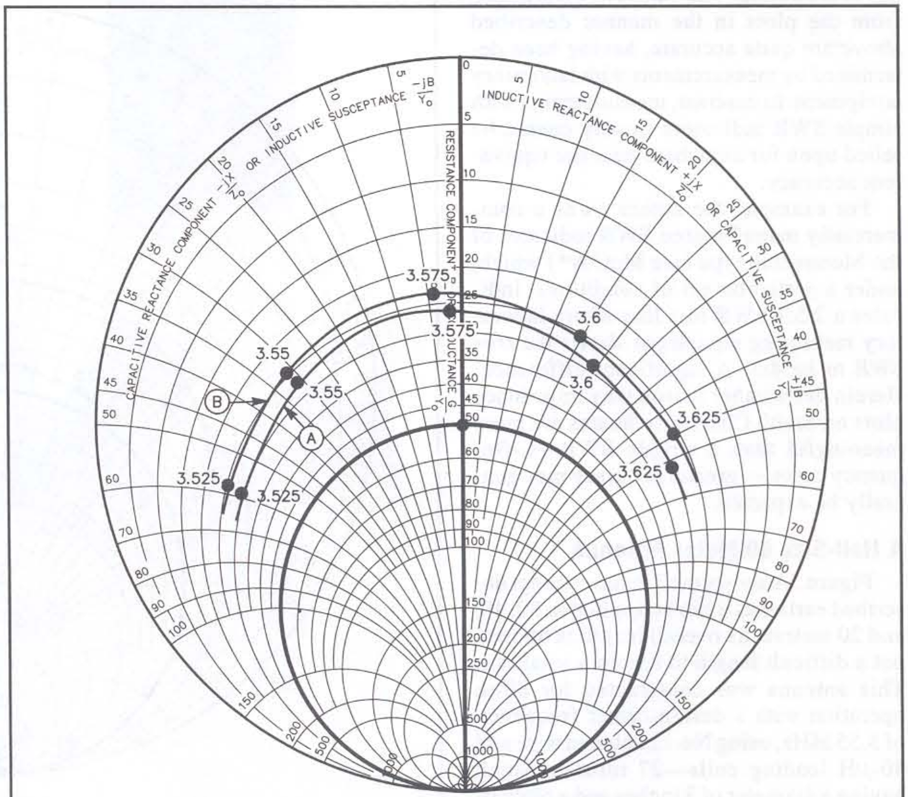

Figure 6—Curve A is the impedance plot of the antenna of Figure 5A with 1.5-foot dangling end sections added and the coil trimmed to restore resonance near the original frequency. Curve B is a plot of the same antenna with X-shaped capacitance hats added at a point just outside the loading coils (dangling sections removed and coils trimmed to reestablish resonance).

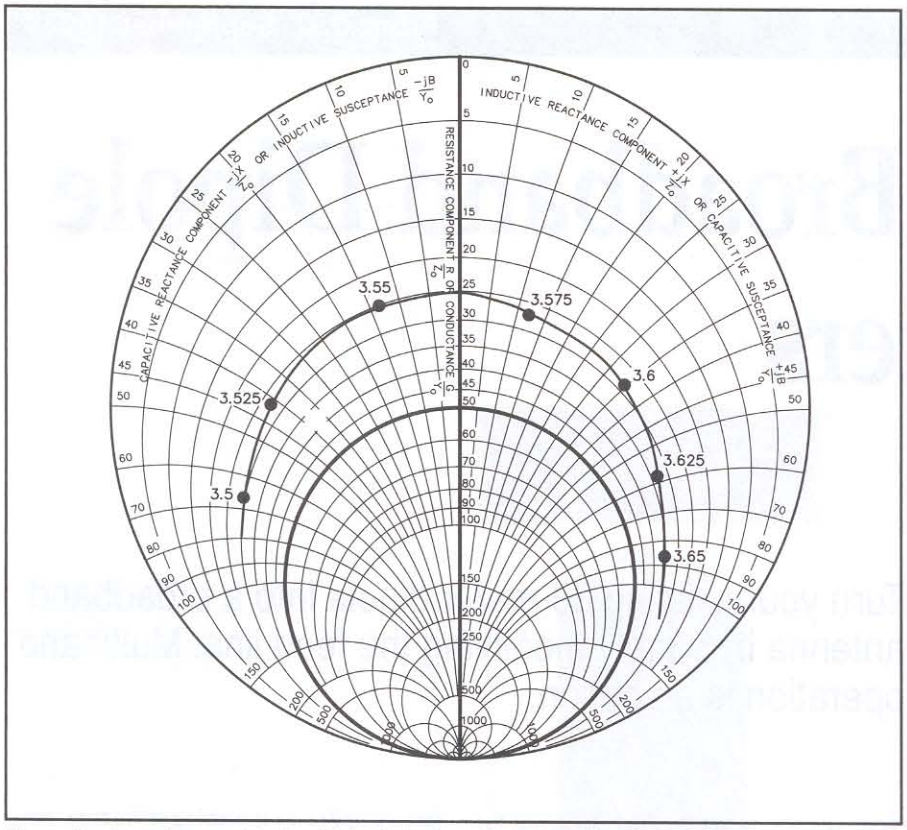

Figure 7—Impedance plot of a 66-foot dipole using a combination of off-center inductive loading and capacitive end loading. Of all the shortened configurations tried, this arrangement offered the greatest bandwidth.

ductive loading is higher (2.15 vs. 1.92), and the 3:1-SWR bandwidths are the same, 60 kHz. Perhaps a significant factor here, though, is that the diameter of the capacitance hats used for these measurements was small, only .011 wavelength. Supporting much larger hats presents mechanical problems with wire antennas, however, as even these were a bit flimsy and would require reshaping after gusty weather.

Capacitance Hats at Antenna Ends:

Finally, the X-shaped capacitance hats were moved to the outside ends of the antenna, just inside the end insulators. With the original 40-µH coils, resonance appeared at 3215 kHz. From calculations, it was as if the 17-foot end sections were actually 21 feet long. With 30-µH coils (22 turns) in place, the resonant frequency was 3560 kHz. At this frequency the resistance was 25 ohms and the SWR 1.98:1. The 3:1-SWR bandwidth is 73 kHz, or 2.05% of the resonant frequency. The impedance plot of this antenna is given in Figure 7.

It is interesting to note that the position and shape of the plot for this antenna on Smith Chart coordinates is nearly identical to that for the same length antenna with loading coils only, the solid curve of Figure 5B. For this antenna, however, the plot points for 25-kHz frequency increments appear closer together, which accounts for the increased bandwidth.

Conclusions:

The measured characteristics of these various configurations of loading for the 80-meter antennas are tabulated in Table 1. Remember that the overall "flat-top" length of each antenna arrangement, is 66 feet, and that the loading coils are always positioned 16 feet each side of the center of the antenna, being pruned for resonance at approximately 3550 kHz. For comparison, information for a half-wave dipole is also included.

Of the various arrangements, capacitive end loading decidedly provides the greatest bandwidth, excepting the full-size half-wave antenna, of course. Although there are slight differences in the resistance value at resonance, all are of the same order of magnitude. These values, as well as those for the 160-m antenna discussed earlier, tend to confirm a *broad* rule of thumb that the writer has formulated for this type of antenna: The feed-point impedance value at resonance is roughly proportional to the length of the antenna. That is, a loaded antenna which is half the size of a half-wave dipole will have approximately half the radiation resistance of the full-sized antenna.

Eq. 1 given earlier or the chart of Figure 2 allows one to calculate loading-coil values for antennas with loading coils only. Additional capacitive loading is not taken into account. Calculating the effects of various capacitive loading arrangements appears to be difficult, and work remains to be done in this area.

Multiband Antennas with Loading Coils

All of the foregoing material has been devoted to the loading of an antenna for resonance at a single frequency. Resonated as described, the antenna is electrically a half wave in length. It will, however, operate well on higher frequencies — frequencies at which it is an odd multiple of half waves, in *electrical* length . . . three half waves, five half waves, etc. Because of the lumped loading of the shortened antenna, these higher frequencies will likely not be closely related to odd-order harmonics of the fundamental frequency, as the case would be for a nonloaded radiator. (For example, it is a well-known fact that a 7-MHz half-wave dipole operates well on its third harmonic, 21 MHz.)

A loaded dipole will become an electrical 3/2-λ antenna at some frequency *below* that which is three times the fundamental resonant frequency. Depending upon the overall antenna length, coil value, and coil position, it is possible for an 80-meter loaded dipole to become a 3/2-λ performer on 40 meters. With such an arrangement, one would have a dual-band antenna without requiring the use of traps. The idea can be expanded upon to arrive at a loaded antenna without traps which will operate on more than two bands. This scheme offers considerable constructional simplification as compared with trap arrangements.

The multiband loading-coil concept has been recognized for better than half a century, but little use of the technique has been made by amateurs. Some years ago a very good article on the subject a was published by William Lattin, W4JRW.* That article is recommended reading for anyone interested in more details on the concept. Supplemental information has been published by Buchanan.* Attempts by this writer to calculate antenna sizes and coil values for dual-band antennas have met with some success. From calculations and experiments to date, it appears that with only two loading coils (one each side of center), the antenna must always be greater than a half wave in physical length for the higher of the two frequency bands. In other words: any 80/40 meter arrangement, for example, apparently would need to be longer than 66 feet from tip to tip. However, much work also remains to be done in this area.

References

Buchanan, "An Inexpensive 40- and 80-Meter Antenna," Hints and Kinks, *QST*, September, 1962, p. 62.

Hall and Kaufmann, "The Macromatcher, An RF Impedance Bridge for Coax Lines," *QST*, January, 1972, p. 14. Information also contained in the 50th and subsequent editions of *The Radio Amateur's Handbook*, and the 13th edition of *The ARRL Antenna Book*.

LaPlaca, "Using the ARRL L/C/F Calculator," *QST*, December, 1973, p.26.

Lattin, "Multiband Antennas Using Loading Coils," *QST*, April, 1961, p.43.

McCoy, "Monimatch, Mark II," *QST*, February, 1957, p. 38. Also, "An Etched-Circuit Monimatch For Checking Your Antenna System," *QST*, October, 1969, p.29.

Sevick, "The Ground-Image Vertical Antenna," *QST*, July, 1971, p. 16. Also, "The W2FMI Ground-Mounted Short Vertical," *QST*, March, 1973; "A High Performance 20- 40- and 80-Meter Vertical System," *QST*, December, 1973, p. 30; and "The Constant-Impedance Trap Vertical," *QST*, March,1974,p.29.

By Frank Witt, AI1H

From *QST*, September 1993

A Simple Broadband Dipole for 80 Meters

Turn your existing 80-meter dipole into a broadband antenna by simply modifying the feed line. Multiband operation is an option.

A conventional coax-fed, half-wave dipole doesn't provide a low SWR over the entire 80-meter band—an inconvenience for those of us who like to operate phone and CW on that band. Several approaches to overcoming this limitation, short of an antenna tuner in the station, have been described.[1,2] The antenna system described here is simpler than any of its predecessors and has the following features:

- A 2:1 SWR or better is achieved over all or most of the 80-meter band.
- Antenna length and appearance are the same as those of a conventional half-wave dipole. Consequently, it's lightweight and has small wind and ice loading.
- The antenna configuration permits multiband operation with a single feed line.
- The losses due to broadband matching are acceptable.
- The cost is about the same as a conventional half-wave dipole.

All the SWR data given in this article were measured at the transmitter end of the feed line. The reference impedance is 50 Ω, since most equipment is designed for this impedance. The term *antenna system* as used throughout this article includes not only the radiating wire, but also the feed line, balun (if used), any lightning-protection measures, antenna tuner and so forth.

The dipole antenna itself is not broadband; the system uses a broadband *match*. The key broadbanding element of this antenna system is the *transmission-line resonator*: Part of the transmission line compensates for the reactance presented by the dipole away from its resonant frequency. This part of the line is a multiple of an electrical half wavelength. Another part of the line presents an appropriate source impedance to the transmission-line resonator (TLR).

First I'll describe a version of the broadband antenna system, along with some practical results. Then I'll cover the important matter of antenna-system loss. Following that are some variations to suit specific requirements, and a method for using the antenna for several bands. I'll also compare transmission-line-resonator broadbanding to other broadbanding methods.

The 80-Meter Broadband Antenna System

Figure 1 shows the simple broadband antenna system as used at my station. The antenna proper is a center-fed half-wavelength dipole. The transmission line is segmented into one electrical wavelength of 50-Ω coax and an electrical quarter wavelength of 75-Ω coax. The calculated and actual lengths are shown in Table 1. Lengths were calculated using the formulas given later in this article, using a center frequency (F_0) of 3.75 MHz and VF (velocity factor) of 0.66. The actual lengths resulted after I performed the tuning procedure described later. Manufacturing variations from the published cable velocity factors, and some stretching of the coax, contributed to the differences between actual and measured values. (The actual lengths were measured on untensioned cable.) The antenna is installed as an inverted V with a 140° included angle and an apex height of 60 feet. The wire size is #14, but is not critical.

Table 1

Calculated and Actual Lengths of the Broadband Dipole Antenna at AI1H

	Calculated	*Actual*
1/4-λ Coax	43.3 feet	43.3 feet
1-λ Coax	173.1 feet	170.5 feet*
Dipole	124.5 feet	122.7 feet

*Includes 11 inches for balun.

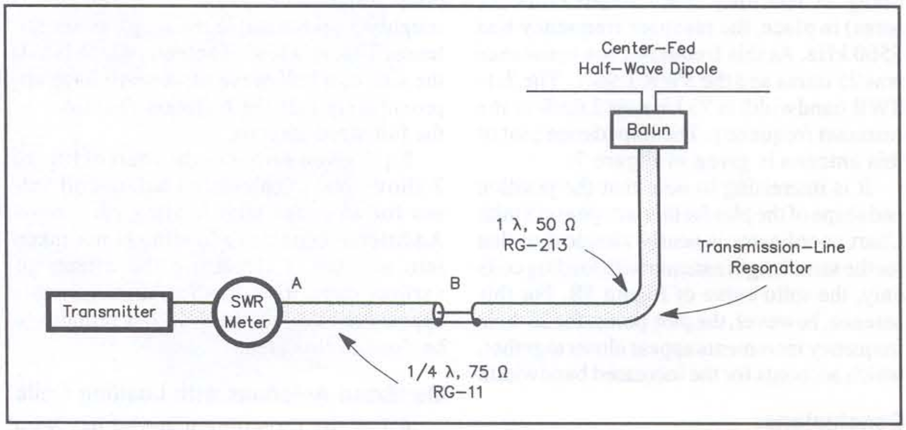

Figure 1—One form of the simple broadband antenna system. It resembles a conventional dipole except for the 1/4-wavelength, 75-Ω segment. Points A and B are discussed in the text.

This system's SWR (at the transmitter) as a function of frequency is shown in Figure 2. For comparison, the SWR for the same dipole fed with about 5/4 wavelengths (214 feet) of RG-213 coax is also shown. (This is the same total length as the RG-213 and RG-11 segments used in the broadband system.) The broadband system's 2:1 SWR bandwidth is 2.2 times that of the conventional system—and the *only* difference is the feedline configuration!

The radiating properties of the broadband antenna over the 80-meter band are essentially identical to those of a dipole cut for any specific frequency in the band. Also, since the antenna system is designed for a 50-Ω transmitter, the feed-line length may be extended by adding the required length of 50-Ω coax between the transmitter and the quarter-wave segment (point A in Figure 1).

A 1:1 current balun should be installed at the antenna's feed point. I use the balun on general principles. Often, it provides no visible difference in operation, but the balun does minimize feed-line radiation. You can determine whether your antenna needs a balun by measuring the SWR versus frequency with and without a balun installed. If the balun is not needed, the two sets of data will be identical.

Antenna-System Losses

It's important to know the losses in any antenna system. This is especially true for broadband antennas, because loss alone can broadband an antenna system. As the next section shows, the configurations presented in this article do not yield a significant loss penalty. Although other loss contributors exist in antenna systems, we will focus on the primary ones: *feed-line loss* and *mismatch loss*. Other losses, such as ohmic loss in the antenna wire, are the same for both the conventional and broadband systems described here.

Feed-line loss is the easiest to understand. It is unavoidable, and is lowest when the feed line is *flat* (when the line SWR is close to 1:1). At HF, feed-line loss results primarily from ohmic losses in the copper conductors.

Mismatch loss occurs when the impedance seen by the transmitter is not the complex conjugate of the transmitter's impedance (when the line SWR at the transmitter is not 1:1). For a 50-Ω transmitter, the mismatch loss is 0 dB when the load impedance is 50 Ω. When the load impedance is not 50 Ω, the mismatch loss can be made to be 0 dB if a transmitter with a tunable output stage (such as a conventional tube-type linear amplifier) is tuned for a conjugate match. An antenna tuner can also provide this match. In this case, however, the antenna-tuner loss (perhaps as much as 1 dB) replaces the mismatch loss in the total-loss equation. That subject isn't discussed here.

If you don't use an antenna tuner and the transmitter has a fixed-tuned 50-Ω output, loads that present the transmitter with an SWR under 2:1 are highly desirable. The impact of high SWR on mismatch loss will become clear in the next section.

Loss must be kept in perspective. All of the broadband antenna systems described here have a worst-case total loss of less than 3 dB—not enough to notice in many 80-meter QSOs. (If the loss is 3 dB, half of the transmitter's output power is radiated and half is lost elsewhere.) The main effect of loss is stress on system components: that on the transmitter due to the mismatched load, and that on the transmission line due to heating.

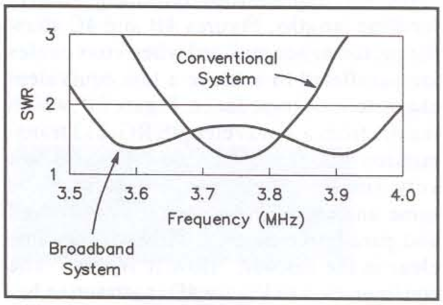

Figure 2—Measured SWR versus frequency for the broadband and conventional antenna systems.

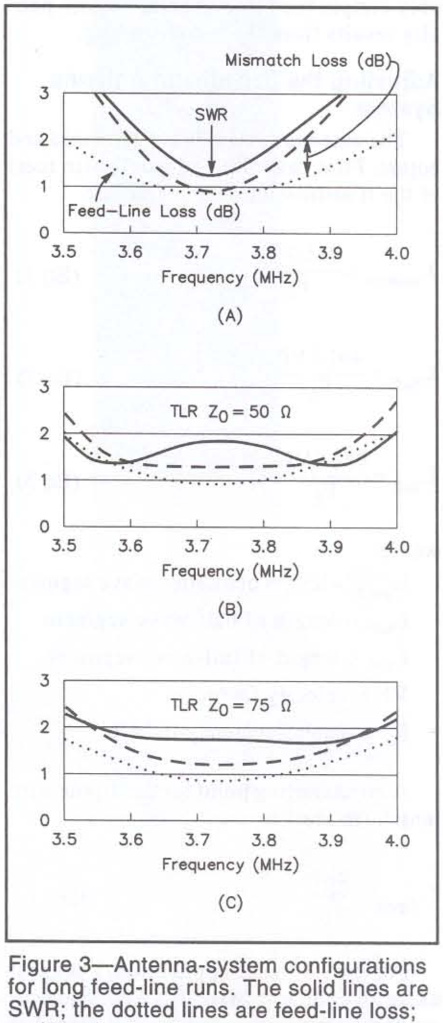

Figure 3—Antenna-system configurations for long feed-line runs. The solid lines are SWR; the dotted lines are feed-line loss; and the dashed lines are feed-line loss plus mismatch loss. At A, a conventional system using an RG-213 feed line; at B, the feed line is a 1/4-λ section of RG-11 (75 Ω) followed by 1 λ of RG-213; and at C, a 1/4-λ segment of RG-11 is followed by 1 λ of RG-11 (one 5/4-λ piece of RG-11). The total feed-line length in each case is 216.4 feet.

Variations

The broadband antenna system described above is well-suited for the installation at my station, where the distance between the shack and the antenna is rela-

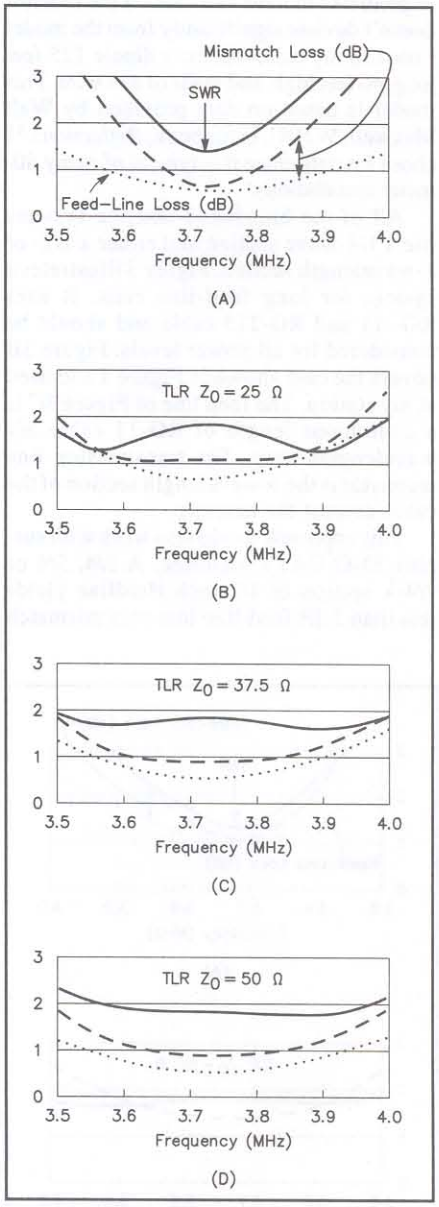

Figure 4—Antenna systems for high power and shorter feed-line runs. The solid lines show SWR; the dotted lines represent feedline loss; and the dashed lines show feed-line loss plus mismatch loss. At A, the feed line is RG-213; at B, it's a 1/4-λ section of RG-11 followed by two paralleled 1/2-λ lengths of RG-213; at C, a 1/4-λ segment of RG-11 is followed by two paralleled 1/2-λ lengths of RG- 11, and at D, 1/4-λ of RG-11 is followed by 1/2-λ of RG-213. The total feed-line length in each case is 129.8 feet.

tively long (more than 200 feet) and because I use a 1-kW amplifier. Other feed-line combinations are better suited to other installations. Some of these are shown in Figures 3 through 5, along with calculated SWR and loss data. From this information, you can select an appropriate feed-line combination for your needs.

The figures also show the characteristics of conventional dipole antenna systems. If you compare them, you'll see that the transmission-line resonator provides broadbanding without a significant loss penalty. I haven't tried all these combinations, but based on my experience, they should perform as predicted in most situations if the radiator doesn't deviate significantly from the model I used in my calculations: a dipole 125 feet long, 40 feet high, and made of #14 wire. This model is based on data provided by Walt Maxwell, W2DU, in his book, *Reflections*.[3] I chose his data since it is typical of many 80-meter installations.

All of the broadband antenna systems use a 1/4-wave section and either a 1/2- or 1-wavelength section. Figure 3 illustrates a system for long feed-line runs. It uses RG- 11 and RG-213 cable and should be considered for all power levels. Figure 3B covers the case shown in Figure 1 and used at my station. The feed line of Figure 3C is a continuous length of RG-11 cable 5/4 wavelengths long. The transmission-line resonator is the 1-wavelength section of the cable nearest the antenna.

This approach would also work with surplus 75-Ω CATV Hardline. A 3/4, 5/4 or 7/4-λ section of 1/2-inch Hardline yields less than 2 dB feed-line loss plus mismatch loss over the entire band, and less than 1 dB total loss over any 300 kHz of the band. This configuration is particularly attractive to contesters and DXers, because even a fairly long line7/4 λ is 372 feet of 1/2-inch CATV Hardline—gives low loss and a very good match over, say, the 3.5- to 3.8-MHz range.

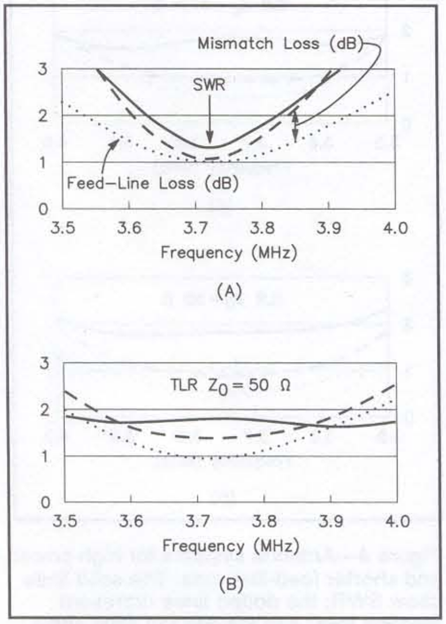

Figure 5—Antenna systems for low power and shorter feed-line runs. The solid lines show SWR; the dotted lines represent feedline loss; and the dashed lines show feed-line loss plus mismatch loss. At A, the feed line is RG-58; at B, it's a 1/4-λ section of RG-59 followed by 1/2 λ of RG-58. The total feedline length in each case is 129.8 feet.

Three broadband antenna systems are shown in Figure 4. All of these are candidates for applications requiring shorter feedline lengths. Figures 4B and 4C show the performance realized when coax cables are paralleled to achieve a low equivalent characteristic impedance. Figure 3B, which results from a 1-wavelength RG-213 transmission-line resonator, and Figure 4B, are very similar. The latter system uses the same amount of cable, but it's cut in half and parallel-connected. This will become clear in the sidebar, "How It Works." The configuration in Figure 4D is attractive because of its simplicity.

Lower-power applications without long feed-line runs can use RG-58 and RG-59 coax. Figure 5B shows how excellent broadbanding is achieved with a remarkably simple feed line. Again, no loss penalty results from the broadbanding.

Adjusting the Broadband Antenna System

The antenna system is easy to build and adjust. First calculate the lengths (in feet) of the transmission-line segments:

$$L_{quarter} = \frac{245.9\,VF}{F_0} \qquad \text{(Eq 1)}$$

$$L_{half} = \frac{491.8\,VF}{F_0} \qquad \text{(Eq 2)}$$

$$L_{full} = \frac{983.6\,VF}{F_0} \qquad \text{(Eq 3)}$$

where

$L_{quarter}$=length of quarter-wave segment

L_{half} = length of half-wave segment

L_{full} = length of full-wave segment

VF = velocity factor

F_0 = center frequency in MHz

A good starting point for the dipole wire length (in feet) is:

$$L_{dipole} = \frac{467}{F_0} \qquad \text{(Eq 4)}$$

For the 80-meter application, I suggest using an F_0 of 3.75 MHz. It's a good idea to cut the wires so that the overall length is 4 feet longer than necessary, in case you need to lengthen the wire during tuning. Pass 2 feet of the extra wire through each end insulator and wrap it back around the antenna wire.

To tune the antenna system, you'll change only the dipole and transmission-line-resonator lengths. The best approach is to build the antenna system as I have outlined here and to measure the SWR at the transmitter end of the system. Any tilt or frequency offset in the SWR characteristic can be removed by increasing or decreasing the dipole or transmission-line resonator length. Start by changing the length of the dipole. To improve the SWR at the high end of the band, the dipole must be shortened; to improve the SWR at the low end of the band, the dipole must be lengthened. Progressively add or subtract 6 inches from both legs of the dipole until the SWR curve is symmetrical about the center frequency.

Frequency offset may be required to center the SWR characteristic in the 80-meter band. You can move the entire curve along the frequency axis without causing asymmetry by changing both the dipole and transmission-line resonator lengths using the following equation:

$$L_{New} = L_{Old}\left(\frac{3750 - \Delta F}{3750}\right) \qquad \text{(Eq 5)}$$

ΔF is the required frequency offset in kilohertz. Shortening the dipole and resonator moves the curve center up in frequency, and lengthening them moves the center down. The length of the quarter-wave segment need not be changed, since the SWR characteristic is not very sensitive to its length.

Lightning Protection

Every antenna system should be designed to minimize the likelihood of a lightning strike. One part of this is keeping all parts of the antenna proper at ground potential. The grounding should be done *outside the shack*, by means of a good ground rod.

I recommend that you install a coaxial lightning protector, which bleeds any static charge from the center conductor, at point B of Figure 1. The protector (and therefore the feed-line shield) should be connected to a high-quality ground rod (the kind electricians use) driven 8 feet into the ground.

Conversion of Existing 80-Meter Dipoles

A study of the cases shown in Figures 3B, 4D and 5B suggests that it's possible to easily convert many existing 80-meter half-wave dipole antennas. Because the most popular way to feed an 80-meter dipole is with a 50-Ω coaxial feed line, the conversion to a broadband antenna system is straightforward. First trim the dipole for resonance at about 3.75 MHz. Then cut the 50-Ω feed line at a multiple of an electrical half-wavelength (at 3.75 MHz) from the antenna. Calculate this length using Eq 2 or Eq 3. Add the 75-Ω quarter-wave section, then complete the run to the shack (if necessary) with 50-Ω coax. Then use the tuning procedure described earlier to optimize the system.

Multiband Operation

Most broadband 80-meter antenna systems are usable only on the 80-meter band, because the broadbanding elements do not allow efficient power transfer on other bands. This is not true with the approach described here, since the structure consists only of a center-fed dipole and a transmission line. Moreover, the transmission-line segments are close to multiples of an electrical half-wavelength near 40 meters and other bands. This opens the possibility for paralleling other half-wave dipoles with the 80-meter dipole and sharing the feed line.

To minimize their interaction, the various dipoles should be spaced from each other away from the feed point. Of course, some interaction will occur and you must tune the multiband system to meet your requirements. I recommend first tuning the 80-meter broadband system and then the next-highest-frequency dipole, and so forth. Only the 80-meter antenna will be broadband, but such broadbanding is not required on the other bands. Figure 6 shows the result of adding a 40-meter dipole to the Figure 1 antenna. Each dipole leg is 34.4 feet long. Note that the SWR on 80 meters changes very little compared to Figure 2. No change was made to the 80-meter dipole or the transmission line.

The multiple-dipole approach described above achieves resonance on several bands and eliminates the need for an antenna tuner on those bands. Of course, if you use an antenna tuner, operation on all HF bands should be possible, but this arrangement is usually not as effective as the multiple-resonance antenna system described here because the feed-line loss is much higher.

Comparison with the Coaxial-Resonator Match

How does the simple broadband dipole described here stack up against other approaches for achieving a good match over the entire 80-meter band? The coaxial resonator match broadband dipole[4,5] represents one of the more efficient designs published to date. It achieves broadband matching at the antenna by the integration of 1/4 wavelength of coaxial cable as a part of the antenna.

Since the coaxial-resonator match achieves a good match at the antenna, the SWR on the feed line is low and the feedline loss is about the same as its matched loss. However, the coaxial cable in the match itself increases the system loss. The net result is that the total loss is about the same with the coaxial-resonator match, but the SWR at the transmitter is lower, never exceeding about 1.6:1 between 3.5 and 4 MHz. Once the SWR is less than 2:1, however, a lower SWR has little value unless you're using a transmitter that significantly reduces power at such SWRs.

Note that the approach described in this article uses a thin wire for the antenna. Most other broadbanding approaches use additional wires or radiators made partly from coaxial cable and are vulnerable to damage from wind and ice loading. Their additional weight and complexity are also limitations.

From the above comparison, the simple broadband antenna system has, by its very simplicity, an edge over the coaxial resonator match, at least in applications where the simpler approach is feasible. Because of the limitations of available coaxial cables, the opportunity for a satisfactory design is constrained. On the other hand, the coaxial resonator match has more adjustment parameters, is useful over a much broader range of applications and yields the lowest SWR over the band.

How It Works

A fundamental way of achieving a broadband match to a resonant dipole antenna involves a parallel-tuned LC network and an appropriate source resistance. In an *RF Design* article,* I described the method for designing such networks, even with lossy resonators. The top of Figure A shows the equivalent circuits of the antenna and matching network. The bottom of Figure A illustrates the corresponding elements in the antenna system.

The role of the resonator is played by the transmission-line segment nearest the antenna. It must be a multiple of an electrical half-wavelength. The quarter-wavelength "Q"-section, made from 75-Ω coax, transforms the 50-Ω transmitter reactance to 112.5 Ω ($75^2/50=112.5$). I won't go into the design details here; they're the subject of another article, "Broadband Matching Using the Transmission-Line Resonator," in preparation for *The ARRL Antenna Compendium, Volume 4.*

For the structure of Figure A to yield a broadband match, the characteristic impedance of the transmission-line resonator and the transmitter resistance must be within a range of values. Fortunately, commonly used transmission lines, which are available in 50- and 75-Ω characteristic impedances, work well in this application. The broadband systems of Figures 3 through 5 show the usefulness of this approach.

Figure B makes another significant point. For this application, the network parameters of a one-wavelength transmission-line resonator (top) are similar to those of a half-wavelength resonator (bottom) with half the characteristic impedance of the upper resonator. Parallel-connecting two identical cables is a convenient way of achieving lower characteristic impedances. This explains the similarity of Figures 3B and 4B and the similarity of Figures 3C and 4C.—*AI1H*

*F. Witt, "Optimum Lossy Broadband Matching Networks for Resonant Antennas," *RF Design*, Apr 1990, pp 44-51 and Jul 1990, p 10.

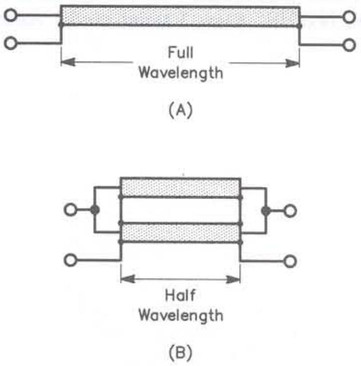

Figure A–Lossy broadband-matching-network equivalent circuit (top), and corresponding simple broadband antenna system elements (bottom).

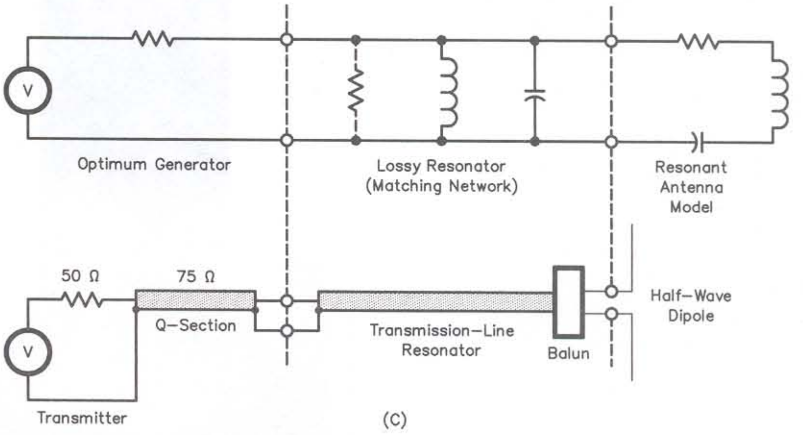

Figure B–These two transmission-line resonators behave essentially the same in this application. The characteristic impedance of each cable segment is the same, making the characteristic impedance of the lower resonator half that of the upper one.

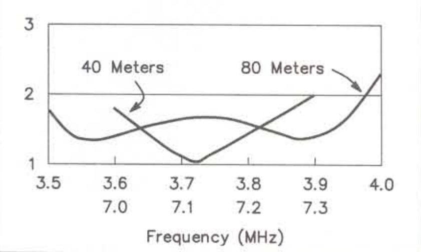

Figure 6—Measured SWR for the 80- and 40-meter multiband antenna system.

Summary

The simple broadbanding technique I've described here capitalizes on the common availability of coaxial cables that fit the application. It overcomes the narrow bandwidth limitations of a conventional 80-meter, half-wave dipole without significant disadvantages. Even parallel dipoles for other bands may be fed with the same feed line.

The limitation of available coaxial cable parameters can be overcome by using the transmission-line resonator as a resonant transformer. Applying this technique is described in an upcoming *ARRL Antenna Compendium* article, "Broadband Matching Using the Transmission-Line Resonator."

This work has benefited from the support and encouragement of my wife, Barbara, N1DIS. Also, I must credit Andrew Griffith, W4ULD, for helping to turn my attention to the approach described here. After reading my *QST* article on match bandwidth of resonant antenna systems,[6] Andy noted that antenna systems should be viewed from their match to a 50-Ω transmitter, even if the feed line does not have a 50-Ω characteristic impedance. He showed examples of the narrowing of match bandwidth to make his point. In my response, published with Andy's letter in *QST*,[7] I pointed out that match bandwidth of an antenna system may actually be *increased* by selecting the right cable length and characteristic impedance. As an example, I showed in Figure 3 of that correspondence the large match bandwidth of a dipole fed with a 5/4-wavelength, 75-Ω RG-11 cable. Note that this is the same case shown in Figure 3C of this article. Thank you, Andy!

Notes

[1] G. Hall, ed, *The ARRL Antenna Book*, 16th ed (Newington: ARRL, 1991), pp 9-1 through 9-12.

[2] M.W. Maxwell, *Reflections: Transmission Lines and Antennas* (Newington: ARRL, 1990), pp 18-1 through 18-6.

[3] See Note 2, p 15-19. I frequency-scaled Walt's data, which is equivalent to changing the wire length. This antenna has a Q of 13 and a resonant resistance of 65-Ω. I took into account the fact that the antenna's radiation resistance increases with frequency.

[4] F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

[5] F. Witt, "The Coaxial Resonator Match," *The ARRL Antenna Compendium, Volume 2* (Newington: ARRL, 1989), pp 110-118.

[6] F. Witt, "Match Bandwidth of Resonant Antenna Systems," *QST*, Oct 1991, pp 21-25.

[7] A. Griffith and F. Witt, "Match Bandwidth Revisited," *QST*, Technical Correspondence, Jun 1992, pp 71-72.

From *QST*, June 1985 (Technical Correspondence)

Once More With the 80-Meter Broadband Dipole

With reference to earlier items by K1TD and W7ZOI in *QST*, it *is* possible to design a simple, adjustment-free matching network that will make an 80-meter dipole have a reasonable SWR across the band.[1,2] When using Hayward's RLC model of the dipole, a simple parallel-tuned circuit across the antenna feed point gives less than 2:1 SWR across all but the bottom 10- to 20-kHz of the band, as shown in Figure 1.

This method *does* require a source impedance of 140 ohms, so a broadband matching transformer will be needed at the input. I see no reason why you couldn't just tap the input down on L1 and use L1 as an autotransformer to accomplish the impedance matching function. In fact, although I have not tried it, it may be possible to include the balanced-to-unbalanced transformation in the same coil, as shown in Figure 2. To tune this network, use a dip meter with no antenna connected. Then install the network on the antenna and adjust the input tap for the best compromise SWR across the band.

By choosing a lower characteristic impedance, you can get much better SWR over a narrower bandwidth (Figure 1). Retune both the antenna and the network to center the response on either the phone or CW portion of the band.

You can get better SWR in both the fullband and narrow-band case by adding a series-tuned circuit to the input (Figure 3). The improvement is marginal, however, and complicates the impedance step-up arrangement, so it is probably not worth the effort. It would be better to take Hayward's suggestion and make the antenna itself inherently more broadband. One method is to "fatten" the elements by using a wire cage arrangement or similar. Of course this would change the values of Zo, L1 and C1.

The original graphs for Figures 1 and 3 were drawn with an HP-87. I would be happy to send a copy of the rogram to anyone who sends me an s.a.s.e.—*Alan Bloom, N1AL*

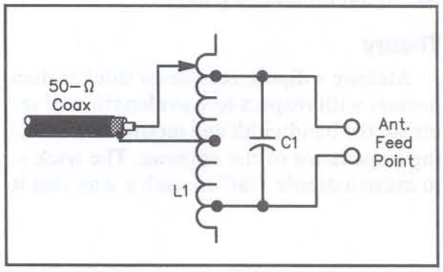

Figure 2—Proposed method of feeding a balanced antenna with unbalanced line. The shielded conductor of the coax line must be connected at the exact center of the coil, which is an electrically neutral point in the antenna system.

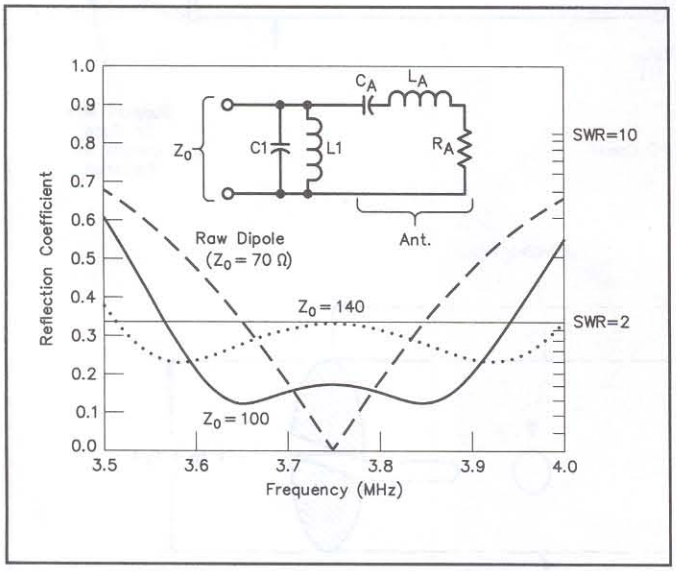

Figure 1—Calculated reflection coefficient and standing-wave ratio vs. frequency for a single-wire dipole and dipoles with lumped constants (C1 and L1) at the feed point.

For the curve where Zo is 140 ohms:
C1 = 2000 pF
L1 = 0.9006 μH

For the curve where Zo is 100 ohms:
C1 = 3500 pF
L1 = 0.5146 μH.

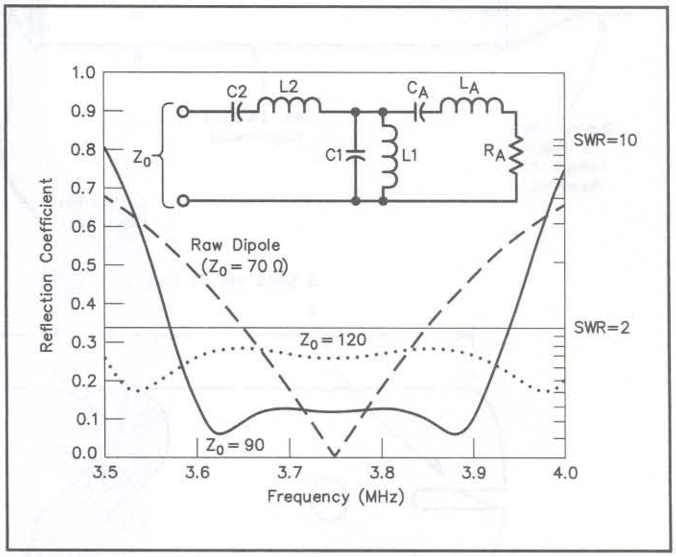

Figure 3—Calculated reflection coefficient and standing-wave ratio vs. frequency for an antenna with shunt and series lumped constants at the feed point (C1-L1 and C2-L2, respectively).

For the curve where Zo equals 120 ohms:
C1–3500 pF
C2–60 pF
L1–0.5146μH
L2–30.0211 μH

For the curve where Zo equals 90 ohms:
C1–7000 pF
C2–60 pF
L1–0.2573 μH
L2–30.0211 μH

---

[1] J. Hall, "The Search for a Simple, Broadband: 80-Meter Dipole," *QST*, April 1983, p. 22.

[2] W. Hayward, "Limitations to Broadband Impedance Matching," Technical Correspondence, *QST*, July 1984, p. 45.

By Robert C. Wilson

From *The ARRL Antenna Compendium, Vol 2*

Fat Dipoles

Antennas without problems make radio communications enjoyable. I design overseas radio stations for a living, so I'd rather not have to fight my own ham station when just relaxing and rag chewing. Fat dipoles do the things I want. They match the coax line well over a wide band, and they launch the signal remarkably well.

Theory

Making a dipole conductor thicker than normal with respect to wavelength will increase the bandwidth and modify the working impedance of the antenna. The trick is to make a dipole "fat" in such a way that it may be easily constructed from cheap materials, be highly efficient and at the same time arrange things so that it will match the transmission line from the lower band edge to the upper band edge.

I started with the assumption that my band of interest would be the 80/75-meter band. From end to end, this requires a 13% bandwidth to the 2:1 SWR points for my broadband-solid-state final. I also assumed that this antenna was going to be at a nominal height of 30 feet or 0.11 wavelength above ground. The calculations indicated that a dipole built of four quarter-wavelength no. 14 wires (0.064 inch) with a spacing of 0.0114 wavelength would produce the necessary results. The correct length would have to be 0.45 wavelength to match a 50-ohm line.

Length = 442.5/f feet
Width = 11.25/f feet
Height = 112.5/f feet
where f = center frequency in MHz

Construction

Very few problems will be encountered in building this simple fat dipole if you follow the drawing (Figure 1). First you will need five good insulators. I prefer egg type

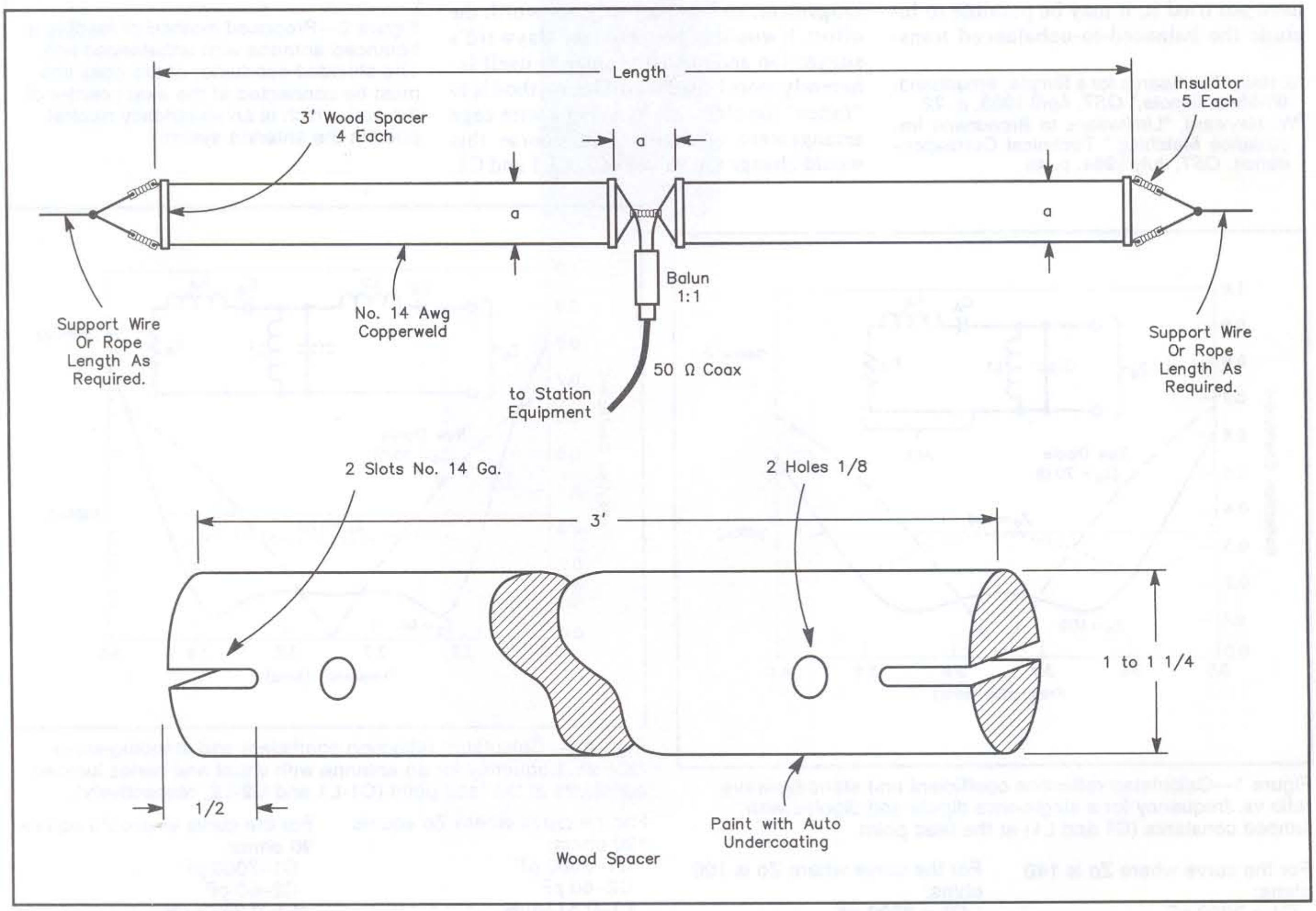
Length
3' Wood Spacer
4 Each
a
a
a
Insulator
5 Each
Balun
1:1
Support Wire
Or Rope
Length As
Required.
No. 14 Awg
Copperweld
50 Ω Coax
to Station
Equipment
Support Wire
Or Rope
Length As
Required.
2 Slots No. 14 Ga.
2 Holes 1/8
3'
1 to 1 1/4
1/2
Paint with Auto
Undercoating
Wood Spacer

Figure 1—The fat dipole with construction details of the spreaders. See Table 1 for the length and spreader dimensions (A).

insulators but there is no critical problem here. One insulator is for the center of the dipole and the others are for the four ends. You will need four 3-foot-long broom sticks or 1-inch wood or plastic rods with good weatherproofing. I painted mine with auto undercoating but outdoor paint or varnish should also work. Copperweld wire is very desirable because it won't stretch and change the tuning of your antenna. This type of wire is available through advertisements in *QST*. The same source may also be able to supply the essential wide-band balun transformer and coax. Either the RG-58 or RG-8 types of coax cable are satisfactory but the latter requires more support because of its greater weight.

Measure your wire carefully and leave enough extra so that the insulators can be attached. The final length will need to be the calculated value, from insulator wire end to insulator wire end. After building the four-wire section, attach dowel rods as shown to act as spreaders. Fasten each rod in place with pieces of wire threaded through the holes and then wrapped around the antenna wire. Wrap these spreader wires tight enough so that the rods will not slip out of place. Snip off projecting wire ends wherever they occur to prevent RF corona power loss. Then, using either wire or rope, make a bridle to hold the ends of the antenna.

Table 1

Dimensions and Bandwidths on Various Bands

Frequency MHz	*Length, ft*	*Spacing (A), ft*	*Bandwidth, MHz*
1.9	233	5.9	0.252
3.75	118	3	0.5
7.15	62	1.6	0.951
14.175	31.2	0.8	1.885
21.225	20.85	0.53	2.823
28.860	15.34	0.39	3.837

Last, solder the balanced end of the balun transformer to the dipole. Each wire from the balun should go to the pair of wires on the same side of the dipole. The solder job should be of the best quality and permanent because it is hard to repair later. The coax needs to be connected to the unbalanced side of the balun. If you use large-diameter coax (3/8 inch) then think about ways to support the weight. Perhaps a piece of nylon rope from the dipole center insulator to the coax will help take the load, but I'll leave the details of the problem up to you. After this final construction step, haul the antenna up in the trees, using care that no twists are allowed.

Operation

For once I had a 75-meter antenna that worked better than predicted. The SWR was 1.6:1 or better from 3.5 to 4.0 MHz. Better yet, reports received were excellent with my old 100-watt solid-state transceiver. Moving up and down the band gave no loading problems from the broadband final. The fat dipole is just what I needed for a good, relaxing rag chew after a hard day with the 500-kW rig at the office.

CHAPTER TWO

OFF-CENTER-FED-DIPOLES

2-1 **The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna**
From August 1990 *QST*

2-9 **Improved Feed for the Off-Center-Fed Dipole**
From May 1996 *QST* (Technical Correspondence)

2-11 **Off-Center-Fed Dipole Comments, Part 2**
From October 1996 *QST* (Technical Correspondence)

2-13 **Off-Center-Fed Dipole Comments, Part 3**
From October 1996 *QST* (Technical Correspondence)

2-14 **Choke the OCFD**
From September 1997 *QST* (Technical Correspondence)

By John Belrose, VE2CV and Peter Bouliane, VE3KLO **From *QST*, August 1990**

The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna

The search for a simple broadband, multiband antenna with acceptable SWR over the entire 80/75meter band continued with the publication of the article by Witt, AI1H, in April 1989 *QST*. This inspired author Belrose, VE2CV, to reopen his files on the "Windom antenna," or, more properly, the off-center-fed dipole. Particularly since the opening of the 10, 18 and 24-MHz bands, this time-honored, somewhat controversial antenna is attracting a revival in interest and usage because of its multiband characteristics. Thus, the off-center-fed dipole deserves an update. Certainly, most North American radio amateurs have not heard about the German version of this antenna.

Even though it is one of the simplest antennas, the drooping dipole (a dipole with drooping ends, popularly referred to as an inverted V) is effective on 80 meters. For good DX performance, the apex of an 80-meter drooping-dipole antenna should be as high as possible above ground (at least 15 meters [50 feet]), because the antenna's vertical radiation angle decreases with height above ground. The arms of an 80-meter drooping dipole should not be too close to the ground (say, greater than 3 m [10 ft]). The input impedance $(R_a \pm jX_a)$ of a drooping dipole depends on the operating frequency, the length of the dipole, the angle between its arms, the dipole's height above ground, and–particularly the resistive component–on the conductivity of the earth beneath the antenna. The included angle between the arms of an 80-meter drooping dipole should be about 127°, since, in this configuration, the antenna's pattern is dipole-like, and its input impedance is about 50 Ω for antenna heights typically employed by radio amateurs. The principal disadvantage of such an antenna is that its SWR bandwidth–less than 200 kHz–is too narrow to cover the 80/75-meter band.

The broadband performance of an antenna can be improved through the use of a matching network at the feed point. This network can comprise discrete components (see Hall,[1] Hayward,[2] Bloom,[3] Hately,[4] and Li, et al[5]); or transmission-line stubs (Snyder,[6] Hansen,[7] and Witt,[8]). Authors Li, et al give a microcomputer program for the design of LCR networks for broadband matching.

Another way to improve broadband performance is to use two dipoles fed in parallel: one dimensioned for the middle of the lower half of the band, and the other for the middle of the upper half of the band. The drooping-dipole configuration is ideal for this arrangement, since the ends of the two dipoles can be fanned (for angular separation) in either the vertical or horizontal plane. An alternative is to use an off-center-fed dipole, which, in addition to broadbandedness, has multiband performance characteristics. This article is concerned with such an antenna.

The "Windom Antenna" and Single-Wire Feed

The original Windom antenna (devised in 1928-29), named after Loren Windom, W8GZ, the amateur who wrote a comprehensive article about it,[9] employed single wire feed at a point of 1/6 to somewhat over 1/7 of the antenna length from the center (see Figure 1A). Windom reported on a detailed experimental study by colleagues John Byrne, W8DKZ, Edward Brooke, W2QZ, Jack Ryder, W8DKJ, and Prof W. L. Everitt of the Ohio State University Dept of Electrical Engineering. They found that if ammeters were placed on the antenna, with the single-wire feeder at the position just described, the current distribution on the dipole was sinusoidal and symmetrical, with no discontinuity in the vicinity of the feed point, and no standing wave on the feeder. Clearly, the feeder was terminated in its characteristic impedance.

This article did not discuss pattern, and it was many years before the computational tools for predicting pattern were available. Parfitt and Griffin[10] have recently analyzed the single-wire-fed dipole (the Windom antenna), and their results show, as anticipated, that radiation from the feed wire does

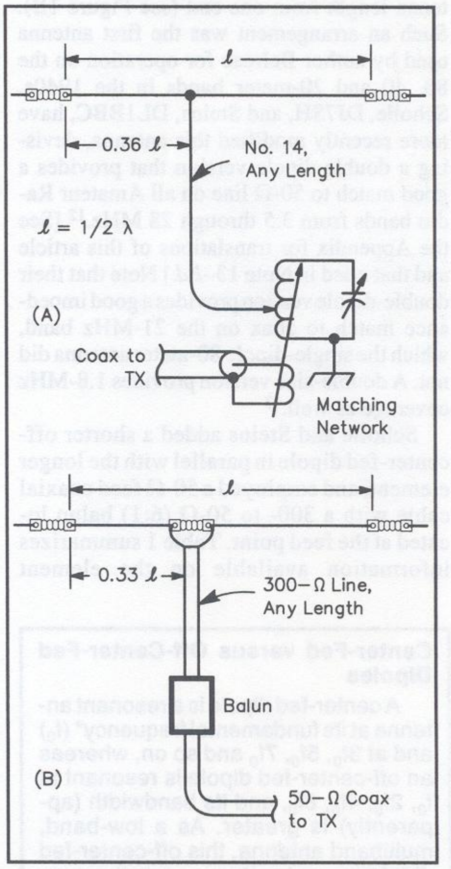

Figure 1—Two versions of the off-center-fed dipole. At A, the original "Windom antenna"–actually a Hertz (1/2 λ) element excited via a single-wire feeder; at B, a dipole off-center-fed via balanced line. The single-wire-fed version dates from the late 1920s; the balanced-line-fed version (sans balun) dates from the 1940s, and the balanced-line-and-balun version appeared in ARRL publications beginning in the late 1950s.

modify the antenna pattern, primarily at low elevation angles with the horizon. The feed line for their modeled antenna was vertical.[11] The pattern was basically dipole-like, with a squint away from the broadside. Current on a wire, whether uniform (a traveling wave) or sinusoidal (a standing wave) leads to radiation, unless the radiation that results from this feedwire current is canceled by an equal amplitude, out-of-phase current on an adjacent parallel wire (as in a balanced transmission line). For the Windom antenna, feeder radiation could be significant, depending on the arrangement and the length of the transmission line.

The Windom antenna was widely used during the 1930s and into the 1940s. For example, in author Belrose's experience, the British Columbia Forest Service were using the single-wire, off-center-fed dipole in the mid-1940s. It was a simple antenna that was easy to tune and match with the π network employed in the output circuit of their transmitter.

The Off-Center-Fed Dipole and Balanced Transmission Line

A later version (1940s) of an off-center fed dipole (*mis*called Windom) employed 300-Ω ribbon feed at a point 1/3 of the antenna length from one end (see Figure 1B). Such an arrangement was the first antenna used by author Belrose for operation on the 80, 40 and 20-meter bands in the 1940s. Scholle, DJ7SH, and Steins, DL1BBC, have more recently modified this antenna, devising a double-dipole version that provides a good match to 50-Ω line on all Amateur Radio bands from 3.5 through 28 MHz.[12] [See the Appendix for translations of this article and that cited in Note 13–*Ed.*] Note that their double-dipole version provides a good impedance match to coax on the 21-MHz band, which the single-dipole 80-meter antenna did not. A double-size version provides 1.8-MHz coverage as well.[13]

Scholle and Steins added a shorter off-center-fed dipole in parallel with the longer element, and employed a 50-Ω feed coaxial cable with a 300- to 50-Ω (6:1) balun located at the feed point. Table 1 summarizes information available on the element lengths for single and double off-center-fed dipoles. Since the dipole must be resonant on harmonic frequencies, its length must be somewhat longer than optimum for the 75-meter band, as noted by Scholle and Steins. It is apparent from their SWR curves that the resonant frequency of their antenna is less than 3.5 MHz.

Table 1

Summary of Element Lengths for Published Off-Center-Fed-Dipole Designs

Figure 1B Version	13.5 m (44.29 ft)	27.94 m (91.67 ft)
cq-DL double-dipole version	13.8 m (45.28 ft) 4.69 m (15.39 ft)	27.7 m (90.88 ft) 9.38 m (30.77 ft)
cq-DL double-sized, double-dipole version	25.88 m (84.91 ft) 4.69 m (15.39 ft)	51.77 m (169.85 ft) 9.38 m (30.77 ft)
Author's double-dipole version, (balanced coaxial-line feed, Figure 4)	23.3 m (76.44 ft) 6.78 m (22.24 ft)	48.2 m (158.13 ft) 13.58 m (44.55 ft)

In the arrangement employed by Scholle and Steins, the longer elements were horizontal and the shorter elements sloped (the included angle between the arms of which was Λ≈100°). The authors' version was a drooping-dipole configuration (Λ≈127°) with the angle between the arms of the two dipoles approximately 45°. This antenna was fed with two 15-meter (50-foot) lengths of RG-62A (foam dielectric) coaxial cable configured as a balanced transmission line.[14]

We fabricated a *single* off-center-fed dipole (dimensionally identical with the longer element of the Scholle and Steins "double dipole") and used it as a Field Day antenna (in 1985) for 75/80 in. The stimulus for this, and for carrying out a detailed study, was a search for a simple all-band antenna conducted for the Canadian National Institute for the Blind.[15] We experimented with 4:1 and 6:1 baluns, primarily because 4:1 baluns are easier to come by (more on baluns later). The antenna was operated in a drooping-dipole configuration (Λ≈127°), with the apex (which, in our case, was the feed point) at about 12 meters (40 feet). (In retrospect, we should have positioned the center of the antenna at apex height, since in this configuration the current maximum on the antenna [for 80-meter operation] would be at maximum height.) Extensive impedance and SWR measurements were made. To facilitate this, we employed a low-loss (foam-dielectric) coaxial feeder consisting of a 32.3-meter (106-foot) length of Belden 8214 (velocity factor, 0.8.) This feeder was 1/2 λ long at 3.5 MHz, 1 λ at 7 MHz, and so on. High-quality commercial baluns (made by Antenna Engineering, Australia[16]) were used between the feed line and the antenna.

Figure 2 shows our SWR measurements for the 80/75-meter band. With the 6:1 balun in line, the 2:1 SWR bandwidth was 3.47 to 3.93 MHz—broadbanded, indeed. The SWR was even lower on 40 and 20 meters, and less than 2:1 for the lower half of the 10-meter band; see Figure 3. Clearly, these bandwidths are not in accord with simple

Center-Fed versus Off-Center-Fed Dipoles

A center-fed dipole is a resonant antenna at its fundamental frequency* (f_0) and at $3f_0$, $5f_0$, $7f_0$ and so on, whereas an off-center-fed dipole is resonant at f_0, $2f_0$, $4f_0$, $8f_0$, and its bandwidth (apparently) is greater. As a low-band, multiband antenna, this off-center-fed dipole has advantages over the center-fed dipole, but a disadvantage (or advantage, depending on the antenna configuration and one's point of view) is that feeder radiation can contribute to the antenna's resultant radiation pattern. The use of a current balun minimizes feeder radiation.—*VE2CV and VE3KLO*

*The frequency at which it is 1/2 λ long.

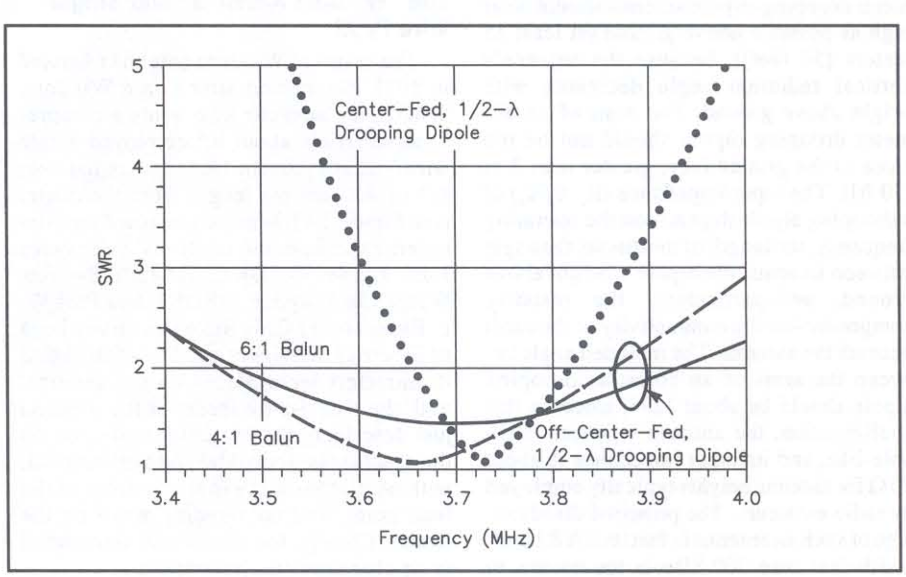

Figure 2—SWR vs frequency for a single off-center-fed dipole employing different baluns (4:1 and 6:1) at the feed point, compared with a center-fed 1/2-λ dipole. Both antennas were operated in a drooping-dipole configuration.

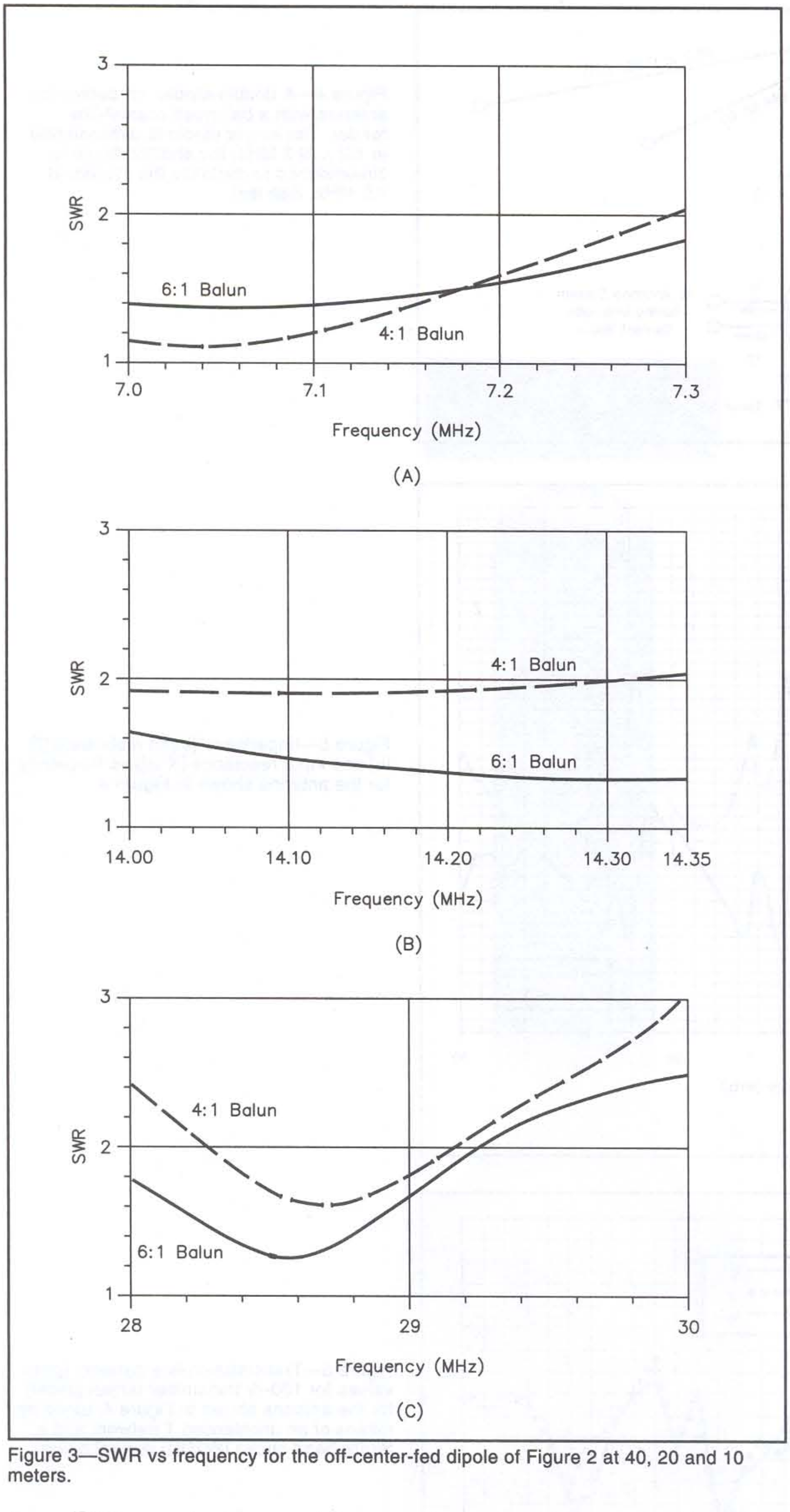

Figure 3—SWR vs frequency for the off-center-fed dipole of Figure 2 at 40, 20 and 10 meters.

theory.[17] Fifteen-meter data are not shown because this single off-center-fed dipole was not resonant on the 15-meter band.

The 4:1 balun provided somewhat sharper resonances and lower SWR at resonance in the 80/75- and 40-meter bands, but a low SWR was not obtainable in the 20-meter band. The measured impedance at resonance in the 80/75-meter band with this balun was exactly 50 Ω; hence the antenna's effective feed-point impedance was 200 Ω. This explains why the minimum SWR with the 6:1 balun was 1.5:1; the 6:1 balun, however, seems to be a better compromise if operation on other bands is wanted.

As noted, we employed high-quality commercial baluns (from the AEA Model 250 series). Radio amateurs have several alternatives in this connection; these include using a 4:1 balun and 75-Ω coax; or purchasing or winding your own 6:1 balun (see Orr, W6SAI[18]),and using 50-Ω coax.

Aside from impedance matching, however, there are other factors to consider in feeding an off-center-fed dipole and selecting a balun for this service. Next, we address some of these aspects.

Balun and Feeder Considerations

In the autumn of 1989, we decided to have another look at off-center-fed dipoles. Because of the multiple-resonant-frequency response of the double off-center-fed dipole, we decided to explore the potential of this antenna system for broadband frequency coverage. If used with a suitable antenna-system tuning unit (ASTU), a double off-center-fed-dipole system could perhaps be used on any frequency from 1.8 to 30 MHz. Since conventional baluns do not perform satisfactorily into reactive loads, we decided to eliminate the balun, at least insofar as the antenna and its feeder were concerned. We fed our antenna with a balanced 190-Ω transmission line consisting of two 15-meter (50-foot) lengths of RG-62A (foam-dielectric) coaxial cable as shown in Figure 4 and described in the work cited in Note 15. Teflon-dielectric cable would have been preferable because of Teflon's superior insulating characteristics but 95-Ω, Teflon-dielectric coax is unavailable.

We started out with a single dipole element dimensioned to 1/2 λ at 2 MHz. This antenna exhibited an antiresonant response (very high input impedance, measured at the transmitter end of the transmission line) at 7.6 MHz. Next, we installed the shorter dipole and dimensioned it to *minimize* the system's impedance at 7.6 MHz. Figure 5 shows the system's impedance v frequency. Except for a narrow band of frequencies near 6 MHz, the input resistance fell in the range of 20 to 400 Ω; and the input reactance fell in the range +*j*100 to –*j*200 Ω. This double off-center-fed dipole system is rather easily tuned and matched, since its reactance is low.

Figure 6 graphs antenna currents for a transmitter output power of 100 W. The antenna system was tuned using an unbalanced T network, with a ferrite-bead current balun of the type described by Maxwell[19] between the ASTU and the antenna. A balanced T network patterned after that described by Belrose was initially used since this is the type of ASTU that, in principle, should be used to tune and match a balanced antenna system. Unfortunately, however, this tuner employed a standard toroidal-core balun—a *voltage* balun, which applies almost-equal *voltages* to the wires of a balanced feeder. Even though the balun was on the "tuned side" of the ASTU—a practice we recommend—and the ASTU it-

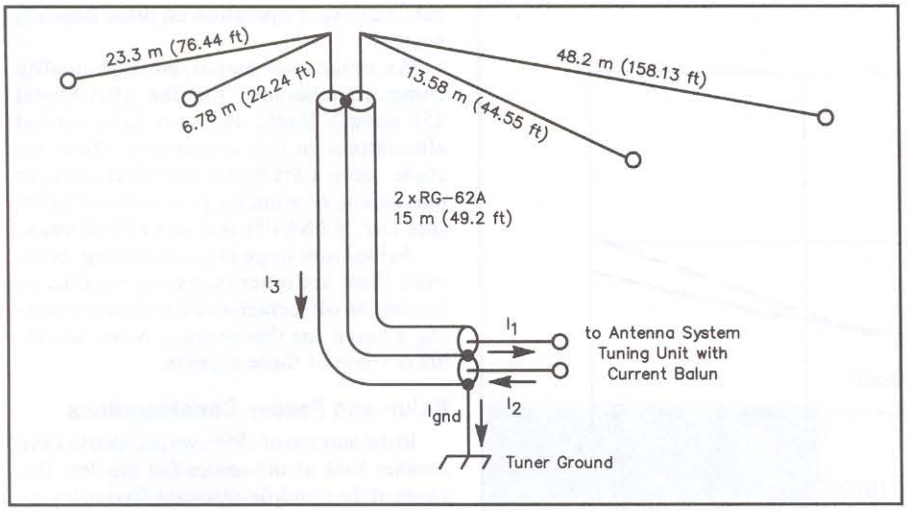

23.3 m (76.44 ft)
6.78 m (22.24 ft)
13.58 m (44.55 ft)
48.2 m (158.13 ft)
2 x RG–62A
15 m (49.2 ft)
I_3
I_1
to Antenna System
Tuning Unit with
Current Balun
I_{gnd}
I_2
Tuner Ground

Figure 4—A double-dipole, off-center-fed antenna with a balanced coaxial-line feeder. The longer dipole is dimensioned to 1/2 λ at 2 MHz; the shorter dipole is dimensioned to minimize the system at 7.6 MHz. See text.

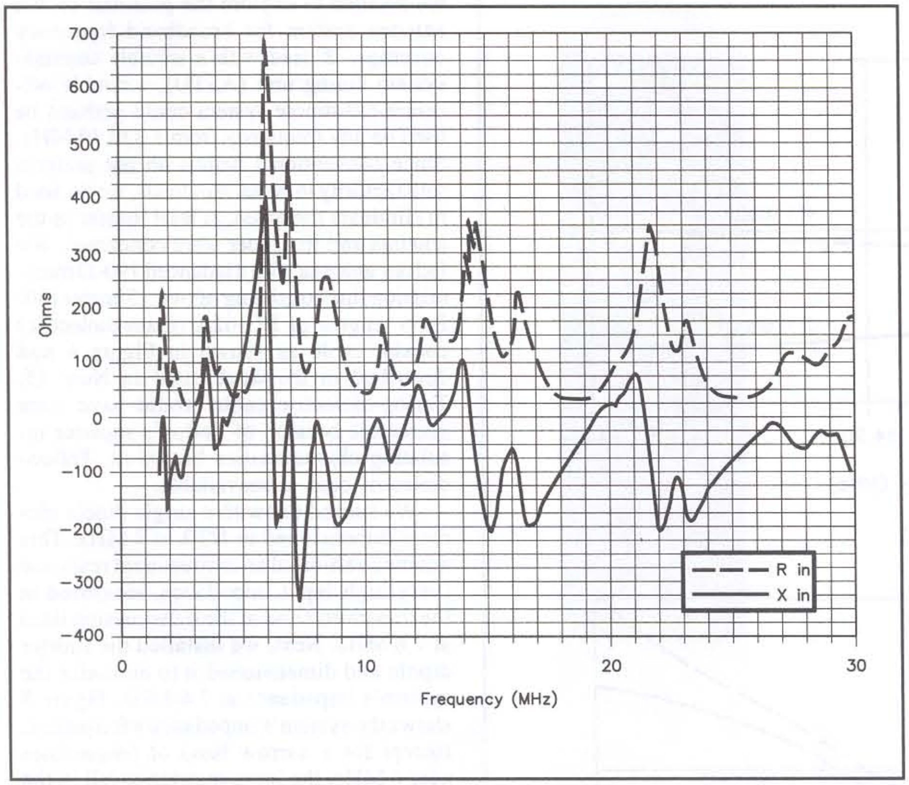

Figure 5—Impedance (input resistance [R in] and input reactance [X in]) vs frequency for the antenna shown in Figure 4.

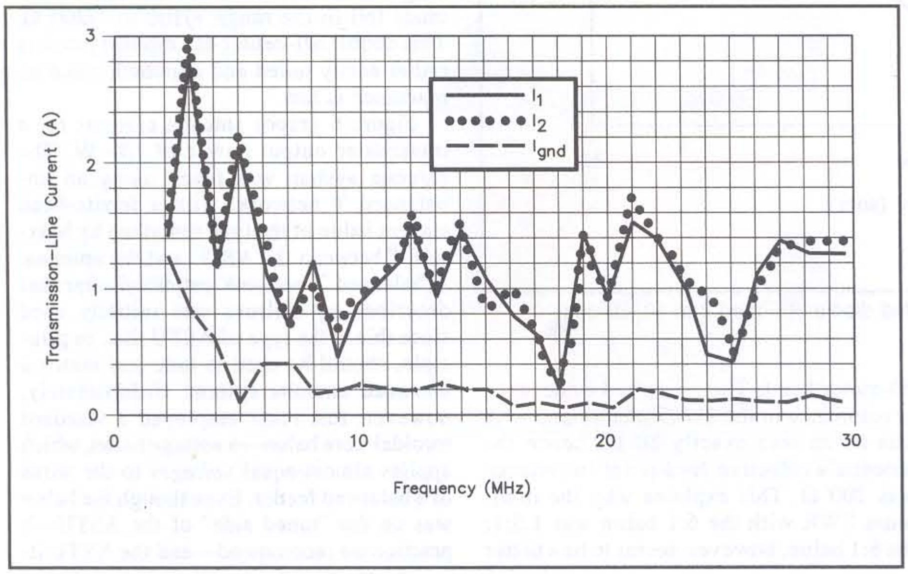

Figure 6—Transmission-line currents (peak values for 100-W transmitter output power) for the antenna shown in Figure 4, tuned by means of an unbalanced T network and a ferrite-bead-choke (W2DU) current balun.

self was balanced, we found that the transmission-line currents (I_1 and I_2 in Figure 4) differed greatly in amplitude at some frequencies. The difference may surprise you; it did us (see Figure 7).

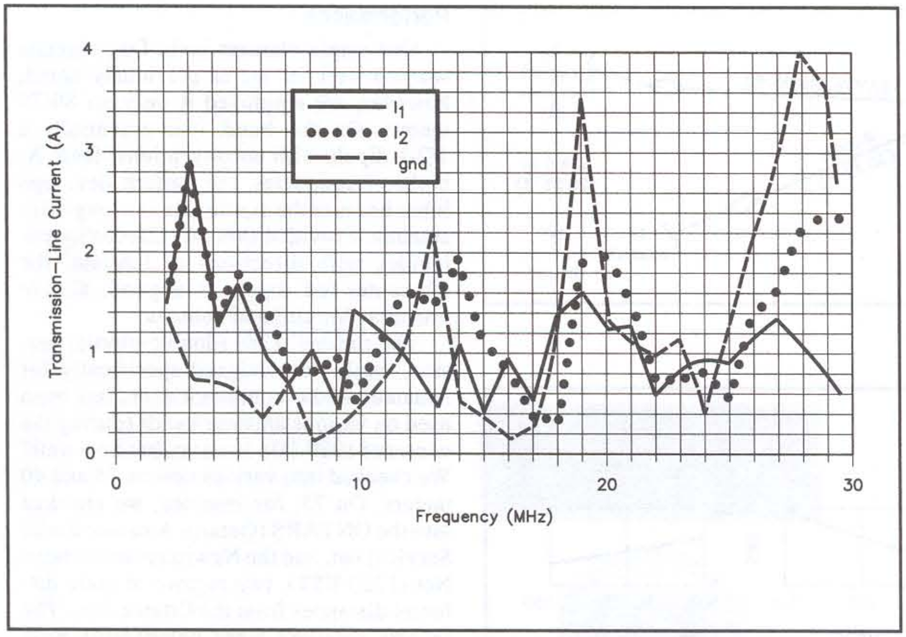

Figure 7—Transmission-line currents (peak values for 100-W transmitter output power) for the Figure 4 antenna, tuned by VE2CV's balanced tuner, which employed a voltage balun.

For an antenna that is asymmetrical with respect to its feeder–such as an off-center-fed dipole–a *current* balun must be used, since this type of balun forces almost equal *currents* into each conductor of the balanced line. This is necessary if transmission-line radiation is to be minimized. If these currents are exactly equal, there will be no difference in current flowing in the ground lead (I_{gnd}), which connects the braids of the coax to the tuner ground. Any current on this lead is then due to radiation-coupled current (I_3) induced on the *outside* surface of the coaxial shield. Clearly, we want this current to be small (to minimize radiation from the transmission line), and indeed it *is* small except at frequencies below 3 MHz (see Figure 6). (Our antenna was suspended from a bracket at the 12-meter [40-foot] level on a 21-meter [70-foot] aluminum-lattice mast. In the final analysis, reradiation by this mast should be considered because current may have been induced on the mast surface.)

Off-Center-Fed Dipoles for the Amateur Radio Experimenter

The off-center-fed dipole used by author Belrose in the 1940s was fed with 300-Ω twin lead via a balanced ASTU that was link coupled to a balanced (push-pull) power amplifier. There was no concern about balun losses and what type of balun to use because the system contained no balun.[20]

The authors' double-dipole, off-center-fed antenna employed a balanced coaxial feed line. This antenna was attractive for the authors because it could be used throughout the HF range if fed via a suitable matching network. If you decide to fabricate and try an off-center-fed dipole system, we suggested that you dimension it in accord with the Scholle and Steins versions, which are optimized for the amateur bands. Furthermore, we suggest that you feed such an antenna with a balanced transmission line. We have used dual RG-62A cables to make a balanced 190-Ω line, but one could use paralleled RG-63 (125-Ω) coaxial cable, which would make a balanced transmission line more in accord ($Z_o \approx 250$ Ω) with the traditional 300-Ω-twin-lead feeder. In such a system, the balun, and the ASTU (if required), can be in the shack to allow experimentation in achieving balanced current feed and reducing losses in the balun. Whatever method you use, feed your off-center-fed antenna via a current balun.

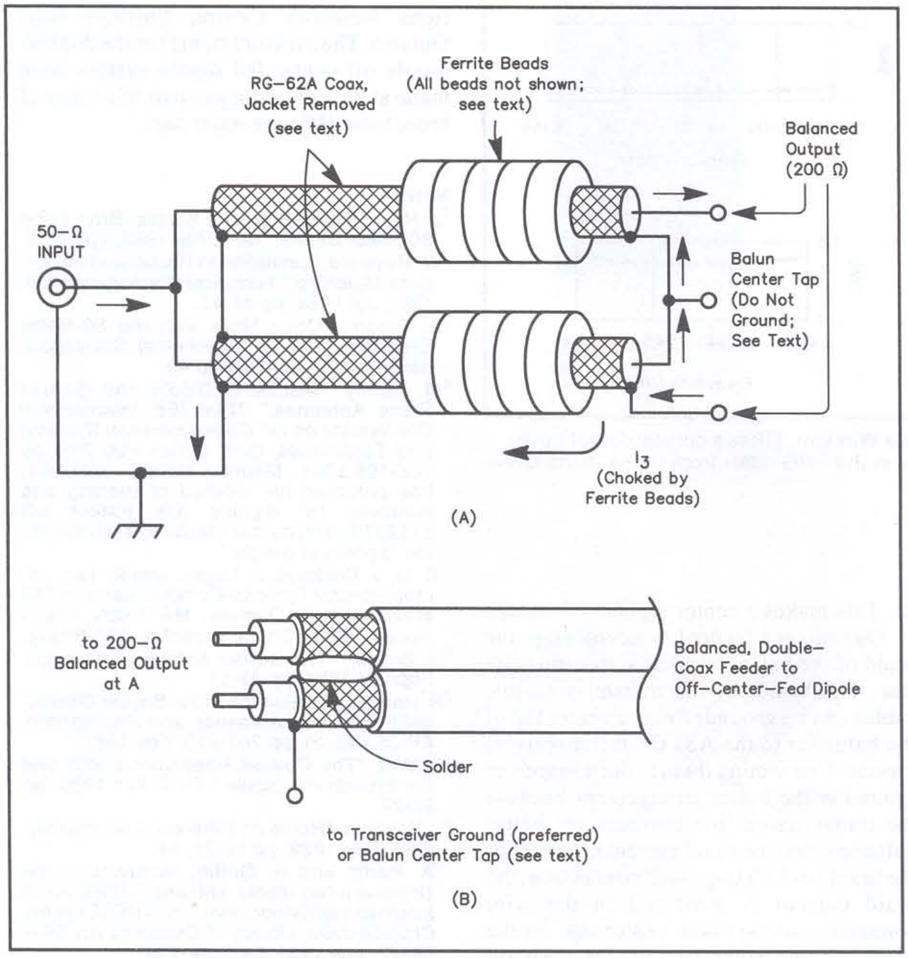

Figure 8—At A, a 4:1, ferrite-bead current balun; at B, connections to the balun end of the dual-coax-balanced feed line.

A shield-choke current balun can be constructed by slipping ferrite beads (–43 or –73 material) over a length of coaxial cable. (Depending on the beads obtainable and the diameter of the cable you use, you may need to remove the cable's outer jacket and install the beads directly around the

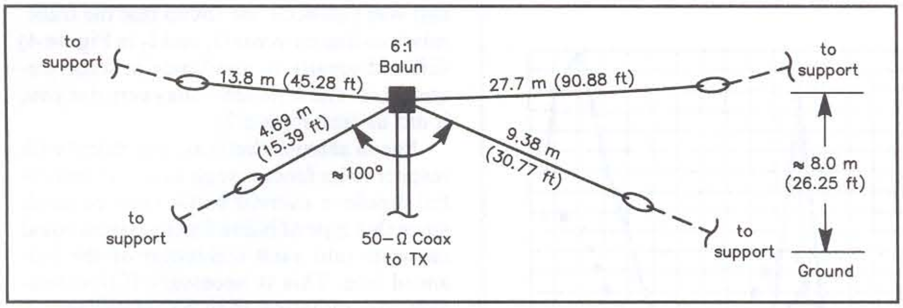

Figure A—Double Windom.

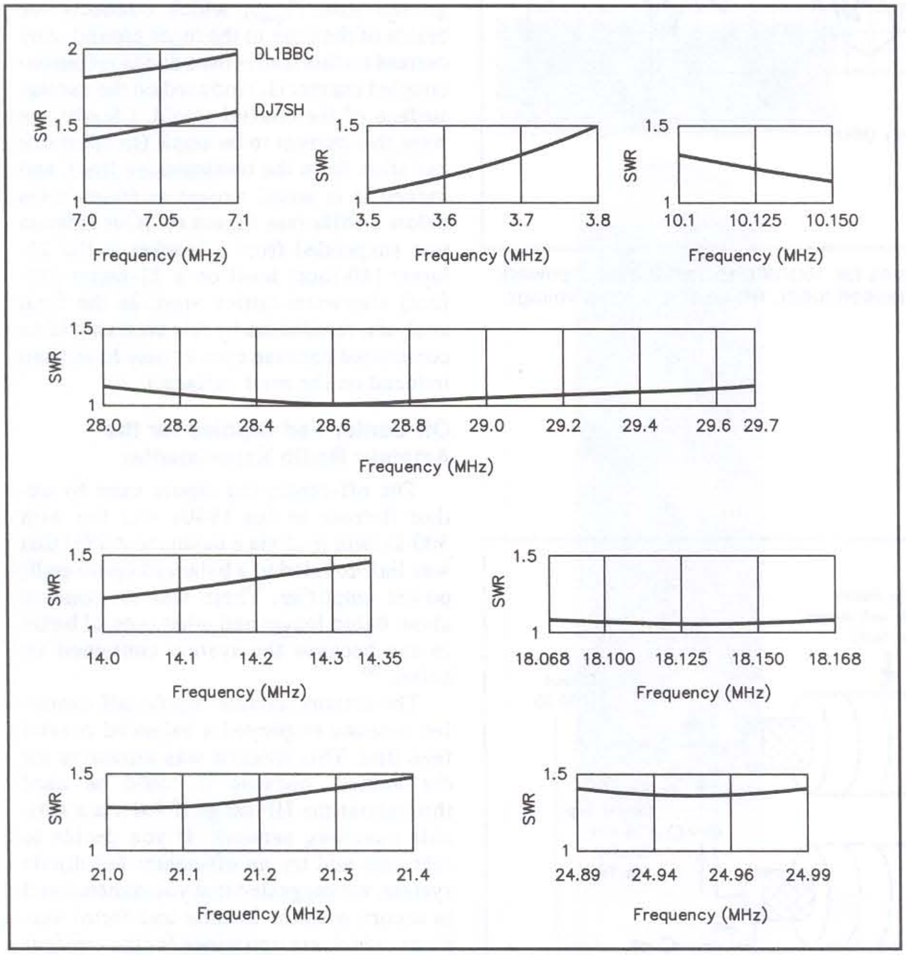

Figure B—SWR curves for the eight-band double Windom. [These curves do not cover all US amateur frequencies because allocations in the FRG differ from those in the US—*Ed.*]

shield.) As described in the work cited at Note 19, Walt Maxwell, W2DU, used 50 beads of no. 73 ferrite (Amidon no. FB-73-2401, Palomar FB-24-73 or equiv) on a piece of Teflon-dielectric cable to make a practical, low-loss, 1.8-to-30-MHz balun about 12 inches long.[21] Such a balun, however, is a 1:1 transformer—not very useful for the present application, where impedance transformation is also required. A 4:1 balun of similar type (see Note 20) can be constructed by using two equal lengths of RG-62 (95-Ω) coaxial cable, each fitted with ferrite beads. The inputs to these two coaxial-cable baluns are connected in parallel, and the outputs in series (see Figure 8). This makes a center-tapped 4:1 balun.

Our antenna feeder has a center tap: the braid of the balanced coaxial transmission line. The braids of the transmission-line cables can be grounded to the center tap of the balun, or to the ASTU (or transceiver) ground. Connecting them to the transceiver ground is the better arrangement because the transmission-line currents are better balanced, and the braid current is less. (For the braid-to-ASTU-ground connection, the braid current is measured in the wire connection to the balun center tap; in the braid-to-transceiver-ground case, the braid current is measured in the connection to the equipment ground.)

Performance

Our single-element Field Day antenna worked well for us; as previously noted, however, we employed it only on 80/75 meters. On this band, it is essentially a 1/2-λ dipole with unconventional feed. At higher frequencies, the pattern develops lobes because the dipole acts as a long-wire antenna. Provided that this directivity coincides with directions of interest, the off-center-fed dipole is a good, simple broadband/multiband antenna.

The authors' 1989 (double-dipole) version, while not designed specifically for Amateur Radio communication, has been used on various amateur bands (during the winter of 1989-90). It works, but how well? We checked into various nets on 75 and 40 meters. On 75, for instance, we checked into the ONTARS (Ontario Amateur Radio Service) net, and the Newfoundland Phone Net (1730 EST), two regions at quite different distances from the Ottawa area. The reports received were comparable with those given to other stations by net control. On the 160-meter band, we found that we could work stations we could hear, provided that they were running comparable power (100 W) and their local noise levels were reasonable. We have not yet determined the antenna's gain and pattern.

Acknowledgment

The authors work at the Communications Research Centre, Shirleys Bay, Ontario. The measurements for the double-dipole off-center-fed dipole system were made at the Laboratory as part of a study of broadband HF wire antennas.

Notes

[1] J. Hall, "The Search for a Simple, Broadband 80-Meter Dipole," *QST*, Apr 1983, pp 22-27.

[2] W. Hayward, "Limitation to Broadband Impedance Matching," Technical Correspondence, *QST*, Jul 1984, pp 45-47.

[3] A. Bloom, "Once More with the 80-Meter Broadband Dipole," Technical Correspondence, *QST*, Jun 1985, p 42.

[4] M. Hately, "Multiband Dipole and Ground Plane Antennas," *Third IEE International Conference on HF Communication Systems and Techniques*, Conf Series Pub 245, pp 102-106,1985. Maurice Hately, GM3HAT, has patented his method of feeding and matching HF dipoles (UK Patent GB 2112579), and he manufactures and markets his "dipoles of delight."

[5] S. Li, J. Rockwell, J. Logan, and D. Tan, *Microcomputer Tools for Communications* (610 Washington St, Dedham, MA 02026: Artech House, 1983), Ch 14: Broadband Matching.

[6] R. Snyder, "The Snyder Antenna," *rf design*, Sep/Oct 1984, pp 49-51.

[7] R. Hansen, "Evaluation of the Snyder Dipole," *IEEE Trans on Antennas and Propagation*, AP-35 (No. 2), pp 207-210, Feb 1987.

[8] F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

[9] L. Windom, "Notes on Ethereal Adornments," *QST*, Sep 1929, pp 19-22, 84.

[10] A. Parfitt and D. Griffin, "Analysis of the singlewire-fed dipole antenna," *IEEE AP-S International Symposium, Vol III* (IEEE cat no. CH2654-2/89, Library of Congress no. 89—84327, Jun 1989, pp 1344-1347.

[11] Such an arrangement is an extreme case because the feeder is perpendicular to the

ground—a configuration probably atypical of singlewire feed in Amateur Radio installations, and one that affords practically zero cancellation between feeder and ground currents. At the other extreme, one of the amateur systems described in Windom's "Notes on Ethereal Adornments" (see Note 9) employed a *1200-foot-long* (365.8-m) feed wire that likely more nearly approached parallelism with the ground. Clearly, "modeling a Windom" is a fuzzy procedure.–*Ed.*

[12] H. Scholle and R. Steins, "Eine Doppel-Windom Antenne für Acht Bänder," *cq-DL*, Sep 1983, p 427.

[13] H. Scholle, "Eine Doppel-Windom Antenne für Neun Bänder," *cq-DL*, Jul 1984, pp 332-333.

[14] J. Belrose, "Tuning and Constructing Balanced Lines," Technical Correspondence, *QST*, May 1981, p 43.

[15] J. Swail, VE3KF, and W. Loucks, VE3AR, private communications, Apr 1985.

[16] Antenna Engineering, Australia Pty Ltd, PO Box 191, Croydon, Victoria 3136, Australia.

[17] W. Wrigley, "Impedance Characteristics of Harmonic Antennas," *QST*, Feb 1954, pp 10-14.

[18] W. Orr, Ham Radio Techniques, *ham radio*, Jan 1983, pp 68-69.

[19] M. Maxwell, "Some Aspects of the Balun Problem," *QST*, Mar 1983, pp 38-40.

[20] J. Belrose, "The Balun Saga Continued," *QST*, in preparation.

[21] As a less-costly alternative, 10≈1-inch-long mix-77 beads—with an inner diameter to pass the coax used—*may* serve as a substitute. (The authors do not have experience with beads of this material, however.) Consult your ferrite-bead supplier on the availability of beads of this type.–*Ed.*

APPENDIX

A Double Windom Antenna for Eight or Nine Bands

This work, which originally appeared as two articles in *cq-DL*, was translated from the German by Dr. George Elliott Tucker, WA5NVI.

Part 1: A Double Windom Antenna for Eight Bands

By Hubert Scholle, DJ7SH, and Rolf Steins, DL1BBC

The asymmetrical dipole antenna developed and described by Windom (W8GZ) in 1929 has been used by many amateurs for many years as the FD4. This has also been the case in Germany.

We discovered in an older periodical (*QRV*) the explanation by F. Spillner (DJ2KY) that this antenna, with the addition of a small one-band Windom for 15 m, can be used as a five-band Windom. After the installation of the additional elements, this antenna worked very well for two years at DL1BBC.

With the opening of new bands (10, 18 and 24 MHz), the thought occurred to try out a new extension of the FD4 to eight bands (3.5 to 29.7 MHz).

What worked for 21 MHz must also be possible for 10 MHz.

So we took off the 21-MHz extension to my antenna and hung two elements of 4.69 and 9.38 m (15.39 and 30.77 ft), respectively, on the FD4 and stretched these downwards from insulators as an inverted V (Figure A).

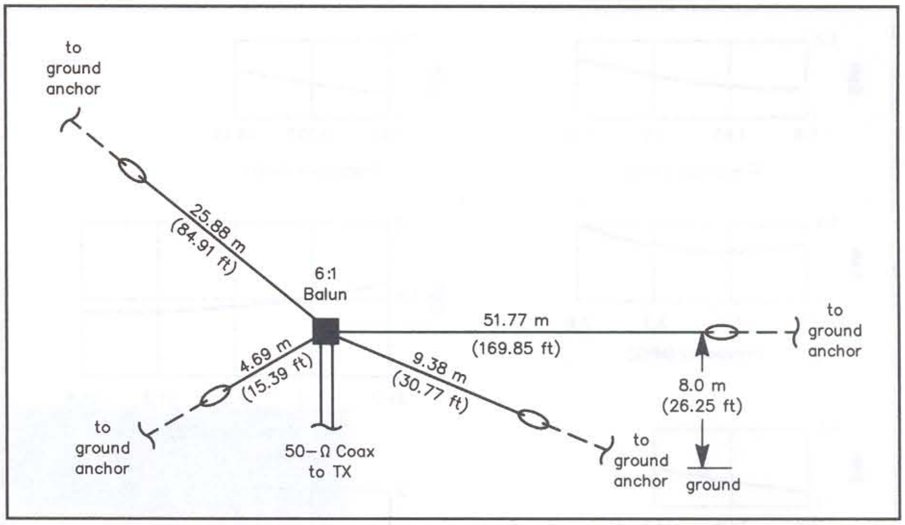

Figure C—A double Windom antenna for nine bands.

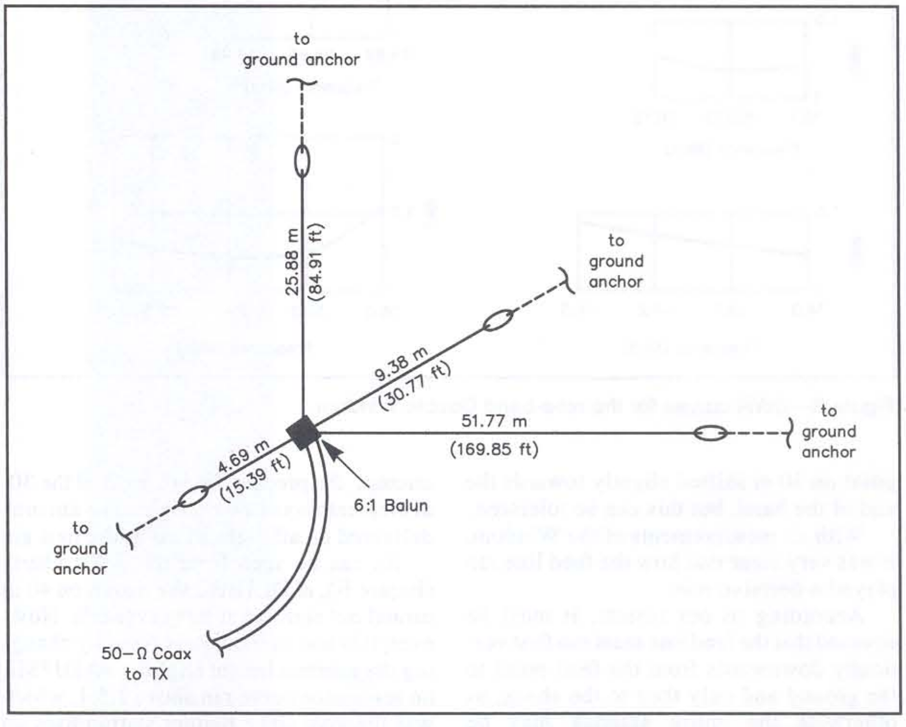

Figure D—Top view of the nine-band double Windom. [The elements are positioned to reduce coupling between the antenna's two off-center-fed dipoles.—*Ed.*]

To calculate the length we used the formula:

$$L/2 = 142.5 \div f \qquad \text{(Eq A)}$$

Whatever would work for 30 m should also work on 15 m.

As suspected, it worked.

As a by-product, it turned out in the measurements that this double Windom resonated just as well on 18 MHz and 24 MHz. So our eight-band Windom came into being with really simple means.

Construction

Thanks to our neighbors, we were able to extend the basic antenna (FD4) to its full length.

At DL1BBC it was installed about 6.9 m (22.63 ft) above the ground, rising to about 8 m (26.25 ft) at each support point. At DJ7SH it hung about 5 m (16.4 ft) above the ground and partly ran over a garage roof. Both extension legs were stretched downwards as an inverted V with an angle of about 100°. Changing this angle allows the whole antenna to be easily tuned during final adjustments.

After construction, the first measurements showed that because of the length of the 30-m elements, the 80, 40 and 20-m bands each had a resonance point that was shifted towards the low end of the band. This effect was eliminated by lengthening slightly the 30-m section, so the resonance points fell more in the middle of the bands.

With this adjustment, the resonance

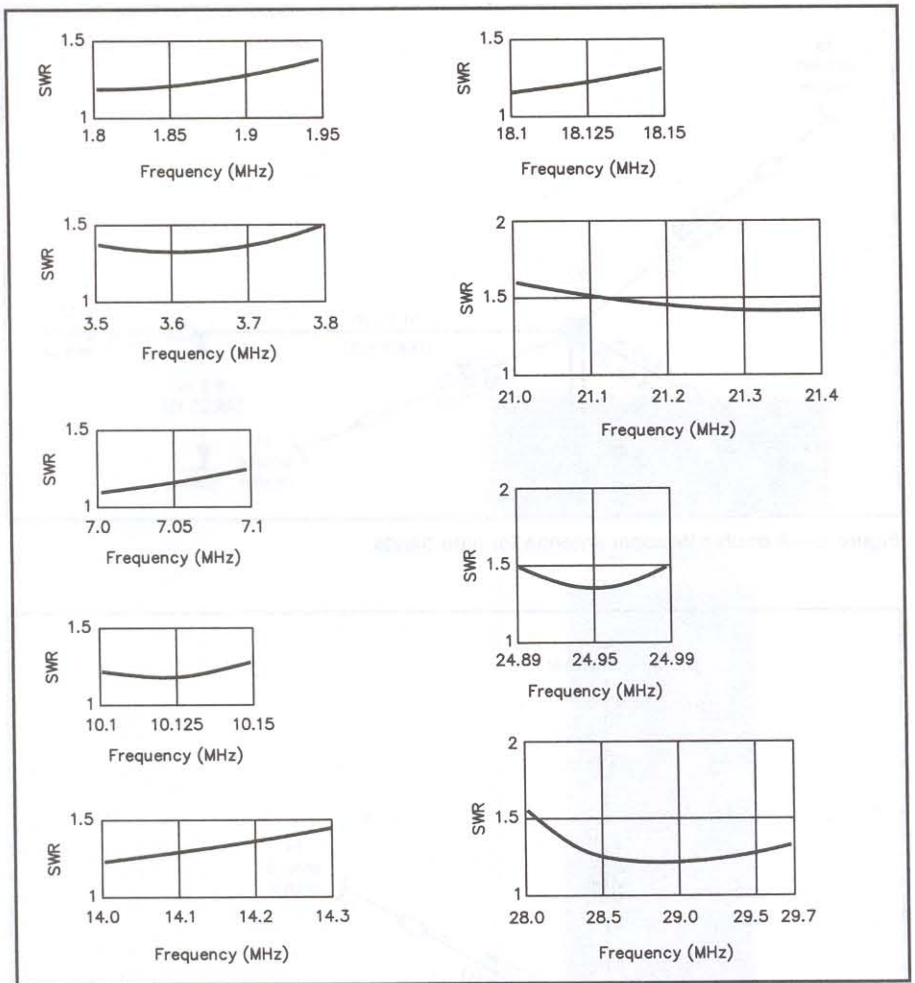

Figure E—SWR curves for the nine-band Double Windom.

point on 30 m shifted slightly towards the end of the band, but this can be tolerated.

With all measurements of the Windom, it was very clear that how the feed line ran played a decisive role.

According to our results, it must be stressed that the feed line must run first vertically downwards from the feed point to the ground and only then to the shack, as otherwise the entire antenna may be detuned. This is especially the case when the height of the antenna is under 10 m (32.8 ft). The 50-ohm-coax feed at DL1BBC was pulled through an old garden hose and then buried under the lawn.

The lower antenna height at DJ7SH had the result that, with the first construction attempt, the precalculated length of the 30-m elements was exactly right. The antenna delivered on all eight bands at the first go.

As can be seen from the SWR charts (Figure B), at DL1BBC the match on 40 m turned out somewhat less favorable. However, this was immediately fixed by changing the antenna height slightly. At DJ7SH, no resonance curve ran above 1.5:1, which was the goal since neither station uses an antenna tuner.

Performance

First contacts were made with both antennas. These showed that the antennas had a good degree of performance for a long wire. Especially the downwards sloping extension elements have a clear advantage over the horizontal basic antenna for DX.

With the first try on 30 m, many contacts were made with the US (East and West coasts), with signal reports between S6 and S7 while running 100 W.

At present, we cannot make a concrete statement about contacts within Europe.

This article makes no scientific claims, but intends to stimulate the long-wire enthusiast, and especially the friends of CW.

Part 2: Adding Another Band

Because the response was unexpectedly great to the publication of the above in *cq-DL*, we went to work again on an extension, as it was worthwhile to add 160 m.

With a half wavelength at 1.835 MHz, we calculated the basic length of the antenna to be 77.65 m (254.75 ft). We tapped the antenna at 25.88 m (84.9 ft) from one end and fed it with 50-ohm coax through a 6:1 balun. The basic antenna of this length was installed horizontally as a reclining L at DL1BBC. The additional elements, with lengths of 4.69 and 9.38 m (15.39 and 30.77 ft), were attached at the balun. This additional Windom for 10 and 21 MHz was again stretched downwards as an inverted V with an angle of about 100°. Here the additional Windom was mounted so that its elements were not extended in the same direction as those of the reclining L, which gave sufficient decoupling (Figures C and D).

For the feed, the Fritzel company made available for testing a new 6:1 balun, series 83, which can also handle high power. The SWR charts (Figure E) were obtained with the wire lengths given in the preceding paragraph. In case builders experience slight resonance shifts, these can be balanced out by lengthening or shortening the additional Windom.

Performance

First contacts were made with the antenna installed at DL1BBC. Here it was once again shown that the antenna has a good degree of performance for a long wire, especially for 1.8 and 3.6 MHz within Europe. The additional Windom again had the degree of performance described in the first part of this article.

The authors welcome questions and exchange of information. (When writing, please include return postage.)

From *QST*, May 1996 (Technical Correspondence)

Improved Feed for the Off-Center-Fed Dipole

By Richard A. Formato, K1POO

Theoretical data suggest that the commonly used feedpoint for the off-center-fed dipole (OCFD) may not be the best. The OCFD is an attractive multiband antenna because it's simple, inexpensive and requires no antenna tuner. Improving its performance simply by moving the feedpoint makes the antenna even more attractive. This note illustrates how the feedpoint influences antenna performance by analyzing computer-modeled SWR data for three different feedpoint locations.

The OCFD (shown schematically in Figure 1), consists of a single wire radiator of length L, fed off center a distance D from one end. The usual implementation uses a "1/3 feed," that is, the RF source is located one-third of the way from the end, so that D ≈ L/3. Why the feedpoint should be located there is not exactly clear. The ninth edition of *The ARRL Antenna Book,*[1] for example, observes that there is not much theoretical justification for this choice. Nevertheless, the 1/3-feed is accepted practice for building an OCFD.

Design details for a 1/3-feed three-band OCFD (80, 40 and 20 meters) appear in the 17th edition of *The ARRL Antenna Book.*[2] A 4:1 current balun at the feedpoint matches this antenna to any length of 50-Ω coax. More recently, Bill Wright, G0FAH, described a four-band, 1/3-feed OCFD (40, 20, 15 and 10 meters) fed with 300-Ω ladder line.[3] Matching 50-Ω coax requires a 4:1 balun on 40, 20 and 10 meters, and a 1:1 balun on 15 meters. Four-band operation, therefore, requires switching baluns. Another minor limitation is that the ladder-line length can be only an odd multiple of the wavelength at 21 MHz because the line is used as an impedance transformer. A simpler approach to achieving four-band operation is to feed the OCFD at a different point along its length.

Table 1

SWR Versus Feedpoint Placement (see Figure 1)

Feedpoint Distance D (m)	*SWR*
40 Meters	
8.65	2.5
6.98	2.1
3.65	2.0
20 Meters	
8.65	2.0
6.98	1.7
3.65	1.7
15 Meters	
8.65	1.2
6.98	21.3
3.65	1.9
10 Meters	
8.65	2.2
6.98	2.2
3.65	2.4

I computer-modeled a 21.03-meter (69-foot) long, 0.2053-cm-diameter (#12 AWG) OCFD in free space. The dimensions are the same as those in the G0FAH design. Free-space results are a good approximation for antennas high enough above the ground (typically a significant fraction of a wavelength). The band-center SWR was computed on 40, 20, 15 and 10 meters at the antenna-input terminals for a feed system impedance of 200 Ω. The theoretical values of input resistance and reactance were used to calculate SWR (the antenna was not assumed to be tuned). Because the feedpoint impedance is 200 Ω, a 4:1 balun is required to feed the antenna with 50-Ω coaxial cable. The results for three different feedpoints appear in Table 1.

For the conventional 1/3-feed (D = 6.98 m), the 40 and 10-meter SWR values are slightly over 2, while the 20-meter SWR is about 1.75. In marked contrast, the 15-meter SWR is off the scale (the actual value > 20). It is this behavior that makes a special feed system necessary on 15 meters, a complication that can be avoided by moving the feedpoint.

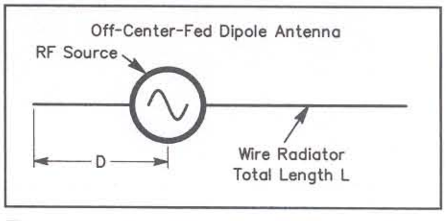

Figure 1—The basic off-center-fed dipole (OCFD).

When the OCFD feed is located 8.65 meters from one end, the 40, 20, and 10-meter SWRs are somewhat higher than they are with the 1/3 feed, but the 15-meter SWR is very low (≈ 1.2). Moving the feedpoint 1.67 meters closer to the antenna's center of the antenna results in a much better average SWR. And, more importantly, special matching is not required to achieve SWR ≤ 2.5 at the antenna terminals on all bands. Balun and coaxial cable losses, which are inevitable, reduce the SWR at the coax input to even lower levels. For most installations, it is probably reasonable to expect SWR at the transmitter to be less than 2 on all bands.

With the feedpoint located 3.65 meters from one end, the SWR on 40, 20 and 15 meters is excellent. The 40-meter SWR is only slightly above 2, and the 20 and 15-meter SWRs are below 2. The highest SWR occurs on 10 meters, where it's approximately 2.4. Because the SWR is reduced by feed-system losses, it will be less than 2.4 at the coax input. And, because balun and cable losses increase with frequency, the SWR reduction will be greatest on 10 meters where it is needed most. Feeding the antenna 3.65 meters from one end may well provide the best overall four-band performance.

In a specific implementation, the OCFD, like any antenna, must be tweaked for optimum SWR. This is accomplished by adjusting the feedpoint location. Other antennas, nearby metallic objects, and the earth are typical factors that influence antenna performance. Since these factors are not included in the computer model, they must be dealt with empirically by adjusting the antenna on-site. The data presented here provide a starting point for experimenting with different feedpoints. Depending on the total antenna length L, height above ground, earth electrical parameters, and feed system Z_0, it should be possible to operate a single OCFD on four or more bands without an antenna tuner or special feed arrangement. It's apparent that the OCFD's SWR varies dramatically as the feedpoint is moved, and that the commonly used 1/3-feed is not necessarily the best. Other feedpoints may therefore produce a better antenna.

---

[1] *The ARRL Antenna Book*, 9th Edition (Newington: ARRL, 1960), pp 191 to 192.

[2] R. Dean Straw, N6BV, Editor, *The ARRL Antenna Book*, 17th Edition (Newington: ARRL, 1994), pp 7-20 to 7-21.

[3] Bill Wright, G0FAH, "Four Bands, Off Center," *QST*, Feb 1996, p 65.

Dean Straw, N6BV, Senior Assistant Technical Editor, comments:

I've modeled what Richard describes and find that he's basically correct in his assertions. However, I would add a couple of caveats:

- The SWR across an individual band will vary. For example, the 40-meter SWR will rise above the level he describes at the ends of the band, because it is a rather large band, percentage-wise. This antenna is no different from an ordinary center-fed 40-meter dipole in that sense.
- The azimuth patterns for this OCFD will not be symmetrical, as is the case for any such off-center-fed antenna. This is nothing new, but should still be mentioned, particularly for those hopeful folks who want a "one-thing-to-all-people" antenna.
- The ham who tries this approach on 30, 17 or 24 meters may burn out the 200:50-Ω balun transformer. The SWR is very high indeed on those bands.

From *QST*, October 1996 (Technical Correspondence)

Off-Center-Fed Dipole Comments, Part 2

By Richard A. Formato, K1POO

In an earlier correspondence,[1] I suggested that the conventional "$^1/_3$-feed" used for the off-center-fed dipole (OCFD) is not the best choice. This letter provides additional information, and responds to comments made by Dean Straw, N6BV, at the end of that correspondence. Before addressing the comments, it is important to examine in more detail the SWR data which are the basis of my earlier letter. Figure 1 plots computer-modeled free-space SWR for the prototype 21.03-meter long, 0.0253-cm-diameter OCFD. SWR for a 200-Ω feed system impedance was calculated at the antenna-input terminals every 50 kHz for source frequencies from 5 to 30 MHz. Three feed-point locations were modeled, 3.65, 6.98 ($^1/_3$ feed), and 8.65 meters from the end of the antenna; the curves are labeled accordingly.

The most important features of the SWR curves are the locations of minima and maxima, and the corresponding SWR.

Table 1
SWR Minima

3.65 m		*6.98 m ($^1/_3$-feed)*		*8.65 m*			
Freq (MHz)	*SWR*	*Freq (MHz)*	*SWR*	*Freq (MHz)*	*SWR*	*Band (Meters)*	*Fc*
6.95	1.27	7.05	2.03	7.05	2.49	40	7.15
14.15	1.70	14.05	1.62	14.00	1.61	20	14.20
21.20	1.91	———	———	21.15	1.14	15	21.20
28.25	1.18	28.30	1.30	28.25	1.41	10	28.85

Table 1 lists minimum SWRs and the frequencies of the minima for the three feed points. For reference, the right-hand columns list the corresponding amateur band and its approximate center frequency.

The main point of my previous correspondence was that the 3.65 and 8.65-meter feeds *might* permit four-band operation without a matching network, because these feed points each result in *four* SWR minima between 5 and 30 MHz (see Figure 1). By contrast, the 6.98-meter SWR curve has only *three* minima in that range. The $^1/_3$-feed antenna cannot possibly operate on more than three bands without some sort of matching, which is why the GØFAH design[2] requires two baluns and a specific transmission line length for 15 meters.

How feasible four-band operation is depends only on where the minimum SWRs occur relative to the ham bands, and how low the SWR is across the bands. Table 1 shows that the prototype OCFD has SWR minima very close to the 40, 20, 15 and 10-meter bands. With some on-site tweaking,

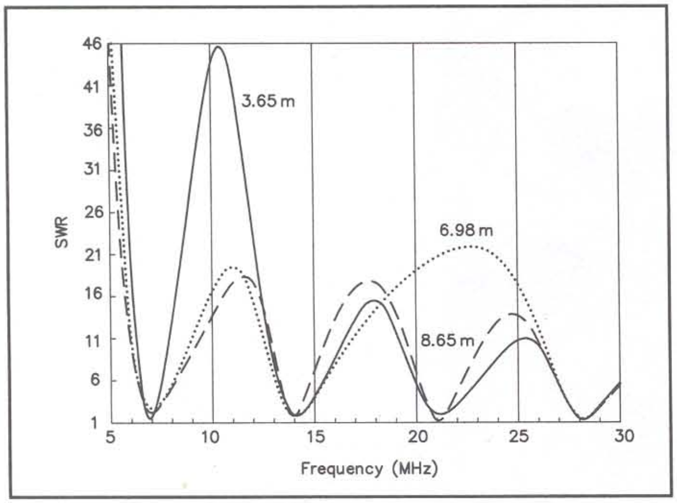

Figure 1—A plot of computer-modeled free-space SWR for the prototype 21.03-meter long, 0.0253-cm-diameter OCFD. SWR for a 200-Ω feed system impedance was calculated at the antenna-input terminals every 50 kHz for source frequencies from 5 to 30 MHz.

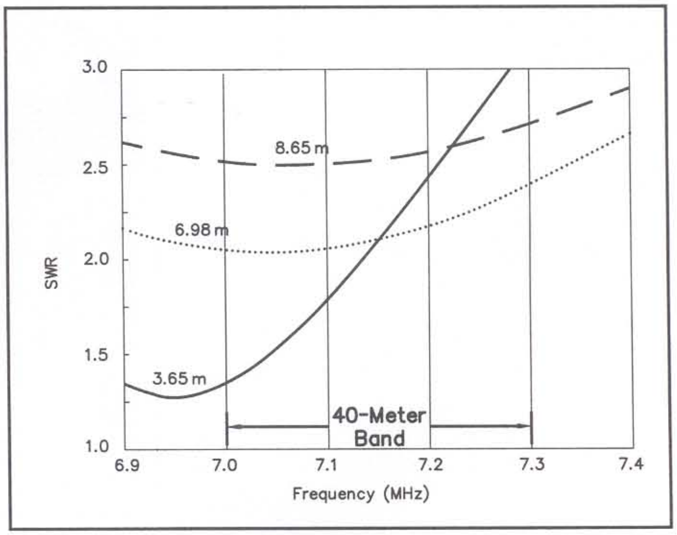

Figure 2—A plot of the prototype OCFD's SWR from 100 kHz below to 100 kHz above the 40-meter band in 5-kHz steps.

---

[1] Richard A. Formato, K1POO, "Improved Feed for the Off-Center-Fed Dipole," *QST*, May 1996, page 76.

[2] Bill Wright, GØFAH, "Four Bands, Off Center," *QST*, Feb 1996, page 65.

it is reasonable to expect that this antenna, fed either 3.65 or 8.65 meters from the end, could provide good SWR across all four amateur bands without a matching network.

The prototype OCFD dimensions were chosen because they correspond to the GØFAH design, thus permitting direct comparison with that antenna, not because they are optimized in any way. In fact, the prototype dimensions are *not* optimum for a four-band OCFD. In free space, the frequencies at which SWR minima occur, and the depth of the minima, are determined by three parameters: radiating-element length and diameter, and feed-point location. Changing any one of these changes both the frequencies of the SWR minima and the minimum SWR values.

Optimizing a four-band, free-space OCFD comes down to determining a set of antenna parameters that produces an acceptably low SWR (typically less than 2) across the 40, 20, 15, and 10-meter bands. Although I have not determined optimum parameter values, the very good predicted performance of the prototype antenna suggests that still better performance is almost certainly achievable. The purpose of my first letter was to encourage experimentation with OCFD designs, which would hopefully advance the state of the art by producing near-optimum designs.

Turning next to the comments at the end of my first correspondence, they are addressed as follows:

SWR Variation Across a Band

In terms of SWR behavior, the OCFD is much different from an ordinary center-fed dipole (CFD). The CFD is intentionally cut (tuned) to place minimum SWR within the band, which is why SWR increases toward the ends of the band (moving away from the minimum). An OCFD may or may not exhibit this behavior, depending on where its SWR minima occur. To illustrate, Figure 2 plots the prototype OCFD's SWR from 100 kHz below to 100 kHz above the 40-meter band in 5-kHz steps. The SWR does *not* increase toward each end of the band. If the SWR minimum is either outside or at one end of the band, as it is in Figure 2, the SWR will increase in one direction (up or down band), but decrease in the other, quite unlike an ordinary CFD. The OCFD's SWR will increase at both band edges *only* when its SWR minimum is inside the band, which generally is not the case.

Azimuth Pattern

The OCFD does indeed have an asymmetrical azimuth pattern, because the radiating element is not symmetrical about the feed point. But feeding the OCFD 8.65 meters from the end provides a *higher* degree of symmetry than the conventional $^1/_3$-feed. Locating the feed 8.65 meters from one end should result in a more symmetrical azimuth pattern than the conventional feed, not less. Of course, feeding the OCFD 3.65 meters from the end increases its asymmetry compared to the $^1/_3$-feed, so that the pattern for this implementation would be expected to be less symmetrical. Even so, pattern asymmetry is not necessarily undesirable. Many operators may want to take advantage of the OCFD's pattern by orienting the antenna to radiate in a preferred direction. This consideration applies to any antenna, even to the single-band CFD, which is an extremely poor radiator in the direction of the antenna axis at low to moderate take-off angles.

Operating the OCFD Out-of-Band

Regardless of where the feed is placed, or how it is implemented, certainly no attempt should be made to operate a 40, 20, 15, 10-meter OCFD on any other band. High SWR conditions may very well result in balun damage.

The data presented here provide additional insight into the advantages of feeding the OCFD at points other than $^1/_3$ of its length from the end. With computer-models for wire antennas widely available, it should be possible to optimize the OCFD in free space and over typical ground so that multiband operation is achievable without a matching network.

From *QST*, October 1996 (Technical Correspondence)

Off-Center-Fed Dipole Comments, Part 3

By Roy O. Hill, Jr, W4PID

QST's May 1996 Technical Correspondence column article "Improved Feed for The Off-Center-Fed Dipole" contained a statement that surprised me: "Why the feed point should be located [$^1/_3$ of the way from the end] is not exactly clear. The ninth edition of *The ARRL Antenna Book,* for example, observed that there is not much theoretical justification for this choice." It seems to me that there is plenty of theoretical justification. I am not an electrical engineer, and I am going to keep this simple. I won't say anything about computer simulations, ground effects, reactance, unbalanced feed-line currents, or adding 15-meter coverage; and I will use the *lowest* frequencies on the 80, 40, and 20-meter bands with full awareness that the frequencies actually used will be higher than that.

Thirty or 40 years ago, it was more or less generally accepted that the impedance of a half-wave antenna was about 4 kΩ at the ends, about 72 Ω at the center, and that the impedance along the antenna could be accurately represented by a straight line on semi-log graph paper. Using these premises, I plotted the impedance along a half-wave 3.5-MHz antenna on the graph of Figure 3. Then I added 7.0 MHz and 14.0 MHz plots for the same antenna.

All three lines cross at two points at about the 280-Ω impedance mark. These two points are $^1/_3$ of the way (within 0.1 percentage point as I measure it) from the ends of the antenna. The antenna can be fed with 300-Ω line at either of the triple-crossing points and it will work on all three frequencies. You could miss the exact crossing point (or change the frequency) somewhat and still have an impedance between 150 and 600 Ω, which would present an SWR less than 2[4]—an allowable value for most transmitters. All this was pretty widely known in the 1950s. What happened to it? Was all the old low-tech knowledge thrown out when the modern high-tech stuff came along?

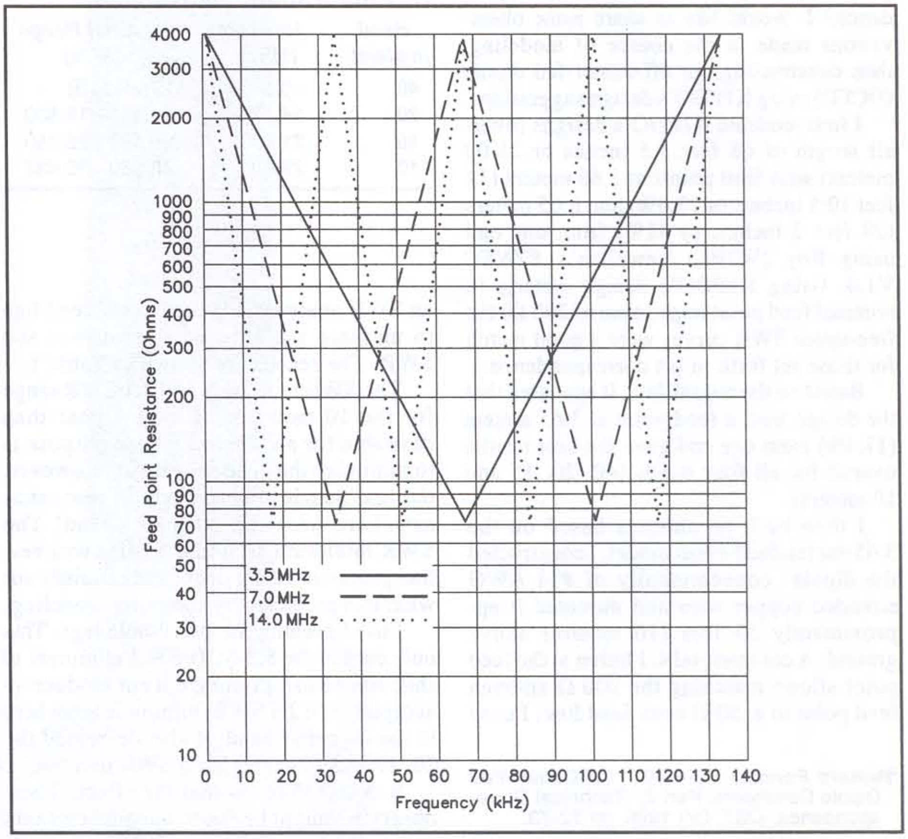

Figure 3—Impedance plot of an OCFD for 3.5, 7 and 14 MHz as a function of distance (in feet) from either end.

Also, the article mentioned that the "$^1/_3$-feed" antenna requires no antenna tuner. That is true, of course, for loading the antenna, but back in the days I've been talking about, the ARRL was stressing that you should *not* use a multiband antenna without an antenna tuner because of the danger of harmonic radiation. Do we not worry about harmonic radiation any more?

A little over 30 years ago, using as a guide a graph like the one shown in Figure 3, but using frequencies more like those on which I would be operating, I put up an antenna fed with 300-Ω ribbon 80 feet from one end and 44 feet from the other. I have used this antenna ever since. I do use an antenna tuner with it. The system works fine and causes no RFI or TVI. I used similar antennas for about 12 years before that, at other locations.

I just couldn't pass up that "not much theoretical justification" quoted from an ARRL publication without comment. The $^1/_3$ feed has, not only plenty of theoretical justification, but plenty of practical justification—from use.

---

[4] This is only accurate if the impedance is purely resistive. With reactive impedances, it is entirely possible to have an SWR higher than 2.—*Zack Lau, KH6CP*

From *QST*, September 1997 (Technical Correspondence)

Choke the OCFD

By Dale Gaudier M0AOP/K4DG

After reading with interest Richard (K1POO) Formato's technical correspondence,[3] I would like to share some observations made in the course of modeling, then constructing, an off-center-fed dipole (OCFD) using K1POO's design suggestions.

I first modeled K1POO's designs (overall length of 68 feet 5.5 inches or 21.03 meters) with feed points at 3.65 meters (11 feet 10.5 inches, or 17.4%) and 8.65 meters (28 feet 2 inches, or 41%) from one end using Roy (W7EL) Lewallen's *EZNEC* V1.0. Using K1POO's design criteria (a nominal feed point impedance of 200 Ω) the free-space SWR curves were a good match for those set forth in his correspondence.

Based on the model data, it appeared that the design with a feed point at 3.65 meters (17.4%) from one end gave the best results overall for all four bands (40, 20, 15 and 10 meters).

I then built an antenna based on the 3.65-meter-feed-point model. I constructed the dipole conventionally of #14 AWG stranded copper wire and mounted it approximately 33 feet (10 meters) above ground. A commercial 4:1 balun at the feed point allows matching the 200 Ω antenna feed point to a 50 Ω coax feed line. I used an SWR analyzer attached to the feed line to measure the antenna's resonance and SWR. The results are shown in Table 1.

Table 1
Results without the RF Choke

Band (meters)	*Minimum (MHz)*	*2:1 SWR Range (MHz)*
40	~8.5	(SWR >>3)
20	14.17	13.250 - 15.500
15	21.58	20.420 - 22.350
10	29.69	28.550 - 30.480

The SWR minimum and 2:1 SWR range for the 10 meter band were higher than desirable for an antenna whose purpose is to minimize the need for a tuner. However, the real puzzle was the lack of resonance anywhere near the 40 meter band! The SWR minimum around 8.5 MHz was neither pronounced nor deep, and certainly not what was predicted by computer modeling.

I tried pruning the two dipole legs. This only caused the 8.5-MHz SWR minimum to shift somewhat. Pruning did not produce an acceptable (< 2:1 SWR) minimum anywhere in the 40 meter band; it also decreased the 20, 15 and 10-meter band SWR minima.

It occurred to me that the effects I was observing might be due to unequal currents flowing at the feed point (possibly due to the asymmetrical design of the OCFD). I made a simple RF choke by coiling several turns of the RG-8X feed line close to the feed point and balun. After some initial adjustments of the choke diameter and number of turns, I had my 40-meter resonance! (See Table 2.) The coil is $5^{1}/_{2}$ turns of RG-8X with an inside diameter of $9^{1}/_{2}$ inches (24 cm). Note that the 40-meter band minimum and 2:1 SWR range can be raised by removing part of a turn from the RF choke.

Table 2
Results with the RF Choke

Band (meters)	*Minimum (MHz)*	*SWR Minima*	*2:1 SWR Range (MHz)*
40	7.00	1.1	6.850 - 7.220
20	14.10	1.2	13.520 - 14.420
15	21.22	1.1	20.420 - 22.340
10	28.34	1.4	27.460 - 29.050

As is apparent, the OCFD design of K1POO is surprisingly broadband, and with a little adjustment, will cover virtually all of the 40, 20, 15 and 10-meter bands with an SWR of 2:1 or less. However, an RF choke appears to be necessary for this design to work on its fundamental frequency.

---

[3] Richard Formato, K1POO, "Off-Center-Fed Dipole Comments, Part 2," Technical Correspondence, *QST*, Oct 1996, pp 72-73.

CHAPTER THREE

MULTIBAND ANTENNAS

3-1 **A Trap Collinear Antenna**
From August 1963 *QST*

3-3 **Antenna Traps of Spiral Delay Line**
From November 1972 *QST*

3-6 **Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters**
From July 1992 *QST*

3-8 **Trap Construction Information for Al (W8NX) Buxton's July 1992 Dipole**
From September 1992 *QST* (Technical Correspondence)

3-9 **Two New Multiband Trap Dipoles**
From August 1994 *QST*

3-13 **A Center-Fed "Zepp" for 80 and 40**
From May 1966 *QST*

3-16 **All-Band Antenna**
From March 1967 *QST* (Technical Correspondence)

3-17 **Multiband Antennas Using Decoupling Stubs**
From December 1960 *QST*

3-20 **"All-Band" Antenna**
From December 1954 *QST* (Hints & Kinks)

3-21 **Three-Band Matching System for a Forty-Meter Doublet**
From January 1975 *QST* (Hints & Kinks)

3-22 **A "Z" Antenna for the 10-160 Meter Bands**
From December 1979 *QST* (Technical Correspondence)

3-23 **Compact Multiband Antenna Without Traps**
From November 1981 *QST*

3-25 **Five Bands, No Tuner**
From June 1995 *QST*

3-26 **The NRY: A Simple, Effective Wire Antenna for 80 through 10 Meters**
From March 1993 *QST*

By Wesley M. Bell, W7QB **From *QST*, August 1963**

A Trap Collinear Antenna

Simple 3-Band Radiator with In-Phase Elements

This antenna covers the 15, 20, and 80-meter bands. On the two higher frequency bands the antenna operates with two extended half waves in phase thereby realizing some gain over the dipole operation of a conventional trap antenna.

As your *ARRL Handbook* tells you, broadside gain over a dipole approximately equivalent to doubling transmitting power may be obtained by using a center-fed antenna about 1¼ wavelengths long (extended double Zepp). Advantage of this is taken in the three-band trap antenna shown in Figure 1. The basic antenna is a dipole for 80 meters. The traps isolate sections of approximately 1¼ wavelengths for 20 and 15 meters. Since the center of a 1¼ wavelength wire is not at a current loop, wire is added in the form of a short open-wire feeder to make the total length about 1½ wavelengths, thereby bringing a current loop at the point where the system is fed by coax line. A balun is used to couple the unbalanced line to the balanced antenna system.

Trap Construction

The coil and capacitor specifications given under Figure 1 should be adequate for transmitters running at 100 watts input or less. For higher power, the inductance and capacitance values should be the same, but coils should be wound with heavier conductor, and capacitors should be of the transmitting type, such as the Centralab 850SL type. I made my own coils by wrapping a 2¼-inch form with waxed paper and winding the turns with double strands of No. 18 wire, unwinding one strand and cementing the remaining turns with strips of model-airplane glue. When the glue was dry, the completed coil was slipped off the form. At least one full extra turn should be wound to allow for pruning.

As shown in Figure 2, the capacitor is placed inside the coil, and the terminals of both capacitor and coil soldered to the heads of brass machine screws. These screws serve to hold the assembly central in a plastic waterproof container as shown in Figure 3. The container is the 1-pint size commonly found filled with chip dip, ice cream, potato salad, oysters and whatnot in grocery stores. The screws are fastened in the top cover and bottom of the container with nuts. The container is suspended from an insulator at the appropriate point in the antenna by short lengths of wire dropped from the insulator to the mounting screws where they are secured by a second set of nuts.

After mounting the traps in the containers, they should be resonated to the designated frequencies by carefully pruning the coils while checking with a grid-dip meter. The strain insulator with its wire wraps must be included, as shown in Figure 3, since the insulator capacitance is in parallel with the trap capacitor. (See *ARRL Handbook*.) As resonance is approached, final adjustment can be made by forming what is left of the last turn into a hairpin, and bend-

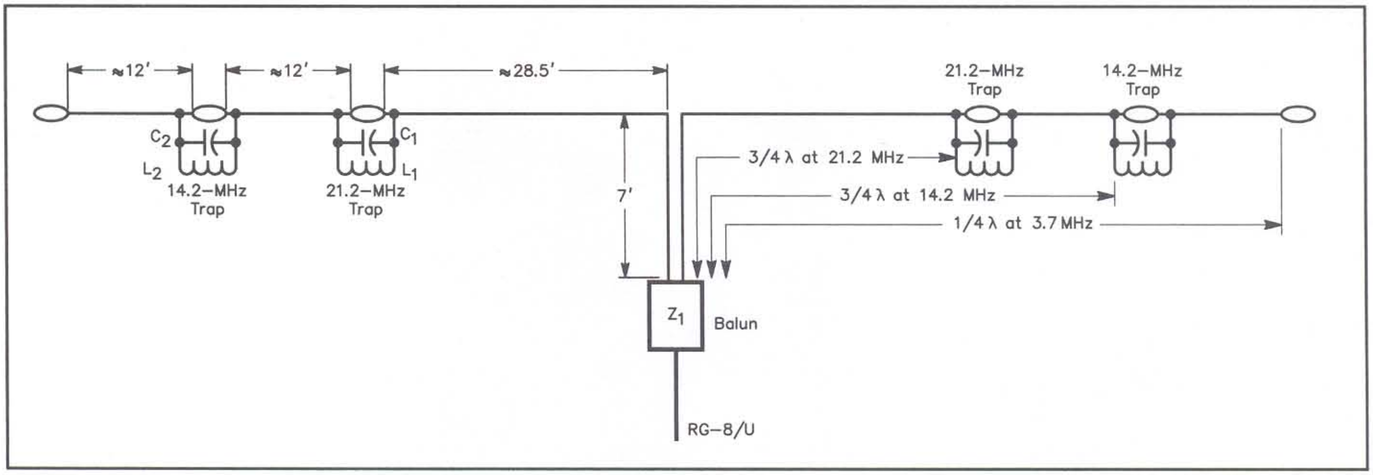

Figure 1—Sketch showing the approximate dimensions of the trap collinear. Wire lengths shown on left side are in terms of feet, while the approximate wavelength equivalents (electrical length) are shown on the right side. Frequencies and approximate wire lengths are for the centers of the three bands as a compromise for full-band coverage. It may be desirable to increase frequency for phone-only operation or lower frequency for c.w.-only operation.

C1, C2—25-pf. 6000-volt disk ceramic. See text.
L_1—Approx. 2 µH-4¾ turns No. 18, 2¼-inches diam., ⅜ inch long, or 5 turns No. 18, 2-inch diam., 16 tpi. See text.
L_2—Approx. 5 µH.—8 turns No. 18, 2¼-inch dia., ⅝ inch long, or 9 turns No. 18, 2-inch diam. 16 tpi. See text.
Z_1—1 to 1 balun.

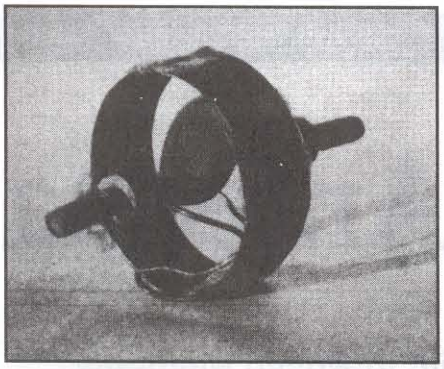

Figure 2—Trap components ready to be mounted in weather-proof container.

ing or twisting the hairpin to alter its inductive relationship to the main part of the coil.

Antenna Adjustment

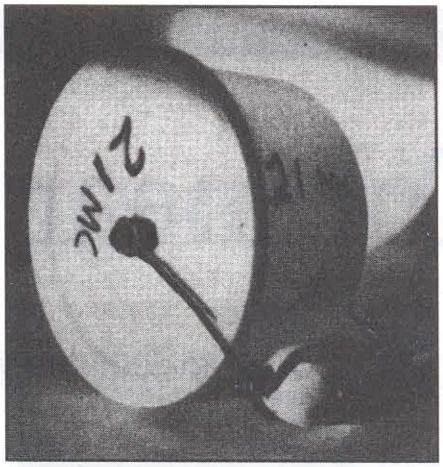

Figure 3—The completed 21-MHz trap and its supporting insulator ready for checking with a grid-dip oscillator.

Antenna resonance can be checked by shorting the ends of the 7-ft. open-wire line and coupling to a grid-dip meter. Initially, the wire lengths should be made a foot or so longer than the lengths shown in Figure 1. Start out with the 21-MHz sections only, anchoring the outer end of each wire section to one side of a 21-MHz trap insulator and connecting it to one side of the trap. An additional insulator should be attached temporarily between the other side of the trap insulator and the antenna-supporting rope. Then gradually shorten the wire until the grid-dip meter shows the desired resonant frequency.

Then add the second sections and traps and adjust similarly for the desired frequency in the 14-MHz band. The end sections of wire are then added and adjusted to show resonance at he desired frequency in the 3.5-MHz band.

In making the antenna adjustments, do not adjust the traps after they have once been set with the g.d.o.; change only the lengths of the wire sections. Devote plenty of time and patience to the adjustments. The job just can't be done correctly in a few minutes.

I use a center supporting pole, and the balun is enclosed in a weather-proof box mounted at the top of the pole.

Results with this antenna have been good. Using it in " inverted-vee" fashion, with one pole at the center and the ends attached to bushes, fences, clotheslines, or whatever else might be handy, I changed directions by simply walking the ends around to different positions. Without half trying, I worked 44 countries with a DX-60. I'm moving to the country soon where I plan to put up several of these antennas. 1 hope to add reflectors too. (Well, a guy can dream, can't he?)

By William J. Lattin, W4JRW **From *QST*, November 1972**

Antenna Traps of Spiral Delay Line

Most "traps" used in amateur radio multiband antennas are made of lumped inductance and capacitance in parallel. These consist of inductors made of coil stock of No. 12 or No. 14 wire and ceramic capacitors having voltage ratings up to 15,000 volts dc, which are relatively expensive. Vacuum capacitors would be the best, of course, but are also rather expensive.

Another type of trap has a capacitor made of two pieces of aluminum tubing arranged with a small-diameter tube inside a larger tube. Some have polystyrene dielectric, others air dielectric. The *ARRL Handbook* has a very complete description of these types.

Quarter-wave stubs of transmission line can be used for isolating sections of an antenna.[1] Loading coils can be used to modify the harmonic responses of a doublet to the second, third, fifth, seventh, and so on, to obtain a multiband antenna.[2] However, the use of loading coils is quite complicated if more than two bands of frequencies are desired. Traps tuned to the desired resonant frequencies make it much easier to adjust the lengths of the antenna sections, and also to obtain closer spacing between bands than can be obtained with loading coils.

Since quarter-wave sections of transmission line can be used as decoupling stubs for isolation of sections of an antenna, the idea occurred that perhaps quarter-wave sections of spiral delay line (SDL) might be used to make a very simple trap, without lumped capacitance. Spiral delay line is coaxial line with a helical inner conductor.

Construction of SDL Traps

Figure 1 shows the coil which is the helical inner conductor of the spiral delay line, along with the polystyrene tubing and end pieces for the coil, and the aluminum tube. The completed assembly is shown in the title photograph and Figure 2.

For a 28.5-MHz trap, a coil of No. 12 magnet wire was wound on a 1/2-inch rod, 37 turns, close wound. The coil was removed from the 1/2-inch rod and it sprang out to about 3/4-inch OD. The coil was 3 1/4 inches long. Enough wire was used to allow end wires straightened out to be 2 inches long on each end of the coil. Two pieces of 3/4-inch OD polystyrene rod cut 1/2-inch thick were drilled in the center to fit over the No. 12 wire ends and one was slipped over each end of the coil. This coil was placed inside a piece of 1-inch OD × 3/4-inch ID polystyrene tube 4 1/4-inches long, and a piece of 1 1/8-inch OD × 1-inch × 4 1/4-inch long aluminum tube slipped over this. The assembly was held together with No. 6-32 × 5/16-inch screws in holes drilled and tapped just far enough into the polystyrene end pieces to hold the screws in place. The short for this quarter-wave section of spiral delay line was made with a solder lug under one screw with a wire soldered between it and one end of the coil, as shown in the photographs.

The curve in the graph, Figure 3, shows the number of turns of No. 12 wire required for quarter-wave sections of the above construction. A close-wound coil of No. 12 magnet wire has approximately 12 turns per inch. The length of the assembly for a particular frequency can be determined approximately by dividing the number of turns on the coil by 12 to get the length in inches.

The traps were adjusted to frequency through the use of a grid-dip meter (checked on a receiver for accuracy of each frequency). The coil can be changed quite easily to the desired frequency by trimming turns if an extra turn or two is put on for this purpose. The coil can also be wound with spacing between turns and compressed or expanded to get the trap exactly on frequency.

After the assembly is completed and tuned to frequency, the coil can be sealed in the polystyrene tube with polystyrene ce-

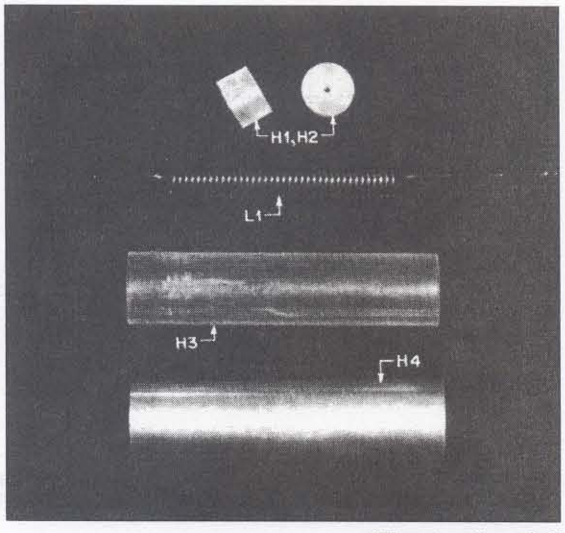

Figure 1 — The parts used in the make-up of a W4JRW SDL trap.

H1, H2 — End pieces of 1/2-inch length of 3/4-inch OD Polystyrene rod with center hole for No. 12 wire.

H3 — 1-inch OD × 3/4-inch ID polystyrene tube, length one inch greater than that of coil turns of L1

H4 — 1 1/8-inch OD × 1-inch ID aluminum tube, length equal to that of H3.

L1 — Close-wound coil of No. 12 enam. or magnet wire, 3/4-inch OD to fit inside H3. See Figure 3 and text for turns information.

[1] Lattin, "Multiband Antennas Using Decoupling Stubs," *QST*, December, 1960, p. 23.

[2] Lattin, "Multiband Antennas Using Loading Coils," *QST*, April, 1961, p. 43.

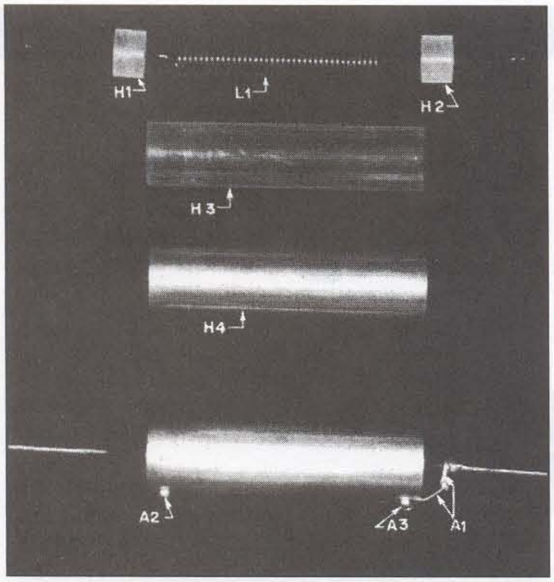

Figure 2—Assembly of the SDL traps. See Figure 1 and text for identification of parts not listed below.

A1—Solder lug and short length of wire assembled and soldered to short one end of H4 to one end of L1.

A2, A3—See text. The screw at A2 must not contact the wire of L1.

ment or coil dope. An inert gas could be sealed inside quite easily, but there seems to be no particular advantage to this.

Trap Ratings and Performance

The thickness of the polystyrene tube used was 1/8 or 0.125 inch. The average voltage rating for polystyrene is given in various handbooks as 500 volts per mil (.001 inch). This would be 62,500 volts for the thickness of 1/8 inch. The maximum power rating of these traps has not been determined. They have been used with a 2-kW PEP ssb transmitter without any failures from either voltage breakdown or heating. Larger wire, polystyrene tubing, and aluminum tubing can be used, of course, but the curve of Figure 3 will be different. Formulas for characteristics of spiral delay lines can be found in radio handbooks and textbooks in which this type of line is described.

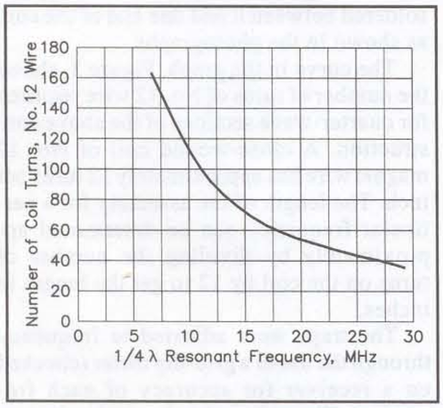

Figure 3—Resonant frequency of spiral-delay-line trap versus number of turns in coil.

The resonant impedance of the traps was measured and found to be approximately 100,000 ohms. For comparison, a trap made of No. 12 wire and a ceramic capacitor gave about the same resonant impedance. Several lumped-constant 15-meter traps borrowed from triband beams were measured and values from 16,000 to 28,000 ohms were found!

The SDL traps can be used in beam antennas, of course, with suitable mechanical modifications to fit the aluminum tube used in the beam. Spiral delay line can be used for other purposes, such as matching transformers, phasing, and any place where coax line is used but short dimensions are needed. The Zo (characteristic impedance) is a function of wire size, diameter, and spacing of the helical coil, dimensions of the insulator and external aluminum tube. Measurement with an rf bridge indicated a Zo of about 250 ohms for the construction used in these traps.

Figures 4 and 5 give the dimensions of two doublet antennas experimented with here. The antenna of Figure 4 has two 7.2-MHz traps and resonates at 3.9 and 7.2 MHz. The antenna of Figure 5 has eight traps, two each for 10 meters, 15 meters, 20 meters, and 40 meters. Resonances are at 3.9, 7.2, 14.3, 21.3, and 28.6 MHz. The flat-top portions were made of No. 12 solid copper wire. The feeder used was RG-8/U. No balun was used in our experiments. It was found that if a trap was not tuned exactly to frequency, it could still be used by changing the wire lengths in the antenna adjacent to the trap to get the desired antenna resonance. An antenna shortened and using traps is sharper in resonance than a full-length doublet. This is generally very well known, but perhaps bears repetition.

SWR curves are also shown in Figures 4 and 5 for these two antennas. During measurements, the antennas were supported in the center about 30 feet high and were 20 feet high at the ends. Measurements were made at the transmitter with 100 feet of RG-8/U coax between the transmitter and the antenna.

It is advisable to support doublet antennas at the center as well as at the ends, with strain relief at the ends – a simple arrangement of a screw eye, plastic rope, and a sash weight or a brick will do.[3] Since RG-8/U coax is fairly heavy, the center support is helpful to reduce the strain on the antenna. The breaking load of No. 12 copper wire is given in handbooks as 197.5 pounds for soft or annealed wire and 261.6 pounds for medium hard-drawn wire. In a high wind any type of support such as trees, towers, push-up masts, and so on, may move a few inches, putting thousands of pounds of tension on a wire stretched between them. Is there any ham who hasn't broken a wire antenna stretched between two trees when no strain reliefs were used?

One spiral-delay-line trap was tested with a hoist and concrete blocks for weight and didn't break at 200 pounds. As the No. 12 wire used in the antenna was softer than

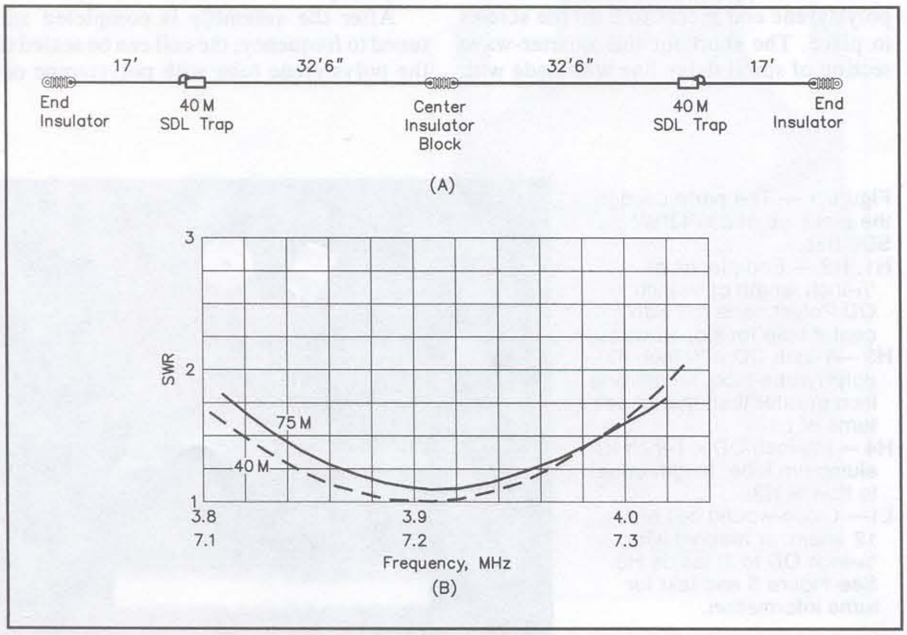

Figure 4—At A, dimensions for a 2-band spiral-delay-line antenna, resonant at 3.9 and 7.2 MHz (not drawn to scale). At B, the measured SWR values with this antenna.

---

[3] [EDITOR'S NOTE: To reduce wear and eventual breaking of the plastic-rope halyard, a pulley should also be used; large-diameter types sold in hardware stores as clothesline pulleys are economical and quite satisfactory.]

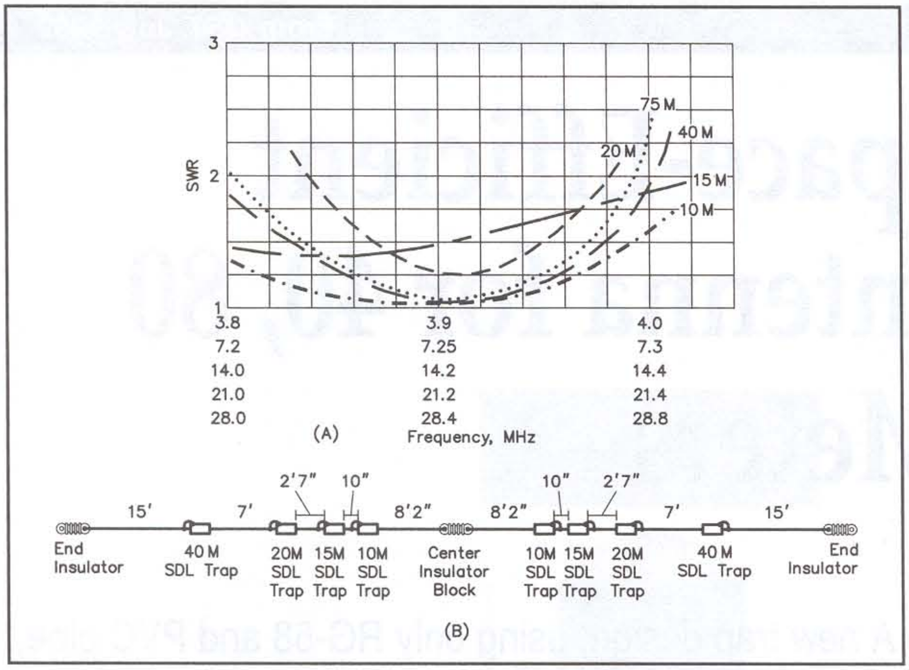

Figure 5—At A, the measured SWR values for a 5-band SDL antenna and at B the dimensions for this antenna.

the No. 12 magnet wires in the SDL traps, it appeared that the antenna wire would probably stretch before the wire in the traps. We have had the antenna up for almost two years, supported by three large trees, with strain reliefs at the ends. The antenna was not damaged at all by an 85-mph wind during a storm which bent the top section of our guyed crank-up tower into an inverted U shape with the triband beam hanging down. We did have one ice storm, but it wasn't severe enough to lift the strain-relief weights at the ends of the antenna.

If one desires, he can make the traps stronger by using two screws 180 degrees apart at each end, or even three screws at 120-degree spacing. Also the ends could be made of copper-weld wire soldered to the inner coil. The materials are not difficult to obtain. Most cities now have plastic supply distributors, and also aluminum tubing suppliers. The magnet wire can be obtained at a motor repair shop or from an electrical supply distributor.

By A. C. Buxton, W8NX **From *QST*, July 1992**

Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters

A new trap design, using only RG-58 and PVC pipe, yields better space efficiency than conventional coaxial traps.

These days more than ever before, many hams who want to work the low bands need an effective antenna that fits on a small lot. I'll show you how to build a shortened dipole for 160, 80 and 40 meters using improved coaxial-cable traps that I call Super Traps. The antenna, which covers the three ham bands below 7.3 MHz, is about the same length as a fullsize 80-meter dipole. If you install the antenna as an inverted V with a 90° included angle, the baseline length is 88 feet. The antenna uses traps that are easily constructed, rugged and weatherproof. They use no exposed capacitors or inductors.

You can feed the antenna directly with balanced 75-Ω line or via a 1:1 balun with either 50- or 75-Ω coaxial cable. Feed-line length is not critical. The antenna resonates at 1.865, 3.825 and 7.225 MHz. I installed such an antenna on my lot as an inverted V, with the apex 38 feet high and the ends at about 15 feet.

As part of this project, I developed a BASIC-language computer program[1] for trap design; a listing is available from the ARRL.[2] You can use this program to design these traps for frequencies of your choice, but you don't need a computer to make the antenna described here.

Figure 1 shows the antenna layout. The antenna is made of #14 stranded wire and two pairs of coaxial traps. Construction is conventional in most respects, except for the high inductance-to-capacitance (L/C) ratio that results from the unique trap construction. Two recent *QST* articles give tips on dipole construction and feeding.[3,4]

The traps use two-layer windings of the core (dielectric and center conductor) of RG-58 coaxial cable. Coaxial cable with flexible, rugged stranded-wire center conductors is preferable to that with a more brittle solid-wire center conductor. Figure 2 shows the traps. The 3.8-MHz trap is shown with the weatherproofing cover of electrical tape removed to show the construction details.

Precautions and Trap Specifications

With this trap-winding configuration, there are two thicknesses of core dielectric material between adjacent turns, which doubles the breakdown voltage of the traps. The transformer action of the two windings gives a second doubling of the trap-voltage rating. Thus, the trap voltage rating is 5.6 kV (four times RG-58's 1.4-kV rating). Conventional coaxial-cable traps made of RG-58 have a rating of 2.8 kV.

The 7-MHz traps have 33 μH of inductance and 15 pF of capacitance, and the 3.8-MHz traps have 74 μH of inductance and 24 pF of capacitance. The trap Qs are over 170 at their design frequencies, as measured on a Boonton Q meter.

These traps are suitable for operation at the 1-kW power level. When making the traps, *do not* use RG-8X or any other foam-dielectric cable. Winding such cables on small-diameter forms causes the center conductor to migrate through the dielectric toward the inside, decreasing the breakdown rating and compromising trap performance. The core diameter also differs from that of RG-58.

Construction

Although these traps are similar in many ways to other coaxial-cable traps, the shield winding of the common coax-cable trap has been replaced by an outer winding that fits snugly into the grooves formed by the inner layer. Capacitance is reduced to 7.1 pF per

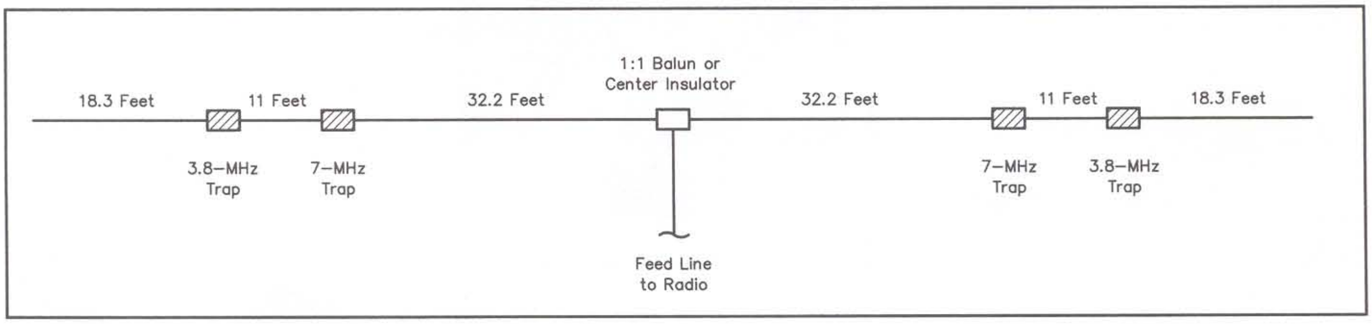

Figure 1—The shortened dipole resonates in the SSB portions of the 40, 80 and 160-meter bands. The antenna is 124 feet long.

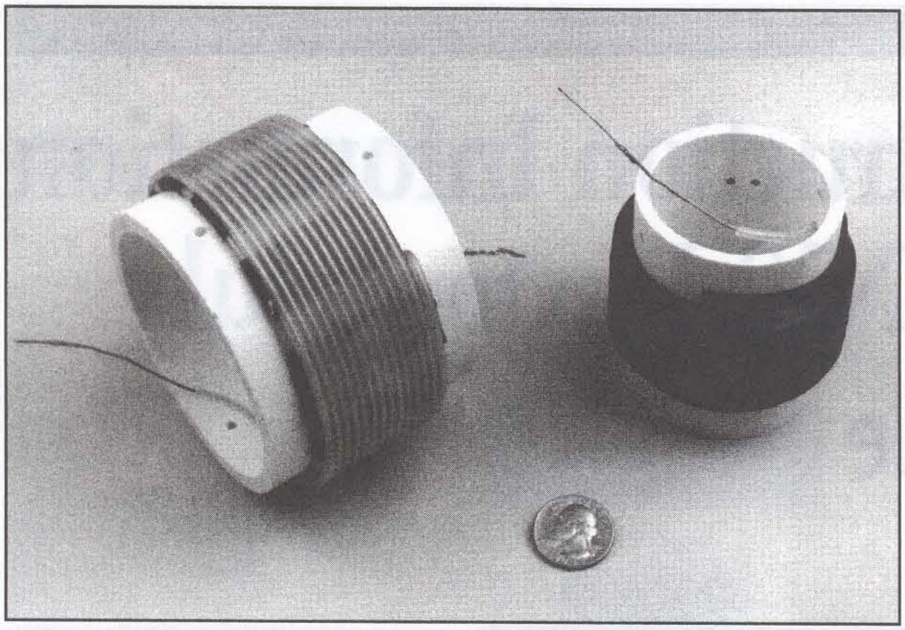

Figure 2—The improved coaxial-cable traps use two layered windings to provide an unusually high inductance-to-capacitance ratio, higher Q, and twice the breakdown voltage of single-layer traps. The 3.8-MHz trap is shown without its protective electrical-tape wrap to show the details of trap construction. This construction method makes for simple, lightweight, rugged and weatherproof traps.

foot, compared to 28.5 pF per foot with conventional coax traps made from RG-58. Trap reactance can be up to four times greater than that provided by conventional coax-cable traps.

The coil forms are cut from PVC pipe available at plumbing-supply stores. The 7-MHz trap form is made from 2-inch-ID pipe with an outer diameter of 2.375 inches. The 3.8-MHz trap form is made from 3-inch pipe with an outer diameter of 3.5 inches. The 7-MHz trap uses a 12.3-turn inner winding and an 11.4-turn outer winding. The 3.8-MHz trap uses a 14.3-turn inner winding and a 13.4-turn outer winding. All turns are closewound. The inner-trap frequency is 7.17 MHz and the outer-trap frequency is 3.85 MHz.

If you are unable to get PVC forms of exactly the same diameters as those called for here, compensate for the effect of form-size differences by taking advantage of the fact that the number of turns varies inversely with the form diameter. Thus, if the form diameter you use is, say, 5% larger than mine, reduce the number of turns by 5%. If necessary, add or remove fractions of a turn at the end of the outer winding. If you have a computer, you can use the BASIC programs[5] to calculate the exact number of turns for other form diameters. Stay as close as you can to the prescribed diameters because too much deviation changes the loading effect of the traps. A small change in trap loading may require a change in the lengths of the tip segments beyond the traps.

Use a #30 (0.128-inch diameter) drill for the feed-through holes in the PVC coil forms. The start and end holes of the 7-MHz traps are spaced 1.44 inches center to center, measured parallel to the trap center line. The holes in the 3.8-MHz traps are 1.66 inches apart. Wind the traps with a single length of coax core. The unspliced lengths are 17.55 feet for the 7-MHz traps and 28.45 feet for the 3.8-MHz traps. These lengths include the trap pigtails and a few inches for fine tuning.

Strip the jacket from the coax. This is easily done using a wood vise with wide jaws to hold the cable while cutting the jacket longitudinally with a sharp knife or razor. The coax outer conductor (braid) is best removed by pushing (not pulling) it off.

Use electrical tape to keep the turns of the inner-layer winding closely spaced during the winding process. This counteracts the tendency of the tension in the outer-layer winding to spread the inner–layer turns. Stick the tape strips directly to the coil form before winding and then tightly loop them over and around the inner layer before winding the outer layer. Use six or more tape strips for each trap.

If possible, check the resonant frequencies of your traps with a dip meter. Try to maintain an accuracy of 50 kHz or better.

For low-noise reception, erect the antenna as close to horizontal as possible. If you let the ends of the antenna droop toward the ground, as I have done with my inverted-V installation, you may have to accept a somewhat higher noise level in the interest of structural simplicity and reduced baseline length. Some feel that the inverted V configuration is better for DXing than a horizontal dipole at the same height. For an inverted V with a 90° included angle (legs that slope downward at 45°), you'll need a minimum apex height of about 55 feet and a baseline length of 88 feet. Get the apex as high as you can and keep the ends at least 10 feet above the ground for safety.

Configuration and Performance Trade-Offs

You seldom get something for nothing. This antenna proves no exception to that rule. As with all trap dipoles, this one has less-than-ideal bandwidth due to the loading effect of the traps. This is the price paid for multiband coverage and physical shortening. This antenna covers 65 kHz of 160 meters, 75 kHz of 80 meters and the entire 40-meter band with SWRs under 2:1. The bandwidth limitations on 160 and 80 meters can be largely offset with an antenna tuner.

It is also important to recognize that the traps are used in low-current portions of the antenna, minimizing I^2R trap losses. A relatively high radiation resistance is therefore also retained.

Good luck with your low-band antennas!

Notes

[1] I wrote the program in GWBASIC 3.2. It uses generic BASIC commands and can be easily converted for use with computers other than IBM PCs and compatibles.

[2] For a copy of the BASIC program, send a business-size SASE to the ARRL Technical Department Secretary, 225 Main St, Newington, CT 06111-1494. Request the July 1992 *QST* BUXTON BASIC PROGRAM listing.

[3] J. Healy, "Antenna Here is a Dipole," *QST*, Jun 1991, pp 23-26.

[4] J. Healy, "Feeding Dipole Antennas," *QST*, Jul 1991, pp 22-24.

[5] See note 2.

From *QST*, September 1992 (Technical Correspondence)

Trap Construction Information for Al (W8NX) Buxton's July 1992 Dipole

Several readers have written asking for clearer instructions on how to build the antenna traps described by A. C. Buxton, W8NX, in his July 1992 *QST* article, "Build a Space-Efficient Dipole Antenna for 40, 80 and 160 Meters." For those of you who didn't write, but aren't sure how to build the traps, here's what you need to know.

The trap-winding technique is deceptively simple. Each trap is simply comprised of one winding atop another, wound in the same direction. First, drill four holes in the form for wire entry and exit, as described in July *QST*. Starting from inside the form, pass the wire through the hole labeled "1" in Figures 1 and 2. Wind a layer of the inner conductor on the form. Then, run the wire end down through a second hole into the form (just below the one labeled "EXIT" in Figure 2) and, inside the form, bring the wire back to a point adjacent to where the first winding started. Bring the wire up through the hole marked "2" and wind another layer on top of the first, in the same direction as the first winding. Pass the wire down through the hole labeled "EXIT" and you're done. The tape you see in the photos, described in the article, helps hold the bottom winding's turns together and the top winding in place as you're assembling the trap.

The wires passing through holes in the form are strain-relieved, but use the two mounting holes you drilled in the form to support the trap, just to be on the safe side.
–Rus Healy, NJ2L

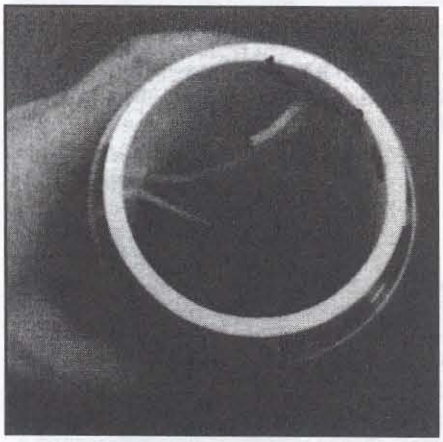

Figure 1—An inside view of a W8NX two-layer trap shows how the windings enter and exit the form. Two holes at each end of the form pass the windings into and out of the form.

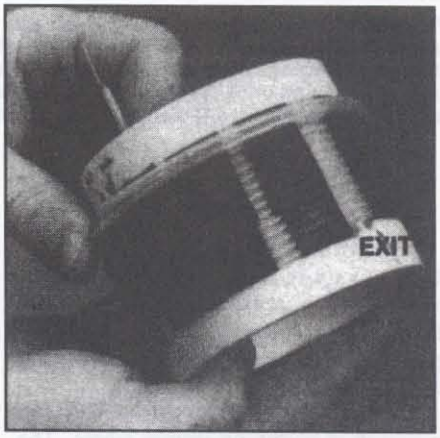

Figure 2—An outside view of a partially assembled W8NX trap. The bottom winding starts at hole "1" and reenters the form just below the "EXIT" hole. The wire then comes back through the inside of the form to hole "T," and is used to make a second winding atop the first. It reenters the form at the "EXIT" hole.

By Al Buxton, W8NX **From *QST*, August 1994**

Two New Multiband Trap Dipoles

W8NX details a new coax trap design used in two multiband antennas; one covering 80, 40, 20, 15 and 10 meters, and the other covering 80, 40, 17 and 12 meters.

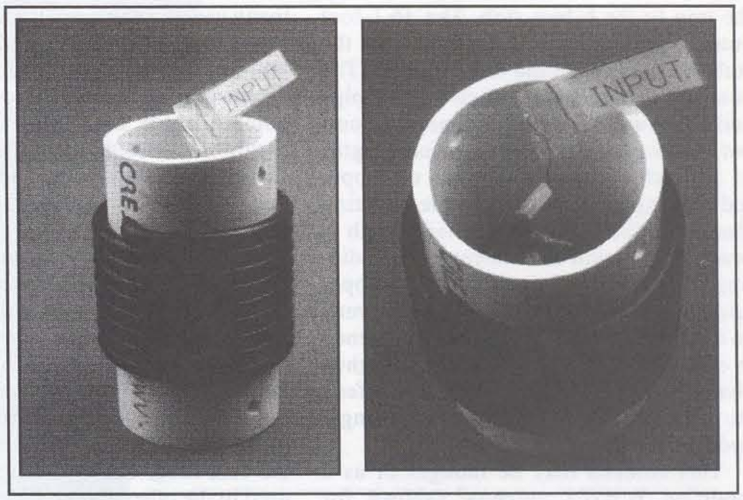

Over the last 60 or 70 years, amateurs have used many kinds of multiband antennas to cover the traditional HF bands. The availability of the 30, 17 and 12-meter bands has expanded our need for multiband antenna coverage. A fortunate few have the space and resources for multiband antennas like rhombics or long Vs, but many hams have employed inverted-L long wires or parallel dipoles. Old-timers will recall the offcenter-fed *Windom* of the '30s–the first version using a single-wire transmission line, and the later design using two-wire feed line. Over the years, random-length dipoles with open-wire feeders and associated tuners have been used successfully as multiband antennas. The G5RV multiband antenna is a specialized example of this approach.[1]

The *log periodic array* represents a kind of brute-force approach to the goal of achieving coverage of multiple HF ham bands. It seems inefficient because of the large gaps between our relatively narrow amateur HF bands.

Over the last few decades, two factors have affected the development of multiband antennas—the popularity of low-impedance (usually 50-Ω) coaxial feed lines, and the appearance of untuned, 50-Ω solid-state amplifiers. The impedance of an antenna is relatively low only at its fundamental frequency and at odd-order harmonics. Although antenna tuners are often necessary to resonate an antenna system, the quest for expanded multiband coverage with simple antennas continues.

At the end of the 1930s, a different technological approach appeared in the form of resonant traps in antennas. The *Mims Signal Squirter* is the grandfather of modern day tribanders.[2] This article discusses in detail an innovative trap design employed in two multiband dipoles.

One W8NX Trap Design—Two Multiband Dipoles

Two different antennas are described here. The first covers 80, 40, 20, 15 and 10 meters, and the second covers 80, 40, 17 and 12 meters. Each uses the same type of W8NX trap—connected for different modes of operation—and a pair of short capacitive stubs to enhance coverage. The new W8NX coaxial-cable traps have two different modes: a high- and a low-impedance mode. The inner-conductor windings and shield windings of the traps are connected in series in the conventional manner for both modes. However, either the low- or high-impedance point can be used as the trap's output terminal. For low-impedance trap operation, only the center conductor turns of the trap windings are used. For high-impedance operation, all turns are used, in the conventional manner for a trap. The short stubs on each antenna are strategically sized and located to permit more flexibility in adjusting the resonant frequencies of the antenna.

Figure 1 shows the configuration of the 80, 40, 20, 15 and 10-meter antenna. The radiating elements are made of #14 stranded copper wire. The element lengths are the wire span lengths in feet. These lengths do not include the lengths of the pigtails at the balun, traps and insulators. The 32.3-foot-long inner 40-meter segments are measured from the eyelet of the input balun to the tension relief hole in the trap coil form. The 4.9-foot segment length is measured from the tension relief hole in

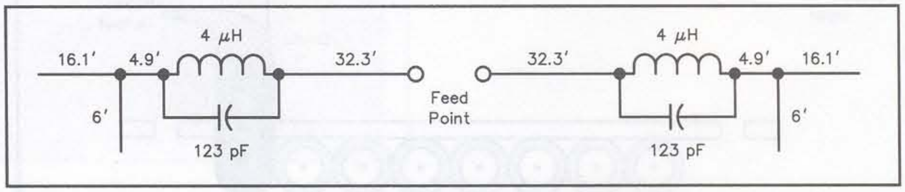

Figure 1—A W8NX multiband dipole for 80, 40, 20, 15 and 10 meters. The values shown (123 pF and 4 µH) for the coaxial-cable traps are for parallel resonance at 7.15 MHz. The low-impedance output of each trap is used for this antenna.

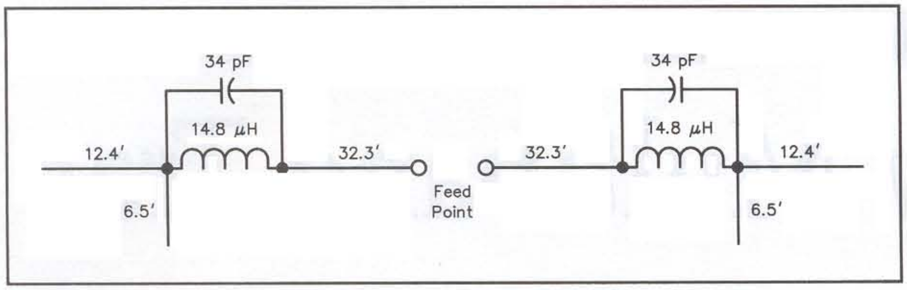

Figure 2—A W8NX multiband dipole for 80, 40, 17 and 12 meters. For this antenna, the high-impedance output is used on each trap. The resonant frequency of the traps is 7.15 MHz.

the trap to the 6-foot stub. The 16.1-foot outer-segment span is measured from the stub to the eyelet of the end insulator. The coaxial-cable traps are wound on PVC pipe coil forms and use the low-impedance output connection. The stubs are 6-foot lengths of 1/8-inch stiffened aluminum or copper rod hanging perpendicular to the radiating elements. The first inch of their length is bent 90° to permit attachment to the radiating elements by large-diameter copper crimp connectors. Ordinary #14 wire may be used for the stubs, but it has a tendency to curl up and may tangle unless weighed down at the end. I recommend that you feed the antenna with 75-Ω coax cable using a good 1:1 balun.

This antenna may be thought of as a modified W3DZZ antenna[3] (shown for many years in various ARRL publications) with the addition of capacitive stubs. The length and location of the stub give the antenna designer two extra degrees of freedom to place the resonant frequencies within the amateur bands. This additional flexibility is particularly helpful to bring the 15 and 10-meter resonant frequencies to more desirable locations in these bands. The actual 10-meter resonant frequency of the W3DZZ antenna is somewhat above 30 MHz, pretty remote from the more desirable low frequency end of 10 meters.

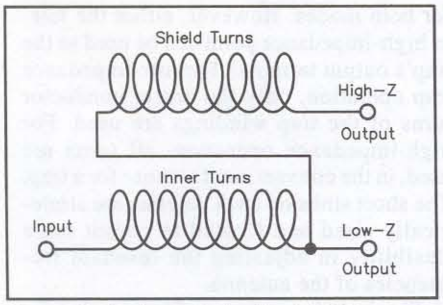

Figure 3—Schematic for the W8NX coaxial-cable trap. RG-59 is wound on a $2\frac{3}{8}$-inch OD PVC pipe.

Figure 2 shows the configuration of the 80, 40, 17 and 12-meter antenna. Notice that the capacitive stubs are attached immediately outboard after the traps and are 6.5 feet long, 0.5 foot longer than those used in the other antenna. The traps are the same as those of the other antenna, but are connected for the high-impedance output mode. Since only four bands are covered by this antenna, it is easier to fine tune it to precisely the desired frequency on all bands. The 12.4-foot tips can be pruned to a particular 17-meter frequency with little effect on the 12-meter frequency. The stub lengths can be pruned to a particular 12-meter frequency with little effect on the 17-meter frequency. Both such pruning adjustments slightly alter the 80-meter resonant frequency. However, the bandwidths of the antennas are so broad on 17 and 12 meters that little need for such pruning exists. The 40-meter frequency is nearly independent of adjustments to the capacitive stubs and outer radiating tip elements. Like the first antennas, this dipole is fed with a 75-Ω balun and feed line.

Figure 3 shows the schematic diagram of the traps. It explains the difference between the low and high-impedance modes of the traps. Notice that the high-impedance terminal is the output configuration used in most conventional trap applications. The low-impedance connection is made across only the inner conductor turns, corresponding to one-half of the total turns of the trap. This mode steps the trap's impedance down to approximately one-fourth of that of the high-impedance level. This is what allows a single trap design to be used for two different multiband antennas.

Figure 4 is a drawing of a cross-section of the coax trap shown through the long axis of the trap. Notice that the traps are conventional coaxial-cable traps, except for the added low-impedance output terminal. The traps are $8\frac{3}{4}$ close-spaced turns of RG-59 (Belden 8241) on a $2\frac{3}{8}$-inch-OD PVC pipe (schedule 40 pipe with a 2-inch ID) coil form. The forms are $4\frac{1}{8}$ inches long. Trap resonant frequency is very sensitive to the outer diameter of the coil form, so check it carefully. Unfortunately, not all PVC pipe is made with the same wall thickness. The trap frequencies should be checked with a dip meter and general coverage receiver and adjusted to within 50 kHz of the 7150 kHz resonant frequency before installation. One inch is left over at each end of the coil forms to allow for the coax feed-through holes and holes for tension-relief attachment of the antenna radiating elements to the traps. Be sure to seal the ends of the trap coax cable with RTV sealant to prevent moisture from entering the coaxial cable.

Also, be sure that you connect the 32.3-foot wire element at the start of the inner conductor winding of the trap. This avoids detuning the antenna by the stray capacitance of the coaxial-cable shield. The trap output terminal (which has the shield stray capacitance) should be at the outboard side of the trap. Reversing the input and output terminals of the trap will lower the 40-meter frequency by approximately 50 kHz, but there will be negligible effect on the other bands.

The title-page photos show a coaxial-cable trap. Details of the trap installation are shown in Figure 5. This drawing applies specifically to the 80, 40, 20, 15 and 10-meter antenna, which uses the low-impedance trap connections. Notice the lengths of the trap pigtails: 3 to 4 inches at each terminal of the trap. If you use a different arrangement, you must modify the span lengths accordingly. All connections can be made using crimp connectors rather than by soldering. Access to the trap's interior is attained more easily with a crimping tool than with a soldering iron.

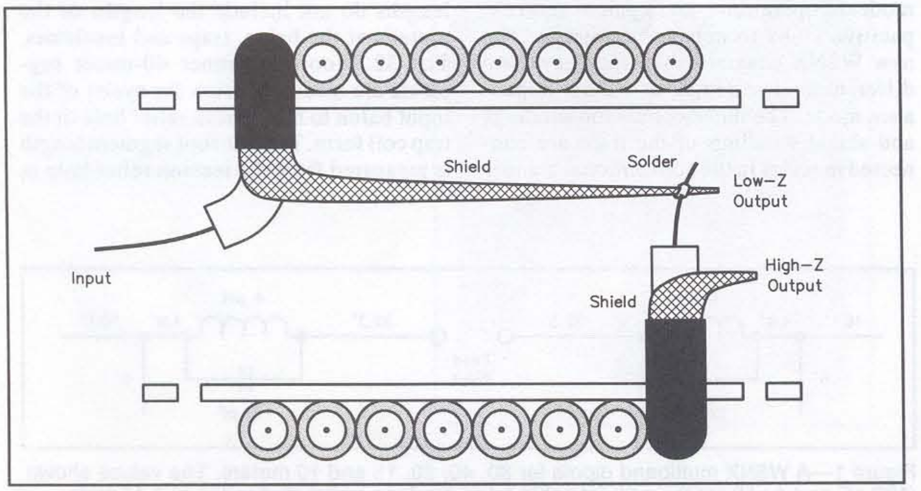

Figure 4—Construction details of the W8NX coaxial-cable trap.

Antenna Patterns

The performance of both antennas has

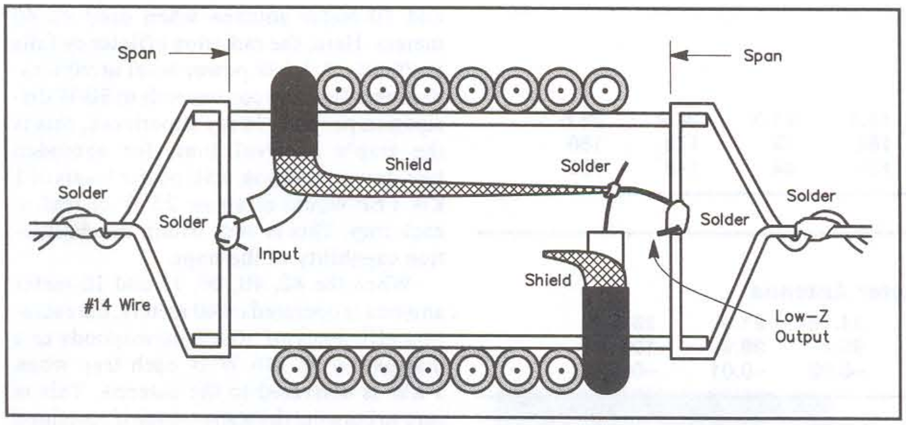

Figure 5—Additional construction details for the W8NX coaxial-cable trap.

been very satisfactory. I am currently using the 80,40, 17 and 12-meter version because it covers 17 and 12 meters. (I have a tribander for 20, 15 and 10 meters.) The radiation pattern on 17 meters is that of 3/2-wave dipole. On 12 meters, the pattern is that of a 5/2-wave dipole. At my location in Akron, Ohio, the antenna runs essentially east and west. It is installed as an inverted V, 40 feet high at the center, with a 120° included angle between the legs. Since the stubs are very short, they radiate little power and make only minor contributions to the radiation patterns. The pattern has four major lobes on 17 meters, with maxima to the northeast, southeast, southwest, and northwest. These provide low-angle radiation into Europe, Africa, South Pacific, Japan and Alaska. A narrow pair of minor broadside lobes provides north and south coverage into Central America, South America and the polar regions.

There are four major lobes on 12 meters, giving nearly end-fire radiation and good low-angle east and west coverage. There are also three pairs of very narrow, nearly broadside, minor lobes on 12 meters, down about 6 dB from the major end-fire lobes. On 80 and 40 meters, the antenna has the usual figure-8 patterns of a half-wave-length dipole. I have some pattern distortion and input impedance effects from aluminum siding on my house. Nevertheless, DX is easily workable on either of these antennas using a 100-W transceiver, when the high-frequency bands are open.

Both antennas function as electrical half-wave dipoles on 80 and 40 meters with a low SWR. They both function as odd harmonic current-fed dipoles on their other operating frequencies, with higher, but still acceptable, SWR. The presence of the stubs can either raise or lower the input impedance of the antenna from those of the usual third and fifth harmonic dipoles. Again, I recommend that 75-Ω, rather than 50-Ω, feed line be used because of the generally higher input impedances at the harmonic operating frequencies of the antennas.

The SWR curves of both antennas were carefully measured. A 75 to 50-Ω transformer from Palomar Engineers was inserted at the junction of the 75-Ω coax feed line and my 50-Ω SWR bridge. The transformer prevents an impedance discontinuity, with attendant additional undesired line reflections appearing at the 75 to 50-Ω junction. The transformer is required for accurate SWR measurement if a 50-Ω SWR bridge is used with a 75-Ω line. No harm is done to any equipment, however, if the transformer is omitted. Most 50-Ω rigs operate satisfactorily with a 75-Ω line, although this requires different tuning and load settings in the final output stage of the rig or antenna tuner. I use the 75 to 50-Ω transformer only when making SWR measurements and at low power levels. The transformer is rated for 100 W, and when I run my 1-kW PEP linear amplifier the transformer is taken out of the line. (I hope my absent-mindedness doesn't catch up with me some day!)

Figure 6 gives the SWR curves of the 80, 40, 20, 15 and 10-meter antenna. Minimum SWR is nearly 1:1 on 80 meters, 1.5:1 on 40 meters, 1.6:1 on 20 meters, and 1.5:1 on 10 meters. The minimum SWR is slightly below 3:1 on 15 meters. On 15 meters, the stub capacitive reactance combines with the inductive reactance of the outer segment of the antenna to produce a resonant rise that raises the antenna input resistance to about 220 Ω, higher than that of the usual 3/2-wavelength dipole. An antenna tuner may be required on this band to keep a solid-state final output stage happy under these load conditions.

Figure 7 shows the SWR curves of the 80, 40, 17 and 12-meter antenna. Notice the excellent 80-meter performance with a

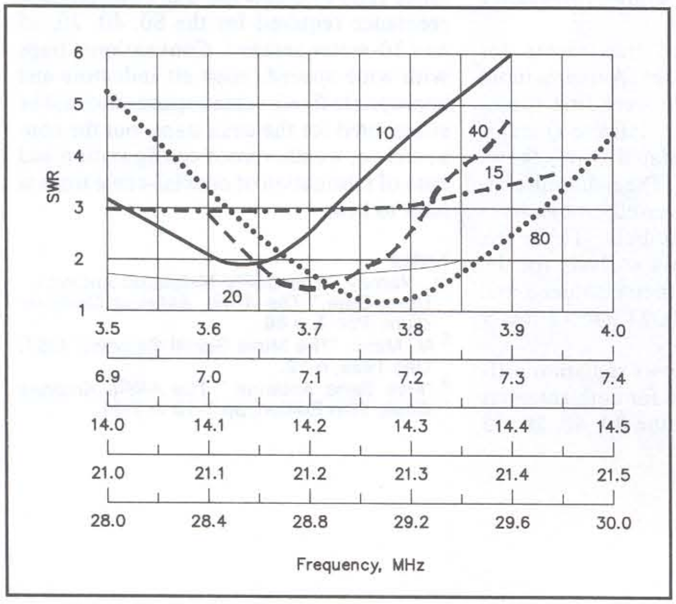

Figure 6—Measured SWR curves for an 80, 40, 20, 15 and 10-meter antenna, installed as an inverted-v with 40-ft apex and 120° included angle between legs.

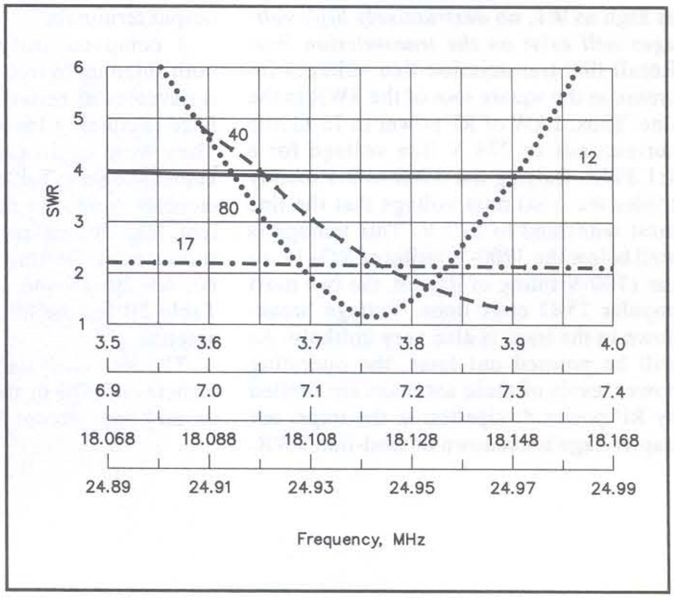

Figure 7—Measured SWR curves for an 80, 40, 17 and 12-meter antenna, installed as an inverted-v with 40-ft apex and 120° included angle between legs.

Table 1
Trap Q

Frequency (MHz)	3.8	7.15	14.18	18.1	21.3	24.9	28.6
High Z out (Ω)	101	124	139	165	73	179	186
Low Z out (Ω)	83	103	125	137	44	149	155

Table 2A
Trap Loss Analysis: 80, 40, 20,15, 10-Meter Antenna

Frequency (MHz)	3.8	7.15	14.18	21.3	28.6
Radiation Efficiency (%)	96.4	70.8	99.4	99.9	100.0
Trap losses (dB)	–0.16	–1.5	–0.02	–0.01	–0.003

Table 2B
Trap Loss Analysis: 80, 40, 17, 12-Meter Antenna

Frequency (MHz)	3.8	7.15	18.1	24.9
Radiation Efficiency (%)	89.5	90.5	99.3	99.8
Trap losses (dB)	–0.5	–0.4	–0.03	–0.006

nearly unity minimum SWR in the middle of the band. The performance approaches that of a full-size 80-meter wire dipole. The short stubs and the very low inductance traps shorten the antenna somewhat on 80 meters. Also, observe the good 17-meter performance, with the SWR being only a little above 2:1 across the band.

But notice the 12-meter SWR curve of this antenna, which shows 4:1 SWR across the band. The antenna input resistance approaches 300 Ω on this band because the capacitive reactance of the stubs combines with the inductive reactance of the outer antenna segments to give resonant rises in impedance. These are reflected back to the input terminals. These stub-induced resonant impedance rises are similar to those on the other antenna on 15 meters, but are even more pronounced.

Too much concern must not be given to SWR on the feed line. Even if the SWR is as high as 9:1, *no destructively high voltages will exist on the transmission line.* Recall that transmission-line voltages increase as the square root of the SWR in the line. Thus, 1 kW of RF power in 75-Ω line corresponds to 274 V line voltage for a 1:1 SWR. Raising the SWR to 9:1 merely triples the maximum voltage that the line must withstand to 822 V. This voltage is well below the 3700-V rating of RG- 11, or the 1700-V rating of RG-59, the two most popular 75-Ω coax lines. Voltage breakdown in the traps is also very unlikely. As will be pointed out later, the operating power levels of these antennas are limited by RF power dissipation in the traps, not trap voltage breakdown or feed-line SWR.

Trap Losses and Power Rating

Table 1 presents the results of trap Q measurements and extrapolation by a two-frequency method to higher frequencies above resonance. I employed an old, but recently calibrated, Boonton Q meter for the measurements. Extrapolation to higher frequency bands assumes that trap resistance losses rise with skin effect according to the square root of frequency, and that trap dielectric losses rise directly with frequency. Systematic measurement errors are not increased by frequency extrapolation. However, random measurement errors increase in magnitude with upward frequency extrapolation. Results are believed to be accurate within 4% on 80 and 40 meters, but only within 10 to 15% at 10 meters. Trap Q is shown at both the high- and low-impedance trap terminals. The Q at the low-impedance output terminals is 15 to 20% lower than the Q at the high-impedance output terminals.

I computer-analyzed trap losses for both antennas in free space. Antenna-input resistances at resonance were first calculated, assuming lossless, infinite-Q traps. They were again calculated using the Q values shown in Table 1. The radiation efficiencies were also converted into equivalent trap losses in decibels. Table 2A summarizes the trap loss analysis for the 80, 40, 20, 15 and 10-meter antenna and Table 2B for the 80, 40, 17 and 12-meter antenna.

The loss analysis shows radiation efficiencies of 90% or more for both antennas on all bands except for the 80, 40, 20, 15 and 10-meter antenna when used on 40 meters. Here, the radiation efficiency falls to 70.8%. A 1-kW power level at 90% radiation efficiency corresponds to 50-W dissipation per trap. In my experience, this is the trap's survival limit for extended key-down operation. SSB power levels of 1 kW PEP would dissipate 25 W or less in each trap. This is well within the dissipation capability of the traps.

When the 80, 40, 20, 15 and 10-meter antenna is operated on 40 meters, the radiation efficiency of 70.8% corresponds to a dissipation of 146 W in each trap when 1 kW is delivered to the antenna. This is sure to burn out the traps—even if sustained for only a short time. Thus, the power should be limited to less than 300 W when this antenna is operated on 40 meters under prolonged key-down conditions. A 50% CW duty cycle would correspond to a 600-W power limit for normal 40-meter CW operation. Likewise, a 50% duty cycle for 40-meter SSB corresponds to a 600-W PEP power limit for the antenna.

I know of no analysis where the burnout wattage rating of traps has been rigorously determined. Operating experience seems to be the best way to determine trap burn-out ratings. In my own experience with these antennas, I've had no traps burn out, even though I operated the 80, 40, 20, 15 and 10-meter antenna on the critical 40-meter band using my AL-80A linear amplifier at the 600-W PEP output level. I have, however, made no continuous, keydown, CW operating tests at full power purposely trying to destroy the traps!

Summary

Some hams may suggest using a different type of coaxial cable for the traps. The dc resistance of 40.7 Ω per 1000 feet of RG-59 coax seems rather high. However, I've found no coax other than RG-59 that has the necessary inductance-to-capacitance ratio to create the trap characteristic reactance required for the 80, 40, 20, 15 and 10-meter antenna. Conventional traps with wide-spaced, open-air inductors and appropriate fixed-value capacitors could be substituted for the coax traps, but the convenience, weatherproof configuration and ease of fabrication of coaxial-cable traps is hard to beat.

Notes

[1] L. Varney, "The G5RV Multiband Antenna... Up-to-Date," *The ARRL Antenna Compendium, Vol. 1*, p 86.

[2] M. Mims, "The Mims Signal Squirter," *QST*, Dec 1939, p 12.

[3] "Five Band Antenna," *The ARRL Antenna Book*, 16th Edition, pp 7-10 to 7-11.

By William C. Gann, W4NML **From *QST*, May 1966**

A Center-Fed "Zepp" for 80 and 40

Fast QSY for the Phone-C.W. Operator

The center-fed "Zepp" antenna is reviewed by W4NML, showing how complete coverage of a single band is made easy by using old concepts. Although the author shows how to use the Zepp on 80 and 40, only, the system can be used from 80 through 10 meters by employing an all-band transmatch.

Multiband antennas fed with resonant feeders were very popular in the pre-coax cable days. This article is presented to review a good, but seemingly forgotten system. This antenna should be of interest to traffic and contest operators, and to the casual operator who likes to use both the c.w. and phone portions of the 80- and 40-meter bands.

Our section s.s.b. net meets on 3965 kHz and the c.w. net meets on 3575 kHz. Many schemes were tried to make one antenna usable on both ends of the band so that a low s.w.r. could be maintained while securing efficient operation at the different frequencies. None of the antennas tried would permit an excursion of more than 300 kHz without a serious s.w.r. problem between the transmitter and the line. Getting from 80 to 40 meters with such an antenna was even more perplexing. The writer's dilemma was finally solved by the installation of the old reliable center-fed Zepp antenna.

Choosing the Dimensions

In order to use the antenna on 40, 75, and 80 meters, tuned feeders are required[1]. So that the feeders can be matched to the transmitter, a transmatch is used at the "shack" end of the line. Parallel tuning is used to minimize the complexity of the transmatch. This requires that the transmission line presents a high impedance to the transmatch on both bands.

The charts in the handbooks did not give a set of Zepp antenna dimensions that were suitable for the author's installation. Because of the existing tower, which would permit the antenna to be supported at the 50-foot level, and because the ham shack was adjacent to the tower, the prescribed feeder lengths were not practical. A graph was plotted to show the frequency extremes to which the antenna would be tuned, showing the minimum and maximum impedance points across the bands. It was determined that the combined length of one leg of the feed line and one section of the dipole would be 114 feet.[2] A length of 53 feet was used for the feed line and each leg of the driven element was cut to 61 feet.

To broaden the antenna's response, the driven element's effective area was made larger by paralleling two lengths of No. 12 copper wire as shown in Figures 1 and 2. With this arrangement, the Q of the antenna is lower, permitting the operator to QSY approximately 200 kHz without readjusting the transmatch.

Construction Notes

The driven element and the feed line are made from No. 12 copper wire. The 4-inch wide ceramic spreaders used to hold the feeder wires apart are made by the E. F. Johnson Co. Light-weight poly spacers are used to spread the driven-element wires and are sold as TV "clothespins" by the Telco Co. (Figure 2). All of the spreaders are attached to the No. 12 wire by short pieces of No. 18 copper wire. The distance between the spreaders is 4 feet for both the driven element and the feed line.

Sections of 1 × 4-inch lumber are used to hold the feeders away from the steel tower (Figure 3). Each piece is 24 inches long, notched at one end, and is fastened to the tower with U-bolts. Porcelain telephone-type insulators are attached to the feed-line end of each board, offering low-loss anchor

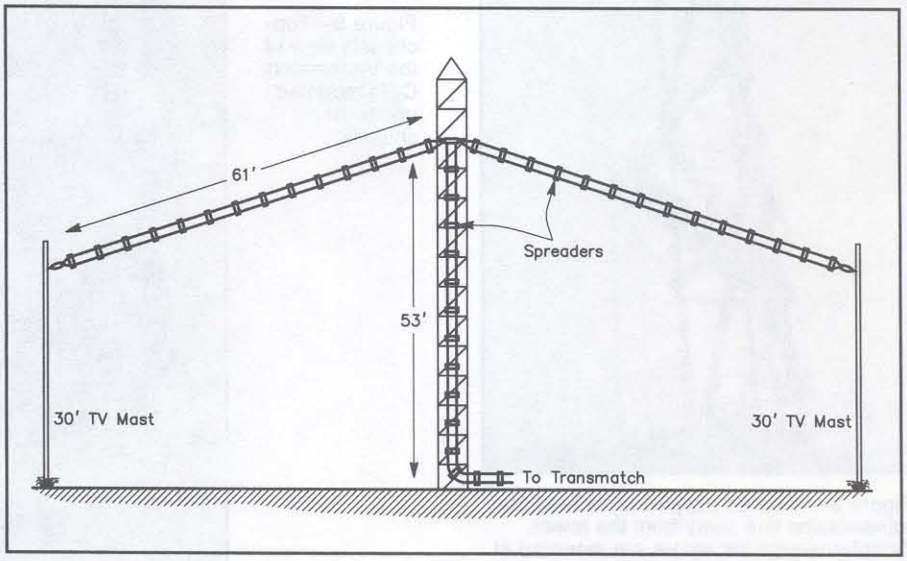

Figure 1—Layout of the 2-band Zepp antenna. Dimensions for each part of the antenna are shown. Feed point is anchored to one of the wooden support arms (See Figure 3).

---

[1] Center-fed Antennas, *A.R.R.L. Antenna Book*, Chapter 6.

[2] This length is generally 145 feet for operation in the 3.5- to 30-MHz range when the antenna is mounted horizontally (no droop), away from steel towers, and with a singlewire driven element.–*Editor*

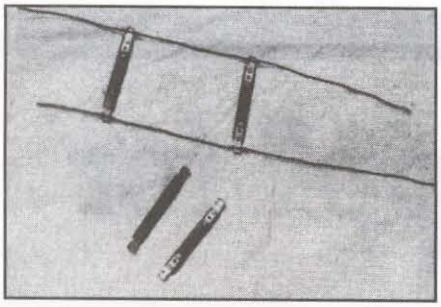

Figure 2—Details showing how the driven element spreaders are attached to the No. 12 wire. No. 18 copper wire is wrapped above and below each spreader to hold it in place.

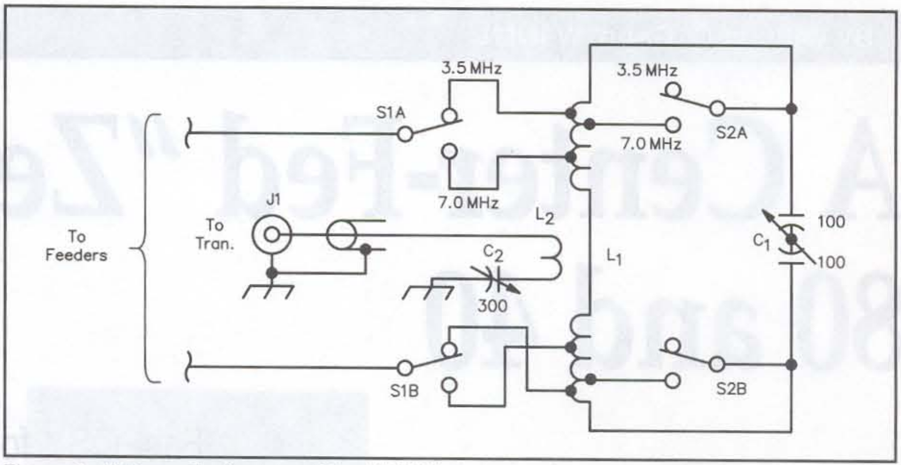

Figure 4—Schematic diagram of the W4NML transmatch.

C_1—100 pf. per section transmitting variable. (Split stator type with 0.175-inch spacing between plates.)

C_2—300-pf. variable capacitor(0.078 spacing or greater).

J_1—SO-239 coax connector.

L_1—56 turns, 3-inch diam., 8 turns per inch coil. S_1 taps are 9 turns from ends of coil for 80 meters, and are 22 turns from ends of coil for 40 meters. S_2 taps are 5 turns from ends of coil for 80 meters and 17 turns from ends of coil for 40 meters. (Air Dux 2408T or Polycoils 1779 usable.)

L_2—8 turns of Air Dux 2408T (center portion of L_1).

S_1, S_2—Ceramic rotary, 2 poles, 2 positions, 2 sections.

points for the transmission line. The uppermost support arm, at the 50-foot level, is used as a mount for the center of the driven element.

The far ends of the antenna are supported by 30-foot TV masts. A pulley and halyard arrangement is used for raising and lowering the ends of the antenna. Because the end supports are not as high as the feed point of the antenna, the dipole has a slight droop, but this does not seem to impair the performance.

The transmission line is brought into the operating position by means of feed-through insulators, mounted on a plywood strip which fits under a partially raised window. Insulated No. 12 house wire is used between the feed-through insulators and the transmatch.

Transmatch

Ideas for the author's tuner (Figure 4) were taken from the excellent transmatch article by McCoy.[3] Band changing is made possible by a large ceramic switch (origin unknown) which was obtained at a hamfest. An identical switch is used for selecting the taps for the feed line (Figure 5). Coil L_1 contains 56 turns of No. 14 wire, is 3 inches in diameter, and has 8 turns-per-inch (Air Dux 2408T). A stationary link, L_2, at the center of L_1, contains 8 turns of No. 14 wire and is a part of the Air Dux coil from which L_1 is made. The link is tuned with a 300-pf. variable capacitor. The author did not have a unit of the correct type, so two 150-pf. capacitors were parallel-connected (mounted under the chassis). Capacitor C_1 is a 100-pf-.per-section variable

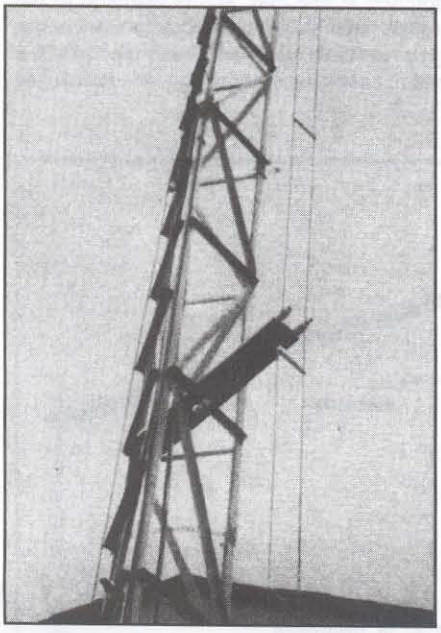

Figure 3—Wooden support arms hold the transmission line away from the tower. Telephone-type insulators are mounted at the end of each board to make the feed line secure.

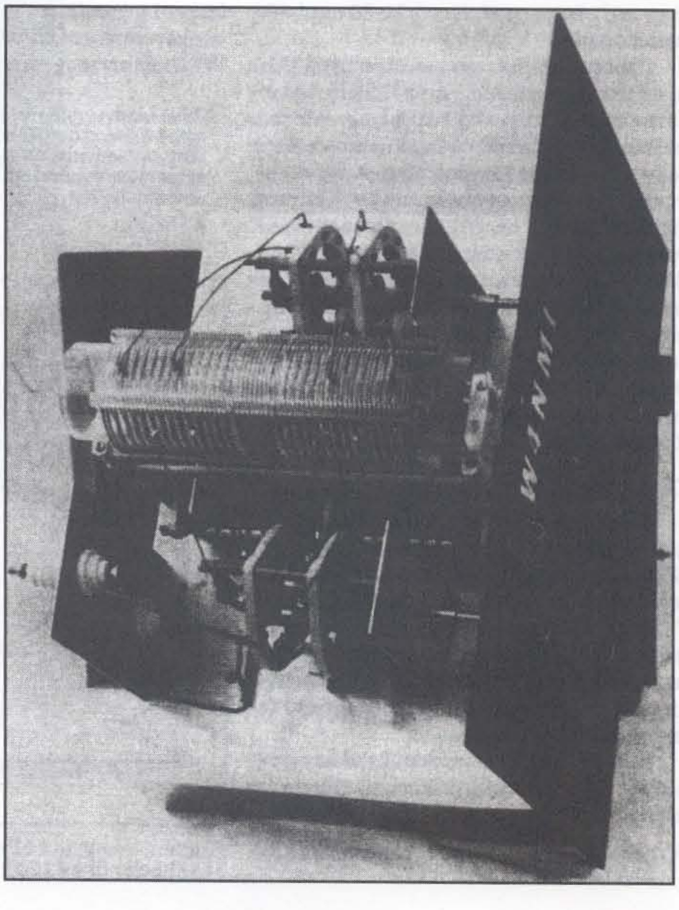

Figure 5—Top-chassis view of the transmatch. C_2 is mounted under the chassis.

---

[3] *QST*, July 1965.

with wide spacing. To give L_1 some rigidity, it is mounted on a plexiglass tube which is supported by the frame of C_1 with standoff insulators.

Results

While using clip leads, the correct tap points for the feeders were found by operating the transmitter through a Collins wattmeter and tuning C_1 and C_2 for zero reflected power. The transmatch permitted the transmitter to "see" 50 ohms in any part of either band. After establishing the correct tap points for the feed line, permanent connections were made between L_1 and the switches.

Next, the tuner was used with the 30L-1 amplifier at an output level of 700 watts. After a 30-minute QSO, no evidence of coil heating could be detected.

When compared to other antenna systems used by the author, the new skywire showed improved performance. It was believed that some sacrifice in efficiency would result from changing to the new antenna. Happily, it was found that we could have our cake and eat it too! Extended use indicated that the performance was, indeed, better than with previous antennas used.

I wish to thank three friends for their help in making this article possible: K4WWN for his tower climbing and photography work, K4ADK for building the transmatch cabinet, and Roy LeCrone for additional darkroom and photographic assistance.

Although this antenna system is an old standard, it may be the answer to your QSY problems. The cost is nominal and the results are most rewarding.

From *QST*, March 1967 (Technical Correspondence)

All-Band Antenna

With reference to the article on a Center-fed Zepp for 80 and 40 in May 1966 *QST*:

I set out to accomplish several things with an antenna to be installed on a California lot which runs east-west:

1) One pole.
2) No guys.
3) Good for short skip up and down the West Coast on 80, 75 and 40 meters.
4) Throw lobes across populated DX areas on 20 and 15 meters with a fairly low radiation angle.
5) Allow some onmidirectional DX on 40 meters.
6) Keep away from anything with critical antenna length or critical tuning.
7) Minimum cost.

Figure 1 shows the arrangement I ended up with. Results have been exceptional for a simple system of low height.

I use No. 14 wire for the antenna. The feeders are also No. 14. The center is mounted on an unguyed wooden pole about 34 feet high. Each section of the antenna is 65 feet long and the ends are only 14 feet high. Thirty-foot feeders are used with series tuning on 80 and 40 and parallel tuning on 20, 15 and 10 meters. Loading from 3.5 to 30 MHz is excellent and not at all critical in tuning. The fact that the feeders are less than 1/8 wave on 80 allows reactance to be tuned out in the feeder-tuning arrangement on that band. The antenna is a bit long for the high end of 75 meters, but tuning there is good (this length was picked because of the slightly longer physical length required on the upper bands for end effect).

I have the antenna itself running east-west, giving some directivity north-south for QSOs with short skip up and down the west coast on 80 and 40 meters (it was found in an earlier antenna . . . vertical . . . that a vertical was not satisfactory for high-radiation-angle short-skip operation). On 20 and 15 the lobes tend to cut across major population DX areas. The tilt of the wire, which lowers the vertical radiation angle plus apparently some lobe addition, seems to give better results in the desired DX directions on 20 and 15 meters than 33-feet-high half-wave horizontal antennas oriented in the correct directions. Quite a bit of omnidirectional DX has been worked on 40 meters, undoubtedly because of the antenna tilt. DX operation on 15-meter c.w. has been really exceptional. Quite often I hook a DX station through the pileup when local beam stations miss (power output is about 150 watts). Since 10 has opened up I have used the antenna quite a bit on that band with very good results for both North American and DX contacts.

As a result of playing around on 160 meters with the bottom of the feeders connected together and working the antenna as a " T " against ground, I decided to see what happened when it was operated as a top-loaded vertical on 80 and 40 meters. The feeders were tied together and the antenna worked against a ground consisting of two 8-foot rods in water-soaked earth. On 80 this places the maximum-current point directly at the top of the vertical section (1/8 wave long) and on 40 gives the effect of a 1/4-wave vertical with maximum current at the bottom. This arrangement gave much better results than the Zepp where low-angle radiation was required, and less effective results than the Zepp where medium- and high-angle radiation was required. An exception is directly off the ends on 80 meters, where the vertical and Zepp seem to give the same results. The Zepp arrangement is therefore now used for short and medium skip on 80 and 40 and the vertical arrangement for long skip or DX. In receiving, the signal-to-noise ratio decreases greatly with the vertical arrangement (vs horizontal), thus somewhat offsetting the overall advantage of the vertical for DX operation. If this condition is extreme, I use the Zepp for receiving and the vertical for transmitting, for DX operation. It appears that the vertical transmitter and horizontal receiver is by far the best DX arrangement for metropolitan areas, but most likely the vertical for both would be best for rural areas where the QRN is lower. – *Dave Hardacker, W6PIZ*

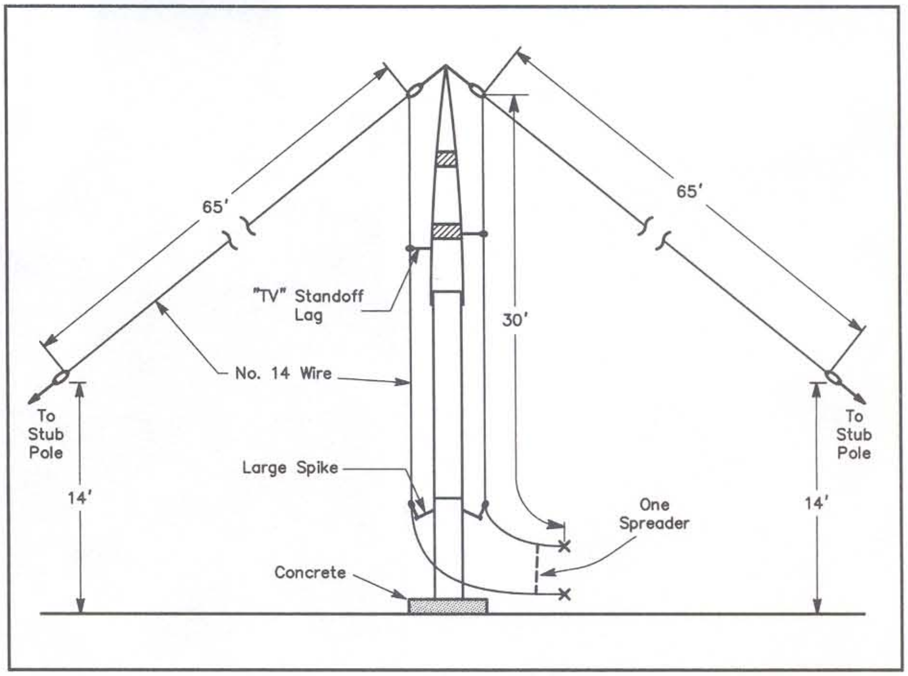

Figure 1–The W6PIZ all-band antenna. Power is applied through a series- or parallel-tuned link-coupled matching circuit at X-X for operation on 3.5 through 30 MHz (see text). For 2 MHz, and also for certain types of work on 3.5 and 7 MHz, points X-X can be connected together and the antenna worked against ground.

By William J. Lattin, W4JRW **From *QST*, December 1960**

Multiband Antennas Using Decoupling Stubs

Substituting Transmission Line Sections for Lumped-Constant Traps

Since W4JRW obtained a patent on this multi-frequency antenna system nearly ten years ago we can't call it "new," but at least it should be welcome news to those seeking a simple way to get good radiation on several bands, Shorted 1/4-wavelength stubs provide r.f. insulation and also serve as part of the antenna.

Since amateurs usually desire to operate on more than one band, several methods have been devised to use a single antenna on several bands. The earliest arrangements employed various combinations of feeder lengths, antenna lengths, and series or parallel tuning of the coupling circuit. Later on, the use of parallel-tuned "traps" with lumped constants which act as insulators at a particular frequency was invented.[1] A practical arrangement of this system for amateur use was developed[2] and is in rather wide use.

It is well known that the parallel-tuned circuit and quarter-wavelength shorted stub of Figure 1 are very similar electrically. Both configurations show a high impedance across points A and B. However, if a stub is connected to an antenna in this manner it does not act as an insulator but rather as a phase changer. The collinear antenna uses such stubs to operate a series of halfwave sections in phase.

There is a different connection possible for the stub, that is from A to C, which will result in insulator action or decoupling in an antenna.[3] For instance, shorted stubs a quarter wavelength long at 28 MHz can be attached to the ends of a 28 MHz dipole as in Figure 2. The 28 MHz dipole is effectively isolated or decoupled from the balance of the antenna which can be made long enough to resonate at 14, 7 or 3.5 MHz. If another pair of stubs is added for 14 MHz,

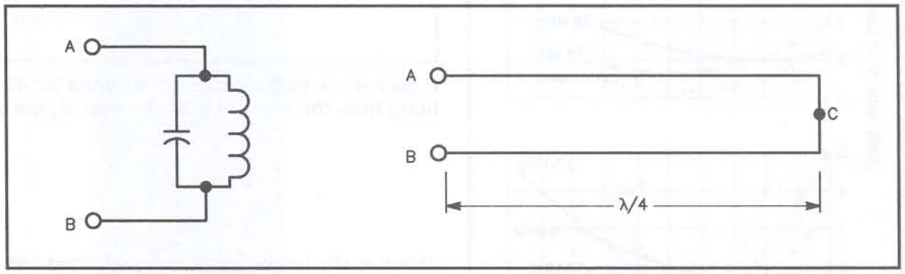

Figure 1—A parallel-tuned circuit has a high impedance at its resonant frequency, and so does a $^1/_4$-wavelength shorted transmission line.

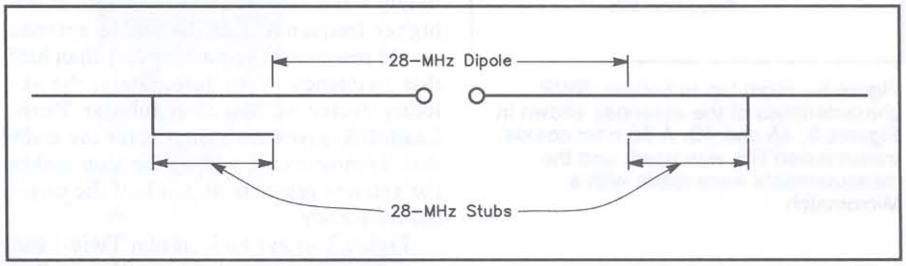

Figure 2—A two-band antenna for 28 MHz and some lower frequency. The center portion is an ordinary 10-meter dipole. The shorted stubs are $^1/_4$ wavelength long at 28 MHz and look like an open circuit at that frequency when connected to the dipole as shown. Extensions on the ends of the stubs can be used to resonate the antenna at any frequency less than half of 28 MHz.

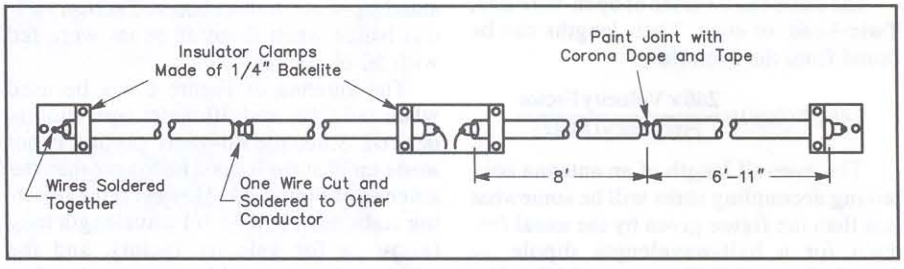

Figure 3—Construction and dimensions of an antenna for 10 and 20 meters using 300-ohm tubular Twin-Lead for both the dipole and stubs. Either a 50- or 75-ohm transmission line can be connected at the center of the dipole.

---

[1] Morgan, "A Multifrequency Tuned Antenna System," *Electronics*, August, 1940.

[2] Buchanan, "The Multimatch Antenna System," *QST*, March, 1955.

[3] Lattin, Patent No. 2,535,298.

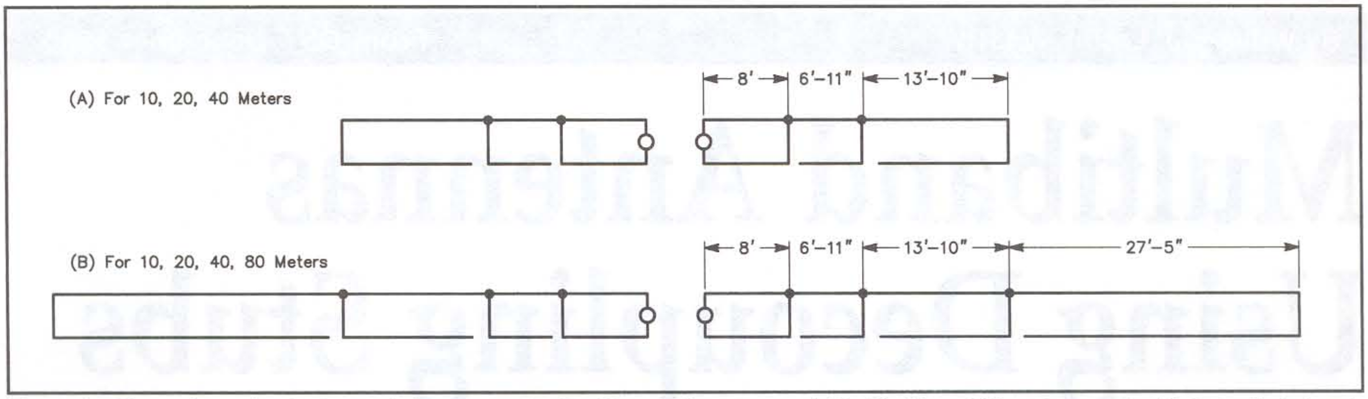

Figure 4—Dimensions of stub-decoupled antennas for 10, 20 and 40 meters and 10, 20, 40 and 80 meters made of tubular Twin-Lead. Either antenna can also be used on 15 meters where the 40-meter section is 3/4 wavelength long.

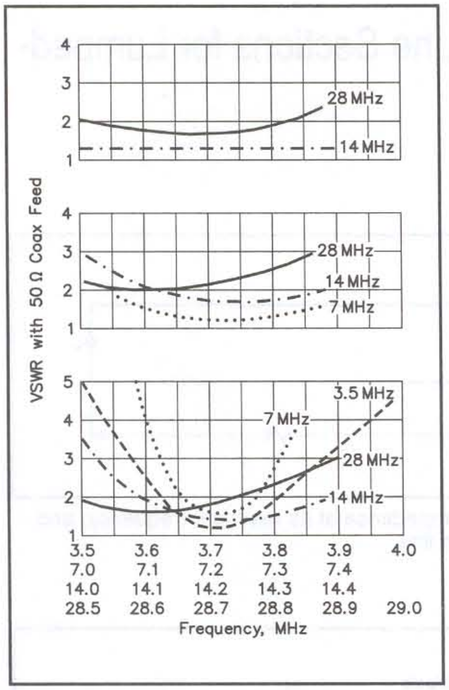

Figure 5—From top to bottom, SWR characteristics of the antennas shown in Figures 3, 4A and 4B. A 50-ohm coaxial transmission line was used, and the measurements were made with a Micromatch.

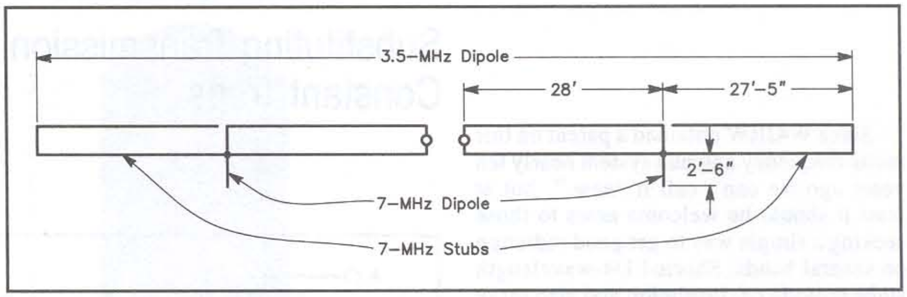

Figure 6—A stub-decoupled antenna for 40 and 80 meters. In this case wires must be hung from the ends of the 40-meter dipole to resonate the antenna in that band.

there will be isolation at both 28 and 14 MHz, and a 10-20-40-meter or 10-20-80-meter antenna can be made.

The stubs can be made of open-wire line, Twin-Lead, or coax. Their lengths can be found from the formula

$$\text{Length (feet)} = \frac{246 \times \text{Velocity Factor}}{\text{Frequency (MHz)}}$$

The over-all length of an antenna containing decoupling stubs will be somewhat less than the figure given by the usual formula for a half-wavelength dipole — *Length* (feet) = 468/*Frequency* (MHz). For instance, an antenna for 10 and 20 meters must be 29 feet, 10 inches long for resonance at the lower frequency, whereas the formula gives a length of 33 feet.

If open line with a velocity factor of nearly unity is used for the stubs, the overall length of a two-band antenna would be nearly a full free-space wavelength at the higher frequency and the whole antenna would resonate at something *less* than half that frequency. Very fortunately, the velocity factor of 300-ohm tubular Twin-Lead (0.8) gives such lengths for the stubs that, in most cases, adding the stub makes the antenna resonate at just half the original frequency.

Figure 3 shows how tubular Twin-Lead can be used for the antenna itself as well as the stubs and includes dimensions for 10- and 20-meter operation. The foam-filled type of Twin-Lead is recommended to keep out moisture. Lengths for three- and four-band antennas using the same construction are given in Figure 4. Figure 5 indicates the standing-wave ratios observed across various bands when these antennas were fed with 50-ohm coax.

The antenna of Figure 6 can be used when only 40- and 80-meter operation is desired. Since the 40-meter portion is not made up of stubs it must be longer than the antenna of Figure 4A. However, the isolating stubs must still be 1/4 wavelength long (allowing for velocity factor), and the whole antenna would resonate at a frequency below 3.5 MHz if the stubs were simply added to the ends of the 7-MHz dipole. To get around this, the dipole is shortened until the whole antenna tunes to 80 meters. Then resonance at 40 meters is restored by adding extra lengths of wire at the stub junctions. These wires are short and can just hang down from the antenna as shown.

Any of the antennas which will operate on 40 meters can be used on 15 meters as the 40-meter stubs will be approximately 3/4 wavelength long and will provide decoupling. The result is equivalent to operating a 7-MHz dipole at three times its resonant frequency, and we have found the s.w.r. is usually not lower than 3 to 6 when using 40-meter antennas of any type on 15 meters[4].

The power rating of the antenna will depend on the insulation at the stub junctions. These junctions can be painted with corona dope and covered with vinyl tape. It has been our experience over several years that the insulation will not break down with a kilowatt-input transmitter, 100 percent modulated, except when wet or very damp. In this case, the input should be reduced to

---

[4] Theoretically, a center-fed antenna working on its third harmonic shouldn't be more than about 50 percent higher in resistance than on the fundamental. One would expect an s.w.r. on the order of 2 to 1 rather than such high figures. – Ed.

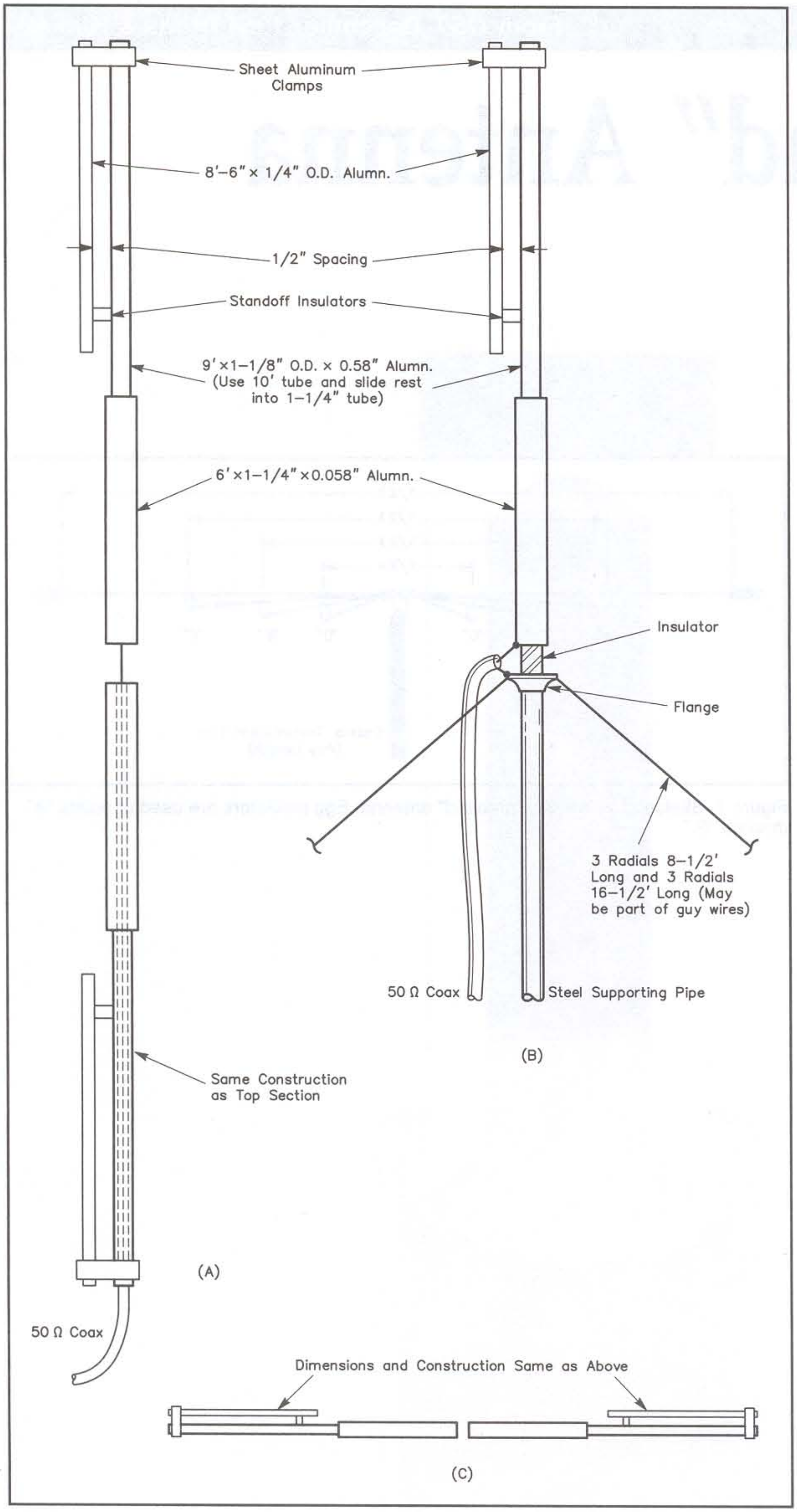
Sheet Aluminum Clamps
8'-6" × 1/4" O.D. Alumn.
1/2" Spacing
Standoff Insulators
9'×1-1/8" O.D. × 0.58" Alumn. (Use 10' tube and slide rest into 1-1/4" tube)
6'×1-1/4"×0.058" Alumn.
Insulator
Flange
3 Radials 8-1/2' Long and 3 Radials 16-1/2' Long (May be part of guy wires)
50 Ω Coax
Steel Supporting Pipe
(B)
Same Construction as Top Section
(A)
50 Ω Coax
Dimensions and Construction Same as Above
(C)

Figure 7—Dimensions and suggested construction for coaxial, ground-plane and tubing dipole antennas for 10 and 20 meters. The arrangement in A might be mounted with standoff insulators attached to the 1¼-inch sections near the center of the antenna. The dipole in C could be closed at the center and fed with a gamma or "T"-matching system. Similarly-constructed parasitic elements could be added to make a multiband beam.

perhaps 500 watts unless special precautions have been taken to seal up the junctions at the open ends of the stub. Of course, on the lowest band for which the antenna is designed the stubs do not have voltage across them and will not be subject to breakdown or flashover. The high voltage across the open end of a stub occurs only at the resonant frequency of that stub.

Figure 7 shows the construction of several 10 and 20-meter antennas which have been built and the dimensions required for resonance in these bands. The spacing between the rods forming the shorted stubs is not critical—the same lengths were obtained with 1-inch instead of 1/2-inch spacing. Insulators should be made of low-loss material. Reflectors and directors for a multiband beam could be made up the same way.

"All-Band" Antenna

Figure 1 is a sketch of an "all-band" antenna system that I have been using with success for some time. The idea is not a new one, having appeared in *QST* at least 10 years ago. However, I feel that there are many newcomers since that time who would be interested in a simple system that can be fed with a single 70-ohm transmission line.

The arrangement consists of dipoles, cut for each band and connected in parallel at the center. Although I have not checked standing-wave ratios, the results seem to indicate that it gets out as well as a bunch of individually-fed doublets. If you haven't tried it, you're in for some surprises.– *R. L. Cope, W8MOK*

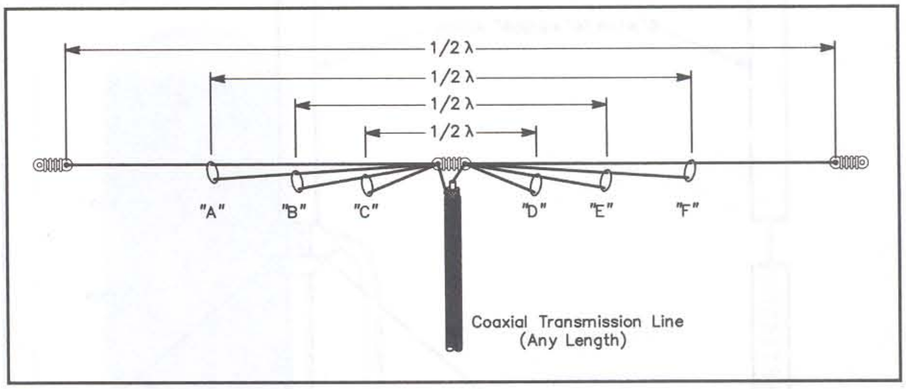

Figure 1–Sketch of W8MOK's "all-band" antenna. Egg insulators are used at points "A" through "F."

From *QST*, January 1975 (Hints & Kinks)

Three-Band Matching System for a Forty-Meter Doublet

A common method for energizing a half-wave antenna is to feed it at the center with parallel-conductor TV lead-in, or Twin-Lead as it is usually called, and to use an open stub for matching the 50- to 70-Ω antenna resistance to the 300-Ω impedance of the line. However, this technique, as described in the latest edition of *The ARRL Antenna Book*, generally gives proper matching on only one band.

After a number of trial-and-error calculations on a Smith Chart, along with lots of cut-and-try experimenting, I devised a three-stub matching scheme so that I could operate my 40-meter doublet on 40, 20 and 15 meters. Figure 1 illustrates this method and gives the lengths of the stubs and their positions along the feed line. The dimensions shown are for standard Twin-Lead, with a velocity factor of 0.82. All of the stubs are open at the ends and are made from the same type of line as the feed line. Note that the two lower stubs are connected at the same point on the feed fine.

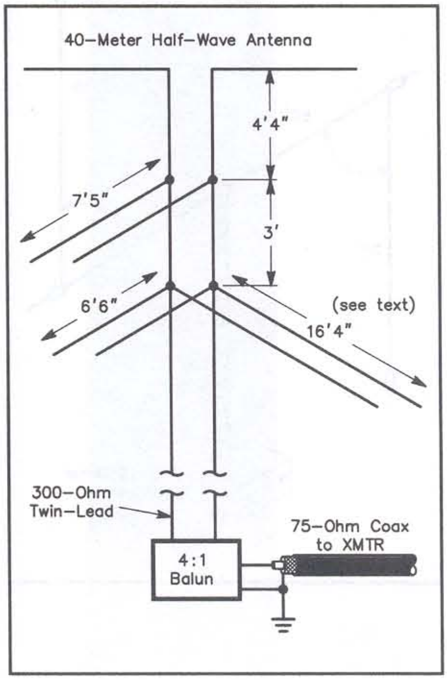

The length of the longest stub is fairly critical. It should first be cut to 17 feet and then trimmed no more than two inches at a time, until the SWR is minimum in the center or the desired portion of the 20-meter band. The feed fine can then be matched with a 4:1 broad-band balun to a 75-Ω coaxial cable from the transmitter. With my antenna, the described matching system gives an SWR of less than 2.5 to 1 over all of the three bands, with minimum values of 1.3 to 1 on 40 and 15 meters, and 1.7 to 1 on 20 meters. As with any multiband antenna, one must guard against harmonic radiation. – *Frank Stuart, K7UUC*

From *QST*, December 1979 (Technical Correspondence)

A "Z" Antenna For the 10-160 Meter Bands

One of my interests in ham radio is designing and constructing antennas for both general amateur use and for the Army MARS system. For the amateur who has limited space, I have designed a "Z" antenna that covers the bands from 10 through 160 meters. It is easily constructed from wire. Spreaders for the transmission line are fabricated from Lucite strips or Plexiglas rods. (Refer to the yellow pages of telephone directories for the names of dealers who handle Plexiglas or the equivalent.) For two no. 14 wires, a 2-inch (51-mm) spacing is adequate.

Although a height of 100 feet (30 meters) is indeed desirable for this antenna, hams who settle for elevations between 30 and 50 feet (9 and 15 meters) will still obtain good results. The angles α between the wire segments will depend on individual situations such as the placement of trees or other supports. Generally, the wider the angle, the better the performance.

W1NH, the New Hampshire SCM, who is really into antennas, says my design is "FB." My evaluation of the antenna is that the "aerial" is great.–*John N. MacInnes III, WB1FPD*

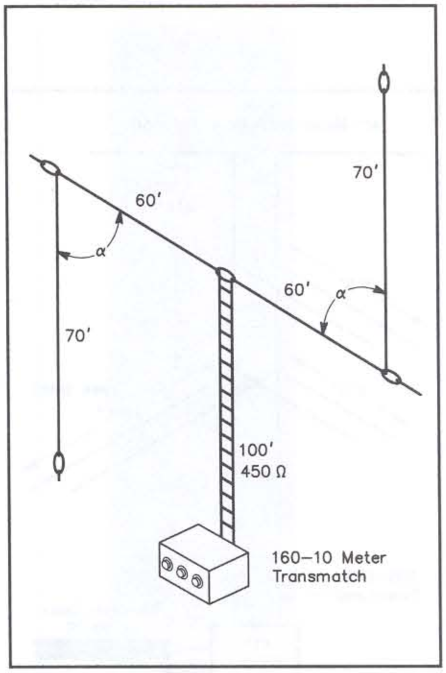

A 10- to 160-meter horizontal Z antenna. Inexpensive no. 14 copper-clad (electric fence) wire may be used. For complete information on open-wire transmission lines, see *The ARRL Antenna Book*, any recent edition.

By Taft Nicholson, W5ANB/AAR6AG **From *QST*, November 1981**

Compact Multiband Antenna Without Traps

Looking for an inexpensive, easy-to-build, portable antenna to use with your tube-type transmitter? This first cousin of the G5RV may be the aerial of your dreams.

Do you need an antenna for portable operation? I do! My preference is for making things as simple as possible. I don't like traps and matching units. The antenna shown in Figure 1 has no traps, needs no matching unit for the 10-, 20- and 40-meter bands (when used with vacuum-tube PAs), is lightweight and installs easily. What more could you ask for?

Construction

Construction is simple. Cut two lengths of stranded copper wire (such as Radio Shack 278-1292) to 44 feet, 2 inches each (m = ft × 0.3048). Attach 36 feet, 8 inches of 300-ohm twin lead as shown in Figure 1. Coaxial cable attaches to the other end of the twin lead at the points marked A and B in the diagram. I wound 7 feet, 2 inches of RG-58/U coaxial cable into an rf choke to minimize problems with rf flowing on the outside of the coaxial cable. This length of cable in the choke evolved from an attempt to match the antenna to the transmitter for operation on 15 meters.

Alternatively, you could use open-wire feeders in place of the twin lead. If you choose that method, you will need to make the section 42 feet, 6 inches long. Or you could attach the twin lead or openwire conductors to a matching unit. When using a matching unit, there is no particular merit in the lengths given in the diagram.

SWR

The SWR of the antenna is less than 3:1 on 10, 20 and 40 meters. I have no difficulty loading transmitters with tube-type PAs (e.g., Galaxy V, Swan 350 or Drake T4X). This antenna will work with some transmitters on 15 meters, but tuning is quite critical.

However, 80-meter operation is a problem. I have not been able to obtain full output power with these transmitters on frequencies below 3.750 MHz. The SWR measures between 5:1 and 8:1 for the lower part of the band. I constructed a loading coil to go between the choke and the twin lead (Figure 2). The coil consists of 44 inches of twin lead wound on a 7/8-inch diameter form (my left thumb). Once the coil was wound, I removed it from my thumb and used electrical tape to secure it. For 80-meter operation, attach ends C and D to the choke and ends E and F to the twin lead. Remove the coil for operation on the other bands. Banana plugs and sockets can be used to facilitate the insertion and removal of the coil.

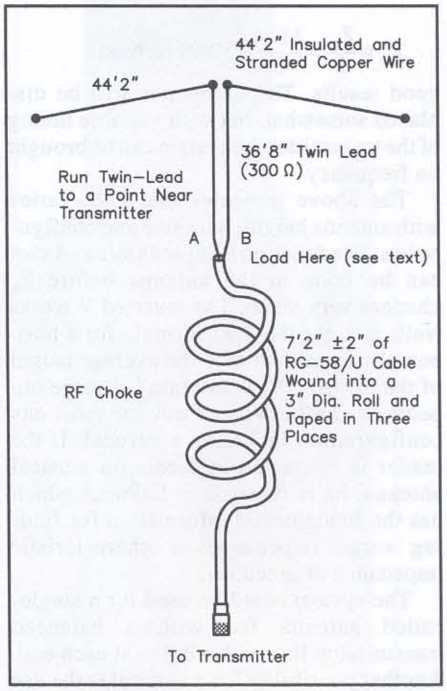
44'2" Insulated and Stranded Copper Wire
44'2"
36'8" Twin Lead (300 Ω)
Run Twin—Lead to a Point Near Transmitter
A
B
Load Here (see text)
7'2" ±2" of RG–58/U Cable Wound into a 3" Dia. Roll and Taped in Three Places
RF Choke
To Transmitter

Figure 1—Diagram of the compact multiband antenna. For 80-meter operation the loading coil is inserted at points A and B. Banana plugs and jacks may be added here to facilitate insertion and removal of the loading coil.

Installation

The antenna should be as high as practical. I've had satisfactory results with the center of the antenna only 25 feet above ground, with the ends tied to fences or other convenient supports. Telescoping TV mast sections make a good support if nothing else is available. The legs of the antenna serve as two of the guy wires. One or two additional guy supports should be added (nonconducting material such as nylon rope is best).

This compact multiband antenna works satisfactorily on all bands from 20 through 80, and 10, meters. It has no traps and requires no matching unit when used with tube-type equipment. I have used it for portable operation in and out of the country. It is easy to pack, carry and erect. Perhaps you might want to try one. I think you'll like it! A brief discussion of the theory of operation follows in the appendix.

[Editor's Note: A description of the G5RV appears in the RSGB *Radio Communications Handbook*. Gray described it in the June issue of *Ham Radio Horizons*. A similar design was depicted in the Collins Radio manuals of the 1930s.]

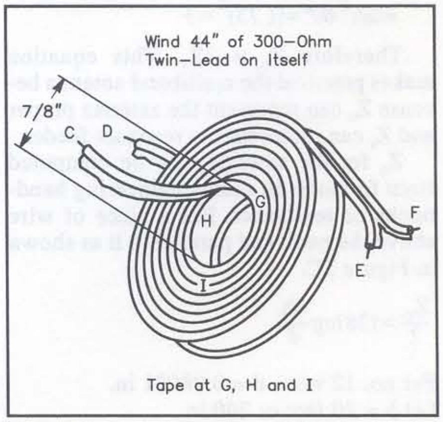

Figure 2—Loading coil for 3500-3750 kHz operation. Points E and F on the coil are connected to the coaxial cable. Points C and D are connected to the twin lead.

Appendix

This multiband antenna evolves from two connected transmission lines with critical length and ratios of surge impedances. The system is self-resonant at a fundamental frequency and at most of the even harmonics and several of the odd harmonics. The first five are 2nd, 4th, 5th, 7th and 8th.

Consider the transmission lines in Figure 3.

The two lines are of equal length, "ℓ" and different surge impedances, Z_o and Z_s.

Looking into the lines:

$$X_o = \frac{Z_o}{\tan(2\pi\ell/\lambda)} \text{ and } X_s = Z_s \tan(2\pi\ell/\lambda)$$

where

λ = wavelength

X_o = reactance looking into open line

X_s = reactance looking into shorted line

From the theory of resonant circuits, we know if we connect the lines the system will be resonant at all frequencies where $\overline{X}_O = \overline{X}_S$ provided the two reactances are of equal value and opposite signs. The open and shorted line provide this condition except at some harmonics.

Joining the lines as depicted in Figure 4 we find that

$$\frac{Z_o}{\tan(2\pi\ell/\lambda)} = Z_s \tan(2\pi\ell/\lambda)$$

$$\frac{Z_o}{Z_s} = \tan^2(2\pi\ell/\lambda) = \tan^2(360°\,\ell/\lambda)$$

If the angle $2\pi\ell/\lambda$ is made 60°, then the amplitude of the tangent at 120° or second harmonic will be the same. This will be true for 240° (4th harmonic) and 300° (5th harmonic). Similarly, these harmonic responses will continue at discrete angles above 360°, e.g., 7th, 8th, 10th, 11th, 13th, 14th and so on. The signs of the tangents wash out when squared.

The angle $2\pi\ell/\lambda$ becomes 60° by making $\ell = \lambda/6$ (1/6 of a wavelength) at the fundamental frequency.

$$\frac{Z_o}{Z_s} = \tan^2(360° \times 1/6)$$

$$= \tan^2 60° = (1.73)^2 = 3$$

Therefore $Z_o = 3Z_s$. This equation makes practical the multiband antenna because Z_o can represent the antenna proper and Z_s can represent the resonant feeder.

Z_o for the antenna may be computed from formulas in radio engineering handbooks or textbooks. For a piece of wire above the earth and parallel to it as shown in Figure 5C.

$$\frac{Z_o}{2} = 138 \log \frac{4h}{d}$$

For no. 12 wire, d = 0.08081 in.

Let h = 20 feet or 240 in.

$$\frac{Z_o}{2} = 138 \times \log\left(\frac{960}{0.08081}\right)$$

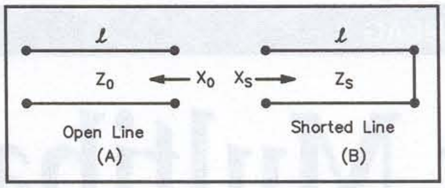

Figure 3—Two identical lengths of transmission line. Looking into the lines, the opposite ends are open and shorted at A and B, respectively.

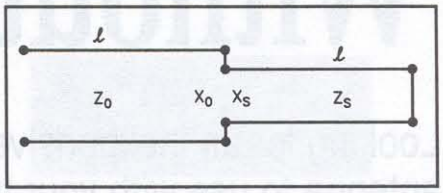

Figure 4—The two transmission lines from Figure 3 are joined to form one line.

138 log (11879) = 562 ohms

This value is not critical. One can use 300-ohm twin lead or 400-ohm open line with

If the wire size had been no. 18,

$$\frac{Z_o}{2} = 605 \text{ ohms}$$

If we use no. 12 wire, then we would have 562 ohms on one side and + 562 ohms on the other side.

$$Z_o = \frac{Z}{2} + \frac{Z}{2} = 1124 \text{ ohms}$$

From the formula $\frac{Z_o}{Z_s} = 3$

$$Z_s = \frac{Z_o}{3} = \frac{1124}{3} = 374.6 \text{ inches}$$

good results. The harmonics will be displaced somewhat, but with variable tuning of the transmitter the system can be brought on frequency.

The above indicates that $Z_o/2$ varies with antenna height, wire size and configuration. The function is logarithmic and a lot can be done to the antenna before Z_o changes very much. The inverted V works well; just use the $Z_o/2$ formula for a horizontal wire and let h be the average height of the inverted V. A formula for surge impedance can be worked out for most any configuration, including a vertical. If the reader is interested in feeding a vertical antenna, he is referred to LaPort,[1] which has the fundamental information for finding surge impedance or characteristic impedance of antennas.

The system could be used for a single-ended antenna fed with a balanced transmission line with a balun at each end. Another possibility for a vertical is the use of a two-wire, grounded, open transmission line, as discussed in LaPort's book.[2] The ground system would be critical.

When experimenting with these multiband lines, it is convenient to have some "stock" numbers to apply to the lines (see Figure 6). One-sixth of a wavelength is one-third of a half wavelength. A convenient length for a half wave on 80 meters is 135 feet. One-third of that is 45 feet or λ/6 for 80 meters. One-sixth of a wavelength on 40 meters is 22-1/2 feet. When you are designing an antenna, these lengths need to be multiplied by the propagation constant of the line. After construction and testing, the dimensions can be pruned for end effect, etc.

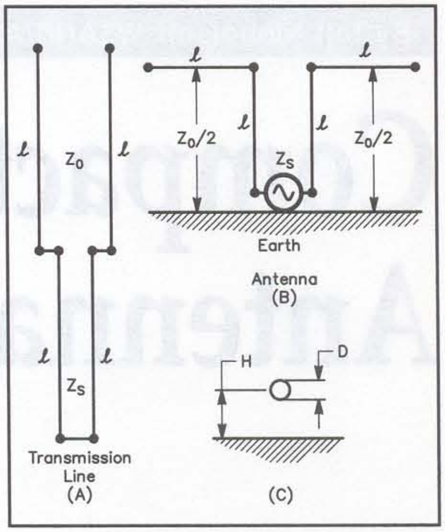

Figure 5—At A and B, the open portion of the transmission line evolves into the flat-top portion of the antenna. At C, diagram illustrating the formula for calculating Z_o for the antenna.

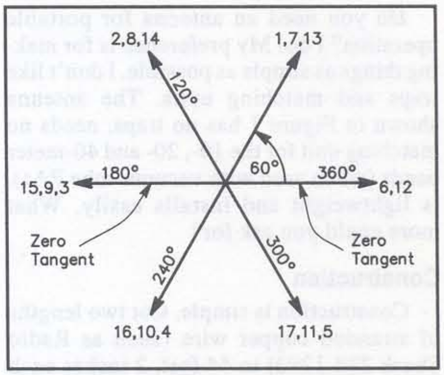

Figure 6—Angular position chart useful for determining "stock numbers" to apply to the chart.

When operated as a transmission line, the system as described may have application in end-feeding half-wave antennas, especially two half waves in phase. The system transforms a high impedance to a low impedance as a quarter-wave line will; however, it will do this at several even harmonics, in contrast to the quarter-wave line that is only responsive to odd quarter wavelengths.

The author wishes to thank Walt Maxwell, W2DU, for his detailed analysis of the theory section of this article.

Notes

[1] E. LaPort, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co., 1952).

[2] See note 1.

By Bill Wright, G0FAH **From *QST*, June 1995**

Five Bands, No Tuner

Enjoy some of the advantages of a multiband, ladder-line-fed antenna without an antenna tuner.

Reading "The Doctor is IN" (*QST*, January 1995) reminded me that the search continues for a simple backyard antenna. A wire *dipole* antenna fed at the center with 450-Ω ladder line is a good choice. The ladder line keeps your losses low—even at moderately high SWRs. All you need is an antenna tuner and you're in business. No coils or traps necessary.

But can you do away with the tuner and still keep the ladder line? That would certainly make life simpler. To achieve this, your transceiver needs to "see" an impedance that looks reasonably close to 50 Ω on as many bands as possible. Without an antenna tuner acting as the middleman between the 450-Ω ladder line and your 50-Ω radio, this could be a problem.

Table 1
Calculated SWRs for a 94-foot Dipole Fed with 41 Feet of 450-Ω Ladder Line

Note: Although the antenna is cut for the CW portions of the band, expect similar results at other frequencies.

Frequency	*SWR*
3.56	7.6:1
7.1	2.4:1
14.2	1.5:1
18.1	2:1
24.9	1.5:1
29	2.4:1

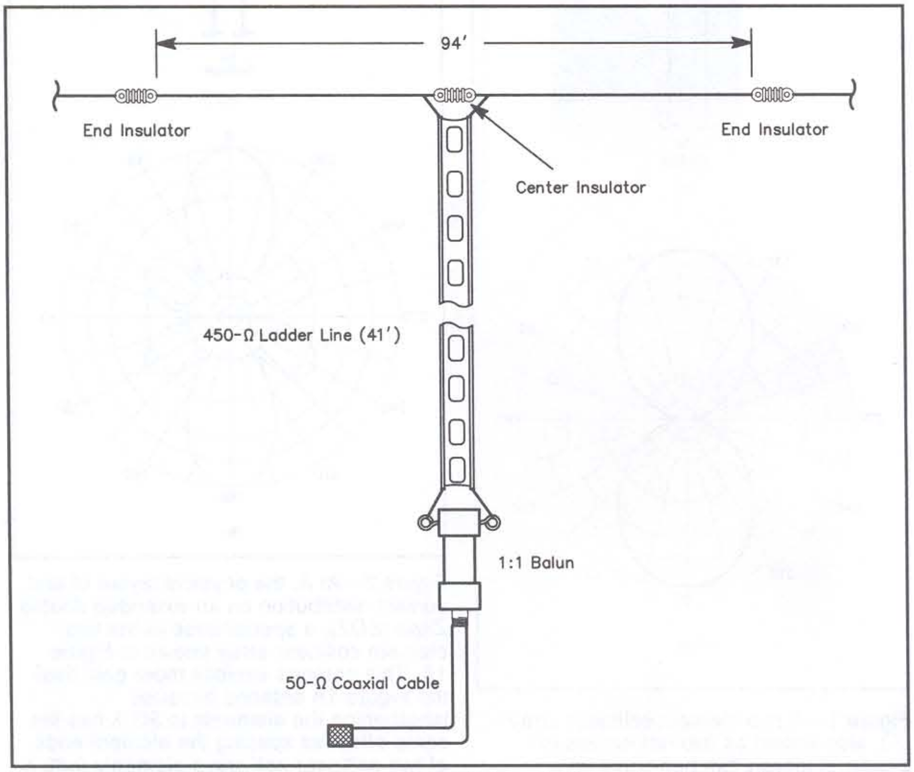

Figure 1—The wire dipole antenna is 94 feet in length. If you don't have 94 feet of open space, don't hesitate to droop the ends of the dipole to make it fit. Feed the antenna with 41 feet of 450-Ω ladder line that is connected to a 1:1 balun. From the balun to your radio, use 50-Ω coax.

Can It Be Done?

A few years ago I attempted to design an antenna that would work on several HF bands from 80 to 10 meters. Full details were published in the spring 1992 edition of *SPRAT*, the journal of the G-QRP club.

My inspiration was the venerable G5RV. I took a 94-foot-long dipole and fed it with ladder line (see Figure 1). By cutting the ladder line to a specific length and using a 1:1 balun to make the transition to coaxial cable, I found that I could get close to 50-Ω (and thus achieve reasonably low SWRs) on at least five bands: 40, 20, 17, 12 and 10 meters (see Table 1).

The on-air results were better than I expected. My radio was happy and I didn't need to meddle constantly with an antenna tuner.

Of course, you'll need an antenna tuner to work the bands where the SWR exceeds 3:1.

A simple tuner will do the job, though. Because you're using *unbalanced* coax ahead of the balun, you won't need one of the more expensive tuners designed for *balanced* feed lines.

For best results, put your antenna as high as possible. If the ends must bend downward to accommodate the size of your lot, don't worry. Run the ladder line to your balun and take your coax from there to your radio. Keep the coax portion as short as possible.

Conclusion

By eliminating the antenna tuner completely, you lose the flexibility of loading your ladder-line-fed antenna on virtually any band. In return, however, you gain the convenience of operating on several bands without making tuner adjustments each time you change frequency. Your losses are held to a minimum, which means that most of the power your radio generates is radiated by your antenna. Not a bad compromise! Of course, the results I achieved will vary when used at other locations. Still, it's a simple, fun project in the experimental spirit of Amateur Radio.

By Rick Olsen, N6NR

From *QST*, March 1993

The NRY: A Simple, Effective Wire Antenna for 80 through 10 Meters

Known as a *broadside collinear curtain array*, this antenna is simple to build, rakes in DX signals and has gain over a dipole on all the bands it covers!

Hams' thoughts ultimately turn to antennas. So it was with me a while ago. I had just moved to the emerald beauty of the Pacific Northwest and began pondering the 160-foot-tall Douglas firs in my backyard. I had considered putting up a crank-up tower, but decided that neighbors, architectural committees and crank-ups don't mix. Then I thought about topping a tree. Nope–I moved here because I love these trees.

I started sniffing around for alternatives. I obtained a copy of fellow ARRL Technical Advisor Roy (W7EL) Lewallen's antenna modeling program, *ELNEC*, along with a copy of *The ARRL Antenna Book*, and began to incubate the seeds of what would become my most enjoyable antenna project since I built a four-element quad back in 1974.

Hanging ropes from big trees is not a new idea.[1,2] 1 began by pacing off the distances between the monsters in my yard so I would know how much space I had to work with and how I could optimally direct my signal to places like Europe, Africa and the Caribbean.

Next came the selection of a configuration. Because the trees are approximately 120 feet apart, my first choice was a multielement collinear array. But a good friend, Terry Conboy, N6RY (in *NRY*, I'm the *NR* and he's the *RY*), suggested stacking a pair of *extended double Zepps*. I filed his suggestion away in my memory for the time being. I started my analysis.

The ARRL Antenna Book, on page 8-32, describes the basic function of the collinear array. In sum, it concentrates energy perpendicular to the wire, hence the term *broadside array*. This antenna's physical layout and pattern are shown in Figure 1. Such an antenna's gain can be increased by varying the end-to-end spacing of the elements.[3]

Chapter 8 goes on to explain three- and four-element arrays. Disappointment set in when I sized a four-element in-line array for 20 meters: It wouldn't fit between the trees! What *would* fit, however, was a "two-over-two." In fact, depending on the vertical separation, this array might even have some gain over the four-element collinear arrangement.[4]

My choice was made. I wanted to maximize performance on 20 meters and take what I could get on the other bands. I chose two 1-λ, center-fed wires vertically spaced 5/8 λ apart. According to *ELNEC*, with the array 100 feet high at the top and modeled

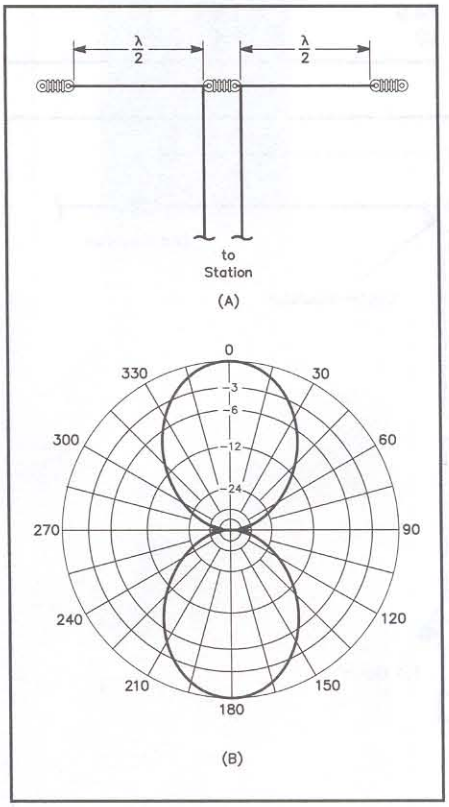

Figure 1—A two-element collinear array (A), also known as *two half-waves in phase*, is merely two half-wave wire elements placed end to end. This antenna has about 1.5 dB gain over a dipole in the same surroundings, and B shows its free-space pattern.

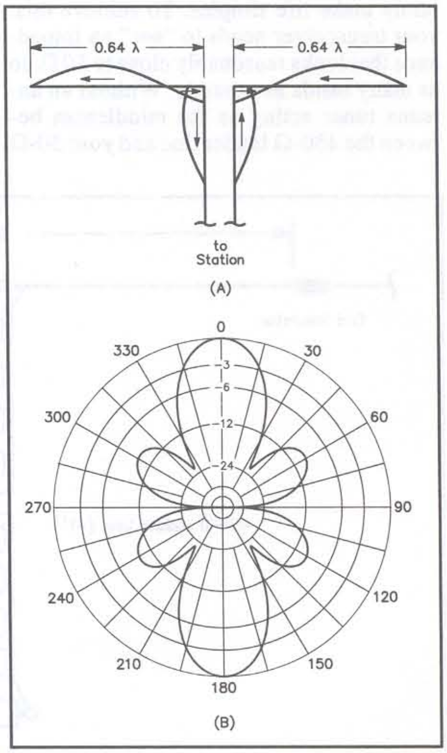

Figure 2—At A, the physical layout of and current distribution on an *extended double Zepp* (EDZ), a special case of the two-element collinear array shown in Figure 1A. This antenna exhibits more gain than the Figure 1A antenna because lengthening the elements to 5/8 λ has the same effect as spacing the element ends of two collinear half-wave elements 0.28 λ apart. (See Note 3.) At B, the antenna's azimuth-plane pattern in free space. This antenna has about 3 dB gain over a dipole in the same surroundings.

over lossy ground, this antenna's gain is approximately 12 dBi.[5] Not bad! The only problem is that this antenna's 20-meter feedpoint impedance is very high.

Reenter N6RY and the extended double Zepp. Take a look at Figure 2A. The extended double Zepp is simply a 1.25-λ, center-fed wire (two end-to-end 5/8-λ elements). There's a hidden message here! Have you found it?

Well, at first, I didn't either. Terry explained to me that the extended double Zepp (EDZ) is essentially two half-wave collinear elements spaced just over 1/4 λ apart. The extended double Zepp's added gain over two half waves in phase comes *not* from the extra 0.28 λ of wire between them, but from the *increased end-to-end spacing of the outermost pair of halfwave elements*. (Radiation from the inner 0.28-λ actually works *against* that from the outer halfwaves, but the gain increase from the new element spacing more than overcomes this disadvantage.) Figures 1B and 2B are similar. The EDZ has broader lobes and useful smaller lobes, both of which are attractive for a fixed array.

Sure enough, it was right in front of me all the time in the *Antenna Book* (page 8-34, fifth paragraph). In fact, the *whole darned array* was right in front of me in Chapter 8! I just needed some help piecing it all together. I modeled this array every way I could imagine, looking for the best possible performance. My wife didn't see me in the evenings for two weeks! Figure 3 shows the final configuration.

Figures 4, 5 and 6 show some of the results of my modeling. I used a design frequency of 14.2 MHz, a height of 100 feet above ground (for the top wire), elements made of #14 copper wire, and ground coefficients typical of the Pacific Northwest. All the elements in this array are fed in phase, as described in a bit.

As the figures show, the antenna's gain and patterns are impressive. On the 80- and 40-meter bands, the antenna has remarkably dipole-like patterns and gains and seems to work at least as well as a dipole in the same height range. On the higher bands, it works *much* better than a dipole. Its predicted gains in my installation are 9, 11 and 14 dBi at 7, 10.1 and 14 MHz, respectively. At 18.1 MHz (Figure 6) and above, the array begins to take on long-wire properties. It has several major lobes in directions closer to the wire axis, which the lower-frequency patterns don't provide.

To help you evaluate the NRY's patterns, Figures 4-6 superimpose the patterns of halfwave dipoles cut for 10.1, 14 and 18.1 MHz, respectively, on the NRY's patterns. These comparisons assume that the dipole is the same height as the NRY's feed point (halfway between the elements, or about 22 feet lower than the top wire), and over the same ground.

You may wonder about this antenna's effectiveness as a function of height. According to the *Antenna Book*, arrays like this have the best gain when the lower element is at least 1/2 λ above the ground (33 feet at 14 MHz). And, of course, as with all horizontally polarized antennas, this array's radiation angle decreases as you raise the antenna, which makes for better DX performance. The bottom line is that for DX work, you don't have to get the array up 100 feet for it to work well—you just need to put it up as high as you can. For closer-in coverage, it will work well at lower heights.

Another advantage of this array is that its feed-point impedances are manageable. To maintain my goal of broadband use, I decided to feed the antenna as shown in Figure 3. This is the "lazy-H" configuration described starting on page 8-37 of *The ARRL Antenna Book*. The attractiveness of this configuration is that it doesn't rely on specific phasing-line lengths to work properly. To the extent made possible by the different mutual impedances between each element and the ground, the phase relationship between elements remains constant regardless of frequency.

Final Dimensions

The antenna's physical dimensions (86 feet, 10 inches long and 43 feet, 5 inches high, not counting insulators) are easy to remember because the height is half the length. Because the antenna is basically 1.25 (5/4) λ long, center-fed and vertically spaced 5/8 λ apart, you can easily scale it to whatever band you like. For instance, to scale it to 10 meters, divide all the dimensions by two (28 MHz/14 MHz = 2).

Construction

Now for the really fun stuff: putting this thing up in the air! Building the antenna is a breeze. It took me all of about an hour to cut and solder the wires and attach the tethers. The harder part was preparing the trees.

If you have a friend like Bernie Olshausen, N6RUX, you have it made. He's a crack shot with a slingshot, similar to that described by Wade Calvert, WA9EZY, in *QST* a couple of years back.[6]

Once Bernie landed the fishing lines over the right tree limbs, the next job was to pull up some lightweight twine (we used seiner twine—the stuff fish nets are made of) with the fishing line. This is important; rope sometimes snags as it goes over limbs, and you don't want to try to pull it over with 6-pound-test fishing line! With the twine, we pulled up some 800-pound-test nylon rope with a pulley on one end, using another piece of nylon rope through the pulley to act as a catenary for the array.

It's important to use a strong, weather-resistant catenary rope to hang the array, because the antenna loses its desirable properties quickly when it takes on the shape of a V. You can expect a 2-dB gain

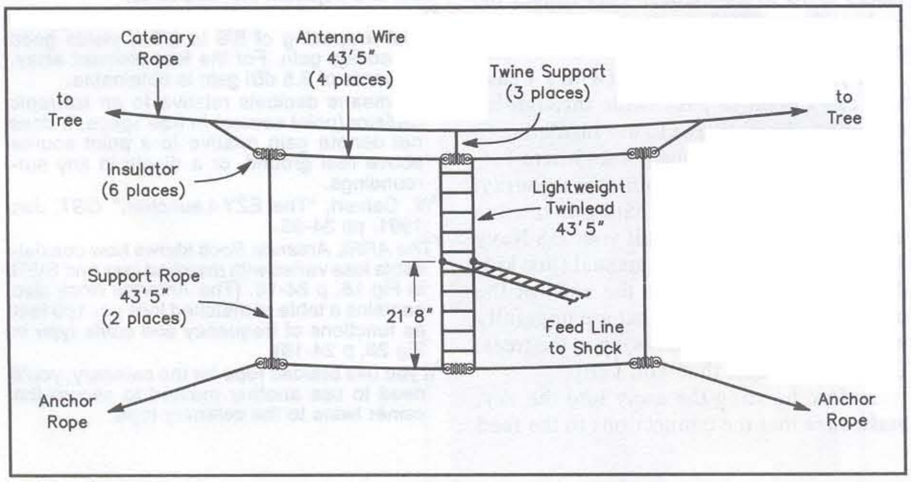

Figure 3—The broadside collinear curtain array—the NRY—in its final configuration. This antenna is cut for 20 meters, but works well on all the HF ham bands. The array hangs from a catenary rope via three short lengths of twine. The lower end insulators are supported by ropes to those of the top wire, and the center insulators are joined by the phasing line, which is 43 feet, 5 inches of parallel-wire feed line. The feeder to the radio is made of the same type of line connected at the halfway point on the phasing line.

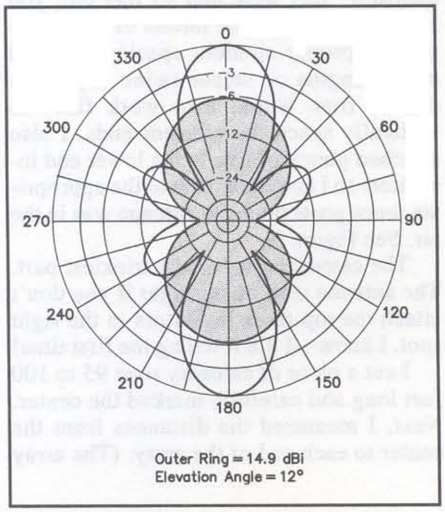

Figure 4—The NRY's 14-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 14-MHz dipole at 78 feet (the same height as the NRY's feed point).

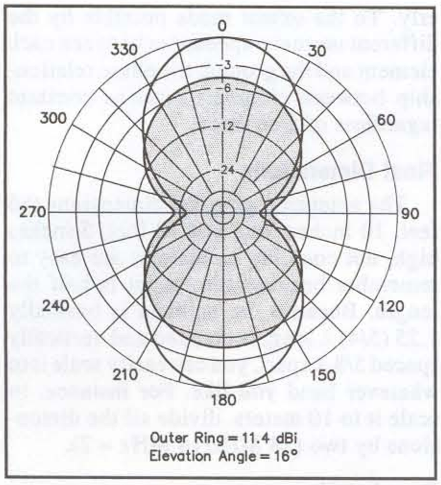

Figure 5—The NRY's 10.1-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 10.1-MHz dipole at 78 feet (the same height as the NRY's feed point).

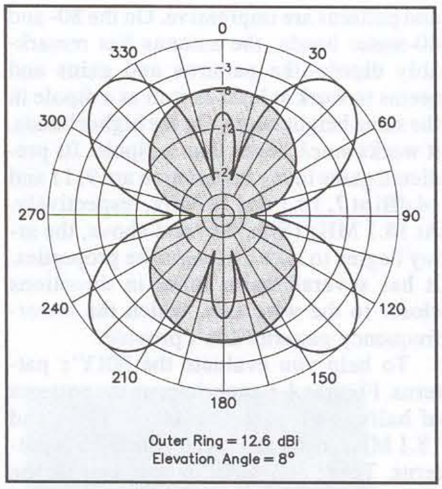

Figure 6—The NRY's 18.1-MHz azimuth-plane pattern with the bottom wire at 56.5 feet and top wire at 100 feet. For comparison, the shaded part shows the pattern of a half-wavelength, 18.1-MHz dipole at 78 feet (the same height as the NRY's feed point).

decrease at 20 meters with a 20° variance from the flat-top configuration. The rope, therefore, must be strong enough to minimize droop and maintain integrity in the sun and wind. We'll get back to the rope in a minute.

For the phasing harness I used low-loss, 300-Ω twinlead. Any parallel-wire feeder should work fine. Don't use coaxial cable in the feed system. It's too heavy, unbalances the array and has very high loss when operated at a high mismatch.[7]

To assemble the feed system, I used porcelain insulators and equal lengths of 300-Ω line. I then cut the #14 antenna wire to the proper length and soldered it to the feed system in a standard dipole configuration.

Next came the tethers. For an antenna almost 87 feet wide and 43 feet tall, you need to provide some means of maintaining the proper element spacings. I used equal lengths of seiner twine (nylon or dacron rope would also work fine) to vertically space the element ends. I also attached plenty of line to the lower end insulators so I could tie them to the appropriate fence posts once the antenna was in the air. See Figure 3.

The catenary rope is the trickiest part. The antenna won't hang right if you don't attach the top three insulators in the right spot. I know—I did it wrong the first time!

I cut a piece of catenary rope 95 to 100 feet long and carefully marked the center. Next, I measured the distances from the center to each end of the array. (The array itself is a handy measuring line.) Then I added about 10 more inches and marked the line. I'll tell you why in a minute.

Then I cut three pieces of twine, each about 24 inches long. At the center of the catenary rope, I carefully opened the rope weave and passed about 6 inches of twine through it, then let the catenary rope retake its natural shape. Wrapping the twine around the rope several times on either side of the joint relieved the tension placed on the rope introduced by passing the twine through it.[8] I then tied the other end of the twine to the top center insulator so that the spacing between the rope and the insulator is 12 inches.

After doing that, I attached the twine at the end marks on the catenary rope in the same manner. The distance to the end insulators is 18 to 20 inches. This causes the ends to hang off the rope at an angle and provides strain relief.

If you use a braided rope (with a weave you can't open to pass twine through it), one good alternative is to use insulators in the catenary rope (at the points where I attached the twine) as tie points for the array.

Now you're ready to haul the antenna up on the lanyards. Consult your US Navy *Marlinspike Seamanship* manual (just kidding, of course) and join the ends of the catenary to the lanyards that are hopefully by now through the pulleys, up in the trees. Use a strong knot that won't slip.

Before hoisting the array into the sky, *make sure* that the connections to the feed harness at the top and bottom are the same (ie, that the left elements both attach to the same wire in the twinlead). This thing does not work right if the top and bottom are fed 180° out of phase!

Now you're ready to haul it up and connect the feed line to the antenna tuner. Check to make sure that the symmetry and top-to-bottom wire spacings are within reasonable limits, then tie the bottom tethers off so that the bottom wire is as flat as you can make it.

You're Done!

This is quite an effective antenna. I've installed two of them: One is boresighted on the Middle East and works like gangbusters into Europe as well; the other favors the Caribbean and most of the US. In the first three months after putting them up, I worked 105 countries on CW, including some rare ones, with a 100-watt transceiver. I'm surprised at how easily I can bust pile-ups—especially on 30 and 20 meters.

I hope you have an opportunity to try this array on for size. Don't limit yourself to just the backyard, either: It's a great Field Day antenna, to say the least. It also comes in handy for emergency work where you need a gain antenna.

You don't have to be a rocket scientist to design effective antennas! Your old friend *The ARRL Antenna Book*, and other ARRL publications, like Walt (W2DU) Maxwell's *Reflections*, the *Handbook*, the *Antenna Compendium* series and Wilfred Caron's *Antenna Impedance Matching*, are serious reference guides and sources of numerous ideas. It doesn't hurt to have your favorite antenna-modeling software loaded on your PC, either!

Notes

[1] D. Brede, "The Care and Feeding of an Amateur's Favorite Antenna Support—the Tree," *QST*, Sep 1989, pp 26-28, 40.

[2] J. Hall, ed, *The ARRL Antenna Book*, Chapter 22 and pp 4-3 and 15-3.

[3] *The ARRL Antenna Book*, Fig 8-38, p 8-32, illustrates how gain varies with the spacing of the two adjacent element ends.

[4] The *Antenna Book*'s Fig 8-45, p 8-35, suggests that a spacing of 5/8 to 2/3 λ yields good broadside gain. For the four-element array, upwards of 8.5 dBi gain is obtainable.

[5] dBi means decibels relative to an *isotropic radiator* (point source) in free space. It *does not* denote gain relative to a point source above real ground, or a dipole in any surroundings.

[6] W. Calvert, "The EZY-Launcher," *QST*, Jun 1991, pp 34-35.

[7] *The ARRL Antenna Book* shows how coaxial-cable loss varies with matched loss and SWR in Fig 18, p 24-18. (The *Antenna Book* also contains a table of matched loss per 100 feet as functions of frequency and cable type in Fig 26, p 24-18).

[8] If you use braided rope for the catenary, you'll need to use another method to secure the seiner twine to the catenary rope.

CHAPTER FOUR

LOOP ANTENNAS

4-1 **Dual Full-Wave Loop Antennas**
From March 1982 *QST*

4-2 **A Gain Antenna for 28 MHz**
From July 1994 *QST*

4-3 **10-Meter "Lazy Quad"**
From July 1968 *QST* (Technical Correspondence)

4-4 **The Full-Wave Delta Loop at Low Height**
From October 1984 *QST*

4-7 **A Triband 75/40/30-Meter Delta Loop**
From *The ARRL Antenna Compendium Vol 5*

4-9 **A Two-Band Loop for 30 and 40 Meters**
From March 1989 *QST* (Hints & Kinks)

4-10 **Feeding an 80-Meter Delta Loop at 160 Meters**
From April 1990 *QST* (Hints & Kinks)

By John Griggs, W6KW **From *QST*, March 1982**

Dual Full-Wave Loop Antenna

Achieving gain from an antenna at 7 and 3.5 MHz normally requires a rather large piece of real estate, a high tower or both. To obtain significantly improved performance over a dipole at these frequencies, and to do it inside an average city lot, is a goal worth pursuing! With this in mind I decided to replace my inverted-V dipoles on 7 and 3.5 MHz with a dual full-wave loop antenna, one inside the other.

I had noticed the strong signals on 7 MHz from Pat Kearins, W7UI (now a silent key), and from Carl Winter, W6OAW, on 3.9 MHz, and was interested to find that both were using full-wave loops. Both signals would pin my S-meter and my QTH is well over 200 miles from either station. The antenna at W7UI was a loop suspended by a very high supporting structure. This loop resembled a square, with one corner facing up, one facing the ground, and the other two corners pulled outward by guy wires.

My attention was attracted to a horizontal loop antenna used by another ham on 3.9 MHz that produced extremely strong signals. It was a square loop, 65 feet on a side, with each leg parallel to the ground but only 14 feet high. Fed by a tuned line, it functioned well, but it was most effective for relatively short ranges, up to 200 miles or so.

Of more interest to me, however, was the 75-meter rectangular loop used at W6OAW. I learned that this was a dual antenna, i.e., a full-wave rectangular loop for 3.9 MHz suspended about 40 feet above ground, with another full-wave loop for 7 MHz inside the first, in the same plane. The 3.9-MHz loop is a closed circuit, whereas the 7-MHz loop is an open circuit. This permits operation on 10, 15 and 20 meters as well as 40 meters when used with open-wire feed line and a matching network. The outer loop is fed with coaxial cable.

Design Planning

In considering loop antennas for my lot, which is 75 feet wide by 125 feet deep, I found that I could use two 40-foot high pipe masts, 130 feet apart along a diagonal across my lot. I duplicated the W6OAW arrangement, except that I hung the antenna vertically and made both loops closed-circuit designs. I could use them only on the bands for which they were cut, 7.2 MHz and 3.8 MHz. Each is fed with a quarter-wave matching section of RG-59/U, which provides a 50-ohm match for the RG-8/U cables leading in to my operating position.

Construction

See Fig. 1 for construction details and dimensions. Matching the 104-ohm antenna impedance to the 50-ohm line impedance requires a quarter-wave coaxial line having a characteristic impedance of 72 ohms. The formula for a $^1/_4$-wave section is:

$$\frac{246}{\text{f (MHz)}} = \ell\,\text{(ft)} \qquad \text{(Eq. 1)}$$

This result must be multiplied by the velocity-factor of the coaxial cable, 0.66 for RG-59/U 73-ohm cable. The quarterwave section was determined to be 42.7 feet for 3.8 MHz and 22.6 feet for 7.2 MHz. These connect to RG-8/U cables for a 50-ohm match to the transceiver.

The top section of the 3.8-MHz loop is fed at the center because of the length of the quarter-wave line, but the shorter matching section for the 7.2-MHz loop allows me to feed the bottom portion of that antenna.

The inner loop is supported from the corners of the outer loop by means of nylon rope and suitable insulators. It is best to use lightweight, high-quality insulators and wire no larger than no. 14 to reduce the weight of the array. Pulleys (and rope) will be required at the tops of the 40-foot poles and also at the 18-foot level. The fact that the bottom horizontal section of the antenna narrowly misses the top of my house roof seems to have no deleterious effect on its operation.

Measure the antenna sections carefully, and cut them to length. Pay particular attention to locate the feed point at the exact center of the horizontal wire sections. This provides horizontal polarization. Feeding the loop at the center of either vertical section will provide vertical polarization.

Tuning

If trimming is needed to resonate the antenna to your favorite operating frequency, cut or add equal length pieces at each vertical section. Do not shorten the sides to the point at which the top and bottom sections are less than 0.1 wavelength apart. Careful pruning of the horizontal sections would be required if such a problem develops.

Information given to me by those using this antenna seems to confirm my results. A typical comment is "Boy are you ever loud!" A gain of 1.8 dB over a dipole is claimed for this type of antenna. The principal difference appears to be a lower angle of radiation. While it is possible to work stations off the ends of the antenna, maximum radiation occurs broadside to the element. What I like about the antenna is the ability to hear weak stations. It is outstanding in this regard.

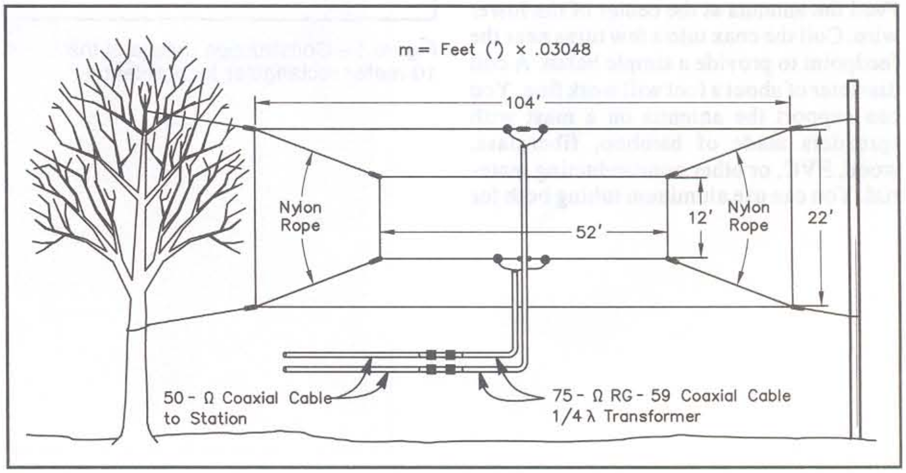
m = Feet (') × .03048
104'
Nylon Rope
52'
12'
Nylon Rope
22'
50 - Ω Coaxial Cable to Station
75 - Ω RG - 59 Coaxial Cable 1/4 λ Transformer

Fig 1—Dimensions for the dual full-wave loop antenna. Input impedance is on the order of 104 ohms. A quarter-wave matching section of RG-59/U is used to provide a match to 50 ohms.

By Brian Beezley, K6STI From *QST*, July 1994

A Gain Antenna for 28 MHz

Give your 10-meter signal a boost with this simple antenna.

Although in coming years the 10-meter band won't provide the excitement it did at the peak of sunspot cycle 22, DX openings will still occur, especially during spring and fall. And who knows? With the large number of 10-meter operators and wide availability of inexpensive, single-band radios, the band may be the hangout for local ragchews that it was before the advent of 2-meter FM.

I'd like to present a simple antenna for 10 meters that provides gain over a dipole or inverted V. Similar designs have been published before, but because I never hear them on the air, I think they must not be fully appreciated. The antenna is a resonant, rectangular loop with a particular shape. It provides 2.1 dB gain over a dipole at low radiation angles when mounted well above ground. This represents a power increase of 62%. The antenna is simple to feed–no matching network is necessary. When fed with 50-Ω coax, SWR is close to 1:1 at the design frequency. SWR is less than 2:1 from 28.0 to 28.8 MHz for an antenna resonant at 28.4 MHz.

The antenna is made from #12 wire (see Figure 1). For horizontal polarization at 28.4 MHz, the loop is 73 inches wide and 146 inches high (just larger than 6×12 feet). Feed the antenna at the center of the lower wire. Coil the coax into a few turns near the feedpoint to provide a simple balun. A coil diameter of about a foot will work fine. You can support the antenna on a mast with spreaders made of bamboo, fiberglass, wood, PVC, or other nonconducting material. You can use aluminum tubing both for support and conductors, but you'll have to readjust antenna dimensions for resonance.

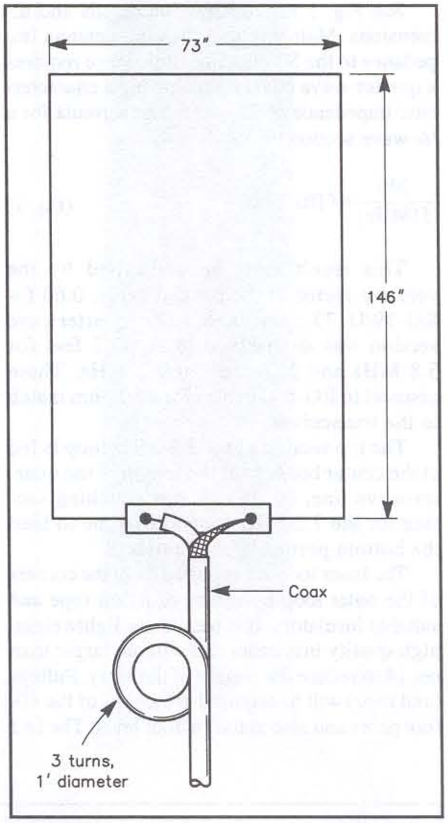

Figure 1—Construction details of the 10-meter rectangular loop antenna.

This rectangular loop has two advantages over a resonant square loop. First, a square loop has just 1.1 dB gain over a dipole. This is a power increase of only 29%. Second, the input impedance of a square loop is about 125 Ω. You must use a matching network to feed a square loop with 50-Ω coax. The rectangular loop achieves gain by compressing its radiation pattern in the elevation plane. The azimuth pattern is slightly wider than that of a dipole (it's about the same as that of an inverted V). A broad pattern is an advantage for a general-purpose, fixed antenna. The rectangular loop provides bidirectional gain over a broad azimuth region.

You should mount the loop as high as possible. To provide 1.7 dB gain at low angles over an inverted V, the top wire must be at least 30 feet high. The loop will work at lower heights, but its gain advantage disappears. For example, at 20 feet the loop provides the same gain at low angles as an inverted V.

A small, 3-element Yagi can provide 6 dB gain over a dipole and great rejection of signals to the rear. If you can install a beam and rotor, you'll find it much more effective than a loop. But for a simple, cheap, gain antenna that can be thrown together quickly, the rectangular loop is hard to beat.

Note: I used the AO 6.0 Antenna Optimizer program to automatically optimize the dimensions of a rectangular loop for maximum forward gain and unity SWR. I used NEC/Wires 1.5 to verify the design with the Numerical Electromagnetics Code.

From *QST*, July 1968 (Technical Correspondence)

10-Meter "Lazy Quad"

Among your readers that like to tinker with antennas there may be some that would like to try the antenna shown in Figure 1. The basic idea came from a station using a somewhat similar configuration on 15 meters, and I make no claim of originating the idea. However, I have never heard of anyone using the antenna on 10 meters. Since it is currently nameless, and has features similar to both a Lazy H and a quad, perhaps it should be known as a "Lazy Quad."

The major advantages of the antenna are: (1) extreme simplicity, (2) feasibility of installation on a light unguyed pole, (3) small horizontal space requirements (as compared to a horizontal dipole), (4) low QRN in receiving (as compared to a vertical dipole), (5) ability to withstand high wind loads and (6) broadband operation.

When looking at loop 1 we see a horizontally-polarized full-wave loop radiating broadside, with maximum currents along the top and bottom horizontal wires H_1 and H_2. When looking at loop 2 we see a similar horizontally-polarized loop with maxiinum current on wires H_2 and H_3. When looking at wires H_1 and H_3 we notice two in-phase horizontal wires spaced a half wave, both carrying maximum current. Minimum currents appear on the vertical portions of the loops where the tuning stubs are inserted.

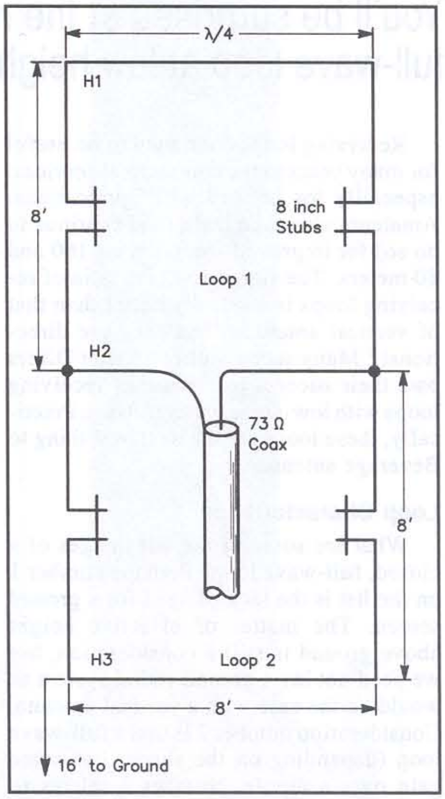

Figure 1—W6PIZ's "Lazy Quad" for 10 meters. The loops are No. 14 wire, with the horizontal sections mounted an 8-foot lengths of 1×2 wood. The adjustable stubs are of the same type wire and are self-supporting since the length is only 6 inches. The antenna at W6PIZ is mounted on an unguyed wooden pole 32 feet high.

I operate mostly on 10-meter c.w. and have the antenna peaked at 28,050 MHz. In my specific case the length of each stub is 6 inches for this frequency. Loading is almost constant from 28 to 30 MHz.

The coax feed line is run straight down the pole (wood) to the ground, and there is little antenna effect on the feeder.

Results have been consistently better than had been obtained with vertical or horizontal dipoles previously installed at the same effective height at the same location. Indications are that it outperforms some beams of the same approximate height, perhaps because of the broad vertical pattern (which allows longer QSOs under critical skip conditions), as compared to the beams.

Perhaps some of your readers might be urged to try this basic idea with a reflector of the same basic configuration as the driven element. It would appear that the Double Lazy Quad would be anything but lazy in operation. — *Dave Hardacker, W6PIZ*

By Doug DeMaw, W1FB and Lee Aurick, W1SE **From *QST*, October 1984**

The Full-Wave Delta Loop at Low Height

You'll be surprised at the results you'll get from a full-wave loop at low heights.

Property size and antenna-support height are ever-present concerns of the urban amateur. Many good antennas are untried because the radio amateur is unable to imagine how a large wire antenna could be squeezed onto a small lot. Certainly, this is typical in the case of full-wave loop antennas. But there is no rule that dictates using a symmetrical loop. It can be distorted rather severely without spoiling the performance. The same philosophy is appropriate with regard to height above ground and the plane in which the antenna is erected. In most instances, a less-than-optimum full-wave loop will outperform a dipole or inverted V antenna that is close to the ground in terms of wavelength. It is possible that such a loop will give comparable or better performance than a vertical antenna that is less than 90 degrees (with respect to ground), or one with a substandard ground screen.

We want to discuss the Practical considerations of loops that can be supported from low supports on small pieces of property. The results we have obtained are noteworthy with respect to all-around "solid" communications within and outside the USA. Perhaps you will be inspired to unroll some wire and try a loop at your QTH.

Some Loop History

Loops were used first as receiving antennas. While single- and multiturn small loops worked well for receiving, they were not satisfactory for transmitting: They were inefficient in terms of gain, and the feed impedance was generally a fraction of an ohm, making them difficult to match. The losses were significant. But, it was possible to use a compact loop (less than 0.5 wavelength) for receiving in place of a fullsize version that could require thousands of feet of conductor. One of us owned a portable broadcast-band receiver in the 1930s. The loop antenna was stored in the lid of the cabinet, and needed to be mounted atop the radio during reception periods! The radio was heavy: it weighed 91 pounds, including the various dry batteries.

Receiving loops continued to be useful for many years in the commercial services, especially for LF and VLF applications. Amateurs also used them (and continue to do so) for improved reception on 160 and 80 meters. The signal-to-noise ratio of receiving loops is markedly better than that of vertical antennas, and they are directional.[1] Many successful 160-meter DXers owe their success to the use of receiving loops with low-noise preamplifiers. Practically, these loops are the next best thing to Beverage antennas.[2]

Loop Characteristics

What are some of the advantages of a closed, full-wave loop? Perhaps number 1 on the list is the lack of need for a ground screen. The matter of effective height above ground is still a consideration, but we need not lay a ground-radial system as would be the case with a vertical antenna. Consideration number 2 is that a full-wave loop (depending on the shape) has some gain over a dipole. Number 3 relates to noise factor. A closed loop is a much "quieter" receiving antenna than are most vertical and some horizontal antennas.

To illustrate this point, the 160-meter antenna at W1FB is a $^3/_8$-wavelength inverted-L with twenty $^3/_8$-wave radials. Since this is essentially a vertically polarized antenna, it is noisy (man-made and atmospheric noise). There are times when an S9 signal is unreadable because of the ambient noise being S9 or greater in strength. Upon switching to the 75-meter Delta loop, the same signal will rise above the noise by 1 or 2 S units, while the noise and signal will drop well below S9. For example, the received signal may drop to S6 on the loop, but the noise will decline to S4.

Feed-point selection will permit the choice of vertical or horizontal polarization. Various angles of radiation will result from assorted feed-point selections. The system is rather flexible when we want to maximize close-in or faraway communications (high angle versus low angle). Figure 1 illustrates various configurations that can be used. The arrangement at C is used at W1SE, and the shape at D is being applied

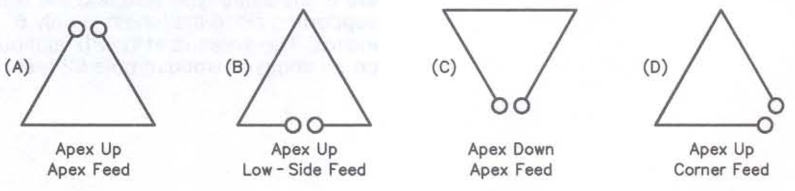

	Polarization	Radiation Angle
A	Horizontal	Moderately High
B	Horizontal	High
C	Horizontal	Moderately High
D	Vertical	Low

Overall Loop Size in feet $\approx \frac{1005}{1\ \text{MHz}}$

Feed Impedance ≈ 100 Ω

Figure 1—Various configurations for a full-wave Delta loop. Radiation angles and polarization are affected by the feed-point placement and location of the apex.

at W1FB. Both antennas are cut for 80-meter operation. The bandwidth at resonance is on par with that of a dipole. A Transmatch is used for matching the system to the transmitter in those parts of the band (75 and 80 meters) where the SWR is too high to deal with.

Our loops are not deployed in a vertical plane, owing to the lack of tower height. A 60-foot tower and 50-foot tree support the W1SE antenna. A single 50-foot tower is used at W1FB. Both loops are tilted away from the supports at roughly 45 degrees (Figure 2). This shows the present W1FB system. The loop is broadside northeast and southwest for maximum radiation in those directions at 80 meters. More on this later.

When these low-to-the-ground experiments began in the summer of 1983, we were joined by Bill Martinek, W8JUY, near Traverse City, Michigan. Bill experimented with various loop configurations so that he and W1FB could make signal comparisons locally and afar. He finally adopted the W1SE format with the apex down (Figure 1C, with the flat top strung between two 50-foot trees). In order to keep the loop completely vertical (not sloping), he chose a triangle that was not equilateral. The upper side of his triangle is substantially longer than the two downward sides. His signal on 75 meters is consistently 10 to 20 dB stronger than with his inverted V. The point of this discussion is that you need not use an equilateral triangle if it will not fit on your property. Erect whatever you can, then give it a try!

Feed Methods

A Q section is used for feeding the W1SE loop. A Q section is a quarter-wavelength line with an impedance that is somewhere between the antenna feed impedance and that of the feed line. Calculation is a simple matter:

$$Z \text{ (Q section)} = \sqrt{Z1\ Z2} \text{ ohms} \qquad \text{(Eq. 1)}$$

where Z1 is the antenna impedance, and Z2 is the feeder impedance in ohms. In this case, assuming approximately 100 ohms for the antenna. feed impedance, we would have $\sqrt{100 \times 50} = 70.7$ ohms for the Q-section impedance. This represents a close match to 52-ohm coaxial cable. The Q-section length (made from RG-59/U) can be determined from L(feet) = 246 V/f(MHz), where V is the velocity factor of the coaxial line for the matching section. (The length should be verified using a dip meter.) For operation at the W1SE-chosen frequency of 3.825 MHz, the calculation calls for a Q section of 42 feet 5 inches (Figure 3).

Open-wire feed is used at W1FB (Fig 6B) to permit multiband operation through 10 meters. Unfortunately, a short run of RG-8/U was needed to bring the feed line to the ham station–under the driveway. The coaxial cable was buried in the ground for this reason. A homemade 4:1 toroidal balun transformer (two stacked T200-2 Amidon

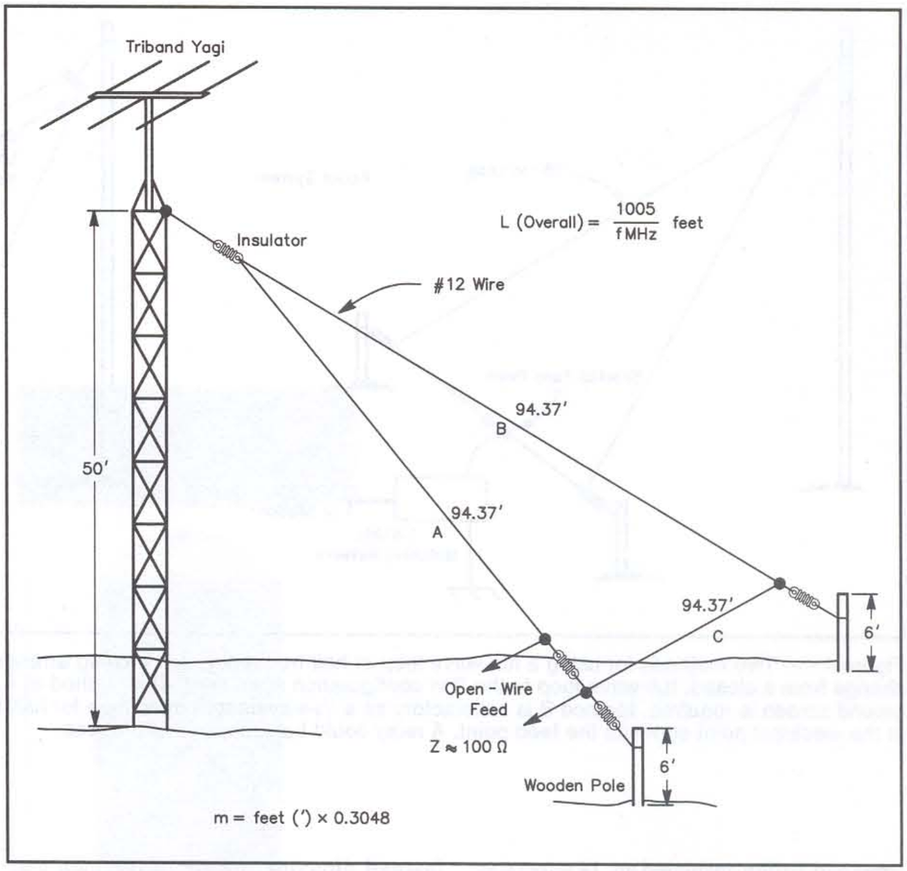

Figure 2—A tilted Delta loop for 80 meters is used at W1FB. The tower height is only 50 feet. Homemade open-wire line is used as the feeder to permit multiband use with vertical polarization and a low radiation angle.

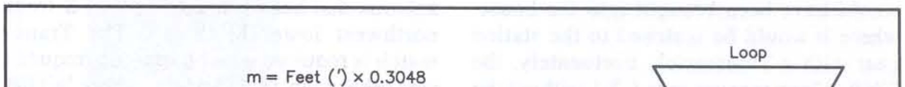

Figure 3—At A is the feed method used at W1SE. A coaxial Q section closely matches the 100-ohm feed impedance to a 52-ohm coaxial line. Illustration B shows the W1FB feed arrangement. Open-wire line, a balun transformer and a short length of RG-8/U cable permit multiband use with a Transmatch. Ideally, the open-wire line would continue all the way to the Transmatch, and the balun transformer would be located at the Transmatch.

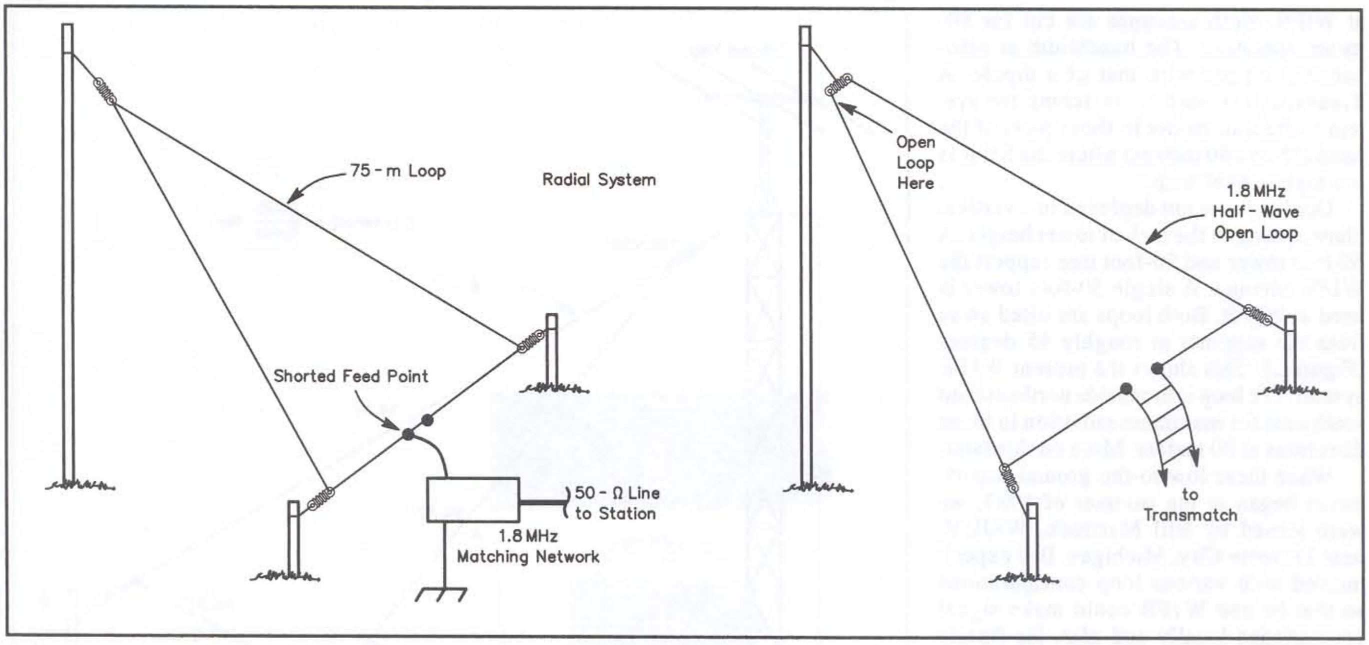

Figure 4 — Two methods for using a full-wave loop at half frequency. A switching arrangement could be applied at the feed point to change from a closed, full-wave loop to the first configuration seen here. The method at A performs as a 1/4-wavelength radiator, but a ground screen is required. Method B is satisfactory as a 1/2-wavelength open loop for half-frequency use. It requires opening the loop at the electrical point opposite the feed point. A relay could be used for this purpose.

cores and Teflon-insulated no. 14 wire) was enclosed in a weatherproof box and mounted on one of the support poles for the 450-ohm open-wire line. The RG-8/U was run underground from that location (about 25 feet). Ideally, the openwire line would have been brought into the house, where it would be matched to the station gear with a Transmatch. Fortunately, the SWR at loop resonance is 1.3:1 without the Transmatch in use.

Performance

This is the part of our article that many of you have been waiting to read. Well, the W1FB results have been entirely gratifying. The loop replaced an inverted V with an apex height of 50 feet. This led to a pronounced improvement in all-around communications on 75 and 80 meters out to 500-600 miles. But, the loop proved to be very effective also for DX communications to Europe on 80 meters. The first version was that of Figure 1B. Although the antenna was outstanding for close-in 75- and 80-meter work, it offered dismal DX performance. The configuration at D of Figure 1 seems to offer a good compromise in performance for local and DX work. The theoretical launch angle to the horizon at the loop fundamental frequency is 10 degrees, as reported by VE2CV in a letter to W1FB. This assumes that the loop is erected vertically and at a reasonable height above ground.

Harmonic operation of the loop, as depicted in Figure 1D, is superb. At times it outperforms the trap tribander atop the tower during DX operation to Europe and Africa. The loop shows an average 6-dB signal increase on 20 and 15 meters in the favored direction, owing to the gain and lower radiation angle of the loop. Radiation at the harmonics is in the plane of the loop rather than broadside to it. This makes it ideal for contacts into Africa. It is perhaps the most effective 40-meter DX antenna that has been used at W1FB from northwest lower Michigan. The Transmatch is required on all harmonic frequencies other than 18.111 MHz, where W1FB has been conducting propagation studies with Bill Orr, W6SAI, Prose Walker W4BW, Bob Haviland, W4MB and Stu Cowan, W2LX, under special experimental/research licenses (KM2XQV). The loop has worked very well on 24.9 MHz as well during these tests. At 18.111 MHz, the SWR is 1.4:1.

The operating results at W1SE also indicate that a tilted loop, close to the ground, functions quite well. With loop resonance at 3825 kHz, the 2:1 SWR points occur at 3734 and 3934 kHz, respectively. This 200-kHz bandwidth spectrum can be shifted up or down the band by lengthening or shortening the loop conductor and Q section accordingly. From the W1SE location in Newington, the loop has delivered impressive performance for local and DX work.

A 40-meter Delta loop was constructed for use at W1SE after noting the fine performance of the 80-meter system. It was cut for resonance at 7016 kHz. This model was erected in a completely vertical format, using 143 feet 3 inches of wire. The Q section is 23 feet 2 inches long. The apex (feed point) is 4 feet above ground. The SWR on 40 meters is less than 2:1 across all of the band. The 80- and 40-meter W1SE loops showed resonance slightly apart from the design frequency, perhaps because of the proximity of the antennas to ground. Resonance on 40 meters was checked as 7050 kHz. Both loops are performing better for local and DX contacts than any of the many antenna types tested at W1SE. We would be even more impressed if we could elevate our Delta loops so the lower portions were a half wavelength or greater above ground.

In Conclusion

There is no rule that dictates the shape of a full-wave loop. The triangular format is convenient for mounting the radiator. If the apex is at the top, only one high support structure is needed. You may have one or more tall trees that can be used as supports. Circular, square or rectangular shapes have been used by many amateurs and the results were good. Certainly, a loop is an impressive receiving antenna, in terms of noise reduction. In some urban locations, that may be more important than transmitting a "death-ray" signal! There is something to be said about the age-old expression, "If you can't hear 'em, You can't work 'em."

An 80-meter Delta loop can be used on 160 meters by adopting one of two simple methods (Figure 4). A closed loop does not, however, offer good results when the overall length is a half wavelength. Either of the techniques in Figure 4 will work, but the method at A requires a ground radial system for best results.

Notes

[1] D. DeMaw, "Beat the Noise with a Scoop Loop," *QST*, July 1977, and "Maverick Trackdown," *QST*, July 1979.

[2] H. H. Beverage and D. DeMaw, "The Classic Beverage, Revisited," *QST*, Jan. 1982.

By David J. Crockett, WB4DFW **From *The ARRL Antenna Compendium Volume 5***

A Triband 75/40/30-Meter Delta Loop

WB4DFW wanted to use his Delta Loop on other bands besides 75 meters, so he installed some coaxial traps. He also reveals an effective technique to prevent breakage when support trees sway in the wind.

I've been using horizontally mounted, coax-fed Delta Loop antennas on 75 meters for many years. I've found them to be superior to dipoles—they are generally quieter on reception (perhaps because they are dc-grounded), exhibit broader bandwidth (typically less than 2:1 SWR over more than 200 kHz using a coaxial matching section), and are often more easily adaptable to available tree supports. They also don't fall down very often when installed correctly!

Outside of using an antenna tuner, a 265-foot-long 75-meter loop doesn't lend itself to working other bands–and I don't have a tuner. So while listening to a virtually dead 75-meter phone band one day and wishing I had a way to hit 30 meters for a long-overdue dose of CW, I had a brainstorm. Why not use traps? I'd homebrewed coaxial traps based on Sommer's article[1] several times in the past, and I knew they were easy to make and adjust.

The Traps

A quick examination of the latest incarnation of my 75-meter loop found it to be a reasonably equilateral triangle (a little over 85 feet per leg). I calculated that a 30-meter dipole should be something over 46 feet in total length (465/frequency in MHz). It appeared to be a simple enough matter to station some homebrewed coaxial traps 23 feet on each side of the feedpoint. A little more calculating found that it might even be practical to put in a pair of 40-meter traps as well.

Sommer gives a useful set of nomographs for computing trap dimensions on PVC pipe forms, using either RG-58 or RG-174 coax (the former capable of handling a kW with few problems). I opted for the larger cable. The 30-meter traps each require 46.8 inches of coax on a 1.62-inch OD section of PVC (6.75 turns). The 40-meter versions each use 61 inches of coax on a 2.25-inch OD PVC section (6.0 turns). Both PVC sections can be any convenient length. I used four-inch long pieces.

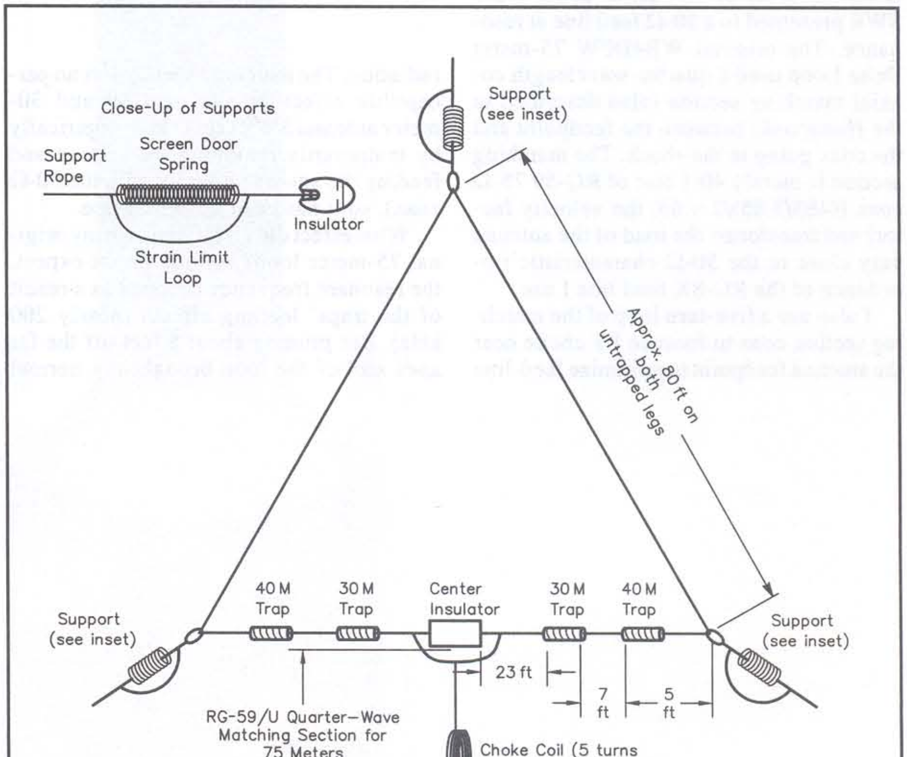
Close-Up of Supports
Screen Door Spring
Support Rope
Insulator
Strain Limit Loop
Support (see inset)
Approx. 80 ft on both untrapped legs
40 M Trap
30 M Trap
Center Insulator
30 M Trap
40 M Trap
Support (see inset)
Support (see inset)
23 ft
7 ft
5 ft
RG-59/U Quarter-Wave Matching Section for 75 Meters (40.1 ft, including choke coil)
Choke Coil (5 turns approx. 1 ft diameter)
RG-8X (any length) to shack

Figure 1—Overhead view of the Triband 75/40/30-meter Delta Loop, which WB4DFW mounted 40 feet off the ground. The traps were constructed with RG-58 coax. See text for winding details.

I wound the 30-meter traps in less than an hour and inserted them in the loop antenna 23 feet either side of the feedpoint. A quick SWR check found them resonant well below the bottom of the 30-meter band, but spacing the turns slightly apart and securing them with electrical tape raised the resonance easily enough, with the SWR bottoming out at 1.2:1 at 10.125 MHz.

The 40-meter traps went together with similar ease. Placement in the existing antenna was a little more challenging than with the first set of traps. The loading caused by the 30-meter traps made determining the physical location for the 40-meter traps uncertain. Grebenkemper notes that traps in a half-wave antenna tend to "pull" an antenna toward resonance,[2] even if they are not placed at exactly the optimum location. So, I simply took the formula length for 40-meter resonance, generously subtracted the physical length of the coax in each of the 30-meter traps and installed the 40-meter traps accordingly.

The approach was unscientific, but SWR was 1.3:1 at 7.1 MHz on the first try. Dumb luck, I guess, and I opted to make no further adjustments. The final antenna layout is shown in Figure 1, and a close-up photo of a trap is shown in Figure 2.

The Matching Section

The ARRL Handbook notes that the typical impedance of a full-wave loop antenna is on the order of 100 Ω, or about a 2:1 SWR presented to a 50-Ω feed line at resonance. The original WB4DFW 75-meter Delta Loop used a quarter-wavelength coaxial matching section (also described in the *Handbook*) between the feedpoint and the coax going to the shack. The matching section is merely 40.1 feet of RG-59 75-Ω coax [(465/3.85)/2 ×.66, the velocity factor] and transforms the load of the antenna very close to the 50-Ω characteristic impedance of the RG-8X feed line I use.

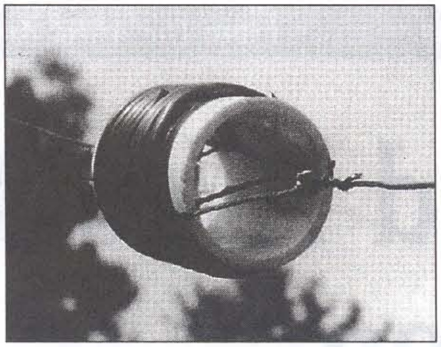

Figure 2—Close-up photo of one of the 40-meter traps. The ends of the coax cable used in each trap have been sealed against moisture with silicone caulking. See text for construction details.

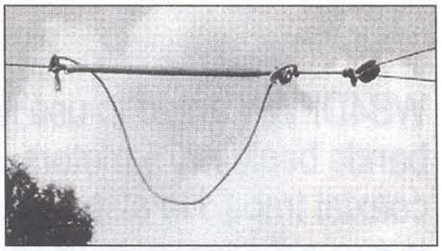

Figure 3—Photo of one of the support system springs, with rope used to limit total stretch. The rope is twice the length of the unstretched spring.

I also use a five-turn loop of the matching section coax to form an RF choke near the antenna feedpoint to minimize feed-line radiation. The matching section has no perceptible effect on the new 40 and 30-meter antenna SWRs (checked empirically by temporarily removing the section and feeding the antenna directly with the 50-Ω coax), so it has been left in service.

What effect did all this have on my original 75-meter loop? As one might expect, the resonant frequency dropped as a result of the traps' loading effects (nearly 200 kHz). But pruning about 8 feet off the far apex end of the loop brought my normal 1.2:1 resonant SWR back to my regular 3.842-MHz stomping grounds.

Keeping It Up in the Air

Here in South Carolina, pines are a great source of antenna support but their swaying during summertime thundershowers can wreak havoc on long, unprotected wire antennas. Over the years, I've found a simple shock-absorber system using a run-of-the-mill screen-door spring placed between a support rope and each antenna insulator can save a lot of grief. See Figure 3 for a closer look at the details.

A special aspect of my system is the employment of a "strain limit rope" in parallel with each spring. This piece of rope, approximately twice the length of the unstretched spring, keeps swaying pines from overly taxing the spring and gives a nice visual reference for gauging how much tension the antenna is operating under. It also provides a fail-safe for the inevitable corrosion-caused failure of the spring after several years aloft. It makes a fine bird perch, too.

So, How Does It Play?

Performance on all three bands has been most satisfactory. The 2:1 SWR bandwidth of the 30-meter section far exceeds the band edges. Similar 40-meter bandwidth is about 100 kHz (noticeably narrower than dipoles I've used, and I suspect partially attributable to my guesswork on the 40-meter trap placement). The 75-meter bandwidth is now approximately 150 kHz (slightly narrower than before the traps were installed). Signal reports both ways on all three bands are comparable to any other non-directional HF antennas I've used.

Notes and References

[1] Robert C. Sommer, "Optimizing Coaxial Cable Traps," *QST*, Dec 1984, pp 38-42.

[2] John Grebenkemper, "Multiband Trap and Parallel HF Dipoles-A Comparison," *QST*, May 1985, pp 29-30.

From *QST*, March 1989 (Hints & Kinks)

A Two-Band Loop for 30 and 40 Meters

After trying to find a way to place a 30-m delta loop inside an existing 40-m loop, I remembered an article in *All About Cubical Quad Antennas*[1] describing a $1\frac{1}{2}$-λ, or "Mini X-Q," loop. The gain of this antenna was said to be about 1 dB more than a 1-λ loop. I installed a large, ceramic SPST knife switch in the center of the delta-loop's bottom leg (see Fig 11). With this switch open, the full-wave, 40-m loop becomes a $1\frac{1}{2}$-λ, 30-m loop! The resonant frequency of this arrangement was 10.5 MHz. By adding 18-inch wires to the loop at both sides of the switch, I obtained resonance at 10.125 MHz.

Since the bottom of the loop is only 12 ft above ground, it's a simple matter to reach the band switch from ground level. (Caution: High RF voltage appears at the switch when the antenna is used for transmitting on 30 m.) Incidentally, the loop also works well on 15 m (SWR under 2:1 across the band) when set for 40 m, and I have used the 30-m configuration successfully on 80 m with the help of an antenna tuner.—*James Brenner, NT4B*

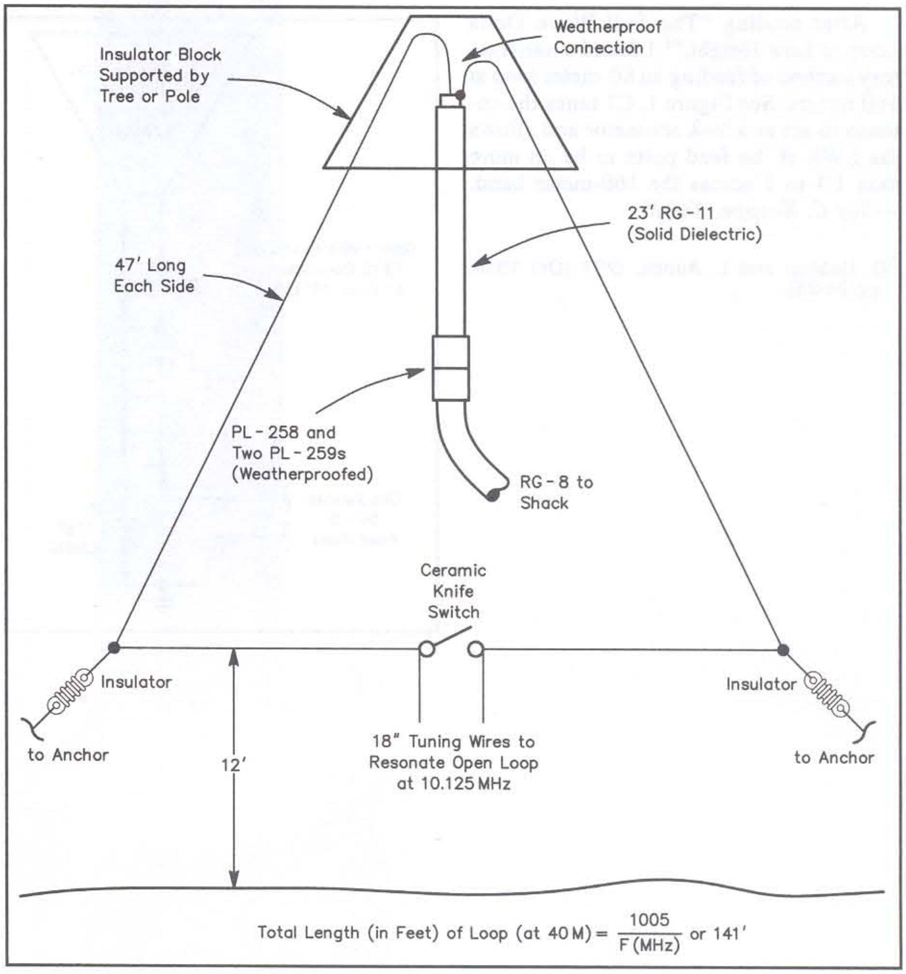
Insulator Block
Supported by
Tree or Pole
Weatherproof
Connection
23′ RG - 11
(Solid Dielectric)
47′ Long
Each Side
PL - 258 and
Two PL - 259s
(Weatherproofed)
RG - 8 to
Shack
Ceramic
Knife
Switch
Insulator
Insulator
to Anchor
to Anchor
12′
18″ Tuning Wires to
Resonate Open Loop
at 10.125 MHz
Total Length (in Feet) of Loop (at 40 M) = 1005 / F (MHz) or 141′

Fig 11—Jim Brenner's 30- and 40-meter loop. Note the 18-inch tuning wires used to lower the antenna's 30-m resonance from 10.5 to 10.125 MHz. The antenna is top-fed via a $\frac{1}{4}$-λ 40-m matching section. See text.

---

[1] W Orr and S. Cowan, *All About Cubical Quad Antennas* (Wilton, CT: Radio Publications, 1970).

Feeding an 80-Meter Delta Loop at 160 Meters

After reading "The Full-Wave Delta Loop at Low Height,"[1] I found a satisfactory method of feeding an 80-meter loop at 160 meters. See Figure 1. C1 tunes the antenna to act as a $^3/_4$-λ resonator and allows the SWR at the feed point to be no more than 1.1 to 1 across the 160-meter band. —*Roy C. Koeppe, K6XK*

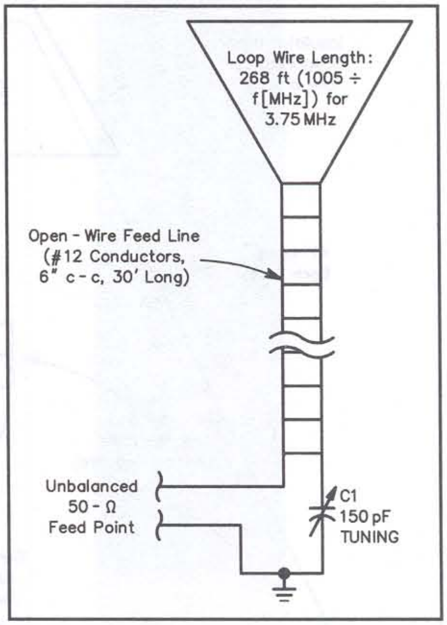
Loop Wire Length: 268 ft (1005 ÷ f[MHz]) for 3.75 MHz
Open - Wire Feed Line (#12 Conductors, 6″ c - c, 30′ Long)
Unbalanced 50 - Ω Feed Point
C1 150 pF TUNING

Figure 1—Roy Koeppe enjoys satisfactory 160-meter operation with an 80-meter delta loop by feeding the loop as shown here. On 80 through 10 meters, Roy takes C1 out of the circuit and feeds the loop via its open-wire feed line and a balanced tuner.

---

[1] D. DeMaw and L. Aurick, *QST* (Oct 1984, pp 24-25).

CHAPTER FIVE

COLLINEAR ANTENNAS

5-1 **The Extended Double-Zepp Antenna**
From June 1938 *QST*

5-7 **Try an Extended Double-Zepp Antenna**
From February 1992 *QST* (Hints & Kinks)

5-8 **Notes on Wire-Antenna Construction**
From March 1992 *QST* (Hints & Kinks)

5-10 **Collinear Arrays**
From *The ARRL Antenna Book* 18th edition

By Hugo Romander, W2NB **From *QST*, June 1938**

The Extended Double-Zepp Antenna

Simple Antenna Structures Having Improved Gain and Horizontal Diversity

The questions of antenna directivity and antenna "gain" are becoming increasingly popular in discussions on antenna systems among amateurs, and it would therefore seem proper to preface an article dealing with directive antenna arrays with a warning, or perhaps a reminder, that one cannot have high gain and radiate in all directions at the same time. To most amateurs this is an obvious fact, but it may not be as well known that this principle is almost equally important in the *vertical* plane. Those of us fortunate enough to have available large open spaces in which to hang wires will find long-wire antennas, such as the "V" and the horizontal diamond, the easiest way to obtain high gain, but the radiation from such antennas is restricted in the vertical plane fully as much as in the horizontal plane. The result may well be that in the very direction such an antenna is supposed to work best, a simple horizontal doublet will put in a far better signal at certain distances.

It would seem, therefore, that the most practical antenna for a variety of distances is one with a fairly wide radiation pattern in the vertical plane.[1] The "stacking" of elements or the use of long wires is not recommended for distances under 1000 miles, except for the very short distances normally reached by the ground wave. The most universal high-gain antenna must restrict its radiation in the horizontal plane only and its height above ground must also be considered to obtain the best compromise in the vertical plane. It is the purpose of this article to discuss the merits of a simple antenna array which combines these desirable features.

THE DOUBLET

Let us discuss, first, the simple doublet, since this antenna will serve admirably as our basis of reference or comparison. To be more specific, consider this doublet as suspended horizontally and fed at one end in the time-honored fashion of the Zepp feeder, as in Figure 1A. Ignoring, for the moment, the fact that the open-ended feeder wire will be at somewhat higher potential than the other feeder wire and will therefore radiate, the horizontal radiation pattern about this doublet, if you're lucky, will be about as shown on Figure 2. Of course, it is assumed the antenna is sufficiently remote from power-line wires and house plumbing to be unaffected by such linear conductors, since our problem is complicated enough without having to consider the mutual impedance between our doublet and the neighbor's b.c.l. antenna!

This pattern of the horizontal doublet is, no doubt, familiar to most amateurs. Less familiar, perhaps, is the vertical radiation pattern (in the direction of maximum horizontal radiation) as shown in Figure 3. Here it is assumed that the height of the doublet above ground is one-half wavelength and that the earth has perfect conductivity. Fortunately, even the relatively low-conductivity soil and sand on Long Island reflects a high percentage of the horizontally polarized waves radiated at angles less than 50° to the earth's surface. Such reflected

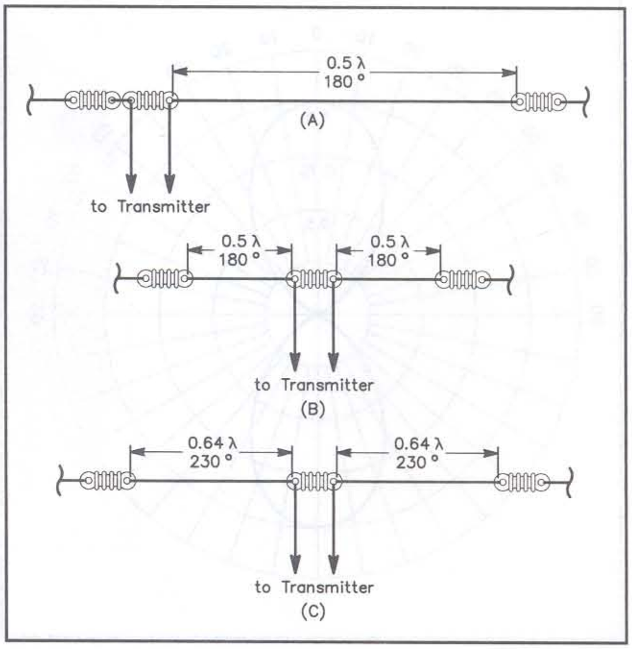

Figure 1—Basic Antenna Configurations (A) The half-wave Zepp; (B) Double Zepp or "two half-waves in phase"; (C) Extended double-Zepp.

---

[1] Although this viewpoint appears to be opposed to that expressed in the article "Simple Directional Arrays Using Half-Wave Elements," by N. C. Stavrou, in May, 1938 *QST* there is no actual conflict. As the present author points out, ihe broad vertical characteristic is to be preferred when the antenna is to give optimum results over short as well as long distances; the former article was concerned with long-distance transmissions, where the lowest possible angle gives best results under nearly all conditions. The type of work to be carried on naturally will be a determining factor. In any event, the simpler structures are not likely to be too sharp in either plane for satisfactory general work. —EDITOR

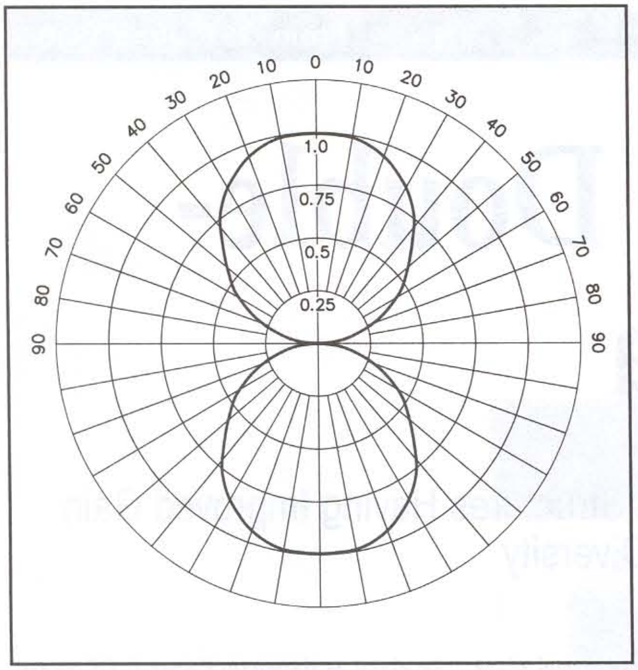

Figure 2—Horizontal pattern (relative field strength) of a doublet half-wave antenna.

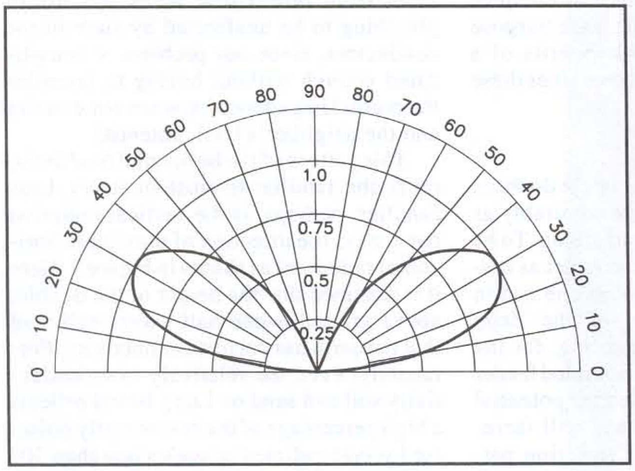

Figure 3—Vertical-plane pattern for a horizontal antenna, ½-wave above perfect earth in direction of maximum radiation.

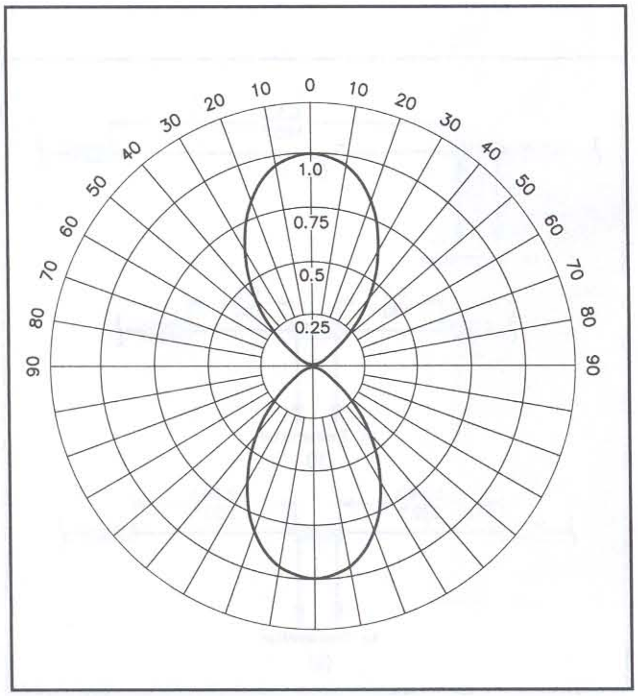

Figure 4—Horizontal pattern of a double-Zepp or two half-waves in phase.

waves, combining with the direct radiation from the antenna within the range of vertical angles in which we are most interested, are responsible for a maximum gain of nearly 6 db over the same doublet in "free space"; that is, without the presence of the earth.

This gift of 6 db must not be taken too much for granted, however. Consider the fact that most amateur communications at high frequencies utilize vertical angles ranging from 10 degrees to 50 degrees. That 6 db must, therefore, come from radiations reflected from the earth's surface quite some distance from the doublet, and in the desired direction of transmission. If the ground slopes sharply upward, or houses or wires are in this area, it becomes questionable, indeed, whether any great portion of the available 6 db is realized. This is especially true for the lower angles of radiation, but usually nothing can be done about this situation, so let us see if we can make up for our ground reflection losses by increasing the horizontal directivity.

THE DOUBLE-ZEPP

By the simple expedient of attaching another doublet to the open-ended terminal of the Zepp feeder, as in Figure 1B, and hanging this doublet parallel and coaxial with the original doublet, an appreciable gain may be obtained. The horizontal radiation pattern will be as shown in Figure 4. The gain should measure about 1.9 db, corresponding to a 55 per cent increase in power. This antenna is widely used among amateurs and is popularly known as the "double-Zepp" antenna, or "two half-waves in phase." A fair amount of gain and a reduction in interfering signals from end-wise directions is easily realized, but the gain is rather disappointing, since it would seem one is entitled to 100 per cent gain in power when doubling the number of radiating members.

The reason why only 1.9 db is realized with the double-Zepp is made clear only after a mathematical study of the situation. Briefly, the close proximity of the two doublets causes a mutual coupling between them, and this coupling (mutual impedance) has an adverse effect on the radiation resistance insofar as gain in the broadside direction is concerned. Obviously, the thing to do is to move the doublets further apart, but this complicates the method of feed. A much simpler way of obtaining increased gain was evolved by Mr. A. A. Alford, of Mackay Radio and Telegraph Company, from the principles discussed in Mr. G. H. Brown's article on broadcast antennas in the Proceedings of the I.R.E. for January, 1936. Mr. Alford presented this idea in a paper delivered at an I.R.E. meeting in Washington.

EXTENDED DOUBLE ZEPP

The gain of the double-Zepp may be increased from 1.9 to 3.0 db by the simple expedient of increasing the length of each doublet until its electrical length is 0.64

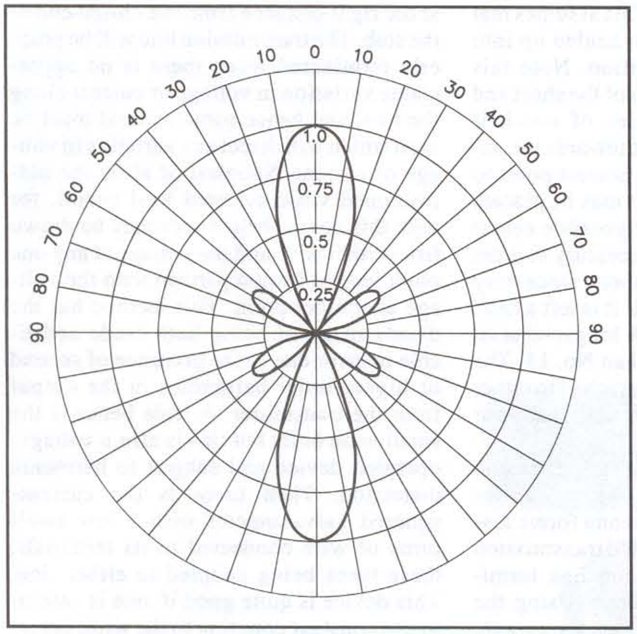

Figure 5—Horizontal pattern of double 230-degree Zepp

wavelength instead of 0.5 wavelength. In electrical degrees the double-Zepp consists of two 180-degree elements; the *extended* double-Zepp for maximum gain should consist of two 230-degree elements. See Figure 1C. In this way the power gain may be increased from 55 per cent to 100 per cent. The gain decreases rapidly for extensions beyond 230 degrees, and therefore, when operating over a band of frequencies, each of the two elements should not exceed 240 degrees for the highest frequency. The horizontal pattern for the double 230-degree Zepp is shown in Figure 5. The vertical pattern in a plane perpendicular to the antenna will be the same as for the simple doublet.

ANTENNA IMPEDANCE

The impedance of this antenna at the termination of the transmission line is of interest, since it has an important bearing on the standing-wave ratio of current or voltage in the line, and it will be compared with that of the ordinary double-Zepp. The double 180-degree antenna presents an impedance of approximately 4400 ohms of almost pure resistance as a termination for the transmission line. This value will be slightly affected by the size of wire used in the antenna and, to a moderate extent, by the height above ground or the influence of nearby conductors, and so the "free space" value is given for No. 14 wire. With this antenna and a 600-ohm surge-impedance line the ratio of maximum to minimum current along the line will be 4400 divided by 600, or 7.3. Incidentally, the terminating resistance of the simple Zepp-fed doublet is about 12,000 ohms, resulting in a standing wave ratio of 20 on our 600-ohm line. For calculation of the surge impedance of the line the reader is referred to *The Radio Amateur's Handbook*.

The impedance at the center of the double 230-degree antenna is not a pure resistance, and hence its effect upon the transmission line be such that maxima or minima of voltage and current along the line will not be odd multiples of a quarter wave from the antenna, as with the ordinary Zepp or double-Zepp antennas. As might be expected, the current or voltage maxima will be shifted towards the antenna, since the two doublets are longer than normal, and this shift is approximately 0.13 times the wavelength. At any rate, the antenna impedance is such that the equivalent pure resistance at any voltage maximum will be about 6000 ohms; that is, the standing-wave ratio will be 10 on a 600-ohm line.

Knowing the standing-wave ratio, it becomes an easy matter to calculate the input resistance to the transmission line if it was cut, let us say, at any current maximum. Thus, for the simple Zepp-fed doublet, this resistance would be 600 divided by 20, or 30 ohms. For the ordinary double-Zepp antenna this resistance would be 600 divided by 7.3, or 92 ohms. For the double 230 degree antenna this resistance would be 600 divided by 10, or 60 ohms. Those who do not use correcting stubs or other methods to obtain a "flat" line (standing wave ratio of 1), but who cut their transmission lines at the maximum current point after the manner of the "tuned feeder," will find these figures useful in calculating the power output of their transniitters.

FEED LINES

However, the use of flat lines, in amateur circles usually referred to as untuned lines, is becoming increasingly popular as their merits are more generally known, and the figures given are essential in calculating the length and position, for example, of a correcting stub, the latter being one of the simplest devices for reducing the standing-wave ratio to unity. A discussion of the use of the correcting stub may be found on page 307 of the *Handbook* (15th edition) where a table indicates the lengths and positions of the stub for various standing-wave ratios. The shorted stub or "loop" will generally be most practical, and some idea of dimensions may be obtained from the fact that when using the double 230-degree Zepp on 14,200 kHz the stub should be 3.5 feet long; that is, it should be composed of the same kind of wire as is used in the transmission line, this wire being 7 feet long and bent to form a U-shaped rectangle with a width equal to the transmission line. At the frequency mentioned, where each extended doublet is 43 feet long, the correct location for the stub will be about 8 feet from the antenna.

Ordinarily the best place to introduce power into an antenna is at a point symmetrically located with respect to the opposite ends of the system. This will not always be practical, because of space limitations, and hence it may become necessary to feed the extended double-Zepp antenna at one end in true Zepp fashion. Some sacrifice in gain will result, due, for the most part, to unequal distribution of current between the two doublets and to radiation from the feeder because it has an unbalanced load. Physical dimensions of feeder, extended doublets, and phasing stub are shown in Figure 6. Note that the phasing-stub length is shorter than a quarter wavelength, but by an amount not exactly

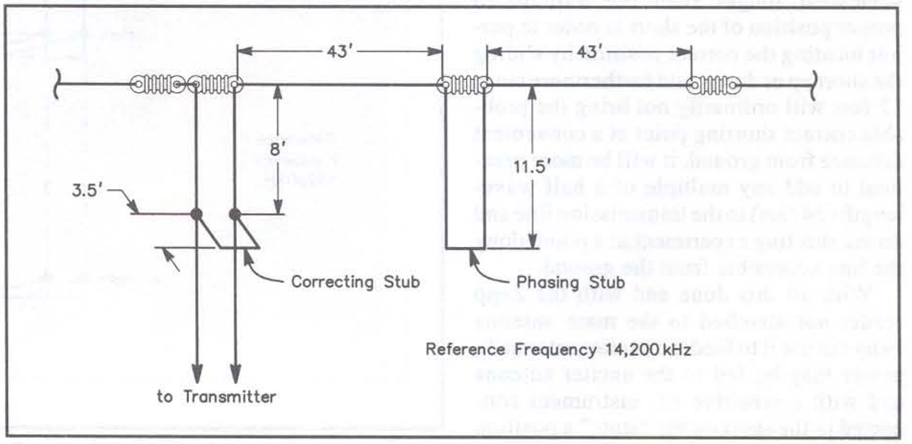

Figure 6—End-fed double 230-degree Zepp with 600-ohm non-resonant transmission line.

equal to the extension of each doublet beyond a half wavelength. This discrepancy is due to interaction between the doublet wires and the wires of the phasing stub. Again a reference frequency is given to simplify calculation of lengths at other frequencies by inverse proportion. The figures given should be regarded as approximations, and tests made individually to determine dimensions and positions of important elements.

ANTENNA ADJUSTMENTS

One of the most effective methods of adjusting an antenna system is to use the transmitter to excite, at the desired frequency, another antenna stretched out, perhaps temporarily, at least 8 or 10 feet from the ground. This "exciter" antenna should preferably be at least one-half wavelength away from the antenna to be adjusted. If the two are parallel to each other, so much the better. Using the end-fed extended double-Zepp antenna as an example, the first consideration is the length of each doublet, and this brings up the problem of "end-effect." We have approximately solved this by cutting our doublets 5 per cent short of the theoretical length. Actually the end effect, which normally exists only at the ends of the antenna, depends on the wire size and the size of the insulator fitting (if metal caps are used), as well as the frequency. Moreover, the end-effect does not vary quite as rapidly as a direct function of the wavelength, so that if an end-effect of 1 foot is correct at 14,200 kHz, it is more nearly 1.5 feet than 2 feet at 7100 kHz.

The theoretical "uncorrected" length of each half of an extended double-Zepp antenna at 14,200 kHz is 44.3 feet. The end-effect may be no more than 1 foot, but since it is better that the antenna be a little too short than too long, let's make the correction 1.3 feet, so that each half will be exactly 43 feet long. At the center insulator which separates the two halves a "stub" must be connected, if the system is to be end-fed. Now, this stub will, take the form of a transmission line shorted at the end opposite from the antenna, and for the frequency under consideration its shortest length will be about 12 feet. Since it is our intention to make this transmission line somewhat longer than the anticipated proper position of the short in order to permit locating the correct position by sliding the short up or down, and furthermore since 12 feet will ordinarily not bring the probable correct shorting point at a convenient distance from ground, it will be more practical to add any multiple of a half wavelength (34 feet) to the transmission line and do the shorting experiment at a point along the line accessible from the ground.

With all this done and with the Zepp feeder *not* attached to the main antenna (why not use it to feed the exciter antenna?), power may be fed to the exciter antenna and with a sensitive r.f. instrument connected in the short on the "stub," a position of the short may easily be found where maximum current flows. This assumes that the main antenna has been hauled up into its normal operating position. Note this maximum-current position of the short and measure in exact multiples of one-half wavelength from this point towards the antenna, thus arriving at the nearest point to the antenna at which a short may be placed. This will give the shortest possible center stub. Of course, it is not necessary that the center stub be made that short, since very little loss will be incurred if it is left a half-wave or even a wavelength longer–that is, if the wire is not smaller than No. 14. The Zepp feeders may now be attached to either end of the antenna, as it is now ready for operation.

The method of adjusting the center stub just outlined is also useful where it is desired to feed any of the antenna forms discussed above with an untuned transmission line—that is, a transmission line terminated by its surge impedance. Using the example cited in the previous paragraph, let us assume the point of maximum current through the short has been found and, after soldering a piece of wire across the line at this point, the extra length of transmission line is chopped off. The two-wire line from the transmitter may now be tapped in above this shorted end of the stub, and it is only a question of position of the feeder tap from the shorted end to terminate the line properly. See Figure 7. This point will be approximately 3.3 feet from the short for a 600-ohm transmission line. Of course, when all this is done, we will have the same transmission line with correcting stub discussed earlier in this article.

LINE-CURRENT MEASURING DEVICES

The same principle holds true for the double Zepp and for center feeding the simple doublet, the only ambiguous point being that of knowing when the transmission line from the transinitter is connected at the right distance from the closed end of the stub. The transmission line will be properly terminated when there is no appreciable variation in voltage or current along the line, and hence some method must be used which will detect any variation in voltage or current. Simplest of all is the old-fashioned wood-covered lead pencil, for with this device bright arcs may be drawn from either wire and the voltage at any one point judged by comparison with the voltage at another point. This method has the disadvantage of being both crude and liable to error due to the presence of second or higher-order harmonics in the output from the transmitter. A little better is the small neon bulb, but this is also a voltage-operated device and subject to harmonic distortion. Then there is the current-squared galvanometer with a few small turns of wire connected to its terminals, these turns being coupled to either line. This device is quite good if one is careful to use constant coupling to the wire, but its disadvantages are that coupling between the meter and line is both inductive and capacitive, causing the meter to read differently according to how it is held up to the line, and that the meter will also be influenced by harmonics.

The most satisfactory device seems to be a sensitive galvanometer in a miniature tuned circuit coupled to the line. Such a device combines sensitivity with freedom from harmonic distortion, but the effects of capacity are still there to some extent. The tuned circuit should therefore be carried along the line with its position relative to the plane of the two wires maintained constant. See Figure 8. Checking along one of the two wires only is usually sufficient, but if one suspects an unbalance exists, the other line should also be checked. If the antenna system is reasonably symmetrical, unbalanced currents will most likely be due to improper coupling to the transmitter, but that is another story. At any rate, if a cur-

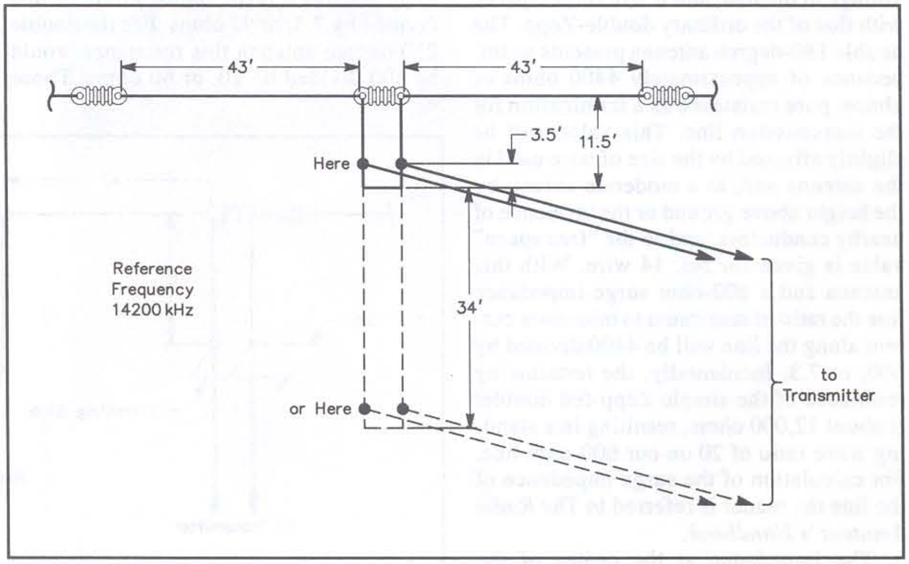

Figure 7—Center-fed double 230-degree Zepp with 600-ohm transmission line.

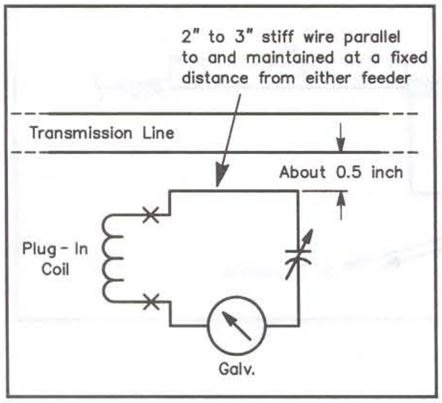
2" to 3" stiff wire parallel to and maintained at a fixed distance from either feeder
Transmission Line
About 0.5 inch
Plug - In Coil
Galv.

Figure 8—Portable line-current indicator.

rent maximum (voltage minimum) appears at the junction of transmission line and stub, the connection to the stub is too near the shorted end, and if the current is a minimum the opposite, of course, will be true. If the current maximum or minimum (assuming standing waves actually exist) does not occur near the stub, there has been a slip-up somewhere in the procedure, for this will be evidence that the stub itself did not tune the antenna to resonance.

PARASITIC ELEMENTS

The use of spreaders several feet in length to support each end of a "flat-top" antenna composed of two wires is again becoming popular, and for a good reason. Gains of around 5.5. db for a bi-directional array and around 7 db for a unidirectional array are thus made possible. One of the most promising arrangements is that of two double-230-degree antennas supported parallel to each other only 0.2 wavelength apart, as illustrated in Figure 9. Antenna A is excited by the transmitter using the same technique as when exciting a single double-230-degree antenna. Antenna B may be tuned by proper positioning of a short on its stub to become either a director or a reflector. In fact, this stub might be extended by multiples of a half wavelength to enter the operating room where, by means of a single switch the shorting position may be changed from that corresponding to a director to the position corresponding to a reflector, thus reversing the directivity of the antenna system. Information as to the gain of such an antenna system when one antenna is used as a director is not available, but when it is used as a reflector and adjusted to give a *minimum* signal to the rear, the signal forward will be 7 db better than a simple doublet, and the signal backwards will be 7 db less than from a doublet, resulting in a front-to-back ratio of 14 db. Adjustment of the reflector shorting bar should be on the basis of minimum backward signal from the antenna, or minimum signal received from a station in that direction when using the antenna system for reception, since this adjustment can be made more accurately than one resulting in maximum front signal. This will minimize QRM for both transmitting and receiving. The same principles hold true when adjusting the auxiliary antenna for use as a director.

If adjustments of the auxiliary antenna are to be made when transmitting and an ammeter is temporarily inserted in the shorting bar, minimum backwards signal will occur for the reflector when the short is moved slightly further from the antenna than that position corresponding to maximum current through the short. For the director, optimum conditions will be obtained when the short is slightly closer to the antenna than the maximum shorting current position. From this it can be seen that the proper adjustment of the auxiliary antenna is quite critical and certainly very important. Since field strength instruments are not generally available, a receiver fitted with a signal strength indicator of some sort located within a few miles of the antenna and in the desired direction should prove to be the next best thing. Connecting your own receiver to the antenna, with the other fellow transmitting, is perhaps even more practical in search for that minimum-signal adjustment.

The presence of the auxiliary antenna only 0.2 wavelength away from the driven antenna and adjusted properly for maximum forward or backward radiation will obviously affect the radiation resistance and impedance of the driven antenna. As a result, if the auxiliary antenna is left open-circuited and the stub on the driven antenna is adjusted in accordance with the method previously given, this adjustment will not be correct when the auxiliary antenna is, in turn, properly tuned. Moreover, the impedance of the driven antenna will differ slightly according to whether the auxiliary antenna is tuned as a director or as a reflector. If quick change from reflector to director is contemplated, some sort of compromise is indicated in the adjustment of the stub on the driven antenna. From a practical standpoint, however, the adjustment of the auxiliary antenna should be made first, be it reflector or director, with the driven antenna excited by the two-wire transmission line without a correcting stub. This may require temporary adjustment of the transmission-line length, either physically or by means of series coils or condensers, to bring a low-impedance point at the coupling coil to the transmitter in order to load the latter satisfactorily. In fact, if the line length from antenna to transmitter is not more than a wavelength or so, the reduction in line losses resulting from the use of a correcting stub is hardly worth the trouble of installing it. The point is, if a "flat" line is desired, adjustments of the stub at the driven antenna should not be made until after the auxiliary antenna has been tuned to give the desired radiation pattern.

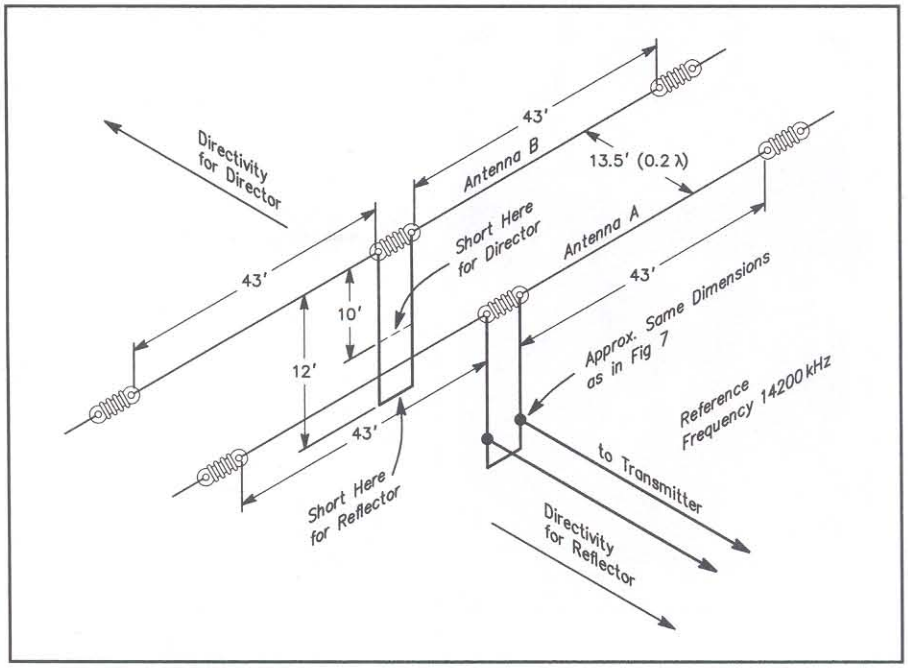

Figure 9—Double 230-degree Zepp with parasitic reflector-director.

LARGER COLINEAR ARRAYS

Now let us suppose our backyard is big enough to hang up more than two doublets end to end. An interesting possibility would be a four-element array of 230-degree elements. But here the principles of the extended double-Zepp must be carefully considered in designing the length of each element and the length of the phasing stub separating them. The actual arrangement will be as shown in Figure 10. Both degrees and dimensional designations are used to indicate the electrical and physical length of each element. The length of the stubs indicated is approximate only and must be adjusted for best results. The lengths of the various antenna elements as shown, however, may be assumed correct. The transmission line can be tapped on to any of the stubs, but connection to the center one will give the greatest gain.

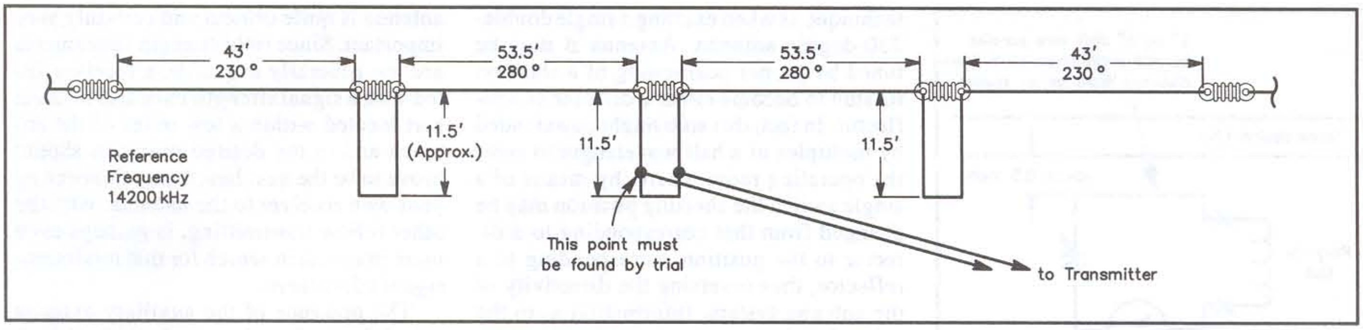

Figure 10—Four-element array with 230-degree elements.

The principle of the design shown for four elements is to provide the same separation in space between each doublet center as is provided between the two doublet centers in the double-230-degree antenna. Each phasing stub will, of course, be considerably shorter than a quarter wavelength, and when adjusting their length to tune the whole antenna to resonance, all three stubs must be made the same length; that is, if one is shortened two inches, the other two must also be shortened by two inches, assuming they were all the same length to start with.

This may prove to be an awkward and tedious method of adjustment, but the procedure may be simplified by adjusting the length of the center stub only as a first approximation. The use of an "exciter" antenna temporarily rigged nearby is assumed. If it is found the center stub must be shorter than the other two stubs by 3 feet, for instance, in order to obtain maximum current through its shorting bar, the other two stubs should be shortened by 1 foot and a new position found for maximum current through the center stub. This new position should correspond closely to equal length for all three stubs. This procedure of tuning the four-element array should be followed even though a stub at the center will not be used, finally, its place being taken by the transmission line itself. Such a connection will result in standing waves along the line and it will have to be "tuned" to permit easy coupling to the transmitter or receiver.

The horizontal pattern of the four-element array just described will have two major lobes at right angles to the antenna and several minor lobes of small amplitude. The gain in actual practice has proved to be greater than anticipated, and probably is more than 7 db. The major lobes will be much narrower than for the double-230-degree antenna, thus requiring more careful "aim" in erecting the array, or provisions for swinging it about. It is assumed this array, as well as the others described, is hung horizontally. It may be hung vertically, of course, but the possible restrictions due to sharp directivity in the vertical plane should then be considered. In general, the horizontally polarized antenna will prove most practical, chiefly because the ground reflection usually encountered with vertically polarized radiations is inferior, and because high masts are not often available.

The adjustment of the antennas described in this article may seem a bit involved as compared with the simple doublet or the double-Zepp antenna, but the extra gain thus made available should make their use worthwhile. For that matter, the double-230-degree antenna need not be tuned at all, its construction being made in accordance with the dimensions given and the transmission line tuned to fit the transmitter. In this respect the simplest of all directive antennas, size and gain also taken into consideration, is the extended double-Zepp.

From *QST*, February 1992 (Hints and Kinks)

Try an Extended Double-Zepp Antenna

Although the extended double Zepp (EDZ) antenna (Figure 1) has been in just every antenna handbook since the year one, hams seldom use it. Its overall length is 1.28 wavelengths (1.28 λ), and it's bidirectional broadside. Fed with open-wire line and a balanced antenna tuner, an EDZ also makes a fine multiband antenna. Let's look at an extended double Zepp for 17 meters. We can calculate the overall length of its two wire elements with the formula

$$\frac{984}{f\,(\text{MHz})} \times 1.28 = \text{length in feet} \qquad \text{(Eq 1)}$$

Using this formula, an 18.15-MHz EDZ works out to be 69.4 feet long. At this frequency, the EDZ exhibits, 3 dBd gain [*AB* page 8-34][1] in a figure-8 pattern with two major and four minor lobes. It still performs usefully, when operated on several bands lower in frequency than 17, however. At 20 meters, a 17-meter EDZ acts as two slightly long half waves in phase, exhibiting between 1.6 and 2 dBd gain [*AB* page 8-32]. At 40 meters, it's a slightly long ½-λ dipole [*AB* page 2-16,2-17,3-11 and 3-12]. All of these modes are directional broadside if the EDZ is positioned at least ½-λ high at 40 meters [*AB* pages 3-8 and 3-9]. At 15 meters, it exhibits a four-leaved-clover pattern, with minor lobes broadside; at 10 and 12 meters, it's close to two full waves in phase and produces a pattern similar to that at 15 meters. It can even be used as a short 75-meter dipole–not bad for a 70-foot piece of wire.

Scaled for 28.7 MHz–43 feet, 10 inches long—the extended double Zepp gives a 3-dBd-gain, figure-8 pattern at 10 meters and a similar pattern with a bit less gain at 12 meters. It acts as two halfwaves in phase at 15 meters, with about 1.6 dBd gain and a figure-8-pattern. At 17 and 20 meters, it's somewhat long for a half-wave dipole; a tuner can make it work at 40. On all of these bands, assuming that it's at least ½ λ high at its lowest band of operation, the 10-meter EDZ is directional broadside.

There's nothing magic about the extended double Zepp. It's a tried-and-true dipole that offers useful gain at its design frequency and good multiband performance. *–Bob Baird, W7CSD*

---

[1] This and the other AB references in this item refer to pages in G. Hall, ed, *The ARRL Antenna Book*, 16th ed (Newington: ARRL, 1991).

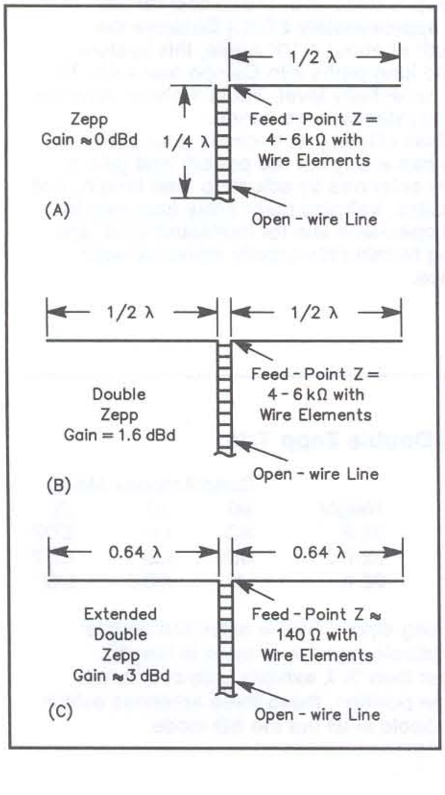

Figure 1—Evolution of the extended double Zepp antenna. At A, the classic Zepp—a ½-λ wire end-fed via an open-wire feeder operated with one of its two wires unterminated. (Antenna lore tells us that this feed method received its name from its use in zeppelin radio installations.) Commonly, Zepp systems use a feeder length of ¼ λ (or an odd multiple thereof) to transform the antenna's high feed-point impedance to a low impedance at the feeder terminals. Assuming efficient feed, this system exhibits 0 dBd gain—that is, *no gain* relative to a half-wave dipole cut for the same frequency and erected in the same position.

B shows a double Zepp, also known as *two half waves in phase* or (incorrectly) a center-fed Zepp. This is a form of collinear antenna because each ½-λ wire acts as an element, and because both elements lie along one line. It's also a *dipole*—a *full-wave* dipole that exhibits about 1.6 dB gain relative to a half-wave dipole in the same position. As with the Zepp, feeders cut to an odd multiple of ¼ λ make for least hassle feed with the double Zepp, but a good balanced antenna tuner should be able to handle whatever impedance appears at the feed-line input terminals.

A double Zepp becomes an *extended* double Zepp when you make its wires 0.64 λ long instead of 0.5 λ. Its gain increases, too—to about 3 dB relative to a half-wave dipole in the same position and cut for the same frequency.

The rough feed-point impedances shown here are for wire elements (fatter elements, such as those made of aluminum tubing, will lower these values somewhat) and convey only the resistive portion of the antennas' impedance. For more on the extended double Zepp antenna, see *The ARRL Antenna Book* and J. Reh, "An Extended Double Zepp Antenna for 17 Meters," *QST*, December 1987, pp 25-27.

From *QST*, March 1992 (Hints and Kinks)

Notes on Wire-Antenna Construction

Robert J. Zavrel, W7SX, seconds last month's extended-double-Zepp hint as he reports on how he keeps three EDZs high in trees for world-encompassing coverage (Figure 1) at 80, 40 and 20 meters—*Ed.*

What is a terminal DX hound to do when he suddenly moves to a rural area and has access to the stuff dreams are made of: several 100-foot-plus ponderosa pine trees atop a gentle rise in a reasonably clear area, only 500 feet from his house? The answer: *Wire antenna arrays!* A few months spent confronting the practical problems of installing three extended-double-Zepp antennas in these trees has yielded some solutions that I'd like to share.

Counterweights. Although my pines are wonderful supports for wire antennas, they have a few minor disadvantages. Among these is the fact that they sway in the wind. Because too much tension can snap antenna wires and supporting cords in strong winds, a wire-antenna engineer must compensate for variable tension in a horizontal wire antenna strung between two of these majestic plants. The textbook solution to this problem involves installing pulleys where the wires meet the trees. Ropes attached to the wire-end insulators pass over the pulleys and connect to counterweights hanging near the ground. Installed with antenna-to-counterweight lines of sufficient length, such a system also allows the antenna to be raised and lowered without further climbing.

My system includes three extended double Zepps, each fed with 450-Ω, low-loss TV twinlead. Including its supporting lines, each EDZ spans 165 feet. Counterweighting the ends of these antennas serves two purposes: It keeps tree movement from snapping them, and it keeps their wires reasonably horizontal. But how much weight is required, and what should the weights consist of? Charts exist for computing tension necessary to keep a singlewire span taut, but the presence of a twinlead feeder at the span's center renders such charts useless.

Enter experimentation! I strung the extended double Zepp, minus feed line and plus pulleys, between the two trees at about 15 feet above ground. Next, I weighed the feed line. Then I located an object of simi-

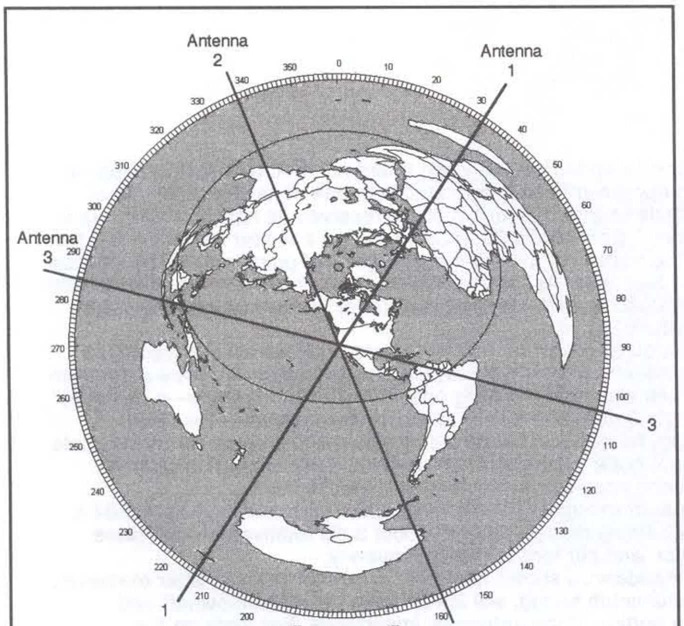

Figure 1—Proving the multiband effectiveness of the extended-double-Zepp antenna, "The W7SX Wire Farm" consists of three center-fed wires supported by three ponderosa pines that form a nearly equilateral triangle atop a small rise. (Total cost for the system, including all hardware, wire, feeders, control lines and $150 fee for the tree climber, approximately $700.) Because the ground slopes to the south at about a 10° angle, this system works best for VK, ZL and long paths into Europe and Asia. To the north, the horizon is essentially level. Table 1 shows how the antennas function on the system's three bands.

You're not a second-class citizen just because your best antenna is a dipole. You can easily set the pattern and gain of center-fed horizontal wire antennas by adjusting their length. Get the wires as high as possible, keeping them away from nearby objects, use a tuner and open-wire line for multiband feed, and have a blast! Surrounding terrain may greatly influence your antenna's DX performance.

Table 1
The W7SX Extended Double Zepp Trio

			Band/Antenna Mode		
	Length	*Height*	*80*	*40*	*20*
Antenna 1	82 ft	98 ft	SD	LD	EDZ
Antenna 2	82 ft	98 ft	SD	LD	EDZ
Antenna 3	164 ft	82 ft	LD	EDZ	LW

SD = short dipole; LD = long dipole; EDZ = extended double Zepp, LW = long wire (dipole legs 1 λ or more in length). Because a dipole longer than 1/2 λ exhibits gain over a half-wave dipole in the same position, these three antennas exhibit gain over a half-wave dipole in all but the SD mode.

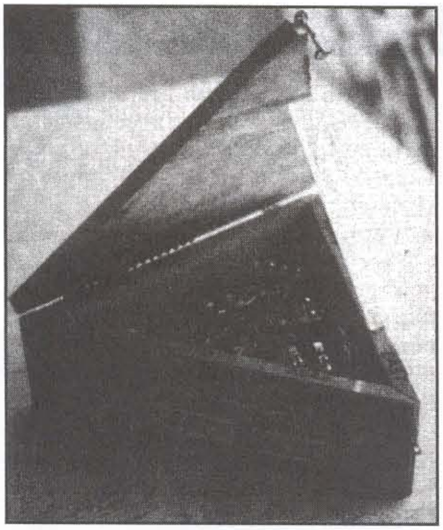

Figure 2—Bob Zavrel uses a triangular relay box to switch feeders between antennas mounted in a triangular pattern; a square box in this situation twists feeders, en-couraging antenna sway to break feeder wires. *(W7SX photo)*

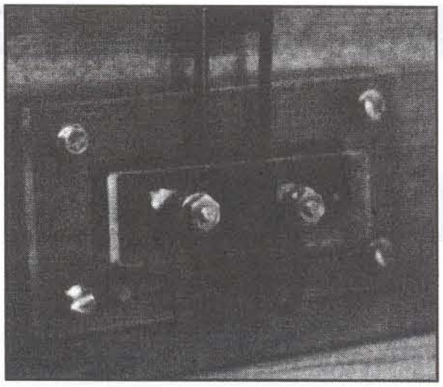

Figure 3—A combination feedthrough insulator/strain relief made of fasteners and two plexiglass plates.

lar weight, put it in a plastic bag, tied the bag at the antenna feed point, and experimented with various weights. Two 15-lb weights kept the antenna wire horizontal without overtension.

Next, I needed to construct multiple 15-lb weights at minimum cost—weights that had to be long and narrow to avoid collisions with the supporting trees and other nearby objects. I decided that slumpblock bricks inside a 2 × 4 wood frame would be satisfactory. (Slump blocks are long, narrow and about 12 lbs each; some are just as wide as a standard 2 × 4.) Including bricks, wood and nails, each weight cost me about $1.50, so the price was right—and 2 × 4s can be easily drilled for attaching ropes.

The relay box. A square relay box centered between three antennas twists the antennas' open-wire feeders at the feed points. I solved this problem for a triangular, relay-switched vertical array[1] and my three-EDZ system by building triangular relay boxes (Figure 2). (Use redwood, cedar or cypress and weatherproof the box with sealant.[2]) Plexiglass plates and screws insulate the feeder wires from the box, continue the feeders to the relays and provide strain relief (Fiureg 3).

Results with the system. Using a 600-W amplifier, I often get through pileups first on long path. Short path to Europe is more difficult because the ground slopes slightly upward in that direction. My blissful encounters with nature's gift to the RF enthusiast continue!–*W7SX*

---

[1] See Volume 3 of *The ARRL Antenna Compendium*.

[2] In my work with wire antennas, I discovered a reference on wood that may be of interest to amateurs: *Wood Handbook: Wood as an Engineering Material* (Washington, DC: US Department of Agriculture, rev 1987), 466 pages (for sale by the Superintendent of Documents, US Government Printing Office, Washington, DC 20402). This text describes, among many mechanical and structural considerations, electrical qualities of wood that come into play when wood supports are used in RF fields.

Collinear Arrays

Collinear arrays are always operated with the elements in phase. (If alternate elements in such an array are out of phase, the system simply becomes a harmonic type of antenna.) A collinear array is a broadside radiator, the direction of maximum radiation being at right angles to the line of the antenna.

Power Gain

Because of the nature of the mutual impedance between collinear elements, the feed point resistance is increased as shown earlier in this chapter (Figure 9). For this reason the power gain does not increase in direct proportion to the number of elements. The gain with two elements, as the spacing between them is varied, is shown by Figure 38. Although the gain is greatest when the end-to-end spacing is in the region of 0.4 to 0.6 λ, the use of spacings of this order is inconvenient constructionally and introduces problems in feeding the two elements. As a result, collinear elements are almost always operated with their ends quite close together–in wire antennas, usually with just a strain insulator between.

With very small spacing between the ends of adjacent elements the theoretical power gain of collinear arrays is approximately as follows:

2 collinear elements—1.6 dB
3 collinear elements—3.1 dB
4 collinear elements—4.2 dB

More than four elements are rarely used.

Directivity

The directivity of a collinear array, in a plane containing the axis of the array, increases with its length. Small secondary lobes appear in the pattern when more than two elements are used, but the amplitudes of these lobes are low enough so that they are not important. In a plane at right angles to the array the directive diagram is a circle, no matter what the number of elements. Collinear operation, therefore, affects only E-plane directivity, the plane containing the antenna. At right angles to the wire the pattern is the same as that of the individual 1/2-λ elements of which it is composed.

When a collinear array is mounted with the elements vertical, the antenna radiates equally well in all geographical directions. An array of such "stacked" collinear elements tends to confine the radiation to low vertical angles.

If a collinear array is mounted horizontally, the directive pattern in the vertical plane at right angles to the array is the same as the vertical pattern of a simple $^1/_2$-λ antenna at the same height (Chapter 2).

TWO-ELEMENT ARRAY

The simplest and most popular collinear array is one using two elements, as shown in Figure 39. This system is commonly known as "two half-waves in phase." The manner in which the desired current distribution is obtained is described in Chapter 26. The directive pattern in a plane containing the wire axis is shown in Figure 40.

Depending on the conductor size, height, and similar factors, the impedance at the feed point can be expected to be in the range from about 4 to 6 kΩ, for wire antennas. If the elements are made of tubing having a low λ/dia (wavelength to diameter) ratio, values as low as 1 kΩ are representative. The system can be fed through an open-wire tuned line with negligible loss for ordinary line lengths, or a matching section may be used if desired.

THREE- AND FOUR-ELEMENT ARRAYS

When more than two collinear elements are used it is necessary to connect "phasing" stubs between adjacent elements in order to bring the currents in all elements in phase. It will be recalled from Chapter 2 that in a long wire the direction of current flow reverses in each $^1/_2$-λ section. Consequently, collinear elements cannot simply

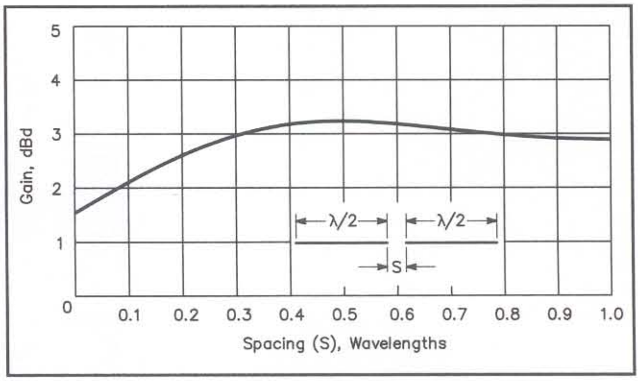

Figure 38—Gain of two collinear $^1/_2$-λ elements as a function of spacing between the adjacent ends.

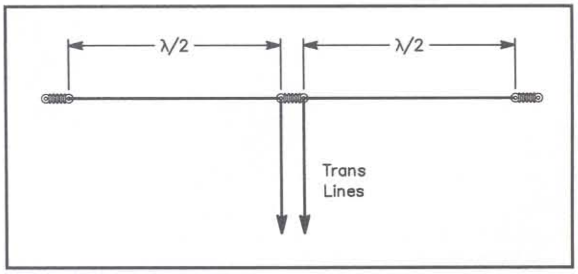

Figure 39—A two-element collinear array (two half-waves in phase). The transmission line shown would operate as a tuned line. A matching section can be substituted and a nonresonant line used if desired.

be connected end to end; there must be some means for making the current now in the same direction in all elements. In Figure 41A the direction of current flow is correct in the two left-hand elements because the transmission line is connected between them. The phasing stub between the second and third elements makes the instantaneous current direction correct in the third element. This stub may be looked upon simply as the alternate $^1/_2$-λ section of a long-wire antenna folded back on itself to cancel its radiation. In Figure 41A the part to the right of the transmission line has a total length of three half wavelengths, the center half wave being folded back to form a $^1/_4$ λ phase-reversing stub. No data are available on the impedance at the feed point in this arrangement, but various considerations indicate that it should be over 1 kΩ.

An alternative method of feeding three collinear elements is shown in Figure 41B. In this case power is applied at the center of the middle element and phase-reversing

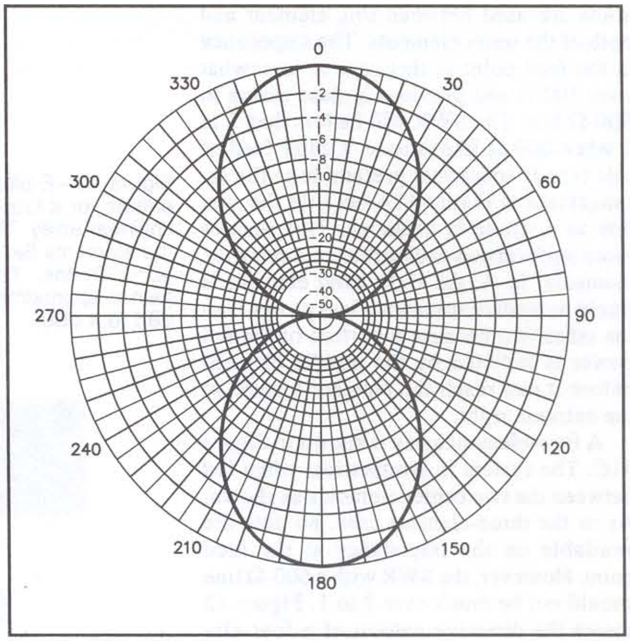

Figure 40—Free-space E-plane directive diagram for the two-element collinear array of Fig 39. The axis of the elements lies along the 90°-270° line. This is the horizontal pattern at low wave angles when the array is horizontal. The array gain is approximately 1.6 dBd (3.8 dBi).

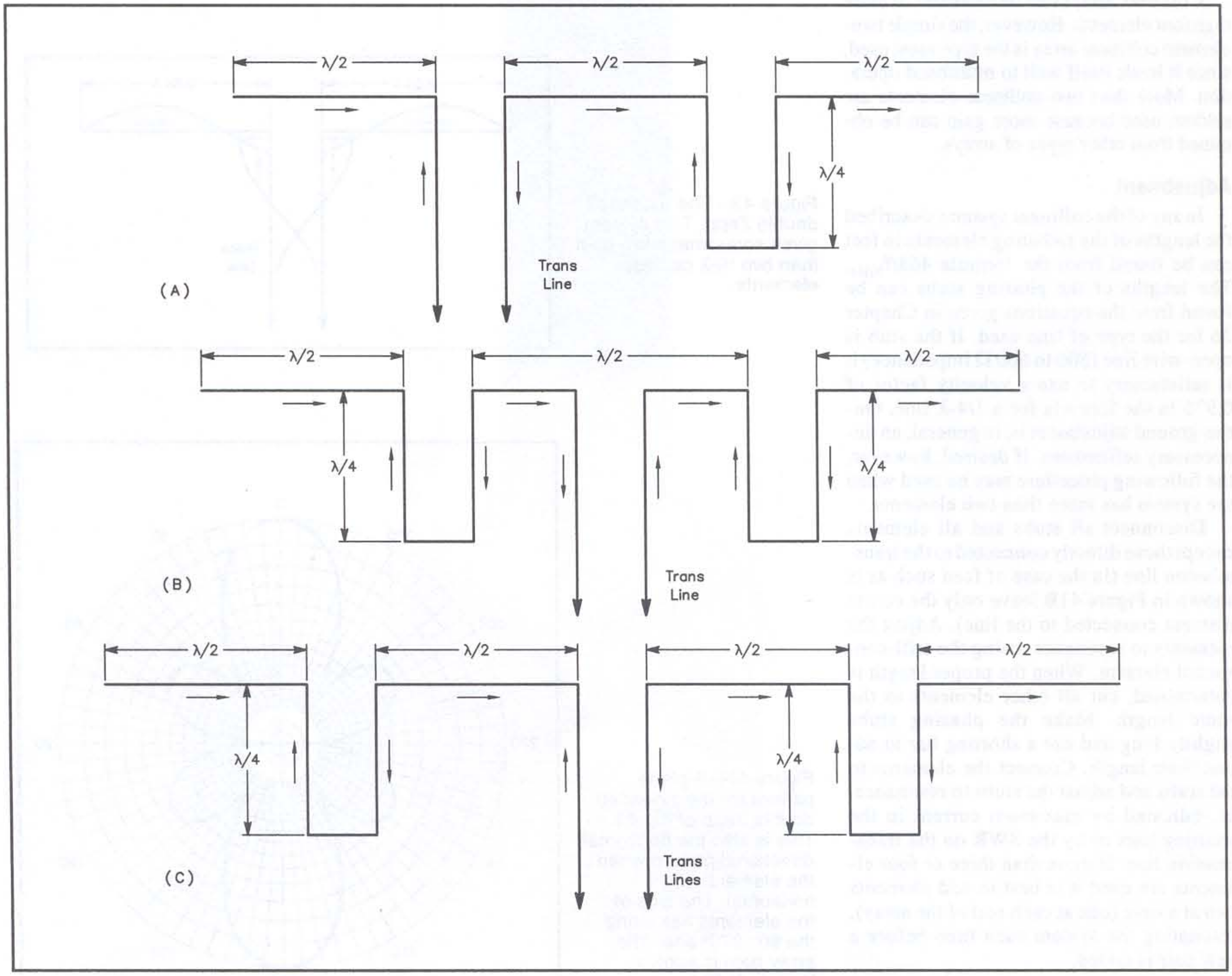

Figure 41—Three and four-element collinear arrays. Alternative methods of feeding a three-element array are shown at A and B. These drawings also show the current distribution on the antenna elements and phasing stubs. A matched transmission line can be substituted for the tuned line by using a suitable matching section.

stubs are used between this element and both of the outer elements. The impedance at the feed point in this case is somewhat over 300 Ω and provides a close match to 300-Ω line. The SWR will be less than 2 to 1 when 600-Ω line is used. Center feed of this type is somewhat preferable to the arrangement in Figure 41A because the system as a whole is balanced. This assures more uniform power distribution among the elements. In A, the right-hand element is likely to receive somewhat less power than the other two because a portion of the fed power is radiated by the middle element before it can reach the element located at the extreme right.

A four-element array is shown in Figure 41C. The system is symmetrical when fed between the two center elements as shown. As in the three-element case, no data are available on the impedance at the feed point. However, the SWR with a 600-Ω line should not be much over 2 to 1. Figure 42 shows the directive pattern of a four-element array. The sharpness of the three-element pattern is intermediate between Figures 40 and 42, with four small minor lobes at 30° off the array axis.

Collinear arrays can be extended to more than four elements. However, the simple two-element collinear array is the type most used, since it lends itself well to multiband operation. More than two collinear elements are seldom used because more gain can be obtained from other types of arrays.

Adjustment

In any of the collinear systems described the lengths of the radiating elements in feet can be found from the formula $468/f_{MHz}$. The lengths of the phasing stubs can be found from the equations given in Chapter 26 for the type of line used. If the stub is open-wire line (500 to 600 Ω impedance) it is satisfactory to use a velocity factor of 0.975 in the formula for a 1/4-λ line. On-the-ground adjustment is, in general, an unnecessary refinement. If desired, however, the following procedure may be used when the system has more than two elements.

Disconnect all stubs and all elements except those directly connected to the transmission line (in the case of feed such as is shown in Figure 41B leave only the center element connected to the line). Adjust the elements to resonance, using the still-connected element. When the proper length is determined, cut all other elements to the same length. Make the phasing stubs slightly long and use a shorting bar to adjust their length. Connect the elements to the stubs and adjust the stubs to resonance, as indicated by maximum current in the shorting bars or by the SWR on the transmission line. If more than three or four elements are used it is best to add elements two at a time (one at each end of the array), resonating the system each time before a new pair is added.

THE EXTENDED DOUBLE ZEPP

An expedient that may be adopted to

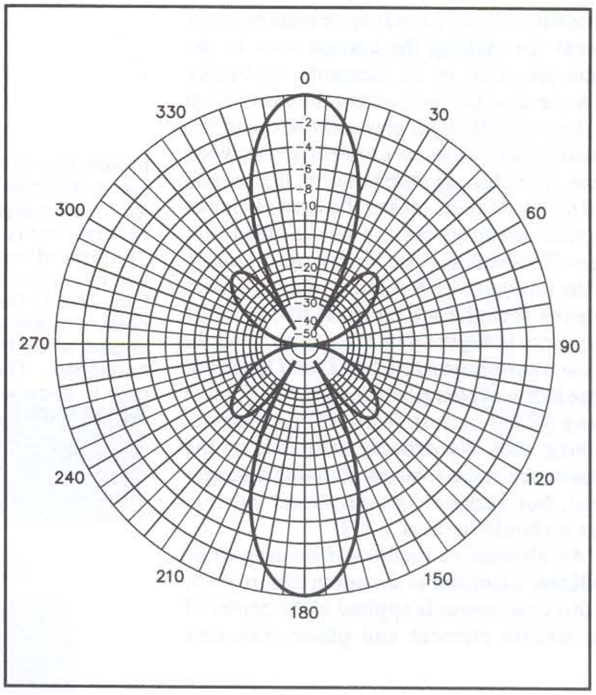

Figure 42—E-plane pattern for a four-element collinear array. The axis of the elements lies along the 90°-270° line. The array gain is approximately 4.2 dBd (6.4 dBi).

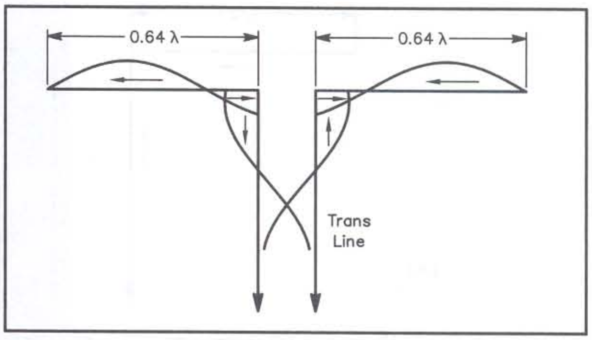

Figure 43—The extended double Zepp. This system gives somewhat more gain than two ½-λ collinear elements.

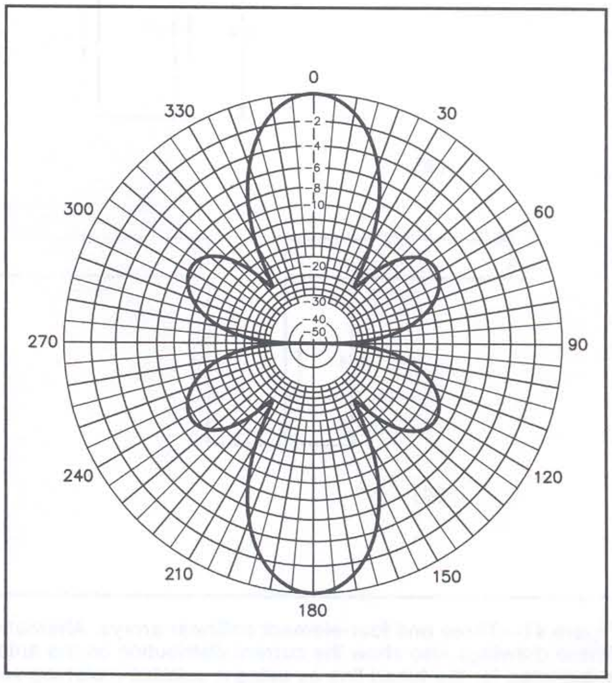

Figure 44—E-plane pattern for the extended double Zepp of Fig 43. This is also the horizontal directional pattern when the elements are horizontal. The axis of the elements lies along the 90°-270° line. The array gain is approximately 3 dBd.

obtain the higher gain that goes with wider spacing in a simple system of two collinear elements is to make the elements somewhat longer than $^1/_2$ λ. As shown in Figure 43, this increases the spacing between the two in-phase $^1/_2$-λ sections at the ends of the wires The section in the center carries a current of opposite phase but if this section is short the current will be small; it represents only the outer ends of a $^1/_2$-λ antenna section. Because of the small current and short length, the radiation from the center is small. The optimum length for each element is 0.64 λ. At greater lengths the system tends to act as a long-wire antenna, and the gain decreases.

This system is known as the "extended double Zepp." The gain over a $^1/_2$-λ dipole is approximately 3 dB, as compared with approximately 1.6 dB for two collinear $^1/_2$-λ dipoles. The directional pattern in the plane containing the axis of the antenna is shown in Figure 44. As in the case of all other collinear arrays, the free-space pattern in the plane at right angles to the antenna elements is the same as that of a $^1/_2$-λ antenna—circular.

CHAPTER SIX

WIRE BEAMS

6-1 The K6NA 80-Meter Wire Beam
From *The ARRL Antenna Compendium Volume 5*

6-10 Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters
From August 1986 *QST*

6-14 Input Impedance of LPDA Antennas
From October 1986 *QST* (Technical Correspondence)

6-15 The WA1AKR 40- and 75-Meter Slopers
From August 1980 *QST* (Technical Correspondence)

6-16 Inexpensive 30-Meter Beam Antenna
From June 1983 *QST* (Hints & Kinks)

6-17 Try the "FD Special" Antenna
From June 1984 *QST*

6-21 Build a 4X Array for 160 Meters
From February 1985 *QST*

6-25 The 'BRD Zapper: A Quick, Cheap and Easy "ZL Special" Antenna
From June 1990 *QST*

6-27 Combination Driven Arrays
From *The ARRL Antenna Book* 18th edition

By Glenn Rattmann, K6NA **From *The ARRL Antenna Compendium Volume 5***

The K6NA 80-Meter Wire Beam

Ever get crunched in a DX pileup on 80/75 meters by K6NA? Lots of folks have! K6NA reveals the secrets behind his two-element "killer" Yagi.

This article describes a practical design for a two-element, horizontal wire Yagi for the 75/80-meter band using just two support towers. Designed with DXing and contesting in mind, this radiator-and-reflector combination features moderate gain and good front-to-back ratio, combined with instant beam reversal and band-segment (mode) switching. It uses tuned feeders and a balanced-output antenna tuner at the operating position for matching at the transmitter. Remote relays at ground level handle reversal and parasitic-tuning requirements, taking advantage of the convenient properties of half-wave transmission lines.

I designed this antenna empirically, before PC-modeling programs were readily available. I've confirmed that it really works, through years of successful 80-meter operation. Later, I verified this with computer modeling. Figure 1 shows the general layout of my system.

Background

At my former location I had two towers, 70 and 100-feet high, fairly close together on a half-acre of property. There was an assortment of quads and Yagis for 10-40 meters,[1] but as is often the case with all-band DXers, the 80-meter antenna was an afterthought. Through the years I experimented with many types, looking for the "silver bullet" antenna for 80 that would finally make this challenging band a pleasure to operate.

Each season numerous magazine articles and endless discussions by experts extolled the virtues of this or that magic 80-meter DX antenna. I even tried some of them: a diamond quad loop, a delta loop (bottom-fed or corner-fed), a pair of parasitic deltas, phased quarter-wave slopers (tilted ground planes), a $\lambda/4$ vertical with lots of radials. I always came back to a basic truth: with the tower heights available, and faced with the reality of rough, rocky, lossy earth, my 95 foot high inverted-V dipole was usually as good as–or better than—any of the "trick" antennas I had spent hundreds of hours building. As a bonus, usually the dipole was quieter on receive.

Even Jim Lawson, W2PV, in an early *QST* article about broadbanding an 80-meter antenna, mentioned almost in passing that his high (110-foot) inverted-V was superior to his four-element phased vertical array, at least on transmit.[2] What I observed about 80-meter DXing in the 1970s was that most attempts using vertically polarized systems to achieve low-angle radiation were not satisfactory. This was due to excessive ground losses (both near-field return-current losses and far-field reflective losses); in addition, verticals were generally noisy on receive.

Neither vertically polarized antennas nor relatively low horizontal antennas were really getting thejob done. I resolved to construct my next station thinking about the 80-meter band from the start."High and horizontal" would be the goal.

Design Criteria

My new station would have two main towers, each 140 feet high. Because I had all-band contesting in mind, there would be numerous other antennas on the towers. I wanted to put a three-element 40-meter beam on top of one tower and a large 20-meter Yagi on the top of the other. Though it likely would be an excellent performer, a rotating beam of some type for 80 meters was ruled out because I did not want to dedicate the top of one tower to it. Also, there was the near-certainty of difficult maintenance problems with such a gigantic antenna. Instead, I settled on a design for a fixed-wire beam. These would be the main criteria for reliable, day-to-day operation:

- Oriented properly for best results in both DX and domestic contests
- Moderate gain, with a wide lobe for good azimuth coverage
- Instant beam reversal with a decent front to-back ratio
- Near-instant band-segment (phone or CW mode) switching, especially for casual, daily DXing
- Easy, low-loss matching anywhere in the band
- Horizontally polarized to minimize both ground-reflection losses and noise on receive
- High enough to produce reasonably low angle radiation
- Reliable and easy to maintain without disrupting other antennas on the towers.

Design Discussion

Some discussion of these interrelated design criteria will aid in understanding the trade-offs and choices I made while planning this antenna. For example, from San Diego the typical short path to Europe on 80 meters *is not* the expected true Great-Circle path of about 25°. Rather, the short path toward Europe is almost always a "bent" or skewed path, where signals pass to the south or southeast of the highly absorptive auroral oval, propagating by means of a scatter mode that uses ionized patches over the central Atlantic Ocean. Consequently, our pseudo short-path heading toward Europe is commonly 50° to 90° in azimuth. Occasionally, European signals on 80 meters arrive in California from the southeasterly direction. Rarely, around the time of the Equinox, they may even arrive by scatter path from straight south, as reported by some rotary beam users in California.

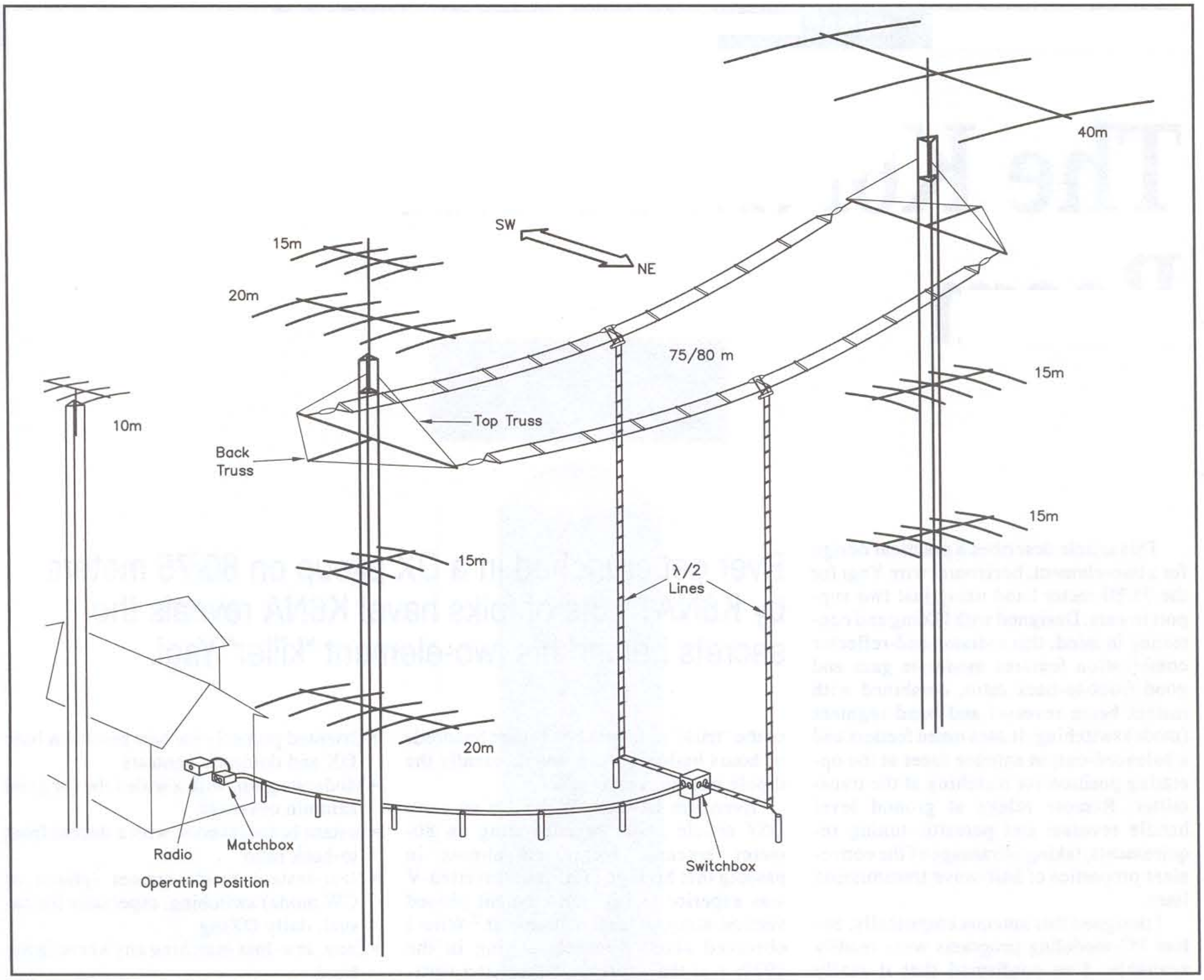

Figure 1—The antenna farm at K6NA, showing the 80-meter wire beam installed on the pair of 140-foot towers, using 35-foot long crossarms made of aluminum tubing and fiberglass. The rearward truss and upper truss wires make each crossarm rigid. For clarity, the tower guy wires are not shown.

I decided to orient the two support towers so that the reversible wire beam, slung between the towers on a pair of horizontal crossarms, would point at 60°/240°. This would provide good coverage across the USA and into Europe (both short and long path), Africa, the Caribbean and South Pacific areas, which together make up about 85% of my annual QSOs on 80 meters.

I could build a simple two-element parasitic design, using $\lambda/2$ elements on a reasonable boomlength of about 35 feet, or $\lambda/8$.[3] Figure 2B shows that such a design has a half-power azimuth beamwidth of about 74° in the azimuth plane. I was willing to accept the slightly lower gain and front-to-back ratio compared to a three-element design (which would be more difficult mechanically), or a two-wire phased array using double-extended Zepp elements. The Zepp would require the towers to be placed about 350 feet apart. Either of these antennas conceivably would provide more gain, but at the price of reduced azimuthal coverage. A reversible, all-driven array has the additional disadvantage that its phasing and matching networks become overly complicated to make it work properly at both ends of the band.

To achieve the twin requirements of easy mode-switching and beam reversal, I decided to construct a symmetrical, two-element Yagi beam. The elements are identical dipoles, self-resonant near 3.675 MHz. This frequency is about 3% to 4% lower than the phone operating area around 3.800 MHz, meaning that each element could be used as a reflector in the phone band with no additional loading. Each dipole is fed with $\lambda/2$ of open-wire line. These lines come together in a relay box at ground level. Power is applied to one or the other driven element, and a short is applied across the base of the remaining feeder to force the remaining element to act as a parasitic reflector.

To operate near 3.5 MHz, the reflector simply is retuned to about 3.4 MHz, using an inductor in place of the direct short in the relay box. Supplying RF to the relay box is accomplished using low-loss, open-wire feeders. A balanced-output tuner in the shack is simply touched up to move between phone and CW subbands, and to accommodate the change in SWR. There is no requirement for an exact match at the line input to the driven element because of the low-loss transmission lines.

With the crossarm supports mounted at 130 ft, the dipoles are about $\lambda/2$ high and thus produce fairly low-angle radiation. A horizontally polarized antenna is less responsive to man-made noise sources, which have a large vertical component. The height allows $\lambda/2$ lines to reach down to the switchbox near ground level. All switching relays are in this weatherproof box, so the system is reliable and easy to maintain. No inaccessible relays or matching networks are up at the elements, where failures might occur.

Construction

No doubt, the 80-meter wire beam can be built and integrated into many existing

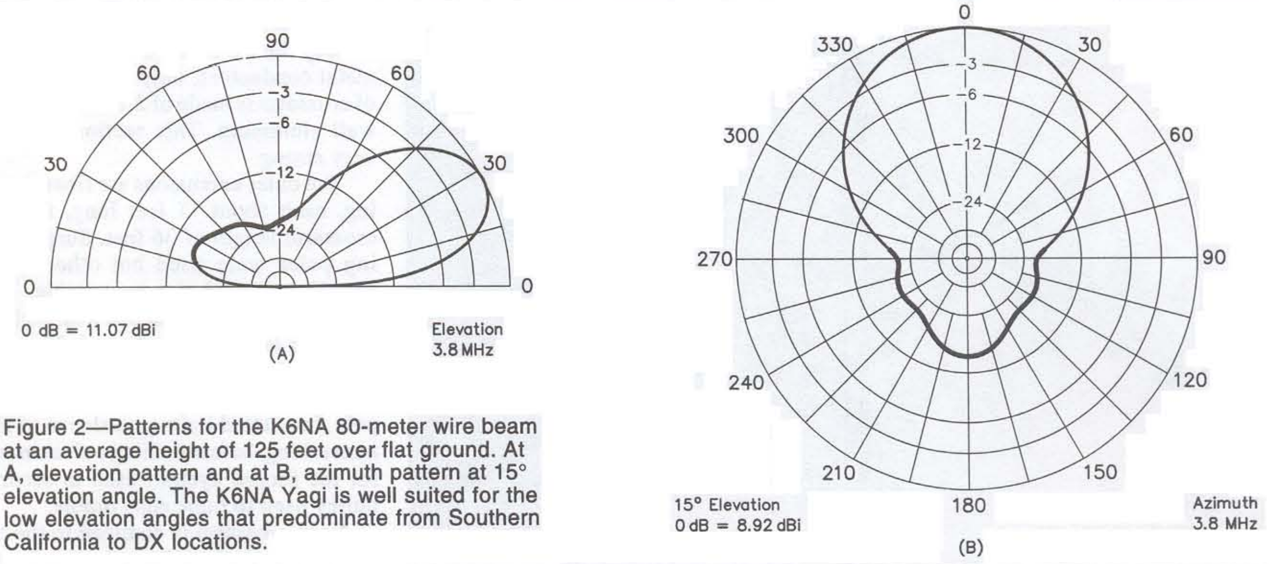

Figure 2—Patterns for the K6NA 80-meter wire beam at an average height of 125 feet over flat ground. At A, elevation pattern and at B, azimuth pattern at 15° elevation angle. The K6NA Yagi is well suited for the low elevation angles that predominate from Southern California to DX locations.

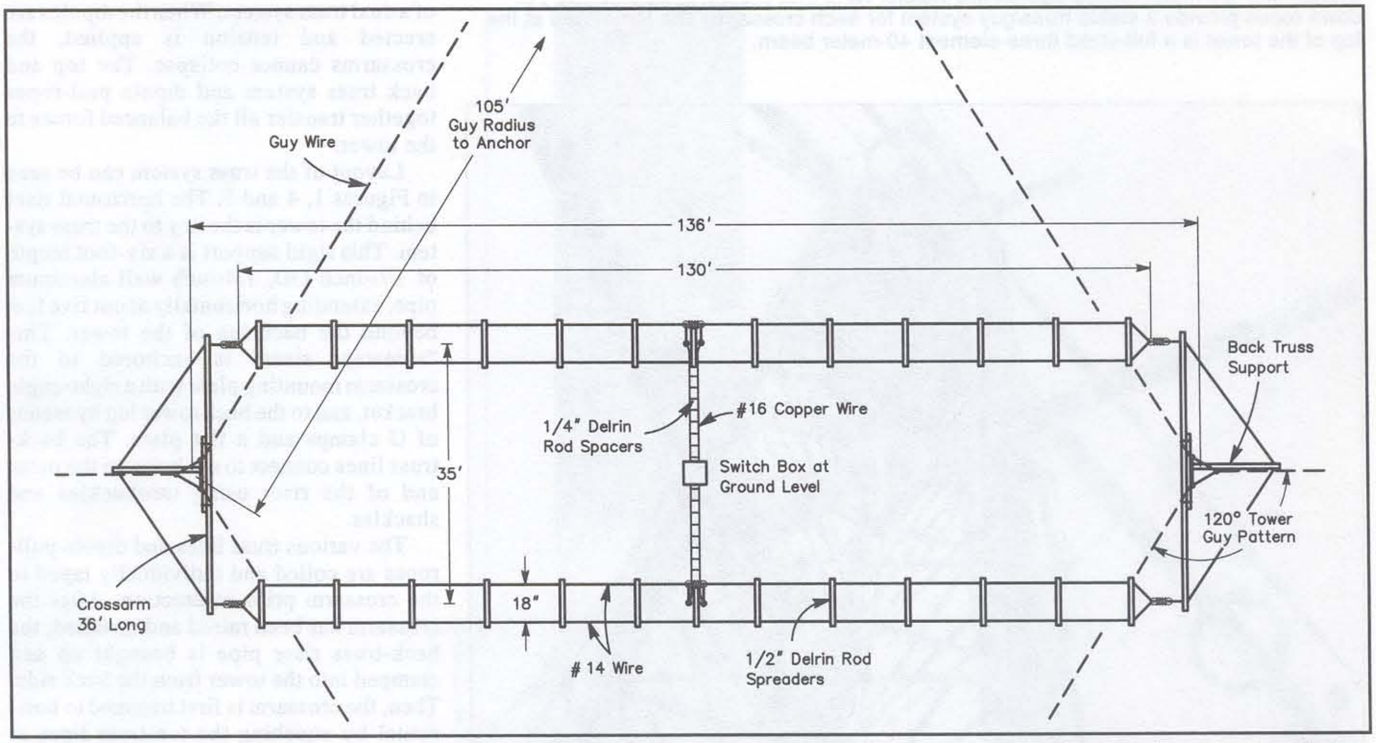

Figure 3—Top view of the 80-meter wire beam mounted on the towers, with details of the crossarms and rearward truss system that prevents the tensioned antenna wires from bending the crossarms inward. The tower-guy placement geometry provides several advantages for antenna-farm management! See text.

multi-tower systems. I had the luxury of planning the tower and guy arrangement well in advance of building the antenna. This helped make erection and maintenance simpler. Such planning contributes to convenient erection of other antennas too!

Figure 3 shows the recommended two-tower layout that has served very well for this and other projects. Through symmetry of guy-anchor placement, any forces due to loading by wire antennas or temporary trams are balanced out. Note that each tower has an opposing, backward-facing guy exactly in line with its counterpart and parallel with the centerline of the towers. When the 36-foot crossarm is mounted about 10 feet below the top of the tower and the topmost guy bracket, the ends extend out beyond and above the sloping, upper guys. This arrangement keeps the dipoles in the clear, allowing them to be raised and lowered using ropes and pulleys with minimal guy-wire interference, without having to remove or move the crossarms.

This tower/guy layout also presents two parallel tower faces on which to mount the booms and to correctly orient the wire antenna. Non-parallel tower faces can work, but the mounting brackets would be more complex and some advantages of symmetry would be lost. For safety, significant sideloads should be distributed across *at least two* tower legs. In addition, guys passing through the central area between the towers would complicate the raising and lowering of the wire antennas, so this should be avoided if possible.

Another advantage of the layout in Figure 3 concerns the overall use and maintenance of other antennas in the system. Note that the 140-foot towers are 136 feet apart, and the guy radius (distance from tower to each of its guy anchors) is 105 feet. These facts, together with the generally uncluttered work area between the towers, make

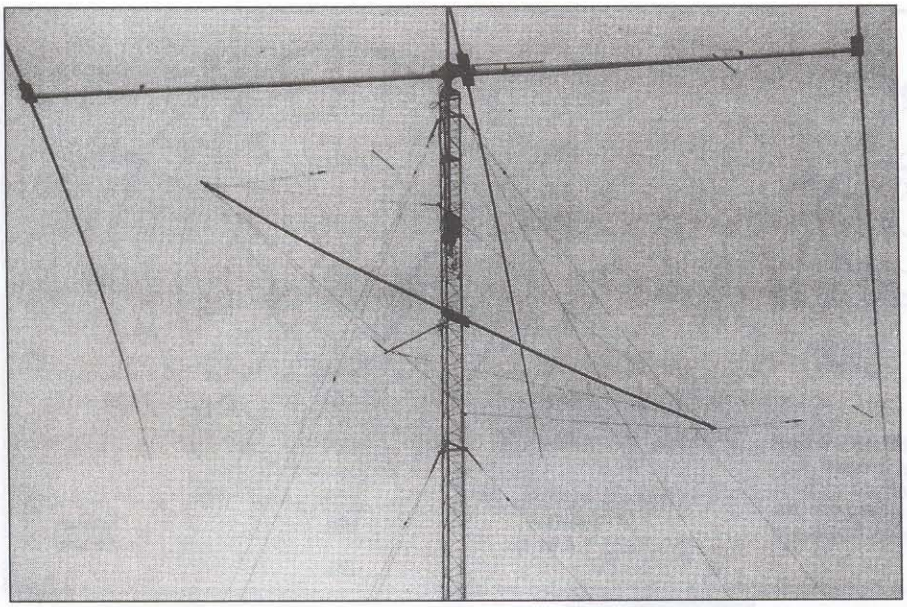

Figure 4—Photo of one of the crossarms (booms) mounted on the tower. The truss system and 80-meter folded dipoles are visible. Note that the truss, dipoles and pull-down ropes provide a stable truss/guy system for each crossarm. The large Yagi at the top of the tower is a full-sized three-element 40-meter beam.

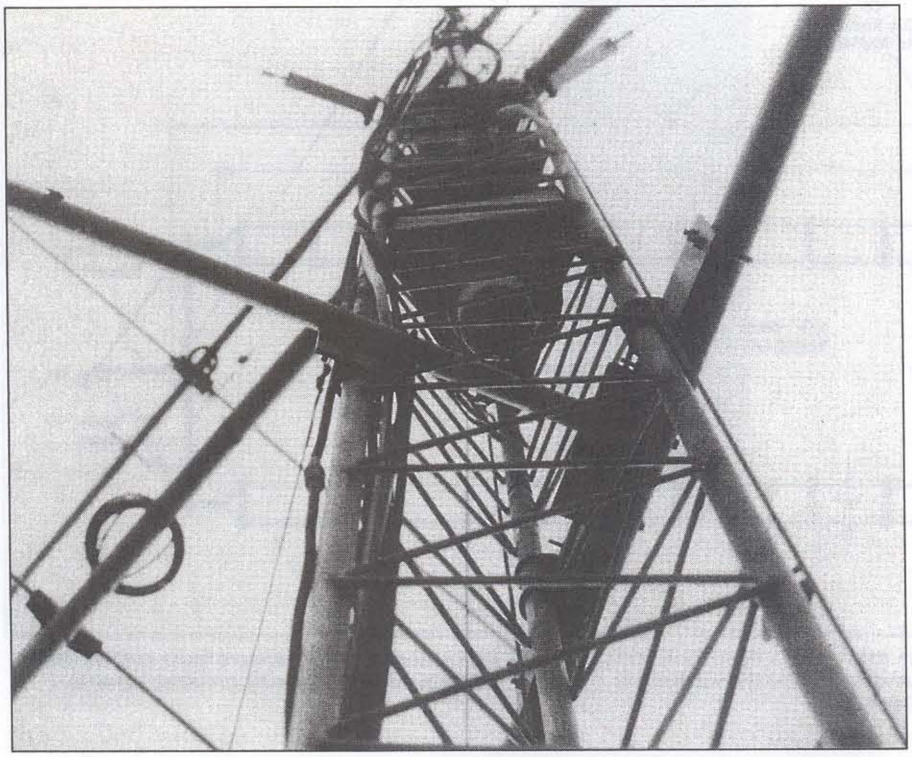

Figure 5—Photo showing mounting method for installing the 80-meter crossarm and associated horizontal back-truss pipe. The horizontal riser prevents the crossarm from collapsing when the side load (antenna tension) is applied. Note how a plate with U-bolts is used to secure the back-truss pipe to one of the tower legs, while the other end is secured with an L-shaped bracket and U-bolts to the mounting plate for the crossarm.

it convenient to run a temporary tram wire anchored at the base of one tower to the top (or anywhere along the length) of the other tower. The tram can be used to raise the 80-meter crossarms themselves, or to move other HF Yagis. If an upper-mast-mounted beam is to be erected, the 80-meter wires first can be lowered easily to the ground. If the tram has to reach only to the area of a side-mounted beam, the wire antenna may stay in place above it. I can detach the 80-meter feeders from the feed-through insulators at the switchbox, and pull them away from the vicinity of the tram wire, if necessary.

Crossarms and Trusses

I made each crossarm (boom) using aluminum and fiberglass. In order to eliminate the possibility of "boom resonance" that might interfere with nearby HF Yagis and to keep the 80-meter wire ends away from metal conductors, only the middle 14 feet of crossarm is made of 2-inch OD, 1/4-inch wall aluminum. This section is stiff and very strong.

The outer extensions are fiberglass tubing, each about 11 feet long, for a total crossarm length of 36 feet. Surplus vaulting poles were used but other types of fiberglass would also be suitable. The fiberglass should be wrapped with high-quality black tape (such as Scotch 33 or 88) to prevent decomposition due to ultraviolet radiation. At the end of each fiberglass pole, I mounted a six-inch sleeve of aluminum tubing, before drilling a clearance hole for the forged eyebolt used to secure the pulley used to hoist each dipole.

Each crossarm is clamped to the tower face using an aluminum plate and muffler clamps. See Figures 4 and Fig 5. Once in place, the crossarm is made rigid by means of a dual truss system. When the dipoles are erected and tension is applied, the crossarms cannot collapse. The top and back truss system and dipole pull-ropes together transfer all the balanced forces to the tower.

Layout of the truss system can be seen in Figures 1, 4 and 5. The horizontal riser behind the tower is the key to the truss system. This rigid support is a six-foot length of $1^{1}/_{2}$-inch OD, 1/4-inch wall aluminum pipe, extending horizontally about five feet beyond the back leg of the tower. This "sideways riser" is anchored to the crossarm mounting plate with a right-angle bracket, and to the back tower leg by means of U clamps and a flat plate. The back-truss lines connect to eyebolts on the outer end of the riser using turnbuckles and shackles.

The various truss lines and dipole pull-ropes are coiled and individually taped to the crossarm prior to erection. After the crossarm has been raised and installed, the back-truss riser pipe is brought up and clamped into the tower from the back side. Then, the crossarm is first trimmed to horizontal by attaching the top-truss lines to the tower legs and adjusting the turnbuckles until the crossarm looks good, much like a Yagi boom truss is adjusted. Then the backtruss lines are untaped from the crossarm, one at a time, and connected to the far end of the horizontal back-truss pipe. This is an interesting maneuver, which should be accomplished only by an experienced climber/rigger! The turnbuckles are adjusted initially for moderate tension. Properly adjusted, the result is a crossarm that doesn't sag downward or fold inward when antenna tension is applied.

After the truss system has been rigged, the installer untapes one of the coiled-up dipole pull-ropes, which has been prerigged to extend along the arm to the outer crossarm pulley and back again to the vicinity of the tower. By attaching either a weight or tag line to the thimble, the dipole-end of the rope is released to hang vertically from the pulley. The rigger then pays out the coiled rope until a ground man can reach the thimble end to attach the dipole insulator. The pull-ropes ultimately are brought to the tower legs through pulleys and thence brought down to ground level. The angle of the pull-ropes completes what is essentially a four-point guy on each end of the crossarm (antenna tension, back-truss, up-truss, and angled pull-down rope). This configuration presents a stable structure when the dipoles are raised and tension applied.

I must emphasize that all subassemblies should first be fabricated and fitted to the tower at ground level. Make sure that all plates, clamps and hardware mate to the tower, referenced to diagonal braces, welds, or leg bolts. Premark the tower legs up at the target location ahead of time to preclude surprises on the tower!

Construction of the Dipoles

Two-wire folded dipoles were selected to help bring the feedpoint impedance up closer to the characteristic impedance of the feeders and to improve bandwidth. This helps minimize tuner adjustments when moving within a subband, but there is no need to go overboard trying to effect a perfect match at the load.

Each folded dipole is built with #14 stranded, non-stretch copper wire. The two parallel wires are held apart by seven horizontal spreaders located about every 20 feet along the wire length. These spreaders of black Delrin are 1/2-inch diameter rod, 18 inches long (See Figure 6). They are held in position by tie wires, as are the ribbed end insulators. I used an unbroken length (about 260 feet) of wire for each two-wire dipole, making the free ends come together at the center-feed insulator. The free ends of the wires and feeders are tied to the center insulator for strain-relief, and then the loose ends are soldered. This technique prevents breakage, because soldered connections (which are weaker due to wire heating) are not under strain, or flexing. After each dipole is assembled on flat ground, it should be stretched temporarily at low level between the towers in order to check the symmetry and mechanical behavior. When adjustments are complete, the feeder can be soldered to the antenna.

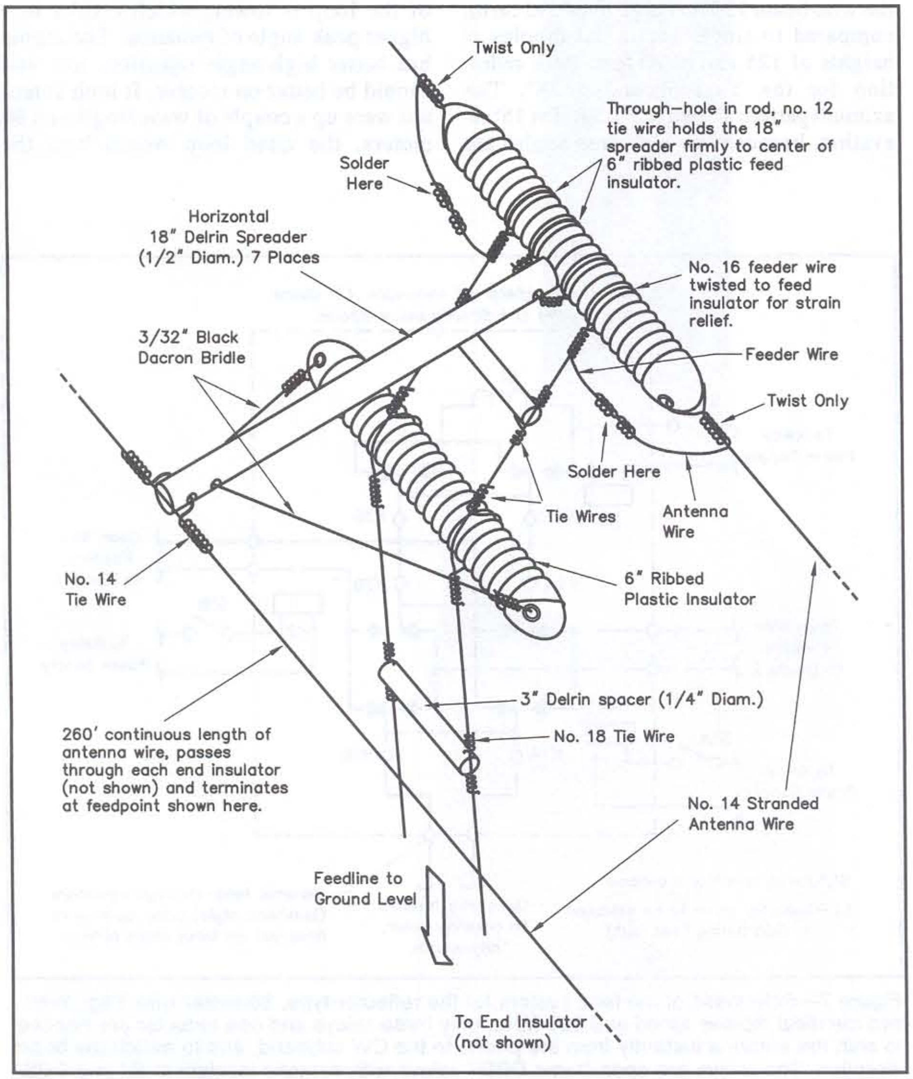
Twist Only
Through–hole in rod, no. 12 tie wire holds the 18" spreader firmly to center of 6" ribbed plastic feed insulator.
Solder Here
Horizontal 18" Delrin Spreader (1/2" Diam.) 7 Places
No. 16 feeder wire twisted to feed insulator for strain relief.
3/32" Black Dacron Bridle
Feeder Wire
Twist Only
Solder Here
Tie Wires
Antenna Wire
No. 14 Tie Wire
6" Ribbed Plastic Insulator
3" Delrin spacer (1/4" Diam.)
No. 18 Tie Wire
260' continuous length of antenna wire, passes through each end insulator (not shown) and terminates at feedpoint shown here.
No. 14 Stranded Antenna Wire
Feedline to Ground Level
To End Insulator (not shown)

Figure 6—Folded-dipole center feedpoint with stabilizing cradle. The cradle encourages stability in the large, two-wire dipoles, preventing twist. See text.

There is little point in making any sort of analyzer measurements at five feet off the ground because the results will be misleading. When the dipole is installed in the clear at a half-wavelength in height, it will perform as predicted. However, just to be sure, I grid-dipped one dipole at its final height through the half-wave line. I had first built the $\lambda/2$ transmission line and had checked for resonance near 3.650 MHz with a grid-dip meter. When the first dipole and its feeder are on frequency, the second dipole and feeder was cloned and the array was erected. The dipoles at K6NA are about 130 feet long.

Feed Line Construction

Each feed line was constructed using old techniques but modern materials. Minimum weight is very important. The lines were made about 130 feet long, using #16, prestretched solid copper wire. Spacers are 1/4-inch Delrin rods, three inches long and spaced every four feet along the feeder. Spacers were notched and drilled, and held in position using #18 copper tie wires through the holes. The result is a very lightweight, low-loss feeder with a characteristic impedance of about 575 Ω.[4] To minimize sag in the dipole, avoid using heavy ceramic spacers.

A good technique for constructing the long, open-wire feeder is to stretch two wires tightly between the towers about four feet above the ground. Space the wires three inches apart in a plane parallel to the earth. Precut all the tie wires and make a simple, four-foot long, interval-measuring rod from a piece of wood or tubing. With all the spacers, tie wires, rod and long-nose pliers in a tool bag, the installer begins at one end, moving easily along the line, installing spacers at the correct intervals. The feeder wires sit in the notched ends, while the tie wires pass through the holes and are twisted in place over the feeder wires. When complete, the suspended line is in a good position for checking resonance with a grid-dip meter, assuming the line is insulated from the towers. The final length of each feeder described here is about 128 feet and each weighs less than three pounds.

Some additional comments about feedpoint construction will be helpful to those planning assembly of such a large, wide-spaced folded dipole and associated feeder. Initially, the dipole was built with upper and lower wires, as depicted in most antenna books. An 18-inch Delrin rod connected the lower feed insulator to the upper wire in order to distribute the weight of the feeder onto both wires and to maintain correct spacing. This worked, but with unforeseen side effects. When the dipole was raised under tension, the result was a half-twist in the parallel wires about halfway back toward the end insulator. This twist was not due to "live" or "kinky" wire! The twist was due to a catenary sort of effect.

The solution to the twist problem is to avoid the upper wire/lower wire configuration, and allow all the spreaders (including the center spreader) to lay over into the horizontal plane. To encourage this, a simple cradle was constructed to distribute the weight of the feeder at the center of the two antenna wires. This forces the wires (and all the other balanced spacers) to form a plane parallel to the earth. Fig 6 shows the centering cradle below the Delrin rod. The feeder passes over the cradle insulator (tie wires take the load) and continues up to the feedpoint. Small Dacron lines tie the cradle insulator to the other end of the rod, keeping the load centered. These large folded dipoles have been well behaved, even in high winds.

Remote Switching

A plywood "doghouse" was constructed under the antenna center, and mounted about four feet above ground for convenience. Fig 7 is a schematic of the switching network housed in the box. Two DPDT relays control the direction of the beam pattern. The main feeder connects through K1 and K2 to one dipole, while the other dipole becomes a phone-band reflector due to the short across K3. In the shack, switch S1 activates K1/K2 for reversal. S2 is a phone/CW selector switch that allows K3 to retune the reflector element to about 3.4 MHz by means of an inductor. Relays are large, ceramic-base open-frame types.

The center-loading inductor for a CW reflector is best chosen experimentally. A coil is connected across K3, S2 is set for CW, and the tuner in the shack is adjusted for a match low in the CW band. From the tuner, a coax cable is routed temporarily back to the switchbox. At the box, the temporary coax feeds a receiver, which is used to monitor signals arriving off the back of the antenna around 3.515 MHz. A variable tap is moved along the coil until the best front-to-back ratio is obtained. After a period of testing, the tap can be soldered. With the dipoles self-resonant around 3.650 MHz, a CW reflector coil of about 17 µH was required. This consisted of a 2-inch OD coil, $2^3/_4$-inches long, on which were wound about 25 turns of #16 wire.

Operation and Maintenance

The 80-meter wire beam performs as expected. Typical F/B is 11 to 14 dB on both phone and CW modes. For high-band operators accustomed to quiet bands with low atmospherics, this level of F/B may not seem like much. However, on a consistently noisy band like 80 meters, this kind of rejection on receive is very important. The high wire beam makes it possible to work long-path Europeans (southwest heading, before sunrise in San Diego) on a regular basis in the winter months, due to effective rejection of storm noise over North America. The antenna performance is roughly comparable to that of a nearby, two-element rotary beam (195 feet high) at N6ND—we hear and work many of the same weak stations during DX sessions. In domestic contests, reports from the East Coast are excellent on both modes.

The wire beam has been very reliable for over 10 years. The first version of the antenna used # 16 solid-copper wire for the elements, and there was a broken wire or two after perhaps six years. The antenna was rebuilt in 1992 with #14 stranded wire and has been perfect since then. The original dipole pull-ropes of black, 1/4-inch polypropylene lasted about seven years before they were replaced. Black Dacron, readily available from several *QST* advertisers, would be an excellent replacement for the pull-ropes, but it is more expensive than polypropylene. Do not use any color except black, and do not use polyethylene lines for permanent antenna supports.

Antenna Comparisons and Modeling

Recently, I wanted to investigate this 80-meter wire beam using computer modeling. By comparing the predicted performance of this antenna with that of several other common types of 80-meter DX antennas, the relative merits of each can be described briefly. Example antennas were modeled over average earth, with a dielectric constant of 13 and a ground conductivity of 5 mS/m. Gain figures listed are referenced to an isotropic antenna in free space.

Fig 8A shows the elevation pattern of the wire beam 125 feet high over real earth, compared to single horizontal dipoles at heights of 125 and at 70 feet. Peak radiation for the Yagi occurs at 28°. The azimuth pattern shown in Fig 8B is at 15° elevation, because low elevation angles are required for success on most DX paths from Southern California. For example, the W6 short path to Europe supports signals between 8° and 18° elevation, essentially 100% of the time.[5] The Yagi has about 5 dB of gain over the high dipole.

Note that a single high dipole is an excellent DX performer, and likely will beat any sort of beam antenna on a 70 to 80-foot tower. In addition, a high dipole has better rejection of local (high-angle) signals, so it will be a superior DX receiving antenna. Inspection of Figure 8A indicates that the high wire beam probably will outperform a 70-foot high dipole by nearly 10 dB at low elevation angles.

Some people might say: "With 140-foot towers, you should be using quad loops!" Not so fast. Figure 9A shows that with the top of a diamond-like, full-wave loop at 135 feet, the $\lambda/2$ dipole at 125 feet is superior by at least 1 dB at all angles lower than 40°, in spite of the small stacking gain expected from the loop. The expected gain from the bigger loop cannot be realized fully because the upper and lower wires are at drastically different heights (in terms of fractional wavelengths) relative to each other and the earth. This prevents proper phase addition. Further, the average height of the loop is lower, which results in a higher peak angle of radiation. The dipole has better high-angle rejection, too, and should be better on receive. If both antennas were up a couple of wavelengths on 80 meters, the quad loop would beat the

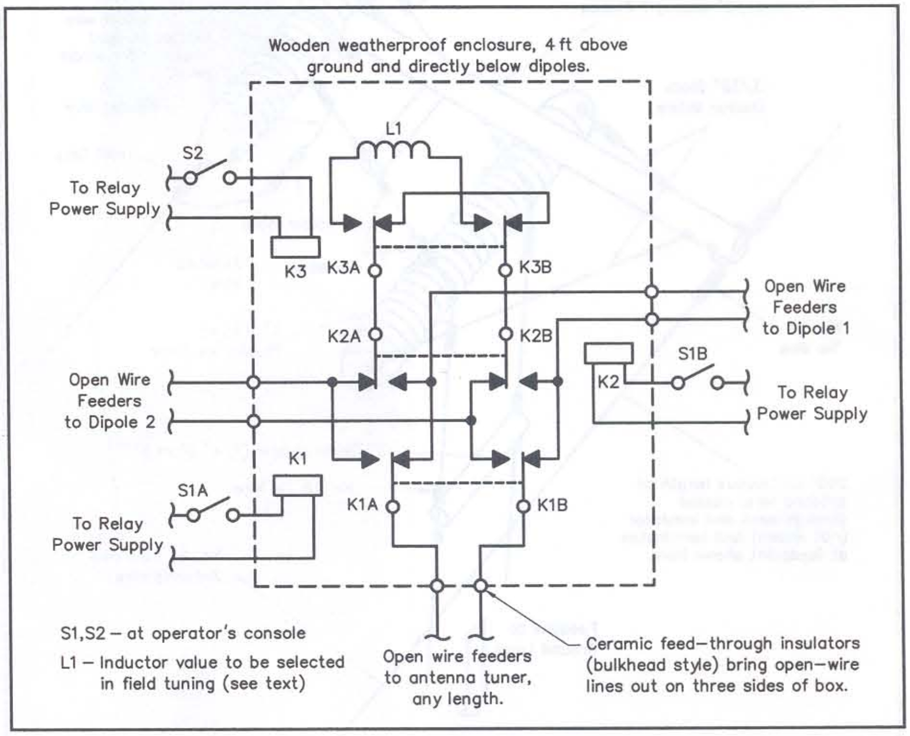
Wooden weatherproof enclosure, 4 ft above ground and directly below dipoles.
S2
To Relay Power Supply
L1
K3
K3A
K3B
Open Wire Feeders to Dipole 1
K2A
K2B
S1B
K2
To Relay Power Supply
Open Wire Feeders to Dipole 2
K1
S1A
To Relay Power Supply
K1A
K1B
S1,S2 – at operator's console
L1 – Inductor value to be selected in field tuning (see text)
Open wire feeders to antenna tuner, any length.
Ceramic feed–through insulators (bulkhead style) bring open–wire lines out on three sides of box.

Figure 7—Schematic of the feed system for the reflector-type, 80-meter wire Yagi. With two identical dipoles tuned at 3.650 MHz, only three relays and one inductor are needed to shift the antenna instantly from the phone to the CW subband, and to switch the beam direction. The relays are open-frame DPDT relays with ceramic insulation. S1 is a DPST switch used to reverse the beam direction by activating K1 and K2. S2 activates K3 to tune the parasitic element as a CW reflector.

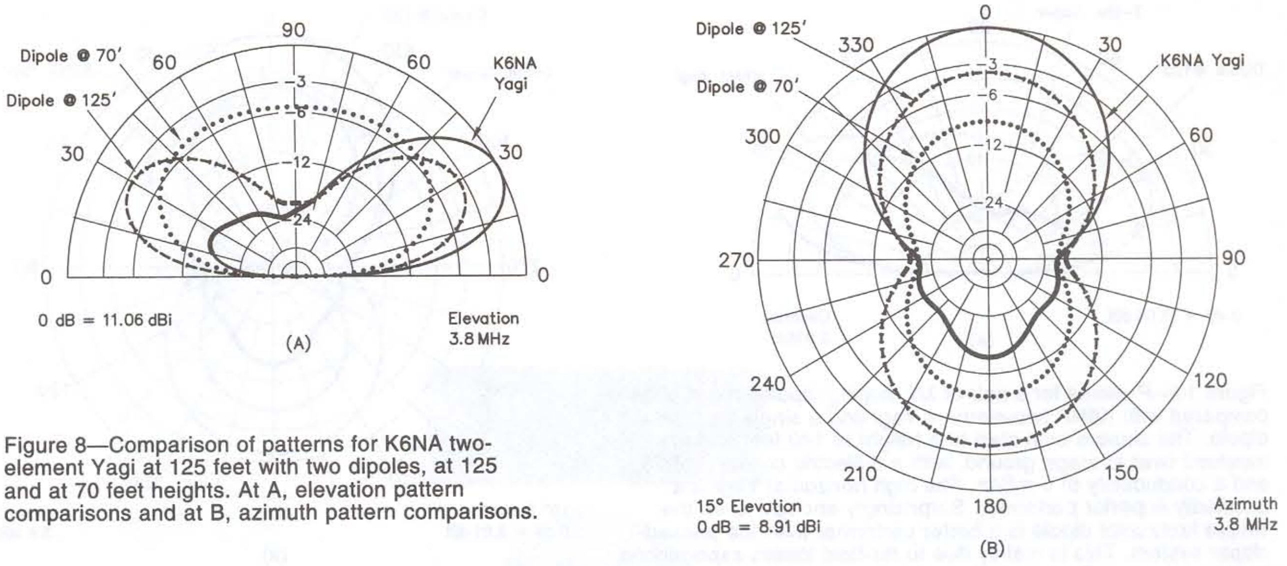

Figure 8—Comparison of patterns for K6NA two-element Yagi at 125 feet with two dipoles, at 125 and at 70 feet heights. At A, elevation pattern comparisons and at B, azimuth pattern comparisons.

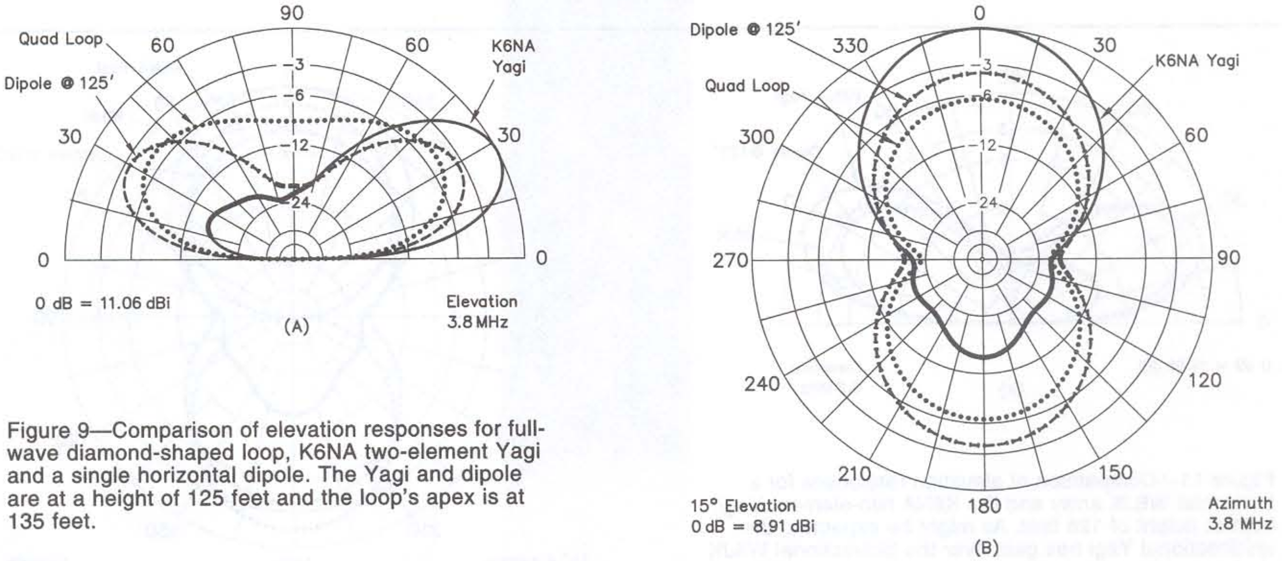

Figure 9—Comparison of elevation responses for full-wave diamond-shaped loop, K6NA two-element Yagi and a single horizontal dipole. The Yagi and dipole are at a height of 125 feet and the loop's apex is at 135 feet.

dipole...but who has a 550-foot tower?

How about a pair of half-wave slopers (that is, tilted vertical dipoles) tied from the tops of the 140-foot towers one λ/2 apart, and fed in phase for broadside gain? See Figure 10A. All the antenna books show the phased-sloper array to be a 4 dB gain, low-angle monster. The author had this exact antenna up for a few months in 1980. It performed fairly well, but was very responsive to powerline noise. For comparison, a single horizontal dipole was hung between the towers a few feet above the top ends of the slopers. With this single dipole, my local noise level dropped by 8 to 10 dB, and stations in New Zealand reported a consistent transmit advantage over the sloper array. This was a real eye-opener. As seen in Fig 10A, the high wire beam *buries* the phased slopers by 7 dB at the peak angle of 28°, and by 5 dB at 15° elevation!

The famous W8JK is a bi-directional, all-driven wire beam made from two dipoles fed 180° out of phase. At 125 feet in the air, this antenna has excellent rejection of high-angle signals, and is only slightly down from the peak gain of the Yagi configuration. See Figure 11A. It requires no relay box. For DXing, it is inferior to the Yagi because there is no way to reject signals (noise) off the back. Years ago, the author fed the two new folded dipoles as a W8JK for a few months prior to building the switchbox. It performed much like the patterns show, but once the switchbox was available to configure the array as a Yagi with good F/B ratio, the W8JK configuration was abandoned. However, a W8JK antenna 100-feet high in Kansas City would be an outstanding Sweepstakes antenna.

Figure 12 compares the horizontal wire beam to a similar antenna with inverted-V elements with a 90° included angle. The V-style gain is down a bit and the pattern is not as clean. But if you have a single tall tower and install this antenna (only one crossarm is required), you will be in the 95th percentile of effective 80-meter DXing antennas.

In Figure 13 the K6NA Yagi is compared to a quarter-wave vertical mounted over average ground. Since there are less losses due to ground reflection characteristics, horizontal polarization allows the Yagi to outperform substantially the verti-

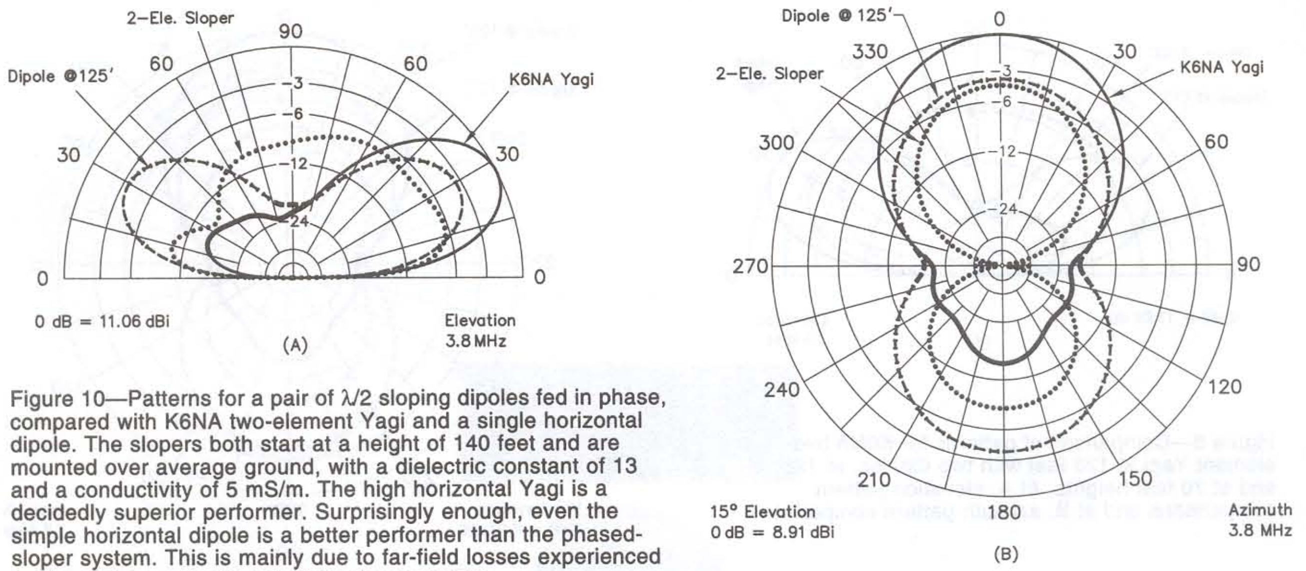

Figure 10—Patterns for a pair of λ/2 sloping dipoles fed in phase, compared with K6NA two-element Yagi and a single horizontal dipole. The slopers both start at a height of 140 feet and are mounted over average ground, with a dielectric constant of 13 and a conductivity of 5 mS/m. The high horizontal Yagi is a decidedly superior performer. Surprisingly enough, even the simple horizontal dipole is a better performer than the phased-sloper system. This is mainly due to far-field losses experienced by the mainly vertically polarized sloper system.

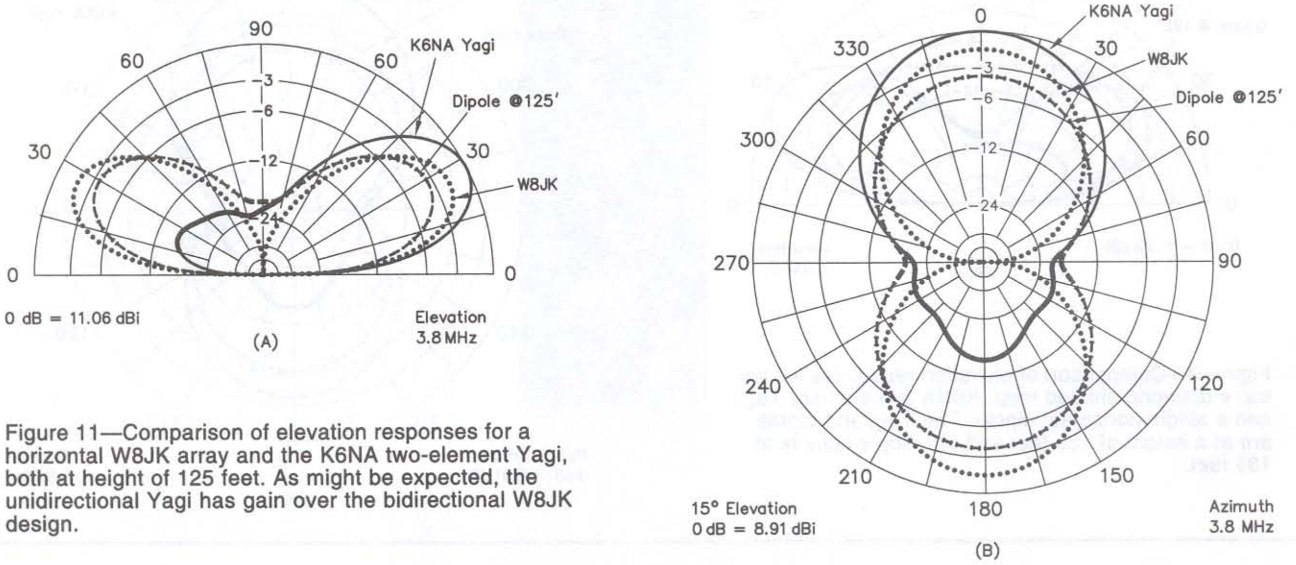

Figure 11—Comparison of elevation responses for a horizontal W8JK array and the K6NA two-element Yagi, both at height of 125 feet. As might be expected, the unidirectional Yagi has gain over the bidirectional W8JK design.

cal at all angles of interest, even down to 10°. In most locations, the vertical will be noisy on receive. *An extensive local ground screen (or elevated radials) will not make any significant change in the pattern relationship depicted here.* Only by installing the vertical in, say, a saltwater marsh extending perhaps 100 λ would its low-angle performance be improved substantially.[6]

The "Four-Square" has become popular in recent years. Figure 13 also shows the beautiful pattern of this array, along with that of the two-element beam and a single vertical. In a quiet location this vertical system with a switching matrix makes a terrific receiving antenna, allowing full azimuth coverage in four directions. However, the horizontal wire Yagi still has superior gain at all elevation angles of interest—note the 4 dB advantage at 15°—and the Yagi likely will be quieter on receive in all but the most remote locations. The remarks in the paragraph above about attempts to improve the lowangle performance of the λ/4 vertical also apply to the Four-Square array.

Conclusions

Someone once said: "All antennas work... some more so." The sometimes-large differences seen during model comparisons can be misleading, as we know from our day-to-day operating. Propagation, pileup dynamics, and operator skills (one skill is picking a quiet location ...) are important factors that determine 80-meter DXing success. Virtually every antenna discussed in the comparison section of this report is a "good" antenna. The author's first 180 countries on 80 meters were worked with an inverted-V dipole at 70 feet.

If *tall* supports are available, a horizontally polarized, gain antenna will certainly provide a statistical increase in performance over the others. In lieu of really tall

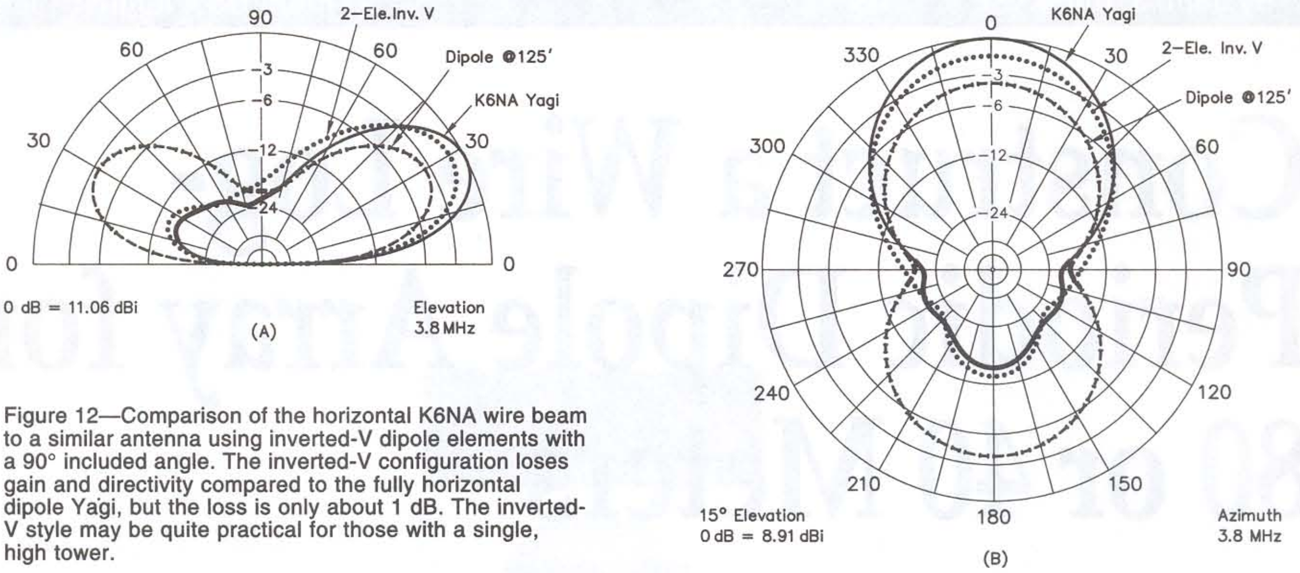

Figure 12—Comparison of the horizontal K6NA wire beam to a similar antenna using inverted-V dipole elements with a 90° included angle. The inverted-V configuration loses gain and directivity compared to the fully horizontal dipole Yagi, but the loss is only about 1 dB. The inverted-V style may be quite practical for those with a single, high tower.

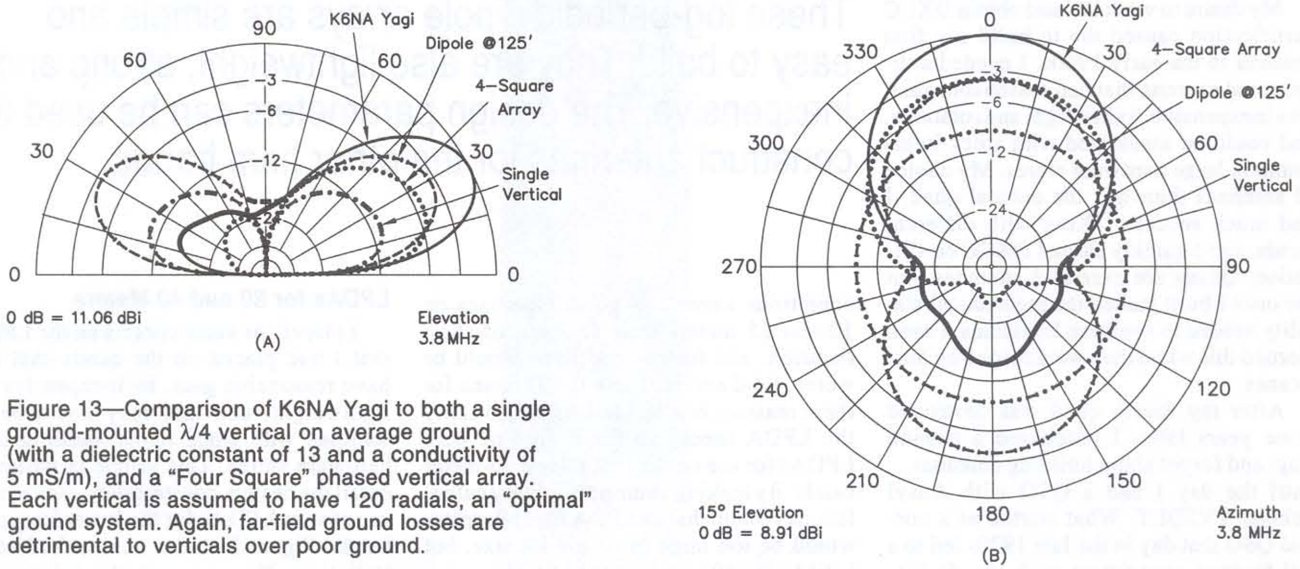

Figure 13—Comparison of K6NA Yagi to both a single ground-mounted λ/4 vertical on average ground (with a dielectric constant of 13 and a conductivity of 5 mS/m), and a "Four Square" phased vertical array. Each vertical is assumed to have 120 radials, an "optimal" ground system. Again, far-field ground losses are detrimental to verticals over poor ground.

towers, a well-installed vertical system can be an excellent antenna, especially in a quiet receiving location. As the computer models show, height above ground is the most significant variable when designing a lowband station using horizontally polarized antennas. [Hills help too!–*Ed.*]

It is hoped that some of the construction techniques for the crossarms, wire antennas, and feed lines described in this report will stimulate others to build even better horizontal arrays. You will enjoy special satisfaction using homemade open-wire lines, and a 40-year old Johnson Matchbox tuner makes a nice addition to your modern shack.

The author would like to thank H. Shepherd, W6US, for his valuable advice regarding wire antennas and open-wire feeder construction; R. Craig, N6ND, for mechanical suggestions; E. Andress, W6KUT, for manuscript review; and N6ND and J. McCook, W6YA, for modeling assistance.

Notes and References

[1] Wayne Overbeck, N6NB, "Quads vs. Yagis Revisited," *Ham Radio*, May 1979, p 12.

[2] James L. Lawson, W2PV, "160/80/75-Meter Broad-Band Inverted-V Antenna," *QST*, Nov 1970, p 17.

[3] James L. Lawson, W2PV, *Yagi Antenna Design*, W. Myers, K1GQ, editor (Newington: ARRL, 1986), p 2-2.

[4] Doug DeMaw, W1FB, editor, *ARRL Electronics Data Book* (Newington: ARRL, 1976), p 82.

[5] Frank Witt, AI1H, "Optimizing the 80-Meter Dipole," *The ARRL Antenna Compendium, Vol 4* (Newington: ARRL, 1995), p 39.

[6] Charles J. Michaels, W7XC, "Some Reflections on Vertical Antennas," *QST*, Jul 1987, p 18.

By John J. Uhl, KV5E **From *QST*, August 1986**

Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters

These log-periodic dipole arrays are simple and easy to build. They are also lightweight, strong and inexpensive. The design parameters can be used to construct antennas for the other ham bands.

My desire to work DX and obtain DXCC certification caused me to build my first antenna in the early 1960s. I needed a directional antenna that had reasonable gain, was inexpensive, lightweight and rotatable, and could be assembled with stock items found in large hardware stores. My choice of antennas then was the cubical quad. I had much success DXing with different quads, and I quickly earned DXCC certification. Quads are excellent antennas, but the ones I built lacked the mechanical stability needed in southern Louisiana. I soon learned this when they were ruined by hurricanes.

After my fourth quad was destroyed some years later, I purchased a triband Yagi and forgot about building antennas ... until the day I had a QSO with Ansyl Eckols, YV5DLT. What started as a normal QSO that day in the late 1970s led to a full-fledged experiment with the design, construction, erection and use of log-periodic dipole arrays made of wire. At that time, YV5DLT was using a triband log-periodic dipole array (LPDA) for 20, 15 and 10 meters. What immediately piqued my interest was that his beam was made of wire, and that his signal had outstanding quality and strength.

During the QSO, I asked Ansyl for construction details of his antenna. His response was generous. He mailed me diagrams, schematics and photographs of the LPDA that he had named Telerana. (He subsequently published his design in *QST.*[1]) After reading and studying all of his data, I was convinced that his design had the mechanical stability to withstand hurricanes, and I began plans in my mind to build a copy of Telerana.

I began a search of the literature, reading all of the LPDA articles that I could find.[2-8] By the time I gathered and read several references, three years had passed. Sunspot activity had diminished and band conditions weren't as good. Openings on 10 and 15 meters were few and of short duration, and future conditions would be worse. I did not duplicate the Telerana for these reasons, but decided instead to apply the LPDA theory to the design of wire LPDAs for use on the 160, 80 and 40-meter bands. By making some preliminary calculations I found that an LPDA for 160 meters would be too large to fit my lot size, but LPDAs for 80 and 40 would fit. However, it would not be possible for me to rotate these LPDAs.

LPDAs for 80 and 40 Meters

I placed the same criteria on the LPDAs that I had placed on the quads–that they have reasonable gain, be inexpensive and lightweight, and that they could be assembled with stock items found in large hardware stores. This article is written to detail the design, construction, erection and use of wire LPDAs for the lower frequency bands. Figure 1 shows one method of installation. You can use the information presented here as a guide and point of reference for building similar LPDAs.

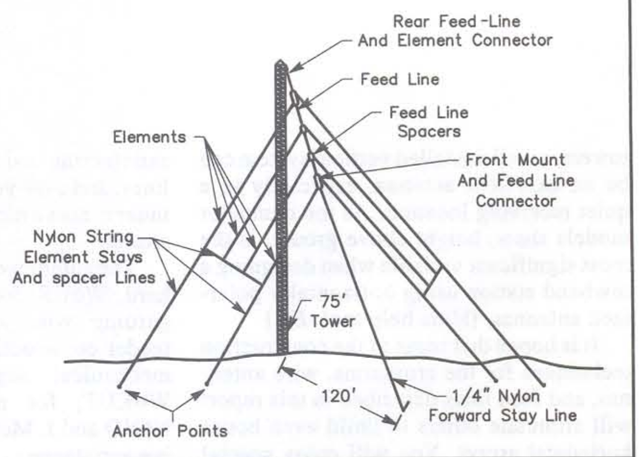

Figure 1—Typical 4-element log-periodic dipole array erected on a tower.

If space is available, the antennas can be "rotated" or repositioned in azimuth after they are completed. A 75-ft tower and a clear turning radius of 120 ft around the base of the tower are needed. The task is simplified if only three anchor points are used, instead of the five shown in Figure 1. Omit the two anchor points on the forward element, and extend the two nylon strings used for element stays all the way to the forward stay line.

For the design procedure, refer to *The ARRL Antenna Book.*[9] By using the formulas given there and other data in the text, all of the dimensions can be obtained and the LPDAs will take shape on paper. The design results are summarized in the concluding section of this article. The next step is to fabricate the fittings; see Figure 2 for details. Cut the wire elements and feed lines to the proper sizes and mark them for identification. After the wires are cut and placed aside, it will be difficult to remember which is which unless they are marked. When you have finished fabricating the connectors and cutting all of the wires, the antenna can be assembled. Use your ingenuity when building one of these antennas; it isn't necessary to duplicate my LPDAs exactly.

The elements are made of standard no. 14 stranded copper wire. The two parallel feed lines are made of no. 12 solid copper-coated steel wire, such as Copperweld. This will not stretch when placed under tension. The front and rear connectors are cut from 1/2-in-thick Lexan sheeting, and the feed-line spacers from 1/4-in Plexiglas sheeting.

Study the plans carefully and be familiar with the way the wire elements are connected to the two feed lines, through the front, rear and spacer connectors. Details are sketched in Figure 4. Connections made this way prevent the wire from breaking. All of the rope, string and connectors must be made of materials that can withstand the effects of tension and weathering. Use nylon rope and strings, the type that yachtsmen use. Figure 1 shows the front stay rope coming down to ground level at a point 120 ft from the base of a 75-ft tower. It may not be possible to do this in all cases. In my installation I put a pulley 40 ft up in a tree and ran the front stay rope through the pulley and down to ground level at the base of the tree. The front stay rope will have to be tightened with a block and tackle at ground level.

Putting an LPDA together is not difficult if it is assembled in an orderly manner. It is easier to connect the elements to the feeder lines when the feed-line assembly is stretched between two points. Use the tower and a block and tackle. Attaching the rear connector to the tower and assembling the LPDA at the base of the tower makes raising the antenna into place a much simpler task. Tie the rear connector securely to the base of the tower and attach the two feeder lines to it. Then thread the two feed-line spacers onto the feed line. The spacers will be loose at this time, but will be positioned properly when the elements are connected. Now connect the front connector to the feed lines. A word of caution: Measure accurately and carefully! Double-check all measurements before you make permanent connections.

Connect the elements to the feeder lines through their respective plastic connectors, beginning with element 1, then element 2, and so on. Keep all of the element wires securely coiled. If they unravel, you will have a tangled mess of kinked wire. Check that the element-to-feeder connections have been made properly. (See Figure 4.) Once you have completed all of the element connections, attach the 4:1 balun to the underside of the front connector. Connect the feeder lines and the coaxial cable to the balun.

You will need a separate piece of rope and a pulley to raise the completed LPDA into position. First secure the eight element ends with nylon string, referring to Figures 1 and 3. The string must be long enough to reach the tie-down points. Connect the front stay rope to the front connector, and the completed LPDA is now ready to be raised into position. While raising the antenna, uncoil the element wires to prevent their getting away and balling up into a mess. Use care! Raise the rear connector to the proper height and attach it securely to the tower, then pull the front stay rope tight and secure it. Move the elements so that they form a 60-degree angle with the feed lines, in the direction of the front, and space them properly relative to one another. By adjusting the end positions of the elements as you walk back and forth, you will be able to align all the elements properly. Now it is time to hook your rig to the system and make some QSOs.

Performance

The reports I received using the LPDAs were compared with an inverted-V dipole. All of the antennas are fixed; the LPDAs radiate to the northeast and the dipole to the northeast and southwest. The apex of the

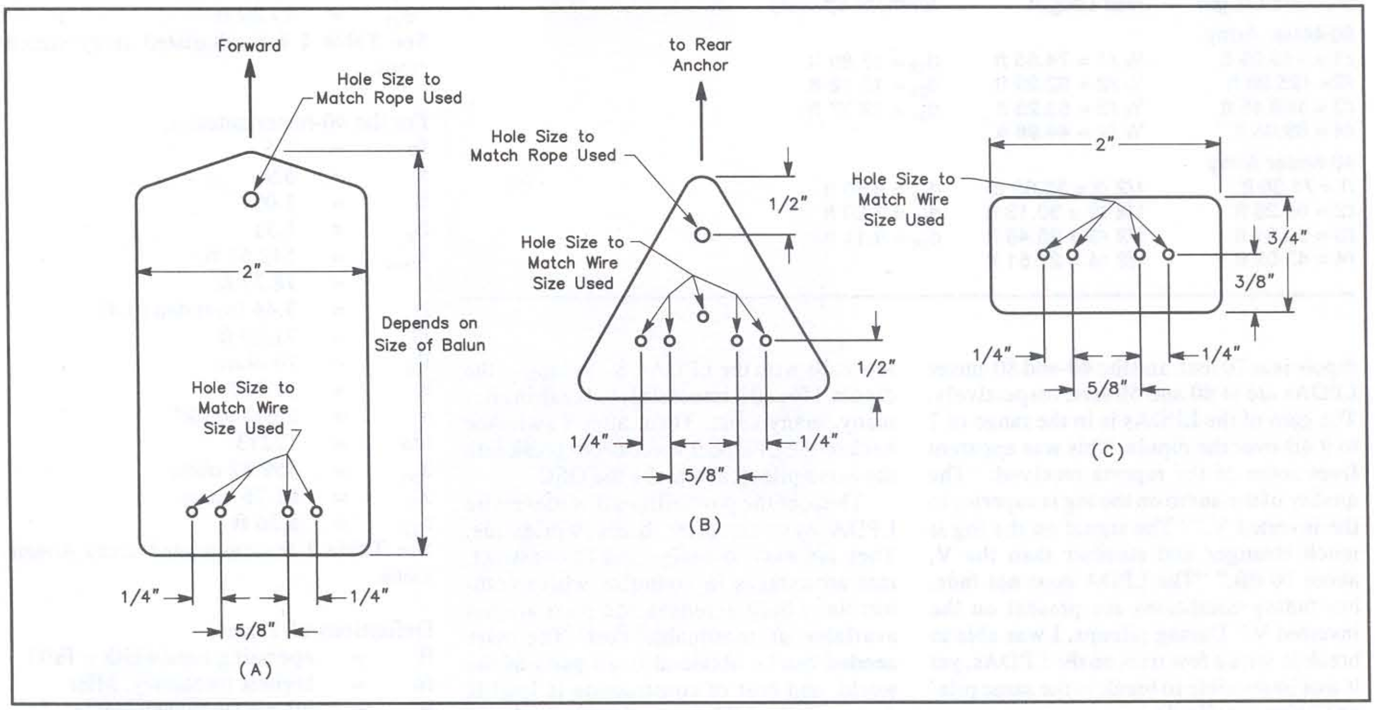
Forward
Hole Size to Match Rope Used
2"
Depends on Size of Balun
Hole Size to Match Wire Size Used
1/4"
5/8"
1/4"
(A)
to Rear Anchor
Hole Size to Match Rope Used
1/2"
Hole Size to Match Wire Size Used
1/2"
1/4"
5/8"
1/4"
(B)
2"
Hole Size to Match Wire Size Used
3/4"
3/8"
1/4"
5/8"
1/4"
(C)

Figure 2—Pieces to be fabricated for the LPDA. At A, the forward connector, made from 1/2-inch Lexan. At B, the rear connector, also made from 1/2-inch Lexan. At C is the pattern for the feed-line spacers, made from 1/4-inch Plexiglas. Two of these spacers are required.

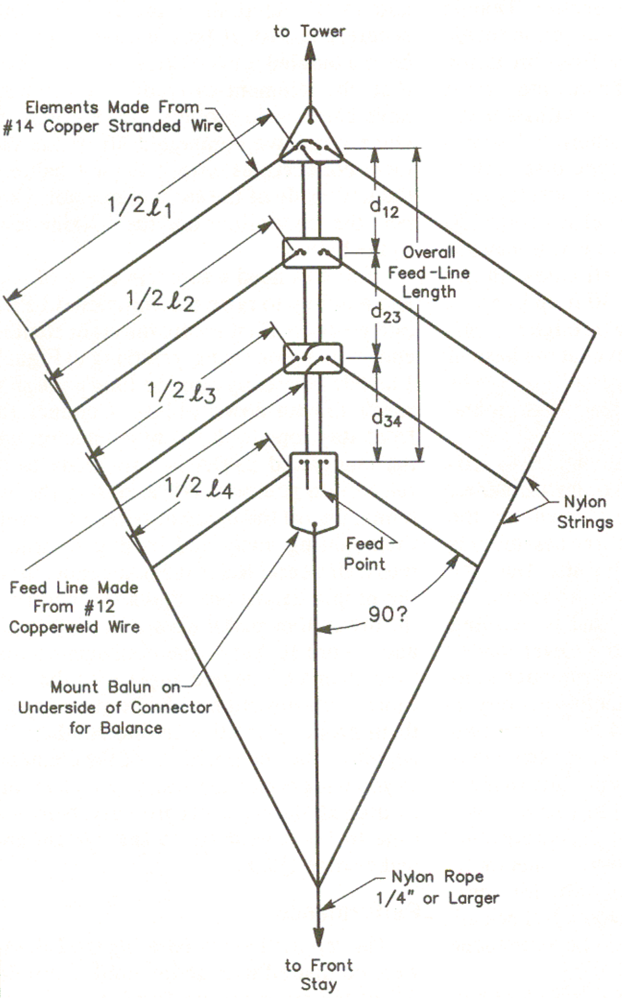
to Tower
Elements Made From #14 Copper Stranded Wire
1/2ℓ1
1/2ℓ2
1/2ℓ3
1/2ℓ4
d12
Overall Feed-Line Length
d23
d34
Nylon Strings
Feed Point
90?
Feed Line Made From #12 Copperweld Wire
Mount Balun on Underside of Connector for Balance
Nylon Rope 1/4" or Larger
to Front Stay

Figure 3—Typical layout for the LPDA. Use a 4:1 balun at the point indicated. See Table 1 for dimensions.

Table 1
Calculated Array Dimensions

Element Length	*Half Length*	*Element Spacing*
80-Meter Array		
ℓ1 = 149.09 ft	½ ℓ1 = 74.55 ft	d_{12} = 17.89 ft
ℓ2= 125.98 ft	½ ℓ2 = 62.99 ft	d_{23} = 15.12 ft
ℓ3 = 106.45 ft	½ ℓ3 = 53.23 ft	d_{34} = 12.77 ft
ℓ4 = 89.95 ft	½ ℓ4 = 44.98 ft	
40-Meter Array		
ℓl = 71.30 ft	1/2 ℓl = 35.65 ft	d_{12} = 8.56 ft
ℓ2 = 60.25 ft	1/2 ℓ2 = 30.13 ft	d_{23} = 7.23 ft
ℓ3 = 50.91 ft	1/2 ℓ3 = 25.46 ft	d_{34} = 6.11 ft
ℓ4 = 43.02 ft	1/2 ℓ4 = 21.51 ft	

dipole is at 70 feet, and the 40-and 80-meter LPDAs are at 60 and 50 feet, respectively. The gain of the LPDAs is in the range of 7 to 9 dB over the dipole. This was apparent from some of the reports received: "The quality of the audio on the log is superior to the inverted V." "The signal on the log is much stronger and steadier than the V, about 10 dB." "The LPDA does not fade, but fading conditions are present on the inverted V." During pileups, I was able to break in with a few tries on the LPDAs, yet it was impossible to break in the same pileups using the dipole.

During the CQ WW DX Contest I was able to break into some *big* pileups after a few calls with the LPDAs. Switching to the dipole, I found it impossible to break in after many, many calls. Then, after I switched back to the LPDA, it was easy to break into the same pileup and make the QSO.

Think of the possibilities that these wire LPDA systems offer hams worldwide. They are easy to design and to construct, real advantages in countries where commercially built antennas and parts are not available at reasonable cost. The wire needed can be obtained in all parts of the world, and cost of construction is low! If damaged, the LPDAs can be repaired easily with pliers and solder. For those who travel on DXpeditions where space and weight are large considerations, LPDAs are lightweight but sturdy, and they perform well. They'll even withstand a hurricane!

Calculations for Log-Periodic Dipole Arrays

Design constants and the results of design procedures follow. (Terms are defined at the end of this section.)

τ = 0.845
σ = 0.06
σ_{opt} = 0.152
cot α = 1.548
α = 32.86°
Gain = 7.5 dBi (5.35 dBd). [By sloping the elements forward, the gain may be increased 3 to 5 dB over this figure. *–Ed.*]
σ' = 0.065
B_{ar} = 1.39; see note 10.
R_0 = 70

Fed with 50-ohm coaxial cable and a 4:1 balun

For the 80-meter antenna,
fn = 4.1
f1 = 3.3
B = 1.24
B_S = 1.72
λ_{max} = 298.18 ft
L = 48.42 ft
N = 4.23 (rounded to 4)
ℓ1 = 149.09 ft
R_0 = 70 ohms
h = 62.4
a = 2.667×10^{-3}
h/a = 23400
Z_{av} = 937.26 ohms
Z_0 = 80.72 ohms
d_{12} = 17.89 ft
See Table 1 for calculated array dimensions.

For the 40-meter antenna,
fn = 7.5
fl = 6.9
B = 1.09
B_s = 1.51
λ_{max} = 142.61 ft
L = 18.57 ft
N = 3.44 (rounded to 4)
ℓ1 = 71.30 ft
R_0 = 70 ohms
h = 32.727
a = 2.667×10^{-3}
h/a = 12273
Z_{av} = 859.82 ohms
Z_0 = 81.76 ohms
d_{12} = 8.56 ft
See Table 1 for calculated array dimensions.

Definitions of Terms
B = operating bandwidth = fn/f1
fn = highest frequency, MHz
fl = lowest frequency, MHz
τ = design constant
σ = relative spacing constant

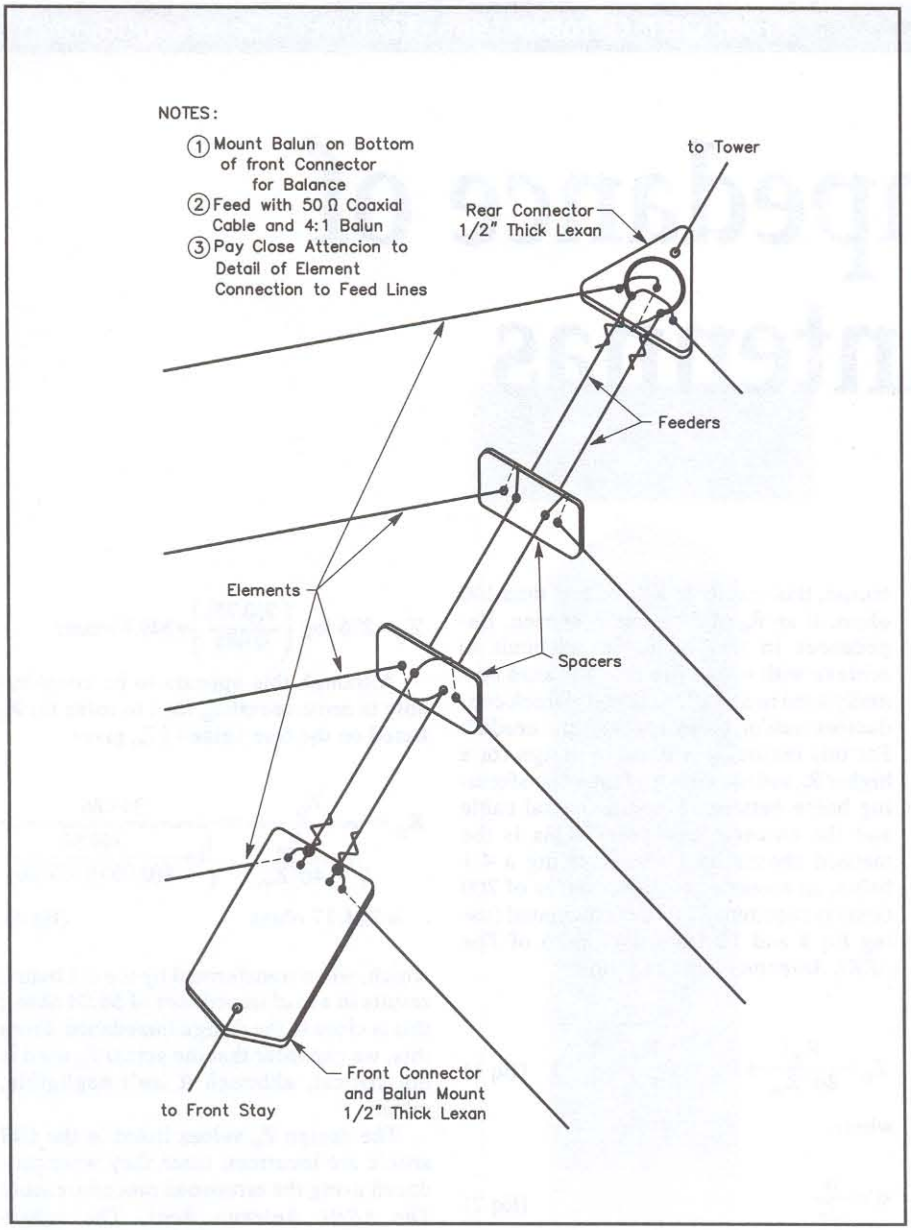
NOTES:
① Mount Balun on Bottom of front Connector for Balance
② Feed with 50 Ω Coaxial Cable and 4:1 Balun
③ Pay Close Attencion to Detail of Element Connection to Feed Lines
to Tower
Rear Connector— 1/2" Thick Lexan
Feeders
Elements
Spacers
Front Connector and Balun Mount 1/2" Thick Lexan
to Front Stay

Figure 4—Details of electrical and mechanical connections of the elements to the feed line. Knots in the nylon stay lines are not shown.

σ_{opt} = value of σ for optimum gain
σ' = mean spacing factor
α = apex half-angle
B_{ar} = bandwidth of the active group. See note 10.
B_s = structure (array) bandwidth
L = boom length for N elements
N = number of elements
$\ell 1$ = longest element = 492/f1
λ_{max} = longest free-space wavelength = 984/f1
Z_0 = characteristic impedance of feeder
R_0 = mean radiation resistance level of required input impedance of active region
Z_{av} = average characteristic impedance of a dipole

$$= 120\left(\ln\frac{h}{a} - 2.25\right)$$

h = element half length
a = radius of element
ℓ = length of elements
d = spacing between elements

Notes

[1] A. Eckols, "The Telerana—A Broadband 13- to 30-MHz Directional Antenna," *QST*, Jul 1981, pp 24-27.

[2] C. T. Milner, "Log Periodic Antennas," *QST*, Nov 1959, p 11.

[3] D. E. Isbell, "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation*, Vol AP-8, No. 3, May 1960, pp 260-267.

[4] P. D. Rhodes, "The Log-Periodic Dipole Array," *QST*, Nov 1973, pp. 19-22.

[5] *The GIANT Book of Amateur Radio Antennas* (Blue Ridge Summit, PA: Tab Books, 1979), pp 55-85.

[6] C. A. Balanis, *Antenna Theory, Analysis and Design* (New York: Harper and Row, 1982), pp 427-439.

[7] G. L. Hall, ed., *The ARRL Antenna Book* (Newington, CT: The American Radio Relay League, Inc, 1982), pp 6-24 to 6-26, 9-12 to 9-14.

[8] D. A. Mack, "A Second-Generation Spiderweb Antenna," *The ARRL Antenna Compendium Vol 1* (Newington, CT: The American Radio Relay League, Inc, 1985), pp 55-59.

[9] See note 7.

[10] See Ref 6, p 435. B_{ar} is found from the following equation:
$B_{ar} = 1.1 + 7.7(1 - \tau)^2 \cot \alpha$

From *QST*, October 1986 (Technical Correspondence)

Input Impedance of LPDA Antennas

The log-periodic dipole array (LPDA) has intriguing possibilities, some of which have been explored by Uhl.[1] Unfortunately, some misleading information in *The ARRL Antenna Book* has kept the subject from being more accessible to the amateur.[2] The design procedure called out by *The Antenna Book* and by Rhodes refers to graphs of R_0, the mean antenna input resistance, versus the log-periodic design parameters τ and α.[3] It would appear that R_0 is dependent solely on these parameters; such is not the case. As shown by Eq 2 and 4, R_0 is dependent on τ, σ, Z_{av} and Z_0. This relationship is demonstrated by Carrel.[4] The graphs in question are from the seminal experimental study of the LPDA by Isbell, and are used to demonstrate the effect on R_0 of changing τ and σ; Z_{av} and Z_0 were constant.[5] Since τ and σ are usually chosen to provide the desired antenna geometry and directivity (gain), R_0 should be controlled by either Z_{av} or Z_0. Z_{av} is usually constrained by mechanical considerations, which leaves Z_0 as the easiest parameter to adjust for desired R_0.

The erroneous procedure determines R_0 from a graph. The correct procedure is to calculate the necessary Z_0 based on the *desired* R_0. For most practical LPDA antennas, this results in a Z_0 of less than 100 ohms, if an R_0 of 50 ohms is chosen. Impedances in this range are difficult to achieve with open-wire line, the kind normally used in an LPDA, because thick conductors and/or close spacing are needed. For this reason, it is usual to design for a higher R_0 and use an impedance-transforming balun between 50-ohm coaxial cable and the antenna feed point. This is the method chosen by Uhl. Assuming a 4:1 balun, an antenna-input impedance of 200 ohms is required. Z_0 is then calculated (using Eq 9 and 11 from page 6-25 of *The ARRL Antenna Book*) as follows:

$$Z_0 = \frac{R_0^{\,2}}{8\sigma' Z_{av}} + R_0\sqrt{\left(\frac{R_0}{8\sigma' Z_{av}}\right)^2 + 1} \qquad \text{(Eq 1)}$$

where

$$\sigma' = \frac{\sigma}{\sqrt{\tau}} \qquad \text{(Eq 2)}$$

$$Z_0 = \frac{200^2}{8(0.065)\,(937.26)} +$$

$$200\sqrt{\left(\frac{200}{8(0.065)\,(937.26)}\right)^2 + 1}$$

$$= 298.26 \text{ ohms}$$

The actual impedance of the feed line used by Uhl can be computed from the formula on p 3-16 of *The ARRL Antenna Book*:

$$Z_0 = 276 \log\left(\frac{2S}{d}\right) \qquad \text{(Eq 3)}$$

where S is the spacing between conductors, 0.75 inch, and d is the conductor diameter, 0.081 inch for no. 12 AWG wire. Therefore:

$$Z_0 = 276 \log\left(\frac{2(0.75)}{0.081}\right) = 349.86 \text{ ohms}$$

Although this appears to be considerably in error, recasting Eq 1 to solve for R_0 based on the true value of Z_0 gives:

$$R_0 = \frac{Z_0}{\sqrt{1 + \dfrac{Z_0}{4\sigma' Z_{av}}}} = \frac{349.86}{\sqrt{1 + \dfrac{349.86}{4(0.065)(937.26)}}}$$

$$= 224.17 \text{ ohms} \qquad \text{(Eq 4)}$$

which, when transformed by the 4:1 balun, results in a load impedance of 56.04 ohms; this is close to the design impedance. From this, we can infer that the actual Z_0 used is not critical, although it isn't negligible, either.

The design Z_0 values listed in the Uhl article are incorrect, since they were produced using the erroneous procedure from *The ARRL Antenna Book*. The values should be 298.26 and 308.56 ohms for the 80- and 40-meter antennas, respectively. In addition, R_0 could be made closer to the design value of 200 ohms by using a different feeder spacing to change the feeder impedance. From Eq 2:

$$S = \left(\frac{d}{2}\right)10^{(Z_0/276)} = \left(\frac{0.081}{2}\right)10^{(298.26/276)}$$

$$= 0.488 \text{ inch}$$

Rounding this value to 0.5 inch results in a Z_0 of 301.26 ohms and an R_0 of 201.46 ohms for 80 meters and 196.62 ohms for 40 meters. These translate to load impedance of 50.37 and 49.15 ohms, respectively.—*Jon Bloom, KE3Z*

---

[1] J.J. Uhl, "Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters," *QST*, Aug 1986, p 21.

[2] G.L. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1982), pp 6-24 to 6-26.

[3] P.D. Rhodes, "The Log-Periodic Dipole Array," *QST*, Nov 1973, pp 16-22.

[4] R.L. Carrel, "The Design of Log-Periodic Dipole Antennas," *1961 IRE International Convention Record, Part 1, Antennas and Propagation*, pp 61-75; also PhD thesis, University of Illinois, Urbana, 1961.

[5] D. E. Isbell, "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation*, Vol AP-8, No. 3, May 1960, pp 260-267.

From *QST*, August 1980 (Technical Correspondence)

The WA1AKR 40- and 75-Meter Slopers

Several amateurs have suggested that I submit a description of my sloper antenna system for publication in "Hints and Kinks." Other amateurs may be interested in this adaptation of the 8JK beam. Construction information is shown for both the 75- and 40-meter bands.

As shown in the accompanying diagram, the array has two half-wave sloping elements joined by a 1/8-wave, 300-ohm phasing line. Transposing the phasing line should bring the element currents into phase. I find the antenna is broadbanded. There appears to be no need for a Transmatch.

If one desires to suspend an additional sloper from the tower for a directional change, installation of remote switching at the top of the tower will permit the use of a single transmission line. Otherwise, separate transmission lines will be required.

Ends of the antenna are suspended by ropes with the tops placed roughly 1 foot away from the tower. An angle of 45 degrees between the antenna and ground should be maintained. Do not use an angle greater than 50 degrees. Resonance with the dimensions shown should occur near 3.8 MHz for the 75-meter sloper and 7.150 for the 40-meter antenna.

How well do my antennas work? I have contacted stations "across the pond" while competing with the big boys who sport three and four-element beams. I have also experienced little difficulty in working VKs and ZLs. – *Carl Bissonnette, WA1AKR*

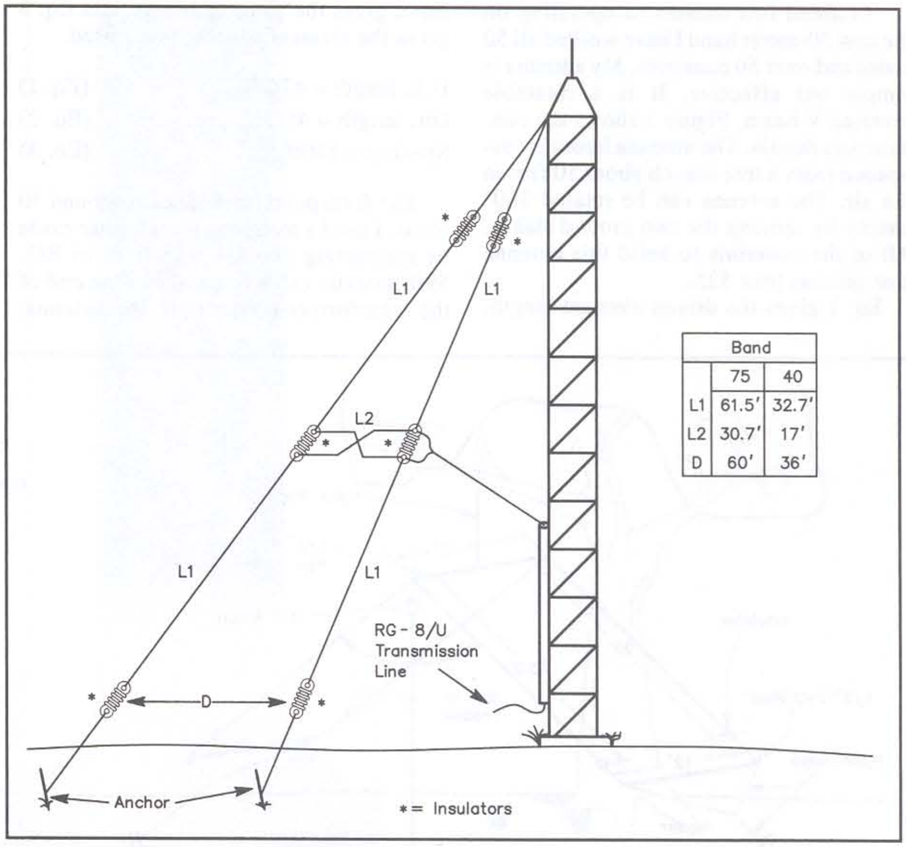

	Band	
	75	40
L1	61.5′	32.7′
L2	30.7′	17′
D	60′	36′

When Carl Bissonnette, WA1AKR, chases DX he uses a sloper like the one illustrated. Carl's arrangement is fashioned after the famous 8JK beam. The feed system resembles that of the ZL Special.

From *QST*, June 1983 (Hints & Kinks)

Inexpensive 30-Meter Beam Antenna

In about two months of operating on the new 30-meter band I have worked all 50 states and over 50 countries. My antenna is simple but effective. It is a rotatable inverted V beam. Figure 1 shows the construction details. The antenna boom is suspended from a tree branch about 50 feet in the air. The antenna can be rotated 360° simply by moving the two ground stakes. All of the materials to build this antenna cost me less than $25.

Eq. 1 gives the driven element length, Eq. 2 gives the director length, and Eq. 3 gives the element spacing that I used.

$$\text{D.E. length} = 476/f_{\text{MHz}} \qquad \text{(Eq. 1)}$$

$$\text{Dir. length} = 450/f_{\text{MHz}} \qquad \text{(Eq. 2)}$$

$$\text{Spacing} = 120/f_{\text{MHz}} \qquad \text{(Eq. 3)}$$

The feed-point impedance is around 30 ohms. I used a matching transformer made by connecting two 1/4-λ sections of RG-59/U coaxial cable in parallel. One end of the transformer connects to the antenna, and the other end goes to 50-ohm cable to the shack. Figure 1B shows how this is wired. Perhaps the easiest method to join the two pieces of 75-ohm cable is to use coaxial T connectors. You should use a balun at the antenna feed point to prevent rf from flowing on the outside of the shield braid.

I also built an antenna of this type for 40 meters, and it works great. I guess the key word is rotatable! —*Jon Ferrara, N9DWR, Chattanooga, Tennessee*

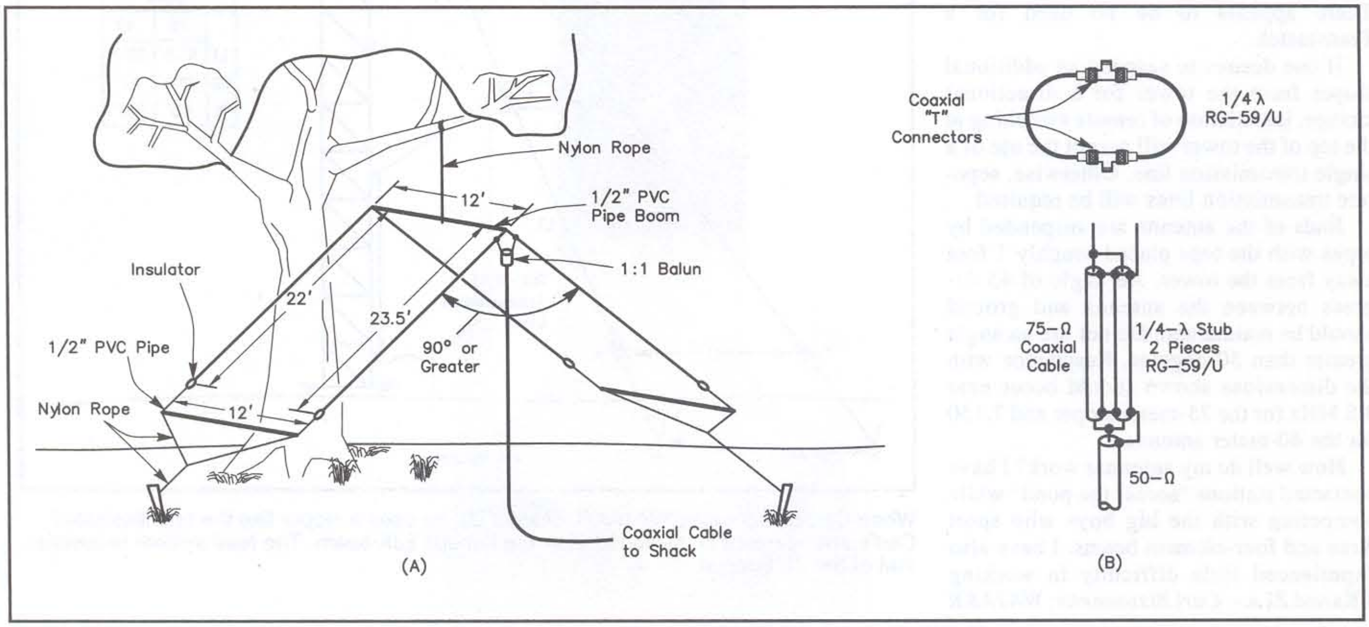
Nylon Rope
12'
1/2" PVC Pipe Boom
Insulator
1:1 Balun
22'
23.5'
1/2" PVC Pipe
90° or Greater
Nylon Rope
12'
Coaxial Cable to Shack
(A)
Coaxial "T" Connectors
1/4 λ RG-59/U
75-Ω Coaxial Cable
1/4-λ Stub 2 Pieces RG-59/U
50-Ω
(B)

Figure 1—Construction details for a 30-meter inverted V beam are given at A. A coaxial-cable impedance-matching transformer is shown at B.

By Roy W. Lewallen, W7EL **From *QST*, June 1984**

Try the "FD Special" Antenna

Looking for an antenna that's simple, inexpensive, lightweight and easy to install? Here's one that fits the description.

While the Field Day "juices" were flowing last year, my Field Day partners and I decided to become more competitive without compromising our general philosophy: Use all homemade gear, pack it in to the site, and don't let operating interfere with watching the scenery. Since most of our gear already ran the accepted QRP power limit of 10-W dc input and had been designed to provide high output efficiency, the antenna seemed like the reasonable point of attack (our high quality superheterodyne receivers have not been the limiting item). The constraints dictated that an antenna be lightweight, portable and easy to put up. It should also have substantial gain.

Because of our West Coast location, the front-to-back ratio was not a concern. But having a reasonably wide lobe toward the East was. We desired a low SWR because of the relatively high loss of our RG-58/U and/or RG-174/U feed line. We were interested only in the CW portion of the band, but this antenna works well over all of the band. It seemed that 20 meters would be our main "money-maker," so we designed the antenna for that band. It can be scaled for other bands, too.

The Research

Although a number of antenna types might have done the job, I settled quickly on a horizontal, close-spaced, driven array. Experience has shown that driven arrays are generally more tolerant of imperfect construction and erection than are parasitic arrays. Experience and much measurement have convinced me that horizontal arrays outperform vertical ones in the high-frequency bands, except perhaps from an exceptional location. In addition, we didn't want the nuisance of establishing a decent ground system — which most vertical arrays require.

The theoretical gain and front-to-back ratio of 2-element arrays with 1/8-wavelength spacing between the elements are shown in Figures 1 and 2. Note the lower curve of Figure 1. It shows the effect of losses on the gain (losses don't affect the front-to-back ratio, and change only the scaling of the pattern). Figure 3 shows the patterns of arrays with 135, 160 and 180-degree relative spacing. All are drawn to the same scale. The 1/8-wavelength spaced, 135-degree-fed array is frequently called the "ZL Special."[1] The close-spaced, 180-degree-fed array is known as an "8JK."[2] From 135 to 160 degrees, phasing was chosen because of the combination of relatively high insensitivity to loss, reasonable gain and wide forward lobes. Note that the gain stays about the same in this range, ensuring good performance if the phasing isn't exactly as predicted. Actually, it's much easier to generate and maintain precise 180-degree phasing than the angles I've chosen — particularly over a wide frequency range.

There's one major flaw (usually fatal) in a simple analysis like the one presented here: It assumes that equal-magnitude currents are flowing in the elements. This is not easy to realize, for even in arrays with elements spaced 1/2 wavelength or greater, mutual coupling has a profound effect on element impedances. This changes them dramatically and unequally, as a rule. This

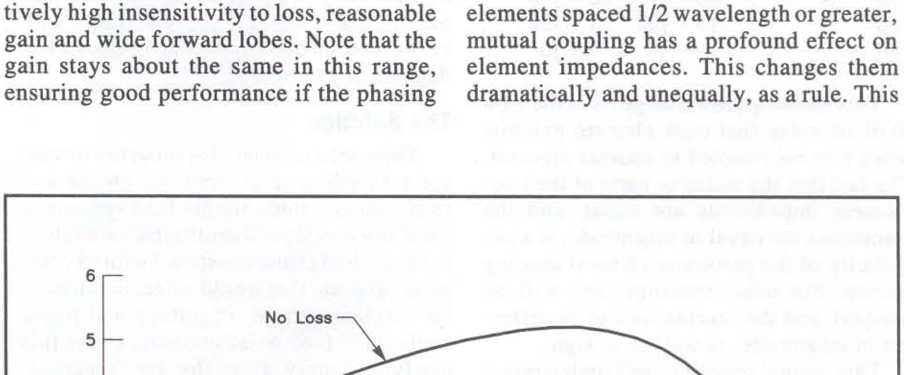

Figure 1—Curves that show gain versus phase angle for two-element arrays with 1/8-wavelength spacing.

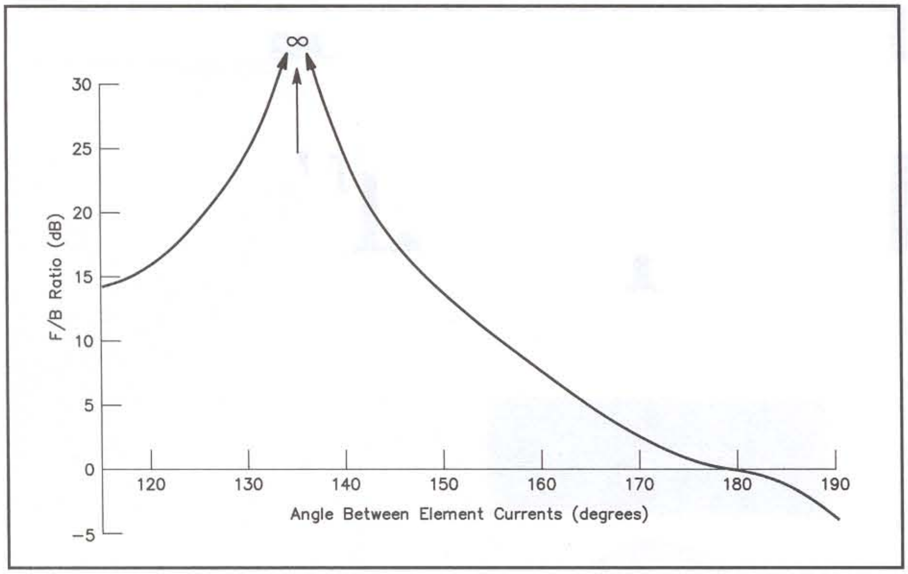

Figure 2—Theoretical F/B ratio for a two-element array with 1/8-wavelength spacing.

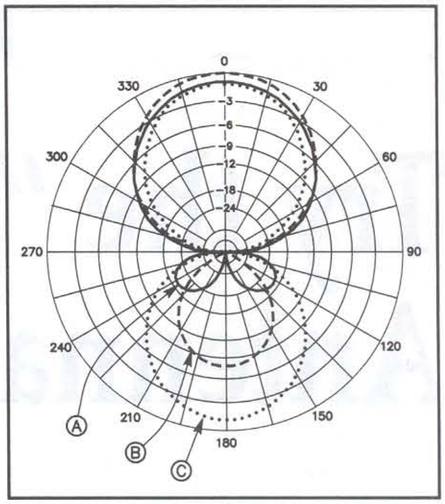

Figure 3—Dipole array patterns for 135, 160 and 180-degree relative phasing at $\lambda/8$ spacing. Curves A, B and C, respectively, represent these conditions. Add 5 dB for dBd. These curves are based on the array being fed with equal currents.

impedance change is a function of not only "mutual impedance," but also the relative magnitudes and phases of the currents flowing in the elements. In an array as closely spaced as these, coupling is so intimate that it could be argued that the term "driven array" is a misnomer. For example, the feed-point impedances of the elements in a 1/8-wavelength-spaced array, assuming equal currents can be made to flow, are

Phase Angle Between Currents (Degrees)	*Loading Element (Ohms)*	*Lagging Element (Ohms)*
135	28 – j46	28 + j46
160	13 – j22	13 + j22
180	9 + j0	9 + j0

This shows quite a change from the 74 + j0 ohms value that each element exhibits when it is not coupled to another element. The fact that the resistive parts of the two-element impedances are equal, and the reactances are equal in magnitude, is a peculiarity of the particular element spacing chosen. For other spacings they will be unequal, and the reactances can be different in magnitude, as well as in sign.

This mutual coupling isn't undesirable; in fact, it's essential for obtaining gain in the presence of rather severe pattern cancellation that is common in these closely spaced arrays. The lower impedances cause more element current to flow for a given power input, thereby increasing the fields from the elements. In these arrays, the increased field strength is sufficient to compensate for the fact that the fields from the elements don't add in phase in any direction. They partially or completely cancel instead. But, the lower feed-point impedances make them more sensitive to losses, and the low resistance with relatively high reactance makes them tricky to feed properly.

Why do these different and reactive feed impedances make feeding the arrays so difficult? The first problem is that, with few exceptions, the magnitude of current out of a line not terminated in its characteristic impedance won't be equal to the current into the line. In classic "If you can't fix it, feature it!" fashion, this impedance-transforming property is put to good use in the form of the 1/4-wavelength Q section.[3] The second (and almost always overlooked) difficulty is that, again with only a few exceptions, the phase delay of current in an imperfectly terminated transmission line doesn't equal the electrical length of the line. This effect isn't minor: The phasing of a casually designed array can easily be off by tens of degrees. In one design I investigated, an 80-degree line produced 139 degrees of phase shift.

The Solution

There are a number of approaches toward correct feeding of an array. My choice was to investigate some simple feed systems to see if any would yield results that came close to the desired characteristics. I wrote a computer program that would solve, iteratively, for element-current magnitude and phase angle, plus feed-point impedances for this one type of array, given the array specifics. Several configurations looked promising, and one of the simplest proved adequate. This was an array of two folded dipoles that were self-resonant, spaced 1/8 wavelength apart and connected by a taut piece of 300-ohm TV ribbon with one half twist. The feed impedance was close to 50 ohms resistive. There was some inductive reactance that could be corrected by adding two small-value capacitors at the feed point. The element current ratio was 1.13:1, with element phasing that was 154 degrees. This was not the 124 degrees one might expect from the 56 electrical degrees of line — assuming a velocity factor of 0.8 — minus the 180 degrees caused by the half twist.

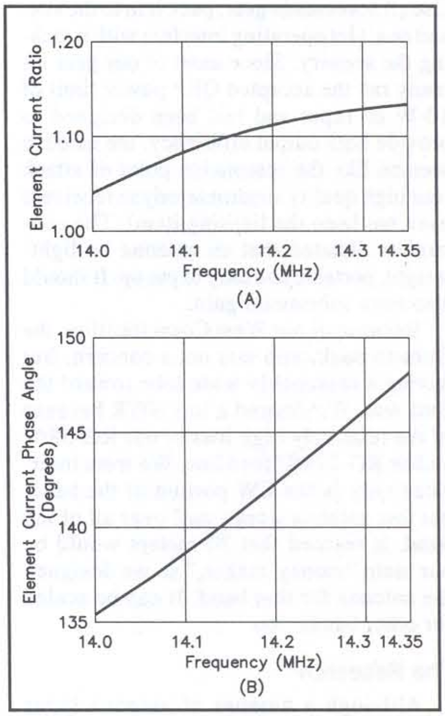

Figure 4—Element current ratio (A) and phase angle (B) as a function of frequency.

It was this array that we used for Field Day, with very good results (see section on performance). However, when the array was reconstructed at the home QTH, a dramatic rise in SWR was noticed when operating the antenna at other than the low end of the band. Computer analysis showed that, above the design frequency, the phase angle increased. This caused a substantial lowering of the element feedpoint impedance, plus narrowing of the forward lobe. The analysis also showed the antenna to be well-behaved *below* the design frequency. Consequently, a similar array was designed (figuratively speaking) for 14.5 MHz. It gave good results over the 20-meter band.

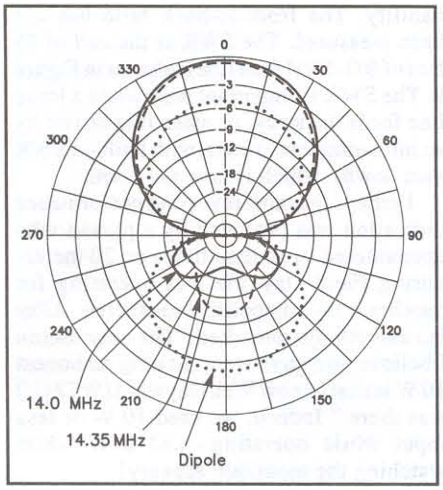

Figure 5—Calculated antenna patterns for the high and low ends of the 20-meter band. Add 5 dB for dBd.

Element phasing varies from 135 degrees at 14.0 MHz to 148 degrees at 14.35 MHz, with current ratios from 1.04 to 1.13:1 (see Figure 4). The gain can be calculated as fairly constant from 4.5 to 4.6 dBd across the band. Again, the array feed-point impedance can be corrected easily to provide a low SWR. The calculated patterns for the antenna at the top and bottom ends of the 20-meter band are shown in Figure 5. These take into account the changes in element phasing, spacing, current magnitude and element self-impedance with frequency.

Construction

The antenna is made from quality 300-ohm TV line to the dimensions given in Figure 6. Sketches of the insulators are provided in Figure 7. They are made from scrap pieces of epoxy-glass PC-board material. This results in ruggedness and minimum weight. The spreaders are readily available 10-foot lengths of "1 inch" ($1^5/_{16}$ in OD)[4] schedule 40 PVC pipe. The capacitors are used only to provide a good match to 50-ohm feed line: They don't otherwise affect the performance of the array. Small 500-V mica or monolithic ceramic units may be used for power levels up to a few hundred watts, since they are at a relatively low-voltage part of the system. Open-ended stubs could probably be substituted for the capacitors, if desired. I recommend that a balun transformer be used with this antenna. Attempts to measure the impedance of one element of this array resulted in a unique experience — the first substantial evidence of the need to use a balun transformer. The antenna-bridge readings varied greatly as the measuring equipment was moved, or as I placed my hand around the feed line. This ceased when I added a balun transformer. The phenomenon is explained by Maxwell in a recent paper.[5] Nearly any style of balun transformer will prevent the unwanted flow of current on the coaxial cable outer conductor. I use a choke type of balun transformer. It consists of 10 turns of small-diameter coaxial cable

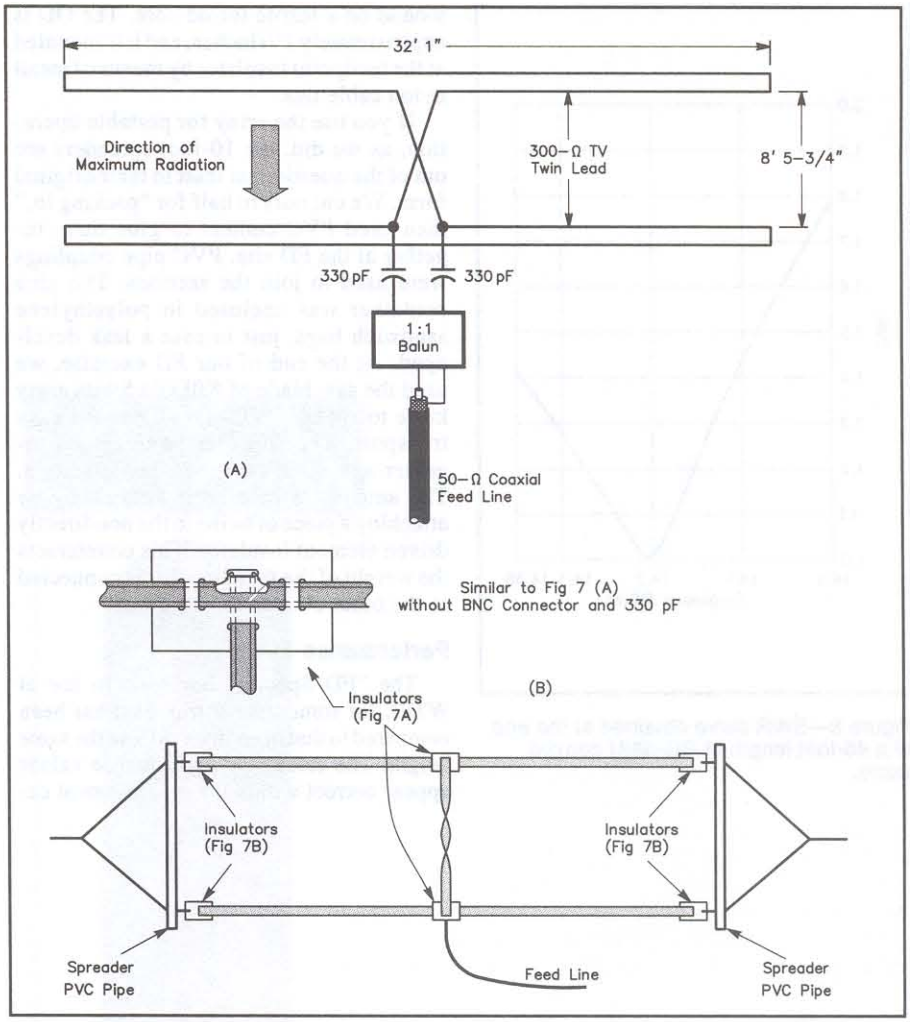
32' 1"
Direction of Maximum Radiation
300-Ω TV Twin Lead
8' 5-3/4"
330 pF
330 pF
1:1 Balun
50-Ω Coaxial Feed Line
(A)
Similar to Fig 7 (A) without BNC Connector and 330 pF
Insulators (Fig 7A)
(B)
Insulators (Fig 7B)
Insulators (Fig 7B)
Spreader PVC Pipe
Feed Line
Spreader PVC Pipe

Figure 6—Electrical dimensions for the W7EL array (A). Illustration B shows how the antenna is assembled on spreaders of PVC pipe.

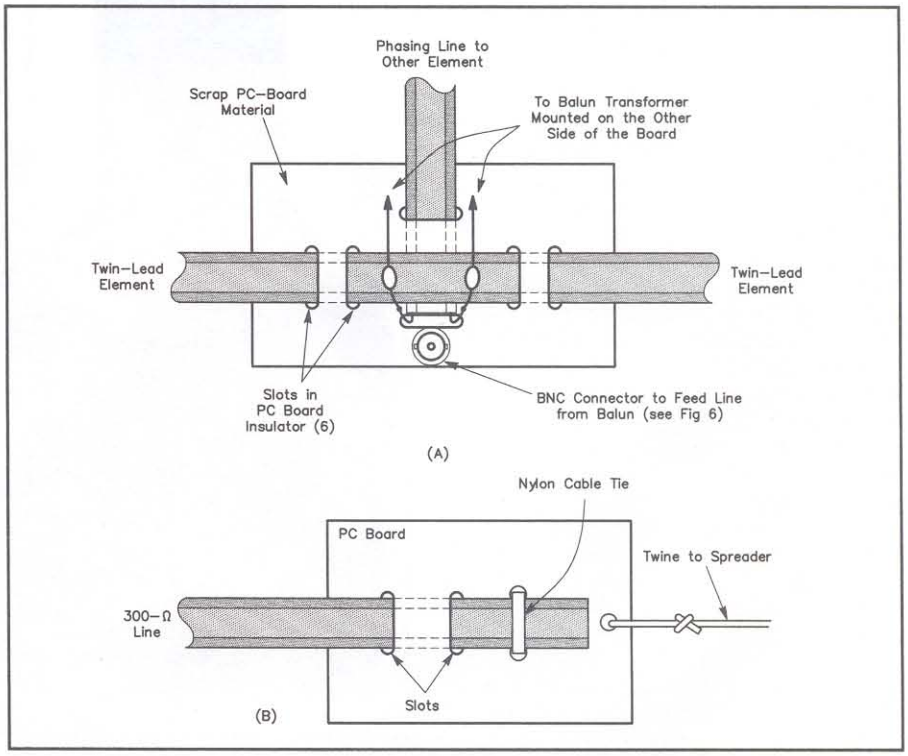
Phasing Line to Other Element
Scrap PC-Board Material
To Balun Transformer Mounted on the Other Side of the Board
Twin-Lead Element
Twin-Lead Element
Slots in PC Board Insulator (6)
BNC Connector to Feed Line from Balun (see Fig 6)
(A)
Nylon Cable Tie
PC Board
Twine to Spreader
300-Ω Line
Slots
(B)

Figure 7—Details for the antenna insulators used in the two-element 20-meter array. The drawing at A shows how the TV ribbon is affixed to the feed-point/phasing-line insulating block. The example at B provides details for the end insulators.

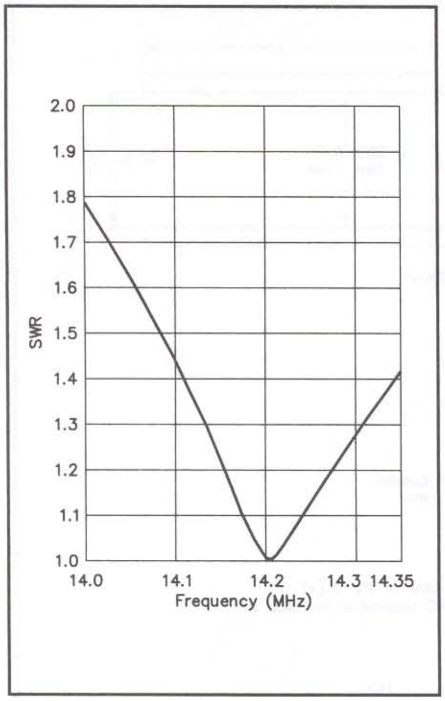

Figure 8—SWR curve obtained at the end of a 45-foot length of RG-58/U coaxial cable.

wound on a ferrite toroid core. The OD is approximately 1¼ inches, and it is mounted at the feedpoint insulator by means of small nylon cable ties.

If you use the array for portable operation, as we did, the 10-foot spreaders are out of the question, at least in their original form. We cut ours in half for "packing in," then used PVC cement to glue them together at the FD site. PVC pipe couplings were used to join the sections. The glue container was enclosed in polyethylene sandwich bags, just in case a leak developed. At the end of our FD exercise, we used the saw blade of K0ED's Swiss army knife to cut the PVC pipes again for easy transport. The spreaders were glued together again for use at the home station. The antenna is held horizontal easily by attaching a piece of twine to the nondirectly driven element insulator. This counteracts the weight of the feed line that is connected to the other element.

Performance

The "FD Special" has been in use at W7EL for some time. Array gain has been compared to that of an inverted V at the same height. The calculated performance values appear correct within the measurement capability. The front-to-back ratio has not been measured. The SWR at the end of 45 feet of RG-58/U feed line is shown in Figure 8. The SWR is important only when a lossy line feeds the array, or when it is driven by an intolerant transmitter (with built-in SWR shut down), which is now the norm.

Perhaps the most revealing performance indication was provided by a person who encountered us several times on 20 meters during Field Day. He was operating for another, very competitive local club. After the exercise he remarked, "The only reason I believe you guys were running an honest 10 W is that I know Wes Hayward (W7ZOI) was there." Indeed, we used 10 W or less input while operating—and 0 W while watching the mountain scenery!

Notes

[1] My apologies to the first person who described or named this antenna. I don't know its history. (See L.A. Moxon, "Two-Element Driven Arrays," *QST*, July 1952, p 28.)

[2] Named after W8JK, Dr. John Kraus, "Antenna Arrays with Closely-Spaced Elements," *Proc. IRE*, Feb. 1940.

[3] G. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1982).

[4] mm = in x 25.4; m = ft x 0.0348.

[5] W. Maxwell, "Some Aspects of the Balun Problem," *QST*, March 1983.

By Riki Kline, 4X4NJ

From *QST*, February 1985

Build a 4X Array for 160 Meters

Low-angle radiation and electrical rotation of directivity are the features of this vertically polarized top-band antenna. If you are interested in 160-meter DX, this system could be your secret weapon.

Is your lament a common one — no room for an effective DX antenna on 160 meters? This complaint is voiced frequently by amateurs who live in urban areas, or who are programmed toward horizontal wire antennas. But, a number of successful top-band operators have adopted the philosophy, "If you can't go out, go up!" It is no secret that a physically short vertical antenna is generally more effective than a horizontal antenna that is close to the ground electrically, at least for DX work.

My 4X array is electrically rotatable. It is compact and is effective as a low-angle radiator. Let's examine how my antenna evolved from some basic designs. I will also cover the practical details of construction and system performance.

The Tilted Ground-Plane Look

The tilted ground plane is almost identical to the usual vertical. The physical format of this antenna resembles a four-conductor ground plane. The major difference is that the radiating elements tilt up toward the supporting structure. The 4X array contains four sloping ground planes. Each of the slope wires is 100 feet long.[1] They are supported at the high end by an 80-foot tower. A four-element, 20-meter Yagi antenna is atop the tower.

Each of the sloping wires is fed separately near ground by means of a tapped-coil matching device (Figure 1) that is returned to radial wires and ground rods. In effect, each radiator is a ground-plane vertical antenna that is slightly less than 0.25 wavelength. The matching inductor provides resonance and effects an impedance match to the coaxial feed line.

I believe that the metal tower and 20-meter antenna may possibly be functioning as a reflector because the Yagi antenna and tower combined with ground wires are resonant slightly below 1.8 MHz.

Two Tilted Ground-Plane Verticals in Phase

I had excellent results with one sloping vertical. Next, I installed a second system in the opposite direction. Switching between the two antennas (north-south sloping radiators) showed considerable front-to-back ratio (a relative reading of 15-20 dB). Subsequently, I connected the two antennas in phase. This gave a bidirectional pattern, east and west. Although I did not gather extensive data on the performance, I observed a 6-dB signal improvement with stations about 700 miles to the

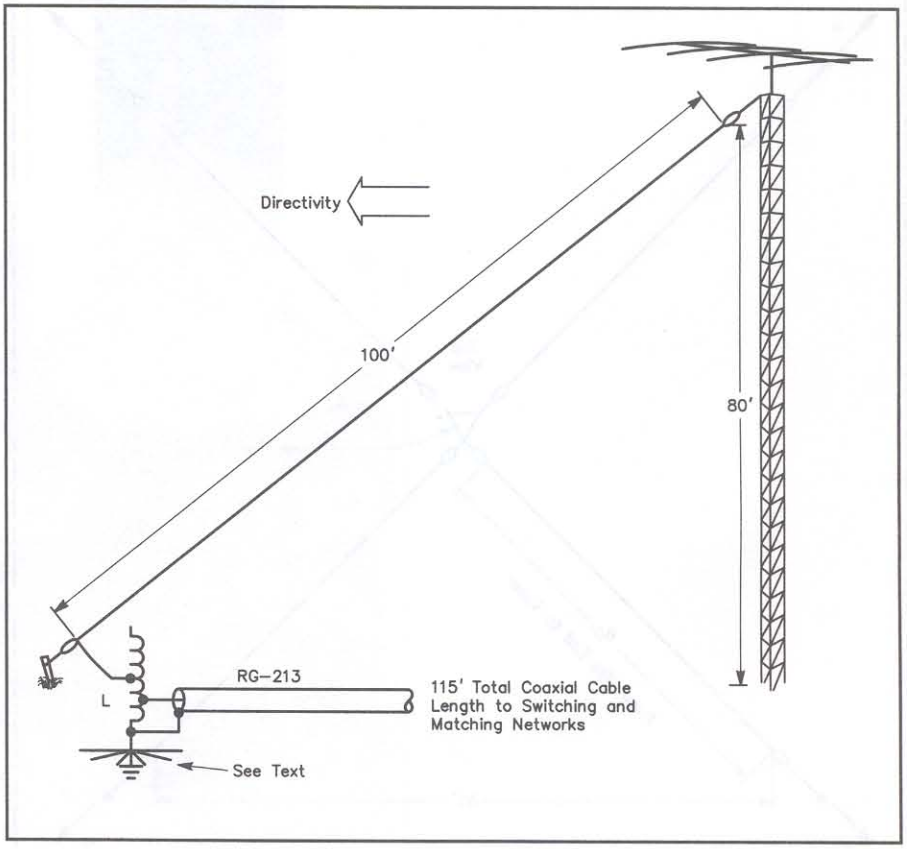
Directivity
100′
80′
L
RG–213
115′ Total Coaxial Cable Length to Switching and Matching Networks
See Text

Figure 1—Basic tilted ground-plane vertical. L has 25 turns of heavy conductor (see text); length is $5\frac{1}{2}$ inches, and diameter is 3 inches. Radiator is tapped $15\frac{1}{2}$ turns above ground.

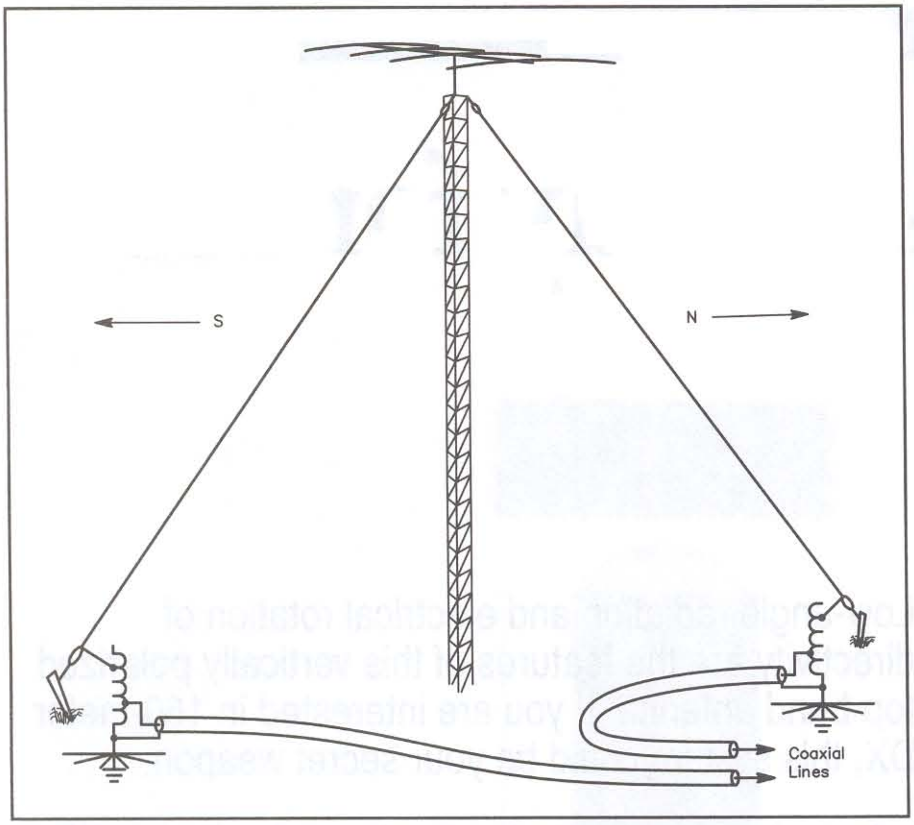

Figure 2—Two tilted ground-plane antennas that can be fed separately or in phase. The feed method is shown in Figure 1, with the method of Figure 7 used for feeding the antennas in phase.

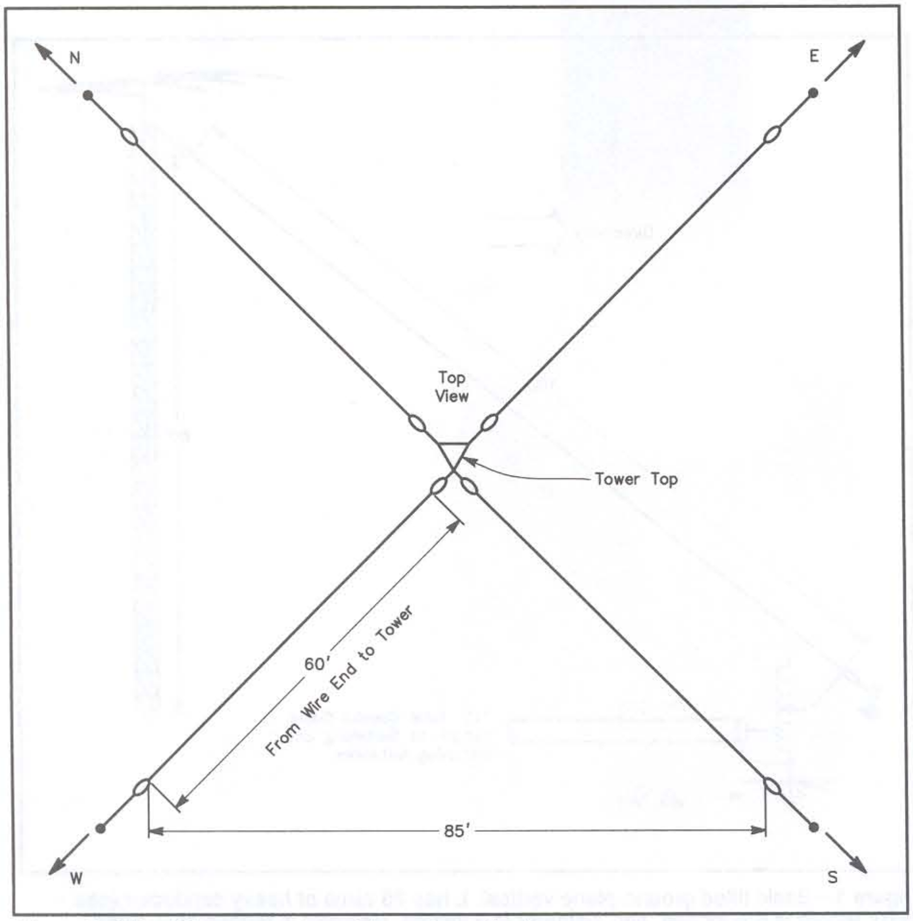

Figure 3—The 4X array as viewed from above the tower.

east. Some of you may want to explore the possibilities further. Figure 2 shows the details of the two-element phased system.

The 4X Configuration

Two more tilted ground-plane verticals were added, thereby providing east-west sloping radiators (Figure 3). A switching and phasing arrangement was added to my 4X array. It allows me to feed any of the slope wires separately, adjacent pairs in phase, or all four wires in phase. When using adjacent pairs in phase, maximum radiation is along a line that bisects the angle between the two antennas (NE, SE, SW or NW directions). When I feed all four wires in phase I note that the radiation is essentially omnidirectional. All of the unfed radiators are resonated to serve as reflectors. This concept is described in *The ARRL Antenna Book*.[2] My switching network is shown in Figure 4.

Phasing Networks

Most phasing methods call for long lengths of non-50-ohm coaxial cable.[3] I found this economically prohibitive. This negative factor inspired the approach I am using.

Each of my radiators is fed by means of 115 feet of RG-213 coaxial cable (formerly RG-8A/U 50-ohm line). The coil at the base of each wire is adjusted for the same resonance and SWR as the remaining three coils. The verticals to be fed in phase have their transmission lines connected in parallel through a suitable network for changing the reflected impedance back to 50 ohms. My networks are shown in Figures 7 and 8.

Tapped-Coil Matching

By using inductance and no intentional parallel capacitance for my matching coils, I am able to obtain greater effective antenna bandwidth because of reduced Q. Stray capacitance and antenna capacitance to the tower and ground are present, however. All electrical connections are soldered. A simple rain cover is used over each coil to protect it from moisture and dirt. The absence of switches, variable capacitors and rotary inductors enables construction of a highly reliable matching system without the need for weatherproof boxes.

This system is relatively easy to tune to obtain nearly identical performance from each antenna branch. This becomes a necessity when using "brute-force" parallel feed in the phased-pair and omnidirectional modes. Otherwise, the power distribution and phasing would be disturbed. This would distort the radiation pattern. Large-diameter, heavy-conductor, air-wound coils are best for this job. Two of my coils are made from silver-plated 1/4-inch-diameter copper tubing. The two remaining coils are made from large, flat conductor material of the kind found in some rotary inductors.

I used a dip meter to adjust the coils for resonance (coaxial cables disconnected). My coaxial cables were tapped initially one

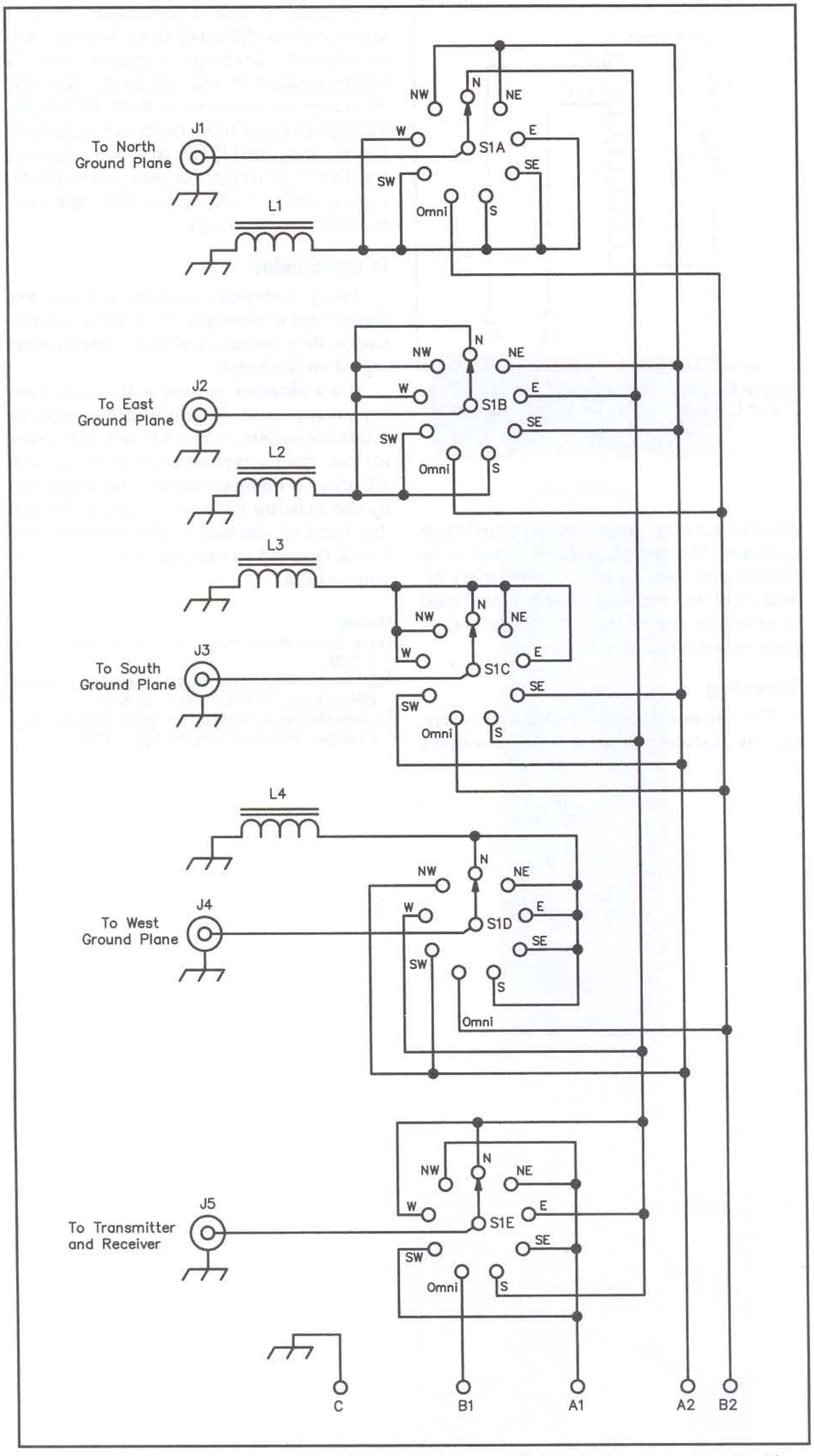

Figure 4—Switching system for the 4X array. L1 through L4, inclusive, are described in the text. A1, A2 and C are connected to the matching circuit shown in Figure 7. B1, B2 and C are connected to the matching circuit shown in Figure 8. S1 is a heavy-duty ceramic rotary switch, five poles, nine positions. J1-J5, inclusive, are coaxial connectors of the builder's choice.

third of the way up from the ground ends of the coils. Final tap placement is made while feeding power to the antenna and observing an SWR meter. Alligator clips make this an easy matter to accomplish. When the SWR bottoms out at the same frequency for all four radiators, remove the alligator clips and solder the coil taps in place. Some interaction between the four antennas will occur, so make certain that all of the taps are where they belong before soldering them.

Direction	SWR vs. Frequency				
	1800 kHz	1825 kHz	1850 kHz	1875 kHz	1900 kHz
N	1.4	1.0	1.2	1.8	2.6
NE	1.5	1.1	1.0	1.1	1.2
E	2.2	1.1	1.1	1.4	1.9
SE	1.2	1.0	1.0	1.0	1.1
S	1.4	1.0	1.1	1.4	2.3
SW	1.5	1.1	1.0	1.0	1.1
W	1.4	1.1	1.0	1.2	1.6
NW	1.2	1.0	1.0	1.0	1.2
Omni	1.4	1.1	1.0	1.0	1.0

Figure 5—Chart that shows SWR versus frequency in kilohertz.

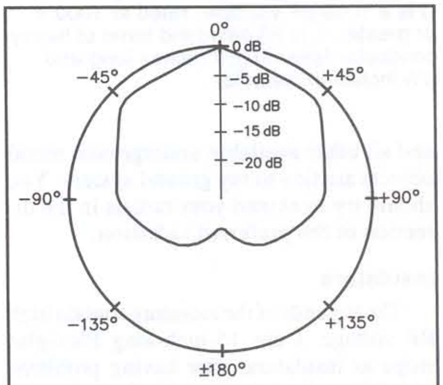

Tabulated Data

Direction	Relative Strength
Forward	0 dB
±45°	−3 dB
±90°	−8 dB
±135°	−14 dB
±180°	−19 dB
Omni	−7 dB

Figure 6—Directivity pattern of the array. The pattern is a composite average of measurements made while receiving 10 different stations. There is no apparent difference between the two directive modes—single radiator or phased pair.

Reflector Tuning

The radiators not being fed are used as reflectors. This is done by switching small inductors in parallel with the ends of the coaxial feed lines. My inductors contain three or four turns of no. 16 wire wound around the center part of a 3/8-inch-diameter ferrite rod from a built-in AM broadcast receiver antenna. The coils are adjusted to give a resonance that is four percent lower than the resonant frequency of the radiators.

Ground Conditions

The efficiency and performance of the antenna depends on the quality of the ground system. Note 3 provides a good reference for ground systems, and a bibliography. Each of my radiators is worked against a counterpoise that contains two or three 1/4-wavelength wires, bent to fit in the boundaries of my property. In addition, I use a 10-foot rod in the ground. Water pipes

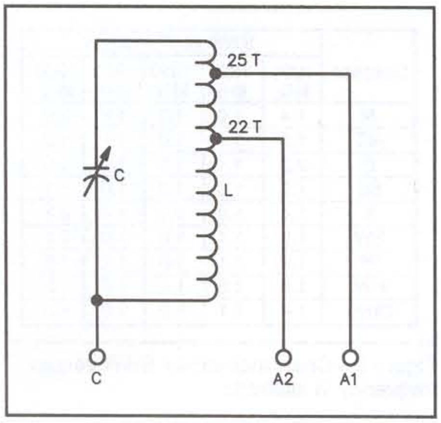

Figure 7—Phased-pair matching network. C is a 1000-pF variable, rated at 1000 V or greater. L is 30 airwound turns of heavy conductor (see text), 7 inches long and 3½ inches in diameter.

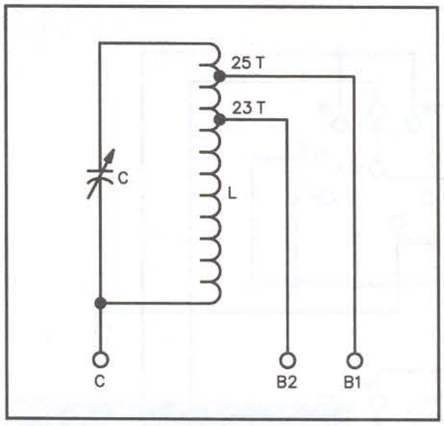

Figure 8—Network for omnidirectional use. C and L are the same as for Figure 7.

and all other available underground metal objects are tied to my ground system. You should try to extend your radials in the direction of the preferred radiation.

Insulators

The top ends of the radiators contain high RF voltage. I use 15-inch-long Plexiglas strips as insulators, after having problems with 8-inch-long commercial plastic insulators. Moisture and air pollutants caused these problems. The present insulators need to be cleaned periodically. High-quality glass insulators of the type used aboard ships should be excellent and, with luck, should not require periodic cleaning.

Receiving

The antenna directivity enhances reception by rejecting signals from unwanted directions. I was encouraged when I compared my 4X array to an 800-foot unterminated Beverage antenna that is bidirectional NW and SE. In the past, the Beverage wire showed an 8-10 dB S/N advantage over the tilted north vertical alone. During long-haul QSOs to North America, the 4X array, used in the phased-pair mode, comes within 3 dB of the S/N ratio provided by the Beverage.

In Conclusion

Many European stations tell me my signals are as strong as local ones. On occasion they remark that I have the loudest signal on the band.

It's a pleasure to have a directional array on top band. I simply turn a knob to rotate the pattern — much faster than a motor can rotate a typical beam antenna! The 4X array is compact and can be supported by the existing HF-antenna tower. I hope that some of you will try this antenna, and I look forward to hearing from you about your results.

Notes

[1] m = ft × 0.3048; mm = in × 25.4; km = mi × 1.609.

[2] G. Hall, ed., *The ARRL Antenna Book* (Newington: ARRL, 1984), p. 8-12.

[3] J. Devoldere, *80-Meter DXing* (Greenville, NH: Communications Technology, 1978).

By Rod Newkirk, W9BRD **From *QST*, June 1990**

The 'BRD Zapper: A Quick, Cheap and Easy "ZL Special" Antenna

Here's a multipurpose directional wire antenna that has its origins in Kraus, Windom and Newkirk *(who?–Ed.)*.

Working Europe from Chicago with low power on 21 MHz, using an indoor, bent-ends W8JK bidirectional wire beam (Figure 1)[1], is fairly straightforward fun. Voltage feed via a λ/4 stub, gamma-matched to coax, is hard to beat for spartan simplicity. No worry about balancing element currents, no nit-picking with element lengths. Just make the whole system symmetrical and dip it to your favorite frequency by adjusting the stub-shorting point.[2] The 8JK has a similar pattern and good gain all the way from the fundamental to the second harmonic, so it's really a multiband antenna. You can roll the whole thing up in three minutes when it's not in use.

One problem: The bidirectional characteristics of the 8JK cause my ears to be flattened regularly by undesired signals from the direction opposite Europe. Unless you're in the geographical center of a three-way QSO, the unused lobe of this antenna can be a nuisance. Question: Is it feasible to convert easy 180° bidirectionality to tricky 135° *uni*directionality without resorting to clumsy center-feeding, multi-wire-dipole elements, balanced-gammas, etc?[3]

Yes—because there's a *uni*directional ZL Special lurking in our little 8JK. The principal difference between the W8JK and the ZL Special lies in the phasing: In the 8JK, the elements are fed 180° out of phase; in the ZL Special, they're driven 135° out of phase. Therefore, instead of feeding the antenna near the stub's shorting point, we'll need to feed it λ/16 from the stub's shorted end. That's where the feed path to one element is λ/8 (45°) longer or shorter than the other, which, after the stub's 180° phase reversal, produces the 135° phase shift that were looking for between the elements.

You can find the proper feed point on the stub by "sniffing" signals of known origin along *one side* of the stub with the insulated center conductor of some coax hooked to a receiver. (Start looking for this point by measuring λ/16 up from the bottom of the stub.) It's refreshing to hear Europeans rolling in while most of the murderous rearward signals now barely budge the S meter! Directivity is reversed at the opposite stub point (Figure 2).

So he's in there all right, that ZL Special, but coaxing the wily rascal out for an honest day's work is a challenge. The ZL

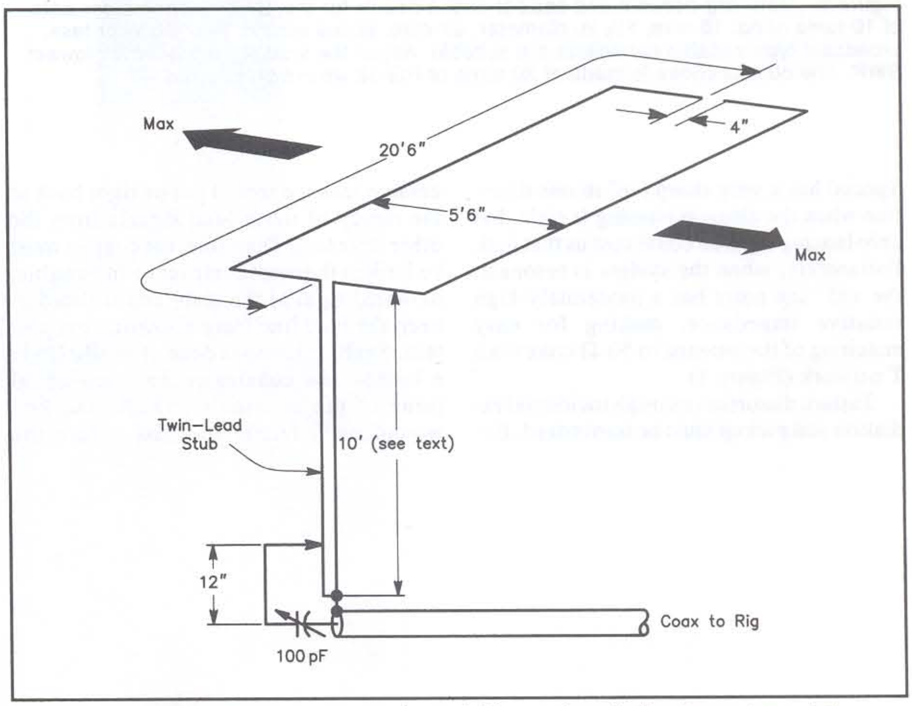

Figure 1—W9BRD's W8JK wire beam for 21 MHz, made to fit the dimensions of the townhouse bedroom that it occupies. Its dimensions aren't critical, although the element lengths should be close to λ/2 (total, each), the stub should be λ/4 or odd multiples, and the element spacing should be close to λ/8. At the bottom of the stub, connect both stub wires and the coax braid together, after resonating the system as outlined in Note 2. The 12-inch tap distance and 100-pF variable capacitor constitute a gamma match that brings the stub impedance to 50 Ω.

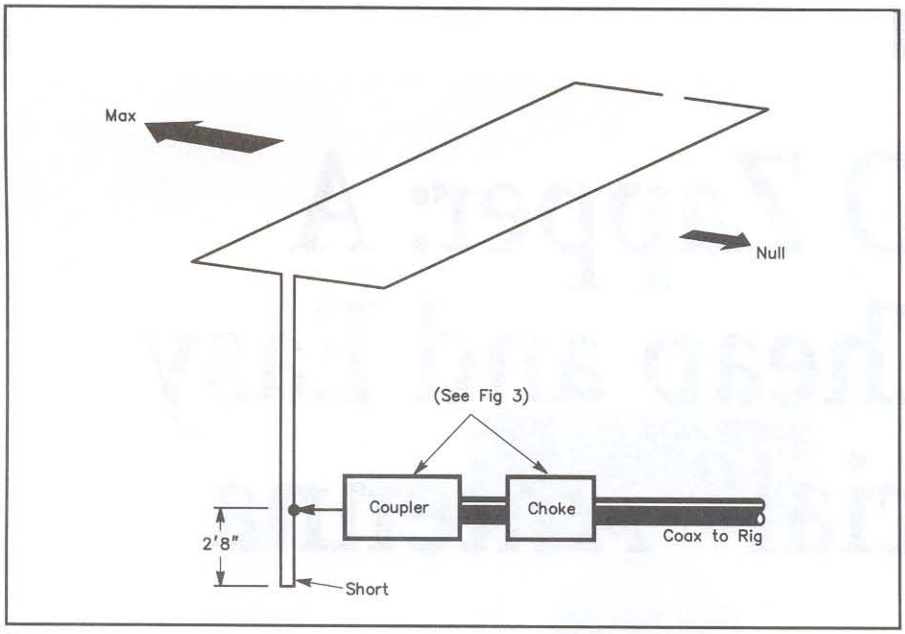

Figure 2—The antenna of Figure 1 configured for unidirectional operation with 135° element phasing. Resonate the system as described in Note 2 and connect the coupler via a very short wire to one leg of the stub about λ/16 above the stub-shorting point. The coaxial choke is needed to eliminate feed-line radiation, which can ruin the rearward null that this system provides.

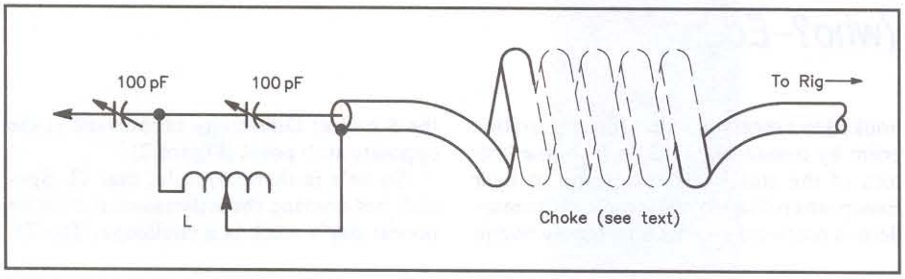

Figure 3—Matching-network and coaxial-choke details for the 'BRD Zapper. L consists of 10 turns of no. 16 wire, $1\frac{1}{2}$ in. diameter, air core, space wound. For 100 W or less, broadcast-type variable capacitors are suitable. Adjust the inductor-tap point for lowest SWR. The coaxial choke is made of 30 turns of RG-58 wound on a ferrite rod.

Special has a very sharp null in one direction when the element phasing is right, but unbalancing the stub could cost us that null. Fortunately, when the system is resonant, the 135° tap point has a moderately high resistive impedance, making for easy matching of the antenna to 50-Ω coax via a T network (Figure 3).

Pattern distortion through incidental radiation and pickup must be minimized. Excessive leakage would put us right back at the mercy of those loud signals from the other direction. Therefore, the coupler must be built in the most compact form possible, mounted right at the stub, and isolated to keep the feed line from distorting the pattern. Such isolation is done at W9BRD via a home-brew coaxial choke made of 30 turns of the antenna's RG-58 feed line wound on a ferrite rod just before the matching network.

Because feeder radiation could upset the stub balance—and thus spoil the antenna's rearward null—the single-wire run from the matching network to the antenna had better be no more than an inch or two. Here, I'm borrowing on the single-wire-feed theme by Windom. Any circuitry above the choke will be hot with RF, so a bulky meters-bells-and-whistles matching unit will not do. It's best to use a compact, barebones, dedicated network. Run the stub all the way down from the flat top if possible. Incidentally, the tap point on the stub corresponding to the best rearward null is higher on the stub for element spacings wider than λ/8. System Q is also lower with wider spacings, but gain is maximum at λ/8.

As for the stub itself, TV twin lead is okay indoors, but close-spaced, end-fire elements mean high voltages and currents. If the stub is to be longer than a quarter wavelength (odd multiples only) or outside in the weather, low-loss, open-wire line is a must. You can run the stub of an outdoor version directly into the shack for handy directivity reversal.

Though we've been talking horizontal, this setup should be equally interesting as a vertical. By the way, when the elements are placed horizontally, one above the other, 135° phasing accentuates radiation and reception at high angles. Neat for local traffic nets, Sweepstakes and Field Day!

Notes

[1] Dr John D. Kraus, W8JK, first described this antenna in *QST* in "Directional Antennas with Closely Spaced Elements," *QST*, Jan 1938, pp 21-23, 37. Kraus also discussed this antenna in three articles in *Radio* magazine in 1937 and 1939, and in *Proceedings of the IRE*, Feb 1940.

[2] The easiest way to do this is with a dip meter. Jab a straight pin into each leg of the stub near the bottom, short the pins together and couple the dip meter Into the system by bringing the dip motor close to the shorted pins. Locate resonance using the dip meter and measure the dip-meter frequency on your receiver. Move the pins up or down the stub and repeat the process until the antenna is resonant at or near your frequency of Interest. When you've done this, cut the stub, strip and solder a short in the wires where the pins were located at resonance.

[3] Of course, this antenna can be center fed in either Its bidirectional or unidirectional modes. Such flat tops are described in J. Hall, Ed., *The ARRL Antenna Book*, 15th ed, (Newington: ARRL, 1988), Chapter 8, and J. Devoldere, *LowBand DXing* (Newington: ARRL, 1987), pp 2-101 through 2-102.

Combination Driven Arrays

Broadside, end-fire and collinear elements can readily be combined to increase gain and directivity, and this is in fact usually done when more than two elements are used in an array. Combinations of this type give more gain, in a given amount of space, than plain arrays of the types just described. The combinations that can be worked out are almost endless, but in this section are described only a few of the simpler types.

The accurate calculation of the power gain of a multi-element array requires a knowledge of the mutual impedances between all elements, as discussed in earlier sections. For approximate purposes it is sufficient to assume that each set (collinear, broadside, end-fire) will have the gains as given earlier, and then simply add up the gains for the combination. This neglects the effects of cross-coupling between sets of elements. However, the array configurations are such that the mutual impedances from cross-coupling should be relatively small, particularly when the spacings are 1/4 λ or more, so the estimated gain should be reasonably close to the actual gain.

FOUR-ELEMENT END-FIRE AND COLLINEAR ARRAY

The array shown in Figure 1 combines collinear in-phase elements with parallel out-of-phase elements to give both broadside and end-fire directivity. It is popularly known as a "two-section W8JK" or "two-section flat-top beam." The approximate gain calculated as described above is 6.2 dB with 1/8-λ spacing and 5.7 dB with 1/4-λ spacing. Directive patterns are given in Figures 2 and 3.

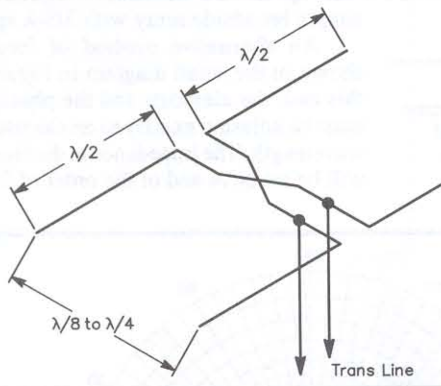

Figure 1—A four-element array combining collinear broadside elements and parallel end-fire elements, popularly known as the W8JK array.

The impedance between elements at the point where the phasing line is connected is of the order of several thousand ohms. The SWR with an unmatched line consequently is quite high, and this system should be constructed with open-wire line (500 or 600 Ω) if the line is to be resonant. With 1/4-λ element spacing the SWR on a 600-Ω line is estimated to be in the vicinity of 3 or 4 to 1.

To use a matched line, a closed stub 3/16 λ long can be connected at the transmission-line junction shown in Figure 1, and the transmission line itself can then be tapped on this matching section at the point resulting in the lowest line SWR. This point can be determined by trial.

This type of antenna can be operated on two bands having a frequency ratio of 2 to 1, if a resonant feed line is used. For example, if designed for 28 MHz with 1/4-λ spacing between elements it can be operated on 14 MHz as a simple end-fire array having 1/8-λ spacing.

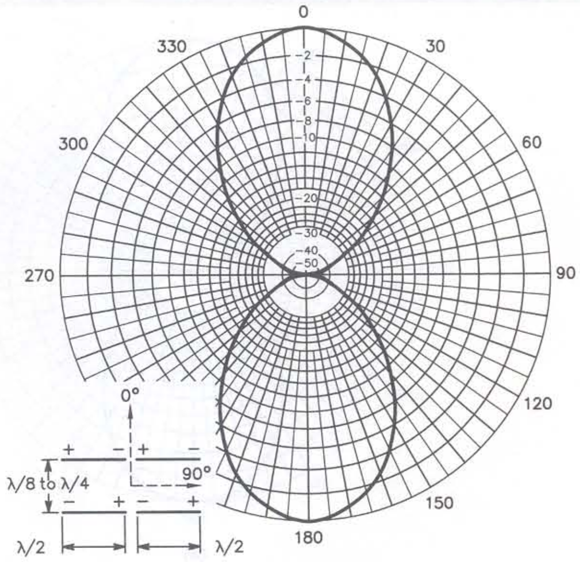

Figure 2—E-plane pattern for the antenna shown in Figure 1. The elements are parallel to the 90°-270° line in this diagram. Less than a 1° change in half-power beamwidth results when the spacing is changed from 1/8 to 1/4 λ.

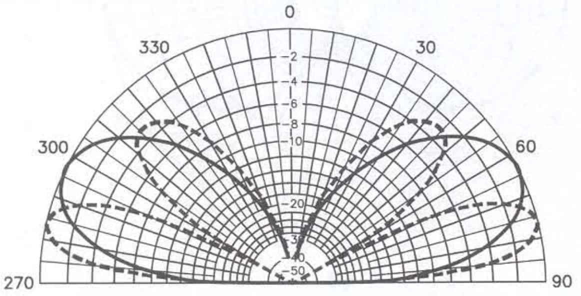

Figure 3—Vertical pattern for the four-element antenna of Figure 1 when mounted horizontally. Solid curve, height 1/2 λ; broken curve, height 1 λ above a perfect conductor. Figure 2 gives the horizontal pattern.

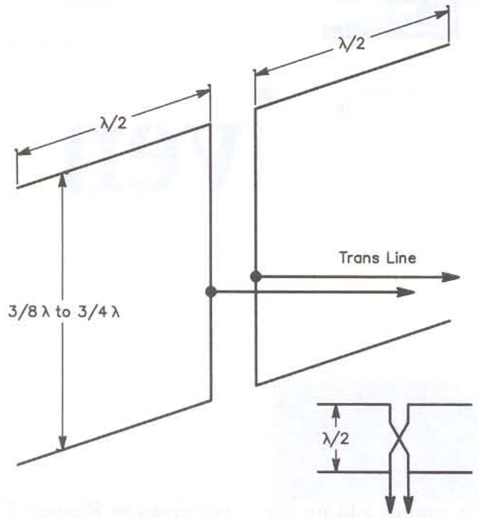

Figure 4—Four-element broadside array ("lazy H") using collinear and parallel elements.

FOUR-ELEMENT BROADSIDE ARRAY

The four-element array shown in Figure 4 is commonly known as the "lazy H." It consists of a set of two collinear elements and a set of two parallel elements, all operated in phase to give broadside directivity. The gain and directivity will depend on the spacing, as in the case of a simple parallel-element broadside array. The spacing may be chosen between the limits shown on the drawing, but spacings below 3/8 λ are not worthwhile because the gain is small. Estimated gains are as follows

3/8-λ spacing—4.4 dB
1/2-λ spacing—5.9 dB
5/8-λ spacing—6.7 dB
3/4-λ spacing—6.6 dB

Half-wave spacing is generally used. Directive patterns for this spacing are given in Figures 5 and 6.

With 1/2-λ spacing between parallel elements, the impedance at the junction of the phasing line and transmission line is resistive and is in the vicinity of 100 Ω. With larger or smaller spacing the impedance at this junction will be reactive as well as resistive. Matching stubs are recommended in cases where a nonresonant line is to be used. They may be calculated and adjusted as described in Chapter 26.

The system shown in Figure 3 may be used on two bands having a 2-to-1 frequency relationship. It should be designed for the higher of the two frequencies, using 3/4-λ spacing between parallel elements. It will then operate on the lower frequency as a simple broadside array with 3/8-λ spacing.

An alternative method of feeding is shown in the small diagram in Figure 3. In this case the elements and the phasing line must be adjusted exactly to an electrical half wavelength. The impedance at the feed point will be resistive and of the order of 2 kΩ.

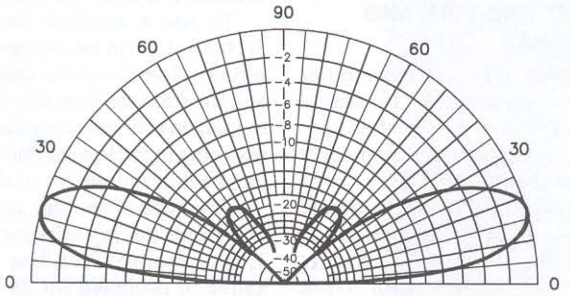

Figure 6—Vertical pattern of the four-element broadside antenna of Figure 3, when mounted with the elements horizontal and the lower set 1/2 λ above a perfect conductor. "Stacked" arrays of this type give best results when the lowest elements are at least 1/2 λ high. The gain is reduced and the wave angle raised if the lowest elements are close to ground.

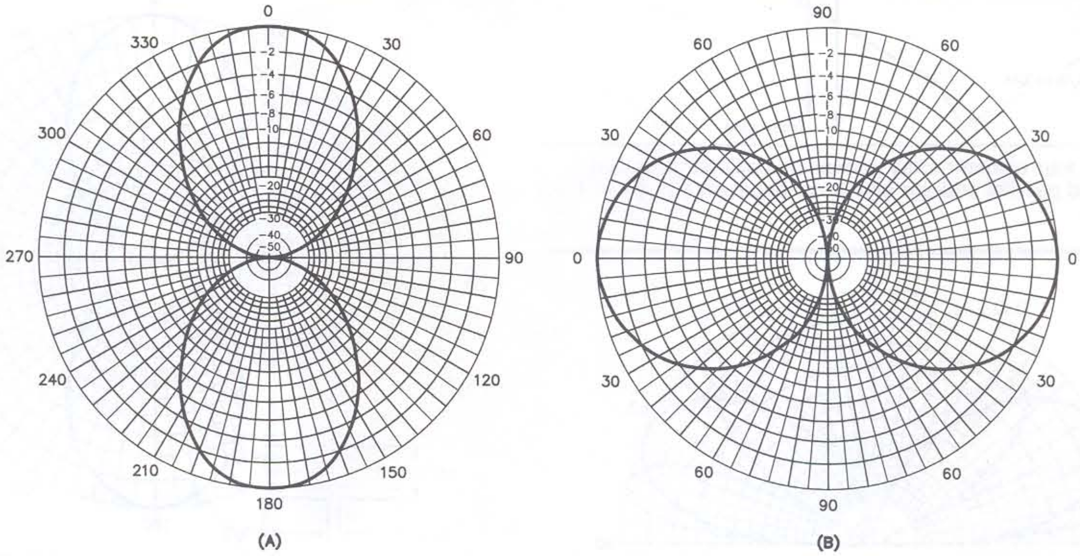

Figure 5—Free-space directive diagrams of the four-element antenna shown in Figure 4. At A is the E-plane pattern, the horizontal directive pattern at low wave angles when the antenna is mounted with the elements horizontal. The axis of the elements lies along the 90°-270° line. At B is the free-space H-plane pattern, viewed as if one set of elements is above the other from the ends of the elements.

CHAPTER SEVEN

VERTICALLY POLARIZED

7-1 **More on "Half Slopers" and Half-Sloper Variation for 160 Meters**
From December 1979 *QST* (Technical Correspondence)

7-3 **A Two-Band Half-Sloper Antenna**
From June 1980 *QST*

7-7 **The KI6O 160-Meter Linear-Loaded Sloper**
From April 1986 *QST*

7-8 **The Super Sloper**
From December 1995 *QST*

7-12 **The Inverted L Revisited**
From January 1983 *QST*

7-14 **160-Meter Antennas**
From July 1991 *QST* (Technical Correspondence)

7-16 **A Truly Broadband Antenna for 80/75 Meters**
From April 1986 *QST*

7-19 **Improved Broadband Antenna Efficiency**
From August 1987 *QST* (Technical Correspondence)

7-20 **Simple, Effective, Elevated Ground-Plane Antennas**
From June 1994 *QST*

From *QST*, December 1979 (Technical Correspondence)

More on "Half Slopers"

The July 1979 *QST* article on quarter-wave slopers has prompted me to write you on the subject. I have had considerable experience using them on 160 and 80, and I have tried them on 40.

The first one I know who used this type antenna was W3AU. His success with the system on 160 caused him to recommend it to W7RM and me. The original configuration used more than one sloping element fed in parallel against the tower. It resembled a top-loaded vertical with a simulated ground system. All my results have supported the view that this is an omnidirectional, vertically polarized radiating system, even when only one sloping element is used.

All the antennas of this type I have tried have been at my father's station, W6UA, and have been on a single tower. The tower is 82 feet (25 m) high, approximately 1½ feet (0.45 m) in uniform triangular cross section, and has a single set of guys at the 70-foot (21.3-m) level. The guys are broken with insulators at about 20-foot intervals, and the top section of each guy is electrically connected to the tower. The base has whatever insulation the concrete provides and there arc 10-, 15- and 20-meter Yagis on top of the tower. (This is an AB105 tower, the same as is used by W3AU and W7RM.)

The first antenna of the type in question was tried on 80 meters. It was attached at the 60-ft level, so that top-loading effects of the guys, as well as the rotaries, were probably effective. The attached wire came off at 30° to vertical and was trimmed for minimum VSWR. The resultant length was within 2 feet of the formula length, and the antenna then had a VSWR less than 1.5 from 3.5 to 4.0 MHz. It worked well on transmit but was extremely noisy on receive. The original wire sloped east, so I put up an identical system, attached to the same tower level, sloping west. Electrical characteristics were nearly identical, and no directivity difference between the two could be detected either on transmit or receive.

This is logical considering how little of the antenna current in the sloping wire has any horizontal component, other antenna systems used on 80 at W6UA have been dipoles, inverted Vs, and a delta loop. The quartcr-wave sloper is better than any of the others for transmitting, but it is poor on receiving. I have taken down all but the two slopers. They are used for transmitting and receiving, with frequent use of a 40-meter collinear antenna on receiving weak signals.

I also tried this system on 40, with a wire attached to the tower at 30 feet, coming off at 30°. Perhaps because there was so much tower electrically above the attachment point – at least half a wavelength – this one was much harder to tune. The wire length, which was very critical, was some feet shorter than expected, and the VSWR behavior across the band less benign. Not much time was spent on this particular antenna because it seemed no better than a λ/2 vertical. Both worked well, but neither was the equal of the collinear types available.

The most recent quarter-wave sloper system I have used has been on 160 meters. My first attempt was a single sloping element from the 75-foot level. This places whatever contribution there is from the top guy sections below the point of attachment. I couldn't get this antenna to match by altering the length of the sloping wire, even while making rather severe adjustments. There did seem to be some promise for the system, so the next season I used two sloping elements of about the "correct" length and fed the arrangement with a matching network at the bottom of the tower, with 300-ohm twin-lead going up the tower (Figure 1). This was the best antenna I have used on 160 at W6UA. Previously I had used an inverted V, a center-fed 80-meter dipole at 90 feet and a λ/2 inverted L. Since the wires must come off the tower at about 45-50° to vertical, the rationale of using two wires instead of one was to attempt to cancel the horizontal current component. As a matter of curiosity I tried adding a third wire and could find no change in setting of the matching components or in the

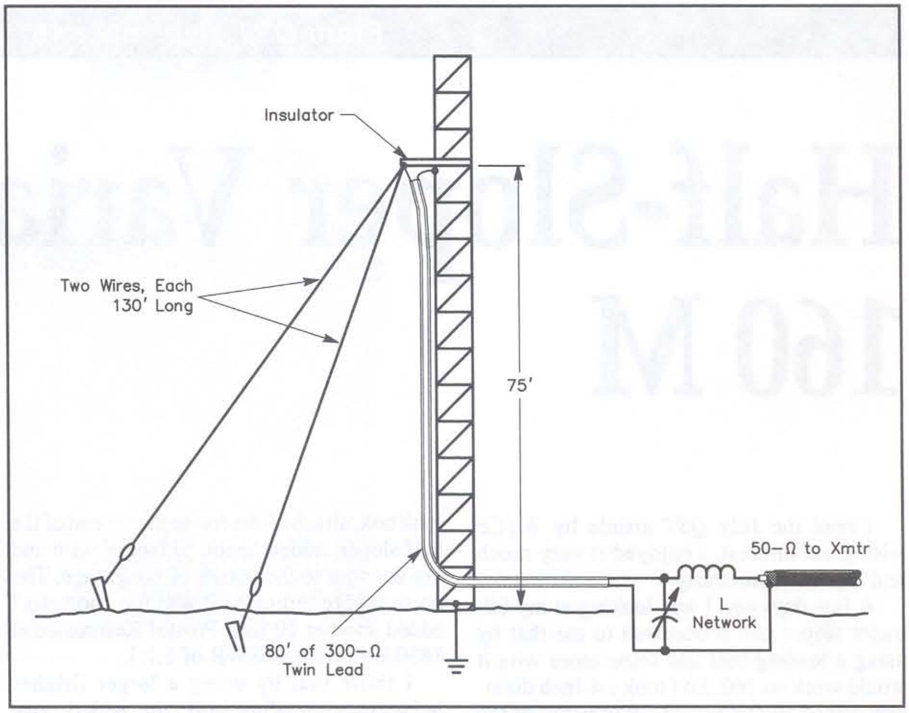

Figure 1—Illustration of the arrangement used at W6UM for a 160-meter half sloper.

overall effectiveness. Using some rough estimates of component values required for matching, length of feed line, etc., I calculate the feed impedance of the antenna itself to be approximately 100–*j*300 ohms.

My estimate of how well this antenna performs must be tempered by the realization that conditions during the past two years have been much worse than the period before. During the 160-meter contests, openings to W1, W2 and W3 have been marginal. I have worked nothing with this antenna I hadn't worked before, but the impression remains that the antenna is better than those used earlier. It has done a consistent job into the Caribbean, SA, Pacific and Japan, as well as across the USA. Receiving remains a problem because of the noise. I have taken down the 160-meter version of this antenna. If I put it up again I will definitely install some auxiliary receiving antennas.

The net result seems to be that this is an effective antenna. Getting a good match directly into coax appears easier if the sloping wire is more nearly vertical and if there is not too much loading above the attachment point. The system should be more efficient, up to a point, as there is more vertical structure above the feed point. There is no compelling reason to insist upon using a resonant configuration. The next attempt I make on 160 might well be with the existing 80-meter system, fed at 60 feet and with 80-meter sloping elements, but using a transmission line and matching network at the bottom of the tower as before.

Given that the antenna does work, the areas I feel need further investigation are better characterization and description of it for different parameters, e.g., measurement of the feed impedance as a function of frequency for different sloper attachment points, lengths, angles and conditions of top loading. I would also like to see the current flow on all of the structure mapped via a probe. If some of the results are forthcoming I hope to see them in an early issue of *QST*. — *Charles Weir Jr., W6UM*

From *QST*, December 1979 (Technical Correspondence)

Half-Sloper Variation For 160 M

I read the July *QST* article by W1CF with great interest. I enjoyed it very much and learned something.

A few days ago I was looking at my 80-meter sloper and it occurred to me that by using a loading coil and some more wire it would work on 160. So I took a 4-inch diameter coil of 10 turns of no. 10 wire out of the junk box, attached the coil to the far end of the half sloper, added about 30 feet of wire and ran the wire to the corner of my garage. The noise bridge indicated it was too short, so I added another 10 feet. Presto! Resonance at 1830 kHz with an SWR of 1.1:1.

I think that by using a larger (higher inductance) loading coil, the added wire could be shortened considerably without seriously affecting the 2-to-1 SWR points on the curve.

I have not made too many contacts on 160 yet, as there is very little activity this early in the season around here. I have had several contacts though, always getting at least as good a report as I have given. *—Philip True, W7AQB.*

By Gary E. Myers, K9CZB **From *QST*, June 1980**

A Two-Band Half-Sloper Antenna

When an off-the-wall empirical design like this works well, suspicion and skepticism are warranted. Maybe you'll agree that this sloper idea is an exception!

The popularity of the half-sloper antenna seems to be increasing, as evidenced by recent articles in *QST*.[1,2,3] This type of antenna has some worthwhile advantages, particularly for the lower frequency operator – low-angle radiation for antennas of modest height, compactness and simplicity of construction. On the minus side, narrow bandwidths and difficulties in resonating the system have been reported.[3]

The antenna system to be described here evolved from a simple-minded attempt to design a two-band half sloper for 80- and 40-meter operation. A trap-type of antenna was selected as the design basis because of previous experience with trap antennas and because the inductive loading of the trap on the lower-frequency band allows a somewhat shorter overall length. The same inductive loading, however, was also expected to increase the antenna Q and thereby further decrease the bandwidth. For this reason I was prepared to experiment.

It is well that I was so prepared because the final form of the antenna bears little resemblance to the initial concept. But most interesting — and exciting — is the impressive bandwidth on 80 meters.

The Design

A diagram of the two-band half-sloper antenna system is shown in Figure 1. It can be seen that there is no obvious quarterwave dimension in the entire system. In fact, the radiator itself is a nonresonant device.

Initial attempts to prune the wires to resonance resulted only in a mound of wire clippings and one frustrated amateur. After many hours of cut-and-try experimentation, accompanied by a growing, gut-level appreciation for what apparently was happening, the magic combination of wire lengths and trap component values was found. Impedances of 40 $-j80$ ohms at 7.2 MHz and 60 $-j40$ ohms at 3.6 MHz were measured with a noise bridge. It was then a simple matter to cancel out these capacitive reactances with an inductor.

For convenience, the inductor was placed in the transmission line, rather than at the feed point – final "tweaking" of the system is more easily performed on the ground than at 40 feet (12.2 m) in the air. A Smith Chart exercise shows that the SWR on the transmission line between the feed point and inductor L is 5:1 at 7.2 MHz and 2:1 at 3.6 MHz. When RG-8/U is used, the additional loss incurred because of these SWRs is less than 0.5 dB at 7.2 MHz and is almost nonexistent at 3.6 MHz.

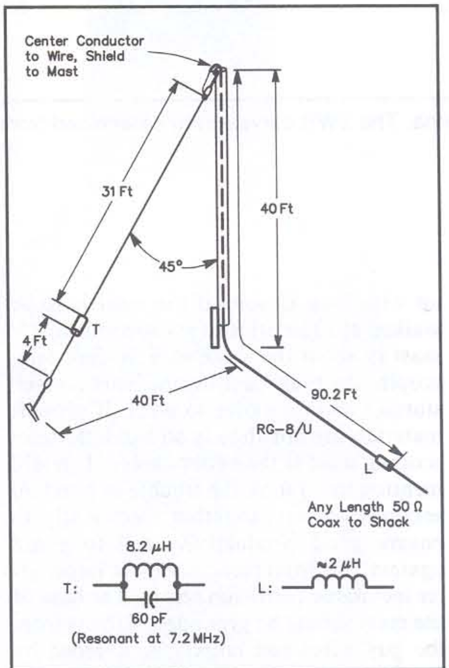

Figure 1—The two-band half sloper. To shield the transmission line from the antenna field, the line is routed up the inside of the support mast, which is grounded. Inductor L is needed to resonate the system.

Performance

This antenna performs very well. Operation at K9CZB is primarily 80-meter cw, with some 40-meter ssb. Power output is nominally 100 watts. Signal reports on 80 have been uniformly good, with comments such as U R LOUDEST 9 ON BAND AND VY FB SIG, VY STRONG. Voice operation on 40 has also resulted in good signal reports, although the praise has not been so lavish. This is my only 80/40-meter antenna, so direct comparisons were not possible. However, it appears to greatly outperform two previous antennas, a 160-foot end-fed wire and a trap dipole, both strung 30 feet above ground. All in all, it is about what I expected from a half sloper.

This kind of performance is nice, but nothing to write an article about, since that has already been done. The real performance story about this antenna can be summed up in one word: bandwidth. A glance at the SWR curves in Figure 2 will open the eyes of any 80-meter operator. As far as I know, this is unheard-of bandwidth for such a simple and compact antenna. It is a real treat to QSY 400 kHz And see the SWR meter needle barely move. This isn't a low-Q antenna —it's a *no*-Q antenna! The bandwidth on 40 is far less impressive and is, in fact, similar to what has been reported previously for half slopers.

Construction

During experiments with prototypes of this antenna, I noticed some sensitivity to feed-line placement and length. Therefore,

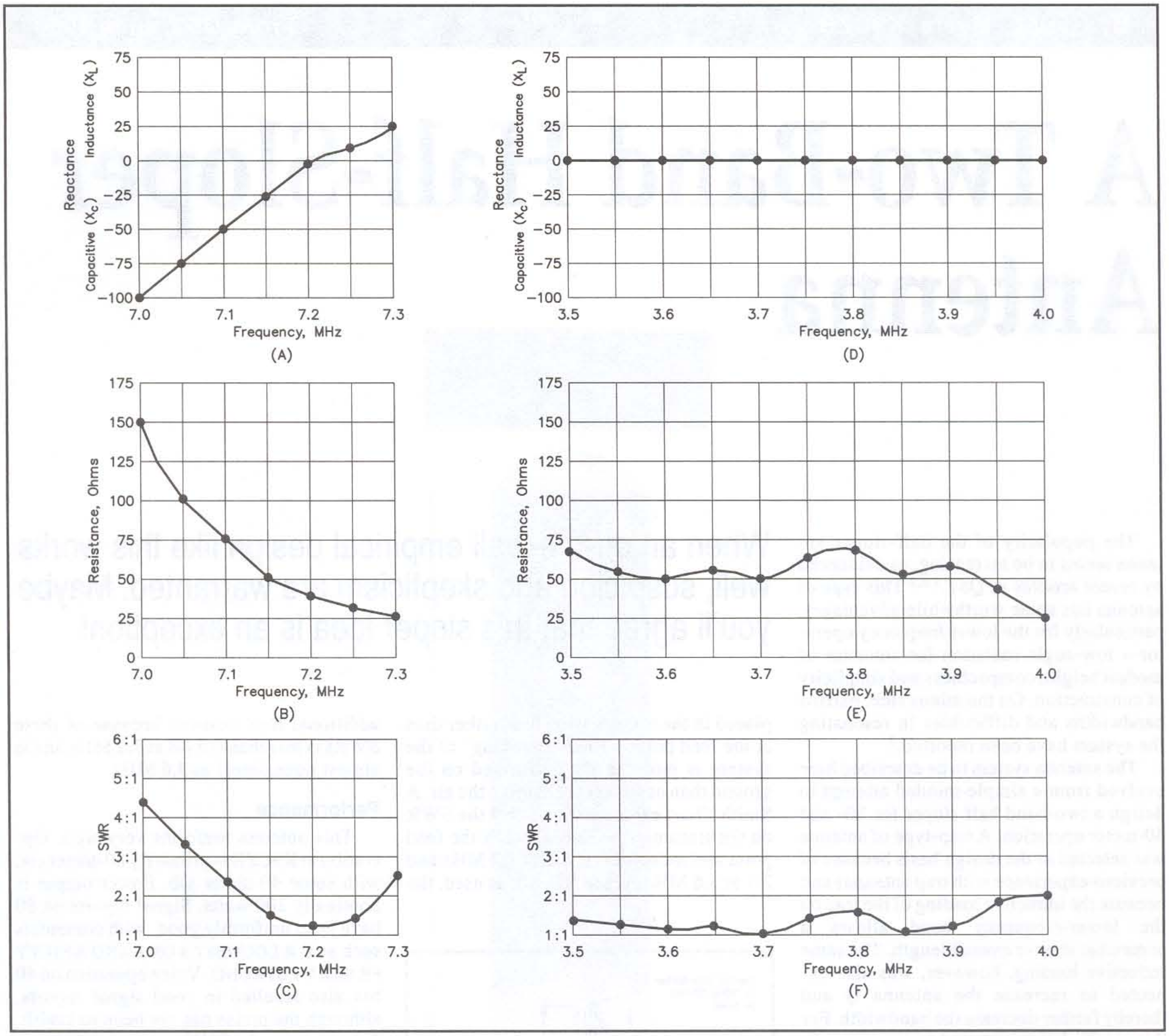

Figure 2—Loading characteristics of the antenna. The SWR curves were determined from a Smith Chart. Impedance values were measured with a noise bridge.

in later versions I ran the feed line up the *inside* of the support mast to shield it from the antenna field. This precaution seems to be effective, for no such sensitivity has been observed since. (I can't help but wonder if this might improve the behavior of any cantankerous half sloper.) If a nonmetallic support is used, or if there is no possibility of placing the feed line inside the support, double-shielded coaxial cable should serve equally well, provided the outer braid is connected to the inner braid at the top of the tower and a ground is connected to the outer braid at the bottom of the tower.

I used standard 10-foot TV mast sections for my support, simply because they were on hand. This results in a very flimsy and flexible mast in a 40-foot length, though. The first 20 feet must be doubled up with long U-bolts if the mast is to be walked up. The price of six sections of TV mast is about the same as a 36-foot telescoping push-up mast, but the latter is much sturdier and far easier to erect. If support materials are not already on hand, the telescoping mast is the better choice. I should mention that I took the trouble to bond all sections of mast together electrically to ensure good conductivity and to guard against TVI from rectification at joints after inevitable corrosion sets in. The base of the mast should be grounded. Effects from the guy wires can largely be avoided by breaking them into nonresonant lengths with strain insulators placed at the mast and every 19 feet thereafter.

The inductance and capacitance values shown in Figure 1 must be used for the trap. Construction techniques for the trap are covered in *The ARRL Antenna Book*.[4] A novel and inexpensive method of trap construction has been described by WB9OQM.[5] I built my trap using the method shown in Figure 3. Traps made in this fashion are much stronger than they appear. I've never had one break, even in high winds that caused property damage. In this antenna, however, the radiator also serves as one of the top guys. For that reason the trap was reinforced. Two 3/16-inch thick pieces of plastic were used with three layers of glass cloth and epoxy sandwiched between them. A rotary wire brush serves well to rough up the inner surfaces of the plastic to ensure good adhesion. However, one may use coarse sandpaper for that purpose. Glass cloth and epoxy are sold as a repair kit in many hardware stores.

Before the antenna wires are connected,

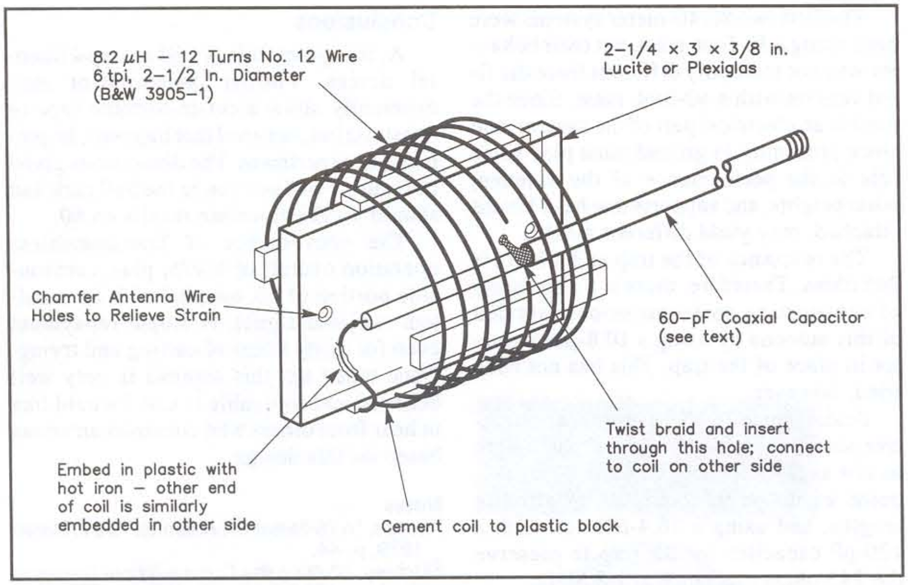

Figure 3—A simple, sturdy trap. The coaxial capacitor should be taped to the antenna wire after installation. It is not necessary to enclose the trap.

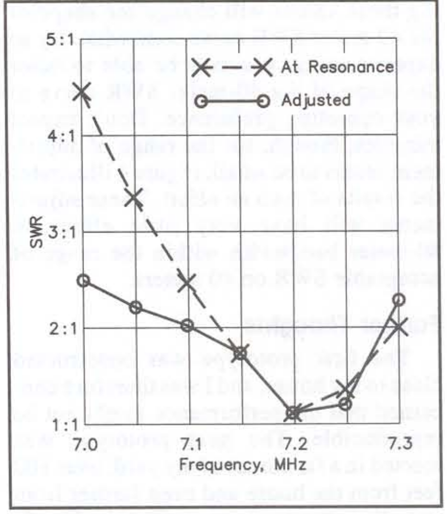

Figure 4—SWR curve for 40 meters, showing the effect of tapping L and adjusting feed-line length between L and the transmitter to obtain the best SWR curve. Such adjustments have little effect on 80-meter characteristics.

the trap must be tuned to resonance. A dip meter or noise bridge can be used to measure the resonant frequency. Start with about 30 inches of RG-8/U for the coaxial capacitor. After connecting it to the coil, as shown, 26 or 27 inches of braid will remain. At this point, the resonant frequency should be below 7.0 MHz. Trim the braid at the far end, a little at a time, snipping off the center conductor as you go. Recheck the resonant frequency each time. As 7.2 MHz is approached, continue trimming the braid, but stop cutting the center conductor. To increase the leakage path, the polyethylene dielectric should extend beyond the braid 1/8 to 3/16 inch (3.2 to 4.8 mm) when the trap is resonated at 7.2 MHz. Very close to 24 inches of braid should remain at completion. Tightly tape this end with several layers of plastic electrical tape.

The component values for this trap are exactly the same as those used in the W3DZZ trap dipole,[4] so there are several commercially made traps that may be suitable for this antenna. Traps made for a five-band, two-trap dipole, 108 feet long should have the proper values of capacitance and inductance.

In any antenna system, the radiator feed point impedance repeats itself every half wavelength along the transmission line. Inductor L must be inserted in the transmission line at a half-wave point in order to exactly cancel the capacitive reactance of this antenna system. It is, of course, advantageous to place L as close to the feed point as possible in order to minimize losses. A half wavelength at the lower frequency is as close as you can get without going to the feed point itself. The 90.2-foot length of RG-8/U shown in Figure 1 is an electrical half wavelength at 3.6 MHz for solid polyethylene dielectric coaxial cable *only*. If cable having a velocity factor other than 0.66 (e,g., foam-dielectric coaxial cable) is used, this length will have to be recalculated from the equation

$\ell = 492\,V/3.6$ feet

(Feet × 0.305 = m)

In this equation, V is the velocity factor of the cable to be used. Only RG-8/U or a similar type such as RG-213/U should be employed for this section of the transmission line in order to keep the losses low. If you don't mind a dB or so of loss on 40 meters, RG-58/U is acceptable. The loss on 80 meters will be negligible in any case.

The value of inductor L should be 1.75 μH, but I recommend that a coil having about 3-μH inductance be used to allow some latitude for final tune-up of the system. I mounted 12 turns of a no. 3018 Miniductor (1¼ inch in diameter, 8 turns per inch) in a small Minibox with SO-239 coaxial connectors placed at each end. After tapping the coil for the best SWR curve on 40 meters, the entire assembly was sealed and waterproofed with bathtub caulk.

Tune-up

As seems to be characteristic of half slopers, this antenna can be very touchy to tune up. If the length of transmission line between the radiator and L is not an exact integral multiple of a half wavelength at 3.6 MHz, tune-up can be a real "can of worms." Since the oft-quoted value of 0.66 for the velocity factor of standard RG-8/U is only a nominal value and can vary appreciably from brand to brand (in cheap cable from lot to lot), this length should be determined with a noise bridge.

If a noise bridge is not available, the following procedure may be tried. Cut this section of cable about 6 feet shorter than the calculated length. Prepare a section of RG-58/U, 12 feet long, with solderless connectors on each end. Connect it to both the shortened feed line, using a PL-258 double female connector, and to L. Tap L to obtain the best combination of SWRs at 3.6 and 7.2 MHz. Record the SWR figures and tap position.

Now shorten the RG-58/U by 6 inches and repeat — and repeat — until you are certain you have passed through the point where the SWR values simultaneously bottom out at both frequencies. Prepare a length of RG-58/U (from the same lot) exactly as long as the best experimental length, using permanent coaxial connectors. Seal and waterproof all connections. RG-58/U is recommended for relative ease of pruning. If you don't mind unsoldering a PL-259 each time, RG-8 could be used. However, the additional loss from such a short section of RG-58/U will be infinitesimal at these frequencies.

This procedure is obviously tedious, but it is necessary to obtain good performance on 40 if a noise bridge is not available.[6] In fact, some adjustment of this section of transmission line may be necessary even if a noise bridge is used to measure the electrical length to obtain optimum two-band performance. The exact half wavelength should always be used as a starting point, in any case.

Strangely enough, the above procedures are necessary only to optimize 40-meter performance. My experience has been that merely cutting the half-wavelength section of transmission line to the calculated length, then tapping L to obtain the best SWR at 3.6 MHz, is sufficient to obtain a ratio of 2:1 or less over the entire 80/75-meter band. So tune for 40, and 80 should take care of itself.

Once the antenna system has been resonated, it may pay to experiment with the value of L and the length of transmission line between L and the transmitter. Changing these values will change the shape of the 40-meter SWR curve somewhat. By so experimenting, you may be able to tailor the shape of the 40-meter SWR curve to your operating preference. Don't expect miracles, though, for the range of adjustment seems to be small. Figure 4 illustrates the results of such an effort. These adjustments will have very little effect on 80-meter bandwidth within the range of acceptable SWR on 40 meters.

Further Thoughts

The first prototype was constructed close to my house, and I was therefore concerned that the performance might not be reproducible. The next prototype was erected in a far corner of my yard, over 100 feet from the house and even further from any other structures or conductors. The final version was similarly located. Except for final tune-up parameters, all three behaved almost identically, even though a number of physical changes was made each time. As a final test of the soundness of the design, I built scaled versions for 40/20 and 20/10 meters. They exhibited very similar characteristics, although all parameters of the system seem to become very critical as the design frequency is increased.

The first two 80/40-meter systems were built using a 30-foot mast, yet their behavior was not markedly different from the final version with a 40-foot mast. Since the mast is an electrical part of the system, and since proximity to ground must play some role in the performance of the antenna, other heights, and supports that have beams attached, may yield different results.

The reactance of the trap at 3.6 MHz is 245 ohms. Therefore, there is a possibility of constructing an 80-meter-only version of this antenna by using a 10.8-µH inductor in place of the trap. This has not been tried, however.

Scaling up to 160/80 meters is an attractive possibility. Conceivably, a mast height as low as 50 feet could be used. A starting point would be the doubling of all wire lengths, and using a 16.4-µH coil with a 120-pF capacitor for the trap to preserve the 245-ohms reactance at 1.8 MHz.

There also seems to be a possibility that a slight increase in trap capacitor value, to resonate the tap at 7.1 or 7.15 MHz, might allow better coverage of 40 cw, but probably at the expense of the phone portion of the band. Such a change might have little effect on 80-meter bandwidth, but another round of cut-and-try could prove necessary.

Conclusions

At this point, this is still an experimental design. Further development may eventually allow a cut-to-formula type of construction, but until that happens, be prepared to experiment. The dimensions given in Figure 1 will put you in the ball park and should yield immediate results on 80.

The convenience of Transmatchless operation overall of 80/75, plus a reasonable portion of 40, coupled with an excellent radiated signal, is ample repayment even for many hours of cutting and trying. Once tuned up, this antenna is very well behaved and enjoyable to use. I would like to hear from others who construct antennas based on this design.

Notes

[1] Hopps, "A 75-Meter DX Antenna," *QST*, March 1979, p. 44.

[2] Atchley, "Putting the Quarter-Wave Sloper to Work on 160," *QST*, July 1979, p. 19.

[3] DeMaw, "Additional Notes on the Half Sloper," *QST*, July 1979, p. 20.

[4] *The ARRL Antenna Book*, 13th edition, 1974.

[5] Mathison, "Inexpensive Traps for Wire Antennas," *QST*, February 1977.

[6] [Editor's Note: A third alternative is to use a dip meter to determine an exact half wavelength of line. See Downs, "Measuring Transmission-Line Velocity Factor," *QST*, June 1979.)

By Deane J. Yungling, KI6O **From *QST*, April 1986**

The KI6O 160-Meter Linear-Loaded Sloper

No room for a top-band antenna? Try this one on for size!

After having good success using my linear-loaded, inverted-L antenna over the winter 1984-85 160-meter season, I decided to try linear loading on a different type of 160-meter wire antenna.[1] I have had considerable success with a quarter-wave sloper on 80 meters, but my city-sized lot isn't deep enough to accommodate a full-sized sloper for the 160-meter band. About 120 feet would be required for the sloper to be 10 feet off the ground at the low end; I have about half that distance to work with. Linear loading solved the problem.

This antenna provides an effective bandwidth of about 70 kHz with an SWR of 2:1 or less. At the design frequency of 1.840 MHz, SWR is 1.1:1 with 50-ohm feed line and no matching network or tuner.

Although the individual dimensions are not critical, the sloping wire and the ladder line must resonate at the desired frequency. If the sloping wire is less than 65 feet long, the ladder must be longer, and vice versa, If you use a different length for the sloping wire, you will need to experiment a bit to see how much to add or remove from the ladder length.

Construction

My sloper is hooked onto the tower at about the 55-ft level using a strain insulator (see Figure 1). The coaxial-cable feed fine is securely taped to one tower leg, and the shield is connected to the tower leg with a radiator hose clamp. The tower is grounded at the base with several ground rods. The center conductor of the coaxial cable is soldered to the sloper at the strain insulator and is taped for weather protection where the center conductor and shield separate.

The ladder portion of the antenna is made with the same type of wire as the sloping portion. The ladder spacers are made of 3/8-inch hardwood dowels that have small holes drilled 1 inch from each end to hold the wires. The dowels should be soaked or sprayed with a wood preservative, prior to assembly, for weather protection. Plastic spreaders could also be used, if desired. The wires are firmly tied to the dowels with waxed lacing twine or similar material where the wires pass through the holes. The dowel spacing can be adjusted, with some difficulty, after the dowels are tied.

One end of the ladder is tied to the tower with heavy monofilament fishing line, with the other end tied to a tree or whatever else is handy. The tie-off lines are fanned vertically to keep the ladder from twisting and should be tensioned enough to eliminate any sagging in the ladder. Monofilament fishing fine provides some cushioning if the supporting tree sways or if the wind is strong.

Adjustment

Initially, the ladder should be a few feet longer than shown, to allow for adjustment. The resonant frequency should be checked, and a few inches removed at a time from the tower end of the ladder until the desired frequency is achieved. At this point, the wires should be soldered together and then taped to the end dowel. It appears that there must be an BF beam, or something similar, on the tower for any quarter-wave sloper to work properly. This sloper is no different in this respect. Finally, the antenna, particularly the sloping section, should be kept as far as possible from surrounding guy wires and other objects.

This antenna requires no ground system, balun or matching network over its operational bandwidth. The linear-loaded sloper is simple to construct and easy to adjust, and its performance is superior to my linear-loaded, inverted L.

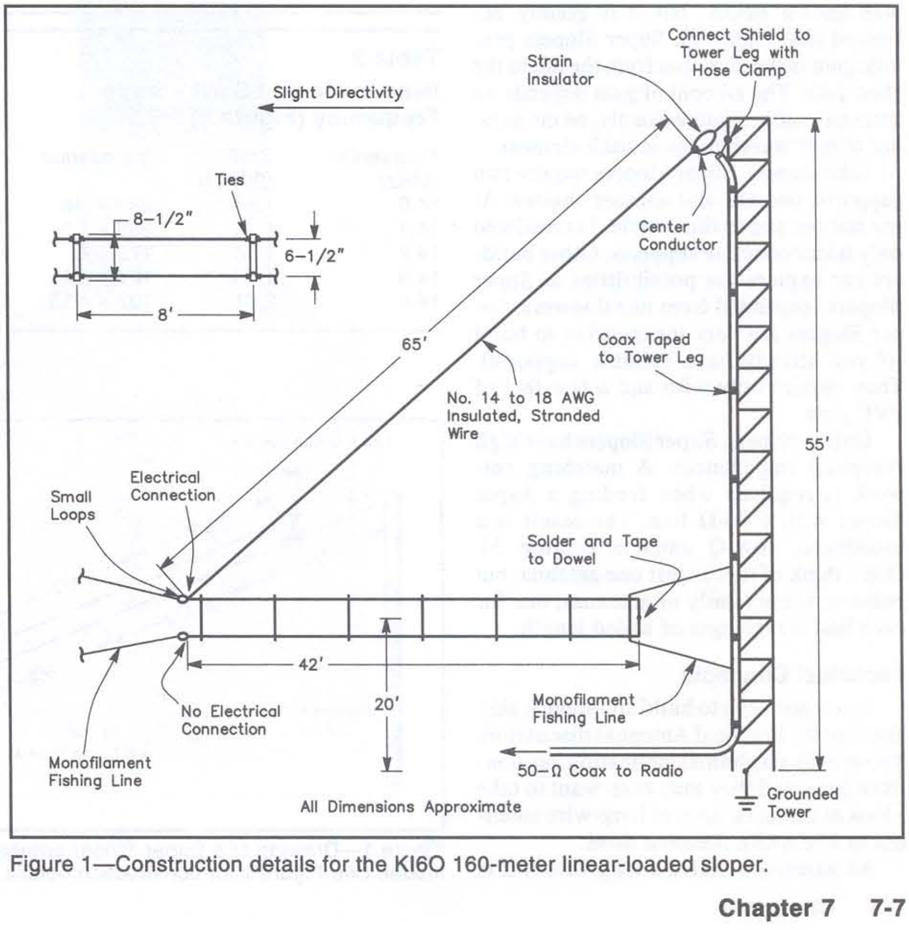

Figure 1—Construction details for the KI6O 160-meter linear-loaded sloper.

---

[1] D.J. Yungling, "The KI6O Top-Linear-Loaded 160 Meter Inverted 'L' Antenna," *CQ*, Apr 1985, pp 38-39.

By Roger Sparks, W7WKB **From *QST*, December 1995**

The Super Sloper

Dramatically improve the front-to-back of a sloper.

You can imagine my surprise when *ELNEC*[1] revealed that a parasitic element can be combined with a long wire to form a whole family of directional antennas. There on my screen, I was looking at useful combinations not found in that bible of antennas, *The ARRL Antenna Book*. I had been exploring combinations of full-wave parasitic elements with wire arrays. How far off center can I slide a parasitic element, I wondered? "A lot," was the answer coming from *ELNEC* [and confirmed by *EZNEC—Ed.*]. As I offset a closely spaced (0.015 to 0.046 λ) parasitic element by 1/8 to 5/8 λ, a whole family of antennas appeared (see Figure 1).

I call the configuration a Super Sloper because the pattern resembles that of the well-known sloper, but it is greatly enhanced (see Figure 2). Super Slopers provide gain in the direction from the tall to the short pole. The amount of gain depends on antenna length, or more exactly, on the number of half wavelengths in each element.

Like slopers, Super Slopers require two supports, one tall and another shorter. At my station, and in this article, I considered only nonconductive supports. Other builders can explore the possibilities of Super Slopers suspended from metal towers. Super Slopers are very inexpensive to build (if you already have suitable supports). They require only wire and a few feet of PVC pipe.

Unlike slopers, Super Slopers have high feedpoint impedances. A matching network is required when feeding a Super Sloper with a 50-Ω line. The result is a broadband, low-Q antenna (Figure 3). Don't think of this as just one antenna, but rather a whole family of antennas, one for each half wavelength of added length.

Technical Concepts

If you just want to build an antenna, skip ahead to the Practical Antennas discussion. Those with a technical inclination can continue here, and they may even want to take a look at the discussion of long-wire antennas in *The ARRL Antenna Book*.

An antenna is called a long wire if it is one wavelength or longer. The radiation pattern can be described as the surfaces of two opposed cones that are coaxial with the wire and have their apexes meeting at the feed-point. The apex angle becomes smaller and the lobes grow stronger as antenna length increases. An azimuth plot of radiation pattern for single long-wire antennas over ground shows four principal lobes at low radiation angles. There are two principal lobes, *in one direction*, when the antenna is terminated in a matched, resistive load.

Table 1
Spacing and Phase Angle for Various Antenna Lengths

At the spacings listed, the current in the resonant parasitic element is equal to driven-element current ±10%. (Derived from *ELNEC*, under free-space conditions, 0.25-λ offset, at 7.1 MHz.)

Antenna Length (λ)	*Spacing (λ)*	*Phase Angle (degrees)*
0.5	0.046	−156
1.0	0.030	−147
1.5	0.023	−144
2.0	0.018	−142
2.5	0.015	−142
3.0	0.015	−144

Table 2
Impedance and SWR versus Frequency (Figure 5)

Frequency (MHz)	*SWR (200-Ω)*	*Impedance*
14.0	1.54	293 + *j*48
14.1	1.25	248 + *j*13
14.2	1.16	172 + *j*0
14.3	1.75	127 + *j*54
14.4	2.61	107 + *j*113

ELNEC shows that a parasitic element added to a single long-wire antenna creates a *unidirectional* radiation pattern similar to that created by a resistive termination. Energy eliminated from the back goes into useful forward gain. The parasitic element can be tuned as either a director or a reflector.[2]

Experience with Yagis (and many other applications of parasitic elements) led me to believe that all parasitic elements must be nearly a half wavelength long and located within the span of the driven element. That concept is completely *incorrect*. Parasitic elements can be *any* resonant length and offset from the driven element, as long as there is adequate coupling between the elements.

Here's how Super Slopers produce front-to-back ratio (F/B) and gain. The best

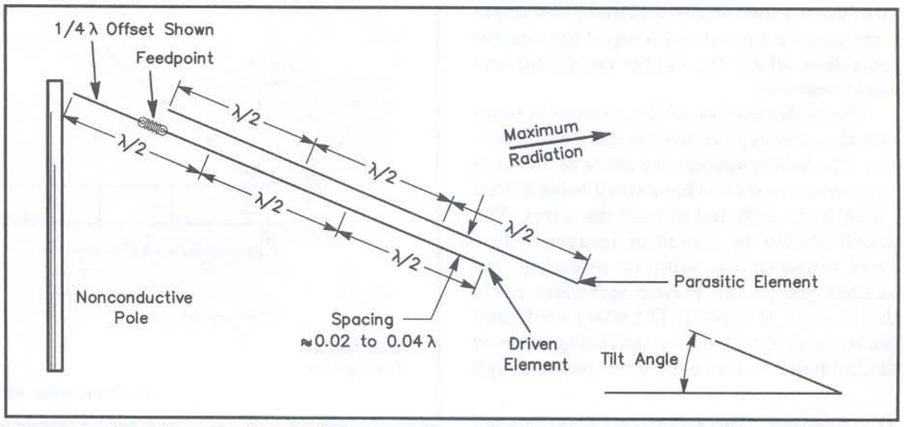

Figure 1—Drawing of a Super-Sloper antenna showing nomenclature. This is a 3-λ/2 model. See Figure 7 for construction details.

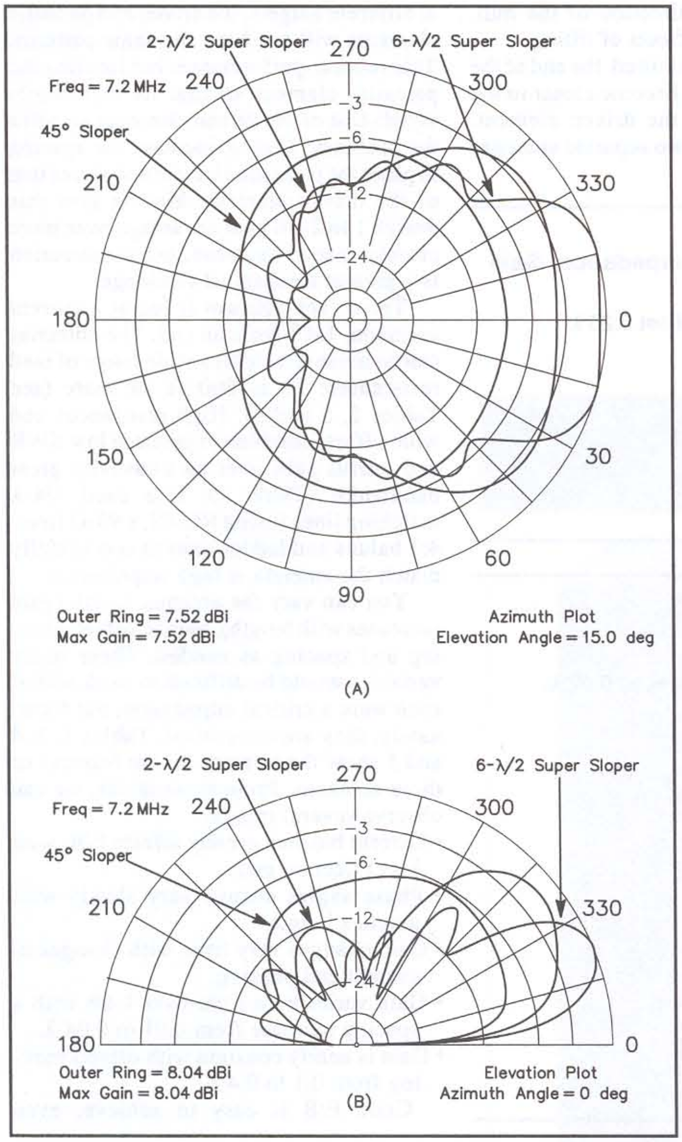

Figure 2—Azimuth (A) and elevation (B) plots for a 45° sloper, 2-λ/2 (0.25-λ offset, spaced 0.04 λ, low end 0.119 λ above ground) and 6λ/2 (0.25-λ offset, spaced 0.015 λ, low end at 0.066 λ above ground) Super Slopers. The high end of each is λ/2 high.

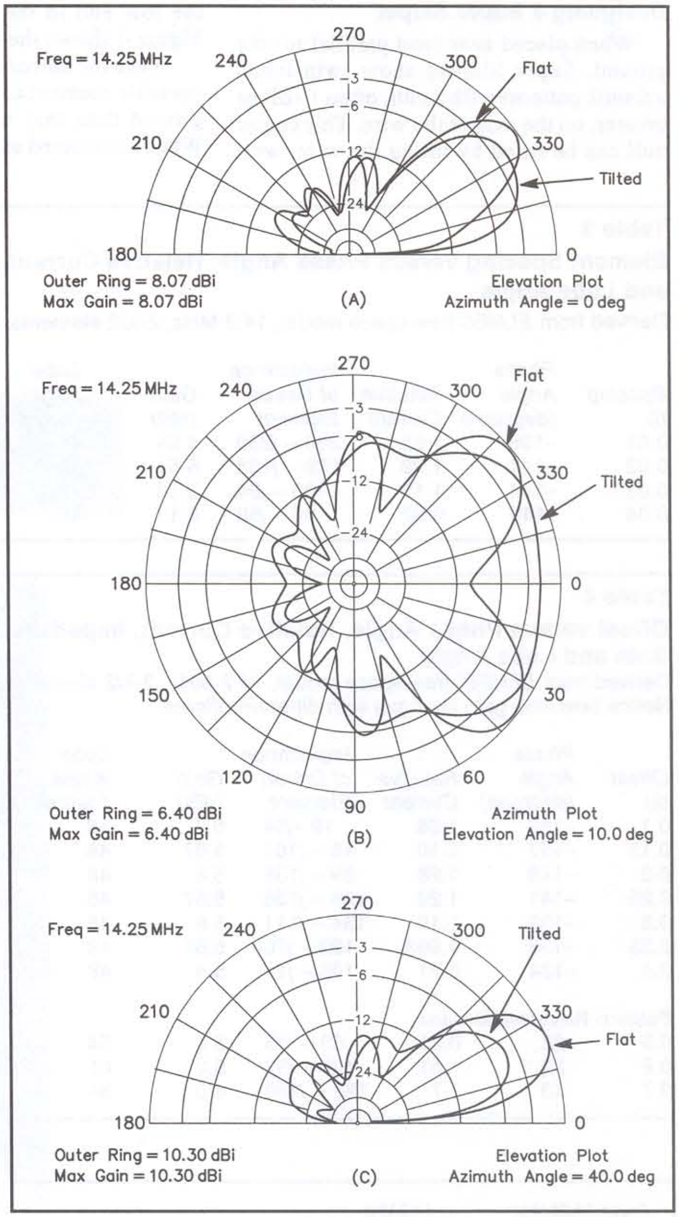

Figure 4—Patterns of a 3-λ/2 antenna with 0.36-λ offset and spaced 0.04-λ oriented flat at 1/2 λ vs sloping from 1/2 λ to 0.045 λ. A shows the elevation patterns at 0°; B shows the azimuth patterns; C shows elevation plots at azimuth=40°.

F/B results from a 180° phase shift between the currents in the driven and parasitic wires, as seen from a distant location where a pattern null is desired. A phase difference of 180° means that the signals completely cancel each other, if each element delivers equal signal strength. A phase shift of 90° is possible with a 1/4-λ offset, and the remaining 90° phase shift can come from the tuning (length) of the parasitic element. If the fields cancel in one direction, they will reinforce in some other direction, to produce gain.

By increasing the phase angle of the current in the parasitic element, the angle of maximum cancellation can be moved. This is particularly useful in Super Slopers because the principal lobes of long wires lie at some small angle to the direction of the wire. From Table 1, notice that the current in a resonant parasitic element has about –144° to –156° phasing, which is about right for correct reverse-lobe canceling.

If we want each element to deliver equal signal strengths to a distant location, nearly equal currents must flow in both the driven and parasitic elements: The coupling between elements must be very close. Table 1 suggests the approximate spacings for currents to be equal within ±10%. (The antenna also has gain at closer spacings because gain is affected less by unequal currents than is F/B.)

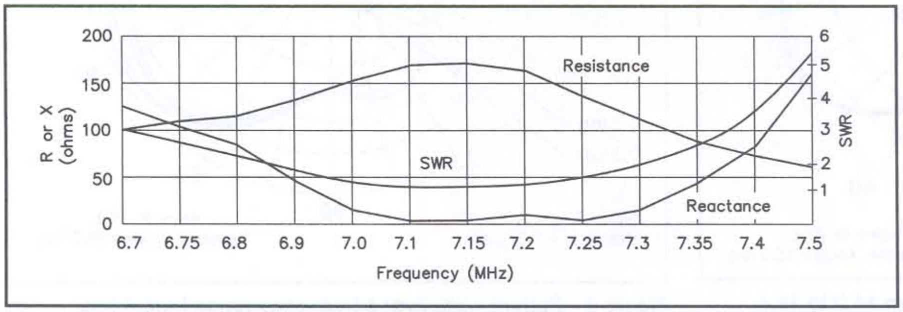

Figure 3—A plot of impedance and SWR versus frequency for the 2-λ/2 antenna described in Table 6, placed as described in Figure 6. The data source is *NEC/WIRES 1.5* (see Note 3).

Designing a Super Sloper

When placed near (and parallel to) the ground, Super Slopers show twin-lobed azimuth patterns with a null, often 10 dB or greater, on the axis of the wire. This center null can be filled by tilting the wire, with the low end in the direction of the null. Figure 4 shows the effects of tilting.

When the antenna is tilted, the end of the parasitic element can become closer to the ground than that of the driven element. When considered as two separate antennas at different heights, the driven and parasitic elements will not have the same patterns. This reduces performance, but locating the parasitic element so that its end height equals that of the driven element can solve the problem. That is, increase the spacing to place the parasitic element end above that of the driven element. Models give this design 1 to 2-dB gain advantage over more closely spaced antennas, but construction is a greater mechanical challenge.

Table 3
Element Spacing versus Phase Angle, Relative Current, Impedance, Gain and Lobe Angle

Derived from *ELNEC* free-space model, 14.2 MHz, 2-λ/2 elements, offset 0.25 λ

Spacing (l)	*Phase Angle (degrees)*	*Relative Current*	*Impedance of Driven Element*	*Gain (dBi)*	*Lobe Angle (degrees)*
0.01	–136	1.62	205 – *j*230	4.98	48
0.02	–141	1.29	128 – *j*138	5.54	48
0.03	–144	1.11	103 – *j*94	5.78	48
0.04	–149	0.97	80 – *j*59	6.1	48

Table 4
Offset versus Phase Angle, Relative Current, Impedance, Gain and Lobe Angle

Derived from *ELNEC*, free-space model, 14.2 MHz, 2-λ/2 elements, spaced 0.02 λ. Notice how little gain changes with different offsets.

Offset (λ)	*Phase Angle (degrees)*	*Relative Current*	*Impedance of Driven Element*	*Gain (dBi)*	*Lobe Angle (degrees)*
0.1	–167	1.08	19 –*j*54	5.7	48
0.15	–157	1.18	48 – *j*101	5.67	48
0.2	–148	1.26	89 – *j*135	5.6	48
0.25	–141	1.29	128 – *j*138	5.57	48
0.3	–139	1.16	134 – *j*111	5.6	48
0.35	–138	0.963	123 – *j*70	5.58	48
0.4	–134	0.71	108 – *j*14	5.4	48
Pattern Reversal Begins					
0.5	–26	0.25	83 + *j*46	3.0	54
0.6	39	1.31	126 – *j*57	5.2	61
0.7	43	1.71	192 – *j*180	4.0	64

The driven element is fed at a current antinode, 1/4 λ from an end. The antennas can be made to exhibit a wide range of feed resistances: 20 to 300 Ω, or more (see Tables 2, 3 and 4). High resistances and wide offsets combine to produce low SWR and useful gain over an unusually great bandwidth (Table 2). I've used 1/4-λ matching lines (using RG-62, a 93-Ω line), 4:1 baluns and ladder lines to successfully match the antenna at high impedances.

You can vary the antenna length (gain increases with length), height, offset, phasing and spacing as needed. These many variables would be difficult to work with if each were a critical adjustment, but fortunately, they are not critical. Tables 1, 3, 4 and 5 show the tolerant design features of these antennas. From these tables, we can observe several trends:

- Current balance greatly affects F/B, with less effect on gain.
- Phase angles change very slowly with antenna length.
- Gain changes very little with changes of current and phasing.
- Gain varies by a little over 1 dB with a spacing increase from 0.01 to 0.04 λ.
- Gain is nearly constant with offsets moving from 0.1 to 0.4 λ.

Good F/B is easy to achieve, even

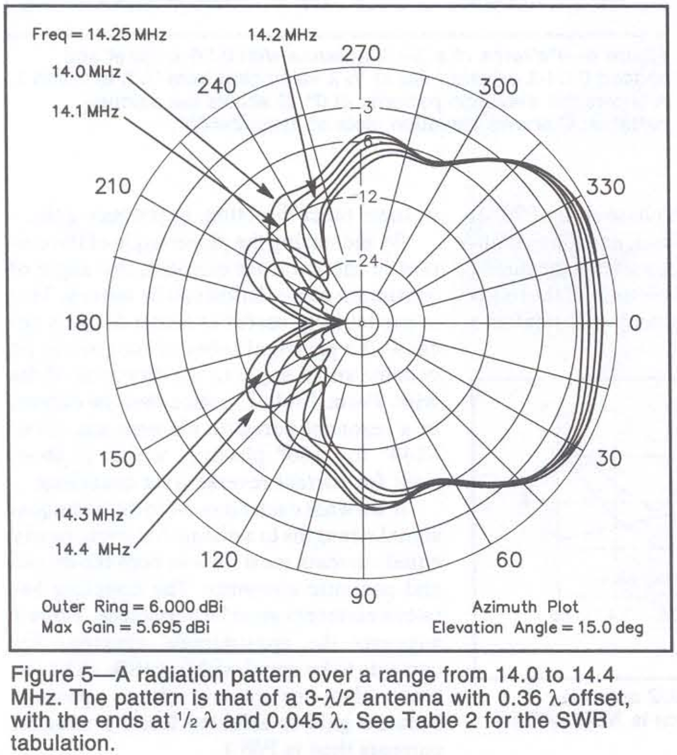

Figure 5—A radiation pattern over a range from 14.0 to 14.4 MHz. The pattern is that of a 3-λ/2 antenna with 0.36 λ offset, with the ends at ½ λ and 0.045 λ. See Table 2 for the SWR tabulation.

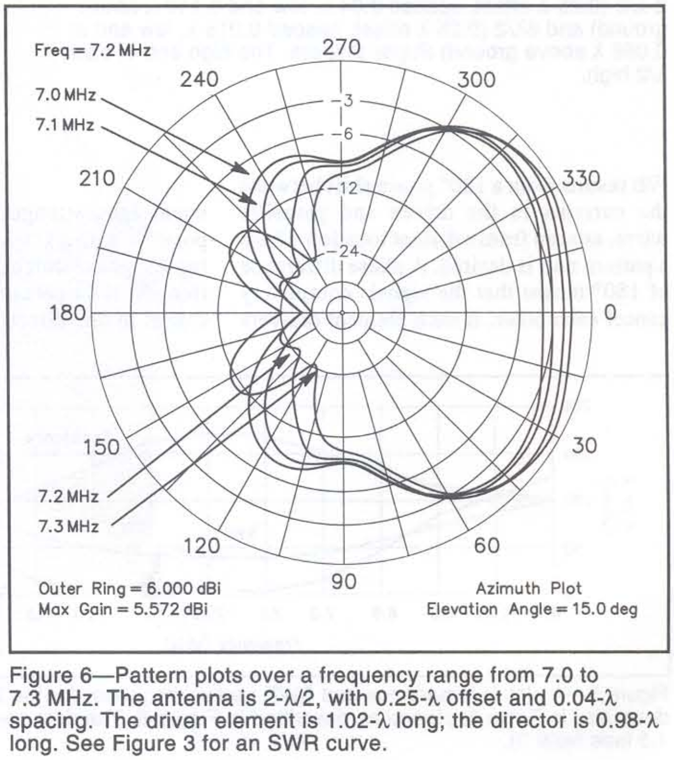

Figure 6—Pattern plots over a frequency range from 7.0 to 7.3 MHz. The antenna is 2-λ/2, with 0.25-λ offset and 0.04-λ spacing. The driven element is 1.02-λ long; the director is 0.98-λ long. See Figure 3 for an SWR curve.

though the antenna is designed for best gain or easy construction. Improved F/B ratios and reverse-oriented patterns are possible from designs optimized at specific frequencies. Computer modeling is the best way to optimize reverse patterns. Figures 5 and 6 show some possible patterns and how the patterns vary with frequency. Table 2 and Figure 3 show predicted SWR.

Practical Antennas

Table 6 describes Super Slopers constructed at my station (refer to Figure I for nomenclature). The principal difference between the two 3-λ/2, 20-meter antennas in Table 6 is the increased offset from 17½ feet to 22 feet. This changes the phasing, resulting in a slightly improved F/B ratio and slightly improved gain. The trade-off is a narrower pattern and a longer antenna.

Table 5
Parasitic Element Length versus Phase Angle, Relative Current, Gain and Lobe Angle

Derived from *ELNEC*, under free-space conditions with 14.2 MHz, 2-λ/2 elements, spaced 0.02 λ apart.

Percent Short (λ)	*Phase Angle (degrees)*	*Relative Current*	*Gain (dBi)*	*Lobe Angle (degrees)*
0	–140	1.30	5.54	48
1	–125	1.47	5.33	49
2	–107	1.53	5.02	51
3	–89	1.43	4.70	52
4	–75	1.23	4.44	53
5	–65	1.05	4.19	53

Table 6
Working Antenna Dimensions

Use a 4:1 step-down transformation to match 50-Ω line.

Antenna Number	*1*	*2*	*3*	*4*	*5*	*6*
Band	40 m	20 m	20 m	20 m	15 m	10m
Length*	2	3	3	2	2	2
Driven (ft)	139.3	106.5	105.6	70	46	34.5
Director (ft)	134.3	104.0	102.2	68	45	32.8
Offset (ft)†‡	34.3	17.5	22	17.5	11.5	8.6
Spacing (ft)	6	1.5	1.5	1.5	1.0	1.0
Feedpoint (ft)‡	34.3	17.5	17.25	17.5	11.5	8.6

*Length expressed as a multiple of λ/2.
†Offset is λ/4, except for antenna 3, where the offset is 0.31 λ.
‡Feedpoint and offset are both measured from the high end of the driven element.

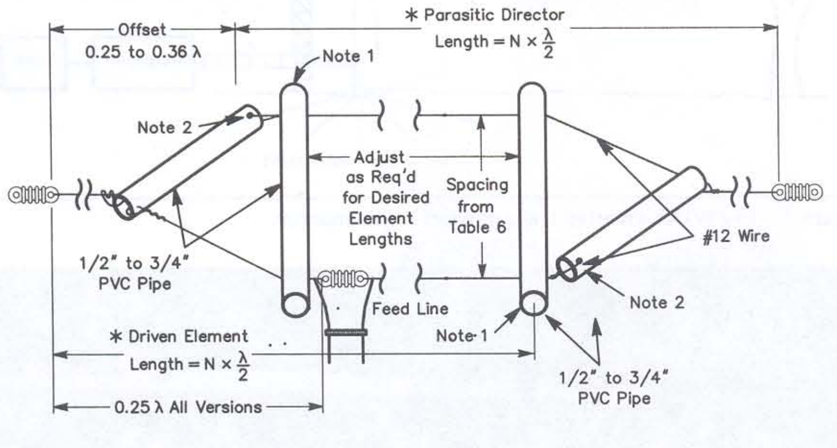

Note 1. Spreader length is 2" longer than spacing. Wires are threaded through holes drilled in spreaders.

Note 2. Tension member is 2" longer than 0.60 × spacing. Wires are threaded through holes drilled 1" from each end.

* Length must be adjusted for optimum gain or F/B ratio. Begin with lengths given in Table 6.

λ= Wavelength

Figure 7—A method of constructing the long-wire parasitic antenna. Additional center spacers reduce the tendency of the wires to twist in the wind.

A 60-foot mast supported the 3-λ/2 and 2-λ/2 20-meter versions. They were mounted back-to-back and tilted to fill the center null. Another 3-λ/2, 20-meter unit was mounted from a 38-foot mast. The 40-meter version sloped from 60 feet down to 6 feet. All antennas performed as predicted. Additional height at either end definitely increases low-angle radiation, but it reduces high-angle radiation.

Figure 7 shows how I assembled and spread the wires. The departure from straight lines (as depicted in Figure 1) has no practical effect on the antenna. Be sure to place the director over the driven element when erecting the antenna. A director placed at the side will skew the pattern, favoring the side of the director. Keep the Super-Sloper support lines tight to minimize sagging. Severe sagging leads to improper phasing and degraded results. I put a support under my 40-meter Super Sloper at midspan, but I just pull the 20-meter antenna support lines tight. The two elements have a tendency to twist and wrap together in the wind. You can prevent this by using two support ropes at the low end or by using spacers as on transmission lines. Both methods work well. Super Sloper performance suffers from excessive ground losses when the low end is at ground level. I strive for an antenna slope (tilt angle) of 10° to 20° and a minimum height of 6 feet. Higher is better.

A 200-Ω feedpoint impedance is easily transformed to 50 Ω with a 4:1 balun. Ladder line and a tuner are another option. Refer to *The ARRL Antenna Book* for other methods of transforming high feed impedances to values acceptable for modern transceivers. Do not feed the Super Sloper directly with coax unless the coax is part of an impedance-matching section.

Results and Conclusions

I've built and used six Super Slopers very successfully. The F/B is often dramatic. A difference of six S units is not uncommon when switching between two antennas built to favor opposite directions.

This family of antennas has not yet been well researched and studied. Use an antenna modeling program before building designs that are substantially different (eg, longer or made from tubing) from those described in Table 6.

Notes

[1] *ELNEC*, and its successor *EZNEC*, are computer antenna modeling programs available from Roy Lewallen, W7EL. The description files for the antenna plots shown in this article are available as a self-extracting archive file named SUPSLOPE.EXE. This file can be found on the Internet (FTP to OAK.OAKLAND.EDU, directory pub/hamradio/arrl/qst-binaries).

[2] Parasitic elements offset more than 1/2 λ from the driven element, become reflectors.

[3] *NECIWIRES 1.5* is a computer antenna modeling program available from Brian Beezley, K6STI.

By John F. Lindholm, W1XX **From *QST*, January 1983**

The Inverted L Revisited

City dwellers, don't despair. Here is a *good* 160-meter antenna that should fit on your lot!

I'll never know what inspired me to make a few contacts in the ARRL 160Meter Contest. My antenna was made by tying together the open-wire feeders of my 80-meter dipole. The performance was not fantastic, but it was the first step in getting me "hooked" on the "gentlemen's band." Working three dozen European stations from a friend's house got my interest up. My friend has a good 160-meter antenna system – and *lots* of property to fit it on!

Returning to my 60 × 150-foot lot made me feel depressed.[1] I suffered all winter while listening to the others working VKs, ZLs and even JAs at daybreak. What could I do to improve *my* signal?

Many hours the following summer were spent trying to figure out how to cram a 160-meter antenna within the confines of my small lot. Space restrictions dictated that my wire be no longer than a standard 80-meter dipole. I began to consider alternative antennas.

Shunt feeding my 50-foot tower was investigated but dismissed for various reasons. This arrangement would require disconnecting the shunt feed when cranking down the tower, and all guy wires would have to be broken up into nonresonant lengths with insulators. Additionally, my cables would have to be rerouted to ground level. I then considered a full-length dipole originating in a neighbor's yard three houses to the east, crossing my property and finally terminating in the yard two houses to the west! This idea was rejected. The legal negotiation fees would have run into six figures! After much head scratching, I settled on the inverted L, an antenna made popular by the grand master of 160-meters, Stew Perry, W1BB. I credit Stew for coming to my rescue!

The Inverted L

The inverted L was selected because it requires no more space than an 80-meter dipole and I could utilize my 50-foot crank-up tower for attachment. The vertical part is 50 feet long, and the horizontal part measures 130 feet, for an overall length of 180 feet (Figure 1). This makes the antenna approximately 3/8 λ, with the horizontal part providing top-loading. W1BB advises making the vertical section as long as possible (depending on tower height), and that an overall length of 160 to 180 feet works well.

Construction

With a bow and arrow, I successfully attached the far horizontal end of the antenna to the top of a 60-foot fir tree. From there, I ran the wire back to the top of the tower, where I bracketed a 30-inch-long two-by-four with an insulator screwed in the end (Figure 2). No. 14 Copperweld wire is used for the horizontal section, and no. 10 copper wire is used for the vertical section, which is spaced about 2 feet from the tower. These two wires are soldered to-

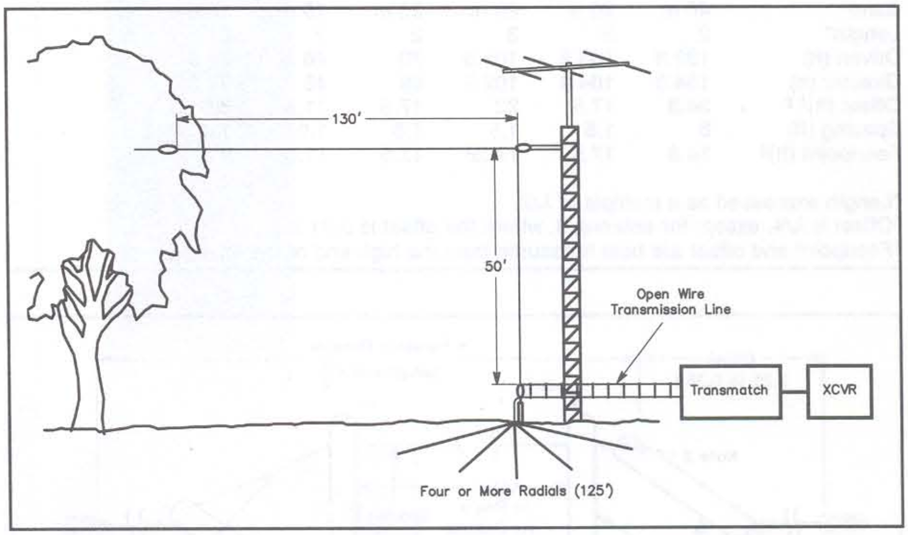

Figure 1—The W1XX inverted L is arranged in this manner.

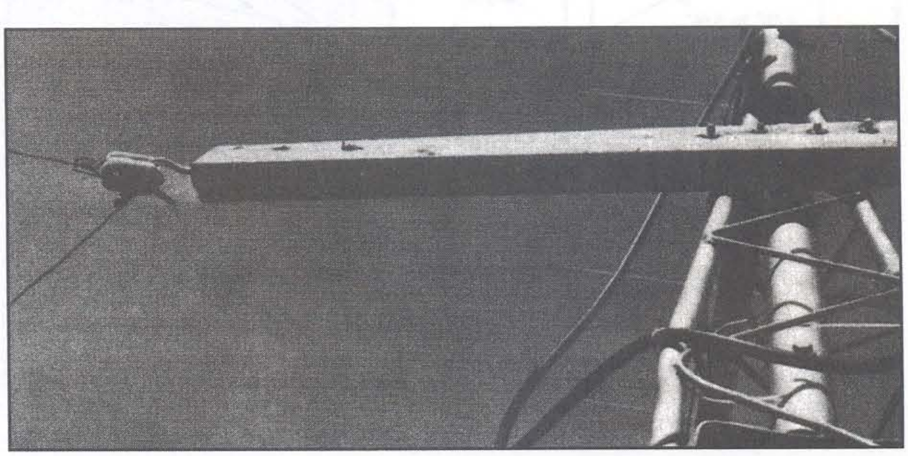

Figure 2—Detail of how the antenna is mounted to the top of the tower. TV-mast clamps are used to secure the wooden insulator.

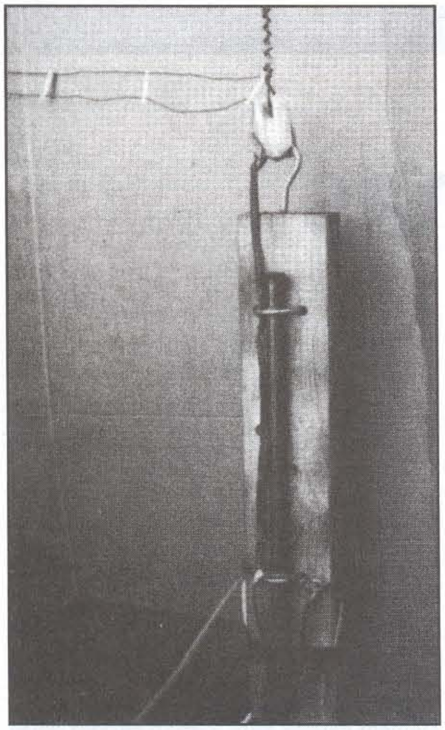

Figure 3—A ground rod at the tower base serves as a physical support for the insulator, and as an rf ground. Radial wires are connected to the rod by means of a ground bus, as described in the text.

gether at the insulator.

A plumb line was used to locate a point on the ground directly below the 90° bend in the "L." At this point, a 10-foot copper-clad ground rod was pounded into the earth, leaving 18 inches sticking out. A two-by-four is clamped to the ground rod by means of TV-mast U bolts. The lower end of the vertical wire is attached to an insulator that is screwed into the top of the wooden block (Figure 3). Next, I stripped some coaxial cable (RG-8/U) of its outer braid and used this to make a ground bus around the antenna base. One side of the open-wire transmission line is soldered to the base of the vertical antenna element, and the other side to the bus, which is attached to the ground rod with a clamp. This bus also serves as a connection point for the radials. So far, less than an afternoon of work had been invested.

Radials — the More the Better

The next day, my objective was to install radial wires, which are necessary because the inverted L is essentially a top-loaded vertical radiator. Previous meditation convinced me that several radials would fit on my lot. I'd *make* them fit! All radials were cut to 1/4 λ (125 feet), using scrap wire. About 300 feet of surplus telephone ground wire provided a good start. Stripping some old coaxial cable with a single-edged razor blade produced two radials from one length of wire (outer and center conductors). My technique for radial installation consisted of creasing the earth with a spade and shoving the wire in; afterwards the turf was pressed back into place with my heel.

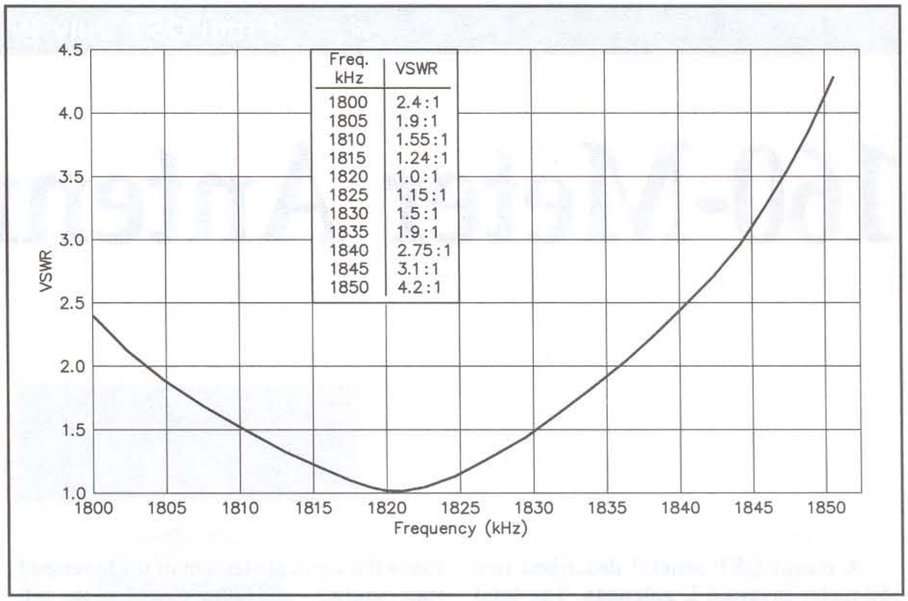

Freq. kHz	VSWR
1800	2.4:1
1805	1.9:1
1810	1.55:1
1815	1.24:1
1820	1.0:1
1825	1.15:1
1830	1.5:1
1835	1.9:1
1840	2.75:1
1845	3.1:1
1850	4.2:1

Figure 4—SWR curve for the inverted L, with the Transmatch adjusted for an SWR of 1 at 1820 kHz.

With space a problem, it may be impossible to place all your radials in a straight line; don't worry, because it is not necessary. My installation followed a zigzag path to avoid fixtures like the house and driveway. W1BB advises putting some radials under the horizontal part of the antenna. Unfortunately, the location of my garage prevented this. Initially, only four radials were planted, but more were added later. As with all vertical antennas, the more radials you can put in the ground, the better the performance!

Matching System

Voltage-fed antennas approximately $^1/_2$-λ long, such as the inverted L, will have a fairly high feed-point impedance. They will also exhibit inductive or capacitive reactance, depending on whether the antenna is slightly longer or shorter than $^1/_2$ λ. Since the inverted L does not have a 50-ohm impedance, a matching system is needed. To make tuning adjustments easier, I opted for locating the Transmatch *in* the shack, using open-wire line to the antenna. Being able to adjust the antenna match conveniently is recommended, since the SWR climbs rapidly as you shift frequency. For example, adjusting the Transmatch for a 1:1 SWR at 1820 kHz produces an SWR of 2.4:1 at 1800 kHz (see Figure 4). With the Transmatch located in the shack, you can easily adjust for a 1:1 SWR no matter where you operate in the band. My Transmatch consists of a plug-in, link coupled coil and variable capacitors – all scrounged at flea markets. Any of the configurations found in the *ARRL Antenna Book*[2] should work well.

Performance

Does the antenna work? Having no comparison antenna, my conclusions are subjective. But I've been on the air enough to know when, as they say, "it plays." With only four radials in place, my first night of operation yielded plenty of U.S. contacts, plus a Caribbean DXpedition on the first call. Subsequently, many European stations have been worked from my northeast location with good signal reports. Contest activity has yielded some respectable scores, including many QSOs with the Caribbean, and South and Central America — even Antarctica!

By adjusting my Transmatch, I made limited tests with the L on 75-meter ssb. Comparisons were made to a 75-meter dipole at 50 feet. For signals close in, the L was down by some 3 to 5 dB, but equal or superior to the dipole for signals from eastern Europe. Apparently, the 160-meter inverted L also provides a low angle of radiation on this band too. On bands higher than 75 meters, the radiation angle will be tilted *upwards*, rendering the antenna inefficient for DX work. This phenomenon is explained in *The ARRL Antenna Book*.[3]

Giving up the 160-meter band for lack of sufficient real estate is unwarranted. With the inverted L, you can work Top Band from your urban lot. Installation is a breeze, and the performance is admirable. *Now* what's your excuse for missing out on the excitement of 160 meters?

Notes

[1] m = ft × 0.3048.

[2] *The ARRL Antenna Book*, 14th ed. (Newington: ARRL, Inc., 1982), pp. 4-1 through 4-8.

[3] *Ibid.*, pp. 2-23 through 2-24.

From *QST*, July 1991 (Technical Corespondence)

160-Meter Antennas

A recent QST article[1] described two 160-meter inverted-L antennas. The total length of one of the antennas (the vertical element and the horizontal arm) is approximately 1/4 λ; the other antenna is 1/2 λ long. The current-driven 1/4-λ antenna requires a fairly extensive ground screen to realize good radiation efficiency. On the other hand, the voltage-driven 1/2-λ antenna requires only a ground stake-or no ground at all. Here's some additional information to aid you in selecting between these alternatives.

Because the current distribution of these two antennas is quite different, the radiation characteristics are also different. That explains why some signals are ". . . very loud or very weak," depending on station distance, propagation conditions and which antenna is being used.

In the captions for the figures illustrating the two antennas, it's stated that the support poles could be metal or wood. Because the vertical element of the inverted L runs parallel—and rather close to—the support pole, one may wonder whether, if the pole is metal, this proximity will influence antenna performance.

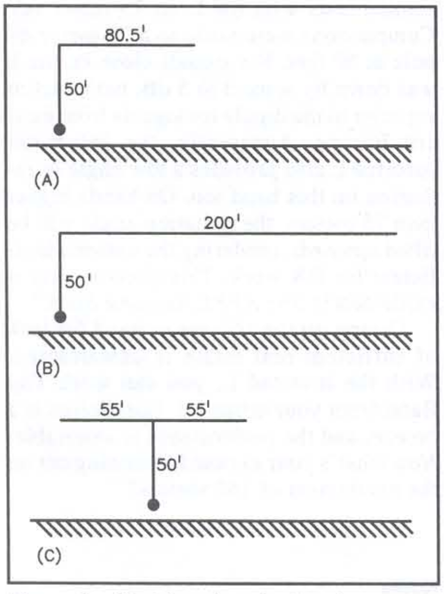

Figure 1—Sketches for simple wire antennas for 160 meters (1.9 MHz); A, 1/4-λ inverted L; B, 1/2-λ inverted L; and C, 1/4-λ T antenna.

The three antenna types I'll discuss are illustrated in Figure 1; they are (A) the 1/4-λ inverted L; (B) the 1/2-λ inverted L; and (C) the T antenna. Figures 2 and 3 show the calculated[2] radiation patterns for the two inverted-L antennas, assuming average ground ($\sigma = 3$ mS/m, $\varepsilon = 13$). The support poles are trees (or other wooden supports) 50 feet tall. Note that the horizontal arm of the antenna is close to the ground, in terms of wavelength (0.1 λ), and that, in this case, *ELNEC* overestimates the gain of horizontally polarized antennas at a height of 0.1 λ by about 3 dB.[3]

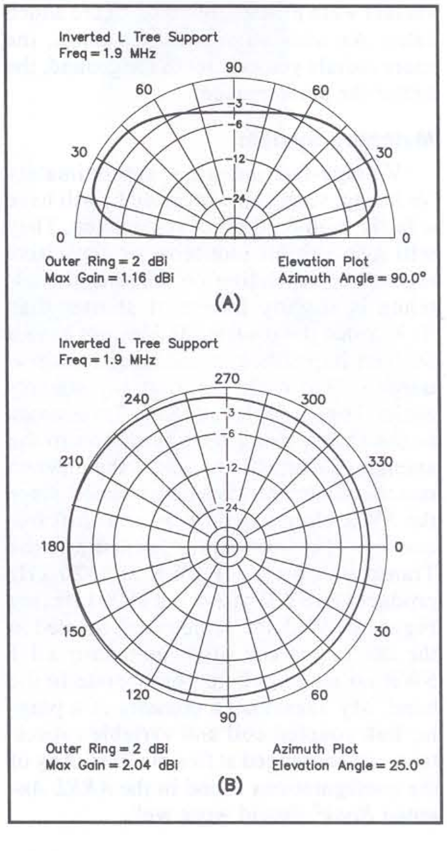

Figure 2—Radiation patterns for Figure 1A's antenna over average ground. The radiated field is dominantly vertically polarized—a monopole-like pattern. The absence of an overhead null in the vertical-plane pattern (compare Figure 2A with Figure 4A) is due to the horizontally polarized component of the field radiated by current on the antenna's horizontal arm.

The reason the patterns are so different is because the vertical antenna element carries a heavy current when the antenna length is 1/4 λ, whereas the vertical antenna element carries only a small current when the antenna length is 1/2 λ. Hence, there is a great difference between the vertically and horizontally polarized components of the radiation field for these two inverted Ls. The former antenna has a monopole-like pattern, whereas the latter antenna has a dipole-like pattern.

The lengths given in Figure 1 are those for resonance at 1.9 MHz. DeMaw's dimensions give wire lengths that are a bit too short for resonance (according to *ELNEC*). (In practice, one would trim the antenna length for resonance.)

If you want to work DX, it is advantageous to deploy an antenna that has a deep overhead null in its radiation pattern. This minimizes high-angle sky-wave signals, noise and interference, and improves the signal-to-noise ratio of distant weak signals. The T antenna exhibits this feature. Figure 1C provides dimensions for a T antenna with a resonant frequency of 1.9 MHz; the radiation patterns are given in Figure 4. This antenna's radiation field is almost entirely vertically polarized.

For these plots, the antenna was in the X-Z plane (Z is the vertical axis and 0° azimuth is in the + X direction). The 0° azimuth for the inverted Ls is the direction that the horizontal arm points (away from the feed). There is a slight azimuthal pattern asymmetry for these antennas because the antenna structure is not symmetrical

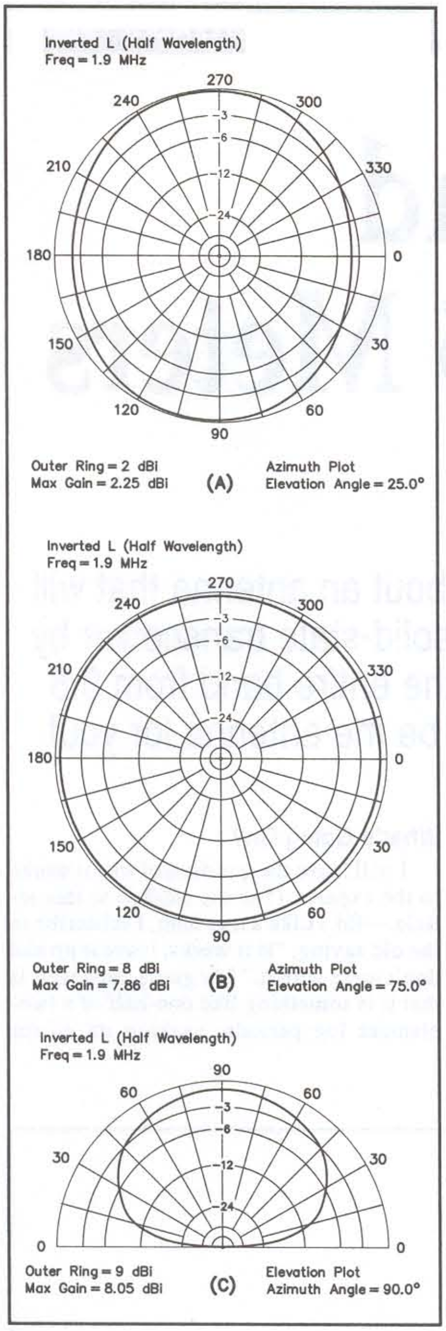

Figure 3—Radiation patterns for Figure 1B's antenna over average ground. The radiated field is dominantly horizontally polarized in the plane broadside to the antenna, and vertically polarized in the plane of the antenna—a dipole-like pattern. Hence the almost circular pattern at the high elevation angle, 75° (B), and the broadside directivity at the lower elevation angle, 25° (A).

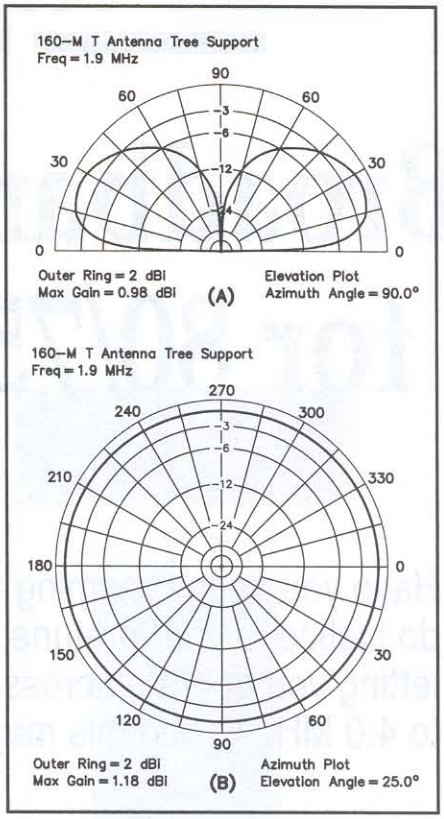

Figure 4—Radiation patterns for Figure 1C's antenna over average ground. The radiated field is almost entirely vertically polarized. The horizontally polarized component of the field radiated by current on one horizontal arm of the T is canceled by that radiated by the other arm.

with respect to the feed.

The effect of a well-grounded metal support tower was also studied. Surprisingly (to me), the effect is rather small, particularly for the 1/4 λ inverted L and 1/4 λ T antennas. The effect on the impedance of the 1/2-λ inverted L looked at first to be significant, but this is only because a small change in the resonant frequency of the antenna (the metal tower couples reactance into the antenna system and so changes the system's resonant frequency) results in a significant impedance change at frequencies near the anti-resonant frequency. Only a slight retuning of the matching circuit is necessary; the radiation patterns are affected relatively little. A $6^1/_2$-foot nonconductive rope connecting the top ends of the horizontal arm to the support towers is assumed. For 50-foot support poles, it matters little (for 160-meter antennas) whether the poles are metal or wood.

But this is not the case if the support tower is also the mount for a 20-meter Yagi, because the tower and Yagi are much more nearly resonant than the tower alone. In this case, the tower carries a heavy current (0.8 A compared with the 1-A base current in the 1/4-λ inverted L). Because the phase of the tower current is +145° with respect to the phase of the current on the vertical element of the inverted L, the antenna's vertically polarized field is almost canceled. Also, the inverted L is no longer resonant: In fact, it is far from resonant; its impedance is 1547 –*j*2635 Ω, compared with the inverted L's impedance of 14 Ω with tree support.–*John S. Belrose, VE2CV*

Notes

[1] D. DeMaw, "The 160-Meter Antenna Dilemma," *QST*, Nov 1990, pp 30-32.

[2] *ELNEC* was used for all calculations. This program is available from Roy Lewallen, W7EL.

[3] *ELNEC* is a version of *MININEC*. *MININEC* assumes that the antenna is over a perfect earth for the purpose of calculating the antenna's impedance and the current distribution. The result of this approximation is that for horizontally polarized antennas, *MININEC* overestimates the gain of the antenna by an amount that increases as the height of the dipole decreases below a height of about 0.2 λ. The effect of real ground on the shape of the radiation pattern is correctly predicted, however, since for the calculation of the far-field pattern the program employs Fresnel reflection coefficients for the specified earth type.

By Brian L. Wermager, K0EOU **From *QST*, April 1986**

A Truly Broadband Antenna for 80/75 Meters

Have you been dreaming about an antenna that will do justice to your no-tune, solid-state transceiver by letting you operate across the entire band from 3.5 to 4.0 MHz? Then this may be the antenna for you!

With declining sunspots and poor conditions on the higher HF bands, 80 meters has suddenly become very popular. But, unfortunately, many hams are not able to use this band to its full potential. It offers every kind of ham activity from CW to phone, from nets and ragchewing to great DXing, but many hams are too limited by the frequency range of their antennas to enjoy this band completely.

Antenna-matching networks are one answer, but they spoil the advantage of the no-tune feature of modern transceivers. Matching networks also are often less effective than many hams think; they introduce losses. The losses can be significant at some settings which provide a match. With these things in mind, I decided to try some ideas that might give me a more broadbanded antenna. The prime requirement was that it be fed with common 50-ohm coaxial cable, with no traps, coils or capacitors.

First, I tried a quarter-wave sloper. This antenna worked very well, with a bandwidth of 300 kHz between the 2:1 SWR points. It still, however, limited me from operating CW DX at the bottom of the band and the phone nets at the top of the band. There had to be a better antenna.

Antennas can be broadbanded by using large-diameter elements. With this in mind, I began experimenting with two-wire slopers, attached to a common feed point, but with the wire ends fanned out from each other. (See Figure 1.) This seemed to help, but not as much as I had hoped. It did, however, shorten the length required for the sloper. For those with a short tower, this idea could make an 80 or 160-meter sloper possible when a single-wire sloper would be too long.

The Fickle Finger of Fate Strikes!

While I was trying one of these two-wire antennas at a low height on my tower, the SWR was less than 2:1 from 3.5 to 4.0 MHz! After several attempts to get it to work the same way at the top of the tower, I discovered that these results could be attained only when my old quarter-wave sloper was at the top of the tower *and grounded* to the tower. (See Figure 2.) The two antenna elements were obviously interacting with each other, broadening the bandwidth tremendously. Further pruning of the lengths of both the sloper and the two-wire element resulted in the amazing SWR curves shown in Figure 3.

What's Going On?

I will leave the question of *why* it works to the experts. (See the sidebar to this article.—Ed.) Like a true ham, I subscribe to the old saying, "If it works, leave it up and don't mess with it." My guess, however, is that it is something like one-half of a two-element log periodic seeking its mirror

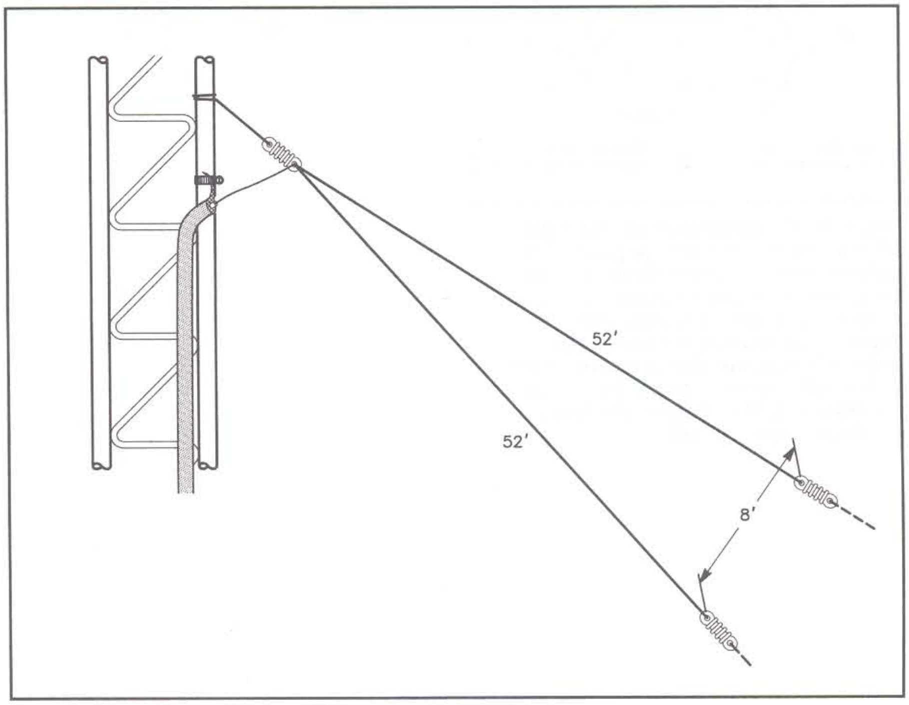

Figure 1—Arrangement of the original K0EOU experiment. The coaxial cable shield is connected to the tower, with both wires of the two-wire element connected to the center conductor. The wires are spread approximately 8 feet at the ends and are each approximately 52 feet long.

The MININEC Analysis of the KØEOU Three-Element Half Sloper

The Mini-Numerical Electronics Code (MININEC) analyzes thin-wire antennas, solving an integral equation representation of the electric fields using a method-of-moments technique. MININEC solves for the currents, impedance and patterns for antennas composed of wires in arbitrary orientations in free space and over perfectly conducting ground.

The impedance at the feed point calculates to be 79.4 + *j*1859.8 ohms. This assumes a perfect ground beneath the structure and simply a 3-foot extension of the tower above the connection of the upper wire. In practice, the impedance will be affected by both the ground conductivity and the top-loading effect of a beam antenna atop the tower.

A fair amount of current flows in the top section above the upper wire connection point. Top loading will affect the phase of this current, which will be reflected as a change in impedance at the feed point. In other words, the calculated data is not absolute. Use it as an approximation only.

A relatively high current flows at the base of the tower to ground—more than in any other part of the system. This indicates that a good earth connection, and even a radial system, would offer highest efficiency.

The antenna patterns, Figures A, B and C, are also approximations. Polarization is predominantly vertical–at low angles it may be considered to be almost completely vertical. Broadside to the direction of the wires, the polarization becomes horizontal at high radiation angles, ie, above 75 degrees. At 60-degrees elevation, the vertical component is almost 16 dB greater than the horizontal. The vertical component increases significantly at lower elevation angles, being in excess of 30 dB above the horizontal component at a 5-degree elevation. These figures all apply in a direction broadside to the wires. In the direction of the wires, both "front" and "back," the radiation is entirely vertically polarized.–*Gerald L. Hall, K1TD*

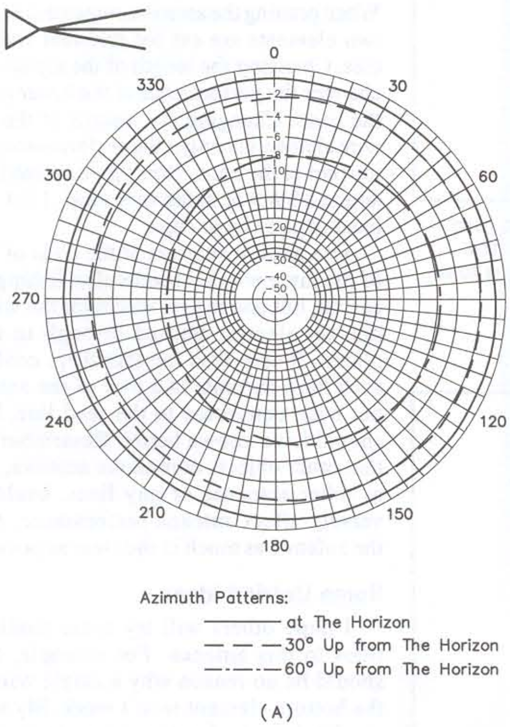

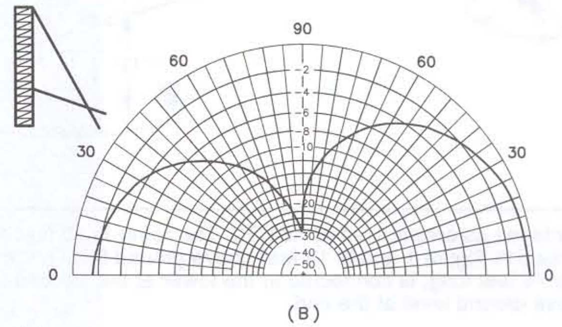

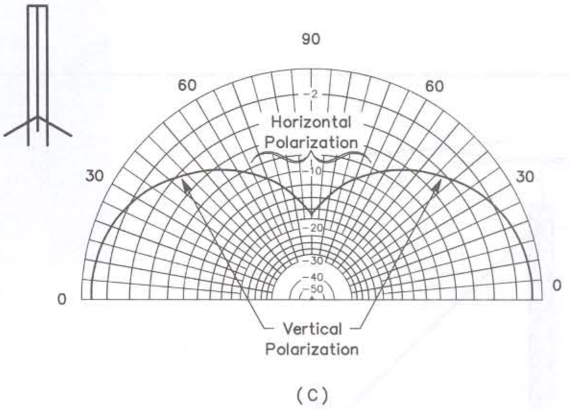

Figure A–Antenna azimuth radiation pattern for the KØEOU three-element half sloper antenna. Values are in dBi. Add 6.0 dB to the values shown.

Figure B—Antenna elevation radiation pattern, in the direction of the wires, for the KØEOU antenna. Values are in dBi. Add 6.0 dB to the values shown.

Figure C—Antenna elevation radiation pattern, in a direction broadside to the wires, for the KØEOU antenna. Values are in dBi. Add 6.0 dB to the values shown.

image in the grounded tower. The top element is tuned for the lower portion of the band and the two-wire element for the upper portion. In fact, there is a little SWR "bump" in the middle of the band that seems to give further evidence of this.

How Well Does It Work?

Although I have no way of scientifically plotting the antenna pattern, it does seem to be vertically polarized. Good DX performance from the antenna seems to verify this. Because many contacts have been made in all directions, the antenna probably has a fairly omnidirectional pattern. On-the-air comparisons with a quarter-wave sloper across town show that the antenna performs at least as well as the sloper. It also seems to have a little less noise on receive than the sloper.

Getting One Up for Yourself

If you have a tower over 40 feet high, you should be in business. The element dimensions will vary according to the height of your tower. My friend Kelly Davis, KD7XY, constructed one of these antennas on his 50-foot tower so we could see how the dimensions would change.

Measurements of his antenna are shown in Figure 4.

It is interesting to note that the height of the feed point on the tower does not appear to be critical at all. The angle of the wires in the two-wire element does not seem to be critical either. The sloper element, however, should come down *between* the two wires of the lower element. The sloper angle is about 45 degrees. As you attempt to get the

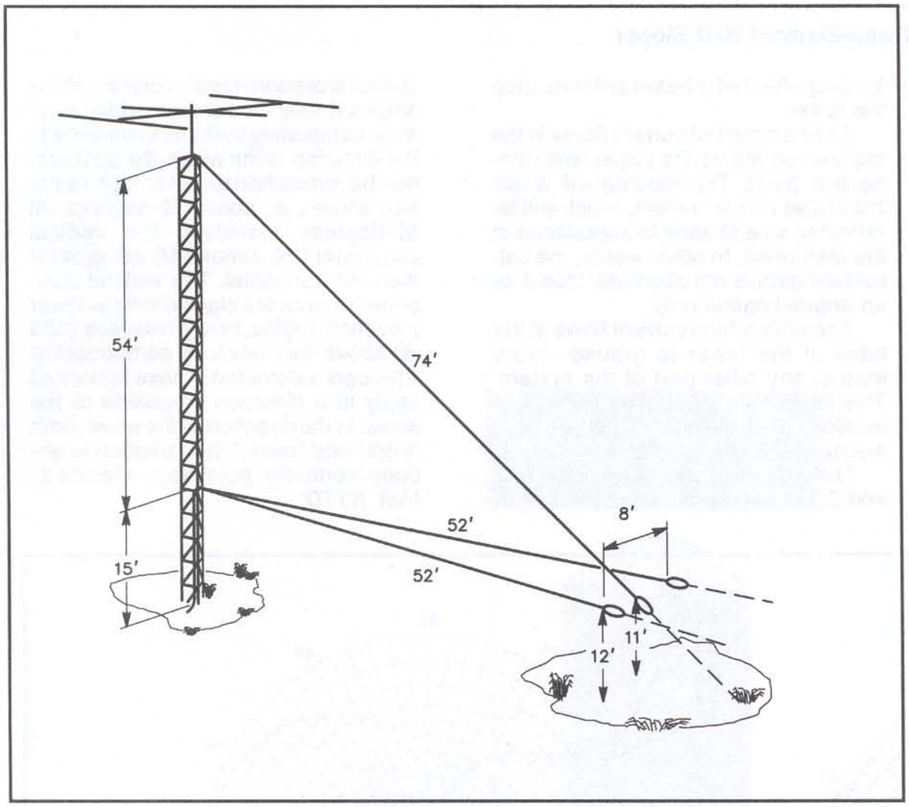

Figure 2—The antenna as constructed at KØEOU. The tower is 70 feet high. The feed point is as described in Figure 1 and is 15 feet above ground level on the tower. The sloper element is 74 feet long, is connected to the tower at the top end and slopes to a point 11 feet above ground level at the end.

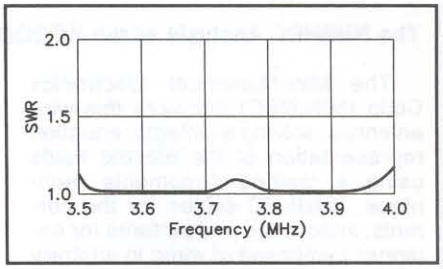

Figure 3—SWR measurements for the antenna at KØEOU. The highest SWR measurement between 3.5 and 4.0 MHz is 1.2:1.

lowest possible SWR from the antenna, remember that the angle of the sloper to the tower and its distance from the ground at the end will have an effect on the bandwidth. Because these antennas can be pruned from the ground, the trial-and-error method is easy. When pruning the antenna, remember that the two elements are cut for different frequencies. Changing the length of the top element changes the performance at the lower part of the band. Changing the lengths of the two-wire elements changes the performance at the top part of the band. Don't give up until your antenna SWR is, at the very most, 1.5:1 from band edge to band edge.

Use good insulators at the ends of each of the three wires. I used nylon fishing line at first, but one foggy, wet night the wire in the top element burned through in three places. It surprised me that there could be such high currents in a part of the antenna not even connected to the feed line, but I should have known better. Remember also that other objects around the antenna, such as other antennas or guy lines, could adversely affect antenna performance. Keep the antenna as much in the clear as possible.

Some Untried Ideas

I hope others will try some modifications to this antenna. For example, there should be no reason why a single wire for the bottom element won't work. My small city lot doesn't give me room to try a single wire, as it would surely need to be longer than the two-wire element. I would also like to see someone try cutting the top element for the high end of the band, and the bottom element for the low end. This could shorten the sloper element for someone with a shorter tower and may even give the antenna some gain in the direction of the two-wire element. It might also be possible to construct an antenna for another band inside of this one (40 meters, for example). The same feed point could be used, but with another sloper element for the second band. Another possibility is to construct the antenna with two dipoles. It is my guess (and only a guess), that it is the merging of the ends of the elements that causes the 50-ohm impedance of the antenna.

Winter gives us the best conditions on the 80/75-meter band. You can be ready to use the whole band with this simple-to-construct and very broadband antenna.

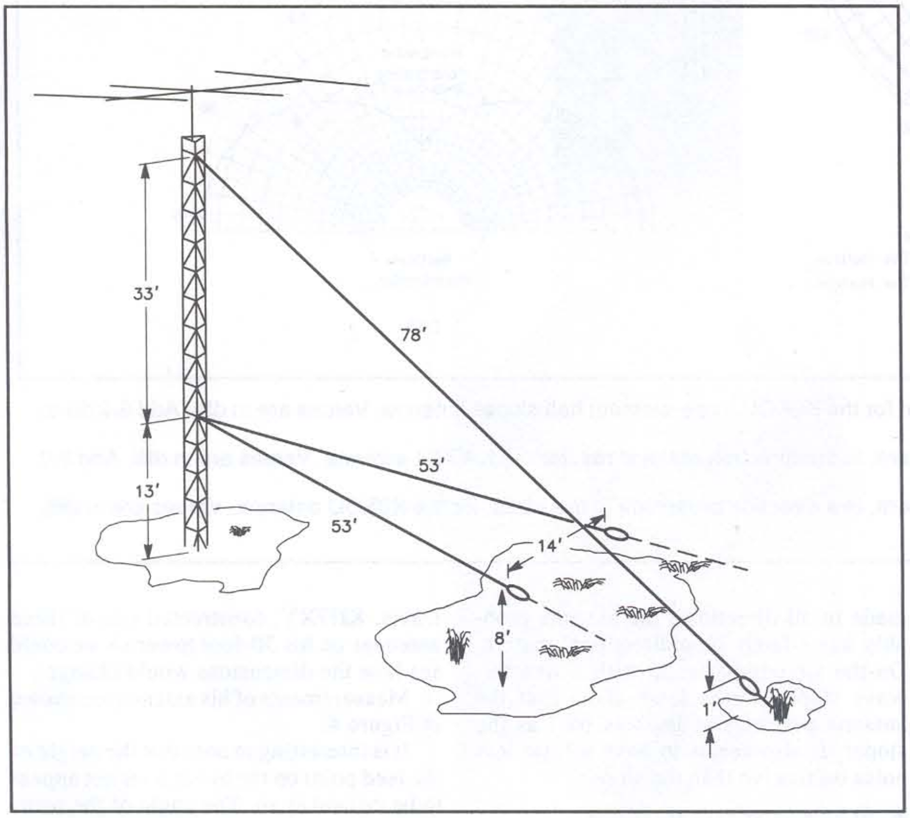

Figure 4—The antenna installation at KD7XY. The tower is 50 feet high. The sloper element is attached at 46 feet, is 78 feet long, and is only 1 foot above ground level at the end. The two-wire element is 53 feet long and is attached to the tower at 13 feet above ground level. Like KØEOU's antenna, it is virtually "flat" across the band; highest measured SWR is 1.4:1.

From *QST*, August 1987 (Technical Corespondence)

Improved Broadband Antenna Efficiency

When Brian Wermager's article. "A Truly Broadband Antenna for 80/75 Meters," arrived at HQ, I wanted to see if the antenna worked as well as claimed before publishing the information in *QST*.[1] I stopped tuning the antenna when the SWR was less than 1.6 across the entire band.

For working DX the antenna seemed to work as well as, or better than, a dipole at 50 feet. Casual contacts were not difficult to make, but in contests, it was a different matter. The results were satisfactory, but could "satisfactory" be changed to "outstanding"?

The key to improving the efficiency of this antenna is found in the sidebar accompanying the article. The MiniNumerical Electronics Code (MININEC) computer analysis done by Gerald Hall, K1TD, shows that: "A relatively high current flows at the base of the tower to ground—more than in any other part of the system. This indicates that a good earth connection, and even a radial system, would offer highest efficiency."

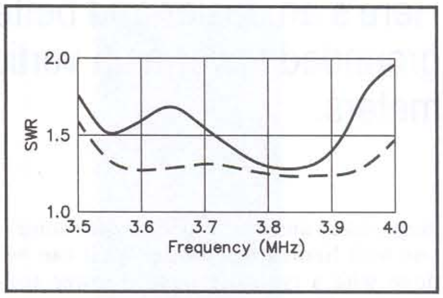

Figure 1–Graph showing SWR vs frequency for the K0EOU 80/75-meter antenna. The lower curve depicts results before adding ground radials; the upper curve was plotted after adding radials and retuning the antenna.

A system of radial wires improves antenna efficiency by cutting ground resistance losses. The decreased resistance in ground losses will also narrow the antenna bandwidth. The question now becomes: Is the narrowing of the bandwidth too much? I decided to investigate that question next.

With help from my brother-in-law, I installed 42 radials around, and bonded to, the tower base. As expected, SWR increased across the entire band. After I retuned the antenna, the SWR was below 2:1 across the entire band (see Figure 1). As for results, DX stations are typically 2 to 6 S units stronger on the K0EOU antenna than on the dipole at 50 feet. Contest QSOs are now a reality. I am not the loudest East Coast station on the band, but I am very pleased with the performance of the broadband antenna. (The dipole has been removed and stored!)—*Chuck Hutchinson, K8CH*

[1] B. Wermager. "A Truly Broadband Antenna for 80/75 Meters," *QST*, Apr 1988, pp 23-25.

By Thomas Russell, N4KG

From *QST*, June 1994

Simple, Effective, Elevated Ground-Plane Antennas

Here's an easier and better way to use your grounded tower as a vertical antenna on 160 or 80 meters.

This article describes a simple and effective means of using a grounded tower, with or without top-mounted antennas, as an elevated ground-plane antenna for 80 and 160 meters.

Grounded towers have been used as shunt-fed verticals on the low-frequency amateur bands for many years. Generally. they required a gamma- or omega-type matching network and an extensive radial system for efficient operation. Recent computer studies reveal that simple elevated radial systems consisting of only four wires can produce results equivalent to 120 buried radials. Typically, these antennas are modeled as isolated monopoles. Presumably, grounded towers could be used with an appropriate shunt-fed matching network. I've found an even easier method!

From Sloper to Vertical

Recall the quarter-wave length sloper, also known as the half-sloper. It consists of an isolated quarter wavelength of wire, sloping from an elevated feedpoint on a grounded tower. Best results were usually obtained when the feedpoint was somewhere below a top-mounted Yagi antenna. You feed a sloper by attaching the center conductor of a coaxial cable to the wire and the braid of the cable to the tower leg. Now, imagine four (or more) slopers, but instead of feeding each individually, connect them together to the center conductor of a single feed line. *Voila*! Instant elevated ground plane.

Now, all you need to do is determine how to tune the antenna to resonance. With no antennas on the top of the tower, the tower can be thought of as a fat conductor and should be approximately 4% shorter than a quarter wavelength in free space. Calculate this length and attach four insulated quarter-wavelength radials at this distance from the top of the tower. For 80 meters, a feedpoint 65 feet below the top of an unloaded tower is called for, The tower guys must be broken up with insulators for all such installations. For 160 meters, 130 feet of tower above the feedpoint is needed.

That's a lot of tower to dedicate to a single-band antenna, especially for someone with limited real estate. What can be done with a typical grounded-tower-and Yagi installation?

A top-mounted Yagi acts as a large capacitance hat, top loading the tower. Fortunately, top loading is the most efficient means of loading a vertical antenna. The amount of loading can be approximated by using an empirical formula developed by John Devoldere, ON4UN.[1]

Devoldere found that the electrical height of a top-loaded tower can be approximated by:

$$L=0.38F(H+\sqrt{2S-H/500}) \qquad \text{(Eq 1)}$$

where

L is the approximate electrical length in degrees
F is the frequency in MHz
H is the height of the tower under the Yagi in feet
S is the area of the Yugi in square feet

Table 1
Effective Loading of Common Yagi Antennas

Antenna	*Boom Length (feet)*	*S (area, ft²)*	*Equivalent Loading (feet)*
3L 20	24	768	39
51 15	26	624	35
4L 15	20	480	31
31 15	16	384	28
5L 10	24	384	28
4L 10	18	288	24
3L 10	12	192	20
TH7	24	—	40 (estimated)
TH3	14	—	27 (estimated)

To check Eq 1, consider the case of no antenna on top, where S = 0. Then L = 0.38 × 3.6 × 65 = 88.9°, which is very close to the desired 90° quarter wavelength.

The effective loading of a Yagi is the portion of the equation under the radical. The examples in Table 1 should give us an idea of how much top loading might be expected from typical amateur antennas. The term H/500 is ignored as insignificant compared with 2S.

The values listed in the Equivalent Loading column of Table 1 tell us the approximate vertical height replaced by the antennas listed in a top-loaded vertical antenna. To arrive at the remaining amount of tower needed for resonance, subtract these numbers from the nonloaded tower height needed for resonance. Note that for all but the 10-meter antennas, the equivalent loading equals or exceeds a quarter wavelength on 40 meters. For typical HF Yagis, this method is best used only on 80 and 160 meters.

Construction Examples

Consider this example: A TH7 Yagi mounted on a 40-foot tower. The TH7 has approximately the same overall dimensions as a full-sized 3-element 20-meter beam, but has more interlaced elements. I estimate its equivalent loading to be 40 feet. At 3.6 MHz, 65 feet of tower is needed without loading. Subtracting 40 feet of equivalent loading, the feedpoint should be 25 feet below the TH7 antenna.

I ran 10 quarter-wavelength (65-foot) radials from a nylon rope tied between tower legs at the 15-foot level, to various

---

[1] J. Devoldere, *Antennas and Techniques for Low Band DXIng* (Newington: ARRL, 1994).

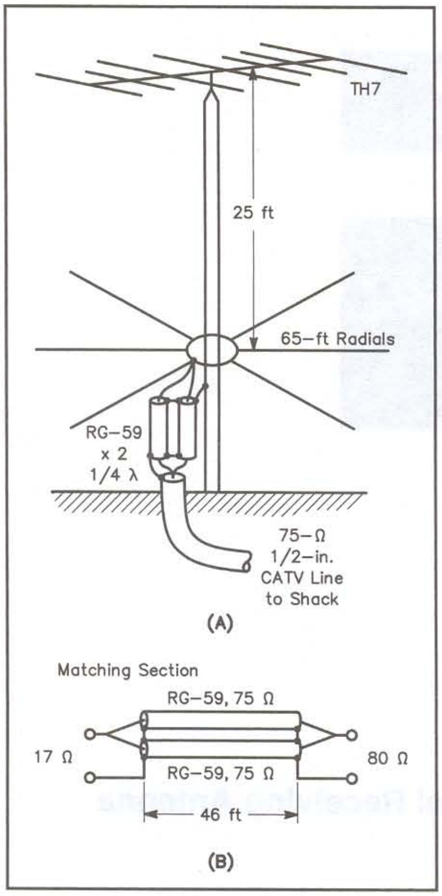

Figure 1—At A, an 80-meter top-loaded, reverse-fed elevated ground plane, using a 40-foot tower carrying a TH7 triband Yagi antenna. At B, dimensions of the 3.6 MHz matching network, made from RG-59.

supports 10 feet high. I tied nylon cord to the insulated, stranded, 18-gauge wire, without using insulators. The radials are all connected together and to the center of an exact half wavelength (at 3.6 MHz) of RG-213 coax, which will repeat the antenna feed impedance at the other end. Figure 1 is a drawing of the installation. I used a Hewlett-Packard low-frequency impedance analyzer to measure the input impedance across the 80-meter band.

An exact resonance (zero reactance) was seen at 3.6 MHz just as predicted. The radiation resistance was found to be 17 Ω. The next question is how to feed and match the antenna.

My approach to 80-meter antennas is to tune them to the low end of the band, use a low-loss transmission line, and switch an antenna tuner in line for operation in the higher portions of the band. With a 50-Ω line, the 17-Ω radiation resistance represents a 3:1 SWR, meaning that an antenna tuner should be in-line for all frequencies. For short runs, it would be permissible to use RG-8 or RG-213 directly to the tuner. Since I have a plentiful supply of low-loss 75-Ω CATV rigid coax, I took another approach.

I made a quarter-wave (70 feet × 0.66 velocity factor = 46 foot) 37-Ω matching line by paralleling two pieces of RG-59 and connecting them between the feedpoint and a run of the rigid coax to the transmitter. The magic of quarter-wave matching transformers is that the input impedance (R_i) and output impedance (R_o) are related by

$$Z_o^2 = R_i \times R_o \qquad \text{(Eq 2)}$$

For $R_i = 17\ \Omega$ and $Z_o = 37\ \Omega$, $R_o = 80\ \Omega$, an almost perfect match for the 75-Ω CATV coax. The resulting 1.6:1 SWR at the transmitter is good enough for CW operation without a tuner.

The Proof is In the Log

How effective is this antenna? Well, I used to install a 60-foot aluminum tower and 100 radials, 100 feet long in a clear one-acre field every winter, and remove it every spring for mowing. The top-loaded reverse-fed elevated ground-plane antenna has replaced that antenna with no regrets. My only other 80-meter antenna is a dipole at 110 feet broadside to Europe and the South Pacific,

I use the elevated ground plane for South Africa, South America. the Caribbean, parts of the Pacific and Asia, both long and short path. With it I have worked everything I can hear, with 1200 W output, including HL (rare in Alabama); HS; UA9; UA0; UI; UJ; UL; most of the VK9s; XV; ZS1; ZS8MI; ZS9; 3Y5X; 8Q; 9M2; and 9V1. While running 5 W output. I have even worked two JAs with this antenna. which may say more for its effectiveness than anything else.

Will It Work on 160 Meters?

You bet it will, but it takes another tower. For the 160-meter band, a resonant quarter-wavelength requires 130 feet of tower above the radials. That's a pretty tall order. Subtracting 40 feet of top loading for a 3-element 20-meter or TH7 antenna brings us to a more reasonable 90 feet above the radials. Additional top loading in the form of more antennas will reduce that even more.

Recently, a friend moved to the country, and he needed a 160-meter antenna in a hurry for an upcoming contest. He had stacked TH6s on a 75-foot tower. I suggested he try four elevated radials at 10 feet above ground, with a tuner if necessary. He connected four radials about 120 feet long and a piece of RG-58. The SWR measured under 2:1 and he worked everything he heard in the contest. Figure 2 is a drawing of this installation.

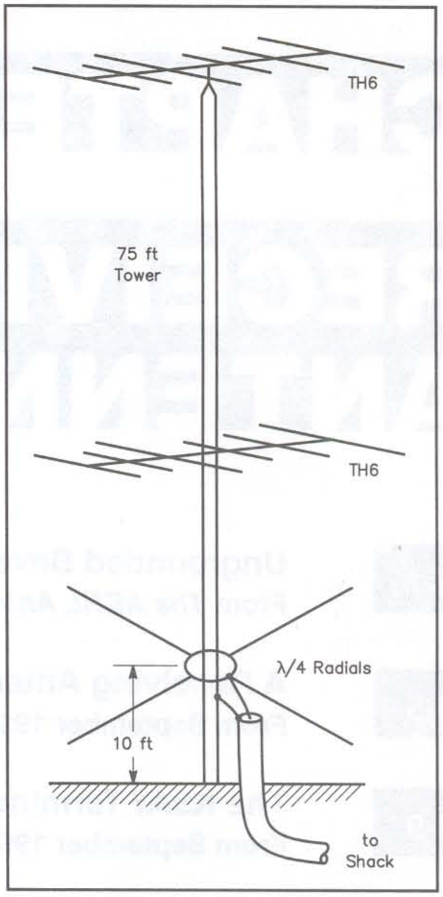

Figure 2—A 160-meter antenna using a 75-foot tower carrying stacked triband Yagis.

Another friend had a 120-foot tower with no antennas on it. He ran four elevated radials at 10 foet and obtained an SWR below 1.5:1 with a 50-Ω feed line. During the contest, he even beat out some big guns.

Elevated ground-plane antennas work! This simple, reverse-feed system makes it possible to feed grounded towers easily and efficiently.

References

A. Christman, "Elevated Vertical Antenna Systems," *QST*, Aug 1988, pp 35-42 (Feedback, *QST*, Oct 88, p 44).

A. Christman, "More on Elevated Radials," *QST*, Mar 1993, p 72 (Technical Correspondence).

CHAPTER EIGHT

RECEIVING ANTENNAS

8-1 **Ungrounded Beverage Antennas**
From *The ARRL Antenna Compendium Volume 4*

8-5 **A Receiving Antenna that Rejects Local Noise**
From September 1995 *QST*

8-9 **The K9AY Terminal Loop–A Compact, Directional Receiving Antenna**
From September 1997 *QST*

8-13 **Hum Problems When Switching the K9AY Loops**
From May 1998 *QST* (Technical Correspondence)

By Brian Beezley, K6STI **From *The ARRL Antenna Compendium Volume 4***

Ungrounded Beverage Antennas

So you've heard how Beverages can help you hear better on the lower bands, but you're still not convinced they're worth the bother. K6STI analyzes the Beverage, especially an easy-to-construct "ungrounded" version.

Beverages are amazing receiving antennas. A good Beverage can turn a signal barely audible on your transmit antenna into solid, Q5 copy. A Beverage antenna (named after 1920s inventor H. H. Beverage) is a long, low wire used for receiving. They're most popular on the 160, 80, and 40-meter bands. Although occasionally pressed into service, Beverages are much too lossy for regular use as transmitting antennas.

Conventional Beverages are fed against ground. Although very long wires may develop a few dB of front-to-back ratio, unterminated Beverages are essentially bidirectional. Terminating a Beverage through a resistance to ground at the far end creates a highly unidirectional pattern. Beverages typically are 1 to 15 feet high and 1 to 4 wavelengths long.

Beverages are broadband, traveling-wave antennas. Power traveling toward the feedpoint is absorbed by the receiver, while that traveling in the opposite direction is absorbed by the termination resistance. Standing waves can't develop. Because waves traveling toward the feedpoint accumulate substantially in phase, while those arriving from other directions tend to cancel, the azimuth pattern of a long Beverage is highly directive. Beverages are nonresonant antennas and can be used over a wide frequency range.

A very useful Beverage variation uses two parallel wires, with one grounded at the far end. Waves traveling toward the far end reflect and return to the feedpoint differentially as transmission-line currents. With suitable termination and switching circuitry at the feedpoint, you can receive signals from either direction. A two-wire Beverage thus can replace two single-wire antennas.

For simplicity, I'll consider just single-wire Beverages here.

Eliminating Ground Connections

Beverages normally connect to ground at the feedpoint and far end. See Figure 1A. However, obtaining a good, reliable ground can be difficult. A short ground rod may exhibit thousands of ohms resistance to ground when installed in dry soil or when corroded. Since Beverages require a termination resistance of just several hundred ohms, a poor far-end ground can grossly misterminate an antenna and destroy its pattern. A poor feedpoint ground can reduce signal levels and encourage coax-shield pickup. You can use multiple ground rods, a ground screen, or a ground-radial system to lower ground resistance, but this greatly increases the work required to install a Beverage.

A ground connection isn't essential for Beverage operation. You can place the termination resistance about one-quarter wavelength from the far end of the wire. Since the impedance of a quarter-wave wire is low and resistive, the remainder of the wire just sees the termination resistor in series with a low resistance. Similarly, you can feed a Beverage about one-quarter wavelength from the near end. Both ends of the wire are left unconnected. See Figure 1B.

Ungrounded Beverages are simple to install and have stable properties. Their radiation patterns often are superior to those of their grounded counterparts because they don't suffer unwanted response from vertical wires. (This response can be reduced by lowering wire height. One way to do this is to slant the feed and termination wires over a long distance.)

The only real disadvantage of ungrounded Beverages is that they're frequency-dependent. For example, the low-impedance, quarter-wave wire sections become high-impedance, half-wave sections at twice the design frequency. I'll show you how to multiband an ungrounded Beverage later.

Antenna Modeling

While you can use a *MININEC*-based antenna- analysis program to model Beverage antennas, you'll have to settle for approximate results at best. The *MININEC* algorithm uses perfect-conductivity ground during current calculations. It takes lossy ground into account only later when calculating radiation patterns. Because Beverages are so close to ground, their current distribution is strongly affected by ground characteristics. *NEC*'s Sommerfeld-Norton ground model takes ground dielectric constant and conductivity into account when calculating wire currents. This lets you accurately analyze antennas in close proximity to earth. I used the *NEC/Wires 1.0* program to obtain all results presented here. I modeled antennas over average-quality earth (dielectric constant= 13, conductivity = 5 mS/m).

Wire Height

Figures 2A and 2B shows the azimuth and elevation patterns for a 1000-foot long, 10-foot high, ungrounded Beverage on 160 meters. The antenna uses # 18 wire and a 650-Ω termination resistance. It's fed and terminated 120 feet from each end. The remarkable pattern is better than that of many HF Yagis.

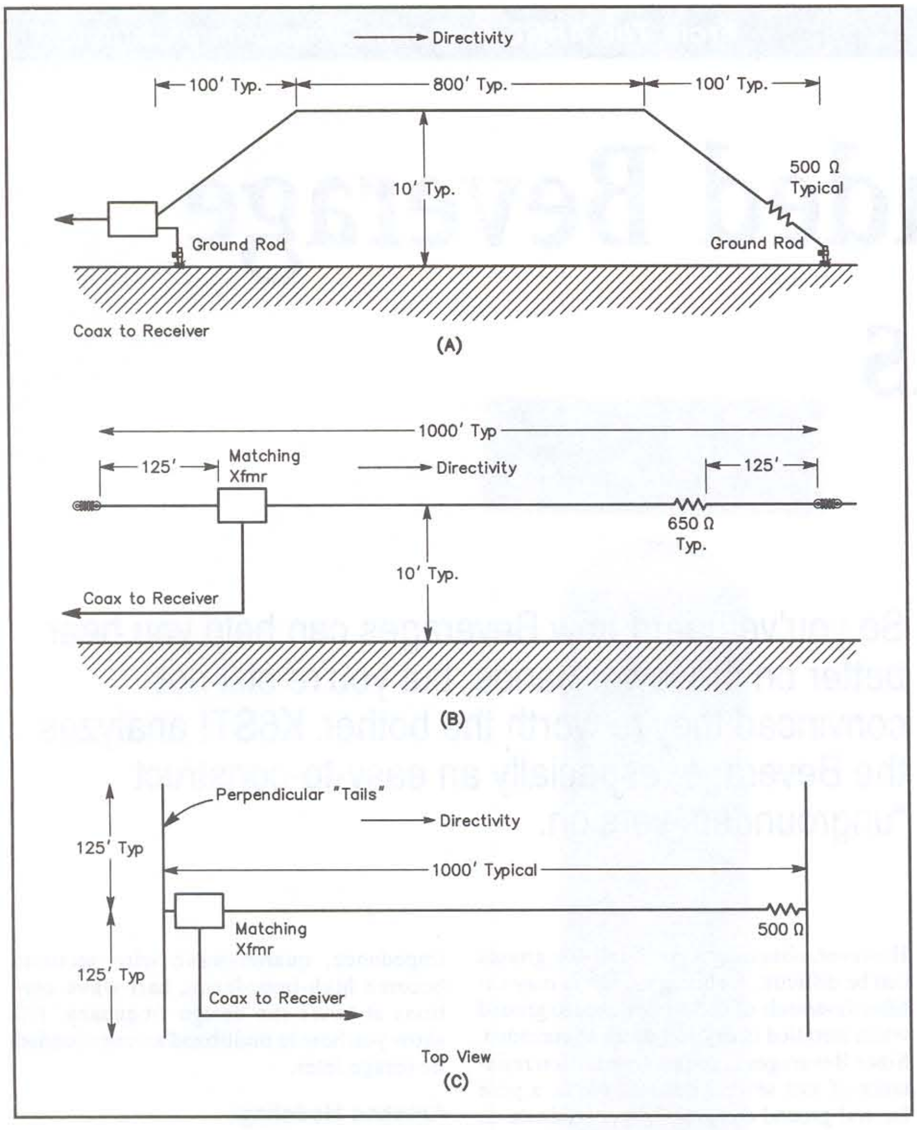

Figure 1—Single-wire Beverage antenna feeding and termination configurations. At A, conventional method is shown using slanted-wire terminations. At B, ungrounded Beverage is shown with matching transformer and terminating resistor inserted λ/4 in from either end of wire. At C, ungrounded Beverage using full span of wire by employing perpendicular "tails" at both ends. The tails function as elevated counterpoise radials at 160 meters.

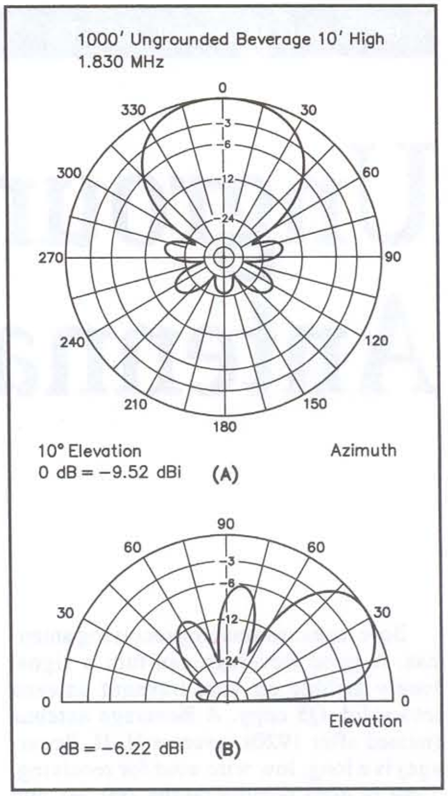

Figure 2—Azimuth and elevation patterns at 1.83 MHz for 1000-foot long ungrounded Beverage mounted 10 feet off flat ground with average conductivity and dielectric constant. The worst-case rearward lobes are better than 18 dB down compared to peak response. Note that the gain is only −9.52 dBi at 10° elevation, peaking at −6.22 dBi at about 30° elevation. If the transmission line to the receiver is very long, a preamplifier may be needed.

Figure 3 shows the azimuth pattern of a similar Beverage 1 foot above ground. The termination resistance for this antenna is 750 Ω and it's fed and terminated 125 feet from the wire ends. The pattern is even sharper than that of Figure 2. Output is about 1 dB lower.

If you're tempted as I was to just lay a Beverage on the ground, take a close look at Figure 4. This is the Beverage of Figure 3 but just 1 inch off ground (*NEC* provides accurate results for wire heights down to within several wire radii of ground). While still useful, the pattern has degraded considerably. Signal output is quite a bit lower. (Nevertheless, some users report excellent results from on-the-ground Beverages. Don't hesitate to try one if it's your only option.)

In the West you can just lay an ungrounded Beverage in the chaparral a few feet off ground. In the East you can unroll one on top of a foot or two of snow (snow has very low dielectric constant and conductivity and should be transparent to RF at low frequencies). Beverages like these are very easy to install (no support poles or ground rods). You can roll one out just before a contest.

Occasionally a Beverage may have to clear a height-sensitive area. Figure 5 shows what can happen when wire height varies. The first 200 feet of this antenna is 3 feet high (chaparral height). Next, a 50-foot section slopes from 3 to 15 feet. This is followed by a 200-foot span 15 feet high, a 50-foot downslope, and finally the remainder of the wire at 3 feet. This model is similar to a Beverage used by K6TQ which must clear an agricultural area with produce trucks. While the rear part of the pattern is relatively unaffected, the height variation causes a large sidelobe bulge. It doesn't matter much if your Beverage is

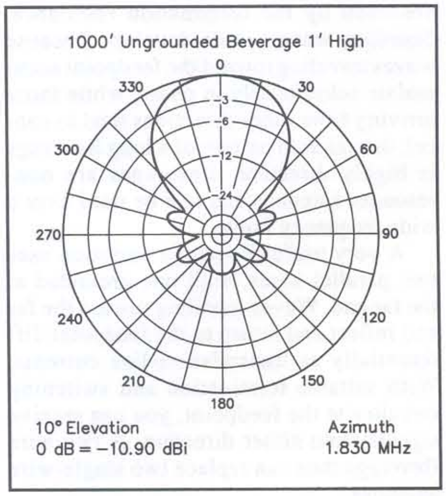

Figure 3—Azimuth response for 10° elevation angle at 1.83 MHz for 1000-foot long ungrounded Beverage mounted only 1 foot off flat ground. Compared to the higher antenna in Figure 2, this low Beverage has about 1 dB less gain, but the worst-case backlobes are suppressed even better.

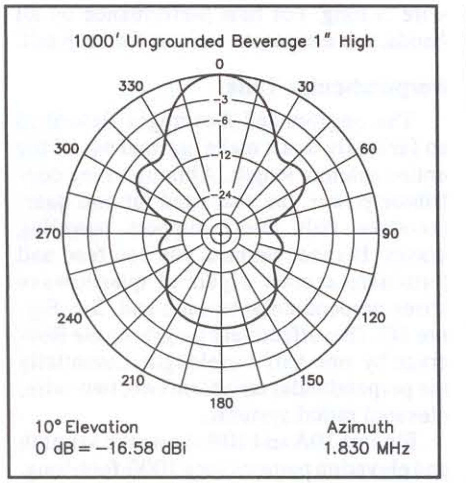

Figure 4—Azimuth response for same antenna as in Figure 3, except that the height is now only 1 inch off ground. The rearward pattern suffers and the gain fails about 5.4 dB compared to the Beverages in Figures 2 and 3.

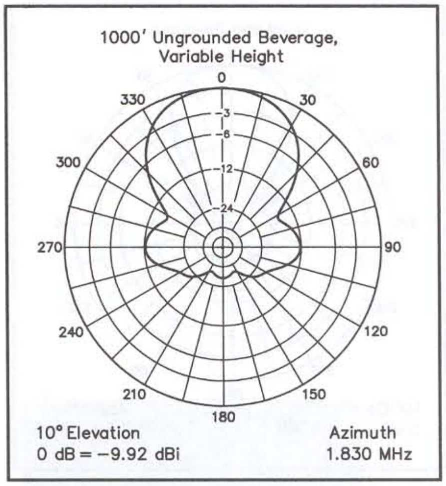

Figure 5—Azimuth response at 1.83 MHz for 1000-foot long, variable height ungrounded Beverage. The sidelobes at 90° and 270° are larger than the antenna in Figure 2, but the pattern and gain are still quite acceptable.

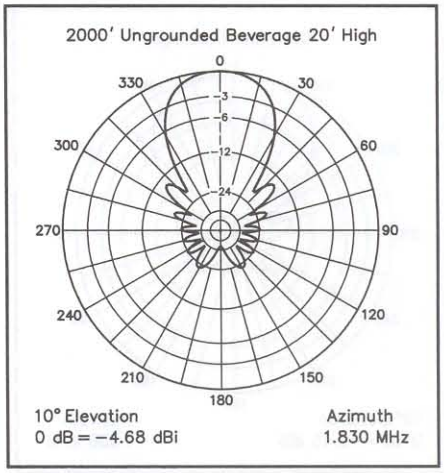

Figure 6—Azimuth response at 1.83 MHz for 2000-foot long ungrounded Beverage mounted 10 feet over flat ground. The gain is almost 5 dB higher than that of Figure 2, and the backlobes are all suppressed in excess of 23 dB. This is a great 160-meter receiving antenna, but must be oriented carefully at desired receiving bearing because of narrow frontal lobe.

high or low, but try to keep the height as constant as possible.

Wire Length

Figure 6 shows the azimuth pattern of a 2000-foot ungrounded Beverage. The termination resistance is 550 Ω. The pattern of a very long Beverage like this may be too narrow for complete coverage of a general direction.

Figure 7 shows the pattern of a 500-foot ungrounded Beverage. The termination resistance is 1100 Ω. While much broader than the patterns of longer antennas, it's still quite directive and can provide a substantial receive improvement.

Feeding Ungrounded Beverages

Conventional Beverages usually are fed with an autotransformer having a 3:1 turns ratio. This matches a 450-Ω impedance to 50 Ω. Ungrounded Beverages typically have an input resistance of 500 to 1000 Ω (roughly equal to the termination resistance) and an input reactance of up to 100 Ω. An exact impedance match isn't important; a better match just delivers a higher signal. Although usually not required to improve signal-to-noise ratio, a receive preamplifier can equalize Beverage and transmit-antenna signals for easier direct comparisons.

Both grounded and ungrounded Beverages can be sensitive to unwanted coupling at the feedpoint. If you connect coax through an autotransformer to a Beverage, the outside of the coax shield can become part of the antenna system if it exhibits an impedance similar to or less than whatever connects to it. It's the same thing that can happen with a coax-fed dipole. To prevent this, use a matching transformer with separate primary and secondary windings and low interwinding capacitance. This will isolate the outside of the coax shield from the antenna. Alternatively, you can use an autotransformer followed by a current-type balun. (Because the feed tail of an ungrounded Beverage is likely to provide a lower-impedance termination than a typical ground rod, these Beverages should be less sensitive to coax-shield pickup.)

Terminating Ungrounded Beverages

Optimal termination resistance depends on wire length, height, diameter, and ground constants. The termination resistance normally just affects the rear part of the pattern. It's easy to find the optimal value by modeling the exact geometry of your antenna with *NEC*. If you use long spans of wire with considerable sag, model the sag using multiple wires.

You can determine optimal termination resistance experimentally using a signal source to the rear of the antenna. Simply adjust the termination resistance for minimum signal.

(Actually, minimum response directly to the rear isn't quite optimal for most designs. This condition usually causes the pattern to degrade slightly over a broad region elsewhere. Nevertheless, adjusting for minimum response is close enough for all but the most fanatical Beverage enthusiast.)

Multibanding Ungrounded Beverages

If you provide a termination resistance one-quarter wavelength from the far end for each band of interest, you can use a Beverage on several bands. At the near end use a separate quarter-wave tail for each band and connect them in parallel at the feedpoint.

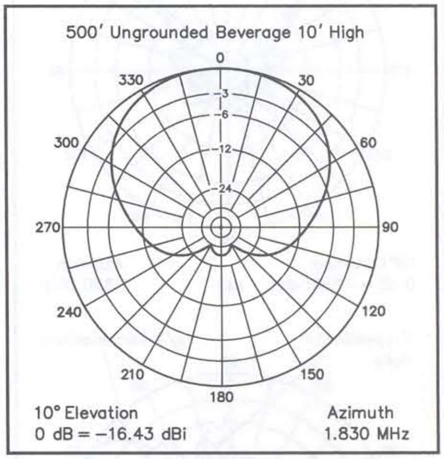

Figure 7—Azimuth response at 1.83 MHz for 500-foot long ungrounded Beverage mounted 10 feet over average ground. The gain is almost 7 dB less than Figure 2 at 10° elevation. The side lobes are large, even though the rear lobes are well suppressed.

Figure 8 and Figure 9 show 160-meter and 80-meter patterns of a 1000-foot-long, 10-foot-high, dualband Beverage. This antenna uses a 700-Ω resistor 63 feet from the far end and a 350-Ω resistor 125 feet out. The two termination resistances provide near-optimum performance on 160. The backlobe of the 80-meter pattern is suboptimal but the sidelobes are still well down. I didn't experiment much with termination-resistance values or positions. I'll bet both can be improved.

Another alternative is to use a single termination resistance with multiple quarter-wave tails. The resistor value will be optimal on just one band since antenna length isn't constant in wavelengths. But the value won't be far off on any band if the

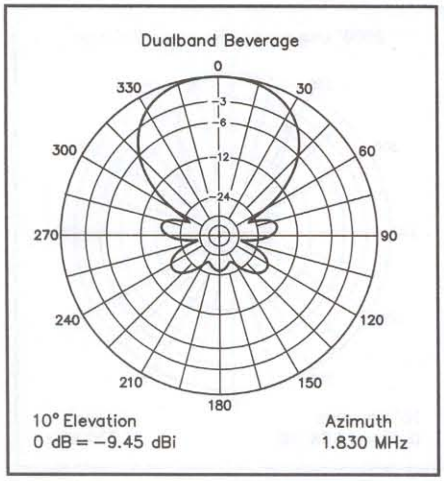

Figure 8—Dualband Beverage on 1.83 MHz, using a 700-Ω resistor 63 feet from the far end, and a 350-Ω resistor 125 feet out.

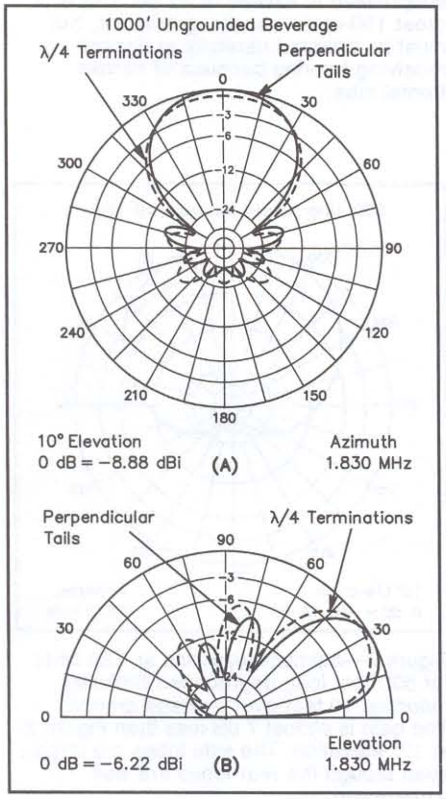

Figure 10—Ungrounded Beverage, 1000 feet long with two λ/4 perpen-dicular termination "tails." Pattern from ungrounded Beverage in Figure 2 is overlaid for comparison. At A, the azimuthal patterns at 10° elevation are shown; at B, the elevation patterns are shown.

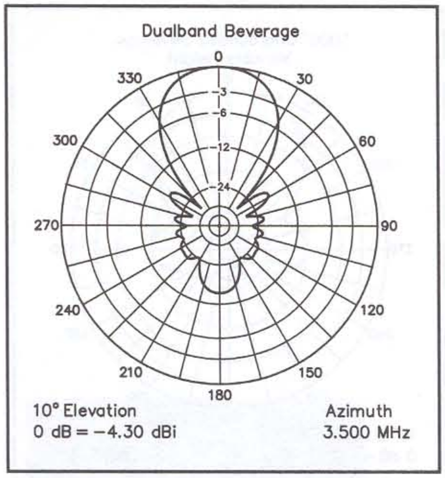

Figure 9—Dualband Beverage on 3.5 MHz, using same terminations as in Figure 7. The termination is not optimal for this band, but still yields desirable directivity.

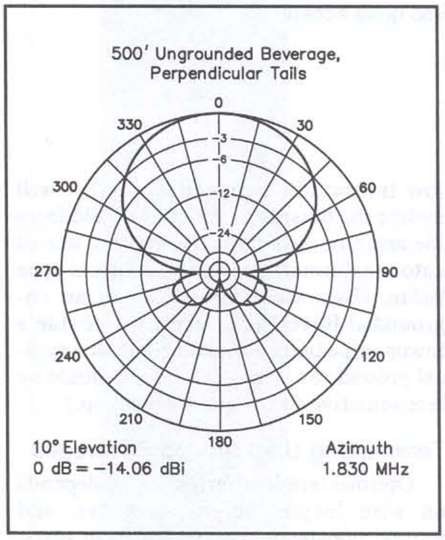

Figure 11—Azimuthal pattern for ungrounded Beverage, 500 feet long with two λ/4 perpendicular termination tails.

wire is long. For best performance on all bands, use a separate resistor for each tail.

Perpendicular Tails

The ungrounded Beverages described so far really don't make optimal use of the entire antenna length. Although they contribute power, the near- and far-end quarter-wave tails don't support traveling waves. If space permits you can feed and terminate against a pair of quarterwave wires perpendicular to each end. See Figure 1C. This effectively lengthens the Beverage by one-half wavelength. Essentially the perpendicular tails form two, two-wire, elevated radial systems.

Figures 10A and 10B shows the azimuth and elevation patterns for a 1000-foot-long, 10-foot-high Beverage with perpendicular tails, with the patterns from Figure 2 overlaid for comparison. This antenna uses a 500-Ω termination. The pattern is somewhat narrower than that of Figure 2 and the backlobes are even smaller. The elevation plot shows that overhead response is about 10 dB less with perpendicular tails and that the forward lobe cants at a lower angle.

Since the quarter-wave tails of a 500-foot Beverage comprise half its length, you might expect a more dramatic improvement for perpendicular tails on this antenna. Figure 11 shows the pattern of a 500-foot-long, 10-foot-high Beverage with perpendicular tails. This antenna also uses a 500-Ω termination. Side response is down quite a bit from Figure 7 and forward response is 2 dB greater.

While they provide better performance, designs with perpendicular tails are more complex and occupy additional space. You can make better use of the additional space by phasing two Beverages with in-line tails to obtain a much better azimuth pattern . . . but that's another story.

By Brian Beezley, K6STI

From *QST*, September 1995

A Receiving Antenna that Rejects Local Noise

Simplicity and performance combine to give birth to a compact antenna you'll want to have!

Noise can make a ham's life miserable on any amateur band. As we approach the minimum of the sunspot cycle, many hams are discovering that noise can be particularly frustrating on the low bands. In summer, static crashes caused by thunderstorm lightning can totally mask weak signals on the 160, 80, and 40-meter bands. During other seasons, power-line noise, noise from household appliances, and incidental radiation from home electronic products often limits reception.

A recent *QST* article by Floyd Koontz, WA2WVL,[1] describes a small receiving antenna for the low bands that provides a cardioid directional pattern. This pattern can reduce noise and QRM from the rear. As I marveled at the elegance and simplicity of Floyd's design, I realized that the antenna did have one shortcoming: Because it is vertically polarized, the antenna responds strongly to local noise propagated by ground waves. I wondered whether it was possible to devise a receiving antenna to better reject local noise.

The Ground Wave

Most hams who operate HF are familiar with the sky wave (or space wave) that's responsible for long-distance ionospheric propagation. See Figure 1A. The space wave has two components: The *direct wave* propagates along a straight line from the transmit antenna toward the ionosphere. The *ground-reflected wave* bounces off the earth's surface and heads in the same direction.

The space wave also exists for local propagation, as shown in Figure 1B. The direct wave travels in a straight line between the transmit and receive antennas, while the ground-reflected wave takes a midpoint bounce. But when the antennas are close to ground, the direct and reflected waves nearly cancel, leaving a very small residual space wave. When the antennas are right at the earth's surface, the waves cancel completely. So what makes local communication possible? Answer: A third wave, called the *surface wave*, that exists for antennas close to ground. This wave diminishes in intensity as you increase antenna height. The surface wave exists *only near the surface of the earth*. The combination of the direct wave, the ground-reflected wave, and the surface wave is called the *ground wave*.[2]

Surface-wave intensity varies with frequency and ground conductivity. It's stronger at low frequencies and for highly conductive ground. But the most important property of the surface wave is its polarization sensitivity. The surface wave is much weaker for horizontal fields. For example, on 80 meters the broadside ground-wave response of a short piece of wire is about 34 dB lower when oriented horizontally. The difference is about 42 dB at 160 meters.[3]

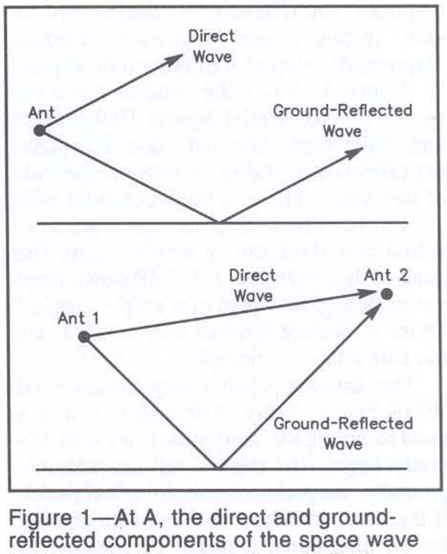

Figure 1—At A, the direct and ground-reflected components of the space wave for ionospheric propagation. At B, the space-wave components for local propagation.

The surface wave makes local AM radio broadcasting possible. Because of the poor propagation of horizontally polarized ground-wave signals, AM broadcasters universally use vertical polarization. The same phenomenon causes vertical antennas to pick up much more local noise than horizontal antennas do. Even if a noise source has a stronger horizontal component, by the time the field reaches the receive antenna, the vertical component almost always dominates.

These facts suggest that the first requirement of a receiving antenna with low response to local noise is insensitivity to vertically polarized radiation, the dominant component of the ground wave. Surprisingly, simply avoiding the use of vertical wires isn't enough. An antenna composed only of horizontal wires can still respond to vertical fields.

A Low Dipole

Figure 2 shows the ground-wave response of an 80-meter dipole 10 feet high. An easy-to-install, inconspicuous dipole like this is sometimes used for receiving when the transmit antenna is vertically polarized. The pattern shows the electric field strength 10 meters above ground at a distance of 1000 meters for an input power of 1 kW (the dipole exhibits the same pattern on receive). This geometry might be representative of that for a noisy power pole. Although the pattern may look similar to that of a free-space dipole, I think

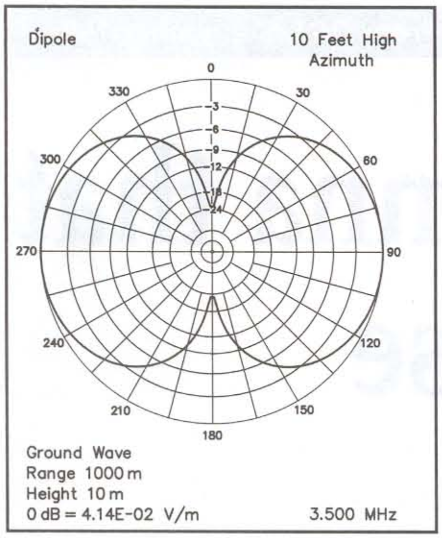

Figure 2—Azimuthal plot of the ground-wave response of a 10-foot-high 80-meter dipole. The input power to the antenna is 1 kW. The peak electric field is shown.

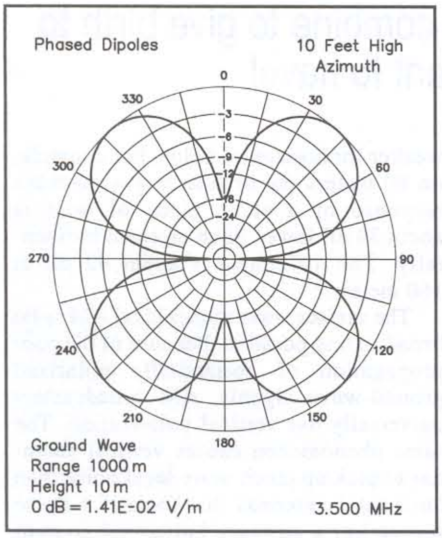

Figure 3—Azimuthal plot of the ground-wave response of two parallel dipoles 10 feet apart, 10 feet high and fed out of phase. The input power to the antenna is 1 kW. The peak electric field is shown.

you'll be surprised to know that the wire is oriented broadside to the pattern null. A low dipole actually responds to ground-wave fields best off its ends!

Here's an explanation for this peculiar behavior: The dipole has no response to the vertical component of a broadside ground wave because the electric field is perpendicular to the wire. The antenna responds only to the weak horizontal component. The vertical component also is perpendicular to the mirror image of the antenna formed by the ground-reflected wave. But away from broadside, the dipole, its image, or both, have a nonvanishing projection in the vertical plane. This enables the dipole to respond to the vertical component of a ground wave. In addition, lossy earth causes the surface wave to develop a radial component in the direction of propagation. Away from broadside, the radial component also projects onto the wire and induces current. The vertical and radial components induce maximum current when the surface wave arrives in line with the wire.

When compared to a quarter-wave vertical with four radials elevated 10 feet above ground, the low dipole has a 12.5 dB lower ground-wave response in its most sensitive direction in line with the wire. For sky-wave signals arriving at 20° elevation from their weakest direction (also off the ends), the dipole has 10 dB lower response than the vertical. Therefore, the low dipole has a signal-to-noise ratio advantage of 2.5 dB for signals and noise arriving from their worst-possible directions. In the most favorable directions broadside to the wire, the S/N advantage peaks sharply at 26.1 dB.

Raising the dipole broadens the broadside S/N peak and improves S/N off the ends. For example, for signals arriving at 20° elevation, a dipole at 50 feet has an S/N advantage over the reference vertical of 6.3 dB in line with the wire and 25.6 dB broadside. If you have just a single noise source and you can rotate a high dipole, you should be able to come within a few decibels of the latter figure most of the time. But when multiple noise sources in different directions arise (typical for power-line noise in times of low humidity), rotating the antenna won't help much. The S/N advantage of the high dipole then is likely to be near the worst-case figure.

These numbers illustrate the advantage of horizontal receiving antennas and substantiate the notion long held by amateurs that "verticals are noisy." But you can reduce ground-wave noise much more effectively if you don't rely on a simple horizontal wire.

Canceling Ground-Wave Components

Although a horizontal wire responds only weakly to a broadside ground wave, its response is substantial off the ends. If you could somehow eliminate the end response, you'd be left with the low broadside response and whatever residual response developed at intermediate angles.

Figure 3 shows the ground-wave response of two parallel dipoles 10 feet apart and 10 feet high fed out of phase. The phasing cancels everything arriving off the ends of the wires. The end nulls combine with the low broadside response to create a residual cloverleaf ground-wave pattern. The peak of the cloverleaf is 9.3 dB down from the peak end-response of a single dipole.[4] (Wire losses are ignored here to illustrate the cancellation principle.)

This antenna is just a very-close-spaced W8JK endfire array. Although it makes a good receiving antenna for the low bands, it's pretty large. And there's still considerable ground-wave pickup in the cloverleaf peaks. If the wires somehow could remain parallel for all directions, it might be possible to achieve complete cancellation of the vertical and radial components of the ground wave.

In some sense, the sides of a circular loop are parallel everywhere. Current amplitude and phase vary little in small loops of regular shape. Therefore, the currents in opposite sides of such loops are nearly equal and out of phase. Unlike a W8JK array, a small horizontal loop does not have a null anywhere along the ground. But its ground-wave response is uniformly low in all directions because the antenna responds only to the weak horizontal component. Everything else cancels out (or nearly so).

A small loop usually is defined as one with a total conductor length of less than 0.1 λ. But unless you use large-diameter conductors to minimize RF resistance, loops this small are inefficient. A preamplifier may be needed to overcome receiver noise. You can increase the output of a loop by increasing its size, but the larger you make it, the less constant the current becomes. This reduces groundwave cancellation.

The antenna of Figure 4 overcomes this difficulty by using two feedpoints, each on opposite sides of the loop to force current balance. One of the phasing lines is twisted to maintain proper phase.[5] This loop can be made quite large and still exhibit very low response to ground-wave noise.

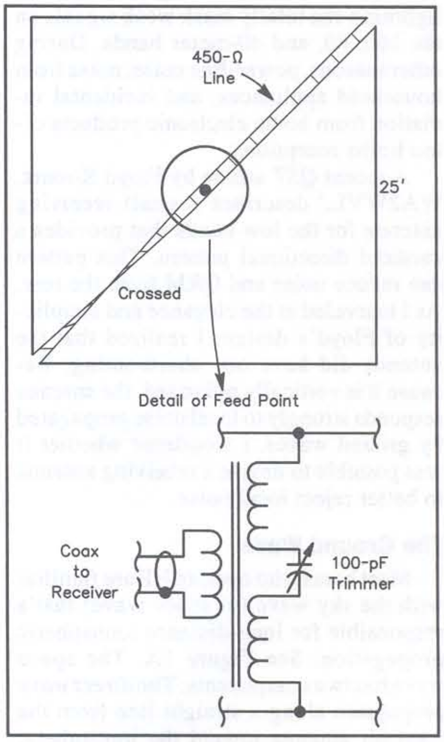

Figure 4—Basic diagram of the 80-meter low-noise loop antenna showing detail of the feedpoint arrangement. The antenna measures 25 feet on a side, is 10 feet high, and made of #14 wire. It's fed at opposite corners with phasing lines made of #14 wire spaced 1.5 inches. A small ferrite transformer at the junction of the phasing lines matches the antenna to 50 Ω coaxial feed line and also functions as a balun. The trimmer capacitor (a capacitance of about 40 pF is required) in series with the antenna-side winding resonates the loop at 3.5 MHz.

A Practical Design

The 80-meter loop of Figure 4 has a perimeter of 0.36 λ. It's 25 feet on a side, 10 feet high, and made of #14 wire. It's fed at opposite corners with phasing lines made of #14 wire spaced 1.5 inches. A small ferrite transformer at the junction of the phasing lines matches the antenna to 50 Ω and also functions as a balun. A trimmer capacitor (about 40 pF is needed) in series with the antenna-side winding resonates the loop at 3.5 MHz.

The ground-wave response of this particular loop is shown in Figure 5 and its sky-wave response in Figure 6. For signals arriving at 20°elevation, the worst-case skywave response is 20.5 dB below that of the reference vertical. This signal level is quite usable on 80 meters without a preamp. Because the loop reduces ground wave noise at least 45.1dB, S/N improvement is 24.6 dB for the worst-case combination of signal and noise directions. if you use a preamp and adjust for equal signal levels, ground-wave noise will be four S units lower on the loop *no matter what direction it comes from!* For signals arriving at higher angles, S/N enhancement approaches 30 dB.

Both the ground-wave noise pattern and sky-wave signal pattern are very uniform in azimuth. The antenna is essentially omnidirectional with an overhead null, just like a vertical. An overhead null is useful for reducing near-vertical-incidence skywave signals from nearby stations. (The loop rejects their ground-wave signals along with local noise.)

Although the S/N performance of this loop is not particularly sensitive to height, you can increase output level substantially by raising the antenna well above ground. For example, if you raise the loop to 50 feet, the output increases 15.2 dB for signals arriving at 20° elevation. At this height, the output level is only 5.3 dB below that of the reference vertical. You'll never need a preamp with a loop this high. The S/N advantage drops 0.1 dB. At a height of 20 feet, output increases 7.2 dB and S/N drops 0.6 dB.

You can shrink the loop to 10 feet on a side. The S/N advantage increases 0.5 dB, but the signal level drops 6.5 dB. You can raise the signal level 3.4 dB by doubling the loop side lengths to 50 feet, but the S/N advantage then drops 7.9 dB. You can thus trade-off loop size, S/N enhancement, and signal level. If you use smaller wire, output drops. For example, it's 1.7 dB lower for #22 wire.

Although this loop is a narrowband device and must be carefully resonated, the resonance is much broader than that of a typical small loop. You can get away with a single capacitor setting for both 3.5 and 3.8 MHz if you're willing to accept somewhat lower signal levels. The capacitor setting does not affect the patterns or skywave-to-ground-wave ratio—it simply alters output level.

The input resistance at the junction of the phasing lines varies over a wide range with loop size, height, and phasing-line characteristic impedance. The input resistance is about 40 Ω for the loop of Figure 4 at a height of 10 feet. A transformer using a type-77 ferrite core (such as an FT-82-77, FT-114-77 or FB-77-1024) with 9 turns of any size enameled wire on the coax side and 8 wire turns on the antenna side provides a good match to 50-Ω coax.[6] The input resistance drops to about 20 Ω when the loop is raised to 20 feet and to about 15 Ω at 50 feet. Use 13 turns of wire on the coax side for 20 feet and 15 for 50 feet.

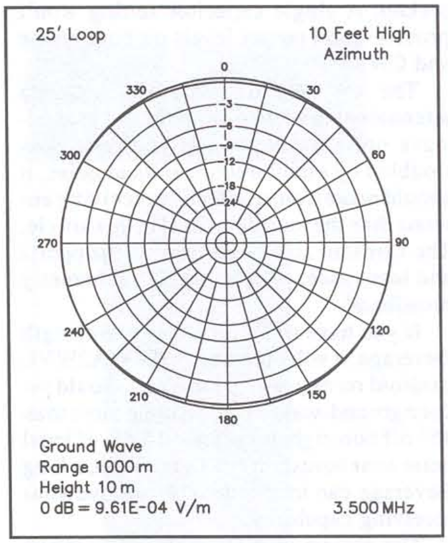

Figure 5—Azimuthal plot of the ground-wave component of a 25-foot-square, 80-meter loop at a height of 10 feet. The input power to the antenna is 1 kW. The peak electric field is shown.

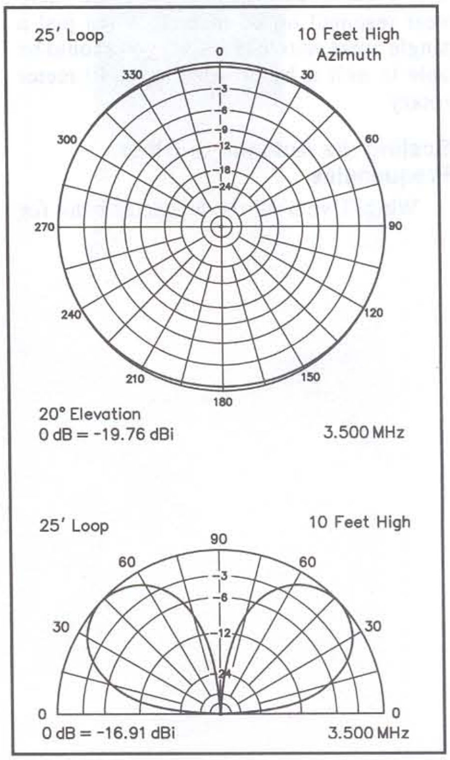

Figure 6—The sky-wave responses of the 10-foot-high, 25-foot-square, 80-meter loop. At A, the azimuthal plot; at B, the elevation plot.

Although the optimum transformer turns ratio varies with antenna height, little output is sacrificed if you use a fixed ratio.

For example, output is only 1 dB less than optimal when a loop designed for 10 feet is used at 50 feet. If you're determined to obtain the best possible match, use 16 turns of wire on the coax side and tap the winding for lowest SWR (alternatively, a switch can be used to select any of two or more taps). Use the lowest possible power when measuring SWR. It's easy to puncture the dielectric of a small trimmer capacitor with a momentary blast.

The antenna's resonant frequency shifts when the phasing lines get wet. If you use true open-wire line with plastic spacers, the frequency shift will be less than 100 kHz. But if you use 450-Ω line with segmented polyethylene dielectric, the resonant frequency decreases more than 200 kHz when the line becomes thoroughly damp. Although this won't affect signal-to-noise ratio, output drops in the desired frequency range. You may be tempted to try phasing lines of 300-Ω twinlead routed inside PVC tubing to avoid moisture effects, but line impedances this low work only for smaller loops.

This antenna should be constructed as symmetrically as possible to maximize cancellation of the vertical and radial components of the ground wave. Make the loop perfectly square and accurately align it in the horizontal plane. Cut the phasing lines to the same length. Although these loops perform well near houses, fences, and towers, try to install the antenna as far from other conductors as possible to maximize current balance. Use the shortest possible leads to interconnect the matching components. Although it's probably unnecessary, I like to split the antenna-side transformer winding and put the tuning capacitor in the center to promote equal currents in the phasing-line conductors.

To minimize the number of supports, a loop about 17 feet on a side can be constructed using a 20-meter quad spreader. You can mount the spreader well up on a tower to increase output. Alternatively, you may be able to eliminate supports altogether by stringing a loop in your attic or garage. However, the current balance of indoor loops may be degraded by electrical wiring, plumbing, heating ducts, or other nearby conductors.

On-the-Air Performance

Ed Andress, W6KUT, located in the San Diego suburb of Poway, constructed a loop 21 feet on a side and 10 feet high. Ed's location is subjected to strong, chronic power-line noise. On 80 meters, Ed uses a pair of phased quarter-wave verticals for transmitting. We used the verticals as a reference when evaluating the loop. A 20-dB preamp was available. With the preamp, signals near loop resonance were about equal to those from the verticals.

The loop performed as expected. It

dramatically enhanced the signal-to-noise ratio of most sky-wave signals. Sometimes the verticals did better during a momentary change in propagation; occasionally a particular noise arose that the loop didn't attenuate much. But overall, the loop was far superior. It made little difference on strong signals. It made listening to moderately strong signals much more pleasant. It let us copy weak signals that were buried in the noise and unreadable on the verticals.

During times of no detectable powerline noise, we often noticed a curious effect: The loop still enhanced signal-to-noise ratio by one or two S units, making copy of moderately weak signals more pleasant. On these occasions, we were unable to hear the telltale, raspy buzz of power-line noise (or any other noise signature) when we listened with the transceiver's AM detector. Unless noise happens to arrive at low angles and signals at high, there's no reason for the loop to enhance skywave S/N. We believe that the unidentified noise is local and propagates by the ground wave. We speculate that it may be the sum of hundreds of weak man-made noise sources in the densely populated suburb. (The superposition of a large number of noise sources tends to be characterless even when the individual sources aren't.)

Total Ground-Wave Cancellation

If you stack two of these loops vertically and bring both feed lines into the shack, you can form a deep null on the horizon for all azimuth angles by combining the signals with a fixed amplitude and phase offset. Except for azimuth-response irregularities caused by nearby conductors and small elevation-angle differences due to range, this system will cancel all ground-wave components. This noise canceler was to have been the original subject of this article. However, a single component loop worked so well in practice that I decided total ground-wave cancellation was overkill.

Comparison with Other Antennas

A conventional, single-feedpoint, small loop about two feet on a side—oriented horizontally—yields roughly the same S/N enhancement as the loop of Figure 4. However, output will be down about 46 dB from the reference vertical. You'll need a low-noise preamp with a loop this small unless your receiver has a very low noise figure. Still, even with a preamp, a conventional small loop makes an attractive, low-profile alternative. You must construct a low-output antenna like this carefully to avoid stray pickup. A single capacitor setting won't provide good output levels on both phone and CW.

The WA2WVL cardioid antenna attenuates thunderstorm static and ground-wave noise to the rear. If you're seldom troubled by omnidirectional local noise, it should make a more effective receiving antenna than the loop described in this article. The cardioid also requires fewer supports and less space. As a bonus, it's inherently broadband.

If you have room for a two-wavelength Beverage, it will outperform the WA2WVL cardioid on sky-wave noise and should reduce ground-wave noise arriving more than 45° off boresight by at least 15 dB. If local noise near boresight isn't a problem, a long Beverage can tremendously improve your receiving capability.

The easiest way to improve reception on 80 or 160 meters is to use the most sensitive horizontal antenna available at your antenna switch. Many hams with 80-meter verticals find that switching to a 40-meter dipole or beam improves copy of weak signals even though the antenna is nowhere near resonant on 80 meters. When just a single noise source is active, you should be able to null it by broadsiding a 40-meter rotary.

Scaling the Antenna to Other Frequencies

While I've used the 80-meter band for illustration in this article, it's easy to scale the design to other frequencies. Simply multiply lengths, heights, transformer turns and capacitor values by the number you get when you divide 3.5 by the target frequency in MHz.

That said, I don't recommend this antenna for use above 40 meters. If you're using a vertical antenna on the upper HF bands, do yourself a favor and replace it with the highest horizontally polarized antenna you can manage. Not only will your receive noise decrease, your transmit signal almost certainly will improve due to higher ground-reflection gain.[7] (Exception: If your vertical radiates over saltwater, keep it!) If you're already using a horizontal wire on upper HF, I think your next antenna project should be a rotary beam rather than a receiving loop.

Notes

[1] Floyd Koontz, WA2WVL, "Is this Ewe for You?" *QST*, Feb 1995, pp 31-33.

[2] Frederick Terman, *Radio Engineers' Handbook*, 1st ed (New York: McGraw-Hill, 1943), pp 674-709.

[3] *MININEC*-based antenna-analysis programs do not compute the surface wave, nor can they accurately model antennas close to ground. I used *NEC/Wires 2.0* with its surface-wave and Sommerfeld-Norton ground options for all antenna models in this article. All models assumed average ground characteristics (dielectric constant 13, conductivity 5 mS/m).

[4] A small cloverleaf peak requires very close wire spacing. For example, the peak is down only 5.4 dB for a spacing of 50 feet.

[5] The feed arrangement is like that of an Alford Loop. In fact, the low-noise receiving antenna really is just a variant of the VHF/UHF loop devised by Andrew Alford for a different purpose in 1940. See Terman, pp 814-815.

[6] Don't be tempted to eliminate the transformer and connect 50-Ω coax directly to a loop. The transformer functions as a current balun. It keeps noise current induced on the coax shield from entering the receiver. Just a little pickup by the shield can pollute the low-noise output of a loop.

[7] For example, on 20 meters, a horizontal dipole only 35 feet high has 3 to 6 dB broadside gain over a vertical dipole at low angles over average ground.

By Gary Breed, K9AY From *QST*, September 1997

The K9AY Terminated Loop—A Compact, Directional Receiving Antenna

Wish you had enough room for an effective low-band receiving antenna? You do! This four-direction system fits in a 30-foot circle!

Low-band operators are always looking for ways to improve their hearing. As a low-band fan, I was impressed with the EWE antenna developed by Floyd Koontz, WA2WVL.[1,2] Koontz shows us how to build a compact, directional antenna—a design that quickly became very popular. But when I sat down at the computer to figure out the best way to install my own EWEs, a surprising new design emerged from my modeling experiments.

Ladies and Gentlemen...

Allow me to introduce you to the *terminated loop*, a concept that further shrinks the space required for a good receiving antenna *without sacrificing performance*! Figure 1 compares the real estate requirements of my four-direction loop system to that of an equivalent EWE array. The new system is not only smaller, it's easier to install, needing only one support instead of five.

The terminated loop (see Figure 2) is physically and electrically quite simple. It consists of a wire loop of any convenient shape (diamond, delta, etc), hung from a single support and with a ground rod at the bottom. A 9:1 impedance-matching transformer connects from one end of the loop to ground; a terminating resistor connects the other end of the loop to ground. This antenna is directional, favoring signals arriving from the feed point end, rejecting by several S units any signals arriving from the end connected to the terminating resistor, R_{TERM}.

Here's the real news: A very small terminated loop maintains its directional pattern. As the loop's size gets smaller, however, the desired signal strength is also reduced. The antenna described here is small enough to fit into a corner of almost *any* backyard. The loop could be made smaller, but I wanted to collect enough signal energy so that

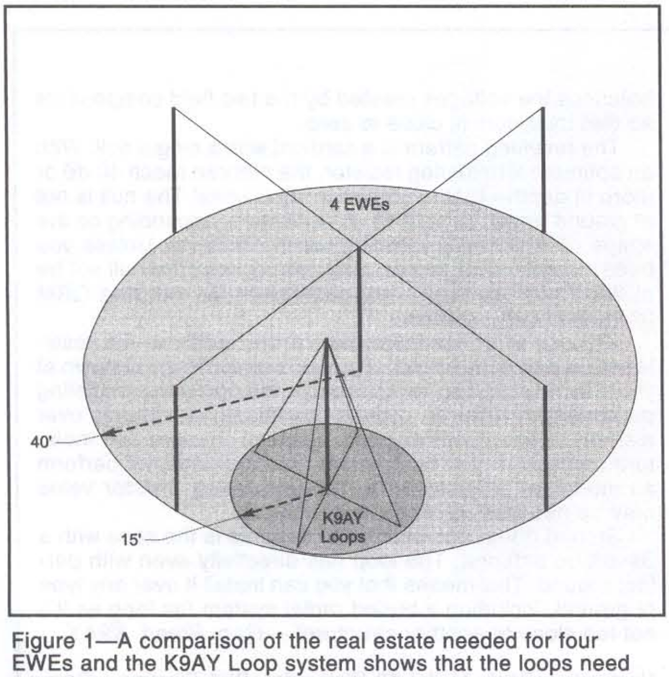

Figure 1—A comparison of the real estate needed for four EWEs and the K9AY Loop system shows that the loops need only 1/7 the area of the EWEs, yet they provide the same directional patterns.

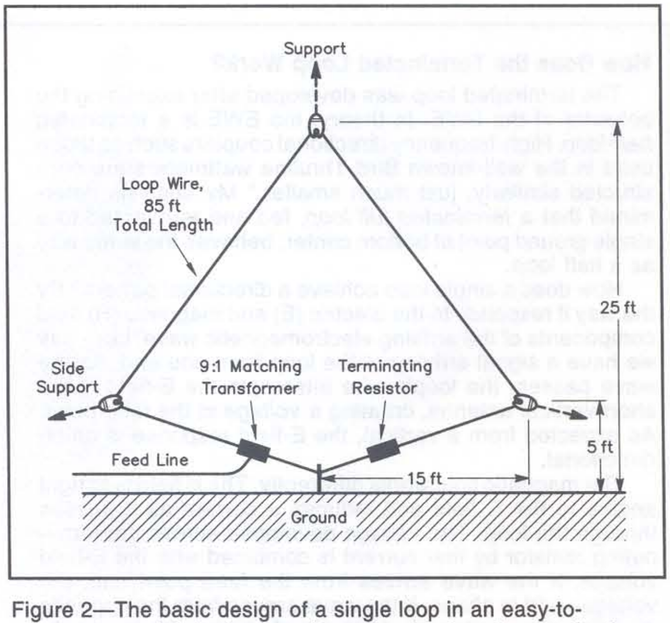

Figure 2—The basic design of a single loop in an easy-to-construct quasi-delta loop configuration. Exchanging the feed point and termination reverses the pattern. A four-direction system uses two of these loops installed at right angles to one another (as shown in Figure 1) and a relay controlled switching system.

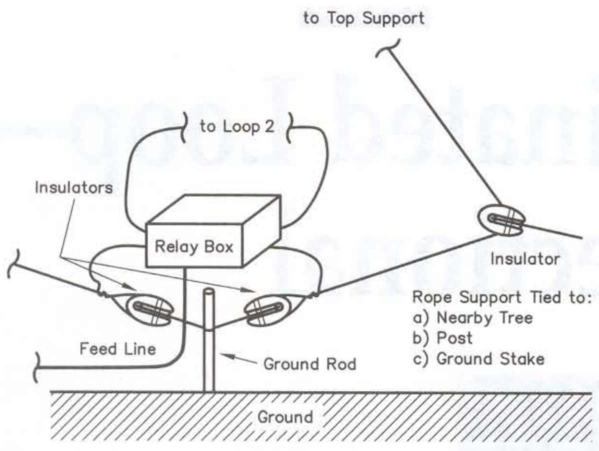

Figure 3—Construction details of a terminated loop, showing the base connections and corner supports of one of the loops. The relay box also houses the 9:1 matching transformer.

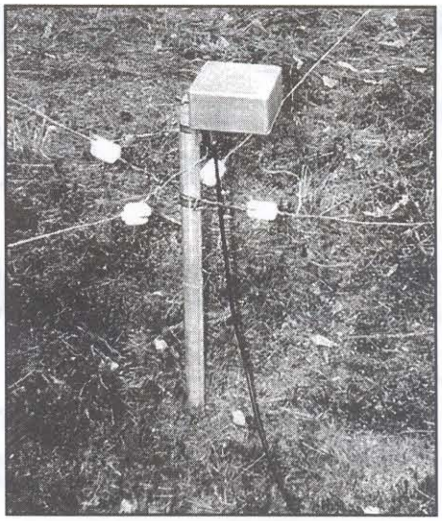

Figure 4—A photo of the central connection point at the base of a two-loop system, as installed at K9AY.

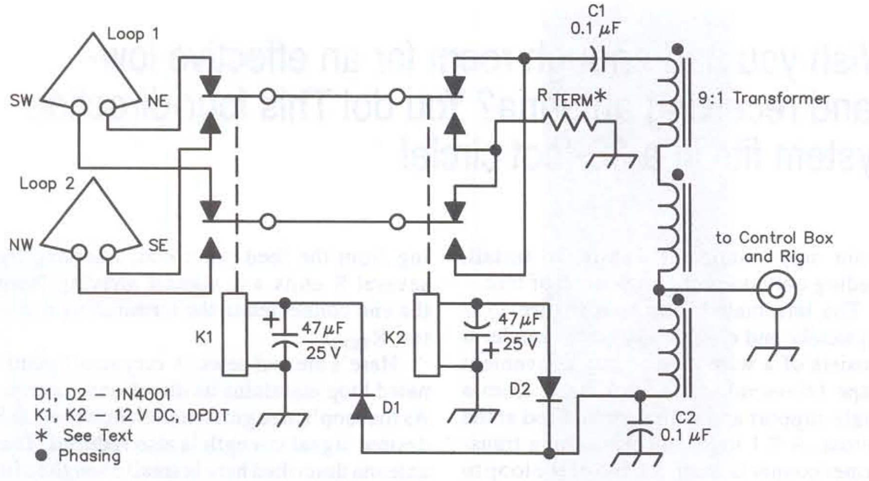

Figure 5—Schematic of the relay box located at the base of the system. The relays switch between the two loops, reversing the feed point and termination connections.

even a modest preamp (such as those included in most HF rigs) can be used. More information on the antenna's operation can be found in the sidebars "How Does the Terminated Loop Work?" and "Summary of Characteristics."

To cover all directions, two loops using the same support are oriented at right angles to each other. Each loop provides reception from two directions when the feed point and termination are reversed, for a total of four separate patterns. Here are all the details you need to build the K9AY Loop antenna system, including a relay-controlled pattern-switching system that uses the coaxial feed line to carry the switching system control voltage.

How Does the Terminated Loop Work?

The terminated loop was developed after examining the behavior of the EWE. In theory, the EWE is a *terminated half-loop*. High-frequency directional couplers such as those used in the well-known Bird Thruline wattmeters are constructed similarly, just much smaller.* My analysis determined that a *terminated full loop,* fed and terminated to a single ground point at bottom center, behaves the same way as a half loop.

How does a single loop achieve a directional pattern? By the way it responds to the electric (E) and magnetic (H) field components of the arriving electromagnetic wave. Let's say we have a signal arriving at the loop from one end. As the wave passes, the loop's wire intercepts the E-field like a short vertical antenna, creating a *voltage* at the feed point. As expected from a vertical, the E-field response is omnidirectional.

The magnetic field works differently. The H field is at right angles to the E field and induces a *current* as it passes through the loop. The voltage developed across the terminating resistor by that current is combined with the E-field voltage. If the wave arrives from the feed point end, the voltages add in phase. If the wave arrives from the opposite direction, the voltage is 180° out of phase with (and subtracted from) the E-field voltage. To maximize the front-to-back ratio, the terminating resistor must have a value that balances the voltages created by the two field components so that their sum is close to zero.

The resulting pattern is a cardioid with a single null. With an optimum terminating resistor, the null can reach 40 dB or more in depth—that's more than six S units! The null is not at ground level, but at 20 to 55° elevation, depending on the shape of the loop and local ground conditions. Unless you build a short, wide loop or a tall, skinny loop, the null will be at 30 to 40° elevation, very convenient for reducing QRM from in-country stations.

Ground is an essential part of the antenna—its resistance is part of the circuit. If ground conductivity is known at your location, it can be included in the computer modeling parameters. However, ground conditions can change over a short distance and vary with seasonal changes and moisture content. If you find that the antenna does not perform as modeled, adjustment of the terminating resistor value may be needed, as noted in the article.

Ground does not need to be lossy, as is the case with a Beverage antenna. The loop has directivity even with perfect ground. This means that you can install it over *any* type of ground, including a buried radial system (as long as it's not too close to another structure).—*Gary Breed, K9AY*

**Instruction Book*, Model 43 Wattmeter, Bird Electronic Corp, Solon, OH.

Let's Build It!

First, gather the following materials to construct the loops:

- Two lengths of **wire** about 85 feet each. Although almost any wire size will do, #14 copper is probably the best for a long-lasting installation.
- Ten simple **insulators**—anything from old toothbrush handles to fancy porcelain insulators will do.
- A three or four-foot length of **copper pipe**—the ground rod doesn't need to be driven deeply—a depth of three to four feet is sufficient. Because copper-plated steel rods eventually rust, I prefer to use ½ or ¾-inch-diameter copper water pipe.
- One **support** 25 feet above ground—my antenna hangs from a limb of a Georgia Pine, but anything that gets the top of the loops up about 25 feet will work. The wooden **A** frame support described in *The ARRL Handbook* and *Antenna Book* is an excellent choice. A metal mast can be used, but make it only as long as necessary and insulate it from ground. Large metal objects affect the antenna's performance, so install it in the clear.
- You'll also need some **rope**—nothing fancy, just something strong enough to keep things in place.

Now, grab your tools and head for the backyard. Attach the midpoint of each wire length to an insulator that will be positioned at the top of the loop. The two loops must not touch each other, so leave some room between their insulators. I separated my insulators with about a foot of rope.

Aligned directly under the tops of the loops, drive a ground rod, leaving a foot or so above ground as an attachment point. Using short ropes or wires, connect four insulators a few inches away from this post to receive the lower ends of the loops. Leave a foot of wire for a pigtail after twisting the ends of the loop wires around the insulators. Figure 3 sketches the important installation details, and the photo of Figure 4 shows how my system is installed.

The lower corners of the loops are supported by insulators and ropes tied off to nearby trees, fence posts, or stakes driven into the ground. Pull out the corners with enough tension to maintain the loop shapes. At this point, the major mechanical work is finished and you should have something resembling an eggbeater, as shown in Figure 1.

The Relay Box and Controller

Figure 5 shows the relay circuit used to switch the feed and termination points of the loop wires. House the components in a weatherproof box having external connection points for the four ends of the loops, a coax connector for the feed line and a ground-wire attachment.

K1 is a DPDT relay that switches between the two loops. K2, another DPDT relay, swaps the connections of the terminating resistor and matching transformer, thereby reversing the pattern. Relay power is supplied via the coaxial feed line. C1 keeps the control voltage from reaching the antenna, and C2 provides an RF ground for the transformer.

Most builders will likely choose to orient the loops for northeast/southwest and northwest/southeast directions. The relay box has four switching modes, one for each direction: 1) northeast, when *neither* relay is energized; 2) southeast, when K1 *only* is energized; 3) southwest, when K2 *only* is energized and 4) northwest, when *both* relays are energized. Switching is accomplished using a single power connection through steering diodes D1 and D2. D1 allows K1 to operate when +12 V is applied, while D2 blocks the voltage from reaching K2. When −12 V is applied, K2 operates, but not K1. When 12 V ac is applied, the diodes rectify it to operate *both* relays.

The matching transformer is a 9:1 impedance, 3:1 turns ratio type that should be familiar to many readers. Five trifilar turns of ordinary hookup wire are wound on a ¾-inch-diameter (0.825 inch) 43 material toroid. The terminating resistor (R_{TERM}) value will be between 390 to 560 Ω, depending on your band preference. With average ground conductivity, a value of 390 Ω provides the optimum F/B at 160 meters, while 560 Ω optimizes the loops for 80 meters. A value of 470 Ω splits the difference for "pretty good" performance on both bands. I chose to optimize the antenna for 160 meter operation, so used a 390 Ω resistor. Use an R_{TERM} power rating of at least 1 W in case some transmitter power ends up being coupled to the loops. I use a 2-W carbon resistor, but two parallel ½-W or four ¼-W units of appropriate ohmic value can also be substituted.

Figure 6 is a diagram of the control box that is located in the shack. One 12 V ac transformer provides the ac relay power and feeds two half-wave rectifiers and filter capacitors to generate ±12 V. One SP4T switch selects the proper voltage to apply to the coax. The RF choke and 0.1 µF capacitor keep the RF and control voltage separated at the shack end.

Once the control unit and relay box are built and operating properly, mount the relay box at the ground rod and connect the four ends of the loops to their proper terminals. A length of 50 Ω coax carries the signal and power between the antenna and the switch box in the shack. I highly recommend keep-

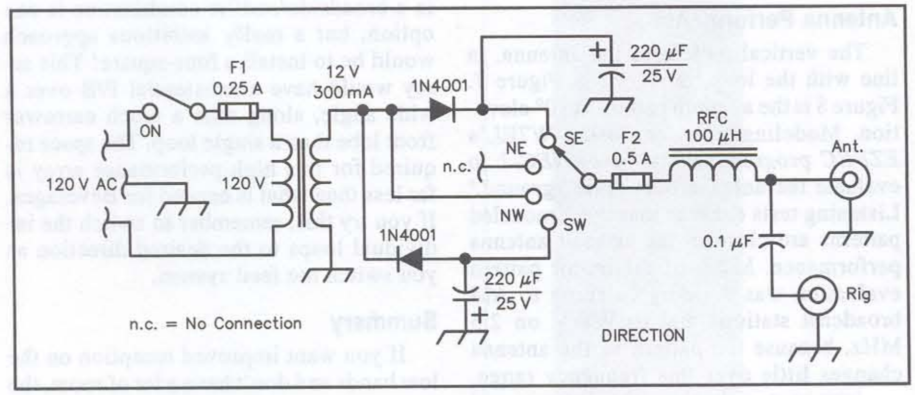

Figure 6—Schematic of the control box located in the shack. The control voltage to operate the relays is delivered via the coaxial line.

Summary of Characteristics

When choosing a specific configuration or location for your K9AY Loop system, consider this list of characteristics:

- Like all antennas, this one can be affected by nearby metallic objects or structures. Do your best to keep the antenna in the clear, away from your tower, house and power lines.
- Noise reduction in this antenna is achieved mainly by the directional pattern, although being grounded offers some reduction of wind, rain and snow static. Compared to the typical omnidirectional vertical or low inverted **V**, the reduction in overall noise and interference can be dramatic.
- The maximum circumference of the loop is a little over ¼ λ at the highest frequency of operation. If the loop is larger, the E and H-field responses of the antenna can no longer be balanced. Smaller loops (or same-size loops at lower frequencies) retain the directional pattern, which makes this an excellent antenna for AM broadcast reception. Unfortunately, the received signal voltage is proportional to the area enclosed by the loop, so sensitivity decreases rapidly as the antenna becomes smaller. Unless you have a very good preamp, keep the loop sizes near the maximum.
- The T-Loop is not just a low-frequency antenna! It can be used at high frequencies for shortwave listening, or to provide improved reception over an omnidirectional antenna such as a vertical. Just scale the size to the desired frequency using the guidelines presented in the accompanying article.—*Gary Breed, K9AY*

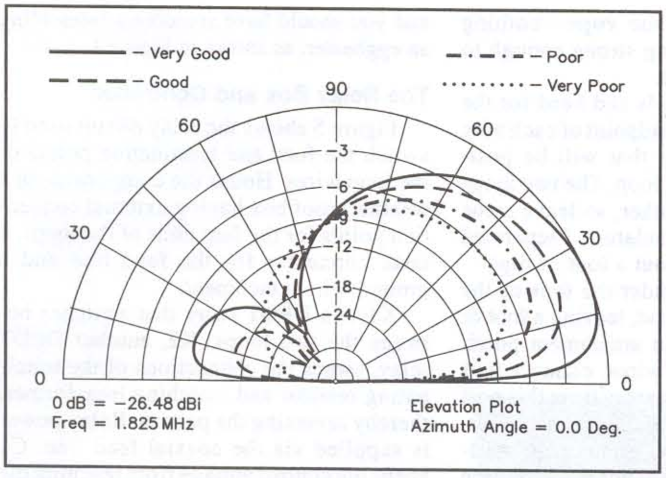

Figure 7—Vertical radiation pattern of the loops along the plane of the loop.

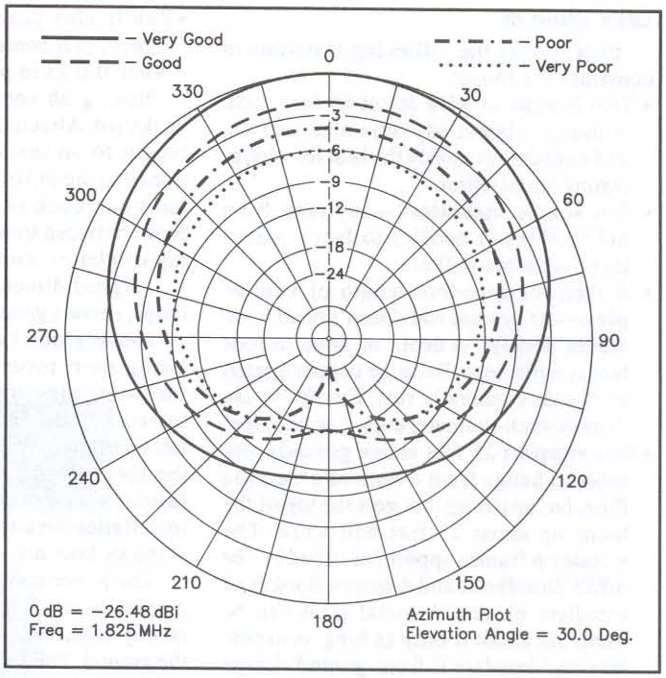

Figure 8—Horizontal radiation pattern of the loops at 30° elevation.

ing the coax on the ground (buried is better) to minimize pickup of noise or energy coupled from your transmit antenna. A receiving preamplifier is almost certainly needed, either your rig's internal preamp or an external preamp.

Evaluation and Adjustment

Next, verify that the antenna is working correctly. Some listening is probably what you'll do first, but fading makes it almost impossible to determine the antenna's actual performance. At best, you will be able to confirm that the antenna has reasonable directivity.

More accurate listening tests can be performed several ways. The best way is to enlist the aid of a nearby ham whose station is very close to being in line with one of the two loops. If such help is not available, the next-best option is to identify a local AM radio station high in the band (1400-1600 kHz) and use that as your test signal. Switch the loops to their various directions. If the test station is directly in line with one loop, you should see a front-to-back (F/B) ratio of about 2 to 3 S units as the antenna is switched toward and away from the station. You won't see a *huge* F/B because the deepest null is up at 30 to 40°.

Local ground conductivity can affect performance. You might not get an optimum pattern at your particular location with the "normal-value" terminating resistor. If you aren't getting the expected performance, substitute a 1 kΩ potentiometer for the terminating resistor and adjust it for best F/B while listening to your test station. Then, measure the pot's resistance and install a fixed-value resistor of the same value.

Antenna Performance

The vertical pattern of the antenna, in line with the loop, is shown in Figure 7. Figure 8 is the azimuth pattern at 30° elevation. Modeling was done using W7EL's *EZNEC* program,[3] which uses *NEC-2* to evaluate the antenna over "real" ground.[4] Listening tests confirm that these modeled patterns are close to the as-built antenna performance. Much of the on-air pattern evaluation was done by listening to AM broadcast stations and to WWV on 2.5 MHz, because the pattern of the antenna changes little over this frequency range. Stations at exactly the right distance and directions have verified the deep null off the back. For example, New York City area radio stations are exactly in line with my northeast/southwest loop, and show 40 dB F/B ratio when the arrival angle is just right!

The front of the pattern is quite broad, with maybe one S unit F/S. The advantage of this antenna is its rearward null, which reduces local noise and distant QRM. I installed my system just 2½ weeks before the 1996 ARRL 160 Meter Contest, so it got a thorough evaluation in a short time. During the contest, changing the pattern direction often made the difference between Q5 copy and a busted contact. For example, pointing southeast to hear Caribbean stations reduced stateside QRM by 2 or 3 S units, enough to easily hear the DX through unruly pileups.

Higher-Performance Ideas

The compact size of the K9AY Loop makes it easy to build more than one for a multielement array. Installing two of them in a broadside/endfire combination is one option, but a really ambitious approach would be to install a four-square! This array would have a substantial F/B over a wide angle, along with a much narrower front lobe than a single loop. The space required for this high-performance array is far less than what is needed for Beverages. If you try this, remember to switch the individual loops to the desired direction as you switch the feed system.

Summary

If you want improved reception on the low bands and don't have a lot of room, the terminated loop is an excellent choice. It is small, easy to build and its directional pattern makes DX much easier to hear.

Notes

[1] Floyd Koontz, WA2WVL, "Is This Ewe for You?" *QST*, Feb 1995, pp 31-33.

[2] Floyd Koontz, WA2WVL, "More EWEs for You," *QST*, Jan 1996, pp 32-34

[3] *EZNEC* 1.0 by Roy Lewallen, W7EL, PO Box 6658, Beaverton, OR 97007; tel: 503-646-2885, fax: 503-671-9046.

[4] The modeling accuracy of antennas with a direct connection to ground is somewhat uncertain, even with the power of *NEC-2*. The high accuracy ground characterization of the Sommerfield method gives answers that vary with the number of segments chosen for each wire. The *MININEC*-type ground model gives more consistent results, but on-air tests show that the deepest null is obtained with a lower value terminating resistor than this model indicates. The plots presented here are the "best fit" between modeled and observed performance.

Photos by the author.

Hum Problems When Switching the K9AY Loops

By Gary Breed, K9AY

I am pleased to report that many hams have successfully built the receiving antenna described in the September 1997 issue of *QST*.[1] I had some concern that variations in local ground conditions and nearby structures could reduce the antenna's performance, but the performance I obtained seems to be readily duplicated.

There is one problem that needs to be addressed. When the loops are switched to the northwest direction, an ac voltage is sent down the coax to the relay box. A few hams have reported hum or distortion when the antenna is switched in this direction. Two explanations are possible: a ground loop due to widely separated antenna and station grounds, or modulation of the core of the matching transformer. The presence of hum was not evident in my prototype, but to avoid either cause of the problem, I recommend using a *separate* three-conductor control wire to operate the relays. The power is solely 12 V dc, and no current flows through the transformer. The modifications are shown in the accompanying figures, which can replace Figures 5 and 6 in the original article.

Secondary benefits of this arrangement include the option of using an existing 12 V power supply to provide the operating voltage, and the ability to use an additional conductor to carry power to an antenna-mounted preamplifier.

---

[1] Gary Breed, K9AY, "The K9AY Terminated Loop—A Compact, Directional Receiving Antenna," *QST*, Sep 1997, pp 43-46.

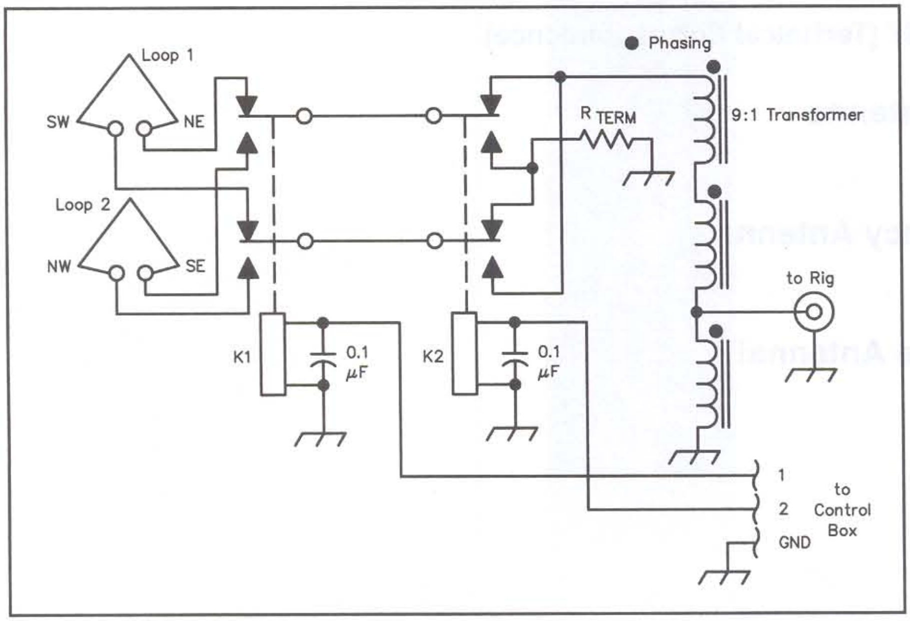
Loop 1
SW
NE
Loop 2
NW
SE
● Phasing
R TERM
9:1 Transformer
to Rig
K1
0.1 µF
K2
0.1 µF
1
2
GND
to Control Box

Figure 1—A revision of the schematic shown in Figure 5 of the September 1997 *QST* article.

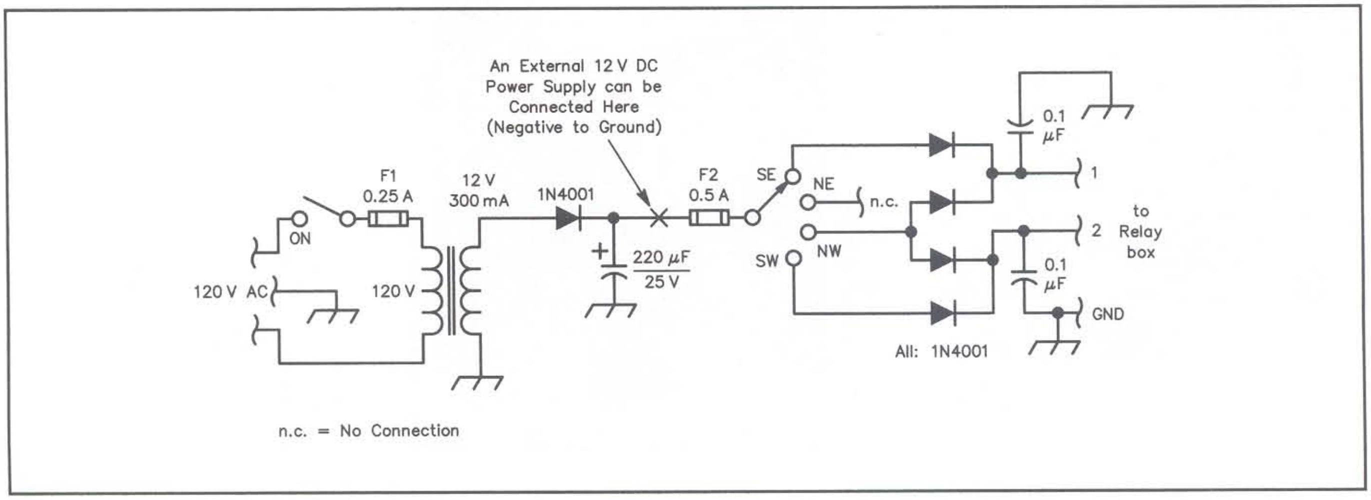
An External 12 V DC Power Supply can be Connected Here (Negative to Ground)
F1 0.25 A
ON
120 V AC
120 V
12 V 300 mA
1N4001
220 µF 25 V
F2 0.5 A
SE
NE
n.c.
NW
SW
0.1 µF
0.1 µF
1
2
GND
to Relay box
All: 1N4001
n.c. = No Connection

Figure 2—Changes to Figure 6 of the original article.

CHAPTER NINE

MISCELLANEA

9-1 **A Skyhook for the '90s**
From May 1997 *QST*

9-4 **Balloon Skyhooks**
From August 1997 *QST* (Technical Correspondence)

9-5 **The Clothesline Antenna**
From July 1998 *QST*

9-8 **A Variable-Frequency Antenna**
From July 1949 *QST*

9-12 **Honey, I Shrunk the Antenna!**
From July 1993 *QST*

By Don Daso, K4ZA **From *QST*, May 1997**

A Skyhook for the '90s

Is your group looking to put up a BIG vertical antenna for Field Day? The crew at N4ZC started their own air corps. Their full-size 160-meter quarter-wavelength vertical is *not* supported by hot air!

PHOTO BY DON DASO, K4ZA

Inspiration, on a grand scale.
What if we had one of *these* to play with?

Some History, as Introduction

By 1936, the Zeppelin company was without rival in the design, construction and operation of rigid airships. Flying successfully around the world, providing regular passenger service between Germany and South America, the *Graf Zeppelin* airship was nonetheless unsuited for the hardship of crossing the North Atlantic. The *Hindenburg* was designed and built for this route. This huge airship was one of the largest flying machines the world has ever known. Her crash and destruction at Lakehurst on May 6, 1937, ended the era of passenger transport by airship.[1]

Later, growing up on a farm near Akron, I sometimes saw the Goodyear airships passing silently overhead. Twenty-odd

PHOTO BY SCOTT DOUGLASS, K2SD

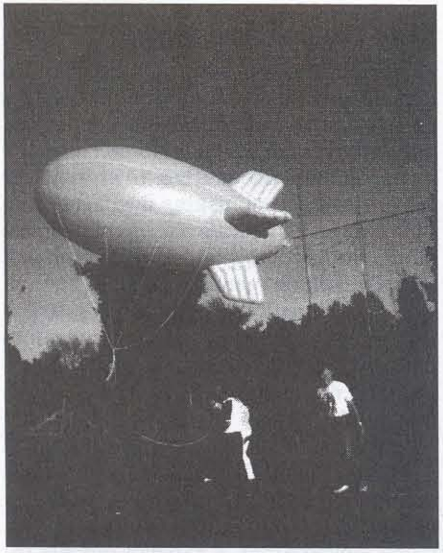

Figure 1—K4ZA fastens the antenna and tether to the miniature blimp prior to launch.

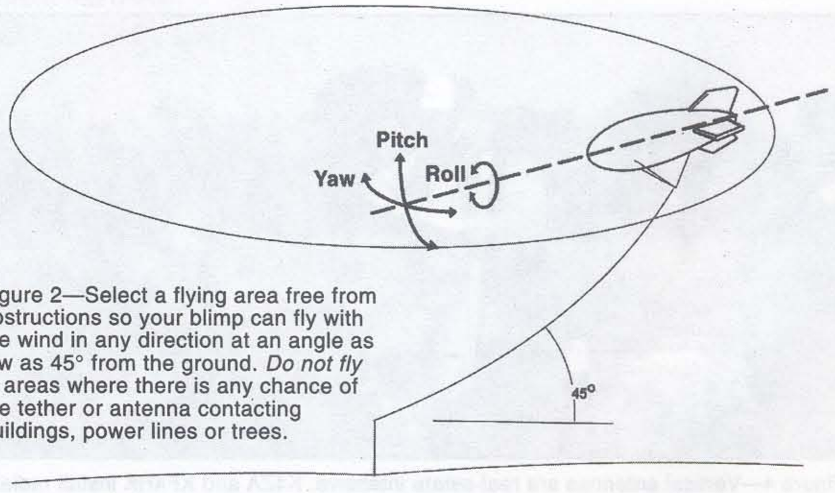

Figure 2—Select a flying area free from obstructions so your blimp can fly with the wind in any direction at an angle as low as 45° from the ground. *Do not fly* in areas where there is any chance of the tether or antenna contacting buildings, power lines or trees.

---

[1] The terms concerning lighter-than-air (LTA) craft are often misapplied. *Webster's Ninth Collegiate Dictionary* indicates that *dirigible*, *airship* and *blimp* all denote an LTA with *steering and propulsion* capabilities. Hence, the terms do not strictly apply to the kytoon, which has neither steering nor propulsion systems. Nonetheless, *blimp* is so commonly used by manufacturers, sellers and users to describe the kytoon that we continue that usage here, as well. —*Ed.*

PHOTO BY DON DASO, K4ZA

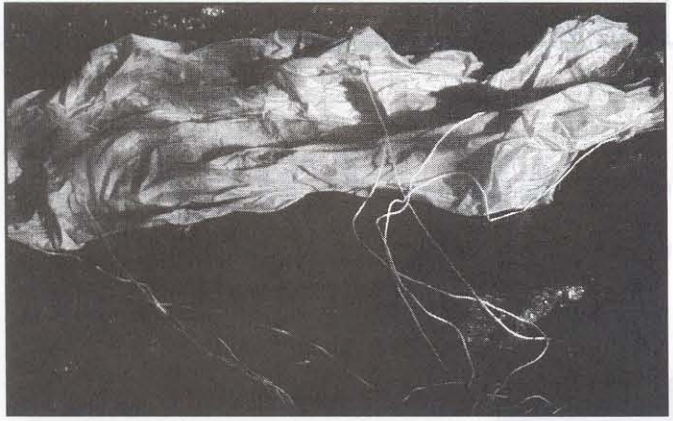

Figure 3—Here's our miniature version of the real thing—fresh out of the box and ready for inflation.

years after the *Hindenburg* disaster, I began to experiment with balloons myself—originally using surplus weather balloons to support wires, and more recently, with small blimps holding contest antennas. Many hams have considered using balloons or kites for antenna supports. Indeed, they come closer than anything else to that "skyhook" we all jokingly refer to on occasion. Yet, each one has limitations.

The Kytoon

Balloon-supported antennas tend to lose altitude in the wind, however slight. Kites cannot fly without wind. Satisfactory service from either support requires a limited range of wind and weather conditions—something that's not too common. Combining features from each—the balloon and kite—solves several of these limitations. Indeed, the Kytoon is a small, inflatable blimp-shaped balloon with kite-like surfaces that cause it to fly into the wind. An early Kytoon—with a rigid skeleton—is registered with the US Patent Office and enjoyed some success as a skyhook. (See Bibliography for further information.)

Today, the early Kytoon's bladder-and-frame construction has been replaced by completely inflatable balloons, made from 2 to 3-mil polyurethane film (see Figure 1). Brilliant skin colors are common because these miniature blimps are intended for advertising purposes. These small blimps are available from various sources, with prices varying according to size—typically from 10 to 30 feet long and starting at $300. [Beware: Many suppliers specialize in advertising Kytoons and sell them for twice the price of those sold for scientific uses.—*Ed.*]

The plastics used in their construction have been specially formulated for high elasticity, helium retention and paint application. Helium (a safe, inert gas) is available from party suppliers or welding supply companies. One tank will easily support an antenna for the typical contest weekend. If you have some room, but maybe not the resources to put up a full-sized radiator, consider a tethered Kytoon. As a modern antenna support, these miniature blimps are the skyhook of the '90s.

Tethered Flight

Most hams considering these small blimps as antenna supports concentrate on their antennas. This natural inclination can create problems. Most hams have little knowledge or experience with the basic principles of LTA flight, knots and rigging rules, with FAA regulations, with weather and especially with how these all relate to tethered flight. It's a complex process and serious business to lift an antenna aloft with a small balloon or blimp.

These blimps consist of a closed envelope of gas—usually helium—that is lighter than the atmosphere. Gas balloons can only descend by losing lift (that is, by venting gas) or when some force overcomes their lift (retrieval by tether or powered flight). As a gas balloon rises, it will reach an equilibrium altitude, where it will remain until the lift-to-weight ratio changes. These principles—which are mostly associated with larger, cargo-carrying blimps—are important, especially because our balloon will be tethered. As with any tower or antenna project, use common sense and make safety your highest priority!

Because space is three-dimensional, any "skyhook" can turn in the air—rotating in three axes: a longitudinal roll, a lateral yaw and a vertical pitch. The shape of these blimps allows them to "fly" in the wind, countering these motions, and the tether allows the blimp to move easily enough within those three axes. Generally, you will experience something like Figure 2.

A 10 to 12-foot blimp is sufficient for most antennas suitable for low-band use. Most blimp applications will be for 160 meters, because the sheer size of effective antennas for this band makes them difficult (and costly) to erect. A 10 to 12-foot-long blimp of 3 to 4 feet in diameter holds about 70 cubic feet of helium and is capable of lifting two pounds. This is sufficient for an effective $\lambda/4$ wire vertical for 160 meters.

Setup, Antenna and Ground

You'll need a working area on the ground to inflate your blimp. A 20×20-foot area should suffice. It's important to protect the thin skin of the balloon. I usually lay the balloon out on a tarp to prevent any stubs or twigs from poking the plastic skin (see Figure 3). While the plastic seems sturdy and elastic, I'm always careful and treat the balloon gently. I hold the helium tank upright, and I usually lash it securely to a two-wheeled dolly, so I can move it to the field easily. Follow the manufacturer's inflation directions carefully (each will be slightly different).

Inflation temperature is important: If possible, inflate the balloon at the same temperature it will experience in flight. Helium comes out of the tank cold. With rising temperatures, it expands quickly inside the balloon. Sunlight and warm air cause further expansion. (In very hot weather, a balloon loses some of its lifting power, but this is usually not a problem.) Do not overinflate the balloon. Finger pressure will barely dent the surface of a properly inflated polyurethane blimp. Pay special attention while closing the neck of the blimp. With many blimps, you can simply fold the neck back on itself several times and then fasten it in place with heavy rub-

PHOTO BY SCOTT DOUGLASS, K2SD

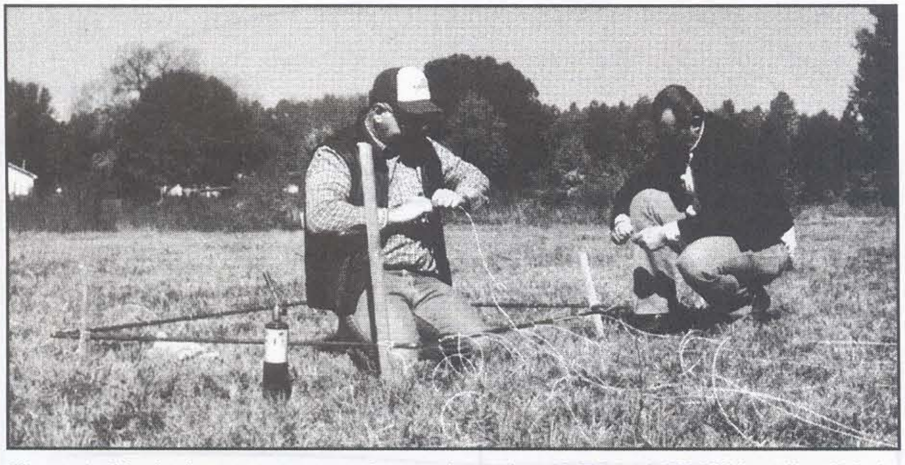

Figure 4—Vertical antennas are real-estate intensive. K4ZA and KF4HK install radials in the horse pasture at N4ZC's contest station.

PHOTO BY SCOTT DOUGLASS, K2SD

Figure 5—WA4UNZ and K4ZA work inside the large copper ground ring installing the feed system.

ber bands (folding them, if necessary) to provide a tight, leak-free fit. Refills are much easier when rubber bands, rather than knotted rope, seal the balloon.

Use a quality ball-bearing swivel between the end of the tether and the balloon's attachment line. This allows the blimp to easily turn into the wind. Fishing line works well as a tether; so do some kite lines and other synthetic ropes. I've used 100-pound fishing line and 1/8-inch nylon rope.

Stranded wire is a lightweight antenna conductor. Aluminum welding wire is light and strong enough to use. (I use phosphor-bronze wire that was originally intended for use with the "Gibson Girl" rescue unit's kite- or balloon-supported antenna. It comes on a small reel. The irony, weight and low cost appeal to me.) To save time, measure and mark the antenna length before construction begins. I have found no instances where the addition of a tether created a problem, in launch, retrieval or during flight. The antenna has never broken; and our balloon has never broken free. Obviously, it's simple insurance, and you don't want your antenna wire touching power lines or other hazards, which might happen with a runaway blimp.

The original installations at N4ZC were all within a working pasture (see Figure 4). On-ground radials were easier to install than elevated radials, albeit more labor-intensive (easier because they need no supports, labor-intensive because it requires more radials—4 versus 60). A large diameter ring of copper water pipe served as the radial connection point and made soldering easier. The ring also created a work space to stand in during set up and launch (see Figure 5). (Future installations will use a high, central wooden post, with attachment points for the antenna, tether and four elevated radials.) Our installation is over 150 feet from the nearest tower (the 20-meter array). RG-213 runs from the feed point to the base of that tower, where it connects to Hardline running to the shack.

Hints

Always wear gloves when launching and retrieving the blimp. Let the blimp rise slowly; retrieve it slowly. The lift can be terrific, and the antenna wire and tether can easily cut your hands. Always have helpers coil up the antenna and tether on spools as you walk the balloon down. (I recommend reeling in the blimp during daylight hours. It's a tempting target, and we hear gunshots in the woods all the time at N4ZC's somewhat country location. Make certain winds cannot push the blimp into nearby objects, even when it's near the ground.)

FAA regulations for flying tethered balloons—those under five pounds in weight—are clear: No person may operate a kite or balloon in a manner that creates a hazard to persons, property or other aircraft.

Avoid all dangerous situations: operation near airports, in wet or stormy weather, near electric power lines, over public streets or areas congested with people or at extreme heights (which is probably the first thing a ham will think of). Use common sense and play it safe. Using less than the legal limit with a single 1/4 vertical supported by a 10-foot blimp, I've been able to work any station heard on 160 meters.

Sources

Here are names and contact information for two suppliers of these inflatable blimps:

Toy-Tex Novelty Company, 7315 N Linder, Skokie IL 60077.

The Blimp Works, 156 Barnes Airship Dr, Statesville NC 28677.

Annotated Bibliography

There's a wide variety of applicable information and material available, but not much has appeared in Amateur Radio publications. This is new territory, in many ways, that's worthy of further experimentation. I recommend the following books and articles for further reference and study—before buying, building and flying:

R. Carleton Greene, W8PWU, "More On Balloon Supported Antennas," *QST*, Nov 1940, pp 38, 39 and 82. One of the earliest articles on using meteorological balloons as antenna supports, complete with lifting charts.

David T. Ferrier, W1LLX and William G. Baird, W9RCQ, "A New Kind of Skyhook," *QST*, Oct 1946, pp 24-25. The original use, as best I can determine, of the Kytoon as an antenna support. Although focused on emergency or Field Day use, the final line of the article presents this challenge: "Perhaps long-wire vertical antennas can now be exploited with outstanding results."

Stan Gibilisco, W1GV, "Balloons as Antenna Supports," *The ARRL Antenna Compendium*, Volume 2, (Newington: ARRL, 1989). A brief overview of the topic, with good descriptions and several safety tips.

Maxwell Eden, *Kiteworks*, (New York: Sterling Publishing Company, 1989). An excellent book about kites of all kinds—including building and flying them. It's filled with tips and techniques.

Will Hayes, *The Complete Ballooning Book*, (Mountain View, California: World Publications, 1977). Although focused on large hot-air balloons, the history, basics of flight and safety chapters are well worth your time. It also includes the FAA regulations.

John Belrose, VE2CV, "A Kite-Supported 160- (or 80-) Meter Antenna," *QST*, Mar 1981, pp 40-42. A unique application of the Parafoil—a special type of kite. Interesting tips and techniques applicable to blimp-supported antennas.

"160-Contest Results," *QST*, Jun 1976, pp 71-74. Pictures and text from W8LT, the Ohio State University Amateur Radio Club, using an original Kytoon to support a vertical antenna.

Balloon Skyhooks

By Jack M. Hughes, WB6SOI

I appreciated and enjoyed "A Skyhook for the '90s."[1] The company I work for, TCOM, L.P., has been doing this sort of thing commercially since 1981, but on a much larger scale. Since 1972, TCOM, L.P. has been an authority on and builder of aerostats.[2] We manufacture and market an aerostat system called the Tethered Aerostat Antenna Platform (TAAP).

VLF and LF transmitters and receivers are used by the military and governments worldwide for high-reliability communication of strategic information. A typical antenna system for VLF/LF communication is a complex array of wires suspended on insulated towers that take months to erect and place into operation. A unique scheme for the rapid deployment of a complete VLF/LF (up to 100 kW) communications system uses the TAAP.

An aerostat is employed to elevate an encapsulated vertical antenna/tether to a height approaching one quarter wavelength for the frequency in use. At VLF, this technique enables the system to operate with a high antenna-radiation efficiency unobtainable by other land or sea-based antennas.

The radiated-power efficiency is further enhanced by the installation of a simple groundplane consisting of eight radial wires emanating from beneath the mooring-system trailer. The radial ends are secured by ground rods and copper wire. The entire system—antenna/tether, aerostat, mooring system, transmitter and equipment shelter—is designed to be transportable by land, air or sea and to be rapidly erected at a selected site.

The aerostat is a helium-filled, aerodynamically stable tethered balloon typically 32 meters long. Its antenna/tether is composed of Kevlar with optional fiber optics and two concentric outer sheaths of aluminum braid. The Kevlar construction provides up to five times the strength of a similar-weight steel cable. A fiber-optic core provides for secure communication between the ground and the aerostat, and the double-aluminum sheaths act as the radiating elements—as well as a path for the 400 Hz electrical power to the aerostat.

The aerostat is launched and retrieved at a rate of 200 feet per minute and it does not vent helium. In addition to the helium chamber, it uses an air-filled ballonet system that expands and contracts depending on the outside air pressure to keep the aerostat rigid at all times. It can operate in winds of up to 50 knots and will survive in winds of greater than 70 knots. The system is designed to fly continuously for up to 21 days before replenishing it with helium.

I commend author Don Daso for his creativity and resourcefulness. He seems to have just as much of a thrill as we do in handling LTA vehicles.

---

[1] Don Daso, K4ZA, "A Skyhook for the '90s," *QST*, May 1997, pp 31-33.

[2] Aerostats are essentially helium-filled, aerodynamically shaped balloons that are tethered to ground. They're not cheap—a 15-meter-long Aerostat system costs about $175,000.

By Robert Victor, VA2ERY From QST, July 1998

The Clothesline Antenna

Dry your laundry or work DX. Could this be the first dual-purpose antenna?

Every once in a while, you run across an idea that seems so simple, so obvious, you can't believe it hasn't been done before. Surely (you say to yourself), you're not the first person in the universe to have thought of this…

That's the case with the Clothesline antenna and me. While mulling over a variety of ideas for an antenna suited to my apartment, I started drawing some graphs of sine waves at various frequencies. One thing led to another and I wound up with a terrific solution for my antenna needs—and it was one that I haven't found any references to anywhere. As far as I'm concerned, *I've* invented the Clothesline. Still, I'm not going to be surprised if someone shows me that it's been done before!

But even if this antenna design has been around since hydrogen, it may still be new to you. And even if you have seen something similar to the Clothesline, this design may be worth a second look. For hams in a variety of situations, this could be just what the doctor ordered to cure your DX dilemma.

What It Is ... And Isn't

I will tell you what the Clothesline is, but first I'll tell you what it is not. The Clothesline isn't some dubious trick for loading up an actual clothesline with 12 cubits of RG-213 looped four turns to the foot around your washing machine. This is an antenna, it works all the HF bands, and it gets great results. It doesn't need a tuner because it's dead-on resonant on the 160, 80, 40, 20, 15 and 10 meter bands. A tiny tweak will bring in 12 and 17 meters, too.

It's remarkably easy to build. There are no traps, no stubs, no loading coils, no variable or fixed capacitors, no screws, no clamps, and you don't have to drill anything. This antenna is so simple to put up, it hurts. The Clothesline consists of little more than a piece of wire, a center insulator, some feed line, and a couple of $2 hardware-store fittings.

Too good to be true, you say? What's the catch, you ask?

Well, you *do* have to adjust this antenna for most band changes. But before you get too excited by the word "adjust," I'm not talking about unwrapping yards of black tape to get at a loading coil, or tweaking the bare shaft (whoops, hot side—sorry!) of a 50-year-old fleamarket variable capacitor.

No, the Clothesline makes band changes a piece of cake. When you see how easy it is, you're going to chuckle!

The Concept

The easiest way to explain how the Clothesline works is to lead you through the simple reasoning I used to come up with the thing. I started with a drawing like the one in Figure 1. It shows a simple dipole for 80 meters. It's 132 feet long and fed in the middle with coax. Next I drew the graph of the voltage distribution along the antenna (at its fundamental resonant frequency). Note the voltage is at a maximum (with reference to ground) at the antenna ends. That's important for any dipole radiator. The ends of the antenna have to correspond to the high voltage points of the curve.

I already knew that the curve the voltage wave describes is a quarter of a full cycle—a quarter wavelength—on each half of the antenna. And that the distance between the two places where the voltage is highest, or in other words, from one end of the antenna to the other, is $\frac{1}{4}$ plus $\frac{1}{4}$—a half wavelength. That's why they call it a half-wave dipole.

I also understood why the feed point is in the middle. It's located precisely where the voltage curve is lowest with respect to ground. To me, an antenna is just like a transformer—it transforms the low impedance coming out of the coax up to the impedance needed to couple to the cosmos. Come in low, go out high. So much for the plain dipole.

BEASLEY
K6BJH

THAT LAST PAIR OF LONG JOHNS YOU REELED OUT MOVED MY FEEDPOINT

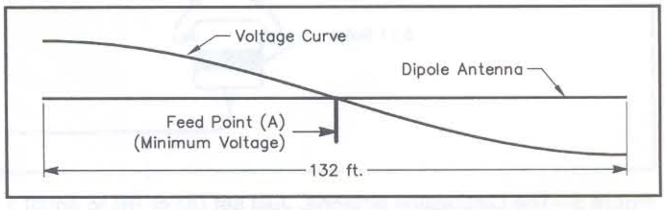

Figure 1—The voltage distribution along an 80-meter dipole antenna. Note that the voltage is at maximum (with respect to ground) at the ends of the antenna.

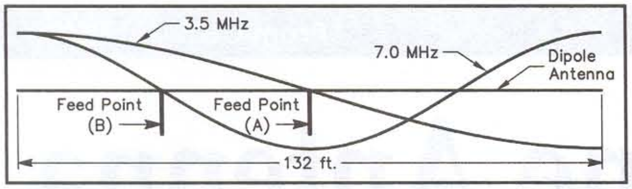

Figure 2—Take the 80-meter dipole shown in Figure 6 and overlay the voltage distribution for a 7-MHz signal. Note how the 7-MHz feed point (B) has shifted to the left of the 3.5-MHz feed point (A).

I then dropped a 7-MHz voltage curve over the original 3.5 MHz length (see Figure 2). Because the new frequency is a whole multiple of our original frequency (3.5 MHz × 2 = 7 MHz), there is still a whole number of half wavelengths from one end to the other, so the voltage peaks are still at the ends. This means this same length of wire will resonate at 7 MHz, which means it will radiate perfectly well at this frequency too.

But I noticed right away that something had changed from our earlier situation—the point where the voltage minimum crossed ground was no longer in the middle. There were now two such points, neither of them anywhere near the middle. So the antenna was resonant—the voltage peaks were at the ends—but my original feed point was now useless! Alright, I reasoned. If I shift the feed line to the new minimum I'll have the perfect setup—high voltage at the ends, low voltage at the feed point. Look out DX, here I come!

I thought about how far I'd moved the feed point. Looking at the curve, you can see that from (A) to (B) is a quarter wavelength at the new frequency, right? We doubled the frequency, which means we cut the old wavelength in half. So what used to be a quarter wave at the old frequency—from the middle to one end—will be twice that, or one-half wavelength, at the new frequency. And I wanted to get the feed point over to the new location at point (B), so I move it half that half—a quarter wavelength—over to the new minimum at (B).

I also realized that I'd added a quarter wave to the right half of the antenna, making the right-hand side $^1/_2$ plus $^1/_4$ wavelengths, or a total of $^3/_4$ wavelengths long. Since $^3/_4$ on the right plus $^1/_4$ on the left equals 1, I now had a full-wavelength antenna with some gain compared to a dipole, fed one-quarter wavelength from one end.

I kept going. I doubled the frequency again, and the same thing happened. The voltage peaks would stay at the ends, but I'd see the feed point moving further to the left by one-quarter wavelength at the new frequency for each time I doubled frequency.

I tried it for other multiples—by 3, 5, 6, 7 and so on. Just as I expected, I always wound up with some whole number of half waves, and voltage peaks at the ends. All I ever had to do was to move the feed point to the new quarter-wave point of the antenna. And for that matter, for any band other than 80, I had a choice of feed points. Each higher-band antenna had more than one voltage minimum point in its length, and I could feed at any one!

Now I drew a bunch of these curves like the ones shown in Figure 2 for 80, 40, 20, 15 and 10 meters. I dropped reference lines down from each voltage minimum to show where the feed points would lie along this single piece of wire, depending on the frequency. Because there are multiple feed points for each of the higher bands, I could see that any time I want to change frequency, I would just shift to the closest feed point for the next frequency, and away we go!

But so what? How was I going to make a simple antenna with all these feed points? For that matter, what was simple about an antenna with all these feed points anyway? The answer was, "nothing." What about a *single* feed point that somehow travels or slides along the antenna to anywhere I want it to go? Could I make that happen? I was looking out the window at my 40-meter folded dipole when I came up with the answer.

Other than the fact that it's fed with 300-Ω line rather than coax, the folded dipole is almost identical to the standard dipole. It resonates over the same lengths at the same frequencies, so my marvelous (but yet to be invented) multi-band, sliding feed-point antenna would be just as comfortable with a folded dipole 132 feet long as it would be with a plain dipole. But what about the sliding feed point?

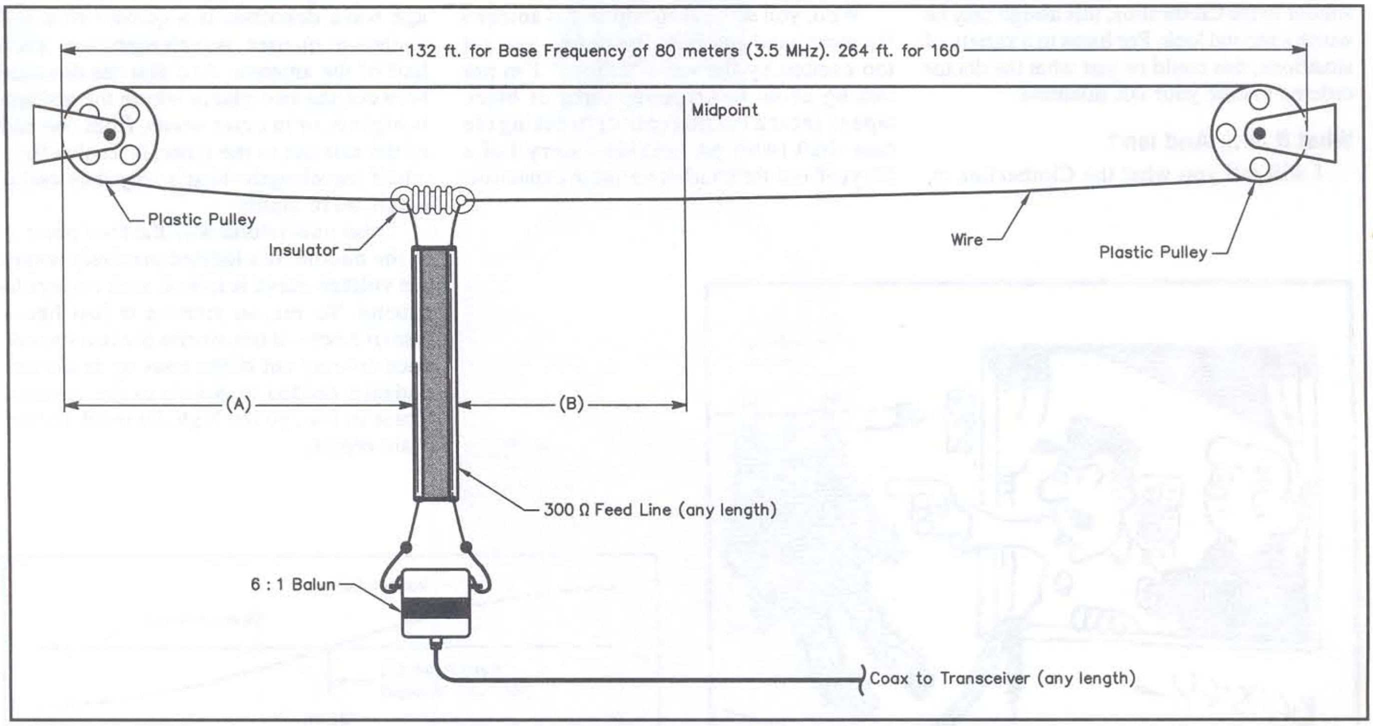
132 ft. for Base Frequency of 80 meters (3.5 MHz). 264 ft. for 160
Midpoint
Plastic Pulley
Insulator
Wire
Plastic Pulley
(A)
(B)
300 Ω Feed Line (any length)
6 : 1 Balun
Coax to Transceiver (any length)

Figure 3—The Clothesline antenna. Just set (A) or (B) to equal a quarter wavelength at your operating frequency. As long as it is a multiple of your base frequency (most HF ham bands are), your antenna is tuned. You can measure out your favorite bands, peak them by observing the SWR, and put color marks or tags on the bottom wire. Roll up to the mark and you're ready to couple to the cosmos!

What would happen with my folded dipole if I just grabbed the feed line and tried to slide it along toward one end? It hit me: If the endpoints were run over pulleys the feed point would effectively slide! It would just run along the bottom of the antenna to anywhere you cared to put it. As far as the electrical nature of the antenna is concerned, absolutely nothing has changed—except the position of the feed point! And there it was—the birth of the multi-band, sliding feed-point antenna, otherwise known as the Clothesline! See Figure 3.

But wait a minute, I thought. It's a clothesline. It can't possibly work. It's too simple!

Yet no amount of poking about on paper could find fault with it. Still very skeptical, I strung up a miniature version (for 10 and 20 meters) running down my hallway. It worked! It loaded like a breeze, with the feed about 8 feet from the end on 10 meters and about 16 feet from the end on 20 meters, just where those graphs said it should. I bought some decent wire, a couple of plastic pulleys, 75 feet of 300-Ω twin lead, and threw one up on the roof to cover 40 through 10 (66 feet is all the horizontal space I have up there). Within a couple of days I had racked up dozens of DX contacts on every band. Most remarkable was the apparent gain on the higher bands. I consistently surprised other stations (and myself) with my signal strength, usually trading equal signal reports and getting lush praise from guys running far greater power and fancier antennas. My 100 W and a Clothesline are reaching the world!

How to Build and Use the Clothesline

Measure out enough wire for twice the lowest frequency. If this happens to be 80 meters, you'll need 2 × 132 feet (plus a bit for trimming). String it up like a clothesline, using nylon or rope leaders to attach the plastic pulleys to their supports. Just make sure that all the wire is used, so that the antenna is full-length. You can use long leaders if you want to hang one end off a distant support. Run each end of the wire through the pulleys top-to-bottom, attach the two ends to a center insulator and hook up your feed line. That's it!

If you're building a version for 80 or 160 meters, you'll want to install a third little pulley to ride along the top run, attached by a short piece of cord to the center insulator, to help keep the top and bottom runs of the antenna roughly parallel. Build it for 40 as the lowest frequency and you won't even need this little extra.

There is a full set of feed points between the middle of the antenna and half-way out to either end, so if you want to keep the feed line short, you can cut it to provide just enough slack to move within this range. Alternatively, because different feed points for the same frequency can affect the gain pattern, you may want the flexibility of having a choice of feed points for each frequency, and decide to leave more line slack. Losses in twin lead are so much lower than in regular coax, you can use all you want (within reason). Do be careful to stand it off from any metal it encounters on the way into the shack.

Inside, the simplest way to match the twin lead to your rig is with a balun. I use a 6:1 balun to get the impedance down to the 50 Ω my radio likes. A 4:1 balun works fine too, though the 6:1 usually presents a better match. Both baluns are sold by a number of *QST* advertisers. Ditto for the 300-Ω twin lead.

Speaking of the shack, if yours is on an upper level at your home, consider setting the Clothesline up with one end near a shack window. You'll be able to reach out, retune (that is, haul in the line), then go back to your rig to check your SWR meter.

If you don't have this luxury, don't sweat it. I calibrated my Clothesline by setting my rig on tune-up power for each band, and going up to the roof of my apartment building to peak the antenna with a field-strength meter. I confirmed each setting by checking the SWR back in the shack. On every band I tuned for, it was well below 1.5:1. As I found each feed-point setting, I marked them with various colored indelible markers for each band. Now when I want to change bands, I just go up to the apartment roof, run the Clothesline out to the right color, and I'm tuned! It takes seconds.

I can imagine other settings where you'd run the Clothesline right outside a window, out to some convenient support. Or vertical, up or down the side of a building, a tree, or a flagpole (tie a flag to it, if you like). If you're camped out in suburbia with difficult, anti-antenna neighbors, a detachable feed line using alligator clips makes it the perfect disguised antenna. Who's going to suspect a clothesline? You could even use it during the day to… wait for it… *dry clothes!* You could even make it out of your standard, garden-variety hardware-store clothesline kit—plastic pulleys, vinyl-covered wire, the works. This makes the Clothesline the only ham antenna kit I know of that you can buy complete (minus the feed line) from your local hardware store. (In a perfect world, your local hardware store would stock twin lead.) Yes, I know I said at the beginning that this wasn't a trick for loading up a real clothesline and I've been true to my word. This is an antenna that happens to *look* like a clothesline.

I mentioned covering 17 and 12 meters. If you cut the original length to resonate just a little up from the bottom of 80, at 3.615 MHz, the fifth harmonic is smack dab on 17 meters. Cut it to 3.55 MHz and harmonic number seven is on 12 meters. In fact, these are such minor variations that you can surely find a convenient center frequency on 80 that'll put you right where you want to be on all bands. Though you don't need it, a tuner (either in or out-board) can get you right down to zero reflected power. Remember, feeding with twin lead or other balanced lines keeps your losses way down, so a little elevated SWR at the band edges is no big deal.

I'm now thinking about some kind of motor drive that'll permit me to tune up from inside the shack. My first reaction to this idea was, if I'm going to get into motor drives and the like, is that any simpler than a beam? I now think the answer is "yes." The drive would be simpler, in that I wouldn't need any position feedback—all I need to do is watch my reflected power meter to know when I've hit resonance. It doesn't need a whole lot of travel because a full range of feed points is available over just a quarter-length of the antenna. As well, few beams are all-band. And since the antenna you can put up is always simpler than the antenna you can't, for those of us in apartment settings where a beam or other, more complex antenna systems are out of the question, a motorized Clothesline might well be worth the effort. I've got a sneaking suspicion that, somewhere out there, there's a cheap, off-the-shelf drive unit with "Clothesline" written all over it!

By Ralph O. Williams, W8AJC **From *QST*, July 1949**

A Variable-Frequency Antenna

One to Ten Meters with a Single Antenna System

Multiband transmitters and band-switching receivers make it easy for us to jump from band to band, but the antenna has not kept pace. Many of us are prevented from operating on several bands by the thoughts of the multiplicity of antennas demanded by the conventional approach to multiband work. Here, to go with the VFO, is the VFA — tunable to resonance from the operating position. It covers 11, 10, 6 and 2, and all the television and f.m. bands in between, with the optimum performance all along the line. Ideas for the lower frequencies are included, too.

It has been the writer's lifelong ambition as a radio amateur to have a universal antenna; one that would not only work on several bands, but also tune within a band, providing optimum operating conditions on any frequency. The memory of endless trips to the rooftop or out to the mast to lower the antenna and cut off or splice on a few inches of wire to hit a special spot in the band is still fresh in mind. The old Zepp was pretty good but it required spaced feeders and tuning at the transmitter end. The half-wave aerial split in the center for a 72-ohm transmission line required no tuning at the station end but necessitated a different antenna for every band. Often the resonant frequency of these antennas varied widely from the values indicated by the formulae because of conditions not always apparent, and under some circumstances it was difficult to get adequate loading over an entire band. As the years passed matching systems were introduced, and they, too, are usually one-band devices. With all their disadvantages the antenna systems were not too bad back in the days of crystal control, but now we have VFO and often operate anywhere within the band. It goes without saying that what we need to go with VFO is a good VFA!

The group, other than amateur, most in need of a VFA, is that vast population trying to receive the various television and f.m. channels on a single antenna. A good half-wave dipole will outperform most of the existing elaborate receiving antennas providing, of course, it is possible to adjust the antenna accurately for each channel.

For years I have been giving thought to ways and means of feeding out and retrieving wire to make an adjustable center-fed half-wave antenna. The increased use of the VFO and the advent of television and f.m. broadcasting have made such an antenna practically a necessity. The folded dipole makes it a possibility, with rigging less complicated than the dial drives on some broadcast receivers! It is the purpose of this discussion to describe a remotely-controlled center-fed antenna capable of continuous adjustment from band to band and within the bands, using an untuned transmission line.

The Folded Dipole as a VFA

The folded dipole lends itself admirably to rigging with cords and pulleys so that the length of the flat top may be varied simply by pulling down on the feeders (see Figure 1). If the spacing between the center pulleys is adjusted so that the characteristic impedance of the two parallel conductors is 300 ohms the feeders will match 300 ohm ribbon regardless of their length and the length of the antenna. In this way, we have a tuned antenna with a flat line for any frequency, accomplished without resorting to sliding contacts at any point in the system. Figure 2 shows a practical hand-operated rigging that will enable the experimenter to make a set-up and observe its characteristics. Antennas of this type may be made large or small but it is suggested

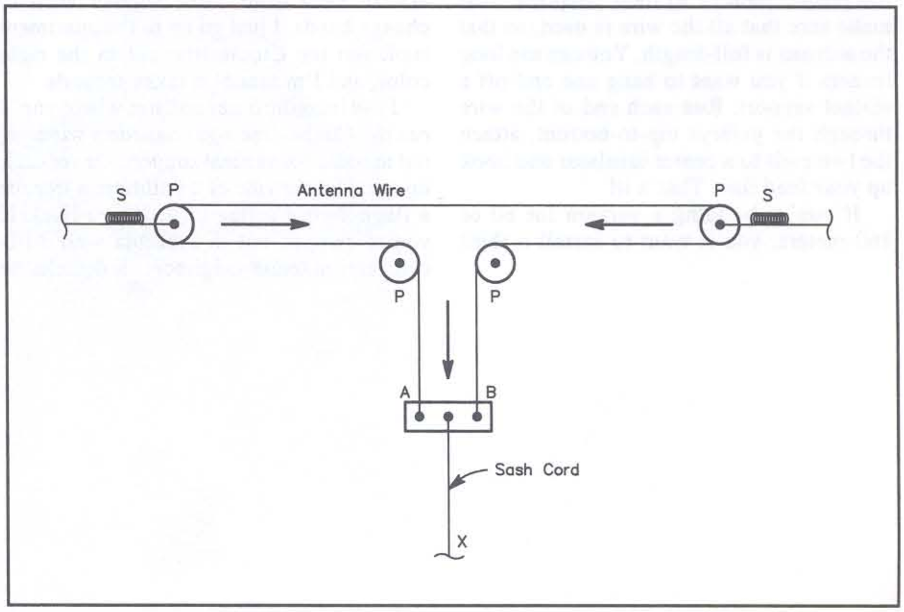

Figure 1—Basic principle of the adjustable folded dipole. Dipole and feeder section are made of one piece of flexible wire. Antenna length is changed by pulling down on the feeder at point X. Spacing of the pulleys at the center is such that the characteristic impedance of the feeder section is 300 ohms. Twin-Lead is connected at points A and B.

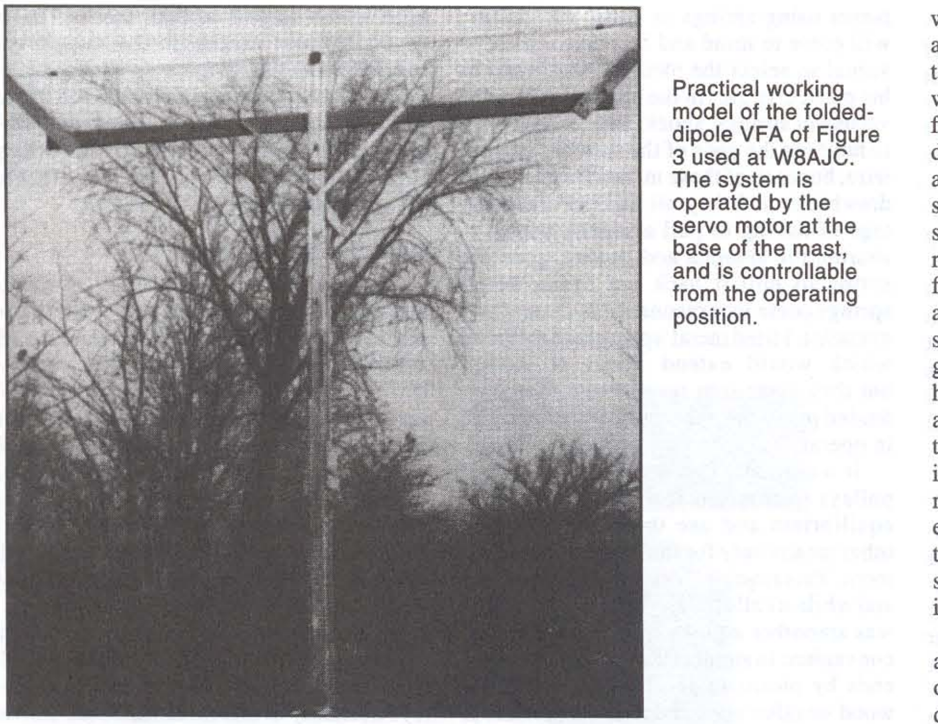

Practical working model of the folded-dipole VFA of Figure 3 used at W8AJC. The system is operated by the servo motor at the base of the mast, and is controllable from the operating position.

that the beginner make up small models for 2 and 6 meters, or simply for the f.m. and television bands, to prove the merit of the antenna.

The photograph and Figure 3 show the details of an experimental working model, made long enough to tune to 10 meters, but designed so it could be pulled down to a flat-top length of only a few inches. The tuning was fairly sharp and the results over conventional antennas for the reception of f.m. broadcasting were gratifying. A considerable improvement was noticed even within the 88-108 MHz f.m. band when the receiver was tuned to different stations and the antenna adjusted for maximum response. It was a real thrill to couple the feeders to the 10-meter transmitter and watch the plate milliammeter go up as the antenna came into resonance, and then pass through, and return for maximum; then without changing antennas, to switch on the f.m. broadcast receiver, run the antenna down to about four and a half feet to pick up a Detroit station. It was interesting to observe the effects on the received signal strength as the antenna was shortened from resonance at 10 meters to the proper length for the f.m. band, with the receiver tuned to a station on 98.5 MHz. Reception was possible with the long antenna and became good as the flat top hit fourteen feet (three half waves), falling off to a very sharp null at 121.5 inches (critical), after which it returned to full signal strength at 56.25 inches, approximately a half wave for the received signal. This ability to tune to an extremely critical null might find application in the elimination of an undesired, strong nearby signal under certain receiving conditions.

The uses to which a continuously-variable antenna may be put are limited only by the operator's imagination. Once the mechanical details have been worked out they may be used singly or in multiple, as antennas or reflectors, driven by common or separate servo motors and in various phase relations. Such antennas may be used for transmitting or receiving, or for special applications such as field operations covering a wide band of frequencies, signal-strength measurements, target transmitters for lining up rotary beams, and antenna studies. Their greatest commercial application will no doubt be in the f.m. and television fields where simplified versions, adjustable from the receiver, should find wide acceptance.

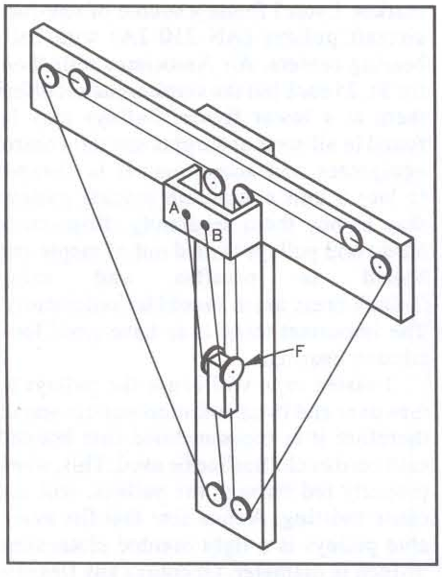

Figure 2—Hand-operated version of the variable-frequency antenna mounted on a wooden frame. To minimize mounting space, the feeders are folded back over ratio pulleys, F. The 300-ohm line connects at A and B. With this arrangement the vertical movement of pulley F is equal to the end movement of the antenna pulleys. The supporting fishline or sash cord must be taut.

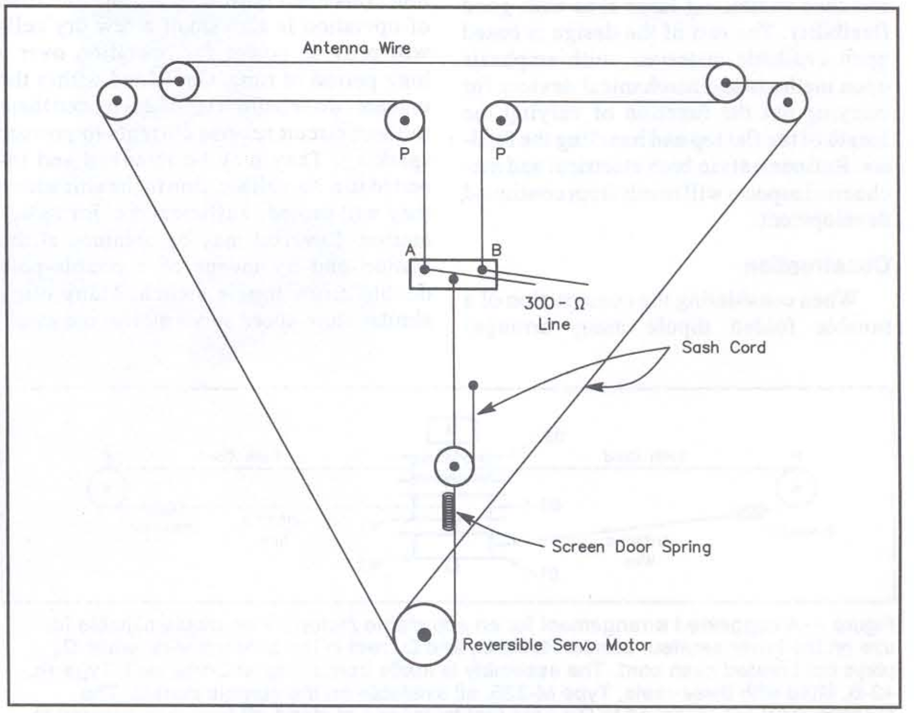

Figure 3—Diagram of the remotely-controlled antenna shown in the photograph. This system is now in use at W8AJC for 28 MHz and up. It may be reduced to a flat top of a few inches.

Other Types of Adjustable Antennas

Although the folded doublet lends itself most easily to continuous adjustment it might also be desirable to have a variable center-fed half-wave antenna suitable for use with coaxial or other low-impedance feeders. This may be done by the use of pulleys, insulators and wire, but requires a sliding contact at the center where the wind-up drums are located, and unless operated frequently it is subject to all the ills of exposed slip rings. This type of antenna is shown in Figure 4 and although at present, untried at W8AJC, it may prove worth-while on the lower frequencies, filling the long-felt need for a tunable 80-meter antenna that will work on 40 and 20 also!

Folded-Dipole Design

Useful information relative to the design of folded dipoles may be found in the ARRL *Handbook* for 1948, War Department *TM11-466*, RCA's little book *A Practical Analysis of U.H.F.* and in a paper by W. Van B. Roberts appearing in the *RCA Review* for June, 1947, page 289. The article by Roberts, in which he tells of the work done by his group at Princeton during the war under Government contract for the study of the folded dipole, is very helpful in providing the reader with a mechanism for analysis of the antenna. The literature indicates that spacing of the conductors composing the flat top should be close, in the order of 1/100 wavelength. Since our antenna is to have flexible conductors it is desirable to work with a type of wire that has a large diameter for low r.f. resistance, and at the same time has a high degree of flexibility. Such a conductor may be composed of insulated wire with a braided copper shield, the shield acting as the antenna and thus combining large area with good flexibility. The rest of the design is based upon available materials, with emphasis upon methods and mechanical devices for carrying out the function of varying the length of the flat top and handling the feeders. Refinements in both electrical and mechanical aspects will result from continued development.

Construction

When considering the construction of a tunable folded dipole many arrangements using springs or cords and pulleys will come to mind and it is up to the individual to select the method best suited to his particular use. In the beginning I tried a variety of springs, shock cords and weights to hold out the ends of the flexible antenna wire, but each of these methods had its own drawbacks and they all had the disadvantage of having to pull against a spring to shorten the antenna and depend upon the spring to pull it back out again. Metal springs come into resonance at certain frequencies. I tried metal-spring sash supports which would extend about 40 inches but they came into resonance in their extended positions and were not very smooth in operation.

It was finally decided to use cords and pulleys so arranged as to be in mechanical equilibrium and use the servo motor or other means only for the purpose of adjustment. This required less power in the servo and while it calls for more pulleys the result was smoother adjustment. I have found it convenient to support the antenna from the ends by means of insulated pulleys on a wood or other nonconducting structure. If the dipole is to be operated in a horizontal position there is no objection to using a vertical metal support pipe, but horizontal metal rods or pipes should not be used. For long antennas a center support must be provided for the feeder pulleys and the wind-up mechanism. The ends may be supported by poles, trees or buildings.

Servo Motors

The servo motor shown in the photograph is from a surplus Azon bomb tail assembly. It has plenty of power and may be reversed at will. It requires a 4-wire cable to the battery or other d.c. source. The current consumed is small and since the time of operation is also small a few dry cells will provide power for operation over a long period of time. Contained within the unit are two selenium rectifiers placed there to short circuit reverse currents to prevent sparking. They may be removed and inserted in a 30-volt a.c. line to the unit where they will provide sufficient d.c. for its operation. Reversal may be obtained at the station end by means of a double-pole double-throw toggle switch. Many other similar slow-speed servo motors are available on the surplus market, most of them reversible, and varying in size and power requirements. In some cases where d.c. is not available advantage may be taken of the gear train by connecting a universal coupling to the motor end and driving with a reversible universal fan or vacuum-cleaner motor operated from 115 volts a.c.

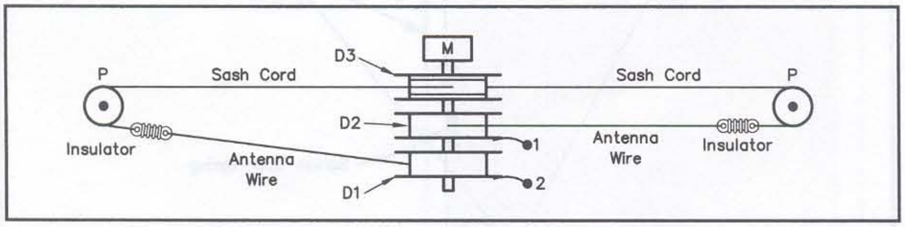

Figure 4–A suggested arrangement for an adjustable motor-driven dipole suitable for use on the lower amateur bands. Reels D_1 and D_2 reel in the antenna wire, while D_3 plays out braided sash cord. The assembly is made from a Signal Corps reel, Type RL-42-B, fitted with three reels, Type M-235, all available on the surplus market. The antenna reels are fastened to the cord reel by means of stand-off bushings secured to the center insulation. Brushes 1 and 2 connect to a 72-ohm line running to the transmitter. Success of this system depends upon maintenance of good contact at the brushes.

Sources of Materials

Antenna wire should be light, durable, flexible, of large diameter and a good conductor. For ease of adjustment it should pull around a one-inch pulley readily. Superflexible stranded copper wire of large diameter would be quite heavy whereas an insulated stranded wire, if size 20 or so and covered by a braided tinned-copper shield, would be light in weight, adequately flexible and of sufficient diameter. Belden No. 8885 shielded grid wire having an o.d. of 0. 1 inch has been found satisfactory. Too-stiff wire will make the system unwieldy.

Pulleys must be free-running for smooth operation and have as little friction as possible. A number of different kinds of pulleys normally available at hardware stores were tried and all had mechanical imperfections. Usually, though they seemed free-running when tried at the store, they turned out to have prohibitive friction when loaded and in the system. Since the number of pulleys required is fairly large and the accumulated friction may be excessive, ball-bearing pulleys are recommended. The first ones used here were homemade and turned out of fiber and used small ball bearings in the center. The ball bearings had 1/4-inch holes and a 5/8-inch o.d. and were obtained from disassembly of surplus gear trains, bomb sights, computers and other equipment so plentiful on the surplus market. Later I found a source of one-inch aircraft pulleys (AN-210-1A) with ball-bearing centers. Air Associates sells them for $1.25 each but the surplus market offers them at a lower figure. Pulleys may be found in all sorts of surplus aircraft control equipment and sometimes it is cheaper to buy a unit containing several pulleys than to buy them separately. Homemade hardwood pulleys turned out of maple and boiled in paraffin and using 1/4-inch brass axles should be satisfactory. The important thing is to have good low-friction bearings.

Twisted rope will cause the pulleys to turn over and twist and short out the aerial; therefore it is recommended that braided sash cord or clothesline be used. This, when properly fed through the pulleys, will not cause twisting. A nice size that fits available pulleys is a light braided clothesline 1/8-inch in diameter. Of course any flexible insulating line may be used such as dial cable, fishline or upholsterer's twine. Spring loading to prevent slipping because of stretching is advisable in some cases.

Limit Switches: If the antenna is not visible from the operating point Micro-Switches may be so placed that when the end of travel of the antenna is reached the circuit will be opened and the motor will stop, it being possible then to reverse it and run it to the other limit where another switch will furnish protection from overtravel. The switches should not be connected in the common lead to the motor but in the circuit controlling that direction only. (Switches and associated wiring must be placed so as not to interfere with the electrical operation of the antenna.)

Conclusion

The examples shown are but a few of many possible ways of setting up remotely-controlled variable-frequency antenna systems. It is hoped that this article will serve as a basis for further development of adjustable antennae for amateur and commercial use.

By Rod Newkirk, W9BRD **From *QST*, July 1993**

Honey, I Shrunk the Antenna!

Think small! Communicate with multiconductor miniature loop antennas.

Great energy and ingenuity go into efforts to boost the effectiveness of compact transmitting antennas—*compact* meaning configurations with circumferences of $^1/_8$ wavelength or less. By this definition, a 40-meter design could be a square with 4-foot sides. In the following discussion, I depart slightly from the square. I prefer $3^1/_2$ feet high by $4^1/_2$ feet wide, suspended in the vertical plane, for convenience in my antenna's location.

The traditional small-loop approach is to insert capacitance in series with a one-turn loop, tune it to the desired frequency, and then attempt to feed power to it as efficiently as possible–no simple matter. The antenna's Q is astronomical, current and voltage are monstrous, bandwidth is razor sharp, and its radiation resistance is ridiculously low. This lossy situation can be improved by reducing the resistance of the loop material through the use of piping, foil, etc, but the resulting plumbers' nightmares are hardly worth the pains.

A more promising avenue toward practical, simple compact antennas would appear to be in the direction of varied configuration. The goal is to "cool" the system with better power distribution, and at the same time simplify the feed. For example, vast improvement in the key parameters is possible by application of the venerable "folded dipole" technique. This can be done by adding one more turn, tuned identically, connecting the two turns *in series*, and re-resonating the antenna. Now, looking into one of the twin current nodes, the input impedance will have roughly quadrupled. The one-turn loop's skimpy 5 or 6 Ω rises to approximately 25 Ω. Adding yet another such turn doubles this to about 50 Ω–a convenient match to common coax. See Figure 1. As with the folded dipole, more than four turns are unprofitable; diminishing returns raise the input impedance. little faster than ohmic resistance. Note that the tuning capacitances, each about 40 pF, must be kept nearly equal for overall system balance.

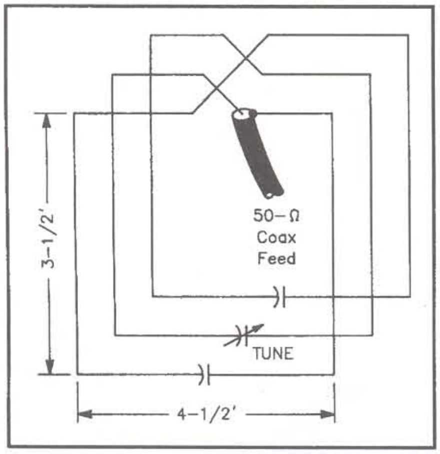

Figure 1—The three-turn loop antenna for 40 meters. Each capacitor is about 40 pF to resonate and match the antenna to a 50-Ω feed line.

Increasing Multiturn Loop Efficiency

Lots of hams use small loops in situations where they can't erect a larger antenna. Making contacts with these antennas largely depends on high radiation efficiency. Here are some things you can do to increase efficiency and effectiveness of such antennas:

- Use large-diameter copper wire. Keep the copper surface clean. Tin-plated wire is noticeably worse than copper in this application because of skin effect–most of the RF travels in the tin, which is far lossier than copper at RF.
- Use low-loss capacitors. The key here is minimizing dielectric loss, which is most favorable in vacuum- and air-dielectric capacitors. You can also parallel several capacitors to decrease losses.

Increasing the wire size used in these loops raises the antenna Q, narrowing the bandwidth and increasing the voltages and currents present. For this reason, you may have to use higher-quality capacitors than those used in the experimental versions if you increase the wire size.

Remember that high voltages can exist on a multiturn loop antenna's wires at power levels of a few tens of watts or more. Thus, it's best to keep the antenna well away from other objects.—*Roy Lewallen, W7EL*

Construction techniques, conductor material, insulation and components now become less critical because voltage peaks and current nodes are distributed throughout the system. I use ordinary zip cord, two lengths entwined, giving four available turns.[1] For proper continuity, the four wires on each side of the loop are clearly tagged at alligator-clip connectors, top and bottom centers. For 7 MHz, one turn is left open (at top *and* bottom), but will become useful on the lower bands as discussed later. A more classy construction might feature braided hook-up wire of contrasting colors. Small cardboard spreaders at the clips keep adjacent capacitors separated.

Midget receiver-type capacitors are adequate for CW power output up to 100 watts. One of these should be conveniently variable, preferably a double-gang or splitstator type, for tweaking the system to the desired center frequency as indicated by a 1: 1 SWR. At 7 MHz, each capacitance is about 40 pF in my antenna, so the old war surplus Hammarlund APC midgets do nicely. Bandwidth can approach that of a normal linear dipole if a slightly higher center frequency SWR is accepted by stagger-tuning.

Excellent 40-meter results indicated that the same loop, dimensions unchanged, could have a shot at 80 meters. At 3.5 MHz, it becomes truly a miniloop, only $^1/_{16}$ wavelength in circumference. Adding the fourth turn (Figure 2) raises the feed impedance at a slightly higher rate than ohmic resistance, so I included it. Armed with a bargain bag

The Ingredient Called *Belief*

Even though my busy family life keeps hamming time rare, the time I spend hamming has to be spent *right*. The idea isn't just to get on the air and contact anyone anywhere—that's too *easy*. The idea is to do something a hair off-the-wall and see if anyone shares the same orbit. So I'm one of those guys who calls CQ QRP for ten minutes at 10149.5 kHz even as game clogs the band's low end—someone who looks for (and *expects*) ragchews on 3560 at noon. And so I naturally was the guy to call a 5-watt, 3 × 2 × 1 CQ at 7138 kHz at around 10 PM one night last fall when the broadcasters had long since muscled straight-thinking NovTechs off to safer terrain. And so I naturally was *of course* immediately answered by W9BRD.

It figures. We've worked on schedule many times before, but he and I—he's the father; I'm the son—have never had to prearrange each new "first contact" after changes of station setup or locale. *Chance* is too facile a word for this, but then I haven't looked too hard for a better word, either. I don't need to; we just do our individual quirky things and click.

This time, we *both* had new setups: I'd just relocated the armful of gear constituting my "shack" to the bedroom and was talking to a low dipole through 100+ feet of mongrel coax; *he* was trying out an indoor, largely below-ground loop antenna—one of the versions shrunken in this article—with just a few tens of watts, tunneling past those steamroller broadcast signals as soundly as any medium-scale ham signal from Chicago.

Enjoying the magic of yet another first contact, I hardly gave his setup a second thought. But you may wonder if such antennas really work. Gooch's Paradox explains it thus: "RF gotta go *somewhere*." Seeking to soothe the linearless, towerless masses who toil with "*only* 100 watts," Newkirk's Tiresome Chant puts it, "Successful MF/HF radio communication may proceed at astoundingly low received-signal strengths." Reduced to generic essentials, it's this: The most important single factor in success with *whatever* you use is overcoming your own disbelief.—*David Newkirk, WJ1Z*

of small 1-kV ceramic capacitors, I reached 3.6-MHz resonance with a value of 100 pF for each of the four capacitances. One capacitor was then replaced with an old broadcast-style, two-gang 300-300-pF variable for tweaking. I measured the feed impedance, at the center of one of the four turns, at about 18 Ω–well within the range of simple gamma matching. So another junk-box broadcast variable, three 350-pF gangs paralleled, was suspended at the

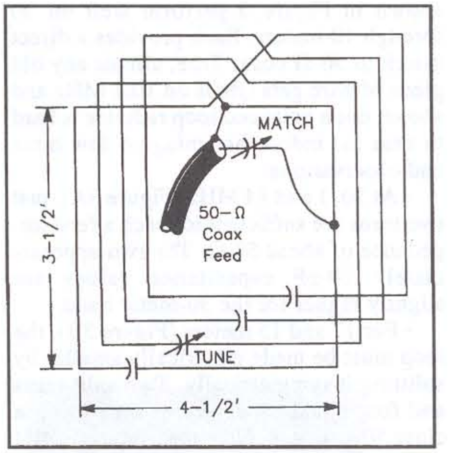

Figure 2—Extending the Fig 1 concept to 80 and 160 meters requires adding another turn to the antenna and increasing the capacitances. See the text for capacitor values.

loop's top center. Only about 650 pF was needed for a 1:1 match to coax after a 30-inch gamma lead was dangled to a tap at the center of one side of the feed turn. Then the system was tweakable over the entire 80-meter CW range. The remaining ceramic capacitors get slightly warm at 80 watts, which indicates significant power loss. These capacitors should be replaced with low-loss units for best performance, but I left them alone for my experiments. I easily contacted the East Coast and as far west as Arizona in the following few nights on the air.

My next inclination, as you may surmise, was to try the little gem on 160 meters. A four-foot-square transmitting loop for 1.8 MHz? I wasn't overly optimistic. Years of tinkering at W9BRD had failed to produce a decent indoor compact antenna, even of much larger size, for top-band work. But I dug into my bag of ceramics and gave it a go. The four-turn loop, now only 1/32 wavelength in circumference, resonated at about 1.8 MHz with four 350-pF capacitors. One of these was replaced with an old broadcast-type 500-pF variable for tweaking the center frequency between 1.8 and 1.85 MHz. I double- and triple-checked the measured feed impedance—16 Ω. Now I knew it would work. The gamma-match SWR dipped to 1:1 at about 800 pF using the same lead, tap and broadcast capacitor as on 80 meters. Immediate solid 40-watt CW QSOs with WN9W, KC4WWV, WK0B and W4VZB were most gratifying.

Variations and the Higher Bands

If such a 1/16-wavelength loop can function at 80 meters, the same should be true at 20 meters. So I wound a 1-foot-square, three-turn model on a cardboard box, hung it on a wooden bulkhead, and resonated it to 14.050 MHz with three 50-pF ceramic capacitors. This time, instead of the usual fudging capacitor, I

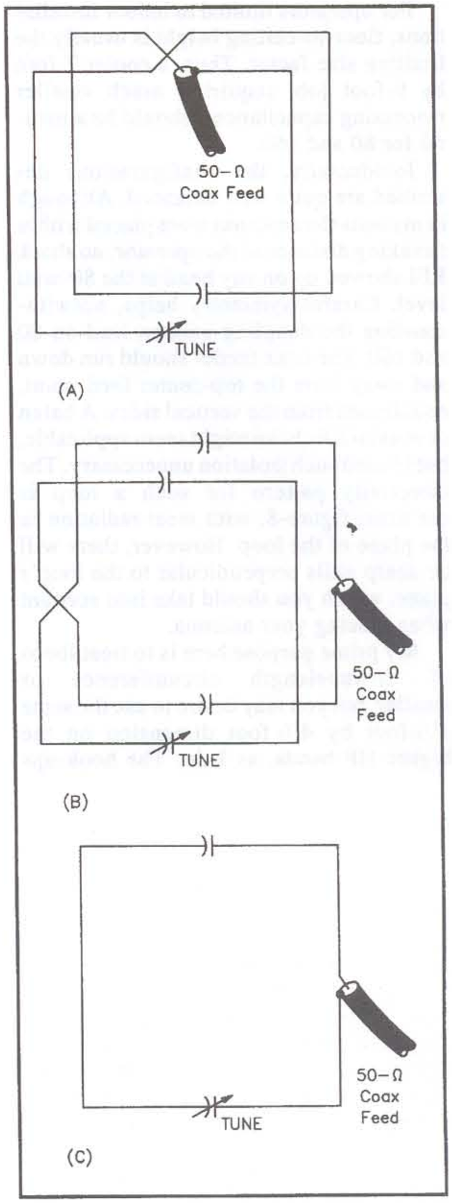

Figure 3—At A, the 30/20-meter loop; at B, the 15/17-meter version, and at C, the 10-meter variant.

spaced the turns 1½ inches apart. Center-frequency tweaking thus could be done by uniformly edging the turns away from or toward each other. This spacing approach works fine for box-wound loops, but for the lower bands, I prefer closewound zip cord. Input impedance turned out to be about 11 Ω—I should have used a fourth turn—and gamma coupling worked okay, as on 80 and 160. The 1:1 feed-line dip occurred with a 120-pF series capacitor, with a gamma lead of 13 inches. In my installation, this antenna is no low-angle DX radiator, to be sure, but I've received S9 reports during many stateside QSOs. Anyone for a 3-inch-square 2-meter version?

Antennas smaller than necessary are interesting stunts that help prove a design. However, for the lowest HF bands I recommend the largest practicable circumference.

For operators limited to indoor installations, floor-to-ceiling height is usually the limiting size factor. Thus, a cooler 7-foot by 9-foot job, requiring much smaller resonating capacitances, should be a natural for 80 and 160.

Incidentally, the configurations described are quite well balanced. Although in my tests the antennas were placed within tweaking distance of the operator, no shack RFI showed up on any band at the 80-watt level. Careful symmetry helps, notwithstanding the dangling gamma lead on 80 and 160. The coax feeder should run down and away from the top-center feed point, equidistant from the vertical sides. A balun or coaxial RF choke might seem applicable, but I found such isolation unnecessary. The directivity pattern for such a loop is the usual figure-8, with most radiation in the plane of the loop. However, there will be sharp nulls perpendicular to the loop's plane, which you should take into account when placing your antenna.

My prime purpose here is to treat loops of 1/8-wavelength circumference or smaller, but you may desire to use the same 3 1/2-foot by 4 1/2 foot dimension on the higher HF bands, as I do. The hook-ups shown in Figure 3 perform well on 30 through 10 meters. Each provides a direct match to 50-Ω coax. True, almost any old piece of wire gets QSOs on 10.1 MHz and above, but a balanced loop radiator is hard to beat for indoor hamming. A few hints and observations:

- At 10. 1 and 14 MHz (Figure 3A), just two turns are sufficient to reach a feed impedance of about 50 Ω. The two approximately 20-pF capacitance values are slightly higher for the 30-meter band.
- For 17 and 15 meters (Figure 3B), the loop must be made electrically smaller by splitting it symmetrically. Two split turns and four equal capacitances then bring a close 50-Ω match. Note that the feed point is moved to the center of one of the four vertical sides. Capacitances are in the 15-pF range.
- On 12 and 10 meters (Figure 3C), one split turn will suffice for a near-50-Ω match, with two capacitors of about 12 pF. Here, splitstator midgets can ease adjustment by minimizing hand capacitance, but it's a fancy junk box that includes them. The old APC midgets, whose rotors have little more mass than their stators, will maintain enough system balance.

Cautions and Conclusions

Like any indoor antenna, the loops I describe here generate substantial electromagnetic fields in operation. Thus, they have considerable potential to generate RFI. For this reason and to prudently avoid placing yourself or others in large RF fields, you should keep all antennas as far from consumer electronic devices and people as possible, and use the least RF power necessary to conduct the desired communications. The current editions of *The ARRL Handbook* and *The ARRL Antenna Book* cover this subject in more detail.[2]

A final comment: For overall results with compact antennas, like almost any skyhook, the higher above ground the better. In my case, radiation and reception strongly favor higher skywave propagation angles. All antenna configurations I describe here were tested and operated in a cellar ham shack, where half of the system is below ground level!

Notes

[1] See the sidebar, "Increasing Multiturn Loop Efficiency," for some ideas on optimizing such an antenna's performance.

[2] See Chapter 9 of the *Handbook* and Chapter 1 of the *Antenna Book*.

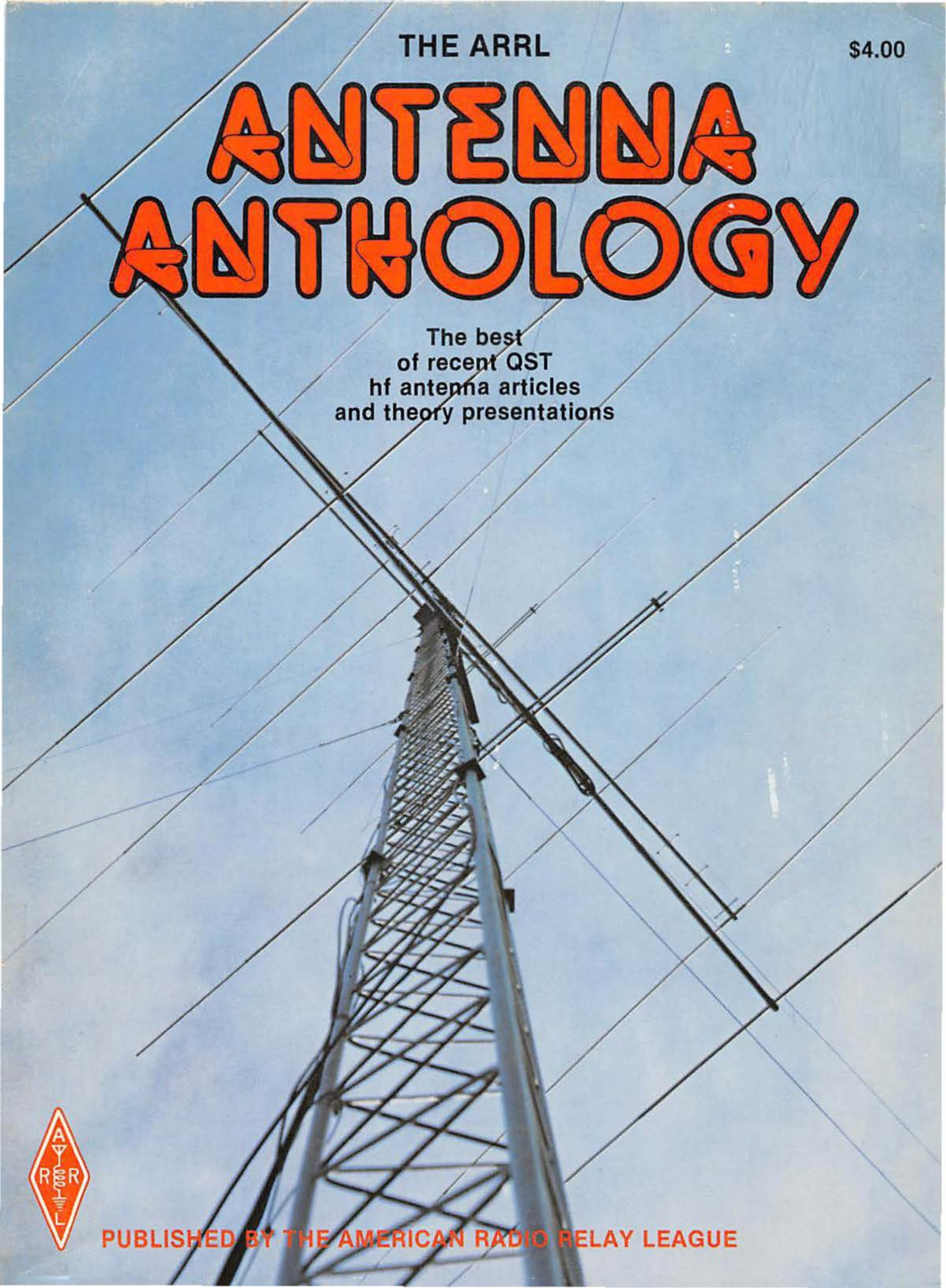

THE ARRL

$4.00

ANTENNA ANTHOLOGY

The best
of recent QST
hf antenna articles
and theory presentations

PUBLISHED BY THE AMERICAN RADIO RELAY LEAGUE

The ARRL
Antenna Anthology

Edited by
Marian S. Anderson, WB1FSB

American Radio Relay League, Inc.
Newington, CT 06111

Copyright © 1978 by

The American Radio Relay League, Inc.

Copyright secured under the Pan-American Convention

International Copyright secured

This work is Publication No. 36 of the Radio Amateur's Library, published by the League. All rights reserved. No part of this work may be reproduced in any form except by written permission of the publisher. All rights of translation are reserved.

Printed In USA

Quedan reservados todos los derechos

Library of Congress Catalog Card Number: 78-71955

$4.00 in USA
$4.50 elsewhere

Foreword

What type of amateur antenna best suits your needs? Does the available space on your property permit the installation of an effective DX antenna? Is there a compromise antenna system that will provide acceptable results over short and long distances while being reasonable in price and modest in complexity? No doubt you have come to grips with these matters repeatedly since obtaining your amateur license. So, where does one turn to find the answers to these sometimes perplexing questions? What one pre-eminent source might one consult among all of the antenna booklets published for amateurs? Why, *QST*, naturally!

This anthology contains what the ARRL hq. staff considers the best of the *QST* antenna articles and theory presentations in recent years. The contributions of numerous amateur authors and designers are contained in this volume for your convenience. This publication was developed not only to serve as a compilation of good articles, but to reach those amateurs who do not have access to past issues of the League's official monthly journal, *QST*.

The editors of this booklet have attempted to select the types of antennas that have greatest appeal to those who like to construct their own radiators, and for the many amateurs who can't justify the expense of purchasing commercial counterparts. A book of this type contains something for everyone who enjoys building and experimenting with antennas. But, if you aren't a builder you will still enjoy the hours of interesting reading provided by these works.

We wish to extend our gratitude to those *QST* authors whose articles have been selected for publication in this volume. Their enthusiasm, ability and generosity typifies clearly what Amateur Radio is all about. Should you have some original antenna ideas of proven performance, send an article to the ARRL, describing your work. Not only may your story be published in *QST* — it might appear in a future edition of this book.

Richard L. Baldwin, W1RU
General Manager

Contents

1 **Vertical Antennas**
Page 7

2 **Yagi Antennas**
Page 41

3 **Quad Antennas**
Page 53

4 **Miscellaneous Antenna Types**
Page 76

5 **Antenna Theory and Test Methods**
Page 100

Index
Page 150

Chapter 1

Vertical Antennas

One of the most practical antennas for use where limited space exists is the vertical. This type of antenna is particularly useful to amateurs who live in dense urban areas where city lots are traditionally small. The same general restriction prevails for those who dwell in apartment buildings. In such locations it is often easier to go "up" with an antenna than to erect one horizontally. But the virtues of vertical antennas extend beyond the foregoing considerations: "Verticals," as they often are called, are relatively inexpensive, easy to erect and are capable of performing well for local (stateside) and DX work. A vertical antenna exhibits a low angle of radiation when worked against a good ground system. This makes it an effective DX type of antenna.

This chapter contains reprints of some popular vertical-antenna systems that have appeared in *QST* during recent years. For the most part, the presentations contained herein have been "lifted" in their entirety, thus making it unnecessary for the reader to research the original articles for additional information.

A Two-Band Vertical for the Novice

Build this inexpensive Novice antenna. It is easy to construct and is made from readily available materials.

Knowing of a Novice who had space problems, the author set out to design an antenna for him that could be used on 40 and 15 meters. His previous antenna was a horizontal "V" whose performance wasn't worth the time it took to put it up. Because of his narrow property, the angle between the two wires was only around 40 degrees, no doubt causing considerable signal cancelation. Horizontal antennas were ruled out, so the best antenna for him was a vertical.

Knowing (or caring) little about multiband trap-loaded antennas, it was decided the simplest antenna would be a quarter-wavelength vertical on 40 meters. After a little paper work, and using formulas given in *The ARRL Antenna Book,* the correct length would be 31 feet 10 inches (9.70 m) for 7.175 MHz. This length would also work as a 3/4-wavelength antenna on 15 meters, providing two-band coverage. However, a 32-foot (9.75-m) vertical antenna made roof-mounting too difficult! Mounting the vertical to the side of the house alleviated many guying and support problems.

Materials

The first concern was what to use for antenna materials. Being almost 32 feet high, the vertical would need strength, so the writer thought of steel tubing, but that was much too expensive. With conduit there aren't enough sizes offered, and that left only one more material — aluminum tubing. The latter comes in just about any size desired, and the author had no trouble in finding just what he wanted. Going to the nearest aluminum dealer, the author found his stock to be more than ample.[1]

In order to keep the cost minimal, the author decided to make the base piece from 1-inch (25-mm) diameter thin-wall tubing. Into 12 feet (3.66 m) of this was inserted a 12-foot piece of 7/8-inch (22-mm) diameter thin-wall tubing. This piece was telescoped 2 feet (0.610 m), then bolted and clamped (after the end had been slit twice by a pair of tin snips). It was necessary to use hose clamps on the tubing because the galvanized bolts didn't provide a good bond between the pieces. The top section was 13-1/2-foot (4.11-m) piece of 3/4-inch (19-mm) diameter thin-wall tubing inserted far enough into the lower piece to make the overall antenna length about 31 feet, 10 inches (9.70 m). Because of its large size, all the work must be done while the antenna is

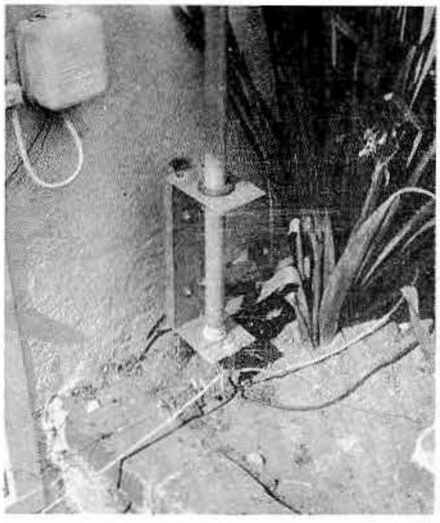
Here is the base mount, installed on the side of the house. The two white wires coming out from the base are the radials.

---

[1] [Editor's Note: TV mast material is also well suited for verticals and comes in 10-foot (3.05-m) lengths, either steel or, preferably, aluminum.]

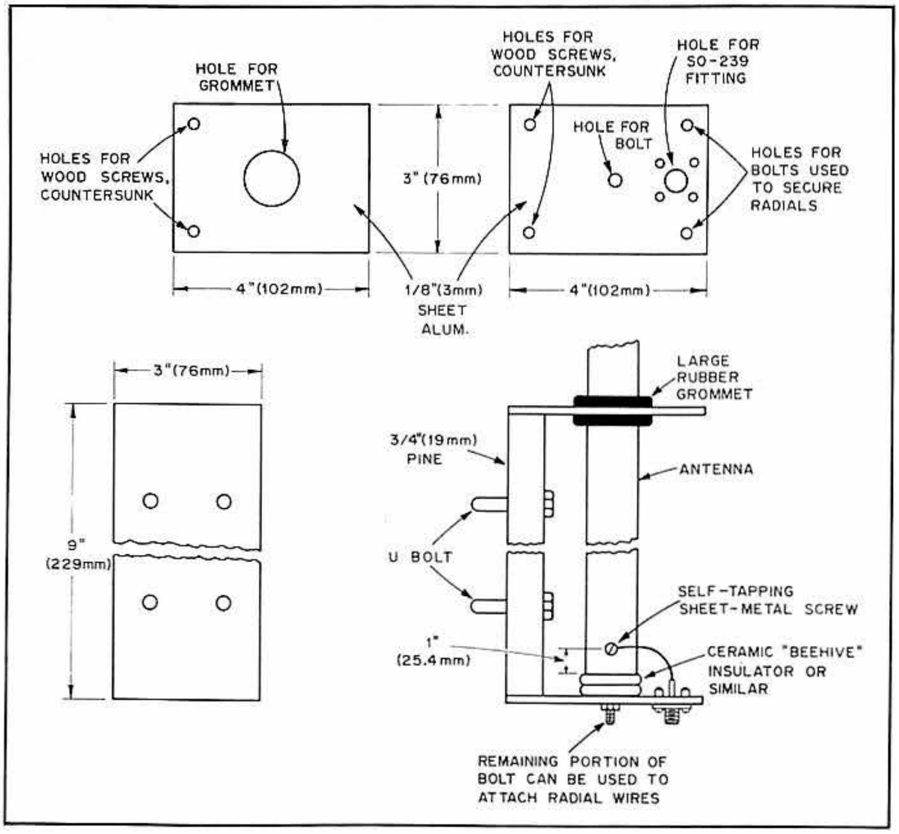
HOLE FOR GROMMET
HOLES FOR WOOD SCREWS, COUNTERSUNK
4"(102mm)
HOLES FOR WOOD SCREWS, COUNTERSUNK
HOLE FOR SO-239 FITTING
HOLE FOR BOLT
3" (76mm)
HOLES FOR BOLTS USED TO SECURE RADIALS
1/8"(3mm) SHEET ALUM.
4"(102mm)
3"(76mm)
9" (229mm)
LARGE RUBBER GROMMET
3/4"(19mm) PINE
ANTENNA
U BOLT
SELF-TAPPING SHEET-METAL SCREW
1" (25.4mm)
CERAMIC "BEEHIVE" INSULATOR OR SIMILAR
REMAINING PORTION OF BOLT CAN BE USED TO ATTACH RADIAL WIRES

Fig. 1 — Details for making the base-mounting section of the vertical.

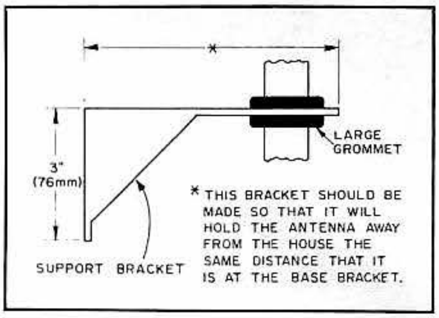
*
3" (76mm)
LARGE GROMMET
* THIS BRACKET SHOULD BE MADE SO THAT IT WILL HOLD THE ANTENNA AWAY FROM THE HOUSE THE SAME DISTANCE THAT IT IS AT THE BASE BRACKET.
SUPPORT BRACKET

Fig. 2 — Details for the support bracket.

lying on the ground. Then, the antenna must be disassembled and put up piece by piece. (Do not attempt to put the antenna up in one piece! While the weight isn't excessive, the tubing cannot stand the strain.)

Putting It Up

The original base bracket was made from 3/4-inch (19-mm) pine, but during the first rain the wood got wet and the bracket turned out to be a very poor insulator. So the unit was redesigned to use two pieces of 1/8-inch (3-mm) thick sheet aluminum, a large grommet, and a ceramic standoff insulator (see Fig. 1).

Assembling the two-band vertical was quite simple. First, a 4-foot (1.22-m) stake was driven into the ground about 2-1/2 feet (0.76 m), and about four or five inches (102 or 127 mm) away from the side of the house. Next, the bottom section (with the bracket attached) was slipped over the stake and tightened in place with the U bolts. Then the second section was inserted into the first, clamped and bolted in place. A support bracket was made from scrap aluminum, and a large grommet was installed on it to insulate the antenna from the brace (see Fig. 2). This assembly was then attached to the side of the house (near the gable) with wood screws. Last, the third section was forced through the grommet in the second bracket, and was then inserted into the second piece and clamped. (If adjustment of resonance is desired, it is better not to bolt the final section). A suitable cap was mounted on top of the antenna to keep out moisture.

Tune Up

RG-58/U coax was then connected between the transmitter and antenna with an SWR bridge in the line, and power was applied at 7175 kHz. The SWR was found to be 3:1. This was without any ground system on the antenna. In order to obtain a better match and reduce the SWR, it was apparent that a ground system was required.

According to *The ARRL Antenna Book*, two methods of making a suitable ground system can be used. One way is to connect the coax outer shield directly to ground at the base of the antenna, via a ground rod. However, this doesn't work too well unless the soil is extremely conductive (his wasn't). The other method is to make up some sort of artificial ground in the form of radials.

The author cut a radial from some surplus wire to the same length as the vertical section, attached the radial to the base plate and the SWR dropped to 2:1. Unknown to the author at the time, *The Antenna Book* says that the radials should be 2.5 percent longer than the vertical itself, so actually, they should have been made 32-feet 7-3/4-inches (9.95-m) long. A second radial, similar to the first, was attached and the SWR dropped further (to about 1.4:1). When the soil was wet, such as after a rain or after watering, the SWR takes a nose dive to less than 1.2:1. Since the property space was limited, two radials were all that could be used. Ideally, the most perfect system would consist of a metal disk, one quarter wavelength in radius, buried a few inches below the earth's surface, but this approach is rather impractical! But if the antenna was to be erected in an area *away* from buildings, and guyed, a ground system could be made of chicken wire or other similar material. However, the more radials, the better the artificial ground. Also, they should be laid out in as nearly a spoke-like pattern as possible. Hy-Gain recommended for their AVQ series antennas that a radial system be used consisting of at least two radials for *each* band used. However, for this installation the 40-meter radials operate on their third harmonic at 15 meters, requiring but one set of wires.[2]

The length specified makes the antenna resonant at about 7175 kHz on 40 meters. On 15 meters, at the third harmonic, the antenna is resonant at about 21,425 kHz, but power was applied as low as 21,105 kHz,

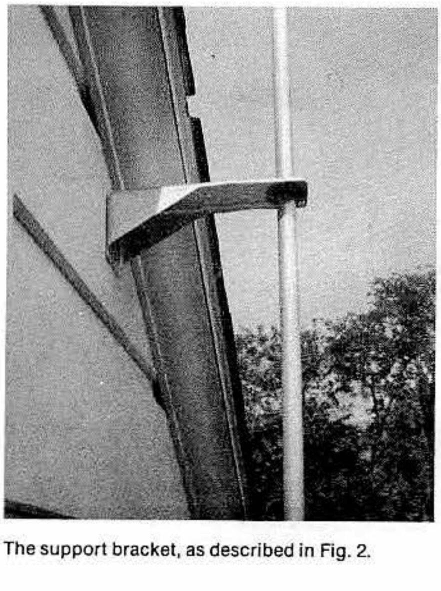

The support bracket, as described in Fig. 2.

---

[2] [Editor's note: Actually, *any* radials, regardless of length, will help the performance of the system. Also when used on 15 meters, the antenna would be three quarter wavelengths long with low-impedance feed, approximately 50 ohms. However, the angle of radiation from the antenna will be higher than that of a quarter-wavelength vertical.]

and the SWR appeared to be only 1.4:1, well within tolerable limits. Bandwidth on both Novice bands was essentially flat.

Additional Information

Total cost will depend greatly on the individual's ability to scrounge parts. If the builder prefers to purchase the principal materials, this list will be useful.

3 — pieces thin-wall aluminum tubing, 1- (25-mm), 7/8- (22-mm), and 3/4-inch (19-mm) diameters. 12 (3.66 m), 12 (3.66 m) and 13-1/2 feet (4.11 m) long, respectively.
2 — U bolts.
1—SO-239 connector.

The other material, including three bolts and nuts, six wood screws, two grommets, two hose clamps, wood and aluminum for brackets, ground stake, and cap will vary from no cost to a few dollars, dependent on available materials. If used or surplus tubing is available one could save a substantial amount of money.

In this instance, the top of the antenna extends about 16 feet (4.88 m) above the peak of the roof, but because of the bracing no guy wires were necessary. If the antenna is to be erected in a high-wind area, guys made from nonconducting monofilament fishing line (or other nonconducting line) could be used.

The owner has received nothing but glowing reports so far. Running about 70 watts input, the first CQ on 40 meters raised W9SVJ in northern Indiana from Glendale, CA. His report, RST 589. On 15 meters the owner has worked ZL3JC, AX7SM and JH1EZZ, and received very good reports.

This vertical has no gain, and probably couldn't come near a Yagi or a quad in performance, but for an antenna costing less than $10 it does one heck of an FB job. This material was originally presented in *QST* by Jerry Arnold, WA6MBP.

Build This Novice Four-Band Vertical

Putting your first amateur station together can be an expensive proposition. One way to cut costs is to keep the antenna simple. Here's how to shave the price and provide four-band operation.

Is operation with one antenna acceptable if it covers the 80-, 40-, 15- and 10-meter bands? For a new Novice that's a reasonable approach. The author's backyard is smaller than that of most urban homes, so full-size dipole or inverted-V antennas were out of the question. Not owning a tower (yet!), it seemed that a ground-mounted vertical antenna would be worth trying.

After reading *The ARRL Antenna Book*, it was decided that a ground-mounted vertical antenna would be easiest to build. Some radial wires could be buried, and the metal fence which encloses the backyard could also be hooked up to enlarge the ground system. The author preferred this type of antenna to one installed above ground, because radials of specific lengths for each of the four bands would have been needed for a roof-mounted, groundplane type of vertical. The buried wires for the ground-mounted antenna could be any convenient length, as long as the available space would permit. From what was read about these antennas, the author believed that reasonable performance could be had even if the ground radials weren't numerous and long, although generally the more, the better.

With the help of W1FB some used aluminum tubing was purchased that would telescope together and give the author a 25-foot (7.62-meter) antenna. The wall thickness of the tubing is 0.058 inch (1.5 mm). Three 10-foot (3.05 m) sections are used. The largest diameter is 1 inch (25 mm). The center telescoping section has a diameter of 7/8 inch (22 mm) and the top piece of tubing has a 3/4-inch (19-mm) diameter. This material, plus hose clamps for holding the sections together, came to $8. An old ceramic rotary switch, a coaxial connector, a feed-through bushing, and a piece of Air Dux coil stock were acquired at a flea market for an additional $3. Two medium-size, ceramic standoff insulators were donated by W1FB. He said they cost him 50 cents each at a swap session. All that remained to collect was a weatherproof box for the loading coil, some 50-ohm coaxial cable and six U bolts. The author's OM, Bob, found some used 1-1/2-inch (38-mm) steel pipe which is 7 feet (2.13 m) long. It is used as a support for the vertical.

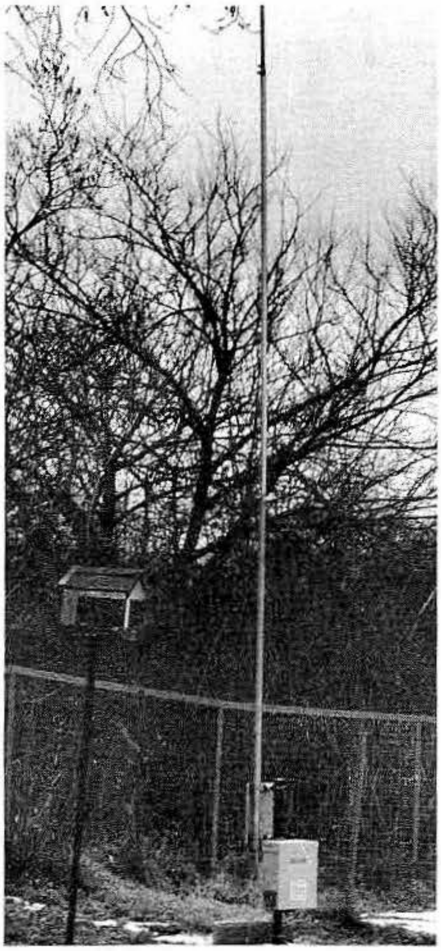

The completed 4-band vertical antenna used at WB1FSB. The box at the base of the antenna contains a loading coil.

Constructing the Antenna

A lawn-edger tool was used to make slits in the lawn, out from the base of the antenna toward the edges of the backyard. The slits were cut to a depth of 2 inches (51 mm). A total of 10 radials were buried in the slits. Some are only 15 feet (4.57 m) in length, while others are 25 feet (7.62 m) long. The metal yard fence was bonded together as needed, using wire jumpers between the fence sections. A single buried wire joined the fence to the common ground point at the base of the antenna.

The OM drove the steel pipe into the ground to a depth of 4 feet (1.22 m), leaving 3 feet (0.91 m) above ground for attaching the vertical antenna and weatherproof box. Construction details are shown in Fig. 3.

Although a wooden box could have been used to house the loading coil, switch and other hardware, an old electrical housing that the author's OM had in his junk box was used (Fig. 4.). It was drilled and

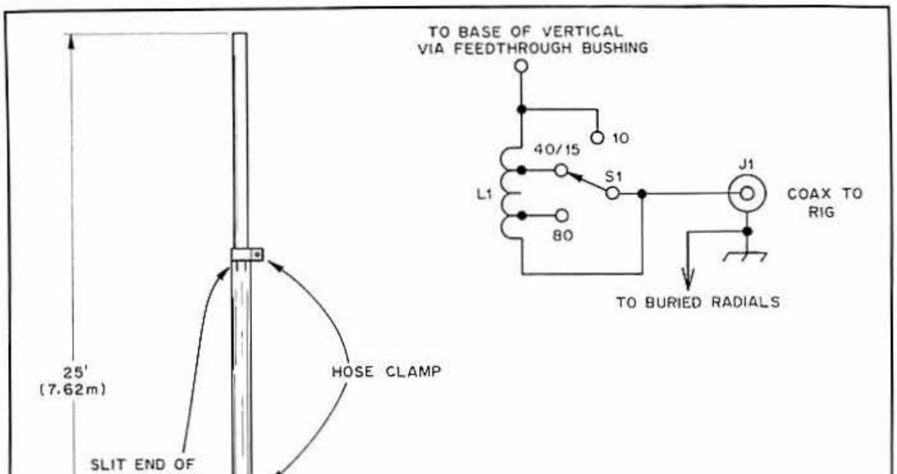
TO BASE OF VERTICAL
VIA FEEDTHROUGH BUSHING
10
40/15
S1
J1
L1
COAX TO RIG
80
TO BURIED RADIALS
25' (7.62m)
HOSE CLAMP
SLIT END OF TUBING WITH HACK SAW (4 CUTS)
PLATES,1/4 X12X6-INCH(6X304X154mm)
SMALL METAL PLATE AND ONE U BOLT FOR EACH INSULATOR
STEEL PIPE
U BOLT
STANDOFF INSULATOR
METAL OR WOODEN BOX (ATTACHED TO PIPE WITH U BOLTS)
JUMPER WIRE TO VERTICAL
UNDERGROUND COAX CABLE TO RIG
FEEDTHROUGH BUSHING
BAND SWITCH
WIRE TO BURIED RADIALS

Fig. 3 — Dimensional drawing of the four-band Novice antenna. The top ends of the two lower tubing sections are slit four times each by means of a hack saw. This permits a tight joint when the hose clamp is compressed. The vertical is attached to one metal plate with U bolts. The large standoff insulators with metal feet connect this plate to a,pair of small ones. The latter are attached to the steel support pipe by means of two U bolts. The metal box is also affixed to the steel pipe with two U bolts. The band switch is a single-pole three-position ceramic wafer type. L1 can be a 5-inch (127-mm) length of B & W 3029 Miniductor, 2-1/2 inches (64 mm) in diameter, 6 turns per inch of no. 12 wire. See text for alternative mounting methods.

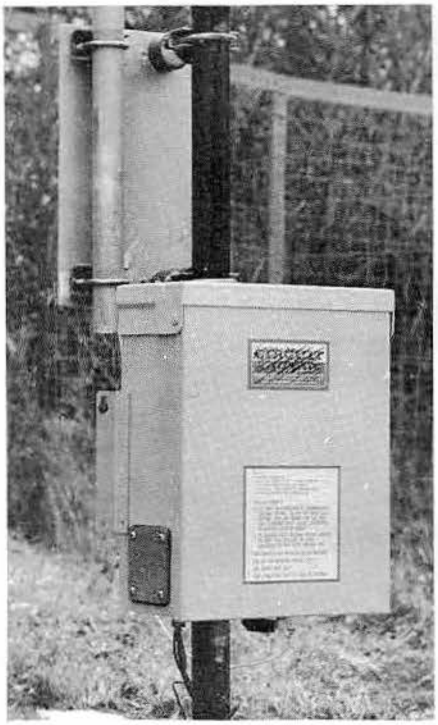

Fig. 4 — Closeup view of the base of the vertical. The aluminum tubing is affixed to a metal plate. The latter is attached to the iron support pipe by means of two surplus standoff insulators. Small aluminum plates are attached to the ends of the insulators to permit them to be fastened to the iron pipe by means of U bolts. The radial wires are connected to the bottom of the coil-housing box.

Fig. 5 — The author checks the SWR of the antenna during final adjustment of the system.

punched on the bottom surface to hold the feedthrough bushing, coaxial connector, switch and ground terminal for the radials.

W1FB designed the antenna, but he wasn't sure that an acceptable impedance match could be had on all four bands without a complex matching network. It was decided to try his idea, so the installation was completed.

Adjusting the Vertical

The author helped her OM install the antenna, then called W1FB for some assistance in tuning the system. He thought we could tune the vertical for 40 meters and make it work okay as a 3/4-wavelength vertical on 15 meters. For use on 80 meters it would be fairly short (63 feet or 19.20 m is the correct length for 3.7 MHz). With base loading it should offer adequate service out to a few hundred miles on 80 meters. Finally, it would operate as a 3/4-wavelength vertical on 10 meters.

A homemade SWR indicator was hooked in the coaxial line at the base of the vertical.[1] A small amount of transmitter power (5 W) was applied at 3725 kHz and the 80-meter switch lead was touched on the turns of the coil until minimum reflected power was indicated (Fig. 5). An SWR of 1:1 was obtained. The wire was then soldered in place on the coil. Next power was fed to the antenna on 7125 kHz, and the 40-meter switch lead was touched to the coil turns until an SWR of 1:1 was read. While using the same coil tap power was supplied to the antenna on 21.1 MHz and the SWR checked. It was approximately 3:1. By moving the coil tap just one turn we were able to get an SWR of 1.5:1 on 15 meters. A recheck on 40 meters followed. The SWR for that band was less than 2:1 — not a bad compromise! The coil was bypassed entirely for operation on 10 meters: an SWR of 2:1

---

[1] DeMaw, "A QRP Man's RF Power Meter," *QST*, June 1973.

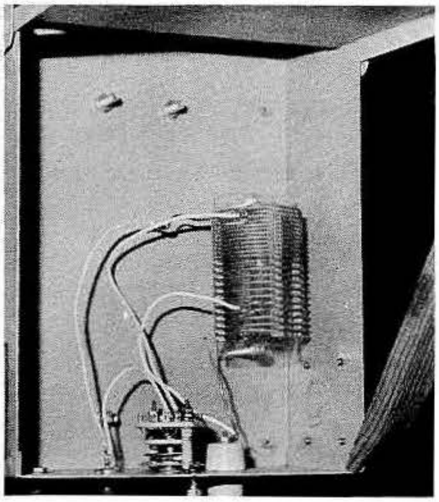

Fig. 6 — Interior view of the coil housing showing the switch, feedthrough bushing, coaxial connector, and ground post for the radials. The coil shown is a piece of Air Dux stock with a tapered pitch. It was obtained at a flea market.

was indicated at 28,000 kHz. The length of the overall antenna for operation on 28,100 kHz should be 25 feet or 7.62 m (3/4-wavelength radiator). However, the switch leads inside the coil housing add to the antenna length. If an SWR of less than 2:1 is desired, break the 10-meter switch lead and insert a 100-pF air variable capacitor. The unwanted reactance can be tuned out by this means and a low SWR will result.

Opening and closing the cover of the metal box had only a minor effect on the SWR. At last an on-the-air test of the system was possible. Fig. 6 shows the interior of the coil and switch housing.

Results

Good signal reports have been received on all bands. The first QSO on 40 meters netted an RST 599 report from North Carolina and many similar reports followed on 80, 15 and 10 meters. The author feels that a WAS award is not too far away now that this antenna is in operation.

An Alternative

There are many ways you can duplicate this design using substitute materials. For example, electrical conduit with couplers between the sections should be satisfactory in place of the aluminum tubing. The entire structure could be made from 2 x 4 inch (51 x 102-mm) lumber. If that is done, the radiator could even be a 25-foot (7.62 m) piece of no. 10 wire, supported on the side of the wood with standoff insulators.

Instead of the mounting method shown in Fig. 3, the vertical pipe could probably be inserted into a 2-foot (.61-m) length of PVC tubing, then clamped to the mounting plate. This would eliminate the need for the two standoff insulators. Better still, four or five wraps of Teflon sheeting (10 mil or 0.25 mm thickness) could be placed over the bottom end of the vertical before clamping it in place on the mounting plate. Teflon can be purchased at most plastic-supply houses.

The author hopes this idea is useful to other Novices who are trying to keep the budget within reasonable limits. This writer likes the way the antenna is working. Others should have good luck with this antenna also. Oh, by the way, the ground radials are made from various scraps of wire. The size isn't important, and they can be insulated or bare. There is quite an assortment of wire types buried in the lawn! This material was originally presented in *QST* by Marian Anderson, WB1FSB.

The Cheapie GP

Keeping an ear on 10 and 15 meters? With the trend of rising solar flux, 15 meters "opens" nearly every day — 10 meters occasionally. Don't miss these openings. Build a Cheapie GP!

Looking for an antenna to cover the 10- and 15-meter bands that won't pinch your pocketbook? Should it contain easy to obtain items and be relatively simple to construct while using ordinary hand tools? Must it be physically small and not a neighborhood eyesore? If your answer to these questions is "yup," you've turned to the right page! Detailed here is a duo-band coaxial-cable-fed, trap groundplane antenna system.

A groundplane antenna consists of a vertical quarter-wavelength radiator and an artificial metallic ground system extending radially from its base. For good results the antenna should be mounted at least one-quarter wavelength above ground. This is approximately 12 feet (3.66 m) on 15 meters. At this height and higher, only four radials for each band are necessary to provide a good ground system. If this antenna was ground mounted, many more radials would be needed to reduce earth losses to an acceptable low level. The antenna has an omnidirectional horizontal or *H*-plane pattern which means radiation from the antenna is essentially the same in all horizontal directions as shown in Fig. 7A. In Fig. 7B the approximate vertical or *E*-plane pattern of the antenna is shown for the antenna mounted one-quarter wavelength above ground. As can be seen, maximum radiation occurs at the low angles above the horizontal plane of the antenna. The antenna has an overhead null so little energy is radiated at high angles. This does not detract from the usefulness of the antenna, since DX signals seldom arrive at high angles on these bands. Instead, maximum radiation is concentrated at the *lower angles* where it will be more effective. Conversely, a horizontal dipole mounted one-quarter wavelength or less above ground will have maximum radiation straight up or nearly so. Lower-angle radiation will not occur unless the dipole is raised to a height of one-half wavelength or more above ground. If it is not possible to mount your horizontal antenna a half wavelength above ground, the vertically polarized groundplane is perhaps the logical choice over a dipole. As a general rule, the higher the groundplane can be located above power lines, gutters, house wiring and the like, the better it will perform.

A Duo-Band System

For the antenna to perform properly on two bands, it must function as a *separate* groundplane vertical on each band. This requirement is met quite simply in the case of the radials since four radials are used for each band. The 10-meter radials are 8 feet, 5 inches (2.56 m) long and the 15-meter radials are 11 feet, 7 inches (3.53 m). When the antenna is operated in the 10-meter mode, the 15-meter radials have little effect on the system. Similarly, when the antenna is used on 15 meters, the 10-meter radials do not appreciably alter the operation of the antenna. The effect of the additional radials is to change slightly the feed point resistance and the resonant frequencies of the antenna.

The vertical member of the antenna must act as a quarter-wavelength radiator on both 10 and 15 meters. To accomplish this a trap is inserted in the 15-meter vertical section. Its approximate location is a

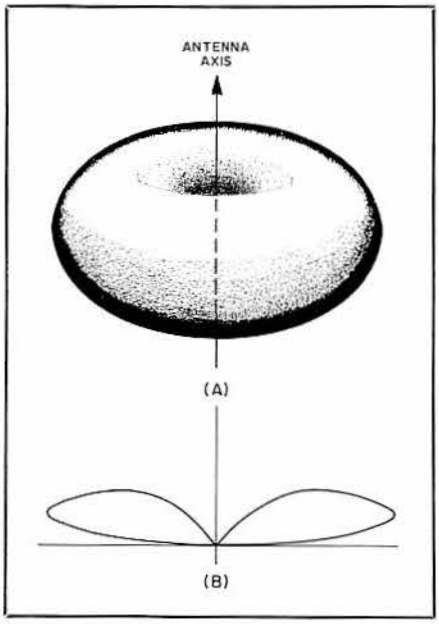

Fig. 7 — (A) Horizontal or H-plane pattern of the groundplane. (B) Approximate vertical or E-plane pattern of the antenna.

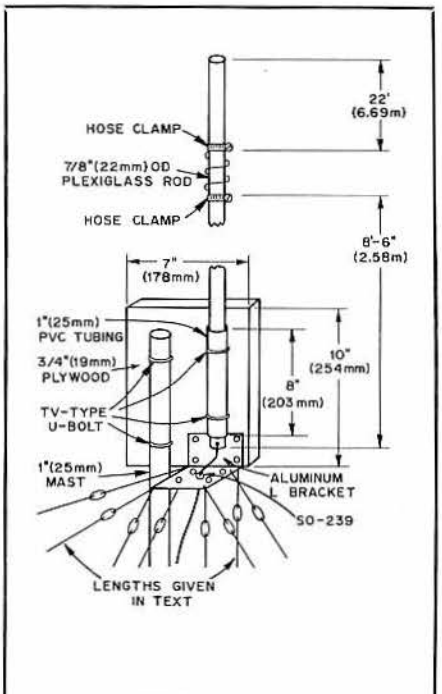

Fig. 8 — Constructional details of the duo-band antenna system.

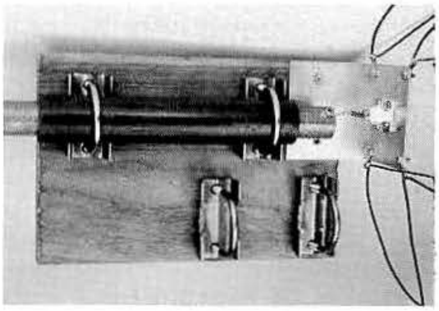

A photograph of the base assembly. The SO-239 coaxial connector and hood can be seen in the center of the aluminum L bracket. The U bolts are TV-type antenna hardware. The plywood should be coated with varnish or similar material.

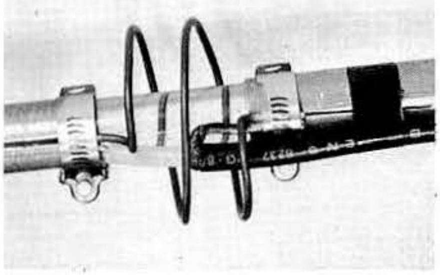

A close-up view of the trap. The leads from the coaxial-cable capacitor should be soldered directly to the pigtails of the coil. These connections should be coated with varnish after they have been secured under the hose clamps.

10-meter quarter wavelength above the base. The trap has a high impedance on 10 meters, thereby electrically divorcing the top section of the antenna when it is operated on that band. The length of the 10-meter section (section below the trap) is somewhat shorter than that of a simple 10-meter-only quarter-wavelength vertical. This is because the trap and top section of the duo-band system add top loading to the 10-meter portion of the antenna, reducing the length required for quarter-wavelength resonance. A coaxial-cable capacitor and a coil made from copper tubing form the trap which is resonant at approximately 28.150 MHz. On 15 meters the trap has a low impedance, effectively connecting together those portions of the antenna above and below the trap. The overall length of the 15-meter radiator is a bit shorter than a simple quarter-wavelength radiator would normally be for that band. This is because at 21 MHz the trap introduces inductance into the radiator — similar to the effects of a loading coil.

Construction

The vertical section of the antenna is mounted to a 3/4-inch (19-mm) thick piece of plywood board that measures 7 X 10 inches (178 mm X 254 mm). Several coats of exterior varnish or similar material will help protect the wood from inclement weather. Both the mast and the radiator are mounted to the piece of wood by means of TV U-bolt hardware. The vertical is electrically isolated from the wood with a piece of 1-inch (24-mm) diameter PVC tubing. A piece approximately 8 inches (203 mm) long is required, and it is of the schedule-80 variety. To prepare the tubing it must be slit along the entire length on one side. A hacksaw will work quite well. The PVC fits rather snugly on the aluminum tubing and will have to be "persuaded" on with the aid of a hammer. The mast is mounted directly to the wood with no insulation. An SO-239 coaxial connector and four solder lugs are mounted to an L-shaped bracket made from a piece of aluminum sheet. A short length of test probe wire, or inner conductor of RG-58/U cable, is soldered to the inner terminal of the connector. A UG-106/U connector hood is then slid over the wire and onto the coaxial connector. The hood and connector are bolted to the aluminum bracket. Two wood screws are used to secure the aluminum bracket to the plywood as shown in the drawing and photograph. The free end of the wire coming from the connector is soldered to a lug which is mounted to the bottom of the vertical radiator. Any space between the wire and where it passes through the hood is filled with GE silicone glue and seal or similar material to keep moisture out. The eight radials are soldered to the four lugs on the aluminum bracket. The two sections of the vertical member are separated by a piece of clear acrylic rod. Approximately 8 inches (203 mm) of 7/8-inch (22-mm) OD material is required. The aluminum tubing must be slit lengthwise for several inches so that the acrylic rod may be inserted. The two pieces of aluminum tubing are separated by 2-1/4 inches (57 mm).

The trap capacitor is made from RG-8/U coaxial cable and is 30.5 inches (775 mm) long. RG-8/U cable has 29.5 pF of capacitance per foot and RG-58/U has 28.5 pF per foot. RG-8/U cable is recommended over RG-58/U because of its higher breakdown-voltage characteristic. The braid should be pulled back 2-inches (51 mm) on one end of the cable, and the center conductor soldered to one end of the coil. Solder the braid to the other end of the coil. Compression-type hose clamps are placed over the capacitor/coil leads and put in position at the edges of the aluminum tubing. When tightened securely, the clamps serve a two-fold purpose — they keep the trap in contact with the vertical members and prevent the aluminum tubing from slipping off the acrylic rod. The coaxial-cable capacitor runs upward along the top section of the antenna. This is the side of the antenna to which the braid of the capacitor is connected. A cork or plastic cap should be placed in the very top of the antenna to keep moisture out.

Installation and Operation

The antenna may be mounted in position using a TV-type tripod, chimney, wall or vent mount. Alternatively, a telescoping mast or ordinary steel TV masting may be used, in which case the radials may be used as guys for the structure.

Any length of 50-ohm cable may be used to feed the antenna. The SWR at resonance should be on the order of 1.2 to 1.5:1 on both bands. The SWR is not 1 because the feedpoint resistance is something other than 50 ohms — closer to 35 or 40 ohms. Nearby metallic objects may also have an effect on the impedance of the antenna. The antenna is resonant at approximately 21.150 MHz and 28.150 MHz.

Some amateurs place too much importance on obtaining a "no reflected power" reading on their SWR indicators. Most ham transmitters will load into lines which exhibit an SWR of up to 2:1, the exception being some of the new broadband, no-tuning transmitters. To demonstrate how little effect an SWR of 1.5 or 2:1 will have on the system, consider the following: If the antenna is fed with 100 feet (30.48 m) of RG-8/U cable and if the antenna is perfectly matched to the line (an SWR of 1), the power loss along the cable will be approximately 0.98 dB on 10 meters. If the SWR on the line is increased to 1.5:1, the additional loss on the line due to the higher SWR is less than 0.1 dB. If the SWR is increased to 2:1, the additional loss over that of the matched condition will be on the order of 0.2 dB. This amount of loss is insignificant, and for this reason a complicated matching network is not necessary. This material was originally presented in *QST* by Jay Rusgrove, W1VD.

A High-Performance, 20-, 40- and 80-Meter Vertical System

Urban dwellers with small lots take heart! Here are the details for building your own DX-effective three-band mini-vertical antenna. Cost-conscious amateurs will be happy with the low cost of this system.

In another article on vertical antennas,[1] we tried to point out some fundamental characteristics of ground-mounted verticals, namely: (1) a good image plane is necessary for efficient operation, (2) a vertical over a good image plane compares favorably with a dipole at an elevation of one-half to one wavelength, (3) a short vertical compromises little in the way of performance.

This paper describes a highly efficient, three-band vertical system for 20, 40 and 80 meters, using elements of the order of an eighth wavelength. The system consists of an 80-meter vertical in parallel with a 20/40-meter trap vertical. Actually, either the 80-meter or the 20/40-meter vertical can be constructed and used alone if one is not interested in triband operation. The input impedances of both antennas are 12-1/2 ohms and they use the same 4:1 matching transformer.[2] The antennas also use the same radial system consisting of 100 radials of no. 15 aluminum wire 50 feet (15.24 m) in length (a lesser number of radials can be used as is discussed later in the article). Because of expected lower sunspot activity and, hence, poorer propagation conditions on the higher bands, the 40- and 80-meter portions of this antenna system, in particular, should prove very effective in DX communication over the next few years.

The first part of this paper deals with the design and tune-up considerations of the 80-meter element; the second part with the 20/40-meter element and the way it is used with the 80-meter vertical forming an efficient triband system. This is followed by reports on performance. Reference is also made to other alternatives for a multiband vertical system.

80-Meter Vertical

The two considerations in designing the 80-meter element of this vertical system were: (1) a good bandwidth for a reasonable height (a height one person can handle); and (2) proper spacing between the 80 and 20/40 portions such that coupling is negligible and both can be used over the same radial system.

Prior to building the shortened vertical described in this article, two others were constructed and tested on the air. One was a 22-foot (6.71 m) vertical[3] which had a 65-kHz bandwidth. The other was a trap vertical[4] with only a 20-kHz bandwidth. Both of these antennas were used separately over the same image plane of 100 radials and were efficient radiators. Therefore, the main reason in going to a slightly longer antenna was simply *to obtain a broader bandwidth.* The eighth-wavelength vertical on 40 meters from the previous work[5] promised considerable improvement by adding only a few feet. The results are shown in Figs. 9 and 10. The total height turned out to be 29 feet (8.83 m). This resulted in a bandwidth of about 140 kHz, a little more than twice the bandwidth of the 22-footer (6.71 m).

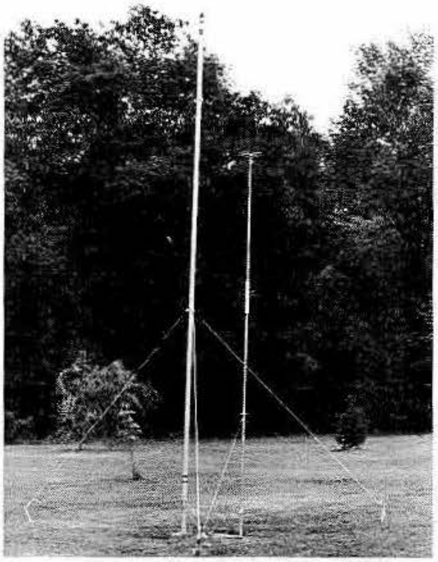
The triband vertical showing the polypropylene guys which provide an extra margin of support.

The vertical, in Fig. 9, uses a 20-foot (6.10-m) section of thick-wall aluminum tubing. (It was used some years ago as a gin pole for erecting a beam on a 40-foot (12.2 m) tower.) An 8-foot (2.44 m) extension was constructed with 1-inch (25-mm) tubing bolted in place using spacers for centering. The insulator at the bottom is phenolic tubing with a canvas base.[6] It has a 1/2-inch (13-mm) thick wall, is 9 inches

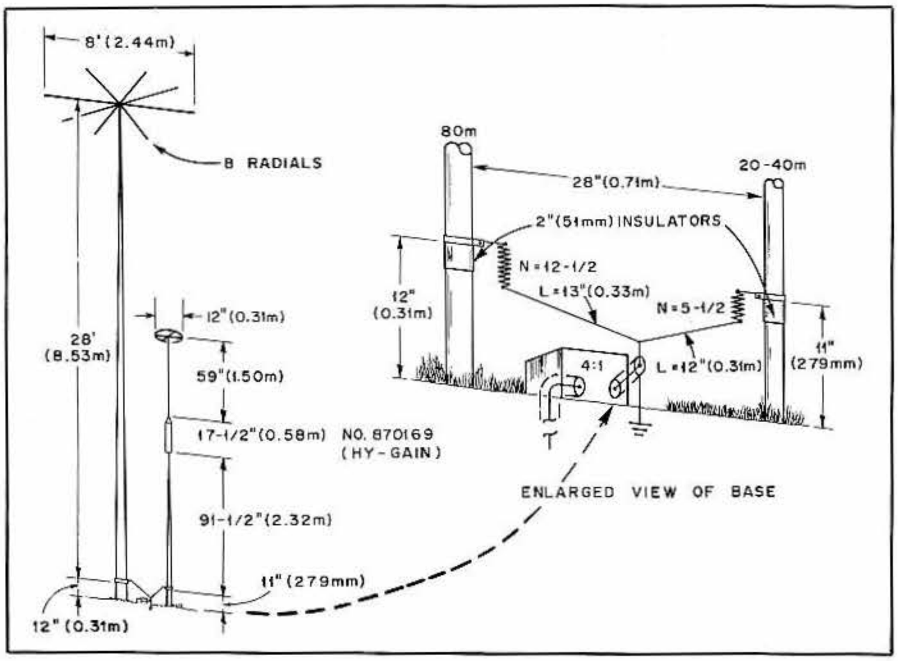

Fig. 9 — The 20-, 40- and 80-meter vertical antenna system. Tuning and construction details are given in the text.

Close-up of the base of the triband vertical antenna.

(229 mm) long and has an ID of 1 inch (25-mm). The bottom aluminum tubing supporting the antenna is 3-1/2 feet (1.06 m) long with 2-1/2 feet (0.76 m) of it placed in cement. The diameter of the hole in the ground is about 1 foot (0.31 m). Even though this construction could probably be self-supporting, three simple polypropylene guys at about the 7-foot (2.13-m) level are used for extra margin of support. The radials at the top use 1/2-inch (13-mm) aluminum tubing.[7] The base loading consists of 12-1/2 turns of a B&W 3029 coil.[8] Actually, 14 turns are on the coil. A shorting stub, as shown in the close-up picture of the base of the two antennas, is used for adjustment. To place the minimum SWR near the low end of the phone band, where much of the DX is worked, 1-1/2 turns were shorted out. The final number of turns employed depends, to some extent, on the number of radials used in the image plane. A simple check is to set the shorting tap at some convenient point, like 12 or 13 turns, and plot the SWR. If the minimum value appears too high in frequency, then add about a half a turn of coil. This half turn chould change the position of the minimum SWR value by about 50 kHz. As can be seen in Fig. 10, the minimum value of SWR is practically 1:1 and occurs at 3.840 MHz. This also verifies the input impedance value of 12-1/2 ohms, as expected from the previous work on short verticals.

20/40-Meter Vertical

In extending the operation of a vertical system over other bands, many alternatives are available. A trap vertical with a 12-1/2-ohm input impedance can be connected in parallel with the 80-meter vertical. Alternatively, the trap vertical can be designed to present an impedance of 50 ohms and thus be connected to the input side of the 4:1 transformer.[9] If broadband operation is desired on only 80 and 40 meters, then one-eighth wavelength verticals on both bands can be used.[10]

For this work, a 20/40-meter trapped vertical having an input impedance of 12-1/2 ohms was used. It offered a rather simple mechanical form of parallel operation as is shown in Fig. 9 and the close-up picture. The bandwidth on 40 meters of 155 kHz appeared acceptable. This bandwidth can be extended by about 50 percent by using the 50-ohm design of a trap vertical.

In the first attempt of parallel operation, the 20/40-meter vertical was placed only 14 inches (355 mm) away from the 80-meter vertical. The coupling appeared excessive. The 80-meter vertical was detuned by approximately 50 kHz. The 20/40-meter vertical also required excessive base loading in order to present an acceptable input impedance. By doubling the spacing between the verticals to 28 inches (711 mm), the interaction between them became negligible. The final values of heights and loadings were practically the same as if the elements were operating alone.

The adjustment of the 20/40-meter vertical is somewhat more complicated than the 80-meter vertical. An impedance bridge, as described in the *ARRL Handbook*, is of considerable help. In this case, there are two degrees of freedom: (1) varying the number of turns at the base, and (2) adjusting the lengths of the vertical sections.

Basically, the tuning is as follows: The tap is set at about 5-1/2 turns and the 20-meter section adjusted to give an acceptable value of SWR, both in position in the band and in magnitude. If the impedance is too high, it can be lowered by increasing the number of turns and lowering the height of the 20-meter section for resonance. After this, the 40-meter portion is then tuned. A plot of the SWR vs. frequency can immediately give an indication of the necessary adjustment of the section above the 20-meter trap. If the minimum value appears too low in frequency, a shortening of the top section is required. In no case should large adjustments be made. A change of a few

Table 1

Some Results With Triband Vertical

Date	*Station Contacted*	*W2FMI Signal Report*	*Freq.(MHz)*	*Input Power SSB (Peak)*	*Comments*
3/13	K5LWL/YV6	59 + 10 dB	7	2 kW	very, very strong signal, only one signal stronger — he used a Yagi at 120 feet (36.58 m)
4/3	WB5HJY	59 + 40 dB	7	2 kW	superior to anything on band
4/4	ZF1SP	59 +	14	2 kW	very, very nice signal
4/4	WA4MUR/4	59 + 40 dB	14	2 kW	best signal on whole band
4/18	KV4HW	59 + 10 dB	7	2 kW	loudest on band
5/9	K6YIY	59	4	2 kW	K1GZL and I are only ones they hear
5/9	W9LZX	59 + 30 dB	4	2 kW	strongest (very consistent signal)
5/12	W2DU	59 +	4	2 kW	tremendous signal
5/17	VK5PB	56-7	4	2 kW	remarkably strong
5/17	W4JNY	59 + 20 dB	4	2 kW	outstanding, certainly one of best
5/22	VK5PB	59 + 10 dB	7	2 kW	really amazing
5/23	ZL3RJ	57-8	4	2 kW	one of the strongest he's heard for some time — pinned the S meter
5/27	WA2BQL	59 + 30 dB	4	200 W	very potent — Stronger than most locals — couldn't believe you were using 200 W

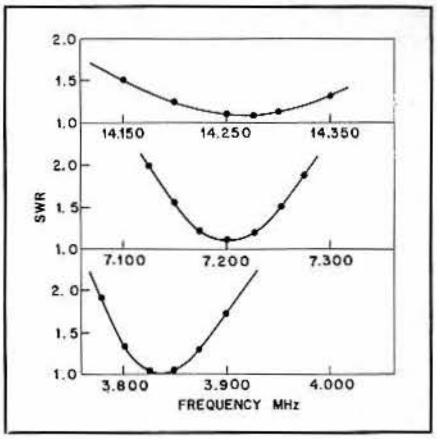

Fig. 10 — Standing wave ratio vs. frequency for the 20-, 40-, and 80-meter vertical antenna system.

inches has considerable effect. If the input impedance on 40 meters appears too low, then the 20-meter section has to be lengthened. This requires that the whole procedure be repeated. In any case, the initial adjustment should be started on the highest band of the trap vertical.

Results

Short verticals have been used by the author during the past year with considerable success. Many DX contacts were made on 40 meters with antennas varying in length from 6 to 33 feet (1.83 to 10.05 m). Since an extensive ground system was used, very little difference in effectiveness was noticed between the antennas.[11] This even includes redesigned trap verticals. As was stated before, the objective of the present investigation was to design a three-band vertical system which not only yielded competitive antennas on the lower bands, but one that was capable of covering a considerable portion of 80 meters. As was seen, a separate one-eighth wavelength antenna connected in parallel with a trap vertical not only gave a bandwidth on 80 meters of 140 kHz where the SWR was less than 2:1, but was short enough to not require considerable help in erection.

During all this time of operation of short verticals on 40 meters, and as of this writing, including several months on 80 meters, very few signal reports were received which did not indicate one of the best signals on the band. Table 1 gives some of the reports and comments received. In only three specific cases on 40 meters have the short verticals been bested by other antenna systems. One is shown in Table 1 where K5LWL/YV6 reported a stronger stateside signal by an amateur using a Yagi at 120 feet (36.58 m). The other cases include a comparison with W2GO of Linden, NJ. On 40 meters, VK5PB reported 6 dB and VK2WC 10 dB in favor of W2GO's signals. He was using a two-element Yagi at 60 feet (18.28 m). The elements were 44-1/2 feet (13.56 m) in length and the boom, 20 feet (6.10 m). On 80 meters, only one other station received a stronger report on direct comparison. This was by W2HCW, when comparing my signals with VK5PB. The difference was 2 to 3 S units. His antenna was an 80-meter Yagi at 120 feet (36.58 m).

Invariably, most amateurs were surprised by the performance of these verticals. In many instances, questions were asked regarding the minimum number and length of radials required for efficient operation of ground-mounted verticals. As was noted in previous articles,[12] the answer depends to some extent on the conductivity of the soil at the respective locations. Poorer soils not only require more radials, but ones that are also considerably longer. Although more experimental work is required in this area, it appears that about 50 radials, 0.2 wavelength long, should generally give good operation. The loss in this case will be approximately 1 to 2 ohms. Doubling the number to 100 radials should reduce the loss to less than 1 ohm. It should be noted that even 1 to 2 ohms of loss are appreciable with these short antennas since their radiation resistances are only 12-1/2 ohms.

Again, the author would like to acknowledge the help, encouragement, and interest shown by the many amateurs during our experimental studies on verticals. Very few antenna laboratories can boast of a greater number of willing and competent field stations. In particular, we would like to thank Al Jones, W2GO, for his considerable help in obtaining comparative reports in Australia and New Zealand. This material was originally presented in *QST* by J. Sevick, W2FMI.

References

[1] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. Also appears elsewhere in this chapter.

[2] Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, June 1972.

[3] See footnote 1.

[4] Commercial unit.

[5] See footnote 1.

[6] Cadillac Plastic and Chemical Co., P. O. Box 810, Detroit, MI 48232

[7] Construction details on the top hat are also given in the reference in footnote 1.

[8] 2-1/2-inch (64-mm) diameter, 6 tpi, no. 12 wire.

[9] The characterization and design of these trap verticals will be published later.

[10] See footnote 1 for details on the 40-meter vertical.

[11] See footnote 1.

[12] See footnote 1.

A Multiband Vertical Radiator

An efficient multiband radiator with no traps and no loading coils? Impossible; you can't build one without them and have it work! "Oh, yes, sure you can." I doubt it. "You wanna bet? Come on, I'll show you."

What would you say if the author told you this vertical antenna system was 89-percent efficient but had only six radials? (Efficiency is defined as the ratio of radiated power to total power delivered to the antenna.) No, it's no joke. The vertical antenna that is described here is designed to work against ground systems that are less than optimum.

By increasing the radiation resistance by about 200 ohms, we have been able to improve the effectiveness substantially of a vertical radiator with a poor ground system. The value of 200 ohms was selected as the feed-point radiation resistance because a toroid autotransformer with a 1:4 ratio will match 50-ohm coax line very nicely. An added bonus when using such an arrangement is the Faraday screen effect that reduces noise pickup from being transferred to the feed line. The noise reduction over a conventional network was determined to be about 6 dB. See Figs. 11 and 12.

The system is based on a 135-degree antenna height. The radiation resistance at the base of the structure is about 200 ohms,[1] ideal for our matching system. And at the same time the radiation pattern retains a reasonable form. The 10-meter groundplane on top of the structure acts as a top hat. On 7 MHz the phase loops cause the antenna to look like a fat radiator and between the two, top hat and phase loops, the electrical height of the antenna is raised to within the tuning range of the toriodal loading network. On 14 and 21 MHz the phase loops keep the instantaneous current on the structure in phase and eliminate high-angle lobes. When the system is switched to operate on 28 MHz, the groundplane at the top of the structure is fed through the coax line running up the lower portion of the structure.

Construction and Tuning

The photographs and the drawings show how the antenna is constructed. The system was designed so that "storebought" items could be used for the assembly, rather than fabrication of the antenna components. A list of required parts is given in Table 2.

The tuning adjustments of the system are very noncritical. Tuning can be done with an SWR indicator. The best sequence is to start with the highest frequency and work down. The 28-MHz groundplane used as a top hat can be adjusted before it is attached to the rest of the antenna. The Radio Shack 11-meter groundplane that we used comes with a hairpin or "beta" match. With a short length of 50-ohm coax for the feed line, the groundplane should be adjusted for the lowest SWR in the 10-meter band. After the groundplane has been adjusted and attached to the top of the antenna, the entire system can be raised into position for final tune-up on the lower frequencies.

At this point the SWR indicator should be inserted in the 50-ohm feed line from the transmitter, as close to the antenna matching network at the base of the structure as possible. All that remains is the adjustment of the 200-pF capacitor. The

This multiband radiator, just over 36 feet (10.97 m) tall, operates on five amateur bands, 80 through 10 meters.

Table 2

Parts Required for the Multiband Vertical Radiator

3 —10-ft (3.05 m) lengths TV mast; Radio Shack 15-5066.
1 —27-MHz ground plane; Radio Shack 21-901.
3 —100-ft (30.48 m) rolls of 22-gauge hook-up wire; Radio Shack 278-1295.
1 —Amidon Balun kit (see Fig. 12).
1 —Aluminum box for tuner.
1 —28-ft (8.53 m) length of RG-8/U coax line.
4 —8-ft (2.44 m) sections of aluminum tubing (see Fig. 11).
1 —Base insulator; soft drink or (?) bottle.
1 —200-pF capacitor, air variable, surplus.
Misc. — See text and figures.

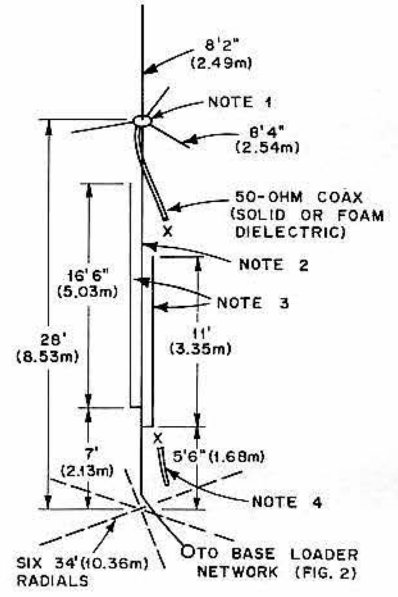

Notes

1) Radio Shack 27-MHz groundplane cut to dimensions shown. Radials are at right angle and connected to mast.
2) Three heavy-duty 10-ft (3.05-m) sections TV mast.
3) Phase loops, sections of telescoping 5/8-(16-mm) and 1/2-inch (13-mm) aluminum tubing spaced 4-1/2 inches (114 mm) from mast. See text.
4) 50-ohm coax such as RG-8/U. At upper end connect center conductor to top element, braid to ground-plane elements. Connect lower end to base loader network (Fig 12).

Fig. 11 — Details of the multiband radiating structure. See Table 2 for parts list. It is very important that a uniform spacing be maintained for the phasing loops, which may be supported at the bottom by shelf brackets. To support the upper portions of these loops, the author used homemade wooden spacers, 1/2 X 1 inch (13 X 25 mm), which were soaked in paraffin and held in place by screws passing through the loop tubing and the main radiator tubing. Homemade Plexiglas spacers may also be used.

---

[1] *The ARRL Antenna Book*, 13th Ed. See Fig. 2-74 and the accompanying text on page 60.

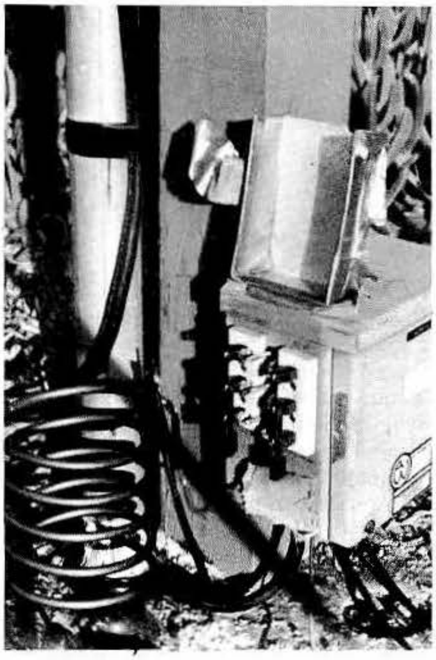

The base of the multiband antenna system. The 200-pF capacitor of the matching network is mounted inside a weatherproofed box, and the 80-meter loading coil is mounted outside. The porcelain knife switches for bandswitching are mounted outside the aluminum box under a homemade hinged cover. The structure itself is supported by a 10-foot (3.05-m) wooden post set 2 feet (0.61 m) in the ground (concrete optional).

Although heavy-duty insulators were used to support the radiating mast from the post as shown here, experience has indicated that insulators are not required. The author suggests that a modified pair of chimney-mount brackets be fastened to the post with lag screws. A short length of PVC pipe or Mylar film may be placed around the mast where it is supported, if you feel it necessary to have the mast insulated from the post.

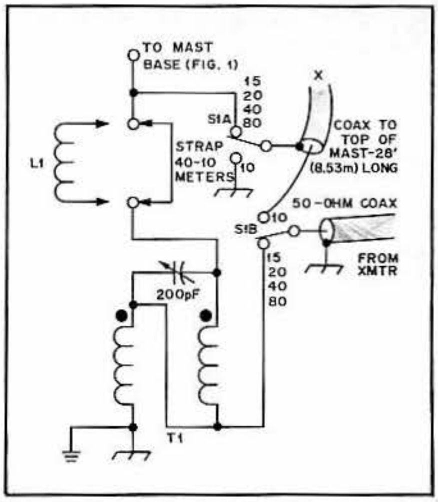

Fig. 12 — Schematic diagram of the base loading network. The connection to the mast base should be made with a short wire, no longer than one foot (0.31-m). The 200-pF capacitor should have a minimum of 1/16-inch (1.6-mm) plate spacing.

L1 — 80-meter loading coil, 0.67 μH; 4 turns of aluminum grounding wire, 3-inch (76-mm) diameter, 1 turn per inch, air core.

S1 — Dpdt, heavy-duty type. Two porcelain knife switches from the local hardware store will suffice, one for each pole.

T1 — Bifilar-wound toroidal autotransformer, 1:4 ratio; 11 turns spaced around 2-inch (51-mm) ID toroidal core: Amidon Balun Kit or equiv. (Amidon Associated, 12033 Otsego St., North Hollywood, CA 91607).

tuning will be different for each band. On 21 MHz the capacitor will be only about five percent meshed, on 14 MHz about half, and on 7 MHz nearly fully meshed. The capacitor on 3.5 MHz, along with the 3.5-MHz coil, can be adjusted to whatever section of the band you wish to work. The bandwidth will be about 400 kHz with less than 2:1 VSWR. To switch bands, all that is required is to set the double-pole, double-throw switch for the proper band and adjust the 200-pF variable capacitor.

All of the antennas constructed so far were cut to the dimensions shown in Fig. 11 and have not required any change. They have been installed in a wide variety of locations and the system design has been broad enough in tolerance to handle all of them. The writer's antenna has been in operation since 1972 and has performed exceptionally well. In the first year on the air, I managed to work 100 countries with less than 250 watts. In several experiments with just the six radials, we made comparative tests between this design and a properly matched quarter-wave vertical over the same radial system. The radiation efficiency was about 60 percent for the quarter wave vs. 89 percent for the "little monster." In the far field the signal improvement was better than 6 dB.

Now there are eight of these antennas on the air, and several more are under construction. All of the hams who have built the antenna have had comparable results to mine.

There are three hams to whom this author would like to give credit. They helped in testing and measuring the system and encouraged me to get it down on paper. They are Bob Crawford, WA6RYZ; Frank Scott, W6WOP and John Campbell, W6NVV. This material was presented originally in *QST* by A. E. Collins, K6VV.

References

The ARRL Antenna Book, ARRL, Newington, CT.
King, *The Theory of Linear Antennas*, Harvard University Press, Cambridge, MA.
Terman, *Radio Engineering*, McGraw-Hill Book Co., New York, NY.

A Triband Vertical Antenna

Keep your vertical antenna simple. Operate three bands without using traps!

The most common vertical is the quarter-wave type, which radiates well and is easily matched to a coax feed line, but has the weakness that it is a single-band antenna. Methods of adapting a vertical for multiband operation include the use of traps, loading coils and elements of different length fed in parallel. The antenna system to be described here was arrived at by treating separately the problems of designing the antenna and of designing the matching system.

The antenna design was easy. A height of 22 feet (6.71 m) is optimum for a vertical antenna for 10, 15 and 20, and several commercial verticals are of this height. It corresponds to about five-eighths wavelength on 10 meters. Greater height results in reduced low-angle radiation on 10, besides being more difficult to erect. A 22-foot (6.71-m) vertical gives better low-angle radiation than a quarter-wave vertical on all three bands and is especially good on 10 meters.

The antenna consists of 22 feet (6.71 m) of 300-ohm twin-lead, with both conductors tied together, held about an inch from a bamboo pole by means of small blocks of insulating material spaced every two feet or so along the pole and held in place with plastic electrical tape. The pole is guyed with nylon fishing line, and the antenna is operated against a groundplane of four quarter-wave radials for each band. The use of twin-lead rather than a single wire simulates a thicker conductor and reduces the impedance, making the antenna easier to match. Even so, the antenna impedances are vastly different on the three bands. The presumed values of impedance are listed in Table 3. These values are taken from published data[1] and assume a cylindrical antenna of specific radius operated against a perfect ground. The antenna is not cylindrical, and it operates against a wire groundplane, so that the impedances listed must be considered approximate only. The results obtained indicate that they are at least fairly close.

Table 3

Approximate impedances at base of 22-foot (6.71-m) ground-plane vertical antenna

Frequency	*Resistance*	*Reactance*
14.250 MHz	100 ohms	+200 ohms
21.375 MHz	1200 ohms	—500 ohms
28.500 MHz	60 ohms	—220 ohms

Table 4

Impedances at input end of 27.8-foot (8.47 -m) matching section of 300-ohm twin-lead terminated by the impedances listed in Table 3.

Frequency	*Resistance*	*Reactance*
14.250 MHz	93 ohms	+177 ohms
21.375 MHz	63 ohms	0 ohms
28.500 MHz	75 ohms	—282 ohms

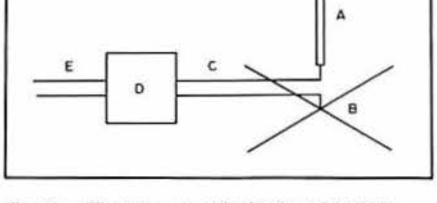

Fig. 13 — The triband vertical antenna system.

A — Vertical radiator, 22 ft (6.71 m) of 300-ohm twin-lead with conductors in parallel.
B — Four radials, each consisting of 1/4λ of wire, for each band (12 wires in all).
C — Matching section, 27.8 ft (8.47 m) of 300-ohm twin-lead.
D — Reactance unit; see Fig. 14.
E — 75-ohm coax to transmitter, any length.

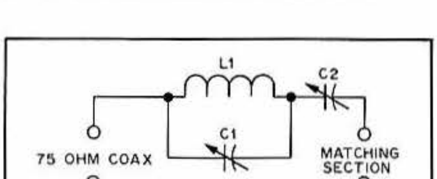

Fig. 14 — Reactance-unit circuit.

L1 — 0.86 μH; 7 turns no. 16 (or heavier). 1-inch (25-mm) dia, 1-inch (25-mm) winding length.
C1 — 30-pF air trimmer set to 23.1 pF.
C2 — 100-pF air trimmer set to 41.7 pF.

If the three antenna matching impedances are normalized to 300 ohms and plotted on a Smith chart, it can be seen that it is possible to obtain a reasonably good match to a coax line on any of the three bands by using an appropriate length of 300-ohm line as a matching section connected between the antenna and the coax line. What is not possible is to find a single length of line that will provide suitable matching on all three bands.

It turns out, however, that a 27.8-foot (8.47 m) section of 300-ohm line having the 0.82 velocity factor which is usual for twin-lead has interesting characteristics. Such a section of line, connected to an antenna having the impedances listed in Table 3, has the input impedances listed in Table 4. This combination of antenna and matching section is shown in Fig. 13.

It will be noted, first of all, that this combination may be used as it stands on 15 meters and fed with any usual coax. What is even more interesting, however, is that the input resistance is near 75 ohms on all three bands, though on 10 and 20 there is also considerable reactance. If a reactance unit could be made which had just the correct value of reactance to cancel out the input reactance of the matching section on all three bands, the unit could be connected in series with the matching-section input and the system fed with 75-ohm coax on all three bands.

Reactance Units

Can such a unit be made? Certainly! It's not hard at all. It consists of a simple seven-turn coil and two trimmer capacitors and is shown in Fig. 14. The trimmers are most easily adjusted with the aid of a grid-dip meter before the reactance unit is installed. C1 should be set so that L1C1 resonates at 35.85 MHz, and C2 can be subsequently adjusted by temporarily connecting it in parallel with L1C1 and varying it until the combination resonates at 21.37 MHz.

With the reactance unit adjusted as described, the antenna when first assembled showed SWRs of 2.1, 1.4, and 1.5 on 10, 15, and 20 meters respectively. Connecting the unit into the feed system apparently increased its shunt capacitance somewhat, for a small reduction in the value of C1 changed the SWR readings to 1.3, 1.5, and 1.4. These values could probably have been improved still more by further adjustment of C1 and C2, but it was not considered worth the trouble. If it is undertaken it should be borne in mind that the adjustments interact, but that C1 has its greatest effect on 10 meters and C2 on 20 meters. In any case it is not possible to obtain a perfect 1:1 match with this antenna system.

An SWR of 2, however, which can easily be bettered on all three bands, is within the loading capability of almost any transmitter and causes an *additional* loss over a perfectly matched line of less than half a dB even when the transmission line is very long. Since such a change in signal strength is undetectable, an SWR of 2 is in practice as good as an SWR of 1, as far as losses are concerned.

---

[1] Jordan, E. C., *Electromagnetic Waves, and Radiating Systems*, Figs. 13-12 and -13, New York: Prentice-Hall, Inc., 1950, pp. 482-3.

A Few Notes

A few observations are in order. The matching section has a fairly high SWR (between 4:1 and 8:1), and it should be spaced well clear of metal objects and should not have sharp bends. TV standoffs are useful here. It is important that the same side of the matching section be connected to the coax shield and to the groundplane. If such a thing were available, 300-ohm coax would probably be preferred for the matching section, but for medium power (180 watts PEP) TV-type twin-lead for the matching section and receiving-type trimmers in the reactance unit have proved satisfactory. The parallel circuit L1C1 does not operate at its resonant frequency, and thus high circulating currents are not encountered. High power might necessitate heavier components. The reactance unit needs protection from the weather and can be built into a small plastic box and sealed with plastic electrical tape.

Although the antenna is matched at three widely separated frequencies, it is not matched at intermediate frequencies and is thus not a broad-band antenna in the usual sense. Nevertheless, it is broad enough for normal ham use. The SWR remains below 2:1 over the entire 15- and 20-meter bands, and also between 28.3 and 29.4 MHz on 10.

All SWRs mentioned in this article have been measured at the transmitter and are thus probably a little optimistic. Transmission-line losses are not known, but the line is not long (about 40 ft. [12.19 m]), and it is doubtful that the SWRs at the reactance unit are much higher than those measured at the transmitter.

This antenna is no match for a quad or a Yagi, but it has provided numerous transatlantic QSOs on all three bands. Because of its effectiveness, convenience, and the cheapness and availability of everything used in its construction, it should appeal to many hams. This material was originally presented in *QST* by Frank A. Regier, OD5CG.

A Multiband Groundplane Vertical Antenna with Tuned Feeders

Who says tuned feeders are used only with horizontal wire antennas? Go multiband with a simple vertical radiator by using a tuned-feeder system — and do it with ease!

For several years, the author has experimented with several versions of this antenna and has found them to be most satisfactory. The antenna shown in Fig. 15 provides not only outstanding performance on 10, 15 and 20 meters, but performance on 40 and 80 meters equivalent to most mobile-antenna installations. Full efficiency on these latter bands can be realized by making the vertical and radial portions proportionately longer.

The antenna system consists of the radiating element, groundplane radials, an open-wire feeder of any convenient length, a Transmatch capable of either series or parallel tuning, and an antenna SWR bridge. This combination is shown in Fig. 16.

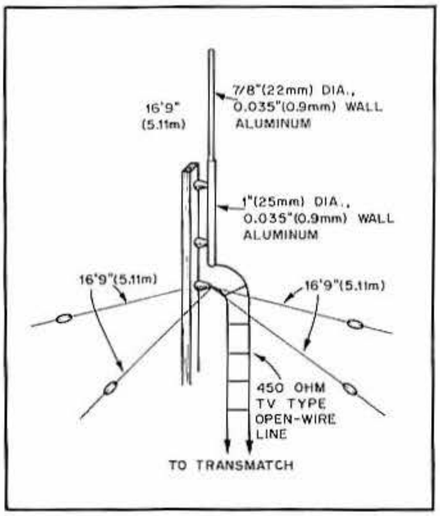

Fig. 15 — Sketch of W4VON's simple multiband antenna.

Antenna and Radial Lengths

The author preferred to cut the antenna for resonance on one particular ham band (20 meters), but it is not necessary that the antenna be resonant on any band. Efficiency will suffer, of course, on bands where the length is significantly shorter than one-quarter wavelength. Whatever the length chosen for the antenna, the radials should be of about the same length.

Construction

Constructional details are shown in the sketch of Fig. 15. In the event the specified aluminum tubing is not available, thin-walled galvanized electrical conduit, aluminum conduit, or copper pipe may be used.

It is recommended that the vertical element not be supported by drilling holes through the tubing, as this will substantially weaken the structure. An alternative mounting arrangement is shown in Fig. 17. The insulators are porcelain standoff types with a lag-screw insert. They will be found at electrical supply houses (also at Sears), and are sometimes referred to as "saddle" type insulators. The hose clamps are stainless-steel gear type.

Adjustment

Operation of the antenna is simple. The Transmatch (a suitable one, including SWR bridge, is described in the ARRL *Handbooks)* and SWR bridge are connected, and the tuning network adjusted for

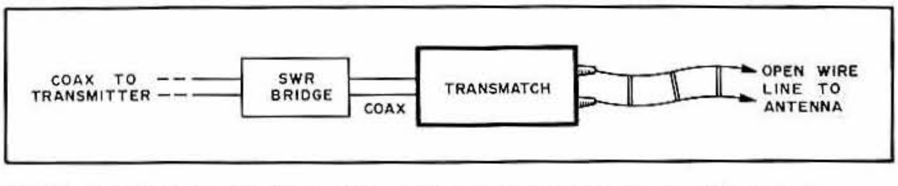

Fig. 16 — Recommended setup for coupling antenna line to low-impedance transmitter output.

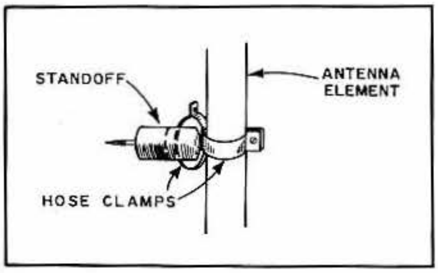

Fig. 17 — Suggested method of mounting radiator element without drilling holes in aluminum tubing.

maximum forward and minimum reflected power. Initially, it will be necessary to determine experimentally whether series or parallel tuning is required for the particular combination of band, feeder length, and antenna length selected.

The tuned-line-fed groundplane vertical antenna gives excellent performance, can be easily constructed in just a few hours even by a beginner, requires a minimum of installation space and costs less than $5, excluding the Transmatch and SWR bridge. The tuned circuit of the Transmatch provides excellent discrimination against harmonic output from the transmitter. Ease of construction and portability make this antenna an ideal one for Field Day use.

In case one wonders about the mismatch between the line and the antenna, the secret is in the use of open-wire line. The loss in such a line with an SWR of 25:1, at 10 meters, is less than the loss in RG-58/U when the latter is matched. This material was originally presented in *QST* by Arthur S. Gillespie, Jr., K4TP.

A Helically Wound Vertical Antenna for the 75-Meter Band

"Around and around she goes, and where she comes out — nobody knows." True, perhaps, with a wheel of fortune, but with a helically wound vertical the rf comes out effectively for DX and local work!

In order to have a full-size vertical antenna for the 75-meter band, the antenna would have to be around 65 feet (19.81 m) tall. That, alone, would require guying, or some other means of support which would result in a structure with no aesthetic appeal to any nonhams (who comprise almost 100 percent of the author's neighborhood). By using a helical configuration, the overall antenna height is reduced, thus eliminating the need for guying. With the helically wound antenna, more equal current and voltage distribution is accomplished (as compared with a lumped impedance from a loading coil). As a result, a better radiation pattern is produced. With this system very little reactance has to be cancelled out, eliminating the need for an overly large "top hat."[1] This antenna will easily take the full legal power limit.

Construction details for the antenna are shown in Fig. 19. The following is a list of parts used in constructing the antenna system.

1 — 20-ft (6.10-m) section of 4-inch (102-mm) OD plastic pipe (obtained at a local plumber)

130 ft. (39.62 m) no. 12 plastic-insulated

The dark pole to the immediate left of the tower is the WA0WHE helically wound vertical antenna.

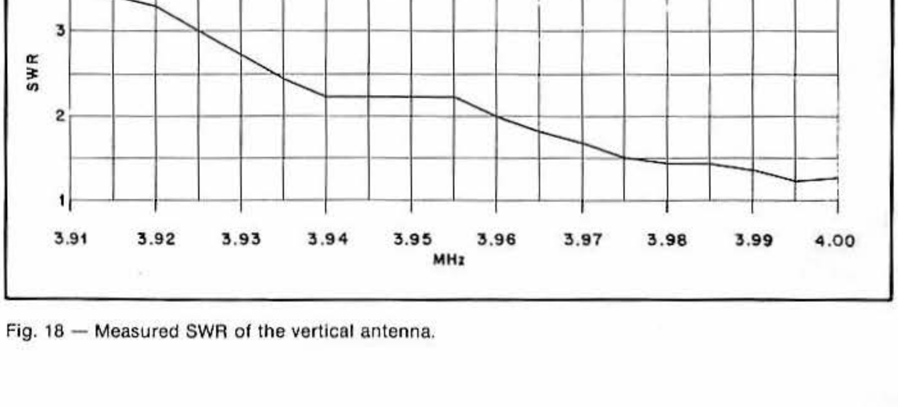

Fig. 18 — Measured SWR of the vertical antenna.

---

[1] [Editor's Note: Some form of metal top hat should be connected to the last turn of the driven element at the high-impedance end of the radiator. Extremely high levels of rf voltage can develop at the end of the helix, sometimes causing the tip of the antenna to burn. The top hat tends to lower the Q of the antenna, thus reducing the voltage level at the far end. An aluminum pie tin mounted on a ceramic cone insulator works well in this application.]

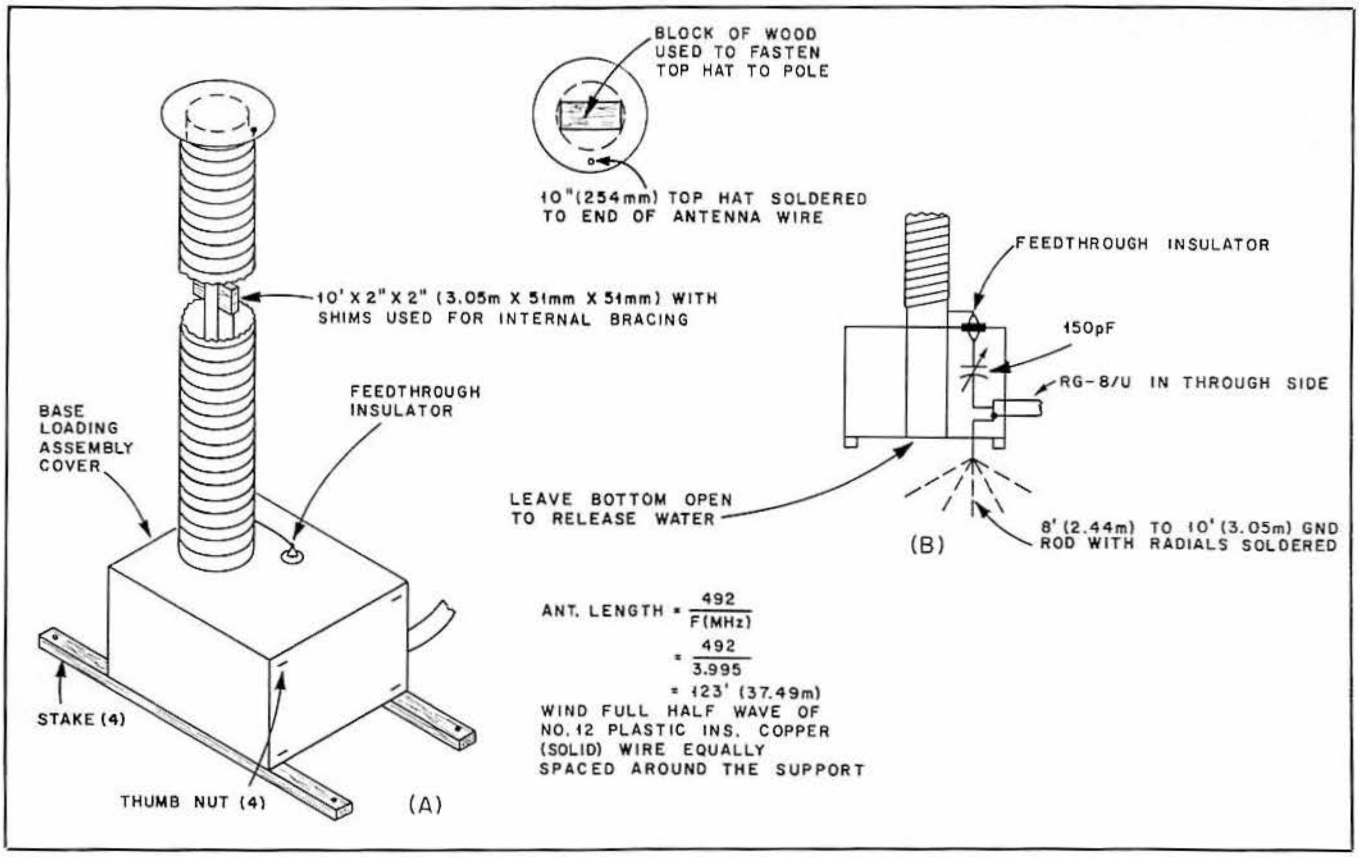
BLOCK OF WOOD USED TO FASTEN TOP HAT TO POLE
10"(254mm) TOP HAT SOLDERED TO END OF ANTENNA WIRE
10' X 2" X 2" (3.05m X 51mm X 51mm) WITH SHIMS USED FOR INTERNAL BRACING
BASE LOADING ASSEMBLY COVER
FEEDTHROUGH INSULATOR
STAKE (4)
THUMB NUT (4)
(A)
LEAVE BOTTOM OPEN TO RELEASE WATER
ANT. LENGTH = 492/F(MHz) = 492/3.995 = 123' (37.49m)
WIND FULL HALF WAVE OF NO. 12 PLASTIC INS. COPPER (SOLID) WIRE EQUALLY SPACED AROUND THE SUPPORT
FEEDTHROUGH INSULATOR
150pF
RG-8/U IN THROUGH SIDE
8' (2.44m) TO 10' (3.05m) GND ROD WITH RADIALS SOLDERED
(B)

Fig. 19 — Construction details of the helically wound antenna.

solid copper wire (allow a little extra, about 2 ft. [0.61 m])

1 — 10-in. (254-mm) disk of sheet metal (top hat; stop at a sheet metal or furnace repairman — I got this free)

260 ft. (79.25 m) no. 14 solid bare copper wire (I used only four radials in this system; use as many as you have room or money for)

20 sq. ft. (1.86 sq. m) of marine plywood for base

2 — 2 × 2 10-ft. (3.05 m) sections, for internal bracing

1 — 8-ft (2.44-m) copper ground rod (for radial junction)

The cost can be reduced considerably if all materials are readily located or already owned.

Fig. 18 shows the results of SWR measurements. The SWR indicator used to make the measurements was a Midland model. The antenna was adjusted for minimum SWR at resonant frequency with the tuning capacitor at the base. The frequency for this antenna is 3995 kHz with 123 feet (37.49 m) of wire wound on the support and one 8-ft. (2.44-m) ground rod and four 65-ft. (19.81-m) radials. There are numerous configurations possible with this system including a broadcasting favorite such as phasing two or more structures. I forgot to mention . . . give the entire structure one or two coats of marine varnish to seal the turns. This material was originally presented in *QST* by Gary L. Ellingson, WA0WHE.

The Ground-Image Vertical Antenna

The name of the game with vertical antennas is the "ground system." Not just any hastily contrived radial ground will do the job right. Read these details for building an effective ground that provides positive results.

Even though a beam antenna supported at the modest height of 40 feet (12.19 m) is a compromise (a 20-meter beam should be higher to be really effective), it still can be an obstacle to good neighborly relations, at least if your neighbors don't appreciate the ecological beauty of such an installation. Moreover, even that modest sort of an installation presents quite a number of engineering problems.

This report presents the results of the first phase of an investigation to find a less conspicuous but equally effective antenna for use at a new location. The author hopes it will provide suggestions for those faced with a similar problem. At the least, it may be of some value to those with a general interest in the subject of antennas.

This first part of this section deals briefly with theoretical considerations, the second with experimental results on quarter-wavelength and five-eighths-wavelength verticals, and the third on the test equipment. It should be pointed out at the outset that the information presented here is the work of a hobby and as such cannot be exhaustive. It is hoped that others will repeat some of the experiments, extend the work, and report the effect in practice of a groundplane system above 3 MHz. To my knowledge such practical data have not been reported.

Theoretical Considerations

A beam antenna possesses the advantage of gain and directivity. Nevertheless, its DX capability is determined primarily by its vertical radiation pattern. A large portion of the radiated energy should be directed between 5 and 25 degrees from the horizon.[1,2] Horizontally polarized antennas yield lower angles of radiation with increase in height above the ground. This is a result of the interference pattern created by reradiation from the earth's surface. Since the earth is a somewhat conducting medium, the electric field tangential to the surface must be approximately zero. This "boundary" condition is brought about by the induced surface currents which create an electric field of opposite phase. This field then combines constructively and destructively with the initial radiation from the antenna. A model for this condition is an image antenna of opposite phase below the earth's surface as a depth equivalent to the height above the surface.[3] In order to get a lobe below 15 degrees, the antenna height must exceed a wavelength. This is greater than 60 feet (18.29 m) on 20 meters.

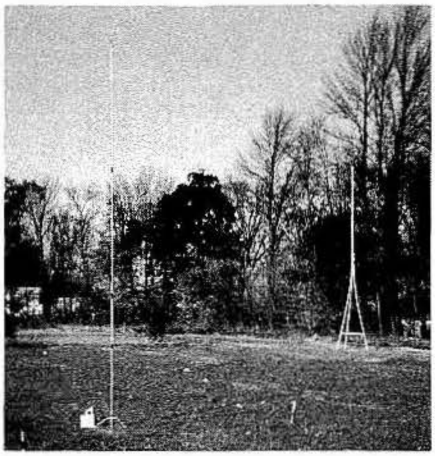

The 20-meter vertical antenna in the foreground, and the 26-foot (7.93-m) test tower.

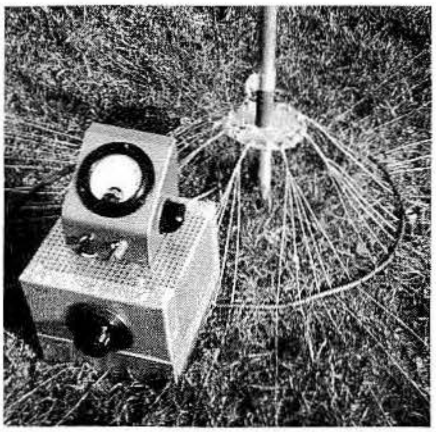

The base of the vertical element of the antenna and the impedance bridge. Forty radials, in bundles of five are fastened to the aluminum base plate and are tied down with a ring made from copper tubing.

On the other hand, a vertical antenna, in the ideal case, possesses an image which is in phase to produce a lobe tangential to the earth's surface. Only when the antenna length is increased to a wavelength, or multiple thereof, does the tangential lobe disappear. This is true whether the antenna is on the ground or suspended in space. Therefore, a vertical beam, i.e., an array of vertical antennas coupled together with an appropriate feed system, on the surface of the earth, seemed to be the logical choice for the new installation since it would appear to meet the following objectives: The system should

1) Exhibit a minimum profile.
2) Be easy to install and tune.
3) Offer a low angle of radiation.
4) Not require a large outlay of money.

The author proceeded, therefore, to construct an array with four vertical elements. When the tests were begun, it immediately became apparent that the simple procedures the writer was using were inadequate to cope with such a complex system. The author had to start anew to develop a test procedure and to build some suitable test equipment. The logical step was to backtrack to a single vertical antenna and use it as a standard on which to develop some basic test standards. As discussed subsequently, the writer found that relatively simple equipment, e.g., a simple impedance bridge, a field-strength meter, and a test oscillator, gave this author all of the data about the system that he needed.

It is important to note that a true ground-image antenna differs substantially from a ground-plane antenna system which relies merely on a few λ/4 radials above the ground. A true ground-image system results when a sufficient number of radials are used and an image of only the vertical section is sufficient to describe it. Considerable information[4,5] is available on ground-image systems for verticals operating below 3 MHz. The results have shown that some

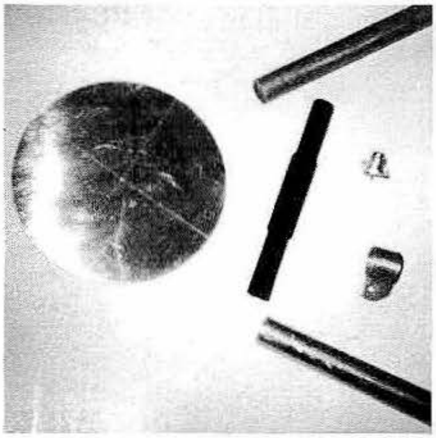

Base hardware for the antennas being tested. The insulator was made from 1-inch (25-mm) maple dowel, turned down to accept the 7/8-inch (22-mm) inside diameter of the ground and antenna tubing.

Table 5

Comparison of responses of 1/4-wavelength and 5/8-wavelength vertical antennas at low radiation angles. Data were taken by field-strength meter mounted on a wooden tower at a distance of 7.5 wavelengths at 14.25 MHz. Field strength (E) is normalized to maximum value obtained with 5/8-wavelength case.

θ	$E(\lambda/4)$	$E(5/8\lambda)$	GAIN OF 5/8-λ ANTENNA
0.1°	0	0.58	∞
0.4°	0	0.62	∞
0.75°	0.62	0.80	3.4 dB
1.1°	0.69	1.0	3.2 dB
1.5°	0.62	0.92	3.4 dB
2.25°	0.48	0.80	4.3 dB
3°	0.41	0.69	4.5 dB

The triband trap vertical antenna, which has an overall height of only 12.5 feet (3.81 m). With the system of 40 radials, performance of this antenna on 20 meters was about the same as that of the 1/4 and 5/8-wavelength vertical elements.

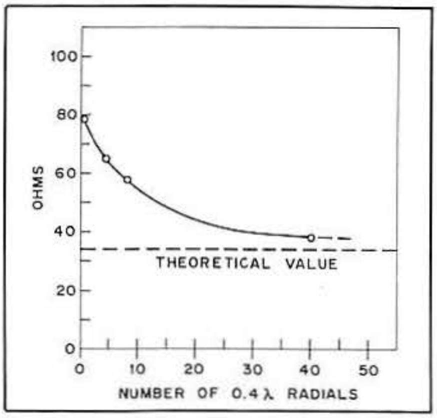

Fig. 20 — The input impedance of a 20-meter quarter-wave vertical antenna as a function of the number of radials 0.4-wavelength long. The four and eight radials consisted of bundles of five wires of no. 18 guage. The 40-radial point was obtained by fanning out the eight bundles.

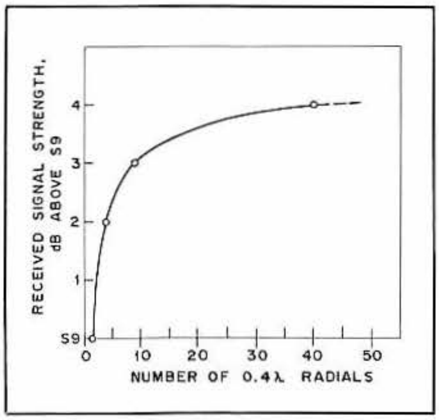

Fig. 21 — The improvement of low-angle radiation of a quarter-wavelength vertical antenna on 20 meters as a function of the number of added radials. A test oscillator was mounted on a wooden tower four wavelengths away at an elevation angle of six degrees from the base of the vertical.

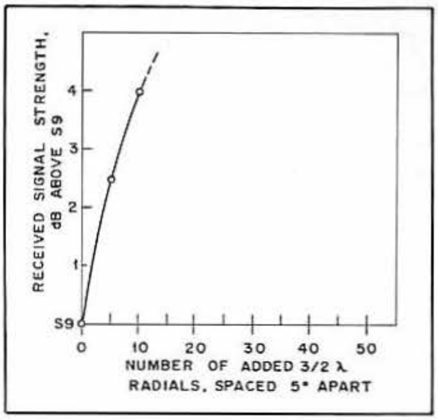

Fig. 22 — The result of interlacing 3/2-wavelength radials in a particular direction. Data were obtained by using a test oscillator on a wooden tower eight wavelengths away, at an elevation angle of 3 degrees from the base of the vertical. The results indicate the advantage of longer wires and the possibility of directional properties of a nonsymmetrical radial system.

100 radials of λ/2 in length, buried just below the surface, provide an adequate ground system. At higher frequencies, the dielectric effect of the earth becomes important, resulting in severe discrimination of radiation or reception at very low angles.[6,7] At low angles, the waves not only suffer by absorption, but also by a change in phase which results in destructive interference. Since little specific information was available at higher frequencies, the objective of the present work was to measure the input impedance and vertical radiation pattern at low angles of a vertical antenna as a function of the number of radials, in order to determine the feasibility of a vertical-array system as a competitor to beams at high elevation. The results of the tests were most gratifying. The author found that many of the "rules of thumb" which have developed and have been perpetuated to the point where they are practically taken for granted were more myth than truth.

Experimental Results

The classical paper on radial systems, which reports experimental results at 1 and 3 MHz, indicates that a large number of radials 0.4λ long should be used.[8] This appeared to be a good starting point for a 20-meter vertical. Accordingly, to check the number of radials needed, this author used eight bundles of wires, each 25 feet (7.62 m) long, and each made up of five no. 18 copper wires. Each bundle was bolted to a 5 × 1/4-inch (127 × 6-mm) aluminum plate as shown in the photograph. The writer was then able to use each bundle as a radial, measure the input impedance of the system, and then separate the bundles, wire by wire, to increase the number of radials in the system.

Fig. 20 shows the input impedance as a function of the number of radials used. Measurements were made with a simple impedance bridge. (Its construction is discussed later.) The antenna was resonated before each measurement was made, and the difference between the 40-radial system and the 8-radial system resulted only by fanning out the bundles of wires. This technique points out another important feature of ground systems and refutes one of the old myths: since the current carried by the radial system is equally divided among n radial elements, each radial is required to carry only 1/n of the total current. This means that relatively small diameter wire is perfectly adequate.

Moreover, it was found that at the higher frequencies it is best to keep the radials near the surface of the ground. Radials buried more than a few feet become less effective! Thus, you should not rely on the old admonition that radials should be 6 feet (1.83 m) down to be effective. This writer found that burying the wires slightly below the surface is the best way of installing the system. Mechanically the radials can be nailed down, electrically they are most effective, and esthetically they provide little interference to a healthy stand of grass.

It will be seen from Fig. 20 that the input impedance for a vertical antenna is drastically affected by the number of radials. Many radial wires are necessary to prevent an excessive loss of power and to provide a convenient input impedance. Fig. 20 also shows the theoretical impedance limit for an antenna having an effective height-to-radius ratio of 300:1.[9]

Fig. 21 shows the effect on low-angle radiation as a function of the number of radials. These data were obtained by placing a test oscillator on a wooden tower, four wavelengths away. A photograph shows the 20-meter vertical and the 26-foot wooden test tower.

The effect of using longer radials in a particular direction is shown in Fig. 22. Radials of no. 18 wire, 3/2 wavelengths long, were put down between the existing 40 radials. The spacing between the longer wires was five degrees. The considerable improvement indicates the need for longer wires and the directional properties a groundplane could give.

A 5/8-wavelength vertical was also constructed and tested. It consisted of a 40-ft. (12.19-m) telescoping aluminum pole and a loading coil of eight turns of no. 12 wire with a diameter of 2.5 inches (64 mm). The comparison in low-angle radiation at a distance of 7.5 wavelengths is presented in Table 5. A field-strength meter was mounted at different heights on the 26-foot (7.93-m) tower. The groundplane for these data consisted of the 40 0.4-λ radials plus the 11 3/2-λ ones. Measurements were taken in the direction of the added longer radials. The results show the improvement in low angle radiation offered by the 5/8-λ vertical. The input impedance of this longer antenna was found to be 76 ohms.

On-the-Air Checks

From these experiments and measurements the author decided to settle on a 1/4-λ antenna with approximately 40 radials. The installation is shown in the photograph. This writer then proceeded to make on-the-air tests to compare its effectiveness with an inverted-V antenna having its apex at 0.4 wavelength, and with a 5/8-λ vertical using the same ground system. Surprisingly, the 1/4-λ vertical seemed to perform just as well as the much taller 5/8-λ vertical, This could result from the fact that most signals arrive after several hops and the optimum lobe angle is probably as high as 15 to 20 degrees.[10] At that angle, the λ/4 antenna actually enjoys an advantage. With practically all DX contacts, the verticals had a 6- to 8-dB improvement over the inverted V! The only exceptions were at intermediate distances and for local contacts. At about 500 or 600 miles, the inverted V with its higher angle of radiation gave better results. Locally, the verticals gave far superior performance. Improvements of 10 to 15 dB were recorded. A triband trap vertical antenna was also tested on 20 meters and found to be practically the same in impedance and performance. This antenna had an overall height of only 12.5 feet (3.81 m)!

Test Equipment

A most interesting aspect of antenna measurements is that the equipment can be rather simple and in many cases constructed from items most amateurs have in their junk boxes. All that is really needed is some standard to compare with, such as a wattmeter or an S meter that is known to have reasonable accuracy. This writer has used both — the wattmeter in a Drake L-4B linear amplifier and the S meter in a R-4B receiver.

The impedance bridge shown in the photograph was patterned after the one shown in the *ARRL Handbook*. Care was taken in shielding the input and output circuits. The meter was mounted externally in order to minimize stray pickup. A calibration curve was obtained at 14.25 MHz by using many carbon resistors of known values as the load.

The field-strength meter is a simple diode detector and dc amplifier. The instrument was constructed to cover the 10-, 15- and 20-meter bands. Its meter was also mounted externally, The antenna length for the field-strength meter was determined by the strength of the available field. In addition to these pieces of test equipment, a 20-meter transistorized crystal oscillator was used for many of the tests.

Conclusion

The performance of a vertical antenna on the ground is highly dependent upon a good ground system, and, properly installed, the antenna can be a very good performer indeed for DX and local contacts. This is particularly true at the higher frequencies where the dielectric property of the earth plays a major role. Forty radials of no. 18 wire, 0.4λ long, will increase the total radiated power by about 3 dB. Radials can be thin if a sufficient number are used. The thought of using thick wires buried deeply, probably a carryover from lightning grounds, is not valid at higher frequencies. The idea of using only four buried radials, as commonly recommended, is a serious error, and if this article does nothing more than eliminate that misconception, this author will be satisfied. Since the electric field only penetrates the ground for a foot or two at the higher frequencies, the radial wires need be buried only as deep as necessary to escape children's feet and the lawn mower. Finally, these results from the single vertical antenna indicate that a vertical array could be a competitor to the more elaborate horizontal beams and warrants further investigation. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] Terman, *Radio Engineering*, McGraw-Hill, New York, 1947, p. 651.

[2] Friis, Feldman, and Sharpless, "The Determination of the Direction of Arrival of Short Radio Waves," *Proceedings of the IRE*, Vol. 22, No. 1, January 1934.

[3] Williams, "Antenna Theory and Design," *The Electrical Design of Antennae*, Vol. 2, Second Edition, Sir Isaac Pitman and Sons Ltd., London, p. 126.

[4] Brown, Lewis, and Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proceedings of the IRE*, Vol. 25, No. 6, June 1937.

[5] Wait, "Input Resistance of L. F. Unipole Aerials," *Wireless Engineer*, May 1955.

[6] Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proceedings of the IRE*, Vol. 21, No. 6, June 1933.

[7] Jager, "Effect of the Earth's Surface on Antenna Patterns in the Short Wave Range," *Internat Elekitron Rundschau 1970*, Nr. 4.

[8] See footnote 4.

[9] King and Harrison, "The Impedance of Short, Long, and Capacitively Loaded Antennas with a Critical Discussion of the Antenna Problem," *Journal of Applied Physics*, Vol. 15, February 1944.

[10] See footnote 2.

The W2FMI Ground-Mounted Short Vertical

Just because the vertical is physically short doesn't mean it won't provide excellent results in DX work. Theory says it can be done. Here's how!

A short vertical antenna, properly designed and installed, approaches the efficiency of a full-size resonant quarter-wave antenna. Even a six-foot (1.83 m) vertical on 40 meters can produce an exceptional signal. Theory tells us that this should be possible, but the practical achievement of such a result requires an understanding of the problems of ground losses, loading, and impedance matching. This paper covers design principles which are applicable to all of our amateur bands.

Background

Earlier work by the author[1] on the radiation efficiency of a ground-mounted vertical resulted from the desire to design a simple, inexpensive, low-profile DX beam for 20 meters. This first work reviewed the theoretical considerations involved and pointed out that a good ground system is essential to achieve efficient operation of a resonant quarter-wave antenna. The second work applied these results to a 20-meter beam.[2] In the process of trying to extend the results of these investigations to the 40-, 80- and 160-meter bands, and to possible multiband operation, the need arose to understand the operation of a shortened vertical and the effects of different loading schemes on the input impedance.

The first part of this paper deals with the theoretical considerations of a short vertical antenna and the experimental procedures involved in measuring the various parameters. This is followed by experimental results which show the trade-offs involved in shortening antennas by various loading schemes. Finally, specific designs are given for the 40- and 80-meter bands.

Theoretical and Experimental Considerations

There are several old familiar axioms in Amateur Radio that are not completely understood by the amateur. The author refers specifically to the two sayings: *Make the antenna as long as possible and erect it as high as possible,* and, *A full-size beam is better than a smaller beam, and a tall vertical is always better than a short one.* The first axiom is particularly relevant to horizontal antennas, where height is most important for reliable DX operation.[3] Increasing the length tends to increase the gain in certain directions. But with the vertical antenna, taking it off the ground and feeding it at a voltage point (like the J antenna) in order to eliminate radials can result in poorer performance because of increased earth loss.[4,5] More experimental work remains to be done on this phase of vertical antennas. The second axiom noted above is generally true, but misinterpreted by many. The author refers to those important properties: power gain — important in transmitting and capture cross-section — important in receiving. (Power gain is the gain, in magnitude, compared to an isotropic radiator; cross-section is a fictitious area related to the ability of the antenna to intercept the radiated power and make it available to the receiver for detection.) As will be seen below, very little is compromised in these properties by shortening the antenna.

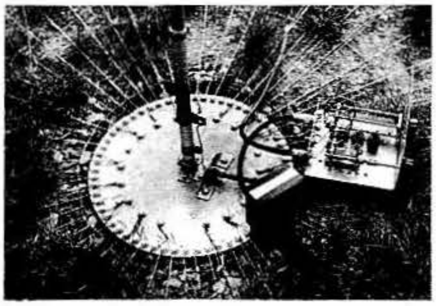

The matching network and base hardware showing 115 radials with the 6-foot (1.83-m) vertical. The pi network is covered with a plastic bag.

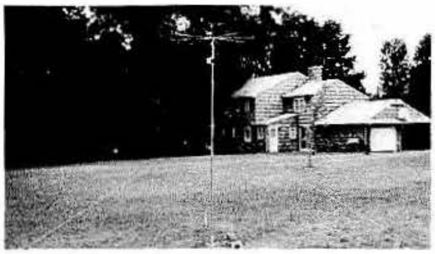

The 10-foot (3.05-m), 40-meter vertical designed for an input impedance of 12.5 ohms in order to use a 4:1 step-up bifilar transformer. The top hat has a diameter of 4 feet (1.22 m) and the coil, placed one foot (0.31 m) below, has 14 turns. Shown at the base is the transformer and impedance bridge.

The author with the 6-foot (1.83-m), 40-meter vertical.

The short antenna has been defined as one that is small compared to a wavelength. In a more exact form, it is defined in such a manner as to simplify the mathematics in the theoretical calculations. King[6] has used the following inequality as the definition where

$$\beta_0 h = \leq 0.5$$

$$\beta_0 = \frac{2\pi}{\lambda}$$

h = half-length of a center-fed antenna or the height of a ground-mounted vertical.

λ = the wavelength

$\beta_0 h$ is actually a quantity which is used to express the height of an antenna in terms of an angle in radians. Thus, this quantity is independent of the frequency. Since 40

Table 6

Inductive Loading

1) *Base Loading*

	*No. of Turns**	*h*	$R_{rad.}$
a)	7	24 ft 3 in.	14.5 ohms
b)	10	18 ft 9 in.	7.5 ohms
c)	12	15 ft 7 in.	5.5 ohms
d)	14	13 ft	4 ohms
e)	18	8 ft 10 in.	2 ohms

2) *Midpoint Loading*

	*No. of Turns**	*h*	$R_{rad.}$
a)	6	28 ft 5 in.	28.5 ohms
b)	11	24 ft 3 in.	25.2 ohms
c)	15	19 ft 8 in.	16.5 ohms
d)	18	15 ft 8 in.	12.3 ohms
e)	24	14 ft 6 in.	10.5 ohms

3) *Three-Quarter Point Loading*

	*No. of Turns**	*h*	$R_{rad.}$
a)	10	29 ft 4 in.	32.5 ohms
b)	18	23 ft	26 ohms
c)	23	21 ft 2 in.	23.5 ohms
d)	24	19 ft 2 in.	22 ohms

*B & W 3029, 2-1/2 in. dia, 6 TPI, no. 12 wire.
1 ft = 0.3048 m

Table 7

Top Hat and Distributed Loading

1) *Top Hat Loading* (4-spoked wheel with 1/8-inch Al wire rim)

	Diameter	*h*	*$R_{rad.}$*
a)	1 ft	30 ft 10 in.	34 ohms
b)	2 ft	28 ft 7 in.	32.5 ohms
c)	4 ft	24 ft	30 ohms
d)	7 ft	19 ft 2 in.	23.5 ohms
e)	4 ft*	23 ft 4 in.	29.4 ohms

2) *Distributed Loading* (Helical Antenna)

	No. of Turns	*h*	*$R_{rad.}$*	*Top Hat at 1.5 ft Above Coil*
a)	111	12 ft	8 ohms	1 ft dia.
b)	105	12 ft	10 ohms	2 ft dia.
c)	113	7 ft	6 ohms	2 ft dia.
d)	75	7 ft	7.5 ohms	4 ft dia.

*8-spoked wheel
1 ft = 0.3048 m

Table 8

Parameters of the 40-Meter Short Vertical Designs

Total Height	*$R_{rad.}$* (ohms)	*No. Turns**	*Dia. Top* (ft)	*Bandwidth*
6 ft	3.5	14	7	100 kHz
8-1/2 ft	7.5	at 6 in. below top hat		
8-1/2 ft	7.5	75 on 7 ft dowel 1-5/8 in. in dia.	4	100 kHz
10 ft	12.5	14 1 ft below top hat	4	125 kHz
15 ft 10 in.	12.5	7 at base	4	540 kHz

*Except for helical antenna, coil wire is same as shown in Table 6.
1 ft = 0.3048 m

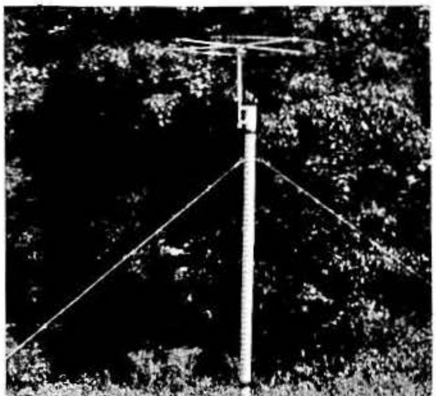

Eight-and-one-half-foot (2.59-m) helical antenna using a 2-foot (0.61-m) top hat 1-1/2 feet (0.46 m) above the 7-foot (2.13-m) helix. The 75 turns have approximately a 3-inch (76-mm) pitch below the midpoint and 1.5-inch (38-mm) pitch above. Very little difference was noted in reversing the pitches. The input impedance was 7.5 ohms.

meters was used mostly for experimental results presented in this article, the inequality above assures accurate theoretical calculations for verticals of 11 feet (3.35 m) or less.

The theoretical results show that the power gain for a *very* short antenna, even less than 1-foot (0.31-m) high on 40 meters, is 1.5.[7] This increases slowly to 1.513 for an 11-foot (3.35-m) antenna. These gains are to be compared to about 1.62 for a resonant quarter-wave vertical. It can be seen that this difference amounts to less than 0.4 dB or 0.07 S unit, based on 6 dB per S unit.

In a similar fashion, the capture cross-sections differ by relatively small amounts. For the *very* short antenna, a cross-section value is $0.119\ \lambda^2$, while for the quarter-wave vertical it is $0.13\ \lambda^2$. This is surprising to many since it is difficult to visualize a vertical of a foot or two (0.31 or 0.61 m) in height on 40 meters, having practically the same receiving ability as a 33-foot (10.06-m) quarter-wave antenna!

But the important property of a short vertical that makes its capture cross-section nearly the equivalent of a full $1/4$-λ antenna is its *very small value of input resistance.* Fig 23 shows the theoretical curve and the 40-meter experimental results for the input resistance of a ground-mounted vertical as a function of height.[8] The experimental data were obtained by essentially canceling out the capacitive reactance of the vertical by an inductance at the base of the antenna and measuring the resistive value with an impedance bridge. Since an extensive, low-loss radial system was used, the resistance measured was actually that of the antenna itself, which is called the *radiation resistance.*

Fig. 24 shows the input resistance of a quarter-wave resonant vertical as a function of the number of radials. At the 115-radial point, the input resistance approaches the theoretical value of 35 ohms, which strongly indicates low earth loss and, hence, reliable data in short antenna measurements. The picture also shows the base hardware for 60 radials. The wire is 15-gauge aluminum. There is nothing sacred about 15 gauge. No. 22 wire, or even no. 28, would be just as good.

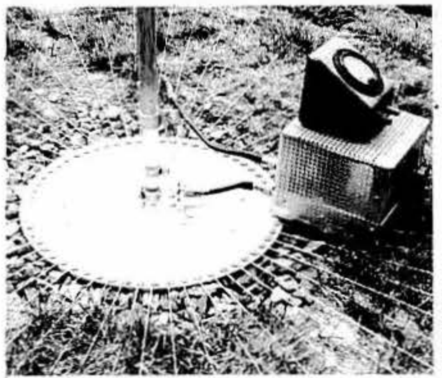

Base of the vertical antenna with 60 radials. The aluminum disk is 15 inches (0.38 m) in diameter and 1/4 inch (6 mm) thick. Sixty tapped holes for 1/4-20 aluminum hex-head bolts form the outer ring and 20 form the inner ring. The insulator is polystyrene material with a 1-inch (25-mm) diameter. Also shown is the impedance bridge for measuring input resistance.
Originally, circular wires were used to connect all radials together. These were positioned every 2 feet (0.61 m), starting from the antenna base. Then, one at a time, they were removed, meanwhile keeping a check on the antenna radiation resistance. There was no apparent change in the radiation resistance, so it was concluded that the interconnecting rings of wire were not needed in the ground system.

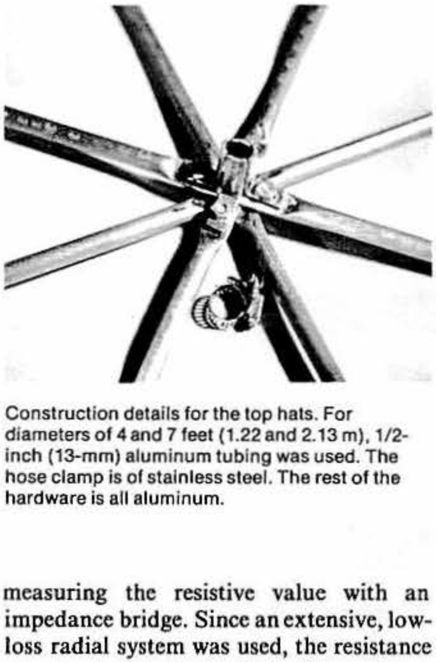

Construction details for the top hats. For diameters of 4 and 7 feet (1.22 and 2.13 m), 1/2-inch (13-mm) aluminum tubing was used. The hose clamp is of stainless steel. The rest of the hardware is all aluminum.

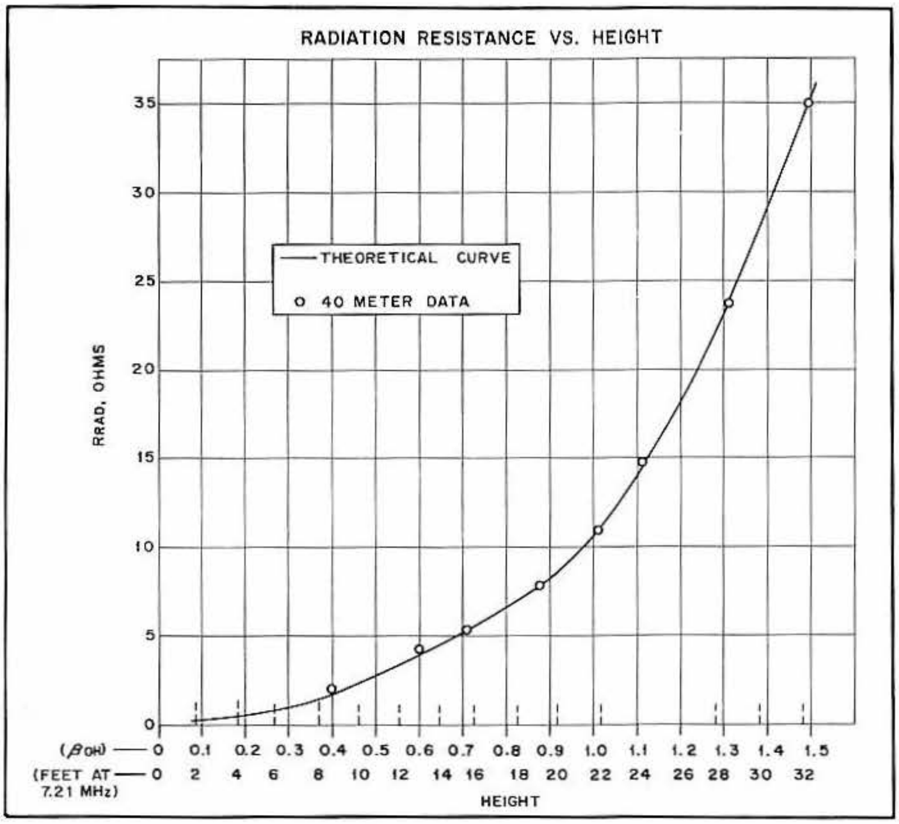

Fig. 23 — Theoretical curve and experimental results for the radiation resistance as function of height.

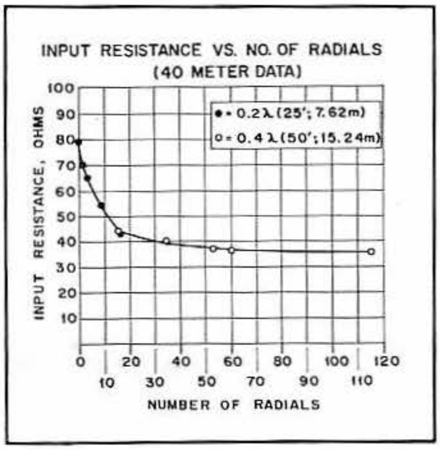

Fig. 24 — Input resistance of a quarter-wave ground-mounted resonant vertical as a function of the number of radials.

It should also be noted that very little difference was noted whether 0.2-λ or 0.4-λ radials were used. At 16 radials, the results are practically identical. In fact, this curve is also similar to the one obtained on 20 meters which was reported earlier.[9] More work remains to be done on the trade-offs in performance versus length of radials for ground-mounted verticals.[10]

Experimental Results

In designing a shortened vertical beam for 40, 80 or 160 meters, a most important consideration is the value of the input impedance of the driven element. As seen in Fig. 23, a shortened vertical using base loading has an extremely low value, particularly at one-eighth wavelength and less. This input impedance is then usually lowered in the presence of other elements in a beam array. The resulting very low value of input impedance makes it difficult to design matching networks. Therefore, an experimental investigation was undertaken to see what increases in radiation resistance could be obtained by other methods of loading, i.e., top hat, three-quarter point, midpoint, and distributed (helical antenna). The results of these experiments are shown in Fig. 25. Table 6 shows the individual points for inductive loading and Table 7 for top hat and distributive loading. Several interesting points were brought out by these experimental results. Fig. 25 shows that top-hat loading yielded the largest value of radiation resistance for a particular height. Surprisingly, the helical antenna[11] yielded a value less than midpoint loading. The three-quarter point and midpoint loading curves were not extended to lower values of height because data were very difficult to obtain below the points shown on the respective curves. The combinations of inductances and lengths below the heights shown on these two curves were probably beyond resonant conditions at the frequency used in the measurements. The other curves were extended by dashed lines indicating no difficulties were encountered in the measurements and other lengths were very possible.

Another interesting aspect of the top-hat loading curve is that a four-spoked wheel approaches to a good degree a solid disk. Doubling to eight spokes only improves the loading by about nine percent as noted in Table 7. Thus, *a few radials on the top of a vertical are very effective.*

Four radials at the base, approximating a ground system, are practically *useless* as noted in Fig. 24. Also, it can be seen from Table 7 that the reduction in length because of top-hat loading is approximately equal to twice the diameter of the disk.

On 160 meters, top hats have been made with sloping wires or struts with success.[12] Because of the sloping nature of this top hat, some cancellation of the antenna current takes place, thus reducing the radiation resistance further.

Although the curves in Fig. 25 were obtained from experiments at 7.21 MHz, these data can be applied to the other bands by proper scaling. For example, by doubling all dimensions, including the number of turns of the loading coils, the radiation resistance values would apply at a frequency of 3.6 MHz. By increasing the dimensions by only 1.85 instead of 2, the results would then apply to a frequency of 3.9 MHz. In like manner, a proper scaling factor could be used to apply these results to a portion of any of the amateur bands.

40- and 80-Meter Short-Vertical Designs

As was stated before, a most important consideration in designing short verticals is a knowledge of the input resistance and how it varies with different types of loading. The objective is to obtain a resistance value large compared to earth losses, so that efficient operation is obtained. In the author's specific case, using 115 radials, practically no earth loss was measured, and, hence, any radiation resistance above a few ohms assured good operation and an opportunity to verify the theoretical predictions for very short vertical antennas.

Since a broad-band, bifilar, 4:1 step-up transformer was available from previous work,[13] the first design was for the shortest vertical having an input resistance of 12.5 ohms. From the curves of Fig. 25, it appeared that a 16- or 17-foot (4.88- or 5.18-m) antenna with a coil of some 13 turns (extrapolated from Table 6) at 8 or 9 feet (2.44 or 2.74 m) from the base would provide the proper impedance at 40 meters. Further, it was decided that some 7 or 8 feet (2.13 or 2.44 m) of length above the coil could be replaced by an eight-spoked top hat having a diameter of 4 feet (1.22 m). The actual design that resulted, after proper tuning, is shown in the picture. The antenna had a total height of only 10 feet (3.05 m) and a 14-turn coil placed 1 foot (0.31 m) below the top hat. This height was about one foot (0.31 m) higher than first expected, but upon careful examination it was noted that the top hat also reduced the radiation resistance, while replacing a section of the vertical portion. Therefore, the height had to be increased somewhat when considering top-loading effects. Also shown in a picture are the construction details for the top hat.

Two other shortened verticals for 40 meters were investigated. One was an 8-1/2-foot (2.59-m) helical using a 4-foot (1.22-m) top hat and 75 turns on a 1-5/8-inch (41-mm), 7-foot (2.13-m) long, wooden dowel. The input impedance was 7.5 ohms and it was matched with a standard pi network. Pictures are shown with the network. Several tests were made by doubling the winding pitch below and above the midpoint, keeping the number of turns constant, with very little difference in results.

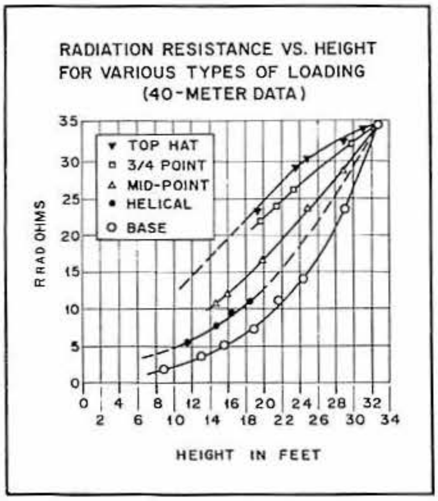

Fig. 25 — Experimental results of radiation resistance as a function of height of antenna for various types of loading (40-meter data).

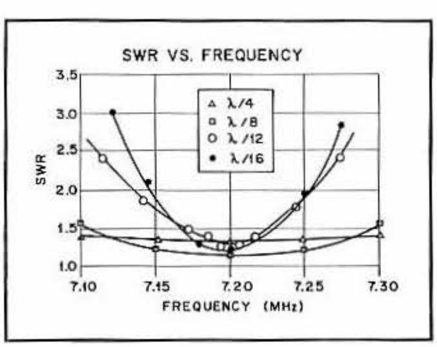

Fig. 26 — Standing wave ratio of various short verticals compared to a resonant quarter-wave antenna. The sixteenth-wavelength antenna is the 8-1/2-foot (2.59-m) helical design, the twelfth-wavelength is the 10-foot (3.05-m) antenna, and the eighth-wavelength is the 15-foot 10-inch (4.83-m) antenna. (All three are described in the text.)

The second antenna, also for 40 meters, using a 7-foot (2.13-m) top-hat, resulted in a matched vertical only 6 feet (1.83 m) high. It had a 14-turn coil, 6 inches (152 mm) below the top hat and an input impedance of only 3.5 ohms. Matching was accomplished with the 4:1 transformer and the pi network. A picture of the antenna and matching network is shown.

A larger 40-meter vertical antenna was also investigated. It had a 4-foot (1.22-m) top hat and 7 turns of base loading which resulted in a height of 15 feet 10 inches (4.83 m) (approximately 1/8 wavelength) and an input impedance of 12.5 ohms. The purpose of this design was to compare its low-angle radiation and bandwidth[14] with the other antennas. Table 8 shows the parameters of these various short vertical designs. Fig. 26 shows the SWR curves of three of the short antennas compared to a quarter-wave resonant vertical and Fig. 27 shows the SWR of the 6-foot (1.83 m) antenna.

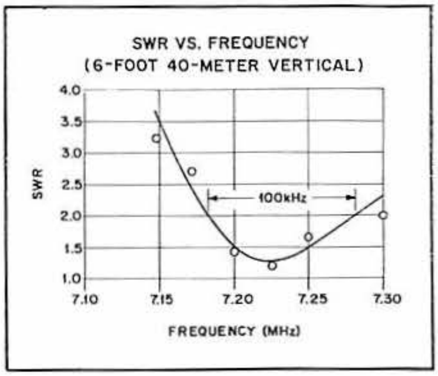

Fig. 27 — Standing wave ratio of the 6-foot (1.83-m) vertical using a 7-foot (2.13-m) top hat and 14 turns of loading 6 inches (152 mm) below the top hat.

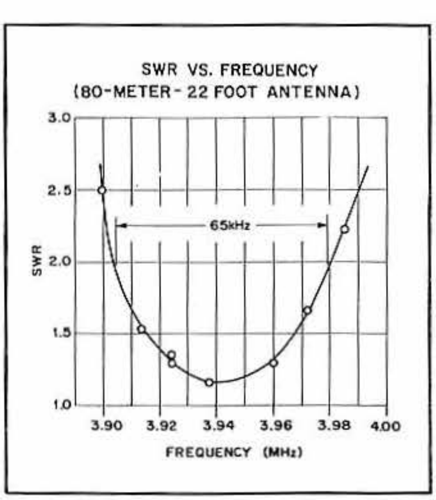

Fig. 28 — The standing wave ratio for the 22-foot (6.71-m), 80-meter vertical.

As can be seen by the SWR curves, shortening an antenna generally decreases its bandwidth. Also by comparing the 8-1/2- and 6-foot (2.59 and 1.83-m) antennas top-hat loading appears to affect the SWR the least. It should also be noted that an eighth-wavelength antenna appears to have a reasonable bandwidth and would probably result in a practical beam design.[15]

And finally, the four short-vertical designs described above were compared with a resonant quarter-wavelength antenna for low-angle radiation and on the air with other amateurs signals. The low-angle radiation measurements were made at three wavelengths in distance and at heights of 1, 3, 6 and 8 feet (0.31, 0.91, 1.83, 2.44 meters). This relates to vertical angles of 0.15, 0.45, 0.90 and 1.2 degrees, respectively. All measurements were made under matched conditions and with a constant 100 watts fed into the antennas under test. In no case were there any appreciable differences noted in the field strength measurements. In fact, the 6-foot (1.83-m) antenna seemed to give slightly higher readings! These measurements certainly tend to verify the theory on the power gain predicted for short verticals. On-the-air checks were again very gratifying and exciting. Over 200 contacts with the 6-foot (1.83-m) antenna strongly indicated the efficiency and capability of a short vertical. Invariably at distances greater than 500 or 600 miles, the short verticals yielded excellent signals.

As was stated previously, all of the results obtained by measurements on 40 meters can be scaled, by the appropriate ratios of frequencies, to other bands. This was tried out on 80 meters. Since a 7-foot (2.13-m) top hat was available (instead of an 8-foot [2.44-m] one), the height turned out to be 22 feet (6.71 m) instead of 20 (6.10 m). The loading coil had 24 turns and was placed 2 feet (0.61 m) below the top hat. On-the-air results duplicated those on 40 meters. The bandwidth was 65 kHz (half of the 40-meter value) as shown in Fig. 28.

Conclusions

Several interesting results came out of this investigation which, even to one schooled in antenna theory, are difficult to believe. The author refers specifically to the ability of very short verticals to radiate and receive as well as a full-size quarter-wave antenna. The differences are practically negligible. But, as was seen, the trade-offs are in lowered input impedances and bandwidths. However, with a good image plane and a proper design, these trade-offs can be entirely acceptable.

As a result of this investigation, other problems were noted which indicate the need for further experimentation. The writer refers specifically to the trade-offs encountered when shortening the radials (to 0.1 wavelength, for example), the efficiency of ground-mounted versus elevated verticals as a function of height above ground, the characteristics of a multiband vertical over a low-loss image plane, and a shortened vertical beam using elements of the order of an eighth wavelength.

Another point should be mentioned in relation to the results reported here. Even though they were obtained on short, ground-mounted verticals, they are valid for center-fed verticals or horizontal antennas as well. The only differences are that the impedance values should be doubled and the effect of the image antenna accounted for.

The author would like to acknowledge the support of many amateurs for their fine words of encouragement and excellent reporting during on-the-air contacts. In this study, some 350 amateurs reported on comparisons between signals from these short verticals and those of other stations. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] Sevick "The Ground-Image Vertical Antenna," *QST,* July 1971. Also appears elsewhere in this chapter.

[2] Sevick, "The W2FMI 20-Meter Vertical Beam," *QST,* June 1972.

[3] See footnote 1.

[4] Miller, Poggio, Burke, Seldon, "Analysis of Wire Antennas in the Presence of a Conducting Half-Space. Part II. The Vertical Antenna in Free Space," *Canadian Journal of Physics,* Vol. 50, 1972.

[5] Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. of the Inst. of Radio Engineers,* Vol. 21, no. 6, June 1933.

[6] King, *The Theory of Linear Antennas,* Harvard Univ. Press, Cambridge, MA, 1956.

[7] See reference in footnote 6, p. 497.

[8] See reference in footnote 6, p. 190.

[9] See footnote 1.

[10] Weeks, *Antenna Engineering,* McGraw-Hill Book Company, 1968, p. 48.

[11] The diameter of the helical antenna was 1-5/8 inches (41 mm). Each helical antenna had a top hat 18 inches (457 mm) above the helix in order to terminate the upper turns with sufficient capacitive reactance.

[12] See reference in footnote 10, p. 44.

[13] See footnote 2.

[14] Bandwidth is defined as the range in frequency where SWR is less than 2:1.

[15] See footnote 2.

Optimizing Vertical Antenna Performance

Do you have tower restrictions, a small yard or patio, almost no room for antennas? If so, here are some suggestions for making your signal bigger in spite of the handicaps.

Antennas never cease to be a hot topic of conversation among amateurs; the interesting articles by John Stanley, K4ERO/HC1[1] and Roger Hostenback, W5EGS,[2] on the optimum grounding of vertical antennas have led the author to communicate these thoughts on the subject. Obstacles at the author's present location (typical of many suburban dwellings) place severe limitations upon where ground wires can be placed. These obstacles are large boulders, areas of concrete, the house, vegetable and flower beds, and the neighbors' property lines. It has not been possible to lay out radials of uniform length or of uniform angular distribution. Nevertheless, the author has had successful results with vertical antennas. Although the writer must make a number of statements which are based upon fragmentary evidence and are lacking in scientific rigor, they should be useful for suggesting how others might proceed to optimize their particular situations.

The writer has three very dissimilar vertical antennas. Much of the discussion will pertain to their individual ground systems, although strictly speaking they cannot be independent; a change of the grounding conductors of one vertical is bound to affect the others. For brevity the writer shall not discuss the methods of impedance matching of the antennas as this is a topic thoroughly covered elsewhere. The author *shall* mention some problems of using dissimilar antennas in a phased array, although the results are preliminary.

The three antennas are (1) a shunt-fed crank-up tower having a three-element triband Yagi on top, the height being variable from 27 to 41 feet (8.23 to 12.50 m); (2) a 40-foot (12.19-m) insulated base-loaded antenna consisting of a 30-foot (9.14-m) pipe mast topped by a whip, located approximately 70 feet (21.34 m) south of the tower and hereafter referred to as the "south antenna"; (3) about 20 feet (6.10 m) east of the tower, a vertical antenna termed the "north antenna," which was described in an earlier *QST* article.[3] Fig. 29 shows the layout. The north antenna has been used with various heights up to 40 feet (12.19 m) but in the present experiments it is used in its "summer" configuration with a height of 25 feet (7.62 m) and with a 7/14-MHz trap located 10 feet (3.05 m) from the base. For 3.5-MHz operation additional loading is supplied at the base.

The Grounding Systems

None of the three verticals have anything approaching an optimum ground system; while the south antenna has the best on paper, it is not necessarily the best in performance. When the south antenna was first placed into operation the grounding system consisted of about 200 feet (60.96 m)

How do you ground a vertical antenna when there is no ground?

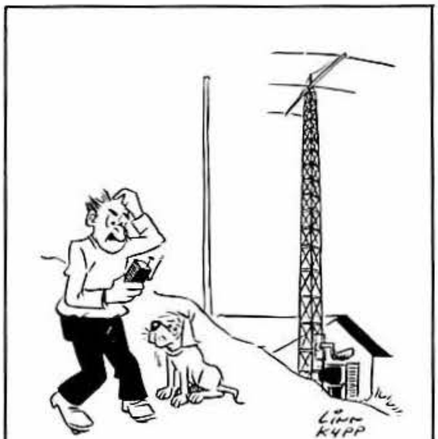

Optimizing ground-wire currents is well worth the time and effort required.

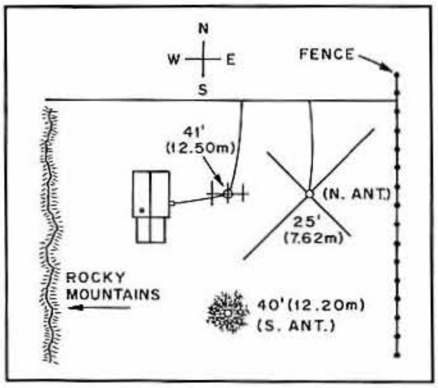

Fig. 29 — The W0JF backyard, with antenna farm.

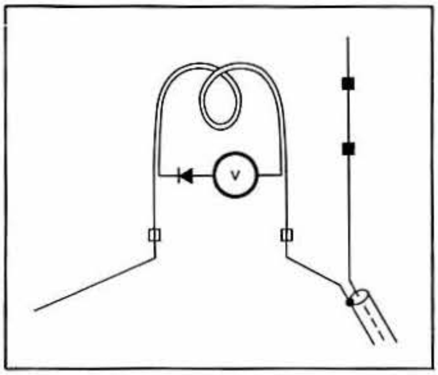

Fig. 30 — Suggested substitute for a thermocouple ammeter.

of wire distributed among a few conductors which went out more or less radially. One was about 60 feet (18.29 m) long, the others 30 feet (9.14 m) or less. On 7 MHz the performance was generally inferior to that of a horizontal half-wave dipole 30 feet (9.14 m) above ground. Later, the ground system was augmented to utilize about 1000 feet (304.80 m) of wire, and the vertical then began outperforming the dipole.

While the author foresaw Stanley's statement that possibly not all of this additional wire would be effective, it was assumed that any wires added at least would do no harm, and the writer tried to put wires in the entire immediate area of the antenna, subject to the constraints mentioned above. Several of the radials have bends to avoid obstacles. In a few cases a single wire goes out from the antenna to a junction some distance away, and then several wires fan outward from the junction. Three wires are connected at the far ends to a few square feet of chicken wire. Where the wires are in shrubbery they are above ground; otherwise they are an inch or two (25 or 51 mm) below. None of the additional wires were as long as 60 feet (18.29 m), and the lengths were generally determined by the obstacles they encountered. Haphazard as this system was, the improvement was spectacular.

While making this change in the south antenna's grounding, the author carried out two simple experiments. First, when operating at 7 MHz, a 33-foot (10.06-m) piece of wire was added, which in free space would be 1/4-λ long, but which when laid on the ground would be longer (in wavelengths). The writer connected it to the rest of the ground system through a thermocouple ammeter, and read the current while applying power. With this wire then cut in half and laid out as separate radials in opposite directions, it was found that the total current increased. Cutting these 1/8-λ pieces in half and distributing the resulting fragments as four radials in directions at 90 degrees to each other caused the total current to decrease. These observations led the author to conclude that for a given total amount of wire the most efficient layout would be radials having lengths of about 1/8 a free-space wavelength. This conclusion is consistent with Stanley's Table 1.[1]

In the other experiment, the author also used a 33-foot (10.06-m) piece of wire, but half of it was no. 34 and half no. 12 wire. With the thick end adjacent to the antenna, the current was higher than with the wire turned around so that the thin end was adjacent. The experiment showed that wire diameter is of some consequence, and it is more important to have heavy conductors close to the antenna.

The Tower and the North Antenna

Grounding of the tower is very heterogeneous; first of all, of course, the bottom of the tower is buried in the ground some 8 feet (2.44 m). There are two radials about 25 feet (7.62 m) long going in a more or less southerly direction, then a short wire going to a water faucet on the wall of the house. Finally, and probably most important, there is a wire that goes northward about 20 feet (6.10 m) and makes a junction with the middle of a 120-foot (36.58-m) east/west wire. The eastern end of this wire is located at the common corner with the property of the neighbor to the northeast and at that point is bonded to his extensive chain link fence. About 25 feet (7.62 m) west of this bond to the fence, there is a junction with a southward wire which is connected to the base of the north antenna. Otherwise, the grounding system of the north antenna, consisting of a few radial wires, is as described in the earlier *QST* article.

Some observations were made by connecting a thermocouple ammeter and a selection of capacitors and inductors in series with various conductors connected to the base of the tower. On 3.5 MHz, with the wire that goes to the fence, the author found a very strong resonance using a 300-pF capacitor. No such resonance was found with this wire of 7 MHz nor on 1.8 MHz, nor did there seem to be any resonance with any of the other wires.

No attempt has been made to bond the upper and lower sections of the tower (where they telescope) together. No such bonding seems necessary, except in the rare case of an extremely strong wind, when there is some receiver noise resulting from poor conductivity between the sections. Cranking the tower up and down seems to have little effect upon the degree of impedance mismatch.

On 3.5 MHz, using a power of about 150 watts, the author had contacts all over the world using this shunt-fed tower, usually not cranked all the way up. The south antenna has also given fairly good account of itself, but with its thousand feet (304 m) of radial wires its performance is still not quite as good as that of the tower with its meager "black magic" grounding.

On 7 MHz, the behavior of the tower is the reverse even though the impedance match is somewhat superior to that on 3.5 MHz. The performance, despite the good match, is far inferior to the south antenna or even the north antenna. Offhand, one would expect the performance, such as it is, to be better with the tower cranked down to minimum height, when it would be close to 1/4λ, but the few reports received on the air indicate better performance is obtained at maximum height. The overall performance of the tower on 7 MHz is a complete mystery.

The author has used the tower on 1.8 MHz only rarely, but has worked as far as 1000 miles, even though there is a very bad impedance mismatch on the feed line. One of the author's future projects is to cure that, since the tower may be magic on the top band, too!

What if You Don't Have a Thermocouple Ammeter?

The preceding discussion has, the author hopes, suggested that individuals may be able to optimize their ground systems by measuring the currents in individual ground wires. In this regard the writer has been fortunate in having available a number of thermocouple ammeters, instruments not commonly available in ham shacks. However, it was found easy to make a substitute which can serve to measure relative current and to detect resonances. Take a short length of twin-lead, speaker cord or similar cable, about a foot (0.305 m) long. Connect one wire of the pair in series with the circuit to be studied (the ground wire). Connect the two ends of the other wire of the pair to a semiconductor diode and a sensitive dc meter in series (Fig. 30). For convenience and for reproducibility in results, coil the cable up and secure it to a terminal board.

Using Dissimilar Antennas in a Phased Array

The author can use any two of the three verticals as a phased array by connecting the feed lines in parallel at the transmitter. In the shack the author has a switchboard that can switch additional lengths of line in series with one antenna or the other. The writer does not attempt to change the impedance match at the bases of the individual antennas, but does generally retune the transmitter as the phasing is changed. Some of the effects observed must be attributed to variations in impedance mismatches rather than to basic changes in directivity. Nevertheless, some gain can usually be obtained by combining the verticals in one way or another.

In the use of two dissimilar antennas, two questions arise which may not apply when identical antennas are used. First, is there really an advantage in using the two of them in a phased array over using the better of the two alone? The relevance of this question may be seen by considering the extreme case where one is very much better than the other. In such a case the field strength of the poorer one is so much less that there is little difference whether it is in or out of phase with the field of the better one, and there can be little or no observable directional effect when the two are used together. Furthermore, the power supplied to the better antenna is cut in half while the other half of the power is largely wasted.

This simple argument suggests that in most cases the two antennas should be nearly equal in the field strengths they produce if they are to be effective in a phased array. However, the actual situation is sufficiently complex that exceptions may be found. For one thing, the better antenna may have "blind" directions in its pattern because of obstructions, and the poorer one may help to fill those in. Also, the impedance effects which have been mentioned before are hard to predict. The author can only suggest that one may find it worthwhile to try using two such antennas together, and may find unexpected advantages in some directions.

On 3.5 MHz the north antenna is generally much poorer than the tower. However, in some directions observers have consistently reported a small gain in using the two together over the tower alone. Incidentally, it should be mentioned that observations made on receiving are not reliable tests, partly because of impedance effects. Also, some of the noise received may be generated locally and its effect may mask the true pattern of the array. The only reliable evaluation is by transmitting to distant observers.

The second question concerning dissimilar antennas involves solving for an unknown quantity: the relative phase of the two antennas. The relative phase of the field of an antenna is *not* determined solely by the length of the transmission line connecting it to a transmitter. The impedance of the antenna usually contains a reactive component, and its presence gives rise to a phase shift. With two identical antennas these phase shifts are the same and do not affect the relative phase between the two antennas, which is determined only by the difference in feed-line lengths. With dissimilar antennas these phase shifts are generally not identical, and they may vary in different ways with shifts in frequency or with variations in the moisture content of the earth. Not unrelated was a discovery made when changing the grounding of the south antenna from 200 feet (60.96 m) of wire to 1000 (304.80 m). There was a substantial change in the phasing when it was used in an array with the tower.

The author has experimentally answered this phasing question by setting up a small probe antenna at a point equally distant from the two antennas and connecting it to a field-strength meter. The line lengths are varied until a minimum (usually quite sharp) is observed. Then it is known that the antennas are 180 degrees out of phase, and the array is a bidirectional end-fire one (W8JK). If the author adds or subtracts 1/2 λ of line from one antenna, the antennas should be in phase and a maximum should be observed. The experimental observations are in agreement with such predictions. If a length of line equal in electrical wavelengths to the spacing between the antennas is then added or subtracted from the feed line of one of them, the combination approximates a unidirectional end-fire array. The favored direction is away from the element which has the longer equivalent feed line (larger phase shift from the transmitter).

To reiterate, the results with phasing dissimilar verticals are very preliminary, and there is little that can be said about them except for the fact that at times very noticeable gains over using the better single antenna are observed.

Vertical Versus Horizontal Antennas

The author used some *horizontal* antennas and observed the differences in performance between them and the vertical antennas at a particular location. Horizontal antennas were not tried on 7 and 3.5 MHz any higher than 30 feet (9.14 m) above ground, however.

The first foothills of the Rocky Mountains rise up immediately to the west of the author's QTH, the nearest ones subtending a vertical angle of about 10 or 15 degrees. As expected, propagation to the west is usually difficult. To the east there are some power wires running north and south. The land slopes off to the north, and has a line-of-sight view for 12 miles from the base of the antenna tower.

The writer finds a tendency for horizontal antennas to be superior to the west, over the mountains, and for verticals to be superior to the east. Standard theory explains these findings as follows: Probably many of the signals coming over the mountains are scattered by the surface of the ground. Those familiar with advanced electromagnetic theory remember that according to the Fresnel equations, the reflection coefficient is greater when the wave is polarized with the electric vector parallel to the surface (i.e., horizontal). Therefore, this scattered radiation is probably horizontally polarized. On the other hand, the electric field from the vertical antenna is perpendicular to the direction of the power wires and is less disturbed by them than the field of a horizontal antenna.

Although very few firm results were presented, the author hopes he has suggested some procedures which will help those facing restrictions on grounding systems they can use. With a little persistence and ingenuity, they should be able to build effective antennas. It *is* possible to evaluate simple arrays yourself and to improve their performance according to your own communication requirements. This material was originally presented in *QST* by Yardley Beers, W0JF.

References

[1]Stanley, "Optimum Ground Systems for Vertical Antennas," *QST*, December 1976.

[2]Hoestenback, "Improving Earth-Ground Characteristics," *QST*, December 1976.

[3]Beers, "Short Antennas for the Lower Frequencies," *QST*, September 1970.

Some Plain Facts about Multiband Vertical Antennas

Does engineering jargon, page-long equations and references to the *IEEE Proceedings* have you gnawing your nails in frustration? Take heart, for these bits of knowledge about multiband vertical antennas are given in plain lingo.

During discussions with newcomers, and old-timers for that matter, it becomes apparent that there is considerable confusion as to what exactly a multiband vertical antenna is. The confusion concerns the method of feed, how much mismatch one can expect, how many radials are required, how the particular antenna is built for multiband use, plus some other points.

This article breaks the subject into simple language and provides the reader with sufficient expertise to assure him that he won't wind up with a system he really doesn't want. Before going into a discussion of the different types of multiband "verticals" we will offer some simple antenna facts.

Some Basic Theory

The term "multiband antenna" has come to mean many things to hams. With trap antennas, tapped coils, random wires, and so forth, there is plenty of reason for the confusion. Simply, a multiband antenna is one that can be used on more than one band. How we make it work on different bands is another story.

Basically, any piece of wire of *any* length can be classed as a multiband antenna. For example, a length of wire 4 feet (1.22 m) long *could* by used on *any* amateur band, from 160 meters on up. But how well the piece of wire would work is a completely different matter.

In the feed point of any antenna there is *radiation resistance*. The energy supplied to an antenna is dissipated in the form of radio waves and in heat losses in the wire and nearby insulating materials. The radiated energy is the useful part, but so far as the transmitter is concerned it represents a power consumption just as much as does the energy lost in heating the wire. In either case the dissipated power is equal to I^2R; in the case of heat losses, R is a real resistance (ohmic losses), but in the case of radiation, R is an *assumed* resistance. This fictitious resistance is the radiation resistance. This brings us to our first important point about multiband antennas.

Whenever one reduces the size (length) of an antenna physically, the radiation resistance is reduced also. As an example, assume we have a 20-meter quarter-wave antenna, which is approximately 16 feet (4.88 m) long. Let's imagine we made it out of no. 40 wire, which has a resistance of about 1 ohm per foot, (0.31 m). The radiation resistance of a resonant quarter-wave vertical operated against a perfect ground is on the order of 35 ohms. In this case, the feed-point impedance of our antenna would be roughly 35 ohms in radiation resistance plus 16 ohms in ohmic resistance. If we were to feed 51 watts into this antenna 16 watts would be dissipated as heat (lost power) and the remainder — 35 watts — would be radiated. Now, suppose we use this same antenna on 80 meters. As mentioned above, when we reduce the size of an antenna physically the radiation resistance is also reduced. On 80 meters our 16-foot (4.88-m) antenna would have a radiation resistance on the order of 1 ohm! However, we would still have the ohmic resistance of 16 ohms. It doesn't take much figuring to realize that just about all of our power would be lost as heat.

Of course we wouldn't use no. 40 wire for such an antenna. More likely the antenna would be made from aluminum tubing and the ohmic losses would be very low, but probably still more than the radiation resistance. There is an old axiom in Amateur Radio that offers some pretty good advice: Always make the antenna as long as possible and erect it as high as possible. Also, there is a joke that goes with that axiom — if such an antenna stays up, it is too small!

At this point we have only mentioned radiation and ohmic resistance in the antenna feed point. These are the two resistances that exist when the antenna is resonant. When the antenna is not resonant, there is reactance present in the feed point. Reactance is also expressed in ohms, but it isn't a real resistance in the sense that power can be dissipated therein. We won't go into a long discussion on reactance because it would take up too much space. An excellent explanation can be found in the League publication, *Understanding Amateur Radio*. Simply, reactance can be likened to a gate or door that stops or hinders the flow of current into a circuit. When an antenna is operated at some frequency other than the resonant frequency there will always be reactance present. Keep in mind that with *any* antenna, multiband or otherwise, we always have a condition on some band or frequency where the antenna is not resonant. Therefore, there will be reactance at the feedpoint.

Types of Vertical Antennas

The basic and most popular type of vertical is one that is a quarter wavelength long and is operated against ground or in a groundplane configuration. The antenna is usually made from tubing and the radials are wire. An ideal groundplane (simulated earth ground) would be a sheet of metal with

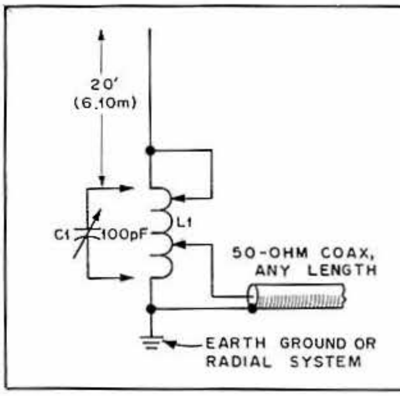

Fig. 31 — A typical multiband vertical antenna. A description of the system is given in the text.

a radius of one-quarter wavelength or more. However, this is only practical at vhf so the customary method is to use wires as the radials. Probably the number one question asked about groundplane antennas is, "how many radials are required?" The answer is simply, the more radials used, the better the antenna will perform, at least up to a certain point. This should not be construed to mean that an antenna with only two or three radials won't work. Such an antenna will work, but for *maximum* performance one should consider 40 or more radials. If the reader is interested in performance data for a few radials versus many, he should read the article in *QST* by Sevick.[1]

The feed-point impedance of quarter-wave groundplane is on the order of 35 ohms. The impedance can be raised by drooping the radials down until a 50-ohm match is obtained. Exactly how much droop is required depends on the number of radials.

The quarter-wave groundplane is essentially a single-band antenna. However, a 40-meter quarter-wave vertical can also be used on 15 meters, a happy circumstance for the Novice. In this case, a 40-meter quarter wave works out to be three quarter waves on 15 and any *odd* multiple of quarter waves will provide a relatively low-impedance feed.

Multiband Verticals

When we get into the field of multiband verticals we find that considerable confusion exists. As pointed out previously, *any* antenna can be *called* a multiband antenna, but how we get power into such an antenna is another matter.

Up until the '50s any amateur multiband antenna was a system that usually consisted of an antenna, tuned feeders, and an antenna coupler. In the early '50s more and more amateurs started to use coaxial cable for feeders, along with band-switching transmitters. The next logical step was the use of a multiband antenna system that required no adjustments and always presented a matched condition to the feed line — in other words, an antenna that had a 50-ohm feed-point impedance on every desired band and frequency within a given band. A logical development was the multiband trap antenna.

By inserting traps in an antenna it was possible to make an antenna "look" like a resonant half-wave dipole in whichever band was used; or, in the case of multiband verticals, making the vertical look like a resonant quarter-wave antenna for the desired band. However, and this is important as far as the newcomer is concerned, to our knowledge there is *no* multiband trap antenna that will provide a perfect match on all bands, *regardless* of what some antenna manufacturers may tell you. Many hams have spent countless hours trying to adjust trap antennas for that "perfect" match when actually, it is just about impossible to obtain such a condition.

Nontrap Multiband Verticals

Several antenna manufacturers sell multiband antennas that consist of a vertical piece of tubing, usually 16 to 20 feet (4.88 to 6.10 m) long. The tubing is used with a loading coil at the ground end. By making appropriate taps and adjustments on the coil the antenna can be matched (or closely matched) on any given band. This type antenna has *no* traps. This in turn means that the coil taps and adjustments *must* be altered when one changes bands. Some misguided amateurs buy these antennas expecting all they need do is put them up and the antenna will work on all bands, automatically. Let's make one point clear: such an antenna *is* a multiband antenna, but requires adjustment *at the antenna* when one changes bands.

Of course, the next question should be, "If the antenna is that simple can't I build my own?" Yes, it is a very simple multiband antenna to make and install. Two or three sections of inexpensive 10-foot (3.05-m) TV mast sections can serve as the vertical radiating element. The mast can be supported on an insulator, such as a beverage bottle, and the mast guyed with nylon line. Fig. 31 shows a diagram of the antenna system. L1 should be a coil made of bare wire, no. 12 or 14, so that it can be tapped at every turn. A convenient coil size is 2-1/2 inches (64 mm) in diameter, six turns per inch (25 mm), such as B & W 3905-1 stock. The number of turns required, assuming 80 meters as the lowest band to be used, should be about 30 turns with an antenna length of 25 feet (7.62 m).

Adjustment of the antenna requires the use of an SWR bridge. Connect the coax line across a few turns of L1 and make a trial position of the shorting tap. Measure the SWR, then try various positions of the shorting tap until the SWR reaches its lowest value. Then vary the line tap similarly. This should bring the SWR down to a low value. Small adjustments of both taps should provide an SWR close to 1. If not, try adding C1 and repeat the adjustment procedure, varying C1 each time until a match is achieved. Radials will enhance the performance of the antenna. The number of radials is up to the individual amateur.

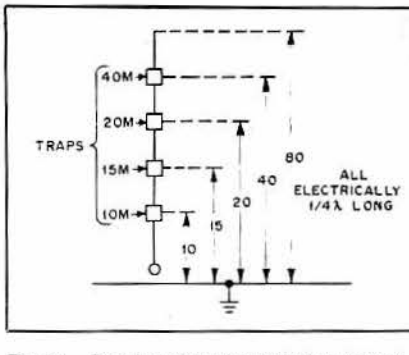

Fig. 32 — Drawing of the theoretical multiband trap vertical. In commercial practice, certain traps may be grouped together, giving the impression that only a single trap is used.

Trap Verticals

As mentioned earlier, traps can be installed in a multiband vertical. These traps are usually parallel-tuned circuits and the objective is to make each section of the antenna work as a quarter-wave vertical or odd multiple thereof on the desired band. Fig. 32 shows an example of this type antenna.

The purpose in using this type antenna is to provide a system that always presents a matched condition for the feed line. Unfortunately, there is so much interaction between various sections of the antenna that it is impossible to come up with a *perfect* match on each band. What is an *acceptable* match is another story.

Amateurs as a whole are inclined to attach too much importance to an SWR of 1. They feel that if their SWR bridge isn't showing an absolute zero reflected power that something is horribly wrong and they won't work out. The plain fact is that using a feed line such as RG-8/U (assume a 100-foot [30.48 m] length) one could have an SWR of as much as 5:1 and have *no* appreciable loss in the system. However, there is one clinker in this thinking!

In many instances commercially made transmitters and transceivers are designed by the manufacturer to work into a 50-ohm load only. They don't allow much leeway from this figure. When there is a mismatch in the antenna system, it can become impossible to load and tune the final amplifier of the transmitter. There just isn't enough tuning range in the tank circuit of the amplifier to handle the reactance that may be present in the load. There is a way around this problem however, and that is using a Transmatch in the feed line to disguise the mismatch.[2] The Transmatch can be adjusted so that the transmitter "sees" a 50-ohm load regardless of the mismatch at the antenna.

The Harmonic Problems

Another consideration should be mentioned. As pointed out earlier, *any* antenna can be a multiband antenna. By the same token any harmonics generated in the transmitter that reach the antenna can be radiated. It is true that a single-band antenna will reject harmonic energy, but *not* completely. In the case of a multiband trap antenna there is no rejection of some harmonics, simply because the antenna is designed to be resonant on all hf amateur bands. The solution to this problem is the use of a selective circuit installed in the feed line. A Transmatch is such a circuit and should provide adequate harmonic rejection.

Some Other Thoughts

The question is frequently asked, "Should I mount my vertical on the ground, or get the base up in the air?" Getting the antenna up in the clear is always better than having it mounted at earth level and surrounded by rain gutters, house wiring, trees, power lines and so forth. However, getting the vertical antenna up in the air also means that radials, as many as possible, should be used. The average installation (if there is such a thing) usually consists of three or four radials (or more) cut for the lowest operating frequency. Such a system should give a good performance.

Another important matter is that of the earth ground. When verticals are mounted at ground level the ground losses can be very important. Too many amateurs buy their verticals, get a 5-foot (1.52-m) long TV-type ground rod and drive it into the earth at the base of the antenna. They think this provides a good ground connection. As a matter of fact, the TV-type ground rods are practically worthless for amateur work. A *good* ground rod is the type used by the power company for home installations. This is a rod that is heavily galvanized, 5/8 inch (16 mm) in diameter, and about 10 feet (3.05 m) long. The amateur should be able to buy these rods from any wholesale electrical supply house. If possible, tie your ground connection to the water-system piping, assuming metal piping is used.

You'll hear the statement from fellow hams that verticals are poor antennas and radiate poorly in all directions. This isn't true, because a vertical *can* by a good antenna, but you have to give it a fighting chance. This material was originally presented in *QST* by Lewis McCoy, W1ICP.

References

[1] Sevick, "The Ground-Image Vertical Antenna" *QST,* July 1971. Also appears elsewhere in this chapter.

[2] A suitable Transmatch is described in the transmission-line chapter of any edition of *The Radio Amateur's Handbook.*

Shunt-Feeding Towers for Operation on the Lower Amateur Frequencies

The 160- and 80-meter DX antenna you've dreamed about may already be standing on your property — just waiting to be used as a DX vertical. Read this presentation and learn how to shunt-feed your tower on 1.8 and 3.5 MHz!

Shunt-feeding a tower used to support beam antennas for 14 MHz or higher frequencies has obvious advantages for the amateur who wants to work all lower bands. If a good ground system is installed, the result may be a very effective antenna for DX work on 80 and 160. The shorter tower installations may work very well on 40 meters, if the effective height above ground is less that 5/8 wavelength. Beyond this height the radiation angle goes higher, and the effectiveness for DX goes down.

The shunt-fed tower is at its best on 160, where a full quarter-wavelength vertical antenna is rarely possible. Almost any tower height can be used. If the beam structure provides some top loading, so much the better — but anything can be made to radiate, if it is fed properly. A self-supporting, aluminum, crank-up, tilt-over tower is used at W5RTQ with a TH6DXX tribander mounted at 70 feet (21.34 m). Measurements showed that the entire structure has about the same properties as a 125-foot (38.10-m) vertical. It thus works quite well on 160 and 80 in DX work requiring low-angle radiation. It can also be used on 40, but results are rather poor because of high radiation angle. WB5FKX, with a 37-foot (11.28-m) tower and a 21-MHz beam, finds that shunt feed on 40 enables him to work DX that he never knew existed before.

Close-up view of the base of the shunt-fed tower, showing the two bottom arms and their insulators.

Preparing the Structure

Usually some work on the tower system must be done before shunt-feeding is tried. Metallic guys should be broken up with insulators. They can be made to simulate top loading, if needed, by judicious placement of the first insulators. Don't overdo it; there is no need to "tune the radiator to resonance" in this way. If the tower is fastened to a house at a point more than about one-fourth of the height of the tower, it may be desirable to insulate the tower from the building. Plexiglas sheet, 1/4-inch (6-mm) or more thick, can be bent to any desired shape for this purpose, if it is heated in an oven and bent while hot.

All cables should be taped tightly to the tower, preferably on the inside, and run down to the ground level. It is not necessary to bond shielded cables to the tower

Table 9

Shunt-Feed Details for Several Towers Used on 160 Meters

Station	*Height of Tower* (ft)	*"Top-Hat"* (Beams)	*Average Height of Beam* (ft)	*Height of Wire-to-Tower Connection* (ft)	*Wire-to-Tower Spacing* (in.)	*Tuned Value of Gamma Capacitor*
W5RTQ	69	TH6DXX	70	68	30	150 pF
K5PFL	86	Stacked 40, 20, 15, 10 M beams	95	75	24	125 pF
K4PUZ	64	TH6DXX	66	63	8	300 pF
K8KAS	70	TH6DXX	77	68	24	225 pF
W1CER	50	QST 20-M DX Weasel	55	47	44	400 pF

1 ft = 0.3048 m
1 in. = 25.4 mm

electrically, but there should be no exceptions to the down-to-the-ground rule.

No rf problems have developed with rotators, beam traps, or even TV sets where the TV antenna is part of the radiating structure, as is the case at WB5FKX. His only precaution was use of coax on the TV antenna. It would be well to proceed with caution, as every installation could be different in this respect.

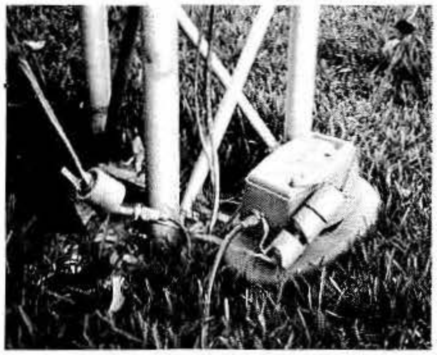

Though variable capacitors are shown in Fig. 33, they were all replaced with fixed-value units except one, which is inside a plastic box at the tower base.

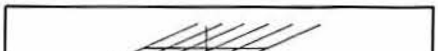
CONNECTION TO TOWER
NO. 8 COPPER SPACED 30" (0.76m) MAX. FROM TOWER
80/75-METER FEED WIRE, NO. 8 ALUMINUM SPACED 15"(0.38m) FROM TOWER
STANDOFF INSULATOR
STANDOFF INSULATOR
150-pF VARIABLES
250-pF VARIABLE
BURIED RADIALS CONNECTED TO BASE OF TOWER
COAXIAL FEED

Fig. 33 — Principal details of the shunt-fed tower at W5RTQ. The 160-meter feed, left side, connects to the top of the tower through a horizontal arm of 1-inch (25-mm) diameter aluminum tubing. The other arms have stand-off insulators at their outer ends, made of 1-foot (0.31-m) lengths of plastic water pipe. The connection for 80/75, right, is made similarly, at 28 feet (8.53 m), but two variable capacitors are used, to permit adjustment of matching with large changes in frequency.

Though the effects of ground losses are less severe with the shunt-fed vertical than with the simple quarter-wave antenna, a good system of buried radials is very desirable. The ideal would be 120 radials, each 250 feet (76.20 m) long, but fewer and/or shorter ones must often suffice. You can sneak them around corners of houses, along fences or sidewalks, wherever they can be put a few inches under the surface, or even on the earth surface. Aluminum clothesline wire is used extensively at W5RTQ, and it stands up well. Neoprene-covered aluminum wire may be safer in highly acid soils. Contact with the soil is not important. Deep-driven ground rods, and connection to underground copper water pipes, are good, if usable.

Installing the Shunt Feed

Principal details of the shunt-fed tower for 80 and 160 meters are shown in Fig. 33. Rigid rod or tubing can be used for the feed portion, but heavy gauge aluminum or copper wire is easier to work with. Flexible stranded no. 8 copper wire is used for the 160-meter feed at W5RTQ, because when the tower is cranked down, the feed wire must come down with it. Connection is made at the top, 68 feet (20.73 m), through a 4-foot (1.22-m) length of aluminum tubing clamped to the top of the tower, horizontally. The wire is clamped to the tubing at the outer end, and runs down vertically through standoff insulators. These are made by fitting 12-inch (0.31-m) lengths of PVC plastic water pipe over 3-foot (0.91-m) lengths of aluminum tubing. These are clamped to the tower at 15 to 20-foot (4.57 to 6.10-m) intervals, with the bottom one about 3 feet (0.91 m) above ground. The lengths given allow for adjustment of the tower-to-wire spacing over a range of about 12 to 36 inches (0.31 to 91 m), for impedance matching.

The gamma-match capacitor for 160 is a 250-pF variable with about 1/16-inch (1.6-mm) plate spacing, which is adequate for the power levels presently authorized. The omega match used for 80 and 75 permits retuning for large excursions in frequency encountered in using both cw and phone on this band. Two capacitors are required, each about 150 pF, with plate spacing of about 3/16 inch (4.8 mm), if full power is used. They can be mounted in plastic refrigerator containers for protection against the weather. Use well-insulated knobs to avoid rf burns during the adjustment process.

Separate 50-ohm lines run underground to the station. The shield side of their connectors should be grounded to the base of the tower and to the buried radials as directly as possible.

Tuning Procedure

It is suggested that the 160-meter wire be connected to the top of a structure 75 feet (22.86 m) tall or less. Note, from Table 9, that the monster at K5PFL was fed at 75 feet (22.86 m) above ground. Mount the standoff insulators so as to have a spacing of about 24 inches (0.61 m) between wire and tower. Pull the wire taut and clamp it in place at the bottom insulator. Leave a little slack below to permit adjustment of the wire spacing, if necessary.

Adjust the series capacitor in the 160-meter line for minimum reflected power, as indicated on an SWR meter connected between the coax and the connector on the capacitor housing. Make this adjustment at a frequency near the middle of your expected operating range. If a high SWR is indicated, try moving the wire closer to the tower. Just the lower part of the wire need be moved for an indication as to whether reduced spacing is needed. If the SWR drops, move all insulators closer to the tower, and try again. If the SWR goes up, increase the spacing. There will be a practical range of about 12 to 36 inches (0.31 to .91 m). If going down to 12 inches (0.31 m) does not give a low SWR, try connecting the top a bit farther down the tower. If wide spacing does not make it, the omega match shown for 80-meter work should be tried. No adjustment of spacing is needed with the latter arrangement which may be necessary with short towers or installations having little or no top loading.

The two-capacitor arrangement is also useful for working in more than one 25-kHz segment of the 160-meter band. Tune up on the highest frequency, say 1990 kHz, using the single capacitor, making the settings of wire spacing and connection point permanent for this frequency. To move to the lower frequency, say 1810 kHz, connect the second capacitor into the circuit and adjust it for the new frequency. Switching the second capacitor in and out then allows changing from one segment to the other, with no more than a slight retuning of the first capacitor.

The omega match is recommended for 80-meter operation, because of the wide tuning range required. It was found that the point of connection could be at 28 feet (8.53 m), using a single support at the top of the first tower section. This served as a mount for the second standoff insulator for the 160-meter feed, as well.

A 40-meter feed, not shown, was connected at about 20 feet (6.10 m) up, with about 8 inches (203 mm) spacing between wire and tower.

Substituting Fixed-Value Capacitors

Depending on the frequencies used and the matching problems in a given installation, it may be possible to put fixed-value capacitors in place of some or all of the variables shown in Fig. 33, and then switch them remotely, or by hand, when changing frequency. This was done at W5RTQ. The 160-meter gamma capacitor is now a fixed-value, 150-pF type designed for high rf current service. It does not have to be weatherproofed, and it gives a satisfactory match over the small frequency range used on this band.

Two or three similar capacitors connected in series are used in covering the cw and phone frequencies in the 80-meter band. Operation around 3800 kHz requires about 50 pF in the omega (parallel) capacitor, so the three 150-pF units are used in series. Moving to cw operation near 3500 kHz requires about 75 pF, so only two 150-pF capacitors are used in series. The only variable now in the system is the 150-pF capacitor mounted in the plastic refrigerator box at the base of the tower. It is the series or gamma capacitor in the 80-meter feed. Even this could be replaced with a fixed unit, as it has been found unnecessary to adjust it in the course of normal operation of the station, if only small changes in frequency are made in using the two modes.

Because every tower installation is likely to be different from every other one in some respects, the values and dimensions given here may be subject to change, but the general principles should hold. Proof of this was found in the May 1975 issue of *Ham Radio*,[1] wherein W4OQ describes shunt-feeding in very similar terms. He also presented estimates of the electrical height of various tower-beam combinations that prospective shunt-feed users should find of interest.

It is hoped that this information will inspire others to try shunt-feeding their towers. This method offers an effective way to fire up on the lower frequencies that is useful to many who may have wanted to work these amateur frequencies but have shied away from doing so because of limited antenna space. This material was originally presented in *QST* by Earl W. Cunningham, K6SE.

---

[1] True, "Shunt-Fed Vertical Antennas," *Ham Radio*, May 1975.

The Minooka Special

This 160-meter vertical antenna could be just the one for you, if you reside on property of city-lot size or are a mobile operator. You can make it from readily available, inexpensive material.

Inspiration for the design offered here resulted in part from the author's survey of antennas in use by 160-meter operators, worldwide. An extensive report was compiled, which filled a loose-leaf notebook! A boil-down of the most useful information resulted in an 11-page report which was made available to numerous amateurs. Subsequently, the writer was encouraged to submit some of the data for publication. Approximately 60 operators were polled. The following information from that inquiry should be of interest to those who are curious about "preferred" antennas for top-band use.

Question 1: "If you could put up any antenna for 160, what would it be?" *Result:* 60% favored verticals, 30% said horizontals, and 10% indicated mixed feelings. (The term "vertical" includes various configurations — 1/4, 1/2, and 5/8 wavelength, vertical arrays, and inverted Ls.)

Question 2: "Comparing antennas that an average ham could build, do you prefer verticals or horizontals for 160?" *Result:* 70% said vertical, 17% favored horizontal, 5% inverted L, 2% horizontal and vertical, 2% inverted V, and 3% had no opinion. Reasons given for the responses were, "Because of signal comparisons and past experience. A backyard-compatible vertical is more effective than a backyard-compatible horizontal."

Question 3: "Do you operate mobile on 160? If so, what is the antenna used?" *Result:* Three used base-loaded verticals, eight used center-loaded ones, and one employed a Heliwhip. Some of the operators used capacitance hats. Other questions in the survey dealt with receiving antennas, types of soil and terrain, besides requesting complete details of the present transmitting antenna.

Fig. 34 shows some of the vertical systems used when the poll began in 1969. Additional configurations used by successful 160-meter operators are illustrated in Fig. 35. It can be seen that a great deal of useful "backyard theory" is represented in the systems shown. The experience and trials of others can save countless years of experimenting with antennas (not that some of us can't use the exercise)! It is hoped that these data will be of value to future 160-meter operators.

The Minooka Special

The following is a description of an effective vertical antenna for 160 meters, designed with these objectives in mind:

1) Highly effective for 160-meter DX and local work.
2) Easy to build and adjust.
3) Very economical.
4) Fits neatly into back yard.
5) Reasonable bandwidth.
6) Good for portable and DXpedition work.
7) Can be scaled down for mobile operation.

The resultant antenna (Fig. 36) is top loaded inductively and can be built by anyone from readily available material. Only a dip meter and SWR indicator are needed for tune-up. Many versions were built and tested, ranging from 7-foot (2.13-m) mobile types to 60-foot (18.29-m) backyard

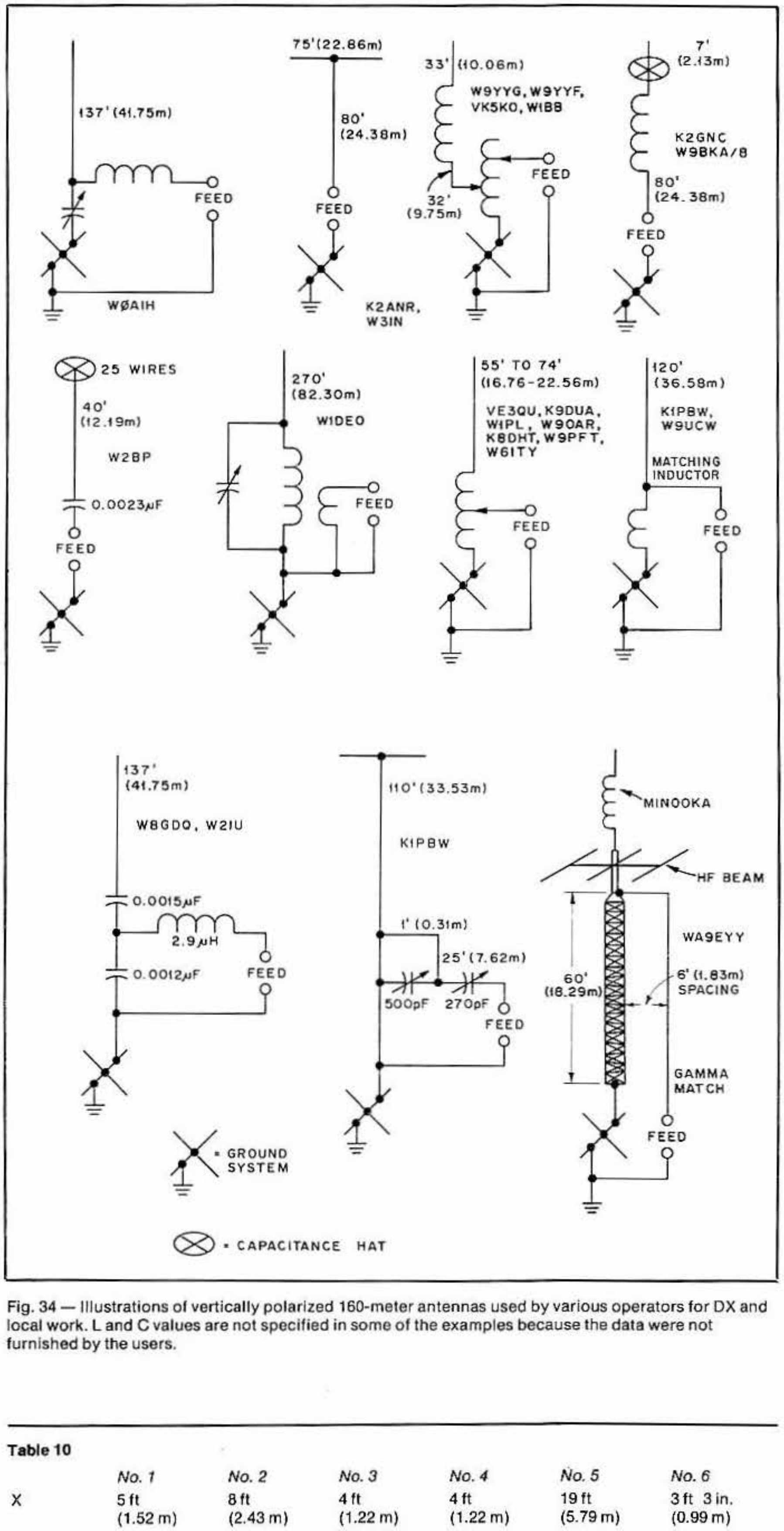
137' (41.75m)
FEED
WØAIH
75'(22.86m)
80'
(24.38m)
FEED
K2ANR,
W3IN
33' (10.06m)
W9YYG, W9YYF,
VK5KO, W1BB
32'
(9.75m)
FEED
7'
(2.13m)
K2GNC
W9BKA/8
80'
(24.38m)
FEED
25 WIRES
40'
(12.19m)
W2BP
0.0023μF
FEED
270'
(82.30m)
W1DEO
FEED
55' TO 74'
(16.76-22.56m)
VE3QU, K9DUA,
W1PL, W9OAR,
K8DHT, W9PFT,
W6ITY
FEED
120'
(36.58m)
K1PBW,
W9UCW
MATCHING
INDUCTOR
FEED
137'
(41.75m)
W8GDO, W2IU
0.0015μF
2.9μH
0.0012μF
FEED
110'(33.53m)
K1PBW
1' (0.31m)
25'(7.62m)
500pF
270pF
FEED
MINOOKA
HF BEAM
WA9EYY
60'
(18.29m)
6' (1.83m)
SPACING
GAMMA
MATCH
FEED
= GROUND SYSTEM
= CAPACITANCE HAT

Fig. 34 — Illustrations of vertically polarized 160-meter antennas used by various operators for DX and local work. L and C values are not specified in some of the examples because the data were not furnished by the users.

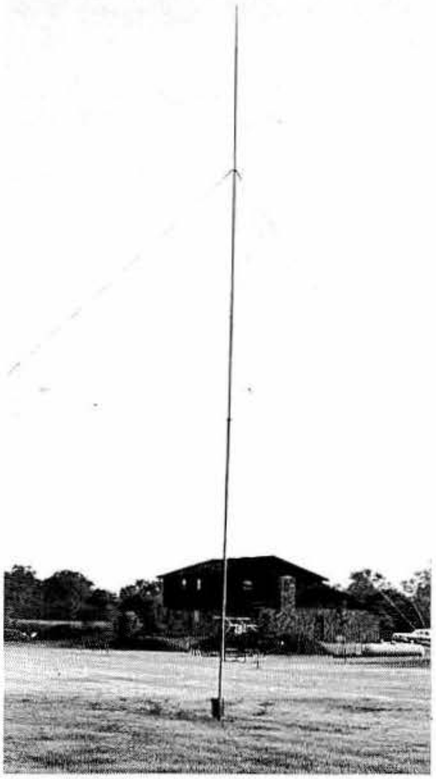

Here is a 50-foot (15.24-m) version of the Minooka Special. It is tuned remotely at K9SKX, and is used on 160 through 20 meters. Nylon or polypropylene guy lines should be used.

or DXpedition models. They have been used with good results from 20-foot (6.10-m)-wide backyards in cluttered Chicago, to vast beaches on Carribean islands, and in South America. The fixed-tuned bandwidths vary from 10 kHz for mobile versions, to 50 kHz for the larger fixed-station models (SWR 2:1 or better).

Construction

The physical layout of the antenna is centered around the use of 3/4-inch (19-mm) diameter rigid PVC water pipe. This sturdy tubing has an outside diameter of 1 inch (25 mm). The inner diameter provides an "interference fit" for 1/2-inch (13-mm) EMT conduit (thinwall). A 3/8-24 nut is driven into a 5-inch (127-mm) length of conduit, then aligned and braized into position. Next, the conduit is taper-ground and polished on the opposite end, then driven into the PVC tubing (see table of dimensions).

In the construction of mobile antennas another piece of conduit is driven into the bottom of the PVC coil form and a 3/8-24 bolt is brazed into the bottom end. The bolt will mate with standard mobile antenna mounts. For larger versions of the antenna the PVC material can be mounted in TV masting, or whatever. A standard 8-foot (2.44-m) stainless-steel whip between 3 and 8 feet (0.91 and 2.44 m) in length can be screwed into the top section. For home-installed versions of the antenna a

Table 10

	No. 1	*No. 2*	*No. 3*	*No. 4*	*No. 5*	*No. 6*
X	5 ft (1.52 m)	8 ft (2.43 m)	4 ft (1.22 m)	4 ft (1.22 m)	19 ft (5.79 m)	3 ft 3 in. (0.99 m)
Y	2 ft (0.61 m)	15 in. (381 mm)	3 ft 6 in. (1.07 m)	4 ft (1.22 m)	11 in. (279 mm)	3 ft (0.91 m)
Z	As long as possible					
Wire Size	No. 20	No. 19	No. 18	No. 16	No. 19	No. 22

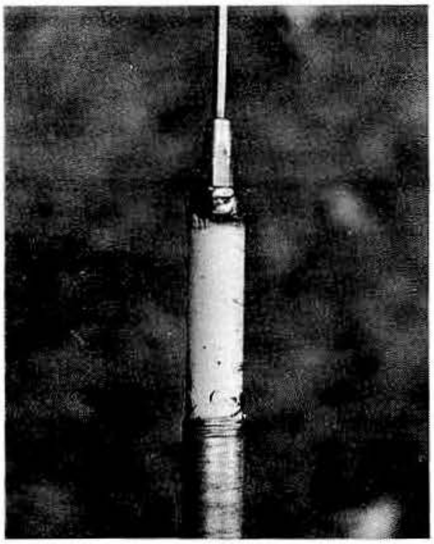

Close-up view of the top section of a coil assembly. This unit was used during mobile operation for over three years (40,000 miles or 64,374 km).

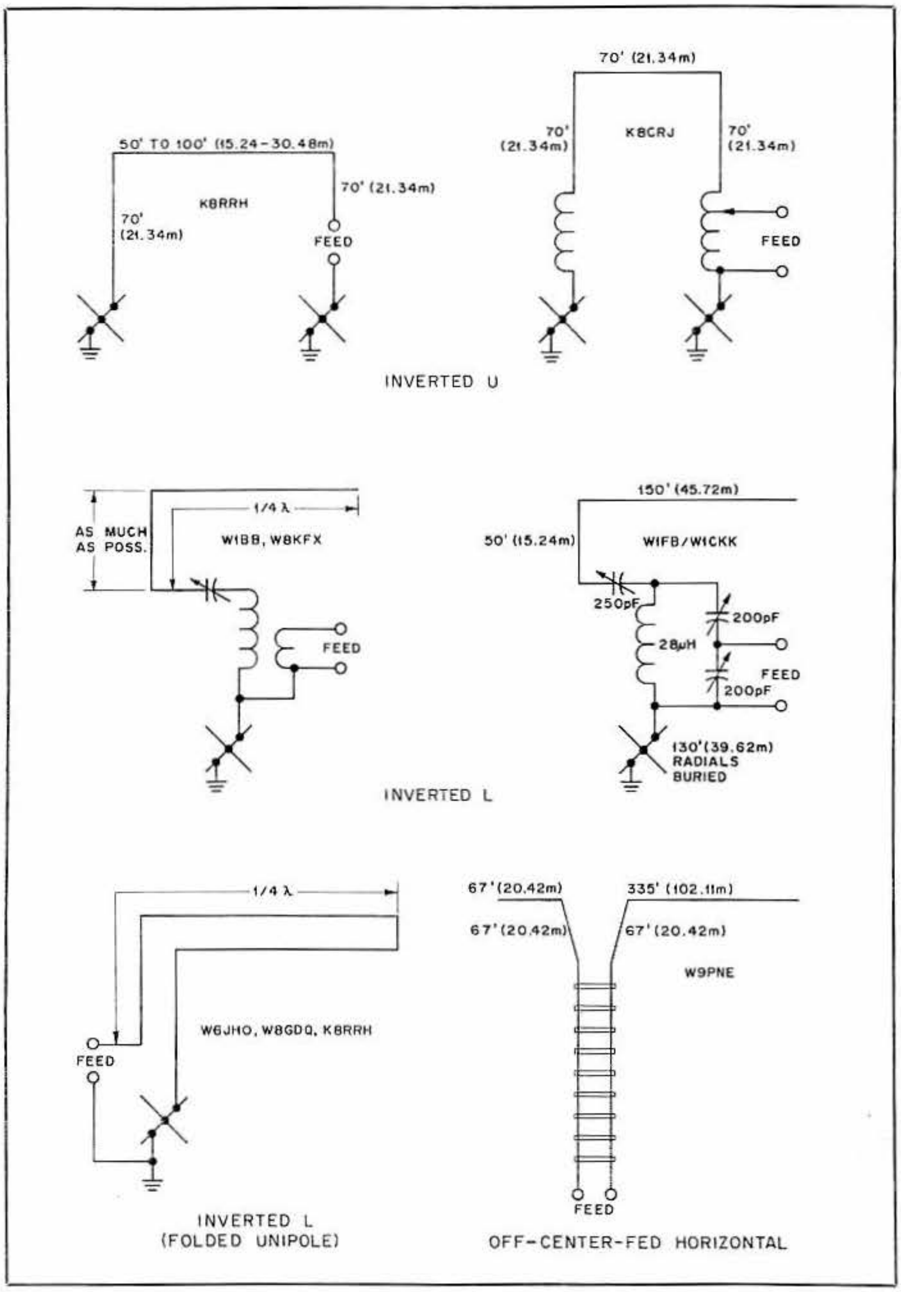

50' TO 100' (15.24-30.48m)
70' (21.34m)
K8RRH
70'
(21.34m)
FEED
70' (21.34m)
70'
(21.34m)
K8CRJ
70'
(21.34m)
FEED
INVERTED U
AS MUCH AS POSS.
1/4 λ
W1BB, W8KFX
FEED
150' (45.72m)
50' (15.24m)
W1FB/W1CKK
250pF
200pF
28μH
FEED
200pF
130' (39.62m) RADIALS BURIED
INVERTED L
1/4 λ
W6JHO, W8GDQ, K8RRH
FEED
INVERTED L
(FOLDED UNIPOLE)
67' (20.42m)
335' (102.11m)
67' (20.42m)
67' (20.42m)
W9PNE
FEED
OFF-CENTER-FED HORIZONTAL

Fig. 35 — Examples of additional antennas for 160 meters which proved effective for long-haul work.

A mobile version of the Minooka Special is shown mounted on the author's car.

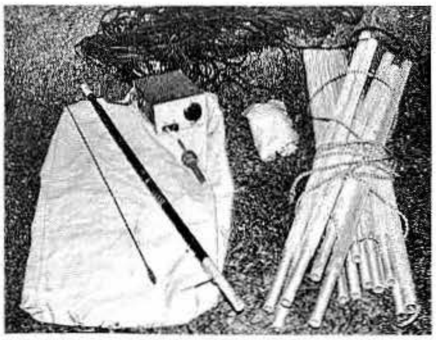

A portable version of the 160-meter vertical antenna. It was used by W9YYG and W9UCW at HK0BKX in January 1974.

3- or 4-foot (0.91- or 22-m) piece of thin-wall tubing can be used in place of the whip. This will save on the cost of materials, and will eliminate the need to have brazing done. The coil wire should be soldered to the conduit to assure a good electrical connection.

Adjustment

Pick a set of dimensions from the table which suits your application, but add a few inches more of coil turns (all turns close wound) than are recommended. This will allow leeway for pruning the system to resonance.

To simplify adjustment it is suggested that the system be assembled first with only the coil and top section (no base section). Place the antenna where it is in the clear (on the car or fixed-station site), and where it can be tuned against the proposed ground system — car body or ground radials.

A three-turn link should be connected temporarily between the lower end of the coil and the ground system. This will permit rough tuning of the system to resonance by inserting a dipper coil into the link and adjusting the coil turns on the antenna until a dip is noted in the desired part of the band. Upon completion of the pruning the constructor can, if he wishes, cover the coil with weatherproof tape or shrink tubing. The antenna should be tuned for roughly 2000 kHz if the entire band is to be used. Adjust the resonance for 1850 kHz if only the low segment will be utilized.

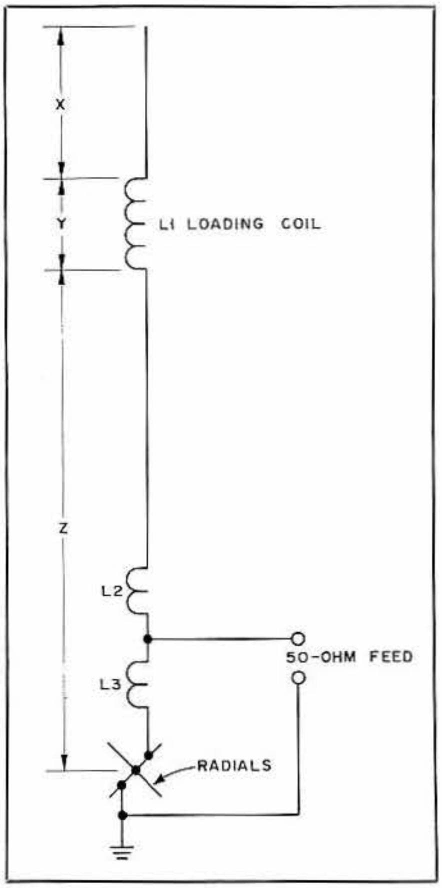

Fig. 36 — Electrical details of the Minooka Special. Table 10 gives specific information concerning dimensions X, Y and Z. L2 may vary in size from 1 to 20 turns, and L3 will contain between 5 and 10 turns, typically. L2 and L3 are made from no. 18 wire, spaced 1/8 inch (3 mm) between turns. The coil diameter is 1-1/2 inches (38 mm). Refer to text for tuning instructions.

Erect the antenna with all of its parts — coil, top and bottom section — and insert inductors L2 and L3 of Fig. 36 as shown. With L3 temporarily out of the system, adjust L2 for the lowerest value of SWR obtainable at the desired operating frequency. Then, place L3 back in the circuit and adjust it for an SWR reading of 1. Addition of the base section later on will not affect the resonant frequency of the overall system materially, provided the base section does not exceed, say, 60 feet (18.29 m).

The foregoing steps are used also in adjusting the mobile version of the antenna. However, because of the small size it is possible to adjust the antenna in completely assembled form. Only L3 of Fig. 36 is needed for mobile antennas. The main coil, L1, is adjusted for resonance, then L3 is set for lowest SWR. L3 can be mounted inside the trunk or under the bumper in a weatherproof enclosure.

For fixed-station operation it is recommended that a good ground system be employed. One should use at least 10 radials of, say, no. 18 or larger wire, 10 to 50 feet (3.05 to 15.24 m) in length. If you can manage 40 radials, 60 feet (18.29 m) in length, all the better.[1]

Concluding Comments

Three fixed-station versions of the Minooka Special have been tested and used at W9UCW. Each was compared against the regular antenna, which is a one-quarter wavelength vertical (130 feet [39.62 m] high), and operates against a radial system that contains 12,000 feet (3657.6 m) of wire. The short verticals were always inferior to the big antenna by approximately 5 dB.

Using version no. 4 from the table (4-foot [1.22-m] base section), the author worked W1s through W6s from the mobile setup. The rig was an HW-18 transceiver. Contacts were also made with VP9 and KV4 stations from the car.

While DXing from San Andres (HK0) reports were received of signal strengths exceeding S9 plus 20 dB from Maine to Washington. Good reports were received from Europe also.

Because the loading coil acts as an rf choke at 3.5 MHz and above, several versions of the antenna have been used successfully from 160 through 20 meters with an appropriate L-network matching section installed at the base of the system.

Acknowledgements

The author wishes to express his gratitude to all who helped make this article possible, . . . particularly K9SKX, WA9EYY, and W9YYF, who assisted with the initial experiments on the Minooka Special. The participants in the antenna survey (a list too large to include here) deserve considerable thanks for their help as well. This material was originally presented in *QST* by Barry A. Boothe, W9UCW.

---

[1] [Editor's Note: The reader should be aware that 1/4-wavelength radials (or longer ones) are recommended for best efficiency in a system of this kind. Detailed information on the subject is given in *The ARRL Antenna Book*, 13th Edition. The recommendations given by the author are no doubt based on the lack of space available in the typical urban lot, around which his design was established. See Hills, "The Grounds Beneath Us." *RSGB Bulletin*, June 1966.]

Inexpensive 5/8-Wave Vertical Antenna with Coaxial Transformer

The author's hastily constructed 5/8-wave vertical antenna (inspired by W0JF's results[1]) consists of a supporting structure made from sections of 2 × 2-inch (51 × 51-mm) lumber on which four lengths of 300-ohm TV ribbon are affixed. Each section of ribbon line is 41 feet 6 inches (12.65 m) long. A length of this line is attached to each side surface of the 2 × 2 support mast, then the ends (top and bottom) of all four line sections are connected in parallel to form a single vertical conductor. The coax-line dimensions are near those of W2JTJ,[2] and are shown in Fig. 37. The SWR varies from a flicker of reflected power at 14,000 kHz to 1.5:1 at 14,350 kHz. (The next task is to get that lowest-SWR point shifted to 14,275 kHz.) The whole thing is leaning against a tree, almost vertical, and seems to work well. This material was originally presented in *QST* by Robert J. Earl, W1DRV.

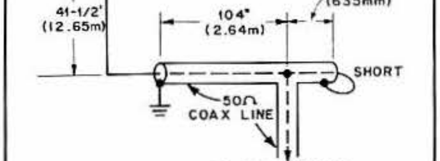

Fig. 37 — Coax-line transformer for 20-meter 5/8-wave vertical antenna.

References

[1] Beers, "Short Antennas for the Lower Frequencies," *QST*, August and September 1970.

[2] Czerwinski, "Coaxial Transformer for Voltage-Fed Antennas," *QST*, June 1961.

A Backyard 160-Meter Vertical

A 160-meter band was the favorite one at the author's QTH this winter. The same could be said for many other amateurs and the simple antenna described here may be of interest.

Some of the old-timers will tell the newcomer to 160 meters that the best antenna for local and DX work is the vertical. Since a full-size vertical would be very large, the question is finding a type that will fit into a closely packed residential area.

As a starter, the writer took an old 40-foot (12.19-m) telescoping mast and insulated it from the ground. (This can be done by setting it on top of a large soft-drink bottle, or clamping it to a wooden post.) The antenna was fed at the base with RG-58A/U 52-ohm coaxial cable. The shield was wired to a radial system and a ground rod. A hint for the radials — buy some aluminum clothesline and cut a number of slits in your yard with a flat spade. Next, push the wire into the slits and tamp the grass back together. If there are any swampy parts of the yard, these seem to make the best ground areas.

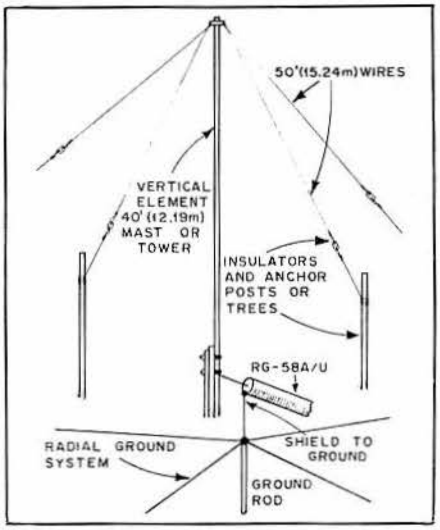
50'(15.24m)WIRES
VERTICAL ELEMENT 40' (12.19m) MAST OR TOWER
INSULATORS AND ANCHOR POSTS OR TREES
RG-58A/U
RADIAL GROUND SYSTEM
SHIELD TO GROUND
GROUND ROD

Fig. 38 — A space-saving vertical for 160 meters. While not always possible, it is desirable to have the wires for the top hat run off at right angles to the mast.

Rather than using a loading network at the base of the antenna, the author constructed a capacitive hat to load it at the top instead. The top hat was made from four 50-foot (15.24-m) pieces of wire strung out from the top of the mast in guy-wire fashion. They are connected electrically to the top, and stretched until they are as close as possible to a 90-degree angle with respect to the mast.

The system described works well with a 40-foot (12.19-m) mast, but other lengths could be used instead. However, the top hat may have to be adjusted in order to bring the system to resonance. While my vertical occupies very little space, it seems to perform as well as do many "full-size" antennas on 160 meters. This material was originally presented in *QST* by Mike Mussler, WB8JJA.

Chapter 2

Yagi Antennas

The antenna systems treated in this chapter are variations of the basic directional gain antenna devised many years ago by Yagi and Uda. Over the years the fundamental design and its variations have been referred to simply as the Yagi. How Mr. Uda was excluded from the credits remains a mystery, as he was a co-developer of the antenna under discussion.

Yagi types of antennas appeal to all manner of hf- vhf- and uhf-band operators. The primary response characteristic of the Yagi is unidirectional, discounting minor response lobes off the front, sides and rear of the antenna. The major lobe concentrates the radiated energy in a desired direction to provide gain in dB and discrimination against interference from unwanted directions.

Yagis can be used singly or in array combinations. In the latter case, each time the size of an array is doubled the system gain can be made to increase by 3 dB — assuming that the phasing, impedance matching, and aperture parameters are correct. Thus, a pair of 3-element Yagis which are "stacked" in an array to meet the foregoing criteria will exhibit 3 dB more gain than a single 3-element Yagi. To effect another 3 dB increase in gain, four 3-element Yagis would need to be placed in a given array of Yagis, and so on.

This chapter deals with the cream of the Yagi crop, so to speak, as taken from past issues of *QST*. When zoning laws and available property permit, the Yagi is preferred over the single-element vertical antennas described in chapter 1.

The 20-Meter DX Weasel

Short antennas have always captured the interest of amateurs.

The antenna presented here is a monoband version of the Bite-Size Beam for 20 and 15 meters published elsewhere in this chapter. It is designed for 20-meter operation only. Most of the parts were "lifted" from the original array.

The reader is cautioned against attempting to figure why some of the dimensions were selected. As a point of interest, they just happened that way. One new twist was added to this system — capacitance hats. The hats provided one noticeable improvement in performance which became obvious during the initial tune-up phase of the project. The hats lowered the Q of the elements and hence broadened the overall response of the system. While only conjecture, one might say that it is this broad tuning which causes the parasitic elements to perform as they do. Element tuning on the Bite-Size Beam[1] was critical. If the dimensions given here are followed carefully, the antenna should provide excellent characteristics for gain and pattern. Front-to-back measurements indicate approximately 20 dB. The front-to-side ratio was measured in excess of 25 dB. There are no objectionable side lobes. The SWR is less than 2:1 at each end of the band when the gamma capacitor and rod are set for proper operation at 14.1 MHz. There is no need to select phone or cw as the primary mode of operation.

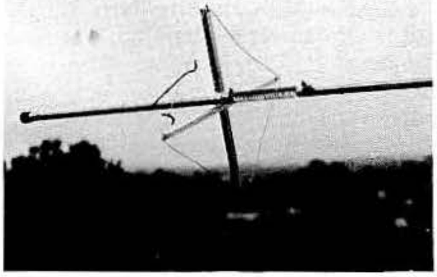

Details of the loading coil and capacitance hat on one of the elements.

Construction

The boom and all the elements are made from 1-1/4-inch (32-mm) diameter aluminum tubing available at most hardware stores. The two boom sections and the two pieces which make up the center portion of the driven element are coupled using 15-inch (0.38-m) sleeves of 1-3/8-inch (35-mm) OD aluminum tubing. Sheet-metal screws should be used to secure the sections within the coupling sleeves.

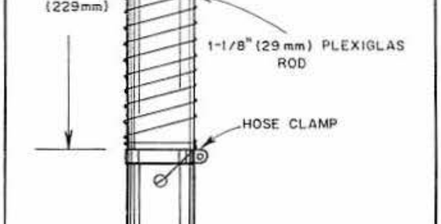

Fig. 1 — Details for joining sections of aluminum tubing and Plexiglas rod.

The loading coils are wound on 1-1/8-inch (29-mm) diameter Plexiglas rod. Details are shown in Fig. 1. Be sure to slit the ends of the aluminum tubing where the compression clamps are placed. The coils are made from no. 14 enameled wire. The specified number of turns are equally spaced to cover the entire 9 inches (229 mm) of Plexiglas.

The capacitance hats are constructed from 3/4-inch (19-mm) angle aluminum. Two pieces 2 feet (0.61 m) in length are required for each hat. The model shown in the diagrams has the angle aluminum

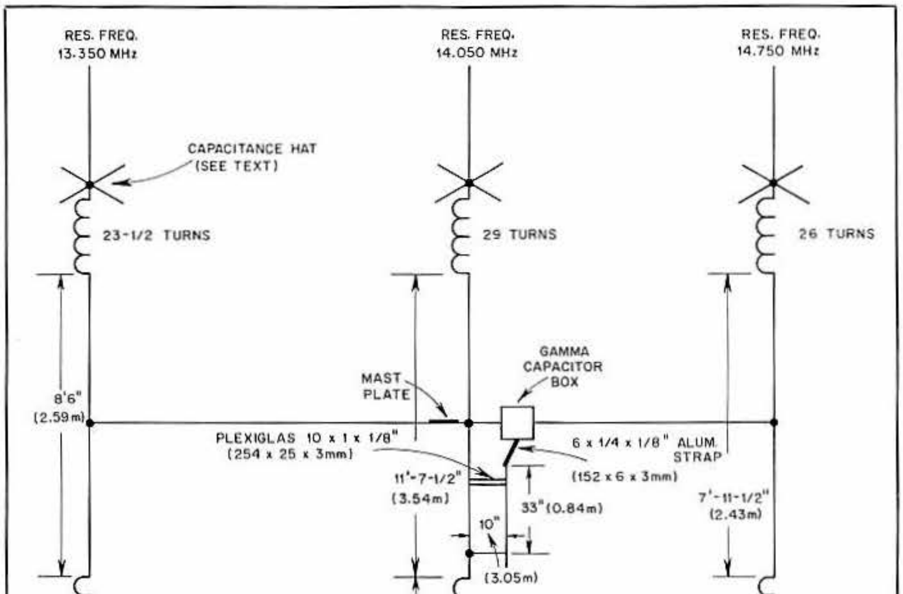

Fig. 2 — Constructional details for the 20-meter beam. The coils on each side of the element are identical. The gamma capacitor is a 140-pF variable unit manufactured by E. F. Johnson Co.

Table 1

Complete Parts List for the Short Beam

Qty	*Material*
2	10-foot (3.05-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (one for the reflector center section, one for the reflector end sections).
3	8-foot (2.44-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (two lengths for the boom, one length for the director element center).
4	6-foot (1.83-m) lengths of 1-1/4-inch (32-mm) dia. aluminum tubing (two lengths for the driven element center, two lengths for the director and driven element ends).
2	15-inch (0.38-m) lengths of 1-5/8-inch (41-mm) dia. aluminum tubing.
1	40-inch (12.19-m) length of 1-3/8-inch (10-mm) dia. aluminum tubing.
4	6-foot (1.83-m) lengths of 3/4-inch (19-mm) angle aluminum.
6	12-inch (0.31-m) lengths of 1-1/8-inch (29-mm) dia. Plexiglas rod.
1	9-inch (229-mm) square, 1/4-inch (6-mm) thick aluminum plate.
8	U bolts.
12	Compression hose clamps.
8	Crutch caps.
38 ft	No. 12 enameled copper wire.
60 ft	No. 14 enameled copper wire.

fastened to the element using aluminum strips and no. 8 sheet-metal screws, but sheet-metal screws alone can be used. Solder lugs are fastened to the ends of the angle aluminum and no. 12 or 14 wire connects the ends of the aluminum, resulting in a square loop. The wires should be soldered at each of the solder lugs.

All of the elements are secured to the boom with TV U-bolt hardware. Plated bolts are desirable to prevent rust from forming. An aluminum plate 9 inches (229 mm) square by 1/4-inch (6-mm) thick was used as the boom-to-mast plate.

A boom strut is recommended because the weight of the elements is sufficient to cause the boom to sag. A 1/8-inch (3-mm) diameter nylon line is plenty strong. A U-bolt clamp is placed on the mast several feet above the antenna and provides the attachment point for the center of the truss line. To reduce the possibility of water accumulating in the element tubing and subsequently freezing, crutch caps are placed over the ends. Rubber feet suitable for keeping furniture from scratching hardwood floors would serve the same purpose.

A piece of Plexiglas was mounted inside an aluminum Minibox to provide support and insulation for the gamma capacitor. A plastic refrigerator box would serve the purpose just as well. The capacitor housing is mounted to the boom by means of a U bolt. The gamma rod is made of 3/8-inch (10-mm) aluminum 40 inches (1.016 m) long and is connected to the gamma capacitor by a 6-inch (152-mm) length of strap aluminum.

Tune-Up and Operation

The builder is encouraged to follow the dimensions given in Fig. 2 as a starting point for the position of the gamma rod shorting strap. Connect the coaxial cable and install the antenna near or at the top of the tower. The gamma capacitor should be adjusted for minimum SWR at 14.1 MHz as indicated by an SWR meter (or power meter) connected in the feed line at the gamma capacitor box. If a perfect match cannot be obtained, a slight repositioning of the gamma short might be required.

Evaluation of performance is always a difficult task. The 20-Meter DX Weasel is mounted on top of a 50-foot (15.24-m) high tower at the QTH of W1FB. The transmitter input power for the testing period was 3 watts. Many European stations as well as a ZL have been contacted with ease. In fact, the Weasel has been responsible for cracking a few pileups on the first call!

Obtaining the Material

Table 1 lists all of the major parts used in the 20-meter Weasel. Most of the U-bolt hardware as well as the aluminum tubing should be available from large hardware stores. The Plexiglas rod, however, presents a bit more of a problem as indicated by mail received by the ARRL Technical Department. Many large plastic supply dealers stock the cast rod. The problem is that there are not very many large dealers! Additionally, Plexiglas is a petroleum-base product which means the cost is no doubt higher than one would expect. Those amateurs who are unable to locate aluminum hardware or Plexiglas should consult the yellow pages to see who their nearest supplier is. This material was originally presented in *QST* by Robert Myers, W1XT and Jay Rusgrove, W1VD.

---

[1] See page 45.

The HW-40 Micro Beam

A helically wound 2-element 40-meter Yagi.

Disconsolate pleas from apartment dwellers and other amateurs who live on property that is too small in area to permit the use of large 40-meter antennas prompted the authors to develop the antenna described here. The particular dimensions represented in this design were dictated primarily by the kind of fiberglass tubing that was immediately available. Kirk quad spreaders were obtained for use as forms on which to wind the helical elements. Longer (or shorter) elements can be used with good results. This article is intended as a guide in the design of similar antennas.

The HW-40 is scaled down to 28 percent of full size. Two elements are used, each being 18 feet (5.49 m) in length, tip-to-tip. Thus, the elements are just 2 feet (0.61 m) longer than those of a full-size 10-meter Yagi. A 16-foot (4.88-m)-long Reynolds aluminum-tubing boom (two 8-foot [2.44-m] lengths joined) provides 0.12-wavelength spacing between the driven element and reflector. The weight of the array is approximately 25 pounds, making it practical to use a TV-type antenna rotator with the system.

Design Considerations

Some may argue that helically wound antennas will tend to radiate off the ends of the elements (axial mode), and that the condition will prevent the Yagi from exhibiting normal front-to-side and front-to-back characteristics. It was established by Kraus,[1] that axial-mode radiation occurs only when a helix is one wavelength or greater in circumference. Generally, when the circumference is less than two-thirds wavelength, a nearly sinusoidal type of current distribution exists. This is effected by alternate reinforcement and cancellation of the two traveling waves which are directed oppositely. Each is of nearly identical current amplitude (I_O), hence the T_O transmission mode results for both traveling waves on the helix. Because of the foregoing condition the helices of this antenna radiate in the *normal mode* and will exhibit linear polarization characteristics. The diameter of these helices is merely 0.0009469 wavelength, eliminating any opportunity for axial-mode radiation. Front-to-side and front-to-back checks of the beam indicate, indeed, that normal-mode radiation is occurring. A number of relative tests were performed locally at distances up to 20 miles, and the front-to-back characteristics indicate a 10-dB figure. Front-to-side checks confirm a figure of approximately 28 dB. The effect is not so pronounced when receiving high-angle incoming signals. Depending upon the time of day, distance involved, and other propagation factors, various directivity traits will be observed.

The HW-40 beam, just 28 percent of full-size, is light enough to require only a TV-rotator.

Helically wound elements were chosen in preference to lumped-inductance elements in order to obtain linear voltage and current distribution across the elements, and to make the matching technique less difficult. Even if end-loaded elements with capacitance hats were used, the 18-foot (5.49-m) dimension would make it impossible to use. a gamma match of conventional design. Because of the very low value of radiation resistance of short loaded antennas, other conventional matching schemes would not lend ease to matching a 50-ohm coaxial feeder to the system.

There are certain penalties one must accept when using physically shortened antennas of this kind, and an in-depth treatment of the subject was given by Sevick in *QST.*[2] It was a fortunate circumstance in this case that the feed-point impedance of the beam turned out to be approximately 12 ohms. Therefore, a 4:1 toroidal balun was installed at the feed point to provide a near-perfect match to the 50-ohm transmission line. In addition to encountering low values of feed impedance, the bandwidth of loaded elements is greatly less than with full-size antennas. This beam was tuned for operation at a center frequency of 7050 kHz. It works well from 7025 to 7075 kHz, with an SWR no greater than 2:1. It is flat at 7050 kHz. Those wishing to duplicate this Yagi should be willing to accept this design trade-off in the interest of having an effective directional antenna. It is recommended, therefore, that it be tuned for one's favorite part of the 40-meter band. The use of a Transmatch will, however, permit excursions into portions of the band where the SWR becomes greater than 1:1.

Construction Details

The construction of the 40-meter beam is very simple and requires no special tools or hardware. Two fiberglass 15-meter quad arm spreaders are mounted on an aluminum plate with U bolts, as shown in the photograph. A wooden dowel is inserted approximately 6 inches (152 mm) in the end of each fiberglass arm to prevent the U bolts from crushing the poles. The aluminum mounting plate is equipped with U-bolt hardware for attachment to the 1-1/4-inch (32-mm) diameter boom.

A plastic refrigerator box is mounted on each element support plate and is used to house a Miniductor coil. No. 14 copper wire is used for the elements. The wire is wound directly on the fiberglass poles at a density of *40 turns per foot* (not turns per inch!) for a total of 360 evenly spaced turns. The wire is attached at each end with an automotive hose clamp of the proper size to fit the fiberglass rods. Since the fiberglass is tapered, care must be taken to keep the turns from sliding in the direction of the tips. Several pieces of plastic electrical tape were wrapped around the pole and wire at intervals of 1 foot (0.31 m). All of the element half sections are identical in terms of wire and pitch. Coil dimensions and type are given in Fig. 3.

The driven-element matching system consists of a 4:1 balun transformer and a link tightly coupled to the Miniductor. Complete details are given in Fig. 4.

Mounted at the end of each element (held in place by the hose clamp) is a short section of stiff copper wire used to allow for final tuning of the system as well as to broaden the frequency response of the antenna. Since the overall antenna is very small in relation to a full-sized array, the SWR points of 2:1 are rather close to each other, as mentioned earlier. The antenna shown in the photograph provides an SWR of less than 2:1 within about 25 kHz either side of resonance. Tuning the antenna for phone-band operation should not be difficult and the procedure outlined below should be suitable.

Tuning

The parasitic element is adjusted to be about four percent lower in frequency than

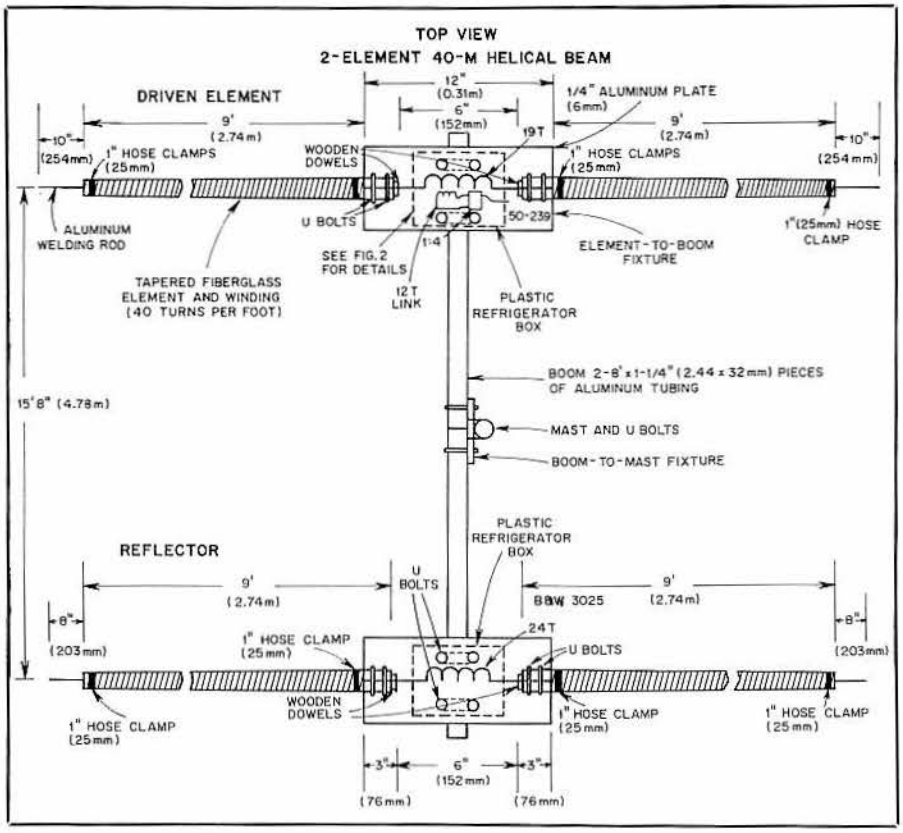
TOP VIEW
2-ELEMENT 40-M HELICAL BEAM
DRIVEN ELEMENT
12" (0.31m)
6" (152mm)
1/4" ALUMINUM PLATE (6mm)
9' (2.74m)
9' (2.74m)
10" (254mm)
10" (254mm)
1" HOSE CLAMPS (25mm)
WOODEN DOWELS
19T
1" HOSE CLAMPS (25mm)
ALUMINUM WELDING ROD
U BOLTS
L
50-239
1:4
ELEMENT-TO-BOOM FIXTURE
1"(25mm) HOSE CLAMP
SEE FIG. 2 FOR DETAILS
TAPERED FIBERGLASS ELEMENT AND WINDING (40 TURNS PER FOOT)
12T LINK
PLASTIC REFRIGERATOR BOX
BOOM 2-8'x1-1/4" (2.44 x 32mm) PIECES OF ALUMINUM TUBING
15'8" (4.78m)
MAST AND U BOLTS
BOOM-TO-MAST FIXTURE
PLASTIC REFRIGERATOR BOX
REFLECTOR
9' (2.74m)
U BOLTS
B&W 3025
9' (2.74m)
8" (203mm)
8" (203mm)
1" HOSE CLAMP (25mm)
24T
U BOLTS
WOODEN DOWELS
1" HOSE CLAMP (25mm)
1" HOSE CLAMP (25mm)
1" HOSE CLAMP (25mm)
3" (76mm)
6" (152mm)
3" (76mm)

Fig. 3 — Overall dimensions for the 40-meter short beam. The boom consists of two pieces of standard 1-1/4-inch (32-mm) dia do-it-yourself aluminum tubing.

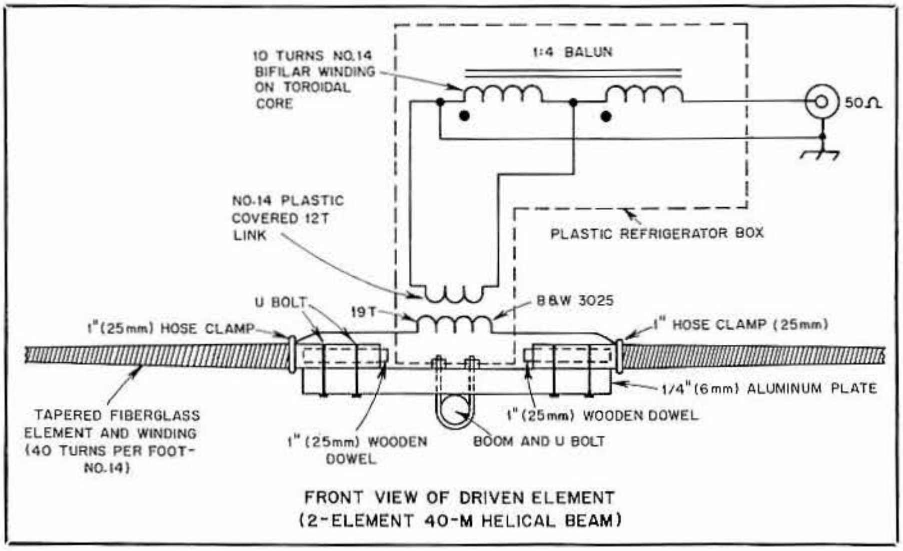
10 TURNS NO.14 BIFILAR WINDING ON TOROIDAL CORE
1:4 BALUN
50Ω
NO.14 PLASTIC COVERED 12T LINK
PLASTIC REFRIGERATOR BOX
U BOLT
19T
B&W 3025
1"(25mm) HOSE CLAMP
1" HOSE CLAMP (25mm)
1/4"(6mm) ALUMINUM PLATE
TAPERED FIBERGLASS ELEMENT AND WINDING (40 TURNS PER FOOT-NO.14)
1" (25mm) WOODEN DOWEL
BOOM AND U BOLT
1" (25mm) WOODEN DOWEL
FRONT VIEW OF DRIVEN ELEMENT
(2-ELEMENT 40-M HELICAL BEAM)

Fig. 4 — Schematic diagram of the balun assembly mounted inside the plastic utility box. The core is a single T-200-2 Amidon. The 12-turn link is wound directly over the 19-turn Miniductor.

the driven element. A grid-dip oscillator was coupled to the center loading coil and the element tips were trimmed (a quarter of an inch [6 mm] at a time!) until the GDO indicated resonance to be at 6.678 MHz. For phone-band use, the ends could be snipped for 6.840 MHz. Adjusting the driven element is simple. Place an SWR meter or power meter at the input connector and cut the end wires (or add some if necessary) to obtain a proper match between line and antenna.

Performance Characteristics

Once the initial near-ground tuning of the elements was completed, the HW-40 was carried aloft and mounted on the 40-foot (12.19-m) tower at W1FB. Final touching up of the matching and element resonances was effected after the beam was in place on the tower. It is important to realize that shortened antennas are sensitive to changes in environment because of their narrow-band characteristics. Thus, height above ground and distance from nearby objects will cause some change in resonance of the elements. The resonant frequency of the HW-40 changed some 150 kHz between tuning at several feet above ground to final tweaking atop the tower.

It should be noted that the 40-meter beam has yet to be tested at an ideal height above ground. A distance of 40 feet (12.19 m) above the earth is not much more than one-quarter wavelength at 7 MHz, though performance thus far has been better than one might expect when using a beam or dipole less than one-half wavelength above ground.

Comparisons were made between the beam and three types of wire antennas . . . an inverted V, a 3/4-wavelength end-fed wire, and a full-wave 40-meter loop (with its theoretical 2-dB gain).[3] The results have been interesting, though at times rather confusing. The confusion results from the time-of-day/propagation syndrome mentioned earlier. At no time, however, did the inverted-V or 3/4-wavelength antennas surpass the performance of the HW-40. The beam was always as good as or better than the other two radiators, regardless of band conditions or time. During contacts with European or South American DX stations, the beam was always the best antenna, usually by two S units or more. During daytime contacts out to, say, 1000 miles, the wire antennas equalled the beam at times.

When comparing the beam to the full-wave loop, results became less easy to evaluate. That is, the antennas took turns being best. Frequently, the loop outperformed the beam (in the favored direction of the loop) both on short-haul and DX contacts. At other times the beam had the edge by a couple of S units. At this time it is impossible to say which antenna would provide the best all-year, all-around service. But, it has been observed that the beam is far less subject to pickup of man-made and atmospheric noise because of its narrow-band characteristics (the loop exhibits the opposite bandwidth trait— broad). Reduction in noise during reception can be a blessing, especially when working with the weaker signals! Another beneficial effect of the narrow-band feature is that strong, near-frequency signals do not so severely affect the receiver with respect to overloading and cross modulation, as is the situation with other types of antennas. The proof of this came when trying to operate while W1AW (two blocks away from W1FB) was transmitting code practice. Until the beam was erected it was impossible to use the 40-meter band, but with the beam connected to the receiver, and with the antenna pointed away from W1AW, it became possible to operate within 10 kHz of W1AW without knowing that station was on the air! So, if some kW operator lives down the street from you, this little beam could put you back in business.

While running approximately 80 watts output, the short beam has been extremely effective in working DX. DX stations now answer our CQs, whereas with the wire antennas (loop excepted) this seldom happened. Signal reports from DX stations average 579, with many 589 and 599 reports. During QRP operation with the beam (2 watts output), a number of DX stations have been raised and worked. The signal reports are lower of course, but RST 559 and 569 indicates that something good must be happening! One unrelated event occurred while testing the beam and loop with the QRP setup: a station in the 4th call area called CQ and was answered. He came back and said, "No QRP lids, CQ kWs only." It's unlikely that the HW-40 beam can be blamed for that, especially since we went on to work five European stations later that day with the 2-watt equipment!

Acknowledgements

The authors wish to express gratitude to those who helped construct, test and erect the 40-meter beam. Special thanks to K1JX, who "clomped" up and down the ARRL test tower and the W1FB station tower numerous times with test gear (and the beam) until the final design goals were met. Thanks also to W1NPG (SK) who labored silently while winding the helices, and subsequently endured the unbearable Connecticut heat-wave temperatures of August, while putting two coats of exterior spar varnish on the completed elements. The element dimensions and turns information were developed by K1TD to test a newly developed calculator program. This material was originally presented in *QST* by Robert Myers, W1XT and Doug DeMaw, W1FB.

References

[1] Kraus, *Antennas*, First Edition, McGraw-Hill Book Co., Chapter 7.

[2] Sevick, "The Ground-Mounted Short Vertical," *QST*, March 1973, p. 13. (See Chapter 1.)

[3] DeMaw, "The Novi-Loop," *QST*, October 1973. (See Chapter 4.)

A Bite-Size Beam for 20 and 15 Meters

Do you want to work DX from your space-restricted QTH with increased effectiveness? If the answer is yes, perhaps this two-band Yagi system is just the one for you.

There is no doubt that a hundred-foot (30.48-m) tower with stacked monobanders for 20 and 15 meters is one way to generate an effective signal in the pileups. But what are the characteristics of this big system which make it perform better than a dipole strung between two trees? First, the tall tower places the antenna up in the clear, away from most of the surrounding ground clutter. Even a relatively short tower (40-foot [12.2-m] variety) provides this advantage in many instances. The ability to point the major radiation lobe of the antenna in the correct direction is another reason the high monobanders outperform the kinky wire. But if the antenna is capable of supporting its own elements, the directional characteristics can be used to good advantage whether the beam is mounted at 100 feet (30.5 m), or at 40 feet (12.2 m). And without any question, the ability to reject interference aids substantially in the reception of weak signals.

A typical one-foot (0.31-m)-per-side triangle tower is self-supporting (when properly installed) to 40 feet in height and requires less than one square foot of property. For many amateurs, the major limiting factor is the inability to turn a monoband beam (or even a large tribander) without hitting some object like a tall tree. Hesitation on behalf of the amateur to place what might be considered by his neighbors an offensive amount of hardware in the sky near his home is another reason beams are ruled out. The 2-element-per-band, 20- and 15-meter, interlaced Yagi presented here, overcomes these problems. It is lightweight and is less than 16-feet (4.88-m) square. And it's probably no more obtrusive than a large TV antenna.

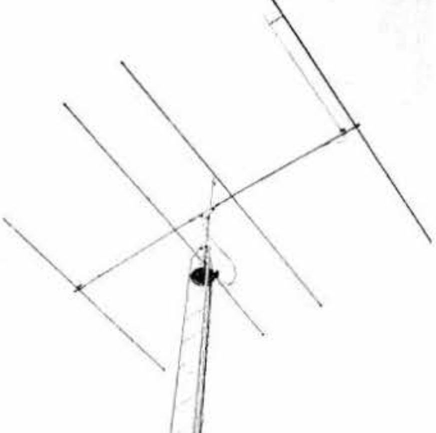

So What if It's Short!

A popular misconception among amateurs is that any element short of full size is no good in an antenna system. Reducing the size of an antenna by 50 percent does lower the efficiency by a decibel or two, but the gain capability of a parasitic array outweighs this small loss in efficiency. Mounting the antenna above the interference-generating neighborhood can greatly reduce susceptibility to man-made noise and certainly aids in the reduction of rf heating to trees, telephone poles, and buildings. It has been said that a brick wall, or dense foliage, can attenuate a signal passed through it by as much as 20 dB. Placing the antenna above these energy-absorbing objects is very desirable.

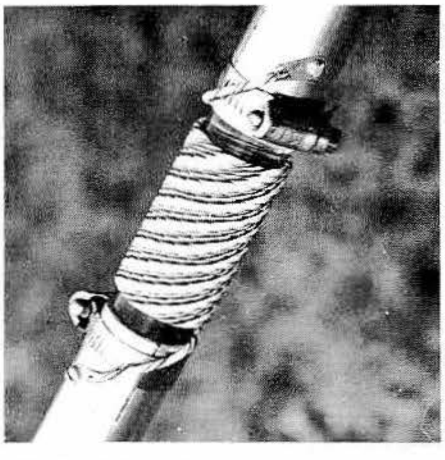

The coils are wound on Plexiglas rod. Compression clamps are used to hold the Plexiglas in position Sheet-metal screws and solder lugs provide attachment points for the ends of the loading coils.

View of the short beam from the front yard at W1FB.

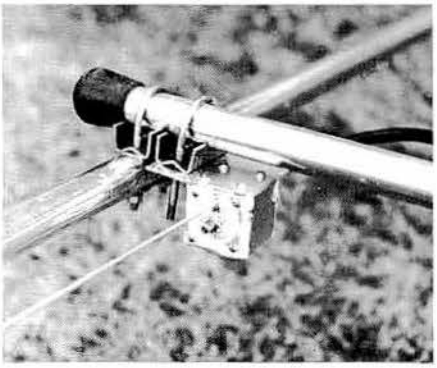

The gamma assembly is held in place by means of a small U bolt. The capacitors are mounted on etched circuit board.

It is a generally accepted fact that traps are difficult to build and adjust, as well as their being lossy. Loading coils, on the other hand, are easy to wind and require essentially no adjustment. If the wire size is large, losses are not a major factor. W2FMI lauded the merits of short antennas in an earlier article explaining that in reality, short (50 percent) elements do not materially depreciate the gain.[1] With these ideas in mind, the authors designed and built a shortened-element, but wide spaced, director array for 20 meters interlaced with a similar 15-meter reflector configuration also having optimum spacing for maximum gain. By placing the 15-meter driven element in front of the 20-meter array, the longer 20-meter elements tend to act as reflectors on 15 meters. There are no unwanted reflectors "in front" nor any directors "behind" the active array — elements which could cause pattern distortion and poor front-to-back ratio.

Construction

The dual-band beam has four elements, the longest of which is 16 feet (4.88 m). All of the elements and the boom are made from 1-1/4-inch (32-mm) diameter aluminum tubing available at most hardware stores. Element sections and boom pieces are joined together by slotting a 10-inch (254-mm) length of 1-1/4-inch (32-mm) tubing with a nibbling tool and compressing it for a snug fit inside the element and boom tubing. Coupling details are shown in Fig. 6.

The loading coils are wound on 1-1/8-inch (29-mm) diameter Plexiglas rod. The rod slips into the element tubing and is held in place with compression clamps. Be sure to slit the end of the aluminum where the compression clamps are placed. The model shown in the photographs has coils made of surplus Teflon-insulated miniature audio coaxial cable with the shield braid and inner conductor shorted together. A suitable substitute would be no. 14 enameled copper wire wound to the same dimensions as those given in Fig. 5.

All of the elements are secured to the boom with common TV U-bolt hardware. Plated bolts are desirable to prevent rust from forming. A 1/4-inch (6-mm) thick boom-to-mast plate is constructed from a few pieces of sheet aluminum cut into 10-inch (254-mm) square sheets and held together with no. 8 hardware. Several cookie tins could be used if sheet aluminum is not available. One local amateur used a plate from a large electrical box as a boom-to-mast bracket. Since it is galvanized, it is quite resistant to the harsh New England winters.

A boom strut (sometimes called a truss) is recommended because the weight of the elements is sufficient to cause the boom to sag a bit. A 1/8-inch (3-mm) diameter nylon line is plenty strong. A U-bolt clamp is

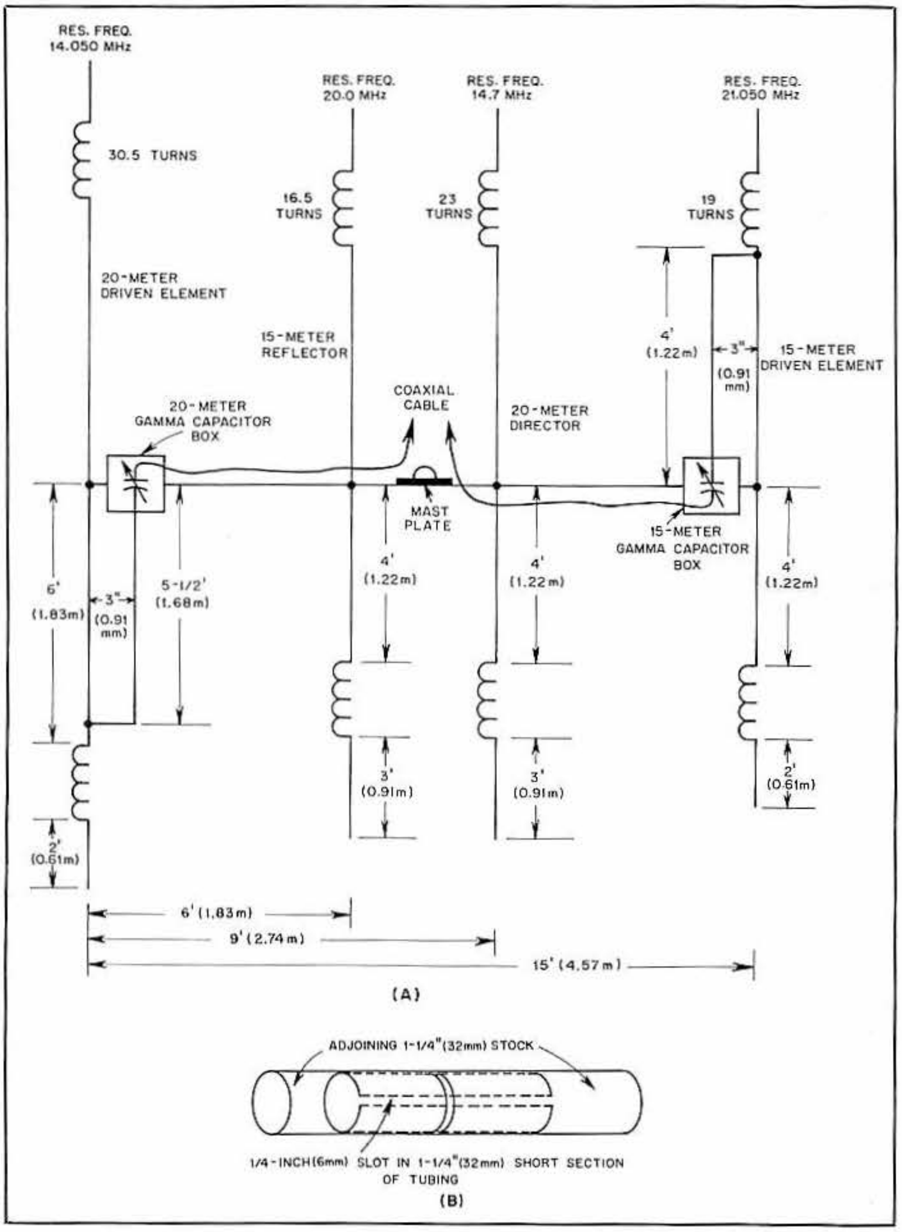
RES. FREQ. 14.050 MHz
RES. FREQ. 20.0 MHz
RES. FREQ. 14.7 MHz
RES. FREQ. 21.050 MHz
30.5 TURNS
16.5 TURNS
23 TURNS
19 TURNS
20-METER DRIVEN ELEMENT
15-METER REFLECTOR
4' (1.22m)
3" (0.91 mm)
15-METER DRIVEN ELEMENT
COAXIAL CABLE
20-METER GAMMA CAPACITOR BOX
20-METER DIRECTOR
MAST PLATE
15-METER GAMMA CAPACITOR BOX
6' (1.83m)
3" (0.91 mm)
5-1/2' (1.68m)
4' (1.22m)
4' (1.22m)
4' (1.22m)
3' (0.91m)
3' (0.91m)
2' (0.61m)
2' (0.61m)
6' (1.83m)
9' (2.74m)
15' (4.57m)
(A)
ADJOINING 1-1/4"(32mm) STOCK
1/4-INCH(6mm) SLOT IN 1-1/4"(32mm) SHORT SECTION OF TUBING
(B)

Fig. 5 — Constructional details for the 20- and 15-meter beam. The coils for each side of the element are identical. The gamma capacitors are each 140-pF variable units manufactured by E. F. Johnson Co. The capacitors are insulated from ground within the container. Since the design is one-half size for each band, the tuning is somewhat critical. The builder is encouraged to follow carefully the dimensions given in the text.

Table 2

Complete Parts List for the Short Beam

Qty	*Material*
9	8-foot (2.44-m) lengths of aluminum tubing, 1-1/4-in. (32 mm) dia
11	U bolts
2	Variable capacitors, 140 pF (E. F. Johnson)
4 ft	Plexiglas cast rod, 1-1/8-in. (29 mm) dia
16	Stainless steel hose clamps, 1-1/2-in. (38 mm) dia
1	Aluminum plate, 8-in. (203 mm) square
10 ft	Aluminum solid rod, 1/4-in. (6 mm) dia
2	Refrigerator boxes, 4 X 4 X 4 in. (102 X 102 X 102 mm)
25 ft	Nylon rope, 1/8-in. (3 mm) dia
16	No. 8 sheet metal screws
16	No. 8 solder lugs
8	Plastic (or rubber) end caps, 1-1/4-in. (32 mm) dia

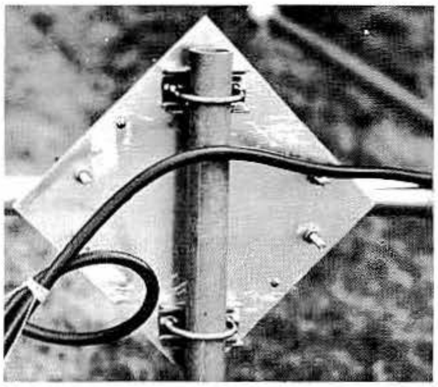

The boom-to-mast plate.

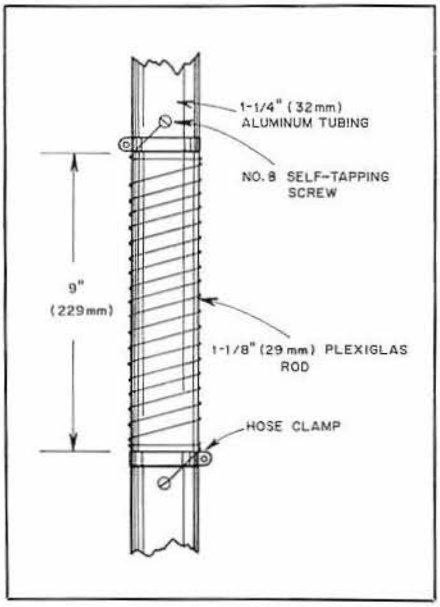

Fig. 6 — Coupling details for joining sections of aluminum tubing.

placed on the mast several feet above the antenna and provides the attachment point for the center of the truss line. To reduce the possibility of water accumulating in the element tubing and subsequently freezing (rupture may be the end result), crutch caps are placed over the element ends. Rubber feet suitable for keeping furniture from scratching hardwood floors would serve the same purpose. In fact, the rubber tips prevent the element ends from damaging surrounding objects during installation.

A heavy-duty steel mast should be used, such as a 1-inch (25-mm)-diameter galvanized water pipe. Steel TV mast is also acceptable. Any conventional TV-type antenna rotator should hold up under load conditions presented by this antenna. Nevertheless, certain precautions should be taken to assure continued trouble-free service. For instance, whenever possible, mount the rotator inside the tower and extend the mast through the tower top sleeve. This procedure relieves the rotator from having to handle lateral pressures during windy weather conditions. A thrust bearing is desirable to reduce downward forces on the rotator bearings.

Hookup and Operation

The monoband nature of the beam requires the use of two coaxial feed lines. The coaxial cable is attached to the 15-meter element (at the front of the beam) at the gamma-capacitor box. The other end of the cable is connected to a surplus 28-V dc single-pole coaxial switch.[2] The cable for the 20-meter element is connected in a similar fashion. The switch allows the use of a single feed line from the shack to a point just below the antenna where the switch is mounted. It is a simple matter to provide voltage to the switch for operation on one of the two bands. At the price of coaxial cable today, a double run of feed line represents a substantial investment and should be avoided if possible.

An etched-circuit board was mounted inside an aluminum Minibox to provide support and insulation for each of the gamma tuning capacitors. Plastic refrigerator boxes available from most department stores would serve just as well. The capacitor housing is mounted to the boom by means of U bolts.

The builder is encouraged to follow the dimensions given in Fig. 5 as a starting point for the position of the gamma rods and shorting bar. Placing the antenna near the top of the tower and then tilting it to allow the capacitors to be reached makes it possible to adjust the capacitors for minimum SWR as indicated by an SWR meter (or power meter) connected in the feed line at the relay. If the SWR cannot be reduced below some nominal figure of approximately 1.4:1, a slight repositioning of the gamma short might be required. The dimensions given are for operation at 14.050 MHz and 21.050 MHz. The SWR climbs above 2:1 about 50 kHz in either direction from the center frequency. Although tests were not conducted at more than 150 watts input to the transmitter there is no reason why the system would not operate correctly with a kilowatt of power supplied to it.

After many months of testing at the QTH of W1FB, several characteristics were noted. First, the antenna withstood several wind and ice storms common to Connecticut. Performance turned out to be what can be expected from a 2-element Yagi. Since the 20-meter portion of the antenna is a director array, the front-to-back ratio is a bit less than 10 dB. On 15 meters, where the system operates with a reflector parasitic element, the front-to-back ratio is considerably better — on the order of 15 dB.

This antenna was mounted atop a 40-foot (12.2-m) tower where previously there was connected a 20-meter sloping dipole pointed at Europe (a multielement array for 2 meters graced the top of the tower!). The number of European stations contacted increased rather dramatically with the installation of the beam. Except in pileups on rare DX stations, DX is now worked as a routine, even though W1FB is a QRP fan and typically runs less than 2 watts of output power.

No doubt the increased effectiveness of the signal was caused by several factors. The antenna certainly has gain (although at ARRL hq. we have no way of making accurate measurements) and is well above the nearby small trees. More importantly, the antenna is now far from the aluminum house siding which once was directly in the pattern of the 20-meter "sloper." Of course, the front-to-back characteristic has the tendency of making DX signals louder in relation to stateside stations and therefore makes receiving much easier.

If previously you've felt that a beam and tower were not possible at your QTH, perhaps re-evaluation of the situation is in order. This antenna, as shown in the photograph, does not appear offensive when viewed from the front yard. Yet the overall amount of property occupied (at ground level) is less than one square foot. And the turning radius is slightly over 11 feet (3.35 m)! This material was originally presented in *QST* by Robert Myers, W1XT and Clarke Greene, K1JX.

References

[1] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. (See Chapter 1.)

[2] Purchased from Fair Radio Sales, Lima, OH.

Linear Loaded 20-Meter Beam

Build this novel minibeam for 20 meters . . . it works!

Full-sized antennas are fine if you have the height and room for them. But much of the advantage is usually lost with average installations. So why bother with a sad-looking monster a few feet atop the house when a trim, smaller model will do the job. This article deals with a method that eliminates the messy and expensive loading coils as well.

Having built an HW-32A transceiver, the author's 20-meter activity was restricted to the phone band. A homemade vertical antenna did a fair job in the states, but not for DX. This writer's thoughts ran to a beam and the aluminum tubing that could be obtained for full-size construction. A 2-element beam was constructed, using a gamma match and information found in the *Handbook*. The beam worked electrically, but the mechanical construction left much to be desired. It flopped in the wind like a wounded butterfly. The author tried to correct this with a wooden framework, but it was just as bad. Something with shorter elements was desired, something that was rigid and would still perform well. But all of the author's research turned up only loading coils or other methods involving a loss compromise.

The author's aluminum supply included four 12-ft (3.66-m) lengths of 7/8-inch (22-mm) diameter tubing. This would give two elements of 24 ft (7.32-m) length and a minimum of sag. The problem was how to mount them and make up the difference in physical and electrical length for a half-wave length. Loading coils were ruled out. The possibility of folding wire under the elements seemed to have merit. (Later, this was found to be called linear loading.) This dictated a split element and standoff supports for the tubing.

Two 2-ft (0.61-m) lengths of aluminum channel were obtained for the element supports (Fig. 7). each 12-ft (3.66-m) section was attached using 1-in. (25-mm) ceramic standoff insulators. A 34-in. (19-mm) birch dowel was used to stiffen and prevent crushing of the tubing where the attachment was made. Element supports were grooved and attached to a 10-ft (3-m) TV mast section using a single 5/16-in. (8-mm) U bolt.

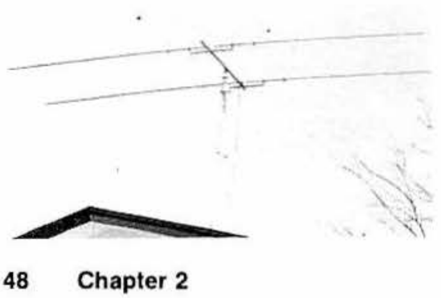

The folded additional loading wire was no. 12 copper house wire with the insulation stripped off, and was supported by two bakelite blocks on each element section. The blocks were anchored to the tubing by small sheet metal screws. Ends of the wire sections were formed into a loop to hold their position.

Tune-Up and Adjustments

My thinking in regard to coaxial-cable-fed dipoles is that they should use a balun or matching device of some kind. This beam, having a split driven element, seemed to call for a 1:1 balun since it was to be fed with RG-58/U (Fig. 8). Formulas found in handbooks for beam-element lengths did not seem to work for the linear-loading design, so cut-and-try was used many times. In final tests, a grid-dip meter was utilized with a calibrated receiver to determine the driven-element total length of 35 ft (10.67 m). A reflector with 10-ft (3.05-m) spacing was used.

It was found that adding a reflector of five-percent greater length lowered the frequency (with the beam near the ground) but on raising the beam to rooftop, the frequency had increased. The 10-ft spacing gave a 3:1 VSWR which was very difficult to reduce. Testing this arrangement on the air seemed to indicate more gain in the back direction than in the front. After many tests and much rebuilding of the wire lengths, the director element approach seemed to give the most encouraging results. The VSWR was down to 1.5:1 and the five-percent shorter director raised the frequency to the phone band. Spacing was reduced to 7 ft 6 in. (2.29 m).

More research brought out the possibility of inductive matching to get the VSWR

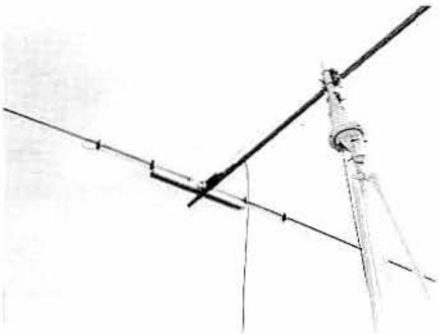

The beam atop the author's tower.

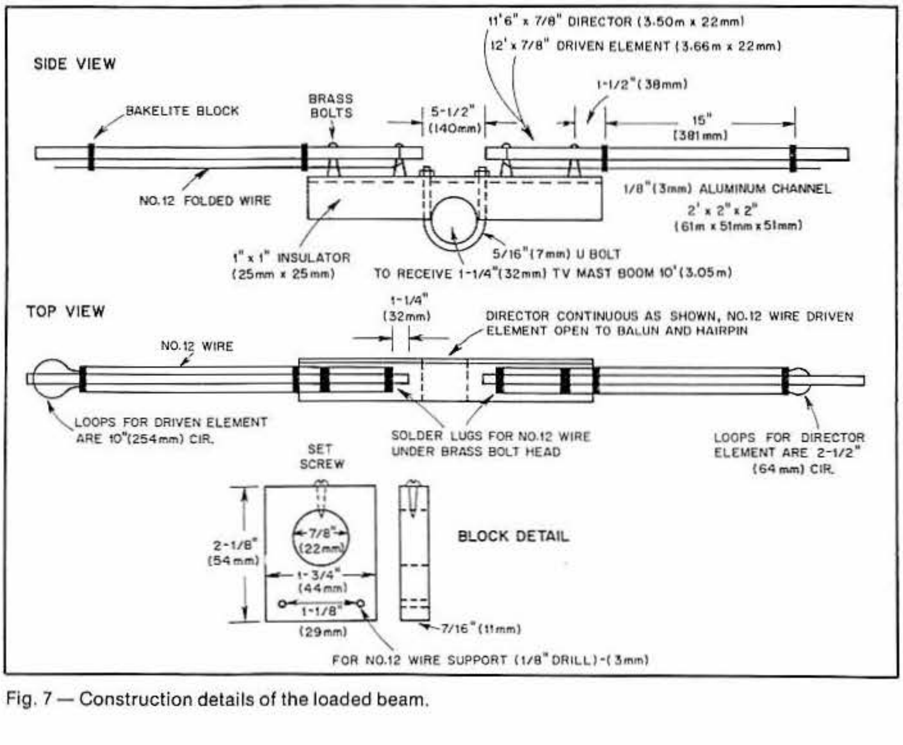

SIDE VIEW
11'6" x 7/8" DIRECTOR (3.50m x 22mm)
12' x 7/8" DRIVEN ELEMENT (3.66m x 22mm)
1-1/2" (38mm)
BAKELITE BLOCK
BRASS BOLTS
5-1/2" (140mm)
15" (381mm)
NO. 12 FOLDED WIRE
1/8" (3mm) ALUMINUM CHANNEL 2' x 2" x 2" (61m x 51mm x 51mm)
1" x 1" INSULATOR (25mm x 25mm)
5/16" (7mm) U BOLT
TO RECEIVE 1-1/4" (32mm) TV MAST BOOM 10' (3.05m)
TOP VIEW
1-1/4" (32mm)
DIRECTOR CONTINUOUS AS SHOWN, NO.12 WIRE DRIVEN ELEMENT OPEN TO BALUN AND HAIRPIN
NO. 12 WIRE
LOOPS FOR DRIVEN ELEMENT ARE 10" (254mm) CIR.
SOLDER LUGS FOR NO.12 WIRE UNDER BRASS BOLT HEAD
LOOPS FOR DIRECTOR ELEMENT ARE 2-1/2" (64 mm) CIR.
SET SCREW
2-1/8" (54mm)
7/8" (22mm)
1-3/4" (44mm)
1-1/8" (29mm)
BLOCK DETAIL
7/16" (11mm)
FOR NO.12 WIRE SUPPORT (1/8" DRILL)-(3mm)

Fig. 7 — Construction details of the loaded beam.

down to a more reasonable level. Many hairpin lengths and widths were tried with varying results. The final hairpin was 23-in. long by 1-1/2 in. (580 × 38 mm) wide with a slight fan shape to fit the balun connections. This reduced the VSWR at 14,200 kHz and 14,350 kHz with a flat response in between.

On-the-Air Tests

The help of other amateurs in tests for front-to back ratio was very helpful. Average reports have been 2.5 S units in the forward position over the back position. The author did not have a reference dipole for gain tests, but has heard and worked more DX than in the past 20 years. The beam is mounted on a rooftop 10-ft (3-m) steel tower 38 ft (11.6-m) above the ground. It weighs around 20 pounds and is rotated by an older-type TV rotator. This material was originally presented in *QST* by Cole Collinge, W0YNF.

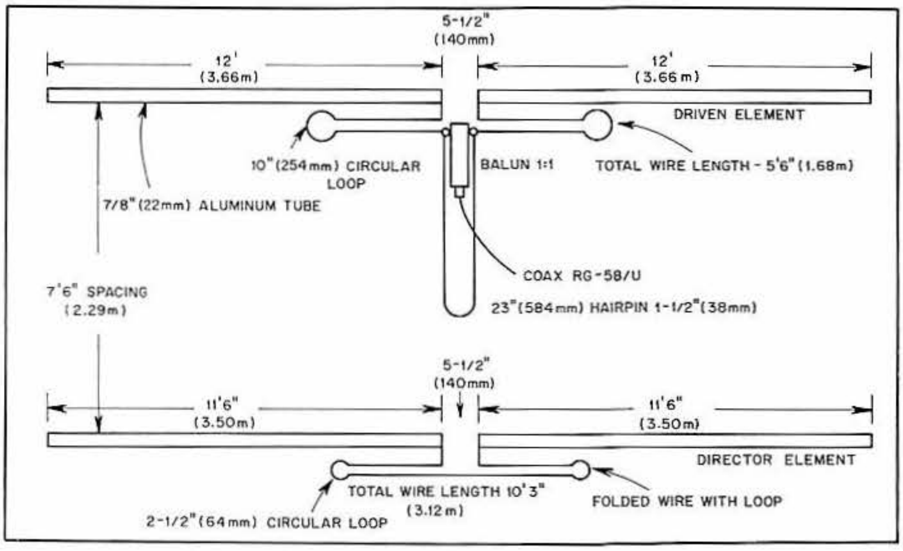

Fig. 8 — Critical electrical dimensions.

The Log-Yag Array

The Yagi antenna array has been around for years and years. A relative newcomer to hams is the log-periodic dipole array (LPDA), which offers nearly constant gain over a greater bandwidth than the Yagi. Guess what happens when you cross a Yagi with an LPDA . . . .

With the decline in sunspot activity, a number of amateurs have considered monoband Yagi arrays. The first problem encountered seems to be array length, that is, overall size for a desired gain and bandwidth. The Log-Yag principle, as will be discussed shortly, has produced a system which will provide the amateur with another alternative to the long-boom Yagi, stacked Yagis, or loop-antenna systems. The L-P Yagi (Log-Yag) array is not a new system; many such arrays have been designed and developed by Oliver Swan[1] and others.[2] This article, however, will provide the basic theory of operation, design procedure, and the construction of a practical antenna.

Theory of Operation

The Log-Yag array utilizes an LPDA-driven[3] group of elements, designed to cover a desired bandwidth, in conjunction with parasitic elements to achieve higher gains and greater directivity than would be realized with either the LPDA or Yagi array alone. The Yagi array requires a long boom and wide element spacing for wide bandwidth and high gain. This is because the Q of the Yagi system increases as the number of elements is increased and/or as the spacing between adjacent elements is decreased.[4,5] An increase in the Q of the Yagi array means that the total bandwidth of that array is decreased; optimum gain, front-to-back ratio, and side lobe rejection are obtainable only over small portions of the band. Dr. I. L. Morris, using a high-speed digital computer, has completed extensive research on 4-, 8-, and 10-director Yagi-Uda arrays.[6] His work is comprehensive and is recommended reading for all technically minded amateurs. The parameters varied in his study were element length, spacing, radius and number. As can be seen in Fig. 9, the forward gain and front-to-back ratio

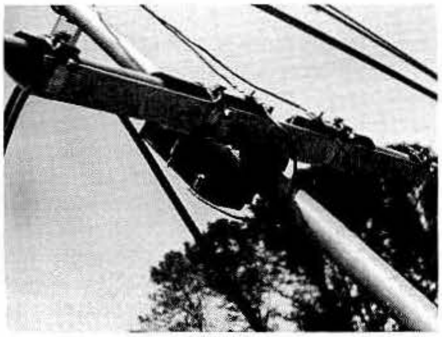

The attachment of the elements to the boom.

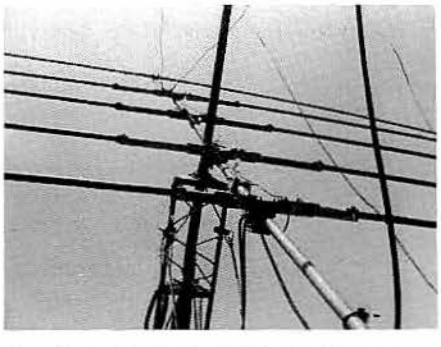

From the front to the back of the Log-Yag array. Note the truss provides lateral and vertical support.

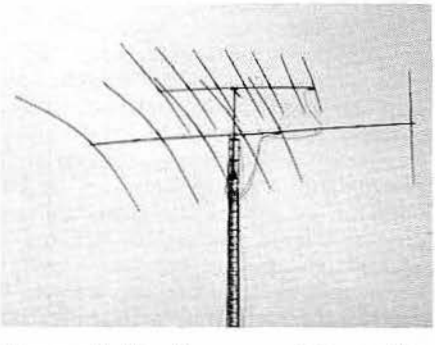

The completed Log-Yag array ready for use. The array mounted above the Log-Yag is a 7-element LPDA for 21 to 30 MHz.

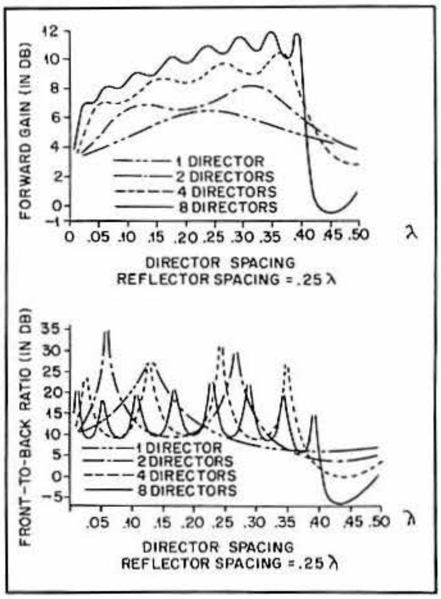

Fig. 9 — The effects of director spacing on various Yagi arrays. The change in gain and front-to-back ratio is plotted for the change in director spacing. The reflector spacing has been held constant.

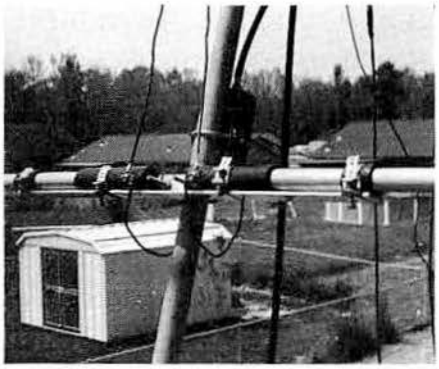

The connections between the balun and the input terminals.

deteriorate sharply as element spacings decrease. If the elements are closely spaced, then as the frequency is shifted either side of the array design frequency the electrical spacing between adjacent elements changes rapidly. This causes a higher SWR and a deterioration of forward gain and front-to-back ratio.

The Log-Yag system overcomes this difficulty by using a multiple driven element "cell" designed in accordance with the principles of the log-periodic dipole array.[7,8] Since this log cell exhibits both gain and directivity by itself, it is a more effective radiator than a simple dipole driven element. The front-to-back ratio and gain of the log cell can be improved with the addition of a parasitic reflector and director. It is not necessary for the parasitic element spacings to be large with respect to wavelength, as in the Yagi array, since the log cell is the determining factor in the array bandwidth. In fact, the element spacings within the log cell may be small with respect to a wavelength without appreciable deterioration of the cell gain. For example, decreasing the relative spacing constant (σ) from 0.1 to 0.5 λ will decrease the gain by less than 1 dB. Hence, a further reduction in boom length. It can be seen that the Log-Yag array will exhibit high theoretical gain (11 dBd), high front-to-back ratio (30 dB), high cross polarization (front-to-side ratio — 45 dB), and a wideband response utilizing boom lengths approximately one half that of a Yagi with similar characteristics.[2]

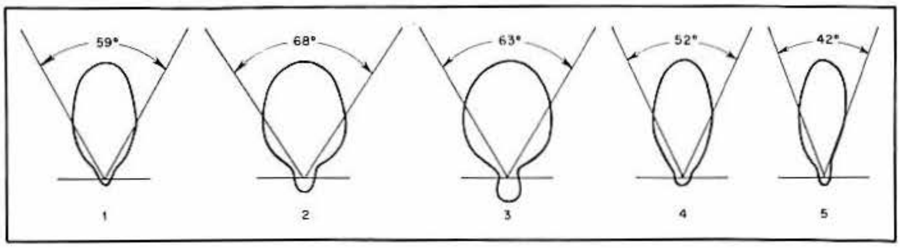

Fig. 10 — Beam patterns of 20-meter arrays.
No. 1 — 3-el. log cell, dir. @ 0.1λ, ref. @ 0.2λ.
No. 2 — 5-el. log periodic, σ = 0.1.
No. 3 — 3-el. log cell, 1st dir. @ 0.1λ, 2nd dir. @ 0.2λ.
No. 4 — 4-el. log cell, ref. @ 0.15λ.
No. 5 — 4-el. log cell, dir. @ 0.15λ, ref. @ 0.085λ (described in this article).

The author has built many monoband 14-MHz Log-Yag arrays in an attempt to find an optimum combination of elements, while holding the boom length to that of a full-sized, three-element monobander Yagi. Relative radiation patterns for various element combinations are found in Fig. 10. The final array design takes the form of a four-element log cell, parasitic reflector spaced at 0.085 λ_{max} and parasitic director spaced at 0.15 λ_{max} where λ_{max} is the longest free-space wavelength within the array passband. It has been found that array gain is almost unaffected with reflector spacings from 0.08 λ to 0.25 λ and the increase in boom length is not justified.[9] The function of the reflector is to improve the front-to-back ratio of the log cell while the director sharpens the forward lobe and decreases the half-power beamwidth. As the spacing between the parasitic elements and the log cell decreases, the parasitic elements must increase in length.[10]

The log cell is designed to meet upper and lower band limits with σ = 0.05λ. The design parameter τ is dependent on the structure bandwidth, B_s. When the log-periodic design parameters have been found, the element length and spacings can be determined. A review of the "Log-Periodic Dipole Array" is recommended though not necessary for the design of the Log-Yag array.[3,7,8]

The method of feeding the antenna is identical to that of feeding the log-periodic dipole array without the parasitic elements. As shown in Fig. 11, a balanced feeder is required for each log-cell element, and all adjacent elements are fed with a 180° phase shift by alternating connections. Since the Log-Yag array will be covering a relatively small bandwidth, the radiation resistance of the narrow-band log cell will vary from 80 to 90 ohms (tubing elements) depending on the operating bandwidth. The addition of parasitic elements lowers the log-cell radiation resistance. Hence, it is recommended that a 1-to-1 balun be connected at the log-cell input terminals and 52-ohm coaxial cable be used for the feed line. The measured radiation resistance of the 14-MHz Log-Yag installed at the author's QTH is 37 ohms, 14.0 to 14.35 MHz. It is assumed that tubing elements will be used. However, if a wire array is used then the radiation resistance R_o and antenna-feeder input impedance Z_o must be calculated so that the proper balun and coax may be used. The procedure is outlined in detail in *The ARRL Antenna Book*.[3,7]

Design Procedure

The following step-by-step design procedure may be used to design any monoband Log-Yag for any desired bandwidth.

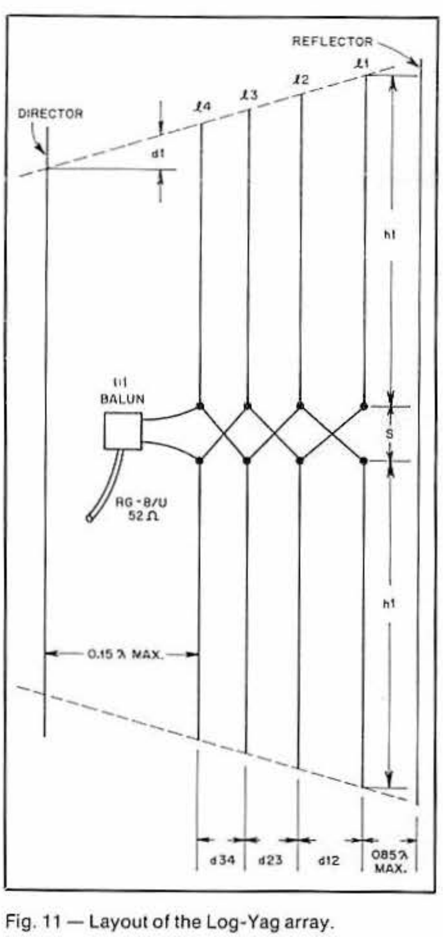

Fig. 11 — Layout of the Log-Yag array.

1) Determine the operating bandwidth, B, between f_1, lowest frequency (band edge), and f_n, highest frequency (band edge).

$$B = \frac{f_n}{f_1}$$

2) Determine the structure bandwidth (log-cell array) B_s.

$$B_s = 1.15B$$

3) Determine the design parameter τ (based on 4-element log cell, note 1).

$$\tau = \frac{1}{\sqrt[3]{B_s}}$$

Note 1. The design parameter τ is chosen for a four-element log cell since it provides the best bandpass for most amateur bands. For log cells with any number of elements

$$\tau = \frac{1}{\sqrt[(n-1)]{B_s}}$$

where n = number of elements within the log cell.

4) Determine the apex half-angle α: Since $\sigma = 0.05$ (relative spacing constant), then

$$\cot \alpha = \frac{0.2}{1 - \tau}$$

5) Determine the longest free-space wavelength λ_{max}, log-cell boom length, L (ft) and longest element length within the log cell l_1 (ft).

$$\lambda_{max} = \frac{984}{f_1 \text{MHz}}$$

$$L = \left[\frac{1}{4}\left(1 - \frac{1}{B_s}\right) \cot \alpha\right] \lambda_{max}$$

and

$$l_1 = \frac{492}{f_1 \text{MHz}}$$

$$l_2 = \tau l_1$$

$$l_3 = \tau l_2$$

$$l_4 = \tau l_3$$

6) Determine the element spacing (d_{12}), distance between elements l_1 and l_2 (ft).

$$d_{12} = \frac{1}{2}\left(l_1 - l_2\right) \quad \cot \alpha$$

and

$$d_{23} = \tau d_{12}$$

$$d_{34} = \tau d_{23}$$

7) Determine the parasitic element lengths (ft) and spacings (ft).

$$l_{REF} = \frac{509.6}{f_1 \text{ MHz}}$$

$$d_{REF} = \frac{84}{f_1 \text{ MHz}}$$

$$l_{DIR} = \frac{450.8}{f_1 \text{ MHz}}$$

$$d_{DIR} = \frac{148}{f_1 \text{ MHz}}$$

This completes the design.

The Finished Log-Yag

The proof is always to be found in the completed and operating product. The

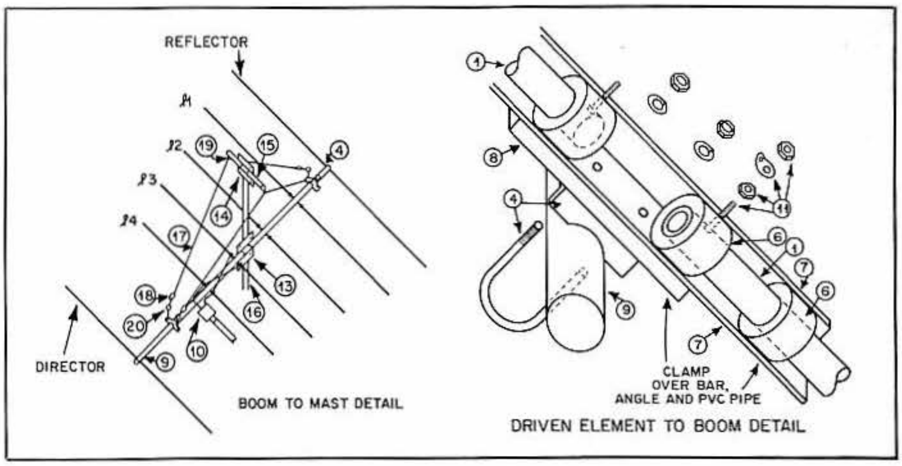

Fig. 12 — Assembly details. The numbered components refer to Table 6.

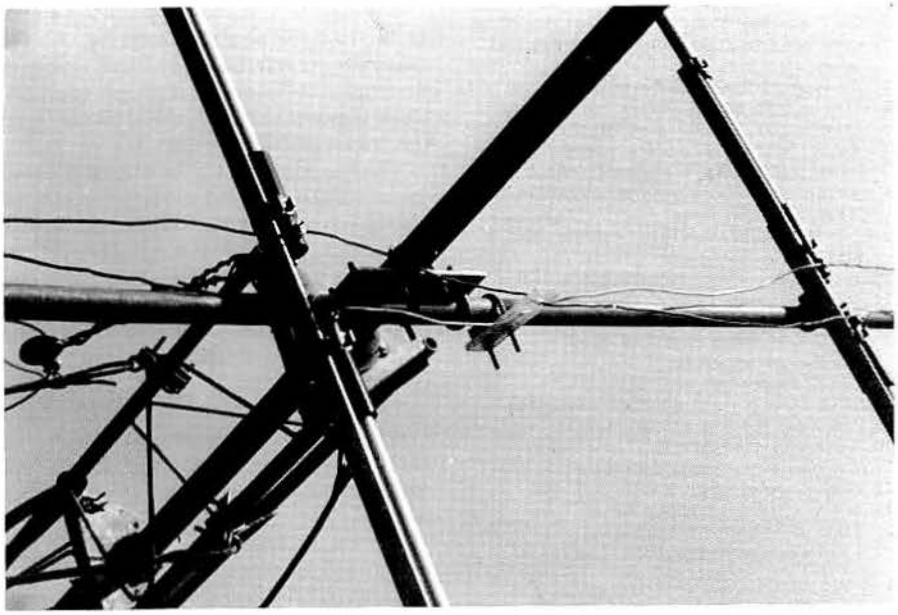

The interconnection between the elements of the log cell changes sides between each element.

Table 3

Array Characteristics

1. Frequency range	14–14.35 MHz
2. Operating bandwidth	B = 1.025
3. Design parameter	$\tau = 0.946657$
4. Apex half angle	a = 14.92°, cot = 3.753
5. Half-power beam width	42° (14-14.35 MHz)
6. Bandwidth of structure	$B_s = 1.17875$
7. Free-space wavelength	$\lambda_{max} = 70.28$ ft
8. Log cell boom length	L = 10.0 ft
9. Longest log element	$l_1 = 35.14$ ft (a tabulation of element lengths and spacings given in Table 2)
10. Forward gain over dipole	11.5 dB (theoretical)
11. Front-to-back ratio	32 dB (theoretical)
12. Front-to-side ratio	45 dB (theoretical)
13. Input impedance	$Z_O = 37$ ohms
14. SWR	1.3 to 1 (14-14.35 MHz)
15. Total weight	96 pounds
16. Wind-load area	8.5 sq. ft
17. Feed-point impedance	$Z_O = 37$ ohms
18. Reflector length	36.4 ft @ 6.0 ft spacing
19. Director length	32.2 ft @ 10.5 ft spacing
20. Total boom length	26.5 ft

The mechanical construction of the log cell is identical to that described in *The ARRL Antenna Book*, except for the lengths and spacings.[3,7] Fig. 12 shows how the log cell is constructed as well as the addition of the parasitic elements.

1 ft = 0.30478 m

Table 4

Array Dimensions

Element	*Length Ft*	*Spacing Ft*
Reflector	36.4	6.0 (Ref. to l_1)
l_1	35.14	3.51 (d_{12})
l_2	33.27	3.32 (d_{23})
l_3	31.49	3.14 (d_{34})
l_4	29.81	10.57 (l_4 to dir.)
Director	32.2	

1 ft = 0.30478 m

Table 5

Element Material Requirements

Element	*1-in. Tubing Length (Ft)*	*Qty.*	*7/8-in. Tubing Length (Ft)*	*Qty.*	*3/4-in. Tubing Length (Ft)*	*Qty.*	*1 1/4-in. Angle Length (Ft)*	*1 X 1/4-in. Bar Length (Ft)*
Reflector	12	1	6	2	8	2	None	None
l_1	6	2	6	2	8	2	3	1
l_2	6	2	6	2	8	2	3	1
l_3	6	2	6	2	6	2	3	1
l_4	6	2	6	2	6	2	3	1
Director	12	1	6	2	6	2	None	None

1 ft = 0.30478 m

Table 6

Materials List

1. Aluminum tubing — 0.047 in. wall thickness
    1 in. — 12 ft lengths, 24 lin. ft
    1 in. — 12 ft or 6 ft lengths, 48 lin. ft
    7/8 in. — 12 ft or 6 ft lengths, 72 lin. ft
    3/4 in. — 8 ft lengths, 48 lin. ft
    3/4 in. — 6 ft lengths, 36 lin. ft
2. Stainless steel hose clamps — 2 in. max., 8 ea.
3. Stainless steel hose clamps — 1-1/4 in. max., 24 ea.
4. TV-type U bolts — 1-1/2 in., 6 ea.
5. U bolts, galv. type: 5/16 in. X 1-1/2 in., 4 ea.
6. U bolts, galv. type: 1/4 in. X 1 in., 2 ea.
7. 1 in. ID water-service polyethylene pipe 160 lb/in.^{2} test, approx. 1-3/8 in. OD, 7 lin. ft
8. 1-1/4 in. X 1-1/4 in. X 1/8 in. aluminum angle — 6 ft lengths, 12 lin. ft
9. 1 in. X 1/4 in. aluminum bar — 6 ft lengths, 6 lin. ft
10. 1-1/4 in. top rail of chain-link fence, 26.5 lin. ft
11 1:1 toroid balun, 1 ea.
12. No. 6-32 X 1 in. stainless steel screws, 8 ea.
    No. 6-32 stainless steel nuts, 16 ea.
    No. 6 solder lugs, 8 ea.
13. No. 12 copper feed wire, 22 lin. ft
14 12 in. X 6 in. X 1/4 in. aluminum plate, 1 ea.
15. 6 in. X 4 in. X 1/4 in. aluminum plate, 1 ea.
16. 3/4 in. galv. pipe, 3 lin. ft
17. 1 in. galv. pipe — mast, 5 lin. ft
18. Galv. guy wire, 50 lin. ft
19. 1/4 in. X 2 in. turnbuckles, 4 ea.
20. 1/4 in. X 1-1/2 in. eye bolts, 2 ea.
21. TV guy clamps and eyebolts, 2 ea.

1 in. = 25.4 mm
1 ft = 0.30478 m

author's 14-MHZ Log-Yag on-the-air performance on cw and ssb substantiates the theory. The characteristics of the array are given in Table 3.

The materials needed are given in Table 5. In the construction diagram, Fig. 12, the materials are referenced by their respective material list number. The photographs show the overall construction picture, and the drawings show the details.

The materials should be available from most hardware and electronic stores. However, some have found difficulty in obtaining aluminum tubing. This can be solved by writing to the manufacturer and asking for the name of their distributor nearest your locality. Commercial antenna manufacturers will sell their tubing, but the cost is at a premium.

This array is in operation at K4EWG and W4BBP. The results on the air are nothing short of fantastic! It will give the stacked Yagis and long-boom Yagis a run for their money.

It is the authors' hope that this antenna design will stimulate additional work and research by other amateurs. The field seems wide open, and Yag-Log combinations are endless. The optimum design is by no means achieved in this article. It does seem, however, that a log cell of more than four elements would be necessary only where the array bandwidths, B, exceed 1.03 ($B = f_n/f_1$).

The authors wish to thank George Smith, W4AEO, for his work in substantiating a consideration for log periodic gain.[8] This material was originally presented in *QST* by P. D. Rhodes, K4EWG and J. R. Painter, W4BBP.

References

[1] "The Swan Multidrive 2-Meter Antenna," *QST*, Oct. 1969.

[2] Orr, *Radio Handbook*, 18th edition, pp. 547-548.

[3] Rhodes, "The Log-Periodic Dipole Array," *QST*, November 1973.

[4] *The ARRL Antenna Book*, 13th edition, p. 151.

[5] See ref. 2 above, p. 558.

[6] King, Mack and Sandler, "Arrays of Cylindrical Dipoles," *Yagi-Uda Programme Apdx. V*, 1968, pp. 218-232, 468-470.

[7] See ref. 4 above: pp. 160-164, 208-210, and Fig. 9-18.

[8] Smith, "Yes, I've Built Sixteen Log Periodic Antennas!" *73*, March 1975. See pp. 98-99.

[9] Orr, "Antennas," *CQ*, March 1975. See section headed "The KLM Antenna."

[10] See ref. 4 above, p. 204 and Fig. 9-5.

Chapter 3

Quad Antennas

A currently popular form of directional gain antenna is called the "quad." In general performance principles, it compares favorably to the Yagi beams discussed in chapter 2. The proper name for this antenna is the "cubical quad." It was developed by Clarence Moore, W9LZX, in 1942 when he was an engineer for HCJB in Quito, Ecuador. The design came as a solution to large corona discharges at the ends of existing beam-antenna elements. The problem was caused by the highly ionized air at the transmitter location, some two miles above sea level. Operation was in the commercial 25-meter band, which for the intentions of the broadcaster, required a gain type of antenna.

Moore concluded that a gain antenna with no ends was required to handle the 10-kW power without corona or flashover. Thus, the quad was born. Not only did it handle the power without difficulty, it later became a popular amateur antenna after he developed a model for use on 20 meters.

The primary limitation of quad antennas is the unwieldy nature of the "beast," especially at 40 and 20 meters, compared to a Yagi. The supporting framework is subject to extreme stress from ice and wind loading. Nonetheless, the quad has proven to be one of the most effective DX beams that an amateur can build at modest cost. This chapter contains data on various types of quads for a number of amateur bands. It is a compendium of the best on the subject, as published in past issues of *QST*.

Quads and Yagis: Comparisons, Patterns and Working Dimensions

Want a fresh look at the Yagi vs. quad question? This is "must" reading for the antenna-minded.

The material in the following sections will present a discussion of two time-tested antenna arrays.

In this class of arrays the Yagi (more properly the Yagi-Uda) antenna is the grandfather of them all. There are very few antennas which can compete with the electrical operation and simplicity of construction of this array. However, in the early 1940s C. C. Moore, of radio station HCJB, brought into being a parasitic array of loop elements with properties very comparable to those of the Yagi-Uda array. Since its conception this array has become very popular in Amateur Radio circles. It has come to be known as the "cubical quad" or "quad" antenna because of the common use of square loops and box-like construction.[1]

The material to follow will attempt to present an unbiased comparison of these two arrays based upon the experimental and theoretical results of many workers. This material is prefaced by a short section of introductory antenna concepts and definitions so that certain terms and notation will have more meaning when the comparisons between the quad and Yagi antennas are made. Finally, in the last section, some of the experimental patterns of modeled quad and Yagi arrays are presented. A compilation of some of the working dimensions for both types of arrays are also given. This latter material should be of interest to the builder and experimenter.

Introductory Antenna Theory and Definitions: Polarization

Inspection of the radiation field of an antenna will show it to have plane-wave character and to be made up of electric and magnetic field lines which are perpendicular to each other. Polarization of an electromagnetic wave is defined as that direction in which the electric field lines are oriented. If the electric field vector is fixed (in direction) in a plane it is said to be linearly polarized. For example, the polarization of a dipole is linear. If the plane in which this dipole lies is horizontal or parallel to some reference plane (e.g., the surface of the earth), it is called a horizontally polarized antenna. Vertical polarization would result if the dipole were placed perpendicular to the reference plane. It should be noted that if the electric vector does not stay fixed in position but rotates (its tip tracing out an ellipse), then the wave is said to be elliptically polarized. Circular polarization is a special case of elliptical polarization. Elliptical polarization will have a right- or left-hand sense, depending upon the direction of rotation of the electric vector.

Directivity and Gain of an Antenna

The degree to which a particular antenna can concentrate the antenna pattern into a beam is known as *directivity*. Mathematically, it is the ratio of the power density (in watts per square unit of measure) that would be available at an observation point to the power density one would have if the total radiated power were radiated *equally* well in *all* directions.[2] Directivity is a function of the antenna pattern alone and does not take into account antenna efficiency or losses.

The gain of an antenna can be defined in terms of directivity when antenna efficiency is taken into consideration. If K is the antenna efficiency, a number between 0 and 1, then antenna gain is given by

$$G = KD$$

where

G = antenna gain
K = antenna efficiency
D = directivity

The gain in decibels is then given by

$$\text{dB gain} = 10 \log_{10} G.$$

Please note that the above definition has been such that it is a function of the observation point. In amateur circles gain is usually measured at that observation point where the maximum power density is observed. Further, when considering practical Yagi and quad arrays for the hf range, K is normally taken as 1. Gain measured in the above manner is normally called gain with respect to an isotropic antenna. It is interesting to note that since a half-wave dipole concentrates its pattern into certain directions, it will have some gain with respect to an isotropic antenna. This gain with respect to isotropic is found to be

$$G = 1.64$$

or, in dB,

$$G_{dB} = 2.15 \text{ dB}$$

If the vertical and horizontal half-power beam widths are known for a given antenna the following formula will give (somewhat on the generous side) the gain of the antenna with respect to an isotropic antenna

$$G = \frac{40{,}000}{\theta_H \theta_V}$$

where

θ_H = horizontal half-power beam width in degrees.

θ_V= vertical half-power beam width in degrees.

This formula assumes low-side and back-lobe levels for the antenna pattern. If the gain with respect to isotropic is known then the gain with respect to a half-wave dipole is obtained by subtracting 2.15 dB.

Antenna Aperture or Area

This concept is of most use when discussing a receiving antenna. It is a measure of the antenna's ability to gather in available radiated power at some receiving position. Antenna area and gain are *directly* related

$$G = 4\pi \frac{A}{\lambda^2}$$

where

G = antenna gain

λ = operating wavelength

A = area (based upon square units of length measurement, with the length measurement the same as that assumed for λ). It should be noted that one *cannot* have two antennas with the same gain and different areas. One hears this concept misused (in the sense of the above sentence) quite often on the amateur bands. That is to say, if you feel that one antenna has more capture area than another, then it must also have more gain.

Antenna Patterns

The pattern of an antenna may be plotted by recording the received electric field strength or power density as a function of viewing (observation) angle. The properties of most antennas used by radio amateurs (below 30 MHz) can be described by two basic patterns. Consider a pattern due to a radiating half-wave dipole, this pattern to be obtained by moving a sampling or test antenna about the dipole in the horizontal plane which contains the dipole. This pattern is called the horizontal-plane or E-plane pattern (the pattern was taken in a plane containing the electric E-lines, hence the name E-plane pattern). If the sampling antenna is placed directly in front of the horizontal dipole (and in the horizontal plane) and moved up over the top of the dipole (in a plane perpendicular to the horizontal plane), and on around until it returns to the starting position, the result is the vertical-plane or H-plane pattern (since the test antenna moved in a plane parallel to the magnetic H-lines it is called an H-plane pattern).

For Yagi and quad arrays these two patterns will give enough essential information to determine their expected radiating properties.

Comparison of Quad and Yagi Arrays

The cubical quad antenna was first introduced to the radio world in 1942 by Clarence C. Moore at HCJB to alleviate a serious corona problem occurring on a large commercial shortwave Yagi array. It is a parasitic array formed with loop elements which are approximately one wavelength in circumference.

Important to proper operation of a quad antenna are considerations of proper element sizes, spacing between elements, boom length, feed impedance and so forth. The following material presents a discussion of these considerations and a comparison of the quad and Yagi arrays.

Polarization of a Quad Antenna

The polarization of a quad antenna can be determined by considering the current distribution on the driven element. To stress the point that square, diamond, or circular loop all produce essentially the same polarization, consider the current distribution shown in Fig. 1. A little thought reveals that fact that the vertical components of current produce radiated fields that cancel each other, and the horizontal components of current produce a radiated field (broadside to the loop) with all components adding. The polarization of any of these loops is then linear and parallel to an imaginary horizontal line drawn through the current minima. Hence, if a quad loop is fed at the bottom it produces horizontal polarization, and if fed on the side it will have vertical polarization. The geometrical form of the loop, square, diamond or circular, does not mean that the (broadside) radiation has a possible combination of both horizontal and vertical polarization or circular polarization. This seems to be a common misconception in amateur circles. The linear polarization properties introduced above can be verified easily experimentally. It should be further noted that there are no noticeable differences in the radiating properties or impedance of a loop 1 wavelength in circumference whether it be of the circle, diamond or square configuration.

Angle of Radiation for a Horizontally Polarized Quad Antenna

Placement of an antenna above a conducting ground will alter its free-space radiation pattern, positioning the maximum of its main lobe at possibly a new angle with respect to a horizontal reference line. The angular position of this main lobe is called the *angle of radiation.* The angle of radiation of an antenna is a function of its height above the conducting ground and its polarization. Given the height and the polarization, the angle of radiation can be determined by using the concept of the antenna's virtual image or ray theory and reflection of the wave by the ground.

The following conclusions can now be drawn:

The angle of radiation of a *horizontally* polarized quad antenna, at a given boom height, is essentially the *same* as that of a *horizontally* polarized Yagi at the same boom height.

This must be so because the boundary conditions imposed upon the electromag-

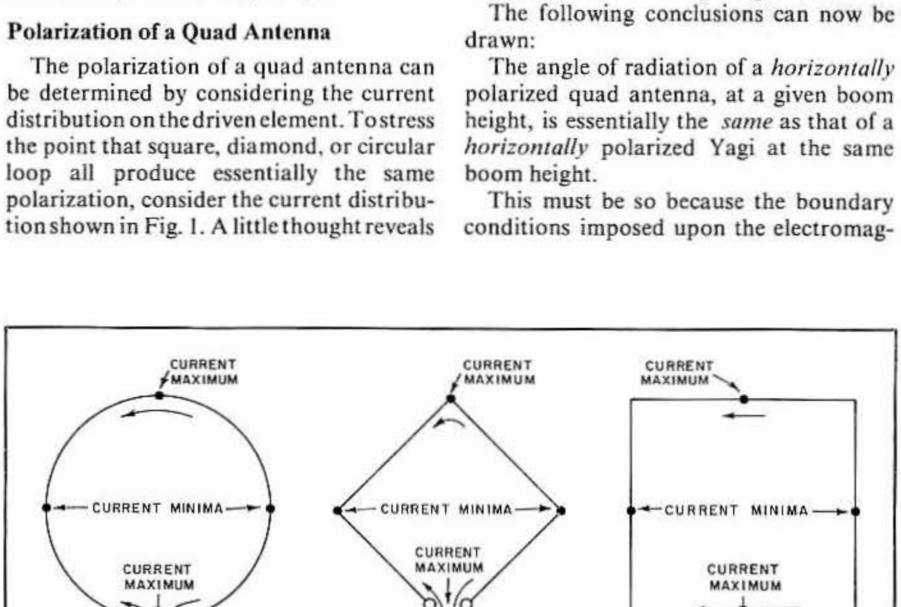

Fig. 1 — Current distribution on circular, diamond and square loops 1 wavelength in circumference.

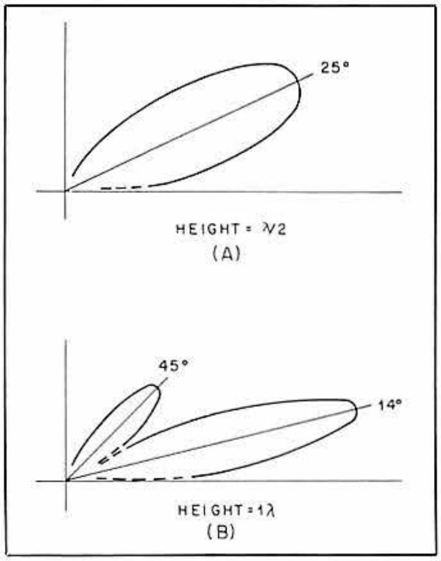

Fig. 2 — Vertical-plane radiation patterns for a horizontally polarized quad or Yagi array above perfectly conducting ground.

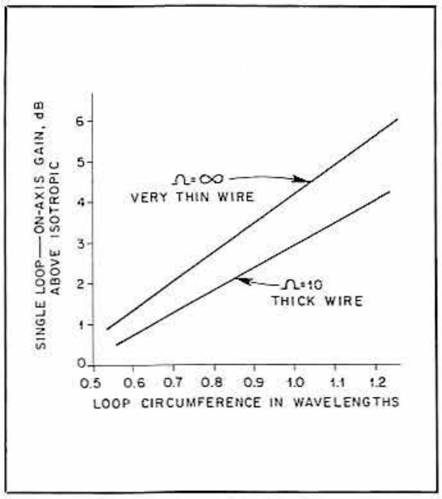

Fig. 3 — On-axis gain of a single loop as a function of loop circumference. (From "Studies of Large Circular Loop Antennas" by S. Adachi and Y. Mushiake *Sci. Rep. Ritu*, B-[Elect. Comm.] vol. 9, no. 2, Tohoku Univ., Sendai, Japan, 1957.)

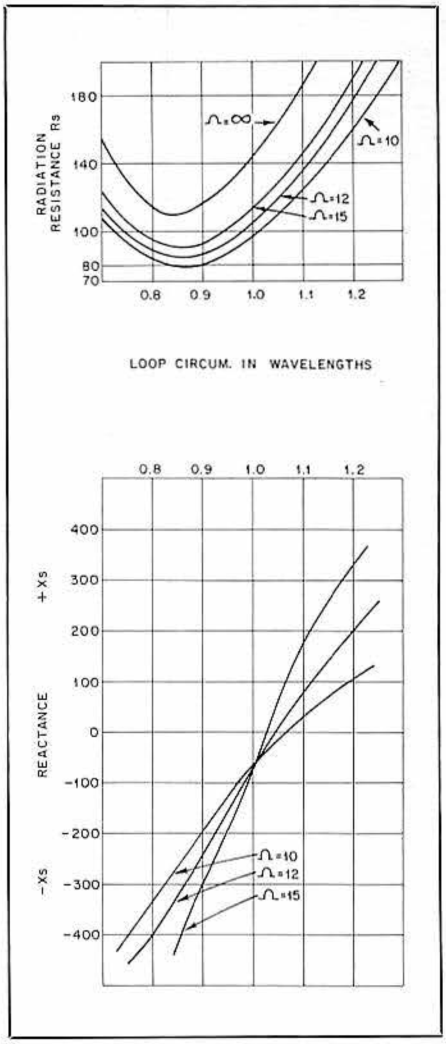

Fig. 4 — Reactance and radiation resistance of a single loop versus circumference. Note: Ω is a measure of the ratio of loop circumference to wire diameter. $\Omega = 2\ln 8a/\rho$ where a = loop radius and ρ = wire radius.

netic wave at the surface of ground are completely insensitive to the fact that the wave is produced by a quad antenna as opposed to the wave produced by a Yagi antenna.

Typical patterns of a quad *or* Yagi antenna placed at a height of 1/2 λ or 1 λ above a perfect ground are shown in Fig. 2 at A and B.

Gain Comparisons of the Quad Antenna with the Yagi Antenna

The gain of a quad array (of given boom length) seems to be somewhat higher than that of a Yagi array of the same boom length. This gain difference can be explained by considering the theoretical work of Adachi and Mushiake of Tohoku University in Japan. In 1952 and 1957 they published two papers which discussed the current distribution, impedance, and radiating properties of loops of large circumference.[3]

Fig. 3 shows the gain of a single loop as a function of wavelength and wire diameter. Fig. 4 shows the variation of reactance and radiation resistance for a single loop. Please note the dependence upon the circumference-to-wire-diameter ratio in these two figures. They showed that the broadside gain of a single loop 1λ in circumference was approximately 2 dB above that of a half-wave dipole. This result implies that a parasitic array of loop elements should have the same gain advantage when compared to a Yagi array of comparable boom length.

This gain differential for the quad array does appear and can be verified experimentally. Fig. 5 demonstrates this comparison. The points on the figure were obtained as follows: For the antenna under consideration E and H plane patterns were made. The gain was then calculated using the E and H-plane half-power beam widths of these radiation patterns and the formula discussed above. This gain and the boom length of the antenna were then used to determine a point on the graph given in Fig. 5, where gain and boom length represent the parameters under consideration. The data points for many, many antennas, both quad and Yagi, were then placed on the graph and a smooth curve representing an average for each type of antenna was then drawn.

The expected gain difference of 2 dB does indeed appear in favor of the quad antenna. This gain differential is essentially that predicted by the results of the Japanese workers cited above.

The measurements used to obtain the above gain comparisons were performed at 440 MHz. A measuring frequency of 440

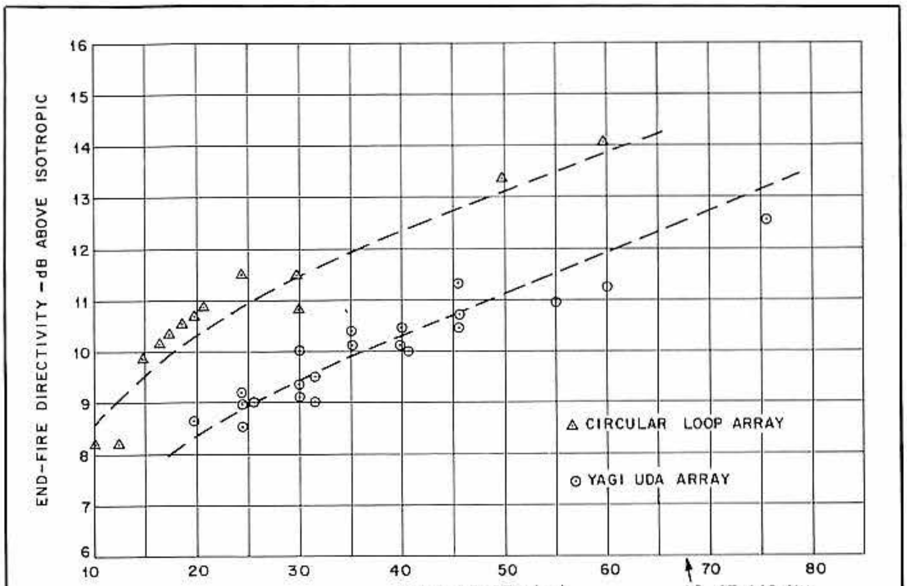

Fig. 5 — Quad and Yagi array gain as a function of boom length.

Table 1

Quad Dimensions

2-element Quad (W0HTH)
Spacing (given below)
Boom length (given below)

Band	*40 M*	*20 M*	*15 M*	*10 M*
Reflector	144 ft 11-1/2 in.	72 ft 4 in.	48 ft 8 in.	35 ft 7 in.
Driven Element	140 ft 11-1/2 in.	70 ft 2 in.	47 ft 4 in.	34 ft 7 in.
Spacing	30 ft	13 ft	10 ft	6 ft 6 in.
Boom length	30 ft	13 ft	10 ft	6 ft 6 in.
Feed method	Directly with 23 ft of RG-11, then any length RG-8 coax.	Directly with 11 ft 7 in. RG-11, then any length RG-8 coax.	Directly with 7 ft 8-1/2 in. RG-11, then any length RG-8 coax.	Directly with 5 ft 8 in. RG-11, then any length RG-8 coax.

(Note that a spider or boomless quad arrangement could be used for the 10/15/20 meter parts of the above dimensions yielding a triband antenna)

4-element Quad* (W0AIW [20 M] /W0HTH** /K0KKU/K0EZH/W6FXB)
Spacing: equal; 10 ft
Boom length: 30 ft

Band	*20 M*		*15 M*	*10 M*
	Phone	CW		
Reflector	72 ft 1-1/2 in.	72 ft 5 in.	48 ft 8 in.	35 ft 8-1/2 in.
Driven Element	70 ft 1-1/2 in.	70 ft 5 in.	47 ft 4 in.	34 ft 8-1/2 in.
Director 1	69 ft 1 in.	69 ft 1 in.	46 ft 4 in.	33 ft 7-1/4 in.
Director 2	69 ft 1 in.	69 ft 1 in.	46 ft 4 in.	33 ft 7-1/4 in.
Feed Method	Directly with 50-ohm coax.		Directly with 50-ohm coax.	Directly with 5 ft 9 in. RG-11, then any length RG-8 coax.

*Common boom used to form a triband array.
**The 2-element 40-meter quad given above is added to form a four-band quad array.

4-element Quad (W0HTH/K8DYZ*/K8YIB*/W7EPA*)
Spacing: equal; 13 ft 4 in.
Boom length: 40 ft

Band	*20 M*	*15 M*	*10 M*
Reflector	72 ft 5 in.	48 ft 4 in.	35 ft 8-1/2 in.
Driven Element	70 ft 5 in.	47 ft 0 in.	34 ft 8-1/2 in.*
Director 1	69 ft 1 in.	46 ft 1 in.	(Directors 1-3 all 33 ft 7 in.)*
Director 2	69 ft 1 in.	46 ft 1 in.	
Feed method	Directly with 50-ohm coax.	Directly with 7 ft 9 in. RG-11, then any length 50-ohm coax.	Directly with 50-ohm coax.

*For the 10-meter band the driven element is placed between the 20/15 reflector and 20/15 driven element. The 10-meter reflector is placed on the same frame as the 20/15-meter reflectors and the remaining 10-meter directors are placed on the remaining 20/15-meter frames. The 10-meter portion is then a 5-element quad.

6-element Quad (W0YDM, W7UMJ)
Spacing: equal; 12 ft
Boom length: 60 ft

Band: 20 M	
Reflector	72 ft 1-1/2 in.
Driven Element	70 ft 1-1/2 in.
Directors 1, 2 and 3	69 ft 1 in.
Director 4	69 ft 4 in.
Feed Method	Directly with 50-ohm coax.

1 ft = 0.3048 m

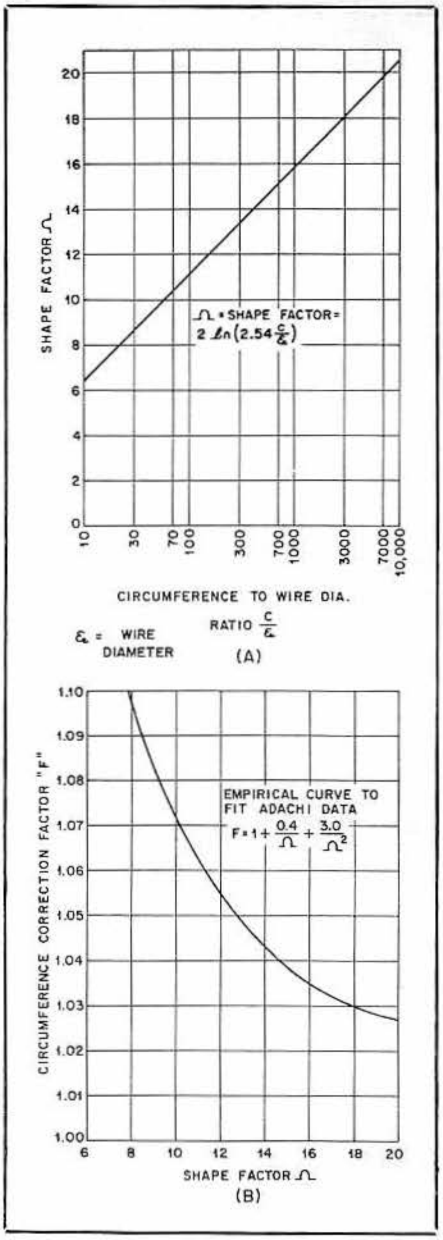

Fig. 6 A: Shape factor as a function of the loop circumference-to-wire diameter ratio. B: Loop "resonant length correction factor" versus shape factor. (Circled values obtained from S. Adachi, and Y. Mushiake, "Studies of Large Circular Loop Antennas" *Sci. Rep. Ritu*, B-[Elect. Comm.] vol. 9, no. 2, Tohoku Univ., Sept. 1957.)

MHz was chosen so that 1 cm (at this frequency) is equivalent to 1 foot of measurement at 14 MHz. Hence, to relate these results directly to the 20-meter amateur band, boom length in cm from the curve can be read as boom length in feet. For example, the gain of a quad on a 40-foot (12.19-m) boom should be slightly more than 1 dB above a quad on a 25-foot (7.62-m) boom. Comparison of the signals from these two antennas at a receiving station will not reveal much difference; however, in the DX pile-ups the longer antenna should have a slight advantage. Another interesting interpretation of these results is to note that a quad on a 30-foot (9.14-m) boom is equivalent to a Yagi on a 55-foot (16.76-m) boom. Hence the boom of a Yagi antenna must be about 1.8 times longer than the boom of a quad in order for the two gains to be comparable.

Impedance Properties and Matching

The driving-point impedance of a quad antenna depends upon the element spacing and number of elements. For 2-element quad antennas the results of Bill Orr in his *Quad Antenna* handbook can be used to determine the driving-point impedance. For multielement quads (3 or more elements)[4] the impedance will range from 40 to 60 ohms. Matching of these multielement quads is most easily done by direct feed with a 50-ohm coaxial line. A 1:1 balun can be used but experimental results show that it does not give performance any better than that obtained with direct feed. A gamma match can be used to obtain an exact match to any chosen coaxial line. For details of using a gamma match on a quad antenna see Bill Orr's *Quad Antenna* handbook.

If a single boom is used to construct a multiband quad antenna, then it is recommended that a separate feed line to each driven element be used.

Approximate Formulas for Quad Antenna Element Lengths

The multielement quad antenna gained its first real popularity following the work of Lee Bergren, W0AIW, in the early 1960s. He published the results of his work in *QST* in 1963.[5] Many experimentalists have used his dimensions for antennas in the 20-meter amateur band and have scaled them to other bands. Based upon Bergren's work and others the following empirical formulas can be used to determine the element lengths of a quad antenna

$$\text{Circumference of driven element} = \frac{1005}{f_{MHz}} \text{ feet}$$

$$\text{Circumference of reflector element} = \frac{1030}{f_{MHz}} \text{ feet}$$

$$\text{Circumference of director elements} = \frac{975}{f_{MHz}} \text{ feet}$$

These results should only be used for quad antennas operating below 30 MHz. It is interesting to note that for quad antennas the reflector and directors are only 2-1/2 to 3 percent different in length from the driven element. This is a smaller change than that normally encountered in Yagi arrays.

The above figures should not be used for vhf or uhf quad antennas. At these higher frequencies the circumference-to-wire-diameter ratio becomes small enough that its effect upon the resonant properties of the loop must be considered. Here again, the results of Adachi and Mushiake are useful and the effects of wire size on the resonant length of the loop can be taken into account. These results are given in graphical form in Fig. 6 where, based upon the circumference-to-wire-diameter ratio, a circumference correction factor can be obtained. Once the correct size of the driven element is determined (using the results of Fig. 6) a quad antenna can be designed by making the reflector approximately 3 percent longer and the directors 3 percent shorter.

The above results can be used to obtain workable dimensions but, just as with approximate formulas for Yagi antenna element lengths, the "best" dimensions would have to be determined experimentally by tuning the array.

A Collection of Quad and Yagi Patterns and Dimensions: E and H Plane Patterns for Quad and Yagi Arrays

The previous section referred to a group of experimental tests, where a series of model antennas were constructed and the E and H plane patterns obtained.

It should be noted that these patterns have direct interpretation in light of the 20-meter band. Recall that the experiments were performed at 440 MHz, where dimensions for boom length, element size, and spacing in cm are directly equivalent to dimensions in feet for the 20-meter band. For example, the patterns for a 440-MHz, 5-element Yagi on a 55-cm boom can also be taken as the pattern for a 14-MHz, 5-element Yagi on a 55-foot boom.

Perusal of these patterns will give a

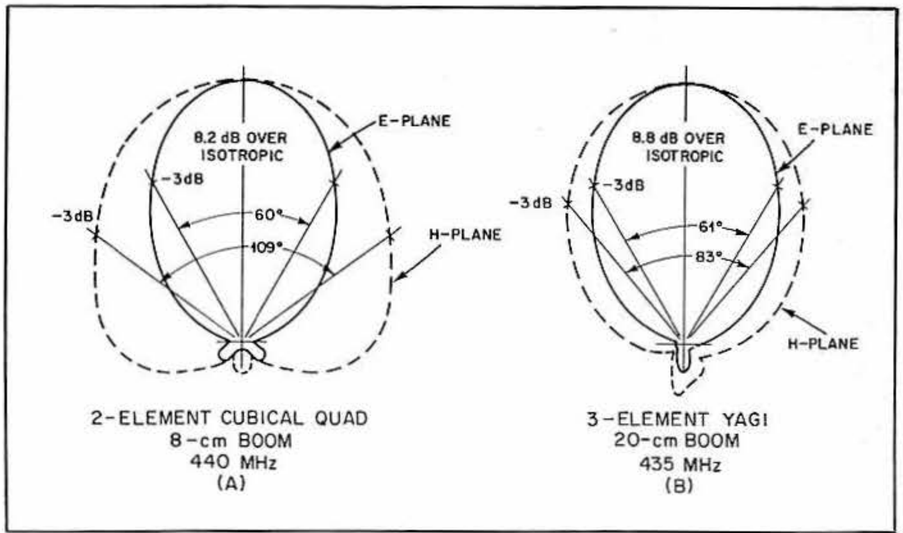

2-ELEMENT CUBICAL QUAD
8-cm BOOM
440 MHz
(A)

3-ELEMENT YAGI
20-cm BOOM
435 MHz
(B)

Fig. 7 — E and H plane radiation patterns for (A) 2-element quad antenna and (B) 3-element Yagi antenna.

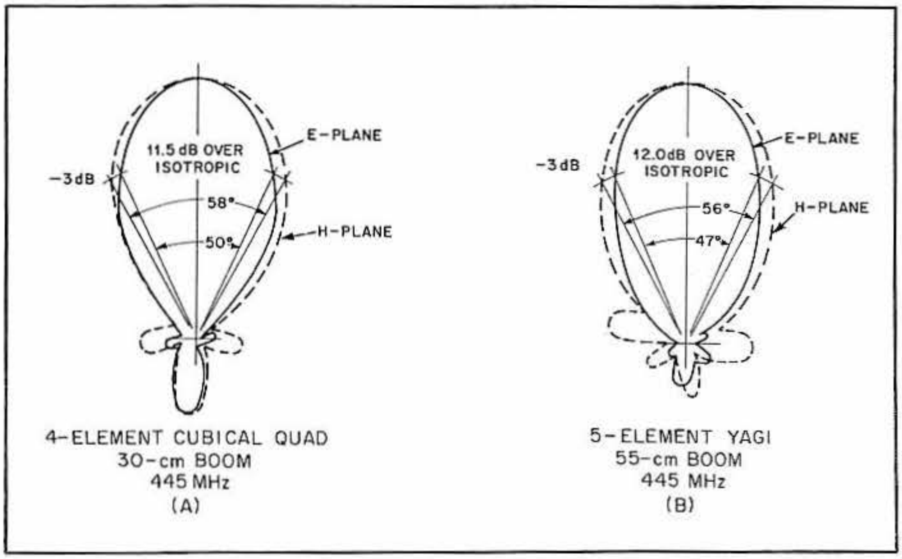

4-ELEMENT CUBICAL QUAD
30-cm BOOM
445 MHz
(A)

5-ELEMENT YAGI
55-cm BOOM
445 MHz
(B)

Fig. 8 — E and H plane radiation patterns for (A) 4-element quad antenna and (B) 5-element Yagi antenna.

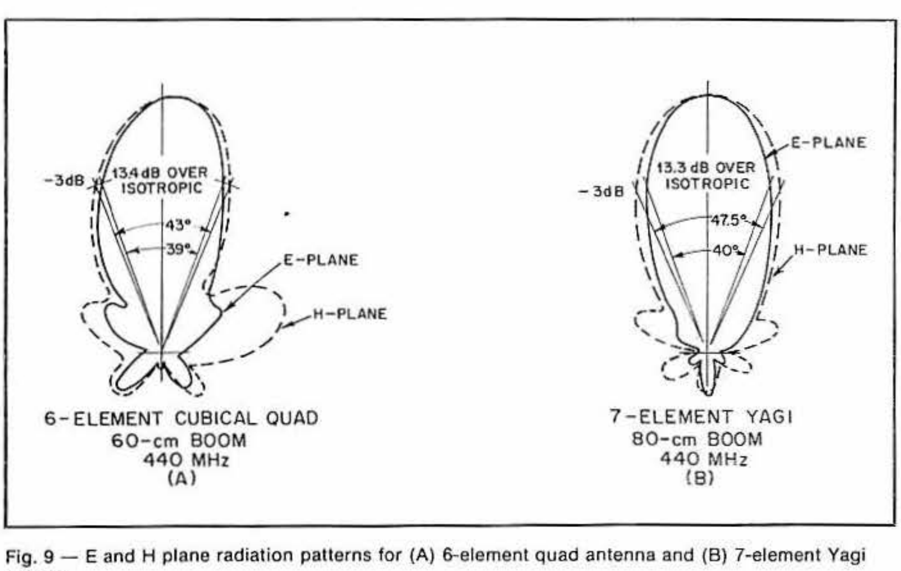

6-ELEMENT CUBICAL QUAD
60-cm BOOM
440 MHz
(A)

7-ELEMENT YAGI
80-cm BOOM
440 MHz
(B)

Fig. 9 — E and H plane radiation patterns for (A) 6-element quad antenna and (B) 7-element Yagi antenna.

visual means of comparing the quad and Yagi arrays. It is interesting to note the dependence of gain and pattern beam width upon array boom length. Without further comment, typical patterns for both types of antennas are given in Figs. 7 through 9.

A Tabulation of Dimensions of Working Arrays

Presented in tabular form are a few representative dimensions for both quad and Yagi antennas. These are dimensions presently in use by a number of amateurs. Most of the Yagi dimensions are due to amateurs across the country who sent their dimensions to Mr. Hal Wolff or the author to be modeled on the antenna range at the University of Denver. The quad dimensions also are those now in use by a number of amateurs in this country and abroad. It should be noted that most of these dimensions, and the resulting model antennas, provided the necessary data used in the previous section to compare the quad and Yagi. Where known, either the user(s) or the originator of a set of dimensions is indicated. The first set of dimensions is for quad antennas, the second set for Yagi antennas.

Summary

A comparison between the quad and Yagi antennas, on the basis of gain, has been given, this comparison being based upon both theoretical and experimental results. The results indicate that if both antennas have the same boom length the gain of the quad array will be about 2 dB higher than that of the Yagi. For either antenna, gain is a function of boom length and will increase approximately 2.5 dB each time the boom length is doubled. Hence, a 2-dB gain advantage is equivalent to a boom-length advantage of 1.8. In other words, for the quad and Yagi arrays to have comparable gain, the boom length of the Yagi antenna must be 1.8 times longer than that of the quad.

The size of the wire used for the loop element (relative to the loop circumference) will have an influence upon the resonance and radiating properties of the loop. The influence of this shape factor — i.e., the ratio of loop circumference to wire diameter — can be accounted for by using the referenced results of Adache and Mushiake.

Table 2

Yagi Dimensions

3-element Yagi (W6SAF)
Boom length: 24 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 10-3/4 in.	R—DE: 13 ft 9-1/2 in.
Driven Element	33 ft 2-1/2 in.	DE—D: 10 ft 4-1/2 in.
Director	31 ft. 9-1/2 in.	

3-element Yagi (W0OKC)
Boom length: 26 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D: 12 ft
Director	31 ft 7 in.	

3-element Yagi (W0CM, formerly W0JYW)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4-1/2 in.	R—DE: 13 ft
Driven Element	32 ft 2-1/2 in.	DE—D: 17 ft
Director	31 ft 10 in.	
Feed Method	Folded dipole	

4-element Yagi (K0KKU)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	(all elements equally spaced)
Driven Element	33 ft 4 in.	
Director 1	31 ft 9-1/2 in.	
Director 2	31 ft 2 in.	

4-element Yagi (W4PLL)
Boom length: 30 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 4-1/2 in.	(all elements equally spaced)
Driven Element	33 ft 4-1/2 in.	
Director 1	31 ft 9-1/2 in.	
Director 2	31 ft 9-1/2 in.	

4-element Yagi (K6VIZ)
Boom length: 32 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 10 ft
Driven Element	32 ft 8 in.	DE—D1: 10 ft
Director 1	31 ft 8 in.	D1—D2: 12 ft
Director 2	31 ft 1 in.	

4-element Yagi (W0OKC)
Boom length: 36 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 10 ft 8 in.
Director 1	31 ft 8-1/2 in.	D1—D2: 11 ft 4 in.
Director 2	31 ft 1 in.	

4-element Yagi
Boom length: 40 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 6 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 12 ft
Director 1	31 ft 7 in.	D1—D2: 14 ft
Director 2	31 ft 0 in.	

5-element Yagi (WA6ZZK)
Boom length: 40 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	34 ft 10 in.	R—DE: 10 ft
Driven Element	33 ft 4 in.	DE—D1: 9 ft
Director 1	32 ft 6 in.	D1—D2: 10 ft 6 in.
Director 2	32 ft 2 in.	D2—D3: 10 ft 6 in.
Director 3	31 ft 10 in.	

5-element Yagi
Boom length: 46 ft
Band: 20 M

	Element Length	*Spacing*
Reflector	35 ft 7-3/4 in.	
Driven Element	34 ft 5-3/8 in.	(All elements equally spaced at 11 ft 6 in.)
Director 1	33 ft 2 in.	
Director 2	31 ft 9-1/2 in.	
Director 3	30 ft 4-1/2 in.	

1 ft = 0.3048 m

In conclusion, it might be stated that one type of antenna (quad or Yagi) is not necessarily better than the other. The choice between a quad or Yagi array will depend critically upon many factors. For example, if single-band operation is of main importance and the geographical area where the antenna is to be used does not have high-wind problems, the best and simplest choice would be a long-boom Yagi antenna. If multiband operation is desired, or if wind problems were to limit the length of the boom to some nominal value, then the quad antenna would be a good choice. Likewise, the choice of available materials in your locale, and the acceptance of the appearance of the antenna by your neighbors or family, constitute other determining factors. The best approach is to survey your requirements and

5-element Yagi (K6EVR)
Boom length: 46 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 4 in.	R—DE: 12 ft
Driven Element	33 ft 5 in.	DE—D1: 10 ft
Director 1	31 ft 8 in.	D1—D1: 12 ft
Director 2	31 ft 8 in.	D2—D3: 12 ft
Director 3	31 ft 8 in.	

5-element Yagi (W8PWH)
Boom length: 52 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 11 in.	R—DE: 13 ft 10 in.
Driven Element	33 ft 4-1/2 in.	DE—D1: 10 ft 5 in.
Director 1	32 ft 10 in.	D1—D2: 13 ft 10 in.
Director 2	31 ft 2 in.	D2—D3: 13 ft 10 in.
Director 3	30 ft 9 in.	

5-element Yagi (W4EX, formerly W4DQH)
Boom length: 54 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 0 in.	R—DE: 13 ft 6 in.
Driven Element	33 ft 4 in.	DE—D1: 11 ft 6 in.
Director 1	32 ft 3 in.	D1—D2: 13 ft
Director 2	32 ft 1 in.	D2—D3: 16 ft
Director 3	31 ft 11 in.	

5-element Yagi (W6RR, formerly W6ITA)
Boom length: 55 ft
Band: 20 M

	Element Length	Spacing
Reflector	34 ft 9 in.	R—DE: 16 ft 9 in.
Driven Element	33 ft 7 in.	
Director 1	31 ft 11 in.	(All elements equally spaced)
Director 2	31 ft 9-1/2 in.	
Director 3	31 ft 8 in.	

6-element Yagi
Boom length: 46 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 5 in.	
Driven Element	34 ft 5-1/2 in.	(All elements equally spaced at 9 ft 2-3/8 in.)
Director 1	32 ft 7 in.	
Director 2	32 ft 3 in.	
Director 3	31 ft 4-1/2 in.	
Director 4	31 ft 1/2 in.	

6-element Yagi (W0OKC)
Boom length: 46 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 4 in.	R—DE: 10 ft
Driven Element	33 ft 4 in.	DE—D1: 8 ft
Director 1	31 ft 11 in.	D1—D2: 9 ft
Director 2	31 ft 3-1/2 in.	D2—D3: 9 ft 6 in.
Director 3	31 ft 1/2 in.	D3—D4: 9 ft 6 in.
Director 4	30 ft 11 in.	

6-element Yagi (W0OKC)
Boom length: 74 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 5 in.	R—DE: 16 ft 6 in.
Driven Element	33 ft 4 in.	DE—D1: 9 ft 9 in.
Director 1	31 ft 9-1/2 in.	D1—D2: 14 ft
Director 2	31 ft 2-1/2 in.	D2—D3: 17 ft 6 in.
Director 3	30 ft 11 in.	D3—D4: 21 ft
Director 4	30 ft 8 in.	

7-element Yagi (W6HAW)
Boom length: 80 ft
Band: 20 M

	Element Length	Spacing
Reflector	35 ft 5 in.	R—DE: 14 ft
Driven Element	33 ft 4 in.	DE—D1: 11 ft
Director 1	31 ft 8 in.	D1—D2: 12 ft
Director 2	31 ft 4 in.	D2—D3: 13 ft
Director 3	30 ft 8 in.	D3—D4: 14 ft
Director 4	30 ft 8 in.	D4—D5: 16 ft
Director 5	30 ft 8 in.	

8-element Yagi (WA6EKD)
Boom length: 100 ft
Band: 20 M

	Element Length	Spacing
Reflector	34 ft 9 in.	(All elements equally spaced at 14 ft 4 in.)
Driven Element	33 ft 5 in.	
Director 1	32 ft 2 in.	
Director 2	31 ft 11-3/4 in.	
Director 3	31 ft 9-7/8 in.	
Director 4	31 ft 8 in.	
Director 5	31 ft 2 in.	
Director 6	30 ft 11-3/4 in.	

environmental restrictions, make your choice and proceed accordingly.

Acknowledgements

It is very difficult to acknowledge systematically and completely all those who have made the quad antenna a successful communications antenna. However, surely credit must be given to C. C. Moore, the father of the quad antenna, to William Orr, who presented the first written material on the theory and construction of 2-element quads; to Lee Bergren, who showed the feasibility and advantages of using multielement quad arrays; to the Japanese workers, Adache and Mushiake, of Tohoku University, who presented a detailed analysis and experimental verification of the electromagnetic properties of loop antennas of large circumference; and finally, to all the radio amateurs, in all corners of the world, who developed constructional and design techniques and proved or disproved their validity.

The author would like also to acknowledge the help of Mr. Hal Wolff (WA0IOR) for his aid in constructing the model Yagi and quad arrays and running the patterns. These patterns were obtained by using the antenna range of the Denver Research Institute of the University of Denver.[6] Many Denver-area amateurs have entered into these quad experiments, notably Messrs. Jim Snyder (K0ZCM), Butch Ford (WA0IMX), Ed Wood (K0KKU), Keith Farris (W0YDM), Warren Wheeler (W0R-EQ) and many others. The author gratefully acknowledges their assistance, help and encouragement. This material originally presented in *QST* by J. E. Lindsay, Jr., W7ZQ.

References

[1] For a colorful description of C. C. Moore's invention of the cubical quad antenna see the introduction of W. I. Orr's *Quad Antennas*, Radio Publications, Inc.

[2] An antenna which radiates equally well in all directions is called an isotropic radiator or isotropic antenna.

[3] Adachi and Mushiake, "Theoretical formulation for large circular loop antennas by integral equation method," *Repts. of the Research Institute of Elec. Commun.*, Tohoku University, Sendai, Japan, Sci. Rept. Ritu, B(Elec. Commun.), vol. 9, no. 1, 1957.

Idem., "Studies of large circular loop antennas," *Repts. of the Research Institute of Elec. Commun.*, Tohoku University, Sci. Rept. Ritu, B(Elec. Commun.), vol. 9, September 1952.

[4] Element spacing of 0.14λ to 0.2λ being considered.

[5] Bergren, "Multielement Quads," *QST*, May 1963.

[6] At the time these tests were performed (1962-64), the author was head of the Antenna Laboratory, Denver Research Institute, University of Denver. Mr. Hal Wolff was the research technician at this facility.

Additional References

Jasik, *Antenna Engineering Handbook*. New York: McGraw-Hill, 1961, pp. 16-25.

Lindsay, "A circular loop antenna with non-uniform current distribution," *IRE Trans. Antennas and Propagation (Communications)*, vol. AP-8, pp. 439-441, July 1960.

Lindsay, "A parasitic end-fire array of circular loop elements," *IEEE Trans. Antennas and Propagation (Communications)*, vol. AP-15, No. 5, pp. 697-698, Sept. 1967.

A Triband One-Loop Cubical Quad Element

Why not trade gain for economy and structural reliability? This three-band "DX Getter" consists of a single quad loop. It will stay aloft in high winds and won't look like an ugly monster to those who dwell nearby.

The advantages of a rotatable beam antenna are well known to all DXers. A problem arises in many situations regarding just how a reasonable compromise between antenna performance and cost, size and weight can be obtained. Because this author likes to live in peace with the XYL, the neighbors and the local council, the antenna must not be considered dangerous or an eyesore as far as appearance is concerned. In addition, the antenna must not hang over the fence of the 50- × 150-foot (15- × 46-m) block, half of which is already occupied by buildings. With these limitations firmly implanted in mind, the author added such factors as cost and difficulty of erecting and supporting a full-size, 20-meter Yagi or cubical quad. After weighing all the facts, it was felt that the only way out was to construct an economy-style, reduced-size antenna.

A single-loop, cubical-quad triband element is shown in Fig. 10. Such a cubical quad has only one-fourth the weight and wind resistance of a full-sized, three-element cubical quad of the same mechanical strength. The mechanical and installation difficulties are many times smaller; the cost of mast and rotor are also considerably reduced. Yes, the antenna is a compromise and the gain is less than that of a 3-element quad, but this antenna still puts a signal in good DX company. This quad, because of its smaller physical size, will be tolerated in many more locations than its big brother. If desired, this quad could be used as an indoor antenna strung between two opposite room walls.

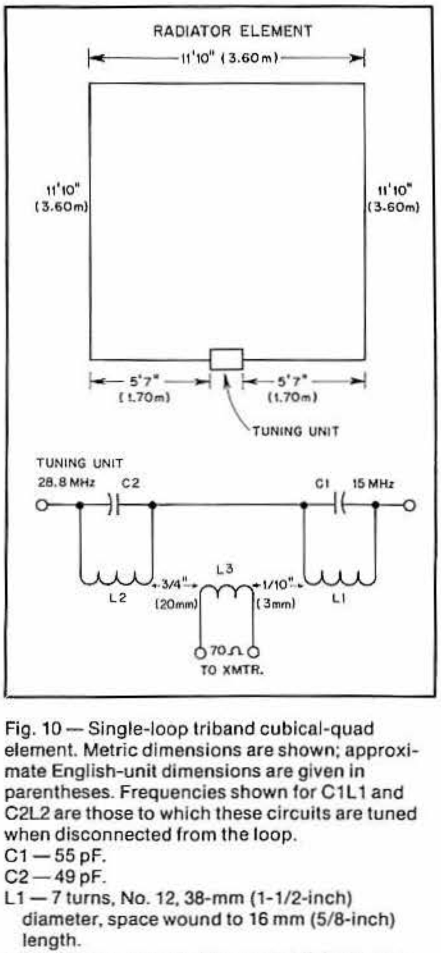

Fig. 10 — Single-loop triband cubical-quad element. Metric dimensions are shown; approximate English-unit dimensions are given in parentheses. Frequencies shown for C1L1 and C2L2 are those to which these circuits are tuned when disconnected from the loop.

C1 — 55 pF.

C2 — 49 pF.

L1 — 7 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 16 mm (5/8-inch) length.

L2 — 4 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 18 mm (11/16-inch) length.

L3 — 4 turns, No. 12, 38-mm (1-1/2-inch) diameter, space wound to 10-mm (3/8-inch) length.

Rather than going to a complicated switching arrangement to change bands, it was decided to apply what the author called the "triband-antenna principle" to this antenna. The theory behind this principle evolves around the same general ideas outlined by Pichitino in *QST*[1] several years ago. Antennas constructed along these guidelines do not operate harmonically and have only the desired resonances. The quad that was built is shortened only at the lowest operating frequency. It is full size at the middle operating frequency and much longer at the highest operating frequency. In addition, no heavy blocking tuned circuits are used near the element ends.

With C1 giving 20-meter resonance depends mainly on the total element length and L1. Fifteen-meter resonance is controlled principally by C1 and L2. Ten-meter resonance is determined by C2 and L2. Once the antenna is constructed, it may be fine-tuned on each band by varying the constants that control the resonant frequencies. After the quad was tuned up, the author found the SWR on each band to be very satisfactory. On 20 meters, the antenna resonated at 14.15 MHz with a resulting SWR of approximately 1:1. With a 15-meter resonance at 21.3 MHz, an SWR of 1.3:1 was measured. On 10 meters, 28.6 MHz was selected as the resonant frequency. The resulting SWR at resonance on this band was 1.5:1. The phase relationship of the fields in L1 and L2 varies from band to band so that too much direct coupling between these coils should be avoided.

Although the author used copper wire to make the loop element, you may desire to use aluminum tubing. The actual constructional details are left to the builder. The materials used and the methods of supporting the antenna are limited only by your imagination. The techniques employed may be as simple or as sophisticated as you may desire. DJ2UT used short pieces of coaxial cable as capacitors and placed them inside the tubing which he used for the element. The coils and inner capacitor ends were sealed and molded in resin. Fifteen- and 10-meter resonances are obtained by shifting copper rings over the ends of L1 and L2. This can be done without affecting the sealed coils in any way.

The author concluded this antenna to be rather efficient based on observations that the tuning network does not become warm and a low SWR has been measured across each of the three bands. This writer wishes to thank OM Sommer, DJ2UT, and his co-workers for the very considerable amount of work carried out, and the many practical ideas which have made successful antennas with the triband principle possible. This material was originally presented in *QST* by Hans F. Ruckert, VK2AOU.

---

[1] Pichitino, "A New Principle in Two-Band Rotary-Beam Design," *QST*, October 1948.

A 2-Element, 15-Meter Quad for the Novice

Why not build this 2-element quad for 15 meters? It is easy to construct, inexpensive and in many ways quite novel.

Not being blessed with a large area for antennas, the author decided on an optimum-spaced, 2-element monoband quad. The standard formulas were used for sides and spacing, and the antenna is fed with 50-ohm cable through a quarter-wave "Q" section of 75-ohm coax cut to the standard formula. The reflector has a tuning stub.

The driven element grid-dips at 21,150 kHz and the reflector was tuned by using the grid-dip oscillator as a signal generator. The grid-dip meter was placed in a neighbor's home. Then, the back of the quad was aimed at the dip meter, and the reflector was tuned for the lowest indication on the S-meter.

One of the checks made on the antenna was with VE5XH, in Swift Current, SK. Doug gave the following information: The quad has 2 S-units gain over my 1/4-wave groundplane which was at the same height. Strength off the front was S9 and to the rear was S2, indicating a good front-to-back ratio. Because the constructional details are rather novel, the author felt that other hams might be interested in how this antenna was built.

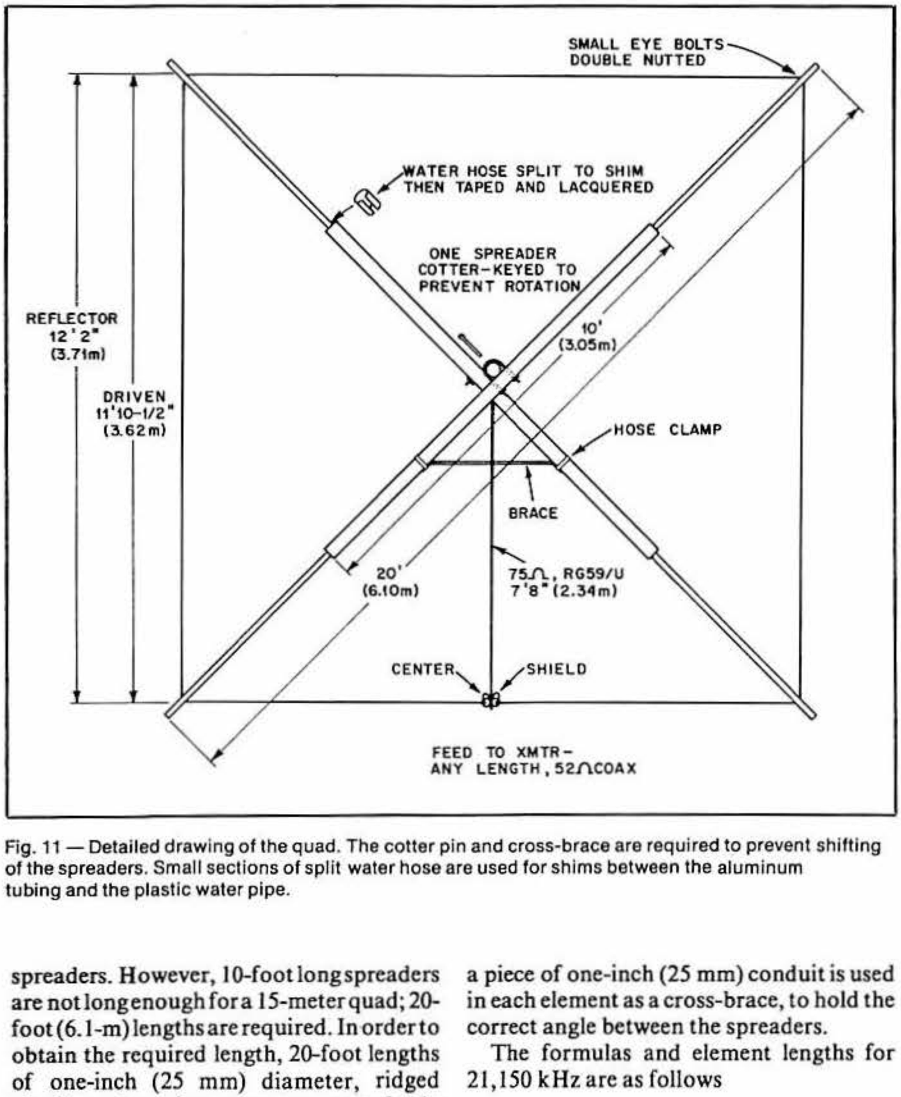
SMALL EYE BOLTS DOUBLE NUTTED
WATER HOSE SPLIT TO SHIM THEN TAPED AND LACQUERED
ONE SPREADER COTTER-KEYED TO PREVENT ROTATION
REFLECTOR 12'2" (3.71m)
DRIVEN 11'10-1/2" (3.62m)
10' (3.05m)
HOSE CLAMP
BRACE
20' (6.10m)
75Ω, RG59/U 7'8" (2.34m)
CENTER
SHIELD
FEED TO XMTR - ANY LENGTH, 52ΩCOAX

Fig. 11 — Detailed drawing of the quad. The cotter pin and cross-brace are required to prevent shifting of the spreaders. Small sections of split water hose are used for shims between the aluminum tubing and the plastic water pipe.

Construction

The author used 1-1/4-inch (32-mm) TV steel mast pipe, 10 feet (3 m) long, for the boom. Also, 10-foot lengths of TV mast sections, but aluminum, are used for the spreaders. However, 10-foot long spreaders are not long enough for a 15-meter quad; 20-foot (6.1-m) lengths are required. In order to obtain the required length, 20-foot lengths of one-inch (25 mm) diameter, ridged plastic water pipe were run completely through the 10-foot aluminum pipe sections. Pieces of rubber hose were used as shims between the aluminum pipe and the plastic pipe and these were taped where the plastic pipe emerges from the aluminum.

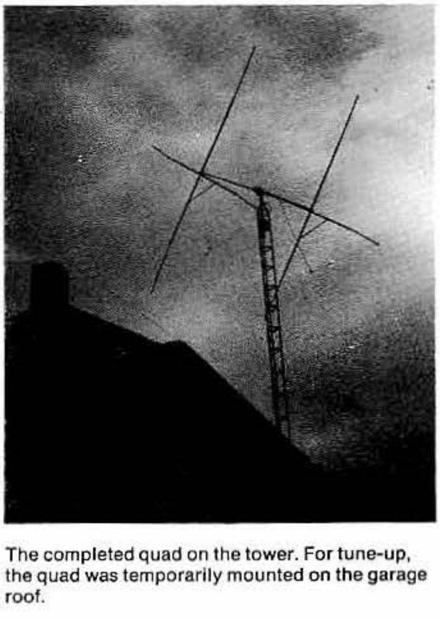

The completed quad on the tower. For tune-up, the quad was temporarily mounted on the garage roof.

The plastic water pipe comes in 20-foot lengths, 80 feet (24.4 m) being required. This pipe was purchased from a local plumbing and lighting outlet; the cost was $10 for the 80 feet. It is ridged plastic, but is not strong enough to support itself; so the aluminum mast pipe is used for the center portion of the spreaders.

The spreaders are clamped to the boom with standard U bolts, and one spreader in each element is pinned to the boom with a 5/16-inch (8 mm) cotter key. Additionally, a piece of one-inch (25 mm) conduit is used in each element as a cross-brace, to hold the correct angle between the spreaders.

The formulas and element lengths for 21,150 kHz are as follows

Driven element

$$\frac{1005}{f(\text{MHz})} = 47 \text{ feet, 6 inches (14.48 m)}$$

Reflector

$$\frac{1030}{f(\text{MHz})} = 48 \text{ feet, 8 inches (14.83 m)}$$

Boom spacing

$$0.2 \text{ wavelength} = 0.2 \times \frac{984}{f(\text{MHz})} = 9 \text{ feet, 3 inches (2.84 m)}$$

75-ohm matching section

$$\frac{246 \times V}{f(\text{MHz})} = \frac{162.4}{f(\text{MHz})} = 7 \text{ feet, 8 inches (2.34 m)}$$

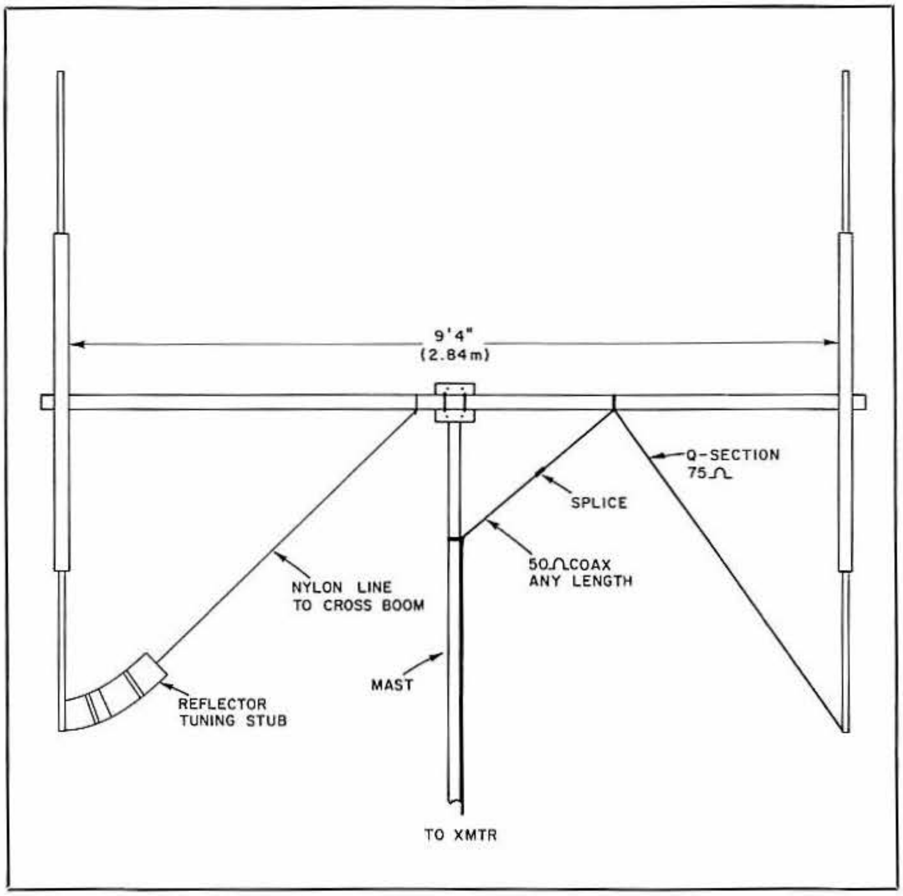

Fig. 12 — Side view, showing the mounting plate, Q matching section, and coax feed.

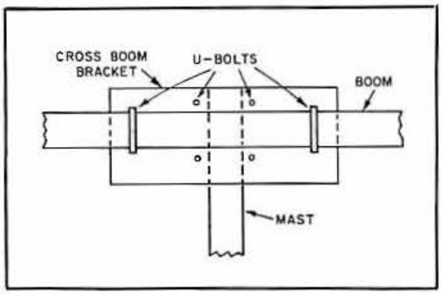

Fig. 13 — The cross-boom bracket, made from 3/8-inch (10 mm) aluminum plate. Ordinary TV U bolts are used to secure it to the boom and mast.

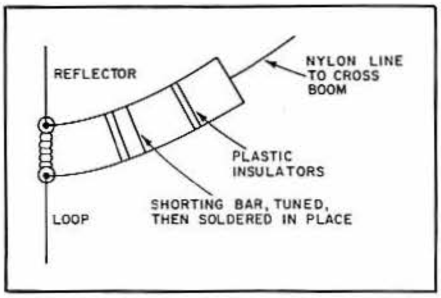

Fig. 14 — The reflector tuning stub is 36 inches (0.91 m) long. The shorting bar, a short piece of wire the same size as used in the stub was approximately 10 inches (254 mm) from the insulator when adjusted for maximum front-to-back ratio. Spacing between wires of the stub may be of the order of a few inches; this spacing is not critical since the stub length is adjusted by the shorting stub in the tune-up procedure.

The author used no. 14 soft-drawn enameled wire for loops and they were secured to the plastic pipe with double-nutted eye bolts. The reflector tuning stub was made from no. 12 copperweld wire. The total length of my stub is 36 inches (0.91 m), and after tuning, the shorting bar ended up approximately 10 inches (254 mm) from the loop.

The SWR showed almost a perfect match across the entire band. Since the quad was put up, the author has worked all continents and 35 countries on the Novice band. Stateside contacts always run S8 to S9. The total cost of the antenna was $28 at the time the system was built. This material was originally presented in *QST* by John Daebelliehn, WA9BJC.

A Practical 40-Meter Quad

One of the most perplexing problems continuously plaguing Field Day committees is how to install a good antenna system. For most clubs, finding a good location is generally the easiest job. The location must then be surveyed by the club antenna specialists. Deciding how many transmitters to use depends mostly on how many indivudual antennas can be set up for simultaneous operation as well as how many operators are available for the 24-hour stretch. Of course, an abundance of equipment is desirable.

Propagation conditions for Field Day and Sweepstakes generally favor the lower bands since they usually remain open all night. Success on the higher frequency bands (10 and 15 meters) depends primarily on catching the short-duration openings. The old trick of tying a rock to the end of a long piece of wire, tossing it into the tallest tree and loading it with a coupler just won't "cut the mustard" anymore. Gain and directivity into the dense population areas is the goal. However, the antenna must be simple to construct, easy to tune, and moderately stable against the wrath of Murphy and Mother Nature.

Our club decided to try a 2-element full-sized cubical quad for 40 meters. The immediate reaction within the more conservative ranks of our club was, "A 40-meter quad? It'll never work — the first breeze will knock it down!" With these objections in mind, we went ahead with our plans. Our location was approximately rectangular in shape, 60 feet wide, 500 feet long (18 × 152 m) and above the rest of the terrain. Of course, it just happened to be one of the highest locations in Winnebago County, IL!

The sketch illustrates the basic construction of the quad. Dimensions were estimated using the formula

$$L(\text{feet}) = \frac{248}{f(\text{MHz})}$$

where L is the length of one side of the diamond.

The parasitic element was constructed as a reflector using an extra long stub. A small relay shorts out the stub, thus converting the parasitic element from a reflector to a director (see Fig. 15).

The resonant frequency of the driven element is 7150 kHz, the middle of the band. The distance L for the driven element is 34 feet 8 inches (10.57 m). The parasitic element dimension is 5 percent smaller than the driven element. The stub is constructed so that the total circumference of the parasitic element with stub in the line is 5 percent greater than the driven element. L for the

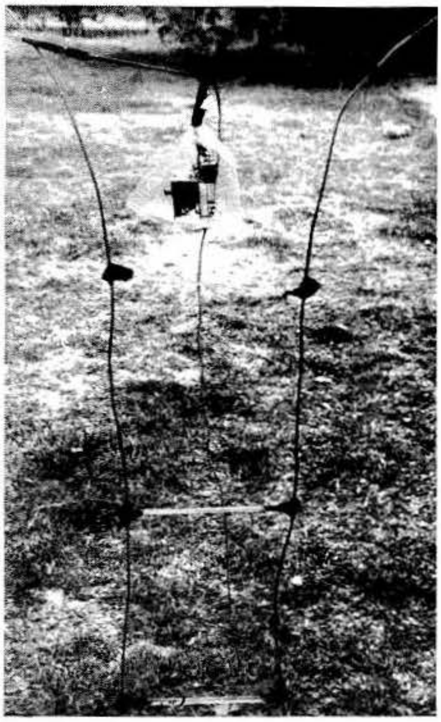

Fig. 15 — A simple relay arrangement is used to short out the stub and change the pattern of the antenna from one direction to another. A small plastic bag provides weatherproofing in the event it rains. (Doesn't it always?)

director is 33 feet (10.06 m). The stub length is 6 feet 9 inches (2.06 m). Eight-inch (203-mm) plastic spacers are placed between the two wires of the stub.

Supporting the quad is very simple. The mast is 50 feet (15.24 m) high, and guyed with nylon ropes. The boom is constructed from two telescoping sections of electrical conduit and is 18 feet (5.49 m) long. The elements are made from no. 14 stranded wire. The driven element is fed directly with RG-58/U coaxial cable. Location of the stub short is determined by the point of maximum front-to-back ratio. The boom-to-mast construction utilizes the support base from a Ham-M rotator. All it takes is a couple of U bolts and the boom-to-mast assembly is complete (see Fig. 16). The wire is strung from the end of the boom and is supported by a small piece of nylon rope and an egg insulator.

Results

Orientation for the quad was east and west. Favoring the eastern direction was the driven element/reflector combination, whereas the driven element/director combination was broadside to the western direction. A local amateur, approximately 5 miles (8 km) west of the Field Day site, assisted in tuning the quad for maximum front-to-back ratio. The end result was approximately 25 dB.

Results were a tremendous surprise! The antenna worked just as planned. The class of entry for simultaneous operation of two transmitters was used. One transmitter operated on 20 and 80 meters; the second transmitter was set up for 15 and 40 meters. Our club, W9AXD/9, scored third place in the two-transmitter class. Our biggest totals came from 40 meters — 693 contacts, over one third of our total number. For the first four hours of the contest, the antenna was aimed east. A rate of 90 contacts per hour prevailed for five solid hours. From that point on it was a steady decline after a sharp drop to about 45 contacts per hour. It was later surmised that the reason for the drop in contact rate was lack of activity and difficulty in finding stations that hadn't been worked before. The East Coast QRM was drastically reduced by changing the switch from "east" to "west." By midmorning we were surprised to find only a dozen W5 stations listed in the log. The antenna may not have been too effective in the southern direction.

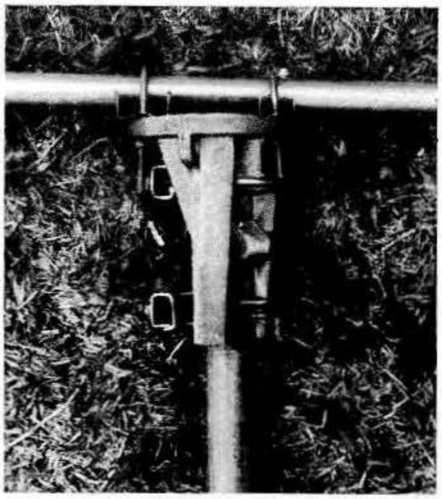

Fig. 16 — The bottom section of a CDE Ham-M rotor can be adapted to serve as a boom-to-mast bracket.

There is no reason why this arrangement could not be used for DX operating. The antenna could be positioned just underneath a beam for the higher frequency bands with the peak of the diamond supported from a boom placed through the top section of the tower.

This material was originally presented in *QST* by Peter H. Grillo, W6RTT.

Build This C-T Quad Beam for Reduced Size

This quad-beam antenna, while smaller in size than a conventional quad, performs like a veteran. The trade secret is to use capacitive tuning.

One of the main drawbacks of the quad beam is its large size compared to the Yagi beam. Previous efforts to reduce the size of the quad have utilized systems of inductive tuning. Loading coils have been used on one or two sides[1] or in the corners. Linear inductors on the vertical sides have also been used.[2] The purpose of this article is to explain a simple method of capacitive tuning to reduce the size of the quad. The C-T (capacitor-tuned) loop is diagrammed in Fig. 17.

It is well-known that the physical size of the dipole antenna can be reduced by replacing a portion of the antenna with a capacitive body and adding an inductor to restore the system to resonance. It seems less well-known that the full-wave loop antenna also may be reduced in physical size while holding the resonant frequency of the system constant by adding capacitance at the voltage loops. No inductors are needed. The capacitance may be added by any of the methods shown in Fig. 18.

The outside dimensions of the C-T loop may be anything chosen by the builder with the larger limit being the full-size, full-wave loop. The limit on the amount of size reduction effectively attainable depends both upon the electrical resistance in the circuit and upon the tendency in loop antennas for the current to become equal at all points in the circuit when the loop becomes very small in proportion to the free-space wavelength at the frequency under consideration. The geometrical shape of the loop is unimportant to basic operation provided that the capacitance is placed at or between the voltage loops. The bottom half of the chart in Fig. 21 shows the effective dimensional range of the C-T loop. The top half shows the tuning range that can

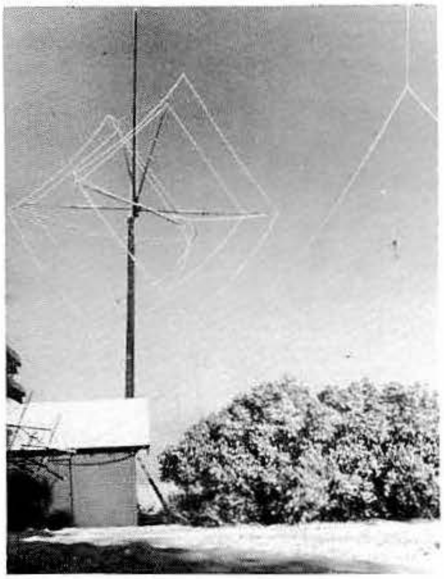

A winter snow sets the scene and accentuates the diamonds of the C-T quad-beam antenna at W7WKB.

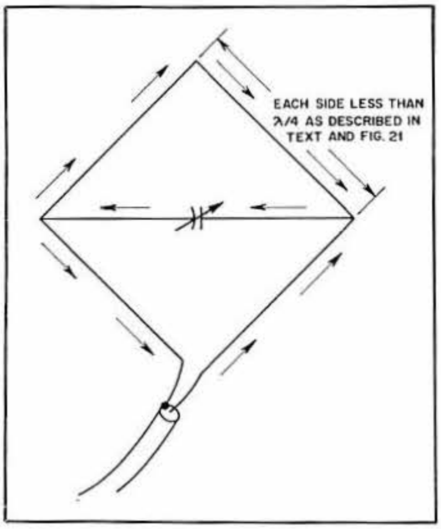

Fig. 17 — The C-T quad design with arrows indicating current flow for 1/2 cycle. The center capacitor is located at the minimum-current (highest-voltage) point in the loop. Each side is less than 1/4-wavelength long. See text and Fig. 21.

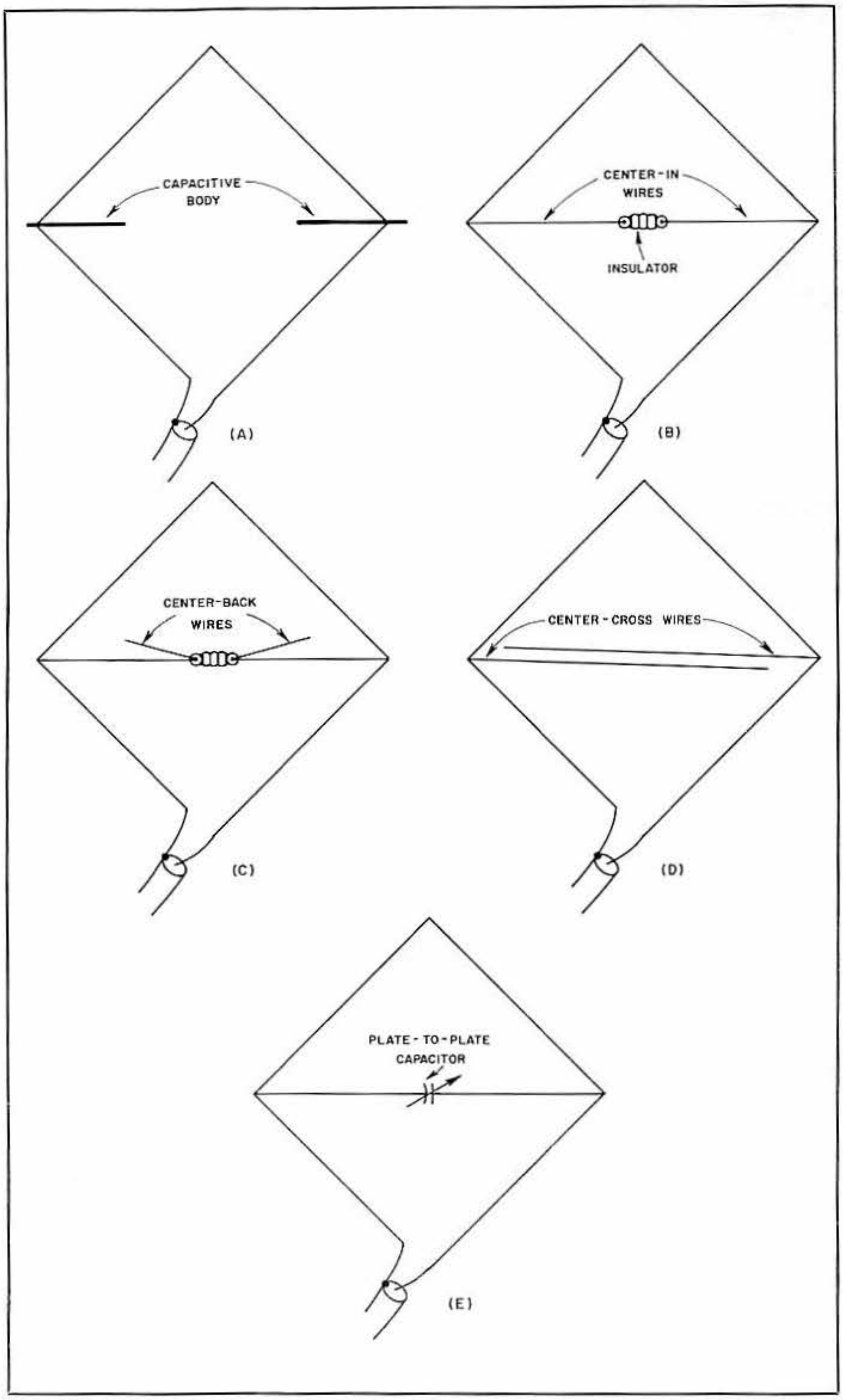

Fig. 18 — Methods of adding capacitance to the C-T loop are shown here. In Fig. 18C, if the center-back wires are placed too closely to the center-in wires, they become ineffective. A 12-inch (0.3048 m) spacing from the end of the center-back wire to the center-in wire is satisfactory. Wires running back toward the loop center may be attached to the outside ends of the center-back wires if more capacitance is needed. In Fig. 18D, the capacitance added by the center cross wires varies with both the length of wire and spacing between wires.

be expected by using various methods of adding capacitance to the diamond quad. These dimensions are intended as guidelines only.

The feedpoint resistance of the full-size, 1-wavelength loop is about 100 ohms. The feedpoint resistance of the C-T loop decreases as it is made smaller but will offer a good match to a 50-ohm coaxial line over the entire usable range as shown in the chart. A gamma match may be used if an exact match is desired.

More than one C-T loop may be used to form the C-T quad beam. A parasitic reflector was used in this installation because of the higher radiation resistance attainable while maintaining good gain. Maintaining the radiation resistance as high as possible is particularly important in small beams, both to keep efficiency high and to prevent the bandwidth from becoming too narrow.

Tuning and Mounting

The use of a grid-dip meter to attain initial band resonance is very helpful. By using a grid-dip oscillator, this beam was tuned initially with the driven element to resonate at the center of the band of interest. The reflector was then tuned to resonate at a frequency three percent lower than the driven element. Final adjustments were made using a test oscillator located about 1000 feet (300 m) away as a reference signal.

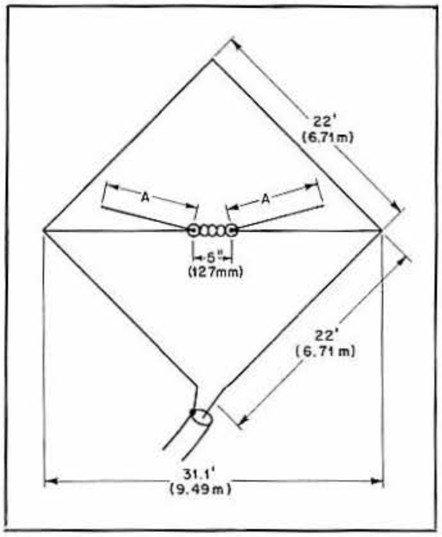

Fig. 19 — Dimensions for the 40-meter C-T quad are shown above. For the 2-element beam the spacing between the driven element and the reflector is 20 feet (6.10 m). The length of the center-back wire (dimension A) is 12 feet, 9 inches (3.89 m) for the reflector, and 9 feet, 11 inches (3.02 m) for the driven element.

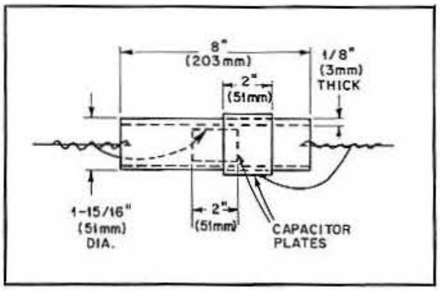

Fig. 20 — A variable capacitor is illustrated here. It can be used to tune the C-T quad. The eight-inch (203-mm) tubular portion is made from plastic water pipe. The cylindrical capacitor plates are formed from a copper sheet. One plate slides inside the eight-inch (203-mm) tube and the other over the outside of the tube. Maximum capacitance is approximately 25 pF.

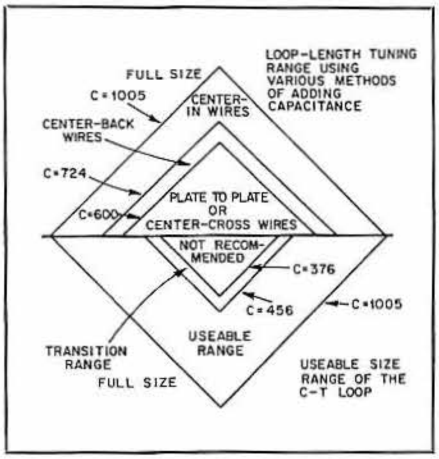

Fig. 21 — The chart shown above, providing dimensional guidelines for a single C-T loop, is based on the equation LF = C, where L is the outside dimension in feet, F is the frequency in MHz, and C is a constant chosen for the size of the C-T loop to be constructed. Use the largest size possible for the best electrical results. This is particularly true when using C-T loop elements to form a parasitic beam.

To use the chart, pick the largest C (up to C = 1005) feasible for your application. Use bottom half of the chart to see if the C-T loop will operate at the dimensional size selected. Then refer to the top half of the chart for the recommended capacitance method.

Table 3

Data for plotting SWR curve. Information in this table is used for plotting the SWR curve for the 40-meter C-T quad beam, as described in the text.

Frequency (MHz)	*SWR*
7.500	4.87
7.450	4.00
7.400	3.08
7.350	2.33
7.300	1.82
7.275	1.78
7.250	1.94
7.200	3.44
7.150	5.25
7.100	8.08

The director was tuned for lowest SWR and the reflector for best gain. If you lack a grid-dipper, initial resonance can be established by listening to your station receiver for a strong signal while making the adjustments. A dramatic increase in signal strength will be noticed at resonance. Tuning is accomplished by shortening or lengthening the capacitance wires.

The single loop or beam may be mounted either horizontally or vertically for horizontal or vertical polarization, but in either case it should be mounted as high above the ground as possible. Loop antennas operating close to ground will show a much higher feedpoint resistance than normal and will waste considerable power in groundheating effects. A rough guideline is to mount the antenna so that the bottom point is at least half the loop diameter above the ground.

The working 40-meter version of the C-T quad beam had a diamond configuration because of the greater mechanical strength and because of the longer center wire for tuning purposes. Final dimensions are shown in Fig. 19. A good quality center insulator should be used because of the high voltages present at the center.

A tubular tuning capacitor was originally installed in the driven element and worked well until the kW power level was reached. It then promptly arced over, leaving a conductive carbon deposit. A version using 2-inch (51-mm) plates as shown in Fig. 20 adequately tuned the reflector even at the kW level for some time. The capacitors, as shown, were not tried under wet conditions, but would probably be satisfactory at power levels up to 400 watts. For long-term use, the method shown in Fig. 18D is preferred over plate-to-plate capacitors because the high voltages are handled more easily.

Because some dielectric heating was noticed in this capacitor, the wire-tuned version was adopted. The wire version is recommended over the methods shown in Fig. 18D and 18E, any time the loop is large enough to be tuned by no more than one doubled-back wire (as shown in Fig. 18C). Tuning wires eliminate the capacitor with its heat loss and allow somewhat better current distribution. Less center radiation occurs because of partial cancelingeffects in the doubled-back wires. It would be wise initially to cut each center-back wire 12 inches (0.31 m) longer than shown to allow tuning. Pruning each center-back wire by 1 foot (304.8 m) increases the resonant frequency by roughly 100 kHz in the 40-meter version. No. 14 Copperweld wire was used throughout.

Performance

The antenna described performed up to expectations and has prompted many complimentary reports. The SWR curve information shown in Table 3 was attained using a 75-ohm coaxial feed line without any matching network. The SWR would be lower if a 50-ohm line or if a matching network were used. Forward signal strength as compared to the strength to the rear showed a maximum 30-dB difference as measured by means of the receiver S meter. Typical difference, as finally adjusted, was 12 dB.

The gain of the C-T quad over a comparable double Zepp (full wave on 40) was typically at least one S unit better when both antennas were at a height of 45 feet (13.7 m). The Zepp was better on very short skip but the C-T quad consistently was favored on longer skip. When elevated to 65 feet (19.8 m), the C-T quad gave outstanding results. At this height, two S-units improvement over a vertical antenna at distant points was noted many times.

The C-T quad may be built for any band. It should be most useful in meeting requirements for an inexpensive easily tuned wire-beam antenna where the full-size quad is unsuitable for mechanical or economic reasons. This material was orginally presented in *QST* by Roger Sparks, W7WKB.

References

[1]Pinner, "The Short Quad," *QST*, February 1964.
[2]Courtier-Dulton, "Some Notes on a 7-MHz Linear-Loaded Quad." See Chapter 4.

A Nearly Full-Size, Rotatable, 2-Element Quad for 80 Meters

We don't expect readers to rush right out and duplicate this antenna system — but this doesn't mean a lot of hams wouldn't like to!

Many hams have been thinking about new 80-meter antennas to improve their DX capabilities. On the higher amateur frequencies, antenna gain is relatively easy to acquire with a compact Yagi or quad. It is much harder in this respect on 40 meters. On 80 meters, however, the problem of securing any increase in antenna gain over a conventional dipole or groundplane is very difficult indeed, particularly where space is limited. Rhombics, Vs or multi-element collinear arrays become completely impractical on the normal urban or suburban lot. Even a phased vertical array is hard to handle on a city lot, and too often, performance is marginal because it is impossible to install an optimum ground system at such a location.

The author solved this problem at his QTH by constructing an almost full size 2-element quad for 80 meters. A unique tuning arrangement permits this antenna to be operated at any frequency within the 80-meter band with an SWR of close to 1. While certainly not adaptable to every individual's situation or pocketbook, a description of this antenna should be of interest to many amateurs. As far as it is known, this is the first and only set of beams, on one rotatable tower, covering all bands from 2 to 80 meters. By connecting the 2 quad elements in series to form a rotatable, bidirectional loop, the frequency coverage has been extended to 160 meters.

At this point, a few comments might be in order on the circumstances which led to this sizeable 80-meter antenna project. In June 1965, the writer returned to ham radio activity after a QRT of almost 30 years. During the first year of operation, the antennas of K3JH were a conventional commercial tribander for 10, 15 and 20 meters mounted on a 60-foot (18.3-m) tilt-over tower, and a trapped, inverted V for 40 and 80 meters. A bit of DX chasing soon led to the conclusion that better antennas were needed. Since the QTH is on a suburban lot about 175-feet (53.3-m) square, and heavily wooded, a "Christmas tree" array seemed the best alternative. Consequently, in the summer of 1966, with a great deal of help from Bob Scully, W2FXN, a 115-foot (35-m) rotary tower was installed along with full-size monoband beams for all bands from 2 through 40 meters.

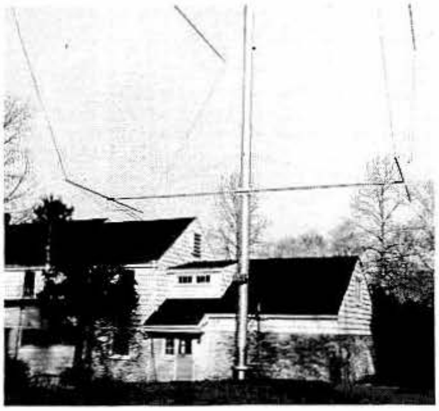

The bottom support of the quad elements is shown in this view, along with the tuning-network box, which is mounted on the mast.

This rotary steel tower is 16 inches (0.41 m) OD at the base with 1-inch (25-mm) thick walls. The tower rests in a 1/2-inch thick steel bearing-tube 15 feet (4.57 m) long and 20-inches (0.51-m) OD. The bearing tube is imbedded in a block of concrete 8 × 8 × 16 feet (2.44 × 2.44 × 4.88 m) which weighs 70 tons. The tower tapers to 5-1/2 inches (140 mm) at the top, and was designed to carry nine full-size monoband beams through 125-mph winds.

Initially, the antenna complement was as follows

2/6meters — Vert. groundplane at 115 feet (35 m)
6 " — 6 el at 113 feet (34.5 m)
2 " — 15 el at 109 feet (33.2 m)
20 " — 5 el at 104 feet (31.7 m)
15 " — 5 el at 96 feet (29.3 m)
10 " — 6 el at 86 feet (26.2 m)
40 " — 3 el at 77 feet (23.5 m)

A rotary ball-bearing ring and clamp at the 70-foot (2.13-m) level on the tower was used to support one end of an inverted L for 80 meters. This long wire, which could be used also on the higher frequencies, was fed through a Matchbox at the base.

Performance on 40 through 2 meters with this antenna system was generally excellent. The L and inverted V on 80, however, left much to be desired, especially when compared to the antenna performance on the other bands.

In attempts to improve this situation, several different 80-meter antennas were installed at various times between 1966 and the spring of 1969. Included in this effort were a top-loaded groundplane and a pair of phased verticals. None of these antennas provided sufficient improvement in DX performance on 80 to be considered satisfactory.

With the advent of the 5-Band DXCC, the question of how to do better on 80 meters again became a matter of concern. It was at this time that the concept of a quad of some sort, to be mounted on the rotary tower, began to emerge. Since 40-meter quads had been constructed previously, and were mechanically feasible, the thought at first was to build a half-size quad for 80 meters, using loading coils. Preliminary calculations indicated the possibility of achieving some gain over a dipole in addition to the obvious advantage of being able to rotate the array.

It was quickly determined that a half-size quad mounted at 57 feet (17.4 m) would fit underneath the 40-meter Yagi. Construction could be quite conventional using fiberglass X frames in each element, and a spacing of 0.15 wavelength.

Design Considerations

The proposed design was discussed with a number of amateurs including Jim Lindsay, W0HJ, Dunc Carter, W5IOU, and Claus Moeller, DL7CM, who were most helpful with advice and suggestions. A search of the literature disclosed that a number of loaded antennas of various types had been built. In most cases, however, performance had been judged empirically, and there was little in the way of specific comparative data on the performance of loaded versus unloaded configurations, or versus a reference dipole. Because of this, the decision was made to build and test a 14-MHz model of a half-size quad. This would enable a direct comparison between the performance of a miniaturized quad and other antennas. Of particular interest was a comparison with a dipole, since this was the more normal antenna used by amateurs on 80 meters.

Henry Pemberton, W3PN, who had become interested in the project, provided the X frames and supports from an old 20-meter quad, for use in the test model. Tom Consalvi, W3EOZ, provided some suitable coil stock for the test design. With this help, the 14-MHz model was quickly constructed. No trouble was experienced in pruning the coils and resonating the loops. Except for one bad piece of insulation on one coil, which promptly burst into flame when rf power was applied to the antenna, the driven element could be fed with a full kilowatt at an SWR of 1. The SWR, however, would rise sharply when the coils, which had no protection from the weather, became wet from rain. All testing, therefore, had to be done on dry days.

Since the 80-meter version was to be mounted on the tower one-quarter wavelength above ground, the 14-MHz model was mounted for the tests at the same relative height. Standard procedures for tuning up the quad were used, and will not be detailed here. Impedance of the loaded loop was measured at 60 ohms, so it could be driven nicely with a 50-ohm line. The bandwidth was 125 kHz, measured between frequencies each side of the resonant frequency where the SWR was 3:1. This was the result expected, due to the heavy loading.

Extensive on-the-air comparisons were made between the model and various other antennas. These included a dipole at the same height above ground as the model, a 3-element tribander at 60 feet (18.3 m), two different inverted Vs at 50 feet (15.2 m) and a 5-element monoband Yagi at 104 feet (31.7 m). The regular antenna switching arrangements in the station were such that almost instantaneous comparisons could be made between the different antennas, minimizing QSB differences in the receiver. In the transmit mode, many amateurs compared signal differences between the test model and the other antennas. It is impossible to list the calls of all those who helped in this way.

After considerable testing, adjustment, readjustment, and minor configuration changes in the model, the results were found to be quite discouraging. The simple conclusion was that a loaded half-size quad had insufficient gain compared to a dipole to warrant proceeding any further with an 80-meter version.

Further Exploration

Before abandoning the 80-meter project, however, we decided to explore some practical ways to increase the size of the elements, preferably without using the normal-type quad X frames, which would be difficult if not impossible to handle when the antenna grew too large. This thinking developed the concept of using the 14-MHz boom at 104 feet (31.7 m) to support diamond-shaped elements at the top, and using a boom and spreaders rather than an X frame to hold the elements at the center. The bottom of the quad elements could be supported easily with a relatively light boom and spreaders near the base of the tower.

In order to reduce the size of the center boom and spreaders, the first model was kite-shaped with an included angle at the top of about 50 degrees. Unfortunately, the results were poor, and compared to a dipole there was no gain. We concluded that this configuration was so squashed together that the antenna was acting like a dipole instead of a quad. A series of configurations was then drawn up on paper. The boom and spreader sizes were varied, but the included angle at the top of each diamond element was kept at 75 degrees or more. This helped to pin down a configuration which would be mechanically feasible, and which would be a reasonable compromise among the various considerations to be taken into account. The one selected is shown in Fig. 22.

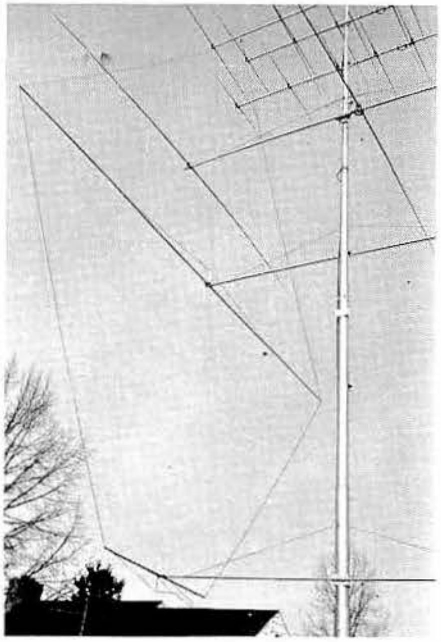

A detailed view of one of the quad elements, showing the method of supporting the loop.

Because these elements are almost full size at 4 MHz, making a 20-meter model quite large, it was decided to build and test a model of this configuration on 15 meters. The test procedures for this model were the same as those previously described for the 20-meter model of the half-size quad. This model showed substantial gain compared to the reference dipole at the same height. Also, the model mounted only 15 feet (4.6 m) above the ground compared favorably in performance with the tribander at 60 feet (18.3 m). The 5-element monobander at 96 feet (29.3 m) consistently provided better gain, as was to be expected. The impedance of the driven element was about 80 ohms. Bandwidth between SWR points of 3:1 was 250 kHz, or about twice what was measured on the half-size model.

Mechanical Considerations

After a lengthy test period, during which many on-the-air comparisons were made, the results were good enough to make the decision to build an 80-meter version of this antenna. Because of the size of the proposed antenna, considerable thought had to be given to the mechanical design to assure longevity comparable to the other beams which are rated for 125-mph winds. The wire in the quad elements is no. 12 stranded copperweld. All other metal in the antenna is aluminum or stainless steel. The two quad elements are suspended from the 20-meter boom at 104 feet (31.7 m). The insulators are five glazed porcelain knobs. Spacing is one-eighth wavelength, 36 feet (10.97 m). The 20-meter boom is 46 feet (14.02 m) long, and is made from 4-inch (102-mm) OD, 1/4-inch (6-mm) wall T-6 aluminum tubing at the center, and similar material tapering from 3-1/2 inches (89 -mm) to 3-inch (76-mm) OD at the ends. A 1/4-inch stainless steel cable supports the boom 18 feet (5.49 m) out from each side of the tower. Originally, the plan was to slide the quad elements down this cable from the tower. Unfortunately, the steel plates holding the Yagi elements to the boom on the other beams were found to be rusting because of poor plating. Rather than dismantle the beams, a crane was brought in so that these plates could be cleaned and painted by a man carried up in a boatswain's chair. At the same time, the elements were hung from the 20-meter beam, and the center boom and spreaders for the quad were installed with relative ease.

Dimensions

The 36-foot (10.97-m) center boom at the 57-foot (17.37-m) level of the tower is made of a single 24-foot (7.32-m) section of 3-1/2-inch (13-mm), 1/4-inch (6-mm) wall, T-6 tubing with 7-foot (2.13-m) sections of 3-inch (76-mm) OD pipe telescoped and bolted at each end. The boom is supported from the tower with a 3/16-inch (4.76-mm) stainless-steel cable. The spreaders at each end of this boom are 70 feet (21.34 m) long. Each consists of a 24-foot (7.32-m) piece of 2-1/2-inch (64-mm) OD, 1/8-inch (3-mm) wall, T-6 tubing at the center, with two short pieces of 2-1/8 × 1/8-inch (54 mm × 3-mm) tubing telescoped and bolted at each end to make up a length of 30 feet (9.14 m). To complete the spreaders, 20-foot (6.10-m) sections of 1-1/2- and 1-1/4-inch (38- and 32-mm) fiberglass poles are attached to each end of the aluminum centerpiece. Use of the fiberglass reduces weight and eliminates a one-quarter wavelength piece of metal from the middle of the quad element. A 1/8-inch (3-mm) cable and strut supports each spreader to minimize sag, which is very slight as may be noted in the photograph. A welded aluminum T structure and stainless steel clamps are used to hold the spreaders on to the ends of the boom. A stainless steel clamp and Teflon grommet is attached to the ends of each spreader to hold the element wires in place. The only function of the spreaders is to hold the two opposite sides of each quad element apart without too much fore and aft flopping around when the antenna is rotated. The length of the spreaders, wind loading, and safety factor dictates the heavy mechanical design.

The lower boom, which is 23 feet (7 m) off the ground, is made of a 24-foot (7.32-m) section of 2-1/4 × 1/4-inch (57- × 6-mm) T-6 tubing with shorter 2- × 1/4-inch (51- × 6-mm) pieces telescoped and bolted at each end to make the total length of 36 feet (10.97 m). To maintain symmetry, 17-foot (5.18-m) spreaders are attached to each end of the lower boom. These spreaders are made of 1-1/2-inch (38-mm) fiberglass. They are attached to the boom with aluminum angles and stainless-steel clamps. A light stainless-steel cable and strut holds the lower spreaders firm against the pull of the quad-element wires, which are attached to each end of the lower spreaders with stainless-steel clamps. The wires are taped along the fiberglass almost to the center of the spreaders, and then go off at right angles to the tuning box, which is mounted on the tower 13 feet (4 m) above ground. The distance from the end of the lower boom to the tower is about 18 feet (5.5 m), and the 36 feet (11 m) of wire which connects the quad proper to the tuning box represents loading.

Tuning

Each element, including the 36 feet (10.97 m) of connecting wire just mentioned, resonates at 4050 kHz. Thus, a small amount of additional inductive loading is required to tune the antenna to resonance within the 80-meter band. This is accomplished by putting two relatively small motor-driven coils in series with each loop of the antenna. The reversible motor for each pair of coils in each loop is controlled by a two-way toggle switch mounted on the antenna control panel in the shack. In the case of the driven element, tuning is accomplished simply by applying power to the antenna and then adjusting the loading inductances by the motor-control switch to the point where the SWR is minimum, usually very close to 1. Tuning the reflector can be accomplished by turning the back of the quad toward a distant signal, and adjusting the reflector loading coils for minimum received signal. This seemed a bit cumbersome to do each time the operating frequency was changed from cw to ssb and vice-versa, so a microswitch was added to the reflector coils and motor assembly. Each revolution of the coils flicks the microswitch, which actuates a light on the control panel in the shack. By counting the blinks of the light, and referring to a chart which shows the resonant frequency of the reflector versus the number of turns of inductance in the loop, the reflector can be tuned to any desired frequency. With this tuning scheme, it is quite easy to tune the quad for best performance at any point in the 80-meter band. The bandwidth between the 3:1 SWR points is close to 200 kHz, so retuning is required only for large frequency changes within the band. Frequency flexibility otherwise is quite good.

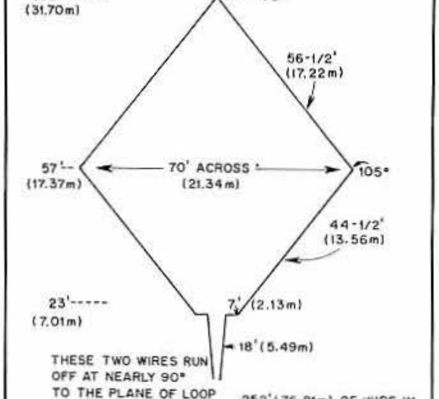
QUAD LOOP DIMENSIONS
104'
(31.70m)
75°
56-1/2'
(17.22m)
57'
(17.37m)
70' ACROSS
(21.34m)
105°
44-1/2'
(13.56m)
23'
(7.01m)
7'
(2.13m)
18' (5.49m)
THESE TWO WIRES RUN OFF AT NEARLY 90° TO THE PLANE OF LOOP TO THE TUNING BOX
252' (76.81m) OF WIRE IN EACH LOOP, INCLUDING LEADS TO TUNING BOX

Fig. 22 — Dimensions of one of the quad loops.

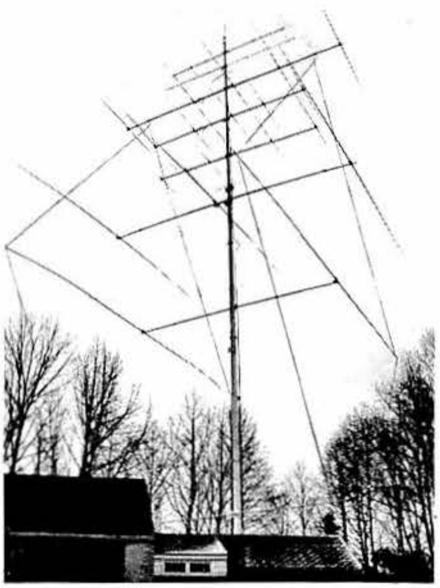

Here is the complete antenna system, 2 through 160 meters, all rotatable, directional arrays.

The impedance of the driven element measures about 75 ohms. The antenna is coupled to the transmitter through a balun and 50-ohm line. The difference between the resonant frequency of the driven element and the reflector, when the latter is tuned for maximum front-to-back ratio, is in the order of only 1-1/2 percent. This probably results from the quad being only a one-quarter wavelength above the ground. The front-to-back ratio is about 20 dB, and the front-to-side, 50 dB. All measurements have an indicated gain over a dipole in the order of 6 to 7 dB. No degradation in performance, or SWR, has been observed during heavy rain. The small loading coils are protected from the weather, of course, in their aluminum tuner box, so the problem, which was encountered with unprotected load coils in the 20-meter test model, has been eliminated.

Tests were conducted with the second element tuned as a director instead of a reflector; there was no noticeable improvement in gain, whereas the front-to-back ratio was diminished. It is felt that the best results are obtained with the reflector.

Some thought was given to driving both elements at a 135-degree phase difference to obtain a cardioid pattern. It is felt, however, that the conventional quad pattern, with deep nulls on each side, and a reasonably good front-to-back ratio, is more desireable for DX. So, nothing further has been done with such a phasing arrangement.

Results

DX results on 80 meters have improved considerably since the new antenna was put on the air. On cw, the gain seems to drop off somewhat as compared to the phone end of the band. Probably this is due to the fact that the relative size of the antenna is smaller at the cw operating frequencies. Nevertheless, reports usually are from one to several S units higher than other U.S. stations with comparable power input, and conventional antennas. In most instances during pile-ups in a recent cw contest, it took only a call or two to get through, whereas, previously, the station was usually last. It was found also that contacts could be made earlier, as the band opened. In several instances like this, it was amazing to get a response from a DX station, and then to hear other U.S. stations calling without success.

Reports from overseas on ssb are outstanding. During a recent phone contest, several dozen DX contacts were made in just a few hours — everyone on the first call.

During the day, on 80 meters, it is quite easy to work into Canada or the Carolinas with good signals at both ends of the circuit. On a dipole, very often the other station is completely unreadable, if not inaudible. This points up the fact that one notices great improvement in reception with the quad compared to a dipole, which on 80 is just as important as being able to transmit a better signal.

Once it was determined that the quad worked well on 80 meters, the possibility of operating the antenna on 160 meters was explored. Since the loops independently resonated at 4050 kHz, and there was considerable inductive loading available in the coils, it was felt that it might be possible to resonate the antenna on 1.8 MHz by putting the two loops in series. Actually, with the load coils tuned to minimum, the series-connected loops resonated at 1775 kHz because of the mutual coupling. By shortening out the coils entirely, the pair of loops were resonated at 1805 kHz. Therefore, the SWR at the low end of the 160-meter band is very low.

Domestic reports on 160 meters have been excellent. Consistent directional effects have been noted by several observers when the two-turn loop is rotated. The pattern is the typical figure eight, with deep nulls off each side of the loop. It remains to evaluate how well the antenna works on 160-meter DX.

Following completion of the antenna in September, there have been several severe storms, with winds gusting as high as 65 mph. These velocities barely moved the wires and spreaders around. It is apparent that the heavy construction is adequate to handle the winds of much higher velocity for which the antenna was designed. Although it is probably one of the largest quads in existence, it certainly appears as though it will stay up a long time, even through rough weather.

In conclusion, this project could never have been completed successfully without the help of many domestic and overseas amateurs. Their reports, advice, and assistance will always be appreciated. Although this has been a sizeable and difficult antenna task, a great deal of satisfaction has been derived from carrying it through from concept to on-the-air operation. The simple lesson has been relearned with regard to antennas — that one cannot get something for nothing. Mini-size antennas are better than no antennas, but there is no substitute for full-size antennas if one wants full-size results. This material was originally presented in *QST* by Joe Hertzberg, N3EA.

The Folded Mini Quad

A full-wavelength loop can take a variety of forms. Despite the shape suggested here, the author claims good results.

An unusual type of full-wavelength wireloop antenna is in use here on 2 and 15 meters. The design, which the author had not seen previously, provides an antenna that occupies minimum space; the antenna is omnidirectional, vertically polarized, broadbanded and requires no groundplane.

It evolves from a full-wavelength loop.[1] The loop, if fed on the side (x-y) as in Fig. 23A, radiates a vertically polarized signal. The radiation pattern is bidirectional, with maximum radiation broadside to the plane of the loop and with the highest field strength in the plane A-A′.

The loop can be deformed into a square, such as in the elements of a quad, and without much loss of efficiency into a rectangle (Fig. 23C). Vertical polarization, bidirectional signal pattern and maximum radiation from the plane A-A′ are retained.

A 90 degree bend at A-A′ and two further right angle bends at B-B′ and C-C′ transforms the flat loop into a cube. The circumference is still a full wavelength and each side of the cube measures 1/10 wavelength. Again, vertical polarization is retained, but the appreciable bidirectional radiation from the sides of the cube adjacent to A-A′ leads to a nearly omnidirectional radiation pattern.

My 2-meter antenna was constructed from aluminum clothesline wire; the length in feet was chosen by 1005/f MHz. It is fed with 50-ohm coaxial cable through a Pomona Electronics no. 1699 adapter and no matching device is used. The SWR is less than 1.2:1 from 144 to 146.5 MHz and climbs to 1.4:1 at 148 MHz. The radiation pattern was measured with a simple field-strength meter, using a half-wave dipole as pickup element and a 20-kΩ resistance in series with the 20-μA meter to improve linearity.[2] The front-to-back ratio is 1.8 dB with no prominent sidelobes apparent, but a rather sharp 3-dB dip is caused by the "shadow" of the feed line and hardware.

Performance

Fed with 10 watts and used as an indoor antenna, it gives consistent communication with 2-meter fm base and mobile stations within a 10- to 15-mile radius. It performs better on receive and transmit than a 1/4-wave whip at the same location.

A 15-meter version of the antenna was then constructed and suspended in the attic with string and tacks. It is operated with 100 watts and the rig loads almost as well as the 2-meter version, but with a slightly higher SWR. The first contacts were with California and it has since given excellent performance to the West, Midwest and South as well as to South America and Europe.

The close proximity of antenna sections radiating out of phase may lower the efficiency over that of a loop, yet, in terms of space requirement, construction cost and ease of installation (no radials!) and tuning, the antenna is hard to beat. As a full-wave loop it should have a slight edge over a vertical. The author recommends further experimentation and modification, e.g., into the even more space-saving configuration of a cylinder. This material was originally presented in *QST* by Max Blumer, WA1MKP, SK.

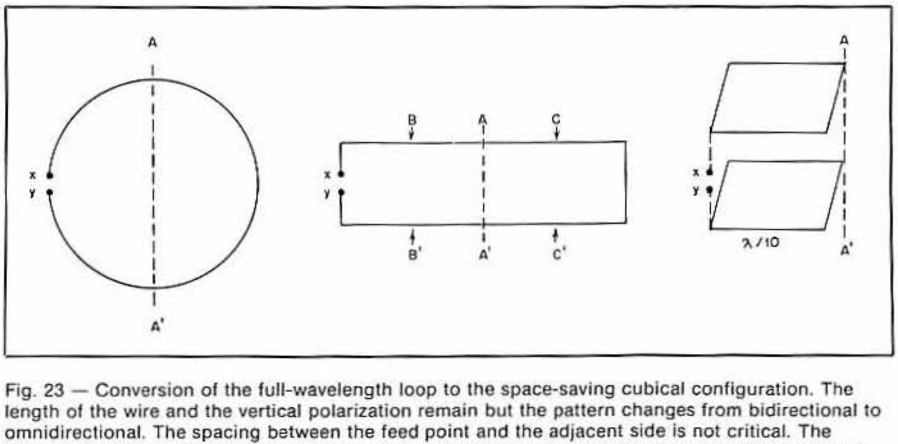

Fig. 23 — Conversion of the full-wavelength loop to the space-saving cubical configuration. The length of the wire and the vertical polarization remain but the pattern changes from bidirectional to omnidirectional. The spacing between the feed point and the adjacent side is not critical. The spacing on 2 meters is about 1-1/2 inches (38 mm), and on 15 meters is about 4 inches (102 mm).

References

[1] *The ARRL Antenna Book*, 13th ed., 1974, p. 65.

[2] *The Radio Amateur's Handbook*, 46th ed., 1969, p. 553.

A Convenient Stub-Tuning System for Quad Antennas

The cubical quad has been a popular antenna with amateurs for a number of reasons — relatively light weight, small turning radius, low cost, and good DX performance at rather low heights when compared to other antennas. The authors' experience with quads has demonstrated, however, that best possible performance requires careful parasitic-element tuning. Cutting elements according to various "established" length formulas has often yielded less than satisfactory results. While adjustments for maximum forward gain are generally uncritical, a mediocre front-to-back ratio often results unless time and care are taken to tune the parasitic elements. Factors such as spacing between elements, proximity of the antenna to ground, or influence of other objects (including the other concentric loops in a multiband array) require that the electrical length of the parasitic elements be adjusted using empirical methods if optimum performance is to be realized from the system. It was felt that if one was going to the trouble of erecting a good antenna system, it was worth seeing to it that the antenna was delivering the performance of which it was capable.

These considerations led to the development of the parasitic-element tuning system described here. While the scheme is applicable to quads in general, whether rotary or fixed, the original intended application was in a fixed two-element, full-size 40-meter quad. The descriptions here apply to this 40-meter system, but the general information allows one to follow the same scheme with a quad for any band.

A 40-Meter System

This 40-meter antenna consisted of two non-rotatable full-wave wire loops suspended from a boom which was mounted near the top of a tower. Spacing between elements was 22 feet (6.71 m). The corners of the loops were tied down to convenient anchor points by means of ropes, thereby eliminating the need for spreaders. With the elements thus fixed in place, the antenna was oriented in one general direction and could not be steered elsewhere. A driven-element and reflector combination was used.

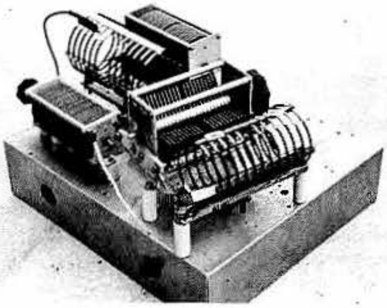

The matching network built by the authors for the modified 40-meter system. Relays, used in the switching circuitry, are not visible. The builder is left to choose a suitable network of his own to meet his operating requirements. Basic design information can be found in *The Radio Amateur's Handbook* or *The ARRL Antenna Book.*

Early experiments with the antenna demonstrated the need for careful reflector tuning. Several trips were made up the tower to alternately prune and add wire to the reflector before the antenna exhibited good performance at the frequencies of interest. But when tuned properly, the quad really performed! In many instances signals which were inaudible while using an 80-foot (24.4-m) high reference dipole were perfectly readable with the quad (at a median height of about 100 feet [30.5 m]). Reception was further enhanced by rejection of unwanted signals off the back of the quad.

Now if only the quad could be made rotatable. . . . Unfortunately, a full-size rotating system (spreaders would be needed) was out of the question. The cost, difficulties of construction and installation, and the severe effects of New England winters discouraged such ideas. The authors, being avid contesters and DXers, were willing to settle for the next best alternative — keeping the fixed quad, but electrically switching its pattern.

The solution involved a tunable stub attached to one of the elements to lengthen or shorten it electrically. With the appropriate stub switched in, the single element could be made to look like either a reflector or director and thus cause the pattern to be reversed. In order that adjustments be made quickly, easily, and from a convenient location (such as the shack or the bottom of the tower), it involved more than the conventional short stub often used with quads. Rather, the situation called for a long stub — long enough to reach the desired remote location — terminated in a variable reactance. Tuning the element, then, involved nothing more than adjustment of the value of reactance which could be provided conveniently by means of a variable capacitor.

The electrical principle is simple. One adds inductive reactance to lengthen an element electrically, or capacitive reactance to shorten it. Typically, the former arrangement is used with quads in a driven-element and reflector combination in which both elements are cut to the same physical length. A stub, shorted at the end, is attached at the center of the bottom section of the reflector. By varying the position of a shorting bar along the stub, the amount of inductive reactance presented to the element can be varied and the electrical length altered correspondingly. Because the stub is short in length, this adjustment must be made at the antenna itself, preferably near or at its final height.

Now, if the stub is cut to some different length which happens to be longer than in the above case, the parasitic element can still be made to see the proper amount of inductive reactance (if it is a reflector) provided the stub is terminated properly with some other reactance (in general, something other than a short). Varying the reactive termination is then equivalent to moving the shorting bar along the conventional stub. This is a consequence of the impedance-transforming property of transmission lines. The long stub can be treated as a transmission line. If we know its characteristic impedance, its length, and the impedance that must be seen at the parasitic-element end of the line, the Smith Chart can tell us what kind of termination is required at the opposite end.

A reflector cut to the same physical dimensions as the driven element — self-resonant at the desired operating frequency — must "see" roughly 150 ohms of inductive reactance at the terminals where the stub is attached. Although little information is available about the use of directors with quads, it is assumed initially that a director, also cut to the same length as the driven element, must see a roughly equal but opposite amount of reactance, or about 150 ohms of capacitive reactance. If one desires to make the termination at the end of the stub a variable capacitor for convenience of tuning, which was the case in the 40-meter system under consideration, the problem can be rephrased slightly: what length of line of a given characteristic impedance is required so that a variable capacitor (a 10- to 350-pF unit was used) causes the parasitic element to see the above reactances? This is easily solved with a Smith Chart. Two somewhat different systems evolved from these ideas. The second, a modification of the first, will be described last.

The design procedure for the 40-meter version went as follows: The driven and parasitic elements were cut to the same length, given by the more or less standard formula, $L = 1005/f$, where L is the length in feet, and f the frequency in MHz. This resulted in a loop circumference of 143 feet, 6 inches (43.74 m) at a frequency of 7.0 MHz. In the practical installation, the loops were triangular in shape rather than

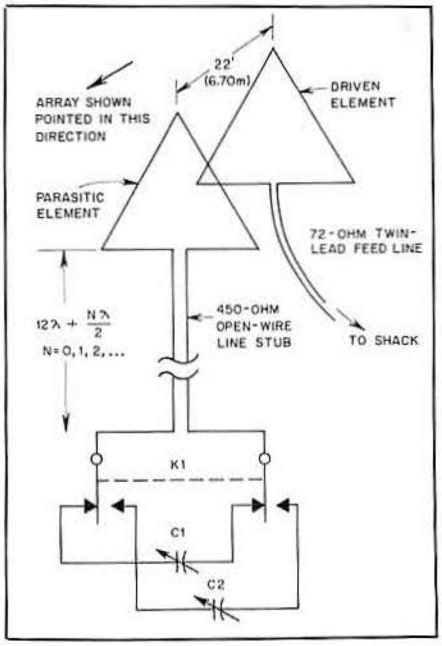
22'
(6.70m)
ARRAY SHOWN POINTED IN THIS DIRECTION
DRIVEN ELEMENT
PARASITIC ELEMENT
72-OHM TWIN-LEAD FEED LINE
12λ + Nλ/2
N=0, 1, 2, ...
450-OHM OPEN-WIRE LINE STUB
TO SHACK
K1
C1
C2

Fig. 24 — Configuration of 2-element switchable 40-meter quad. Triangular loops are shown, although square or diamond loops can be used too. Both elements are cut to the same dimensions: 143 feet, 6 inches (43.74 m) for 7.0-MHz operation. The selection of either C1 or C2 by means of K1 determines which direction the array is pointed. Switching components K1, C1 and C2, can be installed inside the shack or on the ground. The stub line is cut long enough to reach them. Feeding the driven element with 72-ohm Twin Lead results in a low SWR.

C1 — 10- to 350-pF air-variable capacitor, adjusted for director operation as described in text.

C2 — Same as C1, except adjusted for reflector operation.

K1 — Dpdt relay (Potter and Brumfield KA11DG or equivalent).

square or diamond. Shape is relatively uncritical — performance is approximately equivalent for the various geometric configurations, and ease of installation was considered more important. The triangle has the advantage of requiring only three tie-off points — one at the top where it hangs from the boom and two at the bottom — whereas the other two forms require support at four corners. Furthermore, the triangle requires less vertical height — approximately 41 feet (12.5 m) on 40 meters, assuming equal-length sides — as opposed to the diamond which needs about 50 feet (15.25 m). The square requires the least vertical height but is the most difficult to support mechanically. The resultant configuration is shown in Fig. 24.

The stub line was made from 450-ohm open-wire transmission line because of its relatively low cost in commercially made form and its low-loss characteristics. The reactance range of the 10- to 350-pF capacitor by itself at 7 MHz was then plotted on the Smith Chart as shown in Fig. 25. (Recall that $X^L = 2\pi fL$ and $X^C = 1/2\pi fC$ where X^L and X^C represent values of inductive and capacitive reactance respectively, given in ohms, L and C the corresponding values of inductance in henrys and capacitance in farads, and f, the frequency in hertz.) The low-capacitance end of the range is indicated as point B in Fig. 25 with the Smith Chart normalized to 450 ohms in this example, while the other extreme comes out at point A. If the capacitor was then connected at the end of a 0.12-wavelength section of the 450-ohm line (approximately 16 feet [4.88 m] in open-wire line), the impedance seen at the other end of the line would vary between the range of slightly greater than +j300 to -j300, represented by points C and D. This would meet the +j150 and -j150 reactive loading requirements of the parasitic element for reflector and director operation, respectively. Actually a stub length of 0.12 wavelength, plus any multiple of a half wavelength, would work too, because any half-wavelength section of transmission line merely repeats at one end the impedance that appears at the other. Thus, the distance to the desired remote location dictates the minimum required line length. Another advantage of using 450-ohm line, as opposed to a line with a lower characteristic impedance, becomes apparent when working examples on the Smith Chart — for a given capacitor tuning range, the 450-ohm line produces a greater range of reactance change at the antenna end of the line. A 300-ohm line would serve almost as well, however, providing only slightly less tuning flexibility.

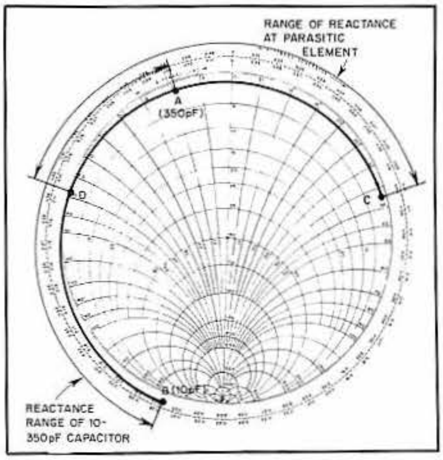

Fig. 25— Example of Smith Chart calculations required to arrive at correct stub dimensions. A 10- to 250-pF capacitor is used as the variable-reactance element which terminates the 450-ohm transmission line used for a parasitic-element stub. The chart is normalized to 450 ohms. The arc between points A and B represents the range of reactance of the capacitor as it is tuned between the extremes of its range. If the capacitor is connected to a 0.12-wavelength long stub (plus any multiple of one-half wavelength) with a 450-ohm characteristic impedance, the range of reactance represented by the arc between points C and D is presented to the parasitic element. This range is sufficient to cause a properly cut parasitic element to look like either a director or reflector. Further discussion of Smith Chart calculations can be found in the latest edition of *The ARRL Antenna Book*.

The stub was terminated at a dpdt relay which was controlled from the station operating position and which was used to select one of two variable capacitors — one adjusted to provide reactance for reflector operation and the other for director operation, as shown in Fig. 24. In this manner, the pattern could be flipped around 180 degrees instantly from the shack. The driven element was fed directly with 72-ohm transmitting-type Twin-Lead. A low SWR was obtained with no special provisions made for matching.

The director and reflector tuning should be done empirically, as was stated before. Experience has shown that quads tune rather broadly for maximum forward gain. It was found that fairly large changes in the settings of the variable capacitors did not alter the gain significantly. It is best, therefore, to tune for maximum front-to-back ratio, on which the settings had considerably more effect. On-the-air signals arriving from the rear direction can be nulled by tuning the capacitors. The nulls should be fairly pronounced if the

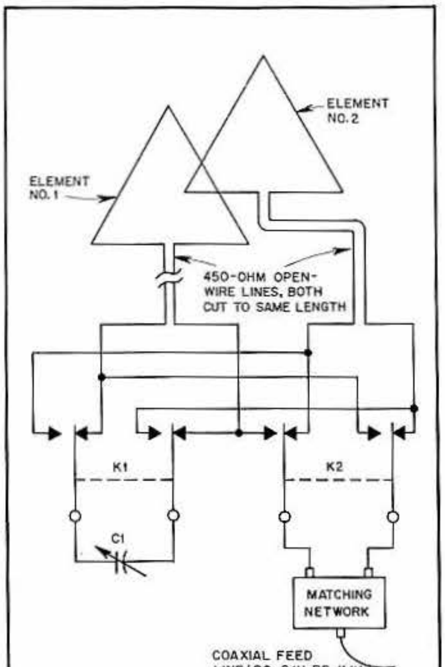
ELEMENT NO. 2
ELEMENT NO. 1
450-OHM OPEN-WIRE LINES, BOTH CUT TO SAME LENGTH
K1
K2
C1
MATCHING NETWORK
COAXIAL FEED LINE (RG-8/U, RG-11/U, ETC.) TO STATION

Fig. 26 — Configuration of modified 2-element switchable 40-meter quad. The lines from each element are both cut to the same length, which can be the same as the parasitic element line in Fig. 24 or other suitable length. Both elements are cut to the same dimensions as in Fig. 24. The positions of K1 and K2 determine whether an element is fed as a driven element or terminated as a reflector. The array pattern can be switched by determining which element operates as a reflector or driven element. Only one tuning capacitor is required, but a matching network is needed to match a coaxial feed line.

C1 — 10- to 250-pF air-variable capacitor, adjusted for reflector operation.

K1, K2 — Dpdt relays operated simultaneously, or both may be replaced by a single four-pole, double-throw relay (Potter and Brumfield KA11DG or equivalent).

system is set up as described here. The exact amount of attenuation of signals off the back of the antenna varies from signal to signal. It was observed that the 40-meter antenna exhibited the greatest front-to-back ratio for signals arriving from distant locations (at low radiation angles). Alternatively, field-strength measurements can be made to determine the proper settings for minimum radiation or reception off the rear of the antenna. In this fashion, one capacitor is adjusted for director operation and the other for reflector operation.

When making large excursions in operating frequency on 40 meters, it is necessary to touch up the tuning. A director tuned for operation in the low end of the cw segment is self-resonant approximately in the middle of the phone portion and provides little usable gain at those higher frequencies. Retuning is done easily, however — something which is not possible using fixed-tuned elements.

A Modified 40-Meter System

The 40-meter system described above provided very good results over a period of months, including stateside and DX contest work and casual DXing. After some thought, however, a modified system was constructed. The new configuration provided some advantages over the orginal scheme.

The same principle of operation was retained. The new system shown in Fig. 26, used only a reflector-type parasitic element, but the switching system now allowed either element to operate as the driven element or reflector. In this manner, the antenna pattern could still be reversed by choosing the appropriate element for the desired function.

Several factors prompted the change. It was found with the original system that a considerably higher front-to-back ratio could be had using a reflector rather than a director. The latter yielded no more than about 10-dB discrimination against signals received off the back of the antenna, even with careful tuning, whereas with the former the figure was approximately as high as 25 dB. Although most available literature made little mention of quad performance using a director, this observation agreed with what little published data could be found on the subject.[1] The reflector-only model also had the advantage that just one tuning adjustment was required, since one stub reflector served both elements in an identical fashion. Furthermore, because the system behaved the same no matter which direction the array was switched, the SWR did not change when switching, and the transmitter was thus always tuned properly. With the original scheme the antenna feed-point impedance changed when switched between director and reflector operation. Although the resultant SWR was not objectionable in either case (it did change, though), it was necessary to retune the transmitter after switching, or tolerate a moderate amount of mistuning.

The design procedure for the new arrangement is basically the same as before. Both elements are cut to the same length. Only the stub requirements for a reflector need be considered now in doing the Smith Chart calculations. Once the stub line length is determined, two such identical-length lines, both made from the same type of transmission line, are cut, one for each element. A four-pole, double-throw relay, or two dpdt relays operated simultaneously, are required in the hookup shown in Fig. 26, to switch the two lines between the parasitic-element termination and the main feed-line matching network.

The new system has one disadvantage, however. A matching network is required between the transmitter and driven element because of the mismatch between the main feedline, which is presumably coax, and the driven-element open-wire feed line. The original system required no special matching. This presented no problem in the 40-meter system considered here since a suitable network was easily constructed. The builder is left to choose a suitable matching network of his own as there are numerous possibilities to take care of the varied matching requirements which may arise under different situations. Depending upon one's preferences and requirements, the modified scheme may or may not prove more attractive.

Additional Remarks

A few general comments can be made at this point. Although a variable capacitor has been designated thus far as the termination for the stub, there is no reason for not using a coil or some other combination of inductance and capacitance if it will work with a stub length which happens to be convenient. The builder can choose for himself. The 10- to 350-pF capacitor described earlier had approximately 0.05-inch (1.27 mm) spacing between plates. While it is difficult to predict the voltages present in the parasitic element (and across the capacitor), no arcing between plates was observed during transmitting periods, even while running one kilowatt transmitter dc input. This suggested the plate spacing was adequate.

The lines from each element should be brought away perpendicular to the elements as much as possible, and open-wire line should be kept reasonably clear of nearby objects, especially metal ones. Also, the velocity factor of the line being used (0.95 for "ladder" line) must be remembered when computing physical lengths.

The authors erected a single diamond-shaped 80-meter quad loop with its bottom close to the ground. The antenna was tried as both a vertically and horizontally polarized radiator. For DX work, the vertically polarized version yielded unquestionably superior results in terms of greater signal strengths at long distances and the beneficial feature of rejection of unwanted high-angle radiation from stateside stations when receiving. A parasitic element to accompany the single loop was not tried. These observations suggest interesting possibilities for vertically polarized ground-mounted quads for low frequencies.

The ideas presented here are applicable to any quad antenna, but should be especially useful in designing large arrays which cannot be turned physically. The design procedure is simple in any case. The feature of continuously variable tuning eliminates the guesswork in cutting fixed-length, fixed-tuned elements. The system will serve well in any situation where, say, climbing the tower to work on the antenna is not convenient.

A fixed, switchable array is an asset to the contestant or DXer who finds it necessary to change antenna headings in a hurry. From the northeastern U.S., a single such array provides good coverage of Europe and Africa in one direction and of the western U.S. and the Pacific in the other. If two suitable supports spaced a reasonable distance apart are available, two arrays of this type, perpendicular to each other, can provide a wide range of coverage. This material was originally presented in *QST* by John E. Kaufmann, WA1CQW and Gary E. Kopec, K8RX.

---

[1] "Technical Topics," *Radio Communication*, February 1973, p. 101.

Some Notes on a 7-MHz, Linear-Loaded Quad

If you like to experiment with antennas, here are enough new ideas about the construction of quad antennas to keep you going.

Most amateurs who have a serious interest in working DX in the 3.5- to 28-MHz bands would like to have antennas providing them with forward gain on all of these bands. Today, a triband trap Yagi or cubical quad may be employed to achieve this objective on 14, 21 and 28 MHz. Rotary beams are, for all practical purposes, out of the question for 3.5 MHz, but it is possible to achieve some gain with a 7-MHz rotary beam without having to pay a fortune for a supporting structure.

The Hy-Gain linear-loaded 2-element Yagi or the Mosley inductively loaded beam for 7 MHz can achieve this objective. The writer, however, prefers quads to Yagis, based on his own experience. Difficulties often arise when one seeks to mount beams for several different bands on the same tower without marked loss of performance on any band. The primary problem is, of course, that a full-sized cubical quad for 7 MHz is a monstrous structure. Such antennas have been erected, but they are not suited for the average amateur with a limited budget.

The first step in building a 7-MHz quad is, therefore, to design a single 7-MHz quad loop of reasonable size which is mechanically strong enough to survive most weather conditions.

Linear Loading

As a general rule, a reduction in antenna size to two-thirds of full size can be effected without a noticeable drop in performance, providing care is taken to limit the losses introduced by the loading devices. This means that one can think in terms of a loop for 7 MHz with 24 feet (7.32 m) on a side instead of 36 feet (10.97 m) on a side. An additional factor in choosing a 24-foot-per-side length was that the longest bamboo spreaders that the writer could obtain were 18 feet (5.5 m). Attempts were made to lengthen these spreaders, but results were not satisfactory mechanically.

If a 24-foot loop is to be used, the next questions are where to put the loading and what form of loading to use. A full-sized quad loop has symmetrical current distribution. The current maxima occur in the middle of the horizontal sections, and the current minima develop in the middle of the vertical sections. The current distribution in the vertical elements is such that most vertical radiation is canceled out. It is desirable to retain this current distribution in the shortened loop. It is also desirable not to put the loading in the center of the horizontal elements, since maximum radiation occurs at the current maxima, and it does not make sense to cut out the portion of an antenna that does the most effective work. The writer, therefore, chose to put the loading in the vertical sides.

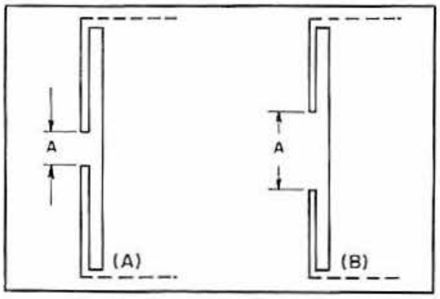

Fig. 27 — A method for linear-loading of one side of a quad element.

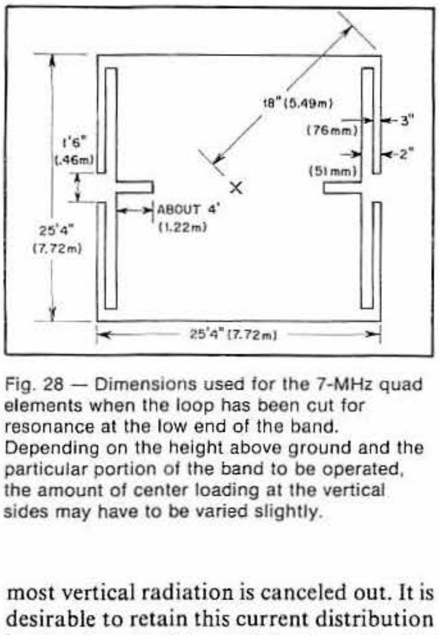

Fig. 28 — Dimensions used for the 7-MHz quad elements when the loop has been cut for resonance at the low end of the band. Depending on the height above ground and the particular portion of the band to be operated, the amount of center loading at the vertical sides may have to be varied slightly.

There are two popular methods of loading, stubs or coils. The writer dislikes coils for loading since, if they are wound with a heavy-gauge wire (desirable in order to avoid undue losses), they are bulky and difficult to weatherproof. The choice was therefore stub loading. It would have been possible to insert a stub at each corner of the loop, tied back to the spreaders. But, the current flowing in such stubs would be unbalanced, and radiation in unwanted directions would take place from the stubs themselves. There is also the consideration that if the same spreaders are used for 14-, 21- and 28-MHz loops, stubs at the corners of the 7-MHz loop would have to be mounted well clear of the loops for the other bands.

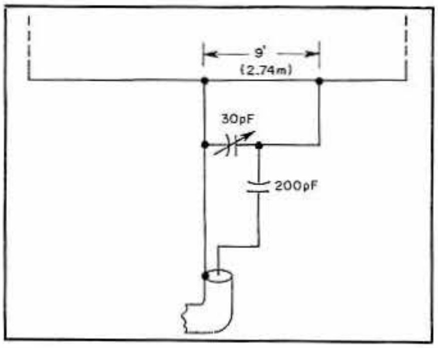

Fig. 29 — Diagram of the matching section used on the 40-meter driven element. The gamma capacitor employs a 100-pF air variable and a 100-pF high-voltage silver mica connected in parallel.

The original linear-loaded 7-MHz quad element was tried by the writer in a four-band quad some years ago. A single stub in the center of each vertical side was used. The loading stubs were made 14 feet (4.27 m) long and were tied on to a point near the center of the boom. Performance of this quad on 7 MHz was excellent, but the 21-MHz quad mounted on the same boom was almost useless, probably due to pickup at 21 MHz by the 7-MHz loop with subsequent reradiation. An attempt was made to find an efficient method of loading which did not involve long stubs "floating around." The next step was to try folding the wires in the vertical sections back on themselves as shown in Fig. 27A. Current distribution either side of the center part of the vertical elements is symmetrical, as desired. This change resulted in a loop that was resonant near 7 MHz, although its resonant frequency was slightly high. Shortening of the stubs slightly, as shown in Fig. 27B, was tried. This had virtually no effect on resonant frequency. This is no doubt because the increased end-loading effect of the portion A almost equalled the reduction in stub length. Short stubs (abour 4 feet [1.22 m] long) were then

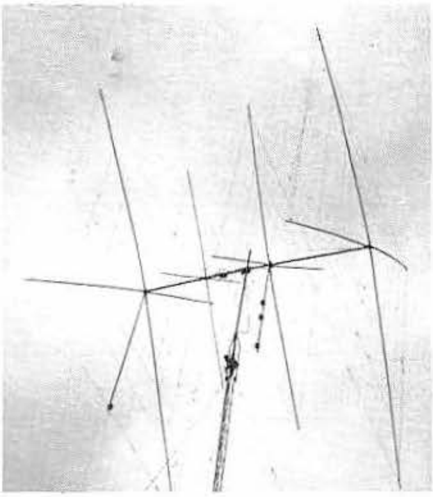

G3FPQ's compact 4-band quad has 2 elements for 7 MHz, 3 elements for 14 MHz, and 4 elements for 21 and 28 MHz. Pairs of Yagi-type directors are used in place of the usual loop director for the three higher frequency bands.

inserted at the center of the vertical sides, at right angles to the vertical sides as shown in Fig. 28, and resonance at 7 MHz was achieved.

A four-band quad for 7, 14, 21 and 28 MHz was then constructed. Much to the writer's delight, there appeared to be no noticeable interference with the pattern of the 21-MHz quad by the 7-MHZ elements. The loading effect of the stubs would be different at 21 MHz since they would include current maxima; the result is to give a loop which, while resonant at 7 MHz, does not appear to have a third-harmonic resonance in or near the 21-MHz band.

Yagi-Type Directors

The reader can probably see from the photograph that the directors for 14, 21 and 28 MHz consist of pairs of stacked Yagi-type elements level with the driven elements. There are two pairs of Yagi directors for the 10- and 15-meter elements at approximately 6-foot, 6-inch (2 m) spacing, and one pair of Yagi directors for 20 meters at approximately 13-foot (4 m) spacing. The overall length of the directors is 31 feet, 3 inches (9.52 m) on 20 meters; 20 feet, 10 inches (6.35 m) on 15 meters; and 15 feet, 2-1/2 inches (4.63 m) on 10 meters. Performance of the author's array on 10, 15, and 20 meters at the moment is little, if any, better than an ordinary 2-element quad. The writer has only recently had an opportunity to start checking out the directors to see what adjustment they need. Current indications are that they require lengthening, but there is a good deal more experimental work to be done. Anyone who wants to try this system must be prepared to do a good deal of cut and try to tune the directors, if any real increase in performance over a standard two-element quad is to be achieved.

One also has to bear in mind that when dealing with multiband quads the resonant frequency of Yagi-type directors is likely to be affected to a far greater extent by the proximity of other elements than would be the case for a quad-loop director. This is because a Yagi director is a comparatively high-Q element. From the theoretical point of view, the drawback to quad-loop directors is that the reactances involved are such that there is a comparatively low degree of mutual coupling between the quad-driven element and the quad-loop director. This, of course, reduces the efficiency of the loop as a director. This problem should not apply to the same extent with Yagi-type directors, and, in theory at least, they should prove to be more efficient than quad-loop directors. The question now is whether practice follows theory or not. This writer won't have an opportunity to conduct any detailed experiments until the summer. When deciding the original length for the Yagi directors, allowance was made for a rise in frequency of approximately 100 kHz. But, the only way to cope with the effects on the director resonant frequency of the proximity of other elements is by field-strength measurements, coupled with a good deal of cut and try. As in any multiband quad, separate feed lines should be used for each band to minimize interaction between bands.

The author's quad has been up for six months now, and final tuning remains to be done to the 7-MHz reflector loop to achieve maximum front-to-back ratio. But the results, even on a cut-and-try basis, are well worth the work that went into the array. This material was originally presented in *QST* by David. L. Courtier-Dutton, G3FPQ.

Spider Quad Mount: Simplified

There seems to be a growing interest in the spider quad. The photograph shows a version of a hub or "spider" mount. Using a 2-foot (0.61-m) length of 2-inch (51-mm) diameter steel pipe, a mount can be constructed easily. Two cuts, 10 inches (254 mm) long, are made through the diameter at each end of the pipe. This leaves 4 inches (102 mm) in the middle for the boom. The cut ends are then bent to an angle to suit the desired spacing. The unit shown in the photograph is bent to 108 degrees. The spreaders may be attached to the boom mount with hose clamps, bolts or wire.

The beauty of this design is that it can be made to any size desired. Also, it can be enlarged to a 4-element version, as shown in Fig. 30.

Fig. 30

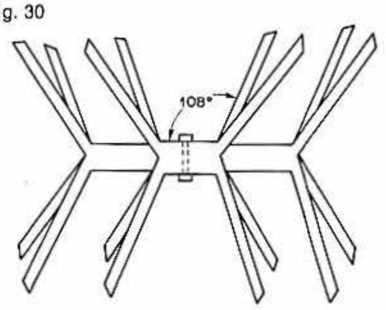

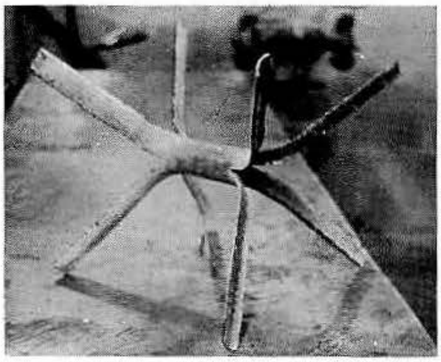

One of these spider mounts is in use at KP4DJI (constructed by KP4DIO and KP4DJI) using bamboo spreaders. The results have been excellent. This material originally presented in *QST* by Lynda B. Crowley, KP4DIP.

Two-Band Quad Matching

While experimenting with a 2-element, 2-band cubical quad antenna the problem of matching the antenna to a 50-ohm transmission line arose. The antenna is a "plain vanilla" quad built around information in W6SAI's book, *Quad Antennas*. The author's method of feeding the antenna *did not* produce fantastic front-to-back ratio or forward gain. It merely made the standing-wave ratio more acceptable to the 32S-3B used at the home QTH.

After building the 2-band quad, the author checked it with a General Radio 1606 impedance bridge and found the 10-meter antenna to be about 118 ohms and the 15-meter antenna to be about 91 ohms at resonance. This is 2.36:1 and 1.82:1, or approximately 2:1 for a 50-ohm transmission line. The feed system described here made the VSWR much nearer the desired 1:1.

The heart of the system is a ferrite-core autotransformer. The measured impedance ratio is very close to 50:107 ohms. The insertion loss is less than 0.2 dB from 15 to 30 MHz when terminated in 107 ohms. When it is terminated in 50 or 200 ohms the insertion loss is less than 0.35 dB from 15 to 30 MHz. Several transformers were built before these figures were obtained, and after a suitable design was achieved it was connected to the quad as shown in Fig. 31. You do not have to sacrifice F/B ratio for a more acceptable VSWR by detuning the reflector stubs; simply adjust the tuning stubs for maximum F/B ratio and then design the transformer for the proper impedance ratio. Also, lowering the VSWR places a larger portion of the desired band within an impedance range the average pi-section tuner will accept.

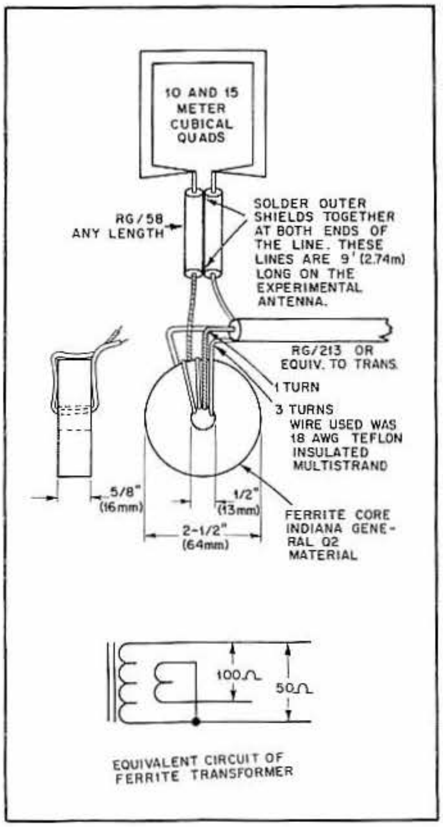
10 AND 15 METER CUBICAL QUADS
RG/58 ANY LENGTH
SOLDER OUTER SHIELDS TOGETHER AT BOTH ENDS OF THE LINE. THESE LINES ARE 9' (2.74m) LONG ON THE EXPERIMENTAL ANTENNA.
RG/213 OR EQUIV. TO TRANS.
1 TURN
3 TURNS
WIRE USED WAS 18 AWG TEFLON INSULATED MULTISTRAND
5/8" (16mm)
1/2" (13mm)
2-1/2" (64mm)
FERRITE CORE INDIANA GENERAL Q2 MATERIAL
100Ω
50Ω
EQUIVALENT CIRCUIT OF FERRITE TRANSFORMER

Fig. 31 — Two-band quad matching.

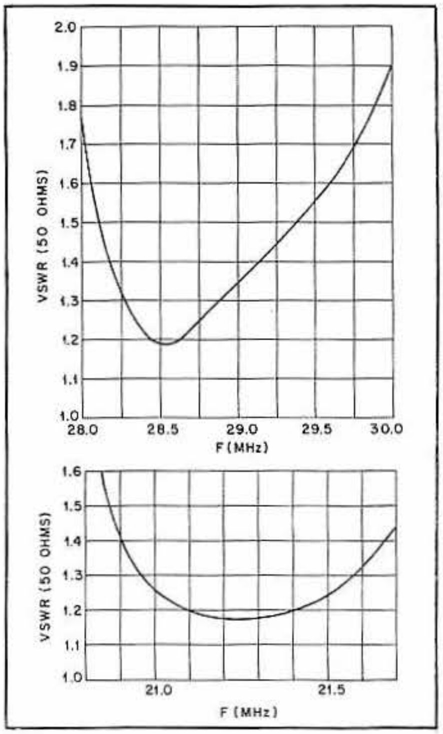

Fig. 32 — Standing-wave ratio curves taken with a laboratory-type VSWR bridge. Measurements were made at the transmitter ends of the transmission line with the antenna array mounted on a steel tower 44 feet (13.4 m) above ground.

In answer to on-the-air inquiries: Yes, performance of the antenna varies considerably from one end of the band to the other.

This antenna feed system is a broadband device, so don't forget to use a low-pass filter. This material was originally presented in *QST* by Toney L. Magnino, W5MVK.

Chapter 4

Miscellaneous Antenna Types

Here is a potpourri of useful antennas for the hf-band enthusiast. This chapter contains information on how to build and operate loop antennas, limited-space antennas and variations of the quad. Once again, the work of numerous *QST* authors has been collected for a timeless volume of useful construction details that can benefit amateurs whose operating objectives are numerous and varied.

For the most part, the antennas described in this section are simple and inexpensive to build. The performance characteristics are consistent with the operating needs of ragchewers, contesters and DX-chasers alike, assuming that the builder does not strive to have the loudest signal on the amateur bands.

Beat the Noise with a "Scoop Loop"

Operators on 160 and 80 meters fight a common foe — noise. Atmospheric QRN is compounded by the presence of man-made noise. Loop antennas for receiving may offer an economical solution to your noise problems. Here are some timely details.

If you've been tempted to massage your receiver with a 10-pound hammer because you can't copy some of the weak signals being heard by your colleagues, you may need a receiving loop! A high ambient noise level on 160 meters has stifled the author's DX work for several years, and more than once was ready to capitulate in favor of the man-made and atmospheric noises that the 75-foot (22.9-m) shunt-fed vertical responded to. "Grrrr," would be the snarl as the author listened to stations with Beverage antennas: They were handing out RST 579 signal reports to DX stations, and the author couldn't find the DX signals in the noise, let alone copy them.

Something had to be done. It was a matter of giving up on 1.8-MHz efforts toward DXCC, or putting forth a zealous attempt toward an improved signal-to-noise ratio. Some effective outdoor receiving antennas were erected (half-wavelength open loop — parallel to the ground, and a 200-foot [61-m] end-fed wire a few feet off the ground), but they did not always respond well to DX signals: It depended on band conditions. Also, there was the danger of receiver front-end damage caused by unwanted excitation of the large receiving antennas (resonant) during the transmit period. They were in close proximity to the vertical antenna, which was used for transmitting purposes. Lethargy prevented the author from installing a protective circuit at the receiver front end, and as a consequence the writer "blew" a JFET preamp and two protective diodes in the front end of the FT-301D transceiver.

Revisiting the Past

While for many years experimenting with small receiving loops for 40, 80 and 160 meters, the efforts were casual. The best of the loops was an outdoor one of the shielded variety, five feet (1.52 m) on a side, and fashioned from RG-59/U coaxial cable.[1] However, it was necessary to go out-of-doors each time to rotate it, and the author wasn't enthusiastic about investing in a rotator to use with the loop.

Recently, it seemed that some work could be done with indoor types of receiving loops. After all, loops of that kind have been around since radio was invented, and plenty of amateurs are using them effectively. After scanning available literature (Terman, Kraus, Jasik and Keen),[2] the author set about the task of building a framework for the chosen design. This time a multiturn loop would be tested (Fig. 1).

Construction

The photograph shows the general structure of the loop. In Fig. 2 are the dimensions of the wooden frame used to contain the antenna wires. There is nothing magic about the size of the antenna: Two pieces of wood that were stored in the workshop were used, each being 48 inches long, 1-1/2 inches wide and 3/8 inch thick (1.22 m × 38 mm × 10 mm).

Arbitrarily, the author decided to make L1 of Fig. 1 a five-turn winding. Holes were drilled (five) in each end of the cross-frame to accommodate the loop turns and hold them in place. Actually, there should be six holes at the bottom point of the frame to permit the final end of L1 to be secured during the threading-through process.

Five additional holes were drilled (Fig. 2) 3/4 inch (19 mm) below the holes for L1. These are used to contain L2, the one-turn output link of the antenna. L2 requires two holes at the bottom member of the frame — again to provide an anchor point for the ends of L2. Alternatively, V-shaped notches can be filed on the ends of the cross-frame. The wire for L1 can be laid in the grooves rather than threaded through the holes shown in Fig. 2. L2 should, however, be wound inside L1, through the five holes shown in the drawing.

A mounting base for the assembled antenna can be made to the builder's specifications. This writer used a piece of 3/4-inch (19 mm) plywood which has dimensions of 30 × 12 inches (0.76 m × 305 mm). A 1-inch (25 mm) diameter hole was drilled in the center of the board to accept an 18-inch (457 mm) length of 1-inch diameter dowel rod. The latter is lashed to the framework of the loop to enable the

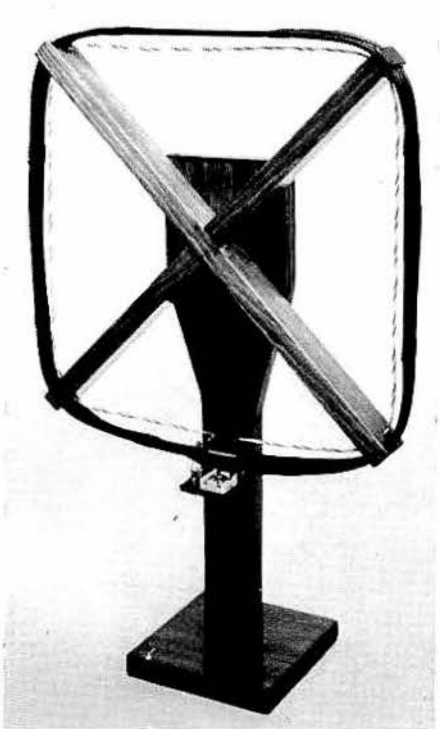

The silhouette is of an indoor version of the four-turn shielded loop described in this article. A wooden framework is used to support the main loop winding and the one-turn pickup link. The resonating capacitor is mounted on an aluminum L bracket along with a phono jack and an 860-pF disk capacitor. The latter is used between the output link and the phono jack instead of a trimmer capacitor. The framework was built by Technical Secretary Marian Anderson's OM, Bob. It has been stained and coated with clear lacquer.

operator to rotate the loop as desired. Although this is a "kluge" type of assembly, it serves its purpose adequately. A *carpenter* should be able to conceive a more practical base and support system for the antenna!

Number 22 insulated hookup wire was used for the loop turns. There is nothing critical about the wire size, but it is best to use the larger gauges. This will make the loop more rugged.

Performance

A variety of feed systems were tried before the author settled for the method illustrated in Fig. 1. Direct connection to the terminals of L1 required careful balancing to assure a uniform loop-response pattern. Unbalance causes pattern distortion, and if it's bad enough the maximum response can occur off the broad side of the loop rather than in the plane of the antenna.

Addition of a one-turn coupling loop, L2, proved to be the easiest method for assuring good balance. Bandwidth of the loop is dependent upon the Q. The latter can be controlled by the amount of capacitance used at C2: the lighter the coupling, the greater the loaded Q. Loop resonance is effected by means of C1. Since there is interaction between L1 and L2, it is necessary to readjust C1 each time C2 is tweaked. In this application it should be stressed that C2 is not used for the purpose of obtaining an SWR of 1. It is purely a *coupling* device. Hence, a 50-ohm condition does not necessarily exist at the loop output terminals.

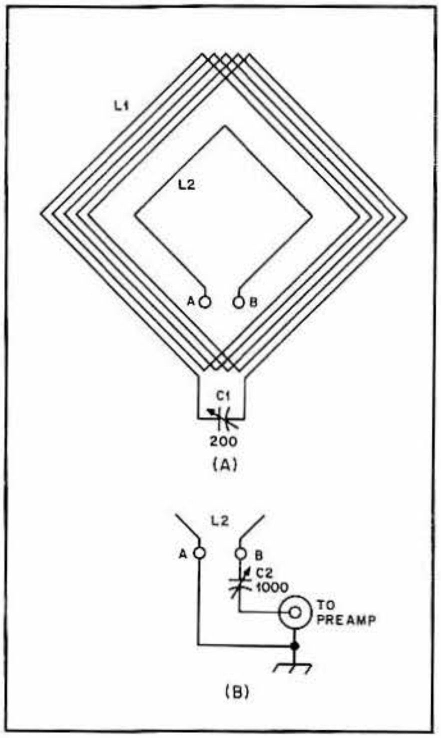

Fig. 1 — Schematic diagram of the receiving loop. C1 and C2 are mica compression trimmers. The illustration at A shows the 1-turn link inside the 5-turn one. At B is an expanded diagram of the feed point for L2.

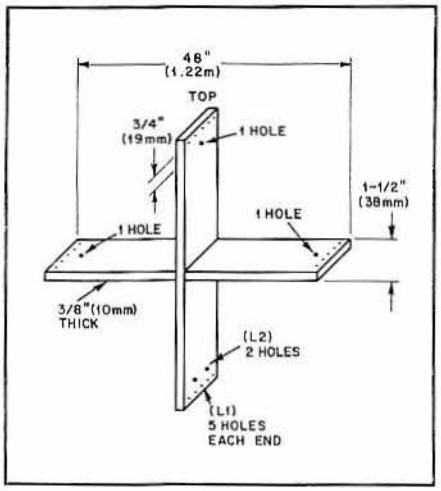

Fig. 2 — Dimensional drawing of the wooden frame which contains the Scoop Loop. Further information concerning the structure is given in the text.

During initial experiments the pickup loop, L2, was connected directly to the 50-ohm input port of a 40-dB broadband preamplifier which was designed by W7ZOI (Fig. 3). Without C2 in the circuit the coupling was excessive. This resulted in low Q, and the loop bandwidth was approximately 150 kHz at the 3-dB points of the response curve. After C2 was added and adjusted to provide a bandwidth of 25 kHz, the loop gain increased 10 dB. Alternatively, one might make L2 smaller and locate it closer to the center of the loop frame in order to lighten the coupling between the two windings. This could be a somewhat tedious mechanical procedure: The writer's inherent lassitude favored the addition of C2 for the purpose!

Performance Characteristics

A test range was set up for checking the null properties and relative gain of the loop. The equipment was arranged in an outdoor location, well away from power wiring and other station antennas. A crystal-controlled weak-signal source (2N2222A oscillator and 9-volt battery) was placed some 100 feet (30.48 m) from the test loop. It contained a 36-inch (0.91-m) whip antenna (vertically mounted) for radiation of the test signal. The 75-foot (22.86-m) transmitting vertical was shorted out at the feed point to prevent it from being resonant and affecting the loop antenna during the test period.

A 50-ohm Tektronix step attenuator was placed in the coaxial line between the loop and the receiver. The attenuator was used to measure accurately the changes in S-meter reading on the FT-301D transceiver which was used as a detector during the measurements.

A null depth of —21 dB was obtained with the loop. An extremely sharp null was noted. As the broad side of the loop was moved toward the signal source, the null began slowly, then fell deeply to —21 dB over the final 10 degrees of rotation. No evidence of pattern distortion was observed. The null was dead off the broad side of the antenna, and maximum response was directly off the thin side of the loop.

A comparison between the 75-foot (22.86-m) transmitting vertical (shunt-fed tower, top loaded and with buried radials) and the wire loop indicated that on groundwave signals the loop (without preamp in line) was —18 dB respective to the vertical. This suggested that in a practical installation a single JFET or dual-gate MOSFET preamp would provide unity gain with the transmitting vertical.

Using the Loop Indoors

The author's home has aluminum siding, and as is the case with all modern homes there is a network of water pipes, phone wires, TV lead-in conductors and house wiring. The author wondered how well the loop would perform within all of these unwanted signal obstacles, some of which form loops by themselves! The ham shack is at ground level inside the family room, and the loop was placed on a stand at floor level. The clutter of unwanted conductors disturbed the directivity of the loop somewhat, but the signal-pickup ability of the antenna was not impaired.

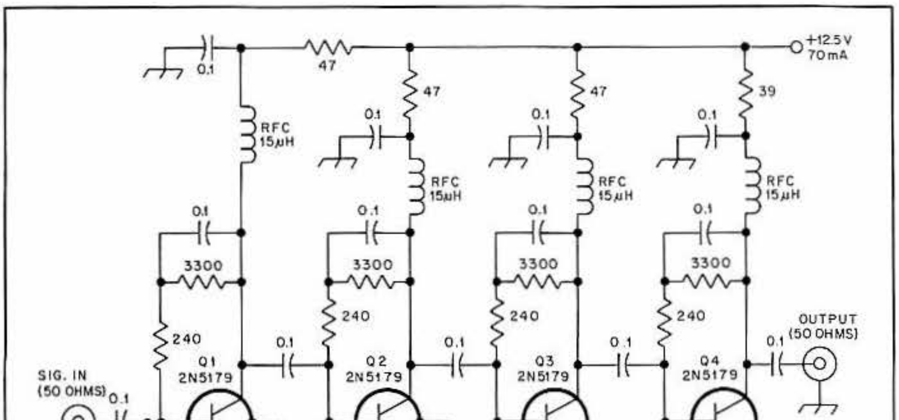
+12.5 V
70 mA
0.1
47
RFC
15µH
39
3300
240
0.1
SIG. IN
(50 OHMS)
Q1
2N5179
Q2
2N5179
Q3
2N5179
Q4
2N5179
OUTPUT
(50 OHMS)
1000
10
150
120
91
(A)
EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF) ; OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS ; k = 1 000, M = 1000 000.
1.8 – 1.9 MHz
50 OHMS
68
220
50 S.M.
C1 200
1000 S.M.
C2 200
L1
L2
-1dB
5.8
870
BWL = 100 kHz
IL = 0.5 dB
f₀ = 1.85 MHz
S.M. = SILVER MICA
(B)
(C)

Fig. 3 — The W7ZOI broadband 40-dB preamplifier is shown at A. Capacitors are disk ceramic or chip types. Resistors are 1/4-w composition. The upper 3-dB point of the amplifier is at 65 MHz. Heavy feedback is used to stabilize gain and provide a 50-ohm characteristic. At B is a 1-dB attenuator pad which can be used between an antenna of unknown impedance and the filter at C to provide a proper load for the filter. The Butterworth band-pass filter at C is suitable for use with the preamp at A. It will help reject out-of-band signals to improve the IMD characteristics of the preamp. When the preamp is used with nonresonant receiving antennas, it will be necessary to employ the filter to protect the preamp from signals across its response range (bc band through vhf). The preamp and filter would be fine for use with Beverage antennas. C1 and C2 are mica trimmers. L1 and L2 consist of 31 turns of no. 22 enameled wire on Amidon T68-6 toroid cores (5.1 µH).

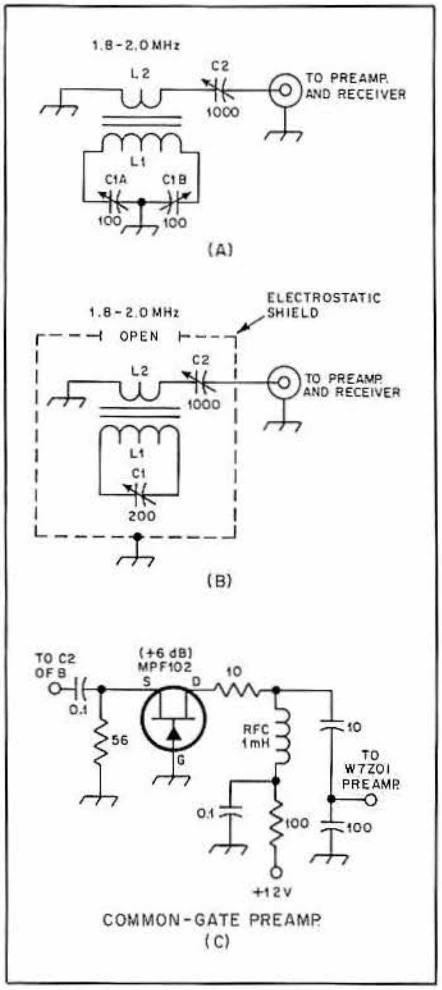
1.8 - 2.0 MHz
L2
C2
1000
TO PREAMP. AND RECEIVER
L1
C1A
C1B
100
(A)
ELECTROSTATIC SHIELD
OPEN
C1
200
(B)
TO C2 OF B
0.1
(+6 dB)
MPF102
S
D
G
10
56
RFC 1mH
TO W7ZOI PREAMP.
100
+12V
COMMON-GATE PREAMP.
(C)

Fig. 4 — The drawing at A shows the circuit of a 7 X 5/8-inch (178 X 16-mm) ferrite-rod loop (see text). At B is the final version which employs an electrostatic shield. C1 and C2 are mica compression trimmers (see text for other details). The circuit at C was used ahead of the W7ZOI preamplifier to improve the system noise figure.

A step attenuator was inserted between the W7ZOI preamplifier and the station receiver so that unity gain could be established between the loop and the 160-meter vertical (22 dB of gain had to be gotten rid of). An antenna switch was added to the system to permit rapid comparisons between the antennas.

Extensive tests proved that the loop outperforms the vertical by a substantial margin when it comes to digging weak signals out of the noise (man-made and atmospheric). The author has copied PA0HIP, PY0ZAE, G3MYI, W6RW, W6BLZ, W5AB and others perfectly Q5 on the loop (indoors) when they couldn't be found, or were unreadable on the vertical. It is only fair to say that on some occasions the vertical was the better antenna for receiving. Propagation conditions and ambient noise levels at a given time will determine which antenna can provide the best reception. Typically, the ambient noise level from the W1FB vertical is S3 on a receiver that is calibrated to indicate S9 on a 50-μV signal.[3] The same receiver shows no noise response when the loop is connected to it. However, static crashes (atmospheric noise peaks) are as loud with the loop as they are with the vertical when a storm is in the immediate area. Noise from distant storms can be discriminated against by turning the loop null toward the storm front. It appears that the major part of the author's ambient noise is man-made in nature. At times it peaks as high as S8.

When the loop is used indoors it responsds to unwanted 15.750-kHz TV spurs, and when appliances are actuated the transients kick the receiver S meter rather hard. This is because the loop is contained within loops formed by house wiring. This type of interference has not been observed while using the vertical antenna, even though it is affixed to the rear of my house. To prove that the house wiring was causing the problem with TV birdies and transients, the loop was located out-of-doors for two nights. The interference vanished!

Some Other Loop Experiments

Ferrite-rod loops have been used to advantage for many years (bc-band radios, direction-finder receivers for boats and aircraft, etc.), so it seemed worthwhile to investigate a model of that type of antenna before concluding the tests. Two versions were tried, each with a preamplifier.[4] The first consisted of a single 7 × 5/8-inch (178 × 16 mm) Amidon Assoc. ferrite rod on which was wound a 40-turn solenoidal coil of no. 22 insulated hookup wire. The loop was tuned to resonance by means of a 100-pF per section split-stator variable. A four-turn output link was wound over the center

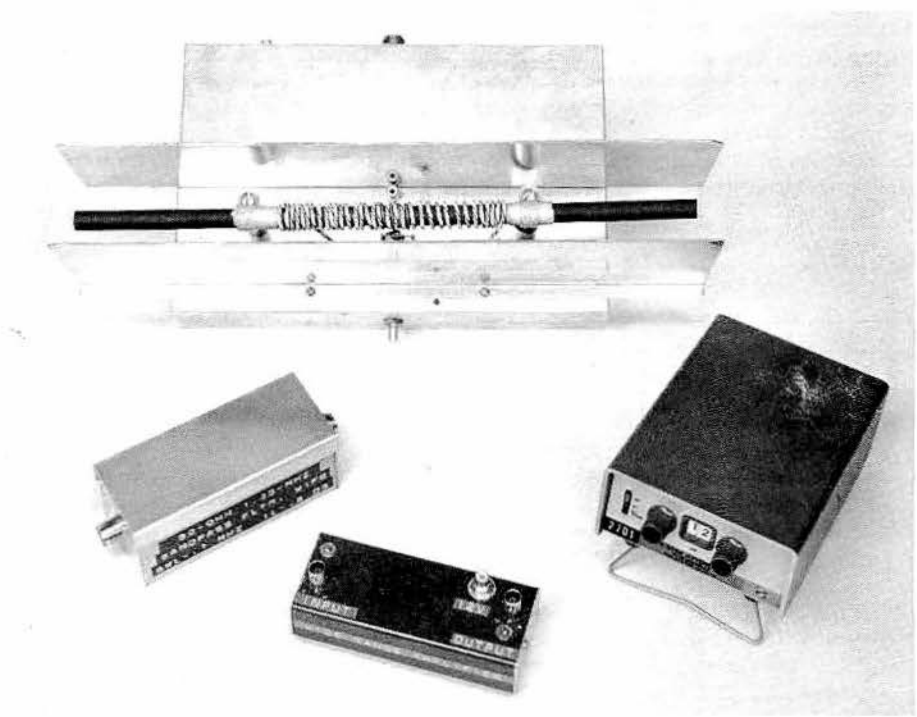

View of the ferrite-rod loop and test setup accessories. In the foreground (left) is the Butterworth band-pass filter. Below it at the center is the broadband preamplifier, and the Tektronix 2701 step attenuator is at the far right.

of the primary winding. The latter occupied approximately 3 inches (76 mm) of the center portion of the rod. A small aluminum chassis served as the foundation for the antenna, and the rod was spaced some 3 inches (76 mm) above the chassis. Performance tests showed the antenna to be less efficient than the large 5-turn wire loop. As referenced to the 75-foot (22.86-m) vertical, the ferrite loop was —38 dB. Furthermore, it was prone to the same induction effects which plagued the wire loop when both were used indoors. The rod-loop null was dismal — some 10 dB!

An improved model of the loop was built by using epoxy cement to join together a pair of 7-inch (178-mm) ferrite rods, end to end. The sensitivity (μV per meter) increases with the length of the rod, but is dependent upon the length of the coil which is wound on the rod. A multilayer coil contained at the center of the rod should be considerably better than a solenoidal coil spread over the major portion of the rod: The shorter the coil, the higher the Q of the antenna.

The author did not have Litz wire available, and was unable to build a multilayer coil, so a short solenoidal winding was placed on the center 5 inches (127 mm) of the 14 × 5/8-inch (356 × 16 mm) rod. It consisted of 32 turns of no. 12 insulated wire, spaced one wire diameter between each turn. The ferrite rods are made of Q2 material ($\mu = 125$).

Fig. 4 shows the schematic diagram of the rod loop. A four-turn output link is wound over the center of L1. C1 is a mica compression trimmer which tunes the loop to resonance. C2 is used as a coupling capacitor to control the loop loading.

It seemed worthwhile to try shielding the loop, so a pair of aluminum L brackets were bolted to the chassis as shown in the photograph. Each bracket is 16 × 3-3/4 inches (406 × 95mm). The shields are spaced 1 inch (25 mm) either side of the loop.

The top of the U-shaped channel must be left open to prevent the shield from acting as a shorted turn. Proof of the effect can be seen by shorting across the open end of the trough (over the center of the rod) with a screwdriver. The received signal will vanish! An electrostatic shield, when used with a loop antenna, should minimize induced fields from nearby conductors and precipitation static, thereby making the system better for indoor use.

Rod-Loop Tests

The same outdoor test setup was employed for checking the performance of the rod loop. The nulls occur off the ends of the rod, and maximum response is found off the broad sides of the rod. A null of —23 dB was obtained with the loop, but efficiency was less than that obtained with the 5-turn wire loop. When compared to the 75-foot (22.86-m) vertical a reading of —31 dB was obtained, 13 dB inferior to the wire loop. The noise figure of the W7ZOI broadband preamplifier was too high for use with the new antenna, so a common-gate JFET pre-preamplifier was added between the loop output and the input to the 'ZOI unit. Ideally, a 30- or 35-dB preamplifier (two JFETs) should be used with the rod loop to assure a low noise figure and ample gain.

With the coupling set for maximum loop gain, the bandwidth (measured with an HP spectrum analyzer) is 10 kHz at the 3-dB points. The loop was measured independent of the shields and preamp for unloaded

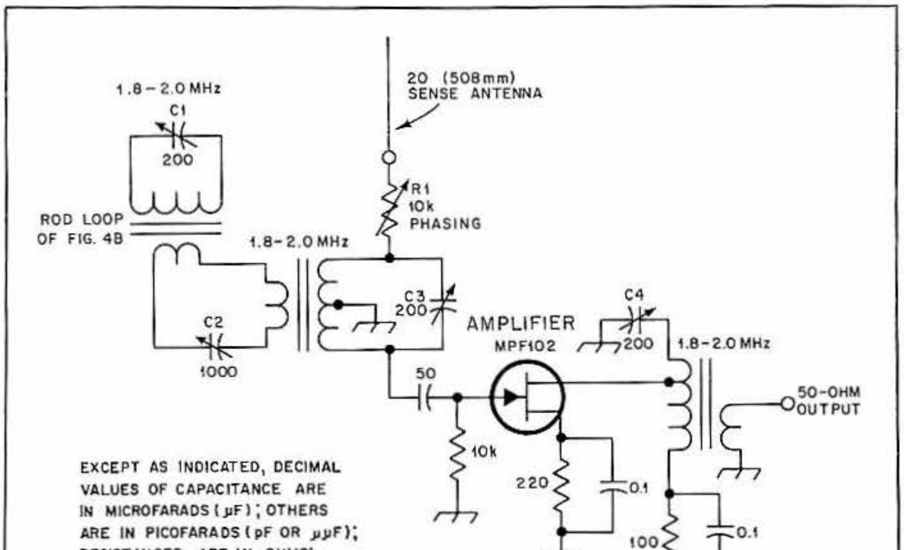

Fig. 5 — Suggested circuit for using a sense antenna with the rod loop. The sense antenna can be mounted 6 to 10 inches (152 to 254 mm) in front of the broadside of the rod loop (centered). C1, C2, C3 and R1 are adjusted alternately for the deepest null possible off the rejection side of the loop (side opposite the sense antenna). It may be necessary to experiment with the length of the sense antenna to obtain the best null in the cardioid response pattern. The tuned circuits for the JFET preamp can be wound on T-68-2 toroid cores.

Q by means of an HP Q meter. A reading of 430 was obtained.

The electrostatic shield cured the unwanted pickup of TV spurs and appliance transients during indoor use. The author would like to mention that the wire and rod loops are affected to some extent by the presence of the 75-foot (22.86-m) vertical. The latter is only 20 feet (6.10 m) away from the operating position. It tends to reradiate noise and signals. Some of that energy is picked up by the loop antennas. If an indoor loop is to be used regularly, or if it is outdoors, but within 100 feet (30.48 m) or so of the vertical, the latter should be shorted out or detuned during the receive period. An spst relay could be slaved to the regular antenna change-over relay to accomplish this.

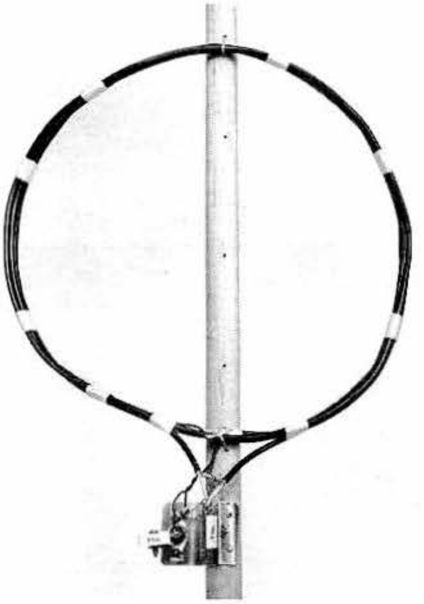

The 4-turn coax loop with electrostatic shielding. The trimmers are visible on aluminum brackets below the loop.

An improvement in performance for either type of loop could be had by adding a sense antenna and appropriate phasing circuit (Fig. 5). This would provide a cardioid pattern and result in a unidirectional response. Greater reduction of QRN and QRM would be effected through that technique. A weatherproof system of that type could be placed out-of-doors for year-round use. A TV type of rotator would be useful for changing the loop position, as desired.

The W1FB 4T-ES Loop

During a final investigation of loop types and performance characteristics, an attempt was made to devise a shielded loop small enough to be practical indoors. The primary objective was to develop a loop from ordinary materials — something that was inexpensive and easy to build. It should have efficiency comparable to that of the 5-turn wire loop described earlier in this report. A reasonable null depth and an undistorted pattern were sought also. From these criteria evolved the 4T-ES (four turn, electrostatic shield) loop shown in Fig. 6.

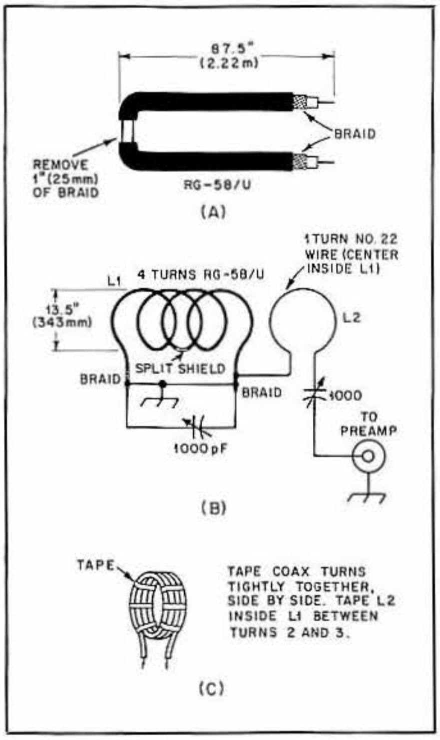
87.5"
(2.22 m)
BRAID
REMOVE 1" (25 mm) OF BRAID
RG-58/U
(A)
1 TURN NO. 22 WIRE (CENTER INSIDE L1)
L1
4 TURNS RG-58/U
13.5" (343 mm)
L2
SPLIT SHIELD
BRAID
BRAID
1000
TO PREAMP
1000 pF
(B)
TAPE
TAPE COAX TURNS TIGHTLY TOGETHER, SIDE BY SIDE. TAPE L2 INSIDE L1 BETWEEN TURNS 2 AND 3.
(C)

Fig. 6 — Structural details of the W1FB 4T-ES loop antenna. At A is an illustration of the total length of RG-58/U before the loop is formed. An expanded view of the loop turns and pickup link is given at B. Drawing C shows how the loop turns are taped together side by side as the antenna takes its final form.

It consists of four turns of RG-58/U cable which are formed into a circular configuration to provide a diameter of 13-1/2 inches (343 mm). There is nothing critical about the diameter, and RG-59/U coaxial cable could be used instead of the type specified. The turns are formed side by side, then taped at several points to maintain the loop form.

It is necessary to split the shield braid at the electrical midpoint of the loop. This prevents the shield from acting as a shorted turn. The principle is the same as that described for the ferrite-rod loop with its aluminum-channel shield. The braids are joined at the loop feed point and made common to one end of L2. The three grounded elements are then attached to the coaxial connector ground terminal on the loop support frame.

Loop performance was checked by using the test procedure discussed earlier. A null depth of —26 dB resulted. There was no discernible pattern warping, and the null occurred deeply over just a few degrees of rotation. The 4T-ES was 18 dB inferior to the 75-foot (22.86-m) vertical antenna when checking distant signals. The unloaded Q was measured by means of the HP Q meter. A value of 85 was obtained. Because C2 was set for light loading of the loop, a bandwidth (3-dB points) of 20 kHz was obtained when the preamplifier and receiver were connected to the antenna.

Numerous comparisons were made on DX signals from Europe and the Caribbean area during the ARRL DX Contest in March of 1977. In all instances the 4T-ES was as good as or better than the 75-foot (22.86-m) vertical. Ambient noise was considerably lower in amplitude while using the loop, even though all tests were performed with the loop indoors at ground level. There was no reduction in atmospheric noise (which was high that weekend) because the maxima of the loop favored the N.E. and S.W. directions: The storm front was S.W. of Connecticut. However, rotation of the loop caused a drop of 26 dB in the static crashes when the loop nulls were S.W. and N.E. Because of the electrostatic shielding there was minimum noise induction from the house wiring — a contrast to the performance characteristics of the unshielded 5-turn wire loop. The W7ZOI broadband preamplifier and an attenuator were used during the tests. There were some instances when the loop provided a stronger signal response than the vertical, even though the two systems were adjusted for unity gain.

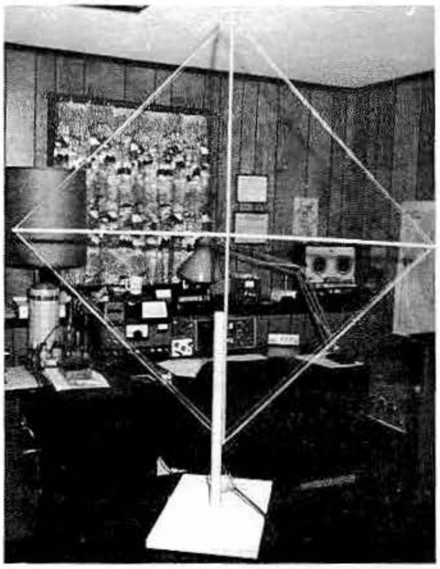

Close-up view of the 160-meter indoor receiving loop at W1FB. The antenna is resting on a piano bench, and is situated near the operating position so that it can be hand-rotated as necessary.

The 4T-ES is suitable for use indoors,

but it could be weatherproofed for outside installations. Epoxy cement could be used to seal the open ends of the coaxial cable of L1, and the trimmer capacitors would survive the natural elements if housed in a metal or plastic container.

In Summary

The loops that have been discussed here should be suitable for use on 40 or 80 meters if scaled accordingly. They may not show much directivity on sky-wave signals because of the "tumbling effect" at the higher frequencies. But QRM from nearby amateurs could be reduced greatly, and man-made noise sources could be rejected significantly by correct orientation of the loop.

The author has attempted to discuss some of the characteristics of small loops which are not treated in most of the references contained in amateur journals. There is plenty of latitude remaining for experimentation, and the reader is encouraged to apply his or her skills in developing better loops for amateur use. Meanwhile, why not build a small loop and scoop those weak signals out of the noise? It could mean some new countries for you on "top band" or 80 meters. This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] DeMaw, "A Receiving Loop for 160 Meters," *QST*, March 1974.

[2] Terman, *Electronic & Radio Engineering*, McGraw-Hill. Kraus, *Antennas*, McGraw-Hill. Jasik, *Antenna Engineering Handbook*, McGraw-Hill. Keen, *Wireless Direction Finding*, 3rd Edition, Iliffe & Sons, Ltd.

[3] DeMaw, "His Eminence — the Receiver," Parts I and 2, *QST*, June and July 1976. Receiver used in combination with an Autek Research QF-1 RC active audio filter.

[4] DeMaw, "Build This Quickie Preamp," *QST*, April 1977.

The Half Square Antenna

Do those 100 countries still elude you for 80-meter DXCC? In terms of dollars versus decibels, a half-square antenna may be your best bet for putting punch into your signal.

With 5-band DXCC as bait, considerable interest has arisen in DX and antennas for the lower frequency ham bands. The author's attempts to get a decent 80-meter DX antenna have gone through an evolutionary cycle with a lot of iterative tries and with steadily improving results. The half-square antenna that is the latest of these has some interesting properties and gives about as many "dBs per buck" as any this author has seen.

Before describing this configuration, it might be interesting to cover briefly the different antennas tried which didn't work as well. The writer's available options were set up by (1) a yard full of tall trees and (2) very little money to spend. The first antenna was an inverted V hung on the 75-foot (22.9-m) tower. This is really a quite good antenna — primarily because it has the high current part of the antenna at the highest support point. It provided a standard against which all succeeding antennas were checked. The next step was to add a parasitic element (director) which didn't seem to help at all. This is perhaps not so surprising with the ground so close. Viewed simply, normal parasitic coupling is dependent upon a fair proportion of the current in the driven element being coupled to the parasitic element. When they are both close to the ground, the reverse polarity of the image of the driven element starts to couple heavily to both the driven element and the parasitic element, making it practically impossible to get the desired parasitic current at any useful spacings.

This effect has been shown experimentally to be less with closed-loop type antennas such as the quad. Based on this supposition, the author next constructed a 2-element quad, supported at the corners by ropes tied off to trees. With only 75-foot (22.9-m) trees and the consequent sag, the bottom was only about 6 feet (1.83 m) off the ground. However, the quad did work, had an acceptable feedpoint impedance, and gave a few dB (about 1/2 an S unit) improvement over the inverted V on European contacts. Obviously, the only useful low-angle radiation in this antenna came from the upper half of it, the lower half serving to feed or excite the upper half. A major problem with the quad was keeping it up. It had to be very tightly stretched between trees to keep it from sagging to the ground and as a result the wire often broke.

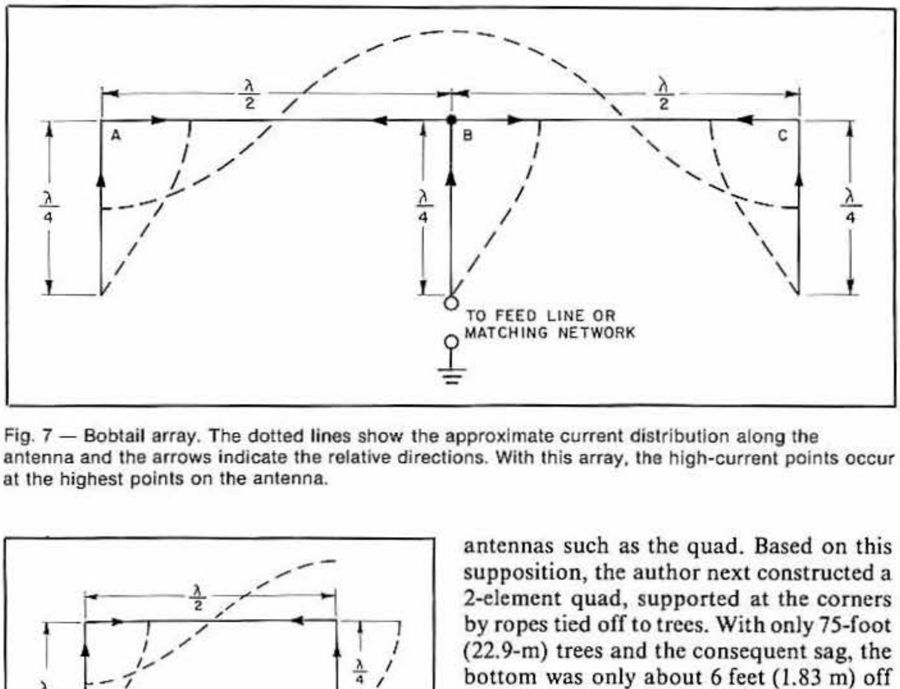

Fig. 7 — Bobtail array. The dotted lines show the approximate current distribution along the antenna and the arrows indicate the relative directions. With this array, the high-current points occur at the highest points on the antenna.

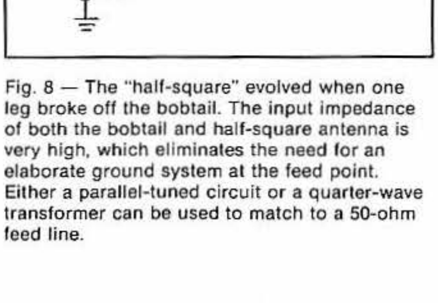

Fig. 8 — The "half-square" evolved when one leg broke off the bobtail. The input impedance of both the bobtail and half-square antenna is very high, which eliminates the need for an elaborate ground system at the feed point. Either a parallel-tuned circuit or a quarter-wave transformer can be used to match to a 50-ohm feed line.

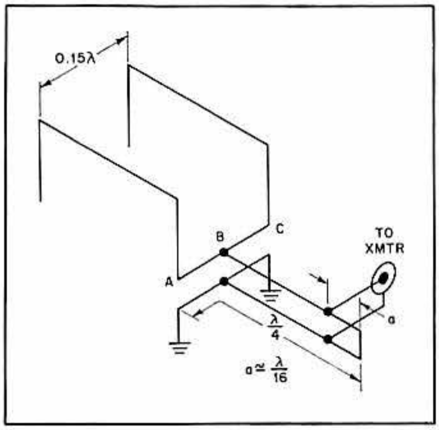

Fig. 9 — Two-element half square array gives greater gain and a very wide bandwidth.

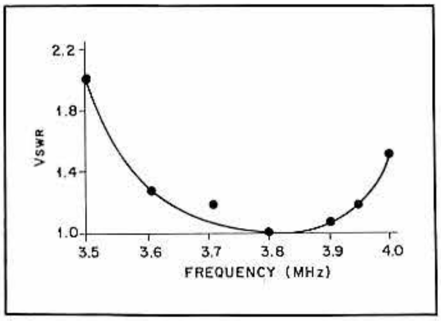

Fig. 10 — Measured SWR of the antenna shown in Fig. 9.

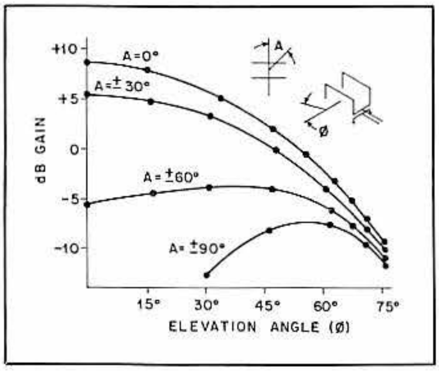

Fig. 11 — Computer-pattern plot for a 2-element half square array with 0.15-wavelength spacing between elements. The angles referred to in the drawing are shown in the inset.

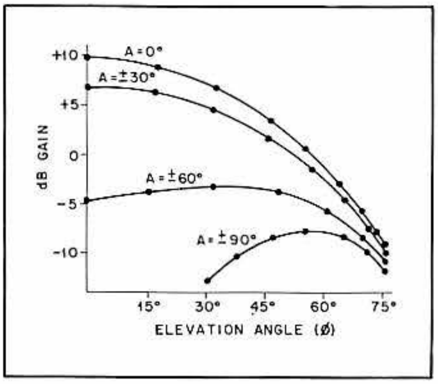

Fig. 12 — Computer-pattern plot for the antenna of Fig. 11 with 0.25-wavelength spacing.

To try another tack which would avoid the parasitic coupling problem an 8JK array with four half-wave elements was tried, all elements being fed. With the span of this being so long and with lots of high-voltage points which had to be kept from touching the trees, the author only managed to get it to a height of about 40 feet (12.19 m). The signal-report results with this were essentially identical to the quad on the path to Europe. The horizontal beamwidth was noticeably less than the quad. It probably would have done better if the writer could have gotten it to 75 or 80 feet (22.86 or 24.38 m) in height.

The next antenna tried was the so-called bobtail which is shown in Fig. 7. This is a broadside vertical array with approximately twice the current in the center leg as the two outer legs and with current distribution in the horizontal wires which tend to cancel horizontally polarized emission.[1] A major reason for trying this was the author's frustration with trying to use the maximum available height of the trees and still be able to feed the antenna and keep the high-voltage points clear of the tree branches. The author had avoided vertical antennas before because of the difficulty in getting a good low-loss ground. Also, the trees looked like great vertically polarized rf energy eaters. The bobtail is particularly nice for tree hanging since its highest support points (A, B, and C in Fig. 7) are all at low impedance and looked as if they could be directly slung over a tree limb with no insulators. It is fed from the bottom at a high-impedance point so the ground system doesn't need to be as good as it does with low-impedance feedpoints. Also, like the inverted V, the bobtail has its maximum current at the highest support point. It can be excited from a parallel-resonant tank circuit or a quarter-wavelength stub.

The bobtail has been an outstanding performer. It's at least an S unit better than the inverted V and works great on DX paths to the Far East (JA, KG6, and VS6). The author put up two of these at right angles and found their interaction to be practically nil even though their center legs are quite close. When switching, the author ground the feed point of the one not in use.

The Half-Square Antenna

With such fine results, it was thought the ultimate "tree-hung" antenna had been discovered. However, after a particularly bad storm, it was noticed that one leg of the bobtail had broken off and yet the antenna feed impedance hadn't changed noticeably. Furthermore, the results compared with the inverted V seemed about the same as the full bobtail. Fig. 8 shows the configuration. It is a 2-element (instead of 3) vertical broadside array with the same current in both legs. It has all the advantages of the bobtail except that the radiation from the horizontal part of the antenna doesn't cancel as well in all directions and some end-fire, horizontally polarized energy does spill off. This might be considered an advantage or a disadvantage, depending on whether you're seeking coverage or discrimination! In discussing this with Ed Watters, WA3LGX, he pointed out the kinship this antenna has with the bisquare array. It's sort of half of a bisquare laid sideways with the ground image antenna making up the other half. That, plus its obvious shape, led the author to dub it the half-square antenna.

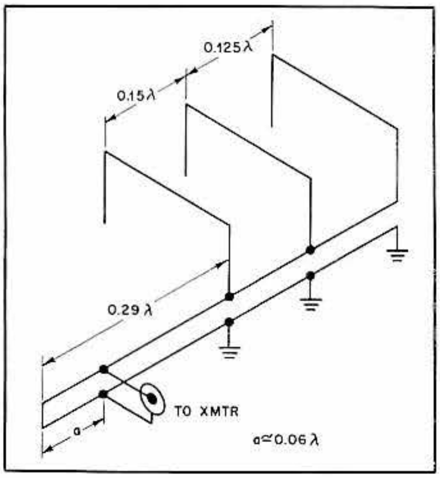

Fig. 13 — A 3-element half-square array.

After using a single half-square for awhile, it was decided to add another element with about 0.15-wavelength spacing. With the new element in place, a check was made to see what parasitic coupling there might be between elements. With several hundred watts fed into the driven element there was no evidence of rf on the parasitic element (using a neon bulb). This was a very gross measurement which tended to confirm what was expected. The very close proximity of the antenna to ground tends to reduce greatly parasitic coupling to a nearby element. Being puzzled why it should appear so small, the only hope was to devise a means of measuring this coupling more accurately.

Having had to tear down the 8JK antenna to get enough wire for this antenna, as a first try for feeding it, the author just used the 8JK open-wire line and stub match as it existed. The one wavelength of wire for each element was taken directly from the 8JK also. Fig. 9 shows the configuration. Unlike the 8JK, which was narrowband (centered around the middle of the band), this configuration gave the very flat SWR shown in Fig. 10 — on the first try! The SWR was not very different whether the feed-line stub was connected to point A, B, or C (Fig. 9). B gives a bidirectional characteristic while connection to A or C gives some unidirectivity. The author is still suspicious of the wide bandwidth and suspects the two elements may be different effective lengths, giving a stagger-tuned effect. If the spacing between these two were increased to a quarter wave and fed at point A or point C, it should theoretically add another couple of dB gain

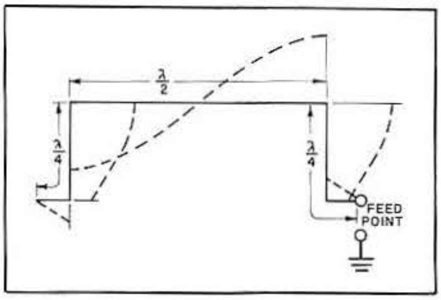

Fig. 14 — Any of the antennas mentioned can be bent as shown in order to keep the current loops at the upper corners.

and give a more significant front-to-back ratio than with the 0.15-wavelength spacing. WA3LGX plugged the dimensions and shape of this antenna into a generalized computer program he has developed for wire antennas and got the curves shown in Figs. 11 and 12. Fig. 11 defines the forward pattern for 0.15-wavelength spacing assuming a perfectly conducting earth. Fig. 12 is for quarter-wave spacing. The actual patterns will be modified at angles below 10° to 15° by the earth's resistivity[2] as is the case with any vertical antenna. The 0.15-wavelength-spaced antenna was used for some time and consistently gave better results than the bobtail on the path to Europe.

The writer took one other step beyond the 2-element array and added another element at 1/8-wavelength spacing (see Fig. 13). The extra element is fed, as are the other two, with an extension of the same open wire line. This antenna is considerably quieter (less QRN) than any of the other 80-meter antennas the writer had and has better than half an S unit advantage over the bobtail on European signals. It doesn't require any retuning to cover the whole band.

With these antennas not insulated at their support points, one might expect some effects when the tree sap rises or in wet conditions. The author hasn't been able to notice any differences from dry weather to monsoon-type rains or wet snow covering everything. The trees are slightly over 1/4-wavelength high on 80 meters. Considered as antennas or absorbers, they would look like grounded 1/4-wavelength vertical radiators. The vertical sections of the half-square antenna that run close to the trees are just the opposite. That is, at the bottom of the antenna the voltage is high and at the top, the voltage is low. This should minimize coupling to the trees, as opposed to, for example, using 1/4-wavelength grounded verticals for the antenna elements.

160-Meter Operation

Now if you look closely, you can see that the 80-meter half-square whether 1, 2 or 3 elements, should make a good antenna for 160 meters. Each element is 1/2-wavelength long on 160 meters, voltage fed on the end, and the high current part of the antenna is at the maximum height. True, a higher impedance part of the antenna touches the tree, but you can't have everything free! While not having managed to work any DX with the few watts available on 160 meters, the author can *hear* Central European ham stations quite well on this band for the first time ever at this location. Of course, a different stub or tuning arrangement is necessary for getting a low SWR on 160 meters. The writer just unshorted the stub and put a tapped coil (about 14 microhenries) in series with the coax feed line to tune it.

Construction Details

These antennas were all built with no. 14 soft-drawn copper wire. The elements are pulled over the top limbs of trees that have approximately the correct spacing and tied to nylon strings at about shoulder level. If your trees aren't a full quarter wave high, the extra wire can be stretched out parallel to the ground at each end as shown in Fig. 14. Note that you still get the maximum current point at the highest support point you have.

The best technique found for getting a line over the highest tree is to use a fisherman's casting reel (I use a Zebco 202) with an 8-pound nylon line and a 1- or 1-1/2-ounce sinker. The reel is put in the RELEASE position and a sling shot is used to shoot the sinker and line over the desired tree. This is a very accurate technique and you can easily put a line over 90-foot (27.43-m) trees. The light fishing line is used to pull over a heavier line. It's a good idea to paint the sinker a bright color since the line is almost invisible. You can waste a lot of time looking for the sinker at the end of the line.

Be sure to leave enough droop in the horizontal section of the antenna to allow for tree motion in high winds. The open-wire feed line used for the harness and stub sections is made with the same no. 14 wire spaced 3-1/2 inches (89 mm).

The avid Field Day or camping fan will notice the obvious advantages of the half-square antenna for that type of operation. It uses minimum materials, goes up very quickly, will really put out a booming signal on 80 meters, and can be taken down in just a few minutes. You can also make good use of the fishing rod in between your turns at the key. This material was originally presented in *QST* by Ben Vester, K3BC.

References

[1] [Editor's Note: While complete cancellation of the horizontally polarized component occurs at a point broadside to the array, some radiation from the flat-top portion will exist at other angles. This is because of incomplete cancellation of the components from the opposite ends of the flat top.]

[2] *The ARRL Antenna Book*, 12th Edition, p. 46.

The HRH Delta-Loop Beam

Here's an antenna that should excite the interest of the amateur fraternity. The Delta-Loop beam has all the features of a quad without many of the disadvantages.

The evolution and design of the antenna shown in the photographs came about primarily because of a lack of confidence in the mechanical durability of quads. Anyone who has ever built and used quads knows how difficult it is to keep them up in icing and high wind conditions.

The author being an old-time mechanic, felt there had to be a better design for an antenna that used full-wave elements in a parasitic array, so some years ago a start was made to find that better design. One thing that nearly all readers will agree is that Mother Nature is about the best designer there is. Along these lines, just observe the growth of branches on tree limbs. You'll find that many of the smaller branches will grow up in a semivertical pattern from the limb; see Fig. 15. It takes a considerable amount of icing and wind to break such branches, indicating the soundness of the design. This same idea should apply to elements mounted on a boom.

To test the idea, an element of aluminum tubing was mounted on a boom. The element was mounted as shown in Fig. 17. The boom was then secured to the family car with the element in a vertical position. The assembly was then tested at 65- to 75-mph. Even though the element was only secured to the boom as shown in Fig. 17, there was no "set" or permanent bending. This test opened the door to several possibilities in beam design.

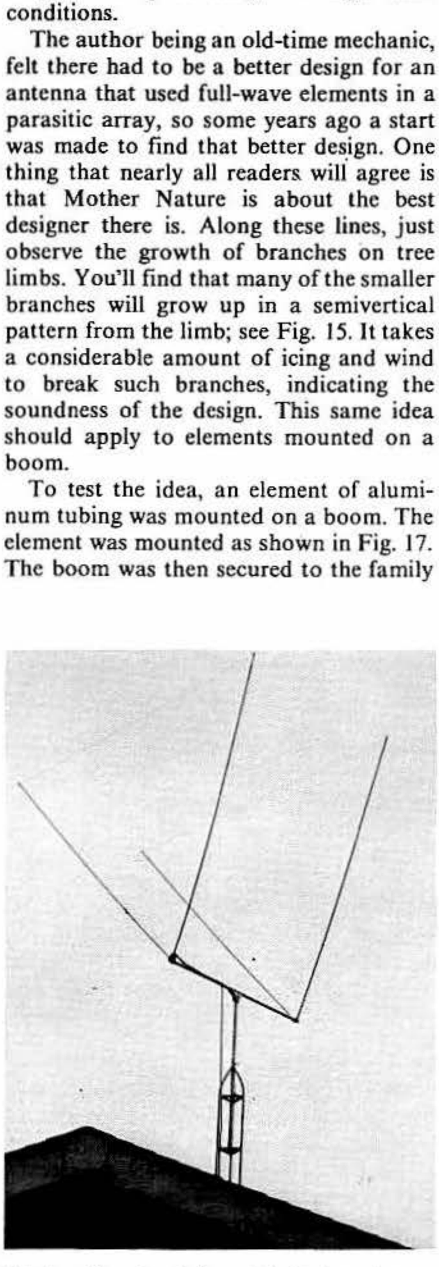

This is a 10-meter, 2-element Delta Loop that is very similar in construction to the antenna described in the article.

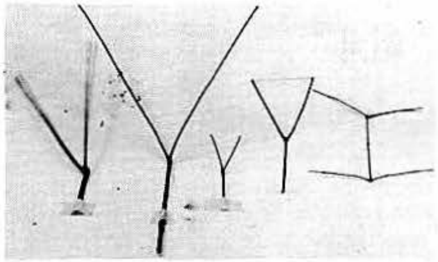

Fig. 15 — The basic idea for the Delta-Loop beam was conceived from the growth of tree limbs and branches. Here are just a few of nature's configurations that lend themselves to antenna design!

A loop aerial can take many shapes. Anyone who studies the history of the quad will find that many different configurations have been used. Why not use a triangle?[1] It should be possible to achieve a very good design using a vertical triangle. See Fig. 16. This shape appears ideal for a beam with full-wave elements, and has most of the advantages of the quad without many of the disadvantages. This led to the construction of several antennas of the type shown in the photographs, and some of these advantages become quickly apparent.

First, the entire antenna is *above* the boom. Second, the antenna is constructed primarily of aluminum tubing, which provides extra strength as compared with wire elements. Third, the antenna has Plumber's Delight type construction, meaning that the antenna is at ground potential for lightning protection, plus the fact that this type construction lends itself to gamma matching of the feed point, eliminating the need for a balun. Last, but not least, we find the antenna is very attractive in its symmetry. Of course, the important point is how well does the antenna perform?

Several 10-meter models have been tested and the unit shown in the roof-mounted position is 20-meter Delta Loop. Using a gamma match with 50-ohm coaxial feed, a 10-meter model was matched at 28.8 MHz. The antenna consisted of a driven element plus reflector with the elements spaced 6-1/2 feet (2 m) apart, or approximately 0.2 wavelength. Surprisingly, the antenna was extremely flat across the entire 10-meter band, the worst mismatch being *less* than 1.2 to 1.[2]

The author has no means of measuring gain or front-to-back ratios. However, with the spacing and element sizes the same as a quad, the gain should be the same, or at least so close the difference would be insignificant. On-the-air tests have shown the antenna to be as good as or better than quads the author has had up. Front-to-back and front-to-side reports have been outstanding, both on ground wave and skip.

Element Lengths

Many tests have been made on driven elements to determine the effect of element lengths on SWR. It was found that the flattest curve was obtained with the formula

$$\frac{1005}{f_{MHz}} \text{ (feet)}$$

for the driven element. The reflector should be about three percent longer, or

$$\frac{1030}{f_{MHz}}$$

If directors are desired it would appear that a length three percent shorter than the driven element would be adequate.

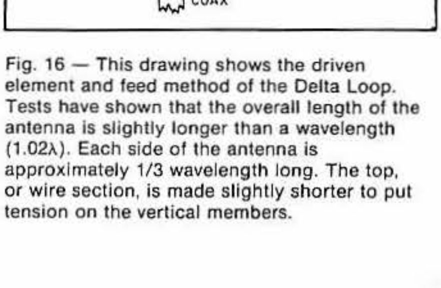

Fig. 16 — This drawing shows the driven element and feed method of the Delta Loop. Tests have shown that the overall length of the antenna is slightly longer than a wavelength (1.02λ). Each side of the antenna is approximately 1/3 wavelength long. The top, or wire section, is made slightly shorter to put tension on the vertical members.

Table 1

Bill of Materials

Tubing used is 6061 T6.
4 — 12-foot lengths of 3/4-inch OD, 0.035 wall (elements).
1 — 12-foot length 1-3/4-inch OD, 0.065 wall (boom).
8 — 3/4-inch diameter stainless steel hose clamps.
1 — 9-inch length of 3/8-inch OD by 5/16-inch ID.
1 — 9-inch length of 1/2-inch OD to slide over 3/8-inch section.
25 feet of copper wire no. 12 or 14, or equivalent in stranded wire.

Feet X 0.3048 = m.
Inches X 25.4 = mm.

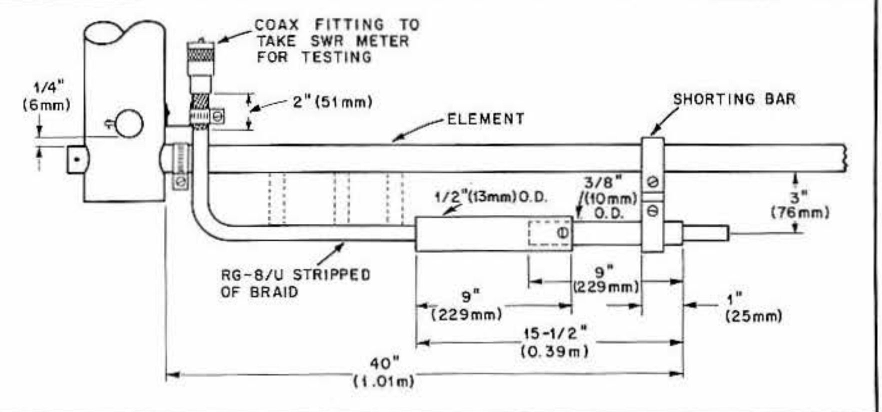

Fig. 18 — Here are the details on the gamma-matching section. As mentioned in the text, the coax fitting can be a female type SO-239, mounted on a small bracket which is mounted on the boom. Spacing insulators for spacing the gamma line from the element can be made up from pieces of Lucite or Plexiglas.

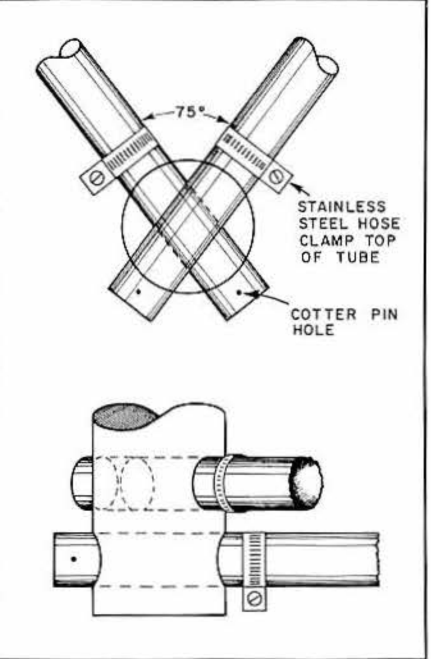

Fig. 17 — This drawing shows the method of mounting the element to the boom. A cotter pin is used on the under side of the boom and a hose clamp on the top to hold the element securely to the boom.

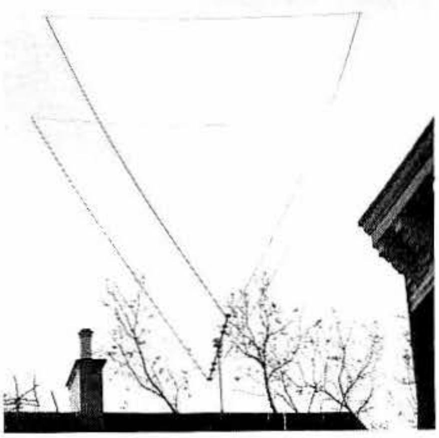

Here is our roof-mounted 20-meter Delta-Loop beam that is now undergoing a series of tests.

As to element spacing, the author has long been an advocate of wide spacing (0.17 to 0.2 wavelength) whenever possible. In the 10-meter tests 0.2 wavelength spacing was used.

For readers who are interested, complete construction information on a 10-meter model, including a list of the required tubing, follows.

Construction Notes

Fig. 17 shows the method of mounting the element tubing to the boom. The angle between the semi-vertical elements is shown as 75 degrees but this can vary a few degrees either way without any appreciable effect on the performance of the array. It is difficult to drill and line up the element support holes in the boom and come out to *exactly* 75 degrees.

There are a couple of methods of making the holes in the boom to hold the elements. One of the simplest is to use a Greenlee-type chassis punch. This makes a clean hole in the boom. Another method is to drill holes large enough to take the end of a 3/4-inch (19-mm) reamer and then ream out the holes to the necessary diameter.

One not accustomed to working with angles might find it difficult to drill holes at a 75-degree angle. A simple method is to first drill holes to take a single element. Next, make a jig from a piece of stiff cardboard or similiar material, using a protractor to get the 75-degree angle. Then insert the single element into the boom and lay the jig along the element; this will give you the correct alingment for the other element.

The elements are inserted through the boom just far enough to take a cotter pin on the underside of the boom. On the top side, a hose clamp is used to hold the element to the boom and prevent slippage. A length of copper wire can be wrapped around both elements just below the clamp to insure a good contact between the elements and the boom. Hose clamps are also used at the tops of the elements to hold the horizontal wire.

Fig. 18 shows the details of the gamma matching section. The gamma section is made up from the inner conductor of the coax, including its insulation, and two sections of telescoping aluminum tubing. The outer covering and braid is removed from a 52-inch (1.32-m) length of RG-8/U coax. In Fig. 18, a male coax fitting is shown with a short length of braid, 2 inches (51 mm) long, which is grounded to the boom via a clamp and bracket. If desired, a chassis-type coax fitting, type SO-239, could be mounted on the boom with a small metal bracket. The capacitor section is made of two lengths of tubing 3/8-inch (10 mm), and 1/2-inch (13 mm) diameter, respectively, and both 9 inches (229 mm) long. The 3/8-inch (10 mm) tubing fits over the insulation around the coax inner conductor while the 1/2-inch (13 mm) tubing slides over the 3/8-inch (10 mm) material. The larger tube is drilled and tapped to take a locking screw. Spacing of the section from the element is 3 inches (76 mm).

In adjusting the gamma, figures given in

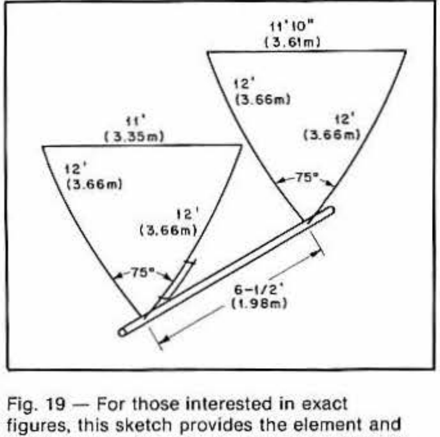

Fig. 19 — For those interested in exact figures, this sketch provides the element and boom lengths for a Delta-Loop beam for 28.8 MHz.

Fig. 19 can be used as a guide. In order to avoid matching errors, the gamma should be adjusted with the SWR bridge right at the beam. Set the shorting bar to the dimension given in the diagram and slide the 1/2-inch (13 mm) tubing over the smaller tubing, looking for a setting that gives a match. If one cannot be obtained, move the shorting bar and sliding tubing assembly a short distance, say an inch (25 mm), and then try different settings of the 1/2-inch (13 mm) tubing until you find a match. We found that the settings were not critical in getting a match. Once the beam is matched, the regular feed line to the station can be attached to the feed point. The connection should be taped to prevent moisture from getting into the fitting and coax.

As stated earlier, the antenna is extremely flat across the entire 10-meter band. The present model was designed for 28.8 MHz and matched at this frequency. The SWR at the highest point, at 29.7 MHz, was about 1.2 to 1. This dropped to 1 to 1 at about 29 MHz and stayed at 1 to 1 all the way down to 28 MHz. This feature of the antenna certainly is a help when using a transceiver designed for 50-ohm output. The beam has proved its ruggedness through winds and icing conditions. All in all, we think the Delta-Loop beam is one of the best performers we have seen. This material was originally presented in *QST* by Harry R. Habig, K8ANV.

References

[1] [Editor's Note: A triangular loop configuration also was described in *QST* in April 1968, "Technical Correspondence" by Norman Watson, W6DL, independently of K8ANV's design. Mr. Habig has been working for some years on this design and has a patent pending.]

[2] [Editor's Note: A 10-meter model of the Delta Loop beam tested at ARRL headquarters was matched at 28.8 MHz and an SWR curve was made using a Bird Wattmeter. The largest mismatch also was less than 1.2 to 1 MHz (at 29.7 MHz).]

The Delta-Loop Beam on 15

DX capability, low SWR and readily available construction materials are highlights of this 15-meter beam antenna. Similar to a quad, but rugged as a Yagi — those are the virtues of the Delta Loop.

Some of the features of the 15-meter beam described in this article include a very low SWR across the entire band, use of readily available materials, "Plumber's Delight" type construction — which is always an appealing feature of any beam antenna — and last but not least, excellent performance. On the last point, the antenna shown in the photographs was tested during a DX contest with a transmitter input power of 100 watts. The boom of the antenna was mounted only eight feet (2.44 m) above the ground, but in about five hours of contest operation 58 different countries were worked — and, believe it or not, the majority of the reports received were 59. Possibly conditions were above par, but even so the antenna shows it can do a real job under crowded band conditions.

As stated above, the beam uses Plumber's Delight type construction. For the newcomer's information, this type of construction has the antenna elements connected directly to the boom which in turn can be connected directly to earth ground without having any effect on the performance of the antenna. This is desirable because of the lightning protection offered by the system.

Fig. 20 is a drawing of the Delta-Loop driven element. The overall length around the driven element is slightly over one wavelength. Each side is approximately 1/3 wavelength long. The reason we say "approximately" is that the top section of the antenna is made of wire and is slightly shorter than 1/3 wavelength in order to put tension on the wire. A gamma-match feed system is used, eliminating the need for a balun.

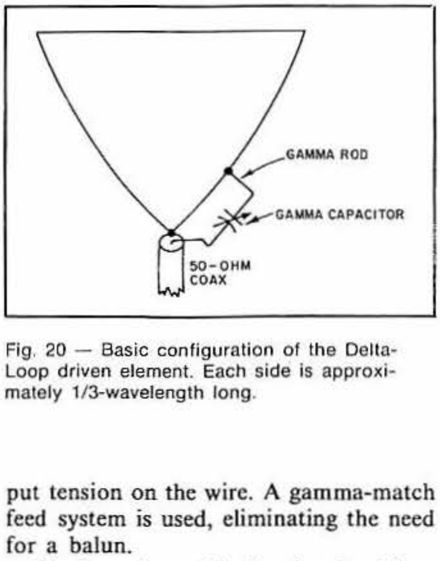

Fig. 20 — Basic configuration of the Delta-Loop driven element. Each side is approximately 1/3-wavelength long.

The formula used in figuring the driven element length is

$$\frac{1005}{f_{MHz}},$$

where f is the desired resonant frequency of the driven element. The answer will be in feet. The reflector length formula is

$$\frac{1030}{f_{MHz}}.$$

For a frequency of 21.2 MHz this figures out to a driven element 47 feet 5 inches (14.45 m) long and a reflector that is 48 feet 7 inches (14.81 m) long. Element spacing is approximately 0.2 wavelength or about 9 feet (2.74 m). Fig. 21 shows the two elements with the lengths of each segment.

Material Requirements

The beam elements were built primarily from Reynolds Do-It-Yourself aluminum tubing, which *should* be available from most hardware dealers. The reason we say "should" is because in some sections of the country hardware stores don't stock the tubing. However, most dealers will order the tubing for you. If you live in an area near or in a city of reasonable size, a look through the Yellow Pages will show any aluminum tubing dealers. The Reynolds

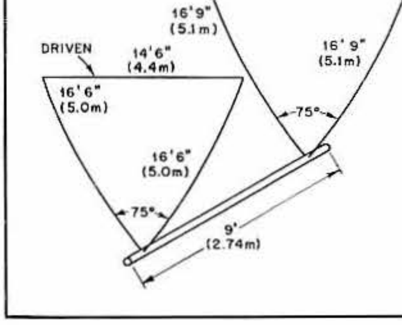

Fig. 21 — The element dimensions and the boom length. All measurements on the elements are made from where the element enters the boom; the portion of the element extending through the boom is not counted.

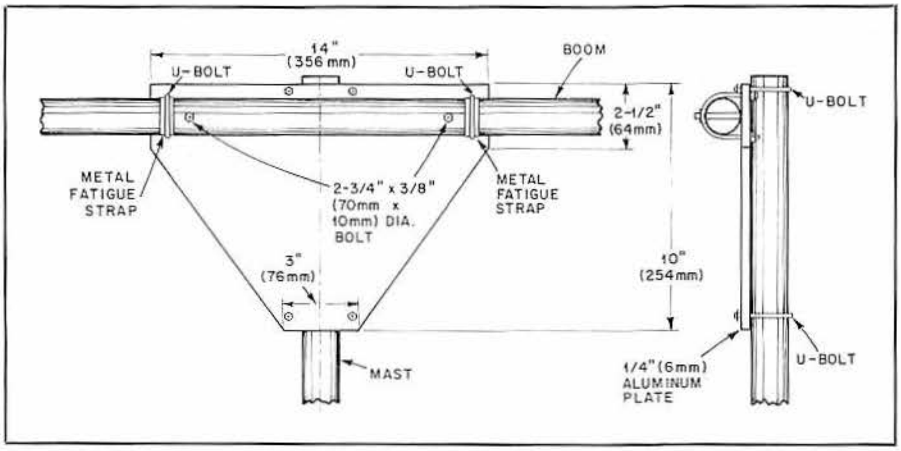

Fig. 22 — Construction details for the boom-to-mast mounting bracket. The U bolts holding the boom are 2-inch (51-mm) type. Those holding the mast will depend on the mast diameter — we used 1-1/2-inch (38-mm) U bolts for a 1-1/2-inch diameter mast.

Table 2

Shopping List

Quantity	*Length (ft)*	*Diameter (in.)*	*Reynolds no.*
4	8	1	4242
4	8	3/4	4222
2	6	7/8	4231
1	9	2	See Text.

3 feet of 3/8- or 1/2-inch diameter aluminum tubing.

12, 1-inch diameter hose clamps, stainless steel.

35 feet of no. 12 or 14 copper wire, solid or stranded.

1 variable capacitor, 100 pF maximum, 0.025 spacing or greater.

1 one-quart freezer container.

1 SO-239 Coax chassis connector.

5 1/4-inch diameter aluminum nuts and bolts, 1-1/4-inches long.

Feet X 0.3048 = m.
Inches X 25.4 = mm.

Do-It-Yourself tubing comes in 8-foot (2.44-m) lengths but most of the commercial tubing available from metal distributors comes in 12-foot (3.66-m) lengths. If you obtain the 12-foot types, be sure to specify type 6061 (61S) alloy as this material affords excellent strength for antenna elements. The boom used in this antenna was made from this alloy and is 2 inches (51 mm) in diameter with a 0.065-inch (1.65-mm) wall. Another source of aluminum or steel tubing is electrical supply houses; although this type of aluminum is a little too soft for the elements it would be suitable for a boom. Also, electrician's thin-wall steel tubing is available in various sizes, including the 2-inch diameter, for a boom. In a pinch, the boom could be made of wood, such as a length of 2 × 4. In such a case, the bottoms of the elements could be connected together with a length of wire to maintain the Plumber's Delight feature. The shopping list included here should prove of help in purchasing the parts.

Construction

The first step in building the antenna is to make up the vertical elements. The description given here is based on using the Reynolds tubing. Using a hack saw, cut the 7/8-inch (22-mm) tubing into 2-foot (0.61-m) sections. This particular type is made by Reynolds for the purpose of telescoping the 1- and 3/4-inch (25-mm and 19-mm) diameter sections together and comes in 6-foot (1.83-m) lengths. After cutting off four 2-foot (0.61 m) lengths you'll have one 4-foot (1.22-m) piece remaining. Cut this into four 1-foot (0.31-m) lengths and set these pieces aside for the time being.

Next, cut two lots, the thickness of the hacksaw blade, about 2 inches (51 mm) deep in one end of each of the 1-inch (25-mm) and 7/8-inch (22-mm) diameter lengths. Slide the 7/8-inch tubing inside the 1-inch diameter to a length of about 6 inches (152 mm) and then slip the hose clamps over the slots and tighten them. The 3/4-inch (19-mm) tubing can then be inserted into the slotted ends of the 7/8-inch (22-mm) tubing and these sections clamped with hose clamps. Two of the elements should be adjusted to the proper length for the driven element and two for the reflector, as per Fig. 21 (or whatever length you wish from the formulas). Some amateurs may prefer to cut their antennas for the cw portion of the band and others for the phone section.

The next step is to cut the 1-inch (25-mm) holes in the boom to take the ends of the 1-inch diameter elements. We used a 1-inch chassis punch to make the holes, but a 1-inch metal reamer could also be used. The aluminum is easy to cut or punch so making the holes should be no problem. What can be a problem is lining up the holes so that the 75-degree angle between the elements is obtained. We made a jig from a piece of cardboard, using a protractor. Once one set of holes for the director and reflector were made the elements were temporarily mounted in the boom. The cardboard jig was laid on the elements and the 75-degree holes were marked off and drilled.

When the holes are all drilled, the boom can be laid on the ground and the element ends inserted into the boom holes. We allowed about 1/4-inch (6-mm) extension of the base of the elements through the boom so as to allow space for cotter-pin holes. However, before drilling the cotter-pin holes, insert the four 1-foot (0.31-m) lengths of the 7/8-inch (22-mm) tubing into the elements at the boom. This will serve to give added strength to the elements at the support point. Drill each of the four ends to take the cotter pins and install the pins. Tighten down the hose clamps on the top side of the boom to secure the elements in place.

Drill the tops of the elements to take the 1/4- × 1-inch (6-mm × 25-mm) aluminum bolts. The top wires can then be installed.

The close-up view shows the gamma installation. The gamma capacitor is mounted inside a plastic freezer container which is held in place by a small metal L-shaped plate mounted on the box and the boom, using four self-tapping screws. This plate also holds an SO-239 coax chassis fitting to take the feed line. We like to install an SWR bridge directly at the beam when making SWR or matching tests to reduce matching errors, and having the fitting there simplifies the procedure. The gamma rod, made from a length of 3/8-inch (10-mm) aluminum tubing 36 inches (0.91 m) long, is flattened at one end for a length of 1-1/2 inches (38 mm).The flatted portion is bent over at right angles and drilled to take one of the 1/4- × 1-1/4-inch (6-mm × 32-mm) aluminum bolts, which is mounted through the top of the freezer box. The gamma rod is held in place by this bolt and by the shorting bar between the rod and the element. The shorting bar is made from a piece of aluminum, 1 inch (25 mm) wide and long enough to fit around the two pieces of tubing and provide a separation of 3 inches (76 mm) between the rod and the elements.

How you mount the antenna boom to the supporting mast will depend a great deal on what type of tower or supporting structure you have. The method shown in the photo is a fairly simple one, and easy to make up. Fig. 22 shows details of the construction of the mounting unit. The one shown in the photo supporting the beam was made from two pieces of 1/16-inch (2 mm) thick aluminum plate. However, we found that even with the double plate there was still too much "give" in the mounting setup, so a replacement was made from 1/4-inch (6 mm) thick aluminum. This made the mount competely steady. Either a plate of 1/4-inch aluminum or 1/8-inch (3 mm) steel should be adequate for the jobs. The 2-inch (51 mm) U bolts that hold the boom to the mounting plate were adequate, but just for added insurance, 3/8-inch (10 mm), 2-3/4-inch (70 mm) long machine bolts (two required) were installed through the boom and mounting plate for greater holding strength.

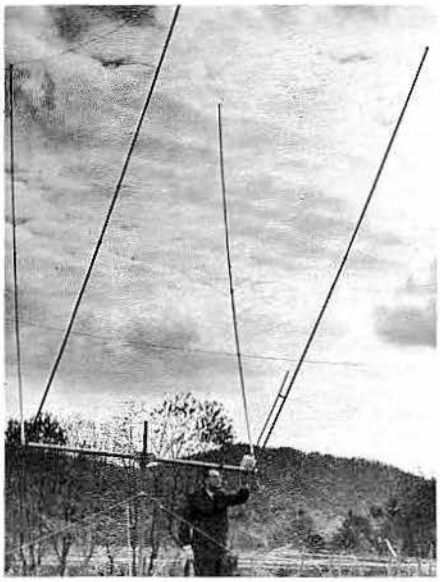

This view shows the completed antenna being matched.

During our tests, a weather front moved through the area and we had wind gusts up to 40 mph plus icing and snow loading on the antenna. The antenna and elements showed very little movement in these winds, and the entire antenna appeared very strong.

Adjustments

Adjustment of the gamma section is quite simple. Install an SWR bridge at the gamma feed point and tune up the rig on the desired frequency, Set the gamma shorting bar 24 inches (0.61 m) from the base of the gamma rod and tighten the gamma-bar nuts and bolts just enough to make electrical contact. Set the SWR bridge to read reflected power and adjust the gamma capacitor for a dip. The object is to find a setting of the capacitor and the gamma bar that gives a reading of zero in the reflected position. This may take a few tries, but you'll find the settings are not critical. If you adjust the antenna near the ground you may find it will require a new adjustment when it is up in its permanent location.

How you mount the antenna in its permanent location will depend on what you have to hold the antenna. If you use a boom mounting-plate-to-mast assembly such as already described, the U bolts holding the boom could be left slightly loose and the 3/8-inch (10-mm) bolts left out. The antenna could then be installed in an upside-down position and then rotated into place. Two men are better than one for this job, but one man *can* do the job. If a hole, say 1/2-inch (13-mm) diameter, is drilled through the boom a 1/2-inch rod could be slid through to serve as a lever to rotate the boom and antenna into an upright position and the U bolts and hardware could be then tightened down.

We think the Delta-Loop beam is an excellent antenna and offers many possibilities. Certainly the results obtained to this point prove the antenna is a real performer. Our tower is a crank-up job, making antenna testing a fairly simple project. The antenna was first matched at the 8-foot (2.44 m) level (boom height) and at the cranked-up height, about 55 feet (16.76 m) above ground. A very slight readjustment of the gamma capacitor was required at the greater height in order to get down to a 1-to-1 SWR. However, the change in SWR from the 8-foot to 55-foot (2.44- to 16.76-m) heights was very small. Also, the highest SWR was at 21,450 kHz, the top end of the band, with a mismatch no worse than 1.2 to 1. At 21,300 kHz the SWR dropped to 1 to 1 and remained at that figure all the way to the lower band edge.

The entire antenna weighs about 15 pounds (using the aluminum boom) so the antenna could be rotated with a heavy-duty TV rotator. One last note: We sprayed the cotter pins and hose clamps with a clear acrylic spray to reduce corrosion. This simple precaution should be taken in all antenna construction to prevent or reduce rusting. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

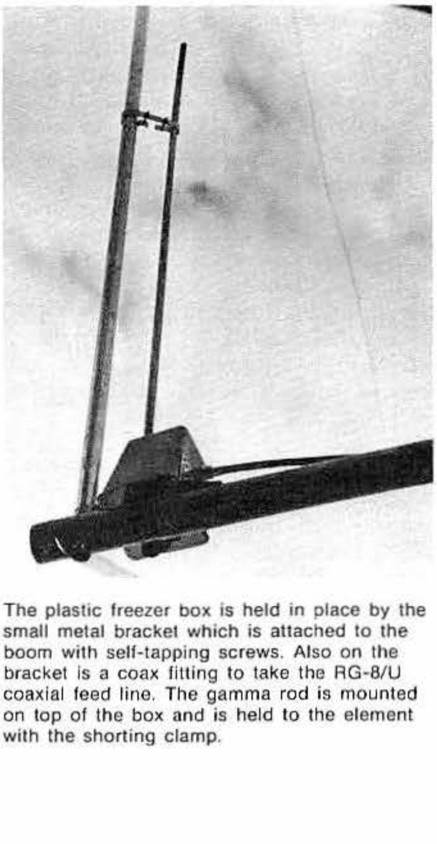

The plastic freezer box is held in place by the small metal bracket which is attached to the boom with self-tapping screws. Also on the bracket is a coax fitting to take the RG-8/U coaxial feed line. The gamma rod is mounted on top of the box and is held to the element with the shorting clamp.

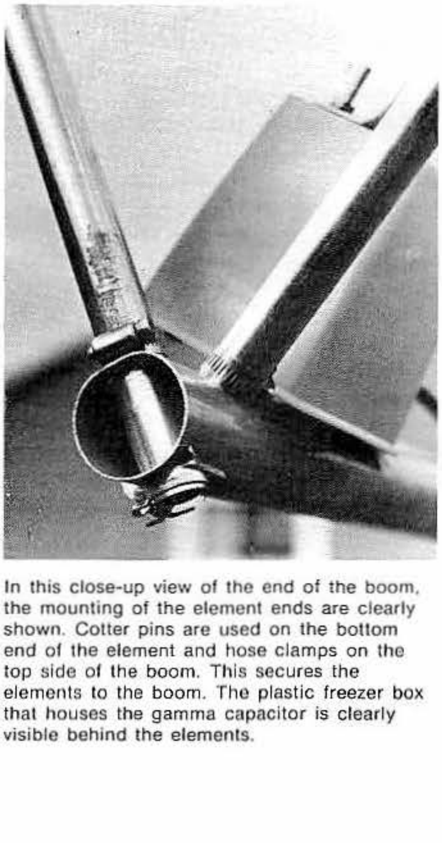

In this close-up view of the end of the boom, the mounting of the element ends are clearly shown. Cotter pins are used on the bottom end of the element and hose clamps on the top side of the boom. This secures the elements to the boom. The plastic freezer box that houses the gamma capacitor is clearly visible behind the elements.

Antennas for Travel Trailers and Campers

Pounding brass and rendering "lip service" from an RV or mobile home can be disappointing as all get-out if the antenna doesn't "play" well. Here are some tips for rig-mounted radiators that get the job done.

Who knows why so many hams like to go camping? Perhaps it's their Field Day training; perhaps it's the next logical step after mobile operation, or perhaps they just cannot leave home without their rigs. At any rate, it is nice to have your own means of communication with the outside world when you are away from home in the Maine woods, at the Grand Canyon, or even down in Mexico.

One thing all hams know is that no matter how good a transmitter you may have, you cannot get out without an antenna; and the better the antenna is, the better you get out. The antenna is usually no problem for the home station. The sky is the limit, so to speak. A 65-foot (19.81 m) crank-up tilt-over tower with a triband beam on top, and a couple of inverted Vs for 40 and 75 are not at all uncommon. But did you ever contemplate transporting such an antenna system to a state park, or some other campground, and setting it up while your XYL cooked suppper? Obviously, the mobile operator is faced with vastly different problems than those confronting the ham who stays at home, when it comes to antennas.

In 1929, the author built a 14-foot travel trailer, which was christened "Black Maria." Shortly thereafter, the writer got a ham license, and one of the first thoughts was toward a rig for the trailer. This was in the days when mobile operation was still a novelty, and the story of Black Maria was the author's first contribution to *QST*, in 1933. The rig used a pair of 33s in parallel operating from a B-battery pack.

Over the ensuing 35 years, many portable antenna designs have been tried, and much has been learned about their relative performance. The degree to which a trailer traveler is restricted in a choice of an antenna varies widely, of course, depending upon the facilities available at each stopping point. These are seldom known in advance, but to get the most out of the equipment, the trailer operator should be prepared to take maximum advantage of whatever facilities he may find on each occasion.

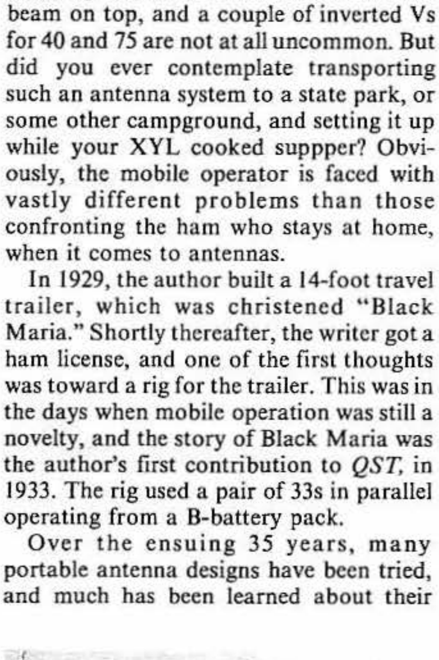
The author's 30-foot (9.14 m) Airstream trailer with the 75-meter Airstream Loop in place. The "tuning" mast is the one carrying the flag.

Dipole Antennas

Long ago, it was found that a simple dipole for 40 or 75 would far outperform the author's mobile whip, even with the dipole only a few feet above ground. While writing this, sitting in the latest Black Maria — a 30-foot Airstream job — up in Port Colburne, ON, the author was listening to the 75-meter phone band. The dipole is only 8 feet (2.44 m) above ground. When checking into the Airstream Travel Trailer Net, which meets every Sunday on 3963 kHz at 8 A.M. local time, the writer worked other trailer stations in Connecticut, New Jersey, Maryland, Pennsylvania, New York and Ohio. Several times the author switched to the mobile whip and could not be heard through the QRM. Such a dipole can be prefabricated and, when rolled up, occupies very little storage space.

When the length of the stay warrants it, and trees or other supports are available, the antenna will be even more effective when raised to a respectable height. A bow and arrow is kept on hand for shooting lines over taller trees. Incidentally, Hy-Gain makes a deluxe portable dipole consisting of two stainless-steel tapes in a double plastic housing. The tapes are calibrated in meters, and can be quickly reeled out to the proper length for use on any desired band. The unit is fitted with a coax connector, making installation very easy.

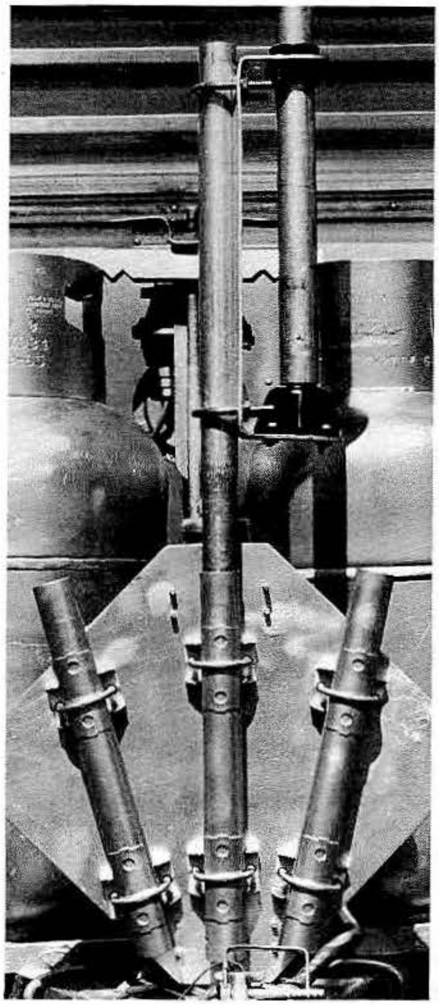
Insulated mounting at one end of the "Airstream Loop" 75-meter antenna. The "tuning" mast can be mounted directly in one of the other sockets. The tuning will vary with the spacing between masts as well as the length of the "tuning" mast. The insulated mounting can be transferred to one of the side sockets to obtain maximum spacing.

If only a single support is available, the dipole can be used in "inverted-V" fashion. A convenient way of putting up such an antenna is to use three or four 5-foot (1.52 m) lengths of 1-1/4-inch (32 mm) aluminum TV masting as a center pole, mounting it on the front of the trailer. The author has a mounting for such an arrangement clamped to the bumper, as shown in one of the photographs. The mast can be equipped with light nylon halyards for hoisting the center of the dipole. The ends of the dipole can be attached to whatever may be handy, but the higher, the better.

Loaded Antennas

Of course, there are many times when the surroundings, or the shortness of the stopover, make erection of a conventional dipole impractical. On these occasions, the author used one or another of several arrangements shown in the accompanying photographs and sketches. Where the

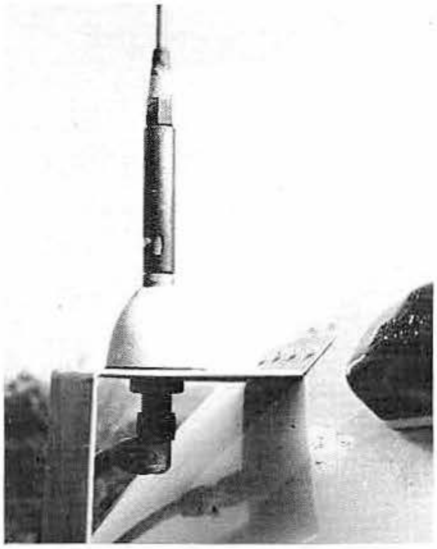

An elevated mounting for a mobile whip at a top rear corner of the trailer body. The bracket is fastened by pop rivets. This point may also be used, with the whip removed, for feeding wire dipoles.

A check with a 20-watt fluorescent lamp by W1WKZ indicates the "hot" points on the Airstream Loop 75-meter antenna system.

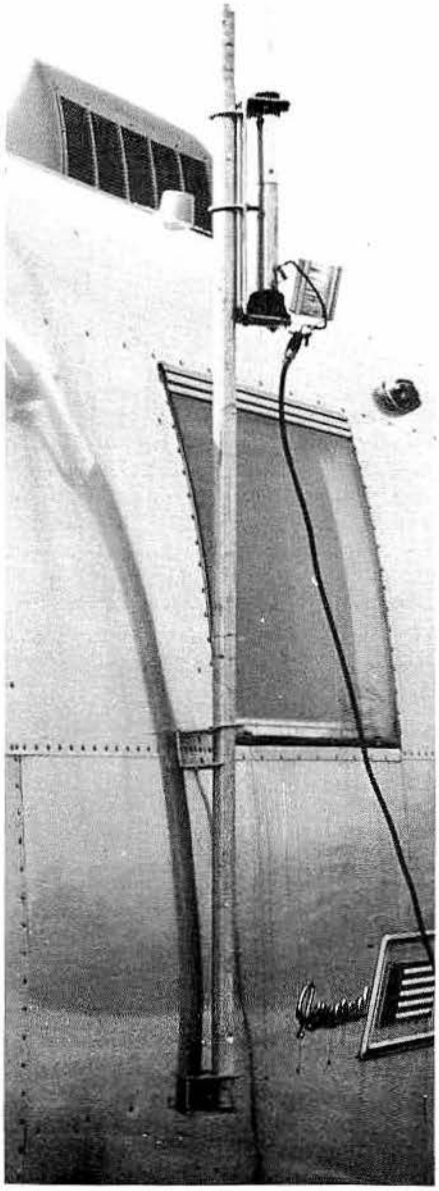

A 25-foot (7.62 m) base-loaded antenna mounted on brackets attached to the side of the trailer.

antenna system is one that requires a ground for operation, it was found that a quarter-wave wire counterpoise connected to the trailer, and strung out a few feet above ground, is usually a much better "ground" than the trailer alone, even if the trailer is actually grounded to a driven rod. The author carries two each of wires cut to lengths of 8, 16, 32 and 60 feet (2.44, 4.88, 9.75 and 18.29 m) for use as counterpoise on the 10-, 20-, 40- and 75-meter bands, respectively. One end of each wire is fitted with a large battery clip, while the other is terminated in an insulator. Trailer awning poles come in handy as counterpoise supports when nothing else is available.

"Airstream Loop" Antenna

Of particular interest is the 75-meter folded configuration shown in the sketch of Fig. 23, which is called the "Airstream Loop," although, strictly speaking, it does not function as a true loop. With the fittings that are provided, this antenna can be put up in a few minutes, and its performance seems to be about equal to that of a low dipole. It has the advantage that it is completely supported by the trailer, and requires no external supports. It is contained entirely within your own campsite. There are no wires strung out for you or your neighboring campers to trip over or be garroted by in the dark. A 30-foot (9.14-m) length of wire is strung between, and connected to, two poles made of aluminum TV mast sections. A third mast (A), of approximately the same length as the others, and connected to the trailer, is an essential part of the system. This arrangement appears to work as a folded half-wave antenna, the 30-foot (9.14-m) wire and the two poles connected to it forming one half of the antenna, while the other half is formed by the trailer body and mast (A). Probing with a fluorescent lamp shows that the end of the trailer opposite the feed point is "hot" while the feed-point end is "cold." The system can be tuned by adjustment of the length of pole A, and/or changing the spacing between A and the adjacent pole. The spacing can be changed by a selection of the pole mounting sockets shown in one of the photographs. In this manner, the resonant frequency can be changed by almost 150 kHz. For any adjustment, the transmitter can be tuned about 30 kHz either side of the resonant point before the SWR exceeds 3 to 1. Since the "hot" end of the trailer is the front end in my case, the tow car must be unhitched, and an insulating block used under the customary supporting jack.

In a similar arragement for 40 meters, mast A is omitted, and a wire approximately 21 feet (6.40 m) long is connected to the pole at the feed-point end. The open end of the wire is supported from the pole at the opposite end of the trailer by a length of nylon rope.

40-Meter Monopole

Another 40-meter antenna that has been used with considerable success is a 34-foot (10.36 m) length of 300-ohm ribbon line connected as a folded monopole, and strung up in the same manner as the antenna just described. One conductor of the ribbon is grounded to the trailer at the feed point, while the other conductor is connected to the center conductor of the coax line. The two ribbon conductors are connected together at the far end.

Matching

Standing-wave ratios up to 5 or 6 to 1 are hardly worth worrying about at 4 MHz so

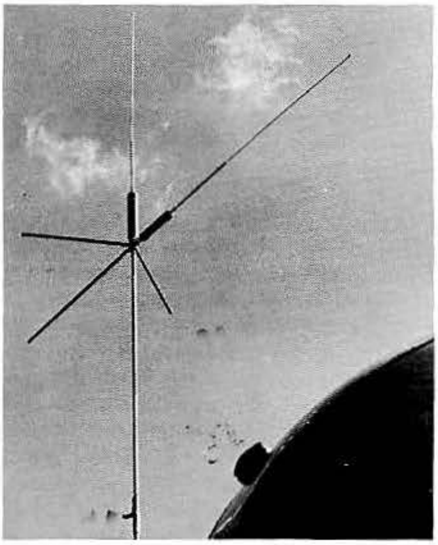

A five-band antenna system. A New-Tronics mobile antenna base section is fitted with a Waters Add-a-Band adapter. The short resonators for 10, 15 and 20 meters are Waters. The 80-meter resonator (top) is New-Tronics. A threaded stud has been added to the vacant side of the adapter to take a New-Tronics 40-meter resonator. The system is fed at the base with a single coax line.

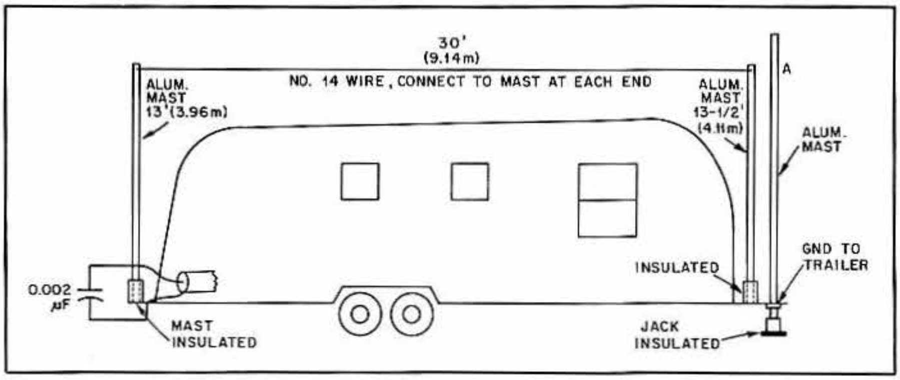

Fig. 23 — 75-meter "Airstream Loop"

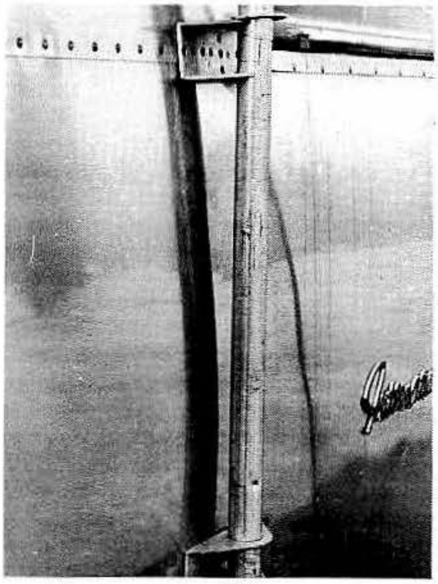

Brackets for mounting TV masting on the side of the trailer. The upper bracket is fitted with a removable U-shaped yoke bolted to the bracket. The brackets are fastened to the trailer with pop rivets.

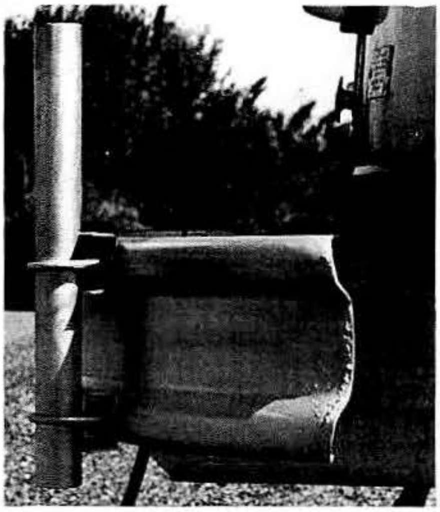

A bumper-mounted socket for TV masting.

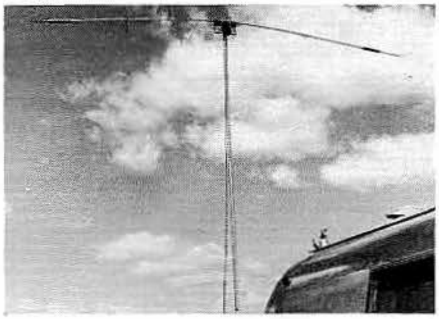

An 80-meter dipole made of two New-Tronics 75-meter mobile whip back to back. The 15-foot (4.57 m) mast is bumper-mounted. A 1-to-1 balun is mounted on the center insulating panel, which is secured to the mast with U bolts. This antenna is quite narrow-band when operated on 75 meters, making it necessary to retune the antenna for excursions of more than 30 or 40 kHz.

far as additional power loss is concerned. However, an SWR greater than 1 to 1 causes the line input impedance to vary, depending on the length of the line. The output circuits of many transmitters and transceivers are not designed to work into loads departing significantly from 50 ohms. With these units, it may not be possible to load the final stage, or components in the output stage not designed for the voltages that develop with higher-impedance loads may be damaged. Use a Monimatch, or other reflectometer, and adjust the antenna for minimum SWR. The author carries along several mica capacitors ranging in value from 50 pF to 0.002 µF with battery clips attached, which are used to shunt the line at the feed point. Start out with the lowest value, and double the capacitance in steps until you get an SWR of 1 to 1 somewhere in the band, and then adjust the antenna length for resonance at the desired frequency.[1]

To get some idea of the relative performances, a series of checks was made at 3825 kHz on several of the antennas described. A receiver with a horizontal antenna was set up at a distance of eight miles from the trailer, and S-meter readings were taken with the results shown in the accompanying tabulation. All readings were made with identical input power to the transmitter, and with the antennas adjusted for an SWR of 1 to 1.

To sum it all up:

1) Carry a complete set of mobile whips for lunch stops, or other occasions where you just do not have the time or room for anything better.

2) If at all possible, use a 1/4-wave counterpoise clipped to the trailer when using a mobile whip.

3) If you have the time or space, clip a 1/4-wave wire, instead of a whip, to your mobile antenna mount, and run it out to a bush or tree.

4) In addition to (3), clip a 1/4-wave counterpoise to the trailer, and run it in the opposite direction.

5) If you are stopping for a short time, or for the night, rig up the 5-band mobile whip with five 1/4-wave counterpoises, one of appropriate length for each band.

6) If you want to concentrate on one band, put up the 16-foot (4.88 m) mobile-whip dipole, if your space is limited.

Table 3

Hustler 75-meter mobile whip mounted vertically on top rear corner of trailer	S7
Same as above, with 60-foot counterpoise wire connected to trailer	S9
Two Hustler 75-meter mobile whips back to back as a horizontal loaded dipole	S9 + 5 dB
60-foot horizontal wire 8 feet high, using trailer as ground	S9 + 10 dB
Hustler 4BTV trap vertical with 75-meter resonator	S9 + 10 dB
120-foot dipole, 15 feet high at center	S9 + 20 dB
"Airstream Loop" antenna	S9 + 20 dB
Home-station dipole 50 feet high	S9 + 30 dB

Feet = 0.3048 = m.

7) If you have plenty of room, put up a 1/2-wave dipole for your favorite band.

8) If you want to work all bands, and space for dipoles is limited, then put up your trap vertical with the 75-meter resonator tuned to your favorite frequency.

9) If you are going to spend a week in a particular campground, shoot some lines over a couple of high trees with a bow and arrow, and get your dipoles up in the air as high as possible.

10) The use of balloons and kites is good, but depends completely on the wind.

11) Next best to a high dipole, put up the "Airstream Loop" antenna for 75 meters.

12) If all else fails, bring your 2-kW PEP linear and your 2.5 kW ac gas generator along. This will at least keep the other campers awake. Don't forget a low-pass filter for TVI! This material was originally presented in *QST* by Philip S. Rand, W1DBM.

---

[1] [Editor's Note: This method is equivalent to the use of a stepdown L network. It can be used when the antenna feed-point impedance is lower than the line impedance. To obtain a match, the antenna length must be somewhat longer than a resonant length, so that the antenna shows inductive reactance at the feed point. (This reactance serves as a substitute for the reactance of the coil of a conventional L network.) For an accurate match, it is necessary to adust the antenna length and the value of the capacitor until the appropriate transformation ratio is obtained.]

The Apartment Dweller's Dilemma

Having trouble getting a signal out from "concrete city"? Nonresonant, random-length wires can provide surprising results if a Transmatch is used to make the rig happy with the load it sees.

It is apparent that many would-be amateurs living in apartment buildings believe it is impossible to get an effective signal on the air. They feel the need to have an outdoor antenna in order to make contacts. This isn't true. Of course, the apartment dweller can always go mobile, but there comes a time when the amateur wants a home station. This article treats the problem and, hopefully, will get a few more hams on the air.

Some General Considerations

There are exceptions to the following rules but, in general, they can be depended upon.

1) An outdoor antenna will work better than an indoor one.

2) An antenna inside a frame building with wood exteriors is better than the same antenna in a steel-and-concrete building.

3) The higher above ground, inside or out, the better the antenna will work.

4) The bigger (or longer) you can make an indoor antenna, the better — even if it means running the wire around corners.

5) Even a poor antenna *should* produce some contacts.

The Coupling Problem

Most transmitters are designed to work into a 50-ohm load, and contain little or no provision for adjusting the transmitter when the load is other than 50 ohms. Unfortunately, there is no random-length wire antenna that will present a 50-ohm load on all bands. What is required is a Transmatch. A Transmatch is simply an adjustable LC network that converts the unknown antenna impedance to 50 ohms. This unit, shown in Fig. 24 and the photographs, will cover the 80- through 10-meter bands and can handle 1 kW of rf power. In addition to matching the transmitter to the antenna load, the Transmatch helps reduce harmonic radiation. Many Novices get in trouble with the FCC because of 2nd-harmonic radiation from the 80-meter signal. This unit should reduce such harmonic radiation. Also, the use of a Transmatch can provide better front-end selectivity for the receiver. Some receivers are subject to severe cross modulation when operated near a broadcast station. The Transmatch should reduce this effect.

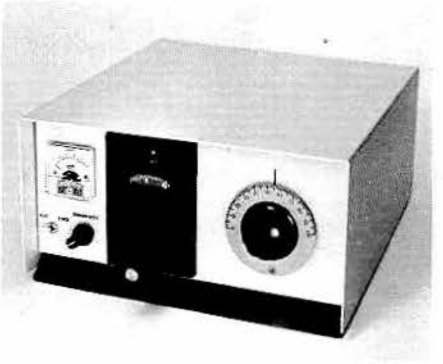

The completed Transmatch. The counter dial and knob are James Millen Mfg. Co., Inc. products. Check with G. R. Whitehouse, Amherst, NH 03031.

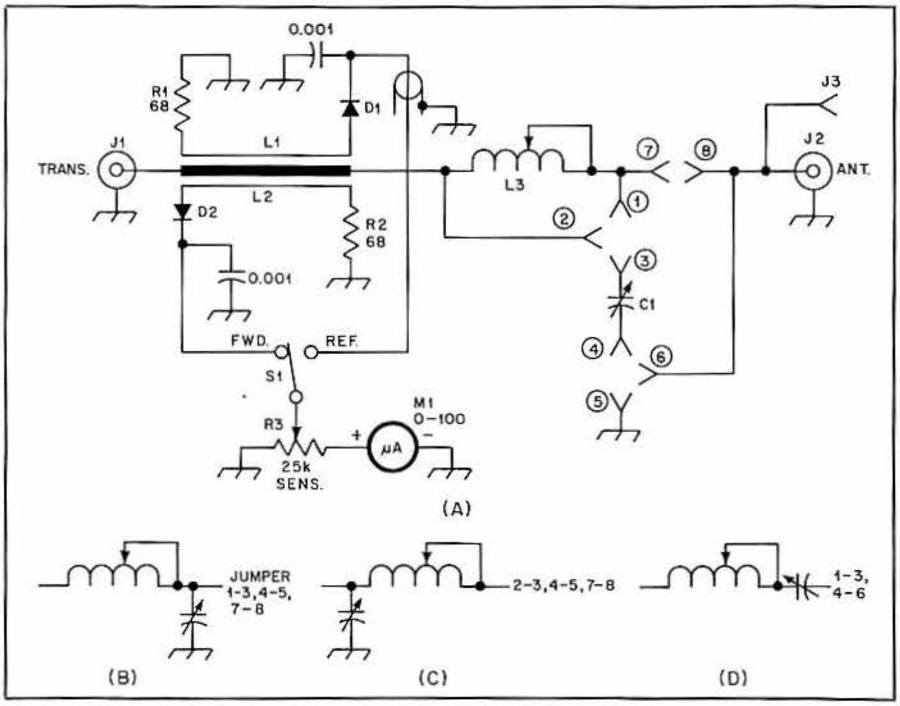

Fig. 24 — Circuit diagram of the L-network Transmatch. The eight banana jacks are E. F. Johnson type 108-900, and three dual banana plugs are required, E. F. Johnson type 108-200.

C1 — Variable capacitor. 350 pF (E. F. Johnson 154-10).

D1, D2 — 1N34A germanium diode.

J1, J2 — Chassis connector, type SO-239.

J3 — Feedthrough terminal, isolantite.

L1, L2 — See Fig. 25, part of etched circuit assembly.

L3 — Variable inductor, 28 μH (E. F. Johnson 220-203).

M1 — 100-μA meter.

R1, R2 — 68-ohm, 1/2-watt carbon or composition, *not wirewound.*

R3 — 25,000-ohm carbon control, linear taper.

S1 — Spst toggle.

Circuit Details

The unit shown in Fig. 24 is designed to be used in three configurations. They are shown at B, C and D. With one of the three hookups, it should be possible to match practically any antenna to the transmitter.

In order to get complete band coverage and avoid the complexities of bandswitching, banana and jack plugs are used to change the circuit to the configuration needed. For example, if one wanted the setup shown at B, he would jumper terminals 7 and 8, 1 and 3, and 4 and 5. Using the banana plugs makes for easy changing of the circuit.

Whenever a Transmatch is used, the operator should have a way of knowing when the unit is adjusted correctly. The answer to this need is a Monimatch or other SWR indicator. If you have an SWR meter, it can be connected in the line between the transmitter and the coupler. If

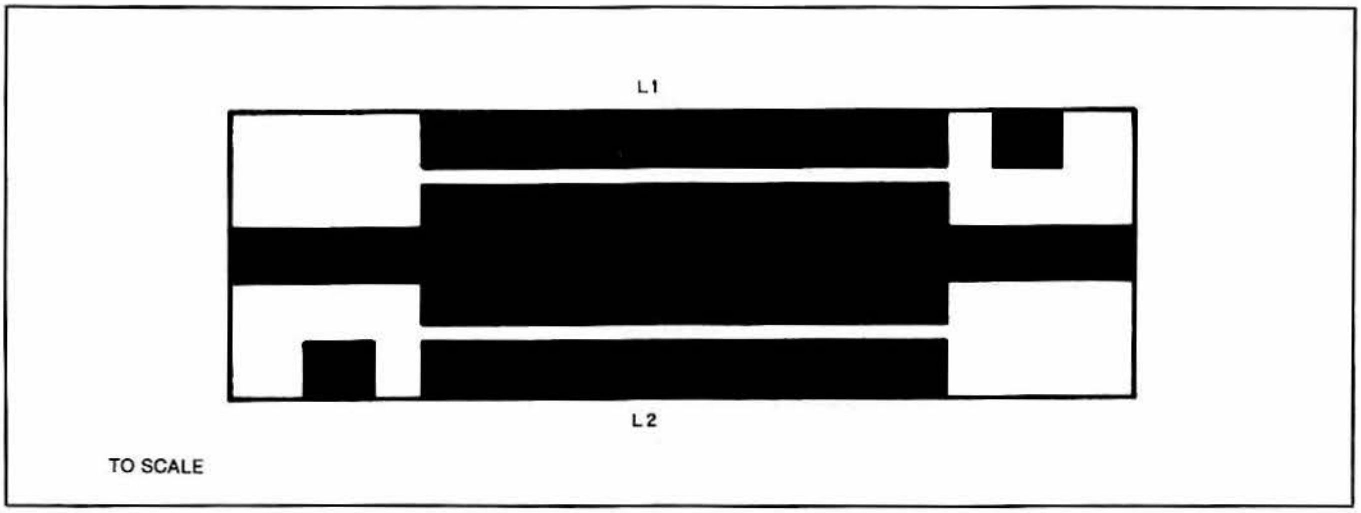

Fig. 25 — Etched circuit-board template. The foil side is shown, with etched portion shaded.

not, you can build the Monimatch[1] shown in Fig. 24, and include it in the cabinet with the Transmatch.

Construction Details

The chassis for mounting the Transmatch is made from a piece of aluminum measuring 10 × 19 inches (254 × 482 mm). The ends of the 19-inch length of aluminum are bent up to form a U-shaped chassis, the ends being 4-1/2 inches high to form a chassis 10 × 10 × 4-1/2 inches (254 × 254 × 114 mm). The back side of the U has an opening cut out, 3-1/4 inches high by 4-1/2 inches (83 × 114 mm) long. A piece of Plexiglas is mounted over this opening. The jack-plug sockets are installed directly on the plastic. Connections from the roller inductor, L3, and variable capacitor, C1, are made to the banana jacks. Be careful when drilling the holes for the jacks to insure that they will mate with the plugs. Fig. 25 shows the details for a pc-board Monimatch.

Methods for making etched circuit boards are given in detail in the construction chapter of The Radio Amateur's Handbook, so we won't treat the process here. When installing D1 and D2 on the board, be sure to use a heat sink while soldering the leads. Too much heat can ruin the diodes. Shielded leads are used from the circuit board to S1 to prevent unwanted rf pickup.

Today's equipment builder soon finds out that locating parts can be a tough problem. Fortunately, we have found a distributor who will furnish all the parts. With the exception of the chassis, the parts for the Transmatch can be purchased from Barry Electronics.[2]

Some Antenna Ideas

With this Transmatch one should be able to match any random-length antenna to his transmitter. To test the system, we strung up 25 feet (7.62 m) of wire in the ARRL lab (a steel-and-concrete building, first floor). The Transmatch provided the desired 50-ohm impedance on all bands. We didn't get any "60-over-9" reports, but we *did* make contacts.

As to your antenna installation, we suggest running the wire up to the ceiling, around the room, perhaps into the next room, making the antennas as long as possible. If the XYL objects to the wire, use an invisible antenna. Nope, we're not

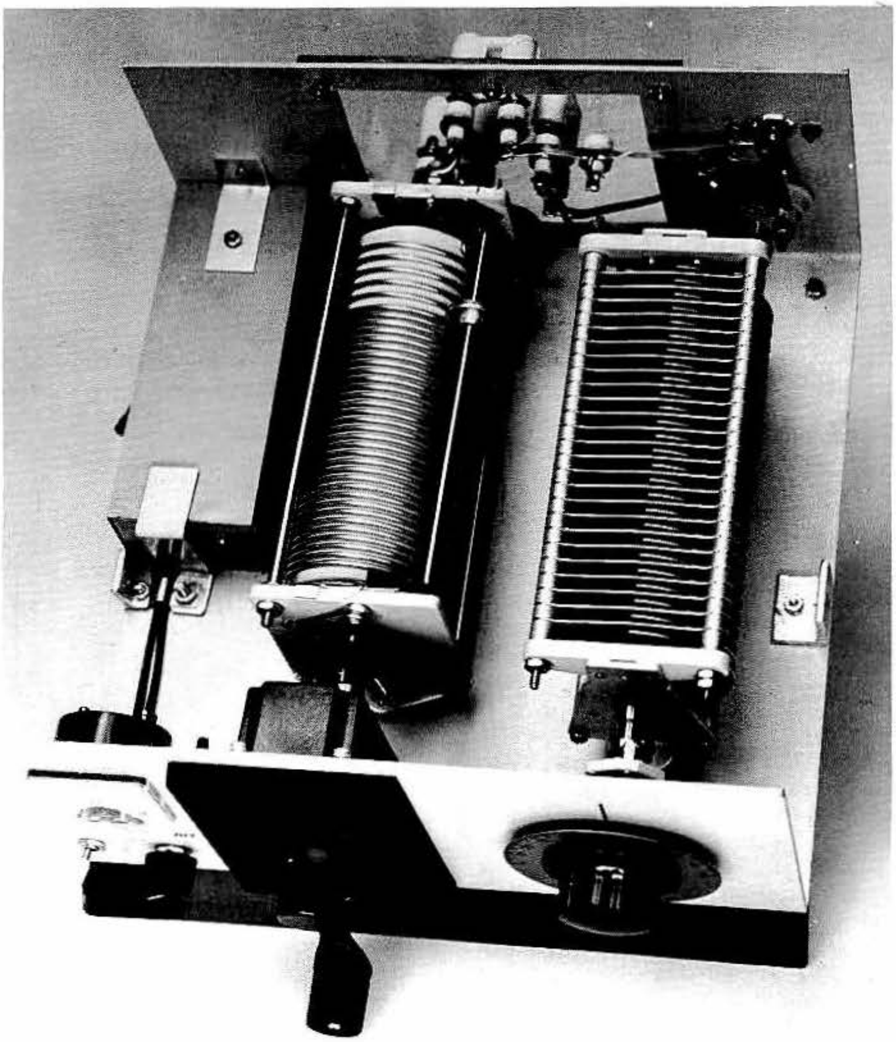

The Monimatch is at the upper left, covered by a metal enclosure. Connections from the roller inductor and the variable capacitor to the terminals on the jacks are made with thin strips of copper, although no. 12 or 14 wire can be used instead. The two antenna terminals are at the rear right. The top terminal is for use with a coax-fed antenna, if desired.

joking! You can make the antenna from no. 26 or 28 wire, which will be practically invisible. Also, you can conceal the wire behind molding. Be sure to insulate the wire if it runs near any metal. In some cases, it is impossible to get the wire outside. However, if you have a screen in one of your windows, terminate the end of the antenna at the screen. The screen *is* outside and it will help the system radiate. While it may be difficult to install, the antenna length to shoot for is approximately 120 to 130 feet (36.58 to 39.62 m) because this is a good length for multiband operation.

If you cannot put an antenna on the apartment roof, there is still a trick or two for having an outside antenna. You can drop a wire out the window and let it hang down. How long the wire should be will depend on how high up you are. You can make the wire long enough to reach nearly to the earth, but high enough to be beyond the reach of passersby. *Safety first,* always! There is no electrical law that states that a vertical must be fed at the bottom. Just connect the end of the wire to your Transmatch and tune up the system. You can use small-diameter wire. Use a lead sinker on the end of the wire and possibly imbed the sinker in a small sponge-rubber ball. (That way, you won't break your neighbor's window.)

If there is a support to which you can attach the far end of the antenna you can use the invisible-wire trick. Use rubber bands for insulators. Erect no. 26 or 28 wire. Don't be too discouraged if the wire gets broken by passing birds!

We know one ham who lived in a basement flat. He ran a wire through the wall to the bottom of a rain gutter, tuned up the system, and managed to work DXCC. (Safety first, again. The downspout should be beyond the reach of human beings!) The general idea is to use ingenuity. With a Transmatch you never can tell *what* will prove to be a useful antenna.

How to Tune Up

Using the Transmatch is not complicated. Although it takes some time to find the correct combination of settings, once determined, they can be logged for later reference. Use a short length of 50-ohm coax to connect the Transmatch to the transmitter. Attach the antenna to the Transmatch. Tune up your transmitter on the desired band, making sure that the final amplifier is resonated, but with the power output reduced. With the Monimatch in the forward-reading position, set the sensitivity control for a full-scale reading. You may have to increase the output of the transmitter to get full deflection, but be sure to keep the final amplifier tank in resonance. Switch the meter to the reflected position, and then adjust L1 and C1, until you get the lowest indication of reflected power. It should be possible to get the meter to read zero. When you have a zero reading in the reflected position, versus full scale in the forward setting, the Transmatch is correctly adjusted, and the SWR is 1. You may have to change the circuit to one of the other configurations in order to get a match, but one combination should work. Once you have the Transmatch set properly, you can then adjust the transmitter to its rated power input. One other point: It isn't always possible to get a good ground connection in an apartment. Therefore, a connection to a cold-water pipe or earth ground should be used. This material was originally presented in *QST* by Lewis G. McCoy. W1ICP.

References

[1] McCoy, "An Etched-Circuit Monimatch for Checking Your Antenna System," *QST,* October 1969. (See Chapter 5.)

[2] Barry Electronics, 512 Broadway, New York, NY 10012.

A 40-Meter DDRR Antenna

This antenna consists of a 1/4-wavelength element grounded at one end and wound into a single-turn coil, a few conductor diameters above the ground (see Fig. 26).

Shortly after the author's article on a practical DDRR Antenna appeared,[1] a number of inquiries were received regarding low-frequency versions of this antenna for operation at ground level. After a little experimenting, the author came up with this version of a 40-meter DDRR which works quite well. In early tests very fine QSOs were held with stations as far away as Phoenix, while the antenna was located inside the garage and surrounded by myriad metallic objects. When the antenna was relocated to the backyard, marked improvement was noted, even though this antenna site was marred by a sharp rise of 4 feet (1.22 m) in ground elevation, plus close proximity to house wiring and power lines. In spite of these obstacles solid contact was easily achieved with stations up and down the West Coast. Encouraged by these results, we moved the antenna to a rooftop location which put it well in the clear insofar as metallic objects above the groundplane were concerned. In this location, the DDRR really proved its worth. In spite of its low profile, a little over 1 foot (0.31 m), and its very small span, a few inches over 9 feet (2.74 m), it was more than adequate when competing with any signal on the band. The obvious conclusion from my experiments is that the 40-Meter DDRR is the apartment-dweller's dream. It is principally for that group that this article is prepared. Other interested amateurs might be those who are limited, as this author is, by too much house on too little real estate; or those who for other reasons cannot cope with high towers, masts and guy wires.

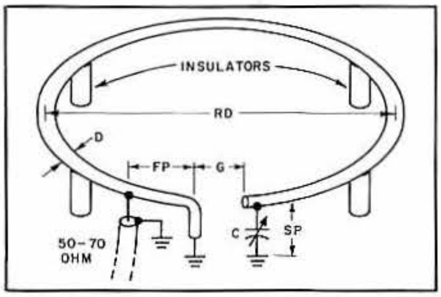

Fig. 26
RD = 0.078 λ (28°)
SP = 0.11D (2.5°)
FP = 0.25 h (See Note 1)
C = (See Note 2)
D = (See Note 3)
G = (See Table 4)

Notes: (1) Actual dimension must be found experimentally. (2) Value to resonate the antenna to the operating freq. (3) d ranges upward from 1/2" (13 mm). The larger "d" is the higher efficiency is. Use largest practical size, e.g., 1/2" (13 mm) for 10 meters, 5" (127 mm) or 6" (152 mm) for 80 or 160 meters.

Before this writer's enthusiasm sends you out to rip down your inverted V or to dismantle your beam, remember this: The DDRR, for all of its capabilities, will not supplant a full-size single-frequency antenna which is properly erected over clear terrain. What it will do is provide an

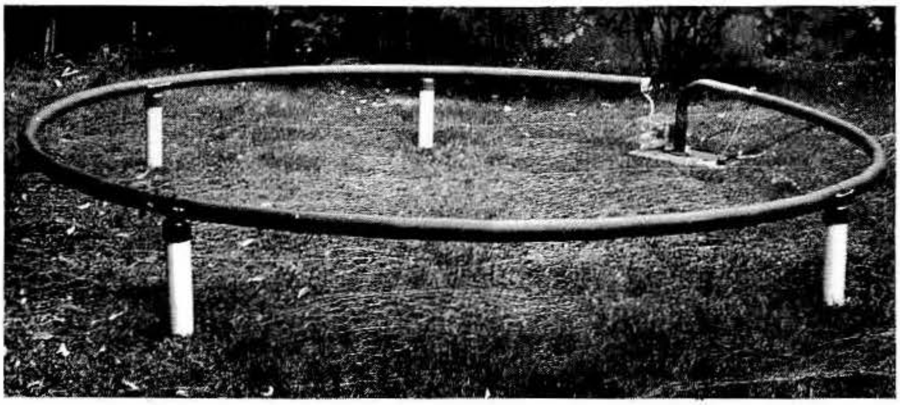

The chicken-wire groundplane is evident in the background. The base plate can be seen at lower right. Note the relative positions of the 52-ohm coaxial feed at the left end of the plate, the flange on the foot of the post, and the tuning unit at the right-hand end of the plate.

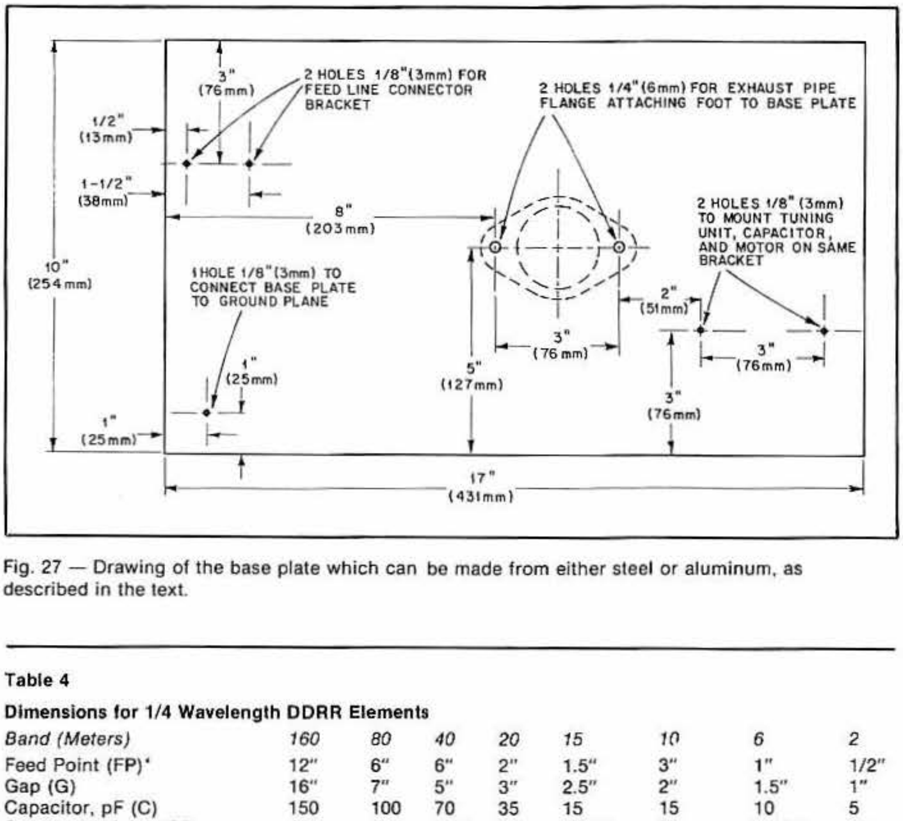
3" (76mm)
2 HOLES 1/8"(3mm) FOR FEED LINE CONNECTOR BRACKET
2 HOLES 1/4"(6mm) FOR EXHAUST PIPE FLANGE ATTACHING FOOT TO BASE PLATE
1/2" (13mm)
1-1/2" (38mm)
8" (203mm)
2 HOLES 1/8" (3mm) TO MOUNT TUNING UNIT, CAPACITOR, AND MOTOR ON SAME BRACKET
10" (254mm)
1 HOLE 1/8"(3mm) TO CONNECT BASE PLATE TO GROUND PLANE
2" (51mm)
3" (76mm)
3" (76mm)
1" (25mm)
5" (127mm)
3" (76mm)
1" (25mm)
17" (431mm)

Fig. 27 — Drawing of the base plate which can be made from either steel or aluminum, as described in the text.

Table 4

Dimensions for 1/4 Wavelength DDRR Elements

Band (Meters)	*160*	*80*	*40*	*20*	*15*	*10*	*6*	*2*
Feed Point (FP)*	12"	6"	6"	2"	1.5"	3"	1"	1/2"
Gap (G)	16"	7"	5"	3"	2.5"	2"	1.5"	1"
Capacitor, pF (C)	150	100	70	35	15	15	10	5
Spacing (Height) (SP)	48"	24"	11"	6"	4 3/4"	3"	1 1/2"	1"
Tubing Diameter (D)	5"	4"	2"	1"	3/4"	3/4"	1/2"	1/4"
Ring Diameter (RD)	36'	18'	9'	4.5'	3'4"	2'4"	16 1/4"	6"

*See Fig. 26 for explanation of designations.

Feet X 0.3048 = m.

Inches X 25.4 = mm.

The braided lead across the flared joint is to assure electrical continuity. The screws used are self-tapping sheet-metal screws. The top end of the insulator (2-inch [51 mm] diameter PVC pipe) is a standard PVC pipe cap; the bottom is closed off with a cap from an aerosol can (2-inch [51 mm] ID).

antenna which will enable communications of respectable quality, where heretofore it has been impossible because of inadequate space to erect a 40-meter antenna.

Some consideration should also be given to the fact that the high-Q nature of the DDRR and its resultant narrow-band characteristics serve to reduce the noise level. Boyer[2,3] reports that in the initial experiments it was found that DX stations which could copy signals from either a vertical or a DDRR could only be heard on the DDRR due to the reduction in background noise. So if you have a noisy location, it might be to your benefit to try the DDRR, regardless of what antenna you are presently using.

Constructing the 40-Meter DDRR

In this application, 2-inch (51 mm) diameter automobile exhaust pipe was used as the radiating element. The local muffler shop not only supplied the material, but also undertook to bend it to specifications. This was an obvious course since the material and the power bender were right at hand.

The dimensions for 40 meters are:

Ring — 9 foot (2.74 m) diameter, center to center.

Height — 12 inches (0.31 m) from ground plant to element center.

Gap — 6 inches (152 mm) from upright post center to open end of ring.

In forming the ring to these dimensions, four 10-foot (3.05 m) lengths of tubing were used. A 10-degree bend was made at 9-inch (229 mm) intervals in three of the lengths. The fourth length was similarly treated, except for the last 18 inches (0.46 m) which were bent at right angles to form the upright leg of the ring. One end of each section was flared so that the sections could be coupled together by slipping the end of one into the flare of its mate.

The required flares are easily made at the muffler shop with the aid of the forming tools. Another task which can best be completed at the shop is to weld a flange onto the end of the upright leg. This flange is to facilitate attaching the leg to the mounting plate which provides a chassis for the tuning mechanism and the coaxial-feed coupler. After bending and flaring is complete, the ring is assembled and minor adjustments made to bring it into round and to the proper dimensions. This can best be done by drawing a circle on the floor with chalk and fitting the ring inside the circle. The circle must be slightly larger than the center-to-center diameter so that the reference line can be seen easily. For example, with two-inch (51-mm) tubing the actual diameter of the reference circle must be 9 feet, 2 inches (2.79 m). When you

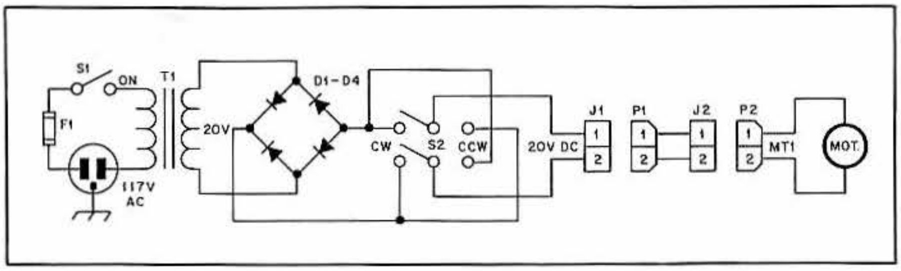

Fig. 28 — Circuit diagram of the power supply and control motor.

D1 through D4 — 1 A, 50 PRV, or equiv.
F1 — Fuse, 1 A.
J1, J2 — Two-terminal jack.
MT1 — Motor (see text).
P1, P2 — Two-terminal plug.
S1 — Spst.
S2 — Dpdt, center off.
T1 — Transformer (see text).

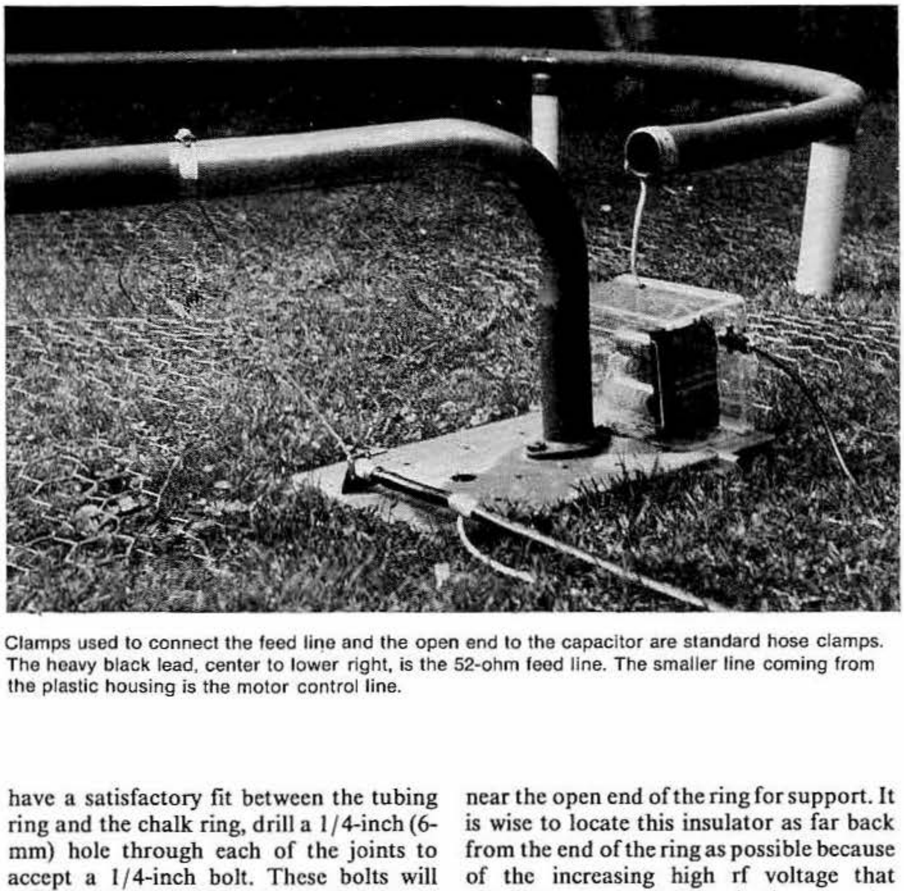

Clamps used to connect the feed line and the open end to the capacitor are standard hose clamps. The heavy black lead, center to lower right, is the 52-ohm feed line. The smaller line coming from the plastic housing is the motor control line.

have a satisfactory fit between the tubing ring and the chalk ring, drill a 1/4-inch (6-mm) hole through each of the joints to accept a 1/4-inch bolt. These bolts will clamp the sections together. Also, they can be used to attach the insulators which support the ring at a fixed height above the groundplane.

Making and Attaching the Insulators

Insulators for the antenna were made from 11-inch (279-mm) lengths of 2-inch (51-mm) PVC pipe inserted into a standard cap of the same material. The PVC caps are first drilled through the center to accept the 1/4-inch (6-mm) bolt previously installed at the joints. The caps are then slipped onto the bolts and nuts are installed and tightened to secure the caps in place. The 11-inch length of pipe, when inserted into the cap and pressed firmly until it touches bottom, results in a total insulator length of 12 inches (0.31 m). Four insulators are required: one at each of the joints and one near the open end of the ring for support. It is wise to locate this insulator as far back from the end of the ring as possible because of the increasing high rf voltage that develops as the end of the ring is approached.[4] As a final measure, the bottom ends of the insulators were sealed to prevent moisture from forming on the inside surfaces. Standard PVC caps could be used here, but we found that plastic caps from 15-ounce aerosol cans fit well.

Making the Mounting Plate

The mounting plate is required to provide good mechanical and electrical connections for the grounded leg of the radiator, the coaxial feed-line connection, and the tuning mechanism. If you are using aluminum tubing, you should use an aluminum plate, and for steel tubing, a steel plate to lessen corrosion from the contacting of dissimilar metals. Dimensions for the plate are shown in Fig. 27. The important consideration here is that good, solid mechanical and electrical connections are made between the ground side at the coaxial connector, the ring base, and the tuning capacitor.

The Tuning Unit

We found that the 9-foot (2.74 m) ring resonated easily with approximately 20 pF of capacitance between the high end of the ring and the base plate or ground. A 35-pF double-spaced variable from the junk box was pressed into service here (Cardwell NG-35-DS). Any variable which will tune the system to resonance and which will not arc under full power should be satisfactory. Remember, the rf voltage at the high impedance end of this antenna can reach 20 to 30 kV with high power, so if you are using the maximum legal limit, you would do well to consider using a vacuum variable capacitor. Since we limited our power to 500 watts PEP, the double-spaced Cardwell unit was satisfactory. To provide for full band coverage, the capacitor was coupled to a reversible, slow-speed motor which enabled the antenna to be remotely tuned from the operating position. An indicated SWR of 1.1 to 1 was easily achieved over the entire 40-meter band. The motor used was a surplus item made by Globe Industries of Dayton, OH.[5] At 20 volts dc the shaft of this motor turns at about 1 rpm which is ideal for DDRR tuning. The gears used were surplus items. If you cannot obtain gears, string and pulley drive will do almost as well, or you can mount both the motor and the capacitor in line and use direct coupling. Of course, if you operate on a fixed frequency or within a 40- to 50-kHz segment of the band, you can dispense with the motor entirely and simply tune the capacitor manually. In any case, the tuning unit must be protected from the weather. We used a plastic refigerator box to house the tuning capacitor and its drive motor.

Fig. 28 shows the electrical connection for the motor. A small train transformer or power supply for toy slot-cars will work admirably as a tuning motor power source. Standard ac zip cord was used for the connection between the control unit and the motor.

Electrical Connections and the Groundplane

The connection between the open end of the ring and the tuning capacitor is made with no. 12 wire or larger. On the end of the base plate opposite the tuning unit, and directly under the ring about 8 inches (203 mm) from the grounded post, install a bracket for a coaxial connector. The connector should be oriented so that the feed line will lead away from the ring at close to 90 degrees. Install a clamp on the ring directly above the coaxial connector. Connect a lead of no. 12 or larger wire from the coaxial connector to the clamp. This wire must have a certain amount of flexibility to accommodate the movement necessary when adjusting the match. The matching point must be found by experimentation. It will be affected by the nature and quality of the groundplane over which the antenna is operating. The antenna will function over earth ground; however, in our location we found the electrical ground to be unpredictable. A groundplane surface of chicken wire (laid under the antenna and bonded to the base plate) provided a constant ground reference and improved performance. In a rooftop location sheet metal roofing should provide an excellent groundplane. A poor ground usually results in a matching point for the feed line far out along the circumference of the circle. In our installation a near-perfect match was obtained with the feed line connected to the ring about 12 inches (0.31 m) from the grounded post. During testing, when the antenna was set up on a concrete surface without the groundplane, a match was found when the feed line was connected nearly 7 feet (2.13 m) from the post!

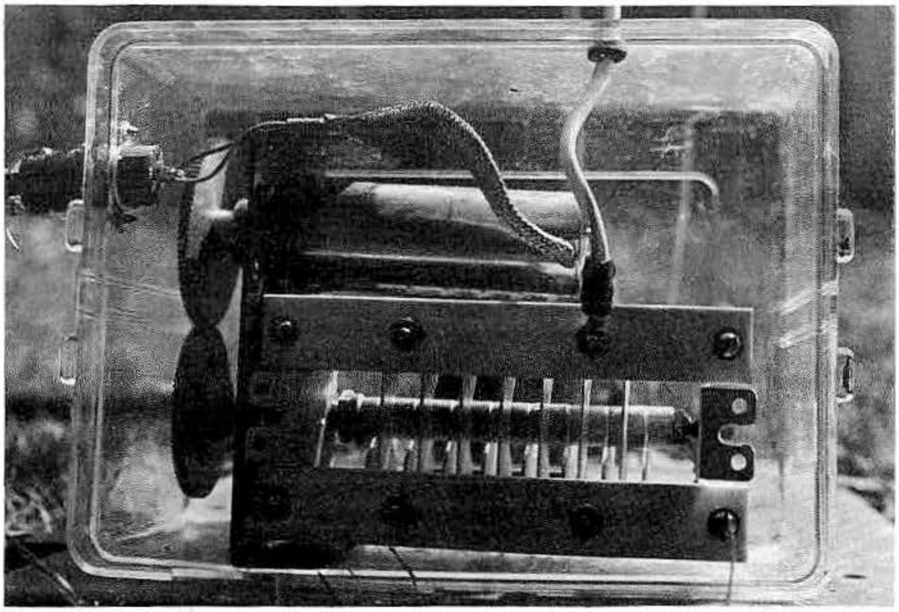

The tuning motor, above the capacitor, couples to the capacitor shaft through the gears at left. Both the motor and the capacitor are mounted on a common bracket which attaches to the base plate by two bolts through the weatherproof housing. The lead passing upward through the grommet connects the stator of the capacitor to the open end of the ring.

As shown in photos, the compactness of the antenna is readily apparent. The groundplane is made up of three 12-foot (3.66-m) lengths of chicken wire, each 4 feet (1.22 m) wide, which are bonded along the edges at about 6-inch (152-mm) intervals. In our installation the antenna, with the groundplane, could be dismantled in about 30 minutes. If portability is not important, it is best to bond all of the joints in the tubing so that good electrical continuity is assured.

After all construction is completed, the antenna should be given a coat of primer paint to minimize rust. If it suits you, there is no reason why a final coat of enamel could not be applied.

Tuning Procedures

Once the mechanical construction is completed, the antenna should be erected in its intended operating location. Coupling to the station may be accomplished with either 52-ohm or 72-ohm coaxial cable. Tune and load the transmitter as with any antenna. While observing an SWR meter in the line, operate the tuning motor. Indication of resonance is the noticeable decrease in indicated reflected power. At this point, note the loading of the transmitter; it will probably increase markedly as antenna resonance is approached. Retune the transmitter and move the feed-point tap on the antenna for a further reduction in indicated reflected power. There is interaction between the movement at the feed tap and the resonance point; therefore, it will be necessary to operate the tuning motor each time the tap is adjusted until the lowest SWR is achieved. Don't settle for anything less than 1.1 to 1. With a good ground and proper tuning and matching, this ratio can be achieved and maintained over the entire band. Once the proper feed point has been located, the only adjustment necessary when changing frequency is retuning the antenna to resonance by means of the motor. If the antenna is to be fixed tuned, provide an insulated shaft extension of 18 inches (457 mm) or so to the tuning-capacitor shaft for manual adjustment. This not only provides insulation from the high rf voltage but also minimizes body-capacitance effects during the tuning process.

Alternatives

A number of materials other than the steel tubing used here are well suited for the ring element. Standard E.M.T. or electrical conduit would work as well with a slight increase in weight. One advantage to be gained through the use of conduit is the elimination of the flaring operation since standard couplings would serve to connect each segment of the ring to its adjacent member. Another suitable material is copper tubing. This material is superior to either exhaust pipe or conduit in terms of its conductivity characteristics. Another advantage of copper is that it is available in continuous lengths and the joints could be omitted entirely.

The 2-inch (51 mm) dimension is by no means mandatory. Smaller diameter tubing has been used with satisfactory results.[6] In fact, DDRR antennas have been fabricated with wire elements. But, if the element diameter is reduced, the antenna tunes more *sharply*. Some experimenters may wish to go in the other direction and use a larger diameter. The author recommends the use of aluminum downspout with a diameter of about 4 inches (2 mm). This material does not lend itself to bending, however, and the ring must be configured as a regular polygon of eight or more sides. Because of the large number of joints involved, welding is about the only practical means of joining the segments. Unless you are equipped to do this work yourself, the cost of welding might be prohibitive. Anyone who undertakes to make a DDRR antenna of a large element diameter will be rewarded in terms of improved performance.

Performance

Results have been quite encouraging, and it is hoped that more and more hams will equip themselves with the DDRR in the future. The antenna has proved its worth and deserves more investigation by the amateur fraternity than it has been given in the past. No intensive efforts have been made to work DX with the antenna described here: however, a low angle of radiation is conducive to DX, and this antenna demonstrates a low radiation angle. It was found that distant areas, such as the East Coast, are more easily contacted than are stations nearby. All of the results could be attributed to peculiarities of individual stations or skip conditions, but since they are charateristics which can be anticipated with high-efficiency antennas having low radiation angle, we prefer that interpretation. Besides, where else will you find an antenna for 40 meters that is small enough to fit into a corner of the backyard and not protrude above the fence; or for that matter, which could be mounted on the roof of an apartment building or even a ranch-type house and not be visible from the street?

Nearly everyone who listened to the description of this new antenna was enthusiastic. The author hopes to hear many hams on the air working with the DDRR in the near future. This material was originally presented in *QST* by W. E. English, W6WYQ.

Parts List

4 — 10-ft (3.05 m) lengths of 2-in. (51 mm) tubing, exhaust pipe, conduit or copper tubing.
1 — base plate 7-1/2 X 19 X 1/8 inches (191 X 482 X 3 mm), steel, aluminum, or copper to match tubing.
4 — PVC pipe caps for 2-inch (51 mm) pipe.
4 feet (1.22 m) of 2-inch PVC pipe.
2 clamps for 2-inch tubing.
4 — 1/4 X 4-inch (6 X 102 mm) bolts with nuts.
1 reversible motor with 1-rpm shaft output.
1 wide spaced variable capacitor. 5 to 35 pF (Millen 16550 or equiv.).
1 coaxial connector SO-239.
2 sets 2-connector plug and socket for motor control.
36 feet (10.97 m) of 4-foot (1.22 m) chicken wire or equivalent.
1 flange to attach tube to base plate.
2 — 1/2 X 1-inch (13 X 25 mm) bolts with nuts (flange mounting).
8 — 3-inch (76 mm) bonding strips no. 8 braided wire.
4 plugs 2-inch (51 mm) ID for insulator bases.

References

Hicks, "The DDRR Antenna: A New Approach to Compact Antenna Design," *CQ*, June 1964.
Horn, "The Half Wavelength DDRR Antenna," *CQ*, September 1967.
The Radio Amateur's Handbook, 46th ed., The American Radio Relay League.

Notes

[1] English, "A Practical DDRR Antenna," *73*, June 1970.

[2] Boyer, U.S. Patent Nos. RD 26196, RE 3,151,328 (All rights assigned to the Northrop Corporation.)

[3] Boyer, "Hula-Hoop Antennas: A Coming Trend?" *Electronics*, January 11, 1963.

[4] Editor's Note: Because of the danger of rf burns, in the event of accidental contact with the antenna, precautions should be taken to prevent random access to the completed installation.

[5] Globe C-5A-1106 (available from Lectronic Research Labs Inc., 75 Arch St., Philadelphia, PA 19106).

[6] See footnote 1.

Novi-Loop

A medium-size roll of wire, some insulators and a day of good antenna "puttin' up" weather can net you a Novi Loop. What will it do for you? Well, follow these instructions and get set to work 40-meter DX. Nuff said!

One of the sad facts of life for many beginners is the matter of poor antenna performance. Many enthusiastic Novices have fallen behind in their efforts to work DX because of bitter disappointment caused by poor performance in the antenna department. All too often it is suggested that a simple end-fed hunk of wire is ample for all-around Novice band operation, or that a dipole will suffice for most Novice operation. Sure, almost any kind of radiator that can be tuned to the operating frequency and marched to the impedance of the feed line and transmitter will enable the operator of a low-power station to make some contacts from time to time. It depends on how high the system is above ground, how free and clear it is of surrounding objects that can detune the radiator and absorb the energy radiated from it, and upon the condition of the band at a given time. It should stand as a firm rule, then, that any amateur who wants good results should erect the best antenna system he can, consistent with available space, finances and structural practicality. It is possible to advance beyond the plateau of simple dipoles and random-length wires in the quest for better performance, and without going to the expense and effort of erecting large directional beams such as the cubical quad or Yagi.

What will be described here can be applied to any of the high-frequency amateur bands. However, the example given relates to only the 40-meter Novice band. The antenna is a full-wave loop of wire, mounted through necessity at W1FB in a rectangular format, and matched with a simple quarter-wavelength coaxial transformer. It has given excellent performance in DX work and for making "solid" stateside contacts. Total cost for the antenna system (wire, matching transformer and insulators) is under $12, exclusive of the feed line used to connect the antenna to the rig.

Antenna Features

Full-wave, closed-loop antennas are broadband, low-Q devices. This is a handy feature because it permits the operator to move his transmitter frequency within a particular band without need to compensate for an increase in standing-wave ratio (SWR) which might otherwise occur if other types of antennas were used. (The higher the Q of an antenna, the narrower its bandwidth, even though an SWR of 1 can be obtained at some frequency within the band for which it is built.) The foregoing rules out the need for a Transmatch with the loop antenna, provided the feed method recommended here is applied.

Another interesting feature of this kind of antenna is the theoretical gain of approximately 2 dB it exhibits over a half-wave dipole. The angle of radiation from a properly erected full-wave loop is considered to be lower than that of a dipole when both are less than one-half wavelength above ground. This feature suggests the superiority of the loop in situations where significant height above ground is not a practical goal for the chap wishing to put up a 40-meter antenna.

A full-wavelength closed loop need not be square. It can be trapezoidal, rectangular, circular, or some distorted configuration in between those shapes. For best results, however, the builder should attempt to make the loop as square as possible. The more rectangular the shape, the greater the cancellation of energy in the system, and the less effective it will be. The effect is similar to that of a dipole whose effectiveness becomes impaired as the ends of the dipole are brought closer and closer together. The practical limit can be seen in the "inverted-V" antenna, where a 90-degree apex angle between the legs is the minimum value used. Angles that are less than 90 degrees cause serious cancellation of the rf energy.

The loop can be fed in the center of one of the vertical sides if vertical polarization is desired. For horizontal polarization it is necessary to feed either of the horizontal sides at the center. At the time of this writing there have been no data compiled to provide a comparison between the performance, vertical versus horizontal. Such an experiment could be interesting (and possibly productive) for the Novice.

Erecting the Antenna

Optimum directivity occurs at right angles to the plane of the loop, or in more

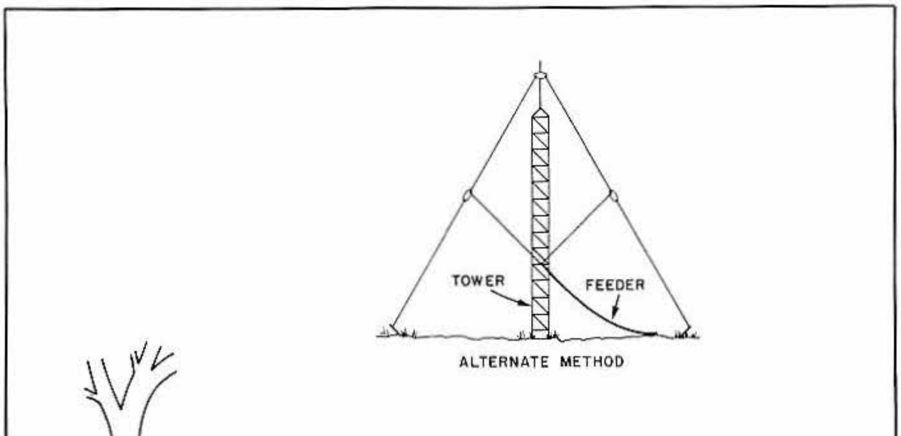

Fig. 29 — Details of the full-wave loop. The dimensions given are for operation at the low end of 40 meters (7050 kHz). The height above ground was 7 feet (2.13 m) in this instance, though improved performance should result if the builder can install the loop higher above ground without sacrificing length on the vertical sides. The inset illustrates how a single supporting structure can be used to hold the loop in a diamond-shaped configuration. Feeding the diamond at the lower tip provides radiation in the horizontal plane. Feeding the system at either side will result in vertical polarization of the radiated signal.

simple terms, broadside from the loop. Therefore, one should try to hang the system from available supports which will enable the antenna to radiate the maximum amount in some favored direction. The bidirectional pattern is maximized NE and SW at the writer's QTH. This gives good results in working Europeans and the stations "down under." Excellent signal reports have been obtained off the sides of the loop when working within the USA.

Just how the wire is erected will depend on what is available in one's yard. Trees are always handy for supporting antennas, and in many instances the house is high enough to be included in the lineup of solid objects from which to hang a radiator. If only one supporting structure is available it should be a simple matter to put up an A frame or pipe mast to use as a remaining support. (Also, tower owners see Fig. 29 inset.)

The overall length of the wire used in a loop is determined from the formula

$$\frac{1005}{f\ (\mathrm{MHz})}.$$

Hence, for operation at 7125 kHz the overall wire length will be 141 feet (42.98 m). The matching transformer, an electrical quarter wavelength of 75-ohm coax cable, can be computed by dividing 234 by the operating frequency in MHz, then multiplying that number by the velocity factor of the cable being used. Thus, for operation at 7125 kHz

$$\frac{234}{7.125\ (\mathrm{MHz})} = 32.84 \text{ feet}$$

If coax with solid polyethylene insulation is used, a velocity factor of 0.66 must be employed. Foam-polyethylene coax has a velocity factor of 0.80. Assuming we are to use the solid dielectric coax, RG-59/U, the length of the matching transformer becomes 32.84 (feet) × 0.66 = 21.67 feet, or 21 feet 8 inches (6.605 m). The transformer will convert the 100-ohm feed-point impedance of the loop to 50 ohms, a convenient value for connecting the system to the transmitter. The loop described here was measured with the K1TD Macromatcher[1] and found to exhibit an impedance of 110 ohms. Some inductive reactance was indicated. By removing 1 foot (0.31 m) of wire from the overall loop the reactance dropped to zero, though the antenna loaded up fine prior to the final trimming. The feed impedance at 21 MHz was also 110 ohms, but considerable inductive reactance was observed. Most certainly one could use the antenna on 15 meters with reasonable success, despite the SWR which might be present.

Results

During two weeks of testing the loop gave an excellent account of itself. It was compared with the writer's two-element short beam for 40 meters, and the results indicated that the loop was better than the beam on some occasions. At other times the beam provided better performance, or was at least equal to the loop. The differences noted were the result of band conditions and the time of day. High-angle signals can cause misleading results when comparing antennas for hf-band use. More meaningful tests can be made on ground-wave contacts, however, and several checks made between the loop and the beam indicated that radiation at right angles to the plane of the loop compared closely with that from the front of the beam. The beam is rotatable and installed 40 feet (12.19 m) above ground.

With 150 watts dc input to the transmitter PA, and operating at the low end of 40 meters (loop was cut for 7050 kHz), the first night of operation resulted in the logging of three Russian stations, two Polish ones, two Yugoslavians, one Italian station and two English ones. The lowest report was RST 549. The highest report was RST 599. Generally speaking, the loop has outperformed the beam during most long-haul contacts. Though no effort has yet been made to work South American stations (off the side of the loop), they have been heard loud and clear, as have some VK stations via long path at approximately 2300 UTC.

It is quite likely that the Novice who chooses to erect his own Novi-Loop will have a commanding signal on 40 meters. If backyard space permits, it should be a worthwhile venture. Don't forget, the formulas and guidelines given here are applicable to the other hf Novice bands as well. This material was originally presented in *QST* by Doug DeMaw, W1FB.

---

[1] Hall, "The Macromatcher," *QST*, January 1972.

Chapter 5

Antenna Theory and Test Methods

In order to obtain top performance from a given style of amateur antenna, it is important to understand the basic principles of operation. Additionally, some test procedures are necessary if one is to ensure that his or her antenna is functioning at its best. This chapter contains assorted data on antenna theory, plus a number of common-sense test methods.

These reprints of popular *QST* articles were chosen for all amateurs, regardless of their technical-aptitude levels. There's something to honor in the phrase, "Learn by doing." This chapter will help to make that philosophy a reality.

Quad vs. Triband Yagi

The controversy about quads working better than Yagis, or vice versa, may never end. But, we'll bet our newest roll of RG-8/U that you'll find some interesting reading on the subject in this report.

While stationed in Japan, a sort of DX crossroads of the world, this writer had the opportunity to observe, firsthand, the excellent performance of the cubical quad in competition with the Yagi, dipoles and an assortment of other antenna systems. In pursuing this undertaking certain steps were necessary to insure than any conclusions made would be meaningful. With this in mind a play emerged.

Objectives

In the many articles written on the cubical quad, it is noteworthy that only on a few occasions have the authors been privileged to compare the quad with other types of antennas on a real-time basis, and from the same operating location. Furthermore, when such comparisons were made, the authors generally compared against some type of monoband antenna system. A casual scanning of the 10-, 15- and 20-meter phone bands would lead one to conclude that the triband Yagi enjoys a rather high position of popularity among the antennas in general use. This being the case, it appeared that a worthwhile contribution to the data already available on the Yagi and quad might be made by conducting a series of controlled comparative tests, employing the triband Yagi and the quad. The test objectives were then defined: to compare various configurations of a cubical quad antenna with a representative commercial triband Yagi; such tests to be conducted over short, medium, and long transmission paths, and to arrive at conclusions regarding the relative merits of each antenna.

Table 1

	Mod. 1	*Mod. 2*	*Mod. 3*	*Mod. 4*
Reflector Element	72'3"	70'4"	72'5"	72'5"
Driven Element	69'	70'4"	70'5"	70'5"
Director Element	—	—	—	69'1"
Spacing "A"	7'6-1/2"	8'5"	13'4"	13'
Spacing "B"	—	—	—	13'
Stub	20-30"	34-38"	—	—

Feet X 0.3048 = m.

Inches X 25.4 = mm.

Test Plan and Procedure

Every effort was made to conduct the tests in a manner which would lessen the possibility of compromising the techniques employed by either the writer or participating stations:

1) The test to be performed by establishing communications with Amateur Radio stations located throughout the world on a random and scheduled basis.

2) Amateur Radio stations volunteering to assist in this effort to be briefed on conduct of test and data desired.

3) A voice ssb transmission to be made to the participating station, identifying the first antenna used as antenna "A."

4) The voice transmission to be followed immediately by an unmodulated carrier for a period of approximately 5 seconds.

5) The antennas would be switched, and a voice transmission made identifying the antenna as "B," and the procedures above repeated.

6) Participating stations will note signal strength related to each antenna, and provide a numerical value as observed on his S meter or other indicating device. These values to be logged, and the test reinitiated with another volunteer station.

Equipment Preparation

1) Antenna heights to be as nearly identical as possible.

2) Center of antenna horizontal lobe patterns to be as nearly identical as possible when pointing the antennas toward a participating station.

3) Resonant frequency of each antenna to be matched as closely as possible.

4) Transmission lines to be matched to antennas and transmitter loading to be as nearly identical as possible with each antenna.

5) Instantaneous transfer of antennas.

6) Relative power and SWR to be monitored continuously.

7) Prior to and after each data gathering session, equipment parameters will be verified. If a significant deviation in any parameter is noted, data collected will be discarded.

Antenna performance conclusions to be based on an analysis of data derived from a minimum of 50 unmodulated-carrier observations with each antenna configuration, and supplemented with data gathered during conventional ssb QSOs.

Table 2

	Mod. 1	*Mod. 2*	*Mod. 3*	*Mod. 4*
Total Observations	50	60	60	52
Less than 2100 miles	12	2	3	3
2100 to 4800 miles	33	31	33	32
Greater than 4800 miles	5	27	24	17
Signal Difference				
More than 1 S unit better	—	—	—	—
Less than 1 S unit better	—	—	7	9
No discernible difference	1	5	51	43
Less than 1 S unit poorer	27	46	2	—
More than 1 S unit poorer	22	9	—	—

Antenna Selection

This writer had been using a four-element commercial triband Yagi (boom length 24 feet [7.32 m], and 55 feet [16.76 m] above ground) for approximately 1-1/2 years, so the properties of this antenna were fairly well established. Furthermore, in on-the-air comparisons with competitive models of triband Yagis in use by other U.S. amateurs operating from the Tokyo area, the antenna appeared representative of commercial triband antennas in general use by the amateur community. Therefore, the Yagi in use at the author's station was selected as the reference antenna.

Text material concerning quad antennas, available to the author in Japan, was reviewed. It became evident that there are almost as many variations in quad design as there are writers on the subject. After much deliberation, and many discussions with amateurs throughout the world, the decision was made to test three models of the quad (a fourth model was tested as will be noted later). Since the physical characteristics of the quad are fairly standard, only the dimensions of the elements and the spacing between them was considered. The dimensions for the three models tested were obtained from a Japanese manufacturer of cubical quads, from Orr's book, *All About Cubical Quad Antennas,*[1] and from Dr., J. E. Lindsay, Jr., W0HJ.[2]

Preliminary Testing

Several days were spent "dry running" the test plan to validate the concept, and to smooth out the operating procedures and techniques. Of particular concern was the possible time required to make a valid data-gathering observation. If data were to be reasonably accurate, the transmission path had to be stable, and the signal-strength observations must be taken on each antenna during a short period of time. The dry runs were valuable in this respect.

A problem became evident during the first day of testing. It appears that those of us who speak and understand English do not always convey the same message when using the same words. As a result it was necessary to modify the verbal format, utilizing simple sentences and placing them in a logical sequence.

It also became apparent that the test could not be conducted under all transmission-path conditions; that even under ideal conditions several observations were often necessary before a conclusive report could be compiled. It was decided to conduct the tests only on 20 meters. The operating time available to the writer favored openings on 20 meters to Europe via the long path, and to Australia, the U.S. and various islands in the Pacific. It was also decided to orient the test antennas so that the topography and obstructions seen by each antenna would be essentially the same. (Physical separation between the two antennas was in the order of one wavelength.)

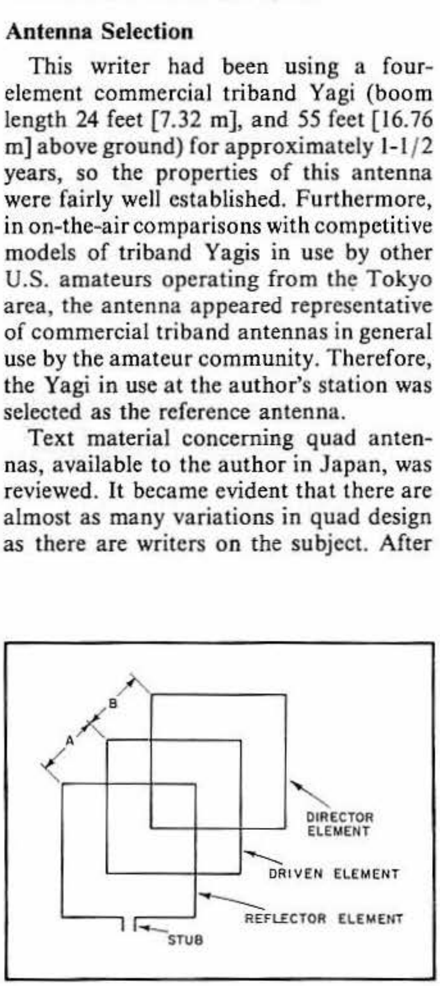

Fig. 1 — Element spacing information for Table 1.

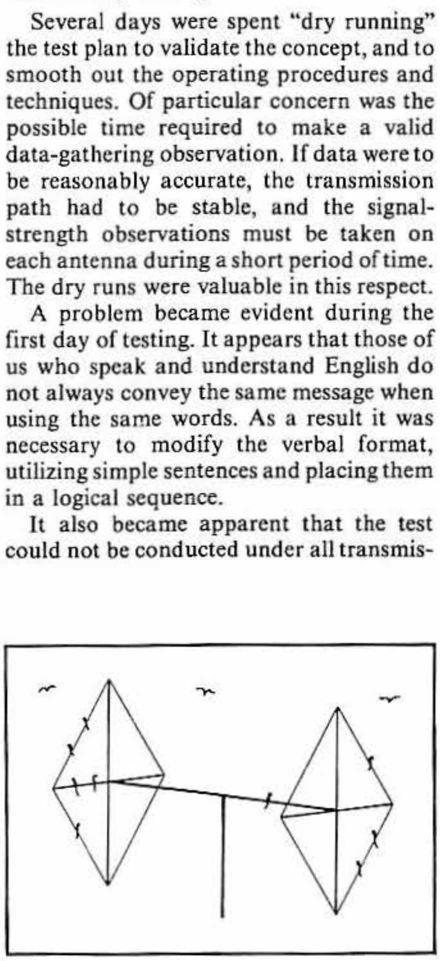

Fig. 2 — CLOSE SPACING IS FOR THE BIRDS!

Testing

Dimensions of the first quad model selected were furnished by a Japanese manufacturer of cubical-quad antennas. (see Table 1). The antenna was assembled, utilizing commercially manufactured heavy-duty hardware and fiberglass spreaders. It was tuned to a center frequency of 14,200 kHz. Testing of the first model began in November 1967 and continued for one month. The results for this period are given in Table 2.

In mid-December 1967 the first quad was replaced by a model constructed according to the formula and dimensions given in Orr's book. The results obtained with model 2 are contained in Table 2.

Construction of the third model (with wider element spacing) was carried out next. Two matching systems (Gamma and Q-section) were experimented with on this antenna. A satisfactory match could be had with either system. However, the Q-section was used for the test because it was the technique used with the previous two quad models (SWR with each antenna was never more than 1.3:1 with a difference between antennas no greater than 0.1). The results conducted with this model were most enlightening, as shown in Table 2. The model antenna was also used extensively during the first weekend of the 1968 ARRL DX Contest. Though these contacts were not used in tabulating test samplings, it is interesting to observe that openings to the U.S. (using the quad) lasted 15 to 30 minutes longer on each end of the period than with the Yagi. It is assumed that this phenomenon would also apply to each of the other quad models.

The fourth quad tested was a 3-element wide-spaced model constructed according to more dimensions furnished by W0HJ. The results of the samplings were somewhat disappointing and are given in Table 2. (Frankly, the author felt that the 3-element quad would show a substantial improvement over the Yagi in every case.) The 3-element model did appear to have a better front-to-back and front-to-side ratio than either the Yagi or the other quad models. One positive comment: The 3-element model is a monster to assemble and put up! In the author's opinion the difference in performance isn't worth the small improvement. Perhaps, on the other hand, if one accepts the 2-element model as the departure point between a simple mechanical structure and a major project, a 4-element model might be more worth the effort. However, this is purely conjecture on the part of the author.

Summary

The antenna tests indicate that:

1) One can expect to achieve the same or better results with a 2-element quad of proper dimensions than with a 3- or 4-element triband Yagi.

2) A wide-spaced quad will perform substantially better than a close-spaced quad.

3) Dollar for dollar, the quad appears to be a better investment than a Yagi.

Acknowledgements

The writer wishes to thank all of the amateurs who participated in the series of tests, particularly the VK gang, who night after night tolerated the request for observations. The support couldn't have been better, and on many occasions, upon completing a check with a particular station, several other stations would call to give their observations (which were taken during the same transmission test). This material was originally presented in *QST* by Col John H. Parrott, Jr., W4FRU.

References

[1] Orr, *All About Cubical-Quad Antennas*, Radio Publications, Wilton, CT 06897.

[2] Dimensions later published: Lindsay, "Quads and Yagis," *QST*, May 1968. (See chapter 3).

High Versus Low Antennas Performance Tests Using Identical Arrays

This section compares the performances of identical antennas mounted side-by-side at different heights, and produces some experimental evidence to support the practical importance of antenna height for most types of Amateur Radio communication.

"The higher the better" is the traditional rule of thumb for ham radio antennas. Few amateurs would disagree with this axiom, and most experienced antenna builders can recall how much better they got out after "raising the beam another 15 feet (4.57 m)." However, few empirical studies of the effect of antenna height on signal strength have been published in amateur circles.

This is unforunate, since several amateurs have studied the question methodically, including Dr. J. E. Lindsay, who has now published his definitive treatise on quads and Yagis[1] but not his excellent work in this area. And while both *The ARRL Antenna Book* and Orr's *Beam Antenna Handbook*[2] discuss the importance of antenna height in theoretical terms, neither reports the results of practical studies in the field.

With this in mind, the author set out to study the effect of height on signal strength at various distances, on several bands, and with various forms of propagation. Only amateur equipment was available for the study, but in other respects, the author strived to keep the methodology as rigorous as possible.

The Method

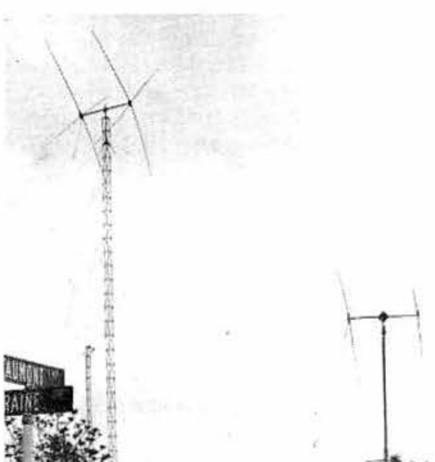

The two cubical quads used by the author to study the effects of antenna height on performance. The antenna at right is atop a 34-foot (10.36 m) tilt-over mast described by the author in June 1969 *QST*. The quad at left is on a 72-foot (21.95 m) tower.

The author's approach was to erect identical antennas atop two towers of different heights and obtain comparative signal reports — with some safeguards to minimize reporting errors and to assure proper statistical treatment for the resulting data. For the main tests, one tower was 72 feet (21.95 m) high (the practical maximum for the author's residential backyard at the time of the tests), while the other stood 34 feet (10.36 m) high (about the minimum usable height for DX work, according to Orr). The two towers were placed as far apart as possible without trespassing — about 50 feet (15.24 m). With this separation, no interaction was evident on any band.

A pair of 2-element cubical quads were selected for the 10-, 15- and 20-meter experiments. Each had an eight-foot (2.44 m) boom with all elements mounted concentrically on two sets of spreaders, and each antenna was fed with a single feed line (85 feet [25.9 m] of RG-8/U in both cases). This design obviously involves some compromises in element spacing and impedance matching, but both quads developed good SWR curves and front-to-back ratios when tuned. And, more important for our study, whatever compromises existed were essentially the same for both the high and low antennas.

It became apparent during the tests that many amateurs regard cubical quads as exceptional performers at low heights, but (unlike Yagis) not much better if raised higher. This popular impression is contrary to Lindsay's findings, which suggest that quads and Yagis respond quite comparably to changes in their height. Accepting Lindsay's conclusions, we assumed that our findings would be roughly the same if Yagis had been used for the primary experiments.

Table 3

Comparative Performance Data for High and Low Antennas

Band/path	No. reports	No. favoring high antenna	No. favoring low antenna	No. reporting no difference	Mean high-antenna advantage
20/DX	16	16	0	0	2.1 S units
20/U.S.	18	15	0	3	1.4 S units
15/DX	12	12	0	0	1.8 S units
15/U.S.	36	20	0	16	0.9 S units
15/ground-wave	12	12	0	0	2.2 S units
10/DX	8	7	0	1	1 S unit
6/ground-wave	5	5	0	0	2 S units
6/E-skip	10	2	7	1	−1.3 S units
2/ground-wave	16	16	0	0	3.3 S units

The author wanted to determine how well high and low antennas would perform not only on DX work, but on "Stateside" F-layer communications, ground-wave work, and E-skip as well. Thus, separate tabulations were kept for DX reports (i.e., those from stations more than 4000 miles away) and for reports from "Stateside" stations (i.e., F-layer reports over 1500—3000-mile paths) on each frequency band.[3] The author was careful to avoid seeking reports during anything that resembled an E-skip opening on 10, 15, and 20 meters, in the interest of studying E skip vs. antenna height separately on 6 meters.

To study the effect of height on vhf groundwave signals, the author replaced the cubical quads on the two towers with two pairs of seven-element 2-meter Yagis, each fed in phase with ratio-dipole driven elements. A move to a new home prevented the use of the same two towers for 6-meter experiments, but the author later set up two 3-element 6-meter Yagis — one atop a new 90-foot (27.43 m) tower, and the other on a 42-footer (12.8 m) — to continue the study.

On each band, the tests were conducted on ssb (except on 2 meters, where a-m was used), with an audio tone serving as a reference signal. The author switched between the high and low antennas repeatedly on each test, in an effort to counteract the effects of QSB.

This test procedure does introduce some sources of error, including the variability of hams' skills in observing the results, and the lack of receiver S-meter standardization. To cope with this kind of sampling fluctuation, the author obtained a number of reports on each band and then used t-distribution statistical methodology to evaluate the resulting data. Even with these efforts to make the study as rigorous as possible, however, the author makes no claim to finality for the results obtained. The findings on each band did prove to be statistically significant (at the 0.05 level of significance or better), but you don't "prove" things or arrive at exact parameters this way. The only claim the author makes for his results is that they suggest general trends that will probably be confirmed if this study is replicated elsewhere. There is no implication here that a 72-foot (21.95 m) high antenna produces "12-dB gain" over a 34-foot (10.36 m)-high antenna in 20-meter DX work just because the mean advantage in our tests was about two S units in this instance.

The Results

The results of these tests, summarized in the accompanying table, were interesting in several respects. With the exception of one case on 10 meters, every single DX station reported a substantially better signal from the high quad than the low one. This is thoroughly predicted in the literature, but we also found that the high quad produced a significantly stronger signal (taking the mean, or average, figure) on "Stateside" work. And on ground-wave work, the high antennas were absolutely consistent in outperforming lower ones. However, E-skip contacts proved to be another story, a point to be covered in detail later. The mean advantage of the high quad over the low quad in DX work was 2.1 S units on 20 meters, 1.8 S units on 15, and one S unit on the 10-meter band.

Since the author's linear amplifier (a commercial unit using two 3-400Zs) produces a signal gain of about two S units over the exciter alone on most S meters, it became apparent that many DX stations could not distinguish the kilowatt feeding the lower quad from the exciter "barefoot" with the higher array. The implications of this dramatic difference in signal strength for DX-contest operations are obvious: it's no wonder the "big guns" using antennas 130 feet (39.62 m) high get through the pileups quickly, with or without big linears.

On Stateside F-layer work, the pattern was less consistent, although the high array continued to hold a significant edge. While a majority of the reports still favored the high quad by one or two S units, more than a third of the stations dissented, reporting equal or virtually equal signals from the two quads. No one reported a stronger signal from the lower antenna, perhaps because of the author's effort to avoid E-ship contacts in this phase of the study. The mean advantage of the higher quad was 1.4 S units on 20-meter Stateside work, and 0.9 S unit on 15 — still a substantial advantage, but not enough to render a low antenna useless for portable work, as generations of Field Day operators can attest. In fact, the author previously described a portable quad[4] that stood only 19 feet (5.79 m) high, but nevertheless worked very well for "back-home" contacts during a long vacation. An antenna only 19 feet (5.97 m) high is clearly undesirable, but if higher antennas are not feasible, much successful hamming can still be done. (This is especially true on an occasion like Field Day, when almost everyone's antenna is low, and the few beams that are 40 [12.19 m] or more feet high seem to dominate the bands; if the competition is using a low antenna you can too!)

Ground-wave Tests

Although the plan was to study the effect of antenna height on ground-wave communication only on vhf bands, a number of stations within 100 miles volunteered reports on the hf bands. These reports were not tabulated except on 15 meters, but they consistently favored the higher array by at least two S units.

In the vhf experiments, height again proved decisive for gound-wave work. On both 2 and 6 meters, every station favored the high antenna. On two meters, the mean advantage was an amazing 3.3 S units. Surprisingly, the higher array appeared to be equally superior on contacts with San Bernardino-Riverside stations (some 30 miles from Beaumont, CA, where these tests were conducted) and on contacts into the Los Angeles basin (almost 100 miles away).[5]

After moving and acquiring the 90-foot (27.43 m) tower, the author put a single 7-element 2-meter Yagi up and took an S-meter reading on a nearby repeater with the new tower cranked down. As the tower ascended to its full height, the signal rose from S8 to 15 dB over S9 on a popular transceiver's S meter.

On 6 meters, only a few ground-wave tests were conducted, but the results were similar: Everyone favored the higher antenna substantially.

6-Meter E-Skip Tests

However, on 6-meter E skip an entirely different relationship developed: The *lower* antenna had the advantage in seven out of 10 tests during single-hop openings.[6] Only one station favored the high antenna, and the low array had a mean advantage of 1.3 S units over the high array!

Although this finding is surprising and contrary to every other test conducted, it is consistent with the one previous practical study of antenna height vs. signal strength the author could locate in ham radio literature. Working in the late 1940s, Leroy May[7] set up identical 6-meter Yagis at heights of 35 and 75 feet (10.67 and 22.8 m), and compared 350 received signals on the two arrays. He found that the low array was superior for distances below 1000 miles while the higher array began to pull ahead at distances exceeding 1100 miles.

The author made no attempt to fully replicate May's work, but all of our 6-meter E-skip experience supports him: For contacts into many major cities from greater Los Angeles, a high antenna is a handicap! Much to his chagrin, the author has frequently seen his 90-foot (27.43 m)-high Yagi outperformed on 6-meter E skip by a beam sitting 10 feet (3.05 m) above somebody's roof on a TV mast.

Conclusions and Recommendations

It is customary when reporting a study such as this to conclude by summarizing the findings and making some recommendations. The conclusion here must be that, for every application except E-skip work, there are worthwhile advantages of having an antenna at least 70 feet (21.34 m) high. This height advantage is most pronounced on ground-wave and long-haul DX work.

The only logical recommendation we can make, then, is that all hams should own towers 70 feet (21.34 m) high. May all of the XYLs, frightened neighbors, bill collectors, and zoning-conscious city officials of the world forgive us!

Alternately, maybe we should suggest that all hams (except the author) use 35-foot (10.67 m) towers to give the author a better chance in DX contests. Or better yet, how about an antenna-height multiplier for the DX contests? Let's see, that's 1.5 if your antenna is below 75 feet (22.86 m), 2.0 if it's below 40 feet (12.19 m), and 3.0 if you bury it in the ground . . . This material was originally presented in *QST* by Wayne E. Overbeck, N6NB.

References

[1] Lindsay, "Quads and Yagis," *QST*, May 1968 (See chapter 3).

[2] Orr, William I., *Beam Antenna Handbook*, 3rd Edition.

[3] No 10-meter "Stateside" tests are reported because the muf was too low for reliable F-layer work across the U.S. when these tests were run.

[4] Overbeck, "The 20-Minute Portable Quad," *QST*, May 1967.

[5] On the hf bands, it is normally assumed that a high antenna has a greater advantage over a low antenna on long-haul DX than it does on shorter paths, but the opposite may well be true for vhf ground-wave work. This is suggested in the *ARRL VHF Manual*, (1st Edition, pages 24,25). Thus, one would expect that an increase in height would result in greater improvement in signals over the 30-mile path than over the 100-mile path.

[6] At the time of the 6-meter tests, E skip was barely "in season." The author is continuing this study during seasons of more frequent E-skip openings.

[7] May, "Some Practical Results on the Question of Antenna Height," *CQ*, April 1953.

Limited-Space Antennas and Methods Of Coupling

When short, random-length wire antennas prove disappointing, chances are the installer didn't follow a few basic rules for good performance. These tips should lead the way to better results with "randoms."

One very popular type of antenna, probably because of its simplicity, is a wire of either a predetermined or random length, fed at one end. It is common procedure among amateurs these days to refer to such an antenna as a "long wire." Actually, this term is usually incorrect because, by definition, a long-wire antenna is a wire that is several wavelengths long at the operating frequency. But no matter what you call it, the end-fed wire can be a very simple antenna to put up and use, if certain steps are taken.

Transmitter Loads

Nearly all transmitters described these days are designed to work into 50- to 70-ohm loads, with little thought given to any load that departs very far from these figures. If the load is something different than the above figure it may be impossible to get the amplifier in a transmitter to load up. In order for the final amplifier stage to operate at its best efficiency, the load (the antenna system) must be within the design range of the amplifier.

An end-fed wire may present a load of 50 to 70 ohms, but if it does it will be a case of pure luck on the part of the user — and what's more important, that load won't stay the same across a band. In other

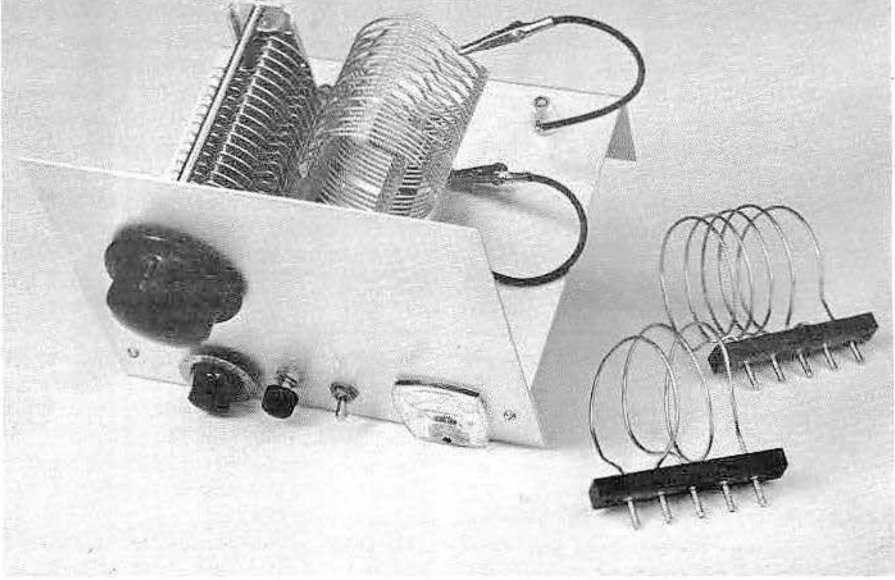

The completed transmatch with the 80/40-meter coil plugged in. At the rear is the 20-meter coil and the 10/15-meter unit is to the front. In regular operation, the antenna would be attached to the standoff insulator at the top rear of the chassis. The clip lead visible below the coil is the one attached to the jack bar.

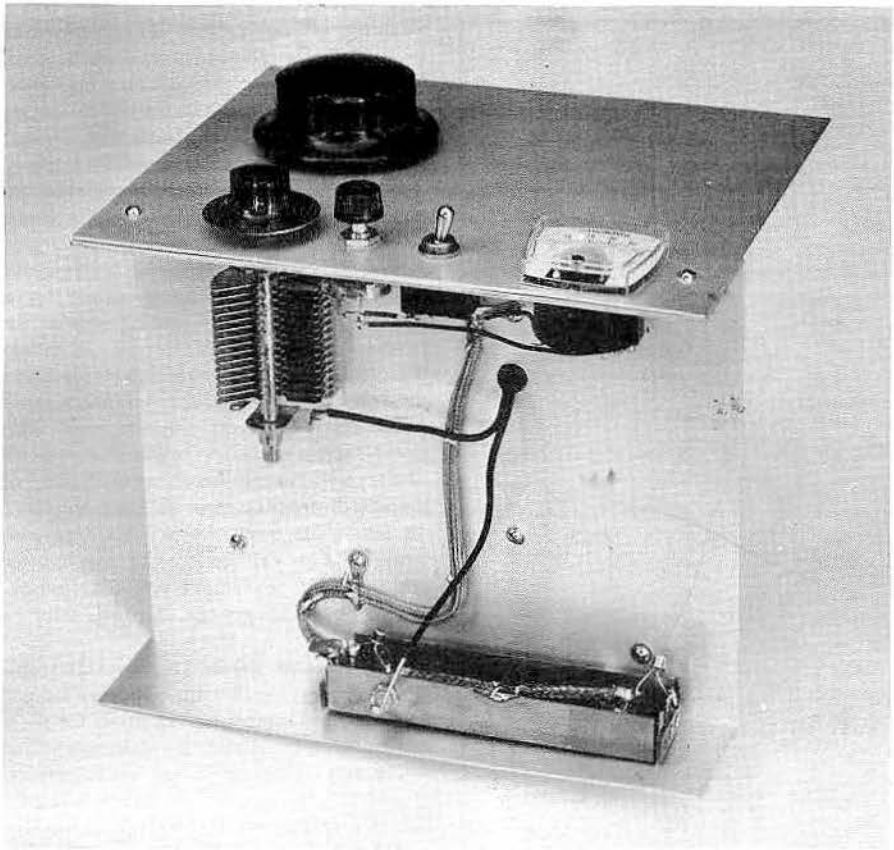

The wiring below deck. At the rear of the chassis is the Varimatcher section. C2 is at the left front of the chassis.

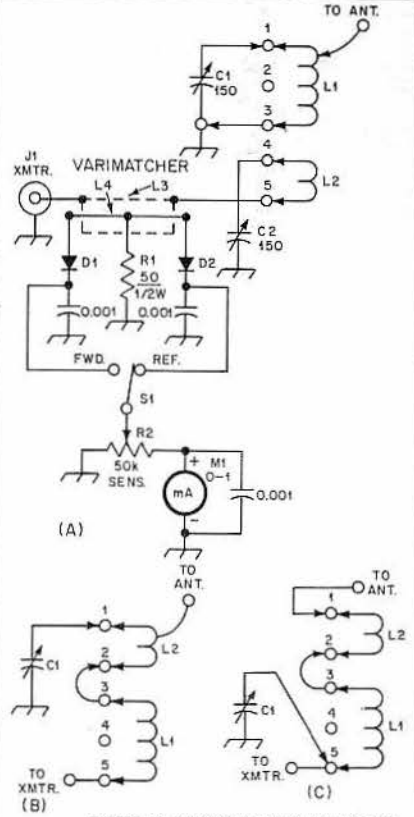

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (µF); OTHERS ARE IN PICOFARADS (pF OR µµF); RESISTANCES ARE IN OHMS; k = 1000, M = 1000 000.

C2 — 150-pF variable (Millen 19140 or similar).
D1, D2 — 1N34A germanium diode.
J1 — Coax chassis fitting, type SO-239.
L1, L2 — See Fig. 7.
L3, L4 — See Fig. 6.
C1 — 150-pF variable, (air gap 0.077 inch [2 mm] or larger for high power, Millen 12515 or similar; air gap 0.040 inch [1 mm] for low power, Millen 19140W or similar).
M1 — 0-1 milliammeter (more sensitive meter can be used).
R1 — 50-ohms, 1/2-watt carbon or composition, not wire-wound.
R2 — 50 kΩ control, linear taper.
S1 — Single-pole, single-throw toggle switch.

Fig. 4 — Circuit diagram of the Transmatch and Varimatcher. All 0.001-µF capacitors are disk ceramic.

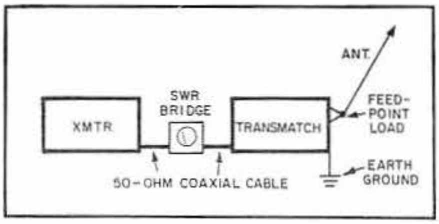

Fig. 3 — This block diagram shows the set-up for using a Transmatch. Not shown is an antenna relay which normally would be inserted in the coaxial line. If a low-pass filter is used, it should be the last item in the line before the Transmatch.

words, the antenna will present a 50- to 70-ohm load for only a very limited frequency range. Once we QSY the load may change to where it is no longer within the tuning range of the amplifier. Our problem boils down to having the transmitter *always* see a load of 50 to 70 ohms, regardless of the band or frequency in use. This in turn means we must make the antenna system look like a 50- to 70-ohm load, no matter which band we use it on.

The Transmatch, an Adjustable RF Transformer

Let's make one point clear about end-fed wires, or any antenna, for that matter: There is no way of predicting what the feed-point impedance (load) will be before making and installing the antenna. We can come close, but there are many factors that get into the act to determine what the impedance will be. The antenna's height, its proximity to nearby objects, its length, and the antenna's relation to earth ground are just a few of the factors. If we don't know what the feed-point impedance is, how can we make the transmitter always work into a 50-ohm load? The easiest way is to insert a Transmatch between the transmitter and the feed point. A Transmatch is simply an adjustable rf transformer that takes our unknown feed-point load and converts it to a desired load, 50 ohms in the case shown in Fig. 3.

Note in Fig. 3 that we show a random-length antenna and an earth ground connection. In this case, our unknown load exists between the end of the antenna and ground. On the transmitter side of the Transmatch we have 50-ohm coaxial cable and inserted in this cable is a 50-ohm reflectometer or SWR (standing-wave-ratio) bridge. The reflectometer is simply a measuring instrument that tells us when our Transmatch is properly adjusted to convert our unknown load to a 50-ohm load. If the Transmatch has adequate matching range, we can convert any feed-point load to 50 ohms. And this means that we can take any random length wire and make a working antenna out of it for any amateur band or frequency. There are some practical limits as to how short a wire we can use, but actually the antenna can be quite short in relation to a wavelength and still work.

Limited-Space Antennas

Many hams live in apartments or locations where it is impossible to put up an "antenna farm" or even a half-wave dipole for 80 or 40 meters. Random-length wires are sometimes the best bet in these circumstances. For hams who live under such conditions there are some general rules that can be followed for best performance. If possible, always get the antenna up on the roof. If not, get the antenna as high as possible. Make the wire as long as possible, even if it has to go around corners. Don't overlook existing antenna possibilities, such as rain gutters or roof flashing. In a moment we are going to describe a Transmatch that can be used to couple practically any wire or metal structure to the transmitter, including rain gutters.

If it is impossible to get an outdoor antenna up (and don't overlook "invisible" antennas[1]), an indoor antenna will work, but not as well as an outdoor one. With the Transmatch to be described, you can run a wire around the ceiling molding and couple the wire to the rig, maintaining that 50-ohm load we mentioned earlier. When you hear hams talking about loading up the bed spring, don't laugh — many a ham has made contacts that way. All you need is some method of coupling the bed spring to the rig; that's what our Transmatch will do.

The Transmatch

Shown in Fig. 4 is the circuit for a Transmatch that will do the job we've been talking about. This unit will match practically any wire to our 50-ohm rigs on any band from 80 through 10 meters. Depending on which capacitor plate spacing you use for C1 the Transmatch can be used for Novice powers or up to the legal limit of 1 kW. Actually, the closer-spaced capacitor will handle 200 watts or so. The same type coil stock is used for either power level, high or low.

Also included in the Transmatch shown in Fig. 4 and the photographs is an SWR bridge. This is a necessary item in order to know when the Transmatch is correctly adjusted. If you already have a Monimatch or reflectometer, the unit described can be eliminated from the Transmatch.

With the random-wire antenna we'll be dealing with various loads, depending on the frequency and band in use. Shown in Fig. 5 are the three basic circuit configurations used with the Transmatch. The circuit can be changed to conform with type A, B, or C by reversing the plug-in coil in the jack bar or by changing the capacitor C1 from the output side of the L configuration, as in B, to the input side, as in C. The capacitor is changed by means of a clip lead connected to the stator.

Some loads may require the circuit setup of A. In this configuration, a link-coupled parallel-tuned circuit is used, the antenna being tapped onto L1.

Construction Details

The chassis used to hold the parts is 2 × 7 × 9 inches (51 mm × 178 mm × 229 mm) and the panel measures 7 × 9 inches (178 mm × 229 mm). Any chassis large enough to hold the parts can be used. C1 and the jack bar for the plug-in coils are mounted on top the chassis, and the Varimatcher reflectometer and C2 are mounted below the chassis top. The remaining components are mounted on the chassis front.

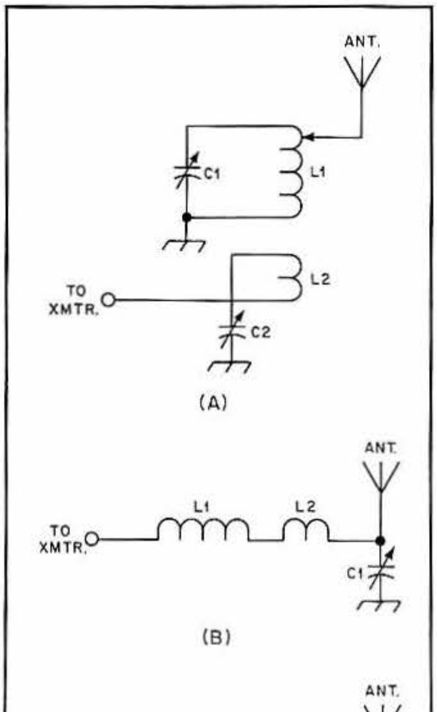

Fig. 5 — Three possible configurations with the Transmatch as described in the text. To set up types B or C the plug-in coil is reversed in the jack bar and terminal no. 3 of the jack is clipped to terminal no. 2.

A 4-inch (102 mm) length of RG-58/U coaxial cable is used for the pickup and conductor sections of the SWR bridge. Details for making this section are given in Fig. 6. Study these details and the photograph of the bottom view carefully before starting construction. The black outer covering of the coax is removed, leaving the braid and inner conductor with its insulation. The outer braid will be L3 and the inner conductor L4. You'll notice that there is an open area at the center of the coax braid. This is to allow room for the connection of one end of the 50-ohm resistor to the center of the inner part of the coaxial line. Use a probe or pointed instrument to open up the area in the braid and then carefully tin the edges of the opening with solder so as to prevent any of the hair-like wires from shorting to the resistor lead or inner conductor. The ends of the braid can be wrapped with a few turns of solid wire and tinned. The wire ends will provide a support for the connection to the feed-through terminal and the inner pin of J1.

The metal trough that houses the bridge section is made from a piece of copper flashing, although tin or aluminum could be used. It is a good idea to drill the holes for J1 and the feed-through bushing before bending up the trough into a U. Also, when soldering the leads for the diodes, D1 and D2, use a heat sink between the body of diode and the point being soldered, because too much heat can ruin the diode.

Using the Transmatch

In making the antenna, shoot for a length of at least 1/4 wavelength at the lowest operating frequency. In other words, if you plan on working 80 meters as the lowest band, then the antenna should be 60 feet (18.29 m) long, at least. This doesn't mean that shorter wires won't work — they will, but not nearly as well. A lot of hams will say they can't get up a length of 60 feet! The wire doesn't *have* to run straight. You can go around corners, up and down, or what have you. The important point is to get up as much wire as possible.

By the same token, many hams think (mistakenly) that they have to have a good earth ground connection in order for their station to work. It's fine if you have one but if you don't, forget it. If you have access to a water pipe or other earth ground connection make a connection to the rig and the Transmatch, but if you can't, don't worry. The antenna will still work.

With the antenna connected and the rig on the desired frequency, feed enough power through the system to get at least a half-scale reading on the Varimatcher meter, and have the Transmatch set up as in (A). Tap the antenna onto the first turn of the coil L1 at the "hot" end, or end opposite the ground. Incidentally, always turn the rig off when making any adjust-

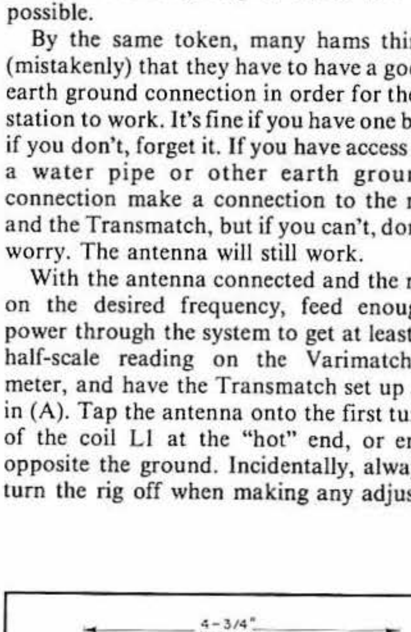

Fig. 6 — Constructional details of the Varimatcher bridge section. The braided portion of the coax is L3 and the inner portion is L4.

---

[1] Any wire that is no. 28 or smaller is practically impossible to see when strung up as an antenna. Many hams use "invisible" antennas to prevent a touchy landlord situation.

ments on the Transmatch. There may not be any dc voltages present but some very, very high rf voltages can be developed on the Transmatch coils and can cause a nasty rf burn.

With the antenna tapped on near the hot end of L1 switch S1 to read reflected power and adjust C1 and C2 for a dip or minimum reading. If you cannot get the reflected reading down to zero, or close to it, turn off the power and move the antenna tap closer to the ground end of the coil. You'll eventually find a tap point that will give a reading of zero in the reflected position versus full scale in the forward position of S1. If you find that the tap point on L1 is very close, say within 25 percent, to the ground end, you may want to try configuration B or C.

In eight of these configurations, start at the antenna end of the coil and, using another short clip lead, short out turns or portions of turns as you move the antenna tap towards the transmitter end of the circuit. In other words, the antenna is always on the end of the coil, but you may have some of the coil shorted out in order to get a match as indicated by M1.

Once you get a match for any given band, make a chart of the taps and so forth so you can return to the setup when you change bands.

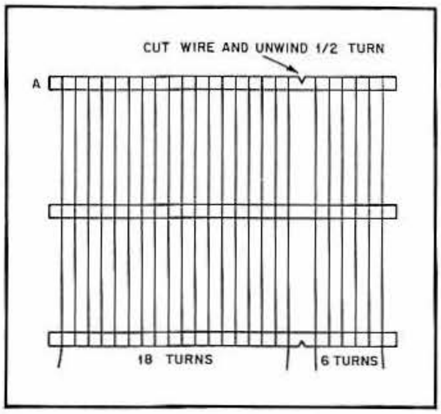

Fig. 7 — The 80/40-meter coil. To use the coil on 40 meters, short out 10 turns, counting from the "A" end of the coil. Shorting clips should be copper and the shorting leads should be insulated. Coil stock is no. 14 wire, 8 turns per inch (25 mm), 2-1/2 inches (64 mm) in diameter (Air Dux 2008T, B & W 3906-1, or Polycoils 1775).
The 20-meter coil consists of 4 turns of no. 12 solid wire, 2-1/2 inches in diameter, with the turns spaced so that L1 is 1-1/2 inches (38 mm) long. L2 is one turn of no. 12, same diameter as L1, spaced 1/2 inch (13 mm) from the end of L1.
The 15/10-meter L1 coil is 2 turns of no. 12 solid wire, 2-1/2 inches (64 mm) in diameter, turns spaced to cover 1-1/2 inches (38 mm). L2 is a single turn, same diameter as L1, spaced 1/2 inch (13 mm) from L1.
All coils are mounted on Millen type 40305 plugs and the socket (jack bar) is a Millen 41305. See G. R. Whitehouse ads in *QST*.

The main reason for using configurations B and C is that if the antenna is tapped near ground on L1 in configuration A, extremely high rf voltages can develop across C1 even with Novice powers. We don't recommend playing around with "arcs" from coils but it might be wise to prove to yourself just how much voltage can be developed. With the antenna tapped near the ground end of L1, take a pencil and hold it by the wood, not at the eraser end but in the center. With the key closed and the rig loaded up to about 75 watts, touch the lead end of the pencil to the stator of C1. You'll quickly see why caution is necessary. Such an arc can make a very nasty and painful burn.

If you want to try using antennas that are *very* short for the frequency, such as 10 feet (3.05 m) long on 80 meters (similar to a mobile whip) certain additional steps may be required to obtain a match with the Transmatch. When you make the coil for 80 and 40 meters you'll have some coil stock left over. With very short antennas, this can be used as a "loading" coil. Connect the antenna to one end of the coil and connect the other end to the antenna terminal on the Transmatch. Then go through the tune-up procedure. You may have to use a clip lead to short out turns on the coil but you'll find a setup that will permit matching with the Transmatch. Again, avoid coming in contact with the series loading coil because very high rf voltages can develop in such a setup.

Some beginners may not be aware of it, but a fluorescent lamp or neon bulb will light up when in an rf field, if the rf voltage is high enough. If you have a fluorescent lamp (a burned-out one is OK), the lamp can be laid along the antenna wire where it leaves the Transmatch. This makes a good output indicator, showing when power is flowing in the antenna. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

Off-Center-Loaded Dipole Antennas

So you can't manage a buried radial system on your city lot, eh? Then why not scrap those plans for a vertical antenna and try K1TD's version of a short dipole? Here are the design tips you need.

In these times when much of our amateur population lives in urban areas, the subject of shortened antennas for the lower frequency amateur bands is a very popular one. Physically short ground-mounted vertical antennas with lumped-constant loading to make them resonant can be quite efficient radiators, if a good radial system has been installed. This has certainly been evidenced in Sevick's series of *QST* articles.[1] To many amateurs, however, the "hitch" in constructing such a system is the installation of a good radial system. It must be admitted that for the "top" amateur bands, 160 and 80/75 meters, an efficient system of buried radials requires a sizable amount of real estate, even for a physically short radiator. On the average city-size lot, 50 or 75 by 120 to 150 feet (15.24 - 22.86 m by 36.58 - 45.72 m), it's almost impossible to install a highly efficient radial system for 80/75 meters, much less for 160 meters, when structures like a house and perhaps a separate garage exist. Or to some amateurs, just the thought of burying hundreds or maybe thousands of feet of wire is enough to turn off any enthusiasm for the project.

What's the alternative? A dipole type of antenna with lunped-constant loading. At modest heights, 30 or 40 feet (9.14 or 12.19 m), such an antenna will prove to be quite satisfactory if it is physically longer than about 0.2 wavelength. Shorter lengths may also be used, at reduced efficiency. Such an antenna can be fed directly with 50-ohm coaxial line, and it can be operated with no earth ground. (Of course the chassis of the transmitter and/or receiver should be grounded adequately for protection against shock hazard.)

Nearly all of us are familiar with the concept behind the use of inductive loading. A vertical antenna which is shorter than a quarter wave (or a dipole antenna which is

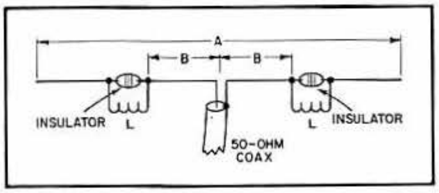

Fig. 8 — A dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance.

shorter than a half wave) will exhibit capacitive reactance at its base (or center) feed point. To cancel such capacitive reactance, a coil having the proper inductive reactance may be connected in series with the base feed point of the vertical. The same result will be obtained through the use of two such coils for a dipole, one coil connected in series with each half. It is not necessary for the inductor to be installed at the feed point, however. In fact greater radiating efficiency results through improved current distribution if the inductor is located along the radiator some distance away from the low-impedance feed point, *viz,* in the manner of a center-loaded mobile whip antenna. Fig. 8 shows this concept extended to a dipole element, with off-center loading. The inductors resonate the antenna to the operating frequency, but do little actual radiating themselves. (This is in contrast to helically wound or continuously loaded elements, where a long thin inductor is the radiator as well as the loading element.)

In the antenna represented by Fig. 8, there are many variable factors to be considered when a practical antenna for a given frequency is being constructed. Of primary consideration from an efficiency standpoint is the overall length, shown as dimension A. Another consideration for efficiency is the distance of the coils from center, dimension B. The longer the overall length (A), up to a half wave, and the farther the loading coils are placed from the center (B), the greater is the efficiency of the antenna. However, the greater is distance B (for a fixed overall antenna size), the larger the inductors must be to maintain resonance. Theoretically, if the coils were placed at the outer ends of the dipole, they should be infinite in value to maintain resonance. Capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce this requirement to a more practical value.

What Inductance Values?

As a matter of personal interest, this writer has been doing experimental work for a number of years with off-center-loaded antennas. One big drawback to such experimentation was the ever-present need for a large amount of cut-and-try work to arrive at resonance whenever a new set of dimensions was to be used. Probably the number of pruned-off turns from coil stock from such experiments, if straightened out and soldered end to end, would make up several full-sized half-wave antennas for the 160-meter band. Therefore, most of the writer's work of late in this area has been in going through paperwork exercises, looking for a way whereby at least "ball-park" values of inductance needed for a particular system could be calculated.

(Eq. 1):

$$L_{\mu H} = \frac{10^6}{68\pi^2 f^2}\left\{\frac{\left[\ln\frac{24\left(\frac{234}{f} - B\right)}{D} - 1\right]\left[\left(1 - \frac{fB}{234}\right)^2 - 1\right]}{\frac{234}{f} - B} - \frac{\left[\ln\frac{\left(24\,\frac{A}{2} - B\right)}{D} - 1\right]\left[\left(\frac{\frac{fA}{2} - fB}{234}\right)^2 - 1\right]}{\frac{A}{2} - B}\right\}$$

where

$L_{\mu H}$ = inductance required for resonance
ln = natural log
f = frequency, MHz
A = overall antenna length, feet
B = distance from center to each loading coil, feet
D = diameter of radiator, inches

The equation contained in the Mobile chapter of *The ARRL Antenna Book* for determining the capacitance of a vertical antenna shorter than a quarter wavelength looked promising in early computations, and, indeed, it became the basis for the calculation procedure which finally resulted. This procedure has been found to produce results much closer than mere "ball-park" values for the necessary inductance — for wire antennas "in the clear" at moderate heights, the final inductance values found by cut-and-try pruning for lowest SWR at the desired frequency have been so close to the value from calculations that a laboratory bridge was necessary to measure the difference. The results are equally good for elements using tubing. Once the needed inductance value is determined by calculations, it is generally found sufficient to obtain coil dimensions from an ARRL L/C/F Calculator (see LaPlaca[2]) or by equation. Any significant pruning which has been found necessary could always be attributed to objects in proximity to the ends of the antenna.

The complete set of calculations is expressed in the mathematical relationship as Eq. 1, presented here primarily for mathematics buffs or those having access to electronic computers. This equation yields the inductance required, in microhenrys, for single-band resonance of a shortened antenna of a particular physical size at a given frequency, for a specific position of the loading coils from the center of the antenna. To spare the reader the task of performing some rather tedious calculations, Fig. 9 has been prepared from Eq. 1. The curves of the chart have been normalized, and may be used for any frequency of resonance. The chart is based on a half-wavelength/diameter ratio of the radiator of approximately 24,000. (This corresponds to no. 14 wire on 80 meters or no. 8 wire on 160 meters.) For "thinner" conductors, the required inductance will be somewhat greater than that determined from Fig. 9, and less inductance will be required for "thicker" conductors.

The use of the chart is as follows: At the intersection of the appropriate curve from the body for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimensions A and B are shown in Fig. 8, and for use with the chart are expressed as percentages. Dimension A is taken as percent length of the shortened antenna with respect to the length of a resonant half-wave dipole of the same conductor material. Dimension B is taken as the percent of coil distance from the feed point to the end of the shortened antenna. For example, resonating an antenna which is 50 percent or half the size of a half-wave dipole (one-quarter wavelength overall), with loading coils positioned midway between the feed point and each end (50 percent out), would require loading coils having an inductive reactance of approximately 950 ohms at the operating frequency. If the antenna is hung "in the clear," and if the length/diameter ratio of the conductor is near 24,000, inductance values as determined from the chart will be very close to actual values required. (Eq. 1 above takes the diameter of the radiator into account, and thus may be used for any length/diameter ratio.) For practical

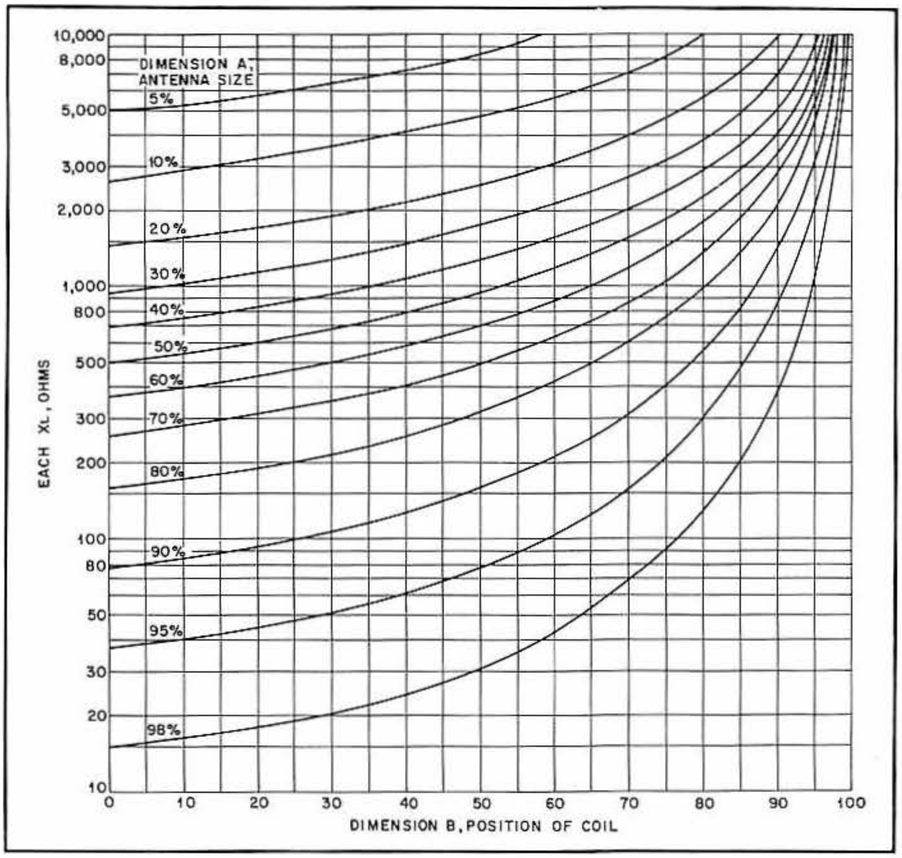

Fig. 9 — A chart for determining approximate inductance values for off-center-loaded dipoles. At the intersection of the appropriate curve from the body of the chart for dimension A and the proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at left. See Fig. 8 regarding dimensions A and B.

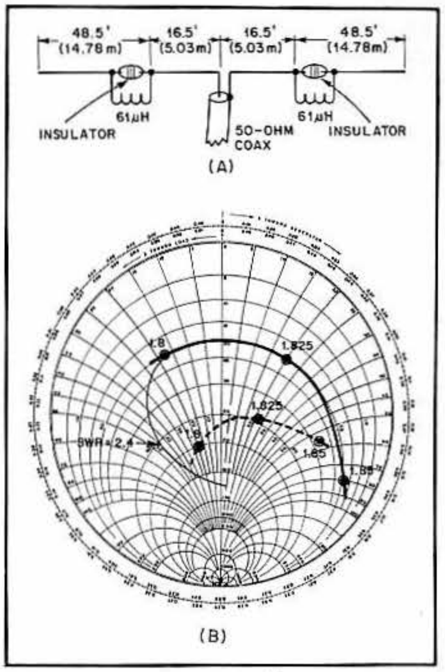

Fig. 10 — At A, an 80-meter dipole loaded for 160-meter operation. The inductors are 37 turns of coil stock having a 3-inch (76 mm) dia and 10 turns per inch (25 mm) (B & W 3035). At B, the impedance plot of this antenna installed at a height of 50 feet (15.24 m) (solid curve), and the plot of a 160-meter half-wave dipole (broken curve).

purposes, dimension B may be taken as that distance from the center of the feed-point insulator to the inside eye of the loading-coil insulator, and dimension A as the eye-to-eye distance inside the end insulators (which are not drawn in Fig. 8).

Proximity of surrounding objects in individual installations may require some pruning of the coils, and the exact amount of final inductance required should be determined experimentally. If the antenna is hung in inverted-V style, with the ends brought near the earth, the required inductance will almost always be somewhat less than that determined from the chart or equation. A grid-dip meter, Macromatcher (see Hall and Kaufmann[3]), or SWR indicator may be used during the final adjustment procedure.

Practical Antennas

Although one might erect an inductively loaded antenna that is cut for a single amateur band, it is possible to use the antenna itself for two, three or more bands of operation, if provision is made to lower the antenna for band changes. A simple rope halyard and pulley arrangement at one of the supports will do the trick. Fig. 10A shows a three-band antenna of this nature, for 160, 80 and 20 meters. If the insulators shown are left open, with nothing bridging them, the antenna is a simple half-wave dipole cut for 14.18 MHz. (The 48.5-foot [14.78 m] lengths act merely as support wires, and have negligible effect on operation of the antenna.) If the insulators are bridged with short lengths of antenna wire, the antenna becomes a center-fed 80-meter dipole, resonant at about 3.6 MHz. For 160-meter operation the 20-meter insulators may be bridged with loading coils to resonate the antenna at 1.8 MHz, as shown in Fig. 10A. Burndy or other manufacturers' "Servit" type of electrical connectors may be used for ease in making band changes quickly, as shown in Fig. 11.

The calculation procedure for determining loading-coil values for the antenna of Fig. 10A, using the chart of Fig. 9, goes like this. If operation is desired on 1.8 MHz, the length of a full-sized half-wave dipole is found from the relationship 468/f to be 260 feet (79.25 m). The 130-foot (39.62 m) length of Fig. 10A represents 50 percent of this size, meaning that the dimension-A curve marked "50 percent" in Fig. 9 is to be used. The position of the coils is $16.5/(16.5 + 48.5) \times 100$, or 25 percent of the distance out from center, dimension B. From the intersection of 25 (horizontal scale at bottom) and the 50 percent curve, the required inductive reactance is read from the scale at the left of Fig. 9 to be 650 ohms. The inductance, L, is $650/2\pi f$ or 57.5 microhenrys, if no. 8 wire is to be used. For smaller diameter wire, the inductance should be somewhat larger. (Calculations from Eq. 1 for no. 12 wire indicate the required inductance is 60.99 µH.)

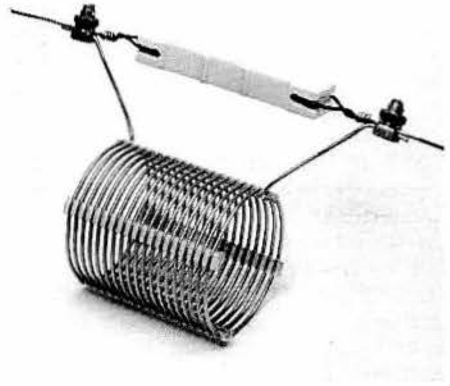

Fig. 11 — Copper electrical service connectors, sold under one trade name of Servit, provide a simple means of installing the loading coils. The antenna wire and the ends of the coil wires should be tinned to prevent corrosion. In addition, a protective coating of acrylic spray may be used at each connection.

The radiation resistance of a shortened antenna loaded to resonance is less than that of a full-sized antenna. Further, the shortened antenna is "sharper," meaning that the change in reactance versus frequency is greater. In other words, the shortened antenna acts as a tuned circuit having a higher Q than a full-sized antenna.

To check these characteristics, the line input impedances for the antenna of Fig. 10A were measured with a laboratory bridge, and the electrical line length at the measurement frequency was then taken into account to determine the impedance at the antenna feed point. The antenna was constructed of no. 12 wire and hung at a height of 50 feet (15.24 m) as a "flat-top" radiator.

The solid curve of Fig. 10B is a plot of the feed-point impedance versus frequency for this antenna. The plot on Smith Chart coordinates is more meaningful than a simple SWR-vs.-frequency curve because the magnitudes of the resistive and reactive components are shown, as well as the sign of the reactance. (Capacitive reactance is negative, plotted to the left of the vertical center line, and inductive reactance is positive, plotted to the right.) In this presentation, a 50-ohm nonreactive impedance will appear at the exact center of the chart. The SWR in 50-ohm line for a given frequency may be determined by first noting the distance from the center of the chart to the particular impedance plot on the curve, and next measuring this same distance down the vertical center line from chart center (a drawing compass is helpful for this task), and finally dividing 50 into the value read at that point on the center line. For example, the SWR at 1.8 MHz equals 120/50 or 2.4, as indicated by the segment of the 2.4 SWR circle in Fig. 10 B. It may be seen that resonance (zero reactance) occurs at approximately 1810 kHz, where the resistance is about 22 ohms. The SWR at resonance is 2.33:1, and climbs to 3:1 at 1825 kHz. At 1850 kHz, the SWR is 10:1. Without any matching provisions the antenna is relatively sharp, as mentioned earlier. If one sets the usable bandwidth as the frequency range where the SWR is 3:1 or less, it is approximately 35 kHz, or 1.9 percent of the resonant frequency. As far as efficiency is concerned, ohmic losses are low, and the antenna is a good performer on 160 meters. Because of its horizontal polarization, it has proved to be most effective at night, and stations several hundred miles away have been worked with S 9 reports received for the 50-watt signal.

For a comparison of impedances, the broken curve of Fig. 10B is a plot of measured impedances of a full-size half-wave dipole, 260 feet (79.25 m) long overall, hung in place of the shortened antenna. From this curve it may be seen that resonance occurs at 1810 kHz, where the resistance is 59 ohms. The 3:1-SWR bandwidth for the half-wave antenna is in the order of 60 kHz, or 3.3 percent of the resonant frequency. It is interesting to note on this curve that the SWR at resonance is 1.18:1, and that it is a somewhat lower value, 1.15:1, at a frequency a few kilohertz above resonance. (Measurements were made every 5 kHz across this band, but plot points are shown only for 25-kHz increments to avoid crowding of the data.) This evidence refutes the oft-heard statement that the SWR-vs.-frequency curve is *always* lowest at antenna resonance. Points to remember are that the SWR in a transmission line is completely dependent upon the characteristic-impedance value of the line in use. Using a line of different impedance may shift the position of the SWR curve along the frequency axis in a simple SWR-vs.-frequency plot. This is definitely true in this case — if the 160-meter half-wave dipole were to be fed with 75-ohm line, the lowest SWR would occur at a frequency about 5 kHz *below* antenna resonance, whereas with 50-ohm line the lowest SWR is at a frequency slightly *above* resonance. The reason this happens is that the resistive component of the impedance, which consists of the radiation resistance plus any loss resistance, is not constant with frequency, even over a rather narrow frequency range. It must be acknowledged that the differences here are very slight, however, and for practical purposes the frequency of lowest SWR is (within a few kilohertz) the resonant frequency of the antenna.

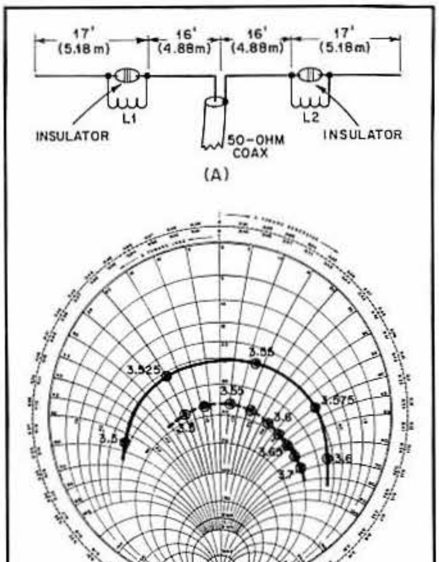

Fig. 12 — At A, a 40-meter dipole loaded for 80-meter operation. For resonance at 3.55 MHz the coils should be approximately 40 μH (27 turns of stock); 3.75 MHz, 35 μH (24 turns); 3.9 MHz, 31 μH (22 turns); and 4.0 MHz, 29 μH (21 turns). These are calculated inductance values for no. 12 antenna wire. Coil stock referenced above is 3-inch (76 mm) diameter, 10 turns per inch (25 mm) (B & W 3035). At B, the impedance plot of the 3.55-MHz version (solid curve) and of an 80-meter half-wave dipole (broken curve).

Another point concerning the SWR values bears noting. The values as determined from the plots in the manner described above are quite accurate, having been determined by measurements with laboratory equipment. In contrast, measurements with simple SWR indicators usually cannot be relied upon for anywhere near the equivalent accuracy.

For example, the author owns a commercially manufactured SWR indicator of the Monimatch type (see McCoy[4]) which, under a particular set of conditions, indicates a 2.5:1 SWR in a line where laboratory measuring equipment shows the *true SWR to be 4:1.* A significant difference! Herein lies another reason why impedance plots on Smith Chart coordinates are more meaningful than a simple SWR-vs.-frequency curve — greater accuracy may generally be expected.

A Half-Size 80-Meter Antenna

Fig. 12A shows the 3-band concept described earlier as it can be applied to 80, 40 and 20 meters. Its overall length is 66 feet (20.12 m), not a difficult length to use on a small lot. This antenna was constructed for 80-meter operation with a design-center frequency of 3.55 MHz, using no. 12 antenna wire and 40-μH loading coils — 27 turns of stock having a diameter of 3 inches (76 mm) and a pitch of 10 turns per inch (25 mm) (tpi). Feed-point impedances versus 80-meter frequency for the antenna, hung at a height of 50 feet (15.24 m), are shown by the solid curve at B of Fig. 12. Actual resonance occurred at 3.54 MHz, where the resistance was about 26 ohms. The bandwidth within which the SWR is 3:1 is 60 kHz, or 1.69 percent of the resonant frequency.

Also shown in Fig. 12B, by the broken curve, are the feed-point impedances of a half-wave dipole, 132 feet (40.23 m) overall length, hung in place of the shortened antenna. Resonance occurs at 3.54 MHz, where the resistance is 43.5 ohms and the SWR is 1.15:1. The broader nature of the half-wave antenna is exhibited by the "tighter" curve which swings closer to the 50-ohm center point of the chart than the shorter, loaded antenna. The SWR at 3.5 MHz is 1.6:1, and remains below 3:1 to 3.67 MHz.

Capacitive and Inductive Loading

One would assume that a combination of capacitive and inductive loading might provide a different feed-point impedance than would inductive loading alone, because of different current distributions in the radiators. To check out this assumption, the antenna of Fig. 12A was used as a "test bed" for comparative measurements. Capacitance hats were attached at different points along the 17-foot (5.18 m) lengths of wire outside the coils, and the coils were pruned to reresonate the antenna at about the same frequency as before. The impedance measuremnts were then repeated.

Dangling End Sections

First, "hats" consisting of 18 inches (457 mm) of no. 12 wire were affixed to the antenna ends and permitted to dangle. This lowered the resonant frequency to 3435 kHz. By calculations, this was approximately the same effect as that of extending the 17-foot portions of the antenna by the same amount as the dangling lengths, so it

Table 4

Characteristics of Various Loading Techniques, 66-Foot, 80-Meter Dipole

Loading	*Approx. feed-point resistance, resonance*	*SWR at resonance*	*3:1-SWR bandwidth, % of resonant freq.*
40-μH coils only	26 ohms	1.92:1	1.69
36.5-μH coils, 18" dangling ends	26	1.90:1	1.79
36" hats outside 32.5-μH coils	23	2.15:1	1.68
30-μH coils, 36" hats at ends	25	1.98:1	2.05
None (λ/2 dipole)	43.5	1.15	Greater than 3.6

Coil positions for each loaded antenna were 16 feet from antenna center. All antennas were constructed of no. 12 wire and installed at a height of 50 feet.

Feet X 0.3048 = m.

Inches X 25.4 = mm.

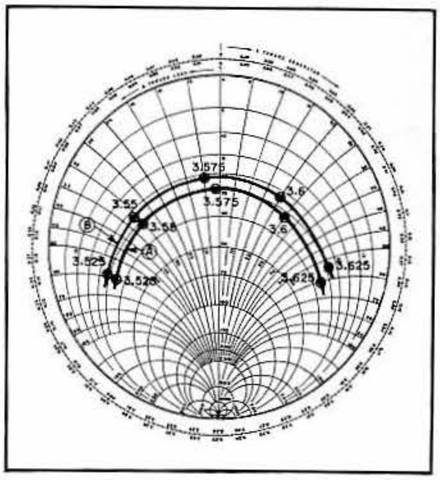

Fig. 13 — Curve A is the impedance plot of the antenna of Fig. 12A with 18-inch (457 mm) dangling end sections added and the coil trimmed to restore resonance near the original frequency. Curve B is a plot of the same antenna with X-shaped capacitance hats added at a point just outside the loading coils (dangling sections removed and coils trimmed to reestablish resonance.).

would seem to make little difference whether short sections of extra length are added inside the supporting insulators or are at the ends, suspended at right angles to the main antenna wire.

The inductors were reduced from 40 to 36.5 μH (25-turn coils replaced the original 27-turn coils), and resonance occurred at about 3575 kHz. At this frequency the resistance was 26 ohms and the SWR 1.90:1. The 3:1-SWR bandwidth, 64 kHz, is 1.79 percent of the frequency of resonance. The impedance plot for this arrangement is shown as Curve A in Fig. 13. The resistance at resonance for this antenna is identical to that with the coils alone, and the bandwidth is only 4 kHz greater, 64 kHz vs. 60. From these results, one would conclude that the main advantage offered by the "danglers" is a small saving of space over a flat-top antenna.

Capacitance Hats near Loading Coils

Next the dangling end sections were removed and a pair of capacitance hats was formed, each from two 36-inch (0.91 m) lengths of no. 12 solid wire. The two wires for a single hat were attached at their centers to the antenna wire at a point just outside one of the loading coils. The hat wires were then bent radially to form an X at right angles to the antenna wire, like four spokes of a wheel with the main antenna wire at the hub. The diameter of the X-shaped hat was thus 36 inches. The second hat was placed in a like manner just outside the second coil. Burndy connectors were used to affix the hat wires. The resonant frequency of this configuration with the original 40-μH loading coils was found to be 3290 kHz. The effect of adding the hats was about the same as that of extending the 17-foot (5.18 m) lengths to 19 feet (5.79 m).

When the inductors were replaced with 23-turn coils (32.5 μH), the antenna resonated at about 3.575 MHz, the resistance being 23 ohms. The SWR at resonance is 2.15:1, and the 3:1-SWR bandwidth for this configuration is 60 kHz, 1.68 percent of the resonant frequency. The impedance of this arrangement versus frequency is shown by Curve B of Fig. 13.

It is surprising to note that, by the standards of most amateurs, the characteristics of this antenna are not as good as those of the same length antenna with loading coils alone. The SWR at resonance for the antenna with combination capacitive and inductive loading is higher (2.15 vs. 1.92), and the 3:1-SWR bandwidths are the same, 60 kHz. Perhaps a significant factor here, though, is that the diameter of the capacitance hats used for these measurements was small, only 0.011 wavelength. Supporting much larger hats presents mechanical problems with wire antennas, however, as even these were a bit flimsy and would require reshaping after gusty weather.

Capacitance Hats at Antenna Ends

Finally, the X-shaped capacitance hats were moved to the outside ends of the antenna, just inside the end insulators. With the original 40-μH coils, resonance appeared at 3215 kHz. From calculations, it was as if the 17-foot (5.18 m) end sections were actually 21 feet (6.40 m) long. With 30-μH coils (22 turns) in place, the resonant frequency was 3560 kHz. At this frequency the resistance was 25 ohms and the SWR 1.98:1. The 3:1-SWR bandwidth is 73 kHz, or 2.05 percent of the resonant frequency. The impedance plot of this antenna is given in Fig. 14.

It is interesting to note that the position and shape of the plot for this antenna on Smith Chart coordinates is nearly identical to that for the same length antenna with loading coils only, the solid curve of Fig. 12B. For this antenna, however, the plot points for 25-kHz frequency increments appear closer together, which accounts for the increased bandwidth.

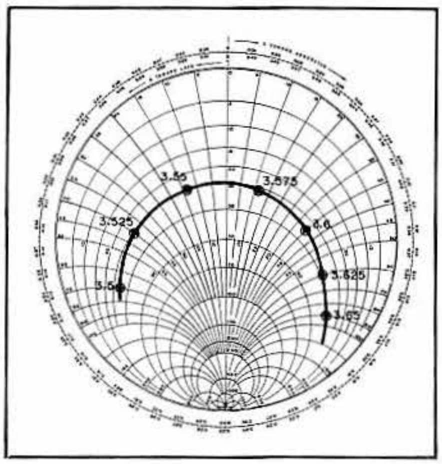

Fig. 14 — Impedance plot of a 66-foot (20.12 m) dipole using a combination of off-center inductive loading and capacitive end loading. Of all the shortened configurations tried, this arrangement offered the greatest bandwidth.

Conclusions

The measured characteristics of these various configurations of loading for the 80-meter antennas are tabulated in Table 4. Remember that the overall "flat-top" length of each antenna arrangement is 66 feet (20.12 m), and that the loading coils are always positioned 16 feet (4.88 m) each side of the center of the antenna, being pruned for resonance at approximately 3550 kHz. For comparison, information for a half-wave dipole is also included.

Of the various arrangements, capacitive end loading decidedly provides the greatest bandwidth, excepting the full-size half-wave antenna, of course. Although there are slight differences in the resistance value at resonance, all are of the same order of magnitude. These values, as well as those for the 160-meter antenna discussed earlier, tend to confirm a *broad* rule of thumb that the writer has formulated for this type of antenna: The feed-point impedance value at resonance is roughly proportional to the length of the antenna. That is, a loaded antenna which is half the size of a half-wave dipole will have approximately half the radiation resistance of the full-sized antenna.

Eq. 1 given earlier or the chart of Fig. 9 allows one to calculate loading-coil values for antennas with loading coils only. Additional capacitive loading is not taken into account. Calculating the effects of various capacitive loading arrangements appears to be difficult, and work remains to be done in this area.

Multiband Antennas with Loading Coils

All of the foregoing material has been devoted to the loading of an antenna for resonance at a single frequency. Resonated as described, the antenna is electrically a half wave in length. It will, however, operate well on higher frequencies — frequencies at which it is an odd multiple of half waves in *electrical* length . . . three half waves, five half waves, etc. Because of the lumped loading of the shortened antenna, these higher frequencies will likely not be closely related to odd-order harmonics of the fundamental frequency, as the case would be for a nonloaded radiator. (For example, it is a well-known fact that a 7-MHz half-wave dipole operates well on its third harmonic, 21 MHz.)

A loaded dipole will become an electrical 3/2-λ antenna at some frequency below that which is three times the fundamental resonant frequency. Depending upon the overall antenna length, coil value, and coil position, it is possible for an 80-meter loaded dipole to become a 3/2-λ performer on 40 meters. With such an arrangement, one would have a dual-band antenna without requiring the use of traps. The idea can be expanded upon to arrive at a loaded antenna without traps which will operate on more than two bands. This scheme offers considerable constructional simplification as compared with trap arrangements.

The multiband loading-coil concept has been recognized for better than half a century, but little use of the technique has been made by amateurs. Some years ago a very good article on the subject was published by William Lattin, W4JRW.[5] That article is recommended reading for anyone interested in more details on the concept. Supplemental information has been published by Buchanan.[6] Attempts by this writer to calculate antenna sizes and coil values for dual-band antennas have met with some success. From calculations and experiments to date, it appears that with only two loading coils (one each side of center), the antenna must always be greater than a half wave in physical length for the higher of the two frequency bands. In other words any 80/40 meter arrangement, for example, apparently would need to be longer than 66 feet (20.12 m) from tip to tip. However, much work also remains to be done in this area. This material was originally presented in *QST* by Jerry Hall, K1TD.

References

[1] Sevick, "The Ground-Image Vertical Antenna," *QST,* July 1971. (See chapter 1.) Also, "The W2FMI Ground-Mounted Short Vertical," *QST,* March 1973. (See chapter 1.) "A High Performance 20-, 40- and 80-Meter Vertical System," *QST,* December 1973. (See chapter 1.) "The Constant-Impedance Trap Vertical," *QST,* March 1974, p. 29.

[2] LaPlaca, "Using the ARRL L/C/F Calculator," *QST,* December 1973, p. 26.

[3] Hall and Kaufmann, "The Macromatcher, An Rf Impedance Bridge for Coax Lines," *QST,* January 1972, p. 14. Information also contained in *The Radio Amateur's Handbook* and *The ARRL Antenna Book.*

[4] McCoy, "Monimatch, Mark II," *QST,* February 1957, p. 38. Also, "An Etched-Circuit Monimatch For Checking Your Antenna System," *QST,* October 1969. (See chapter 4.)

[5] Lattin, "Multiband Antennas Using Loading Coils," *QST,* April 1961, p. 43.

[6] Buchanan, "An Inexpensive 40- and 80-Meter Antenna, Hints and Kinks," *QST,* September 1962, p. 62.

Efficient Short Radiators

If you're cramped for space to put up a beam antenna, you'll want to investigate these ideas as one way to shrink 'er down to usable size.

Shortened antennas have been popular with amateurs for either esthetic, space limitation or economic reasons. And most are inefficient radiators with narrow bandwidth. This section will first define some limitations and problems of conventional short antenna designs. A design for a half-size 2-element Yagi, the performance of which is competitive with its full-size equivalent, will follow.

It has been proved that no appreciable difference exists between the directive patterns of "short" and full half-wave horizontal dipoles. Since the gain of a half-wave dipole over an isotropic radiator is determined by the dipole directive pattern and since the length of the dipole (up to λ/2) has negligible effect on its directive pattern, the length of the dipole (up to λ/2) has a negligible effect on its gain. An assumption is made that the efficiencies of both the short and half-wave dipoles are equal. The efficiency of an antenna can be expressed as

$$\text{Eff} = \frac{R_r}{R_r + R_{loss}} \qquad \text{Eq. 1}$$

where R_r = radiation resistance
R_{loss} = loss resistance of conductors, coils, etc.

The radiation resistance (R_r) of a horizontal dipole is determined by its length and height above ground. In free space, the R_r of a half-wave dipole is 73 ohms. As the length of the dipole is reduced, the R_r decreases as the square of the length. Thus, for a half-size dipole (as compared with λ/2), the R_r decreases to a value one-fourth that of the λ/2 dipole or approximately 13 ohms.

The efficiency of a half-wave dipole is very high, being on the order of 95 percent. The R_r is large and R_{loss} small by comparison. If we reduce the overall length by a factor of two, the R_r decreases by a factor of four. Input impedance at the center of the dipole is comprised of a low resistance and large capacitive reactance, so inductive loading is required in order to resonate the dipole. The inductive reactance needed to center-load the antenna will be approximately 1 kΩ (depending on conductor size used for the antenna). Assuming a coil Q of 200 and computing efficiency using Eq 1: This means that 28 percent of the transmitting power applied to the antenna is dissipated in the loading coil. The author has neglected any additional losses in the matching network which would result if the above antenna were driven by a 50-ohm source.

Let's suppose we decide to build a 2-element Yagi using shortened center-loaded elements. The driven element by itself has an R_r of 13 ohms. Adding a close-

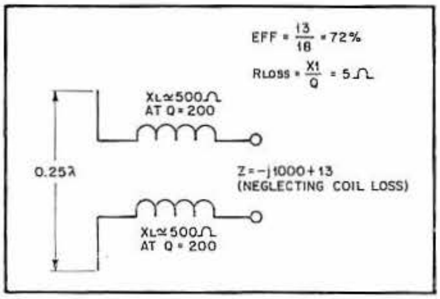

Fig. 15 — Electrical equivalent for the center-loaded dipole discussed in the text. Efficiency is 72 percent, as determined by Eq. 1.

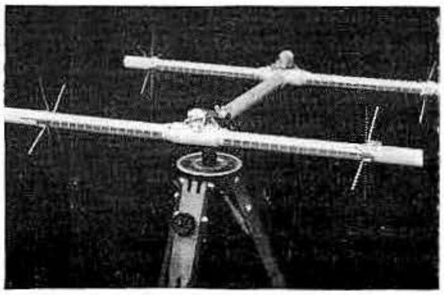

The 100-MHz "model" used by the author for testing shortened elements in a Yagi antenna design.

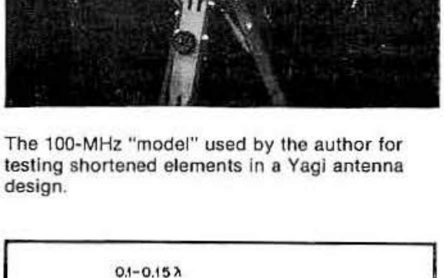

Fig. 16 — Electrical equivalent of a loaded 2-element antenna. The efficiency is 47 percent.

spaced parasitic element reduces the radiation resistance by approximately a factor of three. Thus, the driven element has an impedance composed of approximately 4.5 ohms resistive and 1 kΩ reactive. Loading both the driven and parasitic elements with the same inductors that were used for the dipole and computing efficiency

$$\text{Eff} = \frac{4.5}{4.5 + 5} = 47 \text{ percent}$$

This efficiency equates to slightly more than a 3-dB power loss in the driven element. The writer has neglected the sizable loss in the matching network since it must transform the 50-ohm source to the 9.5-ohm impedance of the antenna at resonance. In addition, the loading coil of the parasitic element will contribute approximately 1 dB of loss.

Taking into account the losses on loading coils of both the driven and parasitic elements and the loss on the driven element to feed-line matching network, we can easily account for a 5-dB power loss. Such a power loss would nullify the gain of a properly tuned Yagi, making it little better than a full-size dipole.

It should be obvious from the previous examples that the poor performance of short dipoles, and particularly Yagis, is a direct result of losses in loading inductors. If we can reduce these losses while raising the radiation resistance, the increase in efficiency will (might?) make the short dipole or Yagi competitive with its full-size counterpart.

Test Antenna

The helically wound whip antenna has always been appealing because the R_r tends to be higher than that of the base-loaded whip. This is because it is continuously loaded. In fact, using end-loading (large capacitance hats), the current distribution will be much more constant, yielding an R_r of up to four times that of a base-loaded whip. The author decided to use the helical design for a short test dipole and investigate the properties of this antenna versus that of a full-size dipole.

Conventional design for a helically wound antenna uses a form (wood, bamboo, fiberglass) wound with approximately $\lambda/2$ of no. 14 wire per $\lambda/4$ element. The writer felt that the resistive R_{loss} of this large amount of wire would be as great as that of the loading coils in the center-loaded dipole, making the helical design ineffective. As it is desirable to reduce conductor resistance, the surface area of the conductor must be substantially increased. Tubing is not mechanically suitable, and because skin depth is only on the order of 0.001 inch (0.03 mm),[1] the use of tubing is not required. Some 1/2-inch (13 mm) wide copper tape[2] was on hand and a study of its skin resistance revealed a loss per unit length of 12 percent less than that of the no. 14 wire.

A helically wound dipole was constructed (using the 1/2-inch [13 mm] wide tape) for use at 100 MHz. This frequency was chosen so that field-strength measurements could be performed using a local fm broadcast station as the signal source. A 28-inch (711 mm) length of 3/4-inch (19 mm) CPVC tubing (plastic water pipe) was wound with 38 equally spaced turns of the 1/2-inch (13 mm) wide tape. Using a grid-dip meter, resonance was measured at 104 MHz. Next, two 5-1/2-inch (140 mm) diameter six-spoke capacitance hats were attached to the ends. Resonance was again measured and found to be 84 MHz. A C-match[3] was constructed to transform the low impedance of the short dipole to 50 ohms. The C-match requires that the element be made to look inductive by lowering its resonant frequency. A variable capacitor is then shunted across the feed point and adjusted for resonance. This L/C ratio determines the impedance transformation. A 100-pF variable capacitor was shunted across the feed point of the test dipole and a 50-ohm coaxial cable was attached. Since the dipole is a balanced antenna, feed-line decoupling was necessary. The author used a quarter-wave sleeve (bazooka).

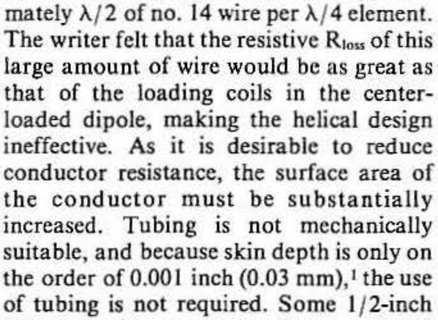

Fig. 17 — Curves depicting gain and FBR (front-to-back ratio) vs. frequency when the Yagi antenna is tuned for operation at 100 MHz.

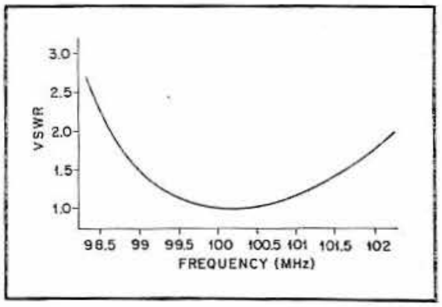

Fig. 18 — VSWR vs. frequency when the Yagi antenna is tuned for operation at 100 MHz.

Antenna Tuning

Rf power was applied to the test antenna through an SWR meter (measurements were taken in a screen room). Turns were removed, one at a time, and the shunt capacitor adjusted until a 1:1 match occurred. The capacitance hats were always installed at the tips of the element. When adjusted for a 1:1 match, the dipole consisted of 32 turns of copper tape, center fed, with 40 pF shunting the feed point and an overall element length of 21 inches (533 mm) which is 37 percent of the length of a full-size dipole. Tuning was very simple and took but 15 minutes to complete.

Measurements

The feed-point impedance was measured (with C-match disconnected) to be 20 ohms. The skin depth resistance was computed to be 0.24 ohm. Computing efficiency

$$\text{Eff} = \frac{20}{20 + 0.24} = 98 \text{ percent}$$

which is essentially the same as a full-size dipole. Bandwidth between 2:1 SWR points was measured to be greater than 6 percent of the operating frequency. Large capacity hats (for 100 MHz) and wide tape play an equally important role in reducing the Q of the antenna sufficiently to obtain a wide bandwidth. Next, a low-level signal generator and dipole were set up 5 wavelengths away from the test antenna. Using an Empire NF-105 noise and field-strength meter, a comparison was made on field strength between the test antenna and a full-size dipole. Results indicated no apparent measurable difference between the short helically wound dipole and a full-size dipole. Field-strength measurements were then taken on the signal from a local fm broadcast station. Again, there was no apparent difference; not bad for an antenna that is less than half size. The next question to be answered was whether gain could be secured by the addition of a parasitic element.

A boom and a parasitic element were added to the shortened test dipole. Although a director might provide better gain and front-to-back ratio (FBR), the decrease in radiation resistance, to a very low value (dipole by itself equals 20 ohms) because of the close spacing, was thought to decrease efficiency and bandwidth. A parasitic reflector was constructed, similar to that of the driven element. The element was split in the center and a 100-pF capacitor installed for tuning purposes. Element spacing was adjusted to 0.15 λ. With the capacitor fully meshed, the six-spoke capacitance hats were physically positioned on the element (both equidistant from element center) so that the reflector resonated at 93 MHz. Nominal reflector tuning calls for reflector resonance approximately 5-percent lower in frequency than the driven element, or 95 MHz. With the above configuration, the author was able to adjust reflector resonance, using the variable capacitor, anywhere in the range of 93 to 96.5 MHz or 3.5 to 7 percent lower in frequency than the driven element.

Using the same field-strength measurement set up as with the dipole, the reflector was aimed at the source signal and adjusted (using the variable capacitor at element center) for minimum pickup. The driven element was then readjusted for a 1:1 SWR and the process repeated until minimum rear pickup and a 1:1 SWR occurred simultaneously. Gain and FBR measurements were made at the design frequency and at points two percent of the design frequency on either side. The results are described in Fig. 17. The bandwidth measurements are given in Fig. 18.

As can be seen from the curves in Fig. 17, a gain of 4 dBd, and a 15-dB FBR can be secured quite easily with a physically short Yagi. As with other Yagis, the gain is relatively constant over a wide frequency range. However, the FBR drops quite fast especially when trying to use the antenna below its design frequency. From curves given in Fig. 18, the bandwidth between the 2:1 VSWR points is shown to be about 3.7 percent. Translated to 20 meters this bandwidth would be 500 kHz, equal to or greater than the average 3-band trap Yagi. If indeed the design were scaled to 14 MHz, the antenna would have element lengths of slightly over 12 feet (3.66 m) and a boom length of 10 feet, 6 inches (3.20 m).

Adapting Design to Ham Bands

Since it has been shown that a physically short Yagi can provide significant gain and FBR, a few design pointers are in order. First, make the element length no shorter than physically necessary for your particular situation. Bandwidth and efficiency will improve with greater lengths. Second, the use of large capacitance hats is recommended, as this reduces the helical conductor length and thus R_{loss}. Third, use large forms of good dielectric quality (fiberglass). The larger the form diameter, the greater the length reduction for a given number of turns. Also, with large diameter forms, 1.5 inch (38 mm), the width of the conductor can be increased, thus reducing R_{loss}. Fourth, regardless of the matching network used, construction of capacitance hats, etc., *solder all* joints and *seal* with silicone rubber. *Do not rely on pressure joints.* Fifth, to secure a reasonable FBR with any Yagi using a split driven element and fed with coaxial cable, some form of balun transformer is required to keep currents from flowing on the outside of the coax braid. Sixth, the use of a grid-dip meter is the only practical method to secure resonance with a fixed physical length. Construction will require some cut and try. Proximity to surrounding objects will affect resonance, and should be avoided when tuning the array.

No doubt the author has left out items, which may prove to be a problem for the reader. This article was intended to be more thought provoking than constructional, although there should be enough information here to build your own antenna with a reasonable amount of experimental effort. My current plans call for construction of a 2-element 10-meter Yagi (for reception of OSCAR signals) with 6-foot (1.83 m) elements and a 5-foot (1.52 m) boom. The author wishes to thank Mike Povlich, WB9HGS and Russ Mills for their assistance with measurements. This material was originally presented in *QST* by Ronald J. Gorski, N9AU.

References

[1] Kuecken, *Antennas and Transmission Lines*, Howard W. Sams and Company, Indianapolis, IN, 1st Edition (1969), p. 274.

[2] Copper tape available from Minnesota Mining and Mfg. Check Yellow Pages for the address of a local distributor.

[3] Orr, *Beam Antenna Handbook*, Radio Publications Inc., Wilton, CT, 3rd Edition, p. 63.

Simple Arrays of Vertical Antenna Elements

Work DX with a vengeance! Simple arrays of vertical elements will provide good low-angle radiation. DXCC may be closer than you think if you can manage a system like this one!

Because of the excellent low-angle radiation properties of vertical antenna elements and because of the benefits to be gained by horizontal-plane directivity, it is interesting to attempt to determine "optimum" arrangments of elements to produce "beam" antenna arrays. Because of the enormous variety of configurations possible, and the great difficulty of either experimental modeling or calculating exact theoretical performance, it has not been practical in the past to arrive at valid conclusions. However, it is now quite easy to simulate such antenna arrays by computer programs, and these can produce relatively quick and precise answers to a very large number of postulated configurations. After a number of computer trial runs, "optimum" configurations can be found quite readily. Such a computer program for quarter-wave elements has been written and supplied by H. Hurwitz, WA2VBW. Using a modification of this program, the author has investigated a

Table 5

Case	Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	1	0	0	0	0°	3 dB	-3 dB	-α
	2	1	-90	0.25	0				
2	1	1	0	0	0	0°	4.1 dB	5.7 dB	-11.5
	2	1	-120	0.25	0				
3	1	1	0	0	0	90°	1.1 dB	0 dB	0 dB
	2	1	0	0.25	0				

Table 6

I1	I2	φ1	φ2	Beam Gain	Back Radiation	Minor Lobes
1	0.8	0	-120	4.0 dB	-5.5 dB	-10.8 dB
1	1	0	-120	4.1	-5.7	-11.5
1	1.2	0	-120	4.1	-5.6	-12.5
1	1	0	-100	3.4	-3.8	-11.1
1	1	0	-120	4.1	-5.7	-11.5
1	1	0	-140	4.6	-6.6	-6.6

Table 7

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	0°	5.6 dB	-8.7 dB	-8.7 dB
2	1	-120	0.25	0				
3	1	-240	0.5	0				

Table 8

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	1	0	0	0	90°	2.7 dB	0 dB	0 dB
2	1	0	0.25	0				
3	1	0	0.5	0				

Table 9

Element	Current	Phase	X	Y	Beam Direction	Beam Gain	Back Radiation	Minor Lobes
1	0.5	0	0	0.145	0°	4.35 dB	-6.1 dB	-15. dB
2	1	-110	0.25	0				
3	0.5	0	0	0.145				

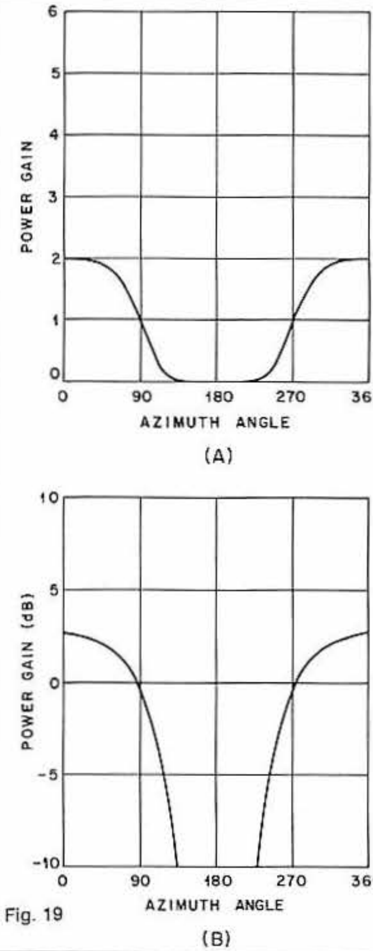

Fig. 19

large number of simple array configurations, which are potentially quite useful at the lower amateur frequencies, with the salient results given here. It is hoped that these will prove interesting to those considering construction of such arrays.

Technique for Optimizing Arrays

We must first try to define the problem and the parameters leading to an optimum design. The horizontal-plane pattern of an array of (vertical) elements depends only on the magnitudes of the individual element currents and their phases, and not on the way the currents are produced (e.g., driven or parasitic). The computer program allows one to specify the locations (x,y coordinates) of any number, n, of quarter-wave elements; also one must specify the magnitude of current and its phase in each element (usually normalized to unity current and zero phase for a given fiducial element). With these data inputs the radiation pattern is calculated and plotted out, giving the relative power gain in each azimuth angle interval, say, 5 or 10 degrees. The power gain is normalized to that power which would have been radiated at the same azimuth by a *single* vertical element using the same input electrical drive power. By using trial values of the current in all elements, in both amplitude and phase, one soon develops a feel for the behavior of the system, and can usually arrive at a "best" value for all the currents and phases. One can then try different geometrical separations of the elements and in this way arrive at the "best" spacing. Usually this best configuration is a compromise between power gain and good pattern control, e.g., front-to-back ratio, or more generally, main lobe to minor lobe (side or back) ratio.

In addition to the optimization described above it must be noted that if the array is to be used primarily to produce a "beam" in a given general direction with little emphasis on other directions, one usually finds a linear array of elements best (equivalent to an end-fire Yagi array). However, if the array is to be used for all directions, a more (circularly) symmetrical array should be considered. These points will become clear in the discussion of results.

Two-Element Array

In the case of two elements there is only one possible configuration. Consider one element at the origin of the x,y plane (coordinates 0,0) and the other element along the x axis (coordinate x,0). The parameters available are x, which we specify in units of wavelength, λ, the rf currents I1 and I2 in each element (we shall normalize one of the currents, e.g., I1, to unity) and the phases, φ1 and φ2, of the currents (we shall generally normalize the phase of the current in element 1 as zero and specify the relative phase of the current in element 2 in electrical degrees, either positive for leading currents or negative for lagging currents). It soon becomes apparent that the region of main interest is where the spacing x is in the neighborhood of 0.25, i.e. λ/4, and there the currents are nearly equal in magnitude. The reasons are that, for spacings of 0.5 or greater, other strong lobes appear in the pattern, and for grossly unequal currents relatively poor

Table 10

Element Spacing	*Optimum $\phi 2$*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
0.2	-130°	4.65 dB	-6.5 dB	-18 dB
0.25	-120°	4.6	-6.5	-16
0.3	-110°	4.4	-6.2	-14
0.35	-110°	4.2	-5.8	-12

Table 11

Element	*Current*	*Phase*	*X*	*Y*	*Beam Direction*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
1	1	0	0	0	0°	7.4 dB	-9.9 dB	-9.9 dB
2	1	-120	0.25	0				
3	1	-240	0.5	0				
4	1	0	0.75	0				

Table 12

Element	*Current*	*Phase*	*X*	*Y*	*Beam Direction*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
1	1	0	0	0	90°	4.2 dB	0 dB	0 dB
2	1	0	0.25	0				
3	1	0	0.5	0				
4	1	0	0.75	0				

Table 13

Element	*Current*	*Phase*	*X*	*Y*	*Beam Direction*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
1	1	0	0	0	45°	6.4 dB	-18 dB	-18 dB
2	1	-110	0.25	0				
3	1	-220	0.25	0.25				
4	1	-110	0	0.25				

Table 14

Element	*Current*	*Phase*	*X*	*Y*	*Beam Direction*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
1	1	0	0	0	0°	5.1 dB	-11.5 dB	-11.5 dB
2	1	-120	0.25	0				
3	1	-120	0.25	0.25				
4	1	0	0	0.25				

Table 15

Square side Length (inλ)	*ϕA opt.*	*ϕB opt.*	*Beam Gain*	*Back Radiation*	*Minor Lobes*
0.167	-130	-260	6.6 dB	-22 dB	-22 dB
0.25	-110	-220	6.4	-18	-18
0.333	-80	-160	5.6	-18	-18

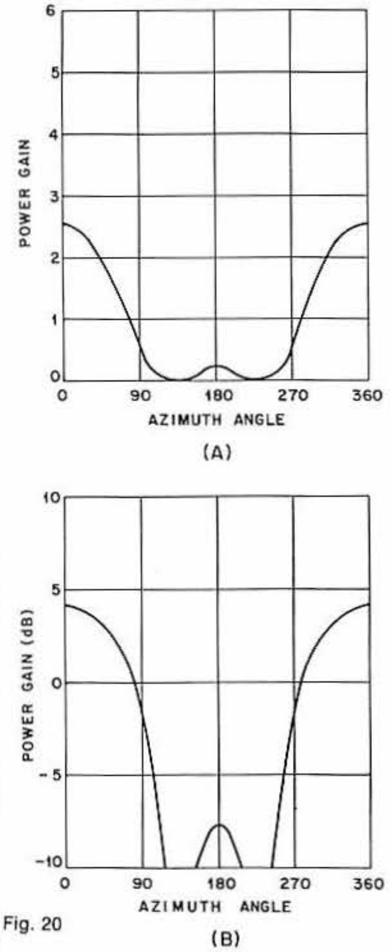

Fig. 20

peak gain and pattern discrimination result. Although the spacing x is not especially critical in the neighborhood of 0.25 (one can essentially compensate by adjustment of phase angle) we shall now examine the case where x = 0.25, and where I1 = I2 = 1. Best overall pattern discrimination (minimum back lobe) for the endfire case occurs where $\phi 1$ = 0, and $\phi 2$ = 90 degrees, and the power-gain pattern is shown in Fig. 19A.

Peak gain is 2 (or 3 dB) and occurs along the x-axis direction. Fig. 19A shows the power pattern of the array, i.e., it shows quantitatively where the radiated power goes. However, if one is looking at what an S meter would indicate, one can present exactly the same information on a vertical *logarithmic* scale as shown in Fig. 19B.

The vertical scale is shown in dB relative to the value for a single vertical radiator. This type of display is useful in judging excellence of pattern with respect to minor lobes. For all future cases discussed here, both displays will be shown.

It is possible to increase the gain of this array at the expense of some back lobe by increasing the phase lag in element 2, e.g., if $\phi 2$ = −120 degrees the peak gain, again along the x axis, is 2.59 (or 4.1 dB) and the entire pattern is shown again in Figs. 20A and B.

Note that the back lobe now shows a gain of 0.186, which is −11.5 dB relative to the forward lobe.

These two elements can be used to form either a beam along the +x axis as shown, or reversed along the −x axis (by reversing the phase $\phi 2$, i.e., making it +120 instead of −120 degrees) or if the two elements are driven in the same phase, i.e., $\phi 1 = \phi 2 = 0$, one produces a bilateral broadside beam as shown in Fig. 21.

Here the peak gain is 1.1 dB, but there is a "back" lobe of equal magnitude to the front lobe. Note that all patterns will be symmetrical around the x axis, or more generally around the line of any set of linearly-positioned elements. This example has illustrated some of the aspects of all of the results to be presented; namely, that:

a) The best element spacing will be found to be in the neighborhood of 0.25 wavelength.
b) Current ratios and phases can be adjusted to produce either a best *pattern discrimination* (minor lobe rejection) or best *main lobe gain.* It is generally not possible to get both at the same time, and a compromise must be made.

It may be useful to tabulate the salient results of each "best" case by noting the coordinates of the elements in the x,y plane, the individual drive currents, the individual phases, main beam heading in the x,y plane, main beam gain, and the

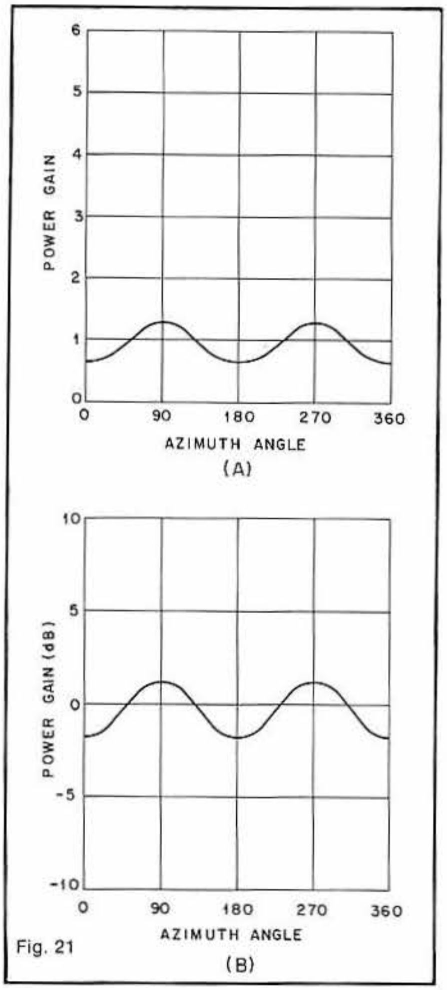

Fig. 21

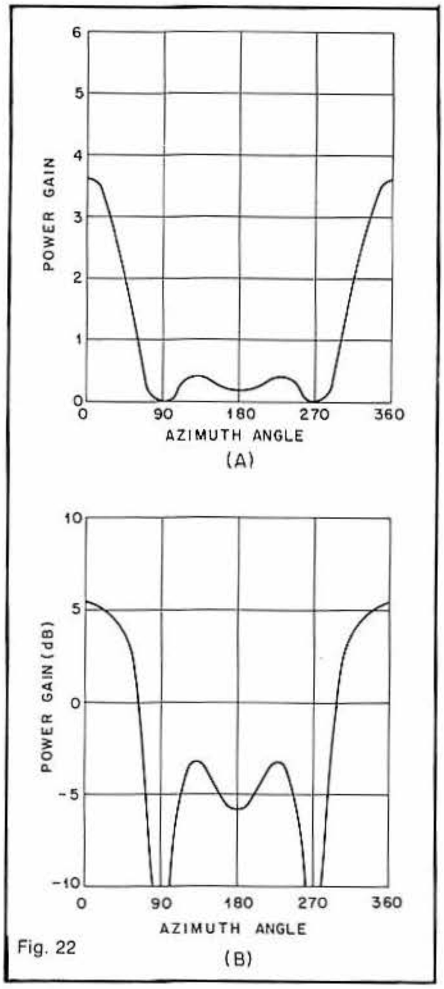

Fig. 22

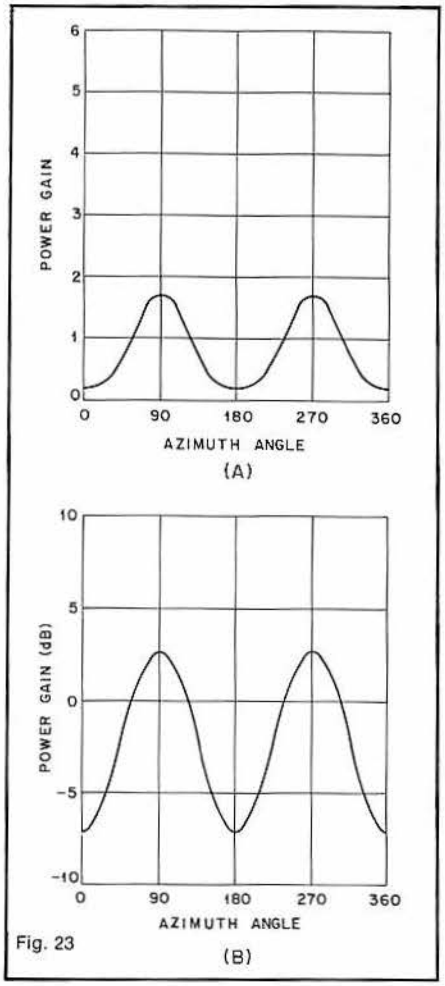

Fig. 23

largest pattern gain (such as given by either the side of the main lobe or by minor lobes) observed anywhere in the entire 180-degree sector centered just opposite in direction to the main lobe. We shall designate this quantity as "back radiation" and express it in dB relative to the main lobe gain. We shall also tabulate the largest minor lobe in this sector and label it "minor lobe," again expressed in dB relative to the main lobe. Thus for the cases just described, Table 5 is presented.

The sensitivity of these results to variations of currents and phases is not high. For example, one can summarize case 2 above with somewhat altered I2, or $\phi 2$, as shown in Table 6.

The author feels that end-fire case 2, where $\phi 2 = -120$ degrees, probably represents the "best" design with a good pattern and gain. Other more complicated cases presented below are optimized in this same way and only the "best" configuration will be shown with actual results.

Three-Element Arrays

For three elements, two different geometries suggest themselves: the linear array, best for a single preferred directional line, and an equilateral triangle configuration, the closest approach to circular symmetry. The latter, as we shall see, could be switched to give "beams" in any one of six angular positions (every 60 degrees). Let us consider first the linear array.

Linear Array — As in the 2-element linear array just discussed, the optimum arrangement seems to center around an element spacing of about 0.25 wavelength with approximately equal current drives. The optimum end-fire case is shown in Table 7; the broadside case is shown in Table 8.

The patterns for these cases are shown in Figs. 22 and 23.

Remember that for the end-fire case, which one can think of as a 3-element Yagi on its side, the rather large side lobe results because the pattern lies in a plane perpendicular to the elements. Side radiation is hard to avoid, as one cannot take advantage of the nulls off of the ends of excited elements.

Equilateral Triangular Array — For this case, the optimum situation appears to be as shown in Table 9. The behavior is very similar to the case of two elements in end-fire. See Table 5 (we have just split one of the two elements and have moved the separate pieces to the triangle corners). The pattern is shown in Fig. 24.

Reversing the phases will reverse the beam direction, and by rotating the drive currents (and phases) around the triangle, one can produce six beam directions (every 60 degrees). This is an attractive feature, but it does not come free; the maximum gain of the system is somewhat less than one obtains with the three linearly-disposed elements (which are chiefly useful in a direction along the line of the array).

The exact size of the triangular array is not critical — one must simply choose the best phase angle for the particular geometry. The "best" phase angles, $\phi 2 (\phi 1 = \phi 3 = 0)$, as determined by computer runs, for different spacings (triangle side lengths expressed in terms of λ) are shown in Table 10.

Four-Element Arrays

With four elements, there are at least three configurations which are interesting to investigate. As before, the linear array with a preferred beam direction is one type, but there are two configurations which possess a measure of circular symmetry. One of these is described by three antenna elements at the corners of an equilateral triangle with the fourth element at the geometric center of the triangle. This one we shall designate as a center-filled triangle. The other type is a square array.

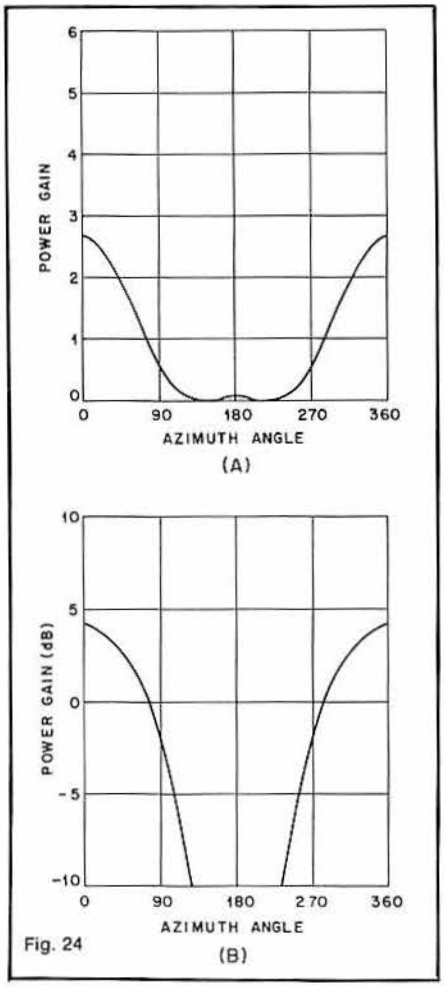

Fig. 24

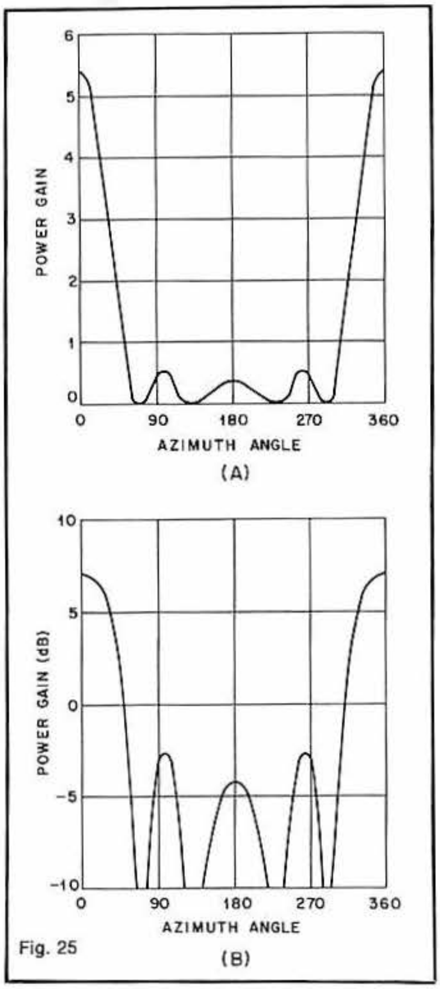

Fig. 25

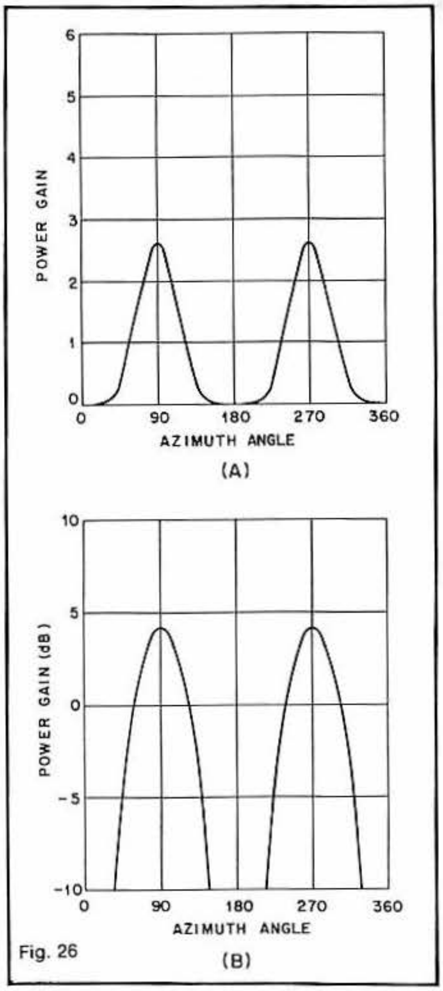

Fig. 26

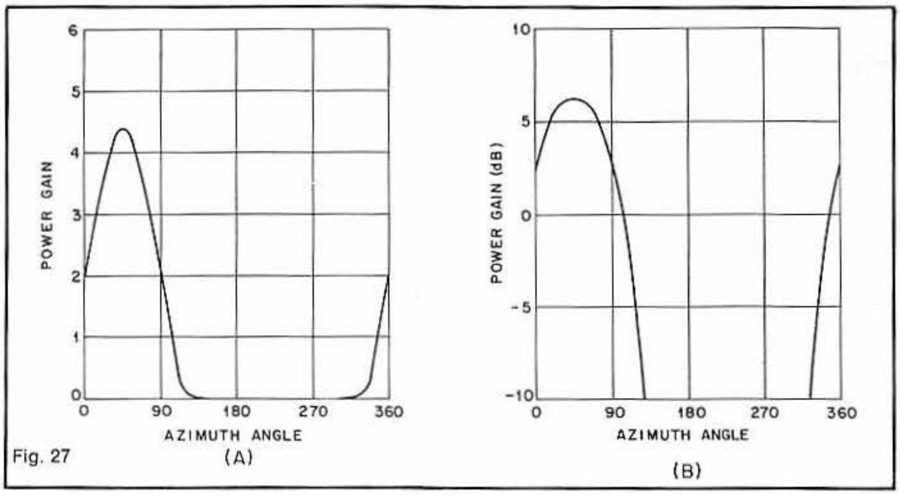

Fig. 27

Let us take these three types in order.

Four-Element Linear Array — For this case the optimum end-fire solution appears to be as shown in Table 11. The gain and major lobe shape of this configuration seems to be excellent. The broadside case is shown in Table 12. Patterns for this case are shown in Figs. 25A, B, and 26A, B.

Note that for this array there is poor coverage in some azimuthal directions (e.g., ±45 degrees) if only the two tabulated phase arrangements are used.

Center-Filled Triangle — A large number of situations were examined for this configuration, but unfortunately no combinations of spacing, currents or phases were found that looked very good. The best pattern was not particularly better than that of the three-element triangle alone, and was substantially inferior to that of the 4-element square to be discussed next.

Four-Element Square Array — For this case the best situation appears to be as shown in Table 13. Note that this arrangement fires diagonally and produces a respectable gain and excellent pattern. The "broadside" arrangement for the same square is shown in Table 14. The computed patterns are shown in Figs. 27 and 28.

The diagonal-fire pattern is exceptional, providing extremely low back radiation and minor lobes. The broadside pattern is not nearly as good, somewhat superior but reminiscent of the end-fire pattern obtainable with just 2 elements (see Fig. 20). In fact, the power-gain coverage at all angles (including the broadside angle) just using the switched four diagonal beams is large enough that it probably is not worthwhile using the broadside arrangement at all even to "fill" between diagonal beams.

This square array is superior to the center-filled triangle, and to all other simpler configurations (circular) in both gain and pattern discrimination. It is sufficiently good that one might ask what particular properties of the square have led to the nearly ideal performance. The

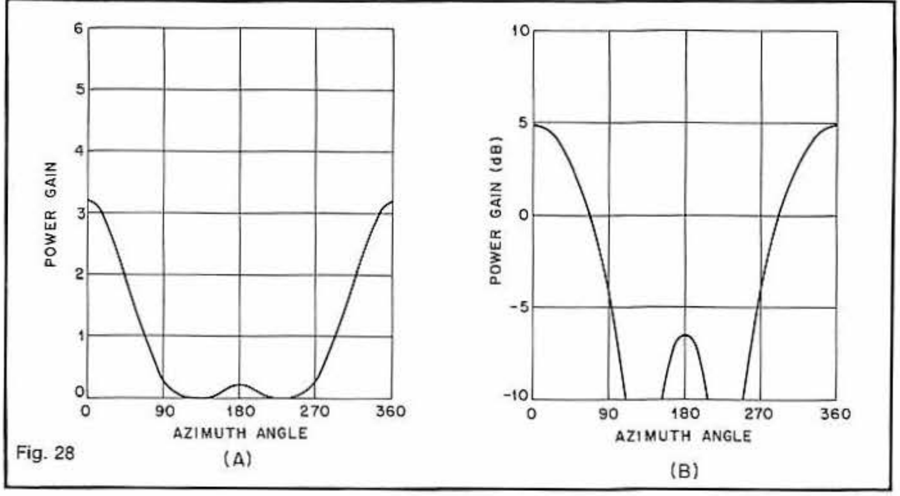

Fig. 28

author believes that there are two parameters which are noteworthy: first, the gain is respectable in diagonal fire because one can view the square along its diagonal as a three-element linear array in which the center element is split, each half being displaced to the other corners of the square. Thus the array behaves something like a 3-element (Yagi-like) linear array which is known to have a good gain. Secondly, the lobes are much smaller than the linear array due to the tendency for the phase cancellation of waves produced by the spatially-separated corner elements.

Again, as in previous examples, the actual dimensions of the square are not very critical, as long as one adjusts the phase (delay), ϕA in elements 2 and 4 and the larger phase delay, ϕB, in element 3 to the "best" value. Examples of such "best" phase values for different size squares are shown in the Table 15.

It turns out that the calculated gain figures for the smaller arrays are actually slightly larger than the gain for the larger squares, but as a practical matter this fact is probably offset by the larger (reactive) currents required for radiation. The larger currents are necessary because of the mutual coupling and phase relations between elements and the required phase relationships; in practice with the inefficiencies usually caused by ground currents when using vertical antennas, the gain differences for the various squares are probably inconsequential. The patterns for all cases are quite good, but the best pattern is obtained for the $\lambda/4$ square. Incidentally, the 0.167λ and 0.33λ square allows the possibility that a single square can be used for two amateur bands, and this has indeed become the subject of a practical antenna system.[1]

Conclusion

Computer trial runs, through a program developed by WA2VBW, have facilitated the investigation of a large number of $\lambda/4$ vertical antenna-array configurations. "Best" configurations, drive conditions, power gains and patterns have been found for the cases of two, three and four elements. Especially interesting is the case of four elements arranged in a square whose side dimension is about one-quarter wavelength.

Cases using more than 4 elements have been investigated, but the complexity, probable cost, and difficulty of installation of such configurations, reduces their utility and so they will not be reported here.

The author wishes to acknowledge the contribution of Henry Hurwitz, WA2VBW, who supplied the original computer program for this investigation. This material was originally presented in *QST* by James L. Lawson, W2PV.

References

[1] Lawson, W2PV, "75/80 Meter Vertical Antenna Square Array," *QST*, March 1971.

360-Degree Steerable Vertical Phased Arrays

Microwave power-splitting and phasing techniques were applied to the design of this 80-meter array. It has established an outstanding record for performance on long and difficult paths.

The possibility of providing effective 360-degree horizontal coverage with an array switched electrically, rather than rotated mechanically, has long intrigued many designers. Such a solution to the directional antenna problem is of particular interest to amateurs operating on the lower frequencies, where any mechanical rotation system is difficult and expensive to build. One of the authors, now W1CF, writing under his old call, W1HKK, described an 80-meter switchable phased array of four quarter-wave in-line elements more than 10 years ago.[1] The present article describes several novel array designs, using up-to-date technology, which provide greatly improved operation. Also covered are techniques used for effective power division, the beam-forming phasing system, switching, and impedance matching. Ground radials, cable and connector selection, and dc continuity testing are discussed.

One of these arrays recently constructed for 80-meter use has a single 97-degree lobe with a gain over a single vertical of approximately 7 dB at low angles. Its front-to-back ratio is 25 dB, and its front-to-side ratio is more that 12 dB, over the whole 80-meter band. It is switchable over 360 degrees in four 90-degree steps.

Computer-Modeled Arrays and Patterns

In the spring of 1975, while sitting over a cup of coffee, the three authors discussed the possibility of rebuilding the 1965 4-

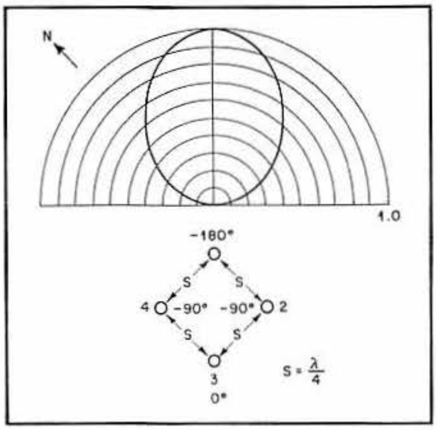

Fig. 29 — Polar plot of relative power, and planar view of the four-element diamond array showing the pattern obtained with no dc voltage on the switching relays, as in Fig. 31. Minor lobes are too far down to show on this scale.

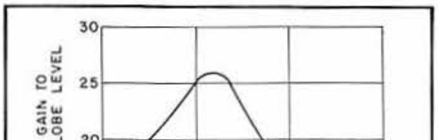

Fig. 30 — Bandwidth and side-lobe study of the four-element diamond array.

element array, using many improvements in technology that have evolved in the last 10 years. Dr. Joseph White accelerated the effort by coming up with a computer program written in Fortran IV, called ARRAY, which allows the user to plot patterns of arrays of from 2 to 99 elements in various spacings and combinations of phase and power, without ever leaving the computer terminal. Over 30 iterations using 4 elements were explored, and the computer-printed polar plots and other gain data analyzed, before we arrived at a configuration which provided good forward gain over a broad beamwidth, with no measurable unwanted lobes.

The configuration selected has four quarter-wavelength vertical elements in a square, with quarter-wave spacing between adjacent elements, as shown, with its predicted pattern, in Fig. 29. All elements are fed with equal amplitudes, the rear element at 0°, the two side elements at -90° and the lead element at -180°. The beam is transmitted along the diagonal from the rear to lead element. Gain due to horizontal beam formation alone is about 5.3 dB over a single vertical element. Front-to-back ratio is 25 dB. Front-to-side ratio is 12 dB, at 90° either side, increasing to much higher levels at 135° either side. Since most of the vertical energy is concentrated at low angles by the array, as much as 4 dB of gain additional to the predicted 5.3 dB in the horizontal plane can be achieved in theory, with perfectly conducting ground. The authors estimate that a good radial system and less-than-perfect ground yield an additional gain of 2 dB, or a total gain just over 7 dB for the system described.

The computer predicts a half-power beamwidth of 97°, so a suitable switching matrix can be used to direct the beam to four different quadrants, with only a slight loss of forward gain at the cross-over points, and virtually no deterioration of the front-to-back and front-to-side suppression. A computer study of the effect of variations in element spacing on forward-gain to maximum side-lobe level, plotted in Fig. 30, led to selection of $\lambda/4$ spacing of the corners of the square array. The 80-meter beam erected according to the computer-synthesized array design has proven to be remarkably successful. On-the-air results clearly approximate the performance anticipated by the program.

RF Power Dividers

Good power splitters are essential to the operation of phased arrays. Two-way and three-way power dividers will be discussed. Amateurs in broadcasting are well aware that tapped coils are the mainstay for power division in their industry. This reactive technique presents many problems.[2] The authors, being more versed in microwave techniques, thought it logical to scale up in wavelength the methods of power division used most successfully in the microwave region, particularly what is now known as the Wilkinson Power Divider,[3] shown in Fig. 31. Power from the transmitter is fed through a 50-ohm line of any length to a coaxial T, feeding two quarter-wavelength 70-ohm lines, W1 and W2. The two inner conductors of the 70-ohm lines are connected through a 100-ohm noninductive resistor, R1. This type of divider gives an equal power split, matches to 50-ohm loads at the two outputs, and has the unique property that any energy returning to the two outputs out of phase, due to mismatches or mutuals, is absorbed in the 100-ohm resistor.

The 90-degree phase-delay cable (DL) used in one side of each of the three power-splitting hybrids serves to assure that equal power reflections from the antennas are absorbed. Theoretically the resistors absorb none of the forward power. This technique provides approximately 30 dB of isolation from one output terminal to the other, to unwanted energy.

Reference is made in the 1965 article to difficulty in adjusting a phased array, because of element interaction due to mutual impedances between elements. The Wilkinson Power Divider, when used to feed phased arrays, reduces these problems and those resulting from imperfect match at the antenna inputs, and contributes to the realization of the predicted patterns. The "Wilkinson," is a remarkably simple and uncritical solution to the problem of power division. The 3-element vertical array described later uses a three-way "Wilkinson," shown in Fig. 33. It is constructed by splitting the 50-ohm input three ways, into three quarter-wave lengths of 93-ohm coaxial cable. The center conductors are connected together through 150-ohm noninductive resistors. An alternative approach is to return each center conductor to an ungrounded common point, through 50-ohm noninductive resistors.

Feeding, Switching and Phasing

The 4-element diamond array erected at W1CF uses three Wilkinson two-way

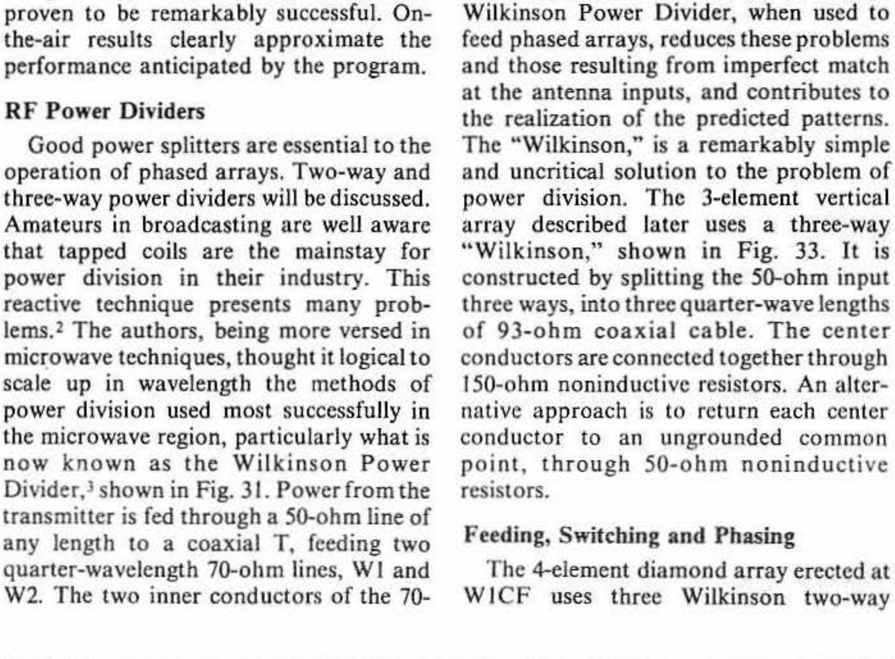

Terminations and switching relays at the center of W1CF array.

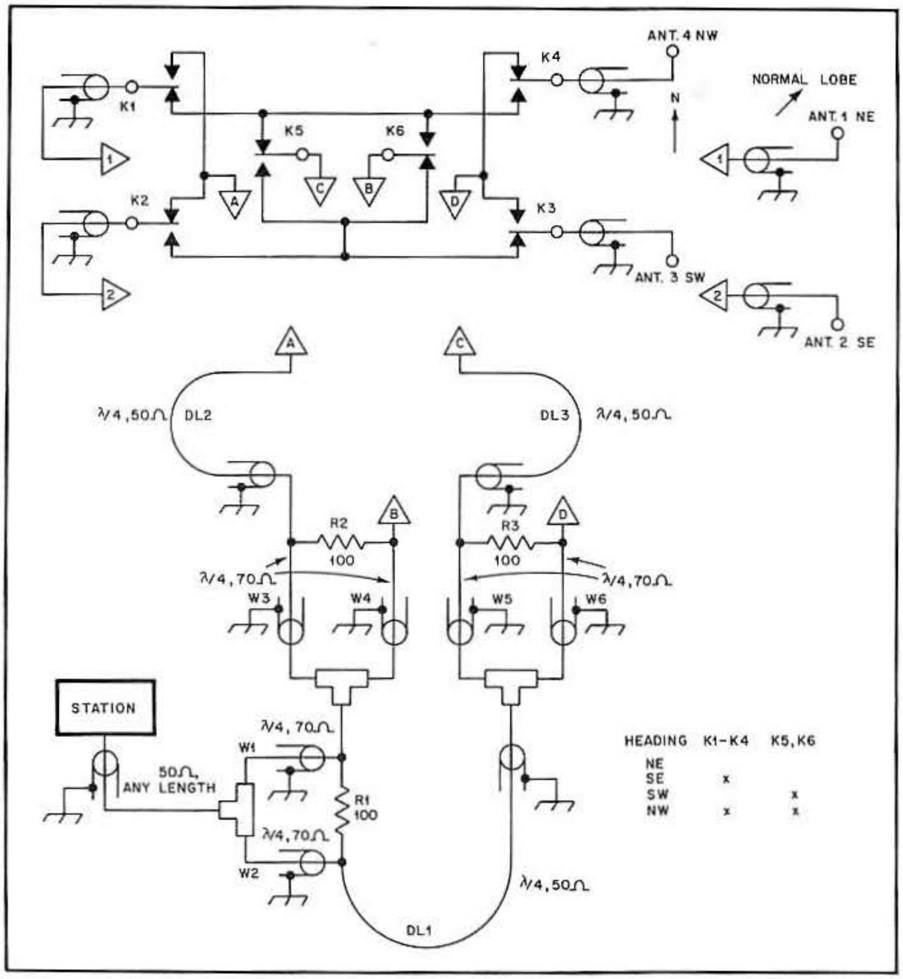

HEADING	K1-K4	K5, K6
NE		
SE	x	
SW		x
NW	x	x

Fig. 31 — Schematic diagram of the Wilkinson Power Dividers, phasing lines, and switching relays (rf connections) for the four-element array. Dc switching commands for the four pattern headings are given at the lower right.

power dividers, as shown in Fig. 31. Phasing is accomplished with three 90-degree sections of RG-8/U cable, DL1, DL2, DL3. The four outputs are at 0°, -90°, -90° and -180° respectively in phase relationship, with power from the transmitter divided into four equal parts. Every effort was made to preserve symmetry throughout the system, in the hybrids, phase shifters, rf switching, feeds, and antenna placement, in order to have the array perform uniformly as it is switched between the four headings.

The switching can be done with six spdt coaxial relays, as in Fig. 31, or with one transfer relay and four spdt relays. With no voltage on the relay coils, the arms are in the positions shown in Fig. 31, giving northeasterly directivity (57° true, on Paris) the heading used most of the time at W1CF. The array is about 500 feet (152.40 m) from the transmitter, necessitating use of remote relays. A prospective user should consider the tradeoff between cable and relay costs. If the array is close to the station, four equal lengths of 50-ohm line can be brought into the station, and the switching done with a 4-pole, 4-position switch.

The long run of line here was dictated by the desire to take advantage of ground qualities and concealment afforded by swampland 500 feet (152.4 m) from the house. Not wanting to be electrocuted, we used 28 volts dc for operation of the switching relays. By judicious use of binary arithmetic, the switching commands require only three wires. Selection of the four beam headings is done at the operating position with a 4-wafer, 4-position switch. Relays should be selected for reliability and ease of weatherproofing. The cluster of relays and teminations shown in the photograph is at the geometrical center of the W1CF array. The connectors are protected against moisture by 3/4-inch (19 mm) heat-shrink tubing, which should extend from the bottom half inch (13 mm) of the connector 2 inches (51 mm) down the cable. In addition, wide plastic wastebaskets are used to protect these components. The expensive precautions were taken as the result of hard experience with moisture in the cables and connectors in the 1965 array.

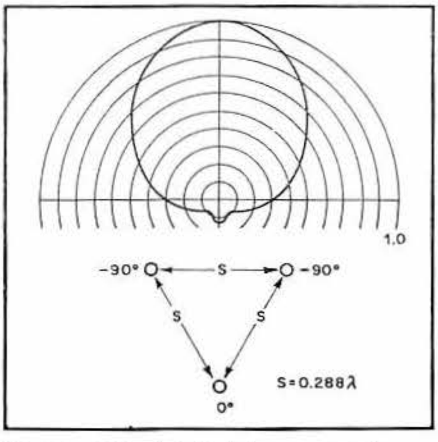

Fig. 32 — Polar plot of relative power, and planar view of the three-element triangular array.

Radiators and Radials

Relays K1 through K4 (Fig. 31) connect to their respective radiators through equal lengths of 50-ohm line, with no special attempt made for matching. The radiators are 57.5 feet (17.53 m) tall, above the base insulators. Constructional details follow the 1965 article, except as outlined below. Base insulators are 2-inch (51-mm) ID PVC pipe, 1/4 inch (6-mm) wall, 18 inches (457 mm) long, affixed to the aluminum with PVC cement and self-tapping screws. Each radiator is guyed at three levels with 3/16-inch (5-mm) galvanized cable, broken every 30 feet (9.14 m) with egg insulators.

Though 120 radials are considered to be the optimum number, only 40 per element are used here. Most are 65 feet (19.81 m) long, no. 12 galvanized wire, lying directly on the ground. This presents no hazard, as there is no foot traffic through the swamp area. The center radials are interconnected to two common busses in the central region, as recommended in reference 2. The authors feel intuitively that the square geometry of the array, in addition to providing considerable symmetry of the mutuals, allows a higher packing density for the radials, reducing ground losses.

Though the array has quarter-wavelength elements, all the feedpoints are at 50 ohms, and any type of vertical element can be used, provided an effort is made to match the inputs, and networks used for matching are *the same for all 4 elements.* Obviously, a height of 5/8 wavelength would be desirable, particularly for arrays built for higher frequencies. An indication of match would be to put rf voltmeters across the 100-ohm resistors, and tune the matching networks simultaneously for minimum voltage across the resistors. A nice thing about the λ/4 case is that the match is close enough that this step has not been necessary.

Testing

Before the array is fired up somebody should go to each element and short the input, while another person watches an ohmmeter placed across the main line at the station end. Make sure that a very low dc resistance is measured. Then, with the array in the normal position, as in Fig. 31, make sure that the resistance across the

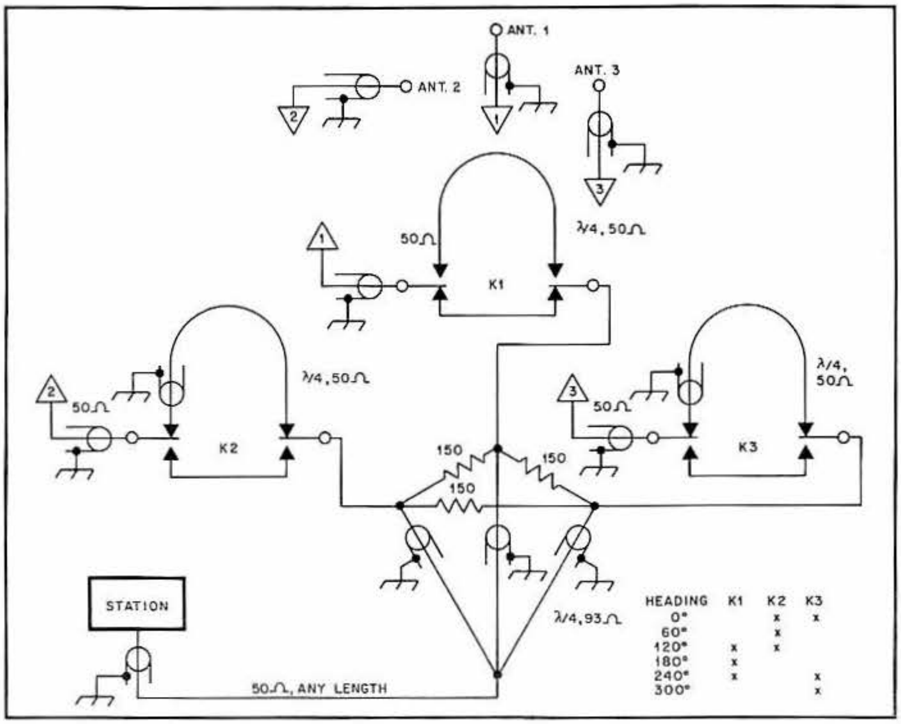

HEADING	K1	K2	K3
0°		x	x
60°		x	
120°	x	x	
180°	x		
240°	x		x
300°			x

Fig. 33 — Schematic diagram of a three-element vertical array and its directivity switching and power dividing circuitry. Interconnecting leads should be as short and symmetrical as possible. The 50-ohm lines between the relays and antenna elements should be all the same length.

input is high. If it is low, check for moisture in the cables or connectors or for other leakage resistance. Silicone grease in the connectors is a good moisture preventive measure. It is recommended that these resistance checks be repeated periodically to be sure that all is well.

When the array is ready for use, go easy at first, as any reflected energy will be dissipated in the 100-ohm resistors. If they become hot, better matching is necessary, and if they blow up, perhaps higher-wattage resistors are needed. At W1CF, resistor banks were made up of 6, 20-watt, 150-ohm surplus resistors in series-parallel to give 100 ohms. These handle the dissipated power when the system is driven from a 4CX1000A amplifier, and after a month of operation there was no evidence of resistor damage. No attempt should be made to hot-switch the directivity. This was done inadvertently, and the result was the need for extensive emergency repairs — in midwinter.

Performance

In the short time that the array has been in use, all continents have been worked on 80-meter phone, including several stations in Japan and India, tough paths from New England. Gain and side and back rejection on all signals except those arriving from very high angles are just what the computer program predicted.

In receiving in the "search" mode, one hand tunes the receiver while the other operates the lobe selector switch, to see which position "listens" best. The big W1CF rotary array for 10, 15 and 20 takes 45 seconds to rotate 360°, which tends to discourage frequent directional checks. With the phased array a complete scan takes but a few seconds. The high front and side rejection eliminates most of the interference from signals in unwanted directions, and in transmitting the clean patterns help to prevent ruffled feathers.

A surprising by-product has been the reduced atmospheric noise pickup from unwanted directions. In particular, when listening toward Europe atmospheric noise coming from electrical storms in the southwest is greatly reduced, improving the signal-to-noise ratio on signals arriving from across the pond. The array is an order of magnitude better on both transmitting and receiving than any antenna previously used on 80 at W1CF. It has more than held its own with the competition, in all four quadrants. Of special interest to operators who like to use both phone and cw on 80 is the fact that the SWR is close to unity, from 3.5 to 4 MHz.

Triangle 3-Element Array

Anticipating that some amateurs may wish to use three rather than four elements, the ARRAY program was used to explore equilateral triangle configurations. Though it was not reduced to practice, a 3-element vertical array with 0.288λ spacing between adjacent radiators looks promising. The lead element is driven at 0° and the two side elements at -90°. The forward lobe is 135° wide, with a gain of close to 5.5 dB, including an estimated 1 dB contributed by vertical focusing. Front-to-back ratio is 12 dB. With suitable switching, the pattern can be aimed in six different directions, again providing 360-degree coverage. The three-way Wilkinson mentioned earlier must be used with three 90-degree sections of 50-ohm line, to provide the proper phase relationships. Several other configurations appear promising, but involve more complex switching and power-dividing circuitry.

Applications for Other Frequencies

The diamond and triangle antenna configurations may find uses at other frequencies, as was done with the 1965 design.[4] Formulas are given below for determining the line lengths and element spacings. A 2-meter phased array can be mounted on the roof of a car and wafer-switched to the desired directions. For people who like to experiment with logic and rf switching with semiconductors, circuits can be devised so that the receiver can scan periodically, and the antenna steered to the desired signal.

The authors wish to thank the many people who have contributed to the success of this project, including W1FRR, K1GXT, W1HMV, W1FXT, K1CCK, M. E. Hines, F. Howe, H. Wells, R. Rearwin, H. P. Scott, and many others. We also wish to thank Microwaves Associates, Inc., for the use of their superb computer facilities.

Formulas for length of quarter-wave (90°) phasing lines, and spacing of radiators, are given below:

Electrical λ/4, solid-dielectric coaxial line, 50 or 70 ohms (RG-8, -11, -58, -59):

$$L_{ft} = \frac{162.36}{f_{MHz}}$$

Electrical λ/4, solid-dielectric coaxial line, 93 ohms (RG-62, -71):

$$L_{ft} = \frac{206.64}{f_{MHz}}$$

300-ohm Twin-Lead — parallel three λ/4 sections to approximate 93-ohm cable (requires 1 input and 3 output 1:1 baluns):

$$L_{ft} = \frac{201.72}{f_{MHz}}$$

Radiator spacing, diamond array:

$$S_{ft} = \frac{246}{f_{MHz}} > 0.25\lambda$$

Radiator spacing, triangular array:

$$S_{ft} = \frac{283.4}{f_{MHz}} > 0.288\,\lambda$$

This material was originally presented in *QST* by Dana W. Atchley, Jr., W1CF, Harold E. Stinehelfer and Joseph F. White, Ph.D.

References

[1] Atchley, "Switchable 4-Element 80-Meter Phased Array," *QST*, March 1965.
[2] National Association of Broadcasters, *Engineering Handbook*, 6th Edition.
[3] Wilkinson, "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*, January 1960.
[4] Botts, "A Four-Element Vertical Beam for 40/15 Meters," *QST*, June 1975.

An Etched-Circuit Monimatch for Checking Your Antenna System

Why buy an expensive SWR indicator when you can build this slick little unit at home? Not only will you save money, but you'll learn something about how SWR meters operate. Try the W1ICP circuit — you'll like it!

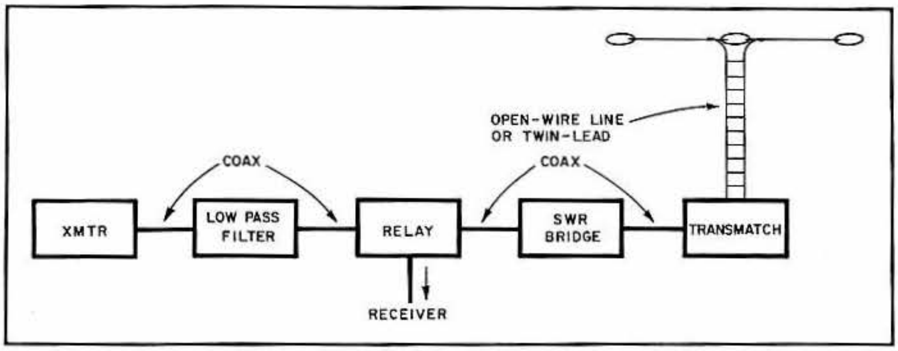

Fig. 34 — A typical setup for using a Monimatch in a multiband antenna system using a single dipole. The length of the dipole is not critical but it should be at least 1/4 wavelength overall at the lowest operating frequency for good efficiency. The feed line can be any length. The antenna relay and low-pass filter may be omitted if not needed.

A Novice who reads the instruction manual that comes with his transmitter and uses a 50-ohm dummy antenna can follow the tune-up procedures fairly accurately. With the dummy antenna he will find that the settings for the tuning controls will be fairly close to those specified by the manufacturer. However, when an antenna system is attached to the rig, in many instances the adjustments are far removed from any "book" setting. When this happens the Novice finds that he cannot get proper tuning of the rig, or worse yet, actually damages the equipment by trying to "force" it to work.

Nearly every transmitter these days, whether commercial or home-built, has a final amplifier stage that is designed to work into a 50-ohm load. If the load is something other than 50 ohms it may be impossible to tune the amplifier stage correctly. Of course, an important part of the problem is finding out what the load is — or, rather, how far from 50 ohms it happens to be. The piece of measuring gear described in this article is a device for doing just this. However, before describing the Monimatch and what it can do, let's take a little closer look at antenna-system loads.

The "50-Ohm" Load

The evolution of transmitter design since WW II has been influenced by several factors that have led to design that is more or less standard these days. First off, television came along right after the war and the hams quickly discovered that extremely tight shielding of a transmitter was needed to prevent undesired radiation that could cause TVI. However, when tight shielding was installed, band changing without band-switching became a real chore because there were so doggone many screws to unscrew and rescrew.

The one type of tank circuit that lent itself very well to the problem was the pi network. It was a fairly simple job to design a tightly shielded bandswitching transmitter, using the pi network, that would work into a 50-ohm load. Why 50 ohms? Simply because at this time 50-ohm coaxial cable had become a very popular type of transmission line. During the war, techniques were developed that made the manufacture of flexible coaxial cable a

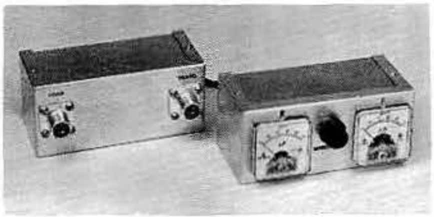

Here is the completed Monimatch with the two meters and the sensitivity control in the box at the right. The sensing unit is at the left.

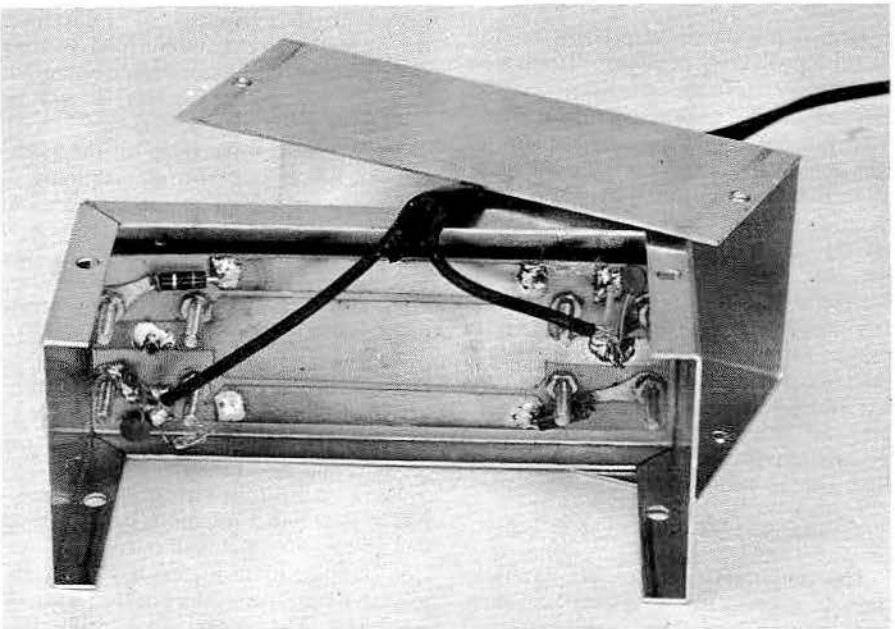

This is the sensing unit of the etched circuit Monimatch. As pointed out in the text, be sure to use a heat sink when soldering the diodes and resistors to the circuit board. The two shielded pickup leads are routed out the back of the Minibox, through a rubber grommet.

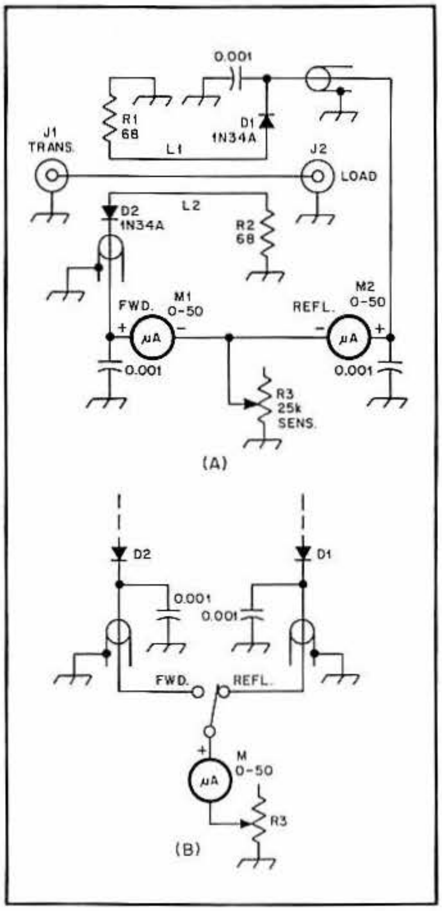

Fig. 35 — Circuit details of the etched-circuit Monimatch. The 0.001 μF capacitors are disk ceramic.
D1, D2 — 1N34A germanium diodes.
J1, J2 — Coax chassis fittings, type SO-239.
L1, L2 — See text and Fig. 36.
M1, M2 — 0-50 μA meter (Lafayette 99 H 5049).
R1, R2 — 68-ohm, 1/2-watt carbon or composition.
R3 — 25 kΩ control, linear taper.
S1 — Spdt switch.

reliable and economical process. So TVI and the availability of coax feed lines were the primary contributing factors that led to our present-day transmitter design.

If the load that is attached to the transmitter is something other than 50 ohms then the transmitter may be difficult to load, depending on a couple of other factors. While it is possible to design a pi network that will handle quite a wide variety of loads, many present day manufacturers, in order to compete in given price ranges, use a minimum number of parts in the tank circuit of the amplifier. For such rigs to operate properly the load must be between 25 and 75 ohms. This of course means that the user *must* furnish a load that will fall within this range.

Transmission Lines

The output terminal on all rigs these days is a coax fitting, which of course implies that a coaxial line must be attached to the rig. This doesn't mean that the coaxial line has to go all the way to the antenna. It could be connected to a Transmatch or a balun. What is important is that the first thing in the antenna system is the coaxial line that is attached to the rig.

Many Novices mistakenly believe that if you attach a 50-ohm coaxial cable to rig you *automatically* have a 50-ohm load. This is not true. The 50-ohm designation on the cable merely means that 50 ohms is the *characteristic impedance* of the cable. The characteristic impedance of a transmission line is determined by the size of the conductors used, the spacing of the conductors, and the dielectric material used to separate and support the conductors. The only time you would have a 50-ohm load using 50-ohm cable is when the line is terminated in its characteristic impedance. In other words, if the antenna has an impedance of 50 ohms then you will have a 50-ohm load on the rig. If the antenna has an impedance of other than 50 ohms then the load at the transmitter will be something other than 50 ohms. This in turn leads us up to a short discussion of standing-wave ratio.

SWR

If a transmission line is terminated in its characteristic impedance, all the power fed into the line from the transmitter will be delivered to the load end — in this case, the antenna. Actually, not quite all the power will reach the antenna because there is always some loss in the transmission line itself. However, what is important is that when the line is terminated in its characteristic impedance none of the power that reaches the end is reflected back toward the transmitter; all of it is used up in the antenna.

When the antenna impedance is different from the line impedance some of the power will be reflected back toward the transmitter end. Standing waves of voltages and currents will then exist on the transmission line. When this happens, the transmitter will no longer "see" a 50-ohm load. Exactly what the load will be will depend on several factors,[1] but suffice to say it will be something other than 50 ohms.

The standing-wave ratio on the transmission line is the ratio of maximum to minimum voltage or maximum to minimum current that exists along the line. If the line were matched in its characteristic impedance the voltage would be the same along the line and of course the SWR would be 1 to 1. The SWR is determined by dividing the resonant antenna impedance into the line impedance, or vice versa. For example, if the antenna impedance were 25 ohms and a 50-ohm line were used, the SWR would be 2 to 1.

For a moment, let's assume that regardless of how bad a mismatch exists, we are still able to tune and load our transmitter. The question then arises, how does the mismatch affect the losses in the transmission line? The answer to the question depends on how efficient the transmission line is.

Remember earlier we said there are always some power losses in every transmission line. If we have a mismatch at the antenna end, some of the power that reaches the end will be reflected back down the line. In traveling back, some of this power will be dissipated in the line, and the higher the SWR the higher these additional losses will be, because a higher SWR means that a greater proportion of the power will be reflected. In a transmission line that is 100 percent efficient (one that has no losses) it follows that regardless of how high an SWR exists, we wouldn't have any losses due to the SWR. Unfortunately, there "ain't no such" line, although some types of lines are much less lossy than others.

Also unfortunately, coaxial lines fall into the class that can be considered to be lossy lines. Just as an example, let's assume that you are using 100 feet (30.48 m) of RG-58/U on the Novice 15-meter band, and you are getting 50 watts out of your 75-watt Novice rig. This 50 watts is what is leaving your transmitter on the way to antenna via the 100 feet (30.48 m) of line. The loss for 100 feet (30.48 m) of RG-58/U at 21 MHz is 1.9 decibels. Translating this figure to power, we would lose about 20 of our 50 watts in the losses in the feed line, leaving only 30 watts to reach the antenna and be radiated. This is assuming the antenna impedance to be 50 ohms, the same as that of the line. If there is a mismatch the losses will be higher, as pointed out earlier. Suppose the SWR is 3 to 1, using the same setup. The *additional* loss in the system because of the SWR would be 1 dB, or a total of close to 3 dB. A 3-dB loss represents almost exactly one-half the power — that is, only 25 watts reach the antenna to be radiated. RG-8/U cable has the same characteristic impedance as RG-58/U but has less loss because it has larger conductors and more spacing between the conductors (and of course is more expensive).

The closest thing to a lossless transmission line is open-wire line. An open wire line on 21 MHz has only 0.08 dB loss per 100 feet (30.48 m). Even with a very high mismatch — for example, an SWR of 20 to 1, the additional losses are still less than 1 db!

This should not be interpreted to mean that coax is an undesirable type of line to use. For beam antennas it is difficult to beat the ease and convenience of using coax. However, for a single antenna, such as a dipole that is to be used on all bands and all frequencies, the best system is one consisting of a Transmatch and a feed line of open-wire line, such as shown in Fig. 34. With this system you can forget about line losses, SWR on the line, and mismatches between the antenna and feed line. By correctly adjusting the Transmatch, you can always give your rig a 50-ohm load regardless of what the load is on the antenna side of the Transmatch. If we use an SWR bridge in the short length of 50-ohm coaxial line that connects the rig to the

Transmatch, we can adjust the Transmatch so that the 50-ohm line shows a match, or SWR of 1 to 1, and the transmitter always sees a 50-ohm load.

The SWR bridge in the photographs and drawings is simple to build, and when installed in 50-ohm cable, will show the relative mismatch in the line, and also will indicate when you get the Transmatch properly adjusted for a match. The SWR bridge can also be used as an output indicator, which is very handy when tuning up the rig.

SWR Bridge Circuit Details

The etched-circuit Monimatch shown here is a reflectometer that samples the forward and reflected voltage in a 50-ohm line. Fig. 35 shows the circuit diagram. L1 and L2 are the pickup lines. In operation, a very small amount of power is coupled into the pickup lines and the rf voltages are rectified by D1 and D2. The rectified voltages are then fed to the two meters, M1 and M2, and the SWR then determined from the readings.

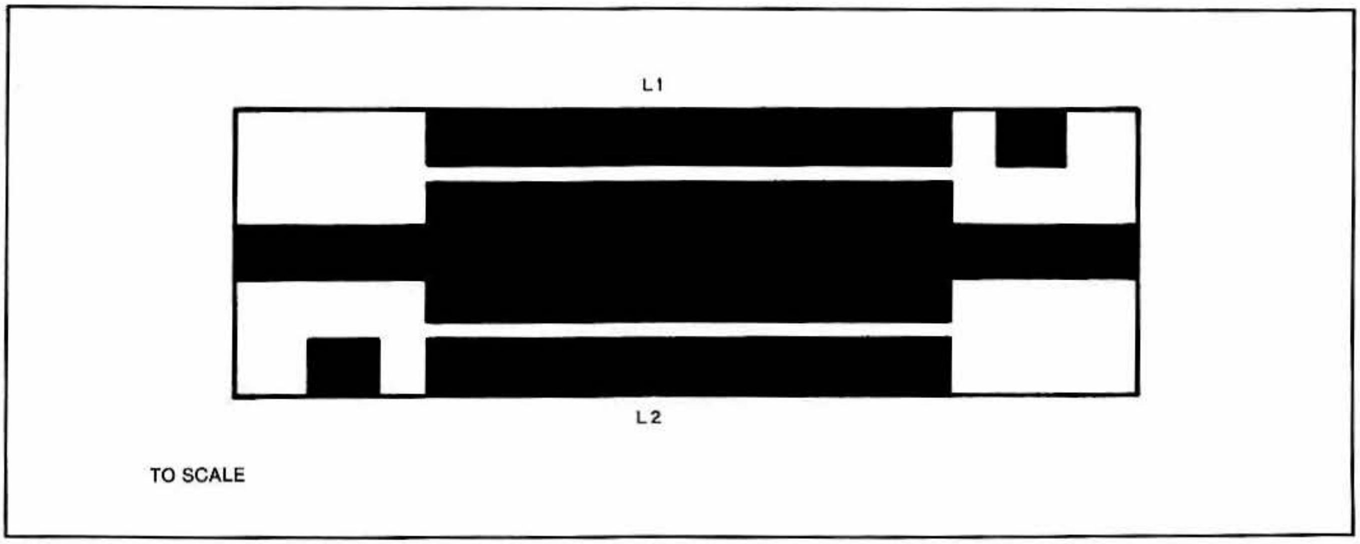

Fig. 36 — Etched circuit board template. The foil side is shown; the etched portion is shaded.

While a Monimatch is not a precise piece of measuring equipment, the SWR readings will be close enough for practical purposes. In order to determine the SWR, the forward-reading meter is set to full scale by adjusting R3, the sensitivity control, and then the reading on the "reflected" meter is noted. The formula for the SWR using this system is

$$\frac{F+R}{F-R}.$$

For example, let's assume the "reflected" reading is 5, with the "forward" reading being 10. Then 15 divided by 5 (that is, 10 + 5 divided by 10 − 5) would mean the SWR is 3 to 1. The closer the "reflected" reading is to zero, versus full-scale reading on "forward," the closer you come to 1 to 1, or a matched condition.

In the unit shown, two meters are used, one for the forward and the second for the reflected reading. However, if desired a single meter can be used and switched as shown in Fig. 35 at B. We used two meters in the indicator as this permits constant monitoring of what is happening in the line. The meters are inexpensive ones made in Japan.

Construction Details

Fig. 36 is a full-sized template of the etched circuit board. A *QST* article[2] went into detail showing simple methods for making etched circuits, so we won't treat the process here. In making this board, it is suggested that the board be covered with masking tape and then the pattern of Fig. 36 transferred to the tape. Using a sharp knife edge or razor blade and a straight edge, the masking tape can be carefully and accurately cut to the pattern.

After the board is etched, it can be positioned in the Minibox over the chassis connector holes and the board can then be marked at the drilling points for the mounting holes and the center conductor pins of the coax fittings. When installing the mounting screws, be sure they don't short to the center — conductor portion of the foil on the board.

There are a couple of other construction points that should be stressed. The lead lengths on R1 and R2 should be kept as short as possible. Also, be sure to use carbon or composition resistors, *not* wire-wound. When mounting the resistor and diode ends to the pickup sections, L1 and L2, the connections should be at the very ends of the sections. Also, use a heat sink when soldering the leads on any of the components mounted on the board, as too much heat from the iron can ruin the component.

The Monimatch and meters are mounted in separate Miniboxes, 2-1/4 × 2-1/4 × 5 inches (57 mm × 57 mm × 127 mm). The two connectors on the Monimatch sensing unit, J1 and J2, are mounted with their center pins 3-3/4 inches (95 mm) apart, center-to-center. In order to avoid an impedance "bump" in the feed line when the bridge is inserted in the line, the circuit board should be mounted 1/4 inch (6 mm) above the base of the Minibox. Quarter-inch (6 mm) spacers can be used under the circuit board at the screws holding both the board and the coax fittings to position the board accurately.

Shielded conductors should be used for the connections from the diodes to the meter enclosure. The shields should be grounded to the chassis at both boxes. These lead lengths are not critical, and the Monimatch can be remote from the meter indicator.

Using the Bridge

If you are using coax feed from the rig to the antenna, the bridge can be installed at any convenient spot in the line. If you are using a Transmatch, similar to the system as shown in Fig. 34, the Monimatch should be installed on the transmitter side of the Transmatch. Any relays or filters should be installed between the bridge and the transmitter, as shown.

Set R3 so that the arm of the control is at the top of the resistance — in other words, with all the resistance in series with the meter circuit. Tune up your rig in the normal fashion, and once tuned up adjust the sensitivity of the "forward" meter by moving the arm of R3 until the meter reads full scale. You can then determine the SWR by the formula mentioned earlier.

When adjusting a Transmatch, feed just enough power through the system to obtain about half-scale reading on the "forward" meter and then adjust the Transmatch controls for a match, as indicated by zero reflected power. You may have to adjust R3 as you adjust the Transmatch to keep the "forward" meter from reading more than full scale. Once you have the Transmatch adjusted for a 1-to-1 ratio as indicated by the bridge, the transmitter can be loaded up in the normal manner. We usually reduce the forward reading to about half scale, and then tune the rig for maximum output, as indicated by the meter. When doing this, you may notice that maximum output as indicated by the bridge meter occurs at some setting other than the normal transmitter plate meter "dip" reading. (Normally, the instruction manuals tell you to tune for a plate dip if the transmitter has no output meter.) However, the amplifier stage will work better if you tune for maximum output rather than the dip — keeping the plate loading within the transmitter ratings, of course.

Once you become familiar with the use of the bridge and interpreting the readings, you'll find it a very valuable device in your station. This material originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] [Editors Note: It is recommended that the newcomer obtain a copy of *The ARRL Antenna Book* for more detailed information on transmission line operation than can be provided in this article.]

[2] Schiebold, "Fast'n' Easy Printed Circuit Boards," *QST*, August 1969.

In-Line RF Power Metering

Problems are frequently encountered in designing and building reflected-power meters and SWR bridges. Examples of practical in-line rf wattmeters are given here, along with complete details for building a unit that will provide two power ranges, forward and reflected, for use from 3.5 to 30 MHz.

It is neither costly nor difficult to build an rf wattmeter. And, if the instrument is equipped with a few additional components it can be switched to read reflected power as well as forward power. With the foregoing feature the instrument can be used as an SWR meter for antenna matching and Transmatch adjustments.

Perhaps the most difficult task faced by the constructor is that of calibrating the power meter for whatever wattage range he desires to have. The least difficult method is to use a commercial wattmeter as a standard. If one is not available, the power output of the test transmitter can be computed by means of an rf ammeter in series with a 50-ohm dummy load, using the standard formula, $P = I^2R$. Or, if one is not interested in obtaining power readings the bridge can be used solely as an SWR indicator, as is done with the Monimatch-style SWR bridge.[1]

µA
ANTENNA
REF
FWD

The advantage of the circuits shown here over those of Monimatch bridges is that these instruments are not frequency-sensitive. Monimatch indicators become more sensitive as the operating frequency is increased, thus making it impractical to calibrate them in watts for more than one band, or for more than one portion of a given band. The units described here are more sensitive than Monimatches are. This makes it possible to calibrate them for power levels as low as 1 watt, full scale, in any part of the hf spectrum.

All of the circuits shown in this article are similar to the basic one which was described in *QST*.[2] Some of the circuits are those of commercial power meters, and are used to illustrate variations in the basic Bruene design. The reader may wish to experiment with some of these circuits.

Design Philosophy

Referring to the circuit of Fig. 37B, the circuit used by Collins Radio Company, the transmission-line center conductor passes through the center of a toroid core and becomes the primary of T1. The multi-turn winding on the core functions as the transformer secondary. Current flowing through the line-wire primary induces a voltage in the secondary which causes a current to flow through resistors R5 and R6. The voltage drops across these resistors are equal in amplitude, but 180 degrees out of phase with respect to common or ground. They are thus, for practical purposes, respectively in and out of phase with the line current. Capacitive voltage dividers, C3C7 and C4C8, are connected across the line to obtain equal-amplitude voltages *in* phase with the line voltage, the division ratio being adjusted so that these voltages match the voltage drops across R5 and R6 in amplitude. (As the current/voltage ratio in the line depends on the load, this can be done only for a particular value of load impedance. Load

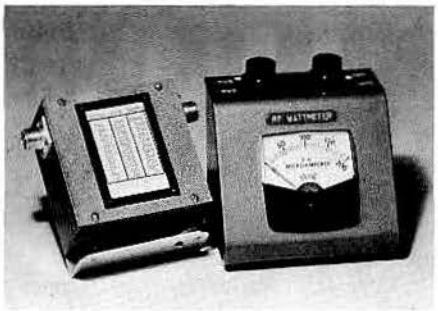
RF WATTMETER

values chosen for this standardization are pure resistances that match the characteristic impedance of the transmission line with which the bridge is to be used, usually 50 or 75 ohms.) Under these conditions, the voltages rectified by D1 and D2 represent, in the one case, the vector *sum* of the voltages caused by the line current and voltage, and in the other, the vector *difference*. With respect to the resistance for which the circuit has been set up, the sum is proportional to the forward component of a traveling wave such as occurs on a transmission line, and the difference is proportional to the reflected component.

The Collins circuit uses two 8-μF capacitors, C5 and C6, to permit the meter to approach the PEP level during ssb operation. The dc voltages in the forward and reflected lines charge the capacitors to permit a near-peak reading. The discharge rate is set by the series calibrating resistors, R1 through R4, and is dependent upon which of them is switched into the metering line at a given time. The circuit of Fig. 37B uses two 43-pF capacitors, C1 and C2, to cancel the inductive reactances of R5 and R6. Such reactance may become manifest at the high end of the range for which the instrument is built. If reactance is present in that part of the circuit the meter readings may not be accurate, especially at 10 and 15 meters. The capacitors were not needed in the circuit of Fig. 38, perhaps because the resistor leads were very short when they were mounted on the etched-circuit board.

Some Design Hints

It is important that the layout of any rf bridge be as symmetrical as possible if good balance is to be had. The circuit-board layout for the instrument of Fig. 38 meets this requirement. Also, the input and output parts of the equipment should be isolated from the remainder of the circuit so that only the sampling circuits feed voltage to the bridge. A shield across the end of the box which contains the input and output jacks, and the interconnecting line between them, is necessary. If stray rf gets into the bridge circuit it will be impossible to obtain a complete zero reflected-power reading on M1 even though a 1:1 SWR exists.

Referring again to Fig. 38, resistors R1 and R2 should be selected for the best null reading when adjusting the bridge into a resistive 50- or 75-ohm load. Normally the value will be somewhere between 10 and 47 ohms. The 10-ohm value worked well with the homemade instruments shown here. It was found that half-watt resistors exhibited somewhat less inductive reactance at 30 MHz than did some 1-watt units tried. R1 and R2 should be as closely matched in resistance as possible. They need not be exactly 10 ohms, so a VTVM can be used to match them. The resistors used for the circuit of Fig. 38 were actually 10.5 ohms each, and were chosen from an assortment of "10-percenters" on hand.

Silver-mica capacitors C3 and C4 were close enough in value so that special selection was not required. There should be enough leeway in the ranges of C1 and C2 to compensate for any difference in the values of the 330-pF capacitors. Ideally, however, C3 and C4 should be matched in value.

Diodes D1 and D2 should also be matched for best results. An ohmmeter can be used to select a pair of diodes whose forward dc resistances are within a couple of ohms of being the same. Similarly, the back resistances of the diodes can be matched. The matched diodes will help to assure equal meter readings when the bridge is reversed. (The bridge should be perfectly bilateral in its performance characteristics.) Germanium diodes are used in the bridges described here, but

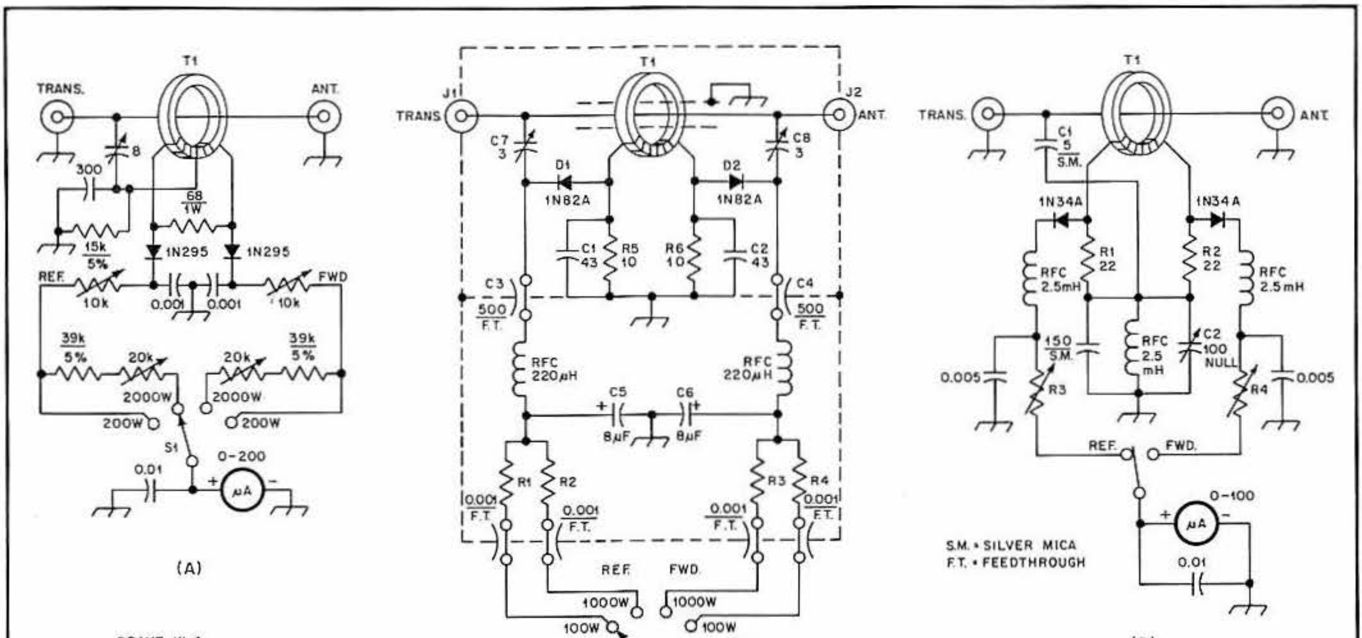
T1
TRANS.
ANT.
8
300
68
1W
15k
5%
1N295
1N295
REF.
FWD
10k
0.001
0.001
10k
39k
5%
20k
20k
39k
5%
2000W
2000W
200W
200W
S1
0-200
0.01
μA
(A)
DRAKE W-4
T1
J1
J2
TRANS
ANT.
C7
3
C8
3
D1
D2
1N82A
1N82A
C1
43
R5
10
R6
10
C2
43
C3
500
F.T.
C4
500
F.T.
RFC
220μH
RFC
220μH
C5
C6
8μF
8μF
R1
R2
R3
R4
0.001
F.T.
0.001
F.T.
0.001
F.T.
0.001
F.T.
REF.
FWD.
1000W
1000W
100W
100W
S1
M1
0-200
μA
COLLINS 302C-1/2
(B)
T1
TRANS.
ANT.
C1
5
S.M.
1N34A
1N34A
R1
22
R2
22
RFC
2.5mH
RFC
2.5mH
150
S.M.
RFC
2.5
mH
C2
100
NULL
0.005
0.005
R3
R4
REF.
FWD.
0-100
μA
0.01
S.M. • SILVER MICA
F.T. • FEEDTHROUGH
(C)

Fig. 37 — Schematic diagrams of typical in-line power meters. At A, the R. L. Drake W-4 instrument. It uses a center-tapped transformer at T1 and has but one capacitive voltage divider in the sensing circuit. The circuit at B is discussed in the text, and is used by Collins Radio Company. The capacitive voltage dividers in this circuit use two 500-pF feedthrough capacitors in place of the silver-mica capacitors specified in Fig. 38. Capacitors C5 and C6 permit a charge time that enables the meter to read near-peak power on ssb. Calibrating resistances R1 through R4 are factory selected. The circuit at C is similar to one used by Comdel in their power meter. In this circuit C1 is a fixed-value (small) capacitor, and the bridge is nulled by the larger capacitor in the divider, C2.

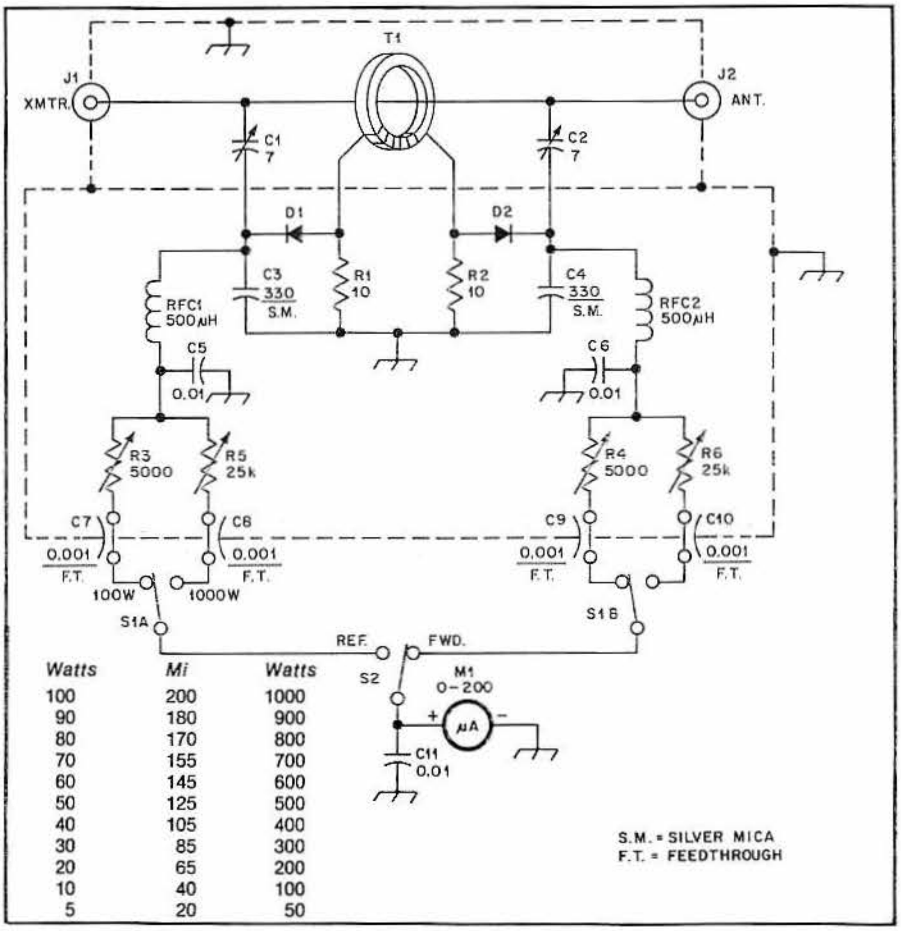

Watts	*Mi*	*Watts*
100	200	1000
90	180	900
80	170	800
70	155	700
60	145	600
50	125	500
40	105	400
30	85	300
20	65	200
10	40	100
5	20	50

Fig. 38 — Schematic diagram of a practical power wattmeter. A calibration scale for M1 is shown also. Fixed-value resistors are 1/2-watt composition. Fixed-value capacitors are disk ceramic unless otherwise noted. Decimal-value capacitances are in µF. Others are pF. Resistance is in ohms; k = 1000.

C1, C2 — 1.3- to 6.7-pF miniature trimmer (E. F. Johnson 189-502-4. Available from Newark Electronics, Chicago, IL).
C3-C11, incl. — Numbered for circuit-board identification.
D1, D2 — Small-signal germanium diode. IN34A, etc. (see text).
J1, J2 — Chassis-mount coax connector of builder's choice. Type SO-239 used here.
M1 — 0- to 200-µA meter (Triplett type 330-M used here.)
R1, R2 — Matched 10-ohm resistors (see text).
R3, R4 — 5000-ohm printed-circuit carbon control (IRC R502-B).
R5, R6 — 25,000-ohm printed-ciruit carbon control (IRC R253-B).
RFC1, RFC2 — 500-µH rf choke (Millen 34300-500 or similar).
S1 — Dpdt single-section phenolic wafer switch (Mallory 3222J).
S2 — Spdt phenolic wafer switch (Centralab 1460).
T1 — Toroidal transformer; 35 turns of no. 26 enam. wire to cover entire core of Amidon T-68-2 toroid (Amidon Assoc., 12033 Otsego St., No. Hollywood, CA 91607).

silicon diodes can also be used. Silicon diodes conduct at a higher voltage than germanium diodes do — approximately 0.7 volt — and will not work too well in low-power wattmeters. Some silicon diodes were tried, but ceased to conduct at approximately 8 watts in the circuit of Fig. 38. This effect can cause misleading results when low values of reflected power are present during antenna adjustments. The SWR can appear to be zero when actually it isn't. The germanium diodes conduct at approximately 0.3 volt, making them more suitable for low-power readings.

Any meter whose full-scale reading is between 50 microamperes and 1 milliampere can be used at M1. The more sensitive the meter, the more difficult it will be to get an absolute reflected-power reading of zero. Some residual current will flow in the bridge circuit no matter how carefully the circuit is balanced, and a sensitive instrument will detect this current flow. Also, the more sensitive the meter, the larger will have to be the calibrating resistances, R3 through R6, to provide high-power readings. A 0- to 200-microampere meter represents a good compromise for power ranges between 100 and 2000 watts.

Construction

The power meter of Fig. 38 is built in two sections. The rf circuit and the calibrating resistors are housed in a 4 × 4 × 2-inch (102 × 102 × 51 mm) aluminum utility box. All components other than J1, J2, and the feedthrough capacitors, are assembled on the etched-circuit board.³ Switches S1 and S2, and the meter, M1, are installed in a sloping-panel utility box which measures 5 × 4 inches (127 × 102 mm). Four-conductor shielded cable — the shield serving as the common lead — is used to join the two pieces. There is no reason why the entire instrument cannot be housed in one container, but it is sometimes awkward to have coaxial cables attach to a unit that occupies a prominent place in the operating position. Built as shown, the two-piece instrument permits the rf pickup head to be concealed behind the transmitter, while the control head can be mounted where it is accessible to the operator.

Toroidal transformer T1 fits into a cutout area on the circuit board. A 1-inch (25 mm) long section of RG-8/U cable — vinyl jacket and shield braid removed — provides a snug fit in the center hole of the toroid, and is used to complete the line between J1 and J2. The inner conductor of the RG-8/U section solders to the circuit board, thus holding T1 in place.

A flashing-copper shield divides T1 and its center-conductor line from the remainder of the circuit. This partition is shown in dotted lines in Fig. 38. It is mounted on the non-foil side of the circuit board and is secured at each end to solder lugs which are mounted under the retaining screws for J1 and J2.

The circuit board is held in place, at the end near T1, by means of an aluminum L bracket. The circuit-board end nearest the feedthrough capacitors is held in place by a no. 6 spade bolt. A solder lug is mounted under the no. 6 nut (outside the case) which secures the spade bolt. The lug serves as a connection point for the common lead between the rf head and the control box. Two solder lugs are mounted under the bottom two retaining screws of each coax connector. The free ends of the lugs are soldered to the copper foil of the circuit board.

A partition is visible in the foil-side view of the rf head. It can be eliminated if desired, since it did not prove necessary when the unit was tested. Similarly, an extra shield partition is shown on the top side of the board. It too can be eliminated, for it turned out to be unnecessary. The flashing-copper shield discussed earlier is the only one required for the circuit of Fig. 38.

Checkout and Tune-up

Once the instrument is wired and ready to test it should be inspected for unwanted solder bridges between the circuit-board foils. It is usually a good idea to scrape out the rosin buildup between the foils, and this can be done with the blade of a small screwdriver. A continuity check for "opens" and "shorts" should also be made before power is applied to the unit. Make certain that the diodes are installed for the correct polarity — the banded ends (cathodes) toward C1 and C2.

Connect a noninductive 50-ohm dummy load to J2. A Heath Cantenna or similar load will serve nicely for adjustment purposes. Place S2 in the FORWARD position, and set S1 for the 100- range. An rf ammeter or calibrated power meter

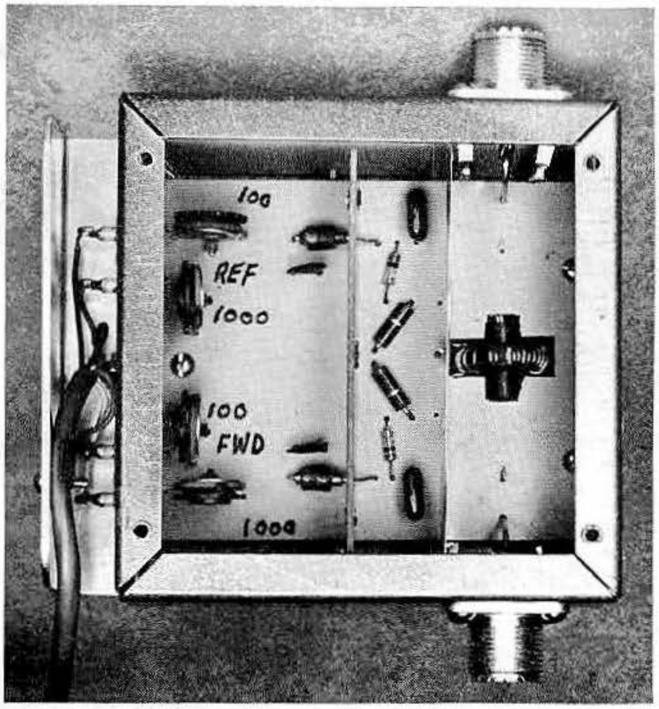

Top view of the rf head for the circuit of Fig. 38. A flashing-copper shield isolates the through-line and T1 from the rest of the circuit. The second shield (thicker) is not required and can be eliminated from the circuit. If a 2000-watt scale is desired, fixed-value resistors of approximately 22 kΩ can be connected in series with high-range printed-circuit controls. Or, the 25 kΩ controls shown here can be replaced by 50 kΩ units.

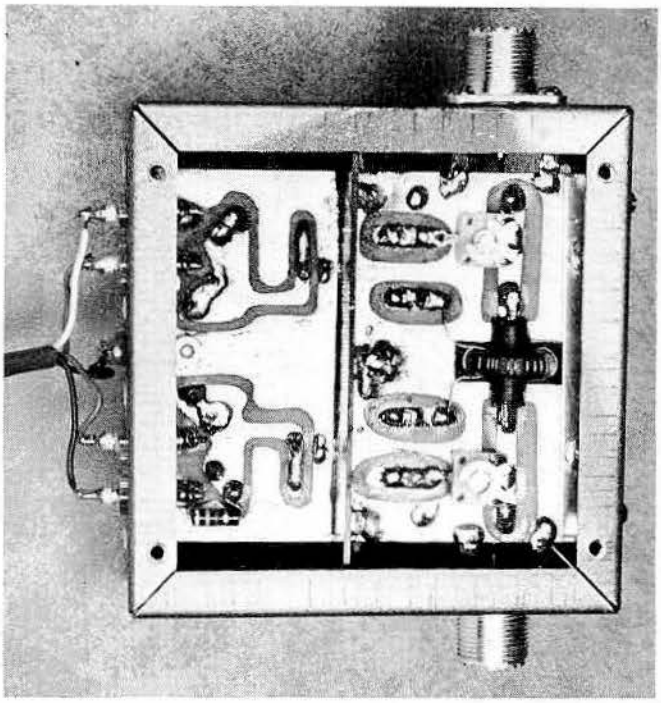

Bottom view of the rf head for the circuit of Fig. 38. The fixed-value resistor at the lower left does not belong in the circuit, but was added as a shunt for one of the calibrating controls which was too high in value — a 50 kΩ unit that was on hand. The shield partition shown here proved unnecessary and can be eliminated.

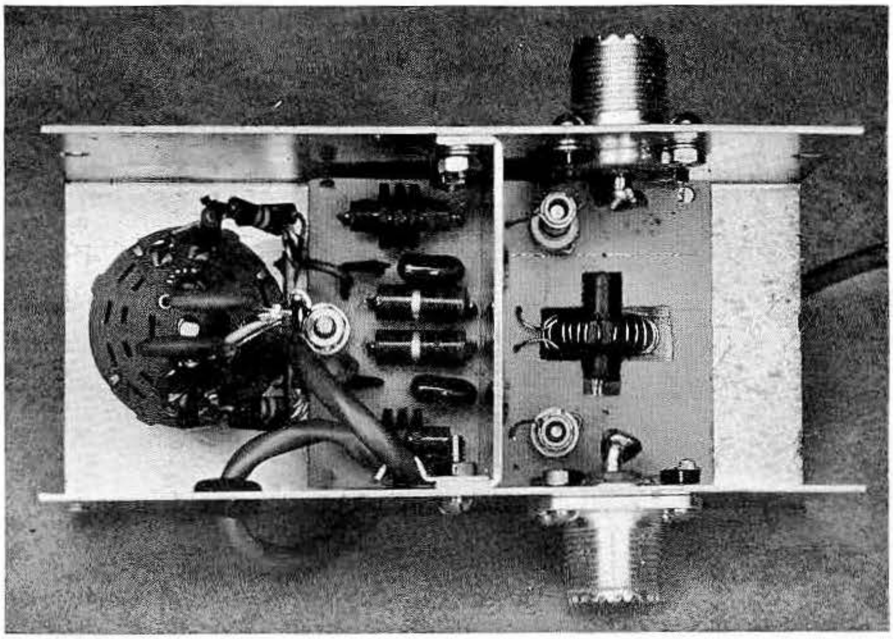

Fig. 39 — Inside view of the 5- and 50-watt power meter rf head. Component values are the same as in the circuit of Fig. 38, except for the calibrating resistances (see text). An aluminum shield isolates the through-line and toroid from the remainder of the circuit. A 50-μA meter is used in this model.

should be connected between J2 and the dummy load during the tests, providing power calibration points against which to plot the scale of M1. Apply transmitter output power to J1, gradually, until M1 begins to deflect upward. Increase transmitter power and adjust R4 so that a full-scale meter reading occurs when 100 watts is indicated on the rf ammeter or other standard in use. Next, switch S2 to REFLECTED and turn the transmitter off. Temporarily short across R3, turn the transmitter on, and gradually increase power until a meter reading is noted. With an insulated screwdriver adjust C2 for a null in the meter reading.

The next step is to reverse the coax connections to J1 and J2. Place S2 in the REFLECTED position and apply transmitter power until the meter reads full scale at 100 watts output. In this mode the REFLECTED position actually reads forward power because the bridge is reversed. Calibrating resistance R3 is set to obtain 100 watts full scale during this adjustment. Now, switch S2 to FORWARD and temporarily place a short across R4. Adjust C1 for a null reading on M1. Repeat the foregoing steps until no further improvement can be obtained. It will not be necessary to repeat the nulling adjustments on the 1000-watt range, but R5 and R6 will have to be adjusted to provide a full-scale meter reading at 1000 watts. If insufficient meter deflection is available for nulling adjustments on the 100-watt range, it may be necessary to adjust C1 and C2 at some power level higher than 100 watts. If the capacitors tune through a null, but the meter will not drop all the way to zero, chances are that some rf is leaking into the bridge circuit through stray coupling. If so, it may be necessary to experiment with the shielding of the through-line section of the rf head. If only a small residual reading is noted it will be of minor importance and can be ignored. In the circuit of Fig. 38 there remained approximately one-half a meter division when the null was reached, and this occurred only on the 1000-watt range. Since this was representative of less than 2 watts of power, it was deemed inconsequential.

With the component values given in Fig. 38 the meter readings track for both power ranges. That is, the 10-watt level on the

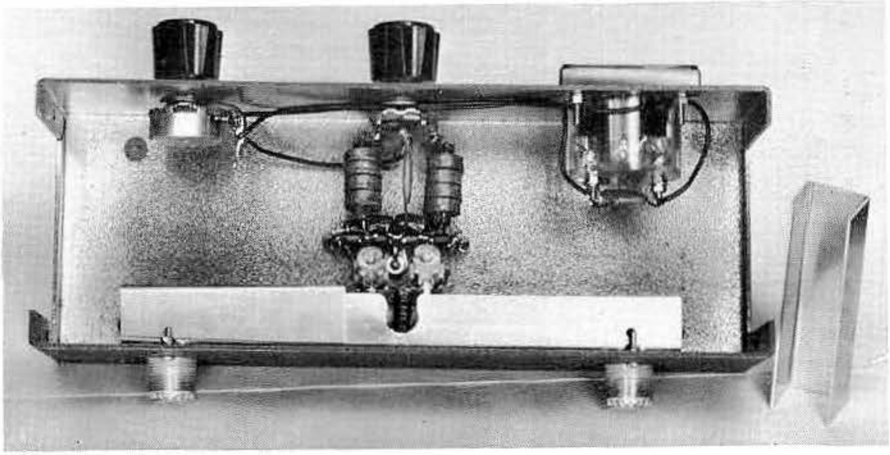

Inside view of the modified Heath HM-15 bridge. The new components are grouped at the center of the chassis on a 5-lug terminal strip. The nulling capacitors are connected between the inner line and the terminal strip. Press-fit aluminum shield covers are slipped over the trough line to aid in rf isolation. One cover is in place; the other is at the right of the photo. Improved shielding might be effected by installing an aluminum plate between the terminal strip and the two nulling capacitors.

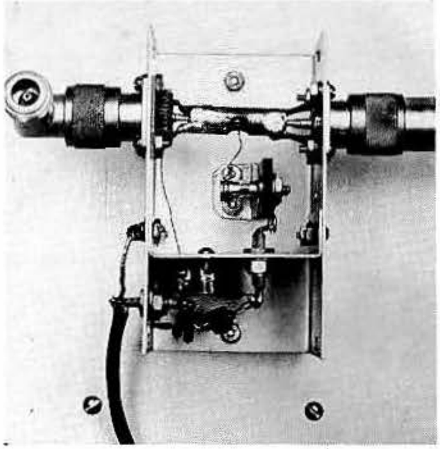

Inside view of a 2000-watt power meter built by N1RM. This bridge is patterned after the circuit of Fig. 37B. Point-to-point wiring is used throughout, thus avoiding the need for a circuit board. Two piston trimmers are used for the nulling capacitors and are mounted one above the other on a phenolic block. The two 500-pF feedthrough capacitors are part of the capacitive voltage dividers.

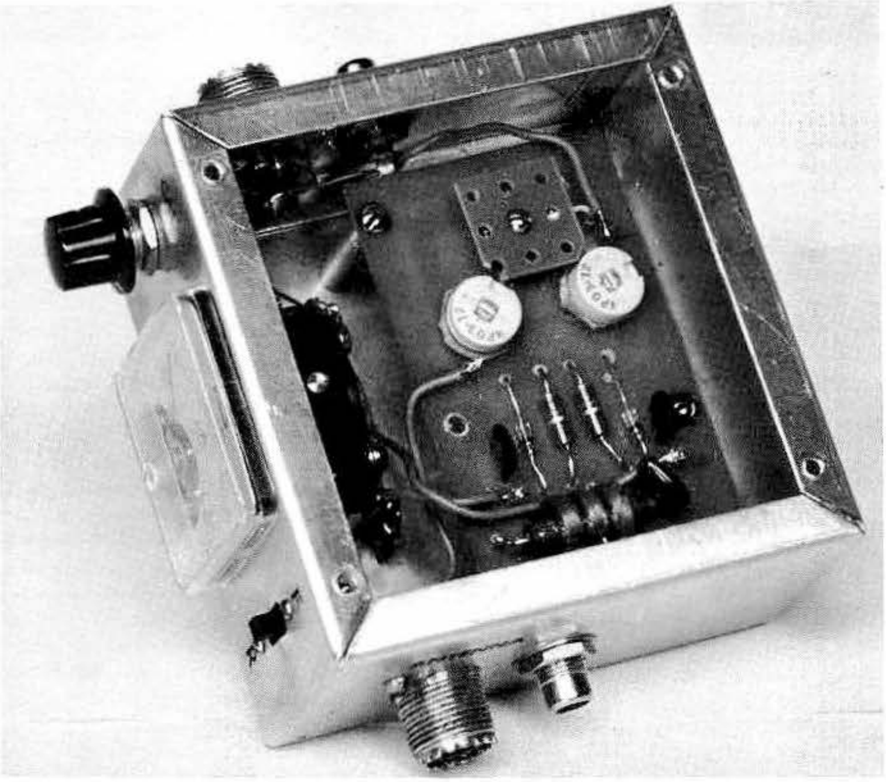

Fig. 40 — Inside view of a 3-watt power meter for QRP rigs. Its circuit is given on page 16 of June 1969 *QST*. Ceramic trimmers are used for nulling the bridge. Type SO-239 connectors are paralleled with phono jacks to add versatility. A 4 X 4 X 2 inch (102 X 102 X 51 mm) utility box houses the entire unit.

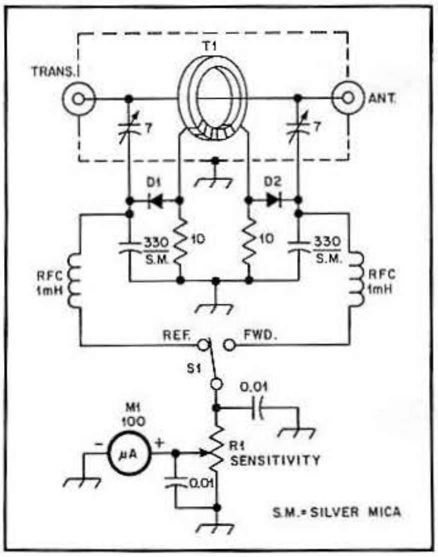

Fig. 41 — Schematic diagram of the modified Heath HM-15 SWR meter. D1, D2, M1, R1 and S1 are original components from the HM-15. T1 is the same as in Fig. 38. See text for additional details.

100-watt range, and the 100-watt point on the 1000-watt range fall at the same place on the meter scale, and so on. This no doubt results from the fact that the diodes are conducting in the most linear portion of their curve. Ordinarily, this desirable condition does not exist, making it necessary to plot separate scales for the different power ranges.

Tests indicate that the SWR caused by insertion of the power meter in the transmission line is negligible. It was checked at 28 MHz and no reflected-power could be noted on a Bird wattmeter. Similarly, the insertion loss was so low that it could not be measured with ordinary instruments.

Other Circuits

Additional circuits and photos are shown for variations in the basic design used at Fig. 38. A low-power model, having scales for 5 and 50 watts, is shown in Fig. 39. It uses fixed-value resistors for meter calibration. The required values of resistance were first determined by temporarily inserting a potentiometer in the meter line, obtaining the required full-scale reading, then substituting fixed-value resistances of the proper ohmage. The meter readings for the two power ranges do not track in this model.

A low-power meter was designed for use with the QRP transmitter described in June 1969 *QST*. It is shown in Fig. 40, and has a full-scale calibration of 3 watts. To obtain additional sensitivity, the primary of the toroidal transformer consists of a one-turn link instead of the single wire that would normally pass through the hole in the toroid core.

Some experiments were conducted to see if a Heath HM-15 SWR bridge (a Monimatch type) could be modified to work in a Bruene circuit. The results were satisfactory, and the circuit is given in Fig. 41. No attempt was made to obtain a calibration scale for the meter. The unit is being used as a simple SWR indicator, but now has better sensitivity in the lower part of the hf spectrum — 7 watts, full scale, from 3.5 to 30 MHz. Also, the instrument is no longer "frequency-conscious" as was the case before modification. The original pickup lines were discarded, the FWD-REV panel switch was rotated 180 degrees so that the labels were correct for the new circuit, and press-fit shield covers were installed on the trough line as shown in the photo. A power scale could be plotted by setting the sensitivity control in a fixed position — possibly replacing the existing control with a screwdriver-adjust type. A new 100-μA meter could be installed to provide a better scale for calibration in watts.

It was necessary to dismantle the trough line so that the toroidal transformer could be slipped over the inner line. A few wraps of mylar tape were wound over the center of the inner line to insulate the toroid winding from the line, and to provide a snug fit to keep the toroid in place. The trough was notched out with a nibbling tool to allow clearance for the toroidal transformer. Additional shielding can be added between the line and the rest of the circuit to further assure a zero meter reading in the reflected position.

It is hoped that the experimenter will find sufficient information here to enable him to build a power meter that will satisfy his specific needs. This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] "Monimatch Mark II," *QST,* February 1957.
[2] Bruene, "An Inside Picture of Directional Wattmeters," *QST,* April 1959.
[3] The etched-circuit board pattern and parts layout sheet for this power meter are available from ARRL hq. Send 25 cents and an s.a.s.e. Check with Circuit Board Specialists, P. O. Box 969, Pueblo, CO 81001 for negative and circuit boards.

A Resistive Antenna Bridge — Simplified

SWR indicators are neat, but when it comes to making measurements of resonant impedances, you need a true bridge type of instrument. Measure those low antenna resistances directly with this W2FMI circuit, and do it up to 30 MHz with ease!

As shown previously,[1] a bridge using a single variable capacitor can be used to measure resonant impedances of antennas and losses in coils with reasonable accuracy. With addition of an L/C network, the bridge can also be very useful in adjusting antennas for resonance conditions. Recently it was brought to the author's attention that a bridge using a variable resistance could also perform well. This is possible because the range of antenna impedances is generally low enough to allow for low values of resistance in the arms of the bridge. Under these conditions parasitic capacitances are negligible up to 30 MHz.

Resistive bridges have the advantages of small size, readily available components, and dial calibrations which can be made practically linear. Two bridges are shown with their respective calibration curves. The first is a linear device. The second expands the lower portion of the calibration curve, 0 to 30 ohms, to measure lower values of antenna impedance and coil losses more accurately.

Bridge Details

The resistive bridge with a variable resistance in one arm and equal resistances in the other two is shown in the photographs, and schematically in Fig. 42A. The nonlinear bridge is similar mechanically. The small enclosures and compact layout minimize lead lengths. With the linear version the dial practically reads directly in

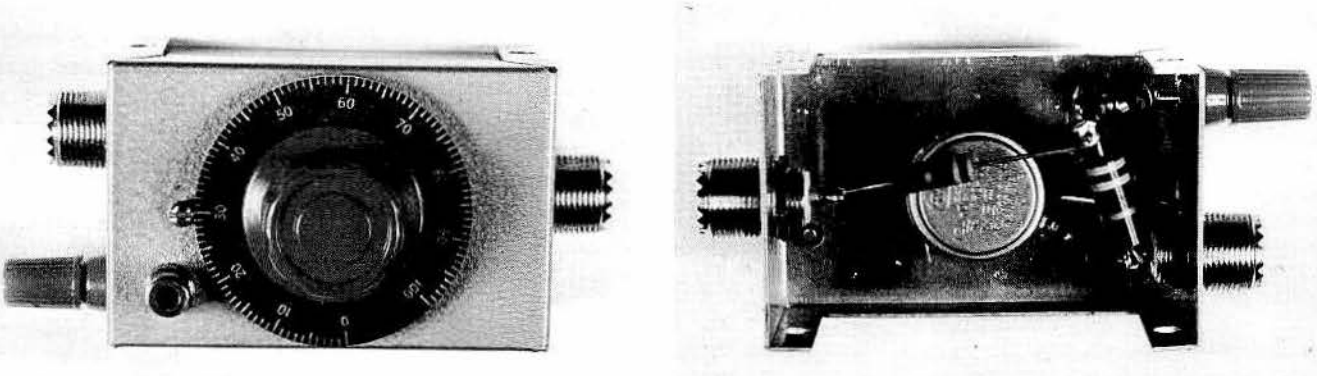

Exterior and interior view of the resistive bridge. Appearance and component arrangement are essentially the same for the two types described.

---

[1] Sevick, "Simple RF Bridges," April 1975 *QST,* page 11.

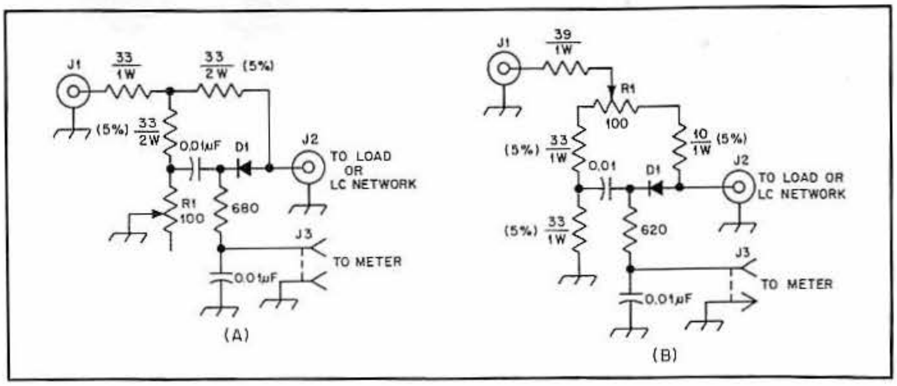

Fig. 42 — Schematic diagram and parts information for the linear (A) and nonlinear (B) resistance bridges. Resistor wattages are the minimum recommended.

D1 — Germanium diode.
J1, J2 — Coaxial connector, chassis type.
J3 — Phono jack.
R1 — 100-ohm, 2-watt linear control (Allen-Bradley, Type J).

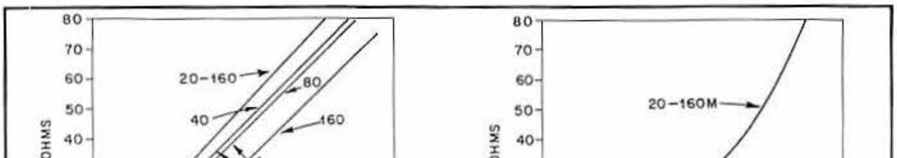

Fig. 43 — Calibration curves for the linear (A) and nonlinear (B) resistance bridges. The add-on L/C network used was described in the author's *QST* article on capacitive bridges.

ohms. In using the L/C network described in the previous article, a constant offset occurs, because of ohmic loss in the coils. This can be as high as 20 ohms, as is shown by the 160-meter curve with the L/C network, Fig. 43A.

The variable resistor used in the bridges has a parasitic (stray) capacitance of the order of 10 to 20 pF, which has little effect even at 14 MHz. Other potentiometers of similar design can probably be substituted. The calibration curves provide a good check. If the difference between the calibrations for the lowest and highest frequencies to be used is greater than an ohm or two, parasitic capacitances are probably excessive.

The nonlinear bridge uses most of the same components, but the control is connected in the differential mode, as shown in Fig. 42B. Resistance is subtracted from one arm and added to the other, tending to spread out the low end of the resistance scale for increased accuracy. In this version, the dial can be read easily to 0.5 ohm.

Results

The author has found these resistance bridges to be entirely satisfactory for measurements in the range of 1.8 to 30 MHz. Because of their greatly expanded scales, as compared with the capacitive bridges, and generally lower values of resistance in all arms of the bridge, they tend to have less marked nulls when used with very low-power signal sources. Increasing the minimum power to 20 or 30 milliwatts, when using the sensitive meter described previously, should take care of this.

In closing, the writer wishes to thank Roger A. Sykes, W1PP, for bringing the resistive bridge to the author's attention. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

Why a Transmatch?

Let's cut away some of the mumbo jumbo about antennas. Here's some basic information to help the newcomer make the decision that best suits his need.

These days, nearly all transmitters have a pi-network tank circuit in the final amplifier. There are good reasons for this. With the pi tank it is possible to design a band-switching transmitter with excellent shielding (in order to prevent undesired harmonic radiation as far as TVI is concerned) and to accomplish the job with a minimum of complications. Additionally, the tank circuit easily can be set up to work into 50- or 70-ohm loads, the characteristic impedance of the popular types of coaxial feed lines. However, as we will see, this is where a "clinker" can get into the act.

When a pi network is designed to work into a 50-ohm load, the load *must* be 50 ohms in order for the transmitter to work at full efficiency. Merely connecting 50-ohm coax to the tank circuit does *not* accomplish this. For the tank to "see" a 50-ohm load, the far end of the coax — the load or antenna end — must also have the same impedance as the coax, 50 ohms. If the load is not 50 ohms the pi network will see some other value than 50 ohms. Usually, too, the tank will see reactance as well as resistance when the line is not matched. While reactance is expressed in ohms, it isn't a true resistance, and what it does is make it more difficult to put power into the line unless certain steps are taken to compensate for it.

Pi-Network Tank Circuits

In the conventional transmitter pi tank we have two controls, customarily referred to as the "tuning" and "loading" controls. The tuning control is the one that resonates the tank circuit to the desired frequency. The loading control, when adjusted properly, permits the final amplifier stage to be loaded to the desired input. To a limited extent it can also be used to compensate for the reactance in the load.

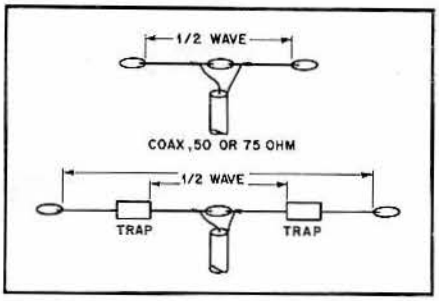

Fig. 44 — At the top is a simple half-wave dipole using coax feed. At the bottom is a drawing of a trap dipole. When a signal is fed to this antenna, for example an 80-meter one, the feed line "sees" a more-or-less resonant antenna on 80. When the transmitter is switched to 40, the traps "divorce" the 80-meter portion and the feed line sees a resonant 40-meter half-wave.

The clinker in this is that the reactance of many antenna-system loads may have too wide a range for the loading capacitor to handle. If you've ever been in the situation where is was impossible to load (or unload) the amplifier stage, the likely reason is that the load is too reactive to be handled by the tank circuit.

Such a condition can lead to some serious difficulties. While a correctly tuned pi network will offer good harmonic attenuation, an improperly adjusted tank may provide *no* harmonic attenuation. In fact, many Novices who receive second-harmonic violation notices for their 80-meter operation can lay the blame on poor loads and incorrectly adjusted tank circuits. This leads us to the real crux of the problem. How do we get a 50-ohm load on the various bands and frequencies desired?

Antenna Loads

As we said earlier, in order for a transmitter to see a 50-ohm load the far end of the 50-ohm coax must also see 50 ohms. The point where you feed the antenna has a certain impedance; in a half-wave dipole, for example, the impedance will be in the neighborhood of 50 ohms, depending somewhat on the dipole's height above ground. Also, at resonance there will be no reactance in the impedance. When we change frequency, the antenna is no longer resonant and reactance is introduced into the impedance. In turn, reactance is present at the transmitter end of the line. Whether or not the transmitter tank circuit can compensate for it will depend on the range of the loading capacitor.

How do we go about getting an antenna that will look like 50 ohms over a wide range of frequencies within a band — and, for that matter, on several bands? It isn't exactly easy.

For example, if we cut an 80-meter half-wave dipole for the center of the band, 3750 kHz, it is possible to obtain a fairly good match for 50-ohm coax at that frequency and approximately 50 kHz each side of 3750 kHz. However, at the band edges, 3500 and 4000 kHz, the impedance will be very reactive and the mismatch between the coax and the feed point may be as high as 10 to 1. It would be difficult, if not impossible, for our pi network to handle that type of mismatch. Actually, there is no "simple" antenna, by ham standards, that can be coax fed on 80 meters and cover the entire band with a reasonable match. The 40-, 20- and 15-meter bands are narrow enough so that the pi tank will handle a coax-fed dipole under normal circumstances. However, a coax-fed dipole is essentially a one-band system so such an antenna is not the answer if we want to work "all" bands and all frequencies with a single antenna.

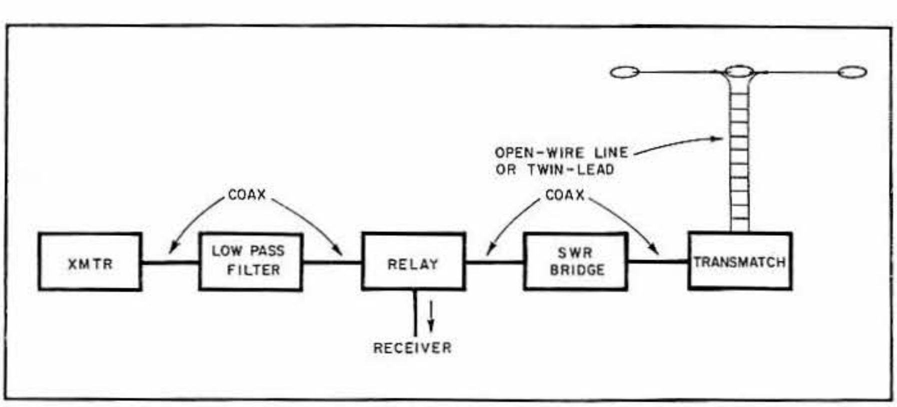

Fig. 45 — This drawing shows an ideal multiband antenna system. The units can be connected together by either 50- or 70-ohm coaxial line. The SWR bridge must match the line used. By putting the antenna relay in this section of the line, the Transmatch can be used to advantage for receiving as well as transmitting.

Trap Antennas

One type of multiband antenna with coax feed is the trap antenna, either horizontal or vertical. The horizontal variety is essentially a dipole with traps added to divorce parts of the antenna so the coax "sees" an electrically resonant dipole on the band in use. However, the trap dipole is no "broader" than a single dipole. Fig. 44 shows such a trap dipole antenna.

There are a couple of drawbacks to trap dipoles. First, a coax-fed single-band dipole is essentially a selective circuit, in that it tends to discourage undesired harmonics from being accepted and radiated. For example, a 40-meter harmonic would have a tough time in an 80-meter coax-fed dipole because the antenna and line are badly mismatched on 40. A trap system, however, does not discriminate against such undesired signals. Also, traps introduce power loss — probably not a great deal, but nevertheless loss. But if you want to use coax line and have multiband operation, the trap dipole is an answer. On the other hand, a Novice may ask, "Is there a better multiband system, one without the trap losses and harmonic problems?" The answer is yes, but before describing it let's take a look at feed lines.

Feed Lines

Three common types of feed lines are used by amateurs: coaxial, twin-lead, and open-wire. Each has a certain amount of loss, with coax the lossiest, twin-lead next, and open-wire line the most efficient. An important point to keep in mind is that the greater the mismatch, or standing-wave ratio, on the transmission line, the greater the line loss. Also, the loss in any line increases with frequency: The loss will be least at 80 meters and will become progressively greater as we approach vhf or uhf. As an example, the loss in 100 feet (30.48 m) of RG-58/U line at 144 MHz is about 6 decibels. Translating this to power, if we had a transmitter that was putting out 100 watts only about 25 watts would reach the antenna, at the end of a 100-foot (30.48 m) line, to be radiated. The remaining 75 watts would be dissipated as heat in the line! And, most important, this figure is based on the coax being *matched* at the antenna end. The mismatch or SWR will depend entirely on the impedance of the transmission line and the impedance of the antenna. If there is a mismatch, the losses increase. On 80 meters, this same coax has a loss of less than 1 dB for 100 feet (30.48 m) of line, so we could tolerate more of a mismatch.

But don't overlook the fact that if the line and antenna are mismatched it may be impossible to make the amplifier load, even if the extra loss can be tolerated. The question that then comes to mind is if we use a low-loss line so we can tolerate a high degree of mismatch, how can we make the transmitter see a 50-ohm load? The answer to this is to use a Transmatch between the transmitter and antenna.

The Transmatch

A Transmatch is simply an adjustable rf transformer that converts an unknown load (antenna side) to a desired load (transmitter side). It consists of inductances and capacitances which can be adjusted to provide a perfect 50-ohm load for the transmitter on any band or frequency.

Let's take an example to show how a Transmatch can do the job. The impedance at the center of an 80-meter dipole is

Approximate Attenuation in Decibels/100 ft. (30.48 m)

Type	Z_0	3.5	7.0	14.0	21.0	28.0	50.	144	220	420
RG-58/U	52.5	0.68	1.0	1.5	1.9	2.2	3.1	5.7	7.2	10.4
100W. Output		8	18	28	35	38	51	72	81	92
RG-8/U	52.	0.30	0.45	0.66	0.83	0.98	1.35	2.5	3.3	4.8
RG-59/U	73	0.64	0.90	1.3	1.6	1.8	2.4	4.2	5.2	7.2
RG-11/U	75	0.38	0.55	0.80	0.98	1.15	1.55	2.8	3.8	4.9
Twin-lead	300	0.18	0.28	0.41	0.52	0.60	0.85	1.55	1.9	2.8
Open-wire	—	0.03	0.05	0.07	0.08	0.1	0.13	0.25		

This chart shows the attenuation figures in decibels for common types of feed lines. Z_0 is the nominal impedance. In the case of open-wire line, the impedance will depend on the size of the conductors and their spacing. The second line of the chart indicates the amount of power that would be lost in a 100-foot (30.48 m) run of RG-58/U when properly matched. For example, on 21 MHz, the loss per 100 feet (30.48 m) is 1.9 dB. With 100 watts output from the transmitter, 1.9 dB represents a loss of 35 watts in the line. A mismatch will increase the line loss.

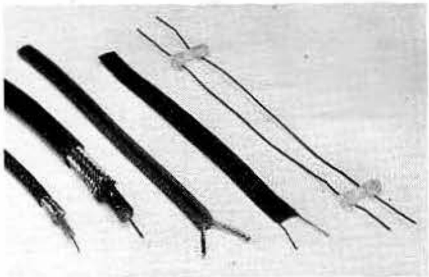

At the left are two examples of coaxial feed lines. The three lines at the right are balanced-type lines, 70-ohm, 300-ohm twin-lead, and TV open-wire line.

between 50 and 70 ohms on 80 meters. On 40 meters, the impedance at the same point on this same antenna is about 4000 ohms. If we feed this antenna with 300-ohm twin-lead the mismatch on 40 would be on the order of 13 or 14 to 1. However, bear in mind that 300-ohm line has relatively low loss, so we can probably tolerate the high SWR if the line is fairly short — less than 100 feet (30.48 m). Our only problem in making this antenna work on 40 is to take care of coupling the line to the transmitter. To do this the feed line is connected to the Transmatch and the Transmatch is then adjusted so that the transmitter sees a 50-ohm load. What we have now is an 80-meter dipole being used efficiently on 40, and our transmitter is working into the load it was designed for.

Now let's suppose that we feed this same antenna with 50-ohm coax. On 40, the mismatch would be 4000/50, or about 80 to 1! Because coax is not a low-loss line, the loss in the transmission line would be prohibitive. However, with twin-lead — or, even better, open-wire line — the loss is usually insignificant.

Also, and most important, a Transmatch adds selectivity at the output of the transmitter. It will discriminate against any undesired harmonics or spurious signals. And, if the antenna switching is set up so that the Transmatch is in the circuit on receiving, it will provide more selectivity for the receiver. In fact, a common problem for hams who live near broadcast stations is cross-modulation of 80- and 40-meter ham signals by the strong bc signal. A Transmatch will provide enough selectivity in most cases to eliminate bc cross modulation when it occurs in the receiver's front end.

Many hams are reluctant to use a Transmatch because it requires additional adjustments. However, the advantages are so great in a multiband system that any additional adjustments are more than worth the effort. With a Transmatch, the antenna doesn't have to be a half wave long to work on any band. For example, we described a complete multiband system just recently[1] using an antenna only 100 feet (30.48 m) long. Most hams can manage to get up a 60-foot (18.29-m) dipole, either horizontal or in an inverted-V configuration, and such an antenna also will work on 80 through 10. One method of putting up a multiband system is to find two supports for each end of the antenna, cut a wire long enough to run between them, put an insulator in the center of the antenna and at each end, and attach open-wire or twin-lead feeders long enough to reach the station. That's all there is to it. Bear in mind that a dipole fed this way doesn't have to be a half wave long; it can be any length, and it will work. Of course it is always a good idea to make it as long as possible and get it up as high as possible. Also, if possible, make it at least 1/4-wavelength long at its lowest operating frequency — about 60 feet (18.29 m) for 80 meters, for example.

A Few Notes

Many hams are reluctant to use open-wire line because of the problems in getting it into the shack from outdoors. One simple answer to this is to bring the line to the entrance — window sill, doorway, or whatever — and at that point attach 300-ohm twin-lead (which is insulated) and run the twin-lead in to the Transmatch. The fact that the two lines are not alike is of no concern because the load at the antenna side of the Transmatch is going to be some unknown value anyhow.

You may have read or heard that you must use coaxial lines in order to avoid TVI. Don't believe it. If you have a well-shielded transmitter and use a low-pass filter with it you won't have any harmonics coming out. Ask youself this simple question: "If there are harmonics, what difference can the kind of transmission line make?" The answer is, of course, "No difference."

One of the beauties in using a Transmatch is that the coaxial line between it and the rig can always be matched, and this matched section is where you put your low-pass filter. This eliminates any danger to the filter from excessive voltage or current caused by a mismatch.

One typical Novice question is "If my transmitter loads properly and I have no TVI problem, do I need more matching between the transmitter and antenna? I am using coax feed. What good would it do me to pay for the additional complexity of changing from band to band?" The answer to such a question depends on several factors. Is the harmonic attenuation — not TVI, but low-frequency harmonics — adequate? This could be a problem the Novice isn't aware of. As to the complexities of band changing, this will depend on how much the Novice desires to work other bands.

What we have tried to do in this article is to show the reader the problems and the recommended cure. There is nothing new about the multiband system using tuned feeders; it has been used for many years, and the interesting thing about it is that it is still the best multiband system. For further information it is recommended that the newcomer study *Understanding Amateur Radio* and *The ARRL Antenna Book* to gain a better insight as to how antennas work. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] McCoy, "A Complete Multiband Antenna System," *QST*, November 1967.

The Ultimate Transmatch

Most hams would like to own a Transmatch that could cover 80 through 10 meters, and work during coax-to-coax or balanced-lines conditions. Here is just such a unit. It will match the proverbial bedsprings, or any other nonresonant or resonant antenna — without plug-in coils or band switching — even on 160 meters.

Some amateurs assume that because they use coaxial feed lines they don't need a Transmatch. This can be an incorrect assumption.

The Harmonic Problem

A large number of hams have found to their sorrow that the harmonic suppression in their transmitters may leave much to be desired. The FCC issues many citations to Novices each year for second-harmonic violations.

A typical Novice station might consist of a transmitter and receiver, with direct coax feed to a trap dipole or vertical antenna. In a properly adjusted transmitter of typical design, often there is insufficient attenuation of the second-harmonic energy. This can mean that the second harmonic of the signal can reach the antenna and be radiated. What is required to eliminate or reduce the second harmonic, and higher-order harmonics, is more selectivity between the transmitter and the antenna. A Transmatch is an excellent aid in taking care of the harmonic problem.

The 50-Ohm-Load Problem

Still another reason for using a Transmatch is that nearly all manufacturers of transmitters design their rigs to work into a nonreactive, 50-ohm load. Any departure from this load impedance causes difficulty in tuning and loading the final-amplifier stage. It might be pertinent to point out that simply because a ham uses 50-ohm coax, it doesn't necessarily follow that he will have a 50-ohm load. Basically, in order to have a 50-ohm load when using 50-ohm coax, the coax must be terminated in a 50-ohm resistive impedance. In other words, the antenna impedance must be 50 ohms.

When the terminating impedance of the coaxial cable is other than 50 ohms, then the transmitter will be "looking" at some value other than 50 ohms. It is practically impossible to design a single antenna that will have an impedance of 50 ohms on all of the bands a ham might want to use. One answer to the problem is to employ a device that will "match" the unknown load to the 50-ohm impedance required by the transmitter. One such circuit or device is a Transmatch, simply an adjustable rf transformer that can match the unknown load presented by the transmitter end of the feed line to the required 50-ohm transmitter impedance.

The kW version of the Ultimate Transmatch. The knobs and dial counter are standard Millen components. The front bottom of the cabinet is tilted by means of two 2-inch (51 mm) standoff insulators used as feet.

Receiver Cross-Modulation

In addition to taking care of the problems just outlined, a Transmatch can serve another very useful function at the receiver. Some receivers are susceptible to what is known as cross-modulation, or front-end overloading, by strong nearby radio stations. This is particularly true on 160 or 80 meters when a broadcast station is nearby. The bc station overloads the front end of the communications receiver, causing "birdies" and other unwanted responses across the band. A Transmatch will eliminate or greatly reduce this problem in nearly all instances.

The Ultimate Transmatch

We shall describe two Transmatches — one for the Novice power limit, and another for the legal 2-kW PEP limit. We have found that many amateurs on the verge of taking their General, or higher-class licenses, would rather invest initially in a unit that will take care of all of their future needs.

Some years ago, a Transmatch was described that was designed for use with coaxial feed lines.[1] It was called "the 50-Ohm Transmatch." This unit handles mismatches on the order of 4:1. It can transform a load of about 200 ohms, or 12 ohms, to a 50-ohm impedance. The circuit of the Transmatch described in this article, Fig. 46, is similar, but with two notable exceptions. In place of a tapped, switched inductor, a variable inductor is used. (The idea for this innovation came from Al LaPlaca, K2DDK.) The variable inductor is a coil wound on an insulated form that can be rotated so that a movable wheel shorts out and grounds the unused portion of the coil. Using the roller inductor, along with the two variable capacitors, the Transmatch provides an almost unlimited matching range, plus the feature of 80- through 10-meter coverage without band switching. With the values shown in the units, it is even possible to add one fixed capacitor to obtain 160-meter coverage.

The other exception is the use of a 1:4 (unbalanced-to-balanced) balun transformer on the output side of the Transmatch, which permits the use of balanced feed lines. This means that a Transmatch can cover 160 through 10 meters (including the MARS frequencies). It can be used to match into coaxial lines, random-length end-fed wires, or balanced feeders, and can do the job without using switches.

The two units shown in the photographs underwent extensive tests as to efficiency, power-handling capabilities, and methods of adjustment. In coax-to-coax tests, power inputs and outputs through the Transmatch were carefully measured while using Bird Thruline wattmeters and an adjustable dummy load. As carefully as could be measured, and with indicated mismatches of up to 15:1, the Transmatches showed an efficiency of over 95 percent. This small loss in efficiency more than outweighs the usefulness of the device.

Many tests were made with the high-power version to check heating of the roller inductor and the balun. Under a key-down condition, with 1200 watts indicated power through the Transmatch, the transmitter was kept on for 15-minute periods. In one test, the measured rf current through the coil was in excess of 15 amperes. The roller inductor used in the unit is rated at 5 amperes, so this was considered a fairly rugged test. The coil did get warm, but not enough to be a cause for concern. Keep in mind that in regular amateur work, 15

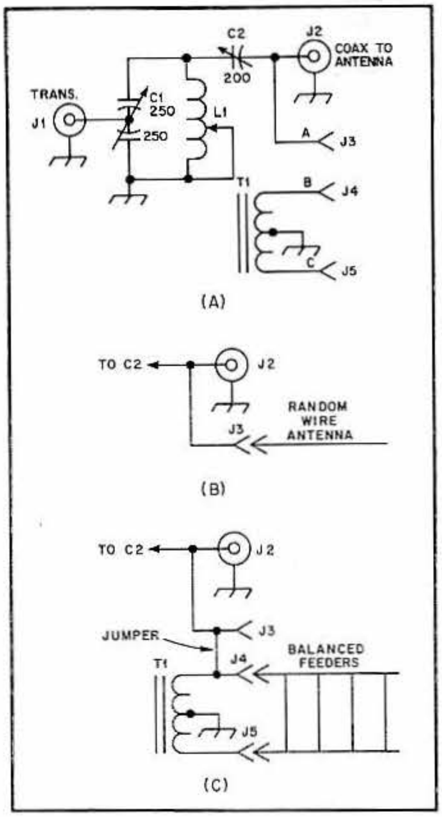

Fig. 46 — Circuit diagram of the Ultimate Transmatch.
C1 — Split-stator variable, 250 pF per section; see text. For low-power version, E. F. Johnson type 16250; or similar, for high-power version, Millen type 16250, or similar.
C2 — 200-pF variable, for low power, E. F. Johnson type 167-12 or similar, for high power, Millen type 16520 0.171-inch (0.43 cm) spacing, 16520A, 0.077-inch (2 mm) spacing, or similar.
J1, J2 — Coax chassis fitting, type SO-239.
J3, J4, J5 — Feed through insulators.
L1 — Roller inductor, see text. If 160-meter operation is desired, total inductance should be 28 μH, E. F. Johnson type 229-203; otherwise, 18 μH is adequate, E. F. Johnson type 229-202.
T1 — 1-to-4 balun; see text for details, cores are Amidon type T-200-2.

minutes would be a ridiculous figure for continuous operation.[2] In cw and ssb operation, the duty cycle would be in terms of seconds or even fractions of seconds.

The balun used for balanced operation in the high-power unit was made up from three Amidon 1-kW cores — more about that in a moment. N1RM has made extensive checks on a single-core balun. It was found that a single core would easily handle 500 watts of power during severe mismatch conditions. Failure of the core, most likely due to saturation and heating, took place at the 700-watt level. In the high-power Transmatch shown here, three cores are used. They are wrapped with glass insulating tape and then wound with Teflon covered wire. Wire insulation material can be important because of the extremely high rf voltages that may be developed with some load conditions. In many tests of long duration, using up to 1400 watts output into a dummy load, it was almost impossible to detect any heating of the balun.

Bear in mind that the balun would only be needed if balanced-line operation were desired. For coax-to-coax or random-wire feed, the balun can be omitted.

Construction Information

A look at Fig. 46, at A, will show that the circuit is a very simple one. The input line from the transmitter is connected to the rotor of C1, a split-stator capacitor. This means that the capacitor must be insulated from the chassis, using standoff pillars, and from the panel by means of an insulated shaft coupler. The same is true for C2. In the kW unit shown, we have installed a power bridge. Construction details are not given here for the bridge because such a unit was described in detail in *QST*[3]. Also, many amateurs already have a Monimatch[4] or some other type of SWR indicator. An SWR bridge or matching indicator is needed in the 50-ohm line between the rig and the Transmatch to show when the Transmatch is correctly adjusted.

Layout of the components should follow the same arrangement shown in the photographs — C1 at the input side, the roller inductor at the center, and then C2. If a balun is going to be installed, allow enough room behind the roller coil for mounting the balun. In the two units shown, the cabinets are made up from aluminum sheet stock. If desired, a commercial chassis and cabinet can be used. In the low-power Transmatch, leave at least 1/2-inch (13 mm) spacing between the edges of the capacitors and any part of the chassis and cabinet. Allow 1-inch (25 mm) spacing in the kW unit.

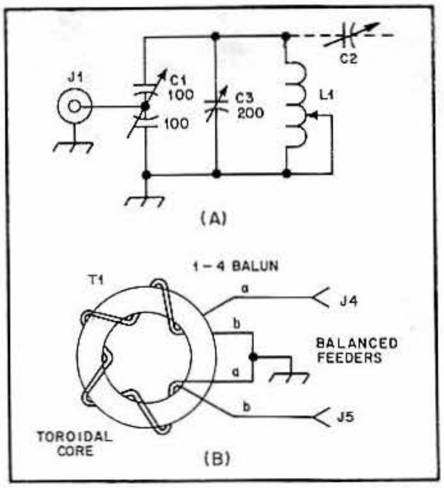

Fig. 47 — At A, details for padding a low-value variable for C1 by installing a second variable, C3, the same value as the low-power unit C2. In the case of tuning 160 meters, using an inductor of 28 μH for L1, and 250 pF per section for C1, a fixed capacitor of 100 pF can be used at C3, Centralab type 850-100N, or similar. At B, details for connecting the bifilar windings on the balun are given. The ungrounded leads a and b are connected to J3 and J4, the feedthrough insulator terminals.

Balun Details

For the low-power Transmatch, a single toroid core will more than handle the power. Incidentally, the low-power unit was tested under all conditions at about 100 watts through the Transmatch. The

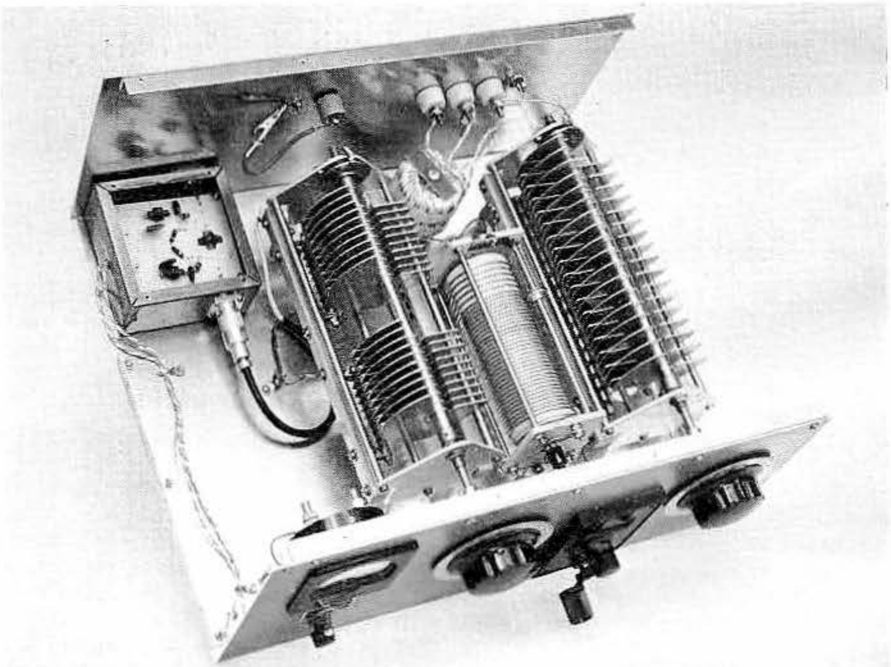

At the rear of the roller inductor is the 1:4 balun. The balun is set on top of a standoff insulator and is held in place with a piece of insulated board and a screw. The fixed capacitor mounted on the rear wall is the 160-meter unit. Both the variable capacitors and inductor are mounted on 1-inch (25 mm) high standoff insulators.

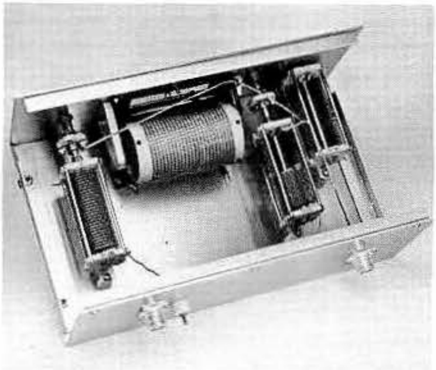

This is the low-power Transmatch. The salvaged roller inductor is mounted on the front panel, making use of the original mounting brackets.

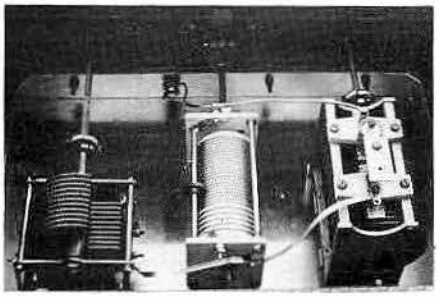

Here is an inside view of the Ultimate Transmatch as built by Al LaPlaca, K2DDK, who came up the idea of using the roller inductor. The installation on top the right-hand capacitor is the homemade 160-meter switch and fixed loading capacitor.

power-handling capabilities of the units depend primarily on the plate spacing used for C1 and C2, aside from the balun requirements mentioned earlier. At the 100-watt level, a plate spacing of 0.03 inch (0.8 mm) is adequate. A 0.045-inch (1 mm) spacing should handle 500 watts. In the kW unit shown, C1 has a 0.077-inch (2 mm) spacing. While a 0.077-inch spacing capacitor wasn't tried at C2, such plate spacing should work at the kW level.

In making the low-power balun, wind two layers of Scotch brand electrical tape around the core before putting on the wire. This will provide extra insulation, and will protect the edges of the wire to prevent the enamel from being scraped away. Fig. 47, at B, shows a sketch of the winding. The winding consists of 10 bifilar turns of no. 14 Formvar insulated wire. Connected as shown, this will provide a 1:4, unbalanced-to-balanced configuration.

Three cores are needed in the high-power balun. Each core is covered with two layers of 3-M type 27 glass-cloth insulating tape. Next, the three cores are stacked, then covered with two more layers of the tape. The winding consists of 15 bifilar turns of either no. 12 or 14 insulated wire. The insulation should have a minimum rating of 1000 volts and can be either Teflon or vinyl-nylon (or any other suitable covering with a minimum voltage rating of 1000 volts).

In making the low-power Transmatch, a roller coil from a junked 40-meter ARC-5 transmitter was salvaged and installed in the unit. Many amateurs know other hams who have old ARC-5 gear, and these units are a good source for the roller coils. However, some hams may not be able to locate used ARC-5 transmitters, so we have included some information on sources of surplus and new roller inductors. There may be surplus outlets in your area worth investigating. Look for a roller coil that has a minimum inductance of 10 μH when used with C1 having a value of 200 pF per section. Such a combination will cover a low frequency of 3500 kHz. In the unit shown, in addition to the ARC-5 roller, we used a value of 100 pF per section for C1 and padded that capacitor with another of 200 pF (single section) in order to provide the required coverage. This added capacitor is shown in Fig. 47 at A. By using this approach, it is sometimes easier to save money then when trying to locate a 200-pF-per-section variable for C1. (Incidentally, you can use a variable at C1 of more than 200 pF per section for good coverage with the roller inductance specified above.)

An even better approach to get the required capacitance for C1, because it is better to maintain the symmetry of the circuit as shown in Fig. 46 at A, would be to gang two 200-pF variables to make up your own split-stator capacitor, with the desired plate spacing. (There may be some around, but we couldn't find any manufactured split-stator with the required plate spacing and capacitance). Still another approach would be to clip fixed transmitting-type mica capacitors across each section of C1 to obtain the required capacitance for use on 160 and 80.

Finding the Parts

As most hams are finding out, it is becoming more and more difficult to buy, or find, components for construction projects. Some searching was done, and the following information is worth passing on. Barry Electronics of New York City has informed us that they will fill any order for any amateur component at the manufacturer's current amateur net price. The parts specified in Fig. 46 are currently manufactured items. Also, Fair Radio Sales in Lima, OH has some surplus roller inductors and counter dials in stock. The roller-inductor catalog number is 5950-199-6469, and the counter dial is listed as RT-45. Also, Millen components can be purchased directly from G. R. Whitehouse, Amherst, NH 03031.

Using the Transmatch

One of the first considerations in using a setup for "coax-to-coax" is exactly how much, or how high an SWR can be tolerated. As has been pointed out, this Transmatch will handle *any* mismatch on the line. Without getting into a lot of details, a few observations about coaxial lines are in order. First, the higher the SWR, the greater the losses in coaxial lines. The amount of losses depends on the line construction and the *frequency* at which the line is used — the lower the frequency, the less will be the losses. For example, RG-8/U has a loss per 100 feet (30.48 m) of line of only 0.33 -dB at 3.5 MHz with an SWR of 1 (a matched condition). At 28 MHz, this same line would have a loss of slightly less than 1 dB. However, as the SWR rises, so do the losses. With an SWR of 10:1, for example, the additional losses at 3.5 MHz would amount to about 0.6 dB, or an overall loss of about 1 dB. For the same SWR at 28 MHz, the losses would be about 2.5 dB, or a total loss of about 3.5 dB. Since a loss of 3 dB would mean half of the ouput power from the rig would be dissipated in 100 feet (30.48 m) of line, one could easily tolerate the condition at 80 meters, but the losses would be prohibitive at 28 MHz. This means that the amateur would be wise to keep the match at the antenna feed point as close to 50 ohms as possible, and thereby keep the SWR low.

With trap or multiple dipoles, the SWR should be less than 5 to 1; the Transmatch would be ideal. Keep in mind that you must provide a load for a transceiver or amplifier that is close to 50 ohms. The Transmatch will provide a means.

In a coax-to-coax system, set both C1 and C2 at maximum capacitance (plates fully meshed), then feed enough power through the Transmatch to get a full-scale forward reading on your SWR indicator. It is always important to use low power when adjusting a Transmatch. Next, switch the SWR bridge to read reflected power, then run the roller inductor slowly through its range. At some point you will notice a sharp dip in the reflected-power reading. Once this point is reached you should adjust C1 and C2, with possibly a slight adjustment of the roller coil, to get a perfect match. Power can then be increased to the desired level. You'll find that several different settings of the Transmatch may provide a perfect match. The one to shoot for is the condition that requires the most capacitance at C1 and C2 for a matched condition.

If a random-length, end-fed antenna is used, merely connect the end of the antenna to terminal J3, Fig. 46B, and go through the adjustment procedures outlined above. It is helpful to have an earth ground connection on the Transmatch. Also, if possible, use an antenna length that comes out to a quarter wavelength, or odd multiple thereof. Such an antenna will be of low impedance, and will reduce the chances of high rf voltages appearing on the Transmatch or associated equipment.

If balanced feeders are used, the feeders should be connected to terminals J4 and J5 with a short jumper connected between terminals J3 and J4. Again, the previously discussed tune-up and adjustment procedures should be followed in setting up the Transmatch.

Some Other Thoughts

Some amateurs might be concerned about the amount of voltage that can be developed in a mismatched coaxial line. In a matched condition, with 1000 watts in a 50-ohm line, the rf voltage would be about 220 volts. With a standing-wave ratio or 20:1 (!), the increase in voltage would be 4-1/2 times, or approximately 990 volts. This however, is well within the rating of the RG-8/U cable.

This Transmatch should satisfy the needs of those hams who desire a multiband unit without switches or plug-in coils. Once the settings for the controls are determined, a chart can be made up for any given istallation. Band changing is then quickly done. This material was originally presented in *QST* by Lewis G. McCoy, W1ICP.

References

[1] "The 50-Ohmer Transmatch," *QST*, July 1961.

[2] [Editor's Note: However, teletype operation would be continuous (to say nothing of those long-winded a-m operators!).]

[3] "In-Line RF Power Metering," *QST*, December 1969. (See chapter 4.)

[4] *Understanding Amateur Radio*, p. 209.

Build a Baby Ultimate

Got a match? Not a classic "match," but one between the antenna and your QRP rig may be what you need. If so, this article is for you.

Nobody needs or wants a 50-pound Transmatch during a QRP junket to a favorite lake, campsite or other distant location. The entire point of QRP operation is negated when the low-power station can be held on the palm of one's hand, but the accessory gear needs to be transported by hand truck to the car, airport or whatever! Operators who prefer to use a random-length wire antenna for field work will require a matching network between the 50-ohm transmitter output and the end of the antenna, whose characteristic impedance can be anything from a few ohms to a couple of thousand. For low-power work (25 watts or less) it doesn't take a big Transmatch to do the job, and some novel ideas can be applied to make the matcher cover 80 through 10 meters without a roller inductor or messy tapped-coil switching arrangement. This article shows some tricks that can be applied to meet that goal.

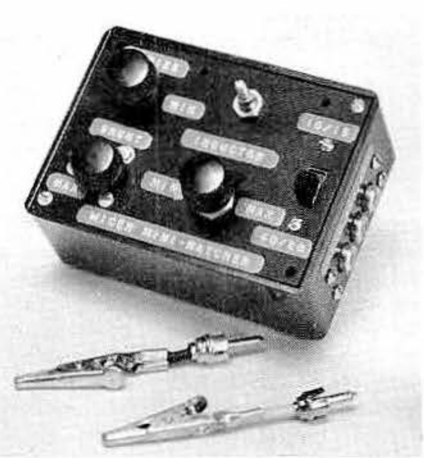
Exterior view of the Baby Ultimate. Dymo tape labels are used to identify the controls.

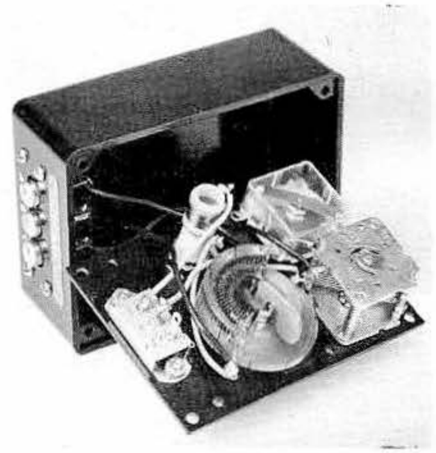
Interior view of the matcher showing the K1KLO variable inductor (lower center).

The heart of the Baby Transmatch is the half-toroid variable inductor. It is the by-product of a few "skull sessions" between Andy Pfieffer, K1KLO, and the writer. Both amateurs are QRP "freaks," and finding the proper mechanical technique for building a miniature panel-operated variable inductor of reasonable inductance range was challenging and productive. The machine-shop skill and impagination of K1KLO can be realized when viewing the accompanying close-up photograph. Early efforts along this line provided a slider type of full-toroid variable inductor (rheostat style), and dismal results were had: the inductor Q was very low, approximately 5, as a result of the shorted-turn syndrome. Finally, a powdered-iron Toroid core of Q1 material, 1-inch (25 mm) diameter, with 3/16 by 3/16-inch (5mm by 5mm) walls, was sawed in half and cemented to a plastic arm. The arm was glued to the shaft of the control so that it would enable the operator to change the coil inductance by moving the semicircular toroid section in and out of the coil.

The coil in this unit was air wound, then cemented to a semicircular piece of plastic which is affixed to a plastic disk of 1-1/2-inch (38 mm) diameter. The overall coil-and-plastic assembly is glued to the 3/8-inch (10 mm) bushing and 1/4-inch (6 mm) diameter control shaft. Although all of the metal work was done from scratch on a lathe, an old volume control or rheostat bushing and shaft can be modified to perform the function required. All that's needed here is some amateur ingenuity! The tuning range of the variable inductor is 3 to 9 μH, and the unloaded Q was measured as 150 at 7.9 MHz.

It is reasonable to conclude that the K1KLO "giz," as he calls it, could be made to cover a greater inductance range by making the coil bigger in ID and cementing both halves of the toroid core together — one atop the other — to provide greater permeability overall. Ferrite material is unsuitable for this component because of the hardness of ferrite. Without special cutting tools the latter will resist all efforts to saw through it. One-half of an Amidon 1-kW balun-kit, powdered-iron toroid should be excellent for larger units of this kind.

Other Circuit Features

As is shown in Fig. 48A, S1 is used to switch an additional variable inductor (L1) in parallel with the K1KLO unit, L2. This provides a lower minimum-inductance amount for operaion on 15 and 10 meters.

Fig. 48B illustrates some variations which will extend the Transmatch range down to 80 meters. The model shown photographically was designed for 40 through 10 meters. In the second example, S2 and L3 have been added to increase the total inductance to permit 80-meter matching. Also, C1 has been changed to a

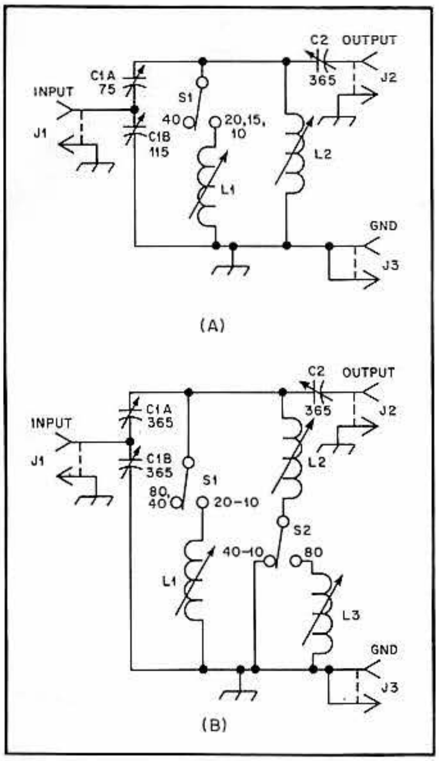

Fig. 48 — The diagram at A is for the 40- through 10-meter Transmatch. At B, a suggested circuit for coverage from 80 through 10 meters.

C1 — Dual-section air variable (Miller 2109, J. W. Miller Co., 19070 Reyes Ave., Compton, CA 90224). See text.
C2 — Calectro or Archer single-section miniature 365-pF variable.
J1-J3, incl. — Phono jack.
L1 — 3.1-to 4.8-μH slug-tuned inductor (Miller 4504 with red core).
L2 — See text. Contains 32 turns of no. 22 enam. wire, air wound, 7/16-inch (11 mm) OD.
L3 — 5.5-to 8.6-μH slug-tuned inductor (Miller 4505 with red core.)
S1, S2 — Spdt slide or toggle switch.

two-section 365-pF, bc-style variable, also for the same reason. A dual 365-pF capacitor can be used for the circuit at A, but the tuning rate will be much sharper than with the unit specified.

Three phono jacks are used for external connections. The first photograph shows a pair of clip leads which have been soldered to two phono plugs. These are used for making connections to a single-wire antenna and ground.

The slug screw of L1 protrudes from the top panel of the small plastic meter case. This feature enables the operator to have additional control of the overall inductance amount during operation on 15 or 10 meters. All internal ground connections are brought to a common point — J3, which has both of its terminals joined.

The "Ultimate" design is based on one which was popularized by W1ICP when he described in *QST* a 1-kW version of the circuit.[1] The network will match a wide range of impedances, but it should be pointed out that this circuit is not capable of providing much (if any) harmonic attenuation. Examination of the circuit will show that the network is basically a high-pass type.

Building the Transmatch

A 1-1/2 × 2-3/4 × 4-inch (38 mm × 70 mm × 102 mm) plastic meter style of case is used to house the Transmatch. The box used by the writer was made by RCA as an enclosure for a 9-volt regulated power-supply kit. Similar boxes are available at Radio Shack and other parts stores, but most have a metal cover. If the latter are used, it will be necessary for the builder to isolate the rotors of C1 and C2 from the metal plate. Adhesive-backed plastic feet are affixed to the bottom of the case. They will keep it from moving about on the operating table.

Using the Transmatch

When matching the 50-ohm transmitter to a given antenna (fed with coaxial line or single wire in nature), an SWR indicator will be necessary between the transmitter and the Transmatch.[2] C1, C2 and L2 are adjusted one at a time for the lowest reflected-power reading on the meter. There will be interaction between the controls, so it will be necessary to go over them several times to effect an SWR of 1.

Some difficulty may be encountered while working with a voltage-fed (high impedance) wire antenna. The symptom is one of hand capacitance affecting the SWR reading. In such instances the tune-and-try method will be required, moving the hands away from the box after each adjustment. For this reason the builder may wish to house the circuit in a metal box and use insulating shaft couplings on C1, C2 and L2. Of course, this will make the unit somewhat larger.

The writer used this Transmatch for two weeks in October 1975 while on a QRP sortie to Barbados (8P6EU). Maximum power used was 20 watts output, and minimum power was 2 watts ouput. The 40-, 20- and 15-meter bands were used, and all manner of antennas (end fed and coaxially fed) were employed. In all cases the SWR could be set for a 1:1 condition. Insertion loss through the Transmatch was measured at 14 MHz while delivering 20 watts to a test dummy load. The loss was approximately 0.1 dB at a matched condition. No heating of the components was noted.

One final word is in order. Always use the maximum possible amount of capacitance at C2 when adjusting the Transmatch for an SWR of 1. There can be several settings of C1, C2 and L2 which will provide a match. Best efficiency will always occur with maximum usable C at C2, and this is true of all variations of the Ultimate Transmatch.

If you're a QRP enthusiast, this matcher should be part of your setup when 50-ohm antennas aren't used. It is lightweight, small of size, and can be tucked away in the XYL's handbag when enroute, if need be! This material was originally presented in *QST* by Doug DeMaw, W1FB.

References

[1] McCoy, "The Ultimate Transmatch," *QST*, July 1970. (Included in this chapter.)

[2] DeMaw, "A QRP Man's RF Power Meter," *QST*, June 1973, p. 13.

Simple Broadband Matching Networks

This article gives complete details on constructing broadband matching transformers for ratios of 4:1, 6:1, 9:1 and 16:1.

Are you looking for a physically small DX antenna? To obtain low-angle radiation on 80 and 160 meters, many amateurs use vertical antennas. Horizontal antennas would have to be at elevations of a half-wavelength or greater to be comparable. Suitable supporting structures are beyond the reach of most amateurs at these lower-frequency bands. Quarter-wavelength verticals require some 60 feet (18.29 m) on 80 meters and 120 feet (36.58 m) on 160 meters — still very tall structures! However, short verticals over a good image plane have been shown theoretically[1] and experimentally[2] to perform practically as well as their full-size quarter-wavelength counterparts. The only problems are very low input impedances and narrower bandwidths.

This paper describes broadband matching networks which can efficiently match 50-ohm coaxial cable to verticals as short as 1/25 wavelength. Also included is information on short vertical antenna designs incorporating large top hats in order to improve bandwidth, such as a very short 40-meter vertical and a dual-band 80- and 160-meter short vertical with remote switching. In addition, loading coils using powdered-iron toroidal cores are discussed.

Broadband Matching Transformers

Broadband transformers have been widely used because of their inherent bandwidth ratios (as high as 20,000:1) from a few tens of kilohertz to over a thousand megahertz. This is possible because of the transmission-line nature of the windings. The interwinding capacitance is a component of the characteristic impedance and therefore, unlike the conventional transformer, forms no resonances which seriously limit the bandwidth.[3] At low frequencies, where interwinding capacitances can be neglected, these transformers are similar in operation to the conventional transformer. The main difference (and a very important one from a power standpoint) is that the windings tend to cancel out the induced flux in the core. Thus, high-permeability ferrite cores, which are not only highly nonlinear but also suffer serious damage even at flux levels as low as 200 to 500 gauss, can be used. This greatly extends the low-frequency range of performance. Since higher permeability also permits fewer turns at the lower frequencies, high-frequency performance is also improved since the upper cutoff is determined mainly from transmission-line considerations. At the high-frequency cutoff, the effect of the core is negligible.

Connie, W2FMI's better half, is making use of the shade under Jerry's 40-meter umbrella antenna. It is hoped that Jerry warns her about touching the antenna when he is running high power!

The applications of broadband transformers to antennas and broadband circuits have been covered by a number of authors.[4,5] This paper extends the application of these transformers to include the operation of short ground-mounted vertical antennas with impedances as low as 3 to 4 ohms. Further design considerations illustrate variable matches to vertical antennas having impedances even greater than 35 ohms.

Bifilar matching transformers for ground-mounted verticals lend themselves to unbalanced operation. That is, both input and output terminals can have a common ground connection. This eliminates the third magnetizing winding required in balanced-to-unbalanced (balun) operation. By adding third and fourth windings, as well as tapping windings at appropriate points, various combinations of broadband matching can be obtained. Fig. 49 shows a 4:1 unbalanced-to-balanced configuration. This transformer can be used with verticals of 12.5 or 200 ohms when using 50-ohm cable. No. 14 wire can be used and it will easily handle 1000 watts of power. By tapping at points 1/4, 1/2 and 3/4 of the way along the top winding, ratios of approximately 1.5:1, 2:1, and 3:1 can also be obtained. It should be noted that one of the wires was covered with vinyl electrical tape in order to prevent voltage breakdown between the windings. This was necessary when a step-up ratio was used at high power to match antennas with impedances greater than 50 ohms.

Fig. 50 shows a transformer with four windings, permitting wide-band matching ratios as high as 16:1. A picture shows a four-winding transformer with taps at 4:1, 6:1, 9:1 and 16:1. In tracing the current flow in the windings when using the 16:1 tap, one sees that the top three windings carry the same current. The bottom winding, in order to maintain the proper potentials, sustains a current three times greater. The bottom current cancels out the core flux due to the other three windings. If this transformer is used to match into low impedances, like 3 to 4 ohms, the current in the bottom winding can be as high as 15 amperes if the high side of the transformer is fed with 50-ohm cable handling a kilowatt of power. If one needs a 16:1 match like this at high power, then cascading two 4:1 transformers is recommended. In this case, the transformer at the lowest impedance side only requires each winding to handle 7.5 A. Thus even no. 14 wire would suffice in this application.

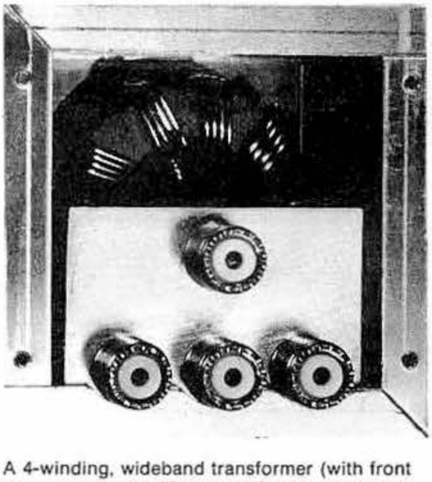

A 4-winding, wideband transformer (with front cover removed) with connections made for matching ratios of 4:1, 6:1, 9:1, and 16:1. The 6:1 ratio is the top coaxial connector and, from left to right, 16:1, 9:1, and 4:1 are the others. There are 10 (quadrifilar) turns of no. 14 enameled wire on a Q1, 2.5 inch (64 mm) ferrite core.

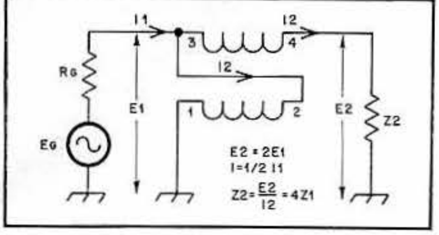

Fig. 49 — Four-to-one broadband bifilar transformer. Upper winding can be tapped at appropriate points to obtain other ratios such as 1.5:1, 2:1 and 3:1.

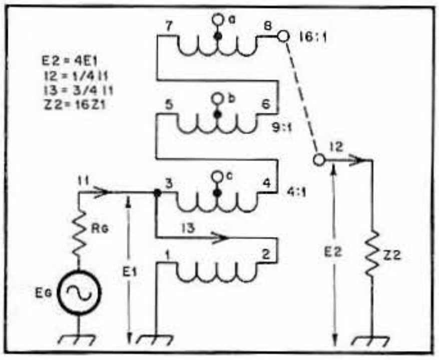

Fig. 50 — Four-winding, broadband, variable impedance transformer. Connections a, b and c can be placed at appropriate points to yield various ratios from 1.5:1 to 16:1.

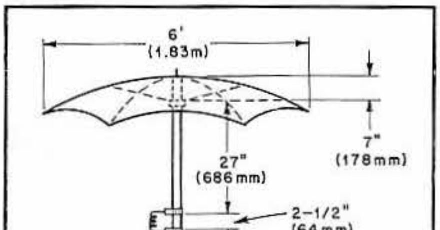

Fig. 51 — Final design and bandwidth characteristic of a short 40-meter vertical having an input impedance of 3.5 ohms and utilizing a commercial beach umbrella. This antenna established its worth as a prototype for the lower frequencies.

The popular cores used in these applications are 2.5 inch (64 mm) OD ferrites[6] of Q1 and Q2 material and powdered iron cores,[7] of 2-inch (51 mm) OD. The permeabilities of these cores, μ, are nominally 125, 40 and 10, respectively. Powdered iron cores of permeabilities 8 and 25 are also available.

In all cases these cores can be made to operate over the 10- to 160-meter bands with full power capability and very low loss. The main difference in their design is that lower permeability cores require more turns at the lower frequencies. For example, Q1 material required 10 turns to cover the 160-meter band, Q2 required 12 turns, and powdered iron (μ = 10) required 14 turns. Since the more common powdered-iron core is generally smaller in diameter and requires more turns because of lower permeability, higher ratios are sometimes difficult to obtain because of physical limitations. Although most of the author's experimentation on these transformers was done on the 40-, 80- and 160-meter bands, the results indicated that they should operate on the higher bands equally well. When working with low impedance levels, unwanted parasitic inductances come into play, particularly on 14 MHz and above. In this case lead lengths should be kept to a minimum.

All the results obtained by the author have shown that the transformers can be used to match a variety of loads and impedance ratios, they can be used over many bands, and that they exhibit negligible loss in operation. More importantly, the core materials are readily available and low in cost.

A Short Vertical Antenna Design

These transformers permit efficient matching to impedances as low as 3 to 4 ohms. This means, with some top-hat loading, antennas could be 4 to 5 feet (1.22 to 1.52 m) tall on 40, 8 to 10 feet (2.44 to 3.05 m) on 80, and 16 to 20 feet (4.88 to 6.10 m) on 160 meters. With the capability of tapping off any portion of the transformer windings, thus providing variable matching, the heights of these antennas can be increased at will depending upon the bandwidth desired.

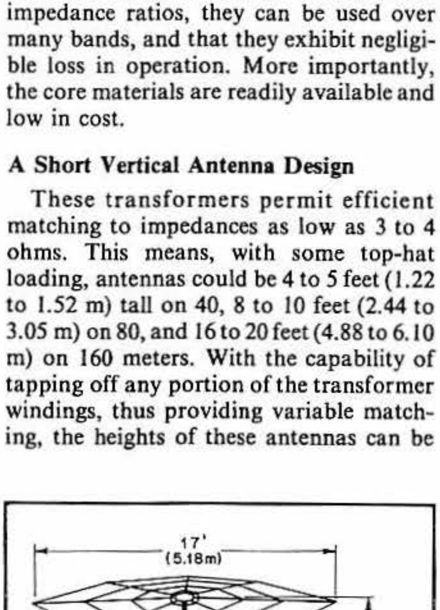

Fig. 52 — The 80- and 160-meter dual-band vertical system. The remotely switched matching network is described in the text. Not shown are three sets of polypropylene guy ropes at heights of about 9, 18 and 26 feet (2.74, 5.49 and 7.92 m).

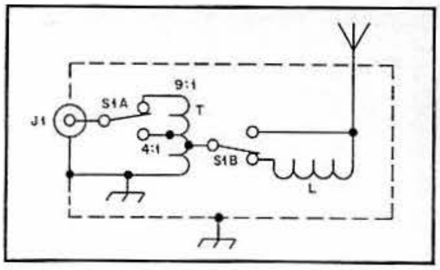

Fig. 53 — The remotely switched matching network for the 80- and 160-meter dual-band vertical antenna. The indicated switch positions are for 160-meter operation.

S1 — Dpdt antenna relay.

J1 — Coaxial connector, chassis type.

L — Load coil for 160 meters, 22-1/2 turns, no. 14 wire on 4 cores of 2-inch (51 mm) OD powdered-iron, permeability of 10 (Amidon T-200-2).

T — Broadband transformer (see text).

The first short vertical designed by the author was brought about by a trip to a local shopping center. The author saw a beach unbrella that had about the right proportions one would use in designing a short 40-meter vertical. The "antenna" was purchased (at a nominal price) and a 14-turn coil, connected across an insulated section about three feet (0.91 m) from the ground, was used to tune the vertical to resonance. The measured input impedance was 3.5 ohms and was matched easily to 50-Ω line with a 16:1 transformer. Fig. 51 shows the final design and the bandwidth characteristic. Several months of operation on 40 meters verified the outstanding capability of this short vertical antenna (with an extensive radial system) and established its worth as a prototype for the lower frequencies. A further study showed that the 6-foot (1.83 m) diameter top-section was about the equivalent of a 4-1/2-foot (1.37 m) diameter solid disk.

A Dual-Band 160- and 80-Meter System

The antenna shown in Fig. 52 is a dual-band design using a remotely switched matching network (see photograph) which not only changed a tap on a broadband transformer from 4:1 to 9:1, but also introduced a powdered-iron base-loading coil for 160-meter operation.

The circuit diagram for the remote matching network is shown in Fig. 53. The loading coil was made with four powdered-iron (μ = 10, 2-inch [51 mm] OD) cores. They were first wrapped individually with several layers of vinyl electrical tape before being combined and wrapped again. The necessary inductance was obtained with 22-1/2 turns. A slight adjustment of the loading coil and the height of the antenna was found necessary in order to put the minimum SWR at 1812 kHz. In general, a 1-foot (0.31 m) change in height of the antenna changes the 80-meter resonance by 60 kHz and the 160-meter resonance by 15 kHz. A change of 1/2 turn on the

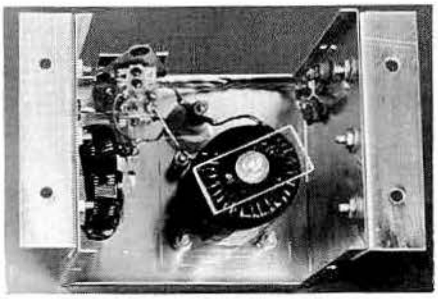

Bottom view of the matching network for the dual-band 80- and 160-meter vertical. The loading coil is on the right and the trifilar 4:1 and 9:1 matching transformer on the left. The dpdt 24-volt antenna relay switches in or out the loading coil and changes the taps on the transformer to the appropriate ratio. The enclosure is the conventional 4 X 5 X 6-inch (102 mm X 127 mm X 152 mm) aluminum Minibox.

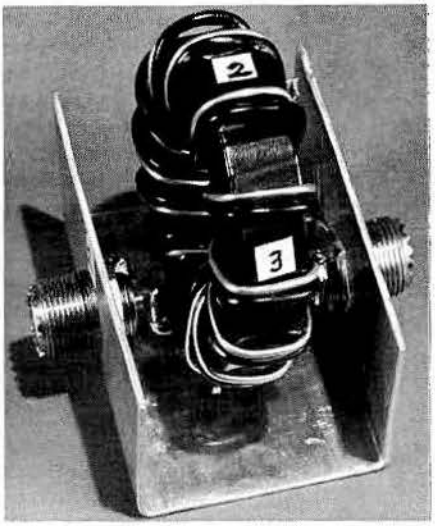

The broadband bifilar transformer shown has 12 bifilar turns, no. 14 wire, on a ferrite (Q1 material, 2.5 inch [64 mm] OD) core. Taps are also placed at 3, 6, and 9 turns, thus yielding ratios of 1.5:1, 2:1, 3:1 and 4:1. This transformer has a connection at the 1.5:1 ratio tap. Extra insulation was provided for the enameled wire in order to prevent voltage breakdown when matching into loads greater than 50 ohms.

powdered-iron core coil changes the 160-meter resonance by about 50 kHz. It should be noted that the windings had extra insulation (about one layer of vinyl tape) added on the inside part of the windings in order to prevent breakdown between them.

Low-permeability ferrite material Q2 (μ = 40, but still four times larger than the powdered iron material) was also investigated for use as a core for a 160-meter loading inductance. It was hoped that with several stacked cores of this material, the flux density would be reduced to a level where nonlinearities and permanent damage could be avoided. Two cores and about 10 turns of wire were used. The result was negative. Large changes in inductance were noted on voice peaks by the sudden large changes in the SWR. Also, small but permanent changes in the SWR were observed, indicating permanent damage. In contrast to ferrite, powdered iron maintains its linearity at high flux densities and suffers no permanent damage. As noted above, four cores of permeability 10, and 22-1/2 turns, provided the proper inductance (approx. 30 μH). The loss in this coil was about equal to the air-wound coil (33 turns, 6 tpi, no. 12 wire, 2-1/2-inch [64 mm] OD). Two stacked powdered-iron cores of permeability 25 were also investigated. This arrangement required about 17 turns, but was rejected because the loss, which was mainly in the two cores, went up to 3 ohms. This is certainly intolerable since it is of the order of magnitude of the radiation resistance of the very short vertical.

Concluding Remarks

The predominant characteristic of the short antenna is probably its very low value of radiation resistance. One must be careful of excessive ground and ohmic losses in coils and matching networks. A good image plane can eliminate the ground-loss factor. The application of a top hat eases the problem somewhat because it increases the effective length and thus the radiation resistance. The radiation resistance can be increased five-fold[8] by resorting to complete top-hat loading of a vertical. Combinations of top-hat and inductive loading can be used depending upon the radiator height and the bandwidth desired. One can also resort to a short folded unipole. With these combinations at his disposal, the designer is left with the problem of efficiently matching his transmission line to various impedance levels. The wide-band transmission-line transformer can provide the answer. It is capable of operating very efficiently over a wide range of impedances and impedance ratios.

The author would like to acknowledge the generosity of the following companies who supplied cores and information for the experiments: Indiana General Corp., Amidon Associates, and the Arnold Engineering Co.

Finally, the author would like to thank his colleague Mort Fagen, WA2EIT, for his encouragement and discussions of the antenna papers the author has written. This material was originally presented in *QST* by Jerry Sevick, W2FMI.

References

[1] King, *The Theory of Linear Antennas*, Harvard University Press, Cambridge, MA, 1956.

[2] Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, March 1973. (See chapter 1.)

[3] Ruthroff, "Some Broad-Band Transformers," *Proc. IRE*, Vol. 47, p. 137, August 1959.

[4] Turrin, "Application of Broadband Balun Transformers," *QST*, April 1969.

[5] Krauss and Allen, "Designing Toroidal Transformers to Optimize Wideband Performance," *Electronics*, August 16, 1973.

[6] Indiana General Corp., Electronics Div., Keasby, NJ 08832.

[7] Amidon Associates, 12033 Otsego Street, North Hollywood, CA 91607; The Arnold Engineering Co., Marengo, IL.

[8] See reference 2.

An RF-Sensed Antenna Changeover Relay

Flick of the switch, sir? Not with this system — it's automatic!

One subject that can be confusing for the newcomer to Amateur Radio is that of antenna-changeover systems. Several systems have been devised over the years and most can be classified into three or four main categories. Before delving into a discussion on these systems, a brief look at the simplest method of all might be in order, that is, no antenna changeover at all. With this setup two separate antennas are used, one for the transmitter and one for the receiver. This system is impractical for the most part since duplicate antennas would be required for each band of operation. There are instances, however, when separate antennas with different directive patterns can be used to an advantage, as in the case of specialized receiving antennas. Generally speaking, most stations use the same antenna for transmit and receiver purposes, so this is more the exception than the rule.

Several antenna-changeover systems, designed to transfer a single antenna from the receiver to the transmitter, are shown in Fig. 54. At A, a manual switch is used to transfer the antenna connection. This scheme is simple and works well, but is not without several drawbacks. Good-quality, high-power rf switches are often costly and

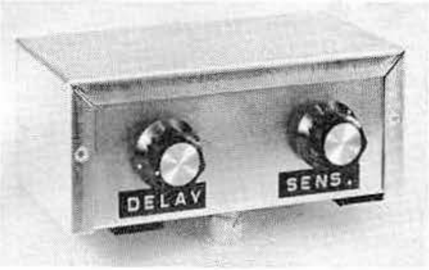

hard to locate and, of course, the operator is required to switch manually from receive to transmit.

One of these problems was eliminated with the advent of the coaxial relay. These relays were, in most cases, mounted external to the transmitter and receiver with connecting cables running to and from the unit, as shown in Fig. 54B. Many of these relays were rated for a kilowatt a-m signal, and maintained the 50-ohm line impedance. The voltage needed for the relay field was switched on or off with a low-powered toggle or foot switch. Alternatively, the transmit-receive switch might be located on the front panel of the transmitter with control voltage or a normally open or closed circuit available at a socket on the rear apron. Several extra sets of relay contacts were usually furnished with the relay to perform various functions, including that of muting the station receiver during transmit periods. Although this system of antenna transfer represented an improvement over the previous setup, it was still necessary for the operator to switch the antenna manually via the relay-control switch.

It's been said that laziness is the mother of invention and so it seemed natural that a more automatic system be developed. The relatively modern scheme at Fig. 54C has the transfer relay built into the transmitter. This system is quite popular and is used by most commercial manufacturers of amateur equipment that offer a separate transmitter and receiver. When the key is pressed the relay activates automatically and switches the antenna to the transmitter. A variable-length delay is usually included to hold the relay closed, thereby keeping the antenna connected to the transmitter during code characters. After the last group of code characters is sent, the relay will remain closed for a second or two — the amount of time controllable. When the relay deactivates, the antenna is reconnected to the receiver.

Unfortunately, some amateurs who build their own equipment or purchase transmitters and receivers of different manufacture, sometimes resort to switching methods A or B. These setups, of course, do not include the automatic features of system C. The circuit described here and shown in simplified form in Fig. 54D was designed with this in mind. Basically, this system detects the presence of rf at the output of the transmitter and immediately breaks the antenna connection to the receiver. At the same time the antenna connection to the transmitter is made. One of the features of this system is that no modifications to either the transmitter or receiver are required.

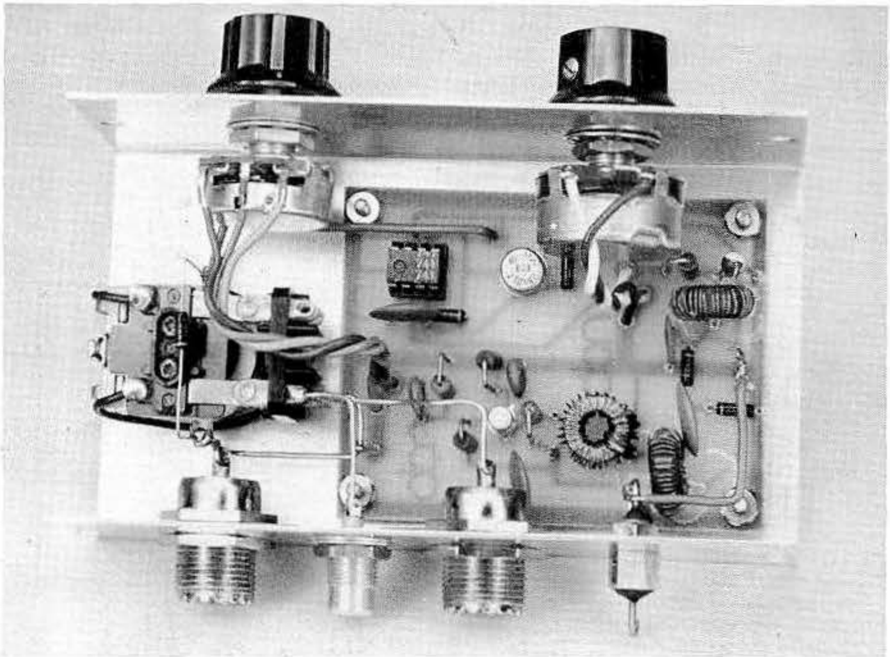

The inner layout of the antenna-changeover relay. Wiring or placement of circuit board and other components is not critical.

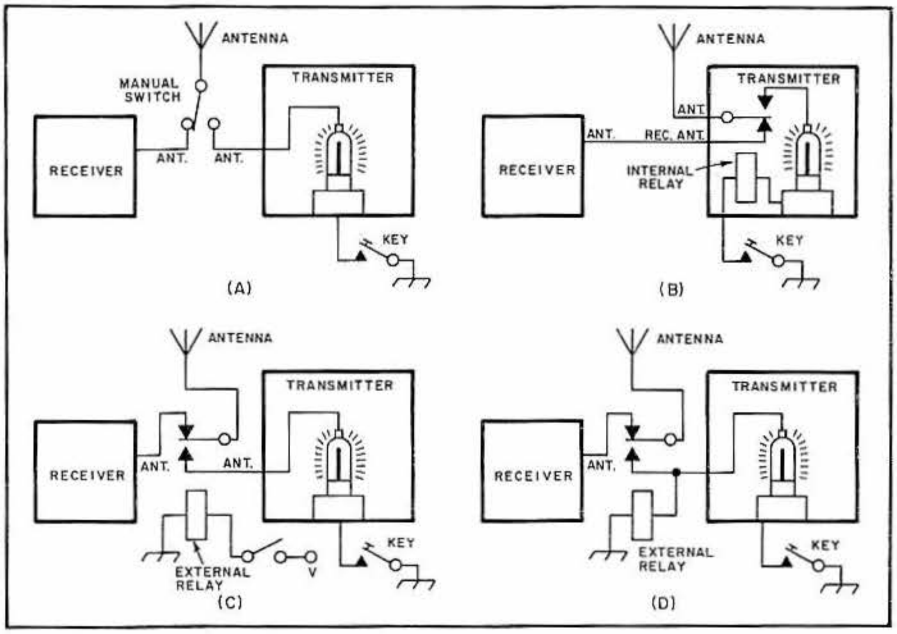

Fig. 54 — Several different antenna-changeover systems in block diagram form.

Circuit Description

R1 of Fig. 55 serves as an rf-voltage divider to permit the circuit to be used with transmitters of various power-output amounts. Rf energy is routed through C1 to the base of broadband amplifier Q1. The amplified hf-band energy is supplied to a voltage-doubler (D1 and D2) through a broadband toroidal step-down transformer, T1. The rectified rf voltage at the output of D1 and D2 is filtered by means of RFC2, C5 and C6. This prevents unwanted rf from reaching U1 and affecting its performance.

C6, R7 and R6 comprise a timing network (variable) which governs the hold-in time of the relay, K1. The smaller the resistance amount at R6, the shorter will be the time delay.

U1 functions as an inverting amplifier. When the input dc voltage at pin 2 increases, the output dc voltage at pin 6 decreases. The output voltage causes the base of relay driver Q2 to be forward biased negatively when it drops below approximately 1.4 volts. Diodes D5 and D6, by virtue of their combined barrier voltages (0.7 V each), establish the 1.4-V fixed bias level. Without the diodes, Q2 would conduct sufficiently to prevent the relay from dropping out during no-signal periods. D4 is used to suppress transients caused by the field coil of K1. When no rectified rf reaches U1, Q2 is cut off because of the high positive base voltage it receives from U1, and the relay contacts to the transmitter are open.

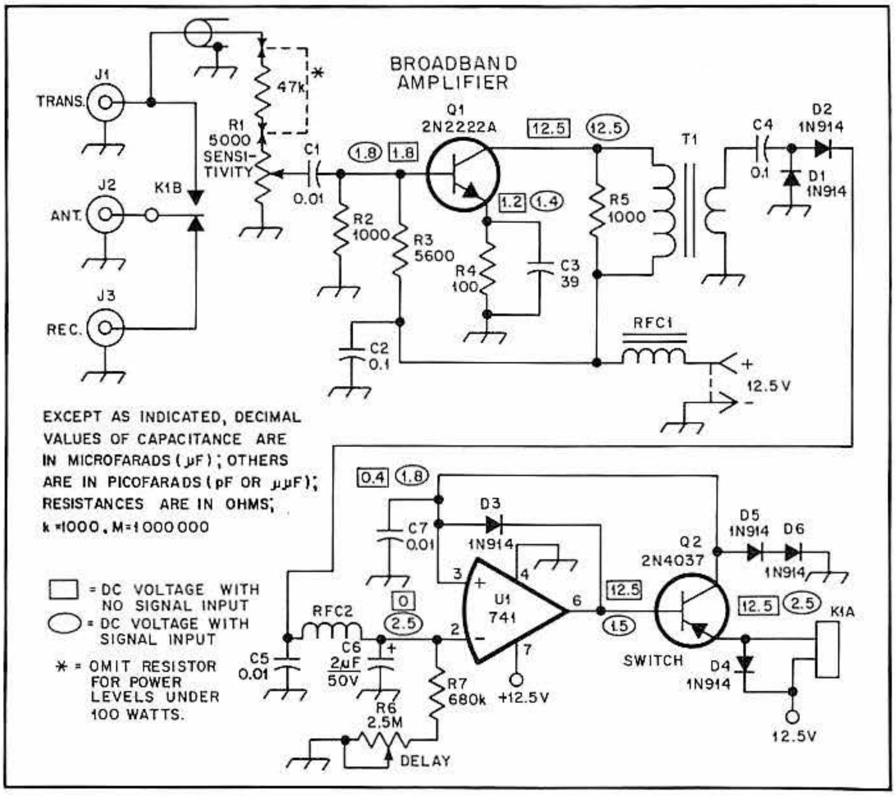
J1
TRANS.
J2
ANT.
K1B
J3
REC.
47k
*
R1
5000
SENSI-
TIVITY
C1
0.01
R2
1000
R3
5600
R4
100
C3
39
C2
0.1
BROADBAND
AMPLIFIER
Q1
2N2222A
12.5
12.5
1.8
1.8
1.2
1.4
R5
1000
T1
C4
0.1
D2
1N914
D1
1N914
RFC1
+
12.5V
-
0.4
1.8
C7
0.01
D3
1N914
3
4
U1
741
6
2
7
+12.5V
12.5
1.5
Q2
2N4037
SWITCH
D5
1N914
D6
1N914
12.5
2.5
K1A
D4
1N914
12.5V
RFC2
0
2.5
C5
0.01
C6
2μF
50V
R7
680k
R6
2.5M
DELAY

EXCEPT AS INDICATED, DECIMAL VALUES OF CAPACITANCE ARE IN MICROFARADS (μF) ; OTHERS ARE IN PICOFARADS (pF OR μμF); RESISTANCES ARE IN OHMS; k =1000, M=1 000 000

□ = DC VOLTAGE WITH NO SIGNAL INPUT

◯ = DC VOLTAGE WITH SIGNAL INPUT

* = OMIT RESISTOR FOR POWER LEVELS UNDER 100 WATTS.

Fig. 55 — Schematic diagram of the rf-sensed antenna-changeover relay. All resistors are half-watt composition type. Capacitors are disk-ceramic unless noted otherwise. Component numbers not appearing in the parts list are for identification purposes only.

J1, J2 — Connector, SO-239.
J3 — Connector, phono-type.
K1 — Dpdt relay, 5-ampere contacts, 12-volt field.
R1 — Linear potentiometer, 5000 ohms.
R2 — Linear potentiometer, 2.5 megohms.
RFC1, RFC2 — Rf choke, 42 turns no. 28 enam. wire on an Amidon FT-50-43 core.
T1 — Broadband transformer. Primary: 25 turns no. 28 enam. wire on an Amidon FT-50-43 core; secondary: 5 turns no. 28 enam. wire wound over primary winding.
U1 — Op amp, 741.

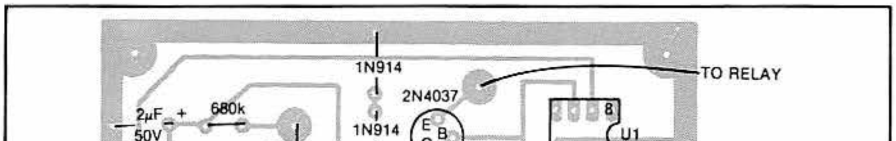
1N914
2N4037
TO RELAY
2μF +
50V
680k
1N914
E
B
C
8
U1
1
R6
RFC 2
0.01μF
1N914
0.01μF
J
1N914
39pF
100Ω
1000Ω
R1
T1
1N914
0.1μF
0.01μF
E
B
C
1000Ω
5600Ω
2N2222A
TO RELAY AND J2
RFC 1
0.1ΩF
J = JUMPER

Fig. 56 — Circuit-board etching pattern for the changeover relay. The pattern is shown at actual size from the foil side of the board with gray areas representing copper. Decimal-value numbers alone represent capacitance in microfarads. Whole-number values with no units represent resistances in ohms; k = 1000. J = wire jumper. All components are mounted on the nonfoil side of the board. K indicates the cathode of a diode.

Construction

In Fig. 56 is a suggested printed-circuit board layout showing the foil pattern and parts placement. R1 and R6 may be circuit-board-mounted potentiometers if the builder desires, as the board layout was designed to accommodate such units. Potentiometers of this type are entirely suitable if the builder does not anticipate frequent changes in power levels and length of delay, as these controls must be readjusted for such changes. If the builder is interested in having access to these controls, panel-mounted potentiometers can be used at the front apron, in which case wires will have to be run from the circuit board to the controls.

Our finished model was housed in a 5 × 3 × 2-inch (127 mm × 76 mm × 51 mm) aluminum Minibox. R1 and R6 were mounted on the front panel and Kurz-Kasch knobs were placed on the control shafts to dress up the appearance of the finished product. D4 is attached directly to the relay terminals. The rear panel supports the terminals for the transmitter, antenna, receiver antenna and power supply. SO-239 connectors were used for the transmitter and antenna connections, and a phono connector was employed for the receiver antenna. Any type of connectors may be substituted to match those used in the builder's station. The power-supply lead is brought out through a feedthrough capacitor. This type of feedthrough is not necessary; a simple feedthrough insulator or wire could be used instead. In our case, the capacitor lead served as a convenient tie point during preliminary testing of the unit.

Operation

Connect the antenna, transmitter, receiver antenna lead and the power-supply connections to the appropriate terminals. Set R6, the time-delay control, to its minimum delay position (ccw) and the sensitivity control to its midrotation position. While keying the transmitter, decrease the setting of the sensitivity control to a point just above that where the transmitter no longer makes the relay energize. Once this has been set, it may be left in that position until a moderate change in power level has been made. The delay control should then be adjusted for the length of delay desired. The circuit shown has been tested from 1.8 to 30 MHz and at dc power input levels up to and including 1000 watts. This material was originally presented in *QST* by Doug DeMaw, W1FB, and Jay Rusgrove, W1VD.

Measuring Antenna Gain with Amateur Methods

Beat the cost of expensive test gear. Make your antenna measurements using this unique method! A VU meter, a handful of parts and this simple explanation are all you need.

Several excellent articles have been written about the techniques used at vhf conferences to measure antenna gains. Two such articles in particular, by Clark[1] and Knadle,[2] suggest very good antenna-range procedures. Both describe techniques which, if properly applied, will produce accurate relative antenna-gain data (in decibels) for vhf-uhf antennas.

However, these antenna range-measurement systems require equipment most amateurs lack — ratiometers, precisely calibrated attenuators, or at least specialized detectors and commercial grade audio or VSWR meters. Moreover, setting up a near-field antenna range is a tricky business in itself. One text[3] simply warns amateurs not to expect accurate antenna-gain measurements if they try it! The problem, of course, is that outdoor antenna ranges invite all sorts of error-causing reflections from various objects, not the least of them being the earth itself.

Ground reflections alone can completely invalidate the results on an otherwise good antenna range. At the 1973 West Coast VHF Conference, for instance, the author's 8-element quagi for 432 MHz[4] was measured at precisely the same gain as a well-built 15-element Yagi more than twice its size, despite the fact (established in other tests) that the bigger antenna had 2 to 3 dB more actual gain! The problem was that, even with the sophisticated measuring equipment in use, ground reflections were distorting the results. The best "gain" indication came with both antennas within a few inches of the earth, and the smaller quagi with its lower Q didn't mind the detuning effects of the earth's proximity as much as the bigger and higher Q array.

With results like that, wild claims are sometimes made for one vhf antenna or another. And the claims may be even wilder on the hf bands, where an amateur may well be told his signal is two of three S units louder than another ham's signal when the actual gain difference between the two antennas is insignificant. The ham who "won " the comparison by several S units, of course, is likely to go away thinking his antenna is 12 or 18 dB better than the other guy's, since S meters are supposed to show 6-dB signal difference per S point.

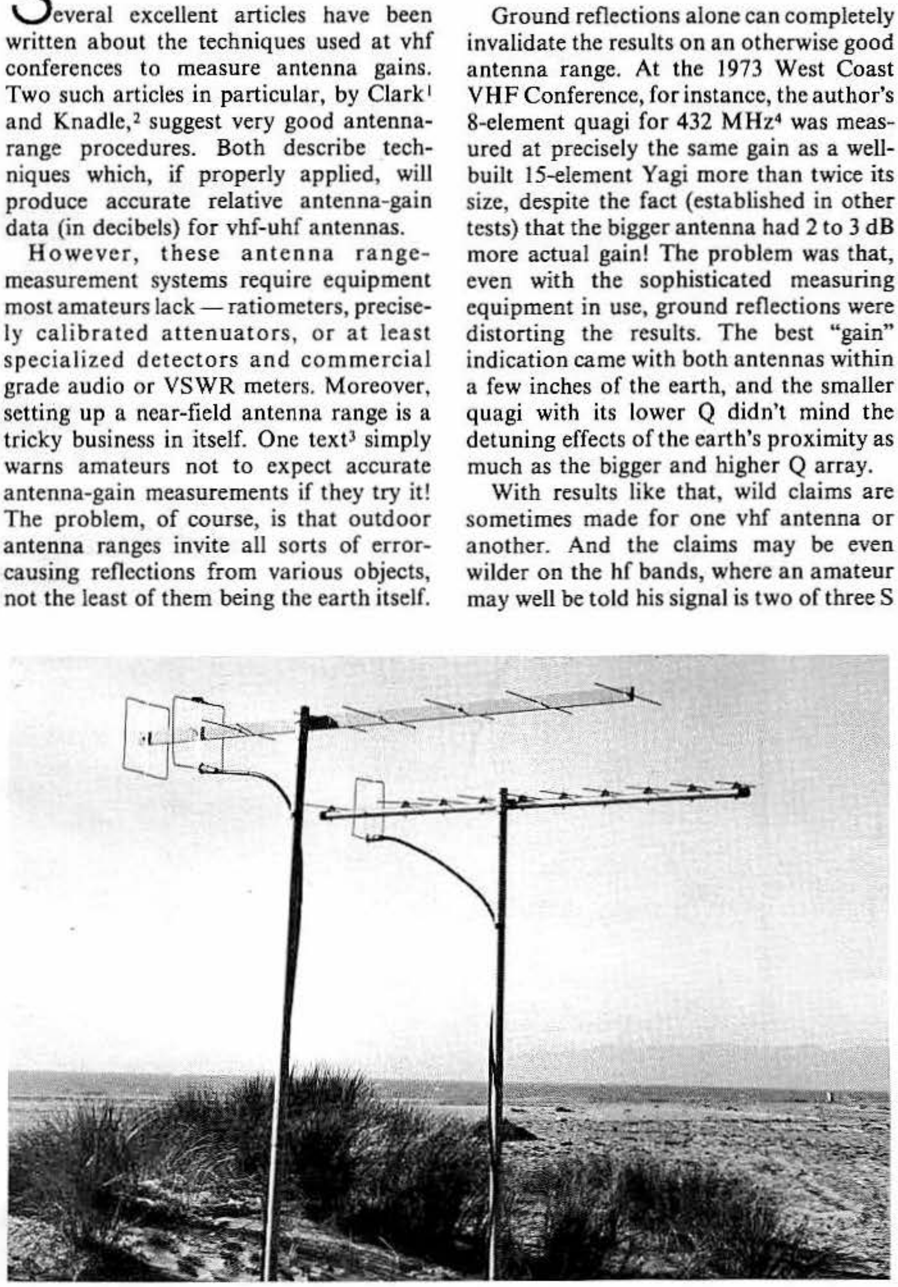

For the antenna gain-measurement technique described here, two identical masts in a clear area (here a southern California beach) are required. With the test and reference antennas at equal heights and fed with identical feed lines, the signal produced by the two antennas is compared by rapid and repeated switching. Here an 8-element quagi is tested against an 11-element Yagi that has been modified for quagi-type drive. The modified Yagi delivers 3 dB more gain than the conventional version, but 2 dB less than the quagi.

A Better Way

This article describes a more reliable and valid way to estimate antenna gain, and without expensive or hard-to-find test equipment. The technique requires only this equipment: a cooperative nearby ham with a transmitter capable of generating a few watts of steady carrier, a receiver whose agc can be turned off, an ordinary audio VU meter, and some antennas to measure.

The technique is nothing more than a more rigorous application of the test procedure radio amateurs have always used when they asked for "comparison" signal reports. It takes advantage of the fact that audio VU meters, unlike S meters, are calibrated directly and fairly accurately in decibels, with easy readout down to fractions of a dB. And with the agc turned off, most receivers will closely reproduce changes in the signal input at the antenna terminals with corresponding changes in audio output until they are badly overloaded.

This means that a nearby amateur's signal transmitted at a reasonably low power level can be used to compare the gain of various antennas by rapidly switching back and forth between them and averaging the results.

The only trick is to use a "test signal" that is coming from a point source — not a signal that is being reflected from surrounding objects. If your test antennas can "see" the signal source a few miles away and are well clear of nearby objects (and clear of the earth itself), valid results should be attainable if the procedures below are followed.

The author has spent many hours working with both local antenna ranges covering a few dozen wavelengths, and on-the-air signals from local stations. It is apparent that ground reflections and other

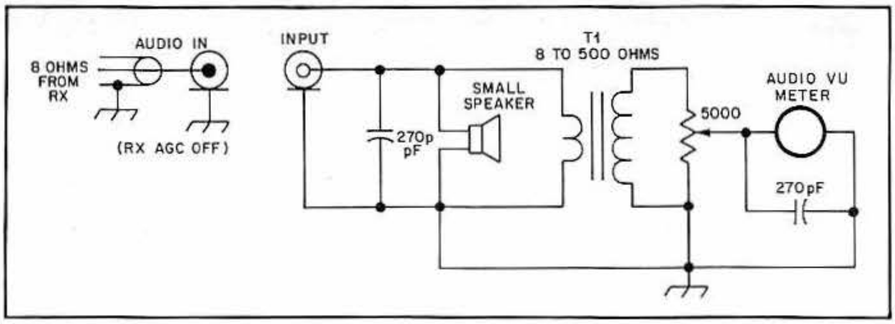

Fig. 57 — The audio VU-meter system for antenna gain measurements.

Fig. 58 — This small box houses all the test equipment required for antenna gain measurements using the technique described here. The front-panel audio VU meter is calibrated directly in decibels, with readout possible to about 0.2 dB in the most expanded part of the scale. Not visible are the meter level potentiometer (on rear), and small speaker which is mounted downward.

such error sources are less serious problems when a steady, on-the-air carrier a few miles away — and not a local signal a few dozen wavelengths away — is the signal source.

Test Procedures

Here's how the procedure works at vhf (at hf, it's a little more complicated because antennas are larger, as we'll see later):

1) Build an audio VU metering circuit like the one shown in the photo and Fig. 57. All parts should be readily available anywhere that consumer electronics parts are sold. The speaker and meter-gain control are handy to have when the receiver itself is located some distance away from the antenna test site, but they aren't absolutely essential.

2) Get a nearby amateur (perhaps one to five miles away) to generate a steady carrier into an antenna of the same polarization as your test antennas. A small beam pointed toward your test site is much better than a dipole or groundplane antenna because it is less likely to invite reflection error. A beam pointed somewhere else is worst of all!

3) Two of the antennas to be tested should be set up side by side, with identical lengths of feed line going to a coaxial relay or well-isolated switch. The two supporting poles should be high enough that the test antennas are at least two wavelengths above ground (13 feet [3.96 m] or more at 2 meters) and clear of nearby obstructions. To minimize interaction between the two antennas, the two masts should also be at least one to two wavelengths apart and standing side by side in relation to the signal source.

4) Make sure you can raise and lower your antennas and move them from side to side by at least a full wavelength or so (81 inches [2.06 m] at 2 meters) without the indicated signal varying more than about 1 dB. If you can't do that, you probably have reflection problems that will have to be solved before you can make valid readings. If this is the case, try different sites until you do achieve nonspotty reception.

5) Check the receiver agc to be sure it is off; then rapidly switch between the two antennas and record the difference in decibels. Do it several times and calculate the average figure (the mean, for statisticians).

6) Next reverse the two antennas. Put antenna no. 1 on mast no. 2 and attach it to feed line no. 2. Move antenna no. 2 over to mast no. 1 and feed line no. 1. Do not eliminate this step and expect valid results.

Now run the test again. If there are more than very slight differences from the previous test, one antenna or the other is seeing some obstruction or reflection source. Your site isn't clear enough for good results; try a different one.

Since this test procedure requires only receiving, there is no reason not to perform it somewhere other than at home — anywhere that is clear of obstructions. Use a receiver or transceiver that runs on battery power, if necessary.

7) If you have followed steps one through six, you now have an accurate indication of the relative gain of your two antennas. Now you can make one of your antennas a "reference" antenna and compare any other antennas you or

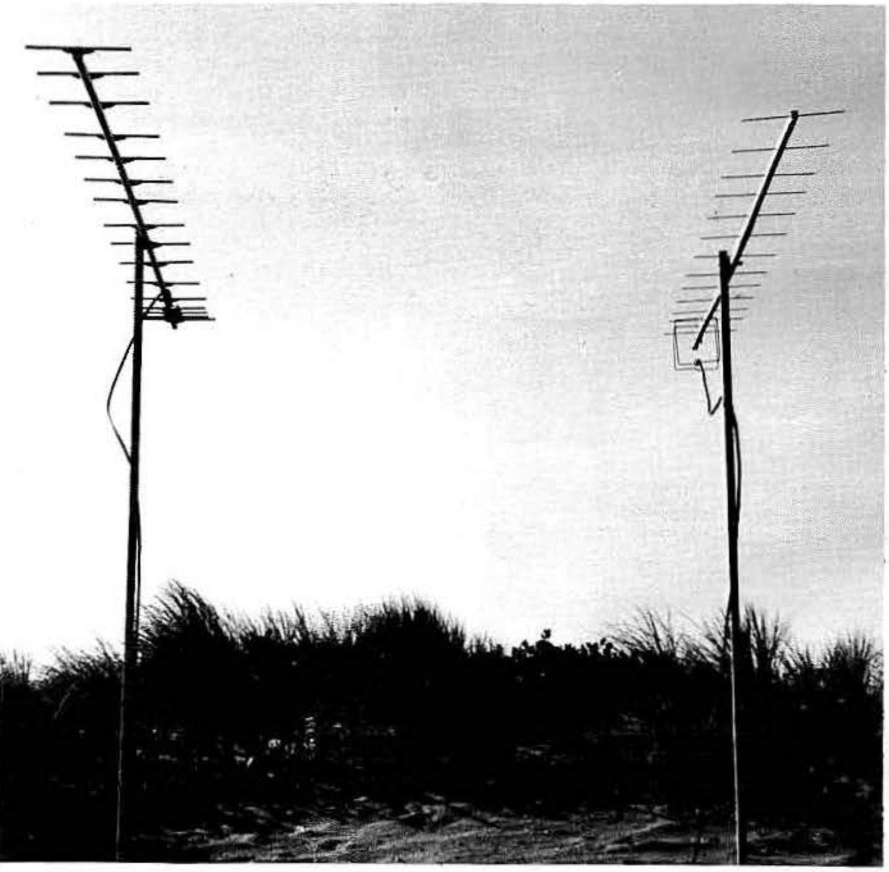

Fig. 59 — Here a 15-element long-boom quagi for 432 MHz is compared to a 16-element log-periodic Yagi, using the test procedure described in this article. The two antennas both deliver 15 dB gain over a dipole, decisively better than any other 432-MHz antenna tested here. The antennas to be tested are set up side by side and fed with identical lengths of coax. The antennas should be at least two wavelengths above ground and clear of nearby obstructions. To minimize interaction between the two antennas, the masts should also be at least one to two wavelengths apart, and standing side by side in relation to the signal source.

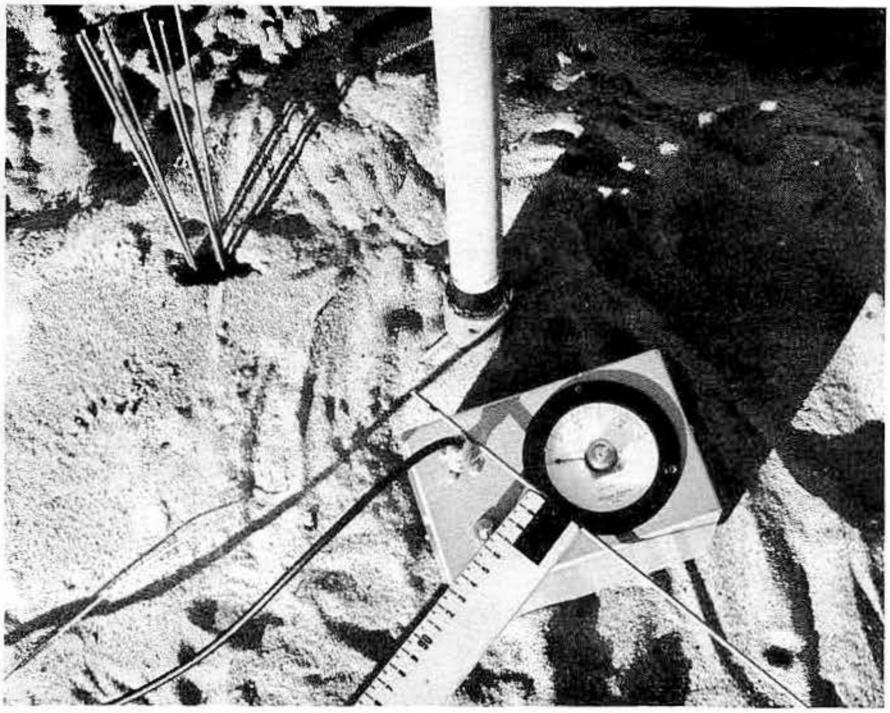

Fig. 60 — An earlier version of the metering system in Fig. 57, with assorted elements of various lengths and the yardstick used to make temporary placements on the boom.

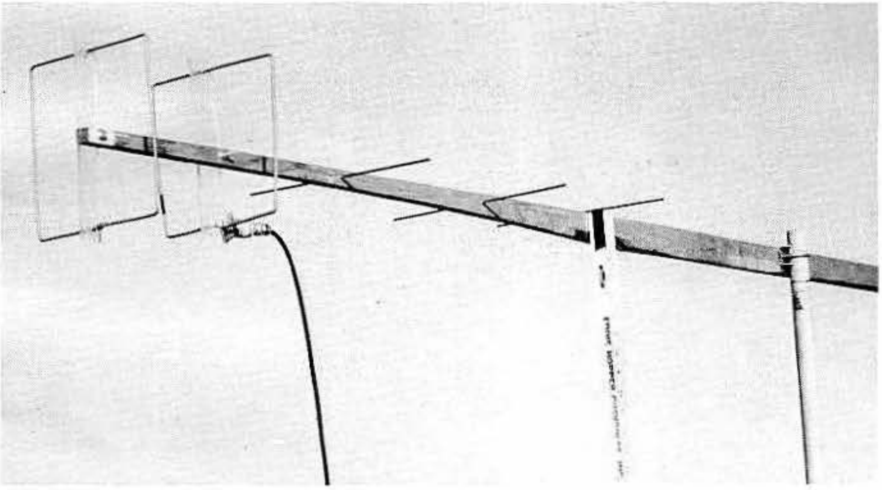

Fig. 61 — The technique used to design a new vhf or uhf antenna experimentally. The boom is shown with the first few elements already in place, and each additional element is adjusted back and forth on the boom until a point of maximum indication of the VU meter is found.

neighboring hams have against this reference.

So how do you determine how much gain your antennas have over a dipole? Some people would say, "Simple, make the reference antenna a dipole." But it may not be that easy, because dipoles are notorious for finding reflections that more directional antennas can't "see." At vhf conference gain-measurement seminars, dipoles have sometimes shown as much as 3 or 4 dB gain over a dipole!

The best reference antenna is probably a small beam of known (or approximately known) gain, assuming you don't have the National Bureau of Standards type of reference antenna that professionals use for this purpose. A 3-element Yagi or a 2-element quad built to handbook dimensions will usually deliver about 6 dB gain over a dipole at 2 meters, if the SWR is okay at the design frequency.

Thus, if you find that a particular 8-element antenna consistently produces 4.5 more dB on the VU meter than your little reference beam, you'll know you have an antenna with about 10.5 dB gain. You won't be exactly correct, but you'll be close to the right figure. Conversely, if your bigger beam doesn't show much gain over the small one, and there are no measurement errors of the sort just described, you'll know for sure that something is wrong with the big antenna.

Once you have established the integrity of your test site, you can even design your own antennas very accurately, using the audio VU-meter technique. To do that, set up your reference antenna and begin designing your new antenna by putting an empty boom on the other pole. Now start with 2 or 3 elements, get a decent SWR and try adding more elements.

As you add elements, you will be able to actually see where each new element "wants to be" on the boom, and you can see your new antenna gain over the reference antenna increase as you go along. Try all sorts of element lengths and spacings as you progress.

Antenna design professionals call this process "perturbation" and they do it with sophisticated mathematics, but you can design antennas that work just as well as theirs on your own antenna range, once you master this technique. Then, if you do go to a vhf conference with your best home-designed antennas, you'll know about how much gain they have before you enter the gain contest. And you'll never come away from such an event with an antenna that has "minus 3 dB gain!" Moreover, you'll know that your measurements, made without any of the sophisticated equipment you saw at the vhf conference, are about as accurate as anybody's. In fact, you may find your own measurements are more valid than the ones at the conference, especially if the antenna-gain figure jumps all over the place when you raise and lower it at the conference antenna range.

HF Antenna Measurements

So far, all of this has been devoted to vhf antenna-gain measurements. Many readers will say, "That's fine, but I want to measure the gain of my 5-element 20-meter beam!" Of course, an audio VU meter and a receiver without agc are a much better way than the notoriously inaccurate S meter to compare two dissimilar hf antennas. The author has seen an S meter show three or four *S units* of signal difference between two antennas that are really only 3 or 4 dB apart in gain.

However, the only accurate way to calculate the gain of an antenna is to compare it to a reference antenna at the same height and then switch the two antennas to opposite masts to validate the data. That may be impractical at hf, but perhaps not as much so as it seems at first. Most amateurs motivated enough to install a 5-element beam on a crankup tower can probably find a way to temporarily set up a smaller beam (not a dipole) on a 20-foot (6.10 m) mast above a roof. Then the two antennas at identical heights can be compared on stable, line-of-sight signals.

The results of this test can be invalidated by obstructions or even by other antennas mounted on the same tower with the test antenna, but at least this will be more valid than a comparison against another ham's readings, when the two locations and rigs may be entirely different.

And if you can switch your test reference hf antennas to opposite masts, you can measure the actual gain of one over the other. One suggestion before you put that giant, long-boom array on the big tower: Put a little beam on the tower and an identical one on the rooftop mast and perform the audio VU-meter tests to see if the two antenna sites do seem to be identical. If they are comparable, then you can put the big antenna on the tower and find out how much gain it has, if any, over the small one on your roof when the two are at the same height.

This may seem cumbersome, and in fact it is! That's why most antenna designers do their research at vhf where antennas are so nice and small, and then scale their best designs up for the hf bands. But if you really want to know your big beam's actual gain, this is probably the easiest way to find out.

Conclusions

The antenna gain-measuring technique described in this article has repeatedly been proved to be reliable and valid. The author has used it to design new antennas that work well enough for all kinds of contest and DX work, including moonbounce. For a dedicated amateur, working DX on a home-designed antenna far surpasses the thrill of working the same DX with an antenna purchased at a store! But remember, whether you're working 20 meters or 23 centimeters, a signal source, a receiver with its agc disabled and an audio VU meter are the only pieces of equipment you need to measure antenna gain with this technique. This material was originally presented in *QST* by Wayne Overbeck, PhD, N6NB.

References

[1] Clark, "Direct Methods for Measuring Antenna Gain," *Ham Radio,* July 1969.

[2] Knadle, "UHF Antenna Ratiometry: An Outdoor Range Improvement Technique," a paper presented to IEEE Intercon session L, April 1975.

[3] Brier and Orr, *VHF Handbook for Radio Amateurs,* Radio Publications, Inc., Wilton, CT p. 225.

[4] Overbeck, "The VHF Quagi," *QST,* April 1977.

[5] Knadle, "UHF Antenna Ratiometry," *QST,* February 1976.

Index

"Airstream Loop," 75-meter: 90
Antenna ideas: 93, 109
Apartment dweller's dilemma: 92
Aperture: 54

Baby ultimate Transmatch: 138
Balun: 136
Base loading network, vertical: 17
Beams: 41, 53
Bobtail array: 81
Bridge, resistive antenna: 131
Bridge, SWR: 125
Broadband matching networks, simple: 140
Broadband matching transformers: 140

Capacitance hats: 111
Capacitive loading: 110
Capacitor-tuned quad: 63
Changeover relay, rf-sensed: 142
Coil housing box: 10
Comparisons:
  High vs. low antennas: 102
  Quad vs. Yagi: 53, 100
Computer-modeled arrays and patterns: 119
Coupling methods: 104
Cross-modulation, receiver: 135

DDRR, 40-meter: 95
Delta-Loop beam, HRH: 84
Delta Loop, 10-meter: 84
Delta Loop, 15-meter: 86
Delta Loop, 20-meter: 85
Dipoles: 89
Dipole, off-center-loaded: 107
Directive arrays: 41, 53
Directivity: 53
Director spacing, effects of: 50
Distributed loading: 26
Dual-band 160- and 80-meter vertical: 141
DX antenna, 160- and 80-meter: 34
DX weasel, 20-meter: 41

Efficient short radiators: 13, 25, 27, 112, 141
E-skip tests, 6-meter: 103
Etched-circuit Monimatch: 123

Ferrite-rod loop: 78
Folded mini quad: 69

Gain, measuring with amateur methods: 145
Gain, quad vs. Yagi: 55
Ground-image vertical: 22
Grounding systems, verticals: 29
Ground-mounted short verticals: 25
Groundplane, H-plane pattern: 11
Groundplane with tuned feeders, multiband: 19
Ground-wave tests: 103

Half-size 80-meter: 110
Half square array, three-element: 82
Half square array, two-element: 82
Half square, 80-meter: 81
Harmonic problems: 33, 135
Helical, 8-1/2 ft: 26
Helically wound vertical, 75-meter: 20
Hf measurements: 147
High vs. low antennas: 102
Horizontal vs. vertical: 32
H-plane pattern, groundplane: 11
HW-40 micro beam: 43

Impedance properties and matching, quad: 56
Inductive loading: 110
In-line rf power metering: 126
Input resistance vs. number of radials: 27

L-network Transmatch: 92
Loaded antennas for travel: 89
Limited-space antennas: 7, 104
Linear array, vertical four-element: 118
Loading coils with multiband antennas: 112
Loading, capacitive and inductive: 110
Loading, distributed: 26
Loading, top-hat: 26
Loads, antenna: 133
Loads, transmitter: 104
Load, 50-ohm: 123, 135
Log-Yag array: 49
Loop:
  Antenna, W1FB 4T-ES: 80
  Antennas for receiving: 76
  Experiments: 78
  Ferrite-rod: 78
  Four-turn coax: 80
  Four-turn shielded: 77
  Indoor: 77
  40-meter full wave: 98
  160-meter indoor receiving: 80

Matching networks, simple broadband: 140
Matching, two-band quad: 75
Measuring gain with amateur methods: 145
Metering, in-line rf power: 126
Micro beam, HW-40: 43
Minooka special: 36
Minooka special, mobile: 38
Miscellaneous antenna types: 76
Monimatch, etched-circuit: 123
Monopole, 40-meter: 90
Multiband antennas: 7, 9, 11, 13, 16, 18, 19, 32, 45, 60, 92, 133
Multiband antennas with loading coils: 112
Multiband ground plane with tuned feeders: 19
Multiband verticals: 16, 32
Multiband verticals, nontrap: 33

Noise, beating: 76
Novice four-band vertical: 9
Novice two-band vertical: 7
Novi-Loop: 98

Off-center-loaded dipole: 107
Optimizing verticals: 29, 115

Phased array, dissimilar antennas: 30
Pi-network tank circuits: 132
Polarization, quads: 54
Portable 160-meter vertical: 38
Power dividers, rf: 120
Preamplifier, broadband 40-dB: 78

Quads: 53
  C-T, reduced size: 63
  Folded mini: 69
  Impedance properties and matching: 56
  Matching, two-band: 75
  Polarization: 54
  Radiation angle: 54
  Spider mount: 74
  Stub tuning: 70
  Triband one-loop cubical: 60
  Yagi-type directors: 74
  7-MHz, linear-loaded: 73
  15-meter, two-element: 61
  40-meter: 62
  80-meter two-element: 66
Quad vs Yagi: 53, 100

Radials, input resistance vs. number of: 27
Radiation angle, quad: 54
Radiation resistance/height curve: 27
Random-length wires: 92, 104
Reactance compensation: 18
Receiver cross-modulation: 135
Resistive antenna bridge: 131
Rf power dividers: 120
Rf power metering, in-line: 126
Rf-sensed changeover relay: 142
Rod-loop tests: 79
Rod loop, sense antenna with: 79

Scoop loop beats noise: 76
Sense antenna: 79
Short radiators, efficient: 27, 112
Short vertical design: 13, 25, 27, 141
Short verticals, SWR: 28
Shunt-feeding towers: 34,
Simple arrays of vertical elements: 114
Spider quad mount: 74
Square array, vertical four-element: 118
Stub tuning, quads: 70
SWR: 124
SWR bridge: 125

Tank circuits, pi-network: 132
Theory and test methods: 100
Top hat loading: 26
Towers, shunt-feeding: 34
Transformers, broadband matching: 140
Transmatch: 92, 105, 132, 135, 138
Transmatch, baby: 138

Transmatch, L-network: 92
Transmatch, Ultimate: 135
Transmission lines: 124
Transmitter loads: 104
Trap antennas: 33, 133
Trap verticals: 33
Travel trailers and campers: 89
Triangle three-element vertical array: 122
Tribanders: 13, 18, 60, 100
Triband one-loop cubical quad: 60

Ultimate Transmatch: 135
Ultimate transmatch, baby: 138

Varimatcher: 105
Verticals: 7
  Four-element arrays: 117
  Ground-image: 22
  Ground-mounted short: 25
  High-performance 20-, 40- and 80-meter: 13
  Linear array, four-element: 118
  Multiband: 32
  Multiband groundplane with tuned feeders: 19
  Nontrap multiband: 33
  Novice four-band: 9
  Novice two-band: 7
  Optimizing: 29, 115
  Radiator, multiband: 16
  Short design: 13, 25, 27, 141
  Simple arrays: 114
  Square array, four-element: 118
  Three-element arrays: 117
  Trap: 33
  Triangle three-element array: 122
  Types of: 32
  Two-element array: 115
  5/8-wave, 20-meter: 39
  10- and 15-meter groundplane: 11
  10-, 15- and 20-meter: 18
  75-meter, helically wound: 20
  160-meter: 36
  160- and 80-meter system: 141
  360-degree steerable phased arrays: 119
Vertical vs. horizontal: 31

Why a Transmatch?: 132
Wilkinson power dividers: 121
Wire, nonresonant, random-length: 92

Yagis: 41
  Director spacing, effects of: 50
  Helically wound, two-element, 40-meter: 43
  Log-Yag array: 49
  20- and 15-meter bite-size beam: 45
  20-meter DX Weasel: 41
  20-meter minibeam: 48
Yagi-type directors for quads: 74
Yagi vs. quad: 53, 100

6-meter E-skip tests: 103
7-MHz, linear-loaded quad: 73
10- and 15-meter groundplane: 11
10-meter Delta Loop: 84
10-, 15- and 20-meter one-loop quad: 60
10-, 15- and 20-meter vertical: 18
15-meter Delta Loop: 86
15-meter quad, two-element: 61
20- and 15-meter bite-size beam: 45
20-meter arrays, beam patterns: 50
20-meter beam, linear loaded: 48
20-meter Delta Loop: 85
20-meter DX Weasel: 41
20-meter minibeam: 48
20-meter, 5/8-wave vertical: 39
20-, 40- and 80-meter vertical system: 13
40- and 15-meter vertical: 7
40-meter DDRR: 94
40-meter full-wave loop: 98
40-meter monopole: 90
40-meter quad: 62
40-meter vertical, short: 25
40-meter Yagi, helically wound two-element: 43
50-ohm load: 123, 135
75-meter "Airstream Loop"' 91
75-meter, helically wound vertical: 20
80-meter DX antenna: 34
80-meter, half-size: 110
80-meter half-square: 81
80-meter quad, two-element: 66
80-meter receiving loop: 76
80-meter vertical, short: 27
160-meter DX antenna: 34
160-meter indoor receiving loop: 80
160-meter, space-saving vertical: 40
160-meter vertical: 36
160-meter vertical, portable: 38
160-meters, receiving loop: 76
160- and 80-meter vertical: 141
160 on 80-meter half-square: 83
360-degree steerable vertical phased arrays: 119

The Radio Amateur's Handbook
THE RADIO AMATEUR'S VHF MANUAL
THE ARRL ANTENNA BOOK

ARRL publications . . .

the foundation of a dynamic hobby

Open the doors to amateur radio.

Whether novice or old-time amateur, student or engineer, League publications will help you to keep abreast of the times in the ever-expanding field of electronics.

THE RADIO AMATEUR'S HANDBOOK
Internationally recognized, universally consulted. The all-purpose volume of radio. Packed with information useful to the amateur and professional alike. Written in a clear, concise manner. Contains hundreds of photos, diagrams, charts and tables. **$9.75**

ARRL HAM RADIO OPERATING GUIDE
Written for the amateur who prides himself on good operating procedures. A ready reference source and guide to good operating practices. Ideal for the amateur who wishes to brush up on operating procedures and who wishes information on all facets of amateur operating. **$4.00**

TUNE IN THE WORLD WITH HAM RADIO
The complete beginner's package. Everything needed to obtain a Novice license: Theory, rules, how to assemble a station, and operating practices. With one-hour code cassette. **$7.00**

SOLID STATE BASICS
Enter the fascinating world of solid state with this exciting book. Perfect as a study guide for the beginner or as a refresher for the more experienced builder. Complete with a ten watt transmitter as a workshop exercise. **$5.00**

THE RADIO AMATEUR'S VHF MANUAL
A thorough treatment of vhf. Covers receiving and transmitting principles, techniques and construction, antenna and feed system design, uhf and and microwaves, test equipment, interference causes and cures. **$4.00**

GETTING TO KNOW OSCAR — FROM THE GROUND UP
This booklet is a reprint of an extremely popular series appearing in *QST*. **$3.00**

UNDERSTANDING AMATEUR RADIO
For the beginner. Explains in simple language the elementary principles of electronic and radio circuits. Includes how-to-build-it information on low cost receivers, transmitters and antennas. A "must" guide for the newcomer. **$5.00**

ARRL CODE KIT
Two top-quality cassettes provide practice at 5, 7-1/2, 10 and 13 wpm. Step-by-step progression and use of random code characters plus proven suggestions and hints in the accompanying instruction book make increasing your code proficiency easy! **$8.00**

FM AND REPEATERS FOR THE RADIO AMATEUR
Complete treatment of fm and repeaters. Chapters on receivers, transmitters, antennas and repeater operation. Of great value to the fm buff as well as to those just becoming interested in the subject. **$5.00**

THE ARRL ANTENNA BOOK
Theoretical explanation and complete instructions for building different types of antennas for amateur work. Simple doublets, multielement arrays, mobile types, rotaries, long wires, rhombics and others. **$5.00**

SOLID STATE DESIGN FOR THE RADIO AMATEUR
Thorough treatment of the use of solid state devices. Provides a wealth of tried and proven circuitry plus practical application data. **$7.00**

ARRL ELECTRONICS DATA BOOK
Includes data on radio frequency circuits, filter design, *L*, *C*, and *R* networks, antennas and feed systems. A catalog of solid-state circuits and much more. This 128-page reference is a useful addition to the amateur's technical library. **$4.00**

THE RADIO AMATEUR'S LICENSE MANUAL
Study guide and reference book. Points the way to the coveted amateur license. Complete with typical questions and answers to the FCC amateur exams for Technician, General, Advanced and Extra Class. **$4.00**

SPECIALIZED COMMUNICATIONS TECHNIQUES
Seven chapters cover the more esoteric forms of amateur radio: ATV, SSTV, FAX, RTTY, Satellite Communications, and advanced communication techniques. **$4.00**

OPERATING KIT
The active ham's favorites — The Ham Radio Operating Guide, Log Book, Message Pad and U.S.A. Call Area Map. **$6.50**

A COURSE IN RADIO FUNDAMENTALS
Twenty-six chapters present the electrical and electronic principles that are basic to understanding radio circuit operation. **$4.00**

RADIO FREQUENCY INTERFERENCE
A must publication for all hams! Full treatment is given to the causes and solutions for those pesky interference problems. Complete with illustrations and photos of the more common kinds of interference. **$3.00**

HINTS AND KINKS
If you build and operate an amateur radio station, you'll find this a mighty valuable book in your shack and workshop. **$4.00**

SINGLE SIDEBAND FOR THE RADIO AMATEUR
A digest of the best articles from *QST*. Includes discussions of theory and practical how-to-build-it descriptions of equipment. **$4.00**

Available from **THE AMERICAN RADIO RELAY LEAGUE, NEWINGTON, CT. 06111**

PRICES ARE VALID IN U.S.A. ONLY AND ARE SUBJECT TO CHANGE WITHOUT NOTICE.

$44.95

The ARRL

ANTENNA BOOK

21st Edition

The *ultimate* reference for Amateur Radio antennas, transmission lines and propagation

Complete CD
for THE ARRL
ANTENNA BOOK

INSIDE: Includes the fully searchable book on CD-ROM

The ARRL Antenna Book

All the information you need to design your own complete antenna system.

Since the first edition in September 1939, radio amateurs and professional engineers have turned to **The ARRL Antenna Book** as THE source of current antenna theory and a wealth of practical how-to construction projects. Use this book to discover even the most basic antenna designs— wire and loop antennas, verticals, and Yagis—and for advanced antenna theory and applications. Many of the antennas in this edition benefit directly from advances in sophisticated computer modeling.

This 21st edition has been extensively revised to include information you can use to build highly optimized or specialized antennas. The book includes new content on Near Vertical Incidence Skywave (NVIS) techniques, phased arrays, S-parameters as used in modern vector network analyzers (VNA), Beverage receiving antennas, mobile “screwdriver” antennas, ionospheric area-coverage maps, *and much…much more.*

Fully searchable CD-ROM included!
Bundled with this book is a CD-ROM containing **The ARRL Antenna Book** in its entirety, using the popular Adobe® Reader® software for Microsoft® Windows® and Macintosh® systems. View, search and print from the entire text, including images, photographs, drawings…**everything!**

The CD-ROM contains additional utility programs, including:

- **YW**—Yagi for Windows
- **TLW**—Transmission Line for Windows
- **HFTA**—HF Terrain Assessment for Windows
- **Range-Bearing**— compute range/bearing or latitude/longitude
- **Arrayfeed1**— designing phased-array feed systems
- **EZNEC-ARRL**— antenna modeling for Windows

PLUS propagation forecast tables for more than 170 locations around the world for all portions of the solar cycle!

CONTENTS

Safety First
Antenna Fundamentals
The Effects of Ground
Antenna Modeling and System Planning
Loop Antennas
Low-Frequency Antennas
Multiband Antennas
Multielement Arrays
Broadband Antenna Matching
Log Periodic Arrays
HF Yagi Arrays
Quad Arrays
Long-Wire and Traveling-Wave Antennas
Direction Finding Antennas
Portable Antennas
Mobile and Maritime Antennas
Repeater Antenna Systems
VHF and UHF Antenna Systems
Antenna Systems for Space Communications
Antenna Materials and Accessories
Antenna Products Suppliers
Antenna Supports
Radio Wave Propagation
Transmission Lines
Coupling the Transmitter to the Line
Coupling the Line to the Antenna
Antenna and Transmission-Line Measurements
Smith Chart Calculations

Includes a comprehensive glossary and index

Published by:
ARRL The national association for AMATEUR RADIO
225 Main Street, Newington, CT 06111-1494 USA
www.arrl.org

ISBN-13: 978-0-87259-987-1
ISBN-10: 0-87259-987-6
54495 >
EAN
9 780872 599871
ISBN 0-87259-987-6 Order No. 9876

The ARRL

Antenna Book

21st Edition

Editor
R. Dean Straw, N6BV

Contributing Editors
L. B. Cebik, W4RNL
Dave Hallidy, K2DH
Dick Jansson, WD4FAB
Roy Lewallen, W7EL
Rudy Severns, N6LF
Frank Witt, AI1H

Production
Michelle Bloom, WB1ENT
Sue Fagan, KB1OKW—Cover Art
Jodi Morin, KA1JPA
David Pingree, N1NAS
Kathy Ford

Software Beta Testers
Ahti Aintila, OH2RZ
Alan Bloom, N1AL
L. B. Cebik, W4RNL
Loney Duncan, W0GZV
Chuck Hutchinson, K8CH
Steve London, N2IC
Dan Maguire, AC6LA
Steve Miller, N8SM (SK)
Bob Neece, K0KR,
Danny Richardson, K6MHE
Rudy Severns, N6LF
Jim Tabor, KU5S
Bill Tippett, W4ZV
Don Toman, K2KQ
Pete Smith, N4ZR
Scott Tuthill, K7ZO
Frank Witt, AI1H.

The Covers:

Tower photo at sunset (front and back): Tower and beam at the station of Francisco R.F. Aragao, PT2TD, in Brasilia, Brazil.

Center image: Kurt Andress, K7NV, working on 40-meter Yagi at N6RO in Oakley, CA. Photo by Dean Straw, N6BV.

Upper right image: UHF antenna. Photo credit: ShutterStock.com.

Lower right image: Some of the members of the C5Z contest DXpedition near Banjul, The Gambia. Photo by Henryk Kotowski, SM0JHF.

ARRL *The national association for* **AMATEUR RADIO**
225 Main Street, Newington, CT 06111-1494 USA
www.arrl.org/

Copyright © 2007 by

The ARRL, Inc.

Copyright secured under the Pan-American Convention

International Copyright secured

This work is publication No. 15 of the Radio Amateur's Library, published by the ARRL. All rights reserved. No part of this work may be reproduced in any form except by written permission of the publisher. All rights of translation are reserved.

Printed in USA

Quedan reservados todos los derechos

21st Edition
First Printing

ISBN: 0-87259-987-6

We strive to produce books without errors. Sometimes mistakes do occur, however. When we become aware of problems in our books (other than obvious typographical errors that should not cause our readers any problems) we post an Adobe Portable Document Format (PDF) file on ARRLWeb. If you think you have found an error, please check **www.arrl.org/notes** for corrections. If you don't find a correction there, please let us know, either using the Feedback Form in the back of this book or by sending an e-mail to **pubsfdbk@arrl.org**.

Foreword

We are pleased to offer the 21st edition of *The ARRL Antenna Book*. Since the first edition in September 1939, each new *Antenna Book* has provided more and better information about the fascinating subject of radio antennas. We've sold more than a million *Antenna Books* over the years to amateurs and professionals alike, making it one of the most successful books in our extensive lineup of publications.

Fundamentals about antennas rarely change from edition to edition, but modern application of these fundamentals can result in more highly optimized or specialized antennas. For example, many of the antennas in this new edition benefit directly from advances in sophisticated computer modeling.

We usually update at least 20% of the material in a new edition, and this book is no exception. There have been major revisions in the following chapters:

- Chapter 2: Updated information the concept of "gain."
- Chapter 6: Further insights into the importance of low elevation angles for the lower frequencies, plus a whole new section on NVIS (Near Vertical Incidence Skywave) operation.
- Chapter 8: Completely new section on feeding of phased arrays by W7EL.
- Chapter 13: Updates on Beverage receiving antennas.
- Chapter 14: New "tape-measure" portable Yagi for fox hunting.
- Chapter 16: New information on mobile "screwdriver" antennas.
- Chapter 23: Expanded section on ionospheric area-coverage maps.
- Chapter 27: New section on S-parameters, as used in Vector Network Analyzers (VNAs)

We are fortunate to have the expertise of some well-known and highly talented authorities, who either wrote or reviewed a number of chapters for technical accuracy:

- Rudy Severns, N6LF, and Roy Lewallen, W7EL—low-frequency antennas.
- LB Cebik, W4RNL—Modeling antennas.
- Dick Jansson, WD4FAB—satellite antennas.
- Dave Hallidy, K2DH—EME arrays.
- Bob Hunsucker, AB7VP, and Carl Luetzelschwab, K9LA—HF propagation.

In addition, some exceptional software writers have contributed programs and data for the *Antenna Book*.

- Roy Lewallen, the author of *EZNEC*, has created a special *EZNEC ARRL* program, just for the *Antenna Book*. *EZNEC ARRL* uses the multitude of specialized modeling files also included on the CD-ROM. These models were used in almost every chapter in the book.
- W7EL has also supplied *Arrayfeed1.exe*, a program to design feed systems for phased-arrays.
- Dr Peter Guth and the US Naval Academy have again graciously allowed ARRL to include the versatile *MicroDEM* mapping program on the CD-ROM. *MicroDEM* can easily and quickly generate customized terrain files for the *HFTA* terrain-assessment program, as well as map terrain all around the country using free US topographic data files from the Internet.
- Jim Tabor, NU5S, wrote *GeoAlert-ARRL*, a wonderful freeware program to track propagation trends and to keep tabs on the latest Internet propagation bulletins.
- Dean Straw, N6BV, editor of *The ARRL Antenna Book* has updated and upgraded his *YW* (Yagi for Windows), *TLW* (Transmission Line for Windows) and *HFTA* (HF Terrain Assessment) programs from the 20th edition. A large number of statistical elevation-angle files for QTHs all around the world are included as well. N6BV has also written a new *Range-Bearing* program that is included on the CD-ROM.
- Also included on the CD-ROM are DOS-based utility programs by several authors that analyze antenna tuners, design mobile antennas and LPDAs, and that scale Yagis for *YW*.
- Are you planning on going on a DXpedition to somewhere you've never been before? The CD-ROM now includes both Simplified and Detailed propagation prediction tables for more than 150 QTHs all around the world. Even if you don't journey to distant lands, these tables will give you plenty of insight on planning contesting or DXing strategies—They can also help you set up that Saturday afternoon schedule with your uncle Harry in Cleveland!

You now have in one place the information you need to design your own complete antenna system scientifically—the elevation angles to aim for from your part of the world and the effects of your own local terrain.

As usual, in a publishing effort of this magnitude, errors creep into the process, despite our best efforts. We appreciate hearing from you, our readers, about errors or about suggestions on how future editions might be made even more useful to you. A form for mailing your comments is included at the back of the book, or you can e-mail us at: **pubsfdbk@arrl.org**.

David Sumner, K1ZZ
Executive Vice President
Newington, Connecticut
February 2007

Contents

1 Safety First
2 Antenna Fundamentals
3 The Effects of Ground
4 Antenna Modeling & System Planning
5 Loop Antennas
6 Low-Frequency Antennas
7 Multiband Antennas
8 Multielement Arrays
9 Broadband Antenna Matching
10 Log Periodic Arrays
11 HF Yagi Arrays
12 Quad Arrays
13 Long-Wire and Traveling-Wave Antennas
14 Direction Finding Antennas
15 Portable Antennas
16 Mobile and Maritime Antennas

17 Repeater Antenna Systems
18 VHF and UHF Antenna Systems
19 Antenna Systems for Space Communications
20 Antenna Materials and Accessories
21 Antenna Products Suppliers
22 Antenna Supports
23 Radio Wave Propagation
24 Transmission Lines
25 Coupling the Transmitter to the Line
26 Coupling the Line to the Antenna
27 Antenna and Transmission-Line Measurements
28 Smith Chart Calculations
A-1 Appendix
AB-1 Index of Advertisers
931 Index

Schematic Symbols Used in Circuit Diagrams

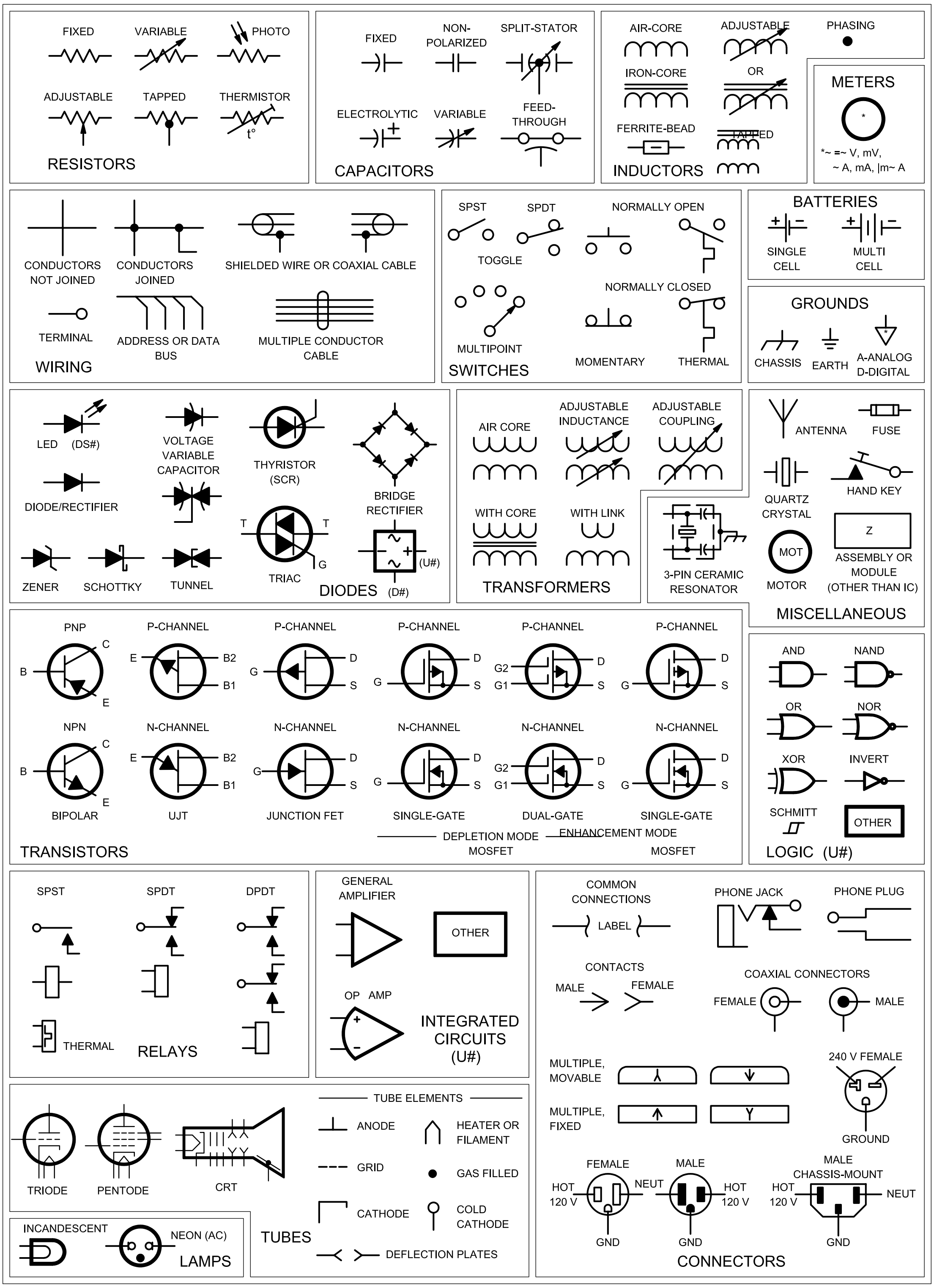
FIXED
VARIABLE
PHOTO
ADJUSTABLE
TAPPED
THERMISTOR
t°
RESISTORS
FIXED
NON-POLARIZED
SPLIT-STATOR
ELECTROLYTIC
VARIABLE
FEED-THROUGH
CAPACITORS
AIR-CORE
ADJUSTABLE
IRON-CORE
OR
FERRITE-BEAD
TAPPED
INDUCTORS
PHASING
METERS
*
*~ =~ V, mV,
~ A, mA, |m~ A
CONDUCTORS NOT JOINED
CONDUCTORS JOINED
SHIELDED WIRE OR COAXIAL CABLE
TERMINAL
ADDRESS OR DATA BUS
MULTIPLE CONDUCTOR CABLE
WIRING
SPST
SPDT
TOGGLE
NORMALLY OPEN
NORMALLY CLOSED
MULTIPOINT
MOMENTARY
THERMAL
SWITCHES
BATTERIES
SINGLE CELL
MULTI CELL
GROUNDS
CHASSIS
EARTH
A-ANALOG
D-DIGITAL
LED (DS#)
VOLTAGE VARIABLE CAPACITOR
THYRISTOR (SCR)
BRIDGE RECTIFIER
DIODE/RECTIFIER
T
T
G
TRIAC
(U#)
ZENER
SCHOTTKY
TUNNEL
DIODES (D#)
AIR CORE
ADJUSTABLE INDUCTANCE
ADJUSTABLE COUPLING
WITH CORE
WITH LINK
TRANSFORMERS
3-PIN CERAMIC RESONATOR
ANTENNA
FUSE
QUARTZ CRYSTAL
HAND KEY
MOT
MOTOR
Z
ASSEMBLY OR MODULE (OTHER THAN IC)
MISCELLANEOUS
PNP
B
C
E
P-CHANNEL
E
B2
B1
P-CHANNEL
G
D
S
P-CHANNEL
G
D
S
P-CHANNEL
G2
G1
D
S
P-CHANNEL
G
D
S
NPN
B
C
E
N-CHANNEL
E
B2
B1
N-CHANNEL
G
D
S
N-CHANNEL
G
D
S
N-CHANNEL
G2
G1
D
S
N-CHANNEL
G
D
S
BIPOLAR
UJT
JUNCTION FET
SINGLE-GATE
DUAL-GATE
SINGLE-GATE
DEPLETION MODE MOSFET
ENHANCEMENT MODE MOSFET
TRANSISTORS
AND
NAND
OR
NOR
XOR
INVERT
SCHMITT
OTHER
LOGIC (U#)
SPST
SPDT
DPDT
THERMAL
RELAYS
GENERAL AMPLIFIER
OTHER
OP AMP
+
-
INTEGRATED CIRCUITS (U#)
COMMON CONNECTIONS
LABEL
PHONE JACK
PHONE PLUG
CONTACTS
MALE
FEMALE
COAXIAL CONNECTORS
FEMALE
MALE
MULTIPLE, MOVABLE
MULTIPLE, FIXED
240 V FEMALE
GROUND
FEMALE
MALE
MALE CHASSIS-MOUNT
HOT 120 V
NEUT
HOT 120 V
HOT 120 V
NEUT
GND
GND
GND
CONNECTORS
TRIODE
PENTODE
CRT
TUBE ELEMENTS
ANODE
HEATER OR FILAMENT
GRID
GAS FILLED
CATHODE
COLD CATHODE
DEFLECTION PLATES
TUBES
INCANDESCENT
NEON (AC)
LAMPS

About the ARRL

The seed for Amateur Radio was planted in the 1890s, when Guglielmo Marconi began his experiments in wireless telegraphy. Soon he was joined by dozens, then hundreds, of others who were enthusiastic about sending and receiving messages through the air—some with a commercial interest, but others solely out of a love for this new communications medium. The United States government began licensing Amateur Radio operators in 1912.

By 1914, there were thousands of Amateur Radio operators—hams—in the United States. Hiram Percy Maxim, a leading Hartford, Connecticut inventor and industrialist, saw the need for an organization to band together this fledgling group of radio experimenters. In May 1914 he founded the American Radio Relay League (ARRL) to meet that need.

Today ARRL, with approximately 150,000 members, is the largest organization of radio amateurs in the United States. The ARRL is a not-for-profit organization that:

- promotes interest in Amateur Radio communications and experimentation
- represents US radio amateurs in legislative matters, and
- maintains fraternalism and a high standard of conduct among Amateur Radio operators.

At ARRL headquarters in the Hartford suburb of Newington, the staff helps serve the needs of members. ARRL is also International Secretariat for the International Amateur Radio Union, which is made up of similar societies in 150 countries around the world.

ARRL publishes the monthly journal *QST*, as well as newsletters and many publications covering all aspects of Amateur Radio. Its headquarters station, W1AW, transmits bulletins of interest to radio amateurs and Morse code practice sessions. The ARRL also coordinates an extensive field organization, which includes volunteers who provide technical information and other support services for radio amateurs as well as communications for public-service activities. In addition, ARRL represents US amateurs with the Federal Communications Commission and other government agencies in the US and abroad.

Membership in ARRL means much more than receiving *QST* each month. In addition to the services already described, ARRL offers membership services on a personal level, such as the ARRL Volunteer Examiner Coordinator Program and a QSL bureau.

Full ARRL membership (available only to licensed radio amateurs) gives you a voice in how the affairs of the organization are governed. ARRL policy is set by a Board of Directors (one from each of 15 Divisions). Each year, one-third of the ARRL Board of Directors stands for election by the full members they represent. The day-to-day operation of ARRL HQ is managed by a Chief Executive Officer.

No matter what aspect of Amateur Radio attracts you, ARRL membership is relevant and important. There would be no Amateur Radio as we know it today were it not for the ARRL. We would be happy to welcome you as a member! (An Amateur Radio license is not required for Associate Membership.) For more information about ARRL and answers to any questions you may have about Amateur Radio, write or call:

ARRL—The national association for Amateur Radio
225 Main Street
Newington CT 06111-1494
Voice: 860-594-0200
Fax: 860-594-0259
E-mail: **hq@arrl.org**
Internet: **www.arrl.org/**

Prospective new amateurs call (toll-free):
800-32-NEW HAM (800-326-3942)
You can also contact us via e-mail at **newham@arrl.org**
or check out *ARRLWeb* at **http://www.arrl.org/**

FEEDBACK

Please use this form to give us your comments on this book and what you'd like to see in future editions, or e-mail us at **pubsfdbk@arrl.org** (publications feedback). If you use e-mail, please include your name, call, e-mail address and the book title, edition and printing in the body of your message. Also indicate whether or not you are an ARRL member.

Where did you purchase this book?

☐ From ARRL directly ☐ From an ARRL dealer

Is there a dealer who carries ARRL publications within:

☐ 5 miles ☐ 15 miles ☐ 30 miles of your location? ☐ Not sure.

License class:

☐ Novice ☐ Technician ☐ Technician with code ☐ General ☐ Advanced ☐ Amateur Extra

---

Name ______________________ ARRL member? ☐ Yes ☐ No

Call Sign ______________________

Daytime Phone ( ) ______________________ Age ______________________

Address ______________________

City, State/Province, ZIP/Postal Code ______________________ E-mail: ______________________

If licensed, how long? ______________________

Other hobbies ______________________

____

Occupation ______________________

For ARRL use only **ANT BK**

Edition 0 21 22 23 24 25 26

Printing 1 2 3 4 5 6 7 8 9 10 11 12

From ______________________________

____

____

Please affix
postage. Post
Office will not
deliver without
postage.

EDITOR, ARRL ANTENNA BOOK
ARRL—THE NATIONAL ASSOCIATION FOR
AMATEUR RADIO
225 MAIN STREET
NEWINGTON CT 06111-1494

— — — — — — — — — — — — — — — — — please fold and tape — — — — — — — — — — — — — — — — —

Chapter 1

Safety First

Safety begins with your attitude. If you make it a habit to plan your work carefully and to consider the safety aspects of a project before you begin the work, you will be much safer than "Careless Carl," who just jumps in, proceeding in a haphazard manner. Learn to have a positive attitude about safety. Think about the dangers involved with a job before you begin the work. Don't be the one to say, "I didn't think it could happen to me."

Having a good attitude about safety isn't enough, however. You must be knowledgeable about common safety guidelines and follow them faithfully. Safety guidelines can't possibly cover all the situations you might face, but if you approach a task with a measure of "common sense," you should be able to work safely.

This chapter offers some safety guidelines and protective measures for you and your Amateur Radio station. You should not consider it to be an all-inclusive discussion of safety practices, though. Safety considerations will affect your choice of materials and assembly procedures when building an antenna. Other chapters of this book will offer further suggestions on safe construction practices. For example, Chapter 22 includes some very important advice on a tower installation.

PUTTING UP SIMPLE WIRE ANTENNAS

No matter what type of antenna you choose to erect, you should remember a few key points about safety. If you are using a slingshot or bow and arrow to get a line over a tree, make sure you keep everyone away from the "downrange" area. Hitting one of your helpers with a rock or fishing sinker is considered not nice, and could end up causing a serious injury.

Make sure the ends of the antenna are high enough to be out of reach of passers-by. Even when you are transmitting with low power there may be enough voltage at the ends of your antenna to give someone nasty "RF burns." If you have a vertical antenna with its base at ground level, build a wooden safety fence around it at least 4 feet away from it. Do not use metal fence, as this will interfere with the proper operation of the antenna. Be especially certain that your antenna is not close to any power wires. That is the only way you can be sure it won't come in contact with them!

Antenna work often requires that one person climb up on a tower, into a tree or onto the roof of a house. Never work alone! Work slowly, thinking out each move before you make it. The person on the ladder, tower, tree or rooftop should wear a safety belt, and keep it securely anchored. It is helpful (and safe!) to tie strings or lightweight ropes to all tools. If your tools are tied on, you'll save time getting them back if you drop them, and you'll greatly reduce the risk of injuring a helper on the ground. (There are more safety tips for climbing and working on towers later in this chapter. Those tips apply to any work that you must do above the ground to install even the simplest antenna.)

Tower Safety

Working on towers and antennas is dangerous, and possibly fatal, if you do not know what you are doing. Your tower and antenna can cause serious property damage and personal injury if any part of the installation should fail. Always use the highest quality materials in your system. Follow the manufacturer's specifications, paying close attention to base pier and guying details. Do not overload the tower. If you have any doubts about your ability to work on your tower and antennas safely, contact another amateur with experience in this area or seek professional assistance.

Chapter 22, Antenna Supports, provides more detailed guidelines for constructing a tower base and putting up a tower. It also explains how to properly attach guy wires and install guy anchors in the ground. These are extremely important parts of a tower installation, and you should not take shortcuts or use second-rate materials. Otherwise the strength and safety of your entire antenna system may be compromised.

Any mechanical job is easier if you have the right tools. Tower work is no exception. In addition to a good assortment of wrenches, screwdrivers and pliers, you will need some specialized tools to work safely and efficiently on a tower. You may already own some of these tools. Others may be purchased or borrowed. Don't start a job until you have assembled all of the necessary tools. Shortcuts or improvised tools can be fatal if you gamble and lose at 70 feet in the air. The following sections describe in detail the tools you will need to work safely on a tower.

CLOTHING

The clothing you wear when working on towers and antennas should be selected for maximum comfort and safety. Wear clothing that will keep you warm, yet allow complete freedom of movement. Long denim pants and a long-sleeve shirt will protect you from scrapes and cuts. (A pull-on shirt, like a sweat shirt with no openings or buttons to snag on tower parts, is best.) Wear work shoes with heavy soles, or better yet, with steel shanks (steel inserts in the soles), to give your feet the support they need to stand on a narrow tower rung.

Gloves are necessary for both the tower climber and all ground-crew members. Good quality leather gloves will protect hands from injury and keep them warm. They also offer protection and a better grip when you are handling rope. In cooler weather, a pair of gloves with light insulation will help keep your hands warm. The insulation should not be so bulky as to inhibit movement, however.

Ground-crew members should have hard hats for protection in case something falls from the tower. It is not uncommon for the tower climber to drop tools and hardware. A wrench dropped from 100 feet will bury itself several inches in soft ground; imagine what it might do to an unprotected skull.

SAFETY BELT AND CLIMBING ACCESSORIES

Any amateur with a tower must own a high-quality safety belt, such as the one shown in **Fig 1**. *Do not attempt* to climb a tower, even a short distance, without a belt. The climbing belt is more than just a safety device for the experienced climber. It is a tool to free up both hands for work. The belt allows the climber to lean back away from the tower to reach bolts or connections. It also provides a solid surface to lean against to exert greater force when hoisting antennas into place.

A climber must trust his life to his safety belt. For this reason, nothing less than a professional quality, commercially made, tested and approved safety belt is acceptable. Check the suppliers' list in Chapter 21, Antenna Product Suppliers, and ads in *QST* for suppliers of climbing belts and accessories. Examine your belt for defects before each use. If the belt or lanyard (tower strap) are cracked, frayed or worn in any way, destroy the damaged piece and replace it with a new one. You should never have to wonder if your belt will hold.

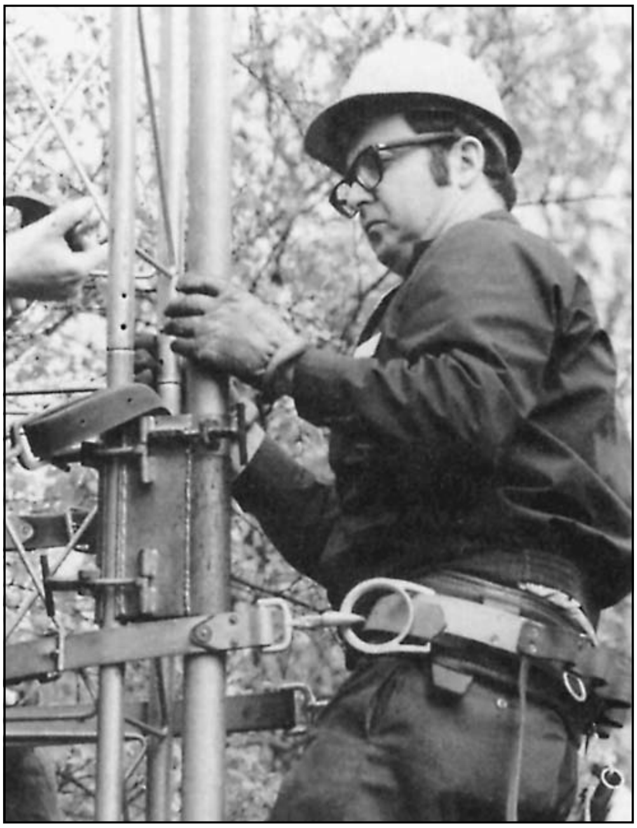

Fig 1—Bill Lowry, W1VV, uses a good quality safety belt, a requirement for working on a tower. The belt should contain large steel loops for the strap snaps. Leather loops at the rear of the belt are handy for holding tools. (*Photo by K1WA*)

Along with your climbing belt, you should seriously consider purchasing some climbing accessories. A canvas bucket is a great help for carrying tools and hardware up the tower. Two buckets, a large one for carrying tools and a smaller one for hardware, make it easier to find things when needed. A few extra snap hooks like those on the ends of your belt lanyard are useful for attaching tool bags and equipment to the tower at convenient spots. These hooks are better than using rope and tying knots because in many cases they can be hooked and unhooked with one hand.

Many hams use climbing belts such as shown in Fig 1. But fully integrated fall-arrest and positioning safety harnesses are preferable. The model ASL-301 in **Fig 2** has a D-ring on the back of the harness to which a safety lanyard is attached. These harnesses are available through Champion Radio Products, Box 572, Woodinville, WA 98072, **www.championradio.com**.

Rope and Pulley

Every amateur who owns a tower should also own a

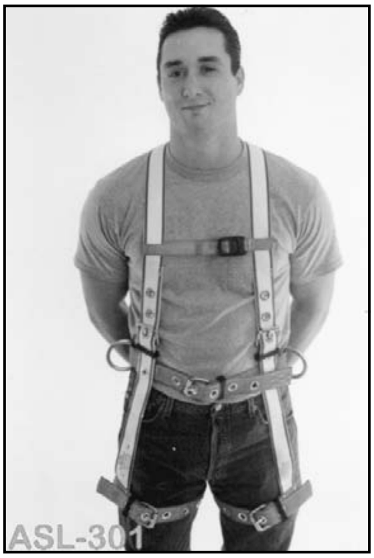

Fig 2—Fall-arrest safety harness integrated with positioning safety belt for tower climbing. *(Courtesy of Champion Radio Products)*

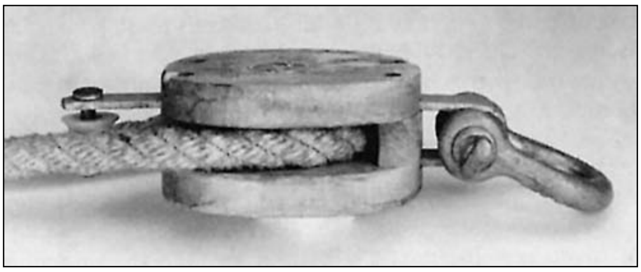

Fig 3—A good quality rope and pulley are essential for anyone working on towers and antennas. This pulley is encased in wood so the rope cannot jump out of the pulley wheel and jam.

good quality rope at least twice as long as the tower height. The rope is essential for safely erecting towers and installing antennas and cables. For most installations, a good quality $^{1}/_{2}$-inch diameter manila hemp rope will do the job, although a thicker rope is stronger and may be easier to handle. Some types of polypropylene rope are acceptable also; check the manufacturer's strength ratings. Nylon rope is not recommended because it tends to stretch and cannot be securely knotted without difficulty.

Check your rope before each use for tearing or chafing. Do not attempt to use damaged rope; if it breaks with a tower section or antenna in mid-air, property damage and personal injury are likely results. If your rope should get wet, let it air dry thoroughly before putting it away.

Another very worthwhile purchase is a pulley like the one shown in **Fig 3**. Use the right size pulley for your rope. Be sure that the pulley you purchase will not jam or bind as the rope passes through it.

THE GIN POLE

A gin pole, like the one shown in **Fig 4**, is a handy device for working with tower sections and masts. This gin pole is designed to clamp onto one leg of Rohn No. 25 or 45 tower. The tubing, which is about 12 feet long,

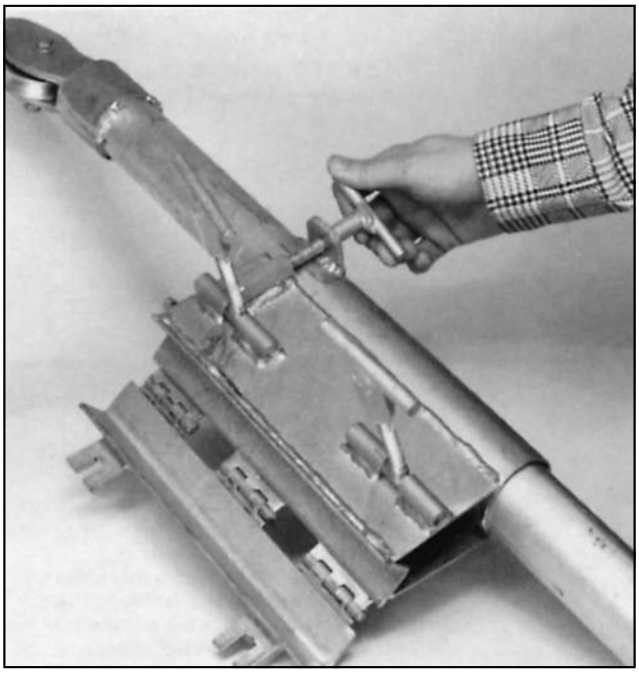

Fig 4—A gin pole is a mechanical device that can be clamped to a tower leg to aid in the assembly of sections as well as the installation of the mast. The aluminum tubing extends through the clamp and may be slipped into position before the tubing clamp is tightened. A rope should be routed through the tubing and over the pulley mounted at the top.

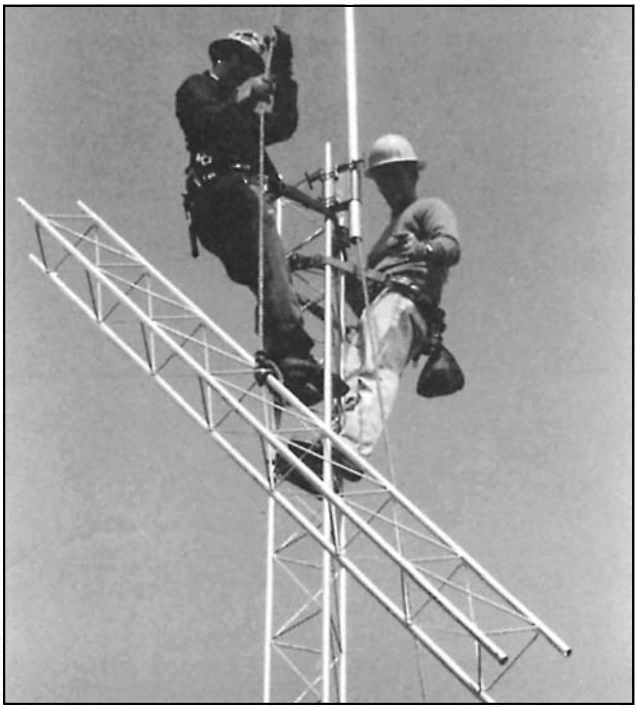

Fig 5—The assembly of tower sections is made simple when a gin pole is used to lift each one into position. Note that the safety belts of both climbers are fastened below the pole, thereby preventing the strap from slipping over the top section. (*Photo by K1WA*)

has a pulley on one end. The rope is routed through the tubing and over the pulley. When the gin pole is attached to the tower and the tubing extended into place, the rope may be used to haul tower sections or the mast into place. **Fig 5** shows the basic process. A gin pole can be expensive for an individual to buy, especially for a one-time tower installation. Some radio clubs own a gin pole for use by their members. Stores that sell tower sections to amateurs and commercial customers frequently will rent a gin pole to erect the tower. If you attempt to make your own gin pole, use materials heavy enough for the job. Provide a means for securely clamping the pole to the tower. There are many cases on record where homemade gin poles have failed, sending tower sections crashing down amidst the ground crew.

When you use a gin pole, make every effort to keep the load as vertical as possible. Although gin poles are strong, you are asking for trouble if you apply too much lateral force.

INSTALLING ANTENNAS ON THE TOWER

All antenna installations are different in some respects. Therefore, thorough planning is the most important first step in installing any antenna. At the beginning, before anyone climbs the tower, the whole process should be thought through. The procedure should be discussed to be sure each crew member understands what is to be done. Plan how to work out all bugs. Consider what tools and parts must be assembled and what items must be taken up the tower. Extra trips up and down the tower can be avoided by using forethought.

Getting ready to raise a beam requires planning. Done properly, the actual work of getting the antenna into position can be accomplished quite easily with only one person at the top of the tower. The trick is to let the ground crew do all the work and leave the person on the tower free to guide the antenna into position.

Before the antenna can be hoisted into position, the tower and the area around it must be prepared. The ground crew should clear the area around the base while someone climbs the tower to remove any wire antennas or other objects that might get in the way. The first person to climb the tower should also rig the rope and pulley that will be used to raise the antenna. The time to prepare the tower is before the antenna leaves the ground, not after it becomes hopelessly entwined with your 3.5-MHz dipole.

SOME TOWER CLIMBING TIPS

The following tower climbing safety tips were compiled by Tom Willeford, N8ETU. The most important safety factor in any kind of hazardous endeavor is the right attitude. Safety is important and worthy of careful consideration and implementation. The right attitude toward safety is a requirement for tower climbers. Lip service won't do, however; safety must be practiced.

The safe ham's safety attitude is simple: *Don't take any unnecessary chances*. There are no exceptions to this plain and simple rule. It is the first rule of safety and, of course, of climbing. The second rule is equally simple: *Don't be afraid to terminate an activity* (climbing, in this case) at any time if things don't seem to be going well.

Take time to plan your climb; this time is never wasted, and it's the first building block of safety. Talk the climb over with friends who will be helping you. Select the date and alternative dates to do the work. Choose someone to be responsible for all activities on the ground and for all communication with the climbers. Study the structure to be climbed and choose the best route to your objective. Plan emergency ascent and descent paths and methods.

Make a list of emergency phone numbers to keep by your phone, even though they may never be used. Develop a plan for rescuing climbers from the structure, should that become necessary.

Give careful thought to how much time you will need to complete the project. Allow enough time to go up, do the work, and then climb down during daylight hours. Include time for resting during the climb and for completing the work in a quality fashion. Remember that the

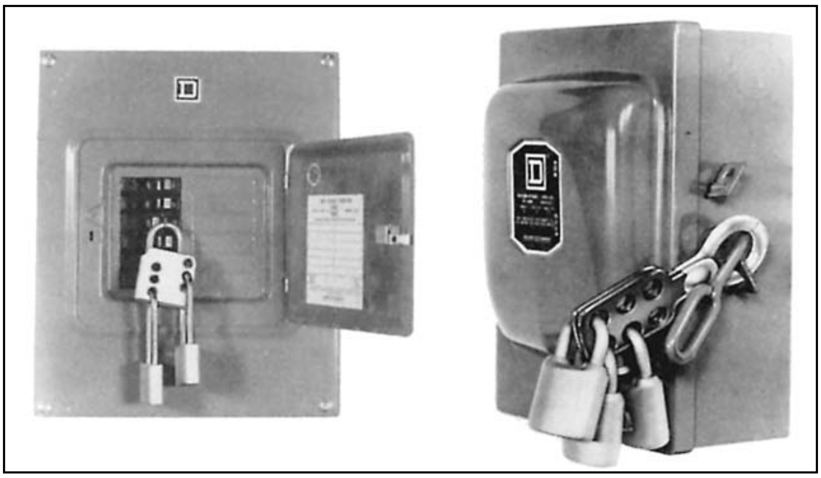

Fig 6—If the switch box feeding power to equipment on your tower is equipped with a lock-out hole, use it. With a lock through the hole on the box, the power cannot be accidentally turned back on. (*Photos courtesy of American ED CO®, at left, and Osborn Mfg Corp, at right.*)

temperature changes fast as the sun goes down. Climbing up or down a tower with cold hands and feet is very difficult—and dangerous.

Give careful consideration to the weather, and climb only in good weather. Investigate wind conditions, the temperature, and the weather forecast. The weather can change quickly, so if you're climbing a really tall tower, it may be a good idea to have a weather alert radio handy during the climb. *Never* climb a wet tower.

The person who is going to do the climbing should be the one to disconnect and tag all sources of power to the structure. All switches or circuit breakers should be labeled clearly with DO NOT TOUCH instructions. Use locks on any switches designed to accept them. (See **Fig 6**.) Only the climber should reconnect power sources.

An important part of the climbing plan is to review notes on the present installation and any previous work. It's a good idea to keep a notebook, listing every bolt and nut size on your tower/antenna installation. Then, when you have to go up to make repairs, you'll be able to take the minimum number of tools with you to do the job. If you take too many tools up the tower, there is a much greater chance of dropping something, risking injury to the ground crew and possibly damaging the tool.

It is also a good idea to review the instruction sheets and take them with you. In other words, plan carefully what you are going to do, and what you'll need to do it efficiently and safely.

It's better to use a rope and pulley to hoist tools. Climbing is hard work and there's no sense making it more difficult by carrying a big load of tools. Always rig the pulley and rope so the ground crew raises and lowers tools and equipment.

Climbing Equipment

Equipment is another important safety consideration. By equipment, we don't just mean tools. We mean safety equipment. Safety equipment should be selected and cared for as if your life depends on it—because it does!

The list of safety equipment essential to a safe climb and safe work on the tower should include:

1) A first class safety belt,
2) Safety glasses,
3) Hard hat,
4) Long-sleeved, pull-over shirt with no buttons or openings to snag (long sleeves are especially important for climbing wooden poles),
5) Long pants without cuffs,
6) Firm, comfortable, steel-shank shoes with no-slip soles and well-defined heels, and
7) Gloves that won't restrict finger movement (insulated gloves if you MUST work in cold weather).

Your safety belt should be approved for use on the structure you are climbing. Different structures may require different types of safety hooks or straps. The belt should be light weight, but strength should not be sacrificed to save weight. It should fit you comfortably. All moving parts, such as snap hooks, should work freely. You should inspect safety belts and harnesses carefully and thoroughly before each climb, paying particular attention to stitching, rivets and weight-bearing mechanical parts.

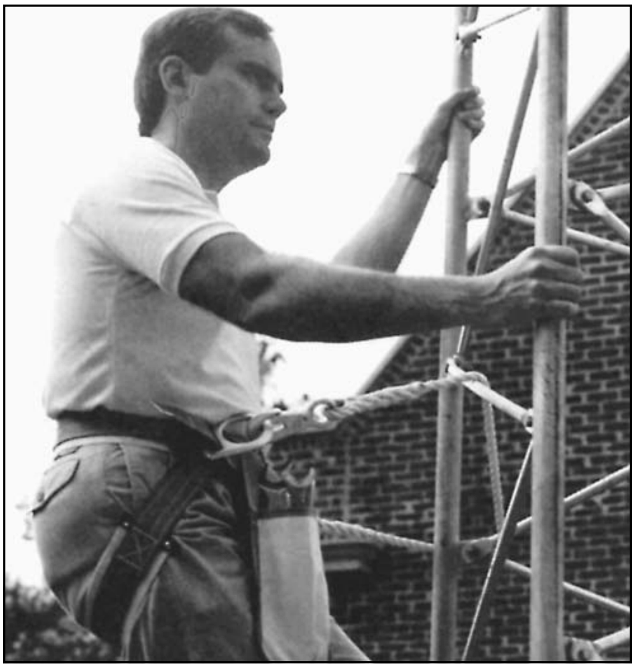

Fig 7—Mark Wilson, K1RO, shows the proper way to attach a safety hook, with the hook opening facing away from the tower. That way the hook can't be accidentally released by pressing it against a tower leg.

Support belt hooks should always be hooked to the D rings in an outward configuration. That is, the opening part of the hook should face away from the tower when engaged in the D rings (see **Fig 7**). Hooks engaged this way are easier to unhook deliberately but won't get squeezed open by a part of the tower or engage and snag a part of the tower while you are climbing. The engagement of these hooks should always be checked visually. A snapping hook makes the same sound whether it's engaged or not. Never check by sound—look to be sure the hook is engaged properly before trusting it.

Remember that the D rings on the safety belt are for support hooks *only*. No tools or lines should be attached to these hooks. Such tools or lines may prevent the proper engagement of support belt hooks, or they may foul the hooks. At best, they could prevent the release of the hooks in an emergency. No one should have to disconnect a support hook to get a tool and then have to reconnect the support hook before beginning to work again. That's foolish.

Equipment you purchase new is best. Homemade belts or home-spliced lines are dangerous. Used belts may have worn or defective stitching, or other faulty components. Be careful of so-called "bargains" that could cost you your life.

Straps, lanyards and lines should be as short as possible. Remember, in general knots reduce the load strength of a line by approximately 50%.

Before actually climbing, check the structure visually. Review the route. Check for obstacles, both natural (like wasp's nests) and man-made. Check the structure supports and add more if necessary. Guy wires can be obstacles to the climb, but it's better to have too many supports than not enough. Check your safety belt, support belts and hooks at the base of the tower. Really test them before you need them. Never leave the ground without a safety belt—even 5 or 10 feet. After all of this, the climb will be a "cake walk" if you are careful.

Climb slowly and surely. Don't overreach or overstep. Patience and watchfulness is rewarded with good hand and foot holds. Take a lesson from rock climbers. Hook on to the tower and rest periodically during the climb. Don't try to rest by wedging an arm or leg in some joint; to rest, hook on. Rests provide an opportunity to review the remainder of the route and to make sure that your safety equipment feels good and is working properly. Rest periods also help you conserve a margin of energy in case of difficulty.

Finally, keep in mind that the most dangerous part of working on a tower occurs when you are actually climbing. Your safety equipment is not hooked up at this time, so be extra careful during the ascent or descent.

You must climb the tower to install or work on an antenna. Nevertheless, any work that can be done on the ground should be done there. If you can do any assembly or make any adjustments on the ground, that's where you should do the work! The less time you have to spend on the tower, the better off you'll be.

When you arrive at the work area, hook on to the tower and review what you have to do. Determine the best position to do the work from, disconnect your safety strap and move to that position. Then reconnect your safety strap at a safe spot, away from joints and other obstacles. If you must move around an obstacle, try to do it while hooked on to the tower. Find a comfortable position and go to work. Don't overreach—move to the work.

Use the right tool for the task. If you don't have it, have the ground crew haul it up. Be patient. Lower tools, don't drop them, when you are finished with them. Dropped tools can bounce and cause injury or damage, or can be broken or lost. It's a good idea to tie a piece of string or light rope to the tools, and to tie the other end to the tower or some other point so if you do drop a tool, it won't fall all the way to the ground. Don't tie tools to the D ring or your safely belt, however!

Beware of situations where an antenna may be off balance. It's hard to obtain the extra leverage needed to handle even a small beam when you are holding it far from the balance point. Leverage can apply to the climber as well as the device being levered. Many slips and skinned knuckles result from such situations. A severely injured hand or finger can be a real problem to a climber.

Before descending, be sure to check all connections and the tightness of all the bolts and nuts that you have worked with. Have the ground crew use the rope and pulley to lower your tools. Lighten your load as much as possible. Remember, you're more tired coming down than going up. While still hooked on, wiggle your toes and move a little to get your senses working again. Check your downward route and begin to descend slowly and even more surely than you went up. Rest is even more important during the descent.

The ground captain is the director of all activities on the ground, and should be the only one to communicate with a person on a tower. Hand-held transceivers can be very helpful for this communication, but no one else should transmit to the workers on the tower. Even minor confusion or misunderstanding about a move to be made could be very dangerous.

"Antenna parties" can be lots of fun, but the joking and fooling around should wait until the job is done and everyone is down safely. Save the celebrating until after the work is completed, even for the ground crew.

These are just a few ideas on tower climbing safety; no list can include everything that you might run into. Check Chapter 22 for additional ideas. *Just remember—you can't be too careful when climbing*. Keep safety in mind while doing antenna work, and help ensure that after you have fallen *for* ham radio, you don't fall *from* ham radio.

THE TOWER SHIELD

A tower can be legally classified as an "attractive nuisance" that could cause injuries. You should take some precautions to ensure that "unauthorized climbers" can't get hurt on your tower. This tower shield was originally described by Baker Springfield, W4HYY, and Richard Ely, WA4VHM, in September 1976 *QST*, and should eliminate the worry.

Generally, the attractive-nuisance doctrine applies to your responsibility to trespassers on your property. (The law is much stricter with regard to your responsibility to an invited guest.) You should expect your tower to attract children, whether they are already technically trespassing

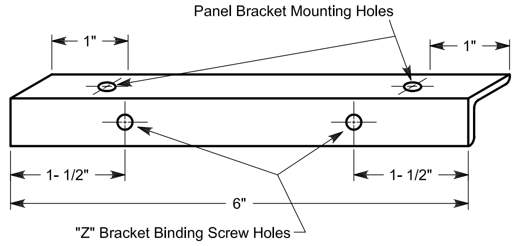

Fig 8—Z-bracket component pieces.

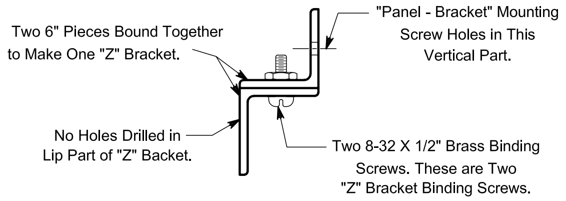
Two 6" Pieces Bound Together
to Make One "Z" Bracket.
"Panel - Bracket" Mounting
Screw Holes in This
Vertical Part.
No Holes Drilled in
Lip Part of "Z" Backet.
Two 8-32 X 1/2" Brass Binding
Screws. These are Two
"Z" Bracket Binding Screws.

Fig 9—Assembly of the Z bracket.

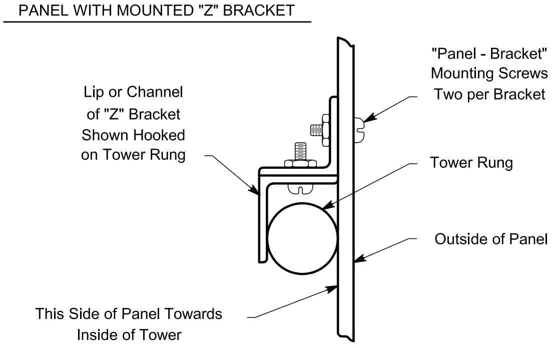
PANEL WITH MOUNTED "Z" BRACKET
"Panel - Bracket"
Mounting Screws
Two per Bracket
Lip or Channel
of "Z" Bracket
Shown Hooked
on Tower Rung
Tower Rung
Outside of Panel
This Side of Panel Towards
Inside of Tower

Fig 10—Installation of the shield on a tower rung.

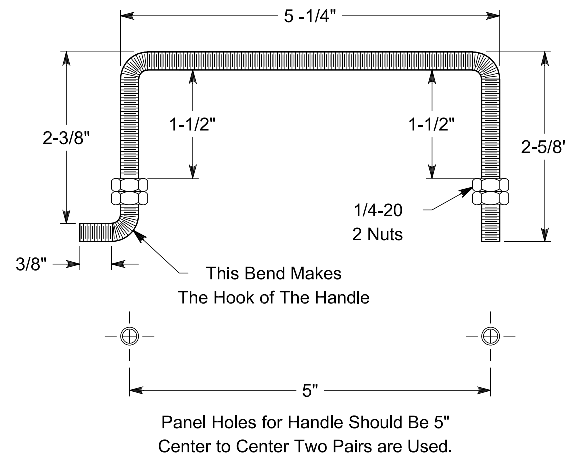
5 -1/4"
2-3/8"
1-1/2"
1-1/2"
2-5/8"
1/4-20
2 Nuts
3/8"
This Bend Makes
The Hook of The Handle
5"
Panel Holes for Handle Should Be 5"
Center to Center Two Pairs are Used.

Notes:

1. Standard Rodstock 1/4"-20 X 36" was Used.
2. Two Pieces were Cut from Rod Stock.
   Each Approximately 10 7/8".
3. Make The Three Bends of The Rod in Vise.
4. Two Nuts 1/4-20 of The Same Threads as
   Rods. Jam or Lock Together.This Makes
   a Handle Stop.

Fig 11—Removable handle construction.

or whether the tower itself lures them onto your property. A tower is dangerous to children, especially because of their inability to appreciate danger. (What child could resist trying to climb a tower once they see one?) Because of this danger, you have a legal duty to exercise reasonable care to eliminate the danger or otherwise protect children against the perils of the attraction.

The tower shield is composed simply of panels that enclose the tower and make climbing practically impossible. These panels are 5 feet in height and are wide enough to fit snugly between the tower legs and flat against the rungs. A height of 5 feet is sufficient in almost every case. The panels are constructed from 18-gauge galvanized sheet metal obtained and cut to proper dimensions from a local sheet-metal shop. A lighter gauge could probably be used, but the extra physical weight of the heavier gauge is an advantage if no additional means of securing the panels to the tower rungs are used. The three types of metals used for the components of the shield are supposedly rust proof and nonreactive. The panels are galvanized sheet steel, the brackets aluminum, and the screws and nuts are brass. For a triangular tower, the shield consists of three panels, one for each of the three sides, supported by two brackets. Construct these brackets from 6-inch pieces of thin aluminum angle stock. Bolt two of the pieces together to form a Z bracket (see **Figs 8**, **9** and **10**). The Z brackets are bolted together with binding head brass machine screws.

Lay the panels flat for measuring, marking and drilling. First measure from the top of the upper mounting rung on the tower to the top of the bottom rung. (Mounting rungs are selected to position the panel on the tower.) Then mark this distance on the panels. Use the same size brass screws and nuts throughout the shield. Bolt the top vertical portion of each Z bracket to the panel. Drill the mounting-screw holes about 1 inch from the end of the Z

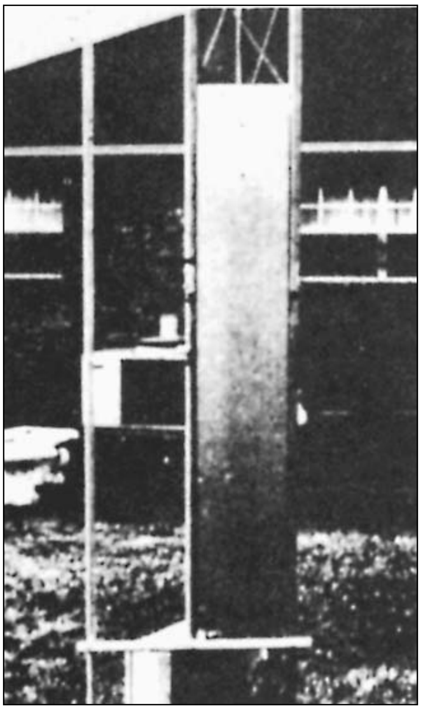

Fig 12—Installed tower shield. Note the holes for using the handles.

brackets so there is an offset clearance between the Z-bracket binding-screw holes and the panel-bracket mounting-screw holes. Drill holes in each panel to match the Z-bracket holes.

The panels are held on the tower by their own weight. They are not easy to grasp because they fit snugly between the tower legs. If you feel a need for added safety against deliberate removal of the panels, this can be accomplished by means of tie wires. Drill a small hole in the panel just above, just below, and in the center of each Z bracket. Run a piece of heavy galvanized wire through the top hole, around the Z bracket, and then back through the hole just below the Z bracket. Twist together the two ends of the wire. One tie wire should be sufficient for each panel, but use two if desired.

The completed panels are rather bulky and difficult to handle. A feature that is useful if the panels have to be removed often for tower climbing or accessibility is a pair of removable handles. The removable handles can be constructed from one threaded rod and eight nuts (see **Fig 11**). Drill two pair of handle holes in the panels a few inches below the top Z bracket and several inches above the bottom Z bracket. For panel placement or removal, you can hook the handles in these panel holes. The hook, on the top of the handle, fits into the top hole of each pair of the handle holes. The handle is optional, but for the effort required it certainly makes removal and replacement much safer and easier.

Fig 12 shows the shield installed on a tower. This relatively simple device could prevent an accident.

Electrical Safety

Although the RF, ac and dc voltages in most amateur stations pose a potentially grave threat to life and limb, common sense and knowledge of safety practices will help you avoid accidents. Building and operating an Amateur Radio station can be, and is for almost all amateurs, a perfectly safe pastime. However, carelessness can lead to severe injury, or even death. The ideas presented here are only guidelines; it would be impossible to cover all safety precautions. Remember, there is no substitute for common sense.

A fire extinguisher is a requirement for the well-equipped amateur station. The fire extinguisher should be of the carbon-dioxide type to be effective in electrical fires. Store it in an easy-to-reach spot and check it at recommended intervals.

Family members should know how to turn the power off in your station. They should also know how to apply artificial respiration. Many community groups offer courses on cardiopulmonary resuscitation (CPR).

AC AND DC SAFETY

The primary wiring for your station should be controlled by one master switch, and other members of your household should know how to kill the power in an emergency. All equipment should be connected to a good ground. All wires carrying power around the station should be of the proper size for the current to be drawn and should be insulated for the voltage level involved. Bare wire, open-chassis construction and exposed connections are an invitation to accidents. Remember that high-current, low-voltage power sources are just as dangerous as high-voltage, low-current sources. Possibly the most-dangerous voltage source in your station is the 120-V primary supply; it is a hazard often overlooked because it is a part of everyday life. Respect even the lowliest power supply in your station.

Whenever possible, kill the power and unplug equipment before working on it. Discharge capacitors with an insulated screwdriver; don’t assume the bleeder resistors are 100% reliable. In a power amplifier, always short the tube plate cap to ground just to be sure the supply is discharged. If you must work on live equipment, keep one hand in your pocket. Avoid bodily contact with any grounded object to prevent your body from becoming the return path from a voltage source to ground. Use insulated tools for adjusting or moving any circuitry. Never work alone. Have someone else present; it could save your life in an emergency.

National Electrical Code

The National Electrical Code® is a comprehensive document that details safety requirements for all types of electrical installations. In addition to setting safety standards for house wiring and grounding, the Code also contains a section on Radio and Television Equipment — Article 810. Sections C and D specifically cover Amateur Transmitting and Receiving Stations. Highlights of the section concerning Amateur Radio stations follow. If you are interested in learning more about electrical safety, you may purchase a copy of *The National Electrical Code* or *The National Electrical Code Handbook*, edited by Peter Schram, from the National Fire Protection Association, Batterymarch Park, Quincy, MA 02269.

Antenna installations are covered in some detail in the Code. It specifies minimum conductor sizes for different length wire antennas. For hard-drawn copper wire, the Code specifies #14 wire for open (unsupported) spans less than 150 feet, and #10 for longer spans. Copper-clad steel, bronze or other high-strength conductors may be #14 for spans less than 150 feet and #12 wire for longer runs. Lead-in conductors (for open-wire transmission line) should be at least as large as those specified for antennas.

The Code also says that antenna and lead-in conductors attached to buildings must be firmly mounted at least 3 inches clear of the surface of the building on nonabsorbent insulators. The only exception to this minimum distance is when the lead-in conductors are enclosed in a “permanently and effectively grounded” metallic shield. The exception covers coaxial cable.

According to the Code, lead-in conductors (except those covered by the exception) must enter a building through a rigid, noncombustible, nonabsorbent insulating tube or bushing, through an opening provided for the purpose that provides a clearance of at least 2 inches or through a drilled window pane. All lead-in conductors to transmitting equipment must be arranged so that accidental contact is difficult.

Transmitting stations are required to have a means of draining static charges from the antenna system. An antenna discharge unit (lightning arrester) must be installed on each lead-in conductor (except where the lead-in is protected by a continuous metallic shield that is permanently and effectively grounded, or the antenna is permanently and effectively grounded). An acceptable alternative to lightning arrester installation is a switch that connects the lead-in to ground when the transmitter is not in use.

Grounding conductors are described in detail in the Code. Grounding conductors may be made from copper, aluminum, copper-clad steel, bronze or similar erosion-resistant material. Insulation is not required. The “protective grounding conductor” (main conductor running to the ground rod) must be as large as the antenna lead-in, but not smaller than #10. The “operating grounding conductor” (to bond equipment chassis together) must be at least #14. Grounding conductors must be adequately supported and arranged so they are not easily damaged. They must run in as straight a line as practical between the mast or discharge unit and the ground rod.

The Code also includes some information on safety inside the station. All conductors inside the building must be at least 4 inches away from conductors of any lighting or signaling circuit except when they are separated from other conductors by conduit or a nonconducting material. Transmitters must be enclosed in metal cabinets, and the cabinets must be grounded. All metal handles and controls accessible by the operator must be grounded. Access doors must be fitted with interlocks that will disconnect all potentials above 350 V when the door is opened.

Ground

An effective ground system is necessary for every amateur station. The mission of the ground system is twofold. First, it reduces the possibility of electrical shock if something in a piece of equipment should fail and the chassis or cabinet becomes “hot.” If connected properly, three-wire electrical systems ground the chassis, but older amateur equipment may use the ungrounded two-wire system. A ground system to prevent shock hazards is generally referred to as “dc ground.”

The second job the ground system must perform is to provide a low-impedance path to ground for any stray RF current inside the station. Stray RF can cause equipment to malfunction and contributes to RFI problems. This low-impedance path is usually called “RF ground.” In most stations, dc ground and RF ground are provided by the same system.

The first step in building a ground system is to bond together the chassis of all equipment in your station. Ordinary hookup wire will do for a dc ground, but for a good RF ground you need a low-impedance conductor. Copper strap, sold as “flashing copper,” is excellent for this application, but it may be hard to find. Braid from coaxial cable is a popular choice; it is readily available, makes a low-impedance conductor, and is flexible.

Grounding straps can be run from equipment chassis to equipment chassis, but a more convenient approach is illustrated in **Fig 13**. In this installation, a $^1/_2$-inch diameter copper water pipe runs the entire length of the operating bench. A thick braid (from discarded RG-8 cable) runs from each piece of equipment to a clamp on the pipe. Copper water pipe is available at most hardware stores and home centers. Alternatively, a strip of flashing copper may be run along the rear of the operating bench.

After the equipment is bonded to a common ground bus, the ground bus must be wired to a good earth ground. This run should be made with a heavy conductor (braid is a popular choice, again) and should be as short and direct as possible. The earth ground usually takes one of two forms.

In most cases, the best approach is to drive one or more ground rods into the earth at the point where the conductor from the station ground bus leaves the house. The best ground rods to use are those available from an electrical supply house. These rods are 8 to 10 feet long and are made from steel with a heavy copper plating. Do not depend on shorter, thinly plated rods sold by some home electronics suppliers. These rods begin to rust almost immediately after they are driven into the soil, and they become worthless within a short time. Good ground rods, while more expensive initially, offer long-term protection.

If your soil is soft and contains few rocks, an acceptable alternative to “genuine” ground rods is $^1/_2$-inch diameter copper water pipe. A 6- to 8-foot length of this material offers a good ground, but it may bend while being driven into the earth. Some people have recommended that you make a connection to a water line and run water down through the copper pipe so that it forces its own

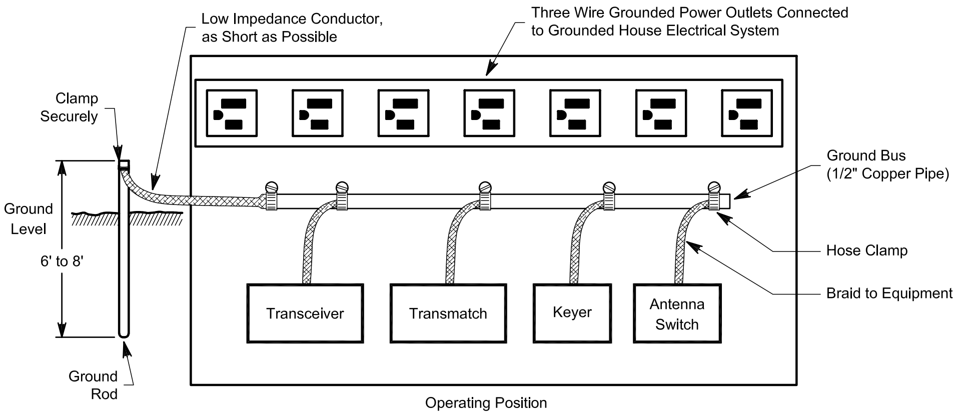
Low Impedance Conductor, as Short as Possible
Three Wire Grounded Power Outlets Connected to Grounded House Electrical System
Clamp Securely
Ground Bus (1/2" Copper Pipe)
Ground Level
6' to 8'
Hose Clamp
Braid to Equipment
Transceiver
Transmatch
Keyer
Antenna Switch
Ground Rod
Operating Position

Fig 13—An effective station ground bonds the chassis of all equipment together with low-impedance conductors and ties into a good earth ground.

hole in the ground. There may be a problem with this method, however. When the ground dries, it may shrink away from the pipe and not make proper contact with the ground rod. This would provide a rather poor ground.

Once the ground rod is installed, clamp the conductor from the station ground bus to it with a clamp that can be tightened securely and will not rust. Copper-plated clamps made especially for this purpose are available from electrical supply houses, but a stainless-steel hose clamp will work too. Alternatively, drill several holes through the pipe and bolt the conductor in place.

Another popular station ground is the cold water pipe system in the building. To take advantage of this ready-made ground system, run a low-impedance conductor from the station ground bus to a convenient cold water pipe, preferably somewhere near the point where the main water supply enters the house. Avoid hot water pipes; they do not run directly into the earth. The advent of PVC (plastic) plumbing makes it mandatory to inspect the cold water system from your intended ground connection to the main inlet. PVC is an excellent insulator, so any PVC pipe or fittings rule out your cold water system for use as a station ground.

For some installations, especially those located above the first floor, a conventional ground system such as that just described will make a fine dc ground but will not provide the necessary low-impedance path to ground for RF. The length of the conductor between the ground bus and the ultimate ground point becomes a problem. For example, the ground wire may be about $^1/_4$ λ (or an odd multiple of $^1/_4$ λ) long on some amateur band. A $^1/_4$-λ wire acts as an impedance inverter from one end to the other. Since the grounded end is at a very low impedance, the equipment end will be at a high impedance. The likely result is RF hot spots around the station while the transmitter is operating. A ground system like this may be worse than having no ground at all.

An alternative RF ground system is shown in **Fig 14**. Connect a system of $^1/_4$-λ radials to the station ground bus. Install at least one radial for each band used. You should still be sure to make a connection to earth ground for the ac power wiring. Try this system if you have problems with RF in the shack. It may just solve a number of problems for you. Be careful, however, to prevent contact with the ends of the radials, where there is high-voltage RF for powers greater than QRP level.

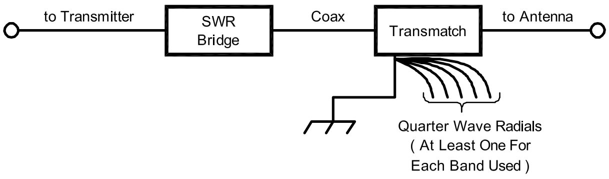
to Transmitter
SWR Bridge
Coax
Transmatch
to Antenna
Quarter Wave Radials (At Least One For Each Band Used)

Fig 14—Here is an alternative to earth ground if the station is located far from the ground point and RF in the station is a problem. Install at least one $^1/_4$-λ radial for each band used.

Ground Noise

Noise in ground systems can affect sensitive radio equipment. It is usually related to one of three problems:

1) Insufficient ground conductor size,
2) Loose ground connections, or
3) Ground loops.

These matters are treated in precise scientific research equipment and some industrial instruments by paying attention to certain rules. The ground conductor should be at least as large as the largest conductor in the primary power circuit. Ground conductors should provide a solid connection to both ground and to the equipment being grounded. Liberal use of lock washers and star washers is highly recommended. A loose ground connection is a tremendous source of noise, particularly in a sensitive receiving system.

Ground loops should be avoided at all costs. A short discussion of what a ground loop is and how to avoid them may lead you down the proper path. A ground loop is formed when more than one ground current is flowing in a single conductor. This commonly occurs when grounds are "daisy-chained" (series linked). The correct way to ground equipment is to bring all ground conductors out radially from a common point to either a good driven earth ground or to a cold water system.

Ground noise can affect transmitted as well as received signals. With the low audio levels required to drive amateur transmitters, and with the ever-increasing sensitivity of our receivers, correct grounding is critical.

Lightning and EMP Protection

The National Fire Protection Association (NFPA) publishes a booklet called *Lightning Protection Code* (NFPA no. 78-1983) that should be of interest to radio amateurs. For information about obtaining a copy of this booklet, write to the National Fire Protection Association, Batterymarch Park, Quincy, MA 02269. Two paragraphs of particular interest to amateurs are presented below:

"3-26 Antennas. Radio and television masts of metal, located on a protected building, shall be bonded to the lightning protection system with a main size conductor and fittings.

"3-27 Lightning arresters, protectors or antenna discharge units shall be installed on electric and telephone service entrances and on radio and television antenna lead-ins."

The best protection from lightning is to disconnect all antennas from equipment and disconnect the equipment from the power lines. Ground antenna feed lines to safely bleed off static buildup. Eliminate the possible paths for lightning strokes. Rotator cables and other control cables from the antenna location should also be disconnected during severe electrical storms.

In some areas, the probability of lightning surges entering homes via the 120/240-V line may be high. Lightning produces both electrical and magnetic fields that vary with distance. These fields can be coupled into power lines and destroy electronic components in equipment that is miles from where the lightning occurred. Radio equipment can be protected from these surges to some extent by using transient-protective devices.

ELECTROMAGMETIC PULSE AND THE RADIO AMATEUR

The following material is based on a 4-part *QST* article by Dennis Bodson, W4PWF, that appeared in the August through November 1986 issues of *QST*. The series was condensed from the National Communications System report NCS TIB 85-10.

An equipment test program demonstrated that most Amateur Radio installations can be protected from lightning and electromagnetic pulse (EMP) transients with a basic protection scheme. Most of the equipment is not susceptible to damage when all external cabling is removed. You can duplicate this stand-alone configuration simply by unplugging the ac power cord from the outlet, disconnecting the antenna feed line at the rear of the radio, and isolating the radio gear from any other long metal conductors. Often it is not practical to completely disconnect the equipment whenever it is not being used. Also, there is the danger that a lightning strike several miles away could induce a large voltage transient on the power lines or antenna while the radio is in use. You can add two transient-protection devices to the interconnected system, however, that will also closely duplicate the safety of the stand-alone configuration.

The ac power line and antenna feed line are the two important points that should be outfitted with transient protection. This is the minimum basic protection scheme recommended for all Amateur Radio installations. (For fixed installations, consideration should also be given to the rotator connections—see **Fig 15**.) Hand-held radios equipped with a "rubber duck" require no protection at the antenna jack. If a larger antenna is used with the hand-held, however, a protection device should be installed.

General Considerations

Because of the unpredictable energy content of a nearby lightning strike or other large transient, it is possible for a metal-oxide varistor (MOV) to be subjected to an energy surge in excess of its rated capabilities. This may result in the destruction of the MOV and explosive rupture of the package. These fragments can cause damage to nearby components or operators and possibly ignite flammable material. Therefore, the MOV should be physically shielded.

A proper ground system is a key factor in achieving protection from lightning and EMP transients. A low-impedance ground system should be installed to eliminate transient paths through radio equipment and to provide a good physical ground for the transient-suppression devices. A single-point ground system is recommended (**Fig 16**) to help prevent lightning from getting into the shack on the shields of antenna feed-line coaxes. Many hams use a well-grounded radio-entrance panel mounted *outside* the shack to ground their coaxes before they enter the house. **Fig 17** shows an entrance panel at K8CH's home in Michigan. All external conductors going to the radio equipment should enter and exit the station through this panel. Install all transient-suppression devices directly on the panel. Use

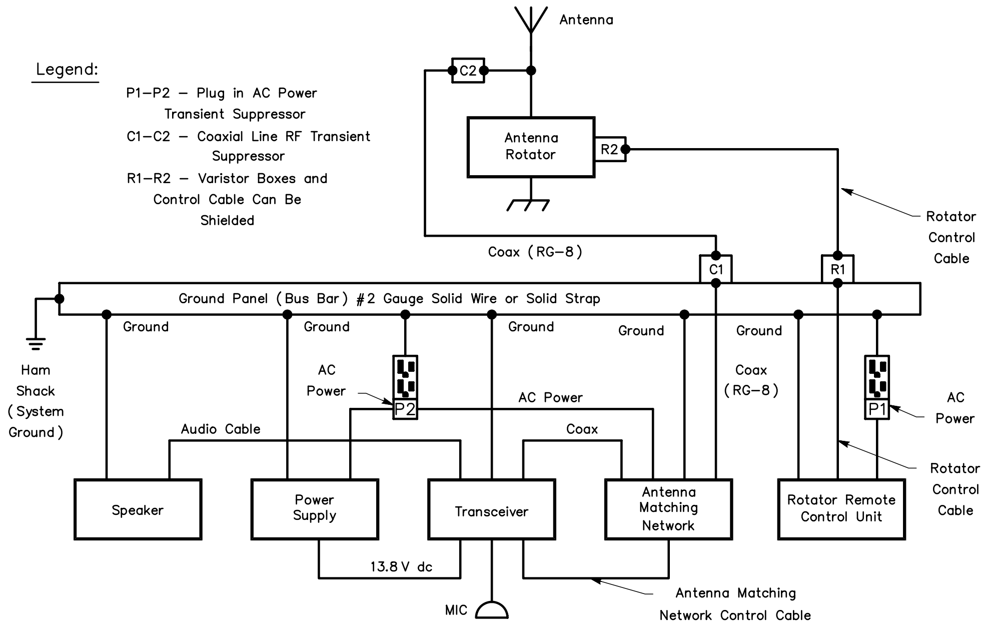
Antenna
Legend:
P1–P2 – Plug in AC Power Transient Suppressor
C1–C2 – Coaxial Line RF Transient Suppressor
R1–R2 – Varistor Boxes and Control Cable Can Be Shielded
C2
Antenna Rotator
R2
Rotator Control Cable
Coax (RG–8)
C1
R1
Ground Panel (Bus Bar) #2 Gauge Solid Wire or Solid Strap
Ground
Ground
Ground
Ground
Ground
Ham Shack (System Ground)
AC Power
P2
AC Power
Coax (RG–8)
P1
AC Power
Audio Cable
Coax
Rotator Control Cable
Speaker
Power Supply
Transceiver
Antenna Matching Network
Rotator Remote Control Unit
13.8 V dc
MIC
Antenna Matching Network Control Cable

Fig 15—Transient suppression techniques applied to an Amateur Radio station.

the shortest length(s) possible of #6 solid wire to connect the radio equipment case(s) to the ground bus.

Ac Power-Line Protection

Tests have indicated that household electrical wiring inherently limits the maximum transient current that it will pass to approximately 120 A. Therefore, if possible, the amateur station should be installed away from the house ac entrance panel and breaker box to take advantage of these limiting effects.

Ac power-line protection can be provided with easy-to-install, plug-in transient protectors. Ten such devices were tested for the article series in 1986. The plug-in-strip units are the best overall choice for a typical amateur installation. They provide the protection needed, they're simple to install and can be easily moved to other operating locations with the equipment.

In their tests, NCS found that the model that provided transient paths to ground from the hot and neutral lines (common mode) as well as the transient path between the hot and neutral lines (normal mode) performed best. The best model used three MOVs and a 3-electrode gas-discharge-tube arrester to provide fast operation and large power dissipation capabilities. This unit was tested repeatedly and operated without failure.

The flood of low-cost computers in the 1990s spawned a host of surge protection devices designed to limit transient voltage spikes coming from the ac line and also through the telephone line into a modem connected to a computer. Many of these devices are well-designed and can be relied upon to provide the protection they claim.

You can, however, easily find a variety of really low-cost bargain strips at flea markets and discount hardware stores. A bargain-brand $6 unit may prove to be a poor bargain indeed if it allows a spike to get through to damage your $2000 computer or $4000 transceiver.

You should be careful to find one that carries a sticker indicating that it meets Underwriters Laboratories safety standard UL 1449. This defines the minimum level of clamping voltage beyond which a surge protector will "fire" to protect the device connected to its output. The UL 1149 limit is 330 V ac. Prices for brand-name units from Tripp Lite, APC or Curtis vary from about $30 to $80, depending on how many ac sockets they have and the number of indicator lights and switched/unswitched sockets. A brand-name device is well worth the small additional cost over the bargain-basement units.

A transient suppressor requires a 3-wire outlet; the outlet should be tested to ensure all wires are properly connected. In older houses, an ac ground may have to be installed by a qualified electrician. The ac ground must be available for the plug-in transient suppressor to function properly. The ac ground of the receptacle should be attached to the station ground bus, and the plug-in receptacle should be installed on the ground panel behind the radio equipment.

Emergency Power Generators

Emergency power generators provide two major transient-protection advantages. First, the station is disconnected from the commercial ac power system. This isolates the radio equipment from a major source of damaging transients. Second, tests have shown that the emergency power generator may not be susceptible to EMP transients.

When the radio equipment is plugged directly into the generator outlets, transient protection may not be

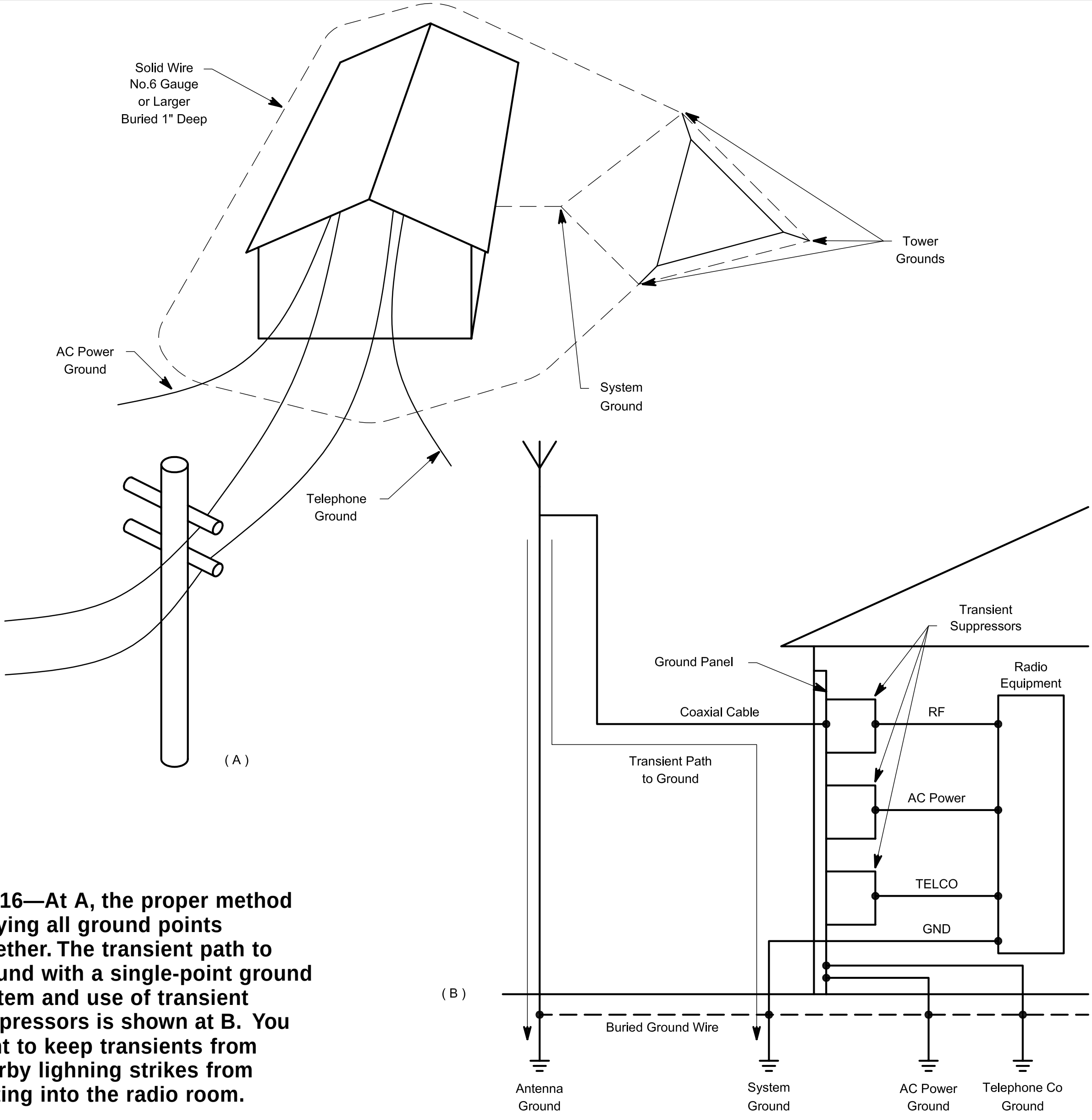
Solid Wire
No.6 Gauge
or Larger
Buried 1" Deep
Tower
Grounds
AC Power
Ground
System
Ground
Telephone
Ground
(A)
Transient
Suppressors
Ground Panel
Radio
Equipment
Coaxial Cable
RF
Transient Path
to Ground
AC Power
TELCO
GND
(B)
Buried Ground Wire
Antenna
Ground
System
Ground
AC Power
Ground
Telephone Co
Ground

Fig 16—At A, the proper method of tying all ground points together. The transient path to ground with a single-point ground system and use of transient suppressors is shown at B. You want to keep transients from nearby lightning strikes from getting into the radio room.

needed. If an extension cord or household wiring is used, transient protection should be employed.

An emergency power generator should be wired into the household circuit only by a qualified electrician. When properly connected, a switch is used to disconnect the commercial ac power source from the house lines before the generator is connected to them. This keeps the generator output from feeding back into the commercial power system. If this is not done, death or injury to unsuspecting linemen can result.

Feed-Line Protection

Coaxial cable is recommended for use as the transmission line because it provides a certain amount of transient surge protection for the equipment to which it is attached. The outer conductor shields the center conductor from the transient field. Also, the cable limits the maximum conducted transient voltage on the center by arcing the differential voltage from the center conductor to the grounded cable shield.

By providing a path to ground ahead of the radio equipment, the gear can be protected from the large currents impressed upon the antenna system by lightning and EMP. A single protection device installed at the radio antenna jack will protect the radio, but not the transmission line. To protect the transmission line, another tran-

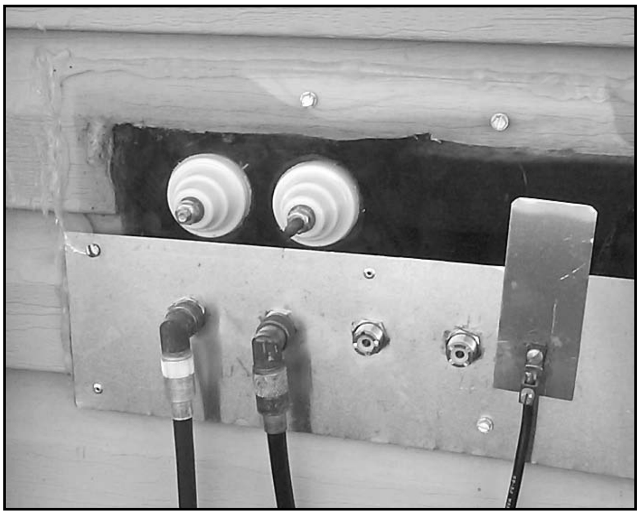

Fig 17—Radio-entrance panel at K8CH. A flat aluminum plate serves as a common grounding point for all coax cables to prevent transients from nearby lightning strikes from getting into the shack on the feed lines. "Spark-gap" protectors to kill transients on the coax inner conductors are located behind the panel. A spark-gap protector is also used on the two open-wire feeder insulators mounted on the Plexiglas sheet behind the aluminum panel.

Table 1
RF Coaxial-Line Protectors

Manufacturer	*Device*	*Approximate Cost (US Dollars)*	*Measured High-Z Clamping Voltage (Volts)*
Fischer	FCC-250-300-UHF	55	393
Fischer	FCC-250-350-UHF	55	260
Fischer	FCC-250-150-UHF	55	220
Fischer	FCC-250-120-UHF	55	240
Fischer	FCC-450-120-UHF	55	120
Polyphaser	IS-NEMP	83	140
Polyphaser	IS-NEMP-1	83	150
Polyphaser	IS-NEMP-2	83	160

Note: The transmitter output power, frequency of operation, and transmission line SWR must be considered when selecting any of these devices.

sient protector must be installed between the antenna and the transmission line. (See Fig 15).

RF transient protection devices from several manufacturers were tested (see **Table 1**) using RG-8 cable equipped with UHF connectors. All of the devices shown can be installed in a coaxial transmission line. Recall that during the tests the RG-8 cable acted like a suppressor; damaging EMP energy arced from the center conductor to the cable shield when the voltage level approached 5.5 kV.

Low price and a low clamping-voltage rating must be considered in the selection of an RF transient-protection device. The lower cost devices have the higher clamping voltages, however, and the higher-cost devices have the lower clamping voltages. Because of this, medium-priced devices manufactured by Fischer Custom Communications were selected for testing. The Fischer Spikeguard Suppressors ($55 price class) for coaxial lines can be made to order to operate at a specific clamping voltage. The Fischer devices satisfactorily suppressed the damaging transient pulses, passed the transmitter RF output power without interfering with the signal, and operated effectively over a wide frequency range.

Polyphaser Corporation devices are also effective in providing the necessary transient protection. The devices available limited the transmitter RF output power to 100 W or less, however. These units cost approximately $83 each.

RF coaxial protectors should be mounted on the station ground bus bar. If the Fischer device is used, it should be attached to a grounded UHF receptacle that will serve as a hold-down bracket. This creates a conductive path between the outer shield of the protector and the bus bar. The Polyphaser device can be mounted directly to the bus bar with the bracket provided.

Attach the transceiver or antenna matching network to the grounded protector with a short (6 foot or less) piece of coaxial cable. Although the cable provides a ground path to the bus bar from the radio equipment, it is not a satisfactory transient-protection ground path for the transceiver. Another ground should be installed between the transceiver case and the ground bus using solid #6 wire. The coaxial cable shield should be grounded to the antenna tower leg at the tower base. Each tower leg should have an earth ground connection and be connected to the single-point ground system as shown in Fig 16.

Antenna Rotators

Antenna rotators can be protected by plugging the control box into a protected ac power source and adding protection to the control lines to the antenna rotator. When the control lines are in a shielded cable, the shield must be grounded at both ends. MOVs of the proper size should be installed at both ends of the control cable. At the station end, terminate the control cable in a small metal box that is connected to the station ground bus. Attach MOVs from each conductor to ground inside the box. At the antenna end of the control cable, place the MOVs inside the rotator case or in a small metal box that is properly grounded.

For example, the Alliance HD73 antenna rotator uses a 6-conductor unshielded control cable with a maximum control voltage of 24.7 V dc. Select an MOV with a clamping voltage level 10% higher (27 V or more) so the MOV won't clamp the control signal to ground. If the control voltage is ac, be sure to convert the RMS voltage value to peak voltage when considering the clamping voltage level.

Mobile Power Supply Protection

The mobile amateur station environment exposes radio equipment to other transient hazards in addition to those of lightning and EMP. Currents as high as 300 A are switched when starting the engine, and this can produce voltage spikes of over 200 V on the vehicle's electrical system. Lightning and EMP are not likely to impact the vehicle's electrical system as much as they would that of a fixed installation because the automobile chassis is not normally grounded. This would not be the case if the vehicle is inadvertently grounded; for example, when the vehicle is parked against a grounded metal conductor. The mobile radio system has two advantages over a fixed installation: Lightning is almost never a problem, and the vehicle battery is a natural surge suppressor.

Mobile radio equipment should be installed in a way that takes advantage of the protection provided by the battery. See **Fig 18**. To do this, connect the positive power lead of the radio directly to the positive battery post, not to intermediate points in the electrical system such as the fuse box or the auxiliary contacts on the ignition switch. To prevent equipment damage or fire, in-line fuses should be installed in the positive lead where they are attached to the battery post.

An MOV should be installed between the two leads of the equipment power cord. A GE MOV (V36ZA80) is recommended for this application. This MOV provides the lowest measured clamping voltage (170 V) and is low in cost.

Mobile Antenna Installation

Although tests indicate that mobile radios can survive an EMP transient without protection for the antenna system, protection from lightning transients is still required. A coaxial-line transient suppressor should be installed on the vehicle chassis between the antenna and the radio's antenna connector.

A Fischer suppressor can be attached to a UHF receptacle that is mounted on, and grounded to, the vehicle chassis. The Polyphaser protector can be mounted on, and grounded to, the vehicle chassis with its flange. Use a short length of coaxial cable between the radio and the transient suppressor.

Clamping Voltage Calculation

When selecting any EMP-protection device to be used at the antenna port of a radio, several items must be considered. These include transmitter RF power output, the SWR, and the operating frequency. The protection device must allow the outgoing RF signal to pass without clamping. A clamping voltage calculation must be made for each amateur installation.

The RF-power input to a transmission line develops a corresponding voltage that becomes important when a voltage-surge arrester is in the line. SWR is important because of its influence on the voltage level. The maximum voltage developed for a given power input is determined by:

$$V = \sqrt{2 \times P \times Z \times SWR} \qquad \text{(Eq 1)}$$

where

P = peak power in W
Z = impedance of the coaxial cable (Ω)
V = peak voltage across the cable

Eq 1 should be used to determine the peak voltage present across the transmission line. Because the RF transient-protection devices use gas-discharge tubes, the voltage level at which they clamp is not fixed; a safety margin must be added to the calculated peak voltage. This is done

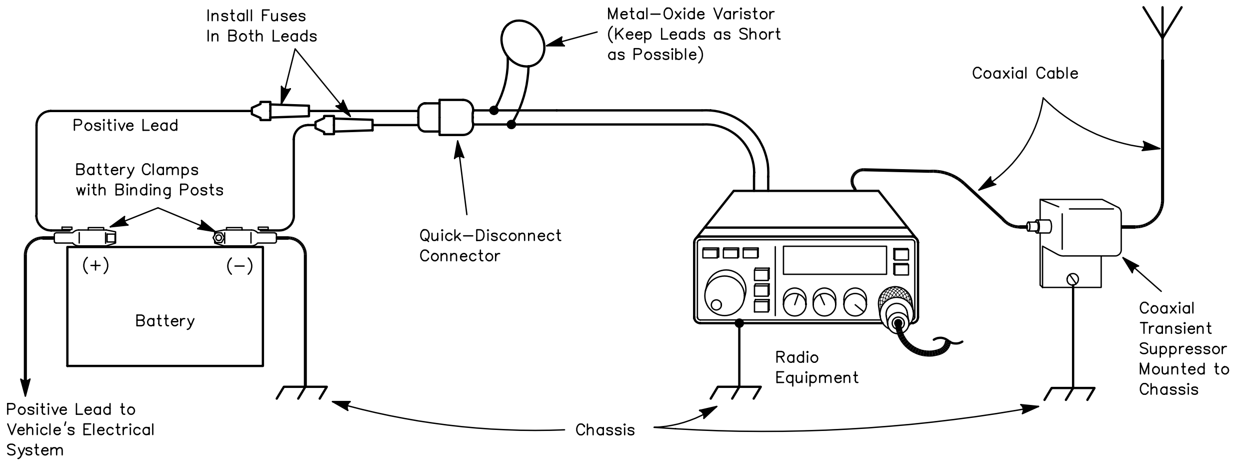
Install Fuses In Both Leads
Metal—Oxide Varistor (Keep Leads as Short as Possible)
Coaxial Cable
Positive Lead
Battery Clamps with Binding Posts
Quick—Disconnect Connector
(+)
(–)
Battery
Radio Equipment
Coaxial Transient Suppressor Mounted to Chassis
Positive Lead to Vehicle's Electrical System
Chassis

Fig 18—Recommended method of connecting mobile radio equipment to the vehicle battery and antenna.

by multiplying the calculated value by a factor of three. This added safety margin is required to ensure that the transmitter's RF output power will pass through the transient suppressor without causing the device to clamp the RF signal to ground. The final clamping voltage obtained is then high enough to allow normal operation of the transmitter while providing the lowest practical clamping voltage for the suppression device. This ensures the maximum possible protection for the radio system.

Here's how to determine the clamping voltage required. Let's assume the SWR is 1.5:1. The power output of the transceiver is 100 W PEP. RG/8 coaxial cable has an impedance of 52 Ω. Therefore:

P = 100 W
Z = 52 Ω
SWR = 1.5

Substituting these values into Eq 1:

$$V = \sqrt{2 \times 100 \times 52 \times 1.5} = 124.89 \text{ V peak}$$

Note that the voltage, V, is the peak value at the peak of the RF envelope. The final clamping voltage (FCV) is three times this value, or 374.7 V. Therefore, a coaxial-line transient suppressor that clamps at or above 375 V should be used.

The cost of a two-point basic protection scheme is estimated to be $100 for each fixed amateur station. This includes the cost of a good quality plug-in power-line protector ($45) and one Fischer coaxial-line protector ($55).

Inexpensive Transient-Protection Devices

Here are two low-cost protection devices you can assemble. They performed flawlessly in the tests.

The radio antenna connection can be protected by means of another simple device. As shown in **Fig 19**, two spark gaps (Siemens BI-A350) are installed in series at one end of a coaxial-cable T connector. Use the shortest practical lead length (about 1/4 inch) between the two spark gaps. One lead is bent forward and forced between the split sections of the inner coaxial connector until the spark gaps approach the body of the connector. A short length of insulating material (such as Mylar) is placed between the spark gaps and the connector shell. The other spark-gap lead is folded over the insulator, then conductive (metallic) tape is wrapped around the assembly. This construction method proved durable enough to allow many insertions and removals of the device during testing. Estimated cost of this assembly is $9. Similar devices can be built using components from Joslyn, General Electric, General Semiconductor or Siemens.

Summary

Amateurs should be aware of which components in their radio system are most likely to be damaged by EMP. They should also know how to repair the damaged equipment. Amateurs should know how to reestablish communications after an EMP event, taking into consideration its adverse effects on the earth's atmosphere and radio equipment. One of the first things that would be noticed, providing the radio equipment is operative, is a sudden silence in radio transmissions across all frequencies below approximately 100 MHz. This silence would be caused in part by damage to unprotected radio gear from the EMP transient. Transmissions from one direction, the direction of the nuclear blast, would be completely out. RF signal loss by absorption and attenuation by the nuclear fireball are the reasons for this.

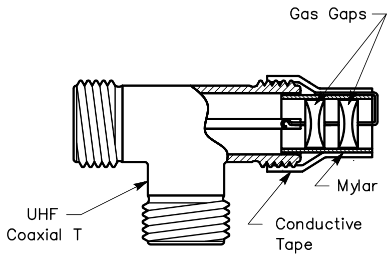

Fig 19—Pictorial diagram of an inexpensive, homemade RF coax transient protector.

After an EMP event, the amateur should be prepared to operate CW. CW gives the most signal power under adverse conditions. It also provides a degree of message security from the general public.

Amateurs should develop the capability and flexibility to operate in more than one frequency band. The lower ground-wave frequencies should be useful for long-distance communications immediately after an EMP event. Line-of-sight VHF would be of value for local communications.

What can be done to increase the survivability of an Amateur Radio station? Here are some suggestions:

1) If you have spare equipment, keep it disconnected; use only the primary station gear. The spare equipment would then be available after an EMP event.
2) Keep equipment turned off and antenna and power lines disconnected when the equipment is not in use.
3) Connect only those external conductors necessary for the current mode of operation.
4) Tie all fixed equipment to a single-point earth ground to prevent closed loops through the ground.
5) Obtain schematic diagrams of your equipment and tools for repair of the equipment.
6) Have spare parts on hand for sensitive components of the radio equipment and antenna system.
7) Learn how to repair or replace the sensitive components of the radio equipment.
8) Use nonmetallic guy lines and antenna structural parts where possible.
9) Obtain an emergency power source and operate from

it during periods of increased world political tension. The power source should be completely isolated from the commercial power lines.

10) Equipment power cords should be disconnected when the gear is idle. Or the circuit breaker for the line feeding the equipment should be kept in the OFF position when the station is off the air.
11) Disconnect the antenna lead-in when the station is off the air. Or use a grounding antenna switch and keep it in the GROUND position when the equipment is not in use.
12) Have a spare antenna and transmission line on hand to replace a damaged antenna system.
13) Install EMP surge arresters and filters on all primary conductors attached to the equipment and antenna.
14) Retain tube type equipment and spare components; keep them in good working order.
15) Do not rely on a microprocessor to control the station after an EMP event. Be able to operate without microprocessor control.

The recommendations contained in this section were developed with low cost in mind; they are not intended to cover all possible combinations of equipment and installation methods found in the amateur community. Amateurs should examine their own requirements and use this report as a guideline in providing protection for the equipment.

RF Radiation and Electromagnetic Field Safety

Amateur Radio is basically a safe activity. In recent years, however, there has been considerable discussion and concern about the possible hazards of electromagnetic radiation (EMR), including both RF energy and power-frequency (50-60 Hz) electromagnetic (EM) fields. FCC regulations set limits on the maximum permissible exposure (MPE) allowed from the operation of radio transmitters. These regulations do not take the place of RF-safety practices, however. This section deals with the topic of RF safety.

This section was prepared by members of the ARRL RF Safety Committee and coordinated by Dr Robert E. Gold, WBØKIZ. It summarizes what is now known and offers safety precautions based on the research to date.

All life on Earth has adapted to survive in an environment of weak, natural, low-frequency electromagnetic fields (in addition to the Earth's static geomagnetic field). Natural low-frequency EM fields come from two main sources: the sun, and thunderstorm activity. But in the last 100 years, man-made fields at much higher intensities and with a very different spectral distribution have altered this natural EM background in ways that are not yet fully understood. Researchers continue to look at the effects of RF exposure over a wide range of frequencies and levels.

Both RF and 60-Hz fields are classified as *nonionizing radiation,* because the frequency is too low for there to be enough photon energy to ionize atoms. (*Ionizing radiation*, such as X-rays, gamma rays and even some ultraviolet radiation has enough energy to knock electrons loose from their atoms. When this happens, positive and negative ions are formed.) Still, at sufficiently high power densities, EMR poses certain health hazards. It has been known since the early days of radio that RF energy can cause injuries by heating body tissue. (Anyone who has ever touched an improperly grounded radio chassis or energized antenna and received an *RF burn* will agree that this type of injury can be quite painful.) In extreme cases, RF-induced heating in the eye can result in cataract formation, and can even cause blindness. Excessive RF heating of the reproductive organs can cause sterility. Other health problems also can result from RF heating. These heat-related health hazards are called *thermal effects.* A microwave oven is a positive application of this thermal effect.

There also have been observations of changes in physiological function in the presence of RF energy levels that are too low to cause heating. These functions return to normal when the field is removed. Although research is ongoing, no harmful health consequences have been linked to these changes.

In addition to the ongoing research, much else has been done to address this issue. For example, FCC regulations set limits on exposure from radio transmitters. The Institute of Electrical and Electronics Engineers, the American National Standards Institute and the National Council for Radiation Protection and Measurement, among others, have recommended voluntary guidelines to limit human exposure to RF energy. The ARRL has established the RF Safety Committee, consisting of concerned medical doctors and scientists, serving voluntarily to monitor scientific research in the fields and to recommend safe practices for radio amateurs.

THERMAL EFFECTS OF RF ENERGY

Body tissues that are subjected to *very high* levels of RF energy may suffer serious heat damage. These effects depend on the frequency of the energy, the power density of the RF field that strikes the body and factors such as the polarization of the wave.

At frequencies near the body's natural resonant frequency, RF energy is absorbed more efficiently, and an increase in heating occurs. In adults, this frequency usually is about 35 MHz if the person is grounded, and about 70 MHz if insulated from the ground. Individual body parts may be resonant at different frequencies. The adult head, for example, is resonant around 400 MHz, while a baby's smaller head resonates near 700 MHz. Body size thus determines the frequency at which most RF energy is absorbed. As the frequency is moved farther from resonance, less RF heating generally occurs. *Specific absorption rate (SAR)* is a term that describes the rate at which RF energy is absorbed in tissue.

Maximum permissible exposure (MPE) limits are based on whole-body SAR values, with additional safety factors included as part of the standards and regulations. This helps explain why these safe exposure limits vary with frequency. The MPE limits define the maximum electric and magnetic field strengths or the plane-wave equivalent power densities associated with these fields, that a person may be exposed to without harmful effect—and with an acceptable safety factor. The regulations assume that a person exposed to a specified (safe) MPE level also will experience a safe SAR.

Nevertheless, thermal effects of RF energy should not be a major concern for most radio amateurs, because of the power levels we normally use and the intermittent nature of most amateur transmissions. Amateurs spend more time listening than transmitting, and many amateur transmissions such as CW and SSB use low-duty-cycle modes. (With FM or RTTY, though, the RF is present continuously at its maximum level during each transmission.) In any event, it is rare for radio amateurs to be subjected to RF fields strong enough to produce thermal effects, unless they are close to an energized antenna or un- shielded power amplifier. Specific suggestions for avoiding excessive exposure are offered later in this chapter.

ATHERMAL EFFECTS OF EMR

Research about possible health effects resulting from exposure to the lower level energy fields, the athermal effects, has been of two basic types: epidemiological research and laboratory research.

Scientists conduct laboratory research into biological mechanisms by which EMR may affect animals including humans. Epidemiologists look at the health patterns of large groups of people using statistical methods. These epidemiological studies have been inconclusive. By their basic design, these studies do not demonstrate cause and effect, nor do they postulate mechanisms of disease. Instead, epidemiologists look for associations between an environmental factor and an observed pattern of illness. For example, in the earliest research on malaria, epidemiologists observed the association between populations with high prevalence of the disease and the proximity of mosquito infested swamplands. It was left to the biological and medical scientists to isolate the organism causing malaria in the blood of those with the disease, and identify the same organisms in the mosquito population.

In the case of athermal effects, some studies have identified a weak association between exposure to EMF at home or at work and various malignant conditions including leukemia and brain cancer. A larger number of equally well designed and performed studies, however, have found no association. A risk ratio of between 1.5 and 2.0 has been observed in positive studies (the number of observed cases of malignancy being 1.5 to 2.0 times the "expected" number in the population). Epidemiologists generally regard a risk ratio of 4.0 or greater to be indicative of a strong association between the cause and effect under study. For example, men who smoke one pack of cigarettes per day increase their risk for lung cancer tenfold compared to nonsmokers, and two packs per day increases the risk to more than 25 times the nonsmokers' risk.

Epidemiological research by itself is rarely conclusive, however. Epidemiology only identifies health patterns in groups—it does not ordinarily determine their cause. And there are often confounding factors: Most of us are exposed to many different environmental hazards that may affect our health in various ways. Moreover, not all studies of persons likely to be exposed to high levels of EMR have yielded the same results.

There also has been considerable laboratory research about the biological effects of EMR in recent years. For example, some separate studies have indicated that even fairly low levels of EMR might alter the human body's circadian rhythms, affect the manner in which T lymphocytes function in the immune system and alter the nature of the electrical and chemical signals communicated through the cell membrane and between cells, among other things. Although these studies are intriguing, they do not demonstrate any effect of these low-level fields on the overall organism.

Much of this research has focused on low-frequency magnetic fields, or on RF fields that are keyed, pulsed or modulated at a low audio frequency (often below 100 Hz). Several studies suggested that humans and animals can adapt to the presence of a steady RF carrier more readily than to an intermittent, keyed or modulated energy source.

The results of studies in this area, plus speculations concerning the effect of various types of modulation, were and have remained somewhat controversial. None of the research to date has demonstrated that low-level EMR causes adverse health effects.

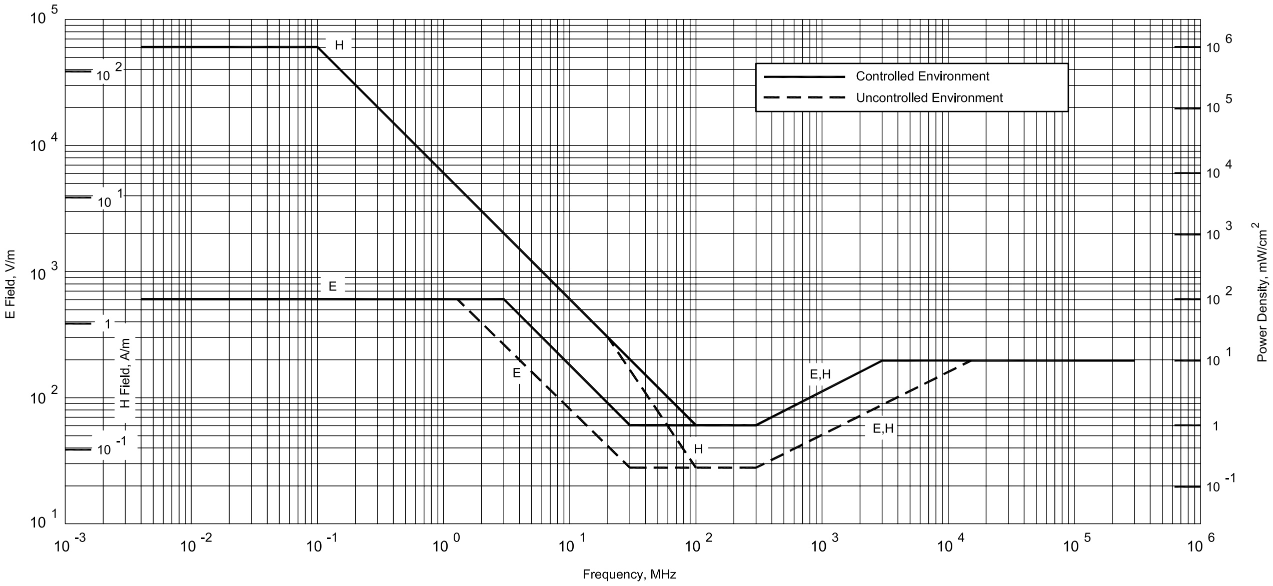

Fig 20—1991 RF protection guidelines for body exposure of humans. It is known officially as the "IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz."

Given the fact that there is a great deal of ongoing research to examine the health consequences of exposure to EMF, the American Physical Society (a national group of highly respected scientists) issued a statement in May 1995 based on its review of available data pertaining to the possible connections of cancer to 60-Hz EMF exposure. This report is exhaustive and should be reviewed by anyone with a serious interest in the field. Among its general conclusions were the following:

1. The scientific literature and the reports of reviews by other panels show no consistent, significant link between cancer and power line fields.
2. No plausible biophysical mechanisms for the systematic initiation or promotion of cancer by these extremely weak 60-Hz fields has been identified.
3. While it is impossible to prove that no deleterious health effects occur from exposure to any environmental factor, it is necessary to demonstrate a consistent, significant, and causal relationship before one can conclude that such effects do occur.

In a report dated October 31, 1996, a committee of the National Research Council of the National Academy of Sciences has concluded that no clear, convincing evidence exists to show that residential exposures to electric and magnetic fields (EMFs) are a threat to human health.

A National Cancer Institute epidemiological study of residential exposure to magnetic fields and acute lymphoblastic leukemia in children was published in the *New England Journal of Medicine* in July 1997. The exhaustive, seven-year study concludes that if there is any link at all, it is far too weak to be concerned about.

Readers may want to follow this topic as further studies are reported. Amateurs should be aware that exposure to RF and ELF (60 Hz) electromagnetic fields at all power levels and frequencies has not been fully studied under all circumstances. "Prudent avoidance" of any avoidable EMR is always a good idea. Prudent avoidance doesn't mean that amateurs should be fearful of using their equipment. Most amateur operations are well within the MPE limits. If any risk does exist, it will almost surely fall well down on the list of causes that may be harmful to your health (on the other end of the list from your automobile). It does mean, however, that hams should be aware of the potential for exposure from their stations, and take whatever reasonable steps they can take to minimize their own exposure and the exposure of those around them.

Safe Exposure Levels

How much EM energy is safe? Scientists and regulators have devoted a great deal of effort to deciding upon safe RF-exposure limits. This is a very complex problem, involving difficult public health and economic considerations. The recommended safe levels have been revised downward several times over the years —and not all scientific bodies agree on this question even today. An Institute of Electrical and Electronics Engineers (IEEE) standard for recommended EM exposure limits was published in 1991 (see Bibliography). It replaced a 1982 American National Standards Institute (ANSI) standard. In the new standard, most of the permitted exposure levels were revised downward (made more stringent), to better reflect the current research. The new IEEE stan-

FCC RF-Exposure Regulations

FCC regulations control the amount of RF exposure that can result from your station's operation (§§97.13, 97.503, 1.1307 (b)(c)(d), 1.1310 and 2.1093). The regulations set limits on the maximum permissible exposure (MPE) allowed from operation of transmitters in all radio services. They also require that certain types of stations be evaluated to determine if they are in compliance with the MPEs specified in the rules. The FCC has also required that five questions on RF environmental safety practices be added to Novice, Technician and General license examinations.

These rules went into effect on January 1, 1998 for new stations or stations that file a Form 605 application with the FCC. Other existing stations have until September 1, 2000 to be in compliance with the rules.

The Rules

Maximum Permissible Exposure (MPE)

All radio stations regulated by the FCC must comply with the requirements for MPEs, even QRP stations running only a few watts or less. The MPEs vary with frequency, as shown in **Table A**. MPE limits are specified in maximum electric and magnetic fields for frequencies below 30 MHz, in power density for frequencies above 300 MHz and all three ways for frequencies from 30 to 300 MHz. For compliance purposes, all of these limits must be considered *separately*. If any one is exceeded, the station is not in compliance.

The regulations control human exposure to RF fields, not the strength of RF fields. There is no limit to how strong a field can be as long as no one is being exposed to it, although FCC regulations require that amateurs use the minimum necessary power at all times (§97.311 [a]).

Environments

The FCC has defined two exposure environments — *controlled* and *uncontrolled*. A controlled environment is one in which the people who are being exposed are aware of that exposure and can take steps to minimize that exposure, if appropriate. In an uncontrolled environment, the people being exposed are not normally aware of the exposure. The uncontrolled environment limits are more stringent than the controlled environment limits.

Although the controlled environment is usually intended as an occupational environment, the FCC has determined that it generally applies to amateur operators and members of their immediate households. In most cases, controlled-environment limits can be applied to your home and property to which you can control physical access. The uncontrolled environment is intended for areas that are accessible by the general public, such as your neighbors' properties.

The MPE levels are based on average exposure. An averaging time of 6 minutes is used for controlled exposure; an averaging period of 30 minutes is used for uncontrolled exposure.

Table A—(From §1.1310) Limits for Maximum Permissible Exposure (MPE)

(A) Limits for Occupational/Controlled Exposure

Frequency Range (MHz)	*Electric Field Strength (V/m)*	*Magnetic Field Strength (A/m)*	*Power Density (mW/cm^2)*	*Averaging Time (minutes)*
0.3-3.0	614	1.63	(100)*	6
3.0-30	1842/f	4.89/f	(900/f^2)*	6
30-300	61.4	0.163	1.0	6
300-1500	—	—	f/300	6
1500-100,000	—	—	5	6

f = frequency in MHz

* = Plane-wave equivalent power density (see Note 1).

(B) Limits for General Population/Uncontrolled Exposure

Frequency Range (MHz)	*Electric Field Strength (V/m)*	*Magnetic Field Strength (A/m)*	*Power Density (mW/cm^2)*	*Averaging Time (minutes)*
0.3-1.34	614	1.63	(100)*	30
1.34-30	824/f	2.19/f	(180/f^2)*	30
30-300	27.5	0.073	0.2	30
300-1500	—	—	f/1500	30
1500-100,000	—	—	1.0	30

f = frequency in MHz

* = Plane-wave equivalent power density (see Note 1).

Note 1: This means the equivalent far-field strength that would have the E or H-field component calculated or measured. It does not apply well in the near field of an antenna. The equivalent far-field power density can be found in the near or far field regions from the relationships: $P_d = |E_{total}|^2 / 3770$ mW/cm^2 or from $P_d = |H_{total}|^2 \times 37.7$ mW/cm^2.

Station Evaluations

The FCC requires that certain amateur stations be evaluated for compliance with the MPEs. Although an amateur can have someone else do the evaluation, it is not difficult for hams to evaluate their own stations. The ARRL book *RF Exposure and You* contains extensive information about the regulations and a large chapter of tables that show compliance distances for specific antennas and power levels. Generally, hams will use these tables to evaluate their stations. Some of these tables have been included in the FCC's information — *OET Bulletin 65* and its *Supplement B*. If hams choose, however, they can do more extensive calculations, use a computer to model their antenna and exposure, or make actual measurements.

Table B—Power Thresholds for Routine Evaluation of Amateur Radio Stations

Wavelength Band	*Evaluation Required if Power* (watts) Exceeds:*
MF	
160 m	500
HF	
80 m	500
75 m	500
40 m	500
30 m	425
20 m	225
17 m	125
15 m	100
12 m	75
10 m	50
VHF (all bands)	50
UHF	
70 cm	70
33 cm	150
23 cm	200
13 cm	250
SHF (all bands)	250
EHF (all bands)	250
Repeater stations (all bands)	*non-building-mounted antennas*: height above ground level to lowest point of antenna < 10 m *and* power > 500 W ERP *building-mounted antennas*: power > 500 W ERP

*Transmitter power = Peak-envelope power input to antenna. For repeater stations ***only,*** power exclusion based on ERP (effective radiated power).

Categorical Exemptions

Some types of amateur stations do not need to be evaluated, but these stations must still comply with the MPE limits. The station licensee remains responsible for ensuring that the station meets these requirements.

The FCC has exempted these stations from the evaluation requirement because their output power, operating mode and frequency are such that they are presumed to be in compliance with the rules.

Stations using power equal to or less than the levels in **Table B** do not have to be evaluated. For the 100-W HF ham station, for example, an evaluation would be required *only* on 12 and 10 meters.

Hand-held radios and vehicle-mounted mobile radios that operate using a push-to-talk (PTT) button are also categorically exempt from performing the routine evaluation. Repeater stations that use less than 500 W ERP or those with antennas not mounted on buildings, if the antenna is at least 10 meters off the ground, also do not need to be evaluated.

Correcting Problems

Most hams are already in compliance with the MPE requirements. Some amateurs, especially those using indoor antennas or high-power, high-duty-cycle modes such as a RTTY bulletin station and specialized stations for moonbounce operations and the like may need to make adjustments to their station or operation to be in compliance.

The FCC permits amateurs considerable flexibility in complying with these regulations. As an example, hams can adjust their operating frequency, mode or power to comply with the MPE limits. They can also adjust their operating habits or control the direction their antenna is pointing.

More Information

This discussion offers only an overview of this topic; additional information can be found in *RF Exposure and You* and on *ARRLWeb* at **www.arrl.org/news/rfsafety/**. *ARRLWeb* has links to the FCC Web site, with *OET Bulletin 65* and *Supplement B* and links to software that hams can use to evaluate their stations.

dard was adopted by ANSI in 1992.

The IEEE standard recommends frequency-dependent and time-dependent maximum permissible exposure levels. Unlike earlier versions of the standard, the 1991 standard recommends different RF exposure limits in controlled environments (that is, where energy levels can be accurately determined and everyone on the premises is aware of the presence of EM fields) and in uncontrolled environments (where energy levels are not known or where people may not be aware of the presence of EM fields). FCC regulations also include controlled/occupational and uncontrolled/general population exposure environments.

The graph in **Fig 20** depicts the 1991 IEEE standard. It is necessarily a complex graph, because the standards differ not only for controlled and uncontrolled environments but also for electric (E) fields and magnetic (H) fields. Basically, the lowest E-field exposure limits occur at frequencies between 30 and 300 MHz. The lowest H-field exposure levels occur at 100-300 MHz. The ANSI standard sets the maximum E-field limits between 30 and 300 MHz at a power density of 1 mW/cm^2 (61.4 V/m) in controlled environments—but at one-fifth that level (0.2 mW/cm^2 or 27.5 V/m) in uncontrolled environments. The H-field limit drops to 1 mW/cm^2 (0.163 A/m) at 100-300 MHz in controlled environments and 0.2 mW/cm^2 (0.0728 A/m) in uncontrolled environments. Higher power densities are permitted at frequencies below 30 MHz (below 100 MHz for H fields) and above 300 MHz, based on the concept that the body will not be resonant at those frequencies and will therefore absorb less energy.

In general, the 1991 IEEE standard requires averaging the power level over time periods ranging from 6 to

30 minutes for power-density calculations, depending on the frequency and other variables. The ANSI exposure limits for uncontrolled environments are lower than those for controlled environments, but to compensate for that the standard allows exposure levels in those environments to be averaged over much longer time periods (generally 30 minutes). This long averaging time means that an intermittently operating RF source (such as an Amateur Radio transmitter) will show a much lower power density than a continuous-duty station—for a given power level and antenna configuration.

Time averaging is based on the concept that the human body can withstand a greater rate of body heating (and thus, a higher level of RF energy) for a short time than for a longer period. Time averaging may not be appropriate, however, when considering nonthermal effects of RF energy.

The IEEE standard excludes any transmitter with an output below 7 W because such low-power transmitters would not be able to produce significant whole-body heating. (Recent studies show that hand-held transceivers often produce power densities in excess of the IEEE standard within the head.)

There is disagreement within the scientific community about these RF exposure guidelines. The IEEE standard is still intended primarily to deal with thermal effects, not exposure to energy at lower levels. A small but significant number of researchers now believe athermal effects also should be taken into consideration. Several European countries and localities in the United States have adopted stricter standards than the recently updated IEEE standard.

Another national body in the United States, the National Council for Radiation Protection and Measurement (NCRP), also has adopted recommended exposure guidelines. NCRP urges a limit of 0.2 mW/cm^2 for nonoccupational exposure in the 30-300 MHz range. The NCRP guideline differs from IEEE in two notable ways: It takes into account the effects of modulation on an RF carrier, and it does not exempt transmitters with outputs below 7 W.

The FCC MPE regulations are based on parts of the 1992 IEEE/ANSI standard and recommendations of the National Council for Radiation Protection and Measurement (NCRP). The MPE limits under the regulations are slightly different than the IEEE/ANSI limits. Note that the MPE levels apply to the FCC rules put into effect for radio amateurs on January 1, 1998. These MPE requirements do not reflect and include all the assumptions and exclusions of the IEEE/ANSI standard.

Cardiac Pacemakers and RF Safety

It is a widely held belief that cardiac pacemakers may be adversely affected in their function by exposure to electromagnetic fields. Amateurs with pacemakers may ask whether their operating might endanger themselves or visitors to their shacks who have a pacemaker. Because of this, and similar concerns regarding other sources of electromagnetic fields, pacemaker manufacturers apply design methods that for the most part shield the pacemaker circuitry from even relatively high EM field strengths.

It is recommended that any amateur who has a pacemaker, or is being considered for one, discuss this matter with his or her physician. The physician will probably put the amateur into contact with the technical representative of the pacemaker manufacturer. These representatives are generally excellent resources, and may have data from laboratory or “in the field” studies with specific model pacemakers.

One study examined the function of a modern (dual chamber) pacemaker in and around an Amateur Radio station. The pacemaker generator has circuits that receive and process electrical signals produced by the heart, and also generate electrical signals that stimulate (pace) the heart. In one series of experiments, the pacemaker was connected to a heart simulator. The system was placed on top of the cabinet of a 1-kW HF linear amplifier during SSB and CW operation. In another test, the system was placed in close proximity to several 1 to 5-W 2-meter hand-held transceivers. The test pacemaker was connected to the heart simulator in a third test, and then placed on the ground 9 meters below and 5 meters in front of a three-element Yagi HF antenna. No interference with pacemaker function was observed in these experiments.

Although the possibility of interference cannot be entirely ruled out by these few observations, these tests represent more severe exposure to EM fields than would ordinarily be encountered by an amateur—with an average amount of common sense. Of course, prudence dictates that amateurs with pacemakers, who use hand-held VHF transceivers, keep the antenna as far as possible from the site of the implanted pacemaker generator. They also should use the lowest transmitter output required for adequate communication. For high power HF transmission, the antenna should be as far as possible from the operating position, and all equipment should be properly grounded.

Low-Frequency Fields

Although the FCC doesn’t regulate 60-Hz fields, some recent concern about EMR has focused on low-frequency energy rather than RF. Amateur Radio equipment can be a significant source of low-frequency magnetic fields, although there are many other sources of this kind of energy in the typical home. Magnetic fields can be measured relatively accurately with inexpensive 60-Hz meters that are made by several manufacturers.

Table 2 shows typical magnetic field intensities of Amateur Radio equipment and various household items. Because these fields dissipate rapidly with distance, “prudent avoidance” would mean staying perhaps 12 to 18 inches away from most Amateur Radio equipment (and

24 inches from power supplies with 1-kW RF amplifiers).

Determining RF Power Density

Unfortunately, determining the power density of the RF fields generated by an amateur station is not as simple as measuring low-frequency magnetic fields. Although sophisticated instruments can be used to measure RF power densities quite accurately, they are costly and require frequent recalibration. Most amateurs don't have access to such equipment, and the inexpensive field-

Table 2

Typical 60-Hz Magnetic Fields Near Amateur Radio Equipment and AC-Powered Household Appliances

Values are in milligauss.

Item	*Field*	*Distance*
Electric blanket	30-90	Surface
Microwave oven	10-100	Surface
	1-10	12″
IBM personal computer	5-10	Atop monitor
	0-1	15″ from screen
Electric drill	500-2000	At handle
Hair dryer	200-2000	At handle
HF transceiver	10-100	Atop cabinet
	1-5	15″ from front
1-kW RF amplifier	80-1000	Atop cabinet
	1-25	15″ from front

(Source: measurements made by members of the ARRL RF Safety Committee)

Table 3

Typical RF Field Strengths Near Amateur Radio Antennas

A sampling of values as measured by the Federal Communications Commission and Environmental Protection Agency, 1990

Antenna Type	*Freq (MHz)*	*Power (W)*	*E Field (V/m)*	*Location*
Dipole in attic	14.15	100	7-100	In home
Discone in attic	146.5	250	10-27	In home
Half sloper	21.5	1000	50	1 m from base
Dipole at 7-13 ft	7.14	120	8-150	1-2 m from earth
Vertical	3.8	800	180	0.5 m from base
5-element Yagi at 60 ft	21.2	1000	10-20	In shack
			14	12 m from base
3-element Yagi at 25 ft	28.5	425	8-12	12 m from base
Inverted V at 22-46 ft	7.23	1400	5-27	Below antenna
Vertical on roof	14.11	140	6-9	In house
			35-100	At antenna tuner
Whip on auto roof	146.5	100	22-75	2 m antenna
			15-30	In vehicle
			90	Rear seat
5-element Yagi at 20 ft	50.1	500	37-50	10 m antenna

Table 4

RF Awareness Guidelines

These guidelines were developed by the ARRL RF Safety Committee, based on the FCC/EPA measurements of Table 3 and other data.

- Although antennas on towers (well away from people) pose no exposure problem, make certain that the RF radiation is confined to the antennas' radiating elements themselves. Provide a single, good station ground (earth), and eliminate radiation from transmission lines. Use good coaxial cable or other feed line properly. Avoid serious imbalance in your antenna system and feed line. For high-powered installations, avoid end-fed antennas that come directly into the transmitter area near the operator.
- No person should ever be near any transmitting antenna while it is in use. This is especially true for mobile or ground-mounted vertical antennas. Avoid transmitting with more than 25 W in a VHF mobile installation unless it is possible to first measure the RF fields inside the vehicle. At the 1-kW level, both HF and VHF directional antennas should be at least 35 ft above inhabited areas. Avoid using indoor and attic-mounted antennas if at all possible. If open-wire feeders are used, ensure that it is not possible for people (or animals) to come into accidental contact with the feed line.
- Don't operate high-power amplifiers with the covers removed, especially at VHF/UHF.
- In the UHF/SHF region, never look into the open end of an activated length of waveguide or microwave feed-horn antenna or point it toward anyone. (If you do, you may be exposing your eyes to more than the maximum permissible exposure level of RF radiation.) Never point a high-gain, narrow-bandwidth antenna (a paraboloid, for instance) toward people. Use caution in aiming an EME (moonbounce) array toward the horizon; EME arrays may deliver an effective radiated power of 250,000 W or more.
- With hand-held transceivers, keep the antenna away from your head and use the lowest power possible to maintain communications. Use a separate microphone and hold the rig as far away from you as possible. This will reduce your exposure to the RF energy.
- Don't work on antennas that have RF power applied.
- Don't stand or sit close to a power supply or linear amplifier when the ac power is turned on. Stay at least 24 inches away from power transformers, electrical fans and other sources of high-level 60-Hz magnetic fields.

strength meters that we do have are not suitable for measuring RF power density.

Table 3 shows a sampling of measurements made at Amateur Radio stations by the Federal Communications Commission and the Environmental Protection Agency in 1990. As this table indicates, a good antenna well removed from inhabited areas poses no hazard under any of the IEEE/ANSI guidelines. However, the FCC/EPA survey also indicates that amateurs must be careful about using indoor or attic-mounted antennas, mobile antennas, low directional arrays or any other antenna that is close to inhabited areas, especially when moderate to high power is used.

Ideally, before using any antenna that is in close proximity to an inhabited area, you should measure the RF power density. If that is not feasible, the next best option is make the installation as safe as possible by observing the safety suggestions listed in **Table 4**.

It also is possible, of course, to calculate the probable power density near an antenna using simple equations. Such calculations have many pitfalls. For one, most of the situations where the power density would be high enough to be of concern are in the near field. In the near field, ground interactions and other variables produce power densities that cannot be determined by simple arithmetic. In the far field, conditions become easier to predict with simple calculations.

The boundary between the near field and the far field depends on the wavelength of the transmitted signal and the physical size and configuration of the antenna. The boundary between the near field and the far field of an antenna can be as much as several wavelengths from the antenna.

Computer antenna-modeling programs are another approach you can use. *MININEC* or other codes derived from *NEC* (Numerical Electromagnetics Code) are suitable for estimating RF magnetic and electric fields around amateur antenna systems.

These models have limitations. Ground interactions must be considered in estimating near-field power densities, and the "correct ground" must be modeled. Computer modeling is generally not sophisticated enough to predict "hot spots" in the near field—places where the field intensity may be far higher than would be expected, due to reflections from nearby objects. In addition, "nearby objects" often change or vary with weather or the season, so the model so laboriously crafted may not be representative of the actual situation, by the time it is running on the computer.

Intensely elevated but localized fields often can be detected by professional measuring instruments. These "hot spots" are often found near wiring in the shack, and metal objects such as antenna masts or equipment cabinets. But even with the best instrumentation, these measurements also may be misleading in the near field.

One need not make precise measurements or model the exact antenna system, however, to develop some idea of the relative fields around an antenna. Computer modeling using close approximations of the geometry and power input of the antenna will generally suffice. Those who are familiar with *MININEC* can estimate their power densities by computer modeling, and those who have access to professional power-density meters can make useful measurements.

While our primary concern is ordinarily the intensity of the signal radiated by an antenna, we also should remember that there are other potential energy sources to be considered. You also can be exposed to RF radiation directly from a power amplifier if it is operated without proper shielding. Transmission lines also may radiate a significant amount of energy under some conditions. Poor microwave waveguide joints or improperly assembled connectors are another source of incidental radiation.

Further RF Exposure Suggestions

Potential exposure situations should be taken seriously. Based on the FCC/EPA measurements and other data, the "RF awareness" guidelines of Table 4 were developed by the ARRL RF Safety Committee. A longer version of these guidelines, along with a complete list of references, appeared in a *QST* article by Ivan Shulman, MD, WC2S ("Is Amateur Radio Hazardous to Our Health?" *QST* , Oct 1989, pp 31-34). For more information or background, see the list of RF Safety References in the next section.

In addition, the ARRL has published a book, *RF Exposure and You*, that is helping hams comply with the FCC's RF-exposure regulations. The ARRL also maintains an RF-exposure news page on its Web site. See **www.arrl.org/news/rfsafety**. This site contains reprints of selected *QST* articles on RF exposure and links to the FCC and other useful sites.

Chapter 2

Antenna Fundamentals

Antennas belong to a class of devices called *transducers*. This term is derived from two Latin words, meaning literally “to lead across” or “to transfer.” Thus, a transducer is a device that transfers, or converts, energy from one form to another. The purpose of an antenna is to convert radio-frequency electric current to *electromagnetic waves*, which are then *radiated* into space. [For more details on the properties of electromagnetic waves themselves, see Chapter 23, Radio Wave Propagation.]

We cannot directly see or hear, taste or touch electromagnetic waves, so it’s not surprising that the process by which they are launched into space from our antennas can be a little mystifying, especially to a newcomer. In everyday life we come across many types of transducers, although we don’t always recognize them as such. A comparison with a type of transducer that you can actually see and touch may be useful. You are no doubt familiar with a *loudspeaker*. It converts audio-frequency electric current from the output of your radio or stereo into acoustic pressure waves, also known as *sound waves*. The sound waves are propagated through the air to your ears, where they are converted into what you perceive as sound.

We normally think of a loudspeaker as something that converts electrical energy into sound energy, but we could just as well turn things around and apply sound energy to a loudspeaker, which will then convert it into electrical energy. When used in this manner, the loudspeaker has become a microphone. The loudspeaker/microphone thus exhibits the principle of *reciprocity*, derived from the Latin word meaning to move back and forth.

Now, let’s look more closely at that special transducer we call an *antenna*. When fed by a transmitter with RF current (usually through a transmission line), the antenna launches electromagnetic waves, which are propagated through space. This is similar to the way sound waves are propagated through the air by a loudspeaker. In the next town, or perhaps on a distant continent, a similar transducer (that is, a receiving antenna) intercepts some of these electromagnetic waves and converts them into electrical current for a receiver to amplify and detect.

In the same fashion that a loudspeaker can act as a microphone, a radio antenna also follows the principle of reciprocity. In other words, an antenna can transmit as well as receive signals. However, unlike the loudspeaker, an antenna does not require a *medium*, such as air, through which it radiates electromagnetic waves. Electromagnetic waves can be propagated through air, the vacuum of outer space or the near vacuum of the upper ionosphere. This is the miracle of radio—electromagnetic waves can propagate without a physical medium.

Essential Characteristics of Antennas

What other things make an antenna different from an ordinary electronic circuit? In ordinary circuits, the dimensions of coils, capacitors and connections usually are small compared with the wavelength of the frequency in use. Here, we define wavelength as the distance in free space traveled during one complete cycle of a wave. The velocity of a wave in free space is the speed of light, and the wavelength is thus:

$$\lambda_{\text{meters}} = \frac{299.7925 \times 10^6 \text{ meters/sec}}{\text{f hertz}} = \frac{299.7925}{\text{f MHz}} \qquad \text{(Eq 1)}$$

where λ_{meters} , the Greek letter lambda, is the free-space wavelength in meters.

Expressed in feet, Eq 1 becomes:

$$\lambda_{feet} = \frac{983.5712}{f\ MHz} \approx \frac{983.6}{f\ MHz} \qquad \text{(Eq 2)}$$

When circuit dimensions are small compared to λ, most of the electromagnetic energy is confined to the circuit itself, and is used up either performing useful work or is converted into heat. However, when the dimensions of wiring or components become significant compared with the wavelength, some of the energy escapes by radiation in the form of electromagnetic waves.

Antennas come in an enormous, even bewildering, assortment of shapes and sizes. This chapter on fundamentals will deal with the theory of simple forms of antennas, usually in *free space*, away from the influence of ground. Subsequent chapters will concentrate on more exotic or specialized antenna types. Chapter 3 deals with the complicated subject of the effect of ground, including the effect of uneven local terrain. Ground has a profound influence on how an antenna performs in the real world.

No matter what form an antenna takes, simple or complex, its electrical performance can be characterized according to the following important properties:

1. Feed-point Impedance
2. Directivity, Gain and Efficiency
3. Polarization

FEED-POINT IMPEDANCE

The first major characteristic defining an antenna is its *feed-point impedance*. Since we amateurs are free to choose our operating frequencies within assigned bands, we need to consider how the feed-point impedance of a particular antenna varies with frequency, within a particular band, or even in several different bands if we intend to use one antenna on multiple bands.

There are two forms of impedance associated with any antenna: *self impedance* and *mutual impedance*. As you might expect, self impedance is what you measure at the feed-point terminals of an antenna located completely away from the influence of any other conductors.

Mutual, or coupled, impedance is due to the parasitic effect of nearby conductors; that is, conductors located within the antenna's reactive near field. (The subject of fields around an antenna will be discussed in detail later.) This includes the effect of ground, which is a lossy conductor, but a conductor nonetheless. Mutual impedance is defined using Ohm's Law, just like self impedance. However, mutual impedance is the ratio of voltage in one conductor, divided by the current in another (coupled) conductor. Mutually coupled conductors can distort the pattern of a highly directive antenna, as well as change the impedance seen at the feed point.

In this chapter on fundamentals, we won't directly deal with mutual impedance, considering it as a side effect of nearby conductors. Instead, here we'll concentrate on simple antennas in free space, away from ground and any other conductors. Mutual impedance will be considered in detail in Chapter 11, HF Yagi Arrays, where it is essential for proper operation of these beam antennas.

Self Impedance

The current that flows into an antenna's feed point must be supplied at a finite voltage. The self impedance of the antenna is simply equal to the voltage applied to its feed point divided by the current flowing into the feed point. Where the current and voltage are exactly in phase, the impedance is purely resistive, with zero reactive component. For this case the antenna is termed *resonant*. (Amateurs often use the term "resonant" rather loosely, usually meaning "nearly resonant" or "close-to resonant.")

Please recognize that an antenna *need not be resonant* in order to be an effective radiator. There is in fact nothing magic about having a resonant antenna, provided of course that you can devise some efficient means to feed the antenna. Many amateurs use non-resonant (even random-length) antennas fed with open-wire transmission lines and antenna tuners. They radiate signals just as well as those using coaxial cable and resonant antennas, and as a bonus they usually can use these antenna systems on multiple frequency bands. It is important to consider an antenna and its feed line as a *system*, in which all losses should be kept to a minimum. See Chapter 24, Transmission Lines, for details on transmission-line loss as a function of impedance mismatch.

Except at the one frequency where it is truly resonant, the current in an antenna is at a different phase compared to the applied voltage. In other words, the antenna exhibits a feed-point *impedance*, not just a pure resistance. The feed-point impedance is composed of either capacitive or inductive reactance in series with a resistance.

Radiation Resistance

The power supplied to an antenna is dissipated in two ways: radiation of electromagnetic waves, and heat losses in the wire and nearby dielectrics. The radiated power is what we want, the useful part, but it represents a form of "loss" just as much as the power used in heating the wire or nearby dielectrics is a loss. In either case, the dissipated power is equal to I^2R. In the case of heat losses, R is a real resistance. In the case of radiation, however, R is a "virtual" resistance, which, if replaced with an actual resistor of the same value, would dissipate the power actually radiated from the antenna. This resistance is called the *radiation resistance*. The total power in the antenna is therefore equal to $I^2(R_0+R)$, where R_0 is the radiation resistance and R represents the total of all the loss resistances.

In ordinary antennas operated at amateur frequencies, the power lost as heat in the conductor does not exceed a few percent of the total power supplied to the antenna. Expressed in decibels, the loss is less than 0.1 dB. The RF loss resistance of copper wire even as small as #14 is very low compared with the radiation resistance of an antenna that is reasonably clear of surrounding objects and is not too close to the ground. You can therefore assume that the ohmic loss in a reasonably well-located antenna is negligible, and that the total resistance shown by the antenna (the feed-point resistance) is radiation resistance. As a radiator of electromagnetic waves, such an antenna is a highly efficient device.

Impedance of a Center-Fed Dipole

A fundamental type of antenna is the *center-fed half-wave dipole*. Historically, the λ/2 dipole has been the most popular antenna used by amateurs worldwide, largely because it is very simple to construct and because it is an effective performer. It is also a basic building block for many other antenna systems, including beam antennas, such as Yagis.

A center-fed half-wave *dipole* consists of a straight wire, one-half wavelength long as defined in Eq 1, and fed in the center. The term "dipole" derives from Greek words meaning "two poles." See **Fig 1**. A λ/2-long dipole is just one form a dipole can take. Actually, a center-fed dipole can be any length electrically, as long as it is configured in a symmetrical fashion with two equal-length legs. There are also versions of dipoles that are not fed in the center. These are called *off-center-fed dipoles*, sometimes abbreviated as "OCF dipoles."

In free space—with the antenna remote from everything else—the theoretical impedance of a physically half-wave long antenna made of an infinitely thin conductor is 73 + *j* 42.5 Ω. This antenna exhibits both resistance and reactance. The positive sign in the + *j* 42.5-Ω reactive term indicates that the antenna exhibits an inductive reactance at its feed point. The antenna is slightly long electrically, compared to the length necessary for exact resonance, where the reactance is zero.

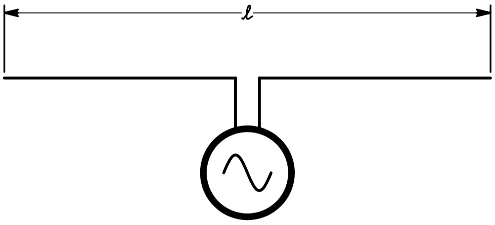

Fig 1—The center-fed dipole antenna. It is assumed that the source of power is directly at the antenna feed point, with no intervening transmission line. Most commonly in amateur applications, the overall length of the dipole is λ/2, but the antenna can in actuality be any length.

The feed-point impedance of any antenna is affected by the wavelength-to-diameter ratio (λ/dia) of the conductors used. Theoreticians like to specify an "infinitely thin" antenna because it is easier to handle mathematically.

What happens if we keep the physical length of an antenna constant, but change the thickness of the wire used in its construction? Further, what happens if we vary the frequency from well below to well above the half-wave resonance and measure the feed-point impedance? **Fig 2** graphs the impedance of a 100-foot long, center-fed dipole in free space, made with extremely thin wire—in this case, wire that is only 0.001 inches in diameter. There is nothing particularly significant about the choice here of 100 feet. This is simply a numerical example.

We could never actually build such a thin antenna (and neither could we install it in free space), but we can model how this antenna works using a very powerful piece of computer software called *NEC-4.1*. See Chapter 4, Antenna Modeling and System Planning, for details on antenna modeling.

The frequency applied to the antenna in Fig 2 is varied from 1 to 30 MHz. The x-axis has a logarithmic scale because of the wide range of feed-point resistance seen over the frequency range. The y-axis has a linear scale representing the reactive portion of the impedance. Inductive reactance is positive and capacitive reactance is negative on the y-axis. The bold figures centered on the spiraling line show the frequency in MHz.

At 1 MHz, the antenna is very short electrically, with a resistive component of about 2 Ω and a series capaci-

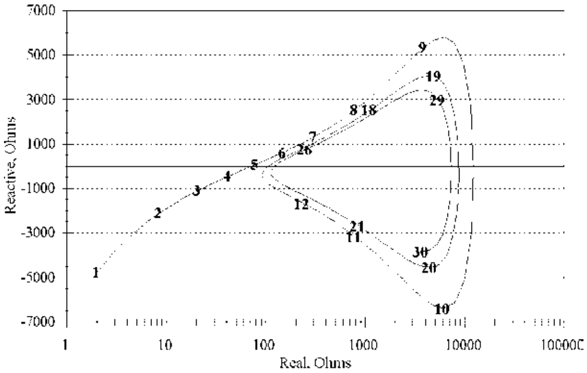

Fig 2—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of extremely thin 0.001-inch diameter wire. The y-axis is calibrated in positive (inductive) series reactance up from the zero line, and negative (capacitive) series reactance in the downward direction. The range of reactance goes from –6500 Ω to +6000 Ω. Note that the x-axis is logarithmic because of the wide range of the real, resistive component of the feed-point impedance, from roughly 2 Ω to 10,000 Ω. The numbers placed along the curve show the frequency in MHz.

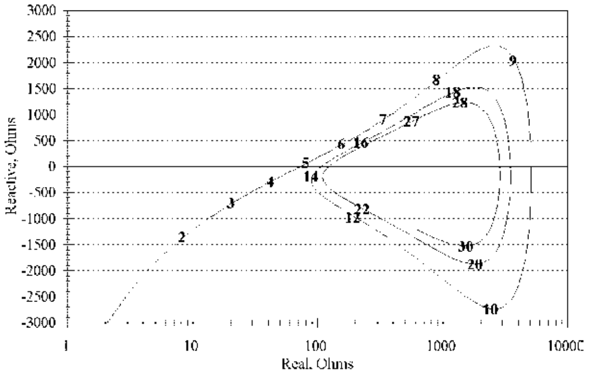

Fig 3—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of thin 0.1-inch (#10) diameter wire. Note that the range of change in reactance is less than that shown in Fig 2, ranging from –2700 Ω to +2300 Ω. At about 5,000 Ω, the maximum resistance is also less than that in Fig 2 for the thinner wire, where it is about 10,000 Ω.

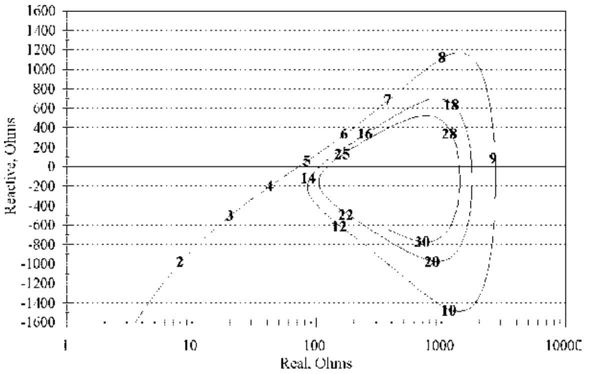

Fig 4—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of thick 1.0-inch diameter wire. Once again, the excursion in both reactance and resistance over the frequency range is less with this thick wire dipole than with thinner ones.

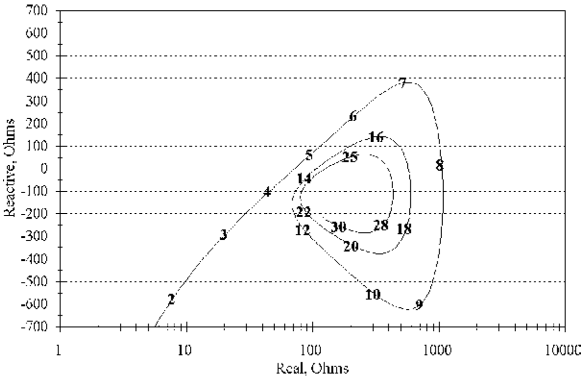

Fig 5—Feed-point impedance versus frequency for a theoretical 100-foot long dipole in free space, fed in the center and made of very thick 10.0-inch diameter wire. This ratio of length to diameter is about the same as a typical rod type of dipole element commonly used at 432 MHz. The maximum resistance is now about 1,000 Ω and the peak reactance range is from about –625 Ω to +380 Ω. This performance is also found in "cage" dipoles, where a number of paralleled wires are used to simulate a fat conductor.

tive reactance about –5000 Ω. Close to 5 MHz, the line crosses the zero-reactance line, meaning that the antenna goes through half-wave resonance there. Between 9 and 10 MHz the antenna exhibits a peak inductive reactance of about 6000 Ω. It goes through full-wave resonance (again crossing the zero-reactance line) between 9.5 and 9.6 MHz. At about 10 MHz, the reactance peaks at about –6500 Ω. Around 14 MHz, the line again crosses the zero-reactance line, meaning that the antenna has now gone through 3/2-wave resonance.

Between 19 and 20 MHz, the antenna goes through 4/2-wave resonance, which is twice the full-wave resonance or four times the half-wave frequency. If you allow your mind's eye to trace out the curve for frequencies beyond 30 MHz, it eventually spirals down to a resistive component somewhere between 200 and 3000 Ω. Thus, we have another way of looking at an antenna—as a sort of *transformer*, one that transforms the free-space impedance into the impedance seen at its feed point.

Now look at **Fig 3**, which shows the same kind of spiral curve, but for a thicker-diameter wire, one that is 0.1 inches in diameter. This diameter is close to #10 wire, a practical size we might actually use to build a real dipole. Note that the y-axis scale in Fig 3 is different from that in Fig 2. The range is from ±3000 Ω in Fig 3, while it was ±7000 Ω in Fig 2. The reactance for the thicker antenna ranges from +2300 to –2700 Ω over the whole frequency range from 1 to 30 MHz. Compare this with the range of +5800 to –6400 Ω for the very thin wire in Fig 2.

Fig 4 shows the impedance for a 100-foot long dipole using really thick, 1.0-inch diameter wire. The reactance varies from +1000 to –1500 Ω, indicating once again that a larger diameter antenna exhibits less of an excursion in the reactive component with frequency. Note that at the half-wave resonance just below 5 MHz, the resistive component of the impedance is still about 70 Ω, just about what it is for a much thinner antenna. Unlike the reactance, the half-wave radiation resistance of an antenna doesn't radically change with wire diameter, although the maximum level of resistance at full-wave resonance is lower for thicker antennas.

Fig 5 shows the results for a very thick, 10-inch diameter wire. Here, the excursion in the reactive component is even less: about +400 to –600 Ω. Note that the full-wave resonant frequency is about 8 MHz for this

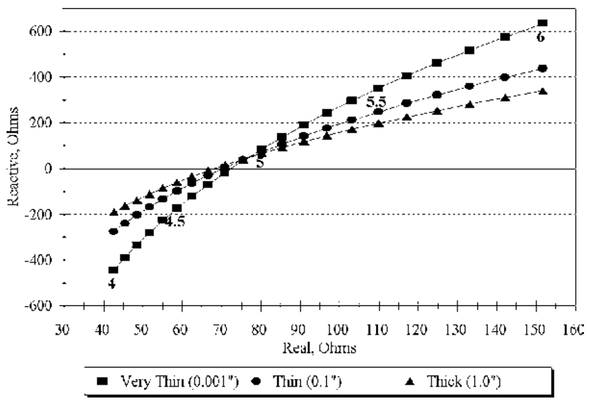

Fig 6—Expansion of frequency range around half-wave resonant point of three center-fed dipoles of three different thicknesses. The frequency is shown along the curves in MHz. The slope of change in series reactance versus series resistance is steeper for the thinner antennas than for the thick 1.0-inch antenna, indicating that the Q of the thinner antennas is higher.

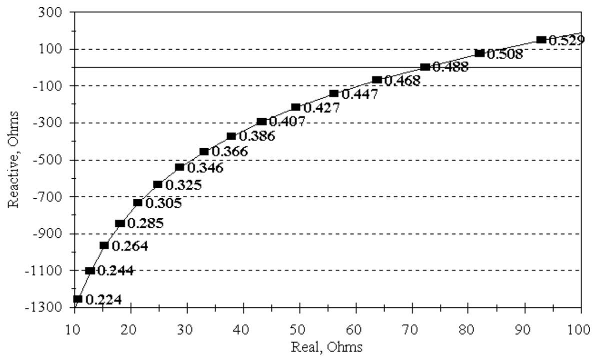

Fig 7—Another way of looking at the data for a 100-foot, center-fed dipole made of #14 wire in free space. The numbers along the curve represent the fractional wavelength, rather than frequency as shown in Fig 6. Note that this antenna goes through its half-wave resonance about 0.488 λ, rather than exactly at a half-wave physical length.

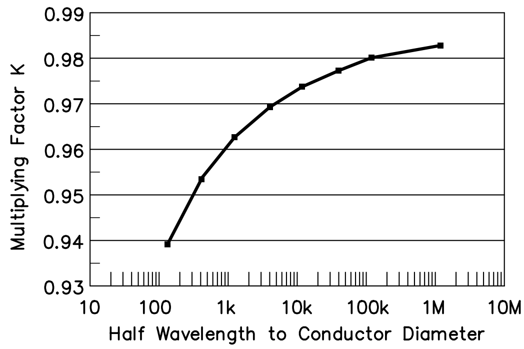

Fig 8—Effect of antenna diameter on length for half-wavelength resonance, shown as a multiplying factor, K, to be applied to the free-space, half-wavelength.

extremely thick antenna, while thinner antennas have full-wave resonances closer to 9 MHz. Note also that the full-wave resistance for this extremely thick antenna is only about 1,000 Ω, compared to the 10,000 Ω shown in Fig 2. All half-wave resonances shown in Figs 2 through 5 remain close to 5 MHz, regardless of the diameter of the antenna wire. Once again, the extremely thick, 10-inch diameter antenna has a resistive component at half-wave resonance close to 70 Ω. And once again, the change in reactance near this frequency is very much less for the extremely thick antenna than for thinner ones.

Now, we grant you that a 100-foot long antenna made with 10-inch diameter wire sounds a little odd! A length of 100 feet and a diameter of 10 inches represent a ratio of 120:1 in length to diameter. However, this is about the same length-to-diameter ratio as a 432-MHz half-wave dipole using 0.25-inch diameter elements, where the ratio is 109:1. In other words, the ratio of length-to-diameter for the 10-inch diameter, 100-foot long dipole is not that far removed from what might actually be used at UHF.

Another way of highlighting the changes in reactance and resistance is shown in **Fig 6**. This shows an expanded portion of the frequency range around the half-wave resonant frequency, from 4 to 6 MHz. In this region, the shape of each spiral curve is almost a straight line. The slope of the curve for the very thin antenna (0.001-inch diameter) is steeper than that for the thicker antennas (0.1 and 1.0-inch diameters). **Fig 7** illustrates another way of looking at the impedance data above and below the half-wave resonance. This is for a 100-foot dipole made of #14 wire. Instead of showing the frequency for each impedance point, the wavelength is shown, making the graph more universal in application.

Just to show that there are lots of ways of looking at the same data, **Fig 8** graphs the constant "K" used to multiply the free-space half-wavelength as a function of the ratio between the half-wavelength and the conductor diameter. The curve approaches the value of 1.00 for an infinitely thin conductor, in other words an infinitely large ratio of half-wavelength to diameter.

The behavior of antennas with different λ/diameter ratios corresponds to the behavior of ordinary series-resonant circuits having different values of Q. When the Q of a circuit is low, the reactance is small and changes rather slowly as the applied frequency is varied on either side of resonance. If the Q is high, the converse is true. The response curve of the low-Q circuit is *broad*; that of the high-Q circuit *sharp*. So it is with antennas—the impedance of a thick antenna changes slowly over a comparatively wide band of frequencies, while a thin antenna has a faster change in impedance. Antenna Q is defined

$$Q = \frac{f_0 \Delta X}{2R_0 \Delta f} \qquad \text{(Eq 3)}$$

where f_0 is the center frequency, ΔX is the change in the reactance for a Δf change in frequency, and R_0 is the resistance the f_0. For the "Very Thin," 0.001-inch diameter dipole in Fig 2, a change of frequency from 5.0 to 5.5 MHz yields a reactance change from 86 to 351 Ω, with an R_0 of 95 Ω. The Q is thus 14.6. For the 1.0-inch-diameter "Thick" dipole in Fig 4, $\Delta X = 131\ \Omega$ and R_0 is still 95 Ω, making Q = 7.2 for the thicker antenna, roughly half that of the thinner antenna.

Let's recap. We have described an antenna first as a transducer, then as a sort of transformer to a range of free-space impedances. Now, we just compared the antenna to a series-tuned circuit. Near its half-wave resonant frequency, a center-fed $\lambda/2$ dipole exhibits much the same characteristics as a conventional series-resonant circuit. Exactly at resonance, the current at the input terminals is in phase with the applied voltage and the feed-point impedance is purely resistive. If the frequency is below resonance, the phase of the current leads the voltage; that is, the reactance of the antenna is capacitive. When the frequency is above resonance, the opposite occurs; the current lags the applied voltage and the antenna exhibits inductive reactance. Just like a conventional series-tuned circuit, the antenna's reactance and resistance determines its Q.

ANTENNA DIRECTIVITY AND GAIN

The Isotropic Radiator

Before we can fully describe practical antennas, we must first introduce a completely theoretical antenna, the *isotropic radiator*. Envision, if you will, an infinitely small antenna, a point located in outer space, completely removed from anything else around it. Then consider an infinitely small transmitter feeding this infinitely small, point antenna. You now have an isotropic radiator.

The uniquely useful property of this theoretical *point-source* antenna is that it radiates equally well in all directions. That is to say, an isotropic antenna favors no direction at the expense of any other—in other words, it has absolutely no *directivity*. The isotropic radiator is useful as a *measuring stick* for comparison with actual antenna systems.

You will find later that real, practical antennas all exhibit some degree of directivity, which is the property of radiating more strongly in some directions than in others. The radiation from a practical antenna never has the same intensity in all directions and may even have zero radiation in some directions. The fact that a practical antenna displays directivity (while an isotropic radiator does not) is not necessarily a bad thing. The directivity of a real antenna is often carefully tailored to emphasize radiation in particular directions. For example, a receiving antenna that favors certain directions can discriminate against interference or noise coming from other directions, thereby increasing the signal-to-noise ratio for desired signals coming from the favored direction.

Directivity and the Radiation Pattern—a Flashlight Analogy

The directivity of an antenna is directly related to the *pattern* of its radiated field intensity in free space. A graph showing the actual or relative field intensity at a fixed distance, as a function of the direction from the antenna system, is called a *radiation pattern*. Since we can't actually see electromagnetic waves making up the radiation pattern of an antenna, we can consider an analogous situation.

Fig 9 represents a flashlight shining in a totally darkened room. To quantify what our eyes are seeing, we might use a sensitive light meter like those used by photographers, with a scale graduated in units from 0 to 10. We place the meter directly in front of the flashlight and adjust the distance so the meter reads 10, exactly full scale. We also carefully note the distance. Then, always keeping the meter the same distance from the flashlight and keeping it at the same height above the floor, we move the light meter around the flashlight, as indicated by the

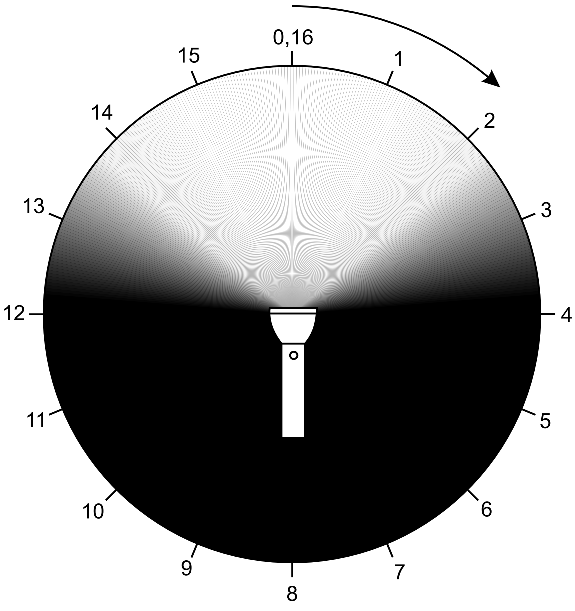

Fig 9—The beam from a flashlight illuminates a totally darkened area as shown here. Readings taken with a photographic light meter at the 16 points around the circle may be used to plot the radiation pattern of the flashlight.

arrow, and take light readings at a number of different positions.

After all the readings have been taken and recorded, we plot those values on a sheet of polar graph paper, like that shown in **Fig 10**. The values read on the meter are plotted at an angular position corresponding to that for which each meter reading was taken. Following this, we connect the plotted points with a smooth curve, also shown in Fig 10. When this is finished, we have completed a radiation pattern for the flashlight.

Antenna Pattern Measurements

Antenna radiation patterns can be constructed in a similar manner. Power is fed to the antenna under test, and a field-strength meter indicates the amount of signal. We might wish to rotate the antenna under test, rather than moving the measuring equipment to numerous positions about the antenna. Or we might make use of antenna reciprocity, since the pattern while receiving is the same as that while transmitting. A source antenna fed by a low-power transmitter illuminates the antenna under test, and the signal intercepted by the antenna under test is fed to a receiver and measuring equipment. Additional information on the mechanics of measuring antenna patterns is contained in Chapter 27, Antennas and Transmission-Line Measurements.

Some precautions must be taken to assure that the measurements are accurate and repeatable. In the case of the flashlight, let's assume that the separation between the light source and the light meter is 2 meters, about 6.5 feet. The wavelength of visible light is about one-half micron, where a micron is one-millionth of a meter.

For the flashlight, a separation of 2 meters between source and detector is $2.0/(0.5 \times 10^{-6}) = 4$ million λ, a very large number of wavelengths. Measurements of practical HF or even VHF antennas are made at much closer distances, in terms of wavelength. For example, at 3.5 MHz a full wavelength is 85.7 meters, or 281.0 feet. To duplicate the flashlight-to-light-meter spacing in wavelengths at 3.5 MHz, we would have to place the field-strength measuring instrument almost on the surface of the Moon, about a quarter-million miles away!

The Fields Around an Antenna

Why should we be concerned with the separation between the source antenna and the field-strength meter, which has its own receiving antenna? One important reason is that if you place a receiving antenna very close to an antenna whose pattern you wish to measure, mutual coupling between the two antennas may actually alter the pattern you are trying to measure.

This sort of mutual coupling can occur in the region very close to the antenna under test. This region is called the *reactive near-field* region. The term "reactive" refers to the fact that the mutual impedance between the transmitting and receiving antennas can be either capacitive or inductive in nature. The reactive near field is sometimes called the "induction field," meaning that the magnetic field usually is predominant over the electric field in this region. The antenna acts as though it were a rather large, lumped-constant inductor or capacitor, storing energy in the reactive near field rather than propagating it into space.

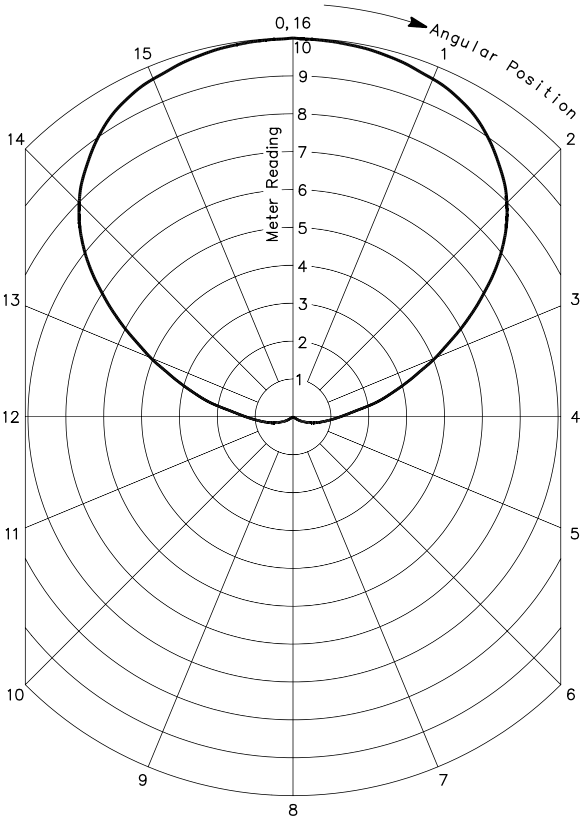

Fig 10—The radiation pattern of the flashlight in Fig 9. The measured values are plotted and connected with a smooth curve.

For simple wire antennas, the reactive near field is considered to be within about a half wavelength from an antenna's radiating center. Later on, in the chapters dealing with Yagi and quad antennas, you will find that mutual coupling between elements can be put to good use to purposely shape the radiated pattern. For making pattern measurements, however, we do not want to be too close to the antenna being measured.

The strength of the reactive near field decreases in a complicated fashion as you increase the distance from the antenna. Beyond the reactive near field, the antenna's radiated field is divided into two other regions: the *radiating near field* and the *radiating far field.* Historically, the terms *Fresnel* and *Fraunhöfer* fields have been used for the radiating near and far fields, but these terms have been largely supplanted by the more descriptive terminology used here. Even inside the reactive near-field

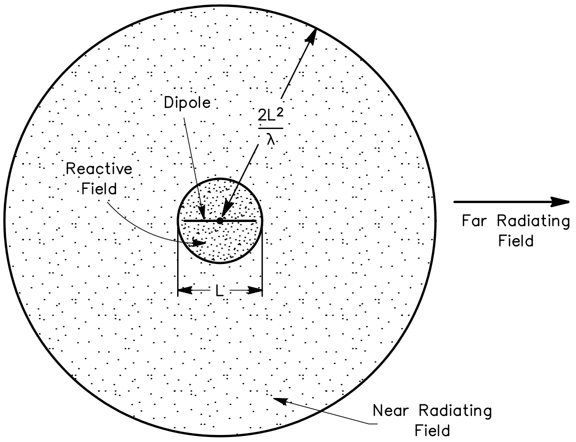

Fig 11—The fields around a radiating antenna. Very close to the antenna, the reactive field dominates. Within this area mutual impedances are observed between antenna and any other antennas used to measure response. Outside of the reactive field, the near radiating field dominates, up to a distance approximately equal to $2L^2/\lambda$, where L is the length of the largest dimension of the antenna. Beyond the near/far field boundary lies the far radiating field, where power density varies as the inverse square of radial distance.

region, both radiating and reactive fields coexist, although the reactive field predominates very close to the antenna.

Because the boundary between the fields is rather fuzzy, experts debate where one field begins and another leaves off, but the boundary between the radiating near and far fields is generally accepted as:

$$D \approx \frac{2L^2}{\lambda} \qquad \text{(Eq 4)}$$

where L is the largest dimension of the physical antenna, expressed in the same units of measurement as the wavelength λ. Remember, many specialized antennas do not follow the rule of thumb in Eq 4 exactly. **Fig 11** depicts the three fields in front of a simple wire antenna.

Throughout the rest of this book we will discuss mainly the radiating far-fields, those forming the traveling electromagnetic waves. Far-field radiation is distinguished by the fact that the intensity is inversely proportional to the distance, and that the electric and magnetic components, although perpendicular to each other in the wave front, are in time phase. The total energy is equally divided between the electric and magnetic fields. Beyond several wavelengths from the antenna, these are the only fields we need to consider. For accurate measurement of radiation patterns, we must place our measuring instrumentation at least several wavelengths away from the antenna under test.

Pattern Planes

Patterns obtained above represent the antenna radiation in just one plane. In the example of the flashlight, the plane of measurement was at one height above the floor. Actually, the pattern for any antenna is three dimensional, and therefore cannot be represented in a single-plane drawing. The *solid* radiation pattern of an antenna in free space would be found by measuring the field strength at every point on the surface of an imaginary sphere having the antenna at its center. The information so obtained would then be used to construct a solid figure, where the distance from a fixed point (representing the antenna) to the surface of the figure is proportional to the field strength from the antenna in any given direction. **Fig 12B** shows a three-dimensional wire-grid representation of the radiation pattern of a half-wave dipole.

For amateur work, *relative* values of field strength (rather than absolute) are quite adequate in pattern plotting. In other words, it is not necessary to know how many microvolts per meter a particular antenna will produce at a distance of 1 mile when excited with a specified power level. (This is the kind of specifications that AM broadcast stations must meet to certify their antenna systems to the FCC.)

For whatever data is collected (or calculated from theoretical equations), it is common to normalize the plotted values so the field strength in the direction of maximum radiation coincides with the outer edge of the chart. On a given system of polar coordinate scales, the *shape* of the pattern is not altered by proper normalization, only its size.

E and H-Plane Patterns

The solid 3-D pattern of an antenna in free space cannot adequately be shown with field-strength data on a flat sheet of paper. Cartographers making maps of a round Earth on flat pieces of paper face much the same kind of problem. As we discussed above, cross-sectional or plane diagrams are very useful for this purpose. Two such diagrams, one in the plane containing the straight wire of a dipole and one in the plane perpendicular to the wire, can convey a great deal of information. The pattern in the plane containing the axis of the antenna is called the *E-plane pattern*, and the one in the plane perpendicular to the axis is called the *H-plane pattern*. These designations are used because they represent the planes in which the electric (symbol E), and the magnetic (symbol H) lines of force lie, respectively.

The E lines represent the *polarization* of the antenna. Polarization will be covered in more detail later in this chapter. As an example, the electromagnetic field pictured in Fig 1 of Chapter 23, Radio Wave Propagation, is the field that would be radiated from a vertically polarized antenna; that is, an antenna in which the conductor is mounted perpendicular to the earth.

When a radiation pattern is shown for an antenna mounted over ground rather than in free space, we automatically gain two frames of reference—an *azimuth angle* and an *elevation angle*. The azimuth angle is usually referenced to the maximum radiation lobe of the antenna, where the azimuth angle is defined at 0°, or it could be referenced to the Earth's True North direction for an antenna oriented in a particular compass direction. The E-plane pattern for an antenna over ground is now called the *azimuth pattern*.

The elevation angle is referenced to the horizon at the Earth's surface, where the elevation angle is 0°. Of course, the Earth is round but because its radius is so large, it can in this context be considered to be flat in the area directly under the antenna. An elevation angle of 90° is straight over the antenna, and a 180° elevation is toward the horizon directly behind the antenna.

Professional antenna engineers often describe an antenna's orientation with respect to the point directly overhead—using the *zenith angle*, rather than the elevation angle. The elevation angle is computed by subtracting the zenith angle from 90°.

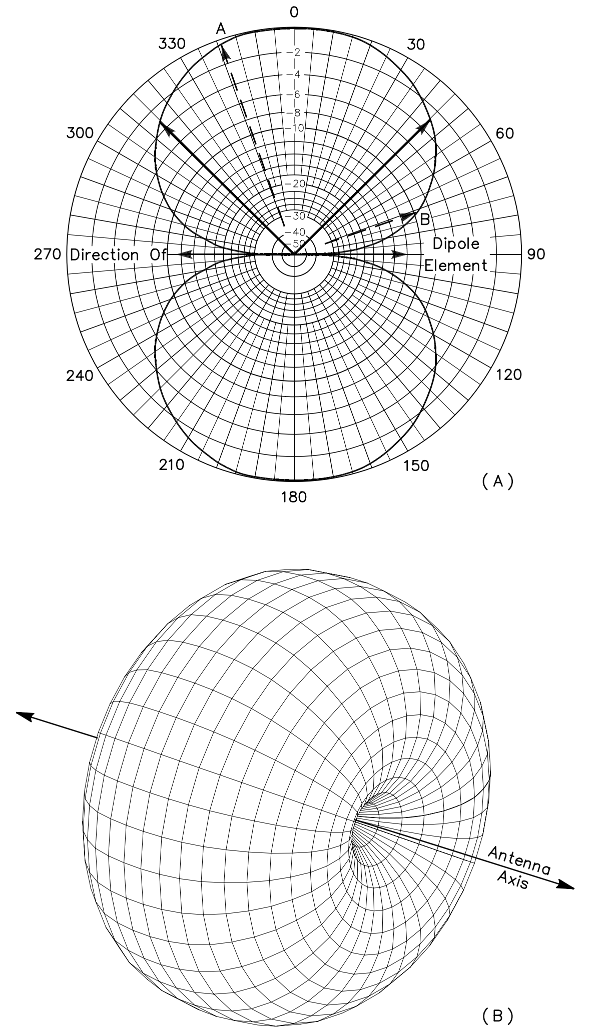

Fig 12—Directive diagram of a free-space dipole. At A, the pattern in the plane containing the wire axis. The length of each dashed-line arrow represents the relative field strength in that direction, referenced to the direction of maximum radiation, which is at right angles to the wire's axis. The arrows at approximately 45° and 315° are the half-power or –3 dB points. At B, a wire grid representation of the solid pattern for the same antenna. These same patterns apply to any center-fed dipole antenna less than a half wavelength long.

Introduction to the Decibel

The power gain of an antenna system is usually expressed in *decibels*. The decibel is a practical unit for measuring power ratios because it is more closely related to the actual effect produced at a distant receiver than the power ratio itself. One decibel represents a just-detectable change in signal strength, regardless of the actual value of the signal voltage. A 20-decibel (20-dB) increase in signal, for example, represents 20 observable steps in increased signal. The power ratio (100 to 1) corresponding to 20 dB gives an entirely exaggerated idea of the improvement in communication to be expected. The number of decibels corresponding to any power ratio is equal to 10 times the common logarithm of the power ratio, or

$$\text{dB} = 10 \log_{10} \frac{P_1}{P_2}$$

If the voltage ratio is given, the number of decibels is equal to 20 times the common logarithm of the ratio. That is,

$$\text{dB} = 20 \log_{10} \frac{V_1}{V_2}$$

When a *voltage* ratio is used, both voltages must be measured across the same value of impedance. Unless this is done the decibel figure is meaningless, because it is fundamentally a measure of a *power* ratio.

The main reason a decibel is used is that successive power gains expressed in decibels may simply be added together. Thus a gain of 3 dB followed by a gain of 6 dB gives a total gain of 9 dB. In ordinary power ratios, the ratios must be multiplied together to find the total gain.

A *reduction* in power is handled simply by subtracting the requisite number of decibels. Thus, reducing the power to ½ is the same as *subtracting* 3 decibels. For example, a power gain of 4 in one part of a system and a reduction to ½ in another part gives a total power gain of 4 × ½ = 2. In decibels, this is 6 – 3 = 3 dB. A power reduction or *loss* is simply indicated by including a negative sign in front of the appropriate number of decibels.

Referenced to the horizon of the Earth, the H-plane pattern is now called the *elevation pattern*. Unlike the free-space H-plane pattern, the over-ground elevation pattern is drawn as a half-circle, representing only positive elevations above the Earth's surface. The ground reflects or blocks radiation at negative elevation angles, making below-surface radiation plots unnecessary.

After a little practice, and with the exercise of some imagination, the complete solid pattern can be visualized with fair accuracy from inspection of the two planar diagrams, provided of course that the solid pattern of the antenna is *smooth*, a condition that is true for simple antennas like λ/2 dipoles.

Plane diagrams are plotted on polar coordinate paper, as described earlier. The points on the pattern where the radiation is zero are called *nulls*. The curved section from one null to the next on the plane diagram, or the corresponding section on the solid pattern, is called a *lobe*. The strongest lobe is commonly called the *main lobe*. Fig 12A shows the E-plane pattern for a half-wave dipole. In Fig 12, the dipole is placed in free space. In addition to the labels showing the main lobe and nulls in the pattern, the so-called *half-power* points on the main lobe are shown. These are the points where the power is 3 dB down from the peak value in the main lobe.

Directivity and Gain

Let us now examine directivity more closely. As mentioned previously, all practical antennas, even the simplest types such as dipoles, exhibit directivity. Here's another picture that may help explain the concept of directivity. **Fig 13A** shows a balloon blown into its usual spherical shape. This represents a "reference" isotropic source. Squeezing the balloon in the middle in Fig 13B produces a dipole-like figure-8 pattern whose peak levels at top and bottom are larger than the reference sphere. Compare this with Fig 13C. Next, squeezing the bottom end of the balloon produces a pattern that gives even more "gain" compared to the reference.

Free-space directivity can be expressed quantitatively by comparing the three-dimensional pattern of the antenna under consideration with the perfectly spherical three-dimensional pattern of an isotropic antenna. The field strength (and thus power per unit area, or *power density*) is the same everywhere on the surface of an imaginary sphere having a radius of many wavelengths and having an isotropic antenna at its center. At the surface of the same imaginary sphere around an antenna radiating the same total power, the directive pattern results in greater power density at some points on this sphere and less at others. The ratio of the maximum power density to the average power density taken over the entire sphere (which is the same as from the isotropic antenna under the specified conditions) is the numerical measure of the directivity of the antenna. That is,

$$D = \frac{P}{P_{av}} \qquad \text{(Eq 5)}$$

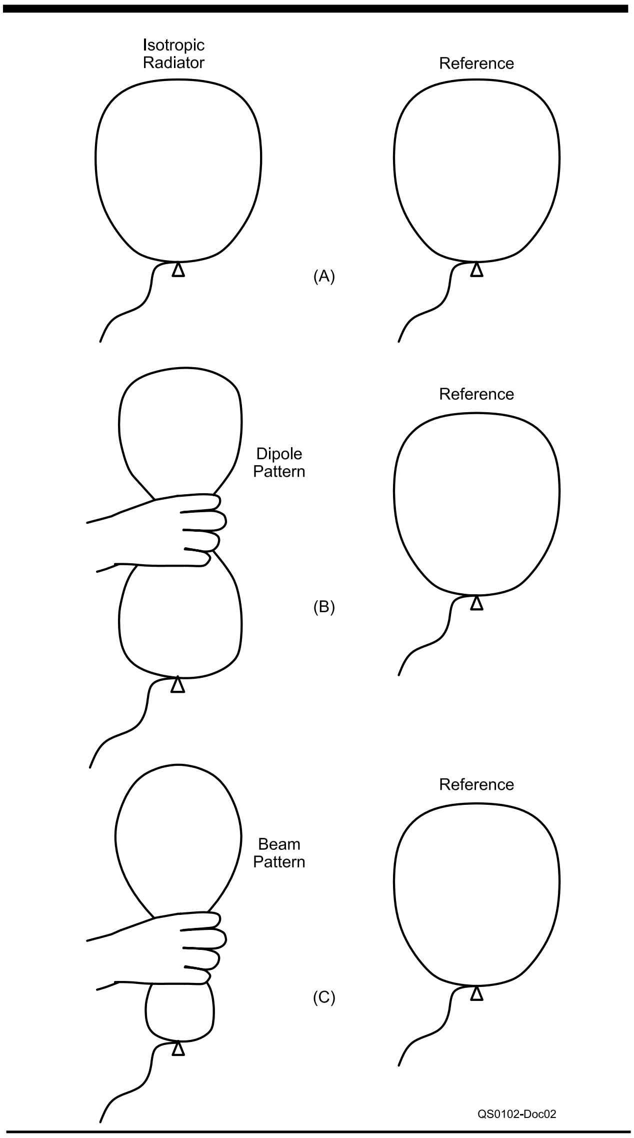

Fig 13—Demonstrating antenna pattern gain with balloons. Take a balloon, blow it up so that it is roughly circular in shape and then declare that this is a radiation pattern from an isotropic radiator. Next, blow up another balloon to the same size and shape and tell the audience that this will be the "reference" antenna (A). Then, squeeze the first balloon in the middle to form a sort of figure-8 shape and declare that this is a dipole and compare the maximum size to that of the reference "antenna" (B). The dipole can be seen to have some "gain" over the reference isotropic. Next, squeeze the end of the first balloon to come up with a sausage-like shape to demonstrate the sort of pattern a beam antenna creates (C).

where

D = directivity
P = power density at its maximum point on the surface of the sphere
P_{av} = average power density

The *gain* of an antenna is closely related to its directivity. Because directivity is based solely on the *shape* of the directive pattern, it does not take into account any power losses that may occur in an actual antenna system. To determine gain, these losses must be subtracted from the power supplied to the antenna. The loss is normally a constant percentage of the power input, so the antenna gain is

$$G=k\frac{P}{P_{av}}=kD \qquad \text{(Eq 6)}$$

where

G = gain (expressed as a power ratio)

D = directivity

k = efficiency (power radiated divided by power input) of the antenna

P and P_{av} are as above

For many of the antenna systems used by amateurs, the efficiency is quite high (the loss amounts to only a few percent of the total). In such cases the gain is essentially equal to the directivity. The more the directive diagram is compressed—or, in common terminology, the *sharper* the lobes—the greater the power gain of the antenna. This is a natural consequence of the fact that as power is taken away from a larger and larger portion of the sphere surrounding the radiator, it is added to the volume represented by the narrow lobes. Power is therefore concentrated in some directions, at the expense of others. In a general way, the smaller the volume of the solid radiation pattern, compared with the volume of a sphere having the same radius as the length of the largest lobe in the actual pattern, the greater the power gain.

As stated above, the gain of an antenna is related to its directivity, and directivity is related to the shape of the directive pattern. A commonly used index of directivity, and therefore the gain of an antenna, is a measure of the width of the major lobe (or lobes) of the plotted pattern. The width is expressed in degrees at the half-power or –3 dB points, and is often called the *beamwidth*.

This information provides only a general idea of relative gain, rather than an exact measure. This is because an absolute measure involves knowing the power density at every point on the surface of a sphere, while a single diagram shows the pattern shape in only one plane of that sphere. It is customary to examine at least the E-plane and the H-plane patterns before making any comparisons between antennas.

A simple approximation for gain over an isotropic radiator can be used, but only if the sidelobes in the antenna's pattern are small compared to the main lobe and if the resistive losses in the antenna are small. When the radiation pattern is complex, numerical integration is employed to give the actual gain.

$$G\approx\frac{41253}{H_{3dB}\times E_{3dB}} \qquad \text{(Eq 7)}$$

where H_{3dB} and E_{3dB} are the half-power points, in degrees, for the H and E-plane patterns.

Radiation Patterns for Center-Fed Dipoles at Different Frequencies

Earlier, we saw how the feed-point impedance of a fixed-length center-fed dipole in free space varies as the frequency is changed. What happens to the radiation pattern of such an antenna as the frequency is changed?

In general, the greater the length of a center-fed antenna, in terms of wavelength, the larger the number of lobes into which the pattern splits. A feature of all such patterns is the fact that the main lobe—the one that gives the largest field strength at a given distance—always is the one that makes the smallest angle with the antenna wire. Furthermore, this angle becomes smaller as the length of the antenna is increased.

Let's examine how the free-space radiation pattern changes for a 100-foot long wire made of #14 wire as the frequency is varied. (Varying the frequency effectively changes the wavelength for a fixed-length wire.) **Fig 14** shows the E-plane pattern at the λ/2 resonant frequency of 4.8 MHz. This is a classical dipole pattern, with a gain in free space of 2.14 dBi referenced to an isotropic radiator.

Fig 15 shows the free-space E-plane pattern for the same antenna, but now at the full-wave (2λ/2) resonant frequency of 9.55 MHz. Note how the pattern has been pinched in at the top and bottom of the figure. In other words, the two main lobes have become sharper at this frequency, making the gain 3.73 dBi, higher than at the λ/2 frequency.

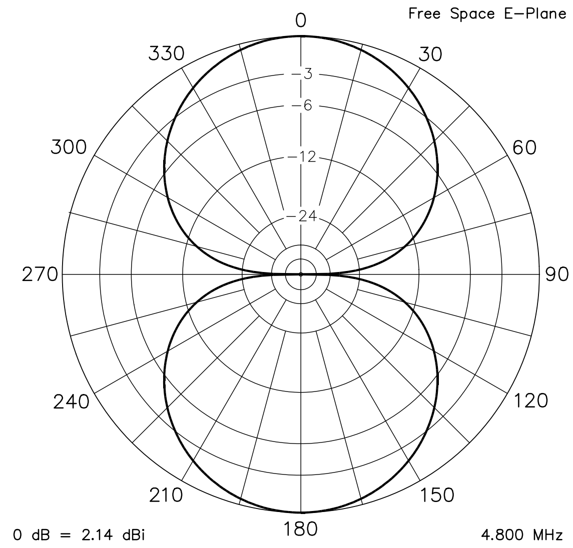

Fig 14—Free-space E-Plane radiation pattern for a 100-foot dipole at its half-wave resonant frequency of 4.80 MHz. This antenna has 2.14 dBi of gain. The dipole is located on the line from 90° to 270°.

Coordinate Scales for Radiation Patterns

A number of different systems of coordinate scales or *grids* are in use for plotting antenna patterns. Antenna patterns published for amateur audiences are sometimes placed on rectangular grids, but more often they are shown using polar coordinate systems. Polar coordinate systems may be divided generally into three classes: linear, logarithmic and modified logarithmic.

A very important point to remember is that the shape of a pattern (its general appearance) is highly dependent on the grid system used for the plotting. This is exemplified in **Fig A**, where the radiation pattern for a beam antenna is presented using three coordinate systems discussed in the paragraphs that follow.

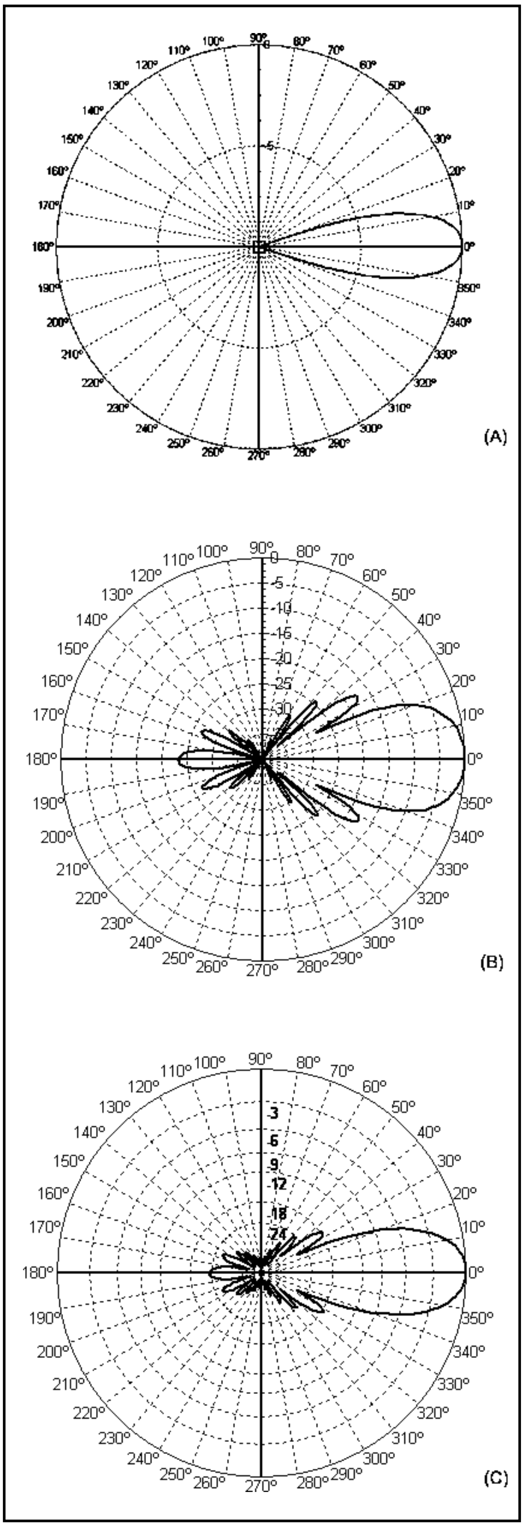

Linear Coordinate Systems

The polar coordinate system for the flashlight radiation pattern, Fig 10, uses linear coordinates. The concentric circles are equally spaced, and are graduated from 0 to 10. Such a grid may be used to prepare a linear plot of the power contained in the signal. For ease of comparison, the equally spaced concentric circles have been replaced with appropriately placed circles representing the decibel response, referenced to 0 dB at the outer edge of the plot. In these plots the minor lobes are suppressed. Lobes with peaks more than 15 dB or so below the main lobe disappear completely because of their small size. This is a good way to show the pattern of an array having high directivity and small minor lobes.

Logarithmic Coordinate System

Another coordinate system used by antenna manufacturers is the logarithmic grid, where the concentric grid lines are spaced according to the logarithm of the voltage in the signal. If the logarithmically spaced concentric circles are replaced with appropriately placed circles representing the decibel response, the decibel circles are graduated linearly. In that sense, the logarithmic grid might be termed a linear-log grid, one having linear divisions calibrated in decibels.

This grid enhances the appearance of the minor lobes. If the intent is to show the radiation pattern of an array supposedly having an omnidirectional response, this grid enhances that appearance. An antenna having a difference of 8 or 10 dB in pattern response around the compass appears to be closer to omnidirectional on this grid than on any of the others. See Fig A-(B).

ARRL Log Coordinate System

The modified logarithmic grid used by the ARRL has a system of concentric grid lines spaced according to the logarithm of 0.89 times the value of the signal voltage. In this grid, minor lobes that are 30 and 40 dB down from the main lobe are distinguishable. Such lobes are of concern in VHF and UHF work. The spacing between plotted points at 0 dB and –3 dB is significantly greater than the spacing between –20 and –23 dB, which in turn is significantly greater than the spacing between –50 and –53 dB.

For example, the scale distance covered by 0 to –3 dB is about $^1/_{10}$ of the radius of the chart. The scale distance for the next 3-dB increment (to –6 dB) is slightly less, 89% of the first, to be exact. The scale distance for the next 3-dB increment (to –9 dB) is again 89% of the second. The scale is constructed so that the progression ends with –100 dB at chart center.

The periodicity of spacing thus corresponds generally to the relative significance of such changes in antenna performance. Antenna pattern plots in this publication are made on the modified-log grid similar to that shown in Fig A-(C).

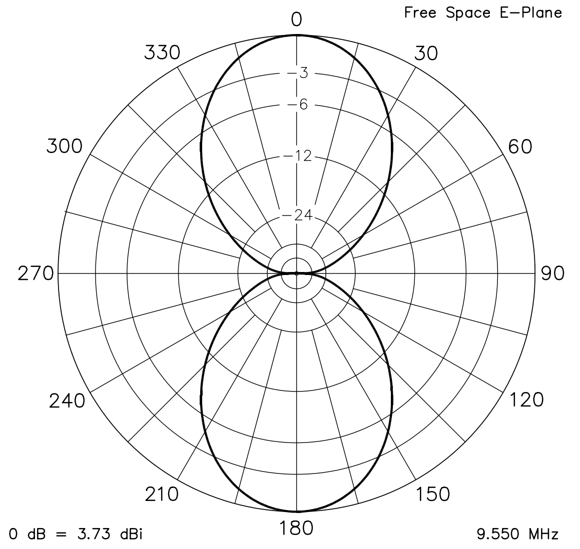

Fig 15—Free-space E-Plane radiation pattern for a 100-foot dipole at its full-wave resonant frequency of 9.55 MHz. The gain has increased to 3.73 dBi, because the main lobes have been focused and sharpened compared to Fig 13.

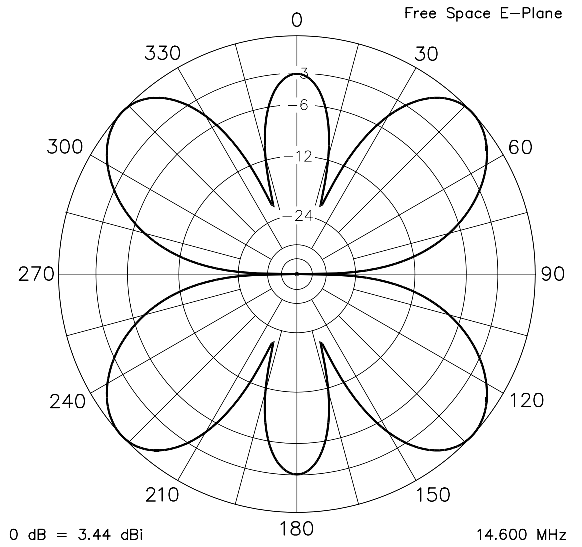

Fig 16—Free-space E-Plane radiation pattern for a 100-foot dipole at its 3/2λ resonant frequency of 14.60 MHz. The pattern has broken up into six lobes, and thus the peak gain has dropped to 3.44 dBi.

Fig 16 shows the pattern at the 3λ/2 frequency of 14.6 MHz. More lobes have developed compared to Fig 14. This means that the power has split up into more lobes and consequently the gain decreases a small amount, down to 3.44 dBi. This is still higher than the dipole at its λ/2 frequency, but lower than at its full-wave frequency. **Fig 17** shows the E-plane response at 19.45 MHz, the 4λ/2, or 2λ, resonant frequency. Now the pattern has reformed itself into only four lobes, and the gain has as a consequence risen to 3.96 dBi.

In **Fig 18** the response has become quite complex at the 5λ/2 resonance point of 24.45 MHz, with ten lobes showing. Despite the presence all these lobes, the main lobes now show a gain of 4.78 dBi. Finally, **Fig 19** shows the pattern at the 3λ (6λ/2) resonance at 29.45 MHz. Despite the fact that there are fewer lobes taking up power than at 24.45 MHz, the peak gain is slightly less at 29.45 MHz, at 4.70 dBi.

Fig A—Radiation pattern plots for a high-gain Yagi antenna on three different grid coordinate systems. At A, the pattern on a linear-power dB grid. Notice how details of sidelobe structure are lost with this grid. At B, the same pattern on a grid with constant 5 dB circles. The sidelobe level is exaggerated when this scale is employed. At B, the same pattern on the modified log grid used by ARRL. The side and rearward lobes are clearly visible on this grid. The concentric circles in all three grids are graduated in decibels referenced to 0 dB at the outer edge of the chart. The patterns look quite different, yet they all represent the same antenna response!

The pattern—and hence the gain—of a fixed-length antenna varies considerably as the frequency is changed. Of course, the pattern and gain change in the same fashion if the frequency is kept constant and the length of the wire is varied. In either case, the wavelength is changing. It is also evident that certain lengths reinforce the pattern to provide more peak gain. If an antenna is not rotated in azimuth when the frequency is changed, the peak gain may occur in a different direction than you might like. In other words, the main lobes change direction as the frequency is varied.

POLARIZATION

We've now examined the first two of the three major properties used to characterize antennas: the radiation pattern and the feed-point impedance. The third general property is *polarization*. An antenna's polarization is defined to be that of its electric field, in the direction where the field strength is maximum.

For example, if a λ/2 dipole is mounted horizontally over the Earth, the electric field is strongest perpendicular to its axis (that is, at right angle to the wire) and parallel to the earth. Thus, since the maximum electric field is horizontal, the polarization in this case is also considered to be *horizontal* with respect to the earth. If the dipole is mounted vertically, its polarization will be *vertical*. See **Fig 20**. Note that if an antenna is mounted in free space, there is no frame of reference and hence its

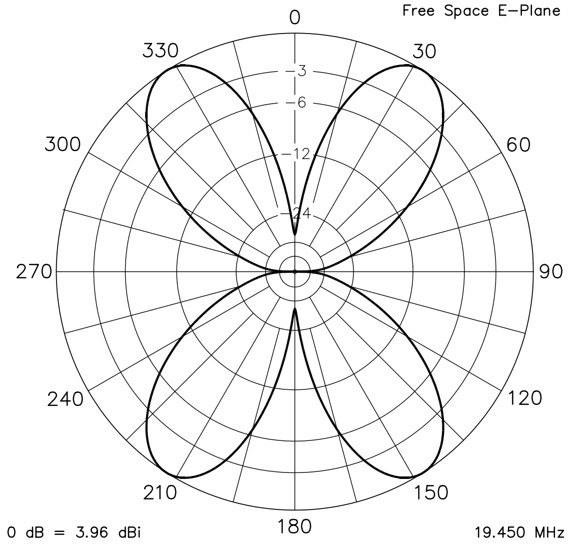

Fig 17—Free-space E-Plane radiation pattern for a 100-foot dipole at twice its full-wave resonant frequency of 19.45 MHz. The pattern has been refocused into four lobes, with a peak gain of 3.96 dBi.

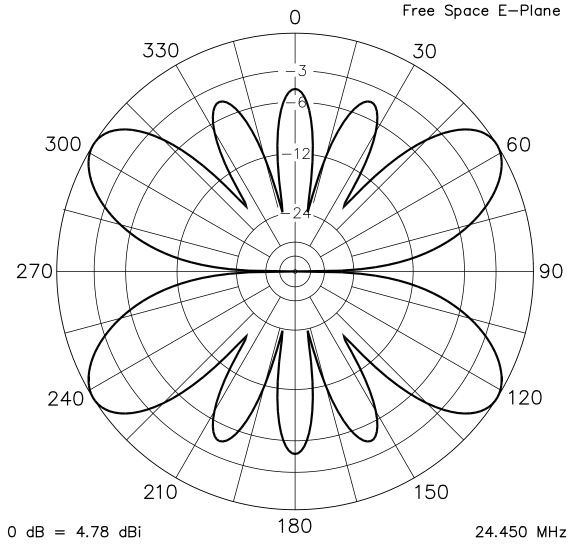

Fig 18—Free-space E-Plane radiation pattern for a 100-foot dipole at its 5/2λ resonant frequency of 24.45 MHz. The pattern has broken down into ten lobes, with a peak gain of 4.78 dBi.

polarization is indeterminate.

Antennas composed of a number of λ/2 elements arranged so that their axes lie in the same or parallel directions have the same polarization as that of any one of the elements. For example, a system composed of a group of horizontal dipoles is horizontally polarized. If both horizontal and vertical elements are used in the same plane and radiate in phase, however, the polarization is the

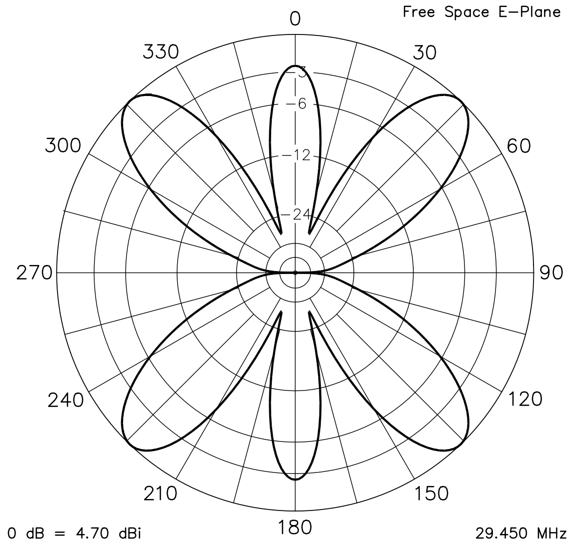

Fig 19—Free-space E-Plane radiation pattern for a 100-foot dipole at three times its full-wave resonant frequency of 29.45 MHz. The pattern has returned to six lobes, with a peak gain of 4.70 dBi.

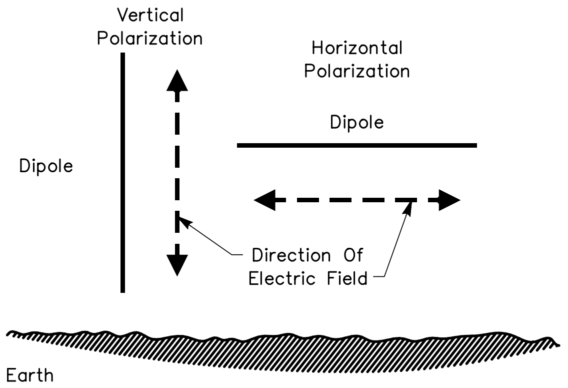

Fig 20—Vertical and horizontal polarization of a dipole above ground. The direction of polarization is the direction of the maximum electric field with respect to the earth.

resultant of the contributions made by each set of elements to the total electromagnetic field at a given point some distance from the antenna. In such a case the resultant polarization is still *linear,* but is tilted between horizontal and vertical.

In directions other than those where the radiation is maximum, the resultant wave even for a simple dipole is a combination of horizontally and vertically polarized components. The radiation off the ends of a horizontal dipole is actually vertically polarized, albeit at a greatly reduced amplitude compared to the broadside horizontally polarized radiation—the sense of polarization changes with compass direction.

Thus it is often helpful to consider the radiation pattern from an antenna in terms of polar coordinates, rather than trying to think in purely linear horizontal or vertical coordinates. See **Fig 21**. The reference axis in a polar system is vertical to the earth under the antenna. The zenith angle is usually referred to as θ (Greek letter theta), and the azimuth angle is referred to as ϕ (Greek letter phi). Instead of zenith angles, most amateurs are more familiar with *elevation angles*, where a zenith angle of 0° is the same as an elevation angle of 90°, straight overhead. Native *NEC* or *MININEC* computer programs use zenith angles rather than elevation angles, although most commercial versions automatically reduce these to elevation angles.

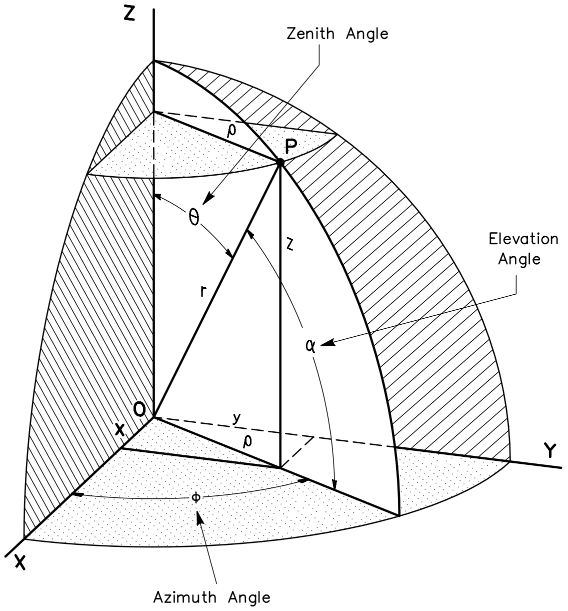

Fig 21—Diagram showing polar representation of a point P lying on an imaginary sphere surround a point-source antenna. The various angles associated with this coordinate system are shown referenced to the x, y and z-axes.

If vertical and horizontal elements in the same plane are fed out of phase (where the beginning of the RF period applied to the feed point of the vertical element is not in time phase with that applied to the horizontal), the resultant polarization is *elliptical*. Circular polarization is a special case of elliptical polarization. The wave front of a circularly polarized signal appears (in passing a fixed observer) to rotate every 90° between vertical and horizontal, making a complete 360° rotation once every period. Field intensities are *equal* at all instantaneous polarizations. Circular polarization is frequently used for space communications, and is discussed further in Chapter 19, Antenna Systems for Space Communications.

Sky-wave transmission usually changes the polarization of traveling waves. (This is discussed in Chapter 23, Radio Wave Propagation.) The polarization of receiving and transmitting antennas in the 3 to 30-MHz range, where almost all communication is by means of sky wave, need not be the same at both ends of a communication circuit (except for distances of a few miles). In this range the choice of polarization for the antenna is usually determined by factors such as the height of available antenna supports, polarization of man-made RF noise from nearby sources, probable energy losses in nearby objects, the likelihood of interfering with neighborhood broadcast or TV reception and general convenience.

Other Antenna Characteristics

Besides the three main characteristics of impedance, pattern (gain) and polarization, there are some other useful properties of antennas.

RECIPROCITY IN RECEIVING AND TRANSMITTING

Many of the properties of a resonant antenna used for reception are the same as its properties in transmission. It has the same directive pattern in both cases, and delivers maximum signal to the receiver when the signal comes from a direction in which the antenna has its best response. The impedance of the antenna is the same, at the same point of measurement, in receiving as in transmitting.

In the receiving case, the antenna is the source of power delivered to the receiver, rather than the load for a source of power (as in transmitting). Maximum possible output from the receiving antenna is obtained when the load to which the antenna is connected is the same as the impedance of the antenna. We say that the antenna is *matched* to its load.

The power gain in receiving is the same as the gain in transmitting, when certain conditions are met. One such condition is that both antennas (usually λ/2-long antennas) must work into load impedances matched to their own impedances, so that maximum power is transferred in both cases. In addition, the comparison antenna should be oriented so it gives maximum response to the signal used in the test. That is, it should have the same polarization as the incoming signal and should be placed so its direction of maximum gain is toward the signal source.

In long-distance transmission and reception via the ionosphere, the relationship between receiving and transmitting, however, may not be exactly reciprocal. This is because the waves do not always follow exactly the same paths at all times and so may show considerable variation in the time between alternations between transmitting and receiving. Also, when more than one ionospheric layer is involved in the wave travel (see Chapter 23, Radio Wave Propagation), it is sometimes possible for reception to be good in one direction and poor in the other, over the same path.

Wave polarization usually shifts in the ionosphere. The tendency is for the arriving wave to be elliptically polarized, regardless of the polarization of the transmitting antenna. Vertically polarized antennas can be expected to show no more difference between transmission and reception than horizontally polarized antennas. On the average, however, an antenna that transmits well in a certain direction also gives favorable reception from the same direction, despite ionospheric variations.

FREQUENCY SCALING

Any antenna design can be scaled in size for use on another frequency or on another amateur band. The dimensions of the antenna may be scaled with Eq 8 below.

$$D = \frac{f1}{f2} \times d \qquad \text{(Eq 8)}$$

where

D = scaled dimension

d = original design dimension

f1 = original design frequency

f2 = scaled frequency (frequency of intended operation)

From this equation, a published antenna design for, say, 14 MHz can be scaled in size and constructed for operation on 18 MHz, or any other desired band. Similarly, an antenna design could be developed experimentally at VHF or UHF and then scaled for operation in one of the HF bands. For example, from Eq 8, an element of 39.0 inches length at 144 MHz would be scaled to 14 MHz as follows: D = 144/14 × 39 = 401.1 inches, or 33.43 feet.

To scale an antenna properly, *all* physical dimensions must be scaled, including element lengths, element spacings, boom diameters and element diameters. Lengths and spacings may be scaled in a straightforward manner as in the above example, but element diameters are often not as conveniently scaled. For example, assume a 14-MHz antenna is modeled at 144 MHz and perfected with $^3/_8$-inch cylindrical elements. For proper scaling to 14 MHz, the elements should be cylindrical, of 144/14 × $^3/_8$ or 3.86 inches diameter. From a realistic standpoint, a 4-inch diameter might be acceptable, but cylindrical elements of 4-inch diameter in lengths of 33 feet or so would be quite unwieldy (and quite expensive, not to mention heavy). Choosing another, more suitable diameter is the only practical answer.

Diameter Scaling

Simply changing the diameter of dipole type elements during the scaling process is not satisfactory without making a corresponding element-length correction. This is because changing the diameter results in a change in the λ/dia ratio from the original design, and this alters the corresponding resonant frequency of the element. The element length must be corrected to compensate for the effect of the different diameter actually used.

To be more precise, however, the purpose of diameter scaling is not to maintain the same resonant frequency for the element, but to maintain the same ratio of self-resistance to self-reactance at the operating frequency—that is, the Q of the scaled element should be the same as that of the original element. This is not always possible to achieve exactly for elements that use several telescoping sections of tubing.

Tapered Elements

Rotatable beam antennas are usually constructed with elements made of metal tubing. The general practice at HF is to taper the elements with lengths of telescoping tubing. The center section has a large diameter, but the ends are relatively small. This reduces not only the weight, but also the cost of materials for the elements. Tapering of HF Yagi elements is discussed in detail in Chapter 11, HF Yagi Arrays.

Length Correction for Tapered Elements

The effect of tapering an element is to alter its electrical length. That is to say, two elements of the same length, one cylindrical and one tapered but with the same average diameter as the cylindrical element, will not be resonant at the same frequency. The tapered element must be made longer than the cylindrical element for the same resonant frequency.

A procedure for calculating the length for tapered elements has been worked out by Dave Leeson, W6NL (ex-W6QHS), from work done by Schelkunoff at Bell Labs and is presented in Leeson's book, *Physical Design of Yagi Antennas*. In the software accompanying this book is a subroutine called EFFLEN.FOR. It is written in Fortran and is used in the *SCALE* program to compute the effective length of a tapered element. The algorithm uses the W6NL-Schelkunoff algorithm and is commented step-by-step to show what is happening. Calculations are made for only one half of an element, assuming the element is symmetrical about the point of boom attachment.

Also, read the documentation SCALE.PDF for the *SCALE* program, which will automatically do the complex mathematics to scale a Yagi design from one frequency to another, or from one taper schedule to another.

The Vertical Monopole

So far in this discussion on Antenna Fundamentals, we have been using the free-space, center-fed dipole as our main example. Another simple form of antenna derived from a dipole is called a *monopole*. The name suggests that this is one half of a dipole, and so it is. The monopole is always used in conjunction with a *ground plane*, which acts as a sort of electrical mirror. See **Fig 22**, where a λ/2 dipole and a λ/4 monopole are compared. The *image antenna* for the monopole is the dotted line beneath the ground plane. The image forms the missing second half of the antenna, transforming a monopole into the functional equivalent of a dipole. From this explanation you can see where the term *image plane* is sometimes used instead of ground plane.

Although we have been focusing throughout this chapter on antennas in free space, practical monopoles are usually mounted vertically with respect to the surface of the ground. As such, they are called *vertical monopoles*, or simply *verticals*. A practical vertical is supplied power by feeding the radiator against a ground system, usually made up of a series of paralleled wires radiating from and laid out in a circular pattern around the base of the antenna. These wires are termed *radials*.

The term *ground plane* is also used to describe a vertical antenna employing a λ/4-long vertical radiator working against a *counterpoise* system, another name for the ground plane that supplies the missing half of the antenna. The counterpoise for a ground-plane antenna usually consists of four λ/4-long radials elevated well above the earth. See **Fig 23**.

Chapter 3, The Effects of Ground, devotes much attention to the requirements for an efficient grounding system for vertical monopole antennas, and Chapter 6, Low-Frequency Antennas, gives more information on ground-plane verticals.

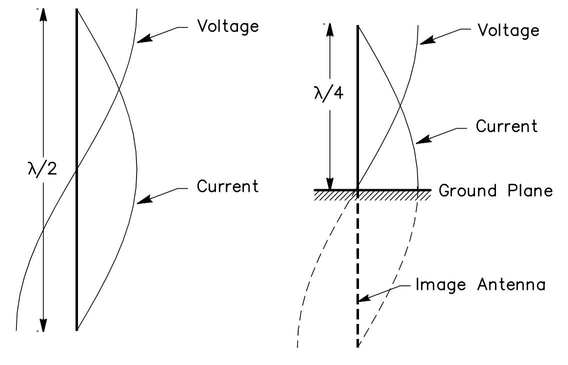

Fig 22—The λ/2 dipole antenna and its λ/4 ground-plane counterpart. The "missing" quarter wavelength is supplied as an image in "perfect" (that is, high-conductivity) ground.

Characteristics of a λ/4 Monopole

The free-space directional characteristics of a λ/4 monopole with its ground plane are very similar to that of a λ/2 antenna in free space. The gain for the λ/4 monopole is slightly less because the H-plane for the λ/2 antenna is compressed compared to the monopole. Like a λ/2 antenna, the λ/4 monopole has an *omnidirectional* radiation pattern in the plane perpendicular to the monopole.

The current in a λ/4 monopole varies practically sinusoidally (as is the case with a λ/2 wire), and is highest at the ground-plane connection. The RF voltage is highest at the open (top) end and minimum at the ground plane. The feed-point resistance close to λ/4 resonance of a vertical monopole over a perfect ground plane is one-half that for a λ/2 dipole at its λ/2 resonance. In this case, a "perfect ground plane" is an infinitely large, lossless conductor.

See **Fig 24**, which shows the feed-point impedance of a vertical antenna made of #14 wire, 50 feet long, located over perfect ground. This is over the whole HF range from 1 to 30 MHz. Again, there is nothing special about the choice of 50 feet for the length of the vertical radiator; it is simply a convenient length for evaluation. **Fig 25** shows an expanded portion of the frequency range above and below the λ/4 resonant point, but now calibrated in terms of wavelength. Note that this particular antenna goes through λ/4 resonance at a length of 0.244 λ, not at exactly 0.25 λ. The exact length for resonance varies with the diameter of the wire used, just as it does for the λ/2 dipole at its λ/2 resonance.

The word *height* is usually used for a vertical monopole antenna whose base is on or near the ground, and in this context, height has the same meaning as *length* when applied to λ/2 dipole antennas. Older texts often refer to heights in electrical degrees, referenced to a free-space wavelength of 360°, but here height is expressed in terms of the free-space wavelength. The range shown in Fig 24 is from 0.132 λ to 0.300 λ, corresponding to a frequency range of 2.0 to 5.9 MHz.

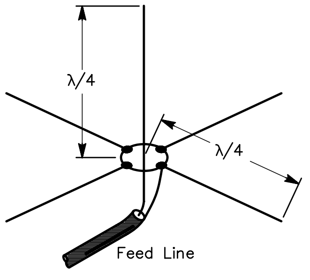

Fig 23—The ground-plane antenna. Power is applied between the base of the vertical radiator and the center of the four ground plane wires.

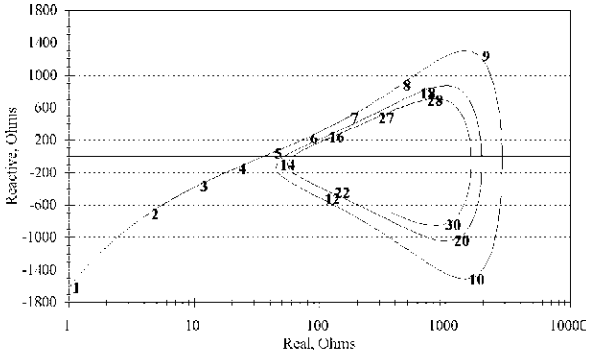

Fig 24—Feed-point impedance versus frequency for a theoretical 50-foot-high grounded vertical monopole made of #14 wire. The numbers along the curve show the frequency in MHz. This was computed using "perfect" ground. Real ground losses will add to the feed-point impedance shown in an actual antenna system.

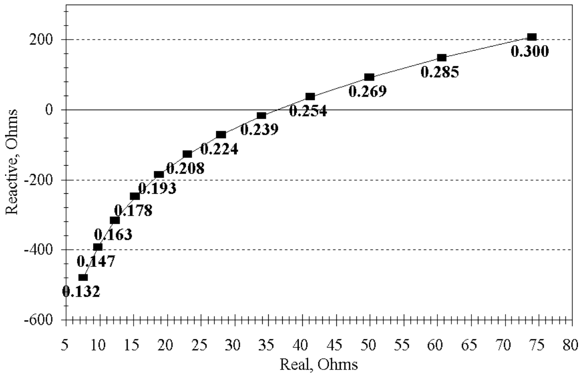

Fig 25—Feed-point impedance for the same antennas as in Fig 21, but calibrated in wavelength rather than frequency, over the range from 0.132 to 0.300λ, above and below the quarter-wave resonance.

The reactive portion of the feed-point impedance depends highly on the length/dia ratio of the conductor, as was discussed previously for a horizontal center-fed dipole. The impedance curve in Figs 24 and 25 is based on a #14 conductor having a length/dia ratio of about 800 to 1. As usual, thicker antennas can be expected to show less reactance at a given height, and thinner antennas will show more.

Efficiency of Vertical Monopoles

This topic of the efficiency of vertical monopole systems will be covered in detail in Chapter 3, The Effects of Ground, but it is worth noting at this point that the efficiency of a real vertical antenna over real earth often suffers dramatically compared with that of a λ/2 antenna. Without a fairly elaborate grounding system, the efficiency is not likely to exceed 50%, and it may be much less, particularly at monopole heights below λ/4.

BIBLIOGRAPHY

J. S. Belrose, "Short Antennas for Mobile Operation," *QST*, Sep 1953.

G. H. Brown, "The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas," *Proc IRE*, Feb 1935.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc IRE*, Jun 1937, pp 753-787.

G. H. Brown and O. M. Woodward, Jr., "Experimentally Determined Impedance Characteristics of Cylindrical Antennas," *Proc IRE*, April 1945.

A. Christman, "Elevated Vertical Antenna Systems," *QST*, Aug 1988, pp 35-42.

R. B. Dome, "Increased Radiating Efficiency for Short Antennas," *QST*, Sep 1934, pp 9-12.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Characteristics of the Counterpoise and Elevated Ground Screen," Professional Program, Session 9, Southcon '83 (IEEE), Atlanta, GA, Jan 1983.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Efficient Ground Systems for Vertical Antennas," *QST*, Feb 1983, pp 20-25.

A. C. Doty, Jr., technical paper presentation, "Capacitive Bottom Loading and Other Aspects of Vertical Antennas," Technical Symposium, Radio Club of America, New York City, Nov 20, 1987.

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Vertical Antennas: New Design and Construction Data," *The ARRL Antenna Compendium, Volume 2* (Newington: ARRL, 1989), pp 2-9.

R. Fosberg, "Some Notes on Ground Systems for 160 Meters," *QST*, Apr 1965, pp 65-67.

G. Grammer, "More on the Directivity of Horizontal Antennas; Harmonic Operation—Effects of Tilting," *QST*, Mar 1937, pp 38-40, 92, 94, 98.

H. E. Green, "Design Data for Short and Medium Length Yagi-Uda Arrays," *Trans IE Australia*, Vol EE-2, No. 1, Mar 1966.

H. J. Mills, technical paper presentation, "Impedance Transformation Provided by Folded Monopole Antennas," Technical Symposium, Radio Club of America, New York City, Nov 20, 1987.

B. Myers, "The W2PV Four-Element Yagi," *QST*, Oct 1986, pp 15-19.

L. Richard, "Parallel Dipoles of 300-Ohm Ribbon," *QST*, Mar 1957.

J. H. Richmond, "Monopole Antenna on Circular Disc," *IEEE Trans on Antennas and Propagation*, Vol. AP-32, No. 12, Dec 1984.

W. Schulz, "Designing a Vertical Antenna," *QST*, Sep 1978, pp 19- 21.

J. Sevick, “The Ground-Image Vertical Antenna,” *QST*, Jul 1971, pp 16-17, 22.

J. Sevick, “The W2FMI 20-Meter Vertical Beam,” *QST*, Jun 1972, pp 14-18.

J. Sevick, “The W2FMI Ground-Mounted Short Vertical,” *QST*, Mar 1973, pp. 13-18, 41.

J. Sevick, “A High Performance 20-, 40- and 80-Meter Vertical System,” *QST*, Dec 1973.

J. Sevick, “Short Ground-Radial Systems for Short Verticals,” *QST*, Apr 1978, pp 30-33.

J. Sevick, *Transmission Line Transformers*, 3rd ed. (Tucker, GA: Noble Publishing, 1996).

C. E. Smith and E. M. Johnson, “Performance of Short Antennas,” *Proc IRE*, Oct 1947.

J. Stanley, “Optimum Ground Systems for Vertical Antennas,” *QST*, Dec 1976, pp 13-15.

R. E. Stephens, “Admittance Matching the Ground-Plane Antenna to Coaxial Transmission Line,” Technical Correspondence, *QST*, Apr 1973, pp 55-57.

D. Sumner, “Cushcraft 32-19 ‘Boomer’ and 324-QK Stacking Kit,” Product Review, *QST*, Nov 1980, pp 48-49.

W. van B. Roberts, “Input Impedance of a Folded Dipole,” *RCA Review*, Jun 1947.

E. M. Williams, “Radiating Characteristics of Short-Wave Loop Aerials,” *Proc IRE*, Oct 1940.

TEXTBOOKS ON ANTENNAS

C. A. Balanis, *Antenna Theory, Analysis and Design* (New York: Harper & Row, 1982).

D. S. Bond, *Radio Direction Finders*, 1st ed. (New York: McGraw-Hill Book Co).

W. N. Caron, *Antenna Impedance Matching* (Newington: ARRL, 1989).

K. Davies, *Ionospheric Radio Propagation*—National Bureau of Standards Monograph 80 (Washington, DC: U.S. Government Printing Office, Apr 1, 1965).

R. S. Elliott, *Antenna Theory and Design* (Englewood Cliffs, NJ: Prentice Hall, 1981).

A. E. Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc, 1941).

K. Henney, *Principles of Radio* (New York: John Wiley and Sons, 1938), p 462.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

W. C. Johnson, *Transmission Lines and Networks*, 1st ed. (New York: McGraw-Hill Book Co, 1950).

R. C. Johnson and H. Jasik, *Antenna Engineering Handbook*, 2nd ed. (New York: McGraw-Hill, 1984).

R. C. Johnson, *Antenna Engineering Handbook*, 3rd ed. (New York: McGraw-Hill, 1993).

E. C. Jordan and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, Inc, 1968).

R. Keen, *Wireless Direction Finding*, 3rd ed. (London: Wireless World).

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ. Press, 1956).

R. W. P. King, H. R. Mimno and A. H. Wing, *Transmission Lines, Antennas and Waveguides* (New York: Dover Publications, Inc, 1965).

King, Mack and Sandler, *Arrays of Cylindrical Dipoles* (London: Cambridge Univ Press, 1968).

M. G. Knitter, Ed., *Loop Antennas—Design and Theory* (Cambridge, WI: National Radio Club, 1983).

M. G. Knitter, Ed., *Beverage and Long Wire Antennas*—Design and Theory (Cambridge, WI: National Radio Club, 1983).

J. D. Kraus, *Electromagnetics* (New York: McGraw-Hill Book Co).

J. D. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co, 1988).

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co, 1952).

J. L. Lawson, *Yagi-Antenna Design*, 1st ed. (Newington: ARRL, 1986).

P. H. Lee, *The Amateur Radio Vertical Antenna Handbook*, 2nd ed. (Port Washington, NY: Cowen Publishing Co., 1984).

A. W. Lowe, *Reflector Antennas* (New York: IEEE Press, 1978).

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington: ARRL, 1990). Out of print.

M. W. Maxwell, *Reflections II—Transmission Lines and Antennas* (Sacramento: Worldradio Books, 2001).

G. M. Miller, *Modern Electronic Communication* (Englewood Cliffs, NJ: Prentice Hall, 1983).

V. A. Misek, *The Beverage Antenna Handbook* (Hudson, NH: V. A. Misek, 1977).

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw-Hill, 1948).

L. A. Moxon, *HF Antennas for All Locations* (Potters Bar, Herts: Radio Society of Great Britain, 1982), pp 109-111.

Ramo and Whinnery, *Fields and Waves in Modern Radio* (New York: John Wiley & Sons).

V. H. Rumsey, *Frequency Independent Antennas* (New York: Academic Press, 1966).

P. N. Saveskie, *Radio Propagation Handbook* (Blue Ridge Summit, PA: Tab Books, Inc, 1980).

S. A. Schelkunoff, *Advanced Antenna Theory* (New York: John Wiley & Sons, Inc, 1952).

S. A. Schelkunoff and H. T. Friis, *Antennas Theory and Practice* (New York: John Wiley & Sons, Inc, 1952).

J. Sevick, *Transmission Line Transformers* (Atlanta: Noble Publishing, 1996).

H. H. Skilling, *Electric Transmission Lines* (New York: McGraw-Hill Book Co, Inc, 1951).

M. Slurzburg and W. Osterheld, *Electrical Essentials of Radio* (New York: McGraw-Hill Book Co, Inc, 1944).

G. Southworth, *Principles and Applications of Waveguide*

Transmission (New York: D. Van Nostrand Co, 1950).

F. E. Terman, *Radio Engineers' Handbook*, 1st ed. (New York, London: McGraw-Hill Book Co, 1943).

F. E. Terman, *Radio Engineering*, 3rd ed. (New York: McGraw-Hill, 1947).

S. Uda and Y. Mushiake, *Yagi-Uda Antenna* (Sendai, Japan: Sasaki Publishing Co, 1954). [Published in English—Ed.]

P. P. Viezbicke, "*Yagi Antenna Design*," NBS Technical Note 688 (U. S. Dept of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

G. B. Welch, *Wave Propagation and Antennas* (New York: D. Van Nostrand Co, 1958), pp 180-182.

The GIANT Book of Amateur Radio Antennas (Blue Ridge Summit, PA: Tab Books, 1979), pp 55-85.

IEEE Standard Dictionary of Electrical and Electronics Terms, 3rd ed. (New York: IEEE, 1984).

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

Radio Communication Handbook, 5th ed. (London: RSGB, 1976).

Radio Direction Finding, published by the Happy Flyer, 1811 Hillman Ave, Belmont, CA 94002

Chapter 3

The Effects of Ground

The ground around and under an antenna is part of the environment in which any actual antenna must operate. Chapter 2, Antenna Fundamentals, dealt mainly with theoretical antennas in free space, completely removed from the influence of the ground. This chapter is devoted to exploring the interactions between antennas and the ground.

The interactions can be analyzed depending on where they occur relative to two areas surrounding the antenna: the *reactive near field* and the *radiating far field*. You will recall that the reactive near field only exists very close to the antenna itself. In this region the antenna acts as though it were a large lumped-constant inductor or capacitor, where energy is stored but very little is actually radiated. The interaction with the ground in this area creates mutual impedances between the antenna and its environment and these interactions not only modify the feed-point impedance of an antenna, but also often increase losses.

In the radiating far field, the presence of ground profoundly influences the radiation pattern of a real antenna. The interaction is different, depending on the antenna's polarization with respect to the ground. For horizontally polarized antennas, the *shape* of the radiated pattern in the elevation plane depends primarily on the antenna's height above ground. For vertically polarized antennas, both the *shape* and the *strength* of the radiated pattern in the elevation plane strongly depend on the nature of the ground itself (its dielectric constant and conductivity at the frequency of operation), as well as on the height of the antenna above ground.

The Effects of Ground in the Reactive Near Field

FEED-POINT IMPEDANCE VERSUS HEIGHT ABOVE GROUND

Waves radiated from the antenna directly downward reflect vertically from the ground and, in passing the antenna on their upward journey, induce a voltage in it. The magnitude and phase of the current resulting from this induced voltage depends on the height of the antenna above the reflecting surface.

The total current in the antenna consists of two components. The amplitude of the first is determined by the power supplied by the transmitter and the free-space feed-point resistance of the antenna. The second component is induced in the antenna by the wave reflected from the ground. This second component of current, while considerably smaller than the first at most useful antenna heights, is by no means insignificant. At some heights, the two components will be in phase, so the total current is larger than is indicated by the free-space feed-point resistance. At other heights, the two components are out of phase, and the total current is the difference between the two components.

Changing the height of the antenna above ground will change the amount of current flow, assuming that the power input to the antenna is constant. A higher current at the same power input means that the effective resistance of the antenna is lower, and vice versa. In other words, the feed-point resistance of the antenna is affected by the height of the antenna above ground because of mutual coupling between the antenna and the ground beneath it.

The electrical characteristics of the ground affect both the amplitude and the phase of reflected signals. For this reason, the electrical characteristics of the ground under the antenna will have some effect on the impedance of that antenna, the reflected wave having been influenced by the ground. Different impedance values may be encountered when an antenna is erected at identical heights but over different types of earth.

Fig 1 shows the way in which the radiation resistance of horizontal and vertical half-wave antennas varies with height above ground (in λ, wavelengths). The height of the vertical half-wave is the distance from the bottom of the antenna to ground. For horizontally polarized half-wave antennas, the differences between the effects of perfect ground and real earth are negligible if the antenna height is greater than 0.2 λ. At lower heights, the feed-point resistance over perfect ground decreases rapidly as the antenna is brought closer to a theoretically perfect ground, but this does not occur so rapidly for actual ground. Over real earth, the resistance actually begins increasing at heights below about 0.08 λ. The reason for the increasing resistance at very low heights is that more and more of the reactive (induction) field of the antenna is absorbed by the lossy ground in close proximity. This results in increased loss that is reflected in the increased value of the feedpoint resistance.

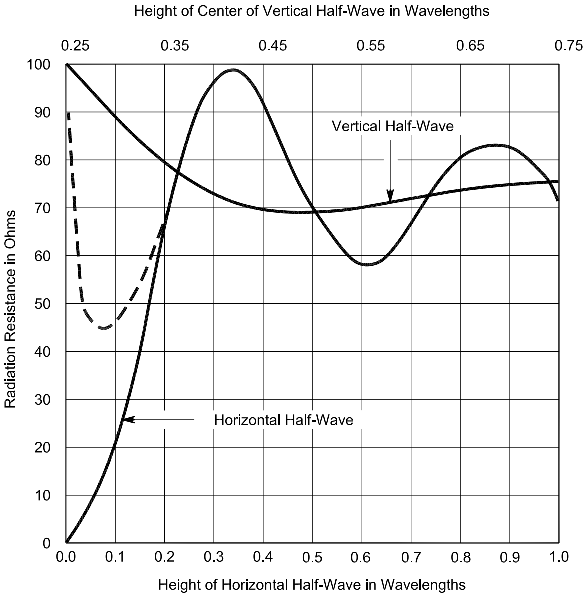

Fig 1—Variation in radiation resistance of vertical and horizontal half-wave antennas at various heights above flat ground. Solid lines are for perfectly conducting ground; the broken line is the radiation resistance of horizontal half-wave antennas at low height over real ground.

For a vertically polarized λ/2-long dipole, differences between the effects of perfect ground and real earth on the feed-point impedance is negligible, as seen in **Fig 1**. The theoretical half-wave antennas on which this chart is based are assumed to have infinitely thin conductors.

GROUND SYSTEMS FOR VERTICAL MONOPOLES

In this section, we'll look at vertical monopoles, which require some sort of ground system in order to make up for the "missing" second half of the antenna and reduce the power lost in the near field. Rudy Severns, N6LF, contributed much of the new material in this chapter.

In Chapter 2, Antenna Fundamentals, and up to this point in this chapter, the discussion about vertical monopoles has mainly been for antennas where *perfect ground* is available. We have also briefly looked at the ground-plane vertical in free space, where the four ground-plane radials form a built-in ground system.

Perfect ground makes a vertical monopole into the functional equivalent of a center-fed dipole, although the feed-point resistance at resonance is half that of the center-fed dipole. But how can we manage to create that elusive perfect ground, or at least a reasonable approximation, for our real vertical antennas?

Simulating a Perfect Ground in the Reactive Near Field

The effect of a perfectly conducting ground (so far as feed-point resistance and losses are concerned) can be simulated under a real antenna by installing a very large metal screen or mesh, such as poultry netting (chicken wire) or hardware cloth, on or near the surface of the ground. The screen (also called a *counterpoise system*, especially if it is elevated off the ground) should extend at least a half wavelength in every direction from the antenna. The feed-point resistance of a quarter-wave long, thin vertical radiator over such a ground screen will approach the theoretical value of 36.6 Ω. Of course on the lower HF bands such a screen is not practical for most amateurs.

Based on the results of a study published in 1937 by Brown, Lewis and Epstein (see Bibliography), a grounding system consisting of 120 wires, each at least λ/2 long, extending radially from the base of the antenna and spaced equally around a circle, is also the practical equivalent of perfectly conducting ground for reactive-field currents. The wires can either be laid directly on the surface of the ground or buried a few inches below.

Another approach to simulating a perfect ground system is to utilize the ground-plane antenna, with its four ground-plane radials elevated well above lossy earth. Heights (between the bottom of the ground-plane and the surface of the ground) greater than λ/8 have proven to yield excellent results. See Chapter 6, Low-Frequency Antennas, for more details on practical ground-plane verticals.

For a vertical antenna, a large ground screen, either made of wire mesh or a multitude of radials, or an elevated system of ground-plane radials will reduce ground losses near the antenna. This is because the screen conductors are solidly bonded to each other and the resistance is much lower than that of the lossy, low-conductivity earth itself. If the ground screen or elevated ground plane were not present, RF currents would be forced to flow through the lossy, low-conductivity earth to return to the base of the radiator. The ground screen or elevated ground plane in effect shield ground-return currents from the lossy earth.

Less-Than-Ideal Ground Systems

Now, what happens when something less than an ideal ground screen is used as the ground plane for a vertical monopole? Typically this will take the form of an on-ground wire radial system. A great deal of mystery and lack of information seems to surround the vertical antenna ground system. In the case of ground-mounted vertical antennas, many general statements such as "the more radials the better" and "lots of short radials are better than a few long ones" have served as rules of thumb, but many questions as to relative performance differences and optimum number for a given length remain unanswered, as is the justification for the rules of thumb. Most of these questions boil down to one: namely, how many radials, and how long, should be used in a given vertical antenna installation?

A ground system with 120 $\lambda/2$ radials is not very practical for many amateur installations, which often must contend with limited space and funding. Unfortunately the ground resistance, R_g, increases rapidly when the number of radials is reduced. To minimize ground loss where a large, optimum ground system is not possible requires that we understand how ground losses occur and how to optimize the design of a ground system that can fit within the space and budget available.

E and H Fields

E and H fields were introduced in Chapter 2, Antenna Fundamentals, to explain some basic concepts concerning antennas. To understand the reasons for ground loss we need to look at the E and H fields in the near field, but we need to have some feeling for what E and H fields are. The following is a brief description of these fields. It is certainly not a rigorous description but should give at least an intuitive feeling for what is happening.

In 1820 Hans Oerstad discovered that a current flowing in a wire would deflect the needle of a nearby compass. We attribute this effect to a *magnetic* or *H-field*, which at any given location is denoted by the bold-faced letter **H**. **H** is a vector, with an amplitude expressed in A/m (Amperes/meter) and a direction. **Fig 2** shows a typical experimental arrangement. The shape of the magnetic field is roughly shown by the distribution of the iron filings. This field distribution is very similar to that for a vertical antenna.

A compass needle (a small magnet itself) will try to align itself parallel to **H.** As the compass is moved around the conductor, the orientation of the needle changes accordingly. The orientation of the needle gives the direction of **H**. If you attempt to turn the needle away from alignment you will discover a torque trying to restore the needle to its original position. The torque is proportional to the strength of the magnetic field at that point. This is called the *field intensity* or amplitude of **H** at that point. If a larger current flows in the conductor going through the piece of paper holding the iron filings, the amplitude of **H** will be larger. Currents flowing in the conductors of an antenna also generate a magnetic field, one component of the near field.

An antenna will also have an *electric* or *E-field*, which can be visualized using a parallel-plate capacitor, as shown in **Fig 3**. If we connect a battery with a potential V_{dc} across the capacitor plates there will be an electric field **E** established between the plates, as indicated by the lines and directional arrows between the plates. The magnitude of vector **E** is expressed in V/m (volts per meter), so for a potential of V volts and a spacing of d meters, E = V/d V/m. The amplitude of **E** will increase with greater voltage and/or a smaller distance (d). In an antenna, there will be ac potential differences between different parts of the antenna and from the antenna to ground. These ac potential differences establish the electric field associated with the antenna.

Conduction And Displacement Currents

If we replace the dc voltage source in Fig 3 with an ac source, a steady ac current will flow in the circuit. In the

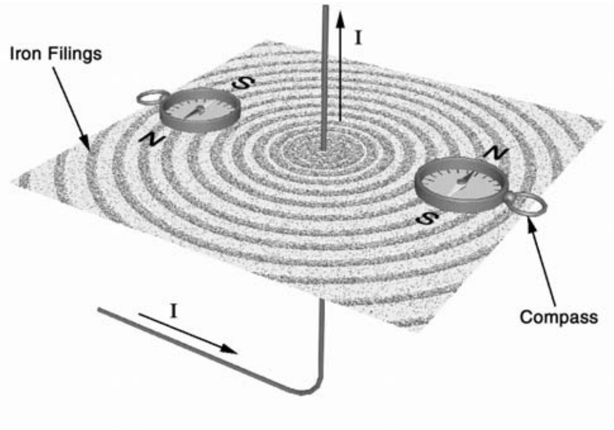

Fig 2—The magnetic lines of force that surround a conductor with an electric current flowing in it are shown by iron filings and small compass needles. The needles point in the direction of the magnetic or H-field. The filings give a general view of the field distribution in the plane perpendicular to the conductor.

conductors between the ac source and the capacitor plates, current (I_c) flows, because of the movement of charge, usually electrons. But in the space between the capacitor plates—particularly in a vacuum—there are no charge carriers available to carry a conduction current. Nonetheless, current still flows in the complete circuit, and we attribute this to a *displacement current* (I_d) flowing between the capacitor plates to account for the continuity of current in the circuit. Displacement and conduction currents are two different phenomena but they both represent current, just two different kinds. Some observers prefer to call conduction currents "currents" and displacement currents "imaginary currents." That terminology is OK, but to account for the current flow in a closed circuit with capacitance you have to keep track of both kinds of current, whatever you call them.

In an antenna over ground, the displacement current represents the current flow from the antenna surface through the air into the ground. The currents flowing in the ground are predominantly conduction currents, but there may also be displacement currents.

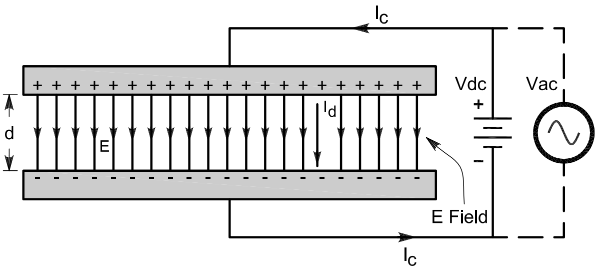

Fig 3—Example of an electric field, $E=V_{dc}/d$. When the dc source is replaced with an ac source there will be a displacement current (I_d) flowing between the capacitor plates.

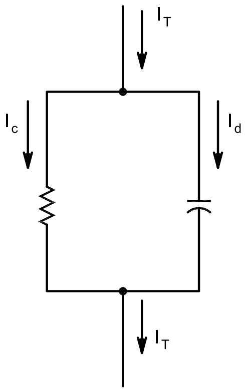

Fig 4—When the capacitor dielectric is less than perfect there will be a conduction current (I_c) in addition to the displacement current (I_d). Soil will typically have both resistive and capacitive components. Power loss in the soil is due to the current flowing through the resistive component.

Where the dielectric material between the capacitor plates is not a perfect insulator, both conduction and displacement currents can flow between the capacitor plates. A good example of this would be a soil dielectric, which has both resistive and capacitive characteristics. Soil can be represented in the circuit of **Fig 4**, where there is a resistor with a conduction current I_c in parallel with a capacitor with a displacement current I_d. The two currents add up (vectorially) as the total current I_T.

A Closer Look at Verticals

A vertical antenna has two field components that induce currents in the ground around the antenna. **Fig 5** shows in a general way the electric-field component (E_z, in V/m) and magnetic-field component (H_ϕ, in A/m) in the region near a vertical. Because the soil near the antenna usually has relatively high resistance, both of these field components will induce currents (I_V and I_H) in the ground, resulting in losses. While the worms may enjoy the heated ground, power dissipated in the ground is subtracted from the radiated power, weakening your signal.

As shown in Fig 5, the tangential component of the H-field (H_ϕ) induces horizontal currents (I_H) flowing radially. The normal component (perpendicular to the ground surface) of the E-field (E_z) induces vertically flowing currents (I_v). Actually, things are more complex than this but we don't need to thrash through that to understand conceptually what's going on.

When modeling an antenna we account for the radiated power (P_r) by assuming there is a resistor we call the *radiation resistance* (R_r) through which the antenna base current (I_o) flows. The radiated power is then $P_r = I_o^2R_r$. Similarly, we can account for the power dissipated in the ground (P_g) by adding a loss resistance (R_g) in series with R_r. The ground loss is then $P_g = I_o^2R_g$. Additional losses due to conductors, loading coils, etc can also be simulated

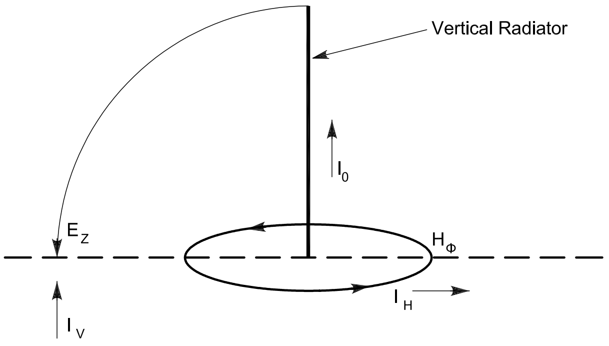

Fig 5—A general view of the fields and ground currents near the base of a vertical antenna. Note that the H-field distribution is equivalent to that shown in Fig 3.

by adding more series loss resistances. Putting aside for the moment these additional losses, the efficiency (η) of a vertical can be expressed as:

$$\eta = \frac{P_r}{P_r + P_g} \qquad \text{(Eq 1)}$$

This can be restated in terms of resistances as:

$$\eta = \frac{R_r}{R_r + R_g} = \frac{1}{1 + \frac{R_g}{R_r}} \qquad \text{(Eq 2)}$$

In essence, efficiency is the ratio of the radiated power to the total input power ($P_T = P_r + P_g$). Another way of saying this is that efficiency depends on the ratio of ground loss resistance (R_g) to radiation resistance (R_r), as Eq 2 shows. The smaller we make R_g the more power will be radiated for a given input power. Reducing R_g is the purpose of the ground system.

A sketch of current flow in the antenna and the surrounding ground (due to H-field), near the base of a vertical, is shown in **Fig 6**. I_z represents the total *zone current* flowing radially through a cylindrical zone at a given radius (r) due to the H field, while I_o is the current at the feed point at the base of the antenna. Technically speaking, the cylinder is infinitely deep, with I_z being the total current integrated over the surface of the cylinder at a given radius.

Fig 7 is a graph of the amplitude of I_z for several antenna heights in wavelengths (h) as we move away from the base of the antenna. Fig 7 shows the zone current that would flow in the ground returning to the base of the antenna, assuming a single ground rod is placed at the feed point for the vertical radiator. The heights indicated are the effective electrical heights. For example, if you use some top loading on the vertical, the effective electrical height will be greater than the physical height.

It is important to recognize that simply adding a top hat to a vertical of a given physical height may reduce ground losses. We can see this from the effect of h on ground current amplitude in Fig 7. Increasing h reduces the ground current. Even something as simple as moving a loading coil from the base up to the center of the antenna may reduce ground losses because it reduces ground current amplitude. But we do have to be careful that the loss introduced by the loading coil does not overcome the reduction in ground loss! Both loading-coil and top-hat schemes also increase the radiation resistance R_r, which further improves efficiency.

The currents in Fig 7 have been adjusted for constant radiated power at the base of the antenna by varying I_o to compensate for the change in R_r as we vary h. To maintain constant radiated power as R_r is falling, you must increase I_o. The base feed-point impedance is a strong function of h. For example, for h = 0.25 λ, R_r will be in the neighborhood of 36 Ω. However, for h = 0.1 λ, R_r will be less than 4 Ω. More information on short antennas can be found in Chapter 16, Mobile and Maritime Antennas.

Fig 7 clearly shows the high currents that flow in the ground near the base of a short antenna due to the antenna's H field. Compared to a 0.25-λ vertical, the 0.1-λ vertical has about three times the base current. As you shorten the antenna further, the zone current increases even more quickly. The ground loss is proportional to the square of the ground current ($P_g = I_g^2 R_g$), so the power loss in the immediate

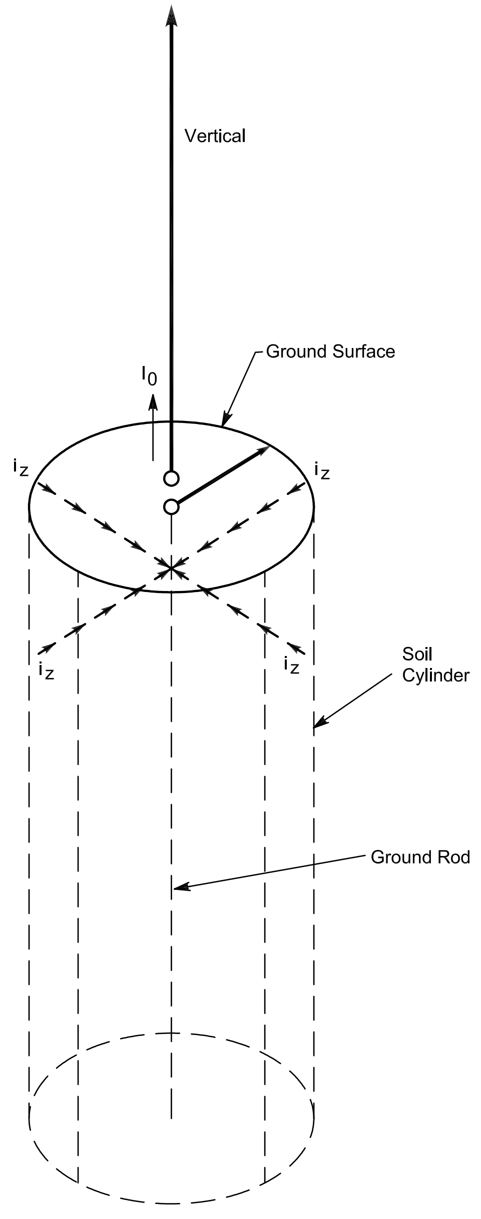

Fig 6—Representation of the zone current near the base of a vertical antenna. Individual i_z components of current flow into a cylinder of soil, with a radius r centered on the base of the vertical. The total current, I_z, thus represents the net current induced in the soil by the H-field for a given radius. (I_z was labeled I_H in Fig 5.)

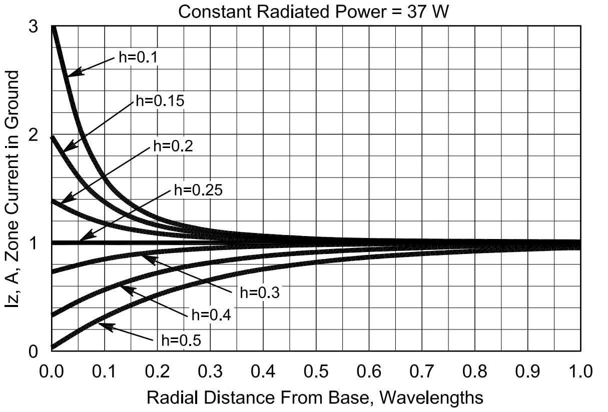

Fig 7—Plot of zone current (I_z) in amperes near the base of a vertical as a function of height (h) and radius (r) in wavelengths. The current in the base of the 0.25-λ antenna is assumed to be 1 A and the current for other values of h is adjusted to maintain the same radiated power ($P_r = I_o^2 R_r = 37$ W) as the radiation resistance (R_r) changes with h.

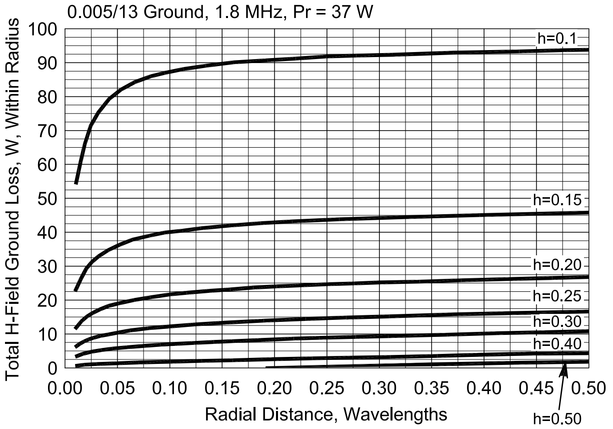

Fig 8—Total H-field induced ground loss within a circle of radius r around the base of a vertical for different values of h and constant P_r = 37 W. Note how the total loss increases rapidly near the base of the antenna indicating high loss. Beyond r = 0.15 λ, however, the additional loss is much lower and the curves flatten out. Note also how much higher the loss is for shorter antennas.

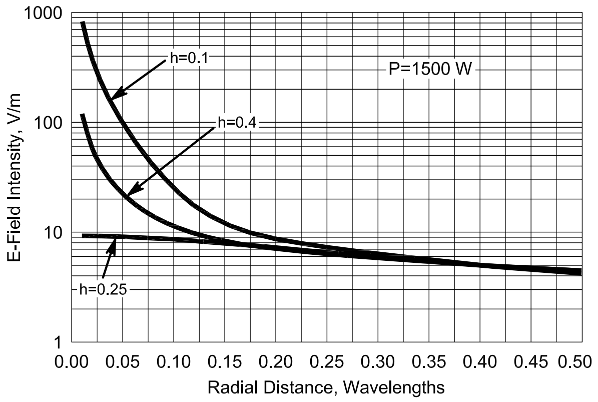

Fig 9—Electric field intensity near the base of a vertical for different values of h. P_r is held constant at 1500 W.

region of the base is much higher for a short antenna operating with the same input power as for a quarter-wave vertical.

We can calculate the losses induced in the soil by either the E- or H-field intensity. **Fig 8** shows an example of the H-field losses for several different antenna heights, given a constant radiated power of 37 W. Note that the total loss within 0.5 λ of the base for h = 0.25 λ is about 16 W (right side of the graph). This gives $\eta = 37/(37 + 16) = 70\%$. However, for h = 0.1 λ, the total loss is about 94 W. Taking into consideration only the H-field losses, $\eta = 37/(37 + 94) = 28\%$. Note that in both cases the majority of the loss is near (< 0.1 λ) the antenna, with the rate of increase of total loss decreasing rapidly as we move farther away from the base, where the lines are almost flat.

Fig 9 is a graph of the E-field intensity around a vertical with 1500 W radiated power, for three values of h. The E-field intensity doesn't depend on the exact type of ground system. (You can see this when you consider that the voltage across a capacitor doesn't depend on the size of the capacitor's plates.) Notice that close to the base, the E-field intensity for the 0.1-λ vertical is almost 100 times that for the 0.25-λ vertical. Because loss is proportional to the square of the voltage, the E-field losses close to the base will be ten thousand times larger in the 0.1-λ vertical! At a 1500-W power level the field intensity near the base of a short vertical is high enough to pose some risk of igniting grass and bushes that grow above any radial system close to the vertical's base. The grass should be kept mowed within 0.1 λ of the base.

Fig 10 shows a computation for the E-field losses, again for a constant radiated power of 37 W and several values of h. For the 0.25-λ vertical, the electric field intensity is quite low and so are the losses associated with it, at only 1.5 W. With any reasonable ground system, the E-field losses for a 0.25-λ vertical will be insignificant.

For shorter or longer verticals, however, the picture is different. This is why we see the very high losses (> 100 W) in Fig 10 for h = 0.1 λ. This loss, when added to the H-field loss, reduces the efficiency of the 0.1-λ vertical to 16% or less without a good ground system. In short antennas the E-field losses cannot be ignored, since they get worse exponentially as the antenna is shortened further.

The presence of a top-loading hat will also increase the E-field intensity in the area below the hat. However, most practical amateur hats will be quite small and the associated E-field loss small. The benefit, however, of reducing I_o because of the addition of the hat—which reduces the field

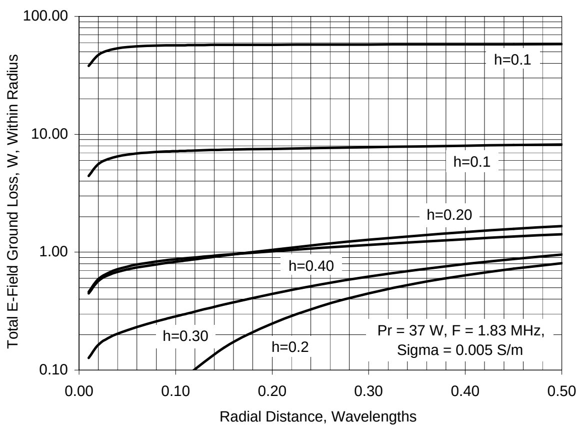

Fig 10—Total E-field induced ground loss within a circle of radius r around the base of a vertical for different values of h and constant P_r = 37 W. Note how the total loss increases rapidly near the base of the antenna, indicating high loss. Beyond r = 0.05 λ, however, the additional loss is much lower and the curves flatten out. Note again how much higher the loss is for shorter antennas. For h = 0.1 λ the E-field loss is greater than the H-field loss.

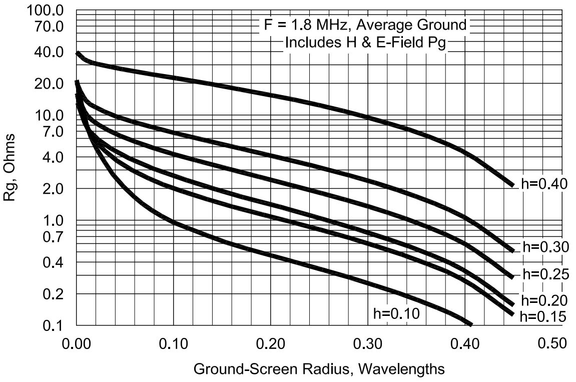

Fig 11—Effective ground resistance (R_g) at the base of the vertical as a function of the radius of a ground screen for several different antenna heights. Note how R_g falls as power dissipation in the soil is eliminated by the highly conducting ground screen.

from the vertical part of the antenna—more than compensates for the small additional E-field loss due to the hat.

Verticals taller than 0.25 λ also display increased E-field intensity, but not nearly so severe as short verticals. In verticals both shorter and longer than a 0.25 λ, the critical loss region is within a radius of about 0.05 λ. We can see this in Fig 10, where the power-loss curves for the shorter antennas flatten out by the time we reach a radius of 0.05 λ.

It's not widely known, but while radial wire systems reduce the H-field losses very effectively, Larsen (see Bibliography) has shown that the E-field losses with the same radial system do not fall in the same fashion as H-field losses. For h > 0.15 λ this doesn't matter much because the E-field loss is so low anyway. However, for short antennas it is very helpful to install either a ground screen or a dense radial system within 0.05 λ of the base.

We can take the data in Figs 8 and 10 and calculate the effective value of the ground resistance R_g. **Fig 11** shows the results of such a computation. Fig 11 assumes a perfect ground screen that varies in radius from 0.001 λ to 0.5 λ. As we would expect, when the ground screen is very small the ground losses are high, meaning R_g is high and the efficiency is low. As we increase the radius of the ground screen, forcing current out of the lossy soil and into the very low-loss screen, R_g drops rapidly and the efficiency increases.

Fig 11 demonstrates why it is desirable to have a good ground system out to at least 0.125 λ, and better yet, even farther. The shorter the antenna, the more important the ground system becomes, especially close to the base. In this example the ground system consists of a highly conductive, bonded ground screen, not always practical for amateur installations. A more typical ground system would consist of a number of individual radial wires. This kind of ground will be inferior to a screen but represents a practical compromise. We'll examine this in more detail shortly.

Note that as we reduce h in Fig 11, R_g actually goes down—even though the ground losses are higher. When h is made smaller the radiation resistance declines rapidly (see Chapter 16, Mobile and Maritime Antennas), so that for a given radiated power I_o must increase. If we measure R_g as we reduce h over a given ground system, we would see that the value for R_g goes down as shown in Fig 11. But because I_o^2 is rising more rapidly than R_g is falling, the power lost in the ground increases and efficiency decreases. The point here is that the value of R_g depends on the ground system, soil characteristics *and* the antenna configuration. You cannot assign an arbitrary value to R_g independent of the antenna system.

Wire Radial Systems

Fig 11 shows R_g for a dense, perfectly conducting ground screen, but what we really need to know is the effect of length and number of individual radials on R_g in a wire radial system. We can calculate the current division between a radial system and the soil and use this to determine R_g. A typical graph of the proportion of the zone current flowing in the radial system, as a function of radius and various numbers of radials (N), is shown in **Fig 12**.

The radial currents decrease as we move away from the base, and the lower the number of radials, the more rapidly the radial current decreases. This means that close to the base of the antenna most of the current is in the radial system, but as we move away from the base the current increasingly flows in the lossy ground. When only a few radials are used, the outer ends of the radials contribute little to reducing ground loss.

Why is this? The problem is that I_z does not go immediately to the nearest radial but may flow for some distance in the soil. This is illustrated in a general way in **Fig 13**. As we move away from the base of the antenna, adjacent radials are further apart from each other and the current must flow further in the soil before it reaches one of the radials. When we use more radials, the distance between radials is less and more of the total current will be in the radials and less in the soil. This reduces ground loss.

When we know the current distribution in the ground we can calculate the power loss and R_g. A typical example for h = 0.25 λ is given in **Fig 14**. We can learn a lot about radial system design from Fig 14 and similar graphs. If we use only a few radials, the radial current drops off very rapidly. Most of the current is flowing in the soil.

Such a ground system is by nature inefficient—that is, R_g is large. We can also see that if we have only 16 radials, R_g falls an ohm or two as we lengthen the radials, but is essentially flat by 0.1 λ. There is no point in making them longer because there is little current in the outer portions and R_g is essentially constant beyond 0.1 λ. As we increase the number of radials, we gather more current further out, making longer radials more useful. The result is cumulative—more radials allow longer radials to be effective and both together reduce ground loss. We can see this in Fig 14, where the initial value of R_g drops as N increases and flattens out at longer radial lengths. For 128 radials, for example, lengths of 0.25 λ or more are useful.

The example in Fig 14 uses #12 wire for the radials. Compared to soil, the resistance of the radial wires is very small, especially if many radials are used, and does not greatly affect overall losses no matter how small the wire. The effect of changing wire size is to slowly change the current division between ground and the radial system. Larger wire results in only a small decrease in R_g. In principle, very small wire could be used for radials but from a mechanical point of view, #18 or #20 wire is about as small as is practical. Any smaller wire breaks too easily to be buried, and also breaks easily when left on the ground surface and is walked on or driven over.

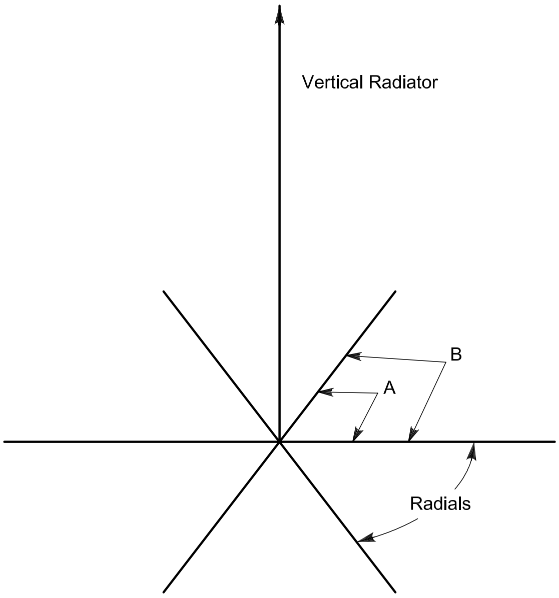

Fig 13—An example of current entering the ground between the radials and flowing for some distance before being picked up by a radial.

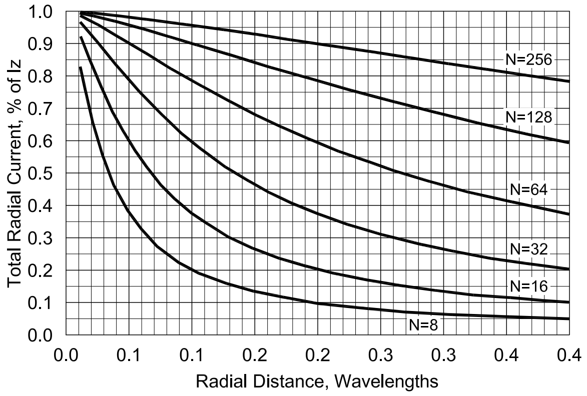

Fig 12—An example of the portion of the zone current flowing in the radial system as you move away from the base of a 0.25 λ vertical for different numbers of radials (N). Note that when more radials are used, more of the zone current flows in the radials and not in the ground, reducing ground loss. The proportion of current in the radial system falls rapidly when only a few radials are used. This leads to high ground loss because most of the zone current is flowing in the ground rather than in the radials.

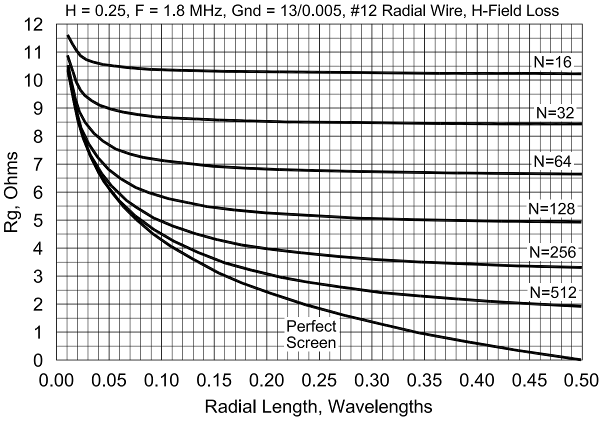

Fig 14—An example of the variation of R_g with radial length and number of radials (N) for h = 0.25 λ. When only a few radials are used there is little point in making them longer than 0.1 λ. Increasing N reduces R_g at a given radius and also makes longer radials useful, further reducing R_g. R_g for other values of h behave similarly.

On the other hand, wire larger than #12 is expensive. Thousands of feet of #8 wire may be affordable for broadcast stations but not for most hams. Increasing the wire size from #20 to #10 would result in only a small reduction in R_g. Of course, if you happen to have a few thousand feet of old RG-8 cable lying around (the diameter is comparable to #0000 wire) then that might indeed help to reduce R_g, as W9QQ has shown (see Bibliography). You are still better off, however, using many radials with small wire than a few radials with large wire. Radial wire size is usually a mechanical and financial issue, not an electrical one.

If the example in Fig 14 were changed for different ground characteristics, then the curves would have a similar shape but would be shifted either up or down. For example, poorer ground will result in higher R_g but the usable length for the radials for a given N would increase somewhat. For better-quality ground, with higher conductivity, R_g will be lower but the usable length of the radials for a given N will be shorter.

For short antennas, the initial drop in R_g will be more rapid and the curves flatten out sooner. This implies that somewhat shorter radials are useful with short antennas. However, given the high losses, it is still a very good idea to use lots of radials with short antennas. As in the above example, increasing N also increases the usable length.

As you go up in frequency from 160 meters, R_g generally rises slowly and then stabilizes around 7 MHz, depending on the ground characteristics. This effect is related to the change in skin depth with frequency, which is discussed in a later section of this chapter. There is also a small shift in current division between the radials and ground as the frequency increases.

A Word Of Caution

In the preceding discussion we presented a number of graphs and the CD-ROM accompanying this book contains some spreadsheets containing the equations from which these graphs were derived. From these graphs we extracted a number of observations on how to design radial systems. Basic to each graph is the assumption that we know the ground characteristics: conductivity and permitivity. In the real world, we amateurs very rarely have more than a rough idea of the ground characteristics under our antennas. Even when careful measurements are made, the characteristics will vary through the year with rainfall or the lack thereof.

Soils are always stratified vertically and can vary by factors or two or more horizontally over distances comparable to radial length, so that even good ground measurements are at best an average. In addition, there will frequently be constraints on the size and shape of the ground system. As a result, we use the calculated information and the previous graphs for general guidance and preliminary design, but when actually installing a ground system we try to measure—or at least estimate—R_g as we go along.

When R_g stops falling, or our patience and/or money run out, we stop adding ground radials. We can measure the feed-point resistance with an impedance bridge to estimate of R_g. The impedance seen at the feed point of the antenna is the sum of the loss and the radiation resistance. To determine R_g you have to estimate R_r (from the antenna height) and other losses due to loading or conductors, and then subtract that from the total measured input resistance. The remainder is R_g, and R_g should fall as we add radials. When R_g stops falling we probably have as many radials of a given length as will be useful. Further reduction in R_g would require more, longer radials.

Practical Suggestions For Vertical Ground Systems

At least 16 radials should be used if at all possible. Experimental measurements and calculations show that with this number, the loss resistance decreases the antenna efficiency by 30% to 50% for a 0.25 λ vertical, depending on soil characteristics. In general, a large number of radials (even though some or all of them must be short) is preferable to a few long radials for a vertical antenna mounted on the ground. The conductor size is relatively unimportant as mentioned before: #12 to #22 copper wire is suitable.

Table 1 summarizes these conclusions. John Stanley, K4ERO, first presented this material in December 1976 *QST*. Another source of information on ground-system design is *Radio Broadcast Ground Systems* (see the Bibliography at the end of this chapter). Most of the data presented in Table 1 is taken from that source, or derived from the interpolation of data contained therein.

Table 1 is based on the number of radials. For each configuration, there is a corresponding optimum radial length. Each configuration also includes the amount of wire used, expressed in wavelengths. Using radials considerably longer than suggested for a given N or using a lot more radials than suggested for a given length, while not adverse to performance, does not yield significant improvement either. That would represent a non-optimum use of wire and construction time. Each suggested configuration represents an optimum relationship between length and number for the given amount of total wire used. Table 1 leads to these conclusions:

- If you install only 16 radials (in configuration A), they need not be very long—0.1 λ is sufficient. The total length of wire will be 1.6 λ, which is about 875 feet at 1.8 MHz.
- If you have the wire, the space and the patience to lay down 120 radials (optimal configuration F), they should be 0.4 λ long. This radial system will gain about 3 dB over the 16-radial case and you'll use 48 λ of wire, or about 26,000 feet at 1.8 MHz.
- If you install 36 radials that are 0.15 λ long, you will lose 1.5 dB compared to optimal configuration F. You will use 5.4 λ of wire, or almost 3,000 feet at 1.8 MHz.

The loss figures in Table 1 assume h = 0.25 λ. A very rough approximation of loss when using shorter antennas can be obtained by doubling the loss in dB each

Table 1
Optimum Ground-System Configurations

	Configuration Designation					
	A	*B*	*C*	*D*	*E*	*F*
Number of radials	16	24	36	60	90	120
Length of each radial in wavelengths	0.1	0.125	0.15	0.2	0.25	0.4
Spacing of radials in degrees	22.5	15	10	6	4	3
Total length of radial wire installed, in wavelengths	1.6	3	5.4	12	22.5	48
Power loss in dB at low angles with a quarter-wave radiating element	3	2	1.5	1	0.5	0*
Feed-point impedance in ohms with a quarter-wave radiating element	52	46	43	40	37	35

Note: Configuration designations are indicated only for text reference.
*Reference: The loss of this configuration is negligible compared to a perfectly conducting ground.

time the antenna height is halved. For taller antennas the losses decrease, approaching 2 dB for configuration A of Table 1 for a half-wave radiator. Even longer antennas yield correspondingly better performance.

Table 1 is based on average ground conductivity. Variation of the loss values shown can be considerable, especially for configurations using fewer radials. Those building antennas over dry, sandy or rocky ground should expect more loss. On the other hand, higher than average soil conductivity and wet soils would make the compromise configurations (those with the fewest radials) even more attractive.

When antennas are combined into arrays, either parasitic or all-driven types, mutual impedances lower the radiation resistance of the elements. This drastically increases the effects of ground loss because I_0 will be higher for the same power level. For instance, an antenna with a 50-Ω feed-point impedance, of which 10 Ω is ground-loss resistance, will have an efficiency of approximately 83%. An array of two similar antennas in a driven array with similar ground losses may have an efficiency of 70% or less.

Special precautions must be taken in such cases to achieve satisfactory operation. Generally speaking, a wide-spaced broadside array presents little problem because R_r is high, but a close-spaced end-fire array should be avoided because R_r is much lower, unless low-loss radial system configurations are used or other precautions taken. Chapter 8, Multielement Arrays, covers the subject of vertical arrays in great detail.

In cases where directivity is desirable or real-estate limitations dictate, longer, more closely spaced radials can be installed in one direction, and shorter, more widely spaced radials in another. Multiband ground systems can be designed using different optimum configurations for different bands. Usually it is most convenient to start at the lowest frequency with fewer radials and add more short radials for better performance on the higher bands.

There is nothing sacred about the exact details of the configurations in Table 1, and small changes in the number of radials and lengths will not cause serious problems. Thus, a configuration with 32 or 40 radials of 0.14 λ or 0.16 λ will work as well as configuration C shown in the table.

If less than 90 radials are contemplated, there is no need to make them a quarter wavelength long. This differs rather dramatically from the case of a ground-plane antenna, where resonant radials are installed above ground. For a ground-mounted antenna, quarter-wave long radials may not be optimum. Because the radials of a ground-mounted vertical are actually on, if not slightly below the surface, they are coupled by capacitance or conduction to the ground, and thus resonance effects are not important. The basic function of radials is to provide a low-loss return path for ground currents.

Radio Broadcast Ground Systems states, "Experiments show that the ground system consisting of only 15 radial wires need not be more than 0.1 wavelength long, while the system consisting of 113 radials is still effective out to 0.5 wavelength." Many graphs in that publication confirm this statement. This is not to say that these two systems will perform equally well; they most certainly will not. However, if 0.1 λ is as long as the radials can be, there is little point in using more than 15 of them unless the vertical radiator's height is also small.

The antenna designer should:

1. Study the cost of various radial configurations versus the gain of each.
2. Compare alternative means of improving transmitted signal and their cost (more power, etc).
3. Consider increasing the physical antenna height (the electrical length) of the vertical radiator, instead of improving the ground system.
4. Use multi-element arrays for directivity and gain, observing the necessary precautions related to mutual impedances discussed in Chapter 8, Multielement Arrays.

The Effect of Ground in the Far Field

The properties of the ground in the far field of an antenna are very important, especially for a vertically polarized antenna, as discussed above. Even if the ground-radial system for a vertical has been optimized to reduce ground-return losses in the reactive near field to an insignificant level, the electrical properties of the ground may still diminish far-field performance to lower levels than "perfect-ground" analyses might lead you to expect. The key is that *ground reflections* from horizontally and vertically polarized waves behave very differently.

Reflections in General

First, let us consider the case of flat ground. Over flat ground, either horizontally or vertically polarized downgoing waves launched from an antenna into the far field strike the surface and are reflected by a process very similar to that by which light waves are reflected from a mirror. As is the case with light waves, the angle of reflection is the same as the angle of incidence, so a wave striking the surface at an angle of, say, 15° is reflected upward from the surface at 15°.

The reflected waves combine with direct waves (those radiated at angles above the horizon) in various ways. Some of the factors that influence this combining process are the height of the antenna, its length, the electrical characteristics of the ground, and as mentioned above, the polarization of the wave. At some elevation angles above the horizon the direct and reflected waves are exactly in phase—that is, the maximum field strengths of both waves are reached at the same time at the same point in space, and the directions of the fields are the same. In such a case, the resultant field strength for that angle is simply the sum of the direct and reflected fields. (This represents a theoretical increase in field strength of 6 dB over the free-space pattern at these angles.)

At other elevation angles the two waves are completely out of phase—that is, the field intensities are equal at the same instant and the directions are opposite. At such angles, the fields cancel each other. At still other angles, the resultant field will have intermediate values. Thus, the effect of the ground is to increase radiation intensity at some elevation angles and to decrease it at others. When you plot the results as an elevation pattern, you will see *lobes* and *nulls*, as described in Chapter 2, Antenna Fundamentals.

The concept of an *image antenna* is often useful to show the effect of reflection. As **Fig 15** shows, the reflected ray has the same path length (AD equals BD) that it would if it originated at a virtual second antenna with the same characteristics as the real antenna, but situated below the ground just as far as the actual antenna is above it.

Now, if we look at the antenna and its image over perfect ground from a remote point on the surface of the ground, we will see that the currents in a horizontally polarized antenna and its image are flowing in opposite directions, or in other words, are 180° out of phase. But the currents in a vertically polarized antenna and its image are flowing in the *same* direction—they are *in* phase. This 180° phase difference between the vertically and horizontally polarized reflections off ground is what makes the combinations with direct waves behave so very differently.

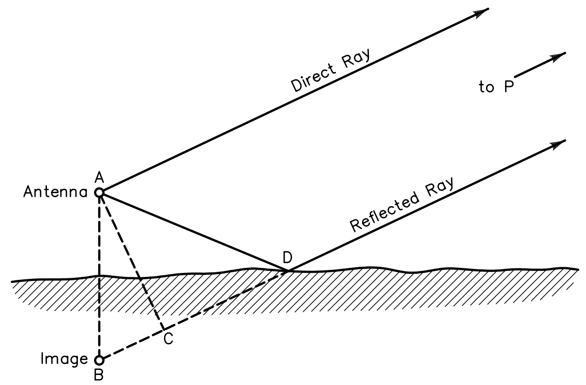

Fig 15—At any distant point, P, the field strength will be the vector sum of the direct ray and the reflected ray. The reflected ray travels farther than the direct ray by the distance BC, where the reflected ray is considered to originate at the *image* antenna.

FAR-FIELD GROUND REFLECTIONS AND THE VERTICAL ANTENNA

A vertical's azimuthal directivity is omnidirectional. A λ/2 vertical over ideal, perfectly conducting earth has the elevation-plane radiation pattern shown by the solid line in **Fig 16**. Over real earth, however, the pattern looks more like the shaded one in the same diagram. In this case, the low-angle radiation that might be hoped for because of perfect-ground performance is not realized in the real world.

Now look at **Fig 17A**, which compares the computed elevation-angle response for two half-wave dipoles at 14 MHz. One is oriented horizontally over ground at a height of λ/2 and the other is oriented vertically, with its center just over λ/2 high (so that the bottom end of the wire doesn't actually touch the ground). The ground is "average" in dielectric constant (13) and conductivity (0.005 S/m). At a 15° elevation angle, the horizontally polarized dipole has almost 7 dB more gain than its vertical brother. Contrast Fig 17A to the comparison in Fig 17B, where the peak gain of a vertically polarized half-wave dipole over seawater, which is virtually perfect for RF reflections, is quite comparable with the horizontal dipole's response at 15°, and exceeds the horizontally polarized antenna dramatically below 15° elevation.

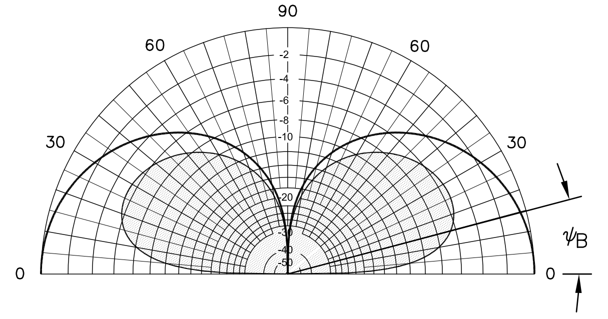

Fig 16—Vertical-plane radiation pattern for a ground-mounted quarter-wave vertical. The solid line is the pattern for perfect earth. The shaded pattern shows how the response is modified over average earth (k = 13, G = 0.005 S/m) at 14 MHz. ψ is the pseudo-Brewster angle (PBA), in this case 14.8°.

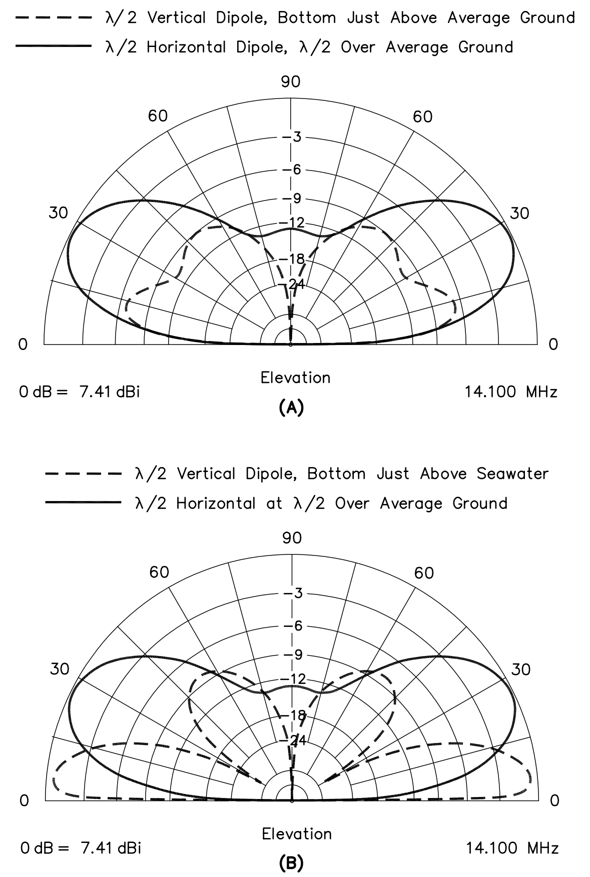

Fig 17—At A, comparison of horizontal and vertical λ/2 dipoles over average ground. Average ground has conductivity of 5 mS/m and dielectric constant of 13. Horizontal dipole is λ/2 high; vertical dipole's bottom wire is just above ground. Horizontal antenna is much less affected by far-field ground losses compared with its vertical counterpart. At B, comparison of 20-meter λ/2 vertical dipole whose bottom wire is just above seawater with λ/2-high horizontal dipole over average ground. Seawater is great for verticals!

To understand in a qualitative fashion why the desired low-angle radiation from a vertical is not delivered when the ground isn't "perfect," examine **Fig 18A**. Radiation from each antenna segment reaches a point P in space by two paths; one directly from the antenna, path AP, and the other by reflection from the earth, path AGP. (Note that P is so far away that the slight difference in angles is insignificant—for practical purposes the waves are parallel to each other at point P.)

If the earth were a perfectly conducting surface, there would be no phase shift of the vertically polarized wave upon reflection at point G. The two waves would add together with some phase difference because of the different path lengths. This difference in path lengths of the two waves is why the free-space radiation pattern differs from the pattern of the same antenna over ground.

Now consider a point P that is close to the horizon, as in Fig 18B. The path lengths AP and AGP are almost the same, so the magnitudes of the two waves add together, producing a maximum at zero angle of radiation. The arrows on the waves point both ways since the process works similarly for transmitting and receiving.

With real earth, however, the reflected wave from a vertically polarized antenna undergoes a change in both *amplitude* and *phase* in the reflection process. Indeed, at a low-enough elevation angle, the phase of the reflected wave will actually change by 180° and its magnitude will then subtract from that of the direct wave. At a zero takeoff angle, it will be almost equal in amplitude, but 180° out of phase with the direct wave.

Note that this is very similar to what happens with horizontally polarized reflected and direct waves at low elevation angles. Virtually complete cancellation will result in a deep null, inhibiting any radiation or reception at 0°. For real-world soils, the vertical loses the theoretical advantage it has at low elevation angles over a horizontal antenna, as Fig 17A so clearly shows.

The degree that a vertical works better than a horizontal antenna at low elevation angles is largely dependent on the characteristics of the ground around the vertical, as we'll next examine.

THE PSEUDO-BREWSTER ANGLE AND THE VERTICAL ANTENNA

Much of the material presented here regarding pseudo-Brewster angle was prepared by Charles J. Michaels, W7XC, and first appeared in July 1987 *QST*, with additional information in *The ARRL Antenna Compendium, Vol 3*. (See the Bibliography at the end of this chapter.)

Most fishermen have noticed that when the sun is low, its light is reflected from the water's surface as glare, obscuring the underwater view. When the sun is high, however, the sunlight penetrates the water and it is possible to see objects below the surface of the water. The angle at which this transition takes place is known as the *Brewster angle*,

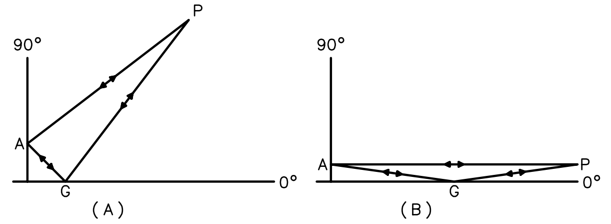

Fig 18—The direct wave and the reflected wave combine at point P to form the pattern (P is very far from the antenna). At A the two paths AP and AGP differ appreciably in length, while at B these two path lengths are nearly equal.

named for the Scottish physicist, Sir David Brewster (1781-1868).

A similar situation exists in the case of vertically polarized antennas; the RF energy behaves as the sunlight in the optical system, and the earth under the antenna acts as the water. The *pseudo-Brewster angle* (PBA) is the angle at which the reflected wave is 90° out of phase with respect to the direct wave. "Pseudo" is used here because the RF effect is similar to the optical effect from which the term gets its name. Below this angle, the reflected wave is between 90° and 180° out of phase with the direct wave, so some degree of cancellation takes place. The largest amount of cancellation occurs near 0°, and steadily less cancellation occurs as the PBA is approached from below.

The factors that determine the PBA for a particular location *are not related to the antenna itself, but to the ground around it*. The first of these factors is earth conductivity, G, which is a measure of the ability of the soil to conduct electricity. Conductivity is the inverse of resistance. The second factor is the dielectric constant, k, which is a unitless quantity that corresponds to the capacitive effect of the earth. For both of these quantities, the higher the number, the better is the ground (for vertical antenna purposes). The third factor determining the PBA for a given location is the frequency of operation. The PBA increases with increasing frequency, all other conditions being equal. **Table 2** gives typical values of conductivity and dielectric constant for different types of soil. The map of **Fig 19** shows the approximate conductivity values for different areas in the continental United States.

As the frequency is increased, the role of the dielectric constant in determining the PBA becomes more significant. **Table 3** shows how the PBA varies with changes in ground conductivity, dielectric constant and frequency. The table shows trends in PBA dependency on ground constants and frequency. The constants chosen are not necessarily typical of any geographical area; they are just examples.

At angles below the PBA, the reflected vertically polarized wave subtracts from the direct wave, causing the radiation intensity to fall off rapidly. Similarly, above the PBA, the reflected wave adds to the direct wave, and the radiated pattern approaches the perfect-earth pattern. Fig 16 shows the PBA, usually labeled ψ_B.

When plotting vertical-antenna radiation patterns over real earth, the reflected wave from an antenna segment is multiplied by a factor called the *vertical reflection coefficient,* and the product is then added vectorially to the direct wave to get the resultant. The reflection coefficient consists of an attenuation factor, A, and a phase angle, ϕ, and is usually expressed as $A\angle\phi$. (ϕ is always a negative angle, because the earth acts as a lossy capacitor in this situation.) The following equation can be used to calculate the reflection coefficient for vertically polarized waves, for earth of given conductivity and dielectric constant at any frequency and elevation angle (also called the wave angle in many texts).

Table 2
Conductivities and Dielectric Constants for Common Types of Earth

Surface Type	*Dielectric Constant*	*Conductivity (S/m)*	*Relative Quality*
Fresh water	80	0.001	
Salt water	81	5.0	
Pastoral, low hills, rich soil, typ Dallas, TX, to Lincoln, NE areas	20	0.0303	Very good
Pastoral, low hills, rich soil typ OH and IL	14	0.01	
Flat country, marshy, densely wooded, typ LA near Mississippi River	12	0.0075	
Pastoral, medium hills and forestation, typ MD, PA, NY, (exclusive of mountains and coastline)	13	0.006	
Pastoral, medium hills and forestation, heavy clay soil, typ central VA	13	0.005	Average
Rocky soil, steep hills, typ mountainous	12-14	0.002	Poor
Sandy, dry, flat, coastal	10	0.002	
Cities, industrial areas	5	0.001	Very Poor
Cities, heavy industrial areas, high buildings	3	0.001	Extremely poor

Table 3
Pseudo-Brewster Angle Variation with Frequency, Dielectric Constant, and Conductivity

Frequency (MHz)	*Dielectric Constant*	*Conductivity (S/m)*	*PBA (degrees)*
7	20	0.0303	6.4
	13	0.005	13.3
	13	0.002	15.0
	5	0.001	23.2
	3	0.001	27.8
14	20	0.0303	8.6
	13	0.005	14.8
	13	0.002	15.4
	5	0.001	23.8
	3	0.001	29.5
21	20	0.0303	10.0
	13	0.005	15.2
	13	0.002	15.4
	5	0.001	24.0
	3	0.001	29.8

$$A_{\text{Vert}} \angle \phi = \frac{k' \sin\psi - \sqrt{k' - \cos^2 \psi}}{k' \sin\psi + \sqrt{k' - \cos^2 \psi}} \qquad \text{(Eq 3)}$$

where

$A_{\text{Vert}} \angle \phi$ = vertical reflection coefficient

ψ = elevation angle

$$k' = k - j\left|\frac{1.8 \times 10^4 \times G}{f}\right|$$

k = dielectric constant of earth (k for air = 1)

G = conductivity of earth in S/m

f = frequency in MHz

j = complex operator ($\sqrt{-1}$)

Solving this equation for several points indicates what effect the earth has on vertically polarized signals at a particular location for a given frequency range. **Fig 20** shows the reflection coefficient as a function of elevation angle at 21 MHz over average earth (G = 0.005 S/m, and k = 13).

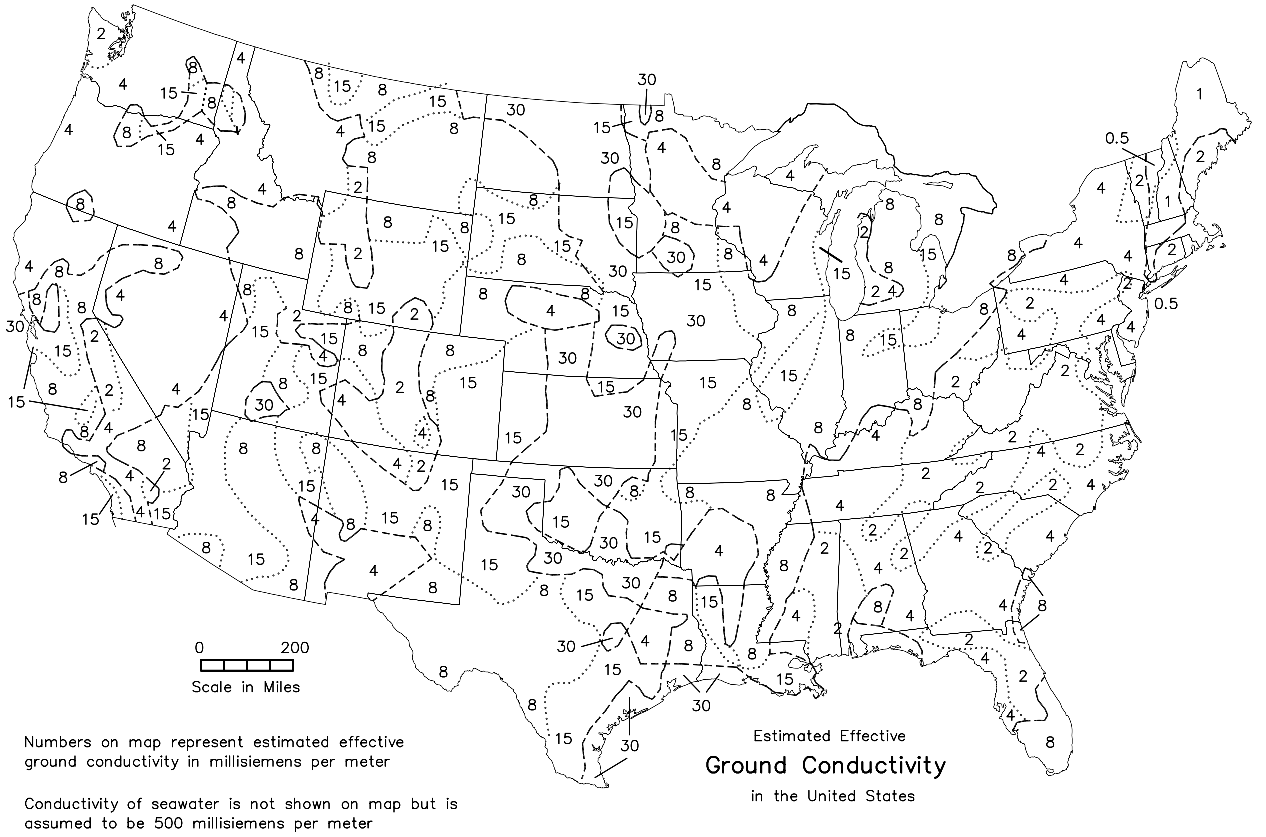

Numbers on map represent estimated effective ground conductivity in millisiemens per meter

Conductivity of seawater is not shown on map but is assumed to be 500 millisiemens per meter

Estimated Effective
Ground Conductivity
in the United States

Fig 19—Typical average soil conductivities for the continental United States. Numeric values indicate conductivities in millisiemens per meter (mS/m), where 1.0 mS/m = 0.001 S/m.

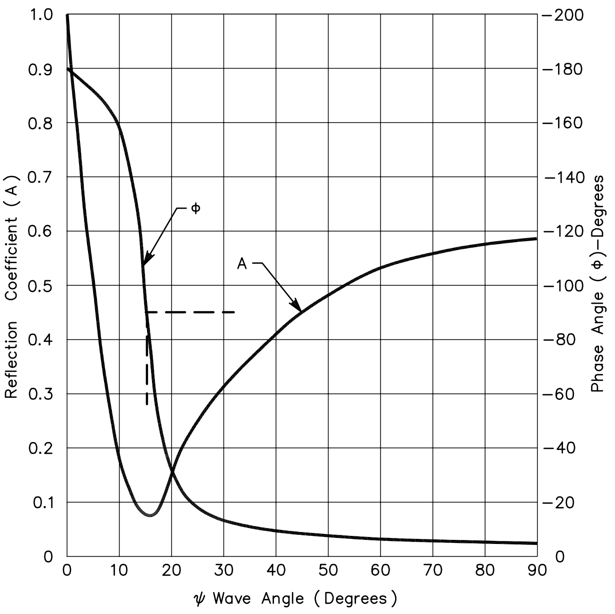

Fig 20—Reflection coefficient for vertically polarized waves. A and ϕ are magnitude and angle for wave angles ψ. This case is for average earth, (k = 13, G = 0.005 S/m), at 21 MHz.

Note that as the phase curve, ψ, passes through 90°, the attenuation curve (A) passes through a minimum at the same wave angle ψ. This is the PBA. At this angle, the reflected wave is not only at a phase angle of 90° with respect to the direct wave, but is so low in amplitude that it does not aid the direct wave by a significant amount. In the case illustrated in Fig 20 this elevation angle is about 15°.

Variations in PBA with Earth Quality

From Eq 3, it is quite a task to search for either the 90° phase point or the attenuation curve minimum for a wide variety of earth conditions. Instead, the PBA can be calculated directly from the following equation.

$$\psi_B = \arcsin \sqrt{\frac{k - 1 + \sqrt{(x^2 + k^2)^2 (k-1)^2 + x^2[(x^2+k^2)^2 - 1]}}{(x^2+k^2)^2 - 1}}$$

(Eq 4)

where k, G and f are as defined for Eq 3, and

$$x = \frac{1.8 \times 10^4 \times G}{f}$$

Fig 21 shows curves calculated using Eq 4 for several different earth conditions, at frequencies between 1.8 and 30 MHz. As expected, poorer earths yield higher PBAs. Unfortunately, at the higher frequencies (where low-angle radiation is most important for DX work), the PBAs are highest. The PBA is the same for both transmitting and receiving.

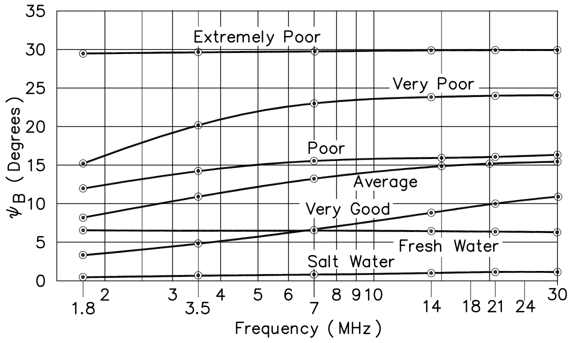

Fig 21—Pseudo-Brewster angle (ψ) for various qualities of earth over the 1.8 to 30-MHz frequency range. Note that the frequency scale is logarithmic. The constants used for each curve are given in Table 2.

Relating PBA to Location and Frequency

Table 2 lists the physical descriptions of various kinds of earth with their respective conductivities and dielectric constants, as mentioned earlier. Note that in general, the dielectric constants and conductivities are higher for better earths. This enables the labeling of the earth characteristics as extremely poor, very poor, poor, average, very good, and so on, without the complications that would result from treating the two parameters independently.

Fresh water and salt water are special cases; in spite of high resistivity, the fresh-water PBA is 6.4°, and is nearly independent of frequency below 30 MHz. Salt water, because of its extremely high conductivity, has a PBA that never exceeds 1° in this frequency range. The extremely low conductivity listed for cities (the last case) in Table 2 results more from the clutter of surrounding buildings and other obstructions than any actual earth characteristic. The PBA at any location can be found for a given frequency from the curves in Fig 21.

FLAT-GROUND REFLECTIONS AND HORIZONTALLY POLARIZED WAVES

The situation for horizontal antennas is different from that of verticals. **Fig 22** shows the reflection coefficient for horizontally polarized waves over average earth at 21 MHz. Note that in this case, the phase-angle departure from 0° never gets very large, and the attenuation factor that causes the most loss for high-angle signals approaches unity for low angles. Attenuation increases with progressively poorer earth types.

In calculating the broadside radiation pattern of a horizontal λ/2 dipole, the perfect-earth image current, equal to the true antenna current but 180° out of phase with it) is multiplied by the horizontal reflection coefficient given by Eq 5 below. The product is then added vectorially to the direct wave to get the resultant at that elevation angle. The reflection coefficient for horizontally polarized waves can be calculated using the following equation.

$$A_{Horiz} \angle \phi = \frac{\sqrt{k' - \cos^2 \psi} - \sin \psi}{\sqrt{k' - \cos^2 \psi} + \sin \psi} \qquad \text{(Eq 5)}$$

where

$A_{Horiz} \angle \phi$ = horizontal reflection coefficient

ψ = elevation angle

$$k' = k - j\left(\frac{1.8 \times 10^4 \times G}{f}\right)$$

k = dielectric constant of earth
G = conductivity of earth in S/m
f = frequency in MHz
j = complex operator ($\sqrt{-1}$)

For a horizontal antenna near the earth, the resultant pattern is a modification of the free-space pattern of the antenna. **Fig 23** shows how this modification takes place for a horizontal λ/2 antenna over a perfectly conducting flat surface. The patterns at the left show the relative radiation when one views the antenna from the side; those at the right show the radiation pattern looking at the end of the antenna. Changing the height above ground from λ/4 to λ/2 makes a significant difference in the high-angle radiation, moving the main lobe down lower.

Note that for an antenna height of λ/2 (Fig 23, bottom), the out-of-phase reflection from a perfectly conducting surface creates a null in the pattern at the zenith (90° elevation angle). Over real earth, however, a *filling in* of this null occurs because of ground losses that prevent perfect reflection of high-angle radiation.

At a 0° elevation angle, horizontally polarized antennas also demonstrate a null, because out-of-phase reflection cancels the direct wave. As the elevation angle departs from 0°, however, there is a slight filling-in effect so that over other-than-perfect earth, radiation at lower angles is enhanced compared to a vertical. A horizontal antenna will often outperform a vertical for low-angle DX work, particularly over lossy types of earth at the higher frequencies.

Reflection coefficients for vertically and horizontally polarized radiation differ considerably at most angles above ground, as can be seen by comparison of Figs 20 and 22. (Both sets of curves were plotted for the same ground constants and at the same frequency, so they may be compared directly.) This is because, as mentioned earlier, the image of a horizontally polarized antenna is out-of-phase with the antenna itself, and the image of a vertical antenna is in-phase with the actual radiator.

The result is that the phase shifts and reflection magnitudes vary greatly at different angles for horizontal and vertical polarization. The magnitude of the reflection coefficient for vertically polarized waves is greatest (near unity) at very low angles, and the phase angle is close to 180°. As mentioned earlier, this cancels nearly all radiation at very low angles. For the same range of angles, the magnitude of the reflection coefficient for horizontally polarized waves is also near unity, but the phase angle is near 0° for the specific conditions shown in Figs 20 and 22. This causes reinforcement of low-angle horizontally polarized waves. At some relatively high angle, the reflection coefficients for

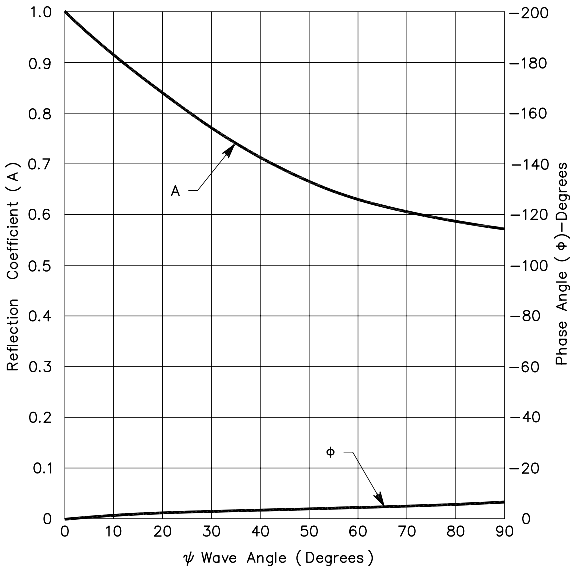

Fig 22—Reflection coefficient for horizontally polarized waves (magnitude A at angle ϕ), at 21 MHz over average earth (k = 13, G = 0.005 S/m).

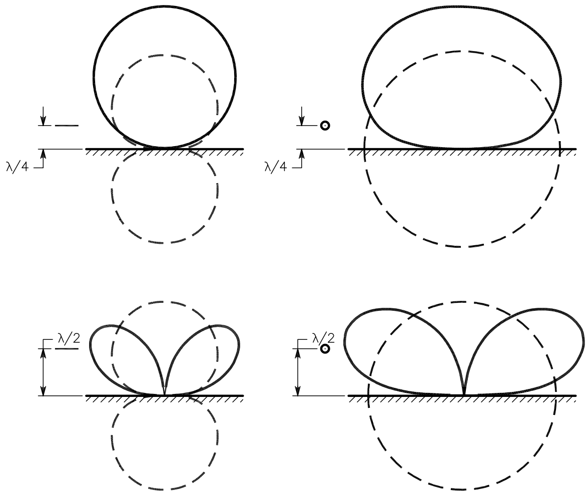

Fig 23—Effect of the ground on the radiation from a horizontal half-wave dipole antenna, for heights of one-fourth and one-half wavelength. Broken lines show what the pattern would be if there were no reflection from the ground (free space).

horizontally and vertically polarized waves are equal in magnitude and phase. At this angle (approximately 81° for the example case), the effect of ground reflection on vertically and horizontally polarized signals will be the same.

DEPTH OF RF CURRENT PENETRATION

When considering earth characteristics, questions about depth of RF current penetration often arise. For instance, if a given location consists of a 6-foot layer of soil overlying a highly resistive rock strata, which material dominates? The answer depends on the frequency, the soil and rock dielectric constants, and their respective conductivities. The following equation can be used to calculate the current density at any depth.

$$e^{-pd} = \frac{\text{Current Density at Depth d}}{\text{Current Density at Surface}} \qquad \text{(Eq 6)}$$

where

d = depth of penetration in cm

e = natural logarithm base (2.718)

$$p = \left(\frac{X \times B}{2} \times \left(\sqrt{1 + \frac{G^2 \times 10^{-4}}{B^2}} - 1 \right) \right)^{1/2}$$

$X = 0.008 \times \pi^2 \times f$

$B = 5.56 \times 10^{-7} \times k \times f$

k = dielectric constant of earth

f = frequency in MHz

G = conductivity of earth in S/m

After some manipulation of this equation, it can be used to calculate the depth at which the current density is some fraction of that at the surface. The depth at which the current density is 37% (1/*e*) of that at the surface (often referred to as *skin depth*) is the depth at which the current density would be zero if it were distributed uniformly instead of exponentially. (This 1/*e* factor appears in many physical situations. For instance, a capacitor charges to within 1/*e* of full charge within one RC time constant.) At this depth, since the power loss is proportional to the square of the current, approximately 91% of the total power loss has occurred, as has most of the phase shift, and current flow below this level is negligible.

Fig 24 shows the solutions to Eq 6 over the 1.8 to 30-MHz frequency range for various types of earth. For example, in very good earth, substantial RF currents flow down to about 3.3 feet at 14 MHz. This depth goes to 13 feet in average earth and as far as 40 feet in very poor earth. Thus, if the overlying soil is rich, moist loam, the underlying rock stratum is of little concern. However, if the soil is only average, the underlying rock may constitute a major consideration in determining the PBA and the depth to which the RF current will penetrate.

The depth in fresh water is about 156 feet and is nearly independent of frequency in the amateur bands below 30 MHz. In salt water, the depth is about seven inches at 1.8 MHz and decreases rather steadily to about two inches at 30 MHz. Dissolved minerals in moist earth increase its conductivity.

The depth-of-penetration curves in Fig 24 illustrate a noteworthy phenomenon. While skin effect confines RF current flow close to the surface of a conductor, the earth is so lossy that RF current penetrates to much greater depths than in most other media. The depth of RF current penetration is a function of frequency as well as earth type. Thus, the only cases in which most of the current flows near the surface are with very highly conductive media (such as salt water), and at frequencies above 30 MHz.

DIRECTIVE PATTERNS OVER REAL GROUND

As explained in Chapter 2, Antenna Fundamentals, because antenna radiation patterns are three-dimensional, it is helpful in understanding their operation to use a form of representation showing the elevation-plane directional characteristic for different heights. It is possible to show selected elevation-plane patterns oriented in various directions with respect to the antenna axis. In the case of the horizontal half-wave dipole, a plane running in a direction along the axis and another broadside to the antenna will give a good deal of information.

The effect of reflection from the ground can be expressed as a separate *pattern factor*, given in decibels. For any given elevation angle, adding this factor algebraically to the value for that angle from the free-space pattern for that antenna gives the resultant radiation value at that

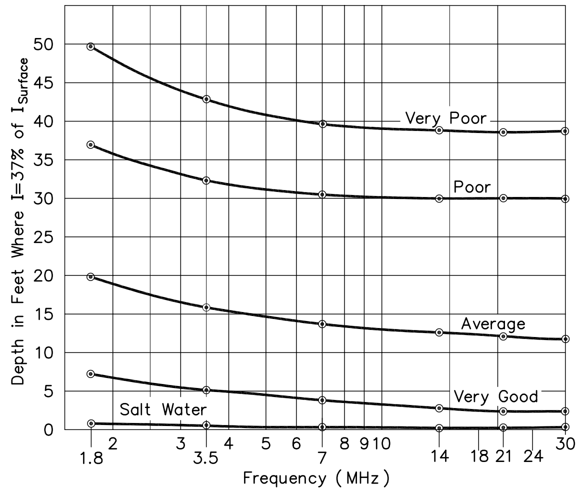

Fig 24—Depths at which the current density is 37% (1/*e*) of that at the surface for different qualities of earth over the 1.8- to 30-MHz frequency range. The depth for fresh water, not plotted, is 156 feet and almost independent of frequency below 30 MHz. See text and Table 2 for ground constants.

angle. The limiting conditions are those represented by the direct ray and the reflected ray being exactly in-phase and exactly out-of-phase, when both (assuming there are no ground losses) have equal amplitudes. Thus, the resultant field strength at a distant point may be either 6 dB greater than the free-space pattern (twice the field strength), or zero, in the limiting cases.

Horizontally Polarized Antennas

The way in which pattern factors vary with height for horizontal antennas over flat earth is shown graphically in the plots of **Fig 25**. The solid-line plots are based on perfectly conducting ground, while the shaded plots are based on typical real-earth conditions. These patterns apply to horizontal antennas of any length. While these graphs are, in fact, radiation patterns of horizontal single-wire antennas (dipoles) as viewed from the axis of the wire, it must be remembered that the plots merely represent pattern factors.

Fig 26 shows vertical-plane radiation patterns in the directions off the ends of a horizontal half-wave dipole for various antenna heights. These patterns are scaled so they may be compared directly to those for the appropriate heights in Fig 25. Note that the perfect-earth patterns in Figs 26A and 25B are the same as those in the upper part of Fig 23. Note also that the perfect-earth patterns of Figs 26B and 25D are the same as those in the lower section of Fig 23. The reduction in field strength off the ends of the wire at the lower angles, as compared with the broadside field strength, is quite apparent. It is also clear from Fig 26 that, at some heights, the high-angle radiation off the ends is nearly as great as the broadside radiation, making the antenna essentially an omnidirectional radiator.

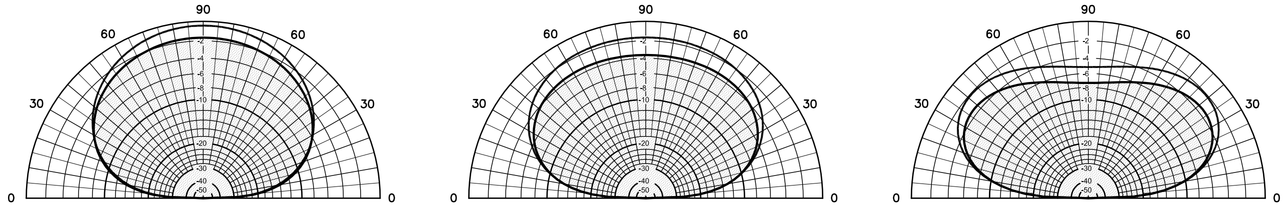

A — Antenna 1/8 λ high B — Antenna 1/4 λ high C — Antenna 3/8 λ high

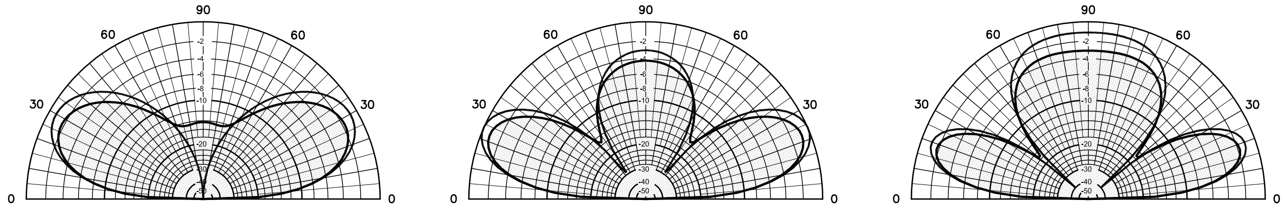

D — Antenna 1/2 λ high E — Antenna 5/8 λ high F — Antenna 3/4 λ high

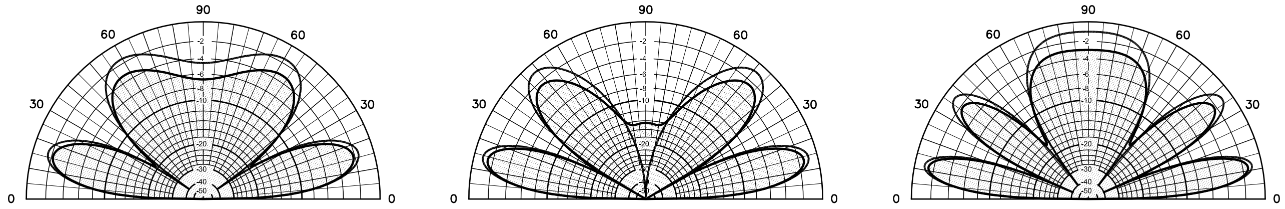

G — Antenna 7/8 λ high H — Antenna 1 λ high I — Antenna 1–1/4 λ high

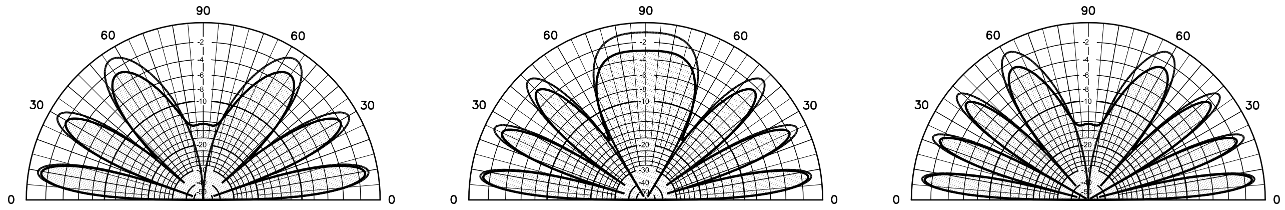

J — Antenna 1–1/2 λ high K — Antenna 1–3/4 λ high L — Antenna 2 λ high

Fig 25—Reflection factors for horizontal dipole antennas at various heights above flat ground. The solid-line curves are the perfect-earth patterns (broadside to the antenna wire); the shaded curves represent the effects of average earth (k = 13, G = 0.005 S/m) at 14 MHz. Add 7 dB to values shown for absolute gain in dBd referenced to dipole in free space, or 9.15 dB for gain in dBi. For example, peak gain over perfect earth at 5/8 λ height is 7 dBd (or 9.15 dBi) at 25° elevation.

In vertical planes making some intermediate angle between 0° and 90° with the wire axis, the pattern will have a shape intermediate between the broadside and end-on patterns. By visualizing a smooth transition from the end-on pattern to the broadside pattern as the horizontal angle is varied from 0° to 90°, a fairly good mental picture of the actual solid pattern may be formed. An example is shown in **Fig 27**. At A, the elevation-plane pattern of a half-wave dipole at a height of λ/2 is shown through a plane 45° away from the favored direction of the antenna. At B and C, the pattern of the same antenna is shown at heights of 3λ/4 and 1λ (through the same 45° off-axis plane). These patterns are scaled so they may be compared directly with the broadside and end-on patterns for the same antenna (at the appropriate heights) in Figs 25 and 26.

The curves presented in **Fig 28** are useful for determining heights of horizontal antennas that give either maximum or minimum reinforcement at any desired wave angle. For instance, if you want to place an antenna at a height so that it will have a null at 30°, the antenna should be placed where a broken line crosses the 30° line on the horizontal scale. There are two heights (up to 2 λ) that will yield this null angle: 1 λ and 2 λ.

As a second example, you may want to have the ground reflection give maximum reinforcement of the direct ray from a horizontal antenna at a 20° elevation angle. The antenna height should be 0.75 λ. The same height will give a null at 42° and a second lobe at 90°.

Fig 28 is also useful for visualizing the vertical pattern of a horizontal antenna. For example, if an antenna is erected at 1.25 λ, it will have major lobes (solid-line crossings) at 12° and 37°, as well as at 90° (the zenith). The nulls in this pattern (dashed-line crossings) will appear at 24° and 53°.

The Y-axis in Fig 28 plots the wave angle versus the height in wavelength above flat ground on the X-axis. Fig 28 doesn't show the elevation angles required for actual

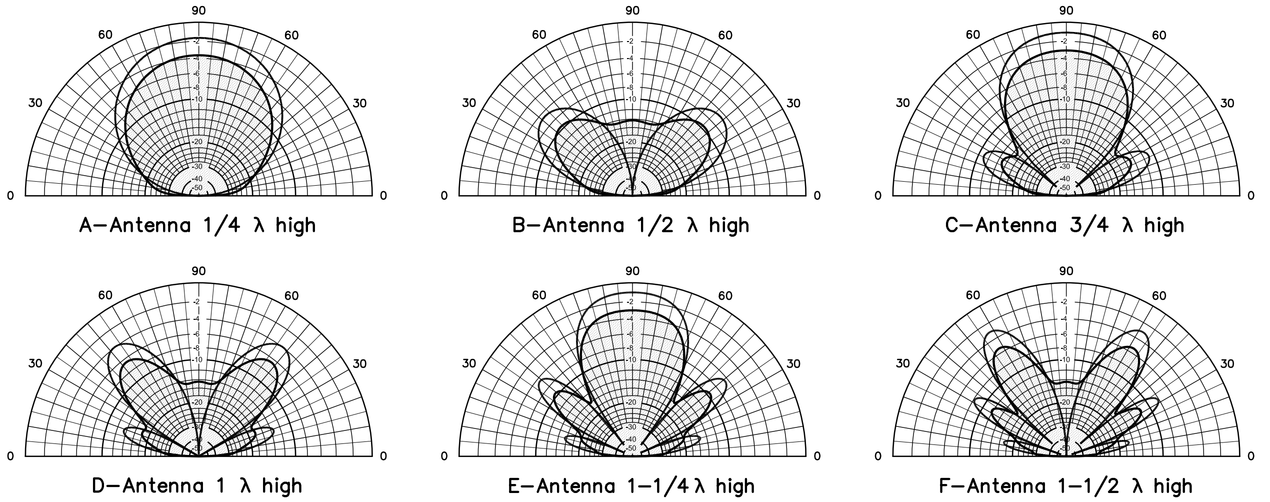

A—Antenna 1/4 λ high
B—Antenna 1/2 λ high
C—Antenna 3/4 λ high
D—Antenna 1 λ high
E—Antenna 1-1/4 λ high
F—Antenna 1-1/2 λ high

Fig 26—Vertical-plane radiation patterns of horizontal half-wave dipole antennas off the ends of the antenna wire. The solid-line curves are the flat, perfect-earth patterns, and the shaded curves represent the effects of average flat earth (k = 13, G = 0.005 S/m) at 14 MHz. The 0-dB reference in each plot corresponds to the peak of the main lobe in the favored direction of the antenna (the maximum gain). Add 7 dB to values shown for absolute gain in dBd referenced to dipole in free space, or 9.15 dB for gain in dBi.

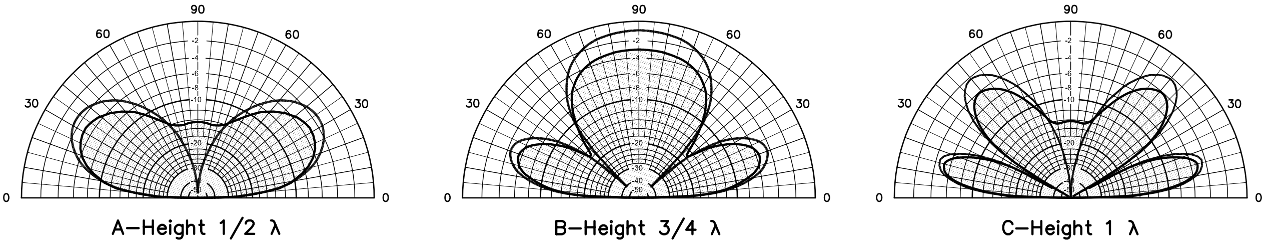

A—Height 1/2 λ
B—Height 3/4 λ
C—Height 1 λ

Fig 27—Vertical-plane radiation patterns of half-wave horizontal dipole antennas at 45° from the antenna wire over flat ground. The solid-line and shaded curves represent the same conditions as in Figs 25 and 26. These patterns are scaled so they may be compared directly with those of Figs 25 and 26.

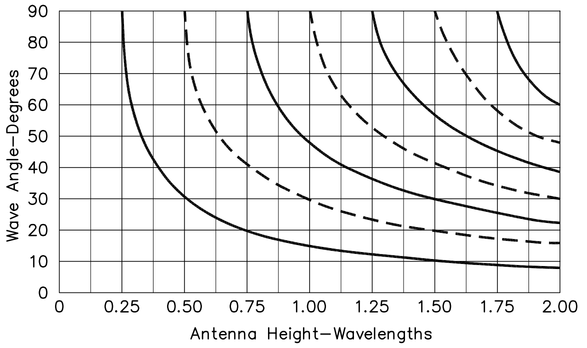

Fig 28—Angles at which nulls and maxima (factor = 6 dB) in the ground-reflection factor appear for antenna heights up to two wavelengths over flat ground. The solid lines are maxima, dashed lines nulls, for all horizontal antennas. See text for examples. Values may also be determined from the trigonometric relationship $\theta = \arcsin (A/4h)$, where θ is the wave angle and h is the antenna height in wavelengths. For the first maximum, A has a value of 1; for the first null A has a value of 2, for the second maximum 3, for the second null 4, and so on.

communications to various target geographic locations of interest. Chapter 23, Radio Wave Propagation, and the CD-ROM in the back of this book give details about the range of angles required for target locations around the world. It is very useful to overlay plots of these angles together with the elevation pattern for horizontally polarized antennas at various heights above flat ground. This will be demonstrated in detail later in this chapter.

Vertically Polarized Antennas

In the case of a vertical $\lambda/2$ dipole or a ground-plane antenna, the horizontal directional pattern is simply a circle at any elevation angle (although the actual field strength will vary, at the different elevation angles, with the height above ground). Hence, one vertical pattern is sufficient to give complete information (for a given antenna height) about the antenna in any direction with respect to the wire. A series of such patterns for various heights is given in **Fig 29**. Rotating the plane pattern about the zenith axis of the graph forms the three-dimensional radiation pattern in each case.

The solid-line curves represent the radiation patterns of the $\lambda/2$ vertical dipole at different feed-point heights over perfectly conducting ground. The shaded curves in Fig 29 show the patterns produced by the same antennas at the same heights over average ground (G = 0.005 S/m, k = 13) at 14 MHz. The PBA in this case is 14.8°.

In short, far-field losses for vertically polarized antennas are highly dependent on the conductivity and dielectric constant of the earth around the antenna, extending far beyond the ends of any radials used to complete the ground return for the near field. Putting more radials out around the antenna may well decrease ground-return losses in the reactive near field for a vertical monopole, but will not increase radiation at low elevation launch angles in the far field, unless the radials can extend perhaps 100 wavelengths in all directions! Aside from moving to the fabled "salt water swamp on a high hill," there is very little that someone can do to change the character of the ground that affects the far-field pattern of a real vertical. Classical texts

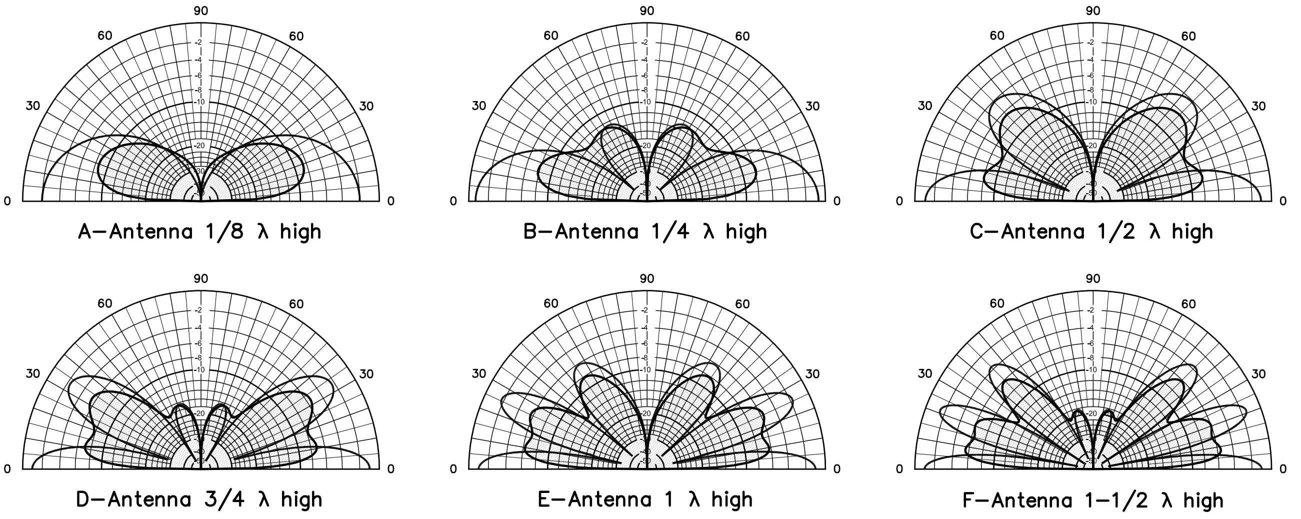

A—Antenna 1/8 λ high

B—Antenna 1/4 λ high

C—Antenna 1/2 λ high

D—Antenna 3/4 λ high

E—Antenna 1 λ high

F—Antenna 1-1/2 λ high

Fig 29—Vertical-plane radiation patterns of a ground-plane antenna above flat ground. The height is that of the ground plane, which consists of four radials in a horizontal plane. Solid lines are perfect-earth patterns; shaded curves show the effects of real earth. The patterns are scaled—that is, they may be directly compared to the solid-line ones for comparison of losses at any wave angle. These patterns were calculated for average ground (k = 13, G = 0.005 S/m) at 14 MHz. The PBA for these conditions is 14.8°. Add 6 dB to values shown for absolute gain in dBd over dipole in free space.

on verticals often show elevation patterns computed over an "infinitely wide, infinitely conducting ground plane." Real ground, with finite conductivity and less than perfect dielectric constant, can severely curtail the low-angle radiation at which verticals are supposed to excel.

While real verticals over real ground are not a surefire method to achieve low-angle radiation, cost versus performance and ease of installation are still attributes that can highly recommend verticals to knowledgeable builders. Practical installations for 160 and 80 meters rarely allow amateurs to put up horizontal antenna high enough to radiate effectively at low elevation angles. After all, a half-wave on 1.8 MHz is 273 feet high, and even at such a lofty height the peak radiation would be at a 30° elevation angle.

The Effects of Irregular Local Terrain in the Far Field

The following material is condensed and updated from an article by R. Dean Straw, N6BV, in July 1995 *QEX* magazine. *HFTA* (HF Terrain Assessment) and supporting data files are included on the CD-ROM at the back of this book. *HFTA* is the latest version of the *YT* program included with earlier editions of *The ARRL Antenna Book.*

Choosing a QTH for DXing

The subject of how to choose a QTH for working DX has fascinated hams since the beginning of amateur operations. No doubt, Marconi probably spent a lot of time wandering around Newfoundland looking for a great radio QTH before making the first transAtlantic transmission. Putting together a high-performance HF station for contesting or DXing has always followed some pretty simple rules. First, you need the perfect QTH, preferably on a rural mountaintop or at least on top of a hill. Even better yet, you need a mountaintop surrounded by seawater! Then, after you have found your dream QTH, you put up the biggest antennas you possibly can, on the highest towers you can afford. Then you work all sorts of DX—sunspots willing, of course.

The only trouble with this straightforward formula for success is that it doesn't always work. Hams fortunate enough to be located on mountain tops with really spectacular drop-offs often find that their highest antennas don't do very well, especially on 15 or 10 meters, but often even on 20 meters. When they compare their signals with nearby locals in the flatlands, they sometimes (but not always) come out on the losing end, especially when sunspot activity is high.

On the other hand, when the sunspots drop into the cellar, the high antennas on the mountaintop are usually the ones crunching the pileups—but again, not always. So, the really ambitious contest aficionados, the guys with lots of resources and infinite enthusiasm, have resorted to putting up antennas at all possible heights, on a multitude of towers.

There is a more scientific way to figure out where and how high to put your antennas to optimize your signal during all parts of the 11-year solar cycle. We advocate a *system approach* to HF station design, in which you need to know the following:

1. The range of elevation angles necessary to get from point A to point B
2. The elevation patterns for various types and configurations of antennas
3. The effect of local terrain on elevation patterns for horizontally polarized antennas.

WHAT IS THE RANGE OF ELEVATION ANGLES NEEDED?

Up until 1994, *The ARRL Antenna Book* contained only a limited amount of information about the elevation angles needed for communication throughout the world. In the 1974 edition, Table 1-1 in the Wave Propagation chapter was captioned: "Measured vertical angles of arrival of signals from England at receiving location in New Jersey."

What the caption didn't say was that Table 1-1 was derived from measurements made during 1934 by Bell Labs. The highest frequency data seemed pretty shaky, considering that 1934 was the low point of Cycle 17. Neither was this data applicable to any other path, other than the one from New Jersey to England. Nonetheless, many amateurs located throughout the US tried to use the sparse information in Table 1-1 as the only rational data they had for determining how high to mount their antennas. (If they lived on hills, they made estimates of the effect of the terrain, assuming that the hill was adequately represented by a long, unbroken slope. More on this later.)

In 1993 ARRL HQ embarked on a major project to tabulate the range of elevation angles from all regions of the US to important DX QTHs around the world. This was accomplished by running many thousands of computations using the *IONCAP* computer program. *IONCAP* has been under development for more than 25 years by various agencies of the US government and is considered the standard of comparison for propagation programs by many agencies, including the Voice of America, Radio Free Europe, and more than 100 foreign governments throughout the world. *IONCAP* is a real pain in the neck to use, but it is the standard of comparison.

The calculations were done for all levels of solar activity, for all months of the year, and for all 24 hours of the day. The results were gathered into some very large databases, from which special custom-written software extracted detailed statistics. The results appeared in summary form in Tables 4 through 13 printed in Chapter 23, Radio Wave Propagation, of the 17th Edition and in more detail on the diskette included with that book. (This book, the 20th Edition, contains even more statistical data, for more areas of the world, on the accompanying CD-ROM.)

Fig 30 shows the full range of elevation angles (represented as vertical bars) for the 20-meter path from New England (centered on Newington, Connecticut) to all of Europe. This is for all openings, in all months, over the entire 11-year solar cycle. The most likely elevation angle occurs at 5° for about 13% of the times when the 20-meter band is open to Europe from New England. From 4° to 6° the band is open a total of about 34% of the times the band is open. There is a secondary peak between 10° to 12°, occurring for a total of about 25% of the times the band is open.

Overlaid on Fig 30 along with the elevation-angle statistics are the elevation-plane responses for three different horizontally polarized Yagi beams, all over flat ground. The first is mounted 140 feet high, 2 λ in terms of wavelength. The second Yagi is mounted 70 feet high (at 1 λ) and the third is 35 feet (0.5 λ). The 140-foot high antenna has a deep null at 15°, but it also has the highest response (13.4 dBi) of the three at the statistical peak elevation angle of 5°. However, at 12°—where the band is open some 9% of the time—the 140-foot high Yagi is down 4 dB compared to the 70-foot antenna.

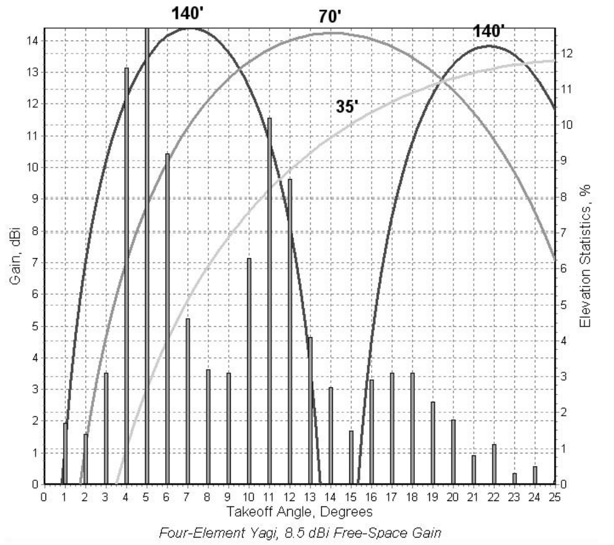

Fig 30—Graph showing 20-meter percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for three 20-meter antenna systems. The most statistically likely angle at which the band will be open is 5°, although at any particular hour, day, month and year, the actual angle will likely be different. Note the deep null exhibited by the 140-foot high antenna centered at 14°.

The 70-foot high Yagi arguably covers the overall range best, since it has no disastrous nulls in the 1° to 25° range, where most of the action is occurring on 20 meters. At 5°, however, its response is only 8.8 dBi, 4.6 dB down from the 140-foot high antenna at that angle. The 35-foot antenna peaks above 26° in elevation angle, and is down some 10.4 dB compared to the 140-foot antenna at 5°. Obviously, no single antenna covers the complete range of elevation angles needed.

Note that the highest Yagi has a strong *second lobe* peaking at 22°. Let's say that you could select between two antennas, one at 140 and one at 70 feet, and that the incoming angle for a particular distant station is 22°. You might be fooled into thinking that the incoming angle is around 6°, favoring the first peak of the higher antenna, when in truth the angle is relatively high. The 70-foot antenna's response would be lower at 22° than the higher one, but only because the 140-foot antenna is operating on its second lobe. (What would clinch a determination of the correct incoming angle—6° or 22°—would be the response of the 35-foot high Yagi, which would be close to its peak at 22°, while it would be very far down at 6°.)

Now, we must emphasize that these elevation angles are *statistical entities*— in other words, just because 5° is the "statistically most likely angle" for the 20-meter path from New England to Europe doesn't mean that the band will be open at 11° at any particular hour, on a particular day, in a particular month, in any particular year. In fact, however, experience agrees with the *IONCAP* computations: the 20-meter path to Europe usually opens at a low angle in the New England morning hours, rising to about 11° during the afternoon, when the signals remain strongest throughout the afternoon until the evening in New England.

What would happen if we were to feed all three Yagi at 140, 70 and 35 feet in-phase as a stack? **Fig 31** shows this situation, along with a more highly optimized stack at 120, 80 and 40 feet that better covers the overall range of elevation angles from Connecticut to Europe.

Now see **Fig 32**, which uses the same 120/80/40-foot stack of 20-meter antennas as in Fig 31, but this time from Seattle, Washington, to Europe. For comparison, the response of a single 4-element Yagi at 100 feet over flat ground is also shown in Fig 32. Just because 5º is the statistically most prevalent angle (occurring some 13% of the time) from Seattle to Europe on 20 meters, this doesn't mean that the actual angle *at any particular moment in time* might not be 10º, or even 2º. The statistics for W7 to Europe say that 5º is the most likely angle, but 20-meter signals from Europe arrive at angles ranging from 1º to 18º. Note that this range of angles is quite a bit less than from W1 to Europe, which is much closer geographically to Europe than is the Pacific Northwest coast of the US. If you design an antenna system to cover all possible angles needed to talk to Europe from Seattle (or from Seattle to Europe) on 20 meters, you would need to cover the full range from 1º to 18º equally well.

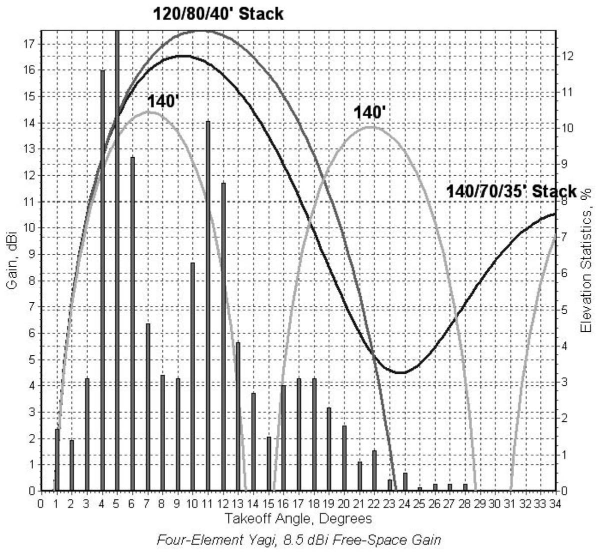

Fig 31—Graph showing results of stacking antennas at different heights on the same tower to cover a wider range of elevation angles, in this case for the path from Connecticut (W1) to all of Europe on 20 meters. The optimized stack at 120/80/40 feet covers the needed range of elevation angles better than the stack at 140/70/35 feet or the single Yagi at 140 feet.

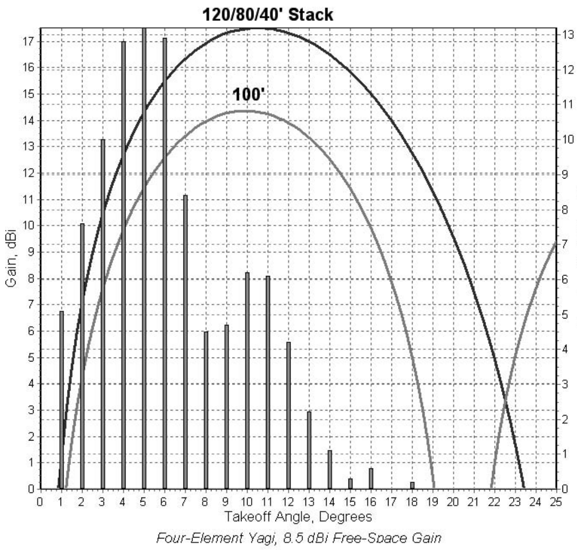

Fig 32—Graph showing 20-meter percentage of all openings, this time from Seattle, WA, to Europe, together with an overlay of elevation patterns over flat ground for two 20-meter antenna systems. The statistically most likely angle on this path is 5°, occurring about 13% of the time when the band is actually open. Higher antennas predominate on this low-angle path.

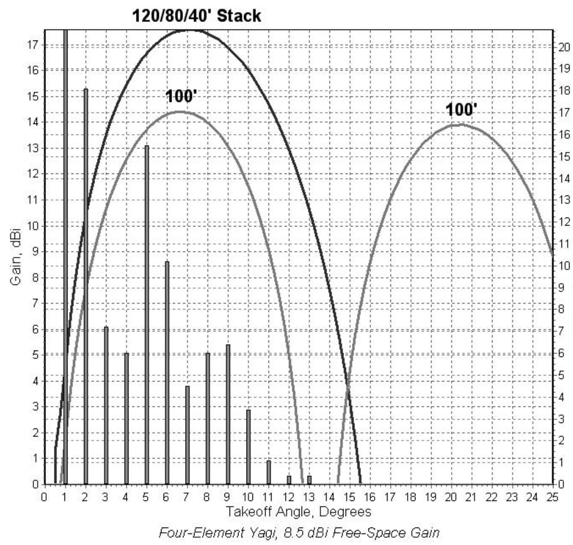

Fig 33—Graph showing 15-meter percentage of all openings from Chicago to Southern Africa, together with overlay of elevation patterns over flat ground for two 15-meter antenna systems. On this long-distance, low-angle path, higher antennas are again most effective.

Similarly, if you wish to cover the full range of elevation angles from Chicago to Southern Africa on 15 meters, you would need to cover 1º to 13º, even though the most statistically likely signals arrive at 1º, for 21% of the time when that the band is open for that path. See **Fig 33**.

It is important to recognize that Figs 30 through 33 are for flat ground. When the antennas are mounted over irregular local terrain, things get much more complicated. First, however, we'll discuss general-purpose antenna modeling programs as they try to model real terrain.

DRAWBACKS OF COMPUTER MODELS FOR ANTENNAS OVER REAL TERRAIN

Modern general-purpose antenna modeling programs such as *NEC* or *MININEC* (or their commercially upgraded equivalents, such as *NEC-Win Plus*, *EZNEC* and *EZNEC ARRL*) can accurately model almost any type of antenna commonly used by radio amateurs. In addition, there are specialized programs specifically designed to model Yagis efficiently, such as *YO* or *YW* (Yagi for Windows, bundled on the CD-ROM with this book) or *YagiMax*. These programs however are all unable to model antennas accurately over anything other than *purely flat ground*.

While both *NEC* and *MININEC* can simulate irregular ground terrain, they do so in a decidedly crude manner, employing step-like concentric rings of height around an antenna. The documentation for *NEC* and *MININEC* both clearly state that diffraction off these steps is not modeled.

Common experience among serious modelers is that the warnings in the manuals are worth heeding.

Although you can analyze and even optimize antenna designs using free-space or flat-earth ground models, it is *diffraction* that makes the real world a very, very complicated place. This should be clarified—diffraction is hard, even tortuous, to analyze properly, but it makes analysis of real world results far more believable than a flat-world reflection model does.

RAY-TRACING OVER UNEVEN LOCAL TERRAIN

The Raytracing Technique

First, let's look at a simple raytracing procedure involving only horizontally polarized reflections, with no diffractions. From a specified height on the tower, an antenna shoots "rays" (just as though they were bullets) in 0.25° increments from +35° above the horizon to –35° below the horizon. Each ray is traced over the foreground terrain to see if it hits the ground at any point on its travels in the direction of interest. If it does hit the ground, the ray is reflected following the classical *law of reflection*. That is, the outgoing angle equals the incoming angle, reflected through the normal to the slope of the surface. Once the rays exit into the ionosphere, the individual contributions are vector-summed to create the overall far-field elevation pattern.

The next step in terrain modeling involves adding *diffractions* as well as reflections. At the Dayton antenna forum in 1994, Jim Breakall, WA3FET, gave a fascinating and tantalizing lecture on the effect of foreground terrain. Later Breakall, Dick Adler, K3CXZ, Joel Young and a group of other researchers published an extremely interesting paper entitled "The Modeling and Measurement of HF Antenna Skywave Radiation Patterns in Irregular Terrain" in July 1994 *IEEE Transactions on Antennas and Propagation*. They described in rather general terms the modifications they made to the *NEC-BSC* program. They showed how the addition of a ray-tracing reflection and diffraction model to the simplistic stair-stepped reflection model in regular *NEC* gave far more realistic results. For validation, they compared actual pattern measurements made on a site in Utah (with an overflying helicopter) to computed patterns made using the modified *NEC* software. However, because the US Navy funded this work the software remained for a long time a military secret.

Thumbnail History of the Uniform Theory of Diffraction

It is instructive to look briefly at the history of how *Geometric Optics* (GO) evolved (and still continues to evolve) into the *Uniform Theory of Diffraction* (UTD). The following is summarized from the historical overview in one book found to be particularly useful and comprehensive on the subject of UTD: *Introduction to the Uniform Geometrical Theory of Diffraction*, by McNamara, Pistorius, and Malherbe.

Many years before the time of Christ, the ancient Greeks studied optics. Euclid is credited with deriving the law of reflection about 300 BC. Other Greeks, such as Ptolemy, were also fascinated with optical phenomena. In the 1600s, a Dutchman named Snell finally figured out the law of refraction, resulting in *Snell's law*. By the early 1800s, the basic world of classical optics was pretty well described from a mathematic point of view, based on the work of a number of individuals.

As its name implies, classical geometric optical theory deals strictly with geometric shapes. Of course, the importance of geometry in optics shouldn't be minimized—after all, we wouldn't have eyeglasses without geometric optics. Mathematical analysis of shapes utilizes a methodology that traces the paths of straight-line *rays* of light. (Note that the paths of rays can also be likened to the straight-line paths of particles.) In classical geometric optics, however, there is no mention of three important quantities: phase, intensity and polarization. Indeed, without phase, intensity or polarization, there is no way to deal properly with the phenomenon of *interference*, or its cousin, *diffraction*. These phenomena require theories that deal with *waves* rather than rays.

Wave theory has also been around for a long time, although not as long as geometry. Workers like Hooke and Grimaldi had recorded their observations of interference and diffraction in the mid 1600s. Huygens had used elements of wave theory in the late 1600s to help explain refraction. By the late 1800s, the work of Lord Rayleigh, Sommerfeld, Fresnel, Maxwell and many others led to the full mathematic characterization of all electromagnetic phenomena, light included.

Unfortunately, ray theory doesn't work for many problems, at least ray theory in the classical optical form. The real world is a lot more jagged, pointy and fuzzy in shape than can be described in a totally rigorous mathematic fashion. Some properties of the real world are most easily explained on the micro level using electrons and protons as conceptual objects, while other macro phenomena (like resonance, for example) are more easily explained in terms of waves. To get a handle on a typical real-world physical situation, a combination of classical ray theory and wave theory was needed.

The breakthrough in the combination of classical geometric optics and wave concepts came from J. B. Keller of Bell Labs in 1953, although he published his work in the early 1960s. In the very simplest of terms, Keller introduced the notion that shooting a ray at a diffraction *wedge* causes wave interference at the tip, with an infinite number of diffracted waves emanating from the diffraction point. Each diffracted wave can be considered to be a point source radiator at the place of generation, the diffraction point. Thereafter, the paths of individual waves can be traced as though they were individual classical optic rays again. What Keller came up with was a reasonable mathematical description of what happens at the tip of the diffraction wedge.

Fig 34 is a picture of a simple diffraction wedge, with an incoming ray launched at an angle of α_r, referenced to the horizon, impinging on it. The diffraction wedge here is considered to be perfectly conducting, and hence impenetrable by the ray. The wedge generates an infinite number of diffracted waves, going in all directions not blocked by the wedge itself. The amplitudes and phases of the diffracted waves are determined by the interaction at the wedge tip, and this in turn is governed by the various angles associated with the wedge. Shown in Fig 34 are the included angle a of the wedge, the angle ϕ' of the incoming ray (referenced to the incoming surface of the wedge), and the observed angle ϕ of one of the outgoing diffracted waves, also referenced to the wedge surface.

The so-called *shadow boundaries* are also shown in Fig 34. The Reflection-Shadow Boundary (RSB) is the angle beyond which no further reflections can take place for a given incoming angle. The Incident-Shadow Boundary (ISB) is that angle beyond which the wedge's face blocks any incident rays from illuminating the observation point.

Keller derived the amplitude and phase terms by comparing the classical Geometric Optics (GO) solution with the exact mathematical solution calculated by Sommerfeld for a particular case where the boundary conditions were well known—an infinitely long, perfectly conducting wedge illuminated by a plane wave. Simply speaking, whatever was left over had to be diffraction terms. Keller combined these diffraction terms with GO terms to yield the total field everywhere.

Keller's new theory became known as the *Geometric Theory of Diffraction* (abbreviated henceforth as GTD). The beauty of GTD was that in the regions where classical GO predicted zero fields, the GTD "filled in the blanks," so to speak. For example, see **Fig 35**, showing the terrain for a hypothetical case, where a 60-foot high 4-element 15-meter Yagi illuminates a wide, perfectly flat piece of ground. A 10-foot high rock has been placed 400 feet away from the tower base in the direction of outgoing rays. **Fig 36** shows the elevation pattern predicted using reflection-only GO techniques. Due to blockage of the direct wave (A) trying to shoot past the 10-foot high rock, and due to blockage of (B) reflections from the flat ground in front of the rock by the rock, there is a *hole* in the smooth elevation pattern.

Now, doesn't it defy common sense to imagine that a single 10-foot high rock will really have such an effect on a 15-meter signal? Keller's GTD took diffraction effects into account to show that waves do indeed sneak past and over the rock to fill in the pattern. The whole GTD scheme is very clever indeed.

However, GTD wasn't perfect. Keller's GTD predicts some big spikes in the pattern, even though the overall shape of the elevation pattern is much closer to reality than a simple GO reflection analysis would indicate. The region right at the RSB and ISB shadow boundaries is where problems are found. The GO terms go to zero at these points because of blockage by the wedge, while Keller's diffraction terms tend to go to infinity at these very spots. In mathematical terms this is referred to as a *caustic problem*. Nevertheless, despite these nasty problems at the ISB and RSB, the GTD provided a remarkably better solution to diffraction problems than did classical GO.

In the early 1970s, a group at Ohio State University under R. G. Kouyoumjian and P. H. Pathak did some pivotal work to resolve this caustic problem, introducing what amounts to a clever *fudge factor* to compensate for the tendency of the diffraction terms at the shadow boundaries to go to infinity. They introduced what is known as a *transition function*, using a form of Fresnel integral. Most importantly, the Ohio State researchers also created several FORTRAN computer programs to compute the amplitude and phase of diffraction components. Now computer hackers could get to work!

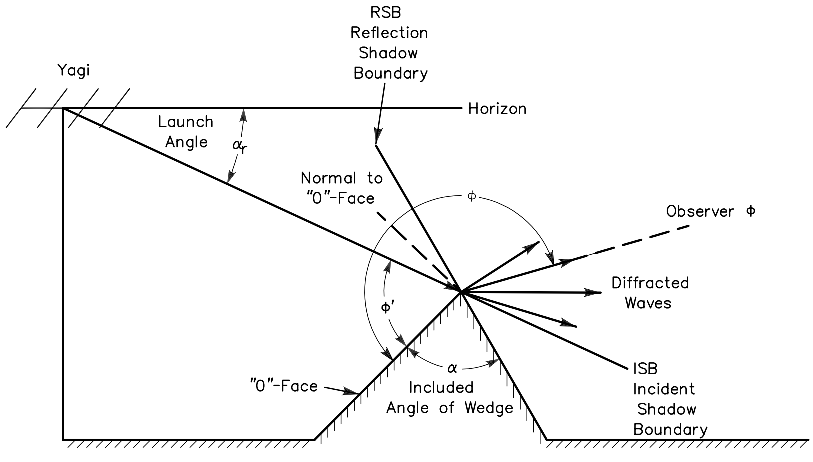

Fig 34—Diagram showing diffraction mechanism of ray launched at angle α_r below the horizon at a diffraction wedge, whose included angle is α. Referenced to the incident face (the *o-face* as it is called in UTD terminology), the incoming angle is ϕ' (phi prime). The wedge creates an infinite number of diffracted waves. Shown is one whose angle referenced to the o-face is ϕ, the so-called *observation angle* in UTD terminology.

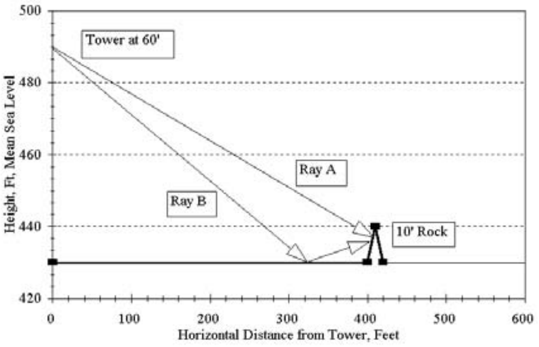

Fig 35—Hypothetical terrain exhibiting so-called “10-foot rock effect.” The terrain is flat from the tower base out to 400 feet, where a 10-foot high rock is placed. Note that this forms a diffraction wedge, but that it also blocks direct waves trying to shoot through it to the flat surface beyond, as shown by Ray A. Ray B reflects off the flat surface before it reaches the 10-foot rock, but it is blocked by the rock from proceeding further. A simple Geometric Optics (GO) analysis of this terrain without taking diffraction into account will result in the elevation response shown in Fig 36.

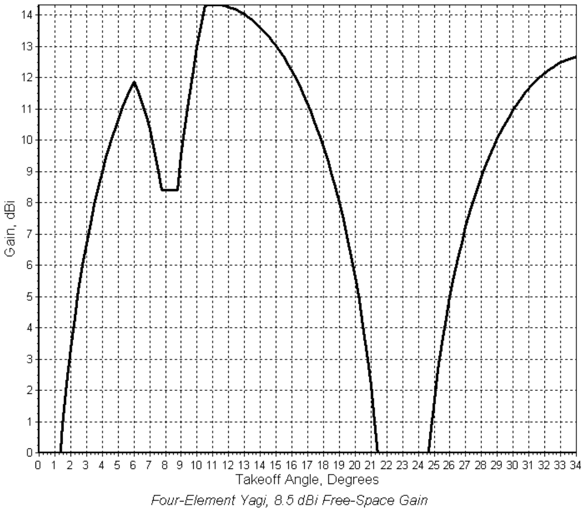

Fig 36—Elevation response for rays launched at terrain in Fig 35 from a height of 60 feet using a 4-element Yagi. This was computed using a simple Geometrical Optics (GO) reflection-only analysis. Note the *hole* in the response between 6° to 10° in elevation. It is not reasonable for a 10-foot high rock to create such a disturbance at 21 MHz!

The program that finally resulted is called *HFTA*, standing for “HF Terrain Assessment.” (The DOS version of *HFTA* was known as *YT*, standing for “Yagi Terrain.”) As the name suggests, *HFTA* analyzes the effect of local terrain on HF propagation through the ionosphere. It is designed for horizontally polarized Yagis, although it will model the effects of a simple flattop dipole also. The accurate appraisal of the effect of terrain on vertically polarized signals is a far more complex problem than for horizontally polarized waves, and *HFTA* doesn’t do verticals.

SIMULATION OF REALITY—SOME SIMPLE EXAMPLES FIRST

We want to focus first on some simple results, to show that the computations do make some sense by presenting some simulations over simple terrains. We’ve already described the “10-foot rock at 400 feet” situation, and showed where a simple GO reflection analysis is inadequate to the task without taking diffraction effects into account.

Now look at the simple case shown in **Fig 37**, where a very long, continuous downslope from the tower base is shown. Note that the scales used for the X and Y axes are different: the Y-axis changes 300 feet in height (from 800 to 1100 feet), while the X-axis goes from 0 to 3000 feet. This exaggerates the apparent steepness of the downwards slope, which is actually a rather gentle slope, at $\tan^{-1} (1000 - 850) / (3000 - 0) = -2.86°$. In other words, the terrain falls 150 feet in height over a range of 3000 feet from the base of the tower.

Fig 38 shows the computed elevation response for this terrain profile, for a 4-element horizontally polarized Yagi on a 60-foot tower. The response is compared to that of an identical Yagi placed 60 feet above flat ground. Compared to the “flatland” antenna, the hilltop antenna has an elevation response shifted over by almost 3° towards the lower elevation angles. In fact, this shift is directly due to the –2.86° slope of the hill. Reflections off the slope are tilted by the slope. In this situation there is a single diffraction at the bottom of the gentle slope at 3000 feet, where the program assumes that the terrain becomes flat.

Look at **Fig 39**, which shows another simple terrain profile, called a “Hill-Valley” scenario. Here, the 60-foot high tower stands on the edge of a gentle hill overlooking a long valley. Once again the slope of the hill is exaggerated by the different X and Y-axes. **Fig 40** shows the computed elevation response at 21.2 MHz for a 4-element Yagi on a 60-foot high tower at the edge of the slope.

Once again, the pattern is overlaid with that of an identical 60-foot-high Yagi over flat ground. Compared to the flatland antenna, the hilltop antenna’s response above 9° in elevation is shifted by almost 3° towards the lower elevation angles. Again, this is due to reflections off the downward slope. From 1° to 9°, the hilltop pattern is enhanced even more compared to the flatland antenna, this time by diffraction occurring at the bottom of the hill.

Now let’s see what happens when there is a hill ahead in the direction of interest. **Fig 41** depicts such a situation, labeled “Hill-Ahead.” Here, at a height of 400 feet above mean sea level, the land is flat in front of the tower, out to a distance 500 feet, where the hill begins. The hill then rises 100 feet over the range 500 to 1000 feet away

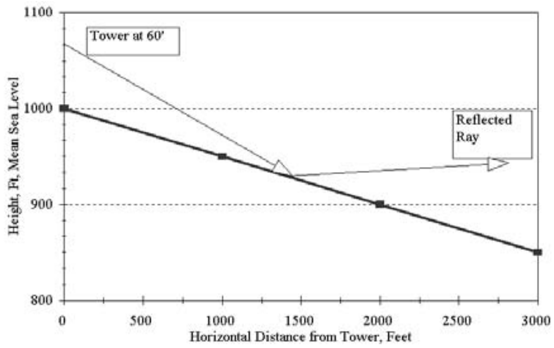

Fig 37—A long, gentle downwards-sloping terrain. This terrain has no explicit diffraction points and can be analyzed using simple GO reflection techniques.

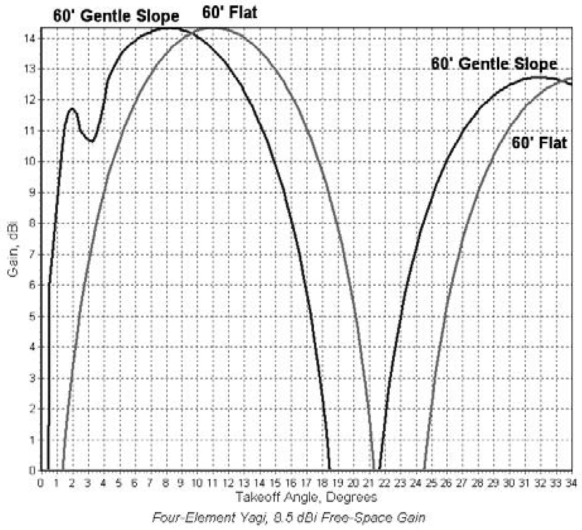

Fig 38—Elevation response for terrain shown in Fig 37, using a 4-element 15-meter Yagi, 60-foot high. Note that the shape of the response is essentially shifted towards the left, towards lower elevation angles, by the angle of the sloping ground. For reference, the response for an identical Yagi placed over flat ground is also shown.

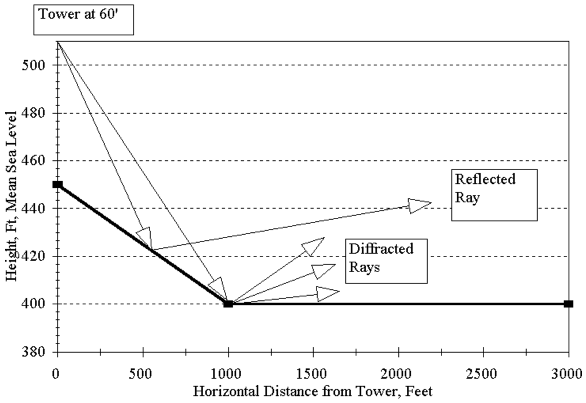

Fig 39—"Hill-Valley" terrain, with reflected and diffracted rays.

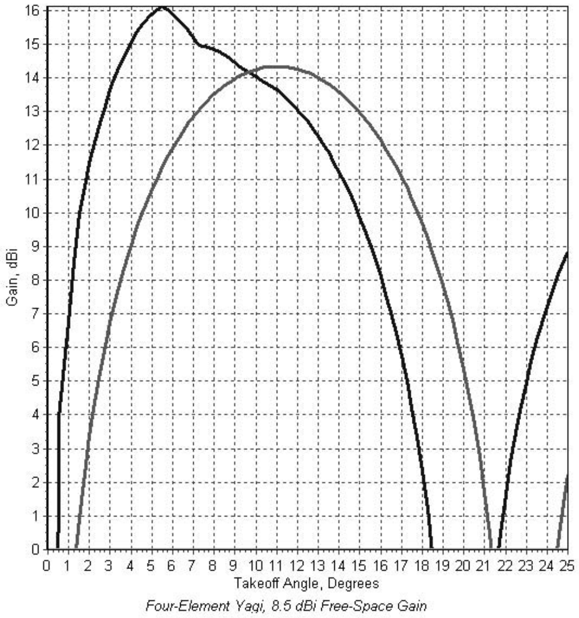

Fig 40—Elevation response computed by *HFTA* program for single 4-element 15-meter Yagi at 60 feet above "Hill-Valley" terrain shown in Fig 39. Note that the slope has caused the response in general to be shifted towards lower elevation angles. At 5° elevation, the diffraction components add up to increase the gain slightly above the amount a GO-only analysis would indicate.

from the tower base. After that, the terrain is a plateau, at a constant 500 feet elevation.

Fig 42 shows the computed elevation pattern for a 4-element 21-MHz Yagi 60-feet high on the tower, compared again with an overlay for an identical 60-foot high antenna over flat ground. The hill blocks low-angle waves directly radiated from the antenna from 0° to 2.3°. In addition, waves that would normally be reflected from the ground, and that would normally add in phase from about 2.3° to 12°, are blocked by the hill also. Thus the signal at 8° is down almost 5 dB from the signal over flat ground, all due to the effect of the hill. Diffracted waves start kicking in once the direct wave rises enough above the horizon to illuminate the top edge of the hill. These diffracted waves tend to augment elevation angles above about 12°, which reflected waves can't reach.

Is there is any hope for someone in such a lousy QTH for DXing? **Fig 43** shows the elevation response for a truly heroic solution. This involves a stack of four 4element Yagis, mounted at 120, 90, 60 and 30 feet on the tower. Now, the total gain at low angles is just about comparable to that from a single 4-element Yagi mounted over flat ground. Where there's a ham, there is a way!

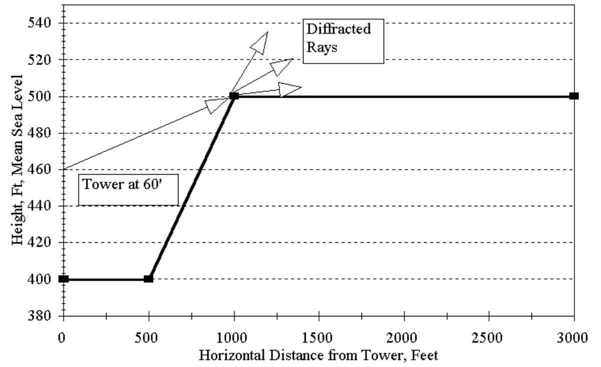

Fig 41—"Hill-Ahead" terrain, shown with diffracted rays created by illumination of the edge of the plateau at the top of the hill.

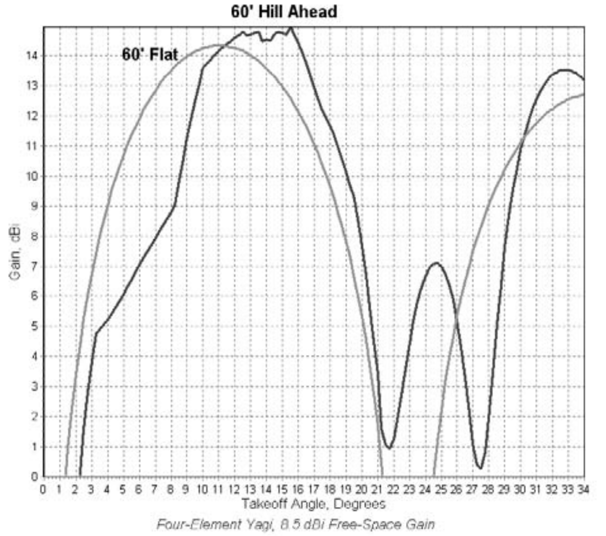

Fig 42—Elevation response computed by *HFTA* for "Hill-Ahead" terrain shown in Fig 41. Now the hill blocks direct rays and also precludes possibility of any constructive reflections. Above 10°, diffraction components add up together with direct rays to create the response shown.

At 5° elevation, four diffraction components add up (there are zero reflection components) to achieve the far-field pattern. This seems reasonable, because each of the four antennas is illuminating the diffraction point separately and we know that none of the four antennas can *see over* the hill directly to produce a reflection at a low launch angle.

At an elevation angle of 5°, 15-meter signals arrive from Europe from New England about 13% of the total time when the band is actually open. We can look at this another way. For about two-thirds of the times when the band is open on this path, the incoming angle is between 3° to 12°. For about one-third of the time, signals arrive above 10°, where the "heroic" four-stack is really beginning to come into its own.

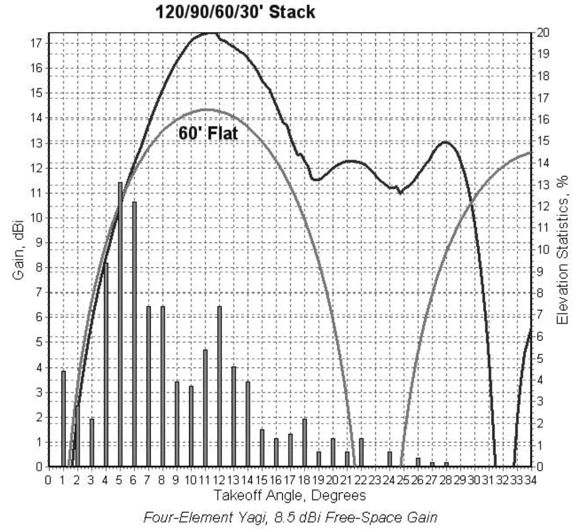

Fig 43—Elevation response of "heroic effort" to surmount the difficulties imposed by hill in Fig 41. This effort involves a stack of four 4-element Yagis in a stack starting at 120 feet and spaced at 30-foot increments on the tower. The response is roughly equivalent to a single 4-element Yagi at 60 feet above flat ground, hence the characterization as being a "heroic effort." The elevation-angle statistics from New England to Europe are overlaid on the graph for reference.

A More Complex Terrain

The results for simple terrains look reasonable; let's try a more complicated real-world situation. **Fig 44** shows the terrain from the New Hampshire N6BV/1 QTH towards Japan. The terrain was complex, with 52 different points *HFTA* identifies as diffraction points. **Fig 45** shows a labeled *HFTA* output for three different types of antennas on 20 meters: a stack at 120 and 60 feet, the 120-foot antenna by itself, and then a 120/60-foot stack over flat ground, for reference. The elevation-angle statistics for New England to Japan are overlaid on the graph also, making for a very complicated looking picture—it is a *lot* easier to decipher the lines on the color CRT, by the way than on a black-and-white printer.

Comparison of the same 120/60-foot stacks over irregular terrain and flat ground is useful to show where the terrain itself is affecting the elevation response. The flatland stack has more gain in the region of 3° to 7° than the same stack over the N6BV/1 local terrain towards Japan. On the other hand, the N6BV/1 local terrain boosts signals in the range of 8° to about 12°. This demonstrates the conservation of energy—you may gain a stronger signal at certain elevation angles, but you will lose gain at others. In this case, the N6BV/1 station always felt "weak" towards Japan on 20 meters, because the dominant angles are low.

Examination of the detailed data output from *HFTA* shows that at an elevation angle of 5°, there are 6159

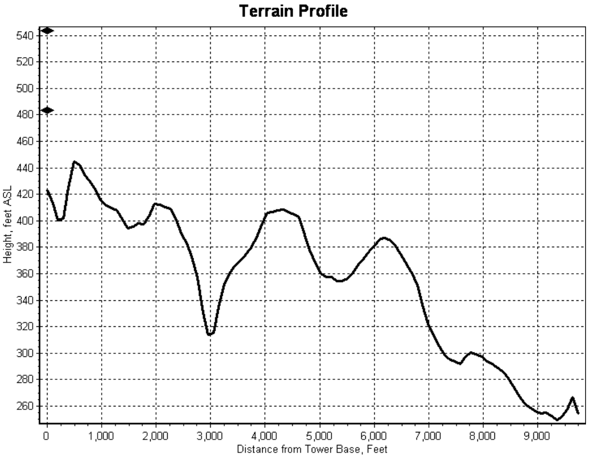

Fig 44—Terrain of N6BV in Windham, NH, towards Japan. *HFTA* identifies 52 different points where diffraction can occur.

diffraction components. There are many, many signals bouncing around off the terrain on their trip to Japan! Note that because of blockage of some parts of the terrain, the 60-foot high Yagi cannot illuminate all the diffraction points, while the higher 120-foot Yagi is able to see these diffraction points.

It is fascinating to reflect on the thought that received signals coming down from the ionosphere to the receiver are having encounters with the terrain, but from the opposite direction. It's not surprising, given these kinds of interactions, that transmitting and receiving might not be totally reciprocal.

The 120/60-foot stack in Fig 45 achieves its peak gain of 17.3 dBi at 11° elevation, where it is about 3 dB stronger than the single Yagi at 120 feet. It maintains this 3-dB advantage over most of the range of incoming signals from Japan. This difference in performance between the stack and each antenna by itself was observed many times on the air. Much of the time when comparisons are being made, however, the small differences in signal are difficult to measure meaningfully, especially when the QSB varies signals by 20 dB or so during a typical QSO. It should be noted that the stack usually exhibited less fading compared to each antenna by itself.

USING *HFTA*

Manually Generating a Terrain Profile

The *HFTA* program uses two distinct algorithms to generate the far-field elevation pattern. The first is a simple reflection-only Geometric Optics (GO) algorithm. The second is the diffraction algorithm using the Uniform Theory of Diffraction (UTD). These algorithms work with a digitized representation of the terrain profile for a single azimuthal direction—for example, towards Japan or towards Europe.

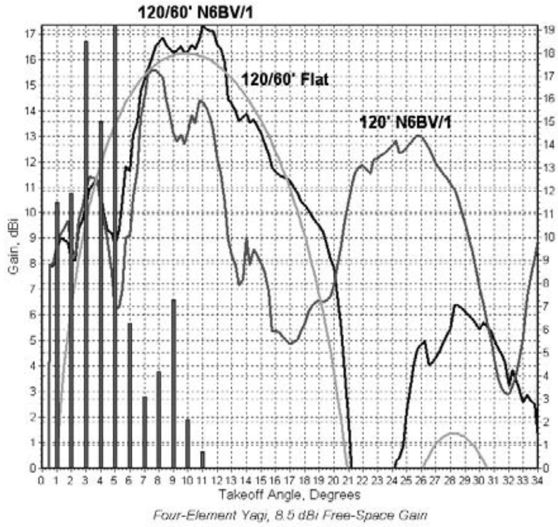

Fig 45—Elevation responses computed by *HFTA* for N6BV/1 terrain shown in Fig 44, for a stack of two 4-element 20-meter Yagis at 120 and 60 feet, together with the response for a single Yagi at 120 feet and a 120/60-foot stack over flat ground for reference. The response due to many diffraction and reflection components is quite complicated!

You can generate a terrain file manually using a topographic map and a ruler or a pair of dividers. The HFTA.PDF file (accessed by clicking on the **Help** button) and on the accompanying CD-ROM gives complete instructions on how to create a terrain file manually (or automatically). The manual process is simple enough in concept. Mark on your US Geological Survey 7.5-minute map the exact location of your tower. You will find 7.5-minute maps available from some local sources, such as large hardware stores, but the main contact point is the U.S. Geological Survey, Denver, CO 80225 or Reston, VA 22092. Call 1-800-MAPS-USA. Ask for the folder describing the topographic maps available for your geographic area. Many countries outside the USA have topographic charts also. Most are calibrated in meters. To use these with *HFTA*, you will have to convert meters to feet by multiplying meters by 3.28 or else inserting a single line at the very beginning of the disk file, saying "meters" for *HFTA* to recognize meters automatically.

Mark off a pencil line from the tower base, in the azimuthal direction of interest, perhaps 45° from New England to Europe, or 335° to Japan. Then measure the distance from the tower base to each height contour crossed by the pencil line. Enter the data at each distance/height into an ASCII computer file, whose filename extension is "PRO," standing for *profile*.

Fig 46 shows a portion of the USGS paper map for the N6BV QTH in Windham, NH, along with lines scribed in several directions towards various parts of Europe and the Far East. Note that the elevation heights of the intermediate contour lines are labeled manually in pencil in order to make sense of things. It is very easy to get confused unless you do this!

The terrain model used by *HFTA* assumes that the terrain is represented by flat *plates* connecting the elevation points in the *.PRO file with straight lines. The model is two dimensional, meaning that range and elevation are the only data for a particular azimuth. In effect, *HFTA* assumes that the width of a terrain plate is wide relative to its length. Obviously, the world is three-dimensional. If your shot in a particular direction involves aiming your Yagi down a canyon with steep walls, then it's pretty likely that your actual elevation pattern will be different than what *HFTA* tells you. The signals must careen horizontally from wall to wall, in addition to being affected by the height changes of the terrain. *HFTA* isn't designed to do canyons.

To get a true 3-D picture of the full effects of terrain, a terrain model would have to show azimuth, along with range and elevation, point-by-point for about two miles in every direction around the base of the tower. After you go through the pain of creating a profile for a single azimuth, you'll appreciate the immensity of the process if you were you try to create a full 360° 3D profile manually.

Terrain Data from the Internet

At one time digitized terrain data commonly available from the Internet didn't have sufficient resolution to be accurate enough for *HFTA*. Nowadays, the complete, accurate set of USGS topographic 7.5-minute maps are available at no cost on the Internet. You can use a program called *MicroDEM*, written by Professor Peter Guth at the US Naval Academy, to quickly and easily produce terrain data files suitable for *HFTA* from topographic data files. Dr Guth and the US Naval Academy have graciously allowed ARRL to include the *MicroDEM* program on the CD-ROM accompanying this book. It should be noted that besides automatically creating terrain profiles for *HFTA*, *MicroDEM* is a full-featured mapping program on its own.

Instructions for using *MicroDEM* are in the Help file for *HFTA* (HFTA.PDF), which you can access from the *HFTA* main window by clicking on the **Help** button. **Fig 47** shows a screen capture of the *MicroDEM* program for the N6BV/1 location in New Hampshire for an azimuth of 45° into Europe. The black/white rendering of the screen capture doesn't do justice to the same information in color. The computed terrain profile is plotted in the window at the right of Fig 47 and the data file is shown in the inset window at the top right.

Using *MicroDEM* and on-line USGS topographic map data, you can also automatically create up to 360 terrain profiles with 1° spacing of azimuths in a few seconds. (Specifying a 1° spacing is really overkill; most operator choose to create 72 profiles with 5° spacing.) On a topographic DEM (digital elevation model) map that covers the geographic

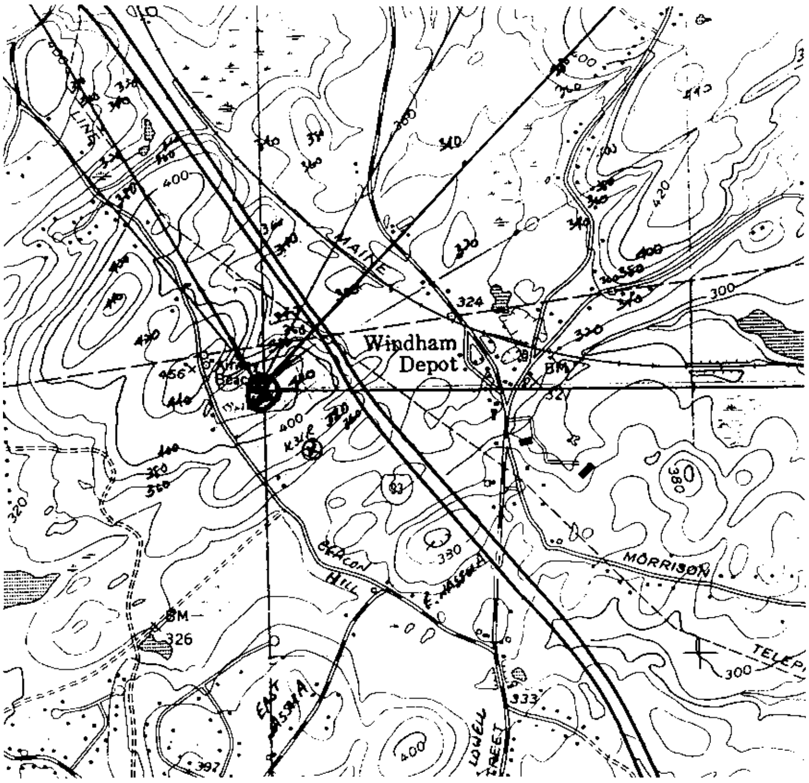

Fig 46—A portion of USGS 7.5 minute topographic map, showing N6BV/1 QTH, together with marks in direction of Europe and Japan from tower base. Note that the elevations contours were marked by hand to help eliminate confusion. This required a magnifying glass and a steady hand!

area of interest, you simply specify the latitude and longitude of a tower's location—found using a GPS receiver—and then ask *MicroDEM* for a **Viewshed**. See the *HFTA* **Help** file for the details.

Compare this automated several-second *MicroDEM* process to creating manual profiles on a paper topographic map—It can take up to an hour of meticulous measurements to manually create a single terrain profile!

Algorithm for Ray-Tracing the Terrain

Once a terrain profile is created, there are a number of mechanisms that *HFTA* takes into account as a ray travels over that terrain:

1. Classical ray reflection, with Fresnel ground coefficients.
2. Direct diffraction, where a diffraction point is illuminated directly by an antenna, with no intervening terrain features blocking the direct illumination.
3. When a diffracted ray is subsequently reflected off the terrain.
4. When a reflected ray encounters a diffraction point and causes another series of diffracted rays to be generated.
5. When a diffracted ray hits another diffraction point, generating another whole series of diffractions.

Certain unusual, bowl-shaped terrain profiles, with sheer vertical faces, can conceivably cause signals to reflect or diffract in a backwards direction, only to be reflected back again in the forward direction by the sheer-walled terrain to the rear. *HFTA* does not accommodate these interactions, mainly because to do so would increase the computation time too much. It only evaluates terrain in the forward direction along one azimuth of interest.

Fig 48 shows a portion of an *HFTA* screen capture in the direction towards Europe from the N6BV/1 location in New Hampshire on 21.2 MHz. It compares the results for a 90/60/30-foot stack of TH7DX tribanders to the same stack over flat land, and to a single antenna at 70 feet over flat ground. The 70-foot single antenna represents a pretty typical station on 15 meters. The terrain produces excellent gain at lower elevation angles compared to the same stack over flat ground. The stack is very close to or superior to the single 70-foot high Yagi at all useful elevation angles. Terrain can indeed exhibit a profound effect on the launch of signals into the ionosphere—for good or for bad.

HFTA's Internal Antenna Model

The operator selects the antenna used inside *HFTA* to be anything from a dipole to an 8-element Yagi. The default assumes a simple cosine-squared mathematic response, equivalent to a 4-element Yagi in free space. *HFTA* traces rays only in the forward direction from the tower along the azimuth of interest. This keeps the algorithms reasonably simple and saves computing time.

HFTA considers each antenna in a stack as a separate *point source*. The simulation begins to fall apart if a traveling wave type of antenna like a rhombic is used, particularly if the terrain changes under the antenna—that is, the ground is not flat under the entire antenna. For a typical Yagi, even a long-boom one, the point-source assumption is reasonable. The internal antenna model also assumes that the Yagi is horizontally polarized. *HFTA* does not do vertically polarized antennas, as discussed previously. The documentation for *HFTA* also cautions the user to work with practical spacings between stacked Yagis—0.5 λ or more

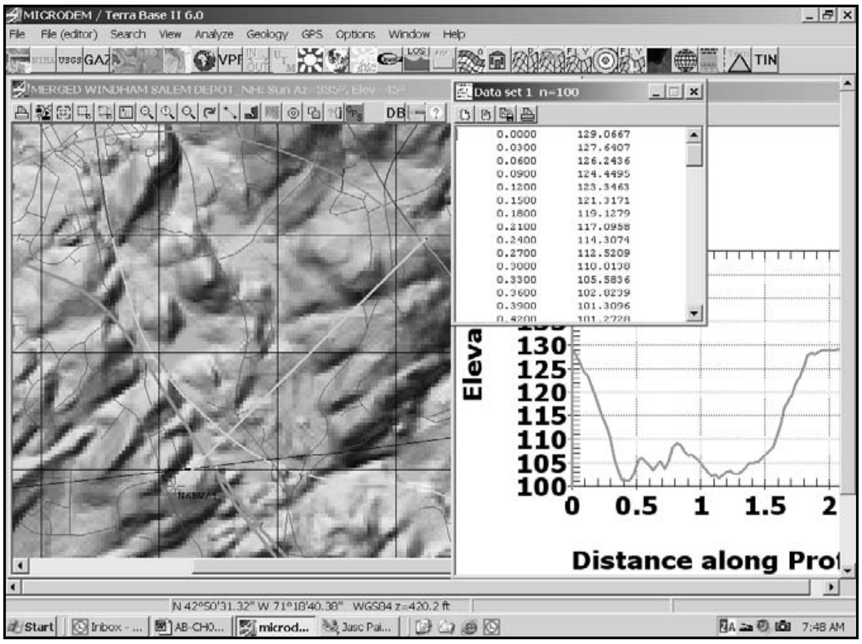

```
        0.0000        129.0667
        0.0300        127.6407
        0.0600        126.2436
        0.0900        124.4495
        0.1200        123.3463
        0.1500        121.3171
        0.1800        119.1279
        0.2100        117.0958
        0.2400        114.3074
        0.2700        112.5209
        0.3000        110.0138
        0.3300        105.5836
        0.3600        102.0239
        0.3900        101.3096
        0.4200        101.2728
```

N 42°50'31.32" W 71°18'40.38" WGS84 z=420.2 ft

Fig 47—A screen-capture of the *MicroDEM* program, showing the topographic map for the same terrain shown in Fig 46, together with the computed terrain profile along an azimuth of 45° on the path towards Europe from the N6BV/1 location in Windham, NH.

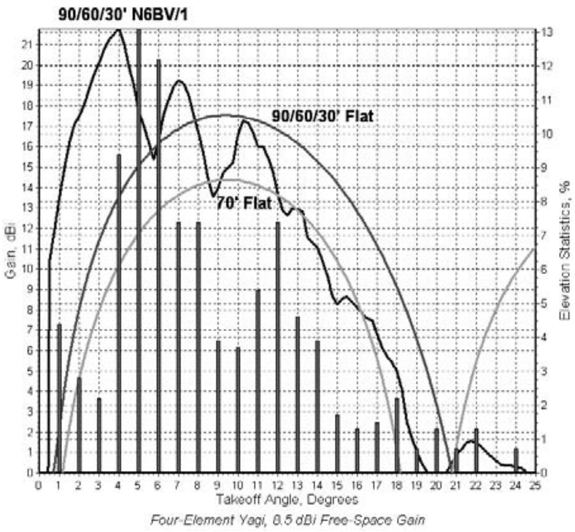

Fig 48—The 21-MHz elevation response for a stack of three TH7DX Yagis mounted on a single tower at 90/60/30 feet, at the N6BV/1 QTH for a 45° azimuth towards Europe. The terrain focuses the energy at low elevation angles compared to the same stack over flat ground. This illustrates once again the conservation of energy—Energy squeezed down into low elevation angles is stolen from other, higher, angles.

because *HFTA* doesn't explicitly model mutual coupling between Yagis in a stack.

HFTA compares well with the measurements for the horizontal antennas described earlier by Jim Breakall, WA3FET, using a helicopter in Utah. Breakall's measurements were done with a 15-foot high horizontal dipole.

More Details About *HFTA*

Frequency Coverage

HFTA can be used on frequencies higher than the HF bands, although the graphical resolution is only 0.25°. The patterns above about 100 MHz thus look rather grainy. The UTD is a *high-frequency-asymptotic* solution, so in theory the results become more realistic as the frequency is raised. Keep in mind too that *HFTA* is designed to model launch angles for skywave propagation modes, including E- and F-layer, and even Sporadic-E. Since by definition the ionospheric launch angles include only those above the horizon, direct line-of-sight UHF modes involving negative launch angles are not considered in *HFTA*.

See the HFTA.PDF documentation file for further details on the operation of *HFTA*. This file, as well as sample terrain profiles for some *big-gun* stations, is located on the CD-ROM accompanying this book.

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references listed below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

B. Boothe, "The Minooka Special," *QST*, Dec 1974.

Brown, Lewis and Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE* , Jun 1937.

G. Brown, "The Phase And Magnitude Of Earth Currents Near Radio Transmitting Antennas," *Proc. IRE*, Feb 1935, pp 168-182.

R. Collin and F. Zucker, *Antenna Theory*, Chap 23 by J. Wait, Inter-University Electronics Series (New York: McGraw-Hill, 1969), Vol 7, pp 414-424.

T. Hullick, W9QQ, "A Two-Element Vertical Parasitic Array For 75 Meters," *QST*, Dec 1995, pp 38-41.

R. Jones, "A 7-MHz Vertical Parasitic Array," *QST*, Nov 1973.

T. Larsen, "The E-Field and H-Field Losses Around Antennas With a Radial Ground Wire System," *Journal of Research of the National Bureau of Standards*, D. Radio Propagation, Vol 66D, No. 2, Mar-Apr 1962, pp 189-204.

D. A. McNamara, C. W. I. Pistorius, J. A. G. Malherbe, *Introduction to the Geometrical Theory of Diffraction* (Norwood, MA: Artech House, 1994).

C. J. Michaels, "Some Reflections on Vertical Antennas," *QST*, Jul 1987.

C. J. Michaels, "Horizontal Antennas and the Compound Reflection Coefficient," *The ARRL Antenna Compendium, Vol 3* (Newington: ARRL, 1992).

R. Severns, N6LF, "Verticals, Ground Systems and Some History," *QST*, Jul 2000, pp 38-44.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971.

J. Sevick, "The W2FMI 20-Meter Vertical Beam," *QST*, Jun 1972.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973.

J. Sevick, "A High Performance 20-, 40- and 80-Meter Vertical System," *QST*, Dec 1973.

J. Sevick, "The Constant-Impedance Trap Vertical," *QST*, Mar 1974.

J. Stanley, "Optimum Ground Systems for Vertical Antennas," *QST*, Dec 1976.

F. E. Terman, *Radio Engineers' Handbook*, 1st ed. (New York, London: McGraw-Hill Book Co, 1943).

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

The last major study that appeared in the Amateur literature on the subject of local terrain as it affects DX appeared in four *QST* "How's DX" columns, by Clarke Greene, K1JX, from October 1980 to January 1981. Greene's work was an update of a landmark Sep 1966 *QST* article entitled "Station Design for DX," by Paul Rockwell, W3AFM. The long-range profiles of several prominent, indeed legendary, stations in Rockwell's article are fascinating: W3CRA, W4KFC and W6AM.

Chapter 4

Antenna Modeling & System Planning

OVERVIEW: ANTENNA ANALYSIS BY COMPUTER

As pointed out in Chapter 3, The Effects of Ground, irregular local terrain can have a profound effect on the launch of HF signals into the ionosphere. A *system approach* is needed to create a scientifically planned station. We pointed out in Chapter 3 that antenna modeling programs do not generally take into account the effects of irregular terrain, and by "irregular" we mean any sort of ground that is not flat. Most modeling programs based on *NEC-2* or *MININEC* do model reflections, but they do not model diffractions.

On the other hand, while a ray-tracing program like *HFTA* (HF Terrain Assessment) does take into account diffraction, it doesn't explicitly factor in the mutual impedance between an antenna and the ground. Instead, *HFTA* makes the basic assumption that the antenna is mounted sufficiently high above ground so that the mutual impedance between an antenna and the ground is minimal.

In this chapter we'll look at modeling the antennas themselves on the PC. We'll evaluate some typical antennas over flat ground and also in free space. Once characterized—or even optimized for certain characteristics—these antennas can then be analyzed over real terrain using *HFTA* and the other tools discussed in Chapter 3.

A Short History of Antenna Modeling

With the proliferation of personal computers since the early 1980s, amateurs and professionals alike have made significant strides in computerized antenna system analysis. It is now possible for the amateur with a relatively inexpensive computer to evaluate even complicated antenna systems. Amateurs can obtain a keener grasp of the operation of antenna systems—a subject that has been a great mystery to many in the past. We might add that modern computing tools allow hams to debunk overblown claims made about certain antennas.

The most commonly encountered programs for antenna analysis are those derived from a program developed at US government laboratories called *NEC*, short for "Numerical Electromagnetics Code." *NEC* uses a so-called *Method of Moments* algorithm. This intriguing name derives from a mathematical convention dealing with how "momentous" the accumulated error becomes when certain simplifying assumptions are made about the current distribution along an antenna wire. If you want to delve into details about the method of moments, John Kraus, W8JK, has an excellent chapter in his book *Antennas*, 2nd edition. See also the article "Programs for Antenna Analysis by the Method of Moments," by Bob Haviland, W4MB, in *The ARRL Antenna Compendium, Vol 4*.

The mathematics behind the method-of-moments algorithm are pretty formidable, but the basic principle is simple. An antenna is broken down into a number of straight-line wire *segments*, and the field resulting from the RF current in each segment is evaluated by itself and also with respect to other mutually coupled segments. Finally, the field from each contributing segment is vector-summed to yield the total field, which can be computed for any elevation or azimuth angle desired. The effects of flat-earth ground reflections, including the effect of ground conductivity and dielectric constant, may be evaluated as well.

In the early 1980s, *MININEC* was written in BASIC for use on personal computers. Because of limitations in memory and speed typical of personal computers of the time, several simplifying assumptions were necessary in *MININEC* and these limited potential accuracy. Perhaps the most significant limitation was that perfect ground was assumed to be directly under the antenna, even though the radiation pattern in the far field did take into account real ground parameters. This meant that antennas modeled closer than approximately 0.2 λ over ground sometimes gave erroneous impedances and inflated gains, especially for horizontal polarization. Despite some limitations, *MININEC* represented a remarkable leap forward in analytical capability. See Roy Lewallen's (W7EL) "*MININEC*—the Other Edge of the Sword" in Feb 1991 *QST* for an excellent treatment on pitfalls when using *MININEC*.

Because source code was made available when *MININEC* was released to the public, a number of programmers produced some very capable commercial versions for the amateur market, many incorporating exciting graphics showing antenna patterns in 2D or 3D. These programs also simplify the creation of models for popular antenna types, and several come with libraries of sample antennas.

By the end of the 1980s, the speed and capabilities

Commercial Implementations of *MININEC* and *NEC-2* Programs

Ever since the source code for *NEC-2* and *MININEC* came into in the public domain, enterprising programmers have been upgrading, extending and improving these programs. There are a number of "freeware" versions available nowadays, and there are also a variety of commercial implementations.

This sidebar deals only with the most popular commercial versions, programs that many hams use. You should keep in mind that whatever program you choose will require an investment in learning time, if not in dollars. Your time is valuable, of course, and so is the ability to swap modeling files you create with other modelers. Other peoples' modeling files, particularly when you are just starting out, are a great way to learn how the "experts" do their modeling. For example, there are archives of *EZNEC/ELNEC* files available on the Internet, since this popular modeling program has been around for a number of years. (*ELNEC* is the DOS-only, *MININEC*-core predecessor of EZNEC.)

The following table summarizes the main features and the pricing as of 2006 for some popular commercial antenna modeling programs. The programs that use the *NEC-4* core require separate licenses from Lawrence Livermore National Laboratories.

Commercial Implementations of *MININEC* and *NEC 2* programs

Name	***EZNEC 4.0* *(+ ver.)***	***EZNEC-M Pro***	***EZNEC/4***	***NEC-Win Plus +***	***NEC-Win Pro***	***GNEC***	***Antenna Model***
Manufacturer	Roy Lewallen	Roy Lewallen	Roy Lewallen	Nittany Scientific	Nittany Scientific	Nittany Scientific	Teri Software
Core	*NEC-2*	*NEC-2*	*NEC-4*	*NEC-2*	*NEC-2*	*NEC-2/NEC-4*	*MININEC*
Operating System	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit	Windows 32-bit
Number Segments	500 (1500, + ver.)	20,000	20,000	10,000	10,000	80,000	Limited by memory
NEC-Card Inputs	No	Yes	Yes	Yes	Yes	Yes	No
Other Input	ASCII (NEC, + ver.)	ASCII, NEC	ASCII, NEC	CAD *.DXF	CAD *.DXF	CAD *.DXF	No
Wires by Equation	No	No	No	Yes	Yes	Yes	Yes
Source Setting	By %	By %	By %	By %	By %	By %	By %
Source Type	Current/ Voltage/Split	Current/ Voltage/Split	Current /Voltage/Split	Current/ Voltage/Split	All types	All types	Current/Voltage
R + *j* X Loads	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RLC Loads	Series, Parallel, Trap	Series, Parallel, Trap	Series, Parallel, Trap	Series, Parallel	Series, Parallel	Series, Parallel	Series, Parallel
True Trap Loads	Yes	Yes	Yes	No	No	No	No
Laplace Loads	Yes	Yes	Yes	Yes	Yes	Yes	No
Conductivity Table	Yes *	Yes *	Yes *	Yes	Yes	Yes	Yes
Average Gain Test	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Transmission Lines	Yes	Yes	Yes	Yes	Yes	Yes	No
View Geometry	Excellent	Excellent	Excellent	Good	Good	Good	Very Good
Geometry Checking	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Easy Height Change	Yes	Yes	Yes	No	No	No	No
Polar Plots	ARRL, linear-dB Az/El, Circ. (+ ver.)	ARRL, linear-dB Az/El, Circ.	ARRL, linear-dB Az/El, Circ.	ARRL, linear-dB Az/El Patterns	ARRL, linear-dB Az/El Patterns	ARRL, linear-dB Az/El Patterns	ARRL, Linear -dB Az/El Patterns
Rectangular Plots	SWR	SWR	SWR	SWR, Zin	SWR, Zin, Az/El, Near/Far Plots, Currents	SWR, Zin, Az/El, Near/Far Plots, Currents	Gain, SWR, F/B, F/R, Rin, Xin
Operating Speed	Fast	Fast	Fast	Very Fast	Very Fast	Very Fast	Slow
Smith Chart	Data for Ext. Smith program	Data for Ext. Smith program	Data for Ext. Smithprogram	No	Yes	Yes	Yes
Near/Far Field Tables	Both	Both	Both	Far	Both	Both	Both
Ground Wave Analysis	No	Yes	Yes	No	Yes	Yes	No
Pricing	$89 Web; $99 CD-ROM, $139 (+ ver.)	$450	$600; must have *NEC-4* license	$150	$425	$795	$85

* Wire conductivity is the same for all wires.

Excellent, Very Good, Good ratings done by *Antenna Book* editor.

of personal computers had advanced to the point where PC versions of *NEC* became practical, and several versions are now available to amateurs. The most recent public-domain version is *NEC-2* and this is the computational core that we'll use as an example throughout this chapter.

Like *MININEC*, *NEC-2* is a general-purpose modeling package and it can be difficult to use and relatively slow in operation for certain specialized antenna forms. Thus, custom commercial software has been created for more user-friendly and speedier analysis of specific antenna varieties, mainly Yagi arrays. See Chapter 11, HF Yagi Arrays. Also see the sidebar, "Commercial Implementations of *MININEC* and *NEC-2* Programs."

For this edition of *The ARRL Antenna Book*, Roy Lewallen, W7EL, has graciously provided a special version of his *EZNEC 3.0* program, called *EZNEC-ARRL*. This version works with the specific antenna models also bundled on the CD-ROM. Please note that this ARRL-specific version of *EZNEC* is limited to a maximum of 20 segments (we'll explain segments later) for all models except for the special ones included on this CD-ROM. You can find information on how to purchase the full-fledged version of *EZNEC* in the **Help** section of the *EZNEC-ARRL* program.

The following material on antenna modeling is by necessity a summary, since entire books have been written on this subject. Serious modelers would be well-advised to enroll in the online Antenna Modeling course, part of the ARRL Certification and Continuing Education series. L. B. Cebik, W4RNL created the ARRL Antenna Modeling course and it contains a great deal of information, tips and techniques concerning modeling by computer. See: **http://www.arrl.org/news/stories/2002/02/06/2/** for more information. We also strongly recommend that you read the Help files in *EZNEC-ARRL*. There is a wealth of practical information on the finer points of antenna modeling there.

THE BASICS OF ANTENNA MODELING

This chapter will discuss the following antenna-modeling topics for an *NEC-2*-based modeling software, using *EZNEC-ARRL* as an example:

- Program outputs
- Wire geometry
- Segmentation, warnings and limitations
- Source (feed point) placement
- Environment, including ground types and frequency
- Loads and transmission lines
- Testing the adequacy of a model

PROGRAM OUTPUTS

Instruction manuals for software programs traditionally start out describing in detail the input data needed by the program. They then demonstrate the output data the program can generate. We feel it is instructive, however, to turn things around and start out with a brief overview of the output from a typical antenna-modeling program.

We'll look at the output from public-domain *NEC-2*. Next, we'll look at the output information available from commercial adaptations of *NEC-2*, using *EZNEC-ARRL* provided by W7EL. After this brief overview of the output data, we'll look in detail at the input data needed to make a modeling program work. In the following discussions it will be very instructive if you to bring up *EZNEC-ARRL* on your computer and open the specific modeling files used in each example. [From now on in this chapter we'll refer merely to *EZNEC* rather than *EZNEC 3.0*, the official name or *EZNEC-ARRL*, a specialized subset of *EZNEC 3.0*. Where there are specific differences between *EZNEC 3.0* and the limited-edition *EZNEC-ARRL* we'll identify them.]

Native *NEC-2*

The native *NEC-2* program produces pages and pages of output formatted for a mainframe "line printer." You may be old enough to remember the stacks of green-and-white, tractor-feed, 132-column computer paper that such a line printer produced. Corporate MIS departments stored untold number of boxes of that paper.

Native *NEC-2* was written in the Fortran language, which stands for *Formula Translation*. Programmers used punched cards to enter the program itself and its accompanying input data into huge mainframe computers. To say that the paper output from *NEC-2* is massive, even intimidating, is putting it mildly. There is a strong distinction between "useful information" and "raw data" and the raw output from native *NEC-2* bombards the user with raw data.

Commercial versions that use the *NEC-2* computational core shield the user from the ugliness of raw line-printer output, as well as punched-card input (or disk surrogates for punched cards). Commercial versions like *EZNEC* do produce output numerical tables where this is useful. These tables show parameters such as the source impedance and SWR at a single frequency, or the characteristics of a load or a transmission line. But as the old saying goes, "One picture is worth a thousand words." This is as true for modeling programs as it is for other endeavors dealing with reams of numbers. Thus, most commercial modeling software packages create *graphs* for the user. *EZNEC* produces the following types of graphs:

- Polar (linear-dB or ARRL-style) graphs of the far-field elevation and azimuth responses.
- 3-D wire-frame graph of the total far-field response.
- Graph of the SWR across a frequency band.
- Graphical display of the RF currents on various conductors in a model.
- Rotatable, zoomable 3-D views of the wires used to make a model.
- Output to programs capable of generating Smith charts.

Fig 1A and 1B shows the computed far-field 2-D elevation and azimuth patterns for a 135-foot long horizontal dipole, mounted in a flattop configuration 50 feet above flat ground. These figures were generated using *EZNEC* at 3.75 MHz. Fig 1C shows a 3-D wire-frame picture of the far-field response, but this time at 14.2 MHz. For comparison, **Table 1** shows a short portion of the line-printer output for the azimuth pattern at 3.75 MHz. The actual printout is many pages long. One picture can indeed replace thousands of numbers!

Fig 2 shows the computed SWR curve over the frequency range 3.0 to 4.0 MHz for this dipole, fed with lossless 50-Ω transmission line. *EZNEC* generated this plot using the "SWR" button. Figs 1 and 2 are typical of the kind of graphical outputs that commercial implementations of the *NEC-2* computing core can produce.

Now, let's get into the details of what kind of input data is required to run a typical method-of-moments antenna-modeling program.

PROGRAM INPUTS: WIRE GEOMETRY

Coordinates in an X, Y and Z World

The most difficult part of using a *NEC*-type of modeling program is setting up the antenna's geometry—you must condition yourself to think in three-dimensional, Cartesian coordinates. Each end point of a wire is represented by three numbers: an x, y and z coordinate. These coordinates represent the distance from the origin (x-axis), the width of an antenna (y-axis), and the height (z-axis).

An example should help sort things out. **Fig 3** shows a simple model of a 135-foot center-fed dipole, made of #14 copper wire placed 50 feet above flat ground. The common term for this antenna is *flattop dipole*. For convenience, the ground is located at the *origin* of the coordinate system, at (0, 0, 0) feet, directly under the center of the dipole. **Fig 4** shows the *EZNEC* spreadsheet-like input data for this antenna. (Use model file: **Ch4-Flattop Dipole.EZ**.) *EZNEC* allows you to specify the type of conductor material from its main window, using the **Wire Loss** button to open a new window. We will click on the **Copper** button for this dipole.

Above the origin, at a height of 50 feet on the z-axis, is the dipole's *feed point*, called a *source* in *NEC* terminology. The width of the dipole goes toward the left (that is, in the "negative-y" direction) one-half the overall length of 135 feet, or –67.5 feet. Toward the right, our dipole's other end is at +67.5 feet. The x-axis dimension of our dipole is zero, meaning that the dipole wire is parallel to and directly above the x-axis. The dipole's ends are thus represented by two points, whose coordinates are (0, –67.5, 50) and (0, 67.5, 50) feet. The use of parentheses with a sequential listing of (x, y, z) coordinates is a common practice among antenna modelers to describe a wire end point.

Fig 3B includes some other useful information about this antenna beyond the wire geometry. Fig 3B overlays the wire geometry, the current distribution along the wire and the far-field azimuth response, in this case at an elevation angle of 30°.

Although not shown specifically in Fig 3, the thickness of the antenna is the diameter of the wire, #14 gauge. Note that native *NEC* programs specify the *radius* of the

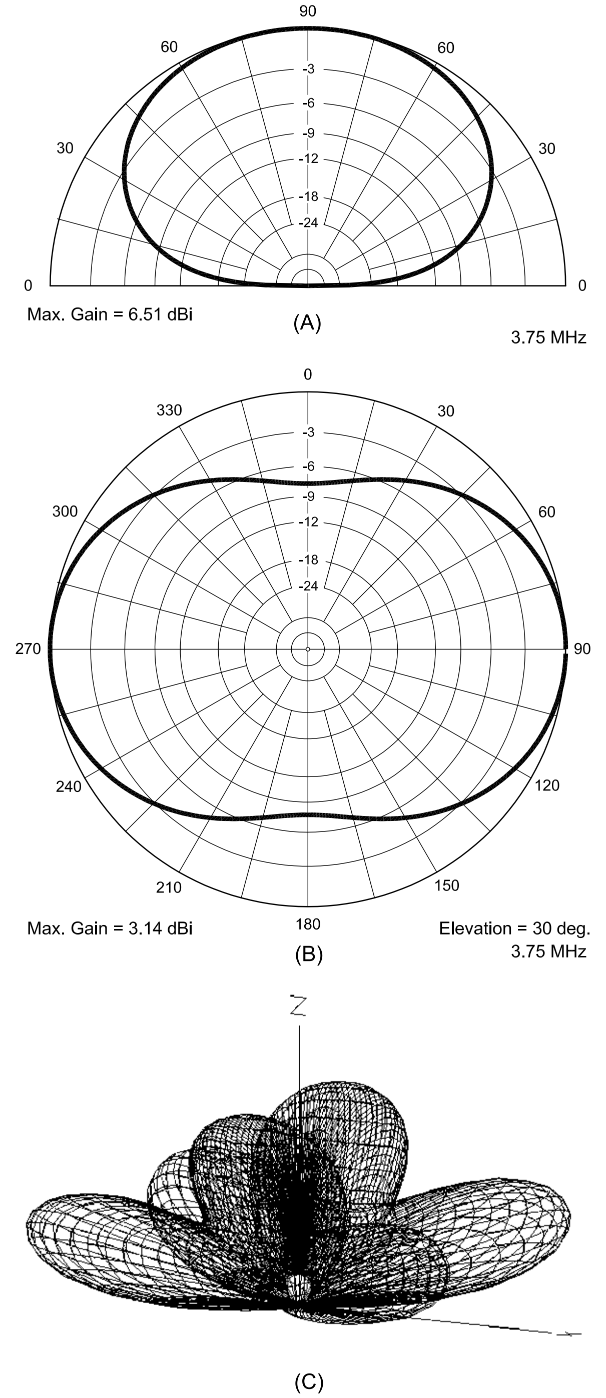

Fig 1—At A, far-field elevation-plane pattern for a 135-foot-long horizontal dipole, 50 feet above flat ground, at 3.5 MHz. At B, the far-field azimuth-plane pattern at an elevation angle of 30°.

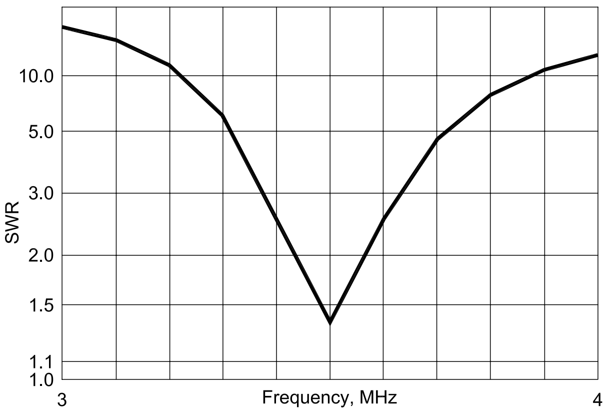

Fig 2—SWR curve for 135-foot flattop dipole over the frequency range 3.0 to 4.0 MHz for a 50-Ω feed line. This antenna is an example and is not optimized for the amateur band.

wire, rather than the diameter, but programs like *EZNEC* use the more intuitive diameter of a wire rather than the radius. *EZNEC* (and other commercial programs) also allows the user to specify the wire as an AWG gauge, such as #14 or #22, for example.

We've represented our simple dipole in Fig 3 using a single, straight wire. In fact, all antenna models created for method-of-moment programs are made of combinations of straight wires. This includes even complex antennas, such as helical antennas or round loops. (The mathematical basis for modeling complex antennas is that they can be simulated using straight-wire *polygons*. A circular loop, for example, can be modeled using an octagon.)

Segmentation and Specifying a Source Segment

We've specified the physical geometry of this simple one-wire dipole. Now several more modeling details surface—you must specify the number of *segments* into which the dipole is divided for the method-of-moment analysis and you must somehow feed the antenna. The *NEC-2* guideline for setting the number of segments is to use at least 10 segments per half-wavelength. This is a general rule of thumb, however, and in many models more dense segmentation is mandatory for good accuracy.

In Fig 3, we've specified that the dipole be divided into 11 segments for operation on the 80-meter band. This follows the rule of thumb above, since the 135-foot dipole is about a half-wavelength long at 3.5 MHz.

Setting the Source Segment

The use of 11 segments, an odd rather than an even number such as 10, places the dipole's feed point (the

Table 1
Portion of line-printer output from *NEC-2* for 135-foot dipole.

```
                                 - - - RADIATION PATTERNS - - -

 - - ANGLES - -     - POWER GAINS -    - - - POLARIZATION - - -      - - - E(THETA) - - -        - - - E(PHI) - - -

Theta    Phi        Vert    Hor   Total  Axial     Tilt     Sense    Magnitude       Phase     Magnitude       Phase
Degrees  Degrees    dB      dB    dB     Ratio     Degrees           Volts           Degrees   Volts           Degrees
60.00     0.00    –999.99   3.14  3.14  0.00000   90.00   LINEAR   0.00000E+00      0.00    6.62073E-01   –66.87
60.00     1.00     –37.87   3.13  3.14  0.00301   89.52   LEFT     5.89772E-03    –86.64    6.61933E-01   –66.87
60.00     2.00     –31.85   3.13  3.13  0.00603   89.04   LEFT     1.17915E-02    –86.64    6.61512E-01   –66.87
60.00     3.00     –28.33   3.12  3.12  0.00904   88.56   LEFT     1.76776E-02    –86.64    6.60812E-01   –66.87
60.00     4.00     –25.84   3.11  3.11  0.01206   88.08   LEFT     2.35520E-02    –86.64    6.59834E-01   –66.87
60.00     5.00     –23.91   3.09  3.10  0.01508   87.59   LEFT     2.94109E-02    –86.64    6.58577E-01   –66.87
60.00     6.00     –22.34   3.07  3.08  0.01810   87.11   LEFT     3.52504E-02    –86.64    6.57045E-01   –66.87
60.00     7.00     –21.01   3.05  3.06  0.02112   86.62   LEFT     4.10669E-02    –86.63    6.55237E-01   –66.87
60.00     8.00     –19.87   3.02  3.04  0.02415   86.14   LEFT     4.68565E-02    –86.63    6.53158E-01   –66.87
60.00     9.00     –18.86   2.99  3.02  0.02718   85.65   LEFT     5.26156E-02    –86.63    6.50808E-01   –66.87
60.00    10.00     –17.96   2.95  2.99  0.03022   85.15   LEFT     5.83405E-02    –86.63    6.48190E-01   –66.87
60.00    11.00     –17.15   2.91  2.96  0.03327   84.66   LEFT     6.40278E-02    –86.63    6.45308E-01   –66.86
60.00    12.00     –16.42   2.87  2.92  0.03631   84.16   LEFT     6.96739E-02    –86.63    6.42165E-01   –66.86
60.00    13.00     –15.75   2.83  2.89  0.03937   83.66   LEFT     7.52755E-02    –86.63    6.38764E-01   –66.86
60.00    14.00     –15.13   2.78  2.85  0.04243   83.16   LEFT     8.08291E-02    –86.63    6.35108E-01   –66.86
60.00    15.00     –14.56   2.72  2.80  0.04550   82.65   LEFT     8.63317E-02    –86.62    6.31203E-01   –66.86
60.00    16.00     –14.03   2.66  2.76  0.04858   82.14   LEFT     9.17800E-02    –86.62    6.27051E-01   –66.86
60.00    17.00     –13.53   2.60  2.71  0.05166   81.62   LEFT     9.71711E-02    –86.62    6.22657E-01   –66.85
60.00    18.00     –13.07   2.54  2.66  0.05475   81.10   LEFT     1.02502E-01    –86.62    6.18027E-01   –66.85
```

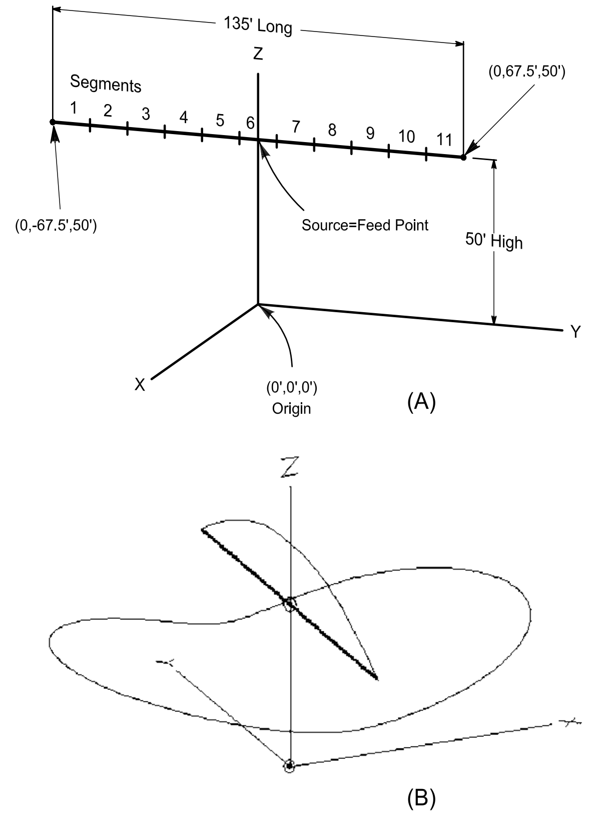

Fig 3—At A, simple model for a 135-foot long horizontal dipole, 50 feet above the ground. The dipole is over the y-axis. The wire has been segmented into 11 segments, with the center of segment number 6 as the feed point. The left-hand end of the antenna is -67.5 feet from the center feed point and that the right-hand end is at 67.5 feet from the center. At B, *EZNEC* "View Antenna" drawing, showing geometry of wire and the x, y and z axes. Overlaid on the wire geometry drawing are the current distribution along the wire and the far-field azimuthal response at an elevation angle of 30°.

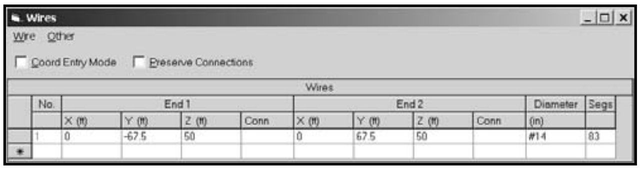

☐ Coord Entry Mode ☐ Preserve Connections

Wires

	No.	End 1				End 2				Diameter	Segs
		X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
	1	0	-67.5	50		0	67.5	50		#14	83
*											

Fig 4—*EZNEC* "Wires" spreadsheet for simple flattop dipole in Fig 3. The numbers shown are in feet, except for the wire diameter, which *EZNEC* allows you to specify as an AWG gauge, in this case #14. Note that 83 segments have been specified for this antenna for analysis over the range from 3.5 to 29.7 MHz.

source in *NEC*-speak, a word choice that can befuddle beginners) right at the antenna's center, at the center of segment number six. In concert with the "EZ" in its name, *EZNEC* makes choosing the source segment easy by allowing the user to specify a percentage along the wire, in this case 50% for center feeding.

At this point you may very well be wondering why no center insulator is shown in the middle of our center-fed dipole. After all, a real dipole would have a center insulator. However, method-of-moment programs assume that a source generator is placed across an infinitely small gap in the antenna wire. While this is convenient from a mathematical point of view, the unstated use of such an infinitely small gap often confuses newcomers to the world of antenna modeling. We'll get into more details, caveats and limitations in source placement later in this chapter. For now, just trust us that the model we've just described with 11 segments, fed at segment 6, will work well over the full amateur band from 3.5 to 4.0 MHz.

Now, let's consider what would happen if we want to use our 135-foot long dipole on all HF amateur bands from 3.5 to 29.7 MHz, rather than just from 3.5 to 4.0 MHz. Instead of feeding such an antenna with coax cable, we would feed it with open-wire line and use an antenna tuner in the shack to create a 50-Ω load for the transmitter. To comply with the segmentation rule above, the number of segments used in the model should vary with frequency—or at least be segmented at or above the minimum recommended level at the highest frequency used. This is because a half wavelength at 29.7 MHz is 16.6 feet, while a half wavelength at 3.5 MHz is 140.6 feet. So the number of segments for proper operation on 29.7 MHz should be 10 × 135/16.6 = 81. We'll be a little more conservative than the minimum requirement and specify 83 segments. **Fig 4** shows the *EZNEC* input spreadsheet for this model. (Use model file: **Ch4-Multiband Dipole.EZ**.)

The penalty for using more segments in a program like *NEC* is that the program slows down roughly as the square of the segments—double the number of segments and the speed drops to a fourth. If we try to use too few segments, we'll introduce inaccuracies, particularly in computing the feed-point impedance. We'll delve into this area of segmentation density in more detail later when we discuss testing the adequacy of a model.

Segment Length-to-Wire-Diameter Ratio

Even if you're willing to live with the slowdown in computing speed for situations involving a large number of wire segments, you should make sure the ratio between the segment length and the diameter of any wire is greater than 1:1. This is to say that the length of each segment is longer than the diameter of the wire. Doing so stays away from internal limitations in the *NEC* program.

For the #14 wire specified in this simple 135-foot long dipole, it's pretty unlikely that you'll bump up against this limitation for any reasonable level of segmentation. After all, #14 wire has a diameter of 0.064 inches and 135 feet is 1620 inches. To keep above a segment length of 0.064 inches, the maximum number of segments is 1620/

0.064 = 25,312. This is a *very* large number of segments and it would take a very long time to compute, assuming that your program can handle that many segments.

Keeping above a 1:1 ratio in segment length to wire diameter can be more challenging at VHF/UHF frequencies, however. This is particularly true for fairly large “wires” made of aluminum tubing. Incidentally, this is another point where newcomers to antenna modeling can be led astray by the terminology. In a *NEC*-type program, all conductors in a model are considered to be *wires*, even if they consist of hollow aluminum or copper tubes. Surface effect keeps the RF current in any conductor confined to the outer surface of that conductor, and thus it doesn’t matter whether the conductor is hollow or solid, or even made using a number of stranded wires twisted together.

Let’s look at a half-wave dipole at 420 MHz. This would be about 14.1 inches long. If you use $^1/_4$-inch diameter tubing for this dipole, the maximum segment length meeting the 1:1 diameter-to-length ratio requirement is also $^1/_4$ inches long. The maximum number of segments then would be 14.1/0.25 = 56.4, rounded down to 56. From this discussion you should now understand why method-of-moment programs are known for using a “thin-wire approximation.” Really fat conductors can get you into trouble, particularly at VHF/UHF.

Some Caveats and Limitations Concerning Geometry

Example: Inverted-V Dipole

Now, let’s get a little more complicated and specify another 135-foot-long dipole, but this time configured as an *inverted V*. As shown in **Fig 5**, you must now specify two wires. The two wires join at the top, at (0, 0, 50) feet. (Again, the program doesn’t use a center insulator in the model.)

If you are using a native version of *NEC*, you may have to go back to your high-school trigonometry book to figure out how to specify the end points of our “droopy” dipole, with its 120° included angle. Fig 5 shows the details, along with the trigonometric equations needed. *EZNEC* is indeed more “easy” here, since it allows you to tilt the ends of each wire downwards an appropriate number of degrees (in this case –30° at each end of the dipole) to automatically create an inverted-V configuration. **Fig 6A** shows the *EZNEC* spreadsheet describing this inverted-V dipole with a 120° included angle between the two wires.

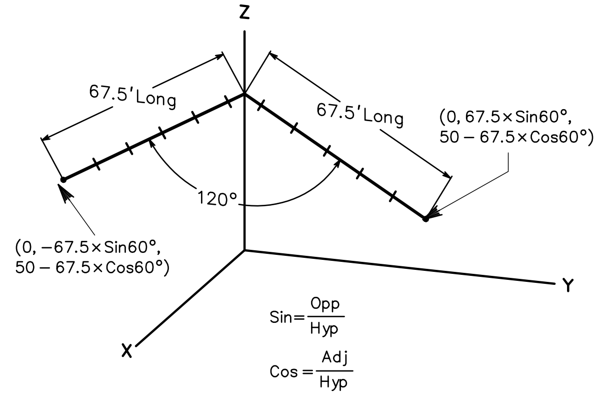

Fig 5—Model for an inverted-V dipole, with an included angle between the two legs of 120°. Sine and cosine functions are used to describe the heights of the end points for the sloping arms of the antenna.

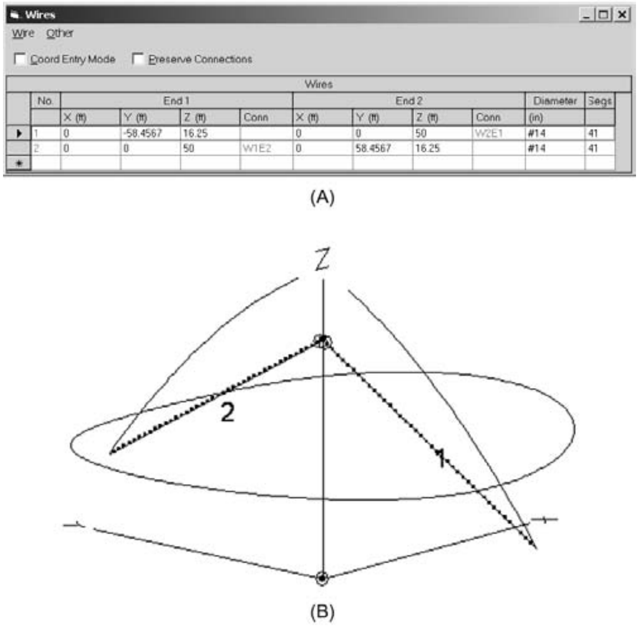

	No.	End 1				End 2				Diameter	Segs
		X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
▶	1	0	-58.4567	16.25		0	0	50	W2E1	#14	41
	2	0	0	50	W1E2	0	58.4567	16.25		#14	41
*											

Fig 6—At A, *EZNEC* spreadsheet for inverted-V dipole in Fig 5. Now the ends of the inverted-V dipole are 16.25 feet above ground, instead of 50 feet for the flattop dipole. At B, *EZNEC* “View Antenna” drawing, with overlay of geometry, current distribution and azimuth plot.

See the *EZNEC* **Help** section under “Wire Coordinate Shortcuts” for specific instructions on how to use the “elevation rotate end” shortcut “RE–30” to create the sloping wires easily by rotating the end of the wire down 30°. Now the specification of the source becomes a bit more complicated. The easiest way is to specify two sources, one on each end segment at the junction of the two wires. *EZNEC* does this automatically if you specify a so-called *split-source* feed. Fig 6B shows the two sources as two open circles at the top ends of the two wires making up the inverted-V dipole. What *EZNEC* is doing is creating two sources, each on the closest segments on either side of the junction of the two wires. *EZNEC* sums up the two source impedances to provide a single readout.

Navigating in the View Antenna Window

At this point it's worthwhile to explore some of the ways you can see what the wire geometry looks like using the *EZNEC* **View Ant** button on the main window. Bring up the file **Ch4-Inverted V Dipole.EZ** in *EZNEC*, and click on the **View Ant** button. You will see a small inverted-V dipole raised over the (0, 0, 0) origin on the ground directly under the feed point of the inverted-V dipole. First, "rotate" the dipole by holding down the left-mouse button and moving the mouse. You can orient the picture any way you wish.

Let's take a closer look at the junction of the two wires at the feed point. Click the **Center Ant Image** checkbox toward the bottom of the window to anchor the center of the image at the center of the window, and then move the **Zoom** slider upwards to zoom in on the image. At some point the junction of the two slanted wires will move up off the edge of the window, so you will need to click on the left-hand side of the **Z Move Image** slider to bring the junction back into view. Now you should be able to see a zoomed view of the junction, along with the two open circles that represent the location of the split sources in the middle of the segments adjacent to the wire junction.

Now put the mouse cursor over one of the slanted wires and double click the left-mouse button. *EZNEC* will now identify that wire and show its length, as well as the length of each segment on that wire. Pretty slick, isn't it?

Short, Fat Wires and the Acute-Angle Junction

Another possible complication can arise for wires with short, fat segments, particularly ones that have only a small included angle between them. These wire segments can end up inter-penetrating within each other's volumes, leading to problems in a model. Once you think of each wire segment as a thick cylinder, you can appreciate the difficulty in connecting two wires together at their ends. The two wires always inter-penetrate each other's volume to some extent. **Fig 7** depicts this problem graphically for two short, fat wires joined at their ends at an acute angle. A rule of thumb is to avoid creating junctions where more than $^1/_3$ of the wire volumes inter-penetrate. You can achieve this by using longer segment lengths or thinner wire diameters.

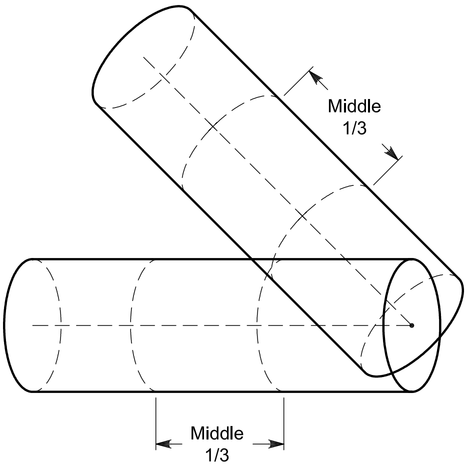

Fig 7—A junction of two short, fat wire segments at an acute angle. This results in inter-penetration of the two wire volumes beyond the middle-1/3 recommended limit.

Some Other Practical Antenna Geometries

A Vertical Half-Wave Dipole

If you turn the 135-foot-long horizontal dipole in Fig 1 on its end you will create a vertical half-wave dipole that is above the origin of the x, y and z axes. See **Fig 8**, where the bottom end of the dipole is placed 8 feet off the ground to keep it away from humans and animals for safety, at (0, 0, 8) feet. The top end is thus at 8 + 135 = 143 feet off the ground at (0, 0, 143). Fig 8 also shows the current distribution and the elevation pattern for this antenna. (Use *EZNEC* model file: **Ch4-Vertical Dipole.EZ**.)

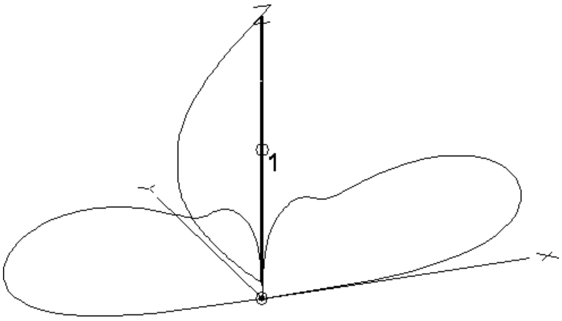

Fig 8—A vertical half-wave dipole, created by turning the dipole in Fig 3 on its end, with a minimum height at the lower end of 8 feet to keep the antenna away from people and animals. The current distribution and the elevation pattern for this antenna are also shown overlaid on the wire geometry.

A Ground-Plane Antenna

The ground-plane model is more complicated than previous ones because a total of five wires are now needed: one for the vertical radiator and four for the radials. **Fig 9** shows the *EZNEC* view for a 20-meter ground plane mounted 15 feet off the ground (perhaps on a garage roof), with the overlay of both the current distribution and the elevation-plane plot. (Use *EZNEC* model file: **Ch4-GP.EZ**.) Note that the source has been placed at the

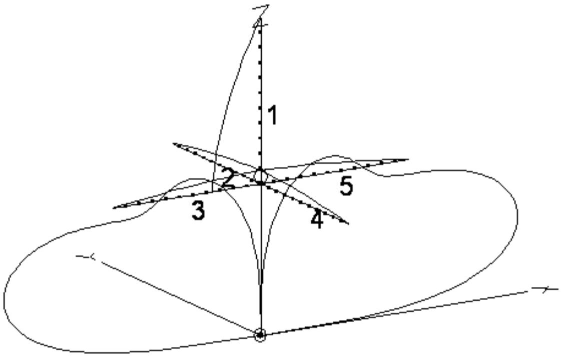

Fig 9—A vertical ground-plane antenna. The radials and the bottom of the vertical radiator are located 15 feet off the ground in this model. The current distribution along each wire and the far-field elevation-plane pattern are overlaid on the antenna geometry.

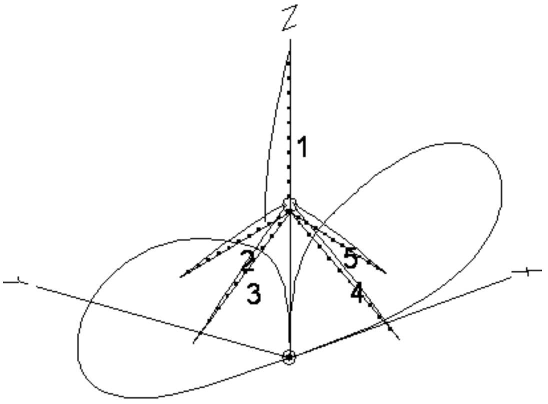

Fig 10—*EZNEC* View Antenna for the ground-plane antenna with its four radials tilted downwards by 40° to improve the SWR at the feed point.

bottom segment of the vertical radiator. Once again, the program needs no bottom insulator, since all five wires are connected together at a common point. *EZNEC* reports that this antenna has a resonant feed-point impedance of about 22 Ω, which would show an SWR of 2.3:1 for a 50-Ω coax feed line if no matching system is used, such as a gamma or hairpin match.

Fig 10 shows the same antenna, except that the radials have now been tilted downwards by 35° to facilitate an almost perfect 50-Ω match (SWR = 1.08:1). In addition, the length of the radiator in this model was shortened by 6 inches to re-resonate the antenna. (Use *EZNEC* model file: **Ch4-Modified GP.EZ**.) The trick of tilting the radials downwards for a ground-plane antenna is an old one, and the modeling programs validates what hams have been doing for years.

A 5-Element Horizontal Yagi

This is a little more challenging modeling exercise. Let's use a 5-element design on a 40-foot boom, but rather than using telescoping aluminum tubing for the elements, we'll use #14 wire. The *SCALE* program included with this book on the CD-ROM converted the aluminum-tubing 520-40.YW to a design using #14 copper wire. **Table 2** shows the element lineup for this antenna. (Later in this chapter we'll see what happens when telescoping aluminum tubing is used in a real-world Yagi design.)

Some explanations of what Table 3 means are in order. First, only one half of each element is shown. The *YW* program (Yagi for Windows), also included on the CD-ROM, computes the other half of the Yagi automatically, essentially mirroring the other half on the opposite side of the boom. Having to enter the dimensions for only half of a real-world Yagi element that uses telescoping aluminum tubing is much easier this way.

Second, the placement of the elements along the boom starts at 0.0 inches for the reflector. The distance between adjacent elements defined in this particular file is the spacing between the element itself and the element just before it. For example, the spacing between the driven element and the reflector is 72 inches, and the spacing between the first director and the driven element is also 72 inches. The spacing between the second director and the first director is 139 inches.

Fig 11A shows the wire geometry for this Yagi array when it is mounted 720 inches (60 feet) above flat ground and Fig 11B shows the *EZNEC* Wires spreadsheet that describes the coordinates. (Use *EZNEC* model file: **Ch4-520-40W.EZ**.) You can see that the x-axis coordinates for the elements have been automatically moved by the *SCALE* program so that the center of the boom is located directly above the origin. This makes it easier to evaluate the effects of stacking different monoband Yagis on a rotating mast in a "Christmas Tree" arrangement. A typical Christmas Tree stack might include 20, 15 and 10-meter monobanders on a single rotating mast sticking out of the top of the tower.

Fig 12 shows the computed azimuth pattern for this Yagi at 14.175 MHz, at an elevation angle of 15°, the angle where the peak of the forward lobe occurs at this

Table 2

520-40W.YW, using #14 wire from 520-40H.YW
14.000 14.174 14.350 MHz

5 elements,	inches
Spacing	.064
0.000	210.923
72.000	200.941
72.000	199.600
139.000	197.502
191.000	190.536

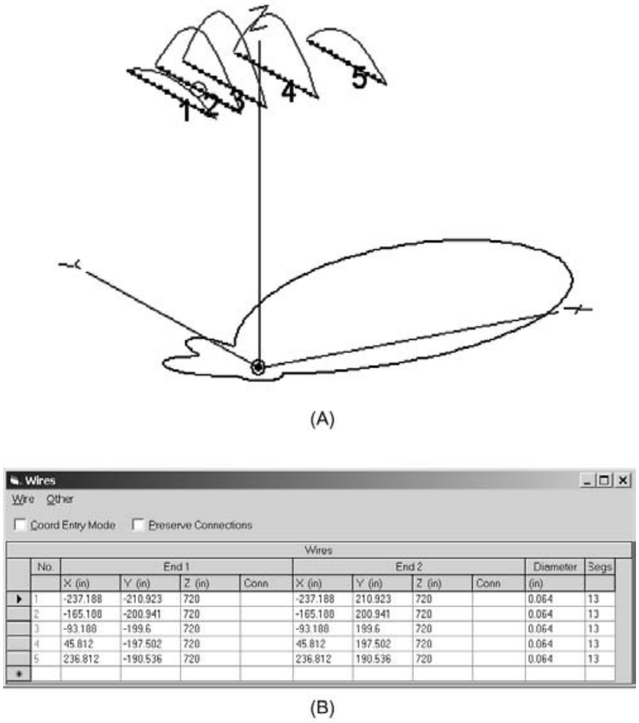

(A)

Wires

Wire Other

☐ Coord Entry Mode ☐ Preserve Connections

Wires

No.	End 1				End 2				Diameter	Segs
	X (in)	Y (in)	Z (in)	Conn	X (in)	Y (in)	Z (in)	Conn	(in)	
1	-237.188	-210.923	720		-237.188	210.923	720		0.064	13
2	-165.188	-200.941	720		-165.188	200.941	720		0.064	13
3	-93.188	-199.6	720		-93.188	199.6	720		0.064	13
4	45.812	-197.502	720		45.812	197.502	720		0.064	13
5	236.812	-190.536	720		236.812	190.536	720		0.064	13

(B)

Fig 11—At A, geometry for 5-element Yagi on a 40-foot boom, mounted 720 inches (60 feet) above flat ground, with an overlay of current and the azimuth pattern. At B, *EZNEC* Wires spreadsheet for this antenna. This design uses #14 wire for simplicity.

height above flat ground. The antenna exhibits excellent gain at 13.1 dBi, as well as a clean pattern behind the main lobe. The worst-case front-to-rear ratio at any point from 90° to 270° in azimuth is better than 23 dB. *EZNEC* says the feed-point impedance is 25 – *j* 23 Ω, just the right impedance suited for a simple hairpin or gamma match.

A Monoband 2-Element Cubical Quad

Unlike a Yagi, with its elements existing only in the x-y plane, a quad type of beam is a three-dimensional sort of antenna. A quad loop has height in the z-axis, as well as width and length in the x-y plane. Each individual loop for a monoband quad consists of four wires, joined together at the corners. **Fig 13** shows the coordinates for a 2-element 15-meter quad, consisting of a reflector and a driven element on a 10-foot boom.

You can see that the axis of symmetry, the x-axis, runs down the center of this model, meaning that the origin of this particular x, y and z-coordinate scheme is in the center of the reflector. The (0, 0, 0) origin is placed this way for convenience in assigning corner coordinates for each element. For actual placement of the antenna at a particular height above real ground, the heights of all z-axis coordinates are changed accordingly. *EZNEC* has a convenient built-in function to change the height of all wires at a single stroke.

Fig 14 shows the input *EZNEC* spreadsheet for this quad in free space, clearly showing the symmetrical nature of the corner coordinates. (Use *EZNEC* model file: **Ch4-Quad.EZ**.) This is a good place to emphasize that you should enter the wire coordinates in a logical

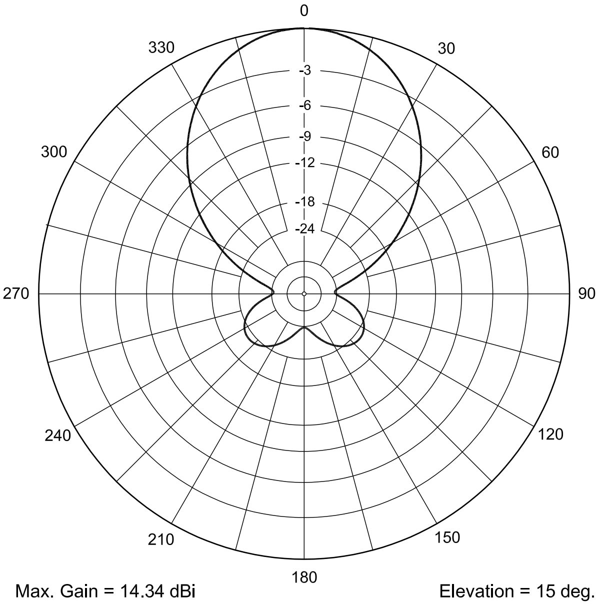

Fig 12—*EZNEC* azimuth-plane pattern at an elevation angle of 15° for #14 wire Yagi described in Fig 11.

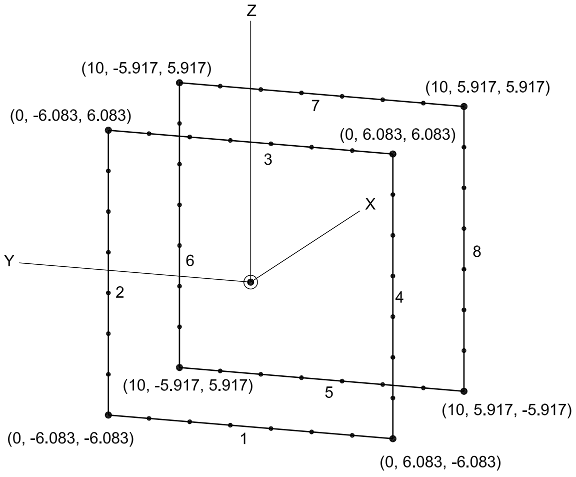

Fig 13—Wire geometry for a 2-element cubical quad, with a reflector and driven element. The x-axis is the axis of symmetry for this free-space model.

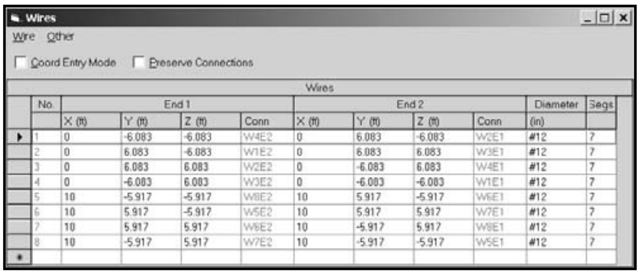

Wires										
No.	End 1				End 2				Diameter	Segs
	X (ft)	Y (ft)	Z (ft)	Conn	X (ft)	Y (ft)	Z (ft)	Conn	(in)	
1	0	-6.083	-6.083	W4E2	0	6.083	-6.083	W2E1	#12	7
2	0	6.083	-6.083	W1E2	0	6.083	6.083	W3E1	#12	7
3	0	6.083	6.083	W2E2	0	-6.083	6.083	W4E1	#12	7
4	0	-6.083	6.083	W3E2	0	-6.083	-6.083	W1E1	#12	7
5	10	-5.917	-5.917	W8E2	10	5.917	-5.917	W6E1	#12	7
6	10	5.917	-5.917	W5E2	10	5.917	5.917	W7E1	#12	7
7	10	5.917	5.917	W6E2	10	-5.917	5.917	W8E1	#12	7
8	10	-5.917	5.917	W7E2	10	-5.917	-5.917	W5E1	#12	7

Fig 14—*EZNEC* Wires spreadsheet showing the coordinates used for the quad in Fig 13. Note how the x-axis describes the position of an element on the 10-foot boom and also is the axis of symmetry for each element. The values for the z-axis and y-axis vary above and below the axis of symmetry.

sequence. The most obvious example in this particular model is that you should group all the wires associated with a particular element together—for example, the four wires associated with the reflector should be in one place. In Fig 14 you can see that all four wires with an x-coordinate of zero represent the reflector.

It's best to follow a convention in entering wires in a loop structure in a logical fashion. The idea to connect the end point of one wire to the starting point of the next wire. For example, in Fig 13 you can see that the left-hand end of Wire 1 is connected to the bottom of Wire 2, and that the top of Wire 2 connects to the left-hand end of Wire 3. In turn, Wire 3 connects to the top of Wire 4, whose bottom end connects to the right-hand end of Wire 1. The pattern is known as "going around the horn" meaning that the connections proceed smoothly in one direction, in this case in a clockwise direction.

You can see that the entry for the wires making up the elements in the 5-element Yagi in Fig 11B also proceeded in an orderly fashion by starting with the reflector, then the driven element, then director 1, then director 2 and finally director 3. This doesn't mean that you couldn't mix things up, say by specifying the driven element first, followed by director 3, and then the reflector, or whatever. But it's a pretty good bet that doing so in this quasi-random fashion will result in some confusion later on when you revisit a model, or when you let another person see your model.

THE MODELING ENVIRONMENT

The Ground

Above, when considering the 135-foot dipole mounted 50 feet above flat earth, we briefly mentioned the most important environmental item in an antenna model—the ground beneath it. Let's examine some of the options available in the *NEC-2* environment in *EZNEC*:

- Free space
- Perfect ground
- *MININEC* type ground
- "Fast" type ground
- Sommerfeld-Norton ground.

The free space environment option is pretty self-explanatory—the antenna model is placed in free space away from the influence of any type of ground. This option is useful when you wish to optimize certain characteristics of a particular antenna design. For example, you might wish to optimize the front-to-rear ratio of a Yagi over an entire amateur band and this might entail many calculation runs. The free-space ground will run the fastest among all the ground options.

Perfect ground is useful as a reference case, especially for vertically polarized antennas over real ground. Antenna evaluations over perfect ground are shown in most classical antenna textbooks, so it is useful to compare models for simple antennas over perfect ground to those textbook cases.

MININEC type ground is useful when modeling vertical wires, or horizontal wires that are higher than 0.2 λ above ground. A *MININEC* type ground will compute faster than either a "Fast" ground or a Sommerfeld-Norton type of ground because it assumes that the ground under the antenna is perfect, while still taking into account the far-field reflections for ground using user-specified values of ground conductivity and dielectric constant. The fact that the ground under the antenna is perfect allows the *NEC-2* user of a *MININEC* type ground to specify wires that touch (but don't go below) the ground surface, something that only users of the advanced *NEC-4* program can do with the more accurate Sommerfeld-Norton type of ground described below. (*NEC-4* is presently not in the public domain and is strictly restricted and licensed by the US government.) The ability to model grounded wires is useful with vertical antennas. The modeler must be wary of the feed-point source impedances reported for either horizontally or vertically polarized wires because of the perfect-ground assumption inherent in a *MININEC*-type ground.

The "fast" type of ground is a hybrid type of ground that makes certain simplifying assumptions that allow it to be used provided that horizontal wires are higher than about 0.1 λ above ground. With today's fast computers the Sommerfeld-Norton model is preferred.

The Sommerfeld-Norton ground (referred to in *EZNEC* as the "high accuracy" ground) is preferable to the other ground types because it has essentially no practical limitations for wire height. It has the disadvantage that it can run about four times slower than a *MININEC* type of ground, but today's fast computers make that almost a non-issue. Again, *NEC-2*-based programs cannot model wires that penetrate into the ground (although there are work-arounds described below).

As mentioned above, for any type of ground other than perfect ground or free space, the user must specify the conductivity and dielectric constant of the soil. *EZNEC* allows the entry by several user-friendly categories, where σ is conductivity in Siemens/meter and ε is dielectric constant:

- Extremely poor: cities, high buildings (σ = 0.001, ε = 3)
- Very Poor: cities, industrial (σ = 0.001, ε = 5)
- Sandy, dry (σ = 0.002, ε = 10)
- Poor: rocky, mountainous (σ = 0.002, ε = 13)
- Average: pastoral, heavy clay (σ = 0.005, ε = 13)
- Pastoral: medium hills and forestation (σ = 0.006, ε = 13)
- Flat, marshy, densely wooded (σ = 0.0075, ε = 12)
- Pastoral, rich soil, US Midwest (σ = 0.010, ε = 14)
- Very Good: pastoral, rich, central US (σ = 0.0303, ε = 20)
- Fresh water (σ = 0.001, ε = 80)
- Saltwater (σ = 5, ε = 80)

Let's use *EZNEC*'s ability to overlay one or more plots together on one graph to compare the response of the vertical ground plane antenna in Fig 9 for two different types of ground: Saltwater and Poor. Open the **Ch4-GP.EZ** file in *EZNEC*. Click the **Ground Descrip** button and then right-click anywhere in the Media window that opens up. Choose first the "Poor: rocky, mountainous" option button, click **OK** and then **FF Plot**. When the elevation plot appears, click the **File** menu at the top of the main window, and then **Save As**. Choose an appropriate name for the trace, perhaps "Poor Gnd.PF."

Go back and select saltwater as your **Ground Descrip** and follow the same procedure to compute the far-field plot for saltwater ground. Now, add the Poor Gnd.PF trace, by clicking menu selection **File**, **Add Trace**. **Fig 15** shows this comparison, which greatly favors the saltwater environment, particularly at low elevation angles. At 5° the ground plane mounted over saltwater has about a 10 dB advantage compared to its landlocked cousin.

You might be wondering what happens if we move the ground-plane antenna down closer to the ground. The lower limit to how far the radials can approach the lossy earth is 0.001 λ or twice the diameter of the radial wire. A distance of 0.001 λ at 1.8 MHz is about 6 inches, while it is 0.4 inches at 30 MHz. While *NEC-2*-based programs cannot model wires that penetrate the ground, radial systems just above the ground with more than about eight radial wires can provide a work-around to simulate a direct-ground connection.

Modeling Environment: Frequency

It's always a good idea to evaluate an antenna over a *range of frequencies*, rather than simply at a single spot frequency. Trends that become quite apparent on a frequency sweep are frequently lost when looking simply at a single frequency. Native *NEC-2* has built-in frequency sweep capabilities, but once again the commercial programs make the process easier to use and understand. You saw in the SWR curve in Fig 2 the result of one such frequency sweep using *EZNEC*. **Fig 16** shows a frequency sweep of the azimuth response for the 5-element Yagi in Fig 11 across the 20-meter band, using steps of 117 kHz so there are four evaluation frequencies. At 14.0 MHz this Yagi's gain is down a small amount compared to the gain at 14.351 MHz but the rearward pattern is noticeably degraded, dropping to a front-to-back ratio of just under 20 dB.

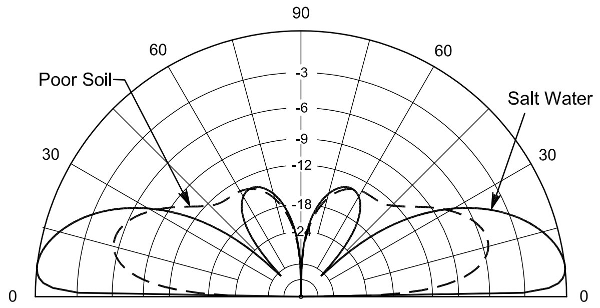

Fig 15—A comparison of the elevation response for the vertical ground plane in Fig 9 over saltwater and over "poor: rocky, mountainous" soil. Saltwater works wonders for verticals, providing excellent low-angle signals.

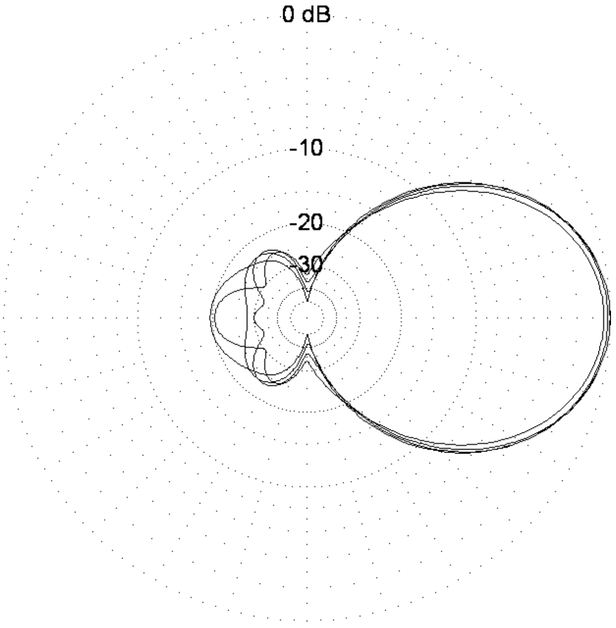

Fig 16—Frequency sweep of 5-element Yagi described in Fig 11, showing how the azimuth pattern changes with frequency.

EZNEC can save to a series of output plot files a frequency sweep of elevation (or azimuth) patterns. In essence, this automates the process described above for saving a plot to disk and then overlaying it on another plot. *EZNEC* can save to a text file for later analysis (or perhaps importation into a spreadsheet) the following parameters, chosen by the user:

- Source data
- Load data
- Pattern data
- Current data
- MicroSmith numeric data
- Pattern analysis summary.

Frequency Scaling

EZNEC has a very useful feature that allows you to create new models *scaled* to a new frequency. You invoke the algorithm used to scale a model from one frequency to another by checking the **Rescale** box after you've clicked the **Frequency** button. *EZNEC* will scale all model dimensions (wire length, height and diameter) except for one specific situation—the wire diameter will stay the same at the new frequency if you originally specified wire size by AWG gauge. For example, #14 copper wire for a half-wave 80-meter dipole will stay #14 copper wire for a 20-meter half-wave dipole. If, however, you specified diameter as a floating point number originally, the diameter will be scaled by the ratio of new to old frequency, along with wire length and height.

Start up *EZNEC* and open up the file **Ch4-520-40W.EZ** for the 5-element 20-meter Yagi on a 40-foot boom. Click the **Frequency** box and then check the **Rescale** check box. Now, type in the frequency of 28.4 MHz and click **OK**. You have quickly and easily created a new 5-element 10-meter Yagi, that is mounted 29.9949 feet high, the exact ratio of 28.4 MHz to 14.1739 MHz, the original design frequency on 20 meters. Click the **FF Plot** button to plot the azimuth pattern for this new Yagi. You will see that it closely duplicates the performance of its 20-meter brother. Click **Src Dat** to see that the source impedance is 25.38 – *j* 22.19 Ω, again very close to the source data for the 20-meter version.

REVISITING SOURCE SPECIFICATION

Sensitivity to Source Placement

Earlier, we briefly described how to specify a source on a particular segment using *EZNEC*. The sources for the relatively simple dipole, Yagi and quad models investigated so far have been in the center of an easy-to-visualize wire. The placement for the source on the vertical ground plane was at the bottom of the vertical radiator, an eminently logical place. In the other cases we specified the position of the source at 50% of the distance along a wire, given that the wire being fed had an odd number of segments. Please note that in each case so far, the feed point (source) has been placed at a relatively low-impedance point, where the current changes relatively slowly from segment to segment.

Now we're going to examine some subtler source-placement problems. *NEC-2* is well-known as being very sensitive to source placement. Significant errors can result from a haphazard choice of the source segment and the segments surrounding it.

Let's return to the inverted-V dipole in Fig 5. The first time we evaluated this antenna (**Ch4-Inverted V Dipole.EZ**) we specified a split source in *EZNEC*. This function uses two sources, one on each of the segments immediately adjacent to the junction of the two downward slanting wires.

Another common method to create a source at the junction of two wires that meet at an angle is to separate these two slanted wires by a short distance and bridge that gap with a short straight wire, which is fed at its center. **Fig 17** shows a close-up of this scheme. In Fig 17 the length of the segments surrounding the short middle wire are purposely made equal to the length of the middle wire. The segmentation for the short middle wire is set to one. **Table 3** lists the source impedance and the maximum gain the *EZNEC* computes for three different models:

1. **Ch4-Inverted V Dipole.EZ** (the original model)
2. **Ch4-Inverted V Dipole Triple Segmentation.EZ**
3. **Ch4-Modified Inverted V Dipole.EZ** (as shown in Fig 17, for the middle wire set to be 2 feet long)
4. **Ch4-Mod Inverted V Poor Segmentation.EZ** (where the number of segments on the two slanted wires have been increased to 200)

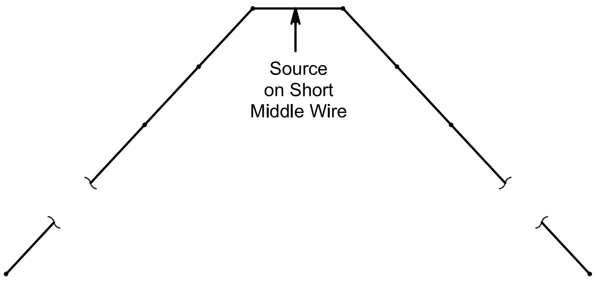

Fig 17—Model of inverted-V dipole using a short center wire on which the source is placed.

Table 3
135-Foot Inverted-V Dipole at 3.75 MHz

Case	*Segments*	*Source Impedance Ω*	*Max. Gain dBi*
1	82	72.64 + *j* 128.2	4.82
2	246	73.19 + *j* 128.9	4.82
3	67	73.06 + *j* 129.1	4.85
4	401	76.21 + *j* 135.2	4.67

Case 2 shows the effect of tripling the number of segments in Case 1. This is a check on the segmentation, to see that the results are stable at a lower level compared to a higher level of segmentation (which theoretically is better, although slower in computation). We purposely set up Case 4 so that the lengths of the segments on either side of the single-segment middle wire are significantly different (0.33 feet) compared to the 2-foot length of the middle wire.

The feed-point and gain figures for the first three models are close to each other. But you can see that the figures for the fourth model are beginning to diverge from the first three, with about a 5% overall change in the reactance and resistance compared to the average values, and about a 3% change in the maximum gain. This illustrates that it is best to keep the segments surrounding the source equal or at least close to equal in length. We'll soon examine a figure of merit called the *Average Gain* test, but it bears mentioning here that the average gain test is very close for the first three models and begins to diverge for the fourth model.

Things get more interesting if the source is placed at a high-impedance point on an antenna—for example, in the center of a full-wave dipole—the value computed for the source impedance will be high, and things will be quite sensitive to the segment lengths. We'll repeat the computations for the same inverted-V models, but this time at twice the operating frequency, at 7.5 MHz.

Table 4 summarizes the results. The impedance is high, as expected. Note that the resistance term varies quite a bit for all four models, a range of about 23% around the average value. Interestingly, the poorly segmented model's resistance falls in between the other three. The reactive terms are closer for all four models but still cover a range of 4% around the average value. The maximum gain shows the same tendancy to be somewhat lower in the fourth model compared to the first three and thus looks as potentially untrustworthy at 7.5 MHz as it does at 3.75 MHz.

This is, of course, but a small sampling of segmentation schemes, and caution dictates that you shouldn't take these results as being representative of all possibilities. Nevertheless, the lesson to be learned here is that the feed-point (source) impedance can vary significantly at a point where the current is changing rapidly, as it does where a high impedance feed is involved. Another general conclusion that can be drawn from Table 5 is that more segments, particularly if they surround the source segment improperly, is not necessarily better.

Table 4
135-Foot Inverted-V Dipole at 7.5 MHz

Case	*Segments*	*Source Impedance Ω*	*Max. Gain dBi*
1	82	2297 – *j* 2668	5.67
2	246	1822 – *j* 2553	5.66
3	67	1960 – *j* 2583	5.66
4	401	2031 – *j* 2688	5.48

Voltage and Current Sources

Before we leave the topic of sources, you should be aware that programs like *EZNEC* and others have the ability to simulate both voltage sources and current sources. Although native *NEC-2* has several source types, voltage sources are the most commonly used by hams. Native *NEC-2* doesn't have a current source, but a current source is nothing more than a voltage source delivering current through a high impedance. Basic network theory says that every Thevenin voltage source has a Norton current source equivalent. Various commercial implementations of *NEC-2* approach the creation of a current source in slightly different fashions. Some use a high value of inductive reactance as a series impedance, while others use a high value of series resistance.

Why would we want to use a current source instead of a voltage source in a model? The general-purpose answer is that models containing a single source at a single feed point can use a voltage source with no problems. Models that employ multiple sources, usually with different amplitudes and different phase shifts, do best with current sources.

For example, *phased arrays* feed RF currents at different amplitudes and phase shifts into two or more elements. The impedances seen at each element may be very different—some impedances might even have negative values of resistance, indicating that power is flowing out of that element into the feed system due to mutual coupling to other elements. Having the ability to specify the amplitude and phase of the current, rather than a feed voltage, at a feed point in a program like *EZNEC* is a valuable tool.

Next, we examine one more important aspect of building a model, setting up loads. After that, we'll look into two tests for the potential accuracy of a model. These tests can help identify source placement, as well as other problems.

LOADS

Many ham antennas, in particular electrically short ones, employ some sort of *loading* to resonate the system. Sometimes loading takes the form of *capacitance hats*, but these can and should be modeled as wires connected to the top of a vertical radiator. A capacitance hat is not the type of loading we'll explore in this section.

Here, the term *loads* refers to discrete inductances, capacitances and resistances that are placed at some point (or points) in an antenna system to achieve certain effects. One fairly common form of a load is a *loading coil* used to resonate an electrically short antenna. Another form of load often seen in ham antennas is a *trap*. *EZNEC* has a special built-in function to evaluate parallel-resonant traps, even at different frequencies beyond their main parallel resonance.

Just for reference, a more subtle type of load is a *distributed* material load. We encountered just such a load in our first model antenna, the 135-foot long flattop dipole—although we didn't identify it specifically as a load at that time. Instead, it was identified as a "wire loss" associated with copper.

The *NEC-2* core program has the capability of simulating a number of built-in loads, including distributed material and discrete loads. *EZNEC* implements the following discrete loads:

- Series R ± *j* X loads.
- Series R-L-C loads, specified in Ω of resistance, µH of inductance and pF of capacitance.
- Parallel R-L-C loads, specified in Ω of resistance, µH of inductance and pF of capacitance.
- Trap loads, specified in Ω of resistance in series with µH of inductance, shunted by pF of capacitance, at a specific frequency.
- Laplace loads, specified as mathematical Laplace coefficients (sometimes used in older modeling programs and left in *EZNEC* for backwards compatibility).

It is important to recognize that the discrete loads in an antenna modeling program *do not radiate* and they have zero size. The *NEC-2* discrete loads are described by L. B. Cebik in his antenna modeling course as being *mathematical loads*. The fact that *NEC-2* loads do not radiate means that the popular mobile antennas that use helical loading coils wound over a length of fiberglass whip cannot be modeled with *NEC-2*, because such coils do radiate.

Let's say that we want to put a air-wound loading coil with an unloaded Q of 400 at the center of a 40-foot long, 50-foot high, flattop dipole so that it is resonant at 7.1 MHz. The schematic of this antenna is shown in **Fig 18**. Examine the modeling file **Ch4-Loaded Dipole.EZ** to see how a discrete series RL load is used to resonate this short dipole at 7.1 MHz, with a feed-point (source) impedance of 25.3 Ω. This requires a series resistance of 1.854 Ω and an inductive reactance of +741.5 Ω. Note that we again used a single wire to model this antenna, and that we placed the load at a point 50% along the length of the wire.

This load represents a 16.62 µH coil with an unloaded Q of 741.5/1.854 = 400, just what we wanted. Let's assume for now that we use a perfect transformer to transform the 25.3-Ω source impedance to 50 Ω. If we now attempt to run a frequency sweep over the whole 40-meter band from 7.0 to 7.3 MHz, the load reactance and resistance will not change, since we specified fixed values for reactance and resistance. Hence, the source impedance will be correct only at the frequency where the reactance and resistance are specified, since the reactance changes with frequency.

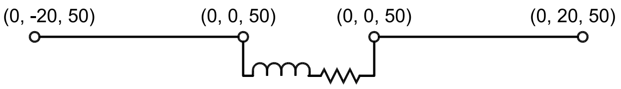

Fig 18—Schematic diagram of a 40-foot long flattop dipole with a loading coil placed at the center. This coil has an unloaded Q of 400 at 7.1 MHz.

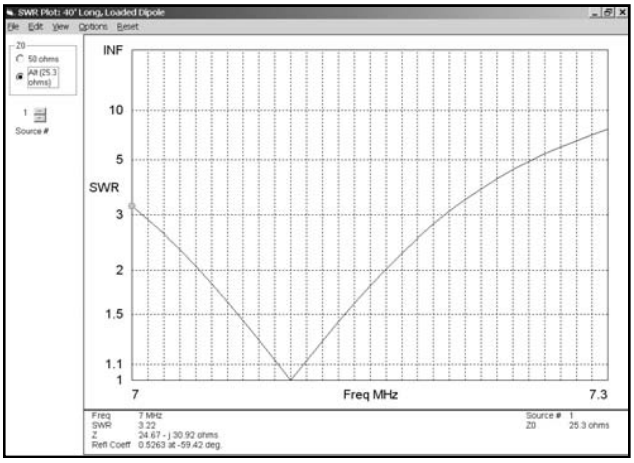

```
Freq       7 MHz                                   Source #  1
SWR        3.22                                    Z0        25.3 ohms
Z          24.67 - j 30.92 ohms
Refl Coeff 0.5263 at -59.42 deg
```

Fig 19—SWR graph of the loaded 40-foot long flattop dipole shown in Fig 18.

So let's use another load capability and substitute a 16.62 µH coil with a series 1.854-Ω resistance at 7.1 MHz. We'll let *EZNEC* take care of the details of computing both the reactance and the changing series resistance at various frequencies. The degree that both reactance and series loss resistance of the coil change with frequency may be viewed using the **Load Dat** button from the main *EZNEC* window.

Fig 19 shows the computed SWR curve for a 25.3-Ω **Alt SWR Z0** reference resistance. The 2:1 SWR bandwidth is about 120 kHz. As could be expected, the antenna has a rather narrow bandwidth because it is electrically short.

ACCURACY TESTS

There are two tests that can help identify accuracy problems in a model:

- The Convergence test.
- The Average Gain test.

Convergence Test

The idea behind the *Convergence* test is simple: If you increase the segmentation in a particular model and the results changes more than you'd like, then you increase the segmentation until the computations converge to a level that is suitable to you. This process has the potential for being subjective, but simple antenna models do converge quickly. In this section, we'll review several more of the antennas discussed previously to see how they converge.

Let's go back to the simple dipole in Fig 3. The original segmentation was 11 segments, but we'll start with a very low value of segmentation of three, well below the minimum recommended level. **Table 5** shows how the source impedance and gain change with increase in segmentation at 3.75 MHz. For this simple antenna, the gain levels off at 6.50 dBi by the time the segmentation has reached 11 segments. Going to ten times the minimum-recommended level (to 111 segments) results in an increase of only 0.01 dBi in the gain.

Arguably, the impedance has also stabilized by the time we reach a segmentation level of 11 segments, although purists may opt for 23 segments. The tradeoff is a slowdown in computational speed.

Let's see how the 5-element Yagi model converges with changes in segmentation level. **Table 6** shows how the source impedance, gain, 180° front-to-back ratio and worst-case front-to-rear ratio change with segmentation density. By the time the segmentation has reached 11 segments per wire, the impedance and gain have stabilized quite nicely, as has the F/R. The 180° F/B is still increasing with segmentation level until about 25 segments, but a relatively small shift in frequency will change the maximum F/B level greatly. For example, with 11 segments per wire, shifting the frequency to 14.1 MHz—a shift of only 0.5%— will change the maximum 180° F/B from almost 50 dB down to 27 dB. For this reason the F/R is considered a more reliable indicator of the adequacy of the segmentation level than is F/B.

Table 5
135-Foot Flattop Dipole at 3.75 MHz

Segments	*Source Impedance Ω*	*Max. Gain dBi*
3	85.9 + *j* 128.0	6.34
5	86.3 + *j* 128.3	6.45
7	86.8 + *j* 128.8	6.48
11	87.9 + *j* 129.5	6.50
23	88.5 + *j* 130.3	6.51
45	89.0 + *j* 130.8	6.51
101	89.4 + *j* 131.1	6.51

Table 6
5-element Wire Yagi at 14.1739 MHz

Segments	*Source Impedance Ω*	*Max. Gain dBi*	*180° F/B dB*	*F/R dB*
3	28.5 – *j* 30.6	12.79	23.2	22.4
5	26.3 – *j* 25.6	13.02	30.5	23.1
7	25.6 – *j* 24.0	13.07	34.8	23.1
11	25.1 – *j* 22.9	13.09	39.9	23.1
25	24.9 – *j* 22.0	13.10	43.7	23.1
99	24.7 – *j* 21.5	13.10	44.2	23.1

Average Gain Test

The theory behind the *Average Gain* test is a little more involved. Basically, if you remove all intentional losses in a model, and if you place the antenna either in free space or over perfect ground, then all the power fed to the antenna should be radiated by it. Internally, the program runs a full 3-D analysis, adding up the power in all directions and then dividing that sum by the total power fed to the antenna. Since *NEC-2* is very sensitive about source placement, as mentioned before, the Average Gain test is a good indicator that something is wrong with the specification of the source.

Various commercial versions of *NEC-2* handle the Average Gain test in different ways. *EZNEC* requires the operator to turn off all distributed losses in wires or set to zero any discrete resistive losses in loads. Next you set the ground environment to free space (or perfect ground) and request a 3-D pattern plot. *EZNEC* will then report the average gain, which will be 1.000 if the model has no problems. The average gain can be lower or higher than 1.000, but if it falls within the range 0.95 to 1.05 it is usually considered adequate.

As L. B. Cebik, W4RNL, stated in his ARRL Certification and Continuing Education Course on antenna modeling: "Like the convergence test, the average gain test is a necessary but not a sufficient condition of model reliability." Pass both tests, however, and you can be pretty well sure that your model represents reality. Pass only one test, and you have reason to worry about how well your model represents reality.

Once again, open the model file **Ch4-Mod Inverted V Poor Segmentation.EZ** and set **Wire Loss** to zero, **Ground Type** to Free Space and **Plot Type** to 3-Dimensional. Click on the **FF Plot** button. *EZNEC* will report that the Average Gain is 0.955 = –0.2 dB. This is very close to the lower limit of 0.95 considered valid for excellent accuracy. This is a direct result of forcing the segment lengths adjacent to the source segment to be considerably shorter than the source segment's length. The gain reported using this test would be approximately –0.2 dB from what it should be—just what Table 4 alludes to also.

Now, let's revisit the basic model **Ch4-Inverted V Dipole.EZ** and look at Case 2 in Table 4. Case 2 amounts to a Convergence test for the basic inverted-V model. Since the impedance and gain changes were small comparing the basic model to the one using three times the number of segments, the model passed the Convergence test. The Average Gain test for the basic model yields a value of 0.991, well within the limits for good accuracy. This model has thus passed both tests and can be considered accurate.

Running the Average Gain test for the 5-element Yagi (using 11 segments per wire and whose convergence we examined in Table 6) yields a value of 0.996, again well within the bounds indicating a good model. And the simple flattop dipole with 11 segments at 3.75 MHz yields an Average Gain result of 0.997, again indicating a very accurate model.

OTHER POSSIBLE MODEL LIMITATIONS

Programs based on the *NEC-2* core computational code have several well-documented limitations that you should know about. Some limitations have been removed in the restricted-access *NEC-4* core (which is not generally available to users), but other limitations still exist, even in *NEC-4*.

Closely Spaced Wires

If wires are spaced too close to each other, the *NEC-2* core can run into problems. If the segments are not carefully aligned, there also can be problems with accuracy. The worst-case situation is where two wires are so close together that their volumes actually merge into each other. This can happen where wires are thick, parallel to each other and close together. You should keep parallel wires separated by at least several diameters.

For example, #14 wire is 0.064 inches in diameter. The rule then is to keep parallel #14 wires separated by more than 2 × 0.064 = 0.128 inches. And you should run the Convergence test to assure yourself that the solution is indeed converging when you have closely spaced wires, especially if the two wires have different diameters. To model antennas containing closely spaced wires, very often you will need many more segments than usual and you must also carefully ensure that the segments line up with each other.

Things can get a little more tricky when wires cross over or under each other, simply because such crossings are sometimes difficult to visualize. Again, the rule is to keep crossing wires separated by more than two diameters from each other. And if you intend to join two wires together, make sure you do so at the ends of the two wires, using identical end coordinates. When any or all of these rules are violated, the Convergence and Average Gain tests will usually warn you of potential inaccuracies.

Parallel-Wire Transmission Lines and LPDAs

A common example of problems with closely spaced wires is when someone attempts to model a parallel-wire transmission line. *NEC-2*-based programs usually do not work as well in such situations as do *MININEC*-based programs. The problems are compounded if the diameters are different for the two wires simulating a parallel-wire transmission line. In *NEC-2* programs, it is usually better to use the built-in "perfect transmission line" function than to try to model closely spaced parallel wires as a transmission line.

For example, a Log Periodic Dipole Array (LPDA) is composed of a series of elements fed using a transmission line that reverses the phase 180° at each element. In other words, the elements are connected to a transmission line that reverses connections left-to-right at each element. It is cumbersome to do so, but you could model such a transmission line using separate wires in *EZNEC*, but it is a potentially confusing and a definitely painstaking process. Further, the accuracy of the resulting model is usually suspect, as shown by the Average Gain test.

It is far easier to use the **Trans Lines** function from the *EZNEC* main window to accurately model an LPDA. See **Fig 20**, which shows the **Trans Lines** window for the **9302A.EZ** 16-element LPDA. There are 15 transmission lines connecting the 16 elements, placed at the 50% point on each element, with a 200-Ω characteristic impedance and with Reversed connections.

Fat Wires Connected to Skinny Wires

Another inherent limitation in the *NEC-2* computational core shows up when modeling several popular ham-radio antennas: many Yagis and some quads.

Tapered Elements

As mentioned before, many Yagis are built using telescoping aluminum tubing. This technique saves weight and makes for a more flexible and usually stronger element, one that can survive wind and ice loading better than a "monotaper" element design. Many vertical antennas are also constructed using telescoping aluminum tubing.

Unfortunately, native *NEC-2* doesn't model accurately such *tapered elements*, as they are commonly called. There is, however, a sophisticated and accurate work-around for such elements, called the *Leeson corrections*. The Leeson corrections, derived by Dave Leeson, W6NL, from pioneering work by Schelkunoff at Bell Labs, compute the diameter and length of an element that is electrically equivalent to a tapered element. This monotaper element is much easier to use in a pro-

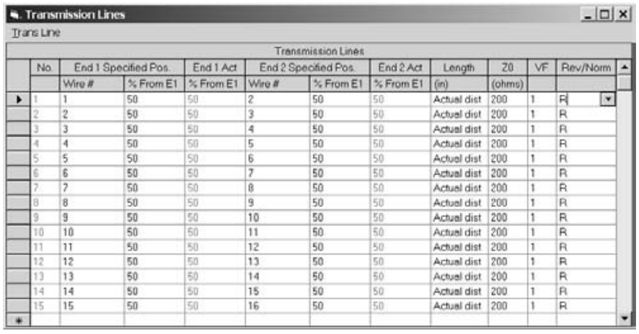

Transmission Lines

No.	End 1 Specified Pos.		End 1 Act	End 2 Specified Pos.		End 2 Act	Length	Z0	VF	Rev/Norm
	Wire #	% From E1	% From E1	Wire #	% From E1	% From E1	(in)	(ohms)		
1	1	50	50	2	50	50	Actual dist	200	1	R
2	2	50	50	3	50	50	Actual dist	200	1	R
3	3	50	50	4	50	50	Actual dist	200	1	R
4	4	50	50	5	50	50	Actual dist	200	1	R
5	5	50	50	6	50	50	Actual dist	200	1	R
6	6	50	50	7	50	50	Actual dist	200	1	R
7	7	50	50	8	50	50	Actual dist	200	1	R
8	8	50	50	9	50	50	Actual dist	200	1	R
9	9	50	50	10	50	50	Actual dist	200	1	R
10	10	50	50	11	50	50	Actual dist	200	1	R
11	11	50	50	12	50	50	Actual dist	200	1	R
12	12	50	50	13	50	50	Actual dist	200	1	R
13	13	50	50	14	50	50	Actual dist	200	1	R
14	14	50	50	15	50	50	Actual dist	200	1	R
15	15	50	50	16	50	50	Actual dist	200	1	R

Fig 20—Transmission-line window for the 9302A.EZ 16-element LPDA. Note that the transmission lines going between elements are "reversed," meaning that they are 180° out-of-phase at each element, a requirement for properly feeding an LPDA.

Table 7

5-element Yagi at 14.1739 MHz with Telescoping Aluminum Elements

	With Leeson Corrections			Without Leeson Corrections		
Freq. MHz	*Source Impedance Ω*	*Gain dBi*	*F/R dB*	*Source Impedance Ω*	*Gain dBi*	*F/R dB*
14.0	23.2 – *j* 26.5	14.82	23.3	22.4 – *j* 12.7	14.92	23.1
14.1	22.7 – *j* 20.5	14.87	22.8	18.6 – *j* 12.5	14.70	21.6
14.2	22.8 – *j* 14.8	14.87	22.7	6.6 – *j* 4.6	14.01	16.2
14.3	22.5 – *j* 11.9	14.76	21.5	1.9 + *j* 10.6	10.61	3.1
14.4	14.5 – *j* 10.5	14.45	19.9	1.6 + *j* 23.7	11.15	–11.4

gram like *NEC-2*. See Chapter 2, Antenna Fundamentals, for more information on the Leeson corrections.

EZNEC and other *NEC-2* programs can automatically invoke the Leeson corrections, providing that some basic conditions are met—and happily, these conditions are true for the telescoping aluminum-tubing elements commonly used as Yagi elements. *EZNEC* gives you the ability to disable or enable Leeson corrections, under the **Option** menu, under **Stepped Diameter Correction**, *EZNEC*'s name for the Leeson corrections. Open the modeling file **520-40H.EZ**, which contains tapered aluminum tubing elements and compare the results using and without using the Leeson corrections.

Table 7 lists the differences over the 20-meter band, with the 5-element Yagi at a height of 70 feet above flat ground. You can see that the non-Leeson corrected figures are very different from the corrected ones. At 14.3 MHz, the pattern for the non-corrected Yagi has degenerated to a F/R of 3.1 db, while at 14.4 MHz, just outside the top of the Amateur band, the pattern for the non-corrected antenna actually has reversed. Even at 14.2 MHz, the non-corrected antenna shows a low source impedance, while the corrected version exhibits smooth variations in gain, F/R and impedance across the whole band, just as the actual antenna exhibits.

Some Quads

Some types of cubical quads are made using a combination of aluminum tubing and wire elements, particularly in Europe where the "Swiss" quad has a wide following. Again, *NEC-2*-based programs don't handle such tubing/wire elements well. It is best to avoid modeling this type of antenna, although there are some ways to attempt to get around the limitations, ways that are beyond the scope of this chapter.

NEAR-FIELD OUTPUTS

FCC regulations set limits on the maximum permissible exposure (MPE) allowed from the operation of radio transmitters. These limits are expressed in terms of the electric (V/m) and magnetic fields (A/m) close to an antenna. *NEC-2*-based programs can compute the electric and magnetic *near fields* and the FCC accepts such computations to demonstrate that an installation meets their regulatory requirements. See Chapter 1, Safety, in this book.

Table 8

E- and H-Field Intensities for 1500 W into 5-Element Yagi at 70 Feet on 14.2 MHz

Height Feet	*H-Field (A/m)*	*E-Field (V/m)*
0	0.04	4.1
10	0.03	13.8
20	0.04	20.6
30	0.06	22.6
40	0.08	25.8
50	0.10	33.8
60	0.12	41.5
70	0.12	44.3

We'll continue to use the 5-element Yagi at 70 feet to demonstrate a near-field computation. Open **Ch4-520-40H.EZ** in *EZNEC* and choose **Setups** and then **Near Field** from the menu at the top of the main window. Let's calculate the E-field and H-field intensity for a power level of 1500 W (chosen using the **Options**, **Power Level** choices from the main menu) in the main beam at a fixed distance, say 50 feet, from the tower base. We'll do this at various heights, using 10-foot increments of height, in order to see the lobe structure of the Yagi at 70 feet height.

Table 8 summarizes the total H- and E-field intensities as a function of height. As you might expect, the fields are strongest directly in line with the antenna at a height of 70 feet. At ground level, the total fields are well within the FCC limits for rf exposure for both fields. In fact, the fields are within the FCC limits if someone were to stand at the tower base, directly under the antenna.

ANTENNA MODELING SUMMARY

This section on antenna modeling is by necessity only a brief introduction to the science of antenna modeling. The subject is partly art as well as science because there are usually several ways of creating a model for a particular antenna or antennas.

Indeed, the presence of other wires surrounding a particular antenna can affect the performance of that antenna. Finally, there are the practical aspects of putting a actual antenna up in the real world. We'll explore this next.

Practical Aspects, Designing Your Antenna System

The most important time spent in putting together an antenna system is the time spent in planning. In Chapter 3, The Effects of Ground, we outlined the steps needed to evaluate how your local terrain can affect HF communications. There we emphasized that you need to compare the patterns resulting from your own terrain to the statistically relevant elevation angles needed for coverage of various geographic areas. (The elevation-angle statistics were developed in Chapter 23, Radio Wave Propagation and are located on the CD-ROM included with this book, as is the terrain-assessment program *HFTA*.)

The implicit assumptions in Chapter 3 are (1) that you know where you want to talk to, and (2) that you'd like the most effective system possible. At the start of such a theoretical analysis, cost is no object. Practical matters, like cost or the desires of your spouse, can come later! After all, you're just checking out all the possibilities. If nothing else, you will use the methodology in Chapter 3 to evaluate any property you are considering buying so that you can build your "dream station."

Next, in the first part of this chapter we described modeling tools used to evaluate different types of antennas. These modeling tools can help you evaluate what type of antenna might be suitable to your own particular style of operating. Do you want a Yagi with a lot of rejection of received signals from the rear? Let's say that terrain analysis shows that you need an antenna at least 50 feet high. Do you really need a steel tower, or would a simple dipole in the trees serve your communication needs just fine? How about a vertical in your backyard? Would that be inconspicuous enough to suit your neighbors and your own family, yet still get you on the air?

In short, using the techniques and tools we've presented in Chapters 3, 23 and here in Chapter 4, you can scientifically plan an antenna *system* that will be best suited for your own particular conditions. Now, however, you have to get practical. Thinking through and planning the installation can save a lot of time, money and frustration. While no one can tell you the exact steps you should take in developing your own master plan, this section, prepared originally by Chuck Hutchinson, K8CH, should help you with some ideas.

WHAT DO YOU REALLY WANT?

Begin planning by spelling out your communications *desires*. What bands are you interested in? Who (or where) do you want to talk to? When do you operate? How much time and money are you willing to spend on an antenna system? What physical limitations affect your master plan?

From the answers to the above questions, begin to formulate *goals*—short, intermediate and long range. Be realistic about those goals. Remember that there are three station effectiveness factors that are under your control. These are: operator skill, equipment in the shack, and the antenna system. There is no substitute for developing operating skills. Some tradeoffs are possible between shack equipment and antennas. For example, a high-power amplifier can compensate for a less than optimum antenna. By contrast, a better antenna has advantages for receiving as well as for transmitting.

Consider your *limitations*. Are there regulatory restrictions on antennas in your community? Are there any deed restrictions or covenants that apply to your property? Do other factors (finances, family considerations, other interests, and so forth) limit the type or height of antennas that you can erect? All of these factors must be investigated because they play a major role determining the type of antennas you erect.

Chances are that you won't be able to immediately do all you desire. Think about how you can budget your resources over a period of time. Your resources are your money, your time available to work, materials you may have on hand, friends that are willing to help, etc. One way to budget is to concentrate your initial efforts on a given band or two. If your major interest is in chasing DX, you might want to start with a very good antenna for the 14-MHz band. A simple multiband antenna could initially serve for other frequencies. Later you can add better antennas for those other bands.

SITE PLANNING

A map of your property or proposed antenna site can be of great help as you begin to consider alternative antennas. You'll need to know the size and location of buildings, trees and other major objects in the area. Be sure to note compass directions on your map. Graph or quadrille paper (or a simple CAD program) can be very useful for this purpose. See **Fig 21** for an example. It's a good idea to make a few photocopies of your *site map* so you can mark on the copies as you work on your plans.

Use your map to plan antenna layouts and locations of any supporting towers or masts. If your plan calls for more than one tower or mast, think about using them as supports for wire antennas. As you work on a layout, be sure to think in three dimensions even though the map shows only two.

Be sensitive to your neighbors. A 70-foot guyed tower in the front yard of a house in a residential neighborhood is not a good idea (and probably won't comply with local ordinances!). You probably will want to locate that tower in the back yard.

ANALYSIS

Use the information earlier in this chapter and in Chapters 3 and 23 to analyze antenna patterns in both

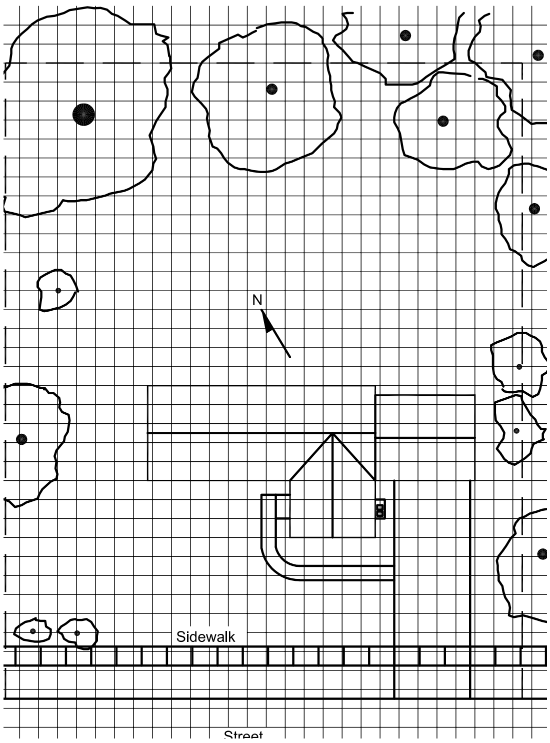

Fig 21—A site map such this one is a useful tool for planning your antenna installation.

horizontal and vertical planes towards geographic areas of interest. If you want to work DX, you'll want antennas that radiate energy at low as well as intermediate angles. An antenna pattern is greatly affected by the presence of ground and by the local topography of the ground. Therefore, be sure to consider what effect ground will have on the antenna pattern at the height you are considering. A 70-foot high antenna is approximately $^{1}/_{2}$, 1, 1$^{1}/_{2}$ and 2 wavelengths (λ) high on 7, 14, 21 and 28 MHz respectively. Those heights are useful for long-distance communications. The same 70-foot height represents only λ/4 at 3.5 MHz, however. Most of the radiated energy from a dipole at that height would be concentrated straight up. This condition is not great for long-distance communication, but can still be useful for some DX work and excellent for short-range communications.

Lower heights can be useful for communications. However, it is generally true that "the higher, the better" as far as communications effectiveness is concerned. This general rule of thumb, of course, should be tempered by an exact analysis of your local terrain. Being located at the top of a steep hill can mean that you can use lower tower heights to achieve good coverage.

There may be cases where it is not possible to install low-frequency dipoles at λ/4 or more above the ground. A vertical antenna with many radials is a good choice for long-distance communications. You may want to install both a dipole and a vertical for the 3.5- or 7-MHz bands. On the 1.8-MHz band, unless extremely tall supports are available, a vertical antenna is likely to be the most useful for DXing. You can then choose the antenna that performs best for a given set of conditions. A low dipole will generally work better for shorter-range communications, while the vertical will generally be the better performer over longer distances.

Consider the azimuthal pattern of fixed antennas. You'll want to orient any fixed antennas to favor the directions of greatest interest to you.

BUILDING THE SYSTEM

When the planning is completed, it is time to begin construction of the antenna system. Chances are that you can divide that construction into a series of *phases* or steps. Say, for example, that you have lots of room and that your long-range plan calls for a pair of towers, one 100-feet high, and the other 70-feet high, to support monoband Yagi antennas. The towers will also support a horizontal 3.5-MHz dipole, for DX work. On your map you've located them so the 80-meter dipole will be broadside to Europe. You decide to build the 70-foot tower with a "triband" beam and 80- and 40-meter inverted-V dipoles to begin the project.

In your master plan you design the guys, anchors and all hardware for the 70-foot tower to support the load of stacked 4-element 10- and 15-meter monobanders Yagis. So you make sure you buy a heavy-duty rotator and the stout mast needed for the monoband antennas later. Thus you avoid having to buy, and then sell, a medium-duty rotator and lighter weight tower equipment later on when you upgrade the station. You could have saved money in the long run by putting up a monoband beam for your favorite band, but you decided that for now it is more important to have a beam on 14, 21 and 28 MHz, so you choose a commercial triband Yagi.

The second step of your plan calls for installing the second tower and stacking a 2-element 40-meter and a 4-element 20-meter monoband Yagi on it. You also plan to replace the tribander on the 70-foot tower with stacked 4-element 10- and 15-meter monoband Yagis. Although this is still a "dream system" you can now apply some of the modeling techniques discussed earlier in this chapter to determine the overall system performance.

Modeling Interactions at Your Dream Station

In this analysis we're going to assume that you have sufficient real estate to separate the 70- and 100-foot towers by 150 feet so that you can easily support an 80-meter dipole between them. We'll also assume that you want the 80-meter dipole to have its maximum response at a heading of 45° into Europe from your location in Newington, Connecticut. The dipole will also have a lobe

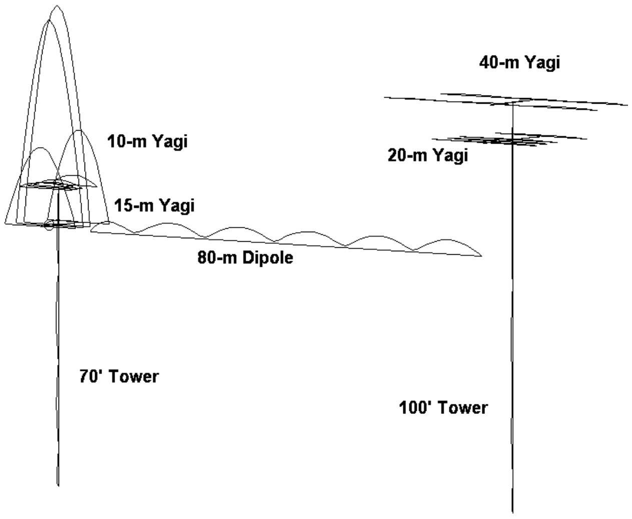

Fig 22—Layout for two-tower antenna system, at 70 and 100 feet high and 150 feet apart. The 70-foot tower has a 4-element 10-meter Yagi at 80 feet on a 10-foot rotating mast and a 4-element 15-meter Yagi at 70 feet. An 80-meter dipole goes from the 70-foot tower to the 100-foot tower, which holds a 2-element 40-meter Yagi at 110 feet and a 4-element 20-meter Yagi at 100 feet. In this figure all the rotatable Yagis are facing the direction of Europe and the currents on the 15-meter Yagi are shown. Note the significant amount of current re-radiated by the nearby 80-meter dipole.

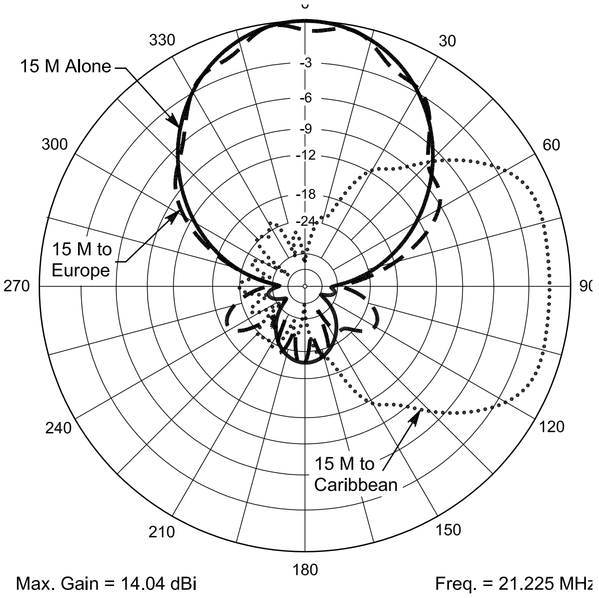

Fig 23—An overlay of azimuth patterns. The solid line is the radiation pattern for the 15-meter Yagi all by itself. The dashed line is the pattern for the 15-meter Yagi, as affected by all the other antennas. The dotted line is the pattern for the 15-meter Yagi when it is pointed toward the Caribbean, with the Yagis on the 100-foot tower pointed toward the 70-foot tower. The peak response of the 15-meter Yagi has dropped by about 1.5 dB.

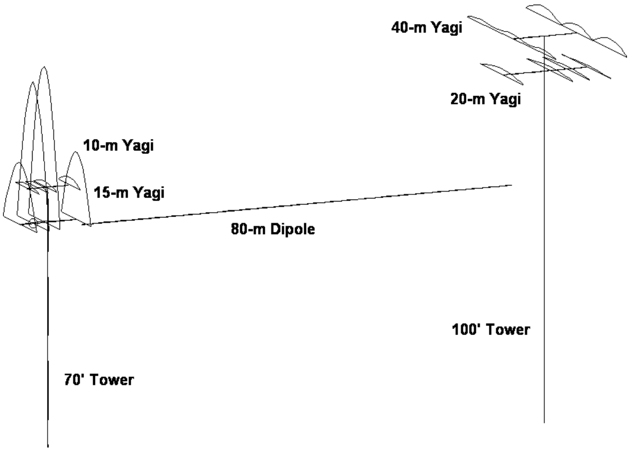

Fig 24—The layout and 15-meter currents when the Yagis on the 100-foot tower are pointed toward the 70-foot tower. The 15-meter Yagi has been rotated to face the direction of the 100-foot tower (toward the Caribbean).

facing 225° towards the USA and New Zealand, making it a good antenna for both domestic contacts and DX work.

Let's examine the interactions that occur between the rotatable Yagis for 10, 15, 20 and 40 meters. See **Fig 22**, which purposely exaggerates the magnitude of the currents on the 4-element 15-meter Yagi mounted at 70 feet. Here, both sets of Yagis have been rotated so that they are pointing into Europe. There is a small amount of current radiated onto the 10-meter antenna but virtually no current is radiated onto the 40- and 20-meter Yagis. This is good.

However, significant current is radiated onto, and then re-radiated, by the 80-meter dipole. This undesired current affects the radiation pattern of the 15-meter antenna, as shown in **Fig 23**, which overlays the pattern of the 4-element 15-meter Yagi by itself with that of the Yagi interacting with the other antennas. You can see "ripples" in the azimuthal response of the 15-meter Yagi due to the effects of the 80-meter dipole's re-radiation. The magnitude of the ripples is about 1 dB at worst, so they don't seriously affect the forward pattern (into Europe), but the rearward lobes are degraded somewhat, to just below 20 dB.

Fig 23 also shows the worst-case situation for the 15-meter Yagi. Here, the 15- and 10-meter stack has been turned clockwise 90°, facing the Caribbean, while the 40- and 20-meter Yagis on the 100-foot tower have been turned counter-clockwise 90° (in the direction of Japan) to face the 70-foot tower holding the 10/15-meter Yagis. You can see the layout and the currents in **Fig 24**. Now the 40- and 20-meter Yagis re-radiate some 15-meter

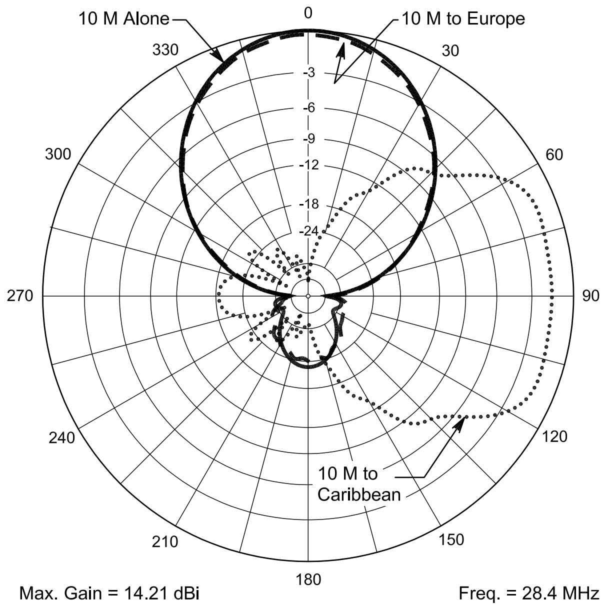

Fig 25—The radiation patterns for the 10-meter Yagi. The solid line is the 10-meter Yagi by itself. The dashed line is for the same Yagi, with all other antenna interactions. The dotted line shows the worst-case pattern, with the stacked Yagis on the 100-foot tower facing the 70-foot tower and the 10-meter Yagi pointed toward the Caribbean. Again, the peak response of the 10-meter Yagi has dropped about 1.5 dB in the worst-case situation.

energy and reduce the maximum gain by about 1.5 dB. Note that in this direction the 80-meter dipole no longer has 15-meter energy radiated onto it by the 15-meter Yagi.

The shape of the patterns will change depending on whether you specify "current" or "voltage" sources in the models for the other antennas, since this effectively opens up or shorts the feed points at the other antennas so far as 15-meter energy is concerned. In practice, this means that the interaction between antennas will vary somewhat depending on the length of the feed lines going to each antenna and whether each feed line is open-circuited or short-circuited when it is not in use.

You can now see that interactions between various antennas pointing in different directions can be significant in a real-world antenna system. In general, higher-frequency antennas are affected by re-radiation from lower-frequency antennas, rather than the other way around. Thus the presence of a 10- or 15-meter stack does not affect the 20-meter Yagi at all.

Modeling can also help determine the minimum stacking distance required between monoband Yagis on the same rotating mast. In this case, stacking the 10- and 15-meter monobanders 10 feet apart holds down interaction between them so that the pattern and gain of the 10-meter Yagi is not impacted adversely. **Fig 25** demonstrates this in the European direction, where the patterns for the 10-meter beam by itself looks very clean compared to the same Yagi separated by 10 feet from the 15-meter Yagi below it. The worst-case situation is pointing towards the Caribbean, when the 40- and 20-meter stack is facing the 70-foot tower. This drops the 10-meter gain down about 1.5 dB from maximum, indicating significant interaction is occurring.

In this situation you might find it best to place the 70-foot tower in the direction closest to the Caribbean if this direction is very important to you. Doing so will, however, cause the pattern in the direction of the Far East to be affected on 10 and 15 meters. You have the modeling tools necessary to evaluate various configurations to achieve whatever is most important to you.

COMPROMISES

Because of limitations, most amateurs are never able to build their dream antenna system. This means that some compromises must be made. Do not, under any circumstances, compromise the *safety* of an antenna installation. Follow the manufacturer's recommendations for tower assembly, installation and accessories. Make sure that all hardware is being used within its ratings.

Guyed towers are frequently used by radio amateurs because they cost less than more complicated unguyed or freestanding towers with similar ratings. Guyed towers are fine for those who can climb, or those with a friend who is willing to climb. But you may want to consider an antenna tower that folds over, or one that cranks up (and down). Some towers crank up (and down) and fold over too. See **Fig 26**. That makes for convenient access to antennas for adjustments and maintenance without climbing. Crank-up towers also offer another advantage. They allow antennas to be lowered during periods of no operation, such as for aesthetic reasons or during periods of high winds.

A well-designed monoband Yagi should outperform a multiband Yagi. In a monoband design the best adjustments can be made for gain, front-to-rear ratio (F/R) and matching, *but only for a single band*. In a multiband design, there are always tradeoffs in these properties for the ability to operate on more than one band. Nevertheless, a multiband antenna has many advantages over two or more single band antennas. A multiband antenna requires less heavy-duty hardware, requires only one feed line, takes up less space and it costs less.

Apartment dwellers face much greater limitations in their choice of antennas. For most, the possibility of a tower is only a dream. (One enterprising ham made arrangements to purchase a top-floor condominium from a developer. The arrangements were made before construction began, and the plans were altered to include a roof-top tower installation.) For apartment and condominium dwellers, the situation is still far from hopeless. A later section presents ideas for consideration.

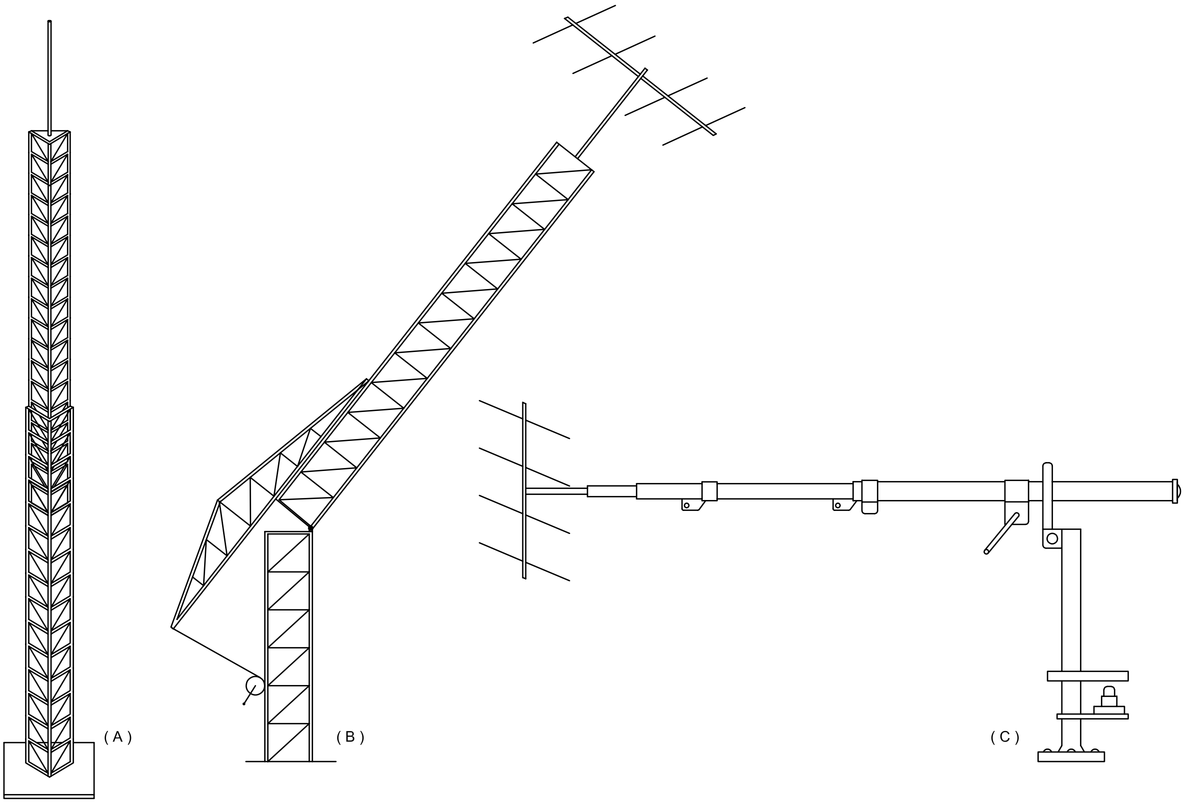

Fig 26—Alternatives to a guyed tower are shown here. At A, the crank-up tower permits working on antennas at reduced height. It also allows antennas to be lowered during periods of no operation. Motor-driven versions are available. The fold-over tower at B and the combination at C permit working on antennas at ground level.

EXAMPLES

You can follow the procedure previously outlined to put together modest or very large antenna systems. What might a ham put together for antennas when he or she wants to try a little of everything, and has a modest budget? Let's suppose that the goals are (1) low cost, (2) no tower, (3) coverage of all HF bands and the repeater portion of one VHF band, and (4) the possibility of working some DX.

After studying the pages of this book, the station owner decides to first put up a 135-foot center-fed antenna. High trees in the back yard will serve as supports to about 50 feet. This antenna will cover all the HF bands by using a balanced feeder and an antenna tuner. It should be good for DX contacts on 10 MHz and above, and will probably work okay for DX contacts on the lower bands. However, her plan calls for a vertical for 3.5 and 7 MHz to enhance the DX possibilities on those bands. For VHF, a chimney-mounted vertical is included.

ANOTHER EXAMPLE

A licensed couple has bigger ambitions. Goals for their station are (1) a good setup for DX on 14, 21 and 28 MHz, (2) moderate cost, (3) one tower, (4) ability to work some DX on 1.8, 3.5 and 7 MHz, and (5) no need to cover the CW portion of the bands.

After considering the options, the couple decides to install a 65-foot guyed tower. A large commercial triband Yagi will be mounted on top of the tower. The center of a trap dipole tuned for the phone portion of the 3.5- and 7-MHz bands will be supported by a wooden yardarm installed at the 60-foot level of the tower, with ends drooping down to form an inverted V. An inverted L for 1.8 MHz starts near ground level and goes up to a similar yardarm on the opposite side of the tower. The horizontal portion of the inverted L runs away from the tower at right angles to the trap dipole. Later, the husband will experiment with sloping antennas for 3.5 MHz. If those experiments are not successful, a $\lambda/4$ vertical will be used on that band.

Apartment Possibilities

A complete and accurate assessment of antenna types, antenna placement and feed-line placement is very important for the apartment dweller. Among the many possibilities for types are balcony antennas, *invisible* ones (made of fine wire), vertical antennas disguised as flag poles or as masts with a TV antenna on top, and indoor antennas.

A number of amateurs have been successful negotiating with the apartment owner or manager for permission to install a short mast on the roof of the building. Coaxial lines and rotator control cables might be routed through conduit troughs or through ductwork. If you live in one of the upper stories of the building, routing the cables over the edge of the roof and in through a window might be the way to go. There is a story about one amateur who owns a triband beam mounted on a 10-foot mast. But even with such a short mast, he is the envy of all his amateur friends because of his superb antenna height. His mast stands on top of a 22-story apartment building.

Usually the challenge is to find ways to install antennas that are unobtrusive. That means searching out antenna locations such as balconies, eaves, nearby trees, etc. For example, a simple but effective balcony antenna is a dangling vertical. Attach a thin wire to the tip of a mobile whip or a length of metal rod or tubing. Then mount the rigid part of the antenna horizontally on the balcony rail, dangling the wire over the edge. The antenna is operated against the balcony railing or other metallic framework. A matching network is usually required at the antenna feed point. Metal in the building will likely give a directivity effect, but this may be of little consequence and perhaps even an advantage. The antenna may be removed and stored when not in use.

Frequently, the task of finding an inconspicuous route for a feed line is more difficult than the antenna installation itself. When Al Francisco, K7NHV, lived in an apartment, he used a tree-mounted vertical antenna. The coax feeder exited his apartment through a window and ran down the wall to the ground. Al buried the section of line that went from under the window to a nearby tree. At the tree, a section of enameled wire was connected to the coax center conductor. He ran the wire up the side of the tree away from foot traffic. A few short radials completed the installation. The antenna worked fine, and was never noticed by the neighbors.

See Chapters 6, Low-Frequency Antennas, and Chapter 15, Portable Antennas, for ideas about low-frequency and portable antennas that might fit into your available space. Your options are limited as much by your imagination and ingenuity as by your pocketbook. Another option for apartment dwellers is to operate away from home. Some hams concentrate on mobile operation as an alternative to a fixed station. It is possible to make a lot of contacts on HF mobile. Some have worked DXCC that way.

Suppose that you like VHF contests. Because of other activities, you are not particularly interested in operating VHF outside the contests. Why not take your equipment and antennas to a hilltop for the contests? Many hams combine a love for camping or hiking with their interest in radio.

Antennas for Limited Space

It is not always practical to erect full-size antennas for the HF bands. Those who live in apartment buildings may be restricted to the use of minuscule radiators because of house rules, or simply because the required space for full-size antennas is unavailable. Other amateurs may desire small antennas for aesthetic reasons, perhaps to keep peace with neighbors who do not share their enthusiasm about high towers and big antennas. There are many reasons why some amateurs prefer to use physically-shortened antennas. This section discusses proven designs and various ways of building and using them effectively. You will find that modeling antennas by computer, even compromised “stealth antennas,” can help you determine the most practical system possible for your particular circumstances—before you go through the effort of stringing up wires.

Few compromise antennas are capable of delivering the performance you can expect from the full-size variety. But the patient and skillful operator can often do as well as some who are equipped with high power and full-size antennas. Someone with a reduced-size antenna may not be able to “bore a hole” in the bands as often and with the commanding dispatch enjoyed by those who are better equipped, but DX can be worked successfully when band conditions are suitable.

INVISIBLE ANTENNAS

We amateurs don’t regard our antennas as eyesores; in fact, we almost always regard them as works of art! But there are occasions when having an outdoor or visible antenna can present problems.

When we are confronted with restrictions—self-imposed or otherwise—we can take advantage of a number of options toward getting on the air and radiating at least a moderately effective signal. In this context, a poor antenna is certainly better than no antenna at all! This section describes a number of techniques that enable us to use indoor antennas or “invisible” antennas outdoors.

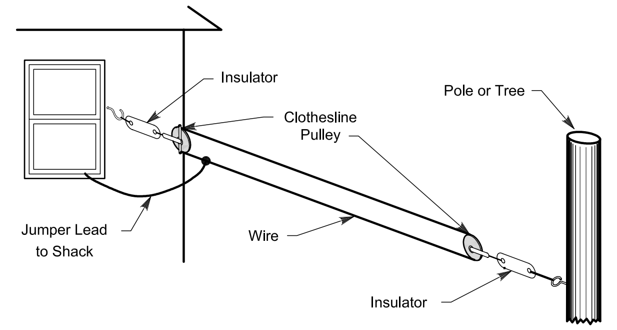

Fig 27—The clothesline antenna is more than it appears to be.

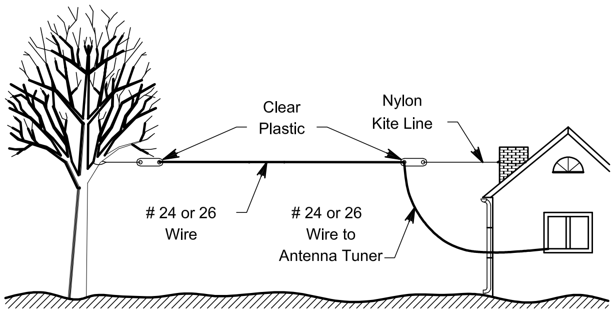

Fig 28—The "invisible" end-fed antenna.

Many of these systems will yield good-to-excellent results for local and DX contacts, depending on band conditions at any given time. *The most important consideration is that of not erecting any antenna that can present a hazard (physical or electrical) to humans, animals and buildings. Safety first!*

Clothesline Antenna

Clotheslines are sometimes attached to pulleys (**Fig 27**) so that the user can load the line and retrieve the laundry from a back porch. Laundry lines of this variety are accepted parts of the neighborhood "scenery," and can be used handily as amateur antennas by simply insulating the pulleys from their support points. This calls for the use of a conducting type of clothesline, such as heavy gauge stranded electrical wire with Teflon or vinyl insulation. A high quality, flexible steel cable (stranded) is suitable as a substitute if you don't mind cleaning it before clothing is hung on it.

A jumper wire can be brought from one end of the line to the ham shack when the station is being operated. If a good electrical connection exists between the wire clothesline and the pulley, a permanent connection can be made by connecting the lead-in wire between the pulley and its insulator. An antenna tuner can be used to match the "invisible" random-length wire to the transmitter and receiver.

Invisible Long Wire

A wire antenna is not actually a "long wire" unless it is one wavelength or greater in length. Yet many amateurs refer to (relatively) long physical spans of conductor as *long wires*. For the purpose of this discussion we will assume we have a fairly long span of wire, and refer to it as an *end-fed* wire antenna.

If we use small-diameter enameled wire for our end-fed antenna, chances are that it will be very difficult to see against the sky and neighborhood scenery. The smaller the wire, the more invisible the antenna will be. The limiting factor with small wire is fragility. A good compromise is #24 or #26 magnet wire for spans up to 130 feet; lighter-gauge wire can be used for shorter spans, such as 30 or 60 feet. The major threat to the longevity of fine wire is icing. Also, birds may fly into the wire and break it. Therefore, this style of antenna may require frequent service or replacement.

Fig 28 illustrates how you might install an invisible end-fed wire. It is important that the insulators also be lacking in prominence. Tiny Plexiglas blocks perform this function well. Small-diameter clear plastic medical vials are suitable also. Some amateurs simply use rubber bands for end insulators, but they will deteriorate rapidly from sun and air pollutants. They are entirely adequate for short-term operation with an invisible antenna, however.

Rain Gutter and TV Antennas

A great number of amateurs have taken advantage of standard house fixtures when contriving inconspicuous antennas. A very old technique is the use of the gutter and downspout system on the building. This is shown in **Fig 29**, where a lead wire is routed to the operating room

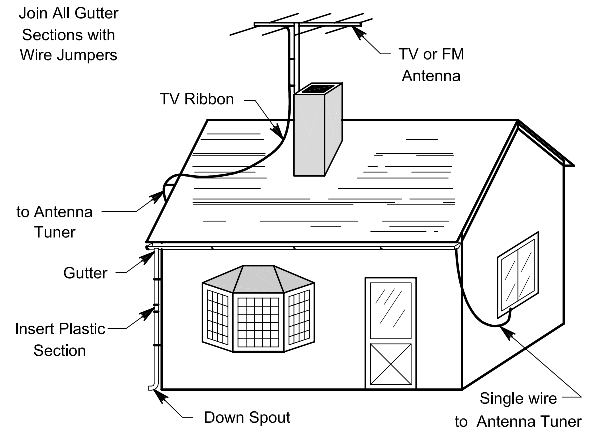

Fig 29—Rain gutters and TV antenna installations can be used as inconspicuous Amateur Radio antennas.

from one end of the gutter trough. We must assume that the wood to which the gutter is affixed is dry and of good quality to provide reasonable electrical insulation. The rain gutter antenna may perform quite poorly during wet weather or when there is ice and snow on it and the house roof.

All joints between gutter and downspout sections must be bonded electrically with straps of braid or flashing copper to provide good continuity in the system. Poor joints can permit rectification of RF and subsequently cause TVI and other harmonic interference. Also, it is prudent to insert a section of plastic downspout about 8 feet above ground to prevent RF shocks or burns to passersby while the antenna is being used. Improved performance may result if you join the front and back gutters of the house with a jumper wire to increase the area of the antenna.

Fig 29 also shows a TV or FM antenna that can be employed as an invisible amateur antenna. Many of these antennas can be modified easily to accommodate the 144- or 222-MHz bands, thereby permitting the use of the 300-Ω line as a feeder system. Some FM antennas can be used on 6 meters by adding #10 bus wire extensions to the ends of the elements, and adjusting the match for an SWR of 1:1. If 300-Ω line is used it will require a balun or antenna tuner to interface the line with the station equipment.

For operation in the HF bands, the TV or FM antenna feeders can be tied together at the transmitter end of the span and the system treated as a random length wire. If this is done, the 300-Ω line will have to be on TV standoff insulators and spaced well away from phone and power company service entrance lines. Naturally, the TV or FM radio must be disconnected from the system when it is used for amateur work! Similarly, masthead amplifiers and splitters must be removed from the line if the system is to be used for amateur operation. If the system is mostly vertical, a good RF ground system with many radials around the base of the house should be used to improve performance.

A very nice top-loaded vertical can be made from a length of TV mast with a large TV antenna on the top. Radials can be placed on the roof or at ground level with the TV “feed line” acting as part of the vertical. There is an extensive discussion of loaded verticals and radial systems in Chapter 6, Low-Frequency Antennas.

Flagpole Antennas

We can exhibit our patriotism and have an invisible amateur antenna at the same time by disguising our antenna as shown in **Fig 30**. The vertical antenna is a wire that has been placed inside a plastic or fiberglass pole.

The flagpole antenna shown is structured for a single amateur band, and it is assumed that the height of the pole corresponds to a quarter wavelength for the chosen band. The radials and feed line can be buried in the ground as shown. In a practical installation, the sealed end of the coax cable would protrude slightly into the lower end of the plastic pole.

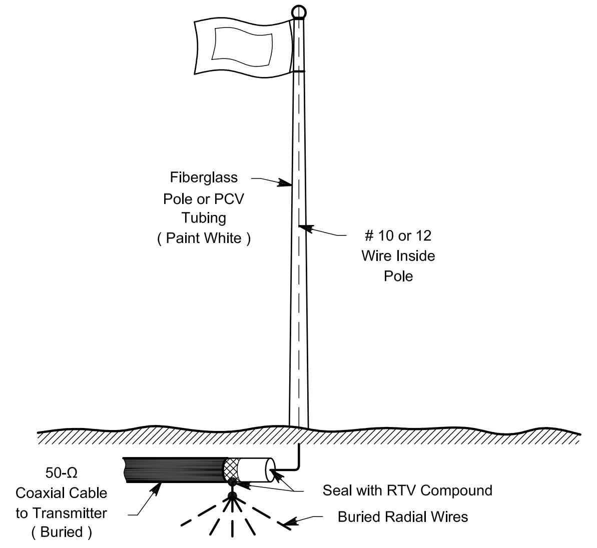

Fig 30—A flagpole antenna.

If a large-diameter fiberglass pole were available, a multiband trap vertical may be concealed inside it. Or you might use a metal pole and bury a water-tight box at its base, containing fixed-tuned matching networks for the bands of interest. The networks could then be selected remotely by means of relays inside the box. A 30-foot flagpole would provide good results in this kind of system, provided it was used in conjunction with a buried radial system.

Still another technique is one that employs a wooden flagpole. A small diameter wire can be stapled to the pole and routed to the coax feeder or matching network. The halyard could by itself constitute the antenna wire if it were made from heavy-duty insulated hookup wire. There are countless variations for this type of antenna, and they are limited only by the imagination of the amateur.

Other Invisible Antennas

Some amateurs have used the metal fence on apartment verandas as antennas, and have had good results on the upper HF bands (14, 21 and 28 MHz). We must presume that the fences were not connected to the steel framework of the building, but rather were insulated by the concrete floor to which they were affixed. These verandah fences have also been used effectively as ground systems (counterpoises) for HF-band vertical antennas put in place temporarily after dark.

One amateur in New York City uses the fire escape on his apartment building as a 7-MHz antenna, and he reports good success working DX stations with it. Another apartment dweller makes use of the aluminum frame on his living room picture window as an antenna for 21 and 28 MHz. He works it against the metal conductors of the baseboard heater in the same room.

Many jokes have been told over the years about *bedspring antennas*. The idea is by no means absurd. Bedsprings and metal end boards have been used to advantage as antennas by many apartment dwellers as 14, 21 and 28 MHz radiators. A counterpoise ground can be routed along the baseboard of the room and used in combination with the bedspring. It is important to remember that any independent (insulated) metal object of reasonable size can serve as an antenna if the transmitter can be matched to it. An amateur in Detroit once used his Shopsmith craft machine (about 5 feet tall) as a 28 MHz antenna. He worked a number of DX stations with it when band conditions were good.

A number of operators have used metal curtain rods and window screens for VHF work, and found them to be acceptable for local communications. Best results with any of these makeshift antennas will be had when the "antennas" are kept well away from house wiring and other conductive objects.

INDOOR ANTENNAS

Without question, the best place for your antenna is outdoors, and as high and in the clear as possible. Some of us, however, for legal, social, neighborhood, family or landlord reasons, are restricted to indoor antennas. Having to settle for an indoor antenna is certainly a handicap for the amateur seeking effective radio communication, but that is not enough reason to abandon all operation in despair.

First, we should be aware of the reasons why indoor antennas do not work well. Principal faults are:

- Low height above ground—the antenna cannot be placed higher than the highest peak of the roof, a point usually low in terms of wavelength at HF
- The antenna must function in a lossy RF environment involving close coupling to electrical wiring, guttering, plumbing and other parasitic conductors, besides dielectric losses in such nonconductors as wood, plaster and masonry
- Sometimes the antenna must be made small in terms of a wavelength
- Usually it cannot be rotated.

These are appreciable handicaps. Nevertheless, global communication with an indoor antenna is still possible, although you must be sure that you are not exposing anyone in your family or nearby neighbors to excessive radiation. See Chapter 1, Safety, in this book.

Some practical points in favor of the indoor antenna include:

- Freedom from weathering effects and damage caused by wind, ice, rain and sunlight (the SWR of an attic antenna, however, can be affected somewhat by a wet or snow-covered roof).
- Indoor antennas can be made from materials that would be altogether impractical outdoors, such as aluminum foil and thread (the antenna need support only its own weight).
- The supporting structure is already in place, eliminating the need for antenna masts.
- The antenna is readily accessible in all weather conditions, simplifying pruning or tuning, which can be accomplished without climbing or tilting over a tower.

Empiricism

A typical house or apartment presents such a complex electromagnetic environment that it is impossible to predict theoretically which location or orientation of the indoor antenna will work best. This is where good old fashioned cut-and-try, use-what-works-best *empiricism* pays off. But to properly determine what really is most suitable requires an understanding of some antenna measuring fundamentals.

Unfortunately, many amateurs do not know how to evaluate performance scientifically or compare one antenna with another. Typically, they will put up one antenna and try it out on the air to see how it "gets out" in comparison with a previous antenna. This is obviously a very poor evaluation method because there is no way to know if the better or worse reports are caused by changing band conditions, different S-meter characteristics or any of several other factors that could influence the reports received.

Many times the difference between two antennas or between two different locations for identical antennas amounts to only a few decibels, a difference that is hard to discern unless instantaneous switching between the two is possible. Those few decibels are not important under strong signal conditions, of course, but when the going gets rough, as is often the case with an indoor antenna, a few dB can make the difference between solid copy and no possibility of real communication.

Very little in the way of test equipment is needed for casual antenna evaluation, other than a communications receiver. You can even do a qualitative comparison by ear, if you can switch antennas instantaneously. Differences of less than 2 dB, however, are still hard to discern. The same is true of S-meters. Signal strength differences of less than a decibel are usually difficult to see. If you want to measure that last fraction of a decibel, you should use a good ac voltmeter at the receiver audio output (with the AGC turned off).

In order to compare two antennas, switching the coaxial transmission line from one to the other is necessary. No elaborate coaxial switch is needed; even a simple double-throw toggle or slide switch will provide more than 40 dB of isolation at HF. See **Fig 31**. Switching by means of manually connecting and disconnecting coaxial lines is not recommended because that takes too long.

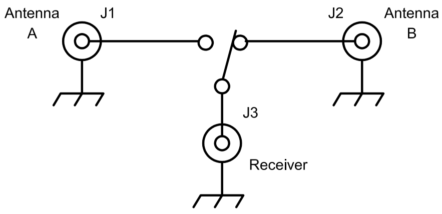

Fig 31—When antennas are compared on fading signals, the time delay involved in disconnecting and reconnecting coaxial cables is too long for accurate measurements. A simple slide switch will do well for switching coaxial lines at HF. The four components can be mounted in a tin can or any small metal box. Leads should be short and direct. J1 through J3 are coaxial connectors.

Fading can cause signal-strength changes during the changeover interval.

Whatever difference shows up in the strength of the received signal will be the difference in performance between the two antennas in the direction of that signal. For this test to be valid, both antennas must have nearly the same feed-point impedance, a condition that is reasonably well met if the SWR is below 2:1 on both antennas.

On ionospheric propagated signals (sky wave) there will be constant fading, and for a valid comparison it will be necessary to take an average of the difference between the two antennas. Occasionally, the inferior antenna will deliver a stronger signal to the receiver, but in the long run the law of averages will put the better antenna ahead.

Of course with a ground-wave signal, such as that from a station across town, there will be no fading problems. A ground-wave signal will enable the operator to properly evaluate the antenna under test in the direction of the source. The results will be valid for ionospheric-propagated signals at low elevation angles in that direction. On 28 MHz, all sky-wave signals arrive and leave at low angles. But on the lower bands, particularly 3.5 and 7 MHz, we often use signals propagated at high elevation angles, almost up to the zenith. For these angles a ground-wave test between local stations may not provide a proper evaluation of the antenna, and use of sky wave signals becomes necessary.

Dipoles

At HF the most practical indoor antenna is usually the dipole. Attempts to get more gain with parasitic elements will usually fail because of close proximity to the ground or coupling to house wiring. Beam antenna dimensions determined outdoors will not usually be valid for an attic antenna because the roof structure will cause dielectric loading of the parasitic elements. It is usually more worthwhile to spend time optimizing the location and performance of a dipole than to try to improve results with parasitic elements.

Most attics are not long enough to accommodate half-wave dipoles for 7 MHz and below. If this is the case, some folding of the dipole will be necessary. The final shape of the antenna will depend on the dimensions and configuration of the attic. Remember that the center of the dipole carries the most current and therefore does most of the radiating. This part should be as high and unfolded as possible. Because the dipole ends radiate less energy than the center, their orientation is not as important. They do carry the maximum voltage, nevertheless, so care should be taken to position the ends far enough from other conductors to avoid arcing.

The dipole may end up being L-shaped, Z-shaped, U-shaped or some indescribable corkscrew shape, depending on what space is available, but reasonable performance can often be had even with such a non-straight arrangement. **Fig 32** shows some possible configurations. Multiband operation is possible with the use of open-wire feeders and an antenna tuner.

One alternative not shown here is the aluminum-foil dipole, which was conceived by Rudy Stork, KA5FSB. He suggests mounting the dipole behind wallpaper or in the attic, with portability, ease of construction and adjustment, and economy in design among its desirable features. This antenna should also display reasonably good bandwidth resulting from the large area of its conductor material. If coaxial feed is used, some pruning of an attic antenna to establish minimum SWR at the band center will be required. Tuning the antenna outdoors and then installing it inside is usually not feasible since the behavior of the antenna will not be the same when placed in the attic. Resonance will be affected somewhat if the antenna is bent.

Even if the antenna is placed in a straight line, parasitic conductors and dielectric loading by nearby wood structures can affect the impedance. Trap and loaded dipoles are shorter than the full-sized versions, but are comparable performers. Trap dipoles are discussed in Chapter 7, Multiband Antennas, and loaded dipoles in Chapter 6, Low-Frequency Antennas.

Dipole Orientation

Theoretically a vertical dipole is most effective at low radiation angles, but practical experience shows that the horizontal dipole is usually a better indoor antenna. A high horizontal dipole does exhibit directional effects at low radiation angles, but you will not be likely to see much, if any, directivity with an attic-mounted dipole. Some operators place two dipoles at right angles to each other with provisions at the operating position for switching between the two. Their reasoning is the radiation patterns will inevitably be distorted in an unpredictable manner by nearby parasitic conductors. There will be little coupling between the dipoles if they are oriented a right

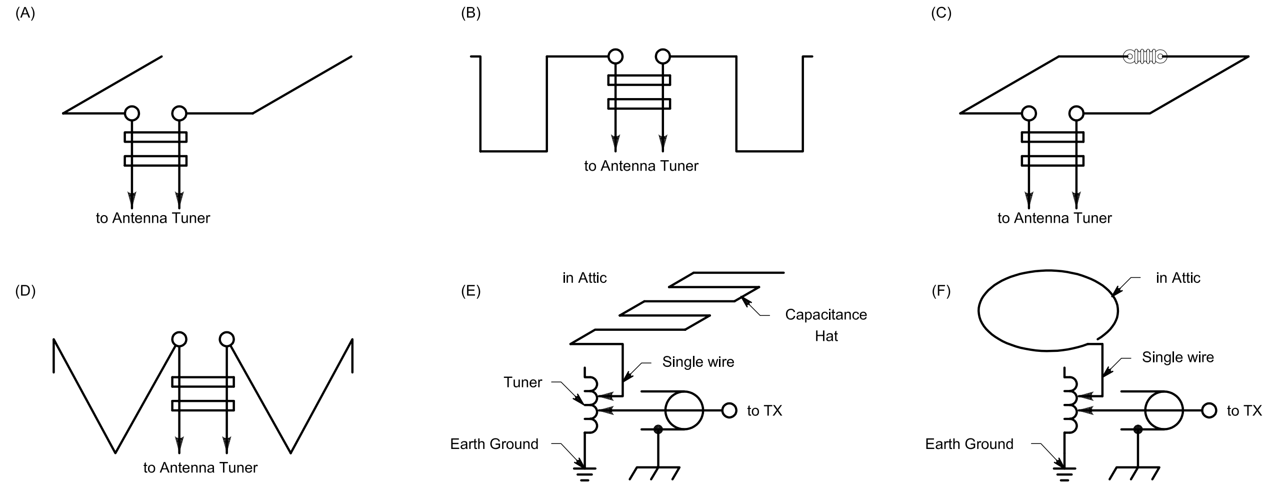
(A)
to Antenna Tuner
(B)
to Antenna Tuner
(C)
to Antenna Tuner
(D)
to Antenna Tuner
(E)
in Attic
Capacitance Hat
Single wire
Tuner
to TX
Earth Ground
(F)
in Attic
Single wire
to TX
Earth Ground

Fig 32—Various configurations for small indoor antennas. See text for discussion.

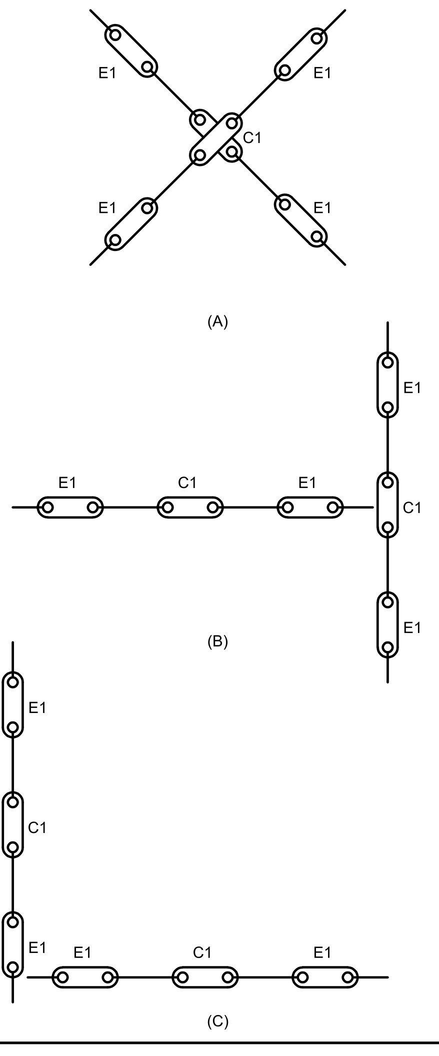
E1
E1
C1
E1
E1
(A)
E1
E1
C1
E1
C1
E1
(B)
E1
C1
E1
E1
C1
E1
(C)

angles to each other as shown in **Figs 33A** and 33B. There will be some coupling with the arrangement shown in Fig 33C, but even this orientation is preferable to a single dipole.

With two antennas mounted 90° apart, you may find that one dipole is consistently better in nearly all directions, in which case you will want to remove the inferior dipole, perhaps placing it someplace else. In this manner the best spots in the house or attic can be determined experimentally.

Parasitic Conductors

Inevitably, any conductor in your house near a quarter wave in length or longer at the operating frequency will be parasitically coupled to your antenna. The word *parasitic* is particularly appropriate in this case because these conductors usually introduce losses and leave less energy for radiation into space. Unlike the parasitic elements in a beam antenna, conductors such as house wiring and plumbing are usually connected to lossy objects such as earth, electrical appliances, masonry or other objects that dissipate energy. Even where this energy is reradiated, it is not likely to be in the right phase in the desired direction; it is, in fact, likely to be a source of RFI.

Fig 33—Ways to orient a pair of perpendicular dipoles. The orientation at A and B will result in no mutual coupling between the two dipoles, but there will be some coupling in the configuration shown at C. End (EI) and center (CI) insulators are shown.

There are, however, some things that can be done about parasitic conductors. The most obvious is to reroute them at right angles to the antenna or close to the ground, or even underground—procedures that are usually not feasible in a finished home. Where these conductors cannot be rerouted, other measures can be taken. Electrical wiring can be broken up with RF chokes to prevent the flow of radio-frequency currents while permitting 60-Hz current (or audio, in the case of telephone wires) to flow unimpeded. A typical RF choke for a power line can be 100 turns of #10 insulated wire close wound on a length of 1-inch diameter plastic pipe. Of course one choke will be needed for each conductor. A three-wire line calls for three chokes. The chokes can be simplified by winding them bifilar or trifilar on a single coil form.

THE RESONANT BREAKER

Obviously, RF chokes cannot be used on conductors such as metal conduit or water pipes. But it is still possible, surprising as it may seem, to obstruct RF currents on such conductors without breaking the metal. The resonant breaker was first described by Fred Brown, W6HPH, in Oct 1979 *QST*.

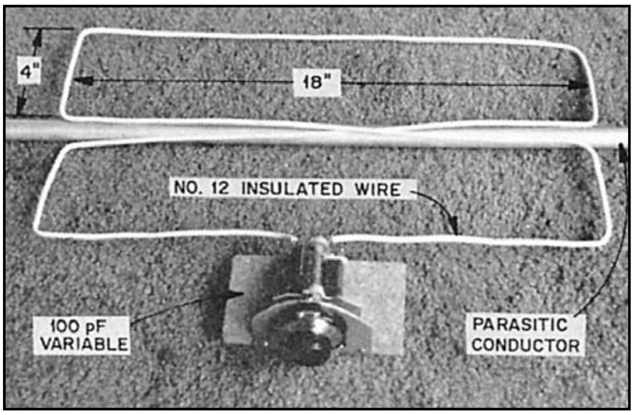

Fig 34—A "resonant breaker" such as shown here can be used to obstruct radio-frequency currents in a conductor without the need to break the conductor physically. A vernier dial is recommended for use with the variable capacitor because tuning is quite sharp. The 100-pF capacitor is in series with the loop. This resonant breaker tunes from 14 through 29.7 MHz. Larger models may be constructed for the lower frequency bands.

Fig 34 shows a method of accomplishing this. A figure-eight loop is inductively coupled to the parasitic conductor and is resonated to the desired frequency with a variable capacitor. The result is a very high impedance induced in series with the pipe, conduit or wire. This impedance will block the flow of radio-frequency currents. The figure-eight coil can be thought of as two turns of an air-core toroid and since the parasitic conductor threads through the hole of this core, there will be tight coupling between the two. Inasmuch as the figure-eight coil is parallel resonated, transformer action will reflect a high impedance in series with the linear conductor.

Before you bother with a *resonant breaker* of this type, be sure that there is a significant amount of RF current flowing in the parasitic conductor, and that you will therefore benefit from installing one. The relative magnitude of this current can be determined with an RF current probe of the type described in Chapter 27, Antenna and Transmission-Line Measurements. According to the rule of thumb regarding parasitic conductor current, if it measures less than $^1/_{10}$ of that measured near the center of the dipole, the parasitic current is generally not large enough to be of concern.

The current probe is also needed for resonating the breaker after it is installed. Normally, the resonant breaker will be placed on the parasitic conductor near the point of maximum current. When it is tuned through resonance, there will be a sharp dip in RF current, as indicated by the current probe. Of course, the resonant breaker will be effective only on one band. You will need one for each band where there is significant current indicated by the probe.

Power-Handling Capability

So far, our discussion have not considered the full power-handling capability of an indoor antenna. Any tendency to flash over must be determined by running full power or, preferably, somewhat more than the peak power you intend to use in regular operation. The antenna should be carefully checked for arcing or RF heating before you do any operating. Bear in mind that attics are indeed vulnerable to fire hazards. A potential of several hundred volts exists at the ends of a dipole fed by the typical Amateur Radio transmitter. If a power amplifier is used, there could be a few thousand volts at the ends of the dipole. Keep your antenna elements well away from other objects. *Safety first*!

Construction Details and Practical Considerations

Ultimately the success of an antenna project depends on the details of how the antenna is fabricated. A great deal of construction information is given in other chapters of this book. For example the construction of HF Yagis is discussed in Chapter 11, Quad arrays in Chapter 12, VHF antennas in Chapter 18, and in Chapter 20 there is an excellent discussion of antenna materials, particularly wire and tubing for elements. Here is still more helpful antenna construction information.

END EFFECT

If the standard expression $\lambda/2 \approx 491.8/f$ (MHz) is used for the length of a $\lambda/2$ wire antenna, the antenna will resonate at a somewhat lower frequency than is desired. The reason is that in addition to the effect of the conductor diameter and ground effects (Chapter 3, The Effects of Ground) an additional "loading" effect is caused by the insulators used at the ends of the wires to support the antenna. The insulators and the wire loops that tie the insulators to the antenna add a small amount of capacitance to the system. This capacitance helps to tune the antenna to a slightly lower frequency, in much the same way that additional capacitance in any tuned circuit lowers the resonant frequency. In an antenna this is called *end effect*. The current at the ends of the antenna does not quite reach zero because of the end effect, as there is some current flowing into the end capacitance. Note that the computations used to create Figs 2 through 7 in Chapter 2, Antenna Fundamentals, did not take into account any end effect.

End effect increases with frequency and varies slightly with different installations. However, at frequencies up to 30 MHz (the frequency range over which wire antennas are most commonly used), experience shows that the length of a practical $\lambda/2$ antenna, including the effect of diameter and end effect, is on the order of 5% less than the length of a half wave in space. As an average, then, the physical length of a resonant $\lambda/2$ wire antenna can be found from:

$$\lambda = \frac{491.8 \times 0.95}{\text{f (MHz)}} \approx \frac{468}{\text{f (MHz)}} \qquad \text{(Eq 1)}$$

Eq 1 is reasonably accurate for finding the physical length of a $\lambda/2$ antenna for a given frequency, but does not apply to antennas longer than a half wave in length. In the practical case, if the antenna length must be adjusted to exact frequency (not all antenna systems require it) the length should be "pruned" to resonance. Note that the use of plastic-insulated wire will typically lower the resonant frequency of a half-wave dipole about 3%.

INSULATORS

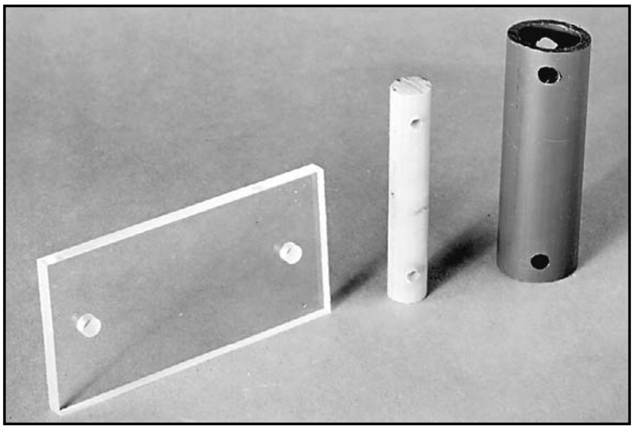

Fig 35—Some ideas for homemade antenna insulators.

Wire antennas must be insulated at the ends. Commercially available insulators are made from ceramic, glass or plastic. Insulators are available from many Amateur Radio dealers. RadioShack and local hardware stores are other possible sources. Acceptable homemade insulators may be fashioned from a variety of material including (but not limited to) acrylic sheet or rod, PVC tubing, wood, fiberglass rod or even stiff plastic from a discarded container. **Fig 35** shows some homemade insulators. Ceramic or glass insulators will usually outlast the wire, so they are highly recommended for a safe, reliable, permanent installation. Other materials may tear under stress or break down in the presence of sunlight. Many types of plastic do not weather well.

INSTALLING TRANSMISSION LINES

Many wire antennas require an insulator at the feed point. Although there are many ways to connect the feed line, there are a few things to keep in mind. If you feed your antenna with coaxial cable, you have two choices. You can install an SO-239 connector on the center insu-

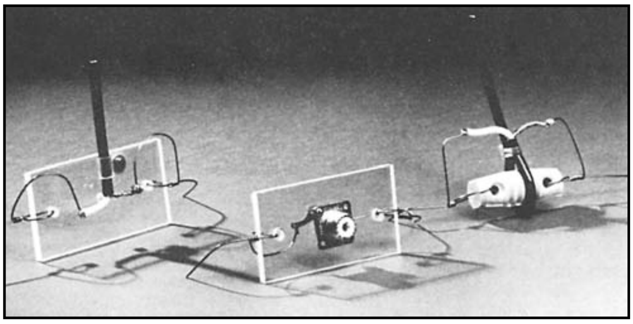

Fig 36—Some homemade dipole center insulators. The one in the center includes a built-in SO-239 connector. Others are designed for direct connection to the feed line.

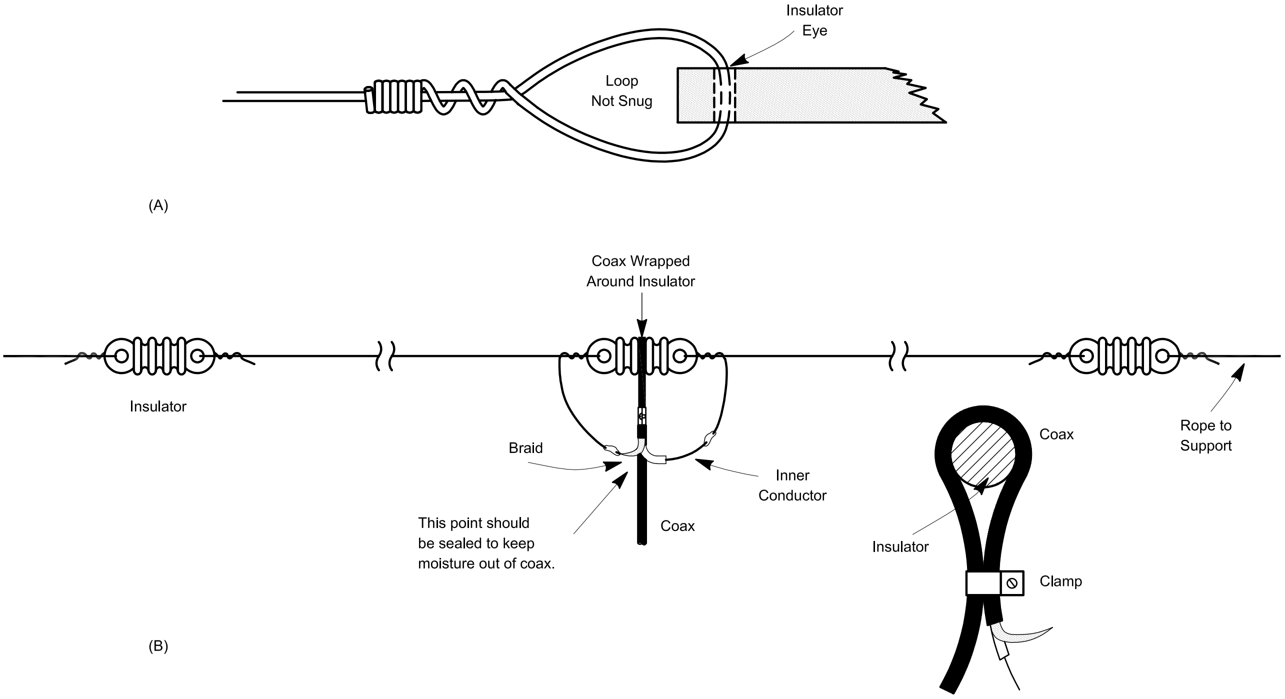
Insulator
Eye
Loop
Not Snug
(A)
Coax Wrapped
Around Insulator
Insulator
Braid
This point should
be sealed to keep
moisture out of coax.
Coax
Inner
Conductor
Coax
Rope to
Support
Insulator
Clamp
(B)

Fig 37—Details of dipole antenna construction. At A, the end insulator connection is shown. At B, the completed antenna is shown. A balun (not shown) is often used at the feed point, since this is a balanced antenna.

lator, as shown by the center example in **Fig 36,** and use a PL-259 on the end of your coax, or you can separate the center conductor from the braid and connect the feed line directly to the antenna wire as shown in the other two examples in Fig 36 and the example in **Fig 37**. Although it costs less to connect direct, the use of connectors offers several advantages. Coaxial cable braid soaks up water like a sponge unless it is very well waterproofed. If you do not adequately seal the antenna end of the feed line, water will find its way into the braid. Water in the feed line will lead to contamination, rendering the coax useless long before its normal lifetime is up. Many hams waterproof the coax, first with vinyl electrical tape, and then using a paint-on material called “PlastiDip,” which is sold by RadioShack (part number 910-5166 for the white variety).

It is not uncommon for water to drip from the end of the coax inside the shack after a year or so of service if the antenna connection is not properly waterproofed. Use of a PL-259/SO-239 combination (or connector of your choice) makes the task of waterproofing connections much easier. Another advantage to using the PL-259/SO-239 combination is that feed-line replacement is much easier, should that become necessary.

Whether you use coaxial cable, ladder line, or twin lead to feed your antenna, an often overlooked consideration is the mechanical strength of the connection. Wire antennas and feed lines tend to move a lot in the breeze, and

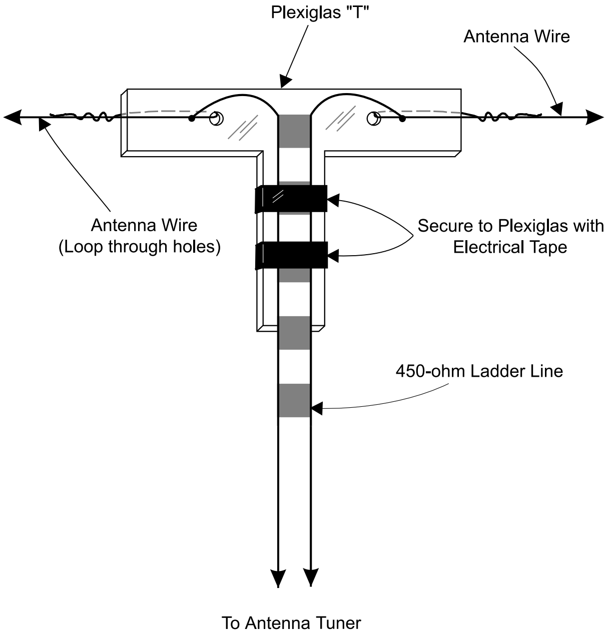
Plexiglas "T"
Antenna Wire
Antenna Wire
(Loop through holes)
Secure to Plexiglas with
Electrical Tape
450-ohm Ladder Line
To Antenna Tuner

Fig 38—A piece of cut Plexiglas can be used as a center insulator and to support a ladder-line feeder. The Plexiglas acts to reduce the flexing of the wires where they connect to the antenna. Use thick Plexiglas in areas subject to high winds.

unless the feed line is attached securely, the connection will weaken with time. The resulting failure can range from a frustrating intermittent electrical connection to a complete separation of feed line and antenna. Fig 37 and **Fig 38** illustrate different ways of attaching either coax or ladder line to the antenna securely.

When open-wire feed line is used, the conductors of the line should be anchored to the insulator by threading them through the eyes of the insulator two or three times, and twisting the wire back on itself before soldering. A slack tie wire should then be used between the feeder conductor and the antenna, as shown in Fig 38. (The tie wires may be extensions of the line conductors themselves.) When window-type line is suspended from an antenna in a manner such as that shown in Fig 38, the line should be twisted—at several twists per foot—to prevent stress hardening of the wire because of constant flexing in the wind.

When using plastic-insulated open-wire line, the tendency of the line to twist and short out close to the antenna can be counteracted by making the center insulator of the antenna longer than the spacing of the line, as shown in Fig 38. In severe wind areas, it may be necessary to use ¼-inch thick Plexiglas for the center insulator rather than thinner material.

RUNNING THE FEED LINE FROM THE ANTENNA TO THE STATION

Chapter 24, Transmission Lines, contains some general guidelines for installing feed lines. More detailed information is contained in this section. Whenever possible, the transmission line should be lead away from the antenna at a 90° angle to minimize coupling from the antenna to the transmission line. This coupling can cause unequal currents on the transmission line, which will then radiate and it can detune the antenna.

Except for the portion of the line in close proximity to the antenna, coaxial cable requires no particular care in running from the antenna to the station entrance, other than protection from mechanical damage. If the antenna is not supported at the center, the line should be fastened to a post more than head high located under the center of the antenna, allowing enough slack between the post and the antenna to take care of any movement of the antenna in the wind. If the antenna feed point is supported by a tower or mast, the cable can be taped to the mast at intervals or to one leg of the tower.

Coaxial cable rated for direct burial can be buried a few inches in the ground to make the run from the antenna to the station. A deep slit can be cut by pushing a square-end spade full depth into the ground and moving the handle back and forth to widen the slit before removing the spade. After the cable has been pushed into the slit with a piece of 1-inch board 3 or 4 inches wide, the slit can be tamped closed. Many hams run coax cables through PVC pipe buried in the ground deeper than the frost line and slanted downwards slightly so that water will drain, rather than pooling inside the length of the pipe.

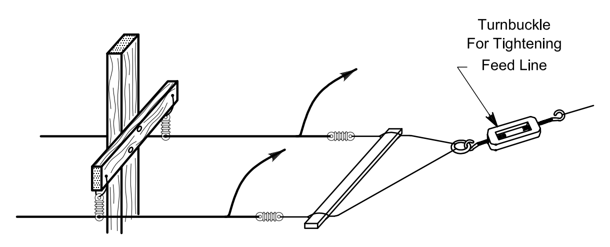

Fig 39—A support for open-wire line. The support at the antenna end of the line must be sufficiently rigid to stand the tension of the line.

Solid ribbon or the newer *window* types of line should be kept reasonably well spaced from other conductors running parallel to it for more than a few feet. TV-type standoff insulators with strap clamp mountings can be used for running this type of line down a mast or tower leg. Similar insulators of the screw-in type can be used in supporting the line on wooden poles for a long run.

Open-wire lines with bare conductors require frequent supports to keep the lines from twisting and shorting out, as well as to relieve the mechanical strain. One method of supporting a long horizontal run of heavy open-wire line is shown in **Fig 39**. The line must be anchored securely at a point under the feed point of the antenna. Window-type line can be supported similarly with wire links fastened to the insulators or with black cable ties (ones not affected by UV radiation from the sun).

To keep the line clear of pedestrians and vehicles, it is usually desirable to anchor the feed line at the eaves or rafter line of the station building (see **Fig 40**), and then drop it vertically to the point of entrance. The points of anchorage and entrance should be chosen to permit the vertical drop without crossing windows for aesthetic reasons.

If the station is located in a room on the ground floor, one way of bringing coax transmission line into the house is to go through the outside wall below floor level, feed it through the basement or crawl space and then up to the station through a hole in the floor. When making the entrance hole in the side of the building, suitable measurements should be made in advance to be sure the hole will go through the sill 2 or 3 inches above the foundation line (and between joists if the bore is parallel to the joists). The line should be allowed to sag below the entrance hole level outside the building to allow rain water to drip off.

Open-wire line can be fed in a similar manner, although it will require a separate hole for each conductor. Each hole should be insulated with a length of polystyrene or Lucite tubing. If available, ceramic tubes salvaged from old-fashioned *knob and tube* electrical

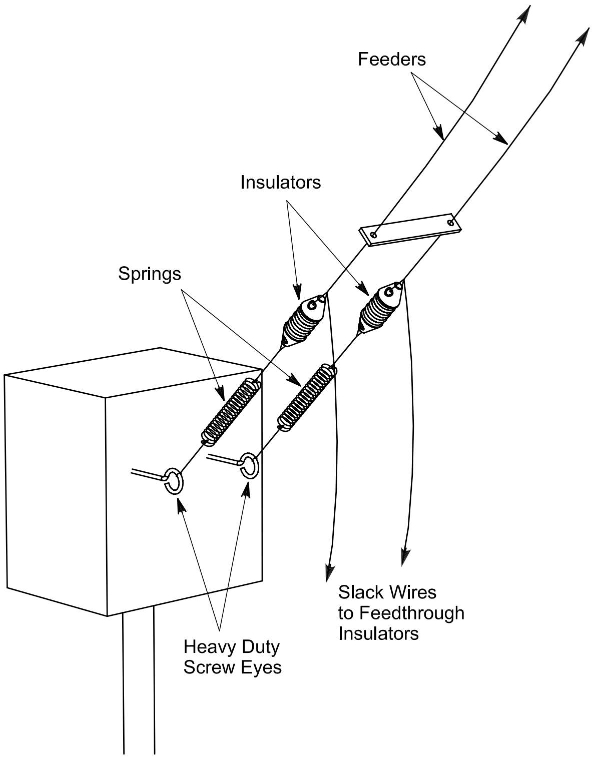

Fig 40—Anchoring open-wire line at the station end. The springs are especially desirable if the line is not supported between the antenna and the anchoring point.

installations, work very well for this purpose. Drill the holes with a slight downward slant toward the outside of the building to prevent rain seepage. With window ladder line, it will be necessary to remove a few of the spreader insulators, cut the line before passing through the holes (allowing enough length to reach the inside) and splice the remainder on the inside.

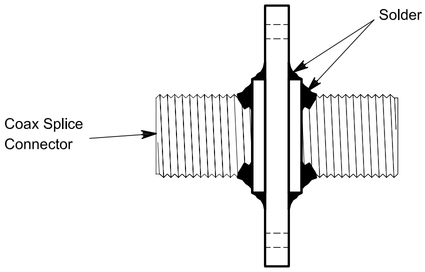

Fig 42—Feedthrough connector for coax line. An Amphenol 83-1J (PL-258) connector, the type used to splice sections of coax line together, is soldered into a hole cut in a brass mounting flange. An Amphenol bulkhead adapter 83-1F may be used instead.

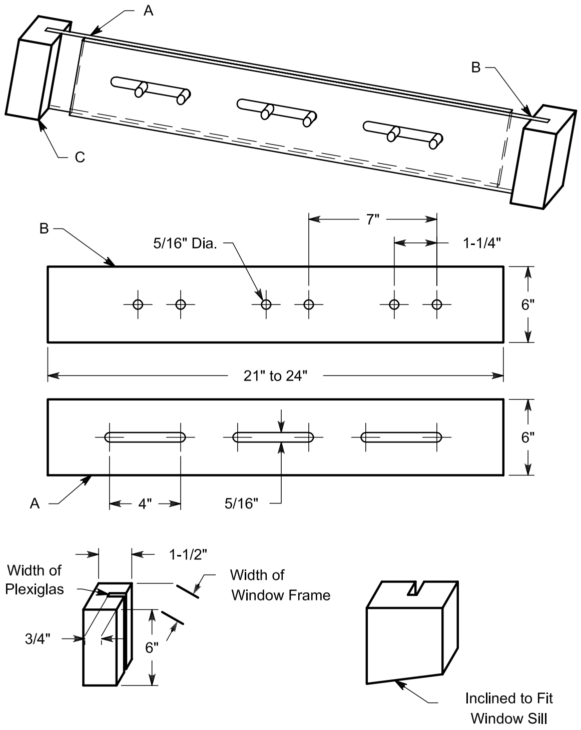

Fig 41—An adjustable window lead-in panel made up of two sheets of Lucite or Plexiglas. A feedthrough connector for coax line can be made as shown in Fig 28. Ceramic feedthrough insulators are suitable for open-wire line. (W1RVE)

If the station is located above ground level, or if there is other objection to the procedure described above, entrance can be made at a window, using the arrangement shown in **Fig 41**. An Amphenol type 83-1F (UG-363) connector can be used as shown in **Fig 42**; ceramic feedthrough insulators can be used for open-wire line. Ribbon line can be run through clearance holes in the panel, and secured by a winding of tape on either side of the panel, or by cutting the retaining rings and insulators from a pair of TV standoff insulators and clamping one on each side of the panel.

LIGHTNING PROTECTION

Two or three types of lightning arresters for coaxial cable are available on the market. If the antenna feed point is at the top of a well-grounded tower, the arrester can be fastened securely to the top of the tower for grounding purposes. A short length of cable, terminated in a coaxial plug, is then run from the antenna feed point to one receptacle of the arrester, while the transmission line is run from the other arrester receptacle to the station. Such arresters may also be placed at the entrance point to the station, if a suitable ground connection is available at that point (or arresters may be placed at both points for added insurance).

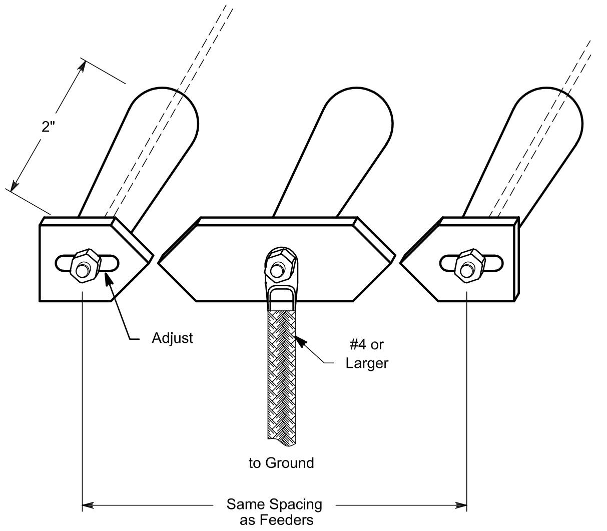

Fig 43—A simple lightning arrester for open-wire line made from three standoff or feedthrough insulators and sections of 1/8 × 1/2-inch brass or copper strap. It should be installed in the line at the point where the line enters the station. The heavy ground lead should be as short and as direct as possible. The gap setting should be adjusted to the minimum width that will prohibit arcing when the transmitter is operated.

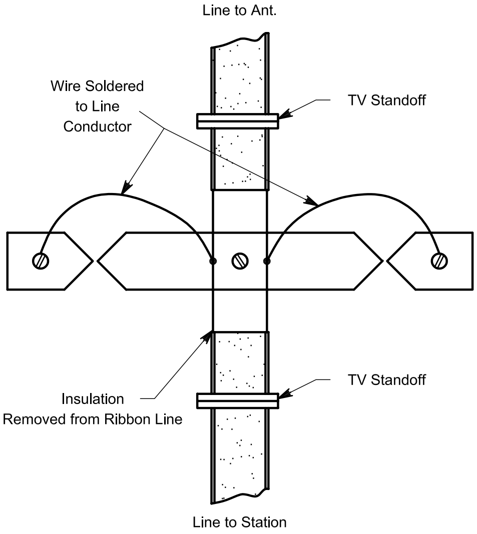
Line to Ant.
Wire Soldered to Line Conductor
TV Standoff
TV Standoff
Insulation Removed from Ribbon Line
Line to Station

Fig 44—The lightning arrester of Fig 39 may be used with 300-Ω ribbon line in the manner shown here. The TV standoffs support the line an inch or so away from the grounded center member of the arrester

The construction of a homemade arrester for open-wire line is shown in **Fig 43**. This type of arrester can be adapted to ribbon line an inch or so away from the center member of the arrester, as shown in **Fig 44**. Sufficient insulation should be removed from the line where it crosses the arrester to permit soldering the arrester connecting leads.

Lightning Grounds

Lightning-ground connecting leads should be of conductor size equivalent to at least #10 wire. The #8 aluminum wire used for TV-antenna grounds is satisfactory. Copper braid 3/4-inch wide (Belden 8662-10) is also suitable. The conductor should run in a straight line to the grounding point. The ground connection may be made to a water pipe system (if the pipe is not plastic), the grounded metal frame of a building, or to one or more 5/8-inch ground rods driven to a depth of at least 8 feet. More detailed information on lightning protection is contained in Chapter 1, Safety.

A central grounding panel for coax cables coming into the house is highly recommended. See **Fig 45** for a photo of the homemade grounding panel installed by Chuck Hutchinson, K8CH, at his Michigan home. The coax cables screwed into dual-female feed-through UHF connectors. K8CH installed this aluminum panel under the outside grill for a duct that provided combustion air to an unused fireplace. He used ground strap to connect to ground rods located under the panel. See the ARRL book *Simple and Fun Antennas* for more information about ground panels.

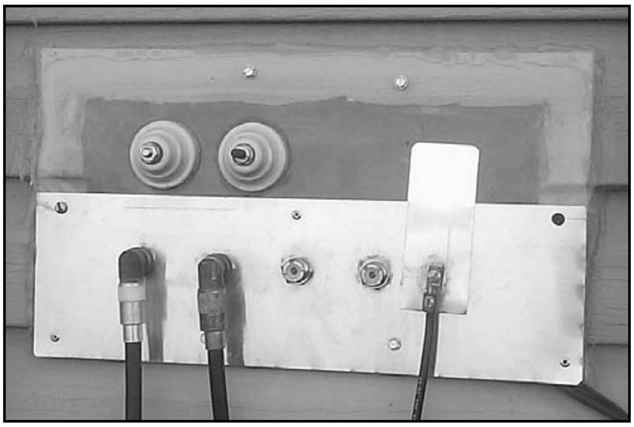

Fig 45—K8CH's coax entry panel mounted on exterior wall (later covered by grill that provides combustion to an unused fireplace). The ground braid goes to a ground rod located beneath the panel. (Photo courtesy: *Simple and Fun Antennas for Hams*)

Before a lightning storm approaches, a prudent ham will disconnect all feed lines, rotor lines and control lines inside the shack to prevent damage to sensitive electronics. When lightning is crashing about outside, you certainly don't want that lightning inside your shack!

Chapter 5

Loop Antennas

A loop antenna is a closed-circuit antenna—that is, one in which a conductor is formed into one or more turns so its two ends are close together. Loops can be divided into two general classes, those in which both the total conductor length and the maximum linear dimension of a turn are very small compared with the wavelength, and those in which both the conductor length and the loop dimensions begin to be comparable with the wavelength.

A "small" loop can be considered to be simply a rather large coil, and the current distribution in such a loop is the same as in a coil. That is, the current has the same phase and the same amplitude in every part of the loop. To meet this condition, the total length of conductor in the loop must not exceed about 0.1 λ. Small loops are discussed later in this chapter, and further in Chapter 14, Direction Finding Antennas.

A "large" loop is one in which the current is not the same either in amplitude or phase in every part of the loop. This change in current distribution gives rise to entirely different properties compared with a small loop.

Half-Wave Loops

The smallest size of "large" loop generally used is one having a conductor length of $^1/_2$ λ. The conductor is usually formed into a square, as shown in **Fig 1**, making each side $^1/_8$ λ long. When fed at the center of one side, the current flows in a closed loop as shown in Fig 1A. The current distribution is approximately the same as on a $^1/_2$ λ wire, and so is maximum at the center of the side opposite the terminals X-Y, and minimum at the terminals themselves. This current distribution causes the field strength to be maximum in the plane of the loop and in the direction looking from the low-current side to the high-current side. If the side opposite the terminals is opened at the center as shown in Fig 1B (strictly speaking, it is then no longer a loop because it is no longer a closed circuit), the direction of current flow remains unchanged but the maximum current flow occurs at the terminals. This reverses the direction of maximum radiation.

The radiation resistance at a current antinode (which is also the resistance at X-Y in Fig 1B) is on the order of 50 Ω. The impedance at the terminals in Fig 1A is a few thousand ohms. This can be reduced by using two identical loops side by side with a few inches spacing between them and applying power between terminal X on one loop and terminal Y on the other.

Unlike a $^1/_2$ λ dipole or a small loop, there is no direction in which the radiation from a loop of the type shown in Fig 1 is zero. There is appreciable radiation in the direction perpendicular to the plane of the loop, as well as to the "rear"—the opposite direction to the arrows shown. The front-to-back (F/B) ratio is approximately 4 to 6 dB. The small size and the shape of the directive pattern result in a loss of about 1 dB when the field strength in the optimum direction from such a

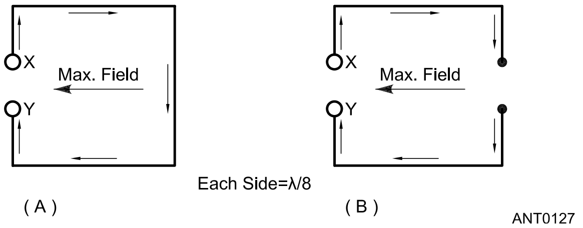

Fig 1—Half-wave loops, consisting of a single turn having a total length of $^1/_2$ λ.

loop is compared with the field from a $^1/_2$ λ dipole in its optimum direction.

The ratio of the forward radiation to the backward radiation can be increased, and the field strength likewise increased at the same time to give a gain of about 1 dB over a dipole, by using inductive reactances to "load" the sides joining the front and back of the loop. This is shown in **Fig 2**. The reactances, which should have a value of approximately 360 Ω, decrease the current in the sides in which they are inserted and increase it in the side having terminals. This increases the directivity and thus increases the efficiency of the loop as a radiator. Lossy coils can reduce this advantage greatly.

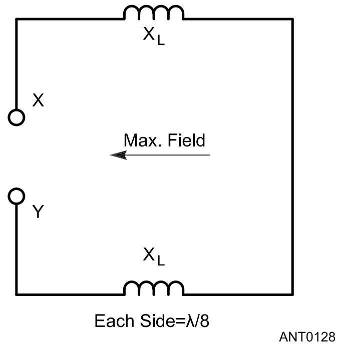

Fig 2—Inductive loading in the sides of a $^1/_2$-λ loop to increase the directivity and gain. Maximum radiation or response is in the plane of the loop, in the direction shown by the arrow.

One-Wavelength Loops

Loops in which the conductor length is 1 λ have different characteristics than $^1/_2$-λ loops. Three forms of 1 λ loops are shown in **Fig 3**. At A and B the sides of the squares are equal to $^1/_4$ λ, the difference being in the point at which the terminals are inserted. At C the sides of the triangle are equal to $^1/_3$ λ. The relative direction of current flow is as shown in the drawings. This direction reverses halfway around the perimeter of the loop, as such reversals always occur at the junction of each $^1/_2$-λ section of wire.

The directional characteristics of loops of this type are opposite in sense to those of a small loop. That is, the radiation is maximum perpendicular to the plane of the loop and is minimum in either direction in the plane containing the loop. If the three loops shown in Fig 3 are mounted in a vertical plane with the terminals at the bottom, the radiation is horizontally polarized. When the terminals are moved to the center of one vertical side in Fig 3A, or to a side corner in B, the radiation is vertically polarized. If the terminals are moved to a side corner in C, the polarization will be diagonal, containing both vertical and horizontal components.

In contrast to straight-wire antennas, the electrical length of the circumference of a 1-λ loop is shorter than the actual length. For a loop made of bare #18 wire and operating at a frequency of 14 MHz, where the ratio of conductor length to wire diameter is large, the loop will be close to resonance when

$$\text{Length}_{\text{feet}} = \frac{1032}{f_{\text{MHz}}}$$

The radiation resistance of a resonant 1 λ loop is approximately 120 Ω, under these conditions. Since the loop dimensions are larger than those of a $^1/_2$-λ dipole, the radiation efficiency is high.

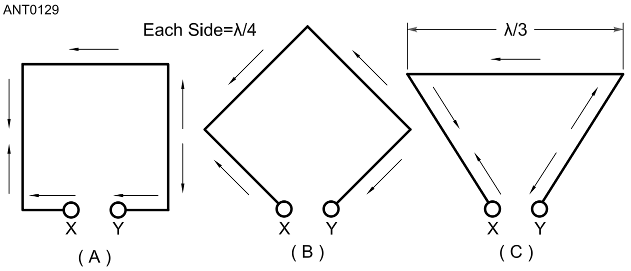

Fig 3—At A and B, loops having sides $^1/_4$ λ long, and at C having sides $^1/_3$ λ long (total conductor length 1 λ). The polarization depends on the orientation of the loop and on the position of the feed point (terminals X-Y) around the perimeter of the loop.

In the direction of maximum radiation (that is, broadside to the plane of the loop, regardless of the point at which it is fed) the 1-λ loop will show a small gain over a $^1/_2$-λ dipole. Theoretically, this gain is about 1 dB, and measurements have confirmed that it is of this order.

The 1-λ loop is more frequently used as an element of a directive antenna array (the quad and delta-loop antennas described in Chapter 12, Quad Arrays) than singly, although there is no reason why it cannot be used alone. In the quad and delta loop, it is nearly always driven so that the polarization is horizontal.

Small Loop Antennas

The electrically small loop antenna has existed in various forms for many years. Probably the most familiar form of this antenna is the ferrite *loopstick* found in portable AM broadcast-band receivers. Amateur applications of the small loop include direction finding, low-noise directional receiving antennas for 1.8 and 3.5 MHz, and small transmitting antennas. Because the design of transmitting and receiving loops requires some different considerations, the two situations are examined separately in this section. This information was written by Domenic M. Mallozzi, N1DM.

The Basic Loop

What is and what is not a small loop antenna? By definition, the loop is considered to be electrically small when its total conductor length is less than 0.1 λ—0.085 is the number used in this section. This size is based on the fact that the current around the perimeter of the loop must be in phase. When the winding conductor is more than about 0.085 λ long, this is no longer true. This constraint results in a very predictable figure-eight radiation pattern, shown in **Fig 4**.

The simplest loop is a 1-turn untuned loop with a load connected to a pair of terminals located in the center of one of the sides, as shown in **Fig 5**. How its pattern is developed is easily pictured if we look at some "snapshots" of the antenna relative to a signal source. **Fig 6** represents a loop from above, and shows the instantaneous radiated voltage wave. Note that points A and B of the loop are receiving the same instantaneous voltage. This means that no current will flow through the loop, because there is no current flow between points of equal potential. A similar analysis of **Fig 7**, with the loop turned 90° from the position represented in Fig 6, shows that this position of the loop provides maximum response. Of course, the voltage derived from the passing wave is small because of the small physical size of the loop. Fig 4 shows the ideal radiation pattern for a small loop.

The voltage across the loop terminals is given by

$$V = \frac{2 \pi A N E \cos \theta}{\lambda} \qquad \text{(Eq 1)}$$

where

V = voltage across the loop terminals
A = area of loop in square meters
N = number of turns in the loop
E = RF field strength in volts per meter
θ = angle between the plane of the loop and the signal source (transmitting station)
λ = wavelength of operation in meters

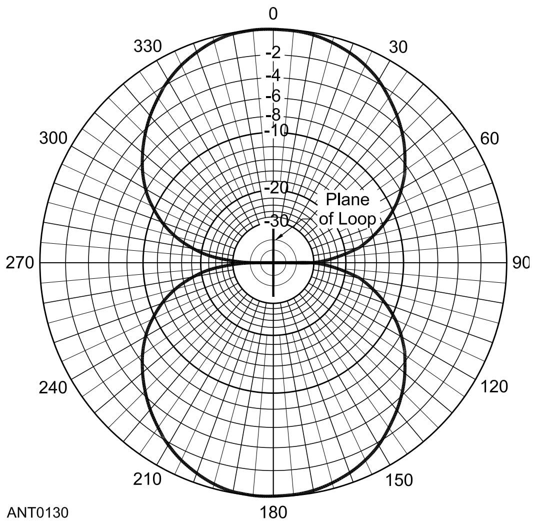

Fig 4—Calculated small loop antenna radiation pattern.

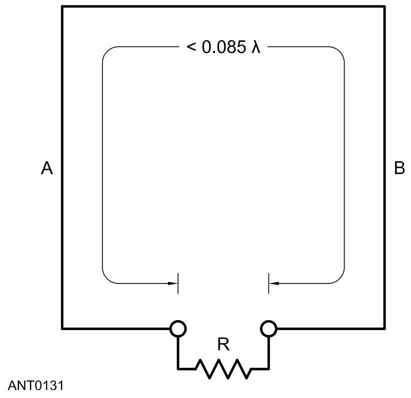

Fig 5—Simple untuned small loop antenna.

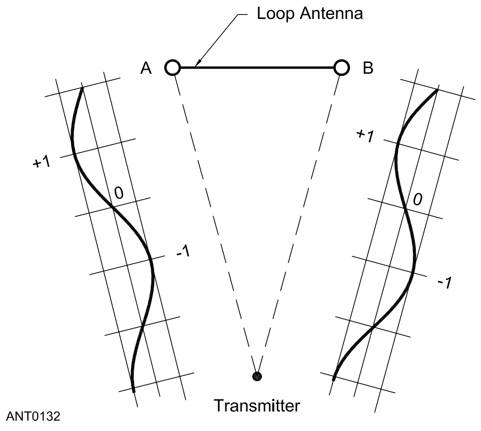

Fig 6—Example of orientation of loop antenna that does not respond to a signal source (null in pattern).

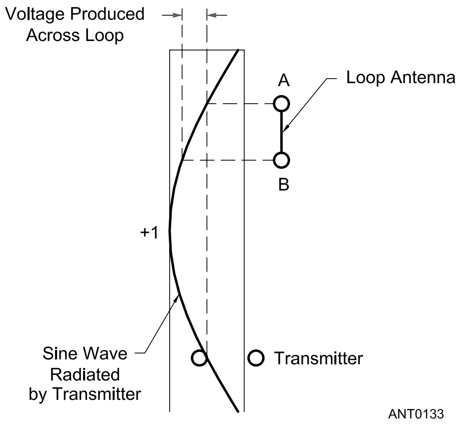

Fig 7—Example of orientation of loop antenna for maximum response.

This equation comes from a term called *effective height*. The effective height refers to the height (length) of a vertical piece of wire above ground that would deliver the same voltage to the receiver. The equation for effective height is

$$h = \frac{2\pi N A}{\lambda} \qquad \text{(Eq 2)}$$

where h is in meters and the other terms are as for Eq 1.

A few minutes with a calculator will show that, with the constraints previously stated, the loop antenna will have a very small effective height. This means it will deliver a relatively small voltage to the receiver, even with a large transmitted signal.

TUNED LOOPS

We can tune the loop by placing a capacitor across the antenna terminals. This causes a larger voltage to appear across the loop terminals because of the Q of the parallel resonant circuit that is formed.

The voltage across the loop terminals is now given by

$$V = \frac{2\pi A N E Q \cos\theta}{\lambda} \qquad \text{(Eq 3)}$$

where Q is the loaded Q of the tuned circuit, and the other terms are as defined above.

Most amateur loops are of the tuned variety. For this reason, all comments that follow are based on tuned-loop antennas, consisting of one or more turns. The tuned-loop antenna has some particular advantages. For example, it puts high selectivity up at the “front” of a receiving system, where it can significantly help factors such as dynamic range. Loaded Q values of 100 or greater are easy to obtain with careful loop construction.

Consider a situation where the inherent selectivity of the loop is helpful. Assume we have a loop with a Q of 100 at 1.805 MHz. We are working a DX station on 1.805 MHz and are suffering strong interference from a local station 10 kHz away. Switching from a dipole to a small loop will reduce the strength of the off-frequency signal by 6 dB (approximately one S unit). This, in effect, increases the dynamic range of the receiver. In fact, if the off-frequency station were further off frequency, the attenuation would be greater.

Another way the loop can help is by using the nulls in its pattern to null out on-frequency (or slightly off-frequency) interference. For example, say we are working a DX station to the north, and just 1 kHz away is another local station engaged in a contact. The local station is to our west. We can simply rotate our loop to put its null to the west, and now the DX station should be readable while the local will be knocked down by 60 or more dB. This obviously is quite a noticeable difference. Loop nulls are very sharp and are generally noticeable only on ground-wave signals (more on this later).

Of course, this method of nulling will be effective only if the interfering station and the station being worked are not in the same direction (or in exact opposite directions) from our location. If the two stations were on the same line from our location, both the station being worked and the undesired station would be nulled out. Luckily the nulls are very sharp, so as long as the stations are at least 10° off axis from each other, the loop null will be usable.

A similar use of the nulling capability is to eliminate local noise interference, such as that from a light dimmer in a neighbor’s house. Just put the null on the offending light dimmer, and the noise should disappear.

Now that we have seen some possible uses of the small loop, let us look at a bit of detail about its design. First, the loop forms an inductor having a very small ratio of winding length to diameter. The equations for finding inductance given in most radio handbooks assume that the inductor coil is longer than its diameter. However, F. W. Grover of the US National Bureau of Standards has provided equations for inductors of common cross-sectional shapes and small length-to-diameter ratios. (See the Bibliography at the end of this chapter.) Grover’s equations are shown in **Table 1**. Their use will yield relatively accurate numbers; results are easily worked out with a scientific calculator or home computer.

The value of a tuning capacitor for a loop is easy to calculate from the standard resonance equations. The only matter to consider before calculating this is the value of distributed capacitance of the loop winding. This capacitance shows up between adjacent turns of the coil because of their slight difference in potential. This causes each turn to appear as a charge plate. As with all other capacitances, the value of the distributed capacitance is based on the physical dimensions of the coil. An exact mathematical analysis of its value is a complex problem. A simple approximation is given by Medhurst (see Bibliography) as:

$$C = HD \qquad \text{(Eq 4)}$$

where

C = distributed capacitance in pF

H = a constant related to the length-to-diameter ratio of the coil (**Table 2** gives H values for length-to-diameter ratios used in loop antenna work.)

D = diameter of the winding in cm

Medhurst's work was with coils of round cross section. For loops of square cross section the distributed capacitance is given by Bramslev (see Bibliography) as

$$C = 60S \qquad \text{(Eq 5)}$$

where

C = the distributed capacitance in pF

S = the length of the side in meters

If you convert the length in this equation to centimeters, you will find Bramslev's equation gives results in the same order of magnitude as Medhurst's equation.

This distributed capacitance appears as if it were a capacitor across the loop terminals. Therefore, when determining the value of the tuning capacitor, the distributed capacitance must be subtracted from the total capacitance required to resonate the loop. The distributed capacitance also determines the highest frequency at which a particular loop can be used, because it is the minimum capacitance obtainable.

Electrostatically Shielded Loops

Over the years, many loop antennas have incorporated an electrostatic shield. This shield generally takes the form of a tube around the winding, made of a conductive but non-magnetic material (such as copper or aluminum). Its purpose is to maintain loop balance with respect to ground, by forcing the capacitance between all portions of the loop and ground to be identical. This is illustrated in **Fig 8**. It is necessary to maintain electrical loop balance to eliminate what is referred to as the *antenna effect*. When the antenna becomes unbalanced it appears to act partially as a small vertical antenna. This vertical pattern gets superimposed on the ideal figure-eight pattern, distorting the pattern and filling in the nulls. The type of pattern that results is shown in **Fig 9**.

Adding the shield has the effect of somewhat reducing the pickup of the loop, but this loss is generally offset by the increase in null depth of the loops. Proper balance of the loop antenna requires that the load on the loop also be balanced. This is usually accomplished by use of a balun transformer or a balanced input preamplifier. One important point regarding the shield is that it cannot form a continuous electrical path around the loop perimeter, or it will appear as a shorted coil turn. Usually the insulated break is located opposite the feed point to maintain symmetry. Another point to be considered is that the shield should be of a much larger diameter than the loop winding, or it will lower the Q of the loop.

Various construction techniques have been used in

Table 1

Inductance Equations for Short Coils (Loop Antennas)

Triangle:

$$L\left(\mu H\right) = 0.006N^2\,s\left[\ln\left(\frac{1.1547\,sN}{\left(N+1\right)\ell}\right) + 0.65533 + \frac{0.1348\left(N+1\right)\ell}{sN}\right]$$

Square:

$$L\left(\mu H\right) = 0.008N^2\,s\left[\ln\left(\frac{1.4142sN}{\left(N+1\right)\ell}\right) + 0.37942 + \frac{0.3333\left(N+1\right)\ell}{sN}\right]$$

Hexagon:

$$L\left(\mu H\right) = 0.012N^2\,s\left[\ln\left(\frac{2sN}{\left(N+1\right)\ell}\right) + 0.65533 + \frac{0.1348\left(N+1\right)\ell}{sN}\right]$$

Octagon:

$$L\left(\mu H\right) = 0.016N^2\left[\ln\left(\frac{2.613sN}{\left(N+1\right)\ell}\right) + 0.75143 + \frac{0.07153\left(N+1\right)\ell}{sN}\right]$$

where

N = number of turns

s = side length in cm

ℓ = coil length in cm

Note: In the case of single-turn coils, the diameter of the conductor should be used for ℓ.

Table 2

Values of the Constant H for Distributed Capacitance

Length to Diameter Ratio	*H*
0.10	0.96
0.15	0.79
0.20	0.78
0.25	0.64
0.30	0.60
0.35	0.57
0.40	0.54
0.50	0.50
1.00	0.46

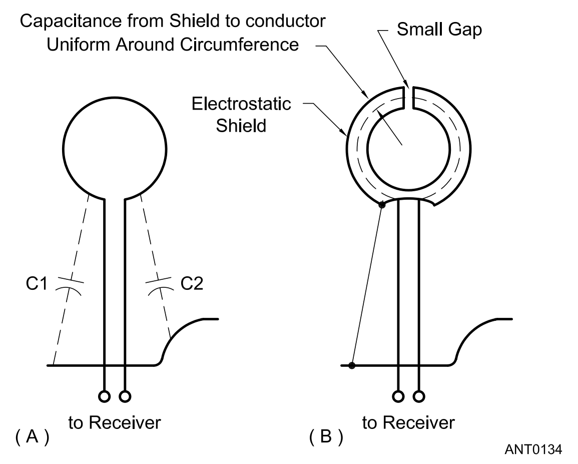

Fig 8—At A, the loop is unbalanced by capacitance to its surroundings. At B, the use of an electrostatic shield overcomes this effect.

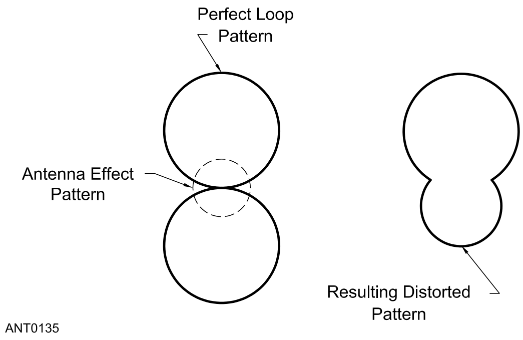

Fig 9—Distortion in loop pattern resulting from antenna effect.

making shielded loops. Genaille located his loop winding inside aluminum conduit, while True constructed an aluminum shield can around his winding. Others have used pieces of Hardline to form a loop, using the outer conductor as a shield. DeMaw used flexible coax with the shield broken at the center of the loop conductor in a multiturn loop for 1.8 MHz. Goldman uses another shielding method for broadcast receiver loops. His shield is in the form of a barrel made of hardware cloth, with the loop in its center. (See Bibliography for above references.) All these methods provide sufficient shielding to maintain the balance. It is possible, as Nelson shows, to construct an unshielded loop with good nulls (60 dB or better) by paying great care to symmetry.

LOOP Q

As previously mentioned, Q is an important consideration in loop performance because it determines both the loop bandwidth and its terminal voltage for a given field strength. The loaded Q of a loop is based on four major factors. These are (1) the intrinsic Q of the loop winding, (2) the effect of the load, (3) the effect of the electrostatic shield, and (4) the Q of the tuning capacitor.

The major factor is the Q of the winding of the loop itself. The ac resistance of the conductor caused by skin effect is the major consideration. The ac resistance for copper conductors may be determined from

$$R = \frac{0.996 \times 10^{-6} \sqrt{f}}{d} \qquad \text{(Eq 6)}$$

where

R = resistance in ohms per foot
f = frequency, Hz
d = conductor diameter, inches

The Q of the inductor is then easily determined by taking the reactance of the inductor and dividing it by the ac resistance. If you are using a multiturn loop and are a perfectionist, you might also want to include the loss from conductor proximity effect. This effect is described in detail later in this chapter, in the section on transmitting loops.

Improvement in Q can be obtained in some cases by the use of Litz wire (short for *Litzendraht*). Litz wire consists of strands of individual insulated wires that are woven into bundles in such a manner that each conductor occupies each location in the bundle with equal frequency. Litz wire results in improved Q over solid or stranded wire of equivalent size, up to about 3 MHz.

Also, the Q of the tuned circuit of the loop antenna is determined by the Q of the capacitors used to resonate it. In the case of air variables or dipped micas this is not usually a problem. But if variable-capacitance diodes are used to remotely tune the loop, pay particular attention to the manufacturer's specification for Q of the diode at the frequency of operation. The tuning diodes can have a significant effect on circuit Q.

Now we consider the effect of load impedance on loop Q. In the case of a directly coupled loop (as in Fig 5), the load is connected directly across the loop terminals, causing it to be treated as a parallel resistance in a parallel-tuned RLC circuit. Obviously, if the load is of a low value, the Q of the loop will be low. A simple way to correct this is to use a transformer to step up the load impedance that appears across the loop terminals. In fact, if we make this transformer a balun, it also allows us to use our unbalanced receivers with the loop and maintain loop symmetry. Another solution is to use what is referred to as an inductively coupled loop, such as DeMaw's four turn electrostatically shielded loop. A one-turn link is connected to the receiver. This turn is wound with the four-turn loop. In effect, this builds the transformer into the antenna.

Another solution to the problem of load impedance on loop Q is to use an active preamplifier with a high impedance balanced input and unbalanced output. This method also has the advantage of amplifying the low-level output voltage of the loop to where it can be used with a receiver of even mediocre sensitivity. In fact, the Q of the loop when used with a balanced preamplifier having high input impedance may be so high as to be unusable in certain applications. An example of this situation would occur where a loop is being used to receive a 5 kHz wide AM signal at a frequency where the bandwidth of the loop is only 1.5 kHz. In this case the detected audio might be very distorted. The solution to this is to put a Q-degrading resistor across the loop terminals.

FERRITE-CORE LOOP ANTENNAS

The ferrite-core loop antenna is a special case of the air-core receiving loops considered up to now. Because of its use in every AM broadcast-band portable radio, the ferrite-core loop is, by quantity, the most popular form of the loop antenna. But broadcast-band reception is far from its only use; it is commonly found in radio-direction-finding equipment and low-frequency-receiving systems (below 500 kHz) for time and frequency standard systems. In recent years, design information on these types of antennas has been a bit sparse in the amateur literature, so the next few paragraphs are devoted to providing some details.

Ferrite-loop antennas are characteristically very small compared to the frequency of use. For example, a 3.5-MHz version may be in the range of 15 to 30 cm long and about 1.25 cm in diameter. Earlier in this chapter, effective height was introduced as a measure of loop sensitivity. The effective height of an air-core loop antenna is given by Eq 2.

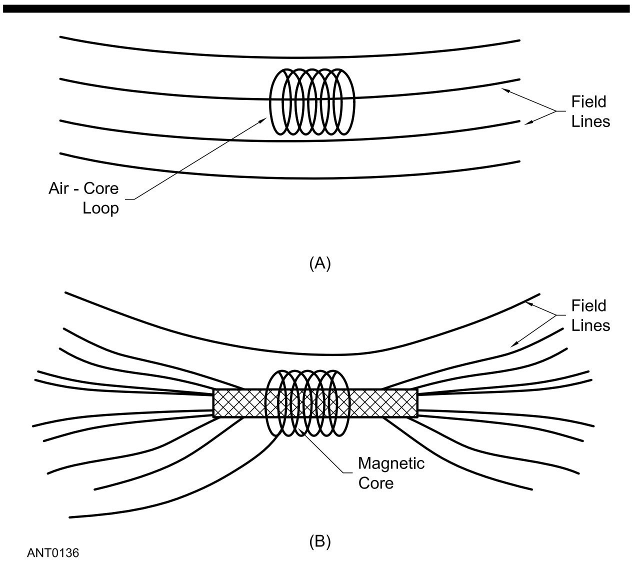

Fig 10—At A, an air-core loop has no effect on nearby field lines. B illustrates the effect of a ferrite core on nearby field lines. The field is altered by the reluctance of the ferrite material.

If an air-core loop is placed in a field, in essence it cuts the lines of flux without disturbing them (**Fig 10A**). On the other hand, when a ferrite (magnetic) core is placed in the field, the nearby field lines are redirected into the loop (Fig 10B). This is because the reluctance of the ferrite material is less than that of the surrounding air, so the nearby flux lines tend to flow through the loop rather than passing it by. (Reluctance is the magnetic analogy of resistance, while flux is analogous to current.) The reluctance is inversely proportional to the permeability of the rod core, μ_{rod}. (In some texts the rod permeability is referred to as effective permeability, μ_{eff}). This effect modifies the equation for effective height of a ferrite-core loop to

$$h = \frac{2\pi N A \mu_{rod}}{\lambda} \qquad \text{(Eq 7)}$$

where

h = effective height (length) in meters
N = number of turns in the loop
A = area of loop in square meters
μ_{rod} = permeability of the ferrite rod
λ = wavelength of operation in meters

This obviously is a large increase in “collected” signal. If the rod permeability were 90, this would be the same as making the loop area 90 times larger with the same number of turns. For example, a 1.25-cm diameter ferrite-core loop would have an effective height equal to an air-core loop 22.5 cm in diameter (with the same number of turns).

By now you might have noticed we have been very careful to refer to rod permeability. There is a very important reason for this. The permeability that a rod of ferrite exhibits is a combination of the material permeability or μ, the shape of the rod, and the dimensions of the rod. In ferrite rods, μ is sometimes referred to as initial permeability, μ_i, or toroidal permeability, μ_{tor}. Because most amateur ferrite loops are in the form of rods, we will discuss only this shape.

The reason that μ_{rod} is different from μ is a very complex physics problem that is well beyond the scope of this book. For those interested in the details, books by Polydoroff and by Snelling cover this subject in considerable detail. (See Bibliography.) For our purposes a simple explanation will suffice. The rod is in fact not a perfect director of flux, as is illustrated in **Fig 11**. Note that some lines impinge on the sides of the core and also exit from the sides. These lines therefore would not pass through all the turns of the coil if it were wound from one end of the core to the other. These flux lines are referred to as *leakage flux*, or sometimes as flux leakage.

Leakage flux causes the flux density in the core to be nonuniform along its length. From Fig 11 it can be seen that the flux has a maximum at the geometric center of the length of the core, and decreases as the ends of the core are approached. This causes some noticeable effects. As a short coil is placed at different locations along a long core, its inductance will change. The maximum inductance exists when the coil is centered on the rod. The Q of a short coil on a long rod is greatest at the center. On the other hand, if you require a higher Q than this, it is recommended that you spread the coil turns along the whole length of the core, even though this will result in a lower value of inductance. (The inductance can be increased to the original value by adding turns.) **Fig 12** gives the relationship of rod permeability to material permeability for a variety of values.

The change in μ over the length of the rod results in an adjustment in the term μ_{rod} for its so called “free ends” (those not covered by the winding). This adjustment factor is given by

$$\mu' = \mu_{rod} \sqrt[3]{\frac{a}{b}} \qquad \text{(Eq 8)}$$

where

μ' = the corrected permeability
a = the length of the core
b = the length of the coil

This value of μ' should be used in place of μ_{rod} in Eq 7 to obtain the most accurate value of effective height.

All these variables make the calculation of ferrite loop antenna inductance somewhat less accurate than for the air-core version. The inductance of a ferrite loop is given by

$$L = \frac{4\pi N^2 A \mu_{rod} \times 10^{-4}}{\ell} \qquad \text{(Eq 9)}$$

where

L = inductance in μH
N = number of turns
A = cross-sectional area of the core in square mm
ℓ = magnetic length of core in mm

Experiments indicate that the winding diameter should be as close to that of the rod diameter as practical in order to maximize both inductance value and Q. By using all this information, we may determine the voltage at the loop terminals and its signal-to-noise ratio (SNR). The voltage may be determined from

$$V = \frac{2\pi A N \mu' Q E}{\lambda} \qquad \text{(Eq 10)}$$

where

V = output voltage across the loop terminals
A = loop area in square meters
N = number of turns in the loop winding
μ' = corrected rod permeability
Q = loaded Q of the loop
E = RF field strength in volts per meter
λ = wavelength of operation in meters

Lankford’s equation for the sensitivity of the loop for a 10 dB SNR is

$$E = \frac{1.09 \times 10^{-10} \lambda \sqrt{f\ L\ b}}{A N \mu' \sqrt{Q}} \qquad \text{(Eq 11)}$$

where

f = operating frequency in Hz
L = loop inductance in henrys
b = receiver bandwidth in Hz

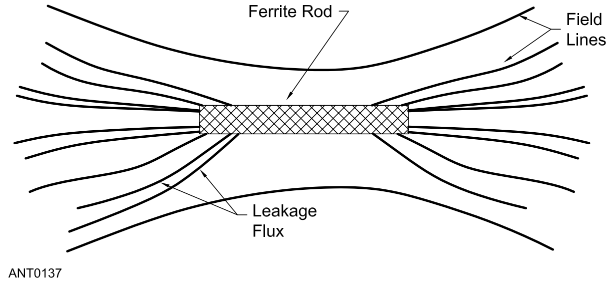

Fig 11—Example of magnetic field lines near a practical ferrite rod, showing leakage flux.

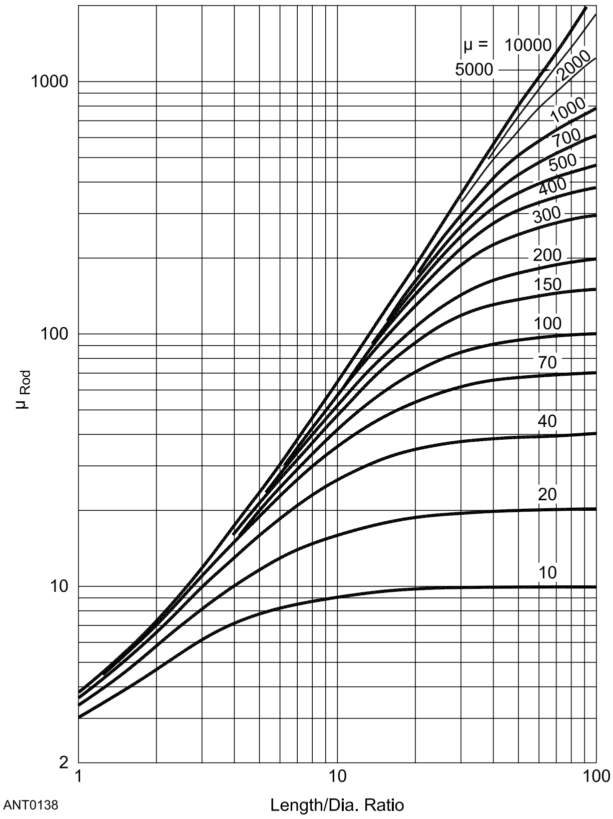

Fig 12—Rod permeability, μ_{rod}, versus material permeability, μ, for different rod length-to-diameter ratios.

Similarly, Belrose gives the SNR of a tuned loop antenna as

$$\mathrm{SNR} = \frac{66.3\,N\,A\,\mu_{rod}\,E}{\sqrt{b}}\sqrt{\frac{Q\,f}{L}} \qquad \text{(Eq 12)}$$

From this, if the field strength E, μ_{rod}, b, and A are fixed, then Q or N must increase (or L decrease) to yield a better SNR. Higher sensitivity can also be obtained (especially at frequencies below 500 kHz) by bunching ferrite cores together to increase the loop area over that which would be possible with a single rod. High sensitivity is important because loop antennas are not the most efficient collectors of signals, but they do offer improvement over other receiving antennas in terms of SNR. For this reason, you should attempt to maximize the SNR when using a small loop receiving antenna. In some cases there may be physical constraints that limit how large you can make a ferrite-core loop.

After working through Eq 11 or 12, you might find you still require some increase in antenna system gain to effectively use your loop. In these cases the addition of a low noise preamplifier may be quite valuable even on the lower frequency bands where they are not commonly used. Chapter 14 contains information on such preamplifiers.

The electrostatic shield discussed earlier with reference to air-core loops can be used effectively with ferrite-core loops. (Construction examples are presented in Chapter 14.) As in the air-core loop, a shield will reduce electrical noise and improve loop balance.

PROPAGATION EFFECTS ON NULL DEPTH

After building a balanced loop you may find it does not approach the theoretical performance in the null depth. This problem may result from propagation effects. Tilting the loop away from a vertical plane may improve performance under some propagation conditions, to account for the vertical angle of arrival. Basically, the loop performs as described above only when the signal is arriving perpendicular to the axis of rotation of the loop. At incidence angles other than perpendicular, the position and depth of the nulls deteriorate.

The problem can be even further influenced by the fact that if the loop is situated over less than perfectly conductive ground, the wave front will appear to tilt or bend. (This bending is not always detrimental; in the case of Beverage antennas, sites are chosen to take advantage of this effect.)

Another cause of apparent poor performance in the null depth can be from polarization error. If the polarization of the signal is not completely linear, the nulls will not be sharp. In fact, for circularly polarized signals, the loop might appear to have almost no nulls. Propagation effects are discussed further in Chapter 14.

SITING EFFECTS ON THE LOOP

The location of the loop has an influence on its performance that at times may become quite noticeable. For ideal performance the loop should be located outdoors and clear of any large conductors, such as metallic downspouts and towers. A VLF loop, when mounted this way, will show good sharp nulls spaced 180° apart if the loop is well balanced. This is because the major propagation mode at VLF is by ground wave. At frequencies in the HF region, a significant portion of the signals is propagated by sky wave, and nulls are often only partial.

Most hams locate their loop antennas near their operating position. If you choose to locate a small loop indoors, its performance may show nulls of less than the expected depth, and some skewing of the pattern. For precision direction finding there may be some errors associated with wiring, plumbing, and other metallic construction members in the building. Also, a strong local signal may be reradiated from the surrounding conductors so that it cannot be nulled with any positioning of the loop. There appears to be no known method of curing this type of problem. All this should not discourage you from locating a loop indoors; this information is presented here only to give you an idea of some pitfalls. Many hams have reported excellent results with indoor mounted loops, in spite of some of the problems.

Locating a receiving loop in the field of a transmitting antenna may cause a large voltage to appear at the receiver antenna terminals. This may be sufficient to destroy sensitive RF amplifier transistors or front-end protection diodes. This can be solved by disconnecting your loop from the receiver during transmit periods. This can obviously be done automatically with a relay that opens when the transmitter is activated.

LOOP ANTENNA ARRAYS

Arrays of loop antennas, both in combination with each other and with other antenna types, have been used for many years. The arrays are generally used to cure some “deficiency” in the basic loop for a particular application, such as a 180° ambiguity in the null direction, low sensitivity, and so forth.

A Sensing Element

For direction-finding applications the single loop suffers the problem of having two nulls that are 180° apart. This leads to an ambiguity of 180° when trying to find the direction to a transmitting station from a given location. A sensing element (often called a *sense antenna*) may be added to the loop, causing the overall antenna to have a cardioid pattern and only one null. The sensing element is a small vertical antenna whose height is equal to or greater than the loop effective height. This vertical is physically close to the loop, and when its omnidirectional pattern is adjusted so that its amplitude and phase are equal to one of the loop lobes, the patterns combine to form a cardioid. This antenna can be made quite compact by use of a ferrite loop to form a portable DF antenna for HF direction finding. Chapter 14 contains additional information and construction projects using sensing elements.

Arrays of Loops

A more advanced array that can develop more diverse patterns consists of two or more loops. Their outputs are combined through appropriate phasing lines and combiners to form a phased array. Two loops can also be formed into an array that can be rotated without physically turning the loops themselves. This method was developed by Bellini and Tosi in 1907 and performs this apparently contradictory feat by use of a special transformer called a *goniometer*. The goniometer is described in Chapter 14.

Aperiodic Arrays

The aperiodic loop array is a wide-band antenna. This type of array is useful over at least a decade of frequency, such as 2 to 20 MHz. Unlike most of the loops discussed up to now, the loop elements in an aperiodic array are untuned. Such arrays have been used commercially for many years. One loop used in such an array is shown in **Fig 13**. This loop is quite different from all the loops discussed so far in this chapter because its pattern is not the familiar figure eight. Rather, it is omnidirectional.

The antenna is omnidirectional because it is purposely unbalanced, and also because the isolating resistor causes the antenna to appear as two closely spaced short monopoles. The loop maintains the omnidirectional characteristics over a frequency range of at least four or five to one. These loops, when combined into end-fire or broadside phased arrays, can provide quite impressive performance. A commercially made end-fire array of this type consisting of four loops equally spaced along a 25-meter baseline can provide gains in excess of 5 dBi over a range of 2 to 30 MHz. Over a considerable portion of this frequency range, the array can maintain F/B ratios of 10 dB. Even though the commercial version is very expensive, an amateur version can be constructed using the information provided by Lambert. One interesting feature of this type of array is that, with the proper combination of hybrids and combiners, the antenna can simultaneously feed two receivers with signals from different directions, as shown in **Fig 14**. This antenna may be especially interesting to one wanting a directional receiving array for two or more adjacent amateur bands.

SMALL TRANSMITTING LOOP ANTENNAS

The electrically small transmitting-loop antenna involves some different design considerations compared to receiving loops. Unlike receiving loops, the size limitations of the antenna are not as clearly defined. For most purposes, any transmitting loop whose physical circumference is less than ¼ λ can be considered "small." In most cases, as a consequence of their relatively large size (when compared to a receiving loop), transmitting loops have a nonuniform current distribution along their circumference. This leads to some performance changes from a receiving loop.

The transmitting loop is a parallel-tuned circuit with a large inductor acting as the radiator. As with the receiving

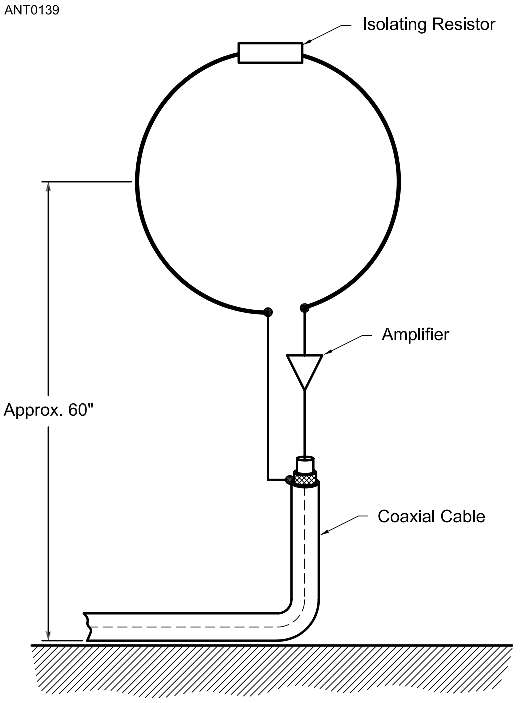

Fig 13—A single wide-band loop antenna used in an aperiodic array.

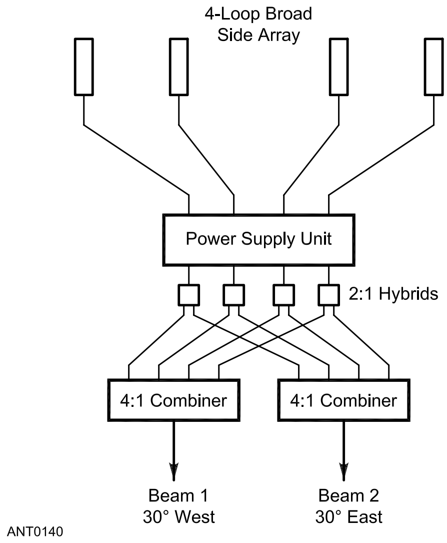

Fig 14—Block diagram of a four-loop broadside array with dual beams separated by 60° in azimuth.

Table 3
Transmitting Loop Equations

$$X_L = 2\pi f L \text{ ohms}$$

$$Q = \frac{f}{\Delta f} = \frac{X_L}{2\left(R_R + R_L\right)}$$

$$R_R = 3.12 \times 10^4 \left[\frac{NA}{\lambda^2}\right]^2 \text{ ohms}$$

$$V_C = \sqrt{P X_L Q}$$

$$I_L = \sqrt{\frac{PQ}{X_L}}$$

where

X_L = inductive reactance, ohms
f = frequency, Hz
Δf = bandwidth, Hz
R_R = radiation resistance, ohms
R_L = loss resistance, ohms (see text)
N = number of turns
A = area enclosed by loop, square meters
λ = wavelength at operating frequency, meters
V_C = voltage across capacitor
P = power, watts
I_L = resonant circulating current in loop

loop, the calculation of the transmitting-loop inductance may be carried out with the equations in Table 1. Avoid equations for long solenoids found in most texts. Other fundamental equations for transmitting loops are given in **Table 3**.

In the March 1968 *QST*, Lew McCoy, W1ICP, introduced the so-called "Army Loop" to radio amateurs. This was an amateur version of a loop designed for portable use in Southeast Asia by Patterson of the US Army and described in 1967. The Army Loop is diagrammed in **Fig 15A**, showing that this is a parallel tuned circuit fed by a tapped-capacitance impedance-matching network.

The Hart "high-efficiency" loop was introduced in the June 1986 *QST* by Ted Hart, W5QJR. It is shown schematically in Fig 15B and has the series-tuning capacitor separate from the matching network. The Hart matching network is basically a form of gamma match. Other designs have used a smaller loop connected to the transmission line to couple into the larger transmitting loop.

The approximate radiation resistance of a loop in ohms is given by

$$R_R = 3.12 \times 10^4 \left(\frac{NA}{\lambda^2}\right)^2 \qquad \text{(Eq 13)}$$

where

N = number of turns
A = area of loop in square meters
λ = wavelength of operation in meters

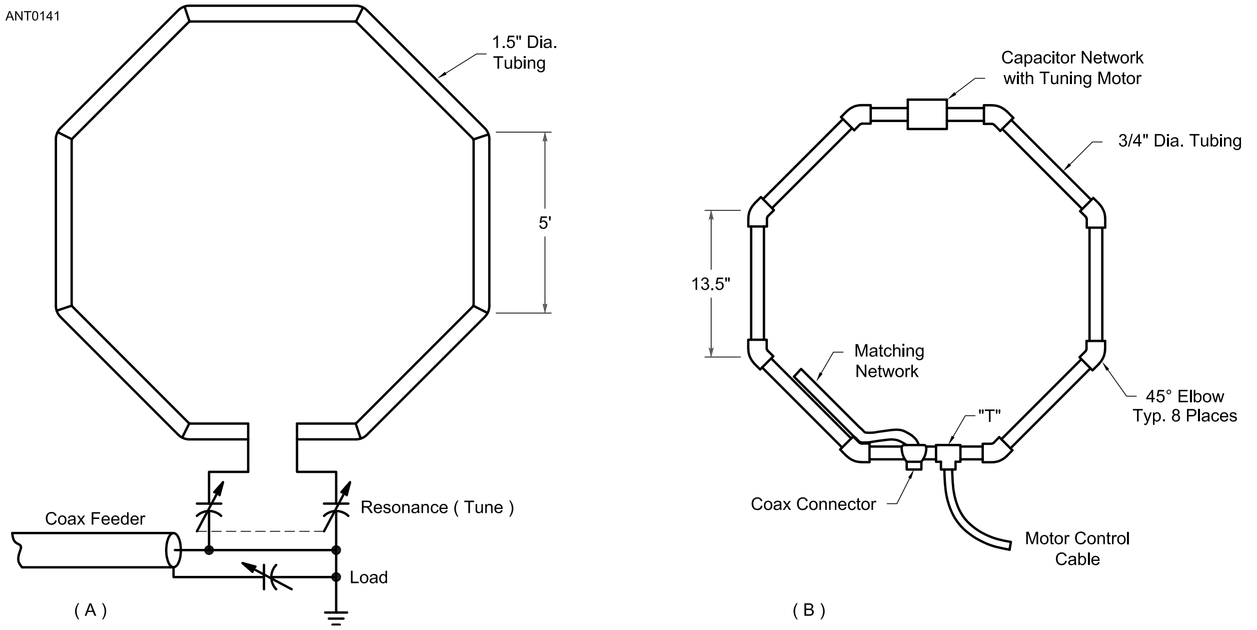

Fig 15—At A, a simplified diagram of the Army Loop. At B, the W5QJR loop, which is described in more detail later in this chapter.

The radiation resistance of a small transmitting loop is usually very small. For example, a 1-meter diameter, single-turn circular loop has a radius of 0.5 meters and an enclosed area of $\pi \times 0.5^2 = 0.785$ m^2. Operated at 14.0 MHz, the free-space wavelength is 21.4 meters and this leads to a computed radiation resistance of only $3.12 \times 10^{-4} (0.785/21.4^2)^2 = 0.092\ \Omega$.

Unfortunately the loop also has losses, both ohmic and from skin effect. By using this information, the radiation efficiency of a loop can be calculated from

$$\eta = \frac{R_R}{R_R + R_L} \qquad \text{(Eq 14)}$$

where

η = antenna efficiency, %

R_R = radiation resistance, Ω

R_L = loss resistance, Ω, which includes the loop's conductor loss plus the loss in the series-tuning capacitor.

A simple ratio of R_R versus R_L shows the effects on the efficiency, as can be seen from **Fig 16**. The loss resistance is primarily the ac resistance of the conductor. This can be calculated from Eq 6. A transmitting loop generally requires the use of copper conductors of at least ¾ inch in diameter in order to obtain reasonable efficiency. Tubing is as useful as a solid conductor because high-frequency currents flow only along a very small depth of the surface of the conductor; the center of the conductor has almost no effect on current flow.

Note that the R_L term above also includes the effect of the tuning capacitor's loss. Normally, the unloaded Q of a capacitor can be considered to be so high that any loss in the tuning capacitor can be neglected. For example, a very high-quality tuning capacitor with no mechanical wiping contacts, such as a vacuum-variable or a transmitting butterfly capacitor, might have an unloaded Q of about 5000. This implies a series loss resistance of less than about 0.02 Ω for a capacitive reactance of 100 Ω. This relatively tiny loss resistance can become significant, however, when the radiation resistance of the loop is only on the order of 0.1 Ω! Practical details for curbing capacitor losses are covered later in this chapter.

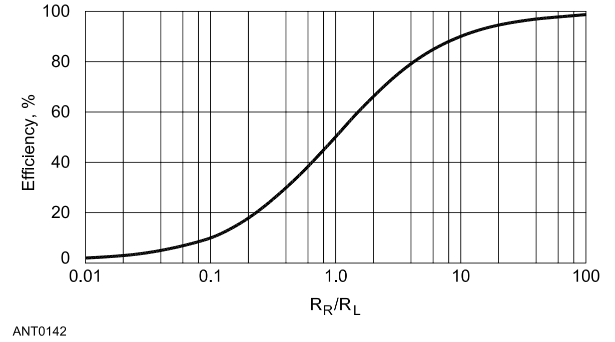

Fig 16—Effect of ratio of R_R/R_L on loop efficiency.

In the case of multiturn loops there is an additional loss related to a term called *proximity effect*. The proximity effect occurs in cases where the turns are closely spaced (such as being spaced one wire diameter apart). As these current-carrying conductors are brought close to each other, the current density around the circumference of each conductor gets redistributed. The result is that more current per square meter is flowing at the surfaces adjacent to other conductors. This means that the loss is higher than a simple skin-effect analysis would indicate, because the current is bunched so it flows through a smaller cross section of the conductor than if the other turns were not present.

As the efficiency of a loop approaches 90%, the proximity effect is less serious. But unfortunately, the less efficient the loop, the worse the effect. For example, an 8-turn transmitting loop with an efficiency of 10% (calculated by the skin-effect method) actually only has an efficiency of 3% because of the additional losses introduced by the proximity effect. If you are contemplating construction of a multiturn transmitting loop, you might want to consider spreading the conductors apart to reduce this effect. G. S. Smith includes graphs that detail this effect in his 1972 IEEE paper.

The components in a resonated transmitting loop are subject to both high currents and voltages as a result of the large circulating currents found in the high-Q tuned circuit formed by the antenna. This makes it important that any fixed capacitors have a high RF current rating, such as transmitting micas or the Centralab 850 series. Be aware that even a 100-W transmitter can develop currents in the tens of amperes, and voltages across the tuning capacitor in excess of 10,000 V. This consideration also applies to any conductors used to connect the loop to the capacitors. A piece of #14 wire may have more resistance than the rest of the loop conductor!

It is therefore best to use copper strips or the braid from a piece of large coax cable to make any connections. Make the best electrical connection possible, using soldered or welded joints. Using nuts and bolts should be avoided, because at RF these joints generally have high resistance, especially after being subjected to weathering.

An unfortunate consequence of having a small but high-efficiency transmitting loop is high loaded Q, and therefore limited bandwidth. This type of antenna may require retuning for frequency changes as little as 5 kHz. If you are using any wide-band mode such as AM or FM, this might cause fidelity problems and you might wish to sacrifice a little efficiency to obtain the required bandwidth.

A special case of the transmitting loop is that of the ferrite-loaded loop. This is a logical extension of the transmitting loop if we consider the improvement that a ferrite core makes in receiving loops. The use of ferrites in a transmitting loop is still under development. (See the Bibliography reference for DeVore and Bohley.)

PRACTICAL COMPACT TRANSMITTING LOOPS

The ideal small transmitting antenna would have performance equal to a large antenna. A small loop antenna can approach that performance except for a reduction in bandwidth, but that effect can be overcome by retuning. This section is adapted and updated from material written by Robert T. (Ted) Hart, W5QJR.

As pointed out above, small antennas are characterized by low radiation resistance. For a typical small antenna, such as a short dipole, loading coils are often added to achieve resonance. However, the loss inherent in the coils can result in an antenna with low efficiency. If instead of coils a large, low-loss capacitor is added to a low-loss conductor to achieve resonance, and if the antenna conductor is bent to connect the ends to the capacitor, a loop is formed.

Based on this concept, the small loop is capable of relatively high efficiency, compared to its coil-loaded cousin. In addition, the small loop, when mounted vertically, can radiate efficiently over the wide range of elevation angles required on the lower frequency bands. This is because it has both high-angle and low-angle response. See **Fig 17**, which shows the elevation response for a compact transmitting loop only 16.2 inches wide at 14.2 MHz. This loop is vertically polarized and its bottom is 8 feet above average ground, which has a conductivity of 5 mS/m and a dielectric constant of 13. For comparison, Fig 17 also shows the responses of three other reference antennas—the same small loop flipped sideways at a height of 30 feet to produce horizontal radiation, a full-sized ¼-λ ground plane antenna mounted 8 feet above average ground using two tuned radials, and finally a simple ½ λ flattop dipole mounted 30 feet above flat ground. The considerably smaller transmitting loop comes to within 3 dB of the larger ¼-λ vertical at a 10° elevation angle, and it is far stronger for high elevation angles because it does not have the null at high elevation angles that the ground plane has. Of course, this characteristic does make it more susceptible to strong signals received at high elevation angles. Incidentally, just in case you were wondering, adding more radials to the λ/4 ground plane doesn't materially improve its performance when mounted at an 8-foot height on 20 meters.

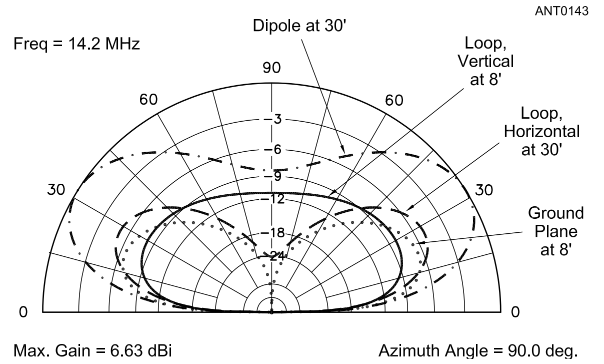

Fig 17—Elevation-plane plot at 14.2 MHz, showing response of an 8.5-foot circumference octagonal copper loop (width of 16.2 inches), compared to a full-sized λ/4 ground-plane vertical with two elevated λ/4 radials, the same small loop flipped horizontally at a height of 30 feet, and lastly, a λ/2 flattop dipole also at a height of 30 feet. Both the λ/4 ground-plane vertical and the vertically polarized loop are elevated 8 feet above typical ground, with σ = 5 mS/m and ε = 13. The low vertically polarized loop is surprisingly competitive, only down about 2.5 dB compared to the far larger ground plane at low elevation angles. Note that the vertical loop has both high-angle as well as low-angle radiation, and hence would be better at working close-in local stations than the ground-plane vertical, with its deep nulls at higher angles. The simple flattop dipole, however, is better than either vertical because of the poor ground reflection for a vertically polarized compared to a horizontally polarized signal.

The simple horizontal dipole in Fig 17 would be the clear winner in any shootout because its horizontally polarized radiation does not suffer as much attenuation at reflection from ground as does a vertically polarized wave. The case is not quite so clear-cut, however, for the small loop mounted horizontally at 30 feet. While it does have increased gain at medium elevation angles, it may not be worth the effort needed to mount it on a mast, considering the slight loss at low angles compared to its twin mounted vertically only 8 feet above ground.

A physically small antenna like the 16.2-inch-wide vertically polarized loop does put out an impressive signal compared to far larger competing antennas. Though somewhat ungainly, it is a substantially better performer than most mobile whips, for example. The main deficiency in a compact transmitting loop is its narrow bandwidth—it must be accurately tuned to the operating frequency. The use of a remote motor drive allows the loop to be tuned over a wide frequency range.

For example, for fixed-station use, two loops could be constructed to provide continuous frequency coverage from 3.5 to 30 MHz. A loop with an 8.5 foot circumference, 16 inches wide, could cover 10 through 30 MHz and a loop with a 20-foot circumference, 72 inches wide, could cover 3.5 to 10.1 MHz.

Table 4 presents summary data for various size loop antennas for the HF amateur bands. Through computer analysis, the optimum size conductor was determined to be $^{3}/_{4}$-inch rigid copper water pipe, considering both performance and cost. Performance will be compromised, but only slightly, if $^{5}/_{8}$-inch flexible copper tubing is used. This tubing can easily be bent to any desired shape, even a circle. The rigid $^{3}/_{4}$-inch copper pipe is best used with 45° elbows to make an octagon.

The loop circumference should be between $^{1}/_{4}$ and $^{1}/_{8}$ λ at the operating frequency. It will become self-resonant above $^{1}/_{4}$ λ, and efficiency drops rapidly below $^{1}/_{8}$ λ. In the frequency ranges shown in Table 4, the high fre-

Table 4

Design Data for Loops

Loop Circumference = 8.5' (Width = 32.4"), Vertically Polarized

Frequency, MHz	10.1	14.2	21.2	29.0
Max Gain, dBi	–4.47	–1.42	+1.34	+2.97
Max Elevation Angle	40°	30°	22°	90°
Gain, dBi @10°	–8.40	–4.61	–0.87	+0.40
Total Capacitance, pF	145	70	29	13
Peak Capacitor kV	23	27	30	30

Loop Circumference = 8.5' (Width = 32.4"), Horizontally Polarized, @30'

Frequency, MHz	10.1	14.2	21.2	29.0
Max Gain, dBi	–3.06	+1.71	+5.43	+6.60
Max Elevation Angle	34°	28°	20°	16°
Gain, dBi @10°	–9.25	–3.11	+2.61	+5.34
Total Capacitance, pF	145	70	29	13
Peak Capacitor kV	23	27	30	30

Loop Circumference = 20' (Width = 6'), Vertically Polarized

Frequency, MHz	3.5	4.0	7.2	10.1
Max Gain, dBi	–7.40	–6.07	–1.69	–0.34
Max Elevation Angle	68°	60°	38°	30°
Gain, dBi @10°	–11.46	–10.12	–5.27	–3.33
Capacitance, pF	379	286	85	38
Peak Capacitor kV	22	24	26	30

Loop Circumference = 20' (Width = 6'), Horizontally Polarized, @30'

Frequency, MHz	3.5	4.0	7.2	10.1
Max Gain, dBi	–13.32	–10.60	–0.20	+3.20
Max Elevation Angle	42°	42°	38°	34°
Gain, dBi @10°	–21.62	–18.79	–7.51	–3.22
Capacitance, pF	379	286	85	38
Peak Capacitor kV	22	24	26	30

Loop Circumference = 38' (Width = 11.5'), Vertically Polarized

Frequency, MHz	3.5	4.0	7.2
Max Gain, dBi	–2.93	–2.20	–0.05
Max Elevation Angle	46°	42°	28°
Gain, dBi @10°	–6.48	–5.69	–2.80
Capacitance, pF	165	123	29
Peak Capacitor kV	26	27	33

Notes: These loops are octagonal in shape, constructed with ¾-inch copper water pipe and soldered 45° copper elbows. The gain figures assume a capacitor unloaded $Q_C = 5000$, typical for vacuum-variable type of tuning capacitor. The bottom of the loop is assumed to be 8 feet high for safety and the ground constants are “typical” at conductivity = 5 mS/m and dielectric constant = 13. Transmitter power is 1500 W. The voltage across the tuning capacitor for lower powers goes down with a multiplier of $\sqrt{\frac{P}{1500}}$. For example, at 100 W using the 38-foot-circumference loop at 7.2 MHz, the peak voltage would be $33\text{ kV} \times \sqrt{\frac{100}{1500}} = 8.5\text{ kV}$.

quency is tuned with a minimum capacitance of about 29 pF—including stray capacitance.

The low frequency listed in Table 4 is that where the loop response is down about 10 dB from that of a full-sized elevated ground plane at low elevation angles suitable for DX work. **Fig 18** shows an overlay at 3.5 MHz of the elevation responses for two loops: one with an 8.5-foot circumference and one with a 20-foot circumference, together with the response for a full-sized 80-meter ground plane elevated 8 feet off average ground with 2 tuned radials. The 20-foot circumference loop holds its own well compared to the full-sized ground plane.

Controlling Losses

Contrary to earlier reports, adding quarter-wave ground radials underneath a vertically polarized transmitting loop doesn’t materially increase loop efficiency. The size of the conductor used for a transmitting loop, however, does directly affect several interrelated aspects of loop performance.

Data for Table 4 was computed for ¾-inch copper water pipe (nominal OD of 0.9 inch). Note that the efficiency is higher and the Q is lower for loops having a circumference near $^1/_4$ λ. Larger pipe size will reduce the loss resistance, but the Q increases. Therefore the bandwidth decreases, and the voltage across the tuning capacitor increases. The voltage across the tuning capacitor for high-power operation can become very impressive, as shown in Table 4. Rigid ¾-inch copper water pipe is a good electrical compromise and can also help make a small-diameter loop mechanically sturdy.

The equivalent electrical circuit for the loop is a parallel resonant circuit with a very high Q, and therefore a narrow bandwidth. The efficiency is a function of radiation resistance divided by the sum of the radiation plus loss resistances. The radiation resistance is much less than 1 Ω, so it is necessary to minimize the loss resistance, which is largely the skin-effect loss of the conductor, assuming that the tuning capacitor has very low loss. Poor construction techniques must be avoided. All joints in the loop must be brazed or soldered.

However, if the system loss is too low, for example by using even larger diameter tubing, the Q may become excessive and the bandwidth may become too narrow for practical use. These reasons dictate the need for a complete analysis to be performed before proceeding with the construction of a loop.

There is another source of additional loss in a completed loop antenna besides the conductor and capacitor losses. If the loop is mounted near lossy metallic conductors, the large magnetic field produced will induce currents into those conductors and be reflected as losses in the loop. Therefore the loop should be as far from other conductors as possible. If you use the loop inside a building constructed with large amounts of iron or near ferrous materials, you will simply have to live with the loss if the loop cannot otherwise be relocated.

The Tuning Capacitor

Fig 19 demonstrates the selection of loop size versus tuning capacitance for any desired operating frequency range for the HF amateur bands. This is for octagonal-shaped loops using ¾-inch copper water pipe with 45° copper elbows. For example, a capacitor that varies from 5 to 50 pF, used with a loop 10 feet in circumference, tunes from 13 to 27 MHz (represented by the left dark vertical bar). A 25 to 150-pF capacitor with a 13.5-foot loop circumference covers the 7 to 14.4-MHz range, represented by the right vertical bar.

Fig 20 illustrates how the 29-MHz elevation pattern becomes distorted and rather bulbous-looking for the 10-foot circumference loop, although the response at low elevation angles is still better than that of a full-sized ground-plane antenna.

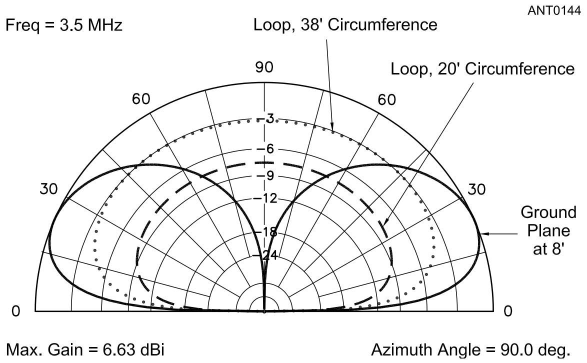

Fig 18—Elevation-plane response of three antennas at 3.5 MHz—a 20-foot circumference octagonal copper loop, a 38-foot circumference copper loop and a full-sized λ/4 ground plane with two elevated radials. The bottom of each antenna is mounted 8 feet above ground for safety. The 38-foot circumference loop (which has a "wingspan" of 11.5 feet) is fairly competitive with the much large ground-plane, being down only about 4 dB at low elevation angles. The 20-foot circumference loop is much more lossy, but with its top only about 14 feet off the ground is very much of a "stealth" antenna.

Air Variable Capacitors

Special care must be taken with the tuning capacitor if an air-variable type is used. The use of a split-stator capacitor eliminates the resistance of wiper contacts, resistance that is inherent in a single-section capacitor. The ends of the loop are connected to the stators, and the rotor forms the variable coupling path between the stators. With this arrangement the value of capacitance is divided by two, but the voltage rating is doubled.

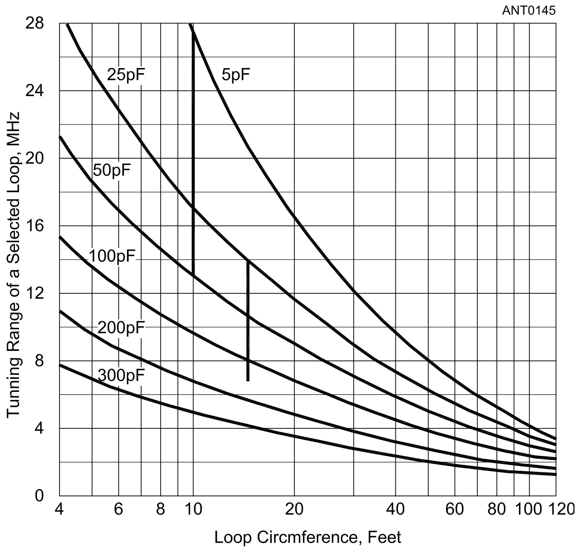

Fig 19—Frequency tuning range of an octagon-shaped loop using $^{3}/_{4}$-inch copper water pipe, for various values of tuning capacitance and loop circumference.

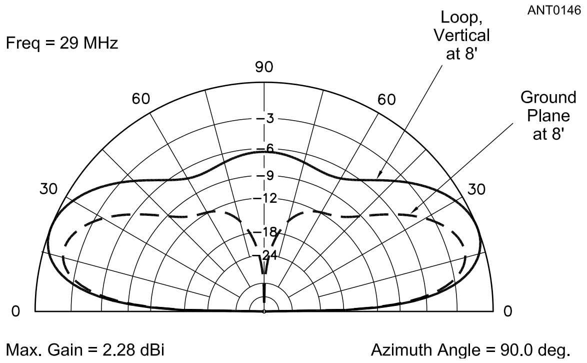

Fig 20—Elevation-plane plot for a 16.2-inch wingspan octagonal copper loop at 29 MHz, compared to a λ/4 ground-plane antenna with two resonant elevated radials. The gains at low angles are almost identical, but the loop exhibits more gain at medium and high elevation angles. Again, the bottom of each antenna is located 8 feet above ground for safety.

You must carefully select a variable capacitor for transmitting-loop application—that is, all contacts must be welded, and no mechanical wiping contacts are allowed. For example, if the spacers between plates are not welded to the plates, there will be loss at each joint, and thus degraded loop efficiency. (Earlier transmitting loops exhibited poor efficiency because capacitors with wiping contacts were used.)

There are several suitable types of capacitors for this application. A vacuum variable is an excellent choice, provided one is selected with an adequate voltage rating. Unfortunately, those capacitors are very expensive.

W5QJR used a specially modified air-variable capacitor in his designs. This had up to 340 pF maximum per section, with ¼-inch spacing, resulting in 170 pF when both sections were in series as a butterfly capacitor. Another alternative is to obtain a large air variable, remove the aluminum plates, and replace them with copper or double-sided PC board material to reduce losses. Connect all plates together on the rotor and on the stators. Solder copper straps to the capacitor for soldering to the loop itself.

The spacing between plates in an air-variable capacitor determines the voltage-handling capability, rated at 75,000 V per inch. For other power ratings, multiply the spacing (and voltage) by the square root of the ratio of your power to 1000 W. For example, for 100 W, the ratio would be = 0.316.

Table 5
KD7S Loop-Tuning Capacitor Parts List for Nominal 50-pF Capacitor

Qty	*Description*
2	10-inch length of ¾-inch-ID type M copper water pipe
2	10-inch length of ½-inch-ID type M copper water pipe
1	3-inch length of ½-inch-ID type M copper water pipe
2	½-inch, 90° copper elbows
2	¾-inch, 90° copper elbows
2	10 × 22-inch piece of 0.005-inch-thick Teflon sheet plastic
1	12-inch length of #8-32 threaded brass rod
1	#8-32 brass shoulder nut
2	22 × 5½ × ¼-inch ABS plastic sheet (top and bottom covers)
3	1 × 5½ × ¼-inch ABS plastic sheet (end pieces and center) brace/guide
2	1 × 22 × ¼-inch ABS plastic sheet (side rails)
1	50 to 200-rpm gear-head dc motor
1	DPDT center-off toggle switch (up/down control)
2	SPDT microswitches (limit switches)
50 feet	3-conductor control cable
1	Enclosure for control switch

A Teflon-Insulated Trombone Variable Capacitor

Another type of variable capacitor discussed in the amateur literature for use with a compact transmitting loop is the so-called "trombone" type of capacitor. **Fig 21** shows a practical trombone capacitor created by Bill Jones, KD7S, for Nov 1994 *QST*. This capacitor uses downward pointing extensions of the two ¾-inch OD main conductor copper pipes, with a Teflon-insulated trombone section made of ½-inch ID copper pipe. The trombone telescopes into the main pipes, driven by a lead screw and a 180-rpm gear-head motor. Like the butterfly air variable capacitor, the trombone works without lossy wiper contacts. Jones' capacitor varied from 12 pF (including strays) to almost 60 pF, making it suitable to tune his 3-foot circumference loop from 14 to 30 MHz at the 100-W level.

KD7S used 5-mil (0.005 inch) thick Teflon sheet as an insulator. Since Teflon is conservatively rated at more than 1 kV per mil of thickness, the voltage breakdown capability of this capacitor is well in excess of 5 kV. The parts list is given in **Table 5**.

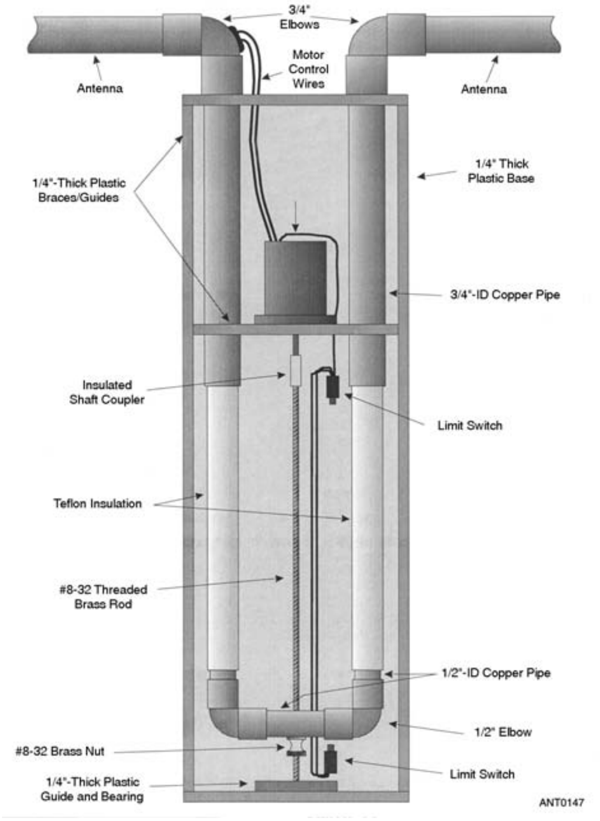

Fig 21—A practical trombone capacitor designed by Bill Jones, KD7S, for his compact transmitting loop. This capacitor has a tuning range from 12 to almost 60 pF, and can withstand at least 5 kV peak. The 10-inch ½-inch ID tubes are covered with Teflon-sheet insulation and slide into the ¾-inch ID copper pipes.

A short length of plastic tubing connects the threaded brass rod to the motor. The tubing acts as an insulator and a flexible coupling to smooth out minor shaft-alignment errors. The other end of the rod is threaded into a brass nut soldered to the crossbar holding the $^1/_2$-inch pipes together. Jones used a 12-V motor rated at 180 rpm, but it has sufficient torque to work with as little as 4 V applied. Instead of a sophisticated variable duty-cycle speed control circuit, he used an LM327 adjustable voltage regulator to vary the motor-control voltage from 4 to 12 V. Tuning speeds ranged from 11 seconds per inch at 12 V to 40 seconds per inch at 4 V. The higher speed is necessary to jump from band to band in a reasonable length of time. The lower speed makes it easy to fine-tune the capacitor to any desired frequency within a band.

When building the capacitor, keep in mind that the smaller tubes must telescope in and out of the larger tubes with silky smoothness. Any binding will cause erratic tuning. For the same reason, the #8-32 brass threaded rod must be straight and properly aligned with the brass nut. *Take your time with this part of the project*.

Perhaps the easiest way to form the insulator is to precut a length of Teflon sheet to the proper size. Place a lengthwise strip of double-sided tape on the tube to secure one end of the Teflon sheet. Begin wrapping the Teflon around the tube while keeping it as tight as possible. *Don't allow wrinkles or ridges to form*. Secure the other end with another piece of tape. Once both tubes are covered, ensure they are just short of being a snug fit inside the larger tubes. Confirm that the insulation completely overlaps the open end of the small tubes. If not, the capacitor is certain to arc internally with more than a few watts of power applied to it.

Route the motor wiring inside the antenna pipes to minimize the amount of metal within the field of the antenna. Bring the wires out next to the coaxial connector. A three-wire system allows the use of limit switches to restrict the movement of the trombone section. Be sure to solder together all metal parts of the capacitor. Use a small propane torch, a good quality flux and 50/50 solid solder. Do not use acid-core solder! Clean all parts to be joined with steel wool prior to coating them with flux.

A Cookie-Sheet and Picture-Frame-Glass Variable Capacitor

In Vol 2 of *The ARRL Antenna Compendium* series, Richard Plasencia, W0RPV, described a clever high-voltage variable capacitor he constructed using readily available materials. See **Fig 22**, which shows Plasencia's homebrew high-voltage variable capacitor, along with the coil and other parts used in his homemade antenna coupler. This capacitor could be varied from 16 to 542 pF and tested at a breakdown of 12,000 V.

The capacitor sits on four PVC pillars and consists of two $4^1/_2 \times 4^1/_2$-inch aluminum plates separated by a piece of window glass that is $8^1/_2 \times 5^1/_2$ inches in size. The lower plate is epoxied to the glass. The upper plate is free to move in a wooden track epoxied to the upper surface of the glass. The motor is reversible and moves the upper capacitor plate by rotating a threaded rod in a wing nut pinned to a tab on the capacitor plate. The four pillars are cut from PVC pipe to insulate the capacitor from the chassis and to elevate it into alignment with the motor shaft.

W0RPV used a piece of 0.063-inch thick single-weight glass that exhibited a dielectric constant of 8. He removed the glass from a dime-store picture frame. In time-honored ham fashion, he improvised his wooden tracks for the upper capacitor plate from a single wooden paint stirrer, and for the capacitor plates, he used aluminum cookie sheets.

The wooden track for the upper plate is made by splitting the wooden paint stirrer with a knife into one narrow and one wide strip. The narrow strip is cemented on top and overhangs the movable plate, creating a slotted track. Since the wood is supported by the glass plate, its insulating qualities are of no importance.

The principle of operation is simple. The reversible motor turns a threaded $^1/_4$-inch rod with a pitch of 20 threads to the inch. This rod engages a wing nut attached to the movable capacitor plate. Although W0RPV grounded his capacitor's movable plate with a braid, an insulator similar to that used in the trombone capacitor above should be used to isolate the lead-screw mechanism. Several pieces of braid made from RG-8 coax shield should be used to connect to the ends of the compact transmitting loop conductors to form low-loss connections.

W0RPV used a 90-rpm motor from a surplus vending machine. It moved his variable capacitor plate $4^1/_2$ inches, taking about a minute to travel from one end to the other. Since he wished to eliminate the complexity and dubious reliability of limit switches when used outdoors, he monitored the motor's dc current through two 3 Ω, 2W resistors

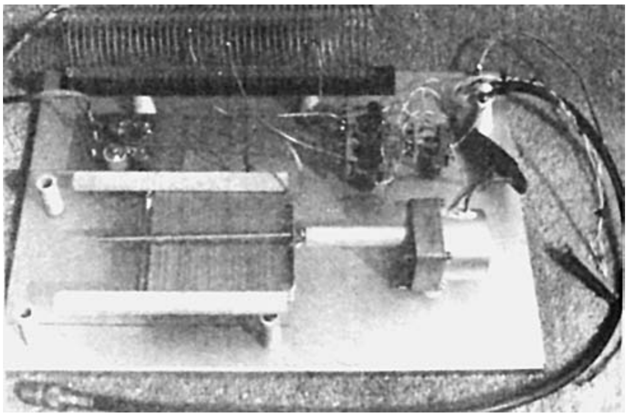

Fig 22—The picture-frame-glass variable capacitor design of Richard Plasencia, W0RPV. Two aluminum plates separated by a piece of glass scavenged from a picture frame create a variable capacitor that can withstand 12,000 V, with a variable range from 16 to 542 pF.

placed in series with each lead of the motor and shunted by red LEDs at the control box. When the motor stalled by jamming up against the PVC limit stop or against the inside of the plastic mounting box, the increased motor current caused one or the other of the LEDs to light up.

TYPICAL LOOP CONSTRUCTION

After you select the electrical design for your loop application, you must consider how to mount it and how to feed it. If you wish to cover only the upper HF bands of 20 through 10 meters, you will probably choose a loop that has a circumference of about 8.5 feet. You can make a reasonably sturdy loop using 1-inch diameter PVC pipe and 5/8-inch flexible copper tubing bent into the shape of a circle. Robert Capon, WA3ULH, did this for a QRP-level transmitting loop described in May 1994 *QST*. **Fig 23** shows a picture of his loop, with PVC H-frame stand.

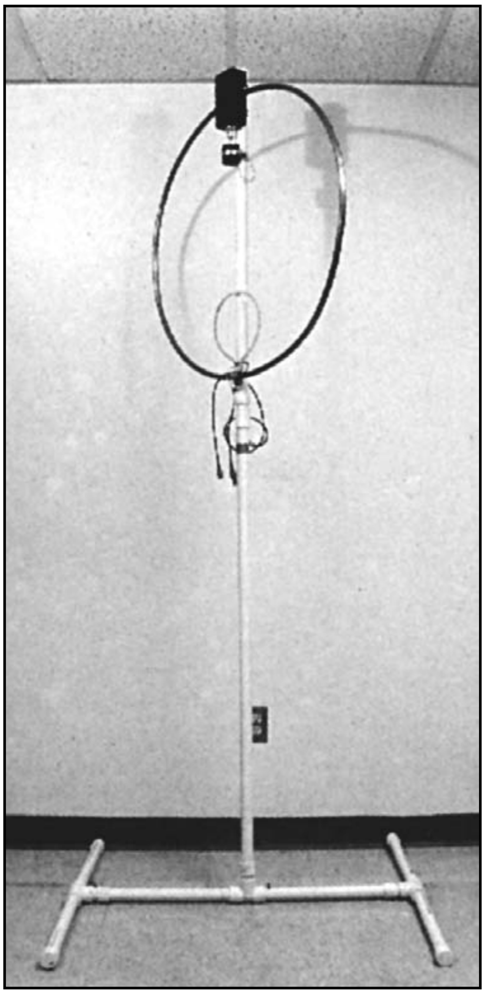

Fig 23—Photo of compact transmitting loop designed by Robert Capon, WA3ULH. This uses a 1-inch PVC H-frame to support the loop made of flexible 5/8-inch copper tubing. The small coupling loop made of RG-8 coax braid couples the loop to the coax feed line. The tuning capacitor and drive motor are at the top of the loop, shown here in the ARRL Laboratory during testing.

This loop design used a 20-inch long coupling loop made of RG-8 coax to magnetically couple into the transmitting loop rather than the gamma-match arrangement used by W5QJR in his loop designs. The coupling loop was fastened to the PVC pipe frame using 2-inch long #8 bolts that also held the main loop to the mast.

A more rugged loop can be constructed using rigid 3/4-inch copper water pipe, as shown in the W5QJR design in **Fig 24**. While a round loop is theoretically a bit more efficient, an octagonal shape is much easier to construct. The values presented in Table 4 are for octagons.

For a given loop circumference, divide the circumference by 8 and cut eight equal-length pieces of 3/4 inch copper water pipe. Join the pieces with 45° elbows to form the octagon. With the loop lying on the ground on scraps of 2 × 4 lumber, braze or solder all joints.

W5QJR made a box from clear plastic to house his air-variable capacitor and drive motor at the top of the loop. The side of the box that mounts to the loop and the capacitor should be at least ¼-inch thick, preferably 3/8-inch. The remainder of the box can be 1/8-inch plastic sheet. He mounted the loop to the plastic using 1/4-inch bolts (two on either side of center) after cutting out a section of pipe 2 inches wide in the center. On the motor side of the capacitor, he cut the pipe and installed a copper T for the motor wiring.

W5QJR's next step was to solder copper straps to the loop ends and to the capacitor stators, then he remounted the loop to the plastic. If you insert wood dowels, the pipe will remain round when you tighten the bolts. Next he installed the motor drive cable through the loop and connected it to the motor. Antenna rotator cable is a good choice for this cable. He completed the plastic box using short pieces of aluminum angle and small sheet-metal screws to join the pieces.

The loop was then ready to raise to the vertical position. Remember, no metal is allowed near the loop. W5QJR made a pole of 2 × 4-inch lumber with 1 × 4-inch boards on either side to form an I section. He held the boards together with 1/4-inch bolts, 2 feet apart and tied rope guys to the top. This made an excellent mast up to 50 feet high. The pole height should be one foot greater than the loop diameter, to allow room for cutting grass or weeds at the bottom of the loop. W5QJR installed a pulley at the top so that his loop could be raised, supported by rope. He supported the bottom of the loop by tying it to the pole and tied guy ropes to the sides of the loop to keep it from rotating in the wind. By moving the anchor points, he could rotate his loop in the azimuth plane.

W5QJR used a gamma-matching arrangement made of flexible 1/4-inch copper tubing to couple the loop to the transmission line. In the center of one leg, he cut the pipe and installed a copper T. Adjacent to the T, he installed a mount for the coax connector. He made the mount from copper strap, which can be obtained by splitting a short piece of pipe and hammering it flat.

While the loop was in the vertical position he cut a piece of 1/4-inch flexible copper tubing the length of one

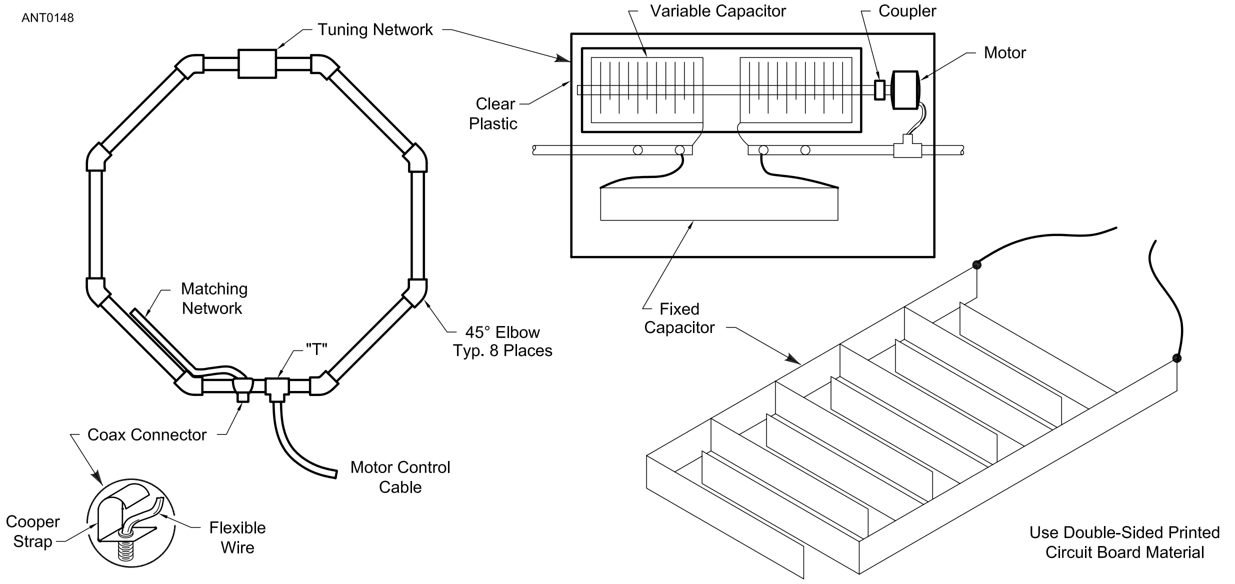

Fig 24—Octagonal loop construction details. Table 4 gives loop design data for various frequency ranges.

of the straight sides of the loop. He then flattened one end and soldered a piece of flexible wire to the other. He wrapped the tubing with electrical tape for insulation and connected the flexible wire to the coax connector. He then installed the tubing against the inside of the loop, held temporarily in place with tape. He soldered the flat part to the loop, ending up with a form of gamma match, but without reactive components. This simple feed provided better than 1.7:1 SWR over a 2:1 frequency range. For safety, he installed a good ground rod under the loop and connected it to the strap for the coax connector, using large flexible wire.

TUNE-UP PROCEDURE

The resonant frequency of the loop can be readily found by setting the receiver to a desired frequency and rotating the capacitor (by remote control) until signals peak. The peak will be very sharp because of the high Q of the loop.

Turn on the transmitter in the tune mode and adjust either the transmitter frequency or the loop capacitor for maximum signal on a field-strength meter, or for maximum forward signal on an SWR bridge. Adjust the matching network for minimum SWR by bending the matching line. Normally a small hump in the $^{1}/_{4}$-inch tubing line, as shown in Fig 24, will give the desired results. For a loop that covers two or more bands, adjust the feed to give equally low SWR at each end of the tubing range. The SWR will be very low in the center of the tuning range but will rise at each end.

If there is metal near the loop, the additional loss will reduce the Q and therefore the impedance of the loop. In those cases it will be necessary to increase the length of the matching line and tap higher up on the loop to obtain a 50-Ω match.

PERFORMANCE COMPARISON

As previously indicated, a compact transmitting loop can provide performance approaching full-size dipoles and verticals. To illustrate one case, a loop 100 feet in circumference would be 30 feet high for 1.8 MHz. However, a good dipole would be 240 feet ($^{1}/_{2}\ \lambda$) in length and at least 120 feet high ($^{1}/_{4}\ \lambda$). A $^{1}/_{4}$-λ vertical would be 120 feet tall with a large number of radials on the ground, each 120 feet in length. The smaller loop could replace both of those antennas with only a moderate degradation in performance and a requirement for a high-voltage variable capacitor.

On the higher frequencies, the same ratios apply, but full-size antennas are less dramatic. However, very few city dwellers can erect good verticals even on 7 MHz with a full-size counterpoise. Even on 14 MHz a loop about 3 feet high can work the world.

Other than trading small size for narrow bandwidth and a high-voltage capacitor, the compact transmitting loop is an excellent antenna and should find use where large antennas are not practical.

The Loop Skywire

Are you looking for a multiband HF antenna that is easy to construct, costs nearly nothing and yet works well? You might want to try this one. The *Loop Skywire* antenna is a full-sized horizontal loop. Early proponents suggested that the antenna could be fed with coaxial cable with little concern for losses, but later analysis proved that this was a bit of wishful thinking—the relatively low values for SWR across multiple bands indicate that cable losses were part and parcel of performance. The best way to feed this versatile antenna is with open-wire ladder line, with an antenna tuner in the shack to present the transmitter with a low value of SWR.

THE DESIGN

The Loop Skywire is shown in **Fig 25**. The antenna has one wavelength of wire in its perimeter at the design or fundamental frequency. If you choose to calculate L_{total} in feet, the following equation should be used:

$$L_{total} = \frac{1005}{f}$$

where f equals the frequency in MHz.

Given any length of wire, the maximum possible area the antenna can enclose is with the wire in the shape of a circle. Since it takes an infinite number of supports to hang a circular loop, the square loop (four supports) is the most practical. Further reducing the area enclosed by the wire loop (fewer supports) brings the antenna closer to the properties of the folded dipole, and both harmonic-impedance and feed-line voltage problems can result. Loop geometries other than a square are thus possible, but remember the two fundamental requirements for the Loop Skywire—its horizontal position and maximum enclosed area.

There is another great advantage to this antenna system. It can be operated as a vertical antenna with top-hat loading on other bands as well. This is accomplished by simply keeping the feed line run from the antenna to the shack as vertical as possible and clear of objects. Both feed-line conductors are then tied together, and the antenna is fed against a good ground.

CONSTRUCTION

Antenna construction is simple. Although the loop can be made for any band or frequency of operation, the following two Loop Skywires are good performers. The 10-MHz band can also be operated on both.

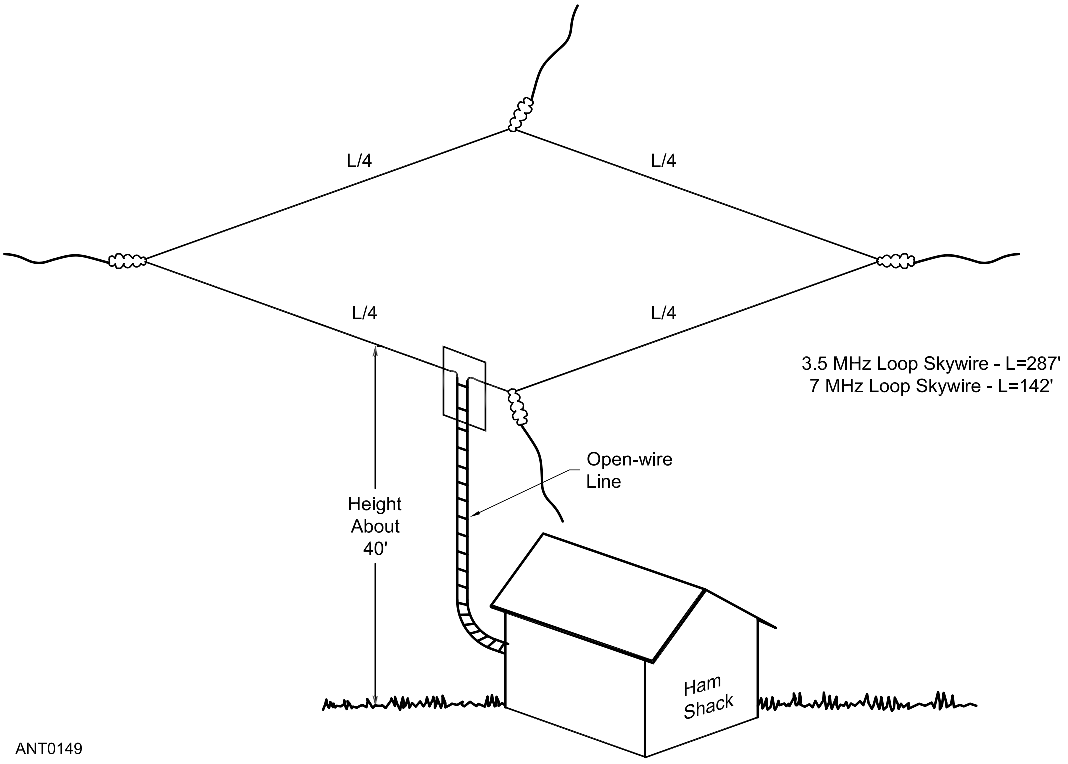

Fig 25—A complete view of the Loop Skywire. The square loop is erected horizontal to the earth.

3.5-MHz Loop Skywire
(3.5-28 MHz loop and 1.8-MHz vertical)
Total loop perimeter: 272 feet
Square side length: 68 feet

7-MHz Loop Skywire
(7-28 MHz loop and 3.5-MHz vertical)
Total loop perimeter: 142 feet
Square side length: 35.5 feet

The actual total length can vary from the above by a few feet, as the length is not at all critical. Do not worry about tuning and pruning the loop to resonance. No signal difference will be detected on the other end when that method is used.

Bare #14 copper wire is used in the loop. **Fig 26** shows the placement of the insulators at the loop corners. Two common methods are used to attach the insulators. Either lock or tie the insulator in place with a loop wire tie, as shown in Fig 26A, or leave the insulator free to "float" or slide along the wire, Fig 26B. Most loop users float at least two insulators. This allows pulling the slack out of the loop once it is in the air, and eliminates the need to have all the supports exactly placed for proper tension in each leg. Floating two opposite corners is recommended.

Fig 27A shows the azimuth-plane performance on 7.2 MHz of a 142-foot long, 7-MHz Loop Skywire, 40 feet high at an elevation angle of 10º, compared to a regular flattop $^{1}/_{2}$-λ dipole at a height of 30 feet. The loop comes into its own at higher frequencies. Fig 27B shows the response at 14.2 MHz, compared again to a $^{1}/_{2}$–λ 14.2-MHz

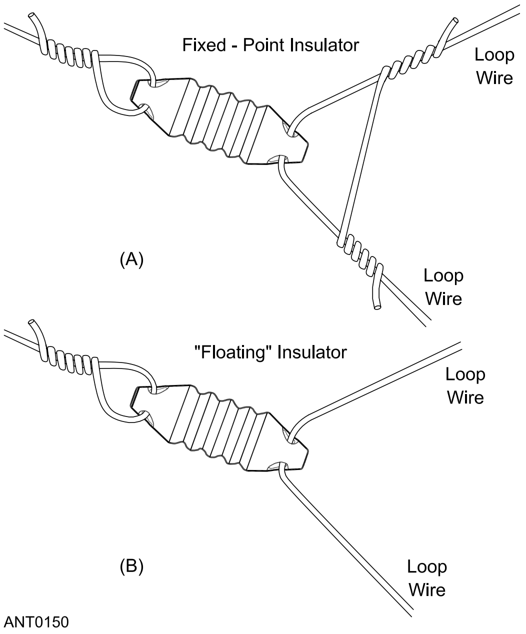

Fig 26—Two methods of installing the insulators at the loop corners.

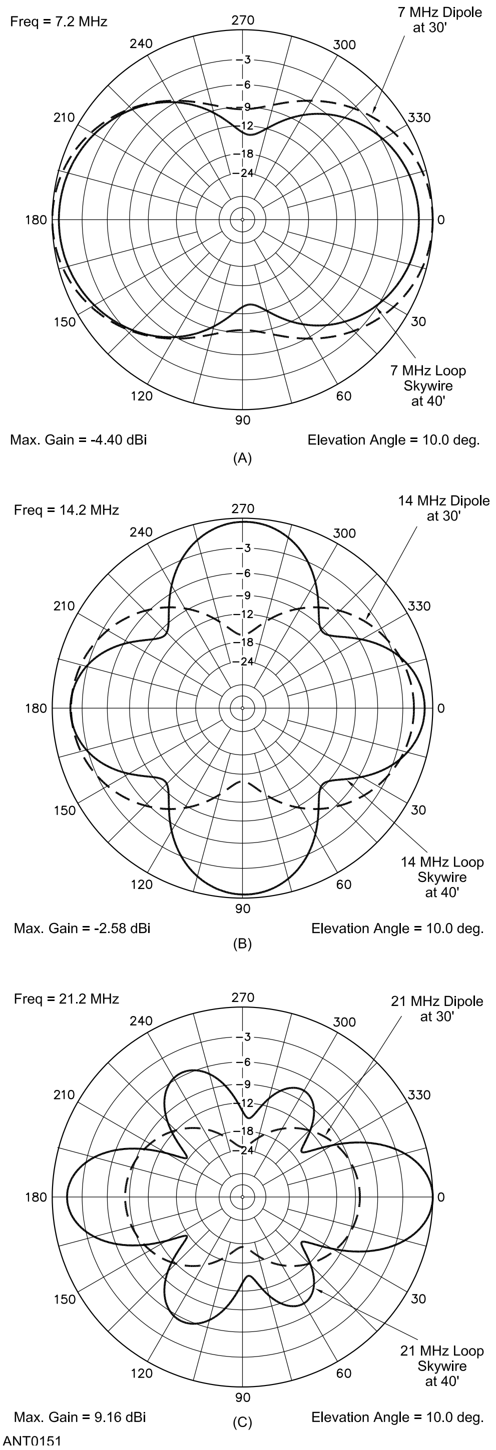

Fig 27—At A, azimuth-plane response of 142-foot long, 7-MHz Loop Skywire, 40 feet in the air at 7.2 MHz, compared with ½-λ dipole 30 feet in the air. At B, response of same Loop Skywire at 14.2 MHz, compared with ½-λ 14.2-MHz dipole 30 feet in the air. Now the loop has some advantage in certain directions. At C, response of the same Loop Skywire at 21.2 MHz compared to a 21.2-MHz dipole at 30 feet. Here, the Loop Skywire has more gain in almost all directions than the simple dipole. All azimuth-plane patterns were made at 10° elevation.

dipole at a height of 30 feet. Now the loop has several lobes that are stronger than the dipole. Fig 27C shows the response at 21.2 MHz, compared to a dipole. Now the loop has superior gain compared to the $^1/_2$-λ dipole at almost any azimuth. In its favored direction on 21.2 MHz, the loop is 8 dB stronger than the dipole.

The feed point can be positioned anywhere along the loop that you wish. However, most users feed the Skywire at a corner. **Fig 28** depicts a method of doing this, using a piece of plexiglass to provide insulation as well as strain relief for the open-wire ladder line. It is advantageous to keep the feed-point mechanicals away from the corner support. Feeding a foot or so from one corner allows the feed line to exit more freely. This method keeps the feed line free from the loop support.

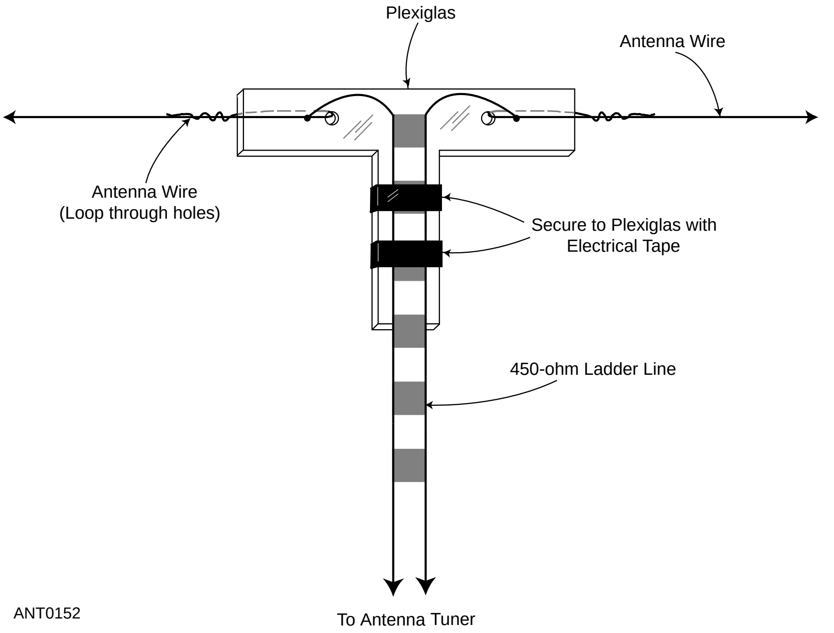

Fig 28—Most users feed the Skywire at a corner. A high-impedance weather-resistant insulator should be used for the feed-point insulator.

Generally a minimum of four supports is required. If trees are used for supports, then at least two of the ropes or guys used to support the insulators should be counterweighted and allowed to move freely. The feedline corner is almost always tied down, however. Very little tension is needed to support the loop (far less than that for a dipole). Thus, counterweights are light. Several such loops have been constructed with bungie cords tied to three of the four insulators. This eliminates the need for counterweighting.

Recommended height for the antenna is 40 feet or more. The higher the better, especially if you wish to use the loop in the vertical mode. However, successful local and DX operation has been reported in several cases with the antenna at 20 feet. **Fig 29** shows the feed arrangement for using the Loop Skywire as a top-loaded vertical fed against ground on the lower bands.

Because the loop is high in the air and has considerable electrical exposure to the elements, proper methods should be employed to eliminate the chance of induced or direct lightning hazard to the shack and operator. Some users simply completely disconnect the antenna from the antenna tuner and rig and shack during periods of possible lightning activity.

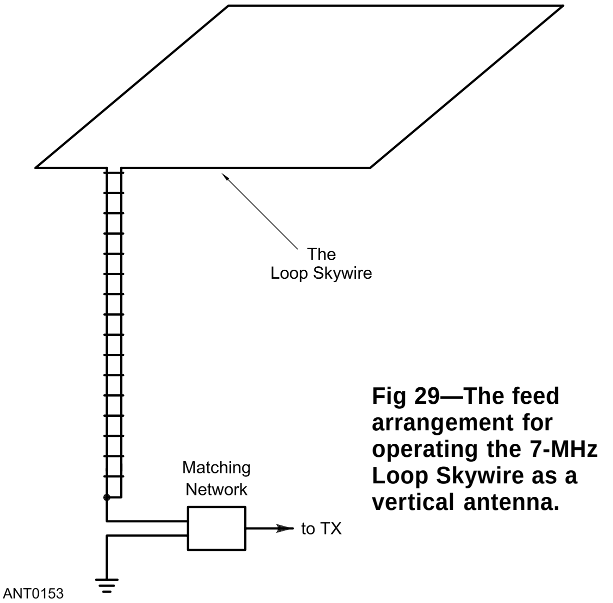

Fig 29—The feed arrangement for operating the 7-MHz Loop Skywire as a vertical antenna.

7-MHz Loop

An effective but simple 7MHz antenna that has a theoretical gain of approximately 1 dB over a dipole is a full-wave, closed vertical loop. Such a loop need not be square, as illustrated in **Fig 30A**. It can be trapezoidal, rectangular, circular, or some distorted configuration in between those shapes. For best results, however, you should attempt to make the loop as square as possible. The more rectangular the shape, the greater the cancellation of energy in the system, and the less effective it will be. In the limiting case, the antenna loses its identity as a loop and becomes a folded dipole.

You can feed the loop in the center of one of the vertical sides if you want vertical polarization. For horizontal polarization, you feed either of the horizontal sides at the center. Since optimum directivity occurs at right angles to the plane of the loop (or in more simple terms, broadside to the loop), you should hang the loop to radiate the maximum amount in some favored direction.

Fig 31A shows the azimuthal response at a takeoff angle of 15°, a typical angle for 40-meter DX, for vertical and horizontal feed systems over ground with “average” conductivity and dielectric constant. Fig 31A includes, for reference, the response of a flattop dipole 50 feet high. For DX work on 40 meters, the vertically polarized loop can perform as well as or substantially better than either a horizontally polarized loop or a flattop dipole, particularly in the azimuthal nulls of the dipole.

For the low elevation angles that favor DX work, the optimal feed point is at the center of one of the vertical wires. Feeding the loop at one of the corners at the bottom gives a

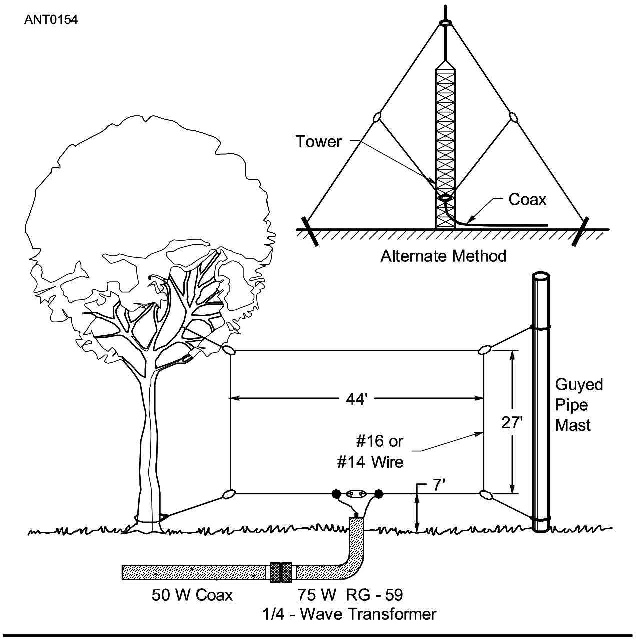

Fig 30—At A, details of the rectangular full-wave loop. The dimensions given are for operation at 7.05 MHz. The height above ground was 7 feet in this instance, although improved performance should result if the builder can install the loop higher above ground without sacrificing length on the vertical sides. At B, illustration how a single supporting structure can be used to hold the loop in a diamond-shaped configuration. Feeding the diamond at the lower tip provides radiation in the horizontal plane. Feeding the system at either side will result in vertical polarization of the radiated signal.

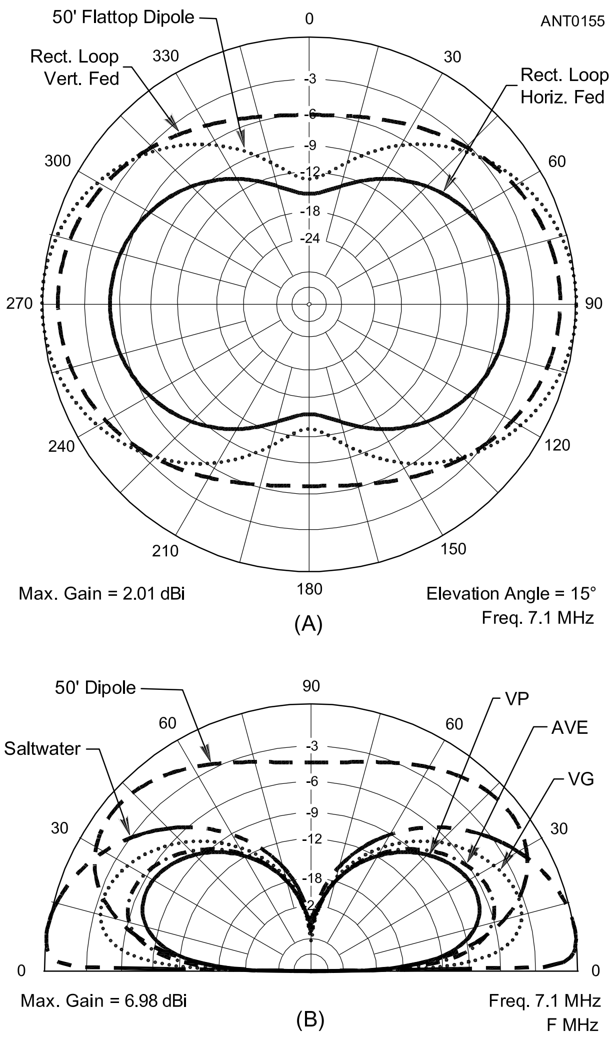

Fig 31—At A, azimuthal plane responses for the vertically and horizontally polarized 7-MHz loop, compared to a flattop 50-foot high dipole, all at a takeoff angle of 15° for DX work. The solid line is for feeding the loop horizontally at the bottom; the dashed line is for feeding the loop vertically at a side, and the dotted line is for a simple flattop horizontal dipole at 50 feet in height. For DX work, the vertically polarized loop is an excellent performer.

compromise result for both local and DX work. The actual impedance is roughly the same at each point: bottom horizontal center, corner or vertical side center.

Fig 31B demonstrates how the gain for vertical polarization changes over different type of grounds: saltwater, very poor ground (conductivity = 1 mS/m, dielectric constant = 5) very good (conductivity = 30 mS/m, dielectric constant = 20) and average ground (conductivity = 5 mS/m, dielectric constant = 13). Again, for reference a 50-foot high flattop dipole's elevation response is included. As has been mentioned previously in other chapters, a seaside location is a wonderful environment for verticals!

Just how you erect such a loop will depend on what is available in your backyard. Trees are always handy for supporting loop antennas. A disadvantage to the rectangular loop shown in Fig 30A is that two 34-foot high supports are needed, although in many instances your house may be high enough to serve as one of these supports. If you have a tower higher than about 50 feet, Fig 30B demonstrates how you can use it to support a diamond-shaped loop for 40 meters. The elevation and azimuthal responses are almost the same for either loop configuration, rectangular- or diamond-shaped.

The overall length of the wire used in a loop is determined in feet from the formula 1005/f (MHz). Hence, for operation at 7.125 MHz the overall wire length will be 141 feet. The matching transformer, an electrical 1/4 λ of 75Ω coax cable, can be computed by dividing 246 by the operating frequency in MHz, then multiplying that number by the velocity factor of the cable being used. Thus, for operation at 7.125 MHz, 246/7.125 MHz = 34.53 feet. If coax with solid polyethylene insulation is used, a velocity factor of 0.66 must be employed. Foam-polyethylene coax has a velocity factor of 0.80. Assuming RG-59 is used, the length of the matching transformer becomes 34.53 (feet) × 0.66 = 22.79 feet, or 22 feet, 9½ inches.

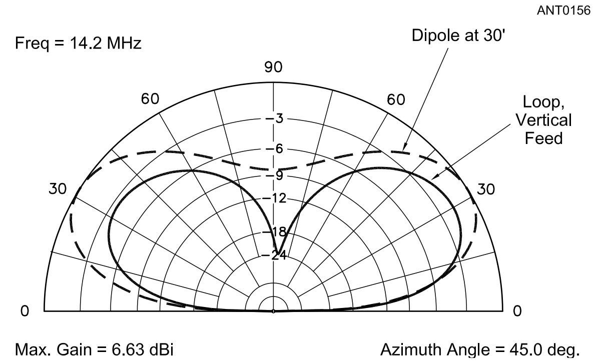

Fig 32—Elevation-plane response of 7-MHz loop used on 14.2 MHz. This is for a feed point at the center of one of the two vertical wires. The dashed line is the response of a flattop 20-meter dipole at 30 feet in height for comparison.

This same loop antenna in Fig 30A fed vertically may be used on the 14 and 21MHz bands, although its pattern will not be as good as that on its fundamental frequency and you will have to use an open-wire transmission line to feed the loop for multiband use. **Fig 32** shows the response at the peak lobe of the loop, at a 45° angle to the plane of the loop, compared to the peak response for a simple halfwave 20-meter dipole, 30 feet high. The gain from a simple flattop dipole, mounted at 30 feet, will be superior to the loop operated on a harmonic frequency.

A Receiving Loop for 1.8 MHz

You can use a small shielded-loop antenna to improve reception under certain conditions, especially at the lower amateur frequencies. This is particularly true when high levels of man-made noise are prevalent, when the second-harmonic energy from a nearby broadcast station falls in the 1.8MHz band, or when interference exists from some other amateur station in the immediate area. A properly constructed and tuned small loop will exhibit approximately 30 dB of front-to-side response, the minimum response being at right angles to the plane of the loop. Therefore, noise and interference can be reduced significantly or completely nulled out, by rotating the loop so that it is sideways to the interference-causing source.

Generally speaking, small shielded loops are far less responsive to man-made noise than are the larger antennas used for transmitting and receiving. But a trade-off in performance must be accepted when using the loop, for the strength of received signals will be 10 or 15 dB less than when using a full-size resonant antenna. This condition is not a handicap on 1.8 or 3.5 MHz, provided the station receiver has normal sensitivity and overall gain. Because a front-to-side ratio of 30 dB may be expected, a shielded loop can be used to eliminate a variety of receiving problems if made rotatable, as shown in **Fig 33**.

To obtain the sharp bidirectional pattern of a small loop, the overall length of the conductor must not exceed 0.1 λ. The loop of **Fig 34** has a conductor length of 20 feet. At 1.81 MHz, 20 feet is 0.037 λ. With this style of loop, 0.037 λ is about the maximum practical dimension if you want to tune the element to resonance. This limitation results from the distributed capacitance between the shield and inner conductor of the loop. RG-59 was used for the loop element in this example. The capacitance per foot for this cable is 21 pF, resulting in a total distributed

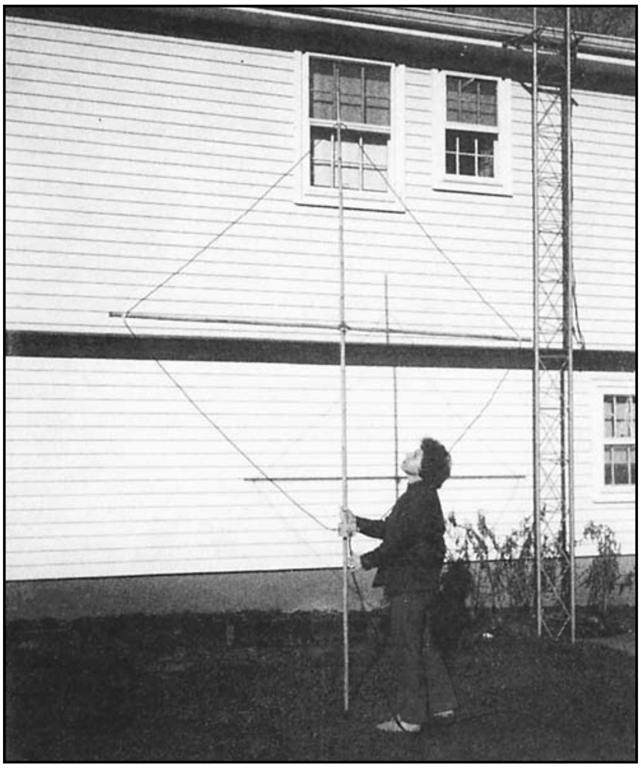

Fig 33—Jean DeMaw, W1CKK, tests the 1.8-MHz shielded loop. Bamboo cross arms are used to support the antenna.

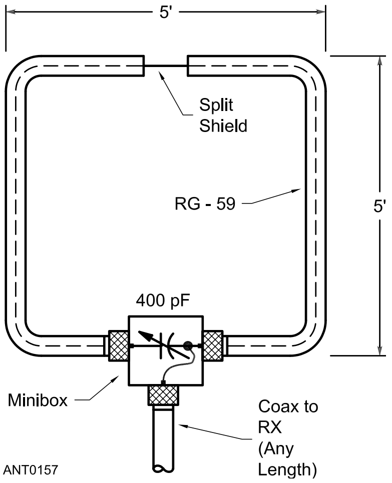

Fig 34—Schematic diagram of the loop antenna. The dimensions are not critical provided overall length of the loop element does not exceed approximately 0.1 λ. Small loops which are one half or less the size of this one will prove useful where limited space is a consideration.

capacitance of 420 pF. An additional 100 pF was needed to resonate the loop at 1.810 MHz.

Therefore, the approximate inductance of the loop is 15 µH. The effect of the capacitance becomes less pronounced at the higher end of the HF spectrum, provided the same percentage of a wavelength is used in computing the conductor length. The ratio between the distributed capacitance and the lumped capacitance used at the feed point becomes greater at resonance. These facts should be contemplated when scaling the loop to those bands above 1.8 MHz.

There will not be a major difference in the construction requirements of the loop if coaxial cables other than RG-59 are used. The line impedance is not significant with respect to the loop element. Various types of coaxial line exhibit different amounts of capacitance per foot, however, thereby requiring more or less capacitance across the feed point to establish resonance.

Shielded loops are not affected noticeably by nearby objects, and therefore they can be installed indoors or out after being tuned to resonance. Moving them from one place to another does not significantly affect the tuning.

You can see in the model shown in Fig 33 that a supporting structure was fashioned from bamboo poles. The X frame is held together at the center with two U bolts. The loop element is taped to the cross-arms to form a square. You could likely use metal cross arms without seriously degrading the antenna performance. Alternatively, wood can be used for the supporting frame.

A Minibox was used at the feed point of the loop to hold the resonating variable capacitor. In this model a 50 to 400-pF compression trimmer was used to establish resonance. You must weatherproof the box for outdoor installations.

Remove the shield braid of the loop coax for one inch directly opposite the feed point. You should treat the exposed areas with a sealing compound once this is done.

In operation this receiving loop has proven very effective for nulling out second-harmonic energy from local broadcast stations. During DX and contest operations on 1.8 MHz it helped prevent receiver overloading from nearby 1.8-MHz stations that share the band. The marked reduction in response to noise has made the loop a valuable station accessory when receiving weak signals. It is not used all of the time, but is available when needed by connecting it to the receiver through an antenna selector switch. Reception of European stations with the loop has been possible from New England at times when other antennas were totally ineffective because of noise.

It was also discovered that the effects of approaching storms (with attendant atmospheric noise) could be nullified considerably by rotating the loop away from the storm front. It should be said that the loop does not exhibit meaningful directivity when receiving sky-wave signals. The directivity characteristics relate primarily to ground-wave signals. This is a bonus feature in disguise, for when nulling out local noise or interference, one is still able to copy sky-wave signals from all compass points!

For receiving applications it is not necessary to match the feed line to the loop, though doing so may enhance the performance somewhat. If no attempt is made to obtain an SWR of 1, the builder can use 50 or 75-Ω coax for a feeder, and no difference in performance will be observed. The Q of this loop is sufficiently low to allow the operator to peak it for resonance at 1.9 MHz and use it across the entire 1.8MHz band. The degradation in performance at 1.8 and 2 MHz will be so slight that it will be difficult to discern.

An Indoor Stealth Loop

Ted Phelps, W8TP, wrote an article in *The ARRL Antenna Compendium, Vol 7* describing his attic-mounted wire loop antenna, fed with an automatic antenna tuner. Here is a shortened version of that article.

If you drive down my street in Whitechapel Village in Newark, DE, trying to find my ham location by looking for my antenna, you wouldn't find it. Even if you pulled up in front of my condo, you wouldn't notice any telltale signs, because my multiband antenna is completely hidden. It's in the attic of my two-bedroom condominium in a small retirement community completed in 1999.

Before the move, I had given considerable thought to what type of antenna I might use, if any. I already knew that permanent outdoor types were out of the question, due to restrictive real estate and condo association rules. So I planned a clause for any sales contract I might sign, specifically mentioning amateur radio and my desire to set up a station in my new living quarters. I decided I would not move where my lifelong hobby would be severely restricted or prohibited.

That meant that to be reasonably sure I could continue enjoying Amateur Radio as before, I would have to install an indoor antenna that could perform as well as a typical outdoor system. What kind? In Ohio, I had tried a horizontally polarized attic dipole made with #14 wire. It didn't work very well—it was just too low to the ground.

I learned about a high-tech method of remote antenna tuning using an antenna coupler which contains a microprocessor. I found this kind of automatic tuner available from two American manufacturers and within a reasonable price range. Although our move was still a few months ahead, I purchased a model SG-230 antenna coupler made by SGC Inc, Bellevue, WA, for use in Delaware.

Fig 35 shows the final dimensions of my hidden loop, which is a single-turn rectangular loop, erected in a north-south vertical plane and made from nearly 78 feet of #6 stranded, aircraft primary wire in a PVC jacket, held taut at the lower corners and supported by a pulley and guy rope at each upper corner. Because it's vertically polarized, it supports low-angle radiation reasonably well. By the way, if you're wondering why I used such a relatively large-gauge wire as #6 for the loop antenna, it was readily available from my son-in-law!

Fig 36 shows my completed condo unit. Note the doghouse dormer on the roof about 12 feet above ground at the attic floor. This is the level of my hidden loop's base leg.

In constructing my system I had to overcome RFI problems on my own premises. Each condo unit has its own electronic security panel on an upper shelf in a closet. As soon as I applied moderate power to my radio and loop, the Fire Alarm sounded and firefighters came to my door! The burglar/intrusion signal was triggered a couple of times, too. Working with a security installation technician, I found that *there was no ground wire connected to my security panel.* "We don't bother with that," said the tech, and then, reacting to my surprise, connected a #14 ground between the security panel and the house water-pipe ground. I then installed a ferrite bead on each lead entering the security panel. I also placed ferrite beads on keyer-paddle leads, GFCI electrical outlets, etc. Those

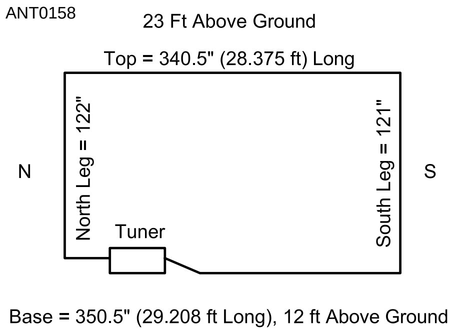
ANT0158
23 Ft Above Ground
Top = 340.5" (28.375 ft) Long
North Leg = 122"
South Leg = 121"
N
S
Tuner
Base = 350.5" (29.208 ft Long), 12 ft Above Ground

Fig 35—Diagram showing layout of W8TP's indoor hidden loop antenna.

Fig 36—Can you see W8TP's antenna in this photograph? Of course you can't—it's hidden from view inside his attic!

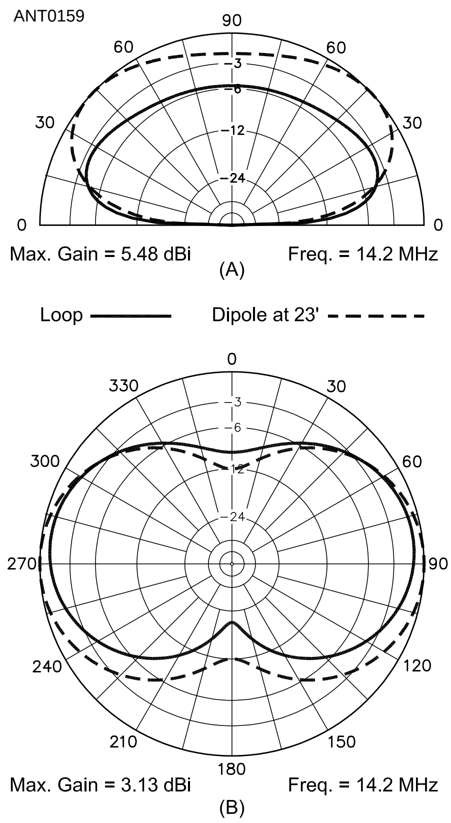

Fig 37—At A, computed elevation pattern at 14.2 MHz for W8TP's hidden loop (solid line), compared to a 20-meter dipole (dashed line) at a height of 23 feet. At B, a comparison of the azimuth patterns at a 20°-elevation angle for W8TP's loop (solid line) and the same 20-meter dipole (dashed line). The loop has a slightly asymmetrical response because it is fed at a corner, but its performance is competitive to an outdoor dipole. In fact, it has superior low-angle performance, typical of a vertically polarized antenna compared to a low horizontal antenna.

measures seem to have eliminated my RFI problems.

When we moved into our condo, I took the obvious precaution of not using a linear amplifier. I took extra care to establish a single-point ground for my station equipment by connecting all equipment grounds to the cover plate of the dedicated metallic outlet box behind the operating position, and thence to a separate ground rod in our front yard. I use a 1-kW RL Drake low-pass filter in the transceiver-antenna feed line.

Is this indoor antenna system safe? I believe so. In the attic it is not at all close to our living space. It is fixed firmly in place and unlike most amateur antennas it is *out of the weather*! I therefore do not use a quick-grounding system for times when a thunderstorm approaches.

Fig 37 shows the computed elevation and azimuth patterns on 20 meters. The tuner is able to hold the SWR down low enough so that my JRC-245 transceiver can operate through its internal antenna tuner.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

C. F. W. Anderson, "A Crossed-Loop/Goniometer DF Antenna for 160 Meters," *The ARRL Antenna Compendium, Vol 1* (Newington: ARRL, 1985), pp 127-132.

J. S. Belrose, "Ferromagnetic Loop Aerials," *Wireless Engineer*, Feb 1955, pp 41-46.

J. S. Belrose, "An Update on Compact Transmitting Loops," *QST*, Nov 1993, pp 37-40.

D. S. Bond, *Radio Direction Finders*, 1st ed. (New York: McGraw-Hill Book Co, 1944).

G. Bramslev, "Loop Aerial Reception," *Wireless World*, Nov 1952, pp 469-472.

R. W. Burhans, "Experimental Loop Antennas for 60 kHz to 200 kHz," *Technical Memorandum (NASA) 71*, (Athens, OH: Ohio Univ, Dept of Electrical Engr), Dec 1979.

R. W. Burhans, "Loop Antennas for VLF-LF," *Radio-Electronics*, Jun 1983, pp 83-87.

D. DeMaw, "Beat the Noise with a Scoop Loop," *QST*, Jul 1977, pp 30-34.

M. F. DeMaw, *Ferromagnetic-Core Design and Application Handbook* (Englewood Cliffs, NJ: Prentice-Hall Inc, 1981).

D. DeMaw and L. Aurick, "The Full-Wave Delta Loop at Low Height," *QST*, Oct 1984, pp 24-26.

R. Devore and P. Bohley, "The Electrically Small Magnetically Loaded Multiturn Loop Antenna," *IEEE Trans on Ant and Prop*, Jul 1977, pp 496-505.

T. Dorbuck, "Radio Direction-Finding Techniques," *QST*, Aug 1975, pp 30-36.

R. G. Fenwick, "A Loop Array for 160 Meters," *CQ*, Apr 1986, pp 25-29.

D. Fischer, "The Loop Skywire," *QST*, Nov 1985, pp 20-22. Also "Feedback," *QST*, Dec 1985, p 53.

R. A. Genaille, "V.L.F. Loop Antenna," *Electronics World*, Jan 1963, pp 49-52.

G. Gercke, "Radio Direction/Range Finder," *73*, Dec 1971, pp 29-30.

R. S. Glasgow, *Principles of Radio Engineering* (New York: McGraw-Hill Book Co, Inc, 1936).

S. Goldman, "A Shielded Loop for Low Noise Broadcast Reception," *Electronics*, Oct 1938, pp 20-22.

F. W. Grover, *Inductance Calculation-Working Formulas and Tables* (New York: D. VanNostrand Co, Inc, 1946).

J. V. Hagan, "A Large Aperture Ferrite Core Loop Antenna for Long and Medium Wave Reception," *Loop Antennas Design and Theory*, M.G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 37-49.

T. Hart, "Small, High-Efficiency Loop Antennas," *QST*, Jun 1986, pp 33-36.

T. Hart, *Small High Efficiency Antennas Alias The Loop* (Melbourne, FL: W5QJR Antenna Products, 1985)

H. Hawkins, "A Low Budget, Rotatable 17 Meter Loop," *QST*, Nov 1997, p 35.

F. M. Howes and F. M. Wood, "Note on the Bearing Error and Sensitivity of a Loop Antenna in an Abnormally Polarized Field," *Proc IRE*, Apr 1944, pp 231-233.

B. Jones, KD7S, "You Can Build: A Compact Loop Antenna for 30 Through 12 Meters," Nov 1994 *QST*, pp 33-36.

J. A. Lambert, "A Directional Active Loop Receiving Antenna System," *Radio Communication*, Nov 1982, pp 944-949.

D. Lankford, "Loop Antennas, Theory and Practice," *Loop Antennas Design and Theory*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 10-22.

D. Lankford, "Multi-Rod Ferrite Loop Antennas," *Loop Antennas Design and Theory*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 53-56.

G. Levy, "Loop Antennas for Aircraft," *Proc IRE*, Feb 1943, pp 56-66. Also see correction, *Proc IRE*, Jul 1943, p 384.

J. Malone, "Can a 7 foot 40m Antenna Work?" *73*, Mar 1975, pp 33-38.

R. Q. Marris, "An In-Room 80-Meter Transmitting Multiturn Loop Antenna," *QST*, Feb 1996, pp 43-45. Feedback. *QST*, May 1996, p 48.

L. G. McCoy, "The Army Loop in Ham Communications," *QST*, Mar 1968, pp 17, 18, 150, 152. (See also Technical Correspondence, *QST*, May 1968, pp 49-51 and Nov 1968, pp 46-47.)

R. G. Medhurst, "HF Resistance and Self Capacitance of Single Layer Solenoids," *Wireless Engineer*, Feb 1947, pp 35-43, and Mar 1947, pp 80-92.

G. P. Nelson, "The NRC FET Altazimuth Loop Antenna," *N.R.C. Antenna Reference Manual Vol. 1*, 4th ed., R. J. Edmunds, Ed. (Cambridge, WI: National Radio Club, 1982), pp 2-18.

K. H. Patterson, "Down-To-Earth Army Antenna," *Electronics*, Aug 21, 1967, pp 111-114.

R. C. Pettengill, H. T. Garland and J. D. Meindl, "Receiving Antenna Design for Miniature Receivers," *IEEE Trans on Ant and Prop*, Jul 1977, pp 528-530.

T. Phelps, W8TP, "A Hidden Loop Antenna," *The ARRL Antenna Compendium, Vol 7* (Newington: ARRL, 2002), pp 160-162.

R. Z. Plasencia, WØRPV, "Remotely Controlled Antenna Coupler," *The ARRL Antenna Compendium, Vol 2* (Newington: ARRL, 1989), pp 182-186.

W. J. Polydoroff, *High Frequency Magnetic Materials—Their Characteristics and Principal Applications* (New York: John Wiley and Sons, Inc, 1960).

E. Robberson, "QRM? Get Looped," *Radio and Television News*, Aug 1955, pp 52-54, 126.

G. S. Smith, "Radiation Efficiency of Electrically Small Multiturn Loop Antennas," *IEEE Trans on Ant and Prop*, Sep 1972, pp 656-657.

E. C. Snelling, *Soft Ferrites—Properties and Applications* (Cleveland, OH: CRC Press, 1969).

G. Thomas, "The Hot Rod—An Inexpensive Ferrite Booster Antenna," *Loop Antennas Theory and Design*, M. G. Knitter, Ed. (Cambridge, WI: National Radio Club, 1983), pp 57-62.

J. R. True, "Low-Frequency Loop Antennas," *Ham Radio*, Dec 1976, pp 18-24.

E. G. VonWald, "Small-Loop Antennas," *Ham Radio*, May 1972, pp 36-41.

"An FET Loop Amplifier with Coaxial Output," *N.R.C. Antenna Reference Manual, Vol 2*, 1st ed., R. J. Edmunds, Ed. (Cambridge, WI: National Radio Club, Oct 1982), pp 17-20.

Aperiodic Loop Antenna Arrays (Hermes Electronics Ltd, Nov 1973).

Reference Data for Radio Engineers, 6th ed. (Indianapolis: Howard W. Sams & Co, subsidiary of ITT, 1977).

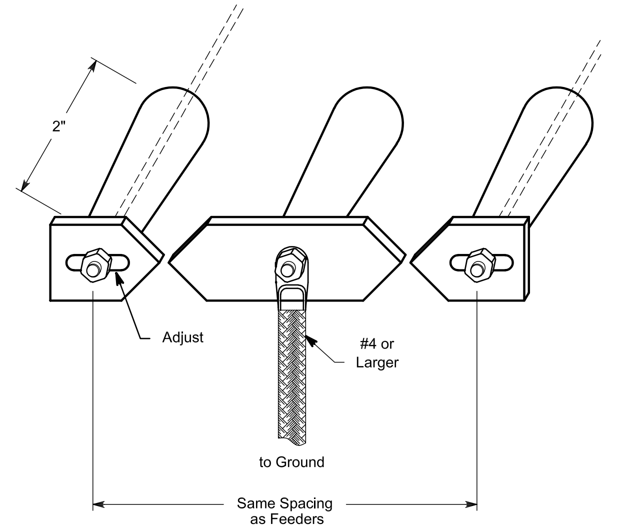

Fig 43—A simple lightning arrester for open-wire line made from three standoff or feedthrough insulators and sections of $^1/_8 \times ^1/_2$-inch brass or copper strap. It should be installed in the line at the point where the line enters the station. The heavy ground lead should be as short and as direct as possible. The gap setting should be adjusted to the minimum width that will prohibit arcing when the transmitter is operated.

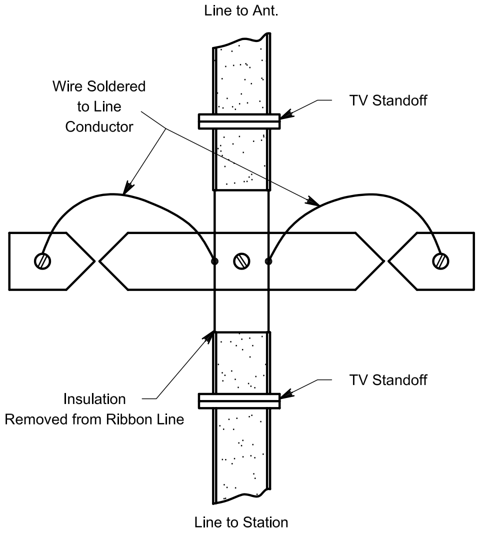
Line to Ant.
Wire Soldered to Line Conductor
TV Standoff
TV Standoff
Insulation Removed from Ribbon Line
Line to Station

Fig 44—The lightning arrester of Fig 39 may be used with 300-Ω ribbon line in the manner shown here. The TV standoffs support the line an inch or so away from the grounded center member of the arrester

The construction of a homemade arrester for open-wire line is shown in **Fig 43**. This type of arrester can be adapted to ribbon line an inch or so away from the center member of the arrester, as shown in **Fig 44**. Sufficient insulation should be removed from the line where it crosses the arrester to permit soldering the arrester connecting leads.

Lightning Grounds

Lightning-ground connecting leads should be of conductor size equivalent to at least #10 wire. The #8 aluminum wire used for TV-antenna grounds is satisfactory. Copper braid $^3/_4$-inch wide (Belden 8662-10) is also suitable. The conductor should run in a straight line to the grounding point. The ground connection may be made to a water pipe system (if the pipe is not plastic), the grounded metal frame of a building, or to one or more $^5/_8$-inch ground rods driven to a depth of at least 8 feet. More detailed information on lightning protection is contained in Chapter 1, Safety.

A central grounding panel for coax cables coming into the house is highly recommended. See **Fig 45** for a photo of the homemade grounding panel installed by Chuck Hutchinson, K8CH, at his Michigan home. The coax cables screwed into dual-female feed-through UHF connectors. K8CH installed this aluminum panel under the outside grill for a duct that provided combustion air to an unused fireplace. He used ground strap to connect to ground rods located under the panel. See the ARRL book *Simple and Fun Antennas* for more information about ground panels.

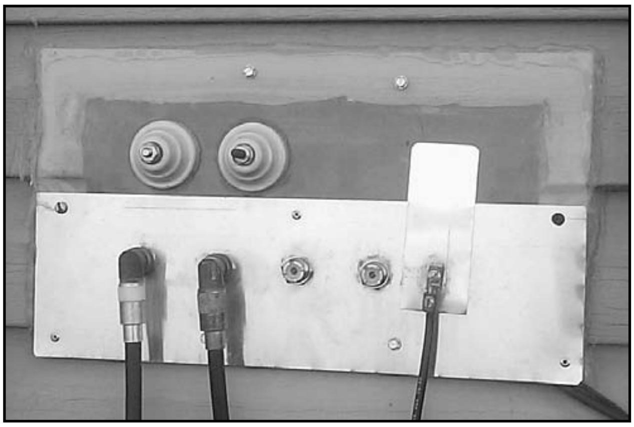

Fig 45—K8CH's coax entry panel mounted on exterior wall (later covered by grill that provides combustion to an unused fireplace). The ground braid goes to a ground rod located beneath the panel. (Photo courtesy: *Simple and Fun Antennas for Hams*)

Before a lightning storm approaches, a prudent ham will disconnect all feed lines, rotor lines and control lines inside the shack to prevent damage to sensitive electronics. When lightning is crashing about outside, you certainly don't want that lightning inside your shack!

Chapter 6

Low-Frequency Antennas

In theory there is no difference between antennas at 10 MHz and up and those for lower frequencies. In reality however, there are often important differences. It is the size of the antennas, which increases as frequency is decreased, that creates practical limits on what can be realized physically at reasonable cost.

At 7.3 MHz, 1λ = 133 feet and by the time we get to 1.8 MHz, 1 λ = 547 feet. Even a λ/2 dipole is very long on 160 meters. The result is that the average antenna for these bands is quite different from the higher bands, where Yagis and other relatively complex antennas dominate. In addition, vertical antennas can be more useful at low frequencies than they are on 20 meters and above because of the low heights (in wavelengths) usually available for horizontal antennas on the low bands. Much of the effort on the low bands is focused on how to build simple but effective antennas with limited resources. This section is devoted to antennas for use on amateur bands between 1.8 to 7 MHz.

The Importance of Low Angles for Low-Band DXing

In Chapter 3, The Effects of Ground, we emphasized the importance of matching the elevation response of your antennas as closely as possible to the range of elevation angles needed for communication with desired geographic areas. **Fig 1** shows the statistical 40-meter elevation angles needed over the entire 11-year solar cycle to cover the path from Boston, Massachusetts, to all of Europe. These angles range from 1° (at 9.6% of the time when the 40-meter band is open to Europe) to 28° (at 0.3% of the time).

Fig 1 also overlays the elevation pattern response of a 100-foot high flattop dipole on the elevation-angle statistics, illustrating that even at this height the coverage is hardly optimum to cover all the necessary elevation angles. While Fig 1 is dramatic in its own right, the data can be viewed in another way that emphasizes even more the importance of low elevation angles. **Fig 2** plots the the *cumulative distribution function*, the total percentage of time 40 meters is open from Boston to Europe, at or below each elevation angle. For example, Fig 2 says that 40 meters is open to Europe from Boston 50% of the time at an elevation angle of 9° or less. The band is open 90%

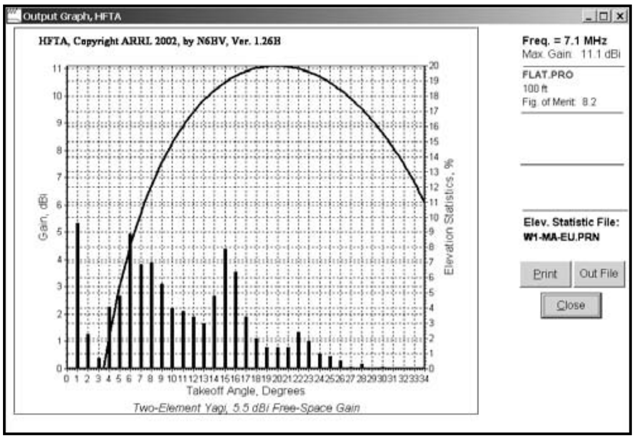

Freq. = 7.1 MHz
Max. Gain: 11.1 dBi
FLAT.PRO
100 ft
Fig. of Merit: 8.2
Elev. Statistic File:
W1-MA-EU.PRN

Fig 1—Screen capture from *HFTA* (HF Terrain Assessment) program showing elevation response for 100-foot high dipole over flat ground on 7.1 MHz, with bar-graph overlay of the statistical elevation angles needed over the whole 11-year solar cycle from New England (Boston) to all of Europe. Even a 100-foot high antenna cannot cover all the necessary angles.

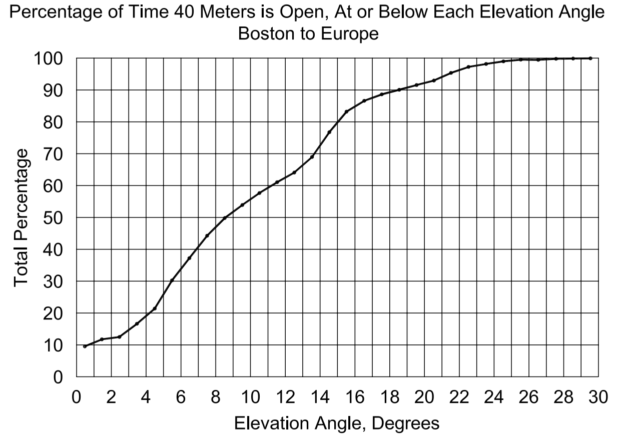

Fig 2—Another way of looking at the elevation statistics from Fig 1. This shows the percentage of time the 40-meter band is open, at or below each elevation angle, on the path from Boston to Europe. For example, the band is open 50% of the time at an angle of 9° or lower. It is open 90% of the time at an angle of 19° or lower.

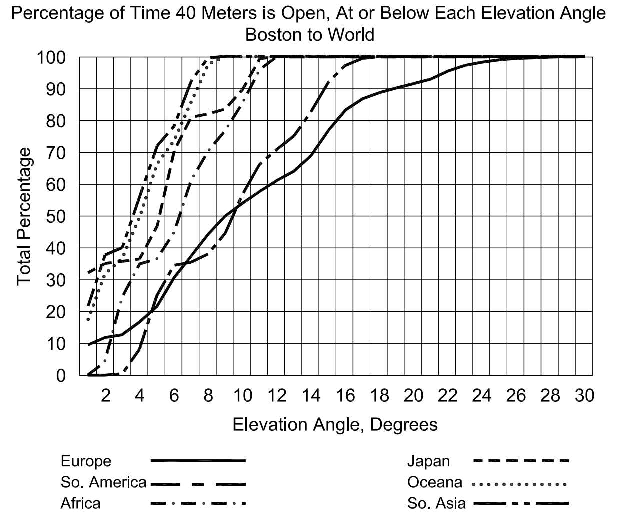

Fig 3—The percentage of time the 40-meter band is open, at or below each elevation angle, for various DX paths from Boston: to Europe, South America, southern Africa, Japan, Oceania and south Asias. The angles are predominantly quite low. For example, on the path from Boston to Japan, 90% of the time when the 40-meter band is open, it is open at elevation angles less than or equal to 10°. Achieving good performance at these low takeoff angles requires very high horizontally polarized antennas, or efficient vertically polarized antennas.

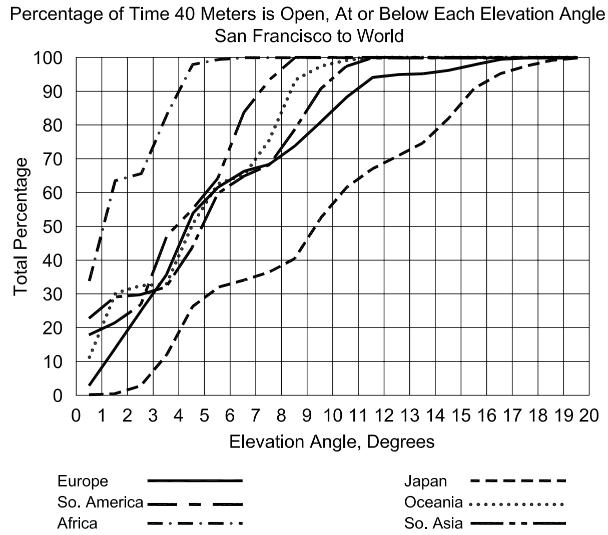

Fig 4—The 40-meter statistics from the West Coast: from San Francisco to the rest of the DX world. Here, 90% of the time the path to Europe is open, it is at takeoff angles less than or equal to 11°. No wonder the hams living on mountain tops do best into Europe from the West Coast.

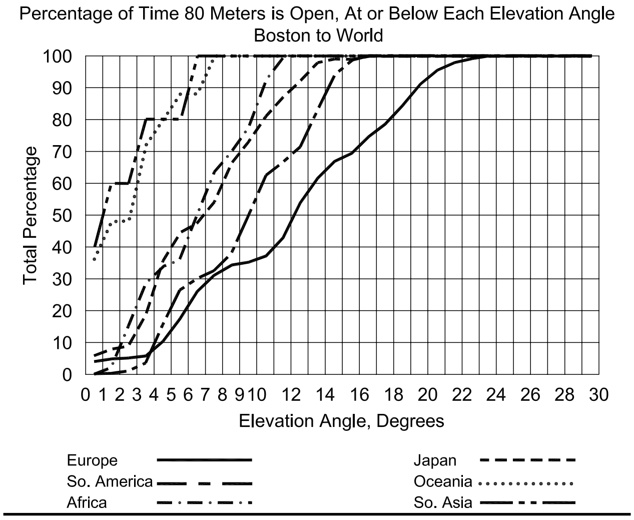

Fig 5—The situation on 80 meters from Boston to the rest of the DX world. Into Europe, 90% of the time the elevation angle is less than or equal to 20°. Into Japan from Boston, 90% of the time the angle is less than or equal to 12°.

of the time at an elevation angle of 19° or less.

Fig 3 plots the 40-meter elevation-angle data for six major geographic areas around the world from Boston. In general, the overall range of elevation angles for far-distant locations is smaller, and the angles are lower than for closer-in areas. For example, from Boston to southern Asia (India), 50% of the time the takeoff angles are 4° or less. On the path to Japan from Boston, the takeoff angles is less than or equal to 6° about 70% of the time. These are low angles indeed.

Fig 4 shows similar data for the 40-meter band from San Francisco, California, to the rest of the world. The path to southern Africa from the US West Coast is a very long-distance path, open some 65% of the time it is open at angles of 2° or less! The 40-meter path to Japan involves takeoff angles of 10° or less more than 50% of the time. If you are

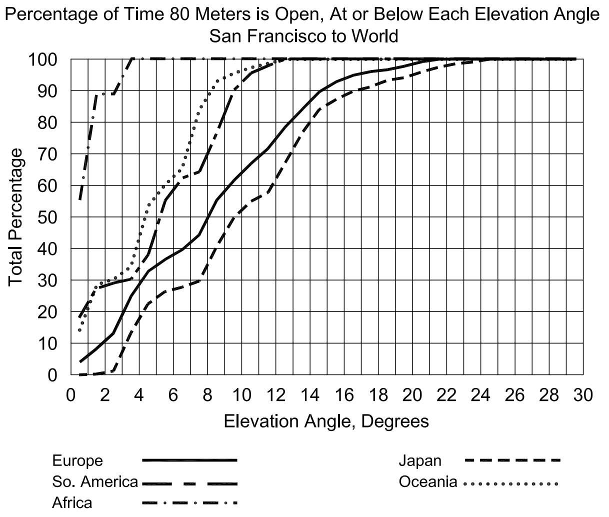

Fig 6—From San Francisco to the rest of the world on 80 meters: 90% of the time on the path to Japan, the takeoff angle is less than or equal to 17°; 50% of the time the angle is less than or equal to 10°; 25% of the time the angle is less than or equal to 6°. A horizontally polarized antenna would have to be 600 feet above flat ground to be optimum at 6°!

fortunate enough to have a 100-foot high flattop dipole for 40 meters, at a takeoff angle of 10° the response would be down about 3 dB from its peak level at 20°. At an elevation angle of 5° the response would be about 8 dB down from peak. You can see why the California stations located on mountain tops do best on 40 meters for DXing.

Fig 5 shows the same percentage-of-time data for the 80-meter band from Boston to the world. Into Europe from Boston, the 80-meter elevation angle is 13° or less more than 50% of the time. Into Japan from Boston, 90% of the time the band is open is at a takeoff angle of 13° or less. (Note that these elevation statistics are computed for "undisturbed" ionospheric conditions. There are times when the incoming angles are affected by geomagnetic storms, and generally speaking the elevation angles rise under these conditions.)

Fig 6 shows the 80-meter data from San Francisco to the world. Low elevation angles dominate in this graph and high horizontal antennas would be necessary to optimal coverage. In fact, 50% of the time for all paths, the elevation angle is less than 10°.

In the rest of this chapter, we'll often compare horizontally polarized antennas at practical heights with vertically polarized antennas, usually at takeoff angles of 5° or 10°, angles useful for DX work. But first, let us look at situations where *high* takeoff angles are most useful.

Short/Medium-Range Communications

Not all hams are interested in working stations thousands of miles from them. Traffic handlers and rag chewers may, in fact, only be interested in *nearby* communications—perhaps out to 600 miles from their location.

For example, a ham in Boston may want to talk with his brother-in-law in Cleveland, OH, a path that is just over 550 miles away. Or an operator in Buffalo, NY, may be the net control station (NCS) for a regional net involving the states of New York and New Jersey. She needs to cover distances up to about 300 miles away.

Depending on the time of day, the most appropriate ham frequencies needed for nearby communications are the 40 and 80/75-meter bands, with 160 meters also a possibility during the night hours, particularly during low portions of the sunspot cycle. The elevation angles involved in such nearby distances are usually high, even almost directly overhead for distances beyond ground-wave coverage (which may be as short as a few miles on 40 meters). For example, the distance between the Massachusetts cities of Boston and Worchester is about 40 miles. On 40 meters, 40 miles is beyond ground-wave coverage. So you will need sky-wave signals that use the ionosphere to communicate between these two cities, where the elevation angle is 83°—very nearly straight up.

Hams using vertical antennas for communications with nearby stations may well find that their signals will be below the noise level typical on the lower bands, especially if they aren't running maximum legal power. Such relatively short-range paths involve so-called *NVIS*, "Near Vertical Incidence Skywave," a fancy name for HF communication systems covering nearby geographic areas. The US military discusses NVIS out to about 500 miles, encompassing the territory a brigade might cover. Elevation angles needed to cover distances from 0 to 500 miles range from about 40º to 90º. This also covers the circumstances involved in amateur communications, particularly in emergency situations.

The following section is adopted from the article "What's the Deal About NVIS?" that appeared in December 2005 *QST*. This article used an example of a hypothetical earthquake in San Francisco to analyze HF emergency communication requirements.

HAM RADIO RESPONSE IN NATURAL DISASTERS

One of San Francisco's somewhat less endearing nicknames is "the city that waits to die." When the *Big Earthquake* does come, you can be assured that all the cell phones and the land-line telephones will be totally jammed, making calling in or out of the San Francisco Bay Area virtually impossible. The same thing occurred in Manhattan on September 11, 2001. The Internet will also be severely affected throughout northern California because of its trunking via the facilities of the telephone network. Commercial electricity will be out in wide areas because power lines will be down. It's virtually certain that water mains will be out of commission too.

If the repeaters on the hills around the San Francisco Bay Area haven't been damaged by the shaking itself, there will be some ham VHF/UHF voice coverage in the intermediate area, at least until the backup batteries run down. But connecting to the dysfunctional telephone system will be difficult at best through amateur repeaters.

With little or no telephone coverage, an obvious need for ham radio communications to aid disaster relief would be from San Francisco to Sacramento, the state capital. Sacramento is 75 miles northeast of the Bay Area, well outside VHF/UHF coverage, so amateur HF will be required on this radio circuit. On-the-ground communications directly between emergency personnel (including the armed-forces personnel who will be brought into the rescue and rebuilding effort) will often be difficult on VHF/UHF since San Francisco is a hilly place. So HF will probably be needed even for short distance, operator-to-operator or operator-to-communications center work. Throughout the city, portable HF stations will have to be quickly set up and staffed to provide such communications.

Hams used to half jokingly call short range HF communications on 40 and 80 meters "cloud warming." This is an apt description, because the takeoff angles needed to launch HF signals up into the ionosphere and then down again to a nearby station are almost directly upwards. **Table 1** lists the distance and takeoff angles from San Francisco to various cities around the western part of the USA. The distance between San Francisco and Sacramento is about 75 miles, and the optimum takeoff angle is about 78°. Launching such a high-angle signal is best done using horizontally polarized antennas mounted relatively close to the ground.

GEOGRAPHIC COVERAGE FOR NVIS

Figure 7A shows the geographic area coverage around San Francisco for a 100-W, 7.2-MHz station using an inverted V dipole. The center of this antenna is 20 feet above flat

Table 1
Average Elevation Angles for Target Destinations from San Francisco

Location	*Distance Miles*	*Average Elevation Angle, Degrees*
San Jose, CA	43	80
Sacramento, CA	75	78
Fresno, CA	160	63
Reno, NV	185	60
Los Angeles	350	44
San Diego	450	42
Portland, OR	530	30
Denver, CO	950	18
Dallas, TX	1500	8

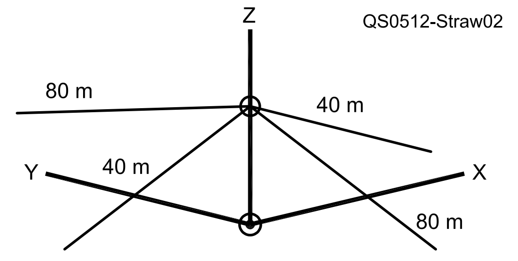

Fig 8—Layout for two band inverted V dipoles for 40 and 80 meters. The two dipoles are fed together at the center and are laid out at right angles to each other to minimize interaction between them. Each end of both dipoles is kept 8 feet above ground for personnel safety.

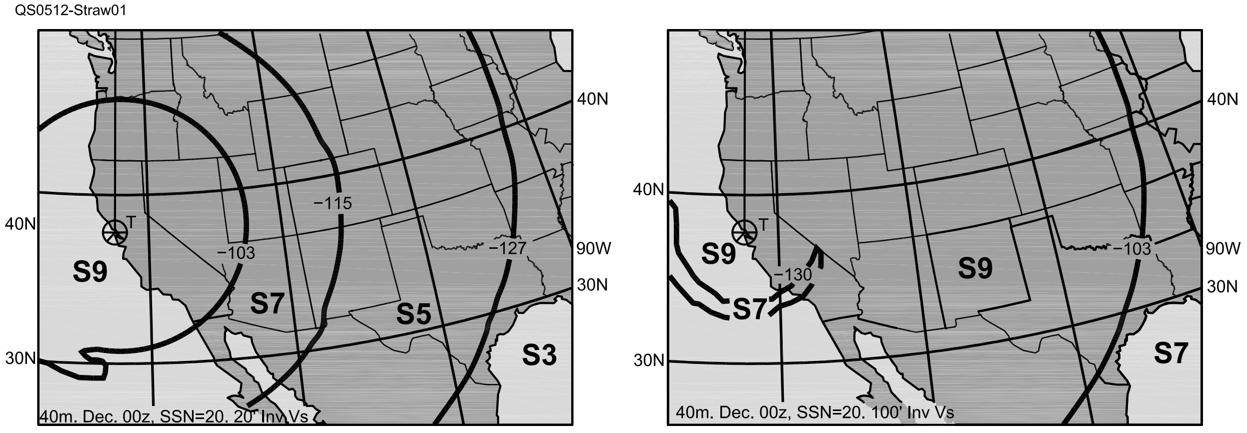

(A)

(B)

Fig 7—At A, Predicted 40 meter geographic coverage plot for a 100 W transmitter in December at 0000 UTC (near sunset), for a SSN (Smoothed Sunspot Number) of 20. The antennas used are 20 foot-high inverted V dipoles. At B, 40 meter coverage for same date and time, but for 100 foot-high flattop dipoles. Most of California is well covered with S9 signals in both cases, but there is more susceptibility in the higher dipole case to thunderstorm crashes coming from outside California, for example from Arizona or even Texas. Such noise can interfere with communications inside California.

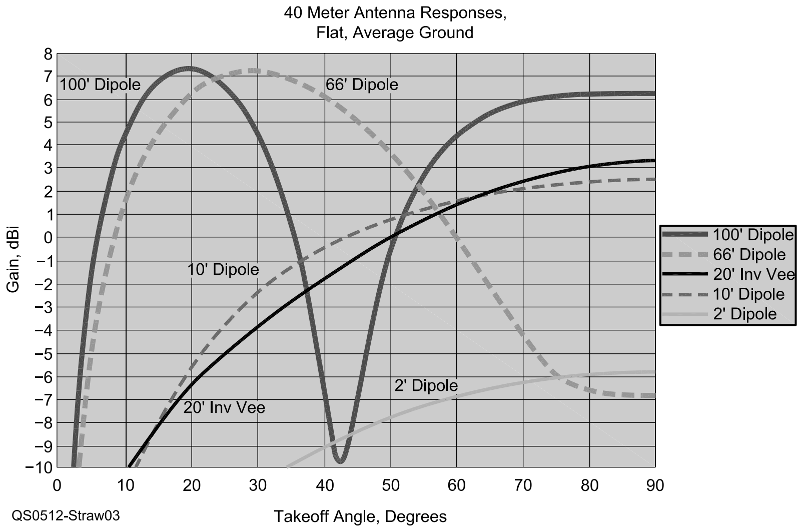

Fig 9—Elevation plots for different 40 meter antennas above flat ground with average ground characteristics (5 mS/m conductivity and dielectric constant of 13). The 10 foot-high flattop dipole and the 20 foot-high inverted V dipoles both have close to the same characteristics. Note that there is a null in the response of the 100 foot-high flattop dipole at a 42° elevation angle. The gain there is roughly that of a 2 foot-high dipole!

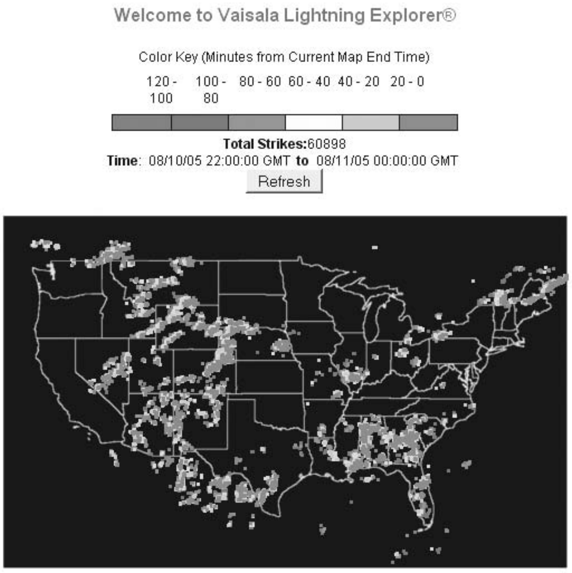
Welcome to Vaisala Lightning Explorer®

Color Key (Minutes from Current Map End Time)

120 - 100 | 100 - 80 | 80 - 60 | 60 - 40 | 40 - 20 | 20 - 0

Total Strikes:60898

Time: 08/10/05 22:00:00 GMT to 08/11/05 00:00:00 GMT

Fig 10—The distribution of lightning strikes across the USA for August 10, 2005 from 2200 to 0000 UTC, in the afternoon California time. There are lots of lightning strikes in the US during the summer—60,898 of them in this two-hour period! (*Courtesy Vaisala Lightning Explorer.*)

ground and the ends are 8 feet high. An actual implementation of such an antenna could be as an 80-meter inverted V, fed in parallel with a 40-meter inverted V dipole at a 90° angle. See **Fig 8**. The 8-foot height puts the ends high enough to prevent RF burns to humans (or most animals). The low height of the antenna above ground means that the azimuthal pattern is omnidirectional for high elevation angles.

Fig 7 was generated using the *VOAAREA* program, part of the *VOACAP* propagation-prediction suite, for the month of December. This was for 0000 UTC, close to sundown, for a low period of solar activity (Smoothed Sunspot Number, SSN of 20). The receiving stations were also assumed to be using identical inverted-V dipoles.

You can see that almost the whole state of California is covered with S9 signals, minus only a thin slice of land near the Mexican border in the southeast portion of the state, where the signal drops to S7. Signals from Texas are predicted to be only S5 or less in strength. Signals (or thunderstorm static) coming from, say, Louisiana would be several S units weaker than signals from central Texas.

Now take a look at Fig 7B. Here, the date, time and solar conditions remain the same, but now the antennas are 100-foot high flattop dipoles. California is still blanketed with S9 signals, save for an interesting crescent-shaped slice near Los Angeles, where the signal drops down to S7. Close investigation of this intriguing drop in signal strength reveals that the necessary elevation angle, 44°, from San Francisco to this part of southern California falls in the first null of the 100-foot high antenna's elevation pattern. See **Fig 9**, which shows the elevation patterns for five 40-meter antennas at different heights. In the null at a 44° takeoff angle, the 100-foot high dipole is just about equal to a 2-foot high dipole. We'll discuss 2-foot high dipoles in more detail later.

For most of California, the problem with 100-foot high 40-meter antennas is that interfering signals from Texas, Colorado or Washington State will *also* be S9 in San Francisco. So will static crashes coming from thunderstorms all over the West and much of the Gulf Coast. (Ed Farmer, AA6ZM, joked once that the Army doesn't have any problem with interfering signals—they just call in an airstrike. We hams don't generally have this ability, although we occasionally call in the FCC.) See **Fig 10**, which shows a typical distribution of thunderstorms across the US in the late afternoon, California time, in mid-August. There certainly are a lot of thunderstorms raging around the country in the summer.

The signal-to-noise and signal-to-interference ratios for a 20-foot high inverted V dipole will be superior for medium-

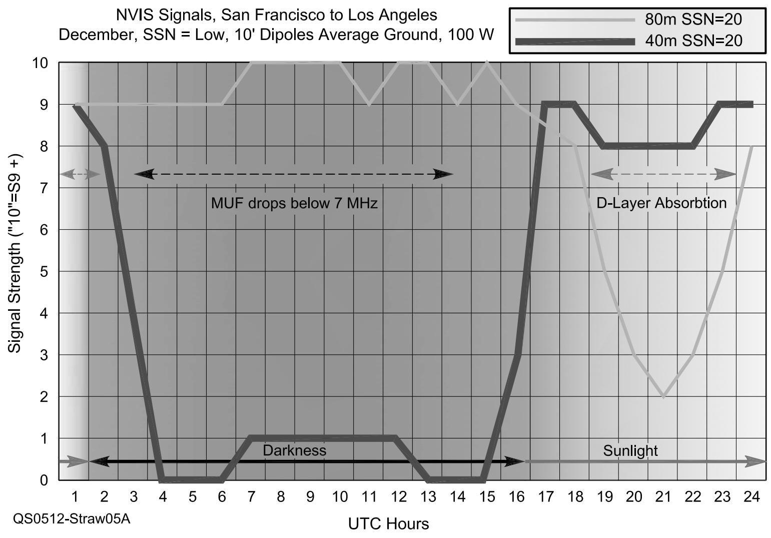

Fig 11—VOACAP calculations for a 350 mile path from San Francisco to Los Angeles, using 10 foot-high flattop dipoles. This plot shows the signal strength in S Units ("S10" = S9+10) for a worst-case month/SSN combination—winter solstice, in December, for a low level of solar activity (SSN = 20). The 40-meter signal drops to a very low level during the night because the MUF drops well below 7.2 MHz. The 80-meter signal drops in the afternoon because of D-layer absorption. For 24-hour communications on this path, the rule of thumb is to select 40 meters during the day and 80 meters during the night.

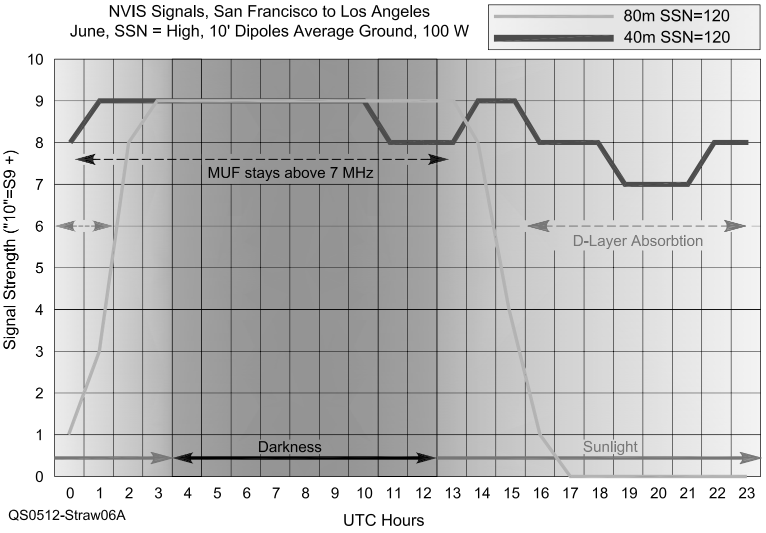

Fig 12—Signal strengths for the San Francisco to Los Angeles path for a worst-case month/SSN combination—summer solstice, in June, for a high level of solar activity (SSN = 120). Now 80 meters drops out more dramatically during the daylight hours, due to increased D-layer absorption. At this high level of solar activity, 40 meters remains open 24 hours with reasonable signal levels. However, the NVIS rule-of-thumb still holds: Use 40 meters during the day; 80 meters at night.

range distances, say out to 500 miles from the center, compared to a 100-foot high antenna. The 20-foot high antenna can discriminate against medium-angle thunderstorm noise in the late afternoon coming from the Arizona desert, although it wouldn't help much for thunderstorms in the Sierra Nevada in central Nevada, which are arriving in San Francisco at high angles, along with the desired NVIS signals.

This is the essence of what NVIS means. NVIS exploits the difference in elevation pattern responses of low horizontally polarized antennas compared to higher horizontal antennas, or even verticals. Over the years, many hams have been lead to believe that higher is always better. This is not quite so true for consistent coverage of medium or short distance signals!

If NVIS only involved putting up a low horizontally polarized antenna on 40 meters the story would end here. However, real cloud warming is more complicated. It also involves the intelligent choice of more than just one operating frequency to achieve reliable all day, all-night communications coverage.

Fig 11 shows the signal strength predicted using *VOACAP* for the 350-mile path from San Francisco to Los Angeles for the month of December for a period of low solar activity (SSN of 20). The antennas used in this case are 10-foot high dipoles, just for some variety. These act almost like 20-foot high Inverted V dipoles. December at a low SSN was chosen as a worst-case scenario because the *winter solstice* occurs on December 21. This is the day that has the fewest hours of daylight in the year. (Contrast this with the *summer solstice*, on June 21, which has the most hours of daylight in the year.) Note that the upper signal limit in Fig 11 is "S10"—a fictitious quantity that allows easier graphing.

S10 is equivalent to S9+, or at least S9+10 dB.

The 40-meter curve in Fig 11 shows that the MUF (maximum usable frequency) actually drops below the 7.2 MHz amateur band after sunset. The signal becomes quite weak for about 14 hours during the night, from about 0300 to 1700 UTC. In a period of low solar activity the 40-meter band thus becomes strictly a *daytime band* on this medium-distance path.

The 80-meter curve in Fig 11 shows strong signals after dusk, through the night and up until about an hour after sunrise. After sunrise, 80 meters starts to suffer absorption in the D layer of the ionosphere and hence the signal strength drops. Here, 80 meters is a true *nighttime band*.

Let's see what happens from San Francisco to Los Angeles during a period of high solar activity (SSN of 120) during the summer solstice in June. **Fig 12** shows that 40 meters now stays open all hours of the day due to the greater number of hours of sunlight in June and because the ionosphere becomes more highly ionized by higher solar activity. Meanwhile, 80 meters still remains a nighttime band during these conditions on this path.

Now, let's look at a shorter-distance path—our 75-mile emergency communications path from San Francisco to Sacramento. We'll again use June during the summer solstice, at a high level of solar activity (SSN of 120) because this represents another worst-case scenario. **Fig 13** shows that 40 meters remains open on this path all day, dropping to a lower signal level just before sunrise. At sunrise, the MUF drops close to 7.2 MHz. 80 meters is still mainly a nighttime band to Sacramento, even though it does yield workable signal levels even during the daylight hours. However, 40 meters is better from 1200 to 0400 UTC, so 40 would be still the right daytime band for this path during the day.

CHOOSING THE RIGHT NVIS FREQUENCY

You can see that a pattern is developing here for efficient NVIS short/medium-distance communications out to 500 miles:

- You should pick a frequency on 40 meters during the day.
- You should pick a frequency on 80 meters during the night.
- You should choose an antenna that emphasizes moderate to high elevation angles, from 40° to almost directly overhead at 90°.

"What about 60 meters?" you might ask. The characteristics on 60 meters fall in-between 40 and 80 meters, although it resembles 40 meters more closely. With characteristics close to that of 40, but with only five channels available and a 50-W power limit, the 60-meter band is of low utility for serious NVIS use.

What about 160 meters? For 100-W level radios, even at the worst-case month or during low solar activity, the critical frequency doesn't fall below 3.8 MHz often enough to destroy the ability to communicate, even for short distances. That is a relief, considering that installing a 160-meter half-wave dipole involves a 255-foot wingspan, and it would need to be elevated at least 30 feet in the center. A short loaded vertical such as a 160-meter mobile whip would have poor response at the high elevation angles needed for NVIS. You could probably put a monster 160-meter horizontal dipole up at a permanent location, but hauling such a thing around in the field would not be an easy task.

SOME OTHER OBSERVATIONS ABOUT NVIS—STRATEGY

You could pose the question about whether NVIS is an operating *mode* or whether it is actually an operating *strategy*. We maintain that NVIS is a strategy. It involves choosing both appropriate frequencies and then appropriate antennas for those frequencies. Fig 13 does show that on short-distance

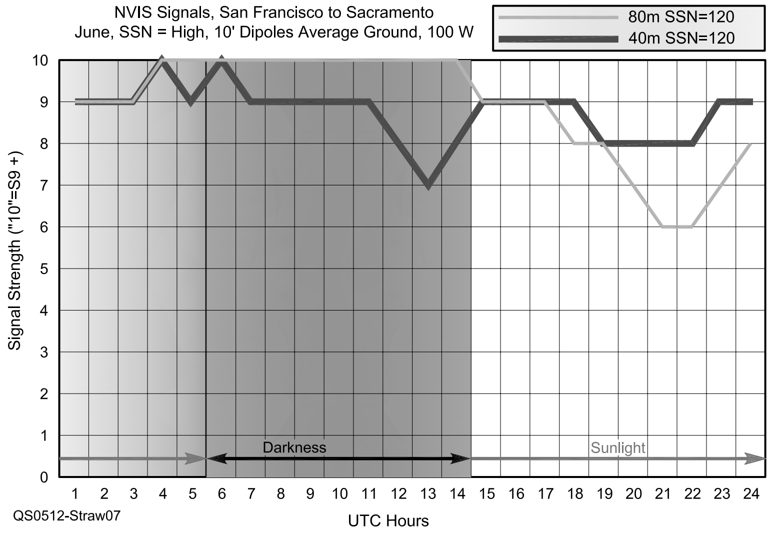

Fig 13—Signal strengths for a 75 mile path—San Francisco to Sacramento. This is for June and SSN = 120. Either band could be used successfully over the full 24-hour period because signal levels are always higher than S6. But the simple NVIS rule-of-thumb still holds: Use 40 meters during the day; 80 meters at night. This simplifies giving instructions to operators unaccustomed to the use of HF.

paths, such as between San Francisco and Sacramento, you could stay on 80 meters all day and night. But if you have to give a single rule-of-thumb to operators who are not very experienced at operating HF, we would tell them to operate on the higher frequency band during the day and on the lower frequency band at night.

SOME OTHER OBSERVATIONS ABOUT NVIS—ANTENNA HEIGHT

Some NVIS aficionados have advocated placing dipoles only a few feet over ground, something akin to saying, "If low is good for NVIS, then lower must be even better." Now we are not claiming that a very low antenna *won't* work in specific instances—for example, covering a small state such as Rhode Island or even just the San Francisco Bay Area.

It certainly is convenient to mount a 40-meter dipole on some 2-foot high red traffic cones! You should be very skeptical, however, about the ability of such antennas to cover all of a large state, such as California or Texas, especially on 80 meters. **Fig 14** shows the computed elevation responses for a number of 80-meter antennas, including a 2-foot-high dipole.

Fig 15B shows the 80 meter geographic coverage plot for 2-foot-high flattop dipoles, compared with the plot in Fig 15A for 20-foot-high inverted V dipoles on both ends of the path. The 2-foot-high dipoles produce about two S-units less signal across all of California than the 20-foot-high inverted V dipoles, at 0300 UTC in December, with an SSN of 20. The reason is that a low dipole will suffer more losses in the ground under it.

The differential between California signals and possible interfering signals from, say, New Mexico, is predicted to be four S-units, the same as it is for the higher inverted V dipole

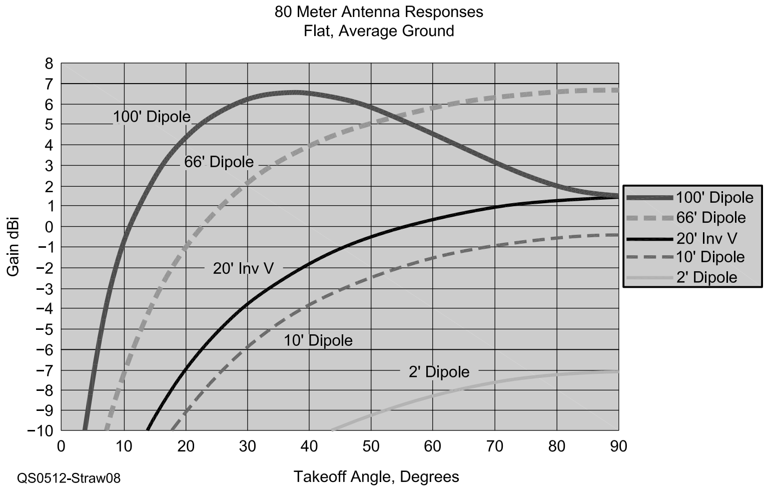

Fig 14—Elevation response patterns for 80 meter antennas over average soil. The shapes track each other rather well, remaining parallel for heights from 2 to 66 feet over flat ground. The 2 foot dipole is substantially down, about 9 dB, from the 20 foot inverted V dipole at all angles.

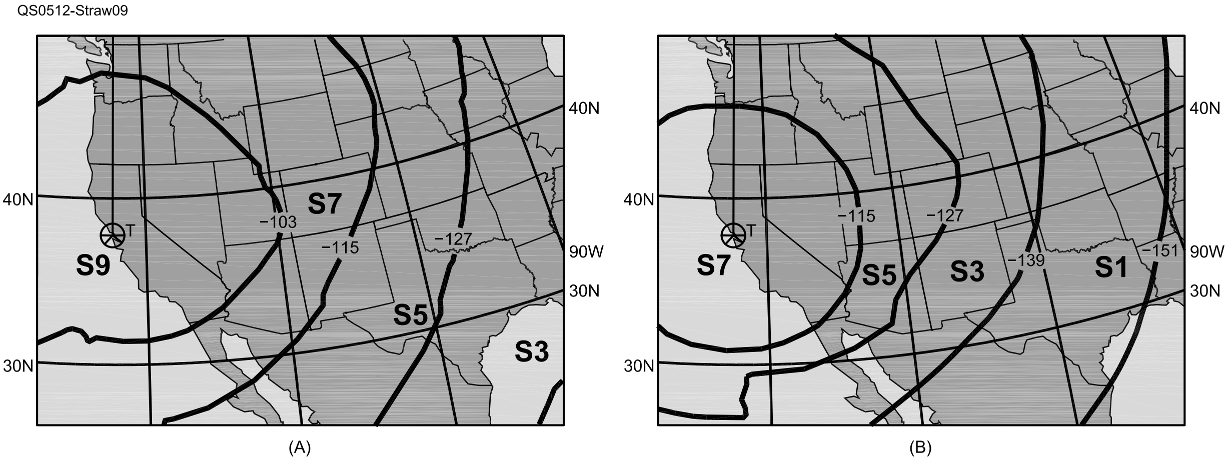

Fig 15—Geographic coverage plots for December, SSN = 20, 0300 UTC. At A, antennas are 20 foot-high inverted V dipoles over Average soil. At B, antennas are 2-foot-high flattop dipoles over Average soil. The response for the 2-foot-high antennas is down about 2 S Units, 8 to 12 dB for a typical communications receiver.

at 20 feet. Thus there is no real advantage in terms of signal-to-interference ratio or signal-to-noise ratio (for thunderstorm static crashes) for either height. This is because the shape of all the response curves in Fig 14 below 20 feet essentially track each other in parallel.

However, the lower the antenna, the lower the transmitted signal strength. Physics remain physics. And if you are in an emergency situation operating on batteries, you could reduce power from 100 W to 10 W with a 20-foot high inverted-V antenna and still maintain the same signal strength as a 2-foot high dipole at 100W.

LOW NVIS ANTENNAS AND LOCAL POWERLINE NOISE

Some advocates of really low antennas have stated that the received noise is much lower than that received from higher antennas, and this therefore leads to better signal-to-noise ratios (SNR). How much this is true depends on the source of the noise. If the noise comes from distant thunderstorms, then the SNR advantage going to a 2-foot antenna from a 20-foot-high one is insignificant, as Fig 15 indicates.

If noise is from an arcing insulator on a HV power line half a mile away, that noise will arrive at the antenna as a ground-wave signal. We calculate that the 2-foot antenna receives 4.4 dB less noise by groundwave than a 20-foot-high inverted V dipole. However, at an incoming elevation angle of 45º—suitable for a signal going from Los Angeles to San Francisco—the signal would be down 7.1 dB on the low dipole compared to the higher antenna. The net loss in SNR for the 2-foot-high dipole is thus 7.1– 4.4 or 2.7 dB. Close, but no cigar. Summarizing about really low NVIS antennas:

- A 2-foot-high dipole yields weaker signals, but without an SNR advantage compared to its more elevated brethren.
- A 2-foot-high dipole is a lot easier to trip over at night. We would call this a “knee biter” (or maybe an “ankle biter” if you’re really tall).
- You (and your dog) can easily get RF burns from an antenna that is only 2 feet off the ground.

This is not a winning strategy to make friends or QSOs, it seems. But still, a really low dipole may serve your short-range communication needs just fine. But remember, that just as “higher is better” isn’t universally true for NVIS (or even longer range) applications, “lower is better” isn’t a panacea either.

ELEVATION ANGLES FOR MODERATE DISTANCES ON 75/80 METERS

Fig 16 shows the elevation angles statistics for a 75-meter, 550-mile path from Boston to Cleveland, together with overlays of the elevation patterns for several different types of antennas. These elevation statistics cover all parts of the 11-year solar cycle for this path. The responses for the popular G5RV antenna (described later in this chapter) are shown for two different heights above flat ground: 50 and 100 feet. An 80-meter half-wave sloper (“full sloper”) and an 80-meter ground-plane antenna are also shown. All antenna patterns are for “average ground” constants of 5 mS/m conductivity and a dielectric constant of 13.

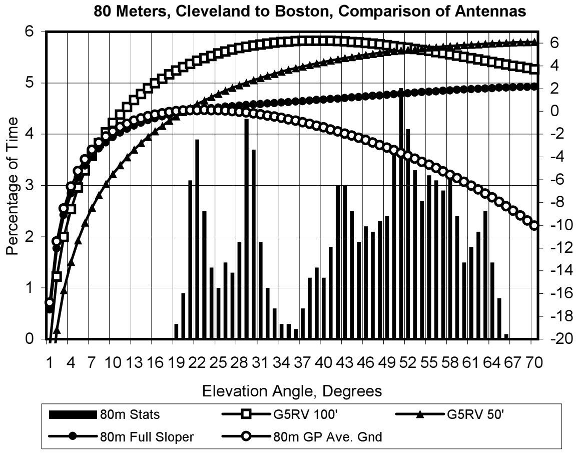

Fig 16—80/75-meter elevation statistics for all portions of the 11-year solar cycle for the path from Cleveland, Ohio, to Boston, Massachusetts, together with the elevation responses for four different multiband antennas. The 100-foot high horizontally polarized G5RV performs well over the entire range of necessary takeoff elevation angles.

At the statistically most significant takeoff angles around 50°, the two horizontally polarized G5RV antennas are about equal. At the second-highest elevation peak near 30°, the 100-foot G5RV has about a 4-dB advantage over its lower counterpart. The full sloper has comparable performance to the 100-foot high G5RV from 1° to about 20° and then gradually rises to its peak at angles higher than 70°. The full sloper is superior to the 50-foot horizontal G5RV at low takeoff elevation angles. The 80-meter ground plane has a deep null directly overhead. At an elevation angle of 70° it is down some 16 dB compared to the 50-foot high horizontal G5RV.

The advantage of antennas suitable for high-angle radiation was vividly demonstrated during a 75-meter QSO one fall evening between N6BV/1 in southern New Hampshire and W1WEF in central Connecticut. This involved a distance of about 100 miles and W1WEF was using his Four Square vertical array. Although W1WEF’s signal was S9 on the Four Square, N6BV/1 suggested an experiment. Instead of connecting the so-called “dump power” connector on his Comtek ACB-4 hybrid phasing coupler to a 50-Ω dummy load (the normal configuration), W1WEF switched the dump power to his 100-foot high 80-meter horizontal dipole. W1WEF’s signal came up more than 20 dB! The approximately 100-W of power that would otherwise be “wasted” in the dummy load was converted to useful signal.

ELEVATION ANGLES FOR MODERATE DISTANCES ON 40 METERS

Fig 17 shows the situation for the 40-meter band, from Boston to Cleveland, together with the same antennas used for 80 meters in Fig 16. Note that the 100-foot high horizontally

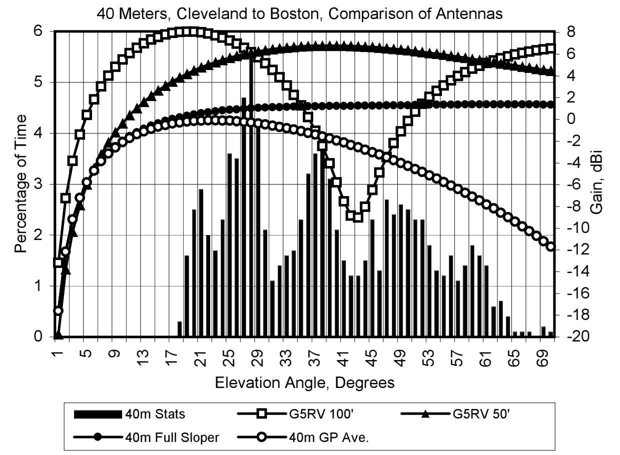

Fig 17—40-meter elevation statistics for the Cleveland to Boston path, together with elevation patterns for four antennas. Here, the 100-foot high horizontally polarized G5RV would have a null in the middle of the range of elevation angles needed for consistent performance on this path. For multiband use on this path to relatively nearby stations, the 50-foot high horizontal antenna would be a better choice than the 100-foot high antenna.

polarized G5RV has about a 16-dB null at an elevation angle of 43°. This doesn't affect things for low elevation angles, but it certainly has a profound effect on signals arriving between about 30° to 60°, especially when compared to the 50-foot high horizontal G5RV. The 40-meter full sloper beats out the high horizontal antenna from about 35° to 50°. And the ground plane is obviously not the antenna of choice for this moderate-range path from Boston to Cleveland, although it is still a good performer on longer-distance paths, with their low takeoff angles.

A 100-foot high multiband dipole is about ⅜-λ high on 75/80 meters. It is an excellent antenna for general-purpose local and DXing operation. But the same dipole used on 40 meters becomes ¾-λ high. At that height, the nulls in its elevation pattern give large holes in coverage for nearby 40-meter contacts. Many operators have found that a 40- to 50-foot high dipole on 40 meters gives them far superior performance for close-in QSOs, when compared to a high dipole, or even a high 2-element 40-meter Yagi.

NVIS SUMMARY

The use of NVIS strategies to cover close-in and intermediate distance communications within about 600 miles involves the intelligent choice of low HF frequencies. As a rule-of-thumb for ham band NVIS, 40 meters is recommended for use during the day; 80 meters during the night.

NVIS involves the choice of antennas suitable for this strategy. Horizontally polarized dual-band 80 and 40-meter flattop dipoles that are mounted higher than about 10 feet high will work adequately for portable operations. Dual-band 80 and 40-meter inverted V dipoles supported 20 feet above the ground at the center can also work well in portable operations.

Single-band 40-meter flattop antennas about 30 feet high and 80-meter flattop antennas about 60 feet high can do a good job for fixed locations.

Horizontal Antennas for the Low Bands

As shown in Chapter 3, The Effects of Ground, and here, radiation angles from horizontal antennas are a very strong function of the height above ground in wavelengths. Typically for DX work heights of λ/2 to 1 λ are considered to be a minimum. As we go down in frequency these heights become harder to realize. For example, a 160-meter dipole at 70 feet is only 0.14 λ high. This antenna will be very effective for local and short distance QSOs but not very good for DX work. Despite this limitation, horizontal antennas are very popular on the lower bands because the low frequencies are often used for short range communications, local nets and rag chewing. Also horizontal antennas do not require extensive ground systems to be efficient.

DIPOLE ANTENNAS

Half-wave dipoles and variations of these can be a very good choice for a low band antenna. A variety of possibilities are shown in **Fig 18**. An untuned or "flat" feed line is a logical choice on any band because the losses are low, but this generally limits the use of the antenna to one band. Where only single-band operation is wanted, the λ/2 antenna fed with open-wire line is one of the most popular systems on the 3.5 and 7-MHz bands.

If the antenna is a single-wire affair, its impedance is in the vicinity of 60 Ω, depending on the height and the ground characteristics. The most common way to feed the antenna is with 50- or 75-Ω coaxial line. Heavy coaxial lines present support problems because they are a concentrated weight at the center of the antenna, tending to pull the center of the antenna down. This can be overcome by using an auxiliary pole to take at least some of the weight of the line. The line should come away from the antenna at right angles, and it can be of any length.

Folded Dipoles

A folded dipole (Fig 18B and C) has an impedance of about 300 Ω, and can be fed directly with any length of 300-Ω line. The folded dipole can be made of ordinary wire spaced by lightweight wooden or plastic spacers, 4 or 6 inches long, or a piece of 300 or 450-Ω twin-lead or ladder line.

A folded dipole can be fed with a 600-Ω open wire line with only a 2:1 SWR, but a nearly perfect match can be ob-

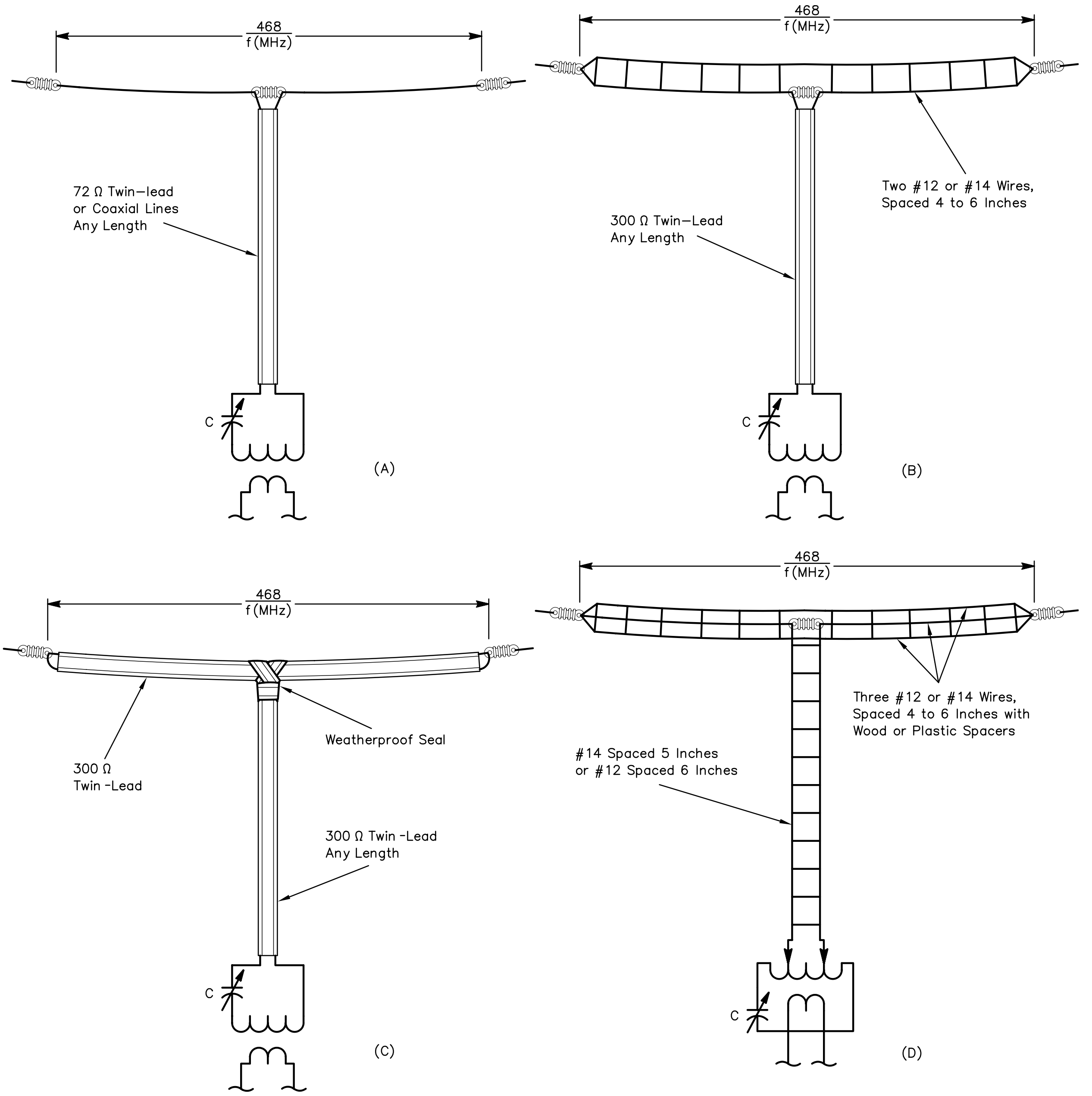
468
f (MHz)
72 Ω Twin—lead
or Coaxial Lines
Any Length
C
(A)
468
f (MHz)
Two #12 or #14 Wires,
Spaced 4 to 6 Inches
300 Ω Twin—Lead
Any Length
C
(B)
468
f (MHz)
300 Ω
Twin -Lead
Weatherproof Seal
300 Ω Twin -Lead
Any Length
C
(C)
468
f (MHz)
Three #12 or #14 Wires,
Spaced 4 to 6 Inches with
Wood or Plastic Spacers
#14 Spaced 5 Inches
or #12 Spaced 6 Inches
C
(D)

Fig 18—Half-wavelength antennas for single band operation. The multiwire types shown in B, C and D offer a better match to the feeder over a somewhat wider range of frequencies but otherwise the performances are identical. The feeder should run away from the antenna at a right angle for as great a distance as possible. In the coupling circuits shown, tuned circuits should resonate to the operating frequency. In the series-tuned circuits of A, B, and C, high L and low C are recommended, and in D the inductance and capacitance should be similar to the output-amplifier tank, with the feeders tapped across at least ½ the coil. The tapped-coil matching circuit shown in Chapter 25 can be substituted in each case.

tained with a three-wire dipole fed with either 450-Ω ladder line or 600-Ω open wire line. One advantage of the two- and three-wire antennas over the single wire is that they offer a better match over a wider band. This is particularly important if full coverage of the 3.5-MHz band is contemplated.

Inverted-V Dipole

The halves of a dipole may be sloped to form an inverted V, as shown in **Fig 19**. This has the advantages of requiring only a single high support and less horizontal space. There will be some difference in performance between a normal horizontal dipole and the inverted V as shown by the radiation patterns in **Fig 20**. There is small loss in peak gain and the pattern is less directional.

Sloping of the wires results in a raising of the resonant frequency and a decrease in feed-point impedance and

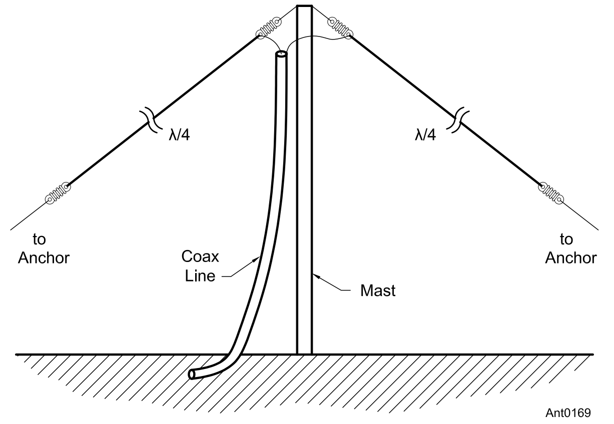

Fig 19—The inverted-V dipole. The length and apex angle should be adjusted as described in the text.

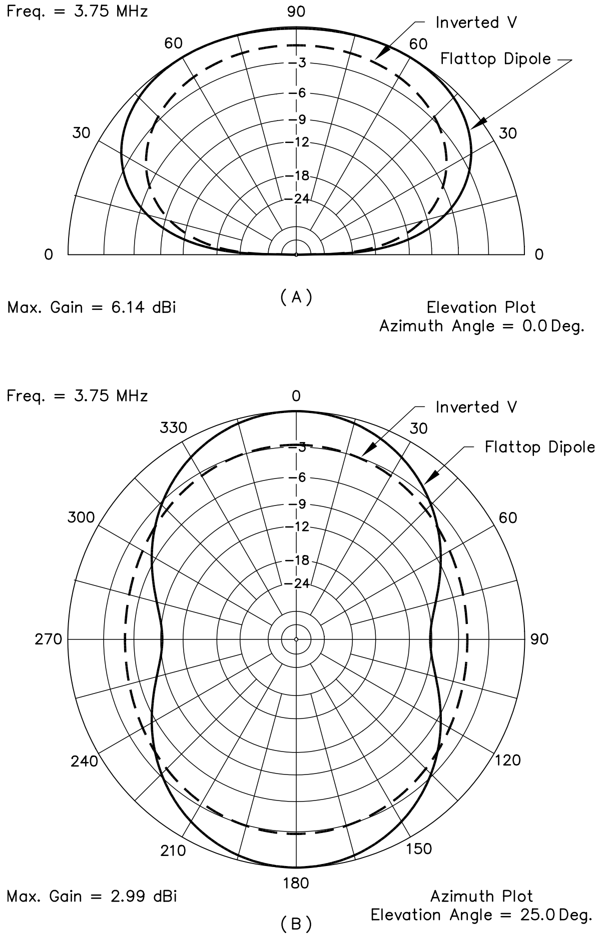

Fig 20—At A, elevation and at B, azimuthal radiation patterns comparing a normal 80-meter dipole and an inverted-V dipole. The center of both dipoles is at 65 feet and the ends of the inverted V are at 20 feet. The frequency is 3.750 MHz.

bandwidth. Thus, for the same frequency, the length of the dipole must be increased somewhat. The angle at the apex is not critical, although it should probably be made no smaller than 90°. Because of the lower impedance, a 50-Ω line should be used. For those who are dissatisfied with anything but a perfect match, the usual procedure is to adjust the angle for lowest SWR while keeping the dipole resonant by adjustment of length. Bandwidth may be increased by using multiconductor elements, such as a cage configuration.

PHASED HORIZONTAL ARRAYS

Phased arrays with horizontal elements, which provide some directional gain, can be used to advantage at 7 MHz, if they can be placed at least 40 feet above ground. At 3.5 MHz heights of 70 feet or more are needed for any real advantage.

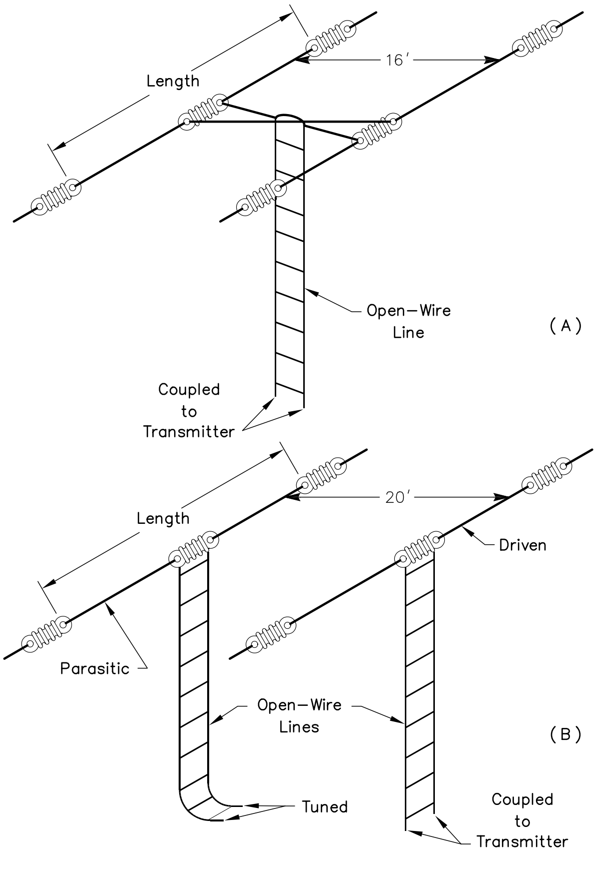

Fig 21—Directional antennas for 7 MHz. To realize any advantage from these antennas, they should be at least 40 feet high. At A, system is bidirectional. At B, system is unidirectional in a direction depending upon the tuning conditions of the parasitic element. The length of the elements in either antenna should be exactly the same, but any length from 60 to 150 feet can be used.
If the length of the antenna at A is between 60 and 80 feet, the antenna will be bidirectional along the same line on both 7 and 14 MHz. The system at B can be made to work on 7 and 14 MHz in the same way, by keeping the length between 60 and 80 feet.

Many of the driven arrays discussed in Chapter 8 and even some of the Yagis discussed in Chapter 11 can be used as fixed directional antennas. If a bidirectional characteristic is desired, the W8JK array, shown in **Fig 21A**, is a good one. If a unidirectional characteristic is required, two elements can be mounted about 20 feet apart and provision included for tuning one of the elements as either a director or reflector, as shown in Fig 21B.

The parasitic element is tuned at the end of its feed line with a series or parallel-tuned circuit (whichever would normally be required to couple power into the line), and the proper tuning condition can be found by using the system for receiving and listening to distant stations along the line to the rear of the antenna. Tuning the feeder to the parasitic element can minimize the received signals from the back of the antenna. This is in effect adjusting the antenna for maximum front-to-back ratio. Maximum front-to-back does not occur at the same point as maximum forward gain but the loss in forward gain is very small. Adjusting the antenna for maximum forward gain (peaking received signals in the forward direction) may increase the forward gain slightly but will almost certainly result in relatively poor front-to-back ratio.

A MODIFIED EXTENDED DOUBLE ZEPP

If the distance between the available supports is greater than λ/2 then a very simple form of a single wire collinear array can be used to achieve significant gain. The *extended double Zepp* antenna has long been used by amateurs and is discussed in Chapter 8, Multielement Arrays. A simple variation of this antenna with substantially improved bandwidth can be very useful on 3.5 and 7.0 MHz. The following material has been taken from an article by Rudy Severns, N6LF, in *The ARRL Antenna Compendium Vol 4*.

The key to improving the characteristics of a standard double-extended Zepp is to modify the current distribution. One of the simplest ways to do this is to insert a reactance(s) in series with the wire. This could either be an inductor(s) or a capacitor(s). In general, a series capacitor will have a higher Q and therefore less loss. With either choice it is desirable to use as few components as possible.

As an initial trial at 7 MHz, only two capacitors, one on each side of the antenna, were used. The value and position of the capacitors was varied to see what would happen. It quickly became clear that the reactance at the feed point could be tuned out by adjusting the capacitor value, making the antenna look essentially like a resistor over the entire band. The value of the feed-point resistance could be varied from less than 150 Ω to over 1500 Ω by changing the location of the capacitors and adjusting their values to resonate the antenna.

A number of interesting combinations were created. The one ultimately selected is shown in **Fig 22**. The antenna is 170 feet in length. Two 9.1 pF capacitors are located 25 feet out each side of the center. The antenna is fed with 450-Ω transmission line and a 9:1 three-core Guanella balun used at the transmitter to convert to 50 Ω. The transmission line can be any convenient length and it operates with a very low SWR.

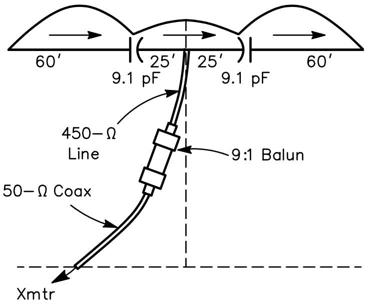

Fig 22—Schematic for modified N6LF Double Extended Zepp. Overall length is 170 feet, with 9.1 pF capacitors placed 25 feet each side of center.

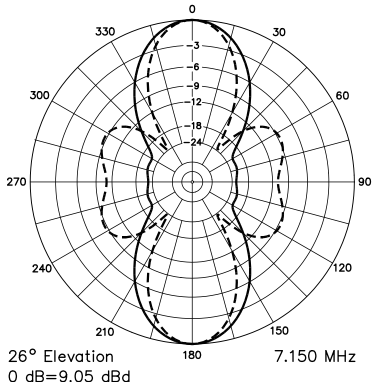

Fig 23—Azimuth pattern for N6LF Double Extended Zepp (solid line), compared to classic Double Extended Zepp (dashed line). The main lobe for the modified antenna is slightly broader than that of the classic model, and the sidelobes are suppressed better.

That's all there is to it. The radiation pattern, overlaid with that for a standard DEZepp for comparison, is shown in **Fig 23**. The sidelobes are now reduced to below 20 dB. The main lobe is now 43° wide at the 3-dB points, as opposed to 35° for the original DEZepp. The antenna has gain over a dipole for > 50° now and the gain of the main lobe has dropped only 0.2 dB below the original DEZepp.

Experimental Results

The antenna was made from #14 wire and the capacitors were made from 3.5-inch sections of RG-213, shown in **Fig 24A**. Note that great care should be taken to seal out moisture in these capacitors. The voltage across the capacitor for 1.5 kW will be about 2000 V so any corona will quickly destroy the capacitor.

A silicon sealant was used and then both ends covered

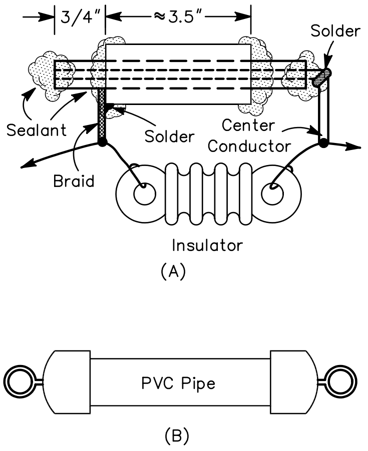

Fig 24—Construction details for series capacitor made from RG-213 coaxial cable. At A, the method used by N6LF is illustrated. At B, a suggested method to seal capacitor better against weather is shown, using a section of PVC pipe with end caps.

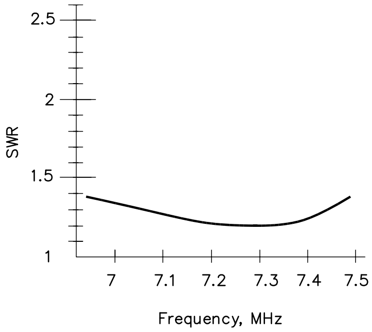

Fig 25—Measured SWR curve across 40-meter band for N6LF DEZepp.

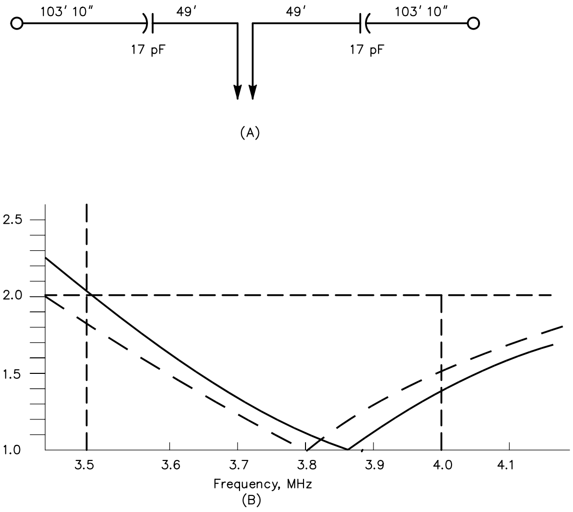

Fig 26—75/80-meter modified Double Extended Zepp, designed using *NEC Wires*. At A, a schematic is shown for antenna. At B, SWR curve is shown across 75/80-meter band. Solid line shows measured curve for W7ISV antenna, which was pruned to place SWR minimum higher in the band. The dashed curve shows the computed response when SWR minimum is set to 3.8 MHz.

with coax seal, finally wrapping it with plastic tape. The solder balls indicated on the drawing are to prevent wicking of moisture through the braid and the stranded center conductor. This is a small but important point if long service out in the weather is expected. An even better way to protect the capacitor would be to enclose it in a short piece of PVC pipe with end caps, as shown in Fig 24B.

Note that all RG-8 type cables do not have exactly the same capacitance per foot and there will also be some end effect adding to the capacitance. If possible the capacitor should be trimmed with a capacitance meter. It isn't necessary to be too exact—the effect of varying the capacitance ±10% was checked and the antenna still worked fine.

The results proved to be close to those predicted by the computer model. **Fig 25** shows the measured value for SWR across the band. These measurements were made with a Bird directional wattmeter. The worst SWR is 1.35:1 at the low end of the band.

Dick Ives, W7ISV, erected an 80-meter version of the antenna, shown in **Fig 26A**. The series capacitors are 17 pF. Since he isn't interested in CW, Dick adjusted the length for the lowest SWR at the high end of the band, as shown in the SWR curve (Fig 26B). The antenna could have been tuned somewhat lower in frequency and would then provide an SWR < 2:1 over the entire band, as indicated by the dashed line.

This antenna provides wide bandwidth and moderate gain over the entire 75/80-meter band. Not many antennas will give you that with a simple wire structure.

Vertical Antennas

On the low bands quarter-wave high vertical antennas become increasingly attractive, especially for DX work, because they provide a means for lowering the radiation angle. This is especially true where practical heights for horizontally polarized antennas are too low. In addition, verticals can be very simple and unobtrusive structures. For example, it is very easy to disguise a vertical as a flagpole. In fact an actual flagpole may be used as a vertical. Performance of a vertical is determined by several factors:

- Height of the vertical portion of the radiator
- The ground or counterpoise system efficiency, if one is used
- Ground characteristics in the near- and far-field regions
- The efficiency of loading elements and matching networks

THE HALF-WAVE VERTICAL DIPOLE (HVD)

The simplest form of vertical is that of a half-wave vertical dipole, an HVD. This is a horizontal dipole turned 90º so that it is perpendicular to the ground under it. Of course, the top end of such an antenna must be at least a half wave above the ground or else it would be touching the ground. This poses quite a construction challenge if the builder wants a free-standing low-frequency antenna. Hams fortunate enough to have tall trees on their property can suspend wire HVDs from these trees. Similarly, hams with two tall towers can run rope catenaries between them to hold up an HVD.

A vertical half-wave dipole has some operational advantages compared to a more-commonly used vertical configuration—the quarter-wave vertical used with some sort of above-ground counterpoise or an on-ground radial system. See **Fig 27** and 27B, which shows the two configurations discussed here. In each case, the lowest part of each antenna is 8 feet above ground, to prevent passersby from being able to touch any live wire. Each antenna is assumed to be made of #14 wire resonant on 80 meters.

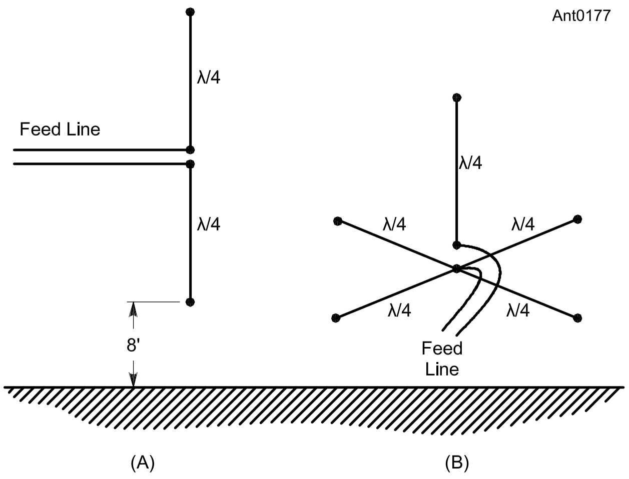

Fig 27—At A, a 80-meter half-wave vertical dipole elevated 8 feet above the ground. The feed line is run perpendicularly away from the dipole. At B, a "ground plane" type of quarter-wave vertical, with four elevated resonant radials. Both antennas are mounted 8 feet above the ground to keep them away from passersby.

Feeding a Half-Wave Vertical Dipole

Fig 28 compares elevation patterns for the two antennas for "average ground." You can see that the half-wave vertical dipole has about 1.5 dB higher peak gain, since it compresses the vertical elevation pattern down somewhat closer to the horizon than does the quarter-wave ground plane. Another advantage to using a half-wave radiator besides higher gain is that less horizontal "real estate" is needed compared to a quarter-wave vertical with its horizontal radials.

The obvious disadvantage to an HVD is that it is taller than a quarter-wave ground plane. This requires a higher support (such as a taller tree) if you make it from wire, or a longer element if you make it from telescoping aluminum tubing.

Another problem is that theory says you must dress the feed line so that it is perpendicular to the half-wave radiator. This means you must support the coax feed line above ground for some distance before bringing the coax down to ground level. A question immediately arises: How far must you go out horizontally with the feed line before going to ground level to eliminate common-mode currents that are radiated onto the coax shield? Such common-mode currents will affect the feed-point impedance as well as the radiation pattern for the antenna system. Quite a bit of distortion in the azimuthal pattern can be created if common-mode currents aren't suppressed, usually by using a common-mode choke, also known as a current balun.

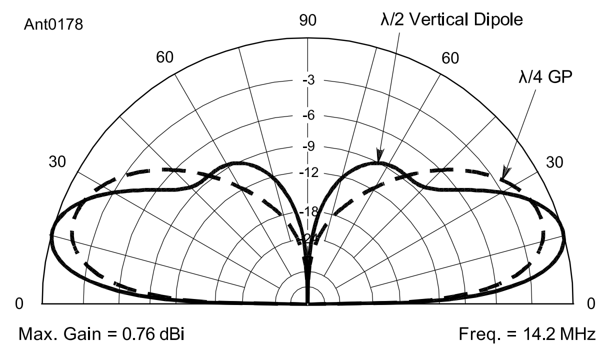

Fig 28—A comparison of the elevation patterns for the two antennas in Fig 27. The peak gain of the HVD is about 1.5 dB higher than that for the quarter-wave ground-plane radiator with radials.

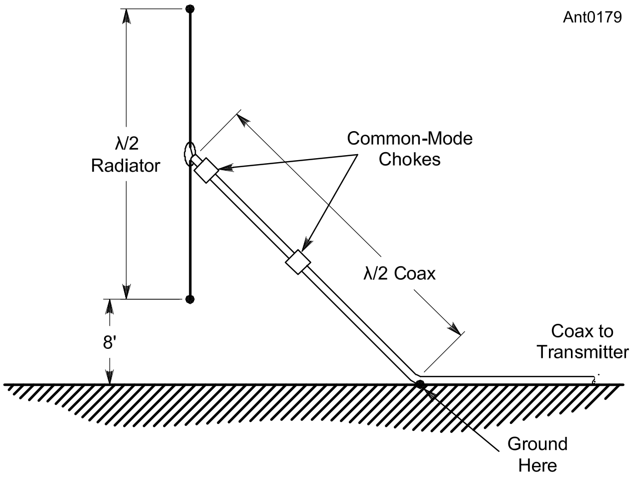

Fig 29—A 20-meter HVD whose bottom is 8 feet above ground. This is fed with a λ/2 of RG-213 coax. This system uses a common-mode choke at the feed point and another λ/4 down the line. The resulting azimuthal radiation pattern is within 0.4 dB of being perfectly circular. The "wingspan" of this antenna system is 27 feet from the radiator to the point where the coax comes to ground level.

Constructing such a common-mode choke is very simple: Three large ferrite beads are slipped over the coax (before the connectors are soldered on or else they won't fit!) and taped in place. The only problem with this scheme is that an additional support (some sort of "skyhook") is required to support the coax horizontally. Let's try to simplify the installation, by slanting the feed-line coax down to ground from the feed point at a fairly steep angle of about 30º from vertical. See **Fig 29**.

Note that the bottom end of the coax in Fig 29 is grounded to a ground rod. This serves several purposes—this serves as a mechanical connection to hold the coax in place and it provides some protection against lightning strikes. Now, as a purely practical matter, just how picky are we being here? What if we skip the second common-mode choke and just use one at the feed point? The computer models predicts that there will be some distortion in the azimuthal pattern—about 1.1 dB worth. Whether this is serious is up to you. However, you may find other problems with common-mode currents on the coax shield—problems such as RF in the shack or variable SWR readings depending on the way coax is routed in the shack. The addition of three extra ferrite beads to suppress the common-mode currents is cheap insurance.

Later in this chapter we'll discuss shortened vertical antennas, ones arranged both as vertical dipoles and as vertical monopoles with radial systems.

MONOPOLE VERTICALS WITH GROUND PLANE RADIALS

For best performance the vertical portion of a ground-plane type of antenna should be λ/4 or more, but this is not an absolute requirement. With proper design, antennas as short as 0.1 λ or even less can be efficient and effective. Antennas shorter than λ/4 will be reactive and some form of loading and perhaps a matching network will be required.

If the radiator is made of wire supported by nonconducting material, the approximate length for λ/4 resonance can be found from:

$$\ell_{\text{feet}} = \frac{234}{f_{\text{MHz}}} \qquad \text{(Eq 1)}$$

For tubing, the length for resonance must be shorter than given by the above equation, as the length-to-diameter ratio is lower than for wire (see Chapter 2, Antenna Fundamentals). For a tower, the resonant length will be shorter still. In any case, after installation the antenna length (height) can be adjusted for resonance at the desired frequency.

The effect of ground characteristics on losses and elevation pattern is discussed in detail in Chapter 3, The Effects of Ground. The most important points made in that discussion are the effect of ground characteristics on the radiation pattern and the means for achieving low ground-loss resistance in a buried ground system. As ground conductivity increases, low-angle radiation improves. This makes a vertical very attractive to those who live in areas with good ground conductivity. If your QTH is on a saltwater beach, then a vertical would be very effective, even when compared to horizontal antennas at great height.

When a buried-radial ground system is used, the efficiency of the antenna will be limited by the loss resistance of the ground system. The ground can be a number of radial wires extending out from the base of the antenna for about λ/4. Driven ground rods, while satisfactory for electrical safety and for lightning protection, are of little value as an RF ground for a vertical antenna, except perhaps in marshy or beach areas. As pointed out in Chapter 3, many long radials are desirable. In general, however, a large number of short radials are preferable to only a few long radials, although the best system would have 60 or more radials longer than λ/4. An elevated system of radials or a ground screen (*counterpoise*) may be used instead of buried radials, and can result in an efficient antenna.

ELEVATED RADIALS AND COUNTERPOISES

Elevated radials, isolated from ground, can be used in place of an extensive buried radial system. Work by Al Christman, K3LC (ex-KB8I), has shown that 4 to 8 elevated radials can provide performance comparable to a 120 λ/4-long buried wires. This is especially important for the low bands, where such a buried ground system is very large and impractical for most amateurs. An elevated ground system is sometimes referred to as a *ground plane* or *counterpoise*. **Fig 30** compares buried and elevated ground systems, showing the difference in current flow in the two systems.

An elevated ground can take several forms. A number of wires arranged with radial symmetry around the base of the antenna is shown in Fig 30B. Four radials are normally

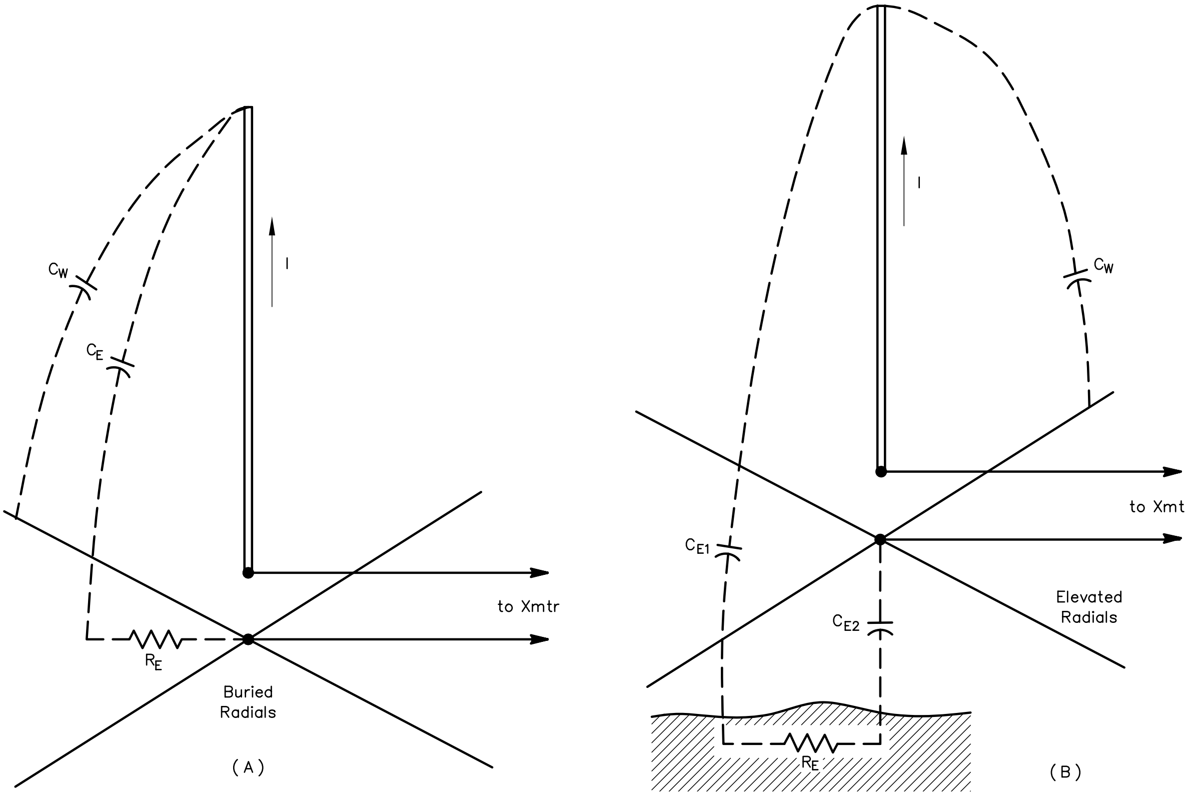

Fig 30—How earth currents affect the losses in a short vertical antenna system. At A, the current through the combination of C_E and R_E may be appreciable if C_E is much greater than C_W, the capacitance of the vertical to the ground wires. This ratio can be improved (up to a point) by using more radials. By raising the entire antenna system off the ground, C_E (which consists of the series combination of C_{E1} and C_{E2}) is decreased while C_W stays the same. The radial system shown at B is sometimes called a *counterpoise*.

used, but as few as two, or as many as eight, can be used. For a given height of vertical, the length of the radials can be adjusted to resonate the antenna. For a λ/4 vertical, the radials are normally λ/4 long.

In the case of a multiband vertical, two or more sets of radials, with different lengths, may be interleaved. The radials associated with each band are adjusted for resonance on their associated band.

A counterpoise is most commonly a system of elevated radials, where the radial wires are interconnected with jumpers, as shown in **Fig 31**. As illustrated in Fig 30, the purpose of the elevated-ground system is to provide a return path for the displacement currents flowing in the vicinity of the antenna. The idea is to minimize the current flowing through the ground itself, which is usually very lossy. By raising the radials above ground most of the current will flow in the radials, which are good conductors. This allows a simple radial system to provide a very efficient ground. However, there is a price to be paid for this.

The ground system now has a direct effect on the feedpoint impedance, introducing reactance as well as resistance, and is relatively narrow band. For a given vertical height, the

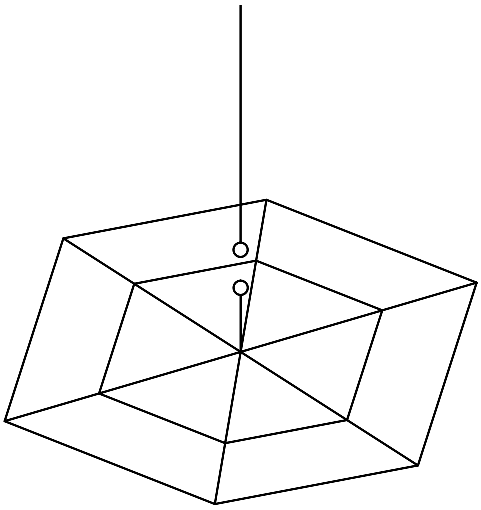

Fig 31—Counterpoise, showing the radial wires connected together by cross wires. The length of the perimeter of the individual meshes should be < λ/4 to prevent undesired resonances. Sometimes the center portion of the counterpoise is made from wire mesh.

Table 2

Illustration of the effect of variable vertical height (L_1) on elevated radial length (L2) and R_R.
#12 wire, elevated 5 feet over average ground at 3.525 MHz.

L_1 (λ)	L_1 (feet)	L_2 (feet)	R_R (Ω)
0.225	62.8	94	28.8
0.25	69.8	67	38.4
0.27	75.0	45	51.0
0.3	83.7	24	75.9

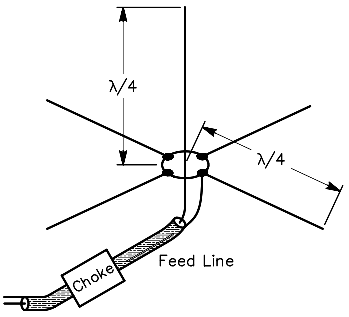

Fig 32—The ground-plane antenna. Power is applied between the base of the vertical radiator and the center of the ground plane, as indicated in the drawing. Decoupling from the transmission line and any conductive support structure is highly desirable.

radial length must be adjusted to resonate the antenna. The length of the radials must be readjusted for each band if a multiband vertical is used. As pointed out above, this usually means the installation of a set of radials for each band. To minimize current flowing in the ground, the antenna, ground plane and feed line must be isolated from ground for RF. More on this later.

The height of the vertical does not have to be exactly λ/4. Other lengths may be used and the antenna may be resonated by adjusting the length of the radials. **Table 2** gives a comparison between three different vertical lengths in an antenna using four elevated radials at 3.525 MHz.

An important feature of Table 2 is the dramatic reduction in radial length (L_2) with even a small increase in vertical height (L_1). For example, increasing the height by 5 feet reduces the radial length by 22 feet on 80 meters. On the other hand even a small decrease in L_1 can cause a substantial increase in L_2. This would be very undesirable, since the area required by the radials is already considerable. Notice also that the small increase in height raises R_R to 51 Ω. This trick of increasing the height slightly to reduce the size of the elevated ground system and to increase the input resistance can be very useful. In a following section the use of *top loading* for short antennas will be discussed. Top loading can also be used on a λ/4 vertical to achieve the same effect as increasing the height—the ability to use shorter radials and a better match.

GROUND-PLANE ANTENNAS

The ground-plane antenna is a λ/4 vertical with four radials, as shown in **Fig 32**. The entire antenna is elevated above ground. A practical example of a 7-MHz ground-plane antenna is given in **Fig 33**. As explained earlier, elevating the antenna reduces the ground loss and lowers the radiation angle somewhat. The radials are sloped downward to make the feed-point impedance closer to 50 Ω.

The feed-point impedance of the antenna varies with the height above ground, and to a lesser extent varies with the ground characteristics. **Fig 34** is a graph of feed-point resistance (R_R) for a ground-plane antenna with the radials parallel to the ground. R_R is plotted as a function of height above ground. Notice that the difference between perfect ground and average ground (ε=13 and σ = 0.005 S/m) is small, except when quite close to ground. Near ground R_R is between 36 and 40 Ω. This is a reasonable match for 50-Ω feed line but as the antenna is raised above ground R_R drops to approximately 22 Ω, which is not a very good match. The feed-point resistance can be increased by sloping the radials downward, away from the vertical section.

The effect of sloping the radials is shown in **Fig 35**. The

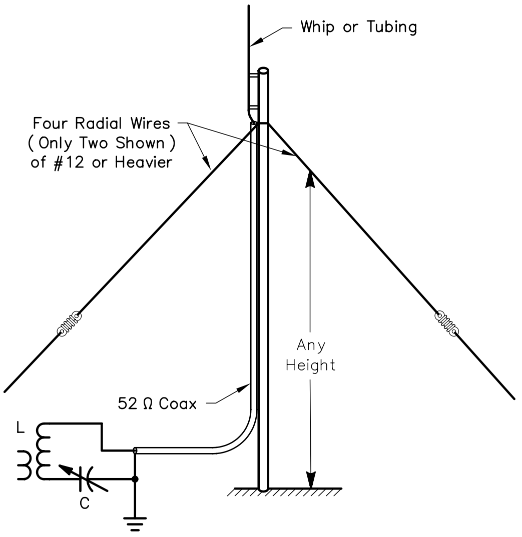

Fig 33—A ground-plane antenna is effective for DX work on 7 MHz. Although its base can be any height above ground, losses in the ground underneath will be reduced by keeping the bottom of the antenna and the ground plane as high above ground as possible. Feeding the antenna directly with 50-Ω coaxial cable will result in a low SWR. The vertical radiator and the radials are all λ/4 long electrically. Contrary to popular myth, the radials need not necessarily be 5% longer than the radiator. Their physical length will depend on their length-to-diameter ratios, the height over ground and the length of the vertical radiator, as discussed in text.

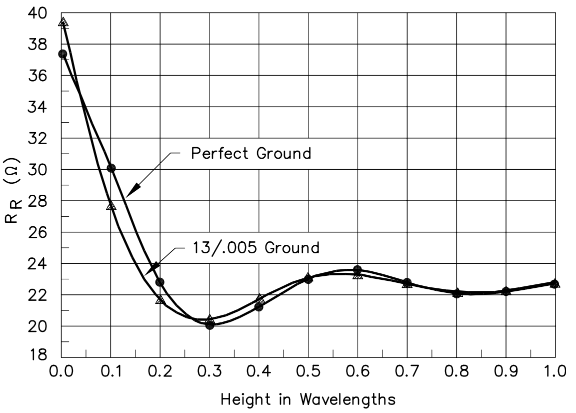

Fig 34—Radiation resistance of a 4-radial ground-plane antenna as a function of height over ground. Perfect and average ground are shown. Frequency is 3.525 MHz. Radial angle (θ) is 0°.

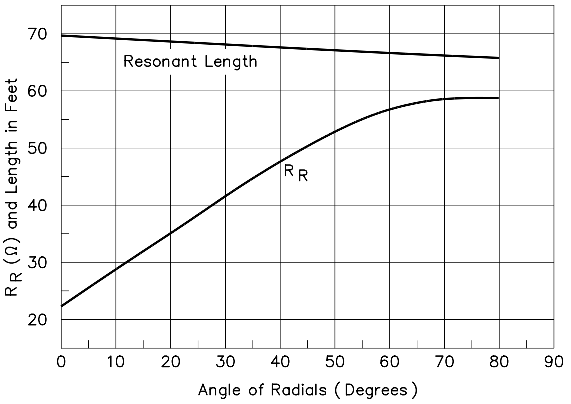

Fig 35—Radiation resistance and resonant length for a 4-radial ground-plane antenna > 0.3 λ above ground as a function of radial droop angle (θ).

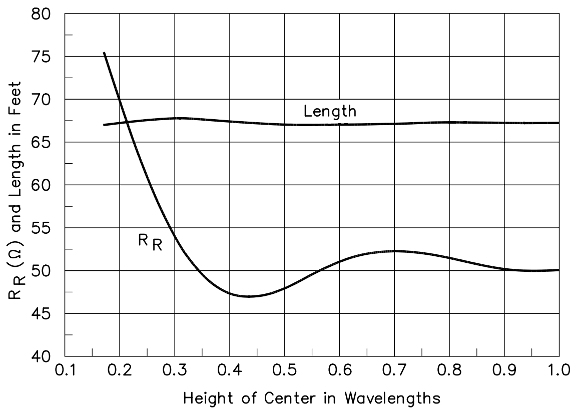

Fig 36—Radiation resistance and resonant length for a 4-radial ground-plane antenna for various heights above average ground for radial droop angle θ = 45°.

graph is for an antenna well above ground (> 0.3 λ). Notice that R_R = 50 Ω when the radials are sloped downward at an angle of 45°, a convenient value. The resonant length of the antenna will vary slightly with the angle. In addition, the resonant length will vary a small amount with height above the ground. It is for these reasons, as well as the effect of conductor diameter, that some adjustment of the radial lengths is usually required. When the ground-plane antenna is used on the higher HF bands and at VHF, the height above ground is usually such that a radial sloping angle of 45° will give a good match to 50-Ω feed line.

The effect of height on R_R with a radial angle of 45° is shown in **Fig 36**. At 7 MHz and lower, it is seldom possible to elevate the antenna a significant portion of a wavelength and the radial angle required to match to 50-Ω line is usually of the order of 10° to 20°. To make the vertical portion of the antenna

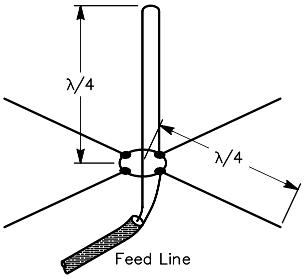

Fig 37—The folded monopole antenna. Shown here is a ground plane of four λ/4 radials. The folded element may be operated over an extensive counterpoise system or mounted on the ground and worked against buried radials and the earth. As with the folded dipole antenna, the feed-point impedance depends on the ratios of the radiator conductor sizes and their spacing.

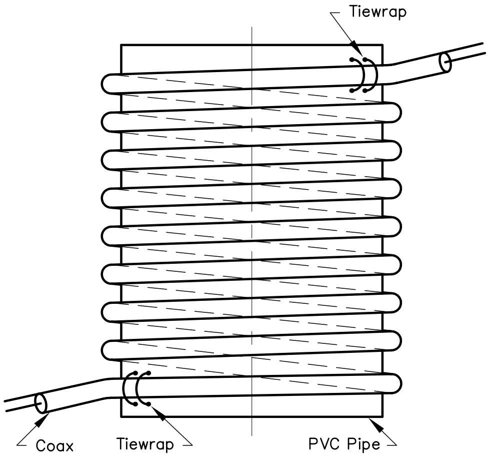

Fig 38—A choke balun with sufficient impedance to isolate the antenna properly can be made by winding coaxial cable around a section of plastic pipe. Suitable dimensions are given in the text.

as long as possible, it may be better to accept a slightly poorer match and keep the radials parallel to ground.

The principles of the folded dipole (Fig 18) can also be applied to the ground-plane antenna, as shown in **Fig 37**. This is the *folded monopole* antenna. The feed-point resistance can be controlled by the number of parallel vertical conductors and the ratios of their diameters.

As mentioned earlier, it is important in most installations to isolate the antenna from the feed line and any conductive supporting structure. This is done to minimize the return current conducted through the ground. A return current on the feed line itself or the support structure can drastically alter the radiation pattern, usually for the worse. For these reasons, a balun (see Chapter 26, Coupling the Line to the Antenna) or other isolation scheme must be used. 1:1 baluns are effective for the higher bands but at 3.5 and 1.8 MHz commercial baluns often have too low a shunt inductance to provide adequate isolation. It is very easy to recognize when the isolation is inadequate. When the antenna is being adjusted while watching an isolated impedance or SWR meter, adjustments may be sensitive to your touching the instrument. After adjustment and after the feed line is attached, the SWR may be drastically different. When the feed line is inadequately isolated, the apparent resonant frequency or the length of the radials required for resonance may also be significantly different from what you expect.

In general, an isolation choke inductance of 50 to 100 μH will be needed for 3.5 and 1.8-MHz ground-plane antennas. One of the easiest ways to make the required isolation choke is to wind a length of coaxial cable into a coil as shown in **Fig 38**. For 1.8 MHz, 30 turns of RG-213 wound on a 14-inch length of 8-inch diameter PVC pipe, will make a very good isolation choke that can handle full legal power continuously. A smaller choke could be wound on 4-inch diameter plastic drain pipe using RG-8X or a Teflon insulated cable. The important point here is to isolate or decouple the antenna from the feed line and support structure.

A full-size ground-plane antenna is often a little impractical for 3.5-MHz and quite impractical for 1.8 MHz, but it can be used at 7 MHz to good advantage, particularly for DX work. Smaller versions can be very useful on 3.5 and 1.8 MHz.

EXAMPLES OF VERTICALS

There are many possible ways to build a vertical antenna—the limits are set by your ingenuity. The primary problem is creating the vertical portion of the antenna with sufficient height. Some of the more common means are:

- A dedicated tower
- Using an existing tower with an HF Yagi on top
- A wire suspended from a tree limb or the side of a building
- A vertical wire supported by a line between two trees or other supports
- A tall pole supporting a conductor
- Flagpoles
- Light standards
- Irrigation pipe
- TV masts

If you have the space and the resources, the most straightforward means is to erect a dedicated tower for a vertical. While this is certainly an effective approach, many amateurs do not have the space or the funds to do this, especially if they already have a tower with an HF antenna on the top. The existing tower can be used as a top-loaded vertical, using shunt feed and a ground radial system. A system like this is shown in **Fig 39B**.

For those who live in an area with tall trees, it may be possible to install a support rope between two trees, or between a tree and an existing tower. (Under no circumstances should you use an active utility pole!) The vertical portion of the antenna can be a wire suspended from the support line to ground, as shown in Fig 39C. If top loading is needed, some or all of the support line can be made part of the antenna.

Your local utility company will periodically have older power poles that they no longer wish to keep in service. These are sometimes available at little or no expense. If you see a power line under reconstruction or repair in your area you might stop and speak with the crew foreman. Sometimes they will have removed older poles they will not use again and will have to haul them back to their shop for disposal. Your offer for local "disposal" may well be accepted. Such a pole can be used in conjunction with a tubing or whip extension such as that shown in Fig 39A. Power poles are not your only option. In some areas of the US, such as the southeast or northwest, tall poles made directly from small conifers are available.

Freestanding (unguyed) flagpoles and roadway illumination standards are available in heights exceeding 100 feet. These are made of fiberglass, aluminum or galvanized steel. All of these are candidates for verticals. Flagpole suppliers are listed under "Flags and Banners" in your Yellow Pages. For lighting standards (lamp posts), you can contact a local electrical hardware distributor. Like a wooden pole, a fiberglass flagpole does not require a base insulator, but metal poles do. Guy wires will be needed.

One option to avoid the use of guys and a base insulator is to mount the pole directly into the ground as originally intended and then use shunt feed. If you want to keep the pole grounded but would like to use elevated radials, you can attach a cage of wires (four to six) at the top as shown in Fig 39D. The cage surrounds the pole and allows the pole (or tower for that matter) to be grounded while allowing elevated radials to be used. The use of a cage of wires surrounding the pole or tower is a very good way to increase the effective diameter. This reduces the Q of the antenna, thereby increasing the bandwidth. It can also reduce the conductor loss, especially if the pole is galvanized steel, which is not a very good RF conductor.

Aluminum irrigation tubing, which comes in diameters of 3 and 4 inches and in lengths of 20 to 40 feet, is widely available in rural areas. One or two lengths of tubing con-

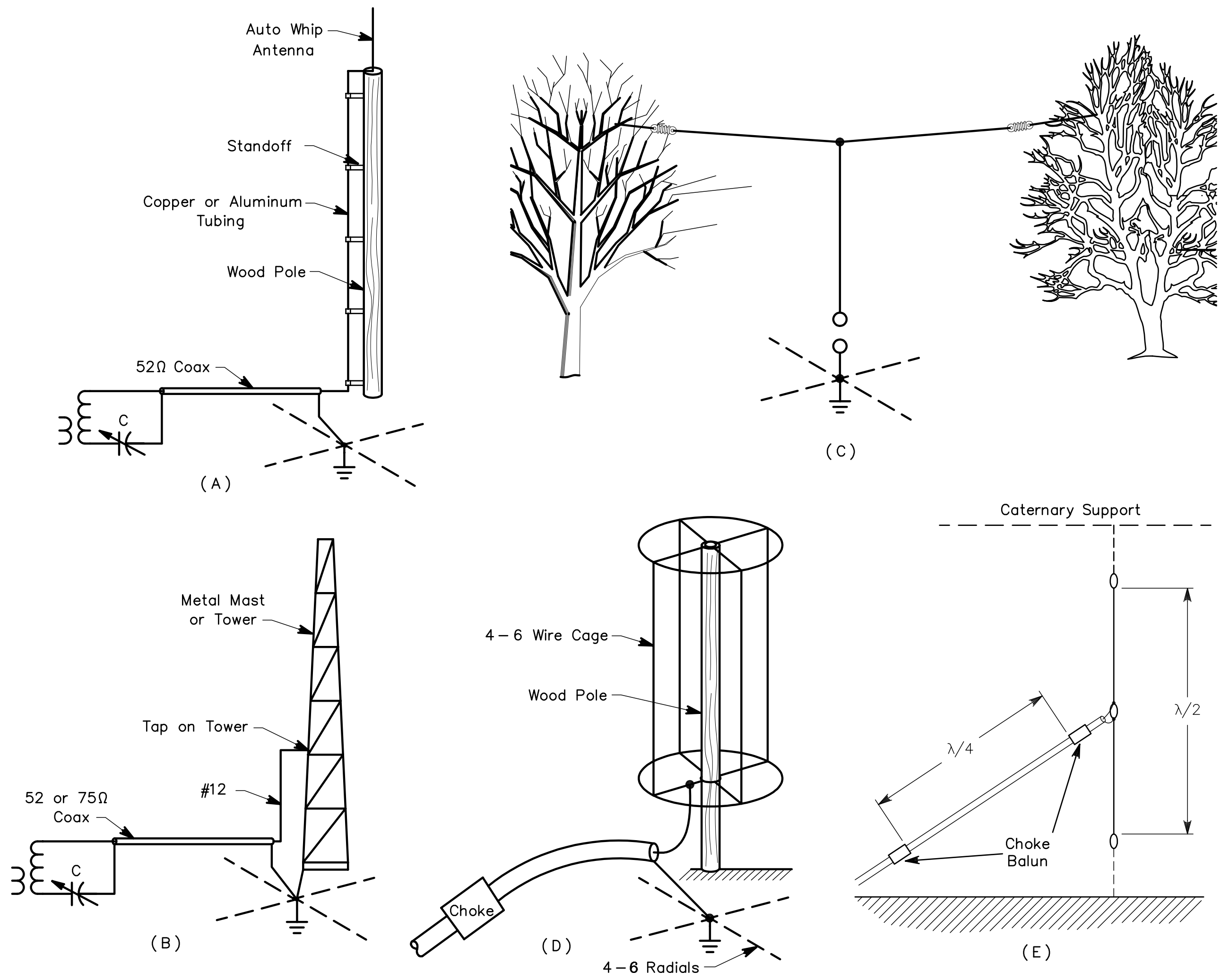

Fig 39—Vertical antennas are effective for 3.5- or 7-MHz work. The λ/4 antenna shown at A is fed directly with 50-Ω coaxial line, and the resulting SWR is usually less than 1.5 to 1, depending on the ground resistance. If a grounded antenna is used as at B, the antenna can be shunt fed with either 50- or 75-Ω coaxial line. The tap for best match and the value of C will have to be found by experiment. The line running up the side of the antenna should be spaced 6 to 12 inches from the antenna. If tall trees are available the antenna can be supported from a line suspended between the trees, as shown in C. If the vertical section is not long enough then the horizontal support section can be made of wire and act as top loading. A pole or even a grounded tower can be used with elevated radials if a cage of four to six wires is provided as shown in D. The cage surrounds the pole which may be wood or a grounded conductor.

nected together can make a very good vertical when guyed with non-conducting line. It is also very lightweight and relatively easy to erect. A variety of TV masts are available which can also be used for verticals.

1.8 TO 3.5-MHz VERTICAL USING AN EXISTING TOWER

A tower can be used as a vertical antenna, provided that a good ground system is available. The shunt-fed tower is at its best on 1.8 MHz, where a full λ/4 vertical antenna is rarely possible. Almost any tower height can be used. If the beam structure provides some top loading, so much the better, but anything can be made to radiate—if it is fed properly. W5RTQ (now K6SE) uses a self-supporting, aluminum, crank-up, tilt-over tower, with a TH6DXX tribander mounted at 70 feet. Measurements showed that the entire structure has about the same properties as a 125-foot vertical. It thus works quite well as an antenna on 1.8 and 3.5 MHz for DX work requiring low-angle radiation.

Preparing the Structure

Usually some work on the tower system must be done before shunt-feeding is tried. If present, metallic guys should be broken up with insulators. They can be made to simulate top loading, if needed, by judicious placement of the first insulators. Don't overdo it; there is no need to "tune the radiator to resonance" in this way since a shunt feed is employed. If the tower is fastened to a house at a point more than about one-fourth of the height of the tower, it may be desirable to insulate the tower from the building. Plexiglas sheet, ¼-inch or more thick, can be bent to any desired shape for this purpose, if it is heated in an oven and bent while hot.

All cables should be taped tightly to the tower, on the inside, and run down to the ground level. It is not necessary to bond shielded cables to the tower electrically, but there should be no exceptions to the down-to-the-ground rule.

A good system of buried radials is very desirable. The ideal would be 120 radials, each 250 feet long, but fewer and shorter ones must often suffice. You can lay them around corners of houses, along fences or sidewalks, wherever they can be put a few inches under the surface, or even on the earth's surface. Aluminum clothesline wire may be used extensively in areas where it will not be subject to corrosion. Neoprene-covered aluminum wire will be better in highly acid soils. Contact with the soil is not important. Deep-driven ground rods and connection to underground copper water pipes may be helpful, if available, especially to provide some protection from lightning.

Installing the Shunt Feed

Principal details of the shunt-fed tower for 1.8 and 3.5 MHz are shown in **Fig 40**. Rigid rod or tubing can be used for the feed portion, but heavy gauge aluminum or copper wire is easier to work with. Flexible stranded #8 copper wire is used at W5RTQ (now K6SE) for the 1.8-MHz feed, because when the tower is cranked down, the feed wire must come down with it. Connection is made at the top, 68 feet, through a 4-foot length of aluminum tubing clamped to the top of the tower, horizontally. The wire is clamped to the tubing at the outer end, and runs down vertically through standoff insulators. These are made by fitting 12-inch lengths of PVC plastic water pipe over 3-foot lengths of aluminum tubing. These are clamped to the tower at 15- to 20-foot intervals, with the bottom clamp about 3 feet above ground. These lengths allow for adjustment of the tower-to-wire spacing over a range of about 12 to 36 inches, for impedance matching.

The gamma-match capacitor for 1.8 MHz is a 250-pF variable with about ⅙-inch plate spacing. This is adequate for power levels up to about 200 W. A large transmitting or a vacuum-variable capacitor should be used for high-power applications.

Tuning Procedure

The 1.8-MHz feed wire should be connected to the top of the structure if it is 75 feet tall or less. Mount the standoff insulators so as to have a spacing of about 24 inches between wire and tower. Pull the wire taut and clamp it in place at the bottom insulator. Leave a little slack below to permit adjustment of the wire spacing, if necessary.

Adjust the series capacitor in the 1.8-MHz line for minimum reflected power, as indicated on an SWR meter connected between the coax and the connector on the capacitor housing. Make this adjustment at a frequency near the middle of your expected operating range. If a high SWR is indicated, try moving the wire closer to the tower. Just the

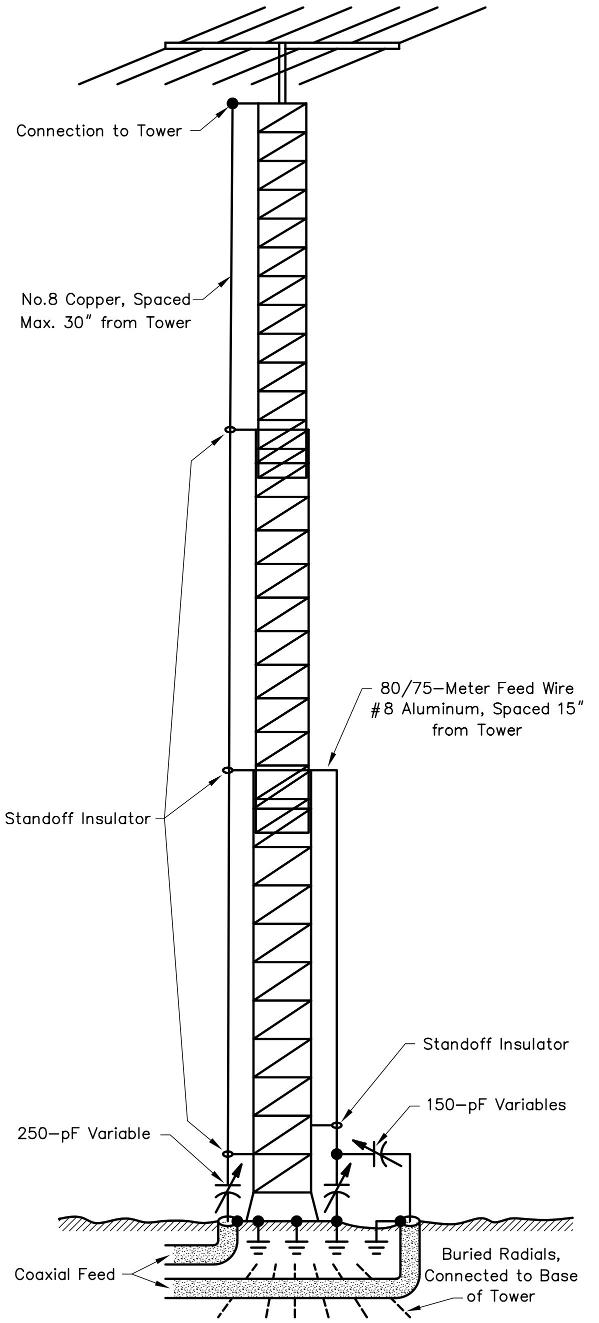
Connection to Tower
No.8 Copper, Spaced Max. 30" from Tower
80/75–Meter Feed Wire #8 Aluminum, Spaced 15" from Tower
Standoff Insulator
Standoff Insulator
150–pF Variables
250–pF Variable
Coaxial Feed
Buried Radials, Connected to Base of Tower

Fig 40—Principal details of the shunt-fed tower at W5RTQ (now K6SE). The 1.8-MHz feed, left side, connects to the top of the tower through a horizontal arm of 1-inch diameter aluminum tubing. The other arms have standoff insulators at their outer ends, made of 1-foot lengths of plastic water pipe. The connection for 3.5-4 MHz, right, is made similarly, at 28 feet, but two variable capacitors are used to permit adjustment of matching with large changes in frequency.

lower part of the wire need be moved for an indication as to whether reduced spacing is needed. If the SWR drops, move all insulators closer to the tower, and try again.

If the SWR goes up, increase the spacing. There will be a practical range of about 12 to 36 inches. If going down to 12 inches does not give a low SWR, try connecting the top a bit farther down the tower. If wide spacing does not make it, the omega match shown for 3.5-MHz work should be tried. No adjustment of spacing is needed with the latter arrangement, which may be necessary with short towers or installations having little or no top loading.

The two-capacitor arrangement in the omega match is also useful for working in more than one 25-kHz segment of the 1.8-MHz band. Tune up on the highest frequency, say 1990 kHz, using the single capacitor, making the settings of wire spacing and connection point permanent for this frequency. To move to the lower frequency, say 1810 kHz, connect the second capacitor into the circuit and adjust it for the new frequency. Switching the second capacitor in and out then allows changing from one segment to the other, with no more than a slight retuning of the first capacitor.

SIMPLE, EFFECTIVE, ELEVATED GROUND-PLANE ANTENNAS

This section describes a simple and effective means of using a grounded tower, with or without top-mounted antennas, as an elevated ground-plane antenna for 80 and 160 meters. It first appeared in a June 1994 *QST* article by Thomas Russell, N4KG.

From Sloper to Vertical

Recall the quarter-wavelength sloper, also known as the *half sloper*. [The half sloper is covered later in this chapter in more detail.—*Ed.*] It consists of an isolated quarter wavelength of wire, sloping from an elevated feed point on a grounded tower. Best results are usually obtained when the feed point is somewhere below a top-mounted Yagi antenna. You feed a sloper by attaching the center conductor of a coaxial cable to the wire and the braid of the cable to the tower leg. Now, imagine four (or more) slopers, but instead of feeding each individually, connect them together to the center conductor of a single feed line. Voilà! Instant elevated ground plane.

Now, all you need to do is determine how to tune the antenna to resonance. With no antennas on the top of the tower, the tower can be thought of as a fat conductor and should be approximately 4% shorter than a quarter wavelength in free space. Calculate this length and attach four insulated quarter-wavelength radials at this distance from the top of the tower. For 80 meters, a feed point 65 feet below the top of an unloaded tower is called for. The tower guys must be broken up with insulators for all such installations. For 160 meters, 130 feet of tower above the feed point is needed.

What can be done with a typical grounded-tower-and-Yagi installation? A top-mounted Yagi acts as a large capacitance hat, top loading the tower. Fortunately, top loading is the most efficient means of loading a vertical antenna.

The examples in **Table 3** should give us an idea of how much top loading might be expected from typical amateur antennas. The values listed in the *Equivalent Loading* column tell us the approximate vertical height replaced by the antennas listed in a top-loaded vertical antenna. To arrive at the remaining amount of tower needed for resonance, subtract these numbers from the non-loaded tower height needed for resonance. Note that for all but the 10-meter antennas, the equivalent loading equals or exceeds a quarter wavelength on 40 meters. For typical HF Yagis, this method is best used only on 80 and 160 meters.

Construction Examples

Consider this example: A TH7 triband Yagi mounted on a 40-foot tower. The TH7 has approximately the same overall dimensions as a full-sized 3-element 20-meter beam, but has more interlaced elements. Its equivalent loading is estimated to be 40 feet. At 3.6 MHz, 65 feet of tower is needed without loading. Subtracting 40 feet of equivalent loading, the feed point should be 25 feet below the TH7 antenna.

Ten quarter-wavelength (65-foot) radials were run from a nylon rope tied between tower legs at the 15-foot level, to various supports 10 feet high. Nylon cord was tied to the insulated, stranded, #18 wire, without using insulators. The radials are all connected together and to the center of an exact half wavelength (at 3.6 MHz) of RG-213 coax, which will repeat the antenna feed impedance at the other end. **Fig 41** is a drawing of the installation. The author used a Hewlett-Packard low-frequency impedance analyzer to measure the input impedance across the 80-meter band. An exact resonance (zero reactance) was seen at 3.6 MHz, just as predicted. The radiation resistance was found to be 17 Ω. The next question is, how to feed and match the antenna.

One good approach to 80-meter antennas is to tune them to the low end of the band, use a low-loss transmission line, and switch an antenna tuner in line for operation in the higher portions of the band. With a 50-Ω line, the 17-Ω radiation resistance represents a 3:1 SWR, meaning that an antenna tuner should be in-line for all frequencies. For short runs, it

Table 3
Effective Loading of Common Yagi Antennas

Antenna	*Boom Length (feet)*	*S (area, ft^2)*	*Equivalent Loading (feet)*
3L 20	24	768	39
5L 15	26	624	35
4L 15	20	480	31
3L 15	16	384	28
5L 10	24	384	28
4L 10	18	288	24
3L 10	12	192	20
TH7	24	—	40 (estimated)
TH3	14	—	27 (estimated)

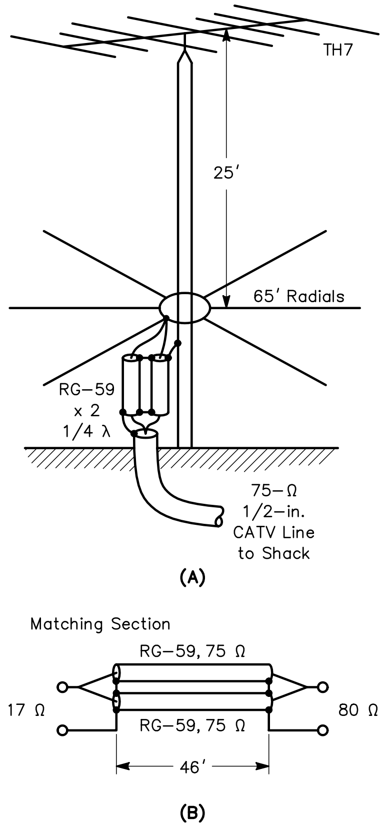

Fig 41—At A, an 80-meter top-loaded, reverse-fed elevated ground plane, using a 40-foot tower carrying a TH7 triband Yagi antenna. At B, dimensions of the 3.6-MHz matching network, made from RG-59.

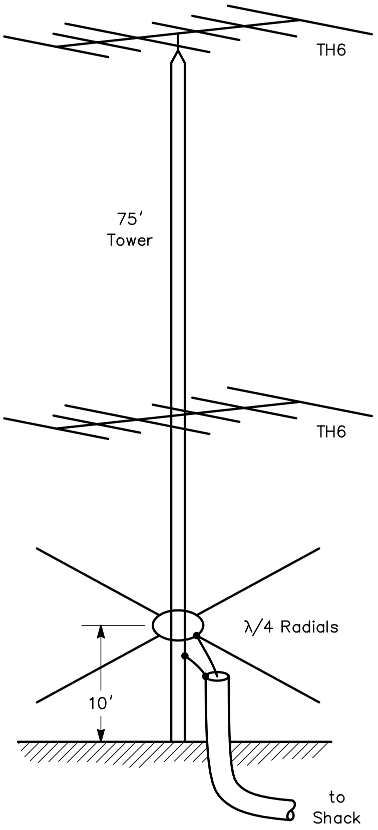

Fig 42—A 160-meter antenna using a 75-foot tower carrying stacked triband Yagis.

would be permissible to use RG-8 or RG-213 directly to the tuner. If you have a plentiful supply of low-loss 75-Ω CATV rigid coax, you can take another approach.

Make a quarter-wave (70 feet × 0.66 velocity factor = 46 feet) 37-Ω matching line by paralleling two pieces of RG-59 and connecting them between the feed point and a run of the rigid coax to the transmitter. The magic of quarter-wave matching transformers is that the input impedance (R_i) and output impedance (R_o) are related by:

$$Z_0^2 = R_i \times R_o \qquad \text{(Eq 2)}$$

For $R_i = 17\ \Omega$ and $Z_0 = 37\ \Omega$, $R_o = 80\ \Omega$, an almost perfect match for the 75-Ω CATV coax. The resulting 1.6:1 SWR at the transmitter is good enough for CW operation without a tuner.

160-Meter Operation

On the 160-meter band, a resonant quarter-wavelength requires 130 feet of tower above the radials. That's a pretty tall order. Subtracting 40 feet of top loading for a 3-element 20-meter or TH7 antenna brings us to a more reasonable 90 feet above the radials. Additional top loading in the form of more antennas will reduce that even more.

Another installation, using stacked TH6s on a 75-foot tower, is shown in **Fig 42**. The radials are 10 feet off the ground.

PHASED VERTICALS

Two or more vertical antennas spaced apart can be operated as a single antenna system to obtain additional gain and a directional pattern. There is an extensive discussion of phased arrays in Chapter 8, Multielement Arrays. Much of the material in Chapter 8 is useful for low-band antennas.

The Half-Square Antenna

The *half-square* antenna is a very simple form of vertical two-element phased array that can be very effective on the low bands. The following section was originally presented in *The ARRL Antenna Compendium Vol 5*, by Rudy Severns, N6LF.

A simple modification to a standard dipole is to add two λ/4 vertical wires, one at each end, as shown in **Fig 43**. This makes a *half-square antenna*. The antenna can be fed at one corner (low-impedance, current fed) or at the lower end of one of the vertical wires (high-impedance, voltage fed).

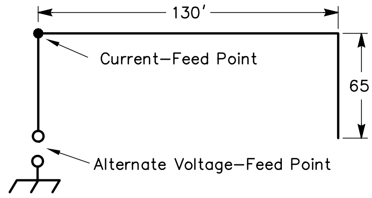

Fig 43—Typical 80-meter half-square, with λ/4-high vertical legs and a λ/2-long horizontal leg. The antenna may be fed at the bottom or at a corner. When fed at a corner, the feed point is a low-impedance, current-feed. When fed at the bottom of one of the wires against a small ground counterpoise, the feed point is a high-impedance, voltage-feed.

Other feed arrangements are also possible.

The "classical" dimensions for this antenna are λ/2 (131 feet at 3.75 MHz) for the top wire and λ/4 (65.5 feet) for the vertical wires. However, there is nothing sacred about these dimensions! They can vary over a wide range and still obtain nearly the same performance.

This antenna is two λ/4 verticals, spaced λ/2, fed in-phase by the top wire. The current maximums are at the top corners. The theoretical gain over a single vertical is 3.8 dB. An important advantage of this antenna is that it does not require the extensive ground system and feed arrangements that a conventional pair of phased λ/4 verticals would.

Comparison to a Dipole

In the past, one of the things that has turned off potential users of the half-square on 80 and 160 meters is the perceived need for λ/4 vertical sections. This forces the height to be > 65 feet on 80 meters and > 130 feet on 160 meters. That's not really a problem. If you don't have the height there are several things you can do. For example, just fold the ends in, as shown in **Fig 44**. This compromises the performance surprisingly little.

It is helpful to compare the examples given in Figs 43 and 44 to dipoles at the same height. Two heights, 40 and 80 feet, and average, very good and sea water grounds, were used for this comparison. It is also assumed that the lower end of the vertical wires had to be a minimum of 5 feet above ground.

At 40 feet the half-square is really mangled, with only 35-foot long (≈ λ/8) vertical sections. The elevation-plane comparison between this antenna and a dipole of the same height is shown in **Fig 45**. Over average ground the half-square is superior below 32° and at 15° is almost 5 dB better. That is a worthwhile improvement. If you have very good soil conductivity, like parts of the lower Midwest and South, then the half-square will be superior below 38° and at 15° will be nearly 8 dB better. For those fortunate few with saltwater frontal property the advantage at 15° is 11 dB! Notice also that above 35°, the response drops off rapidly. This is great for DX but is not good for local work.

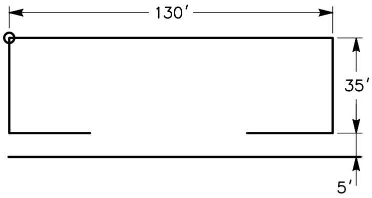

Fig 44—An 80-meter half-square configured for 40-foot high supports. The ends have been bent inward to reresonate the antenna. The performance is compromised surprisingly little.

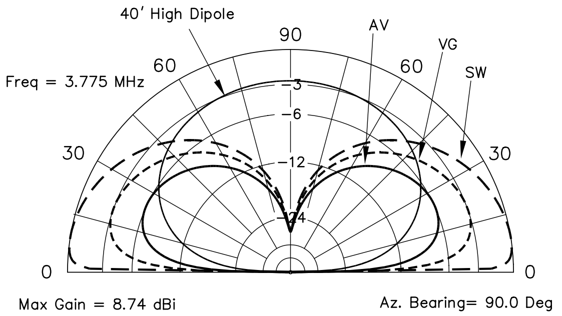

Fig 45—Comparison of 80-meter elevation response of 40-foot high, horizontally polarized dipole over average ground and a 40-foot high, vertically polarized half-square, over three types of ground: average (conductivity σ = 5 mS/m, dielectric constant ε = 13), very good (σ = 30 mS/m, ε = 20) and salt water (σ = 5000 mS/m, ε = 80). The quality of the ground clearly has a profound effect on the low-angle performance of the half-square. Even over average ground, the half-square outperforms the low dipole below about 32°.

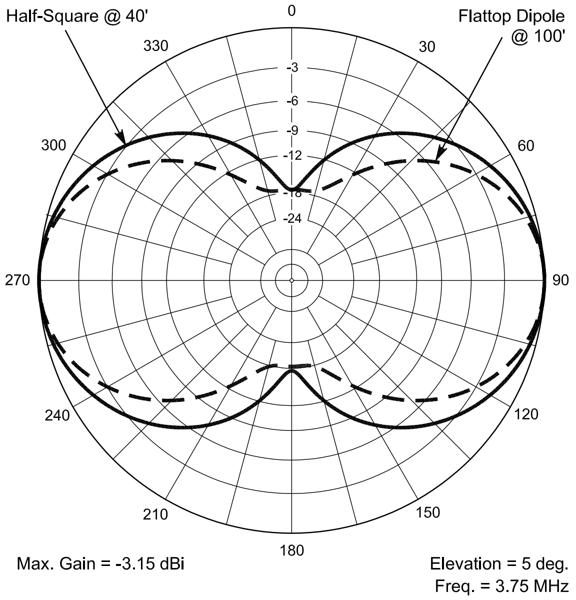

Fig 46—80-meter azimuth patterns for shortened half-square antenna (solid line) shown in Fig 44, compared with flattop dipole (dashed line) at 100 feet height. Average ground is assumed for these cases.

Fig 46 shows the azimuthal-plane pattern for the 80-meter half-square antenna in Fig 44, but this time compared with the response of a flattop horizontal dipole that is 100 feet high. These comparisons are for average ground and are for an elevation angle of 5°. The message here is that the lower your dipole and the better your ground, the more you have to gain by switching from a dipole to a half-square. The half-square antenna looks like a good bet for DXing.

Changing the Shape of the Half Square

Just how flexible is the shape? There are several common distortions of practical importance. Some have very little effect but a few are fatal to the gain. Suppose you have either more height and less width than called for in the standard version or more width and less height, as shown in **Fig 47A**.

The effect on gain from this type of dimensional variation is given in **Table 4**. For a top length (L_T) varying between 110 and 150 feet, where the vertical wire lengths (L_V) readjusted to resonate the antenna, the gain changes only by 0.6 dB. For a 1-dB change the range of L_T is 100 to 155 feet, a pretty wide range.

Another variation results if we vary the length of the horizontal top wire and readjust the vertical wires for resonance, while keeping the top at a constant height. See Fig 47B. **Table 5** shows the effect of this variation on the peak gain. For a range of L_T = 110 to 145 feet, the gain changes only 0.65 dB.

The effect of bending the ends into a V shape, as shown in Fig 47C, is given in **Table 6**. The bottom of the antenna is kept at a height of 5 feet and the top height (H) is either 40 or 60 feet. Even this gross deformation has only a relatively small effect on the gain. Sloping the ends outward as shown in Fig 47D and varying the top length also has only a small effect on the gain. While this is good news because it allows you dimension the antenna to fit different QTHs, not all distortions are so benign.

Table 4

Variation in Gain with Change in Horizontal Length, with Vertical Height Readjusted for Resonance (see Fig 47A)

L_T (feet)	*L_V (feet)*	*Gain (dBi)*
100	85.4	2.65
110	79.5	3.15
120	73.7	3.55
130	67.8	3.75
140	61.8	3.65
150	56	3.05
155	53	2.65

Table 5

Variation in Gain with Change in Horizontal Length, with Vertical Length Readjusted for Resonance, but Horizontal Wire Kept at Constant Height (see Fig 47B)

L_T (feet)	*L_V (feet)*	*Gain (dBi)*
110	78.7	3.15
120	73.9	3.55
130	68	3.75
140	63	3.35
145	60.7	3.05

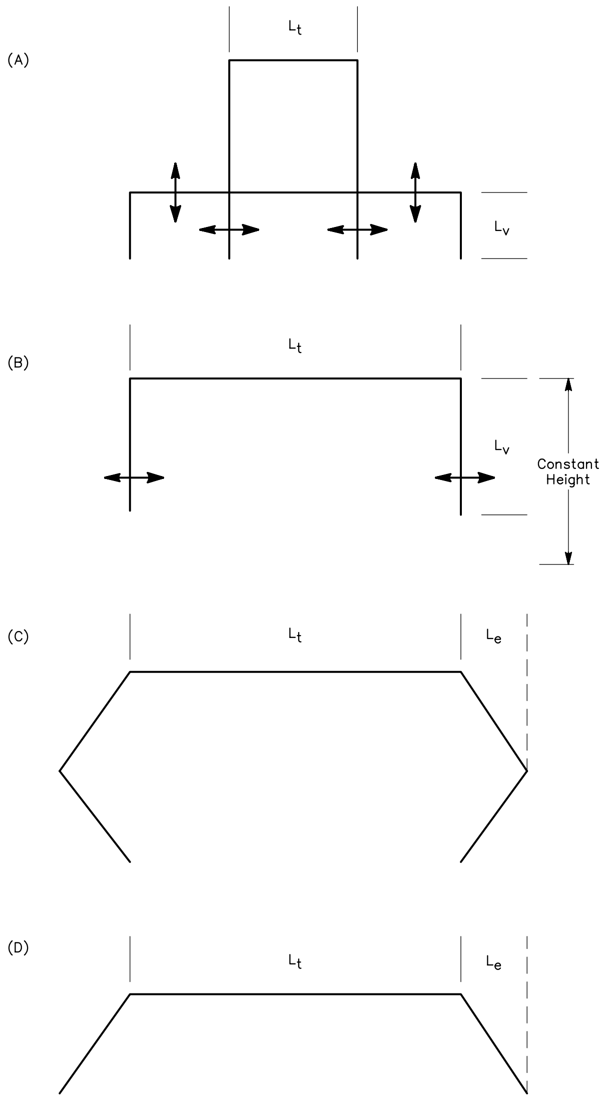

Fig 47—Varying the horizontal and vertical lengths of a half-square. At A, both the horizontal and vertical legs are varied, while keeping the antenna resonant. At B, the height of the horizontal wire is kept constant, while its length and that of the vertical legs is varied to keep the antenna resonant. At C, the length of the horizontal wire is varied and the legs are bent inwards in the shape of "vees." At D, the ends are sloped outward and the length of the flattop portion is varied. All these symmetrical forms of distortion of the basic half-square shape result in small performance losses. At E, a "halfwave vertical dipole" (HVD) with the feed coax isolated with common-mode choke baluns to keep RF current off the coax shield.

Table 6

Gain for Half-Square Antenna, Where Ends Are Bent Into V-Shape (see Fig 47 C)

Height ⇒	*H=40 feet*	*H=40 feet*	*H=60 feet*	*H=60 feet*
L_T (feet)	*L_e (feet)*	*Gain (dBi)*	*L_e (feet)*	*Gain (dBi)*
40	57.6	3.25	52.0	2.75
60	51.4	3.75	45.4	3.35
80	45.2	3.95	76.4	3.65
100	38.6	3.75	61.4	3.85
120	31.7	3.05	44.4	3.65
140	—	—	23	3.05

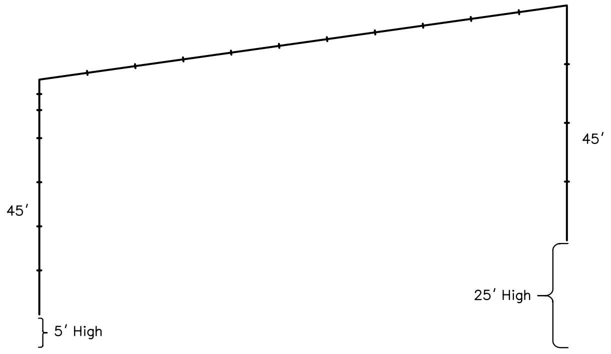

Fig 48—An asymmetrical distortion of the half-square antenna, where the bottom of one leg is purposely made 20 feet higher than the other. This type of distortion does affect the pattern!

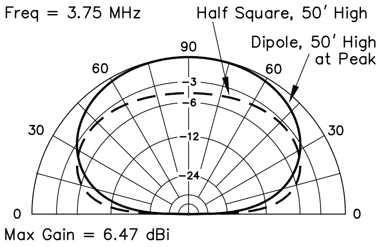

Fig 49—Elevation pattern for the asymmetrical half-square shown in Fig 48, compared with pattern for a 50-foot high dipole. This is over average ground, with a conductivity of 5 mS/m and a dielectric constant of 13. Note that the zenith-angle null has filled in and the peak gain is lower compared to conventional half-square shown in Fig 43 over the same kind of ground.

Suppose the two ends are not of the same height, as illustrated in **Fig 48**, where one end of the half-square is 20 feet higher than the other. The elevation-plane radiation pattern for this antenna is shown in **Fig 49** compared to a dipole at 50 feet. This type of distortion does affect the pattern. The gain drops somewhat and the zenith null goes away. The nulls off the end of the antenna also go away, so that there is some end-fire radiation. In this example the difference in height is fairly extreme at 20 feet. Small differences of 1 to 5 feet do not affect the pattern seriously.

If the top height is the same at both ends but the length of the vertical wires is not the same, then a similar pattern distortion can occur. The antenna is very tolerant of symmetrical distortions but it is much less accepting of asymmetrical distortion.

What if the length of the wires is such that the antenna is not resonant? Depending on the feed arrangement, that may or may not matter. We will look at that issue later on, in the section on patterns versus frequency. The half-square antenna, like the dipole, is very flexible in its proportions.

Half-Square Feed-Point Impedance

There are many different ways to feed the half-square. Traditionally the antenna has been fed either at the end of one of the vertical sections, against ground, or at one of the upper corners as shown in Fig 43.

For voltage feed at the bottom against ground, the impedance is very high, on the order of several thousand ohms. For current feed at a corner, the impedance is much lower and is usually close to 50 Ω. This is very convenient for direct feed with coax.

The half-square is a relatively high-Q antenna (Q ≈ 17). **Fig 50** shows the SWR variation with frequency for this feed arrangement. An 80-meter dipole is not particularly wideband either, but a dipole will have less extreme variation in SWR than the half-square.

Patterns Versus Frequency

Impedance is not the only issue when defining the bandwidth of an antenna. The effect on the radiation pattern of changing frequency is also a concern. For a voltage-fed half-square, the current distribution changes with frequency. For an antenna resonant near 3.75 MHz, the current distribution is nearly symmetrical. However, above and below resonance the current distribution increasingly becomes asymmetrical. In effect, the open end of the antenna is constrained to be a voltage maximum but the feed point can behave less as a voltage point and more like a current maxima. This allows the current distribution to become asymmetrical.

The effect is to reduce the gain by –0.4 dB at 3.5 MHz and by –0.6 dB at 4 MHz. The depth of the zenith null is reduced from –20 dB to –10 dB. The side nulls are also reduced. Note that this is exactly what happened when the antenna was made physically asymmetrical. Whether the asymmetry is due to current distribution or mechanical arrangements, the antenna pattern will suffer.

When current-feed at a corner is used, the asymmetry introduced by off-resonance operation is much less, since both ends of the antenna are open circuits and constrained to be voltage maximums. The resulting gain reduction is only –0.1 dB. It is interesting that the sensitivity of the pattern to changing frequency depends on the feed scheme used.

Of more concern for corner feed is the effect of the transmission line. The usual instruction is to simply feed

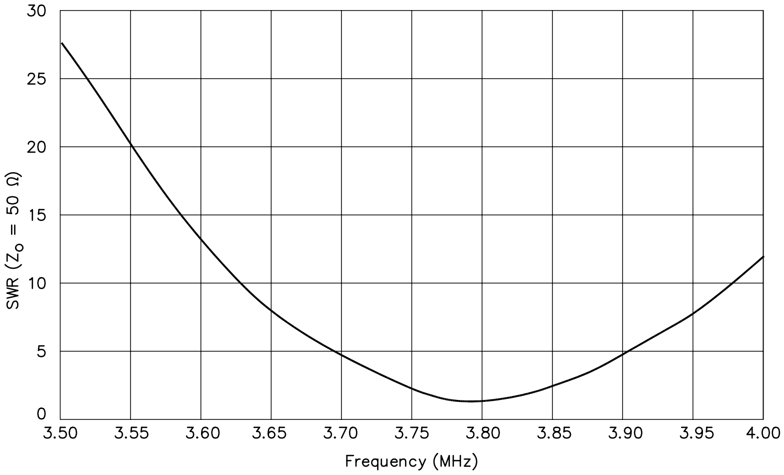

Fig 50—Variation of SWR with frequency for current-fed half-square antenna. The SWR bandwidth is quite narrow.

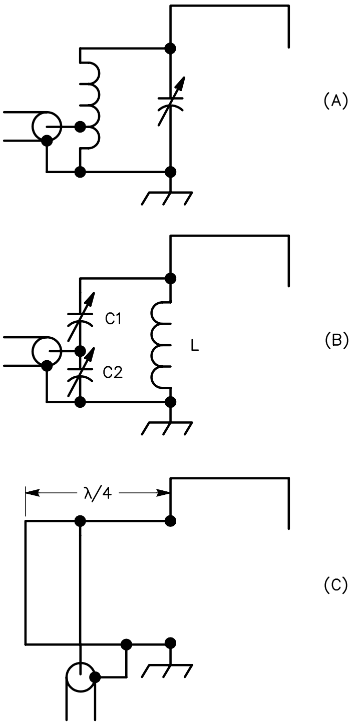

Fig 51—Typical matching networks used for voltage-feeding a half-square antenna.

the antenna using coax, with the shield connected to vertical wire and the center conductor to the top wire. Since the shield of the coax is a conductor, more or less parallel with the radiator, and is in the immediate field of the antenna, you might expect the pattern to be seriously distorted by this practice. This arrangement seems to have very little effect on the pattern. The greatest effect is when the feed line length was near a multiple of λ/2. Such lengths should be avoided.

Of course, you may use a choke balun at the feed point if you desire. This might reduce the coupling to the feed line even further but it doesn't appear to be worth the trouble. In fact, if you use an antenna tuner in the shack to operate away from resonance with a very high SWR on the transmission line, a balun at the feed point would take a beating.

Voltage-Feed at One End of Antenna: Matching Schemes

Several straightforward means are available for narrow-band matching. However, broadband matching over the full 80-meter band is much more challenging. Voltage feed with a parallel-resonant circuit and a modest local ground, as shown in **Fig 51**, is the traditional matching scheme for this antenna. Matching is achieved by resonating the circuit at the desired frequency and tapping down on the inductor in Fig 51A or using a capacitive divider (Fig 51B). It is also possible to use a λ/4 transmission-line matching scheme, as shown in Fig 51C.

If the matching network shown in Fig 51B is used, typical values for the components would be: L = 15 µH, C1 = 125 pF and C2 = 855 pF. At any single point the SWR can be made very close to 1:1 but the bandwidth for SWR < 2:1 will be very narrow at <100 kHz. Altering the L-C ratio doesn't make very much difference. The half-square antenna has a well-earned reputation for being narrowband.

Short Vertical Antennas

On the lower frequencies it becomes increasingly difficult to accommodate a full λ/4 vertical height and full-sized λ/4 radials, or even worse, a full-sized half-wave vertical dipole (HVD). In fact, it is not absolutely necessary to make the antenna full size, whether it is an HVD, a grounded monopole antenna or a ground-plane type of monopole antenna. The size of the antenna can be reduced by half or even more and still retain high efficiency and the desired radiation pattern. This requires careful design, however. And if high efficiency is maintained, the operating bandwidth of the shortened antenna will be reduced because the shortened antenna will have a higher Q.

This translates into a more rapid increase of reactance away from resonance. The effect can be mitigated to some extent by using larger-diameter conductors. Even doing this however, bandwidth will be a problem, particularly on the 3.5 to 4-MHz band, which is very wide in proportion to the center frequency.

If we take a vertical monopole with a diameter of 2 inches and a frequency of 3.525 MHz and progressively shorten it from λ/4 in length, the feed-point impedance and efficiency (using an inductor at the base to tune out the capacitive reactance) will vary as shown in **Table 7**. In this example perfect ground and conductor are assumed. Real ground will not make a great difference in the impedance but will introduce ground loss, which will reduce the efficiency further. Conductor loss will also reduce efficiency. In general, higher R_R will result in better efficiency.

The important point of Table 7 is the drastic reduction in radiation resistance R_R as the antenna gets shorter. This combined with the increasing loss resistance of the inductor (R_L) used to tune out the increasing base reactance (X_C) reduces the efficiency.

BASE LOADING A SHORT VERTICAL ANTENNA

The base of the antenna is a convenient point at which to add a loading inductor, but it is usually not the lowest loss point at which an inductor, of a given Q, could be placed. There is an extensive discussion of the optimum location of the loading in a short vertical as a function of ground loss and inductor Q in Chapter 16 for mobile antennas, which by necessity are electrically and physically short. This information should be reviewed before using inductive loading.

On the accompanying CD-ROM is a copy of the program *MOBILE.EXE*. This is an excellent tool for designing short, inductively loaded antennas. In most cases, where top loading (discussed below) is not used, the optimum point is near or a little above the middle of the vertical section. Moving the loading coil from the base to the middle of the vertical antenna can make an important difference, increasing R_R and reducing the inductor loss. For example, in an antenna operating at 3.525 MHz, if we make L_1 = 34.9 feet (0.125 λ) the amount of loading inductor placed at the center is 25.2 µH. This resonates the antenna. In this configuration R_R will increase from 6.8 Ω (base loading) to 13.5 Ω (center loading). This substantially increases the efficiency of the antenna, depending on the ground loss and conductor resistances.

Instead of a lumped inductance being inserted at some point in the antenna, it is also possible to use "continuous loading," where the entire radiator is wound as a small diameter coil. The effect is to distribute the inductive loading all along the radiator. In this version of inductive loading the coil is the radiator. An example of a short vertical using this principle is given later in this chapter.

OTHER WAYS OF LOADING A SHORT ANTENNA FOR RESONANCE

Inductive loading is not the only, or even the best, way to compensate for reduced antenna height. *Capacitive top loading* can also be used as indicated in **Fig 52** to bring a vertical monopole to resonance. **Table 8** gives information on a shortened 3.525-MHz vertical using top loading. The vertical portion (L_1) is made from 2-inch tubing. The top loading is also 2-inch tubing extending across the top like a T. The length of the top loading T (±L_2) is adjusted to resonate the

Table 7

Effect of Shortening a Vertical Radiator Below λ/4 Using Inductive Base Loading.

Frequency is 3.525 MHz and for the Inductor Q_L = 200. Ground and conductor losses are omitted.

Length (feet)	*Length (λ)*	*R_R (Ω)*	*X_C (Ω)*	*R_L (Ω)*	*Efficiency (%)*	*Loss (dB)*
14	0.050	0.96	–761	3.8	20	–7.0
20.9	0.075	2.2	–533	2.7	45	–3.5
27.9	0.100	4.2	–395	2.0	68	–1.7
34.9	0.125	6.8	–298	1.5	82	–0.86
41.9	0.150	10.4	–220	1.1	90	–0.44
48.9	0.175	15.1	–153	0.77	95	–0.22
55.8	0.200	21.4	–92	0.46	98	–0.09
62.8	0.225	29.7	–34	0.17	99	–0.02

Table 8

Effect of Shortening a Vertical Using Top Loading

L_1 (feet)	*L_2 (feet)*	*Length (λ)*	*R_R (Ω)*
14.0	48.8	0.050	4.0
20.9	38.6	0.075	8.5
27.9	30.1	0.100	14.0
34.9	22.8	0.125	19.9
41.9	17.3	0.150	25.5
48.9	11.9	0.175	30.4
55.8	7.0	0.200	33.9
62.8	2.4	0.225	35.7

antenna. Again, the ground and the conductors are assumed to be perfect in Table 8.

For a given vertical height, resonating the antenna with top loading results in much higher radiation resistance R_R—2 to 4 times. In addition, the loss associated with the loading element will be smaller. The result is a more efficient antenna for low heights. A comparison of R_R for both capacitive top loading and inductive base loading is given in **Fig 53**. For heights below 0.15 λ the length of the top-loading elements becomes impractical but there are other, potentially more useful, top-loading schemes.

A multiwire system such as the one shown in **Fig 54** has more capacitance than the single-conductor arrangement, and thus does not need to be as long to resonate at a given frequency. This design does, however, require extra supports for the additional wires. Ideally, an arrangement of this sort should be in the form of a cross, but parallel wires separated by several feet give a considerable increase in capacitance

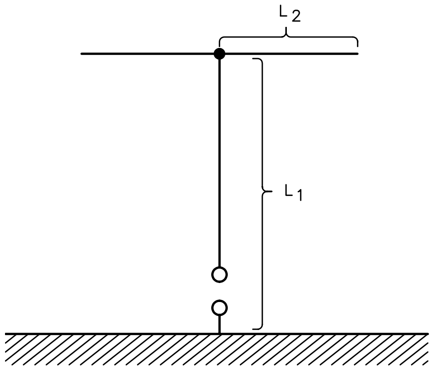

Fig 52—Horizontal wire used to top load a short vertical.

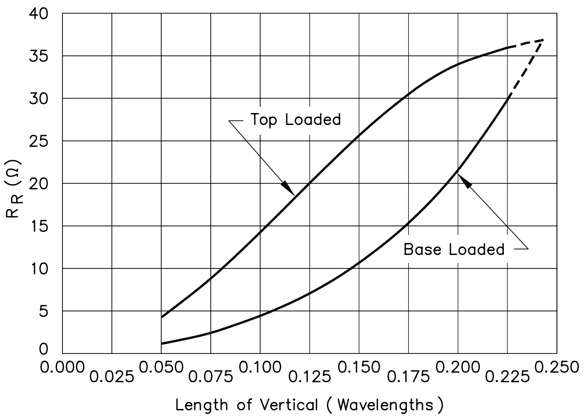

Fig 53—Comparison of top (capacitive) and base (inductive) loading for short verticals. Sufficient loading is used to resonate the antenna.

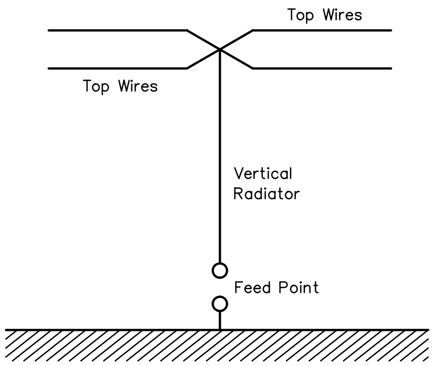

Fig 54—Multiple top wires can increase the effective capacitance substantially. This allows the use of shorter top wires to achieve resonance.

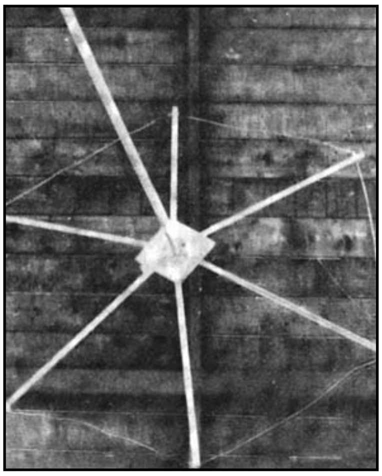

Fig 55—A close-up view of the capacitance hat for a 7-MHz vertical antenna. The ½-in. diameter radial arms terminate in a loop of copper wire.

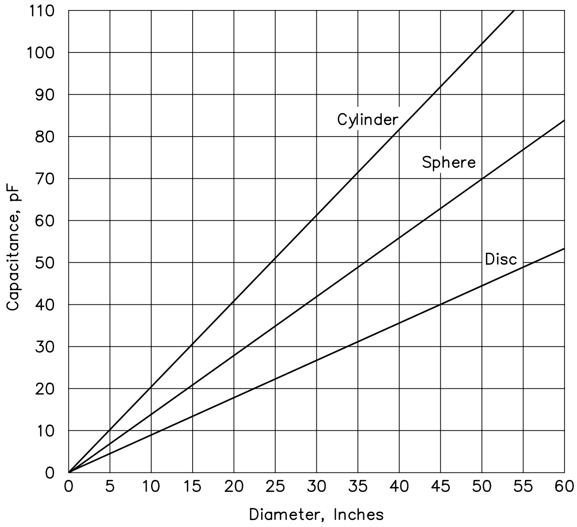

Fig 56—Capacitance of sphere, disc and cylinder as a function of their diameters. The cylinder length is assumed equal to its diameter.

over a single wire.

The top loading can be supplied by a variety of metallic structures large enough to have the necessary self-capacitance. For example, as shown in **Fig 55**, a multi-spoked structure with the ends connected together can be used. One simple way to make a capacitance hat is to take four to six 8-foot fiberglass CB mobile whips, arrange them like spokes in a wagon wheel and connect the ends with a peripheral wire. This arrangement will produce a 16-foot diameter hat that is economical and very durable, even when loaded with ice. Practically any sufficiently large metallic structure can be used for this purpose, but simple geometric forms such as the sphere, cylinder and disc are preferred because of the relative ease with which their capacitance can be calculated.

The capacitance of three geometric forms can be estimated from the curves of **Fig 56** as a function of their size. For the cylinder, the length is specified equal to the diameter. The sphere, disc and cylinder can be constructed from sheet metal, if such construction is feasible, but the capacitance will be practically the same in each if a "skeleton" type of construction with screening or networks of wire or tubing are used.

Finding Capacitance Hat Size

The required size of a capacitance hat may be determined from the following procedure. The information in this section is based on a September 1978 *QST* article by Walter Schulz, K3OQF. The physical length of a shortened antenna can be found from:

$$h_{\text{inches}} = \frac{11808}{F_{\text{MHz}}} \qquad \text{(Eq 3)}$$

where h = length in inches.

Thus, using an example of 7 MHz and a shortened length of 0.167 λ, h = 11808/7 × 0.167 = 282 inches, equivalent to 23.48 feet.

Consider the vertical radiator as an open-ended transmission line, so the impedance and top loading may be determined. The characteristic impedance of a vertical antenna can be found from

$$Z_0 = 60 \left(\ln\left(\frac{4h}{d}\right) - 1 \right) \qquad \text{(Eq 4)}$$

where

ln = natural logarithm

h = length (height) of vertical radiator in inches (as above)

d = diameter of radiator in inches

The vertical radiator for this example has a diameter of 1 inch. Thus, for this example,

$$Z_0 = 60 \left(\ln\left(\frac{4 \times 281}{1}\right) - 1 \right) = 361\ \Omega$$

The capacitive reactance required for the amount of top loading can be found from

$$Z_0 = 60\ \ln\left(\frac{4 \times 281}{1}\right) - 1 = 361\ \Omega \qquad \text{(Eq 5)}$$

where

X = capacitive reactance, ohms

Z_0 = characteristic impedance of antenna (from Eq 4)

θ = amount of electrical loading, degrees.

This value for a 30° hat is 361/tan 30° = 625 Ω. This capacitive reactance may be converted to capacitance with the following equation,

$$C = \frac{10^6}{2\pi f X_C} \qquad \text{(Eq 6)}$$

where

C = capacitance in pF

f = frequency, MHz

X_C = capacitive reactance, ohms (from above).

For this example, the required C = $10^6/(2\ \pi \times 7 \times 625)$ = 36.4 pF, which may be rounded to 36 pF. A disc capacitor is used in this example. The appropriate diameter for 36 pF of hat capacitance can be found from Fig 56. The disc diameter that yields 36 pF of capacitance is 40 inches.

The skeleton disc shown in Fig 55 is fashioned into a wagonwheel configuration. Six 20-inch lengths of ½-inch OD aluminum tubing are used as spokes. Each is connected to the hub at equidistant intervals. The outer ends of the spokes terminate in a loop made of #14 copper wire. Note that the loop increases the hat capacitance slightly, making a better approximation of a solid disc. The addition of this hat at the top of a 23.4-foot radiator makes it quarter-wave resonant at 7 MHz.

After construction, some slight adjustment in the radiator length or the hat size may be required if resonance at a specific frequency is desired. From Fig 53, the radiation resistance of a 0.167-λ high radiator is seen to be about 13 Ω without top loading. With top loading Rr ≈ 25 Ω or almost double.

THE COMPACT VERTICAL DIPOLE

A variation on the HVD (half-wave vertical dipole) theme is the *compact vertical dipole*, or CVD. The CVD uses capacitance-hat loading on each end of a shortened vertical radiator, as shown in **Fig 57C**. Some call this "top hat" and "bottom hat" loading. Les Moxon, G6XN, called this method of loading a shortened antenna simply "end loading." The vertical wire for his 40-meter CVD is 25 feet high, with 15-foot long horizontal loading wires on each side, top and bottom.

K8CH Compact Vertical Dipole

The top loading wires needn't be perpendicular to the vertical radiator, although that is convenient if you construct the antenna from aluminum tubing. K8CH described a wire 30-meter CVD in the 2006 Edition of *The ARRL Handbook.* This used #14 wire throughout, with sloping top loading wires. It was designed to be suspended from a tree at least

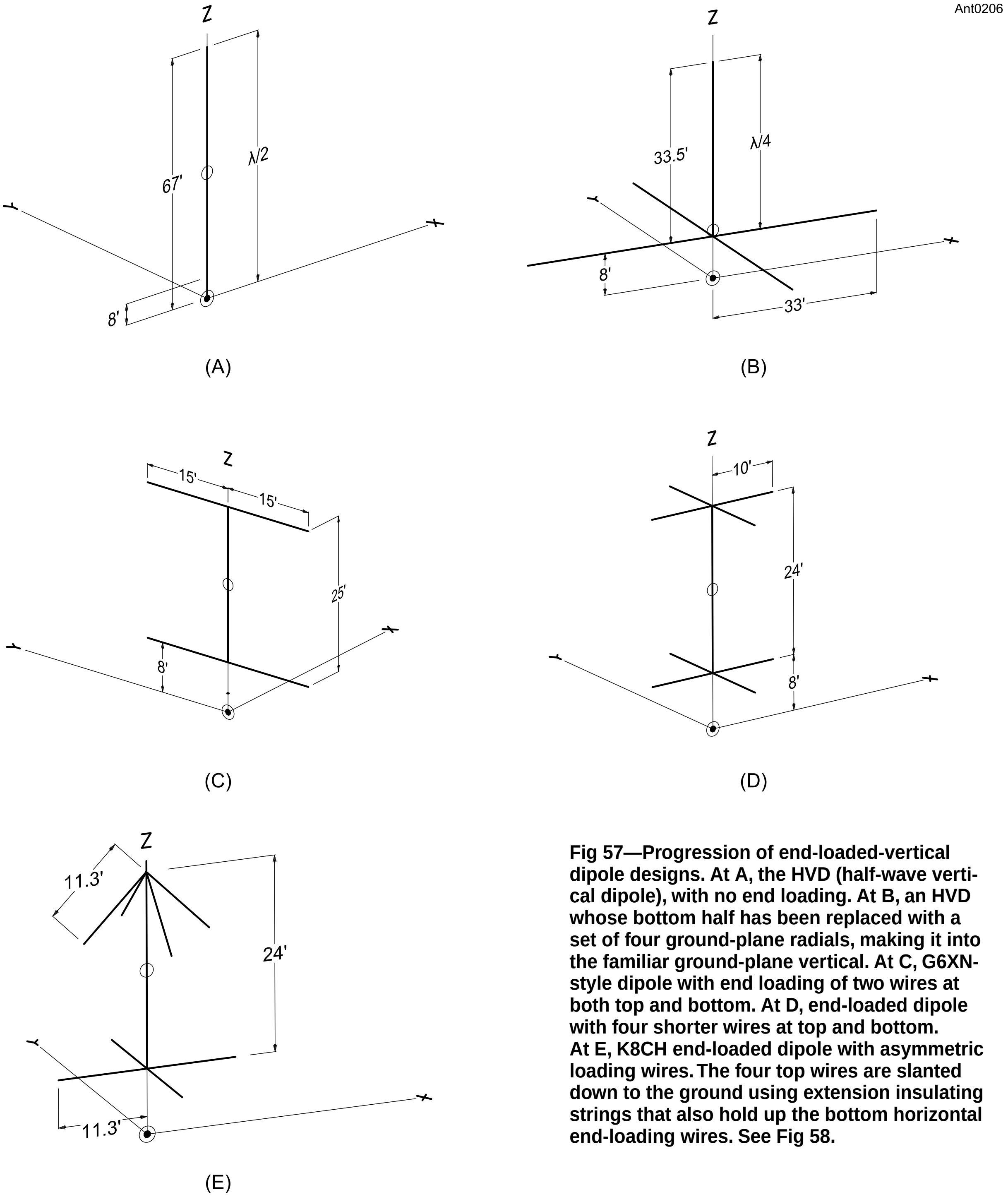

Fig 57—Progression of end-loaded-vertical dipole designs. At A, the HVD (half-wave vertical dipole), with no end loading. At B, an HVD whose bottom half has been replaced with a set of four ground-plane radials, making it into the familiar ground-plane vertical. At C, G6XN-style dipole with end loading of two wires at both top and bottom. At D, end-loaded dipole with four shorter wires at top and bottom.
At E, K8CH end-loaded dipole with asymmetric loading wires. The four top wires are slanted down to the ground using extension insulating strings that also hold up the bottom horizontal end-loading wires. See Fig 58.

32 feet high to keep all wires 8 feet or more above humans and animals. The vertical radiating wire is 24 feet long, and the eight top and bottom end-loading wires are all 5 feet 9 inches long. See Fig 57E and **Table 9**, CVD 1.

The top loading wires slant at an angle of 45º down to the ground, using insulating strings that also support the ends of the bottom loading wires, holding them out so that they are horizontal. See **Fig 58**. There is a small loss of gain because of the "umbrella" shape of the top loading wires, and the 2:1 SWR bandwidth is diminished slightly from the horizontal case. This isn't a problem on a narrow band like 30 meters.

A 40-meter version of this antenna, CVD 6, also uses a 24-foot long vertical radiator, 8 feet high at its bottom. It uses 11.3-foot long radial wires made of #14 wire, again with the top four slanted down at 45º. This CVD has a 2:1

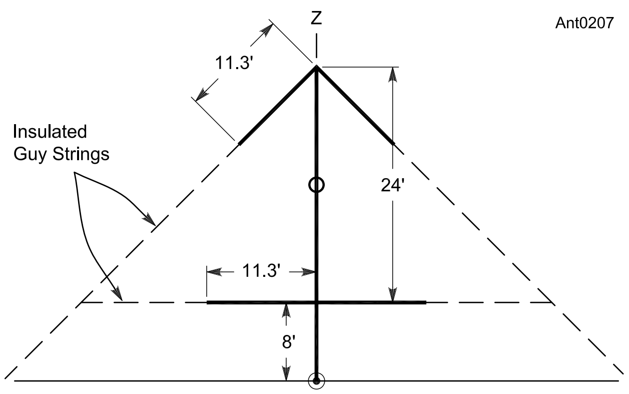

Fig 58—Layout of CVD made using #14 wire suspended from a tree branch.

SWR bandwidth of 250 kHz. It is directly fed with 50-Ω coax with ferrite-core common-mode choke baluns in the middle of the vertical radiator. An additional choke balun is used where the coax reaches ground level in order to knock down common-mode currents that might otherwise radiate onto the coax shield.

You should note the comparison of 40-meter ground-plane vertical antennas in Table 9. GP 1 is for horizontal radials that are 8 feet off the ground, while GP 2 has its radials only 2 feet high. There is some loss in gain because of the proximity of the lossy ground. The 40-meter CVD 1 and CVD 2 cases illustrate the same effect of being close to the lossy ground.

Some of the cases in Table 9 require Center Load coils to bring the antenna to resonance. Where the loading coil inductance is equal to the "Hairpin Coil" inductance, the loading coil also serves as a hairpin matching coil. Where the amount of Hairpin Coil inductance is less than the Load Coil inductance, a match is achieved by tapping the Center Load Coil symmetrically out from the center.

80-Meter CVD

The size of a CVD becomes a real challenge on 80 meters, requiring either very tall support structures or multiple loading methods to keep the vertical radiator to a reasonable length. The CVD 2 design in Table 9 shows a K8CH-style CVD wire antenna whose vertical radiator is 46.5 feet long. It requires a 54.5-foot high tree to keep the bottom end of the vertical radiator 8 feet above ground for safety. Compare this to an HVD that requires a 143-foot support of some sort to keep it 8 feet off the ground at the bottom. The CVD 2 sacrifices some 0.7 dB in gain for this difference in size, and about 75 kHz in 2:1 SWR bandwidth.

The CVD 2 would require retuning when going from

Table 9
Variations on a Vertical Center-Fed Dipole

Name	*Style Fig 57*	*Vertical Length feet*	*Spoke Length feet*	*Min. Ht feet*	*Max. Gain dBi*	*2:1 SWR kHz*	*Hairpin Coil µH*	*Center Load µH*
20 Meters								
GP	B	17.53	16.53	8	0.29	400	—	—
CVD 1	C	13	7.57	8	0.12	625	—	—
CVD 2	D	12	5.1	8	0.00	550	0.68	0.68
CVD 3	E	12.15	5.6	8	–0.01	450	0.5	0.5
30 Meters								
GP	B	24.54	23.14	8	0.04	400	—	—
CVD 1	E	24	5.33	8	–0.2	500	—	—
CVD 2	E	17	7.60	8	–0.36	400	0.82	0.82
40 Meters								
HVD	A	66	—	8	0.13	450	—	—
GP 1	B	35	33	8	–0.12	325	—	—
GP 2	B	34.5	33	2	–0.37	400	—	—
CVD 1	C	25	15	8	–0.42	450	—	—
CVD 2	C	24	15	2	–1.09	400	—	—
CVD 3	D	24	10	8	–0.55	425	—	0.25
CVD 4	D	24	5	8	–0.85	225	—	8.7
CVD 5	D	16	8	8	–1.18	175	0.94	6.8
CVD 6	E	24	11.3	8	–0.59	250	—	—
80 Meters								
HVD	A	135	—	8	0.19	225	—	—
GP	B	65.5	61	8	0.11	200	—	—
CVD 1	C	48.3	30	8	–0.27	200	—	—
CVD 2	E	46.5	21.9	8	–0.50	150	—	—

CW to phone, probably by changing the length of the four bottom horizontal wires equally.

LINEAR LOADING

Another alternative to inductive loading is *linear loading*. This little-understood method of shortening radiators can be applied to almost any antenna configuration—including parasitic arrays. Although commercial antenna manufacturers make use of linear loading in their HF antennas, relatively few hams have used it in their own designs. Linear loading can be used to advantage in many antennas because it introduces relatively little loss, does not degrade directivity patterns, and has low enough Q to allow reasonably good bandwidth. Some examples of linear-loaded antennas are shown in **Fig 59**.

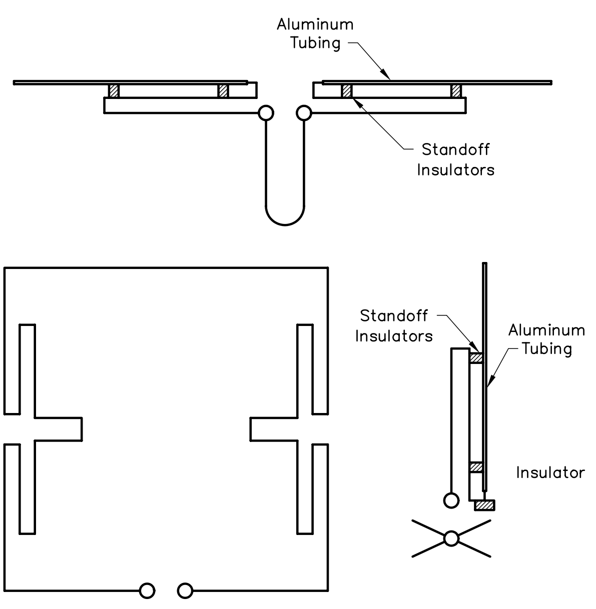

Fig 59—Some examples of linear loading. The small circles indicate the feed points of the antennas.

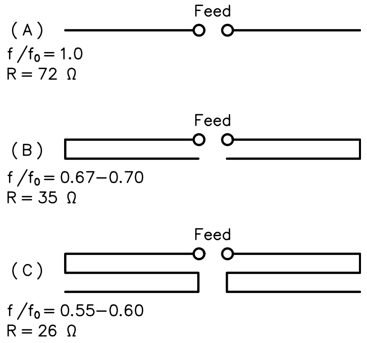

Fig 60—Wire dipole antennas. The ratio f/f_0 is the measured resonant frequency divided by frequency f_0 of a standard dipole of same length. R is radiation resistance in ohms. At A, standard single-wire dipole. At B, two-wire linear-loaded dipole, similar to folded dipole except that side opposite feed line is open. At C, three-wire linear-loaded dipole.

Since the dimensions and spacing of linear-loading devices vary greatly from one antenna installation to another, the best way to employ this technique is to try a length of conductor 10% to 20% longer than the difference between the shortened antenna and the full-size dimension for the linear-loading device. Then use the "cut-and-try" method, varying both the spacing and length of the loading device to optimize the match. A hairpin at the feed point can be useful in achieving a 1:1 SWR at resonance.

Linear-Loaded Short Wire Antennas

More detail on linear loading is provided in this section, which was originally presented in *The ARRL Antenna Compendium Vol 5* by John Stanford, NNØF. Linear loading can significantly reduce the required length for resonant antennas. For example, it is easy to make a resonant antenna that is as much as 30 to 40% shorter than an ordinary dipole for a given band. The shorter overall lengths come from bending back some of the wire. The increased self-coupling lowers the resonant frequency. These ideas are applicable to short antennas for restricted space or portable use.

Experiments

The results of the measurements are shown in **Fig 60** and are also consistent with values given by Rashed and Tai from an earlier paper. This shows several simple wire antenna configurations, with resonant frequencies and impedance (radiation resistance). The reference dipole has a resonant frequency f_0 and resistance R = 72 Ω. The f/f_0 values give the effective reduced frequency obtained with the linear loading in each case. For example, the two-wire linear-loaded dipole has its resonant frequency lowered to about 0.67 to 0.70 that of the simple reference dipole of the same length.

The three-wire linear-loaded dipole has its frequency reduced to 0.55 to 0.60 of the simple dipole of the same length. As you will see later, these values will vary with conductor diameter and spacing.

The two-wire linear-loaded dipole (Fig 60B) looks almost like a folded dipole but, unlike a folded dipole, it is open in the middle of the side opposite where the feed line is attached. Measurements show that this antenna structure has a resonant frequency lowered to about two-thirds that of the reference dipole, and R equal to about 35 Ω. A three-wire linear-loaded dipole (Fig 60C) has even lower resonant frequency and R about 25 to 30 Ω.

Linear-loaded monopoles (one half of the dipoles in Fig 60) working against a radial ground plane have similar resonant frequencies, but with only half the radiation resistance shown for the dipoles.

A Ladder-Line Linear-Loaded Dipole

Based on these results, NNØF next constructed a linear loaded dipole as in Fig 60B, using 24 feet of 1-inch ladder line (the black, 450-Ω plastic kind widely available) for the dipole length. He hung the system from a tree using nylon fishing line, about 4 feet from the tree at the top, and about 8 feet from the ground on the bottom end. It was slanted at about a 60° angle to the ground. This antenna resonated at 12.8 MHz and had a measured resistance of about 35 Ω. After the resonance measurements, he fed it with 1-inch ladder open-wire line (a total of about 100 feet to the shack).

For brevity, this is called a vertical *LLSD* (linear-loaded short dipole). A tuner resonated the system nicely on 20 and 30 meters. On these bands the performance of the vertical LLSD seemed comparable to his 120-foot long, horizontal center-fed Zepp, 30 feet above ground. In some directions where the horizontal, all-band Zepp has nulls, such as toward Siberia, the vertical LLSD was definitely superior. This system also resonates on 17 and 40 meters. However, from listening to various signals, NN0F had the impression that this length LLSD is not as good on 17 and 40 meters as the horizontal 120-foot antenna.

Using Capacitance End Hats

He also experimented with an even shorter resonant length by trying an LLSD with capacitance end-hats. The hats, as expected, increased the radiation resistance and lowered the resonant frequency. Six-foot long, single-wire hats were used on each end of the previous 24-foot LLSD, as shown in **Fig 61**. The antenna was supported in the same way as the previous vertical dipole, but the bottom-end hat wire was only inches from the grass. This system resonated at 10.6 MHz with a measured resistance of 50 Ω.

If the dipole section were lengthened slightly, by a foot or so, to about 25 feet, it should hit the 10.1-MHz band and be a good match for 50-Ω coax. It would be suitable for a restricted space, shortened 30-meter antenna. Note that this antenna is only about half the length of a conventional 30-meter dipole, needs no tuner, and has no losses due to traps. It does have the loss of the extra wire, but this is essentially negligible.

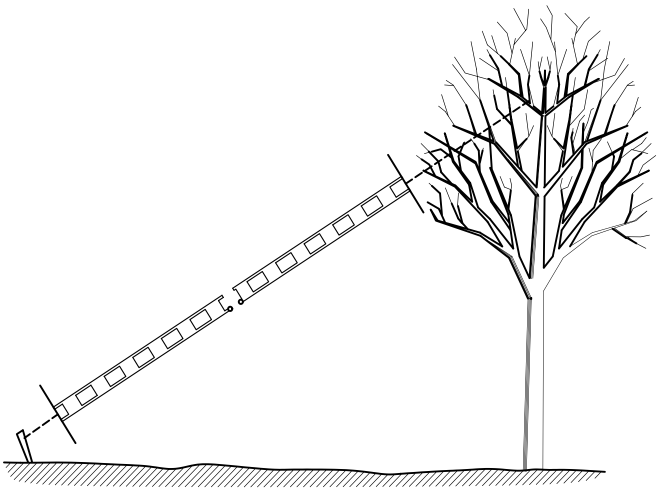

Fig 61—Two-wire linear-loaded dipole with capacitance end hats. Main dipole length was constructed from 24 feet of "windowed" ladder line. The end-hat elements were stiff wires 6 feet long. The antenna was strung at about a 60° angle from a tree limb using monofilament fishing line. Measured resonant frequency and radiation resistance were 10.6 MHz and 50 Ω.

Any of the linear-loaded dipole antennas can be mounted either horizontally or vertically. The vertical version can be used for longer skip contacts—beyond 600 miles or so—unless you have rather tall supports for horizontal antennas to give a low elevation angle. Using different diameter conductors in linear-loaded antenna configurations yields different results, depending on whether the larger or small diameter conductor is fed. NN0F experimented with a vertical ground-plane antenna using a 10-foot piece of electrical conduit pipe (5⁄8 inch OD) and #12 copper house wire.

Fig 62 shows the configuration. The radial ground system was buried a couple of inches under the soil and is not shown. Note that this is not a folded monopole, which would have either A or B grounded.

The two conductors were separated by 2 inches, using plastic spreaders held onto the pipe by stainless-steel hose clamps obtained from the local hardware store. Hose clamps intertwined at right angles were also used to clamp the pipe on electric fence stand-off insulators on a short 2×4 post set vertically in the ground.

The two different diameter conductors make the antenna characteristics change, depending on how they are configured. With the antenna bridge connected to the larger diameter conductor (point A in Fig 62), and point B unconnected, the system resonated at 16.8 MHz and had R = 35 Ω. With the bridge at B (the smaller conductor), and point A left unconnected, the resonance lowered to 12.4 MHz and R was found to be about 24 Ω.

The resonant frequency of the system in Fig 62 can be adjusted by changing the overall height, or for increasing the frequency, by reducing the length of the wire. Note that a 3.8-MHz resonant ground plane can be made with height only about half that of the usual 67 feet required, if the smaller conductor is fed (point B in Fig 62). In this case, the pipe would be left unconnected electrically. The lengths given above can be scaled to determine a first-try attempt for your favorite band. Resonant lengths will, however, depend on the conductor diameters and spacing.

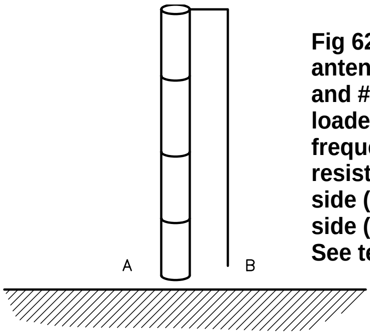

Fig 62—Vertical ground-plane antenna with a 10-foot pipe and #12 wire as the linear-loaded element. Resonant frequency and radiation resistance depend on which side (A or B) is fed. The other side (B or A) is *not* grounded. See text for details.

The same ideas hold for a dipole, except that the lengths should be doubled from those of the ground plane in Fig 62. The resistance will be twice that of the ground plane. Say, how about a shortened 40-meter horizontal beam to enhance your signal?!

COMBINED LOADING

As an antenna is shortened further the size of the top loading device will become larger and at some point will be impractical. In this situation inductive loading, usually placed directly between the capacitance "hat" and the top of the antenna, can be added to resonate the antenna. An alternative would be to use linear loading in place of inductive loading. The previous section contained an example of end loading combined with linear loading.

SHORTENING THE RADIALS

Very often the space required by full-length radials is simply not available. Like the vertical portion of the antenna, the radials can also be shortened and loaded in very much the same way. An example of end loaded radials is given in **Fig 63A**. Radials half the usual length can be used with little reduction in efficiency but, as in the case of top loading, the antenna Q will be higher and the bandwidth reduced. As shown in Fig 63B**,** inductive loading can also be used. As long as they are not made too short (down to 0.1λ) loaded radials can be efficient—with careful design.

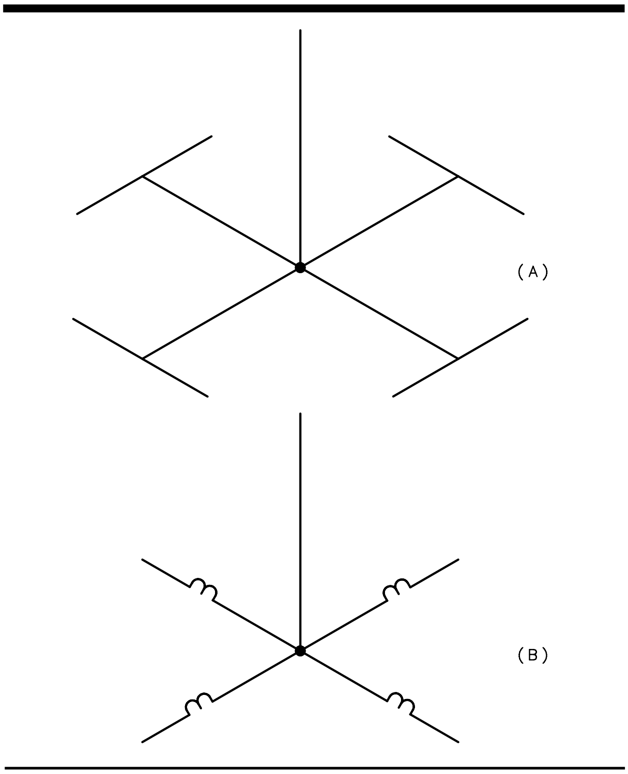

Fig 63—Radials may be shortened by using either capacitive (A) or inductive (B) loading. In extreme cases both may be used but the operating bandwidth will be limited.

GENERAL RULES

The steps in designing an efficient short vertical antenna system are:

- Make the vertical section as long as possible.
- Make the diameter of the vertical section as large as possible. Tubing or a cage of smaller wires will work well.
- Provide as much top and/or bottom loading as possible.
- If the top/bottom loading is insufficient, resonate the antenna with a high-Q inductor placed between the hat and the top of the antenna.
- For buried-ground systems, use as many radials (> 0.2 λ) as possible. 40 or more is best.
- If an elevated ground plane is used, use 4 to 8 radials, 5 or more feet above ground.
- If shortened radials must be used then capacitive loading is preferable to inductive loading.

EXAMPLES OF SHORT VERTICALS

A 6-Foot-High 7-MHz Vertical Antenna

Figs 64 through 67 give details for building short, effective vertical quarter-wavelength radiators. This information was originally presented by Jerry Sevick, W2FMI.

A short vertical antenna, properly designed and installed, approaches the efficiency of a full-size resonant quarter-wave antenna. Even a 6-foot vertical on 7 MHz can produce an exceptional signal. Theory tells us that this should be possible, but the practical achievement of such a result requires an understanding of the problems of ground losses, loading, and impedance matching.

The key to success with shortened vertical antennas lies in the efficiency of the ground system with which the antenna is used. A system of at least 60 radial wires is rec-

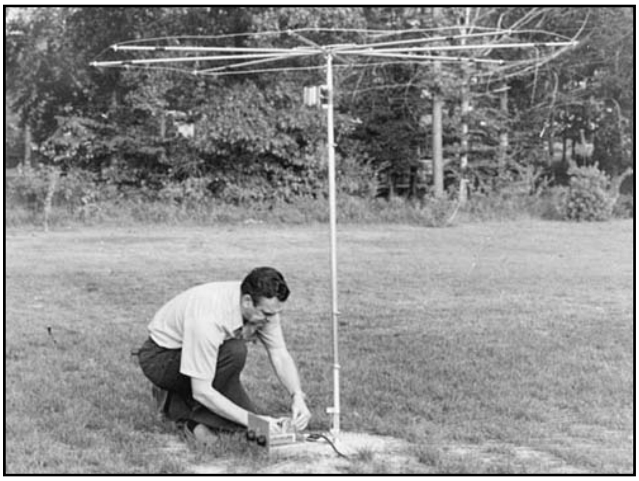

Fig 64—Jerry Sevick, W2FMI, adjusts the 6-foot high, 40-meter vertical.

ommended for best results, although the builder may elect to reduce the number at the expense of some performance. The radials can be tensioned and pinned at the far ends to permit on-the-ground installation, which will enable the amateur to mow the lawn without the wires becoming entangled in the mower blades. Alternatively, the wires can be buried in the ground, where they will not be visible. There is nothing critical about the wire size for the radials. Radials made of 28, 22, or even 16-gauge wire, will provide the same results. The radials should be at least 0.2 λ long (27 feet or greater on 7 MHz).

A top hat is formed as illustrated in **Fig 65**. The diameter is 7 feet, and a continuous length of wire is connected to the spokes around the outer circumference of the wheel. A loading coil consisting of 14 turns of B&W 3029 Miniductor stock (2½-inch dia, 6 TPI, #12 wire) is installed 6 inches below the top hat (see Fig 65). This antenna exhibits a feed-point impedance of 3.5 Ω at 7.21 MHz. For operation above or below this frequency, the number of coil turns must be decreased or increased, respectively. Matching is accomplished by increasing the feed-point impedance to 14 Ω through addition of a 4:1-transformer, then matching 14 Ω to 50 Ω (feeder impedance) by means of a pi network. The 2:1 SWR bandwidth for this antenna is approximately 100 kHz.

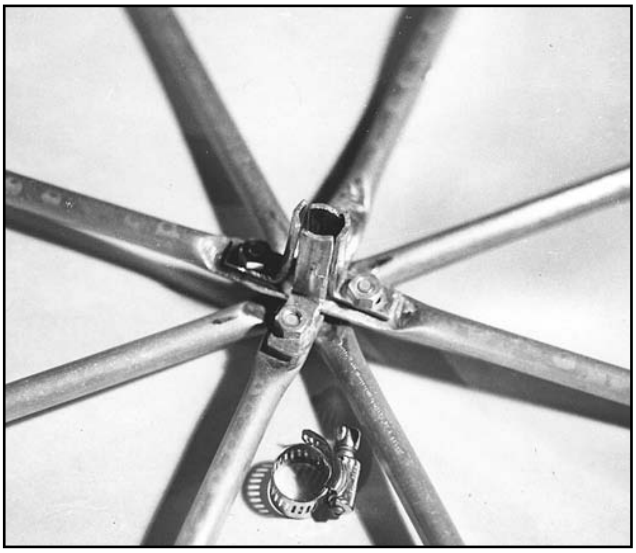

Fig 65—Construction details for the top hat. For a diameter of 7 feet, ½-in. aluminum tubing is used. The hose clamp is made of stainless steel and is available at Sears. The rest of the hardware is aluminum.

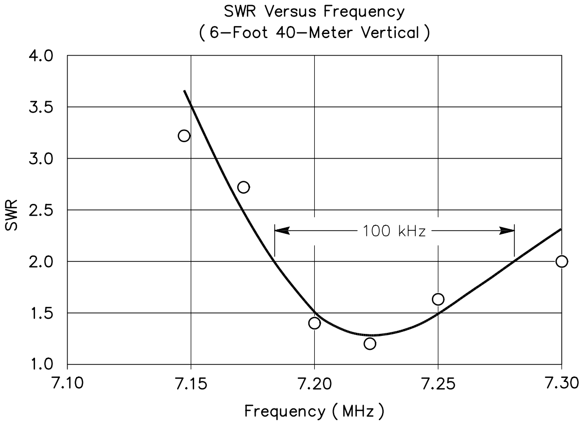

Fig 66—Standing-wave ratio of the 6-foot vertical using a 7-foot top hat and 14 turns of loading 6 inches below the top hat.

More than 200 contacts with the 6-foot antenna have indicated the efficiency and capability of a short vertical. Invariably at distances greater than 500 or 600 miles, the short vertical yields excellent signals. Similar antennas can be scaled and constructed for bands other than 7 MHz. The 7-foot-diameter-top hat was tried on a 3.5-MHz vertical, with an antenna height of 22 feet. The loading coil had 24 turns and was placed 2 feet below the top hat. On-the-air results duplicated those on 40-meters. The bandwidth was 65 kHz.

Short verticals such as these have the ability to radiate and receive almost as well as a full-size quarter-wave. Trade-offs are in lowered input impedances and bandwidths. However, with a good radial system and a proper design, these trade-offs can be made entirely acceptable.

Short Continuously Loaded Verticals

While there is the option of using lumped inductance to achieve resonance in a short antenna, the antenna can also be helically wound to provide the required inductance. This is shown in **Fig 68**. Shortened quarter-wavelength vertical antennas can be made by forming a helix on a long cylindrical insulator. The diameter of the helix must be small in terms of λ to prevent the antenna from radiating in the axial mode.

Acceptable form diameters for HF-band operation are from 1 inch to 10 inches when the practical aspects of antenna construction are considered. Insulating poles of fiberglass, PVC tubing, treated bamboo or wood, or phenolic are suitable for use in building helically wound radiators. If wood or bamboo is used the builder should treat the material with at least two coats of exterior spar varnish prior to winding the antenna element. The completed structure should be given two more coats of varnish, regardless of the material used for the coil form. Application of the varnish will help weatherproof the antenna and prevent the coil turns from changing position.

No strict rule has been established concerning how short a helically wound vertical can be before a significant drop in performance is experienced. Generally, one should use the greatest amount of length consistent with available space. A guideline might be to maintain an element length of 0.05 wavelength or more for antennas which are electrically a quarter wavelength long. Thus, use 13 feet or more of stock for an 80-meter antenna, 7 feet for 40 meters, and so on.

A quarter-wavelength helically wound vertical can be used in the same manner as a full-size vertical. That is, it can be worked against an above-ground wire radial system (four or more radials), or it can be ground-mounted with

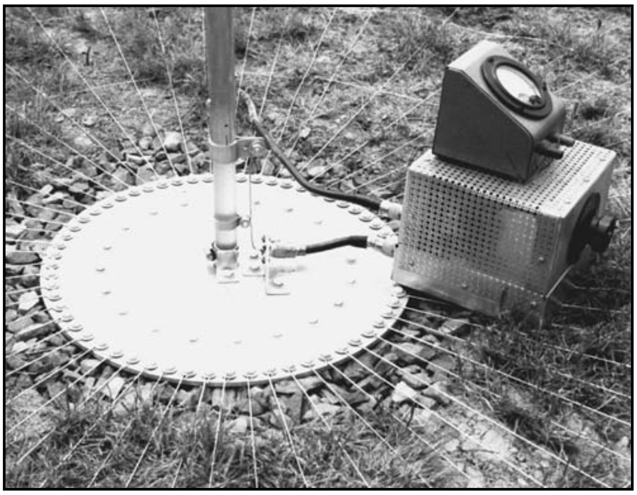

Fig 67—Base of the vertical antenna showing the 60 radial wires. The aluminum disc is 15 inches in diameter and ¼ inches thick. Sixty tapped holes for ¼-20 aluminum hex-head bolts form the outer ring and 20 form the inner ring. The inner bolts were used for performance comparisons with more than 60 radials. The insulator is polystyrene material (phenolic or Plexiglas suitable) with a 1-inch diameter. Also shown is the impedance bridge used for measuring input resistance.

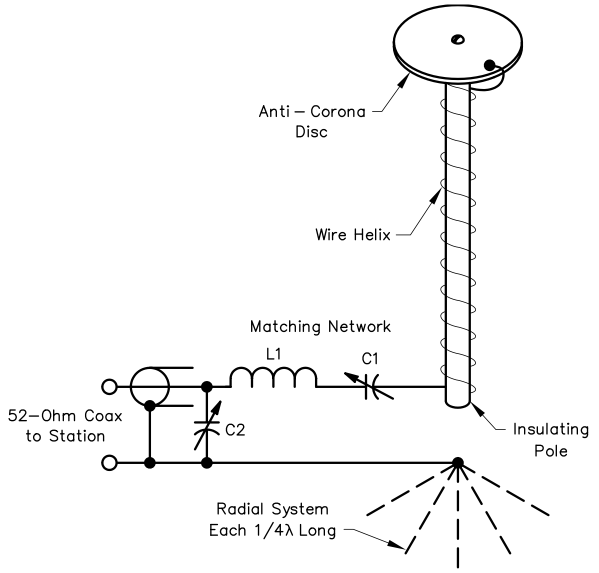

Fig 68—Helically wound ground-plane vertical. Performance from this type of antenna is comparable to that of many full-size $\lambda/4$ vertical antennas. The major design trade-off is usable bandwidth. All shortened antennas of this variety are narrow-band devices. At 7 MHz, in the example illustrated here, the bandwidth between the 2:1 SWR points will be on the order of 50 kHz, half that amount on 80 meters, and twice that amount on 20 meters. Therefore, the antenna should be adjusted for operation in the center of the frequency band of interest.

radials buried or lying on the ground. Some operators have reported good results when using antennas of this kind with four helically wound radials cut for resonance at the operating frequency. The latter technique should capture the attention of those persons who must use indoor antennas.

Winding Information

There is no hard-and-fast formula for determining the amount of wire needed to establish resonance in a helical antenna. The relationship between the length of wire needed for resonance and a full quarter wave at the desired frequency depends on several factors. Some of these are wire size, diameter of the turns, and the dielectric properties of the form material, to name a few. Experience has indicated that a section of wire approximately one half wavelength long, wound on an insulating form with a linear pitch (equal spacing between turns) will come close to yielding a resonant quarter wavelength. Therefore, an antenna for use on 160 meters would require approximately 260 feet of wire, spirally wound on the support.

No specific rule exists concerning the size or type of wire one should use in making a helix. Larger wire sizes are, of course, preferable in the interest of minimizing I^2R losses in the system. For power levels up to 1000 W it is wise to use a wire size of #16 or larger. Aluminum clothesline wire is suitable for use in systems where the spacing between turns is greater than the wire diameter. Antennas requiring close-spaced turns can be made from enameled magnet wire or #14 vinyl jacketed, single-conductor house wiring stock. Every effort should be made to keep the turn spacing as large as is practical to maximize efficiency.

A short rod or metal disc should be made for the top or high-impedance end of the vertical. This is a necessary part of the installation to assure reduction in antenna Q. This broadens the bandwidth of the system and helps prevent extremely high amounts of RF voltage from being developed at the top of the radiator. (Some helical antennas act like Tesla coils when used with high-power transmitters, and can actually catch fire at the high-impedance end when a stub or disc is not used.) Since the Q-lowering device exhibits some additional capacitance in the system, it must be in place before the antenna is tuned.

Tuning and Matching

Once the element is wound it should be mounted where it will be used, with the ground system installed. The feed end of the radiator can be connected temporarily to the ground system. Use a dip meter to check the antenna for resonance by coupling the dipper to the last few turns near the ground end of the radiator. Add or remove turns until the vertical is resonant at the desired operating frequency.

It is impossible to predict the absolute value of feed impedance for a helically wound vertical. The value will depend upon the length and diameter of the element, the ground system used with the antenna, and the size of the disc or stub atop the radiator. Generally speaking, the radiation resistance will be very low—approximately 3 to 10 Ω. An L network of the kind shown in Fig 68 can be used to increase the impedance to 50 Ω. The Q_L (loaded Q) of the network inductors is low to provide reasonable bandwidth, consistent with the bandwidth of the antenna. Network values for other operating bands and frequencies can be determined by using the reactance values listed below. The design center for the network is based on a radiation resistance of 5 Ω. If the exact feed impedance is known, the following equations can be used to determine precise component values for the matching network. (See Chapter 25, Coupling the Transmitter to the Line, for additional information on L-network matching.)

XA QRL (Eq 7)

$$X_{C2} = 50\sqrt{\frac{R_L}{50 - R_L}} \quad \text{(Eq 8)}$$

$$X_{L1} = X_{C1} + \frac{R_L\,50}{X_{C2}} \quad \text{(Eq 9)}$$

where

X_{C1} = capacitive reactance of C1
X_{C2} = capacitive reactance of C2
X_{L1} = inductive reactance of L1
Q = loaded Q of network
R_L = radiation resistance of antenna

Example: Find the network constants for a helical antenna with a feed impedance of 5 Ω at 7 MHz, Q = 3:

$$X_{C1} = 3 \times 5 = 15$$

$$X_{C2} = \sqrt{\frac{5}{50 - 5}} = 16.666$$

$$X_{L1} = 15 + \frac{250}{16.666} = 30$$

Therefore, C1 = 1500 pF, C2 = 1350 pF, and L1 = 0.7 μH. The capacitors can be made from parallel or series combinations of transmitting micas. L_1 can be a few turns of large Miniductor stock. At RF power levels of 100 W or less, large compression trimmers can be used at C1 and C2 because the maximum RMS voltage at 100 W (across 50 Ω) will be 50 V. At, say, 800 W there will be approximately 220 V RMS developed across 50 Ω. This suggests the use of small transmitting variables at C1 and C2, possibly connected in parallel with fixed values of capacitance to constitute the required amount of capacitance for the network.

By making some part of the network variable, it will be possible to adjust the circuit for an SWR of 1:1 without knowing precisely what the antenna feed impedance is. Actually, C1 is not required as part of the matching network. It is included here to bring the necessary value for L1 into a practical range.

Fig 68 illustrates the practical form a typical helically wound ground-plane vertical might take. Performance from this type antenna is comparable to that of many full-size quarter-wavelength vertical antennas. The major design trade-off is in usable bandwidth. All shortened antennas of this variety are narrow-band devices. At 7 MHz, in the example illustrated here, the bandwidth between the 2:1 SWR points will be on the order of 50 kHz, half that amount on 80 meters, and twice that amount on 20 meters. Therefore, the antenna should be adjusted for operation in the center of the frequency spread of interest.

SHORTENED DIPOLES

As shown in preceding sections, there are a number of ways to load antennas so they may be reduced in size without severe reductions in effectiveness. Loading is always a compromise; the best method is determined by the amount of space available and the band(s) to be worked.

The simplest way to shorten a dipole is shown in **Fig 69**. If you do not have sufficient length between the supports, simply hang as much of the center of the antenna as possible between the supports and let the ends hang down. The ends can be straight down or may be at an angle as indicated but in either case should be secured so that they do not move in the wind. As long as the center portion between the supports is at least ¼, the radiation pattern will be very nearly the same

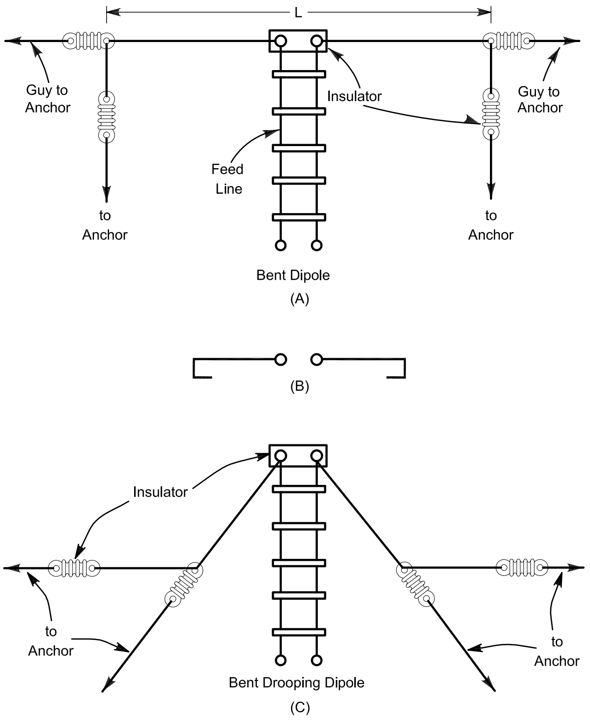

Fig 69—When space is limited, the ends may be bent downward as shown at A, or back on the radiator as shown at B. The bent dipole ends may come straight down or be led off at an angle away from the center of the antenna. An inverted V at C can be erected with the ends bent parallel to the ground when the support structure is not high enough.

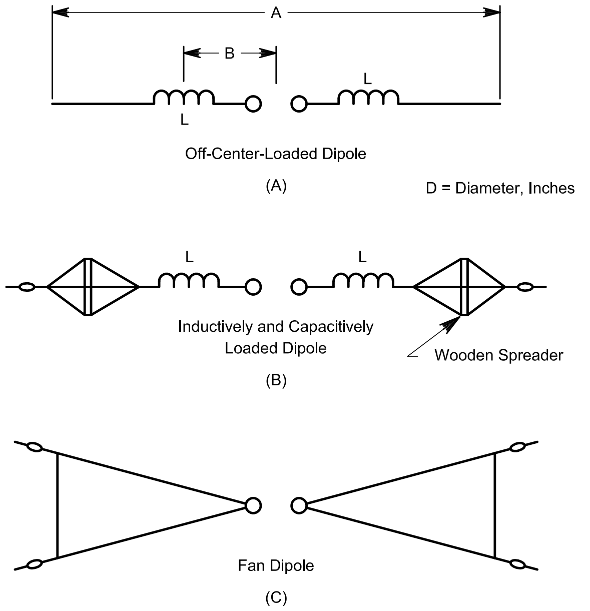

Fig 70—At A is a dipole antenna lengthened electrically with off-center loading coils. For a fixed dimension A, greater efficiency will be realized with greater distance B, but as B is increased, L must be larger in value to maintain resonance. If the two coils are placed at the ends of the antenna, in theory they must be infinite in size to maintain resonance. At B, capacitive loading of the ends, either through proximity of the antenna to other objects or through the addition of capacitance hats, will reduce the required value of the coils. At C, a fan dipole provides some electrical lengthening as well as broadbanding.

as a full-length dipole.

The resonant length of the wire will be somewhat shorter than a full-length dipole and can best be determined by experimentally adjusting the length of ends, which may be conveniently near ground. Keep in mind that there can be very high potentials at the ends of the wires and for safety the ends should be kept out of reach. Letting the ends hang down as shown is a form of capacitive end loading. While it is efficient, it will also reduce the matching bandwidth—as does any form of loading.

The most serious drawback associated with inductive loading is high loss in the coils themselves. It is important that you use inductors made from reasonably large wire or tubing to minimize this problem. Close winding of turns should also be avoided if possible. A good compromise is to use some off-center inductive loading in combination with capacitive end loading, keeping the inductor losses small and the efficiency as high as possible.

Some examples of off-center coil loading and capacitive-end loading are shown in **Fig 70**. This technique was described by Jerry Hall, K1TD in Sep 1974 *QST*. In the equation below, the diameter D is that of the wire used for the antenna, in inches. The frequency f is expressed in MHz.

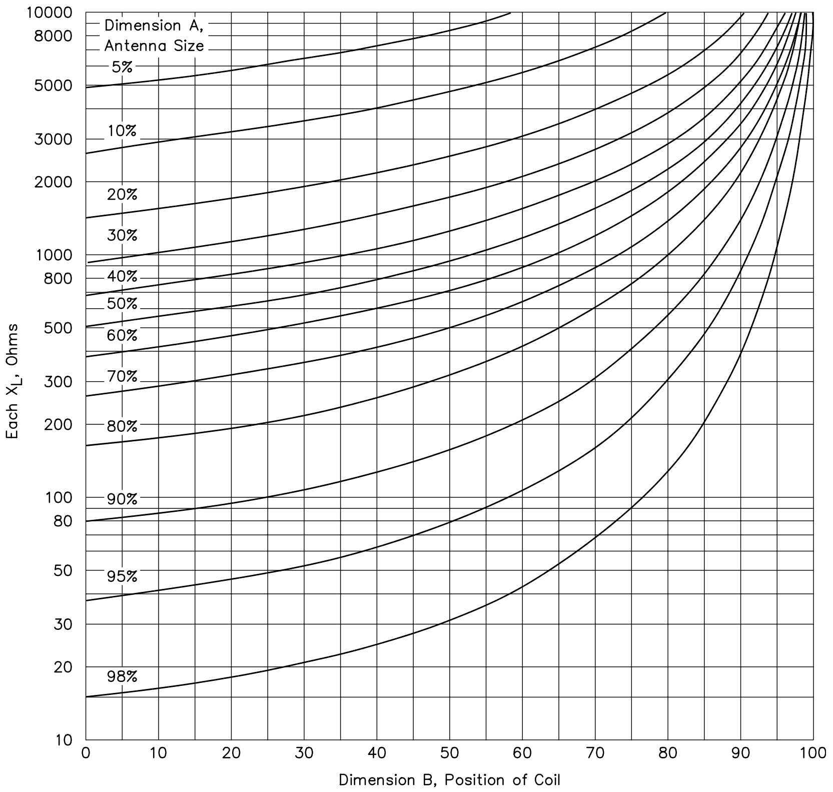

Fig 71—Chart for determining approximate inductance values for off-center-loaded dipoles. See Fig 70A. At the intersection of the appropriate curve from the body of the chart for dimension A and proper value for the coil position from the horizontal scale at the bottom of the chart, read the required inductive reactance for resonance from the scale at the left. Dimension A is expressed as percent length of the shortened antenna with respect to the length of a half-wave dipole of the same conductor material. Dimension B is expressed as the percentage of coil distance from the feed point to the end of the antenna. For example, a shortened antenna, which is 50% or half the size of a half-wave dipole (one-quarter wavelength overall) with loading coils positioned midway between the feed point and each end (50% out), would require coils having an inductive reactance of approximately 950 Ω at the operating frequency for antenna resonance.

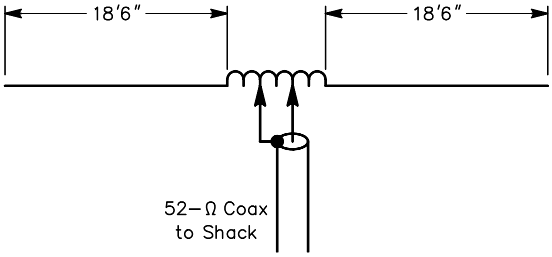

Fig 72—The WØSVM "Shorty Forty" center-loaded antenna. Dimensions given are for 7.0 MHz. The loading coil is 5 inches long and 2½ inches diameter. It has a total of 30 turns of #12 wire wound at 6 turns per inch (Miniductor 3029 stock).

For the antennas shown, the longer the overall length (dimension A, Fig 70A, in feet) and the farther the loading coils are from the center of the antenna (dimension B, also in feet), the greater the efficiency of the antenna. As dimension B is increased, however, the inductance required to resonate the antenna at the desired frequency increases.

Approximate inductive reactances for single-band resonance (for the antenna in Fig 70A only) may be determined with the aid of **Fig 71** or from Eq 10 below. The final values will depend on the proximity of surrounding objects in individual installations and must be determined experimentally. The use of high-Q low-loss coils is important for maximum efficiency.

A dip meter or SWR indicator is recommended for use during adjustment of the system. Note that the minimum inductance required is for a center-loaded dipole. If the inductive reactance is read from Fig 66 for a dimension B of zero, one coil having approximately twice this reactance can be used near the center of the dipole. **Fig 72** illustrates this idea. This antenna was conceived by Jack Sobel, WØSVM, who dubbed the 7-MHz version the "Shorty Forty."

$$X_L = \frac{10^6}{34\pi f}\left(\frac{\ln\frac{24\left(\frac{234}{f}\right)-B}{D}-1\left(\left(1-\frac{fB}{234}\right)^2-1\right)}{\frac{234}{f}-B}-\frac{\left(\ln\frac{24\left(\frac{A}{2}-B\right)}{D}-1\right)\left(\frac{\frac{fA}{2}-fB}{234}\right)^2-1}{\frac{A}{2}-B}\right) \quad \text{(Eq 10)}$$

Inverted-L Antennas

The antenna shown in **Fig 73** is called an *inverted-L* antenna. It is simple and easy to construct and is a good antenna for the beginner or the experienced 1.8-MHz DXer. Because the overall electrical length is made somewhat greater than λ/4, the feed-point resistance is on the order of 50 Ω, with an inductive reactance. That reactance is canceled by a series capacitor as indicated in the figure. For a vertical section length of 60 feet and a horizontal section length of 115 feet, the input impedance is ≈ 40 + *j* 300 Ω. Longer vertical or horizontal sections would increase the input impedance. The azimuthal radiation pattern is slightly asymmetrical with ≈1 to 2-dB increase in the direction opposite to the horizontal wire. This antenna requires a good buried ground system or elevated radials and will have a 2:1 SWR bandwidth of about 50 kHz.

This antenna is a form of top-loaded vertical, where the top loading is asymmetrical. This results in both vertical and horizontal polarization because the currents in the top wire do not cancel like they would in a symmetrical-T vertical. This is not necessarily a bad thing because it eliminates the zenith null present in a true vertical. This allows for good communication at short ranges as well as for DX.

A yardarm attached to a tower or a tree limb can be used to support the vertical section. As with any vertical, for best results the vertical section should be as long as possible. A good ground system is necessary for good results—the better the ground, the better the results.

If you don't have the space for the inverted L shown in Fig 73 (with its 115-foot horizontal section) and if you don't have a second tall supporting structure to make the top

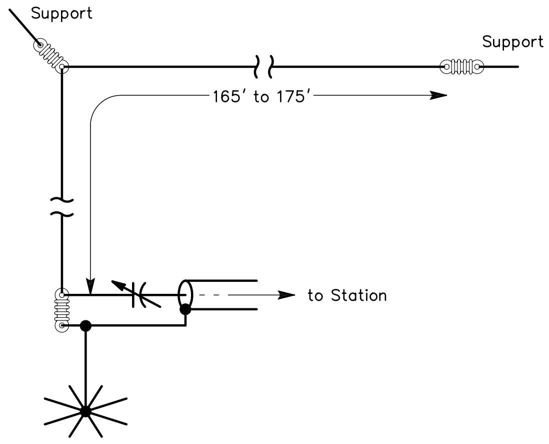

Fig 73—The 1.8-MHz inverted L. Overall wire length is 165 to 175 feet. The variable capacitor has a capacitance range from 100 to 800 pF, at 3 kV or more. Adjust antenna length and variable capacitor for lowest SWR.

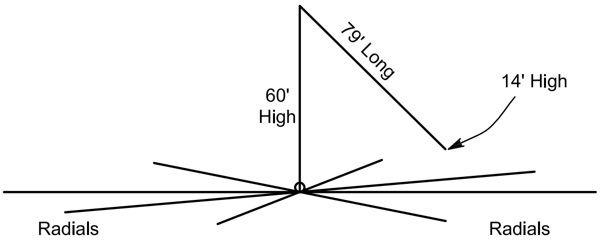

Fig 74—Sketch showing a modified 160-meter inverted L, with a single supporting 60-foot high tower and a 79-foot long slanted top-loading wire. The feed-point impedance is about 12 Ω in this system, requiring a quarter-wave matching transformer made of paralleled 50-Ω coaxes.

wire horizontal, consider sloping the top wire down towards ground. **Fig 74** illustrates such a setup, with a 60-foot high vertical section and a 79-foot sloping wire. As always, you will have to adjust the length of the sloping wire to fine-tune the resonant frequency. For a good ground radial system, the feed-point impedance is about 12 Ω, which may be transformed to 50 Ω with a 25-Ω quarter-wave transformer consisting of two paralleled 50-Ω quarter-wave coaxes. The peak gain will decrease about 1 dB compared to the inverted L shown in Fig 73. **Fig 75** overlays the elevation responses for average ground conditions. The 2:1 SWR bandwidth will be about 30 kHz, narrower than the larger system in Fig 73.

If the ground system suggested for Figs 73 and 74 is not practical, you can use a single elevated radial as shown in **Fig 76**. For the dimensions shown in the figure $Z_i = 50 + j\,498\ \Omega$, requiring a 175-pF series resonating capacitor. The azimuthal radiation pattern is shown in **Fig 77** compared to

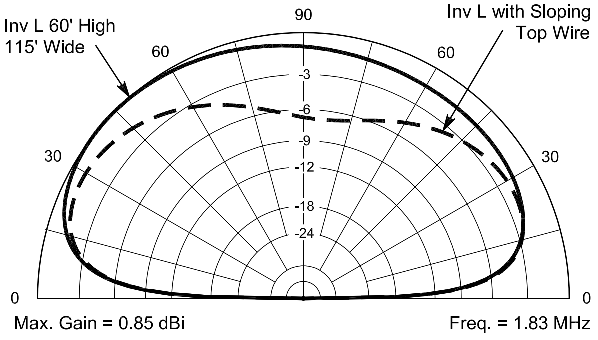

Fig 75—Overlay of the elevation responses for the inverted-L antennas in Fig 73 (solid line) and Fig 74 (dashed line). The gains are very close for these two setups, provided that the ground radial system for the antenna in Fig 74 is extensive enough to keep ground losses low.

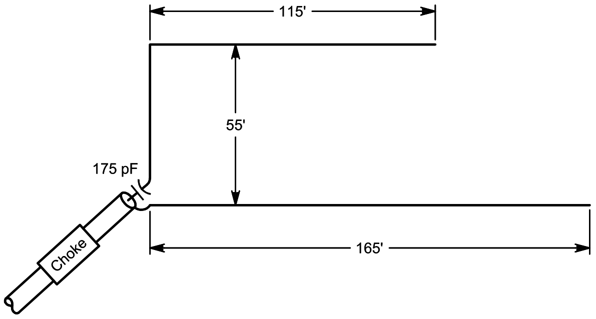

Fig 76—A single elevated radial can be used for the inverted-L. This changes the directivity slightly. The series tuning capacitor is approximately 175 pF for this system.

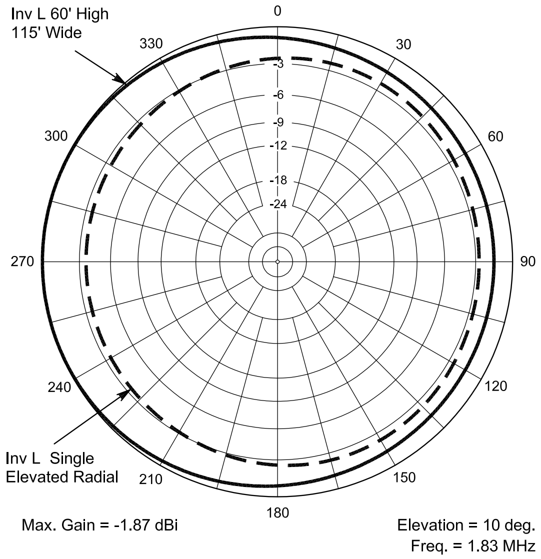

Fig 77—Azimuthal pattern comparison for inverted-L antennas shown in Fig 73 (solid line) and the compromise, single-radial system in Fig 76 (dashed line). This is for a takeoff angle of 10°.

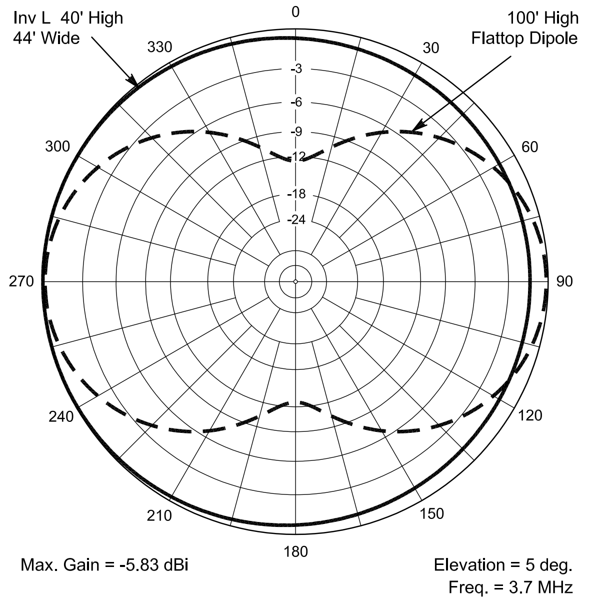

Fig 78—Azimuthal pattern at a takeoff angle of 5° for an 80-meter version of the inverted L (solid line) in Fig 73, compared to the response for a 100-foot high flattop dipole (dashed line).

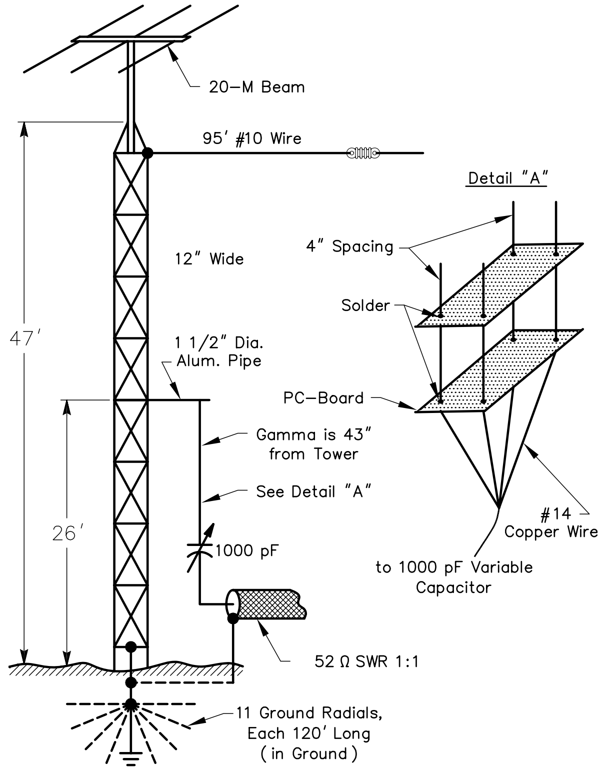

Fig 79—Details and dimensions for gamma-match feeding a 50-foot tower as a 1.8-MHz vertical antenna. The rotator cable and coaxial feed line for the 14-MHz beam is taped to the tower legs and run into the shack from ground level. No decoupling networks are necessary.

the inverted L in Fig 68. Note that the 1 to 2-dB asymmetry is now in the direction of the horizontal wires, just the opposite of that for a symmetrical ground system. The 2:1 SWR bandwidth is about 40 kHz, assuming that the series capacitor is adjusted at 1.83 MHz for minimum SWR.

Fig 78 shows the azimuthal response at a 5° elevation angle for an 80-meter version of the inverted L in Fig 73. The peak response occurs at an azimuth directly behind the direction in which the horizontal portion of the inverted L points. For comparison, the response for a 100-foot high flattop dipole is also shown. The top wire of this antenna is only 40 feet high and the 2:1 SWR bandwidth is about 150 kHz wide with a good, low-loss ground-radial system.

Fig 78 illustrates that the azimuth response of an inverted L is nearly omnidirectional. This gives such an antenna an advantage in certain directions compared to a flattop dipole, which is constrained by its supporting mounts (such as trees or towers) to favor fixed directions. For example, the flattop dipole in Fig 78 is at its weakest at azimuths of 90° and 270°, where it is down about 12 dB compared to the inverted L. Hams who are fortunate enough to have high rotary dipoles or rotatable low-band Yagis have found them to be very effective antennas indeed.

A DIFFERENT APPROACH

Fig 79 shows the method used by Doug DeMaw, W1FB, to gamma match his self-supporting 50-foot tower operating as an inverted L. A wire cage simulates a gamma rod of the proper diameter. The tuning capacitor is fashioned from telescoping sections of 1½ and 1½-inch aluminum tubing with polyethylene tubing serving as the dielectric. This capacitor is more than adequate for power levels of 100 W. The horizontal wire connected to the top of the tower provides the additional top loading.

Sloper Antennas

Sloping dipoles and λ/2 dipoles can be very useful antennas on the low bands. These antennas can have one end attached to a tower, tree or other structure and the other end near ground level, elevated high enough so that passersby can't contact them, of course. The following section gives a number of examples of these types of antennas.

THE HALF-WAVE SLOPING DIPOLE

If you have a sufficiently high support, you can install a halfwave dipole sloping downwards toward ground to provide vertical as well as horizontal polarization. This antenna is popularly known as a *sloper* or a *halfwave sloper*. The amount of slope from horizontal can vary from 0°, where the dipole is in a flattop configuration, all the way to 90°, where the dipole becomes fully vertical. The latter configuration is sometimes called a *Halfwave Vertical Dipole* (HVD).

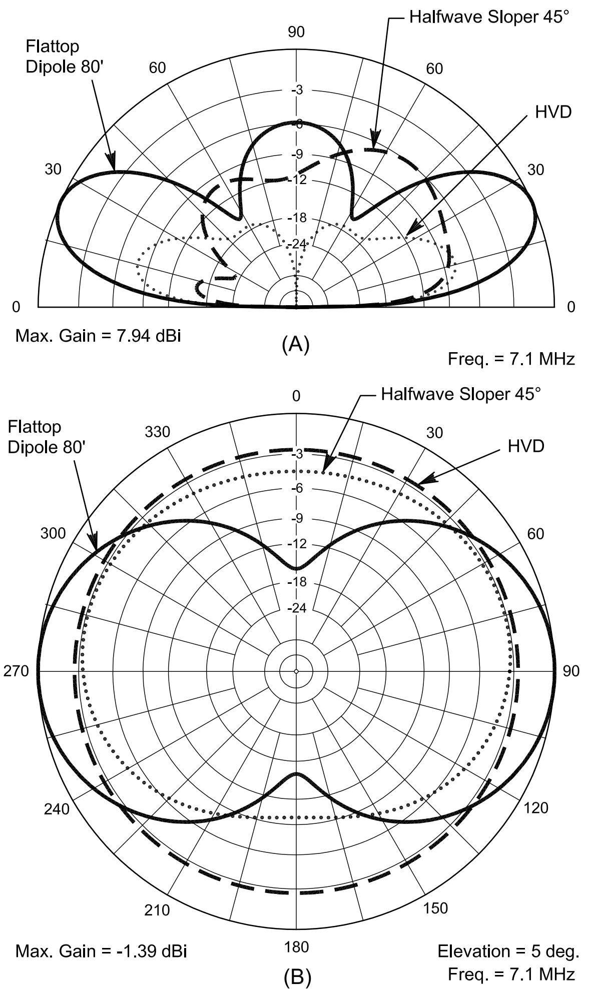

Fig 80—At A, the azimuthal responses for a flattop dipole (solid line), a dipole whose end has been tilted down 45° (dashed line), and a HVD (halfwave vertical dipole, dotted line). All these were modeled over average ground, with a conductivity of 5 mS/m and a dielectric constant of 13. Note that the tilted dipole exhibits about 5 dB front-to-back ratio, although its maximum gain is less than either the HVD or flattop dipole. At B, the elevation-plane patterns for the same antennas. Note that the tilted halfwave dipole (dashed line) has more energy at higher elevation angles than either the flattop dipole or HVD.

The question arises when contemplating a vertical halfwave dipole or a halfwave sloping dipole about how to treat the feed line to make sure it doesn't accidentally become part of the radiating system. The ideal situation would be to bring the feed line out perpendicular to the vertical or sloping wire for an infinite distance. Obviously, that isn't very practical because the feed line eventually has to be connected to a transmitter located near the ground. An intensive modeling study on feeding an HVD was done for the book *Simple and Fun Antennas for Hams*. This study indicated that a slant angle down to the ground of as little as 30° from a vertical radiator can work with only minor interaction, provided that common-mode decoupling chokes were employed at the feed point and a quarter-wavelength down the line from the feed point. These common-mode chokes can consist of either discrete ferrite beads placed over the outer jacket of the coaxial line or multiple turns of the coax itself to form a choke.

Fig 80A compares the 40-meter azimuthal patterns at a DX takeoff angle of 5° for three configurations: a flattop dipole, a dipole tilted down 45° and an HVD (halfwave vertical dipole). These are computed for ground with average conductivity and dielectric constant, and for a maximum height of 80 feet in each configuration. The sloping halfwave dipole exhibits about 5 dB of front-to-back ratio, although even at its most favored direction it doesn't quite have the same maximum gain as the HVD or the flattop dipole.

The reason why the maximum gain for the sloper is less than the other two configurations, even while still exhibiting some front-to-back pattern, is shown in Fig 80A, which shows the elevation-plane patterns for the same antennas, each at the azimuth of maximum gain. The halfwave sloper distributes much of its energy higher in elevation than the HVD, lowering the peak-gain potential of the sloper.

You can also see from Fig 80B that the 80-foot high horizontal dipole would perform much better than either the HVD or halfwave sloper for close-in local contacts, which occur at high elevation angles. On the other hand, except for the greater gain exhibited in the flattop dipole's most favored directions, the HVD has more gain than the other antennas at low elevation angles. While the HVD's omnidirectional pattern is a plus for transmitting, it may be a problem for receiving, where local noise may be coming from specific directions (such as power lines) and may also be predominantly vertically polarized. In such cases, a horizontally polarized flattop dipole may be a considerably better receiving antenna than a vertically polarized antenna of any sort. We've already mentioned the fact that a rotary flattop dipole high in the air can be a very effective antenna on the low bands.

THE QUARTER-WAVELENGTH “HALF SLOPER”

Perhaps one of the easiest antennas to install is the λ/4 sloper shown in **Fig 81**. As pointed out above, a sloping λ/2 dipole is known among radio amateurs as a *sloper* or sometimes as a *full sloper*. If only one half of it is used, it becomes a *half sloper*. The performance of the two types of sloping antennas is similar—They exhibit some directivity in the direction of the slope and radiate vertically polarized energy at low angles respective to the horizon. The amount of directivity will range from 3 to 6 dB, depending upon the individual installation, and will be observed in the slope direction.

The main advantage of the half sloper over the full half-wave-long sloping dipole is that its supporting tower needn’t be as high. Both the half sloper and the full sloper place the feed point (the point of maximum current) high above lossy ground. But the half-sloper only needs half as much wire to build the antenna for a given amateur band. The disadvantage of the half sloper is that it is sometimes difficult or even impossible to obtain a low SWR when using coaxial-cable feed, especially without a good isolating choke balun. (See the section above on isolating ground-plane antennas.)

Other factors that affect the feed impedance are tower height, height of the attachment point, enclosed angle between the sloper and the tower, and what is mounted atop the tower (HF or VHF beams). Further, the quality of the ground under the tower (ground conductivity, radials, etc) has a marked effect on the antenna performance. The final SWR can vary (after optimization) from 1:1 to as high as 6:1. Generally speaking, the closer the low end of the slope wire is to ground, the more difficult it will be to obtain a good match.

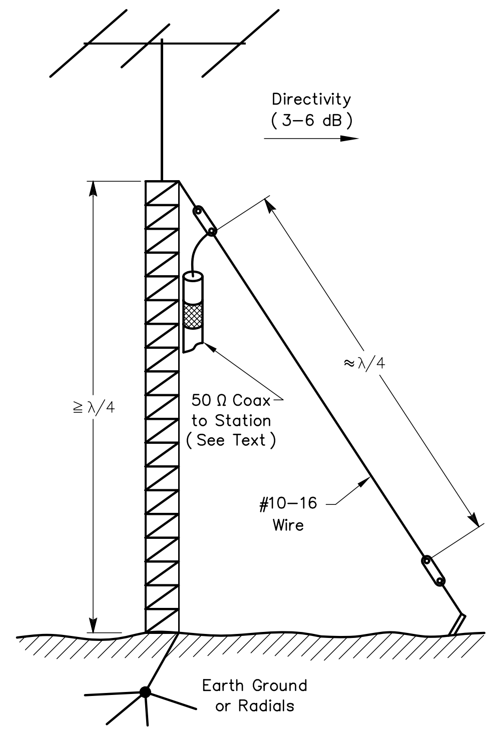

Fig 81—The λ/4 “half sloper” antenna.

Basic Recommendations for a Half Sloper

The half sloper can be an excellent DX type of antenna. Hams usually install theirs on a metal supporting structure such as a mast or tower. The support needs to be grounded at the lower end, preferably to a buried or on-ground radial system. If a nonconductive support is used, the outside of the coax braid becomes the return circuit and should be grounded at the base of the support. As a starting point you can attach the sloper so the feed point is approximately λ/4 above ground. If the tower is not high enough to permit this, the antenna should be fastened as high on the supporting structure as possible. Start with an enclosed angle of approximately 45°, as indicated in Fig 81. Cut the wire to the length determined from

$$\ell = \frac{260}{f_{MHz}} \qquad \text{(Eq 11)}$$

This will allow sufficient extra length for pruning the wire for the lowest SWR. A metal tower or mast becomes an operating part of the half sloper system. In effect, it and the slope wire function somewhat like an inverted-V dipole antenna. In other words, the tower operates as the missing half of the dipole. Hence its height and the top loading (beams) play a significant role.

Detailed modeling indicates that a sufficiently large

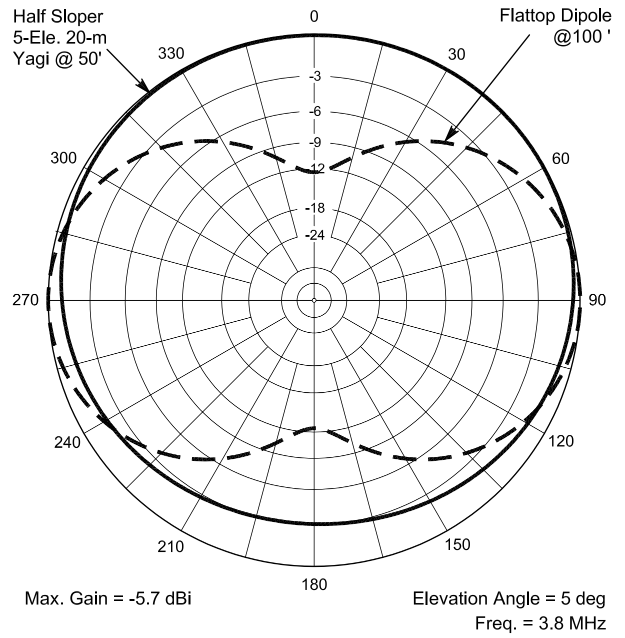

Fig 82—Radiation pattern for a typical half sloper (solid line) mounted on a 50-foot high tower with a large 5-element 20-meter beam on the top compared to that for a flattop dipole (dashed line) at 100 feet. At a 5° takeoff angle typical for DX work on 80 meters, the two antennas are pretty comparable in the directions favored by the high dipole. In other directions, the half sloper has an advantage of more than 10 dB.

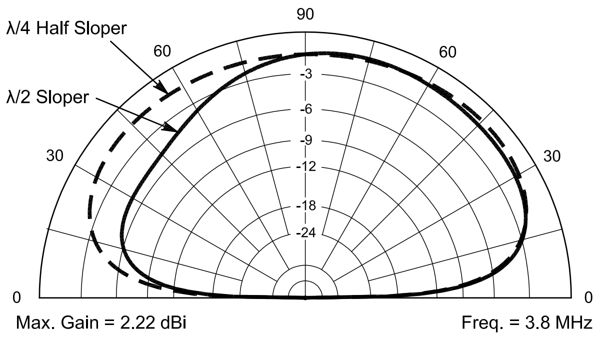

Fig 83—Comparison of elevation patterns for a full-sized halfwave sloper (solid line) on a 100-foot tower and a half sloper (dashed line) on a 50-foot tower with a 5-element 20-meter Yagi acting as a top counterpoise. The performance is quite comparable for these two systems.

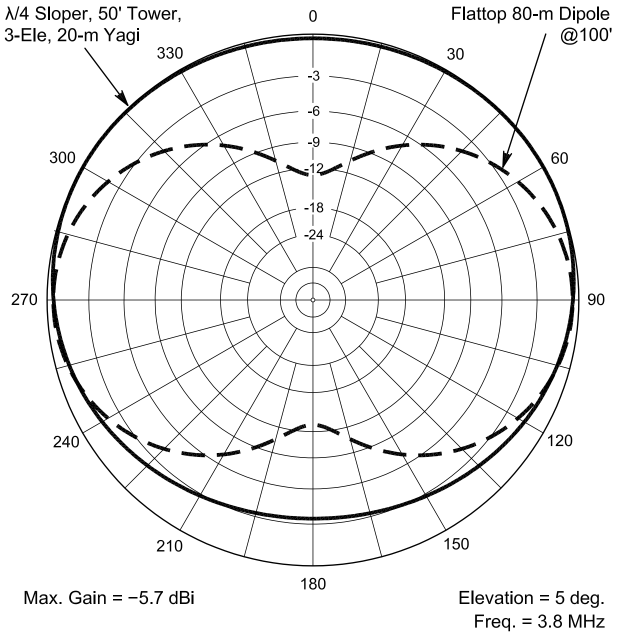

Fig 84—Comparing the azimuthal response of a half sloper (solid line) on a 50-foot tower with a 3-element 20-meter Yagi on top to that of a flattop dipole (dashed line) at 100 feet. The two are again quite comparable at a 5° takeoff angle.

mass of metal (that is, a large, "Plumber's Delight" Yagi) connected to the top of the tower acts like enough of a "top counterpoise" that the tower may be removed from the model with little change in the essential characteristics of the half-sloper system. Consider an installation using a freestanding 50-foot tower with a large 5-element 20-meter Yagi on top. This Yagi is assumed to have a 40-foot boom oriented 90° to the direction of the slanted 80-meter half-sloper wire. The best SWR that could be reached by changing the length and slant angle for this sloper is 1.67:1, representing a feed-point impedance of 30.1 – j 2.7 Ω. The peak gain at 3.8 MHz is 0.97 dBi at an elevation angle of 70°. **Fig 82** shows the azimuth-plane pattern for this half sloper, compared to a 100-foot high flattop dipole for reference, at an elevation angle of 5°.

Removing the tower from the model resulted in a feed-point impedance of 30.1 – j 1.5 Ω and a peak gain of 1.17 dBi. The tower is obviously not contributing much in this setup, since the mass of the large 20-meter Yagi is acting like an elevated counterpoise all by itself. It's interesting to rotate the boom of the model Yagi and observe the change in SWR that occurs on the half-sloper antenna. With the boom turned 90°, the SWR falls to 1.38:1. This level of SWR change could be measured with amateur-type instrumentation.

On the other hand, substituting a smaller 3-element 20-meter Yagi with an 18-foot boom in the model does result in significant change in feed-point impedance and gain when the tower is removed from the model, indicating that the "counterpoise effect" of the smaller beam is insufficient by itself. Interestingly enough, the best SWR for the half sloper/tower and the 3-element Yagi (with its boom inline with the half sloper is 1.33:1), changing to 1.27:1 with the boom turned 90°. Such a small change in SWR would be difficult to measure using typical amateur instrumentation.

In any case, the 50-Ω transmission line feeding a half sloper should be taped to the tower leg at frequent intervals to make it secure. The best method is to bring it to earth level, then route it to the operating position along the surface of the ground if it can't be buried. This will ensure adequate RF decoupling, which will help prevent RF energy from affecting the equipment in the station. Rotator cable and other feed lines on the tower or mast should be treated in a similar manner.

Adjustment of the half sloper is done with an SWR indicator in the 50-Ω transmission line. A compromise can usually be found between the enclosed angle and wire length, providing the lowest SWR attainable in the center of the chosen part of an amateur band. If the SWR "bottoms out" at 2:1 or lower, the system will work fine without using an antenna tuner, provided the transmitter can work into the load. Typical optimum values of SWR for 3.5 or 7-MHz half slopers are between 1.3:1 and 2:1. A 100-kHz bandwidth is normal on 3.5 MHz, with 200 kHz being typical at 7 MHz.

If the lowest SWR possible is greater than 2:1, the attachment point can be raised or lowered to improve the match. Readjustment of the wire length and enclosed angle may be necessary when the feed-point height is changed. If the tower is guyed, the guy wires will need to be insulated from the tower and broken up with additional insulators to prevent resonance.

At this point you may be curious about which antenna is better—a full sloper or a half sloper. The peak gain for each antenna is very nearly identical. **Fig 83** overlays the elevation-plane pattern for the full-sized halfwave sloper on a 100-foot tower and for the half sloper shown in Fig 81 on a 50-foot tower with a 5-element 20-meter Yagi on top. The

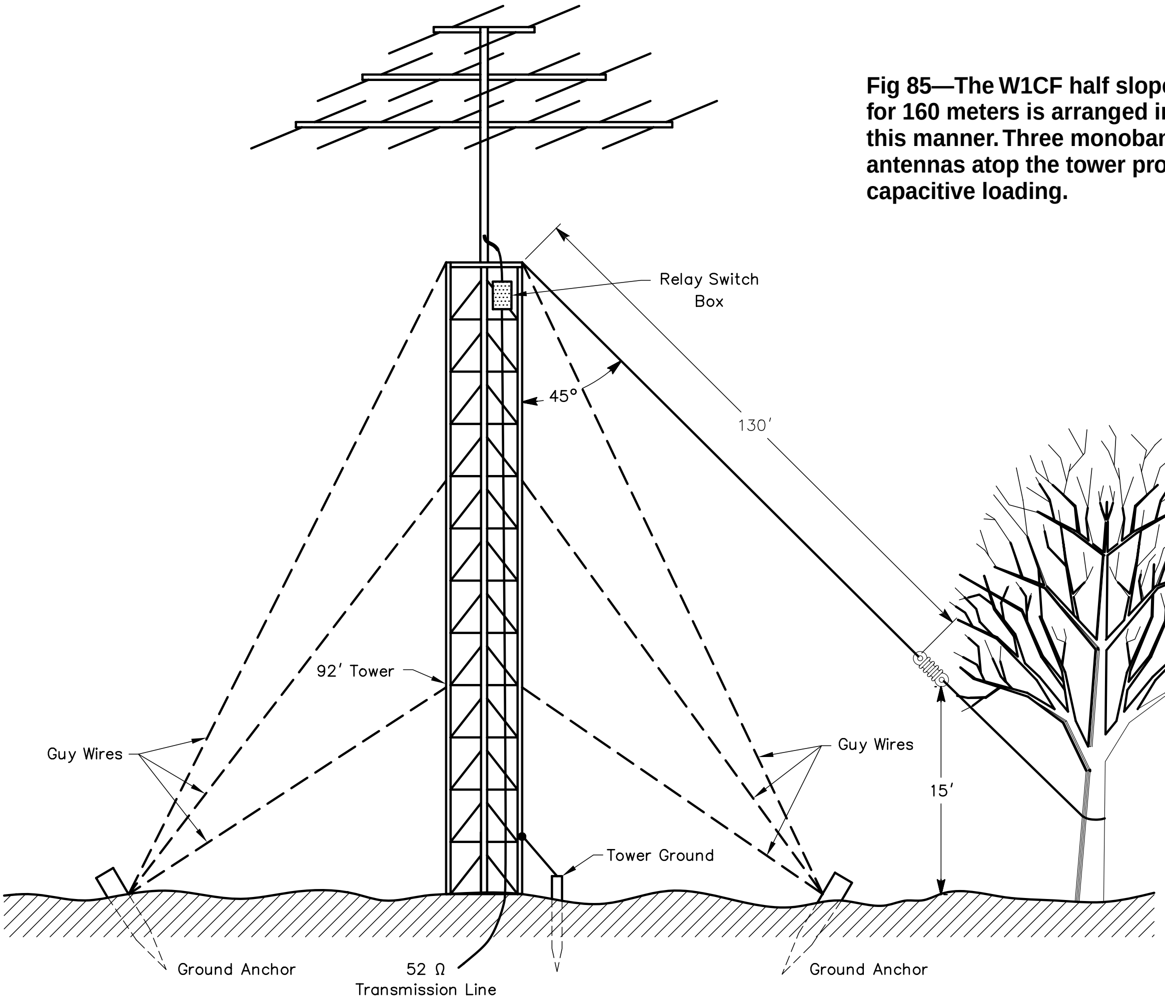
Relay Switch Box
45°
130′
92′ Tower
Guy Wires
Guy Wires
15′
Tower Ground
Ground Anchor
52 Ω Transmission Line
Ground Anchor

Fig 85—The W1CF half sloper for 160 meters is arranged in this manner. Three monoband antennas atop the tower provide capacitive loading.

full-sized halfwave sloper has more front-to-back ratio, but it is only a few dB more than the half sloper. **Fig 84** compares the azimuthal patterns at a 5° takeoff angle for a 100-foot high flattop dipole and a half-sloper system on a 50-foot tower with a 3-element 20-meter Yagi on top.

Despite the frustration some have experienced trying to achieve a low SWR with some half-sloper installations, many operators have found the half sloper to be an effective and low-cost antenna for DX work.

1.8-MHz ANTENNA SYSTEMS USING TOWERS

The half sloper discussed above for 80 or 40-meter operation will also perform well on 1.8 MHz where vertically polarized radiators can achieve the low takeoff angles needed on Topband. Prominent 1.8-MHz operators who have had success with the half sloper antenna suggest a minimum tower height of 50 feet. Dana Atchley, W1CF (SK), used the configuration sketched in **Fig 85**. He reported that the uninsulated guy wires act as an effective counterpoise for the sloping wire. In **Fig 86** is the feed system used by Doug

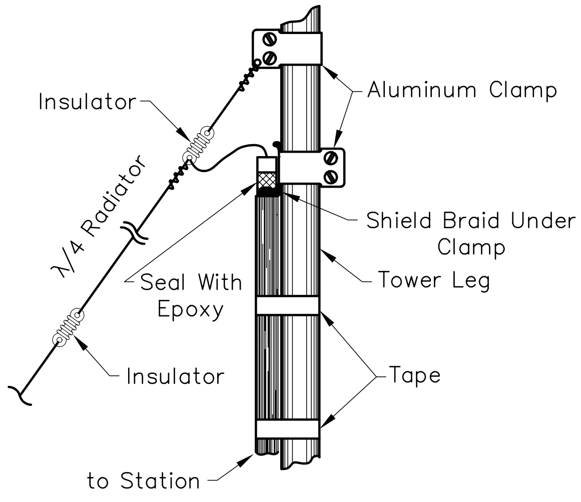
Aluminum Clamp
Insulator
λ/4 Radiator
Shield Braid Under Clamp
Tower Leg
Seal With Epoxy
Insulator
Tape
to Station

Fig 86—Feed system used by W1FB for 1.8 MHz half sloper on a 50-foot self-supporting tower.

DeMaw, W1FB (SK), on a 50-foot self-supporting tower. The ground for the W1FB system is provided by buried radials connected to the tower base.

As described previously, a tower can also be used as a true vertical antenna, provided a good ground system is used. The shunt-fed tower is at its best on 1.8 MHz, where a full λ/4 vertical antenna is rarely possible. Almost any tower height can be used. An HF beam at the top provides some top loading.

THE K1WA 7-MHz "SLOPER SYSTEM"

One of the more popular antennas for 3.5 and 7 MHz is the half-wave long sloping dipole described previously. David Pietraszewski, K1WA, made an extensive study of sloping dipoles at different heights with reflectors at the 3-GHz frequency range. From his experiments, he developed the novel 7-MHz antenna system described here. With several sloping dipoles supported by a single mast and a switching network, an antenna with directional characteristics and forward gain can be simply constructed. This 7-MHz system uses several "slopers" equally spaced around a common center support. Each dipole is cut to λ/2 and fed at the center with 50-Ω coax. The length of each feed line is 36 feet.

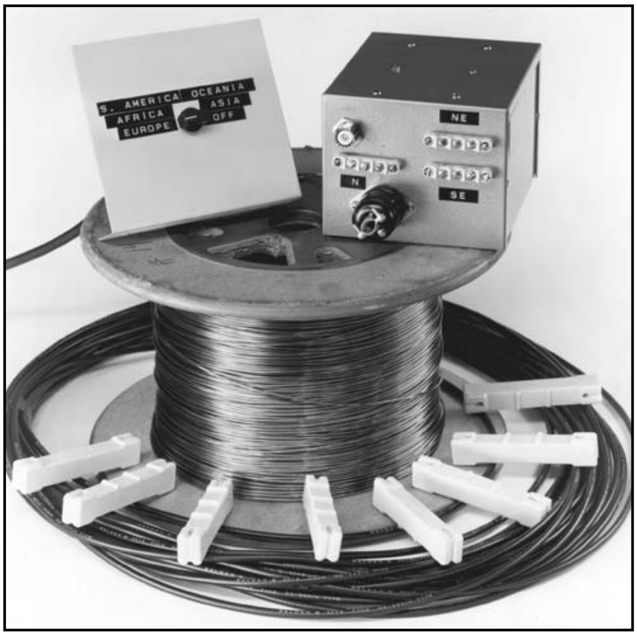

Fig 88—The basic materials required for the sloper system. The control box appears at the left, and the relay box at the right.

All of the feed lines go to a common point on the support (tower) where the switching takes place. The line length of 36 feet is just over 3λ/8, which provides a useful quality. At 7 MHz, the coax looks inductive to the antenna when the end at the switching box is open circuited. This has the effect of adding inductance at the center of the sloping dipole element, which electrically lengthens the element. The 36-foot length of feed line serves to increase the length of the element about 5%. This makes any unused element appear to be a reflector.

The array is simple and effective. By selecting one of the slopers through a relay box located at the tower, the system becomes a parasitic array that can be electrically rotated. All but the driven element of the array become reflectors.

The physical layout is shown in **Fig 87**, and the basic materials required for the sloper system are shown in **Fig 88**. The height of the support point should be about 70 feet, but can be less and still give reasonable results. The upper portion of the sloper is 5 feet from the tower, suspended by rope. The wire makes an angle of 60° with the ground.

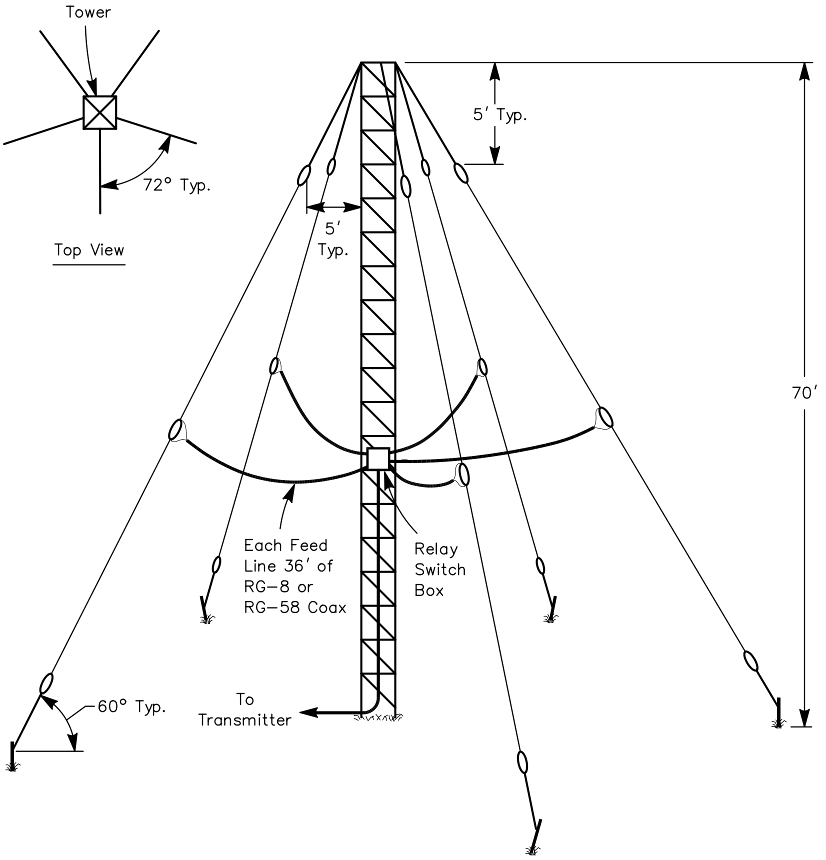

Fig 87—Five sloping dipoles suspended from one support. Directivity and forward gain can be obtained from this simple array. The top view shows how the elements should be spaced around the support.

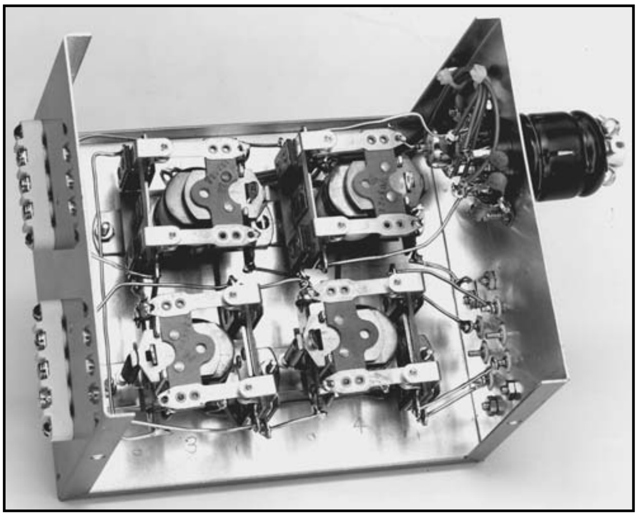

Fig 89—Inside view of relay box. Four relays provide control over five antennas. See text. The relays pictured here are Potter and Brumfeld type MR11D.

In **Fig 89**, the switch box is shown containing all the necessary relays to select the proper feed line for the desired direction. One feed line is selected at a time and the feed lines of those remaining are opened, **Fig 90**. In this way the array is electrically rotated. These relays are controlled from inside the shack with an appropriate power supply and rotary switch. For safety reasons and simplicity, 12-volt dc relays are used. The control line consists of a five conductor cable, one wire used as a common connection; the others go to the four relays. By using diodes in series with the relays and a dual-polarity power supply, the number of control wires can be reduced, as shown in Fig 90B.

Measurements indicate that this sloper array provides up to 20 dB front-to-back ratio and forward gain of about 4 dB over a single half-wave sloper. **Fig 91** shows the azimuthal pattern (at a 5° takeoff angle) for the K1WA array, compared to a 100-foot high flattop dipole and a full sloper suspended from a 50-foot tower. These patterns were calculated for average ground conditions. Just for fun, look at

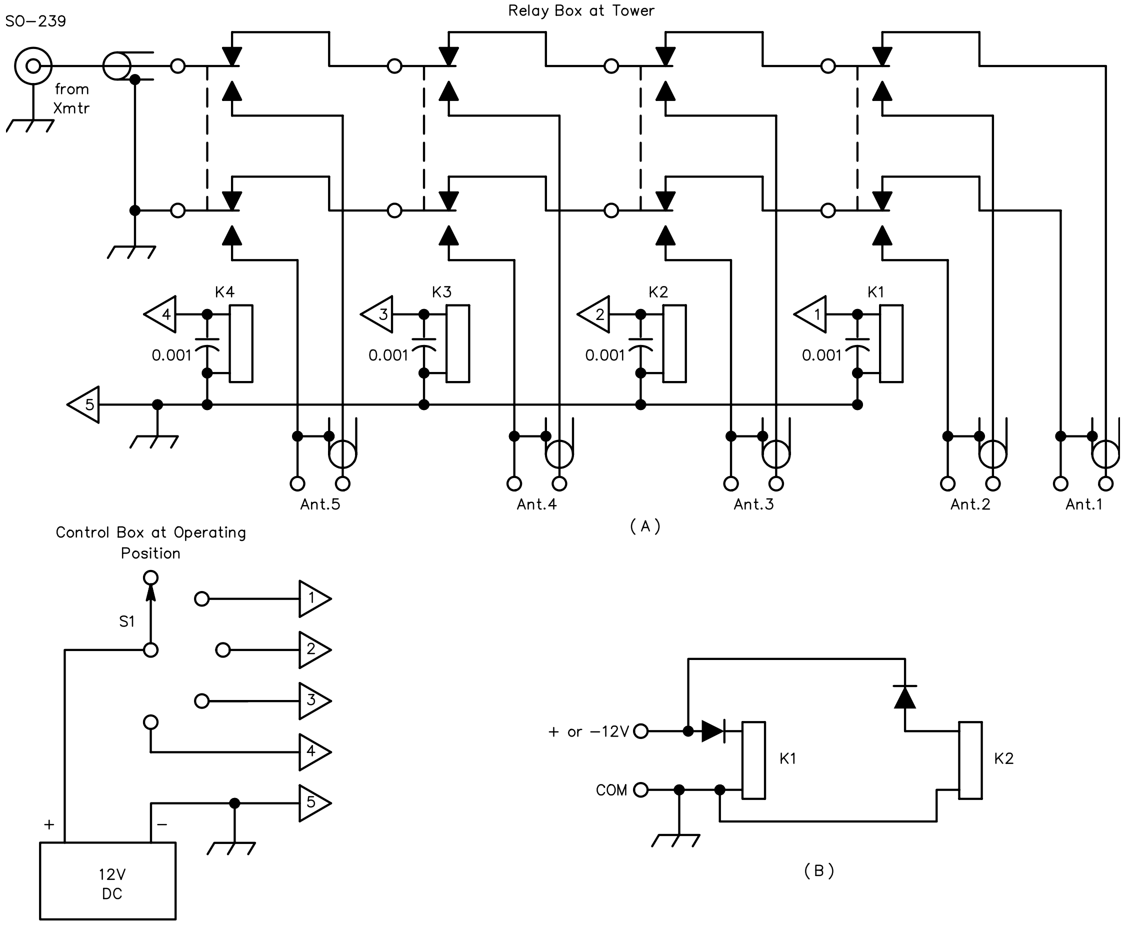

Fig 90—Schematic diagram for sloper control system. All relays are 12-volt dc, DPDT, with 8-A contact ratings. At A, the basic layout, excluding control cable and antennas. Note that the braid of the coax is also open-circuited when not in use. Each relay is bypassed with 0.001-µF capacitors. The power supply is a low current type. At B, diodes are used to reduce the number of control wires when using dc relays. See text.

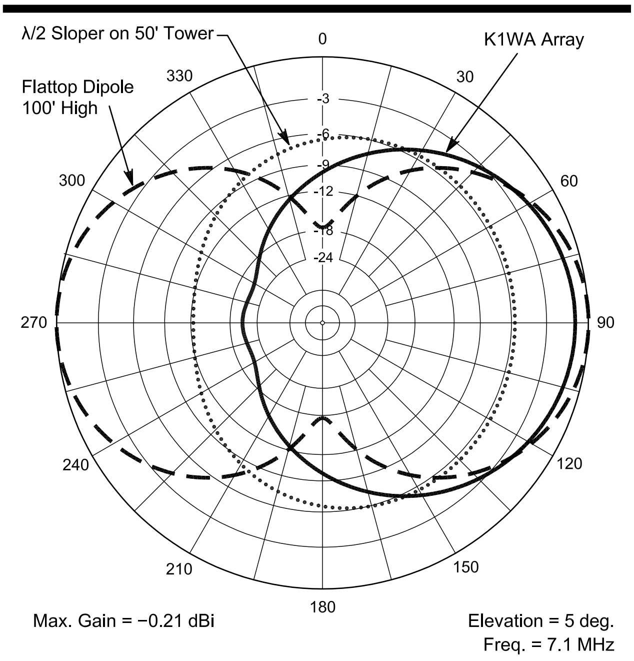

Fig 91—Azimuth pattern for K1WA 40-meter sloper array (solid line), compared to a flattop dipole (dashed line) at 100 feet and a halfwave full sloper on a 50-foot tower (dotted line). The K1WA array has an excellent front-to-back ratio and almost as much gain as the high flattop dipole. These patterns are for average ground.

Fig 92, which shows a comparison between a 100-foot high flattop dipole and a K1WA array placed over saltwater. Now that's a real barnburner at low takeoff angles! Such a seaside system would be very competitive with a rotatable 2-element "shorty-40" type of Yagi.

If one direction is the only concern, the switching system can be eliminated and the reflectors should be cut 5% longer than the resonant frequency. The feature worth noting is the good F/B ratio. By arranging the system properly, a null can be placed in an unwanted direction, thus making it an effective receiving antenna. In the tests conducted with this antenna, the number of reflectors used were as few as one and as many as five. The optimum combination appeared to occur with four reflectors and one driven element. No tests were conducted with more than five reflectors. This same array can be scaled to 3.5 MHz for similar results.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

D. Atchley, Jr., "Putting the Quarter-Wave Sloper to Work on 160," *QST*, Jul 1979, pp 19-20.

J. Belrose, "Transmission Line Low Profile Antennas," *QST*, Dec 1975, pp 19-25.

L. Braskamp, AA6GL, "*MOBILE*, a Computer Program for Short HF Verticals," *The ARRL Antenna Compendium Vol 4*, pp 92-96.

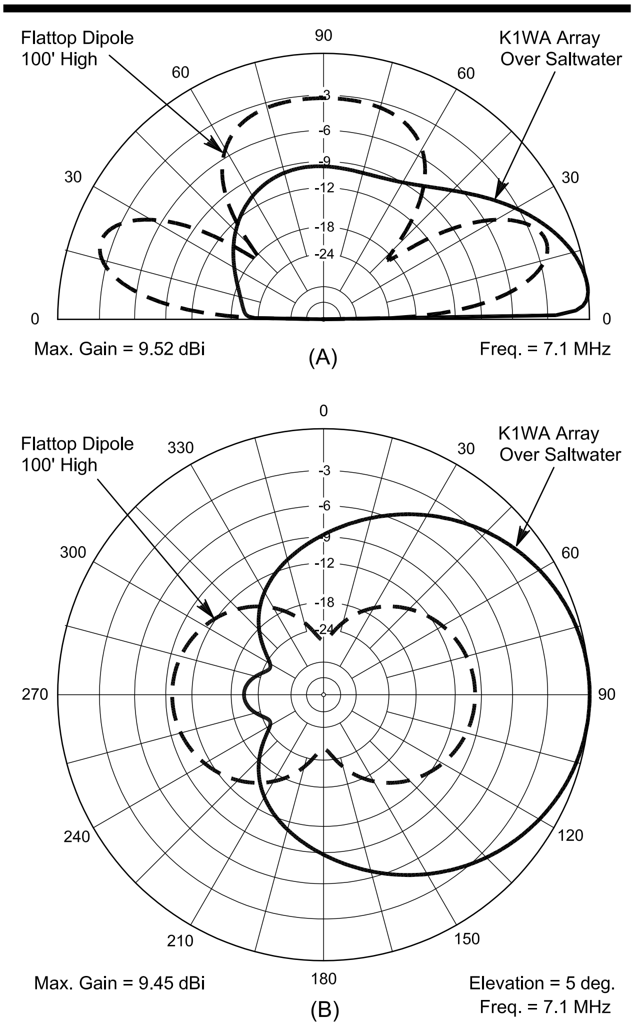

Fig 92—The K1WA 40-meter sloper array over saltwater, compared to a 100-foot high flattop dipole. At A, azimuth patterns. At B, elevation patterns. Now the sloper array really comes into its own!

G. H. Brown, "The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas," *Proc. IRE*.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE* , Jun 1937, pp 753-787.

P. Carr, N4PC, "A Two Band Half-Square Antenna With Coaxial Feed," *CQ*, Sep 1992, pp 40-45.

P. Carr, N4PC, "A DX Antenna For 40 Meters," *CQ*, Sep 1994, pp 40-43.

A. Christman, KB8I, "Elevated Vertical Antennas for the Low Bands," *The ARRL Antenna Compendium Vol 5*, pp 11-18.

Curry Communications, PO Box 1884, Burbank, CA 91507, a kit manufacturer of lowfer CW transceivers.

D. DeMaw, "Additional Notes on the Half Sloper," *QST*, Jul 1979, p 20.

P. Dodd, *The LF Experimenter's Source Book* (RSGB, 1996).

A. C. Doty, Jr., J. A. Frey and H. J. Mills, "Efficient Ground Systems for Vertical Antennas," *QST*, Feb 1983, pp 9-12.

S. Ford, "*Low*fing on 1750 Meters," *QST*, Oct 1993, pp 67-68.

R. Fosberg, "Some Notes on Ground Systems for 160 Meters," *QST*, Apr 1965, pp 65-67.

J. Hall, "Off-Center-Loaded Dipole Antennas," *QST*, Sep 1974, pp 28-34.

G. Hubbell, "Feeding Grounded Towers as Radiators," *QST*, pp 32-33, 140, 142.

C. Hutchinson and R. D. Straw, *Simple and Fun Antennas for Hams* (Newington, Ct: ARRL, 2002).

P. H. Lee, *The Amateur Radio Vertical Antenna Handbook*, 1st edition (Port Washington, NY: Cowan Publishing Corp., 1974).

C. J. Michaels, "Some Reflections on Vertical Antennas," *QST*, July 1987, pp 15-19; feedback *QST*, Aug 1987, p 39.

Rashed and C. Tai, "A New Class of Wire Antennas," 1982 International Symposium Digest, *Antennas and Propagation*, Vol 2, IEEE.

T. Russell, N4KG, "Simple, Effective, Elevated Ground-Plane Antennas," *QST*, June 1994, pp 45-46

F. J. Schnell, "The Flagpole Deluxe," *QST*, Mar 1978, pp 29-32.

R. Severns, N6LF, "Using the Half-Square Antenna for Low-Band Dxing," *The ARRL Antenna Compendium Vol 5*, pp 35-42.

R. Severns, N6LF, "Broadbanding the Half-Square Antenna For 80-meter Dxing," *The ARRL Antenna Compendium Vol 5*, pp 43-44.

R. Severns, N6LF, "An Improved Double Extended Zepp," *The ARRL Antenna Compendium Vol 4*, pp 78-80.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971, pp 16-19, 22.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973, pp 13-18, 41.

J. Sevick, "The Constant-Impedance Trap Vertical," *QST* , Mar 1974, pp 29-34.

J. Sevick, "Short Ground-Radial Systems for Short Verticals," *QST*, Apr 1978, pp 30-33.

J. Sevick, W2FMI, *Transmission Line Transformers* (Noble Publishing, Atlanta, 1996).

W. Smith, W6BCX, "Bet My Money On A Bobtail Beam," *CQ*, Mar 1948, pp 21-23, 92.

J. Stanford, NNØF, "Linear-Loaded Short Wire Antennas," *The ARRL Antenna Compendium Vol 5*, pp 105-107.

J. Swank, W8HXR, "The S-Meter Bender," *73*, Jun 1978, pp 170-173.

J. Tyskewicz, "The Heli-Rope Antenna," *QST*, Jun 1971.

B. Vester, K3BC, "The Half-Square Antenna," *QST*, Mar 1974, pp 11-14.

A. D. Watt, *VLF Radio Engineering* (Pergamon Press, 1967) (out of print).

Prediction of Sky-wave Field Strength at Frequencies Between About 150 and 1700 kHz, ITU Doc 3/14, Radiocommunication Study Groups, Feb 1995.

Chapter 7

Multiband Antennas

For operation in a number of bands, such as those between 3.5 and 30 MHz, it would be impractical for most amateurs to put up a separate antenna for each band. But this is not necessary—a dipole, cut for the lowest frequency band to be used, can be operated readily on higher frequencies. To do so, one must be willing to accept the fact that such harmonic-type operation leads to a change in the directional pattern of the antenna, both in the azimuth and the elevation planes (see Chapter 2, Antenna Fundamentals, and Chapter 3, The Effects of Ground).

You can see from discussions in Chapter 6, Low-Frequency Antennas, that you should carefully plan the height at which you install a multiband horizontally polarized antenna. This is one aspect of multiband antennas. Another important thing to consider is that you should be willing to use so-called *tuned feeders*. A center-fed single-wire antenna can be made to accept power and radiate it with high efficiency on any frequency higher than its fundamental resonant frequency and, with a reduction in efficiency and bandwidth, on frequencies as low as one half the fundamental.

In fact, it is not necessary for an antenna to be a full half-wavelength long at the lowest frequency. An antenna can be considerably shorter than ½ λ, even as short as ¼ λ, and still be a very efficient radiator. The use of such short antennas results in stresses, however, on other parts of the system, for example the antenna tuner and the transmission line. This will be discussed in some detail in this chapter.

Methods have been devised for making a single antenna structure operate on a number of bands while still offering a good match to a transmission line, usually of the coaxial type. It should be understood, however, that a multiband antenna is not *necessarily* one that will match a given line on all bands on which you intend to use it. Even a relatively short whip type of antenna can be operated as a multiband antenna with suitable loading for each band. Such loading may be in the form of a coil at the base of the antenna on those frequencies where loading is needed, or it may be incorporated in the tuned feeders running from the transmitter to the base of the antenna.

This chapter describes a number of systems that can be used on two or more bands. Beam antennas, such as Yagis or quads, are treated separately in later chapters.

Simple Wire Antennas

The simplest multiband antenna is a random length of #12 or #14 wire. Power can be fed to the wire on practically any frequency using one or the other of the methods shown in **Fig 1**. If the wire is made either 67 or 135 feet long, it can also be fed through a tuned circuit, as in **Fig 2**. It is advantageous to use an SWR bridge or other indicator in the coax line at the point marked "X."

If you have installed a 28- or 50-MHz rotary beam, in many cases it may be possible to use the beam's feed line as an antenna on the lower frequencies. Connecting the two wires of the feeder together at the station end will give a random-length wire that can be conveniently coupled to the transmitter as in Fig 1. The rotary system at the far end will serve only to *end-load* the wire and will not have much other effect.

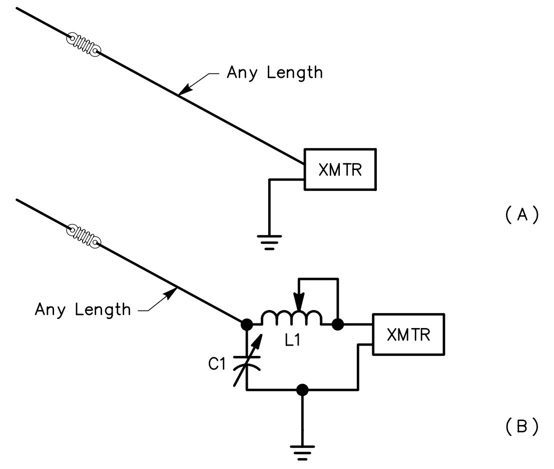

Fig 1—At A, a random-length wire driven directly from the pi-network output of a transmitter. At B, an L network for use in cases where sufficient loading cannot be obtained with the arrangement at A. C1 should have about the same plate spacing as the final tank capacitor in a vacuum-tube type of transmitter; a maximum capacitance of 100 pF is sufficient if L1 is 20 to 25 µH. A suitable coil would consist of 30 turns of #12 wire, 2½ inches diameter, 6 turns per inch. Bare wire should be used so the tap can be placed as required for loading the transmitter.

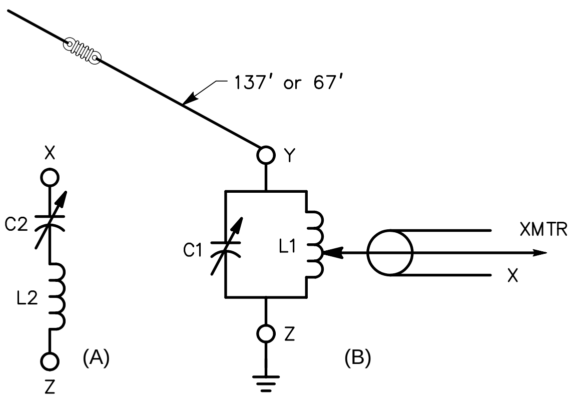

Fig 2—If the antenna length is 137 feet, a parallel-tuned coupling circuit can be used on each amateur band from 3.5 through 30 MHz, with the possible exception of the WARC 10-, 18- and 24-MHz bands. C1 should duplicate the final tank tuning capacitor and L1 should have the same dimensions as the final tank inductor on the band being used. If the wire is 67 feet long, series tuning can be used on 3.5 MHz as shown at the left; parallel tuning will be required on 7 MHz and higher frequency bands. C2 and L2 will in general duplicate the final tank tuning capacitor and inductor, the same as with parallel tuning. The L network shown in Fig 1B is also suitable for these antenna lengths.

One disadvantage of all such directly fed systems is that part of the antenna is practically within the station, and there is a good chance that you will have some trouble with RF feedback. RF within the station can often be minimized by choosing a length of wire so that the low feed-point impedance at a current loop occurs at or near the transmitter. This means using a wire length of $\lambda/4$ (65 feet at 3.6 MHz, 33 feet at 7.1 MHz), or an odd multiple of $\lambda/4$ (¾-λ is 195 feet at 3.6 MHz, 100 feet at 7.1 MHz). Obviously, this can be done for only one band in the case of even harmonically related bands, since the wire length that presents a current loop at the transmitter will present a voltage loop at two (or four) times that frequency.

When you operate with a random-length wire antenna, as in Figs 1 and 2, you should try different types of grounds on the various bands, to see what gives you the best results. In many cases it will be satisfactory to return to the transmitter chassis for the ground, or directly to a convenient metallic water pipe. If neither of these works well (or the metallic water pipe is not available), a length of #12 or #14 wire (approximately $\lambda/4$ long) can often be used to good advantage. Connect the wire at the point in the circuit that is shown grounded, and run it out and down the side of the house, or support it a few feet above the ground if the station is on the first floor or in the basement. It should not be connected to actual ground at any point.

END-FED ANTENNAS

When a straight-wire antenna is fed at one end with a two-wire transmission line, the length of the antenna portion becomes critical if radiation from the line is to be held to a minimum. Such an antenna system for multiband operation is the *end-fed Zepp* or *Zepp-fed* antenna shown in **Fig 3**. The antenna length is made $\lambda/2$ long at the lowest operating frequency. (This name came about because the first documented use of this sort of antennas was on the *Zeppelin* airships.) The feeder length can be anything that is convenient, but feeder lengths that are multiples of $\lambda/4$ generally give trouble with parallel currents and radiation from the feeder portion of the system. The feeder can be an open-wire line of #14 solid

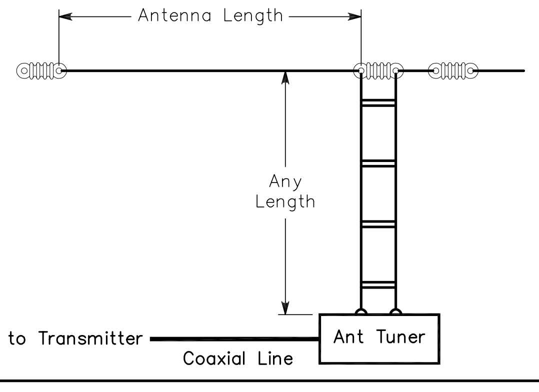

Fig 3—An end-fed Zepp antenna for multiband use.

copper wire spaced 4 or 6 inches with ceramic or plastic spacers. Open-wire TV line (not the type with a solid web of dielectric) is a convenient type to use. This type of line is available in approximately 300- and 450-Ω characteristic impedances.

If you have room for only a 67-foot flat top and yet want to operate in the 3.5-MHz band, the two feeder wires can be tied together at the transmitter end and the entire system treated as a random-length wire fed directly, as in Fig 1. The simplest precaution against parallel currents that could cause feed-line radiation is to use a feeder length that is not a multiple of λ/4. An antenna tuner can be used to provide multiband coverage with an end-fed antenna with any length of open-wire feed line, as shown in Fig 3.

CENTER-FED ANTENNAS

The simplest and most flexible (and also least expensive) all-band antennas are those using open-wire parallel-conductor feeders to the center of the antenna, as in **Fig 4**. Because each half of the flat top is the same length, the feeder currents will be balanced at all frequencies unless, of course, unbalance is introduced by one half of the antenna being closer to ground (or a grounded object) than the other. For best results and to maintain feed-current balance, the feeder should run away at right angles to the antenna, preferably for at least λ/4.

Center feed is not only more desirable than end feed because of inherently better balance, but generally also results in a lower standing wave ratio on the transmission line, provided a parallel-conductor line having a characteristic impedance of 450 to 600 Ω is used. TV-type open-wire line is satisfactory for all but possibly high power installations (over 500 W), where heavier wire and wider spacing is desirable to handle the larger currents and voltages.

The length of the antenna is not critical, nor is the length of the line. As mentioned earlier, the length of the antenna can be considerably less than λ/2 and still be very effective. If the overall length is at least λ/4 at the lowest frequency, a quite usable system will result. The only difficulty that may exist with this type of system is the matter of coupling the antenna-system load to the transmitter. Most modern transmitters are designed to work into a 50-Ω coaxial load. With this type of antenna system a coupling network (an antenna tuner) is required.

Feed-Line Radiation

The preceding sections have pointed out means of reducing or eliminating feed-line radiation. However, it should be emphasized that any radiation from a transmission line is not "lost" energy and is not necessarily harmful. Whether or not feed-line radiation is important depends entirely on the antenna system being used. For example, feed-line radiation is not desirable when a directive array is being used. Such feed-line radiation can distort the desired pattern of such an array, producing responses in unwanted directions. In other words, you want radiation only from the directive array, rather than from the directive array and the feed line. See Chapter 26, Coupling the Line to the Antenna, for a detailed discussion of this topic.

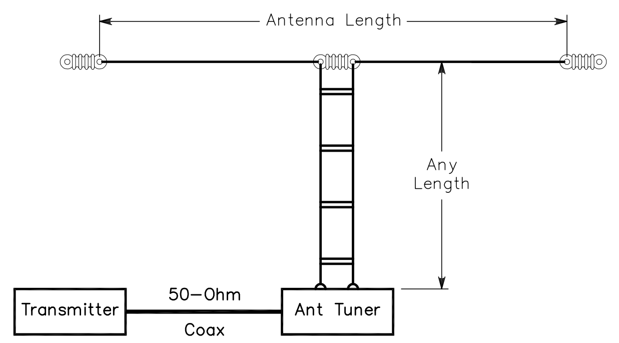

Fig 4—A center-fed antenna system for multiband use.

On the other hand, in the case of a multiband dipole where general coverage is desired, if the feed line happens to radiate, such energy could actually have a desirable effect. Antenna purists may dispute such a premise, but from a practical standpoint where you are not concerned with a directive pattern, much time and labor can be saved by ignoring possible transmission-line radiation.

THE 135-FOOT, 80 TO 10-METER DIPOLE

As mentioned previously, one of the most versatile antennas around is a simple dipole, center-fed with open-wire transmission line and used with an antenna tuner in the shack. A 135-foot long dipole hung horizontally between two trees or towers at a height of 50 feet or higher works very well on 80 through 10 meters. Such an antenna system has significant gain at the higher frequencies.

Flattop or Inverted-V Configuration?

There is no denying that the inverted-V mounting configuration (sometimes called a *drooping dipole*) is very convenient, since it requires only a single support. The flattop configuration, however, where the dipole is mounted horizontally, gives more gain at the higher frequencies. **Fig 5** shows the 80-meter azimuth and elevation patterns for two 135-foot long dipoles. The first is mounted as a flattop at a height of 50 feet over flat ground with a conductivity of 5 mS/m and a dielectric constant of 13, typical for average soil. The second dipole uses the same length of wire, with the center apex at 50 feet and the ends drooped down to be suspended 10 feet off the ground. This height is sufficient so that there is no danger to passersby from RF burns.

At 3.8 MHz, the flattop dipole about 4 dB more peak gain than its drooping cousin. On the other hand, the inverted-V configuration gives a pattern that is more omnidirectional than the flattop dipole, which has nulls off the ends of the wire. Omnidirectional coverage may be more important to net operators, for example, than maximum gain.

Fig 6 shows the azimuth and elevation patterns for the same two antenna configurations, but this time at 14.2 MHz. The flattop dipole has developed four distinct lobes at a 10° elevation angle, an angle typical for 20-meter skywave communication. The peak elevation angle gain of 9.4 dBi occurs at about 17° for a height of 50 feet above flat ground for the flattop dipole. The inverted-V configuration is again nominally more omnidirectional, but the peak gain is down some 6 dB from the flattop.

The situation gets even worse in terms of peak gain at 28.4 MHz for the inverted-V configuration. Here the peak gain is down about 8 dB from that produced by the flattop dipole, which exhibits eight lobes at this frequency with a maximum gain of 10.5 dBi at about 7° elevation. See the comparisons in **Fig 7**.

Whatever configuration you choose to mount the 135-foot dipole, you will want to feed it with some sort of low-loss open-wire transmission line. So-called *window* 450-Ω ladder line is popular for this application. Be sure to twist the line about three or four turns per foot to keep it from twisting excessively in the wind. Make sure also that you provide some mechanical support for the line at the junction with the dipole

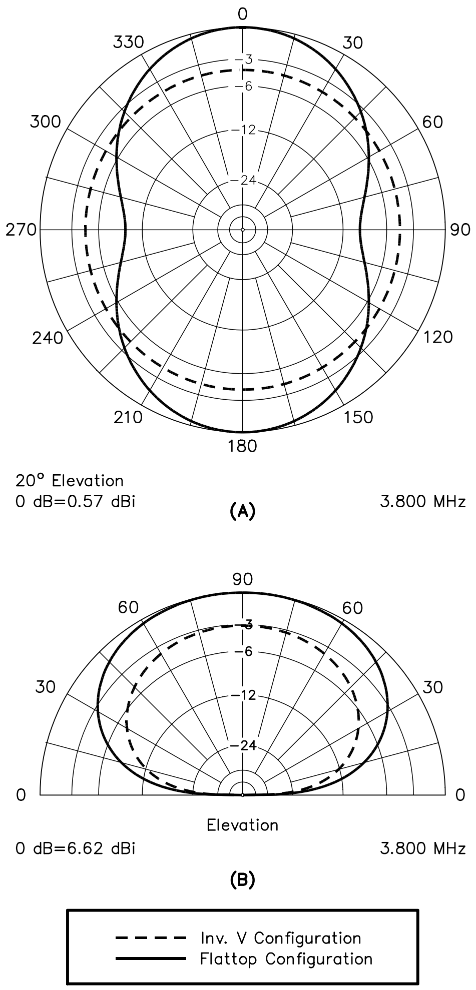

Fig 5—Patterns on 80 meters for 135-foot, center-fed dipole erected as a horizontal flattop dipole at 50 feet, compared with the same dipole installed as an inverted V with the apex at 50 feet and the ends at 10 feet. The azimuth pattern is shown at A, where the dipole wire lies in the 90° to 270° plane. At B, the elevation pattern, the dipole wire comes out of the paper at a right angle. On 80 meters, the patterns are not markedly different for either flattop or inverted-V configuration.

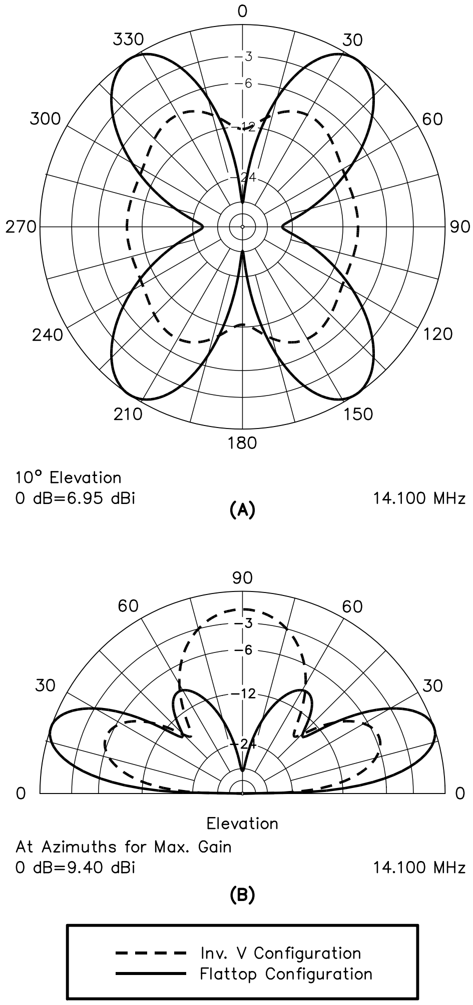

Fig 6—Patterns on 20 meters for two 135-foot dipoles. One is mounted horizontally as a flattop and the other as an inverted V with 120° included angle between the legs. The azimuth pattern is shown in A and the elevation pattern is shown in B. The inverted V has about 6 dB less gain at the peak azimuths, but has a more uniform, almost omnidirectional, azimuthal pattern. In the elevation plane, the inverted V has a fat lobe overhead, making it a somewhat better antenna for local communication, but not quite so good for DX contacts at low elevation angles.

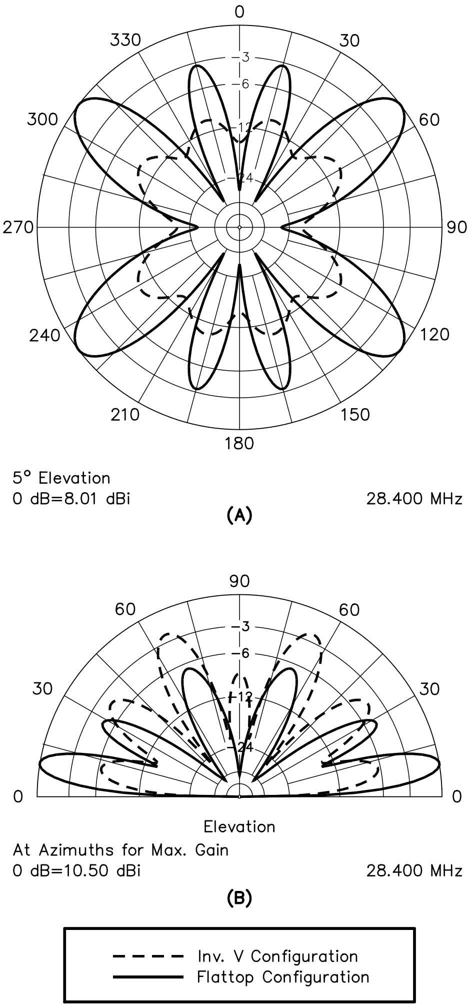

Fig 7—Patterns on 10 meters for same antenna configurations as in Figs 7 and 8. Once again, the inverted-V configuration yields a more omnidirectional pattern, but at the expense of almost 8 dB less gain than the flattop configuration at its strongest lobes.

wires. This will prevent flexing of the transmission-line wire, since excessive flexing will result in breakage.

THE G5RV MULTIBAND ANTENNA

A multiband antenna that does not require a lot of space, is simple to construct, and is low in cost is the G5RV. Designed in England by Louis Varney (G5RV) some years ago, it has become quite popular in the US. The G5RV design is shown in **Fig 8**. The antenna may be used from 3.5 through 30 MHz. Although some amateurs claim it may be fed directly with 50-Ω coax on several amateur bands with a low SWR, Varney himself recommended the use of an antenna tuner on bands other than 14 MHz (see Bibliography). In fact, an analysis of the G5RV feed-point impedance shows there is *no* length of balanced line of *any* characteristic impedance that will transform the terminal impedance to the 50 to 75-Ω range on all bands. (Low SWR indication with coax feed and no matching network on bands other than 14 MHz may indicate excessive losses in the coaxial line.)

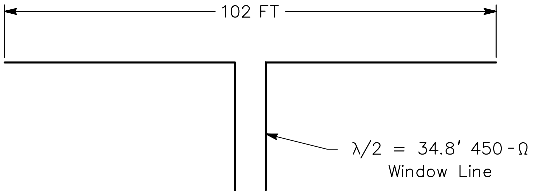

Fig 8—The G5RV multiband antenna covers 3.5 through 30 MHz. Although many amateurs claim it may be fed directly with 50-Ω coax on several amateur bands, Louis Varney, its originator, recommends the use of a matching network on bands other than 14 MHz.

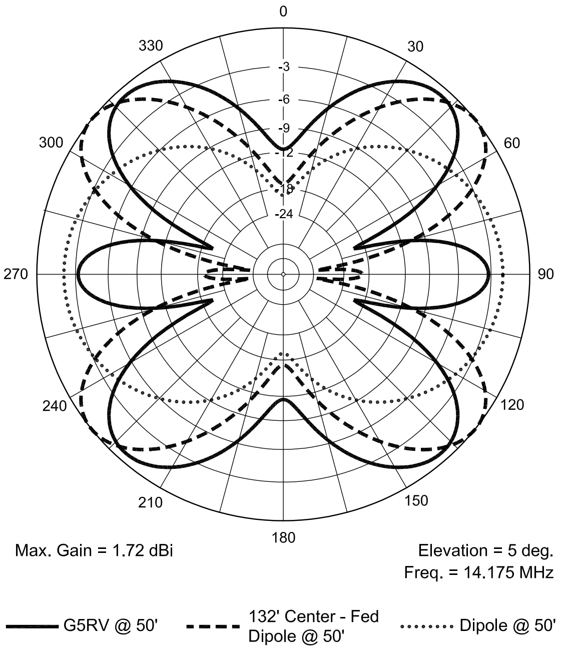

Fig 9—Azimuth pattern at a 5° takeoff angle for a 102-foot long, 50-foot high G5RV dipole (solid line). For comparison, the response for a 132-foot long, center-fed dipole at 50 feet height (dashed line) and a 33-foot long half wave 20-meter dipole at 50 feet (dotted line) are also shown. The longest antenna exhibits about 0.5 dB more gain than the G5RV, although the response is more omnidirectional for the G5RV—an advantage for a wire antenna that is not usually rotatable.

Fig 9 shows the 20-meter azimuthal pattern for a G5RV at a height of 50 feet over flat ground, at an elevation angle of 5° that is suitable for DX work. For comparison, the response for two other antennas is also shown in Fig 9—a standard half wave 20-meter dipole at 50 feet and a 132-foot long center-fed dipole at 50 feet. The G5RV on 20 meters is, of course, longer than a standard half wave dipole and it exhibits about 2 dB more gain compared to that dipole. With four lobes making it look rather like a four-leaf clover, the azimuth pattern is more omnidirectional than the two-lobed dipole. The 132-foot center-fed dipole is longer than the G5RV and it has about 0.5 dB more gain than the G5RV, also exhibiting four major lobes, along with two strong minor lobes in the plane of the wire. Overall, the azimuthal response for the G5RV is more omnidirectional than the comparison antennas.

The G5RV patterns for other frequencies are similar to those shown for the 135-foot dipole previously for other frequencies. Incidentally, you may be wondering why a 132-foot dipole is shown in Fig 9, rather than the 135-foot dipole described earlier. The 132-foot overall length describes another antenna that we'll discus in the next section on Windom antennas.

The portion of the G5RV antenna shown as horizontal in Fig 8 may also be installed in an inverted-V dipole arrangement, subject to the same loss of peak gain mentioned above for the 135-foot dipole. Or instead, up to ⅙ of the total length of the antenna at each end may be dropped vertically, semi-vertically, or bent at a convenient angle to the main axis of the antenna, to cut down on the requirements for real estate.

THE WINDOM ANTENNA

An antenna that enjoyed popularity in the 1930s and into the 1940s was what we now call the *Windom*. It was known at the time as a "single-feeder Hertz" antenna, after being described in Sep 1929 *QST* by Loren G. Windom, W8GZ (see Bibliography).

The Windom antenna, shown in **Fig 10**, is fed with a single wire, attached approximately 14% off center. In theory, this location provides a match for the single-wire transmission line, which is worked against an earth ground. Because the single-wire feed line is not inherently well balanced and because it is brought to the operating position, "RF in the shack" and a potential radiation hazard may be experienced with this antenna.

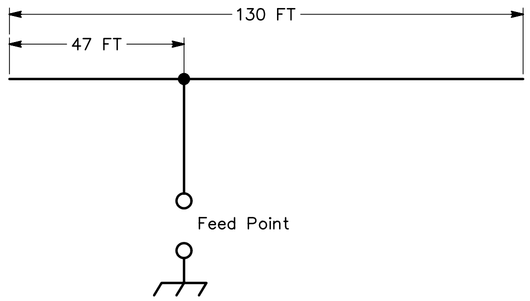

Fig 10—The Windom antenna, cut for a fundamental frequency of 3.75 MHz. The single-wire feeder, connected 14% off center, is brought into the station and the system is fed against ground. The antenna is also effective on its harmonics.

Later variations of the off-center fed Windom moved the attachment point slightly to accommodate balanced 300-Ω ribbon line. One relatively recent variation is called the "Carolina Windom," apparently because two of the designers, Edgar Lambert, WA4LVB, and Joe Wright, W4UEB, lived in coastal North Carolina (the third, Jim Wilkie, WY4R, lived

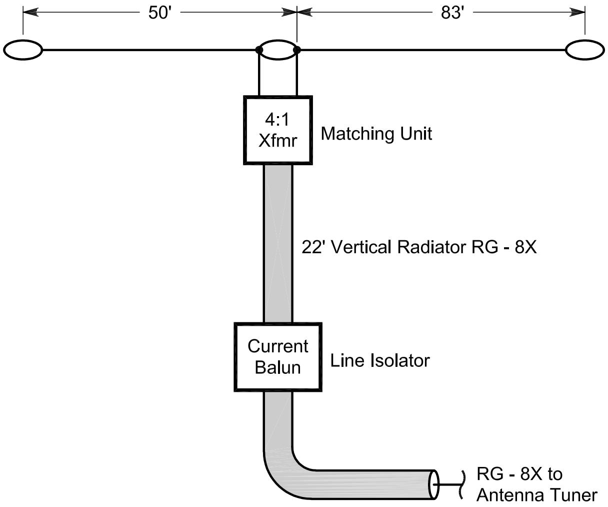

Fig 11—Layout for flattop "Carolina Windom" antenna.

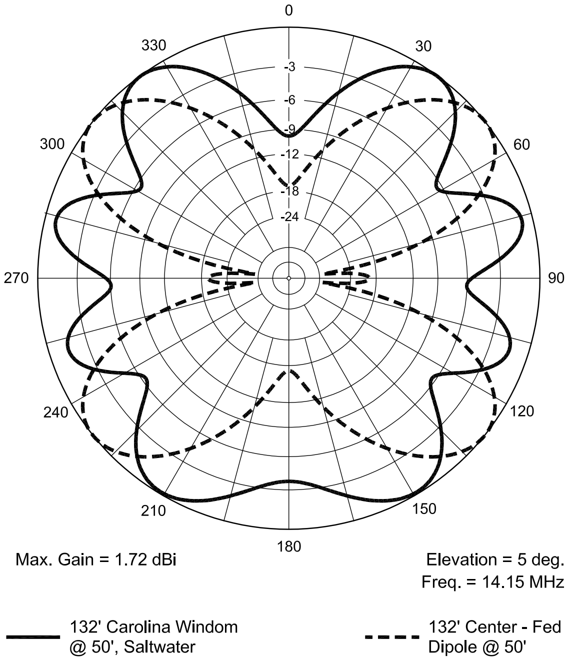

Fig 12—20-meter azimuth patterns for a 132-foot long off-center fed Carolina Windom and a 132-foot long center-fed flattop dipole on 20 meters, both at a height of 50 feet above saltwater. The response for the Carolina Windom is more omnidirectional because the vertically polarized radiation from the 22-foot long vertical RG-8X coax fills in the deep nulls.

in nearby Norfolk, Virginia). One of the interesting parts about the Carolina Windom is that it turns a potential disadvantage—feed line radiation—into a potential advantage.

Fig 11 is a diagram of a flattop Carolina Windom, which uses a 50-foot wire joined with an 83-foot wire at the feed-point insulator. This resembles the layout shown in Fig 10 for the original W8GZ Windom. The "Vertical Radiator" for the Carolina Windom is a 22-foot piece of RG-8X coax, with a "Line Isolator" (current-type choke balun) at the bottom end and a 4:1 "Matching Unit" at the top. The system takes advantage of the asymmetry of the horizontal wires to induce current onto the braid of the vertical coax section. Note that the matching unit is a voltage-type balun transformer, which purposely does not act like a common-mode current choking balun. You must use an antenna tuner with this system to present a 1:1 SWR to the transmitter on the amateur bands from 80 through 10 meters.

The radiation resulting from current induced onto the 22-foot vertical coax section tends to fill in the deep nulls that would be present if the 132-feet of horizontal wire were symmetrically center fed. Over saltwater, the vertical radiator can give significant gain at the low elevation angles needed for DX work. Indeed, field reports for the Carolina Windom are most impressive for stations located near or on saltwater. Over average soil the advantage of the additional vertically polarized component is not quite so evident. **Fig 12** compares a 50-foot high Carolina Windom on 14 MHz over saltwater to a 50-foot high, 132-foot long, flattop center-fed dipole. The Carolina Windom has a more omnidirectional azimuthal pattern, a desirable characteristic in a 132-foot long wire antenna that is not normally rotated to favor different directions.

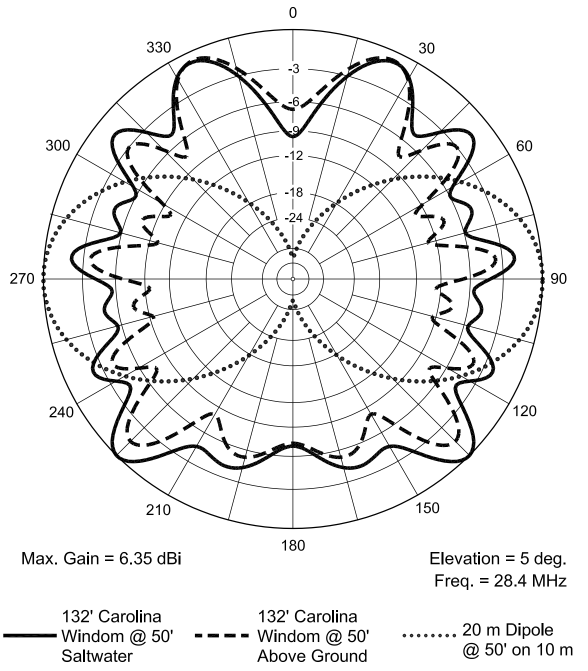

Fig 13—10-meter azimuthal responses for a 132-foot long, 50-foot high Carolina Windom over saltwater (solid line) and over average ground (dashed line), compared to that for a 20-meter half-wave dipole at 50 feet (dotted line).

Another advantage of the Carolina Windom over a traditional Windom is that the coax feed line hanging below the common-mode current choke does not radiate, meaning that there will be less "RF in the shack." Since the feed line is not always operating at a low SWR on various ham bands, use the minimum length of feed coax possible to hold down losses in the coax.

Fig 13 shows the azimuth responses for a 50-foot long flattop Carolina Windom on 28.4 MHz over saltwater and over average soil. The pattern for a 50-foot high, flattop 20-meter dipole operated on 28.4 MHz is also shown, since this 20-meter dipole can also be used as a multiband antenna, when fed with open-wire transmission line rather than with coax. Again, the Carolina Windom exhibits a more omnidirectional pattern, even if the pattern is somewhat lopsided at the bottom.

MULTIPLE-DIPOLE ANTENNAS

The antenna system shown in **Fig 14** consists of a group of center-fed dipoles, all connected in parallel at the point where the transmission line joins them. The dipole elements are *stagger-tuned*. That is, they are individually cut to be λ/2 at different frequencies. Chapter 9, Broadband Antenna Matching, discusses stagger tuning of dipole antennas to attain a low SWR across a broad range of frequencies. An extension of the stagger tuning idea is to construct multiwire dipoles cut for different bands.

In theory, the 4-wire antenna of Fig 14 can be used with a coaxial feeder on five bands. The four wires are prepared as parallel-fed dipoles for 3.5, 7, 14, and 28 MHz. The 7-MHz dipole can be operated on its 3rd harmonic for 21-MHz operation to cover a fifth band. However, in practice it has been found difficult to get a good match to coaxial line

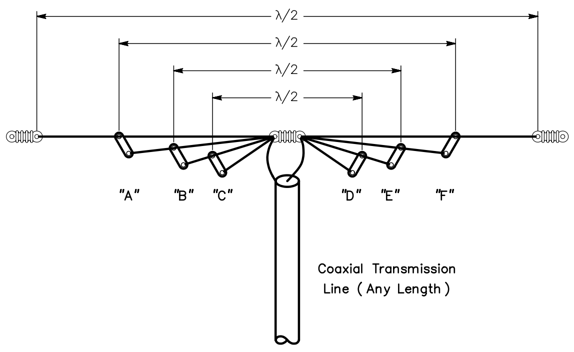

Fig 14—Multiband antenna using paralleled dipoles all connected to a common low-impedance transmission line. The half-wave dimensions may be either for the centers of the various bands or selected to fit favorite frequencies in each band. The length of a half wave in feet is 468/frequency in MHz, but because of interaction among the various elements, some pruning for resonance may be needed on each band.

on all bands. The λ/2 resonant length of any one dipole in the presence of the others is not the same as for a dipole by itself due to interaction, and attempts to optimize all four lengths can become a frustrating procedure. The problem is compounded because the optimum tuning changes in a different antenna environment, so what works for one amateur may not work for another. Even so, many amateurs with limited antenna space are willing to accept the mismatch on some bands just so they can operate on those frequencies using a single coax feed line.

Since this antenna system is balanced, it is desirable to use a balanced transmission line to feed it. The most desirable type of line is 75-Ω transmitting twin-lead. However, either 52-Ω or 75-Ω coaxial line can be used. Coax line introduces some unbalance, but this is tolerable on the lower frequencies. An alternative is to use a balun at the feed point, fed with coaxial cable.

The separation between the dipoles for the various frequencies does not seem to be especially critical. One set of wires can be suspended from the next larger set, using insulating spreaders (of the type used for feeder spreaders) to give a separation of a few inches. Users of this antenna often run some of the dipoles at right angles to each other to help reduce interaction. Some operators use inverted-V-mounted dipoles as guy wires for the mast that supports the antenna system.

An interesting method of construction used successfully by Louis Richard, ON4UF, is shown in **Fig 15**. The antenna has four dipoles (for 7, 14, 21 and 28 MHz) constructed from 300-Ω ribbon transmission line. A single length of ribbon makes two dipoles. Thus, two lengths, as shown in the sketch, serve to make dipoles for four bands. Ribbon with copper-clad steel conductors (Amphenol type 14-022) should be used because all of the weight, including that of the feed line, must be supported by the uppermost wire.

Two pieces of ribbon are first cut to a length suitable for the two halves of the longest dipole. Then one of the conductors in each piece is cut to proper length for the next band higher in frequency. The excess wire and insulation is stripped away. A second pair of lengths is prepared in the same manner, except that the lengths are appropriate for the next two higher frequency bands.

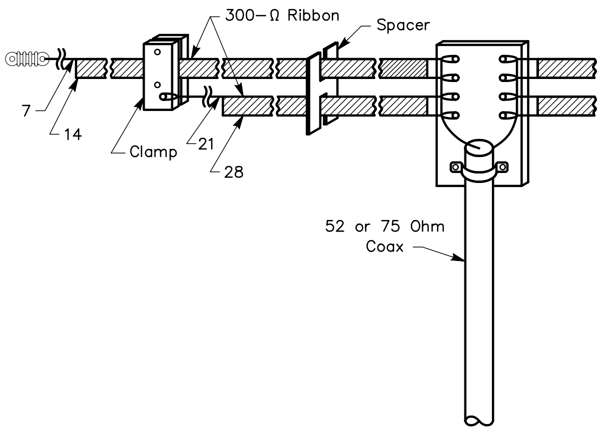

Fig 15—Sketch showing how the twin-lead multiple-dipole antenna system is assembled. The excess wire and insulation are stripped away.

A piece of thick polystyrene sheet drilled with holes for anchoring each wire serves as the central insulator. The shorter pair of dipoles is suspended the width of the ribbon below the longer pair by clamps also made of poly sheet. Intermediate spacers are made by sawing slots in pieces of poly sheet so they will fit the ribbon snugly.

The multiple-dipole principle can also be applied to vertical antennas. Parallel or fanned λ/4 elements of wire or tubing can be worked against ground or tuned radials from a common feed point.

OFF-CENTER-FED DIPOLES

Fig 16 shows an off-center-fed or *OCF* dipole. Because it is similar in appearance to the Windom of Fig 12, this antenna is often mistakenly called a “Windom,” or sometimes a “coax-fed Windom.” The two antennas are not the same, since the Windom is worked against its image in the ground, while one leg is worked against the other in the OCF dipole.

It is not necessary to feed a dipole antenna at its center, although doing so will allow it to be operated with a relatively low feed-point impedance on its fundamental and *odd* harmonics. (For example, a 7-MHz center-fed half-wave dipole can also be used for 21-MHz operation.) By contrast, the OCF dipole of Fig 16, fed ⅓ of its length from one end, may be used on its fundamental and *even* harmonics. Its free-space antenna-terminal impedance at 3.5, 7 and 14 MHz is on the order of 150 to 200 Ω. A 1:4 step-up transformer at the feed point should offer a reasonably good match to 50- or 75-Ω line, although some commercially made OCF dipoles use a 1:6 transformer.

At the 6th harmonic, 21 MHz, the antenna is three wavelengths long and fed at a voltage loop (maximum), instead of a current loop. The feed-point impedance at this frequency

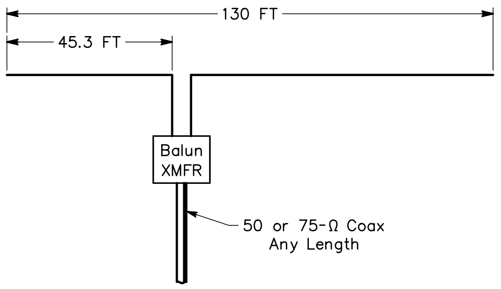

Fig 16—The off-center-fed (OCF) dipole for 3.5, 7 and 14 MHz. A 1:4 or 1:6 step-up current balun is used at the feed point.

is high, a few thousand ohms, so the antenna is unsuitable for use on this band.

Balun Requirements

Because the OCF dipole is not fed at the center of the radiator, the RF impedance paths of the two wires at the feed point are unequal. If the antenna is fed directly with coax (or a balanced line), or if a voltage step-up transformer is used, then voltages of equal magnitude (but opposite polarity) are applied to the wires at the feed point. Because of unequal impedances, the resulting antenna currents flowing in the two wires will not be equal. This also means that antenna current can flow on the feeder—on the *outside* of a coaxial line. (You may recall that this is how the Carolina Windom works, actually inducing current onto a carefully chosen length of coax, choked at its bottom end, so that it acts as a vertical radiator.)

How much current flows on the coax shield depends on the impedance of the RF current path down the outside of the feed line. In general, this is not a desirable situation. To prevent radiation, equal *currents* are required at the feed point, with the same current flowing in and out of the short leg as in and out of the long leg of the radiator. A *current* or *choke* type of balun provides just such operation. (Current baluns are discussed in detail in Chapter 26, Coupling the Line to the Antenna.)

Trap Antennas

By using tuned circuits of appropriate design strategically placed in a dipole, the antenna can be made to show what is essentially fundamental resonance at a number of different frequencies. The general principle is illustrated by **Fig 17**.

Even though a trap-antenna arrangement is a simple one, an explanation of how a trap antenna works can be elusive. For some designs, traps are resonated in our amateur bands, and for others (especially commercially made antennas) the traps are resonant far outside any amateur band.

A trap in an antenna system can perform either of two functions, depending on whether or not it is resonant at the operating frequency. A familiar case is where the trap is parallel-resonant in an amateur band. For the moment, let us assume that dimension A in Fig 17 is 32 feet and that each L/C combination is resonant in the 7-MHz band. Because of its parallel resonance, the trap presents a high impedance at that point in the antenna system. The electrical effect at 7 MHz is that the trap behaves as an insulator. It serves to divorce the outside ends, the B sections, from the antenna. The result is easy to visualize—we have an antenna system that is resonant in the 7-MHz band. Each 33-foot section (labeled A in the drawing) represents λ/4, and the trap behaves as an insulator. We therefore have a full-size 7-MHz antenna.

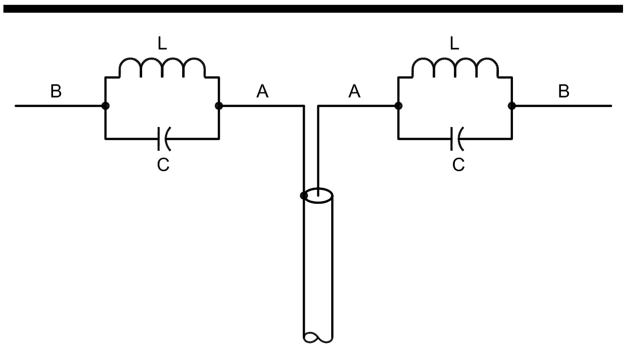

Fig 17—A trap dipole antenna. This antenna may be fed with 50-Ω coaxial line. Depending on the L/C ratio of the trap elements and the lengths chosen for dimensions A and B, the traps may be resonant either in an amateur band or at a frequency far removed from an amateur band for proper two-band antenna operation.

The second function of a trap, obtained when the frequency of operation is *not* the resonant frequency of the trap, is one of electrical loading. If the operating frequency is below that of trap resonance, the trap behaves as an inductor; if above, as a capacitor. Inductive loading will electrically lengthen the antenna, and capacitive loading will electrically shorten the antenna.

Let's carry our assumption a bit further and try using the antenna we just considered at 3.5 MHz. With the traps resonant in the 7-MHz band, they will behave as inductors when operation takes place at 3.5 MHz, electrically lengthening the antenna. This means that the total length of sections A and B (plus the length of the inductor) may be something less than a physical λ/4 for resonance at 3.5 MHz. Thus, we have a two-band antenna that is shorter than full size on the lower frequency band. But with the electrical loading provided by the traps, the overall electrical length is λ/2. The total antenna length needed for resonance in the 3.5-MHz band will depend on the L/C ratio of the trap elements.

The key to trap operation off resonance is its L/C ratio, the ratio of the value of L to the value of C. At resonance, however, within practical limitations the L/C ratio is immaterial as far as electrical operation goes. For example, in the antenna we've been discussing, it would make no difference for 7-MHz operation whether the inductor were 1 µH and the capacitor were 500 pF (the reactances would be just below 45 Ω at 7.1 MHz), or whether the inductor were 5 µH and the capacitor 100 pF (reactances of approximately 224 Ω at 7.1 MHz). But the choice of these values will make a significant difference in the antenna size for resonance at 3.5 MHz. In the first case, where the L/C ratio is 2000, the necessary length of section B of the antenna for resonance at 3.75 MHz would be approximately 28.25 feet. In the second case, where the L/C ratio is 50,000, this length need be only 24.0 feet, a difference of more than 15%.

The above example concerns a two-band antenna with trap resonance at one of the two frequencies of operation. On each of the two bands, each half of the dipole operates as an electrical λ/4. However, the same band coverage can be obtained with a trap resonant at, say, 5 MHz, a frequency quite removed from either amateur band. With proper selection of the L/C ratio and the dimensions for A and B, the trap will act to shorten the antenna electrically at 7 MHz and lengthen it electrically at 3.5 MHz. Thus, an antenna that is intermediate in physical length between being full size on 3.5 MHz and full size on 7 MHz can cover both bands, even though the trap is not resonant at either frequency. Again, the antenna operates with electrical λ/4 sections. Note that such non-resonant traps have less RF current flowing in the trap components, and hence trap losses are less than for resonant traps.

Additional traps may be added in an antenna section to cover three or more bands. Or a judicious choice of dimensions and the L/C ratio may permit operation on three or more bands with just a pair of identical traps in the dipole.

An important point to remember about traps is this. If the operating frequency is below that of trap resonance, the trap behaves as an inductor; if above, as a capacitor. The above discussion is based on dipoles that operate electrically as λ/2 antennas. This is not a requirement, however. Elements may be operated as electrical 3/2 λ, or even 5/2 λ, and still present a reasonable impedance to a coaxial feeder. In trap antennas covering several HF bands, using electrical lengths that are odd multiples of λ/2 is often done at the higher frequencies.

To further aid in understanding trap operation, let's now choose trap L and C components that each have a reactance of 20 Ω at 7 MHz. Inductive reactance is directly proportional to frequency, and capacitive reactance is inversely proportional. When we shift operation to the 3.5-MHz band, the inductive reactance becomes 10 Ω, and the capacitive reactance becomes 40 Ω. At first thought, it may seem that the trap would become capacitive at 3.5 MHz with a higher capacitive reactance, and that the extra capacitive reactance would make the antenna electrically shorter yet. Fortunately, this is not the case. The inductor and the capacitor are connected in parallel with each other.

$$Z = \frac{-jX_L X_C}{X_L + X_C} \qquad \text{(Eq 1)}$$

where j indicates a reactive impedance component, rather than resistive. A positive result indicates inductive reactance, and a negative result indicates capacitive. In this 3.5-MHz case, with 40 Ω of capacitive reactance and 10 Ω of inductive, the equivalent series reactance is 13.3 Ω inductive. This inductive loading lengthens the antenna to an electrical λ/2 overall at 3.5 MHz, assuming the B end sections in Fig 17 are of the proper length.

With the above reactance values providing resonance at 7-MHz, X_L equals X_C, and the theoretical series equivalent is infinity. This provides the insulator effect, divorcing the ends.

At 14 MHz, where $X_L = 40\ \Omega$ and $X_C = 10\ \Omega$, the resultant series equivalent trap reactance is 13.3 Ω capacitive. If the total physical antenna length is slightly longer than 3/2 λ at 14 MHz, this trap reactance at 14 MHz can be used to shorten the antenna to an electrical 3/2 λ. In this way, 3-band operation is obtained for 3.5, 7 and 14 MHz with just one pair of identical traps. The design of such a system is not straightforward, however, for any chosen L/C ratio for a given total length affects the resonant frequency of the antenna on both the 3.5 and 14-MHz bands.

Trap Losses

Since the tuned circuits have some inherent losses, the efficiency of a trap system depends on the unloaded Q values of the tuned circuits. Low-loss (high-Q) coils should be used, and the capacitor losses likewise should be kept as low as possible. With tuned circuits that are good in this respect—comparable with the low-loss components used in transmitter tank circuits, for example—the reduction in efficiency compared with the efficiency of a simple dipole is small, but tuned circuits of low unloaded Q can lose an appreciable portion of the power supplied to the antenna.

The commentary above applies to traps assembled from conventional components. The important function of a trap that is resonant in an amateur band is to provide a high isolating impedance, and this impedance is directly proportional to Q. Unfortunately, high Q restricts the antenna bandwidth, because the traps provide maximum isolation only at trap resonance.

FIVE-BAND W3DZZ TRAP ANTENNA

C. L. Buchanan, W3DZZ, created one of the first trap antennas for the five pre-1979 WARC amateur bands from 3.5 to 30 MHz. Dimensions are given in **Fig 18**. Only one set of traps is used, resonant at 7 MHz to isolate the inner (7-MHz) dipole from the outer sections. This causes the overall system to be resonant in the 3.5-MHz band. On 14, 21 and 28 MHz the antenna works on the capacitive-reactance principle just outlined. With a 75-Ω twin-lead feeder, the SWR with this antenna is under 2:1 throughout the three highest frequency

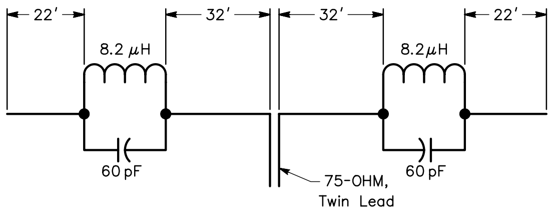

Fig 18—Five-band (3.5, 7, 14, 21 and 28 MHz) trap dipole for operation with 75-Ω feeder at low SWR (C. L. Buchanan, W3DZZ). The balanced (parallel-conductor) line indicated is desirable, but 75-Ω coax can be substituted with some sacrifice of symmetry in the system. Dimensions given are for resonance (lowest SWR) at 3.75, 7.2, 14.15 and 29.5 MHz. Resonance is very broad on the 21-MHz band, with SWR less than 2:1 throughout the band.

bands, and the SWR is comparable with that obtained with similarly fed simple dipoles on 3.5 and 7 MHz.

Trap Construction

Traps frequently are built with coaxial aluminum tubes (usually with polystyrene tubing in-between them for insulation) for the capacitor, with the coil either self-supporting or wound on a form of larger diameter than the tubular capacitor. The coil is then mounted coaxially with the capacitor to form a unit assembly that can be supported at each end by the antenna wires. In another type of trap devised by William J. Lattin, W4JRW (see Bibliography at the end of this chapter), the coil is supported inside an aluminum tube and the trap capacitor is obtained in the form of capacitance between the coil and the outer tube. This type of trap is inherently weatherproof.

A simpler type of trap, easily assembled from readily available components, is shown in **Fig 19**. A small transmitting-type ceramic "doorknob" capacitor is used, together with a length of commercially available coil material, these being supported by an ordinary antenna strain insulator. The circuit constants and antenna dimensions differ slightly from those of Fig 18, in order to bring the antenna resonance points closer to the centers of the various phone bands. Construction data are given in **Fig 20**. If a 10-turn length of inductor is used, a half turn from each end may be used to slip through the anchor holes in the insulator to act as leads.

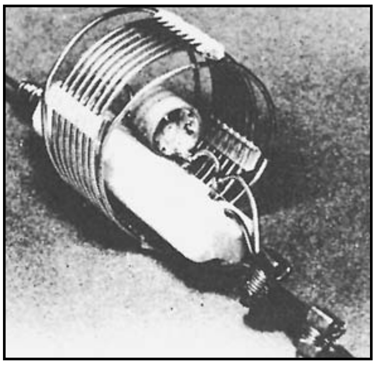

Fig 19—Easily constructed trap for wire antennas (A. Greenburg, W2LH). The ceramic insulator is 4¼ inches long (Birnback 688). The clamps are small service connectors available from electrical supply and hardware stores (Burndy KS90 servits).

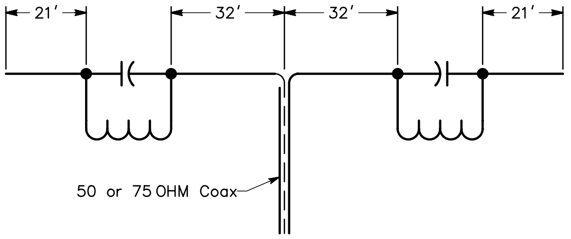

Fig 20—Layout of multiband antenna using traps constructed as shown in Fig 21. The capacitors are 100 pF each, transmitting type, 5000-volt dc rating (Centralab 850SL-100N). Coils are 9 turns of #12 wire, 2½ inches diameter, 6 turns per inch (B&W 3029) with end turns spread as necessary to resonate the traps to 7.2 MHz. These traps, with the wire dimensions shown, resonate the antenna at approximately the following frequencies on each band: 3.9, 7.25, 14.1, 21.5 and 29.9 MHz (based on measurements by W9YJH).

The components used in these traps are sufficiently weatherproof in themselves so that no additional weatherproofing has been found necessary. However, if it is desired to protect them from the accumulation of snow or ice, a plastic cover can be made by cutting two discs of polystyrene slightly larger in diameter than the coil, drilling at the center to pass the antenna wires, and cementing a plastic cylinder on the edges of the discs. The cylinder can be made by wrapping two turns or so of 0.02-inch poly or Lucite sheet around the discs, if no suitable ready-made tubing is available. Plastic drinking glasses and 2-liter soft-drink plastic bottles are easily adaptable for use as impromptu trap covers.

TWO W8NX MULTIBAND, COAX-TRAP DIPOLES

Over the last 60 or 70 years, amateurs have used many kinds of multiband antennas to cover the traditional HF bands. The availability of the 30, 17 and 12-meter bands has expanded our need for multiband antenna coverage.

Two different antennas are described here. The first covers the traditional 80, 40, 20, 15 and 10-meter bands, and the second covers 80, 40, 17 and 12 meters. Each uses the same type of W8NX trap—connected for different modes of operation—and a pair of short capacitive stubs to enhance coverage. The W8NX coaxial-cable traps have two different modes: a high- and a low-impedance mode. The inner-conductor windings and shield windings of the traps are connected in series for both modes. However, either the low- or high-impedance point can be used as the trap's output terminal. For low-impedance trap operation, only the center conductor turns of the trap windings are used. For high-impedance operation, all turns are used, in the conventional manner for a trap. The short stubs on each antenna are strategically sized and located to permit more flexibility in adjusting thc resonant frequencies of the antenna.

80, 40, 20, 15 and 10-Meter Dipole

Fig 21 shows the configuration of the 80, 40, 20, 15 and 10-meter antenna. The radiating elements are made of #14

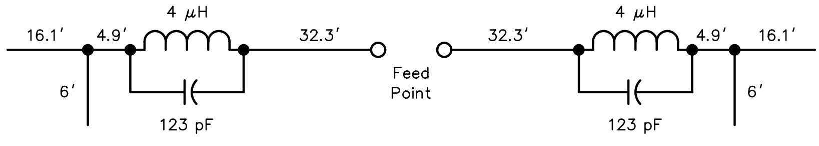

Fig 21—A W8NX multiband dipole for 80, 40, 20, 15 and 10 meters. The values shown (123 pF and 4 μH) for the coaxial-cable traps are for parallel resonance at 7.15 MHz. The low-impedance output of each trap is used for this antenna.

stranded copper wire. The element lengths are the wire span lengths in feet. These lengths do not include the lengths of the pigtails at the balun, traps and insulators. The 32.3-foot-long inner 40-meter segments are measured from the eyelet of the input balun to the tension-relief hole in the trap coil form. The 4.9-foot segment length is measured from the tension-relief hole in the trap to the 6-foot stub. The 16.l-foot outer-segment span is measured from the stub to the eyelet of the end insulator.

The coaxial-cable traps are wound on PVC pipe coil forms and use the low-impedance output connection. The stubs are 6-foot lengths of ⅛-inch stiffened aluminum or copper rod hanging perpendicular to the radiating elements. The first inch of their length is bent 90° to permit attachment to the radiating elements by large-diameter copper crimp connectors. Ordinary #14 wire may be used for the stubs, but it has a tendency to curl up and may tangle unless weighed down at the end. You should feed the antenna with 75-Ω coax cable using a good 1:1 balun.

This antenna may be thought of as a modified W3DZZ antenna due to the addition of the capacitive stubs. The length and location of the stub give the antenna designer two extra degrees of freedom to place the resonant frequencies within the amateur bands. This additional flexibility is particularly helpful to bring the 15 and 10-meter resonant frequencies to more desirable locations in these bands. The actual 10-meter resonant frequency of the original W3DZZ antenna is somewhat above 30 MHz, pretty remote from the more desirable low frequency end of 10 meters.

80, 40, 17 and 12-Meter Dipole

Fig 22 shows the configuration of the 80, 40, 17 and 12-meter antenna. Notice that the capacitive stubs are attached immediately outboard after the traps and are 6.5 feet long, ½ foot longer than those used in the other antenna. The traps are the same as those of the other antenna, but are connected for the high-impedance parallel-resonant output mode. Since only four bands are covered by this antenna, it is easier to fine tune it to precisely the desired frequency on all bands. The 12.4-foot tips can he pruned to a particular 17-meter frequency with little effect on the 12-meter frequency. The stub lengths can be pruned to a particular 12-meter frequency with little effect on the 17-meter frequency. Both such pruning adjustments slightly alter the 80-meter resonant frequency. However, the bandwidths of the antennas are so broad on 17 and 12 meters that little need for such pruning exists. The 40-meter frequency is nearly independent of adjustments to the capacitive stubs and outer radiating tip elements. Like the first antennas, this dipole is fed with a 75-Ω balun and feed line.

Fig 23 shows the schematic diagram of the traps. It explains the difference between the low and high-impedance modes of the traps. Notice that the high-impedance terminal is the output configuration used in most conventional trap applications. The low-impedance connection is made across only the inner conductor turns, corresponding to one-half of the total turns of the trap. This mode steps the trap's impedance down to approximately one-fourth of that of the high-impedance level. This is what allows a single trap design to be used for two different multiband antennas.

Fig 24 is a drawing of a cross-section of the coax trap shown through the long axis of the trap. Notice that the traps are conventional coaxial-cable traps, except for the added low-impedance output terminal. The traps are 8¾ close-spaced turns of RG-59 (Belden 8241) on a 2⅜-inch-OD PVC pipe (schedule 40 pipe with a 2-inch ID) coil form. The forms are 4⅛ inches long. Trap resonant frequency is very sensitive to the outer diameter of the coil form, so check it carefully. Unfortunately, not all PVC pipe is made with the same wall thickness. The trap frequencies should be checked with a dip meter and general-coverage receiver and adjusted to within 50 kHz of the 7150 kHz resonant frequency before installation. One inch is left over at each end of the coil forms to allow for the coax feed-through holes and holes for tension-relief attachment of the antenna radiating elements to the traps. Be sure to seal the ends of the trap coax cable with RTV sealant to prevent moisture from entering the coaxial cable.

Also, be sure that you connect the 32.3-foot wire element at the start of the inner conductor winding of the trap. This avoids detuning the antenna by the stray capacitance of the coaxial-cable shield. The trap output terminal (which has the shield stray capacitance) should be at the outboard side of

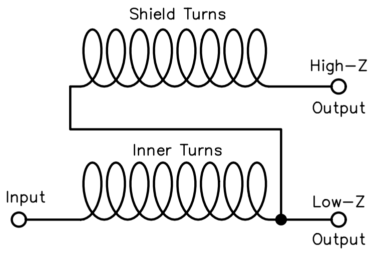

Fig 23—Schematic for the W8NX coaxial-cable trap. RG-59 is wound on a 2⅜-inch OD PVC pipe.

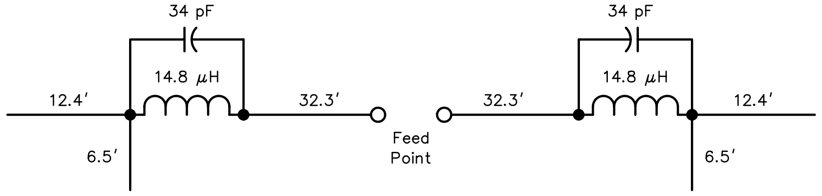

Fig 22—A W8NX multiband dipole for 80, 40, 17 and 12 meters. For this antenna, the high-impedance output is used on each trap. The resonant frequency of the traps is 7.15 MHz.

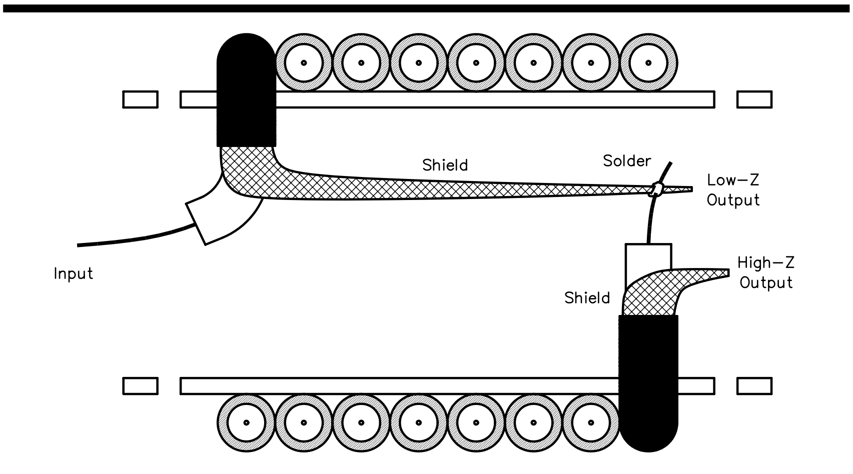

Fig 24—Construction details of the W8NX coaxial-cable trap.

the trap. Reversing the input and output terminals of the trap will lower the 40-meter frequency by approximately 50 kHz, but there will be negligible effect on the other bands.

Fig 25 shows a coaxial-cable trap. Further details of the trap installation are shown in **Fig 26**. This drawing applies specifically to the 80, 40, 20, 15 and l0-meter antenna, which uses the low-impedance trap connections. Notice the lengths of the trap pigtails: 3 to 4 inches at each terminal of the trap. If you use a different arrangement, you must modify the span lengths accordingly. All connections can be made using crimp connectors rather than by soldering. Access to the trap's interior is attained more easily with a crimping tool than with a soldering iron.

Performance

The performance of both antennas has been very satisfactory. W8NX uses the 80, 40, 17 and 12-meter version because it covers 17 and 12 meters. (He has a tribander for 20, 15 and 10 meters.) The radiation pattern on 17 meters is that of a $\frac{3}{2}$-wave dipole. On 12 meters, the pattern is that of a $\frac{5}{2}$-wave dipole. At his location in Akron, Ohio, the antenna runs essentially east and west. It is installed as an inverted V, 40 feet high at the center, with a 120° included angle between the legs. Since the stubs are very short, they radiate little power and make only minor contributions to the radiation patterns. In theory, the pattern has four major lobes on 17 meters, with maxima to the northeast, southeast, southwest and northwest. These provide low-angle radiation into Europe, Africa, South Pacific, Japan and Alaska. A narrow pair of minor broadside lobes provides north and south coverage into Central America, South America and the polar regions.

There are four major lobes on 12 meters, giving nearly end-fire radiation and good low-angle east and west coverage. There are also three pairs of very narrow, nearly broadside, minor lobes on 12 meters, down about 6 dB from the major end-fire lobes. On 80 and 40 meters, the antenna has the usual figure-8 patterns of a half-wave-length dipole.

Both antennas function as electrical half-wave dipoles on 80 and 40 meters with a low SWR. They both function as odd-harmonic current-fed dipoles on their other operating frequencies, with higher, but still acceptable, SWR. The presence of the stubs can either raise or lower the input impedance of the antenna from those of the usual third and fifth harmonic dipoles. Again W8NX recommends that 75-Ω,

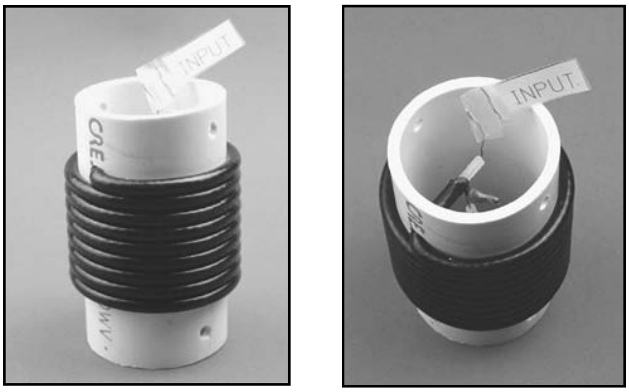

Fig 25—Other views of a W8NX coax-cable trap.

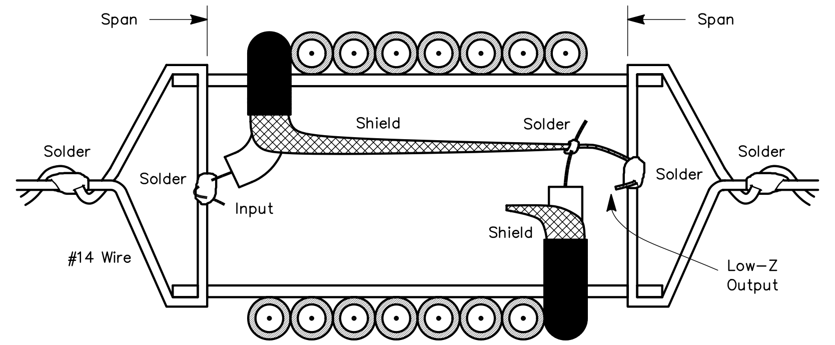

Fig 26—Additional construction details for the W8NX coaxial-cable trap.

rather than 50-Ω, feed line be used because of the generally higher input impedances at the harmonic operating frequencies of the antennas.

The SWR curves of both antennas were carefully measured using a 75 to 50-Ω transformer from Palomar Engineers inserted at the junction of the 75-Ω coax feed line and a 50-Ω SWR bridge. The transformer is required for accurate SWR measurement if a 50-Ω SWR bridge is used with a 75-Ω line. Most 50-Ω rigs operate satisfactorily with a 75-Ω line, although this requires different tuning and load settings in the final output stage of the rig or antenna tuner. The author uses the 75 to 50-Ω transformer only when making SWR measurements and at low power levels. The transformer is rated for 100 W, and when he runs his 1-kW PEP linear amplifier the transformer is taken out of the line.

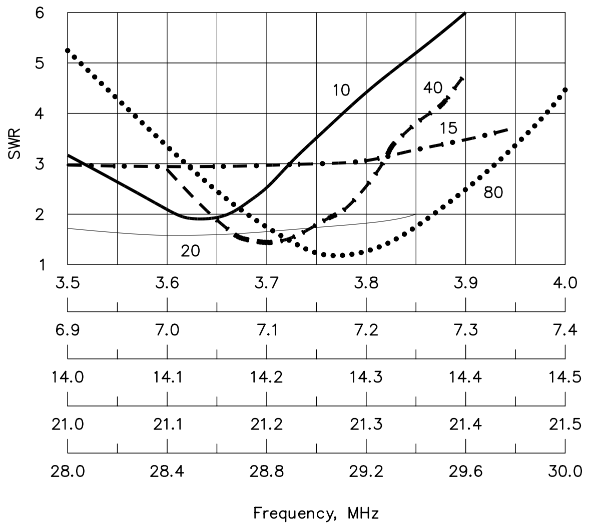

Fig 27—Measured SWR curves for an 80, 40, 20, 15 and 10-meter antenna, installed as an inverted-V with 40-ft apex and 120° included angle between legs.

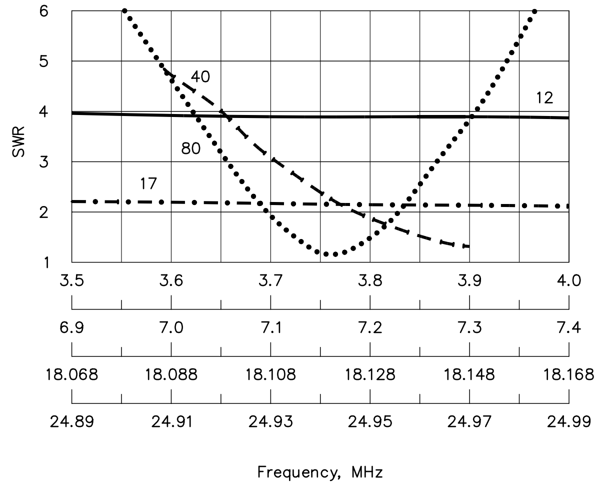

Fig 28—Measured SWR curves for an 80, 40, 17 and 12-meter antenna, installed as an inverted-V with 40-ft apex and 120° included angle between legs.

Fig 27 gives the SWR curves of the 80, 40, 20, 15 and 10-meter antenna. Minimum SWR is nearly 1:1 on 80 meters, 1.5:1 on 40 meters, 1.6:1 on 20 meters, and 1.5:1 on 10 meters. The minimum SWR is slightly below 3:1 on 15 meters. On 15 meters, the stub capacitive reactance combines with the inductive reactance of the outer segment of the antenna to produce a resonant rise that raises the antenna input resistance to about 220 Ω, higher than that of the usual 3/2-wavelength dipole. An antenna tuner may be required on this band to keep a solid-state final output stage happy under these load conditions.

Fig 28 shows the SWR curves of the 80, 40, 17 and 12-meter antenna. Notice the excellent 80-meter performance with a nearly unity minimum SWR in the middle of the band. The performance approaches that of a full-size 80-meter wire dipole. The short stubs and the low-inductance traps shorten the antenna somewhat on 80 meters. Also observe the good 17-meter performance, with the SWR being only a little above 2:1 across the band.

But notice the 12-meter SWR curve of this antenna, which shows 4:1 SWR across the band. The antenna input resistance approaches 300 Ω on this band because the capacitive reactance of the stubs combines with the inductive reactance of the outer antenna segments to give resonant rises in impedance. These are reflected back to the input terminals. These stub-induced resonant impedance rises are similar to those on the other antenna on 15 meters, but are even more pronounced.

Too much concern must not be given to SWR on the feed line. Even if the SWR is as high as 9:1 *no destructively high voltages will exist on the transmission line.* Recall that transmission-line voltages increase as the square root of the SWR in the line. Thus, 1 kW of RF power in 75-Ω line corresponds to 274 V line voltage for a 1:1 SWR. Raising the SWR to 9:1 merely triples the maximum voltage that the line must withstand to 822 V. This voltage is well below the 3700-V rating of RG-11, or the 1700-V rating of RG-59, the two most popular 75-Ω coax lines. Voltage breakdown in the traps is also very unlikely. As will be pointed out later, the operating power levels of these antennas are limited by RF power dissipation in the traps, not trap voltage breakdown or feed-line SWR.

Trap Losses and Power Rating

Table 1 presents the results of trap Q measurements and extrapolation by a two-frequency method to higher frequencies above resonance. W8NX employed an old, but recently calibrated, Boonton Q meter for the measurements. Extrapolation to higher-frequency bands assumes that trap resistance losses rise with skin effect according to the square root of frequency, and that trap dielectric loses rise directly with frequency. Systematic measurement errors are not increased by frequency extrapolation. However, random measurement

Table 1
Trap Q

Frequency (MHz)	*3.8*	*7.15*	*14.18*	*18.1*	*21.3*	*24.9*	*28.6*
High Z out (Ω)	101	124	139	165	73	179	186
Low Z out (Ω)	83	103	125	137	44	149	155

Table 2
Trap Loss Analysis: 80, 40, 20, 15, 10-Meter Antenna

Frequency (MHz)	*3.8*	*7.15*	*14.18*	*21.3*	*28.6*
Radiation Efficiency (%)	96.4	70.8	99.4	99.9	100.0
Trap Losses (dB)	0.16	1.5	0.02	0.01	0.003

Table 3
Trap Loss Analysis: 80, 40, 17, 12-Meter Antenna

Frequency (MHz)	*3.8*	*7.15*	*18.1*	*24.9*
Radiation Efficiency (%)	89.5	90.5	99.3	99.8
Trap Losses (dB)	0.5	0.4	0.03	0.006

errors increase in magnitude with upward frequency extrapolation. Results are believed to be accurate within 4% on 80 and 40 meters, but only within l0 to 15% at 10 meters. Trap Q is shown at both the high- and low-impedance trap terminals. The Q at the low-impedance output terminals is 15 to 20% lower than the Q at the high-impedance output terminals.

W8NX computer-analyzed trap losses for both antennas in free space. Antenna-input resistances at resonance were first calculated, assuming lossless, infinite-Q traps. They were again calculated using the Q values in Table 1. The radiation efficiencies were also converted into equivalent trap losses in decibels. **Table 2** summarizes the trap-loss analysis for the 80, 40, 20, 15 and 10-meter antenna and **Table 3** for the 80, 40,17 and 12-meter antenna.

The loss analysis shows radiation efficiencies of 90% or more for both antennas on all bands except for the 80, 40, 20, 15 and 10-meter antenna when used on 40 meters. Here, the radiation efficiency falls to 70.8%. A 1-kW power level at 90% radiation efficiency corresponds to 50-W dissipation per trap. In W8NX's experience, this is the trap's survival limit for extended key-down operation. SSB power levels of 1 kW PEP would dissipate 25 W or less in each trap. This is well within the dissipation capability of the traps.

When the 80, 40, 20, 15 and 10-meter antenna is operated on 40 meters, the radiation efficiency of 70.8% corresponds to a dissipation of 146 W in each trap when 1 kW is delivered to the antenna. This is sure to burn out the traps—even if sustained for only a short time. Thus, the power should be limited to leas than 300 W when this antenna is operated on 40 meters under prolonged key-down conditions. A 50% CW duty cycle would correspond to a 600-W power limit for normal 40-meter CW operation. Likewise, a 50% duty cycle for 40-meter SSB corresponds to a 600-W PEP power limit for the antenna.

The author knows of no analysis where the burnout wattage rating of traps has been rigorously determined. Operating experience seems to be the best way to determine trap burnout ratings. In his own experience with these antennas, he's had no traps burn out, even though he operated the 80, 40, 20, 15 and 10-meter antenna on the critical 40-meter band using his AL-80A linear amplifier at the 600-W PEP output level. He did not make a continuous, key-down, CW operating test at full power purposely trying to destroy the traps!

Some hams may suggest using a different type of coaxial cable for the traps. The dc resistance of 40.7 Ω per 1000 feet of RG-59 coax seems rather high. However, W8NX has found no coax other than RG-59 that has the necessary inductance-to-capacitance ratio to create the trap characteristic reactance required for the 80, 40, 20, 15 and 10-meter antenna. Conventional traps with wide-spaced, open-air inductors and appropriate fixed-value capacitors could be substituted for the coax traps, but the convenience, weatherproof configuration and ease of fabrication of coaxial-cable traps is hard to beat.

Multiband Vertical Antennas

There are two basic types of vertical antennas; either type can be used in multiband configurations. The first is the ground-mounted vertical and the second, the ground plane. These antennas are described in detail in Chapter 6, Low-Frequency Antennas.

The efficiency of any ground-mounted vertical depends a great deal on near-field earth losses. As pointed out in Chapter 3, The Effects of Ground, these near-field losses can be reduced or eliminated with an adequate radial system. Considerable experimentation has been conducted on this subject by Jerry Sevick, W2FMI, and several important results were obtained. It was determined that a radial system consisting of 40 to 50 radials, 0.2 λ long, would reduce the earth losses to about 2 Ω when a λ/4 radiator was being used. These radials should be on the earth's surface, or if buried, placed not more than an inch or so below ground. Otherwise, the RF current would have to travel through the lossy earth before reaching the radials. In a multiband vertical system, the radials should be 0.2 λ long for the lowest band, that is, 55 feet long for 3.5-MHz operation. Any wire size may be used for the radials. The radials should fan out in a circle, radiating from the base of the antenna. A metal plate, such as a piece of sheet copper, can be used at the center connection.

The other common type of vertical is the ground-plane antenna. Normally, this antenna is mounted above ground with the radials fanning out from the base of the antenna. The vertical portion of the antenna is usually an electrical λ/4, as is each of the radials. In this type of antenna, the system of radials acts somewhat like an RF choke, to prevent RF currents from flowing in the supporting structure, so the number of radials is not as important a factor as it is with a ground- mounted vertical system. From a practical standpoint, the customary number of radials is four or five. In a multiband configuration, λ/4 radials are required for each band of operation with the ground-plane antenna.

This is not so with the ground-mounted vertical antenna, where the ground plane is relied upon to provide an image of the radiating section. Note that even quarter-wave-long radials are greatly detuned by their proximity to ground—radial resonance is not necessary or even possible. In the ground-mounted case, so long as the ground-screen radials are approximately 0.2 λ long at the lowest frequency, the length will be more than adequate for the higher frequency bands.

Short Vertical Antennas

A short vertical antenna can be operated on several bands by loading it at the base, the general arrangement being similar to Figs 1 and 2. That is, for multiband work the vertical can be handled by the same methods that are used for random-length wires.

A vertical antenna should not be longer than about ¾ λ at the highest frequency to be used, however, if low-angle radiation is wanted. If the antenna is to be used on 28 MHz and lower frequencies, therefore, it should not be more than approximately 25 feet high, and the shortest possible ground lead should be used.

Another method of feeding is shown in **Fig 29**. L1 is a loading coil, tapped to resonate the antenna on the desired band. A second tap permits using the coil as a transformer for matching a coax line to the transmitter. C1 is not strictly necessary, but may be helpful on the lower frequencies, 3.5 and 7 MHz, if the antenna is quite short. In that case C1 makes it possible to tune the system to resonance with a coil of reasonable dimensions at L1. C1 may also be useful on other bands as well, if the system cannot be matched to the feed line with a coil alone.

The coil and capacitor should preferably be installed at the base of the antenna, but if this cannot be done a wire can be run from the antenna base to the nearest convenient location for mounting L1 and C1. The extra wire will of course be a part of the antenna, and since it may have to run through unfavorable surroundings it is best to avoid using it if at all possible.

This system is best adjusted with the help of an SWR indicator. Connect the coax line across a few turns of L1 and take trial positions of the shorting tap until the SWR reaches its lowest value. Then vary the line tap similarly; this should bring the SWR down to a low value. Small adjustments of both taps then should reduce the SWR to close to 1:1. If not, try adding C1 and go through the same procedure, varying C1 each time a tap position is changed.

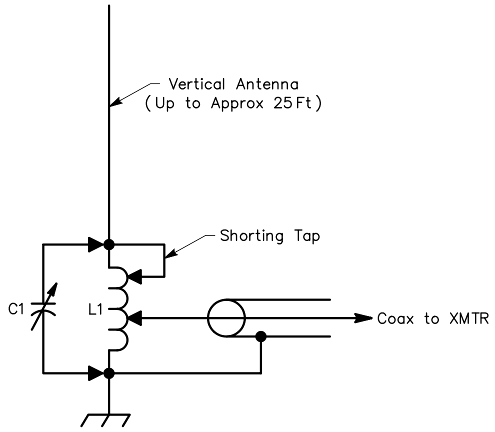

Fig 29—Multiband vertical antenna system using base loading for resonating on 3.5 to 28 MHz. L1 should be wound with bare wire so it can be tapped at every turn, using #12 wire. A convenient size is 2½ inches diameter, 6 turns per inch (such as B&W 3029). Number of turns required depends on antenna and ground lead length, more turns being required as the antenna and ground lead are made shorter. For a 25-foot antenna and a ground lead of the order of 5 feet, L1 should have about 30 turns. The use of C1 is explained in the text. The smallest capacitance that will permit matching the coax cable should be used; a maximum capacitance of 100 to 150 pF will be sufficient in any case.

Trap Verticals

The trap principle described in Fig 17 for center-fed dipoles also can be used for vertical antennas. There are two principal differences. Only one half of the dipole is used, the ground connection taking the place of the missing half, and the feed-point impedance is one half the feed-point impedance of a dipole. Thus it is in the vicinity of 30 Ω (plus the ground-connection resistance), so 52-Ω cable should be used since it is the commonly available type that comes closest to matching.

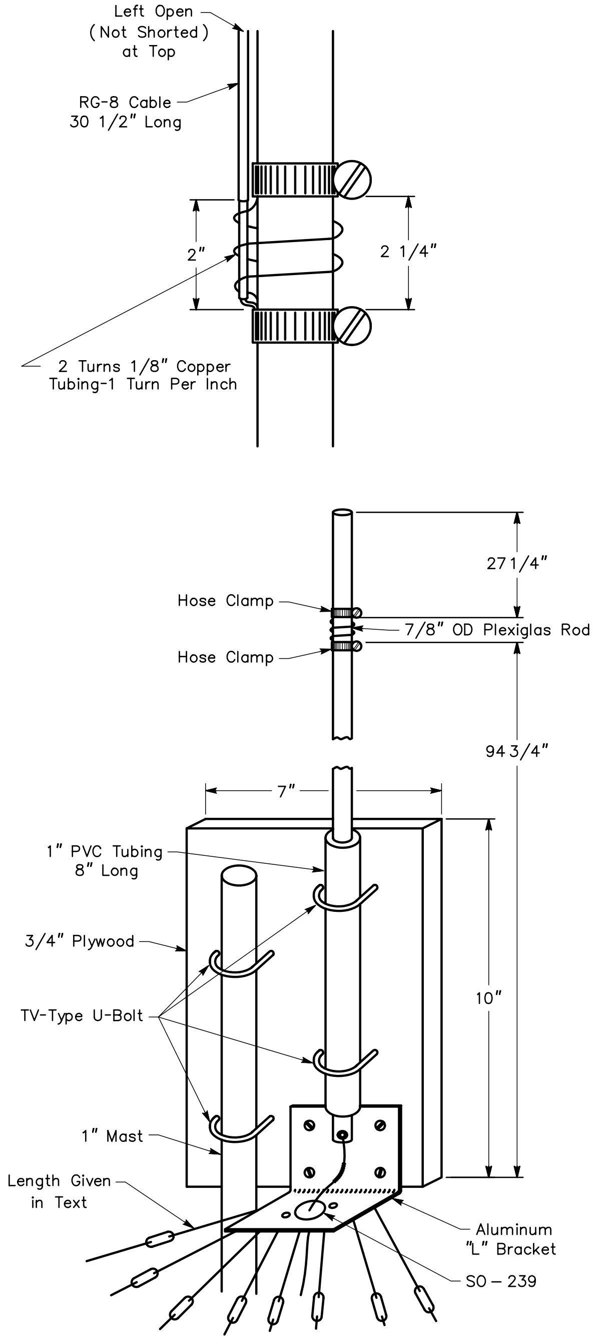
Left Open (Not Shorted) at Top
RG-8 Cable 30 1/2" Long
2"
2 1/4"
2 Turns 1/8" Copper Tubing-1 Turn Per Inch
27 1/4"
Hose Clamp
7/8" OD Plexiglas Rod
Hose Clamp
94 3/4"
7"
1" PVC Tubing 8" Long
3/4" Plywood
TV-Type U-Bolt
10"
1" Mast
Length Given in Text
Aluminum "L" Bracket
SO – 239

Fig 30—Constructional details of the 21- and 28-MHz dual-band antenna system.

A TRAP VERTICAL FOR 21 AND 28 MHZ

Simple antennas covering the upper HF bands can be quite compact and inexpensive. The two-band vertical ground plane described here is highly effective for long-distance communication when installed in the clear.

Figs 30, **31** and **32** show the important assembly details. The vertical section of the antenna is mounted on a ¾-inch thick piece of plywood board that measures 7 × 10 inches. Several coats of exterior varnish or similar material will help protect the wood from inclement weather. Both the mast and the radiator are mounted on the piece of wood by means of TV U-bolt hardware. The vertical is electrically isolated from the wood with a piece of 1-inch diameter PVC tubing. A piece approximately 8 inches long is required, and it is of the schedule-80 variety. To prepare the tubing, you must slit it along the entire length on one side. A hacksaw will work quite well. The PVC fits rather snugly on the aluminum tubing and

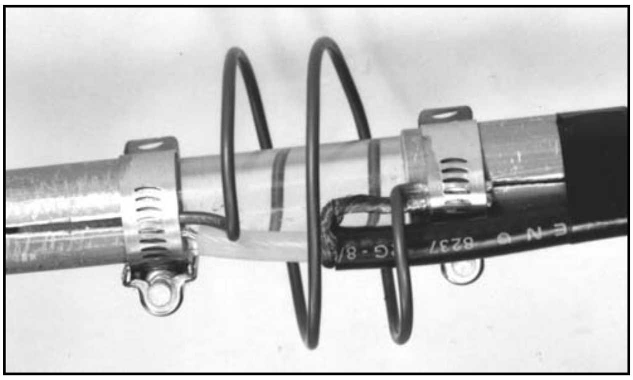

Fig 31—A close-up view of a trap. The coil is 3 inches in diameter. The leads from the coaxial-cable capacitor should be soldered directly to the pigtails of the coil. These connections should be coated with varnish after they have been secured under the hose clamps.

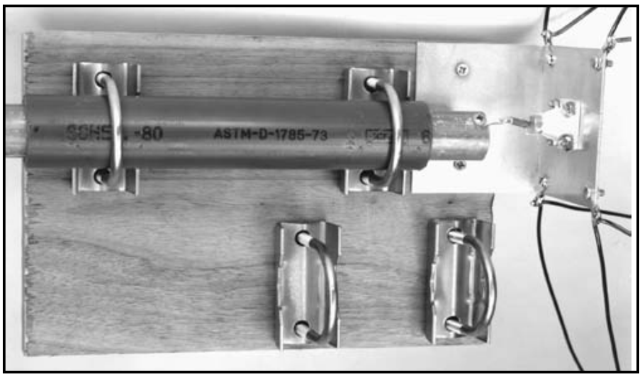

Fig 32—The base assembly of the 21- and 28-MHz vertical. The SO-239 coaxial connector and hood can be seen in the center of the aluminum L bracket. The U bolts are TV-type antenna hardware. The plywood should be coated with varnish or similar material.

will have to be "persuaded" with the aid of a hammer. Mount the mast directly on the wood with no insulation.

Use an SO-239 coaxial connector and four solder lugs on an L-shaped bracket made from a piece of aluminum sheet. Solder a short length of test probe wire, or inner conductor of RG-58 cable, to the inner terminal of the connector. A UG-106 connector hood is then slid over the wire and onto the coaxial connector. Then bolt the hood and connector to the aluminum bracket. Two wood screws are used to secure the aluminum bracket to the plywood, as shown in the drawing and photograph. Solder the free end of the wire coming from the connector to a lug mounted on the bottom of the vertical radiator. Fill any space between the wire and where it passes through the hood with GE silicone sealant or similar material to keep moisture out. The eight radials, four for each band, are soldered to the four lugs on the aluminum bracket. Separate the two sections of the vertical member with a piece of clear acrylic rod. Approximately 8 inches of ⅞-inch OD material is required. You must slit the aluminum tubing lengthwise for several inches so the acrylic rod may be inserted. The two pieces of aluminum tubing are separated by 2¼ inches.

The trap capacitor is made from RG-8 coaxial cable and is 30.5 inches long. RG-8 cable has 29.5 pF of capacitance per foot and RG-58 has 28.5 pF per foot. RG-8 cable is recommended over RG-58 because of its higher breakdown-voltage capability. The braid should be pulled back 2 inches on one end of the cable, and the center conductor soldered to one end of the coil. Solder the braid to the other end of the coil. Compression type hose clamps are placed over the capacitor/coil leads and put in position at the edges of the aluminum tubing. When tightened securely, the clamps serve a two-fold purpose—they keep the trap in contact with the vertical members and prevent the aluminum tubing from slipping off the acrylic rod. The coaxial-cable capacitor runs upward along the top section of the antenna. This is the side of the antenna to which the braid of the capacitor is connected. Place a cork or plastic cap in the very top of the antenna to keep moisture out.

Installation and Operation

The antenna may be mounted in position using a TV-type tripod, chimney, wall or vent mount. Alternatively, a telescoping mast or ordinary steel TV mast may be used, in which case the radials may be used as guys for the structure. The 28-MHz radials are 8 feet 5 inches long, and the 21-MHz radials are 11 feet 7 inches.

Any length of 50-Ω cable may be used to feed the antenna. The SWR at resonance should be on the order of 1.2:1 to 1.5:1 on both bands. The reason the SWR is not 1:1 is that the feed-point resistance is something other than 50 Ω —closer to 35 or 40 Ω.

The Open-Sleeve Antenna

Although only recently adapted for the HF and VHF amateur bands, the open-sleeve antenna has been around since 1946. The antenna was invented by Dr J. T. Bolljahn, of Stanford Research Institute. This section on sleeve antennas was written by Roger A. Cox, WB0DGF.

The basic form of the open-sleeve monopole is shown in **Fig 33**. The open-sleeve monopole consists of a base-fed central monopole with two parallel closely spaced parasitics, one on each side of the central element, and grounded at each base. The lengths of the parasitics are roughly one half that of the central monopole.

Impedance

The operation of the open sleeve can be divided into two modes, an antenna-mode and a transmission-line mode. This is shown in **Fig 34**.

The antenna-mode impedance, Z_A, is determined by the length and diameter of the central monopole. For sleeve lengths less than that of the monopole, this impedance is essentially independent of the sleeve dimensions.

The transmission-line mode impedance, Z_T, is determined by the characteristic impedance, end impedance, and length of the 3- wire transmission line formed by the central monopole and the two sleeve elements. The characteristic impedance, Z_c, can be determined by the element diameters and spacing if all element diameters are equal, and is found from

$$Z_c = 207 \log 1.59 (D/d) \qquad \text{(Eq 2)}$$

where

D = spacing between the center of each sleeve element and the center of the driven element

d = diameter of each element

This is shown graphically in **Fig 35**. However, since the end impedance is usually unknown, there is little need to know the characteristic impedance. The transmission-line mode impedance, Z_T, is usually determined by an educated guess and experimentation.

As an example, let us consider the case where the central monopole is λ/4 at 14 MHz. It would have an antenna mode impedance, Z_A, of approximately 52 Ω, depending upon the ground conductivity and number of radials. If two sleeve elements were added on either side of the central monopole, with each approximately half the height of the monopole and at a distance equal to their height, there would be very little

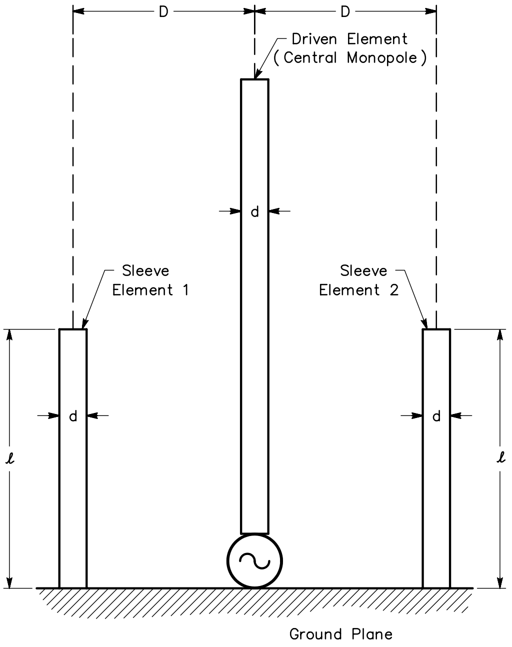

Fig 33—Diagram of an open-sleeve monopole.

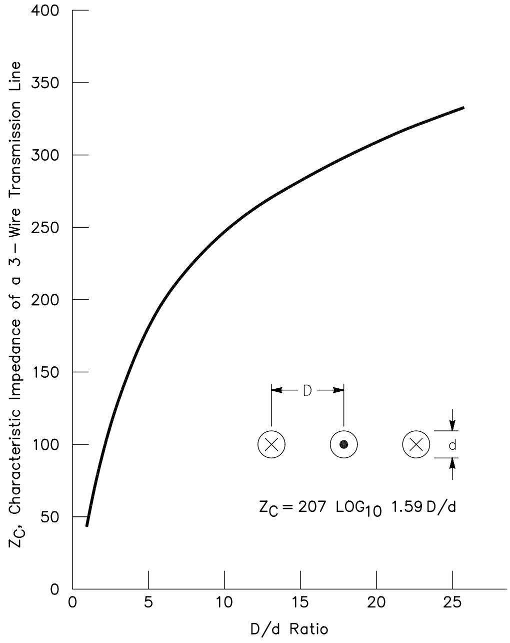

Fig 35—Characteristic impedance of transmission-line mode in an open-sleeve antenna.

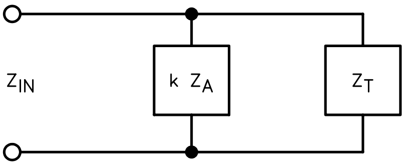

Fig 34—Equivalent circuit of an open-sleeve antenna.

effect on the antenna mode impedance, Z_A, at 14 MHz.

Also, Z_T at 14 MHz would be the end impedance transformed through a λ/8 section of a very high characteristic impedance transmission line. Therefore, Z_T would be on the order of 500-2000 Ω resistive plus a large capacitive reactance component. This high impedance in parallel with 52 Ω would still give a resultant impedance close to 52 Ω.

At a frequency of 28 MHz, however, Z_A is that of an end-fed half-wave antenna, and is on the order of 1000-5000 Ω resistive. Also, Z_T at 28 MHz would be on the order of 1000 to 5000 Ω resistive, since it is the end impedance of the sleeve elements transformed through a quarter-wave section of a very high characteristic impedance 3-wire transmission line. Therefore, the parallel combination of Z_A and Z_T would still be on the order of 500 to 2500 Ω resistive.

If the sleeve elements were brought closer to the central monopole such that the ratio of the spacing to element diameter was less than 10:1, then the characteristic impedance of the 3-wire transmission line would drop to less than 250 Ω. At 28 MHz, Z_A remains essentially unchanged, while Z_T begins to edge closer to 52 Ω as the spacing is reduced. At some particular spacing the characteristic impedance, as determined by the D/d ratio, is just right to transform the end impedance to exactly 52 Ω at some frequency. Also, as the spacing is decreased, the frequency where the impedance is purely resistive gradually increases.

The actual impedance plots of a 14/28-MHz open-sleeve monopole appear in **Figs 36** and **37**. The length of the central monopole is 195.5 inches, and of the sleeve elements 89.5 inches. The element diameters range from 1.25 inches at the bases to 0.875 inch at each tip. The measured impedance of the 14-MHz monopole alone, curve A of Fig 36, is quite high. This is probably because of a very poor ground plane under the antenna. The addition of the sleeve elements raises this impedance slightly, curves B, C and D.

As curves A and B in Fig 37 show, an 8-inch sleeve spacing gives a resonance near 27.8 MHz at 70 Ω, while a 6-inch spacing gives a resonance near 28.5 MHz at 42 Ω. Closer spacings give lower impedances and higher resonances. The optimum spacing for this particular antenna would be somewhere between 6 and 8 inches. Once the spacing is found, the lengths of the sleeve elements can be tweaked slightly for a choice of resonant frequency.

In other frequency combinations such as 10/21, 10/24, 14/21 and 14/24-MHz, spacings in the 6 to 10-inch range work very well with element diameters in the 0.5 to 1.25-inch range.

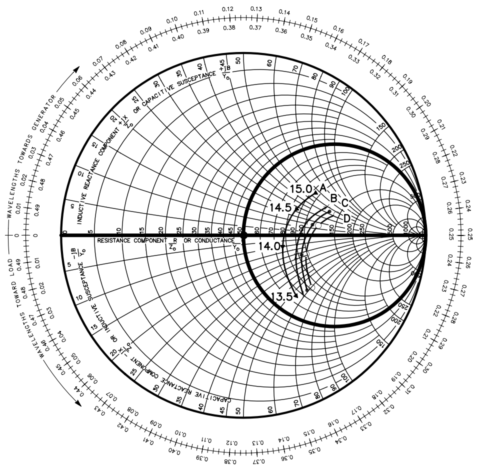

Fig 36—Impedance of an open-sleeve monopole for the frequency range 13.5-15 MHz. Curve A is for a 14 MHz monopole alone. For curves B, C and D, the respective spacings from the central monopole to the sleeve elements are 8, 6 and 4 inches. See text for other dimensions.

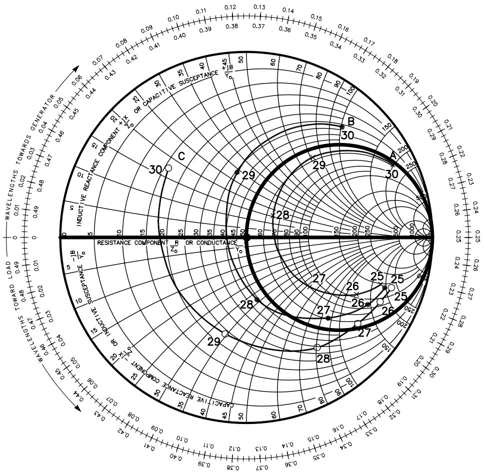

Fig 37—Impedance of the open-sleeve monopole for the range 25-30 MHz. For curves A, B and C the spacings from the central monopole to the sleeve elements are 8, 6 and 4 inches, respectively.

Bandwidth

The open-sleeve antenna, when used as a multiband antenna, does not exhibit broad SWR bandwidths unless, of course, the two bands are very close together. For example, **Fig 38** shows the return loss and SWR of a single 10-MHz vertical antenna. Its 2:1 SWR bandwidth is 1.5 MHz, from 9.8 to 11.3 MHz. Return loss and SWR are related as given by the following equation.

$$SWR = \frac{1+k}{1-k} \qquad \text{(Eq 3)}$$

where

$$k = 10^{\frac{R_L}{20}}$$

RL = return loss, dB

When sleeve elements are added for a resonance near 22 MHz, the 2:1 SWR bandwidth at 10 MHz is still nearly 1.5 MHz, as shown in **Fig 39**. The total amount of spectrum under 2:1 SWR increases, of course, because of the additional band, but the individual bandwidths of each resonance are virtually unaffected.

The open-sleeve antenna, however, can be used as a broadband structure, if the resonances are close enough to overlap. With the proper choices of resonant frequencies, sleeve and driven element diameters and sleeve spacing, the SWR "hump" between resonances can be reduced to a value less than 3:1. This is shown in **Fig 40**.

Current Distribution

According to H. B. Barkley (see Bibliography at the end of this chapter), the total current flowing into the base of the open-sleeve antenna may be broken down into two components, that contributed by the antenna mode, I_A, and that contributed by the transmission-line mode, I_T. Assuming that the sleeves are approximately half the height of the central monopole, the impedance of the antenna mode, Z_A, is very low at the resonant frequency of the central monopole, and

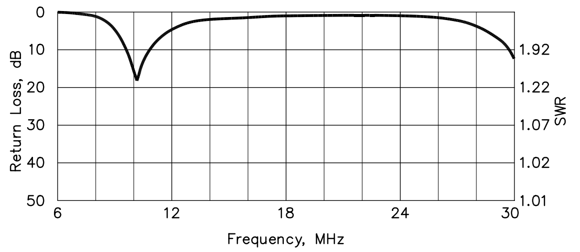

Fig 38—Return loss and SWR of a 10 MHz vertical antenna. A return loss of 0 dB represents an SWR of infinity. The text contains an equation for converting return loss to an SWR value.

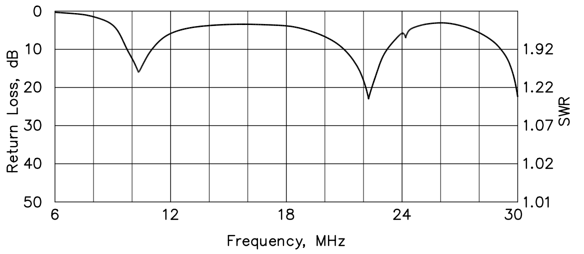

Fig 39—Return loss and SWR of a 10/22 MHz open-sleeve vertical antenna.

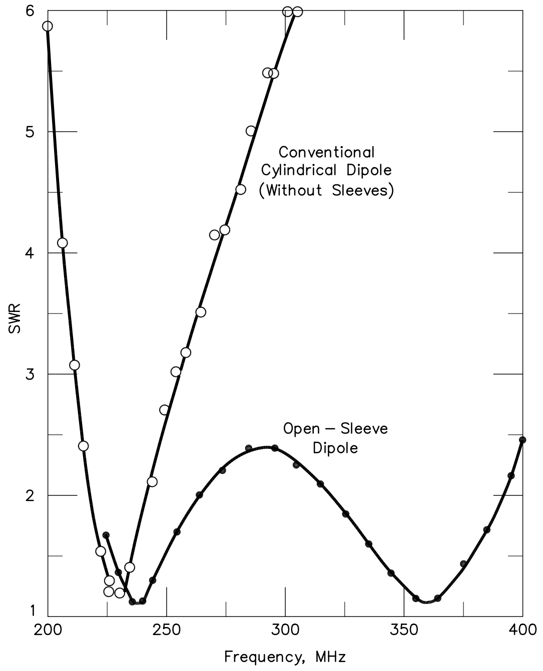

Fig 40—SWR response of an open-sleeve dipole and a conventional dipole.

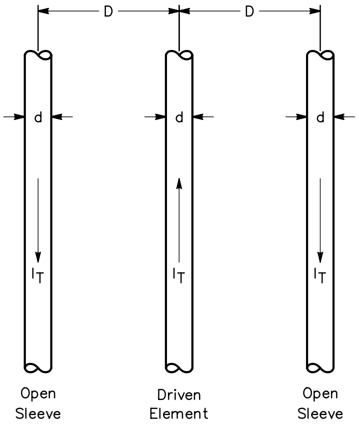

Fig 41—Current distribution in the transmission-line mode. The amplitude of the current induced in each sleeve element equals that of the current in the central element but the phases are opposite, as shown.

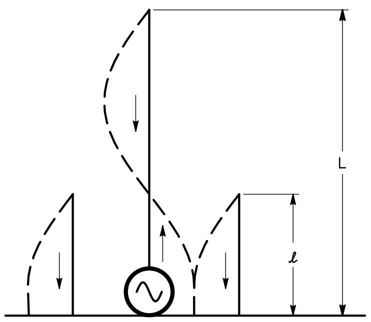

Fig 42—Total current distribution with λ= L/2.

the impedance of the transmission-line mode, Z_T, is very high. This allows almost all of the current to flow in the antenna mode, and I_A is very much greater than I_T. Therefore, the current on the central λ/4 monopole assumes the standard sinusoidal variation, and the radiation and gain characteristics are much like those of a normal λ/4 vertical antenna.

However, at the resonant frequency of the sleeves, the impedance of the central monopole is that of an end fed half-wave monopole and is very high. Therefore I_A is small. If proper element diameters and spacings have been used to match the transmission line mode impedance, Z_T, to 52 Ω, then I_T, the transmission line mode current, is high compared to I_A.

This means that very little current flows in the central monopole above the tops of the sleeve elements, and the radiation is mostly from the transmission-line mode current, I_T, in all three elements below the tops of the sleeve elements. The resulting current distribution is shown in **Figs 41** and **42** for this case.

Radiation Pattern and Gain

The current distribution of the open-sleeve antenna where all three elements are nearly equal in length is nearly that of a single monopole antenna. If, at a particular frequency, the elements are approximately λ/4 long, the current distribution is sinusoidal.

If, for this and other length ratios, the chosen diameters and spacings are such that the two sleeve elements approach an interelement spacing of λ/8, the azimuthal pattern will show directivity typical of two in-phase vertical radiators, approximately λ/8 apart. If a bi-directional pattern is needed, then this is one way to achieve it.

Spacings closer than this will produce nearly circular azimuthal radiation patterns. Practical designs in the 10 to 30 MHz range using 0.5 to 1.5-inch diameter elements will produce azimuthal patterns that vary less than ±1 dB.

If the ratio of the length of the central monopole to the length of the sleeves approaches 2:1, then the elevation pattern of the open-sleeve vertical antenna at the resonant frequency of the sleeves becomes slightly compressed. This is because of the in-phase contribution of radiation from the λ/2 central monopole.

As shown in **Fig 43**, the 10/21-Mhz open-sleeve vertical antenna produces a lower angle of radiation at 21.2 MHz with a corresponding increase in gain of 0.66 dB over that of the 10-MHz vertical alone. At length ratios approaching 3:1, the

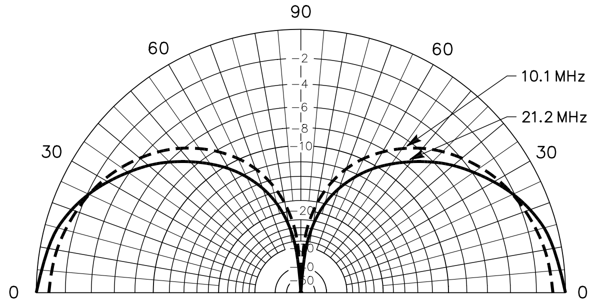

Fig 43—Vertical-plane radiation patterns of a 10/21-MHz open-sleeve vertical antenna on a perfect ground plane. At 10.1 MHz the maximum gain is 5.09 dBi, and 5.75 dBi at 21.2 MHz.

antenna mode and transmission-line mode impedance become nearly equal again, and the central monopole again carries a significant portion of the antenna current. The radiation from the top λ/2 combines constructively with the radiation from the λ/4 sleeve elements to produce gains of up to 3 dB more than just a quarter-wave vertical element alone.

Length ratios in excess of 3.2:1 produce higher level sidelobes and less gain on the horizon, except for narrow spots near the even ratios of 4:1, 6:1, 8:1, etc. These are where the central monopole is an even multiple of a half-wave, and the antenna-mode impedance is too high to allow much antenna-mode current.

Up to this point, it has been assumed that only λ/4 resonance could be used on the sleeve elements. The third, fifth, and seventh-order resonances of the sleeve elements and the central monopole element can be used, but their radiation patterns normally consist of high-elevation lobes, and the gain on the horizon is less than that of a λ/4 vertical.

Practical Construction and Evaluation

The open-sleeve antenna lends itself very easily to home construction. For the open-sleeve vertical antenna, only a feed-point insulator and a good supply of aluminum tubing are needed. No special traps or matching networks are required. The open-sleeve vertical can produce up to 3 dB more gain than a conventional λ/4 vertical. Further, there is no reduction in bandwidth, because there are no loading coils.

The open-sleeve design can also be adapted to horizontal dipole and beam antennas for HF, VHF and UHF. A good example of this is Telex/Hy-Gain's Explorer 14 triband beam which utilizes an open sleeve for the 10/15-meter driven element. The open-sleeve antenna is also very easy to model in computer programs such as *NEC* and *MININEC*, because of the open tubular construction and lack of traps or other intricate structures.

In conclusion, the open-sleeve antenna is an antenna experimenters delight. It is not difficult to match or construct, and it makes an ideal broadband or multiband antenna.

The Coupled-Resonator Dipole

A variation of the open-sleeve system above is the coupled-resonator system described by Gary Breed, K9AY, in an article in *The ARRL Antenna Compendium, Vol 5*, entitled “The Coupled-Resonator Principle: A Flexible Method for Multiband Antennas.” The following is condensed from that article.

In 1995, *QST* published two antenna designs that use an interesting technique to get multiband coverage in one antenna. Rudy Severns, N6LF, described a wideband 80 and 75-meter dipole using this technique, and Robert Wilson, AL7KK, showed us how to make a three-band vertical. Both of these antennas achieve multi-frequency operation by placing resonant conductors very close to a driven dipole or vertical—with no physical connection.

The Coupled-Resonator Principle

As we all know, nearby conductors can interact with an antenna. Our dipoles, verticals and beams can be affected by nearby power lines, rain gutters, guy wires and other metallic materials. The antennas designed by Severns and Wilson use this interaction intentionally, to combine the resonances of several conductors at a single feed point. While other names have been used, I call the behavior that makes these antennas work the *coupled-resonator* (C-R) principle.

Take a look at **Fig 44**, which illustrates the general idea. Each figure shows the SWR at the feed point of a dipole, over a range of frequencies. When this dipole is all alone, it will have a very low SWR at its half-wave resonant frequency (Fig 44A). Next, if we take another wire or tubing conductor and start bringing it close to the dipole, we will see a “bump” in the dipole’s SWR at the resonant frequency of this new wire. See Fig 44B. We are beginning to the see the effects of interaction between the two conductors. As we bring this new conductor closer, we reach a point where the SWR “bump” has grown to a very deep dip—a low SWR. We now have a good match at both the original dipole’s resonant frequency and the frequency of the new conductor, as illustrated in Fig 44C.

We can repeat this process for several more conductors at other frequencies to get a dipole with three, four, five, six, or more resonant frequencies. The principle also applies to verticals, so any reference to a dipole can be considered to be valid for a vertical, as well.

We can write a definition of the C-R principle this

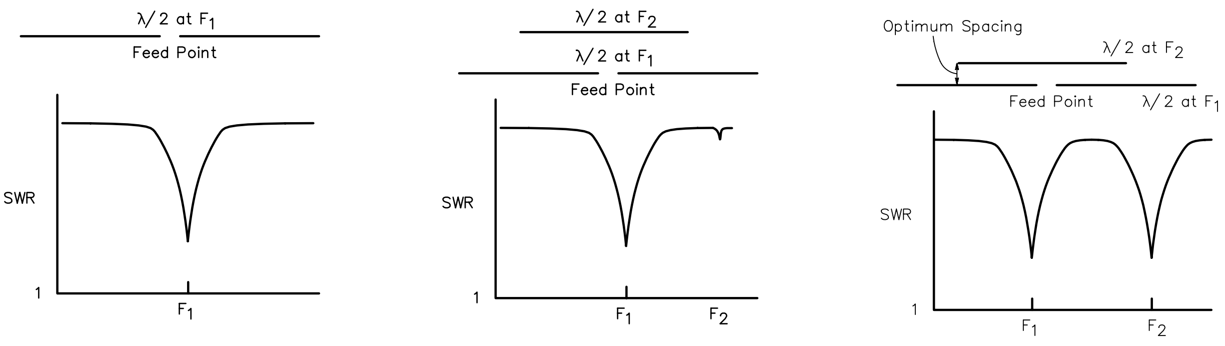

Fig 44—At A, the SWR of a dipole over a wide frequency range. At B, a nearby conductor is just close enough to interact with the dipole. At C, when the second conductor is at the optimum spacing, the combination is matched at both frequencies.

way: *Given a dipole (or vertical) at one frequency and an additional conductor resonant at another frequency, there is an optimum distance between them that results in the resonance of the additional conductor being imposed upon the original dipole, resulting in a low SWR at both resonant frequencies.*

Some History

In the late 1940s, the *coaxial sleeve* antenna was developed (**Fig 45A**), covering two frequencies by surrounding a dipole or monopole with a cylindrical tube resonant at the higher of the desired frequencies. In the 1950s, Gonset briefly marketed a two-band antenna based on this design. Other experimenters soon determined that two conductors at the second frequency, placed on either side of the main dipole or monopole, would make a skeleton representation of a cylinder (Fig 45B). This is called the *open-sleeve* antenna. The Hy-Gain Explorer tribander uses this method in its driven element to obtain resonance in the 10-meter band. Later on, a few antenna developers finally figured out that these extra conductors did not need to be added in pairs, and that a single conductor at each frequency could add the extra resonances (Fig 45C). This is the method used by Force 12 in some of their multiband antennas.

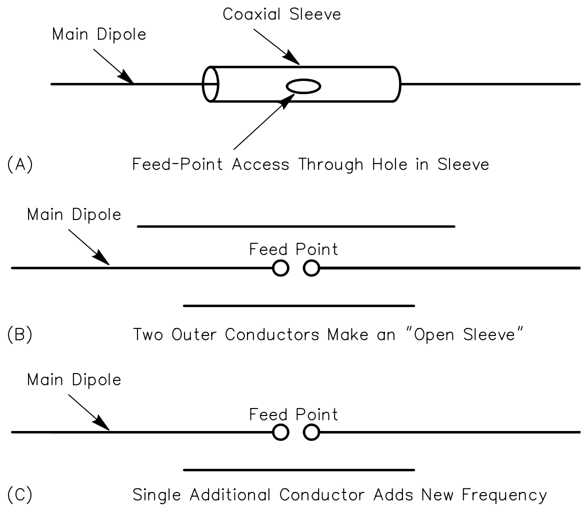
Coaxial Sleeve
Main Dipole
(A)
Feed-Point Access Through Hole in Sleeve
Main Dipole
Feed Point
(B)
Two Outer Conductors Make an "Open Sleeve"
Main Dipole
Feed Point
(C)
Single Additional Conductor Adds New Frequency

Fig 45—Evolution of coupled-resonator antennas: At A, the *coaxial-sleeve* dipole; at B, the *open-sleeve* dipole; and at C, a *coupled-resonator* dipole, the most universal configuration.

This is a perfect example of how science works. A specific idea is discovered, with later developments leading to an underlying general principle. The original coaxial-sleeve configuration is the most specific, being limited to two frequencies and requiring a particular construction method. The open-sleeve antenna is an intermediate step, showing that the sleeve idea is not limited to one configuration.

Finally, we have the coupled-resonator concept, which is the general principle, applicable in many different antenna configurations, for many different frequency combinations. Severns' antenna uses it with a folded dipole, and Wilson uses it with a main vertical that is off-center fed. The author K9AY used it with conventional dipoles and quarter-wave verticals. Other designers have used the principle more subtly, like putting the first director in a Yagi very close to the driven element, broadening the SWR bandwidth the same way Severns' design does with a dipole.

In the past, most antennas built with this single-conductor technique have also been called *open-sleeve* (or *multiple-open-sleeve*) antennas, a term taken from the history of their development. However, the term *sleeve* implies that one conductor must surround another. This is not really a physical or electrical description of the antenna's operation, therefore, K9AY suggests using the term *coupled-resonator,* which is the most accurate description of the general principle.

A Little Math

The interaction that makes the C-R principle work is not random. It behaves in a predictable, regular manner. K9AY derived an equation that shows the relationship between the driven element and the additional resonators for ordinary dipoles and verticals:

$$\frac{\log_{10} d}{\log_{10}(D/4)} = 0.54 \qquad \text{(Eq 4)}$$

where

d = distance between conductors, measured in wavelengths at the frequency of the chosen additional resonator

D = the diameter of the conductors, also in wavelengths at the frequency of the additional resonator.

Eq 4 assumes they are both the same diameter and that the feed point impedance at both frequencies is the same as a dipole in free space (72 Ω) or a quarter-wave monopole over perfect ground (36 Ω).

The equation only describes the impedance due to the additional resonator. The main dipole element is always part of the antenna, and it may have a fairly low impedance at the additional frequency. This is the case when the frequencies are close together, or when the main element is operating at its third harmonic. At these frequencies, the spacing distance must be adjusted so that the parallel combination of dipole and resonator results in the desired feed-point impedance.

K9AY worked out two correction factors, one to cover a range of impedances and another for frequencies close together. These can be included in the basic equation, which is rearranged below to solve for the distance between the conductors:

$$d = 10^{0.54\log_{10}(D/4)} \times \frac{Z_0 + 35.5}{109} \times [1 + e^{-[((F_2/F_1)-1.1)\times 11.3)+0.1)]}] \qquad \text{(Eq 5)}$$

where

d and D are the same as above.

Z_0 = the desired feed point impedance at the frequency of the additional resonator (between 20 and 120 Ω). For a vertical, multiply the desired impedance by two to get Z_0. If you want a 50-Ω feed, use 100 Ω for Z_0.

F_1 = the resonant frequency of the main dipole or vertical.

F_2 = the resonant frequency of the additional conductor. The ratio F_2/F_1 is more than 1.1.

e = 2.7183, the base of natural logarithms.

Eq 5 does not directly allow for conductors of unequal diameters, but it can be used as a starting point if you use the diameter of the driven dipole or vertical element for D in the equation.

Characteristics of C-R Antennas

Here's the important stuff—what's different about C-R antennas, what are they good for and what are their drawbacks? The key points are:

- Multiband operation without traps, stubs or tuners

K9AY's Eq 5 above does indeed yield a good "first-cut" value for the spacing between coupled-resonator elements. **Fig A** shows the spacing, in inches, plotted against the ratio of frequencies, for two coupled resonator elements with different diameters, again expressed in inches. This is for an upper frequency of 28.4 MHz. Beyond a frequency ratio of about 1.5:1 (28.4:18.1 MHz), the spacing flattens out to a fixed distance between elements for each element diameter. For example, if $^1/_2$-inch elements are used at 28.4 and 18.1 MHz, the spacing between the elements is about 3.75 inches.

EZNEC verifies Eq 5's computations. Note that a large number of segments are necessary for each element when they are closely spaced from each other, and the segments on the elements must be closely aligned with each other. Be sure to run the Average Gain test, as well as Segmentation tests. The modeler should also be aware that if mutually coupled resonators are placed along a horizontal boom (as they would be on multiband Yagis using coupled resonators), the higher-frequency elements will act like retrograde directors, producing some gain (or lack of gain, depending on the azimuth being investigated).

For example, in the *EZNEC* file **K9AY C-R 28-21-14 MHz 1 In.EZ**, using 1-inch diameter elements spaced 6 inches apart, if the 28-MHz element is placed 6 inches behind the 14-MHz driven element (with the 21-MHz element placed 6 inches ahead), on 28 MHz the system will have a F/B of 2.6 dB, favoring the rearward direction. On 21 MHz, the system will exhibit a F/B of 1.6 dB, favoring the forward direction. Of course, there are systems where gain and F/B due to the C-R configuration may be put to good use, such as the multiband Yagis mentioned above. However, if the elements are spaced above/below the 14-MHz driven element there is no distortion of the dipole patterns.

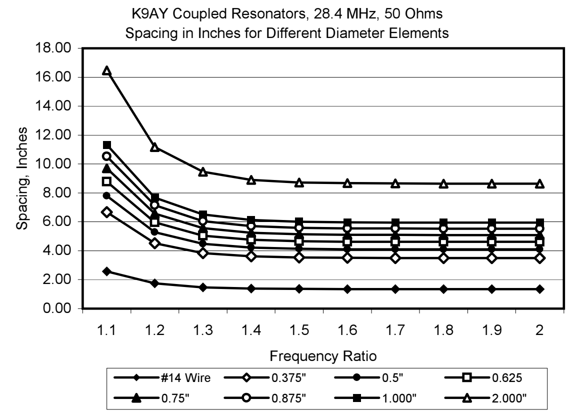

Fig A—Graph of the spacing versus frequency ratio for two Coupled-Resonator elements at 28.4 MHz, for 50-Ω feed-point impedance.

- Flexible impedance matching at each frequency
- Independent fine-tuning at each frequency (little interaction)
- Easily modeled using *MININEC* or *NEC*-based programs
- Pruning process same as a simple dipole
- Can accommodate many frequencies (seven or more)
- Virtually lossless coupling (high efficiency)
- Requires a separate wire or tubing conductor at each frequency
- Mechanical assembly requires a number of insulated supports
- Narrower bandwidth than equivalent dipole
- Capacitance requires slight lengthening of conductors

To begin with, the most obvious characteristic is that this principle can be used to add multiple resonant frequencies to an ordinary dipole or vertical, using additional conductors that are not physically connected. This gives us three variable factors: (1) the diameter of the conductor, (2) its length, and (3) its position relative to the main element.

Having the freedom to control these factors gives us the advantage of *flexibility*; we have a wide range of control over the impedance at each added frequency. Another advantage is that the behavior at each frequency is quite *independent*, once the basic design is in place. In other words, making fine-tuning adjustments at one frequency doesn't change the resonance or impedance at the other frequencies. A final advantage is *efficiency.* With conductors close together, and with a resonant target conductor, coupling is very efficient. Traps, stubs, and compensating networks found on other multiband antennas all introduce lossy reactive components.

There are two main disadvantages of C-R antennas. The first is the relative *complexity* of construction. Several conductors are needed, installed with some type of insulating spacers. Other multiband antennas have their complexities as well (such as traps that need to be mounted and tuned), but C-R antennas will usually be bulkier. The larger size generally means greater windload, which is a disadvantage to some hams.

The other significant disadvantage is *narrower bandwidth,* particularly at the highest of the operating frequencies. We can partially overcome this problem with large conductors that are naturally broad in bandwidth, and in some cases we might even use an extra conductor to put two resonances in one band. It is interesting to note that the pattern is opposite that of trapped antennas. The C-R antenna gets narrower at the highest frequencies of operation, while trap antennas generally have narrowest bandwidth at their lowest frequencies.

There are two special situations that should be noted. First, when the antenna has a resonance near the frequency where the driven dipole is 3/2 λ long (3/4 λ for a vertical), the dipole has a fairly low impedance. The spacing of the C-R element needs to be increased to raise its impedance so that the parallel combination of the main element and C-R element equals the desired impedance (usually 50 Ω). There is also significant antenna current in the part of the main dipole extending beyond the C-R section, contributing to the total radiation pattern. As a result, this particular arrangement radiates as three λ/2 sections in phase, and has about 3 dB gain and a narrower directional pattern compared to a dipole (**Fig 46**). This might be an advantage for antennas covering bands with a frequency ratio of about three, such as 3.5 and 10.1 MHz, 7 and 21 MHz, or 144 and 430 MHz.

The other special situation is when we want to add a new frequency very close to the resonant frequency of the main dipole. An antenna for 80 and 75 meters would be an example of this. Again, the driven dipole has a fairly low impedance at the new frequency. Add the fact that coupling is very strong between these similar conductors and we find that a wide spacing is required to make the antenna work. A dipole resonant at 3.5 MHz and another wire resonant at 3.8 MHz will need to be 3 or 4 feet apart, while a 3.5 MHz and 7 MHz combination might only need to be spaced 4 or 5 inches.

Another useful characteristic of C-R antennas is that they are easily and accurately modeled by computer programs based on either *MININEC* or *NEC*, as long as you stay within each program's limitations. For example, Severns points out that *MININEC* does not handle folded dipoles very well, and *NEC* modeling is required. With ease of computer modeling, a precise answer isn't needed for the design equation given above. An approximate solution will provide a starting point that can quickly be adjusted for optimum dimensions.

The added resonators have an effect on the lengths of all conductors, due to the capacitance between the conductors. Capacitance causes antennas to look electrically shorter, so

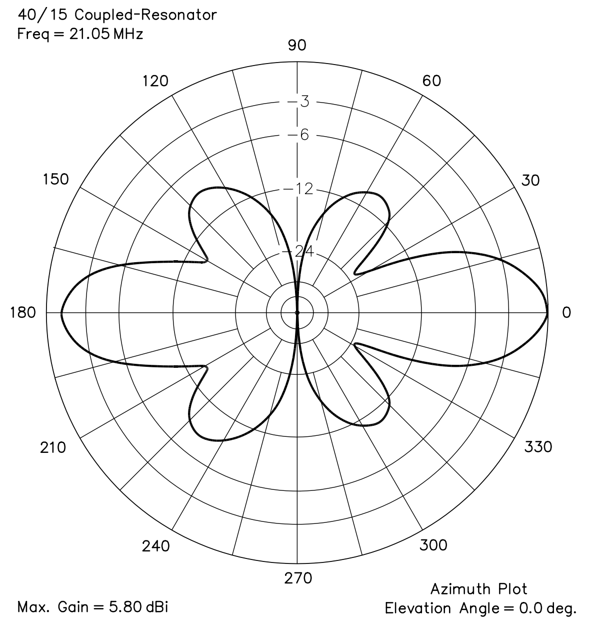

Fig 46—Radiation pattern for the special case of a C-R antenna with the additional frequency at the third harmonic of the main dipole resonant frequency.

each element needs to be about 1% or 2% longer than a simple dipole at the same frequency. As a rule of thumb, use 477/f (in feet) instead of the usual 468/f when calculating dipole length, and 239/f instead of 234/f for a λ/4 vertical.

A 30/17/12-Meter Dipole

To show how a C-R antenna is designed, let's build a dipole to cover all three WARC bands. We'll use #12 wire, which has a diameter of 0.08 inches, and the main dipole will be cut for the 10.1 MHz band. From the equation above, the spacing between the main dipole and the 18-MHz resonator should be 2.4 inches for 72 Ω, or 1.875 inches for 50 Ω. At 24.9 MHz, the spacing to the resonator for that band should be 2.0 inches for 72 Ω, or 1.62 inches for 50 Ω. Of course, this antenna will be installed over real ground, not in free space, so these spacing distances may not be exact. Plugging these numbers into your favorite antenna-modeling program will let you optimize the dimensions for installation at the height you choose.

For those of you who like to work with real antennas, not computer-generated ones, the predicted spacing is accurate enough to build an antenna with minimum trial-and-error. You should use a nice round number just larger than the calculated spacing for 50 Ω. For this antenna, K9AY decided that the right spacing for the desired height would be 2 inches for the 18 MHz resonator and 1.8 inches for the 24.9 MHz resonator. For simplicity of construction, he just used 2 inches for both, figuring that the worst he would get is a 1.2:1 SWR if the numbers were a little bit off. Like all dipoles, the impedance varies with height above ground, but the 2-inch spacing results in an excellent match on the two additional bands, at heights of more than 25 feet.

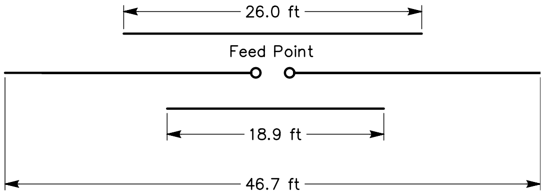

Fig 47—Dimensions of a C-R dipole for the 30, 17 and 12-meter bands.

The final dimensions of the dipole for 10.1, 18.068 and 24.89 MHz are shown in **Fig 47**. These are the final pruned lengths for a straight dipole installed at a height of about 40 feet. If you put up the antenna as an inverted V, you will need each wire to be a bit longer. Pruning this type of antenna is just like a dipole—if it's resonant too low in frequency, it's too long and the appropriate wire needs to be shortened. So, you can cut the wires just a little long to start with and easily prune them to resonance.

A final note: if you want to duplicate this antenna design, remember that the 2-inch spacing is just for #12 wire! The required spacing for a C-R antenna is related to the conductor diameter. This same antenna built with #14 wire needs under 1½-inch spacing, while a 1-inch aluminum-tubing version requires about 7-inch spacing.

Summary

The coupled-resonator principle is one more weapon in the antenna designer's arsenal. It's not the perfect method for all multiband antennas, but what the C-R principle offers is an alternative to traps and tuners, in exchange for using more wire or aluminum. Although a C-R antenna requires more complicated construction, its main attraction is in making a multiband antenna that can be built with no compromise in matching or efficiency.

HF Discone Antennas

The material in this section is adapted from an article by Daniel A. Krupp, W8NWF, in *The ARRL Antenna Compendium, Vol 5*. The name *discone* is a contraction of the words disc and cone. Although people often describe a discone by its design-center frequency (for example, a "20-meter discone"), it works very well over a wide frequency range, as much as several octaves. **Fig 48** shows a typical discone, constructed of sheet metal for UHF use. On lower frequencies, the sheet metal may be replaced with closely spaced wires and/or aluminum tubing.

The dimensions of a discone are determined by the lowest frequency of use. The antenna produces a vertically polarized signal at a low-elevation angle and it presents a good match for 50-Ω coax over its operating range. One advantage of the discone is that its maximum current area is near the top of the antenna, where it can radiate away from ground clutter. The cone-like skirt of the discone radiates the signal—radiation from the disc on top is minimal. This is because the currents flowing in the skirt wires essentially all go in the same direction, while the currents in the disc elements oppose each other and cancel out. The discone's omnidirectional characteristics make it ideal for roundtable QSOs or for a Net Control station.

Electrical operation of this antenna is very stable, with no changes due to rain or accumulated ice. It is a self-contained antenna—unlike a traditional ground-mounted vertical radiator, the discone does not rely on a ground-radial system for efficient operation. However, just like any other vertical

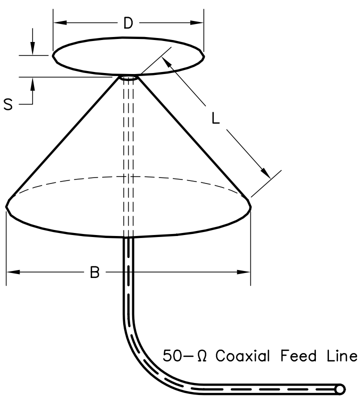

Fig 48—Diagram of VHF/UHF discone, using a sheet-metal disc and cone. It is fed directly with 50-Ω coax line. The dimensions L and D, together with the spacing S between the disc and cone, determine the frequency characteristics of the antenna. $L = 246 / f_{MHz}$ for the lowest frequency to be used. Diameter D should be from 0.67 to 0.70 of dimension L. The diameter at the bottom of the cone B is equal to L. The space S between disc and cone can be 2 to 12 inches, with the wider spacing appropriate for larger antennas.

antenna, the quality of the ground in the Fresnel area will affect the discone's far-field pattern.

Both the disc and cone are inherently balanced for wind loading, so torque caused by the wind is minimal. The entire cone and metal mast or tower can be connected directly to ground for lightning protection.

Unlike a trap vertical or a triband beam, discone antennas are not adjusted to resonate at a particular frequency in a ham band or a group of ham bands. Instead, a discone functions as a sort of high-pass filter, efficiently radiating RF all the way from the low-frequency design cutoff to the high-frequency limits imposed by the physical design.

While VHF discones have been available out-of-the-box for many years, HF discones are rare indeed. Some articles have dealt with HF discones, where the number of disc elements and cone wires was minimized to cut costs or to simplify construction. While the minimalist approach is fine if the sought-after results really are obtained, W8NWF believes in building his discones without compromise.

History of the Discone

The July 1949 and July 1950 issues of *CQ* magazine both contained excellent articles on discones. The first article, by Joseph M. Boyer, W6UYH, said that the discone was developed and used by the military during World War II. (See Bibliography.) The exact configuration of the top disc and cone was the brainchild of Armig G. Kandonian. Boyer described three VHF models, plus information on how to build them, radiation patterns, and most importantly, a detailed description of how they work. He referred to the discone as a type of "coaxial taper transformer."

The July 1950 article was by Mack Seybold, W2RYI. He described an 11-MHz version he built on his garage roof. The mast actually fit through the roof to allow lowering the antenna for service. Seybold stated that his 11-MHz discone would load up on 2 meters but that performance was down 10 dB compared to his 100-MHz Birdcage discone. He commented that this was caused by the relatively large spacing between the disc and cone. Actually, the performance degradation he found was caused by the wave angle lifting upward at high frequencies. The cone wires were electrically long, causing them to act like long wire antennas. See **Fig 49**.

W8NWF's First Discone: the A-Frame Discone

The first discone was one designed to cover 20 through 10 meters without requiring an antenna tuner. The cone assembly uses 18-foot long wires, with a 60° included apex angle and a 12-foot diameter disc assembly. See **Fig 50** and **Fig 51**. The whole thing was assembled on the ground, with the feed coax and all guys attached. Then with the aid of some friends, it was pulled up into position.

The author used a 40-foot tall wooden "A-frame" mast, made of three 22-foot-long 2 × 4s. He primed the mast with sealer and then gave it two coats of red barn paint to make it look nice and last a long time. The disc hub was a 12-inch length of 3-inch schedule-40 PVC plumbing pipe. The PVC is very tough, slightly ductile, and easy to drill and cut. PVC is well suited for RF power at the feed point of the antenna.

Three 12-foot by 0.375-inch OD pieces of 6061 aluminum, with 0.058-inch wall thickness, were used for the 12-foot diameter top disc. These were cut in half to make the center portions of the six telescoping spreaders. Four twelve foot by 0.250-inch OD (0.035-inch wall thickness) tubes were cut into 12 pieces, each 40 inches long. This gave extension tips for each end of the six spreaders.

See **Fig 52** for details on the disc hub assembly. W8NWF started by drilling six holes straight through the PVC for the six spreaders, accurately and squarely, starting

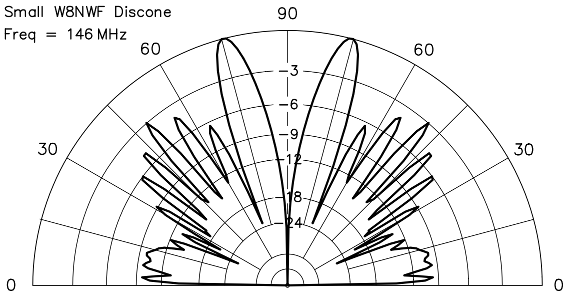

Fig49—Computed elevation plot over average ground for W8NWF's small discone at 146 MHz, ten times its design frequency range. The cone wires are acting as long-wire antennas, distorting severely the low-elevation angle response, even though the feed-point impedance is close to 50 Ω.

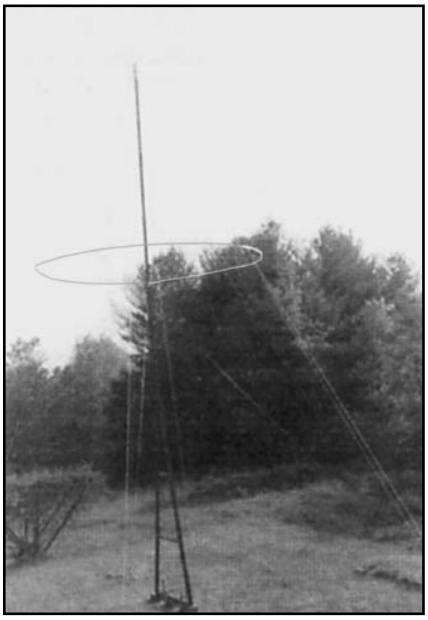

Fig 50—Photo of W8NWF's original A-frame mounted HF discone.

about two inches down from the top and spaced radially every 30°. Each hole is 0.375 inches below the plane of the previous one. Take great care in drilling—a poor job now will look bad from the ground up for a long time! It's a good idea to make up a paper template beforehand. Tape this to the PVC hub and then drill the holes, which should make for a close fit with the elements. If you goof, start over with a new piece of PVC—it's cheap.

Each six-foot spreader tube was secured exactly in the center to clear a 6-32 threaded brass rod that secured the elements mechanically and electrically. A two-foot long by ¼-inch OD wooden dowel was inserted into the middle of each six-foot length of tubing. The dowel added strength and also prevented crushing the element when the nuts on the threaded rod were tightened.

Insert the 40-inch long extensions four inches into each end of the six-foot spreaders. Mark and drill holes to pin the telescoping tips, plus holes big enough to clear #18 soft-drawn copper wire. This was for the inner circumferential wire for the disc. Drill a single hole for #18 wire about ¼ inch from each extension element tip, through which passes the outer circumferential wire. Finally, insert all 6-foot elements into the PVC hub and line up the holes in the center so the brass rod could be inserted through the middle to secure the elements.

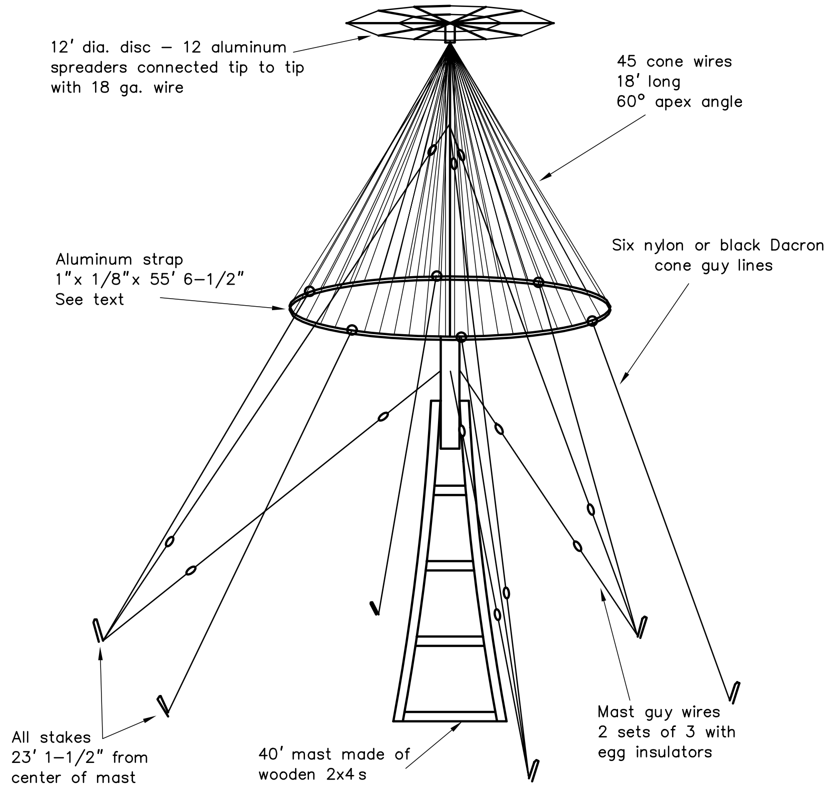
12′ dia. disc — 12 aluminum spreaders connected tip to tip with 18 ga. wire
45 cone wires 18′ long 60° apex angle
Aluminum strap 1″x 1/8″x 55′ 6-1/2″ See text
Six nylon or black Dacron cone guy lines
All stakes 23′ 1-1/2″ from center of mast
40′ mast made of wooden 2x4 s
Mast guy wires 2 sets of 3 with egg insulators

Fig 51—Detailed drawing of the A-frame discone for 14 to 30 MHz. The disc assembly at the top of the A-frame is 12 feet in diameter. There are 45 cone wires, each 18 feet long, making a 60° included angle of the cone. This antenna works very well over the design frequency range.

The next step is to "chisel to fit" the top of my wooden mast to allow the PVC to slide down on it about six or seven inches. For convenience, place the whole mast assembly in a horizontal position on top of two clothesline poles and one stepladder.

Place the disc head assembly over the top of the mast, but don't secure it yet. This allows for rotation while adding the disc spreader extensions. A tip for safety: tie white pieces of cloth to the ends of elements near eye level. Just remember to remove them before raising the antenna.

For a long-lasting installation, use an anti-corrosion compound, such as Penetrox, when assembling the aluminum antenna elements. As the extensions are added, secure them in the innermost of the two holes with a short piece of #18 wire. Then run a wire through the remaining holes looping each element as you go. This gives added support laterally to the elements. Next add a #18 wire to the tips of the extensions in the same fashion. This provides even more

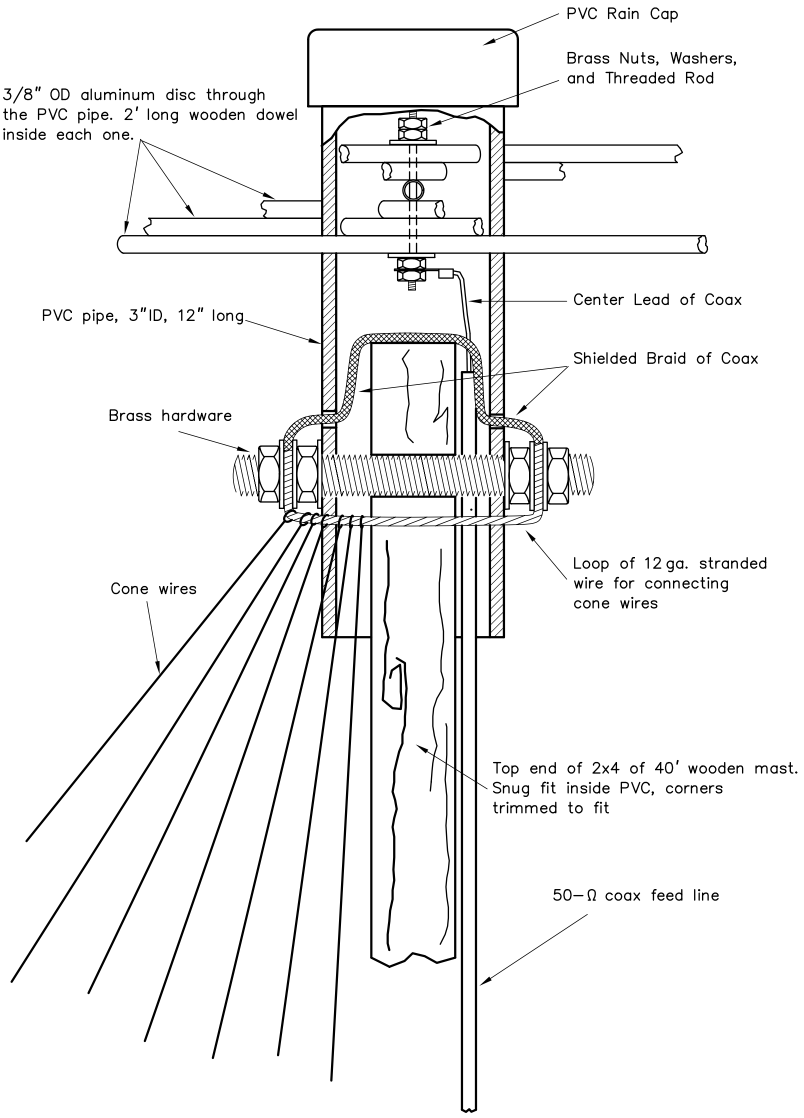
PVC Rain Cap
Brass Nuts, Washers, and Threaded Rod
3/8" OD aluminum disc through the PVC pipe. 2' long wooden dowel inside each one.
Center Lead of Coax
PVC pipe, 3"ID, 12" long
Shielded Braid of Coax
Brass hardware
Loop of 12 ga. stranded wire for connecting cone wires
Cone wires
Top end of 2x4 of 40' wooden mast. Snug fit inside PVC, corners trimmed to fit
50–Ω coax feed line

Fig 52—Details of the top hub for the A-frame discone. The three-inch PVC pipe was drilled to hold the six spreaders making up the top disc. Connections for the shield of the feed coax were made to the disc. The coax center conductor was connected to the cone-wire assembly by means of a loop of #12 stranded wire encircling the outside of the PVC hub.

physical stability as well as making electrical connections.

Next, pin the PVC disk hub to the wooden mast with a ⅜-inch threaded rod. This is also the point where the cone wires are attached, using a loop of #12 stranded copper wire around the PVC. Solder each cone wire to this loop, together with the coax shield braid. Make sure the loop of #12 wire is large enough to make soldering possible without burning the PVC with the soldering iron.

Connect the coax center conductor to the disc assembly by securing it with the same 6-32 threaded rod that ties all the disc elements together. Make sure to use coax-seal compound to keep moisture out of the coax. The coax is then fed down the mast and secured in a few places to provide strain relief and to keep it out of the way of the cone wires.

Use two sets of three guy wires. Break these up with egg insulators, just to be sure there won't be any interaction with the antenna. Use 45 wires of #18 soft-drawn copper wire for the cone, 18 feet long each. Cut them a little long so they can be soldered to the connecting loop.

A difficult task is now at hand—keeping all the cone wires from getting tangled! Solder each of the 45 cone wires to the loop of #12 wire, spacing each wire about ¼ inch from the last one for an even distribution all the way around.

The cone base is 18 feet in diameter to provide a 60° included angle. At the base of the cone, use five 12-foot long aluminum straps, 1 inche wide by ⅛ inch thick, overlapping 8¼ inches and fastened together with aluminum rivets. Drill holes along the strap every 15 inches to secure the cone wires.

Make sure to handle the aluminum strap carefully while fastening the cone wire ends; too sharp of a bend could possibly break it. Fasten six small-diameter nylon lines to the cone-base aluminum strap to stabilize the cone. These cone-guys share the same guy stakes as the mast guy lines.

After cutting the nylon lines, heat the frayed ends of each with a small flame to prevent unraveling. Apply several coats of clear protective spray to the disk head assembly, after checking that all hardware is tight. A rain cap at the top of the PVC disc hub completes construction.

Putting It All Up

You are going to need a lot of help now to raise this antenna. Have the whole process fully thought out before trying to raise it. You should have the spot selected for the base of the mast and some pipes driven into the ground to prevent the mast from slipping sideways as it is being pulled up. The three guy stakes should be in place, 23 feet, 1½ inches from mast center. Of course, the guys should have been cut to the correct length, with some extra. Be sure the coax transmission line will come off the mast where it should. A long length of rope to an upper and lower guy line is used to pull up the whole works.

The author used an old trick of standing an extension ladder vertically near the antenna base with the pull lines looped over the top rung to get a good lift angle. The weight added to the mast from the antenna disc assembly and cone wires is about 26 pounds, most of it from the cone assembly. Use two strong people to pull up the antenna slowly so that the other helpers on the guy wires and cone guy lines have time to move about as required. As the antenna rises to the vertical position, if there are no snafus, the guy lines can be secured. Then tie the six cone lines to stakes.

A Really Big Discone

When an opportunity arose to buy a 64-foot self-supporting TV tower, the author jumped at the chance to implement a full 7 to 30-MHz discone. His new tower had eight sections, each eight feet long. Counting the overlap between sections, the cone wires would come off the tower at about the 61.5-foot mark.

W8NWF took some liberties with the design of this larger discone compared to the first one, which he had done strictly “by the book.” The first change was to make the cone wires 70 feet long, even though the formula said they should be 38 feet long. Further, the cone wires would not be connected together at the bottom. With the longer cone wires, he felt that 75 and 80-meter operation might be a possibility.

The second major change was to widen the apex angle out from 60° to about 78°. Modeling said this should produce a flatter SWR over the frequency spectrum and would also give a better guy system for the tower.

The topside disc assembly would be 27 feet in diameter and have 16 radial spreaders, using telescoping aluminum tubing tapering from ⅝ to ½ to ⅜ inches OD. All spreaders were made from 0.058-inch wall thickness 6063-T832 aluminum tubing, available from Texas Towers. A section of 10-inch PVC plumbing pipe would be used as the hub for construction of the disc assembly.

Construction Details for the Large Discone.

While installing the tower, the author had left the top section on the ground. This allowed him to fit the disc head assembly precisely to it. **Fig 53** shows the overall plan for the large discone. The 10-inch diameter PVC hub was designed to slip over the tower top section, but was a little too large. So a set of shims was installed on the three legs at the top of the tower for a just-right fit. Drilling the PVC pipe for the eight ⅝-inch OD elements was started about an inch down from the top. W8NWF purposely staggered the drilled holes in the same fashion as the hub for the smaller antenna. See **Fig 54**.

Again, three-foot sections of ½-inch wooden dowel were used to strengthen the ⅝-inch center portion of each spreader. Instead of using a loop of #12 wire for connecting the cone wires, as had been done on the smaller discone, he drilled 36 holes in the PVC hub. These holes are small enough so

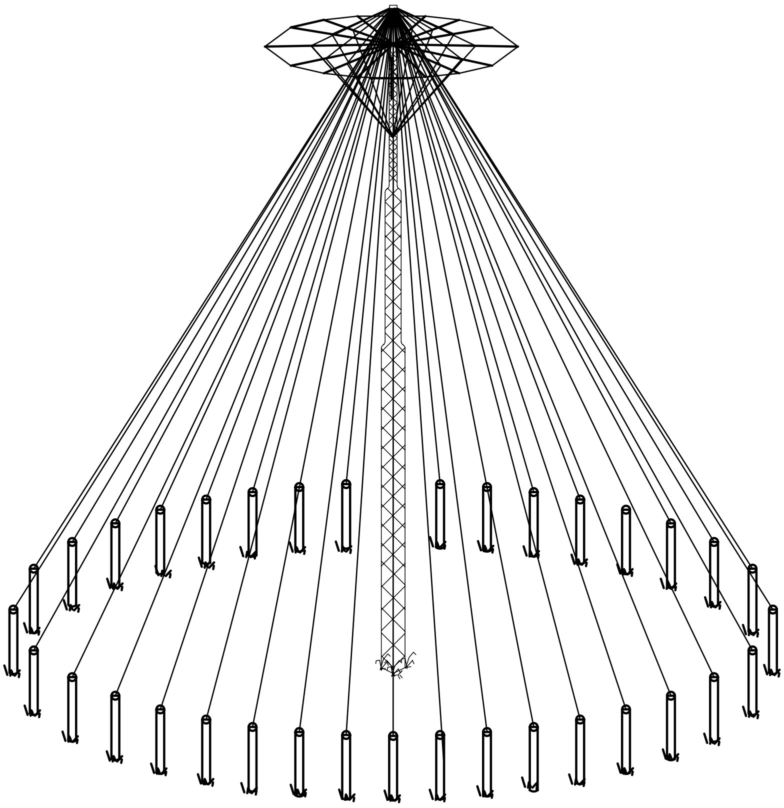

Fig 53—The large W8NWF discone, designed for operation from 7 to 14 MHz, but useable with a tuning network in the shack for 3.8 MHz.

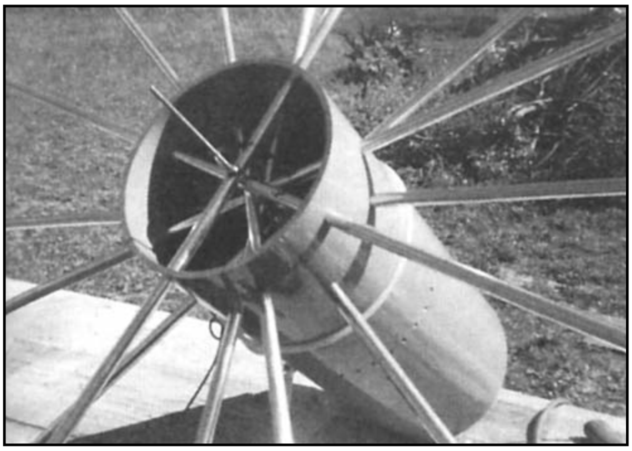

Fig 54—Photo showing details of the hub assembly for the large discone, including the threaded brass rod that connects the radial spreaders together. The 10-inch PVC pipe is drilled to accommodate the radial spreaders. Each spreader is reinforced with a three-foot long wooden dowel inside for crush resistance. Note the row of holes drilled below the lowest spreader. Each of the 36 cone wire passes through one of these holes.

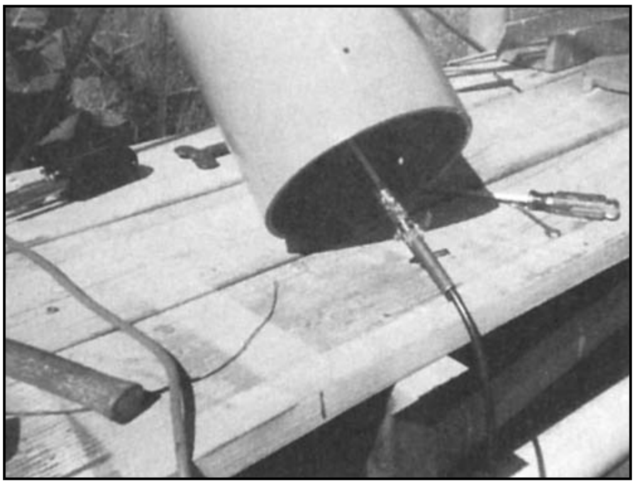

Fig 55—Details of the copper pipe slipped over the feed coax. The coax shield has been folded back over the copper pipe and secured with two stainless-steel hose clamps. The cone wires are also laid against the copper pipe and secured with additional hose clamps.

that the PVC hub would not be weakened appreciably. He drilled the circles of holes for the cone wires about 6 inches below the disc spreaders.

He prepared a three-foot long piece of RG-213 coax, permanently fastened on one end to the antenna, with a female type-N connector at the other end. Type-N fittings were used because of their superior waterproofing abilities. The coax center lead was connected with a terminal lug under a nut on the brass threaded rod securing the disc spreaders. The coax shield braid was folded back over a six-inch long copper pipe and clamped to it with a stainless-steel hose clamp. See **Fig 55** for details.

The plan was that after the top disc assembly had been hoisted up and attached at the top of the tower, individual cone wires would be fed, one at a time, through the small holes drilled in the PVC. They were to be laid against the copper pipe and secured with stainless-steel hose clamps.

The ½ and ⅜-inch OD spreader extension tips were secured in place with two aluminum pop-rivets at each joint. Again, the author used anti-oxidant compound on all spreader junctions. He drilled a hole horizontally near the tip of each ⅜-inch tip all around the perimeter to allow a #8 aluminum wire to circle the entire disc. A small stainless-steel sheet-metal screw was threaded into the end of each element to secure the wire.

In parallel with the aluminum wire, a length of small-diameter black Dacron line was run, securing it in a couple of places between each set of spreaders with UV-resistant plastic tie-wraps. The reason for doing this was to hold the aluminum wire in position and to prevent it from dangling, in case it should break some years in the future. Two coats of clear protective spray were applied for protection.

A truss system helps prevent the disc from sagging due to its own weight. See **Fig 56** for details. This shows the completed disk assembly mounted on the

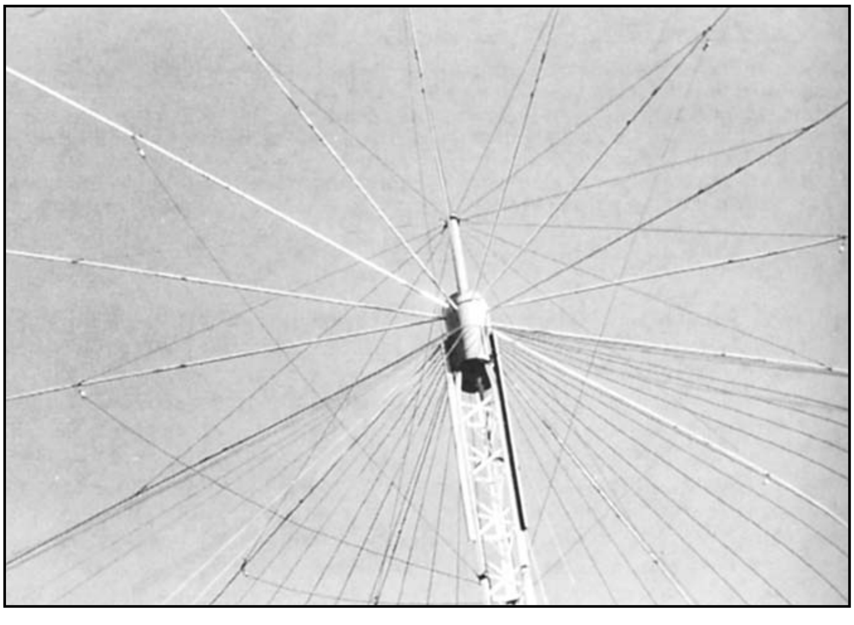

Fig 56—Photo of the spreader hub assembly, showing the truss ropes above and below the radial spreaders. This is a very rugged assembly!

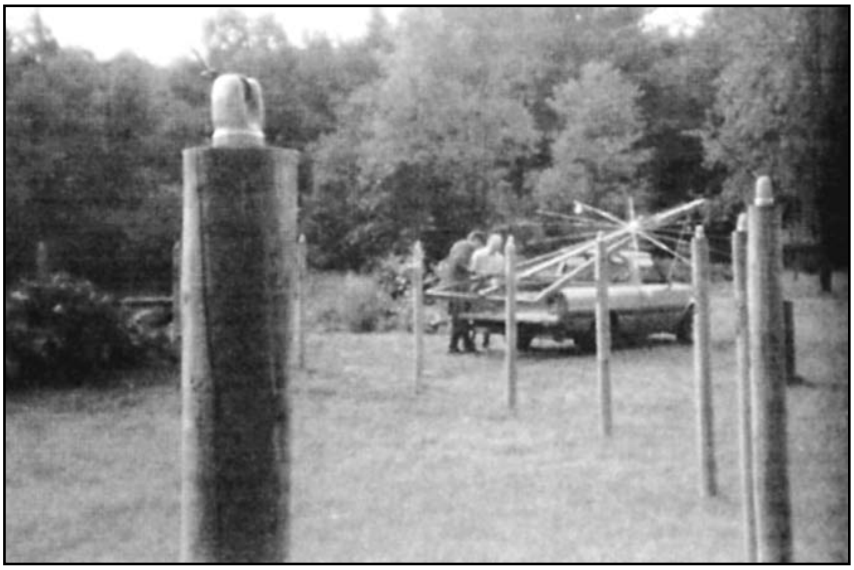

Fig 57—Photo showing some of the fence posts used to hold individual cone wires to keep them off the ground and out of harm's way. The truck in the background is carting away the A-frame discone for installation at KA8UNO's QTH.

top of the tower. A 3-foot length of 2-inch PVC pipe was used for a truss mast above the disc assembly, notching the bottom of the pipe so that it would form a saddle over the top couple of spreaders. This gave a good foothold. He cut a circle of thin sheet aluminum to fit over the 10-inch PVC to serve as a rain cap. The cap has a hole in the center for the two-inch PVC truss mast to pass through, thereby holding it down tight. The author sprayed a few light coats of paint over the PVC for protection from ultraviolet radiation from the sun.

Sixteen small-diameter black Dacron ropes were connected at the top of the truss support mast, with the other ends fastened to the disc spreaders, halfway out. Another rain cap was added to the top of the two-inch PVC truss mast. Eight lengths of the same small diameter Dacron rope were added halfway out the length of every other spreader. These ropes are meant to be tied back to the tower, to prevent updrafts from blowing the disc assembly upwards. Small egg insulators were used near the spot where the eight bottom trusses were tied to the disc spreaders, just to be sure there would be no RF leakage in rainy weather.

Hoisting the completed disc assembly to the top of the tower can be done easily, with the assistance of at least two others. The trickiest part is to get the disc assembly from its position sitting flat on the ground to the vertical position needed for hoisting it up the tower without damaging it. The disc assembly weighs about 35 pounds. Someone at the top of the tower will receive the disc as it is hoisted up by gin pole, and can mount it on the tower top.

You should prepare three 6-foot long metal braces going over the outside of the PVC to fasten to the tower legs. They really beef things up.

In plastic irrigation pipe buried between the house and tower base, the author ran 100 feet of 9086 low-loss coax to the shack. For cone wires, he was able to obtain some #18 copperclad steel wire, with heavy black insulation that looked a lot like neoprene. The cone system takes a lot of wire: 36 × 70 feet = 2520 feet, plus some extra at each end for termination. You'd be well advised to look around at hamfests to save money.

As each cone wire was connected at the top of the tower, a helper should place the other end at its proper spot below. The lower end of each cone wire is secured to an insulator screwed into a fencepost. See **Fig 57**. There are 36 treated-pine fenceposts, each standing about 5½ feet tall, 45 feet from the tower base to hold the lower end of the cone wires. This makes mowing the grass easier and the cone wires are less likely to be tripped over too.

On the final trip down the tower, the eight Dacron downward-truss lines were tied back to the tower about 6 feet below the disc assembly. The author's tower has three ground rods driven near the base, connected with heavy copper wire to the three tower legs.

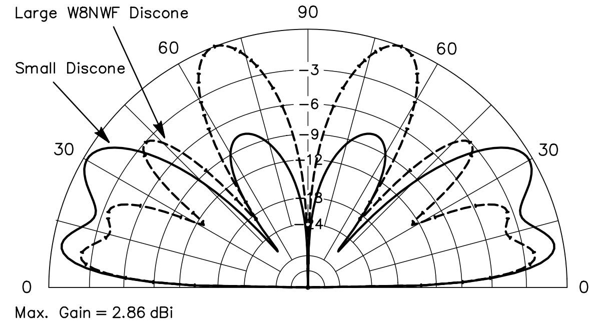

Fig 58—Computed patterns showing elevation response of small discone at 28.5 MHz compared to that of the larger discone at 28.5 MHz. The cone wires are clearly too long for efficient operation on 10 meters, producing unwanted high-angle lobes that rob power from the desirable low-elevation angles.

Performance Tests

On the air tests proved to be very satisfying. Loading up on 40 meters was easy—the SWR was 1:1 across the entire band. W8NWF can work all directions very well and receives excellent signal reports from DX stations. When he switches to his long (333 foot) center-fed dipole for comparison, he finds the dipole is much noisier and that received signals are weaker. During the daytime, nearby stations (less than about 300 to 500 miles) can be louder with the dipole, but the discone can work them just fine also.

The author happily reports that this antenna even works well on 75 meters. As you might expect, it doesn't present a 1:1 match. However, the SWR is between 3.5:1 and 5.5:1 across the band. W8NWF uses an antenna tuner to operate the discone on 75. It seems to get out as well on 75 as it does on 40 meters.

The SWR on 30 meters is about 1.1:1. On 20 meters the SWR runs from 1.05:1 at 14.0 MHz to 1.4:1 at 14.3 MHz. The SWR on the 17, 15, 12 and 10-meter bands varies, going up to a high of 3.5:1 on 12 meters.

Radiation Patterns for the Discones

From modeling using *NEC/Wires* by K6STI, W8NWF verified that the low-angle performance for the bigger antenna is worse than that for the smaller discone on the upper frequencies. See **Fig 58** for an elevation-pattern comparison on 10 meters for both antennas, with average ground constants. The azimuth patterns are simply circles. Radiation patterns produced by antenna modeling programs are very helpful to determine what to expect from an antenna.

The smaller discone, which was built by the book, displays good, low-angle lobes on 20 through 10 meters. The frequency range of 14 through 28 MHz is an octave's worth of coverage. It met his expectations in every way by covering this frequency span with low SWR and a low angle of radiation.

The bigger discone, with a modified cone suitable for use on 75 meters, presents a little different story. The low-angle lobe on 40 meters works well, and 75-meter performance also is good, although an antenna tuner is necessary on this band. The 30-meter band has a good low-angle lobe but secondary high-angle lobes are starting to hurt performance. Note that 30 meters is roughly three times the design frequency of the cone. On 20 and 17 meters there still are good low-angle lobes but more and more power is wasted in high-angle lobes.

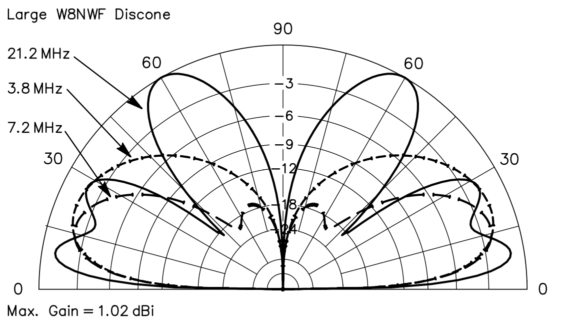

Fig 59—Computed elevation-response patterns for the larger W8NWF discone for 3.8, 7.2 and 21.2 MHz operation. Again, as in Fig 58, the pattern degrades at 21.2 MHz, although it is still reasonably efficient, if not optimal.

The operation on 15, 12, and 10 meters continues to worsen for the larger discone. The message here is that although a discone may have a decent SWR as high as 10 times the design frequency, its radiation pattern is not necessarily good for low-angle communications. See **Fig 59** for a comparison of elevation patterns for 3.8, 7.2 and 21.2 MHz on the larger discone.

A discone antenna built according to formula will work predictably and without any adjustments. One can modify the antenna's cone length and apex angle without fear of rendering it useless. The broadband feature of the discone makes it attractive to use on the HF bands. The low angle of radiation makes DX a real possibility, and the discone is also much less noisy on receive than a dipole.

Probably the biggest drawback to an HF discone is its bulky size. There is no disguising this antenna! However, if you live in the countryside you should be able to put up a nice one.

Harmonic Radiation from Multiband Antennas

Since a multiband antenna is intentionally designed for operation on a number of different frequencies, any harmonics or spurious frequencies that happen to coincide with one of the antenna resonant frequencies will be radiated with very little, if any, attenuation. Particular care should be exercised, therefore, to prevent such harmonics from reaching the antenna.

Multiband antennas using tuned feeders have a certain inherent amount of built-in protection against such radiation, since it is nearly always necessary to use a tuned coupling circuit (antenna tuner) between the transmitter and the feeder. This adds considerable selectivity to the system and helps to discriminate against frequencies other than the desired one.

Multiple dipoles and trap antennas do not have this feature, since the objective in design is to make the antenna show as nearly as possible the same resistive impedance in all the amateur bands the antenna is intended to cover. It is advisable to conduct tests with other amateur stations to determine whether harmonics of the transmitting frequency can be heard at a distance of, say, a mile or so. If they can, more selectivity should be added to the system since a harmonic that is heard locally, even if weak, may be quite strong at a distance because of propagation conditions.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

H. B. Barkley, *The Open-Sleeve As A Broadband Antenna*, Technical Report No. 14, U.S. Naval Postgraduate School, Monterey, CA, Jun 1955.

W. M. Bell, "A Trap Collinear Antenna," *QST*, Aug 1963, pp 30-31.

J. S. Belrose, "The HF Discone Antenna," *QST*, Jul 1975, pp 11-14, 56.

J. Belrose and P. Bouliane, "The Off-Center-Fed Dipole Revisited: A Broadband, Multiband Antenna," *QST*, Oct 1990, p 40.

H. J. Berg, "Multiband Operation with Paralleled Dipoles," *QST*, Jul 1956.

E. L. Bock, J. A. Nelson and A. Dorne, "Sleeve Antennas," *Very High Frequency Techniques*, H. J. Reich, ed. (New York: McGraw-Hill, 1947), Chap 5.

J. T. Bolljahn and J. V. N. Granger, "Omnidirectional VHF and UHF Antennas," *Antenna Engineering Handbook* , H. Jasik, ed. (New York: McGraw-Hill, 1961) pp 27-32 through 27-34.

J. M. Boyer, W6UYH, "Discone--40 to 500 Mc Skywire," *CQ*, July 1949, p 11.

G. A. Breed, K9AY, "Multi-Frequency Antenna Technique Uses Closely-Coupled Resonators," *RF Design*, November 1994. US Patent 5,489,914, "Method of Constructing Multiple-Frequency Dipole or Monopole Antenna Elements Using Closely-Coupled Resonators," Gary A. Breed, Feb 6, 1996.

G. H. Brown, "The Phase and Magnitude of Earth Currents Near Radio Transmitting Antennas," *Proc. IRE*, Feb 1935.

G. H. Brown, R. F. Lewis and J. Epstein, "Ground Systems as a Factor in Antenna Efficiency," *Proc. IRE* , Jun 1937, pp 753-787.

C. L. Buchanan, "The Multimatch Antenna System," *QST*, Mar 1955.

R. A. Cox, "The Open-Sleeve Antenna," *CQ*, Aug 1983, pp 13-19.

D. DeMaw, "Lightweight Trap Antennas—Some Thoughts," *QST*, Jun 1983, p 15.

W. C. Gann, "A Center-Fed 'Zepp' for 80 and 40," *QST*, May 1966.

D. Geiser, "An Inexpensive Multiband VHF Antenna," *QST*, Dec 1978, pp 28-29.

A. Greenberg, "Simple Trap Construction for the Multiband Antenna," *QST*, Oct 1956.

G. L. Hall, "Trap Antennas," Technical Correspondence, *QST*, Nov 1981, pp 49-50.

W. Hayward, "Designing Trap Antennas," Technical Correspondence, *QST*, Aug 1976, p 38.

R. H. Johns, "Dual-Frequency Antenna Traps," *QST*, Nov 1983, p 27.

A. G. Kandoian, "Three New Antenna Types and Their Applications," *Proc IEE*, Vol 34, Feb 1946, pp 70W-75W.

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ Press, 1956), pp 407-427.

W. J. Lattin, "Multiband Antennas Using Decoupling Stubs," *QST*, Dec 1960.

W. J. Lattin, "Antenna Traps of Spiral Delay Line," *QST*, Nov 1972, pp 13-15.

M. A. Logan, "Coaxial-Cable Traps," Technical Correspondence, *QST*, Aug 1985, p 43.

J. R. Mathison, "Inexpensive Traps for Wire Antennas," *QST*, Aug 1977, p 18.

L. McCoy, "An Easy-to-Make Coax-Fed Multiband Trap Dipole," *QST*, Dec 1964.

M. Mims, "The Mims Signal Squirter," *QST*, Dec 1939, p 12.

G. E. O'Neil, "Trapping the Mysteries of Trapped Antennas," *Ham Radio*, Oct 1981, pp 10-16.

B. Orr, W6SAI, "Radio FUNdamentals," The Open-Sleeve Dipole, *CQ*, February 1995.

E. W. Pappenfus, "The Conical Monopole Antenna," *QST*, Nov 1966, pp 21-24.

L. Richard, "Parallel Dipoles of 300-Ohm Ribbon," *QST*, Mar 1957.

R. Severns, N6LF, "A Wideband 80-Meter Dipole," *QST*, July 1995.

R. R. Shellenback, "Try the 'TJ'," *QST*, Jun 1982, p 18.

R. R. Shellenback, "The JF Array," *QST*, Nov 1982, p 26.

H. Scholle and R. Steins, “Eine Doppel-Windom Antenna fur Acht Bander,” *cq-DL*, Sep 1983, p 427. (In English: *QST*, Aug 1990, pp 33-34.)

T. H. Schiller, N6BT, President, Force 12, US Patent 5,995,061, “No loss, multi-band, adaptable antenna,” Nov 30, 1999.

M. Seybold, W2RYI, “The Low-Frequency Discone,” *CQ*, July 1950, p 13.

D. P. Shafer, “Four-Band Dipole with Traps,” *QST*, Oct 1958.

R. C. Sommer, “Optimizing Coaxial-Cable Traps,” *QST*, Dec 1984, p 37.

L. Varney, “The G5RV Multiband Antenna . . . Up-to-Date,” *The ARRL Antenna Compendium, Vol 1* (Newington: ARRL, 1985), pp 86.

R. Wilson, AL7KK, “The Offset Multiband Trapless Antenna (OMTA),” *QST*, October 1995.

L. G. Windom, “Notes on Ethereal Antennas,” *QST*, Sep 1929, pp 19-22, 84.

W. I. Orr, W6SAI, editor, “The Low-Frequency Discone,” *RADIO Handbook*, 14th Edition, (Editors and Engineers, 1956), p 369.

Chapter 8

Multielement Arrays

The gain and directivity offered by an array of elements represents a worthwhile improvement both in transmitting and receiving. *Power gain* in an antenna is the same as an equivalent increase in the transmitter power. But unlike increasing the power of your own transmitter, antenna gain works equally well on signals received from the favored direction. In addition, the directivity reduces the strength of signals coming from the directions not favored, and so helps discriminate against interference.

One common method of obtaining gain and directivity is to combine the radiation from a group of $\lambda/2$ dipoles to concentrate it in a desired direction. A few words of explanation may help make it clear how power gain is obtained.

In **Fig 1**, imagine that the four circles, A, B, C and D, represent four dipoles so far separated from each other that the coupling between them is negligible. Also imagine that point P is so far away from the dipoles that the distance from P to each one is exactly the same (obviously P would have to be much farther away than it is shown in this drawing). Under these conditions the fields from all the dipoles will add up at P if all four are fed RF currents in the same phase.

Let us say that a certain current, I, in dipole A will produce a certain value of field strength, E, at the distant point P. The same current in any of the other dipoles will produce the same field at P. Thus, if only dipoles A and B are operating, each with a current I, the field at P will be 2E. With A, B and C operating, the field will be 3E, and with all four operating with the same I, the field will be 4E. Since the power received at P is proportional to the square of the field strength, the relative power received at P is 1, 4, 9 or 16, depending on whether one, two, three or four dipoles are operating.

Now, since all four dipoles are alike and there is no coupling between them, the same power must be put into each in order to cause the current I to flow. For two dipoles the relative power input is 2, for three dipoles it is 3, for four dipoles 4, and so on. The actual gain in each case is the relative received (or output) power divided by the relative input power. Thus we have the results shown in **Table 1**. The power ratio is directly proportional to the number of elements used.

It is well to have clearly in mind the conditions under which this relationship is true:

1) The fields from the separate antenna elements must be in-phase at the receiving point.

2) The elements are identical, with equal currents in all elements.

3) The elements must be separated in such a way that the current induced in one by another is negligible; that is, the radiation resistance of each element must be the same as it would be if the other elements were not there.

Very few antenna arrays meet all these conditions exactly. However, the power gain of a directive array using dipole elements with optimum values of element spacing is approximately proportional to the number of elements.

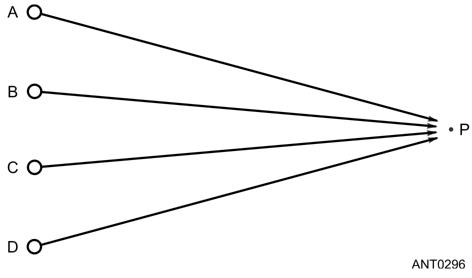

Fig 1—Fields from separate antennas combine at a distant point, P, to produce a field strength that exceeds the field produced by the same power in a single antenna.

Table 1
Comparison of Dipoles with Negligible Coupling (See Fig 1)

Dipoles	*Relative Output Power*	*Relative Input Power*	*Power Gain*	*Gain in dB*
A only	1	1	1	0
A and B	4	2	2	3
A, B and C	9	3	3	4.8
A, B, C and D	16	4	4	6

Another way to say this is that a gain of approximately 3 dB will be obtained each time the number of elements is doubled, assuming the proper element spacing is maintained. It is possible, though, for an estimate based on this rule to be in error by a ratio factor of two or more (gain error of 3 dB or more), especially if mutual coupling is *not* negligible.

DEFINITIONS

An *element* in a multi-element directive array is usually a λ/2 radiator or a λ/4 vertical element above ground. The length is not always an exact electrical half or quarter wavelength, because in some types of arrays it is desirable that the element show either inductive or capacitive reactance. However, the departure in length from resonance is ordinarily small (not more than 5% in the usual case) and so has no appreciable effect on the radiating properties of the element.

Antenna elements in multi-element arrays of the type considered in this chapter are always either *parallel*, as in **Fig 2A** , or *collinear* (end-to-end), as in Fig 2B. Fig 2C shows an array combining both parallel and collinear elements. The elements can be either horizontal or vertical, depending on whether horizontal or vertical polarization is desired. Except for space communications, there is seldom any reason for mixing polarization, so arrays are customarily constructed with all elements similarly polarized.

A *driven element* is one supplied power from the transmitter, usually through a transmission line. A *parasitic element* is one that obtains power solely through coupling to another element in the array because of its proximity to such an element.

A *driven array* is one in which all the elements are driven elements. A *parasitic array* is one in which one or more of the elements are parasitic elements. At least one element must be a driven element, since you must somehow introduce power into the array.

A *broadside array* is one in which the principal direction of radiation is perpendicular to the axis of the array and to the plane containing the elements, as shown in **Fig 3**. The elements of a broadside array may be collinear, as in Fig 3A, or parallel (two views in Fig 3B).

An *end-fire array* is one in which the principal direction of radiation coincides with the direction of the array axis. This definition is illustrated in **Fig 4**. An end-fire array must consist of parallel elements. They cannot be collinear, as λ/2 elements do not radiate straight off their ends. A Yagi is a familiar form of an end-fire array.

A *bidirectional array* is one that radiates equally well in either direction along the line of maximum radiation. A bidirectional pattern is shown in **Fig 5A**. A *unidirectional array* is one that has only one principal direction of radiation, as the pattern in Fig 5B shows.

The *major lobes* of the directive pattern are those in which the radiation is maximum. Lobes of lesser radiation intensity are called *minor lobes*. The *beamwidth* of a directive antenna is the width, in degrees, of the major lobe between the two directions at which the relative radiated power is equal to one half its value at the peak of the lobe. At these *half-power points* the field intensity is equal to 0.707 times its maximum value, or in other words, is down 3 dB from the maximum. **Fig 6** shows a lobe having a beamwidth of 30°.

Unless specified otherwise, the term *gain* as used in this

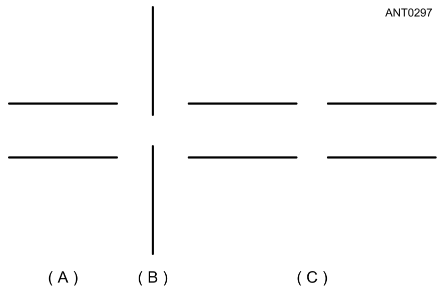

Fig 2—At A, parallel and at B, collinear antenna elements. The array shown at C combines both parallel and collinear elements.

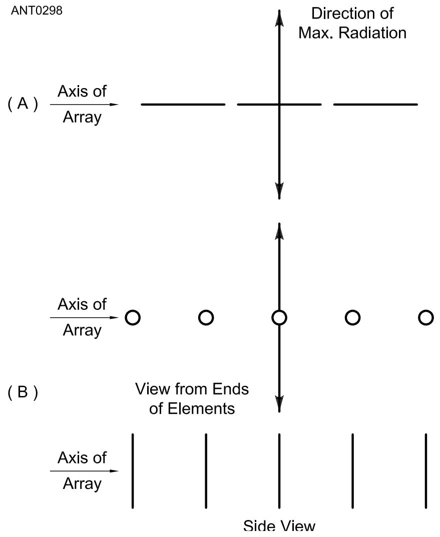

Fig 3—Representative broadside arrays. At A, collinear elements, with parallel elements at B.

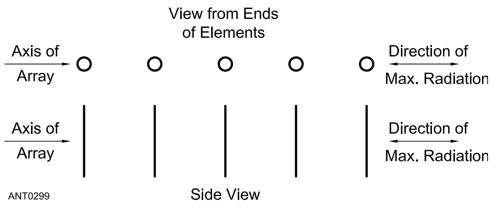

Fig 4—An end-fire array. Practical arrays may combine both broadside directivity (Fig 3) and end-fire directivity, including both parallel and collinear elements.

section is the power gain over an isotropic radiator in free space. The gain can also be compared with a λ/2 dipole of the same orientation and height as the array under discussion and having the same power input. Gain may either be measured experimentally or determined by calculation. Experimental measurement is difficult and often subject to considerable error, for two reasons. First, errors normally occur in measurement because the accuracy of simple RF measuring equipment is relatively poor—even high-quality instruments suffer in accuracy compared with their low-frequency and dc counterparts. And second, the accuracy depends considerably on conditions—the antenna site, including height, terrain characteristics, and surroundings—under which the measurements are made.

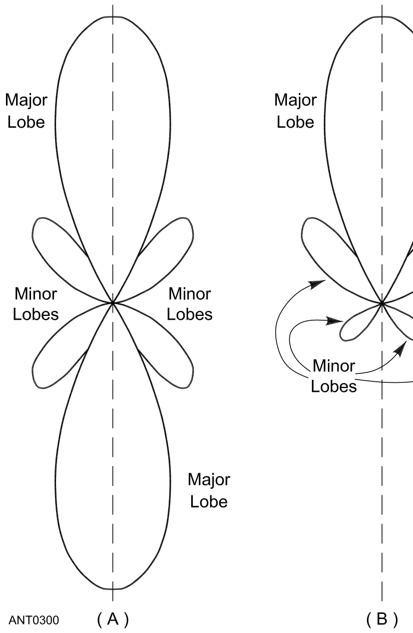

Fig 5—At A, typical bidirectional pattern and at B, unidirectional directive pattern. These drawings also illustrate the application of the terms *major* and *minor* to the pattern lobes.

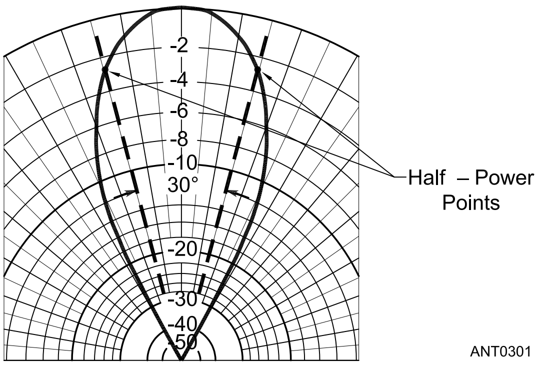

Fig 6—The width of a beam is the angular distance between the directions at which the received or transmitted power is half the maximum power (–3 dB). Each angular division of the pattern grid is 5°.

Calculations are frequently based on the measured or theoretical directive patterns of the antenna (see Chapter 2). The theoretical gain of an array may be determined approximately from:

$$G = 10 \log \frac{41{,}253}{\theta_H \; \theta_V} \qquad \text{(Eq 1)}$$

where

G = decibel gain over a dipole in its favored direction
θ_H = horizontal half-power beamwidth in degrees
θ_V = vertical half-power beamwidth in degrees.

This equation, strictly speaking, applies only to lossless antennas having approximately equal and narrow E- and H-plane beam widths—up to about 20°—and no large minor lobes. The E and H planes are discussed in Chapter 2. The error may be considerable when the formula is applied to simple directive antennas having relatively large beam widths. The error is in the direction of making the calculated gain larger than the actual gain.

Front-to-back ratio (F/B) is the ratio of the power radiated in the favored direction to the power radiated in the opposite direction. See Chapter 11 for a discussion of front-to-back ratio, and its close cousin, *worst-case front-to-rear ratio*.

Phase

The term *phase* has the same meaning when used in connection with the currents flowing in antenna elements as it does in ordinary circuit work. For example, two currents are in-phase when they reach their maximum values, flowing in the same direction, at the same instant. The direction of current flow depends on the way in which power is applied to the element.

This is illustrated in **Fig 7**. Assume that by some means an identical voltage is applied to each of the elements at the ends marked A. Assume also that the coupling between the elements is negligible, and that the instantaneous polarity of the voltage is such that the current is flowing away from the point at which the voltage is applied. The arrows show the assumed current directions. Then the currents in elements 1 and 2 are in-phase, since they are flowing in the same direction in space and are caused by the same voltage. However, the current in element 3 is flowing in the *opposite* direction in space because the voltage is applied to the opposite end of the element. The current in element 3 is therefore 180° out-of-phase with the currents in elements 1 and 2.

The phasing of driven elements depends on the direction of the element, the phase of the applied voltage, and the point at which the voltage is applied. In many systems used by amateurs, the voltages applied to the elements are exactly

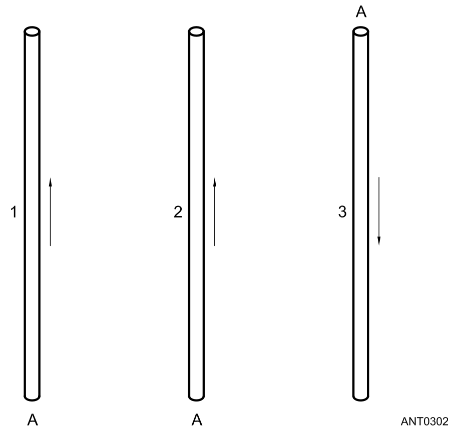

Fig 7—This drawing illustrates the phase of currents in antenna elements, represented by the arrows. The currents in elements 1 and 2 are in phase, while that in element 3 is 180° out of phase with 1 and 2.

in or exactly out-of-phase with each other. Also, the axes of the elements are nearly always in the same direction, since parallel or collinear elements are invariably used. The currents in driven elements in such systems therefore are usually either exactly in or exactly out-of-phase with the currents in other elements.

It is possible to use phase differences of less than 180° in driven arrays. One important case is where the current in one set of elements differs by 90° from the current in another set. However, making provision for proper phasing in such systems is considerably more complex than in the case of simple 0° or 180° phasing, as described in a later section of this chapter.

In parasitic arrays the phase of the currents in the parasitic elements depends on the spacing and tuning, as described later.

Ground Effects

The effect of the ground is the same with a directive antenna as it is with a simple dipole antenna. The reflection factors discussed in Chapter 3 may therefore be applied to the vertical pattern of an array, subject to the same modifications mentioned in that chapter. In cases where the array elements are not all at the same height, the reflection factor for the mean height of the array may be used for a close approximation. The mean height is the average of the heights measured from the ground to the centers of the lowest and highest elements.

MUTUAL IMPEDANCE

Consider two $\lambda/2$ elements that are fairly close to each other. Assume that power is applied to only one element, causing current to flow. This creates an electromagnetic field, which induces a voltage in the second element and causes current to flow in it as well. The current flowing in element 2 will in turn induce a voltage in element 1, causing additional current to flow there. The total current in 1 is then the sum (taking phase into account) of the original current and the induced current.

With element 2 present, the amplitude and phase of the resulting current in element 1 will be different than if element 2 were not there. This indicates that the presence of the second element has changed the impedance of the first. This effect is called *mutual coupling*. Mutual coupling results in a *mutual impedance* between the two elements. The mutual impedance has both resistive and reactive components. The actual impedance of an antenna element is the sum of its self-impedance (the impedance with no other antennas present) and its mutual impedances with all other antennas in the vicinity.

The magnitude and nature of the feed-point impedance of the first antenna depends on the amplitude of the current induced in it by the second, and on the phase relationship between the original and induced currents. The amplitude and phase of the induced current depend on the spacing between the antennas and whether or not the second antenna is tuned to resonance.

In the discussion of the several preceding paragraphs, power is applied to only one of the two elements. Do not interpret this to mean that mutual coupling exists only in parasitic arrays! It is important to remember that mutual coupling exists between any two conductors that are located near one another.

Amplitude of Induced Current

The induced current will be largest when the two antennas are close together and are parallel. Under these conditions the voltage induced in the second antenna by the first, and in the first by the second, has its greatest value and causes the largest current flow. The coupling decreases as the parallel antennas are moved farther apart.

The coupling between collinear antennas is comparatively small, and so the mutual impedance between such antennas is likewise small. It is not negligible, however.

Phase Relationships

When the separation between two antennas is an appreciable fraction of a wavelength, a measurable period of time elapses before the field from antenna 1 reaches antenna 2. There is a similar time lapse before the field set up by the current in number 2 gets back to induce a current in number 1. Hence the current induced in antenna 1 by antenna 2 will have a phase relationship with the original current in antenna 1 that depends on the spacing between the two antennas.

The induced current can range all the way from being completely in-phase with the original current to being completely out-of-phase with it. If the currents are in-phase, the total current is larger than the original current and the antenna feed-point impedance is reduced. If the currents are out-of-phase, the total current is smaller and the impedance is increased. At intermediate phase relationships the impedance will be lowered or raised depending on whether the induced current is mostly in or mostly out-of-phase with the original current.

Except in the special cases when the induced current is exactly in or out-of-phase with the original current, the induced current causes the phase of the total current to shift with respect to the applied voltage. Consequently, the presence of a second antenna nearby may cause the impedance of an antenna to be reactive—that is, the antenna will be detuned from resonance—even though its self-impedance is entirely resistive. The amount of detuning depends on the magnitude and phase of the induced current.

Tuning Conditions

A third factor that affects the impedance of antenna 1 when antenna 2 is present is the tuning of number 2. If antenna 2 is not exactly resonant, the current that flows in it as a result of the induced voltage will either lead or lag the phase it would have if the antenna were resonant. This causes an additional phase advance or delay that affects the phase of the current induced back in antenna 1. Such a phase lag has an effect similar to a change in the spacing between self-resonant antennas. However, a change in tuning is not exactly equivalent to a change in spacing because the two methods do not have the same effect on the amplitude of the induced current.

MUTUAL IMPEDANCE AND GAIN

The mutual coupling between antennas is important because it can have a significant effect on the amount of current that will flow for a given amount of power supplied. And it is the amount of *current* flowing that determines the field strength from the antenna. Other things being equal, if the mutual coupling between two antennas is such that the currents are greater for the same total power than would be the case if the two antennas were not coupled, the power gain will be greater than that shown in Table 1.

On the other hand, if the mutual coupling is such as to reduce the current, the gain will be less than if the antennas were not coupled. The term *mutual coupling*, as used in this paragraph, assumes that the mutual impedance between elements is taken into account, along with the added effects of propagation delay because of element spacing and element tuning or phasing.

The calculation of mutual impedance between antennas is a complex problem. Data for two simple but important cases are graphed in Figs 8 and 9. These graphs do not show the mutual impedance, but instead show a more useful quantity—the feed-point resistance measured at the center of an antenna as it is affected by the spacing between two antennas.

As shown by the solid curve in **Fig 8**, the feed-point resistance at the center of either antenna, when the two are self-resonant, parallel, and operated in-phase, decreases as the spacing between them is increased until the spacing is about 0.7 λ. This is a broadside array. The maximum gain is achieved from a pair of such elements when the spacing is in this region, because the current is larger for the same power and the fields from the two arrive in-phase at a distant point placed on a line perpendicular to the line joining the two antennas.

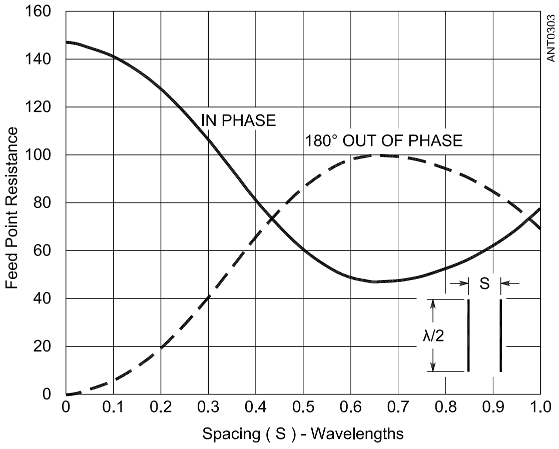

Fig 8—Feed-point resistance measured at the center of one element as a function of the spacing between two parallel ½-λ self-resonant antenna elements. For ground-mounted ¼-λ vertical elements, divide these resistances by two.

The dashed line in Fig 8, representing two antennas operated 180° out-of-phase (end-fire), cannot be interpreted quite so simply. The feed-point resistance decreases with spacing decreasing less than about 0.6 λ in this case. However, for the range of spacings considered, only when the spacing is 0.5 λ do the fields from the two antennas add up exactly in phase at a distant point in the favored direction. At smaller spacings the fields become increasingly out-of-phase, so the total field is less than the simple sum of the two. Smaller spacings thus decrease the gain at the same time that

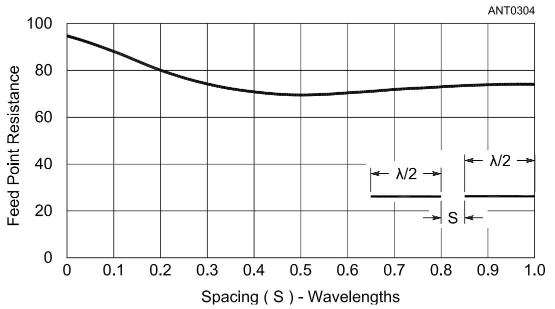

Fig 9—Feed-point resistance measured at the center of one element as a function of the spacing between the ends of two collinear self-resonant ½-λ antenna elements operated in phase.

the reduction in feed-point resistance is increasing it. For a lossless antenna, the gain goes through a maximum when the spacing is in the region of ⅛ λ.

The feed-point resistance curve for two collinear elements in-phase, **Fig 9**, shows that the feed-point resistance decreases and goes through a broad minimum in the region of 0.4 to 0.6-λ spacing between the adjacent ends of the antennas. As the minimum is not significantly less than the feed-point resistance of an isolated antenna, the gain will not exceed the gain calculated on the basis of uncoupled antennas. That is, the best that two collinear elements will give, even with optimum spacing, is a power gain of about 2 (3 dB). When the separation between the ends is very small—the usual method of operation—the gain is reduced.

GAIN AND ARRAY DIMENSIONS

The gain of an array is principally determined by the dimensions of the array, so long as there are a minimum number of elements. A good example of this is the relationship between boom length, gain and number of elements for an array such as a Yagi. **Fig 10** compares the gain versus boom length for Yagis with different numbers of elements. For given number of elements, notice that the gain increases as the boom length increases, up to a maximum. Beyond this point, longer boom lengths result in less gain for a given number of elements. This observation does not mean that it is always desirable to use only the minimum number of elements. Other considerations of array performance, such as front-to-back ratio, minor lobe amplitudes or operating bandwidth, may make it advantageous to use more than the minimum number of elements for a given array length. A specific example of this is presented in a later section in a comparison between a half-square, a bobtail curtain and a Bruce array.

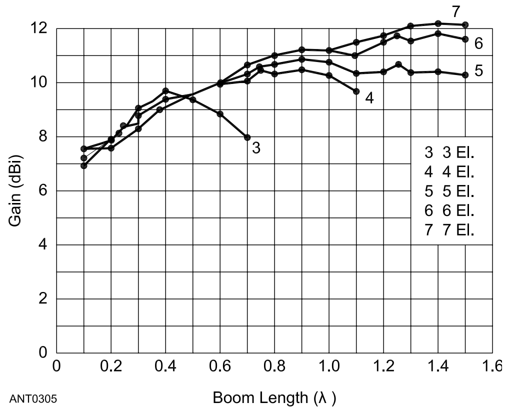

Fig 10—Yagi gain for 3, 4, 5, 6 and 7-element beams as a function of boom length. (From *Yagi Antenna Design*, J. Lawson, W2PV.)

In a broadside array the gain is a function of both the length and width of the array. The gain can be increased by adding more elements (with additional spacing) or by using longer elements (>λ/2), although the use of longer elements requires proper attention to current phase in the elements. In general, in a broadside array the element spacing that gives maximum gain for a minimum number of elements, is in the range of 0.5 to 0.7 λ. Broadside arrays with elements spaced for maximum gain will frequently have significant side lobes and associated narrowing of the main lobe beamwidth. Side lobes can be reduced by using more than the minimum number of elements, spaced closer than the maximum gain distance.

Additional gain can be obtained by expanding the array into a third dimension. An example of this is the stacking of endfire arrays in a broadside configuration. In the case of stacked short endfire arrays, maximum gain occurs with spacings in the region of 0.5 to 0.7 λ. However, for longer higher-gain end-fire arrays, larger spacing is required to achieve maximum gain. This is important in VHF and UHF arrays, which often use long-boom Yagis.

PARASITIC ARRAYS

The foregoing applies to multi-element arrays of both types, driven and parasitic. However, there are special considerations for driven arrays that do not necessarily apply to parasitic arrays, and vice versa. Such considerations for Yagi and quad parasitic arrays are presented in Chapters 11 and 12. The remainder of this chapter is devoted to driven arrays.

Driven Arrays

Driven arrays in general are either broadside or end-fire, and may consist of collinear elements, parallel elements, or a combination of both. From a practical standpoint, the maximum number of usable elements depends on the frequency and the space available for the antenna. Fairly elaborate arrays, using as many as 16 or even 32 elements, can be installed in a rather small space when the operating frequency is in the VHF range, and more at UHF. At lower frequencies the construction of antennas with a large number of elements is impractical for most amateurs.

Of course the simplest of driven arrays is one with just two elements. If the elements are collinear, they are always fed in-phase. The effects of mutual coupling are not great, as illustrated in Fig 9. Therefore, feeding power to each element in the presence of the other presents no significant problems. This may not be the case when the elements are parallel to each other. However, because the combination of spacing and phasing arrangements for parallel elements is infinite, the number of possible radiation patterns is endless.

This is illustrated in **Fig 11**. When the elements are fed in-phase, a broadside pattern always results. At spacings of less than ⅝ λ with the elements fed 180° out-of-phase, an end-fire pattern always results. With intermediate amounts of phase difference, the results cannot be so simply stated. Patterns evolve that are not symmetrical in all four quadrants.

Because of the effects of mutual coupling between the two driven elements, for a given power input greater or lesser currents will flow in each element with changes in spacing and phasing, as described earlier. This, in turn, affects the gain of the array in a way that cannot be shown merely by plotting the *shapes* of the patterns, as has been done in Fig 11. Therefore, supplemental gain information is also shown in Fig 11, adjacent to the pattern plot for each combination of spacing and phasing. The gain figures shown are referenced to a single element. For example, a pair of elements fed 90° apart at a spacing of λ/4 will have a gain in the direction of maximum radiation of 3.1 dB over a single element.

Current Distribution in Phased Arrays

In the plots of Fig 11, the two elements are assumed to be identical and self-resonant. In addition, currents of equal amplitude are assumed to be flowing at the feed point of each element, a condition that most often will not exist in practice without devoting special consideration to the feeder system. Such considerations are discussed in the next section of this chapter.

Most literature for radio amateurs concerning phased arrays is based on the assumption that if all elements in the array are identical, the *current distribution* in all the elements will be identical. This distribution is presumed to be that of a single, isolated element, or nearly sinusoidal. However, information published in the professional literature as early as the 1940s indicates the existence of dissimilar current distributions among the elements of phased arrays. (See Harrison and King references in the Bibliography.) Lewallen, in July 1990 *QST*, pointed out the causes and effects of dissimilar current distributions.

In essence, even though the two elements in a phased array may be identical and have exactly equal currents of the desired phase flowing *at the feed point*, the amplitude and phase relationships degenerate with departure from the feed point. This happens any time the phase relationship is not 0° or 180°. Thus, the field strengths produced at a distant point by the individual elements may differ. This is because the field from each element is determined by the *distribution* of the current, as well as its magnitude and phase.

The effects are minimal with shortened elements—verticals less than λ/4 or dipoles less than λ/2 long. The effects on radiation patterns begin to show at the above resonant lengths, and become profound with longer elements—λ/2 or longer verticals and 1 λ or longer center-fed elements. These effects are less pronounced with thin elements. The amplitude and phase degeneration takes place because the currents in the array elements are not sinusoidal. Even in two-element arrays with phasing of 0° or 180°, the currents are not sinusoidal, but in these two special cases they do remain identical.

The pattern plots of Fig 11 take element current distributions into account. The visible results of dissimilar distributions are incomplete nulls in some patterns and the development of very small minor lobes in others. For example, the pattern for a phased array with 90° spacing and 90° phasing has traditionally been published in amateur literature as a cardioid with a perfect null in the rear direction. Fig 11, calculated for 7.15-MHz self-resonant dipoles of #12 wire in free space, shows a minor lobe at the rear and only a 33-dB front-to-back ratio.

It is characteristic of broadside arrays that the power gain is proportional to the length of the array but is substantially independent of the number of elements used, provided the optimum element spacing is not exceeded. This means, for example, that a five-element array and a six-element array will have the same gain, provided the elements in both are spaced so the overall array length is the same. Although this principle is seldom used for the purpose of reducing the number of elements because of complications introduced in feeding power to each element in the proper phase, it does illustrate the fact that there is nothing to be gained, in terms of more gain, by increasing the number of elements if the space occupied by the antenna is not increased proportionally.

Generally speaking, the maximum gain in the smallest linear dimensions will result when the antenna combines both broadside and end-fire directivity and uses both parallel and collinear elements. In this way the antenna is spread over a greater volume of space, which has the same effect as extending its length to a much greater extent in one linear direction.

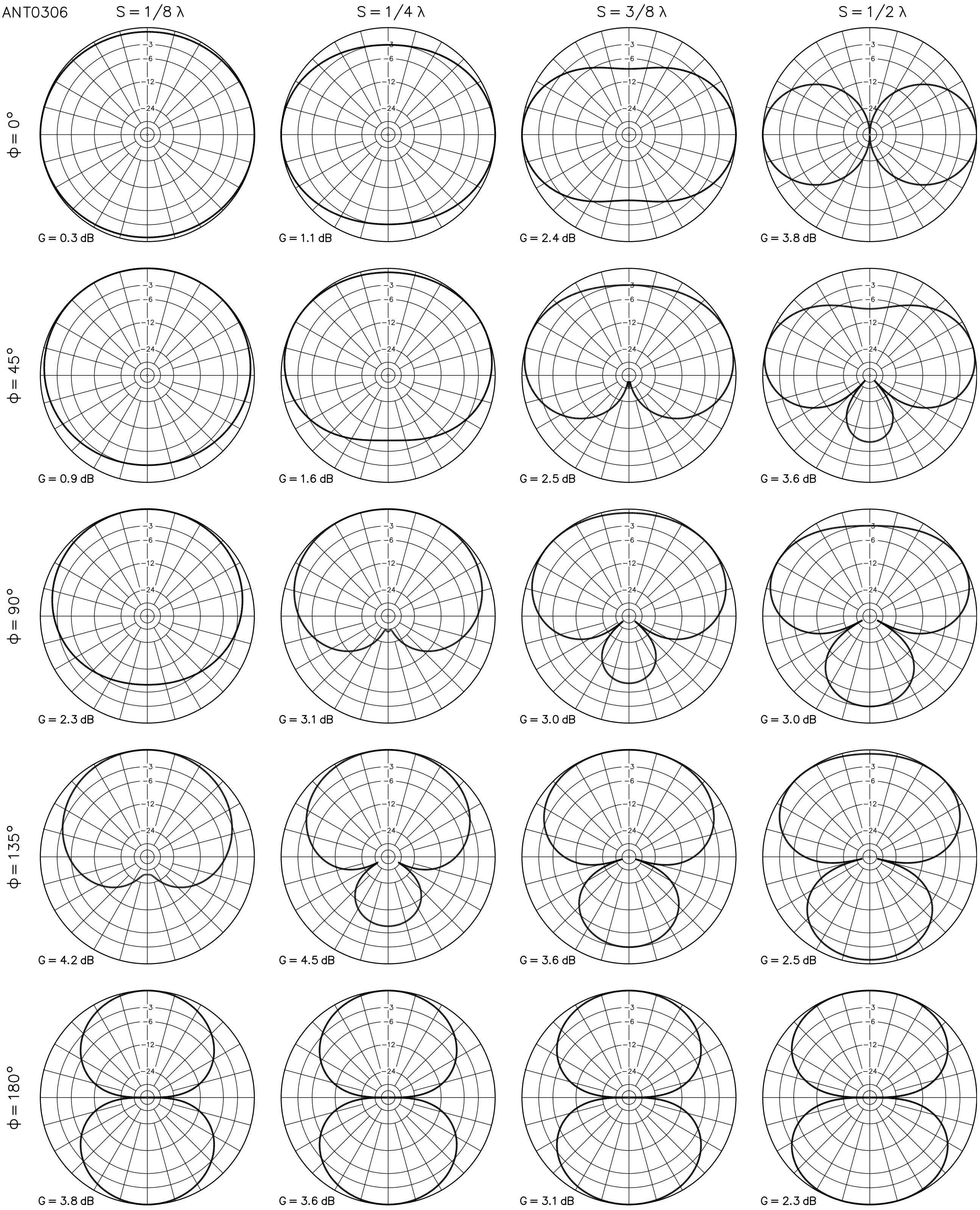

Fig 11—H-plane patterns of two identical parallel driven elements, spaced and phased as indicated (S = spacing, φ = phasing). The elements are aligned with the vertical (0°-180°) axis, and the element nearer the 0° direction (top of page) is of lagging phase at angles other than 0°. The two elements are assumed to be thin and self-resonant, with equal-amplitude currents flowing at the feed point. See text regarding current distributions. The gain figure associated with

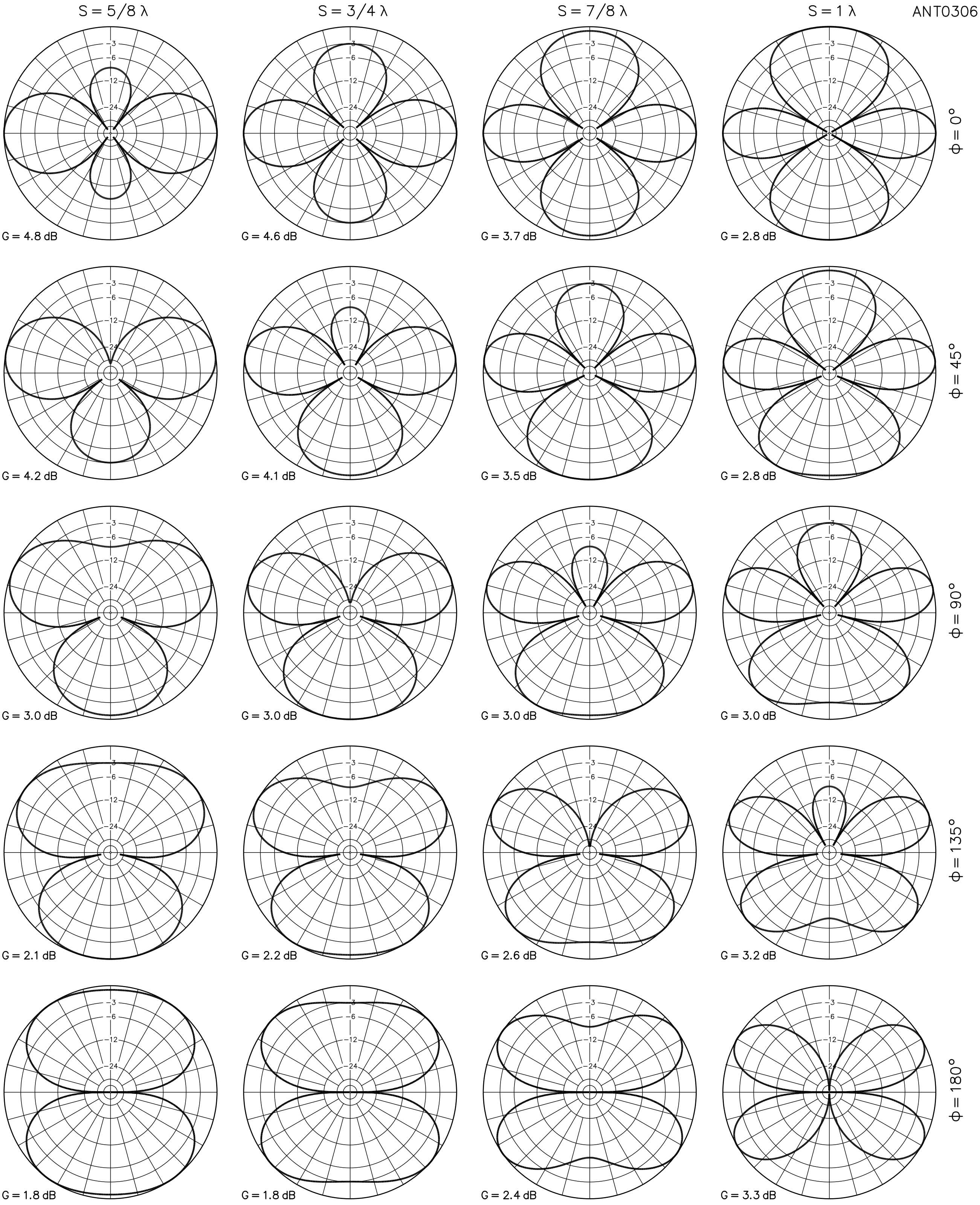

each pattern indicates that of the array over a single element. The plots represent the horizontal or azimuth pattern at a 0° elevation angle of two ¼-λ vertical elements over a perfect conductor, or the free-space vertical or elevation pattern of two horizontal ½-λ elements when viewed on end, with one element above the other. (Patterns computed with *ELNEC*—see Bibliography.)

Phased Array Techniques

Phased antenna arrays have become increasingly popular for amateur use, particularly on the lower frequency bands, where they provide one of the few practical methods to obtain substantial gain and directivity. This section on phased-array techniques was written by Roy Lewallen, W7EL.

The operation and limitations of phased arrays, how to design feed systems to make them work properly and how to make necessary tests and adjustments are discussed in the pages that follow. The examples deal primarily with vertical HF arrays, but the principles apply to VHF/UHF arrays and arrays made from other element types as well.

OVERVIEW

Much of this chapter is devoted to techniques for feeding phased arrays. Many people who have a limited acquaintance with phased array techniques believe this is a simple problem, consisting only of connecting array elements through "phasing lines" consisting of transmission lines of the desired electrical lengths. Unfortunately, except for a very few special cases, this approach won't achieve the desired array pattern.

Other proposed universal solutions, such as hybrid couplers or Wilkinson or other power dividers, also usually fail to achieve the necessary phasing. These approaches sometimes produce—often more by accident than design—results good enough to mislead the user into believing that the simple approach is working as planned. Confusion can result when an approach fails to work in different circumstances. This section will explain why the simple solutions don't work as often thought, and how to design feed systems that do consistently produce the desired results.

Very briefly, the reason why the simple phasing-line approach fails is that the delay of current or voltage in a transmission line equals the line's electrical length only if the line is terminated in its characteristic impedance. And in phased arrays, element feed-point impedances are profoundly affected by mutual coupling.

Consequently, even if each element has the correct impedance when isolated, it won't when all elements are excited. Furthermore, transmission lines that are not terminated in their characteristic impedance will transform both the voltage and current magnitude. The net result is that the array elements will have neither the correct magnitudes nor phases of current necessary for proper operation except in a few special cases. This isn't a minor effect of concern only to perfectionists, but often a major one that causes significant pattern distortion and poor or mislocated nulls. The problem is examined in greater depth later.

Power dividers and hybrid couplers also fail to achieve the desired result for different reasons, which will be discussed below, although in one common application hybrid couplers fortuitously provide results that are acceptable to many users. This chapter will show how to design array feed systems that *will* produce predicted element currents and array patterns.

Various *EZNEC* models are provided to illustrate concepts presented in this chapter. They can all be viewed with the *EZNEC-ARRL* software furnished on the CD included with this book. Step-by-step instructions for the examples are given in **Appendix A**.

FUNDAMENTALS OF PHASED ARRAYS

The performance of a phased array is determined by several factors. Most significant among these are the characteristics of a single element, reinforcement or cancellation of the fields from the elements and the effects of mutual coupling. To understand the operation of phased arrays, it is first necessary to understand the operation of a single antenna element.

Of primary importance is the strength of the field produced by the element. The field radiated from a linear (straight) element, such as a dipole or vertical monopole, is proportional to the sum of the elementary currents flowing in each part of the antenna element. For this discussion it is important to understand what determines the current in a single element.

The amount of current flowing at the base of a ground mounted vertical or ground-plane antenna is given by the familiar formula

$$I = \sqrt{\frac{P}{R}} \qquad \text{(Eq 2)}$$

where

P is the power supplied to the antenna
R is the feed-point resistance.

R consists of two parts, the loss resistance and the radiation resistance. The loss resistance, R_L, includes losses in the conductor, in the matching and loading components and dominantly (in the case of ground-mounted verticals) in ground losses. The power "dissipated" in the radiation resistance, R_R, is actually the power that is radiated, so maximizing the power dissipated by the radiation resistance is desirable. However, the power dissipated in the loss resistance truly is lost as heat, so resistive losses should be made as small as possible.

The radiation resistance of an element can be derived from electromagnetic field theory, being a function of antenna length, diameter and geometry. Graphs of radiation resistance versus antenna length are given in Chapter 2. The radiation resistance of a thin resonant $\lambda/4$ ground-mounted vertical is about 36 Ω. A resonant $\lambda/2$ dipole in free space has a radiation resistance of twice this amount, about 73 Ω. Reducing the antenna lengths by one half drops the radiation resistances to approximately 7 and 14 Ω, respectively.

The radiation resistance of a large variety of antennas can easily be determined by using *EZNEC-ARRL*, which is included on the CD in the back of this book. The radiation

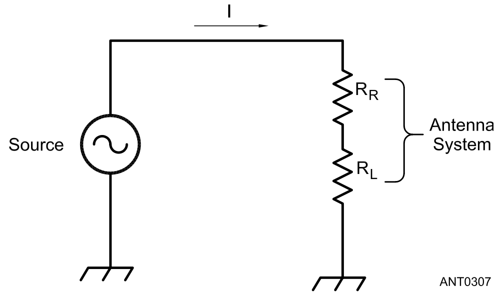

Fig 12—Simplified equivalent circuit for a single-element resonant antenna. R_R represents the radiation resistance, and R_L the ohmic losses in the total antenna system.

resistance is simply the feed-point resistance (the resistive part of the feed-point impedance) when all losses have been set to zero.

Radiation Efficiency

To generate a stronger field from a given radiator, it is necessary to increase the power P (the brute-force solution), decrease the loss resistance R_L (by putting in a more elaborate ground system for a vertical, for instance) or to somehow decrease the radiation resistance R_R so more current will flow with a given power input. This can be seen by expanding the formula for base current as

$$I = \sqrt{\frac{P}{R_R + R_L}} \qquad \text{(Eq 3)}$$

Splitting the feed-point resistance into components R_R and R_L easily leads to an understanding of element efficiency. The efficiency of an element is the proportion of the total power that is actually radiated. The roles of R_R and R_L in determining efficiency can be seen by analyzing a simple equivalent circuit, shown in **Fig 12**.

The power dissipated in R_R (the radiated power) equals I^2R_R. The total power supplied to the antenna system is

$$P = I^2 (R_R + R_L) \qquad \text{(Eq 4)}$$

so the efficiency (the fraction of supplied power that is actually radiated) is

$$\text{Eff} = \frac{I^2 R_R}{I^2 (R_R + R_L)} = \frac{R_R}{R_R + R_L} \qquad \text{(Eq 5)}$$

Efficiency is frequently expressed in percent, but expressing it in decibels relative to a 100%-efficient radiator gives a better idea of what to expect in the way of signal strength. The field strength of an element relative to a lossless but otherwise identical element, in dB, is

$$\text{FSG} = 10 \log \frac{R_R}{R_R + R_L} \qquad \text{(Eq 6)}$$

where FSG = field strength gain, dB.

For example, information presented by Sevick in March 1973 *QST* shows that a λ/4 ground-mounted vertical antenna with four 0.2-λ radials has a feed-point resistance of about 65 Ω (see the Bibliography at the end of this chapter). The efficiency of such a system is 36/65 = 55.4%. It is rather disheartening to think that, of 100 W fed to the antenna, only 55 W is being radiated, with the remainder literally warming up the ground. Yet the signal will be only 10 log (36/65) = –2.57 dB relative to the same vertical with a perfect ground system. In view of this information, trading a small reduction in signal strength for lower cost and greater simplicity may become an attractive consideration.

So far, only the current at the base of a resonant antenna has been discussed, but the field is proportional to the sum of currents in each tiny part of the antenna. The field is a function of not only the magnitude of current flowing at the feed point, but also the distribution of current along the radiator and the length of the radiator. Nothing can be done at the feed point to change the current distribution, so for a given element the field strength is proportional to the feed point current. However, changing the radiator length or loading it at some point other than the feed point will change the current distribution.

More information on shortened or loaded radiators may be found in Chapters 2 and 6 and in the Bibliography references of this chapter. The current distribution is also changed by mutual coupling to other array elements, although for most arrays this has only a minor effect on the pattern. This is discussed later in more detail. A few other important facts follow.

1) If there is no loss, the field from even an infinitesimally short radiator is less than ½ dB weaker than the field from a half-wave dipole or quarter-wave vertical. Without loss, all the supplied power is radiated regardless of the antenna length, so the only factor influencing gain is the slight difference in the patterns of very short and λ/2 antennas. The small pattern difference arises from different current distributions. A short antenna has a very low radiation resistance, resulting in a heavy current flow over its short length. In the absence of loss, this generates a field strength comparable to that of a longer antenna. Where loss is present—that is, in practical antennas—shorter radiators usually don't do so well, since the low radiation resistance leads to lower efficiency for a given loss resistance. If care is taken, reasonably short antennas can achieve good efficiency.

2) Care has to be taken in calculating the efficiency of folded antennas. (See Chapter 6.) Folding transforms both the radiation resistance and loss resistance by the same factor, so their ratio and therefore the efficiency remains the same. It's easy to show that in a ground-mounted vertical array, folding reduces the current flowing from the feed line to the ground system by a factor of two due to the impedance transformation. However, the folded antenna has an additional connection to ground, which also carries half the original ground current. The result is that the same amount of current flows

into the ground system, whether unfolded or folded, resulting in the same ground system loss. Analyses purporting to show otherwise invariably transform the radiation resistance but neglect to also transform the loss resistance and reach an incorrect conclusion.

3) The current flowing in an element with a given power input can be increased or decreased by mutual coupling to other elements. The effect is equivalent to changing the element radiation resistance. Mutual coupling is sometimes thought of as a minor effect, but often it is not minor!

Field Reinforcement and Cancellation

The mechanism by which phased arrays produce gain, and the role of mutual coupling in determining gain, were covered earlier in this chapter. One important point that can't be emphasized enough is that all antennas must abide by the law of *conservation of energy*. No antenna can radiate more power than supplied to it. The total amount of power it radiates is the amount it's supplied, less the amount lost as heat. This is true of all antennas, from the smallest "rubber ducky" to the most gigantic array.

Gain

Gain is strictly a relative measure, so the term is completely meaningless unless accompanied by a statement of just what it is relative to. One useful measure for phased array gain is *gain relative to a single similar element*. This is the increase in signal strength that would be obtained by replacing a single element by an array made from elements just like it. In some instances, such as investigating what happens to array performance when *all* elements become more lossy, it's useful to state gain relative to a more absolute, although unattainable standard: a lossless element.

And the most universal reference for gain is another unattainable standard, the *isotropic radiator*. This fictional antenna radiates absolutely equally in all directions. It's very useful because the field strength resulting from any power input is readily calculated, so if the gain relative to this standard is known, the field strength is also known for any radiated power. Gain relative to this reference is referred to as dBi, and it's the standard used by most modeling programs including *EZNEC-ARRL*. To find the gain of an array relative to a single element or other reference antenna such as a dipole, model both the array and the single element or other reference antenna in the same environment and subtract their dBi gains. Don't rely on some assumption about the gain of a single element—many people assume values that can be very wrong.

Nulls

Pattern nulls are very often more important to users of phased arrays than gain because of their importance in reducing both man-made and natural interference when receiving. Consequently, a good deal of emphasis is, and should be, placed on achieving good pattern nulls. Unfortunately, good nulls are much more difficult to achieve than gain and they are much more sensitive to array and feed-system imperfections.

As an illustration, consider two elements that each produce a field strength of, say, exactly 1 millivolt per meter (mV/m) at some distance many wavelengths from the array. In the direction in which the fields from the elements are in-phase, a total field of 2 mV/m results. In the direction in which they're out-of-phase, zero field results. The ratio of maximum to minimum field strength of this array is 2/0, or infinity.

Now suppose, instead, that one field is 10% high and the other 10% low—1.1 and 0.9 mV/m, respectively. In the forward direction, the field strength is still 2 mV/m, but in the canceling direction, the field will be 0.2 mV/m. The front-to-back ratio has dropped from infinity to 2/0.2, or 20 dB. (Actually, slightly more power is required to redistribute the field strengths this way, so the forward gain is reduced—but by only a small amount, less than 0.1 dB.) For most arrays, unequal fields from the elements have a minor effect on forward gain, but a major effect on pattern nulls. This is illustrated by ***EZNEC* Example: Nulls** in Appendix A.

Even with perfect current balance, deep nulls aren't assured. **Fig 13** shows the minimum spacing required for total field reinforcement or cancellation. If the element spacing isn't adequate, there may be no direction in which the fields are completely out-of-phase (see curve B of Fig 13). Slight physical and environmental differences between elements will invariably affect null depths, and null depths will also vary with elevation angle.

However, a properly designed and fed array can produce very impressive nulls. The key to producing good nulls, like producing gain, is controlling the strengths and phases of the fields from the elements. Just how to accomplish that is the subject of most of the remainder of this section. But be sure to keep in mind that producing good nulls is generally a much more difficult task than producing approximately the predicted gain.

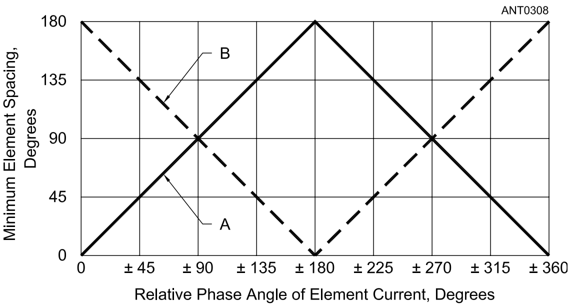

Fig 13—Minimum element spacing required for total field reinforcement, curve A, or total field cancellation, curve B. Total cancellation results in pattern nulls in one or more directions. Total reinforcement does not necessarily mean there is gain over a single element, as the effects of loss and mutual coupling must also be considered.

Mutual Coupling

Mutual coupling was discussed briefly earlier in this chapter. Because it has an important and profound effect on both the performance and feed system design of phased arrays, it will be covered in greater depth here.

Mutual coupling refers to the effects which the elements in an array have on each other. Mutual coupling can occur intentionally or entirely unintentionally. People with multiple antennas on a small lot (or car top) often discover that a better description of their system is a single antenna with multiple feed points. Current is induced in conductors in various antennas by mutual coupling, causing them to act like parasitic elements, which re-radiate and distort the antenna's pattern. The effects of mutual coupling are present whether or not the elements are driven.

Suppose that two driven elements are many wavelengths from each other. Each has some voltage and current at its feed point. For each element, the ratio of this voltage to current is the element self-impedance. If the elements are brought close to each other, the current in each element will change in magnitude and phase because of coupling with the field from the other element. The field from the first element changes the current in the second. This changes the field from the second, which alters the current in the first, and so forth until an equilibrium condition is reached in which the currents in all elements (hence, their fields) are totally interdependent.

The feed-point impedances of all elements also are changed from their values when far apart, and all are dependent on each other. In a driven array, the changes in feed-point impedances can cause additional changes in element currents, because the operation of many feed systems depends on the element feed-point impedances. Significant mutual coupling occurs at spacings as great as a wavelength or more.

Connecting the elements to a feed system to form a driven array does not eliminate the effects of mutual coupling. In fact, in many driven arrays the mutual coupling has a greater effect on antenna operation than the feed system does. All feed-system designs must account for the impedance changes caused by mutual coupling if the desired current balance and phasing are to be achieved.

Several general statements can be made regarding the effects of mutual coupling on phased-array systems.

1) The resistances and reactances of all elements of an array generally will be substantially different from the values when the elements are isolated (that is, very distant from other elements).

2) If the elements of a two-element array are identical and have equal currents that are in-phase or 180° out-of-phase, the feed-point impedances of the two elements will be equal. But they will be different than for an isolated element. If the two elements are part of a larger array, their impedances can be very different from each other.

3) If the elements of a two-element array have currents that are neither in-phase (0°) nor out-of-phase (180°), their feed-point impedances will not be equal. The difference will be considerable in typical amateur arrays.

4) The feed-point resistances of the elements in a closely spaced, 180° out-of-phase array will be very low, resulting in poor efficiency due to ohmic losses unless care is taken to minimize loss. This is also true for any other closely spaced array with significant predicted gain.

It's essential to realize that this is *not* a minor effect and one that can be overlooked or ignored. See ***EZNEC* Example—Mutual Coupling** in Appendix A for an illustration of these phenomena.

Loss Resistance, Mutual Coupling and Antenna Gain

Loss reduces the effects of mutual coupling because the feed-point impedance change resulting from mutual coupling is effectively in series with loss resistance. If the loss is great enough, two important results occur. First, the feed-point impedance becomes independent of the presence of nearby current-carrying elements. This greatly simplifies feed system design—the simple "phasing-line" or hybrid-coupler feed system described below is adequate *provided that all elements are physically identical and the feed point of each element is matched to the Z_0 of the feed line and, if used, the hybrid coupler.*

The impedance matching restrictions are necessary to insure that the phasing line or hybrid coupler performs as expected. Identical elements are needed so that equal element currents will result in equal fields from the elements.

In the absence of mutual coupling effects, the maximum gain of an array of identical elements relative to a single (similarly lossy) element is simply 10 log(N), where N is the number of elements—providing that spacing is adequate for the fields to fully reinforce in some direction. If spacing is less, maximum gain will also be less. Of course, the array gain relative to a single *lossless* element will be very low, most likely a sizeable negative number when expressed in dB. So intentionally introducing loss isn't a wise idea for a transmitting array. It is sometimes an advantageous thing to do for a receiving array, however, as explained in the following section.

High-gain close-spaced arrays, such as the W8JK phased array (see *EZNEC-ARRL* example file **ARRL_W8JK.EZ** and accompanying Antenna Notes file), and most parasitic arrays depend heavily on mutual coupling to achieve their gain. Introduction of any loss to these arrays, which reduces the mutual coupling effects, has a profound effect on the gain. Consequently, parasitic or close-spaced driven arrays often produce disappointing results when made from grounded vertical elements unless each has a fairly elaborate (and therefore very low-loss) ground system.

If you place two low-loss elements very close together and feed them in-phase, mutual coupling reduces the array gain to essentially that of a single element, so there's no advantage to this configuration over a single element. However, if you have a single lossy element, for example a short vertical having a relatively poor ground system, you can improve the gain by up to 3 dB by adding a second, close spaced, element and ground system and feeding the two in-phase. Another way to look at this technique is that you're putting two equal ground system resistances in parallel, which effectively cuts the loss in half. The gain you can realize in practice depends on such things as the ground system overlap, but it might be a practical way to improve transmitting array performance in some situations.

FEEDING PHASED ARRAYS

The previous section explains why the fields from the elements must be very close to the ratios required by the array design. Since the field strengths are proportional to the currents in the elements, controlling the fields requires controlling the element currents. Since the desired current ratio is 1:1 for virtually all two-element and for most (but not all) larger amateur arrays, special attention is paid to methods of assuring equal element currents. But we will examine other current ratios also.

The Role of Element Currents

The field from a conductor is proportional to the current flowing on it. So if we're to control the relative strengths and phases of the fields from the elements, we have to control their currents. We usually do this by controlling the currents at the element feed points. But because the field from an element depends on the current everywhere along the element, elements having identical feed-point currents will produce different fields if they have different current distributions—that is, if the way the current varies along the lengths of the elements is different.

A previous section explained that mutual coupling alters the current distribution, so in many arrays the current distributions will be different on the elements and consequently the relationship between the overall fields won't be the same as that between the feed-point currents. Fortunately, this effect is relatively minor in thin, λ/4 monopole or λ/2 dipole elements. The most common arrays are made from elements in this category, so we can generally get very nearly the desired ratio of fields by effecting the same ratio of feed point currents. Exceptions are detailed immediately below.

Feed-point vs Element Current

For most antennas, environmental factors are likely to cause greater performance anomalies than current distribution differences, and both can be corrected with minor feed system adjustments. The difference between field and feed-point current ratios can become very significant, however, if the elements are very fat and/or close to λ/2 (monopole) or 1 λ (dipole) long. In those cases, most of the feed systems described here won't produce the desired field ratios without major adjustment or modification, except in the special cases of 2-element arrays with identical elements having feed point currents in-phase or 180° out-of-phase. In those special cases, the element current distributions are the same for the same reason the feed-point impedances are equal. This is explained later in the feed system sections.

To get an idea of just how large an element must be to disturb the pattern of an array with correct feed-point currents, a two-element cardioid array of quarter wave vertical elements was modeled at 10 MHz. With thin, 0.1-inch diameter elements, the front/back ratio was 35 dB, the very small reverse lobe caused by slightly unequal element current distributions. Increasing the element diameter to 20 inches decreased the front/back ratio to 20 dB. Returning the front/back ratio of the array of 20-inch elements to >35 dB required changing the feed-point current ratio from the nominal value of 1.0 at an angle of 90° to 0.88 at 83°.

The same array was first modeled with 0.1 inch diameter elements, where it has a front/back ratio of 35 dB, then the elements were lengthened. The front/back ratio dropped to 20 dB at an element length of 36 feet, or about 0.37 λ. In that case, adjustment of the feed-point current ratio to about 0.9 at about 83° restored a good front/back ratio.

In the discussion and development which follow, the assumption is made that the fields will be very nearly proportional to the feed-point currents. If the elements are fat or long enough to make this assumption untrue, some adjustment of feed-point current ratio will be necessary to achieve the desired pattern, particularly nulls. Most feed systems can be designed for any current ratio. Modeling will reveal the ratio required for the desired pattern, and then the feed system can be designed accordingly.

COMMON PHASED-ARRAY FEED SYSTEMS

This section will first describe several popular approaches to feeding phased arrays that often don't produce the desired results. It will describe why they don't work as well as hoped. It also briefly discusses systems that could be used, but that often aren't appropriate or optimum for amateur arrays.

This will be followed in the next section by detailed descriptions of array feed systems that *do* produce the predicted element current ratios and array patterns.

The "Phasing-Line" Approach

For an array to produce the desired pattern, the element currents must have the required magnitude and the required phase relationship. As explained above, this can generally be achieved well enough by causing the feed point currents to have that same relationship.

On the surface, this sounds easy—just make sure that the difference in electrical lengths of the feed lines to the elements equals the desired phase angle. Unfortunately, this approach doesn't necessarily achieve the desired result. The first problem is that the phase shift through the line is not equal to its electrical length. The current (or, for that matter, voltage) delay in a transmission line is equal to its electrical length in only a few special cases—cases which don't exist in most amateur arrays! The impedance of an element in an array is frequently very different from the impedance of an isolated element and the impedances of all the elements in an array can be different from each other.

See the ***EZNEC*** **Example—Mutual Coupling** in Appendix A for a graphic illustration of the effect of mutual coupling on feed-point impedance. Also look at the Four-Square array example in the **Phased Array Design Examples** section. The array in that example has one element with a *negative* feed-point resistance, if ground loss is low. Without mutual coupling, the resistance of that same element would be about 36 Ω plus ground loss.

Because of mutual coupling, the elements seldom provide a matched load for the element feed lines. The effect of mismatch on phase shift can be seen in **Fig 14**. Observe what happens to the phase of the current and voltage on a line terminated by a purely resistive impedance that is lower than the characteristic impedance of the line (Fig 14A). At a point 45° from the load the current has advanced less than 45°, and the voltage more than 45°. At 90° from the load both are advanced 90°. At 135° the current has advanced more and the voltage less than 135°. This apparent slowing down and speeding up of the current and voltage waves is caused by interference between the forward and reflected waves. It occurs on any line that is not terminated with a pure resistance equal to its characteristic impedance. If the load resistance is greater than the characteristic impedance of the line, as shown in Fig 14B, the voltage and current exchange angles. Adding reactance to the load causes additional phase shift. The *only cases in which the current (or voltage) delay is equal to the electrical length of the line are*

1) When the line is flat; that is, terminated in a purely resistive load equal to its characteristic impedance;

2) When the line length is an integral number of half wavelengths;

3) When the line length is an odd number of quarter wavelengths and the load is purely resistive; and

4) When other specific lengths are used for specific load impedances.

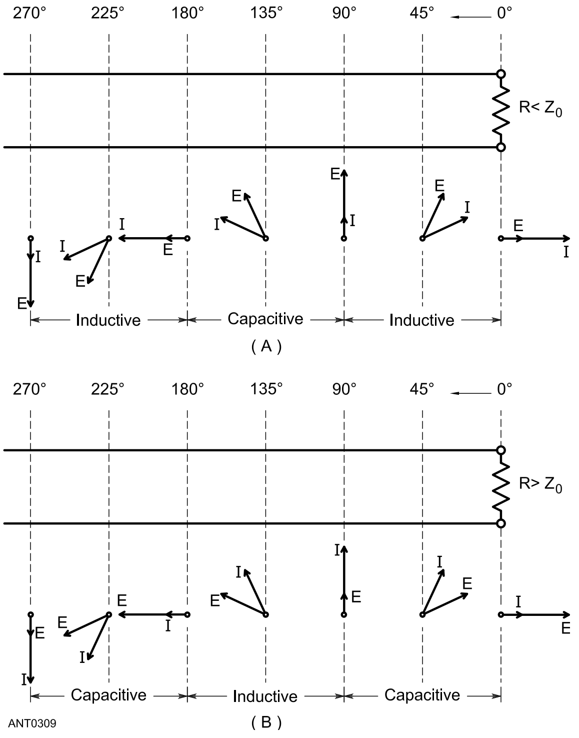

Fig 14—Resultant voltages and currents along a mismatched line. At A, R less than Z_0, and at B, R greater than Z_0.

Just how much phase error can be expected if two feed lines are simply hooked up to form an array? There is no simple answer. Some casually designed feed systems might deliver satisfactory results, but most will not. See the ***EZNEC* Example—"Phasing-Line" Feed** in Appendix A for the typical consequences of using this sort of feed system.

A second problem with simply connecting feed lines of different lengths to the elements is that the lines will change the *magnitudes* of the currents. The magnitude of the current (or voltage) out of a line does not equal the magnitude into that line, except in cases 1, 2 and 4 above. The feed systems presented later in this chapter assure currents that are correct in both magnitude and phase.

The elementary phasing-line approach *will* work in three very special but common situations. If the array consists of only two identical elements and those elements are fed in-phase, mutual coupling will modify the element impedances, but both will be modified exactly the same amount. Consequently, if the two elements are fed through equal-length transmission lines, the lines will transform and delay the currents by the same amount and result in equal, in-phase currents at the element feed points.

Similarly, an array of two identical elements fed 180º out-of-phase will have the same feed-point impedances and can be fed with two lines of any length so long as one line is an electrical half wavelength longer than the other. But this can't be extended to any two elements in a larger array, since mutual coupling to the other elements can result in different feed-point impedances. Methods will be described later which do assure a correct current ratio in this situation.

The third application in which the phasing-line approach works is in receiving arrays where the elements are very short in terms of wavelength and/or very lossy. In either of these cases, mutual coupling between elements is much less than an element's self-impedance. This allows the elements to be individually matched to the feed lines, with no significant change taking place when the elements are formed into an array. Under those conditions, the transmission lines can be matched and the lines used as simple delay lines with easily predictable phase shift and with no transformation of current or voltage magnitude other than cable loss. This is discussed in the later section on receiving antennas.

Many arrays *can* be correctly fed with a feed system consisting of only transmission lines, but the technique requires knowledge of the element feed-point impedances in a correctly fed array. Line lengths can then be computed that provide the correct ratio of currents into those particular load impedances. The line lengths generally differ by amount that's considerably different from the element phase angle difference, and appropriate line lengths can't always be found for all arrays. This technique is described more fully in the **"The Simplest Phased Array Feed System—That Works"** section later in the chapter and illustrated in the examples in **Phased Array Design Examples**.

The Wilkinson Divider

The *Wilkinson divider*, sometimes called the *Wilkinson power divider*, was once heavily promoted as a means to distribute power among the elements of a phased array. While it's a very useful device for other purposes, it won't produce the desired current ratios in antenna elements. In most phased arrays, element feed-point resistances are different and therefore require different amounts of power to achieve the desired equal magnitude currents. (See the section on mutual coupling above.) A Wilkinson divider is intended to deliver equal powers, not currents, to multiple loads. And it won't even do that when the load impedances are different.

The Hybrid Coupler

Hybrid couplers are promoted as solving the problem of achieving equal magnitude currents with a 90º phase difference between elements. Unfortunately, they provide equal magnitude, quadrature (90° phased) currents only when the load impedances are equal and correct. And this simply isn't true of arrays with quadrature-fed elements, except for arrays consisting of short and/or lossy elements, usually suitable only for receiving. In those arrays, the hybrid coupler can be useful for the same reasons as the phasing-line approach, discussed in an earlier section.

At the time of this writing, hybrid couplers are being used in a popular commercial product for phasing at least one type of array. Reports are that it works satisfactorily. However, this shouldn't be taken as proof that the hybrid coupler forces equal magnitude 90º phased currents in loads of arbitrary impedances. No passive network, including the hybrid coupler, is capable of doing that. See **The "Magic Bullet"** below for more information.

Large Array Feed Systems

The author once worked on a radar system where the transmit array consisted of over 5,000 separate dipole elements and the receive array over 4,000 pairs of crossed dipoles, all over a metal reflecting plane, which was the sloping side of a 140 foot high building. In such large arrays, each element is in essentially the same environment as every other element except near the array edges, so almost all elements have very nearly the same feed-point impedance. While producing the phase shifts and magnitude tapers is a considerable mathematical challenge, the problem of unequal element feed-point impedances can largely be ignored. Consequently, feed methods for these large arrays are generally not suitable for typical amateur arrays of a few elements.

The Broadcast Approach

Networks can be designed to transform the element base impedances from their values in an excited array to, say, 50 Ω resistive. Then another network can be inserted at the junction of the feed lines to properly divide the power among the elements (not necessarily equally!). And finally, additional networks must be added to correct for the phase shifts and magnitude transformations of the other networks. This general approach is used by the broadcast industry, in installations that are typically adjusted only once for a particular frequency and pattern.

Although this technique can be used to correctly feed any type of array, design is difficult and adjustment is tedious, since all adjustments interact. When the relative currents and phasings are adjusted, the feed-point impedances change, which in turn affect the element currents and phasings, and so on. A further disadvantage of using this method is that switching the array direction is generally impossible. Information on applying this technique to amateur arrays can be found in Paul Lee's book, listed in the Bibliography.

The "Magic Bullet"

For about 15 years prior to this writing, this *Antenna Book* section has contained a specification for a hypothetical passive circuit that would provide equal-magnitude, 90° phased currents into two loads without respect to the load impedances. This would be a circuit to which we could connect any two elements and guarantee that they'd have exactly the correct currents.

Along with the specifcation was a request that any person knowing of such a circuit would contact the author (Roy Lewallen, W7EL) or the book editors. During this time, only a single response was received, in 1996. It was from Kevin Schmidt, W9CF, who has formulated a mathematical proof that such a circuit—in fact, one resulting in any relative phase other than 0º or 180º—*cannot exist* if restricted to reciprocal elements. (That is, it can't exist unless directional components such as ferrite circulators are used.) This means that, in order to design a network to feed elements at any other phase angle other than 0 or 180º, we *must* know the impedance of at least one element, and correct feed system operation depends on that impedance. There's no way around this requirement. At the time of this writing, Schmidt's proof can be found at **fermi.la.asu.edu/w9cf/articles/magic/index.html**.

RECOMMENDED FEED METHODS FOR AMATEUR ARRAYS

The following feed methods are able to produce element feed-point currents having a desired magnitude and phase relationship, resulting in desirable and predictable patterns. Most methods require knowing the feed-point impedance of one or more array elements *when the array element currents are the correct values*. This isn't possible to measure directly, because if the element currents were correct, the feed system would already be working properly and no further design would be necessary.

By far the easiest way to get this information, if possible, is by computer modeling. Modeling programs such as *EZNEC-ARRL* (included on the CD in the back of this book) allow you to construct an ideal array with perfect element currents, then look at the resulting feed-point impedances. Because of its simplicity and versatility, this approach is highly recommended and it's the one used for the array design examples in this chapter.

Some feed systems allow adjustment, so even an approximate result provides an adequate starting point on which to base the feed-system design. There are several other alternatives to computer modeling. One is to first eliminate the effects of coupling of the element to be measured from all other elements, usually by open circuiting the feed points of the other elements. Then the feed-point impedance of the element is measured. Next, the impedance change due to mutual coupling from all other elements has to be calculated, based on the intended currents in the other elements, their lengths and their distances from the element being measured. Mutual impedance (which is not the same as the impedance *change* due to mutual coupling) between each pair of elements must be known for this calculation and it can be determined by measurement, calculation or from a graph.

The latter two methods are possible only for the simplest element types and measurement is very difficult to do accurately because it involves resolving very small differences between two relatively large values. Accuracy of a calculated result will be reduced if any elements are relatively fat (that is, they have a large diameter, because this impacts the current distribution) or they aren't perfectly straight and parallel.

So the only situations where you're likely to get good results from approaches other than modeling are the very easiest ones to model! And modeling allows determination of the feed-point impedances of many antennas that areimpossible to calculate by manual or graphical methods. Therefore, the manual approach isn't discussed or used here. Appendix B, on the CD, contains equations and manual techniques from previous editions of *The ARRL Antenna Book,* for those who are interested. You can also find a great deal of additional information in many of the texts listed in the Bibliography, particularly Jasik and Johnson.

Current Forcing with λ/4 Lines—Elements In-Phase or 180° Out-of-Phase

The feed method introduced here has been used in its simplest form to feed television receiving antennas and other arrays, as presented by Jasik, pages 2-12 and 24-10 or Johnson, on his page 2-14. However, until first presented in *The ARRL Antenna Book,* this feed method was not widely applied to amateur arrays.

The method takes advantage of an interesting property of λ/4 transmission lines. (All references to lengths of lines are electrical length and lines are assumed to have negligible loss.) See **Fig 15**. The magnitude of the *current out* of a λ/4 transmission line is equal to the *input voltage* divided by the characteristic impedance of the line. This is independent of the load impedance. In addition, the phase of the output current lags the phase of the input voltage by 90°, also independent of the load impedance. These properties can be used to advantage in feeding arrays with certain phase angles between elements.

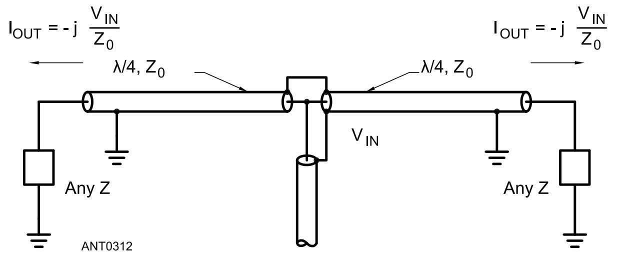
$I_{OUT} = -j\frac{V_{IN}}{Z_0}$
$I_{OUT} = -j\frac{V_{IN}}{Z_0}$
λ/4, Z_0
λ/4, Z_0
V_{IN}
Any Z
Any Z
ANT0312

Fig 15—A useful property of λ/4 transmission lines; see text. This property is utilized in the "current-forcing" method of feeding an array of coupled elements.

If any number of loads are connected to a common driving point through λ/4 lines of equal impedance, the currents in the loads will be *forced* to be equal and in-phase, regardless of the load impedances. So any number of in-phase elements can be correctly fed using this method, regardless of how their impedances might have been changed by mutual coupling. Arrays that require unequal currents can be fed through λ/4 lines of unequal impedances to achieve other current ratios.

The properties of λ/2 lines also are useful. Since the current out of a λ/2 line equals the input current shifted 180°, regardless of the load impedance, any number of half wavelengths of line may be added to the basic λ/4, and the current and phase forcing property will be preserved. For example, if one element is fed through a λ/4 line and another element is fed from the same point through a 3λ/4 line of the same characteristic impedance, the currents in the two elements will be forced to be equal in magnitude and 180° out-of-phase, regardless of the feed-point impedances of the elements.

If an array of two, and only two, identical elements is fed in-phase or 180° out-of-phase with equal magnitude currents, both elements have the same feed-point impedance. The reason is that each element sees exactly the same thing when looking at the other. In an in-phase array, each sees another element with an identical current; in an out-of-phase array, each sees another element with an equal magnitude current that's 180° out-of-phase, the same distance away in both cases. This isn't true in something like a 90° fed array, where one element sees another with a current leading its current by 90°, while the other sees another element with a lagging current.

With arrays fed in-phase or 180° out-of-phase, feeding the elements through equal lengths of feed line (in-phase) or lengths differing by 180° (out-of-phase) will lead to the correct current magnitude ratio and phase difference, regardless of the line length and regardless of how much the element feed-point impedances depart from the lines' Z_0.

Unless the feed-point impedances equal the line Z_0 or the lines are an integral number of half wavelengths long, the magnitudes of the currents out of the lines will not be equal to the input magnitudes, and the phase will not be shifted an amount equal to the electrical lengths of the lines. But both lines will produce the same transformation and phase shift because their load impedances are equal, resulting in a properly fed array. In practice, however, feed-point impedances of elements frequently are different even in these arrays, because of such things as different ground systems (for ground mounted vertical elements), proximity to buildings or other antennas, or different heights above ground (for horizontal or elevated vertical elements).

In many larger arrays, two or more elements must be fed either in-phase or out-of-phase with equal currents, but coupling to other elements can cause their impedances to change unequally—sometimes extremely so. Using the current-forcing method allows the feed system designer to ignore all these effects, while guaranteeing equal and correctly phased currents in any combination and number of 0° and 180° fed elements.

This method is used to develop feed systems for the Four Square and 4-element rectangular arrays in the **Practical Array Design** section. The front and rear elements of a Four-Square antenna provide a good example of elements having very different feed-point impedances that are forced to have equal out-of-phase currents.

"The Simplest Phased Array Feed System—That Works"

This is the title of an article in *The ARRL Antenna Compendium*, *Vol 2*, which describes how arrays can be fed with a feed system consisting of only transmission lines. (The article is available for viewing at **eznec.com/Amateur/Articles/Simpfeed.pdf** and is also on the CD included in the back of this book, along with the program *Arrayfeed1*, which solves the equations presented in the article.)

As explained earlier in the Phasing Line section, this method requires knowing what the element feed-point impedances will be in a correctly fed array. Feed-line lengths can then be computed, for most but not all arrays. These lengths will produce the desired current ratio in array elements that do present those feed-point impedances. If you know the load impedances connected to transmission lines whose inputs are connected to a common source, it's simple to calculate the resulting load currents for any transmission line lengths. However, the reverse problem is much more difficult; that is, given the load impedances and desired currents to calculate the required cable lengths.

One way to solve the problem is to choose some feedline lengths, solve for the currents, examine the answer, adjust the feed-line lengths, and try again until the desired currents are obtained. The author used this iterative approach, using first a programmable calculator and later a computer, for some time before developing a direct way of solving for the transmission-line lengths. The direct solution method is described briefly in the *Compendium* article.

Fig 16 shows the basic so-called "simplest" system applied to a two element array. Although it resembles an elementary phasing-line system as described earlier, the critical difference is that the lengths of Lines 1 and 2 are calculated to provide the correct current relative magnitude transformation and phase shift when terminated with the actual feed-point impedances.

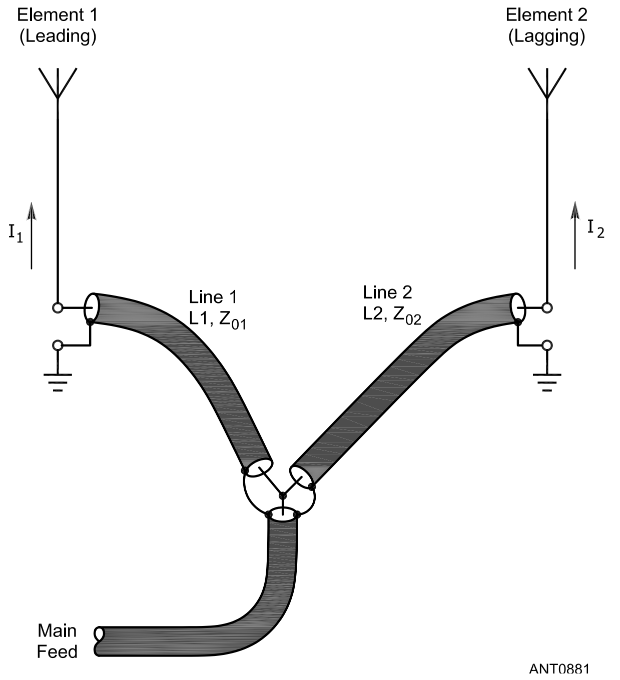

Fig 16—"Simplest" feed system for 2-element array. No matching or phasing network is used here, only transmission lines.

The advantage to using this "simplest" feed system is indeed its simplicity. It's no more complicated than the elementary phasing-line approach but actually works as planned. The disadvantage over some other methods is that there's no convenient adjustment to compensate for environment factors, array imperfections or inaccurately known feed-point impedances.

Also, while unusual, it's possible that no suitable feedline lengths can be found for some arrays, or at least none with practical feed-line characteristic impedances. The difference in electrical feed-line lengths almost never equals the difference in phase angles between element currents. This is because of the different line delays resulting from different feed-point impedances.

Program *Arrayfeed1*, included on the CD included with this book, can do the calculations for any two elements (alone or in a larger array), a Four-Square array or a rectangular array in which two in-phase elements are driven at any current magnitude and phase relative to the other two in-phase elements. These possibilities cover a large number of common arrays.

Arrayfeed1 can also be applied to other types of arrays

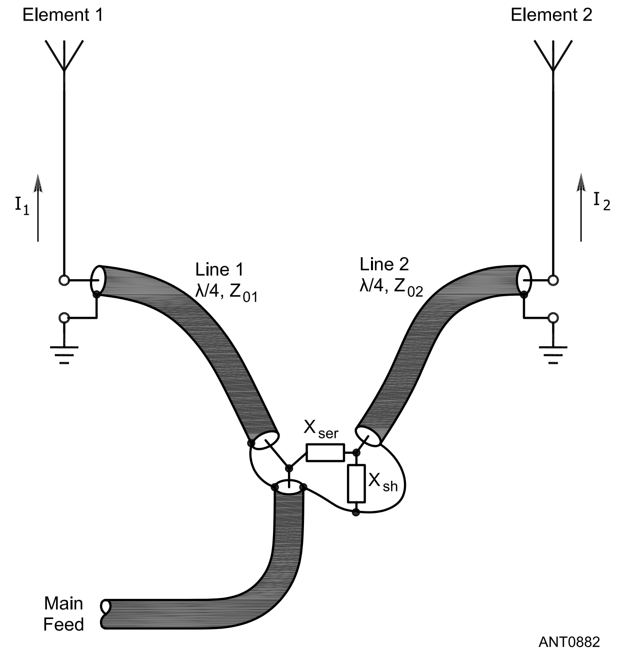

Fig 17— The addition of a simple L-network to Fig 16 allows you to easily adjust feeding of element pairs at other relative phase angles and/or magnitude ratios.

using the method described in the **Feeding Larger Arrays** section in Appendix B (on the CD). The required knowledge of element feed-point impedances in a correctly fed array can be obtained using *EZNEC-ARRL*, also included on the CD. Examples of the design of a "simplest" feed system for several different arrays using *EZNEC-ARRL* and *Arrayfeed1* can be found in the **Phased Array Design Examples** section.

When a solution is possible for a given choice of line characteristic impedances, a second solution with different lengths is always available. See the comments in the introductory part of the **Phased Arrray Design Examples** section about choosing the solution to use.

An Adjustable L-Network Feed System

Adjustment of the current ratio of any two elements requires varying two independent quantities; for example the magnitude and phase of the current ratio. Two degrees of freedom—adjustments that are at least partially independent—are required. The "simplest" all-transmission line feed system described earlier adjusts the lengths of the two transmission lines to achieve the correct ratio.

But if the antenna characteristics aren't well known—for example, if the ground resistance isn't known even approximately—then the initial "simplest" design won't be optimum and adjustment can be difficult and tedious. The current-forcing method produces correct currents independently of the element characteristics, so it doesn't require adjustment as long as the elements are identical. But it's suitable only for feeding elements in-phase or 180° out-of-phase and a few fixed current-magnitude ratios.

The addition of a simple network as shown in **Fig 17** allows you to easily adjust feeding of element pairs at other relative phase angles and/or magnitude ratios. Any desired current ratio (magnitude and phase) can be obtained with two elements fed with any lengths of wire, equal or unequal, by adding a network.

However, calculations for the general case are complex. The problem becomes much simpler if the transmission lines are restricted to lengths of odd multiples of λ/4, forming a modified "forcing" system that includes an added network. There are at least three additional advantages of this scheme. One is that a λ/4 line is easy to measure, even if the velocity factor isn't known. This is described in the **Practical Aspects of Phased Array Design** section.

Second is that the feed system becomes completely insensitive to the feed-point impedance of one of the two elements. And the third is that the transmission lines of "forcing" systems feeding groups of elements in larger arrays can be used in place of the normal λ/4 lines. This greatly simplifies both the design of feed systems for larger arrays and the feed systems themselves. Note that both lines can be changed to 3λ/4 if necessary to span the physical distance between elements, but *both* lines must be the same 3λ/4 length.

This basic feed method can be used for any pair of elements, or for two groups of elements having forced equal currents. (See **Feeding Four Element and Larger Arrays** below.) Many networks can accomplish the desired function, but a simple L network is adequate for most feed systems. The network can be designed to produce a phase lead or phase lag. The basic two-element L-network feed system is shown in Fig 17. Many variations of this general method can be used, but the equations, program, and method to be discussed here apply only to the feed system shown.

If the phase angle of I_2/I_1 is negative (element 2 is lagging element 1), the L network will usually resemble a low-pass network (X_{ser} is an inductor and X_{sh} is a capacitor). But if the phase angle is positive (element 2 lagging element 1), the L network will resemble a high-pass network (X_{ser} is a capacitor and X_{sh} is an inductor). However, some current ratios and feed-point impedances could result in both components being inductors or both being capacitors.

If it's desired to maintain symmetry in the feed system, X_{ser} can be divided into two components, each being inserted in series with a transmission line conductor. If X_{ser} is an inductor, the new components will each have half the value of the original X_{ser}, as shown in **Fig 18**. If X_{ser} is a capacitor, each of the new components will be twice the original value of X_{ser}.

Because of the current-forcing properties of λ/4 lines, we need to make the ratio of *voltages* at the inputs of the lines equal to the desired ratio of *currents* at the output ends of the lines; that is, at the element feed points. The job of the L network is to provide the desired voltage transformation. If the output-to-input voltage ratio of the network is, say, 2.0 at an angle of –60°, then the ratio of element currents (I_2/I_1)

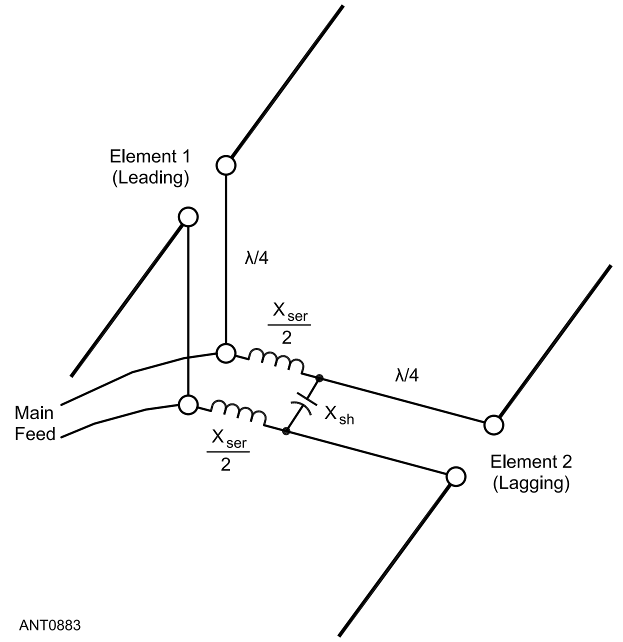

Fig 18—Symmetrical feed system similar to Fig 17, where feed network is split into two symmetrical parts.

will be 2.0 at an angle of –60°. The voltage transformation of the network is affected by the impedance of element 2, but not by the impedance of element 1. So only the impedance of element 2 must be known to design the feed system.

Equations for designing the L network are given in Appendix B, but the program *Arrayfeed1* is included on the CD to make it unnecessary to solve them. The feed-point impedance of the lagging element or group of elements must be known in order to design the network. This can best be determined by modeling the array with *EZNEC-ARRL*. The impedance can be manually calculated for some simple element and array types by using the equations in Appendix B, but those same types of element and array are simple to model.

Examples of the design of L-network feed systems for several different arrays using *EZNEC-ARRL* and *Arrayfeed1* can be found in the **Phased Array Design Examples** section. A similar application of this feed system and a spreadsheet program for calculation was developed by Robye Lahlum, W1MK, and described in *Low-Band DXing* (see the Bibliography). *Arrayfeed1* can be used for the applications of the feed system described in that book if desired.

Additional Considerations

Feeding 4-Element and Larger Arrays

Both the simplest and L-network feed systems described above can be extended to feeding larger arrays having two groups of elements in which all the elements in a group are in-phase or 180° out-of-phase with each other—basically, any group that can be fed with the current-forcing method. The elements in each group are connected to a common point with λ/4 or 3λ/4 lines to force the currents to be in the correct ratio within the group. Then the "simplest" or L-network feed system can be used to produce the correct phasing between the two groups, just as it does between two individual elements.

Two common arrays that fit this description are the Four Square and the 4-element rectangular array. But more elaborate arrays could be constructed and fed using this method, such as a pair of binomial arrays. (A single binomial array is described in the **Phased Array Design Example** section below.) The *Arrayfeed1* program incorporates additional calculations necessary for designing Four Square and 4-element rectangular arrays. The general procedure for adapting the feed methods to other larger arrays can be found in Appendix B.

What If the Elements Aren't Identical?

Getting the desired pattern requires getting the correct relative magnitude and phase of the *fields* from the elements. If the elements are identical, which we've generally assumed up to this point, then producing currents of the desired magnitude and phase will create the desired fields (neglecting mutual coupling current distribution effects, discussed elsewhere).

But what if the elements aren't identical? Fortunately, the feed systems described here can still be used for any 2-element array and some more complex arrays, provided that the system can be accurately modeled. But a slightly different approach is required than for identical elements.

The first step is to model the array with a current source at the feed point of each element. Next, the magnitudes and phases of the model source currents are varied until the desired pattern is achieved. Then the ratio of feed-point (source) currents is calculated and this value, along with the feed-point impedances reported by the model, are used for the feed system design. The feed system will produce the same ratio of currents as the model, resulting in the same pattern.

In general, this approach won't work with shunt fed towers or gamma-fed elements because of the difficulty of accurately modeling those structures. See **Shunt and Gamma Fed Towers and Elements** for more information.

Shunt and Gamma Fed Towers and Elements

In a shunt, gamma, or similarly fed tower or element, the feed-point current isn't the same as the main current flowing in the element. The ratio between the feed-point current and element current isn't a constant, but depends on a number of factors. The ratio of currents in shunt or gamma fed elements are typically different—often vastly different—from the currents at the feed points. This complicates the design of feed systems for arrays of these elements.

An even more limiting problem is that the feed-point impedances are difficult to determine. The feed-point impedances of one or more elements in a properly fed array must be known in order to design a feed system for anything but

2-element in-phase or 180° out-of-phase arrays.

The only practical way to get this information for a shunt or gamma fed array is by modeling an array having the desired element currents. But Cebik has pointed out ("Two Limitations of NEC-4"—see the Bibliography) that many common antenna analysis programs, including *EZNEC-ARRL*, have difficulty accurately modeling folded dipoles with unequal diameter wires. The same problem applies to shunt and gamma-fed elements when the element diameter is significantly different from the diameter of the shunt or gamma feed wire. Without accurate feed-point impedances, feed systems can't be designed to work without adjustment. It might be possible to get reasonably accurate results from a *MININEC*-based modeling program, but there are a number of issues which must be given great care when using one. (See Lewallen, "*MININEC*—The Other Edge of the Sword," listed in the Bibliography.)

If such a *MININEC* program is available, you would have to model the complete array including feed system, with sources at the normal feed points in the shunt or gamma wires. Next, you would have to adjust the magnitudes and phases of the sources to produce the desired pattern. The reported source impedances and currents would be the ones you would use to design the feed system. It's likely that some adjustment would be necessary, so an adjustable system such as the L-network feed system described later would be best.

Loading, Matching and Other Networks

Adding a component such as a loading inductor in series with an element or element feed point won't change the ratio of element current to feed-point current. As a result, a feed system designed to produce a particular ratio of element currents will still function properly if the elements contain series components. The extra feed-point impedance introduced by the loading component(s) must be considered when designing the feed system, however. Similarly, end or top loading won't alter the relationship between feed-point and element current, provided that the current distribution in the elements is essentially the same. (See **Feed-point vs Element Current** in a previous section.)

However, insertion of any *shunt* component, or a network containing a shunt component, *will* alter the relationship between feed-point and element current because it will divert part of the feed-line current that would otherwise flow into the antenna. As a result, a feed system designed to deliver correct currents at the feed points will produce incorrect element currents and therefore an incorrect pattern. Therefore, any components or networks other than a series loading component should be avoided at any place in the feed system on the antenna side of the point at which the feed system splits to go to the various elements.

There are a few exceptions to this rule. If the feed-point impedances of the elements when in the excited array are equal, then identical networks with or without shunt components can be put at the feed points of the elements and the proper element current ratio maintained—so long as the feed system is designed to deliver the proper feed-point current ratio with the networks in place. Equal element impedances occur in arrays having only two identical elements fed in-phase or 180º out-of-phase, or arrays of any number of elements where the elements are electrically short and/or very lossy.

Baluns in Phased Arrays

For purposes of achieving the correct array pattern, baluns aren't usually required when feeding grounded vertical elements with coaxial cable feed lines. However, a balun might be desirable if current induced onto the outside of the feed line by mutual coupling to the elements is causing RF in the shack. And with arrays of dipole or other elevated elements, baluns can be important to achieve the proper element current ratio, as explained below.

First, however, the general rules for using baluns in phased arrays will be stated. Here, "main feed line" means the feedline going from the transmitter or receiver to the common point where the system splits to feed the various elements. "Phasing-system lines" means any transmission lines between that common point and any element. The rules:

- **Rule 1:** A balun or baluns (more specifically, a current, sometimes called a choke balun) should be used as necessary to suppress unbalanced current on the main feed line. This usually isn't required when feeding grounded elements with coaxial feed line from an unbalanced rig or tuner. Unbalanced current can occur on either coax or parallel-conductor line.

"Baluns: What They Do and How They Do It", listed in the Bibliography, describes conducted-imbalance (common-mode) currents. Imbalance can also be caused by mutual coupling to the array elements. Common-mode currents have at least two undesirable effects on array performance. First, the imbalance current can flow from the main feed line to the phasing system lines, not necessarily splitting in the right proportion to maintain the correct element current ratio. This can affect the array pattern. In practice, however, this effect is likely to be small unless the common-mode current is unusually large. Even a small common-mode current, however, results in main feed-line radiation, and even a small amount of radiation can significantly degrade array pattern nulls. Any type of current balun can be used on the main feed line, at any place along the line, without any effect on the array pattern except to the extent that it reduces common mode current.

- **Rule 2:** No balun or any other component or network should be inserted in any phasing system line that will alter the line length or characteristic impedance. This means that baluns in phasing system lines must be of a type made from the phasing line itself. Options are the W2DU type balun, consisting of ferrite cores placed along the outside of the feed line; an air-core balun made by winding part of the line into an approximately self-resonant or otherwise high-impedance coil; or winding part of the line onto a ferrite core or rod to make a several-turn winding. When coaxial cable is used, the feed system characteristics are dictated by the inside of the cable. Any cores or winding of the outside prevents common-mode current on the outside, but otherwise have no effect on the phasing performance. This rule applies equally to parallel wire line, where the balun affects only common-mode current (equivalent to current on the outside of coax) while the phasing performance depends on differential mode current (equivalent to the current on the inside of coax).

Baluns are important when feeding dipole or other elevated arrays, unless a fully balanced tuner is used. This is because common-mode current represents a diversion of some of the current that should be going to the array elements. The presence of common-mode current means that the element currents are being altered from the desired ratio and therefore the pattern won't as intended. A balun should be placed wherever a path for current exists other than along a parallel-line conductor or on the inside of a coaxial line. Such a path exists, for example, where a coaxial cable connects to a dipole, as shown in Fig 1 of the balun article referenced above. Or a path can exist where a parallel-conductor transmission line connects to an unbalanced tuner or to a coaxial line, as shown in Fig 2 of that article. In both those cases, a path exists for a common-mode current to flow on the outside of the coax cable. A balun creates a high impedance to this current, thereby reducing its magnitude. But remember that all baluns must conform to the rules above.

Fig19 shows recommended balun locations for a coax-fed dipole array using an L-network feed system.

Receiving Arrays

While it might not be entirely intuitive, an array designed for a particular gain and pattern for transmitting that considers mutual coupling, element currents, field reinforcement and cancellation, and so forth, will perform exactly the same when receiving. So a receiving array can be designed by approaching the problem as though the array were to be used for transmitting.

However, at HF and below, the system requirements for transmitting and receiving antennas are different, so receiving-only arrays can be designed that aren't suitable for transmitting but are perfectly adequate for receiving in that frequency range. The reason, described in more detail in Chapter 13 of this book, is that at HF and below atmospheric noise is typically much greater than a receiver's internally generated noise. Lowering a receiving antenna's gain and efficiency reduces the signal and atmospheric noise both by the same factor. Because the overall noise is for practical purposes all atmospheric noise, the signal/noise ratio isn't affected by antenna efficiency.

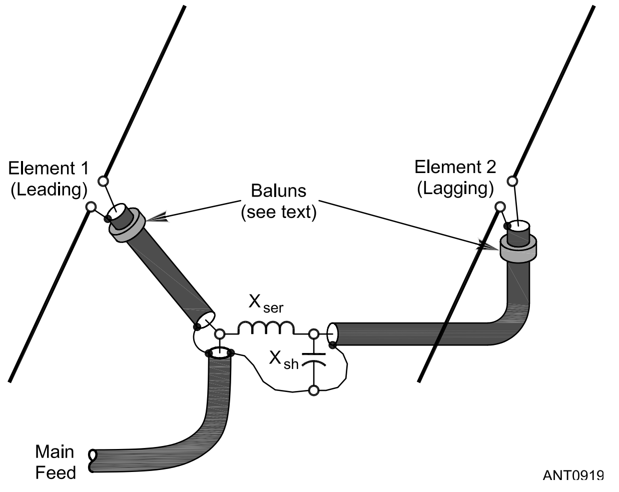

Fig 19—Adding choke baluns to a two-dipole feed system to get rid of common-mode currents radiated onto the coax shields.

Of course, a point can be reached where the atmospheric noise is so reduced by inefficiency that the receiver itself becomes the dominant source of noise, but this typically doesn't happen until the antenna is extremely inefficient. When transmitting, reduced efficiency lowers the transmitted signal, but it has no effect on the receiving station's noise. So reduced efficiency of a transmitting antenna results in a reduced signal/noise ratio at the receiving end, and consequently should be avoided.

Mutual coupling effects can be minimized by increasing the loss (and therefore reducing the efficiency) of the elements, or by reducing the element sizes to a small fraction of a wavelength. Doing the second without the first isn't usually a good idea because the feed-point impedance tends to change rapidly with frequency for very small elements, making an antenna that works well over only a narrow bandwidth. But increasing loss broadens the bandwidth, even for small elements, as well as reducing mutual coupling effects. So this approach is often taken for designing a receiving-only array. With mutual coupling effects minimized because of loss, feed-system design becomes relatively simple, provided a few simple rules are followed. See **Loss Resistance, Mutual Coupling, and Antenna Gain**, above.

Phased Array Design Examples

This section, also written by Roy Lewallen, W7EL, presents examples of feed-system design for several kinds of array using the design principles given in previous sections. All but the last example array are assumed to be made of λ/4 vertical elements. The last example is for a halfwave-dipole array, which illustrates that exactly the same method can be used for arrays of any shape of elements, including dipole, square (quad) and triangular. Likewise, the methods shown here apply equally well to VHF and UHF arrays. The first example includes more detail than the remaining ones, so you should read it before the others.

General Array Design Considerations

If either the "simplest" feed system (Fig 16) or L-network feed system (Fig 17) is used, the feed-point impedance of one or more elements—*when the array elements all have the correct currents*—must be known. By far the best way to determine this is by modeling. If accurate modeling isn't practical for some reason, an estimation should be made from an approximate model, and you should expect to have to adjust the feed system after building and installing it.

Manual calculation methods for some simple configurations are given in Appendix B (on the CD), but calculation is tedious and, as stated earlier, the configurations for which this method works are the very ones which are easiest to model. *EZNEC-ARRL* (also included on the CD) is used in the following examples to determine feed-point impedance. Space doesn't permit detailed instructions here on creating the models, so they are included in complete form. They should provide a convenient starting point for any variations you might like to try. See the *EZNEC-ARRL* manual (accessed by clicking **Help/Contents** in the main *EZNEC-ARRL* window) for help in using this program.

In the following examples, vertical elements are close to λ/4 high and dipole elements close to λ/2, and their lengths have been adjusted for resonance when all other elements are absent or open circuited. There's actually no need in practice to make the elements self-resonant—it's simply used as a convenient reference point for these examples. You'll also find it interesting to see how much reactance is present at the feed points of the elements when in the arrays, knowing that it's very nearly zero when only one element is present.

In any real grounded vertical array, there is ground loss associated with each element. The amount of loss depends on the length and number of ground radials, and on the type and wetness of the ground under and around the antenna. This resistance becomes part of the feed-point resistance, so it must be included in the model used to determine feed-point impedance. The 90° Fed, 90° Spaced Array example below discusses how this is done. **Fig 20** gives resistance values for typical ground systems, based on measurements by Sevick (July 1971 and March 1973 *QST*). The values of feed system components based on Fig 20 will be reasonably close to correct, even if the ground characteristics are somewhat different than Sevick's.

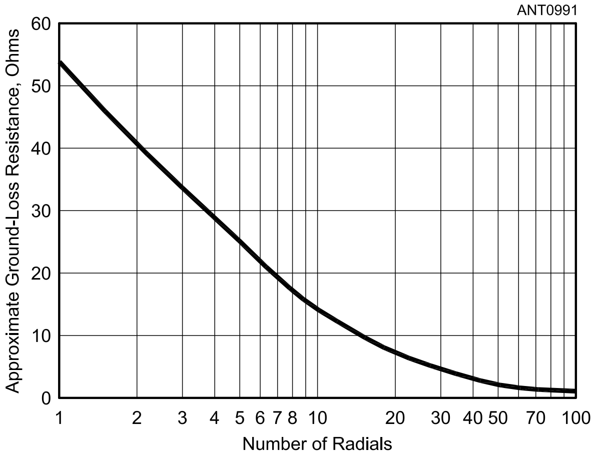

Fig 20—Approximate ground system loss resistance of a resonant λ/4 ground-mounted vertical element versus the number of radials, based on measurements by Jerry Sevick, W2FMI. Moderate length radials (0.2 to 0.4 λ) were used for the measurements. The exact resistance, especially for only a few radials, will depend on the nature of the soil under the antenna. Add 36 Ω for the approximate feed-point resistance of a thin resonant λ/4 vertical.

Feed systems for the design example arrays to follow are based on the resistance values given below.

Number of Radials	*Loss Resistance, Ω*
4	29
8	18
16	9
Infinite	0

Elevated radial systems also have some ground loss, although it can be considerably less than a system with the same number of buried radials. This loss will be automatically included in the feed-point impedance of a model which includes the elevated radials, so no further estimation is required. Be sure to use Perfect, High-Accuracy ground type when modeling an elevated radial system with *EZNEC-ARRL*. In other *NEC-2* based programs, this might be referred to as Sommerfeld type ground. More information can be found in the *EZNEC-ARRL* manual.

The matter of matching the array for the best SWR on the feed line to the station is not dealt with here, since it's a separate problem from that of the main topic, which is designing feed systems to produce a desired pattern. Some of the simpler arrays provide a match that is close to 50 or 75 Ω, so no further matching is required. However, as shown by program *Arrayfeed1*, many larger arrays present a less favorable impedance for direct connection and will require matching if a low SWR on the main feedline is required. If matching is necessary, the appropriate network should be placed in the single feed line running to the station. Attempts to improve the match by adjustment of the phasing L network, individual element lengths, matching at the element feed points or individual element feeder lengths will usually ruin the current balance of the array. Program *TLW*, included on the CD, can be used for designing an appropriate matching network. Additional information on impedance matching may be found in Chapters 25 and 26 of this book.

Choosing Arrayfeed1 Solutions

When designing a feed system for a two element array, *Arrayfeed1* program allows you to choose the characteristic impedances of the two transmission lines going to the elements, which don't have to be equal, so you have your choice of more than one solution. However, directional array switching is much more difficult if the lines have different impedances, so in general you should use the same characteristic impedances.

For larger arrays, *Arrayfeed1* requires the feed lines to all elements to have the same impedance. In choosing the transmission line impedance values, usually you can simply use convenient impedances. But in general, you should avoid solutions where component reactance (X) values are vastly different (say, more than three times or less than one third as large) as the line characteristic impedances. Such networks will become more critical to adjust, and both the impedance and pattern will change more rapidly with changes in frequency. You can usually avoid this situation by choosing feed-line impedances that are in the same ballpark as the element feed-point impedances. The last example in the Practical Array Design section illustrates this problem and its solution.

When designing a "simplest" feed system, the most broadbanded and least critical system is usually one where the difference in electrical feed-line lengths is closest to the relative element phase angle. Here, "broadbanded" means that the pattern changes less with frequency, not necessarily that the SWR changes less. However, an array that's broadbanded in the pattern sense is usually also relatively broadbanded with respect to SWR.

Arrayfeed1 reports the impedance seen at the main array feed point. While it might be tempting to choose the solution producing the lowest SWR on the main feedline, you'll end up with a less critical and more broadbanded system if you base your choice on the criteria given above, and provide separate impedance matching at the array's main feed point when necessary.

90° Fed, 90° Spaced Vertical Array

This example illustrates the design of both "simplest" and L-network feed systems for a 2-element, 90° spaced and fed vertical array. The first task when using either feed system is to determine the feed-point impedances of the elements when placed in an array having the desired element currents. The "simplest" feed system method requires knowledge of both element impedances, while the L-network system requires you to know only one. Actually, it's equally easy to determine both as it is to find just one, using *EZNEC-ARRL*. (Appendix B contains equations for those interested in manual methods or for more insight as to how the impedances come about.) The first step is to specify the antenna we want. For this example, we'll specify:

- Frequency: 7.15 MHz.
- Two identical, one inch (2.54 cm) diameter, 33 feet (10.06 meter) long elements spaced 90 electrical degrees, with element currents equal in magnitude and 90° apart in-phase.
- 8 buried radial wires, 0.3 λ long, under each element.

A model of this antenna has been created and furnished with *EZNEC-ARRL*. So the next step is to start *EZNEC-ARRL*, click the **Open** button, enter **ARRL_Cardioid_Example** in the text box (or double-click it in the file list) to open example file **ARRL_Cardioid_Example.EZ**.

This *EZNEC* example model uses a *MININEC*-type ground, which is the same as perfect ground when calculating antenna currents and impedances. A real antenna would have some additional resistive loss due to the finite conductivity of the ground system. The only way to model a buried radial ground system with an *NEC-2* based program like *EZNEC* is to create radial wires just above the ground (using the Real, High-Accuracy ground type), because *NEC-2* can't handle buried conductors.

This provides only a moderate approximation of a buried system. Another way to estimate ground-system resistance is to measure the feed-point impedance of a single element, then subtract from that the resistance reported for a model of that element over perfect (or *MININEC*-type) ground. For most uses, however, an adequate approximation can be made by simply referring to the graph of Fig 20. As stated previously, the feed system design depends on the feed-point impedances of the elements, which in turn depend on the ground system resistance. So the ground system resistance must be known, approximately anyway, before designing the feed system. At the end of this example we'll investigate the effect of changes in the ground system or errors in estimating the resistance on the pattern.

For 8 radials, Fig 20 shows the ground system resistance to be about 18 Ω. This is included in the example model as a simple resistive load at the feed point of each element. Click the **Src Dat** button to see the feed-point impedances of the two elements. In this model, Source 1 is at the base of Wire 1 (element 1), and Source 2 is at the base of Wire 2 (element 2). Notice in the **Source Data** display that the Source 1 current has been specified at 1 amp at 0°, and Source 2 is 1 amp at –90°. So the Source 2 element is the lagging element. You should see impedances of 37.53 – *j*19.1 Ω for element 1 and 68.97 + *j*18.5 Ω for element 2. These are the feed-point impedances resulting when the array is ideally fed, with equal magnitude and 90° phased currents. Record these values for use in *Arrayfeed1*.

Click the **FF Plot** button to generate a plot of the azimuth pattern at an elevation angle of 10°. In the 2D Plot Window, open the **File** menu and select **Save Trace As**. Enter **Cardioid_Ideal Feed** in the **File Name** box, then click **Save**. This saves the cardioid pattern plot so you can compare it later to the pattern you get with the transmission line feed system.

Now it's time to design the feed system. Refer to the appropriate subheading below for the design of each of the two kinds of feed systems. Both systems use program *Arrayfeed1* program.

"Simplest" (Transmission Line Only) Feed System

Start *Arrayfeed1*. In the **Array Type** frame, select **Two Element**. In Feed System Type, select **"Simplest."** In the **Inputs** frame, enter the following values:

Frequency MHz = 7.15; Feed-point impedances – Leading Element: R ohms = 37.53, X ohms = –19.1; Lagging Element: R ohms = 68.97, X ohms = 18.5 (these are the element R and X values from *EZNEC-ARRL*). We'll be discussing the array input impedance, so check the **Calc Zin** box near the lower left corner of the main window if it's not already checked.

We're free to choose any transmission-line characteristic impedances we want, so long as we can get cables with those impedances. And the two cables don't have to have the same characteristic impedances. Each choice will lead to a different set of solutions. But sometimes a solution isn't possible, which then requires choosing different line impedances. Let's try 50 Ω for both lines. Enter **50** in both **Z0** boxes.

Finally, enter **1** for the lagging:leading **I Mag**, and **–90** for the **Phase**. Click **Find Solutions**. The result is no solution! So enter **75** into both the line impedance boxes and click **Find Solutions** again. You should now see two sets of results in the **Solutions** frame, electrical lengths of 68.80° and 156.03° for the first solution and 131.69° and 185.00° for the second. (Notice that the difference in length between the two lines isn't 90° for either solution, although the first solution is quite close. It's normal for the feed-line length difference to be different than the phase difference, due to the unequal element feed-point impedances caused by mutual coupling.)

The solution with a line length difference closest to the element phase difference is usually preferable. Also, all else being equal, the solution with shortest lines is better providing that the lines will physically reach the elements. This is because the current magnitude and phase will change less with frequency than for a longer-length solution. However, there might be some cases where the change with frequency luckily compensates for the changing electrical distance between elements, so it's not a bad idea to model both solutions unless you plan on using the antenna over only a narrow frequency range.

In this case, the first solution looks best in all respects. The sum of the two lines in the first solution is about 225 electrical degrees. Assuming the lines have a velocity factor of 0.66, the total length of the lines will be more than 148 physical degrees. Since our two elements are spaced 90 physical degrees apart, the lines will comfortably reach. If they didn't, we could either use the second solution's lengths, use cable with a higher velocity factor or add a half wavelength to both the line lengths in the first solution.

The impedance **Zin** shown by *Arrayfeed1* is the impedance at the input to the feed system, so it's the impedance that will be seen by the main feed line. The second solution provides nearly a perfect match for a 50-Ω transmission line. But the first solution is good for nearly all applications. Also a 50-Ω line connected to the first solution's feed system would have an SWR of only 1.65:1, which wouldn't require any matching under most circumstances. Normal line loss would reduce the SWR even more at the transmitter end of the feed coax.

To find the required physical line lengths, enter the cable velocity factor and make your choice of units in the **Physical Lengths** frame. The design is now complete; all you have to do is cut two lines to the specified lengths and connect one from a common feed point to each element as shown in Fig 16, or the screen capture from *Arrayfeed1* shown in **Fig 21**.

Next, we'll design an L-network feed system for the same array.

L Network Feed System

In *Arrayfeed1*, select **L Network** in the **Feed System Type** frame. The program doesn't need to know the leading element impedance to calculate the L-network values, but it does need it to calculate the array input impedance. If you want to know the impedance, check the **Zin** box at the lower left corner of the main window, otherwise you can uncheck it and the input box for the leading element Z will disappear. The values from the "Simplest" analysis should still be present in the appropriate boxes; if not, refer to the "Simplest" feed system design above and re-enter the values. Again, we'll use 75 Ω for the line impedances, since that gave us a

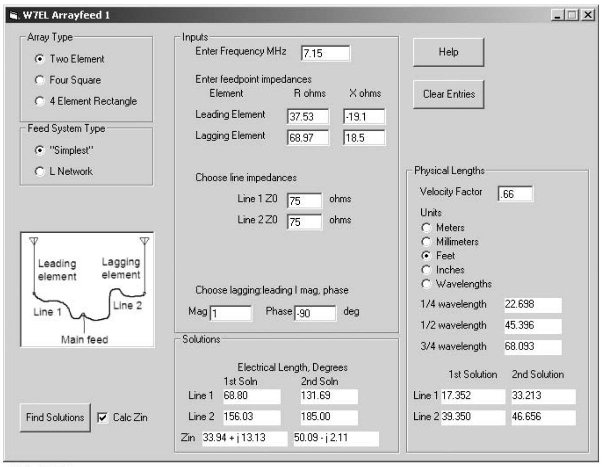

Array Type: ● Two Element ○ Four Square ○ 4 Element Rectangle

Feed System Type: ● "Simplest" ○ L Network

Inputs

Enter Frequency MHz 7.15

Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	37.53	-19.1
Lagging Element	68.97	18.5

Choose line impedances

Line 1 Z0 75 ohms

Line 2 Z0 75 ohms

Choose lagging:leading I mag, phase

Mag 1 Phase -90 deg

Solutions

Electrical Length, Degrees

	1st Soln	2nd Soln
Line 1	68.80	131.69
Line 2	156.03	185.00
Zin	33.94 + j 13.13	50.09 - j 2.11

☑ Calc Zin

Physical Lengths

Velocity Factor .66

Units: ○ Meters ○ Millimeters ● Feet ○ Inches ○ Wavelengths

1/4 wavelength 22.698

1/2 wavelength 45.396

3/4 wavelength 68.093

	1st Solution	2nd Solution
Line 1	17.352	33.213
Line 2	39.350	46.656

ANT0992

Fig 21—Screen capture from *Arrayfeed1* program for "Simplest" 2-element phased array shown in Fig 16 and whose feed-point impedances are modeled by *EZNEC-ARRL*.

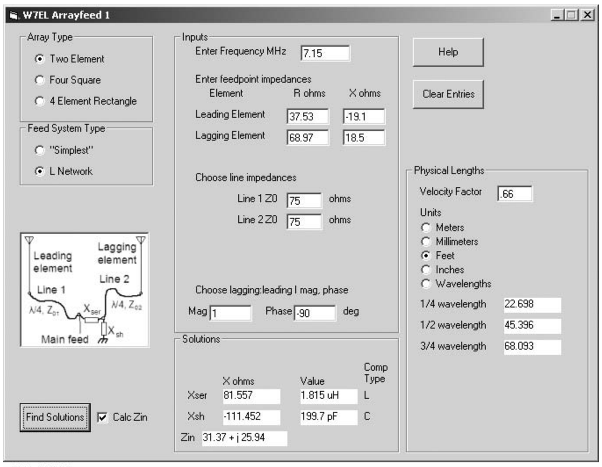

W7EL Arrayfeed 1

Array Type
● Two Element
○ Four Square
○ 4 Element Rectangle

Feed System Type
○ "Simplest"
● L Network

Inputs
Enter Frequency MHz 7.15
Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	37.53	-19.1
Lagging Element	68.97	18.5

Choose line impedances
Line 1 Z0 75 ohms
Line 2 Z0 75 ohms

Choose lagging:leading I mag, phase
Mag 1 Phase -90 deg

Solutions

	X ohms	Value	Comp Type
Xser	81.557	1.815 uH	L
Xsh	-111.452	199.7 pF	C

Zin 31.37 + j 25.94

☑ Calc Zin

Physical Lengths
Velocity Factor .66
Units
○ Meters
○ Millimeters
● Feet
○ Inches
○ Wavelengths
1/4 wavelength 22.698
1/2 wavelength 45.396
3/4 wavelength 68.093

ANT0993

Fig 22—Screen capture from *Arrayfeed1* program for L-network feed system using "current-forcing" properties of λ/4 feed lines.

solution for the "Simplest" feed system. This feed system is more versatile, though, so we could use 50-Ω lines with this feed system if desired.

Click **Find Solution** and see the results in the **Solution** frame. See screen capture in **Fig 22**. With 75-Ω lines, the L network consists of a series inductor of 1.815 µH and a shunt capacitor of 199.7 pF, connected as shown in the diagram in the left part of the program window. To find the physical length of the λ/4 lines, enter the velocity factor and choice of units in the **Physical Lengths** frame.

The main feed-point impedance of 31.37 + *j*25.94 Ω would result in about a 2.2:1 SWR on a 50-Ω feed line, which would be acceptable for many applications. It could easily be reduced to 1.6:1 by the simple addition of a series capacitor of 25.94 Ω reactance (858 pF) at the main feed point or, of course, reduced to 1:1 with a simple L network or other matching system designed with the *TLW* program.

Pattern Verification and Effect of Loss Resistance—"Simplest" System

EZNEC-ARRL doesn't have the capability to model an L-network, so *EZNEC-ARRL* verification of the pattern and the effect of various modifications can be done only for the "simplest" feed system.

EZNEC model **ARRL_Cardioid_TL_Example.EZ** has been created to model the "simplest" feed system just designed. Open it with *EZNEC-ARRL*. In the View Antenna Display, you can see the transmission lines connecting to the source midway between the antennas. In *EZNEC*, the physical locations of the ends of transmission line models don't have to be the same as the physical locations, so the view isn't a precise representation of what the actual setup would look like. (You can find more about this in **ARRL_Cardioid_TL_Example.txt**, the Antenna Notes file that accompanies example file **ARRL_Cardioid_TL_Example.EZ.**)

Click **FF Plot** to generate a 2D pattern of the antenna. In the 2D Plot Window, open the **File** menu and select **Add Trace**. Select **Cardioid—Ideal Feed** (which you saved earlier) and click **Open**. The added plot overlays perfectly, indicating that the pattern using this feed system is identical to the pattern we got with perfect current sources at each feed point.

To check the feed-point currents, click the **Currents** button. In the resulting table, you can see that **Wire 1 Segment 1** current is 0.56467 A at a phase of –56.73° and **Wire 2 Segment 1** current is 0.56467 A at –146.7°. (If you get the correct phase angles but wrong magnitudes, open the main window **Options** menu, select **Power level**, and make sure the **Absolute V, I sources** box is checked.) The ratio is 1.0000 at an angle of –89.97°, which is within normal error bounds for the desired 1 at –90°.

As a check on *Arrayfeed1*, click the **Src Dat** button to find the impedance seen by the source. This would be the impedance at the main feedline connection in the real array. *EZNEC-ARRL* reports 33.96 + *j*13.11 Ω, very close to the 33.94 + *j*13.13 Ω given by *Arrayfeed1* in Fig 21. Small differences of this order are normal and to be expected. This provides a further check that the *EZNEC-ARRL* model is correctly analyzing the *Arrayfeed1* feed system.

This *EZNEC-ARRL* model uses lossless transmission lines of a fixed physical length rather than a fixed electrical length (number of degrees), so they'll behave like real lines as the frequency is changed. By changing the *EZNEC* frequency and re-running the 2D plot, you can see that the front-to-back ratio degrades at 7.0 and 7.3 MHz. A slight adjustment of one or more line lengths, or a new *Arrayfeed1* solution at a slightly different frequency might produce a better compromise for some uses.

Other things you can try are to evaluate the second *Arrayfeed1* solution, or to try using different line impedances. (Keep the two line impedances equal if you anticipate doing array direction switching.) The effect of varying ground system resistance can also be evaluated by clicking the **Loads** line in the main window and changing the load resistance values. For example, if the ground system resistance were 9 Ω instead of the 18 Ω we have assumed, the front/back ratio would drop from about 32 to about 20 dB. Note that changing the *EZNEC* ground conductivity in this model has no effect on the feed-point current ratio. With a *MININEC* type ground, it's used only for pattern calculation—the ground is assumed perfect during impedance and current calculations, and the only ground system loss resistance in the model is what we've specifically put in as loads.

Not surprisingly, the forward gain is affected very little by changes in frequency or ground system loss. To find the gain relative to a single element, compare the reported dBi gain of **ARRL_Cardioid_Example** with the same model with one of the elements deleted. You'll find it's very close to 3.0 dB. The 90° fed, 90° spaced array is a special case of array where the effects of mutual coupling on the two elements are opposite and cancel, resulting in the same gain as if mutual coupling didn't exist. But mutual coupling most certainly does exist!

The second solution presented a more favorable main feed-point impedance, so it would be tempting to use that one instead of the first solution. Replacing the feed-line lengths with the second solution lengths to model the second solution shows that the front/back ratio deteriorates more at the band edges when the second solution is used. This might be tolerable if restricted frequency use is anticipated. But it does illustrate that the solution with shorter lines is generally more broadbanded and that the choice of solution shouldn't in general be based on the one giving the most favorable impedance.

A Three-Element Binomial Broadside Array

An array of three in-line elements spaced λ/2 apart and fed in-phase gives a pattern that is generally bidirectional. If the element currents are equal, the resulting pattern has a forward gain of 5.7 dB (for lossless elements) compared to a single element, but it has substantial side lobes. If the currents are tapered in a binomial coefficient 1:2:1 ratio (twice the current in the center element as in the two end elements), the gain drops slightly to just under 5.3 dB, the main lobes widen and the side lobes disappear.

The array is shown in **Fig 23**, and an *EZNEC-ARRL* model of the antenna over perfect ground to show the ideal pattern is provided as **ARRL_Binomial_Example.EZ**. To obtain a 1:2:1 current ratio in the elements, each end element is fed through a 3λ/4 line of impedance Z_0. Line lengths of 3λ/4 are chosen because λ/4 lines will not physically reach. The center element is fed from the same point through two parallel 3λ/4 lines of the same characteristic impedance. This is equivalent to feeding it through a line of impedance $Z_0/2$. The currents are thus forced to be in-phase and to have the correct ratio. **ARRL_Binomial_TL_Example.EZ** is an *EZNEC-ARRL* model that shows this feed system with lossless transmission lines. The reader is encouraged to experiment with this model to see the effect of changes in frequency, the addition of loss resistance (as resistive loads at the element feed points) and other alterations on the array pattern and gain. You should also replace the perfect ground with *MININEC* type of ground to show how radiation patterns over real ground differ from the theoretical perfect-ground pattern.

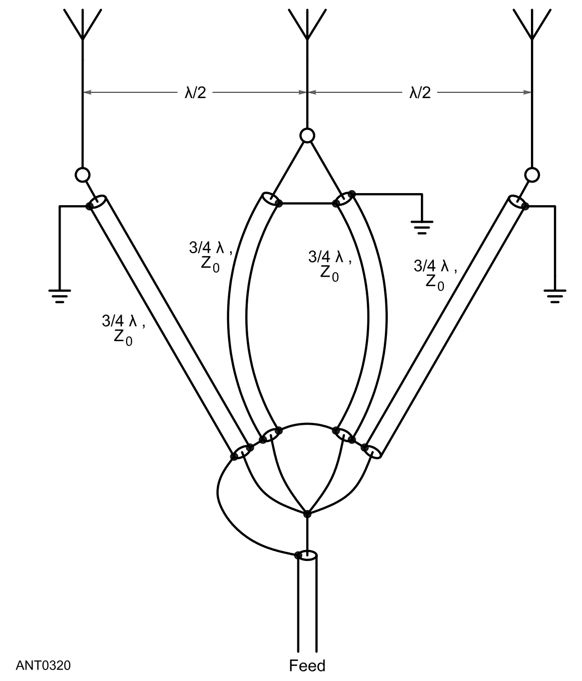

Fig 23—Feed system for the three element 1:2:1 binomial array. All feed lines are ¾ electrical wavelength long and have the same characteristic impedance.

A "Four Square" Array

Several types of feed system are used for feeding this popular array, and most share a common problem—they don't provide the correct element current ratio—although a number of them produce a workable approximation. The feed systems described here are capable of producing exactly the correct current ratio. The only significant variable is the element feed-point impedances, so the quality of the result depends on your ability to model the feed-point impedances of a correctly fed array. As in the examples above, *EZNEC-ARRL* will be used for that purpose and *Arrayfeed1* for the design of the feed system itself.

In this array (see **Fig 24**), four elements are placed in a square with λ/4 sides. (A variation of the Four Square uses wider spacing.) The rear and front elements (1 and 4) are 180° out-of-phase with each other. The side elements (2 and 3) are in phase with each other and 90° delayed from the front ele-

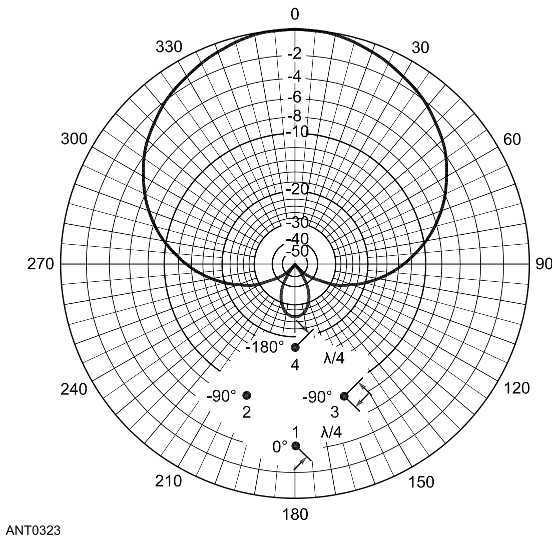

Fig 24—Pattern and layout of the four-element Four-Square array. Gain is referenced to a single similar element; add 5.5 dB to the scale values shown.

ment. The magnitudes of the currents in all four elements are equal. The front and rear elements can be forced to be 180° out-of-phase and to have equal currents by using the current-forcing method described earlier. One element is connected to a line that is either λ/4 or 3λ/4 long, the other to a line that is λ/2 longer, and the two lines to a common point.

Likewise, the two side element currents are forced to be equal by connecting them to a common point via λ/4 or 3λ/4 lines. **Fig 25** shows the basic current-forcing system.

If the pattern is to be electrically rotated, it is necessary to bring lines from all four elements to a common location. If solid-polyethylene dielectric coaxial cable, which has a dielectric constant of 0.66, is used, λ/4 lines won't reach the center of the array. So 3λ/4 lines must be used. Alternatively, you can use λ/4 lines with foam or other dielectric having a velocity factor of more than about 0.71 (plus a little extra margin). These will reach to the center. Whatever your choice, three of the lines must be the same length and the fourth must be λ/2 longer.

In this array, the side elements (2 and 3) have equal impedances, but the rear and front (1 and 4) are different from each other, and both are different from the side elements. We have to know the feed-point impedances of the front, rear and side elements in order to design the "simplest" feed system, but only the side element impedances are needed to design the L-network system. Knowledge of all feed-point impedances is necessary if the array main feedpoint impedance **Zin** is to be calculated. *EZNEC-ARRL* model **4Square_Example.EZ** shows a 40-meter Four Square array with 18 Ω of loss resistance at each element, to approximate an 8-radial per element ground system. (See the cardioid array example above for more information about modeling ground system loss.) Opening the file in *EZNEC-ARRL* and clicking the **Src Dat** button gives the following impedances:

Source 1: 16.4 – *j*15.85 Ω
Sources 2 and 3: 57.47 – *j*19.44 Ω
Source 4: 77.81 + *j*54.8 Ω

It's interesting to note that the resistive part of source 1 is less than the 18 Ω of loss resistance we intentionally added to simulate ground system loss. That means that the element 1 feed-point resistance would be negative if the ground resistance were less than about an ohm and a half. This isn't uncommon in phased arrays and simply means that the element is feeding power *into* the feed system. This power is coming via mutual coupling from the other elements.

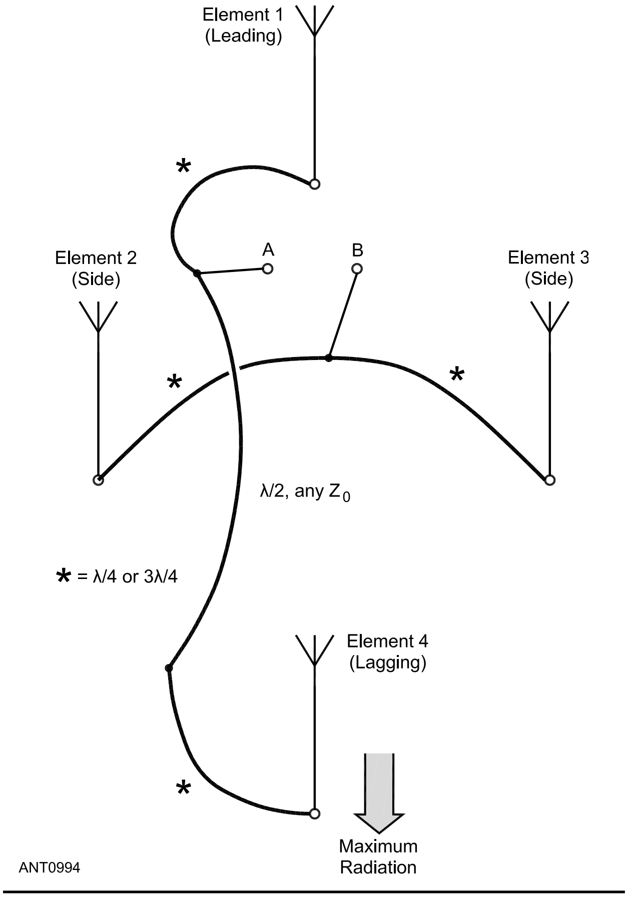

Fig 25—"Simplest" feed system for the Four-Square array in Fig 24. Grounds and cable shields have been omitted for clarity.

"Simplest" (Transmission Line Only) Feed System

To design a "simplest" feed system, start *Arrayfeed1*. In the **Array Type** frame, select **4 Square**, and select **"Simplest"** in the **Feed System Type** frame. In the **Inputs** frame, enter the frequency and the impedances from *EZNEC-ARRL*:

Frequency = 7.15 MHz
Leading Element: R = 16.4, X = –15.85
Side elements: R = 57.47, X = –19.44
Lagging Element: R = 77.81, X = 54.8

We'll try using 50 Ω for all lines, so enter **50** into the next three boxes.

Enter **1** for the lagging:leading I magnitude and **–90** for the phase.

Click **Find Solutions**.

The result is shown in the **Solutions** frame, shown in **Fig 26**. As always when any solution exists, there are two to choose from. The one with the shortest lines is generally preferable, so we'll choose it. For this example, we'll use λ/4 lines with velocity factor of 0.82. So enter **0.82** in the **Velocity Factor** box in the **Physical Lengths** frame, and read the physical lengths from the bottom of that frame. The λ/4 lines (marked in the *Arrayfeed1* diagram with an asterisk) are 28.2 feet, line 1 is 7.483 feet and line 2 is 51.668 feet. The "simplest" feed system is shown in Fig 26, and the complete feed system consists of this connected to the array of Fig 25.

EZNEC-ARRL model **ARRL_4Square_TL_Example.EZ** is simulates the array fed with this system. Comparison of the pattern plot to one from ideal-current model **ARRL_4Square_Example.EZ** and examination of

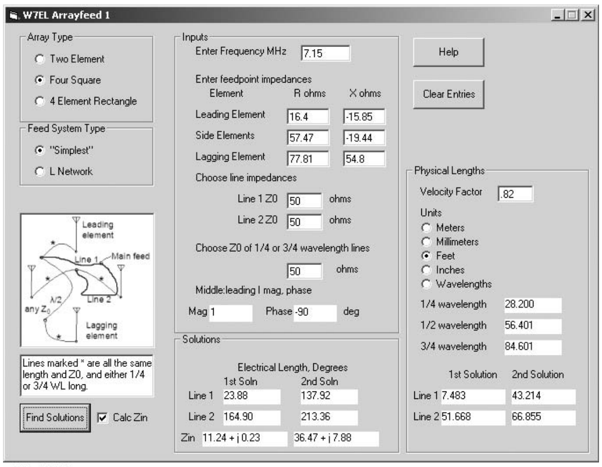

W7EL Arrayfeed 1

Array Type
○ Two Element
● Four Square
○ 4 Element Rectangle

Feed System Type
● "Simplest"
○ L Network

Lines marked * are all the same length and Z0, and either 1/4 or 3/4 WL long.

Find Solutions ☑ Calc Zin

Inputs
Enter Frequency MHz 7.15
Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	16.4	-15.85
Side Elements	57.47	-19.44
Lagging Element	77.81	54.8

Choose line impedances
Line 1 Z0 50 ohms
Line 2 Z0 50 ohms
Choose Z0 of 1/4 or 3/4 wavelength lines 50 ohms
Middle:leading I mag, phase
Mag 1 Phase -90 deg

Solutions
Electrical Length, Degrees

	1st Soln	2nd Soln
Line 1	23.88	137.92
Line 2	164.90	213.36
Zin	11.24 + j 0.23	36.47 + j 7.88

Physical Lengths
Velocity Factor .82
Units
○ Meters
○ Millimeters
● Feet
○ Inches
○ Wavelengths
1/4 wavelength 28.200
1/2 wavelength 56.401
3/4 wavelength 84.601

	1st Solution	2nd Solution
Line 1	7.483	43.214
Line 2	51.668	66.855

Fig 26—Screen capture from *Arrayfeed1* for "Simplest" feed system for Four Square feed system shown in Fig 25.

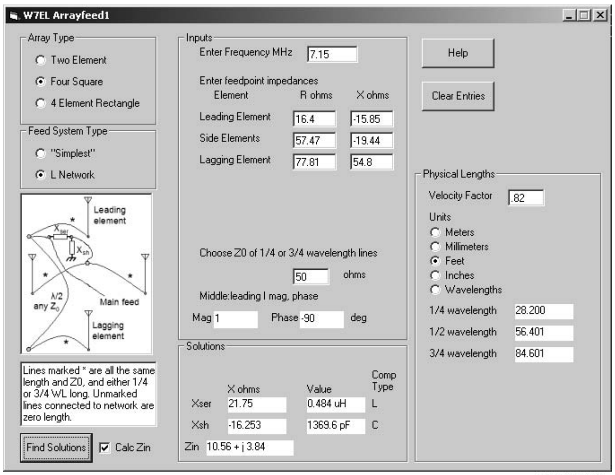

W7EL Arrayfeed1

Array Type
○ Two Element
● Four Square
○ 4 Element Rectangle

Feed System Type
○ "Simplest"
● L Network

Lines marked * are all the same length and Z0, and either 1/4 or 3/4 WL long. Unmarked lines connected to network are zero length.

Find Solutions ☑ Calc Zin

Inputs
Enter Frequency MHz 7.15
Enter feedpoint impedances

Element	R ohms	X ohms
Leading Element	16.4	-15.85
Side Elements	57.47	-19.44
Lagging Element	77.81	54.8

Choose Z0 of 1/4 or 3/4 wavelength lines 50 ohms
Middle:leading I mag, phase
Mag 1 Phase -90 deg

Solutions

	X ohms	Value	Comp Type
Xser	21.75	0.484 uH	L
Xsh	-16.253	1369.6 pF	C

Zin 10.56 + j 3.84

Physical Lengths
Velocity Factor .82
Units
○ Meters
○ Millimeters
● Feet
○ Inches
○ Wavelengths
1/4 wavelength 28.200
1/2 wavelength 56.401
3/4 wavelength 84.601

Fig 27—L-network setup for Four-Square array in Fig 25, fed with λ/4 (or 3λ/4) current-forcing feed system.

the element currents verify that the feed system is producing the desired pattern and element currents. You can use **ARRL_4Square_Example.EZ** to investigate the effect of frequency change, ground loss and other changes on the array gain and pattern.

L-Network Feed System

To design the L-network feed system, simply change the **Feed System Type** to **L Network** and click **Find Solutions**. The results you should see are a 0.484 µH inductor for the series component X_{ser}, and a 1369.6 pF capacitor for the shunt component X_{sh}. The L-network feed system is shown in **Fig 27**, and the complete feed system consists of this L network connected to the array of Fig 25.

EZNEC-ARRL doesn't have the ability to directly model the L network, so it's unable to model the complete system. However, the system has been modeled using the network capability of *EZNEC v.5* and found to work as designed. Arrays have also been built using this feed system and the element currents measured, with exactly the expected results.

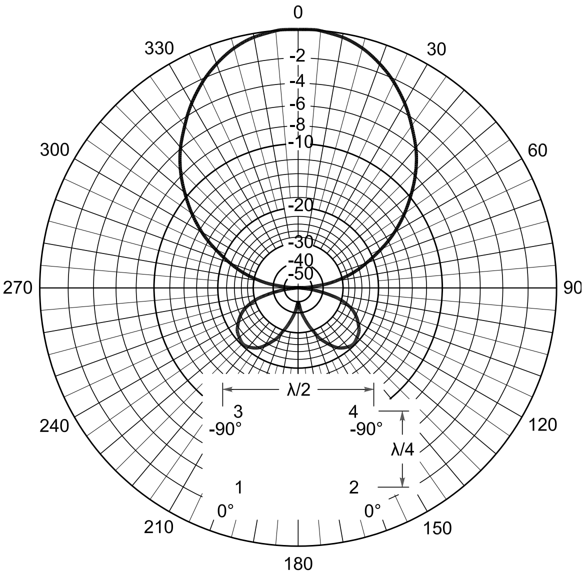

Fig 28—Pattern and layout of the four-element rectangular array. Gain is referenced to a single similar element; add 6.8 dB to the scale values shown.

This array is more sensitive to adjustment than the 2-element 90° fed, 90° spaced array. Adjustment procedures and a method of remotely switching the direction of this array are described in the **Practical Aspects of Phased Array Design** section that follows.

A 4-Element Rectangular Array

The 4-element rectangule array shown with its pattern in **Fig 28** has appeared numerous times in amateur publications. However, many of the accompanying feed systems fail to deliver currents in the proper amounts and phases to the various elements. The array can be correctly fed using the principles discussed in this chapter and the design methods that follow.

Elements 1 and 2 can be forced to be in-phase and to have equal currents by feeding them through 3λ/4 lines. (As in the binomial and Four Square array examples, 3λ/4 lines are chosen because λ/4 lines won't physically reach.) The currents in elements 3 and 4 can similarly be forced to be equal and in-phase. **Fig 29** shows the "current-forcing" feed system. Elements 3 and 4 are made to have currents of equal magnitude but of 90° phase difference from elements 1 and 2 by use of either a "simplest" all-transmission line feed system or an L-network feed system. Both will be designed in this example.

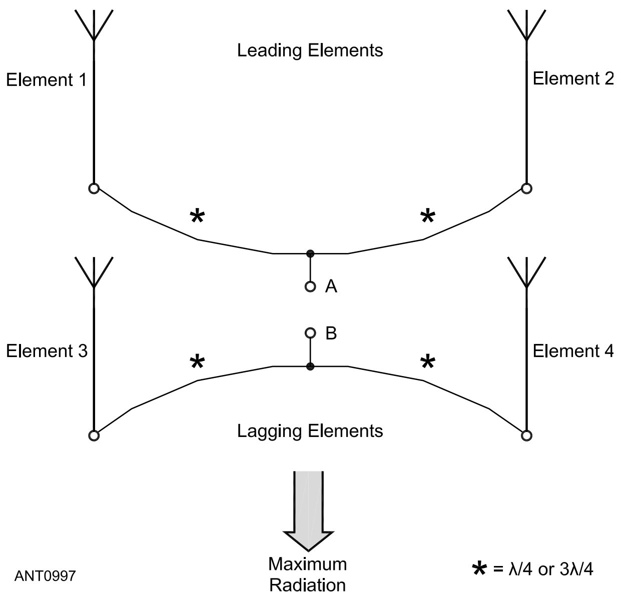

Fig 29—"Simplest" feed system for four-element rectangular array, using four equal-length λ/4 (or 3λ/4) cables.

For this array, we have to know the feed-point impedances of two elements (one of each pair) in order to design either type of feed system. *EZNEC-ARRL* model **Rectangular_Example.EZ** shows a 20-meter rectangular array with 18 Ω of loss resistance at each element, again to approximate an 8-radial per element ground system. (See the cardioid array example above for more information about modeling ground system loss.) Open the file in *EZNEC-ARRL* and click the **Src Dat** button to find the following feed-point impedances:

Sources 1 and 2: 21.44 – *j* 21.29 Ω
Sources 3 and 4: 70.81 – *j* 5.232 Ω

"Simplest" (Transmission- Line Only) Feed System

To design a "simplest" feed system, start program *Arrayfeed1*. In the **Array Type** frame, select **4 Element Rectangle**, and select **"Simplest"** in the **Feed System Type** frame. In the **Inputs** frame, enter the frequency and the impedances from *EZNEC-ARRL*:

Frequency = 14.15 MHz
Leading Elements R = 21.44, X = –21.29
Lagging Elements R = 70.81, X = –5.232

We'll use 50 Ω for all lines, so enter **50** into the next three boxes.

Enter **1** for the lagging:leading I magnitude and **–90** for the phase.

Click Find Solutions.

The result is "No Solution"—This combination of line impedances can't be used. Several other combinations also produce this result, but making lines 1 and 2 each 75 Ω and the 3λ/4 lines 50 Ω does produce a solution. Enter **75** into the **Line 1 Z0** and **Line 2 Z0** boxes, and leave **50** in the **Choose Z0 of ¼ or ¾ wavelength lines** box, then click the Find Solutions button. There won't be any problem making lines 1 and 2 reach, so we'll choose the first solution because the lines are shorter. The physical lengths of all the lines are shown in the **Physical Lengths** frame when the velocity factor is entered in the appropriate box. Assuming that we use coax with a velocity factor of 0.66 (and the example frequency of 14.15 MHz), the lengths are:

Line 1: 4.982 feet
Line 2: 20.153 feet
3λ/4 lines (marked with an asterisk in the *Arrayfeed1* diagram): 34.408 feet

The lines are connected following the diagram in the upper left part of the *Arrayfeed1* window. This completes the "simplest" feed system design. *EZNEC-ARRL* model **Rectangular_TL_Example.EZ** simulates an array fed with this system.

Comparison of the pattern plot to one from ideal-current **Rectangular_Example.EZ**, and examination of the element currents verify that the feed system is producing the desired pattern and element currents.

L-Network Feed System

To design the L-network feed system using *Arrayfeed1*, change the **Feed System Type** to **L Network** and click **Find Solutions**. The resulting L-network values are a 0.484 µH inductor for the series component X_{ser} and a 1369.6 pF capacitor for the shunt component X_{sh}.

120° Fed, 60° Spaced Dipole Array

This example shows the design of "simplest" and L network feed systems for a 2-element 20- meter dipole array, rather than a vertical array. No special accommodation is required for the array made from dipoles rather than vertical elements—the same methods can be used regardless of element shape. This example also shows that both the "simplest" and L-network feed systems can readily be applied to elements that use phase angles other than 90°.

Any 2-element array made with identical elements spaced λ/2 or closer and having equal magnitude currents with a relative phase angle of 180° minus the spacing will produce a unidirectional pattern with a good null to the rear. In practice, very close spacings lead to very low feed-point resistances, with consequent losses and very narrowband characteristics. But this 60° spaced array is well within the range of practical realization. File **ARRL_Dipole_Array_Example.EZ** is a model created for this array, with ideal element currents. Open this file in *EZNEC-ARRL* and click **FF Plot** to show the pattern at an elevation angle of 10°. You can save this pattern for later comparison to the pattern with a "simplest" feed system by opening the **File** menu in the **2D Plot** window, selecting **Save Trace As**, entering a name for the trace file and clicking **Save**.

Following the same procedure as in the previous examples, we begin the array design by finding the element feed-point impedances in the ideally fed array using *EZNEC-ARRL* numbers. Having already opened **ARRL_Dipole_Array_Example.EZ**, all that's needed is to click **Src Dat.** The results are:

Leading element (source 1) : 36.16 – j46.05 Ω
Lagging element (source 2) : 49.56 + j51.47 Ω

"Simplest" (Transmission Line Only) Feed System

Select **Two Element** for the **Array Type** in *Arrayfeed1* and **"Simplest"** for the **Feed System Type**. Enter the frequency of 14.15 MHz and enter the element feed-point impedances from *EZNEC-ARRL* into the appropriate boxes in the **Inputs** frame. For line impedances, the section describing the "simplest" feed system recommends against choosing one which is very different from the element feed-point impedances, but for fun let's try 300 Ω for the two lines and see what happens. Enter **300** in the **Line 1 Z0** and **Line 2 Z0** boxes. Finally, enter the **lagging:leading I mag, phase** of **1** for **Mag** and **–120** for **Phase**.

Click **Find Solutions**. For this example we'll assume that TV type twinlead with a velocity factor of 0.8 is being used. So enter **0.8** for the **Velocity Factor** and read the physical line lengths in the **Physical Lengths** frame. A model of the array using the first solution has been created as **ARRL_Dipole_Array_TL_Example.EZ**. Open this file in *EZNEC-ARRL* and click **FF Tab**. You should see that the plot is virtually identical to the one saved earlier from the ideal-current model. Note the gain and front/back ratio or 8.79 dBi and 31.01 dB respectively reported in the data box below the 2D plot.

Don't subtract 2.15 dB to find the gain relative to a single element! This isn't a free-space model, and the gain of a single dipole over ground is much greater than 2.15 dBi. Instead, delete one of the elements in **ARRL_Dipole_Array_Example.EZ** to find the gain of a single element and subtract that value from the array gain. You can use the undo feature or re-open the file to restore the array.

Now, go back to the model with the "simplest" feed system in *EZNEC-ARRL* and change the Frequency to **14.0** MHz. Click **FF Tab** again. The gain has decreased a little, to 8.54 dBi and the front/back ratio has also decreased, to 21.8 dB. At 14.3 MHz, the gain is slightly higher, 9.04 dBi, but the front/back is again worse, down to 18.64 dB. But this isn't bad overall.

Let's take a look at the second solution. Click the **Trans Lines** line in the main *EZNEC-ARRL* window to open the Transmission Lines Window. Change the length of the first line to **26.856** feet, the second to **28.356** feet, and press the **Enter** key to complete the change. Change the **Frequency** back to **14.15** MHz and click **FF Tab**. You should see exactly the same pattern as for both the first solution and for the ideal current model. But now change the **Frequency** to **14.0** MHz, click FF Tab, and look at the pattern.

What happened? The gain has dropped to 5.95 dBi and the front/back to only 3.1 dB. The array is now nearly bi-directional! It's almost as bad at 14.3 MHz. So we've created a terribly touchy system. The chance of its working correctly even at the design frequency is slim, because there are inevitably some differences between the model and real antenna.

We did have a clue this might happen. As stated in the section describing the "simplest" feed system, the best choices for line Z_0 and for the resulting solution give a difference in electrical line lengths about equal to the desired phase delay of the current. The difference in electrical line lengths for the first solution was about 152°—not as close to the 120° current phase difference as we'd like, but much better than the mere 9.7° difference of the lines for the second solution. While the 300-Ω line Z_0 is quite different from the element feed-point impedances, the first solution result is quite good. If desired, you can try other line impedance values into *Arrayfeed1* and evaluate the results with *EZNEC-ARRL*.

Please see the information about baluns in the **Baluns in Phased Arrays** section. Baluns are placed the same as in Fig 19, which shows the L-network feed system.

L-Network Feed System

To design an L network feed system, change the *Arrayfeed1* **Feed System Type** to **L Network** and click **Find Solutions**. The results aren't good ones to use. The component reactance magnitudes of about 1573 and 2619 Ω are more than five times the 300-Ω Z_0 of the feed lines. As explained in the section describing the L-network feed system, it's undesirable to have such a large ratio of component reactance to line Z_0. Among other problems, the inductor and capacitor values are quite extreme and capacitor stray inductance and inductor capacitance would have a significant impact on performance.

The problem occurs because the feed-line impedance we chose is much larger than the element feed-point impedances, so the $\lambda/4$ lines transform the feed-point impedances to much higher values at the L network and main feed point. This feed system would be extremely critical, narrowbanded and difficult to adjust. We can do better by choosing feed-line impedances that aren't too drastically different than the element feed-point impedances. In this case, 50 or 75 Ω would be a much better choice than 300. Let's try 75.

In *Arrayfeed1*, change the **Line 1 Z0** and **Line 2 Z0** impedances from **300** to **75** and click **Find Solutions**. L-network component reactance magnitudes are now about 98 and 164 Ω, much better than before. This will be a relatively uncritical and broadbanded feed system.

Again, be sure to read the information about baluns in the **Baluns in Phased Arrays** section. Fig 19 shows the completed feed system including baluns.

PRACTICAL ASPECTS OF PHASED ARRAY DESIGN

With almost any type of antenna system, there is much that can be learned from experimenting with, testing and using various array configurations. In this section, Roy Lewallen, W7EL, shares the benefit of years of his experience from actually building, adjusting and using phased arrays. There is much more work to be done in most of the areas covered here, and Roy encourages the reader to build on this work.

Adjusting Phased Array Feed Systems

If a phased array is constructed only to achieve forward gain, adjusting it is seldom worthwhile. This is because the forward gain of most arrays is quite insensitive to either the magnitude or phase of the relative currents flowing in the elements. If, however, good rejection of unwanted signals is desired, adjustment may be required. And achieving very deep nulls will almost surely require some adjustment.

The in-phase and 180° out-of-phase current-forcing method supplies very well-balanced and well-phased currents to elements without adjustment. If the pattern of an array fed using this method is unsatisfactory, it's generally the result of environmental differences—where the elements, even though furnished with correct currents, aren't generating the correct fields. Such an array can be optimized in a single direction, but a more general approach than the current-forcing method must be taken. Some possibilities are described by Paul Lee and Forrest Gehrke (see Bibliography).

Unlike the current-forcing method, the "simplest" and L-network feed systems described earlier in this chapter are dependent on the self and mutual impedance of one or more elements. The required transmission-line lengths or L-network component values can be computed to a high level of precision, but the results are only as good as the knowledge of the relevant feed-point impedances.

While the simplest feed system doesn't readily lend itself to adjustment, the components of an L network can easily be made adjustable or can be experimentally changed in increments. A practical approach is to model the array as accurately as possible, design and build the feed system based on the model results and then adjust the network for the best performance.

Simple arrays such as the two-element 90° fed and spaced array can be adjusted as follows. Place a low-power signal source at a distance from the array (preferably several wavelengths), in the direction a null should be. While listening to the signal on a receiver connected to the array, alternately adjust the two L-network components for the best rejection of the signal.

This has proved to be a very good way to adjust 2-element arrays. However, variable results were obtained when a Four-Square array was adjusted using this technique. The probable reason is that more than one combination of current balance and phasing can produce a null in a given direction but each produces a different overall pattern. So a different method must be used for adjusting more complex arrays. This involves actually measuring the element currents in some way, and adjusting the network until the currents are correct. After adjusting the currents, small adjustments can be made to deepen the null(s) if desired.

Measuring Element Currents

You can measure the element currents two ways. One way is to measure them directly at the element feed points, as shown in **Fig 30**. A dual-channel oscilloscope is required to monitor the currents. This method is the most accurate and it provides a direct indication of the actual relative magnitudes and phases of the element currents. The current probe is shown in **Fig 31**.

Instead of measuring the element currents directly, you could measure them indirectly by measuring the voltages on the feed lines an electrical λ/4 or 3λ/4 distance from the array. The voltages at these points are directly proportional to the element currents. This introduces additional variables that can reduce the accuracy of the result, but the method generally produces adequate performance. The 2-element arrays fed with the L-network system and all the four element arrays presented earlier have λ/4 or 3λ/4 lines from all elements to a common location, making this second measurement method convenient. The voltages can be observed with a dual-channel oscilloscope, or, to adjust for equal-magnitude currents and 90° phasing, you can use the test circuit shown in **Fig 32**.

The test circuit is connected to the feed lines of two elements that are to be adjusted for 90° phasing (such as elements 1 and 2, or 2 and 4 of the Four-Square array of Figs 24 and 25). Adjust the L-network components alternately until both meters read zero. Proper operation of the test circuit can be verified by disconnecting one of the inputs. The *phase* output

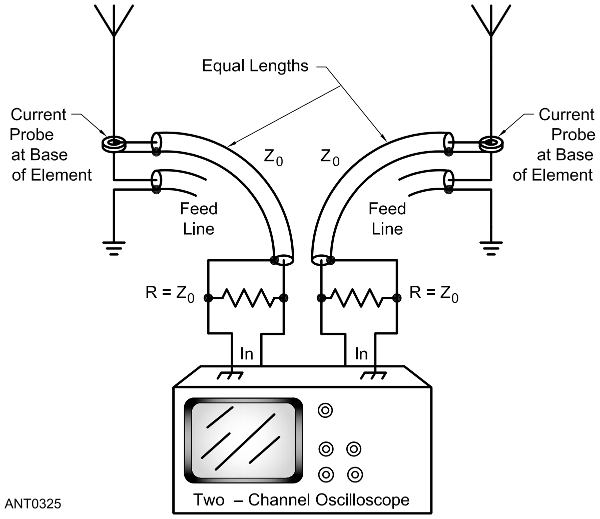
Equal Lengths
Current Probe at Base of Element
Current Probe at Base of Element
Z_0
Z_0
Feed Line
Feed Line
$R = Z_0$
$R = Z_0$
In
In
Two – Channel Oscilloscope
ANT0325

Fig 30—One method of measuring element currents in a phased array. Details of the current probe are given in Fig 31. Caution: Do not run high power to the antenna system for this measurement, or damage to the test equipment may result.

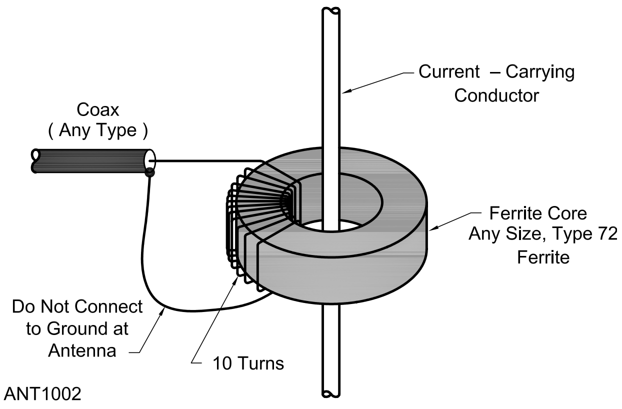
Current – Carrying Conductor
Coax (Any Type)
Ferrite Core Any Size, Type 72 Ferrite
Do Not Connect to Ground at Antenna
10 Turns
ANT1002

Fig 31—The current probe for use in the test setup of Fig 30. The ferrite core is of type 72 material, and may be any size. The coax line must be terminated at the opposite end with a resistor equal to its characteristic impedance. You should build this probe in a plastic or metal box to provide mechanical ruggedness.

should remain close to zero. If not, there is an undesirable imbalance in the circuit, which must be corrected. Another means of verification is to first adjust the L network so the tester indicates correct phasing (zero volts at the *phase* output). Then reverse the tester input connections to the elements. The *phase* output should remain close to zero.

Directional Switching of Arrays

One ideal directional-switching method would take the entire feed system, including the lines to the elements and physically rotate it. The smallest possible increment of rotation would depend on the symmetry of the array—the feed system would need to rotate until the array again *looks* the same to it. For example, any 2-element array can be rotated 180° (although that wouldn't accomplish anything if the array is bidirectional to begin with). The 4-element rectangular array of Figs 28 and 29 can also be reversed, and the Four-Square array of Figs 24 and 25 can be switched in 90° increments.

Smaller switching increments can be accomplished only by reconfiguring the feed system, including any network if used, effectively creating a different kind of array. Switching in smaller increments than dictated by symmetry will create a different pattern in some directions than in others, and must be thoughtfully done to maintain equal and properly phased element currents. The methods illustrated here will deal only with switching in increments related to the array symmetry, except for one: a 2-element broadside/end-fire array.

In all arrays, the success of directional switching depends on the elements and ground systems being identical so that equal element currents result in equal fields. It's even more important in arrays fed with any method other than current forcing, because the effectiveness of those methods depends on the element feed-point impedances. Few of us can afford the luxury of having an array many wavelengths away from all other conductors, so an array will nearly always perform somewhat differently in each direction. The array should be adjusted when steered in the direction requiring the most signal rejection in the nulls. Forward gain will, for all practical purposes, be equal in all the switched directions, since gain is much more tolerant of error than are nulls.

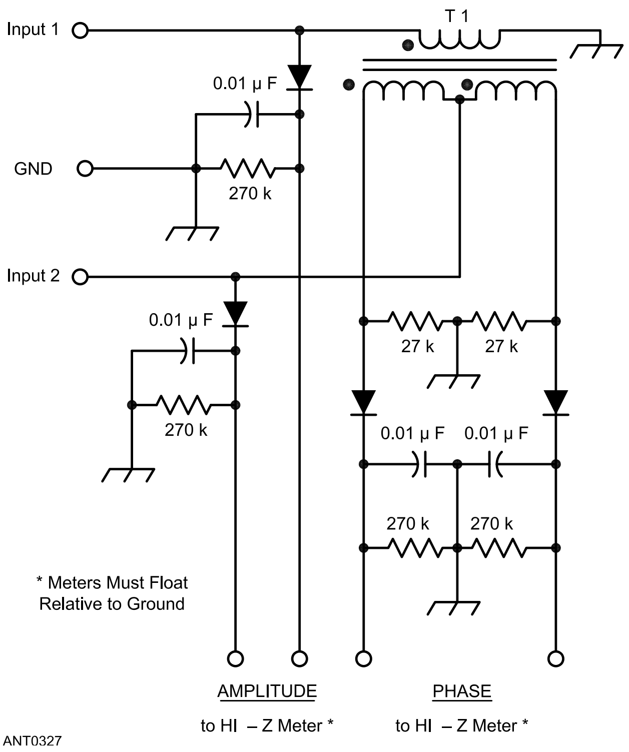

Fig 32—Quadrature test circuit. All diodes are germanium, such as 1N34A, 1N270, or equiv. Hot carrier or silicon diodes can be used at higher power levels. All resistors are ¼ or ½ W, 5% tolerance. Capacitors are ceramic. Alligator clips are convenient for making the input and ground connections to the array.

T1—7 trifilar turns on an Amidon FT-37-43, -75, -77, or equivalent ferrite toroid core.

Basic Switching Methods

Following is a discussion of basic switching methods, how to power relays through the main feed line and other practical considerations. In diagrams, grounds are frequently omitted to aid clarity, but connections of the ground conductors must be carefully made. In fact, it is recommended that the ground conductors be switched just as the center conductors are, as explained in more detail in **Improving Array Switching Systems** below. In all cases, interconnecting lines must be very short.

A pair of elements spaced λ/2 apart can readily be switched between broadside and end-fire bidirectional pat-

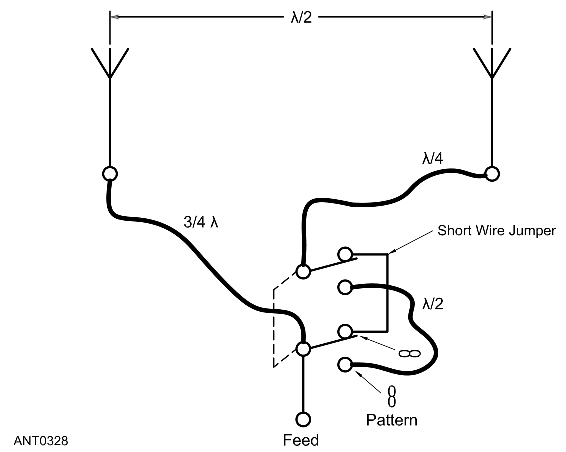

Fig 33—Two-element broadside/end-fire switching. All lines must have the same characteristic impedance. Grounds and cable shields have been omitted for clarity.

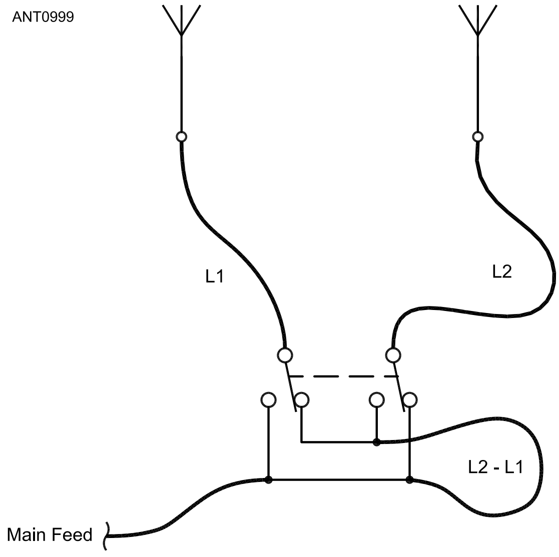

Fig 34—Directional switching for 90°, 90° spaced 2-element array fed with a "simplest" feed system.

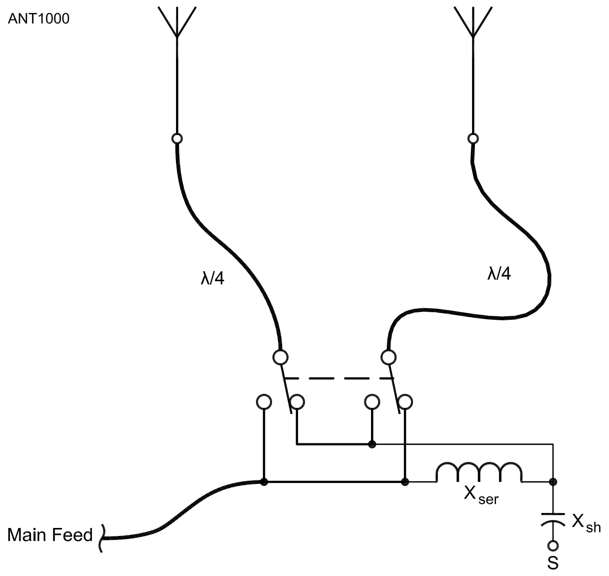

Fig 35—Directional switching for 90°, 90° spaced 2-element array fed with an L-network, current-forcing feed system.

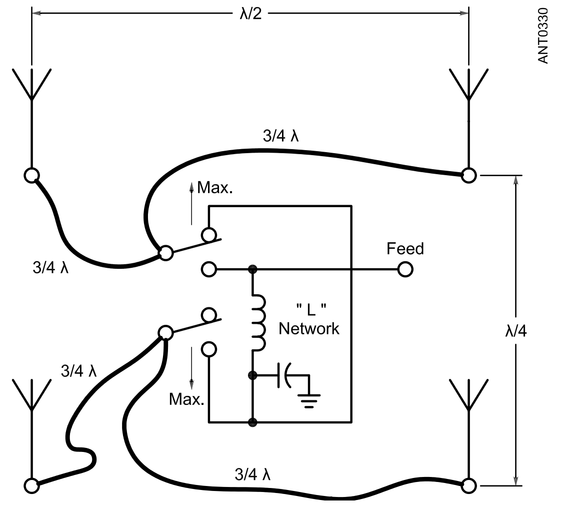

Fig 36—Directional switching of a four-element rectangular array. All interconnections must be very short. As usual, grounds and cable shields have been omitted for clarity.

terns, using the current-forcing properties of λ/4 lines. The method is shown in **Fig 33**. The switching device can be a relay powered via a separate cable or by dc sent along the main feed line.

Fig 34 shows directional switching of a 90° fed, 90° spaced array fed with a "simplest" feed system, where L1 and L2 are the required lengths of the two feed lines. **Fig 35** shows how to switch the same array when fed with an L-network, current-forcing system.

The rectangular array of Fig 28 can be switched in a similar manner, as shown in **Fig 36**. To switch a "simplest" fed rectangular array, use the switching circuit of Fig 34, but connect the two equal length lines to points A and B of Fig 29 in place of the two elements shown in Fig 34.

Switching the direction of an array in increments of 90°, when permitted by symmetry, requires at least two relays. A method of 90° switching of the Four-Square array with L-network feed is shown in **Fig 37**.

Powering Relays Through Feed Lines

All of the above switching methods can be implemented without additional wires to the switch box. A single-relay system is shown in **Fig 38A**, and a two-relay system in

Fig 38B. Small 12 or 24-V dc power relays can be used in either system at power levels up to at least a few hundred watts. Do not attempt to change directions while transmitting, however. Blocking capacitors C1 and C2 should be good quality ceramic or transmitting mica units of 0.01 to 0.1 µF. No problems have been encountered using 0.1 µF, 300-V monolithic ceramic units at RF output levels up to 300 W. C2 may be omitted if the antenna system is an open circuit at dc. C3 and C4 should be ceramic, 0.001 µF or larger.

In Fig 38B, capacitors C5 through C8 should be selected with the ratings of their counterparts in Fig 38A, as given above. Electrolytic capacitors across the relay coils, C9 and C10 in Fig 38B, should be large enough to prevent the relays from buzzing, but not so large as to make relay operation too slow. Final values for most relays will be in the range from 10 to 100 µF. They should have a voltage rating of at least double the relay coil voltage. Some relays do not require this capacitor. All diodes are 1N4001 or similar. A rotary switch may be used in place of the two toggle switches in the two-relay system to switch the relays in the desired sequence.

Improving Array-Switching Systems

The extra circuitry involved in switching arrays can degrade array performance by altering the relative currents fed to each element. One common cause is current sharing in common ground conductors, even when connections are kept very short. The author has seen a 30° phase shift in voltage along a 4-inch piece of #12 wire in a 40-meter array feed system.

When the two conductors of a feedline are physically separated from each other, the impedance increases. This is especially true when the main lines are coaxial cables. If currents from two elements share the ground conductor of a split line, a relatively large voltage drop results. Voltage changes λ/4 from the elements translate to current changes at the elements. Although keeping all leads extremely short is sometimes adequate, the best way to reduce current sharing problems is to keep the two conductors of each transmission line as close together as possible, and switch both conductors of each line rather than just a single or "hot" conductor.

An example of a carefully designed switching system is shown in **Fig 39**. It avoids the problem of shared ground conductor currents, as well as another common problem, namely that effective line lengths are often different along different switching paths. Notice how the path from the main feed point travels through a single line to each element with no common ground connections to other lines except at the

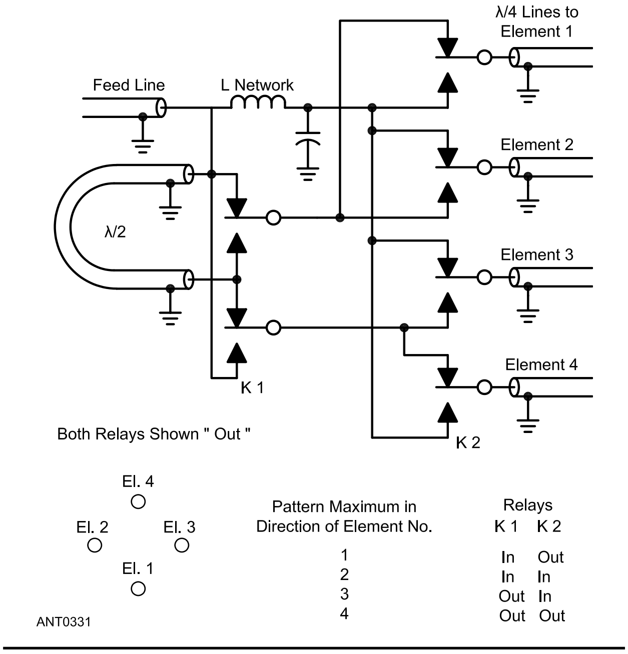

Pattern Maximum in Direction of Element No.	Relays K 1	Relays K 2
1	In	Out
2	In	In
3	Out	In
4	Out	Out

Fig 37—Directional switching of the four-square array. All interconnections must be very short.

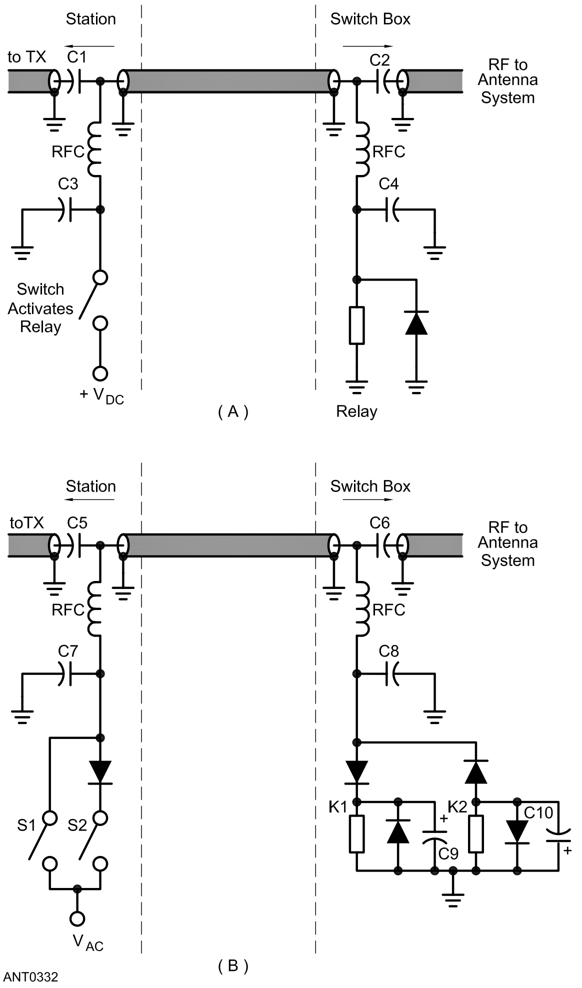

Fig 38—Remote switching of relays. See text for component information. A one-relay system is shown at A, and a two-relay system at B. In B, S1 activates K1, and S2 activates K2.

main feed point. Notice also that the distance doesn't change as the direction is switched. The λ/4 lines going to the two elements must be shortened by the length ℓ of the lines on the feed side of the relays, so that the total line length from the main feed point to each element is λ/4 (or 3λ/4).

You can see that in either relay position, there's an open ended stub of length ℓ connected at the main feed point and another at the output end of the L network. These will add capacitance at those points. Extra C at the main feed point will alter the overall impedance seen by a transmitter, but won't otherwise have any effect on the array or its performance. The one at the output of the L network will, however, change the transformation and phase shift properties of the network. But it's easy to compensate—the value of the shunt capacitor element is simply reduced by the amount of the C added by the stub. The amount of C for any kind of transmission line can be calculated from:

$$C(\text{pf}/\text{ft}) = \frac{1017}{Z_0 \text{VF}}$$

or

$$C(\text{pf}/\text{m}) = \frac{3336}{Z_0 \text{VF}}$$

where Z_0 = the characteristic impedance of the line and VF = the velocity factor. This works out to 31 pF/foot or 101 pF/meter for 50-Ω solid polyethylene insulated coax, which has a velocity factor of 0.66.

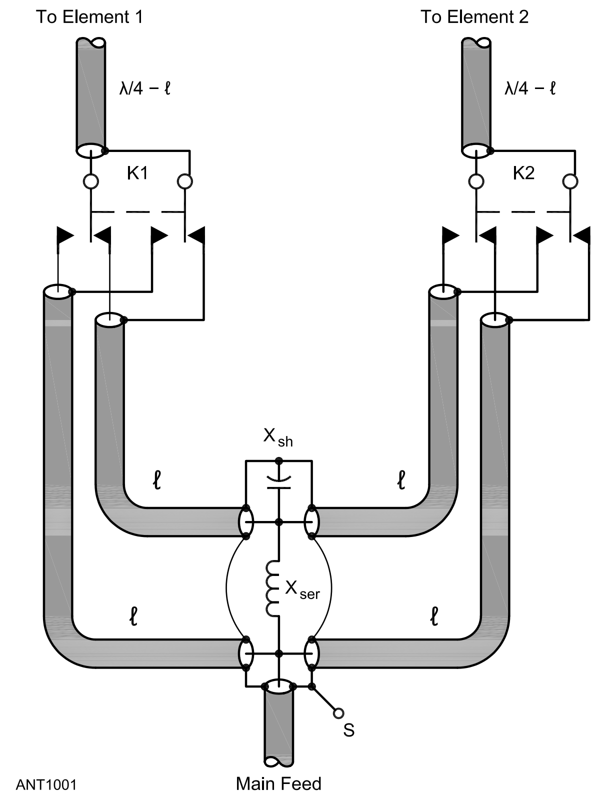

Fig 39—A carefully designed L-network, current-forcing switching system that switches both hot and shield conductors in feed coaxes.

The general principles illustrated in Fig 39 can be extended to other switching systems. If switching the ground conductors as described above isn't practical, use of a metal box for the switching circuitry is recommended, so that the relatively large surface area of the box can be used for the common ground conductors, minimizing their inductance. Always keep leads extremely short.

Measuring the Electrical Length of Feed Lines

When using the feed methods described earlier the feed lines must be very close to the correct length. For best results, they should be correct within 1% or so. This means that a line that is intended to be, say, λ/4 at 7 MHz, should actually be λ/4 at some frequency within 70 kHz of 7 MHz. A simple but accurate method to determine at what frequency a line is λ/4 or λ/2 is shown in **Fig 40A**. The far end of the line is short circuited with a very short connection. A signal is applied to the input and the frequency is swept until the impedance at the input is a minimum. This is the frequency at which the

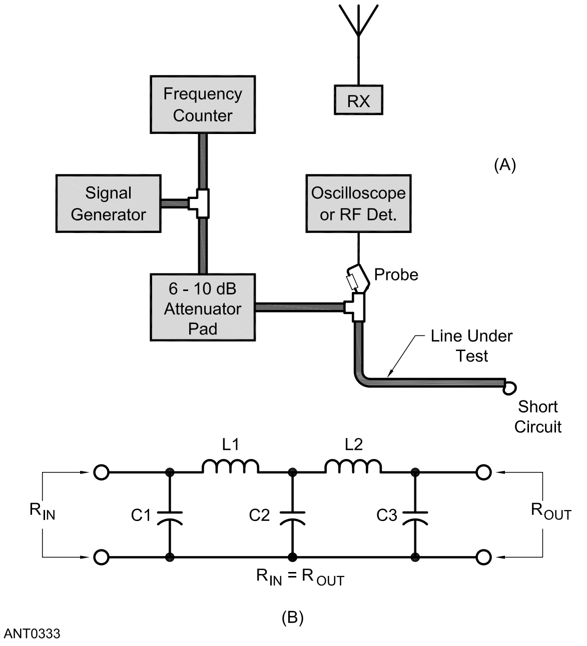

Fig 40—At A, the setup for measurement of the electrical length of a transmission line. The receiver may be used in place of the frequency counter to determine the frequency of the signal generator. The signal generator output must be free of harmonics; the half-wave harmonic filter at B may be used outboard if there is any doubt. It must be constructed for the frequency band of operation. Connect the filter between the signal generator and the attenuator pad.

C1, C3—Value to have a capacitive reactance = R_{IN}.
C2—Value to have a capacitive reactance = ½ R_{IN}.
L1, L2—Value to have an inductive reactance = R_{IN}.

line is λ/2. Either the frequency counter or the receiver may be used to determine this frequency. The line is, of course, λ/4 at one half the measured frequency.

The detector can be a simple diode detector or an oscilloscope may be used if available. A 6 to 10 dB attenuator pad is included to prevent the signal generator from looking into a short circuit at the measurement frequency. The signal generator output must be free of harmonics. If there is any doubt, an outboard low-pass filter, such as a half-wave harmonic filter, should be used. The half-wave filter circuit is shown in Fig 40B, and must be constructed for the frequency band of operation.

Another satisfactory method is to use a noise or resistance bridge or antenna analyzer at the input of the line, again looking for a low impedance at the input while the output is short circuited. Simple resistance bridges are described in Chapter 27.

Dip oscillators have been found to be unsatisfactory. The required coupling loop has too great an effect on measurements.

Measuring Element Self and Mutual Impedances

The need for measuring element self and mutual impedances has been made largely unnecessary with the ready availability of modeling software. Few amateurs appreciate the considerable difficulty of making accurate impedance measurements and accurate mutual impedance measurements are very difficult even with professional test equipment and skills. Despite the limitations of computer modeling, results very often are better than measured values because of the multiple factors affecting measurement accuracy.

Those who are interested in measuring self and mutual impedances can find more detailed information about doing so in Appendix B. The information there is from earlier editions of *The ARRL Antenna Book*.

Broadside Arrays

Broadside arrays can be made up of collinear or parallel elements or combinations of the two. This section was contributed by Rudy Severns, N6LF.

COLLINEAR ARRAYS

Collinear arrays are always operated with the elements in-phase. (If alternate elements in such an array are out-of-phase, the system simply becomes a harmonic type of antenna.) A collinear array is a broadside radiator, the direction of maximum radiation being at right angles to the line of the antenna.

POWER GAIN

Because of the nature of the mutual impedance between collinear elements, the feed-point resistance (compared to a single element, which is ≈73 Ω) is increased as shown earlier in this chapter (Fig 9). For this reason the power gain does not increase in direct proportion to the number of elements. The gain with two elements, as the spacing between them is varied, is shown by **Fig 41**. Although the gain is greatest when the end-to-end spacing is in the region of 0.4 to 0.6 λ, the use of spacings of this order is inconvenient constructionally and introduces problems in feeding the two elements. As a result, collinear elements are almost always operated with their ends

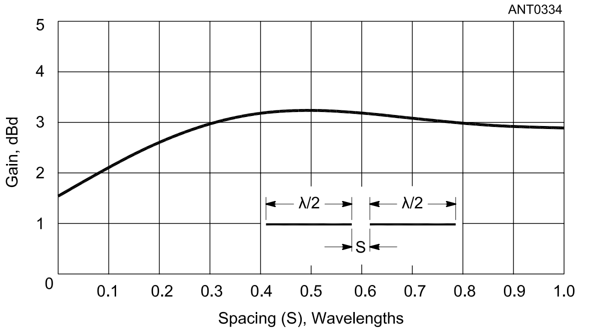

Fig 41—Gain of two collinear λ/2 elements as a function of spacing between the adjacent ends.

quite close together—in wire antennas, usually with just a strain insulator between.

With very small spacing between the ends of adjacent elements the theoretical power gain of collinear arrays, assuming the use of #12 copper wire, is approximately as follows over a dipole in free space:

2 collinear elements—1.6 dB
3 collinear elements—3.1 dB
4 collinear elements—3.9 dB

More than four elements are rarely used.

DIRECTIVITY

The directivity of a collinear array, in a plane containing the axis of the array, increases with its length. Small secondary lobes appear in the pattern when more than two elements are used, but the amplitudes of these lobes are low enough so that they are usually not important. In a plane at right angles to the array the directive diagram is a circle, no matter what the number of elements. Collinear operation, therefore, affects only E-plane directivity, the plane containing the antenna.

When a collinear array is mounted with the elements vertical, the antenna radiates equally well in all geographical directions. An array of such *stacked* collinear elements tends to confine the radiation to low vertical angles.

If a collinear array is mounted horizontally, the directive pattern in the vertical plane at right angles to the array is the same as the vertical pattern of a simple λ/2 antenna at the same height (Chapter 3).

TWO-ELEMENT ARRAYS

The simplest and most popular collinear array is one using two elements, as shown in **Fig 42**. This system is commonly known as *two half-waves in phase*. The directive pattern in a plane containing the wire axis is shown in **Fig 43**, which shows superimposed patterns for a dipole and 2, 3 and 4-element collinear arrays. Depending on the conductor size, height, and similar factors, the impedance at the feed point can be expected to be in the range of 4 to 6 kΩ, for wire antennas. If the elements are made of tubing having a low λ/dia (wavelength to diameter) ratio, values as low as 1 kΩ are representative. The system can be fed through an open-wire tuned line with negligible loss for ordinary line lengths, or a matching section may be used if desired.

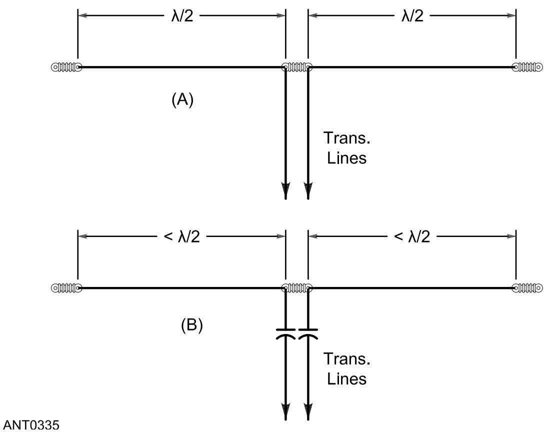

Fig 42—At A, two-element collinear array (two half-waves in phase). The transmission line shown would operate as a tuned line. A matching section can be substituted and a nonresonant line used if desired, as shown at B, where the matching section is two series capacitors.

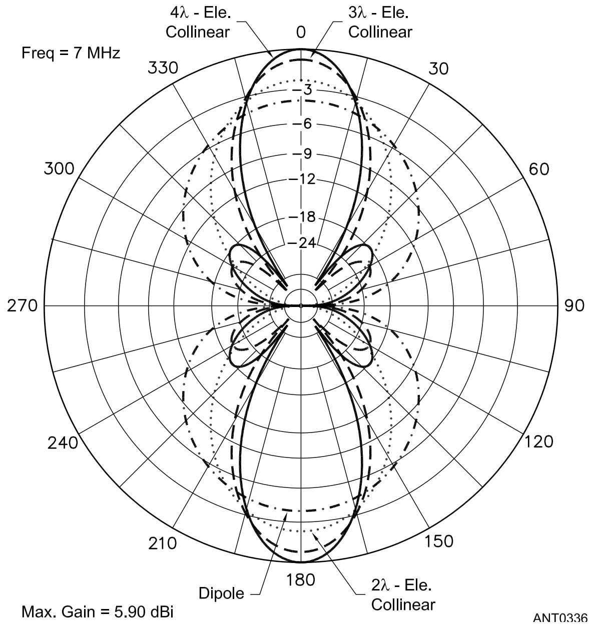

Fig 43—Free-space E-plane directive diagram for dipole, 2, 3 and 4-element collinear arrays. The solid line is a 4-element collinear; the dashed line is for a 3-element collinear; the dotted line is for a 2-element collinear and the dashed-dotted line is for a λ/2 dipole.

A number of arrangements for matching the feed line to this antenna are described in Chapter 26. If elements somewhat shorter than λ/2 are used, then additional matching schemes can be employed at the expense of a slight reduction in gain. When the elements are shortened two things happen—the impedance at the feed-point drops and the impedance has inductive reactance that can be tuned out with simple series capacitors, as shown in Fig 42B.

Note that these capacitors must be suitable for the power level. Small *doorknob* capacitors such as those frequently used in power amplifiers, are suitable. By way of an example, if each side of a 40-meter 2-element array is shortened from 67 to 58 feet, the feed-point impedance drops from nearly 6000 Ω to about 1012 Ω with an inductive reactance of 1800 Ω. The reactance can be tuned out by inserting 25 pF capacitors at the feed-point. The 1012 Ω resistance can be transformed to 200 Ω using a λ/4 matching section made of 450-Ω ladder line and then transformed to 50 Ω with a 4:1 balun. Shortening the array as suggested reduces the gain by about 0.5 dB.

Another scheme that preserves the gain is to use a 450-Ω

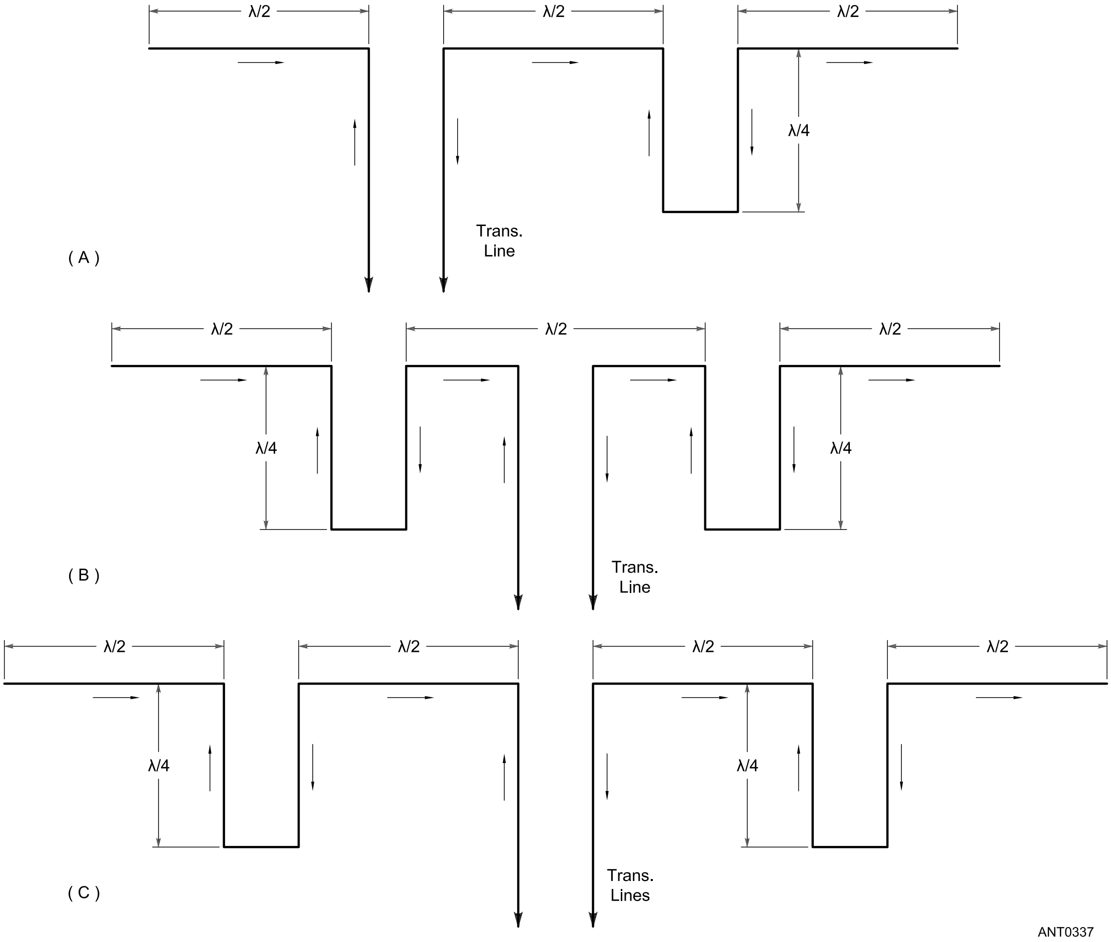

Fig 44—Layouts for 3- and 4-element collinear arrays. Alternative methods of feeding a 3-element array are shown at A and B. These drawings also show the current distribution on the antenna elements and phasing stubs. A matched transmission line can be substituted for the tuned line by using a suitable matching section.

λ/4 matching section and shorten the antenna only slightly to have a resistance of 4 kΩ. The impedance at the input of the matching section is then near 50 Ω and a simple 1:1 balun can be used. Many other schemes are possible. The free-space E-plane response for a 2-element collinear array is shown in Fig 43, compared with the responses for more elaborate collinear arrays described below.

THREE- AND FOUR-ELEMENT ARRAYS

In a long wire the direction of current flow reverses in each λ/2 section. Consequently, collinear elements cannot simply be connected end to end; there must be some means for making the current flow in the same direction in all elements. When more than two collinear elements are used it is necessary to connect *phasing* stubs between adjacent elements in order to bring the currents in all elements in-phase. In **Fig 44A** the direction of current flow is correct in the two left-hand elements because the shorted λ/4 transmission line (*stub*) is connected between them. This stub may be looked upon simply as the alternate λ/2 section of a long-wire antenna folded back on itself to cancel its radiation. In Fig 44A the part to the right of the transmission line has a total length of three half wavelengths, the center half wave being folded back to form a λ/4 phase-reversing stub. No data are available on the impedance at the feed point in this arrangement, but various considerations indicate that it should be over 1 kΩ.

An alternative method of feeding three collinear elements is shown in Fig 44B. In this case power is applied at the center of the middle element and phase-reversing stubs are used between this element and both of the outer elements. The impedance at the feed point in this case is somewhat over 300 Ω and provides a close match to 300 Ω line. The SWR will be less than 2:1 when 600-Ω line is used. Center feed of this type is somewhat preferable to the arrangement in Fig 44A because the system as a whole is balanced. This assures more uniform power distribution among the elements. In Fig 44A, the right-hand element is likely to receive somewhat less power than the other two because a portion of the input power is radiated by the middle element before it can reach the element located at the extreme right.

A four-element array is shown in Fig 44C. The system is symmetrical when fed between the two center elements as shown. As in the three-element case, no data are available on the impedance at the feed point. However, the SWR with a 600 Ω line should not be much over 2:1.

Fig 43 compares the directive patterns of 2, 3 and 4-element arrays. Collinear arrays can be extended to more than four elements. However, the simple 2-element collinear array is the type most frequently used, as it lends itself well to multi-band operation. More than two collinear elements are seldom used because more gain can be obtained from other types of arrays.

ADJUSTMENT

In any of the collinear systems described, the lengths of the radiating elements in feet can be found from the formula $468/f_{MHz}$. The lengths of the phasing stubs can be found from the equations given in Chapter 26 for the type of line used. If the stub is open-wire line (500 to 600 Ω impedance) you may assume a velocity factor of 0.975 in the formula for a λ/4 line. On-site adjustment is, in general, an unnecessary refinement. If desired, however, the following procedure may be used when the system has more than two elements.

Disconnect all stubs and all elements except those directly connected to the transmission line (in the case of feed such as is shown in Fig 44B leave only the center element connected to the line). Adjust the elements to resonance, using the still-connected element. When the proper length is determined, cut all other elements to the same length. Make the phasing stubs slightly long and use a shorting bar to adjust their length. Connect the elements to the stubs and adjust the stubs to resonance, as indicated by maximum current in the shorting bars or by the SWR on the transmission line. If more than three or four elements are used it is best to add elements two at a time (one at each end of the array), resonating the system each time before a new pair is added.

THE EXTENDED DOUBLE ZEPP

One method to obtain higher gain that goes with wider spacing in a simple system of two collinear elements is to make the elements somewhat longer than λ/2. As shown in **Fig 45**, this increases the spacing between the two in-phase λ/2 sections at the ends of the wires. The section in the center carries a current of opposite phase, but if this section is short the current will be small; it represents only the outer ends of a λ/2 antenna section. Because of the small current and short length, the radiation from the center is small. The optimum length for each element is 0.64 λ. At greater lengths the system tends to act as a long-wire antenna, and the gain decreases.

This system is known as the *extended double Zepp*. The gain over a λ/2 dipole is approximately 3 dB, as compared with about 1.6 dB for two collinear λ/2 dipoles. The directional pattern in the plane containing the axis of the antenna

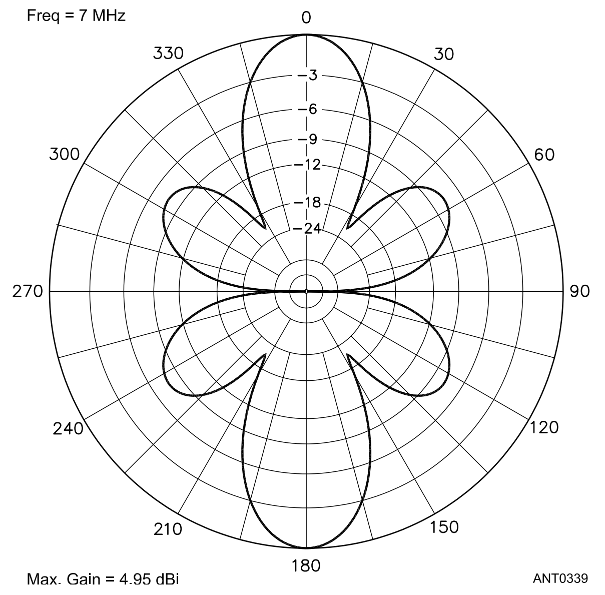

Fig 46—E-plane pattern for the extended double Zepp of Fig 45. This is also the horizontal directional pattern when the elements are horizontal. The axis of the elements lies along the 90°-270° line. The free-space array gain is approximately 4.95 dBi.

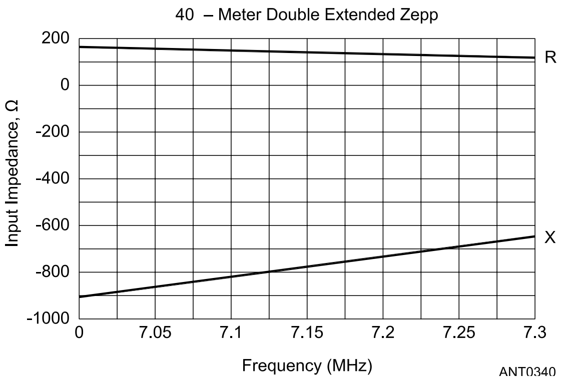

Fig 47—Resistive and reactive feed-point impedance of a 40-meter extended double Zepp in free space.

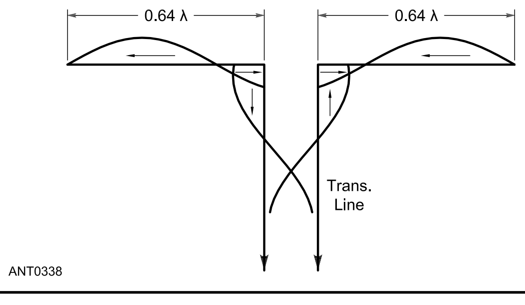

Fig 45—The extended double Zepp. This system gives somewhat more gain than two λ-sized collinear elements.

is shown in **Fig 46**. As in the case of all other collinear arrays, the free-space pattern in the plane at right angles to the antenna elements is the same as that of a λ/2 antenna—circular.

This antenna is not resonant at the operating frequency so that the feed-point impedance is complex (R ± *j* X). A typical example of the variation of the feed-point impedance over the band for a 40-meter double-extended Zepp is shown in **Fig 47**. This antenna is normally fed with open-wire transmission line to an antenna tuner. Other matching arrangements are, of course, possible. A method for transforming the feed-point impedance to 450 Ω and eliminating the minor lobes is given in Chapter 6.

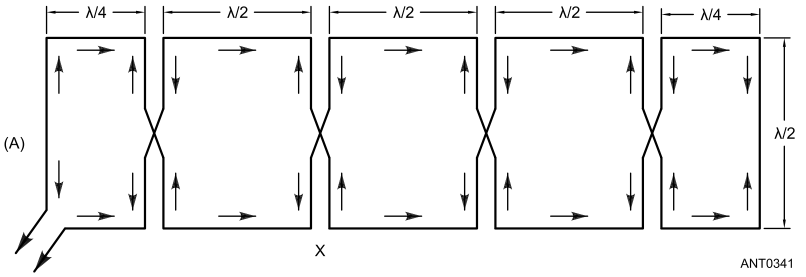

Fig 48—Typical Sterba array, an 8-element version.

THE STERBA ARRAY

Two collinear arrays can be combined to form the *Sterba array*, often called the *Sterba curtain*. An 8-element example of a Sterba array is shown in **Fig 48**. The four λ/4 elements joined on the ends are equivalent to two λ/2 elements. The two collinear arrays are spaced λ/2 and the λ/4 phasing lines connected together to provide λ/2 phasing lines. This arrangement has the advantage of increasing the gain for a given length and also increasing the E-plane directivity, which is no longer circular. An additional advantage of this array is that the wire forms a closed loop. For installations where icing is a problem a low voltage dc or low frequency (50 or 60 Hz) ac current can be passed through the wire to heat it for deicing. The heating current is isolated from RF by decoupling chokes. This is standard practice in commercial installations.

The number of sections in a Sterba array can be extended as far as desired but more than four or five are rarely used because of the slow increase in gain with extra elements, the narrow H-plane directivity and the appearance of multiple sidelobes. When fed at the point indicated the impedance is about 600 Ω. The antenna can also be fed at the point marked X. The impedance at this point will be about 1 kΩ. The gain of the 8-element array in Fig 48 will be between 7 to 8 dB over a single element.

Parallel Broadside Arrays

To obtain broadside directivity with parallel elements the currents in the elements must all be in-phase. At a distant point lying on a line perpendicular to the axis of the array and also perpendicular to the plane containing the elements, the fields from all elements add up in phase. The situation is like that pictured in Fig 1 in this chapter, where four parallel λ/2 dipoles were fed together a broadside array.

Broadside arrays of this type theoretically can have any number of elements. However, practical limitations of construction and available space usually limit the number of broadside parallel elements.

POWER GAIN

The power gain of a parallel-element broadside array depends on the spacing between elements as well as on the number of elements. The way in which the gain of a two-element array varies with spacing is shown in **Fig 49**. The greatest gain is obtained when the spacing is in the vicinity of 0.67 λ.

The theoretical gains of broadside arrays having more than two elements are approximately as follows:

No. of Parallel Elements	*dB Gain with λ/2 Spacing*	*dB Gain with 3λ/4 Spacing*
3	5.7	7.2
4	7.1	8.5
5	8.1	9.4
6	8.9	10.4

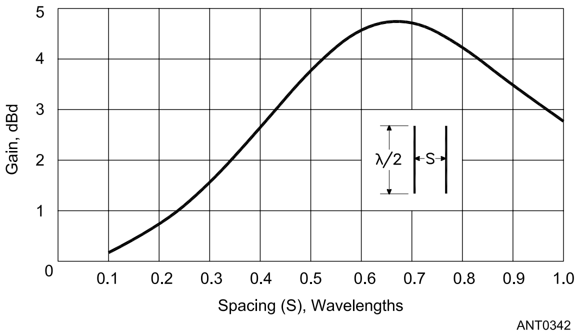

Fig 49—Gain as a function of the spacing between two parallel elements operated in-phase (broadside).

The elements must, of course, all lie in the same plane and all must be fed in-phase.

DIRECTIVITY

The sharpness of the directive pattern depends on spacing between elements and number of elements. Larger element spacing will sharpen the main lobe, for a given number of elements, up to a point as was shown in Fig 41. The two-element array has no minor lobes when the spacing is λ/2, but small minor lobes appear at greater spacings. When three or more elements are used the pattern always has minor lobes.

Other Forms Of Broadside Arrays

For those who have the available room, multi-element arrays based on the broadside concept have something to offer. The antennas are large but of simple design and noncritical dimensions; they are also very economical in terms of gain per unit of cost.

Large arrays can often be fed at several different points. However, the pattern symmetry may be sensitive to the choice of feed point within the array. Non-symmetrical feed points will result in small asymmetries in the pattern but these are not usually of great concern.

Arrays of three and four elements are shown in **Fig 50**. In the 3-element array with λ/2 spacing at A, the array is fed at the center. This is the most desirable point in that it tends to keep the power distribution among the elements uniform. However, the transmission line could alternatively be connected at either point B or C of Fig 50A, with only slight skewing of the radiation pattern.

When the spacing is greater than λ/2, the phasing lines must be 1 λ long and are not transposed between elements. This is shown Fig 50B. With this arrangement, any element spacing up to 1 λ can be used, if the phasing lines can be folded as suggested in the drawing.

The 2-element array at C is fed at the center of the system to make the power distribution among elements as uniform as possible. However, the transmission line could be connected at either point B, C, D or E. In this case the section of phasing line between B and D must be transposed to make the currents flow in the same direction in all elements. The 4-element array at C and the 3-element array at B have approximately the same gain when the element spacing in the array at B is 3λ/4.

An alternative feeding method is shown in Fig 50D. This system can also be applied to the 3-element arrays, and will result in better symmetry in any case. It is necessary only to move the phasing line to the center of each element, making connection to both sides of the line instead of one only.

The free-space pattern for a 4-element array with λ/2 spacing is shown in **Fig 51**. This is also approximately the pattern for a 3-element array with 3λ/4 spacing.

Larger arrays can be designed and constructed by following the phasing principles shown in the drawings. No accurate figures are available for the impedances at the various feed points indicated in Fig 50. You can estimate it to be in the vicinity of 1 kΩ when the feed point is at a junction between the phasing line and a λ/2 element, becoming smaller as the number of elements in the array is increased. When the feed point is midway between end-fed elements as in Fig 50C, the feed-point impedance of a 4-element array is in the vicinity of 200 to 300 Ω, with 600-Ω open-wire phasing lines. The impedance at the feed point with the antenna shown at D should be about 1.5 kΩ.

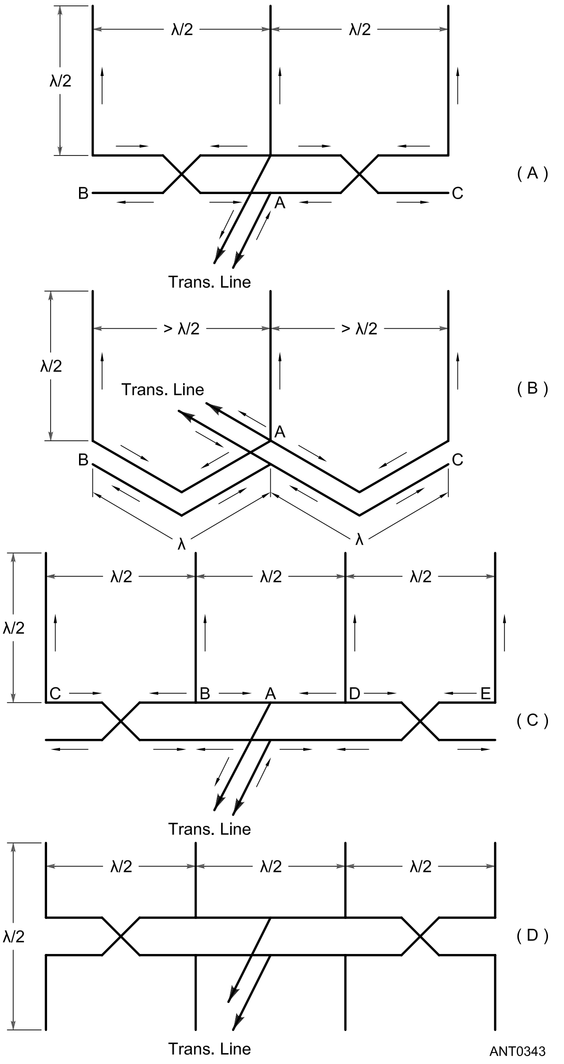

Fig 50—Methods of feeding 3- and 4-element broadside arrays with parallel elements.

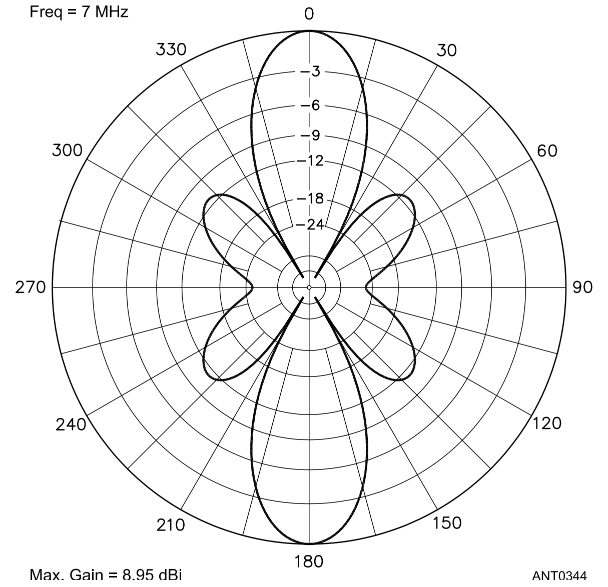

Fig 51—Free-space E-plane pattern of a 4-element broadside array using parallel elements (Fig 50). This corresponds to the horizontal directive pattern at low wave angles for a vertically polarized array over ground. The axis of the elements lies along the 90°-270° line.

NON-UNIFORM ELEMENT CURRENTS

The pattern for a 4-element broadside array shown in Fig 51 has substantial side lobes. This is typical for arrays more than λ/2 wide when equal currents flow in each element. Sidelobe amplitude can be reduced by using non-uniform current distribution among the elements. Many possible current amplitude distributions have been suggested. All of them have reduced current in the outer elements and greater current in the inner elements. This reduces the gain somewhat but can produce a more desirable pattern. One of the common current distributions is called *binomial current grading*. In this scheme the ratio of element currents is set equal to the coefficients of a polynomial. For example:

$$\begin{aligned}
&1\,x+1, \quad \Rightarrow \; 1,1 \\
&(x+1)^2 = 1x^2 + 2x + 1, \quad \Rightarrow \; 1, 2, 1 \\
&(x+1)^3 = 1x^3 + 3x^2 + 3x + 1, \quad \Rightarrow \; 1, 3, 3, 1 \\
&(x+1)^4 = 1x^4 + 4x^3 + 6x^2 + 6x + 1, \quad \Rightarrow \; 1, 4, 6, 4, 1
\end{aligned} \tag{Eq 7}$$

In a 2-element array the currents are equal, in a 3-element array the current in the center element is twice that in the outer elements, and so on.

HALF-SQUARE ANTENNA

On the low-frequency bands (40, 80 and 160 meters) it becomes increasingly difficult to use λ/2 elements because of their size. The half-square antenna is a 2-element broadside array with λ/4-high vertical elements and λ/2 horizontal spacing. See **Fig 52**. The free-space H-plane pattern for this array is shown in **Fig 53**. The antenna gives modest (4.2 dBi) but useful gain and has the advantage of only λ/4 height. Like all vertically polarized antennas, real-world performance depends directly on the characteristics of the ground surrounding it.

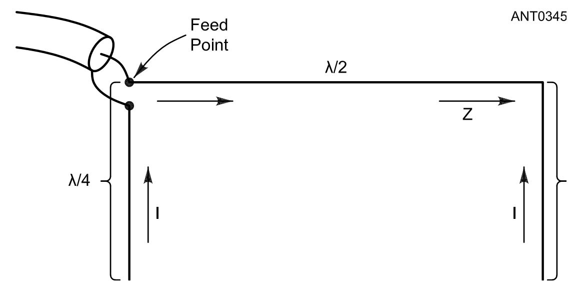

Fig 52—Layout for the half-square antenna.

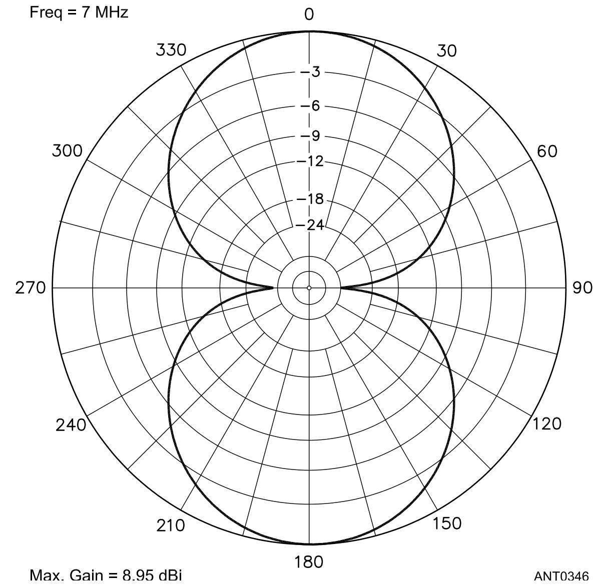

Fig 53—Free-space E-plane directive pattern for the half-square antenna.

The half-square can be fed either at the point indicated or at the bottom end of one of the vertical elements using a voltage-feed scheme, such as that shown in **Fig 54** for the bobtail curtain. The feed-point impedance is in the region of 50 Ω when fed at a corner as shown in Fig 52. A typical SWR plot is shown in **Fig 55**. Chapter 6 has a detailed discussion of the half-square antenna with several variations, together with practical considerations.

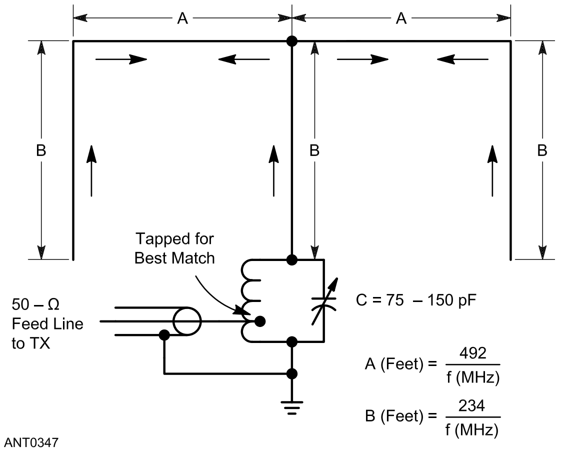

Fig 54—The bobtail curtain is an excellent low-angle radiator having broadside bidirectional characteristics. Current distribution is represented by the arrows. Dimensions A and B (in feet, for wire antennas) can be determined from the equations.

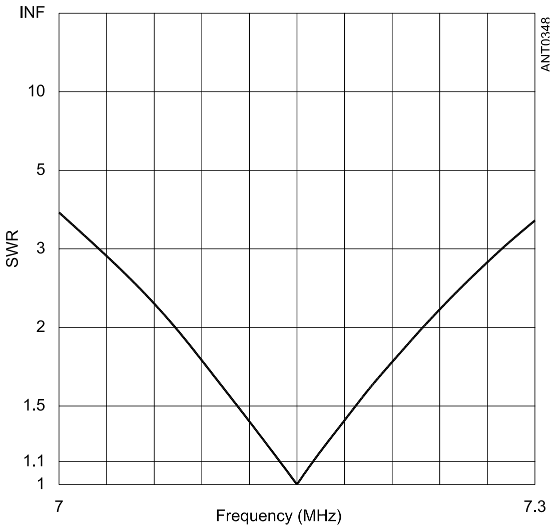

Fig 55—Typical SWR plot for a 40-meter half-square antenna fed at one corner. Antenna in free space.

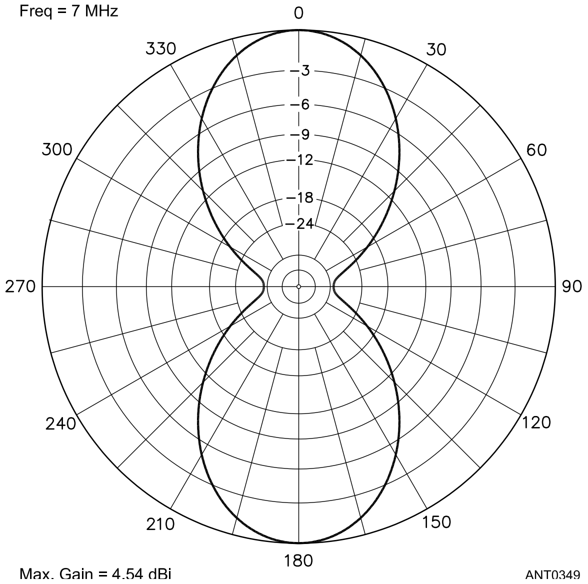

Fig 56—Calculated free-space E-plane directive diagram of the bobtail curtain shown in Fig 54. The array lies along the 90°-270° axis.

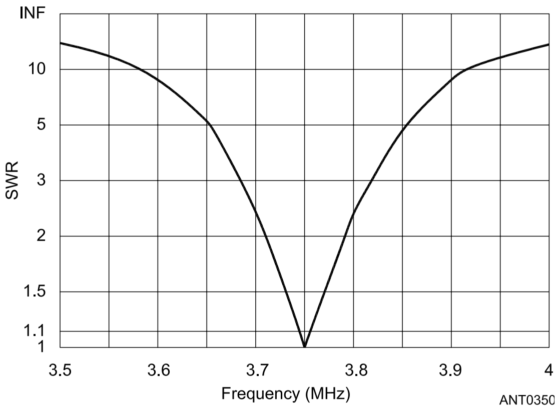

Fig 57—Typical SWR plot for an 80-meter bobtail curtain in free space. This is a narrow-band antenna.

BOBTAIL CURTAIN

The antenna system in Fig 54 uses the principles of co-phased verticals to produce a broadside, bidirectional pattern providing approximately 5.1 dB of gain over a single $\lambda/4$ element. The antenna performs as three in-phase, top-fed vertical radiators approximately $\lambda/4$ in height and spaced approximately $\lambda/2$. It is most effective for low-angle signals and makes an excellent long-distance antenna for 1.8, 3.5 or 7 MHz.

The three vertical sections are the actual radiating components, but only the center element is fed directly. The two horizontal parts, A, act as phasing lines and contribute very little to the radiation pattern. Because the current in the center element must be divided between the end sections, the current distribution approaches a binomial 1:2:1 ratio. The radiation pattern is shown in **Fig 56**.

The vertical elements should be as vertical as possible. The height for the horizontal portion should be slightly greater than B, as shown in Fig 54. The tuning network is resonant at the operating frequency. The L/C ratio should be fairly low to provide good loading characteristics. As a starting point, a maximum capacitor value of 75 to 150 pF is recommended, and the inductor value is determined by C and the operating frequency. The network is first tuned to resonance and then the tap point is adjusted for the best match. A slight readjustment of C may be necessary. A link coil consisting of a few turns can also be used to feed the antenna.

A feeling for the matching bandwidth of this antenna can be obtained by looking at a feed point located at the top end of the center element. The impedance at this point will be approximately 32 Ω. An SWR plot (for Z_0 = 32 Ω) for an 80-meter bobtail curtain at this feed-point is shown in **Fig 57**. However, it is not advisable to actually connect a feed line at this point since it would detune the array and alter the pattern. This antenna is relatively narrow band. When fed at the bottom of the center element as shown in Fig 54, the SWR can be adjusted to be 1:1 at one frequency but the operating bandwidth for SWR < 2:1 may be even narrower than Fig 57 shows. For 80-meters, where operation is often desired in the CW DX window (3.510 MHz) and in the phone DX window (3.790 MHz), it will be necessary to retune the matching network as you change frequency. This can be done by switching a capacitor in or out, manually or remotely with a relay.

While the match bandwidth is quite narrow, the radiation pattern changes more slowly with frequency. **Fig 58** shows the variation in the pattern over the entire band (3.5 to 4.0 MHz). As would be expected, the gain increases with frequency because the antenna is larger in terms of wavelengths. The general shape of the pattern, however, is quite stable.

THE BRUCE ARRAY

Four variations of the Bruce array are shown in **Fig 59**. The Bruce is simply a wire folded so that the vertical sections carry large in-phase currents, while the horizontal sections carry small currents flowing in opposite directions with respect to the center of a section (indicated by dots). The radiation is vertically polarized. The gain is proportional to the length of the array but is somewhat smaller than you can obtain from a broadside array of λ/2 elements of the same length. This is because the radiating portion of the elements is only λ/4.

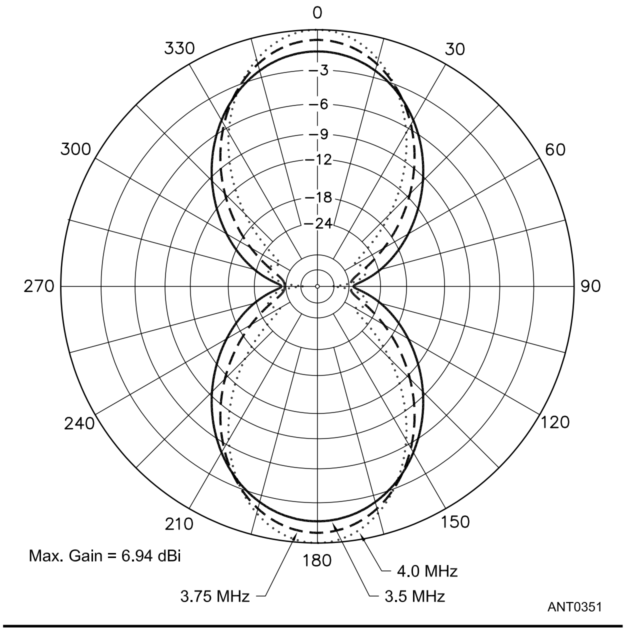

Fig 58—80-meter bobtail curtain's free-space E-plane pattern variation over the 80-meter band.

The Bruce array has a number of advantages:

1) The array is only λ/4 high. This is especially helpful on 80 and 160 meters, where the height of λ/2 supports becomes impractical for most amateurs.

2) The array is very simple. It is just a single piece of wire folded to form the array.

3) The dimensions of the array are very flexible. Depending on the available distance between supports, any number of elements can be used. The longer the array, the greater the gain.

4) The shape of the array does not have to be exactly 1.05 λ/4 squares. If the available height is short but the array can be made longer, then shorter vertical sections and longer horizontal sections can be used to maintain gain and resonance. Conversely, if more height is available but width is restricted then longer vertical sections can be used with shorter horizontal sections.

5) The array can be fed at other points more convenient for a particular installation.

6) The antenna is relatively low Q, so that the feed-point impedance changes slowly with frequency. This is very helpful on 80 meters, for example, where the antenna can be relatively broadband.

7) The radiation pattern and gain is stable over the width of an amateur band.

Note that the nominal dimensions of the array in Fig 59 call for section lengths = 1.05 λ/4. The need to use slightly longer elements to achieve resonance is common in large wire arrays. A quad loop behaves in the same manner. This is quite different from wire dipoles, which are typically *shortened* by 2-5% to achieve resonance.

Fig 60 shows the variations in gain and pattern for 2 to 5-element 80-meter Bruce arrays. **Table 2** lists the gain over a vertical λ/2 dipole, a 4-radial ground-plane vertical and the size of the array. The gain and impedance parameters listed are for free space. Over real ground the patterns and gain will depend on the height above ground and the ground characteristics. Copper loss using #12 conductors in included.

Worthwhile gain can be obtained from these arrays, especially on 80 and 160 meters, where any gain is hard to come by. The feed-point impedance is for the center of a vertical section. From the patterns in Fig 60 you can see that sidelobes start to appear as the length of the array is increased beyond 3λ/4. This is typical for arrays using equal currents in the elements.

It is interesting to compare the bobtail curtain (Fig 54) with a 4-element Bruce array. **Fig 61** compares the radiation patterns for these two antennas. Even though the Bruce is shorter (3λ/4) than the bobtail (1 λ), it has slightly more

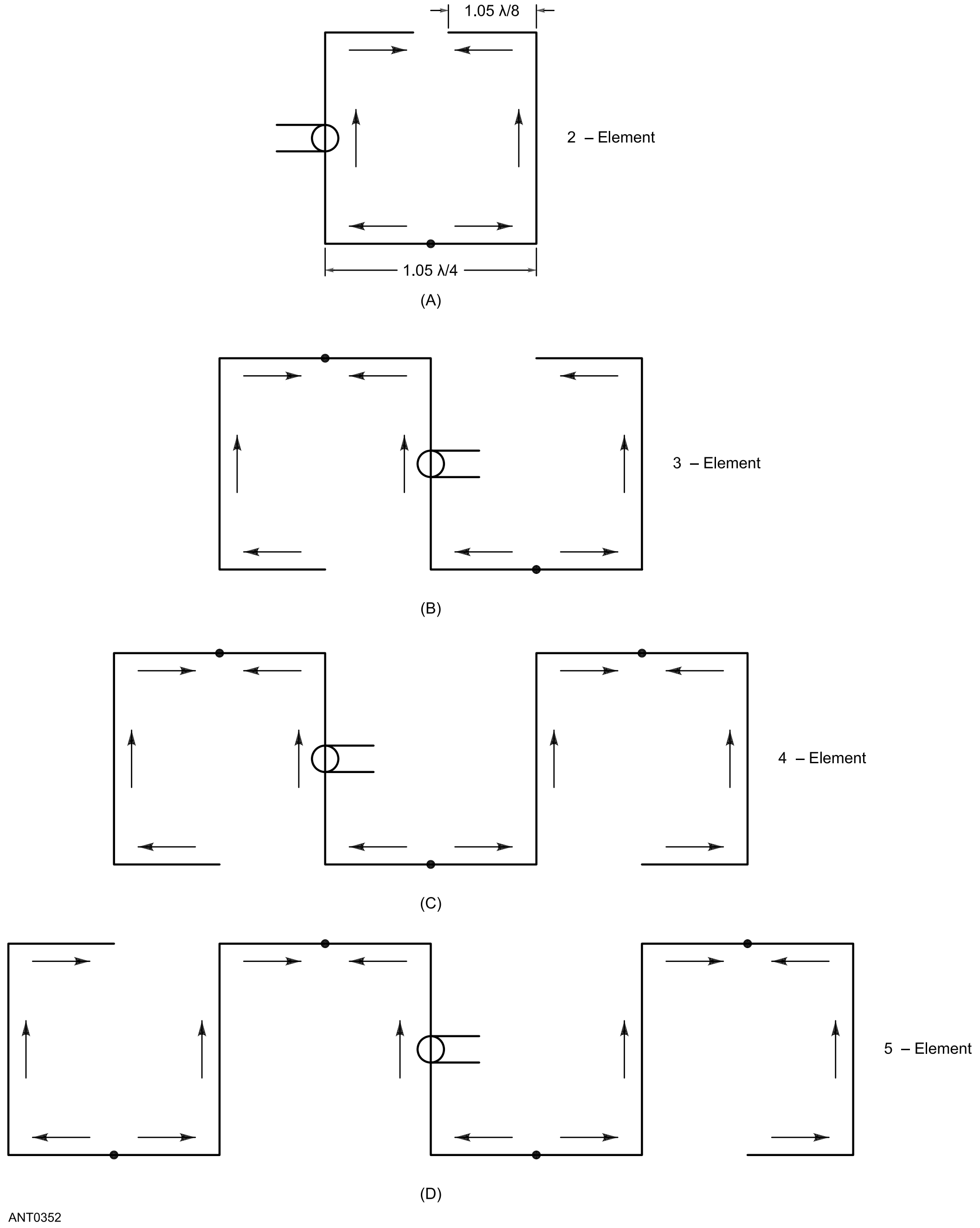
1.05 λ/8
2 – Element
1.05 λ/4
(A)
3 – Element
(B)
4 – Element
(C)
5 – Element
(D)
ANT0352

Fig 59—Various Bruce arrays: 2, 3, 4 and 5-element versions.

Table 2
Bruce array length, impedance and gain as a function of number of elements

Number Elements	*Gain Over λ/2 Vertical Dipole*	*Gain over λ/4 Ground-Plane*	*Array Length Wavelengths*	*Approx. Feed Z, Ω*
2	1.2 dB	1.9 dB	¼	130
3	2.8 dB	3.6 dB	½	200
4	4.3 dB	5.1 dB	¾	250
5	5.3 dB	6.1 dB	1	300

gain. The matching bandwidth is illustrated by the SWR curve in **Fig 62**. The 4-element Bruce has over twice the match bandwidth (200 kHz) than does the bobtail (75 kHz in Fig 57). Part of the gain difference is due to the binomial current distribution—the center element has twice the current as the outer elements in the bobtail. This reduces the gain slightly so that the 4-element Bruce becomes competitive. This is a good example of using more than the minimum number of elements to improve performance or to reduce size. On 160 meters the 4-element Bruce will be 140 feet shorter than the bobtail, a significant reduction. If additional space is available for the bobtail (1 λ) then a 5-element Bruce could

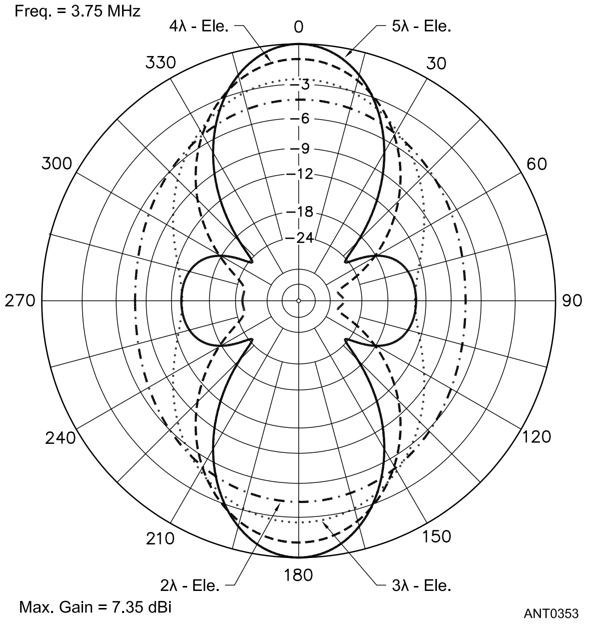

Fig 60—80-meter free-space E-plane directive patterns for the Bruce arrays shown in Fig 59. The 5-element's pattern is a solid line; the 4-element is a dashed line; the 3-element is a dotted line, and the 2-element version is a dashed-dotted line.

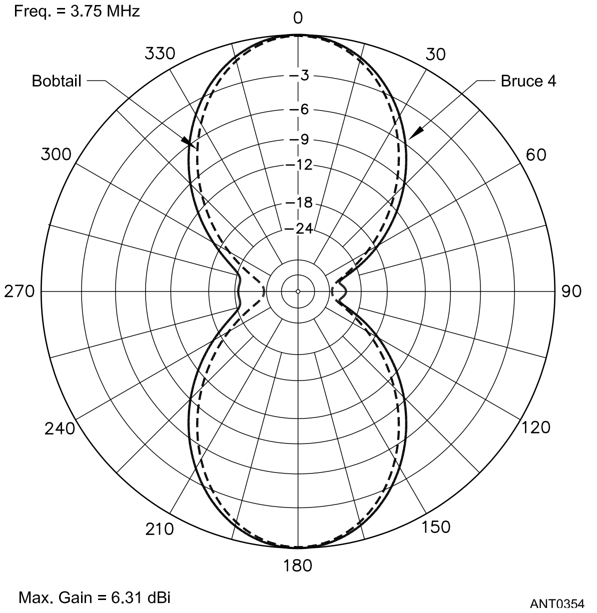

Fig 61—Comparison of free space patterns of a 4-element Bruce array (solid line) and a 3-element bobtail curtain (dashed line).

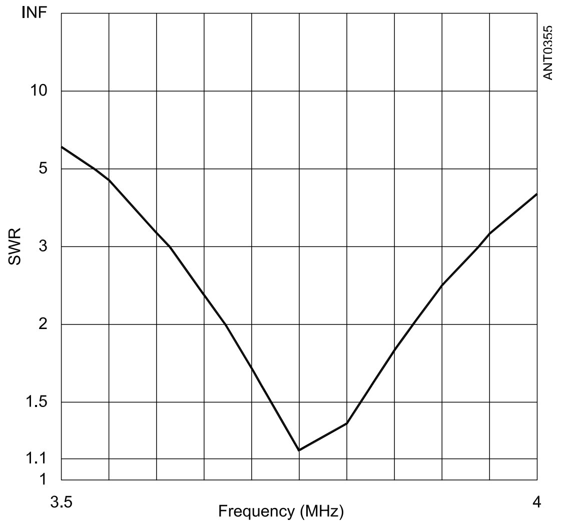

Fig 62—Typical SWR curve for a 4-element 80-meter Bruce array.

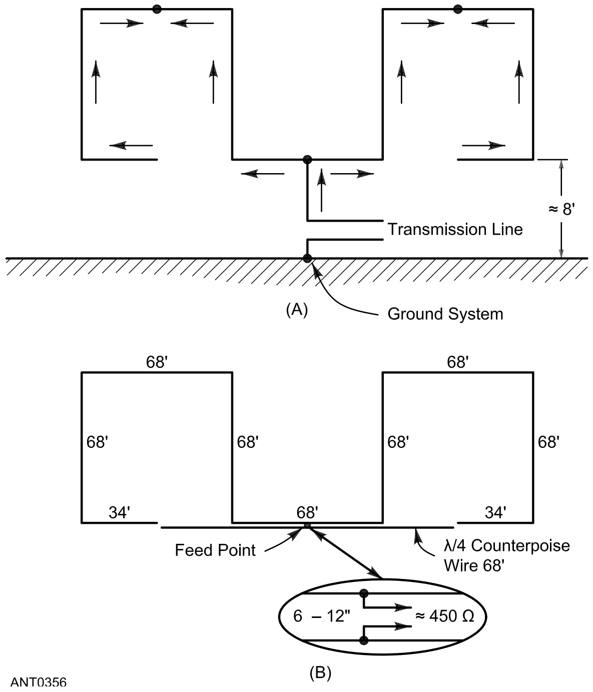

Fig 63—Alternate feed arrangements for the Bruce array. At A, the antenna is driven against a ground system and at B, it uses a two-wire counterpoise.

be used, with a small increase in gain but also introducing some sidelobes.

The 2-element Bruce and the half-square antennas are both 2-element arrays. However, since the spacing between radiators is greater in the half-square (λ/2) the gain of the half-square is about 1 dB greater. If space is available, the half-square would be a better choice. If there is not room for a half-square then the Bruce, which is only half as long (λ/4), may be a good alternative. The 3-element Bruce, which has the same length (λ/2) as the half-square, has about 0.6 dB more gain than the half-square and will have a wider match bandwidth.

The Bruce antenna can be fed at many different points and in different ways. In addition to the feed points indicated in Fig 59, you may connect the feed line at the center of any of the vertical sections. In longer Bruce arrays, feeding at one end will result in some current imbalance among the elements but the resulting pattern distortion is small. Actually, the feed-point can be anywhere along a vertical section. One very convenient point is at an outside corner. The feed-point impedance will be higher (about 600 Ω). A good match for 450-Ω ladder-line can usually be found somewhere on the vertical section. It is important to recognize that feeding the antenna at a voltage node (dots in Fig 59) by breaking the wire and inserting an insulator, completely changes the current distribution. This will be discussed in the section on endfire arrays.

A Bruce can be fed unbalanced against ground or against a counterpoise as shown in **Fig 63**. Because it is a vertically polarized antenna, the better the ground system, the better the performance. As few as two elevated radials can be used as shown in Fig 63B, but more radials can also be used to improve the performance, depending on local ground constants. The original development of the Bruce array in the late 1920s used this feed arrangement.

FOUR-ELEMENT BROADSIDE ARRAY

The 4-element array shown in **Fig 64** is commonly known as the *lazy H*. It consists of a set of two collinear elements and a set of two parallel elements, all operated in-phase to give broadside directivity. The gain and directivity will depend on the spacing, as in the case of a simple parallel-element broadside array. The spacing may be chosen between the limits shown on the drawing, but spacings below 3λ/8 are not worthwhile because the gain is small. Estimated gains compared to a single element are:

3λ/8 spacing—4.2 dB
λ/2 spacing—5.8 dB
5λ/8 spacing—6.7 dB
3λ/4 spacing—6.3 dB

Half-wave spacing is generally used. Directive patterns for this spacing are given in **Figs 65** and **66**. With λ/2 spacing between parallel elements, the impedance at the junction of the phasing line and transmission line is resistive and in the vicinity of 100 Ω. With larger or smaller spacing the impedance at this junction will be reactive as well as resistive. Matching stubs are recommended in cases where a non-resonant line is to be used. They may be calculated and adjusted as described in Chapter 26.

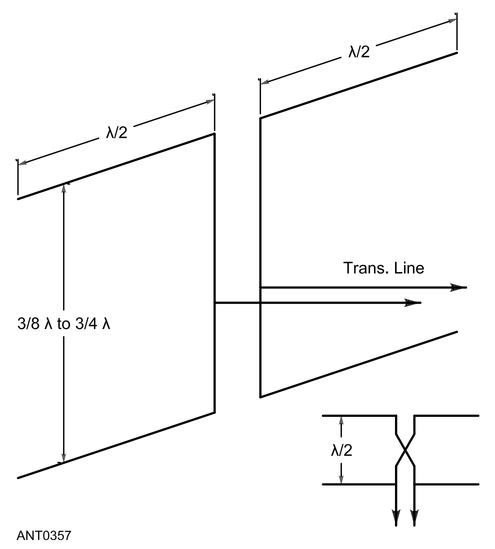

Fig 64—Four-element broadside array ("lazy H") using collinear and parallel elements.

The system shown in Fig 64 may be used on two bands having a 2-to-1 frequency relationship. It should be designed for the higher of the two frequencies, using 3λ/4 spacing between parallel elements. It will then operate on the lower frequency as a simple broadside array with 3λ/8 spacing.

An alternative method of feeding is shown in the small diagram in Fig 64. In this case the elements and the phasing line must be adjusted exactly to an electrical half wavelength. The impedance at the feed point will be resistive and on the order of 2 kΩ.

THE BI-SQUARE ANTENNA

A development of the lazy H, known as the *bi-square antenna*, is shown in **Fig 67**. The gain of the bi-square is somewhat less than that of the lazy-H, but this array is attractive because it can be supported from a single pole. It has a circumference of 2 λ at the operating frequency, and is horizontally polarized.

The bi-square antenna consists of two 1 λ radiators, fed 180° out-of-phase at the bottom of the array. The radiation resistance is 300 Ω, so it can be fed with either 300- or 600-Ω line. The free space gain of the antenna is about 5.8 dBi, which is 3.7 dB more than a single dipole element. Gain may be increased by adding a parasitic reflector or director. Two bi-square arrays can be mounted at right angles and switched to provide omnidirectional coverage. In this way, the antenna wires may be used as part of the guying system for the pole.

Although it resembles a loop antenna, the bi-square is not a true loop because the ends opposite the feed point

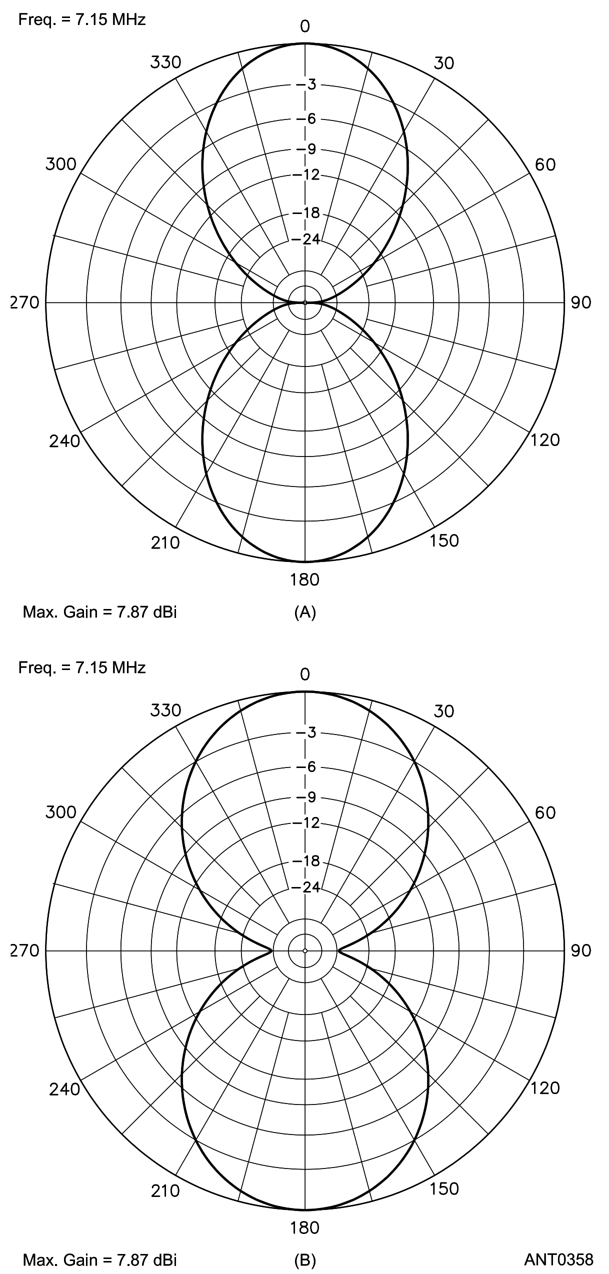

Fig 65—Free-space directive diagrams of the 4-element antenna shown in Fig 64. At A is the E-plane pattern. The axis of the elements lies along the 90°-270° line. At B is the free-space H-plane pattern, viewed as if one set of elements is above the other from the ends of the elements.

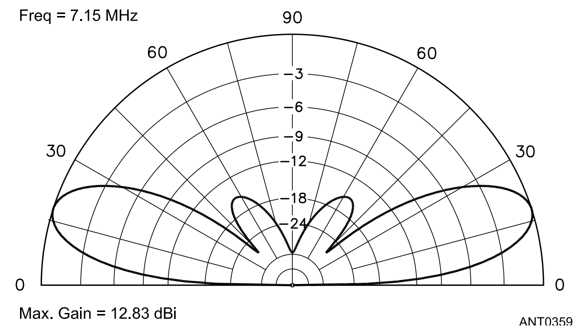

Fig 66—Vertical pattern of the 4-element broadside antenna of Fig 64, when mounted with the elements horizontal and the lower set λ/4 above flat ground. Stacked arrays of this type give best results when the lowest elements are at least λ/2 high. The gain is reduced and the wave angle raised if the lowest elements are too close to ground.

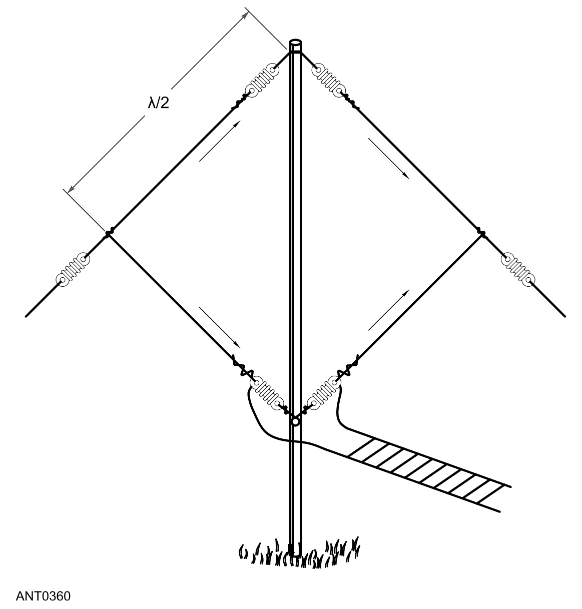

Fig 67—The bi-square array. It has the appearance of a loop, but is not a true loop because the conductor is open at the top. The length of each side, in feet, is 480/f (MHz).

are open. However, identical construction techniques can be used for the two antenna types. Indeed, with a means of remotely closing the connection at the top for lower frequency operation, the antenna can be operated on two harmonically related bands. As an example, an array with 17 feet per side can be operated as a bi-square at 28 MHz and as a full-wave loop at 14 MHz. For two-band operation in this manner, the side length should favor the higher frequency. The length of a closed loop is not as critical.

End-Fire Arrays

The term *end-fire* covers a number of different methods of operation, all having in common the fact that the maximum radiation takes place along the array axis, and that the array consists of a number of parallel elements in one plane. End-fire arrays can be either bidirectional or unidirectional. In the bidirectional type commonly used by amateurs there are only two elements, and these are operated with currents 180° out-of-phase. Even though adjustment tends to be complicated, unidirectional end-fire driven arrays have also seen amateur use, primarily as a pair of phased, ground-mounted λ/4 vertical elements. Extensive discussion of this array is contained in earlier sections of this chapter.

Horizontally polarized unidirectional end-fire arrays see little amateur use except in log-periodic arrays (described in Chapter 10). Instead, horizontally polarized unidirectional arrays usually have parasitic elements (described in Chapter 11) and are called Yagis.

TWO-ELEMENT END-FIRE ARRAY

In a 2-element array with equal currents out-of-phase, the gain varies with the spacing between elements as shown in **Fig 68**. The maximum gain occurs in the neighborhood of 0.1 λ spacing. Below that the gain drops rapidly due to conductor loss resistance.

The feed-point resistance for either element is very low at the spacings giving greatest gain, as shown in Fig 8 earlier in this chapter. The spacings most frequently used are λ/8 and λ/4, at which the resistances of center-fed λ/2 elements are about 9 and 32 Ω, respectively.

The effect of conductor resistance on gain for various spacings is shown in **Fig 69**. Because current along the element is not constant (it is approximately sinusoidal), the resistance shown is the equivalent resistance (R_{eq}) inserted at the center of the element to account for the loss distributed along the element.

The equivalent resistance of a λ/2 element is one half the ac resistance (R_{ac}) of the complete element. R_{ac} is usually >> R_{dc} due to skin effect. For example, a 1.84 MHz dipole using #12 copper wire will have the following R_{eq}:

Wire length = 267 feet
Rdc = 0.00159 [/foot] × 267 [feet] = 0.42
Fr = Rac/Rdc = 10.8
Req =(Rdc/2) × Fr = 2.29

For a 3.75 MHz dipole made with #12 wire, Req = 1.59 Ω. In Fig 69, it is clear that end-fire antennas made with #12 or smaller wire will limit the attainable gain because of losses. There is no point in using spacings much less than λ/4 if you use wire elements. If instead you use elements made of aluminum tubing then smaller spacings can be used to increase gain. However, as the spacing is reduced below λ/4 the increase in gain is quite small even with good conductors. Closer spacings give little gain increase but can drastically reduce the operating bandwidth due to the rapidly increasing Q of the array.

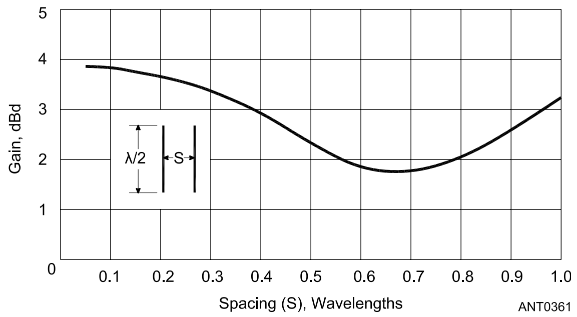

Fig 68—Gain of an end-fire array consisting of two elements fed 180° out-of-phase, as a function of the spacing between elements. Maximum radiation is in the plane of the elements and at right angles to them at spacings up to λ/2, but the direction changes at greater spacings.

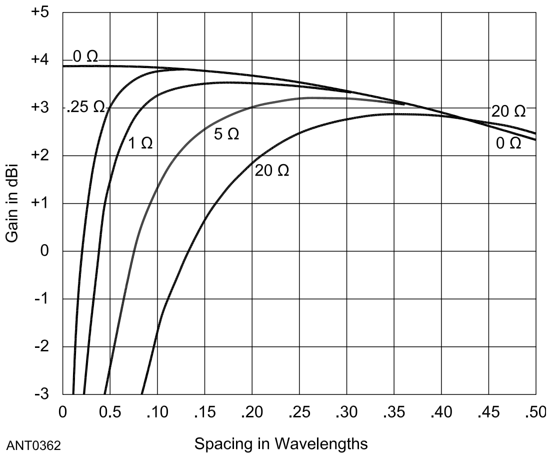

Fig 69—Gain over a single element of two out-of-phase elements in free space as a function of spacing for various loss resistances.

Unidirectional End-Fire Arrays

Two parallel elements spaced λ/4 apart and fed equal currents 90° out-of-phase will have a directional pattern in the plane at right angles to the plane of the array. See **Fig 70**. The maximum radiation is in the direction of the element in which the current lags. In the opposite direction the fields from the two elements cancel.

When the currents in the elements are neither in-phase nor 180° out-of-phase, the feed-point resistances of the elements are not equal. This complicates the problem of feeding equal currents to the elements, as discussed in earlier sections.

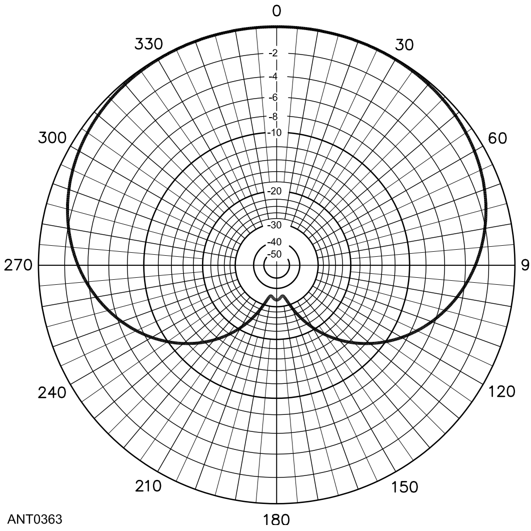

Fig 70—Representative H-plane pattern for a 2-element end-fire array with 90° spacing and phasing. The elements lie along the vertical axis, with the uppermost element the one of lagging phase. Dissimilar current distributions are taken into account. (Pattern computed with *ELNEC*.)

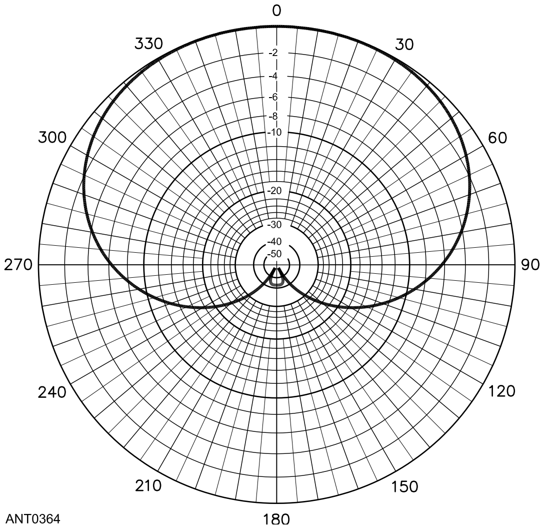

Fig 71—H-plane pattern for a 3-element end-fire array with binomial current distribution (the current in the center element is twice that in each end element). The elements are spaced λ/4 apart along the 0°-180° axis. The center element lags the lower element by 90°, while the upper element lags the lower element by 180° in phase. Dissimilar current distributions are taken into account. (Pattern computed with *ELNEC*.)

More than two elements can be used in a unidirectional end-fire array. The requirement for unidirectivity is that there must be a progressive phase shift in the element currents equal to the spacing, in electrical degrees, between the elements. The amplitudes of the currents in the various elements also must be properly related. This requires binomial current distribution. In the case of three elements, this requires that the current in the center element be twice that in the two outside elements, for 90° (λ/4) spacing and element current phasing. This antenna has an overall length of λ/2. The directive diagram is shown in **Fig 71**. The pattern is similar to that of Fig 70, but the 3-element binomial array has greater directivity, evidenced by the narrower half-power beamwidth (146° versus 176°). Its gain is 1.0 dB greater.

THE W8JK ARRAY

As pointed out earlier, John Kraus, W8JK, described his bidirectional flat-top W8JK beam antenna in 1940. See **Fig 72**. Two λ/2 elements are spaced λ/8 to λ/4 and driven 180° out-of-phase. The free-space radiation pattern for this antenna, using #12 copper wire, is given in **Fig 73**. The pattern is representative of spacings between λ/8 and λ/4 where the gain varies less than 0.5 dB. The gain over a dipole is about 3.3 dB (5.4 dBi referenced to an isotropic radiator), a worthwhile improvement. The feed-point impedance (including wire resistance) of *each* element is about 11 Ω for λ/8 spacing and 33 Ω for λ/4 spacing. The feed-point impedance at the center connection will depend on the length and Z_0 of the connecting transmission line.

Kraus gave a number of other variations for end-fire arrays, some of which are shown in **Fig 74**. The ones fed at the center (A, C and E) are usually horizontally polarized flat-top beams. The end-fed versions (B, D & F) are usually vertically polarized, where the feed point can be conveniently near ground.

A practical variation of Fig 74B is given in **Fig 75**. In this example, the height is limited to λ/4 so the ends can be bent over as shown, producing a 2-element Bruce array. This reduces the gain somewhat but allows much shorter supports, an important consideration on the low bands. If additional height is available, then you can achieve some additional gain. The upper ends can be bent over to fit the available height. The feed-point impedance will greater than 1 kΩ.

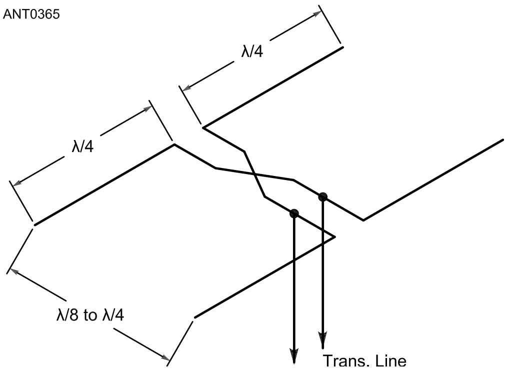

Fig 72—A 2-element W8JK array.

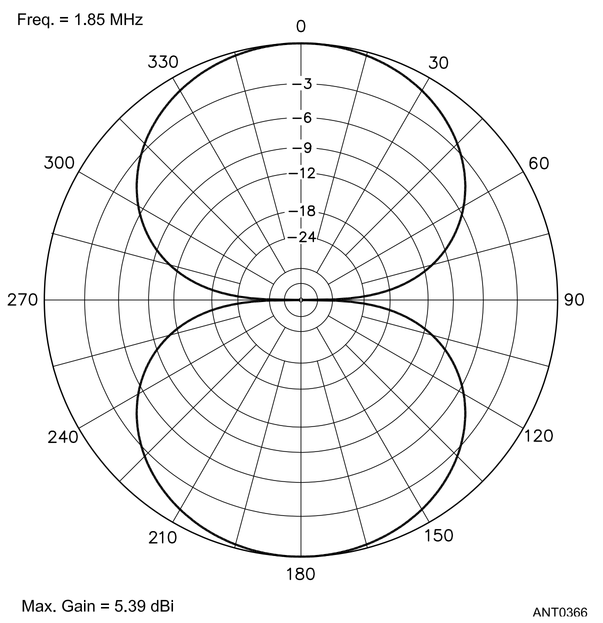

Fig 73—Free-space E-plane pattern for the 2-element W8JK array

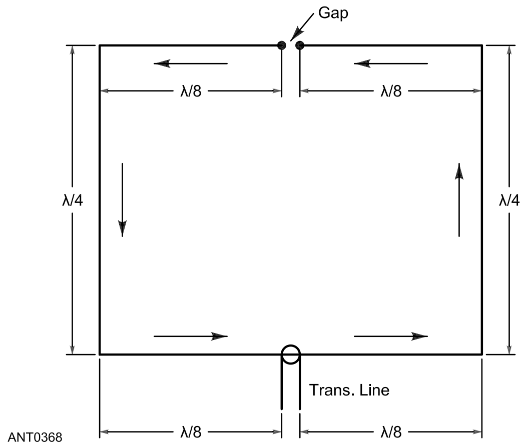
Gap
λ/8
λ/8
λ/4
λ/4
Trans. Line
ANT0368
λ/8
λ/8

Fig 75—A 2-element end-fire array with reduced height.

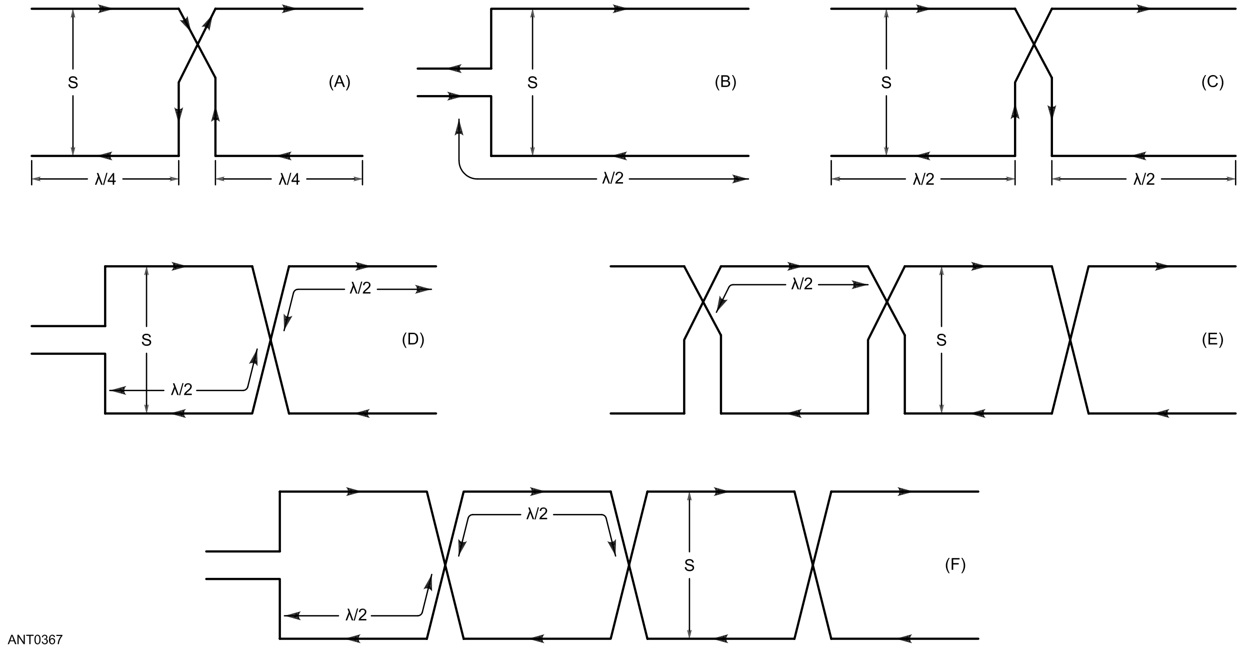
S
(A)
λ/4
λ/4
S
(B)
λ/2
S
(C)
λ/2
λ/2
λ/2
S
(D)
λ/2
λ/2
S
(E)
λ/2
S
(F)
λ/2
ANT0367

Fig 74—Six other variations of W8JK "flat-top beam" antennas.

FOUR-ELEMENT END-FIRE AND COLLINEAR ARRAYS

The array shown in **Fig 76** combines collinear in-phase elements with parallel out-of-phase elements to give both broadside and end-fire directivity. It is a *two-section W8JK*. The approximate free-space gain using #12 copper wire is 4.9 dBi with λ/8 spacing and 5.4 dBi with λ/4 spacing. Directive patterns are given in **Figs 77** for free space, and in **Fig 78** for heights of 1 λ and λ/2 above flat ground.

The impedance between elements at the point where the phasing line is connected is of the order of several thousand ohms. The SWR with an unmatched line consequently is quite high, and this system should be constructed with open-wire line (500 or 600 Ω) if the line is to be resonant. With λ/4 element spacing the SWR on a 600 Ω line is estimated to be in the vicinity of 3 or 4:1.

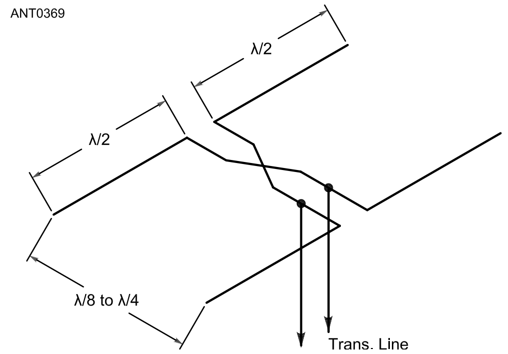

Fig 76—A 4-element array combining collinear broadside elements and parallel end-fire elements, popularly known as a two-section W8JK array.

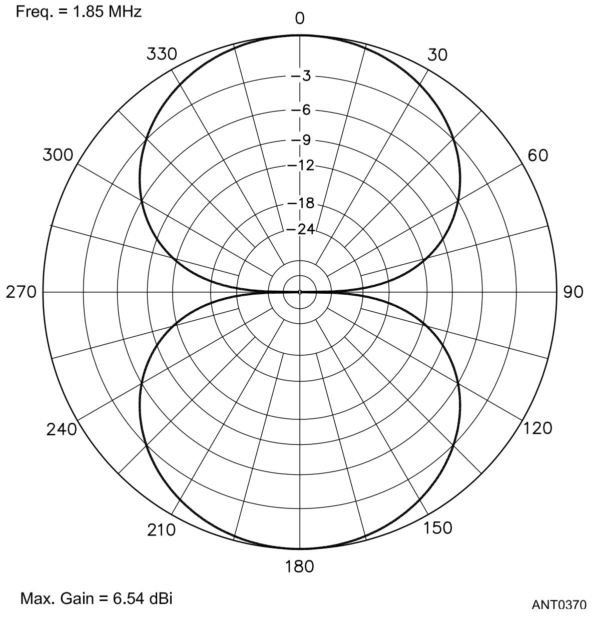

Fig 77—Free-space E-plane pattern for the antenna shown in Fig 76, with λ/8 spacing. The elements are parallel to the 90°-270° line in this diagram. Less than a 1° change in half-power beamwidth results when the spacing is changed from λ/8 to λ/4.

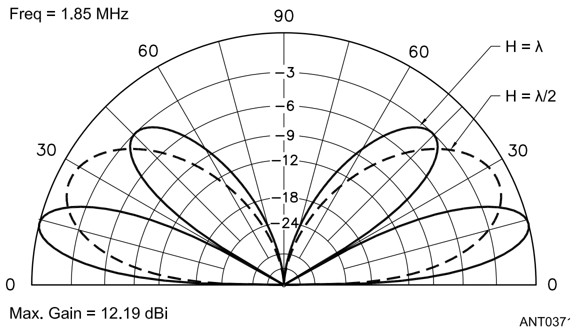

Fig 78—Elevation-plane pattern for the 4-element antenna of Fig 76 when mounted horizontally at two heights over flat ground. Solid line = 1 λ high; dashed line = λ/2 high.

To use a matched line, you could connect a closed stub 3λ/16 long at the transmission-line junction shown in Fig 76. The transmission line itself can then be tapped on this matching section at the point resulting in the lowest line SWR. This point can be determined by trial.

This type of antenna can be operated on two bands having a frequency ratio of 2 to 1, if a resonant feed line is used. For example, if you design for 28 MHz with λ/4 spacing between elements, you can also operate on 14 MHz as a simple 2-element end-fire array having λ/8 spacing.

Combination Driven Arrays

You can readily combine broadside, end-fire and collinear elements to increase gain and directivity, and this is in fact usually done when more than two elements are used in an array. Combinations of this type give more gain, in a given amount of space, than plain arrays of the types just described. Since the combinations that can be worked out are almost endless, this section describes only a few of the simpler types.

The accurate calculation of the power gain of a multielement array requires a knowledge of the mutual impedances between all elements, as discussed in earlier sections. For approximate purposes it is sufficient to assume that each set (collinear, broadside, end-fire) will have the gains as given earlier, and then simply add up the gains for the combination. This neglects the effects of cross-coupling between sets of elements. However, the array configurations are such that the mutual impedances from cross-coupling should be relatively small, particularly when the spacings are λ/4 or more, so the estimated gain should be reasonably close to the actual gain. Alternatively, an antenna modeling program, such as *EZNEC*, can give good estimates of all parameters for a real-world antenna, providing that you take care to model all applicable parameters.

FOUR-ELEMENT DRIVEN ARRAYS

The array shown in **Fig 79** combines parallel elements with broadside and end-fire directivity. The smallest array (physically)—3λ/8 spacing between broadside and λ/8 spacing between end-fire elements—has an estimated gain of 6.5 dBi and the largest—3λ/4 and λ/4 spacing, respectively—about 8.4 dBi. Typical directive patterns for a λ/4 × λ/2 array are given in **Figs 80** and **81**.

The impedance at the feed point will not be purely resistive unless the element lengths are correct and the phasing lines are exactly λ/2 long. (This requires somewhat less than λ/2 spacing between broadside elements.) In this case the impedance at the junction is estimated to be over 10 kΩ. With other element spacings the impedance at the junction will be reactive as well as resistive, but in any event the SWR will be quite large. An open-wire line can be used as a resonant line, or a matching section may be used for non-resonant operation.

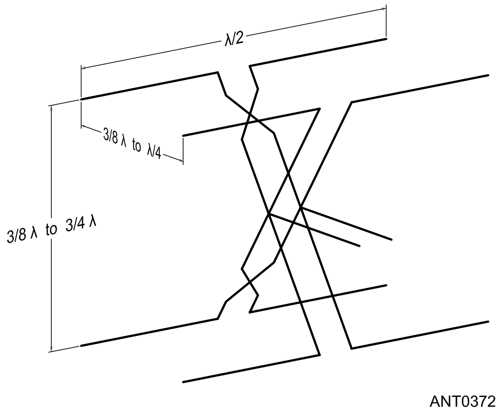

Fig 79—Four-element array combining both broadside and end-fire elements.

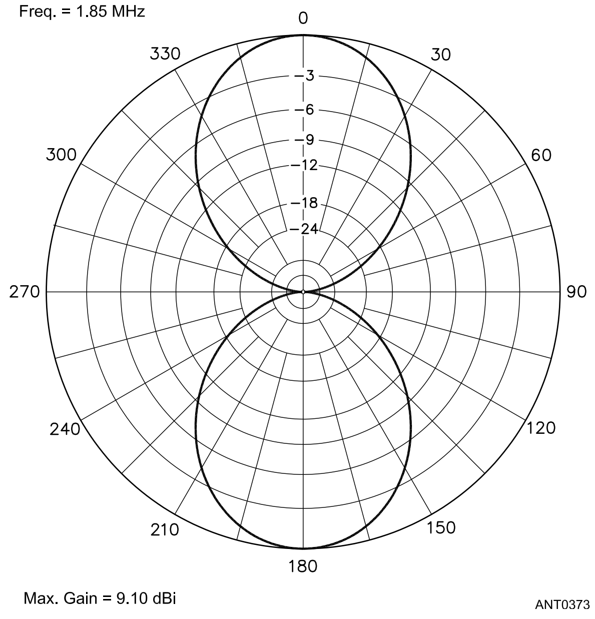

Fig 80—Free-space H-plane pattern of the 4-element antenna shown in Fig 79.

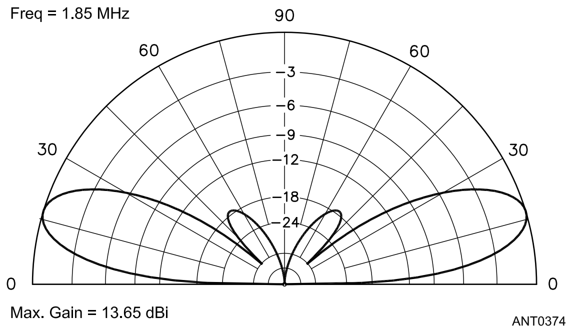

Fig 81—Vertical pattern of the antenna shown in Fig 79 at a mean height of 3λ/4 (lowest elements λ/2 above flat ground) when the antenna is horizontally polarized. For optimum gain and low wave angle the mean height should

EIGHT-ELEMENT DRIVEN ARRAYS

The array shown in **Fig 82** is a combination of collinear and parallel elements in broadside and end-fire directivity. Common practice in a wire antenna is to use λ/2 spacing for the parallel broadside elements and λ/4 spacing for the end-fire elements. This gives a free-space gain of about 9.1 dBi. Directive patterns for an array using these spacings are similar to those of Figs 80 and 81, but are somewhat sharper.

The SWR with this arrangement will be high. Matching stubs are recommended for making the lines non-resonant. Their position and length can be determined as described in Chapter 26.

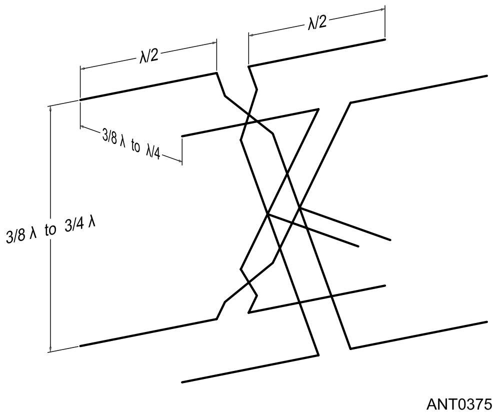

Fig 82—Eight-element driven array combining collinear and parallel elements for broadside and end-fire directivity.

This system can be used on two bands related in frequency by a 2-to-1 ratio, providing it is designed for the higher of the two, with 3λ/4 spacing between the parallel broadside elements and λ/4 spacing between the end-fire elements. On the lower frequency it will then operate as a 4-element antenna of the type shown in Fig 79, with 3λ/8 broadside spacing and λ/8 end-fire spacing. For two-band operation a resonant transmission line must be used.

PHASING ARROWS IN ARRAY ELEMENTS

In the antenna diagrams of preceding sections, the relative direction of current flow in the various antenna elements and connecting lines was shown by arrows. In laying out any antenna system it is necessary to know that the phasing lines are properly connected; otherwise the antenna may have entirely different characteristics than anticipated. The phasing may be checked either on the basis of current direction or polarity of voltages. There are two rules to remember:

1) In every λ/2 section of wire, starting from an open end, the current directions reverse. In terms of voltage, the polarity reverses at each λ/2 point, starting from an open end.
2) Currents in transmission lines always must flow in opposite directions in adjacent wires. In terms of voltage, polarities always must be opposite.

Examples of the use of current direction and voltage polarity are given at A and B, respectively, in **Fig 83**. The λ/2 points in the system are marked by small circles. When current in one section flows toward a circle, the current in the next section must also flow toward it, and vice versa. In the 4-element antenna shown at A, the current in the upper right-hand element cannot flow toward the transmission line because then the current in the right-hand section of the phasing line would have to flow upward and thus would be flowing in the same direction as the current in the left-hand wire. The phasing line would simply act like two wires in parallel in such a case. Of course, all arrows in the drawing could be reversed, and the net effect would be unchanged.

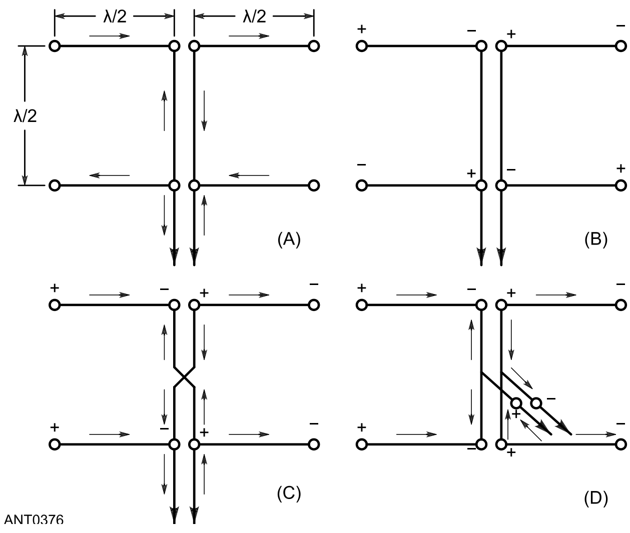

Fig 83—Methods of checking the phase of currents in elements and phasing lines.

C shows the effect of transposing the phasing line. This transposition reverses the direction of current flow in the lower pair of elements, as compared with A, and thus changes the array from a combination collinear and end-fire arrangement into a collinear-broadside array.

The drawing at D shows what happens when the transmission line is connected at the center of a section of phasing line. Viewed from the main transmission line, the two parts of the phasing line are simply in parallel, so the half wavelength is measured from the antenna element along the upper section of phasing line and thence along the transmission line. The distance from the lower elements is measured in the same way. Obviously, the two sections of phasing line should be the same length. If they are not, the current distribution becomes quite complicated; the element currents are neither in-phase nor 180° out-of-phase, and the elements at opposite ends of the lines do not receive the same current. To change the element current phasing at D into the phasing at A, simply transpose the wires in one section of the phasing line. This reverses the direction of current flow in the antenna elements connected to that section of phasing line.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

D. W. Atchley, H. E. Stinehelfer, and J. F. White, “360°-Steerable Vertical Phased Arrays,” *QST*, Apr 1976, pp 27-30.

G. H. Brown, “Directional Antennas,” *Proc. IRE*, Vol 25, No. 1, Jan 1937, pp 78-145.

G. H. Brown, R. F. Lewis and J. Epstein, “Ground Systems as a Factor in Antenna Efficiency,” *Proc. IRE*, Jun 1937, pp 753-787.

G. H. Brown and O. M. Woodward, Jr., “Experimentally Determined Impedance Characteristics of Cylindrical Antennas,” *Proc. IRE*, Apr 1945.

L. B. Cebik, “Two Limitations of NEC-4”, **www.cebik.com/model/fd.html**.

A. Christman, “Feeding Phased Arrays: An Alternate Method,” *Ham Radio*, May 1985, pp 58-59, 61-64.

J. Devoldere, *ON4UN’s Low-Band DXing*, 4th ed. (Newington, CT: ARRL, 2005).

EZNEC is an antenna-modeling computer program for Microsoft Windows. See **eznec.com** for full information.

F. Gehrke, “Vertical Phased Arrays,” in six parts, *Ham Radio*, May-Jul, Oct and Dec 1983, and May 1984.

C. Harrison, Jr, and R. King, “Theory of Coupled Folded Antennas,” *IRE Trans on Antennas and Propagation*, Mar 1960, pp131-135.

W. Hayward and D. DeMaw, *Solid State Design for the Radio Amateur* (Newington, CT: ARRL, 1977).

W. Hayward, *Radio Frequency Design* (Newington, CT: ARRL, 1994).

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961). Later editions are edited by Richard C. Johnson.

R. King and C. Harrison, Jr, "Mutual and Self-Impedance for Coupled Antennas," *Journal of Applied Physics* , Vol 15, Jun 1944, pp 481-495.

R. King, "Self- and Mutual Impedances of Parallel Identical Antennas," *Proc. IRE*, Aug 1952, pp 981-988.

R. W. P. King, *Theory of Linear Antennas* (Cambridge, MA: Harvard Univ Press, 1956), p 275ff.

H. W. Kohler, "Antenna Design for Field-Strength Gain," *Proc. IRE*, Oct 1944, pp 611-616.

J. D. Kraus, "Antenna Arrays with Closely Spaced Elements," *Proc. IRE*, Feb, 1940, pp 76-84.

J. D. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co., 1988).

Johnson, Richard C., *Antenna Engineering Handbook*, 3rd ed. (New York: McGraw-Hill Inc., 1993). This is a later edition of the volume by the same name edited by H. Jasik.

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co, 1952).

J. L. Lawson, "Simple Arrays of Vertical Antenna Elements," *QST*, May 1971, pp 22-27.

P. H. Lee, N6PL *The Amateur Radio Vertical Antenna Handbook*, 2nd ed. (Hicksville, NY: CQ Publishing, Inc., 1984).

R. W. Lewallen, "Baluns: What They Do and How They Do It," *The ARRL Antenna Compendium, Vol 1* (Newington: ARRL, 1985). Also available for viewing at **eznec.com/Amateur/Articles/Baluns.pdf**.

R. Lewallen, "The Impact of Current Distribution on Array Patterns," Technical Correspondence, *QST*, Jul 1990, pp 39-40. Also available for viewing at **eznec.com/Amateur/Articles/Current_Dist.pdf**.

R. Lewallen, "*MININEC*—The Other Edge of the Sword", *QST*, Feb 1991, pp 18-22. *ELNEC*, referenced in the article, is no longer available.

M. W. Maxwell, "Some Aspects of the Balun Problem," *QST*, Mar 1983, pp 38-40.

J. Sevick, "The Ground-Image Vertical Antenna," *QST*, Jul 1971, pp 16-19, 22.

J. Sevick, "The W2FMI Ground-Mounted Short Vertical," *QST*, Mar 1973, pp 13-28,41.

E. J. Wilkinson, "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*, Jan, 1960.

Radio Broadcast Ground Systems, available from Smith Electronics, Inc, 8200 Snowville Rd, Cleveland, OH 44141.

Appendix A—*EZNEC-ARRL* Examples

This appendix contains step-by-step procedures using *EZNEC-ARRL* (included on the *Antenna Book* CD) to illustrate various topics discussed in the main chapter. A standard *EZNEC* program type of v. 4.0 or later may also be used. Different versions, program types and calculating engines may give results that are slightly different from those shown in the examples. However, any differences should be insignificantly small.

EZNEC Example—Mutual Coupling

Illustrates the effect of mutual coupling on feed-point impedance. Open the **ARRL_Cardioid.EZ** file, which is mounted over "perfect" ground. Click the **View Ant** button to see a diagram of the antenna, a 2-element array of vertical elements. Click on the **Wires** line in the main window to open the Wires Window. Click the button at the left of the Wire 2 line, and then press the **Delete** key on your keyboard to delete wire #2. After clicking **Ok**, note that one of the verticals has disappeared from the View Antenna display, leaving a single element. Click **Src Dat** and note that the feed-point impedance of this single vertical is about $37 + j\,1\ \Omega$--it's very nearly resonant.

Next, in the Wires Window, open the **Edit** menu at the top and click **Undo delete wire(s)** to restore the second element. Click **Src Dat** again and notice that the feed-point impedance of wire #1 is now about $21 - j\,19\ \Omega$. The feed-point impedance of the second element, which is identical to the first, is about $52 + j\,21\ \Omega$. This difference, and the change from the self-impedance of $37 + j\,1\ \Omega$, is due to mutual coupling. As you see, it's not at all a minor effect.

As an additional exercise, change the magnitude or phase angle of the source at the base of wire #2 (click **Sources** in the main window), and see how this changes the feed-point impedances of both elements. You should be able to confirm each of the four points enumerated in the **MUTUAL COUPLING** section.

EZNEC Example—Nulls

Illustrates the effect of current magnitude on nulls and gain. Again, open the **ARRL_Cardioid.EZ** file. Click the **FF Plot** button to generate the azimuth pattern of an ideal array. Save the plot for future reference as follows: In the plot window, open the **File** menu and select **Save Trace As**. Enter the name **Cardioid** and click **Save**. Now, in the main window click on the **Sources** line to open the Sources Window. Change the magnitude of source 1 from 1 to 1.1, and of source 2 from 1 to 0.9 and press **Enter** on your keyboard so that *EZNEC-ARRL* will accept the last change.

Click **FF Plot** to generate a pattern with the new currents. In the plot window, open the File menu and select **Add**

Trace. Enter the name **Cardioid** and click Open. You should now see the original plot and new plot overlaid. Notice that the null is much less deep with the altered currents, but the forward patterns are nearly identical. By clicking on the names of the traces, **Primary** and **Cardioid**, you can see in turn the gain and front-to-back ratio of each of the traces. The original, **Cardioid**, has a front-to-back ratio of about 32 dB, while Primary, the new plot, has a ratio of about 22.5 dB. The forward gain, however, differs by only 0.02 dB, a completely insignificant amount.

EZNEC Example—"Phasing-Line" Feed

Illustrates the effect of using a "phasing-line" feed. Open the **ARRL_CardTL.EZ** file. This is a model of an array fed with transmission lines whose lengths were designed using the *Arrayfeed1* program to take into account the actual load impedances of elements in a phased array. This model is mounted over "perfect" ground.

Click the **View Ant** button to show the array. Note that the lengths of the lines from the source (circle) to the elements don't represent the actual physical lengths of the lines. In the main window, click on the **Trans Lines** line to open the Transmission Lines window. In it you can see that the lengths of the feed lines, both of which are connected to the same source, are about 81° and 155°, a difference of 74° rather than 90°.

In the main window, click the **Currents** button and take a look at the current shown for segment 1 of wires 1 and 2. These are the currents at the element feed points. The ratio of the magnitude of currents is 4.577/4.561 = 1.003, and the phase difference is –56.3°–(–147.5°) = 91.2°. (A more accurate determination of feed-line lengths with program *Arrayfeed1* gives lengths of 80.61° and 153.70°, resulting in a current ratio of 1.000 at a phase of 90.02°. But the resulting pattern is very nearly the same.) But let's see what happens when we make the lines exactly 90° different in length.

First, click the **FF Plot** button to generate the azimuth pattern of the original model. Save the plot for future reference as follows: In the plot window, open the **File** menu and select **Save Trace As**. Enter the name **CardTL** and click **Save**. Now in the Transmission Lines Window, change the length of line number 1 from 80.56° to 90°. *Important:* In the line 1 Length box, enter **90d** to make the line 90° long. If you omit the **d**, it will become 90 meters long! Similarly, change the length of line 2 to 180° by entering **180d** in the line 2 Length box, then press **Enter** on your keyboard so that *EZNEC* will accept the last change.

Click **FF Plot** to generate a pattern with the new line lengths. In the plot window, open the **File** menu and select **Add Trace**. Enter the name **CardTL** and click **Open**. You should now see the original plot and new plot overlaid together. Notice that the gain of the modified model is about 1 dB greater than the original, but the front-to-back ratio has deteriorated to about 10 dB.

Experiment with different combinations of line lengths that differ by 90°--for example, 45° and 135° (don't forget the '**d**'!), or change the impedance of one or both lines and you'll see that you can get a wide variety of patterns. None, however, are likely to be as close to the ideal cardioid pattern as the original.

Chapter 9

Broadband Antenna Matching

Antenna systems that provide a good impedance match to the transmitter over a wide frequency range have been a topic of interest to hams for many years. Most emphasis has been focused on the 80-meter band since a conventional half-wave dipole will provide better than 2:1 SWR over only about one-third of the 3.5 to 4.0 MHz band. The advantage of a broadband match is obvious—fewer adjustments during tune-up and an antenna tuner may not be required.

This chapter was written by Frank Witt, AI1H, who has written numerous articles in *QST* and in *The ARRL Antenna Compendium* series on this subject. See the Bibliography for details.

The term *broadband antennas* has frequently been used to describe antenna systems that provide a wideband impedance match to the transmitter. This is something of a misnomer, since most antennas are good radiators over a wide range of frequencies and are therefore "broadband antennas" by definition. The problem is getting the energy to the antenna so it can be radiated. Antenna tuners solve this problem in some cases, although losses in transmission lines, baluns and the antenna tuner itself can be excessive. Also worthy of mention is that antenna directional properties are usually frequency dependent. In this chapter, we discuss broadbanding the impedance match to the transmitter only.

GENERAL CONCEPTS

The Objective

Fig 1 shows a simplified system: the transmitter, an SWR meter and the transmitter load impedance. Modern transceivers are designed to operate properly into a 50-Ω load and they will deliver full power into the load impedance, at their rated level of distortion, when the SWR is less than about 1.5:1. Loads beyond this limit may cause the transceiver to protect itself by lowering the output power.

Many transceivers have built-in automatic antenna tuners that permit operation for loads outside the 1.5:1 SWR range, but the matching range is often limited, particularly on the lower-frequency bands. In practice, barefoot operation is simplified if the load SWR is held to less than 2:1.

For high-power amplifiers, it is also best to keep the load SWR to less than about 2:1, since output tuning components are commonly rated to handle such loads. You can see that the primary function of the SWR meter in Fig 1 is to measure the suitability of the load impedance so far as the transmitter is concerned. Henceforth, we will use the term *load SWR* as a description of the transmitter's load.

The SWR meter is actually reading a circuit condition at a single point in the system. In fact, the meter really measures magnitude of the reflection coefficient, but is calibrated in SWR. The relationships between the complex load impedance, Z_L, the magnitude of the reflection coefficient, $|\rho|$, and the load SWR are as follows:

$$|\rho| = \left| \frac{Z_L - 50}{Z_L + 50} \right| \qquad \text{(Eq 1)}$$

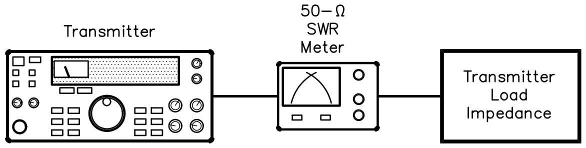

Fig 1—Basic antenna system elements at the output of a transmitter.

$$\mathrm{SWR} = \frac{1 + |\rho|}{1 - |\rho|} \qquad \text{(Eq 2)}$$

where 50 Ω is used as the reference impedance, since the design load impedance for the transmitter is assumed to be 50 Ω.

An important point is that you need not know the output impedance of the transmitter to design a broadband matching network. The issue of the actual value of the output impedance of typical RF power amplifiers is a matter of continuing controversy in amateur circles. Fortunately, this issue is not important for the design of broadband matching networks, since the load SWR is independent of the output impedance of the transmitter. Our objective is to design the matching network so that the load SWR (with a 50-Ω reference resistance) is less than some value, say 2:1, over as wide a band as possible.

Resonant Antennas

The broadband matching techniques described here apply to antennas operating near resonance. Typical resonant antennas include half-wave dipoles, quarter-wave verticals over a ground plane and full-wave loops. To design a broadband matching network, you must know the antenna feed-point impedance near resonance. **Fig 2** shows the antenna-impedance equivalent circuit near resonance. Although the series RLC circuit is an approximation, it is good enough to allow us to design matching networks that significantly increase the band over which a good match to the transmitter is achieved.

Note that the impedance is defined by F_0, the resonant frequency, R_A, the antenna resistance at resonance, and Q_A, the antenna Q. R_A is actually the sum of the radiation resistance and any loss resistance, including conductor losses and losses induced by surrounding objects, such as the ground below the antenna. R_A is frequency dependent, but it is sufficient to assume it is fixed during the matching network design process. Minor adjustments to the matching network will correct for the frequency dependence of R_A.

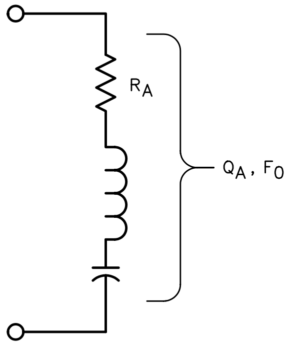

Fig 2—The equivalent circuit of a resonant antenna. The simple series RLC approximation applies to many resonant antenna types, such as dipoles, monopoles, and loops. R_A, Q_A and F_0 are properties of the entire antenna, as discussed in the text.

The antenna resistance and Q depend on the physical properties of the antenna itself, the properties of ground and the height above ground. Consider as an example an 80-meter horizontal half-wave dipole made from #12 wire located over average ground (dielectric constant = 13 and conductivity = 5 mS/m). **Figs 3** and **4** show how feed-point resistance and Q vary with height. It is clear from these figures that there are wide gyrations in antenna parameters. The better these parameters are known, the more successful we will be at designing the broadband matching network. For horizontal dipoles, Figs 3 and 4 may be used to get a good idea of antenna resistance and Q, so long as the height is scaled in proportion to wavelength. For example, a 160-meter dipole at a height of 100 feet would have about the same resistance and Q as an 80-meter dipole at 50 feet.

For optimum results, the resistance and Q can be

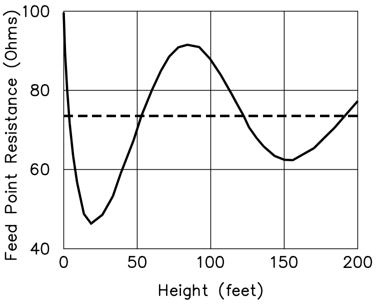

Fig 3—Feed-point resistance (solid line) for 80-meter horizontal dipole at resonance versus height over ground. The free-space value is shown as a dashed line.

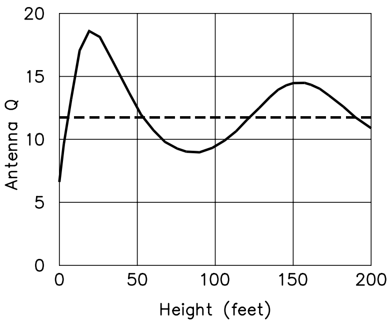

Fig 4—Antenna Q for 80-meter horizontal dipole (solid line) versus height. The free-space value is shown as a dashed line.

determined from computer simulation (using a program like *EZNEC,* for example) or measurement (using a low power SWR/Z meter such as the MFJ-259B or the Autek Research VA1 or RF-1). R_A can be computed using SWR measurements at resonance. **Fig 5** shows the typical bowl-shaped SWR curve as a function of frequency for a resonant antenna. S_0 is the SWR at resonance. You must take into account the loss of the feed line if the measurement must be made at the end of a long 50-Ω transmission line. If the line loss is low or if the measurement is made at the antenna terminals, the formulas given below apply.

For $R_A > 50\ \Omega$

$$R_A = S_0 \times 50 \qquad \text{(Eq 3)}$$

For $R_A < 50\ \Omega$

$$R_A = \frac{50}{S_0} \qquad \text{(Eq 4)}$$

There is an ambiguity, since we do not know if R_A is greater than or less than 50 Ω. A simple way of resolving this ambiguity is to take the same SWR measurement at resonance with a 10-Ω non-inductive resistor added in series with the antenna impedance. If the SWR goes up, $R_A > 45\ \Omega$. If the SWR goes down, $R_A < 45\ \Omega$. This will resolve the ambiguity.

Q_A may be determined by measuring the SWR bandwidth. Define BW_2 as the bandwidth over which the resonant antenna SWR is less than 2:1 (in the same units as F_0). See Fig 5.

For $R_A > 50\ \Omega$

$$Q_A = \frac{F_0}{BW_2}\sqrt{2.5\,S_0 - S_0^{\,2} - 1} \qquad \text{(Eq 5)}$$

For $R_A < 50\ \Omega$

$$Q_A = \frac{F_0}{BW_2 \times S_0}\sqrt{2.5\,S_0 - S_0^{\,2} - 1} \qquad \text{(Eq 6)}$$

Again, you must take into account the loss of the feed line if you make the measurement at the end of a long 50-Ω transmission line.

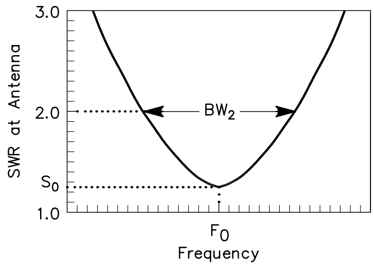

Fig 5—SWR at the antenna versus frequency in the vicinity of resonance.

For horizontal and inverted-V half-wave dipoles, a very good approximation for the antenna Q when R_A is known is given by:

$$Q_A = \frac{93.9\left[\ln\dfrac{8110}{D\,F_0} - 1\right]}{R_A} \qquad \text{(Eq 7)}$$

where D = the diameter of wire, in inches.

Loss

When you build a resonant antenna with some matching scheme designed to increase the bandwidth, you might overlook the loss introduced by the broadband matching components. Loss in dB is calculated from:

$$\text{Loss} = -10\log\frac{\text{Power radiated by antenna}}{\text{Total power from transmitter}} \qquad \text{(Eq 8)}$$

or, alternatively, define efficiency in % as

$$\text{Efficiency} = \frac{100 \times \text{Power radiated by antenna}}{\text{Total power from transmitter}} \qquad \text{(Eq 9)}$$

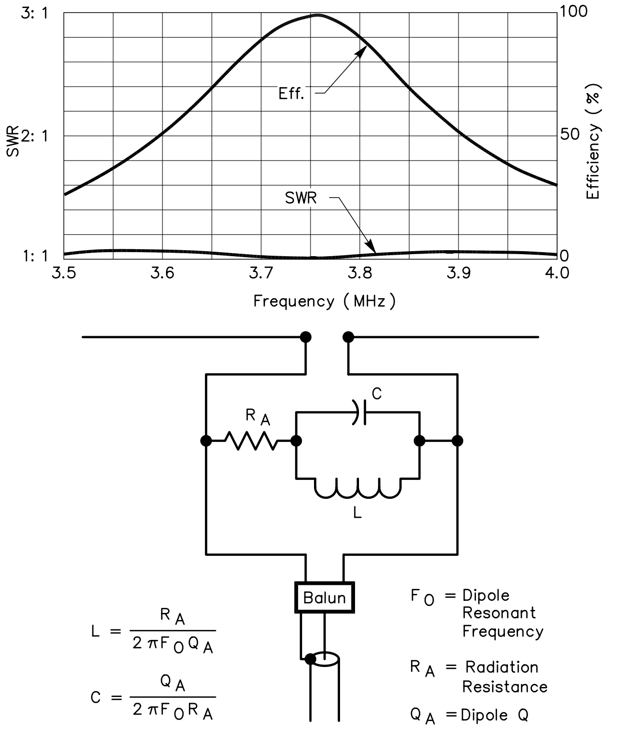
3: 1
2: 1
1: 1
SWR
100
50
0
Efficiency (%)
Eff.
SWR
3.5
3.6
3.7
3.8
3.9
4.0
Frequency (MHz)
C
R_A
L
Balun
F_0 = Dipole Resonant Frequency
R_A = Radiation Resistance
Q_A = Dipole Q
$L = \frac{R_A}{2\pi F_0 Q_A}$
$C = \frac{Q_A}{2\pi F_0 R_A}$

Fig 6—Matching the dipole with a complementary RLC network greatly improves the SWR characteristics, nearly 1:1 across the 3.5-MHz band. However, the relative loss at the band edges is greater than 5 dB.

An extreme degree of bandwidth broadening of an 80-meter dipole is illustrated in **Fig 6**. This approach is not recommended, but it is offered here to make a point. The broadening is accomplished by adding resistive losses. From network theory we obtain the RLC (resistor, inductor, capacitor) matching network shown. The network provides the complement of the antenna impedance. Note that the SWR is virtually 1:1 over the entire band (and beyond), but the efficiency falls off dramatically away from resonance.

The band-edge efficiency of 25 to 30% shown in Fig 6 means that the antenna has about 5 to 6 dB of loss relative to an ideal dipole. At the band edges, 70 to 75% of the power delivered down the transmission line from the transmitter is heating up the matching-network resistor. For a 1-kW output level, the resistor must have a power rating of at least 750 W! Use of an RLC complementary network for broadbanding is not recommended, but it does illustrate how resistance (or losses) in the matching network can significantly increase the apparent antenna SWR bandwidth.

The loss introduced by any broadband matching approach must be taken into consideration. Lossy matching networks will usually provide more match bandwidth improvement than less lossy ones.

SIMPLE BROADBAND MATCHING TECHNIQUES

The Cage Dipole

You can increase the match bandwidth of a single-wire dipole by using a thick radiator, one with a large diameter. The gain and radiation pattern are essentially the same as that of a thin-wire dipole. The radiator does not necessarily have to be solid; open construction such as shown in **Fig 7** may be used.

The theoretical SWR response of an 80-meter cage dipole having a 6-inch diameter is shown in **Fig 8**. BW_2 for this antenna fed with 50-Ω line is 287 kHz, and the Q is approximately 8. Its 2:1-SWR frequency range is 1.79 times broader than a dipole with a Q of 13, typical for thin-wire dipoles.

There are other means of creating a thick radiator, thereby gaining greater match bandwidth. The bow-tie and the fan dipole make use of the same Q-lowering principle as the cage for increased match bandwidth. The broadbanding techniques described below are usually more practical than the unwieldy cage dipole.

Stagger-Tuned Dipoles

A single-wire dipole exhibits a relatively narrow bandwidth in terms of coverage for the 3.5 to 4.0-MHz band. A technique that has been used for years to cover the entire band is to use two dipoles, one cut for the CW portion and one for the phone portion. The dipoles are connected in parallel at the feed point and use a single feeder. This technique is known as *stagger tuning*.

Fig 9 shows the theoretical SWR response of a pair of stagger-tuned dipoles fed with 50-Ω line. No mutual coupling between the wires is assumed, a condition

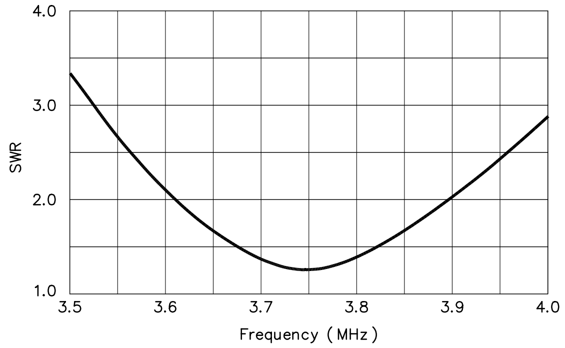

Fig 8—Theoretical SWR versus frequency response for a cage dipole of length 122 feet 6 inches and a spreader diameter of 6 inches, fed with 50-Ω line. The 2:1-SWR bandwidth frequencies are 3.610 and 3.897 MHz, with a resulting BW_2 of 287 kHz.

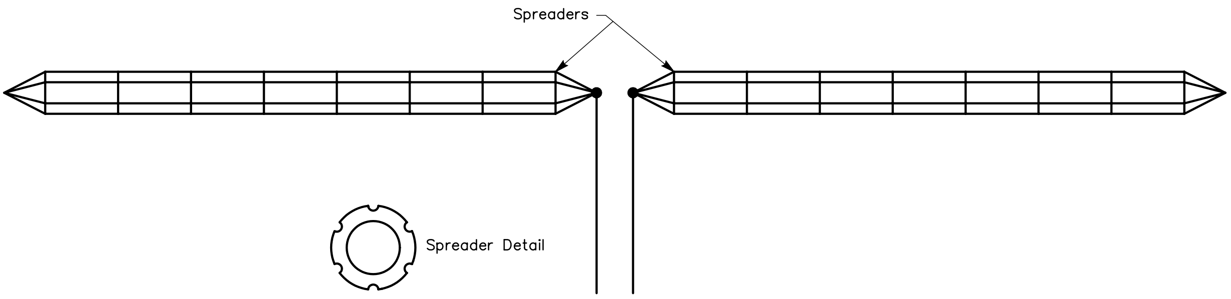

Fig 7—Construction of a cage dipole, which has some resemblance to a round birdcage. The spreaders need not be of conductive material, and should be lightweight. Between adjacent conductors, the spacing should be 0.02 λ or less. The number of spreaders and their spacing should be sufficient to maintain a relatively constant separation of the radiator wires.

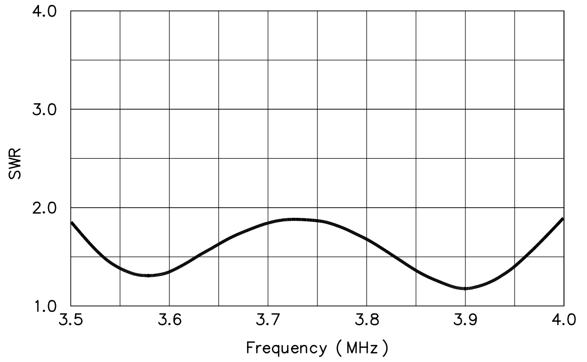

Fig 9—Theoretical SWR response of two stagger-tuned dipoles. They are connected in parallel at the feed point and fed with 50-Ω line. The dipoles are of wire such as #12 or #14, with total lengths of 119 and 132 feet.

that would exist if the two antennas were mounted at right angles to one another. As Fig 9 shows, the SWR response is less than 1.9:1 across the entire band.

A difficulty with such crossed dipoles is that four supports are required for horizontal antennas. A more common arrangement is to use inverted V dipoles with a single support, at the apex of each element. The radiator wires can also act as guy lines for the supporting mast.

When the dipoles are mounted at something other than a right angle, mutual coupling between them comes into play. This causes interaction between the two elements—tuning of one by length adjustment will affect the tuning of the other. The interaction becomes most critical when the two dipoles are run parallel to each other, suspended by the same supports, and the wires are close together. Finding the optimum length for each dipole for total band coverage can become a tedious and frustrating process.

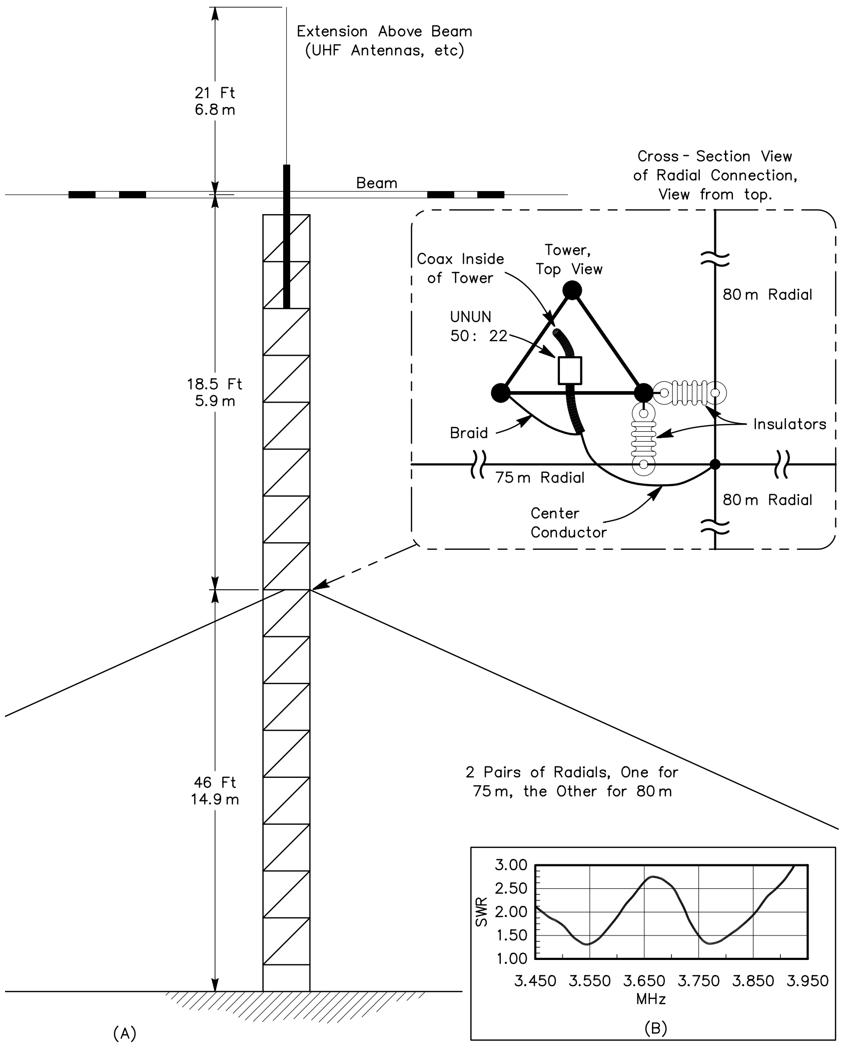
Extension Above Beam
(UHF Antennas, etc)
21 Ft
6.8 m
Beam
18.5 Ft
5.9 m
46 Ft
14.9 m
Cross - Section View
of Radial Connection,
View from top.
Coax Inside
of Tower
Tower,
Top View
UNUN
50: 22
80 m Radial
Braid
Insulators
75 m Radial
Center
Conductor
80 m Radial
2 Pairs of Radials, One for
75 m, the Other for 80 m
3.00
2.50
2.00
1.50
1.00
SWR
3.450
3.550
3.650
3.750
3.850
3.950
MHz
(A)
(B)

Fig 10—At A, details of the inverse-fed ground-plane at W4PK. The inset shows the feed-point details. Keep the two sets of radials as close to 90· as possible. Don't try to resonate the antenna by adjusting radial lengths. At B, SWR curve of the inverse-fed ground plane antenna with stagger-tuned radials. The objective is to have a low SWR for both the CW and SSB 80/75-meter DX windows.

Stagger-Tuned Radials

A variation of the use of stagger tuning has been applied by Sam Leslie, W4PK. See his article entitled “Broadbanding the Elevated, Inverse-Fed Ground Plane Antenna” in *The ARRL Antenna Compendium, Vol 6*. An existing tower with a large beam and VHF/UHF antennas on top is used as an 80-meter monopole. See **Fig 10A**. The feed point is part way up the tower, at a point where the metal above it makes up an electrical quarter wavelength. Four elevated quarter-wave radials are used as a ground plane. The radials droop away from the tower. They are joined together at the tower but not connected to the tower. The antenna is fed with 50-Ω coax and an impedance step-down autotransformer (50:22 Ω unun) located at the tower. The shield of the coax is carried through the transformer and connected to the tower. The coax “hot side” is connected to the junction of the radials. A Q-section made from two paralleled RG-59 coax cables (Z_0 = 75/2 = 37.5 Ω) may be used instead of the autotransformer.

The broadband match is achieved by cutting two opposing radials for one end of the desired match range (75 meters) and cutting the other two for the other end of the range (80 meters). See Fig 10B for the resultant SWR versus frequency characteristic.

FEED-LINE IMPEDANCE MISMATCHING

The simplest broadband matching network is a transformer at the junction of the transmission line and the antenna. See Frank Witt’s article entitled “Match Bandwidth of Resonant Antenna Systems” in October 1991 *QST*. Observe that the impedance of the series RLC resonant antenna model of Fig 2 increases at frequencies away from resonance. The result is that more bandwidth is achieved when SWR at resonance is not exactly 1:1. For example, if the antenna is fed with a low-loss 50-Ω transmission line, the maximum SWR bandwidth is obtained if the effective R_A = 50/1.25 = 40 Ω. The improvement in BW_2, the 2:1 SWR bandwidth, over a perfect match at resonance is only 6%, however. This condition of R_A = 40 Ω can be achieved by installing a transformer at the junction of the feed line and the antenna.

If the line loss is larger, you can gain more benefit in terms of bandwidth by deliberately mismatching at the antenna. If the matched-line loss is 2 dB, the improvement rises to 18%. To achieve this requires an R_A = 28.2 Ω. Again, the improvement over a perfect match at resonance is modest.

It is interesting to compare the 2:1 SWR bandwidth (at the transmitter end of the line) for the case of a lossless line versus the case of 2-dB matched-line loss. BW_2 for the lossy line case is 1.95 times that of the lossless line case, so substantial match bandwidth improvement occurs, but at the cost of considerable loss. The band edge loss for the lossy-line case is 3 dB!

This example is provided mostly as a reference for comparison with the more desirable broadbanding techniques below. It also explains why some installations with long feed lines show large match bandwidths.

PARALLEL-TUNED CIRCUITS AT THE ANTENNA TERMINALS

As mentioned previously, a resonant antenna has an equivalent circuit that may be represented as a series RLC circuit. We can cancel out some of the antenna reactance away from resonance with a parallel-tuned circuit connected across the antenna terminals. The impedance level of the parallel-tuned circuit must be low enough to be effective and must have the same resonant frequency as the antenna. You can use a parallel-tuned circuit made with lumped LC components or with transmission-line segments. A quarter-wave transmission line with a short at the far end, or a half-wave line with an open at the far end, each behave like a parallel-tuned LC circuit.

The Double Bazooka

The response of the controversial *double bazooka* antenna is shown in **Fig 11**. This antenna actually consists of a dipole with two quarter-wave coaxial resonator stubs connected in series. The stubs are connected across the antenna terminals.

Not much bandwidth enhancement is provided by this resonator connection because the impedance of the matching network is too high. The antenna offers a 2:1-SWR bandwidth frequency range that is only 1.14 times that of a simple dipole with the same feeder. And the bandwidth enhancement is partially due to the “fat” antenna wires composed mostly of the coax shield. No

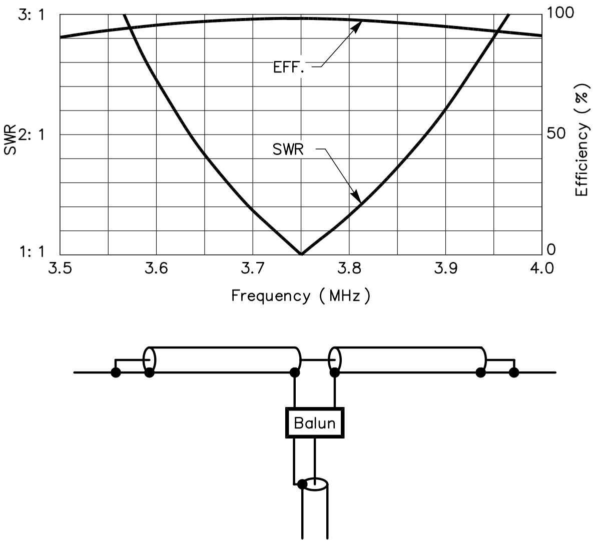

Fig 11—The double bazooka, sometimes called a coaxial dipole. The antenna is self-resonant at 3.75 MHz. The resonator stubs are 43.23-foot lengths of RG-58A coax.

improvement in antenna gain or pattern over a thin-wire dipole can be expected from this antenna.

The Crossed Double Bazooka

A modified version of the double-bazooka antenna is shown in **Fig 12**. In this case, the impedance of the matching network is reduced to one-fourth of the impedance of the standard double-bazooka network. The lower impedance provides more reactance correction, and hence increases the bandwidth frequency range noticeably, to 1.55 times that of a simple dipole. Notice, however, that the efficiency of the antenna drops to about 80% at the 2:1-SWR points. This amounts to a loss of approximately 1 dB. The broadbanding, in part, is caused by the losses in the coaxial resonator stubs (made of RG-58A coax), which have a remarkably low Q (only 20).

MATCHING NETWORK DESIGN

Optimum Matching

You can improve the match bandwidth by using a transformer or a parallel-tuned circuit at the antenna terminals. A combination of these two techniques can yield a very effective match bandwidth enhancement. Again, you can make the impedance transformation and implement the tuned circuit using lumped-circuit elements (transformers, inductors and capacitors) or with distributed-circuit elements (transmission lines).

In 1950, an article by R. M. Fano presented the theory of broadband matching of arbitrary impedances. This classic work was limited to lossless matching networks. Unfortunately, matching networks realized with transmission lines have enough loss to render the Fano results useful as only a starting point in the design process. A theory of optimum matching with lossy matching networks was presented in an article by Frank Witt entitled "Optimum Lossy Broadband Matching Networks for Resonant Antennas" that appeared in April 1990 *RF Design* and was summarized and extended in "Broadband Matching with the Transmission Line Resonator," in *The ARRL Antenna Compendium Vol 4*.

Using a matching network, the 2:1 SWR bandwidth can be increased by a factor of about 2.5. Instead of the familiar bowl-shaped SWR versus frequency characteristic seen in Fig 5, the addition of the matching network yields the W-shaped characteristic of **Fig 13**. The maximum SWR over the band is S_M. The band edges are F_L and F_H. Thus the bandwidth, BW, and the center frequency, F_0, are

$$BW = F_H - F_L \qquad \text{(Eq 10)}$$

$$F_0 = \sqrt{F_H \times F_L} \qquad \text{(Eq 11)}$$

Fig 14 shows the pertinent equivalent circuit for a typical broadband antenna system. The parallel LC matching network is defined by its resonant frequency, F_0 (which is the same as the resonant frequency of the antenna), its Q and its impedance level. The impedance level is the reactance of either the matching network inductance or capacitance at F_0. The optimum transmitter load resistance,

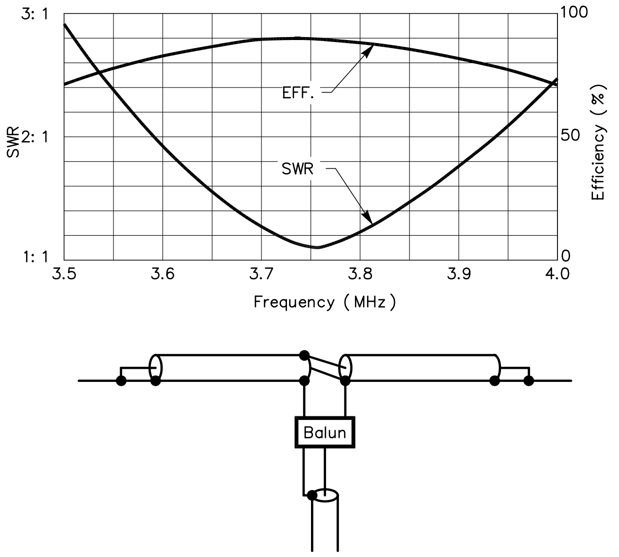

Fig 12—The crossed double bazooka yields bandwidth improvement by using two quarter-wave resonators, parallel connected, as a matching network. As with the double bazooka, the resonator stubs are 43.23-foot lengths of RG-58A coax.

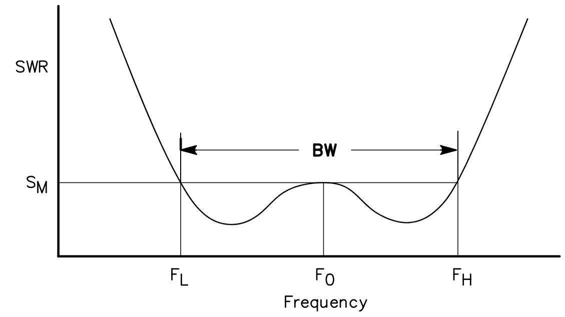

Fig 13—W-shaped SWR versus frequency characteristic that results when a transformer and parallel LC matching network are used.

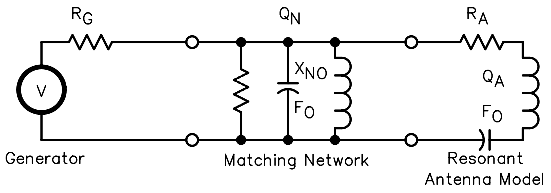

Fig 14—Assumed equivalent circuit for the broadband matching system. The antenna and the matching network have the same resonant frequency.

R_G, is achieved using an impedance transformer between the transmitter and the input to the parallel-tuned circuit. Thus for a transmitter whose optimum load resistance is 50 Ω, the required impedance transformer would have an impedance transformation ratio of 50:R_G.

The goal is to keep the SWR over the band as low as possible. There is a best or minimum value of S_M that may be achieved over the entire band. It depends on the desired match bandwidth and the Q's of the antenna and matching network and is given by the following formula:

$$S_{M\,\min} = \frac{\sqrt{B_N{}^2 + 1} + \sqrt{B_N{}^2 + 1 + \dfrac{2Q_A}{Q_N}\left(1 + \dfrac{Q_A}{2\,Q_N}\right)}}{2\left(1 + \dfrac{Q_A}{2\,Q_N}\right)} \qquad \text{(Eq 12)}$$

where $B_N = \dfrac{BW}{F_0}$ and

Q_A = antenna Q

Q_N = matching network Q.

The reactance at resonance of the parallel tuned LC matching network is given by:

$$X_{N0} = \frac{R_A}{Q_A}\left[\left(1 + \frac{Q_A}{2Q_N}\right) S_{M\,\min}{}^2 - \frac{Q_A}{2Q_N}\right] \qquad \text{(Eq 13)}$$

where R_A = antenna resistance in Ω.

The transmitter load resistance is calculated from:

$$R_G = \frac{S_{M\,\min}\, R_A\, Q_N\, X_{N0}}{R_A + Q_N\, X_{N0}} \qquad \text{(Eq 14)}$$

The loss of the matching network is always greatest at the edges of the band. This loss, L_{MNE}, in dB, is given by:

$$L_{MNE} = 10 \log\left[1 + \frac{R_A}{Q_N\, X_{N0}}\left(B_N{}^2 + 1\right)\right] \qquad \text{(Eq 15)}$$

Now you have all the information needed to design an optimal broadband matching network. An 80-meter dipole will serve as an example to illustrate the procedure. The following parameters are assumed:

F_L = 3.5 MHz
F_H = 4.0 MHz
R_A = 57.2 Ω
Q_A = 13
Q_N = 40.65

The values of R_A and Q_A are reasonable values for an 80-meter inverted V with an apex 60 feet above ground. Q_N = 40.65 is the calculated value of the Q of a resonator made from RG-213 or similar cable in the middle of the 80-meter band. The results are:

BW = 0.5 MHz (from Eq 10)
F_0 = 3.742 MHz (from Eq 11)
S_{Mmin} = 1.8 (from Eq 12)
X_{N0} = 15.9 Ω (from Eq 13)
R_G = 94.8 Ω (from Eq 14)
L_{MNE} = 1.32 dB (from Eq 15)

Shown in **Fig 15** is the plot of SWR and matching-network loss versus frequency for this example. The reference resistance for the SWR calculation is R_G or 94.8 Ω. Since the desired load resistance for the transmitter is 50 Ω, we need some means to transform from 94.8 Ω to 50 Ω. The examples given in the rest of this chapter will show a variety of ways of achieving the necessary impedance transformation.

The meaning of X_{N0} = 15.9 Ω is that the inductor and the capacitor in the matching network would each have a reactance of 15.9 Ω at 3.742 MHz. Methods for obtaining a particular value of X_{N0} are shown in the fol-

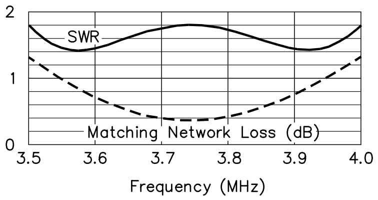

Fig 15—SWR and matching-network loss for example. Note that S_M is exactly 1.8 and the band-edge loss is exactly 1.32 dB, as calculated.

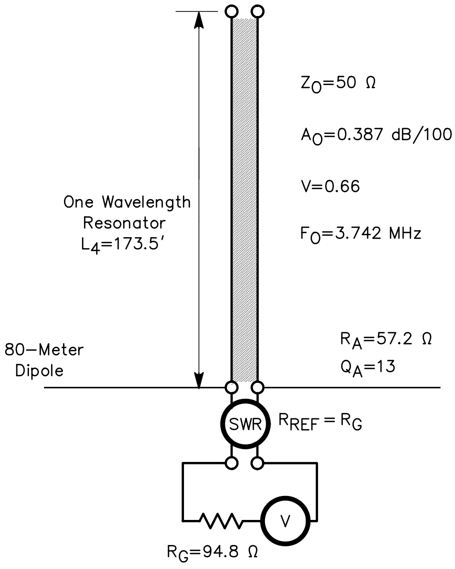

Fig 16—80-meter dipole example with a 1-λ resonator as the matching network.

lowing examples as well. The particular value of X_{N0} of this example may be obtained with a one-wavelength resonator (open-circuited at the far end) made from RG-213. The system described in this example is shown in **Fig 16**.

The broadband antenna system shown in Fig 16 has a desirable SWR characteristic, but the feed line to the transmitter is not yet present. Fortunately, the same cable segment that makes up the resonator may be used as the feed line. A property of a feed line whose length is a multiple of λ/2 (an even multiple of λ/4) is that its input impedance is a near replica of the impedance at the far end. This is exactly true for lossless lines at the frequencies where the λ/2 condition is true. Fortunately, practical lines have low-enough loss that the property mentioned above is true at the resonant frequency. Off resonance, the desired reactance cancellation we are looking for from the resonant line takes place. The very same design equations apply.

In **Fig 17**, the antenna shown in Fig 16 is moved to the far end of the 1λ resonator. The SWR and loss for this case are shown in **Fig 18**, which should be compared with Fig 15. The SWR is only slightly degraded, but the loss is about the same at the band edges. We have picked up a feed line that is 1λ long (173.5 feet) *and* broadband matching. You can make the SWR curve virtually the same as that of Fig 15 if the transmitter load resistance is increased from 94.8 Ω to 102 Ω. From a practical point of view, this degree of design refinement is unnecessary, but is instructive to know how the SWR characteristic may be controlled. You may apply the same approach when the resonator is an odd multiple of a wavelength, but in that case the transmitter or antenna must be connected λ/4 from the shorted end.

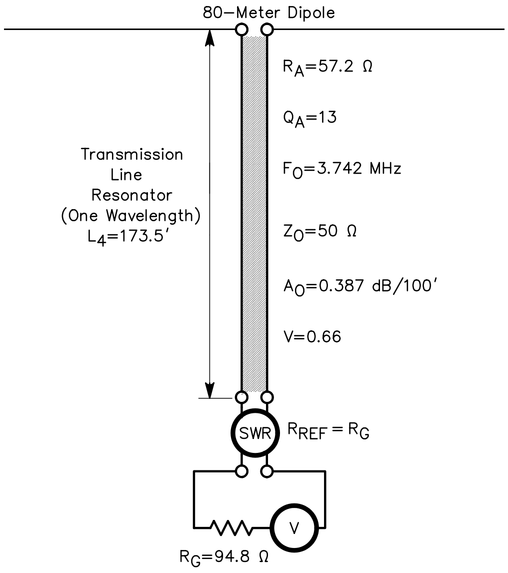

Fig 17—The antenna is at the far end of the 1-λ line. The line does double duty—broadband matching and signal transport.

Chebyshev Matching

Although it is not as efficient from the standpoint of bandwidth improvement and loss, Chebyshev matching is useful for some purposes. This arrangement yields a transmitter load impedance of 50 + *j* 0 Ω at two frequencies in the band. The matching network circuit is the same as the optimum matching case, but the parameters are different. In this case, we assume that the antenna parameters, F_0, R_A and Q_A, the network Q and the maximum SWR over the band are specified. The bandwidth, impedance level and generator resistance are given by:

$$BW = \frac{F_0}{Q_A Q_N} \left\{ \frac{(Q_A + Q_N)\left[(Q_A + 2Q_N)S_M - Q_A\right](S_M - 1)}{S_M} \right\}^{\frac{1}{2}} \qquad \text{(Eq 16)}$$

$$X_{N0} = \frac{R_A}{Q_A}\left[S_M + \frac{Q_A}{Q_N}(S_M - 1)\right] \qquad \text{(Eq 17)}$$

$$R_G = R_A\left(S_M - \frac{Q_A}{Q_A + Q_N}\right) \qquad \text{(Eq 18)}$$

The loss at the band edges is given by Eq 15. The frequencies where the transmitter load is 50 + *j* 0 Ω, F_L and F_H, are given by:

$$F_L = \sqrt{F_0^{\,2} + F_M^{\,2}} - F_M \qquad \text{(Eq 19)}$$

where

$$F_M = \frac{F_0}{2Q_A}\left[(S_M - 1)\left(1 + \frac{Q_A}{Q_N}\right)\right]^{\frac{1}{2}} \qquad \text{(Eq 20)}$$

$$F_H = \frac{F_0^{\,2}}{F_L} \qquad \text{(Eq 21)}$$

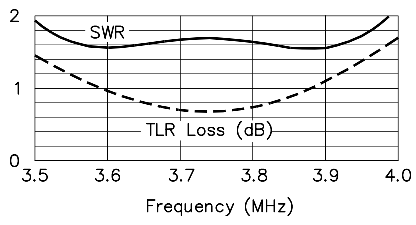

Fig 18—SWR and loss for a 1-λ 50-Ω feed line. The antenna system model is the one shown in Fig 17.

LC-MATCHING NETWORKS

Before the theory of optimal matching above was developed, Frank Witt, AI1H, described in his article "Broadband Dipoles—Some New Insights" in October 1986 *QST* the basic principle and some applications of LC matching networks. **Fig 19** shows an LC matching network that is a modified version of one originally proposed by Alan Bloom, N1AL. Note that the basic ingredients are present: the parallel-tuned circuit for reactance cancellation and a transformer.

The network also functions as a voltage balun by connecting the shield of the coaxial feed line to the center tap of the inductor. The capacitor is connected across the entire coil to obtain practical element values. The SWR response and efficiency offered by a network of lumped components is shown in **Fig 20**. The 2:1-SWR bandwidth with 50-Ω line is 460 kHz. The design is a Chebyshev match, where SWR = 1:1 is achieved at two frequencies. For the same configuration, more match bandwidth and less loss could have been realized if the optimal-match theory had been applied.

The efficiency at the band edges for the antenna system shown in Fig 20 is 90% (Loss = 0.5 dB). This low loss is due to the high Q of the LC matching network ($Q_N \approx 200$). The very low-impedance level required (X_{N0} = 12.4 Ω) cannot be easily realized with practical inductor-capacitor values. It is for this reason that the coil is tapped and serves as an autotransformer with multiple taps. The required impedance transformation ratio of 1:2.8 is easily achieved with this arrangement.

80-Meter DXer's Delight

A version of an antenna system with an LC broadbanding network is dubbed the *80-meter DXer's Delight*. This antenna has SWR minima near 3.5 and 3.8 MHz. A single antenna permits operation with a transmitter load impedance of about 50 Ω in the DX portions of the band, both CW and phone. The efficiency and SWR characteristic are shown in **Fig 21**.

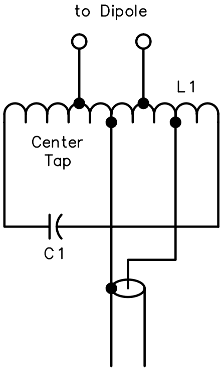

Fig 19—A practical LC matching network that provides reactance compensation, impedance transformation and balun action.

Fig 22 shows the method of construction. The selection of a capacitor for this application must be made carefully, especially if high power is to be used. For the capacitor described in the caption of Fig 22, the allowable peak power (limited by the breakdown voltage) is 2450 W. However, the allowable average power (limited by the RF current rating of 4 A) is only 88 W! These limits apply at the 1.8:1 SWR points.

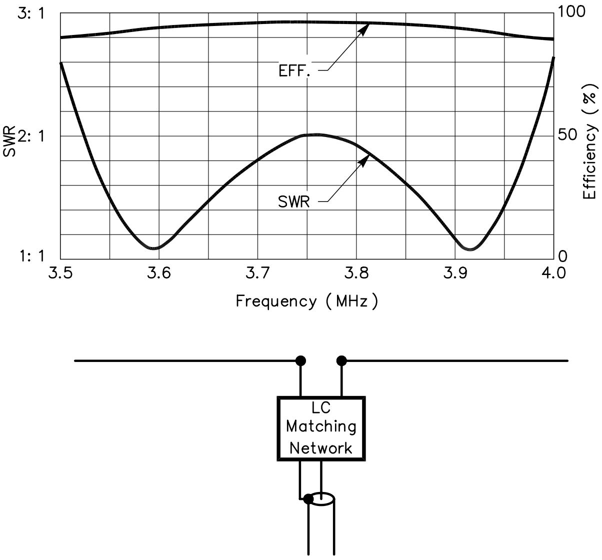

Fig 20—Efficient broadband matching with a lumped element LC network. The antenna in this example has a resistance of 72 Ω and a Q of 12. The matching network has an impedance level of 12.4 Ω, and a Q of 200. The feed line is 50-Ω coax, and the matching network provides an impedance step-up ratio of 2.8:1.

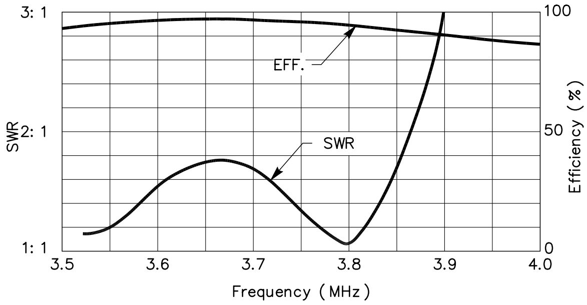

Fig 21—The 80-meter DXer's Delight permits operation with a 50-Ω transmitter load in the DX portions of the band, both CW and phone. This network was designed to provide a broadband match for an inverted V with F_0 = 3.67 MHz, an antenna resistance of 60 Ω and a Q of 13. The matching network has an impedance level of 9 Ω, and a Q of 220. The feed line is 50-Ω coax, and the matching network provides an impedance step-up ratio of 2.0:1.

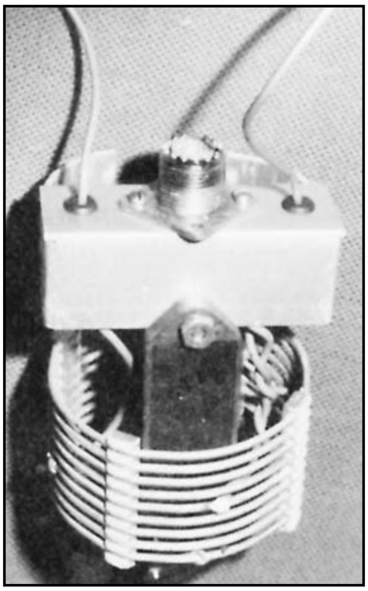

Fig 22—A method of constructing the LC broadband matching network. Components must be chosen for high Q and must have adequate voltage and current ratings. The network is designed for use at the antenna feed point, and should be housed in a weather-resistant package. The component values are as follows:

C1—400 pF transmitting mica rated at 3000 V, 4 A (RF)

L1—4.7 µH, 8 1/2 turns of B&W coil stock, type 3029 (6 turns per in., 2 1/2-in. dia, #12 wire). The LC circuit is brought to midband resonance by adjusting an end tap on the inductor. The primary and secondary portions of the coil have 1 3/4 and 2 1/2 turns, respectively.

PROPERTIES OF TRANSMISSION-LINE RESONATORS

One way around the power limitations of LC broadband matching networks is to use a transmission line as the resonator. Transmission line resonators, TLRs, typically have higher power-handling capability because the losses, though higher, are distributed over the length of the line instead of being concentrated in the lumped-LC components. Transmission-line resonators must be multiples of a quarter wavelength. For a parallel-tuned circuit, the odd-multiple λ/4 lines must be shorted at the far end and the even-multiple λ/4 lines must be open circuited.

The length of the matching network resonator is n electrical quarter wavelengths. The impedance level, Q, and line length are given by:

$$X_{N0} = \frac{4 Z_0}{n \pi}$$

or, solving for Z_0,

$$Z_0 = \frac{n \pi X_{N0}}{4} \quad \text{(Eq 22)}$$

$$Q_N = \frac{2.774\, F_0}{A_0\; V} \quad \text{(Eq 23)}$$

$$L_N = \frac{245.9\; n\; V}{F_0} \quad \text{(Eq 24)}$$

where

Z_0 = line characteristic impedance in Ω

A_0 = matched line loss in dB per 100 feet

V = velocity factor, and

L_N = line length in feet.

These equations may be used to find the properties of the transmission line resonator in the systems shown in Figs 16 and 17. The line is RG-213 with the following properties:

Z_0 = 50 Ω

V = 0.66

A_0 = 0.4 dB/100 feet at 4 MHz

We need the matched loss at 3.742 MHz. Since the line loss is approximately proportional to $\sqrt{\text{frequency}}$,

$$A_{0F2} = A_{0F1} \sqrt{\frac{F_2}{F_1}} \quad \text{(Eq 25)}$$

where A_{0F1} and A_{0F2} are the matched line losses at frequencies F_1 and F_2, respectively. Hence,

A_0 = 0.387 dB/100 feet at 3.742 MHz

Q_N = 40.65 (from Eq 23)

This is the Q of any resonator made from RG-213 for a frequency of 3.742 MHz. It does not matter how many quarter wavelengths (n) make up the cable segment. Earlier we calculated that the required impedance level, X_{N0}, is 15.9 Ω. Therefore, the required impedance level is:

$$Z_0 = \frac{n \pi \times 15.9}{4} = 12.5\, n\, \Omega \quad \text{(from Eq 22)}$$

This means that you could use a quarter-wavelength segment with a short at the far end if Z_0 = 12.5 Ω, or a half-wavelength segment with a open at the far end if Z_0 = 25 Ω. Our example used an open-circuited full-wavelength resonator (n = 4) with a Z_0 of 50 Ω. This example was crafted to make this happen (so that it fits RG-213 properties), but it illustrates that some juggling is required to obtain the right resonator properties when it is made from a transmission line, since we are stuck with the cable's characteristic impedance.

The TLR Transformer

A way around the limitation created by available characteristic impedance values is to use the *transmission-line resonator transformer*. The basic idea is to make the connection to the TLR at an intermediate point along the line instead of at the end of the line. It is analogous to using taps on a parallel-tuned LC resonant circuit to achieve transformer action (See Fig 19). The nature of the transformer action is seen in **Fig 23**, where the ends of the transformer are designated 1-1 and 2-2. The impedance ratio of the transformer, N_Z, is approximated by

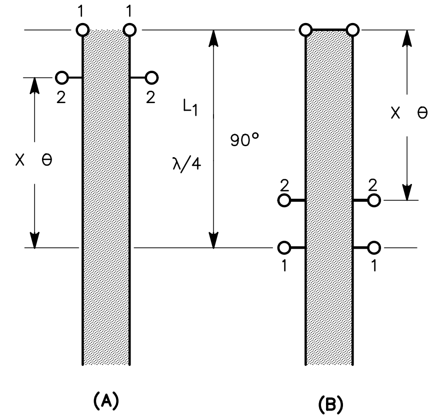

Fig 23—Transformer action in the TLR. The definition of transformer terminals depends on whether the TLR end is open- or short-circuited. θ is the distance between the minimum of the voltage standing wave (at resonance) and the connection point, expressed as an electrical angle. The distance x is that same distance expressed in feet.

$$N_Z = \sin^2 \theta \qquad \text{(Eq 26)}$$

The distance in feet, x, is given by

$$x = \frac{\theta}{90} \times \frac{\lambda}{4} = \frac{\theta}{90} \times \frac{245.9\ V}{F_0} \qquad \text{(Eq 27)}$$

This approach may be used for the connection of the TLR to the transmitter or to the antenna or both. An example will help make the power of the TLR transformer clear. We want to broadband match an 80-meter dipole (where F_0 is 3.742 MHz, R_A is 65 Ω and Q_A is 13). The feed line is RG-213 50-Ω cable, and the total distance between the antenna and transmitter is less than 100 feet.

Since we know that Z_0 = 50 Ω, we will design a prototype system with a 3/4 λ TLR (n=3). Since n is odd, the TLR must be shorted at one end, and we will place the short at the transmitter end of the line. The intermediate step of using a prototype system will enable us to design a system with the final TLR, but without transformer action. Then we'll calculate the proper connection points along the TLR. **Fig 24** shows the prototype system and the system with built-in transformers.

Note that the prototype antenna is connected to the open end of the TLR. The prototype generator (transmitter) is connected at a multiple of λ/2 from the antenna. Here, at resonance, an approximate replica of the antenna will appear. We will use primed variables to indicate prototype antenna system elements. The prototype antenna differs from the actual antenna in impedance level, so only R_A will change. F_0 and Q_A remain the same.

Starting with the 3/4 λ TLR, we have:

X_{N0} = 21.2 Ω (from Eq 22)

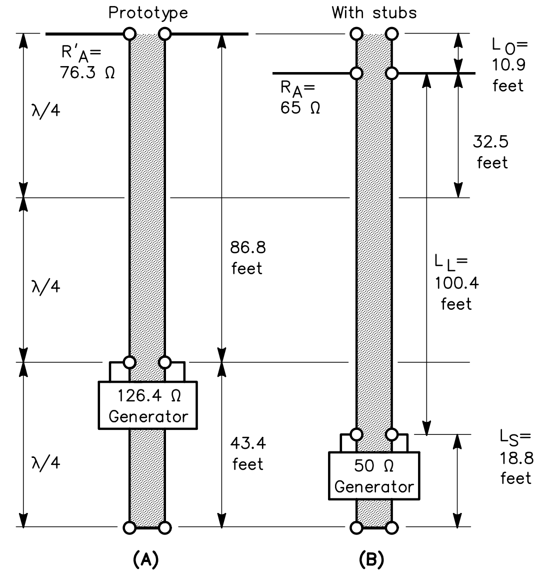

Fig 24—An optimized antenna system with a 3/4 λ TLR made from RG-213 coaxial cable. The prototype system at A is a convenient intermediate step in the design process. At B is the configuration with TLR transformers. It has an open stub, L_O, a link, L_L (which serves as the feed line), and a shorted stub, L_S.

The prototype antenna resistance at resonance, R'_A, is calculated by rearranging Eq 13.

$$R'_G = \frac{X_{N0}\, Q_A}{\left(1 + \frac{Q_A}{2 Q_N}\right) S_{M\,min}{}^2 - \frac{Q_A}{2\, Q_N}} \qquad \text{(Eq 28)}$$

Earlier we calculated that Q_N = 40.65 and S_{Mmin} = 1.8. Now we find

R'_A = 76.3 Ω (from Eq 28)
R'_G = 126.4 Ω (from Eq 14)

The electrical angles (in degrees) and connection locations (in feet) are given by

$$\theta_A = \sin^{-1} \sqrt{\frac{R_A}{R'_A}} \qquad \text{(Eq 29)}$$

$$\theta_G = \sin^{-1} \sqrt{\frac{R_G}{R'_G}} \qquad \text{(Eq 30)}$$

$$X_A = \frac{\theta_A}{90} L_1 \qquad \text{(Eq 31)}$$

$$X_G = \frac{\theta_G}{90} L_1 \qquad \text{(Eq 32)}$$

where

$$L_1 = \frac{\lambda}{4} = \frac{245.9\ V}{F_0}$$

The lengths of the stubs and link of Fig 24 are given by

$L_O = L_1 = x_A$ (Eq 33)

$L_S = x_G$ (Eq 34)

$L_L = 3L_1 - L_O - L_S$ (Eq 35)

Since L_1 = 43.4 feet,

$\theta_A = 67.4°$ (from Eq 29)

$\theta_G = 39.0°$ (from Eq 30)

L_O = 10.9 feet (from Eq 33)

L_S = 18.8 feet (from Eq 34)

L_L = 100.4 feet (from Eq 35)

These dimensions are summarized in Fig 24. If required, additional 50-Ω cable may be added between the transmitter and the junction of the shorted stub and the link. A remarkable aspect of the TLR transformer, as demonstrated in this example, is that it is a transformer with multiple taps distributed over a great distance, in this case over 100 feet.

The calculated SWR and loss of the antenna system with the TLR transformer are shown in **Fig 25A**. We mentioned earlier that the stub calculations are an approximation. Fig 25 bears this out, but the changes necessary to achieve the optimized results are relatively small. By changing the lengths and dipole resonant frequency slightly, the target optimized SWR characteristic of Fig 25B results. See **Table 1** for a summary of the changes required.

To demonstrate the significant improvement in match bandwidth that the TLR transformer provides, the SWR and loss of the same dipole fed with 100.4 feet of RG-213 cable is shown in Fig 25C. The loss added by the two stubs used to obtain match bandwidth enhancement is negligible (0.2 to 0.5 dB).

A TLR transformer may be used by itself to achieve a low-loss narrow-band impedance transformation, functioning like a λ/4 Q-section. It is different in nature from a Q-section, in that it is a true impedance transformer, while a Q section is an *impedance inverter*. The TLR transformer may be designed into the antenna system and is a useful tool when a narrow-band impedance transformation is needed.

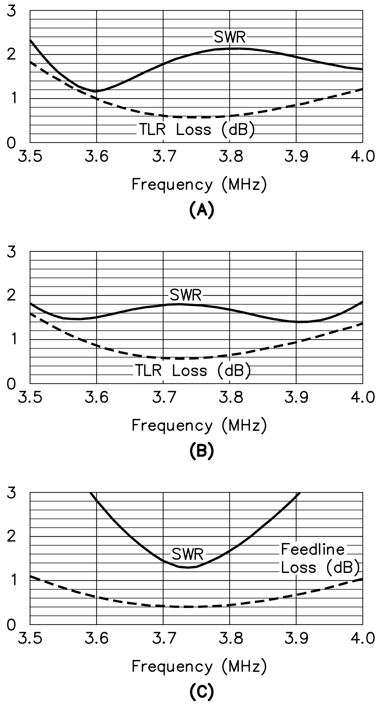

Fig 25—The SWR and loss of the TLR transformer antenna system. At A, the results are shown with the calculated line lengths. At B, the improvement after the lengths were optimized is shown in Table 1. At C, for comparison, SWR and loss for the same setup as at A, but with the stubs removed.

Table 1

Calculated and Optimized Parameters Using TLR Transformers

	Calculated	*Optimized*
Dipole resonant frequency (F_0)	3.742 MHz	3.710 MHz
Open stub (L_O)	10.9 feet	13.1 feet
Shorted stub (L_S)	18.8 feet	20.1 feet
Link (L_L)	100.4 feet	101.0 feet

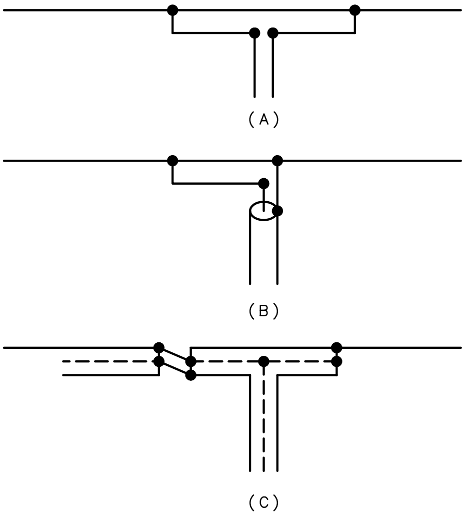

Fig 26—Dipole matching methods. At A, the T match; at B, the gamma match; at C, the coaxial resonator match.

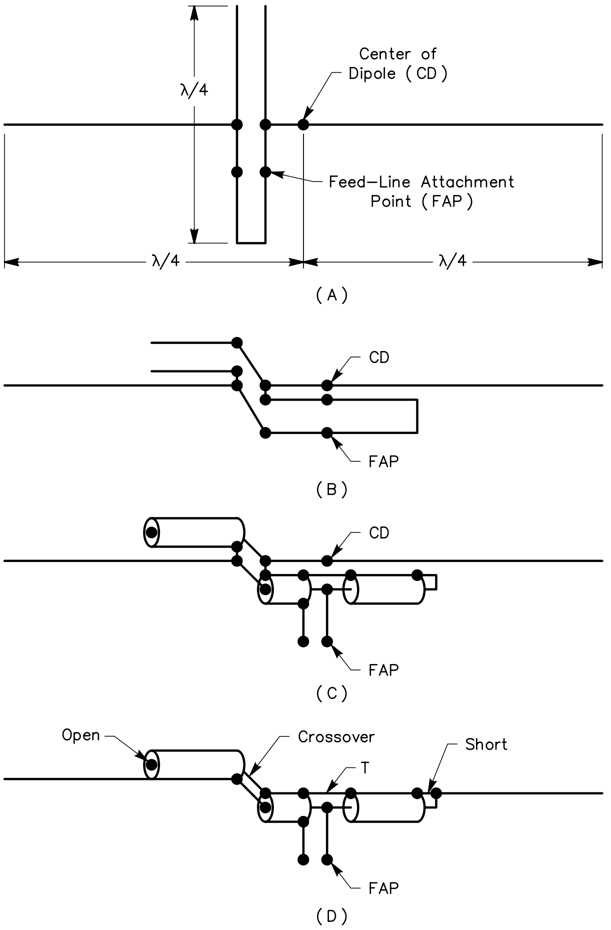

Fig 27—Evolution of the coaxial-resonator-match broadband dipole. At A, a TLR transformer is used to match the feed line to the off-center-fed dipole. The match and dipole are made collinear at B. At C, the balanced transmission-line TLR transformer of A and B is replaced by a coaxial version. Because the shield of the coax can serve as a part of the dipole radiator, the wire adjacent to the coax match may be eliminated, D.

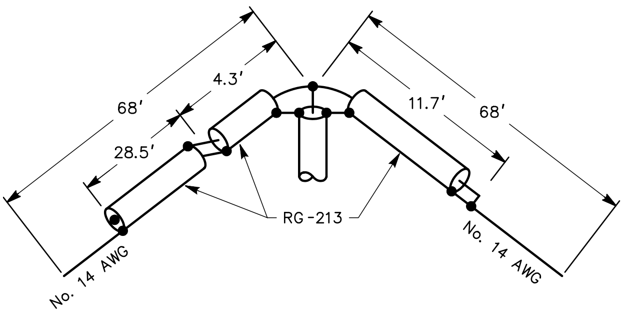

Fig 28—Dimensions for the 80-Meter MHz DX Special, an antenna optimized for the phone and CW DX portions of the 3.5-MHz band. The coax segment lengths total to one quarter-wavelength. The overall length is the same as that of a conventional inverted V dipole.

TRANSMISSION-LINE RESONATORS AS PART OF THE ANTENNA

The Coaxial-Resonator Match

This material is condensed from articles that appeared in April 1989 *QST* and *The ARRL Antenna Compendium, Vol 2*. The coaxial-resonator match, a concept based on the TLR transformer, performs the same function as the T match and the gamma match, that is, matching a transmission line to a resonant dipole. These familiar matching devices as well as the coaxial resonator match are shown in **Fig 26**. The coaxial resonator match has some similarity to the gamma match in that it allows connection of the shield of the coaxial feed line to the center of the dipole, and it feeds the dipole off center. The coaxial resonator match has a further advantage: It can be used to broadband the antenna system while providing an impedance match.

The coaxial resonator match is a resonant transformer made from a quarter-wave long piece of coaxial cable. **Fig 27** shows the evolution of the coaxial-resonator match. Now it becomes clear why coaxial cable is used for the quarter-wave TLR transformer—interaction between the dipole and the matching network is minimized. The effective dipole feed point is located at the crossover, which is an off-center feed point. In effect, the match is physically located "inside" the dipole. Currents flowing on the inside of the shield of the coax are associated with the resonator; currents flowing on the outside of the shield of the coax are the usual dipole currents. Skin effect provides a degree of isolation and allows the coax to perform its dual function. The wire extensions at each end make up the remainder of the dipole, making the overall length equal to one half-wavelength.

A useful feature of an antenna using the coaxial-resonator match is that the entire antenna is at the same dc potential as the feed-line potential, thereby avoiding charge buildup on the antenna. Hence, noise and the potential of lightning damage are reduced.

Fig 28 shows the detailed dimensions of an 80-meter inverted-V dipole using a coaxial-resonator match. This design provides a load SWR better than 1.6:1 from 3.5 to 3.825 MHz and has been named the *80-Meter DX Special*. The coax is an electrical quarter wavelength, has a short at one end, an open at the other end, a strategically placed crossover, and is fed at a tee junction. (The crossover is made by connecting the shield of one coax segment to the center conductor of the adjacent segment and by connecting the remaining center conductor and shield in a similar way.) The antenna is constructed as an inverted-V dipole with a 110° included angle and an apex at 60 feet. The measured SWR versus frequency is shown in **Fig 29**. Also in Fig 29

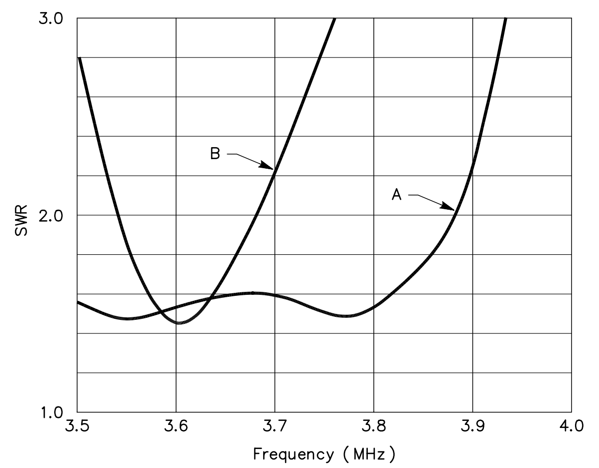

Fig 29—Measured SWR performance of the 80-Meter DX Special, curve A. Note the substantial broadbanding relative to a conventional uncompensated dipole, curve B.

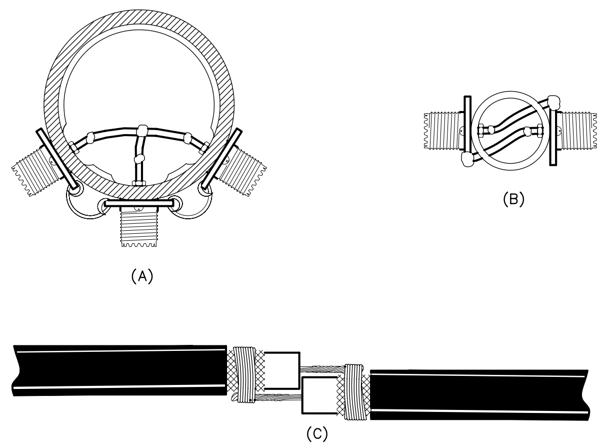

Fig 30—T and crossover construction. At A, a 2-inch PVC pipe coupling can be used for the T, and at B, a 1-inch coupling for the crossover. These sizes are the nominal inside diameters of the PVC pipe that is normally used with the couplings. The T could be made from standard UHF hardware (an M-358 T and a PL-258 coupler). An alternative construction for the crossover is shown at C, where a direct solder connection is made.

is the SWR characteristic for an uncompensated inverted-V dipole made from the same materials and positioned exactly like the broadband version.

The 80-Meter DX Special is made from RG-213 coaxial cable and #14 AWG wire, and is fed with 50-Ω coax. The coax forming the antenna should be cut so that the stub lengths of Fig 28 are within $^1/_2$ inch of the specified values. PVC plastic-pipe couplings and SO-239 UHF chassis connectors can be used to make the T and crossover connections, as shown in **Fig 30A** and 30B. Alternatively, a standard UHF T connector and coupler can be used for the T, and the crossover may be a soldered connection (Fig 30C). RG-213 is used because of its ready availability, physical strength, power handling capability, and moderate loss.

Cut the wire ends of the dipole about three feet longer than the lengths given in Fig 28. If there is a tilt in the SWR-frequency curve when the antenna is first built, it may be flattened to look like the shape given in Fig 29 by increasing or decreasing the wire length. Each end should be lengthened or shortened by the same amount.

A word of caution: If the coaxial cable chosen is not RG-213 or equivalent, the dimensions will have to be modified. The following cable types have about the same characteristic impedance, loss and velocity factor as RG-213 and could be substituted: RG-8, RG-8A, RG-10, RG-10A and RG-215. If the Q of the dipole is particularly high or the radiation resistance is unusually low because of different ground characteristics, antenna height, surrounding objects and so on, then different segment lengths will be required.

What is the performance of this broadband antenna relative to that of a conventional inverted-V dipole? Aside from the slight loss (about 0.75 dB at band edges, less elsewhere) because of the non-ideal matching network, the broadband version will behave essentially the same as a dipole cut for the frequency of interest. That is, the radiation patterns for the two cases will be virtually the same.

The Snyder Dipole

A commercially manufactured antenna utilizing the principles described in the Matching Network Design section is the Snyder dipole. Patented by Richard D. Snyder in late 1984 (see Bibliography), it immediately received much public attention through an article that Snyder published. Snyder's claimed performance for the antenna is a 2:1 SWR bandwidth of 20% with high efficiency.

The configuration of the Snyder antenna is like that of the crossed-double bazooka of Fig 12, with 25-Ω line used for the cross-connected resonators. In addition, the antenna is fed through a 2:1 balun, and exhibits a W-shaped SWR characteristic like that of Fig 13. The SWR at band center, based on information in the patent document, is 1.7 to 1. There is some controversy in professional circles regarding the claims for the Snyder antenna.

The Improved Crossed-Double Bazooka

The following has been condensed from an article by Reed Fisher, W2CQH, that appeared in *The ARRL Antenna Compendium Volume 2*. The antenna is shown in **Fig 31**. Note that the half-wave flat-top is constructed of sections of RG-58 coaxial cable. These sections of coaxial cable serve as quarter-wave shunt stubs that are essentially connected *in parallel* at the feed point in the crossed double-bazooka fashion. At an electrical quarter

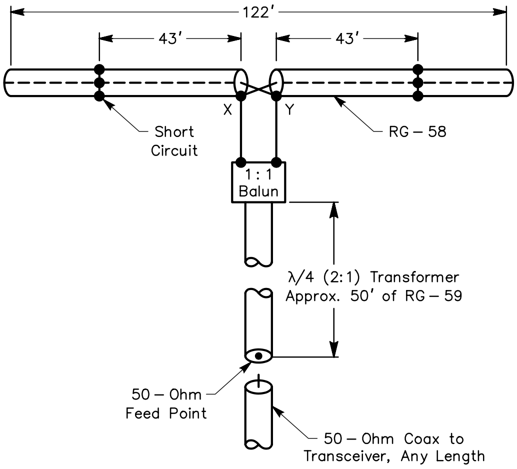

122′
43′
43′
X
Y
Short Circuit
RG – 58
1 : 1 Balun
λ/4 (2:1) Transformer
Approx. 50′ of RG – 59
50 – Ohm Feed Point
50 – Ohm Coax to Transceiver, Any Length

Fig 31—Details of the W2CQH broadband-matched 80-meter dipole.

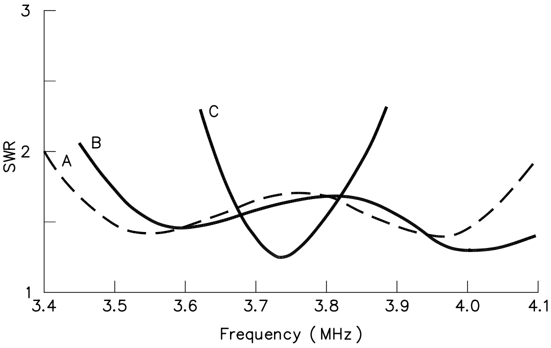

Fig 32—SWR curves for the improved double bazooka. Curve A, the theoretical curve with 50-Ω stubs and a λ/4, 75-Ω matching transformer. Curve B, measured response of the same antenna, built with RG-58 stubs and an RG- 59 transformer. Curve C, measurements from a dipole without broadbanding. Measurements were made at W2CQH with the dipole horizontal at 30 feet.

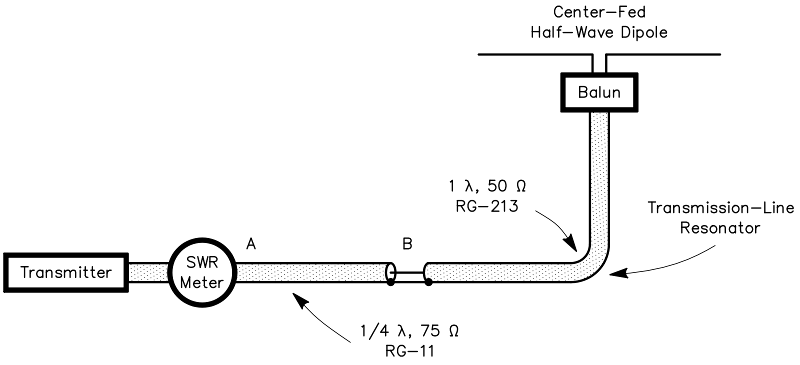

Center–Fed Half–Wave Dipole
Balun
1 λ, 50 Ω RG–213
Transmission–Line Resonator
A
B
Transmitter
SWR Meter
1/4 λ, 75 Ω RG–11

Fig 33—This simple broadband antenna system resembles a conventional 80-meter dipole except for the λ/4-wavelength 75-Ω segment. The lengths of the Q-section, TLR (including balun) and dipole are 43.3 feet, 170.5 feet, and 122.7 feet, respectively.

wavelength (43 feet) from each side of the feed point X-Y, the center conductor is shorted to the braid of the coaxial cable.

The parallel stubs provide reactance compensation. The necessary impedance transformation at the antenna feed point is provided by the quarter-wave Q-section constructed of a 50-foot section of 75-Ω coaxial cable (RG-59). A W2DU 1:1 current balun is used at the feed point. See **Fig 32** for the SWR versus frequency for this antenna with and without the broadband matching network.

This antenna is essentially that of the crossed double bazooka antenna shown in Fig 12, with the addition of the 75-Ω Q-section. The improvement in match bandwidth is substantial. The antenna at W2CQH is straight and nearly horizontal with an average height of about 30 feet.

TRANSMISSION-LINE RESONATORS AS PART OF THE FEED LINE

A Simple Broadband Dipole for 80 Meters

This system was described in an article in September 1993 *QST*. It has the advantage that the radiator is the same as that of a conventional half-wave wire dipole. Thus the antenna is light weight, easy to support and has small wind and ice loading. The broadband matching network is integrated into the feed line.

The system is shown in **Fig 33**. It is a variation of the example shown in the Matching Network Design section of this chapter as shown in Fig 17. The TLR is a 1λ length of RG-213 50-Ω coax and the impedance transformation is accomplished with a 75-Ω Q-section

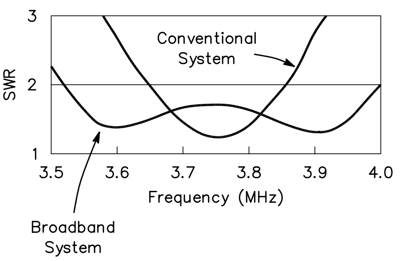

Fig 34—Measured SWR versus frequency for the broadband and conventional antenna systems.

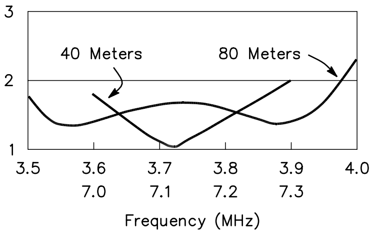

Fig 35—Measured SWR for the 80- and 40-meter multiband antenna system.

made of RG-11. The antenna is an inverted-V with a 140° included angle and an apex height of 60 feet. The wire size is #14, but is not critical. The balun is a W2DU 1:1 current type. It easily handles the legal power limit over the entire 80-meter band. The measured SWR is shown in **Fig 34**.

A unique advantage of this antenna is that by paralleling a 40-meter dipole at the feed point and sharing the feed line, operation on both 80 and 40 meters is possible. The reason for this is that the lengths of the Q-section and the TLR are multiples of a half-wavelength on 40 meters. To minimize the interaction, the two dipoles should be spaced from each other away from the feed point. First tune the 80-meter antenna and then the 40-meter one. **Fig 35** shows the result of adding a 40-meter dipole to the system shown in Fig 33. Each 40-meter dipole leg is 34.4 feet long. Note that the SWR on 80 meters changes very little compared to Fig 34. No change was made to the 80-meter dipole or to the transmission line.

80-Meter Dipole with the TLR Transformer

The total length of the feed line in the system just described above in Fig 33 is quite long, about 214 feet. For installations with shorter runs, the TLR transformer concept may be employed to advantage. An example provided in "Optimizing the 80-Meter Dipole" in *The ARRL Antenna Compendium, Vol 4* is shown in **Fig 36**. Here the TLR is $^{3}/_{4}$ λ long overall and is made from RG-213 50-Ω cable. The design applies for an 80-meter horizontal dipole, 80 feet above average ground, with an assumed antenna F_0, R_A, and Q_A of 3.72 MHz, 92 Ω and 9, respectively. (See Figs 3 and 4). The lengths must be revised for other dipole parameters.

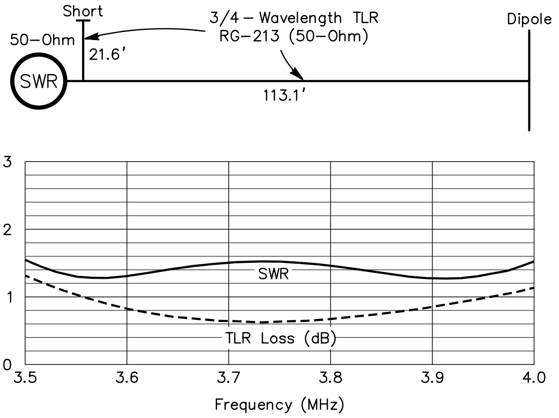

Fig 36—Broadbanding with the TLR transformer. The dimensions shown apply for an 80-meter horizontal dipole 80 feet above average ground.

For this design, the feed line is 113.1 feet long and the shorted stub at the transmitter end is 21.6 feet long. You can add any length of 50-Ω cable between the transmitter and the junction of the shorted stub and the feed line. The shorted stub gives a dc path to ground for both sides of the dipole, preventing charge buildup and some lightning protection. For multiband operation with an antenna tuner, the stub should be removed.

The general application of the TLR transformer concept has a tap at both the transmitter end and a tap at the antenna end. (See Fig 24.) In this case, the concept was applied only at the transmitter end, since very little impedance transformation was required at the antenna end. The location of the TLR transformer is not obvious. Consider a $^{3}/_{4}$ λ TLR with a short at one end, and an open at the other end where the antenna is connected. The generator for no impedance transformation would be located λ/4 (43.6 feet @ 3.72 MHz) from the short. By connecting the transmitter 21.6 feet from the short, the required impedance transformation for a transmitter optimized for 50-Ω loads is achieved.

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

A. Bloom, "Once More with the 80-Meter Dipole," Technical Correspondence, *QST*, Jun 1985, p 42.

R. E. Fisher, "A Simple, Broadband 80-Meter Dipole Antenna," *The ARRL Antenna Compendium, Vol 2* (Newington, CT: ARRL, 1989), pp 119-123.

R. M. Fano, "Theoretical Limitations on the Broadband Matching of Arbitrary Impedances," *Journal of the Franklin Inst*, Jan 1950, pp 57-83, and Feb 1950, pp 139-155.

J. Hall, "The Search for a Simple, Broadband 80-Meter Dipole," *QST*, Apr 1983, pp 22-27.

J. Hall, "Maxcom Antenna Matcher and Dipole Cable Kit," Product Review, *QST*, Nov 1984, pp 53-54.

R. C. Hansen, "Evaluation of the Snyder Dipole," *IEEE Trans on Antennas and Prop*, Vol AP-35, No. 2, Feb 1987, pp 207-210.

A. Harbach, "Broadband 80-Meter Antenna," *QST*, Dec 1980, pp 36-37 (presentation of a cage antenna).

S. Leslie, "Broadbanding the Elevated, Inverse-Fed Ground Plane Antenna," *The ARRL Antenna Compendium, Vol 6* (Newington, CT: 1999), pp 209-211.

R. D. Snyder, "The Snyder Antenna," *RF Design*, Sep/Oct 1984, pp 49-51.

R. D. Snyder, "Broadband Antennae Employing Coaxial Transmission Line Sections," United States Patent no. 4,479,130, issued Oct 23, 1984.

C. Whysall, "The 'Double Bazooka' Antenna," *QST*, Jul 1968, pp 38-39.

F. Witt, "A Simple Broadband Dipole for 80 Meters," *QST*, Sept 1993, pp 27-30, 76.

F. J. Witt, "Broadband Dipoles—Some New Insights," *QST*, Oct 1986, pp 27-37.

F. Witt, "Broadband Matching with the Transmission Line Resonator," *The ARRL Antenna Compendium, Vol 4*, (Newington, CT: 1995), pp 30-37.

F. Witt, "How to Design Off-Center-Fed Multiband Wire Antennas Using that Invisible Transformer in the Sky," *The ARRL Antenna Compendium, Vol 3,* (Newington, CT: ARRL, 1992), p 74.

F. Witt, "Match Bandwidth of Resonant Antenna Systems," *QST*, Oct 1991, pp 21-25

F. Witt, "Optimizing the 80-Meter Dipole," *The ARRL Antenna Compendium, Vol 4* (Newington, CT: 1995), pp 38-48.

F. Witt, "Optimum Lossy Broadband Matching Networks for Resonant Antennas," *RF Design*, Apr 1990, pp 44-51 and Jul 1990, p 10.

F. Witt, "The Coaxial Resonator Match and the Broadband Dipole," *QST*, Apr 1989, pp 22-27.

F. Witt, "The Coaxial Resonator Match, *The ARRL Antenna Compendium, Vol 2* (Newington, CT: ARRL, 1989), pp 110-118.

Chapter 10

Log Periodic Arrays

This chapter was contributed by L. B. Cebik, W4RNL. The *Log Periodic Dipole Array* (*LPDA*) is one of a family of frequency-independent antennas. Alone, an LPDA forms a directional antenna with relatively constant characteristics across a wide frequency range. It may also be used with parasitic elements to achieve specific characteristic within a narrow frequency range. Common names for such hybrid arrays are the *log-cell Yagi* or the *Log-Yag*. We shall look at the essential characteristics of both types of arrays in this chapter.

The LPDA is the most popular form of log-periodic systems, which also include zig-zag, planar, trapezoidal, slot, and V forms. The appeal of the LPDA version of the log periodic antenna owes much to its structural similarity to the Yagi-Uda parasitic array. This permits the construction of directional LPDAs that can be rotated—at least within the upper HF and higher frequency ranges. Nevertheless, the LPDA has special structural as well as design considerations that distinguish it from the Yagi. A number of different construction techniques for both wire and tubular elements are illustrated later in this chapter.

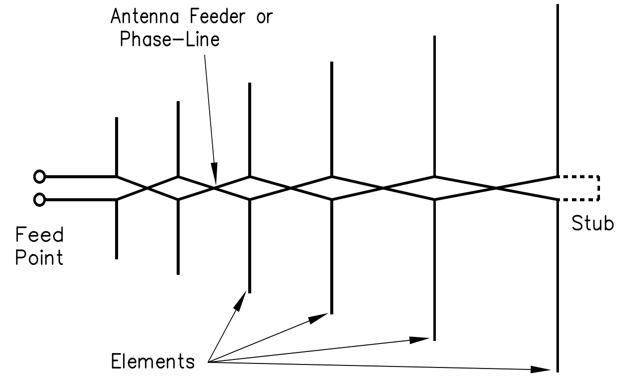

Fig 1—The basic components of a log periodic dipole array (LPDA). The forward direction is to the left in this sketch. Many variations of the basic design are possible.

The LPDA in its present form derives from the pioneering work of D. E. Isbell at the University of Illinois in the late 1950s. Although you may design LPDAs for large frequency ranges—for example, from 3 to 30 MHz or a little over 3 octaves—the most common LPDA designs that radio amateurs use are limited to a one-octave range, usually from 14 to 30 MHz. Amateur designs for this range tend to consist of linear elements. However, experimental designs for lower frequencies have used elements shaped like inverted Vs, and some versions use vertically oriented $^1/_4$ λ elements over a ground system.

Fig 1 shows the parts of a typical LPDA. The structure consists of a number of linear elements, the longest of which is approximately $^1/_2$ λ long at the lowest design frequency. The shortest element is usually about $^1/_2$ λ long at a frequency well above the highest operating frequency. The antenna feeder, also informally called the *phase-line*, connects the center points of each element in the series, with a phase reversal or *cross-over* between each element. A stub consisting of a shorted length of parallel feed line is often added at the back of an LPDA.

The arrangement of elements and the method of feed yield an array with relatively constant gain and front-to-back ratio across the designed operating range. In addition, the LPDA exhibits a relatively constant feed-point impedance, simplifying matching to a transmission line.

BASIC DESIGN CONSIDERATIONS

For the amateur designer, the most fundamental facets of the LPDA revolve around three interrelated design variables: α (alpha), τ (tau), and σ (sigma). Any one of the three variables may be defined by reference to the other two.

Fig 2 shows the basic components of an LPDA. The angle α defines the outline of an LPDA and permits every dimension to be treated as a radius or the consequence of a radius (R) of a circle. The most basic structural dimensions are the element lengths (L), the distance (R) of each element from the apex of angle α, and the distance between elements (D). A single design constant, τ, defines all of these relationships in the following manner:

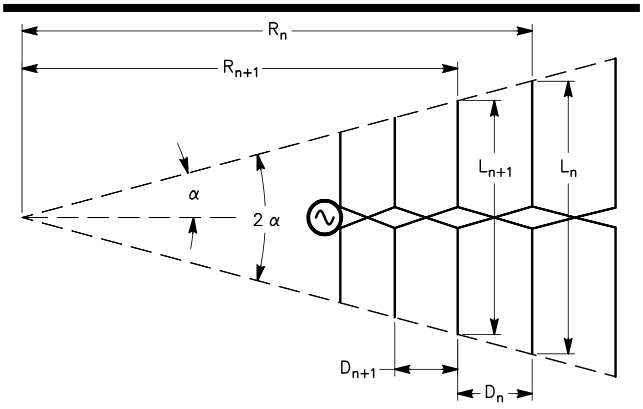

Fig 2—Some fundamental relationships that define an array as an LPDA. See the text for the defining equations.

$$\tau = \frac{R_{n+1}}{R_n} = \frac{D_{n+1}}{D_n} = \frac{L_{n+1}}{L_n} \tag{Eq 1}$$

where element n and n+1 are successive elements in the array working toward the apex of angle α. The value of τ is always less than 1.0, although effective LPDA design requires values as close to 1.0 as may be feasible.

The variable τ defines the relationship between successive element spacings but it does not itself determine the initial spacing between the longest and next longest elements upon which to apply τ successively. The initial spacing also defines the angle α for the array. Hence, we have two ways to determine the value of σ, the relative spacing constant:

$$\sigma = \frac{1 - \tau}{4 \tan \alpha} = \frac{D_n}{2L_n} \tag{Eq 2}$$

where D_n is the distance between any two elements of the array and L_n is the length of the longer of the two elements. From the first of the two methods of determining the value of σ, we may also find a means of determining α when we know both τ and σ.

For any value of τ, we may determine the optimal value of σ:

$$\sigma_{opt} = 0.243\tau - 0.051 \tag{Eq 3}$$

The combination of a value for τ and its corresponding optimal value of σ yields the highest performance of which an LPDA is capable. For values of τ from 0.80 through 0.98, the value of optimal σ varies from 0.143 to 0.187, in increments of 0.00243 for each 0.01 change in τ. However, using the optimal value of σ usually yields a total array length that is beyond ham construction or tower/mast support capabilities. Consequently, amateur LPDAs usually employ compromise values of τ and σ that yield lesser but acceptable performance.

For a given frequency range, increasing the value of τ increases both the gain and the number of required elements. Increasing the value of σ increases both the gain and the overall boom length. A τ of 0.96—which approaches the upper maximum recommended value for τ—yields an optimal σ of about 0.18, and the resulting array grows to over 100 feet long for the 14 to 30 MHz range. The maximum free space gain is about 11 dBi, with a front-to-back ratio that approaches 40 dB. Normal amateur practice, however, uses values of τ from about 0.88 to 0.95 and values of σ from about 0.03 to 0.06.

Standard design procedures usually set the length of the rear element for a frequency about 7% lower than the lowest design frequency and use the common dipole formula ($L_{feet} = 468/f_{MHz}$) to determine its length (5% lower than a free-space half wavelength, where $L_{feet} = 493.56/f_{MHz}$). The upper frequency limit of the design is ordinarily set at about 1.3 times the highest design frequency. Since τ and σ set the increment between successive element lengths, the number of elements becomes a function of when the shortest element reaches the dipole length for the adjusted upper frequency.

The adjusted upper frequency limit results from the behavior of LPDAs with respect to the number of active elements. See **Fig 3**, which shows an edge view of a 10-element LPDA for 20 through 10 meters. The vertical lines represent the peak relative current magnitude for each element at the specified frequency. At 14 MHz, virtually every element of the array shows a significant cur-

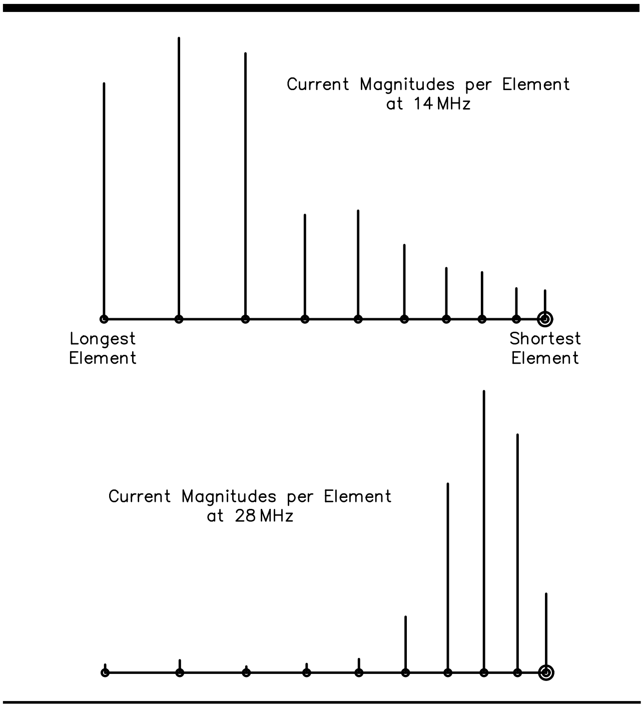

Fig 3—The relative current magnitude on the elements of an LPDA at the lowest and highest operating frequencies for a given design. Compare the number of "active" elements, that is, those with current levels at least 1/10 of the highest level.

rent magnitude. However, at 28 MHz, only the forward 5 elements carry significant current. Without the extended design range to nearly 40 MHz, the number of elements with significant current levels would be severely reduced, along with upper frequency performance.

The need to extend the design equations below the lowest proposed operating frequency varies with the value of τ. In **Fig 4**, we can compare the current on the rear elements of two LPDAs, both with a σ value of 0.04. The upper design uses a τ of 0.89, while the lower design uses a value of 0.93. The most significant current-bearing element moves forward with increases in τ, reducing (but not wholly eliminating) the need for elements whose lengths are longer than a dipole for the lowest operating frequency.

LPDA Design and Computers

Originally, LPDA design proceeded through a series of design equations intended to yield the complete specifications for an array. More recent techniques available to radio amateurs include basic LPDA design software and antenna modeling software. One good example of LPDA design software is *LPCAD28* by Roger Cox, WB0DGF. A copy of this freeware program is on the CD-ROM that accompanies this volume. The user begins by specifying the lowest and highest frequencies in the design. He then enters either his selected values for τ and σ or his choices for the number of elements and the total length of the array. With this and other input data, the program provides a table of element lengths and spacings, using the adjusted upper and lower frequency limits described earlier.

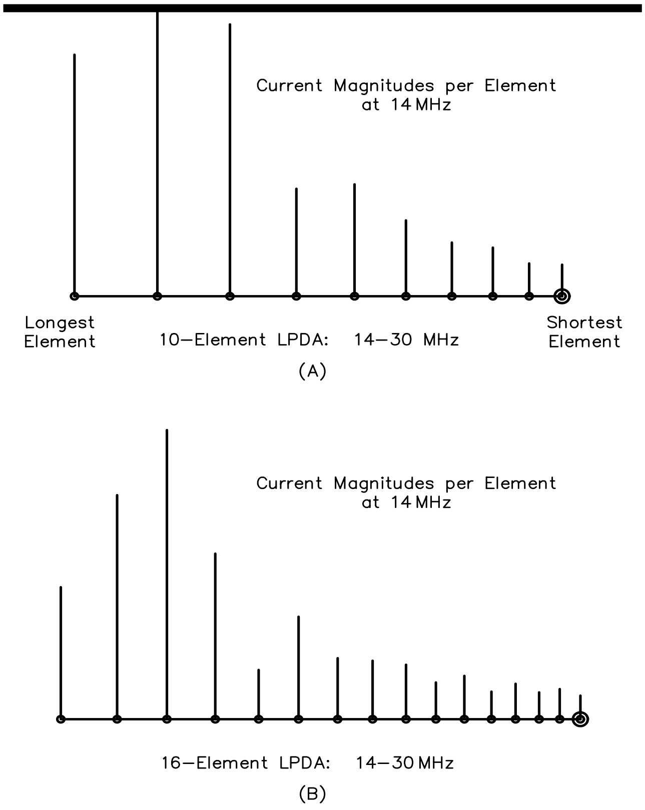

Fig 4—Patterns of current magnitude at the lowest operating frequency of two different LPDA designs: a 10-element low-τ design and a 16-element higher-τ design.

The program also requests the diameters of the longest and shortest elements in the array, as well as the diameter of average element. From this data, the program calculates a recommended value for the phase-line connecting the elements and the approximate resistive value of the input impedance. Among the additional data that *LPCAD28* makes available is the spacing of conductors to achieve the desired characteristic impedance of the phase-line. These conductors may be round—as we would use for a wire phase-line—or square—as we might use for double-boom construction.

An additional vital output from *LPCAD28* is the conversion of the design into antenna modeling input files of several formats, including versions for *AO* and *NEC4WIN* (both *MININEC*-based programs), and a version in the standard *.NEC format usable by many implementations of *NEC-2* and *NEC-4*, including *NECWin Plus*, *GNEC*, and *EZNEC Pro*. Every proposed LPDA design should be verified and optimized by means of antenna modeling, since basic design calculations rarely provide arrays that require no further work before construction. A one-octave LPDA represents a segment of an arc defined by α that is cut off at both the upper and the lower frequency limits. Moreover, some of the design equations are based upon approximations and do not completely predict LPDA behavior. Despite these limitations, most of the sample LPDA designs shown later in this chapter are based directly upon the fundamental calculations. Therefore, the procedure will be outlined in detail before turning to hybrid log-cell Yagi concepts.

Modeling LPDA designs is most easily done on a version of *NEC*. The transmission line (TL) facility built into *NEC-2* and *NEC-4* alleviates the problem of modeling the phase-line as a set of physical wires, each section of which has a set of constraints in *MININEC* at the right-angle junctions with the elements. Although the *NEC* TL facility does not account for losses in the lines, the losses are ordinarily low enough to neglect.

NEC models do require some careful construction to obtain the most accurate results. Foremost among the cautions is the need for careful segmentation, since each element has a different length. The shortest element should have about 9 or 11 segments, so that it has sufficient segments at the highest modeling frequency for the design. Each element behind the shortest one should have a greater number of segments than the preceding element by the inverse of the value of τ. However, there is a further limitation. Since the transmission line is at the center of each element, *NEC* elements should have an odd number of segments to hold the phase-line centered. Hence, each segmentation value calculated from the inverse of τ must be rounded up to the nearest odd integer.

Initial modeling of LPDAs in *NEC-2* should be done with uniform-diameter elements, with any provision for stepped-diameter element correction turned off. Since these correction factors apply only to elements within about 15% of dipole resonance at the test frequency, models with stepped-diameter elements will correct for only a few elements at any test frequency. The resulting combination of corrected and uncorrected elements will not yield a model with assured reliability.

Once one has achieved a satisfactory model with uniform-diameter elements, the modeling program can be used to calculate stepped-diameter substitutes. Each uniform-diameter element, when extracted from the larger array, will have a resonant frequency. Once this frequency is determined, the stepped-diameter element to be used in final construction can be resonated to the same frequency. Although *NEC-4* handles stepped diameter elements with much greater accuracy than *NEC-2*, the process just described is also applicable to *NEC-4* models for the greatest precision.

LPDA Behavior

Although LPDA behavior is remarkably uniform over a wide frequency range compared to narrow-band designs, such as the Yagi-Uda array, it nevertheless exhibits very significant variations within the design range. **Fig 5** shows several facets of these behaviors. Fig 5 shows the free-space gain for three LPDA designs using 0.5-inch diameter aluminum elements. The designations for each model list the values of τ (0.93, 0.89, and 0.85) and of σ (0.02, 0.04, and 0.06) used to design each array. The resultant array lengths are listed with each designator. The total number of elements varies from 16 for "9302" to 10 for "8904" to 7 for "8504."

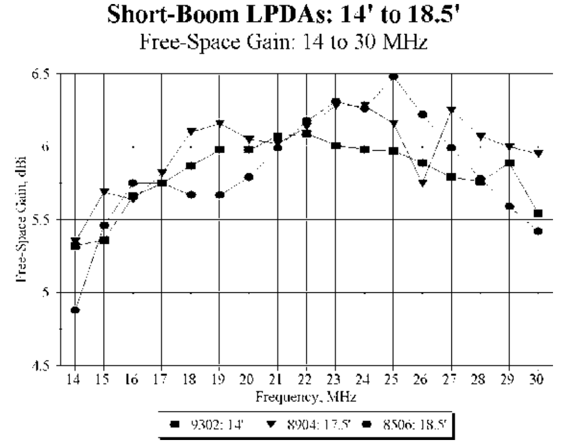

Fig 5—The modeled free-space gain of 3 relatively small LPDAs of different design. Note the relationship of the values of τ and of σ for these arrays with quite similar performance across the 14-30 MHz span.

First, the gain is never uniform across the entire frequency span. The gain tapers off at both the low and high ends of the design spectrum. Moreover, the amount of gain undulates across the spectrum, with the number of peaks dependent upon the selected value of τ and the resultant number of elements. The front-to-back ratio tends to follow the gain level. In general, it ranges from less than 10 dB when the free-space gain is below 5 dBi to over 20 dB as the gain approaches 7 dBi. The front-to-back ratio may reach the high 30s when the free-space array gain exceeds 8.5 dBi. Well-designed arrays, especially those with high values of τ and σ, tended to have well-controlled rear patterns that result in only small differences between the 180° front-to-back ratio and the averaged front-to-rear ratio.

Since array gain is a mutual function of both τ and σ, average gain becomes a function of array length for any given frequency range. Although the gain curves in Fig 5 interweave, there is little to choose among them in terms of average gain for the 14 to 18-foot range of array lengths. Well-designed 20- to 10-meter arrays in the 30-foot array length region are capable of about 7 dBi free-space gain, while 40-foot arrays for the same frequency range can achieve about 8 dBi free-space gain.

Exceeding an average gain of 8.5 dBi requires at least a 50-foot array length for this frequency range. Long arrays with high values of τ and σ also tend to show smaller excursions of gain and of front-to-back ratio in the overall curves. In addition, high-τ designs tend to show higher gain at the low frequency end of the design spectrum.

The frequency sweeps shown in Fig 5 are widely spaced at 1 MHz intervals. The evaluation of a specific design for the 14 to 30-MHz range should decrease the interval between check points to no greater than 0.25 MHz in order to detect frequencies at which the array may show a *performance weakness*. Weaknesses are frequency regions in the overall design spectrum at which the array shows unexpectedly lower values of gain and front-to-back ratio. In Fig 5 note the unexpected decrease in gain of model "8904" at 26 MHz. The other designs also have weak points, but they fall between the frequencies sampled.

In large arrays, these regions may be quite small and may occur in more than one frequency region. The weakness results from the harmonic operation of longer elements to the rear of those expected to have high current levels. Consider a 7-element LPDA about 12.25-foot long for 14 to 30 MHz using 0.5-inch aluminum elements. At 28 MHz, the rear elements operate in a harmonic mode, as shown by the high relative current magnitude curves in **Fig 6**. The result is a radical decrease in gain, as shown in the "No Stub" curve of **Fig 7**. The front-to-back ratio also drops as a result of strong radiation from the long

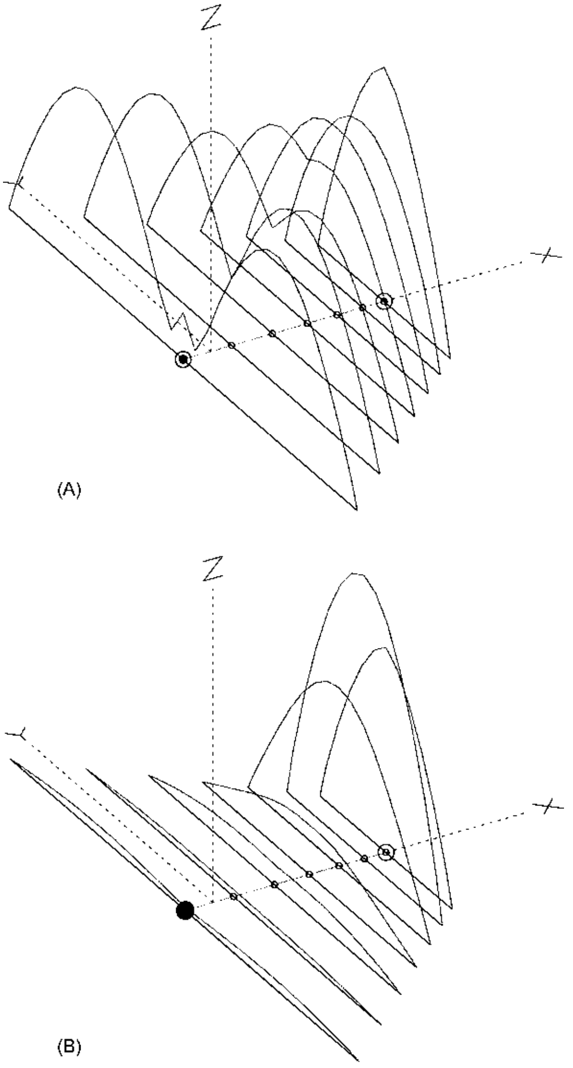

Fig 6—The relative current magnitude on the elements of model "8504" at 28 MHz without and with a stub. Note the harmonic operation of the rear elements before a stub is added to suppress such operation.

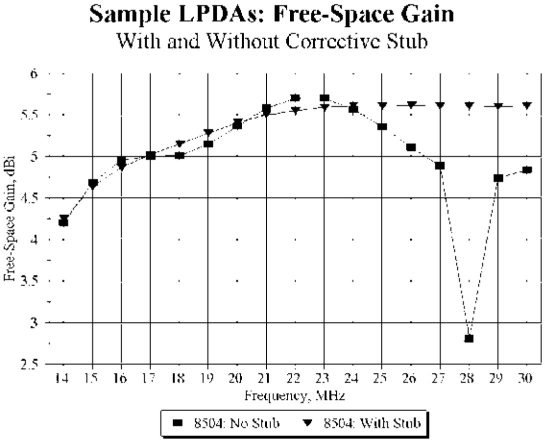

Fig 7—A graph of the gain of model "8504" showing the frequency region in which a "weakness" occurs and its absence once a suitable stub is added to the array.

elements to the rear of the array.

Early designs of LPDAs called for terminating transmission-line stubs as standard practice to help eliminate such weak spots in frequency coverage. In contemporary designs, their use tends to be more specific for eliminating or moving frequencies that show gain and front-to-back weakness. (Stubs have the added function of keeping both sides of each element at the same level of static charge or discharge.) The model dubbed "8504" was fitted (by trial and error) with an 18-inch shorted stub of 600-Ω transmission line. As Fig 6B shows, the harmonic operation of the rear elements is attenuated. The "stub" curve of Fig 7 shows the smoothing of the gain curve for the array throughout the upper half of its design spectrum. In some arrays showing multiple weaknesses, a single stub may not eliminate all of them. However, it may move the weaknesses to unused frequency regions. Where full-spectrum operation of an LPDA is necessary, additional stubs located at specific elements may be needed.

Most LPDA designs benefit (with respect to gain and front-to-back ratio) from the use of larger-diameter elements. Elements with an average diameter of at least 0.5-inch are desirable in the 14 to 30 MHz range. However, standard designs usually presume a constant element length-to-diameter ratio. In the case of *LPCAD28*, this ratio is about 125:1, which assumes an even larger diameter. To achieve a relatively constant length-to-diameter ratio in the computer models, you can set the diameter of the shortest element in a given array design and then increase the element diameter by the inverse of τ for each succeeding longer element. This procedure is often likely to result in unreasonably large element diameters for the longest elements, relative to standard amateur construction practices.

Since most amateur designs using aluminum tubing for elements employ stepped-diameter (tapered) elements, roughly uniform element diameters will result unless the LPDA mechanical design tries to lighten the elements at the forward end of the array. This practice may not be advisable, however. Larger elements at the high end of the design spectrum often counteract (at least partially) the natural decrease in high-frequency gain and show improved performance compared to smaller diameter elements.

An alternative construction method for LPDAs uses wire throughout. At every frequency, single-wire elements reduce gain relative to larger-diameter tubular elements. An alternative to tubular elements appears in **Fig 8**. For each element of a tubular design, there is a roughly equivalent 2-wire element that may be substituted. The

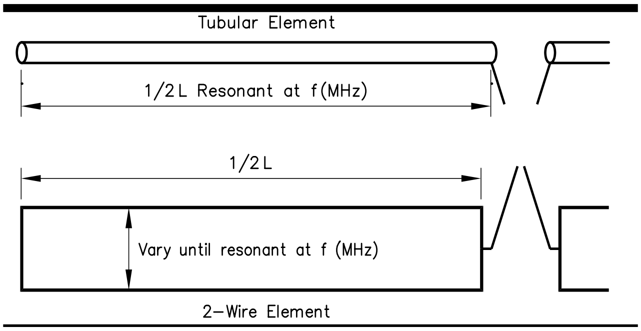
Tubular Element
1/2 L Resonant at f(MHz)
1/2 L
Vary until resonant at f (MHz)
2–Wire Element

Fig 8—A substitute for a large-diameter tubular element composed of two wires shorted at both the outer ends and at the center junction with the phase-line.

spacing between the wire is determined by taking one of the modeled tubular elements and finding its resonant frequency. A two-wire element of the same length is then constructed with shorts at the far ends and at the junctions with the phase-line. The separation of the two wires is adjusted until the wire element is resonant at the same frequency as the original tubular element. The required separation will vary with the wire chosen for the element. Models used to develop these substitutes must pay close attention to segmentation rules for *NEC* due to the short length of segments in the end and center shorts, and to the need to keep segment junctions as exactly parallel as possible with close-spaced wires.

Feeding and Constructing the LPDA

Original design procedures for LPDAs used a single, ordinarily fairly high, characteristic impedance for the phase-line (antenna feeder). Over time, designers realized that other values of impedance for the phase line offered both mechanical and performance advantages for LPDA performance. Consequently, for the contemporary designer, phase-line choice and construction techniques are almost inseparable considerations.

High-impedance phase lines (roughly 200 Ω and higher) are amenable to wire construction similar to that used with ordinary parallel-wire transmission lines. They require careful placement relative to a metal boom used to support individual elements (which themselves must be insulated from the support boom). Connections also require care. If the phase-line is given a half-twist between each element, the construction of the line must ensure constant spacing and relative isolation from metal supports to maintain a constant impedance and to prevent shorts.

Along with the standard parallel-wire line, shown in **Fig 9A**, there are a number of possible LPDA structures using booms. The booms serve both to support the elements and to create relatively low-impedance (under 200 Ω) phase-lines. Fig 9B shows the basics of a twin circular tubing boom with the elements cross-supported by insulated rods. Fig 9C shows the use of square tubing with the elements attached directly to each tube by through-bolts. Fig 9D illustrates the use of L-stock, which may be practical at VHF frequencies. Each of these sketches is incomplete, however, since it omits the necessary stress analyses that determine the mechanical feasibility of a structure for a given LPDA project.

The use of square boom material requires some adjustment when calculating the characteristic impedance of the phase-line. For conductors with a circular cross-section,

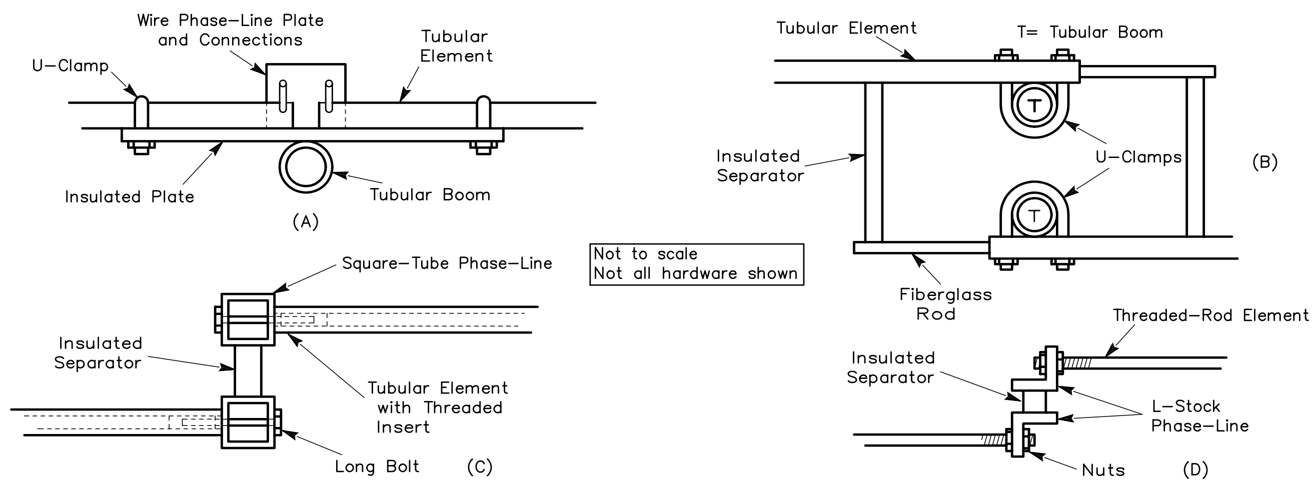
Wire Phase–Line Plate and Connections
Tubular Element
U–Clamp
Insulated Plate
Tubular Boom
(A)
Tubular Element
T= Tubular Boom
Insulated Separator
U–Clamps
(B)
Fiberglass Rod
Not to scale
Not all hardware shown
Square–Tube Phase–Line
Insulated Separator
Tubular Element with Threaded Insert
Long Bolt
(C)
Threaded–Rod Element
Insulated Separator
L–Stock Phase–Line
Nuts
(D)

Fig 9—Four (of many) possible construction techniques, shown from the array end. In A, an insulated plate supports and separates the wires of the phase-line, suitable with wire or tubular elements. A dual circular boom phase-line also supports the elements, which are cross-supported for boom stability. Square tubing is used in C, with the elements joined to the boom/phase-line with through-bolts and an insert in each half element. The L-stock shown in D is useful for lighter VHF and UHF arrays.

$$Z_0 = 120 \cosh^{-1} \frac{D}{d} \qquad \text{(Eq 4)}$$

where D is the center-to-center spacing of the conductors and d is the outside diameter of each conductor, both expressed in the same units of measurement. Since we are dealing with closely spaced conductors, relative to their diameters, the use of this version of the equation for calculating the characteristic impedance (Z_0) is recommended. For a square conductor,

$$d \approx 1.18\ w \qquad \text{(Eq 5)}$$

where d is the approximate equivalent diameter of the square tubing and w is the width of the tubing across one side. Thus, for a given spacing, a square tube permits you to achieve a lower characteristic impedance than round conductors. However, square tubing requires special attention to matters of strength, relative to comparable round tubing.

Electrically, the characteristic impedance of the LPDA phase-line tends to influence other performance parameters of the array. Decreasing the phase-line Z_0 also decreases the feed-point impedance of the array. For small designs with few elements, the decrease is not fully matched by a decrease in the excursions of reactance. Consequently, using a low impedance phase-line may make it more difficult to achieve a 2:1 or less SWR for the entire frequency range. However, higher-impedance phase-lines may result in a feed-point impedance that requires the use of an impedance-matching balun.

Decreasing the phase-line Z_0 also tends to increase LPDA gain and front-to-back ratio. There is a price to be paid for this performance improvement—weaknesses at specific frequency regions become much more pronounced with reductions in the phase-line Z_0. For a specific array you must weigh carefully the gains and losses, while employing one or more transmission line stubs to get around performance weaknesses at specific frequencies.

Depending upon the specific values of τ and σ selected for a design, you can sometimes select a phase-line Z_0 that provides either a 50-Ω or a 75-Ω feed-point impedance, holding the SWR under 2:1 for the entire design range of the LPDA. The higher the values of τ and σ for the design, the lower the reactance and resistance excursions around a central value. Designs using optimal values of σ with high values of τ show a very slight capacitive reactance throughout the frequency range. Lower design values obscure this phenomenon due to the wide range of values taken by both resistance and reactance as the frequency is changed.

At the upper end of the frequency range, the source resistance value decreases more rapidly than elsewhere in the design spectrum. In larger arrays, this can be overcome by using a variable Z_0 phase-line for approximately the first 20% of the array length. This technique is, however, difficult to implement with anything other than wire phase-lines. Begin with a line impedance about half of the final value and increase the wire spacing evenly until it reaches its final and fixed spacing. This technique can sometimes produce smoother impedance performance across the entire frequency span and improved high-frequency SWR performance.

Designing an LPDA requires as much attention to designing the phase-line as to element design. It is always useful to run models of the proposed design through several iterations of possible phase-line Z_0 values before freezing the structure for construction.

Special Design Corrections

The curve for the sample 8504 LPDA in Fig 7 revealed several deficiencies in standard LPDA designs. The weakness in the overall curve was corrected by the use of a stub to eliminate or move the frequency at which rearward elements operated in a harmonic mode. In the course of describing the characteristics of the array, we have noted several other means to improve performance. Fattening elements (either uniformly or by increasing their diameter in step with τ) and reducing the characteristic impedance of the phase-line are capable of small improvements in performance. However, they cannot wholly correct the tendency of the array gain and front-to-back ratio to fall off at the upper and lower limits of the LPDA frequency range.

One technique sometimes used to improve performances near the frequency limits is to design the LPDA for upper and lower frequency limits much higher and lower than the frequencies of use. This technique unnecessarily increases the overall size of the array and does not eliminate the downward performance curves. Increasing the values of τ and σ will usually improve perfor-

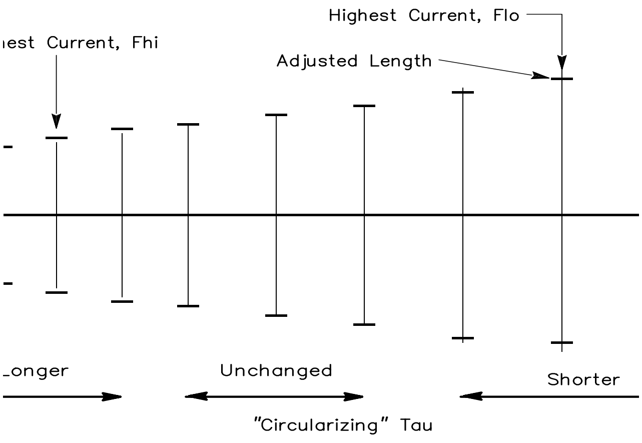

Fig 10—A before and after sketch of an LPDA, showing the original lengths of the elements and their adjustments from diminishing the value of τ at both ends of the array. See the text for the amount of change applicable to each element.

mance at no greater cost in size than extending the frequency range. Increasing the value of τ is especially effective in improving the low frequency performance of an LPDA.

Working within the overall size limits of a standard design, one may employ a technique of *circularizing* the value of τ for the rear-most and forward-most elements. See **Fig 10**, which is not to-scale relative to overall array length and width. Locate (using an antenna-modeling program) the element with the highest current at the lowest operating frequency, and the element with the highest current at the highest operating frequency. The adjustments to element lengths may begin with these elements or—at most—one element further toward the array center. For the first element (counting from the center) to be modified, reduce the value of τ by about 0.5%. For a rearward element, use the inverse of the adjusted value of τ to calculate the new length of the element relative to the unchanged element just forward of the change. For a forward element, use the new value of τ to calculate the new length of the element relative to the unchanged element immediately to the rear of it.

For succeeding elements outward, calculate new values of τ from the adjusted values, increasing the increment of decrease with each step. Second adjusted elements may use values of τ about 0.75% to 1.0% lower than the values just calculated. Third adjusted elements may use an increment of 1.0% to 1.5% relative to the preceding value.

Not all designs require extensive treatment. As the values of τ and σ increase, fewer elements may require adjustment to obtain the highest possible gain at the frequency limits, and these will always be the most outward elements in the array. A second caution is to check the feed-point impedance of the array after each change to ensure that it remains within design limits.

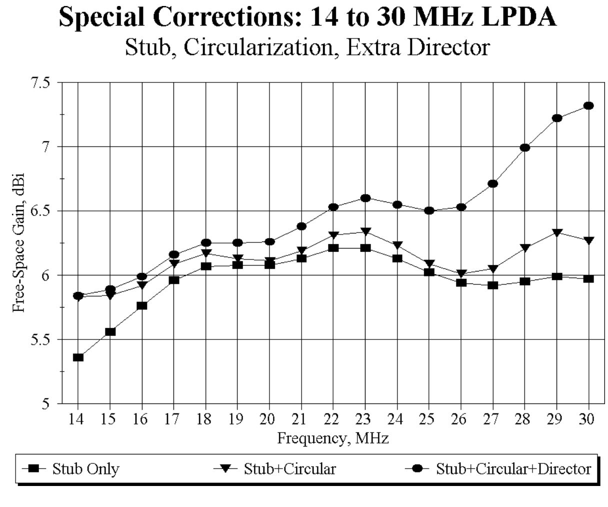

Fig 11—The modeled free space gain from 14 to 30 MHz of an LPDA with τ of 0.89 and σ of 0.04. Squares: just a stub to eliminate a weakness; Triangles: with a stub and circularized elements, and Circles: with a stub, circularized elements and a parasitic director.

Fig 11 shows the free-space gain curves from 14 to 30 MHz for a 10-element LPDA with an initial τ of 0.89 and a σ of 0.04. The design uses a 200-Ω phase-line, 0.5-inch aluminum elements, and a 3-inch 600-Ω stub. The lowest curve shows the modeled performance across the design frequency range with only the stub. Performance at the frequency limits is visibly lower than within the peak performance region. The middle curve shows the effects of circularizing τ. Average performance levels have improved noticeably at both ends of the spectrum.

In lieu of, or in addition to, the adjustment of element lengths, you may also add a parasitic director to an LPDA, as shown in **Fig 12**. The director is cut roughly for the highest operating frequency. It may be spaced between 0.1 λ and 0.15 λ from the forward-most element of the LPDA. The exact length and spacing should be determined experimentally (or from models) with two factors in mind. First, the element should not adversely effect the feed-point impedance at the highest operating frequencies. Close spacing of the director has the greatest effect on this impedance. Second, the exact spacing and element length should be set to have the most desired effect on the overall performance curve of the array. The mechanical impact of adding a director is to increase overall array length by the spacing selected for the element.

The upper curve in Fig 11 shows the effect of adding a director to the circularized array already equipped with a stub. The effect of the director is cumulative, increasing the upper range gain still further. Note that the added parasitic director is not just effective at the highest

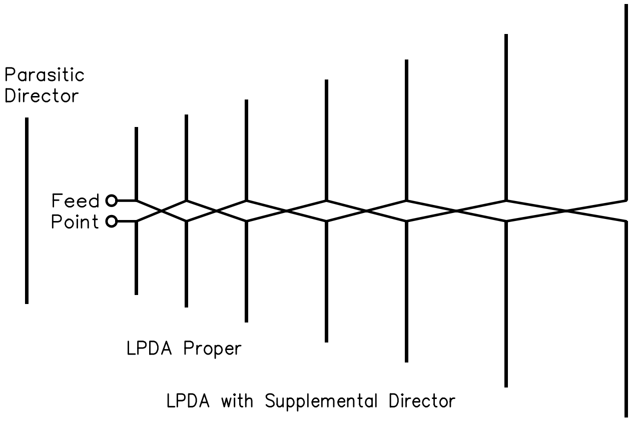

Fig 12—A generalized sketch of an LPDA with the addition of a parasitic director to improve performance at the higher frequencies within the design range.

frequencies within the LPDA design range. It has a perceptible effect almost all the way across the frequency span of the array, although the effect is smallest at the low-frequency end of the range.

The addition of a director can be used to enhance upper frequency performance of an LPDA, as in the illustration, or simply to equalize upper frequency performance with mid-range performance. High-τ designs, with good low-frequency performance, may need only a director to compensate for high-frequency gain decrease. One potential challenge to adding a director to an LPDA is sustaining a high front-to-back ratio at the upper frequency range.

Throughout the discussion of LPDAs, the performance curves of sample designs have been treated at all frequencies alike, seeking maximum performance across the entire design frequency span. Special compensations are also possible for ham-band-only LPDA designs. They include the insertion of parasitic elements within the array as well as outside the initial design boundaries. In addition, stubs may be employed not so much to eliminate weaknesses, but only to move them to frequencies outside the range of amateur interests.

A DESIGN PROCEDURE FOR AN LPDA

The following presents a systematic step-by-step design procedure for an LPDA with any desired bandwidth. The procedure requires some mathematical calculations, but a common calculator with square-root, logarithmic, and trigonometric functions is completely adequate. The notation used in this section may vary slightly from that used earlier in this chapter.

1) Decide on an operating bandwidth B between f_1, lowest frequency and f_n, highest frequency:

$$B = \frac{f_n}{f_1} \qquad \text{(Eq 6)}$$

2) Choose τ and σ to give the desired estimated average gain.

$$0.8 \le \tau \le 0.98 \text{ and } 0.03 \le \sigma \le \sigma_{opt} \qquad \text{(Eq 7)}$$

where σ_{opt} is calculated as noted earlier in this chapter.

3)Determine the value for the cotangent of the apex half-angle α from

$$\cot\alpha = \frac{4\sigma}{1-\tau} \qquad \text{(Eq 8)}$$

Although α is not directly used in the calculations, cot α is used extensively.

4) Determine the bandwidth of the active region B_{ar} from

$$B_{ar} = 1.1 + 7.7\left(1-\tau\right)^2 \cot\alpha \qquad \text{(Eq 9)}$$

5) Determine the structure (array) bandwidth B_s from

$$B_S = B \times B_{ar} \qquad \text{(Eq 10)}$$

6) Determine the boom length L, number of elements N, and longest element length ℓ_1.

$$L_n = \left(1 - \frac{1}{B_S}\right) \cot\alpha \times \frac{\lambda_{max}}{4} \qquad \text{(Eq 11)}$$

$$\lambda_{max} = \frac{984}{f_1} \qquad \text{(Eq 12)}$$

$$N = 1 + \frac{\log B_s}{\log \frac{1}{\tau}} = 1 + \frac{\ln B_S}{\ln \frac{1}{\tau}} \qquad \text{(Eq 13)}$$

$$\lambda_{1ft} = \frac{492}{f_1} \qquad \text{(Eq 14)}$$

Usually the calculated value for N will not be an integral number of elements. If the fractional value is more than about 0.3, increase the value of N to the next higher integer. Increasing the value of N will also increase the actual value of L over that obtained from the sequence of calculations just performed.

Examine L, N and l_1 to determine whether or not the array size is acceptable for your needs. If the array is too large, increase f_1 or decrease σ or τ and repeat steps 2 through 6. Increasing f_1 will decrease all dimensions. Decreasing σ will decrease the boom length. Decreasing τ will decrease both the boom length and the number of elements.

7) Determine the terminating stub Z_t. (Note: For many HF arrays, you may omit the stub, short out the longest element with a 6-inch jumper, or design a stub to overcome a specific performance weakness.) For VHF and UHF arrays calculate the stub length from

$$Z_t = \frac{\lambda_{max}}{8} \qquad \text{(Eq 15)}$$

8) Solve for the remaining element lengths from

$$\ell_n = \tau\,\ell_{n-1} \qquad \text{(Eq 16)}$$

9) Determine the element spacing $d_{1\text{-}2}$ from

$$d_{1-2} = \frac{\left(\ell_1 - \ell_2\right)\cot\alpha}{2} \qquad \text{(Eq 17)}$$

where ℓ_1 and ℓ_2 are the lengths of the rearmost elements, and d_{1-2} is the distance between the elements with the lengths ℓ_1 and ℓ_2. Determine the remaining element-to-element spacings from

$$d_{(n-1)-n} = \tau\, d_{(n-2)-(n-1)} \qquad \text{(Eq 18)}$$

10) Choose R_0, the desired feed-point resistance, to give the lowest SWR for the intended balun ratio and feed-line impedance. R_0, the mean radiation resistance level of the LPDA input impedance, is approximated by:

$$R_0 = \frac{Z_0}{\sqrt{1 + \frac{Z_0}{4\sigma' Z_{AV}}}} \qquad \text{(Eq 19)}$$

where the component terms are defined and/or calculated in the following way.

From the following equations, determine the necessary antenna feeder (phase-line) impedance, Z_0:

$$Z_0 = \frac{{R_0}^2}{8\sigma' Z_{AV}} + R_0 \sqrt{\left(\frac{R_0}{8\sigma' Z_{AV}}\right)^2 + 1} \qquad \text{(Eq 20)}$$

σ' is the mean spacing factor and is given by

$$\sigma' = \frac{\sigma}{\sqrt{\tau}} \qquad \text{(Eq 21)}$$

Z_{AV} is the average characteristic impedance of a dipole and is given by

$$Z_{AV} = 120 \left[1n \left(\frac{\ell_n}{\text{diam}_n} \right) - 2.25 \right] \qquad \text{(Eq 22)}$$

The ratio, l_n/diam_n is the length-to-diameter ratio of the element n.

11) Once Z_0 has been determined, select a combination of conductor size and spacing to achieve that impedance, using the appropriate equation for the shape of the conductors. If an impractical spacing results for the antenna feeder, select a different conductor diameter and repeat step 11. In severe cases it may be necessary to select a different R_0 and repeat steps 10 and 11. Once a satisfactory feeder arrangement is found, the LPDA design is complete.

A number of the LPDA design examples at the end of this chapter make use of this calculation method. However, the resultant design should be subjected to extensive modeling tests to determine whether there are performance deficiencies or weaknesses that require modification of the design before actual construction.

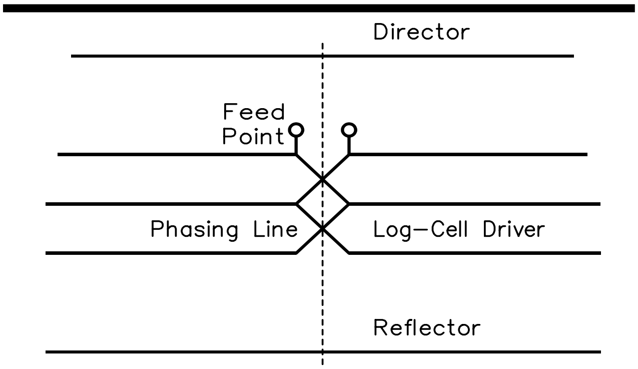

Fig 13—A sketch of a typical monoband log-cell Yagi. The reflector is, in principle, optional. The log-cell may have from 2 to 5 (or more elements). There may be one or more directors.

Log-Cell Yagis

Fig 12 showed an LPDA with an added parasitic director. Technically, this converts the original design into a hybrid Log-Yag. However, the term *Log-Yag* (or more generally the *log-cell Yagi*) is normally reserved for monoband designs that employ two or more elements in a single-band LPDA arrangement, together with (usually) a reflector and one or more directors. The aim is to produce a monoband directive array with superior directional qualities over a wider bandwidth than can be obtained from many Yagi-Uda designs. Log-cells have also been successfully used as wide-band driver sections for multiband Yagi beams.

Fig 13 illustrates the general outline of a typical log-cell Yagi. The driver section consists of a log periodic array designed for the confines of a single amateur band or other narrow range of frequencies. The parasitic reflector is usually spaced about 0.085 λ behind the rear element of the log cell, while the parasitic director is normally placed between 0.13 and 0.15 λ ahead of the log cell.

Early log-cell Yagis tended to be casually designed. Most of these designs have inferior performance compared with present-day computer-optimized Yagis of the same boom length. Some were designed by adding one or more parasitic directors to simple phased pairs of elements. Although good performance is possible, the operating bandwidth of these designs is small, suitable only for the so-called WARC bands. However, when the log-cell is designed as a narrowly spaced monoband log periodic array, the operating bandwidth increases dramatically. Operating bandwidth here refers not just to the SWR bandwidth, but also to the gain and front-to-back bandwidth.

The widest operating bandwidths require log cells of 3 to 4 elements for HF bands like 20 meters, and 4 to 5 elements for bands as wide as 10 meters. (The bandwidth of the 20-meter band is approximately 2.4% of the center frequency, while the bandwidth of the 10-meter band approaches 9.4% of the center frequency.) A practical limit to σ for log cells used within parasitic arrays is about 0.05. Slightly higher gain may be obtained from higher values of σ, but at the cost of a much longer log cell. The limiting figure for σ results in a practical value for τ between 0.94 and 0.95 to achieve a cell with the desired bandwidth characteristics.

An array designed according to these principles has an overall length that varies with the size of the log cell. A typical array with a 4-element log-cell and single parasitic elements fore and aft is a bit over 0.35 λ long, while a 5-element log-cell Yagi will be between 0.4 and 0.45 λ long. Spacing the reflector more widely (for example, up

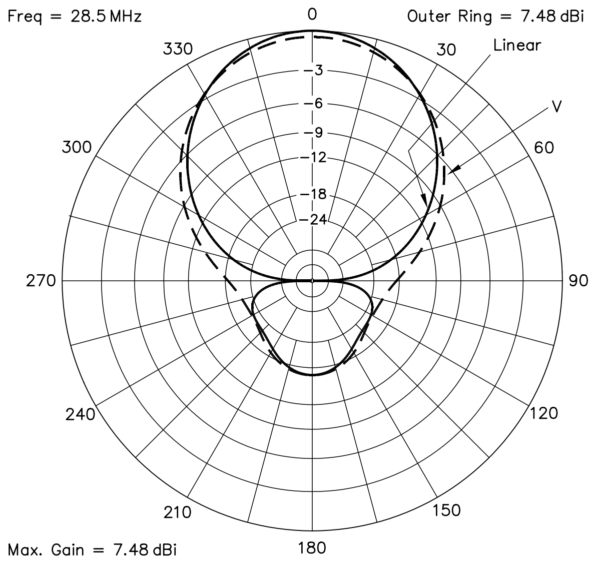

Fig 14—Overlaid free-space azimuth patterns for virtually identical 5-element log-cell Yagis (with 3-element log-cells), one having the elements V'd forward at 40° from linear (dashed), the other with linear elements (solid).

to 0.25 λ) has little effect on either gain or front-to-back ratio. Wider spacing of the director will also have only a small effect on gain, since the arrangement is already close to the boom-length limit recommended for director-driver-reflector arrays. Further lengthening of the boom should be accompanied by the addition of one or more directors to the array, if additional gain is desired from the design.

Compared to a modern-day Yagi of the same boom length, the log-cell Yagi is considerably heavier and exhibits a higher wind load due to the requirements of the log-cell driver. Yagis with 3- and 4-elements within the boom lengths just given are capable of peak free-space gain values of 8.2 to 8.5 dBi, while sustaining a high front-to-back ratio. Peak front-to-back ratios are typically in the vicinity of 25 dB. However, Yagi gain tends to decrease below the design frequency (and increase above it), while the front-to-back ratio tends to taper off as one moves away from the design frequency. For the largest log-cell sizes, log-cell Yagis of the indicated boom lengths are capable of sustaining at least 8.2 dBi free-space gain over the entire band, with front-to-back ratios of over 30 dB across the operating bandwidth.

The feed-point impedance of a log-cell Yagi is a function of both the cell design and the influence of the parasitic elements. However, for most cell designs and common phase-line designs, you can achieve a very low variation of resistance and reactance across a desired band. In many cases, the feed impedance will form a direct match for the standard 50-Ω coaxial cable used by most amateur installations. (In contrast, the high-gain, high-front-to-back Yagis used for comparison here have feed-point impedances ranging from 20 to 25 Ω.)

A common design technique used in some LPDA and log-cell Yagi designs is to bend the elements forward to form a series of Vs. A forward angle on each side of the array centerline of about 40° relative to a linear element has been popular. In some instances, the mechanical design of the array may dictate this element formation. However, this arrangement has no special benefits and possibly may degrade performance.

Fig 14 shows the free-space azimuth patterns of a single 5-element log-cell Yagi in two versions: with the elements linear and with the elements bent forward 40°. The V-array loses about $^1/_2$ dB gain, but more significantly, it loses considerable signal rejection from the sides. Similar comparisons can be obtained from pure LPDA designs and from Yagi-Uda designs when using elements in the vicinity of $^1/_2$ λ. Unless mechanical considerations call for arranging the elements in a V, the technique is not recommended.

Ultimately, the decision to build and use a log-cell Yagi involves balancing the additional weight and wind-load requirements of this design against the improvements in operating bandwidth for all of the major operating parameters, especially with respect to the front-to-back ratio and the feed-point impedance.

Wire Log-Periodic Dipole Arrays for 3.5 or 7 MHz

These wire log-periodic dipole arrays for the lower HF bands are simple in design and easy to build. They are designed to have reasonable gain, be inexpensive and lightweight, and may be assembled with stock items found in large hardware stores. They are also strong—they can withstand a hurricane! These antennas were first described by John J. Uhl, KV5E, in *QST* for August, 1986. **Fig 15** shows one method of installation. You can use the information here as a guide and point of reference for building similar LPDAs.

If space is available, the antennas can be rotated or repositioned in azimuth after they are completed. A 75-foot tower and a clear turning radius of 120 feet around the base of the tower are needed. The task is simplified

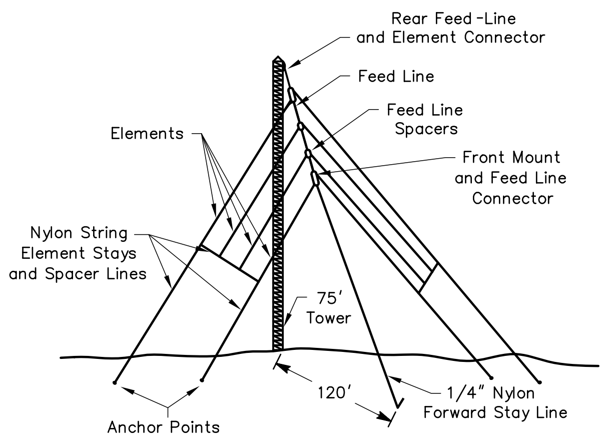

Fig 15—Typical lower-HF wire 4-element log periodic dipole array erected on a tower.

if you use only three anchor points, instead of the five shown in Fig 15. Omit the two anchor points on the forward element, and extend the two nylon strings used for element stays all the way to the forward stay line.

DESIGN OF THE LOG-PERIODIC DIPOLE ARRAYS

Design constants for the two arrays are listed in **Tables 1** and **2**. The preceding sections of this chapter contain the design procedure for arriving at the dimensions and other parameters of these arrays. The primary differences between these designs and one-octave upper HF arrays are the narrower frequency ranges and the use of wire, rather than tubing, for the elements. As design examples for the LPDA, you may wish to work through the step-by-step procedure and check your results against the values in Tables 1 and 2. You may also wish to compare these results with the output of an LPDA design software package such as *LPCAD28*.

From the design procedure, the feeder wire spacings for the two arrays are slightly different, 0.58 inch for the 3.5-MHz array and 0.66 inch for the 7-MHz version. As a compromise toward the use of common spacers for both bands, a spacing of 5/8 inch is quite satisfactory. Surprisingly, the feeder spacing is not at all critical here from a matching standpoint, as may be verified from Z_0 = 276 log (2S/diam) and from Eq 4. Increasing the spacing to as much as 3/4 inch results in an R_0 SWR of less than 1.1:1 on both bands.

Constructing the Arrays

Construction techniques are the same for both the 3.5 and the 7-MHz versions of the array. Once the designs are completed, the next step is to fabricate the fittings; see **Fig 16** for details. Cut the wire elements and feed lines to the proper sizes and mark them for identification. After the wires are cut and placed aside, it will be difficult to remember which is which unless they are marked. When you have finished fabricating the connectors and cutting all of the wires, the antenna can be assembled. Use your ingenuity when building one of these antennas; it isn't necessary to duplicate these LPDAs precisely.

The elements are made of standard #14 stranded cop-

Table 1

Design Parameters for the 3.5-MHz Single-Band LPDA

f1 = 3.3 MHz
f_n = 4.1 MHz
B = 1.2424
τ = 0.845
σ = 0.06
Gain = 5.9 dBi = 3.8 dBd
cot α = 1.5484
B_{ar} = 1.3864
B_s = 1.7225
L = 48.42 feet
N = 4.23 elements (decrease to 4)
Zt = 6-inch short jumper
R_0 = 208 Ω
Z_{AV} = 897.8 Ω
σ' = 0.06527
Z_0 = 319.8 Ω
Antenna feeder:
#12 wire spaced 0.58 inches
Balun: 4:1
Feed line: 52-Ω coax

Element lengths:
$\ell 1$ = 149.091 feet
$\ell 2$ = 125.982 feet
$\ell 3$ = 106.455 feet
$\ell 4$ = 89.954 feet
Element spacings:
d_{12}=17.891 feet
d_{23} = 15.118 feet
d_{34} = 12.775 feet
Element diameters
All = 0.0641 inches
ℓ/diameter ratios:
ℓ/diam_4 = 16840
ℓ/diam_3 = 19929
ℓ/diam_2 = 23585
ℓ/diam_1 = 27911

Table 2

Design Parameters for the 7-MHz Single-Band LPDA

f1 = 6.9 MHz
f_n = 7.5 MHz
B = 1.0870
τ = 0.845
σ = 0.06
Gain = 5.9 dBi = 3.8 dBd
cot α = 1.5484
B_{ar} = 1.3864
B_s = 1.5070
L = 18.57 feet
N = 3.44 elements (increase to 4)
Z_t = 6-inch short jumper
R_0 = 208 Ω
Z_{AV} = 809.3 Ω
σ' = 0.06527
Z_0 = 334.2 Ω
Antenna feeder:
#12 wire spaced 0.66 inches
Balun: 4:1
Feed line: 52-Ω coax

Element lengths:
$\ell 1$ = 71.304 feet
$\ell 2$ = 60.252 feet
$\ell 3$ = 50.913 feet
$\ell 4$ = 43.022 feet
Element spacings:
d_{12} = 8.557 feet
d_{23} = 7.230 feet
d_{34} = 6.110 feet
Element diameters:
All = 0.0641 inches
ℓ/diameter ratios:
$\ell 4/\text{diam}_4$ = 8054
$\ell 3/\text{diam}_3$ = 9531
$\ell 2/\text{diam}_2$ = 11280
$\ell 1/\text{diam}_1$ = 13349

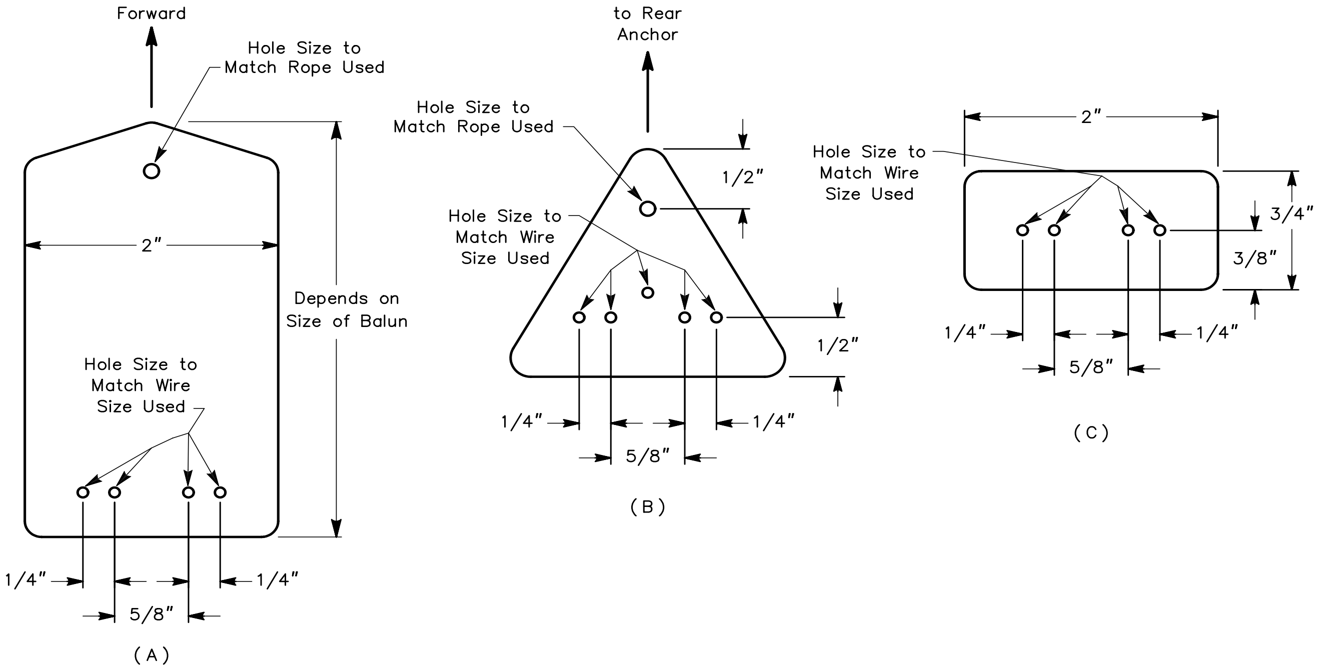
Forward
Hole Size to Match Rope Used
2"
Depends on Size of Balun
Hole Size to Match Wire Size Used
1/4"
1/4"
5/8"
(A)
to Rear Anchor
Hole Size to Match Rope Used
1/2"
Hole Size to Match Wire Size Used
1/2"
1/4"
1/4"
5/8"
(B)
2"
Hole Size to Match Wire Size Used
3/4"
3/8"
1/4"
1/4"
5/8"
(C)

Fig 16—Pieces for the LPDA that require fabrication. At A is the forward connector, made from 1/2-inch Lexan (polycarbonate). At B is the rear connector, also made from 1/2-inch Lexan. At C is the pattern for the phase-line spacers, made from 1/4-inch Plexiglas. Two spacers are required for the array.

per wire. The two parallel feed lines are made of #12 solid copper-coated steel wire, such as Copperweld. Copperweld will not stretch when placed under tension. The front and rear connectors are cut from 1/2-inch thick Lexan sheeting, and the feed-line spacers from 1/4-inch Plexiglas sheeting.

Study the drawings carefully and be familiar with the way the wire elements are connected to the two feed lines, through the front, rear and spacer connectors. Details are sketched in **Figs 17** and **18**. Connections made in the way shown in the drawings prevent the wire from breaking. All of the rope, string, and connectors must be made of materials that can withstand the effects of tension and weathering. Use nylon rope and strings, the type that yachtsmen use. Fig 15 shows the front stay rope coming down to ground level at a point 120 feet from the base of a 75-foot tower. Space may not be available for this arrangement in all cases. An alternative installation technique is to put a pulley 40 feet up in a tree and run the front stay rope through the pulley and down to ground level at the base of the tree. The front stay rope will have to be tightened with a block and tackle at ground level.

Putting an LPDA together is not difficult if it is assembled in an orderly manner. It is easier to connect the elements to the feeder lines when the feed-line

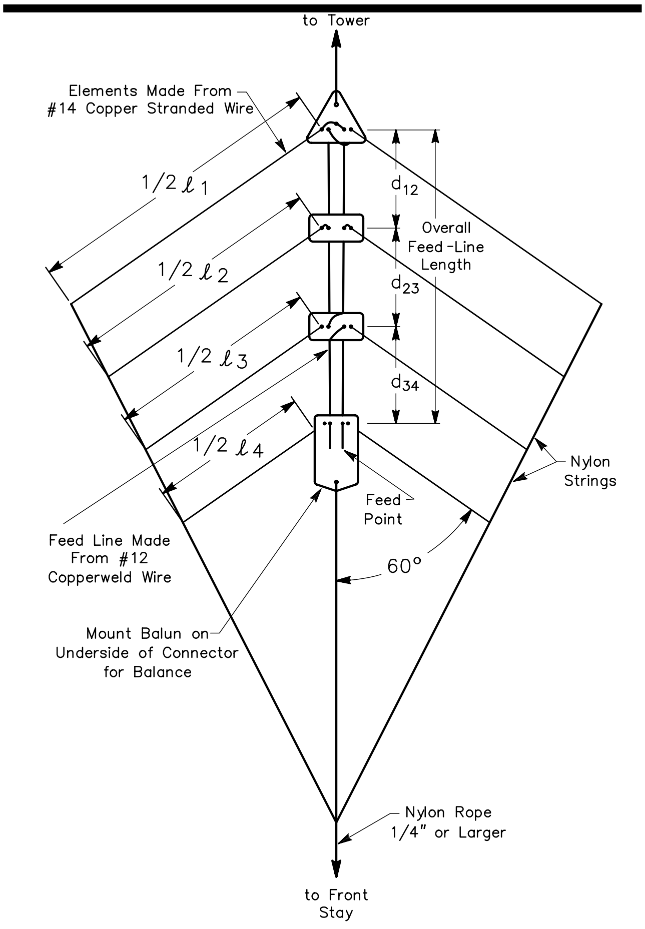
to Tower
Elements Made From #14 Copper Stranded Wire
1/2 ℓ 1
1/2 ℓ 2
1/2 ℓ 3
1/2 ℓ 4
d12
d23
d34
Overall Feed-Line Length
Nylon Strings
Feed Point
60°
Feed Line Made From #12 Copperweld Wire
Mount Balun on Underside of Connector for Balance
Nylon Rope 1/4" or Larger
to Front Stay

Fig 17—The generic layout for the lower HF wire LPDA. Use a 4:1 balun on the forward connector. See Tables 1 and 2 for dimensions.

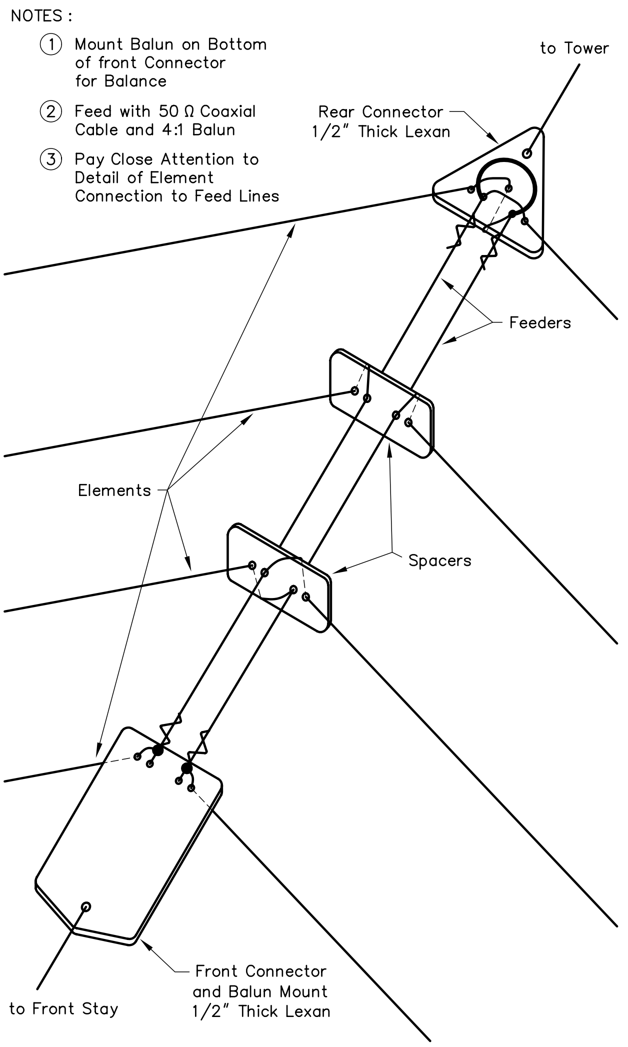
NOTES :
① Mount Balun on Bottom of front Connector for Balance
② Feed with 50 Ω Coaxial Cable and 4:1 Balun
③ Pay Close Attention to Detail of Element Connection to Feed Lines
to Tower
Rear Connector 1/2" Thick Lexan
Feeders
Elements
Spacers
to Front Stay
Front Connector and Balun Mount 1/2" Thick Lexan

Fig 18—Details of the electrical and mechanical connections of the elements to the phase-line. Knots in the nylon rope stay line are not shown.

assembly is stretched between two points. Use the tower and a block and tackle. Attaching the rear connector to the tower and assembling the LPDA at the base of the tower makes raising the antenna into place a much simpler task. Tie the rear connector securely to the base of the tower and attach the two feeder lines to it. Then thread the two feed-line spacers onto the feed line. The spacers will be loose at this time, but will be positioned properly when the elements are connected. Now connect the front connector to the feed lines. A word of caution: Measure accurately and carefully! Double-check all measurements before you make permanent connections.

Connect the elements to the feeder lines through their respective plastic connectors, beginning with element 1, then element 2, and so on. Keep all of the element wires securely coiled. If they unravel, you will have a tangled mess of kinked wire. Recheck the element-to-feeder connections to ensure proper and secure junctions. (See Figs 17 and 18.) Once you have completed all of the element connections, attach the 4:1 balun to the underside of the front connector. Connect the feeder lines and the coaxial cable to the balun.

You will need a separate piece of rope and a pulley to raise the completed LPDA into position. First secure the eight element ends with nylon string, referring to Figs 15 and 17. The string must be long enough to reach the tie-down points. Connect the front stay rope to the front connector, and the completed LPDA is now ready to be raised into position. While raising the antenna, uncoil the element wires to prevent their getting away and tangling up into a mess. Use care! Raise the rear connector to the proper height and attach it securely to the tower, then pull the front stay rope tight and secure it. Move the elements so they form a 60° angle with the feed lines, in the direction of the front, and space them properly relative to one another. By adjusting the end positions of the elements as you walk back and forth, you will be able to align all the elements properly. Now it is time to hook your rig to the system and make some contacts.

Performance

The reports received from these LPDAs were compared with an inverted-V dipole. All of the antennas are fixed; the LPDAs radiate to the northeast, and the dipole to the northeast and southwest. The apex of the dipole is at 70 feet, and the 40- and 80-meter LPDAs are at 60 and 50 feet, respectively. Basic array gain was apparent from many of the reports received. During pileups, it was possible to break in with a few tries on the LPDAs, yet it was impossible to break in the same pileups using the dipole. The gain of the LPDAs is several dB over the dipole. For additional gain, experimenters may wish to try a parasitic director about $^1/_8$ λ ahead of the array. Director length and spacing from the forward LPDA element should be field-adjusted for maximum performance while maintaining the impedance match across each of the bands.

Wire LPDA systems offer many possibilities. They are easy to design and to construct: real advantages in countries where commercially built antennas and parts are not available at reasonable cost. The wire needed can be obtained in all parts of the world, and cost of construction is low. If damaged, the LPDAs can be repaired easily with pliers and solder. For those who travel on DXpeditions where space and weight are large considerations, LPDAs are lightweight but sturdy, and they perform well.

5-Band Log Periodic Dipole Array

A rotatable log periodic array designed to cover the frequency range from 13 to 30 MHz is pictured in **Fig 19**. This is a large array having a free-space gain that varies from 6.6 to over 6.9 dBi, depending upon the operating portion of the design spectrum. This antenna system was originally described by Peter D. Rhodes, WA4JVE, in Nov 1973 *QST*. A measured radiation pattern for the array appears in **Fig 20**.

The characteristics of this array are:

1) Half-power beamwidth, 43° (14 MHz)
2) Design parameter $\tau = 0.9$
3) Relative element spacing constant $\sigma = 0.05$
4) Boom length, L = 26 feet
5) Longest element $\lambda 1$ = 37 feet 10 inches. (A tabulation

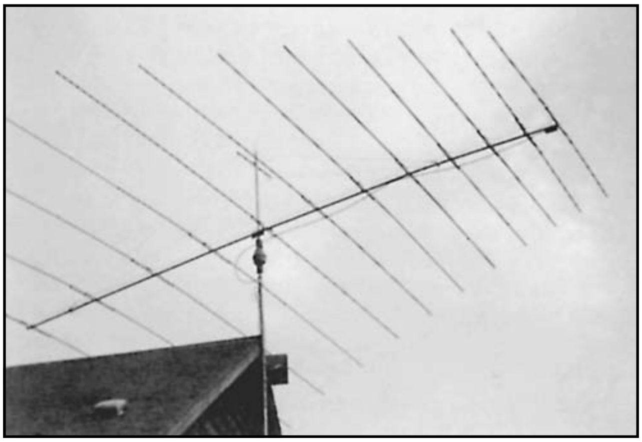

Fig 19—The 13-30 MHz log periodic dipole array.

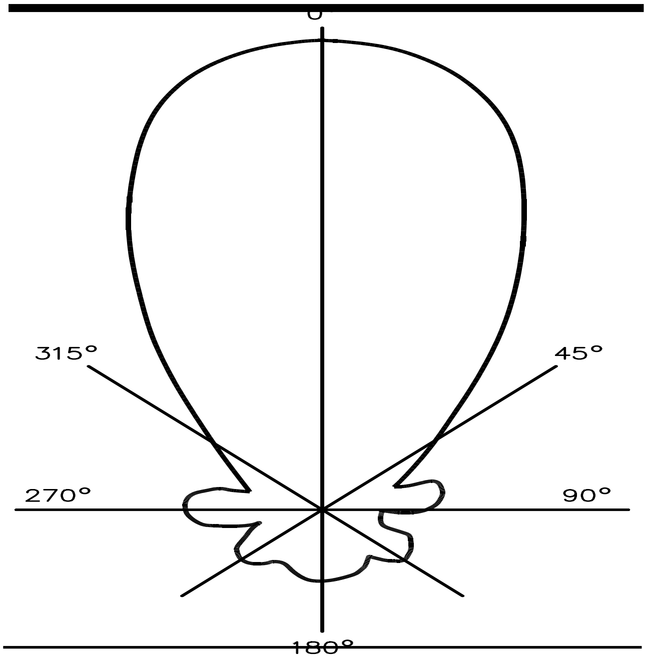

Fig 20—Measured radiation pattern of the 13-30 MHz LPDA. The front-to-back ratio is about 14 dB at 14 MHz and increases to 21 dB at 28 MHz.

Table 3
13-30 MHz LPDA Dimensions, feet

Ele. No.	*Length*	*$d_{n-1,n}$ (spacing)*	*Nearest Resonant*
1	37′ 10.2″	—	
2	34′ 0.7″	3′ 9.4″ = d_{12}	14 MHz
3	30′ 7.9″	3′ 4.9″ = d_{23}	
4	27′ 7.1″	3′ 0.8″ = d_{34}	
5	24′ 10.0″	2′ 9.1″ = d_{45}	18 MHz
6	22′ 4.2″	2′ 5.8″ = d_{56}	21 MHz
7	20′ 1.4″	2′ 2.8″ = d_{67}	
8	18′ 1.2″	2′ 0.1″ = d_{78}	24.9 MHz
9	16′ 3.5″	1′ 9.7″ = d_{89}	28 MHz
10	14′ 7.9″	1′ 7.5″ = $d_{9,10}$	
11	13′ 2.4″	1′ 5.6″ = $d_{10,11}$	
12	11′ 10.5″	1′ 3.8″ = $d_{11,12}$	

Table 4
Materials list: 13-30 MHz LPDA

Material Description	*Quantity*
1) Aluminum tubing—0.047″ wall thickness	
1″—12′ or 6′ lengths	126 lineal feet
7/8″—12′ lengths	96 lineal feet
7/8″—6′ or 12′ lengths	66 lineal feet
3/4″—8′ lengths	16 lineal feet
2) Stainless-steel hose clamps—2″ max	48 ea
3) Stainless-steel hose clamps—1 1/4″ max	26 ea
4) TV type U bolts	14 ea
5) U bolts, galv. type	
5/16″ × 1 1/2″	4 ea
1/4″ × 1″	2 ea
6) 1″ ID polyethylene water-service pipe 160 lb/in.^{2} test, approx. 1 1/4″ OD	20 lineal feet
A) 1 1/4″ × 1 1/4 × 1/8″ aluminum Angle—6' lengths	30 lineal feet
B) 1″ × 1/4″ aluminum bar—6′ lengths	12 lineal feet
7) 1 1/4″ top rail of chain-link fence	26 lineal feet
8) 1:1 toroid balun	1 ea
9) 6-32 × 1″ stainless steel screws	24 ea
6-32 stainless steel nuts	48 ea
#6 solder lugs	24 ea
10) #12 copper feeder wire	60 lineal feet
11A) 12″ × 8″ × 1/4″ aluminum plate	1 ea
B) 6″ × 4″ × 1/4″ aluminum plate	1 ea
12A) 3/4″ galv. Pipe	3 lineal feet
B) 1″ galv. pipe-mast	5 lineal feet
13) Galv. guy wire	50 lineal feet
14) 1/4″ × 2″ turnbuckles	4 ea
15) 1/4″ × 1 1/2″ eye bolts	2 ea
16) TV guy clamps and eye bolts	2 ea

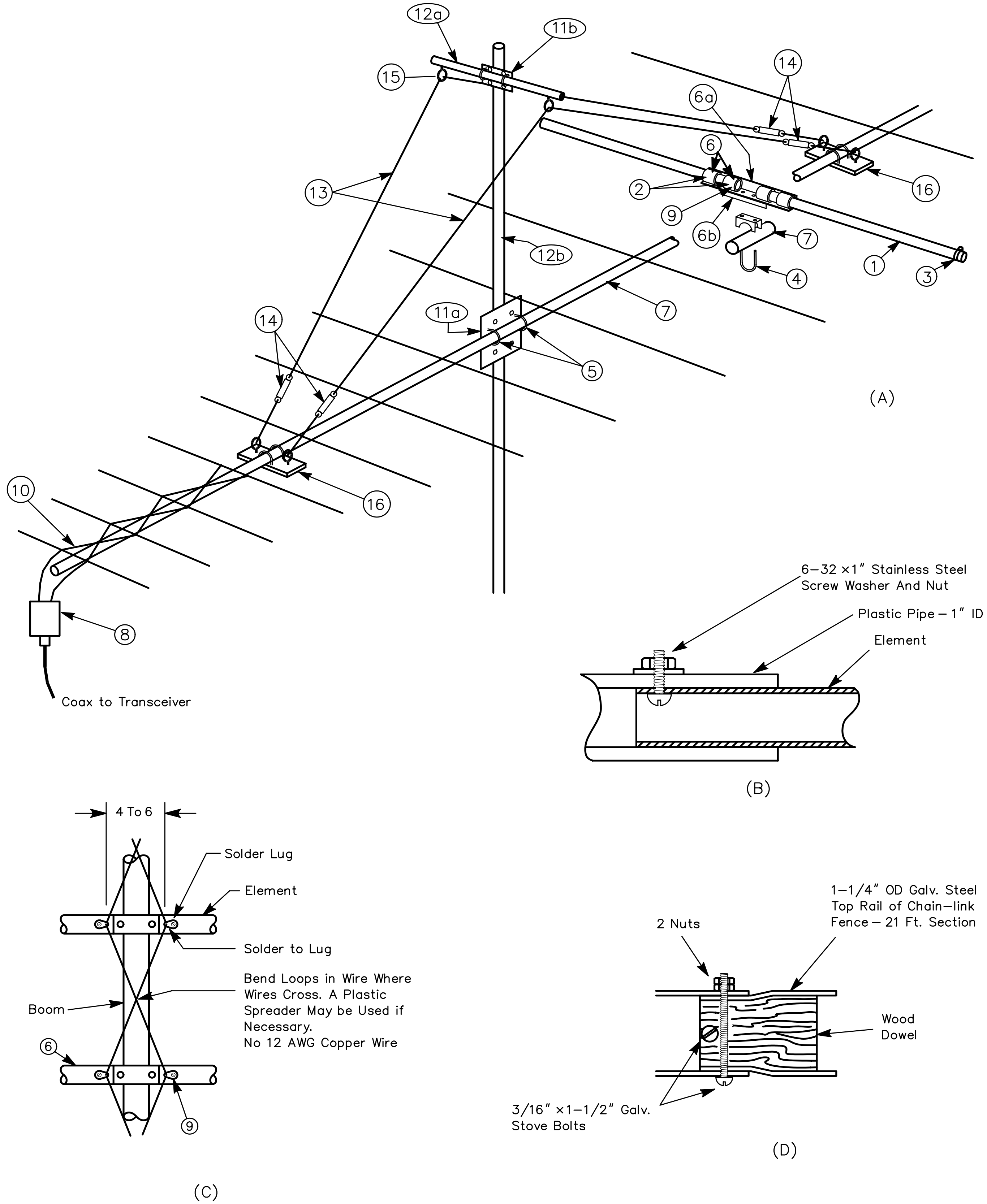
12a
11b
15
14
6a
6
2
9
6b
16
7
4
1
3
13
12b
11a
7
5
14
16
10
8
Coax to Transceiver
(A)
6–32 ×1″ Stainless Steel
Screw Washer And Nut
Plastic Pipe – 1″ ID
Element
(B)
4 To 6
Solder Lug
Element
Solder to Lug
Bend Loops in Wire Where
Wires Cross. A Plastic
Spreader May be Used if
Necessary.
No 12 AWG Copper Wire
Boom
6
9
(C)
1–1/4″ OD Galv. Steel
Top Rail of Chain–link
Fence – 21 Ft. Section
2 Nuts
Wood
Dowel
3/16″ ×1–1/2″ Galv.
Stove Bolts
(D)

Fig 21—Construction diagrams of the 13-30 MHz LPDA. B and C show the method of making electrical connection between the phase-line and each half-element. D shows how the boom sections are joined.

Table 5
Element Material Requirements: 13-30 MHz LPDA

Ele. No.	***1″ Tubing***		***7/8″ Tubing***		***3/4″ Tubing***		***1 1/4″ Angle***	***1″ Bar***
	Len.	*Qty*	*Len.*	*Qty*	*Len.*	*Qty*	*Len.*	*Len.*
1	6′	2	6′	2	8′	2	3′	1′
2	6′	2	12′	2	—	—	3′	1′
3	6′	2	12′	2	—	—	3′	1′
4	6′	2	8.5′	2	—	—	3′	1′
5	6′	2	7′	2	—	—	3′	1′
6	6′	2	6′	2	—	—	3′	1′
7	6′	2	5′	2	—	—	2′	1′
8	6′	2	3.5′	2	—	—	2′	1′
9	6′	2	2.5′	2	—	—	2′	1′
10	3′	2	5′	2	—	—	2′	1′
11	3′	2	4′	2	—	—	2′	1′
12	3′	2	4′	2	—	—	2′	1′

of element lengths and spacings appears in **Table 3**.)
6) Total weight, 116 pounds
7) Wind-load area, 10.7 square feet
8) Required input impedance (mean resistance), R_0 = 72 Ω, Z_t = 6-inch jumper #18 wire
9) Average characteristic dipole impedance, Z_{AV}: 337.8 Ω
10) Impedance of the feeder, Z_0: 117.1 Ω
11) Feeder: #12 wire, close spaced
12) With a 1:1 toroid balun at the input terminals and a 72-Ω coax feed line, the maximum SWR is 1.4:1.

The mechanical assembly uses materials readily available from most local hardware stores or aluminum supply houses. The materials needed are given in **Table 4**. In the construction diagram, **Fig 21**, the materials are referenced by their respective material list numbers. The photograph shows the overall construction, and the drawings show the details. **Table 5** gives the required tubing lengths to construct the elements.

Experimenters may wish to improve the performance of the array at both the upper and lower frequency ends of the design spectrum so that it more closely approaches the performance in the middle of the design frequency range. The most apt general technique for raising both the gain and the front-to-back ratio at the frequency extremes would be to circularize τ as described earlier in this chapter. However, other techniques may also be applied.

The Telerana

The Telerana (Spanish for *spider web*) is a rotatable log periodic antenna that is lightweight, easy to construct and relatively inexpensive to build. Designed to cover 12.1 to 30 MHz, it was co-designed by George Smith, W4AEO, and Ansyl Eckols, YV5DLT, and first described by Eckols in *QST* for Jul 1981. Some of the design parameters are as follow.

1) $\tau = 0.9$
2) $\sigma = 0.05$
3) Gain = 4.5 to 5.5 dBi (free-space) depending upon frequency
4) Feed arrangement: 400-Ω feeder line with 4:1 balun, fed with 52-Ω coax. The SWR is 1.5:1 or less in all amateur bands.

The array consists of 13 dipole elements, properly spaced and transposed, along an open wire feeder having an impedance of approximately 400 Ω. See **Figs 22** and **23**. The array is fed at the forward (smallest) end with a 4:1 balun and RG-8 cable placed inside the front arm and leading to the transmitter. An alternative feed method is to use open wire or ordinary TV ribbon and a tuner, eliminating the balun.

The frame that supports the array (**Fig 24**) consists of four 15-foot fiberglass vaulting poles slipped over short nipples at the hub, appearing like wheel spokes (**Fig 25**). Instead of being mounted directly into the fiberglass, the hub mounts into short metal tubing sleeves that are inserted into the ends of each arm to prevent crushing and splitting the fiberglass. The necessary holes are drilled to receive the wires and nylon.

A shopping list is provided in **Table 6**. The center hub is made from a $1^1/_4$-inch galvanized four-outlet cross or X and four 8-inch nipples (Fig 25). A 1-inch diameter X may be used alternatively, depending on the diameter of the fiberglass. A hole is drilled in the bottom of the hub to allow the cable to be passed through after welding the hub to the rotator mounting stub.

All four arms of the array must be 15 feet long. They should be strong and springy to maintain the tautness of the array. If vaulting poles are used, try to obtain all of them with identical strength ratings.

The forward spreader should be approximately 14.8 feet long. It can be much lighter than the four main arms, but must be strong enough to keep the lines rigid.

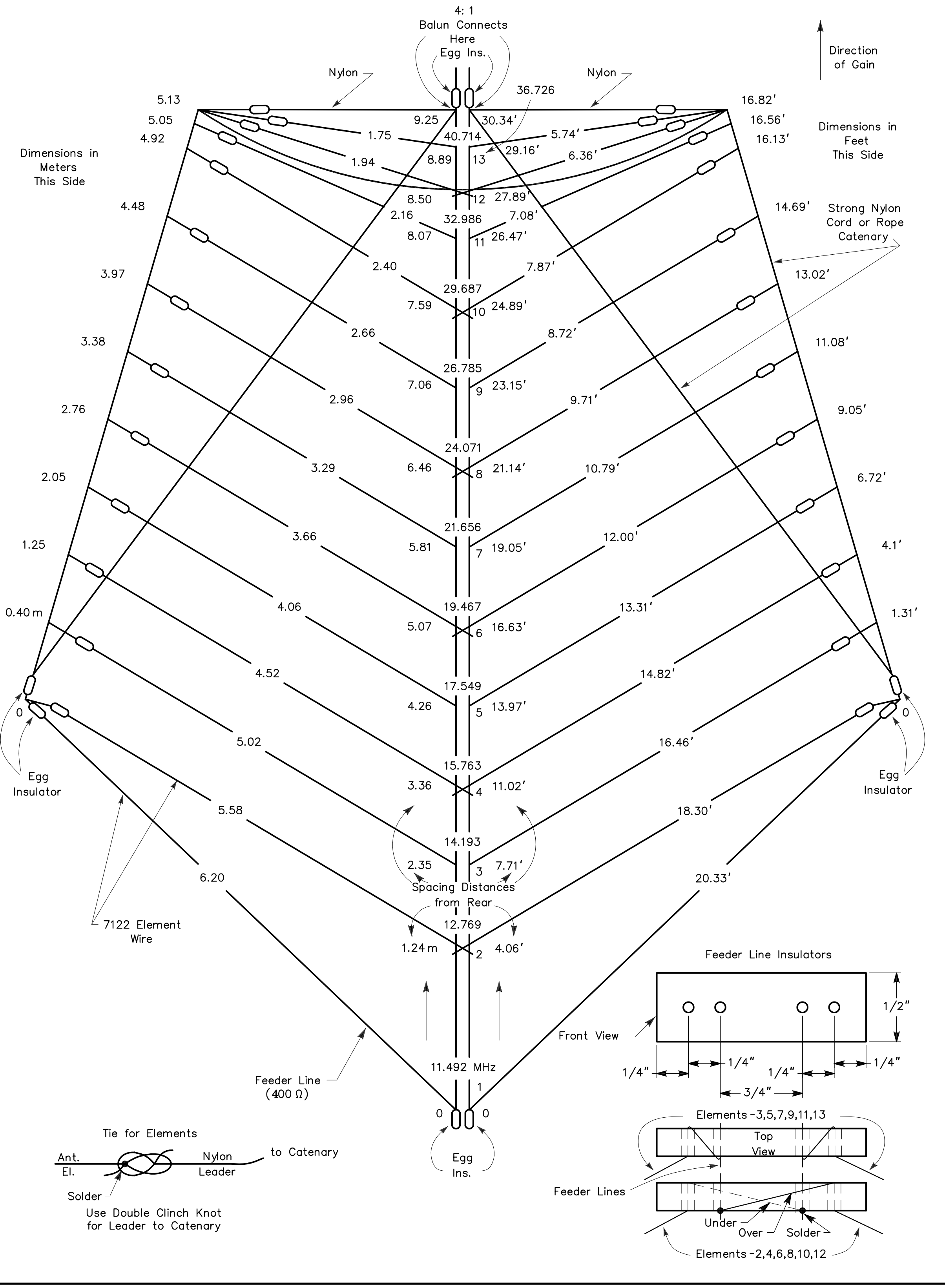
4: 1
Balun Connects
Here
Egg Ins.
Direction
of Gain
Nylon
Nylon
36.726
5.13
5.05
4.92
16.82'
16.56'
16.13'
Dimensions in
Feet
This Side
Dimensions in
Meters
This Side
9.25
30.34'
1.75
40.714
5.74'
1.94
8.89
13
29.16'
6.36'
8.50
12
27.89'
4.48
14.69'
Strong Nylon
Cord or Rope
Catenary
2.16
32.986
7.08'
8.07
11
26.47'
3.97
13.02'
2.40
7.87'
29.687
7.59
10
24.89'
3.38
11.08'
2.66
8.72'
26.785
7.06
9
23.15'
2.76
9.05'
2.96
9.71'
24.071
6.46
8
21.14'
2.05
6.72'
3.29
10.79'
21.656
5.81
7
19.05'
1.25
4.1'
3.66
12.00'
19.467
5.07
6
16.63'
0.40 m
1.31'
4.06
13.31'
17.549
4.26
5
13.97'
4.52
14.82'
0
0
15.763
3.36
4
11.02'
5.02
16.46'
Egg
Insulator
Egg
Insulator
5.58
18.30'
14.193
2.35
3
7.71'
6.20
20.33'
Spacing Distances
from Rear
12.769
1.24 m
2
4.06'
7122 Element
Wire
Feeder Line Insulators
1/2"
Front View
1/4"
1/4"
1/4"
1/4"
3/4"
11.492 MHz
1
Feeder Line
(400 Ω)
0
0
Egg
Ins.
Tie for Elements
Ant.
El.
Nylon
Leader
to Catenary
Solder
Use Double Clinch Knot
for Leader to Catenary
Elements -3,5,7,9,11,13
Top
View
Feeder Lines
Under
Over
Solder
Elements -2,4,6,8,10,12

Fig 22—The overall configuration of the spider web antenna. Nylon monofilament line is used from the ends of the elements to the nylon cords. Use nylon line to tie every point where lines cross. The forward fiberglass feeder lies on the feeder line and is tied to it. Both metric and English measurements are shown, except for the illustration of the feed-line insulator. Use soft-drawn copper or stranded wire for elements 2 through 12. Element 1 should use 7/22 flexible wire or #14 AWG Copperweld.

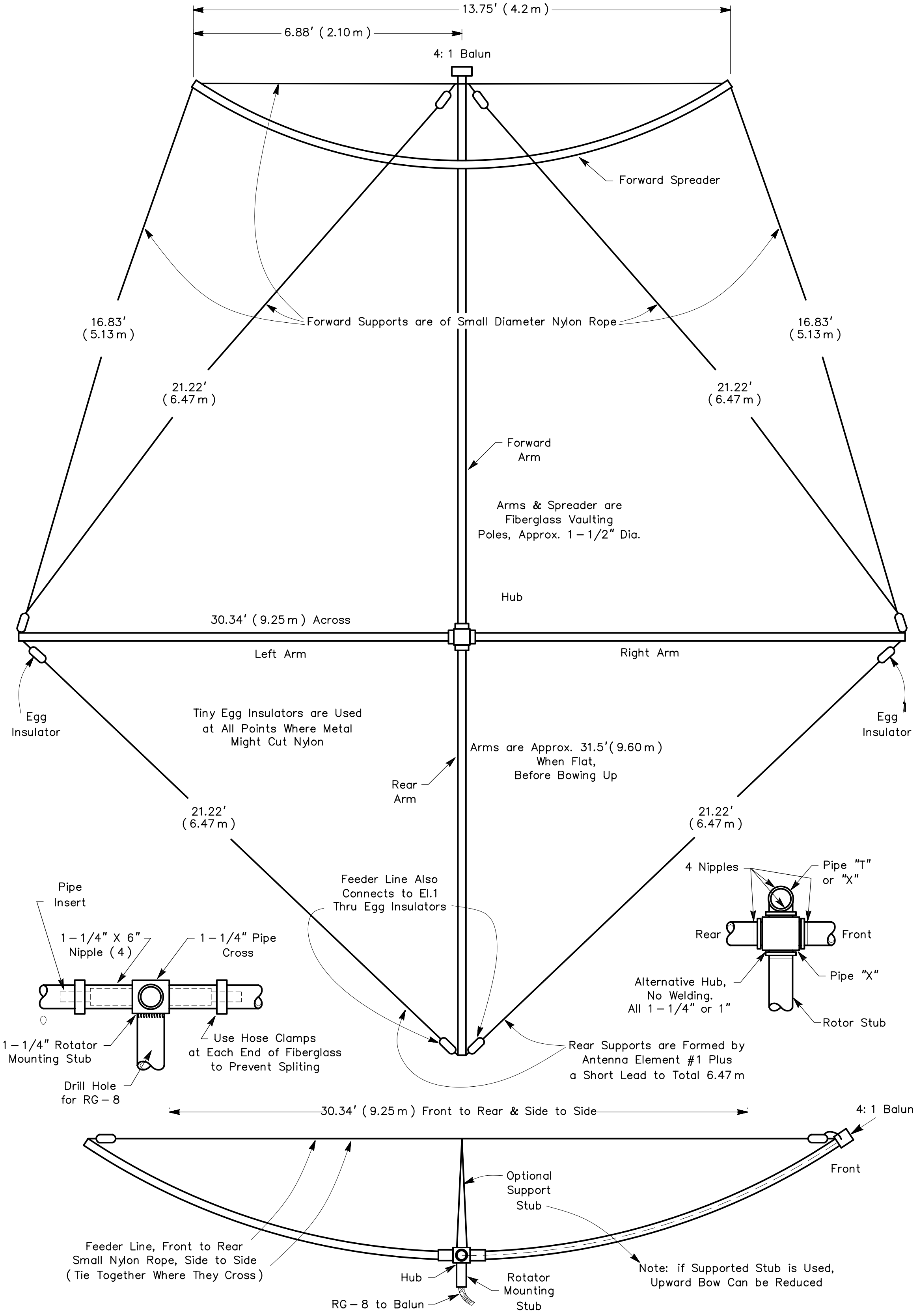
13.75′ (4.2 m)
6.88′ (2.10 m)
4: 1 Balun
Forward Spreader
Forward Supports are of Small Diameter Nylon Rope
16.83′ (5.13 m)
16.83′ (5.13 m)
21.22′ (6.47 m)
21.22′ (6.47 m)
Forward Arm
Arms & Spreader are Fiberglass Vaulting Poles, Approx. 1 – 1/2″ Dia.
Hub
30.34′ (9.25 m) Across
Left Arm
Right Arm
Egg Insulator
Egg Insulator
Tiny Egg Insulators are Used at All Points Where Metal Might Cut Nylon
Arms are Approx. 31.5′(9.60 m) When Flat, Before Bowing Up
Rear Arm
21.22′ (6.47 m)
21.22′ (6.47 m)
Feeder Line Also Connects to El.1 Thru Egg Insulators
4 Nipples
Pipe "T" or "X"
Rear
Front
Pipe "X"
Alternative Hub, No Welding. All 1 – 1/4″ or 1″
Rotor Stub
Pipe Insert
1 – 1/4″ X 6″ Nipple (4)
1 – 1/4″ Pipe Cross
1 – 1/4″ Rotator Mounting Stub
Use Hose Clamps at Each End of Fiberglass to Prevent Spliting
Drill Hole for RG – 8
Rear Supports are Formed by Antenna Element #1 Plus a Short Lead to Total 6.47 m
30.34′ (9.25 m) Front to Rear & Side to Side
4: 1 Balun
Front
Optional Support Stub
Feeder Line, Front to Rear Small Nylon Rope, Side to Side (Tie Together Where They Cross)
Hub
Rotator Mounting Stub
RG – 8 to Balun
Note: if Supported Stub is Used, Upward Bow Can be Reduced

Fig 23—The frame construction of the spider web antenna. Two different hub arrangements are illustrated.

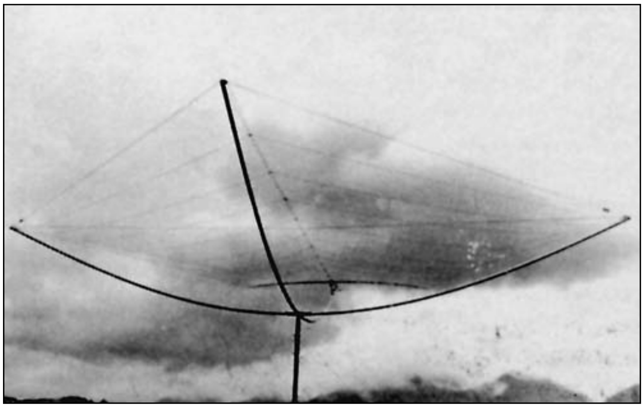

Fig 24—Although the spider web antenna resembles a rotatable clothes line, it is much larger, as indicated by Figs 22 and 23. However, the antenna can be lifted by hand.

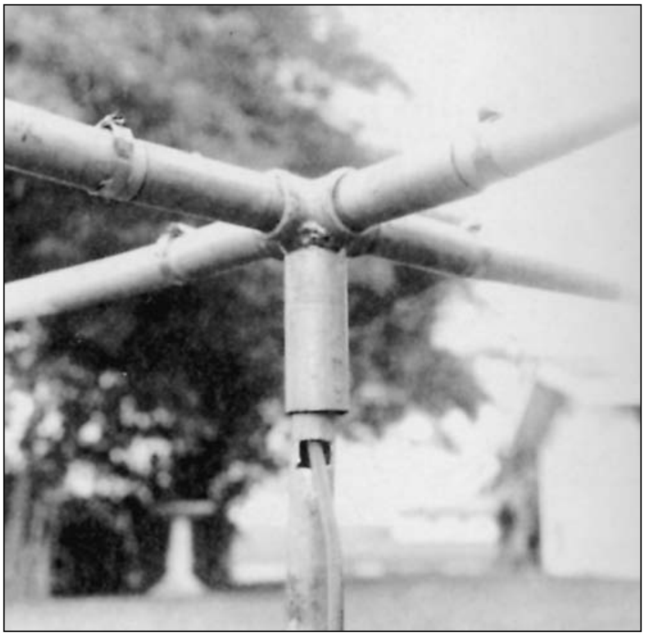

Fig 25—The simple arrangement of the spider web antenna hub. See Fig 23 and the text for details.

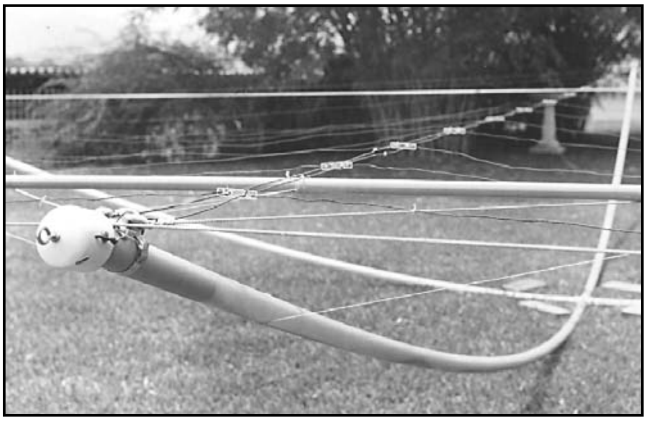

Fig 26—The elements, balun, transmission line and main bow of the spider web antenna.

Table 6
Shopping List for the Telerana

1—$1\frac{1}{4}$-inch galvanized, 4-outlet cross or X.
4—8-inch nipples.
4—15-foot long arms. Vaulting poles suggested. These must be strong and all of the same strength (150 lb) or better.
1—Spreader, 14.8 foot long (must not be metal).
1—4:1 balun unless open-wire or TV cable is used.
12—Feed-line insulators made from Plexiglas or fiberglass.
36—Small egg insulators.
328 feet copper wire for elements; flexible 7/22 is suggested.
65.6 feet (20 m) #14 Copperweld wire for interelement feed line.
164 feet (50 m) strong $\frac{1}{8}$-inch dia cord.
1—Roll of nylon monofilament fishing line, 50 lb test or better.
4—Metal tubing inserts go into the ends of the fiberglass arms.
2—Fiberglass fishing-rod blanks.
4—Hose clamps.

If tapered, the spreader should have the same measurements from the center to each end. *Do not use metal for this spreader*.

Building the frame for the array is the first construction step. Once the frame is prepared, then everything else can be built onto it. Begin by assembling the hub and the four arms, letting them lie flat on the ground with the rotator stub inserted in a hole in the ground. The tip-to-tip length should be about 31.5 feet each way. A hose clamp is used at each end of the arms to prevent splitting. Place the metal inserts in the outer ends of the arms, with 1 inch protruding. The mounting holes should have been drilled at this point. If the egg insulators and nylon cords are mounted to these tube inserts, the whole antenna can be disassembled simply by bending up the arms and pulling out the inserts with everything still attached.

Choose the arm to be at the front end. Mount two egg insulators at the front and rear to accommodate the inter-element feeder. These insulators should be as close as possible to the ends.

At each end of the cross-arm on top, install a small pulley and string nylon cord across and back. Tighten the cord until the upward bow reaches 3 feet above the hub. All cords will require retightening after the first few days because of stretching. The cross-arm can be laid on its side while preparing the feeder line. For the front-to-rear bowstring it is important to use a wire that will not stretch, such as #14 Copperweld. This bowstring is actually the inter-element transmission line. See **Fig 26**.

Secure the rear ends of the feeder to the two rear insulators, soldering the wrap. Before securing the fronts, slip the 12 insulators onto the two feed lines. A rope can be used temporarily to form the bow and to aid in mounting the feeder line. The end-to-end length of the feeder

Improving the Telerana

In *The ARRL Antenna Compendium, Vol 4*, Markus Hansen, VE7CA, described how he modified the Telerana to improve the front-to-back ratio on 20 and 15 meters. In addition, he added a trapped 30/40-meter dipole that functions as a top truss system to stabilize the modified Telerana in strong uprising winds that otherwise could turn the antenna into an "inside-out umbrella."

Fig A shows the layout for the modified Telerana, and Table A lists the lengths and spacings for the #14 wire elements. Note that VE7CA used tuning stubs to tweak the 15 and 20-meter reflector wires for best rearward pattern. The construction techniques used by VE7CA are the same as for the original Telerana. Fig B shows a side view of the additional 40/30-meter-dipole truss sytem.

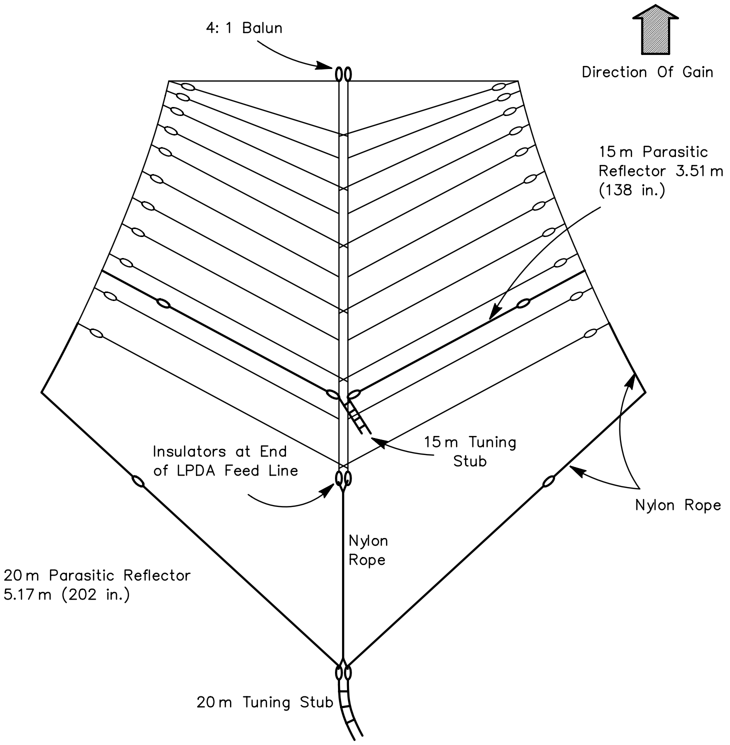
4: 1 Balun
Direction Of Gain
15 m Parasitic Reflector 3.51 m (138 in.)
Insulators at End of LPDA Feed Line
15 m Tuning Stub
Nylon Rope
Nylon Rope
20 m Parasitic Reflector 5.17 m (202 in.)
20 m Tuning Stub

Table A

Element Lengths and Spacings, in Inches

Element Number	*½ Element Length (inches)*	*Total Distance (inches)*
R1	202.0	0.0
L1	210.1	102.0
L2	191.2	140.7
R2	138.0	158.3
L3	174.0	175.9
L4	158.3	207.9
L5	144.1	237.0
L6	131.1	261.5
L7	119.3	285.6
L8	108.6	307.6
L9	98.8	327.6
L10	89.9	345.8
L11	82.4	364.5

Note: the reflector lengths do not include the length of the tuning stubs.

Fig A—Physical layout of modified Telerana with 20 and 15-meter reflectors added (in place of first two elements in original Telerana). Note the tuning stubs for the added reflectors.

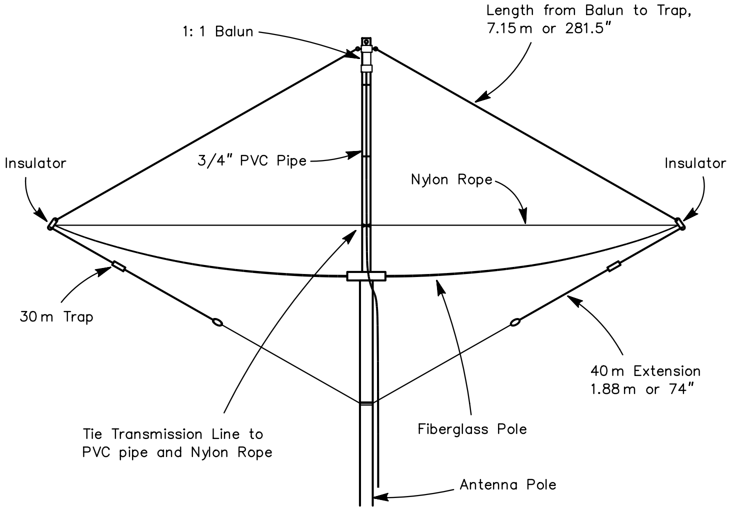
1: 1 Balun
Length from Balun to Trap, 7.15 m or 281.5"
Insulator
3/4" PVC Pipe
Nylon Rope
Insulator
30 m Trap
40 m Extension 1.88 m or 74"
Tie Transmission Line to PVC pipe and Nylon Rope
Fiberglass Pole
Antenna Pole

Fig B—Side view of 30/40-meter addition to the modified Telerana, using $^3/_4$-inch PVC pipe as a vertical stabilizer and support for the 30/40-meter trapped dipole.

should be 30.24 feet.

Now lift both bows to their upright position and tie the feeder line and the cross-arm bowstring together where they cross, directly over and approximately 3 feet above the hub.

The next step is to install the number 1 rear element from the rear egg insulators to the right and left cross-arms using other egg insulators to provide the proper element length. Be sure to solder the element halves to the transmission line. Complete this portion of the construction by installing the nylon cord catenaries from the front arm to the cross-arm tips. Use egg insulators where needed to prevent cutting the nylon cords.

When preparing the fiberglass forward spreader, keep in mind that it should be 14.75 feet long before bowing and is approximately 13.75 feet across when bowed. Secure the center of the bowstring to the end of the front arm. Lay the spreader on top of the feed line, then tie the feeder to the spreader with nylon fish line. String the catenary from the spreader tips to the cross-arm tips.

At this point of assembly, prepare antenna elements 2 through 13. There will be two segments for each element. At the outer tip make a small loop and solder the wrap. The loop will be for the nylon leader. Measure the length plus 0.4 inch for wrapping and soldering the element segment to the feeder. Seven-strand #22 antenna wire is suggested for the element wires. Slide the feedline insulators to their proper position and secure them temporarily.

The drawings show the necessary transposition scheme. Each element half of elements 1, 3, 5, 7, 9, 11 and 13 is connected to its own side of the feeder, while elements 2, 4, 6, 8, 10 and 12 cross over to the opposite side of the transmission line.

There are four holes in each of the transmission-line insulators (see Fig 22). The inner holes are for the transmission line, and the outer ones are for the elements. Since the array elements are slanted forward, they should pass through the insulator from front to back, then back over the insulator to the front side and be soldered to the transmission line. The small drawings of Fig 22 show the details of the element transpositions.

Everywhere that lines cross, tie them together with nylon line, including all copper-nylon and nylon-nylon junctions. Careful tying makes the array much more rigid. However, all elements should be mounted loosely before you try to align the whole thing. Tightening any line or element affects all the others. There will be plenty of walking back and forth before the array is aligned properly. Expect the array to be firm but not extremely taut.

The Pounder: A Single-Band 144-MHz LPDA

The 4-element Pounder LPDA pictured in **Fig 27** was developed by Jerry Hall, K1TD, for the 144-148 MHz band. Because it started as an experimental antenna, it utilizes some unusual construction techniques. However, it gives a very good account of itself, exhibiting a theoretical free-space gain of about 7.2 dBi and a front-to-back ratio of 20 dB or better. The Pounder is small and light. It weighs just 1 pound, and hence its name. In addition, as may be seen in **Fig 28**, it can be disassembled and reassembled quickly, making it an excellent antenna for portable use. This array also serves well as a fixed-station antenna, and may be changed easily to either vertical or horizontal polarization.

The antenna feeder consists of two lengths of $^1/_2 \times {}^1/_2 \times {}^1/_{16}$-inch angle aluminum. The use of two facing flat surfaces permits the builder to obtain a lower characteristic impedance than can be obtained from round conductors with the same spacing. The feeder also serves as the boom for the Pounder. In the first experimental model, the array contained only two elements with a spacing of 1 foot, so a boom length of 1 foot was the primary design requirement for the 4-element version. **Table 7** gives the calculated design data for the 4-element array.

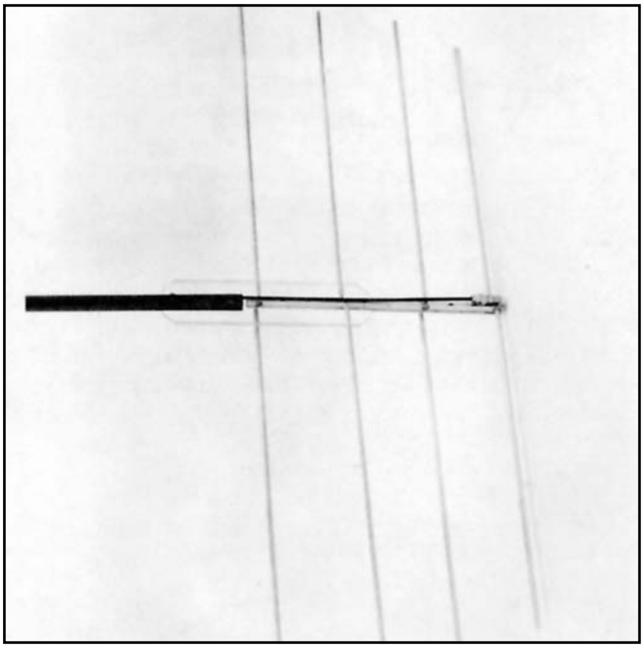

Fig 27—The 144 MHz Pounder. The boom extension at the left of the photo is a 40-inch length of slotted PVC tubing, $^7/_8$-inch outer diameter. The tubing may be clamped to the side of a tower or attached to a mast with a small boom-to-mast plate. Rotating the tubing at the clamp will provide for either vertical or horizontal polarization.

Construction

You can see the general construction approach for

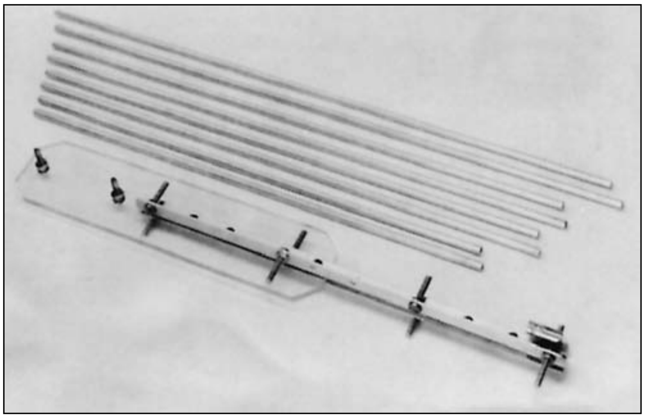

Fig 28—One end of each half element is tapped to fasten onto boom-mounted screws. Disassembly of the array consists of merely unscrewing 8 half elements from the boom. The entire disassembled array creates a small bundle only 21 inches long.

Table 7

Design Parameters for the 144-MHz Pounder

$f1 = 143$ MHz
$f_n = 148$ MHz
$B = 1.0350$
$\tau = 0.92$
$\sigma = 0.053$
Gain = 7.2 dBi = 5.1 dBd
$\cot \alpha = 2.6500$
$B_{ar} = 1.2306$
$B_s = 1.2736$
L = 0.98 feet
N = 3.90 elements (increase to 4)
Zt = none
$R_0 = 52\ \Omega$
$Z_{AV} = 312.8\ \Omega$
$\sigma' = 0.05526$
$Z_0 = 75.1\ \Omega$

Element lengths:
$\ell 1 = 3.441$ feet
$\ell 2 = 3.165$ feet
$\ell 3 = 2.912$ feet
$\ell 4 = 2.679$ feet
Element spacings:
$d_{12} = 0.365$ feet
$d_{23} = 0.336$ feet
$d_{34} = 0.309$ feet
Element diameters:
All = 0.25 inches
ℓ/diameter ratios:
$\ell 4/\text{diam}_4 = 128.6$
$\ell 3/\text{diam}_3 = 139.8$
$\ell 2/\text{diam}_2 = 151.9$
$\ell 1/\text{diam}_1 = 165.1$

Antenna feeder: $^1/_2 \times {}^1/_2 \times {}^1/_{16}''$ angle aluminum spaced $^1/_4''$
Balun: 1:1 (see text)
Feed line: 52-Ω coax (see text)

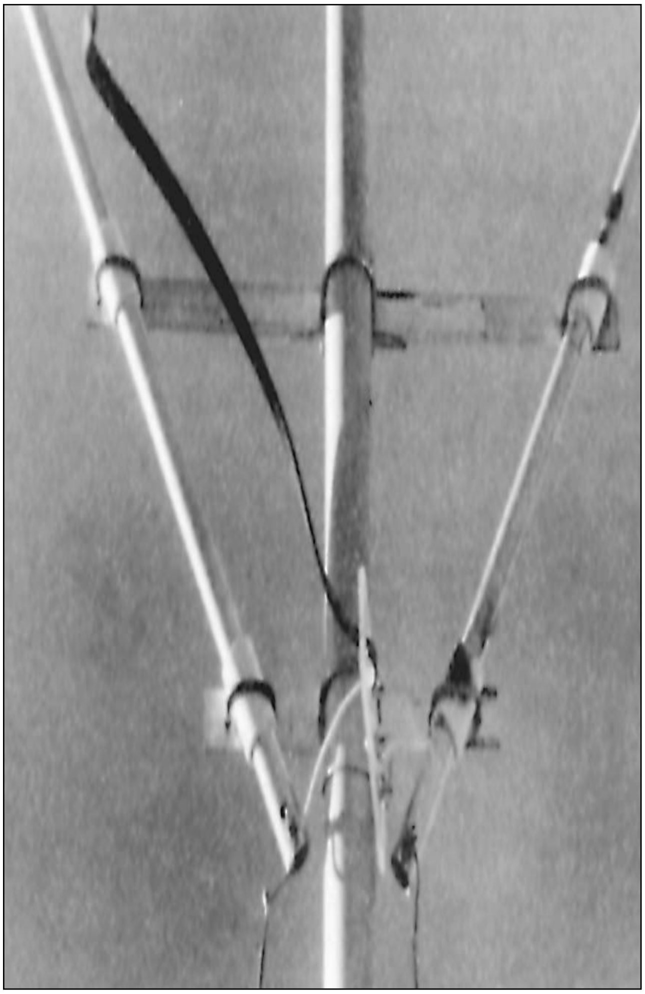

Fig 29—A close-up view of the boom, showing an alternative mounting scheme. This photo shows an earlier 2-element array, but the boom construction is the same for 2 or 4 elements. See the text for details.

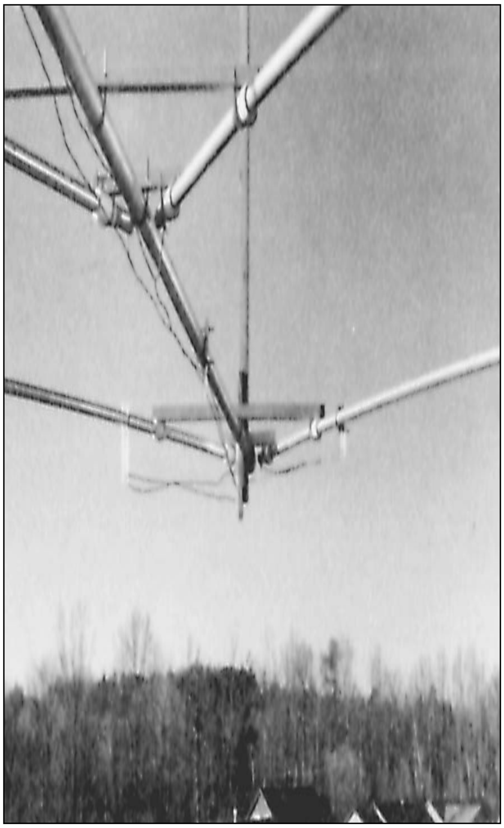

Fig 30—The feed arrangement, using a right-angle chassis-mounted BNC connector, modified by removing a portion of the flange. A short length of bus wire connects the center pin to the opposite feeder conductor.

the Pounder in the photographs. Drilled and tapped pieces of Plexiglas sheet, 1/4-inch thick, serve as insulating spacers for the angle aluminum feeder. Two spacers are used, one near the front and one near the rear of the array. Four #6-32 × 1/4-inch pan head screws secure each aluminum angle section to the Plexiglas spacers, as shown in **Figs 29** and **30**. Use flat washers with each screw to prevent it from touching the angle stock on the opposite side of the spacer. Be sure the screws are not so long as to short out the feeder! A clearance of about 1/16 inch is sufficient. If you have doubts about the screw lengths, check the assembled boom for a short with your ohmmeter on a Megohm range.

Either of two mounting techniques may be used for the Pounder. As shown in Figs 27 and 28, the rear spacer measures 10 × 2 1/2 inches, with 45° corners to avoid sharp points. This spacer also accommodates a boom extension of PVC tubing, which is attached with two #10-32 × 1-inch screws. This tubing provides for side mounting the Pounder away from a mast or tower.

An alternative support arrangement is shown in Fig 29. Two 1/2 × 3-inch Plexiglas spacers are used at the front and rear of the array. Each spacer has four holes drilled 5/8 inch apart and tapped with #6-32 threads. Two screws enter each spacer from either side to make a tight aluminum-Plexiglas-aluminum sandwich. At the center of the boom, secured with only two screws, is a 2 × 18-inch strip of 1/4-inch Plexiglas. This strip is slotted about 2 inches from each end to accept hose clamps for mounting the Pounder atop a mast. As shown, the strip is attached for vertical polarization. Alternate mounting holes, visible on the now-horizontal lip of the angle stock, provide for horizontal polarization. Although sufficient, this mounting arrangement is not as sturdy as that shown in Fig 27.

The elements are lengths of thick-wall aluminum tubing, 1/4-inch OD. The inside wall conveniently accepts a #10-32 tap. The threads should penetrate the tubing to a depth of at least 1 inch. Eight #10-32 × 1-inch screws are attached to the boom at the proper element spacings and held in place with #10-32 nuts, as shown in Fig 28. For assembly, the elements are then simply screwed into place.

Note that with this construction arrangement, the two halves of any individual element are not precisely collinear; their axes are offset by about 3/4 inch. This offset does not seem to affect performance.

The Feed Arrangement

Use care initially in mounting and cutting the elements to length. To obtain the 180° crossover feed arrangement, the element halves from a single section of the feeder/boom must alternate directions. That is, for half-elements attached to one of the two pieces of angle stock, elements 1 and 3 will point to one side, and elements 2 and 4 to the other. This arrangement may be seen by observing the element-mounting screws in Fig 28. Because of this mounting scheme, the length of tubing for an element "half" is not simply half of the length given in Table 7. After final assembly, halves for elements 2 and 4 will have a slight overlap, while elements 1 and 3 are extended somewhat by the boom thickness. The best procedure is to cut each assembled element to its final length by measuring from tip to tip.

The Pounder may be fed with RG-58 or RG-59 coax and a BNC connector. A modified right-angle chassis-mount BNC connector is attached to one side of the feeder/boom assembly for cable connection, Fig 30. The modification consists of cutting away part of the mounting flange that would otherwise protrude from the boom assembly. This leaves only two mounting-flange holes, but these are sufficient for a secure mount. A short length of small bus wire connects the center pin to the opposite side of the feeder, where it is secured under the mounting-screw nut for the shortest element.

For operation, you may secure the coax to the PVC boom extension or to the mast with electrical tape. You should use a balun, especially if the Pounder is operated with vertical elements. A choke type of balun is satisfactory, formed by taping 6 turns of the coax into a coil of 3 inches diameter, but a bead balun is preferred (see Chapter 26). The balun should be placed at the point where the coax is brought away from the boom. If the mounting arrangement of Fig 29 is used with vertical polarization, a second balun should be located approximately 1/4 wavelength down the coax line from the first. This will place it at about the level of the lower tips of the elements. For long runs of coax to the transmitter, a transition from RG-58 to RG-8 or from RG-59 to RG-9 is suggested, to reduce line losses. Make this transition at some convenient point near the array.

No shorting feeder termination is used with the array described here. The antenna feeder (phase-line) Z_0 of this array is in the neighborhood of 120 Ω, and with a resulting feed-point impedance of about 72 Ω. The theoretical mean SWR with 52 Ω line is 72/52 or 1.4 to 1. Upon array completion, the measured SWR (52-Ω line) was found to be relatively constant across the band, with a value of about 1.7 to 1. The Pounder offers a better match to 72-Ω coax.

Being an all-driven array, the Pounder is more immune to changes in feed-point impedance caused by nearby objects than is a parasitic array. This became obvious during portable use when the array was operated near trees and other objects… the SWR did not change noticeably with antenna rotation toward and away from those objects. Consequently, the Pounder should behave well in a restricted environment, such as an attic. Weighing just one pound, this array indeed does give a good account of itself.

Log Periodic-Yagi Arrays

Several possibilities exist for constructing high-gain arrays that use the log periodic dipole array concept. One technique is to add parasitic elements to the LPDA to increase both the gain and the front-to-back ratio for a specific frequency within the passband of the LPDA. The LPDA-Yagi combination is simple in concept. It utilizes an LPDA group of driven elements, along with parasitic elements at normal Yagi spacings from the active elements of the LPDA.

The LPDA-Yagi combinations are endless. An example of a single-band high-gain design is a 2- or 3-element LPDA for 21.0 to 21.45 MHz with the addition of two or three parasitic directors and one parasitic reflector. The name Log-Yag (log-cell Yagi) array has been coined for these hybrid antennas. The LPDA portion of the array is of the usual design to cover the desired bandwidth, and standard Yagi design procedures are used for the parasitic elements. Information in this section is based on a Dec 1976, *QST* article by P. D. Rhodes, K4EWG, and J. R. Painter, W4BBP, "The Log-Yag Array."

THE LOG-YAG ARRAY

The Log-Yag array, with its added parasitic elements, provides higher gain and greater directivity than would be realized with the LPDA alone. Yagi arrays require a long boom and wide element spacing for wide bandwidth and high gain, because the Q of the Yagi system increases as the number of elements is increased or as the spacing between adjacent elements is decreased. An increase in the Q of the Yagi array means that the total operating bandwidth of the array is decreased, and the gain and front-to-back ratio specified in the design are obtainable only over small portions of the band. [Older Yagi designs did indeed exhibit the limitations mentioned here. But modern, computer-aided design has resulted in wideband Yagis, provided that sufficient elements are used on the boom to allow stagger tuning for wide-band coverage. See Chapter 11.—*Ed.*]

The Log-Yag system overcomes this difficulty by using a multiple driven element cell designed in accordance with the principles of the log periodic dipole array. Since this log cell exhibits both gain and directivity by itself, it is a more effective wide-band radiator than a simple dipole driven element. The front-to-back ratio and gain of the log cell can then be improved with the addition of a parasitic reflector and director.

It is not necessary for the parasitic element spacings to be large with respect to wavelength, since the log cell is the determining factor in the array bandwidth. As well, the element spacings within the log cell may be small with respect to a wavelength without appreciable deterioration of the cell gain. For example, decreasing the relative spacing constant (σ) from 0.1 to 0.05 will decrease the array gain by less than 1 dB.

A Practical Example

The photographs and figures show a Log-Yag array for the 14-MHz amateur band. The array design takes the form of a 4-element log cell, a parasitic reflector spaced at 0.085 λ_{max}, and a parasitic director spaced at 0.15 λ_{max} (where λ_{max} is the longest free-space wavelength within the array passband). Array gain is almost unaffected with reflector spacings from 0.08 λ to 0.25 λ, and the increase in boom length is not justified. The function of the reflector is to improve the front-to-back ratio of the log cell, while the director sharpens the forward lobe and decreases the half-power beamwidth. As the spacing between the parasitic elements and the log cell decreases, the parasitic elements must increase in length.

The log cell is designed to meet upper and lower band limits with $\sigma = 0.05$. The design parameter τ is dependent on the structure bandwidth, B_s. When the log periodic design parameters have been found, the element length and spacings can be determined.

Array layout and construction details can be seen in Figs 31 through 34. Characteristics of the array are given in **Table 8**.

The method of feeding the antenna is identical to that of feeding the log periodic dipole array without the parasitic elements. As shown in **Fig 31**, a balanced feeder is required for each log-cell element, and all adjacent elements are fed with a 180° phase shift by alternating connections. Since the Log-Yag array will be covering a relatively small bandwidth, the radiation resistance of the narrow-band log cell will vary from 80 to 90 Ω (tubing

Table 8
Log-Yag Array Characteristics

1) Frequency range	14 to 14.35 MHz
2) Operating bandwidth	B = 1.025
3) Design parameter	$\tau = 0.946657$
4) Apex half angle	$\alpha = 14.921°$; cot $\alpha = 3.753$
5) Half-power beamwidth	42° (14 to 14.35 MHz)
6) Bandwidth of structure	Bs = 1.17875
7) Free-space wavelength	λ_{max} = 70.28 feet
8) Log cell boom length	L = 10.0 feet
9) Longest log element	$\ell 1$ = 35.14 feet (a tabulation of element lengths and spacings is given in Table 9)
10) Forward gain (free space)	8.2 dBi
11) Front-to-back ratio	32 dB (theoretical)
12) Front-to-side ratio	45 dB (theoretical)
13) Input impedance	Z_0 = 37 Ω
14) SWR	1.3 to 1 (14 to 14.35 MHz)
15) Total weight	96 pounds
16) Wind-load area	8.5 sq feet
17) Reflector length	36.4 feet at 6.0 foot spacing
18) Director length	32.2 feet at 10.5 foot spacing
19) Total boom length	26.5 feet

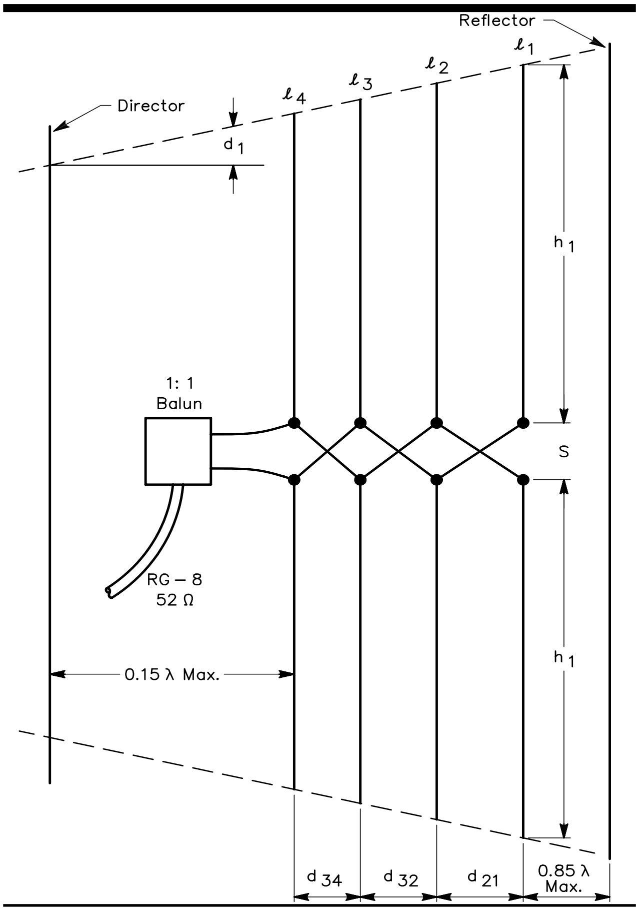
Reflector
ℓ1
ℓ2
ℓ3
ℓ4
Director
d_1
h_1
1: 1
Balun
s
RG – 8
52 Ω
h_1
0.15 λ Max.
d_{34}
d_{32}
d_{21}
0.85 λ
Max.

Fig 31—Log-Yag array layout.

Table 9
Log-Yag Array Dimensions

Element	*Length Feet*	*Spacing Feet*
Reflector	36.40	6.00 (Ref. to ℓ1)
ℓ1	35.14	3.51 (d_{12})
ℓ2	33.27	3.32 (d_{23})
ℓ3	31.49	3.14 (d_{34})
ℓ4	29.81	10.57 (ℓ4 to dir.)
Director	32.20	

elements) depending on the operating bandwidth. The addition of parasitic elements lowers the log-cell radiation resistance. Hence, it is recommended that a 1:1 balun be connected at the log-cell input terminals and 50-Ω coaxial cable be used for the feed line.

The measured radiation resistance of the 14-MHz Log-Yag is 37 Ω over the frequency range from 14.0 to 14.35 MHz. It is assumed that tubing elements will be used. However, if a wire array is used, then the radiation resistance R_0 and antenna-feeder input impedance Z_0 must be calculated so that the proper balun and coax may be used. The procedure is outlined in detail in an earlier part of this chapter. However, programs such as *LPCAD28* are also suitable to automate the calculations.

Table 9 has array dimensions. **Tables 10** and **11** contain lists of the materials necessary to build the Log-Yag array.

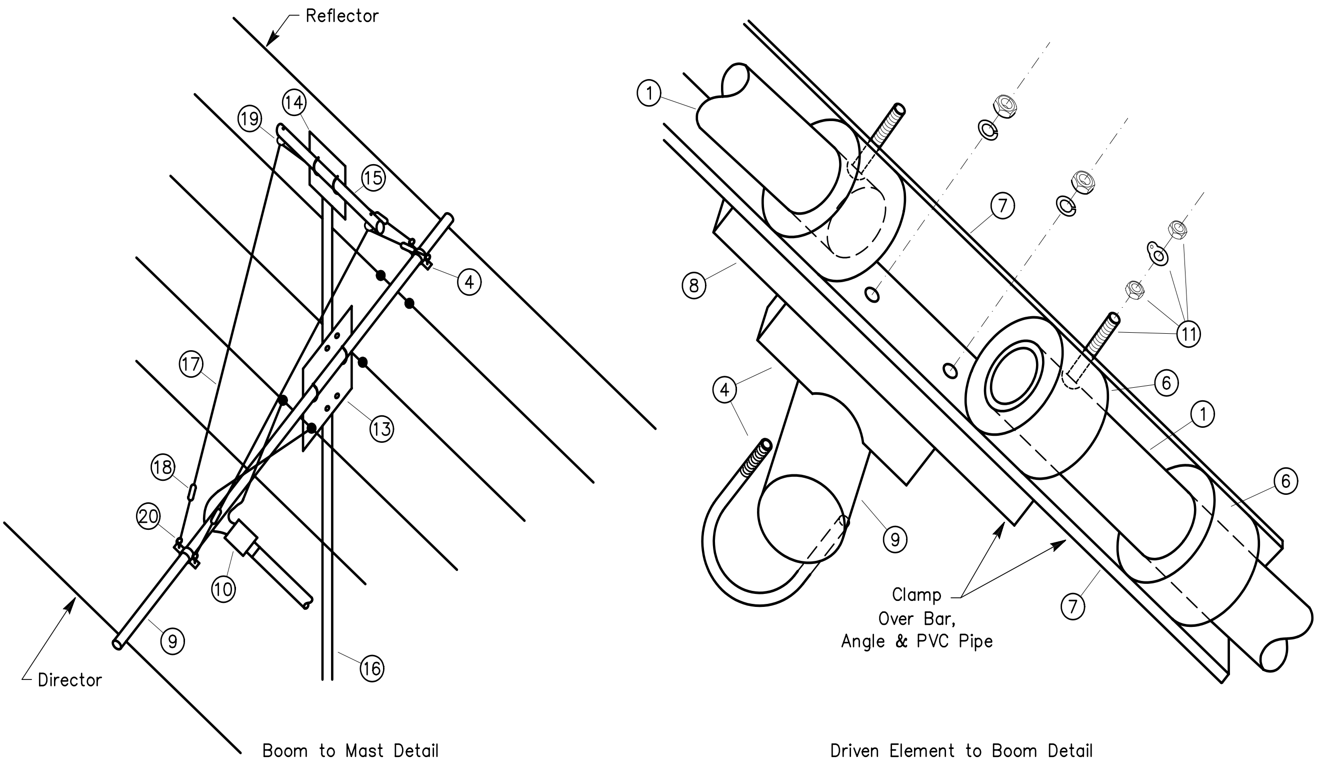
Reflector
⑭
⑲
⑮
④
⑰
⑬
⑱
⑳
⑩
⑨
⑯
Director
①
⑦
⑧
⑪
④
⑥
①
⑥
⑨
Clamp
Over Bar,
Angle & PVC Pipe
⑦

Boom to Mast Detail

Driven Element to Boom Detail

Fig 32—Assembly details. The numbered components refer to Table 11.

Table 10
Element Material Requirements: Log-Yag Array

	1-in. Tubing		*$\frac{7}{8}$-in. Tubing*		*$\frac{3}{4}$-in. Tubing*		*$1\frac{1}{4}$-in. Angle*	*$1\frac{1}{4}$-in. Bar*
	Len. Feet	*Qty*	*Len. Feet*	*Qty*	*Len. Feet*	*Qty*	*Len. Feet*	*Len. Feet*
Reflector	12	1	6	2	8	2	None	None
ℓ1	6	2	6	2	8	2	3	1
ℓ2	6	2	6	2	8	2	3	1
ℓ3	6	2	6	2	6	2	3	1
ℓ4	6	2	6	2	6	2	3	1
Director	12	1	6	2	6	2	None	None

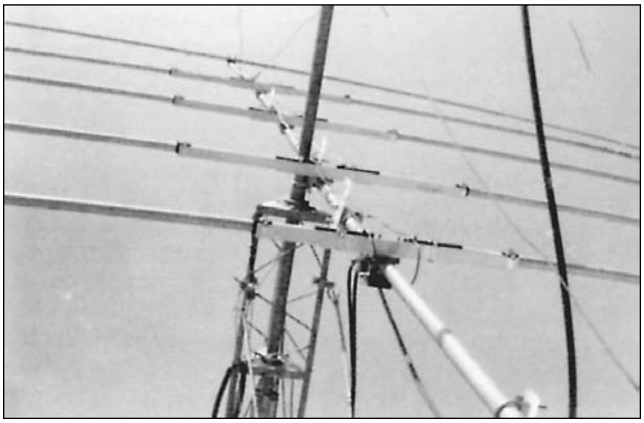

Fig 33—The attachment of the elements to the boom.

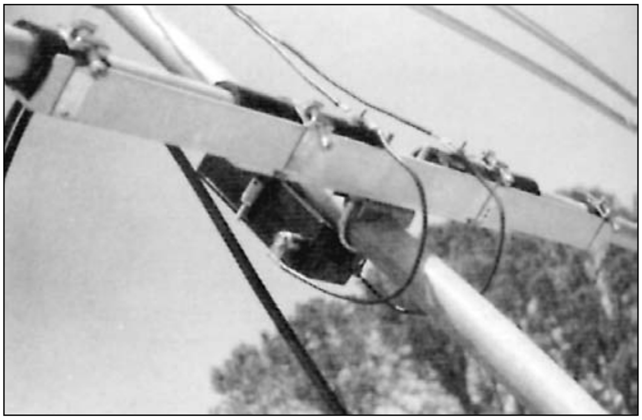

Fig 34—Looking from the front to the back of the Log-Yag array. A truss provides lateral and vertical support.

Table 11
Materials List, Log-Yag Array

1) Aluminum tubing—0.047 in. wall thickness
   1 in.—12 ft lenths, 24 lin. ft
   1 in.—12 ft or 6 ft lengths, 48 lin. ft
   $\frac{7}{8}$ in.—12 ft or 6 ft lengths, 72 lin. ft
   $\frac{3}{4}$ in.—8 ft lengths, 48 lin. ft
   $\frac{3}{4}$ in.—6 ft lengths, 36 lin. ft
2) Stainless steel hose clamps—2 in. max, 8 ea
3) Stainless steel hose clamps—$1\frac{1}{4}$ in. max, 24 ea
4) TV-type U bolts—$1\frac{1}{2}$ in., 6 ea
5) U bolts, galv. type: 5/16 in. × $1\frac{1}{2}$ in., 6 ea
5A) U bolts, galv. type: $\frac{1}{4}$ in. × 1 in., 2 ea
6) 1 in. ID water-service polyethylene pipe 160 lb/in.^{2} test, approx. $1\frac{3}{8}$ in. OD, 7 lin. ft
7) $1\frac{1}{4}$ in. × $1\frac{1}{4}$ in. × $\frac{1}{8}$ in. aluminum angle—6 ft lengths, 12 lin. ft
8) 1 in. × $\frac{1}{4}$ in. × $\frac{1}{4}$ in. aluminum angle—6 ft lengths, 6 lin. ft
9) $1\frac{1}{4}$ in. top rail of chain-link fence, 26.5 lin. ft
10) 1:1 toroid balun, 1 ea
11) No. 6-32 × 1 in. stainless steel screws, 8 ea
    No. 6-32 stainless steel nuts, 16 ea
    No. 6 solder lugs, 8 ea
12) #12 copper feed wire, 22 lin. ft
13) 12 in. × 6 in. × $\frac{1}{4}$ in. aluminum plate, 1 ea
14) 6 in. × 4 in. × $\frac{1}{4}$ in. aluminum plate, 1 ea
15) $\frac{3}{4}$ in. galv. pipe, 3 lin. ft
16) 1 in. galv. pipe—mast, 5 lin. ft
17) Galv. guy wire, 50 lin. ft
18) $\frac{1}{4}$ in. × 2 in. turnbuckles, 4 ea
19) $\frac{1}{4}$ in. × $1\frac{1}{2}$ in. eye bolts, 2 ea
20) TV guy clamps and eyebolts, 2 ea

BIBLIOGRAPHY

Source material and more extended discussion of the topics covered in this chapter can be found in the references listed below and in the textbooks listed at the end of Chapter 2.

D. Allen, "The Log Periodic Loop Array (LPLA) Antenna," *The ARRL Antenna Compendium, Vol 3*, pp 115-117.

C. A. Balanis, *Antenna Theory, Analysis and Design*, 2nd Ed. (New York: John Wiley & Sons, 1997) Chapter 9.

P. C. Butson and G. T. Thompson, "A Note on the Calculation of the Gain of Log-Periodic Dipole Antennas," *IEEE Trans on Antennas and Propagation*, Vol AP-24, No. 1, Jan 1976, pp 105-106.

R. L. Carrel, "The Design of Log-Periodic Dipole Antennas," *1961 IRE International Convention Record*.

L. B. Cebik, “The Monoband Log-Cell Yagi Revisited,” (4-part series) *National Contest Journal*, Jan-Jul 2000.

L. B. Cebik, “Some Preliminary Notes on Standard Design LPDAs for 3-30 MHz,” (2-part series), *QEX*, Mar-May 2000.

R. H. DuHamel and D. E. Isbell, “Broadband Logarithmically Periodic Antenna Structures,” *1957 IRE National Convention Record*, Part 1.

A. Eckols, “The Telerana: A Broadband 13- to 30-MHz Directional Antenna,” *QST*, Jul 1981, pp 24-27.

J. Fisher, “Development of the W8JF Waveram: A Planar Log-Periodic Quad Array,” *The ARRL Antenna Compendium, Vol 1*, pp 50-54.

M. Hansen, “The Improved Telerana, with Bonus 30/40 meter Coverage,” *The ARRL Antenna Compendium, Vol 4*, pp 112-117.

K. Heitner, “A Wide-Band, Low-Z Antenna—New Thoughts on Small Antennas,” *The ARRL Antenna Compendium, Vol 1*, pp 48-49.

D. E. Isbell, “Log-Periodic Dipole Arrays,” *IRE Transactions on Antennas and Propagation*, Vol. AP-8, No. 3, May 1960.

J. D. Kraus, *Antennas*, 2nd Ed. (New York: McGraw-Hill, 1988), Chapter 15.

R. A. Johnson, ed., *Antenna Engineering Handbook*, 3rd Ed. (New York: McGraw-Hill, 1993), Chapters 14 and 26.

K. Luetzelschwab, “Log Periodic Dipole Array Improvements,” *The ARRL Antenna Compendium, Vol 6*, pp 74-76.

D. A. Mack, “A Second-Generation Spiderweb Antenna,” *The ARRL Antenna Compendium, Vol 1*, pp 55-59.

P. E. Mayes and R. L. Carrel, “Log Periodic Resonant-V Arrays,” *IRE Wescon Convention Record*, Part 1, 1961.

P. E. Mayes, G. A. Deschamps, and W. T. Patton, “Backward Wave Radiation from Periodic Structures and Application to the Design of Frequency Independent Antennas,” *Proc. IRE*.

J. J. Meyer, “A Simple Log-Yag Array for 50 MHz,” *The ARRL Antenna Compendium, Vol 1*, pp 62-63.

C. T. Milner, “Log Periodic Antennas,” *QST*, Nov 1959, pp 11-14.

W. I. Orr and S. D. Cowan, *Beam Antenna Handbook*, pp 251-253.

P. D. Rhodes, “The Log-Periodic Dipole Array,” *QST*, Nov 1973, pp 16-22.

P. D. Rhodes and J. R. Painter, “The Log-Yag Array,” *QST*, Dec 1976, pp 18-21.

P. D. Rhodes, “The Log-Periodic V Array,” *QST*, Oct 1979, pp 40-43.

P. D. Rhodes, “The K4EWG Log Periodic Array,” *The ARRL Antenna Compendium, Vol 3*, pp 118-123

V. H. Rumsey, *Frequency Independent Antennas* (New York: Academic Press, 1966).

F. Scholz, “A 14-30 MHz LPDA for Limited Space,” *The ARRL Antenna Compendium, Vol 2*, pp 96-99

W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design*, 2nd Ed. (New York: John Wiley & Sons, 1998), Chapter 6.

J. J. Uhl, “Construct a Wire Log-Periodic Dipole Array for 80 or 40 Meters,” *QST*, Aug 1976, pp 21-24.

R. F. Zimmer, “Three Experimental Antennas for 15 Meters,” *CQ*, Jan 1983, pp 44-45.

R. F. Zimmer, “Development and Construction of ‘V’ Beam Antennas,” *CQ*, Aug 1983, pp 28-32.

Chapter 11

HF Yagi Arrays

Along with the dipole and the quarter-wave vertical, radio amateurs throughout the world make extensive use of the Yagi array. Hidetsugu Yagi and Shintaro Uda, two Japanese university professors, invented the Yagi in the 1920s. Uda did much of the developmental work, while Yagi introduced the array to the world outside Japan through his writings in English. Although the antenna should properly be called a *Yagi-Uda* array, it is commonly referred to simply as a *Yagi*.

The Yagi is a type of endfire multielement array. At the minimum, it consists of a single driven element and a single parasitic element. These elements are placed parallel to each other, on a supporting boom spacing them apart. This arrangement is known as a 2-element Yagi. The parasitic element is termed a *reflector* when it is placed behind the driven element, opposite to the direction of maximum radiation, and is called a *director* when it is placed ahead of the driven element. See **Fig 1**. In the VHF and UHF spectrum, Yagis employing 30 or more elements are not uncommon, with a single reflector and multiple directors. See Chapter 18, VHF and UHF Antenna Systems, for details on VHF and UHF Yagis. Large HF arrays may employ 10 or more elements, and will be covered in this chapter.

The gain and directional pattern of a Yagi array is determined by the relative amplitudes and phases of the currents induced into all the parasitic elements. Unlike the directly driven multielement arrays considered in Chapter 8, Multielement Arrays, where the designer must compensate for mutual coupling between elements, proper Yagi operation *relies on* mutual coupling. The current in each parasitic element is determined by its spacing from both the driven element and other parasitic elements, and by the tuning of the element itself. Both length and diameter affect element tuning.

For about 50 years amateurs and professionals created Yagi array designs largely by "cut and try" experimental techniques. In the early 1980s, Jim Lawson, W2PV, described in detail for the amateur audience the fundamental mathematics involved in modeling Yagis. His book *Yagi Antenna Design* is highly recommended for serious antenna designers. The advent of powerful microcomputers and sophisticated computer antenna modeling software in the mid 1980s revolutionized the field of Yagi design for the radio amateur. In a matter of minutes, a computer can

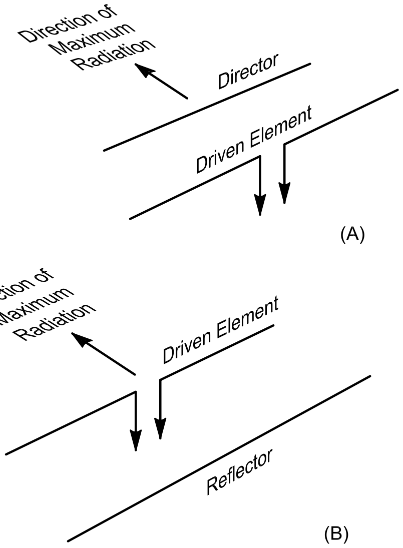
Direction of Maximum Radiation
Director
Driven Element
(A)
Direction of Maximum Radiation
Driven Element
Reflector
(B)

Fig 1—Two-element Yagi systems using a single parasitic element. At A the parasitic element acts as a director, and at B as a reflector. The arrows show the direction in which maximum radiation takes place.

try 100,000 or more different combinations of element lengths and spacings to create a Yagi design tailored to meet a particular set of high-performance parameters. To explore this number of combinations experimentally, a human experimenter would take an unimaginable amount of time and dedication, and the process would no doubt suffer from considerable measurement errors. With the computer tools available today, an antenna can be designed, constructed and then put up in the air, with little or no tuning or pruning required.

Yagi Performance Parameters

There are three main parameters used to characterize the performance of a particular Yagi—*forward gain*, *pattern* and *drive impedance/SWR*. Another important consideration is *mechanical strength*. It is very important to recognize that each of the three electrical parameters should be characterized over the frequency band of interest in order to be meaningful. Neither the gain, SWR nor the pattern measured at a single frequency gives very much insight into the overall performance of a particular Yagi.

Poor designs have been known to reverse their directionality over a frequency band, while other designs have excessively narrow SWR bandwidths, or overly "peaky" gain response. Finally, an antenna's ability to survive the wind and ice conditions expected in one's geographical location is an important consideration in any design. Much of this chapter will be devoted to describing detailed Yagi designs that are optimized for a good balance between gain, pattern and SWR over various amateur bands, and that are designed to survive strong winds and icing.

YAGI GAIN

Like any other antenna, the gain of a Yagi must be stated in comparison to some standard of reference. Designers of phased vertical arrays often state gain referenced to a single, isolated vertical element. See the section on "Phased Array Techniques" in Chapter 8, Multielement Arrays.

Many antenna designers prefer to compare gain to that of an *isotropic radiator in free space*. This is a theoretical antenna that radiates equally well in all directions, and by definition, it has a gain of 0 *dBi* (dB isotropic). Many radio amateurs, however, are comfortable using a dipole as a standard reference antenna, mainly because it is *not* a theoretical antenna.

In free space, a dipole does not radiate equally well in all directions—it has a figure-eight azimuth pattern, with deep nulls off the ends of the wire. In its favored directions, a free-space dipole has 2.15 dB gain compared to the isotropic radiator. You may see the term *dBd* in amateur literature, meaning gain referenced to a dipole in free space. Subtract 2.15 dB from gain in dBi to convert to gain in dBd.

Assume for a moment that we take a dipole out of "free space," and place it one wavelength over the ocean, whose saltwater makes an almost perfect ground. At an elevation angle of 15°, where sea water-reflected radiation adds in phase with direct radiation, the dipole has a gain of about 6 dB, compared to its gain when it was in free space, isolated from any reflections. See Chapter 3, The Effects of the Earth.

It is perfectly legitimate to say that this dipole has a gain of 6 dBd, although the term "dBd" (meaning "dB dipole") makes it sound as though the dipole somehow has gain over itself! Always remember that gain expressed in dBd (or dBi) refers to the *counterpart antenna in free space*. The gain of the dipole over saltwater in this example can be rated at either 6 dBd (over a dipole in free space), or as 8.15 dBi (over an isotropic radiator in free space). Each frame of reference is valid, as long as it is used consistently and clearly. In this chapter we will often switch between Yagis in free space and Yagis over ground. To prevent any confusion, gains will be stated in dBi.

Yagi free-space gain ranges from about 5 dBi for a small 2-element design to about 20 dBi for a 31-element long-boom UHF design. The length of the boom is the main factor determining the gain a Yagi can deliver. Gain as a function of boom length will be discussed in detail after the sections below defining antenna response patterns and SWR characteristics.

RESPONSE PATTERNS—FRONT-TO-REAR RATIO

As discussed in Chapter 2, Antenna Fundamentals, for an antenna to have gain, it must concentrate energy radiated in a particular direction, at the expense of energy radiated in other directions. Gain is thus closely related to an antenna's directivity pattern, and also to the losses in the antenna. **Fig 2** shows the *E-plane* (also called *E-field*, for electric field) and *H-plane* (also called *H-field*, for magnetic field) pattern of a 3-element Yagi in free space, compared to a dipole, and an isotropic radiator. These patterns were generated using the computer program *NEC-2*, which is highly regarded by antenna professionals for its accuracy and flexibility.

In free space there is no Earth reference to determine whether the antenna polarization is horizontal or vertical, and so its response patterns are labeled as E-field (electric) or H-field (magnetic). For a Yagi mounted over ground rather than in free space, if the

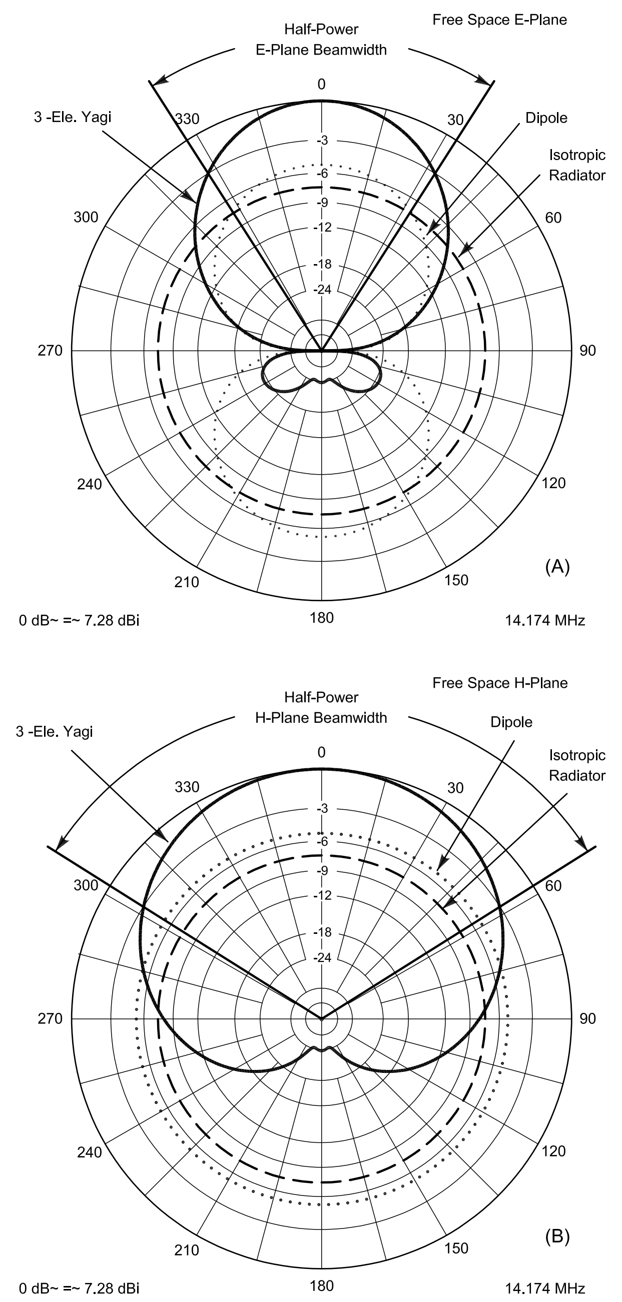

Fig 2—E-Plane (electric field) and H-Plane (magnetic field) response patterns for 3-element 20-meter Yagi in free space. At A the E-Plane pattern for a typical 3-element Yagi is compared with a dipole and an isotropic radiator. At B the H-Plane patterns are compared for the same antennas. The Yagi has an E-Plane half-power beamwidth of 66°, and an H-Plane half-power beamwidth of about 120°. The Yagi has 7.28 dBi (5.13 dBd) of gain. The front-to-back ratio, which compares the response at 0° and at 180°, is about 35 dB for this Yagi. The front-to-rear ratio, which compares the response at 0° to the largest lobe in the rearward 180° arc behind the antenna, is 24 dB, due to the lobes at 120° and 240°.

E-field is parallel to the earth (that is, the elements are parallel to the earth) then the antenna polarization is horizontal, and its E-field response is then usually referred to as its *azimuth* pattern. Its H-field response is then referred to as its *elevation* pattern.

Fig 2A demonstrates how this 3-element Yagi in free space exhibits 7.28 dBi of gain (referenced to isotropic), and has 5.13 dB gain over a free-space dipole. The gain is in the forward direction on the graph at 0° azimuth, and the forward part of the lobe is called the *main lobe*. For this particular antenna, the angular width of the E-plane main lobe at the half power, or 3 dB points compared to the peak, is about 66°. This performance characteristic is called the antenna's azimuthal *half-power beamwidth*.

Again as seen in Fig 2A, this antenna's response in the reverse direction at 180° azimuth is 34 dB less than in the forward direction. This characteristic is called the antenna's *front-to-back ratio*, and it describes the ability of an antenna to discriminate, for example, against interfering signals coming directly from the rear, when the antenna is being used for reception. In Fig 2A there are two sidelobes, at 120° and at 240° azimuth, which are about 24 dB down from the peak response at 0°. Since interference can come from any direction, not only directly off the back of an antenna, these kinds of sidelobes limit the ability to discriminate against rearward signals. The term *worst-case front-to-rear ratio* is used to describe the worst-case rearward lobe in the 180°-wide sector behind the antenna's main lobe. In this case, the worst-case front-to-rear ratio is 24 dB.

In the rest of this chapter the worst-case front-to-rear ratio will be used as a performance parameter, and will be abbreviated as "F/R." For a dipole or an isotropic radiator, Fig 2A demonstrates that F/R is 0 dB. Fig 2B depicts the H-field response for the same 3-element Yagi in free space, again compared to a dipole and an isotropic radiator in free space. Unlike the E-field pattern, the H-field pattern for a Yagi does not have a null at 90°, directly over the top of the Yagi. For this 3-element design, the H-field half-power beamwidth is approximately 120°.

Fig 3 compares the azimuth and elevation patterns for a horizontally polarized 6-element 14-MHz Yagi, with a 60-foot boom mounted one wavelength over ground, to a dipole at the same height. As with any horizontally polarized antenna, the height above ground is the main factor determining the peaks and nulls in the elevation pattern of each antenna. Fig 3A shows the E-field pattern, which has now been labeled as the Azimuth pattern. This antenna has a half-power azimuthal beamwidth of about 50°, and at an elevation angle of 12° it exhibits a forward gain of 16.02 dBi, including about 5 dB of ground reflection gain over relatively poor ground, with a dielectric constant of 13 and conductivity of 5 mS/m. In free space this Yagi has a gain of 10.97 dBi.

The H-field elevation response of the 6-element Yagi has a half-power beamwidth of about 60° in free space,

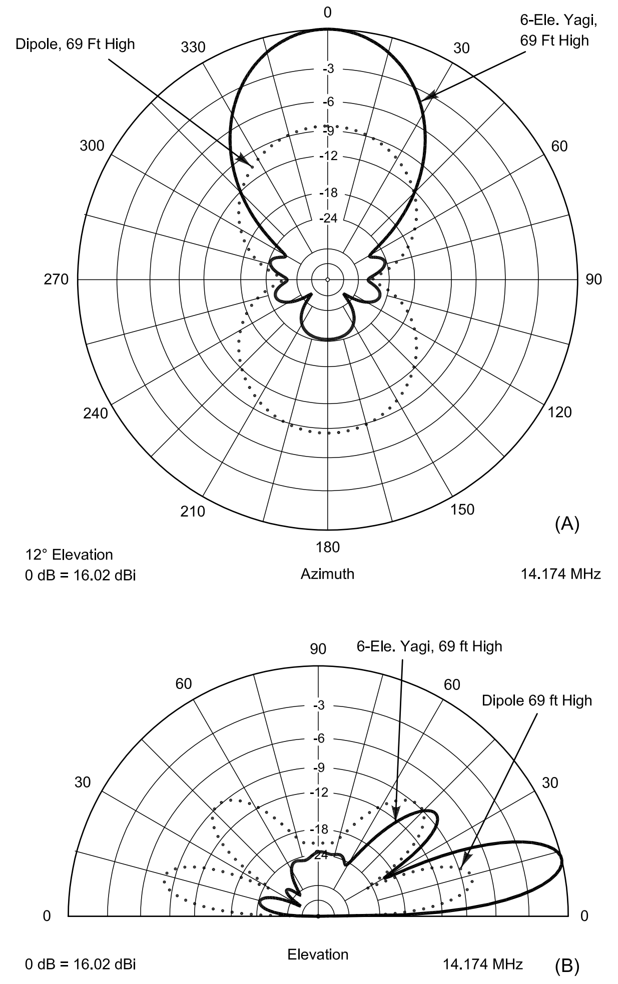

Fig 3—Azimuth pattern for 6-element 20-meter Yagi on 60-foot long boom, mounted 60 feet over ground. At A, the azimuth pattern at 12° elevation angle is shown, compared to a dipole at the same height. Peak gain of the Yagi is 16.04 dBi, or just over 8 dB compared to the dipole. At B, the elevation pattern for the same two antennas is shown. Note that the peak elevation pattern of the Yagi is compressed slightly lower compared to the dipole, even though they are both at the same height over ground. This is most noticeable for the Yagi's second lobe, which peaks at about 40°, while the dipole's second lobe peaks at about 48°. This is due to the greater free-space directionality of the Yagi at higher angles.

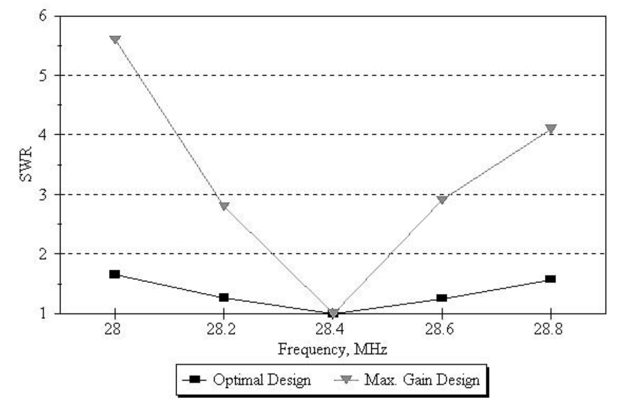

Fig 4—SWR over the 28.0 to 28.8 MHz portion of the 10-meter band for two different 3-element Yagi designs. One is designed strictly for maximum gain, while the second is optimized for F/R pattern and SWR over the frequency band. A Yagi designed only for maximum gain usually suffers from a very narrow SWR bandwidth.

but as shown in Fig 3B, the first lobe (centered at 12° in elevation) has a half-power beamwidth of only 13° when the antenna is mounted one wavelength over ground. The dipole at the same height has a very slightly larger first-lobe half-power elevation beamwidth of 14°, since its free-space H-field response is omnidirectional.

Note that the free-space H-field directivity of the Yagi suppresses its second lobe over ground (at an elevation angle of about 40°) to 8 dBi, while the dipole's response at its second lobe peak (at about 48°) is at a level of 9 dBi.

The shape of the azimuthal pattern for a Yagi operated over real ground will change slightly as the Yagi is placed closer and closer to earth. Generally, however, the azimuth pattern doesn't depart significantly from the free-space pattern until the antenna is less than 0.5 λ high. This is just over 17 feet high at 28.4 MHz, and just under 35 feet at 14.2 MHz, heights that are not difficult to achieve for most amateurs. Some advanced computer programs can optimize Yagis at the exact installation height.

DRIVE IMPEDANCE AND SWR

The impedance at the driven element in a Yagi is affected not only by the tuning of the driven element itself, but also by the spacing and tuning of nearby parasitic elements, and to a lesser extent by the presence of ground. In some designs that have been tuned solely for maximum gain, the driven-element impedance can fall to very low levels, sometimes less than 5 Ω. This can lead to excessive losses due to conductor resistance, especially at VHF and UHF. In a Yagi that has been optimized solely for gain, conductor losses are usually compounded by large excursions in impedance levels with relatively small changes in frequency. The SWR can thus change dramatically over a band and can create additional losses in the feed cable. **Fig 4** illustrates the SWR over the 28 to 28.8 MHz portion of the 10-meter amateur band for a 5-element Yagi on a 24-foot boom, which has been tuned for maximum forward gain at a spot frequency of 28.4 MHz. Its SWR curve is contrasted to that of a Yagi designed for a good compromise of gain, SWR and F/R.

Even professional antenna designers have difficulty accurately measuring forward gain. On the other hand, SWR can easily be measured by professional and amateur alike. Few manufacturers would probably want to advertise an antenna with the narrow-band SWR curve shown in Fig 4!

Monoband Yagi Performance Optimization

DESIGN GOALS

The previous section discussing driven-element impedance and SWR hinted at possible design trade-offs among gain, pattern and SWR, especially when each parameter is considered over a frequency band rather than at a spot frequency. Trade-offs in Yagi design parameters can be a matter of personal taste and operating style. For example, one operator might exclusively operate the CW portions of the HF bands, while another might only be interested in the Phone portions. Another operator may want a good pattern in order to discriminate against signals coming from a particular direction; someone else may want the most forward gain possible, and may not care about responses in other directions.

Extensive computer modeling of Yagis indicates that the parameter that must be compromised most to achieve wide bandwidths for front-to-rear ratio and SWR is forward gain. However, not much gain must be sacrificed for good F/R and SWR coverage, especially on long-boom Yagis. Although 10 and 7-MHz Yagis are not rare, the HF bands from 14 to 30 MHz are where Yagis are most often found, mainly due to the mechanical difficulties involved with making sturdy antennas for lower frequencies. The highest HF band, 28.0 to 29.7 MHz, represents the largest percentage bandwidth of the upper HF bands, at almost 6%. It is difficult to try to optimize in one design the main performance parameters of gain, worst-case F/R ratio and SWR over this large a band. Many commercial designs thus split up their 10-meter designs into antennas covering one of two bands: 28.0 to 28.8 MHz, and 28.8 to 29.7 MHz. For the amateur bands below 10 meters, optimal designs that cover the entire band are more easily achieved.

DESIGN VARIABLES

There are only a few variables available when one is designing a Yagi to meet certain design goals. The variables are:

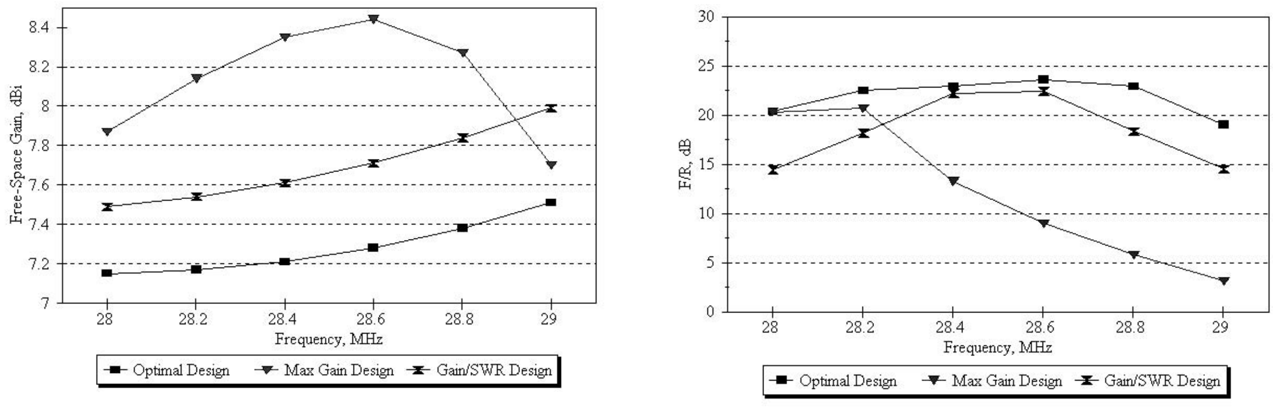

(A)

(B)

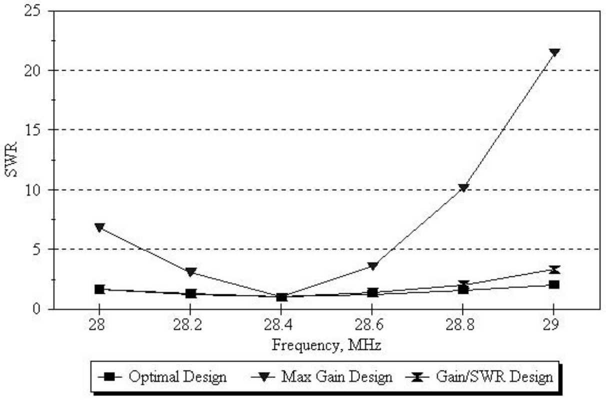

(C)

Fig 5—Comparisons of three different 3-element 10-meter Yagi designs using 8-foot booms. At A, gain comparisons are shown. The Yagi designed for the best compromise of gain and SWR sacrifices an average of about 0.5 dB compared to the antenna designed for maximum gain. The Yagi designed for optimal F/R, gain and SWR sacrifices an average of 1.0 dB compared to the maximum-gain case, and about 0.4 dB compared to the compromise gain and SWR case. At B, the front-to-rear ratio is shown for the three different designs. The antenna designed for optimal combination of gain, F/R and SWR maintains a F/R higher than 20 dB across the entire frequency range, while the antenna designed strictly for gain has a F/R of 3 dB at the high end of the band. At C, the three antenna designs are compared for SWR bandwidth. At the high end of the band, the antenna designed strictly for gain has a very high SWR.

1. The physical length of the boom
2. The number of elements on the boom
3. The spacing of each element along the boom
4. The tuning of each element
5. The type of matching network used to feed the array.

GAIN AND BOOM LENGTH

As pointed out earlier, the gain of a Yagi is largely a function of the length of the boom. As the boom is made longer, the maximum gain potential rises. For a given boom length, the number of elements populating that boom can be varied, while still maintaining the antenna's gain, provided of course that the elements are tuned properly. In general, putting more elements on a boom gives the designer added flexibility to achieve desired design goals, especially to spread the response out over a frequency band.

Fig 5A is an example illustrating gain versus frequency for three different types of 3-element Yagis on 8-foot booms. The three antennas were designed for the lower end of the 10-meter band, 28.0 to 28.8 MHz, based on the following different design goals:

Antenna 1: Maximum mid-band gain, regardless of F/R or SWR across the band
Antenna 2: SWR less than 2:1 over the frequency band; best compromise gain, with no special consideration for F/R over the band.
Antenna 3: "Optimal" case: F/R greater than 20 dB, SWR less than 2:1 over the frequency band; best compromise gain.

Fig 5B shows the F/R over the frequency band for these three designs, and Fig 5C shows the SWR curves over the frequency band. Antenna 1, the design that strives strictly for maximum gain, has a poor SWR response over the band, as might be expected after the previous section discussing SWR. The SWR is 10:1 at 28.8 MHz and rises to 22:1 at 29 MHz. At 28 MHz, at the low end of the band, the SWR of the maximum-gain design is more than 6:1. Clearly, designing for maximum gain alone produces an unacceptable design in terms of SWR bandwidth. The F/R for Antenna 1 reaches a high point of about 20 dB at the low-frequency end of the band, but falls to only 3 dB at the high-frequency end.

Antenna 2, designed for the best compromise of gain while the SWR across the band is held to less than 2:1, achieves this goal, but at an average gain sacrifice of 0.7 dB compared to the maximum gain case. The F/R for this design is just under 15 dB over the band. This design is fairly typical of many amateur Yagi designs before the advent of computer modeling and optimization programs. SWR can easily be measured, and experimental optimization for forward gain is a fairly straightforward procedure. By contrast, overall pattern optimization is not a trivial thing to achieve experimentally, particularly for antennas with more than four or five elements.

Antenna 3, designed for an optimum combination of F/R, SWR and gain, compromises forward gain an average of 1.0 dB compared to the maximum gain case, and about 0.4 dB compared to the compromise gain/SWR case. It achieves its design objectives of more than 20 dB F/R over the 28.0 to 28.8 MHz portion of the band, with an SWR less than 2:1 over that range.

Fig 6A shows the free-space gain versus frequency for the same three types of designs, but for a bigger 5-element 10-meter Yagi on a 20-foot boom. Fig 6B shows the variation in F/R, and Fig 6C shows the SWR curves versus frequency. Once again, the design that concentrates solely on maximum gain has a poor SWR curve over the band, reaching just over 6:1 toward the high end of the band. The difference in gain between the maximum gain case and the optimum design case has narrowed for this size of boom to an average of under 0.5 dB. This comes about because the designer has access to more variables in a 5-element design than he does in a 3-element design, and he can stagger-tune the various elements to spread the response out over the whole band.

Fig 7A, **B** and **C** show the same three types of designs, but for a 6-element Yagi on a 36-foot boom. The SWR bandwidth of the antenna designed for maximum gain has improved compared to the previous two shorter-boom examples, but the SWR still rises to more than 4:1 at 28.8 MHz, while the F/R ratio is pretty constant over the band, at a mediocre 11 dB average level. While the antenna designed for gain and SWR does hold the SWR below 2:1 over the band, it also has the same mediocre level of F/R performance as does the maximum-gain design.

The optimized 36-foot boom antenna achieves an excellent F/R of more than 22 dB over the whole 28.0 to 28.8 MHz band. Again, the availability of more elements and more space on the 36-foot long boom gives the designer more flexibility in broadbanding the response over the whole band, while sacrificing only 0.3 dB of gain compared to the maximum-gain design.

Fig 8A, **B**, and **C** show the same three types of 10-meter designs, but now for a 60-foot boom, populated with eight elements. With eight elements and a very long boom on which to space them out, the antenna designed solely for maximum gain can achieve a much better SWR response across the band, although the SWR does rise to more than 7:1 at the very high end of the band. The SWR remains less than 2:1 from 28.0 to 28.7 MHz, much better than for shorter-boom, maximum-gain designs. The worst-case F/R ratio is never better than 19 dB, however, and remains around 10 dB over much of the band. The antenna designed for the best compromise gain and SWR loses only about 0.1 dB of gain compared to the maximum-gain design, but does little better in terms of F/R across the band.

Contrasted to these two designs, the antenna optimized for F/R, SWR and gain has an outstanding pattern, exhibiting an F/R of more than 24 dB across the entire

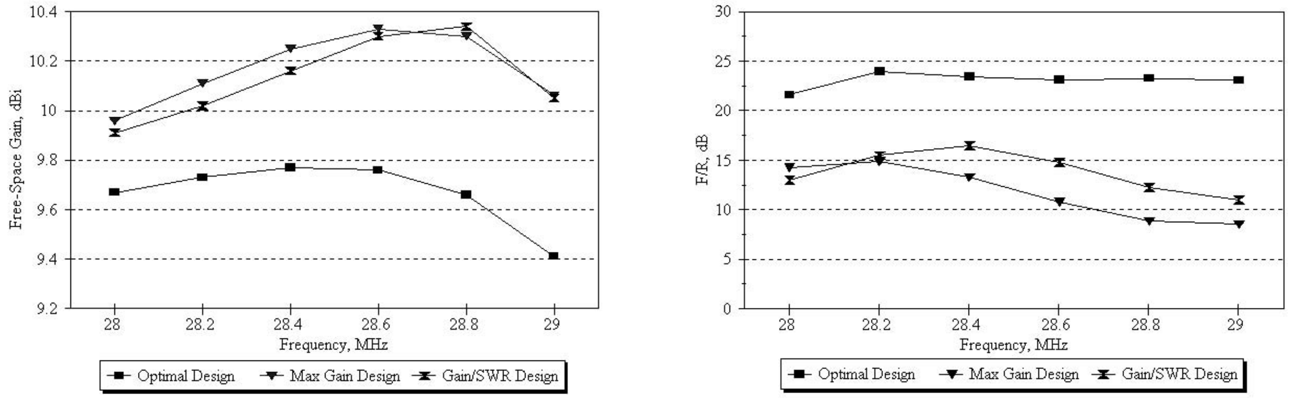

(A)

(B)

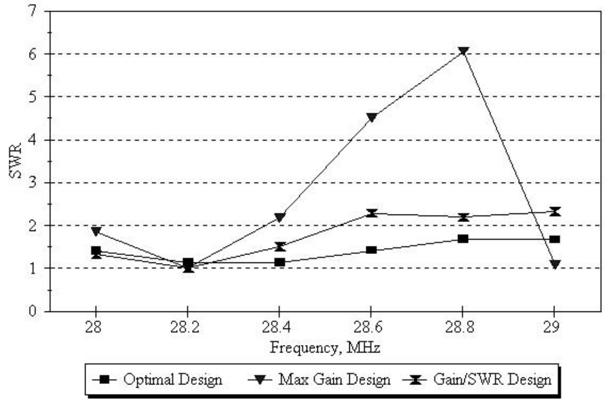

(C)

Fig 6—Comparisons of three different designs for 5-element 10-meter Yagis on 20-foot booms. At A, the gain of three different 5-element 10-meter Yagi designs are graphed. The difference in gain between the three antennas narrows because the elements can be stagger-tuned to spread the response out better over the desired frequency band. The average gain reduction for the fully optimized antenna design is about 0.5 dB. At B, the optimal antenna displays better than 22 dB F/R over the band, while the Yagi designed for gain and SWR displays on average 10 dB less F/R throughout the band. At C, the SWR bandwidth is compared for the three Yagis. The antenna designed strictly for forward gain has a poor SWR bandwidth and a high peak SWR of 6:1 at 28.8 MHz.

band, while keeping the SWR below 2:1 from 28.0 to 28.9 MHz. It must sacrifice an average of only 0.4 dB compared to the maximum gain design at the low end of the band, and actually has more gain than the maximum gain and gain/SWR designs at the high-frequency end of the band.

The conclusion drawn from these and many other detailed comparisons is that designing strictly for maximum mid-band gain yields an inferior design when the antenna is examined over an entire frequency band, especially in terms of SWR. Designing a Yagi for both gain and SWR will yield antennas that have mediocre rearward patterns, but that lose relatively little gain compared to the maximum gain case, at least for designs with more than three elements.

However, designing a Yagi for an optimal combination of F/R, SWR and gain results in a loss of gain less than 0.5 dB compared to designs designed only for gain and SWR. **Fig 9** summarizes the forward gain achieved for the three different design types versus boom length, as expressed in wavelength.

Except for the 2-element designs, the Yagis described in the rest of this chapter have the following design goals over a desired frequency band:

1. Front-to-rear ratio over the frequency band of more than 20 dB
2. SWR over the frequency band less than 2:1
3. Maximum gain consistent with points 1 and 2 above

Just for fun, **Fig 10** shows the gain versus boom length for theoretical 20-meter Yagis that have been designed to meet the three design goals above. The 31-element design for 14 MHz would be wondrous to behold. Sadly, it is unlikely that anyone will build one, considering that the boom would be 724 feet long! However, such a design *does* become practical when scaled to 432 MHz. In fact, a K1FO 22-element and a K1FO 31-element Yagi are the prototypes for the theoretical 14-MHz long-boom designs. See Chapter 18, VHF and UHF Antenna Systems.

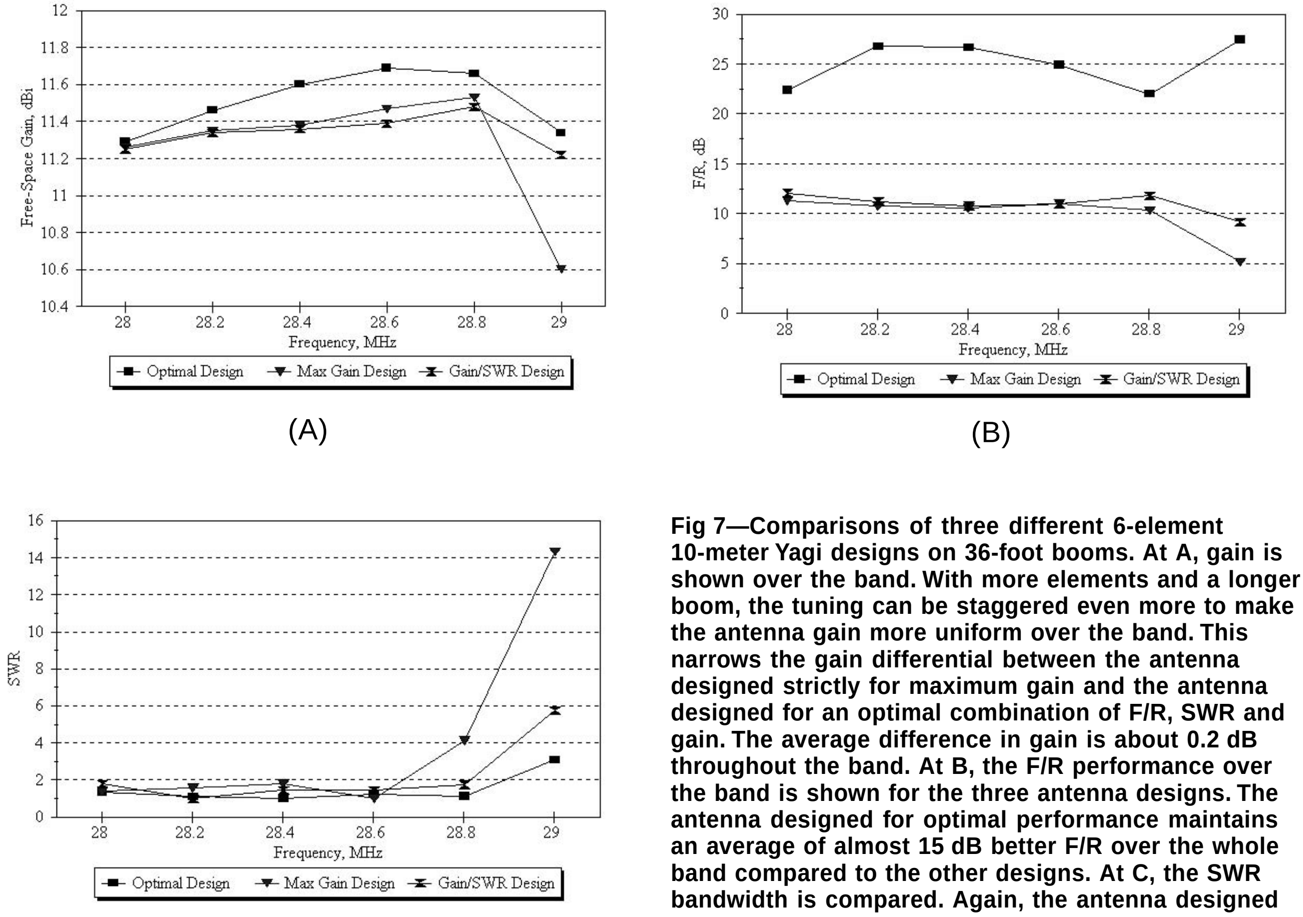

(A)

(B)

(C)

Fig 7—Comparisons of three different 6-element 10-meter Yagi designs on 36-foot booms. At A, gain is shown over the band. With more elements and a longer boom, the tuning can be staggered even more to make the antenna gain more uniform over the band. This narrows the gain differential between the antenna designed strictly for maximum gain and the antenna designed for an optimal combination of F/R, SWR and gain. The average difference in gain is about 0.2 dB throughout the band. At B, the F/R performance over the band is shown for the three antenna designs. The antenna designed for optimal performance maintains an average of almost 15 dB better F/R over the whole band compared to the other designs. At C, the SWR bandwidth is compared. Again, the antenna designed strictly for maximum gain exhibits a high SWR of 4:1 at 28.8 MHz, and rises to more than 14:1 at 29.0 MHz.

OPTIMUM DESIGNS AND ELEMENT SPACING

Two-Element Yagis

Many hams consider a 2-element Yagi to give “the most bang for the buck” among various Yagi designs, particularly for portable operations such as Field Day. A 2-element Yagi has about 4 dB of gain over a simple dipole (sometimes jokingly called a “one-element Yagi”) and gives a modest F/R of about 10 dB to help with rejection of interference on receive. By comparison, going from a 2-element to a 3-element Yagi increases the boom length by about 50% and adds another element, a 50% increase in the number of elements—for a gain increase of about 1 dB and another 10 dB in F/R.

Element Spacing in Larger Yagis

One of the more interesting results of computer modeling and optimization of high-performance Yagis with four or more elements is that a distinct pattern in the element spacings along the boom shows up consistently. This pattern is relatively independent of boom length, once the boom is longer than about 0.3 λ.

The reflector, driven element and first director of these optimal designs are typically bunched rather closely together, occupying together only about 0.15 to 0.20 λ of the boom. This pattern contrasts sharply with older designs, where the amount of boom taken up by the reflector, driven element and first director was typically more than 0.3 λ. **Fig 11** shows the element spacings for an optimized 6-element, 36-foot boom, 10-meter design, compared to a W2PV 6-element design with constant spacing of 0.15 λ between all elements.

A problem arises with such a bunching of elements toward the reflector end of the boom—the wind loading of the antenna is not equal along the boom. Unless properly compensated, such new-generation Yagis will act like windvanes, punishing, and often breaking, the rotators trying to turn, or hold, them in the wind. One successful solution to windvaning has been to employ “dummy elements” made of PVC piping. These nonconducting elements are placed on the boom close to the last director so

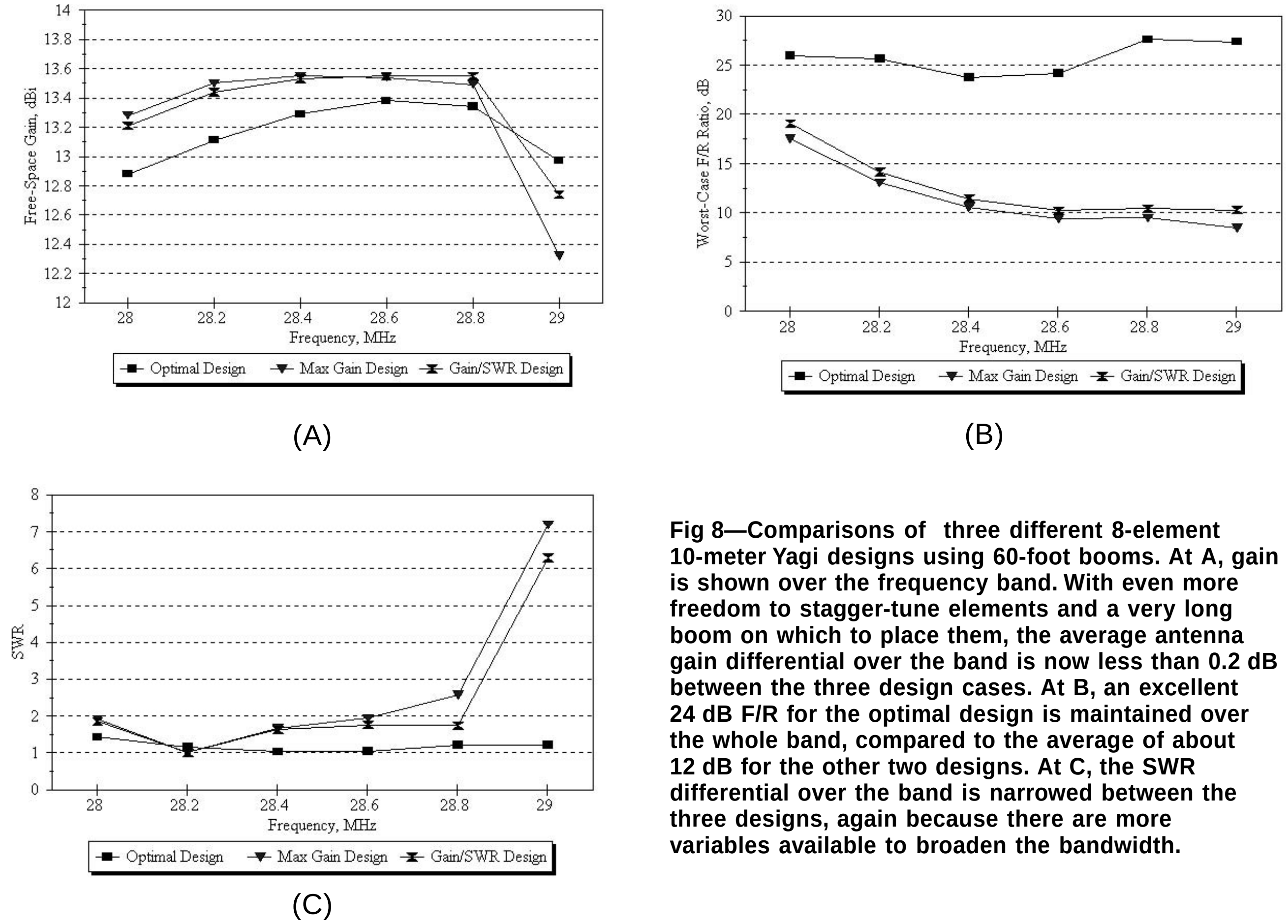

Fig 8—Comparisons of three different 8-element 10-meter Yagi designs using 60-foot booms. At A, gain is shown over the frequency band. With even more freedom to stagger-tune elements and a very long boom on which to place them, the average antenna gain differential over the band is now less than 0.2 dB between the three design cases. At B, an excellent 24 dB F/R for the optimal design is maintained over the whole band, compared to the average of about 12 dB for the other two designs. At C, the SWR differential over the band is narrowed between the three designs, again because there are more variables available to broaden the bandwidth.

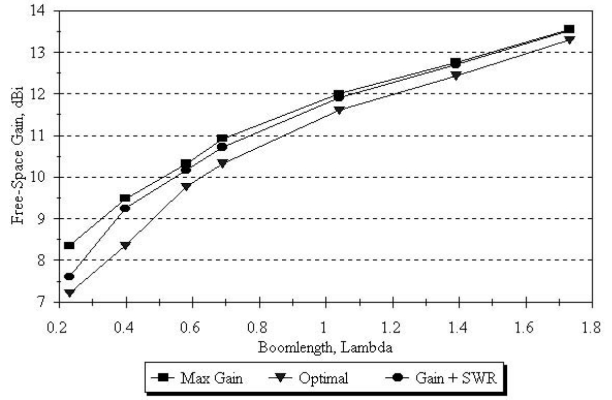

Fig 9—Gain versus boom length for three different 10-meter design goals. The goals are: (1) designed for maximum gain across band, (2) designed for a compromise of gain and SWR, and (3) designed for optimal F/R, SWR and gain across 28.0 to 28.8 MHz portion of 10-meter band. The gain difference is less than 0.5 dB for booms longer than approximately 0.5 λ.

the windload is equalized at the mast-to-boom bracket. In addition, it may be necessary to insert a small amount of lead weight at one end of the boom in order to balance the antenna weight.

Despite the relatively close spacing of the reflector, driven element and first director, modern optimal Yagi designs are not overly sensitive to small changes in either element length or spacing. In fact, these antennas can be constructed from design tables without excessive concern about close dimensional tolerances. In the HF range up to 30 MHz, building the antennas to the nearest $^1/_8$-inch results in performance remarkably consistent with the computations, without any “tweaking” or fine-tuning when the Yagi is on the tower.

ELEMENT TUNING

Element tuning (or *self-impedance*) is a complex function of the effective electrical length of each element and the effective diameter of the element. In turn, the effective length and diameter of each element is related to

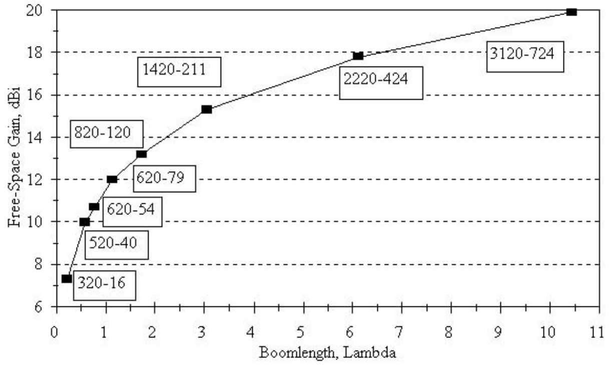

Fig 10—Theoretical gain versus boom length for 20-meter Yagis designed for optimal combination of F/R, SWR and gain across the entire 14.0 to 14.35 MHz band. The theoretical gain approaches 20 dBi for a gigantic 724-foot boom, populated with 31 elements. Such a design on 20 meters is not too practical, of course, but can readily be achieved on a 24-foot boom on 432 MHz.

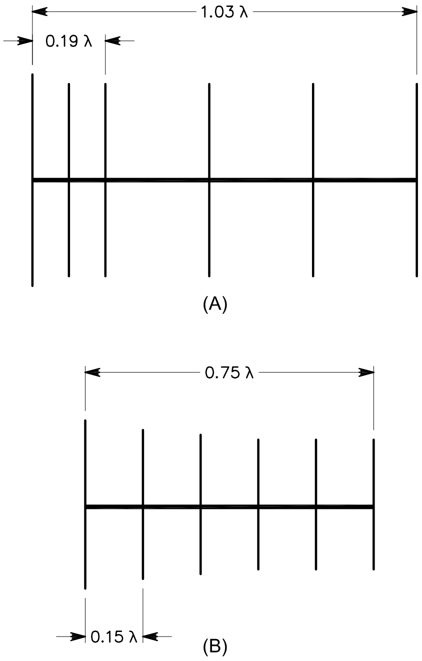

Fig 11—Tapering spacing versus constant element spacing. At A, illustration of how the spacing of the reflector, driven element and first director (over the first 0.19 λ of the boom) of an optimally designed Yagi is bunched together compared to the Yagi at B, which uses constant 0.15 λ spacing between all elements. The optimally designed antenna has more than 22 dB F/R and an SWR less than 1.5:1 over the frequency band 28.0 to 28.8 MHz.

the taper schedule (if telescoping aluminum tubing is used, the most common method of construction), the length of each telescoping section, the type and size of mounting bracket used to secure the element to or through the boom, and the size of the Yagi boom itself. See the section entitled "Antenna Frequency Scaling," and "Tapered Elements" in Chapter 2, Antenna Fundamentals, of this book for details about element tuning as a function of tapering and element diameter. Note especially that Yagis constructed using wire elements will perform very differently compared to the same antenna constructed with elements made of telescoping aluminum tubing.

The process by which a modern Yagi is designed usually starts out with the selection of the longest boom possible for a given installation. A suitable number of elements of a given taper schedule are then placed on this boom, and the gain, pattern and SWR are calculated over the entire frequency band of interest to the operator. Once an electrical design is chosen, the designer must then ensure the mechanical integrity of the antenna design. This involves verifying the integrity of the boom and each element in the face of the wind and ice loading expected for a particular location. The section entitled "Construction with Aluminum Tubing" in Chapter 20, Antenna Materials and Accessories, of this book shows details of tapered telescoping aluminum elements for the upper HF bands. In addition, the ARRL book *Physical Design of Yagi Antennas*, by Dave Leeson, W6NL (ex-W6QHS), describes the mechanical design process for all portions of a Yagi antenna very thoroughly, and is highly recommended for serious Yagi builders.

Specific Monoband Yagi Designs

The detailed Yagi design tables that follow are for two taper schedules for HF Yagis covering the 14 through 30-MHz amateur bands. The heavy-duty elements are designed to survive at least 120-mph winds without icing, or 85-mph winds with $^1/_4$-inch radial ice. The medium-duty elements are designed to survive winds greater than 80 mph, or 60-mph winds with $^1/_4$-inch radial ice.

For 10.1 MHz, the elements shown are capable of surviving 105-mph winds, or 93-mph winds with $^1/_4$-inch radial ice. For 7.1 MHz the elements shown can survive 93-mph winds, or 69-mph winds with $^1/_4$-inch radial ice. For these two lower frequency bands, the elements and the booms needed are very large and heavy. Mounting, turning and keeping such antennas in the air is not a trivial task.

Each element is mounted above the boom with a heavy rectangular aluminum plate, by means of U-bolts with saddles, as shown in Fig 35 in Chapter 18, VHF and UHF Antenna Systems for a 6-meter Yagi. This method of element mounting is rugged and stable, and because the element is mounted away from the boom, the amount of element detuning due to the presence of the boom is minimal. The element dimensions given in each table already take into account any element detuning due to the boom-to-element mounting plate. For each element, the length of the tip determines the tuning, since the inner tubes are fixed in diameter and length.

Half Elements

Each design shows the dimensions for *one-half* of each element, mounted on *one side* of the boom. The other half of each element is symmetrical, mounted on the other side of the boom. The use of a tubing sleeve inside the center portion of the element is recommended, so that the element is not crushed by the mounting U-bolts. Unless otherwise noted, each section of tubing is made of 6061-T6 aluminum tubing, with a 0.058-inch wall thickness. This wall thickness ensures that the next standard size of tubing can telescope with it. Each telescoping section is inserted 3 inches into the larger tubing, and is secured by one of the methods shown in Fig 11 in Chapter 20, Antenna Materials and Accessories.

Matching System

Each antenna is designed with a driven-element length appropriate for a hairpin type of matching network. The driven-element's length may require slight readjustment for best match, particularly if a different matching network is used. *Do not change* either the lengths or the telescoping tubing schedule of the parasitic elements—they have been optimized for best performance and will not be affected by tuning of the driven element!

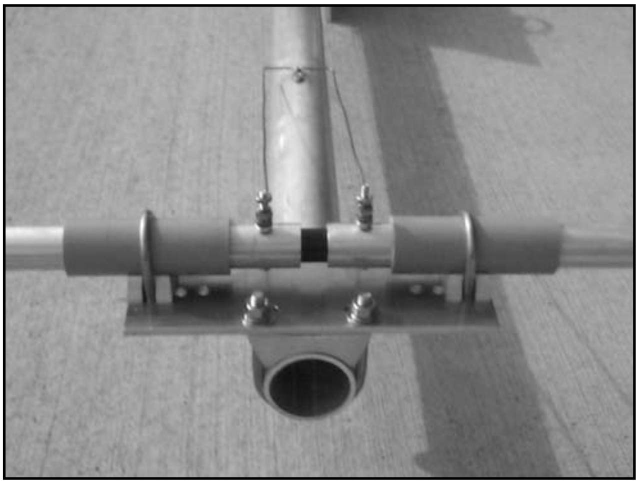

Fig 12—Typical construction techniques for an HF Yagi. This photo shows a hairpin match on a driven element that uses a fiberglass insulator (wrapped in black vinyl tape for protection against UV). Muffler clamps and saddles mount the element to the boom, while U-bolts and saddles mount the element to the boom-to-element plate. The gray PVC sleeves insulate the element from the plate. The feed coax is connected to the two bolts that also connect to the hairpin wire. Note that the hairpin is grounded at its opposite end to dissipate static charges that might otherwise build up.

Fig 12 is a photograph of the driven element for a 2-element 17-meter Yagi built by Chuck Hutchinson, K8CH, for the ARRL book *Simple and Fun Antennas for Hams.* The aluminum tubing on each side of the boom was 1-inch OD, and the two pieces were mechanically joined together with a $^3/_4$-inch OD fiberglass insulator. Chuck wound electrical tape over the insulator to protect the fiberglass from the sun's UV.

Chuck used 3-inch lengths of 1-inch sunlight-resistant PVC conduit, split lengthwise, to make the grey outer insulators for the driven element. The aluminum plates came from DX Engineering, as did the stainless-steel U-bolts and saddle clamps. These saddles ensured that the elements don't rotate on the 2-inch OD boom in the heavy winds in his part of rural Michigan.

You can see the bolts used to pin the center fiberglass insulator to the aluminum tubing, while also providing an electrical connection for the #12 hairpin wire and for the feed-line coax, which uses ferrite beads over the coax's outer vinyl jacket to make a common-mode current-type of balun (not shown in Fig 12). Note that the center of the hairpin is connected to the boom using a grounding lug for some measure of protection from static buildup.

10-METER YAGIS

Fig 13 describes the electrical performance of eight optimized 10-meter Yagis with boom lengths between 6 to 60 feet. The end of each boom includes 3 inches of space for the reflector and last-director (or driven element for the 2-element designs) mounting plates. Fig 13A shows the free-space gain versus frequency for each antenna; 13B shows the front-to-rear ratio, and 13C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the lower half of the 10-meter band from 28.0 to 28.8 MHz, with SWR less than 2:1 and F/R better than 20 dB over that range.

Fig 13D shows the taper schedule for two types of 10-meter elements. The heavy-duty design can survive 125-mph winds with no icing, and 88-mph winds with $^1/_4$-inch of radial ice. The medium-duty design can handle 96-mph winds with no icing, and 68-mph winds with $^1/_4$-inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.250-inch thick flat aluminum plate, 4 inches wide by 4 inches long. Each element except for the insulated driven element, is centered on the plate, held by two stainless-steel U-bolts with saddles. Another set

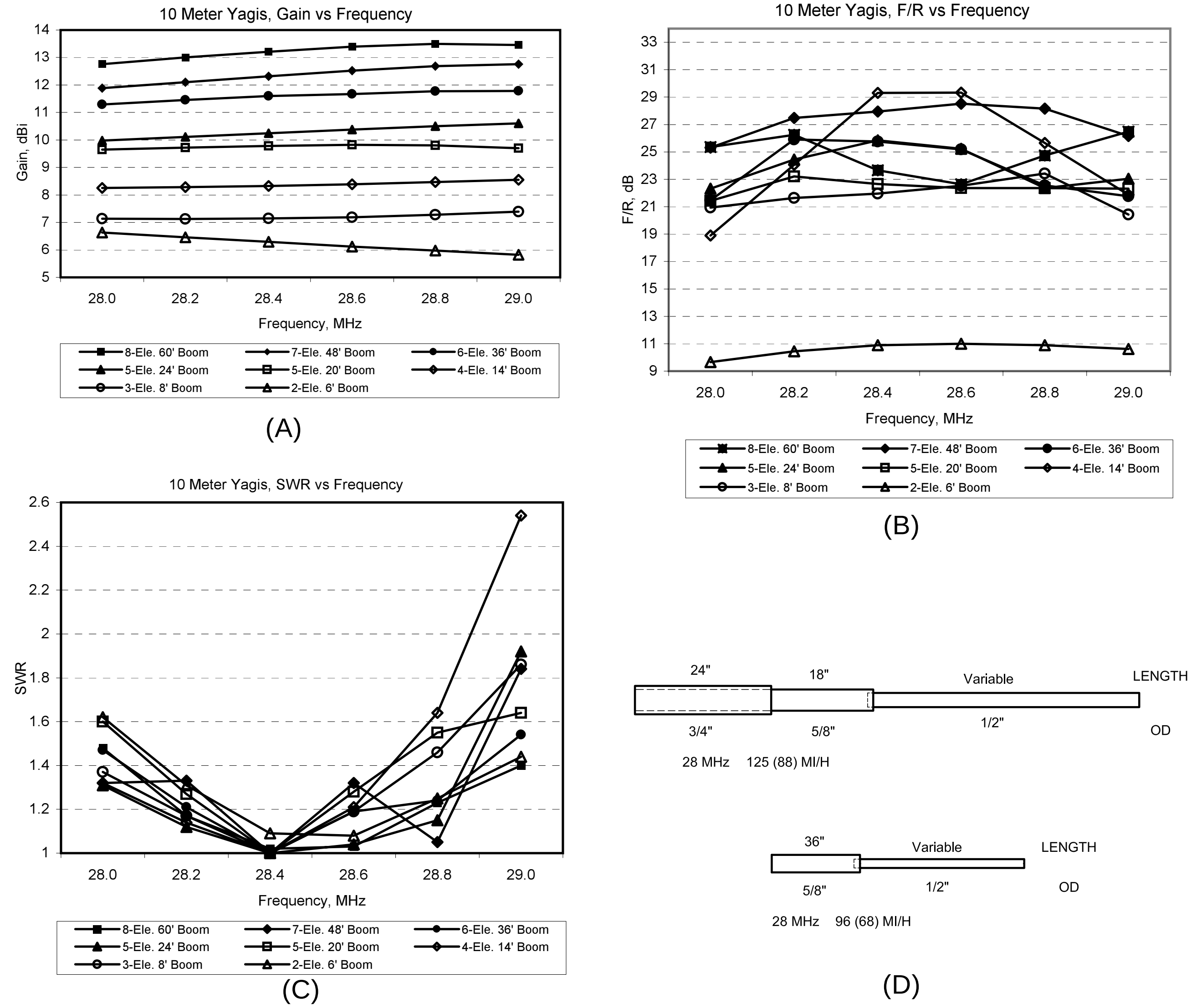

(A)

(B)

(C)

(D)

Fig 13—Gain, F/R and SWR performance versus frequency for optimized 10-meter Yagis. At A, gain is shown versus frequency for eight 10-meter Yagis whose booms range from 6 feet to 60 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 28.0 to 28.8 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR is shown over the frequency range. At D, the taper schedule is shown for heavy-duty and for medium-duty 10-meter elements. The heavy-duty elements can withstand 125-mph winds without icing, and 88-mph winds with $^1/_4$-inch radial ice. The medium-duty elements can survive 96-mph winds without icing, and 68-mph winds with $^1/_4$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

Table 1

Optimized 10-Meter Yagi Designs

Two-element 10-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		210-06H.YW	210-06M.YW
Reflector	0.000"	66.000"	71.500"
Driven Element	66.000"	57.625"	63.000"

Three-element 10-meter Yagi, 8 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		310-08H.YW	310-08M.YW
Reflector	0.000"	66.750"	71.875"
Driven Element	36.000"	57.625"	62.875"
Director 1	54.000"	53.125"	58.500"
Compensator	12" behind Dir. 1	19.000"	18.125"

Four-element 10-meter Yagi, 14 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		410-14H.YW	410-14M.YW
Reflector	0.000"	66.000"	72.000"
Driven Element	36.000"	58.625"	63.875"
Director 1	36.000"	57.000"	62.250"
Director 2	90.000"	47.750"	53.125"
Compensator	12" behind Dir. 2	22.000"	20.500"

Five-element 10-meter Yagi, 24 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		510-24H.YW	510-24M.YW
Reflector	0.000"	65.625"	70.750"
Driven Element	36.000"	58.000"	63.250"
Director 1	36.000"	57.125"	62.375"
Director 2	99.000"	55.000"	60.250"
Director 3	111.000"	50.750"	56.125"
Compensator	12" behind Dir. 3	28.750"	26.750"

Six-element 10-meter Yagi, 36 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		610-36H.YW	610-36M.YW
Reflector	0.000"	66.500"	71.500"
Driven Element	37.000"	58.500"	64.000"
Director 1	43.000"	57.125"	62.375"
Director 2	98.000"	54.875"	60.125"
Director 3	127.000"	53.875"	59.250"
Director 4	121.000"	49.875"	55.250"
Compensator	12" behind Dir. 4	32.000"	29.750"

Seven-element 10-meter Yagi, 48 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		710-48H.YW	710-48M.YW
Reflector	0.000"	65.375"	70.500"
Driven Element	37.000"	59.000"	64.250"
Director 1	37.000"	57.500"	62.750"
Director 2	96.000"	54.875"	60.125"
Director 3	130.000"	52.250"	57.625"
Director 4	154.000"	52.625"	58.000"
Director 5	116.000"	49.875"	55.250"
Compensator	12" behind Dir. 5	35.750"	33.750"

Eight-element 10-meter Yagi, 60 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		810-60H.YW	810-60M.YW
Reflector	0.000"	65.000"	70.125"
Driven Element	42.000"	58.000"	63.500"
Director 1	37.000"	57.125"	62.375"
Director 2	87.000"	55.375"	60.625"
Director 3	126.000"	53.250"	58.625"
Director 4	141.000"	51.875"	57.250"
Director 5	157.000"	52.500"	57.875"
Director 6	121.000"	50.125"	55.500"
Compensator	12" behind Dir. 6	59.375"	55.125"

These 10-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over frequency range from 28.000 to 28.800 MHz, for heavy-duty elements (125 mph wind survival) and for medium-duty (96 mph wind survival). For coverage from 28.8 to 29.7 MHz, subtract 2.000 inches from end of each element, but leave element spacings the same as shown here. Only element tip dimensions are shown, and all dimensions are inches. See Fig 13D for element telescoping tubing schedule. Torque compensator element is made of 2.5" OD PVC water pipe placed 12 inches behind last director. Dimensions shown for compensators is one-half of total length, centered on boom.

of U-bolts with saddles is used to secure the mounting plate to the boom.

Electrically each mounting plate is equivalent to a cylinder, with an effective diameter of 2.405 inches for the heavy-duty element, and 2.310 inches for the medium-duty element. The equivalent length on each side of the boom is 2 inches. These dimensions are incorporated in the files for the *YW* (Yagi for Windows) computer modeling program on the CD-ROM accompanying this book to simulate the effect of the mounting plate.

The second column in **Table 1** shows the spacing of each element relative to the next element in line on the boom, starting at the reflector, which itself is defined as being at the 0.000-inch reference point on the boom. The boom for antennas less than 30 feet long can be constructed of 2-inch OD tubing with 0.065-inch wall thickness. Designs larger than 30 feet long should use 3-inch OD heavy-wall tubing for the boom. Because each boom has extra space at each end, the reflector is actually placed 3 inches from the end of the boom. For example, in the 310-08H.YW design (3 elements on an 8-foot boom), the driven element is placed 36 inches ahead of the reflector, and the director is placed 54 inches ahead of the driven element.

The next columns give the lengths for the variable tips for the heavy-duty and then the medium-duty elements. In the example above for the 310-08H.YW Yagi, the heavy-duty reflector tip, made out of $^{1}/_{2}$-inch OD tubing, sticks out 66.750 inches from the $^{5}/_{8}$-inch OD tubing. Note that each telescoping piece of tubing overlaps 3 inches inside the piece into which it fits, so the overall length of $^{1}/_{8}$-inch OD tubing is 69.750 inches long for the reflector. The medium-duty reflector tip has 71.875 inches protruding from the $^{5}/_{8}$-inch OD tube, and is 74.875 inches long overall. As previously stated, the dimensions are not extremely critical, although measurement accu-

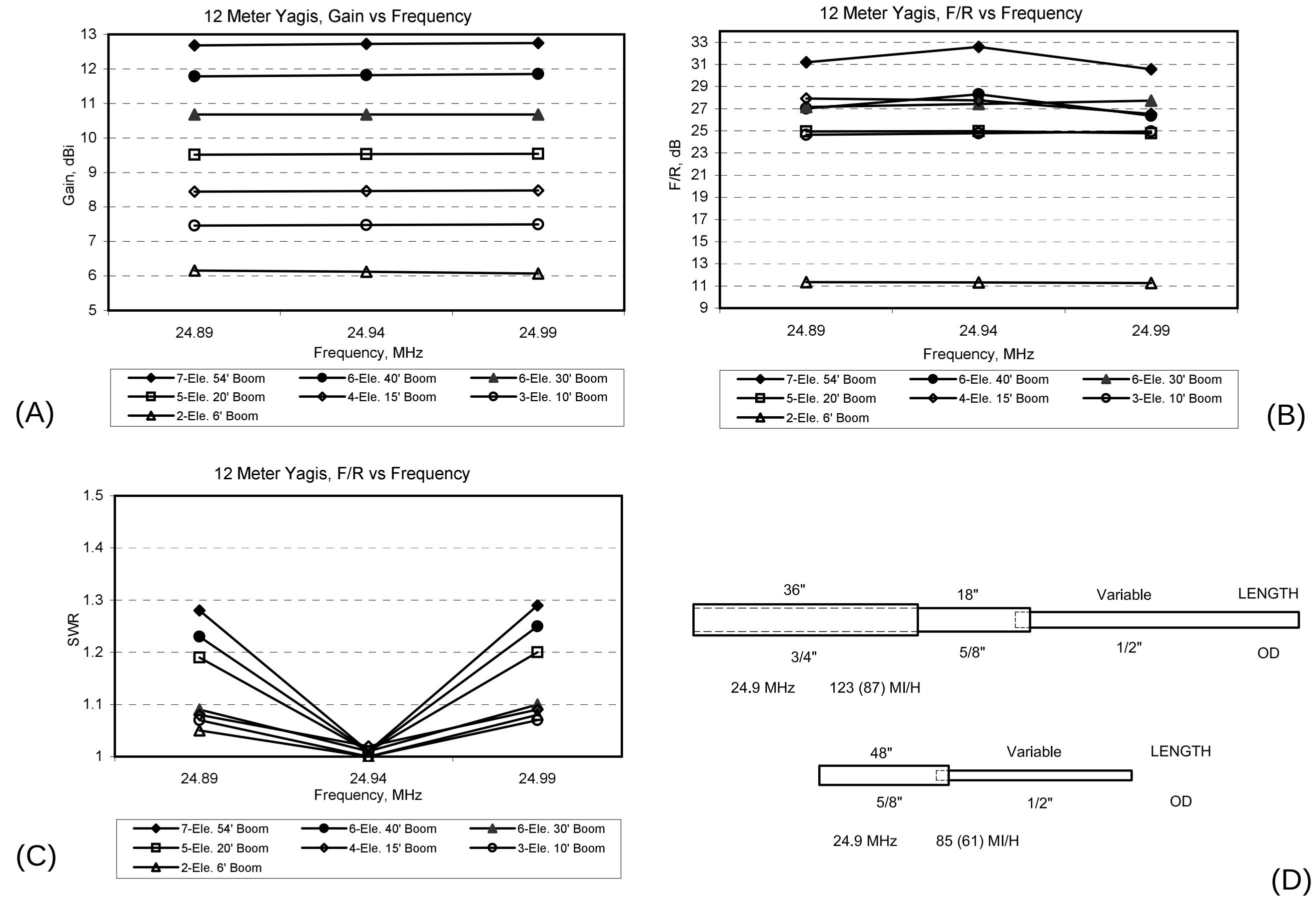

Fig 14—Gain, F/R and SWR performance versus frequency for optimized 12-meter Yagis. At A, gain is shown versus frequency for seven 12-meter Yagis whose booms range from 6 feet to 54 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the narrow 12-meter band 24.89 to 24.99 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 12-meter elements is shown. The heavy-duty elements can withstand 123-mph winds without icing, and 87-mph winds with $^{1}/_{4}$-inch radial ice. The medium-duty elements can survive 85-mph winds without icing, and 61-mph winds with $^{1}/_{4}$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

Table 2
Optimized 12-Meter Yagi Designs

Two-element 12-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		212-06H.YW	212-06M.YW
Reflector	0.000"	67.500"	72.500"
Driven Element	66.000"	59.500"	65.000"

Three-element 12-meter Yagi, 10 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		312-10H.YW	312-10M.YW
Reflector	0.000"	69.000"	73.875"
Driven Element	40.000"	60.250"	65.250"
Director 1	74.000"	54.000"	59.125"
Compensator	12" behind Dir. 1	13.625"	12.000"

Four-element 12-meter Yagi, 15 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		412-15H.YW	412-15M.YW
Reflector	0.000"	66.875"	71.875"
Driven Element	46.000"	61.000"	66.000"
Director 1	46.000"	58.625"	63.750"
Director 2	82.000"	50.875"	56.125"
Compensator	12" behind Dir. 2	16.375"	14.500"

Five-element 12-meter Yagi, 20 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		512-20H.YW	512-20M.YW
Reflector	0.000"	69.750"	74.625"
Driven Element	46.000"	62.250"	67.000"
Director 1	46.000"	60.500"	65.500"
Director 2	48.000"	55.500"	60.625"
Director 3	94.000"	54.625"	59.750"
Compensator	12" behind Dir. 3	22.125"	19.625"

Six-element 12-meter Yagi, 30 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		612-30H.YW	612-30M.YW
Reflector	0.000"	68.125"	73.000"
Driven Element	46.000"	61.750"	66.750"
Director 1	46.000"	60.250"	65.250"
Director 2	73.000"	52.375"	57.625"
Director 3	75.000"	57.625"	62.750"
Director 4	114.000"	53.625"	58.750"
Compensator	12" behind Dir. 4	30.000"	26.250"

Six-element 12-meter Yagi, 40 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		612-40H.YW	612-40M.YW
Reflector	0.000"	67.000"	71.875"
Driven Element	46.000"	60.125"	65.500"
Director 1	46.000"	57.375"	62.500"
Director 2	91.000"	57.375"	62.500"
Director 3	157.000"	57.000"	62.125"
Director 4	134.000"	54.375"	59.500"
Compensator	12" behind Dir. 4	36.500"	31.625"

Seven-element 12-meter Yagi, 54 foot boom

Element	*Spacing, inches*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		712-54H.YW	712-54M.YW
Reflector	0.000"	68.000"	73.000"
Driven Element	46.000"	60.500"	65.500"
Director 1	46.000"	56.750"	61.875"
Director 2	75.000"	58.000"	63.125"
Director 3	161.000"	55.625"	60.750"
Director 4	174.000"	56.000"	61.125"
Director 5	140.000"	53.125"	58.375"
Compensator	12" behind Dir. 5	43.125"	37.500"

These 12-meter Yagi designs were optimized for > 20 dB F/R, and SWR < 2:1 over frequency range from 24.890 to 24.990 MHz, for heavy-duty elements (123 mph wind survival) and for medium-duty (85 mph wind survival). Only element tip dimensions are shown, and all dimensions are inches. See Fig 14D for element telescoping tubing schedule. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director. Dimensions shown for compensators is one-half of total length, centered on boom.

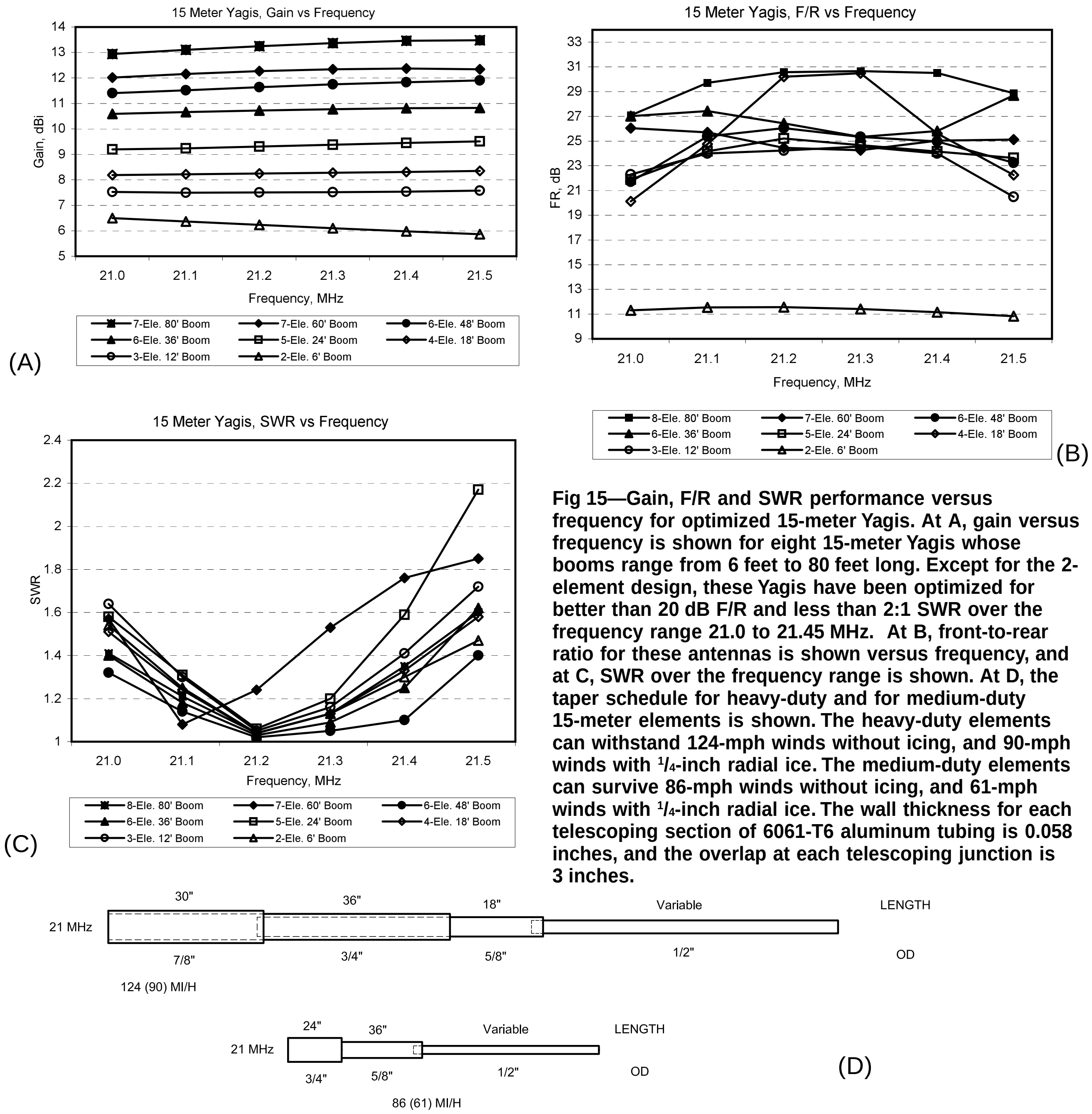

Fig 15—Gain, F/R and SWR performance versus frequency for optimized 15-meter Yagis. At A, gain versus frequency is shown for eight 15-meter Yagis whose booms range from 6 feet to 80 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 21.0 to 21.45 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 15-meter elements is shown. The heavy-duty elements can withstand 124-mph winds without icing, and 90-mph winds with 1/4-inch radial ice. The medium-duty elements can survive 86-mph winds without icing, and 61-mph winds with 1/4-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

racy to 1/8 inch is desirable.

The last row in each variable tip column shows the length of one-half of the “dummy element” torque compensator used to correct for uneven wind loading along the boom. This compensator is made from 2.5 inches OD PVC water pipe mounted to an element-to-boom plate like those used for each element. The compensator is mounted 12 inches behind the last director, the first director in the case of the 3-element 310-08H.YW antenna. Note that the heavy-duty elements require a correspondingly longer torque compensator than do the medium-duty elements.

12-METER YAGIS

Fig 14 describes the electrical performance of seven optimized 12-meter Yagis with boom lengths between 6 to 54 feet. The end of each boom includes 3 inches of space for the reflector and last director (or driven element) mounting plates. The narrow frequency width of the 12-meter band allows the performance to be optimized easily. Fig 14A shows the free-space gain versus frequency for each antenna; 14B shows the front-to-rear ratio, and 14C shows the SWR versus frequency. Each antenna with three or more elements was designed to

Table 3
Optimized 15-Meter Yagi Designs

Two-element 15-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		215-06H.YW	215-06M.YW
Reflector	0.000"	62.000"	85.000"
Driven Element	66.000"	51.000"	74.000"

Three-element 15-meter Yagi, 12 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		315-12H.YW	315-12M.YW
Reflector	0.000"	62.000"	84.250"
Driven Element	48.000"	51.000"	73.750"
Director 1	92.000"	43.500"	66.750"
Compensator	12" behind Dir. 1	34.750"	37.625"

Four-element 15-meter Yagi, 18 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		415-18H.YW	415-18M.YW
Reflector	0.000"	61.000"	83.500"
Driven Element	56.000"	51.500"	74.500"
Director 1	56.000"	48.000"	71.125"
Director 2	98.000"	36.625"	60.250"
Compensator	12" behind Dir. 2	20.875"	18.625"

Five-element 15-meter Yagi, 24 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		515-24H.YW	515-24M.YW
Reflector	0.000"	62.000"	84.375"
Driven Element	48.000"	52.375"	75.250"
Director 1	48.000"	47.875"	71.000"
Director 2	52.000"	47.000"	70.125"
Director 3	134.000"	41.000"	64.375"
Compensator	12" behind Dir. 3	40.250"	35.125"

Six-element 15-meter Yagi, 36 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		615-36H.YW	615-36M.YW
Reflector	0.000"	61.000"	83.375"
Driven Element	53.000"	52.000"	75.000"
Director 1	56.000"	49.125"	72.125"
Director 2	59.000"	45.125"	68.375"
Director 3	116.000"	47.875"	71.000"
Director 4	142.000"	42.000"	65.375"
Compensator	12" behind Dir. 4	45.500"	39.750"

Seven-element 15-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		615-48H.YW	615-48M.YW
Reflector	0.000"	62.000"	84.000"
Driven Element	48.000"	52.000"	75.000"
Director 1	48.000"	51.250"	74.125"
Director 2	125.000"	48.000"	71.125"
Director 3	190.000"	45.500"	68.750"
Director 4	161.000"	42.000"	65.375"
Compensator	12" behind Dir. 4	51.500"	45.375"

Seven-element 15-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		715-60H.YW	715-60M.YW
Reflector	0.000"	59.750"	82.250"
Driven Element	48.000"	52.000"	75.000"
Director 1	48.000"	52.000"	74.875"
Director 2	93.000"	49.500"	72.500"
Director 3	173.000"	44.125"	67.375"
Director 4	197.000"	45.500"	68.750"
Director 5	155.000"	41.750"	65.125"
Compensator	12" behind Dir. 5	58.500"	51.000"

Eight-element 15-meter Yagi, 80 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		815-80H.YW	815-80M.YW
Reflector	0.000"	62.000"	84.000"
Driven Element	56.000"	52.500"	75.500"
Director 1	48.000"	51.500"	74.375"
Director 2	115.000"	48.375"	71.500"
Director 3	164.000"	45.750"	69.000"
Director 4	202.000"	43.125"	66.500"
Director 5	206.000"	44.750"	68.000"
Director 6	163.000"	40.875"	64.250"
Compensator	12" behind Dir. 6	95.000"	83.375"

These 15-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from 21.000 to 21.450 MHz, for heavy-duty elements (124 mph wind survival) and for medium-duty (86 mph wind survival). Only element tip dimensions are shown. See Fig 15D for element telescoping tubing schedule. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

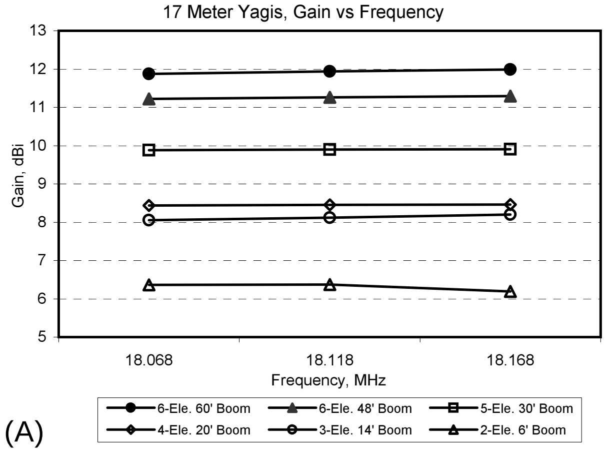

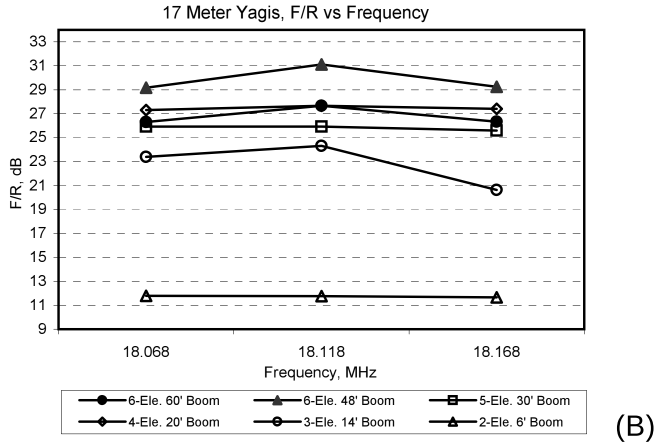

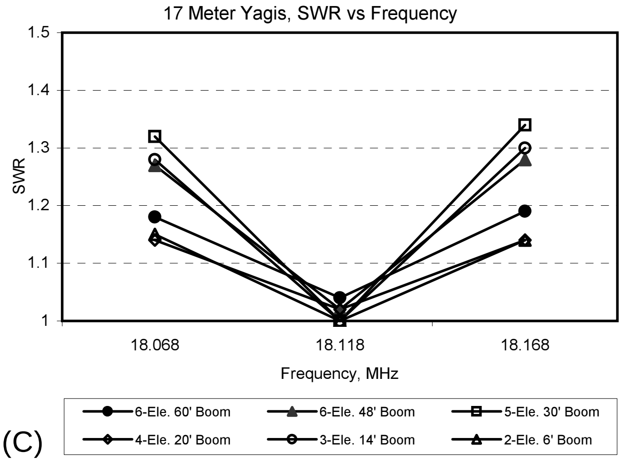

Fig 16—Gain, F/R and SWR performance versus frequency for optimized 17-meter Yagis. At A, gain versus frequency is shown for six 17-meter Yagis whose booms range from 6 feet to 60 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the narrow 17-meter band 18.068 to 18.168 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 10-meter elements is shown. The heavy-duty elements can withstand 123-mph winds without icing, and 89-mph winds with 1/4-inch radial ice. The medium-duty elements can survive 83-mph winds without icing, and 59-mph winds with 1/4-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

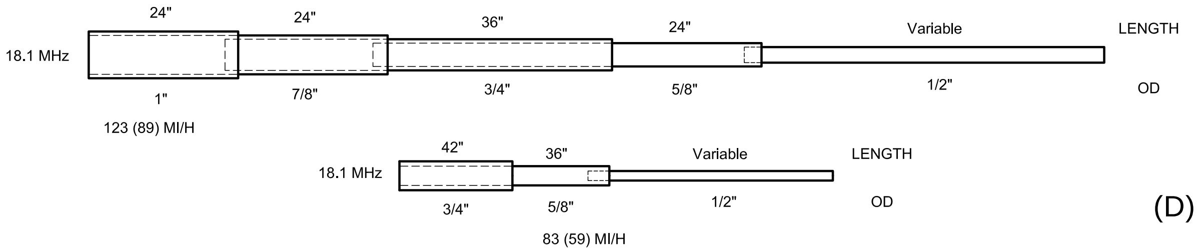

cover the narrow 12-meter band from 24.89 to 24.99 MHz, with SWR less than 2:1 and F/R better than 20 dB over that range.

Fig 14D shows the taper schedule for two types of 12-meter elements. The heavy-duty design can survive 123-mph winds with no icing, and 87-mph winds with 1/4 inch of radial ice. The medium-duty design can handle 85-mph winds with no icing, and 61-mph winds with 1/4 inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375 inch thick flat aluminum plate, 5 inches wide by 6 inches long.

Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 2.945 inches for the heavy-duty element, and 2.857 inches for the medium-duty element. The equivalent length on each side of the boom is 3 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

15-METER YAGIS

Fig 15 describes the electrical performance of eight optimized 15-meter Yagis with boom lengths between 6 feet to a spectacular 80 feet. The end of each boom includes 3 inches of space for the reflector and last-director (or driven element) mounting plates. Fig 15A shows the free-space gain versus frequency for each antenna; 15B shows the worst-case front-to-rear ratio, and 15C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the full 15-meter band from

Table 4

Optimized 17-meter Yagi Designs

Two-element 17-meter Yagi, 6 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		217-06H.YW	217-06M.YW
Reflector	0.000"	61.000"	89.000"
Driven Element	66.000"	48.000"	76.250"

Three-element 17-meter Yagi, 14 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		317-14H.YW	317-14M.YW
Reflector	0.000"	61.500"	91.500"
Driven Element	65.000"	52.000"	79.500"
Director 1	97.000"	46.000"	73.000"
	12" behind Dir. 1	12.625"	10.750"

Four-element 17-meter Yagi, 20 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		417-20H.YW	417-20M.YW
Reflector	0.000"	61.500"	89.500"
Driven Element	48.000"	54.250"	82.625"
Director 1	48.000"	52.625"	81.125"
Director 2	138.000"	40.500"	69.625"
Compensator	12" behind Dir. 2	42.500"	36.250"

Five-element 17-meter Yagi, 30 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		517-30H.YW	517-30M.YW
Reflector	0.000"	61.875"	89.875"
Driven Element	48.000"	52.250"	80.500"
Director 1	52.000"	49.625"	78.250"
Director 2	93.000"	49.875"	78.500"
Director 3	161.000"	43.500"	72.500"
Compensator	12" behind Dir. 3	54.375"	45.875"

Six-element 17-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		617-48H.YW	617-48M.YW
Reflector	0.000"	63.000"	90.250"
Driven Element	52.000"	52.500"	80.500"
Director 1	51.000"	45.500"	74.375"
Director 2	87.000"	47.875"	76.625"
Director 3	204.000"	47.000"	75.875"
Director 4	176.000"	42.000"	71.125"
Compensator	12" behind Dir. 4	68.250"	57.500"

Six-element 17-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		617-60H.YW	617-60M.YW
Reflector	0.000"	61.250"	89.250"
Driven Element	54.000"	54.750"	83.125"
Director 1	54.000"	52.250"	80.750"
Director 2	180.000"	46.000"	74.875"
Director 3	235.000"	44.625"	73.625"
Director 4	191.000"	41.500"	70.625"
Compensator	12" behind Dir. 4	62.875"	53.000"

These 17-m Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from18.068 to 18.168 MHz, for heavy-duty elements (123 mph wind survival) and for medium-duty (83 mph wind survival). Only element tip dimensions are shown. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

21.000 to 21.450 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 15D shows the taper schedule for two types of 15-meter elements. The heavy-duty design can survive 124-mph winds with no icing, and 90-mph winds with ¼ inch of radial ice. The medium-duty design can handle 86-mph winds with no icing, and 61-mph winds with ¼ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 5 inches wide by 6 inches long.

Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.0362 inches for

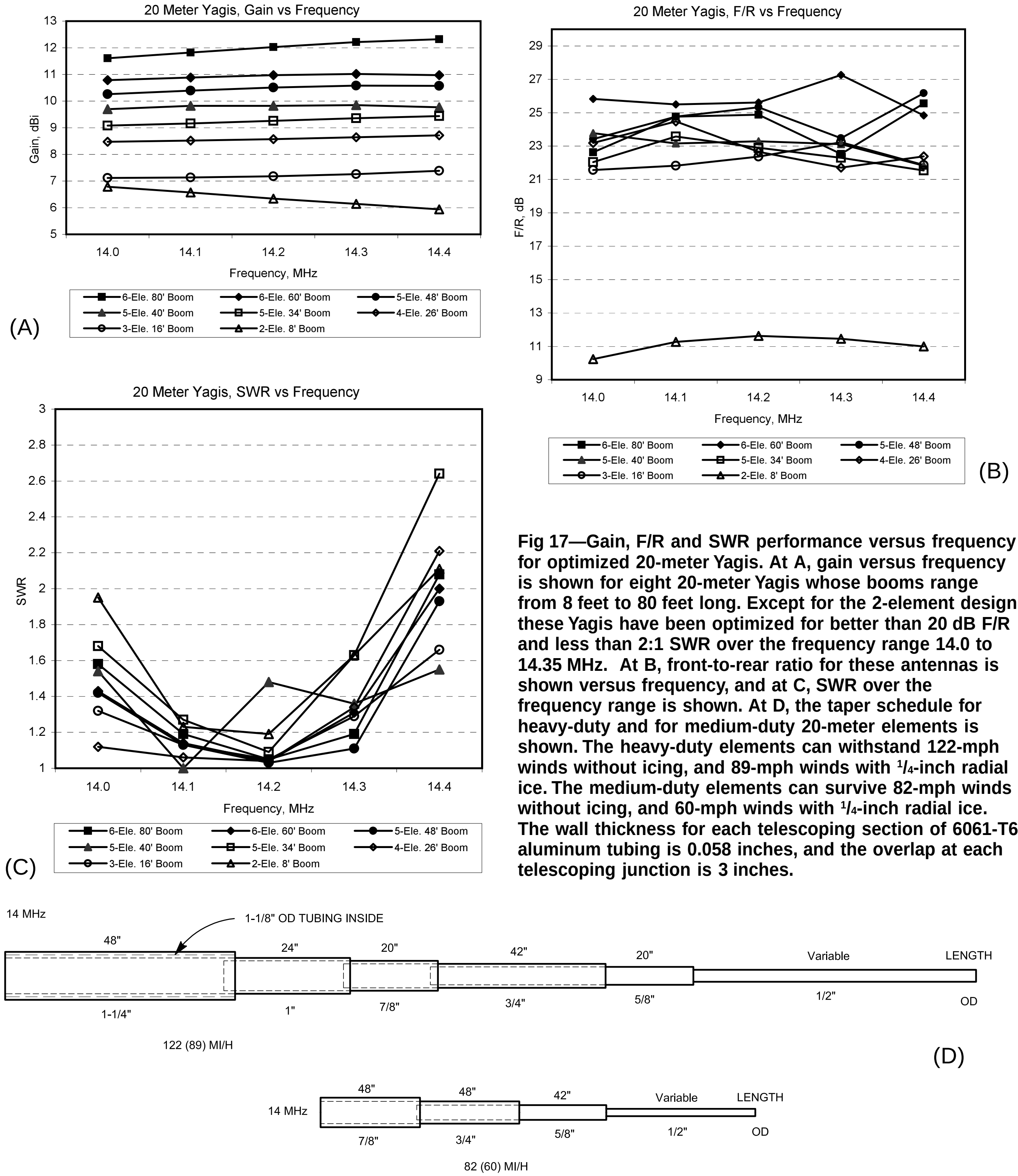

Fig 17—Gain, F/R and SWR performance versus frequency for optimized 20-meter Yagis. At A, gain versus frequency is shown for eight 20-meter Yagis whose booms range from 8 feet to 80 feet long. Except for the 2-element design, these Yagis have been optimized for better than 20 dB F/R and less than 2:1 SWR over the frequency range 14.0 to 14.35 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule for heavy-duty and for medium-duty 20-meter elements is shown. The heavy-duty elements can withstand 122-mph winds without icing, and 89-mph winds with $^1/_4$-inch radial ice. The medium-duty elements can survive 82-mph winds without icing, and 60-mph winds with $^1/_4$-inch radial ice. The wall thickness for each telescoping section of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at each telescoping junction is 3 inches.

the heavy-duty element, and 2.9447 inches for the medium-duty element. The equivalent length on each side of the boom is 3 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

17-METER YAGIS

Fig 16 describes the electrical performance of six optimized 17-meter Yagis with boom lengths between 6 to a heroic 60 feet. As usual, the end of each boom includes 3 inches of space for the reflector and last director (or driven element) mounting plates. Fig 16A shows the free-space gain versus frequency for each an-

Table 5

Optimized 20-Meter Yagi Designs

Two-element 20-meter Yagi, 8 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		220-08H.YW	220-08M.YW
Reflector	0.000"	66.000"	80.000"
Driven Element	90.000"	46.000"	59.000"

Three-element 20-meter Yagi, 16 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		320-16H.YW	320-16M.YW
Reflector	0.000"	69.625"	81.625"
Driven Element	80.000"	51.250"	64.500"
Director 1	106.000"	42.625"	56.375"
Compensator	12" behind Dir. 1	33.375"	38.250"

Four-element 20-meter Yagi, 26 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		420-26H.YW	420-26M.YW
Reflector	0.000"	65.625"	78.000"
Driven Element	72.000"	53.375"	65.375"
Director 1	60.000"	51.750"	63.875"
Director 2	174.000"	38.625"	51.500"
Compensator	12" behind Dir. 2	54.250"	44.250"

Five-element 20-meter Yagi, 34 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-34H.YW	520-34M.YW
Reflector	0.000"	68.625"	80.750"
Driven Element	72.000"	52.250"	65.500"
Director 1	71.000"	45.875"	59.375"
Director 2	68.000"	45.875"	59.375"
Director 3	191.000"	37.000"	51.000"
Compensator	12" behind Dir. 3	69.250"	56.250"

Five-element 20-meter Yagi, 40 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-40H.YW	520-40M.YW
Reflector	0.000"	68.375"	80.500"
Driven Element	72.000"	53.500"	66.625"
Director 1	72.000"	51.500"	64.625"
Director 2	139.000"	48.375"	61.750"
Director 3	191.000"	38.000"	52.000"
Compensator	12" behind Dir. 3	69.750"	56.750"

Five-element 20-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		520-48H.YW	520-48M.YW
Reflector	0.000"	66.250"	78.500"
Driven Element	72.000"	53.000"	66.000"
Director 1	88.000"	50.500"	63.750"
Director 2	199.000"	47.375"	60.875"
Director 3	211.000"	39.750"	53.625"
Compensator	12" behind Dir. 3	70.325"	57.325"

Six-element 20-meter Yagi, 60 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		620-60H.YW	620-60M.YW
Reflector	0.000"	67.000"	79.250"
Driven Element	84.000"	51.500"	65.000"
Director 1	91.000"	45.125"	58.750"
Director 2	130.000"	41.375"	55.125"
Director 3	210.000"	46.875"	60.375"
Director 4	199.000"	39.125"	53.000"
Compensator	12" behind Dir. 4	72.875"	59.250"

Six-element 20-meter Yagi, 80 foot boom

Element	*Spacing*	*Heavy-Duty Tip*	*Medium-Duty Tip*
File Name		620-80H.YW	620-80M.YW
Reflector	0.000"	66.125"	78.375"
Driven Element	72.000"	52.375"	65.500"
Director 1	122.000"	49.125"	62.500"
Director 2	229.000"	44.500"	58.125"
Director 3	291.000"	42.625"	56.375"
Director 4	240.000"	38.750"	52.625"
Compensator	12" behind Dir. 4	78.750"	64.125"

These 20-meter Yagi designs are optimized for > 20 dB F/R, and SWR < 2:1 over entire frequency range from 14.000 to 14.350 MHz, for heavy-duty elements (122 mph wind survival) and for medium-duty (82 mph wind survival). Only element tip dimensions are shown. See Fig 17 for element telescoping tubing schedule. All dimensions are in inches. Torque compensator element is made of 2.5" OD PVC water pipe placed 12" behind last director, and dimensions shown for compensators is one-half of total length, centered on boom.

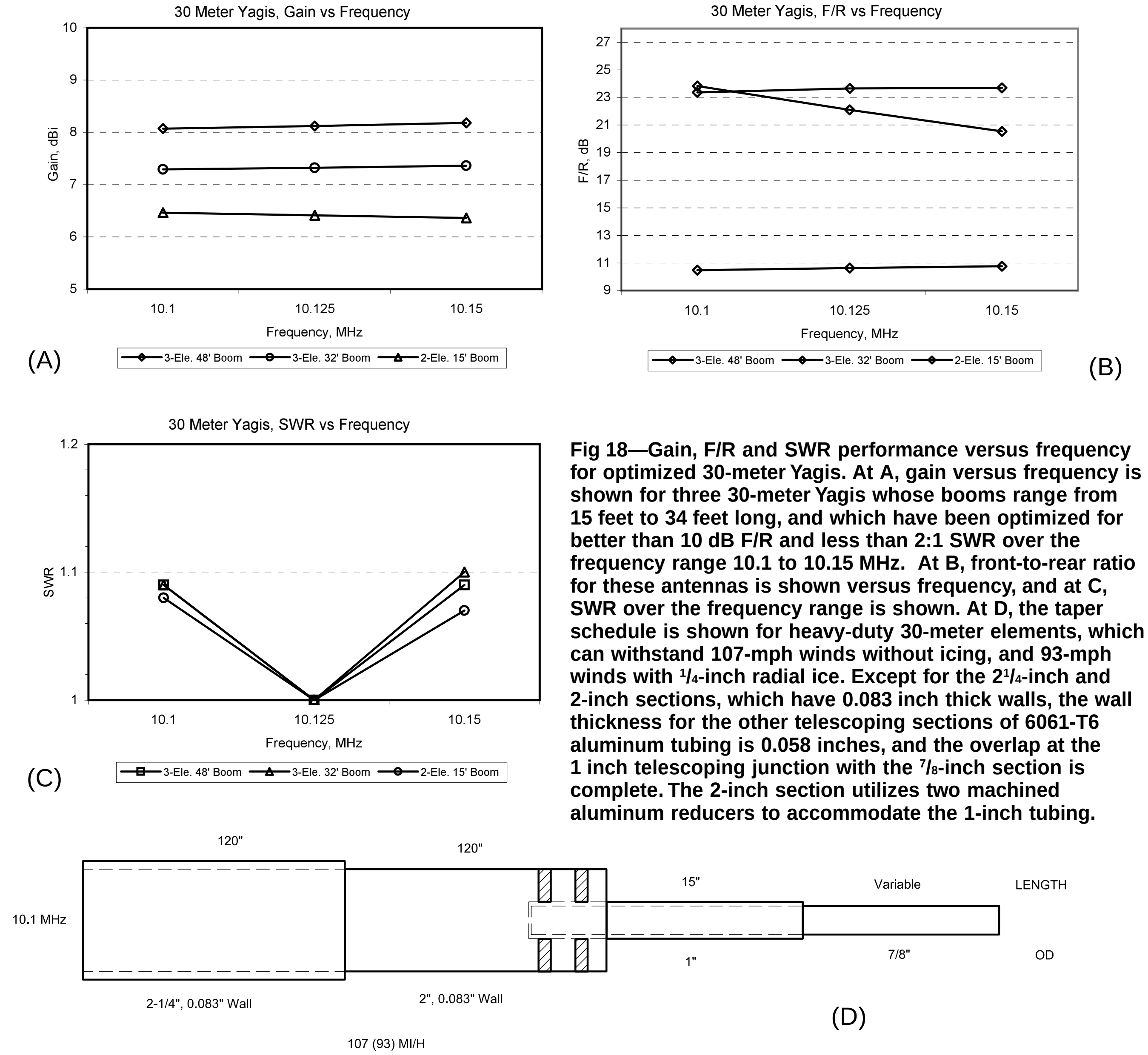

Fig 18—Gain, F/R and SWR performance versus frequency for optimized 30-meter Yagis. At A, gain versus frequency is shown for three 30-meter Yagis whose booms range from 15 feet to 34 feet long, and which have been optimized for better than 10 dB F/R and less than 2:1 SWR over the frequency range 10.1 to 10.15 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule is shown for heavy-duty 30-meter elements, which can withstand 107-mph winds without icing, and 93-mph winds with $^1/_4$-inch radial ice. Except for the $2^1/_4$-inch and 2-inch sections, which have 0.083 inch thick walls, the wall thickness for the other telescoping sections of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at the 1 inch telescoping junction with the $^7/_8$-inch section is complete. The 2-inch section utilizes two machined aluminum reducers to accommodate the 1-inch tubing.

tenna; 16B shows the worst-case front-to-rear ratio, and 16C shows the SWR versus frequency. Each antenna with three or more elements was designed to cover the narrow 17-meter band from 18.068 to 18.168 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 16D shows the taper schedule for two types of 17-meter elements. The heavy-duty design can survive 123-mph winds with no icing, and 83-mph winds with $^1/_4$-inch of radial ice. The medium-duty design can handle 83-mph winds with no icing, and 59-mph winds with ¼ inch of radial ice.

The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 6 inches wide by 8 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.5122 inches for the heavy-duty element, and 3.3299 inches for the medium-duty element. The equivalent length on each side of the boom is 4 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

20-METER YAGIS

Fig 17 describes the electrical performance of eight optimized 20-meter Yagis with boom lengths between 8 to a giant 80 feet. As usual, the end of each boom includes 3 inches of space for the reflector and last director (driven element) mounting plates. Fig 17A shows the free-space gain versus frequency for each antenna; 17B shows the front-to-rear ratio, and 17C shows the SWR versus frequency. Each antenna with three or more ele-

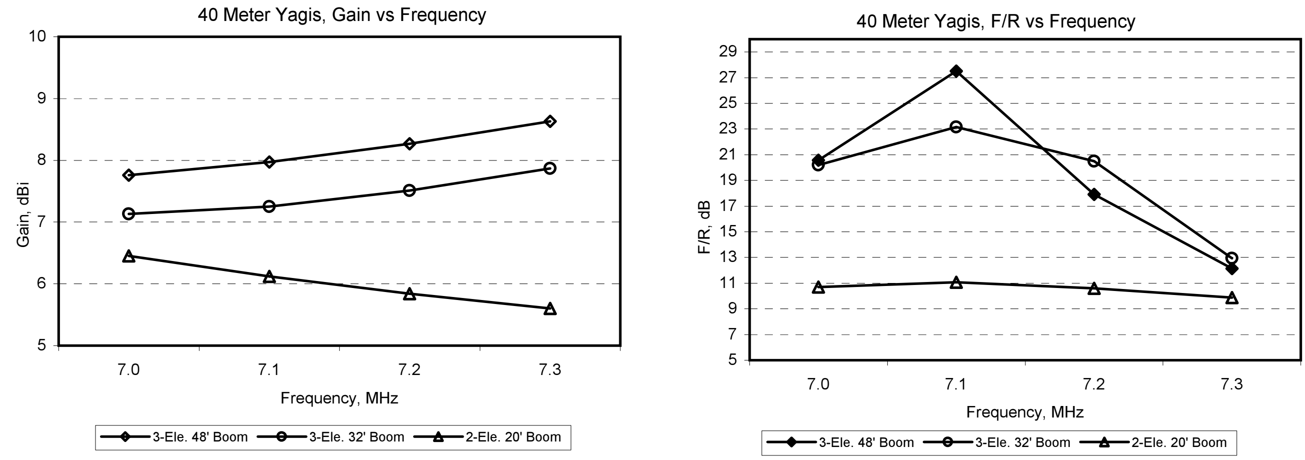

(A) (B)

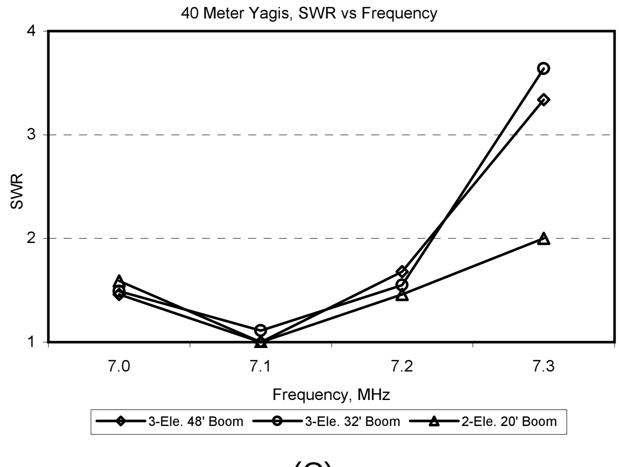

(C)

Fig 19—Gain, F/R and SWR performance versus frequency for optimized 40-meter Yagis. At A, gain versus frequency is shown for three 40-meter Yagis whose booms range from 20 feet to 48 feet long, and which have been optimized for better than 10 dB F/R and less than 2:1 SWR over the frequency range 7.0 to 7.2 MHz. At B, front-to-rear ratio for these antennas is shown versus frequency, and at C, SWR over the frequency range is shown. At D, the taper schedule is shown for heavy-duty 40-meter elements, which can withstand 107-mph winds without icing, and 93-mph winds with 1/4-inch radial ice. Except for the 2 1/4-inch and 2-inch sections, which have 0.083 inch thick walls, the wall thickness for the other telescoping sections of 6061-T6 aluminum tubing is 0.058 inches, and the overlap at the end telescoping junction is 3 inches. The 2-inch section utilizes two machined aluminum reducers to accommodate the 1-inch tubing.

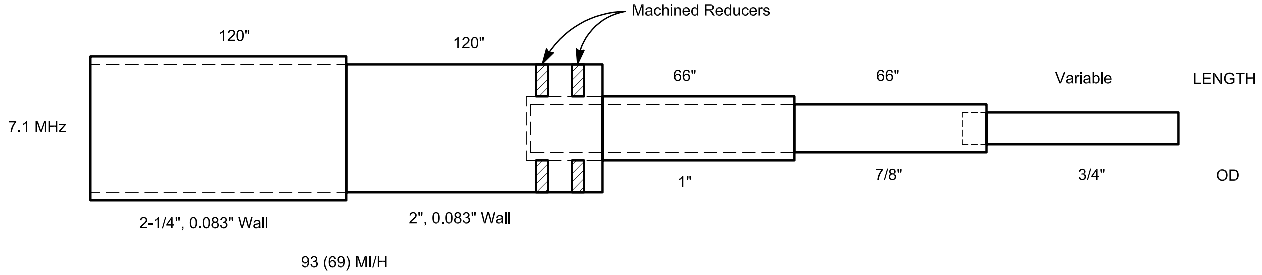

(D)

ments was designed to cover the complete 20-meter band from 14.000 to 14.350 MHz, with SWR less than 2:1 and F/R ratio better than 20 dB over that range.

Fig 17D shows the taper schedule for two types of 20-meter elements. The heavy-duty design can survive 122-mph winds with no icing, and 89-mph winds with 1/4 inch of radial ice. The medium-duty design can handle 82-mph winds with no icing, and 60-mph winds with 1/4 inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.375-inch thick flat aluminum plate, 6 inches wide by 8 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 3.7063 inches for the heavy-duty element, and 3.4194 inches for the medium-duty element. The equivalent length on each side of the boom is 4 inches. As usual, the torque compensator is mounted 12 inches behind the last director.

30-METER YAGIS

Fig 18 describes the electrical performance of three

Table 6
Optimized 30-Meter Yagi Designs

Two-element 30-meter Yagi, 15 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		230-15H.YW
Reflector	0.000"	50.250"
Driven Element	174.000"	14.875"

3-element 30-meter Yagi, 22 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		330-22H.YW
Reflector	0.000	59.375
Driven Element	135.000	35.000
Director 1	123.000	19.625

Three-element 30-meter Yagi, 34 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		330-34H.YW
Reflector	0.000"	53.750"
Driven Element	212"	29.000"
Director 1	190"	14.500"

These 30-m Yagi designs are optimized for > 10 dB F/R, and SWR < 2:1 over entire frequency range from 10.100 to 10.150 MHz for heavy-duty elements (105 mph wind survival). Only element tip dimensions are shown. See Fig 18D for element telescoping tubing schedule. All dimensions are in inches. No torque compensator element is required.

Table 7
Optimized 40-Meter Yagi Designs

Two-element 40-meter Yagi, 20 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		240-20H.YW
Reflector	0.000"	85.000"
Driven Element	234.000"	35.000"

Three-element 40-meter Yagi, 32 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		340-32H.YW
Reflector	0.000"	90.750"
Driven Element	196.000"	55.875"
Director 1	182.000"	33.875"

Three-element 40-meter Yagi, 48 foot boom

Element	*Spacing*	*Heavy-Duty Tip*
File Name		340-48H.YW
Reflector	0.000"	81.000"
Driven Element	300.000"	45.000"
Director 1	270.000"	21.000"

These 40-m Yagi designs are optimized for > 10 dB F/R, and SWR < 2:1 over low-end of frequency range from 7.000 to 7.200 MHz, for heavy-duty elements (95 mph wind survival). Only element tip dimensions are shown. See Fig 19D for element telescoping tubing schedule. All dimensions are in inches. No wind torque compensator is required.

optimized 30-meter Yagis with boom lengths between 15 to 34 feet. Because of the size and weight of the elements alone for Yagis on this band, only 2-element and 3-element designs are described. The front-to-rear ratio requirement for the 2-element antenna is relaxed to be greater than 10 dB over the band from 10.100 to 10.150 MHz, while that for the 3-element designs is kept at greater than 20 dB over that frequency range.

As usual, the end of each boom includes 3 inches of space for the reflector and last director mounting plates. Fig 18A shows the free-space gain versus frequency for each antenna; 18B shows the worst-case front-to-rear ratio, and 18C shows the SWR versus frequency.

Fig 18D shows the taper schedule for the 30-meter elements. Note that the wall thickness of the first two sections of tubing is 0.083 inches, rather than 0.058 inches. This heavy-duty element design can survive 107-mph winds with no icing, and 93-mph winds with $^{1}/_{4}$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.500-inch thick flat aluminum plate, 6 inches wide by 24 inches long. Electrically, each mounting plate is equivalent to a cylinder, with an effective diameter of 4.684 inches. The equivalent length on each side of the boom is 12 inches. These designs require no torque compensator.

40-METER YAGIS

Fig 19 describes the electrical performance of three optimized 40-meter Yagis with boom lengths between 20 to 48 feet. Like the 30-meter antennas, because of the size and weight of the elements for a 40-meter Yagi, only 2-element and 3-element designs are described. The front-to-rear ratio requirement for the 2-element antenna is relaxed to be greater than 10 dB over the band from 7.000 to 7.300 MHz, while the goal for the 3-element designs is 20 dB over the frequency range of 7.000 to 7.200 MHz. It is exceedingly difficult to hold the F/R greater than 20 dB over the entire 40-meter band without sacrificing excessive gain with a 3-element design.

As usual, the end of each boom includes 3 inches of space for the reflector and last director mounting plates. Fig 19A shows the free-space gain versus frequency for each antenna; 19B shows the front-to- rear ratio, and 19C shows the SWR versus frequency.

Fig 19D shows the taper schedule for the 40-meter elements. Note that the wall thickness of the first two sections of tubing is 0.083 inches, rather than 0.058 inches. This element design can survive 93-mph winds with no icing, and 69-mph winds with $^{1}/_{4}$ inch of radial ice. The element-to-boom mounting plate for these Yagis is a 0.500-inch thick flat aluminum plate, 6 inches wide by 24 inches long. Electrically each mounting plate is equivalent to a cylinder, with an effective diameter of 4.684 inches. The equivalent length on each side of the boom is 12 inches. These designs require no torque compensator.

Modifying Monoband Hy-Gain Yagis

Enterprising amateurs have long used the Telex Communications Hy-Gain "Long John" series of HF monobanders as a source of top-quality aluminum and hardware for customized Yagis. Often-modified older models include the 105BA for 10 meters, the 155BA for 15 meters, and the 204BA and 205BA for 20 meters. Newer Hy-Gain designs, the 105CA, 155CA and 205CA, have been redesigned by computer for better performance.

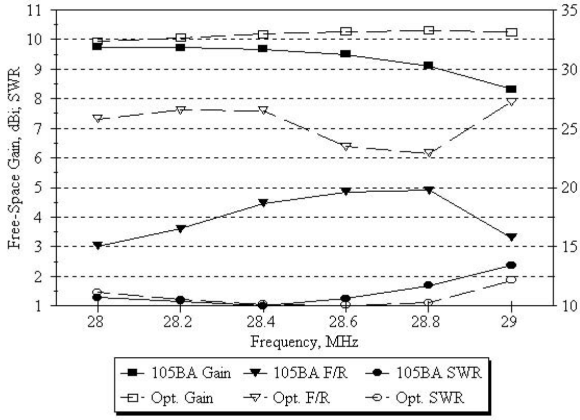

Fig 20—Gain, F/R and SWR over the 28.0 to 28.8 MHz range for original and optimized Yagis using Hy-Gain hardware. Original 105BA design provided excellent weight balance at boom-to-mast bracket, but compromised the electrical performance somewhat because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 23 dB. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 8.

Table 8
Optimized Hy-Gain 20-Meter Yagi Designs

Optimized 204BA, Four-element 20-meter Yagi, 26 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV204CA.YW
Reflector	0.000"	56.000"
Driven Element	85.000"	52.000"
Director 1	72.000"	61.500"
Director 2	149.000"	50.125"

Optimized 205CA, Five-element 20-meter Yagi, 34 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV205CA.YW
Reflector	0.000"	62.625"
Driven Element	72.000"	53.500"
Director 1	72.000"	63.875"
Director 2	74.000"	61.625"
Director 3	190.000"	55.000"

Hy-Gain antennas have historically had an excellent reputation for superior mechanical design, and Hy-Gain proudly points out that many of their monobanders are still working after more than 30 years. In the older designs the elements were purposely spaced along the boom to achieve good weight balance at the mast-to-boom bracket, with electrical performance as a secondary goal. Thus, the electrical performance was not necessarily optimum, particularly over an entire amateur band. Newer Hy-Gain designs are electrically superior to the older ones, but because of their strong concern for weight-balance are still not optimal by the definitions used in this chapter.

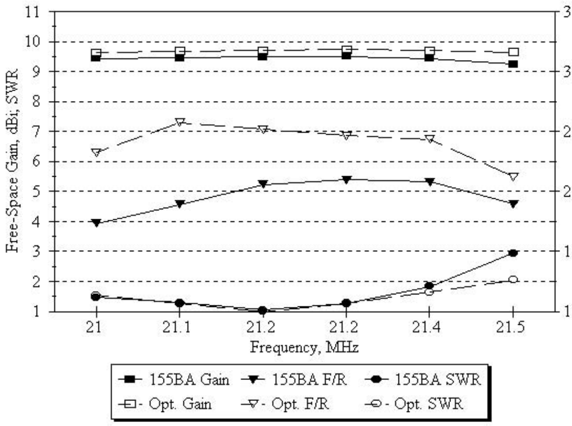

Fig 21—Gain, F/R and SWR over the 21.0 to 21.45 MHz band for original and optimized Yagis using Hy-Gain hardware. Original 155BA design provided excellent weight balance at boom-to-mast bracket, but compromised the electrical performance somewhat because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 22 dB. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 9.

Table 9
Optimized Hy-Gain 15-Meter Yagi Designs

Optimized 155BA, Five-element 15-meter Yagi, 24 foot boom

Element	*Spacing*	*Element Tip*
File Name		BV155CA.YW
Reflector	0.000"	64.000"
Driven Element	48.000"	65.500"
Director 1	48.000"	63.875"
Director 2	82.750"	61.625"
Director 3	127.250"	55.000"

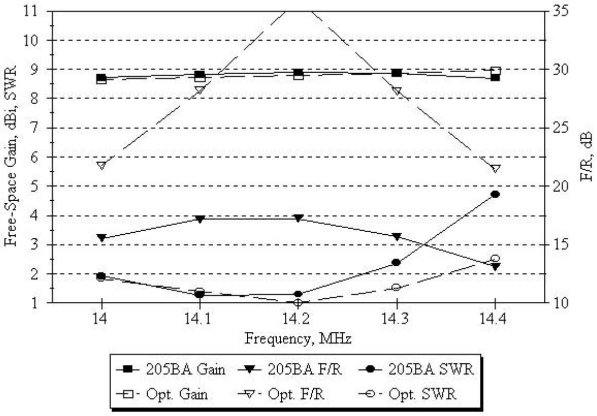

Fig 22—Gain, F/R and SWR over the 14.0 to 14.35 MHz band for original and optimized Yagis using Hy-Gain hardware. Original 205BA design provided good weight balance at boom-to-mast bracket, but compromised the electrical performance because of non-optimum spacing of elements. Optimized design requires wind torque-balancing compensator element, and compensating weight at director end of boom to rebalance weight. The F/R ratio over the frequency range for the optimized design is more than 23 dB, while the original design never went beyond 17 dB of F/R. Each element uses the original Hy-Gain taper schedule and element-to-boom clamp, but the length of the tip is changed per Table 10.

Table 10
Optimized Hy-Gain 10-Meter Yagi Designs

Optimized 105BA, Five-element 10-meter Yagi, 24 foot boom

Element	*Spacing, inches*	*Element Tip*
File Name		BV105CA.YW
Reflector	0.000"	44.250"
Driven Element	40.000"	53.625"
Director 1	40.000"	52.500"
Director 2	89.500"	50.500"
Director 3	112.250"	44.750"

With the addition of wind torque-compensation dummy elements, and with extra lead weights, where necessary, at the director end of the boom for weight-balance, the electrical performance can be enhanced, using the same proven mechanical parts.

Fig 20 shows the computed gain, F/R ratio and SWR for a 24-foot boom, 10-meter optimized Yagi (modified 105BA) using Hy-Gain hardware. **Fig 21** shows the same for a 26-foot boom 15-meter Yagi (modified 155BA), and **Fig 22** shows the same for a 34-foot boom (modified 205BA) 20-meter Yagi. **Tables 8** through **10** show dimensions for these designs. The original Hy-Gain taper schedule is used for each element. Only the length of the end tip (and the spacing along the boom) is changed for each element.

Multiband Yagis

So far, this chapter has discussed monoband Yagis—that is, Yagis designed for a single Amateur-Radio frequency band. Because hams have operating privileges on more than one band, multiband coverage has always been very desirable.

INTERLACING ELEMENTS

In the late 1940s, some experimenters tried interlacing Yagi elements for different frequencies on a single boom, mainly to cover the 10 and 20-meter bands (at that time the 15-meter band wasn't yet available to hams). The experimenters discovered, to their considerable chagrin, that the mutual interactions between different elements tuned to different frequencies are very difficult to handle.

Adjusting a lower-frequency element usually results in interaction with higher-frequency elements near it. In effect, the lower-frequency element acts like a retrograde reflector, throwing off the effectiveness of the higher-frequency directors nearby. Element lengths and the spacing between elements can be changed to improve performance of the higher-frequency Yagi, but the resulting compromise is rarely equal to that of an optimized monoband Yagi. A reasonable compromise for portable operation may be found in Chapter 15, Portable Antennas, by VE7CA.

TRAPPED MULTIBANDERS

Multiband Yagis using a single boom can also be made using traps. Traps allow an element to have multiple resonances. See Chapter 7, Multiband Antennas, for details on trap designs. Commercial vendors have sold trapped antennas to hams since the 1950s and surveys show that after simple wire dipoles and multiband verticals, trapped triband Yagis are the most popular antennas in the Amateur Radio service.

The originator of the trapped tribander was Chester Buchanan, W3DZZ, in his Mar 1955 *QST* article, "The Multimatch Antenna System." On 10 meters this rather unusual tribander used two reflectors (one dedicated and one with traps) and two directors (one dedicated and one with traps). On 20 and 15 meters three of the five elements were active using traps. The W3DZZ tribander employed 12 traps overall, made with heavy wire and concentric tubular capacitors to hold down losses in the traps. Each trap was individually fine tuned after construction before mounting it on an element.

Another example of a homemade tribander was the 26-foot boom 7-element 20/15/10-meter design described by Bob Myers, W1XT (ex-W1FBY) in Dec 1970 *QST*. The W1FBY tribander used only two sets of traps in the driven element, with dedicated reflectors and directors for each frequency band. Again, the traps were quite robust in this design to minimize trap losses, using $^7/_{16}$-inch aluminum tubing for the coils and short pieces of RG-8 coax as high-voltage tuning capacitors.

Only a relatively few hams actually built tribanders for themselves, mainly because of the mechanical complexity and the close tolerances required for such antennas. The traps themselves must be constructed quite accurately for reproducible results, and they must be carefully weatherproofed for long life in rain, snow, and often polluted or corrosive atmospheres.

Christmas Tree Stacks

Another possible method for achieving multiband coverage using monoband Yagis is to stack them in a "Christmas tree" arrangement. See **Fig 23**. For an installation covering 20, 15 and 10 meters, you could mount on the rotating mast just at the top of the tower the 20-meter monobander. Then perhaps 9 feet above that you would mount the 15-meter monobander, followed by the 10-meter monoband Yagi 7 feet further up on the mast. Another configuration would be to place the 10-meter Yagi in between the lower 20-meter and upper 15-meter Yagis. Whatever the arrangement, the antenna in the middle of such a Christmas-tree always suffers the most interaction from the lowest-frequency Yagi.

Dave Leeson, W6NL (ex-W6QHS), mentions that the 10-meter Yagi in his closely stacked Christmas Tree (15 meters at the top, 10 meters in the middle, and 20 meters at the bottom of the rotating mast) loses "substantial gain" because of serious interaction with the 20-meter antenna. (N6BV and K1VR calculated that the free-space gain in the W6NL stack drops to 5 dBi, compared to about 9 dBi with no surrounding antennas.) Monobanders are *definitely not* universally superior to tribanders in multiband installations. In private conversations, W6NL has indicated that he would not repeat this kind of short Christmas Tree installation again.

Forward Staggering

Some hams have built multiband Yagis on a common boom, using a technique called *forward staggering*. This means that that most (or all) of the higher-frequency elements are placed in front of any lower-frequency elements—in other words, most of the elements are not interlaced. Richard Fenwick, K5RR, described his triband Yagi design in Sep 1996 *QEX* magazine. This uses forward-stagger and open-sleeve design techniques and was optimized using several sophisticated modeling programs.

Fenwick's tribander used a 57-foot, 3-inch OD boom to hold 4 elements on 20 meters, 4 elements on 15 meters and 5 elements on 10 meters. **Fig 24** shows the element placement for the K5RR tribander. Most hams, of course, don't have the real-estate or the large rotator needed to turn such a large, but elegant solution to the interaction problem!

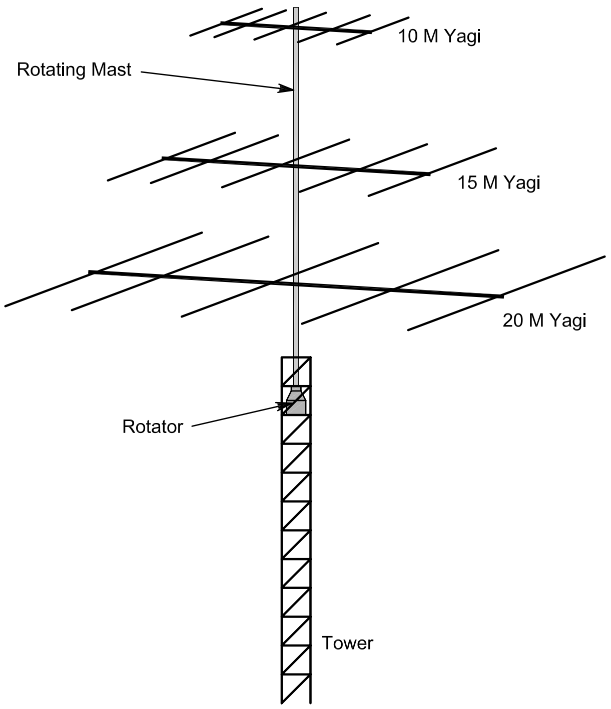

Fig 23—"Christmas Tree" stack of 20/15/10-meter Yagis spaced vertically on a single rotating mast.

Force 12 C3 "Multi-Monoband" Triband Yagi

Antenna manufacturer Force 12 also uses forward-stagger layouts and patented combinations of open- and closed-sleeve drive techniques extensively in their product line of multiband antennas, which they call "multi-monoband Yagis." **Fig 25** shows the layout for the popular Force 12 C3 triband Yagi. The C3 uses no traps, thereby avoiding any losses due to traps. The C3 consists of three 2-element Yagis on an 18-foot boom, using full-sized elements designed to withstand high winds.

The C3 feed system employs open-sleeves, where the 20-meter driver element is fed with coax through a common-mode current balun and parasitically couples to the closely spaced 15-meter driver and the two 10-meter drivers to yield a feed-point impedances close to 50 Ω on all three bands. See the section on open-sleeve dipoles in Chapter 7, Multiband Antennas.

Note the use of the forward-stagger technique in the C3, especially on 10 meters. To reduce interaction with

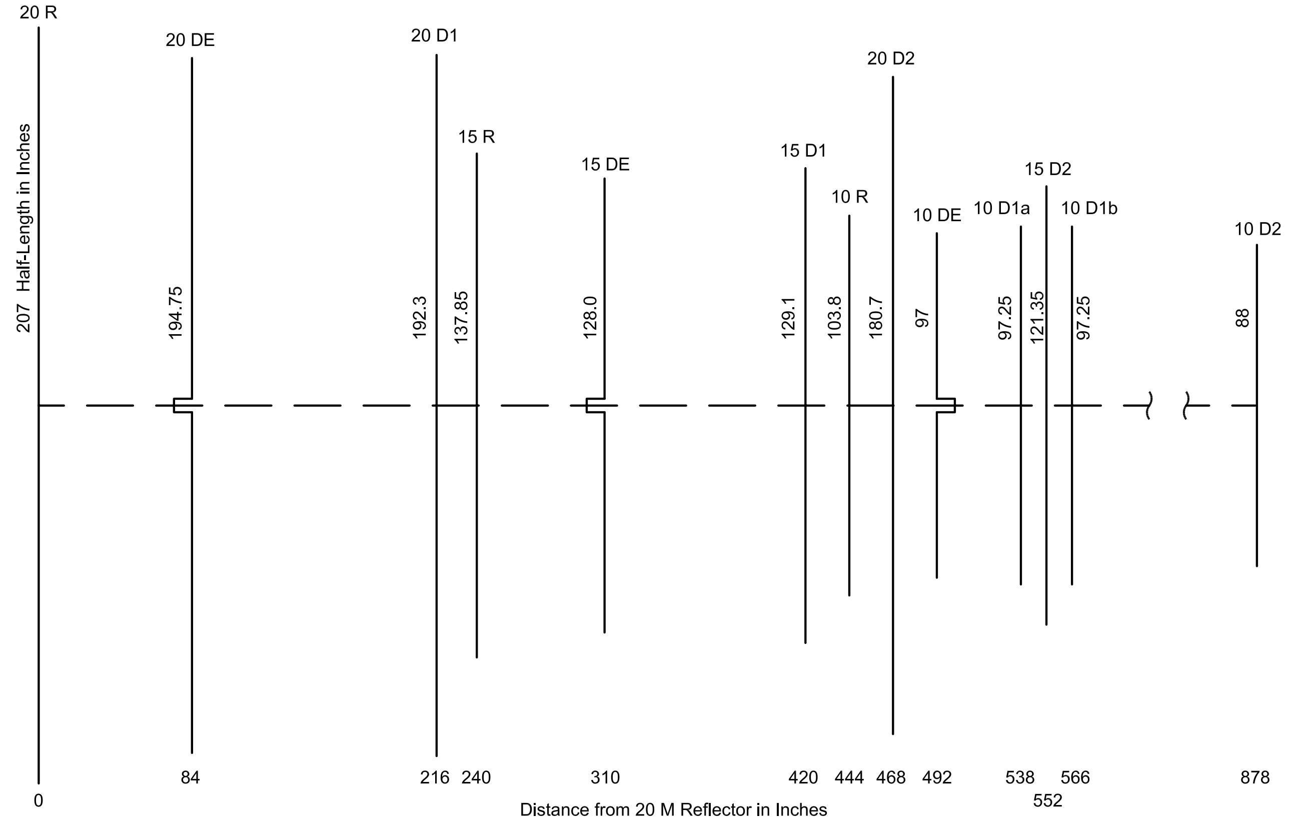
20 R
20 DE
20 D1
20 D2
15 R
15 DE
15 D1
10 R
10 DE
10 D1a
15 D2
10 D1b
10 D2
207 Half-Length in Inches
194.75
192.3
137.85
128.0
129.1
103.8
180.7
97
97.25
121.35
97.25
88
0
84
216
240
310
420
444
468
492
538
552
566
878
Distance from 20 M Reflector in Inches

Fig 24—Dimensions of K5RR's trapless tribander using "forward stagger" and open-sleeve techniques to manage interaction between elements for different frequencies.

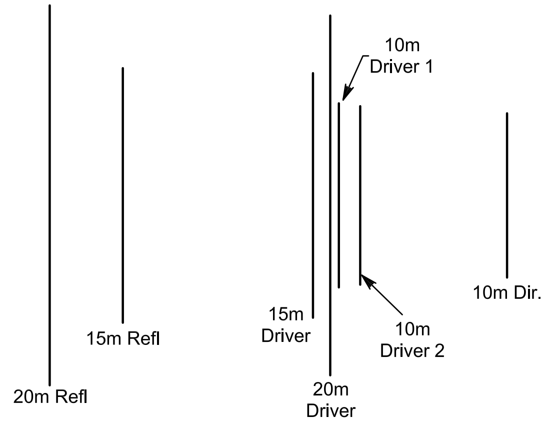
10m Driver 1
10m Dir.
15m Refl
15m Driver
10m Driver 2
20m Refl
20m Driver

Fig 25—Layout of Force 12 C3 multiband Yagi. Note that the 10-meter (driver/director) portion of the antenna is "forward staggered" ahead of the 15-meter (reflector/ driver) portion, which in turn is placed ahead of the 20-meter (reflector/driver) portion. The antenna is fed at the 20-meter driver, which couples parasitically to the 15-meter driver and the two 10-meter drivers.

the lower-frequency elements behind it, the 10-meter portion of the C3 is mounted on the boom ahead of all the lower-frequency elements, with the main 10-meter parasitic element (#7) acting as a director. The lower-frequency elements behind the 10-meter section act as retrograde reflectors, gaining some improvement of the gain and pattern compared to a monoband 2-element Yagi. A simplified *EZNEC* model of the C3 is included on the CD-ROM accompanying this book.

On 15 meters, the main parasitic element (#2) is a dedicated reflector, but the other elements ahead on the boom act like retrograde directors to improve the gain and pattern somewhat over a typical 2-element Yagi with a reflector. On 20 meters, the C3 is a 2-element Yagi with a dedicated reflector (#1) at the back end of the boom.

The exact implementation of any Yagi, of course, depends on the way the elements are constructed using telescoping aluminum tubing. The C3 type of design is no exception.

Stacked Yagis

Monoband parasitic arrays are commonly stacked either in broadside or collinear fashion to produce additional directivity and gain. In HF amateur work, the most common broadside stack is a vertical stack of identical Yagis on a single tower. This arrangement is commonly called a *vertical stack*. At VHF and UHF, amateurs often employ collinear stacks, where identical Yagis are stacked side-by-side at the same height. This arrangement is called a *horizontal stack*, and is not usually found at HF, because of the severe mechanical difficulties involved with large, rotatable side-by-side arrays.

Fig 26 illustrates the two different stacking arrangements. In either case, the individual Yagis making up the stack are generally fed in phase. There are times, however, when individual antennas in a stacked array are purposely fed out of phase in order to emphasize a particular elevation pattern. See Chapter 17, Repeater Antenna Systems, for such a case where elevation pattern steering is implemented for a repeater station.

Let's look at the reasons hams stack Yagis:

- For more gain
- For a wider elevation footprint in a target geographical area
- For azimuthal diversity—two or more directions at once
- For less fading
- For less precipitation static

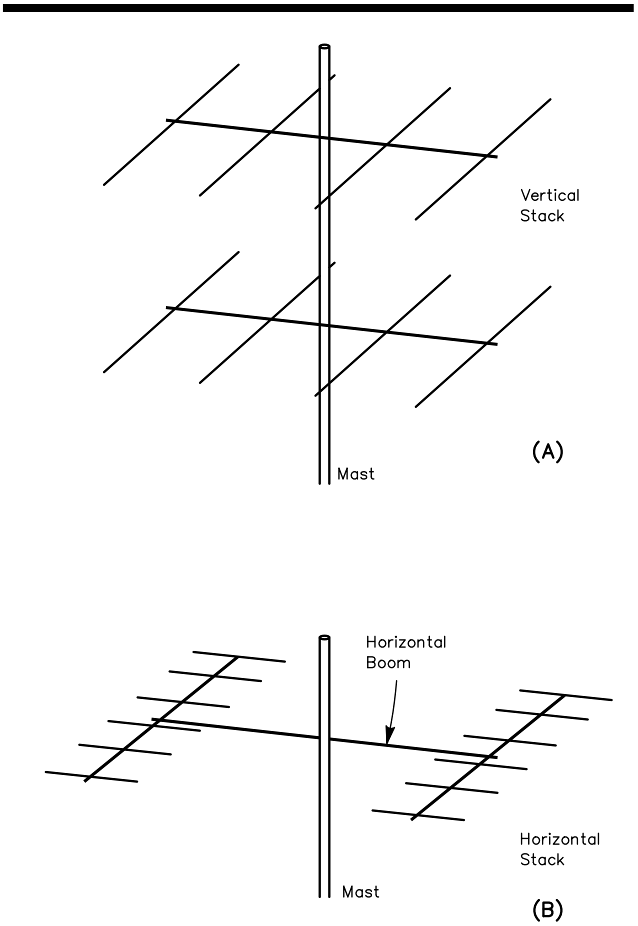

Fig 26—Stacking arrangements. At A, two Yagis are stacked vertically (broadside) on the same mast. At B, two Yagis are stacked horizontally (collinear) side-by-side. At HF the vertical stack is more common because of mechanical difficulties involved with large HF antennas stacked side-by-side, whereas at VHF and UHF the horizontal stack is common.

STACKS AND GAIN

Fig 27 compares the elevation responses for three antenna systems of 4-element 15-meter Yagis. The response for the single Yagi at a height of 120 feet peaks at an elevation of about 5°, with a second peak at 17° and a third at 29°. When operated by itself, the 60-foot high Yagi has its first peak at about 11° and its second peak beyond 34°.

The basic principle of a vertically stacked HF array is that it takes energy from higher-angle lobes and concentrates that energy into the main elevation lobe. The main lobe of the 120/60-foot stack peaks about 7° and is about 2 dB stronger than either the 60- or 120-foot antenna by itself. The shape of the left-hand side of the stack's main lobe is determined mainly by the 120-foot antenna's response. The right-hand side of the stack's main lobe is "stretched" rightwards (toward higher angles) mainly by the 60-foot Yagi, while the shape follows the curve of the 120-foot Yagi.

Look at the second and third lobes of the stack, which appear about 18° and 27°. These are about 14 dB down from the stack's peak gain, showing that energy has indeed been extracted from them. By contrast, look at the levels of the second and third lobes for the individual Yagis at 60 and 120 feet. These higher-angle lobes are almost as strong as the first lobes.

The stack squeezes higher-angle energy into its main elevation lobe, while maintaining the frontal lobe azimuth pattern of a single Yagi. This is the reason why many state-of-the-art contest stations are stacking arrays of relatively short-boom antennas, rather than stacking long-boom, higher-gain Yagis. A long-boom HF Yagi narrows the azimuthal pattern (and the elevation pattern too), making pointing the antenna more critical and making it more difficult to spread a signal over a wide azimuthal area, such as all of Europe and Asiatic Russia at one time.

STACKS AND WIDE ELEVATION FOOTPRINTS

Detailed studies using sophisticated computer models of the ionosphere have revealed that coverage

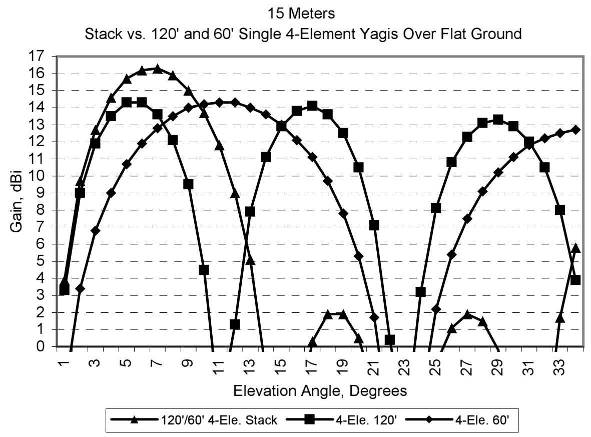

Fig 27—Comparison of elevation patterns on 15 meters for a stack of 4-element Yagis at 120 and 60 feet and individual Yagis at those two heights. The shape of the stack's response is determined mainly by that of the top antenna.

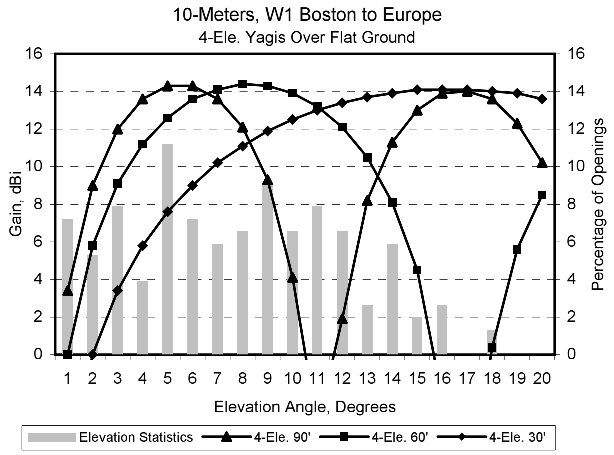

Fig 28—Comparison of elevation patterns and elevation-angle statistics for individual 10-meter TH7DX tribanders mounted over flat ground aiming from New England to Europe. No single antenna can cover the wide range of angles needed—from 1° to 18°.

of a wide range of elevation angles is necessary to ensure consistent DX or contest coverage on the HF bands. These studies have been conducted over all phases of the 11-year solar cycle, and for numerous transmitting and receiving QTHs throughout the world.

Chapter 23, Radio Wave Propagation, covers these studies in more detail, and the CD-ROM accompanying this book contains a huge number of elevation-angle statistical tables for locations all around the world. The *HFTA* (HF Terrain Assessment) program on the CD-ROM can not only compute antenna elevation patterns over irregular local terrain, but it can compare them directly to the elevation-angle statistics for a particular target geographic area.

A 10-Meter Example

Fig 28 shows the 10-meter elevation-angle statistics for the New England path from Boston, Massachusetts, to all of the continent of Europe. The statistics are overlaid with the computed elevation response for three individual 4-element Yagis, at three heights: 90, 60 and 30 feet above flat ground. In terms of wavelength, these heights are 2.60 λ, 1.73 λ and 0.86 λ high.

You can see that the 90-foot high Yagi covers the lower elevation angles best, but it has a large null in its response centered at about 11°. This null puts a big hole in the coverage for some 22% of all the times the 10-meter band is open to Europe. At those angles where the 90-foot Yagi exhibits a null, the 60-foot Yagi would be effective, and so would the 30-foot Yagi. If that is the only antenna you have, the 90-foot high Yagi would be too high for good coverage of Europe from New England.

The peak statistical elevation angle into Europe is 5°, and this occurs about 11% of all the times the 10-meter band is open to Europe from Boston. At an elevation of 5° the 30-foot high Yagi would be down almost 7 dB compared to the 90-foot high Yagi, but at 11° the 90-foot Yagi would be more than 22 dB down from the 30-foot Yagi. There is no single height at which one Yagi can optimally cover all the necessary elevation angles, especially to a large geographic area such as Europe—although the 60-foot high antenna is arguably the best compromise for a single height. To cover all the possibilities to Europe, however, you need a 10-meter antenna system that can cover equally well the entire range of elevation angles from 1° to 18°.

Fig 29 compares elevation-angle statistics for two 10-meter paths from New England to Europe and to Japan. The elevation angles needed for communications with the Far East are very low. Overlaid on Fig 29 for comparison are the elevation responses over flat ground for three different antenna systems, using identical 4-element Yagis:

- Three Yagis, stacked at 90, 60 and 30 feet
- Two Yagis, stacked at 70 and 40 feet
- One Yagi at 90 feet.

The best coverage of all the necessary angles on 10 meters to Europe is with the stack of three Yagis at 90/60/30 feet. The two-Yagi stack at 70 and 40 feet comes in a close second to Europe, and for elevation angles higher than about 9° the 70/40-foot stack is actually superior to the 90/60/30-foot stack.

Both of the stacks illustrated here give a wider *elevation footprint* than any single antenna, so that all the angles can be covered automatically without having to switch from higher to lower antennas manually. This is perhaps the major benefit of using stacks, but not the only

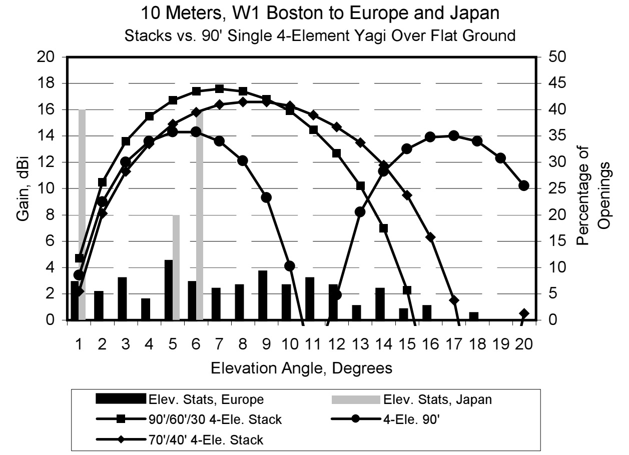

Fig 29—Combinations of 4-element Yagis over flat ground. The elevation-angle statistics into Japan from New England (Boston) are represented by the black vertical bars, while the grey vertical bars represent the elevation-angle statistics to Europe. The 90/60/30-foot stack has the best *elevation footprint* into Japan, although the 70/40-foot stack performs well also.

one, as we'll see.

To Japan, the necessary range of elevation angles is considerably smaller than that needed to a larger geographic target area like Europe. The 90/60/30-foot stack is still best on the basis of having higher gain at low angles, although the two-Yagi stack at 70 and 40 feet is a good choice too. Note that the single 90-foot high Yagi's performance is very close to the 70/40-foot stack of two Yagis at low angles, but the two-Yagi stack is superior to the single 90-foot antenna for angles higher than about 5° on 10 meters.

A 15-Meter Example

The situation is similar on 15 meters from New England to Europe. On 15 meters, the range of angles needed to fully cover Europe is 1° to 28°. This large range of angles makes covering all the angles even more challenging. Ken Wolff, K1EA, a devoted contest operator and the author of the famous *CT* contest logging program, put it very clearly when he wrote in the bulletin for the Yankee Clipper Contest Club:

"Suppose you have 15-meter Yagis at 120 feet and 60 feet, but can feed only one at a time. A 15-meter beam at 120 feet has its first maximum at roughly 5° and the first minimum at 10°. The Yagi at 60 feet has a maximum at 10° and a minimum at 2°. At daybreak, the band is just opening, signals are arriving at 3° or less and the high Yagi outperforms the low one by 5-10 dB. Late in the morning, western Europeans are arriving at angles of 10° or more, while UA6 is still arriving at 4-5°. Western Europe can be 20-30 dB louder on the low antenna than the high! What to do? Stack em!"

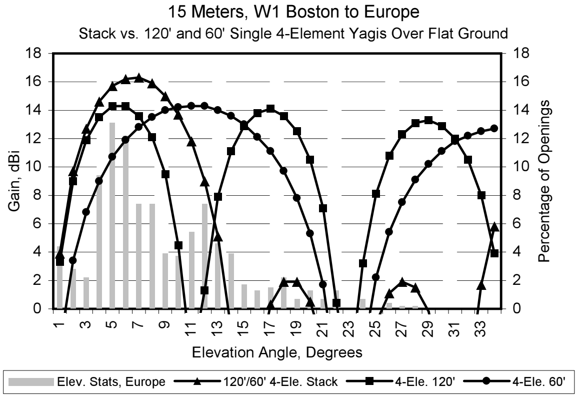

Fig 30—Comparison of elevation patterns for K1EA's illustration about 15-meter Yagis mounted over flat ground, with elevation-angle statistics to Europe added. The stack at 120 and 60 feet yields a better footprint over the range of 3° to 11° at its half-power points, better than either antenna by itself.

Fig 30 illustrates K1EA's scenario, showing the elevation statistics to Europe from Massachusetts and the elevation responses for a 120- and a 60-foot high, 4-element Yagi, both over flat ground, together with the response for both antennas operated as vertical stack. The half-power beamwidth of the stack's main lobe is 6.9°, while that for the 120-foot antenna by itself is 5.5° and that for the 60-foot antenna by itself is 11.1°. The half-power beamwidth numbers by themselves can be deceiving, mainly because the stack starts out with a higher gain. A more meaningful observation is that the stack has equal to or more gain than either of the two individual antennas from 1° to about 10°.

Is such a stack of 15-meter Yagis at 120 and 60 feet optimal for the New England to Europe path? No, it isn't, as we'll explore later, but the stack is clearly better than either antenna by itself for the scenario K1EA outlined above.

A 20-Meter Example

Take a look now at **Fig 31**, which overlays elevation-angle statistics for Europe (gray vertical bars) and Japan (black vertical bars) from Boston on 20 meters, plus the elevation responses for four different sets of antennas mounted over flat ground. Just for emphasis, the highest antenna is a 200-foot high 4-Element Yagi. It is clearly too high for complete coverage of all the needed angles into Europe. A number of New England operators have verified that this is true—a really high Yagi will open the 20-meter band to Europe in the morning and may shut it down in the afternoon, but during the middle of the day the high antenna gets soundly beaten by lower antennas.

To Japan, however, from New England the range of angles needed narrows considerably on 20 meters,

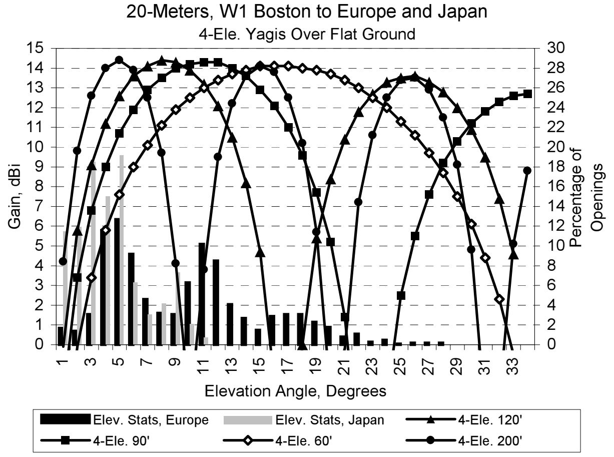

Fig 31—Comparison of elevation patterns for individual 20-meter Yagis over flat ground, compared with the range of elevation angles needed on this band from New England to Europe (gray bars) and to Japan (black bars). For fun, the response of a 200-foot high Yagi is included—this antenna is far too high to cover the needed range of angles to Europe because of its deep nulls at critical angles, like 10°. But the 200 footer is great into Japan!

from 1° to only 11°. For these angles, the 200-foot Yagi is the best antenna to work Japan from New England on 20 meters.

This is true provided that the antenna is aiming out over flat ground. The actual, generally irregular, terrain in various directions can profoundly modify the takeoff angles favored by an antenna system, particularly on steep hills. There will be more discussion on this important topic later on.

SPARE ME THE NULLS!

Now, let's look closely at some other 20-meter antennas in Fig 28, the ones at 120 and 60 feet. At an elevation angle of 8° the difference in elevation response between the 60- and 120-foot high Yagis is just over 3 dB. Can you really notice a change of 3 dB on the air? Signals on the HF bands often rise and fall quickly due to fading, so differences of 2 or 3 dB are difficult to discern. Consequently, the difference between a Yagi at 120 feet and one at 60 feet may be difficult to detect at elevation angles covered well by both antennas. But a *deep null* in the elevation response is very noticeable.

Back in 1990, when editor Dean Straw, N6BV, put up his 120-foot tower in Windham, New Hampshire, his first operational antenna was a 5-element triband Yagi, with 3 elements on 40 and 4 elements on both 20 and 15 meters. Just as the sun was going down on a late August day Straw finished connecting the feed line in the shack. The antenna seemed to be playing like it should, with a good SWR curve and a good pattern when it was rotated. So N6BV/1 called a nearby friend, John Dorr, K1AR, on the telephone and asked him to get on the air to make some signal comparisons on 20 meters into Europe.

Straw was shocked that every European they worked that evening said his signal was several S units weaker than K1AR's. Dorr was using a 4-element 20-meter monobander at 90 feet, which at first glance should have been comparable to Straw's 4-element antenna at 120 feet. But N6BV really shouldn't have been so shocked—in New England, the elevation angles from Europe late in the day on 20 meters are almost always higher than 11°, and that is true for the entire solar cycle.

The N6BV/1 station was located on a small hill, while K1AR was located on flat terrain towards Europe. The elevation response for N6BV/1's 120-foot high Yagi fell right into a deep null at 11°. This was later confirmed many times in the following eight years that the N6BV/1 station was operational. During the early morning opening on 20 meters into Europe, the top antenna was always very close to or equal to the stack of three TH7DX tribanders at 90/60/30 feet on the same tower. But in the afternoon the top antenna was *always* decidedly worse than the stack, so much so that Straw often wondered whether something had gone wrong with the top antenna!

So what's the moral to this short tale? It's simple: *The gain you can achieve, while useful, is not so important as the deep nulls you can avoid by using a stack.*

STACKING DISTANCES BETWEEN YAGIS

So far, we've examined stacks as a means of achieving more gain over an individual Yagi, while also matching the antenna system's response to the range of elevation angles needed for particular propagation paths. Most importantly, we seek to avoid nulls in the elevation response. Earlier we asked whether a 120/60-foot stack was optimal for the path from New England to Europe on 15 meters. Let's examine how the stacking distance between individual antennas affects the performance of a stack.

Fig 32 shows overlays of various combinations of 15-meter Yagis. Just for reference, a plot for a single 60-foot high Yagi is also included. Let's start by looking at the most widely spaced stack in the group: the 120/30-foot stack. Here, the spacing is obviously too large, since the second lobe is actually stronger than the first lobe. In terms of wavelength, the 90-foot spacing between antennas in this stack is 1.94 λ, a large spacing indeed.

There is a great deal of folklore and superstition among amateurs about stacking distances for HF arrays. For years, high-performance stacked Yagi arrays have been used for weak-signal DXing on the VHF and UHF bands. The most extreme example of weak-signal work is EME work (Earth-Moon-Earth, also called *moonbounce*) because of the huge path losses incurred on the way to and from the Moon. The most successful arrays used for

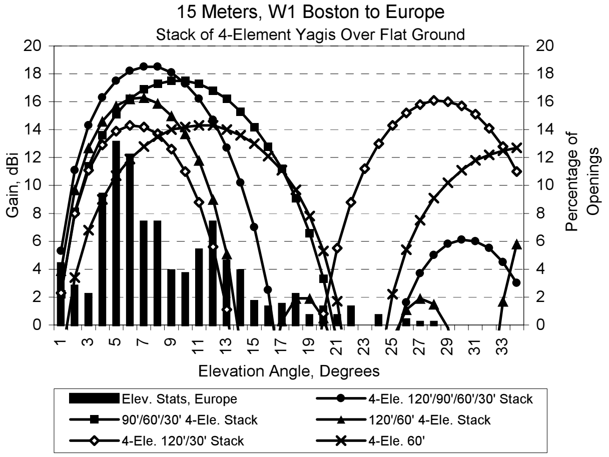

Fig 32—Various stacks towards Europe from New England for 15-meters. The stack at 120 and 30 feet is clearly suboptimal, since the second lobe is higher than the first lobe. The 120/60-foot stack is better in this regard, but is still not as good a performer as the 90/60/30-foot stack. It's debatable whether going to four Yagis in the 120/90/60/30-foot stack is a good idea because it drops below the performance of the 90/60/30-foot stack at about 10° in elevation. The exact distance between practical HF Yagis is not critical to obtain the benefits of stacking. For a stack of tribanders at 90, 60 and 30 feet, the distance in wavelengths between individual antennas is 0.87 λ at 28.5 MHz, 0.65 λ at 21.2 MHz, and 0.43 λ at 14.2 MHz.

moonbounce have low sidelobe levels and very narrow frontal lobes that give huge amounts of gain. The low sidelobes help minimize received noise, since the receive levels for signals that do manage to bounce off the Moon and return to Earth are exceedingly weak.

But HF work is different from moonbounce in that rigorously trying to minimize high-angle lobes is far less crucial at HF, where we've already shown that the main goal is to achieve gain over a wide elevation-plane footprint without any disastrous nulls in the pattern. The gain gradually increases as spacing in terms of wavelength is increased between individual Yagis in a stack, and then decreases slowly once the spacing is greater than about 1.0 λ. The difference in gain between spacings of 0.5 λ to 1.0 λ for a stack of typical HF Yagis amounts to only a fraction of a decibel. Stacking distances on the order of 0.6 λ to 0.75 λ give best gain commensurate with good patterns.

While the stack at 120/60 feet in Fig 32 doesn't have the second-lobe-stronger problem the 120/30-foot stack has, 60 feet between antennas is 1.29 λ, again outside the normal range of HF stack spacings. As a consequence, the 120/60-foot stack doesn't cover the range of elevation angles as well as it could, and is inferior to both the 90/60/30-foot stack and the 120/90/60/30-foot stack. The 120/60-foot two-Yagi stack needs at least one more antenna placed in-between to spread out the elevation-range coverage and to provide more gain.

It could be debated, but the 90/60/30-foot stack seems optimal for coverage of all the angles into Europe from New England on 15 meters. Note that the 30-foot spacing between Yagis is 0.65 λ on 21.2 MHz, right in the middle of the range of typical stack spacings.

Switching Out Yagis in the Stack

Still, the extra gain that is available at low elevation angles from a 120/90/60/30-foot high, four-Yagi stack in Fig 32 is alluring. For those statistically possible, but less likely, occasions when the elevation angle is higher than about 12°, it would be advantageous to switch out the top 120-foot Yagi and operate with only the lower three Yagis in a stack. (This also allows the top antenna to be rotated in another direction, an aspect we'll explore later.) There are even times when the incoming angles are really high and when the top two antennas might be switched out to create a 60/30-foot stack. Later in this chapter we'll explore flexible circuitry for such stack switching.

Stacking Distance and Lobes at HF

Let's look a little more closely at how a stack achieves gain and a wide elevation footprint. **Fig 33** shows a rectangular X-Y graph of the elevation response from 0° to 180° for two 3-element 15-meter Yagis (with 12-foot booms) spaced 30 feet apart (0.65 λ at 21.2 MHz), but mounted at two different heights: 95/65 and 85/55 feet. The rectangular plot gives more resolution than is possible on a polar plot. Note that the heights shown represent typical stacking heights on 15 meters—there's nothing magic about these choices. The free-space H-Plane pattern for the 30-foot spaced stack is also shown for reference.

The worst-case overhead elevation lobe, which ranges from about 60° to 120° in elevation (±30° from straight overhead at 90°), is about 14.7 dB down for the 95/65-foot stack. The overhead lobe peaks broadly at an elevation angle of about 82°. The overhead lobe for the lower 85/55-foot stack occurs at an elevation of about 64°, where it is 19 dB down.

The F/B for both 3-element sets of heights is about 15 dB, well down from the excellent 32 dB F/B for each Yagi by itself. The degradation of F/B is mainly due to mutual coupling to its neighbor in the stack.

The ground-reflection pattern in effect "modulates" the free-space pattern of the individual Yagi, but in a complex and not always intuitive manner. This is quite evident for the 85/55-foot stack at near-overhead angles. In this region things become complicated indeed, because the fourth and fifth lobes due to ground reflections are interacting with the free-space pattern of the stack.

Because the spacing remains constant at 30 feet for

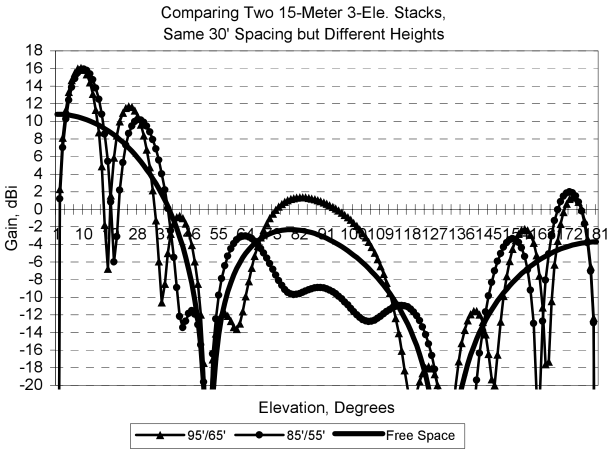

Fig 33—Rectangular plot comparing two 15-meter stacks of 3-element Yagis—each antenna is spaced 30 feet from its partner, but at different heights. The lobes are a complicated function of the antenna height, not the spacing, since that remains constant.

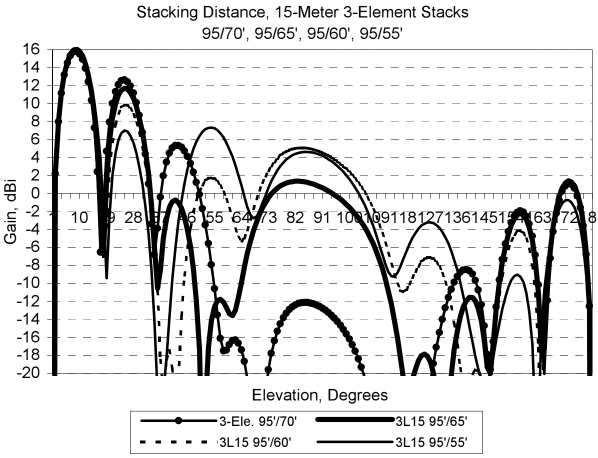

Fig 34—Four spacing scenarios for two 3-element 15-meter Yagis. Things get very complicated. The optimal spacing in terms of stacking gain is 30 feet, which is 0.65 λ. The near-overhead lobes turn out to be ugly looking, but unimportant for skywave propagation.

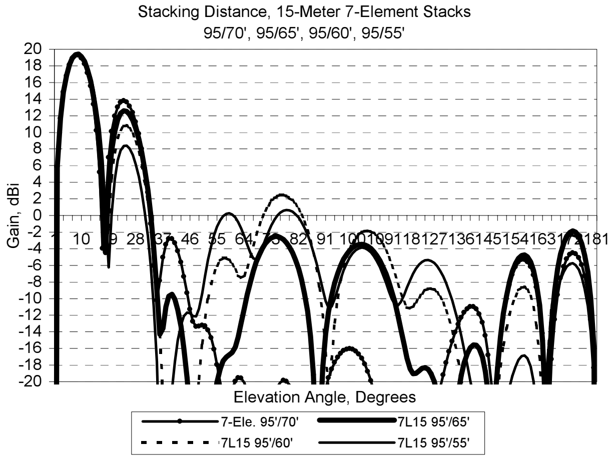

Fig 35—Four spacing scenarios for two large 7-element 15-meter Yagis (on 64-foot booms). Again, a 0.65 λ spacing (30 feet) provides the most stacking gain.

these pairs of antennas, however, the main determinant for the upper-elevation angle lobes is the distance of the horizontally polarized antennas above the ground, not the spacing between them.

Changing the Stack Spacing

Fig 34 demonstrates just how complicated things get for four different spacing scenarios. Here, the lower Yagi in the stack is moved down in 5-foot increments from the 95/70 feet level, to 95/65, 95/60 and 95/55 feet. The closest spacing, 25 feet in the 95/70-foot stack, yields nominally the "cleanest" pattern in the overhead region from 60° to 120°. The worst-case overhead lobe for the 95/70-foot stack is down 28 dB from peak. The F/B is again about 15 dB.

The worst case overhead lobe for the widest spacing, 40 feet in the 95/55-foot stack, is about 11 dB down from peak. The F/B has increased marginally, but is still only about 16 dB. It is difficult to pinpoint directly whether the spacing or the height above ground is the major determinant for the various lobe amplitudes for the 3-element stack. We'll soon look closely at whether the overhead lobe is important or not for HF work.

Longer Boom Length and Stack Spacing

Fig 35 shows the same type overlay of elevation plots, but this time for two 7-element 15-meter Yagis on gigantic 64-foot booms. These Yagis are also spaced 30 feet apart (0.65 λ at 21.2 MHz), mounted at the same four sets of heights in Fig 34. As you'd expect, the free-space elevation pattern for a stacked pair of 7-element Yagis on 64-foot booms is narrower than that for a stacked pair of 3-element Yagis on 12-foot booms. The intrinsic F/B of the longer Yagi is also better than the F/B of the shorter antenna. As a result, all lobes beyond the main lobe of the stacked 7-element pair are lower for both sets of heights than their 3-element counterparts. The worst-case overhead lobe for the 7-element 95/65-foot pair is about 22 dB down at 76° and the F/B at 172° is greater than 21 dB for all four sets of heights.

Table 11 summarizes the main performance characteristics for four sets of stacked Yagis. The first entry for each boom length is for the Yagi by itself at a height of 95 feet. Stacked configurations are next listed in order of gain. The column labeled "Worst lobe, dB re Peak" is the

Table 11
Example, Spacing Between 15–Meter Yagis

Antenna	*Peak Gain dBi*	*Worst Lobe dB re Peak*	*Worst Lobe Angle,°*	*F/B dB*	*Overhead Lobe dB re Peak*
3–Ele., 12' boom					
By itself 95'	13.2	–0.9	21	28.8	–17.5
95'/65' (Δ 30')	16.08	–4.5	25	14.9	–14.7
95'/60' (Δ 35')	16.01	–6.2	24	15.1	–10.9
95'/70' (Δ 25')	15.81	–3.2	24	14.8	–28
95'/55' (Δ 40')	15.71	–8.7	24	16.4	–11
95'/75' (Δ 20')	15.34	–2.3	23	16.3	–17.2
4–Ele., 18' boom					
By itself 95'	13.92	–1	21	28.3	–20.4
95'/65' (Δ 30')	16.63	–4.5	23	18.5	–17.3
95'/60' (Δ 35')	16.6	–6.2	24	18.2	–13.1
95'/55' (Δ 40')	16.36	–8.7	24	19.8	–13.2
95'/70' (Δ 25')	16.36	–3.3	24	20.4	–31.8
95'/75' (Δ 20')	15.92	–2.5	23	25.9	–19
5–Ele., 23' boom					
By itself 95'	14.26	–1.1	21	27.9	–22.3
95'/65' (Δ 30')	16.86	–4.6	24	20.8	–19
95'/60' (Δ 35')	16.86	–6.3	24	20.7	–14.4
95'/55' (Δ 40')	16.67	–8.8	24	23.5	–14.4
95'/70' (Δ 25')	16.59	–3.4	24	24.9	–34.4
95'/75' (Δ 20')	16.18	–2.6	23	34.3	–20.2
7–Ele., 64' boom					
By itself 95'	17.93	–2.2	21	28.9	–17.1
95'/65' (Δ 30')	19.39	–6.9	24.3	21.4	–21.9
95'/60' (Δ 35')	19.38	–8.6	24	21.4	–16.9
95'/55' (Δ 40')	19.29	–10.9	24	25.0	–18.6
95'/70' (Δ 25')	19.26	–5.5	23	24	–35.3
95'/75' (Δ 20')	19.08	–4.6	23	27	–23.4

amplitude of the second lobe due to ground reflections, and the elevation angle of that second lobe is listed as well.

Besides the 3- and 7-element designs discussed above, we've also added 4- and 5-element designs in Table 11. Over the range of stacking distances between 20 and 40 feet on 15 meters (0.43 λ to 0.86 λ), the peak gain for the 3-element stacks changes less than 0.75 dB, with the 30-foot spacing exhibiting the highest gain. The differences between peak gains versus stacking distance become smaller as the boom length increases. For example, for the 64-foot boom Yagi, the gain varies 19.39 – 19.08 = 0.31 dB for stack spacings from 20 to 40 feet.

In other words, changing the spacing from 20 to 40 feet (0.43 λ to 0.86 λ) doesn't change the gain significantly for boom lengths from 12 to 64 feet (0.26 λ to 1.38 λ). From the point of view of gain, the vertical spacing between individual antennas in an HF stack is not critical.

The worst-case lobes (generally speaking, the second lobe due to ground reflections) are highest for a Yagi operated by itself. After all, a single Yagi doesn't benefit from the redistribution of energy from higher-angle lobes into the main lobe that a stack gives. Thus, the 3-element, 12-foot boom Yagi by itself at 95 feet would have a second lobe at 21° that is only 0.9 dB down from the main lobe, while the stack of two such antennas at a 30-foot (0.65 λ) spacing at 95/65 feet would have a second lobe down 4.5 dB. As the spacing between antennas in a vertical stack increases, the second lobe is suppressed more, up to 8.7 dB at a 40-foot (0.86 λ) spacing.

Since the free-space elevation pattern for a 3-element Yagi is wider than that for a 7-element Yagi, the second lobe due to ground reflection will be somewhat reduced. This is true for all longer-boom antennas operating by themselves over ground. Used in stacks, the second lobe's amplitude will vary depending on spacing between antennas, but they range only about 6 dB.

The front-to-back ratio will also tend to increase with longer boom lengths on a properly designed Yagi. Table 11 shows that the F/B is somewhat better for closer spacings between antennas in a stack, a rather non-intuitive result, considering that the mutual coupling should be greater for closer antennas. For example, the 5-element Yagi stack with a 20-foot spacing has a exceptional F/B of 34.3 dB, compared to a F/B of 21.4 dB with the 30-foot spacing distance that gives nominally the most gain. High values of F/B, however, rarely hold over a wide frequency range because of the very critical phasing relationships necessary to get a deep null, so the difference between 34.3 and 21.4 dB would rarely be noticeable in practice.

The near-overhead lobe structure (between 60º to 120º in elevation) tends also to be lower for smaller stack spacings—for all boom lengths—peaking in this example at a spacing of 25 feet for the boom lengths considered here. Since the peak gain actually occurs with smaller spacing between Yagis in this 7-element stack, even relatively large and messy looking overhead lobes are not subtracting from the stacking gain. In the next section we'll now examine whether this overhead lobe is important or not.

Are Higher-Angle Lobes Important?

We've already shown that the exact spacing between HF Yagis is not critical for stacking gain. Further, the heights (and hence spacing) of the individual Yagis in a stack interact in a complicated fashion to determine higher-angle lobes.

Let's examine the relevance of such higher-angle lobes for stacked HF Yagis, this time in terms of interference reduction on receive. As Chapter 23, Radio Wave Propagation, points out, few DX signals arrive at elevation angles greater than about 30°. In fact, DX signals only propagate at elevation angles in the range from 1° to 30° on all the bands where operators might reasonably expect to stack Yagis—nominally from 7 to 29.7 MHz.

You should remember that the definition of the *critical frequency* for HF propagation is the highest frequency for which a wave launched directly overhead at 90° elevation is reflected back down to Earth, rather than being lost into outer space. The maximum critical frequency for extremely high levels of solar flux is about 15 MHz. In other words, high overhead angles do not propagate signals on the upper HF bands.

However, some domestic signals do arrive at relatively high elevation angles. Let's look at some scenarios where higher angles might be encountered and how the elevation patterns of typical HF stacks affect these signals. Let's examine a situation where a medium-range interfering station is on the same heading as a more distant target station.

We'll examine a typical scenario involving stations in Atlanta, Boston and Paris. The heading from Atlanta to Paris is 49º, the same heading as Atlanta to Boston. In other words, the Atlanta station would have to transmit over (and listen through) a Boston station for communication with Paris. The distance between Atlanta and Boston is about 940 miles, while the distance from Atlanta to Paris is about 4350 miles. Ground wave signals obviously cannot travel either of these distances at 21 MHz (ground wave coverage is less than about 10 miles at this frequency), and so the propagation between Atlanta to Boston and Atlanta to Paris will be entirely by means of the ionosphere.

Let's evaluate the situation on 15 meters in the month of October. We'll assume a smoothed sunspot number (SSN) of 100 and that each station puts 1500 W of power into theoretical isotropic antennas that have +10 dBi of gain at all elevation and azimuth angles. [We use such theoretical isotropic antennas because they make it easier to work in *VOACAP*. We will factor in real-world stacks later.] *VOACAP* predicts that the signal from Boston will be S9 + 8 dB in Atlanta at 1400 UTC, arriving at an elevation angle of 21.3º on a single F_2 hop. This elevation angle is higher than commonly encountered angles for DX signals, but it is still far away from near-overhead angles.

The signal from Paris into Atlanta is predicted to be about S6 for the same theoretical isotropic antennas, at an incoming elevation angle of 6.4º on three F_2 hops. The S6 level validates the rule-of-thumb that each extra hop loses approximately 10 dB of signal strength, assuming that each S unit is about 4 dB, typical for modern receivers.

Now look at **Fig 36**, which shows the response for a stack of 3-element Yagis at 90/60/30 feet over flat ground, along with the response for a similar stack of 7-element Yagis. Again, we'll assume that all three stations are using such 3-element 90/60/30-foot stacks. The stations in Atlanta and Boston point their stacks into Europe and the Parisian station points his stack towards the USA. The gain of the Atlanta array at 6.4º into Paris will be about 16 dBi, or 6 dB more than the isotropic array with its +10 dBi of gain selected for use in *VOACAP*. Similarly, the French station's transmitted signal will enjoy a 6 dB gain advantage over the isotropic array used in the *VOACAP* calculation, and thus the French signal into Atlanta will now be S6 + 12 dB, or about S9.

By comparison, the interfering signal from Boston into Atlanta will be reduced by the rearward pattern of his array, which will launch a signal at 180º – 21.3º = 158.7º in elevation at the single F_2 mode from Boston to Atlanta. From Fig 33, the Boston station's gain at this rearward elevation is going to drop from the isotropic's +10 dBi of gain down to –11 dBi, a drop of 21 dB. The signal into the Atlanta receiver will also be reduced by the pattern of the Atlanta array on receive, which has a gain of about 0 dBi at 21.3º, compared to the isotropic's +10 dBi gain at 6.4º, a net drop of 10 dB.

Thus, the Boston station's signal will drop by about 21 + 10 = 31 dB, bringing the interfering signal from Boston, which would be S9 + 8 dB for isotropic antennas, down to about S3 due to the combined effects of the arrays. This is a very significant reduction in interference. But you will note that the reduction has nothing to

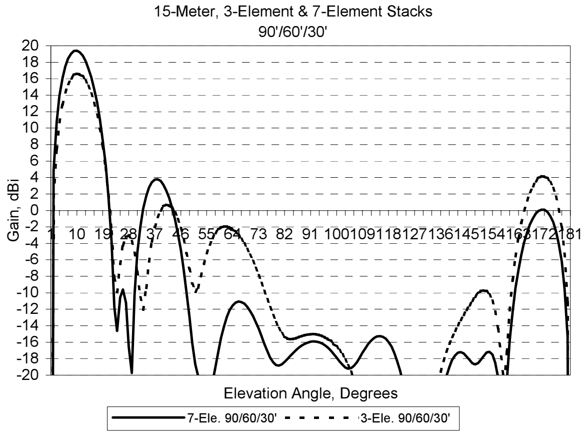

Fig 36—Stacks of three 3-element and 7-element Yagis on 15 meters at 90/60/30 feet heights. The F/B for the 7-element stack is superior to the 3-element stack mainly because the F/B is intrinsically better for the long-boom design.

do with the near-overhead lobes, dealing as it does with the trailing edge of the main lobe and the F/B lobe.

Even Higher Elevation Angles

Now let's evaluate a station that is even closer to Boston, say a station in Philadelphia. The heading from Philadelphia to Paris is 53º and the distance is 3220 miles. On the same day in October as above, *VOACAP* predicts a signal strength of S8 from Paris to Philadelphia, at a 2.7º elevation angle on two F_2 hops. Again, the *VOACAP* computations assume isotropic antennas with +10 dBi gain at all three stations. The gain of the 3-element stacks at both ends of the circuit at 2.7º is also about +10 dBi, so the signal level from Paris to Philadelphia would be S8 with the 3-element stacks.

Now *VOACAP* computes the elevation angle from Philadelphia to Boston as 56.3º, on one F_2 hop launched at an azimuth of 53º, well within the azimuthal beamwidth of the stack. *VOACAP* says the predicted signal strength for isotropic antennas with +10 dBi of gain is less than S1!

What's happening here? Boston and Philadelphia are within the "skip" region on 21 MHz and signals are skipping right over Boston from Philadelphia (and vice versa). Actual signals would be much weaker than they would be with theoretical isotropic antennas because of the actual patterns of the transmitting and receiving stacks. At an elevation angle of 56.3º the receiving stack would have a gain of –10 dBi, while at an elevation of 180º – 56.3º = 123.7º the transmitting stack would be down to –10 dBi as well. The net reduction for the stacks compared to isotropics with +10 dBi gain each would be 40 dB, putting the interfering signal well into the receiver noise.

You can safely say that near-overhead angles don't enter into the picture, simply because signals at intermediate distances are in the ionospheric skip zone and interfering signals are very weak in that zone already.

Even in situations where having a poor front-to-back ratio might be beneficial—because it alerts stations tuning across your signal that you are occupying that frequency—the ionosphere doesn't cooperate for intermediate-distance signals that are in the skip zone. Often two stations may be on the same frequency without either knowing that the other is there.

Ground Wave?

What happens, you might wonder, for ground-wave signals? Let's look at a situation where the interfering station is in the same direction as the desired target, but is only 5 miles away. Unfortunately, his signal is S9 + 50 dB. Even reducing the level by 30 dB, a huge number, is still going to make his signal 20 dB stronger than signals from your desired target location! There is not much you can do about ground-wave signals and fretting about optimizing stack heights to discriminate against local signals is generally futile.

Stacking Distances for Multiband Yagis

By definition, a stack of multiband Yagis (such as a "tribander" covering 20/15/10 meters) has a constant vertical spacing between antennas in terms of feet or meters, but not in terms of wavelength. Tribanders are no different than monobanders in terms of optimal spacing between individual antennas. Again, the difference in gain between spacings of 0.5 λ and 1.0 λ for a stack of triband Yagis amounts to only a fraction of a decibel. Furthermore, the main practical constraint that limits choice of stacking distances between any kind of Yagis, multiband or monoband, is the spacing between guy wire sets on the tower itself.

Summary, Stacking Distances

In short, let us summarize that there is nothing magical about stacking distances for practical HF Yagis—a good rule-of-thumb is a stacking distance of 0.65 λ. This is 23 feet on 10 meters, 30 feet on 15 meters and 45 feet on 20 meters for monoband stacks. Practically speaking, however, you've only got limited places where you can mount antennas on the tower—mainly where guy wires allow you to place them. This is especially applicable if you wish to rotate lower antennas on the tower, where you must clear the guys from up above.

STACKS AND FADING

The following is derived from an article by Fred Hopengarten, K1VR, and Dean Straw, N6BV, in a Feb 1994 *QST* article. Using stacked Hy-Gain TH7DXs or TH6DXXs at their respective stations, they have solicited a number of reports from stations, mainly in Europe, to compare various combinations of antennas in stacks and as single antennas. The peak gain of the stack is usually just a little bit higher than that for the best of the single antennas, which is not surprising. Even a large stack has no more than about 6 dB of gain over a single Yagi at a height favoring the prevailing elevation angle. Fading on the European path can easily be 20 dB or more, so it is very confusing to try to make definitive comparisons. They have noticed over many tests that the stacks are much less susceptible to fading compared to single Yagis. Even within the confines of a typical SSB bandwidth, frequency-selective fading occasionally causes the tonal quality of a voice to change on both receive and transmit, often dramatically becoming fuller on the stacks, and tinnier on the single antennas. This doesn't happen all the time, but is often seen. They have also observed often that the depth of a fade is less, and the period of fading is longer, on the stacks compared to single antennas.

Exactly *why* stacks exhibit less fading is a fascinating subject, for which there exist a number of speculative ideas, but little hard evidence. Some maintain that stacks outperform single antennas because they can afford *space diversity* effects, where by virtue of the difference in physical placement one antenna will randomly pick up signals that another one in another physical location might not hear.

This is difficult to argue with, and equally difficult to prove scientifically. A more plausible explanation about why stacked Yagis exhibit superior fading performance is that their narrower frontal elevation lobes can discriminate against undesired propagation modes. Even when band conditions favor, for example, a very low 3° elevation angle on 10 or 15 meters from New England to Western Europe, there are signals, albeit weaker ones, that arrive at higher elevation angles. These higher-angle signals have traveled longer distances on their journey through the ionosphere, and thus their signal levels and their phase angles are different from the signals traversing the primary propagation mode. When combined with the dominant mode, the net effect is that there is both destructive and constructive fading. If the elevation response of a stacked antenna can discriminate against signals arriving at higher elevation angles, then in theory the fading will be reduced. Suffice it to say: In practice, stacks do reduce fading.

STACKS AND PRECIPITATION STATIC

The top antenna in a stack is often much more affected by rain or snow precipitation static than is the lower antenna. N6BV and K1VR have observed this phenomenon, where signals on the lower antenna by itself are perfectly readable, while S9+ rain static is rendering reception impossible on the higher antenna or on the stack. This means that the ability to select individual antennas in a stack can sometimes be extremely important.

STACKS AND AZIMUTHAL DIVERSITY

Azimuthal diversity is a term coined to describe the situation where one of the antennas in a stack is purposely pointed in a direction different from the main direction of the stack. During most of the time in a DX contest from the East Coast, the lower antennas in a stack are pointed into Europe, while the top antenna is often rotated toward the Caribbean or Japan. In a stack of three identical Yagis, the first-order effect of pointing one antenna in a different direction is that one-third of the transmitter power is diverted from the main target area. This means that the peak gain is reduced by 1.8 dB, not a very large amount considering that signals are often 10 to 20 dB over S9 anyway when the band is open from New England to Europe.

Fig 37 shows the 3D pattern of a pair of 4-element Yagis fed in-phase at 95 and 65 feet, but where the lower antenna has been rotated 180° to fire in the –X direction. The backwards lobe peaks at a higher elevation angle because the antenna doing the radiating in this direction is lower on the tower. The forward lobe peaks at a lower angle because its main radiator is higher.

THE N6BV/1 ANTENNA SYSTEM—BRUTE FORCE FEEDING

The N6BV/1 system in Windham, New Hampshire, was located on the crest of a small hill about 40 miles from Boston, and could be characterized as a good, but not dominant, contesting station. A number of top-10 contest results were achieved from that station in the 1990s before N6BV returned to California.

There was a single 120-foot high Rohn 45 tower, guyed at 30-foot intervals, with a 100-foot horizontal spread from tower base to each guy point so there was sufficient room for rotation of individual Yagis on the tower. Each set of guy wires employed heavy-duty insulators at 57-foot intervals, to avoid resonances in the 80 through 10-meter amateur bands. There were five Yagis on the tower. A heavy-duty 12-foot long steel mast with 0.25-inch walls was at the top of the tower, turned by an Orion 2800 rotator. Two thrust bearings were used above the rotator, one at the top plate of the tower itself, and the other about 2 feet down in the tower on a modified rotator shelf plate. The two thrust bearings allowed the rotator to be removed for service.

At the top of the mast, 130 feet high, was a 5-element, computer-optimized 10-meter Yagi, which was a modified Create design on a 24-foot boom. The element tuning was modified from the stock antenna in order to achieve higher gain and a better pattern over the band. At the top of the tower (120-foot level) was mounted a Create 714X-3 triband Yagi. This was a large tribander, with a 32-foot boom and five elements. Three elements were active on 40 meters, four were active on 20 meters and four were active on 15 meters. The 40-meter elements were loaded with coils, traps and

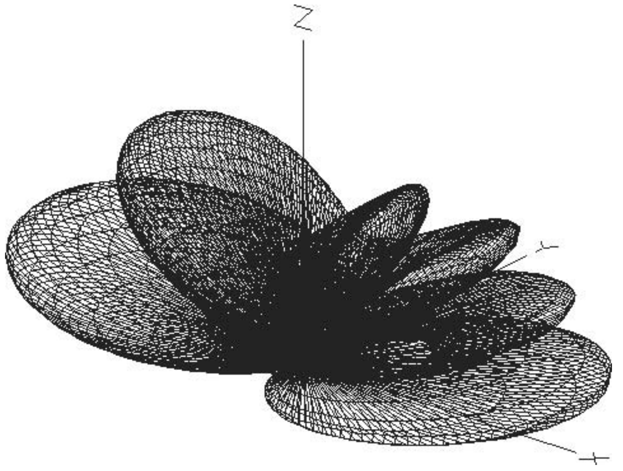

Fig 37—3D representation of the pattern for two 4-element 15-meter Yagis, with the top antenna at 95 and the bottom at 65 feet, but pointed in the opposite direction.

capacitance hats, and were approximately 46 feet long. A triband 20/15/10-meter Hy-Gain TH7DX tribander was fixed into Europe at the 90-foot level on the tower, just above the third set of guys.

At the 60-foot level on the tower, just above the second set of guys, there was a "swinging-gate" side-mount bracket, made by DX Engineering of Oregon. A Hy-Gain *Tailtwister* rotator turned a TH7DX on this side mount. (Note that both the side mount and the element spacings of the TH7DX itself prevented full rotation around the tower—about 280° of rotation was achieved with this system.) At the 30-foot level, just above the first set of guys, was located the third TH7DX, also fixed on Europe.

All five Yagis were fed with equal lengths of Belden 9913 low-loss coaxial cable, each measured with a noise bridge to ensure equal electrical characteristics. At each feed point a ferrite-bead choke balun (using seven large beads) was placed on the coax. All five coaxial cables went to a relay switch box mounted at the 85-foot level on the tower. **Fig 38** shows the schematic for the switch box, which was fed with 250 feet of 75-Ω, 0.75-inch OD Hardline coaxial cable.

The stock DX Engineering remote switch box was modified by adding relay K6, so that either the 130-foot or the 120-foot rotating antenna could be selected through a second length of 0.75-inch Hardline going to the shack. This created a *Multiplier* antenna, independent of the *Main* antennas. A second band could be monitored in this fashion while calling CQ using the main antennas on another band. Band-pass filters were required at the multiplier receiver to prevent overload from the main transmitter.

The 0.75-inch Hardline had very low losses, even when presented with a significant amount of SWR at the switch-box end. This was important, because unlike K1VR's system, no attempt was made at N6BV to maintain a constant SWR when relays K1 through K5 were switched in or out. This seemingly cavalier attitude came about because of several factors. First, there were many

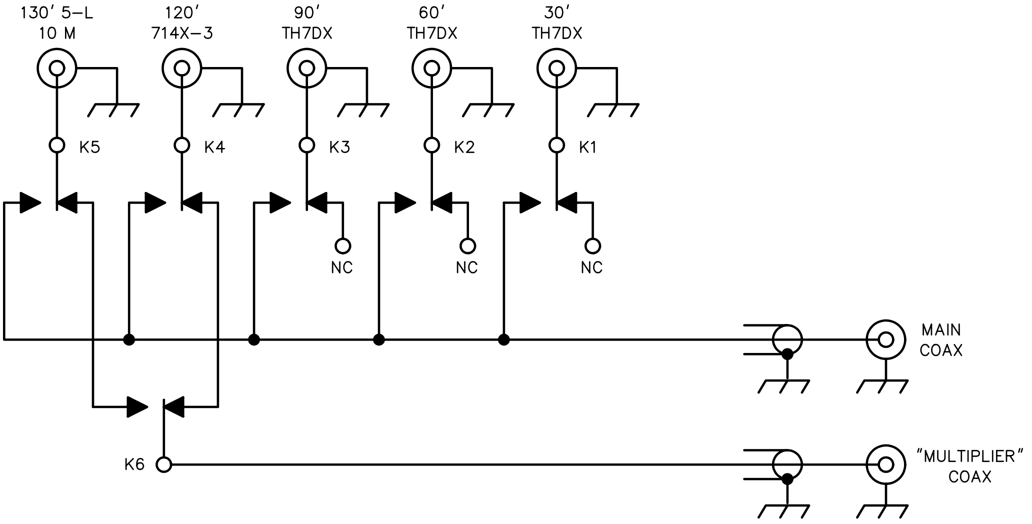

Fig 38—N6BV/1 switch box system. This uses a modified DX Engineering remote switch box, with relay K6 added to allow selection of either of the two top antennas (5-element 10-meter Yagi or 40/20/15-meter triband 714X-3) as a "multiplier" antenna. There is no special provision for SWR equalization when any or all of the Yagis are connected in parallel as a stack fed by the Main coaxial cable. Each of the five Yagis is fed with equal lengths of flexible Belden 9913 coax, so phasing can be maintained on any band. The Main and "Multiplier" coaxes going to the shack are 0.75" OD 75-Ω Hardline cables.

different combinations of antennas that could be used together in this system. Each relay coil was independently controlled by a toggle switch in the shack. N6BV was unable to devise a matching system that did not become incredibly complex because of the numerous impedance combinations used over all the five bands.

Second, the worst-case additional transmission line loss due to a 4:1 SWR mismatch when four antennas were connected in parallel on 10 meters was only 0.5 dB. It was true that the linear amplifier had to be retuned slightly when combinations of antennas were switched in and out, but this was a small penalty to pay for the reduced complexity of the switching and matching networks. The 90/60/30-foot stack into Europe was used for about 95% of the time during DX contests, so the small amount of amplifier retuning for other antenna combinations was considered only a minor irritation.

WHY TRIBANDERS?

Without a doubt, the most common question K1VR and N6BV have been asked is: "Why did you pick *tribanders* for your stacks?" Triband antennas were chosen with full recognition that they are compromise antennas. Other enterprising amateurs have built stacked tribander arrays. Bob Mitchell, N5RM, is a prominent example, with his so-called *TH28DX* array of four TH7DX tribanders on a 145-foot-high rotating tower. Mitchell employed a rather complex system of relay-selected tuned networks to choose either the upper stacked pair, the lower stacked pair or all four antennas in stack. Others in Texas have also had good results with their tribander stacks. Contester Danny Eskenazi, K7SS, has very successfully used a pair of stacked KT-34XA tribanders for years.

A major reason why tribanders were used is that over the years both authors have had good results using TH6DXX or TH7DX antennas. They are ruggedly built, mechanically and electrically. They are able to withstand New England winters without a whimper, and their 24-foot long booms are long enough to produce significant gain, despite trap-loss compromises. Amateurs speculating about trap losses in tribanders freely bandy about numbers between 0.5 and 2 dB. Both N6BV and K1VR are comfortable with the lower figure, as are the Hy-Gain engineers.

Consider this: If 1500 W of transmitter power is going into an antenna, a loss of 0.5 dB amounts to 163 W. This would create a significant amount of heat in the six traps that are on average in use on a TH6DXX, amounting to 27 W per trap. If the loss were as high as 1 dB, this would be 300 W total, or 50 W per trap. Common sense says that if the overall loss were greater than about 0.5 dB, the traps would act more like big *firecrackers* than resonant circuits! A long-boom tribander like the TH6DXX or TH7DX also has enough space to employ elements dedicated to different bands, so the compromises in element spacing usually found on short-boom 3 or 4-element tribanders can be avoided.

Another factor in the conscious choice of tribanders was first-hand frustration with the serious interaction that can result from stacking monoband antennas closely together on one mast in a Christmas Tree configuration. N6BV's worst experience was with the ambitious 10 through 40-meter Christmas Tree at W6OWQ in the early 1980s. This installation used a Tri-Ex SkyNeedle tubular crankup tower with a rotating 10-foot-long heavy-wall mast. The antenna suffering the greatest degradation was the 5-element 15-meter Yagi, sandwiched 5 feet below the 5-element 10-meter Yagi at the top of the mast, and 5 feet above the full-sized 3-element 40-meter Yagi, which also had five 20-meter elements interlaced on its 50-foot boom.

The front-to-back ratio on 15 meters was at best about 12 dB, down from the 25+ dB measured with the bottom 40/20-meter Yagi removed. No amount of fiddling with element spacing, element tuning or even orientation of the 15-meter boom with respect to the other booms (at 90° or 180°, for example) improved its performance. Further, the 20-meter elements had to be lengthened by almost a foot *on each end of each element* in order to compensate for the effect of the interlaced 40-meter elements. It was a lucky thing that the tower was a motorized crankup, because it went up and down hundreds of times as various experiments were attempted!

Interaction due to close proximity to other antennas in a short Christmas Tree can definitely destroy carefully optimized patterns of individual Yagis. Nowadays, such interaction can be modeled using a computer program such as *EZNEC* or *NEC*. A gain reduction of as much as 2 to 3 dB can easily result due to close vertical spacing of monobanders, compared to the gain of a single monoband antenna mounted in the clear. Curiously enough, at times such a reduction in gain can be found even when the front-to-back ratio is not drastically degraded, or when the front-to-back occasionally is actually *improved*.

If you plan on stacking monoband Yagis—for example, putting only 15-meters Yagis on a single tower, with your other monoband stacks on other towers—do make sure you model the system to see if any interactions occur. You may be quite surprised.

Finally, in the N6BV/1 installation, triband antennas were chosen because the system was meant to be as simple as possible, given a certain desired level of performance, of course. Triband antennas make for less mechanical complexity than do an equivalent number of monobanders. There were five Yagis on the N6BV/1 tower, yielding gain from 40 to 10 meters, as opposed to using 12 or 13 monobanders on the tower.

THE K1VR ARRAY: A MORE ELEGANT APPROACH TO MATCHING

The K1VR stacked array is on a 100-foot high Rohn

25 tower, with sets of guy wires at 30, 60 and 90 feet, made of nonconducting Phillystran. Phillystran is a nonmetallic Kevlar rope covered by black polyethylene to protect against the harmful effects of the sun's ultraviolet rays. A caution about Phillystran: Don't allow tree branches to rub against it. It is designed to work in tension, but unlike steel guy wire, it does not tolerate abrasion well.

Both antennas are Hy-Gain TH6DXX tribanders, with the top one at 97 feet and the bottom one at 61 feet. The lower antenna is rotated by a Telex Ham-M rotator on a homemade swinging-gate side mount, which allows it to be rotated 300° around the tower without hitting any guy wires or having an element swing into the tower. At the 90-foot point on the tower, a 2-element 40-meter Cushcraft Yagi has been mounted on a RingRotor so it can be rotated 360° around the tower.

After several fruitless attempts trying to match the TH6DXX antennas so that either could be used by itself or together in a stack, K1VR settled on using a relay-selected broadband toroidal matching transformer. When both triband antennas are fed together in parallel as a stack, it transforms the resulting 25-Ω impedance to 50 Ω. The transformer is wound on a T-200A powdered-iron core, available from Amidon, Palomar Engineering or Ocean State Electronics. Two lengths of twin RG-59 coax (sometimes called Siamese or WangNet), four turns each, are wound on the core. Two separate RG-59 cables could be used, but the Siamese-twin cable makes the assembly look much more tidy. The shields of the RG-59 cables are connected in series, and the center conductors are connected in parallel. See **Fig 39** for details.

Fig 40 shows the schematic of the K1VR switch box, which is located in the shack. Equal electrical lengths of 50-Ω Hardline are brought from the antennas into the shack and then to the switch box. Inside the box, the relay contacts were soldered directly to the SO-239 chassis connectors to keep the wire lengths down to the absolute minimum. K1VR used a metal box that was larger than

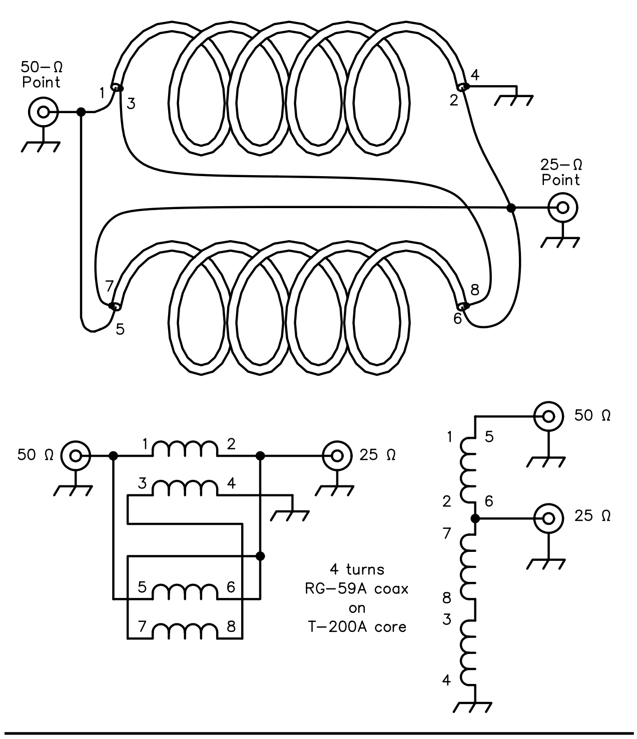
50–Ω Point
1
3
4
2
25–Ω Point
7
8
5
6
50 Ω
1
2
25 Ω
3
4
5
6
7
8
4 turns RG–59A coax on T–200A core
1
5
50 Ω
2
6
25 Ω
7
8
3
4

Fig 39—Diagram for matching transformer for K1VR stacked tribander system. The core is powdered iron-core T-200A, with four turns of two RG-59A or "Siamese" coax cables. Center conductors are connected in parallel and shields are connected in series to yield 0.667:1 turns ratio, close to desired 25-Ω to 50-Ω transformation.

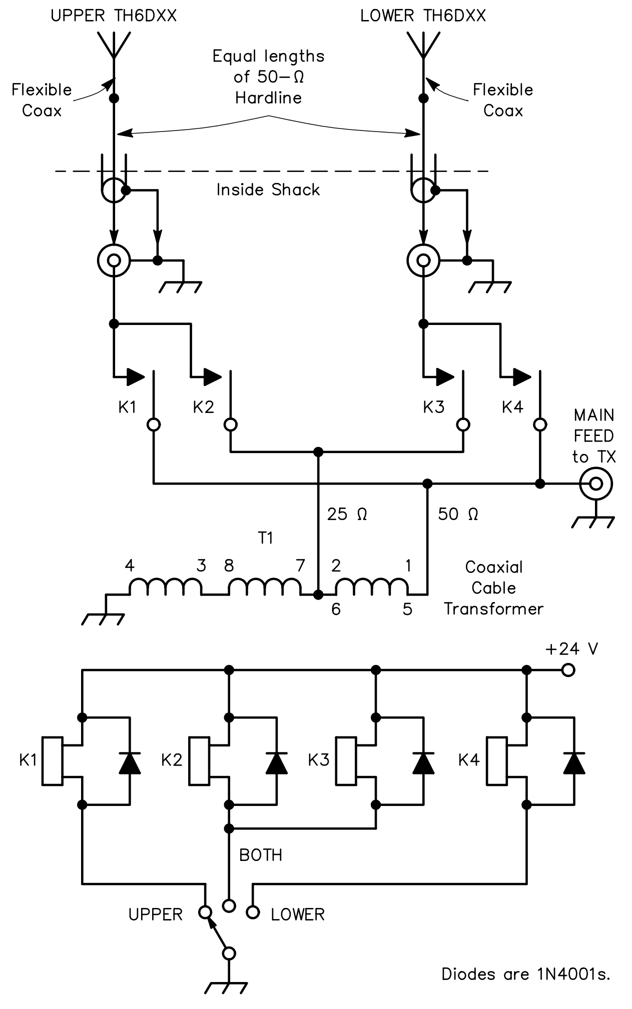
UPPER TH6DXX
LOWER TH6DXX
Equal lengths of 50–Ω Hardline
Flexible Coax
Flexible Coax
Inside Shack
K1
K2
K3
K4
MAIN FEED to TX
25 Ω
50 Ω
T1
4
3
8
7
2
1
6
5
Coaxial Cable Transformer
+24 V
K1
K2
K3
K4
BOTH
UPPER
LOWER
Diodes are 1N4001s.

Fig 40—Relay switch box for K1VR stacked tribander system. Equal lengths of 50-Ω Hardline (with equal lengths of flexible 50-Ω cable at each antenna to allow rotation) go to the switch box in the shack. The SWR on all three bands for Upper, Lower or Both switch positions is very close to constant.

might appear necessary because he wanted to mount the toroidal transformer with plenty of clearance between it and the box walls. The toroid is held in place with a piece of insulation foam board. Before placing the switch box in service, the system was tested using two 50-Ω dummy loads, with equal lengths of cable connected in parallel to yield 25 Ω. The maximum SWR measured was 1.25:1 at 14 MHz, 1.3:1 at 21 MHz and 1.15:1 at 28 MHz, and the core remained cold with 80 W of continuous output power.

One key to the system performance is that K1VR made the electrical lengths of the two hardlines the same (within 1 inch) by using a borrowed TDR (time domain reflectometer). Almost as good as Hardline, K1VR points out, would be to cut exactly the same length of cable from the same 500-foot roll of RG-213. This eliminates manufacturing tolerances between different rolls of cable.

K1VR's experience over the last 10 years has been that at the beginning of the 10 or 15-meter morning opening to Europe the upper antenna is better. Once the band is wide open, both antennas are fed in phase to cast a bigger shadow, or footprint, on Europe. By mid-morning, the lower antenna is better for most Europeans, although he continues to use the stack in case someone is hearing him over a really long distance path throughout Europe. He reports that it is always very pleasant to be called by a 4S7 or HS0 or VU2 when he is working Europeans at a fast clip!

SOME SUGGESTIONS FOR STACKING TRIBANDERS

It is unlikely that many amateurs will try to duplicate exactly K1VR's or N6BV's contest setups. However, many hams already have a tribander on top of a moderately tall tower, typically at a height of about 70 feet. It is not terribly difficult to add another, identical tribander at about the 40-foot level on such a tower. The second tribander can be pointed in a fixed direction of particular interest (such as Europe or Japan), or it can be rotated around the tower on a side mount or a Ring Rotor. If guy wires get in the way of rotation, the antenna can usually be arranged so that it is fixed in a single direction.

Insulate the guy wires at intervals to ensure that they don't shroud the lower antenna electrically. A simple feed system consists of equal-length runs of surplus 0.5-inch 75-Ω Hardline (or more expensive 50-Ω Hardline, if you are really obsessed by SWR) from the shack up the tower to each antenna. Each tribander is connected to its respective Hardline feeder by means of an equal length of flexible coaxial cable, with a ferrite choke balun, so that the antenna can be rotated.

Down in the shack, the two hardlines can simply be switched in and out of parallel to select the upper antenna only, the lower antenna only, or the two antennas as a stack. See **Fig 41**. Any impedance differences can be handled as stated previously, simply by retuning the linear amplifier, or by means of the internal antenna tuner (included in most modern transceivers) when the transceiver is run barefoot. The extra performance experienced in such a system will be far greater than the extra decibel or two that modeling calculates.

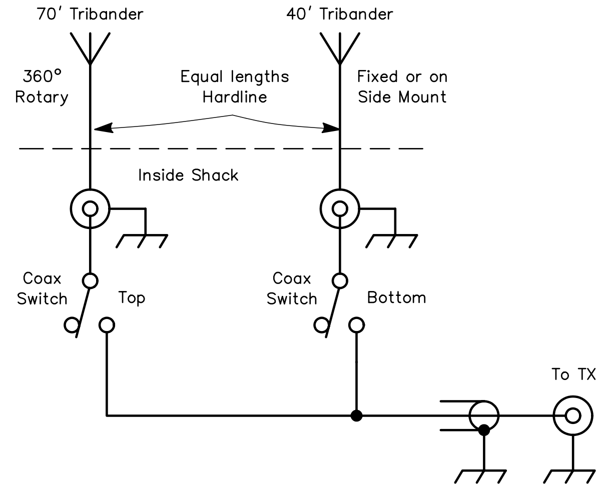

Fig 41—Simple feed system for 70/40-foot stack of tribanders. Each tribander is fed with equal lengths of 0.5-inch 75-Ω Hardline cables (with equal lengths of flexible coax at the antenna to allow rotation), and can be selected singly or in parallel at the operator's position in the shack. Again, no special provision is made in this system to equal SWR for any of the combinations.

THE WX0B APPROACH TO STACK MATCHING AND FEEDING

Earlier we mentioned how useful it would be to switch various antennas in or out of a stack, depending on the elevation angles that need to be emphasized at that moment. Jay Terleski, WX0B, of Array Solutions has designed switchable matching systems, called *StackMatches*, for stacks of monoband or multiband Yagis.

The StackMatch uses a 50-Ω to 22.25-Ω broadband transmission-line transformer to match combinations of up to three Yagis in a stack. See **Fig 42** for a schematic of the StackMatch. For selection of any 50-Ω Yagi by itself, no matching transformer is needed and Relay IN routes RF directly to the common bus going to Relay 1, 2 and 3. For selection of two Yagis together the parallel impedance is 50/2 = 25 Ω and Relay IN routes RF to the matching transformer. The SWR is 25/22.25 = 1.1:1. For three Yagis used together, the parallel impedance is 50/3 = 16.67 Ω, and the SWR is 22.25/16.67 = 1.3:1.

The broadband transformer consists of four trifilar turns of #12 enamel-insulated wire wound on a Ferrite Corporation FT-240 2.4-inch OD core made of #61

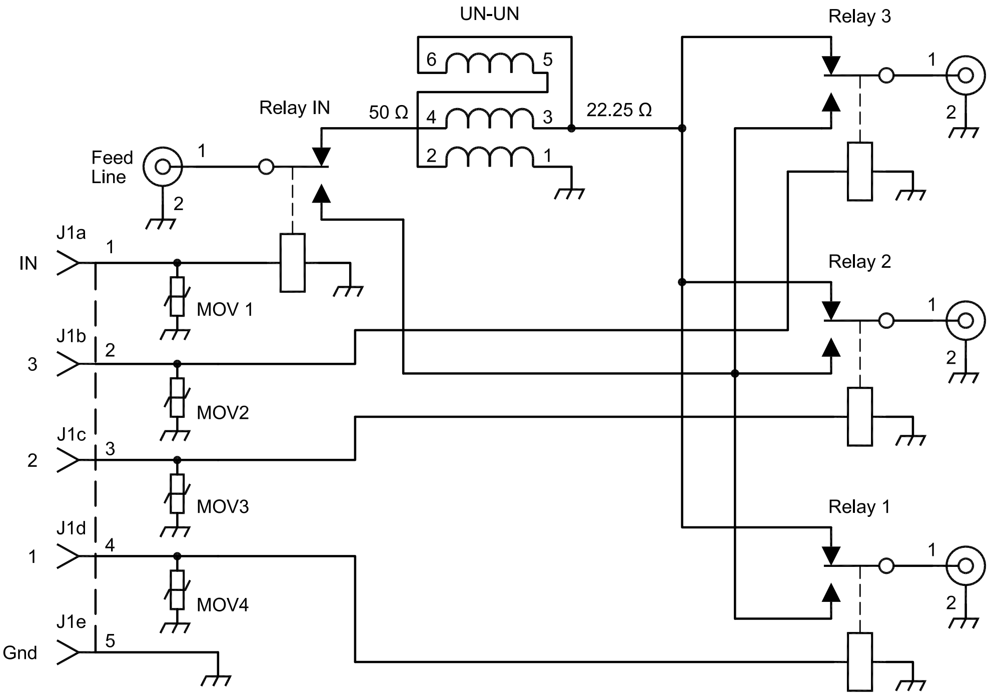

Fig 42—Schematic of WX0B's StackMatch 2000 switchbox, which uses a broadband transmission line transformer using trifilar #12 enamel-insulated wires. (*Courtesy Array Solutions.*)

material ($\mu = 125$). WX0B uses 10-A relays enclosed in plastic cases to do the RF switching, selected by a control box at the operating position. (10-A relays can theoretically handle $10\ \text{A}^2 \times 50\ \Omega = 5000$ W.) **Fig 43** shows a photo of the transmission-line transformer and StackMaster PCB.

The control/indicator box uses a diode matrix to switch various combinations of antennas in/out of the stack. Three LEDs lined up vertically on the front panel indicate which antennas in a stack are selected.

"BIP/BOP" OPERATION

The contraction "BIP" means "both in-phase," while "BOP" means "both out-of-phase." BIP/BOP refer to stacks containing two Yagis, although the term is commonly used for stacks containing more than two Yagis. In theory, feeding a stack with the antennas out-of-phase will shift the elevation response higher than in-phase feeding.

Fig 44 shows a rectangular plot comparing BIP/BOP operation of two 3-element 15-meter Yagis at heights of 2 λ and 1 λ (93 and 46 feet) over flat ground. The BOP pattern is the higher-angle lobe and the two lobes cross over about 14°. The maximum amplitude of the BOP stack's gain is about ½ dB less than the BIP pair. For reference, the pattern of a single 46-foot high Yagi is overlaid on the pattern for the stacks.

The most common method for feeding one Yagi 180° out-of-phase is to include an extra electrical half wavelength of feed-line coax going to one of the antennas. This method obviously works on a single frequency band and thus is not applicable to stacks of multiband Yagis, such as tribanders. For such multiband stacks, feeding only the lower antenna(s)—by switching out higher antenna(s) in the stack—is a practical method for achieving better coverage at medium or high elevation angles.

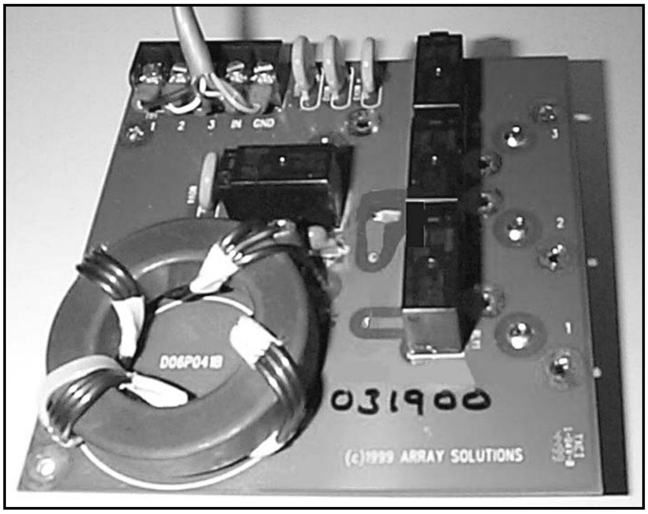

Fig 43—Inside view of StackMatch. (*Photo courtesy Array Solutions.*)

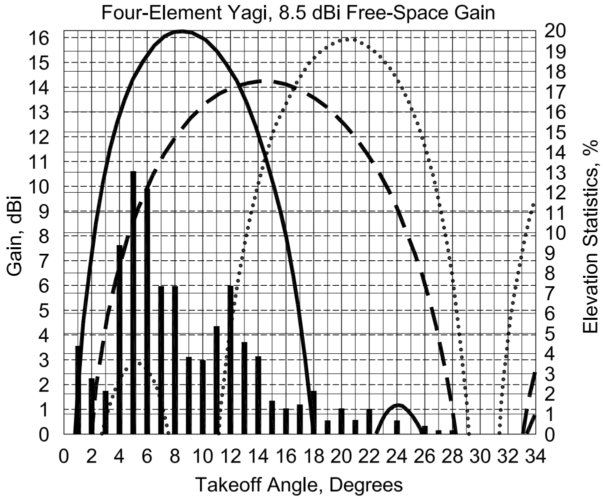

Fig 44—*HFTA* screen shot of "BIP/BOP" operation of two 4-element 15-meter Yagis at 93 and 46 feet above flat ground. The elevation response in BOP (both out-of-phase) operation is shifted higher, peaking at about 21°, compared to the BIP (both in-phase) operation where the peak is at 8°. The dashed line is response of single Yagi at 46 feet.

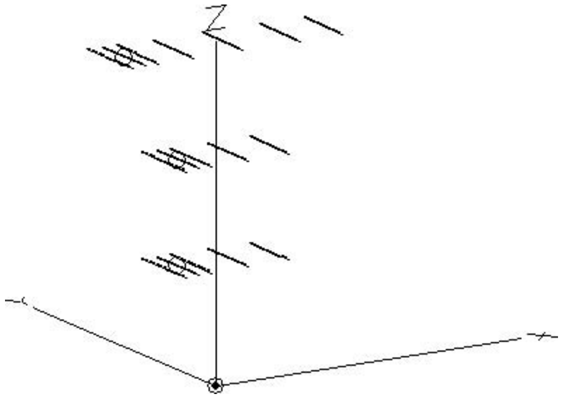

Fig 45—Stacking dissimilar Yagis. In this case a 7-element 10-meter Yagi is stacked over two 5-element Yagis. Note the displacement of the 7-element Yagi's driven element compared to the position of the two 5-element Yagis. This leads to an undesired phase shift for the higher antenna.

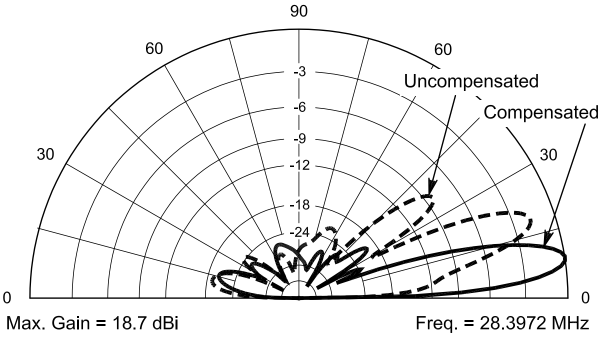

Fig 46—Comparison of elevation responses for 7/5/5-element 10-meter stacks, with and without compensation for driven-element offset.

STACKING DISIMILAR YAGIS

So far we have been discussing vertical stacks of identical Yagis. Less commonly, hams have successfully stacked dissimilar Yagis. For example, consider a case where two 5-element 10-meter Yagis are placed 46 and 25 feet above flat ground, with a 7-element 10-meter Yagi at 68 feet on the same tower. See **Fig 45**, which is a schematic of the layout for this stack. Note that the driven element for the top 7-element Yagi is well behind the vertical plane of the driven elements for the two 5-element Yagis. This offset distance must be compensated for with a phase shift in the drive system for the top Yagi.

Fig 46 shows the elevation-pattern responses for uncompensated (equal-length feed lines) and the compensated (additional 150° of phase shift to top Yagi) stacks. These patterns were computed using *EZNEC ARRL*, which is included with this book. Not only is about 1.7 dB of maximum gain lost, but the peak elevation angle is shifted upwards by 11° from the optimal takeoff angle of 8°—where some 10 dB of gain is also lost. Without compensation, this is a severe distortion of the stack's elevation pattern.

For RG-213 coax, the extra length needed to provide an additional 150° of phase shift = 150°/360° λ = 0.417 λ = 9.53 feet at 28.4 MHz. This was computed using the program *TLW* (Transmission Line for Windows) included on the CD-ROM accompanying this book.

It is not always possible to compensate for dissimilar Yagis in a stack with a simple length of extra coax, so you should be sure to model such combinations to make sure that they work properly. A safe alternative, of course, is to stack only identical Yagis, feeding all of them with equal lengths of coax to ensure in-phase operation.

Real-World Terrain and Stacks

So far, the stacking examples shown have been for flat ground. Things can become a lot more complicated when you deal with real-world irregular terrains! See Chapter 3, The Effects of Ground, for a description of the *HFTA* (High Frequency Terrain Assessment) program that is included with this book.

Fig 47 shows the *HFTA*-computed 20-meter elevation responses towards Europe (at an azimuth of 45°) for three antennas at the N6BV/1 location in Windham, New Hampshire. Overlaid as a bar graph are the elevation-angle statistics for the path to all of Europe from New England (Massachusetts). The stack at 90/60/30 feet clearly covers all the angles needed best at 14 MHz. The N6BV 120-foot Yagi has a severe null in the region from about 7° to about 20°, with the deepest part of that null occurring at about 13° and is roughly comparable to the 90/60/30-foot stack between 2° to 7°.

In practice, the 120-foot Yagi was indeed comparable to the stack during morning openings to Europe on 20 meters, when the elevation angles are typically about 5°. In the New England afternoon, when the elevation angles typically rise to about 11°, the 120-foot Yagi was always distinctly inferior to the stack.

For reference, the response of a single 120-foot high Yagi over flat ground is also shown. Note that the N6BV 120-foot high Yagi has about 3 dB more gain at a 5° takeoff angle than does its flatland counterpart. This additional gain is due to the focusing effects of the local terrain, which had about a 3° downwards slope towards Europe.

Fig 48 shows the *HFTA*-computed 15-meter elevation responses towards Europe for the 90/60/30-foot stack at 90/60/30 feet at N6BV/1, compared to the same 120-foot high Yagi and a 90/60/30-foot stack, but this time over flat ground. Again, the N6BV/1 terrain towards Europe has a significant effect on the gain of the stack compared to that of an identical stack over flat ground. In fact, the peak gain of 20.1 dBi at a 4° elevation angle is close to moon-bounce levels.

OPTIMIZING OVER LOCAL TERRAIN

There are only a small number of possibilities to optimize an installation over local terrain:

- Change the antenna height(s) above ground.
- Stack two (or more) Yagis.
- Change the spacing between stacked Yagis.
- Move the tower back from a cliff (or a hill).
- BIP/BOP (Both In Phase/Both Out of Phase).

The *HFTA* program on the CD-ROM accompanying this book can be used, together with Digital Elevation Model (DEM) topographic data available on the Internet, to evaluate all these options.

SO NEAR, YET SO FAR

It is sometimes very surprising to compare elevation responses for different towers located at various points on the same property, particularly when that prop-

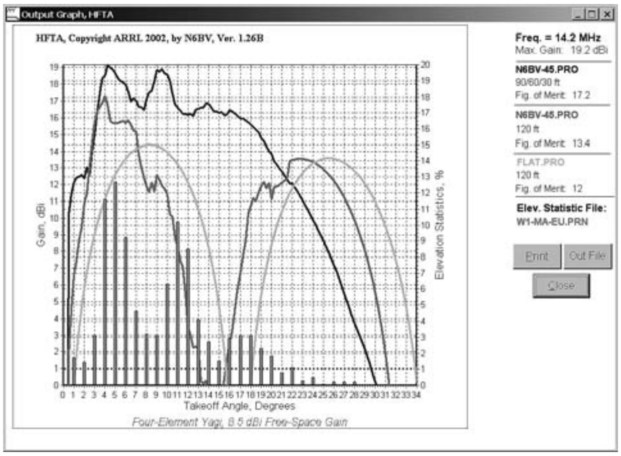

Freq. = 14.2 MHz
Max. Gain: 19.2 dBi

N6BV-45.PRO
90/60/30 ft
Fig. of Merit: 17.2

N6BV-45.PRO
120 ft
Fig. of Merit: 13.4

FLAT.PRO
120 ft
Fig. of Merit: 12

Elev. Statistic File:
W1-MA-EU.PRN

Fig 47—*HFTA* screen shot showing how complicated things become when real-world irregular terrain is analyzed. This is the 20-meter elevation pattern for the N6BV/1 station location in Windham, NH, for the 90/60/30-foot stack of triband TH7DX Yagis and a 4-element Yagi at 120 feet on the same tower. For comparison, the response of a 120-foot Yagi over flat ground is also included.

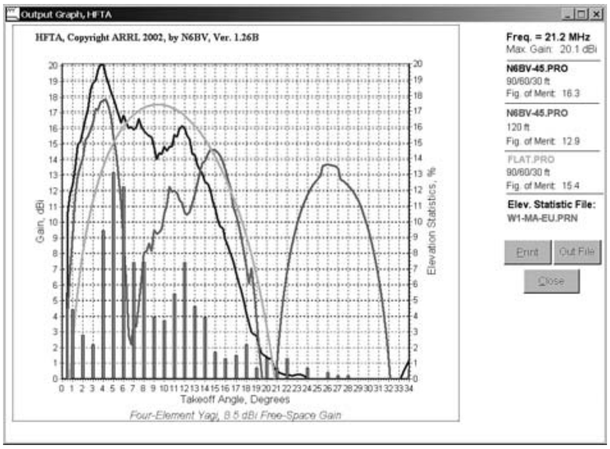

Freq. = 21.2 MHz
Max. Gain: 20.1 dBi

N6BV-45.PRO
90/60/30 ft
Fig. of Merit: 16.3

N6BV-45.PRO
120 ft
Fig. of Merit: 12.9

FLAT.PRO
90/60/30 ft
Fig. of Merit: 15.4

Elev. Statistic File:
W1-MA-EU.PRN

Fig 48—*HFTA* screen shot showing the 15-meter elevation pattern for the N6BV/1 station location in Windham, NH, for the 90/60/30-foot stack of triband TH7DX Yagis and a 4-element Yagi at 120 feet on the same tower. For comparison, the response of a 120-foot Yagi over flat ground is also included.

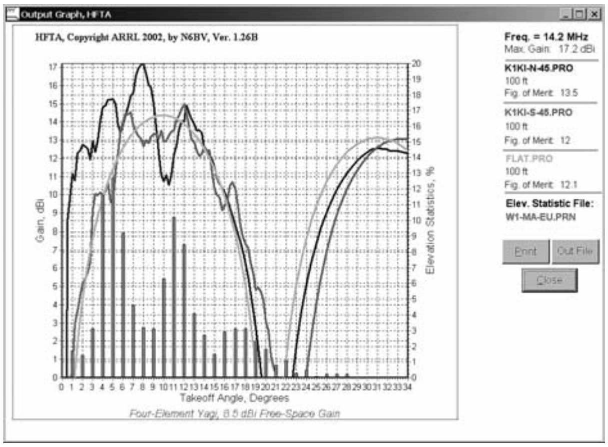
Output Graph, HFTA

HFTA, Copyright ARRL 2002, by N6BV, Ver. 1.26B

Freq. = 14.2 MHz
Max Gain: 17.2 dBi

K1KI-N-45.PRO
100 ft
Fig. of Merit: 13.5

K1KI-S-45.PRO
100 ft
Fig. of Merit: 12

FLAT.PRO
100 ft
Fig. of Merit: 12.1

Elev. Statistic File:
W1-MA-EU.PRN

Gain, dBi
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Elevation Statistics, %
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Takeoff Angle, Degrees
Four-Element Yagi, 8.5 dBi Free-Space Gain
Print
Out File
Close

Fig 49—*HFTA* screen shot showing the 20-meter elevation pattern for K1KI's North and South towers, with 100-foot high 4-element Yagis pointing into Europe at an azimuth of 45°. The responses are surprisingly different for two towers separated by only 600 feet.

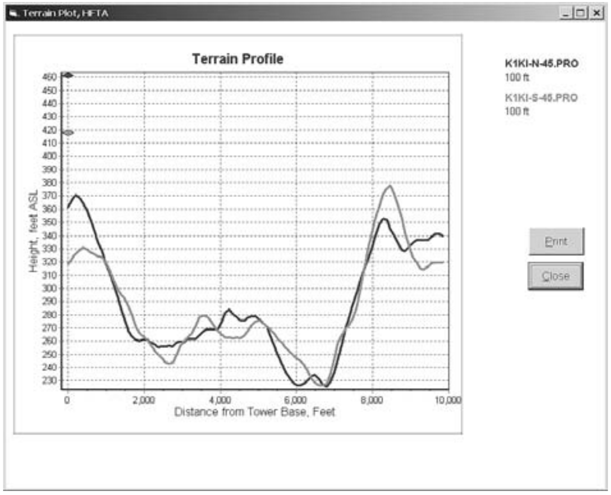
Terrain Plot, HFTA

K1KI-N-45.PRO
100 ft

K1KI-S-45.PRO
100 ft

Terrain Profile
Height, feet ASL
230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450 460
0
2,000
4,000
6,000
8,000
10,000
Distance from Tower Base, Feet
Print
Close

Fig 50—K1KI's terrain profiles for the North and South towers at an azimuth of 45° into Europe.

erty is located in the mountains. **Fig 49** shows the computed elevation responses for three 100-foot high 14-MHz Yagis over three terrains towards Europe: from the North tower at K1KI's location in West Suffield, Connecticut, from the South tower at K1KI, and over flat ground. The elevation response from the South tower follows that over flat ground well, while the response from the North tower is quite a bit stronger at low elevation angles—about 1.5 dB on average, as the Figure of Merit shows from *HFTA*.

Fig 50 shows the reason why this happens—the terrain from the North tower slopes down quickly towards Europe, while the terrain from the South tower goes out almost 900 feet before starting to fall off. These two towers are about 600 feet apart.

Moxon Rectangle Beams

LB Cebik, W4RNL, has written extensively about the *Moxon rectangle*, an antenna invented by Les Moxon, G6XN, derived from a design by VK2ABQ. The Moxon rectangle beam takes less space horizontally than a conventional 2-element Yagi design, yet it offers nearly the same amount of gain and a superior front-to-back ratio. And as an additional benefit, the drive-point impedance is close to 50 Ω, so that it doesn't need a matching section.

For example, rather than a "wingspan" of 17 feet for the reflector in a conventional 2-element 10-meter Yagi, the Moxon rectangle is 13 feet wide, a saving of almost 25%. The Moxon rectangle W4RNL created for *The ARRL Antenna Compendium, Vol 6*, had an SWR less than 2:1 from 28.0 to 29.7 MHz, with a gain over ground of 11 dBi. It had a F/B of 15 dB at 28.0 MHz, more than 20 dB at 28.4 MHz, and 12 dB at 29.7 MHz.

The Moxon rectangle relies on controlling the spacing (hence controlling the coupling) between the ends of the driven element tips and the ends of the reflector tips, which are both bent toward each other. See **Fig 51**, which shows the general outline for W4RNL's 10-meter aluminum Moxon rectangle. The tips of the elements are kept a fixed distance from each other by PVC spacers. The closed rectangular mechanical assembly gives some rigidity to the design, keeping it stable in the wind. W4RNL described other Moxon rectangle designs using wire elements in June 2000 *QST*.

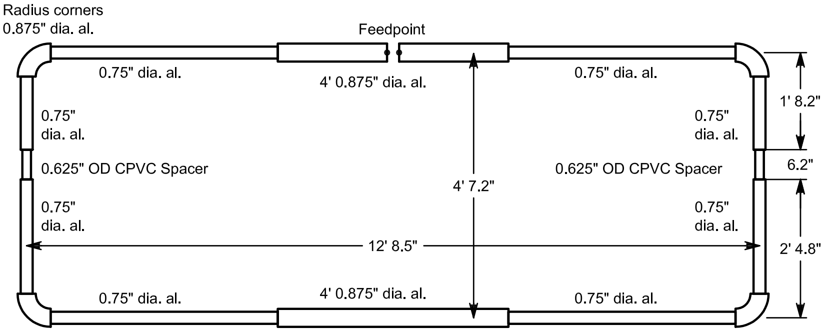
Radius corners
0.875" dia. al.
Feedpoint
0.75" dia. al.
4' 0.875" dia. al.
0.75" dia. al.
1' 8.2"
0.75"
dia. al.
0.75"
dia. al.
0.625" OD CPVC Spacer
0.625" OD CPVC Spacer
6.2"
4' 7.2"
0.75"
dia. al.
0.75"
dia. al.
12' 8.5"
2' 4.8"
0.75" dia. al.
4' 0.875" dia. al.
0.75" dia. al.

Fig 51—General outline of the 10-meter aluminum Moxon rectangle, showing tubing dimensions.

Chapter 12

Quad Arrays

Chapter 11, HF Yagi Arrays, discussed Yagi arrays as systems of approximately half-wave dipole elements that are coupled together mutually. You can also employ other kinds of elements using the same basic principles of analysis. For example, loops of various types may be combined into directive arrays. A popular type of parasitic array using loops is the *quad antenna*, in which loops having a perimeter of about one wavelength are used in much the same way as half-wave dipole elements in the Yagi antenna.

Clarence Moore, W9LZX, created the quad antenna in the early 1940s while he was at the Missionary Radio Station HCJB in Quito, Ecuador. He developed the quad to combat the effects of corona discharge at high altitudes. The problem at HCJB was that their large Yagi was literally destroying itself by melting its own element tips. This occurred due to the huge balls of corona it generated in the thin atmosphere of the high Andes mountains. Moore reasoned correctly that closed loop elements would generate less high voltage—and hence less corona—than would the high impedances at the ends of a half-wave dipole element.

Fig 1 shows the original version of the two-element quad, with a driven element and a parasitic reflector. The square loops may be mounted either with the corners lying on horizontal and vertical lines, as shown at the left, or with two sides horizontal and two vertical (right). The feed points shown for these two cases will result in horizontal polarization, which is commonly used.

Since its inception, there has been controversy whether the quad is a better performer than a Yagi. Chapter 11 showed that the three main electrical performance parameters of a Yagi are gain, response patterns (front-to-rear ratio, F/R) and drive impedance/SWR. Proper analysis of a quad also involves checking all these parameters across the entire frequency range over which you intend to use it. Both a quad and a Yagi are classified as "parasitic, end-fire arrays." Modern antenna modeling by computer shows that monoband Yagis and quads with the same boom lengths and optimized for the same performance parameters have gains within about 1 dB of each other, with the quad slightly ahead of the Yagi.

Fig 2 plots the three parameters of gain, front-to-

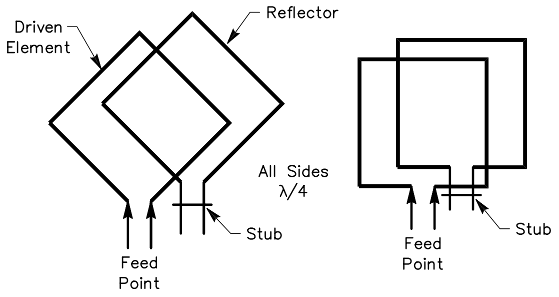

Fig 1—The basic two-element quad antenna, with driven-element loop and reflector loop. The driven loops are electrically one wavelength in circumference ($^1/_4$ wavelength on a side); the reflectors are slightly longer. Both configurations shown give horizontal polarization. For vertical polarization, the driven element should be fed at one of the side corners in the arrangement at the left, or at the center of a vertical side in the "square" quad at the right.

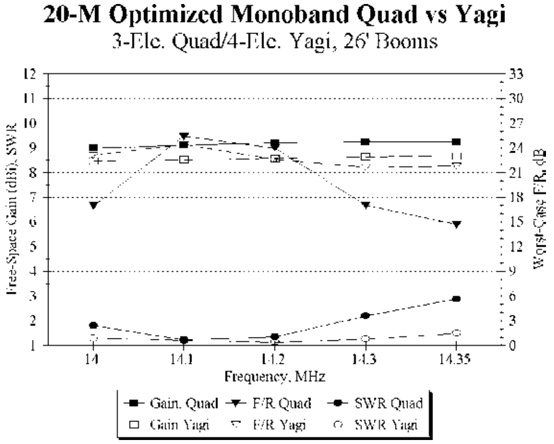

Fig 2—Comparison of gain, F/R and SWR over the 14.0 to 14.35-MHz range for an optimized three-element quad and an optimized three-element Yagi, both on 26-foot booms. The quad exhibits almost 0.5 dB more gain for the same boom length, but doesn't have as good a rearward pattern over the whole frequency range compared to the Yagi. This is evidenced by the F/R curve. The quad's SWR curve is also not quite as flat as the Yagi. The quad's design emphasizes gain more than the other two parameters.

rear ratio (F/R) and SWR over the 14.0 to 14.35-MHz band for two representative antennas—a monoband three-element quad and a monoband four-element Yagi. Both of these have 26-foot booms and both are optimized for the best compromise of gain, F/R and SWR across the whole band.

While the quad in Fig 2 consistently exhibits about 0.5 dB more gain over the whole band, its F/R pattern toward the rear isn't quite as good as the Yagi's over that span of frequencies. This quad attains a maximum F/R of 25 dB at 14.1 MHz, but it falls to 17 dB at the bottom end of the band and 15 dB at the top. On the other hand, the Yagi's F/R stays consistently above 21 dB across the whole 20-meter band. The quad's SWR rises to just under 3:1 at the top end of the band, but stays below 2:1 from 14.0 to almost 14.3 MHz. The Yagi's SWR remains lower than 1.5:1 over the whole band.

The reason the Yagi in Fig 2 has more consistent responses for gain, F/R and SWR across the whole 20-meter band is that it has an additional parasitic element, giving two additional variables to play with—that is, the length of that additional element and the spacing of that element from the others on the boom.

Yagi advocates point out that it is easier to add extra elements to a Yagi, given the mechanical complexities of adding another element to a quad. Extra parasitic elements give a designer more flexibility to tailor all performance parameters over a wide frequency range. Quad designers have historically opted to optimize strictly for gain and, as stated before, they can achieve as much as 1 dB more gain than a Yagi with the same length boom. But in so doing, a quad designer typically has to settle for front-to-rear patterns that are peaked over more narrow frequency ranges. The 20-meter quad plots in Fig 2 actually represent an even-handed approach, where the gain is compromised slightly to obtain a more consistent pattern and SWR across the whole band.

Fig 3 plots gain, F/R and SWR for two 10-meter monoband designs: a five-element quad and a five-element Yagi, both placed on 26-foot booms. The quad now has the same degrees of freedom as the Yagi, and as a consequence the pattern and SWR are more consistent across the range from 28.0 to 28.8 MHz. The quad's F/R remains above about 18.5 dB from 28.0 to 28.8 MHz. Meanwhile, the Yagi maintains an F/R of greater than 22 dB over the same range, but has almost 0.8 dB less gain compared to the quad at the low end of the band, eventually catching up at the high end of the band. The SWR for the quad is just over 2:1 at the bottom of the band, but remains less than 2:1 up to 28.8 MHz. The SWR on the Yagi remains less than 1.6:1 over the whole band.

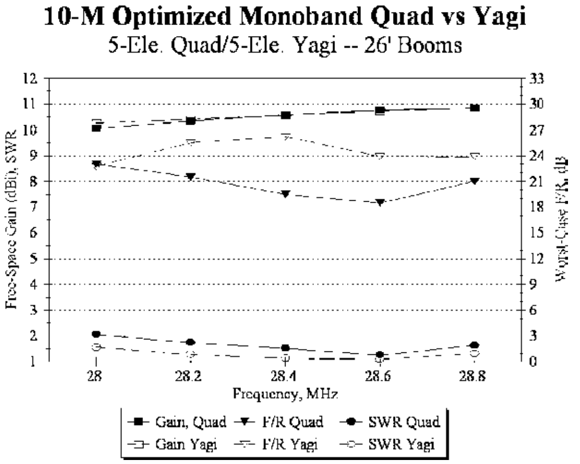

Fig 3—Comparison of gain, F/R and SWR over the 28.0 to 28.8-MHz range for an optimized five-element quad and an optimized five-element Yagi, both on 26-foot booms. The gain advantage of the quad is about 0.25 dB at the low end of the band. The F/R is more peaked in frequency for the quad, however, than the Yagi.

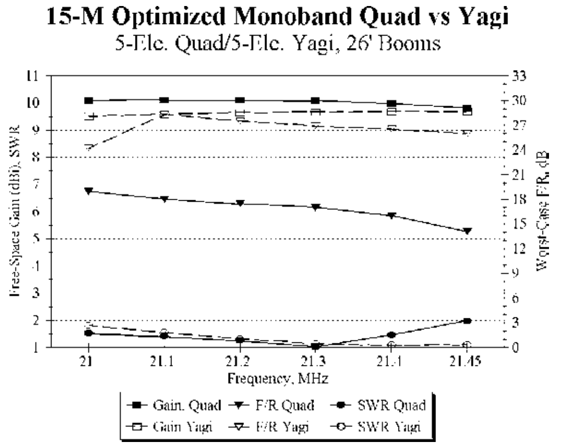

Fig 4—Comparison of gain, F/R and SWR over the 21.0 to 21.45-MHz range for an optimized 5-element quad and optimized 5-element Yagi, both on 26-foot booms. The quad enjoys a gain advantage of about 0.5 dB over most of the band. Its rearward pattern is not as good as the Yagi, which remains higher than 24 dB across the whole range, compared to the quad, which remains in the 16-dB average range.

Fig 4 shows the performance parameters for two 15-meter monoband designs: a five-element quad and a five-element Yagi, both on 26-foot booms. The quad is still the leader in gain, but has a less optimal rearward pattern and a somewhat less flat SWR curve than the Yagi. One thing should be noted in Figs 2 through 4. The F/R pattern on the Yagi is largely determined by the response at the 180° point, directly in back of the frontal lobe. This point is usually referred to when discussing the "front-to-back ratio."

The quad on the other hand has what a sailor might term "quartering lobes" (referring to the direction back towards the "quarterdeck" at the stern of a sailing vessel) in the rearward pattern. These quartering lobes are often

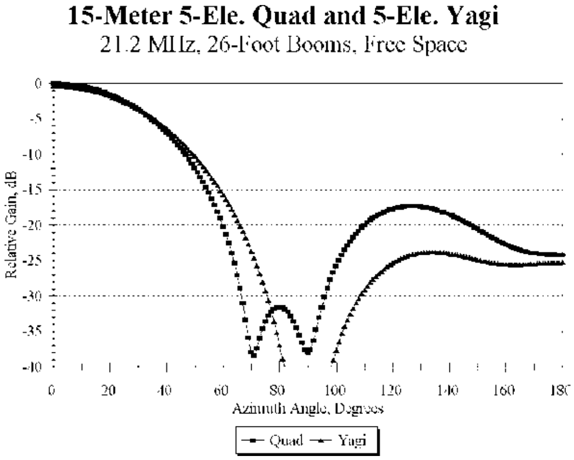

Fig 5—Comparing the pattern of the 15-meter quad and Yagi shown in Fig 4. The quad has a slightly narrower frontal beamwidth (it has 0.5 dB more gain than the Yagi), but has higher "rear quartering" sidelobes at about 125° (with a twin sidelobe, not shown, at 235°). These sidelobes limit the worst-case front-to-rear (F/R) to about 17 dB, while the F/B (at 180°, directly at the back of the quad) is more than 24 dB for each antenna.

worse than the response at 180°, directly in back of the main beam. **Fig 5** overlays the free-space E-Field responses of the 15-meter quad and Yagi together. At 21.2 MHz, the quad actually has a front-to-back ratio (F/B) of about 24 dB, excellent in anyone's book. The Yagi at 180° has a F/B of about 25 dB, again excellent.

However, at an azimuth angle of about 125° (and at 235° azimuth on the other side of the main lobe) the quad's "quartering lobe" is down only some 17 dB, setting the worst-case F/R at 17 dB also. As explained in Chapter 11, the reason F/R is more important than just the F/B is that on receive signals can come from any direction, not just from directly behind the main beam.

Table 1 lists the dimensions for the three computer-optimized monoband quads shown in Figs 2, 3 and 4.

Table 1

Dimensions for Optimized Monoband Quads in Figs 2, 3 and 4, on 26-Foot Booms

	14.2 MHz	*21.2 MHz*	*28.4 MHz*
Reflector	73′ 9"	49′ 6"	37′ 3"
R-DE Spacing	17′ 8"	7′	6′ 4"
Driven Element	71′ 8"	47′ 6"	35′ 9"
DE-D1 Spacing	8′ 3"	5′	5′ 6"
Director 1	68′ 7"	46′ 8"	34′ 8"
D1-D2 Spacing	—	6′ 8"	6′ 9"
Director 2	—	46′ 10"	35′ 2"
D2-D3 Spacing	—	7′ 4"	7′ 5"
Director 3	—	45′ 8"	34′ 2"
Feed method	Direct 50 Ω	Direct 50 Ω	Direct 50 Ω

Is a Quad Better at Low Heights than a Yagi?

Another belief held by some quads enthusiasts is that they need not be mounted very high off the ground to give excellent DX performance. Quads are somehow supposed to be greatly superior to a Yagi at the same height above ground. Unfortunately, this is mainly wishful thinking.

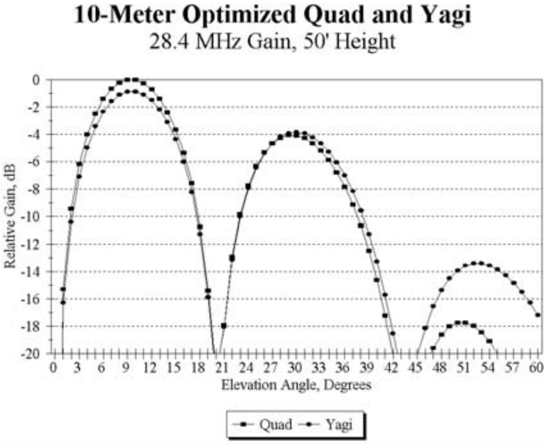

Fig 6—A comparison on 10 meters between an optimized five-element quad and an optimized five-element Yagi, both mounted 50 feet high over flat ground and both employing 26-foot booms. There is no appreciable difference in the peak elevation angle for either antenna. In other words, a quad does not have an appreciable elevation-angle advantage over a Yagi mounted at the same boom height. Note that the quad achieves its slightly higher gain by taking energy from higher-angle lobes and concentrating that energy in the main elevation lobe. This is a process that is similar to what happens with stacked Yagis.

Fig 6 compares the same two 10-meter antennas as in Fig 3, but this time with each one mounted on a 50-foot tower over flat ground, rather than in theoretical free space. The quad does indeed have slightly more gain than a Yagi with the same boom length, as it has in free space. This is evidenced by the very slight compression of the quad's main lobe, but is more obvious when you look at the third lobe, which peaks at about 53° elevation. In effect, the quad squeezes some energy out of its second and third lobes and adds that to the first lobe. However, the difference in gain compared to the Yagi is only 0.8 dB for this particular quad design at a 9° elevation angle. And while it's true that every dB counts, you can also be certain that on the air you wouldn't be able to tell the difference between the two antennas. After all, a 10- to 20-dB variation in the level of signals is pretty common because of fading at HF.

Multiband Quads

On the other hand, one of the valid reasons quads have remained popular over the years is that antenna homebrewers can build multiband quads far more easily than they can construct multiband Yagis. In effect, all you have to do with a quad is add more wire to the existing support arms. It's not quite as simple as that, of course, but the idea of ready expandability for other bands is very appealing to experimenters.

Like the Yagi, the quad does suffer from interactions between wires of different frequencies, but the degree of interaction between bands is usually less for a quad. The higher-frequency bands are the ones that often suffer most from any interaction, for both Yagis and quads. For example, the 10- and 15-meter bands are usually the ones affected most by nearby 20-meter wires in a triband quad, while the 20-meter elements are not affected by the 10- or 15-meter elements.

Modern computer modeling software can help you counteract at least some of the interaction by allowing you to do virtual "retuning" of the quad on the computer screen — rather than clinging precariously to your tower fiddling with wires. However, the programs (such as *NEC-2* or *EZNEC*) that can model three-dimensional wire antennas such as quads typically run far more slowly than those designed for monoband Yagis (such as *YW* included with this book). This makes optimizing rather tedious, but you use the same considerations for tradeoffs between gain, pattern (F/R) and SWR over the operating bandwidth as you do with monoband Yagis.

CONSTRUCTING A QUAD

The parasitic element shown in Fig 1 is tuned in much the same way as the parasitic element in a Yagi antenna. That is, the parasitic loop is tuned to a lower frequency than the driven element when the parasitic is to act as a reflector, and to a higher frequency when it is to act as a director. Fig 1 shows the parasitic element with an adjustable tuning stub, a convenient method of tuning since the resonant frequency can be changed simply by changing the position of the shorting bar on the stub. In practice, it has been found that the length around the loop should be approximately 3.5% greater than the self-resonant length if the element is a reflector, and about 3.0% shorter than the self-resonant length if the parasitic element is a director. Approximate formulas for the loop lengths in feet are:

$$\text{Driven Element} = \frac{1008}{f_{\text{MHz}}}$$

$$\text{Reflector} = \frac{1045}{f_{\text{MHz}}}$$

$$\text{Director} = \frac{977}{f_{\text{MHz}}}$$

These are valid for quad antennas intended for operation below 30 MHz and using uninsulated #14 stranded copper wire. At VHF, where the ratio of loop circumference to conductor diameter is usually relatively small, the circumference must be increased in comparison to the wavelength. For example, a one-wavelength loop constructed of $^1/_4$-inch tubing for 144 MHz should have a circumference about 2% greater than in the above equation for the driven element.

Element spacings on the order of 0.14 to 0.2 free-space wavelengths are generally used. You would employ the smaller spacings for antennas with more than two elements, where the structural support for elements with larger spacings tends to become challenging. The feed-point impedances of antennas having element spacings on this order have been found to be in the 40- to 60-Ω range, so the driven element can be fed directly with coaxial cable with only a small mismatch.

For spacings on the order of 0.25 wavelength (physically feasible for two elements, or for several elements at 28 MHz) the impedance more closely approximates the impedance of a driven loop alone—that is, 80 to 100 Ω. The feed methods described in Chapter 26 can be used, just as in the case of the Yagi.

Making It Sturdy

The physical sturdiness of a quad is directly proportional to the quality of the material used and the care with which it is constructed. The size and type of wire selected for use with a quad antenna is important because it will determine the capability of the spreaders to withstand high winds and ice. One of the more common problems confronting the quad owner is that of broken wires. A solid conductor is more apt to break than stranded wire under constant flexing conditions. For this reason, stranded copper wire is recommended. For 14-, 21- or 28-MHz operation, #14 or #12 stranded wire is a good choice. Soldering of the stranded wire at points where flexing is likely to occur should be avoided.

You may connect the wires to the spreader arms in many ways. The simplest method is to drill holes through the fiberglass at the appropriate points on the arms and route the wires through the holes. Soldering a wire loop across the spreader, as shown later, is recommended. However, you should take care to prevent solder from flowing to the corner point where flexing could break it.

While a boom diameter of 2 inches is sufficient for smaller quads using two or even three elements for 14, 21 and 28 MHz, when the boom length reaches 20 feet or longer a 3-inch diameter boom is highly recommended. Wind creates two forces on the boom, vertical and horizontal. The vertical load on the boom can be reduced with a guy-wire truss cable. The horizontal forces on the boom are more difficult to relieve, so 3-inch diameter tubing is desirable.

Generally speaking, three grades of material can be used for quad spreaders. The least expensive material is bamboo. Bamboo, however, is also the weakest material normally used for quad construction. It has a short life, typically only three or four years, and will not withstand a harsh climate very well. Also, bamboo is heavy in contrast to fiberglass, which weighs only about a pound per 13-foot length. Fiberglass is the most popular type of spreader material, and will withstand normal winter climates. One step beyond the conventional fiberglass arm is the pole-vaulting arm. For quads designed to be used on 7 MHz, surplus "rejected" pole-vaulting poles are highly recommended. Their ability to withstand large amounts of bending is very desirable. The cost of these poles is high, and they are difficult to obtain. See Chapter 21 for dealers and manufacturers of spreaders.

Diamond or Square?

The question of how to orient the spreader arms has been raised many times over the years. Should you mount the loops in a diamond or a square configuration? Should one set of spreaders be horizontal to the earth as shown in Fig 1 (right), or should the wire itself be horizontal to the ground (spreaders mounted in the fashion of an X) as shown in Fig 1 (left)? From the electrical point of view, there is not enough difference in performance to worry about.

From the mechanical point of view there is no question which version is better. The diamond quad, with the associated horizontal and vertical spreader arms, is capable of holding an ice load much better than a system where no vertical support exists to hold the wire loops upright. Put another way, the vertical poles of a diamond array, if sufficiently strong, will hold the rest of the system erect. When water droplets are accumulating and forming into ice, it is very reassuring to see water running down the wires to a corner and dripping off, rather than just sitting there on the wires and freezing. The wires of a loop (or several loops, in the case of a multiband antenna) help support the horizontal spreaders under a load of ice. A square quad will droop severely under heavy ice conditions because there is nothing to hold it up straight.

Of course, in climates where icing is not a problem, many amateurs point out that they like the aesthetics of the square configuration. There are thousands of square-configuration quads in temperate areas around the world.

Another consideration will enter into your choice of orientation for a quad. You must mount a diamond quad somewhat higher on the mast or tower than for an equivalent square array, just to keep the bottom spreader away from the tower guys when you rotate the antenna.

Two Multiband Quads

This section describes two multiband quad designs. The first is a large triband 20/15/10-meter quad built on a 26-foot boom made of 3-inch irrigation tubing. This antenna has three elements on 20 meters, four elements on 15 meters, and five elements on 10 meters. **Fig 7** shows a photograph of the five-element triband quad.

The second project is a compact two-element triband quad on an 8-foot boom that covers 20, 17, 15, 12 and 10 meters. We call this a "pentaband" quad since it covers five bands. This antenna uses five concentric wire loops mounted on each of the two sets of spreaders. Either antenna may be constructed in a diamond or square configuration.

While the same basic construction techniques are employed for both multiband quads, the scale of the larger triband antenna makes it a far more ambitious undertaking! The large quad requires a strong tower and a rugged rotator. It also requires a fair amount of real estate in order to raise the quad to the top of the tower without entangling trees or other antennas.

A FIVE-ELEMENT, 26-FOOT BOOM TRIBAND QUAD

Five sets of element spreaders are used to support the three elements used on 14 MHz, four elements on 21 MHz and five elements on 28-MHz. We chose to use four elements on 15 meters in this design (rather than the five we could have been employed on this length of boom) because the difference in optimized performance wasn't great enough to warrant the extra complexity of using five elements. The dimensions are listed in **Table 2**, and are designed for center frequencies of 14.175, 21.2 and 28.4 MHz.

The spacing between elements has been chosen to provide good compromises in performance consistent with boom length and mechanical construction. You can see that the element spacings for 20 meters are quite different from those for the optimized monoband design. This is because the same set of spreaders is used for all

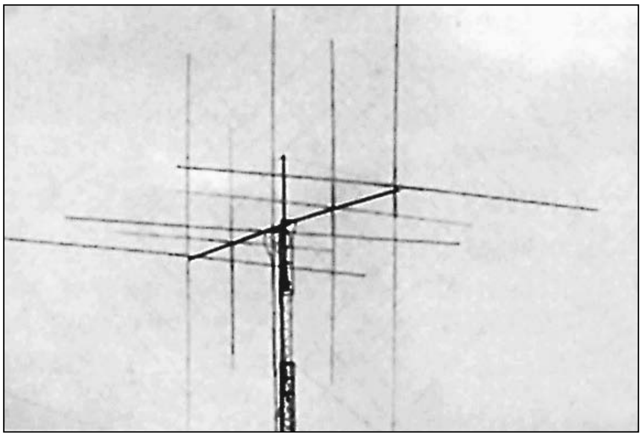

Fig 7—Photo of the three-band, five-element quad antenna.

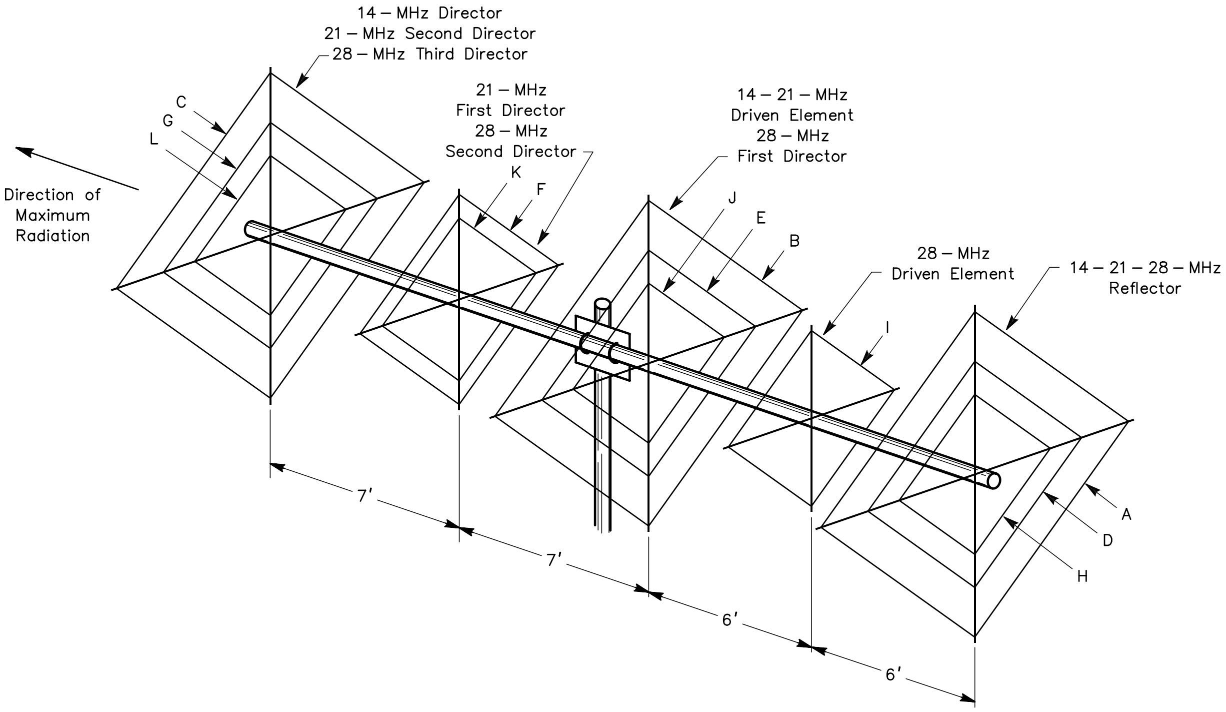
14 – MHz Director
21 – MHz Second Director
28 – MHz Third Director
21 – MHz
First Director
28 – MHz
Second Director
14 – 21 – MHz
Driven Element
28 – MHz
First Director
C
G
L
Direction of
Maximum
Radiation
K
F
J
E
B
28 – MHz
Driven Element
14 – 21 – 28 – MHz
Reflector
I
A
D
H
7′
7′
6′
6′

Fig 8—Layout for the three-band, five-element quad, not drawn to scale. See Table 2 for dimensions.

three bands on three out of the five elements, and the higher-frequency bands dictate the spacing because they are more critical.

Each of the parasitic loops is closed (ends soldered together) and requires no tuning. **Fig 8** shows the physical layout of the triband quad. **Fig 9** plots the computed free-space gain, front-to-rear ratio and SWR response across the 20-meter band. With only a few degrees of freedom in tuning and spacing of the three elements, it is impossible to spread the response out to cover the entire 20-meter band. The compromise design results in a rearward pattern that varies from a worst-case of just under 10 dB at the high end of the band, to a peak F/R of just under 19 dB at 14.2 MHz, in the phone portion of the band. The F/R is about 11 dB at the low end of the band.

The SWR remains under 3:1 for the entire 20-meter band, rising to 2.8:1 at the high end. The feed system for this triband quad consists of three separate 50-Ω coax

Table 2
Three-Band Five-Element Quad on 26-Foot Boom

	14.15 MHz	*21.2 MHz*	*28.4 MHz*
Reflector	72′ 6"	49′ 4"	36′ 8"
R-DE Spacing	12′	12′	6′
Driven Element	71′	47′ 6"	35′ 4"
DE-D1 Spacing	14′	7′	6′
Director 1	68′ 6"	46′ 8"	34′ 8"
D1-D2 Spacing	—	14′	7′
Director 2	—	46′ 5"	34′ 8"
D2-D3 Spacing	—	—	7′
Director 3	—	—	34′
Feed method	Direct 50 Ω	Direct 50 Ω	Direct 50 Ω

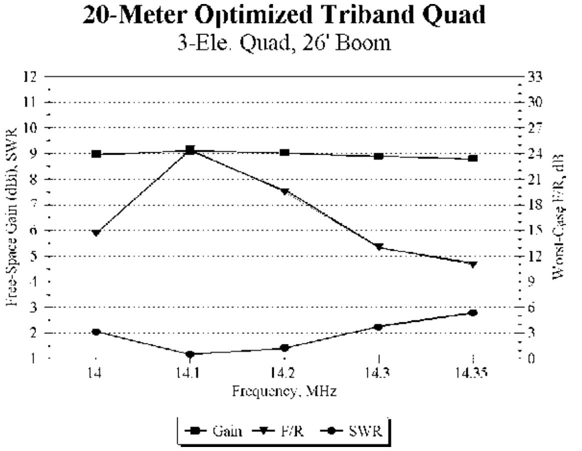

Fig 9—Computed performance of the triband, five-element quad over the 20-meter band. The direct 50-Ω feed system holds the SWR below 2.8: 1 across the whole band. This could be improved with a gamma-match system tuned to 14.1 MHz if the builder really desires a low SWR. The F/R peaks at 14.1 MHz and remains above 10 dB across the whole band.

lines, one per driven element, together with a relay switchbox mounted to the boom so that a single coax can be used back to the operating position. Each feed line uses a ferrite-bead balun to control common-mode currents and preserve the radiation pattern and each coax going to the switchbox is cut to be an electrical three-quarter wavelength on 15 meters. This presents a short at the unused driven elements since modeling indicated that the 15-meter band is adversely affected by the presence of the 20-meter driven element if it is left open-circuited. If you use RG-213 coax, the $^{3}/_{4}$-λ electrical length of each feed line is 23 feet long at 21.2 MHz. This is sufficient physical length to reach each driven element from the switchbox.

Fig 10 shows the free-space response for the 15-meter band. The rearward response is roughly 15 dB across the band. This is a result of the residual interaction between the 20-meter elements on 15 meters, and no further tuning could improve the F/R. Note how flat the SWR curve is. This SWR characteristic is what gives the quad the reputation of being "wideband." A flat SWR curve, however, is not necessarily a good indicator of optimal performance for directional antennas like quads or Yagis, particularly multiband designs where compromises must be made by physical necessity.

Fig 11 shows the characteristics of the 10-meter portion of the two-element triband quad. The response favors the low-phone band, with the F/R falling to about 12 dB at the low end of the frequency range and rising to just about 23 dB at 28.4 MHz. The SWR curve is once again relatively flat across the major portion of the band up to 28.8 MHz.

Construction

The most obvious problem related to quad antennas is the ability to build a structurally sound system. If high winds or heavy ice are a normal part of the environment, special precautions are necessary if the antenna is to survive a winter season. Another stumbling block for would-be quad builders is the installation of a three-dimensional system (assuming a Yagi has only two important dimensions) on top of a tower—especially if the tower needs guy wires for support. With proper planning, however, many of these obstacles can be overcome. For example, a tram system may be used.

Both multiband quad arrays use fiberglass spreaders (see Chapter 21 for suppliers). Bamboo is a suitable substitute (if economy is of great importance). However, the additional weight of the bamboo spreaders over fiberglass is an important consideration. A typical 12-foot bamboo pole weighs about 2 pounds; the fiberglass type weighs less than a pound. By multiplying the difference times 8 for a two-element array, times 12 for a three-element antenna, and so on, it quickly becomes apparent that fiberglass is worth the investment if weight is an important factor. Properly treated, bamboo has a useful life of three or four years, while fiberglass life is probably 10 times longer.

Spreader supports (sometimes called *spiders*) are available from many different manufacturers. If the builder is keeping the cost at a minimum, he should consider building his own. The expense is about half that of a commercially manufactured equivalent and, according to some authorities, the homemade arm supports described below are less likely to rotate on the boom as a result of wind pressure.

A 3-foot length of steel angle stock, 1 inch per side, is used to interconnect the pairs of spreader arms. The steel is drilled at the center to accept a muffler clamp of sufficient size to clamp the assembly to the boom. The fiberglass is clamped to the steel angle stock with auto-

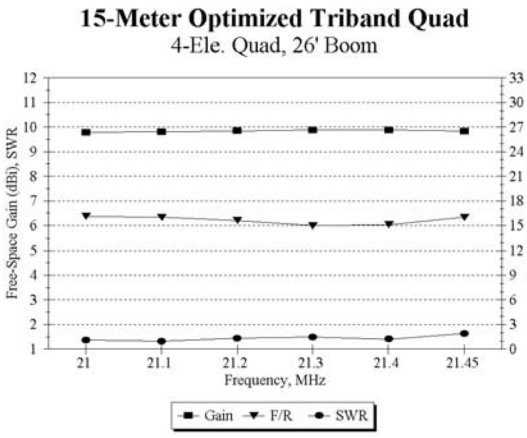

Fig 10—Computed performance of the triband, five-element quad over the 15-meter band. There is some degree of interaction with the 20-meter elements, limiting the worst-case F/R to about 15 dB. The gain and SWR curves are relatively flat across the band.

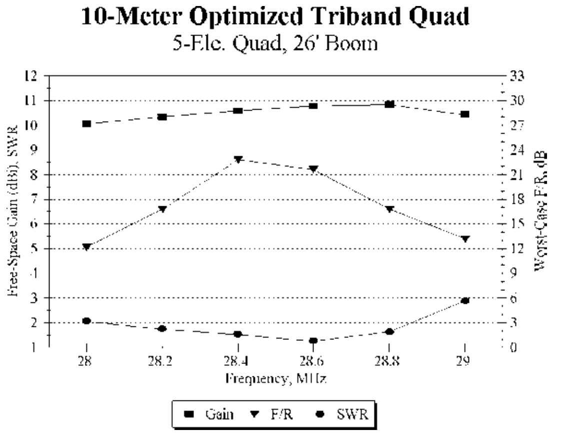

Fig 11—Computed performance of the triband, five-element quad over the 10-meter band. The F/R is higher than 12 dB across the band from 28.0 to 29.0 MHz, but the SWR rises at the top end of the band beyond 2:1. The free-space gain is higher than 10 dBi across the band.

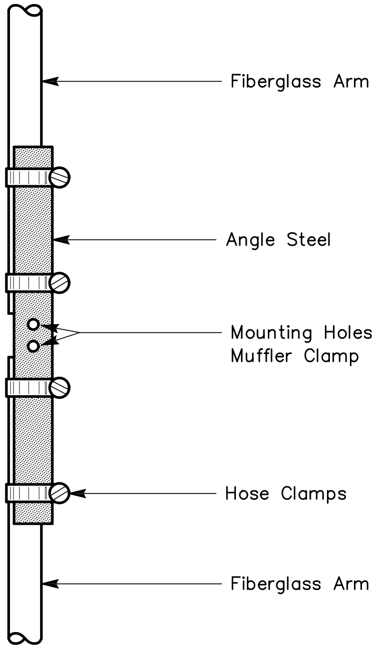
Fiberglass Arm
Angle Steel
Mounting Holes for
Muffler Clamp
Hose Clamps
Fiberglass Arm

Fig 12—Details of one of two assemblies for a spreader frame. The two assemblies are joined back-to-back to form an X with a muffler clamp mounted at the position shown.

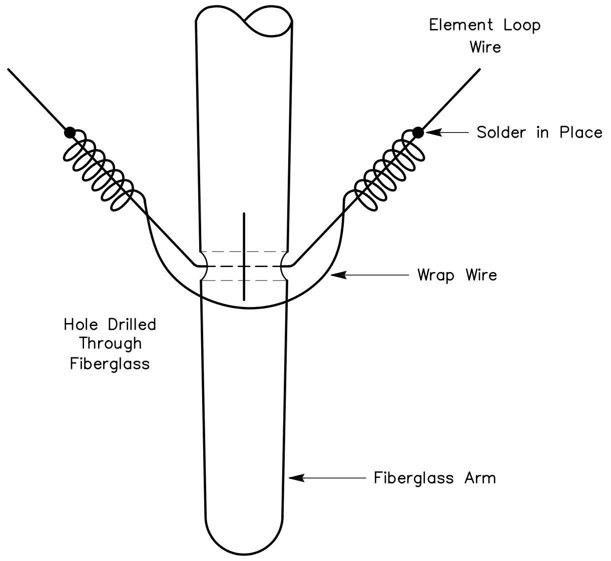
Element Loop
Wire
Solder in Place
Wrap Wire
Hole Drilled
Through
Fiberglass
Fiberglass Arm

Fig 13—A method of assembling a corner of the wire loop of a quad element to the spreader arm.

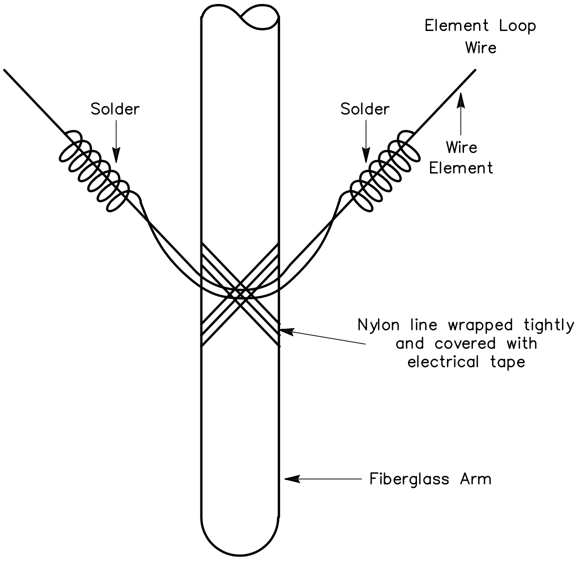
Element Loop
Wire
Solder
Solder
Wire
Element
Nylon line wrapped tightly
and covered with
electrical tape
Fiberglass Arm

Fig 14—An alternative method of assembling the wire of a quad loop to the spreader arm.

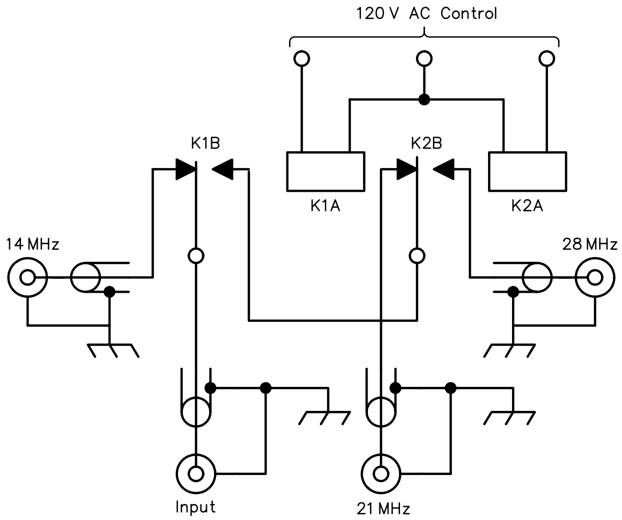
120 V AC Control
K1B
K2B
K1A
K2A
14 MHz
28 MHz
Input
21 MHz

Fig 15—Suitable circuit for relay switching of bands for the three-band quad. A three-wire control cable is required. K1, K2—any type of relay suitable for RF switching, coaxial type not required (Potter and Brumfeld MR11A acceptable; although this type has double-pole contacts, mechanical arrangements of most single-pole relays make them unacceptable for switching of RF).

motive hose clamps, two per pole. Each quad-loop spreader frame consists of two assemblies of the type shown in **Fig 12**.

Connecting the wires to the fiberglass can be done in a number of different ways. Holes can be drilled at the proper places on the spreader arms and the wires run through them. A separate wrap wire should be included at the entry/exit point to prevent the loop from slipping. Details are presented in **Fig 13**. Some amateurs have experienced cracking of the fiberglass, which might be a result of drilling holes through the material. However, this seems to be the exception rather than the rule. The model described here has no holes in the spreader arms; the wires are attached to each arm with a few layers of

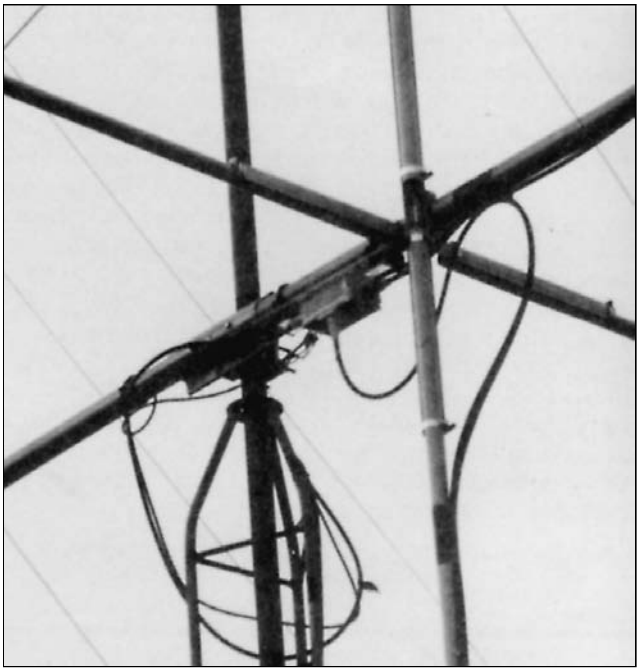

Fig 16—The relay box is mounted on the boom near the center. Each of the spreader-arm fiberglass poles is attached to steel angle stock with hose clamps.

plastic electrical tape and then wrapped approximately 20 times in a crisscross fashion with $^1/_8$-inch diameter nylon string, followed by more electrical tape for UV protection, as shown in **Fig 14**.

The wire loops are left open at the bottom of each driven element where the feed-line coaxes are attached. All of the parasitic elements are continuous loops of wire; the solder joint is at the base of the diamond.

Although you could run three separate coax cables down to the shack, we suggest that you install a relay box at the center of the boom. A three-wire control system may be used to apply power to the proper relay for changing bands. The circuit diagram of a typical configuration is presented in **Fig 15** and its installation is shown in **Fig 16**.

Every effort must be placed upon proper construction if you want to have freedom from mechanical problems. Hardware must be secure or vibration created by the wind may cause separation of assemblies. Solder joints should be clamped in place to keep them from flexing, which might fracture a connection point.

A TWO-ELEMENT, 8-FOOT BOOM PENTABAND QUAD

This two-element pentaband (20/17/15/12/10-meter) quad uses the same construction techniques as its big brother above. Since only two elements are used, the boom can be less robust for this antenna, at 2 inches diameter rather than 3 inches. Those who like really rugged antennas can still use the 3-inch diameter boom, of course.

Table 3 lists the element dimensions for the pentaband quad. The following plots show the performance for each of the five bands covered. The feed system for the pentaband quad uses five, direct 50-Ω coaxes, one to each driven element. These five coaxes are cut to be $^3/_4$-λ electrically on 10 meters (17 feet, 2 inches for RG-213 at 28.4 MHz). In this design the 10-meter band is the one most affected by the presence of the other driven elements if they are left unshorted. The $^3/_4$-λ lines open-circuited at the switchbox are long enough physically to reach all elements from a centrally mounted switchbox. This length assumes that the switchbox open-circuits the unused coaxes. If the switchbox short-circuits unused coaxes (as several commercial switchboxes do), then use $^1/_2$-λ long lines to feed all five driven elements (11 feet,

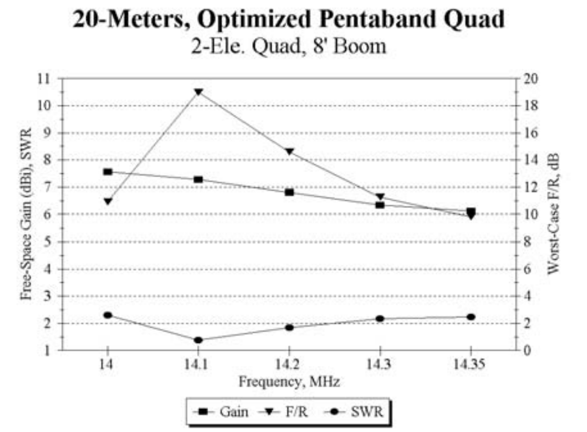

Fig 17—Computed performance of the pentaband two-element quad on 20 meters. With the simple direct-feed system, the SWR rises to about 2.3:1 at the low end of the band. A gamma match can bring the SWR down to 1:1 at 14.1 MHz, if desired.

Table 3
Five-Band Two-Element Quad on 8-Foot Boom

	14.2 MHz	*18.1 MHz*	*21 MHz*	*24.9 MHz*	*28.4 MHz*
Reflector	72′ 4"	56′ 4"	48′ 6"	40′ 11$^1/_4$"	37′ 5$^1/_2$"
R-DE Spacing	8′	8′	8′	8′	8′
Driven Element	69′ 10$^1/_2$"	54′ 10$^1/_2$"	46′ 7"	39′ 10$^1/_2$"	34′ 6"

5 inches for RG-213 at 28.4 MHz).

The SWR curves do not necessarily go down to 1:1 because of this simple, direct feed system. If anyone is bothered by this, of course they can always implement individual matching systems, such as gamma matches. Most amateurs would agree that such a degree of complexity is not warranted. The worst-case SWR is less than 2.3:1 on each band, even with direct feed on 20 meters. With typical lengths of coaxial feed line from the shack to the switchbox at the antenna, say 100 feet of RG-213, the SWR at the transmitter would be less than 2.0:1 on all bands due to losses in the feed line.

Fig 17 shows the computed responses for the pentaband quad over the 20-meter band. With only two degrees of freedom (spacing and element tuning) there is not much that can be done to spread the response out over the entire 20-meter band. Nonetheless, the performance over the band is still pretty reasonable for an antenna this small. The F/R pattern peaks at 19 dB at 14.1 MHz and falls to about 10 dB at either end of the band. The free-space gain varies from about 7.5 dBi to just above 6 dBi, comparable to a short-boom three-element Yagi. The SWR curve remains below 2.3:1 across the band. If you were to employ a gamma match tuned at 14.1 MHz, you could limit the peak SWR to less than 2.0:1, and this would still occur at 14.0 MHz.

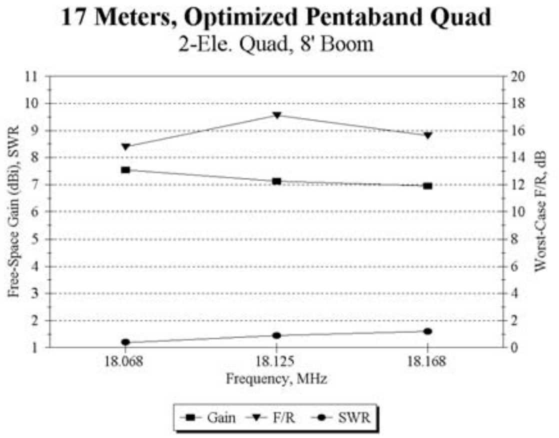

Fig 18—Computed performance of the pentaband two-element quad on 17 meters. There is some interaction with the other elements, but overall the performance is satisfactory on this band.

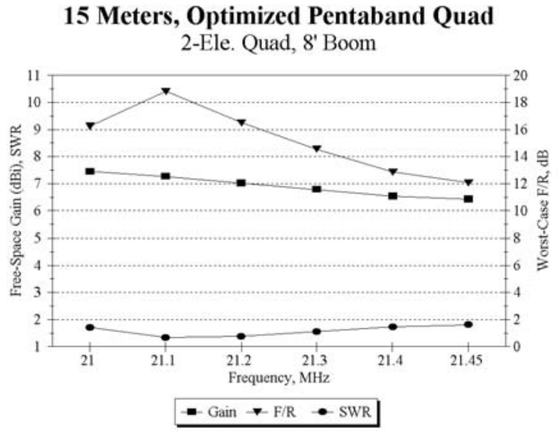

Fig 19—Computed performance of the pentaband two-element quad on 15 meters. The performance is acceptable across the whole band.

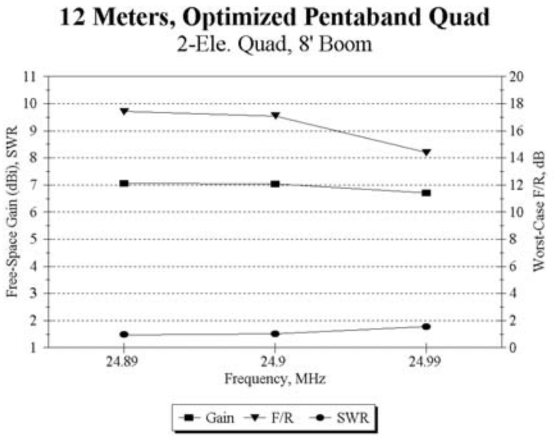

Fig 20—Computed performance of the pentaband two-element quad on 12 meters.

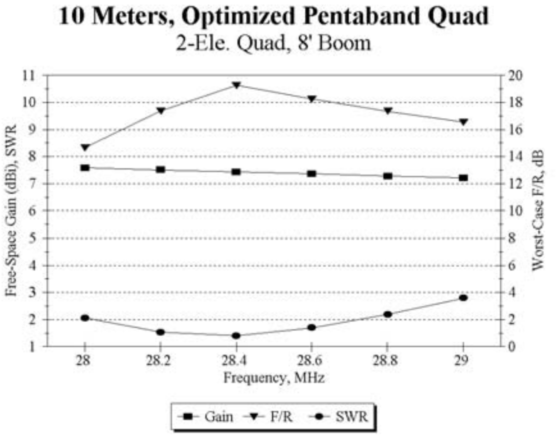

Fig 21—Computed performance of the pentaband two-element quad on 10 meters. The SWR curve is slightly above the target 2:1 at the low end of the band and rises to about 2.2:1 at 28.8 MHz. This unlikely to be a problem, even with rigs with automatic power-reduction due to SWR, since the SWR at the input of a typical coax feed line will be lower than that at the antenna due to losses in the line.

On 17 meters, **Fig 18** shows that the other elements are affecting 18 MHz, even with element-length optimization. Careful examination of the current induced on the other elements shows that the 20-meter driven element is interacting on 18 MHz, deteriorating the pattern and gain slightly. Even still, the performance on 17 meters is reasonable, especially for a five-band quad on an 8-foot boom.

On 15 meters, the interactions seems to have been contained, as **Fig 19** demonstrates. The F/R peaks at 21.1 MHz, at 19 dB and remains better than 12 dB past the top of the band. The SWR curve is low across the whole band.

On 12 meters, the interaction between bands is minor, leading to the good results shown in **Fig 20**. The SWR change across this band is quite flat, which isn't surprising given the narrow bandwidth of the 12-meter band.

On 10 meters, the interaction seems to have been tamed well by computer-tuning of the elements. The F/R remains higher than about 14 dB from 28 to 29 MHz. The SWR remains below 2.2:1 up to about 28.8 MHz, while the gain is relatively flat across the band at more than 7.2 dBi in free space. See **Fig 21**.

Overall, this pentaband quad is physically compact and yet it provides good performance across all five bands. It is competitive with commercial Log Periodic Dipole Array (LPDA) designs and triband Yagi designs that employ longer booms.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

P. S. Carter, C. W. Hansell and N. E. Lindenblad, "Development of Directive Transmitting Antennas by R.C.A. Communications," *Proc. IRE*, Oct 1931.

C. Cleveland, "More'n One Way to Switch an Antenna," Technical Correspondence, *QST*, Nov 1986.

D. Cutter, "Simple Switcher," *73*, May 1980.

D. DeMaw, "A Remote Antenna Switcher for HF," *QST*, Jun 1986.

M. G. Knitter, Ed., *Loop Antennas—Design and Theory* (Cambridge, WI: National Radio Club, 1983).

H. Landskov, W7KAR, "Evolution of a Quad Array," *QST*, Mar 1977.

J. Lindsay, "Quads and Yagis," *QST*, May 1968.

F. E. Terman, *Radio Engineering*, 3rd ed. (New York: McGraw-Hill Book Co, 1947).

E. M. Williams, "Radiating Characteristics of Short-Wave Loop Aerials," *Proc. IRE*, Oct 1940.

Chapter 13

Long-Wire and Traveling-Wave Antennas

The power gain and directive characteristics of electrically long wires (that is, wires that are long in terms of wavelength), as described in Chapter 2, make them useful for long-distance transmission and reception on the higher frequencies. Long wires can be combined to form antennas of various shapes that increase the gain and directivity over a single wire. The term *long wire*, as used in this chapter, means any such configuration, not just a straight-wire antenna.

Long Wires Versus Multielement Arrays

In general, the gain obtained with long-wire antennas is not as great, when the space available for the antenna is limited, as you can obtain from the multielement phased arrays in Chapter 8 or from a parasitic array such as a Yagi or quad (Chapters 11 or 12). However, the long-wire antenna has advantages of its own that tend to compensate for this deficiency. The construction of long-wire antennas is simple both electrically and mechanically, and there are no especially critical dimensions or adjustments. The long-wire antenna will work well and give satisfactory gain and directivity over a 2-to-1 frequency range. In addition, it will accept power and radiate well on any frequency for which its overall length is not less than about a half wavelength. Since a wire is not electrically long, even at 28 MHz, unless its physical length is equal to at least a half wavelength on 3.5 MHz, any long-wire can be used on all amateur bands that are useful for long-distance communication.

Between two directive antennas having the same theoretical gain, one a multielement array and the other a long-wire antenna, many amateurs have found that the long-wire antenna seems more effective in reception. One possible explanation is that there is a *diversity effect* with a long-wire antenna because it is spread out over a large distance, rather than being concentrated in a small space, as would be the case with a Yagi, for example. This may raise the average level of received energy for ionospheric-propagated signals. Another factor is that long-wire antennas have directive patterns that can be extremely sharp in the horizontal (azimuthal) plane. This is an advantage that other types of multielement arrays do not have, but it can be a double-edged sword too. We'll discuss this aspect in some detail in this chapter.

General Characteristics of Long-Wire Antennas

Whether the long-wire antenna is a single wire running in one direction or is formed into a V-beam, rhombic, or some other configuration, there are certain general principles that apply and some performance features that are common to all types. The first of these is that the power gain of a long-wire antenna as compared with a half-wave dipole is not considerable until the antenna is really long (its length measured in wavelengths rather than in a specific number of feet). The reason for this is that the fields radiated by elementary lengths of wire along the antenna do not combine, at a distance, in as simple a fashion as the fields from half-wave dipoles used in other types of directive arrays.

There is no point in space, for example, where the distant fields from all points along the wire are exactly in phase (as they are, in the optimum direction, in the case of two or more collinear or broadside dipoles when fed with in-phase currents). Consequently, the field strength at a distance is always less than would be obtained if the same length of wire were cut up into properly phased and separately driven dipoles. As the wire is made longer, the fields combine to form increasingly intense main lobes, but these lobes do not develop appreciably until the wire is several wavelengths long. See **Fig 1**.

The longer the antenna, the sharper the lobes become, and since it is really a hollow cone of radiation about the wire in free space, it becomes sharper in both planes. Also, the greater the length, the smaller the angle with the wire at which the maximum radiation lobes occur. There are four main lobes to the directive patterns of long-wire antennas; each makes the same angle with respect to the wire.

Fig 2A shows the azimuthal radiation pattern of a 1-λ long-wire antenna, compared with a $^1/_2$-λ dipole. Both antennas are mounted at the same height of 1 λ above flat ground (70 feet high at 14 MHz, with a wire length of 70 feet) and both patterns are for an elevation angle of 10°, an angle suitable for long-distance communication on 20 meters. The long-wire in Fig 2A is oriented in the 270° to 90° direction, while the dipole is aligned at right angles so that its characteristic figure-8 pattern goes left-to-right. The 1-λ long-wire has about 0.6 dB more gain than the dipole, with four main lobes as compared to the two lobes from the dipole.

You can see that the two lobes on the left side of Fig 2A are about 1 dB down compared to the two lobes on the right side. This is because the long-wire here is fed at the left-hand end in the computer model. Energy is radiated as a wave travels down the wire and some energy is also lost to ohmic resistance in the wire and the ground. The forward-going wave then reflects from the open-circuit at the right-hand end of the wire and reverses direction, traveling toward the left end, still radiating as it

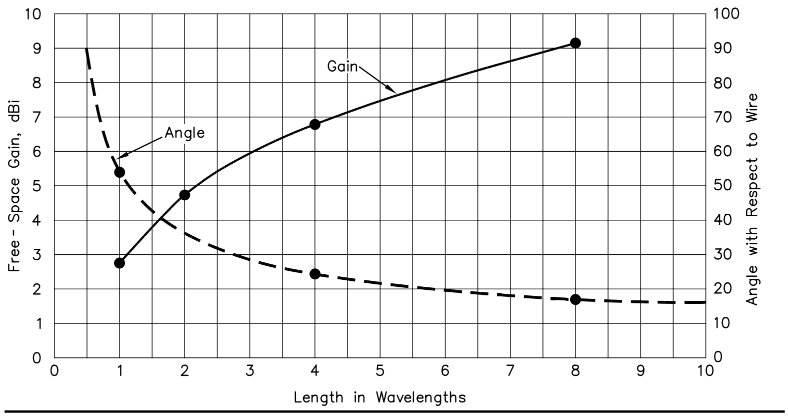

Fig 1—Theoretical gain of a long-wire antenna, in dBi, as a function of wire length. The angle, with respect to the wire, at which the radiation intensity is maximum also is shown.

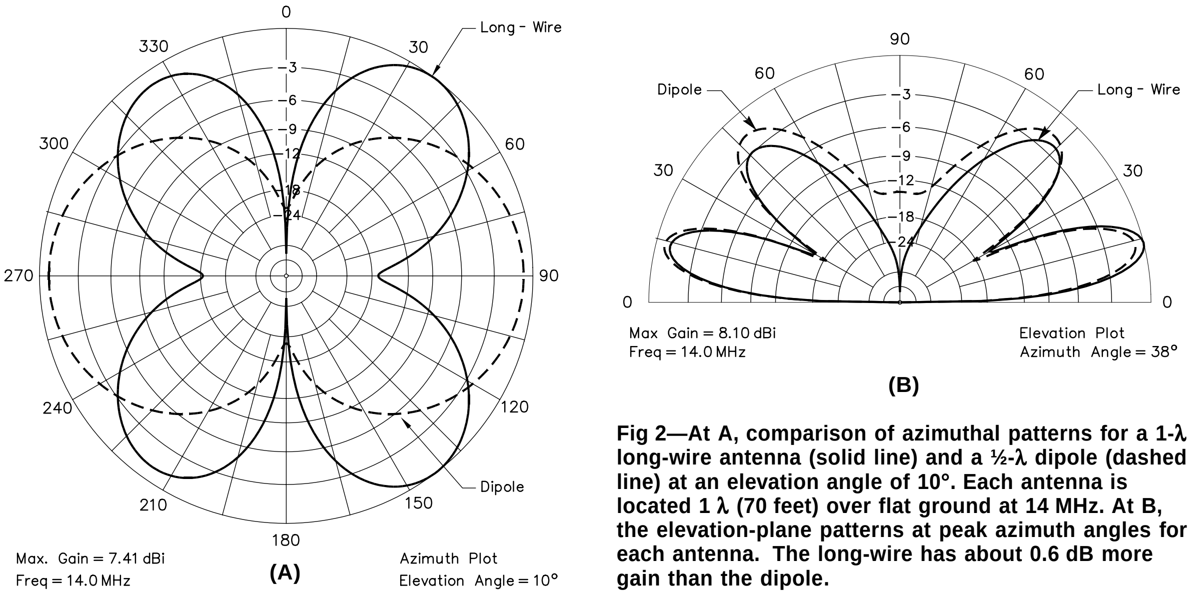

Fig 2—At A, comparison of azimuthal patterns for a 1-λ long-wire antenna (solid line) and a ½-λ dipole (dashed line) at an elevation angle of 10°. Each antenna is located 1 λ (70 feet) over flat ground at 14 MHz. At B, the elevation-plane patterns at peak azimuth angles for each antenna. The long-wire has about 0.6 dB more gain than the dipole.

travels. An antenna operating in this way has much the same characteristics as a transmission line that is terminated in an open circuit—that is, it has standing waves on it. Unterminated long-wire antennas are often referred to as *standing wave antennas*. As the length of a long-wire antenna is increased, a moderate front-to-back ratio results, about 3 dB for very long antennas.

Fig 2B shows the elevation-plane pattern for the long-wire and for the dipole. In each case the elevation pattern is at the azimuth of maximum gain—at an angle of 38° with respect to the wire-axis for the long-wire and at 90° for the dipole. The peak elevation for the long-wire is very slightly lower than that for the dipole at the same height above ground, but not by much. In other words, the height above ground is the main determining factor for the shape of the main lobe of a long-wire's elevation pattern, as it is for most horizontally polarized antennas.

The shape of the azimuth and elevation patterns in Fig 2 might lead you to believe that the radiation pattern is simple. **Fig 3** is a 3-D representation of the pattern from a 1-λ long-wire that is 1 λ high over flat ground. Besides the main low-angle lobes, there are strong lobes at higher angles. Things get even more complicated when the length of the long-wire increases.

Directivity

Because many points along a long wire are carrying currents in different phases (with different current amplitudes as well), the field pattern at a distance becomes more complex as the wire is made longer. This complexity is manifested in a series of minor lobes, the number of which increases with the wire length. The intensity of radiation from the minor lobes is frequently as great as, and sometimes greater than, the radiation from a half-wave dipole. The energy radiated in the minor lobes is not available to improve the gain in the major lobes, which is another reason why a long-wire antenna must be long to give appreciable gain in the desired directions.

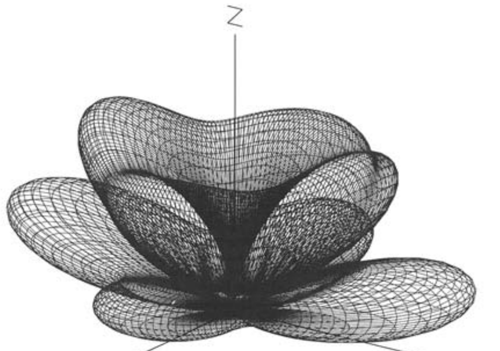

Fig 3—A 3-D representation of the radiation pattern for the 1-λ long-wire shown in Fig 2. The pattern is obviously rather complex. It gets even more complicated for wires longer than 1 λ.

Fig 4 shows an azimuthal-plane comparison between a 3-λ (209 feet long) long-wire and the comparison $^1/_2$-λ dipole. The long-wire now has 8 minor lobes besides the four main lobes. Note that the angle the main lobes make with respect to the axis of the long-wire (also left-to-right in Fig 4) becomes smaller as the length of the long-wire increases. For the 3-λ long-wire, the main lobes occur 28° off the axis of the wire itself.

Other types of simple driven and parasitic arrays do not have minor lobes of any great consequence. For that reason they frequently seem to have much better directivity than long-wire antennas, because their responses in undesired directions are well down from their response in the desired direction. This is the case even if a multielement array and a long-wire antenna have the same peak gain in the favored direction. **Fig 5** compares the same 3-λ long-wire with a 4-element Yagi and a ½-λ dipole, again both at the same height as the long-wire. Note that the Yagi has only a single backlobe, down about 21 dB from its broad main lobe, which has a 3-dB beamwidth of 63°. The 3-dB beamwidth of the long-wire's main lobes (at a 28° angle from the wire axis) is far more narrow, at only 23°.

For amateur work, particularly with directive antennas that cannot be rotated, the minor lobes of a long-wire antenna have some advantages. Although the nulls in the

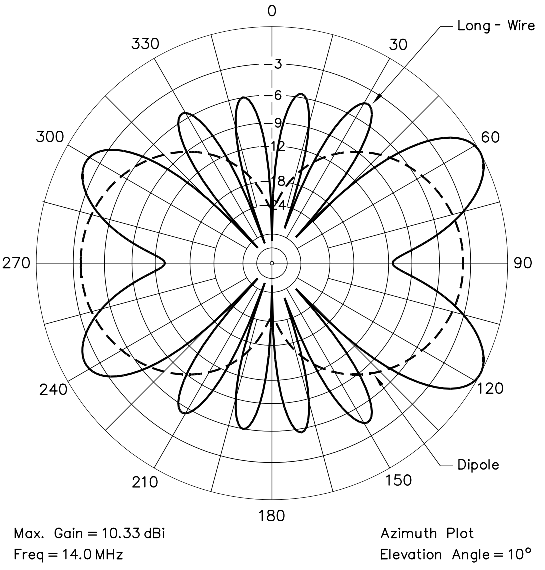

Fig 4—An azimuthal-plane comparison between a 3-λ (209 feet long) long-wire (solid line) and the comparison ½-λ dipole (dashed line) at 70 feet high (1 λ) at 14 MHz.

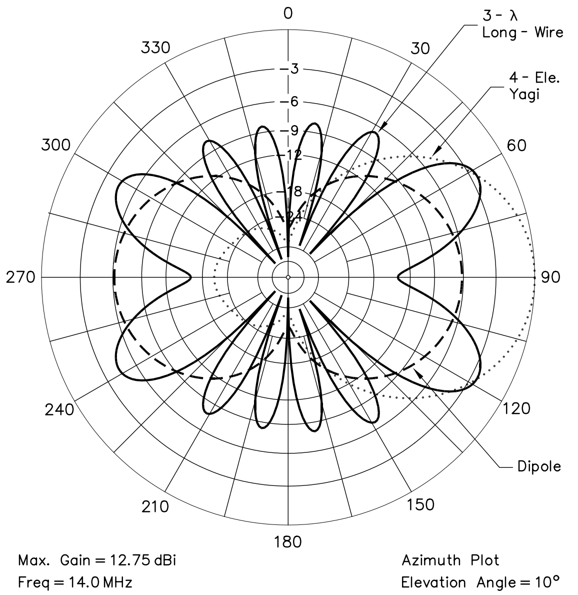

Fig 5—A comparison between the 3-λ long-wire (solid line) in Fig 4, a 4-element 20-meter Yagi on a 26-foot boom (dotted line), and a ½-λ dipole (dashed line), again at a height of 70 feet. The main lobes of the long-wire are very narrow compared to the wide frontal lobe of the Yagi. The long-wire exhibits an azimuthal pattern that is more omnidirectional in nature than a Yagi, particularly when the narrow, deep nulls in the long-wire's pattern are filled-in due to irregularities in the terrain under its long span of wire.

computer model in Fig 5 are deeper than 30 dB, they are not so dramatic in actual practice. This is due to irregularities in the terrain that inevitably occur under the span of a long wire. In most directions the long-wire antenna will be as good as a half-wave dipole, and in addition will give high gain in the most favored directions, even though that is over narrow azimuths.

Fig 6A compares the azimuth responses for a 5-λ long-wire (350 feet long at 14 MHz) to the same 4-element Yagi and dipole. The long-wire now exhibits 16 minor lobes in addition to its four main lobes. The peaks of these sidelobes are down about 8 dB from the main lobes and they are stronger than the dipole, making this long-wire antenna effectively omnidirectional. Fig 6B shows the elevation pattern of the 5-λ long-wire at its most effective azimuth compared to a dipole. Again, the shape of the main lobe is mainly determined by the long-wire's height above ground, since the peak angle is only just a bit lower than the peak angle for the dipole. The long-wire's elevation response breaks up into numerous lobes above the main lobes, just as it does in the azimuth plane.

For the really ambitious, **Fig 7** compares the performance for an 8-λ (571 feet) long-wire antenna with a 4-element Yagi and the $^1/_2$-λ dipole. Again, in actual practice, the nulls would tend to be filled in by terrain irregularities, so a very long antenna like this would be a pretty potent performer.

Calculating Length

In this chapter, lengths are discussed in terms of wave-

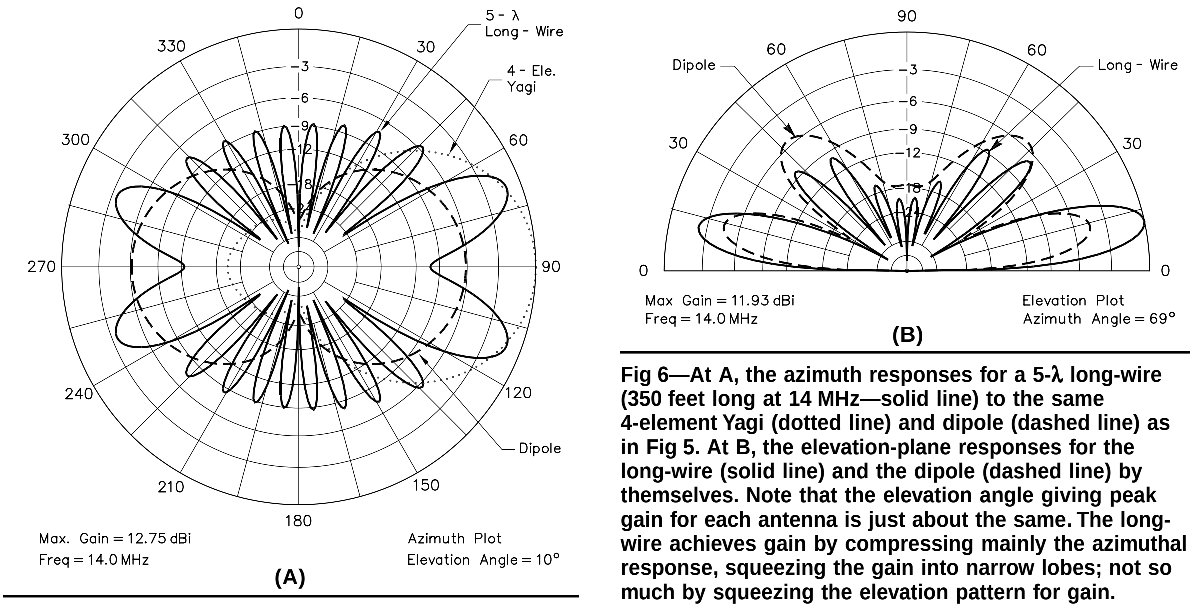

(A)

(B)

Fig 6—At A, the azimuth responses for a 5-λ long-wire (350 feet long at 14 MHz—solid line) to the same 4-element Yagi (dotted line) and dipole (dashed line) as in Fig 5. At B, the elevation-plane responses for the long-wire (solid line) and the dipole (dashed line) by themselves. Note that the elevation angle giving peak gain for each antenna is just about the same. The long-wire achieves gain by compressing mainly the azimuthal response, squeezing the gain into narrow lobes; not so much by squeezing the elevation pattern for gain.

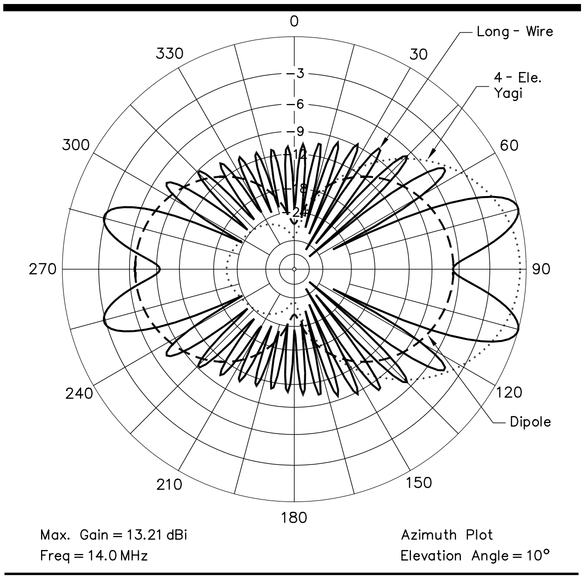

Fig 7—The azimuthal-plane performance for an 8-λ (571 feet) long-wire antenna (solid line), compared with a 4-element Yagi (dotted line) and a ½-λ dipole (dashed line).

lengths. Throughout the preceding discussion the frequency in the models was held at 14 MHz. Remember that a long-wire that is 4 λ long at 14 MHz is 8 λ long at 28 MHz.

There is nothing very critical about wire lengths in an antenna system that will work over a frequency range including several amateur bands. The antenna characteristics change very slowly with length, except when the wires are short (around one wavelength, for instance). There is no need to try to establish exact resonance at a particular frequency for proper antenna operation.

The formula for determining the lengths for harmonic wires is:

$$\text{Length}\left(\text{feet}\right)=\frac{984\left(N-0.025\right)}{f\left(\text{MHz}\right)} \qquad \text{(Eq 1)}$$

where N is the antenna length in wavelengths. In cases where precise resonance is desired for some reason (for obtaining a resistive load for a transmission line at a particular frequency, for example) it is best established by trimming the wire length until the standing-wave ratio on the line is minimum.

Tilted Wires

In theory, it is possible to maximize gain from a long-wire antenna by tilting it to favor a desired elevation takeoff angle. Unfortunately, the effect of real ground under the antenna negates the possible advantages of tilting, just as it does when a Yagi or other type of parasitic array is tilted from horizontal. You would do better keeping a long-wire antenna horizontal, but raising it higher above ground, to achieve more gain at low takeoff angles.

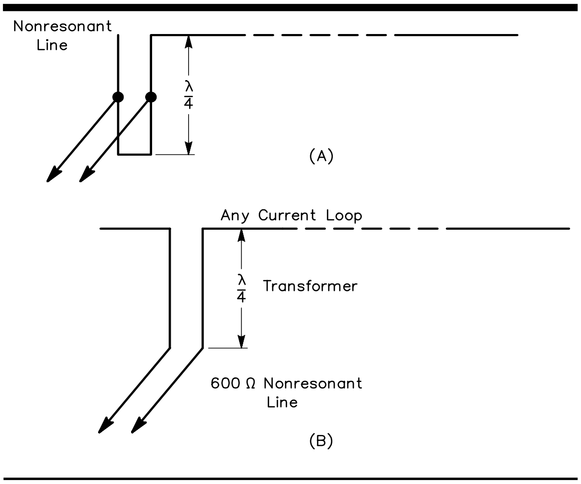

Fig 8—Methods for feeding long single-wire antennas.

Feeding Long Wires

A long-wire antenna is normally fed at the end or at a current loop. Since a current loop changes to a node when the antenna is operated at any even multiple of the frequency for which it is designed, a long-wire antenna will operate as a true long wire on all bands only when it is fed at the end.

A common method of feeding a long-wire is to use a resonant open-wire line. This system will work on all bands down to the one, if any, at which the antenna is only a half wave long. Any convenient line length can be used if you match the transmitter to the line's input impedance using an antenna tuner, as described in Chapter 25.

Two arrangements for using nonresonant lines are given in **Fig 8**. The one at A is useful for one band only since the matching section must be a quarter wave long, approximately, unless a different matching section is used for each band. In B, the λ/4 transformer (Q-section) impedance can be designed to match the antenna to the line, as described in Chapter 26. You can determine the value of radiation resistance using a modern modeling program or you can actually measure the feed-point impedance. Although it will work as designed on only one band, the antenna can be used on other bands by treating the line and matching transformer as a resonant line. In this case, as mentioned earlier, the antenna will not radiate as a true long wire on even multiples of the frequency for which the matching system is designed.

The end-fed arrangement, although the most convenient when tuned feeders are used, suffers the disadvantage that there is likely to be a considerable antenna current on the line. In addition, the antenna reactance changes rapidly with frequency. Consequently, when the wire is several wavelengths long, a relatively small change in frequency—a fraction of the width of a band—may require major changes in the adjustment of the antenna tuner. Also, the line becomes unbalanced at all frequencies between those at which the antenna is resonant. This leads to a considerable amount of radiation from the line. The unbalance can be overcome by using multiple long wires in a V or rhombic shape, as described below.

COMBINATIONS OF LONG WIRES

The directivity and gain of long wires may be increased by using two wires placed in relation to each other such that the fields from both combine to produce the greatest possible field strength at a distant point. The principle is similar to that used in designing the multi-element arrays described in Chapter 8.

Parallel Wires

One possible method of using two (or more) long wires is to place them in parallel, with a spacing of $^1/_2$ λ or so, and feed the two in phase. In the direction of the wires the fields will add in phase. However, the takeoff angle is high directly in the orientation of the wire, and this method will result in rather high-angle radiation even if the wires are several wavelengths long. With a parallel arrangement of this sort the gain should be about 3 dB over a single wire of the same length, at spacings in the vicinity of $^1/_2$ wavelength.

The V-Beam Antenna

Instead of using two long wires parallel to each other, they may be placed in the form of a horizontal V, with the included angle between the wires equal to twice the angle made by the main lobes referenced to the wire axis for a single wire of the same physical length. For example, for a leg length of 5 λ, the angle between the legs of a V should be about 42°, twice the angle of 21° of the main lobe referenced to the long-wire's axis. See Fig 6A.

The plane directive patterns of the individual wires combine along a line in the plane of the antenna and bisecting the V, where the fields from the individual wires reinforce each other. The sidelobes in the azimuthal pattern are suppressed by about 10 dB, so the pattern becomes essentially bidirectional. See **Fig 9**.

The included angle between the legs is not particularly critical. This is fortunate, especially if the same antenna is used on multiple bands, where the electrical length varies directly with frequency. This would normally require different included angles for each band. For multiband V-antennas, a compromise angle is usually chosen to equalize performance. **Fig 10** shows the azimuthal pattern for a V-beam with 1-λ legs, with an included angle of 75° between the legs, mounted 1 λ above flat ground. This is for a 10° elevation angle. At 14 MHz the antenna has two 70-foot high, 68.5-foot long legs, separated at their far ends by 83.4 feet. For comparison the azimuthal patterns for the same 4-element Yagi and ½-λ dipole used previously for the long-wires are overlaid on the same plot. The V has about 2 dB more gain than the dipole but is down some 4 dB compared to the Yagi, as expected for relatively short legs.

Fig 11 shows the azimuthal pattern for the same antenna in Fig 10, but at 28 MHz and at an elevation angle of 6°. Because the legs are twice as long electrically at 28 MHz, the V-beam has compressed the main lobe into a narrow beam that now has a peak gain equal to the Yagi, but with a 3-dB beamwidth of only 18.8°. Note that you could obtain about 0.7 dB more gain at 14 MHz, with a 1.7-dB degradation of gain at 28 MHz, if you increase the included angle to 90° rather than 75°.

Fig 12 shows the azimuthal pattern for a V-beam with 2-λ legs (137 feet at 14 MHz), with an included angle of 60° between them. As usual, the assumed height is 70 feet, or 1 λ at 14 MHz. The peak gain for the V-beam is just about equal to that of the 4-element Yagi, although the 3-dB

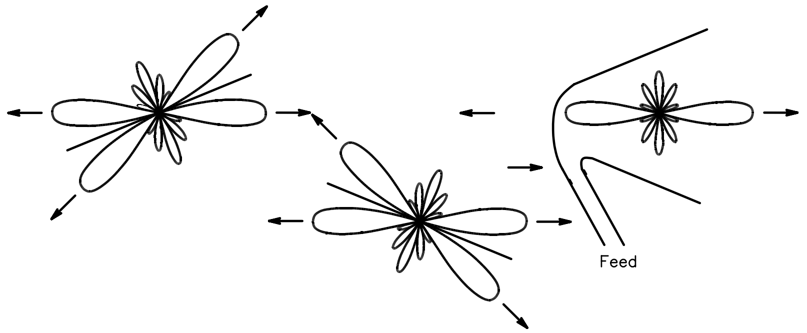

Fig 9—Two long wires and their respective patterns are shown at the left. If these two wires are combined to form a V with an angle that is twice that of the major lobes of the wires and with the wires excited out of phase, the radiation along the bisector of the V adds and the radiation in the other directions tends to cancel.

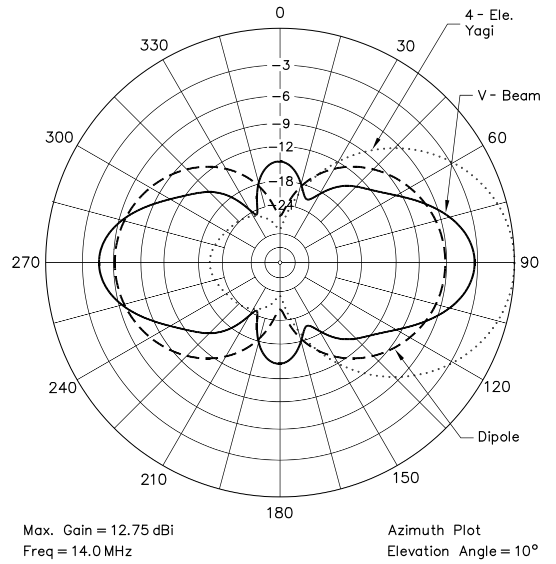

Fig 10— Azimuthal-plane pattern at 10° elevation angle for a 14-MHz V-beam (solid line) with 1-λ legs (68.5 feet long), using an included angle of 75° between the legs. The V-beam is mounted 1 λ above flat ground, and is compared with a ½-λ dipole (dashed line) and a 4-element 20-meter Yagi on a 26-foot boom (dotted line).

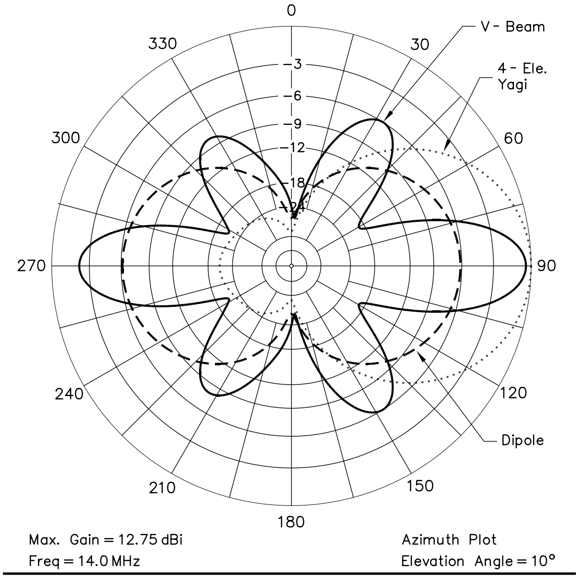

Fig 12—Azimuthal pattern for a V-beam (solid line) with 2-λ legs (137 feet at 14 MHz), with an included angle of 60° between them. The height is 70 feet, or 1 λ, over flat ground. For comparison, the response for a 4-element Yagi (dotted line) and a dipole (dashed line) are shown. The 3-dB beamwidth has decreased to 23.0°.

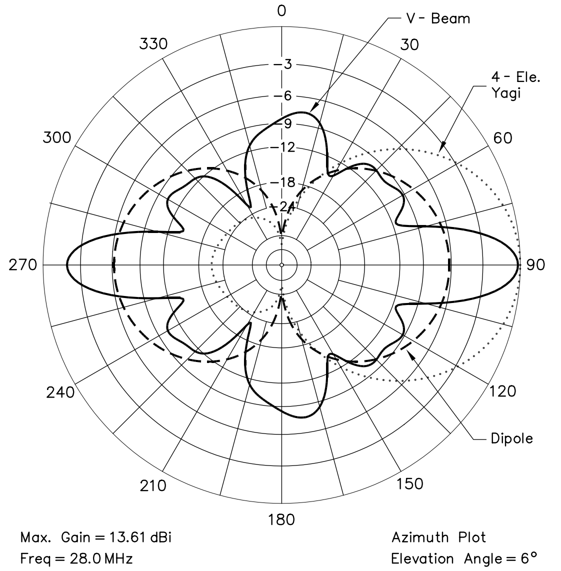

Fig 11—The same V-beam as in Fig 10 at 28 MHz (solid line), at an elevation angle of 6°, compared to a 4-element Yagi (dotted line) and a dipole (dashed line). The V-beam's pattern is very narrow, at 18.8° at the 3-dB points, requiring accurate placement of the supports poles to aim the antenna at the desired geographic target.

nose beamwidth is narrow, at 23°. This makes setting up the geometry critical if you want to maximize gain into a particular geographic area. While you might be able to get away with using convenient trees to support such an antenna, it's far more likely that you'll have to use carefully located towers to make sure the beam is aimed where you expect it to be pointed.

For example, in order to cover all of Europe from San Francisco, an antenna must cover from about 11° (to Moscow) to about 46° (to Portugal). This is a range of 35° and signals from the V-beam in Fig 12 would be down some 7 dB over this range of angles, assuming the center of the beam is pointed exactly at a heading of 28.5°. The 4-element Yagi on the other hand would cover this range of azimuths more consistently, since its 3-dB beamwidth is 63°.

Fig 13 shows the same V-beam as in Fig 12, but this time at 28 MHz. The peak gain of the main lobe is now about 1 dB stronger than the 4-element Yagi used as a reference, and the main lobe has two nearby sidelobes that tend to broaden out the azimuthal response. At this frequency the V-beam would cover all of Europe better from San Francisco.

Fig 14 shows a V-beam with 3-λ (209 feet at 14 MHz) legs with an included angle of 50° between them. The peak gain is now greater than that of a 4-element Yagi, but the 3-dB beamwidth has been reduced to 17.8°, making aim-

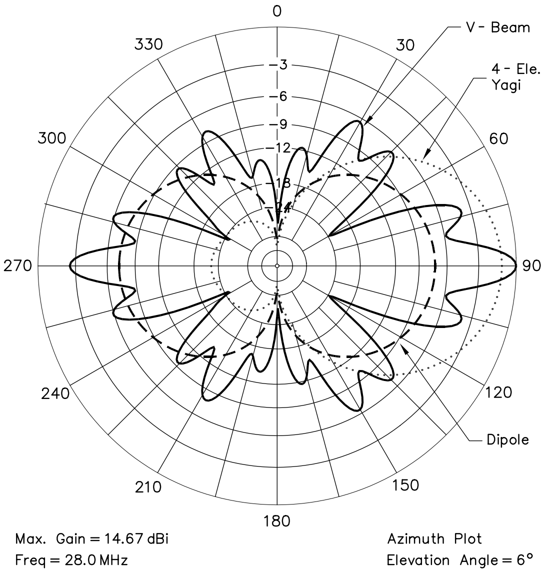

Fig 13—The same 2-λ per leg V-beam (solid line) as in Fig 12, but at 28 MHz and at a 6° takeoff elevation angle. Two sidelobes have appeared flanking the main lobe, making the effective azimuthal pattern wider at this frequency.

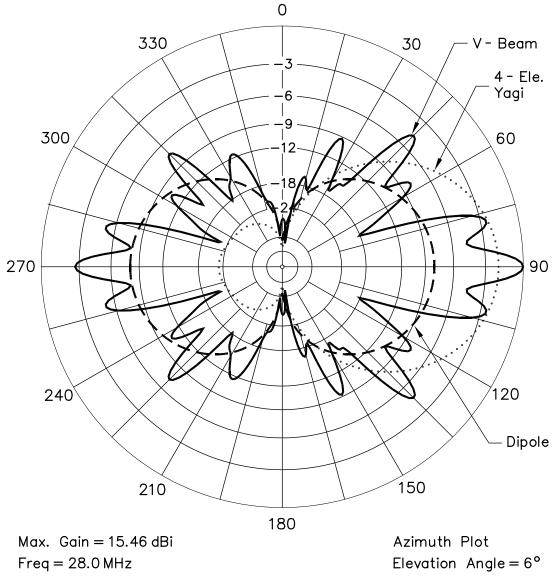

Fig 15—The same 209-foot/leg V-beam as Fig 14, but at 28 MHz. Again, the two close-in sidelobes tend to spread out the azimuthal response some at 28 MHz.

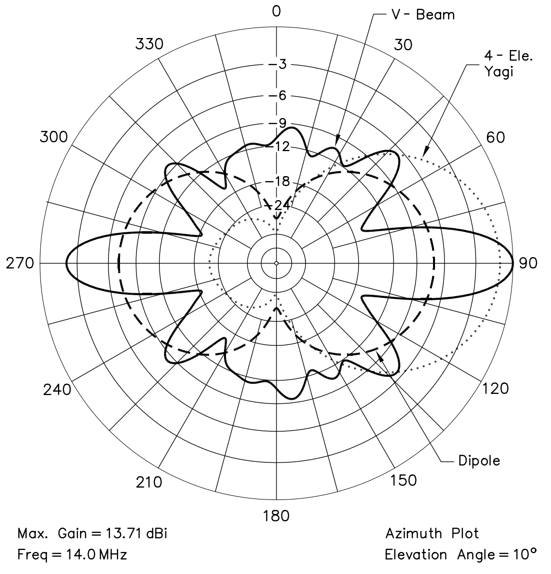

Fig 14—A V-beam (solid line) with 3-λ (209 feet at 14 MHz) legs using an included angle of 50° between them, compared to a 4-element Yagi (dotted line) and a dipole (dashed line). The 3-dB beamwidth has now decreased to 17.8°.

ing the antenna even more critical. **Fig 15** shows the same V-beam at 28 MHz. Here again, the main lobe has nearby sidelobes that broaden the effective azimuth to cover a wider area.

Fig 16 shows the elevation-plane response for the same 209-foot leg V-beam at 28 MHz (3-λ at 14 MHz), compared to a dipole at the same height of 70 feet. The higher-gain V-beam suppresses higher-angle lobes, essentially stealing energy from them and concentrating it in the main beam at 6° elevation.

The same antenna can be used at 3.5 and 7 MHz. The gain will not be large, however, because the legs are not very long at these frequencies. **Fig 17** compares the V-beam versus a horizontal ½-λ 40-meter dipole at 70 feet. At low elevation angles there is about 2 dB of advantage on 40 meters. **Fig 18** shows the same type of comparison for 80 meters, where the 80-meter dipole is superior at all angles.

Other V Combinations

A gain increase of about 3 dB can be had by stacking two V-beams one above the other, a half wavelength apart, and feeding them with in-phase currents. This will result in a lowered angle of radiation. The bottom V should be at least a quarter wavelength above the ground, and preferably a half wavelength. This arrangement will narrow the

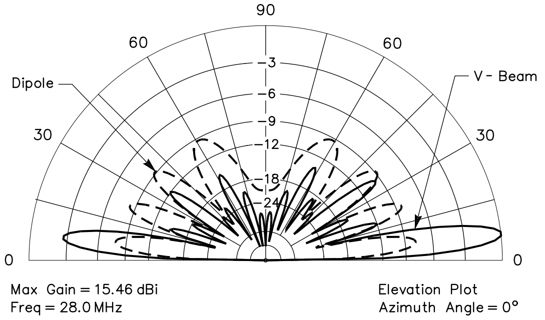

Fig 16—The elevation-plane of the 209-foot/leg V-beam (solid line) compared to the dipole (dashed line). Again, the elevation angle for peak gain corresponds well to that of the simple dipole at the same height.

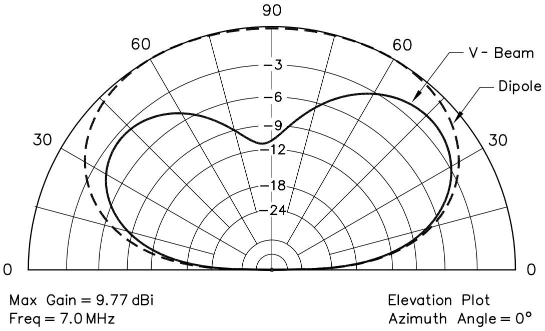

Fig 18—Elevation pattern for the same 209-foot-per-leg V-beam (solid line), at 3.5 MHz, compared to an 80-meter dipole at 70 feet (dashed line).

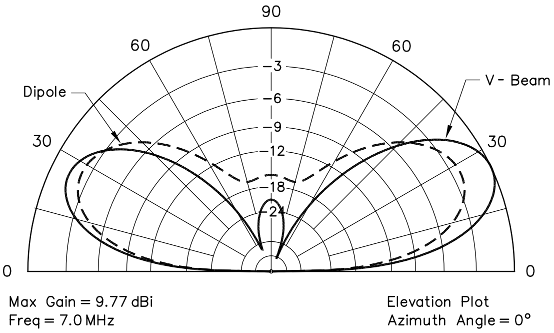

Fig 17—Elevation pattern for the same 209-foot-per-leg V-beam (solid line), at 7 MHz, compared to a 40-meter dipole (dashed line) at the same height of 70 feet.

elevation pattern and it will also have a narrow azimuthal pattern.

The V antenna can be made unidirectional by using a second V placed an odd multiple of a quarter wavelength in back of the first and exciting the two with a phase difference of 90º. The system will be unidirectional in the direction of the antenna with the lagging current. However, the V reflector is not normally employed by amateurs at low frequencies because it restricts the use to one band and requires a fairly elaborate supporting structure. Stacked Vs with driven reflectors could, however, be built for the 200- to 500-MHz region without much difficulty.

Feeding the V Beam

The V-beam antenna is most conveniently fed with tuned open-wire feeders with an antenna tuner, since this permits multiband operation. Although the length of the wires in a V-beam is not at all critical, it is important that both wires be the same electrical length. If a single band matching solution is desired, probably the most appropriate matching system is that using a stub or quarter-wave matching section. The adjustment of such a system is described in Chapter 26.

THE RESONANT RHOMBIC ANTENNA

The diamond-shaped or rhombic antenna shown in **Fig 19** can be looked upon as two acute-angle V-beams placed end-to-end. This arrangement is called a *resonant rhombic*. The leg lengths of the resonant rhombic must be an integral number of half wavelengths to avoid reactance at its feed point.

The resonant rhombic has two advantages over the simple V-beam. For the same total wire length it gives somewhat greater gain than the V-beam. A rhombic with 3 λ on a leg, for example, has about 1 dB gain over a V antenna with 6 wavelengths on a leg. **Fig 20** compares the azimuthal pattern at a 10° elevation for a resonant rhombic with 3 λ legs on 14 MHz, compared to a V-beam with 6 λ legs at the same height of 70 feet. The 3-dB nose beamwidth of the resonant rhombic is only 12.4° wide, but the gain is very high at 16.26 dBi.

The directional pattern of the rhombic is less frequency sensitive than the V when the antenna is used over a wide frequency range. This is because a change in frequency causes the major lobe from one leg to shift in one direction while the lobe from the opposite leg shifts the other way. This automatic compensation keeps the direction the same over a considerable frequency range. The disadvantage of the rhombic as compared with the V-beam is that an additional support is required.

The same factors that govern the design of the V-beam apply in the case of the resonant rhombic. The optimal apex angle A in Fig 19 is the same as that for a V having an equal leg length. The diamond-shaped antenna also can be operated as a terminated antenna, as described later in this chapter, and much of the discus-

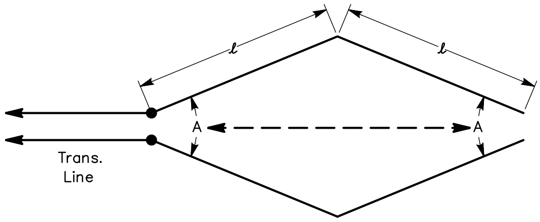

Fig 19—The resonant rhombic or diamond-shaped antenna. All legs are the same length, and opposite angles of the diamond are equal. Length ℓ is an integral number of half wavelengths for resonance.

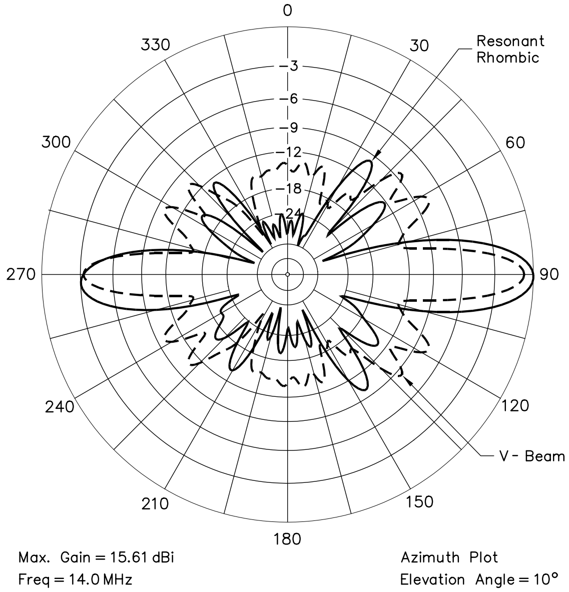

Fig 20—Azimuthal-plane pattern of resonant (unterminated) rhombic (solid line) with 3-λ legs on 14 MHz, at a height of 70 feet above flat ground, compared with a 6-λ per leg V-beam (dashed line) at the same height. Both azimuthal patterns are at a takeoff angle of 10°. The sidelobes for the resonant rhombic are suppressed to a greater degree than those for the V-beam.

sion in that section applies to the resonant rhombic as well.

The resonant rhombic has a bidirectional pattern, with minor lobes in other directions, their number and intensity depending on the leg length. In general, these sidelobes are suppressed better with a resonant rhombic than with a V-beam. When used at frequencies below the VHF region, the rhombic antenna is always mounted with the plane containing the wires horizontal. The polarization in this plane, and also in the perpendicular plane that bisects the rhombic, is horizontal. At 144 MHz and above, the dimensions are such that the antenna can be mounted with the plane containing the wires vertical if vertical polarization is desired.

When the rhombic antenna is to be used on several HF amateur bands, it is advisable to choose the apex angle, A, on the basis of the leg length in wavelengths at 14 MHz. Although the gain on higher frequency bands will not be quite as favorable as if the antenna had been designed for the higher frequencies, the system will still work well at the low angles that are necessary at such frequencies.

The resonant rhombic has lots of gain, but you must not forget that this gain comes from a radiation pattern that is very narrow. This requires careful placement of the supports for the resonant rhombic to cover desired geographic areas. This is definitely not an antenna that allows you to use just any convenient trees as supports!

The resonant rhombic antenna can be fed in the same way as the V-beam. Resonant feeders are necessary if the antenna is to be used in several amateur bands.

TERMINATED LONG-WIRE ANTENNAS

All the antenna systems considered so far in this chapter have been based on operation with standing waves of current and voltage along the wire. Although most hams use antenna designs based on using resonant wires, resonance is by no means a necessary condition for the wire to radiate and intercept electromagnetic waves efficiently, as discussed in Chapter 2. The result of using nonresonant wires is reactance at the feed point, unless the antenna is terminated with a resistive load.

In **Fig 21**, suppose that the wire is parallel with the ground (horizontal) and is terminated by a load Z equal to its characteristic impedance, Z_{ANT}. The wire and its image in the ground create a transmission line. The load Z can represent a receiver matched to the line. The *terminating resistor* R is also equal to the Z_{ANT} of the wire. A wave coming from direction X will strike the wire first at its far end and sweep across the wire at some angle until it reaches the end at which Z is connected. In so doing, it will induce voltages in the antenna, and currents will flow as a result. The current flowing toward Z is the useful output of the antenna, while the current flowing backwards toward R will be absorbed in R. The same thing is true of a wave coming from the direction X'. In such an antenna there are no standing waves, because all received power is absorbed at either end.

The greatest possible power will be delivered to the load Z when the individual currents induced as the wave sweeps across the wire all combine properly on reaching the load. The currents will reach Z in optimum phase when the time required for a current to flow from the far end of the antenna to Z is exactly one-half cycle longer than the time taken by the wave to sweep over the antenna. A half cycle is equivalent to a half wavelength greater than the distance traversed by the wave from the instant it strikes

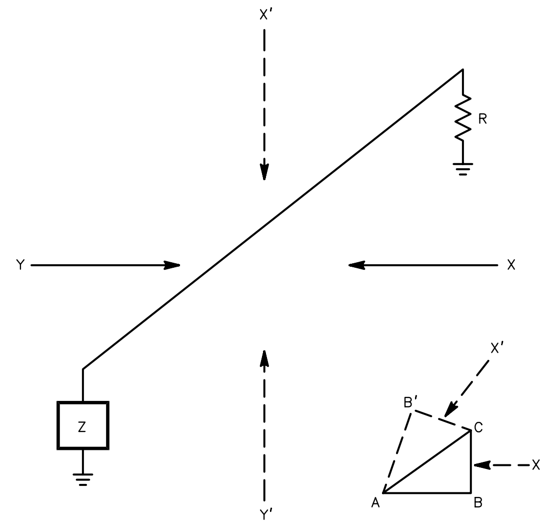

Fig 21—Layout for a terminated long-wire antenna.

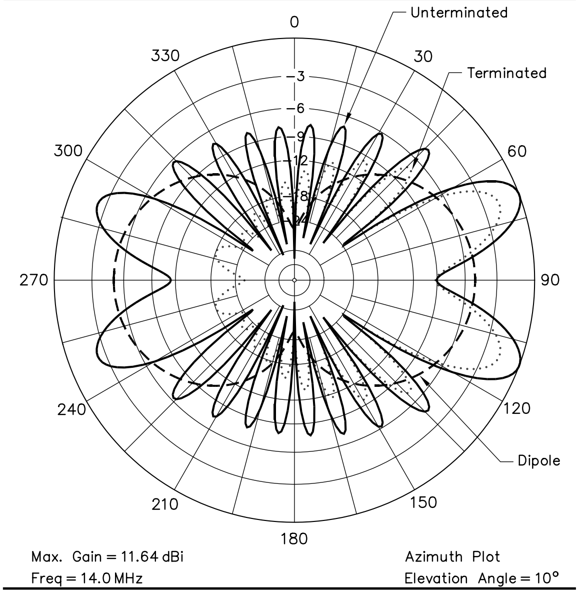

Fig 22—Azimuthal-plane pattern for 5-λ long-wire antenna at 14 MHz and 70 feet above flat ground. The solid line shows the long-wire terminated with 600-Ω to ground, while the dashed line is for the same antenna unterminated. For comparison, the response for a ½-λ dipole is overlaid with the two other patterns. You can see that the terminated long-wire has a good front-to-back pattern, but it loses about 2 dB in forward gain compared to the unterminated long-wire.

the far end of the antenna to the instant that it reaches the near end. This is shown by the small drawing, where AC represents the antenna, BC is a line perpendicular to the wave direction, and AB is the distance traveled by the wave in sweeping past AC. AB must be one-half wavelength shorter than AC. Similarly, AB' must be the same length as AB for a wave arriving from X'.

A wave arriving at the antenna from the opposite direction Y (or Y'), will similarly result in the largest possible current at the far end. However, since the far end is terminated in R, which is equal to Z, all the power delivered to R by the wave arriving from Y will be absorbed in R. The current traveling to Z will produce a signal in Z in proportion to its amplitude. If the antenna length is such that all the individual currents arrive at Z in such phase as to add up to zero, there will be no current through Z. At other lengths the resultant current may reach appreciable values. The lengths that give zero amplitude are those which are odd multiples of $^{1}/_{4}$ λ, beginning at $^{3}/_{4}$ λ. The response from the Y direction is greatest when the antenna is any even multiple of $^{1}/_{2}$ λ long; the higher the multiple, the smaller the response.

Directional Characteristics

Fig 22 compares the azimuthal pattern for a 5-λ long 14-MHz long-wire antenna, 70 feet high over flat ground, when it is terminated and when it is unterminated. The rearward pattern when the wire is terminated with a 600 Ω resistor is reduced about 15 dB, with a reduction in gain in the forward direction of about 2 dB.

For a shorter leg length in a terminated long-wire antenna, the reduction in forward gain is larger—more energy is radiated by a longer wire before the forward wave is absorbed in the terminating resistor. The azimuthal patterns for terminated and unterminated V-beams with 2-λ legs are overlaid for comparison in **Fig 23**. With these relatively short legs the reduction in forward gain is about 3.5 dB due to the terminations, although the front-to-rear ratio approaches 20 dB for the terminated V-beam. Each leg of this terminated V-beam use a 600-Ω non-inductive resistor to ground. Each resistor would have to dissipate about one-quarter of the transmitter power. For average conductor diameters and heights above ground, the Z_{ANT} of the antenna is of the order of 500 to 600 Ω.

THE TERMINATED RHOMBIC ANTENNA

The highest development of the long-wire antenna is the *terminated rhombic*, shown schematically in **Fig 24**. It consists of four conductors joined to form a diamond, or *rhombus*. All sides of the antenna have the same length and the opposite corner angles are equal. The antenna can be considered as being made up of two V antennas placed end to end and terminated by a noninductive resistor to produce a unidirectional pattern. The terminating resistor is connected between the far ends of the two sides, and is made approximately equal to the characteristic impedance of the antenna as a unit. The rhombic may be constructed either horizontally or vertically, but is practically always constructed horizontally at frequencies below 54 MHz,

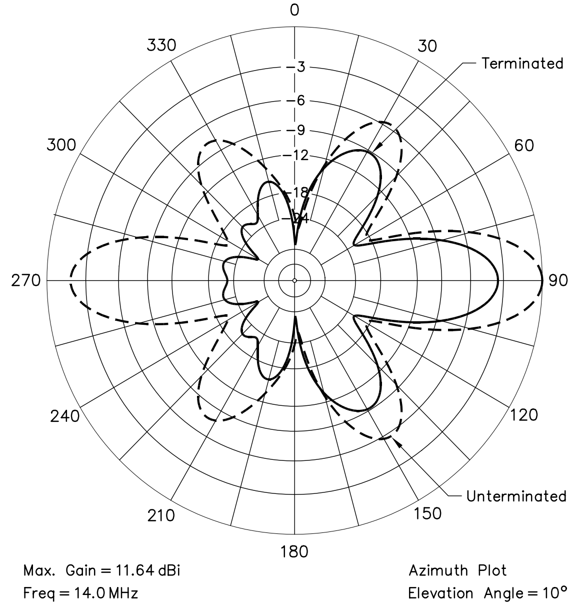

Fig 23—The azimuthal patterns for a shorter-leg V-beam (2-λ legs) when it is terminated (solid line) and unterminated (dashed line). With shorter legs, the terminated V-beam loses about 3.5 dB in forward gain compared to the unterminated version, while suppressing the rearward lobes as much as 20 dB.

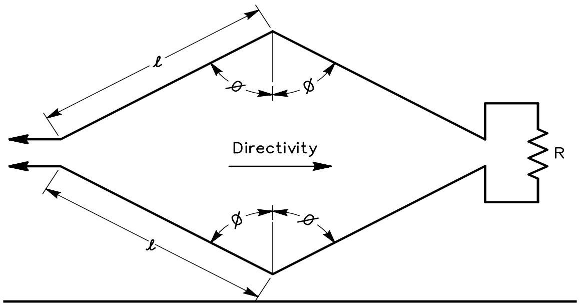

Fig 24—The layout for a terminated rhombic antenna.

since the pole height required is considerably less. Also, horizontal polarization is equally, if not more, satisfactory at these frequencies over most types of soil.

The basic principle of combining lobes of maximum radiation from the four individual wires constituting the rhombus or diamond is the same in either the terminated type or the resonant type described earlier in this chapter.

Tilt Angle

In dealing with the terminated rhombic, it is a matter of custom to talk about the *tilt angle* (ϕ in Fig 24), rather than the angle of maximum radiation with respect to an individual wire. **Fig 25** shows the tilt angle as a function of the antenna leg length. The curve marked "0°" is used for a takeoff elevation angle of 0°; that is, maximum radiation in the plane of the antenna. The other curves show the proper tilt angles to use when aligning the major lobe with a desired takeoff angle. For a 5° takeoff angle, the difference in tilt angle is less than 1° for the range of lengths shown.

The broken curve marked "optimum length" shows the leg length at which maximum gain is obtained at any given takeoff angle. Increasing the leg length beyond the optimum will result in less gain, and for that reason the curves do not extend beyond the optimum length. Note that the optimum length becomes greater as the desired takeoff angle decreases. Leg lengths over 6 λ are not recommended because the directive pattern becomes so sharp that the antenna performance is highly variable with small changes in the angle, both horizontal and vertical, at which an incoming wave reaches the antenna. Since these angles vary to some extent in ionospheric propagation, it does not pay to attempt to try for too great a degree of directivity.

Multiband Design

When a rhombic antenna is to be used over a considerable frequency range, a compromise must be made in the tilt angle. **Fig 26** gives the design dimensions of a suitable compromise for a rhombic that covers the 14 to 30 MHz range well. **Fig 27** shows the azimuth and elevation patterns for this antenna at 14 MHz, at a height of 70 feet over flat ground. The comparison antenna in this case is a 4-element Yagi on a 26-foot boom, also 70 feet above flat ground. The rhombic has about 2.2 dB more gain, but its azimuthal pattern is 17.2° wide at the 3 dB points, and only 26° at the –20 dB points! On the other hand, the Yagi has a 3-dB beamwidth of 63°, making it far easier to aim at a distant geographic location. Fig 27B shows the elevation-plane patterns for the same antennas above. As usual, the peak angle for either horizontally polarized antenna is determined mainly by the height above ground.

The peak gain of a terminated rhombic is less than that of an unterminated resonant rhombic. For the rhombic of Fig 26, the reduction in peak gain is about 1.5 dB. **Fig 28** compares the azimuthal patterns for this rhombic with and without an 800-Ω termination.

Fig 29 shows the azimuth and elevation patterns for the terminated rhombic of Fig 26 when it is operated at 28 MHz. The main lobe becomes very narrow, at 6.9° at the 3-dB points. However, this is partially compensated for by the appearance of two sidelobes each side of the main beam. These tend to spread out the main pattern some. Again, a 4-element Yagi at the same height is used for comparison.

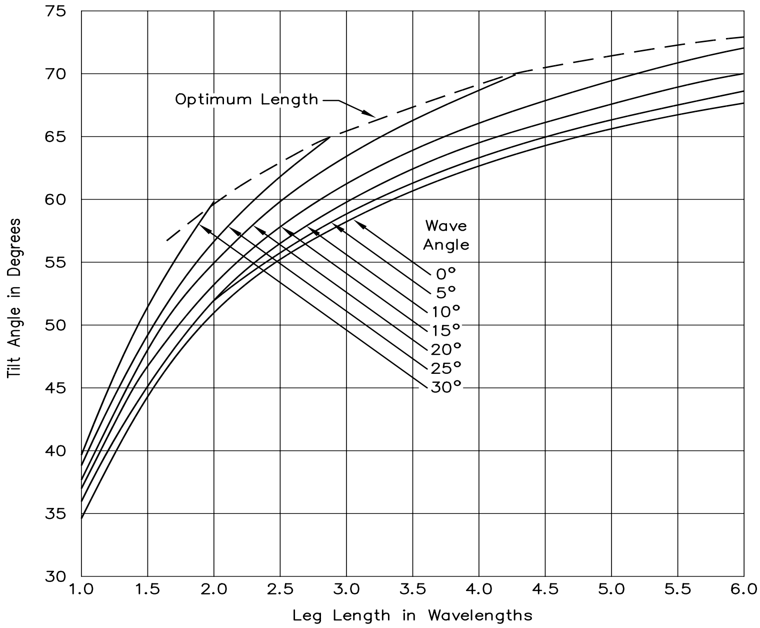

Fig 25—Rhombic-antenna design chart. For any given leg length, the curves show the proper tilt angle to give maximum radiation at the selected takeoff angle. The broken curve marked "optimum length" shows the leg length that gives the maximum possible output at the selected takeoff angle. The optimum length as given by the curves should be multiplied by 0.74 to obtain the leg length for which the takeoff angle and main lobe are aligned.

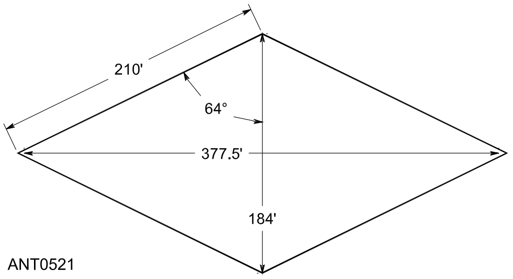

Fig 26—Rhombic antenna dimensions for a compromise design between 14- and 28-MHz requirements, as discussed in the text. The leg length is 6 λ at 28 MHz, 3 λ at 14 MHz.

Termination

Although the difference in the gain is relatively small with terminated or unterminated rhombics of comparable design, the terminated antenna has the advantage that over a wide frequency range it presents an essentially resistive and constant load to the transmitter. In a sense, the power dissipated in the terminating resistor can be considered power that would have been radiated in the other direction had the resistor not been there. Therefore, the fact that some of the power (about one-third) is used up in heating the resistor does not mean that much actual loss in the desired direction.

The characteristic impedance of an ordinary rhombic antenna, looking into the input end, is in the order of 700 to 800 Ω when properly terminated in a resistance at the far end. The terminating resistance required to bring about the matching condition usually is slightly higher than the input impedance because of the loss of energy through radiation by the time the far end is reached. The correct value usually will be found to be of the order of 800 Ω, and should be determined experimentally if the flattest possible antenna is desired. However, for average work a noninductive resistance of 800 Ω can be used with the assurance that the operation will not be far from optimum.

The terminating resistor must be practically a pure resistance at the operating frequencies; that is, its inductance and capacitance should be negligible. Ordinary wire-wound resistors are not suitable because they have far too much inductance and distributed capacitance. Small carbon resistors have satisfactory electrical characteristics but will not dissipate more than a few watts and so cannot be used, except when the transmitter power does not exceed 10 or 20 watts or when the antenna is to be used for reception only. The special resistors designed either for use as dummy antennas or for terminating rhombic antennas should be used in other cases. To allow a factor of safety, the total rated power dissipation of the resistor or resistors should be equal to half the power output of the transmitter.

To reduce the effects of stray capacitance it is desirable to use several units, say three, in series even when one alone will safely dissipate the power. The two end units should be identical and each should have one fourth to one third the total resistance, with the center unit making up the difference. The units should be installed in a weather-

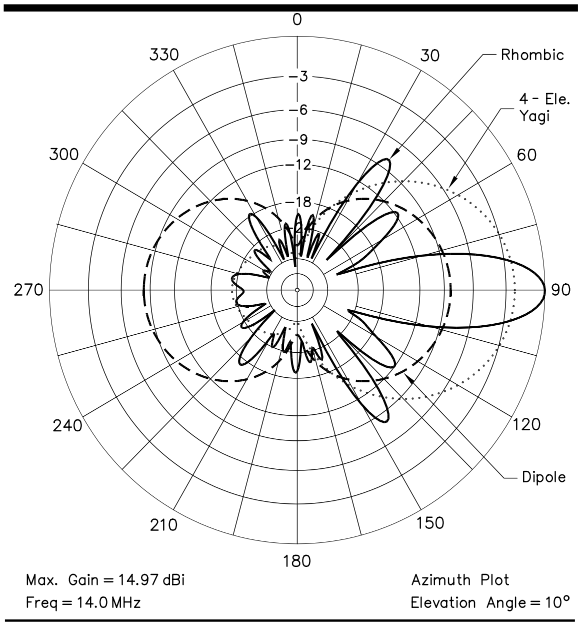

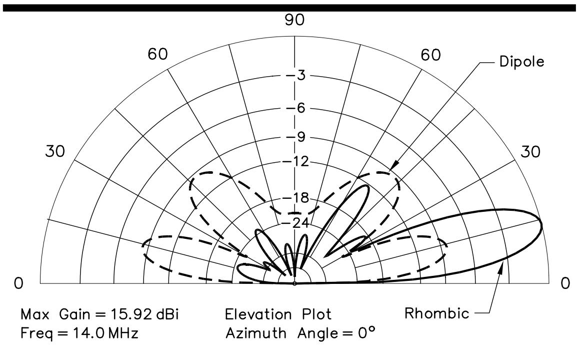

Fig 27—At left, azimuthal pattern for 3-λ (at 14 MHz) terminated rhombic (solid line) shown in Fig 26, compared with 4-element 20-meter Yagi (dotted line) on a 26-foot boom and a 20-meter dipole (dashed line). All antennas are mounted 70 feet (1 λ) above flat ground. The rearward pattern of the terminated rhombic is good and the forward gain exceeds that of the Yagi, but the frontal lobe is very narrow. Above, elevation-plane pattern of terminated rhombic compared to that of a simple dipole at the same height.

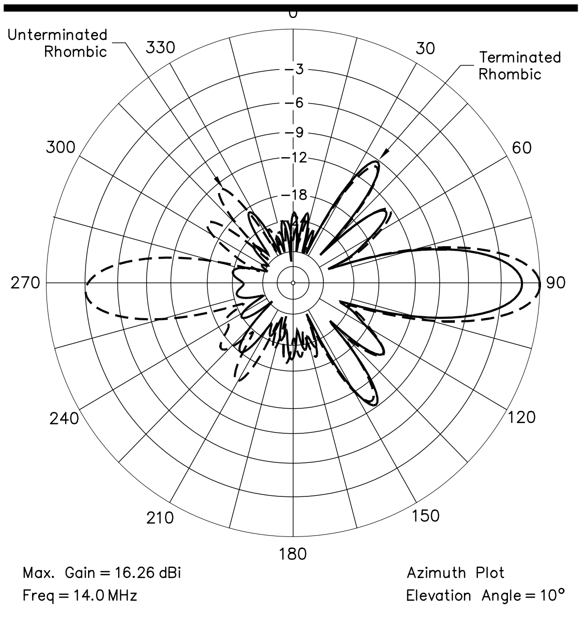

Fig 28—Comparison of azimuthal patterns for terminated (solid line) and unterminated (dashed line) rhombic antennas, using same dimensions as Fig 26 at a frequency of 14 MHz. The gain tradeoff is about 1.5 dB in return for the superior rearward pattern of the terminated antenna.

proof housing at the end of the antenna to protect them and to permit mounting without mechanical strain. The connecting leads should be short so that little extraneous inductance is introduced.

Alternatively, the terminating resistance may be placed at the end of an 800-Ω line connected to the end of the antenna. This will permit placing the resistors and their housing at a point convenient for adjustment rather than at the top of the pole. Resistance wire may be used for this line, so that a portion of the power will be dissipated before it reaches the resistive termination, thus permitting the use of lower wattage lumped resistors.

Multiwire Rhombics

The input impedance of a rhombic antenna constructed as in Fig 26 is not quite constant as the frequency is varied. This is because the varying separation between the wires causes the characteristic impedance of the antenna to vary along its length. The variation in Z_{ANT} can be minimized by a conductor arrangement that increases the capacitance per unit length in proportion to the separation between the wires.

The method of accomplishing this is shown in **Fig 30**. Three conductors are used, joined together at the ends but with increasing separation as the junction between legs is approached. For HF work the spacing between the wires at the center is 3 to 4 feet, which is similar to that used in commercial installations using legs several wavelengths long. Since all three wires should have the same length, the top and bottom wires should be slightly farther from the support than the middle wire. Using three wires in this way reduces the Z_{ANT} of the antenna to approximately 600 Ω, thus providing a better match for practical open-wire line, in addition to smoothing out the impedance variation over the frequency range.

A similar effect (although not quite as favorable) is obtained by using two wires instead of three. The 3-wire system has been found to increase the gain of the

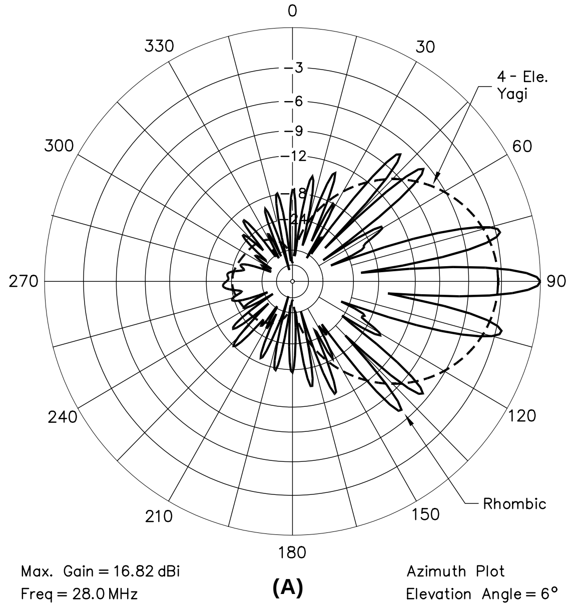

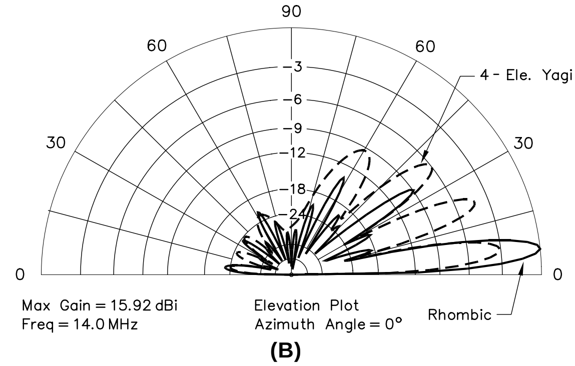

Fig 29—At A, the azimuthal pattern for the same terminated antenna in Fig 26, but now at 28 MHz compared to a 4-element 10-meter Yagi. At B, the elevation-plane pattern comparison for these antennas.

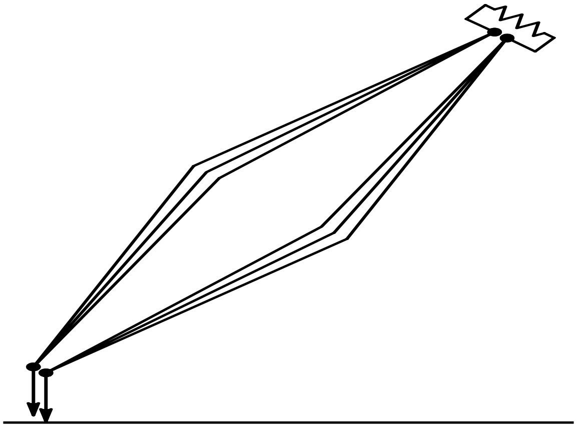

Fig 30—Three-wire rhombic antenna. Use of multiple wires improves the impedance characteristic of a terminated rhombic and increases the gain somewhat.

antenna by about 1 dB over that of a single-conductor version.

Front-to-Back Ratio

It is theoretically possible to obtain an infinite front-to-back ratio with a terminated rhombic antenna, and in practice very large values can be had. However, when the antenna is terminated in its characteristic impedance, the infinite front-to-back ratio can be obtained only at frequencies for which the leg length is an odd multiple of a quarter wavelength. The front-to-back ratio is smallest at frequencies for which the leg length is a multiple of a half wavelength.

When the leg length is not an odd multiple of a quarter wave at the frequency under consideration, the front-to-back ratio can be made very high by decreasing the value of terminating resistance slightly. This permits a small reflection from the far end of the antenna, which cancels out the residual response at the input end. With large antennas, the front-to-back ratio may be made very large over the whole frequency range by experimental adjustment of the terminating resistance. Modification of the terminating resistance can result in a splitting of the back null into two nulls, one on either side of a small lobe in the back direction. Changes in the value of terminating resistance thus permit steering the back null over a small horizontal range so that signals coming from a particular spot not exactly to the rear of the antenna may be minimized.

Methods of Feed

If the broad frequency characteristic of the terminated rhombic antenna is to be utilized fully, the feeder system must be similarly broadbanded. Open-wire transmission line of the same characteristic impedance as that shown at the antenna input terminals (approximately 700 to 800 Ω) may be used. Data for the construction of such lines is given in Chapter 24. While the usual matching stub can be used to provide an impedance transformation to more satisfactory line impedances, this limits the operation of the antenna to a comparatively narrow range of frequencies centering about that for which the stub is adjusted. Probably a more satisfactory arrangement would be to use a coaxial transmission line and a broadband transformer balun at the antenna feed point.

Receiving Wave Antennas

Perhaps the best known type of wave antenna is the *Beverage*. Many 160-meter enthusiasts have used Beverage antennas to enhance the signal-to-noise ratio while attempting to extract weak signals from the often high levels of atmospheric noise and interference on the low bands. Alternative antenna systems have been developed and used over the years, such as loops and long spans of unterminated wire on or slightly above the ground, but the Beverage antenna seems to be the best for 160-meter weak-signal reception. The information in this section was prepared originally by Rus Healy, K2UA (ex-NJ2L).

THE BEVERAGE ANTENNA

A Beverage is simply a directional wire antenna, at least one wavelength long, supported along its length at a fairly low height and terminated at the far end in its characteristic impedance. This antenna is shown in **Fig 31A**. It takes its name from its inventor, Harold Beverage, W2BML.

Many amateurs choose to use a single-wire Beverage because they are easy to install and they work well. The drawback is that Beverages are physically long and they do require that you have the necessary amount of real estate to install them. Sometimes, a neighbor will allow you to put up a temporary Beverage for a particular contest or DXpedition on his land, particularly during the winter months.

Beverage antennas can be useful into the HF range, but they are most effective at lower frequencies, mainly on 160 through 40 meters. The antenna is responsive mostly to low-angle incoming waves that maintain a constant (vertical) polarization. These conditions are nearly always satisfied on 160 meters, and most of the time on 80 meters. As the frequency is increased, however, the polarization and arrival angles are less and less constant and favorable, making Beverages less effective at these frequencies. Many amateurs have, however, reported excellent performance from Beverage antennas at frequencies as high as 14 MHz, especially when rain or snow (precipitation) static prevents good reception on the Yagi or dipole transmitting antennas used on the higher frequencies.

Beverage Theory

The Beverage antenna acts like a long transmission line with one lossy conductor (the earth), and one good conductor (the wire). Beverages have excellent directivity if erected properly, but they are quite inefficient because they are mounted close to the ground. This is in contrast with the terminated long-wire antennas described earlier, which are typically mounted high off the ground. Beverage antennas are not suitable for use as transmitting antennas.

Because the Beverage is a traveling wave, terminated antenna, it has no standing waves resulting from radio signals. As a wave strikes the end of the Beverage from the desired direction, the wave induces voltages along the antenna and continues traveling in space as well. Fig 31B shows part of a wave on the antenna resulting from a desired signal. This diagram also shows the tilt of the wave. The signal induces equal voltages in both directions. The resulting currents are equal and travel in both directions; the component traveling toward the termination end moves *against* the wave and thus builds down to a very low level at the termination end. Any residual signal resulting from this direction of current flow will be absorbed in the termination (if the termination is equal to the antenna impedance). The component of the signal flowing in the other direction, as we will see, becomes a key part of the received signal.

As the wave travels along the wire, the wave in space travels at approximately the same velocity. (There is some phase delay in the wire, as we shall see.) At any given point in time, the wave traveling along in space induces a voltage in the wire in addition to the wave already travel-

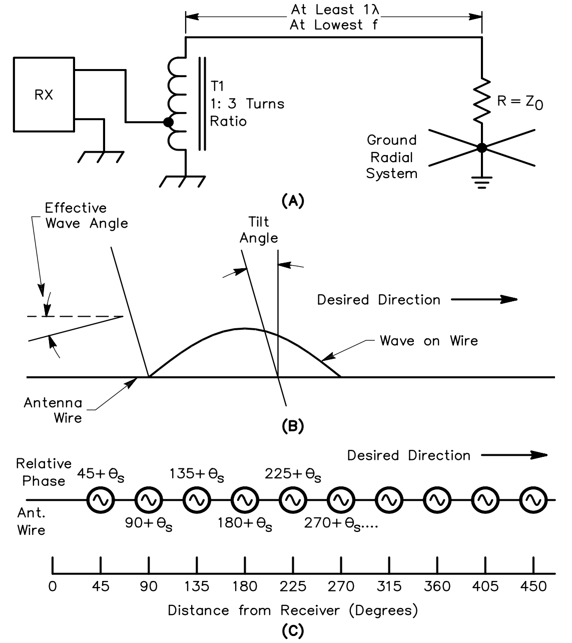

Fig 31—At A, a simple one-wire Beverage antenna with a variable termination impedance and a matching 9:1 autotransformer for the receiver impedance. At B, a portion of a wave from the desired direction is shown traveling down the antenna wire. Its tilt angle and effective takeoff angle are also shown. At C, a situation analogous to the action of a Beverage on an incoming wave is shown. See text for discussion.

ing on the wire (voltages already induced by the wave). Because these two waves are nearly in phase, the voltages add and build toward a maximum at the receiver end of the antenna.

This process can be likened to a series of signal generators lined up on the wire, with phase differences corresponding to their respective spacings on the wire (Fig 31C). At the receiver end, a maximum voltage is produced by these voltages adding in phase. For example, the wave component induced at the receiver end of the antenna will be in phase (at the receiver end) with a component of the same wave induced, say, 270° (or any other distance) down the antenna, after it travels to the receiver end.

In practice, there is some phase shift of the wave on the wire with respect to the wave in space. This phase shift results from the velocity factor of the antenna. (As with any transmission line, the signal velocity on the Beverage is somewhat less than in free space.) Velocity of propagation on a Beverage is typically between 85 and 98% of that in free space. As antenna height is increased to a certain optimum height (which is about 10 feet for 160 meters), the velocity factor increases. Beyond this height, only minimal improvement is afforded, as shown in **Fig 32**. These curves are the result of experimental work done in 1922 by RCA, and reported in a *QST* article (November 1922) entitled "The Wave Antenna for 200-Meter Reception," by H. H. Beverage. The curve for 160 meters was extrapolated from the other curves.

Phase shift (per wavelength) is shown as a function of velocity factor in **Fig 33**, and is given by:

$$\theta = 360 \left(\frac{100}{k} - 1 \right) \qquad \text{(Eq 2)}$$

where k = velocity factor of the antenna in percent.

The signals present on and around a Beverage antenna are shown graphically in A through D of **Fig 34**. These curves show relative voltage levels over a number of periods of the wave in space and their relative effects in terms of the total signal at the receiver end of the antenna.

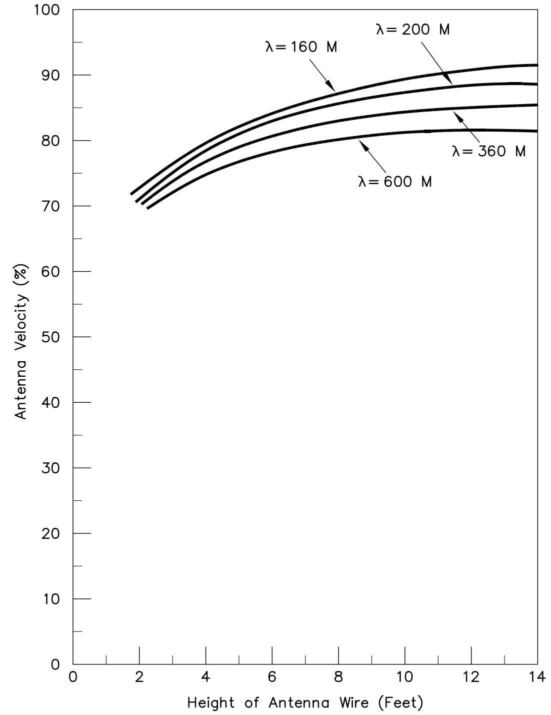

Fig 32—Signal velocity on a Beverage increases with height above ground, and reaches a practical maximum at about 10 feet. Improvement is minimal above this height. (The velocity of light is 100%.)

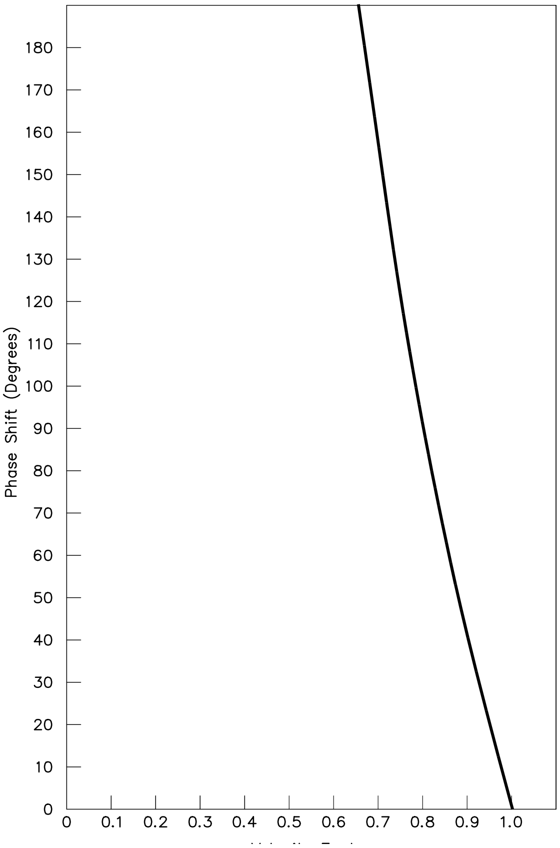

Fig 33—This curve shows phase shift (per wavelength) as a function of velocity factor on a Beverage antenna. Once the phase shift for the antenna goes beyond 90°, the gain drops off from its peak value, and any increase in antenna length will decrease gain.

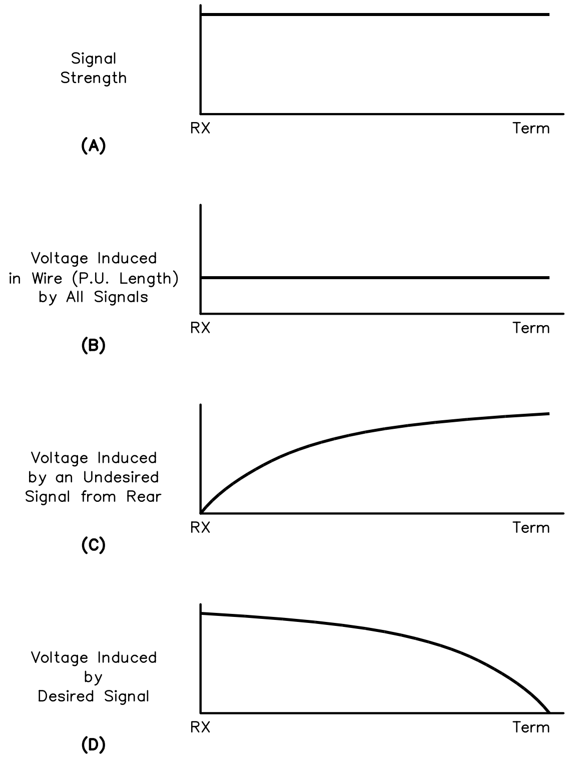

Fig 34—These curves show the voltages that appear in a Beverage antenna over a period of several cycles of the wave. Signal strength (at A) is constant over the length of the antenna during this period, as is voltage induced per unit length in the wire (at B). (The voltage induced in any section of the antenna is the same as the voltage induced in any other section of the same size, over the same period of time.) At C, the voltages induced by an undesired signal from the rearward direction add in phase and build to a maximum at the termination end, where they are dissipated in the termination (if $Z_{term} = Z_0$). The voltages resulting from a desired signal are shown at D. The wave on the wire travels closely with the wave in space, and the voltages resulting add in phase to a maximum at the receiver end of the antenna.

Performance in Other Directions

The performance of a Beverage antenna in directions other than the favored one is quite different than previously discussed. Take, for instance, the case of a signal arriving perpendicular to the wire (90° either side of the favored direction). In this case, the wave induces voltages along the wire that are essentially *in phase*, so that they arrive at the receiver end more or less out of phase, and thus cancel. (This can be likened to a series of signal generators lined up along the antenna as before, but having no progressive phase differences.)

As a result of this cancellation, Beverages exhibit deep nulls off the sides. Some minor sidelobes will exist, as with other long-wire antennas, and will increase in number with the length of the antenna.

In the case of a signal arriving from the rear of the antenna, the behavior of the antenna is very similar to its performance in the favored direction. The major difference is that the signal from the rear adds in phase at the termination end and is absorbed by the termination impedance. **Fig 35** compares the azimuth and elevation patterns for a 2-λ (1062 foot) and a 1-λ (531 foot) Beverage at 1.83 MHz. The wire is mounted 8 feet above flat ground (to keep it above deer antlers and away from humans too) and is terminated with a 500-Ω resistor in each case, although the exact value of the terminating resistance is not very critical. The ground constants assumed in this computer model are conductivity of 5 mS/m and a dielectric constant of 13. Beverage dielectric performance tends to decrease as the ground becomes better. Beverages operated over saltwater do not work as well as they do over poor ground.

For most effective operation, the Beverage should be terminated in an impedance equal to the characteristic impedance Z_{ANT} of the antenna. For maximum signal transfer to the receiver you should also match the receiver's input impedance to the antenna. If the termination impedance is not equal to the characteristic impedance of the antenna, some part of the signal from the rear will be reflected back toward the receiver end of the antenna.

If the termination impedance is merely an open circuit (no terminating resistor), total reflection will result and the antenna will exhibit a bidirectional pattern (still with very deep nulls off the sides). An unterminated Beverage will not have the same response to signals in the rearward direction as it exhibits to signals in the forward direction because of attenuation and reradiation of part of the reflected wave as it travels back toward the receiver end. **Fig 36** compares the response from two 2-λ Beverages, one terminated and the other unterminated. Just like a terminated long-wire transmitting antenna (which is mounted higher off the ground than a Beverage, which is meant only for receiving), the terminated Beverage has a reduced forward lobe compared to its unterminated sibling. The unterminated Beverage exhibits about a 5 dB front-to-back ratio for this length because of the radiation and wire and ground losses that occur before the forward wave gets to the end of the wire.

If the termination is between the extremes (open circuit and perfect termination in Z_{ANT}), the peak direction and intensity of signals off the rear of the Beverage will change. As a result, an adjustable reactive termination can be employed to *steer* the nulls to the rear of the antenna (see **Fig 37**). This can be of great help in eliminating a local interfering signal from a rearward direction (typically 30° to 40° either side of the back direction). Such a scheme doesn't help much for interfering skywave signals because of variations encountered in the ionosphere that constantly shift polarity, amplitude, phase and

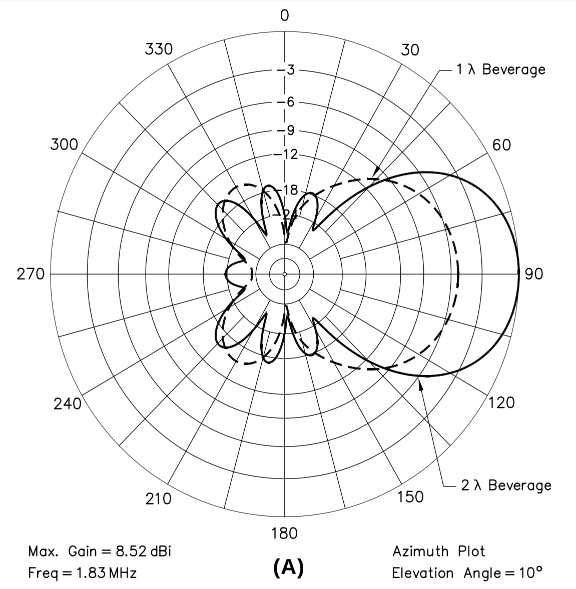

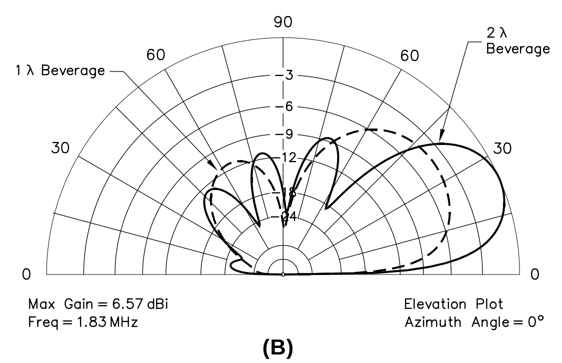

Fig 35—At A, azimuthal patterns of a 2-λ (solid line) and a 1-λ (dashed line) Beverage antenna, terminated with 550-Ω resistor at 1.83 MHz, at an elevation angle of 10°. The rearward pattern around 180° is more than 20 dB down from the front lobe for each antenna. At B, the elevation-plane patterns. Note the rejection of very high-angle signals near 90°.

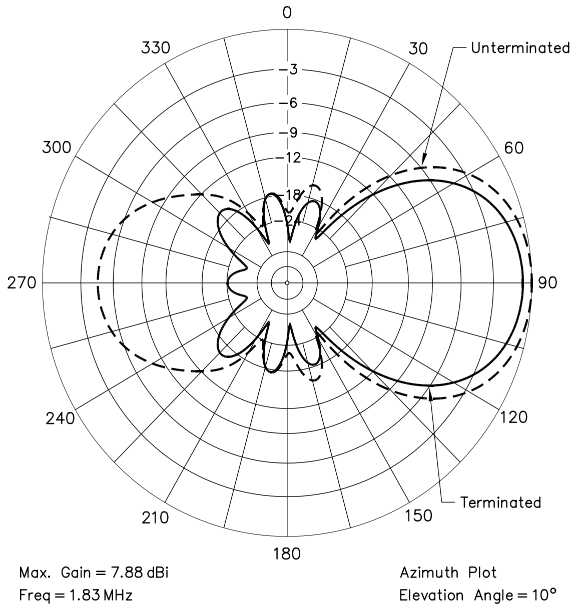

Fig 36—Comparing the azimuthal patterns for a 2-λ Beverage, terminated (solid line) and unterminated (dashed line).

incoming elevation angles.

To determine the appropriate value for a terminating resistor, you need to know the characteristic impedance (surge impedance), Z_{ANT}, of the Beverage. It is interesting to note that Z_{ANT} is not a function of the length, just like a transmission line.

$$Z_{ANT} = 138 \times \log\left(\frac{4h}{d}\right) \qquad \text{(Eq 3)}$$

where

Z_{ANT} = characteristic impedance of the Beverage = terminating resistance needed
h = wire height above ground
d = wire diameter (in the same units as h)

Another aspect of terminating the Beverage is the quality of the RF ground used for the termination. For most types of soil a ground rod is sufficient, since the optimum value for the termination resistance is in the range of 400 to 600 Ω for typical Beverages and the ground-loss resistance is in series with this. Even if the ground-loss resistance at the termination point is as high as 40 or 50 Ω, it still is not an appreciable fraction of the overall terminating resistance. For soil with very poor conductivity, however, (such as sand or rock) you can achieve a better ground termination by laying radial wires on the ground at both the receiver and termination ends. These wires need not be resonant quarter-wave in length, since the ground detunes them anyway. Like the ground counterpoise for a vertical antenna, a number of short radials is better than a few long ones. Some amateurs use chicken-wire ground screens for their ground terminations.

As with many other antennas, improved directivity and gain can be achieved by lengthening the antenna and by arranging several antennas into an array. One item that must be kept in mind is that by virtue of the velocity factor of the antenna, there is some phase shift of the wave on the antenna with respect to the wave in space. Because of this phase shift, although the directivity will continue to sharpen with increased length, there will be some optimum length at which the gain of the antenna will peak. Beyond this

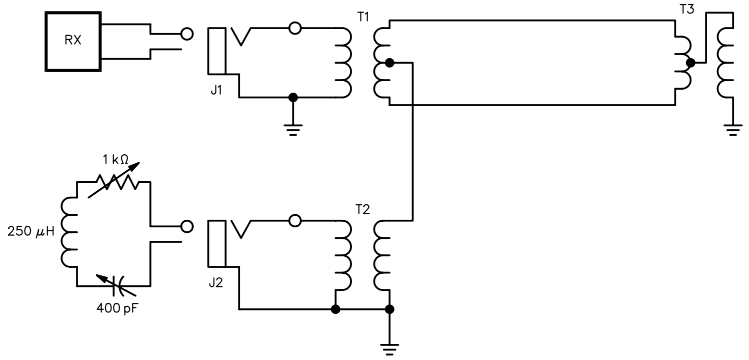

Fig 37—A two-wire Beverage antenna that has provisions for direction switching and null steering in the rear quadrant. Performance improves with height to a point, and is optimum for 1.8-MHz operation at about 10 to 12 feet. Parts identifications are for text reference.

length, the current increments arriving at the receiver end of the antenna will no longer be in phase, and will not add to produce a maximum signal at the receiver end. This optimum length is a function of velocity factor and frequency, and is given by:

$$L = \frac{\lambda}{4\left(\frac{100}{k} - 1\right)} \qquad \text{(Eq 4)}$$

where

L = maximum effective length
λ = signal wavelength in free space (same units as L)
k = velocity factor of the antenna in percent

Because velocity factor increases with height (to a point, as mentioned earlier), optimum length is somewhat longer if the antenna height is increased. The maximum effective length also increases with the number of wires in the antenna system. For example, for a two-wire Beverage like the bidirectional version shown in Fig 37, the maximum effective length is about 20% longer than the single-wire version. A typical length for a single-wire 1.8-MHz Beverage (made of #16 wire and erected 10 feet above ground) is about 1200 feet.

Feed-Point Transformers for Single-Wire Beverages

Matching transformer T1 in Fig 31 is easily constructed. Small toroidal ferrite cores are best for this application, with those of high permeability (μ_i = 125 to 5000) being the easiest to wind (requiring fewest turns) and having the best high-frequency response (because few turns are used). Trifilar-wound autotransformers are most convenient.

Most users are not concerned with a small amount of SWR on the transmission line feeding their Beverages. For example, let us assume that the Z_{ANT} of a particular Beverage is 525 Ω and the terminating resistance is made equal to that value. If a standard 3:1 turns-ratio autotransformer is used at the input end of the antenna, the nominal impedance transformation $50\ \Omega \times 3^2 = 450\ \Omega$. This leads to the terminology often used for this transformer as a *9:1 transformer*, referring to its impedance transformation. The resulting SWR on the feed line going back to the receiver would be 525/450 = 1.27:1, not enough to be concerned about. For a Z_{ANT} of 600 Ω, the SWR is 600/450 = 1.33:1, again not a matter of concern.

Hence, most Beverage users use standard 9:1 (450:50 Ω) autotransformers. You can make a matching transformer suitable for use from 160 to 40 meters using eight trifilar turns of #24 enameled wire wound over a stack of two Amidon FT-50-75 or two MN8-CX cores. See **Fig 38**.

Make your own trifilar cable bundle by placing three 3-foot lengths of the #24 wire side-by-side and twisting them in a hand drill so that there is a uniform twist about one twist-per-inch. This holds the three wires together in a bundle that can be passed through the two stacked cores, rather like threading a needle. Remember that each time you put the bundle through the center of the cores counts as one turn.

After you finish winding, cut the individual wires to leave about 3/4-inch leads, sand off the enamel insulation and tin the wires with a soldering iron. Identify the individual wires with an ohmmeter and then connect them together following Fig 38. Coat the transformer with Q-dope (liquid polystyrene) to finalize the transformer. White glue will work also. See Chapters 25 and 26 and *The ARRL Handbook* for more information on winding toroidal transformers or see Chapter 7 (Special Receiving Antennas) of *ON4UN's Low-Band DXing* book.

The Two-Wire Beverage

The two-wire antenna shown in Fig 37 has the major advantage of having signals from both directions available at the receiver at the flip of a switch between J1 and J2. Also, because there are two wires in the system (equal amounts of signal voltage are induced in both wires), greater signal voltages will be produced.

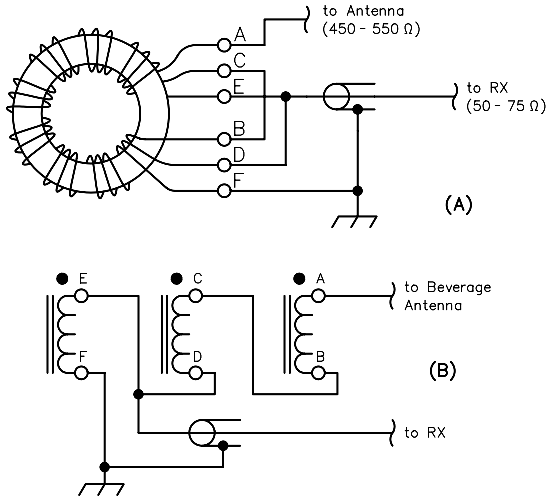

Fig 38—Constructing the feed-point transformer for a single-wire Beverage. See text for details.

A signal from the left direction in Fig 37 induces equal voltages in both wires, and equal in-phase currents flow as a result. The *reflection transformer* (T3 at the right-hand end of the antenna) then inverts the phase of these signals and reflects them back down the antenna toward the receiver, using the antenna wires as a balanced open-wire transmission line. This signal is then transformed by T1 down to the input impedance of the receiver (50 Ω) at J1.

Signals traveling from right to left also induce equal voltages in each wire, and they travel in phase toward the receiver end, through T1, and into T2. Signals from this direction are available at J2.

T1 and T2 are standard 9:1 wideband transformers capable of operating from 1.8 to at least 10 MHz. Like any two parallel wires making up a transmission line, the two-wire Beverage has a certain characteristic impedance—we'll call it Z_1 here—depending on the spacing between the two wires and the insulation between them. T3 transforms the terminating resistance needed at the end of the line to Z_1. Keep in mind that this terminating resistance is equal to the characteristic impedance Z_{ANT} of the Beverage—that is, the impedance of the parallel wires over their images in the ground below. For example, if Z_1 of the Beverage wire is 300 Ω (that is, you used TV twin-lead for the two Beverage wires), T3 must transform the balanced 300 Ω to the unbalanced 500 Ω Z_{ANT} impedance used to terminate the Beverage.

The design and construction of the reflection transformer used in a two-wire Beverage is more demanding than that for the straightforward matching transformer T1 because the exact value of terminating impedance is more critical for good F/B. See Chapter 7 (Special Receiving Antennas) in *ON4UN's Low-Band DXing* for details on winding the reflection transformers for a two-wire Beverage.

Another convenient feature of the two-wire Beverage is the ability to steer the nulls off either end of the antenna while receiving in the opposite direction. For instance, if the series RLC network shown at J2 is adjusted while the receiver is connected to J1, signals can be received from the left direction while interference coming from the right can be partially or completely nulled. The nulls can be steered over a 60° (or more) area off the right-hand end of the antenna. The same null-steering capability exists in the opposite direction with the receiver connected at J2 and the termination connected at J1.

The two-wire Beverage is typically erected at the same height as a single-wire version. The two wires are at the same height and are spaced uniformly—typically 12 to 18 inches apart for discrete wires. Some amateurs construct two-wire Beverages using "window" ladder-line, twisting the line about three twists per foot for mechanical and electrical stability in the wind.

The characteristic impedance Z_{ANT} of a Beverage made using two discrete wires with air insulation between them depends on the wire size, spacing and height and is given by:

$$Z_{ANT} = \frac{69}{\sqrt{\varepsilon}} \times \log \left[\frac{4h}{d} \sqrt{1 + \left(\frac{2h}{S} \right)^2} \right] \qquad \text{(Eq 5)}$$

where

Z_{ANT} = Beverage impedance = desired terminating resistance
S = wire spacing
h = height above ground
d = wire diameter (in same units as S and h)
ε = 2.71828

Beverages in Echelon

The pattern of a Beverage receiving antenna is dependent on the terminating resistance used for a particular antenna, as was demonstrated at the extremes by Fig 36. This compared the patterns for a terminated and an unterminated Beverage. The pattern of even a poorly terminated Beverage can be significantly improved by the addition of a second Beverage. The additional Beverage is installed so that it is operated *in echelon*, a word deriving from the fact that the two wires look like the parallel rungs on a ladder. For a practial 160- and 80-meter setup the second Beverage wire is parallel to the first Beverage, spaced from it by about 5 meters, and also staggered 30 meters ahead. See **Fig 38**.

The forward Beverage is fed with a phase difference of +125° such that the total phase, including that due to the forward staggering, is 180°. This forms the equivalent of an end-fire array fed out-of-phase, but it takes advantage of the natural directivity of each Beverage. **Fig 39** compares the pattern of a single 1-λ 160-meter Beverage that is sloppily terminated with two Beverages fed in echelon. The

Beverages in echelon gives a modest additional gain of almost 2 dB. But where the two Beverages in echelon really shine is how they cleans up the rearward pattern—from an average about 15 dB for the single Beverage to more than 25 dB for the two Beverages.

Even at a spacing of 5 meters, there is very little mutual coupling between the two Beverage wires because of their inherently small radiation resistance when they are mounted low above lossy ground. If you adjust for a low SWR (using proper transformers to match the feed-line coaxes), the phase difference will depend solely on the difference in length of the two coaxes feeding the Beverage wires. **Fig 40** shows a wideband feed system designed by Tom Rauch, W8JI, as a "cross-fire" feed system. The 180° wideband phase-inverting transformer allows the system to work on two bands, say 160 and 80 meters. See Chapter 7, Receiving Antennas, in *ON4UN's Low-Band DXing* book, 4th Edition for transformer details.

Practical Considerations

Even though Beverage antennas have excellent directive patterns if terminated properly, gain never exceeds about −3 dBi in most practical installations. However, the directivity that the Beverage provides results in a much higher signal-to-noise ratio for signals in the desired direction than almost any other real-world antenna used at low frequencies.

A typical situation might be a station located in the US Northeast (W1), trying to receive Topband signals from Europe to the northeast, while thunderstorms behind him in the US Southeast (W4) are creating huge static crashes. Instead of listening to an S7 signal with 10-dB over S9 noise and interference on a vertical, the directivity of a Beverage will typically allow you to copy the same signal at perhaps S5 with only S3 (or lower) noise and interference. This is certainly a worthwhile improvement. However, if you are in the middle of a thunderstorm, or if there is a thunderstorm in the direction from which you are trying to receive a signal, no Beverage is going to help you!

There are a few basic principles that must be kept in mind when erecting Beverage antennas if optimum performance is to be realized.

1) Plan the installation thoroughly, including choosing an antenna length consistent with the optimum length values discussed earlier.
2) Keep the antenna as straight and as nearly level as possible over its entire run. Avoid following the terrain under the antenna too closely—keep the antenna level with the average terrain.
3) Minimize the lengths of vertical downleads at the ends of the antenna. Their effect is detrimental to the directive pattern of the antenna. It is best to slope the antenna wire from ground level to its final height (over a distance of 50 feet or so) at the feed-point end. Similar action should be taken at the termination end. Be sure to seal the transformers against weather.

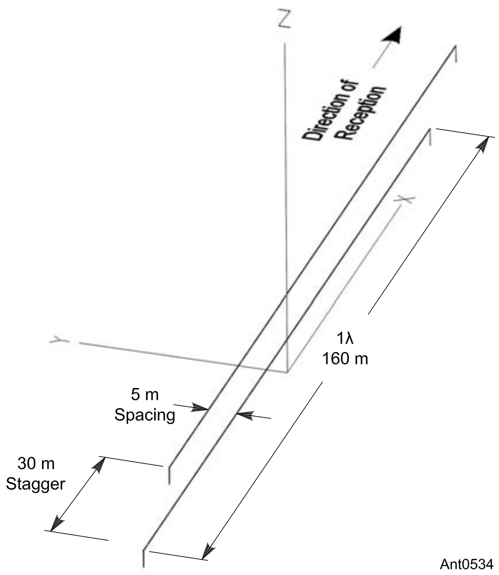

Fig 39—Layout of two 160-meter 1-λ long Beverages in echelon, spaced 5 meters apart, with 30 meter forward stagger. The upper antenna has a 125° phase shift in its feed system.

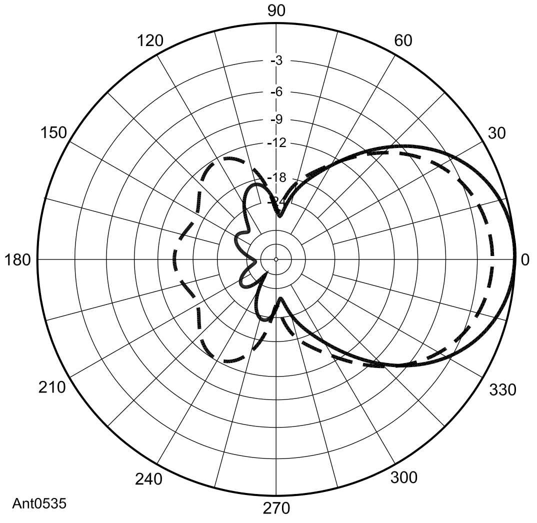

Fig 40—Azimuth pattern at 10° takeoff angle for single Beverage (dashed line) and two Beverages in an echelon end-fire array. The rearward pattern is considerably cleaner on the echelon. Thus, two closely spaced, short Beverages can give considerable improvement over a single short Beverage.

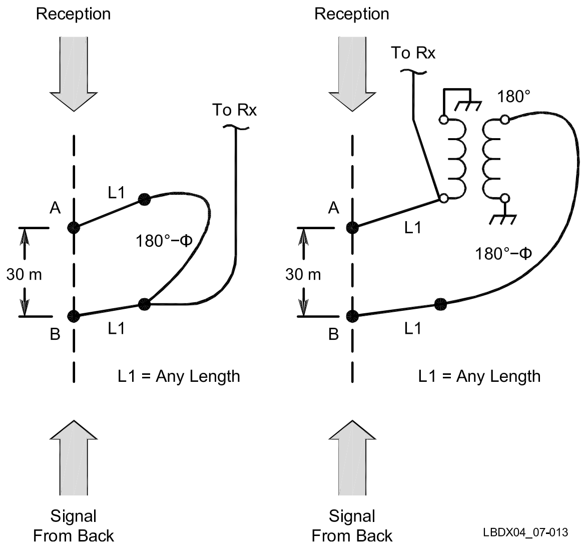

Fig 41—Two ways of feeding the two-Beverage echelon array in Fig 39. On the left, a feed system good for one frequency; on the right, a "cross-fire" feed system good for 1.8 and 3.6 MHz. For this system we want a phase shift due to the coax length of +116° at the back Beverage A. The angle ϕ is thus 180°–116° = 64° long on 160 meters. In the system on the right, a 64° length on 160 meters becomes 128° long on 80 meters. So with the phase-inverting transformer the net phase shift becomes 53° on 80 meters, a reasonable compromise. (*Courtesy W8JI and ON4UN.*)

4) Use a noninductive resistor for terminating a single-wire Beverage. If you live in an area where lightning storms are common, use 2-W terminating resistors, which can survive surges due to nearby lightning strikes.
5) Use high-quality insulators for the Beverage wire where it comes into contact with the supports. Plastic insulators designed for electric fences are inexpensive and effective.
6) Keep the Beverage away from parallel conductors such as electric power and telephone lines for a distance of at least 200 feet. Perpendicular conductors, even other Beverages, may be crossed with relatively little interaction, but do not cross any conductors that may pose a safety hazard.
7) Run the coaxial feed line to the Beverage so that it is not directly under the span of the wire. This prevents common-mode currents from appearing on the shield of the coax. It may be necessary to use a ferrite-bead choke on the feed line if you find that the feed line itself picks up signals when it is temporarily disconnected from the Beverage. See Chapter 26 for details on common-mode chokes.
8) If you use elevated radials in your transmitting antenna system, keep your Beverage feed lines well away from them to avoid stray pickup that will ruin the Beverage's directivity.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

A. Bailey, S. W. Dean and W. T. Wintringham, "The Receiving System for Long-Wave Transatlantic Radio Telephony," *The Bell System Technical Journal*, Apr 1929.

J. S. Belrose, "Beverage Antennas for Amateur Communications," Technical Correspondence, *QST*, Sep 1981.

H. H. Beverage, "Antennas," *RCA Review*, Jul 1939.

H. H. Beverage and D. DeMaw, "The Classic Beverage Antenna Revisited," *QST* , Jan 1982.

B. Boothe, "Weak-Signal Reception on 160—Some Antenna Notes," *QST*, Jun 1977.

E. Bruce, "Developments in Short-Wave Directive Antennas," *Proc IRE*, Aug 1931.

E. Bruce, A. C. Beck and L. R. Lowry, "Horizontal Rhombic Antennas," *Proc IRE*, Jan 1935.

P. S. Carter, C. W. Hansel and N. E. Lindenblad, "Development of Directive Transmitting Antennas by R.C.A. Communications," *Proc IRE*, Oct 1931.

J. Devoldere, *Low-Band DXing* (Newington: ARRL, 1987).

J. Devoldere, *ON4UN's Low-Band DXing* (Newington: ARRL, 1999). See in particular Chapter 7, "Special Receiving Antennas," for many practical details on Beverage antennas.

A. E. Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc).

E. A. Laport, "Design Data for Horizontal Rhombic Antennas," *RCA Review*, Mar 1952.

G. M. Miller, *Modern Electronic Communication* (Englewood Cliffs, NJ: Prentice Hall, 1983).

V. A. Misek, *The Beverage Antenna Handbook* (Wason Rd., Hudson, NH: W1WCR, 1977).

F. E. Terman, *Radio Engineering*, Second Edition (New York: McGraw-Hill, 1937).

Chapter 14

Direction Finding Antennas

The use of radio for direction-finding purposes (RDF) is almost as old as its application for communications. Radio amateurs have learned RDF techniques and found much satisfaction by participating in hidden-transmitter hunts. Other hams have discovered RDF through an interest in boating or aviation, where radio direction finding is used for navigation and emergency location systems.

In many countries of the world, the hunting of hidden amateur transmitters takes on the atmosphere of a sport, as participants wearing jogging togs or track suits dash toward the area where they believe the transmitter is located. The sport is variously known as *fox hunting*, *bunny hunting*, ARDF (Amateur Radio direction finding) or simply transmitter hunting. In North America, most hunting of hidden transmitters is conducted from automobiles, although hunts on foot are gaining popularity.

There are less pleasant RDF applications as well, such as tracking down noise sources or illegal operators from unidentified stations. Jammers of repeaters, traffic nets and other amateur operations can be located with RDF equipment. Or sometimes a stolen amateur rig will be operated by a person who is not familiar with Amateur Radio, and by being lured into making repeated transmissions, the operator unsuspectingly permits himself to be located with RDF equipment. The ability of certain RDF antennas to reject signals from selected directions has also been used to advantage in reducing noise and interference. Through APRS, radio navigation is becoming a popular application of RDF. The locating of downed aircraft is another, and one in which amateurs often lend their skills. Indeed, there are many useful applications for RDF.

Although sophisticated and complex equipment pushing the state of the art has been developed for use by governments and commercial enterprises, relatively simple equipment can be built at home to offer the Radio Amateur an opportunity to RDF. This chapter deals with antennas suitable for that purpose.

RDF by Triangulation

It is impossible, using amateur techniques, to pinpoint the whereabouts of a transmitter from a single receiving location. With a directional antenna you can determine the direction of a signal source, but not how far away it is. To find the distance, you can then travel in the determined direction until you discover the transmitter location. However, that technique can be time consuming and often does not work very well.

A preferred technique is to take at least one additional direction measurement from a second receiving location. Then use a map of the area and plot the bearing or direction measurements as straight lines from points on the map representing the two locations. The approximate location of the transmitter will be indicated by the point where the two bearing lines cross. Even better results can be obtained by taking direction measurements from three locations and using the mapping technique just described. Because absolutely precise bearing measurements are difficult to obtain in practice, the three lines will almost always cross to form a triangle on the map, rather than at a single point. The transmitter will usually be located inside the area represented by the triangle. Additional information on the technique of triangulation and much more on RDF techniques may be found in recent editions of *The ARRL Handbook*.

DIRECTION FINDING SY\STEMS

Required for any RDF system are a directive antenna and a device for detecting the radio signal. In amateur applications the signal detector is usually a transceiver and for convenience it will usually have a meter to indicate signal strength. Unmodified, commercially available portable or mobile receivers are generally quite satisfactory for signal detectors. At very close ranges a simple diode detector and dc microammeter may suffice for the detector.

On the other hand, antennas used for RDF techniques are not generally the types used for normal two-way communications. Directivity is a prime requirement, and here the word *directivity* takes on a somewhat different meaning than is commonly applied to other amateur antennas. Normally we associate directivity with gain, and we think of the ideal antenna pattern as one having a long, thin main lobe. Such a pattern may be of value for coarse measurements in RDF work, but precise bearing measurements are not possible. There is always a spread of a few (or perhaps many) degrees on the *nose* of the lobe, where a shift of antenna bearing produces no detectable change in signal strength. In RDF measure-ments, it is desirable to correlate an exact bearing or compass direction with the position of the antenna. In order to do this as accurately as possible, an antenna exhibiting a *null* in its pattern is used. A null can be very sharp in directivity, to within a half degree or less.

Loop Antennas

A simple antenna for HF RDF work is a small loop tuned to resonance with a capacitor. Several factors must be considered in the design of an RDF loop. The loop must be small in circumference compared with the wavelength. In a single-turn loop, the conductor should be less than 0.08 λ long. For 28 MHz, this represents a length of less than 34 inches (a diameter of approximately 10 inches). Maximum response from the loop antenna is in the plane of the loop, with nulls exhibited at right angles to that plane.

To obtain the most accurate bearings, the loop must be balanced electrostatically with respect to ground. Otherwise, the loop will exhibit two modes of operation. One is the mode of a true loop, while the other is that of an essentially nondirectional vertical antenna of small dimensions. This second mode is called the *antenna effect*. The voltages introduced by the two modes are seldom in phase and may add or subtract, depending upon the direction from which the wave is coming.

The theoretical true loop pattern is illustrated in **Fig 1A**. When properly balanced, the loop exhibits two nulls that are 180° apart. Thus, a single null reading with a small loop antenna will not indicate the exact direction toward the transmitter—only the line along which the transmitter lies. Ways to overcome this ambiguity are discussed later.

When the antenna effect is appreciable and the loop is tuned to resonance, the loop may exhibit little directivity, as shown in Fig 1B. However, by detuning the loop to shift the phasing, a pattern similar to 1C may be obtained. Although this pattern is not symmetrical, it does exhibit a null. Even so, the null may not be as sharp as that obtained with a loop that is well balanced, and it may not be at exact right angles to the plane of the loop.

By suitable detuning, the unidirectional cardioid pattern of Fig 1D may be approached. This adjustment is sometimes used in RDF work to obtain a unidirectional bearing, although there is no complete null in the pattern. A cardioid pattern can also be obtained with a small loop antenna by adding a *sensing element*. Sensing elements are discussed in a later section of this chapter.

An electrostatic balance can be obtained by shielding the loop, as **Fig 2** shows. The shield is represented by the broken lines in the drawing, and eliminates the antenna effect. The response of a well-constructed shielded loop is quite close to the ideal pattern of Fig 1A.

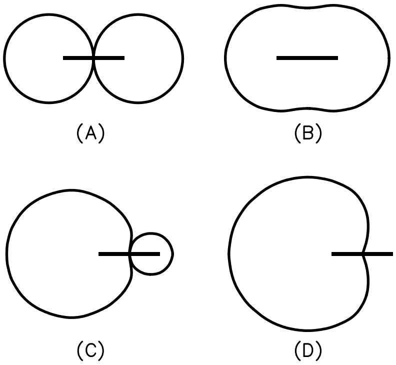

Fig 1—Small-loop field patterns with varying amounts of antenna effect—the undesired response of the loop acting merely as a mass of metal connected to the receiver antenna terminals. The straight lines show the plane of the loop.

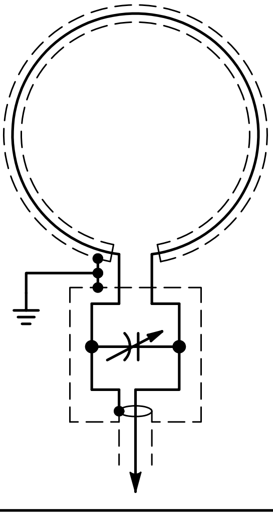

Fig 2—Shielded loop for direction finding. The ends of the shielding turn are not connected, to prevent shielding the loop from magnetic fields. The shield is effective against electric fields.

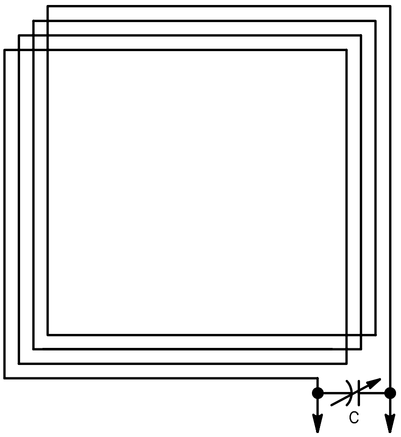

Fig 3—Small loop consisting of several turns of wire. The total conductor length is very much less than a wavelength. Maximum response is in the plane of the loop.

For the low-frequency amateur bands, single-turn loops of convenient physical size for portability are generally found to be unsatisfactory for RDF work. Therefore, multiturn loops are generally used instead. Such a loop is shown in **Fig 3**. This loop may also be shielded, and if the total conductor length remains below 0.08λ, the directional pattern is that of Fig 1A. A sensing element may also be used with a multiturn loop.

Loop Circuits and Criteria

No single word describes a direction-finding loop of high performance better than *symmetry*. To obtain an undistorted response pattern from this type of antenna, you must build it in the most symmetrical manner possible. The next key word is *balance*. The better the electrical balance, the deeper the loop null and the sharper the maxima.

The physical size of the loop for 7 MHz and below is not of major consequence. A 4-foot diameter loop will exhibit the same electrical characteristics as one which is only an inch or two in diameter. The smaller the loop, however, the lower its efficiency. This is because its aperture samples a smaller section of the wave front. Thus, if you use loops that are very small in terms of a wavelength, you will need preamplifiers to compensate for the reduced efficiency.

An important point to keep in mind about a small loop antenna oriented in a vertical plane is that it is vertically polarized. It should be fed at the bottom for the best null response. Feeding it at one side, rather than at the bottom, will not alter the polarization and will only degrade performance. To obtain horizontal polarization from a small loop, it must be oriented in a horizontal plane, parallel to the earth. In this position the loop response is essentially omnidirectional.

The earliest loop antennas were of the *frame antenna* variety. These were unshielded antennas built on a wooden frame in a rectangular format. The loop conductor could be a single turn of wire (on the larger units) or several turns if the frame was small. Later, shielded versions of the frame antenna became popular, providing electrostatic shielding—an aid to noise reduction from such sources as precipitation static.

Ferrite Rod Antennas

With advances in technology, magnetic-core loop antennas came into use. Their advantage was reduced size, and this appealed especially to the designers of aircraft and portable radios. Most of these antennas contain ferrite bars or cylinders, which provide high inductance and Q with a relatively small number of coil turns.

Magnetic-core antennas consist essentially of turns of wire around a ferrite rod. They are also known as *loopstick* antennas. Probably the best-known example of this type of antenna is that used in small portable AM broadcast receivers. Because of their reduced-size advantage, ferrite-rod antennas are used almost exclusively for portable work at frequencies below 150 MHz.

As implied in the earlier discussion of shielded loops in this chapter, the true loop antenna responds to the magnetic field of the radio wave, and not to the electrical field. The voltage delivered by the loop is proportional to the amount of magnetic flux passing through the coil, and to the number of turns in the coil. The action is much the same as in the secondary winding of a transformer. For a given size of loop, the output voltage can be increased by increasing the flux density, and this is done with a ferrite core of high permeability. A $^{1}/_{2}$-inch diameter, 7-inch rod of Q2 ferrite ($m_i = 125$) is suitable for a loop core from the broadcast band through 10 MHz. For increased output, the turns may be wound on two rods that are taped together, as shown in **Fig 4**. Loopstick antennas for con-

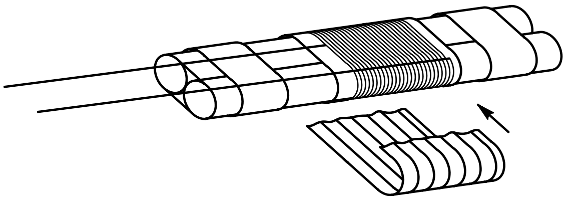

Fig 4—A ferrite-rod or loopstick antenna. Turns of wire may be wound on a single rod, or to increase the output from the loop, the core may be two rods taped together, as shown here. The type of core material must be selected for the intended frequency range of the loop. To avoid bulky windings, fine wire such as #28 or #30 is often used, with larger wire for the leads.

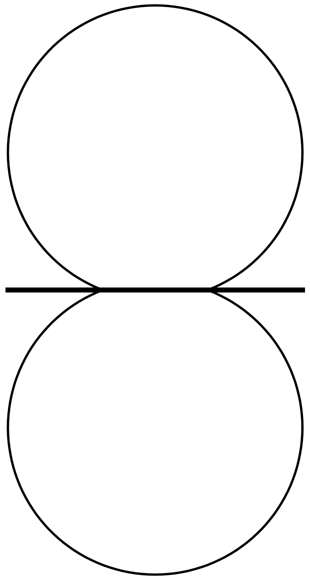

Fig 5—Field pattern for a ferrite rod antenna. The dark bar represents the rod on which the loop turns are wound.

struction are described later in this chapter.

Maximum response of the loopstick antenna is broadside to the axis of the rod as shown in **Fig 5**, whereas maximum response of the ordinary loop is in a direction at right angles to the plane of the loop. Otherwise, the performances of the ferrite-rod antenna and of the ordinary loop are similar. The loopstick may also be shielded to eliminate the antenna effect, such as with a U-shaped or C-shaped channel of aluminum or other form of *trough*. The length of the shield should equal or slightly exceed the length of the rod.

Sensing Antennas

Because there are two nulls that are 180° apart in the directional pattern of a loop or a loopstick, an ambiguity exists as to which one indicates the true direction of the station being tracked. For example, assume you take a bearing measurement and the result indicates the transmitter is somewhere on a line running approximately east and west from your position. With this single reading, you have no way of knowing for sure if the transmitter is east of you or west of you.

If more than one receiving station takes bearings on a single transmitter, or if a single receiving station takes bearings from more than one position on the transmitter, the ambiguity may be worked out by triangulation, as described earlier. However, it is sometimes desirable to have a pattern with only one null, so there is no question about whether the transmitter in the above example would be east or west from your position.

A loop or loopstick antenna may be made to have a single null if a second antenna element is added. The element is called a *sensing antenna*, because it gives an added sense of direction to the loop pattern. The second element must be omnidirectional, such as a short vertical. When the signals from the loop and the vertical element are combined with a 90° phase shift between the two, a cardioid pattern results. The development of the pattern is shown in **Fig 6A**.

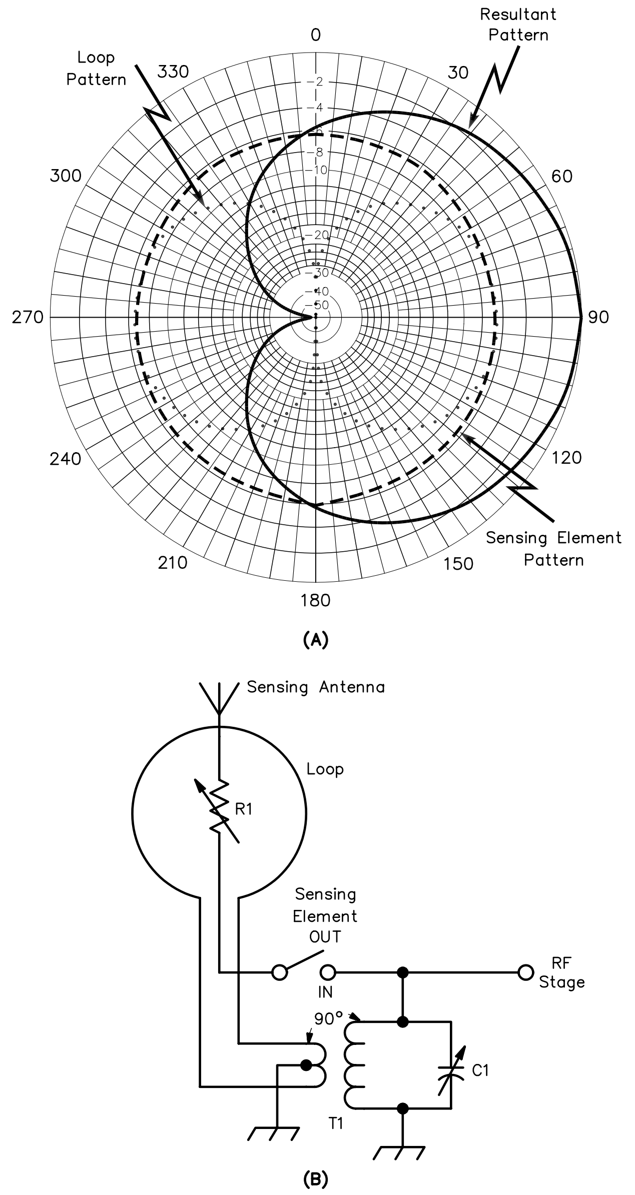

Fig 6—At A, the directivity pattern of a loop antenna with sensing element. At B is a circuit for combining the signals from the two elements. C1 is adjusted for resonance with T1 at the operating frequency.

Fig 6B shows a circuit for adding a sensing antenna to a loop or loopstick. R1 is an internal adjustment and is used to set the level of the signal from the sensing antenna. For the best null in the composite pattern, the signals from the loop and the sensing antenna must be of equal amplitude, so R1 is adjusted experimentally during setup. In practice, the null of the cardioid is not as sharp as that of the loop, so the usual measurement procedure is to first use the loop alone to obtain a precise bearing reading, and then to add the sensing antenna and take another reading to resolve the ambiguity. (The null of the cardioid is 90° away from the nulls of the loop.) For this reason, provisions are usually made for switching the sensing element in an out of operation.

PHASED ARRAYS

Phased arrays are also used in amateur RDF work. Two general classifications of phased arrays are end-fire and broadside configurations. Depending on the spacing and phasing of the elements, end-fire patterns may exhibit a null in one direction along the axis of the elements. At the same time, the response is maximum off the other end of the axis, in the opposite direction from the null. A familiar arrangement is two elements spaced $^1/_4$ λ apart and fed 90° out of phase. The resultant pattern is a *cardioid*, with the null in the direction of the leading element. Other arrangements of spacing and phasing for an end-fire array are also suitable for RDF work. One of the best known is the *Adcock array*, discussed in the next section.

Broadside arrays are inherently bidirectional, which means there are always at least two nulls in the pattern. Ambiguity therefore exists in the true direction of the transmitter, but depending on the application, this may be no handicap. Broadside arrays are seldom used for amateur RDF applications however.

The Adcock Antenna

Loops are adequate in RDF applications where only the ground wave is present. The performance of an RDF system for sky-wave reception can be improved by the use of an Adcock antenna, one of the most popular types of end-fire phased arrays. A basic version is shown in **Fig 7**.

This system was invented by F. Adcock and patented in 1919. The array consists of two vertical elements fed 180° apart, and mounted so the system may be rotated. Element spacing is not critical, and may be in the range from 0.1 to 0.75 λ. The two elements must be of identical lengths, but need not be self-resonant. Elements that are shorter than resonant are commonly used. Because neither the element spacing nor the length is critical in terms of wavelengths, an Adcock array may be operated over more than one amateur band.

The response of the Adcock array to vertically polarized waves is similar to a conventional loop, and the directive pattern is essentially the same. Response of the array to a horizontally polarized wave is considerably different from that of a loop, however. The currents induced in the horizontal members tend to balance out regardless of the orientation of the antenna. This effect has been verified in practice, where good nulls were obtained with an experimental Adcock under sky-wave conditions. The same circumstances produced poor nulls with small loops (both conventional and ferrite-loop models).

Generally speaking, the Adcock antenna has attractive properties for amateur RDF applications. Unfortunately, its portability leaves something to be desired, making it more suitable to fixed or semi-portable applications. While a metal support for the mast and boom could be used, wood, PVC or fiberglass are preferable because they are nonconductors and would therefore cause less pattern distortion.

Since the array is balanced, an antenna tuner is required to match the unbalanced input of a typical receiver. **Fig 8** shows a suitable link-coupled network. C2 and C3 are null-balancing capacitors. A low-power signal source is placed some distance from the Adcock

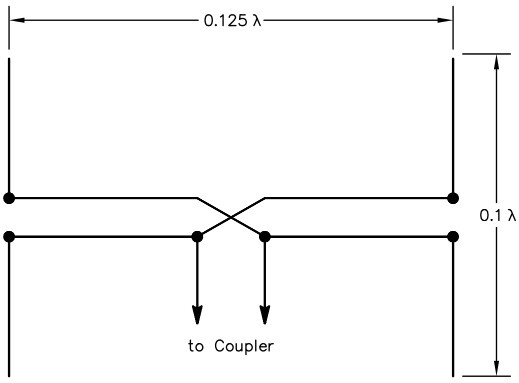

Fig 7—A simple Adcock antenna.

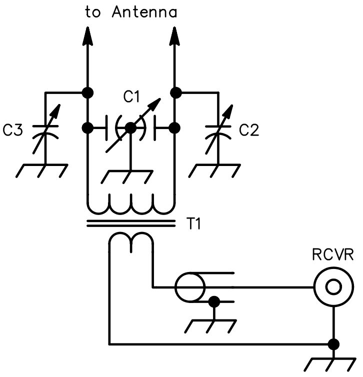

Fig 8—A suitable coupler for use with the Adcock antenna.

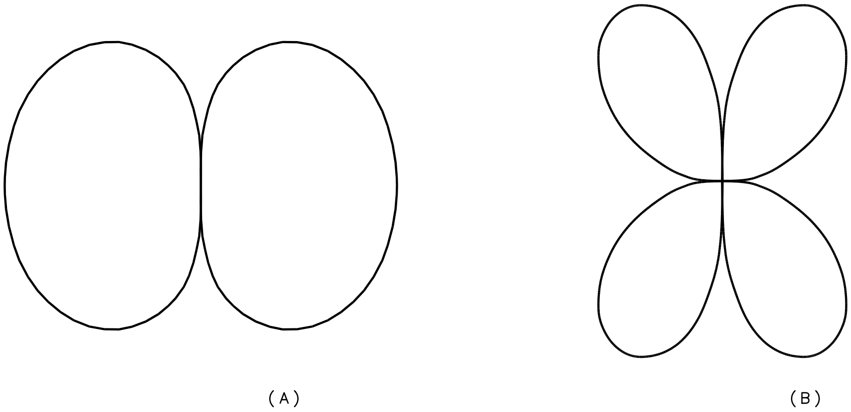

Fig 9—At A, the pattern of the Adcock array with an element spacing of $^1/_2$ wavelength. In these plots the elements are aligned with the horizontal axis. As the element spacing is increased beyond $^3/_4$ wavelength, additional nulls develop off the ends of the array, and at a spacing of 1 wavelength the pattern at B exists. This pattern is unsuitable for RDF work.

antenna and broadside to it. C2 and C3 are then adjusted until the deepest null is obtained. The tuner can be placed below the wiring-harness junction on the boom. Connection can be made by means of a short length of 300-Ω twin-lead.

The radiation pattern of the Adcock is shown in **Fig 9A**. The nulls are in directions broadside to the array, and become sharper with greater element spacings. However, with an element spacing greater than 0.75 λ, the pattern begins to take on additional nulls in the directions off the ends of the array axis. At a spacing of 1 λ the pattern is that of Fig 9B, and the array is unsuitable for RDF applications.

Short vertical monopoles are often used in what is sometimes called the *U-Adcock*, so named because the elements with their feeders take on the shape of the letter U. In this arrangement the elements are worked against the earth as a ground or counterpoise. If the array is used only for reception, earth losses are of no great consequence. Short, elevated vertical dipoles are also used in what is sometimes called the *H-Adcock*.

The Adcock array, with two nulls in its pattern, has the same ambiguity as the loop and the loopstick. Adding a sensing element to the Adcock array has not met with great success. Difficulties arise from mutual coupling between the array elements and the sensing element, among other things. Because Adcock arrays are used primarily for fixed-station applications, the ambiguity presents no serious problem. The fixed station is usually one of a group of stations in an RDF network.

LOOPS VERSUS PHASED ARRAYS

Although loops can be made smaller than suitable phased arrays for the same frequency of operation, the phased arrays are preferred by some for a variety of reasons. In general, sharper nulls can be obtained with phased arrays, but this is also a function of the care used in constructing and feeding the individual antennas, as well as of the size of the phased array in terms of wavelengths. The primary constructional consideration is the shielding and balancing of the feed line against unwanted signal pickup, and the balancing of the antenna for a symmetrical pattern.

Loops are not as useful for skywave RDF work because of random polarization of the received signal. Phased arrays are somewhat less sensitive to propagation effects, probably because they are larger for the same frequency of operation and therefore offer some space diversity. In general, loops and loopsticks are used for mobile and portable operation, while phased arrays are used for fixed-station operation. However, phased arrays are used successfully above 144 MHz for portable and mobile RDF work. Practical examples of both types of antennas are presented later in this chapter.

THE GONIOMETER

Most fixed RDF stations for government and commercial work use antenna arrays of stationary elements, rather than mechanically rotatable arrays. This has been true since the earliest days of radio. The early-day device that permits finding directions without moving the elements is called a *radiogoniometer*, or simply a *goniometer*. Various types of goniometers are still used today in many installations, and offer the amateur some possibilities.

The early style of goniometer is a special form of RF transformer, as shown in **Fig 10**. It consists of two fixed coils mounted at right angles to one another. Inside the fixed coils is a movable coil, not shown in Fig 10 to

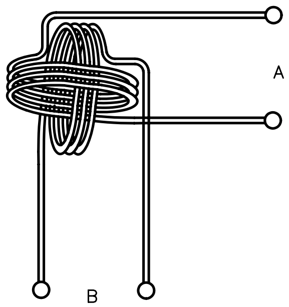

Fig 10—An early type of goniometer that is still used today in some RDF applications. This device is a special type of RF transformer that permits a movable coil in the center (not shown here) to be rotated and determine directions even though the elements are stationary.

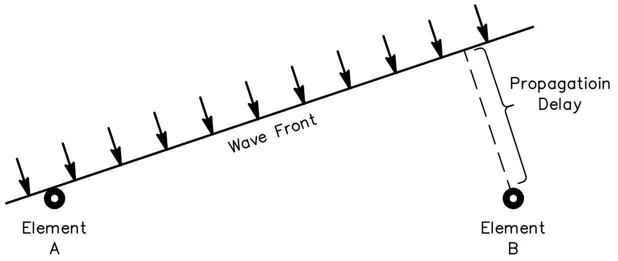

Fig 11—This diagram illustrates one technique used in electronic beam forming. By delaying the signal from element A by an amount equal to the propagation delay, the two signals may be summed precisely in phase, even though the signal is not in the broadside direction. Because this time delay is identical for all frequencies, the system is not frequency sensitive.

avoid cluttering the diagram. The pairs of connections marked A and B are connected respectively to two elements in an array, and the output to the detector or receiver is taken from the movable coil. As the inner coil is rotated, the coupling to one fixed coil increases while that to the other decreases. Both the amplitude and the phase of the signal coupled into the pickup winding are altered with rotation in a way that corresponds to actually rotating the array itself. Therefore, the rotation of the inner coil can be calibrated in degrees to correspond to bearing angles from the station location.

In the early days of radio, the type of goniometer just described saw frequent use with fixed Adcock arrays. A refinement of that system employed four Adcock elements, two arrays at right angles to each other. With a goniometer arrangement, RDF measurements could be taken in all compass directions, as opposed to none off the ends of a two-element fixed array. However, resolution of the 4-element system was not as good as with a single pair of elements, probably because of mutual coupling among the elements. To overcome this difficulty a few systems of eight elements were installed.

Various other types of goniometers have been developed over the years, such as commutator switching to various elements in the array. A later development is the diode switching of capacitors to provide a commutator effect. As mechanical action has gradually been replaced with electronics to “rotate” stationary elements, the word goniometer is used less frequently these days. However, it still appears in many engineering reference texts. The more complex electronic systems of today are called *beam-forming networks*.

Electronic Antenna Rotation

With an array of many fixed elements, beam rotation can be performed electronically by sampling and combining signals from various individual elements in the array. Contingent upon the total number of elements in the system and their physical arrangement, almost any desired antenna pattern can be formed by summing the sampled signals in appropriate amplitude and phase relationships. Delay networks are used for some of the elements before the summation is performed. In addition, attenuators may be used for some elements to develop patterns such as from an array with binomial current distribution.

One system using these techniques is the *Wullenweber* antenna, employed primarily in government and military installations. The Wullenweber consists of a very large number of elements arranged in a circle, usually outside of (or in front of) a circular reflecting screen. Depending on the installation, the circle may be anywhere from a few hundred feet to more than a quarter of a mile in diameter. Although the Wullenweber is not one that would be constructed by an amateur, some of the techniques it uses may certainly be applied to Amateur Radio.

For the moment, consider just two elements of a Wullenweber antenna, shown as A and B in **Fig 11**. Also shown is the wavefront of a radio signal arriving from a distant transmitter. As drawn, the wavefront strikes element A first, and must travel somewhat farther before it strikes element B. There is a finite time delay before the wavefront reaches element B.

The propagation delay may be measured by delaying the signal received at element A before summing it with that from element B. If the two signals are combined directly, the amplitude of the resultant signal will be maximum when the delay for element A exactly equals the propagation delay. This results in an in-phase condition at the summation point. Or if one of the signals is inverted and the two are summed, a null will exist when the element-A delay equals the propagation delay; the signals will combine in a 180° out-of-phase relationship. Either way, once the time delay is known, it may be converted to distance. Then the direction from which the wave is arriving may be determined by trigonometry.

By altering the delay in small increments, the peak of the antenna lobe (or the null) can be steered in azimuth. This is true without regard to the frequency of the incoming wave. Thus, as long as the delay is less than the period of one RF cycle, the system is not frequency sensitive, other than for the frequency range that may be covered satisfactorily by the array elements themselves. Surface acoustic wave (SAW) devices or lumped-constant networks can be used for delay lines in such systems if the system is used only for receiving. Rolls of coaxial cable of various lengths are used in installations for transmitting. In this case, the lines are considered for the time delay they provide, rather than as simple phasing lines. The difference is that a phasing line is ordinarily designed for a single frequency (or for an amateur band), while a delay line offers essentially the same time delay at all frequencies.

By combining signals from other Wullenweber elements appropriately, the broad beamwidth of the pattern from the two elements can be narrowed, and unwanted sidelobes can be suppressed. Then, by electronically switching the delays and attenuations to the various elements, the beam so formed can be rotated around the compass. The package of electronics designed to do this, including delay lines and electronically switched attenuators, is the beam-forming network. However, the Wullenweber system is not restricted to forming a single beam. With an isolation amplifier provided for each element of the array, several beam-forming networks can be operated independently. Imagine having an antenna system that offers a dipole pattern, a rhombic pattern, and a Yagi beam pattern, all simultaneously and without frequency sensitivity. One or more may be rotating while another is held in a particular direction. The Wullenweber was designed to fulfill this type of requirement.

One feature of the Wullenweber antenna is that it can operate 360° around the compass. In many government installations, there is no need for such coverage, as the areas of interest lie in an azimuth sector. In such cases an in-line array of elements with a backscreen or curtain reflector may be installed broadside to the center of the sector. By using the same techniques as the Wullenweber, the beams formed from this array may be slewed left and right across the sector. The maximum sector width available will depend on the installation, but beyond 70° to 80° the patterns begin to deteriorate to the point that they are unsatisfactory for precise RDF work.

RDF SYSTEM CALIBRATION AND USE

Once an RDF system is initially assembled, it should be calibrated or checked out before actually being put into use. Of primary concern is the balance or symmetry of the antenna pattern. A lop-sided figure-8 pattern with a loop, for example, is undesirable; the nulls are not 180° apart, nor are they at exact right angles to the plane of the loop. If you didn't know this fact in actual RDF work, measurement accuracy would suffer.

Initial checkout can be performed with a low-powered transmitter at a distance of a few hundred feet. It should be within visual range and must be operating into a vertical antenna. (A quarter-wave vertical or a loaded whip is quite suitable.) The site must be reasonably clear of obstructions, especially steel and concrete or brick buildings, large metal objects, nearby power lines, and so on. If the system operates above 30 MHz, you should also avoid trees and large bushes. An open field makes an excellent site.

The procedure is to find the transmitter with the RDF equipment as if its position were not known, and compare the RDF null indication with the visual path to the transmitter. For antennas having more than one null, each null should be checked.

If imbalance is found in the antenna system, there are two options available. One is to correct the imbalance. Toward this end, pay particular attention to the feed line. Using a coaxial feeder for a balanced antenna invites an asymmetrical pattern, unless an effective balun is used. A balun is not necessary if the loop is shielded, but an asymmetrical pattern can result with misplacement of the break in the shield itself. The builder may also find that the presence of a sensing antenna upsets the balance slightly, due to mutual coupling. Experiment with its position with respect to the main antenna to correct the error. You will also note that the position of the null shifts by 90° as the sensing element is switched in and out, and the null is not as deep. This is of little concern, however, as the intent of the sensing antenna is only to resolve ambiguities. The sensing element should be switched out when accuracy is desired.

The second option is to accept the imbalance of the antenna and use some kind of indicator to show the true directions of the nulls. Small pointers, painted marks on the mast, or an optical sighting system might be used. Sometimes the end result of the calibration procedure will be a compromise between these two options, as a perfect electrical balance may be difficult or impossible to attain.

The discussion above is oriented toward calibrating portable RDF systems. The same general suggestions apply if the RDF array is fixed, such as an Adcock. However, it won't be possible to move it to an open field. Instead, the array must be calibrated in its intended operating position through the use of a portable or mobile transmitter. Because of nearby obstructions or reflecting objects, the null in the pattern may not appear to indicate the precise direction of the transmitter. Do not confuse this with imbalance in the RDF array. Check for imbalance by rotating the array 180° and comparing readings.

Once the balance is satisfactory, you should make a table of bearing errors noted in different compass directions. These error values should be applied as corrections when actual measurements are made. The mobile or por-

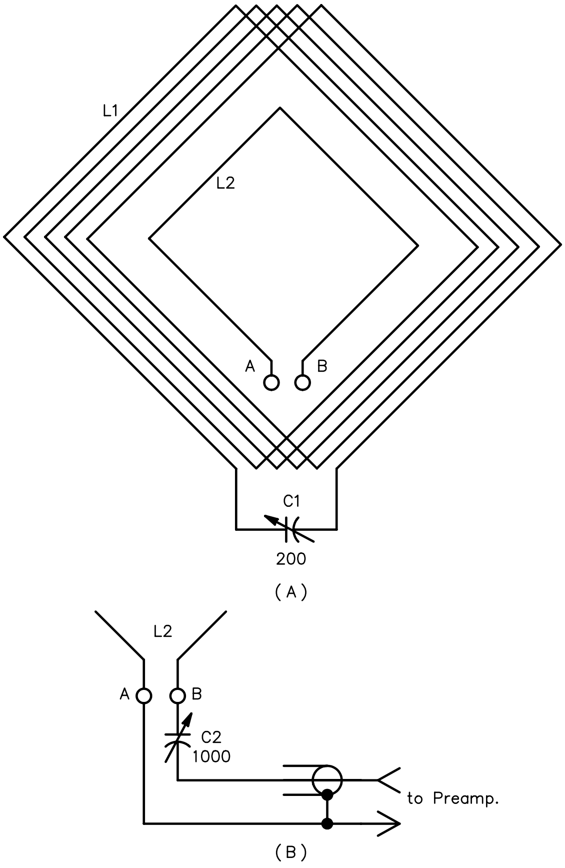

Fig 12—A multiturn frame antenna is shown at A. L2 is the coupling loop. The drawing at B shows how L2 is connected to a preamplifier.

table transmitter should be at a distance of two or three miles for these measurements, and should be in as clear an area as possible during transmissions. The idea is to avoid conduction of the signal along power lines and other overhead wiring from the transmitter to the RDF site. Of course the position of the transmitter must be known accurately for each transmission.

FRAME LOOPS

It was mentioned earlier that the earliest style of receiving loops was the frame antenna. If carefully constructed, such an antenna performs well and can be built at low cost. **Fig 12** illustrates the details of a practical frame type of loop antenna. This antenna was designed by Doug DeMaw, W1FB, and described in *QST* for July 1977. (See the Bibliography at the end of this chapter.) The circuit in Fig 12A is a 5-turn system tuned to resonance by C1. If the layout is symmetrical, good balance should be obtained. L2 helps to achieve this objective by eliminating the need for direct coupling to the feed ter-

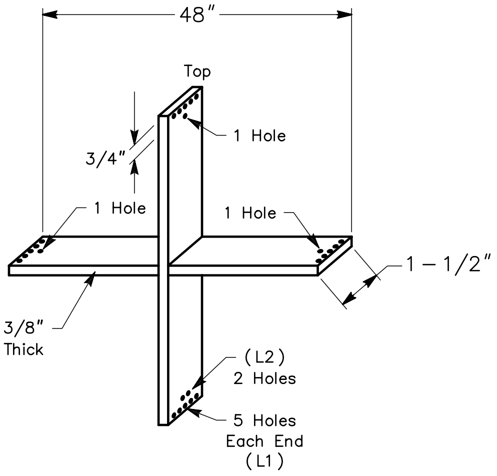

Fig 13—A wooden frame can be used to contain the wire of the loop shown in Fig 12.

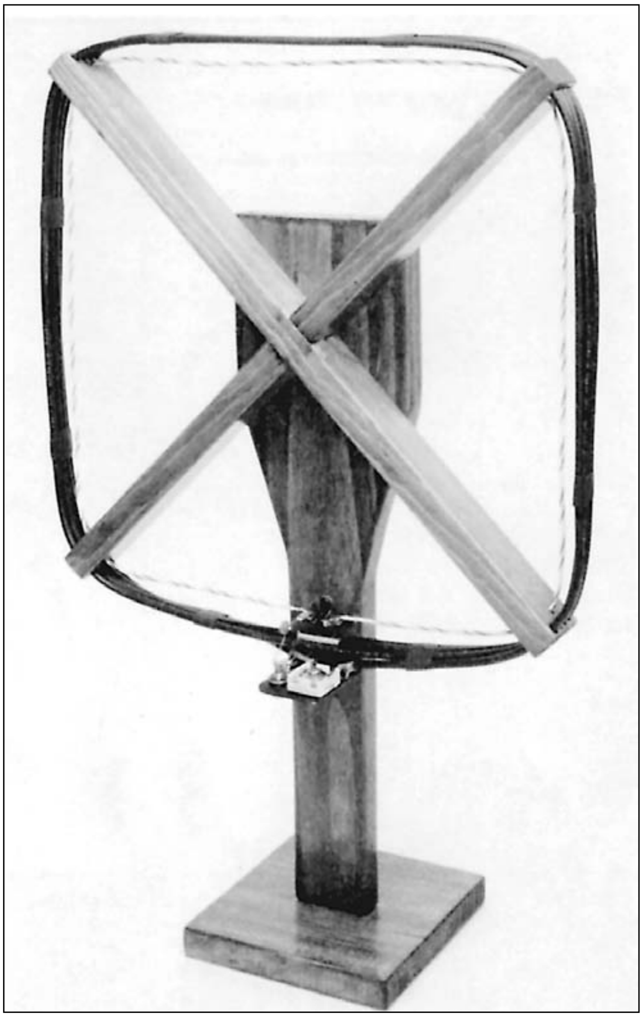

Fig 14—An assembled table-top version of the electrostatically shielded loop. RG-58 cable is used in its construction.

minals of L1. If the loop feed were attached in parallel with C1, a common practice, the chance for imbalance would be considerable.

L2 can be situated just inside or slightly outside of L1; a 1-inch separation works nicely. The receiver or preamplifier can be connected to terminals A and B of L2, as shown in Fig 12B. C2 controls the amount of coupling between the loop and the preamplifier. The lighter the coupling, the higher is the loop Q, the narrower is the frequency response, and the greater is the gain requirement from the preamplifier. It should be noted that no attempt is being made to match the extremely low loop impedance to the preamplifier.

A supporting frame for the loop of Fig 12 can be constructed of wood, as shown in **Fig 13**. The dimensions given are for a 1.8-MHz frame antenna. For use on 75 or 40 meters, L1 of Fig 12A will require fewer turns, or the size of the wooden frame should be made somewhat smaller than that of Fig 13.

SHIELDED FRAME LOOPS

If electrostatic shielding is desired, the format shown in **Fig 14** can be adopted. In this example, the loop conductor and the single-turn coupling loop are made from RG-58 coaxial cable. The number of loop turns should be sufficient to resonate with the tuning capacitor at the operating frequency. Antenna resonance can be checked by first connecting C1 (Fig 12A) and setting it at midrange. Then connect a small 3-turn coil to the loop feed terminals, and couple to it with a dip meter. Just remember that the pickup coil will act to lower the frequency slightly from actual resonance.

In the antenna photographed for Fig 14, the 1-turn coupling loop was made of #22 plastic-insulated wire. However, electrostatic noise pickup occurs on such a coupling loop, noise of the same nature that the shield on the main loop prevents. This can be avoided by using RG-58 for the coupling loop. The shield of the coupling loop should be opened for about 1 inch at the top, and each end of the shield grounded to the shield of the main loop.

Larger single-turn frame loops can be fashioned from aluminum-jacketed Hardline, if that style of coax is available. In either case, the shield conductor must be opened at the electrical center of the loop, as shown in **Fig 15** at A and B. The design example is for 1.8-MHz operation.

To realize the best performance from an electrostatically shielded loop antenna, you must operate it near to and directly above an effective ground plane. An automobile roof (metal) qualifies nicely for small shielded loops. For fixed-station use, a chicken-wire ground screen can be placed below the antenna at a distance of 1 to 6 feet.

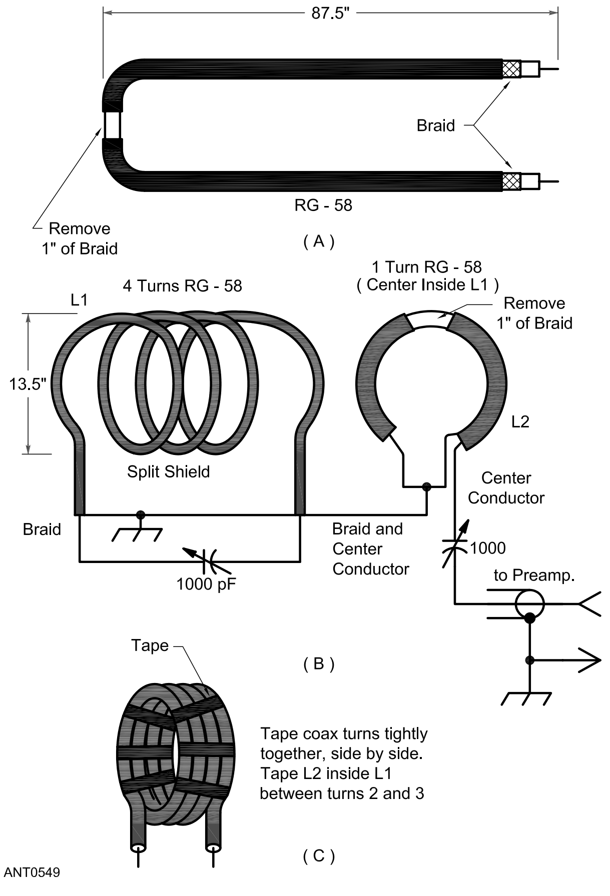
87.5"
Braid
RG - 58
Remove
1" of Braid
(A)
4 Turns RG - 58
L1
1 Turn RG - 58
(Center Inside L1)
Remove
1" of Braid
13.5"
L2
Split Shield
Center
Conductor
Braid
Braid and
Center
Conductor
1000
1000 pF
to Preamp.
Tape
(B)
Tape coax turns tightly
together, side by side.
Tape L2 inside L1
between turns 2 and 3
(C)
ANT0549

Fig 15—Components and assembly details of the shielded loop shown in Fig 14.

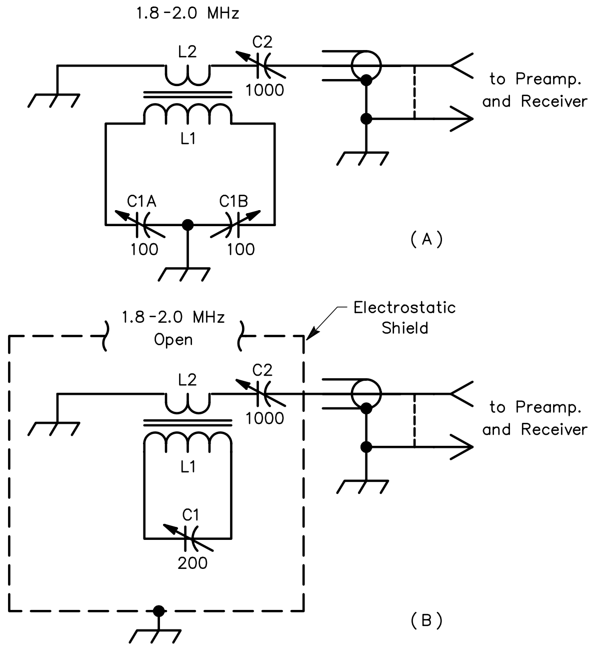
1.8 -2.0 MHz
L2
C2
1000
to Preamp.
and Receiver
L1
C1A
C1B
100
100
(A)
1.8 -2.0 MHz
Open
Electrostatic
Shield
L2
C2
1000
to Preamp.
and Receiver
L1
C1
200
(B)

Fig 16—At A, the diagram of a ferrite loop. C1 is a dual-section air-variable capacitor. The circuit at B shows a rod loop contained in an electrostatic shield channel (see text). A suitable low-noise preamplifier is shown in Fig 19.

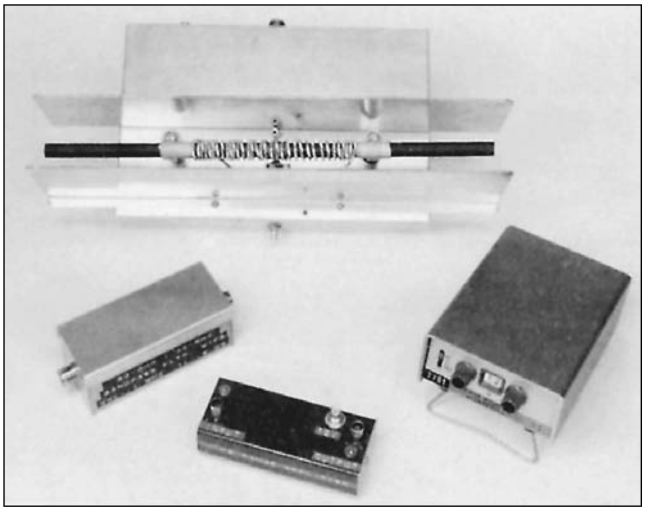

Fig 17—The assembly at the top of the picture is a shielded ferrite-rod loop for 160 meters. Two rods have been glued end to end (see text). The other units in the picture are a low-pass filter (lower left), broadband preamplifier (lower center) and a Tektronix step attenuator (lower right). These were part of the test setup used when the antenna was evaluated.

FERRITE-CORE LOOPS

Fig 16 contains a diagram for a rod loop (loopstick antenna). This antenna was also designed by Doug DeMaw, W1FB, and described in *QST* for July 1977. The winding (L1) has the appropriate number of turns to permit resonance with C1 at the operating frequency. L1 should be spread over approximately $^1/_3$ of the core center. Litz wire will yield the best Q, but Formvar magnet wire can be used if desired. A layer of 3M Company glass tape (or Mylar tape) is recommended as a covering for the core before adding the wire. Masking tape can be used if nothing else is available.

L2 functions as a coupling link over the exact center of L1. C1 is a dual-section variable capacitor, although a differential capacitor might be better toward obtaining optimum balance. The loop Q is controlled by means of C2, which is a mica-compression trimmer.

Electrostatic shielding of rod loops can be effected by centering the rod in a U-shaped aluminum, brass or copper channel, extending slightly beyond the ends of the rod loop (1 inch is suitable). The open side (top) of the channel can't be closed, as that would constitute a shorted turn and render the antenna useless. This can be proved by shorting across the center of the channel with a screwdriver blade when the loop is tuned to an incoming signal. The shield-braid gap in the coaxial loop of Fig 15 is maintained for the same reason.

Fig 17 shows the shielded rod loop assembly. This antenna was developed experimentally for 160 meters and uses two 7-inch ferrite rods, glued together end-to-end with epoxy cement. The longer core resulted in improved sensitivity for weak-signal reception. The other items in the photograph were used during the evaluation tests and are not pertinent to this discussion. This loop and the frame loop discussed in the previous section have bidirectional nulls, as shown in Fig 1A.

Obtaining a Cardioid Pattern

Although the bidirectional pattern of loop antennas

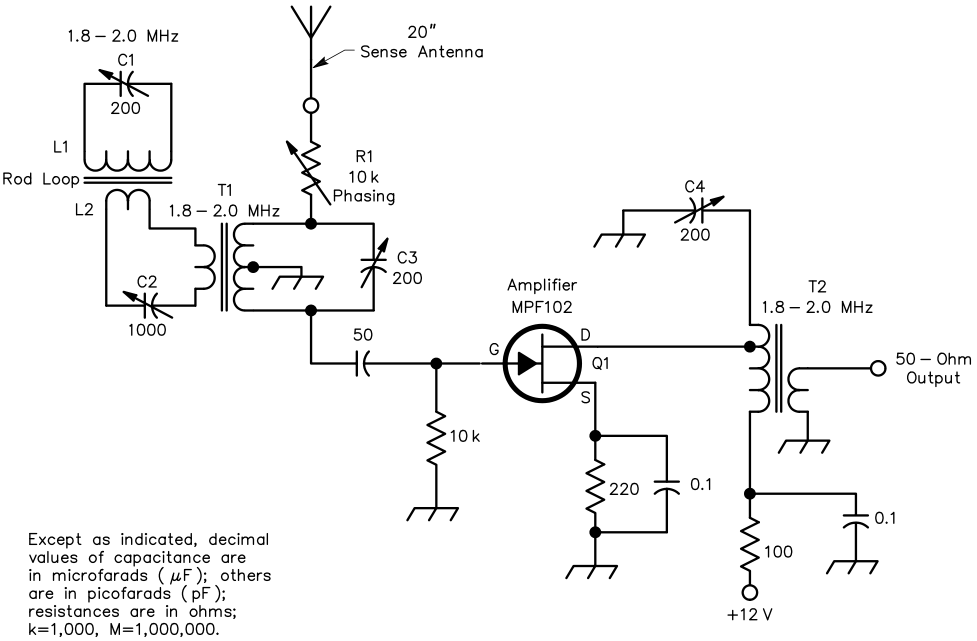

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.

Fig 18—Schematic diagram of a rod-loop antenna with a cardioid response. The sensing antenna, phasing network and a preamplifier are shown also. The secondary of T1 and the primary of T2 are tuned to resonance at the operating frequency of the loop. T-68-2 to T-68-6 Amidon toroid cores are suitable for both transformers. Amidon also sells ferrite rods for this type of antenna.

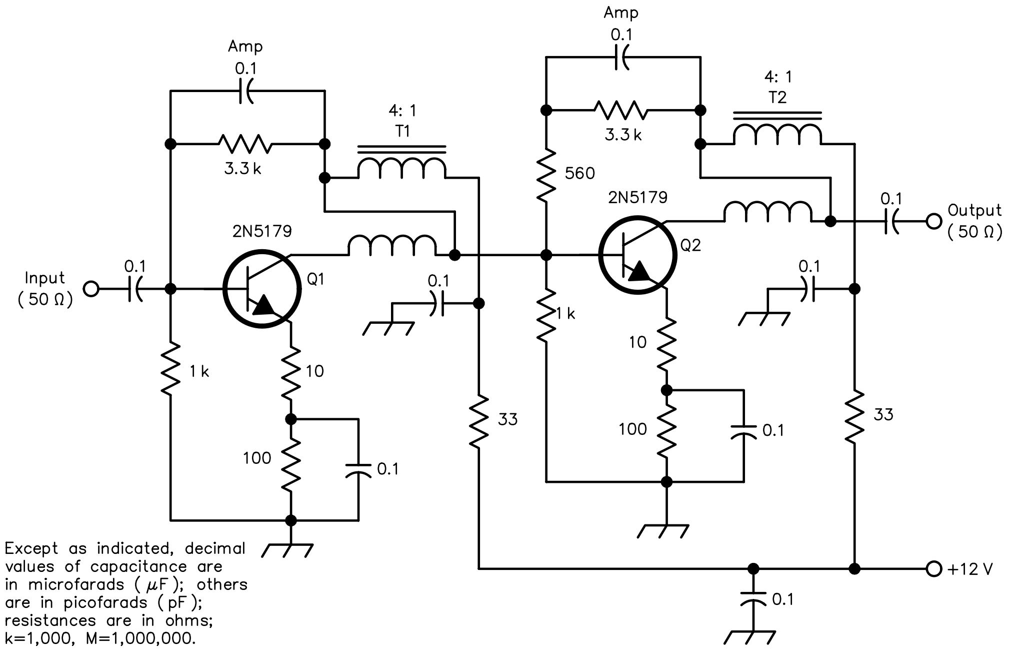

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.

Fig 19—Schematic diagram of a two-stage broadband amplifier patterned after a design by Wes Hayward, W7ZOI. T1 and T2 have a 4:1 impedance ratio and are wound on FT-50-61 toroid cores (Amidon) which have a μ_i of 125. They contain 12 turns of #24 enamel wire, bifilar wound. The capacitors are disc ceramic. This amplifier should be built on double-sided circuit board for best stability.

can be used effectively in tracking down signal sources by means of triangulation, an essentially unidirectional loop response will help to reduce the time spent finding the fox. Adding a sensing antenna to the loop is simple to do, and it will provide the desired cardioid response. The theoretical pattern for this combination is shown in Fig 1D.

Fig 18 shows how a sensing element can be added to a loop or loopstick antenna. The link from the loop is connected by coaxial cable to the primary of T1, which is a tuned toroidal transformer with a split secondary winding. C3 is adjusted for peak signal response at the frequency of interest (as is C4), then R1 is adjusted for minimum back response of the loop. It will be necessary to readjust C3 and R1 several times to compensate for the interaction of these controls. The adjustments are repeated until no further null depth can be obtained. Tests at ARRL Headquarters showed that null depths as great as 40 dB could be obtained with the circuit of Fig 18 on 75 meters. A near-field weak-signal source was used during the tests.

The greater the null depth, the lower the signal output from the system, so plan to include a preamplifier with 25 to 40 dB of gain. Q1 shown in Fig 18 will deliver approximately 15 dB of gain. The circuit of **Fig 19** can be used following T2 to obtain an additional 24 dB of gain. In the interest of maintaining a good noise figure, even at 1.8 MHz, Q1 should be a low-noise device. A 2N4416, an MPF102, or a 40673 MOSFET would be satisfactory. The sensing antenna can be mounted 6 to 15 inches from the loop. The vertical whip need not be more than 12 to 20 inches long. Some experimenting may be necessary in order to obtain the best results. Optimization will also change with the operating frequency of the antenna.

A SHIELDED LOOP WITH SENSING ANTENNA FOR 28 MHz

Fig 20 shows the construction and mounting of a simple shielded 10-meter loop. The loop was designed by Loren Norberg, W9PYG, and described in *QST* for April 1954. (See the Bibliography at the end of this chapter.) It is made from an 18-inch length of RG-11 coax (either solid or foam dielectric) secured to an aluminum box of any convenient size, with two coaxial cable hoods (Amphenol 83-1HP). The outer shield must be broken at the exact center. C1 is a 25-pF variable capacitor, and is connected in parallel with a 33-pF fixed mica padder capacitor, C3. C1 must be tuned to the desired frequency while the loop is connected to the receiver in the same way as it will be used for RDF. C2 is a small differential capacitor used to provide electrical symmetry. The lead-in to the receiver is 67 inches of RG-59 (82 inches if the cable has a foamed dielectric).

The loop can be mounted on the roof of the car with a rubber suction cup. The builder might also fabricate some kind of bracket assembly to mount the loop temporarily in the window opening of the automobile, allowing for loop rotation. Reasonably true bearings may be obtained through the windshield when the car is pointed in the direction of the hidden transmitter. More accurate bearings may be obtained with the loop held out the window and the signal coming toward that side of the car.

Sometimes the car broadcast antenna may interfere with accurate bearings. Disconnecting the antenna from the broadcast receiver may eliminate this trouble.

Sensing Antenna

A sensing antenna can be added to Norberg's loop above to determine which of the two directions indicated by the loop is the correct one. Add a phono jack to the top of the aluminum case shown in Fig 20. The insulated center terminal of the jack should be connected to the side of the tuning capacitors that is common to the center conductor of the RG-59 coax feed line. The jack then takes a short vertical antenna rod of the diameter to fit the jack, or a piece of #12 or #14 solid wire may be soldered to the center pin of a phono plug for insertion in the jack. The sensing antenna can be plugged in as needed. Starting with a length of about four times the loop diameter, the length of the sensing antenna should be pruned until the pattern is similar to that of Fig 1D.

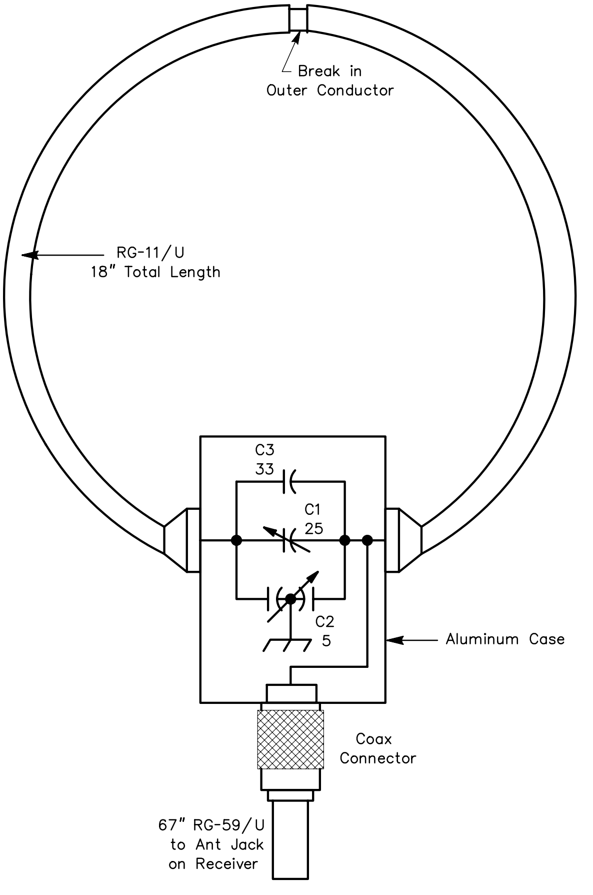

Fig 20—Sketch showing the constructional details of the 28-MHz RDF loop. The outer braid of the coax loop is broken at the center of the loop. The gap is covered with waterproof tape, and the entire assembly is given a coat of acrylic spray.

THE SNOOP LOOP—FOR CLOSE-RANGE HF RDF

Picture yourself on a hunt for a hidden 28-MHz transmitter. The night is dark, very dark. After you take off at the start of the hunt, heading in the right direction, the signal gets stronger and stronger. Your excitement increases with each additional S unit on the meter. You follow your loop closely, and it is working perfectly. You're getting out of town and into the countryside. The roads are unfamiliar. Now the null is beginning to swing rather rapidly, showing that you are getting close.

Suddenly the null shifts to give a direction at right angles to the car. With your flashlight you look carefully across the deep ditch beside the road and into the dark field where you know the transmitter is hidden. There are no roads into the field as far as you can see in either direction. You dare not waste miles driving up and down the road looking for an entrance, for each tenth of a mile counts. But what to do?—Your HF transceiver is mounted in your car and requires power from your car battery.

In a brief moment your decision is made. You park

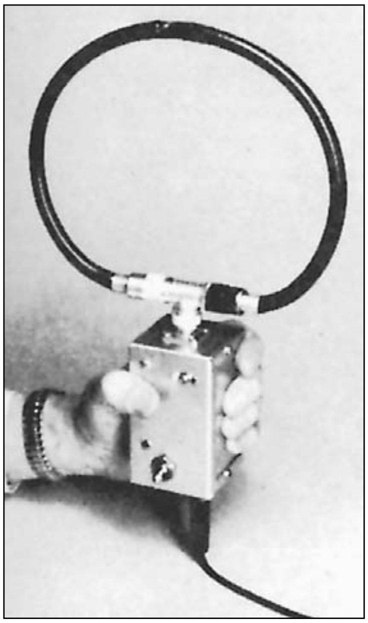

Fig 21—The box containing the detector and amplifier is also the "handle" for the Snoop Loop. The loop is mounted with a coax T as a support, a convenience but not an essential part of the loop assembly. The loop tuning capacitor is screwdriver adjusted. The on-off switch and the meter sensitivity control may be mounted on the bottom.

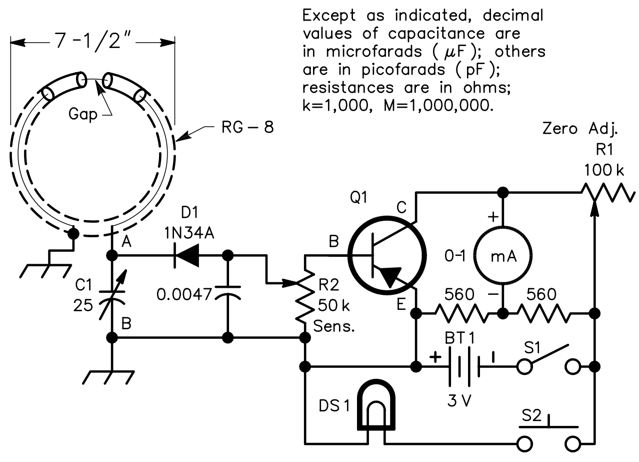
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
7-1/2"
Gap
RG-8
D1
1N34A
A
B
C1
25
0.0047
Q1
B
C
E
R2
50 k
Sens.
0-1
mA
+
560
−
560
Zero Adj.
R1
100 k
BT1
+
3 V
S1
S2
DS1

Fig 22—The Snoop Loop circuit for 28-MHz operation. The loop is a single turn of RG-8 inner conductor, the outer conductor being used as a shield. Note the gap in the shielding; about a 1-inch section of the outer conductor should be cut out. Refer to Fig 23 for alternative connection at points A and B for other frequencies of operation.

BT1—Two penlight cells.
C1—25-pF midget air padder.
D1—Small-signal germanium diode such as 1N34A or equiv.
DS1—Optional 2-cell penlight lamp for meter illumination, such as no. 222.
Q1—PNP transistor such as ECG102 or equiv.
R1—100-kΩ potentiometer, linear taper. May be PC-mount style.
R2—50-kΩ potentiometer, linear taper.
S1—SPST toggle.
S2—Optional momentary push for illuminating meter.

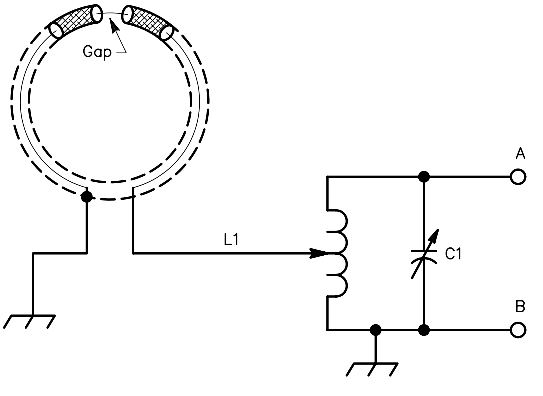

Fig 23—Input circuit for lower frequency bands. Points A and B are connected to corresponding points in the circuit of Fig 22, substituting for the loop and C1 in that circuit. L1-C1 should resonate within the desired amateur band, but the L/C ratio is not critical. After construction is completed, adjust the position of the tap on L1 for maximum signal strength. Instead of connecting the RDF loop directly to the tap on L1, a length of low impedance line may be used between the loop and the tuned circuit, L1-C1.

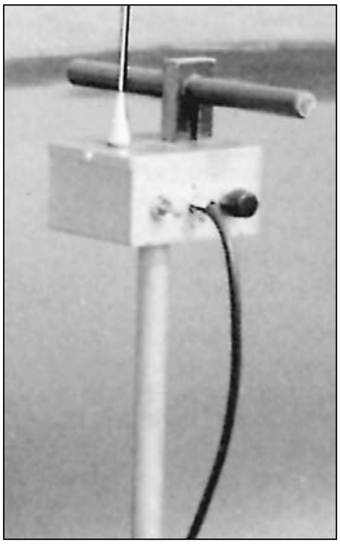

Fig 24—Unidirectional 3.5-MHz RDF using ferrite-core loop with sensing antenna. Adjustable components of the circuit are mounted in the aluminum chassis supported by a short length of tubing.

beside the road, take your flashlight, and plunge into the veldt in the direction your loop null clearly indicated. But after taking a few steps, you're up to your armpits in brush and can't see anything forward or backward. You stumble on in hopes of running into the hidden transmitter—you're probably not more than a few hundred feet from it. But away from your car and radio equipment, it's like the proverbial hunt for the needle in the haystack. What you really need is a portable setup for hunting at close range, and you may prefer something that is inexpensive. The Snoop Loop was designed for just these requirements by Claude Maer, Jr, W0IC, and was described in *QST* for February 1957. (See the Bibliography at the end of this chapter.)

The Snoop Loop is pictured in **Fig 21**. The loop itself is made from a length of RG-8 coax, with the shield broken at the top. A coax T connector is used for convenience and ease of mounting. One end of the coax loop is connected to a male plug in the conventional way, but the center conductor of the other end is shorted to the shield so the male connector at that end has no connection to the center prong. This results in an unbalanced circuit, but seems to give good bidirectional null readings, as well as an easily detectable maximum reading when the grounded end of the loop is pointed in the direction of the transmitter. Careful tuning with C1 will improve this maximum reading. Don't forget to remove one inch of shielding from the top of the loop. You won't get much signal unless you do.

The detector and amplifier circuit for the Snoop Loop

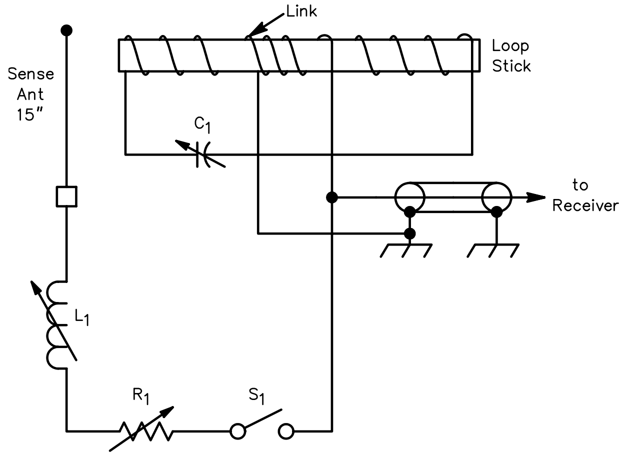

Fig 25—Circuit of the 3.5-MHz direction finder loop.

C1—140 pF variable (125-pF ceramic trimmer in parallel with 15-pF ceramic fixed.
L1—Approximately 140 µH adjustable (Miller No. 4512 or equivalent.
R1—1-kΩ carbon potentiometer.
S1—SPST toggle.
Loopstick—Approximately 15 µH (Miller 705-A, with original winding removed and wound with 20 turns of #22 enamel). Link is two turns at center. Winding ends secured with Scotch electrical tape. This type of ferrite rod may also be found in surplus transistor AM radios.

is shown in **Fig 22**. The model photographed does not include the meter, as it was built for use only with high-impedance headphones. The components are housed in an aluminum box. Almost any size box of sufficient size to contain the meter can be used. At very close ranges, reduction of sensitivity with R2 will prevent pegging the meter.

The Snoop Loop is not limited to the 10-meter band or to a built-in loop. **Fig 23** shows an alternative circuit for other bands and for plugging in a separate loop connected by a low-impedance transmission line. Select coil and capacitor combinations that will tune to the desired frequencies. Plug-in coils could be used. It is a good idea to have the RF end of the unit fairly well shielded, to eliminate signal pickup except through the loop. This little unit should certainly help you on those dark nights in the country. (Tip to the hidden-transmitter operator—if you want to foul up some of your pals using these loops, just hide near the antenna of a 50-kW broadcast transmitter!)

A LOOPSTICK FOR 3.5 MHz

Figs 24 through 26 show an RDF loop suitable for the 3.5-MHz band. It uses a construction technique that has had considerable application in low-frequency marine direction finders. The loop is a coil wound on a ferrite rod from a broadcast-antenna loopstick. The loop was designed by John Isaacs, W6PZV, and described in *QST* for June 1958. Because you can make a coil with high Q using a ferrite core, the sensitivity of such a loop is comparable to a conventional loop that is a foot or so in diameter. The output of the vertical-rod sensing antenna, when properly combined with that of the loop, gives the system the cardioid pattern shown in Fig 1D.

To make the loop, remove the original winding on the ferrite core and wind a new coil, as shown in **Fig 25**. Other types of cores than the one specified may be substituted; use the largest coil available and adjust the winding so that the circuit resonates in the 3.5-MHz band within the range of C1. The tuning range of the loop may be checked with a dip meter.

The sensing system consists of a 15-inch whip and an adjustable inductor that resonates the whip as a quarter-wave antenna. It also contains a potentiometer to control the output of the antenna. S1 is used to switch the sensing antenna in and out of the circuit.

The whip, the loopstick, the inductance L1, the capacitor C1, the potentiometer R1, and the switch S1 are all mounted on a 4 × 5 × 3-inch box chassis, as shown in **Fig 26**. The loopstick may be mounted and protected inside a piece of $^{1}/_{2}$-inch PVC pipe. A section of $^{1}/_{2}$-inch electrical conduit is attached to the bottom of the chassis box and this supports the instrument.

To produce an output having only one null there must a 90° phase difference between the outputs of the loop and sensing antennas, and the signal strength from each must be the same. The phase shift is obtained by tuning the sensing antenna slightly off frequency, using the slug in L1. Since the sensitivity of the whip antenna is greater than that of the loop, its output is reduced by adjusting R1.

Adjustment

To adjust the system, enlist the aid of a friend with a mobile transmitter and find a clear spot where the transmitter and RDF receiver can be separated by several hundred feet. Use as little power as possible at the transmitter. (Make very sure you don't transmit into the loop if you are using a transceiver as a detector.) With the test transmitter operating on the proper frequency, disconnect the sensing antenna with S1, and peak the loopstick using C1, while watching the S meter on the transceiver. Once the loopstick is peaked, no further adjustment of C1 will be necessary. Next, connect the sensing antenna and turn R1 to minimum resistance. Then vary the adjustable slug of L1 until a maximum reading of the S meter is again obtained. It may be necessary to turn the unit a bit during this adjustment to obtain a higher reading than with the loopstick alone. The last turn of the slug is quite critical, and some hand-capacitance effect may be noted.

Now turn the instrument so that one side (not an end) of the loopstick is pointed toward the test transmitter. Turn R1 a complete revolution and if the proper side was chosen a definite null should be observed on the S meter for one particular position of R1. If not, turn the RDF 180° and try again. This time leave R1 at the setting producing

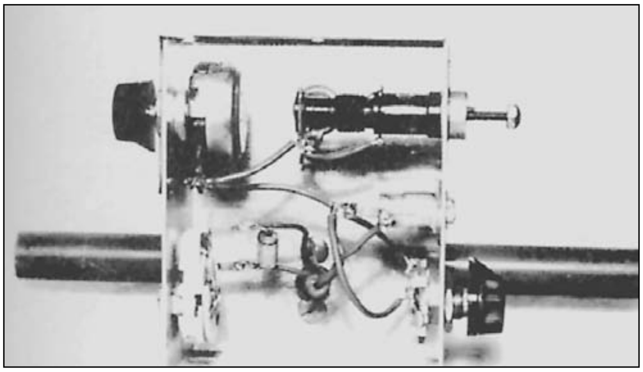

Fig 26—Components of the 3.5-MHz RDF are mounted on the top and sides of a channel-lock type box. In this view R1 is on the left wall at the upper left and C1 is at the lower left. L1, S1 and the output connector are on the right wall. The loopstick and whip mount on the outside.

the minimum reading. Now adjust L1 very slowly until the S-meter reading is reduced still further. Repeat this several times, first R1, and then L1, until the best minimum is obtained.

Finally, as a check, have the test transmitter move around the RDF and follow it by turning the RDF. If the tuning has been done properly the null will always be broadside to the loopstick. Make a note of the proper side of the RDF for the null, and the job is finished.

A 144-MHZ CARIOID-PATTERN RDF ANTENNA

Although there may be any number of different VHF antennas that can produce a cardioid pattern, a simple design is depicted in **Fig 27**. Two $^{1}/_{4}$-wavelength vertical elements are spaced one $^{1}/_{4}$-λ apart and are fed 90° out of phase. Each radiator is shown with two radials approximately 5% shorter than the radiators. This array was designed by Pete O'Dell, KB1N, and described in *QST* for March 1981.

Computer modeling showed that slight alterations in the size, spacing and phasing of the elements strongly impact the pattern. The results suggest that this system is a little touchy and that the most significant change comes at the null. Very slight alterations in the dimensions caused the notch to become much more shallow and, hence, less usable for RDF. Early experience in building a working model bore this out.

This means that if you build this antenna, you will find it advantageous to spend a few minutes to tune it carefully for the deepest null. If it is built using the techniques presented here, then this should prove to be a small task, well worth the extra effort. Tuning is accomplished by adjusting the length of the vertical radiators, the spacing between them and, if necessary, the lengths of the phasing harness that connects them. Tune for the deepest null on your S meter when using a signal source such as a moderately strong repeater.

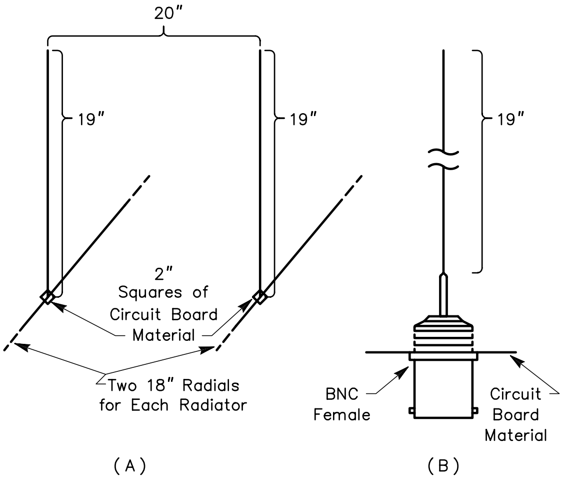

Fig 27—At A is a simple configuration that can produce a cardioid pattern. At B is a convenient way of fabricating a sturdy mount for the radiator using BNC connectors.

This should be done outside, away from buildings and large metal objects. Initial indoor tuning on this project was tried in the kitchen, which revealed that reflections off the appliances were producing spurious readings. Beware too of distant water towers, radio towers, and large office or apartment buildings. They can reflect the signal and give false indications.

Construction is simple and straightforward. Fig 27B shows a female BNC connector (RadioShack 278-105) that has been mounted on a small piece of PC-board material. The BNC connector is held upside down, and the vertical radiator is soldered to the center solder lug. A 12-inch piece of brass tubing provides a snug fit over the solder lug. A second piece of tubing, slightly smaller in diameter, is telescoped inside the first. The outer tubing is crimped slightly at the top after the inner tubing is installed. This provides positive contact between the two tubes. For 146 MHz the length of the radiators is calculated to be about 19 inches. You should be able to find small brass tubing at a hobby store. If none is available in your area, consider brazing rods. These are often available in hardware sections of discount stores. It will probably be necessary to solder a short piece to the top since these come in 18-inch sections. Also, tuning will not be quite as convenient. Two 18-inch radials are added to each element by soldering them to the board. Two 36-inch pieces of heavy brazing rod were used in this project.

The Phasing Harness

As shown in **Fig 28**, a T connector is used with two different lengths of coaxial line to form the phasing harness. This method of feeding the antenna is superior over

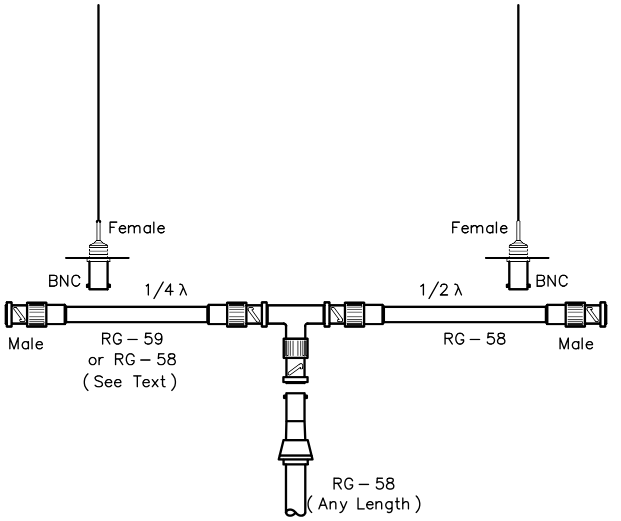

Fig 28—The phasing harness for the phased 144-MHz RDF array. The phasing sections must be measured from the center of the T connector to the point that the vertical radiator emerges from the shielded portion of the upside-down BNC female. Don't forget to take the length of the connectors into account when constructing the harness. If care is taken and coax with solid polyethylene dielectric is used, you should not have to prune the phasing line. With this phasing system, the null will be in a direction that runs along the boom, on the side of the 1/4-wavelength section.

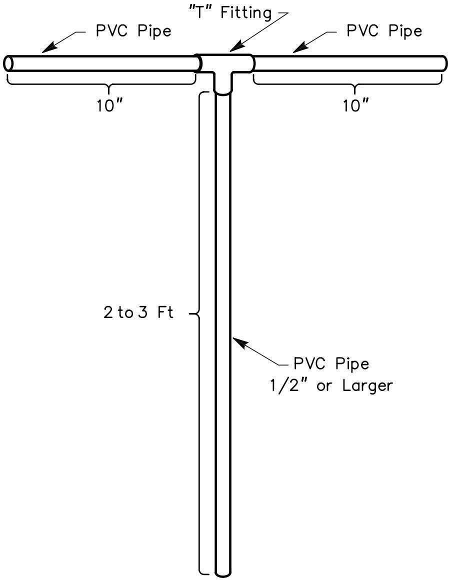

Fig 29—A simple mechanical support for the DF antenna, made of PVC pipe and fittings.

other simple systems to obtain equal currents in the two radiators. Unequal currents tend to reduce the depth of the null in the pattern, all other factors being equal.

The 1/2-wavelength section can be made from either RG-58 or RG-59, because it should act as a 1:1 transformer. With no radials or with two radials perpendicular to the vertical element, it was found that a 1/4-wavelength section made of RG-59 75-Ω coax produced a deeper notch than a 1/4-wavelength section made of RG-58 50-Ω line. However, with the two radials bent downward somewhat, the RG-58 section seemed to outperform the RG-59. Because of minor differences in assembly techniques from one antenna to another, it will probably be worth your time and effort to try both types of coax and determine what works best for your antenna. You may also want to try bending the radials down at slightly different angles for the best null performance.

The most important thing about the coax for the harness is that it be of the highest quality (well-shielded and with a polyethylene dielectric). The reason for avoiding foam dielectric is that the velocity factor can vary from one roll to the next—some say that it varies from one foot to the next. Of course, it can be used if you have test equipment available that will allow you to determine its electrical length. Assuming that you do not want to or cannot go to that trouble, stay with coax having a solid polyethylene dielectric. Avoid coax that is designed for the CB market or do-it-yourself cable-TV market. (A good choice is Belden 8240 for the RG-58 or Belden 8241 for the RG-59.)

Both RG-58 and RG-59 with polyethylene dielectric have a velocity factor of 0.66. Therefore, for 146 MHz a quarter wavelength of transmission line will be 20.2 inches × 0.66 = 13.3 inches. A half-wavelength section will be twice this length or 26.7 inches. One thing you must take into account is that the transmission line is the total length of the cable *and the connectors*. Depending on the type of construction and the type of connectors that you choose, the actual length of the coax by itself will vary somewhat. You will have to determine that for yourself.

Y connectors that mate with RCA phono plugs are widely available and the phono plugs are easy to work with. Avoid the temptation, however, to substitute these for the T and BNC connectors. Phono plugs and a Y connector were tried. The results with that system were not satisfactory. The performance seemed to change from day to day and the notch was never as deep as it should have been. Although they are more difficult to find, BNC T connectors will provide superior performance and are well worth the extra cost. If you must make substitutions, it would be preferable to use UHF connectors (type PL-259).

Fig 29 shows a simple support for the antenna. PVC tubing is used throughout. Additionally, you will need a T fitting, two end caps, and possibly some cement. (By not cementing the PVC fittings together, you will have the option of disassembly for transportation.) Cut the PVC for the dimensions shown, using a saw or a tubing cutter.

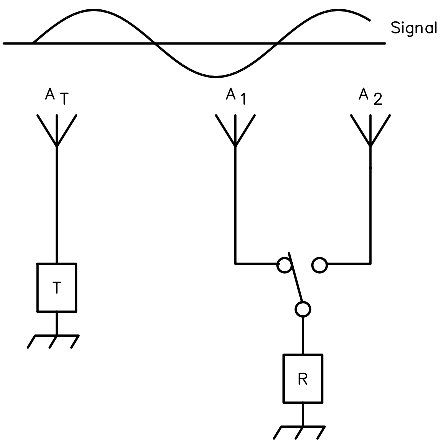

Fig 30—At the left, A_T represents the antenna of the hidden transmitter, T. At the right, rapid switching between antennas A_1 and A_2 at the receiver samples the phase at each antenna, creating a pseudo-Doppler effect. An FM detector detects this as phase modulation.

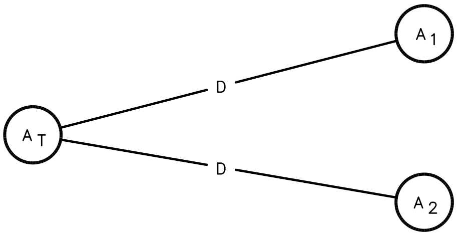

Fig 31—If both receiving antennas are an equal distance (D) from the transmitting antenna, there will be no difference in the phase angles of the signals in the receiving antennas. Therefore, the detector will not detect any phase modulation, and the audio tone will disappear from the output of the detector.

A tubing cutter is preferred because it produces smooth, straight edges without making a mess. Drill a small hole through the PC board near the female BNC of each element assembly. Measure the 20-inch distance horizontally along the boom and mark the two end points. Drill a small hole vertically through the boom at each mark. Use a small nut and bolt to attach each element assembly to the boom.

Tuning

The dimensions given throughout this section are those for approximately 146 MHz. If the signal you will be hunting is above that frequency, then the measurements should be a bit shorter. If you wish to operate below that frequency, then they will need to be somewhat longer. Once you have built the antenna to the rough size, the fun begins. You will need a signal source near the frequency that you will be using for your RDF work. Adjust the length of the radiators and the spacing between them for the deepest null on your S meter. Make changes in increments of $^{1}/_{4}$ inch or less. If you must adjust the phasing line, make sure that the $^{1}/_{4}$-wavelength section is exactly one-half the length of the half-wavelength section. Keep tuning until you have a satisfactorily deep null on your S meter.

THE DOUBLE-DUCKY DIRECTION FINDER

For direction finding, most amateurs use antennas having pronounced directional effects, either a null or a peak in signal strength. FM receivers are designed to eliminate the effects of amplitude variations, and so they are difficult to use for direction finding without looking at an S meter. Most modern HT transceivers do not have S meters.

This classic "Double-Ducky" direction finder (DDDF) was designed by David Geiser, WA2ANU, and was described in *QST* for July 1981. It works on the principle of switching between two nondirectional antennas, as shown in **Fig 30**. This creates phase modulation on the incoming signal that is heard easily on the FM receiver. When the two antennas are exactly the same distance (phase) from the transmitter, as in **Fig 31**, the tone disappears. (This technique is also known in the RDF literature as *Time-Difference-of-Arrival*, or TDOA, since signals arrive at each antenna at slightly different times, and hence at slightly different phases, from any direction except on a line perpendicular to and halfway in-between the two antennas. Another general term for this kind of two-antenna RDF technique is *interferometer*. *—Ed.*)

In theory the antennas may be very close to each other, but in practice the amount of phase modulation increases directly with the spacing, up to spacings of a half wavelength. While a half-wavelength separation on 2 meters (40 inches) is pretty large for a mobile array, a quarter wavelength gives entirely satisfactory results, and even an eighth wavelength (10 inches) is acceptable.

Think in terms of two antenna elements with fixed spacing. Mount them on a ground plane and rotate that ground plane. The ground plane held above the hiker's head or car roof reduces the needed height of the array and the directional-distorting effects of the searcher's body or other conducting objects.

The DDDF is bidirectional and, as described, its tone null points both toward and away from the signal origin. An L-shaped search path would be needed to resolve the ambiguity. Use the techniques of triangulation described earlier in this chapter.

Specific Design

It is not possible to find a long-life mechanical switch operable at a fairly high audio rate, such as 1000 Hz. Yet we want an audible tone, and the 400- to 1000-Hz range is perhaps most suitable considering audio amplifiers and average hearing. Also, if we wish to use the transmit function of a transceiver, we need a switch that will carry per-

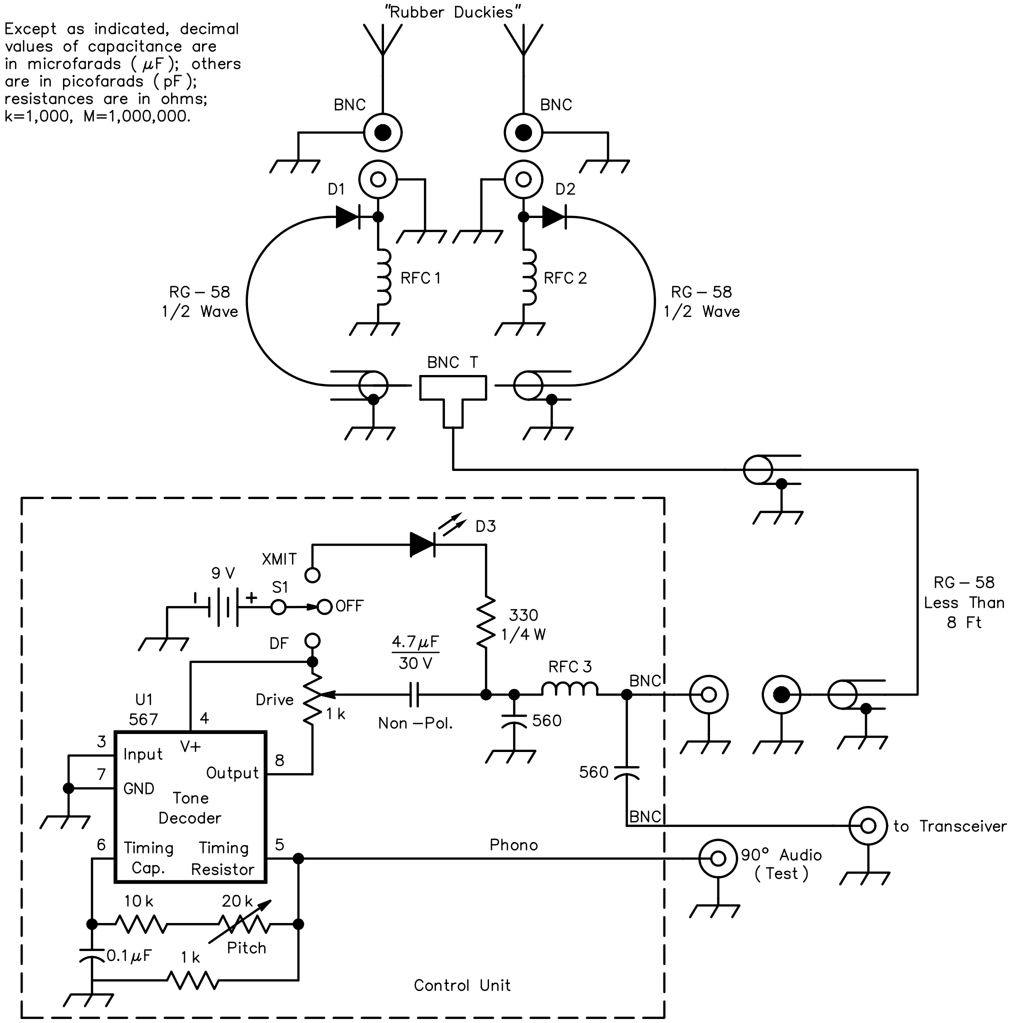
Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
"Rubber Duckies"
BNC
BNC
D1
D2
RFC 1
RFC 2
RG – 58 1/2 Wave
RG – 58 1/2 Wave
BNC T
RG – 58 Less Than 8 Ft
D3
XMIT
9 V
S1
OFF
DF
330 1/4 W
4.7μF 30 V
Non –Pol.
RFC 3
BNC
560
560
BNC
to Transceiver
U1 567
4
3
Input
V+
Drive
1 k
8
Output
7
GND
Tone Decoder
6
Timing Cap.
Timing Resistor
5
Phono
90° Audio (Test)
10 k
20 k
Pitch
0.1μF
1 k
Control Unit

Fig 32—Schematic diagram of the DDDF circuit. Construction and layout are not critical. Components inside the broken lines should be housed inside a shielded enclosure. Most of the components are available from RadioShack, except D1, D2, the antennas and RFC1-RFC3. These components are discussed in the text. S1—See text.

haps 10 W without much problem.

A solid-state switch, the PIN diode is used. The intrinsic region of this type of diode is ordinarily bare of current carriers and, with a bit of reverse bias, looks like a low-capacitance open space. A bit of forward bias (20 to 50 mA) will load the intrinsic region with current carriers that are happy to dance back and forth at a 148-MHz rate, looking like a resistance of an ohm or so. In a 10-W circuit, the diodes do not dissipate enough power to damage them.

Because only two antennas are used, the obvious approach is to connect one diode *forward* to one antenna, to connect the other *reverse* to the second antenna and to drive the pair with square-wave audio-frequency ac. **Fig 32** shows the necessary circuitry. RF chokes (Ohmite Z144, J. W. Miller RFC-144 or similar vhf units) are used to let the audio through to bias the diodes while blocking RF. Of course, the reverse bias on one diode is only equal to the forward bias on the other, but in practice this seems sufficient.

A number of PIN diodes were tried in the particular setup built. These were the Hewlett-Packard HP5082-3077, the Alpha LE-5407-4, the KSW KS-3542 and the Microwave Associates M/A-COM 47120. All worked well, but the HP diodes were used because they provided a slightly lower SWR (about 3:1).

A type 567 IC is used as the square-wave generator. The output does have a dc bias that is removed with a nonpolarized coupling capacitor. This minor inconvenience is more than rewarded by the ability of the IC to work well with between 7 and 15 volts (a nominal 9-V minimum is recommended).

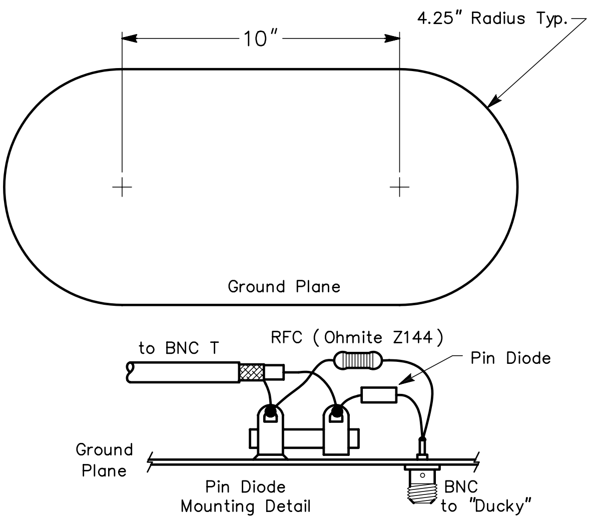

Fig 33—Ground-plane layout and detail of parts at the antenna connectors.

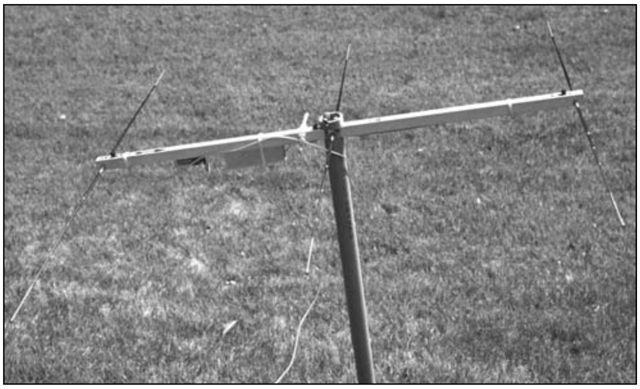

Fig 34—Photo of Fox-Hunting DF Twin 'Tenna set up as a horizontally polarized, 3-element Yagi.

The nonpolarized capacitor is also used for dc blocking when the function switch is set to XMIT. D3, a light-emitting diode (LED), is wired in series with the transmit bias to indicate selection of the XMIT mode. In that mode there is a high battery current drain (20 mA or so). S1 should be a center-off locking type toggle switch. An ordinary center-off switch may be used, but beware. If the switch is left on XMIT you will soon have dead batteries.

Cables going from the antenna to the coaxial T connector were cut to an electrical $^{1}/_{2}$ wavelength to help the open circuit, represented by the reverse-biased diode, look open at the coaxial T. (The length of the line within the T was included in the calculation.)

The length of the line from the T to the control unit is not particularly critical. If possible, keep the total of the cable length from the T to the control unit to the transceiver under 8 feet, because the capacitance of the cable does shunt the square-wave generator output.

Ground-plane dimensions are not critical. See **Fig 33**. Slightly better results may be obtained with a larger ground plane than shown. Increasing the spacing between the pickup antennas will give the greatest improvement. Every doubling (up to a half wavelength maximum) will cut the width of the null in half. A 1° wide null can be obtained with 20-inch spacing.

DDDF Operation

Switch the control unit to DF and advance the drive potentiometer until a tone is heard on the desired signal. Do not advance the drive high enough to distort or "hash up" the voice. Rotate the antenna for a null in the fundamental tone. Note that a tone an octave higher may appear.

If the incoming signal is quite out of the receiver linear region (10 kHz or so off frequency), the off-null antenna aim may present a fairly symmetrical AF output to one side, **Fig 35A**. It may also show instability at a sharp null position, indicated by the broken line on the display in Fig 35B. Aimed to the other side of a null, it will give a greatly increased AF output, Fig 35C. This is caused by the different parts of the receiver FM detector curve used. The sudden tone change is the tip-off that the antenna null position is being passed.

The user should practice with the DDDF to become acquainted with how it behaves under known situations of signal direction, power and frequency. Even in difficult nulling situations where a lot of second-harmonic AF exists, rotating the antenna through the null position causes a very distinctive tone change. With the same frequencies and amplitudes present, the quality of the tone (timbre) changes. It is as if a note were first played by a violin, and then the same note played by a trumpet. (A good part of this is the change of phase of the fundamental and odd harmonics with respect to the even harmonics.) The listener can recognize differences (passing through the null) that would give an electronic analyzer indigestion.

A FOX-HUNTING DF TWIN 'TENNA

Interferometers give sharp bearings, but they lack sensitivity for distant work. Yagis are sensitive, but they provide relatively broad bearings. This project yields an antenna that blends both on a single boom to cover both ends of the hunt. This is a condensation of a October 1998 *QST* article by R. F. Gillette, W9PE.

A good fox-hunting antenna must meet a number of criteria: (1) small size, (2) gain to detect weak signals and (3) high directivity to pinpoint the fox. Small antennas, however, do not normally yield both gain and direc-

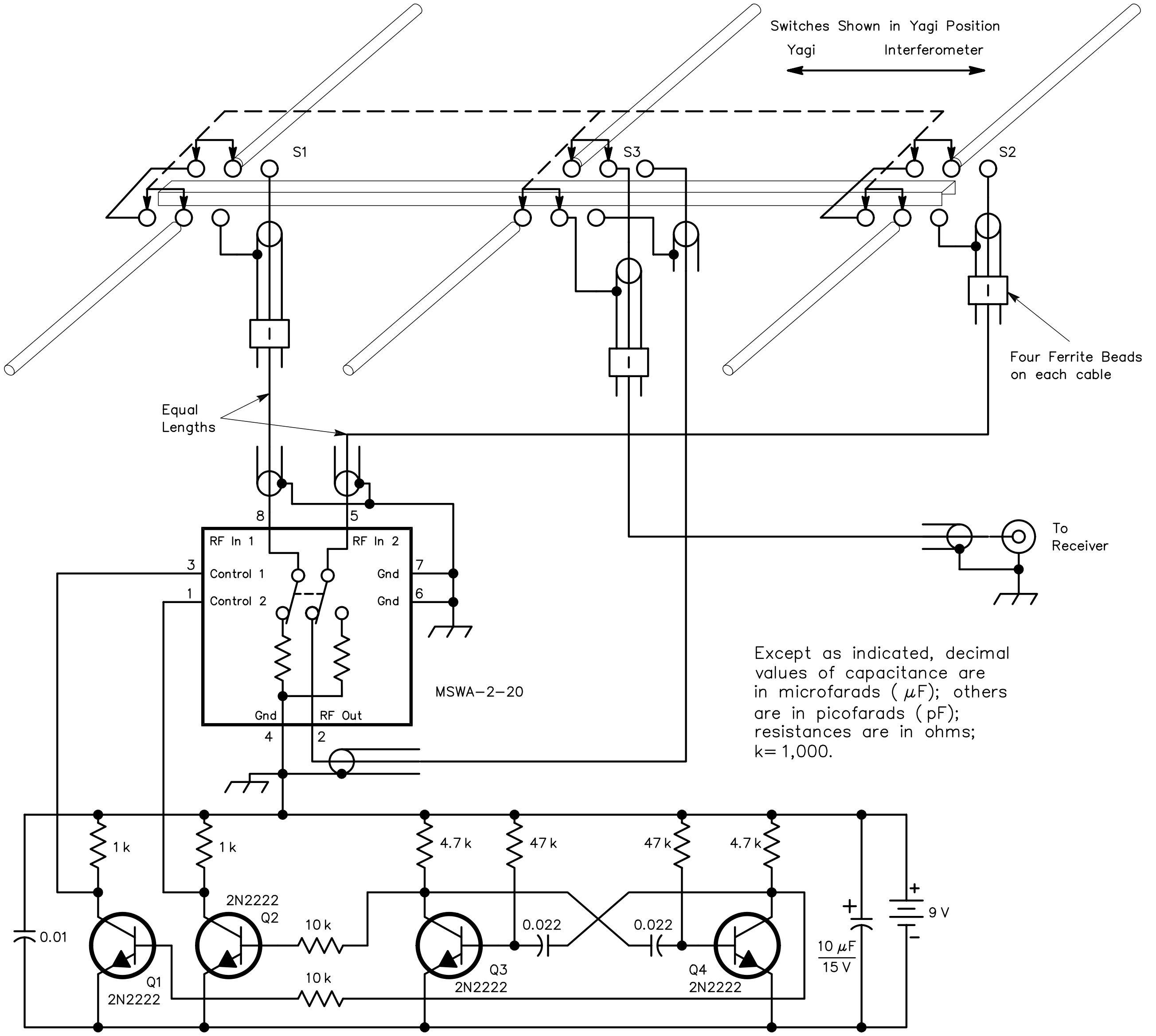
Switches Shown in Yagi Position
Yagi
Interferometer
S1
S3
S2
Four Ferrite Beads on each cable
Equal Lengths
8
5
RF In 1
RF In 2
3
Control 1
Gnd
7
1
Control 2
Gnd
6
MSWA-2-20
Gnd
RF Out
4
2
To Receiver
1 k
1 k
4.7 k
47 k
47 k
4.7 k
2N2222
Q2
10 k
0.01
Q1
2N2222
10 k
0.022
0.022
Q3
2N2222
Q4
2N2222
10 μF
15 V
+
9 V

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k= 1,000.

Fig 35—Schematic of the Yagi/interferometer antenna system.

Table 1
Yagi Design

Item	*Overall Length (Inches)*	*Boom to Element Tip (Inches)*
Director length	34.75	17.00
Director to Driven El. spacing	15.75	16.00
Driven El. length	37.75	18.50
Driven El. to Reflector spacing	15.75	16.00
Reflector length	40.75	20.00

*SWR less than 1.3:1 from 144.5 to 148 MHz

tivity. By combining two antennas, all three requirements are satisfied in a way that makes a nice build-it-yourself project.

This antenna uses slide switches to configure it as either a Yagi or a single-channel interferometer. When used as an interferometer, a GaAs RF microcircuit switches the FM receiver between two matched dipoles at an audio frequency. To make the antenna compact W9PE used hinged, telescopic whips as the elements; they collapse and fold parallel to the boom for storage.

The Yagi

The Yagi is a standard three-element design, based on 0.2-λ spacing between the director, the driven element and the reflector. It yields about 7 dBi gain and a front-to-back ratio of over 15 dB. Because a slide switch is used at the center of each element and the elements have small diameters, their resonant lengths are different from typical ones. **Table 1** shows the sizes used and **Fig 34** shows the Yagi.

To make sure that radiation from the coax does not affect the pattern, the author used some ferrite beads as coaxial choke-baluns. This also prevents objects near the coax from affecting signal-strength readings. The Yagi also has a low SWR; with uncalibrated equipment, he measured less than 1.3:1 over most of the 2-meter band.

This Yagi has a lot more gain than a "rubber ducky," but we need more directivity for fox hunting. That's where the interferometer comes in.

An Interferometer

To form the interferometer, the two end elements are converted to dipoles and the center element is disabled. When the three switches in **Fig 35** are thrown to the right, the feed line to the receiver is switched from the center element to the RF switch output, and the end elements are connected via feed lines to the RF switch inputs. With the Yagi's feed point open and the driven element equidistant from both interferometer antennas, the center element should have no effect on the interferometer. Nonetheless, it's easier to collapse the driven-element whips and get them out of the way than to worry about spacing.

Now if both interferometer coax cables are of equal length (between the antennas and switch) and the two antennas are the same distance from the transmitter (broadside to it), the signals from both antennas will be in phase. Switching from one antenna to the other will have no effect on the received signal. If one antenna is a little closer to the transmitter than the other, however, there will be a phase shift when we switch antennas.

When the antenna switch is at an audio rate, say 700 Hz, the repeated phase shifts result in a set of 700 Hz sidebands. At this point, all that was needed was a circuit to switch from one antenna to the other at an audio rate. W9PE chose a low-cost Mini-Circuits MSWA-2-20 GaAs RF switch driven by a simple multivibrator and buffer. The GaAs switch is rated to 2.0 GHz, hence this switching concept can easily be scaled to other ham bands. The PC board should work through the 440 MHz ham band. The author suggests adding a ground plane under the RF portion of the PC board and testing it before using it at a higher frequency.

The RF switch is controlled by a set of equal-amplitude, opposite-phase square waves: 0 V at one control port and –8 to –12 V at the other. (Mini Circuits is unclear about maximum voltages for this device. For safety, don't power it with more than 9 V.—*Ed*.) The opposite controls the other switch position. A 9-V battery was used as the power supply, with the positive terminal grounded. This results in a 0 V control signal to the RF switch when the buffer transistor is off and a Vsat (about 0.2 V less than the –9 V battery: –8.8 V) signal when the buffer transistor is saturated. The multi-vibrator has two outputs, and each drives a buffer resulting in the required equal-and-opposite-phase drive signals.

Table 2
Bill of Materials

Quantity	*Item*
3 ft	$^3/_4$-inch aluminum U channel
6 sets	insulated shoulder washers for elements
1	9 V battery
1	9 V battery connector
1	10 µF, 16 V electrolytic capacitor
1	0.01 µF, 25 V capacitor
2	0.022 µF, 25 V capacitor
2	1 kΩ $^1/_8$ W resistor
2	4.7 kΩ $^1/_8$ W resistor
2	10 kΩ $^1/_8$ W resistor
2	47 kΩ $^1/_8$ W resistor
2	1 kΩ $^1/_8$ W resistor
4	2N2222 transistors
1	Mini-Circuits MSWA-2-20 (Mini-Circuits Labs, 13 Neptune Ave, Brooklyn, NY 11235; tel 718-934-4500, 417-335-5935, fax 718-332-4661; e-mail **sales@minicircuits.com**; URL **www.minicircuits.com**)
3	DPDT slide switch ($1^1/_8$-inch, 29 mm, mounting centers), Stackpole, 3 A, 125 V used
10 ft	50 Ω coax (0.140-inch maximum OD)
1	coaxial connector (receiver dependent)
1	lot, mounting hardware
1	lot, heat-shrink tubing or equal
4	cable ties
1	2 × 3.5-inch single-sided fiberglass PC board
1	1-inch PVC conduit
12	#FB-20 ferrite beads 0.14-inch ID, 0.5-inch long (All Electronics Corp: PO Box 567, Van Nuys, CA 91408-0567; tel 888-826-5432, fax 818-781-2653, e-mail: **allcorp@allcorp.com**; URL **www.allcorp.com/**.)
6	$20^1/_2$-inch telescoping antenna elements (Nebraska Surplus, tel 402-346-4750; e-mail **grinnell@probe.net**)
1	Special resist film (Techniks Inc, PO Box 463, Ringoes, NJ 08551; tel 908-788-8249, fax 908-788-8837; e-mail **techniks@idt.net**; URL **www.techniks.com/**)

Circuit Construction

After he selected the Mini-Circuits RF switch, W9PE realized that its small size would be best handled with a simple PC board. He made the prototype boards with a photocopy transparency technique.

A power on-off switch was not used, as the 9-V bat-

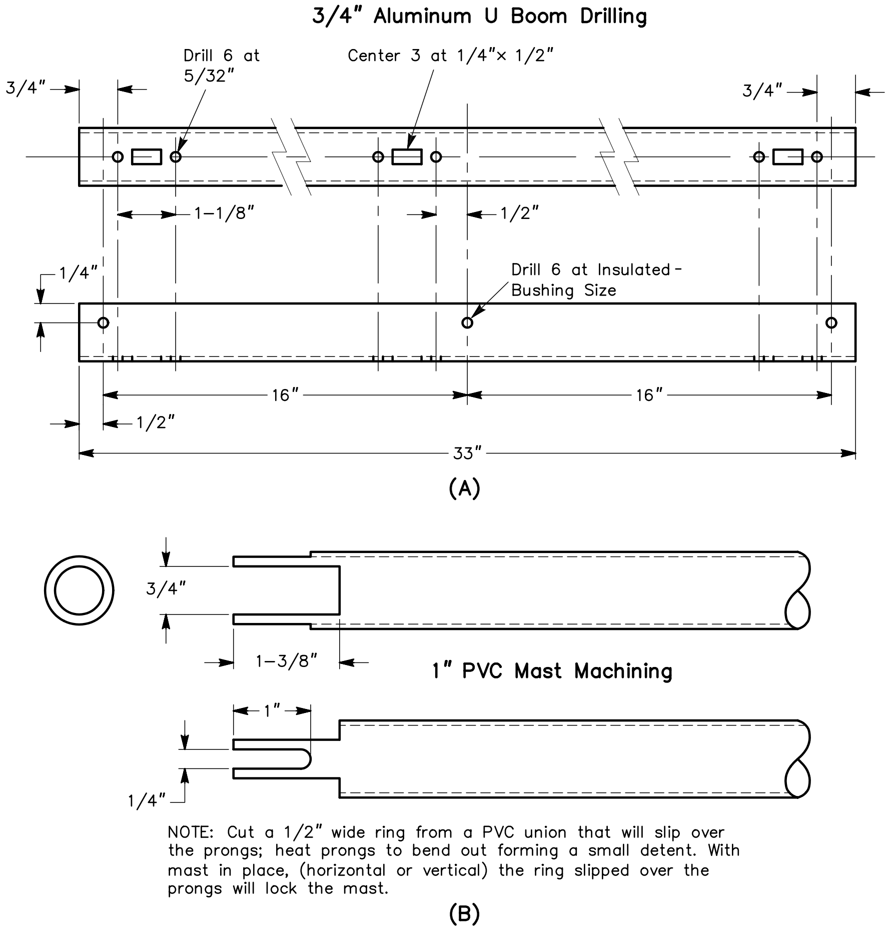

NOTE: Cut a 1/2" wide ring from a PVC union that will slip over the prongs; heat prongs to bend out forming a small detent. With mast in place, (horizontal or vertical) the ring slipped over the prongs will lock the mast.

Fig 36—Boom-drilling and mast-machining details.

tery connector serves the same function. The battery fits tightly in the $^3/_4$-inch U channel. W9PE covered the circuit board with a plastic-lined aluminum cover, but plastic film and some aluminum foil, provide the same function. A cable tie will strap either into the U channel.

Table 2 is a complete bill of materials. You can use any telescoping elements, providing that they extend to over 20 inches and have a mounting stud long enough to accommodate the insulated washers. As an alternate to the stud, they can have ends tapped to receive a screw for the insulated mounting. The author picked up his elements at a hamfest from the vendor listed; they are also available from most electronic parts houses. The Mini-Circuit RF switch is a currently available part.

Antenna Construction

Fig 35 shows the antenna schematic. It shows all three switches in the Yagi position; each would slide to the right for interferometer use. Slide switches work pretty well at 2 meters. Each of the elements is mounted to the boom with insulating washers, and a strip of copper stock connects each element to its slide switch. (You can substitute copper braid, solder wick, coax shield or any short, low resistance, low inductance conductor for the copper stock.) This switching arrangement allows you to switch the reflector and director from being parasitic elements (electrically continuous) to being dipoles (center fed).

Because the elements telescope, you can adjust the interferometer dipoles to exactly equal lengths each time you switch the antenna configuration from Yagi to interferometer. Again, choke baluns block RF on the outside of each element's coax.

Caution: Do not transmit when the RF switch is selected. Transmit only when in the Yagi configuration. RF power will destroy the Mini-Circuits RF switch. To be safe, lock out your transmit function. Most HTs have this capability. When using a mobile radio, disconnect the microphone. It is, however, safe to transmit in the Yagi configuration—which is nice for portable operating.

Fig 36 gives dimensions for drilling a standard $^3/_4$-inch aluminum U channel for the boom and shows how the author cut a 1-inch PVC pipe (plastic conduit) for a mast and a mast locking ring. If PVC conduit is not available in your area, PVC water pipe (and a PVC union for the locking ring) will work. This mast allows mounting the antenna for either vertical or horizontal polarization.

Be sure to test the plastic pipe you use for low RF loss. Do this by heating a sample in a microwave oven. Place a pipe sample and a glass of water in the oven. (The sample is not placed in the glass of water; the water keeps the microwave from operating without a load.) Bring the water to a boil, and then carefully check the sample's temperature. If the sample is not hot, its RF loss is low, and the plastic can be used.

Using the Antenna

When starting a hunt, set up the Yagi antenna by placing all switches in the Yagi position. Swing all of the telescoping elements perpendicular to the boom and set the whip lengths to achieve the proper element lengths, while keeping each element symmetrical about the boom. The cables or boom can be marked with the length data.

While the signal is weak, use the Yagi. It has 7 dBi gain, but its bearing resolution is only about 20°. When the signal gets stronger, use the interferometer. It has less gain, but its bearing resolution is better than 1°. If the fox transmitter begins overloading your receiver, collapse the whips (equally) to reduce the gain and continue triangulating. Near the transmitter, you should triangulate both horizontally (azimuth) and vertically (elevation). The antenna works both ways, and the transmitter may be located above or below you.

Antenna Alternative

As an alternative to the telescoping elements, George Holada, K9GLJ, suggested using fixed-length elements with banana plugs matched to banana jacks on the boom. Three pairs would be used for the Yagi, an extra driven-element pair for the interferometer mode and two short-element pairs to reduce the received signal level if an overload condition occurs. He also suggested a PVC boom allowing the elements to be stored inside the boom.

THE FOUR-WAY MOBILE DF SYSTEM

This innovative, yet simple, RDF antenna system was described in an article by Malcolm C. Mallette, WA9BVS, in November 1995 *QST*. It is derived from the TDOA design shown earlier in this chapter by David T. Geiser, WA2ANU, and by a design by Paul Bohrer, W9DUU. (See Bibliography.)

Direction-finding often involves two different activities: DFing on foot and DFing from a vehicle. Often, you must track the signal using a vehicle, then finish the hunt on foot. Whether on foot or in a vehicle, the primary problem you'll encounter when trying to locate the transmitter is multipath reception. Multipath reception involves receiving the same signal by more than one path, one signal from the true direction of the transmitter and others by reflected paths that may come from widely different directions. VHF and UHF signals bounce off almost any object and hide the true source of a signal. For example, if there's a large metal building north of you, a signal from the south may arrive from the north because the signal bounces off the building and back to you.

Multipath reception effects can be defeated by taking a number of readings from different positions and arriving at an average direction. While moving at road speeds in a vehicle, it's possible to take a number of readings from different positions and average them, and it's also possible to average a number of readings over a distance of travel by electronic means. The true bearing to the transmitter can usually be found by either method.

DFing equipment for use on foot is simpler than systems for use on a vehicle. While afoot, you can turn at will or easily rotate an antenna. Turning a vehicle while going down a street may result in a fender-bender if you're not careful!

The simplest DFing system to use while on foot consists of an S-meter-equipped hand-held receiver and a small, hand-held Yagi with an attenuator in line between the antenna and receiver. The attenuator keeps the S meter from pinning. The direction in which the beam points when the strongest signal is received is the direction of the transmitter. Of course, you'll want to take readings at several locations at least a wavelength apart to obtain an average heading, as multipath reception can still cause false readings in some locations.

Another approach that many hams have taken is the simple WA2ANU DDFF—it is now commonly known simply as the "buzzbox." Various commercial versions of the hand-held buzzbox system are available. Some sys-

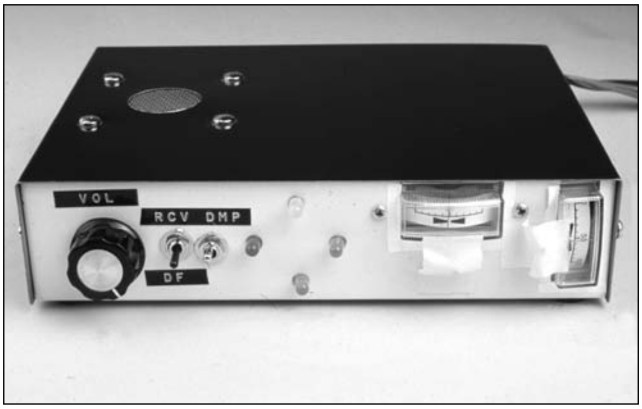

Fig 37—Front panel of the Four-Way DFer. At the extreme left of the front panel is the VOLUME control. Immediately to the right is the RCV/OFF/DF center-off toggle switch, with the damping (DMP) control switch nearby. Four LEDs mounted in a diamond pattern indicate signal direction: front (yellow), right (green), back (orange) and left (red). The horizontally mounted zero-center meter indicates left/right signal reception, the vertically mounted meter displays front/back signal reception. A small speaker is mounted on the top cover.

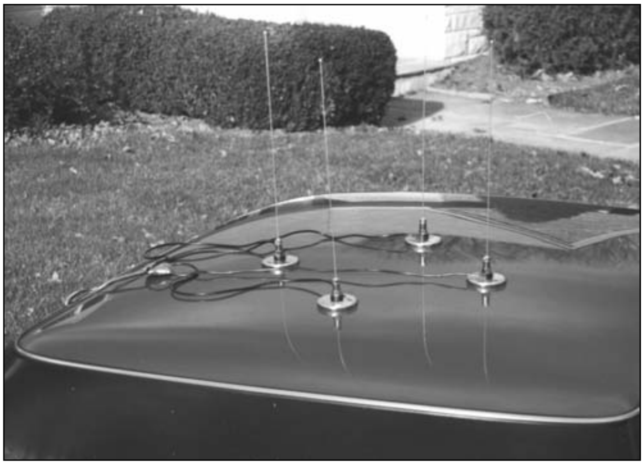

Fig 38—Placement of the four antennas on the author's car roof. The small object to the left of the antennas is the switch board.

tems have been upgraded to indicate whether the signal is arriving from the left or right of your position. The main drawback, however, to the buzzbox is that it's not as sensitive as a simple dipole and not nearly as sensitive as a beam.

In theory, you could take a buzzbox or Yagi/attenuator system in a car, stop periodically, get out and check the direction to the transmitter, then climb back in and drive off. Although this procedure works, it isn't very practical—it takes a long time to find the transmitter.

This design is for a left-right, front-back box (LRFB box) that indicates whether the received signal is to the left or right and whether it is to the front or back of the receiver. The location display consists of four LEDs arranged in a diamond pattern (see the title-page photo). When the top LED is on, the signal is coming from the front. When the top and right LEDs are on, the transmitter is between the front and the right. When only the right LED is on, the signal is directly to the right. When the bottom LED and right LED are on, the transmitter is to the right and to the back. The same pattern occurs around the clock. Therefore, four LEDs indicate eight directions. As most highways and streets have intersections that force a driver to choose moving straight ahead, right or left, the indication is sufficiently precise for practical transmitter hunting.

If the four-LED display is used alone, all parts can be obtained from your local Radio Shack store. Two zero-center 50-µA meters (50-0-50)—M1 and M2—can be used in addition to, or in place of, the LEDs, but RadioShack does not stock such meters. **Fig 37** shows the front panel layout of the LRFB.

The LRFB box uses four mag-mount $^1/_4$-λ antennas placed on the vehicle roof as shown in **Fig 38**. The whips in the mag mounts can be changed to $^1/_4$-λ 440-MHz whips and the antennas placed closer together when switching from 144-MHz to 440-MHz operation.

Circuit Description

See **Fig 39** in the following discussion (pages 26 and 27). U2, a 555 timer, generates a string of square-wave pulses at pin 3. The pulse frequency is determined by the setting of R4. The pulses are fed to the clock input (pin 14) of U3, a 4017 decade counter. On the first pulse, a positive voltage appears at U3, pin 3. On receipt of the second pulse from U2, pin 3 of U3 goes to ground and a positive voltage appears on pin 2. This sequence continues on successive pulses from U2 as pins 3, 2, 4, 7, 10, 1, 5, 6, 9 and 11 go positive in succession.

D1 through D5, and D6 through D10, OR the pulses. The result is that TP3 goes positive on the first pulse from U2, while TP2 is at 0 V. The next pulse of U2 results in TP2 going positive and TP3 going to 0 V. This sequence repeats as the counter goes around to make pin 3 positive again, and recycles.

Antenna Switching

U2 and U3 produce alternating pulses at TP2 and TP3. If we wanted only to alternately turn on and off two antennas, we could use the pulses at TP2 and TP3. The design ensures that the pulses at TP2 are the same length as the pulses at TP3. For the LRFB box, however, we need to switch between the left-right antennas many times, then switch between the front-back antennas many times.

Pin 12 of U3, CARRY OUT, emits a pulse every time U3 counts through its cycle of 10 pulses. The carry pulses from U3 go to U4, pin 14, the clock input of that 4017 counter. As U4 cycles, its output pins pulse; those pulses are, in effect, directed by D11 through D20.

As a result, TP4 is positive during the first 50 pulses from U2 and TP5 is positive during the second 50 pulses of U2. Q1 through Q6 form a quad AND gate. They AND the pulses at TP4 and TP5 with the alternating pulses at TP2 and TP3 so that the result is a pulse at the base of Q9, followed by a pulse at the base of Q10, a pulse at the base of Q9, and so on. The pulses alternate 25 times between Q9 and Q10. Then, as TP4 goes to 0 V, TP5 rises from 0 V to some positive voltage and the alternating pulses appear at the bases of Q7 and Q8. The pulses alternately go to the bases of Q7 and Q8 25 times. Then they alternate between the bases of Q9 and Q10 25 times. This pattern continues as long as the unit is in DF operation.

In **Fig 40**, two leads of a four-conductor-plus-ground cable to the antenna-switch board are connected to points A and B. The same pulses that turn Q7 and Q8 off and on turn on and off the left and right antennas. One of those two antennas is turned on and off in phase with Q7 and the other is turned on and off in phase with Q8. The PHASE switch, S3, determines which antenna is in phase with which transistor. Similarly, the two front/back antennas are turned on and off in phase with Q9 and Q10, and S4 determines which antenna is in phase with which transistor.

The pulses arriving at points A and B turn on and off the diodes connecting the coax of the left and right

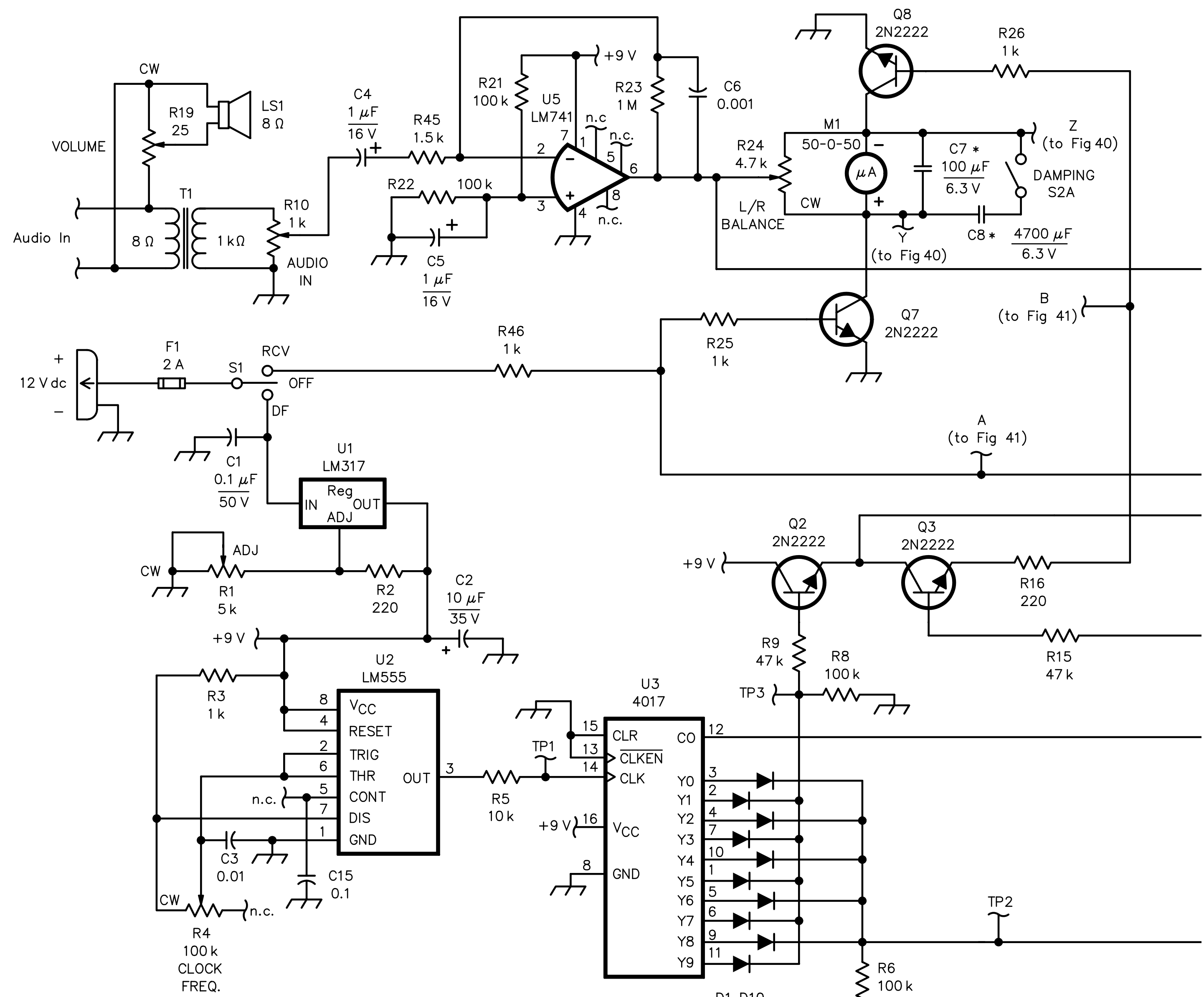
CW
VOLUME
R19
25
LS1
8 Ω
T1
8 Ω
1 kΩ
Audio In
R10
1 k
AUDIO IN
C4
1 µF
16 V
R45
1.5 k
R21
100 k
U5
LM741
n.c
n.c.
+9 V
R23
1 M
C6
0.001
R22
100 k
C5
1 µF
16 V
n.c.
Q8
2N2222
R26
1 k
M1
50-0-50
µA
C7 *
100 µF
6.3 V
Z
(to Fig 40)
DAMPING
S2A
R24
4.7 k
L/R BALANCE
CW
Y
(to Fig 40)
C8 *
4700 µF
6.3 V
B
(to Fig 41)
Q7
2N2222
R25
1 k
R46
1 k
+
12 V dc
–
F1
2 A
S1
RCV
OFF
DF
C1
0.1 µF
50 V
U1
LM317
Reg
IN
OUT
ADJ
A
(to Fig 41)
ADJ
CW
R1
5 k
R2
220
C2
10 µF
35 V
+9 V
Q2
2N2222
Q3
2N2222
R16
220
R15
47 k
R9
47 k
R8
100 k
TP3
R3
1 k
U2
LM555
VCC
RESET
TRIG
THR
CONT
DIS
GND
OUT
n.c.
C3
0.01
C15
0.1
CW
n.c.
R4
100 k
CLOCK FREQ.
TP1
R5
10 k
U3
4017
CLR
CLKEN
CLK
VCC
GND
CO
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
TP2
R6
100 k
D1-D10
1N914

Except as indicated, decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); resistances are in ohms; k=1,000, M=1,000,000.
n.c= not connected
∗See text

Fig 39—Unless otherwise specified, part numbers in parentheses are RadioShack. All fixed-value resistors are $^1/_4$-W, 5%-tolerance units. Equivalent parts can be substituted.

C1, C15—0.1-µF, 50-V (272-1069)
C2—10-µF, 35-V electrolytic capacitor (272-1025)
C3—0.01-µF, 25-V disc-ceramic capacitor (272-131)
C4, C5—1-µF, 16-V electrolytic capacitor (272-1434)
C6—0.001-µF, 25-V disc-ceramic capacitor (272-126)
C7, C9—100-µF, 6.3-V bipolar (nonpolarized) capacitor; Digi-Key P-1102, available from Digi-Key Corp, 701 Brooks Ave S, PO Box 677, Thief River Falls, MN 56701-0677, tel 800-344-4539, 218-681-6674; fax: 218-681-3880; RadioShack stocks 100-µF, 35-V axial (272-1016) and radial-lead (272-1028) electrolytic capacitors.
C8, C10—4700-µF, 6.3-V bipolar capacitor (made of five 1000-µF, 6.3-V bipolar capacitors), Digi-Key P1106. Standard 1000-µF, 35-V radial and axial-lead electrolytic capacitors are available from RadioShack; a 4700-µF, 35-V axial-lead electrolytic capacitor is also available (272-1022). Note that C7, C8, C9 and C10 are non-polarized capacitors because a small reverse voltage can appear across the meter and capacitors when the system is in use. Standard polarized electrolytics have been used in a number of units using this circuit (the same detector circuit used in W9DUU's unit) without any known ill effects, however.
D1-D20—1N914 silicon switching diode (276-1620 or 276-1122)
F1—2-A fuse (270-1007)
LS1—8-W speaker (40-245)

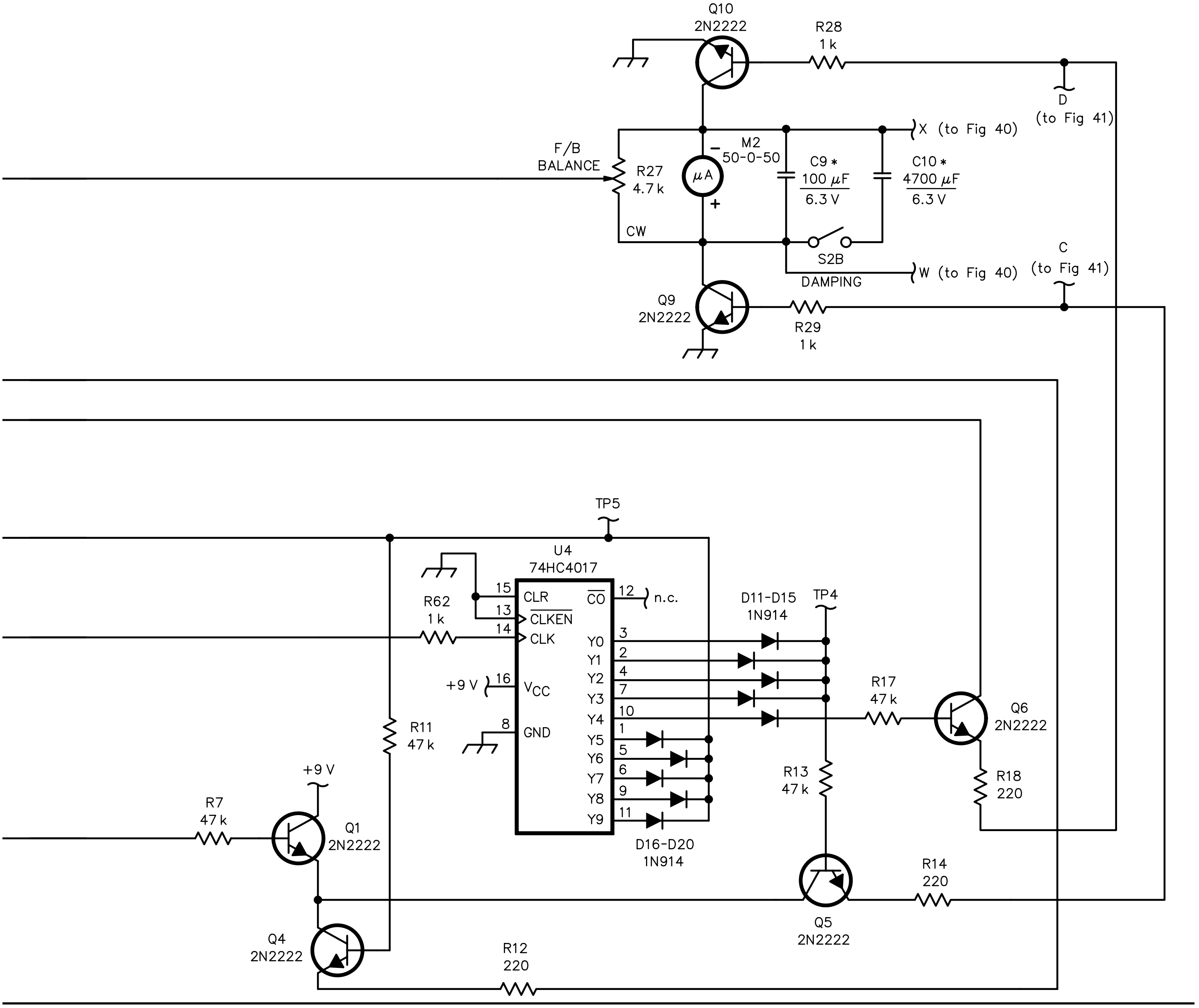
Q10
2N2222
R28
1 k
D
(to Fig 41)
X (to Fig 40)
F/B
BALANCE
R27
4.7 k
M2
50-0-50
μA
C9 *
100 μF
6.3 V
C10 *
4700 μF
6.3 V
CW
S2B
DAMPING
W (to Fig 40)
C
(to Fig 41)
Q9
2N2222
R29
1 k
TP5
U4
74HC4017
15
CLR
13
CLKEN
14
CLK
CO
12
n.c.
R62
1 k
+9 V
16
VCC
8
GND
Y0
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
3
2
4
7
10
1
5
6
9
11
D11-D15
1N914
TP4
D16-D20
1N914
R17
47 k
Q6
2N2222
R18
220
R13
47 k
R11
47 k
+9 V
R7
47 k
Q1
2N2222
Q5
2N2222
R14
220
Q4
2N2222
R12
220

M1, M2—Zero-center, 50-μA meter; optional—see text
Q1-Q10—MPS2222 or 2N2222 NPN silicon general-purpose transistors (276-2009)
R1, R24, R27—4.7 kΩ (271-281); note: many of the fixed-value resistors can be found in RadioShack resistor assortment packages 271-308 and 271-312.
R2, R12, R14, R16, R18—220 Ω (271-1330)
R3, R25, R26, R28, R29—1 kΩ (271-1321)
R4—100-kΩ trimmer potentiometer (271-284)
R5—10 kΩ (271-1335)
R6, R8, R21, R22—100 kΩ (271-1347)
R7, R9, R11, R13, R15, R17—47 kΩ (271-1342)
R10—1-kΩ trimmer potentiometer (271-280)
R19—25-Ω panel-mount potentiometer (271-265A)
R23—1 MΩ (271-1134)
R45—1.5 kΩ (part of 271-312 assortment)
S1—SPDT, center-off switch (275-325)
S2—DPDT switch (275-626)
T1—8-Ω to 1-kΩ audio-output transformer (273-1380)
U1—LM317T, 1.5-A, three-terminal, adjustable voltage regulator (276-1778)
U2—555 timer (276-1723)
U3, U4—4017 decade counter (276-2417)
U5—LM741 op amp (276-007)
Misc: two 8-pin IC sockets (276-1995); two 16-pin IC sockets (276-1992); experimenter's PC board (276-148) or FAR Circuits PC board set; enclosure; four mag-mount antennas, four-conductor shielded cable; in-line fuse holder (270-1281).

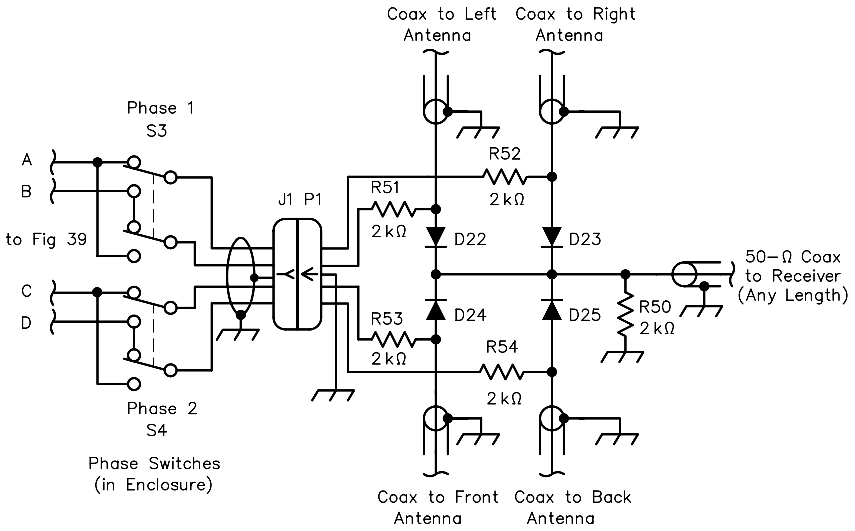
Coax to Left Antenna
Coax to Right Antenna
Phase 1
S3
A
B
to Fig 39
C
D
Phase 2
S4
Phase Switches (in Enclosure)
J1 P1
R51
2 kΩ
R52
2 kΩ
D22
D23
R53
2 kΩ
D24
D25
R50
2 kΩ
R54
2 kΩ
50-Ω Coax to Receiver (Any Length)
Coax to Front Antenna
Coax to Back Antenna

Fig 40—Schematic of the antenna switch board. Part numbers in parentheses are RadioShack. All fixed-value resistors are $^1/_4$-W, 5%-tolerance units. Equivalent parts can be substituted.

D22-D25—1N914 silicon switching diode (276-1620 or 276-1122)
J1—Six-pin female Molex connector (274-236 or 274-155)
P1—Six-pin male Molex connector (274-226 or 274-152)
R50-R54—2.2 kΩ (can be found in RadioShack resistor assortment packages 271-308 and 271-312); also available in pack of five (271-1325)
S3, S4—DPDT switch (275-626)

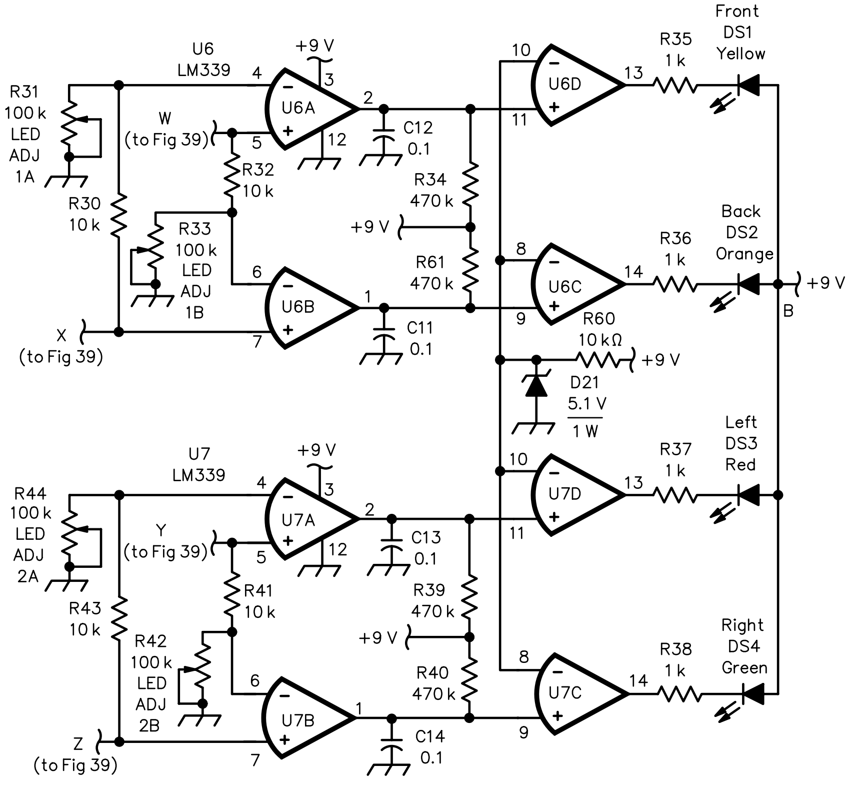
U6
LM339
+9 V
R31
100 k
LED
ADJ
1A
W
(to Fig 39)
U6A
4
3
2
5
12
C12
0.1
R32
10 k
R30
10 k
R33
100 k
LED
ADJ
1B
R34
470 k
+9 V
R61
470 k
U6B
6
7
1
C11
0.1
X
(to Fig 39)
10
U6D
11
13
R35
1 k
Front
DS1
Yellow
8
U6C
9
14
R36
1 k
Back
DS2
Orange
+9 V
B
R60
10 kΩ
+9 V
D21
5.1 V
1 W
U7
LM339
+9 V
R44
100 k
LED
ADJ
2A
Y
(to Fig 39)
U7A
4
3
2
5
12
C13
0.1
R41
10 k
R43
10 k
R42
100 k
LED
ADJ
2B
R39
470 k
+9 V
R40
470 k
U7B
6
7
1
C14
0.1
Z
(to Fig 39)
10
U7D
11
13
R37
1 k
Left
DS3
Red
8
U7C
9
14
R38
1 k
Right
DS4
Green

Fig 41—Schematic of the LED driver circuit. Part numbers in parentheses are RadioShack. All fixed-value resistors are $^1/_4$-W, 5%-tolerance units. Equivalent parts can be substituted.

C11-C14—0.1-µF, 25-V disc-ceramic capacitor (272-135)
D21—1N4733, 5.1-V, 1-W Zener diode (276-565)
DS1-DS4—LEDs; one each red (276-066); green (276-022); yellow (276-021); orange (276-012)
R30, R32, R41, R43—10 kΩ (271-1335)
R34, R39, R40, R61—470 kΩ (271-1354)
R35-R38—1 kΩ (271-1321)
R31, R33, R42, R44—100-kΩ trimmer potentiometer (271-284)
U6, U7—LM339 quad comparator (276-1712)
Misc: two 14-pin IC sockets (276-1999)

antennas to the receiver coax. This occurs 25 times, thereby switching the receiver between the left and right antennas 25 times. Similar switching then occurs between the front and the back antennas from pulses arriving at points C and D.

Detector Circuit

The detector circuit (back again in Fig 39) starts with U5, a 741 op amp that amplifies the receiver's audio output. The audio is fed into R24 and R27, two 4.7-kΩ pots. The zero-center, 50-µA meters across R24 and R27 are optional. The meters, as well as the LEDs, indicate front/back and left/right. Such meters can be expensive unless you find surplus meters, and they're not really necessary.

When the left/right antennas are active, one end of R24 is grounded by Q7 and Q8 on each alternate pulse.

If Q7 is conducting, Q8 is not conducting. On each pulse, one of the left/right antennas is turned on and one end of R24 is grounded. On the next pulse, the other left/right antenna is turned on and the other end of R24 is grounded. If there is a phase difference between the signal received by the left antenna and the signal received by the right antenna, a dc voltage is built up across R24. That voltage causes the quad comparator, U7 in **Fig 41** to turn on DS3 (red) or DS4 (green) L/R LEDs. If the optional left/right meter is installed, it deflects to indicate the direction as do the LEDs.

After 25 cycles, the left and right antennas are both turned off and the front and back antennas are cycled on and off 25 times with the same detection process, producing a voltage across R27 if there is a phase difference between the RF received by the front and the back antennas. That voltage across R27 causes quad comparator U6 to turn on DS1 or DS2.

C9 and C10, for the F/B detector, and C7 and C8, for the L/R detector, damp the voltage swings caused by multipath reception. To control damping, S2A and S2B switch the 4700-µF capacitors in or out of the circuit. You want the greatest amount of damping when you drive through an area with a lot of multipath propagation (as from buildings); a lot of damping helps under those circumstances.

Construction

The prototype was built using a pad-per-hole RadioShack board. However, a PC board makes construction a lot faster. Far Circuits offers a printed-circuit project on their Web site: **www.cl.ais.net/farcir**). Except for the optional meters and the nonpolarized capacitors, most parts are available from RadioShack.

First build the power supply so you can power the unit from your car or another 12-V source. Apply 12 V to the DFer and adjust R1 until U1's output is +9 V. (You can use a 9-V battery and omit the power-supply section, but you'd better take along a spare battery when you go DFing.)

Install U2 and its associated parts. Power up and turn on S1. A string of pulses should appear at TP1. If you have a frequency counter, set R4 for a pulse frequency of 2200 Hz at TP1. If you don't have a counter, connect a 0.1-µF capacitor from TP1 to headphones or a small speaker and set R4 for a tone of about 2 kHz. Later, you'll adjust the clock so the unit works with the passband of your receiver.

Turn off S1 and remove the power source. Install U3 and its diodes. Pin 12 of U3 need not be connected yet. Apply power and turn on S1. At TP2 and TP3, you should find alternating 1100-Hz pulses. If you have a dual-trace scope, you can see that the pulses alternate. If you have a single-trace scope, connect TP3 to TP2 and to the scope input and the trace will appear as a solid line as there is a pulse at either TP2 or TP3 at all times. If you don't have a scope, the tone in a speaker or earphones from TP2 or TP3 will sound half as high (about 1 kHz) as the tone at TP1.

Turn off S1. Install U4 and its diodes. Note that pin 12 of U3 is connected to pin 14 of U4. At TP4 and TP5, there should be long pulses—five times longer than the pulses at TP1, and the pulses should alternate between TP4 and TP5. The pulse frequency should be about 44 Hz. Power down and turn off S1. Install the remaining circuit components. When you power up, 25 alternating pulses should appear at A and B, then 25 alternating pulses should appear at C and D. Use a scope to verify that.

If you're not using the optional panel meters, connect a voltmeter across R24 (L/R BALANCE), using the lowest dc-voltage range. Note that neither end of R24 is grounded. With no audio input, adjust R24 until there is no voltage across it. Do the same with R27 (F/B BALANCE). If you use the optional meters, adjust R24 and R27 so there's no current shown on either meter.

Power down and assemble the rest of the circuit. With power applied, but with no audio input, adjust R31, 33, 42 and 44 so that the four LEDs (DS1 through DS4) are off. The objective of the following adjustments is to get the red and green LEDs to turn on with the same voltage amplitude, but opposite polarity, across R24. Move R24's wiper so a low positive voltage appears across R24, as indicated by the voltmeter connected across R24 or movement of the panel-meter needle. Adjust R44 (LED ADJ 1A) and R42 (LED ADJ 1B) so that the green LED (DS4) comes on when the voltage goes positive at one end of

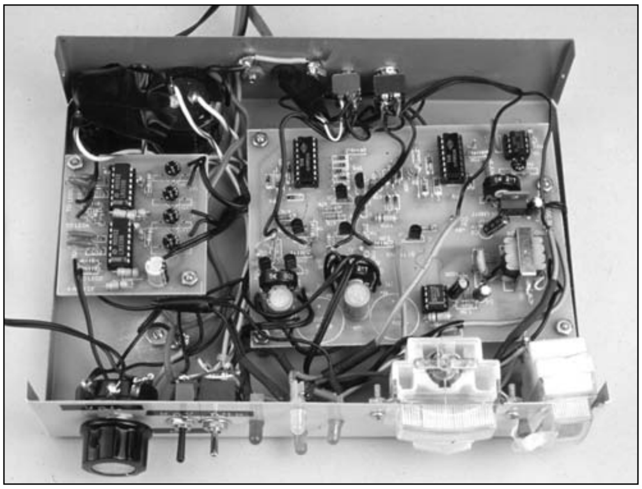

Fig 42—An inside view of one DF unit built into a $2 \times 8 \times 5^3/_4$-inch (HWD) box. Because of the height restriction, the two 4700-µF damping capacitors (C8 and C10) are not mounted on the PC board, but near the rear panel behind the smaller of the two PC boards. One of the 4700-µF damping capacitors is a standard electrolytic, the other is a parallel combination of five 1000-µF, 6.3-V bipolar (nonpolarized) capacitors wrapped in electrical tape.

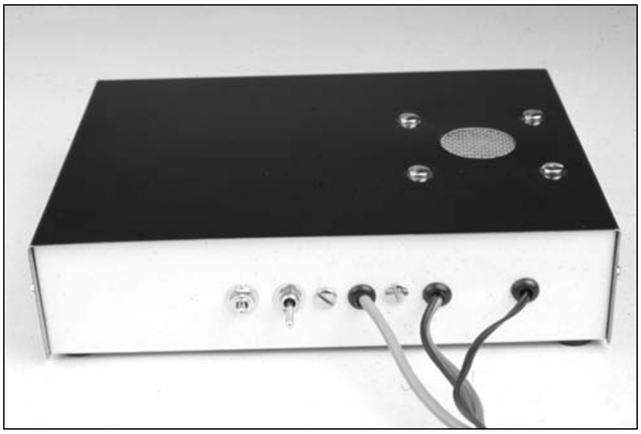

Fig 43—The rear panel of the DF unit supports the two DPDT PHASE toggle switches. Grommets in the panel holes allow abrasion-free passage of the antenna, dc-power and audio cables. The dc power cord is outfitted with an inline fuse holder and a male Jones plug. A six-pin female Molex connector (five pins are used) feeds the four antennas. The audio-input cable is terminated in a $^1/_8$-inch diameter male plug.

R24, but goes off when R24 is adjusted for 0 V across R24.

Next, adjust R24 for a slight negative-voltage indication and adjust R42 and R44 so that the red LED (DS3) comes on, but extinguishes when the voltage across R24 is 0 V. When you're done, adjusting R24's wiper slightly one way should illuminate the red LED. Both LEDs should be off when there's no voltage across R24; rotating R24's wiper slightly in the opposite direction should turn on the green LED.

Connect the voltmeter across R27 or use the panel meter as an indicator. With no audio input, adjust R27 so that there's no voltage across R27. Adjust R31 (F/B LED ADJ 1A) and R33 (F/B LED 1B) so that a F/B LED (DS1) is on when there is a slight positive voltage across R27 and the other F/B LED (DS2) is on when there is a negative voltage across R27.

Switch Board

Assemble the switch board for the four mag-mount antennas (see Fig 40). You can use half of a RadioShack dual pad-per-hole PC board (RS 276-148) as a platform. Lead length is critical only on this board and in the length of the coax from the switching board to the antennas, so avoid wire-wrap construction here.

Feed Lines

The coax lines from the switch board to each of the four antennas must be of equal length. The coax lengths should be long enough to permit each antenna to be placed slightly less than $^1/_4$ λ from its counterpart, at the lowest frequency of operation. (There's a local belief that $27^1/_2$ inches is the best length. The author used that length and it works, but other lengths might work as well.) An attempt to locate the switch board inside the vehicle and run equal-length 12-foot-long cables to the antennas failed. Keep the switch board on the vehicle's roof.

Use the same type of coax for all lines—that is, don't mix foam and polyethylene dielectric coax on the antenna lines. If the velocity factor of the lines is not equal, a phase shift in the signals will exist even when the transmitter is dead ahead and it will lead you astray.

Solder the coax from the antennas directly to the board—don't use connectors. From the switch board to the receiver, use 50-Ω coax of any type and length. Equip the receiver end of the line with a connector that mates with your receiver's antenna-input jack. Make the four-conductor shielded cable from the main DFer box to the switch board long enough to reach from the LRFB box's operating position to the roof of the car. Construct the antennas so that the whips can be changed easily for use on any frequency in which you're interested.

Mechanical Assembly

Mount the finished PC boards in a metal box of your choice. You can follow the construction method used in the prototype (see **Fig 42** and **Fig 43**), but do ensure that R4 can be adjusted easily with a tuning tool from outside the box. A hole drilled in the box at the right point will suffice. Arrange the LEDs in a diamond pattern on the front panel, with the left LED (red) to the left, the right LED (green) on the right, and the front and back LEDs at the top and the bottom.

Wrap the switch board with tape to waterproof it, or place it in a watertight box. Arrange the mag-mount antennas on top of the vehicle in a diamond shape. The distance between each antenna pair should be less than $^1/_4$ λ at the operating frequency. (In limiting the distance between antennas in a pair—F/B or L/R—to less than $^1/_4$ λ, the author followed W9DUU's example.)

Final Adjustments

It's best to start on 2 meters. Install the 2-meter whips in the antenna bases. Mount the antennas on the top of your vehicle. Identify the left and right antenna bases with L and R, and mark the front and back antenna bases, too.

A good way to find out if the antennas are properly installed is to short either the left or right antenna with a clip lead from the whip to the metal base; the L/R meter will deflect one way and the left or right LED will light. If you short one of the front and back antennas with a clip lead, the front or back LED will come on. If you short one of the L/R antennas and it makes the meter go to the right, it does not necessarily mean you have the PHASE switch in the proper position. That depends on the relative phase determined by the number of audio stages in your receiver, each of which may contribute a shift of 180°.

Connect the coax from the switch board to your FM receiver. It must be an FM receiver; an AM VHF or AM aircraft receiver won't suffice. Connect the audio output of your 2-meter receiver to the LRFB box audio input. If you're using a transceiver, disable its transmit function by removing the mike. You don't want to transmit into the switch board!

Turn on the receiver and place the LRFB box in receive by setting S1 to the RCV position. Center R10. Only one of the antennas will be turned on and the DF operation will be disabled. Back off the squelch and notice that you hear the audio from the speaker of the LRFB box. Switch to an unused simplex channel. Have a friend with an HT stand 20 feet or so in front of the vehicle. With the receiver in the car turned off, turn S1 to DF. The four LEDs should be off. If an LED lights, adjust R24 and R27 for zero voltage. If the LED is still on, readjust R31 and R33, or R42 and R44 as explained earlier. When all four LEDs are off with the antennas connected, no audio from the receiver and no RF input signal, you're ready. Turn on your receiver and have your friend transmit on the 2-meter frequency (simplex) that your receiver is set to. When he transmits, one or more of the LEDs should illuminate. Ignore the front/back LEDs, but check to see if the right or left LED is on. If the right LED is on and your friend is standing to the right of the center of the vehicle front, all is well. Have him walk back and forth in front of the vehicle and notice that when he is to the right, the green LED turns on, and when he's to the left, the red LED glows. If the indications are reversed, that is, if the LRFB box indicates left when your friend is to your right, reverse the position of PHASE switch S3.

When the L/R indicators are working properly, have your friend walk back and forth between a position 20 feet to the front and right of the car and a position 20 feet away to the rear and right of the car. The right LED should stay on, but the front LED should be on when the HT is in front of the antennas, and the back LED should be on when the HT is behind the antennas. If the front/back indications are reversed, reverse the position of PHASE switch S4.

While receiving a signal from your friend's HT, adjust R4 for maximum deflection on any one of the two optional meters, or on a voltmeter (set on the lowest dc voltage range) connected across R24 or R27. Look for the maximum deflection of the meter needle as the meter swings both ways while the signal source moves from back to front or left to right (depending, of course, on which panel meter you're looking at or which resistor—R24 or R27—your voltmeter is across). The audio passband of FM receivers varies, and if you switch from an ICOM IC-27 to an ICOM IC-W2A, for example, you'll have to change U2's master clock frequency. You may encounter receivers that don't require readjusting R4, but such readjustment should be expected. When changing receivers, you may also need to change the position of PHASE switches S3 and S4. This may also be necessary when changing from UHF to VHF, or VHF to UHF on the same receiver, as the number of receiver stages (and, hence, the audio phase) may change from band to band.

Audio Level Adjustments

With meter damping (S2) off, adjust R10 so that you have full deflection of one of the two meters (or your multimeter) with the signal source at a 45° angle from the vehicle (halfway between ahead and right), and a reasonable audio level from the speaker. Turn the rig's volume control all the way up to ensure that the audio circuit doesn't overload. If it overloads, the meters won't deflect. From zero to full blast, the meter should deflect more and more (unless the signal is straight ahead or exactly left or right). If you're using an H-T, maximum deflection of the meter should occur before the volume control is $^3/_4$ of maximum, or before the volume control is at $^1/_2$ of maximum if you're using a mobile rig. If R10 is properly adjusted, turning the volume control to maximum won't cause the meter to fall back toward zero. If increasing the volume causes the meter to deflect less, then R10's setting is too high.

Now have your friend walk around the vehicle with the HT transmitting and notice that the LEDs indicate the signal direction. On 2 meters in a clear field, the indications should be correct 80 or 90% of the time. The erroneous readings that occasionally occur are due to multipath propagation caused by the irregular shape of the vehicle. Slight adjustments in the positions of the L/R and F/B antennas may be necessary to make the zero points fall directly in front of the vehicle (neither left nor right LED on) and at the center of the antennas (neither front nor back LED on).

Try the same procedure on 440 MHz. You may have to flip the PHASE switches when you move to another band, even when using the same receiver. Remember that the total length of the 440-MHz antennas must be $^1/_4\ \lambda$ or less, and the antennas must be placed less than $^1/_4\ \lambda$ apart. The results on 440 MHz probably won't be as consistent as the results on 2 meters, as there is likely to be a lot more multipath propagation caused by the irregular shape of the vehicle.

Fox Hunting

Before heading out to find the fox, check to be certain the LRFB box is working properly. Tune to the fox's frequency and drive off. Turning on the DAMPING switch stabilizes the indication as you drive along. Follow indications generated as you travel over the road or street. If an indication is constant for 15 or 20 seconds while you're moving down the road, it's probably the true direction to the fox. It's possible to have a reflection from a mountain over a long distance down the road, however. When you can hear the fox with no antenna, it's time to get out of the car, switch to the hand-held system and hunt for the fox on foot.

If you switch to a new receiver, you may have to readjust R4, CLOCK FREQ. That's because both receivers may not have the same audio bandwidth. Author WA9BVS forgot this while chasing a balloon once and the results were comical. When chasing balloons with ham radio transmitters, the readings you get are likely to be confusing when the balloon is at a high angle with respect to the plane of the car top. Use a hand-held Yagi to verify the balloon location. Even with a simple buzzbox, you should be able to find a keyed transmitter.

A TAPE-MEASURE BEAM FOR RDFING

Joe Leggio, WB2HOL, designed this antenna while searching for a beam with a really great front-to-back ratio to use in hidden transmitter hunts. It exhibits a very clean pattern and is perfect for RDF use. You can construct this beam using only simple hand tools, and it has been duplicated many times.

WB2HOL's first design requirement was to be able to get in and out of his car easily when hunting for a hidden transmitter. He accomplished this by using steel "tape-measure" elements, which fold easily when putting the antenna into a car and yet are self supporting. They also hold up well while crashing through the underbrush on a fox hunt.

WB2HOL decided to use three elements to keep the boom from getting too long. He used inexpensive

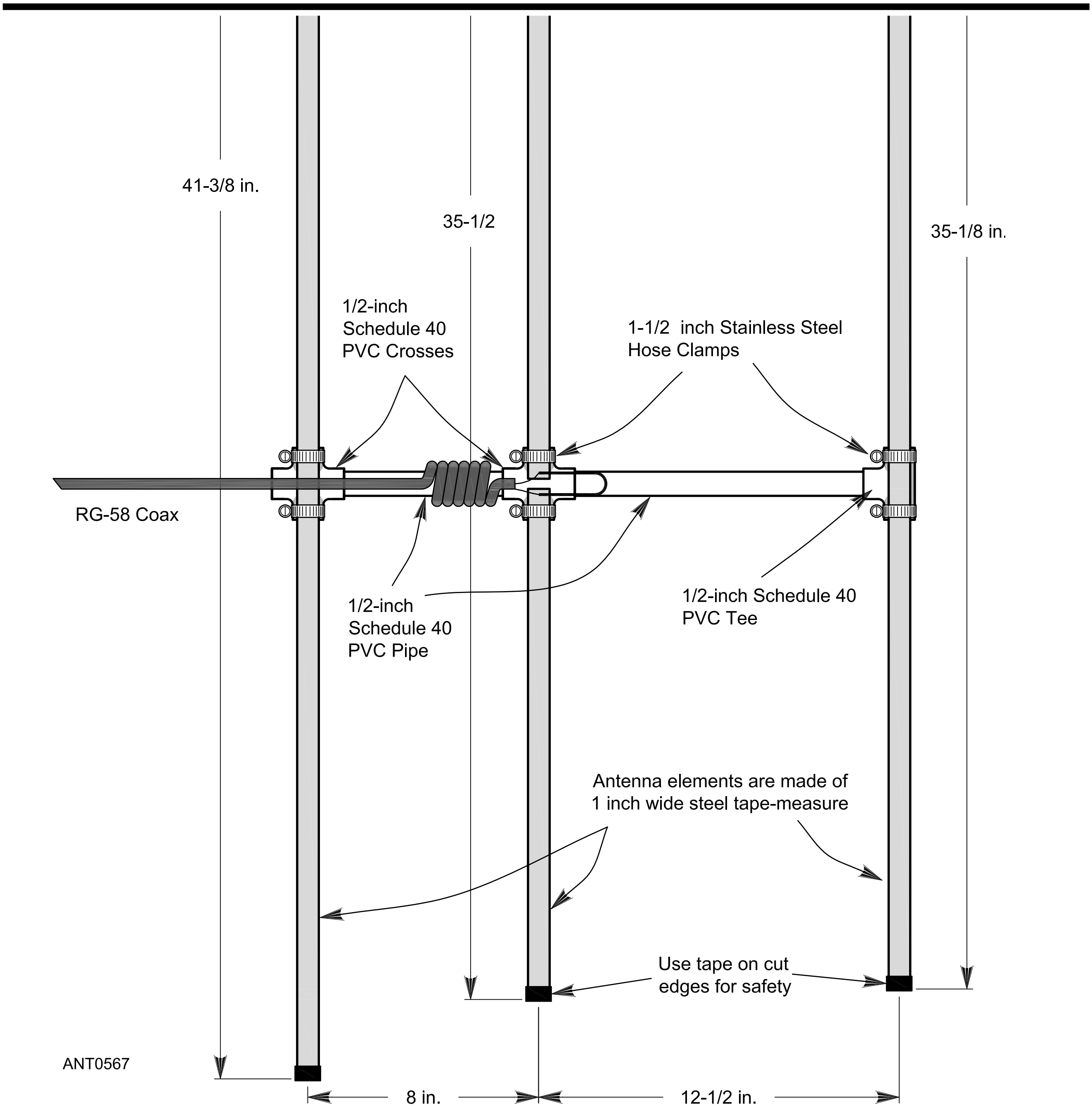
41-3/8 in.
35-1/2
35-1/8 in.
1/2-inch
Schedule 40
PVC Crosses
1-1/2 inch Stainless Steel
Hose Clamps
RG-58 Coax
1/2-inch
Schedule 40
PVC Pipe
1/2-inch Schedule 40
PVC Tee
Antenna elements are made of
1 inch wide steel tape-measure
Use tape on cut
edges for safety
ANT0567
8 in.
12-1/2 in.

Fig 44—Tape-measure beam dimensions.

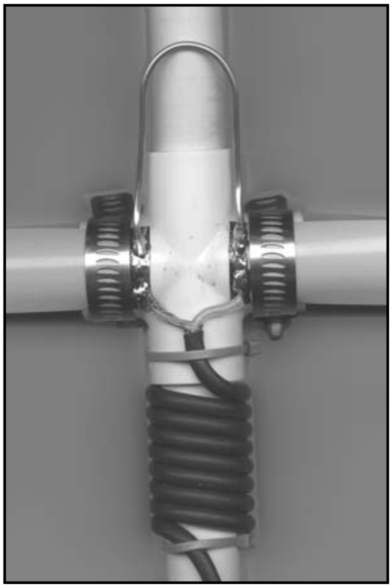

Fig 45—Photo of driven-element mounted to PVC tee using hose clamps. The hairpin match wires is shown here soldered to the tape-measure elements, along with the RG-58 feed line.

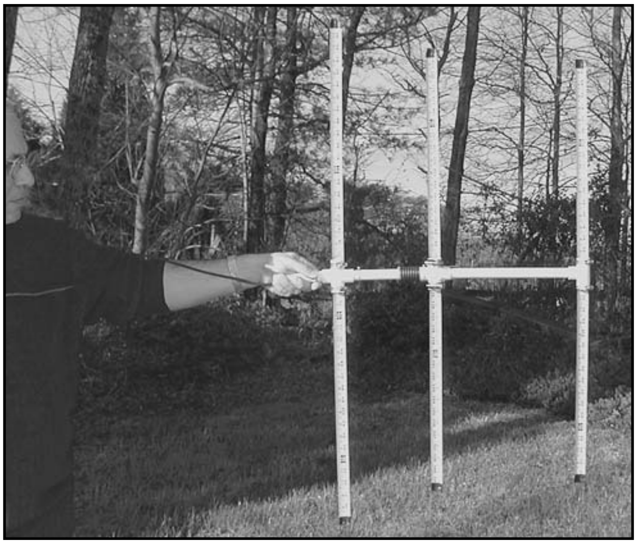

Fig 46—Photo of complete tape-measure beam, ready to hunt foxes!

Schedule-40 PVC pipe, crosses and tees that can be found at any hardware store for the boom and element supports.

He used a simple hairpin match, consisting of a 5-inch length of #14 solid wire bent into the shape of a U, with the two legs about $^{3}/_{4}$ inch apart. This gave in a very good match across the 2-meter band after he tweaked the distance (1-inch on his prototype) between the halves of the driven element for minimum SWR.

You can cut the 1-inch wide tape-measure elements with a pair of shears, chamfering the ends of the elements. Be very careful—the edges are very sharp and will inflict a nasty cut if you are careless. Use some sandpaper to remove the sharp edges and burrs and put some vinyl electrical tape on the ends of the elements to protect yourself from getting cut. Wear safety glasses while cutting the elements. See **Fig 44** for dimensions.

The RG-58 coax feed line is wound into a 8-turn coil along the boom to form a choke balun. The coil is covered with vinyl black tape to secure it to the boom. Make sure you scrape or sand the paint off the tape-measure elements where the feed line is attached. Most tape measures have a very durable paint finish designed to stand up to heavy use. You do not want the paint to insulate your feed-line connection!

If you are careful, you can solder the feed line to the element halves, but take care since the steel tape measure does not solder easily and the PVC supports can be easily melted. Tin the tape-measure elements before mounting them to the PVC cross if you decide to connect the feed line in this fashion.

If you decide not to solder to the tape-measure elements, you can use two other methods to attach the feed line. One method employs ring terminals on the end of the coax. The ring terminals are then secured under self-tapping screws or with 6-32 bolts and nuts into holes drilled in the driven-element halves. However, with this method you cannot fine-tune the antenna by moving the halves of the driven element in and out.

The simplest method is simply to slide the ends of the feed line under the driven element hose clamps and tighten the clamps to hold the ends of the coax. This is low-tech but it works just fine.

WB2HOL used $1^{1}/_{2}$-inch stainless-steel hose clamps to attach each driven-element half to the PVC cross that acts as its support. This allowed him to fine-tune his antenna for lowest SWR simply by loosening the hose clamps and sliding the halves of the driven element in or out to lengthen or shorten the element. He achieved a 1:1 SWR at 146.565 MHz (the local fox-hunt frequency) when the two elements were spaced about 1 inch apart. **Fig 45** shows the hose-clamp method for attaching the driven element to the PVC cross, along with the hairpin wire and feed-line coax.

Some builders have used rubber faucet washers between the tape-measure elements and the PVC-cross fittings on the director and reflector. These allow for the tape to fit the contour of the PVC fitting better and will make the antenna look nicer. It is normal for the reflector and director elements to buckle a bit as they are tightened to the PVC tee and cross if you don't use faucet washers. You can also eliminate the buckling if you use self-tapping screws to attach these elements instead of hose clamps. The beam will not be as rugged, however, as when you use hose clamps.

This beam has been used on fox hunts, on mountain tops, at local public-service events, outdoors, indoors in attics—just about everywhere. The SWR is typically very close to 1:1 once adjusted. Front-to-back performance is exactly as predicted. The null in the rear of the pattern is perfect for transmitter hunts.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

W. U. Amphar, “Unidirectional Loops for Transmitter Hunting,” *QST*, Mar 1955.

G. Bonaguide, “HF DF—A Technique for Volunteer Monitoring,” *QST*, Mar 1984.

D. S. Bond, *Radio Direction Finders*, 1st edition (New York: McGraw-Hill Book Co).

R. E. Cowan and T. A. Beery, “Direction Finding with the Interferometer,” *QST*, Nov 1985.

D. DeMaw, “Beat the Noise with a Scoop Loop,” *QST*, Jul 1977.

D. DeMaw, “Maverick Trackdown,” *QST*, Jul 1980.

T. Dorbuck, “Radio Direction-Finding Techniques,” *QST*, Aug 1975.

D. T. Geiser, “Double-Ducky Direction Finder,” *QST*, Jul 1981.

D. T. Geiser, “The Simple Seeker,” *The ARRL Antenna Compendium*, *Vol 3*, p 126.

N. K. Holter, “Radio Foxhunting in Europe,” Parts 1 and 2, *QST*, Aug and Nov 1976.

J. Isaacs, “Transmitter Hunting on 75 Meters,” *QST*, Jun 1958.

H. Jasik, *Antenna Engineering Handbook*, 1st edition (New York: McGraw-Hill, 1961).

R. Keen, *Wireless Direction Finding*, 3rd edition (London: Wireless World).

J. Kraus, *Antennas*, 2nd edition (New York: McGraw-Hill Book Co, 1988).

J. Kraus, *Electromagnetics*, 4th edition (New York: McGraw-Hill Book Co, 1992).

C. M. Maer, Jr., “The Snoop-Loop,” *QST*, Feb 1957.

M. C. Mallette, “The Four-Way DFer,” *QST*, Nov 1995. A set of two PC boards is available from FAR Circuits for the “Four-Way DFer” project. 18N640 Field Ct, Dundee, IL 60118-9269, tel 708-576-3540 (voice and fax). Note: No component pads for C15 exist. Mount C15 between U2 pin 5 and ground on the bottom (foil) side of the board. A PC-board template package is available from the ARRL free of charge. Send your request for the MALLETTE 4-WAY DFER, along with a business size SASE, to the Technical Department Secretary, 225 Main St, Newington, CT 06111-1494.

L. R. Norberg, “Transmitter Hunting with the DF Loop,” *QST*, Apr 1954.

P. O’Dell, “Simple Antenna and S-Meter Modification for 2-Meter FM Direction Finding,” Basic Amateur Radio, *QST*, Mar 1981.

Ramo and Whinnery, *Fields and Waves in Modern Radio* (New York: John Wiley & Sons, 1944).

F. Terman, *Electronic and Radio Engineering* (New York: McGraw-Hill Book Co, 1955).

Radio Direction Finding , published by the Happy Flyer, 1811 Hillman Ave, Belmont, CA 94002.

For more information on direction finding, see *The ARRL Handbook* and *Transmitter Hunting: Radio Direction Finding Simplified*, by Joe Moell, KØOV, and Thomas Curlee, WB6UZZ. These books are available from your local dealer or can be ordered directly from ARRL.

Chapter 15

Portable Antennas

For many amateurs, the words *portable antennas* may conjure visions of antenna assemblies that can be broken down and carried in a backpack, suitcase, golf bag, or what-have-you, for transportation to some out-of-the way place where they will be used. Or the vision could be of larger arrays that can be disassembled and moved by pickup truck to a Field Day site, and then erected quickly on temporary supports. Portable antennas come in a wide variety of sizes and shapes, and can be used on any amateur frequency.

Strictly speaking, the words "portable antenna" really means *transportable antenna*—one that is moved to some (usually temporary) operating position for use. As such, portable antennas are not placed into service when they are being transported. This puts them in a different class from mobile antennas, which are intended to be used while in motion. Of course this does not mean that mobile antennas cannot be used during portable operation. Rather, true portable antennas are designed to be packed up and moved, usually with quick reassembly being one of the design requisites. This chapter describes antennas that are designed for portability. However, many of these antennas can also be used in more permanent installations.

Any of several schemes can be employed to support an antenna during portable operation. For HF antennas made of wire, probably the most common support is a conveniently located tree at the operating site. Temporary, lightweight masts are also used. An aluminum extension ladder, properly guyed, can serve as a mast for Field Day operation. Such supports are discussed in Chapter 22, Antenna Supports.

A SIMPLE TWIN-LEAD ANTENNA FOR HF PORTABLE OPERATION

The typical portable HF antenna is a random-length wire flung over a tree and end-fed through an antenna tuner. Low-power antenna tuners can be made quite compact, but each additional piece of necessary equipment makes portable operation less attractive. The station can be simplified by using resonant impedance-matched antennas for the bands of interest. Perhaps the simplest antenna of this type is the half-wave dipole, center-fed with 50- or 75-Ω coax. Unfortunately, RG-58, RG-59 or RG-8 cable is quite heavy and bulky for backpacking, and the miniature cables such as RG-174 are too lossy.

A practical solution to the coax problem, developed by Jay Rusgrove, W1VD, and Jerry Hall, K1TD, is to use folded dipoles made from lightweight TV twin-lead. The characteristic impedance of this type of dipole is near 300 Ω, but this can easily be transformed to a 50-Ω impedance. The transformation is obtained by placing a lumped capacitive reactance at a strategic distance from the input end of the line. **Fig 1** illustrates the construction method and gives important dimensions for the twin-lead dipole.

A silver-mica capacitor is shown for the reactive element, but an open-end stub of twin-lead can serve as well, provided it is dressed at right angles to the transmission line for some distance. The stub method has the advantage of easy adjustment of the system resonant frequency.

The dimensions and capacitor values for twin-lead dipoles for the HF bands are given in **Table 1**. To pre-

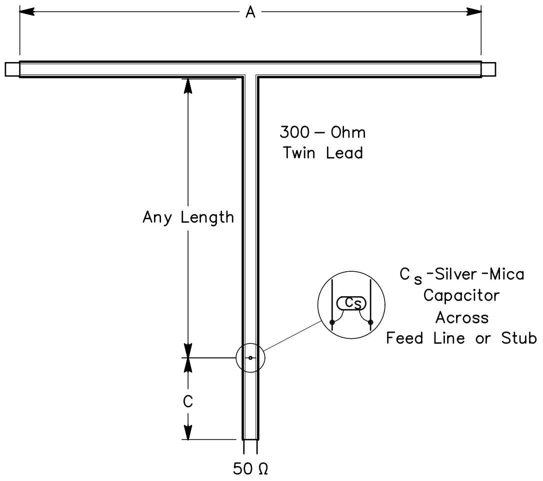

Fig 1—A twin-lead folded dipole makes an excellent portable antenna that is easily matched to 50-Ω equipment. See text and Table 1 for details.

Table 1

Twin-Lead Dipole Dimensions and Capacitor Values

Frequency	*Length A*	*Length C*	C_s	*Stub Length*
3.75 MHz	124' 9½"	13' 0"	289 pF	37' 4"
7.15	65' 5½"	6' 10"	151 pF	19' 7"
10.125	46' 2½"	4' 10"	107 pF	13' 10"
14.175	33' 0"	3' 5½"	76 pF	9' 10½"
18.118	25' 10"	2' 8½"	60 pF	7' 9"
21.225	22' ½"	2' 3½"	51 pF	6' 7"
24.94	18' 9"	1' 11½"	43 pF	5' 7½"
28.5	16' 5"	1' 8½"	38 pF	4' 11"

serve the balance of the feeder, a 1:1 balun must be used at the end of the feed line. In most backpack QRP applications the balance is not critical, and the twin-lead can be connected directly to a coaxial output jack—one lead to the center contact, and one lead to the shell.

Because of the transmission-line effect of the shorted-radiator sections, a folded dipole exhibits a wider bandwidth than a single-conductor type. The antennas described here are not as broad as a standard folded dipole because the impedance-transformation mechanism is frequency selective. However, the bandwidth should be adequate. An antenna cut for 14.175 MHz, for example, will present an SWR of less than 2:1 over the entire 14-MHz band.

ZIP-CORD ANTENNAS

Zip cord is readily available at hardware and department stores, and it's not expensive. The nickname, *zip cord*, refers to that parallel-wire electrical cord with brown or white insulation used for lamps and many small appliances. The conductors are usually #18 stranded copper wire, although larger sizes may also be found. Zip cord is light in weight and easy to work with.

For these reasons, zip cord can be pressed into service as both the transmission line and the radiator section for an emergency dipole antenna system. This information by Jerry Hall, K1TD, appeared in *QST* for March 1979. The radiator section of a zip-cord antenna is obtained simply by "unzipping" or pulling the two conductors apart for the length needed to establish resonance for the operating frequency band. The initial dipole length can be determined from the equation $\ell = 468/f$, where ℓ is the length in feet and f is the frequency in MHz. (It would be necessary to unzip only half the length found from the formula, since each of the two wires becomes half of the dipole.) The insulation left on the wire will have some loading effect, so a bit of length trimming may be needed for exact resonance at the desired frequency.

For installation, you may want to use the electrician's knot shown in **Fig 2** at the dipole feed point. This is a balanced knot that will keep the transmission-line part of the system from unzipping itself under the tension of dipole suspension. This way, if zip cord of sufficient length for both the radiator and the feed line is obtained, a solder-free installation can be made right down to the input end of the line.

(Purists may argue that knots at the feed point will create an impedance mismatch or other complications, but as will become evident in the next section, this is not a major consideration.) Granny knots (or any other variety) can be used at the dipole ends with cotton cord to suspend the system. You end up with a lightweight, low-cost antenna system that can serve for portable or emergency use.

But just how efficient is a zip-cord antenna system? Since it is easy to locate the materials and simple to install, how about using such for a more permanent installation? Upon casual examination, zip cord looks about like 72-Ω balanced feed line. Does it work as well?

Zip Cord as a Transmission Line

To determine the electrical characteristics of zip cord as a radio-frequency transmission line, a 100-foot roll was subjected to tests in the ARRL laboratory with an RF impedance bridge. Zip cord is properly called *parallel power cord*. The variety tested was manufactured for GC Electronics, Rockford, IL, being 18-gauge, brown, plastic-insulated type SPT-1, GC cat. no. 14-118-2G42. Undoubtedly, minor variations in the electrical-characteristics will occur among similar cords from different manufacturers, but the results presented here are probably typical.

The characteristic impedance was determined to be 107 Ω at 10 MHz, dropping in value to 105 Ω at 15 MHz and to a slightly lower value at 29 MHz. The nominal

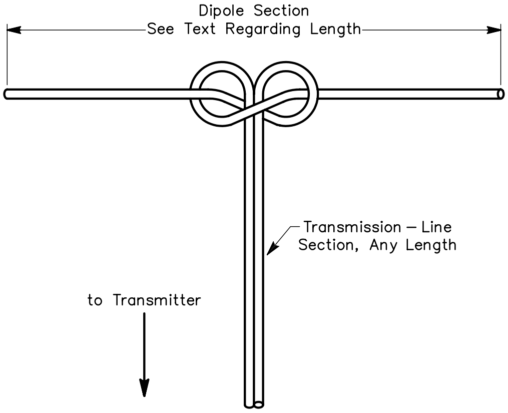

Fig 2—This electrician's knot, often used inside lamp bases and appliances in lieu of a plastic grip, can also serve to prevent the transmission-line section of a zip-cord antenna from unzipping itself under the tension of dipole suspension. To tie the knot, first use the right-hand conductor to form a loop, passing the wire behind the unseparated zip cord and off to the left. Then pass the left-hand wire of the pair behind the wire extending off to the left, in front of the unseparated pair, and thread it through the loop already formed. Adjust the knot for symmetry while pulling on the two dipole wires.

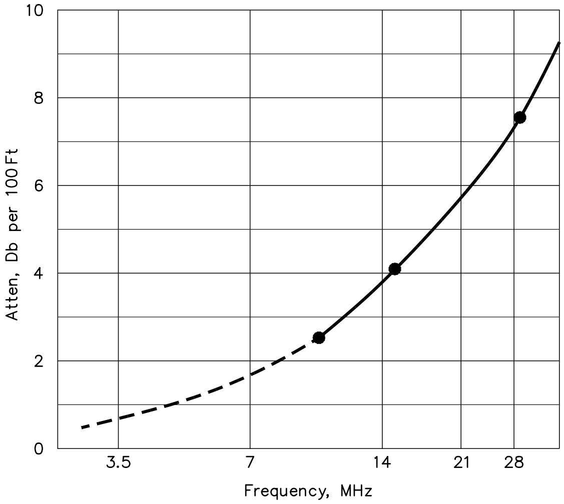

Fig 3—Attenuation of zip cord in decibels per hundred feet when used as a transmission line at radio frequencies. Measurements were made only at the three frequencies where plot points are shown, but the curve has been extrapolated to cover all high-frequency amateur bands.

value is 105 Ω at HF. The velocity factor of the line was determined to be 69.5%.

Who needs a 105-Ω line, especially to feed a dipole? A dipole in free space exhibits a feed-point resistance of 73 Ω, and at heights above ground of less than $^1/_4$ wavelength the resistance can be even lower. An 80-meter dipole at 35 feet over average soil, for example, will exhibit a feed-point resistance of about 35 Ω. Thus, for a resonant antenna, the SWR in the zip-cord transmission line can be 105/35 or 3:1, and maybe even higher in some installations. Depending on the type of transmitter in use, the rig may not like working into the load presented by the zip-cord antenna system.

But the really bad news is still to come—line loss! **Fig 3** is a plot of line attenuation in decibels per hundred feet of line versus frequency. Chart values are based on the assumption that the line is perfectly matched (sees a 105-Ω load as its terminating impedance).

In a feed line, losses up to about 1 decibel or so can be tolerated, because at the receiver a 1-dB difference in signal strength is just barely detectable. But for losses above about 1 dB, beware. Remember that if the total losses are 3 dB, half of your power will be used just to heat the transmission line. Additional losses over those charted in Fig 3 will occur when standing waves are present. (See Chapter 24, Transmission Lines.) The trouble is, you can't use a 50- or 75-Ω SWR instrument to measure the SWR in zip-cord line accurately.

Based on this information, we can see that a hundred feet or so of zip-cord transmission line on 80 meters might be acceptable, as might 50 feet on 40 meters. But for longer lengths and higher frequencies, the losses become appreciable.

Zip Cord Wire as the Radiator

For years, amateurs have been using ordinary copper house wire as the radiator section of an antenna, erecting it without bothering to strip the plastic insulation. Other than the loading effects of the insulation mentioned earlier, no noticeable change in performance has been noted with the insulation present. And the insulation does offer a measure of protection against the weather. These same statements can be applied to single conductors of zip cord.

The situation in a radiating wire covered with insulation is not quite the same as in two parallel conductors, where there may be a leaky dielectric path between the two conductors. In the parallel line, it is the current leakage that contributes to line losses. This leakage current is set up by the voltage potential that exists on the two adjacent wires. The current flowing through the insulation on a single radiating wire is quite small by comparison, and so as a radiator the efficiency is high.

In short, communications can certainly be established with a zip-cord antenna in a pinch on 160, 80, 40, 30 and perhaps 20 meters. For higher frequencies, especially with long line lengths for the feeder, the efficiency of the system is so low that its value becomes questionable.

A TREE-MOUNTED HF GROUND-PLANE ANTENNA

A tree-mounted, vertically polarized antenna may sound silly. But is it really? Perhaps engineering references do not recommend it, but such an antenna does not cost much, is inconspicuous, and it works. This idea was described by Chuck Hutchinson, K8CH, in *QST* for September 1984.

The antenna itself is simple, as shown in **Fig 4**. A piece of RG-58 cable runs to the feed point of the antenna, and is attached to a porcelain insulator. Two radial wires are soldered to the coax-line braid at this point. Another piece of wire forms the radiator. The top of the radiator section is suspended from a tree limb or other convenient support, and in turn supports the rest of the antenna.

The dimensions for the antenna are given in **Fig 5**. All three wires of the antenna are ¼ wavelength long. This generally limits the usefulness of the antenna for portable operation to 7 MHz and higher bands, as temporary supports higher than 35 or 40 feet are difficult to come by. Satisfactory operation might be had on 3.5 MHz with an inverted-L configuration of the radiator, if you can overcome the accompanying difficulty of erecting the antenna at the operating site.

The tree-mounted vertical idea can also be used for fixed station installations to make an invisible antenna. Shallow trenches can be slit for burying the coax feeder and the radial wires. The radiator itself is difficult to see unless you are standing right next to the tree.

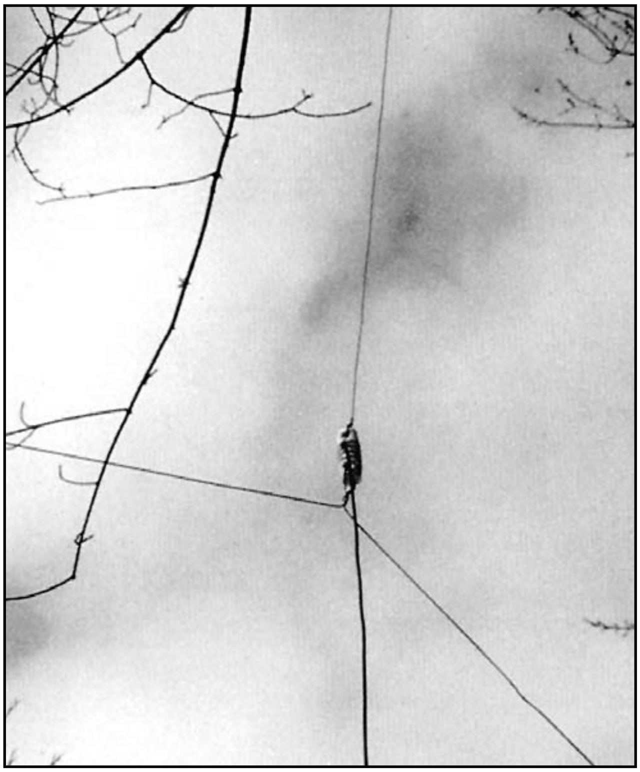

Fig 4—The feed point of the tree-mounted ground-plane antenna. The opposite ends of the two radial wires may be connected to stakes or other convenient points.

A PORTABLE DIPOLE FOR 80 TO 2 METERS

This dipole antenna, described by Robert Johns, W3JIP, in August 1998 *QST*, can be used for any band from 80 through 2 meters. One half of the dipole is an inductively loaded aluminum tube. Its length is adjustable from 4 to 11 feet, depending on how much room is available. The other half is flexible insulated wire that can be spooled out as necessary. The tube is supported by a flagpole bracket attached to a long carpenter's clamp. The clamp mounts the antenna to practically anything: a windowsill, railing, a chair or post. If there is no structure to mount the antenna (a parking lot or the beach), the clamp attaches to two light wooden legs to form a tripod, as shown in **Fig 6**.

The key to mounting flexibility is the large clamp. The key to electrical flexibility is a large adjustable coil that lets you resonate the tube on many bands. The coil is wound with #8 aluminum ground wire from RadioShack. The form is a four-inch ($4^{1}/_{2}$-inch OD) styrene drain coupling from the Home Depot or a large plumbing supply store. A movable tap adjusts the inductance to tune the upright tube to the desired band. The wire half of the antenna is always a bit less than $\lambda/4$ on each band. Hang it from any convenient support or drape it over bushes to keep it off the ground.

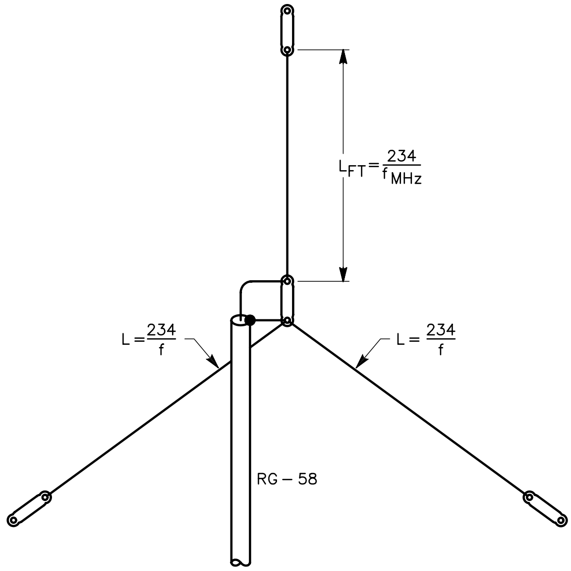

Fig 5—Dimensions and construction of the tree-mounted ground-plane antenna.

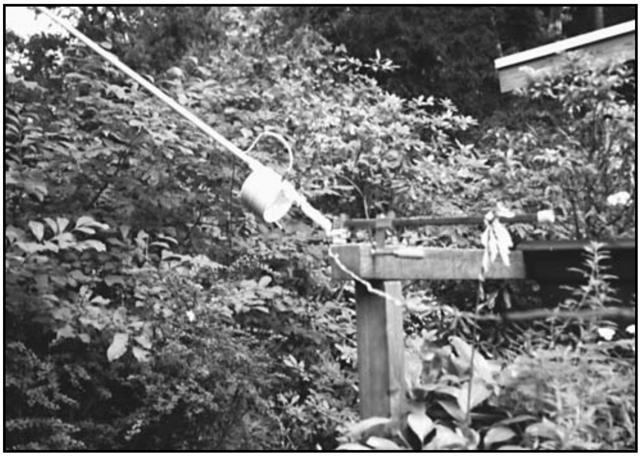

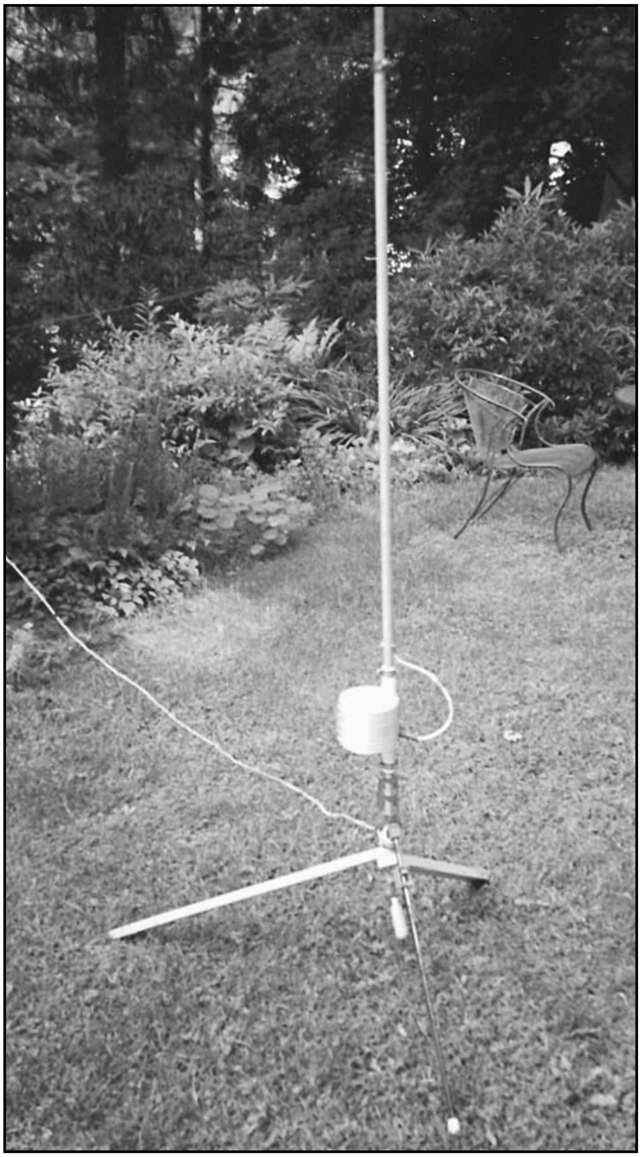

Fig 6—At top, the portable antenna in some of the many places it may be mounted around the house, porch and yard. At bottom, the simple ground-mounted legs that make a tripod.

Construction

The 18-inch carpenter's clamp (sometimes called a *bar clamp*, such as Jorgensen's No. 3718) and flagpole bracket that takes a $^3/_4$-inch pole are common hardware items. Insulate the bracket from the clamp jaw with a $1^1/_2$-inch length of 1-inch PVC pipe (see **Fig 7**). Hammer the PVC over the end clamp jaw to make it take the shape of the jaw. Secure the flagpole bracket to PVC with a large ground clamp (for $^1/_2$ or 1-inch conduit). The ground clamp includes $^1/_4$-inch bolts; enlarge the flagpole bracket holes to accept them. Some flagpole brackets have an integral cleat; the author hammered the cleat ears flat on his.

Mount an SO-239 chassis connector on the flagpole bracket using RadioShack insulated standoffs (276-1381). The standoffs tightly press the center pin of the SO-239 against the bracket surface; no other connection is needed. Other mounting hardware may require a connection from the coax center conductor to the bracket. The spooled wire's inner end wraps around a standoff and connects to a ring terminal under a screw holding the SO-239 flange to the standoff.

The 1 × 2-inch wooden legs for the tripod are each 30 inches long. Bolt them together at one end with a 1 or $1^1/_4$-inch-long bolt. Countersink the bolt head and nut below the surface of the wood so they don't interfere with the clamp jaws.

Aluminum Element

You can make this element from three lengths of telescoping aluminum tubing ($^3/_4$, $^5/_8$ and $^1/_2$ inch, 0.058-inch walls). The author used tubing with thinner walls for less weight and easier handling. A 45-inch-long, $^3/_4$-inch tube fits the flagpole bracket. He chose this length because it and the flagpole bracket make $\lambda/4$ on 6 meters. The two outer tubes are both $^5/_8$ inch, made by cutting a seven-foot aluminum clothesline pole in halves. They are

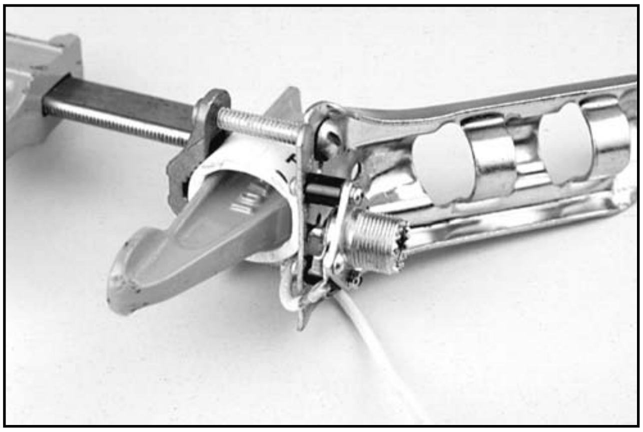

Fig 7—The flagpole bracket that supports the tubing elements is clamped to the long carpenter's clamp, but insulated from it by a small section of 1-inch PVC pipe. A coax connector is mounted to the bracket and the spool of wire is attached to the coax connector.

joined together with a copper coupling (5/8 inch ID) for 1/2-inch copper pipe. The coupling is bolted to one of the 5/8-inch sections, and a slot is cut in the free half of the coupling. The remaining 5/8-inch tube is inserted there and secured with a hose clamp. See **Fig 8**.

One 5/8-inch tube slides into the 3/4-inch tube to provide a continuously variable element length from about 4 to 7½ feet. This extends from 7½ to 11 feet when the two 5/8-inch sections are joined together.

Loading Coil

The loading coil has 12 turns of bare aluminum wire spaced to fill the 3½-inch length of the drain coupling. Drill 9/64-inch holes at the ends of the coil form to accept the ends of the coil wire. (See **Fig 9**.) To wind the coil, feed four inches of wire through the form, make a sharp bend in the wire and start winding away from the nearby mounting hole. Wind 12 turns on the form, spacing them only approximately. Cut the wire for a 4-inch lead and feed it through the other hole in the form. Tighten the wire as best you can and bend it into another acute angle where it passes into the form. Space the turns about equally, but don't fuss with them. Final spacing will be set after the wire is tightened.

Tighten the coil wire by putting a screwdriver or a needle-nose pliers jaw under one turn, and pry the wire up and away from the surface of the coil form. While this can be done anywhere, it's best to put these kinks on the backside, away from the mounting shaft. Put kinks in every other turn, removing any slack from the coil and holding the turns in place. Should the coil ever loosen, simply retighten it with a screwdriver. If you prefer, glue the coil turns in place with epoxy or coil dope. Use a thin bead of glue that won't interfere with the clip that connects to the coil.

The coil form mounts on a nine-inch-long 3/4-inch PVC pipe. (See Fig 9 and **Fig 10**.) The inside diameter is slightly larger than the 3/4-inch aluminum, but slotting the PVC and tightening it with a hose clamp secures the tube. (Use a wide saw to cut these slots, not a hacksaw.) The coil form mounts to the PVC pipe with #6-32 × 1½-inch bolts. A five-inch long, 3/4-inch aluminum tube permanently attaches to one end of the coil assembly and slides into the flagpole bracket. One end of the coil wire connects to this short tube. Flatten the wire end by hammering it on something hard, then drill a 9/64-inch hole in the flattened end and attach it to the short tube with a #6-32 × 1-inch bolt. Tighten the bolt until the 1/2-inch tube starts to flatten. This will keep pressure on the aluminum-to-aluminum joint.

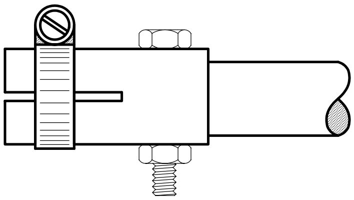

Fig 8—The joint between two sections of 5/8-inch tubing is made from a 1/2-inch copper pipe coupling, bolted to one section and hose clamped to the other.

The aluminum element slides into the opposite end of the coil assembly. The hose clamp there can be tightened until the element just slides in snugly.

A 12-inch clip lead connects the aluminum element to the coil. Bolt the plain wire end to the 3/4-inch tube three inches from its end. Many alligator clips will fit in the space between the turns of the coil (about 3/16 inch), but W3JIP preferred using a solid-copper clip (Mueller TC-1). Cut off most of the jaws, so that only the part close to the hinge grabs the coil. This shorter lever grips very tightly.

Wire and Spool

The lower half of the antenna is insulated wire that is about $\lambda/4$ on the band of operation. The wire is pulled from the spool, and the remaining wire forms an inductance that doesn't add much to the antenna length. The Home Depot sells #12 and #14 insulated stranded copper wire in 50 and 100-foot lengths, on plastic spools. (See **Fig 11**.) A 1/2-inch dowel fits into the spool to make a

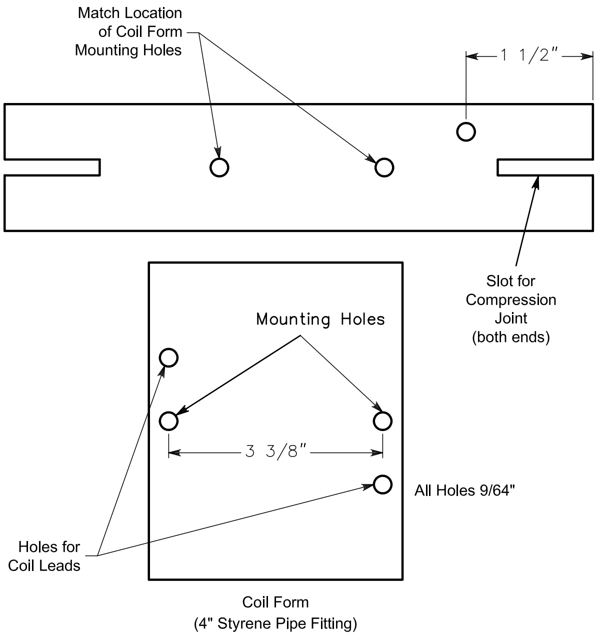
Match Location of Coil Form Mounting Holes
1 1/2"
Slot for Compression Joint (both ends)
Mounting Holes
3 3/8"
All Holes 9/64"
Holes for Coil Leads
Coil Form (4" Styrene Pipe Fitting)

Fig 9—Holes to be drilled in the styrene coupling and the 3/4-inch PVC pipe. All holes are 9/64-inch diameter, to provide clearance for #6-32 bolts. The hole 1½ inches from one end holds a bolt that serves as a stop, so that the antenna tube does not slide in too far. Space the holes for coil leads far enough from the mounting holes to clear the 3/4-inch pipe.

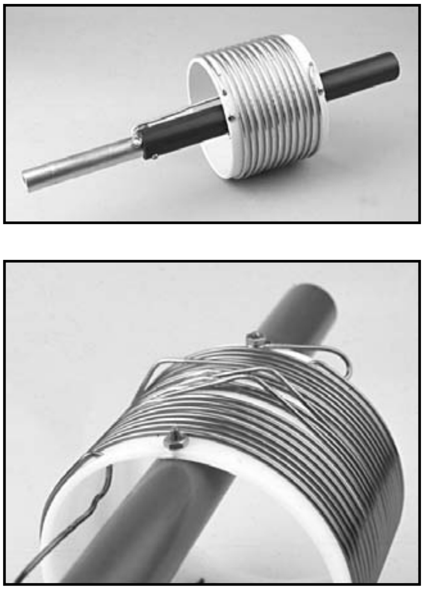

Fig 10—Top photo shows the loading coil for the 20, 30, and 40-meter band coverage. The short aluminum tube on the coil slides into the flagpole bracket, and the tubing element slides into the other end of the PVC pipe. The wire and clip connect the element to the coil. Bottom photo shows the coil for operating the antenna on 80 meters. This is placed in the flagpole bracket and the 40-meter aluminum coil plus the tubing element is inserted into it.

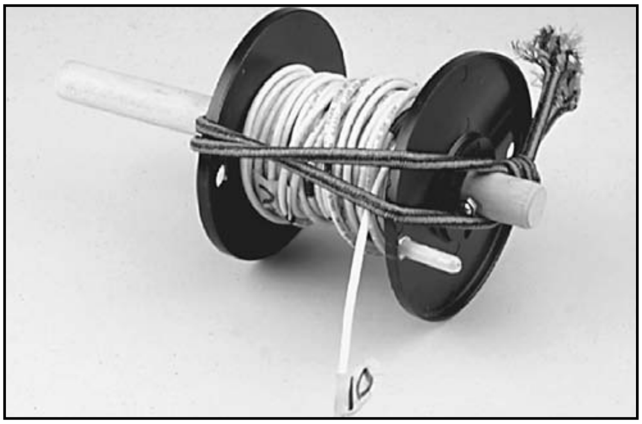

Fig 11—The wire spool has a wooden axle/handle and small handle for winding the wire. A bungee cord stretched over the spool and around the axle prevents the wire from unwinding.

Table 2

	Tubing Length (feet)	*Coil Turns*	*Wire Length (feet)*
6 meters	4	0	4
10 meters	7	0	8
	4	1.7	7
12 meters	8	0	8.3
	4	1.8	9
15 meters	10	0	10.3
	8	1.8	10
	4	3.5	10
17 meters	11	0	13
	8	2.2	13
	4	4.2	13
20 meters	11	1	16
	8	4	16
	4	6.2	16
30 meters	11	4	24
	8	7	24
	4	10.5	24
40 meters	11	9	32
	8	12	32

handle and spool axle. Bolts through the dowel on either side of the spool hold it in place. A crank handle is made by putting a one-inch-long bolt through the spool flange.

The author calibrates the wire on the spool with markers and electrical-tape flags. There's a mark (from a permanent marking pen) at each foot, a black flag every five feet and a length-marked colored flag every 10 feet. Simply mark the length for each band, if you like.

Make sure you prevent the wire from unspooling, especially when it's hanging from a window mounting. A heavy rubber band works, but it doesn't last long. A better solution is a loop of light bungee cord, preferably with a knot for grip. The bungee loop runs from the axle/handle around the spool making a half twist on the way, and then passes over the axle end on the other side of the spool. (See Fig 11.)

Operation

Table 2 lists element length, wire length and coil tap point for various bands. When the number of turns shown is zero, the coil is not needed. On all bands except 6 meters, you can simply bypass the coil with the clip lead—the extra length just lowers the frequency a bit. For 6 meters, the coil *must* be removed. The location of the unspooled wire greatly affects the settings, so these numbers are only starting points. The lengths in the table were taken with the wire one to three feet above ground, draped over and through bushes and flowerbeds. The antenna will still work if the wire is lying on the ground, but it will require less unspooled wire to resonate. A balun is not needed, and an SWR analyzer is very helpful while adjusting this antenna.

The SWR is less than 1.5:1 on all bands, and it's usually below 1.2:1. Occasionally, a band shows a higher SWR (still less than 2:1), but that can always be lowered by adjusting the length or location of the lower wire. Never set up the antenna where it could fall and injure someone, or where the unwary could get an RF burn by touching it.

The author's results with this antenna have been excellent, both from home and on vacation. If you haven't yet operated from a seashore location, be prepared for a pleasant surprise! The good ground afforded by the salt water really makes a difference.

80 Meters

It's easy to add this lower frequency band. Fig 10 shows a 35 µH coil for 80 meters. It's constructed and tightened just like the 40-meter coil, but has 20 turns of #12 magnet wire.

To operate on 75/80 meters, insert the new coil into the flagpole bracket and plug the 40-meter coil into it. Tune across the band with the movable tap on the 40-meter coil. This varies antenna resonance from below 3.5 to above 4.0 MHz, with the full 11 feet of tubing extended. If your version doesn't achieve this tuning range, adjust the spacing of the turns on the 80-meter coil.

The 80-meter coil has a five-inch length of $^3/_4$-inch aluminum tube inserted into one end of the $^1/_2$-inch PVC pipe that supports the coil form. One end of the coil is connected to this aluminum tube. The other end is secured under the bolt that holds the coil form to the PVC pipe. A second clip lead connects the base of the 40-meter coil to the outer end of the larger coil. The length of wire on the spool must also be increased to about 64 feet.

2 Meters

A $\lambda/2$ dipole for 2 meters can be made with about 15 inches of $^1/_2$-inch aluminum tubing in the flagpole bracket, and 18 inches of wire. The tubing element is shorter than normal for 2 meters because the bracket is also part of the antenna. You can also shorten the 6-meter wire a bit and operate the 6-meter antenna as a $3\lambda/2$ on 2 meters, with a somewhat higher SWR.

Continuous Coverage

With easily changed element lengths and a continuously variable loading coil, you may operate the antenna on any frequency from 6.5 to 60 MHz, if coverage for other services is needed. With taps in the 80-meter coil at 8, 11 and 14 turns, the antenna will also tune from 4 to 7 MHz.

THE HF CABOVER ANTENNA

If you have ever had the pleasure of traveling across the country with an HF mobile in a camper, trailer or motor home you may want to duplicate this antenna for use when you park. This antenna was described in *The ARRL Antenna Compendium, Vol* 5, by Jim Ford, N6JF.

The author's camper has limited spots on which to mount an HF mobile antenna. The back ladder is a convenient place to mount a small mobile whip. However, the efficiency of typical mobile center-loaded antennas, depending on coil Q and other assumptions, is often less than 2% for 80 meters and 10% for 40 meters. (These numbers come from the excellent, easy-to-use *MOBILE* antenna design program by Leon Braskamp, AA6GL, which is on the CD-ROM bundled with this book.)

At some locations, N6JF used an 80-meter dipole, which was very efficient and worked great on all bands when fed with open-wire line and an antenna tuner. However, it took over 40 minutes to set up and about 20 minutes to tear down, working with a sling shot and many tree snags. This is too much time to make a schedule or for an early morning departure, although it's OK if you plan to stay for a while. Even more important, there were often too many trees or other barriers (perhaps some even social) to allow putting up the dipole at a campsite.

When this happened he was stuck with the mobile antenna with poor efficiency. There had to be a better antenna for camper operation. The author decided on a large vertical.

A telescoping aluminum extension pole used for roller painting would make a good bottom section for the loaded vertical. These are available at many local home supply centers. The author's was 1 inch in diameter and 6 feet long, telescoping to almost 12 feet. He disassembled both sections and cross-cut a 1-inch slot in the top of the bottom section with a hacksaw to allow compression clamping with a hose clamp. The tip of the top section was fitted with an aluminum plug that had a 3/8-24 hole tapped in it. This procedure was a simple machine-shop operation. The plug was pounded into the top section and is quite snug. He tapped some set screws through the pole into the plug, however, just to be sure. An insulated, lay-down marine antenna mount fit perfectly into the bottom of the aluminum base and was secured with a bolt that also served as the electrical connection from the capacitor matching box. See **Fig 12** showing the aluminum base plate, the laydown mount and the antenna

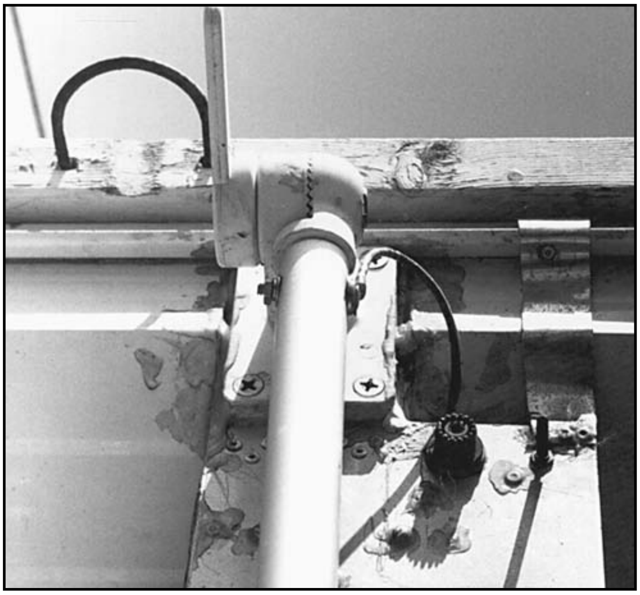

Fig 12—Photo of aluminum base plate, showing details of mounting to the camper. The four banana-plug jacks on the bottom are for extra radials, if desired.

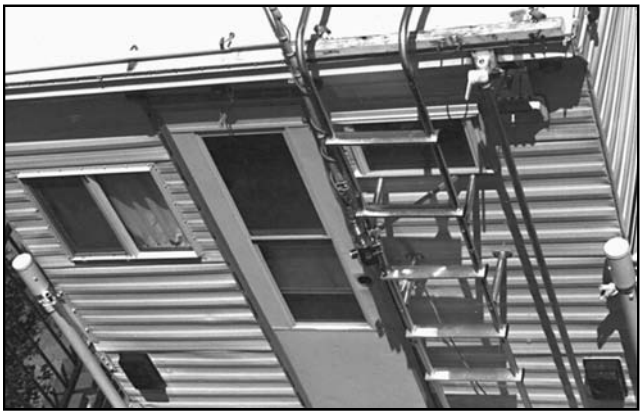

Fig 13—Photo showing the back of the camper, with the antenna in the lowered position, parallel to the ladder. The "Outback" standby mobile antenna is shown clamped to the left side of the ladder.

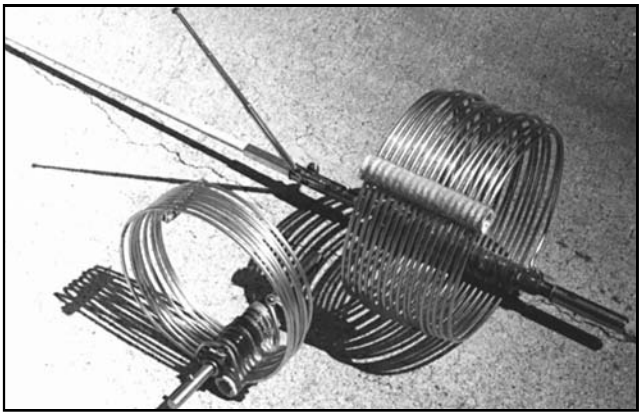

Fig 15—Close-up photo of 80 and 40-meter coils, with top section and telescoping whip antenna with swivel bracket for tuning the top section for the higher bands. Note quick-disconnect connectors at top and bottom of both coils. The top whip is a Fiberglas CB whip, used on 80 and 40 meters.

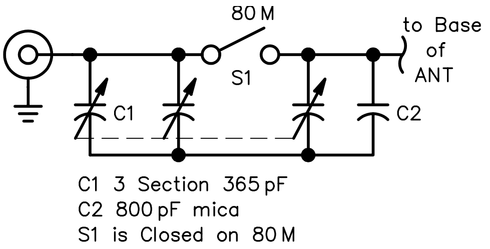

Fig 14—Schematic of tuning capacitor at base plate. C2 is an 800 pF transmitting mica capacitor. C1 is a three-section 365 pF broadcast tuning capacitor. S1 is closed for 80-meter operation.

mast itself lowered down the back of the camper. **Fig 13** shows the layout of the back of the camper, with the antenna on the right-hand side, laid down for travel.

The variable capacitor in the matching box is a surplus three-section 365-pF broadcast tuning capacitor. Two of three sections are connected in parallel and a switch parallels in the third section, along with an extra 800 pF mica capacitor for use on 80 meters. See the schematic in **Fig 14**.

The capacitor assembly was put in a custom-glued Plexiglas box to keep out the weather and mounted to a piece of plate aluminum, along with an SO-239 connector. The aluminum plate was riveted to the camper shell using a lot of aluminum rivets. Do not use steel. N6JF peeled back about a 4-inch wide section of the side of the camper for this direct aluminum-to-aluminum connection. People who are hesitant to modify their campers like this need to find an alternate low-resistance connection method. His camper was old enough not to be an issue.

For an extra low-resistance connection a $1^{1}/_{2}$-inch aluminum strip was added from the top of the base plate to the camper, as shown in Fig 12 near the 80-meter switch. The tuning knob protrudes from the side of the Plexiglas box. The four bottom holes in the plate are for banana plugs to connect ground radials if extra efficiency is desired. However, the roof of a camper is one of the better places for a mobile antenna, so the author seldom hooks up the radials. When he does use them, the tuning changes only a little.

To keep losses down, N6JF used coils wound with aluminum clothes-line wire on old mobile coil center sections with quick disconnect fittings. See **Fig 15**, which shows both the 80 and 40 meter coils, together with the top portion of the antenna. An article by Robert Johns, W3JIP, in October 1992 *QST* described techniques for building your own loading coils. The coils ended with a little more inductance than calculated and the author had to remove some turns. Both coils are spaced at 4 turns per inch. The 8-inch long 80-meter coil has 18 turns. The 7-inch long 40-meter coil has 8 turns.

The matching network is actually an L-section step-up match, using the net inductive reactance of the antenna plus the center loading coil. The PVC coil construction technique was described in W3JIP's *QST* article and a follow-up Technical Correspondence piece in October 1992 *QST*. Basically, it consists of drilling an accurate row of slightly undersized holes along a length of $^{1}/_{2}$-inch PVC pipe and then carefully sawing down the center of the row of holes with a hack saw. Then, you trap the coil wires in the grooves between the two sawed halves and tie the two halves together with string. When you are satisfied that everything is proper, you then tighten the string and apply epoxy glue to make it strong and permanent.

One advantage of aluminum clothes line wire is that it is already coiled at the approximate diameter needed when you buy it and it is easy to position on the coil form.

The clothes line wire had a plastic coating, which wasn't removed except at the contact points. Once the epoxy dries, this method of construction does a good job of holding the finished coil together and it is lightweight.

The computed coil Q from the *MOBILE* antenna program is about 800 for the 80-meter coil and about 400 for the 40-meter coil. The author accidentally made the 40-meter coil 7 inches in diameter instead of a higher-Q 6 inch diameter. Even so, the whole antenna system with a 9-foot top section calculates as being 56% efficient on 80 meters and 85% for 40 meters.

The removable top section for 80 and 40 meters is a full-size fiberglass CB whip from RadioShack. The fiberglass whip has about a #16 hole in the center of it. Be sure to sand and paint the whip for protection against UV and to protect yourself against fiberglass spurs in the hands. N6JF tried a full-size stainless steel CB whip to get a slightly higher capacitance to ground because of larger conductor size but discovered it was far too heavy. That experience reinforced his decision that aluminum was a far more practical coil and base section material for this project.

Quick-disconnect connectors found years ago at a hamfest were used for both coil forms and for the top section. Bands higher than 40 meters don't need any loading coils and the antenna length can be telescoped to get a $^1/_4$ wavelength. Ten meters doesn't require any top section. Be sure to use some NOALOX or similar compound to prevent corrosion and poor connections at all aluminum joints. This is especially true for the telescoping sections and the aluminum rivets. The matching capacitor is in the circuit at all times but when the 80-meter switch is off you can set the capacitor at minimum (about 14 pF) and it is effectively out of the circuit, even at 10 meters. The author has not tried this antenna on power levels greater than about 100 W but the weakest link would probably be the matching capacitor. The voltage at the matching capacitor is low, so 200 W should be no problem.

You can achieve a 1:1 SWR match for 80 or 40 meters and a good SWR is obtained without retuning the base matching capacitor for approximately 100 kHz on 80 meters and most of 40 meters. The top section, however, does not have such low Q and needs to be tuned. The 2:1 SWR bandwidth on 80 meters is about 25 kHz and 150 kHz for 40 meters. Tuning is accomplished by using a telescoping FM portable radio antenna connected to the top-section whip with a stainless steel hose clamp. The maximum length of the telescoping section used was 29 inches, and it collapses down to 7 inches. A whip with an elbow was used to adjust the angle of the whip as well. A telescoping whip half the length would still be long enough for the full adjustment of both bands.

Adjustment of the top section is one of the penalties paid to achieve high efficiency for operation on 80 and 40 meters. Substituting an automatic antenna tuner would likely lose efficiency, particularly on 80 meters since the base resistance calculates to about 10 Ω. N6JF has not tried to make the antenna work on 160 meters. The SWR for the higher bands was good enough without any matching network.

As light as the antenna is, it still won't hold up in a moderate wind without some support guys. N6JF used $^3/_4$-inch and $^1/_2$-inch PVC support pipes in a telescoping arrangement for storage, but expanding to give an approximate 45° support at the top of the bottom section from both directions. One end of the telescoping section was connected to the camper roof with a hinge. The other end formed a snap-fit out of a PVC barrel that was cut lengthwise. See **Fig 16** and **Fig 17** for details. Even though it formed a good snap fit, N6JF didn't trust the joint for strong winds so he glued a piece of Velcro to the joint to close up the open end. Be careful, though, because Velcro deteriorates with exposure.

Another hose clamp near the top of the bottom section holds an $^1/_8$-inch line taken up the ladder to pull up the antenna without the need of an assistant. The disadvantage is you have to climb the roof. Use a non-slip floor mat or something similar to spread the load on the roof and to avoid slipping. Once on the roof, however, the coil is at a height for easy adjustment when the telescoping section is in the down position.

An advantage of being able to assemble the whole antenna on the roof is that you don't need a lot of swing-up room and you can clear trees easily. You can put calibration marks on the upper aluminum section for resonant lengths on the higher bands but just raise the top section up all the way for 80 or 40 meters. Mark also the small

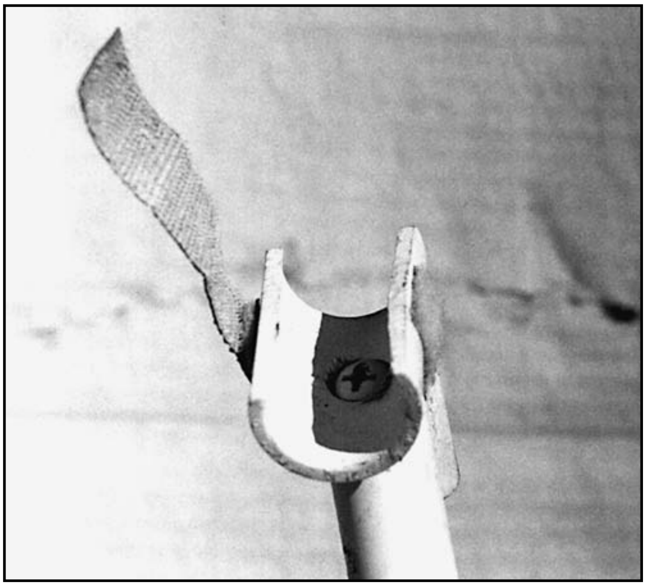

Fig 16—Close-up of one of the snap-on support brackets used to brace the antenna. Note the Velcro pieces used to ensure that the antenna doesn't pop out of the bracket in the wind.

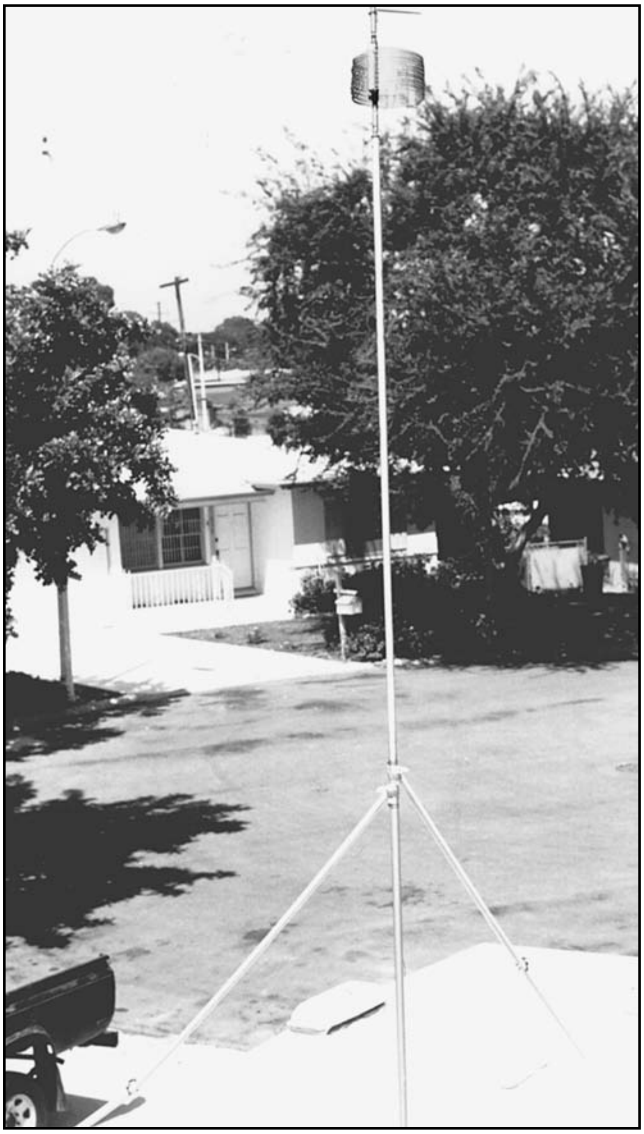

Fig 17—Photo showing the two support bracket poles bracing the bottom section of the antenna. The top tuning whip is evident above the homemade loading coil at the top of the bottom section.

coil tuning whip for 80 or 40 meters, although different locations may require slightly different settings due to detuning from nearby metal objects.

The overall length is about 21 feet. The top of the author's camper is about 10 feet high when on the truck, putting the tip at 31 feet. This antenna is definitely designed for use when you are parked at a fixed location. N6JF can put up this antenna in less than 5 minutes and can take it down in half the time. The success of this project has as much to do with knowing how and where you operate as it does paying attention to mechanical and electrical details. The antenna has been a good compromise between efficiency and convenience.

TWO PORTABLE 6-METER ANTENNAS

These antennas were described by Markus Hansen, VE7CA, in *The ARRL Antenna Compendium, Vol 5*. After years of HF operation, he became enthusiastic about VHF/UHF operation when he found a used Yaesu FT-726R VHF/UHF all-mode transceiver at a reasonable price.

But he became really enthused when he got on 6 meters and discovered the joys of driving to high mountain peaks to operate. Not only does an antenna have to be portable for this kind of operation, it must be easily assembled and disassembled, just in case you have to move quickly to a better location.

A Portable Two-Element Six-Meter Quad

VE7CA's primary objective was to construct a two-element quad using material found in any small town. It should not use a complicated matching network. The Gamma matches commonly used on quads do not hold up well when you are setting up and taking down these antennas in the field. The final design adjusted the distance between the driven and the reflector elements so that the intrinsic feed-point impedance was 50 Ω.

Fig 18 shows the dimensions for the boom and the boom-to-mast bracket. The boom is made from a $27^1/_4$-inch length of 2 × 2. (The actual dimension of 2 × 2 is closer to $1^3/_4$ by $1^3/_4$ inches but it is commonly known in lumber yards as a 2 × 2.) Use whatever material is available in your area, but lightweight wood is preferred, so clear cedar or pine is ideal. Drill the four $^1/_2$-inch holes for the spreaders with a wood bit, two at each end, through one of the faces of the 2 × 2 and the other through the other face. The boom-to-mast bracket is made from $^1/_4$-inch fir plywood.

The spreaders are $^1/_2$-inch dowel. The local lumberyard had a good supply of fir dowels but other species of wood are available. The exact material is not important. Maple is stronger but expensive. Fiberglass would be ideal but it is not always available locally. Cut two of the $^1/_2$-inch dowels to a length of $83^5/_8$ inches for the driven element spreaders and two to 88 inches for the reflector spreaders. **Fig 19** shows one end of the boom with the two spreaders inserted. The mast was made from two six-foot lengths of $1^3/_4$-inch fir dowel. Again, use whatever you may have available. Waterproof all wooden parts with at least two coats of exterior varnish.

While you are at the lumberyard or hardware store look for plastic pipe that fits over the end of the $^1/_2$-inch spreaders. You will need a one-foot length, with some to spare. Cut it into seven equal lengths, approximately one inch long, and one to a length of $1^1/_2$ inches. Drill a $^1/_{16}$-inch hole through the seven equal lengths $^1/_4$ inch from the ends, and two holes one above the other $^1/_4$-inch apart on the $1^1/_2$-inch sleeve. VE7CA used #14 hard-drawn stranded bare copper wire for the elements. Do not use insulated wire unless you are willing to experimentally

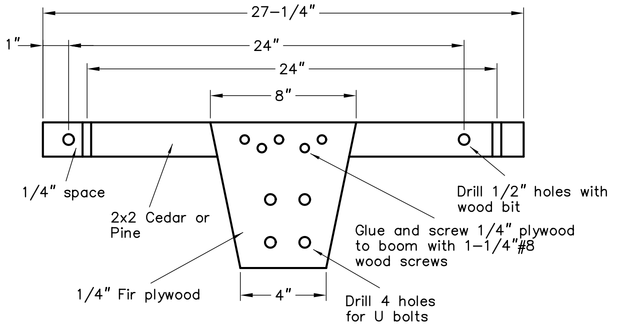
27-1/4"
1"
24"
24"
8"
1/4" space
2x2 Cedar or Pine
1/4" Fir plywood
4"
Drill 1/2" holes with wood bit
Glue and screw 1/4" plywood to boom with 1-1/4"#8 wood screws
Drill 4 holes for U bolts

Fig 18—Dimensions for the boom-to-mast bracket for VE7CA's portable two-element 6-meter quad.

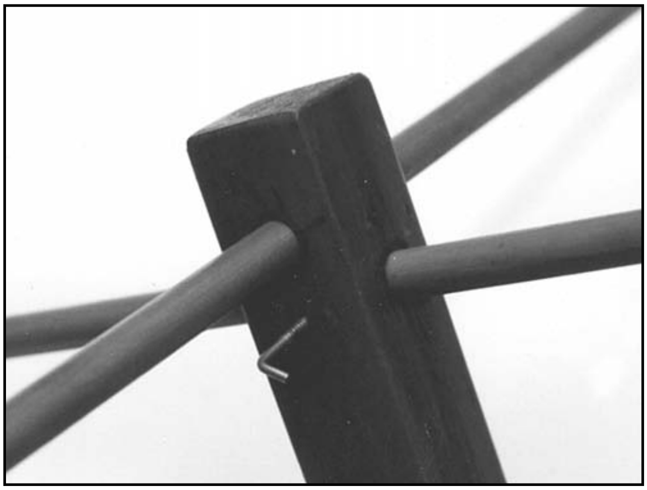

Fig 19—Photo of one end of the VE7CA quad with the two spreaders inserted.

determine the element lengths, since the insulation detunes each element slightly.

Cut the reflector element 251 inches long and slip one end of the wire through the holes you drilled in four of the plastic sleeves. Don't attempt to secure the wire to the plastic sleeves at this point. Cross the end of the reflector elements one inch from their respective ends and twist and solder together. The total circumference of the reflector element should be 249 inches when the ends are connected together.

Cut the driven element wire to 241 inches and slip three of the 1-inch sleeves onto the wire. Again, don't secure the wires to the sleeves yet. Then the ends are passed through the two holes in the $1^{1}/_{2}$-inch pipe. Wrap the ends around the pipe and twist them back onto themselves to secure the wire. The coax feed line is attached directly to the two ends at this point. The circumference of the driven-element loop from the points where the coax is attached should be $236^{5}/_{8}$ inches. Solder the coax feed line to the driven element and waterproof the coax with silicone seal. The author used RG-58, as it is lightweight. The length required for a portable installation is typically not very long, maybe 20 feet, so the loss in the small cable is not excessive. Near the feed point, coil the coax into six turns with an inside diameter of two inches. This is an effective method of choking any RF from flowing on the outside of the coax shield.

Begin assembling the quad by pushing the two reflector spreaders, without wires attached, through one end of the boom and the two shorter driven-element spreaders through the holes in the other end of the boom. Center the spreaders and mark the spreaders with a black felt-tip pen next to the boom. Now insert a $1^{1}/_{2}$ #8 wood screw or a threaded L-hook into the boom so that it just touches one of the spreaders. Take the screw or L-hook out and file the end flat, then reinsert it so that it is just snug against the spreaders. The author only used two L-hooks for the two vertical spreaders; the horizontal spreaders are held in the proper position by the tension of the wire loops. If you use an L-hook, you can unscrew it with your hands—you won't have to worry about leaving the screwdriver at home.

You are now ready to assemble the wire loops. Take the reflector loop and place the four plastic caps over the ends of the reflector spreaders. Equalize the wire lengths between the spreader so that the loop is square. Now, secure the plastic sleeve pipes by tightly wrapping wire around the sleeve and the wire element and soldering the wire in place. See **Fig 20**, a photo showing one of the plastic sleeves slipped over one of the spreader ends, with the wire element through the hole and fastened in place. Follow the same procedure with the driven element.

Fig 21 shows the quad's boom, with the plywood boom-to-mast bracket fastened with wood screws and glue. Two U-bolts are used to attach to the mast. When the quad is raised, the shape of the loop is commonly known as a diamond configuration. The mast consists of two six-foot lengths of doweling joined together with a two-foot length of PVC plastic pipe, held together with wood screws.

Make a slot the width of a #8 wood screw about one inch deep from the top of the plastic PVC pipe and then put the top mast into the plastic pipe. Insert a one inch #8 wood screw into the bottom of the slot you cut into the top of the pipe and tighten only enough so that the top mast can be removed without unscrewing it. VE7CA drove a nail into the end of the lower mast and left it exposed an inch or more. This end is placed in the ground and the nail holds the pole in place. A strip of wood approximately 1 × 3 and long enough to cross over the roof rack of the family van is used to hold the center of the antenna mast to the roof rack of the van with small diameter rope. See **Fig 22** for a photo of the quad in action next to the family van.

To disassemble the quad, lay it on its side, slip the plastic sleeves off the ends of the spreaders and roll up

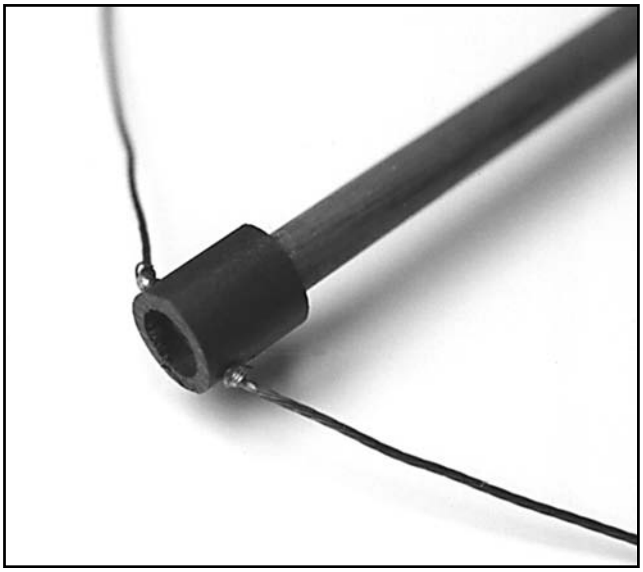

Fig 20—Photo showing one of the plastic sleeves slipped over end of a spreader to provide a mechanical mounting point and support for the wires.

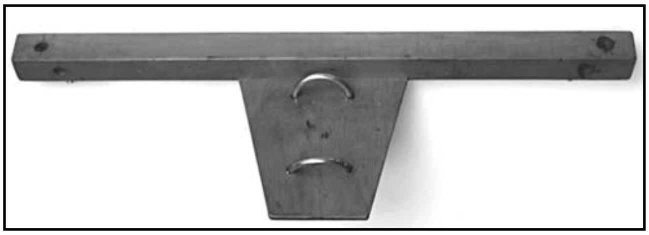

Fig 21—Photo of the two-element quad's boom, with the plywood boom-to-mast bracket secured with wood screws and glue.

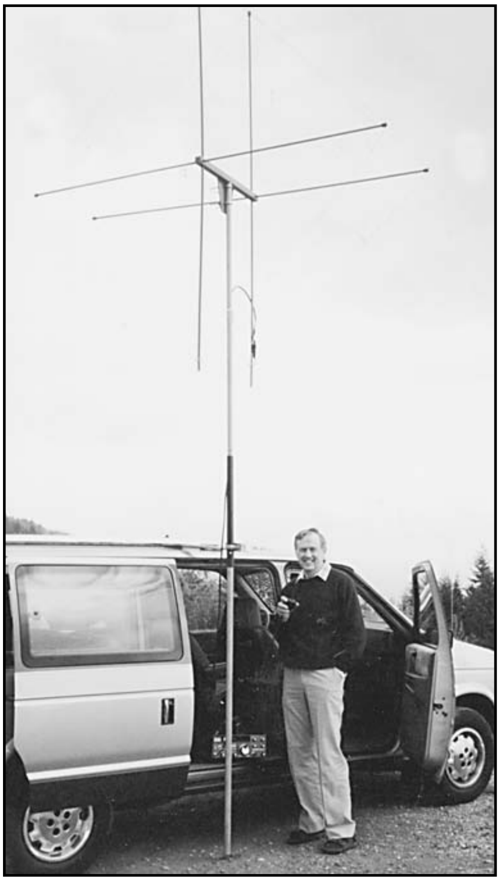

Fig 22—Ready for action! VE7CA has set up his quad next to the family van.

the wire loops. Loosen the L-hooks holding the vertical spreaders in place. Push the spreaders out of the boom, loosen the U-bolts and free the mast from the boom.

That is all there is to it. It takes about two minutes to put it up, or take it down. It is quite sturdy and has survived several high-wind storms.

A Three-Element Portable 6-Meter Yagi

The idea to build a Yagi antenna resulted when the author traded the family van for a compact car. He needed something that would fit into the trunk of the car. At close to 7-feet long, the quad spreaders were too long. Computer modeling showed that a three-element Yagi on a five-foot boom also could pick up about 1.5 dB gain over the short-boom two-element quad. A five-foot boom fits into the trunk or across the back seat of the car, but something had to be done about the nine-foot elements!

One day VE7CA noticed a box of portable-radio telescoping antenna elements at the radio parts store. They were 54 inches long when fully extended. He next found a 60-inch length of aluminum tubing that fit over the end of the telescoping elements. There are many different sizes of telescoping antenna elements, with different diameters. This is where you will have to use your scrounging skills! **Fig 23** shows how the tubing is used as a center section to join two telescoping elements together. It also serves to extend the total length of each element, since two telescoping elements themselves are not long enough to resonate on six meters. See **Table 3** for dimensions and element spacings. Each center section is slotted at both ends with a hacksaw, and stainless-steel hose clamps are used to secure the telescoping elements.

Fig 24 shows the center sections of the three elements with their mounting brackets. A square boom was used to obtain a flat surface to work with. **Fig 25** shows how the reflector is attached to the end of the boom with two $1^1/_2$-inch 10-32 bolts and wingnuts. **Fig 26A** provides the dimensions and details for the reflector and director element-to-boom brackets, which are formed from $^1/_{16}$-inch plate aluminum. The driven element is split in the center and is insulated from the boom. Fig 26B shows details for the driven-element bracket. **Fig 27** is a photo

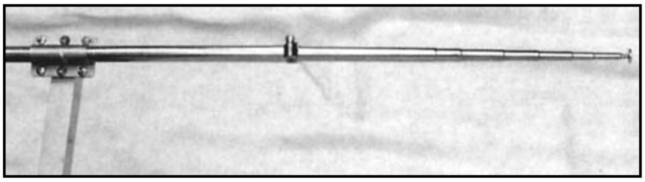

Fig 23—Photo showing a piece of aluminum tubing used as a center section to join the two telescoping tips together.

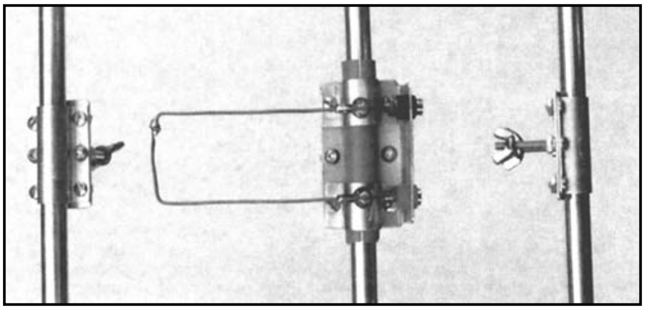

Fig 24—A view of the center sections of the three Yagi elements with their mounting brackets.

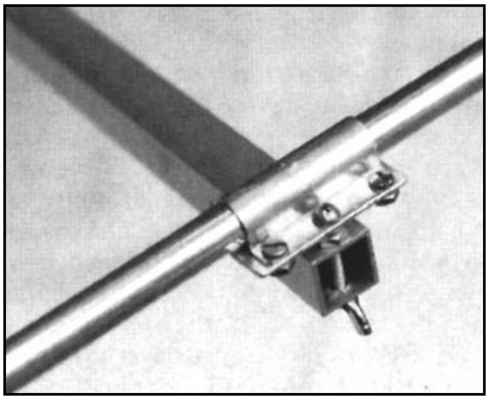

Fig 25—Photo detailing attachment of the reflector to the square-section boom, using two #10 bolts and wingnuts.

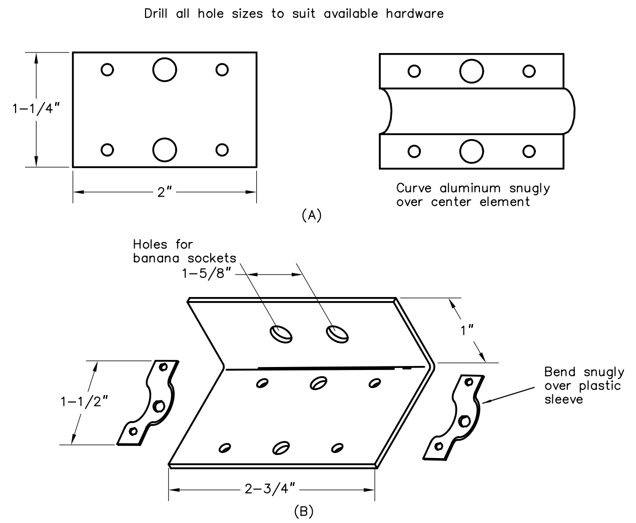
Drill all hole sizes to suit available hardware
1-1/4"
2"
Curve aluminum snugly over center element
(A)
Holes for banana sockets
1-5/8"
1"
1-1/2"
Bend snugly over plastic sleeve
2-3/4"
(B)

Fig 26—At A, details for the reflector and director element-to-boom brackets, made of $^{1}/_{16}$-inch plate aluminum. At B, details for the driven-element bracket. These are screwed to the square boom.

of the driven element with the hairpin matching wire and the banana plugs used to connect the coax to the driven element. You could use a female PL-259 connector if you wish. VE7CA used #14 solid bare copper wire for the hairpin. It is very durable—even after being severely warped in the car trunk, it can be bent back into shape quickly and easily.

The boom is $^{3}/_{4}$-inch square aluminum, 65 inches long. The material was found at a local hardware store. To detach the elements, just loosen the wing nuts and remove the elements from the boom. A similar method was used to attach the support mast to the boom.

As with the quad, a choke balun was used, consisting of a coil of 6 turns of coax with an inside diameter of 2 inches. To tune the hairpin match, assemble the Yagi on its mast and extend the elements. Spray switch contact solution on a cloth and wipe any dirt and grease from the elements. Push the elements together and apart a

Table 3

Three-Element Yagi, Element Lengths and Spacing Along the Boom, and Hairpin Dimensions

Element	Spacing Along Boom (inches)	Center Section Ele. Length (inches)	Telescoping (inches)	Total Length (inches)
Reflector	0	$22^{3}/_{4}$	$51^{1}/_{2}$	$125^{3}/_{4}$
Driven	28	$9^{3}/_{4}$ *	$48^{5}/_{8}$	$58^{3}/_{16}$
Director	$63^{3}/_{8}$	$14^{1}/_{2}$	$51^{1}/_{4}$	117
Hairpin	#14 wire	4 long	$1^{5}/_{8}$ spacing	

* Driven-element uses 2 sections insulated at center

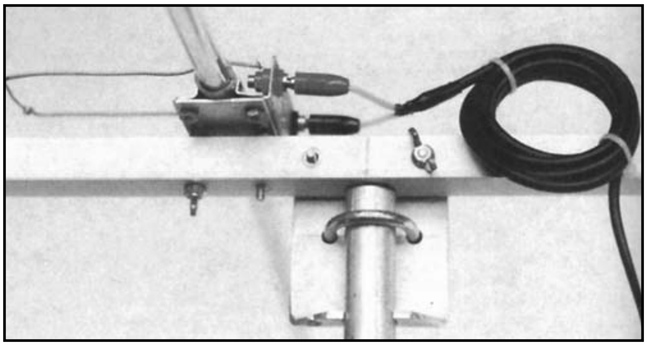

Fig 27—Photo of the driven element, complete with hairpin match and the banana plugs used to connect the coax cable to the driven element.

couple of times so that the contact solution cleans the elements thoroughly. Attach the antenna mast to your vehicle or use whatever method of support you intend to use in the field. Connect an SWR meter and a transmitter to the coax feeding the antenna. VE7CA used two alligator clips soldered together to slide along the two hairpin wires to find the position for the lowest SWR. The dimensions computed by computer were correct! The SWR was below 1.16:1 from 50.05 to 50.2 MHz.

You can take this antenna out of the trunk of the car and assemble it in less than two minutes. One caution: the telescoping elements when fully extended are quite fragile. VE7CA has not broken one as yet, but carrying a spare element just in case would be a good idea.

VE7CA 2-Element Portable HF Triband Yagis

These portable HF wire Yagis were described by Markus Hansen, VE7CA, in November 2001 *QST* and in *The ARRL Antenna Compendium, Vol 7*.

A 20/15/10-Meter Triband 2-Element Yagi

VE7CA wanted a simple 2-element wire Yagi for 20/15/10 meters that could be stored in the ski boot of his car. The basic concept comprises three individual dipole driven elements, one each for 20, 15 and for 10 meters tied to a common feed point, plus three separate reflector elements. The elements are strung between two 2.13-meter (7-foot) long, 2 × 2-inch wood spreaders, each just long enough to fit into the ski boot of his car.

By playing with the reflector-to-driven element spacing and the initial driven-element lengths, VE7CA was able to come up with a feed-point impedance on each band that allowed the use of a single setting for the shorting bar on a hairpin match. The result was a very acceptable match over the lower portions of each band. The layout of the 20/15/10-meter triband wire Yagi is shown in **Fig 28**, with the dimensions provided in **Table 4**.

The dimensions shown in Table 4 are what resulted after tuning for the lowest SWR in the middle of the lower portion of each band. VE7CA set his system up by hanging one end of the antenna from a tree and sloped it downwards at 45º, tying the lower end to a peg in the ground. The height at the feed point was 30 feet.

The feed-point impedance of an antenna is affected by many environmental factors. The presence of a reflector relatively close to the driven element has a major effect, since the impedance at the driven element in a Yagi is affected by the tuning of the driven element itself, by the spacing and length of the reflector element and to a lesser extent the height of the antenna above ground and the character of the soil itself. The real challenge in a multiband Yagi with a single feed line is to obtain a low SWR on all the bands.

The hairpin match is one of the easiest matching systems to make. It is easy to adjust and since wire is the only ingredient, it can be coiled up with the rest of the antenna when the antenna is disassembled. The feed-point impedance of the Yagi with a reflector element spaced 0.1 λ behind the driven element typically produces a resistance around 20 Ω. By shortening the driven element from its resonant length, capacitive reactance is added to the feed-point resistance. This can be cancelled by shunting the feed point with an inductor in the shape of a wire loop resembling a *hairpin*. This causes a step up of the 20-Ω feed-point resistance to 50 Ω.

Table 4
Dimensions for 20/15/10-Meter Tribander

Frequency MHz	*Spacing DE to Refl cm (feet)*	*Driven Element Half Length cm (feet)*	*Reflector Length cm (feet)*	*Hairpin Length cm (inches)*
14.1	213 (6.99)	488 (16.01)	1065 (34.94)	43 (14.9)
21.1	175 (5.74)	335 (10.99)	708 (23.23)	
28.25	125 (4.1)	254 (8.33)	531 (17.42)	

Spacing between hairpin wires is 10 cm (4 inches).

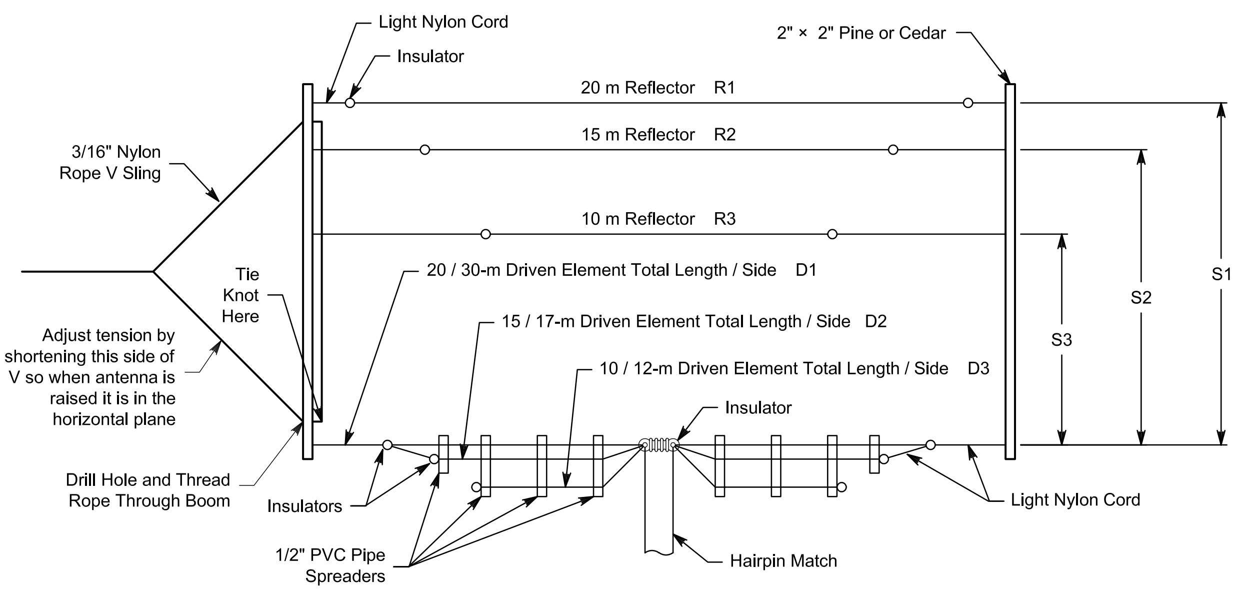
Light Nylon Cord
Insulator
2" × 2" Pine or Cedar
20 m Reflector R1
15 m Reflector R2
3/16" Nylon Rope V Sling
10 m Reflector R3
S1
S2
S3
Tie Knot Here
20 / 30-m Driven Element Total Length / Side D1
15 / 17-m Driven Element Total Length / Side D2
10 / 12-m Driven Element Total Length / Side D3
Adjust tension by shortening this side of V so when antenna is raised it is in the horizontal plane
Insulator
Drill Hole and Thread Rope Through Boom
Insulators
Light Nylon Cord
1/2" PVC Pipe Spreaders
Hairpin Match

Fig 28—Dimensions for VE7CA's 2-element 20/15/10-meter triband Yagi.

The Balun

Fig 29 shows the hairpin match and the common-mode choke balun for the 10/15/20-meter triband wire Yagi. You should let the coax drop straight down from the center insulator and attach it to the center of the hairpin shorting bar. Continue by making a coil using the coax of 8 turns, with a diameter of about 4 inches. This balun will choke off RF flowing along the outside of the coax shield that would otherwise distort the radiation pattern of the antenna. The center of the shorting bar is at a neutral potential, so there is no harm in attaching the coax feed line at that point.

Flattop or Tilted?

If DX is your main interest, then you want to tilt the antenna towards the vertical to emphasize the lower elevation angles. Remember that the radiation pattern is quite dependent on ground conductivity and dielectric constant for a vertically polarized antenna. A location close to saltwater will yield the highest gain and the lowest radiation angle. With very poor soil in the near and far field, the peak radiation angle will be higher and the gain less.

VE7CA has had numerous opportunities to test this out while operating portable at his favorite location. Using two trees as supports he can pull the antenna to horizontal with the feed line about 7 meters above the ground. In this position, with 20 meters open to Europe, he found it difficult to work DX on CW with 3 W of output power. However, when he changed the slope of the antenna so that it is nearly vertical he not only heard more DX stations, but he found it relatively easy to work DX. The sloping antenna always performs much better for working DX than a low horizontal antenna.

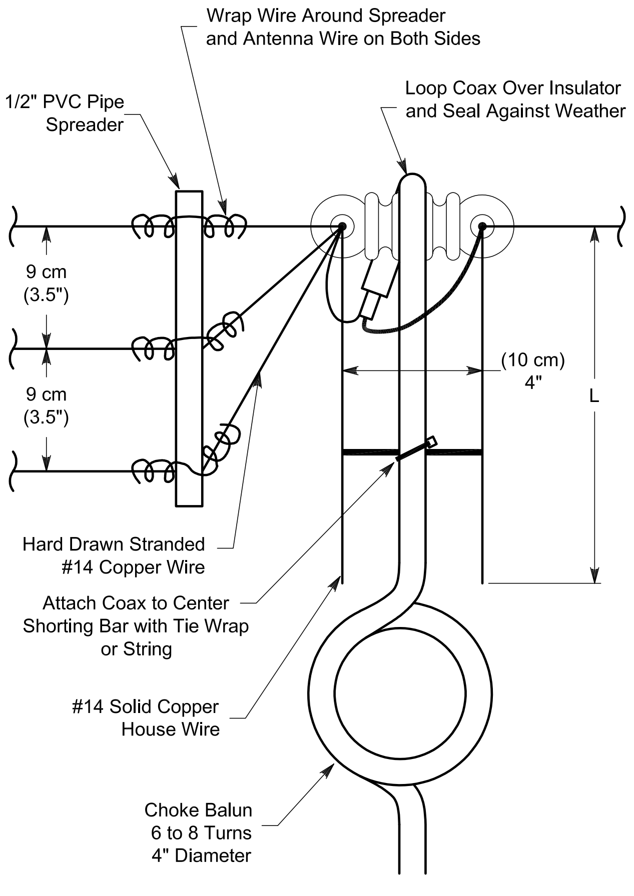
Wrap Wire Around Spreader and Antenna Wire on Both Sides
1/2" PVC Pipe Spreader
Loop Coax Over Insulator and Seal Against Weather
9 cm (3.5")
9 cm (3.5")
(10 cm) 4"
L
Hard Drawn Stranded #14 Copper Wire
Attach Coax to Center Shorting Bar with Tie Wrap or String
#14 Solid Copper House Wire
Choke Balun 6 to 8 Turns 4" Diameter

Fig 29—Details of feed point for 20/15/10-meter triband Yagi. The same mechanical support is provided for the balun and feed coax for the 30/17/12-meter tribander.

A 30/17/12-Meter Triband 2-Element Yagi

When the author attempted to add 17 and 12 meter elements to the existing 20/15/10-meter Yagi model he became exasperated. Adding two more driven elements and reflectors brings many more variables into the equation! It became clear that there was serious interaction between the elements. He could not obtain a workable feed-point impedance on all five bands that could be transformed to 50 Ω using a single hairpin match. There was also serious pattern distortion on 12 meters.

Even building a WARC-only triband Yagi for 30/17/12 meters turned out to be a real challenge. VE7CA had difficulty finding a combination that would allow him to use a single matching system to transform the feed-point impedance of the combined driven elements to 50 Ω. He couldn't create a 30/17/12 triband Yagi using the same design principles as his 20/15/10-meter version. The main problem occurred on 12 meters. Not only was the feed-point impedance unmanageable, but the radiation pattern had four lobes, not the single lobe you'd like from a Yagi.

He decided to try the *Modified Coaxial-Sleeve* method, more aptly termed by K9AY the *Coupled-Resonator (C-R)* in his article in *The ARRL Antenna Compendium*, *Vol 5*. The K9AY method uses a single driven element, with other elements placed in very close proximity (but not physically connected) to the driven element. By starting with the dimensions suggested by K9AY for a triband 30/17/12-meter dipole, VE7CA was able to develop a 2-element Yagi with acceptable feed-point impedances on all three bands using a single hairpin match. Notice that this WARC design uses a 2 × 2-inch wooden boom length that is 230 cm (7.5 feet) long. Of course, the antenna can't fit into a typical ski boot anymore, so VE7CA had to put it on a roof rack for transportation.

The space between the tightly coupled driven elements is only 3.7 cm (1.5 inches), so you need to use more PVC pipe spreaders than in the 20/15/10 design to make sure the driven-element wires stay as close as possible to the desired spacing without physically touching each other. The driven elements lie in the horizontal plane and the hairpin match and feed line hang down vertically from the center of the 30-meter driven element.

The spacing between the 30-meter driven elements and the other two conductors and the size of the wire all played a part in developing this antenna for a single feed line with the common hairpin match. Do not change the wire size from the recommended #14 for the driven elements unless you are willing to spend a considerable amount of time with a *NEC*-based modeling program retweaking the antenna. This is not the case with the 20/15/10 tri-band Yagi, where any convenient sized wire is acceptable.

Using #14 gauge wire allows all the Yagi antennas in this article to be used at the maximum power levels allowed in North America. The only limiting factor is the power handling capability of the feed line. However, even RG-58 should work for the relatively short length from the feed point down to ground level, where you can change to RG-8 or some other higher-power, lower-loss coaxial cable if you wish. **Fig 30** is a detailed drawing of the 30/17/12-meter driven element. The other dimensions for the 30/17/12-meter triband Yagi are shown in **Table 5**.

Assembly

When you are ready to assemble your wire Yagi, start by attaching the longest reflector element and the driven element assembly to the wood end booms. Do this with the wires and the booms on the ground. Next attach the V slings to both of the booms and with ropes attached to the V slings pull the array up off the ground between two supports (perhaps two trees). A height of 1.5 meters (5 feet) above the ground makes it easy to work on the antenna while you add the other reflector elements and adjust the V slings. Pull them tight so that the array is fairly flat. It won't stay horizontal, because the driven elements are heavier than a single reflector element. So you will need to support the 2 × 2-inch spreaders so they are horizontal. Lean the booms on something at a convenient height, such as the rungs of two step ladders. Now add the two other reflector elements, but don't pull them as tight as the longest reflector. Next attach the feed line.

Table 5
Dimensions for 30/17/12-Meter Tribander

Frequency MHz	*Spacing DE to Refl cm (feet)*	*Driven Element Length cm (feet)*	*Reflector Length cm (feet)*	*Hairpin Length cm (inches)*
10.12	230 (7.5)	713 (23.4) Half	1476 (48.4)	24.5 (9.5)
18.11	165 (5.4)	808 (26.5)	822 (27.0)	
24.91	120 (3.9)	570 (18.7)	606 (19.9)	

Spacing between hairpin wires is 10 cm (4 inches). Note that dimensions for 17 and 12-meter driven elements are full lengths, since they are not broken with insulator in the middle, unlike all driven elements for 20/15/10-meter triband design in Table 4.

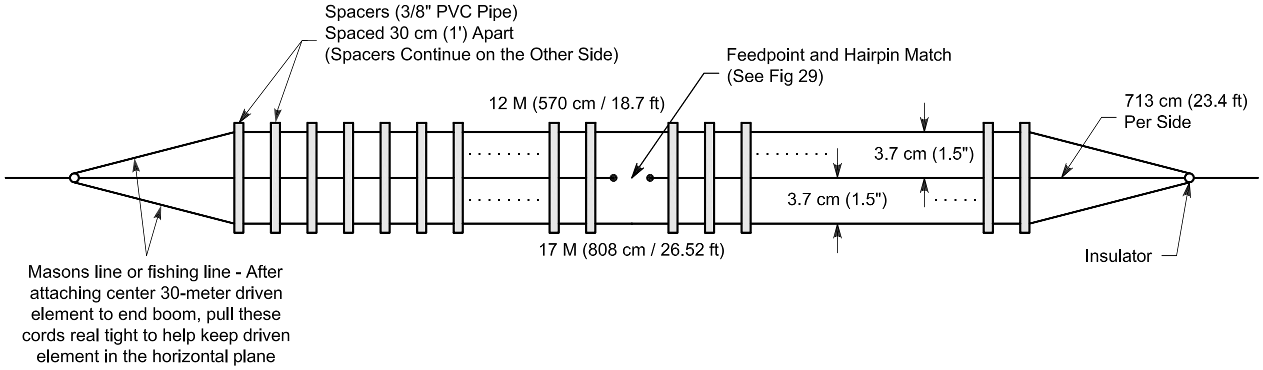
Spacers (3/8" PVC Pipe)
Spaced 30 cm (1') Apart
(Spacers Continue on the Other Side)
Feedpoint and Hairpin Match
(See Fig 29)
12 M (570 cm / 18.7 ft)
713 cm (23.4 ft)
Per Side
3.7 cm (1.5")
3.7 cm (1.5")
17 M (808 cm / 26.52 ft)
Insulator
Masons line or fishing line - After
attaching center 30-meter driven
element to end boom, pull these
cords real tight to help keep driven
element in the horizontal plane

Fig 30—Layout of 30-meter driven element with coupled resonators for 17 and 12 meters.

V Slings

Since the author wanted to be able to raise the 30/17/12-meter triband antenna by himself, he again used only one rope on either end of the array. One end goes over a tree limb and the other end is tied to a stake in the ground or some other nearby support, perhaps a tree trunk. Using only one attachment rope on either end makes it very easy to change beam direction by walking the antenna around the antenna support tree or tower. To accomplish this he used two V slings, one on each end attached to the 2 × 2-inch spreaders.

The secret to keeping the antenna level in the horizontal plane is that the V slings are not equilateral in shape. The combined weight of driven elements, balun and feed line is heavier than the reflectors. If the length of the sides of the V are equal, the array will rotate downwards. The driven elements will end up facing the ground, with the reflectors facing up. Adjust the V slings so that the antenna will stay level in the horizontal plane by shortening the length of the side of the V attached to the driven elements. It is quite easy to adjust in the field, and once you have it adjusted it stays balanced.

Once you raise the antenna to its operating position and in the horizontal plane, you can change direction 180° by pulling on the feed line. As you pull, the whole array will slowly turn over. Stop it from turning by holding onto the feed line once the array has swung over to face the opposite direction.

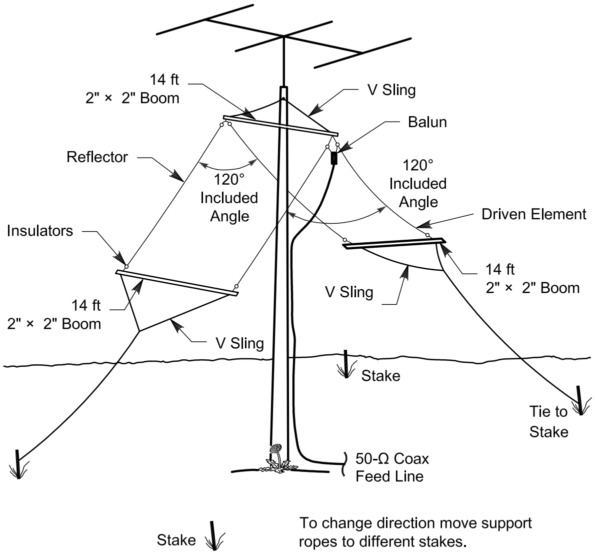
14 ft
2" × 2" Boom
V Sling
Balun
Reflector
120°
Included
Angle
120°
Included
Angle
Driven Element
Insulators
14 ft
2" × 2" Boom
V Sling
14 ft
2" × 2" Boom
V Sling
Stake
Tie to
Stake
50-Ω Coax
Feed Line
Stake
To change direction move support
ropes to different stakes.

Fig 31—Layout for inverted-V 40-meter portable wire Yagi suspended from tower.

SWR Adjustment

Since you may situate your antenna in an entirely different position than VE7CA did, you may need to fine tune your antenna. Begin with the dimensions in Table 5 as a starting point. Make a temporary shorting bar using two alligator clips joined by a piece of wire and attach them at the recommended position. Next raise the antenna to the desired final position. Using an antenna analyzer (or transmitter and SWR meter) plot the SWR over all three bands. Start with the lowest band, 30 meters and adjust the shorting bar up or down to find the lowest SWR point in the portion of the band you plan to operate in. (This procedure also works if you wish to adjust the 20/15/10-meter tribander. Start with the lowest frequency.)

You shouldn't have to move the shorting bar very far from the suggested length. Now that you have determined the right shorting-bar position, adjust the other two driven element lengths for the lowest SWR in the portion of those bands in which you plan to operate. You may have to compromise with the position of the shorting bar to find a satisfactory range where the SWR is acceptable on all three bands. After satisfying himself with the position of the shorting bar, VE7CA replaced the alligator clips simply by folding one side of the parallel hair-pin wire lengths over to the other side and soldering it in the position where the alligator clips had been attached. The author does not recommend changing the reflector element lengths unless you are familiar with antenna modeling programs and are willing to model different spacing or element lengths.

40-Meter Wire Yagis

After his November 2001 *QST* article, VE7CA received several requests for a 40-meter wire Yagi. One ham mentioned that he wanted to be able to pull up a 40-meter Yagi between two towers and to be able to flip it over to change direction. Another wanted a 40-meter Yagi he could pull up on a single tower for the winter DX contests and then put it away during the summer. So VE7CA ran four different 40-meter scenarios in his computer models:

1. A horizontal 2-element wire Yagi at 65 feet.
2. A sloping 2-element wire Yagi, with one end at 65 feet and sloping downward 30° from vertical.
3. A sloping 2-element wire Yagi, with one end at 65 feet and sloping downward 65° from vertical.
4. An inverted-V 2-element wire Yagi with the apex at 65 feet and an included angle between the wires of 120°.

Fig 31 shows the layout for an inverted-V system and **Table 6** lists the element and hairpin lengths. Elevation patterns for the 40 meter antennas are compared in **Fig 32**, with a reference antenna of a single flattop dipole at 65 feet. As they say, a picture is worth a thousand words. If your interest is DX, it is very clear that horizontal and high is a very good rule of thumb for most antennas.

Yes, a $^1/_4$ wave vertical over salt water or 120 $^1/_4$-wave radials over good ground will produce very low radiation angles, but such systems are not exactly portable and we don't all live near the ocean. Mind you, if you can manage to locate antenna Number 3 (the most vertically oriented wire Yagi) next to the ocean, you would be very happy.

The point here is that if you have two towers and you're not fortunate enough to be located on a saltwater marsh, you should pull the 40-meter array up as high and as horizontal as you can. If you have only one tower and don't need to change direction often, then try the inverted-V configuration. You can still change the direction by walking each end around the tower.

However, even the sloping 40-meter Yagi with one end at 65 feet up a tower (or tree) and the other end attached with a long rope as far as possible from the tower will still put out a very respectable signal. It is directional, and you can walk it around the tower to change direction or you can

Table 6
40-Meter Wire Yagi Configurations

Configuration	*Driven Element cm (feet)*	*Reflector cm (feet)*	*Hairpin Length cm (feet)*
1. Horizontal	1978 (64.90)	2098 (68.83)	Approx 50 cm
2. –30° Sloper	1978 (64.90)	2113 (69.32)	(22 inches)
3. –65° Sloper	1978 (64.90)	2101 (68.93)	
4. Inverted V	2040 (66.93)	2126 (69.75)	

Spacing between driven element and reflector is 427 cm (14 feet). Spacing between parallel hairpin wires is 10 cm (4 inches). The lengths shown above are the total wire length for each element.

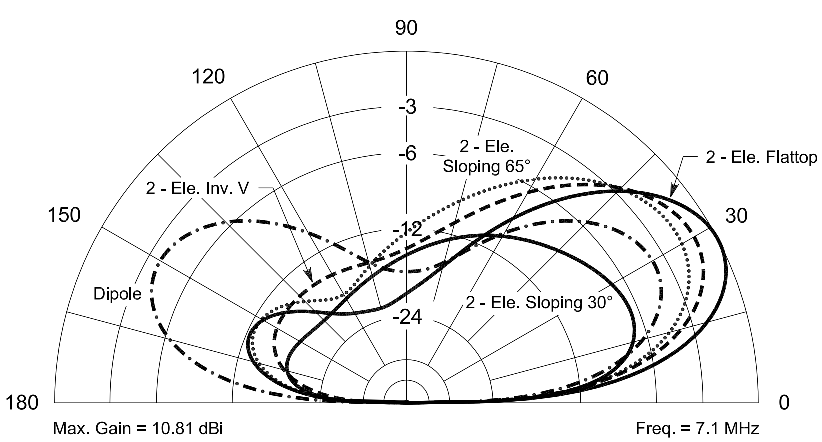

Fig 32—Comparisons of elevation patterns for five 40-meter antennas: a 2-element flattop Yagi at 65 feet; a 2-element inverted-V Yagi at 65 feet; a 2-element Yagi sloped 65° from the vertical plane; a 2-element Yagi sloped 30° from the vertical plane; and a horizontal dipole at 65 feet.

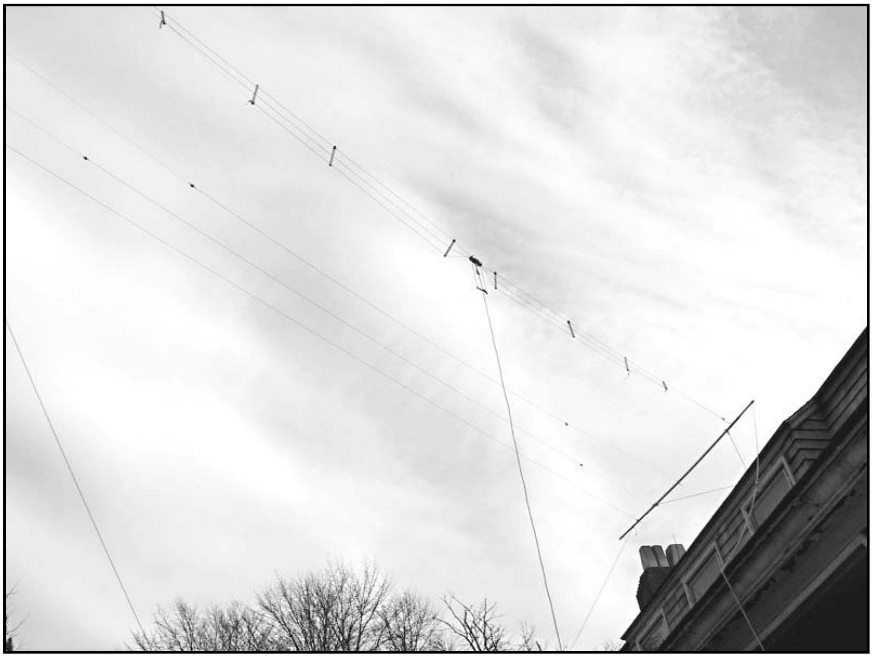

Fig 33 — VE7CA portable 2-element 20/15/10-meter wire Yagi installed at Plimoth Plantation in Plymouth, Massachusetts. (*Photo courtesy K1HTN*).

flip the antenna over and change direction 180° very quickly.

Summary

You don't need a 10-meter (60-foot) high tower, a commercial triband Yagi and a rotator to put out a good signal on the HF bands. Wire Yagis work very well and they can be inexpensive and easy-to-build, using components found at your local hardware store. **Fig 33** is a photograph of a successful portable installation in Massachusetts.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

G. Breed, K9AY, "The Coupled-Resonator Principle: A Flexible Method for Multiband Antennas," *The ARRL Antenna Compendium, Vol 5* (Newington: ARRL, 1996), p 109ff.

R. J. Decesari, "A Portable Quad for 2 Meters," *QST*, Oct 1980, and "Portable Quad for 2 Meters, Part 2," Technical Correspondence, *QST*, Jun 1981, p 39ff.

D. DeMaw, "A Traveling Ham's Trap Vertical," *QST*, Oct 1980, p 28ff.

J. Hall, "Zip-Cord Antennas—Do They Work?" *QST*, Mar 1979, p 31ff.

M. Hansen, "A Portable 2-Element Triband Yagi," *QST*, Nov 2001, p 35ff.

M. Hansen, "Some More VE7CA 2-Element Portable Yagis," *The ARRL Antenna Compendium, Vol7* (Newington: ARRL, 2002, p 183ff.

C. L. Hutchinson, "A Tree-Mounted 30-Meter Ground-Plane Antenna," *QST*, Sep 1974, p 16ff.

C. W. Schecter, "A Deluxe RV 5-Band Antenna," *QST*, Oct 1980, p 38.

Chapter 16

Mobile and Maritime Antennas

Mobile antennas are those designed for use while antennas, most amateurs think of a whip mounted on an automobile or other highway vehicle, perhaps on a recreational vehicle (RV) or maybe on an off-road vehicle. While it is true that most mobile antennas are vertical whips, mobile antennas can also be found in other places. For example, antennas mounted aboard a boat or ship are mobile, and are usually called *maritime antennas*. **Fig 1** shows yet another type of mobile antenna—those for use on handheld transceivers. Because they may be used while in motion, even these antennas are mobile by literal definition.

Pictured in Fig 1 is a telescoping full-size quarter-wave antenna for 144 MHz, and beside it a *stubby* antenna for the same band. The stubby is a helically wound radiator, made of stiff copper wire enclosed in a protective covering of rubber-like material. The inductance of the helical windings provides electrical loading for the antenna. For frequencies above 28 MHz, most mobile installations permit the use of a full-size antenna, but sometimes smaller, loaded antennas are used for convenience. The stubby, for example, is convenient for short-range communications, avoiding the problems of a lengthier, cumbersome antenna attached to a handheld radio.

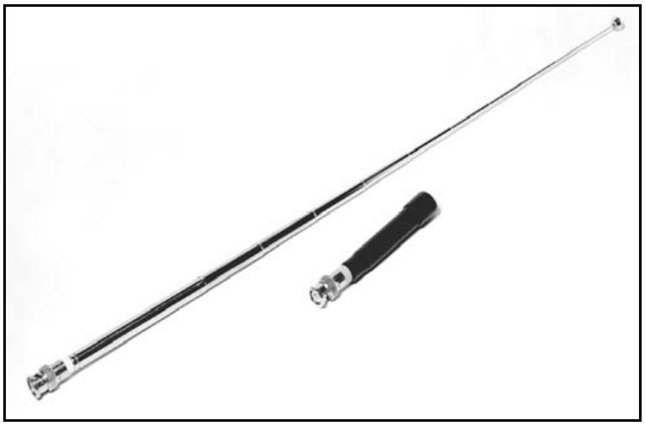

Fig 1—Two mobile antennas—mobile because they may be used while in motion. Shown here are a telescoping $^1/_4$-λ antenna and a "stubby" antenna, both designed for use at 144 MHz. The $^1/_4$-λ antenna is 19 inch long, while the stubby antenna is only $3^1/_2$ inch long. (Both dimensions exclude the length of the BNC connectors. The stubby is a helically wound radiator.

Below 28 MHz, physical size becomes a problem with full-size whips, and some form of electrical loading (as with the stubby) is usually employed. Commonly used loading techniques are to place a coil at the base of the whip (base loading), or at the center of the whip (center loading). These and other techniques are discussed in this chapter.

Few amateurs construct their own antennas for HF mobile and maritime use, since safety reasons dictate very sound mechanical construction. Several construction projects are included, however, in this chapter for those who may wish to build their own mobile antenna. Even if commercially made antennas are installed, most require some adjustment for the particular installation and type of operation desired and the information given here may provide a better understanding of the optimization requirements.

HF-MOBILE FUNDAMENTALS

Fig 2 shows a typical bumper-mounted center-loaded whip suitable for operation in the HF range. Jack Schuster, W1WEF, operates 80 through 2 meters from his car. The antenna could also be mounted on the car body itself (such as a fender), and mounts are available for this purpose. The base spring and tennis ball act as shock absorbers for the bottom of the whip, as the continual flexing while in motion would otherwise weaken the antenna. A short heavy mast section is mounted between the base spring and loading coil. Some models have a mechanism that allows the antenna to be tipped over for adjustment or for fastening to the roof of the car when not in use.

It is also advisable to extend a couple of guy lines from the base of the loading coil to clips or hooks fastened to the roof gutter on the car, or to the trunk and

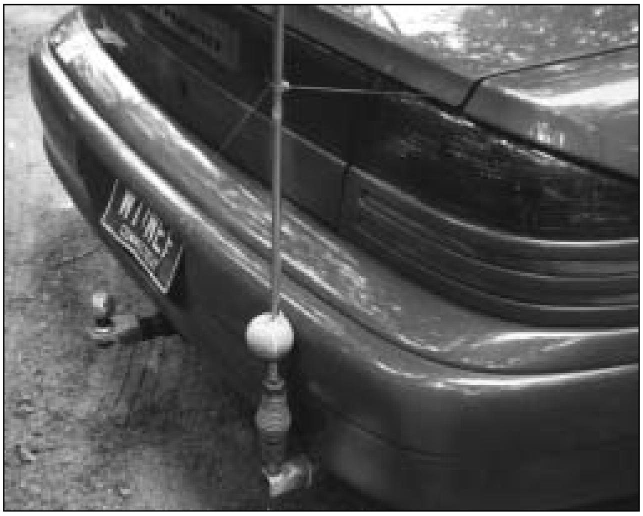

Fig 2—A typical bumper-mounted HF mobile antenna, as used by W1WEF. Note the nylon guy lines and the tennis ball used as a shock absorber. (*Photo courtesy W1WEF.*)

rear bumper, as W1WEF has done. Nylon fishing line (about 40-pound test) is suitable for this purpose. The guy lines act as safety cords and also reduce the swaying motion of the antenna considerably. The feed line to the transmitter is connected to the bumper and base of the antenna. Good low-resistance connections are important here.

Tune-up of the antenna is usually accomplished by changing the height of the adjustable whip section above the precut loading coil. First, tune the receiver and try to determine where the signals seem to peak up. Once this frequency is found, check the SWR with the transmitter on, and find the frequency where the SWR is lowest. Shortening the adjustable section will increase the resonant frequency, and making it longer will lower the frequency. It is important that the antenna be away from surrounding objects such as overhead wires by ten feet or more, as considerable detuning can occur. Once the setting is found where the SWR is lowest at the center of the desired frequency range, the length of the adjustable section should be recorded.

Propagation conditions and ignition noise are usually the limiting factors for mobile operation on 10 through 28 MHz. Antenna size restrictions affect operation somewhat on 7 MHz and much more on 3.5 and 1.8 MHz. From this standpoint, perhaps the optimum band for HF-mobile operation is 7 MHz. The popularity of the regional mobile nets on 7 MHz is perhaps the best indication of its suitability. For local work, 28 MHz is also useful, as antenna efficiency is high and relatively simple antennas without loading coils are easy to build.

As the frequency of operation is lowered, an antenna of fixed length looks (at its feed point) like a decreasing resistance in series with an increasing capacitive reactance. The capacitive reactance must be tuned out, necessitating the use of an equivalent series inductive reactance or loading coil. The amount of inductance required will be determined by the placement of the coil in the antenna system.

Base loading requires the lowest value of inductance for a fixed-length antenna, and as the coil is placed farther up the whip, the necessary value increases. This is because the capacitance of the shorter antenna section (above the coil) to the car body is now lower (higher capacitive reactance), requiring more inductance to tune the antenna to resonance. The advantage is that the current distribution on the whip is improved, increasing the radiation resistance. The disadvantage is that requirement of a larger coil also means the coil size and losses increase. Center loading has been generally accepted as a good compromise with minimal construction problems. Placing the coil $^2/_3$ the distance up the whip seems to be about the optimum position.

For typical antenna lengths used in mobile work, the difficulty in constructing suitable loading coils increases as the frequency of operation is lowered. Since the required resonating inductance gets larger and the radiation resistance decreases at lower frequencies, most of the power is dissipated in the coil's loss resistance and in other ohmic losses. This is one reason why it is advisable to buy a commercially made loading coil with the highest power rating possible, even if only low-power operation is planned.

Coil losses in the higher-power loading coils are usually less (percentage-wise), with subsequent improvement in radiation efficiency, regardless of the power level used. Of course, the above philosophy also applies to homemade loading coils, and design considerations will be considered in a later section.

Once the antenna is tuned to resonance, the input impedance at the antenna terminals will look like a pure resistance. Neglecting losses, this value drops from nearly 15 Ω at 21 MHz to 0.1 Ω at 1.8 MHz for an 8-foot whip. When coil and other losses are included, the input resistance increases to approximately 20 Ω at 1.8 MHz and 16 Ω at 21 MHz. These values are for relatively high-efficiency systems. From this it can be seen that the radiation efficiency is much poorer at 1.8 MHz than at 21 MHz under typical conditions.

Since most modern gear is designed to operate with a 50-Ω transmission line, a matching network is usually necessary, especially with the high-efficiency antennas previously mentioned. This can take the form of either a broadband transformer, a tapped coil, or an LC matching network. With homemade or modified designs, the tapped-coil arrangement is perhaps the easiest one to build, while the broadband transformer requires no adjustment. As the losses go up, so does the input resistance, and in less efficient systems the matching network may not be needed.

The Equivalent Circuit of a Typical Mobile Antenna

In the previous section, some of the general considerations were discussed, and these will now be taken up in more detail. It is customary in solving problems involving electric and magnetic fields (such as antenna systems) to try to find an equivalent network with which to replace the antenna for analysis reasons. In many cases, the network may be an accurate representation over only a limited frequency range. However, this is often a valuable method in matching the antenna to the transmission line.

Antenna resonance is defined as the frequency at which the input impedance at the antenna terminals is purely resistive. The shortest length at which this occurs for a vertical antenna over a ground plane is when the antenna is an electrical quarter wavelength at the operating frequency; the impedance value for this length (neglecting losses) is about 36 Ω. The idea of resonance can be extended to antennas shorter (or longer) than a quarter wave, and means only that the input impedance is purely resistive. As pointed out previously, when the frequency is lowered, the antenna looks like a series RC circuit, as shown in **Fig 3**. For the average 8-foot whip, the capacitive reactance of C_A may range from about –150 Ω at 21 MHz to as high as –8000 Ω at 1.8 MHz, while the radiation resistance R_R varies from about 15 Ω at 21 MHz to as low as 0.1 Ω at 1.8 MHz.

For an antenna less than 0.1 λ long, the approximate radiation resistance may be determined from the following:

$$R_R = 273 \times (\ell f)^2 \times 10^{-8} \qquad \text{(Eq 1)}$$

where ℓ is the length of the whip in inches, and f is the frequency in megahertz.

Since the resistance is low, considerable current must flow in the circuit if any appreciable power is to be dissipated in the form of radiation in R_R. Yet it is apparent that little current can be made to flow in the circuit as long as the comparatively high series reactance remains.

Antenna Capacitance

Capacitive reactance can be canceled by connecting an equivalent inductive reactance, (coil L_L) in series, as shown in **Fig 4**, thus tuning the system to resonance.

The capacitance of a vertical antenna shorter than a quarter wavelength is given by:

$$C_A = \frac{17\ell}{\left[\left(\ln\frac{24\ell}{D}\right)-1\right]\left[1-\left(\frac{f\ell}{234}\right)^2\right]} \qquad \text{(Eq 2)}$$

where

C_A = capacitance of antenna in pF
ℓ = antenna height in feet
D = diameter of radiator in inches
f = operating frequency in MHz

$$\ln\frac{24\ell}{D} = 2.3\log_{10}\frac{24\ell}{D}$$

Fig 5 shows the approximate capacitance of whip antennas of various average diameters and lengths. For 1.8, 4 and 7 MHz, the loading coil inductance required (when the loading coil is at the base) would be approximately the inductance required to resonate in the desired band (with the whip capacitance taken from the graph). For 10 through 21 MHz, this rough calculation will give more than the required inductance, but it will serve as a

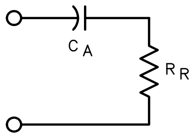

Fig 3—At frequencies below resonance, the whip antenna will show capacitive reactance as well as resistance. R_R is the radiation resistance, and C_A represents the antenna capacitance.

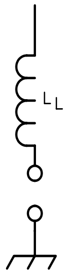

Fig 4—The capacitive reactance at frequencies below the resonant frequency of the whip can be canceled by adding an equivalent inductive reactance in the form of a loading coil in series with the antenna.

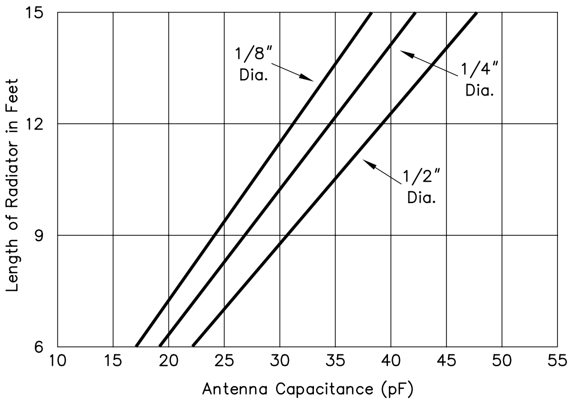

Fig 5—Graph showing the approximate capacitance of short vertical antennas for various diameters and lengths. These values should be approximately halved for a center-loaded antenna.

starting point for the final experimental adjustment that must always be made.

LOADING COIL DESIGN

To minimize loading coil loss, the coil should have a high ratio of reactance-to-resistance (that is, a high unloaded Q). A 4-MHz loading coil wound with small wire on a small-diameter solid form of poor quality, and enclosed in a metal protector, may have a Q as low as 50, with a loss resistance of 50 Ω or more. High-Q coils require a large conductor, air-wound construction, large spacing between turns, and the best insulating material available. A diameter not less than half the length of the coil (not always mechanically feasible) and a minimum of metal in the field of the coil are also necessities for optimum efficiency. Such a coil for 4 MHz may show a Q of 300 or more, with a resistance of 12 Ω or less.

The coil could then be placed in series with the feed line at the base of the antenna to tune out the unwanted capacitive reactance, as shown in Fig 4. Such a method is often referred to as *base-loading*, and many practical mobile antenna systems have been built using this scheme.

Over the years, the question has come up as to whether or not more efficient designs are possible compared with simple base loading. While many ideas have been tried with varying degrees of success, only a few have been generally accepted and incorporated into actual systems. These are center loading, continuous loading, and combinations of the latter with more conventional antennas.

Base Loading and Center Loading

If a whip antenna is short compared to a wavelength *and the current is uniform along the length ℓ*, the electric field strength E, at a distance d, away from the antenna is approximately:

$$E = \frac{120\pi I\ell}{d\lambda} \qquad \text{(Eq 3)}$$

where

I is the antenna current in amperes
λ is the wavelength in the same units as D and ℓ.

A uniform current flowing along the length of the whip is an idealized situation, however, since the current is greatest at the base of the antenna and goes to a minimum at the top. In practice, the field strength will be less than that given by the above equation, because it is a function of the current distribution on the whip.

The reason that the current is not uniform on a whip antenna can be seen from the circuit approximation shown in **Fig 6**. A whip antenna over a ground plane is similar in many respects to a tapered coaxial cable where the center conductor remains the same diameter along its length, but with an increasing diameter outer conductor. The inductance per unit length of such a cable would increase along the line, while the capacitance per unit length would decrease. In Fig 6 the antenna is represented by a series of LC circuits in which C1 is greater than C2, which is greater than C3, and so on. L1 is less than L2, which is less than succeeding inductances. The net result is that most of the antenna current returns to ground near the base of the antenna, and very little near the top.

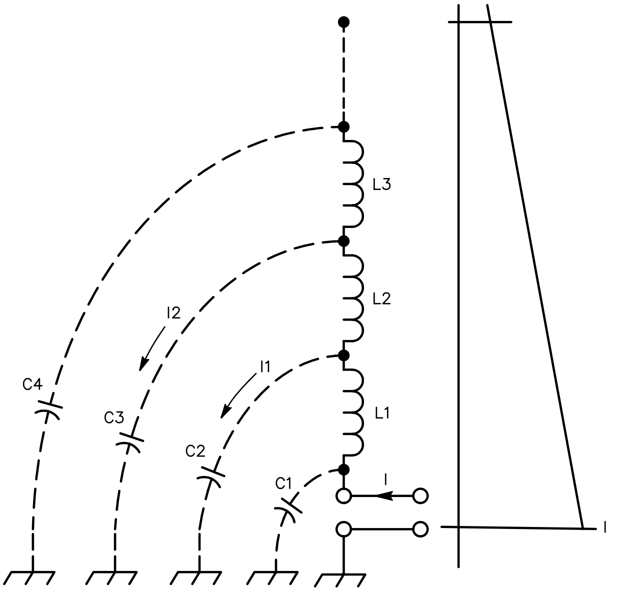

Fig 6—A circuit approximation of a simple whip over a perfectly conducting ground plane. The shunt capacitance per unit length gets smaller as the height increases, and the series inductance per unit length gets larger. Consequently, most of the antenna current returns to the ground plane near the base of the antenna, giving the current distribution shown at the right.

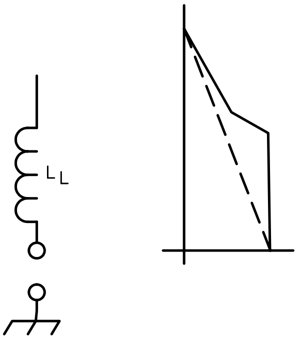

Fig 7—Improved current distribution resulting from center loading.

Two things can be done to improve this distribution and make the current more uniform. One would be to increase the capacitance of the top of the antenna to ground through the use of top loading or a capacitance hat, as discussed in Chapter 6. Unfortunately, the wind resistance of the hat makes it somewhat unwieldy for mobile use. The other method is to place the loading coil farther up the whip, as shown in **Fig 7**, rather than at the base. If the coil is resonant (or nearly so) with the

Table 1
Approximate Values for 8-ft Mobile Whip

f(MHz)	Loading L µH	R_C(Q50) Ω	R_C(Q300) Ω	R_R Ω	Feed R* Ω	Matching L µH
Base Loading						
1.8	345	77	13	0.1	23	3
3.8	77	37	6.1	0.35	16	1.2
7.2	20	18	3	1.35	15	0.6
10.1	9.5	12	2	2.8	12	0.4
14.2	4.5	7.7	1.3	5.7	12	0.28
18.1	3.0	5.0	1.0	10.0	14	0.28
21.25	1.25	3.4	0.5	14.8	16	0.28
24.9	0.9	2.6	—	20.0	22	0.25
29.0	—	—	—	—	36	0.23
Center Loading						
1.8	700	158	23	0.2	34	3.7
3.8	150	72	12	0.8	22	1.4
7.2	40	36	6	3.0	19	0.7
10.1	20	22	4.2	5.8	18	0.5
14.2	8.6	15	2.5	11.0	19	0.35
18.1	4.4	9.2	1.5	19.0	22	0.31
21.25	2.5	6.6	1.1	27.0	29	0.29

R_C = loading coil resistance; R_R = radiation resistance.
*Assuming loading coil Q = 300, and including estimated ground-loss resistance.

Table 2
Suggested Loading Coil Dimensions

Req'd L(µH)	Turns	Wire Size	Dia. Inches	Length Inches
700	190	22	3	10
345	135	18	3	10
150	100	16	2½	10
77	75	14	2½	10
77	29	12	5	4¼
40	28	16	2½	2
40	34	12	2½	4¼
20	17	16	2½	1¼
20	22	12	2½	2¾
8.6	16	14	2	2
8.6	15	12	2½	3
4.5	10	14	2	1¼
4.5	12	12	2½	4
2.5	8	12	2	2
2.5	8	6	2⅜	4½
1.25	6	12	1¾	2
1.25	6	6	2⅜	4½

capacitance to ground of the section above the coil, the current distribution is improved as also shown in Fig 7. The result with both top loading and center loading is that the radiation resistance is increased, offsetting the effect of losses and making matching easier.

Table 1 shows the approximate loading coil inductance for the various amateur bands. Also shown in the table are approximate values of radiation resistance to be expected with an 8-foot whip, and the resistances of loading coils—one group having a Q of 50, the other a Q of 300. A comparison of radiation and coil resistances will show the importance of reducing the coil resistance to a minimum, especially on the three lower frequency bands. **Table 2** shows suggested loading-coil dimensions for the inductance values given in Table 1.

OPTIMUM DESIGN OF SHORT COIL-LOADED HF MOBILE ANTENNAS

Optimum design of short HF mobile antennas results from a careful balance of the appropriate loading coil Q-factor, loading coil position in the antenna, ground loss resistance, and the length-to-diameter ratio of the antenna. The optimum balance of these parameters can be realized only through a thorough understanding of how they interact. This section presents a mathematical approach to designing mobile antennas for maximum radiation efficiency. Bruce Brown, W6TWW, in *The ARRL Antenna Compendium Volume 1*, first presented this approach. (See the Bibliography at the end of this chapter.)

The optimum location for a loading coil in an antenna can be found experimentally, but it requires many hours of designing and constructing models and making measurements to ensure the validity of the design. A faster and more reliable way of determining optimum coil location is through the use of a personal computer. This approach allows the variation of any single variable, while observing the cumulative effects on the system. When plotted graphically, the data reveals that the placement of the loading coil is critical if maximum radiation efficiency

Table 3
Variables used in Eqs 4 through 20

A = area in degree-amperes
a = antenna radius in English or metric units
dB = signal loss in decibels
E = efficiency in percent
f (MHz) = frequency in megahertz
H = height in English or metric units
h = height in electrical degrees
h_1 = height of base section in electrical degrees
h_2 = height of top section in electrical degrees
I = I_{base} = 1 ampere base current
k = 0.0128
k_m = mean characteristic impedance
k_{m1} = mean characteristic impedance of base section
k_{m2} =mean characteristic impedance of top section
L = length or height of the antenna in feet
P_I = power fed to the antenna
P_R = power radiated
Q = coil figure of merit
R_C = coil loss resistance in Ω
R_G = ground loss resistance in Ω
R_R = radiation resistance in Ω
X_L = loading-coil inductive reactance

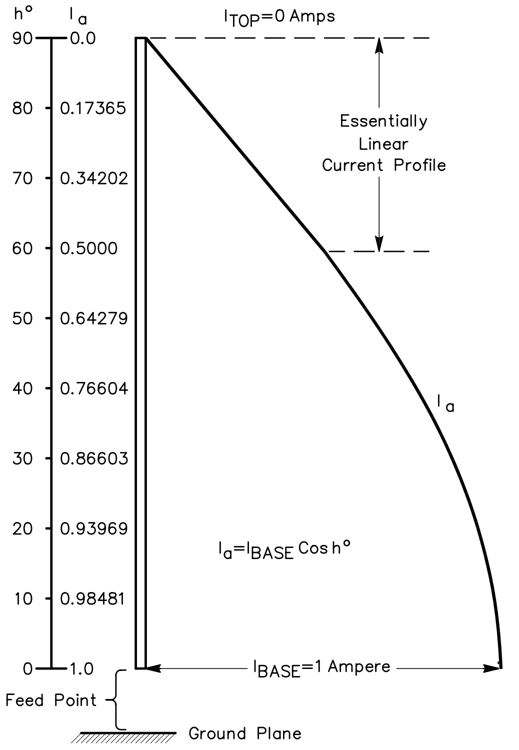

Fig 8—Relative current distribution on a vertical antenna of height h = 90 electrical degrees.

is to be realized. (See the program *MOBILE.EXE*, which is included on the CD-ROM in the back of this book.)

Radiation Resistance

The determination of radiation efficiency requires the knowledge of resistive power losses and radiation losses. Radiation loss is expressed in terms of radiation resistance. Radiation resistance is defined as the resistance that would dissipate the same amount of power that is radiated by the antenna. The variables used in the equations that follow are defined once in the text, and are summarized in **Table 3**. Radiation resistance of vertical antennas shorter than 45 electrical degrees ($^1/_8$ wavelength) is approximately:

$$R_R = \frac{h^2}{312} \qquad \text{(Eq 4)}$$

where

R_R = radiation resistance in Ω
h = antenna length in electrical degrees.

Antenna height in electrical degrees is expressed by:

$$h = \frac{\ell}{984} \times f\left(\text{MHz}\right) \times 360 \qquad \text{(Eq 5)}$$

where

ℓ = antenna length in feet
f (MHz) = operating frequency in megahertz.

End effect is purposely omitted to ensure that an antenna is electrically long. This is so that resonance at the design frequency can be obtained easily by removing a turn or two from the loading coil.

Eq 4 is valid only for antennas having a sinusoidal current distribution and no reactive loading. However, it can be used as a starting point for deriving an equation that is useful for shortened antennas with other than sinusoidal current distributions.

Refer to **Fig 8**. The current distribution on an antenna 90° long electrically ($^1/_4$ wavelength) varies with the cosine of the length in electrical degrees. The current distribution of the top 30° of the antenna is essentially linear. It is this linearity that allows for derivation of a simpler, more useful equation for radiation resistance.

The radiation resistance of an electrically short base-loaded vertical antenna can be conveniently defined in terms of a geometric figure, a triangle, as shown in **Fig 9**. The radiation resistance is given by:

$$R_R = KA^2 \qquad \text{(Eq 6)}$$

where

K is a constant (to be derived shortly)
A = area of the triangular current distribution in degree-amperes.

Degree-ampere area is expressed by

$$A = \frac{1}{2} h \times I_{base} \qquad \text{(Eq 7)}$$

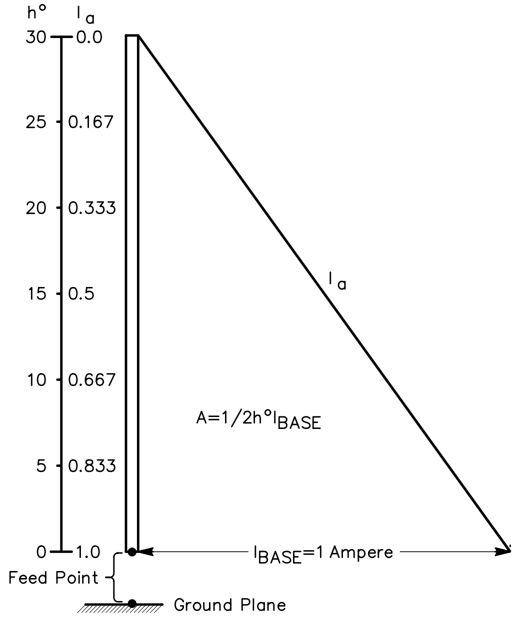

Fig 9—Relative current distribution on a base-loaded vertical antenna of height H = 30 electrical degrees (linearized). The base loading coil is not shown here.

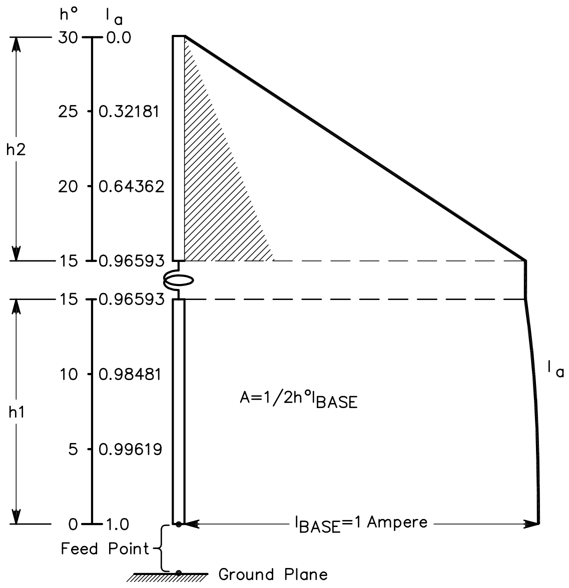

Fig 10—Relative current distribution on a center-loaded antenna with base and top sections each equal to 15 electrical degrees in length. The cross-hatched area shows the current distribution that would exist in the top 15° of a 90°-high vertical fed with 1 ampere at the base.

By combining Eqs 4 and 6 and solving for K, we get

$$K = \frac{h^2}{312 \times A^2} \qquad \text{(Eq 8)}$$

By substituting the values from Fig 9 into Eq 8 we get

$$K = \frac{30^2}{312 \times \left(0.5 \times 30 \times 1\right)^2} = 0.0128$$

and by substituting the derived value of K into Eq 6 we get

$$R_R = 0.0128 \times A^2 \qquad \text{(Eq 9)}$$

Eq 9 is useful for determining the radiation resistance of coil-loaded vertical antennas less than 30° in length. The derived constant differs slightly from that presented by Laport (see Bibliography), as he used a different equation for radiation resistance Eq 4.

When the loading coil is moved up an antenna (away from the feed point), the current distribution is modified as shown in **Fig 10**. The current varies with the cosine of the height in electrical degrees at any point in the base section. Therefore, the current flowing into the bottom of the loading coil is less than the current flowing at the base of the antenna.

But what about the current in the top section of the antenna? The loading coil acts as the lumped constant that it is, and disregarding losses and coil radiation, maintains the same current flow throughout. As a result, the current at the top of a high-Q coil is essentially the same as that at the bottom of the coil. This is easily verified by installing RF ammeters immediately above and below the loading coil in a test antenna. Thus, the coil "forces" much more current into the top section than would flow in the equivalent section of a full 90° long antenna. This occurs as a result of the extremely high voltage that appears at the top of the loading coil. This higher current flow results in more radiation than would occur from the equivalent section of a quarter-wave antenna. (This is true for conventional coils. However, radiation from long thin coils allows coil current to decrease, as in helically wound antennas.)

The cross-hatched area in Fig 10 shows the current that would flow in the equivalent part of a 90° high antenna, and reveals that the degree-ampere area of the whip section of the short antenna is greatly increased as a result of the modified current distribution. The current flow in the top section decreases almost linearly to zero at the top. This can be seen in Fig 10.

The degree-ampere area of Fig 10 is the sum of the triangular area represented by the current distribution in the top section, and the nearly trapezoidal current distribution in the base section. Radiation from the coil is not included in the degree-ampere area because it is small and difficult to define. Any radiation from the coil can be considered a bonus.

The degree-ampere area is expressed by:

$$A = \frac{1}{2}\left[h_1\left(1 + \cos h_1\right) + h_2\left(\cos h_1\right)\right] \qquad \text{(Eq 10)}$$

where

h_1 = electrical length in degrees of the base section

h_2 = electrical height in degrees of the top section.

The degree-ampere area (calculated by substituting Eq 10 into Eq 9) can be used to determine the radiation resistance when the loading coil is at any position other than the base of the antenna. Radiation resistance has been calculated with these equations and plotted against loading coil position at three different frequencies for 8- and 11-foot antennas, **Fig 11**. Eight feet is a typical length for commercial antennas, and 11-foot antennas are about the maximum practical length that can be installed on a vehicle.

In Fig 11, the curves reveal that the radiation resistance rises almost linearly as the loading coil is moved up the antenna. They also show that the radiation resistance rises rapidly as the frequency is increased. If the analysis were stopped at this point, one might conclude that the loading coil should be placed at the top of the antenna. This is not so, and will become apparent shortly.

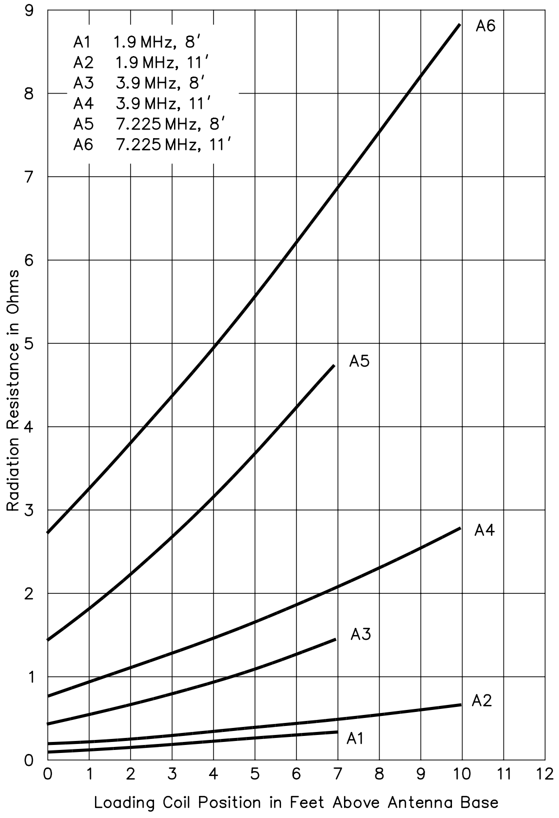

Fig 11—Radiation resistance plotted as a function of loading coil position.

Required Loading Inductance

Calculation of the loading coil inductance needed to resonate a short antenna can be done easily and accurately by using the antenna transmission-line analog described by Boyer in *Ham Radio*. For a base-loaded antenna, Fig 9, the loading coil reactance required to resonate the antenna is given by

$$X_L = -j\,K_m \cot h \qquad \text{(Eq 11)}$$

where

X_L = inductive reactance required

K_m = mean characteristic impedance (defined in Eq 12).

The $-j$ term indicates that the antenna presents capacitive reactance at the feed point. A loading coil must cancel this reactance.

The mean characteristic impedance of an antenna is expressed by

$$K_m = 60\left[\left(\ln \frac{2H}{a}\right) - 1\right] \qquad \text{(Eq 12)}$$

where

H = physical antenna height (excluding the length of the loading coil)

a = radius of the antenna in the same units as H.

From Eq 12 you can see that decreasing the height-to-diameter ratio of an antenna by increasing the radius results in a decrease in K_m. With reference to Eq 11, a decrease in K_m decreases the inductive reactance required to resonate an antenna. As will be shown later, this will increase radiation efficiency. In mobile applications, we quickly run into wind-loading problems if we attempt to use an antenna that is physically large in diameter.

If the loading coil is moved away from the base of the antenna, the antenna is divided into a base and top section, as depicted in Fig 10. The loading coil reactance required to resonate the antenna when the coil is away from the base is given by

$$X_L = j\,K_{m2}\,(\cot h_2) - j\,K_{m1}\,(\tan h_1) \qquad \text{(Eq 13)}$$

In mobile-antenna design and construction, the top section is usually a whip with a much smaller diameter than the base section. Because of this, it is necessary to compute separate values of K_m for the top and base sections. K_{m1} and K_{m2} are the mean characteristic impedances of the base and top sections, respectively.

Loading coil reactance curves for the 3.8-MHz antennas of Fig 11 have been calculated and plotted in **Fig 12**. These curves show the influence of the loading coil position on the reactance required for resonance. The curves in Fig 12 show that the required reactance decreases with longer antennas. The curves also reveal that the required loading coil reactance grows at an increasingly rapid rate after the coil passes the center of the antenna. Because the highest possible loading coil Q

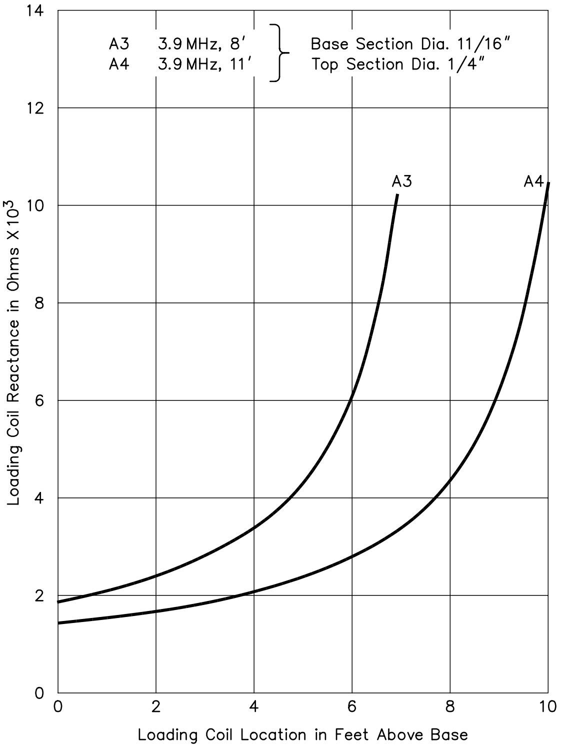

Fig 12—Loading coil reactance required for resonance, plotted as a function of coil height above the antenna base. The resonant frequency is 3.9 MHz.

factor is needed, and because optimum Q is attained when the loading coil diameter is twice the loading coil length, the coil would grow like a smoke ring above the center of the antenna, and would quickly reach an impractical size. It is for this reason that the highest loading coil position is limited to one foot from the top of the antenna in all computations.

Loading Coil Resistance

Loading coil resistance constitutes one of the losses that consumes power that could otherwise be radiated by the antenna. Heat loss in the loading coil is not of any benefit, so it should be minimized by using the highest possible loading coil Q. Loading coil loss resistance is a function of the coil Q and is given by

$$R_C = \frac{X_L}{Q} \qquad \text{(Eq 14)}$$

where

R_C = loading coil loss resistance in Ω
X_L = loading coil reactance
Q = coil figure of merit

Inspection of Eq 14 reveals that, for a given value of inductive reactance, loss resistance will be lower for higher Q coils. Measurements made with a Q meter show that typical, commercially manufactured coil stock produces a Q between 150 and 160 in the 3.8-MHz band.

Higher Q values can be obtained by using larger diameter coils having a diameter-to-length ratio of two, by using larger diameter wire, by using more spacing between turns, and by using low-loss polystyrene supporting and enclosure materials. In theory, loading coil turns should not be shorted for tuning purposes because shorted turns somewhat degrade Q. Pruning to resonance should be done by removing turns from the coil.

In fairness, it should be pointed out that many practical mobile antennas use large-diameter loading coils with shorted turns to achieve resonance. The popular “Texas Bug Catcher” coils come to mind here. Despite general proscriptions against shorting turns, these systems are often more efficient than antennas with small, relatively low-Q, fixed loading coils.

Radiation Efficiency

The ratio of power radiated to power fed to an antenna determines the radiation efficiency. It is given by:

$$E = \frac{P_R}{P_I} \times 100\% \qquad \text{(Eq 15)}$$

where

E = radiation efficiency in percent
P_R = power radiated
P_I = power fed to the antenna at the feed point.

In a short, coil-loaded mobile antenna, a large portion of the power fed to the antenna is dissipated in ground and coil resistances. A relatively insignificant amount of power is also dissipated in the antenna conductor resistance and in the leakage resistance of the base insulator. Because these last two losses are both very small and difficult to estimate, they are here neglected in calculating radiation efficiency.

Another loss worth noting is matching network loss. Because we are concerned only with power fed to the antenna in the determination of radiation efficiency, matching network loss is not considered in any of the equations. Suffice it to say that matching networks should be designed for minimum loss in order to maximize the transmitter power available at the antenna.

The radiation efficiency equation may be rewritten and expanded as follows:

$$E = \frac{I^2\, R_R \times 100}{I^2\, R_R + I^2\, R_G + \left(I \cos h_1\right)^2 R_C} \qquad \text{(Eq 16)}$$

where

I = antenna base current in amperes
R_G = ground loss resistance in Ω
R_C = coil loss resistance in Ω.

Each term of Eq 16 represents the power dissipated in its associated resistance. All the current terms cancel, simplifying this equation to

$$E = \frac{R_R \times 100}{R_R + R_G + R_C \left(\cos^2 h_1\right)} \qquad \text{(Eq 17)}$$

For base-loaded antennas the term $\cos^2 h_1$ drops to unity and may be omitted.

Ground Loss

Eq 14 shows that the total resistive losses in the antenna system are:

$$R_T = R_R + R_G + R_c \left(\cos^2 h_1\right) \qquad \text{(Eq 18)}$$

where R_T is the total resistive loss. Ground loss resistance can be determined by rearranging Eq 18 as follows:

$$R_G = R_T - R_R - R_C \cos^2 h_1 \qquad \text{(Eq 19)}$$

R_T may be measured in a test antenna installation on a vehicle using an R-X noise bridge or an SWR analyzer. You can then calculate R_R and R_C.

Ground loss is a function of vehicle size, placement of the antenna on the vehicle, and conductivity of the ground over which the vehicle is traveling. Only the first two variables can be feasibly controlled. Larger vehicles provide better ground planes than smaller ones. The vehicle ground plane is only partial, so the result is considerable RF current flow (and ground loss) in the ground around and under the vehicle.

By raising the antenna base as high as possible on the vehicle, ground losses are decreased. This results from a decrease in antenna capacitance to ground, which increases the capacitive reactance to ground. This, in turn, reduces ground currents and ground losses.

This effect has been verified by installing the same antenna at three different locations on two different vehicles, and by determining the ground loss from Eq 19. In the first test, the antenna was mounted 6 inches below the top of a large station wagon, just behind the left rear window. This placed the antenna base 4 feet 2 inches above the ground, and resulted in a measured ground loss resistance of 2.5 Ω. The second test used the same antenna mounted on the left rear fender of a mid-sized sedan, just to the left of the trunk lid. In this test, the measured ground loss resistance was 4 Ω. The third test used the same mid-sized car, but the antenna was mounted on the rear bumper. In this last test, the measured ground loss resistance was 6 Ω.

The same antenna therefore sees three different ground loss resistances as a direct result of the antenna mounting location and size of the vehicle. It is important to note that the measured ground loss increases as the antenna base nears the ground. The importance of minimizing ground losses in mobile antenna installations cannot be overemphasized.

Efficiency Curves

With the equations defined previously, a computer

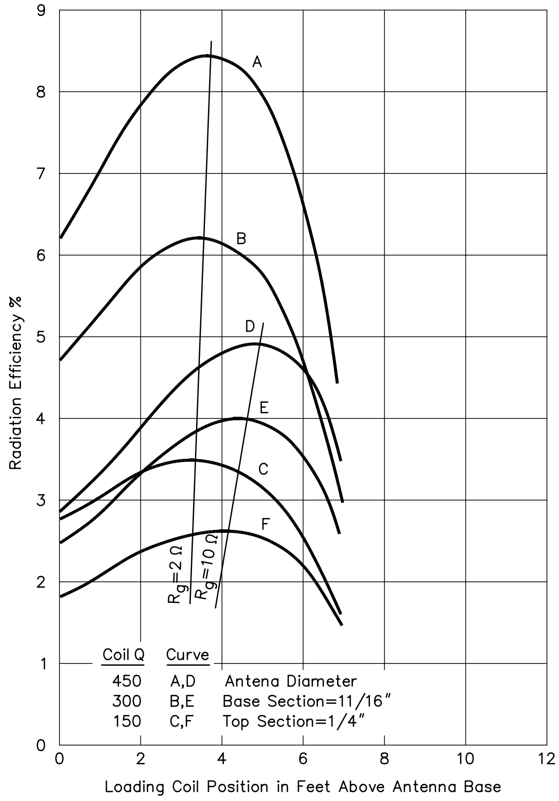

Fig 13—Radiation efficiency of 8-foot antennas at 3.9 MHz.

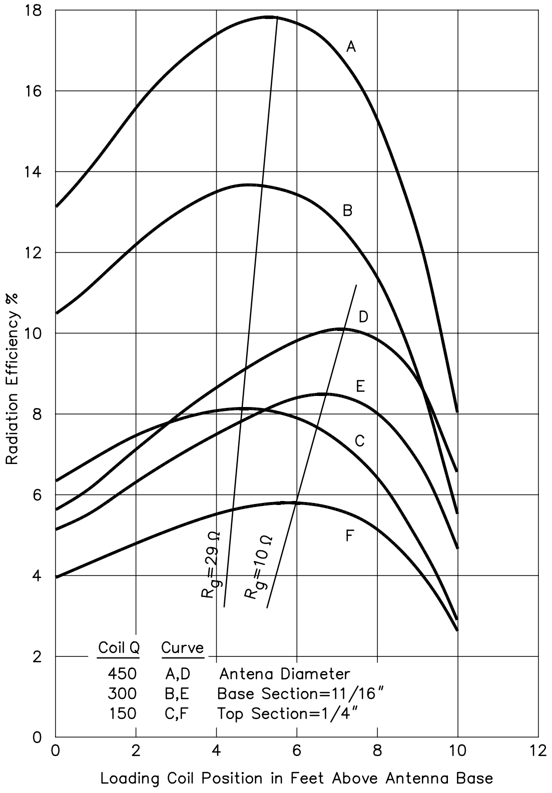

Fig 14—Radiation efficiency of 11-foot antennas at 3.9 MHz.

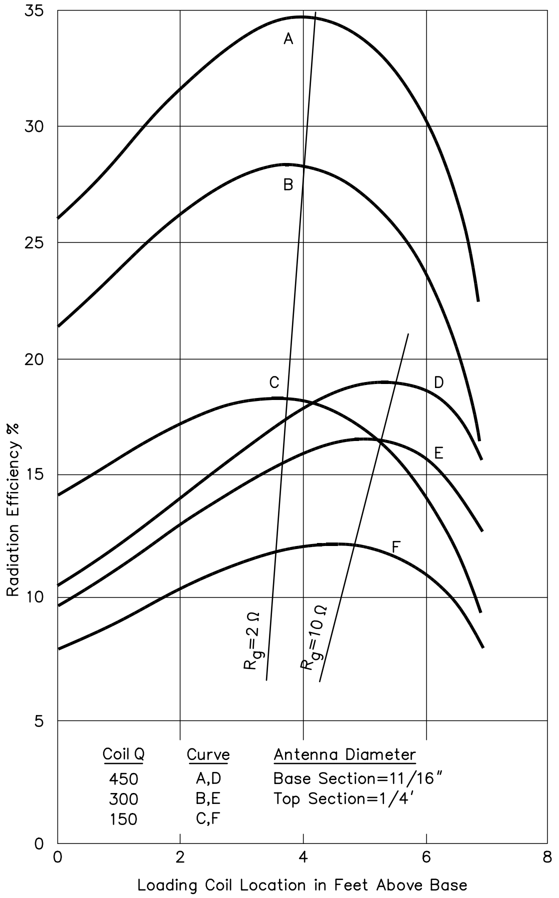

Fig 15—Radiation efficiency of 8-foot antennas at 7.225 MHz.

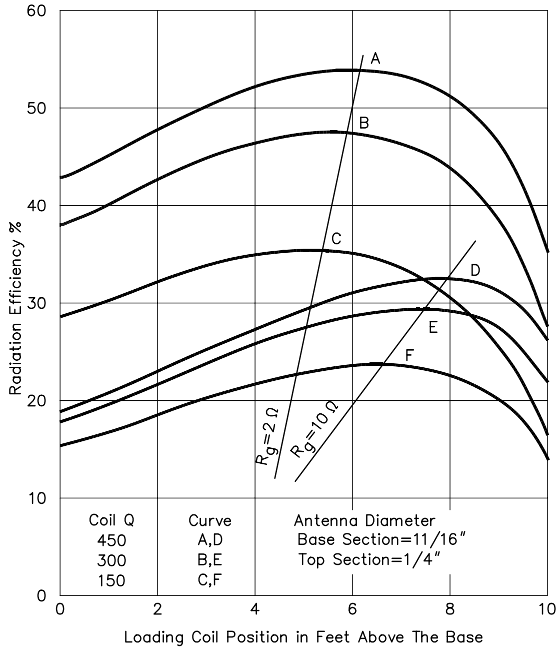

Fig 16—Radiation efficiency of 11-foot antennas at 7.225 MHz.

was used to calculate the radiation efficiency curves depicted in **Figs 13** through **16**. These curves were calculated for 3.8- and 7-MHz antennas of 8- and 11-foot lengths. Several values of loading coil Q were used, for both 2 and 10 Ω of ground loss resistance. For the calculations, the base section is $^1/_2$-inch diameter electrical EMT, which has an outside diameter of $^{11}/_{16}$ inch. The top section is fiberglass bicycle-whip material covered with Belden braid. These are readily available materials, which can be used by the average amateur to construct an inexpensive but rugged antenna.

Upon inspection, these radiation-efficiency curves reveal some significant information:

1) Higher coil Q produces higher radiation efficiencies,
2) longer antennas produce higher radiation efficiencies,
3) higher frequencies produce higher radiation efficiencies,
4) lower ground loss resistances produce higher radiation efficiencies,
5) higher ground loss resistances force the loading coil above the antenna center to reach a crest in the radiation-efficiency curve, and
6) higher coil Q sharpens the radiation-efficiency curves, resulting in the coil position being more critical for optimum radiation efficiency.

Note that the radiation efficiency curves reach a peak and then begin to decline as the loading coil is raised farther up the antenna. This is because of the rapid increase in loading coil reactance required above the antenna center. Refer to Fig 12. The rapid increase in coil size required for resonance results in the coil loss resistance increasing much more rapidly than the radiation resistance. This results in decreased radiation efficiency, as shown in Fig 11.

A slight reverse curvature exists in the curves between the base-loaded position and the one-foot coil-height position. This is caused by a shift in the curve resulting from insertion of a base section of larger diameter than the whip when the coil is above the base.

The curves in Figs 13 through 16 were calculated with constant (but not equal) diameter base and whip sections. Because of wind loading, it is not desirable to increase the diameter of the whip section. However, the base-section diameter can be increased within reason to further improve radiation efficiency. **Fig 17** was calculated for base-section diameters ranging from $^{11}/_{16}$ inch to 3 inches. The curves reveal that a small increase in radiation efficiency results from larger diameter base sections.

The curves in Figs 13 through 16 show that radiation efficiencies can be quite low in the 3.8-MHz band compared to the 7-MHz band. They are lower yet in the 1.8-MHz band. To gain some perspective on what these low efficiencies mean in terms of signal strength, **Fig 18** was calculated using the following equation:

$$\mathrm{dB} = \log \frac{100}{E} \qquad \text{(Eq 20)}$$

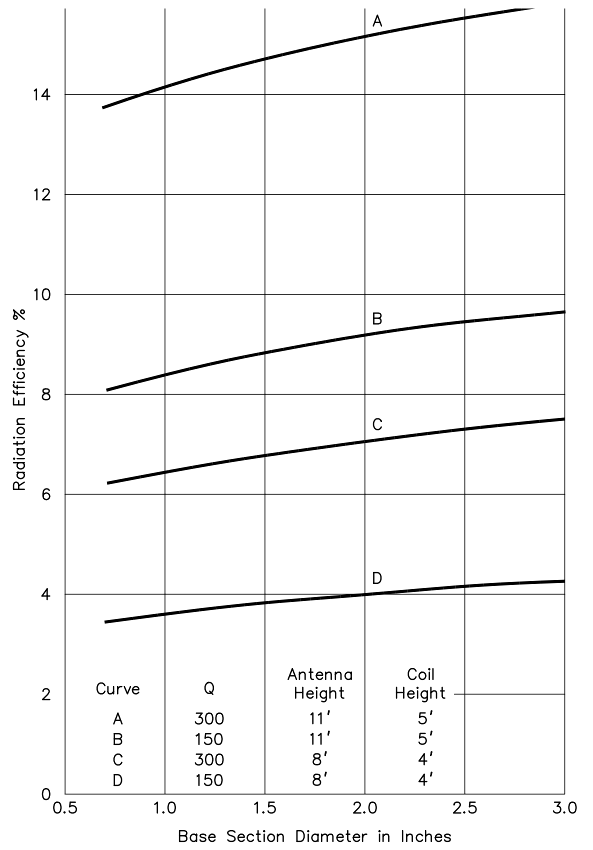

Fig 17—Radiation efficiency plotted as a function of base section diameter. Frequency = 3.9 MHz, ground loss resistance = 2 Ω, and whip section = $^1/_4$-inch diameter.

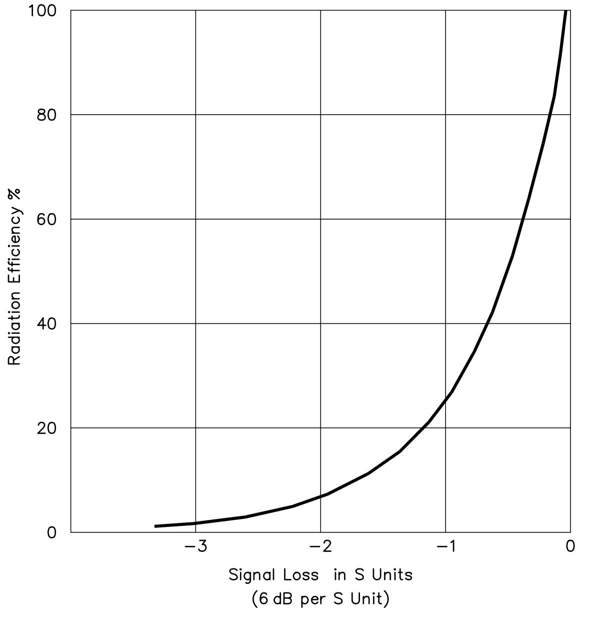

Fig 18—Mobile antenna signal loss as a function of radiation efficiency, compared to a quarter-wave vertical antenna over perfect ground.

where

dB = signal loss in decibels
E = efficiency in percent.

The curve in Fig 18 reveals that an antenna having 25% efficiency has a signal loss of 6 dB (approximately one S unit) below a quarter-wave vertical antenna over perfect ground. An antenna efficiency in the neighborhood of 6% will produce a signal strength on the order of two S units or about 12 dB below the same quarter-wave reference vertical. By careful optimization of mobile-antenna design, signal strengths from mobiles can be made fairly competitive with those from fixed stations using comparable power. And don’t forget: Moving your car near saltwater, with its high conductivity, can result in surprisingly strong signals from a mobile station!

Impedance Matching

The input impedance of short, high-Q coil-loaded antennas is quite low. For example, an 8-foot antenna optimized for 3.9 MHz with an unloaded coil Q of 300 and a ground-loss resistance of 2 Ω has a base input impedance of about 13 Ω. This low impedance value causes a standing wave ratio of 4:1 on a 50-Ω coaxial line at resonance. This high SWR is not compatible with the requirements of solid-state transmitters. Also, the bandwidth of shortened vertical antennas is very narrow. This severely limits the capability to maintain transmitter loading over even a small frequency range.

Impedance matching can be accomplished by means of L networks or impedance-matching transformers, but the narrow bandwidth limitation remains. A more elegant solution to the impedance matching and narrow bandwidth problem is to install an automatic tuner at the antenna base. Such a device matches the antenna and coaxial line automatically, and permits operation over a wide frequency range.

The tools are now available to tailor a mobile antenna design to produce maximum radiation efficiency. Mathematical modeling with a personal computer reveals that loading coil Q factor and ground loss resistance greatly influence the optimum loading coil position in a short vertical antenna. It also shows that longer antennas, higher coil Q, and higher operating frequencies produce higher radiation efficiencies.

End effect has not been included in any of the equations to assure that the loading coil will be slightly larger than necessary. Pruning the antenna to resonance should be done by removing coil turns, rather than by shorting turns or shortening the whip section excessively. Shortening the whip reduces radiation efficiency, by both shortening the antenna and moving the optimum coil position. Shorting turns in the loading coil degrades the Q of the coil.

Shortened Dipoles

Mathematical modeling techniques can be applied to shortened dipoles by using zero ground loss resistance and by doubling the computed values of radiation resistance and feed-point impedance. Radiation efficiency, however, does not double. Rather, it remains unchanged, because a second loading coil is required in the other leg of the dipole. The addition of the second coil offsets the gain in efficiency that occurs when the feed-point impedance and radiation resistance are doubled. There is a gain in radiation efficiency over a vertical antenna worked against ground, though, because the dipole configuration allows ground loss resistance to be eliminated from the calculations.

CONTINUOUSLY LOADED ANTENNAS

The design of high-Q air core inductors for RF work is complicated by the number of parameters that must be optimized simultaneously. One of these factors affecting coil Q adversely is radiation from a discrete loading coil. Therefore, the possibility of cutting down other losses while incorporating the coil radiation into that from the rest of the antenna system is an attractive one.

The general approach has been to use a coil made from heavy wire (#14 or larger), with length-to-diameter ratios as high as 21. British experimenters have reported good results with 8-foot overall lengths on the 1.8- and 3.5-MHz bands. The idea of making the entire antenna out of one section of coil has also been tried with some success. This technique is referred to as linear loading. Further information on linear-loaded antennas can be found in Chapter 6.

While going to extremes trying to find a perfect loading arrangement may not improve antenna performance very much, a poor system with lossy coils and high-resistance connections must be avoided if a reasonable signal is to be radiated.

MATCHING TO THE TRANSMITTER

Most modern transmitters require a 50-Ω output load, and because the feed-point impedance of a mobile whip is quite low, a matching network is usually necessary. Although calculations are helpful in the initial design, considerable experimenting is often necessary in final tune-up. This is particularly true for the lower bands, where the antenna is electrically short compared with a quarter-wave whip. The reason is that the loading coil is required to tune out a very large capacitive reactance, and even small changes in component values result in large reactance variations. Since the feed-point resistance is low to begin with, the problem is even more aggravated.

You can transform the low resistance of the whip to a value suitable for a 50-Ω system with an RF transformer or with a shunt-feed arrangement, such as an L network. The latter may only require a shunt coil or shunt capacitor at the base of the whip, since the net series capacitive or inductive reactance of the antenna and its loading coil may be used as part of the network. The following example illustrates the calculations involved.

Assume that a center-loaded whip antenna, 8.5 feet in overall length, is to be used on 7.2 MHz. From Table 1, earlier in this chapter, we see that the feed-point resistance of the antenna will be approximately 19 Ω, and from Fig 5 that the capacitance of the whip, as seen at its base, is approximately 24 pF. Since the antenna is to be center loaded, the capacitance value of the section above the coil will be cut approximately in half, to 12 pF. From this, it may be calculated that a center-loading inductor of 40.7 µH is required to resonate the antenna, that is, to cancel out the capacitive reactance. (This figure agrees with the approximate value of 40 µH shown in Table 1. The resulting feed-point impedance would then be 19 + *j* 0 Ω—a good match, *if* one happens to have a supply of 19-Ω coax.)

Solution: The antenna can be matched to a 52-Ω line such as RG-8 by tuning it either above or below reso-

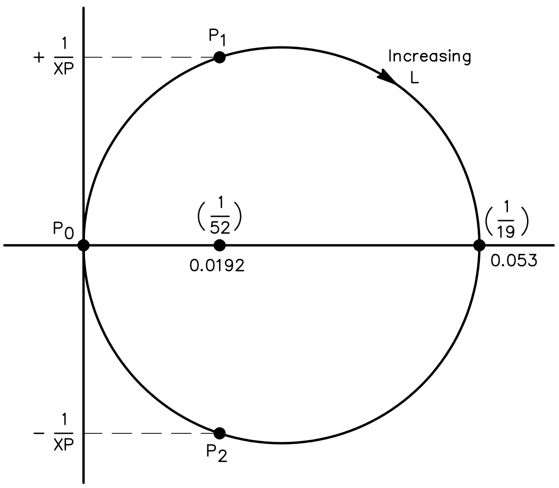

Fig 19—Admittance diagram of the RLC circuit consisting of the whip capacitance, radiation resistance and loading coil discussed in text. The horizontal axis represents conductance, and the vertical axis susceptance. The point P_0 is the input admittance with no whip loading inductance. Points P1 and P2 are described in the text. The conductance equals the reciprocal of the resistance, if no reactive components are present. For a series RX circuit, the conductance is given by

$$G = \frac{R}{R^2 + X^2}$$

and the susceptance is given by

$$B = \frac{-X}{R^2 + X^2}$$

Consequently, a parallel equivalent GB circuit of the series RX one can be found which makes computations easier. This is because conductances and susceptances add in parallel the same way resistances and reactances add in series.

nance and then canceling out the undesired component with an appropriate shunt element, capacitive or inductive. The way in which the impedance is transformed up can be seen by plotting the admittance of the series RLC circuit made up of the loading coil, antenna capacitance, and feed-point resistance. Such a plot is shown in **Fig 19** for a constant feed-point resistance of 19 Ω. There are two points of interest, P1 and P2, where the input conductance is 19.2 millisiemens, corresponding to 52 Ω. The undesired susceptance is shown as $1/X_P$ and $-1/X_P$, which must be canceled with a shunt element of the opposite sign, but with the same magnitude. The value of the canceling shunt reactance, X_P, may be found from the formula:

$$X_p = \frac{R_f\ Z_0}{\sqrt{R_f\left(Z_0 - R_f\right)}} \qquad \text{(Eq 21)}$$

where X_p is the reactance in Ω, R_f is the feed-point resistance, and Z_0 is the feed-line impedance. For Z_0 = 52 Ω and R_f = 19 Ω, X_p = ±39.5 Ω. A coil or good quality mica capacitor may be used as the shunt element. With the tune-up procedure described later, the value is not critical, and a fixed-value component may be used.

To arrive at point P1, the value of the center loading-coil inductance would be less than that required for resonance. The feed-point impedance would then appear capacitive, and an inductive shunt matching element would then be required. To arrive at point P2, the center loading coil should be more inductive than required for resonance, and the shunt element would need to be capacitive.

The value of the center loading coil required for the shunt-matched and resonated condition may be determined from the equation:

$$L = \frac{10^6}{4\pi^2 f^2 C} \pm \frac{X_S}{2\pi f} \qquad \text{(Eq 22)}$$

where addition is performed if a capacitive shunt is to be used, or subtraction performed if the shunt is inductive, and where L is in µH, f is the frequency in MHz, C is the capacitance of the antenna section being matched in pF, and

$$X_S = \sqrt{R_f\left(Z_0 - R_f\right)} \qquad \text{(Eq 23)}$$

For the example given, where Z_0 = 52 Ω, R_f = 19 Ω, f = 7.2 MHz, and C = 12 pF, X_S is found to be 25.0 Ω. The required antenna loading inductance is either 40.2 µH or 41.3 µH, depending on the type of shunt. Various matching possibilities for this example are shown in **Fig 20**. At A, the antenna is shown as tuned to resonance with L_L, a 40.7 µH coil, but with no provisions included for matching the resulting 19-Ω impedance to the 52-Ω line. At B, L_L has been reduced to 40.2 µH to make the antenna appear net capacitive, and L_M, having a reactance of 39.5 Ω, is added in shunt to cancel the capacitive reactance and transform the feed-point impedance to 50 Ω. The arrangement at C is similar to that at B except that L_L has been increased to 41.3 µH, and C_M (a shunt capacitor having a negative reactance of 39.5 Ω) is added, which also results in a 52-Ω nonreactive termination for the feed line.

The values determined for the loading coil in the above example point out an important consideration concerning the matching of short antennas—relatively small changes in values of the loading components will have a greatly magnified effect on the matching requirements. A change of less than 3% in the loading coil inductance value necessitates a completely different matching network! Likewise, calculations show that a 3% change in antenna capacitance will give similar results, and the value of the precautions mentioned earlier becomes clear. The sensitivity of the circuit with regard to frequency variations is also quite critical, and an excursion around practically the entire circle in Fig 19 may represent only 600 kHz, centered around 7.2 MHz, for the above example. This is why tuning up a mobile antenna can be very frustrating unless a systematic procedure is followed.

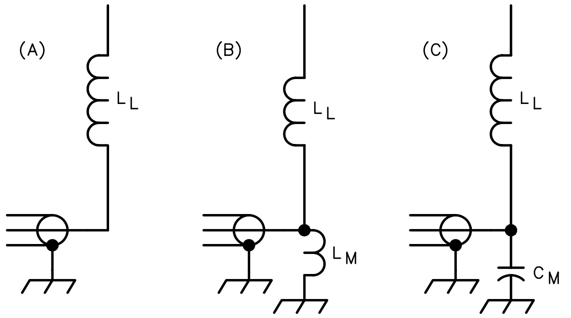

Fig 20—At A, a whip antenna that is resonated with a center loading coil. At B and C, the value of the loading coil has been altered slightly to make the feed-point impedance appear reactive, and a matching component is added in shunt to cancel the reactance. This provides an impedance transformation to match the Z_0 of the feed line. An equally acceptable procedure, rather than altering the loading coil inductance, is to adjust the length of the top section above the loading coil for the best match, as described in the tune-up section of the text.

Tune-Up

Assume that inductive shunt matching is to be used with the antenna in the previous example, Fig 20B, where 39.5 Ω is needed for L_M. This means that at 7.2 MHz, a coil of 0.87 µH will be needed across the whip feed-point terminal to ground. With a 40-µH loading coil in place, the adjustable whip section above the loading coil should be set for minimum height. Signals in the receiver will sound weak and the whip should be lengthened a bit at a time until signals start to peak. Turn the transmitter on and check the SWR at a few frequencies to find where a minimum occurs. If it is below the desired frequency, shorten the whip slightly and check again. It should be moved approximately $^1/_4$ inch at a time until the SWR is minimum at the center of the desired range. If the frequency where the minimum SWR occurs is above the desired frequency, repeat the procedure above, but lengthen the whip only slightly.

If a shunt capacitance is to be used, as in Fig 20C, a value of 560 pF would correspond to the required –39.5 Ω of reactance at 7.2 MHz. With a capacitive shunt, start with the whip in its longest position and shorten it until signals peak up.

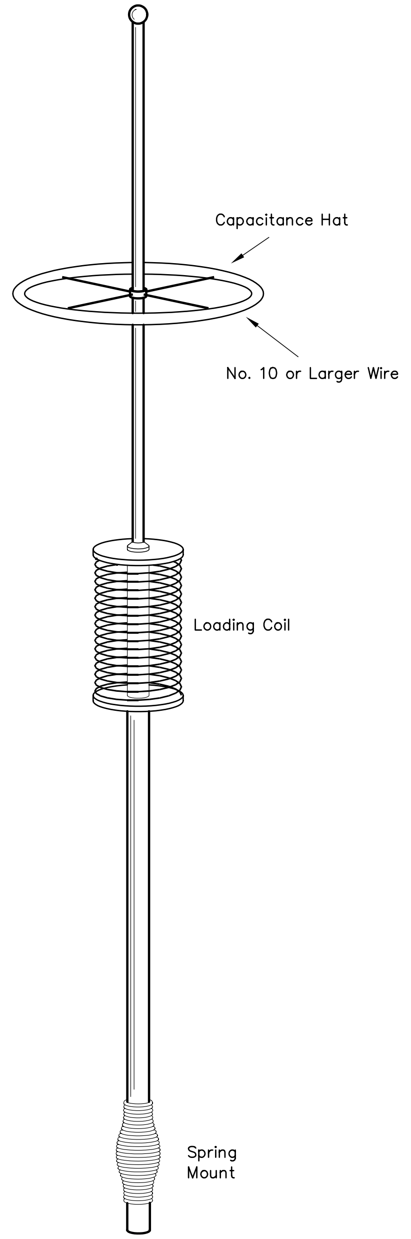
Capacitance Hat
No. 10 or Larger Wire
Loading Coil
Spring
Mount

Fig 21—A capacitance hat can be used to improve the performance of base- or center-loaded whips. A solid metal disc can be used in place of the skeleton disc shown here.

TOP-LOADING CAPACITANCE

Because the loss resistance varies with the inductance of the loading coil, the resistance can be reduced by removing turns from the coil. This must be compensated by adding capacitance to the portion of the mobile antenna that is *above* the loading coil (**Fig 21**). *Capacitance hats*, as they are called, can consist of a single stiff wire, two wires or more, or a disc made up of several wires like the spokes of a wheel. A solid metal disc could also be used, but is less practical for mobile work. The larger the capacitance hat (physically), the greater is the capacitance. The greater the capacitance, the less is the inductance required for resonance at a given frequency.

Capacitance-hat loading is applicable in either base-loaded or center-loaded systems. Since more inductance is required for center-loaded whips to make them resonant at a given frequency, capacitance hats are particularly useful in improving their efficiency.

TAPPED-COIL MATCHING NETWORK

Some of the drawbacks of the L-network can be eliminated by the use of the tapped-coil arrangement shown in **Fig 22**. Tune-up still remains critical, however, although somewhat more straightforward than for an L-network.

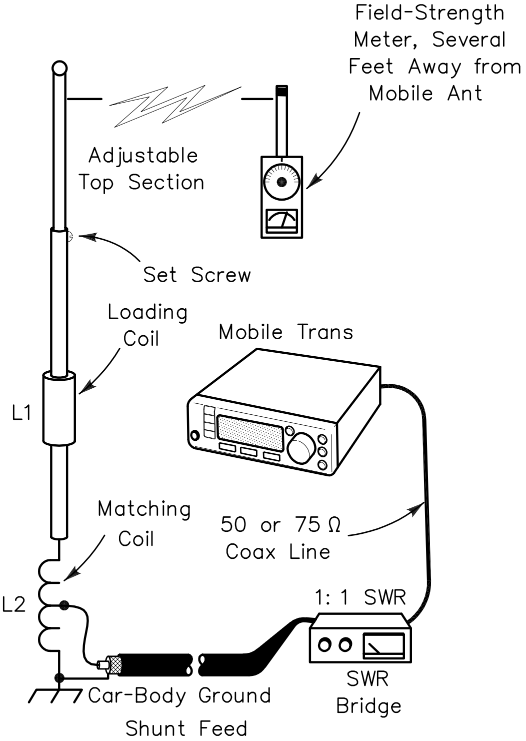
Field-Strength
Meter, Several
Feet Away from
Mobile Ant
Adjustable
Top Section
Set Screw
Loading
Coil
Mobile Trans
L1
Matching
Coil
50 or 75 Ω
Coax Line
1: 1 SWR
L2
SWR
Bridge
Car-Body Ground
Shunt Feed

Fig 22—A mobile antenna using shunt-feed matching. Overall antenna resonance is determined by the combination of L1 and L2. Antenna resonance is set by pruning the turns of L1, or adjusting the top section of the whip, while observing the field-strength meter or SWR indicator. Then adjust the tap on L2 for the lowest SWR.

Coil L2 can be inside the car body, at the base of the antenna, or at the base of the whip. As L2 helps determine the resonance of the antenna, L1 should be tuned to resonance in the desired part of the band with L2 in the circuit. The top section of the whip can be telescoped until a field-strength maximum is found. The tap on L2 is then adjusted for the lowest reflected power. Repeat these two adjustments until no further increase in field strength can be obtained; this point should coincide with the lowest SWR. The number of turns needed for L2 will have to be determined experimentally.

THE "SCREWDRIVER" MOBILE ANTENNA

Imagine QSYing from the bottom of 80 meters to the top of 10 meters, right from the driver's seat. With this antenna you will enjoy a very high Q antenna that has no taps or external adjustments, and that exhibits an SWR under 1.5:1 on all bands. Max Bloodworth, KO4TV, described this antenna in *The ARRL Antenna Compendium, Vol 7*.

The Screwdriver type of antenna was the brainchild of Don Johnson, W6AAQ, who developed it after many years of experimenting with mobile antennas. **Fig 23** is a photo of the completed antenna mounted on KO4TV's truck.

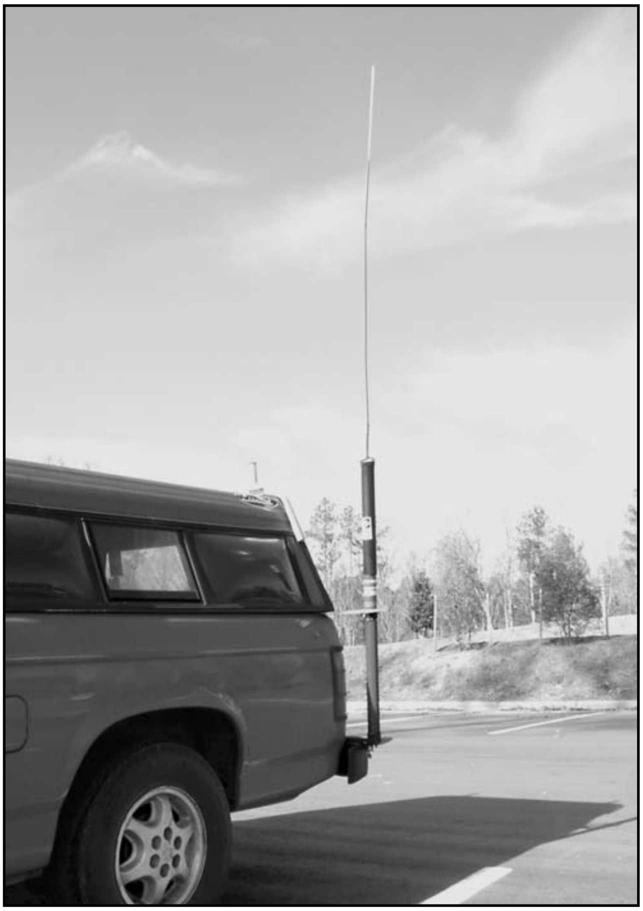

Fig 23—The completed Screwdriver antenna mounted on KO4TV's truck rear bumper. (*Photo courtesy of Gary Pearce, KN4AQ.*)

THE LOADING COIL

The general concept for the Screwdriver is that of a center-loaded antenna with an adjustable loading coil. This is an old idea; however, most previous multi-band antennas required multiple coil taps and an adjustable top whip section, as well as an impedance-matching unit at the base.

The problem of tapping a coil is well known. If you leave one end of the coil open, you have a miniature Tesla Coil, which can cause corona discharge and arcing. If you short the turns, the Q of the loading coil usually drops drastically. The Screwdriver is remarkable in that it does not have any coil taps and yet it can cover a wide range of frequencies with very high Q.

The secret is in the manner of adjusting the coil. It simply slides up or down into a metal tube (the base section), and makes contact with the top of the tube by means of "finger stock," which is made of spring-like Beryllium Copper. The coil is pushed up or down by an ordinary cordless screwdriver; hence the name, Screwdriver. This turns a section of threaded rod and moves the coil form up or down. The section of coil above the finger stock is the active loading coil, and the part of the coil just below the finger stock simply "disappears" into the base tube and is totally out of the circuit. No taps, no loose ends—just a high-Q coil in the middle of two solid metal tubes. Although this antenna is available ready-made from several commercial sources, making one is well within the capabilities of any ham who is relatively handy with ordinary hand tools, and who has access to either a small lathe or a common $1^1/_4$-inch pipe die.

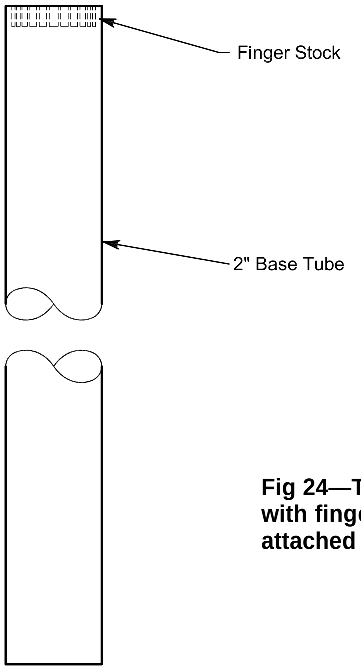

Fig 24—The base tube, with finger stock attached to the top.

CONSTRUCTING YOUR SCREWDRIVER ANTENNA

Construction begins by locating a 2-inch inside-diameter tube, about 3 to 3$^1/_2$ feet long. See **Fig 24**. The author had used such diverse materials as:

- Aluminum irrigation pipe
- Schedule-20 copper tubing
- A stainless-steel hydraulic cylinder
- A section of aluminum from a 6-meter beam that was damaged by a tornado
- And even a brass bedpost salvaged from the local garbage dump.

The brass bedpost looked especially good on his vehicle, which was Burgundy with gold trim. Whatever material you use for the bottom tube, you can either leave it unfinished or painted to your taste.

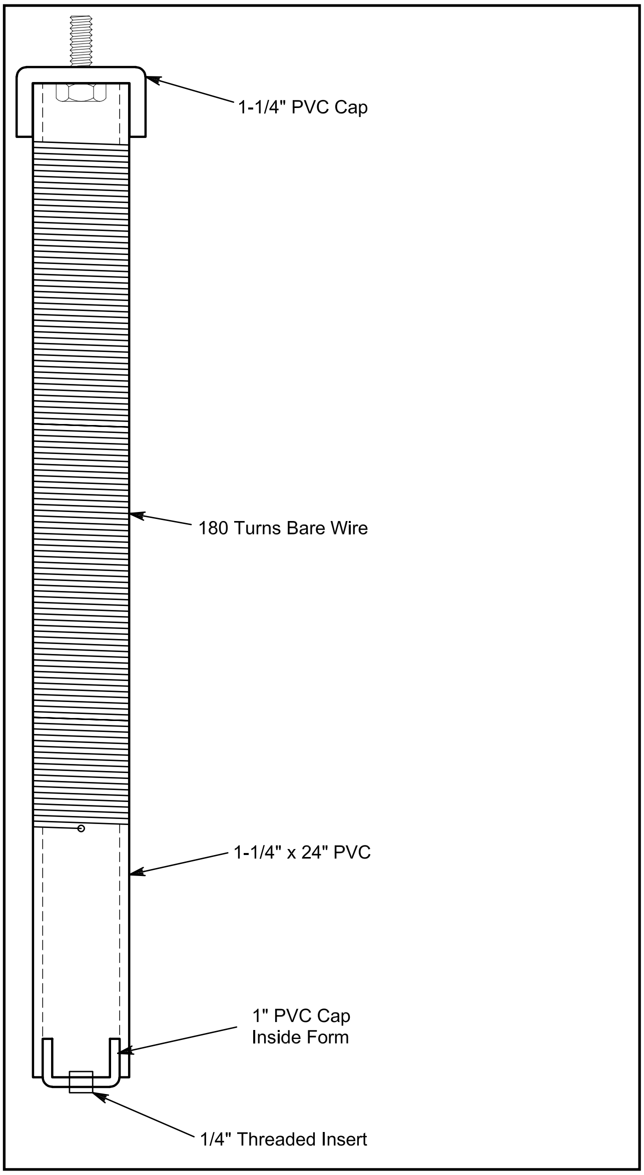

Fig 25—The coil form, with a threaded insert inserted into the bottom PVC cap inside the coil form.

The other essential item is a cordless screwdriver, minus its batteries and switch. You can often find one at yard sales or flea markets for a dollar or so, usually with dead batteries, since it is often just as cheap to buy another one new as it is to buy new batteries. So long as it will fit snugly inside the base tube, the brand name is immaterial. KO4TV has used Skil, Black & Decker, or even a Wal-Mart $8 special, all with equal success. The most difficult part of the job will be making the coil form and winding the loading coil, but with a little patience you should be able to do this satisfactorily.

Bear in mind that it is *not* a "Heathkit" type project, and it will require some innovation and ingenuity on your part. Next, obtain a 2-foot long piece of 1$^1/_4$-inch PVC pipe (*not* CPVC, which is vulnerable to ultraviolet rays from the sun). With either a lathe or a pipe die, thread approximately 20 inches from one end, at 10 to 12 threads per inch (See **Fig 25**). If you use a pipe die, it helps to loosen the cutters slightly, to make the grooves shallower.

When using a lathe, temporarily insert a piece of 1-inch water pipe inside the coil form to hold it steady in the lathe. The grooves should be just deep enough to comfortably hold the wire in place when it is wound. The choice of coil wire is up to you. The author has used #16 or #18 tinned copper, bare copper, or best yet, 17-gauge aluminum electric fence wire. This is available from any

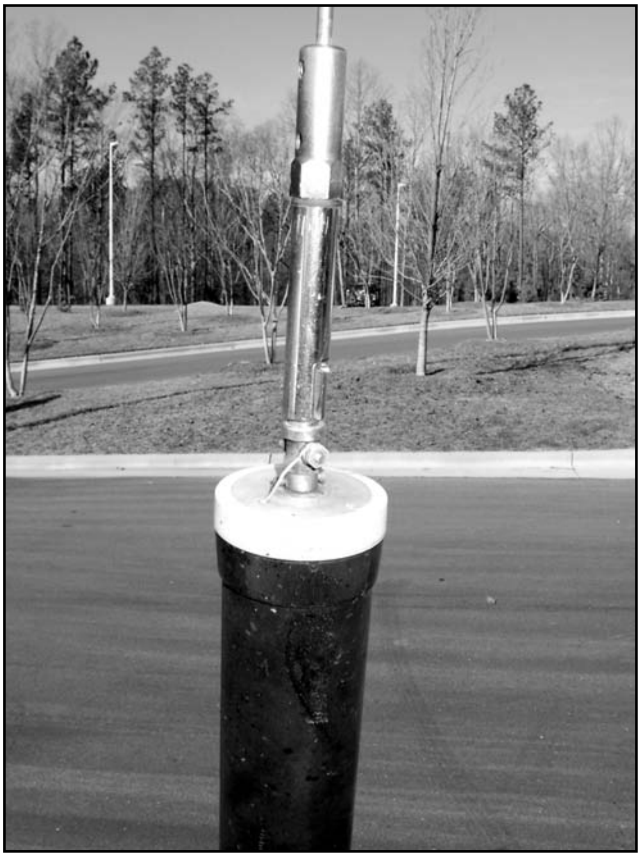

Fig 26—Photo of the top of the loading coil. (*Photo courtesy of Gary Pearce, KN4AQ.*)

farm supply store at a very reasonable cost. It will take about 65 to 70 feet of wire.

Begin by drilling a $^1/_{16}$-inch hole through the PVC pipe at the bottom end of the threaded portion. Slip the end of the wire through this hole and tie a knot in it to serve as a stop. You could also insert a nut and bolt through a loop in the wire. Then carefully wind the coil to within an inch of the end of the threaded portion. The preferred method of winding is to put a stick or pipe through the holes of the wire reel and hold it between your feet and the floor, to provide the necessary tension to allow tight winding.

About an inch below the top of the coil, drill a $^1/_{16}$-inch hole and thread about 6 or 8 inches of wire through it. This will be the top connection from the coil to the whip. The bottom of the coil has no electrical connection.

The method of attaching the whip to the top of the coil form is up to you. I usually drill a $^3/_8$-inch hole through a $1^1/_4$-inch PVC pipe cap and place a $^3/_8$ × $1^1/_2$-inch SAE bolt through the hole, with a matching washer and nut on top. The bolt can be drilled and tapped for mounting the whip, which should be about 5 to 6 feet long. (It will be pruned for resonance later.) See **Fig 26**, a photo showing the top of the coil.

A more elegant way of connecting the whip is to obtain a swivel or quick-disconnect antenna fitting from your nearest truck stop. These are widely available for use with CB antennas and are threaded $^3/_8$-inch SAE. Do not attach the cap to the coil form until after the coil is inserted in the base tube.

Now that you have a coil form and a base tube, the next step will be to plug the bottom of the PVC coil form with a 1-inch PVC pipe cap, which fits snugly inside the coil form. Drill a hole directly in the center of this cap, and install a $^1/_4$-20 threaded insert, available at most hardware stores for about 25 or 30 cents. These are commonly used in wooden furniture to provide a thread for attaching bolts to wood. See **Fig 27**.

It is imperative that you install this insert squarely. The best method is to screw a short bolt into it and chuck the bolt in a drill press, placing the pipe cap

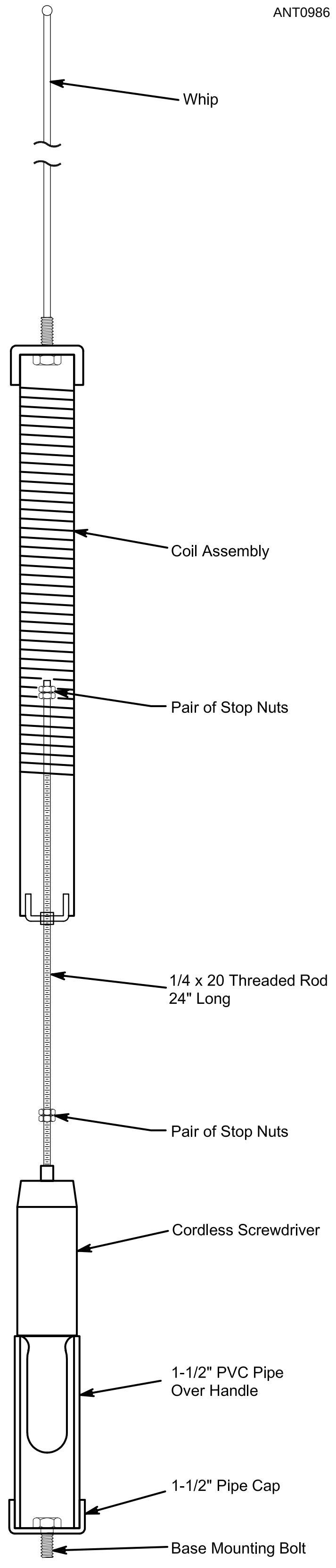
ANT0986
Whip
Coil Assembly
Pair of Stop Nuts
1/4 x 20 Threaded Rod
24" Long
Pair of Stop Nuts
Cordless Screwdriver
1-1/2" PVC Pipe
Over Handle
1-1/2" Pipe Cap
Base Mounting Bolt

Fig 27—The coil assembly and cordless screwdriver drive assembly.

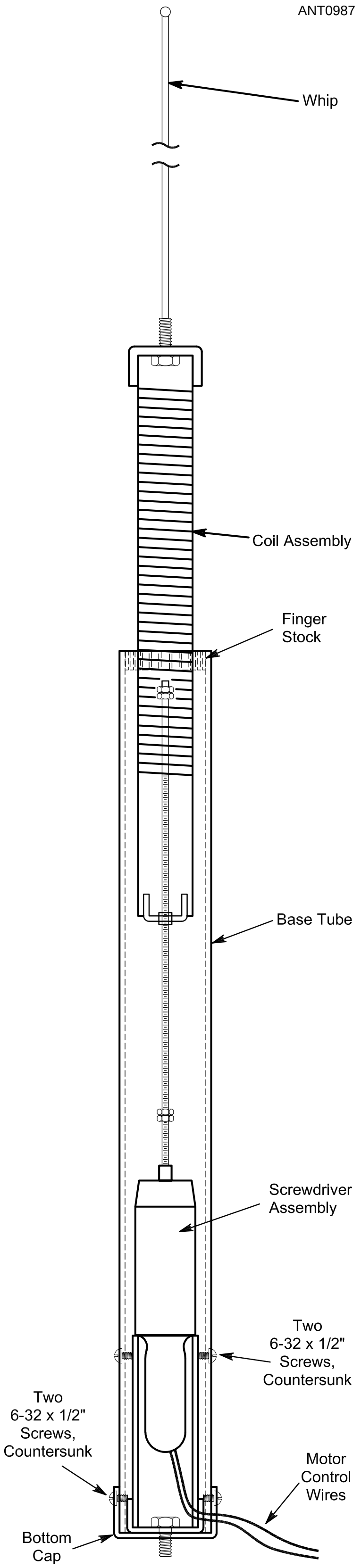
ANT0987
Whip
Coil Assembly
Finger
Stock
Base Tube
Screwdriver
Assembly
Two
6-32 x 1/2"
Screws,
Countersunk
Two
6-32 x 1/2"
Screws,
Countersunk
Motor
Control
Wires
Bottom
Cap

Fig 28—The complete antenna assembly.

squarely on the drill press table. Carefully turn the drill chuck by hand until the insert is properly seated.

The Motor

Now prepare the cordless drill by removing the dead batteries and the switch. Install a pair of 1 or 1.5-Ω, 10-W resistors in series with each motor lead. Connect a 0.1 μF 50-V disc ceramic capacitor across the motor leads to suppress motor noise.

Connect a pair of wires, about a foot long, to the other ends of the resistors. These will power the screwdriver from your car's 12-V system. Next, procure a 20-inch long piece of $^{1}/_{4}$-20 threaded rod (also called "all-thread"). Insert one end of this into the cordless screwdriver chuck, and drill a $^{1}/_{16}$-inch or $^{3}/_{32}$-inch hole through both the chuck and the rod. Secure the rod in place with either a roll or cotter pin, and install a pair of $^{1}/_{4}$-inch nuts on the rod near the screwdriver chuck. This will be the lower stop for the coil travel and will be adjusted later.

Thread the rod into the threaded insert at the bottom of the coil form, and install another pair of stop nuts on the top side of the rod. It will be easier to do this before attaching the bottom plug to the coil form. Again, these nuts will be adjusted later for proper coil travel.

Place the bottom plug into the coil form, drill a couple of holes through the form and the plug and tap them for 6-32 flat head bolts. Countersink the form so that the heads will be flush with the outside. Again, see **Fig 27**.

Finger Stock

Now, let's go back to the base tube. You must install a circle of finger-stock strip on the top of the tube to make contact with the coil, as shown in Fig 24. This can either be soldered or riveted, depending on your choice of tube material. Next, insert the coil and screwdriver assembly through the bottom of the tube, making sure that the coil clears the finger stock without deforming or bending it.

The best method I have found for attaching the screwdriver to the bottom tube is to fit a piece of $1^{1}/_{2}$-inch PVC pipe, about 8 or 10 inches long, over the handle of the screwdriver and attach it with a couple of flat-head 6-32 bolts. You may have to either grind some material off the handle or wrap it with some duct tape to make a snug fit inside the $1^{1}/_{2}$-inch PVC pipe, which should slip snugly into the base tube.

Place a matching $1^{1}/_{2}$-inch PVC pipe cap on the bottom of the pipe, and secure it with a couple of 6-32 screws. Drill a $^{1}/_{4}$-inch hole through one side of the cap, which will serve to pass the wires from the screwdriver motor. Drill a $^{3}/_{8}$-inch hole directly in the center of this cap for the base mount. See **Fig 28**, which is a drawing of the completed antenna assembly.

MOUNTING THE ANTENNA

At this point you must determine exactly how you wish to mount the antenna on your vehicle. You could fasten it to a standard mobile antenna mount, but *do not* use a spring! I prefer to fasten the assembly to a thick metal plate about 6 × 18 inches in size. I mount this to the frame of the car, protruding about 4 or 5 inches from the lower car body, just behind the rear wheel. You could also fasten such a plate to the lower fender sheet metal with large sheet-metal screws. The SO-239 connector is mounted to this plate also. See **Fig 29**.

An upper bracket, made from Plexiglas or similar insulating material, can be mounted from the trunk lip with an L-angle and used to support the top of the lower tube. See **Fig 30**. With a Plexiglas bracket about 4-inches wide, a 2-inch hole can be drilled in it to pass the tube. This will make a snug and rigid fit at the top. Under no circumstances use a metal band around the tube, as this will form a shorted turn and drastically lower the antenna Q.

With a little care, this type of mounting will leave no visible holes in the vehicle, which will likely enhance its trade-in value. If you mount the antenna on a pickup truck, it can be mounted directly on the rear safety bumper with a $1^{1}/_{2}$ inch pipe floor flange and matching adaptor on the bottom of the tube.

Electrical Connections

After mounting the base, your next job is to connect the wires from the screwdriver to a DPDT, spring-return center-off switch, mounted in a convenient location in your car. See the schematic in **Fig 31**. Connect the coax from the radio to the base tube, with the center conductor going to the base tube and the shield connected to ground at the base of the antenna.

If you like, you could install a simple base-matching network to give a perfect match on the lower bands. Even with no matching unit, SWR is usually under 2:1 on all

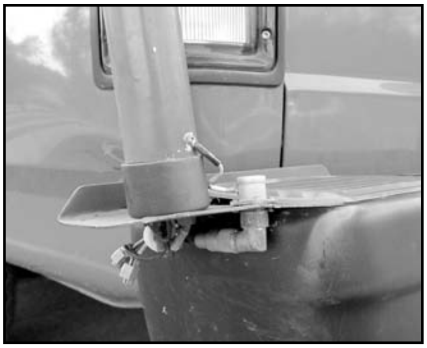

Fig 29—Base-mounting plate for Screw-driver installation on KO4TV's truck. Note the quick-disconnect dc motor con-nections and the SO-239 coax connector.

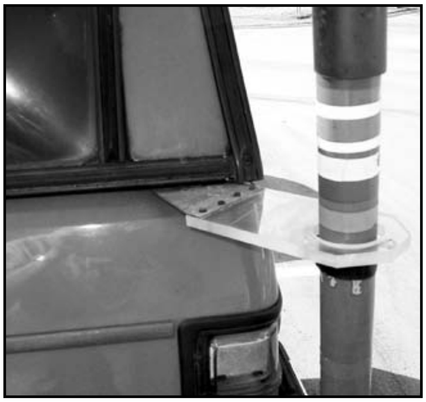

Fig 30—The support Plexiglas plate. (*Photo courtesy of Gary Pearce, KN4AQ.*)

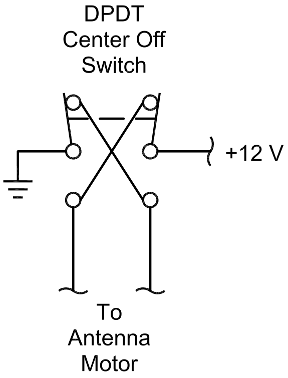

Fig 31—The DPDT center-off spring-loaded switch used to control the screwdriver motor.

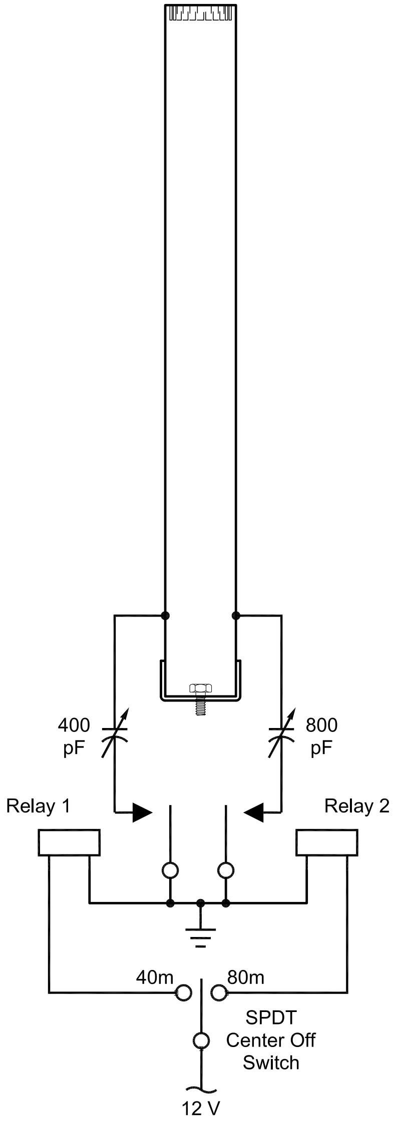

Fig 32—Relay-switched matching capacitors for 40 and 80 meters.

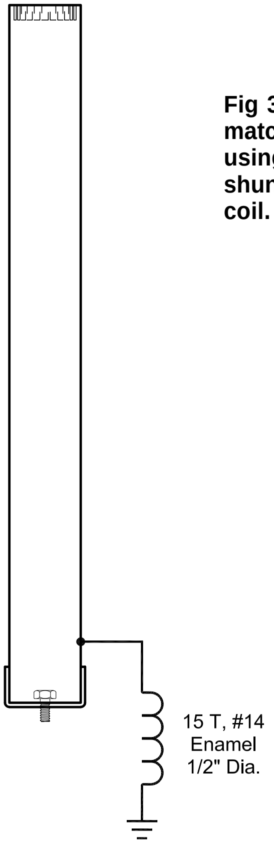

Fig 33—Alternative matching scheme using a single shunt matching coil.

bands. See **Fig 32**. One matching network consists of a couple of trimmer capacitors switched from the bottom of the base tube to ground using a pair of small relays. For 80 meters, approximately 800 pF is required, and for 40 meters about 400 pF. No matching is required for 30 meters or higher.

An alternative matching method is to install a 15-turn coil made from #14 enamel wire, about 1/2 inch diameter, from the base to ground. See **Fig 33**. In either case, adjust the whip length for proper resonance at both ends of the HF spectrum over the range of movement of your coil.

This antenna may have different matching requirements, depending on the vehicle type and mounting location, so you may need to do some experimentation to obtain optimum results. If it doesn't work exactly right at first, be prepared to do a little experimenting.

A COIL COVER

Most builders will want to provide a coil cover, both for protection and appearance. KO4TV's favorite cover is made from a plastic mailing tube, about 24 to 30 inches long and about 1/2 inches in diameter. You can find these at an office supply store. They have a screw-on cap on one end, which can be mounted to the top of the loading coil.

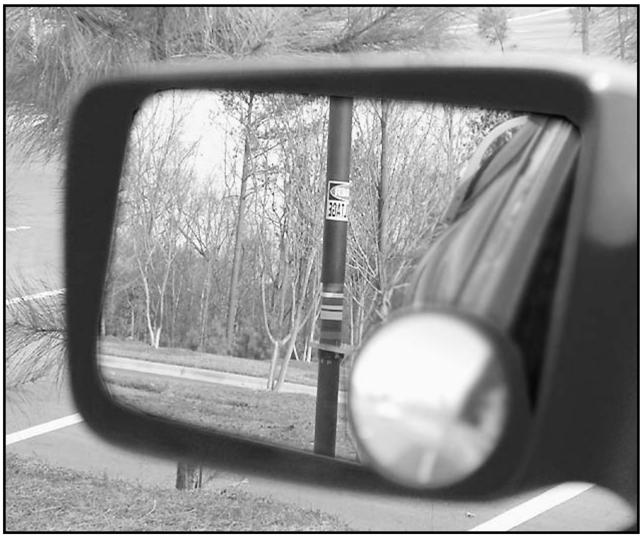

Fig 34—Remote control by the rear-view mirror—tuning by the stripes! (*Photo courtesy of Gary Pearce, KN4AQ.*)

You trim the bottom of the cover to the proper length to allow the coil to be fully retracted into the base tube while still clearing the top bracket. Other builders have used such diverse materials as:

- Several 1-liter soft drink bottles cut off, swaged and glued together
- Sections of Schedule 20 PVC pipe with matching cap
- Empty beef jerky tubes (usually available for the asking at convenience stores)
- Wands from a shop vacuum cleaner, fitted with the cap from an aerosol spray can.

Just let your imagination run wild and you may be surprised at your own ingenuity. Whatever material you do use, paint it to match your vehicle. Just be sure to use non-metallic paint.

TUNING BY THE STRIPES

The author used different colored vinyl tapes to mark the tuning points for the amateur bands along the tuning coil. The lower edge of the supporting Plexiglas middle support acts as a pointer, which he can spot from his rear-view mirror. See **Fig 34**. This may be low-tech tuning method, but it works!

THE PROOF IS IN THE PUDDING

Now, you are ready to enjoy some real mobile communications on all bands, without even stopping the vehicle to QSY. KO4TV's antenna has contacted over 100 DX countries, from as far away as Tasmania and Japan from the East Coast. He enjoys regular contacts with Israel and other stations in the Mid-East, running 100 W or less.

A MOBILE J ANTENNA FOR 144 MHz

The J antenna is a mechanically modified version of the Zepp (Zeppelin) antenna. It consists of a half-wave-length radiator fed by a quarter-wave matching stub. This antenna exhibits an omnidirectional pattern with little high-angle radiation, but does not require the ground plane that $^1/_4$-wave and $^5/_8$-wave antennas do to work properly. The material in this section was prepared by Domenic Mallozzi, N1DM, and Allan White, W1EYI.

Fig 35 shows two common configurations of the J antenna. Fig 35A shows the shorted-stub version that is usually fed with 200- to 600-Ω open-wire line. Some have attempted to feed this antenna directly with coax without a balun, and this usually leads to less than optimum results. Among the problems with such a configuration are a lack of reproducibility and heavy coupling with nearby objects. To eliminate these problems, many

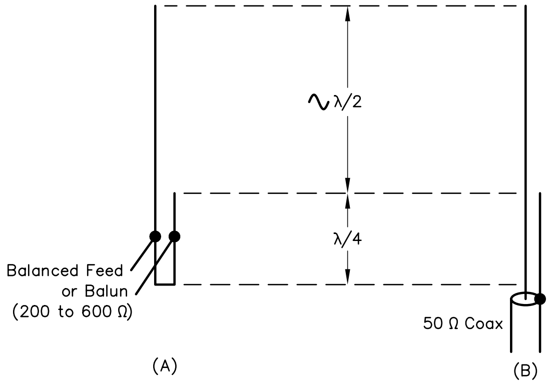

Fig 35—Two configurations of the J antenna.

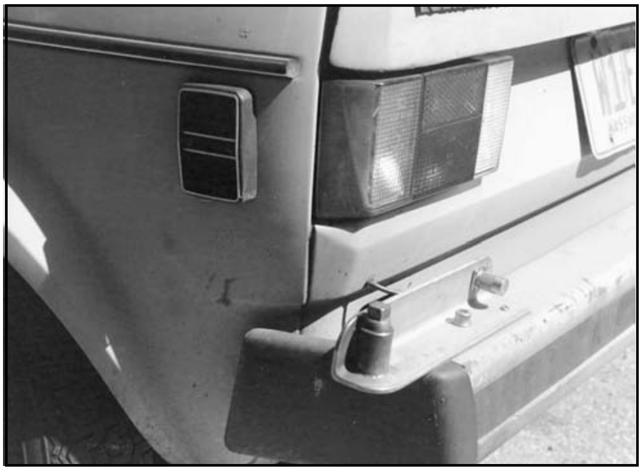

Fig 36—The mount for the mobile J is made from stainless-steel angle stock and secured to the bumper with stainless-steel hardware. Note the $^1/_2$-inch pipe plug and a PL-259 (with a copper disc soldered in its unthreaded end). These protect the mount and connector threads when the antenna is not in use.

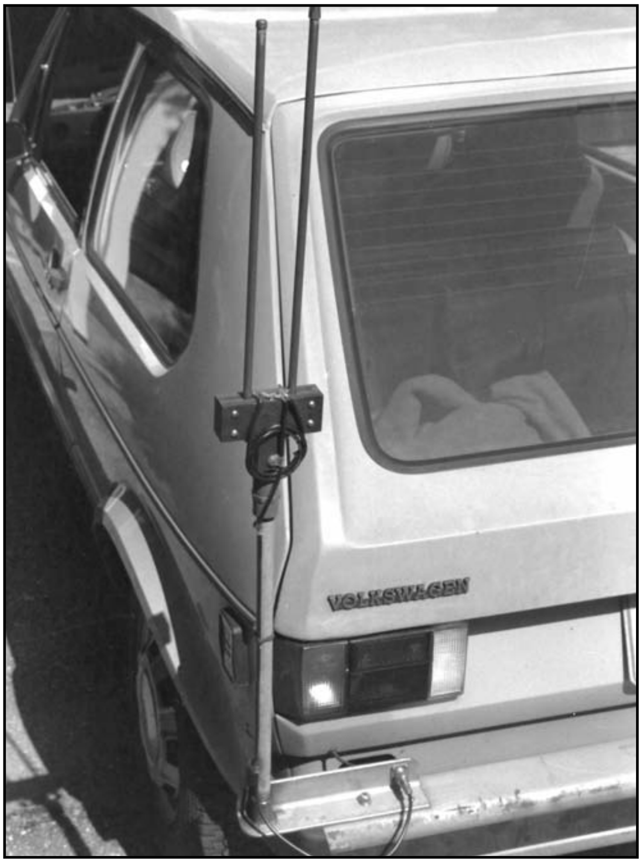

Fig 37—The J antenna, ready for use. Note the bakelite insulator and the method of feed. Tie wraps are used to attach the balun to the mounting block and to hold the coax to the support pipe. Clamps made of flashing copper are used to connect the balun to the J antenna just above the insulating block. The ends of the balun should be weatherproofed.

amateurs have used a 4:1 half-wave balun between the feed point and a coaxial feed line. This simple addition results in an antenna that can be easily reproduced and that does not interact so heavily with surrounding objects. The bottom of the stub may be grounded (for mechanical or other reasons) without impairing the performance of the antenna.

The open-stub-fed J antenna shown in Fig 35B can be connected directly to low-impedance coax lines with good results. The lack of a movable balun (which allows some impedance adjustment) may make this antenna a bit more difficult to adjust for minimum SWR, however.

The Length Factor

Dr. John S. Belrose, VE2CV, noted in *The Canadian Amateur* that the diameter of the radiating element is important to two characteristics of the antenna—its bandwidth and its physical length. (See Bibliography at the end of this chapter.) As the element diameter is increased, the usable bandwidth increases, while the physical length of the radiating element decreases with respect to the free-space half-wavelength. The increased diameter makes the end effect more pronounced, and also slows the velocity of propagation on the element. These two effects are related to resonant antenna lengths by a factor, "k." This factor is expressed as a decimal fraction giving the equivalent velocity of propagation on the antenna wire as a function of the ratio of the element diameter to a wavelength. The k factor is discussed at length in Chapter 2.

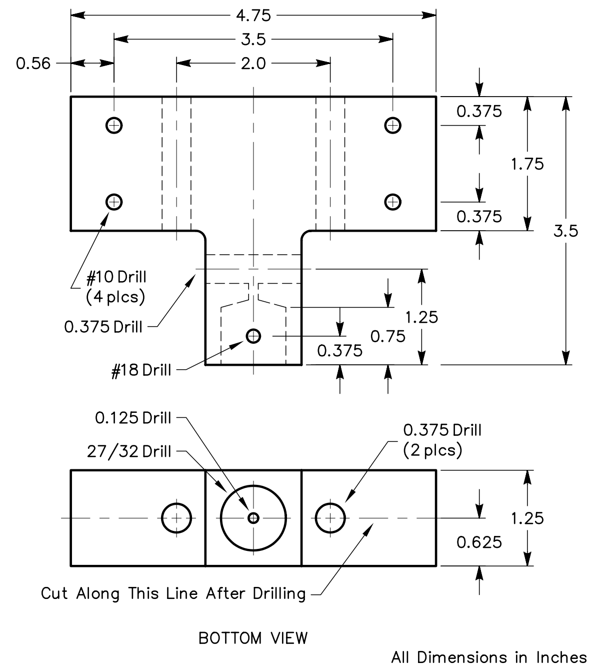

Fig 38—Details of the insulated mounting block. The material is bakelite.

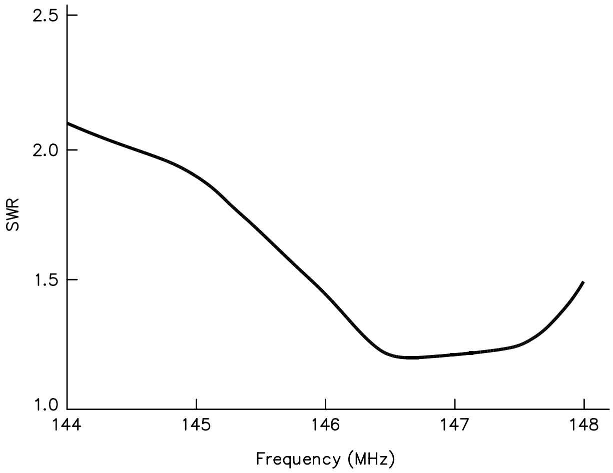

Fig 39—Measured SWR of the mobile J antenna.

The length of the radiating element is given by

$$\ell = \frac{5904 \times k}{f} \qquad \text{(Eq 25)}$$

where

ℓ = length in inches
f = frequency in MHz
k = k factor.

The k factor can have a significant effect. For example, if you use a $^5/_8$-inch diameter piece of tubing for the radiator at 144 MHz, the k value is 0.907 (9.3% shorter than a free-space half wavelength).

The J antenna gives excellent results for both mobile and portable work. The mobile described here is similar to an antenna described by W. B. Freely, K6HMS, in April 1977 *QST*. This design uses mechanical components that are easier to obtain. As necessary with all mobile antennas, significant attention has been paid to a strong, reliable, mechanical design. It has survived not only three New England winters, but also two summers of 370-mile weekend commutes. During this time, it has maintained consistent electrical performance with no noticeable deterioration.

The mechanical mount to the bumper is a 2 × 2-inch stainless steel angle iron, 10 inches long. It is secured to the bumper with stainless steel hardware, as shown in **Fig 36**. A stainless steel $^1/_2$-inch pipe coupling is welded to the left side of the bracket, and an SO-239 connector is mounted at the right side of the bracket. The bracket is mounted to the bumper so a vertical pipe inserted in the coupling will allow the hatchback of the vehicle to be opened with the antenna installed, **Fig 37**.

A $^1/_2$-inch galvanized iron pipe supports the antenna so the radiating portion of the J is above the vehicle roof line. This pipe goes into a bakelite insulator block, visible in Fig 37. The insulator block also holds the bottom of the stub. This block was first drilled and then split with a band saw, as shown in **Fig 38**. After splitting, the two portions are weatherproofed with varnish and rejoined with 10-32 stainless hardware. The corners of the insulator are cut to clear the L sections at the shorted end of the stub.

The quarter-wave matching section is made of $^1/_4$-inch type L copper tubing ($^5/_{16}$ inch ID, $^3/_8$ inch OD). The short at the bottom of the stub is made from two copper L-shaped sections and a short length of $^1/_4$-inch tubing. Drill a $^1/_8$-inch hole in the bottom of this piece of tubing to drain any water that may enter or condense in the stub.

A $^5/_{16}$-inch diameter brass rod, $1^1/_2$ to 2 inches long, is partially threaded with a $^5/_{16}$ × 24 thread to accept a Larsen whip connector. This rod is then sweated into one of the legs of the quarter-wave matching section. A 40-inch whip is then inserted into the Larsen connector.

The antenna is fed with 50-Ω coaxial line and a coaxial 4:1 half-wave balun. This balun is described at the end of Chapter 26. As with any VHF antenna, use high-quality coax for the balun. Seal all open cable ends and the rear of the SO-239 connector on the mount with RTV sealant.

Adjustment is not complicated. Set the whip so that its tip is 41 inches above the open end of the stub, and adjust the balun position for lowest SWR. Then adjust the height of the whip for the lowest SWR at the center frequency you desire. **Fig 39** shows the measured SWR of the antenna after adjustments are completed.

THE SUPER-J MARITIME ANTENNA

This 144-MHz vertical antenna doesn't have stringent grounding requirements and can be made from easy to find parts. The material in this section was prepared by Steve Cerwin, WA5FRF, who developed the Super-J for use on his boat.

Antennas for maritime use must overcome difficulties that other kinds of mobile antennas normally do not encounter. For instance, the transom of a boat is the logical place to mount an antenna. But the transoms of many boats are composed mostly of fiberglass, and they ride some distance out of the water—from several inches to a few feet, depending on the size of the vessel. Because the next best thing to a ground plane (the water surface) is more than an appreciable fraction of a wavelength away at 144 MHz, none of the popular gain-producing antenna designs requiring a counterpoise are suitable. Also, since a water surface does a good job of assuming the earth's lowest mean elevation (at least on a calm day), anything that can be done to get the radiating part of the antenna up in the air is helpful.

One answer is the venerable J-pole, with an extra in-phase half-wave section added on top—the Super-J antenna. The two vertical half waves fed in phase give outstanding omnidirectional performance for a portable antenna. Also, the J-pole feed arrangement provides the desired insensitivity to height above ground (or water) plus added overall antenna height. Best of all, a $^1/_4$-wave CB whip provides enough material to build the whole driven element of the antenna, with a few inches to spare. The antenna has enough bandwidth to cover the entire 144-MHz band, and affords a measure of lightning protection by being a grounded design.

Antenna Operation

The antenna is represented schematically in **Fig 40**. The classic J-pole antenna is the lower portion shown between points A and C. The half-wave section between points B and C does most of the radiating. The added half-wave section of the Super-J version is shown between points C and E. The side-by-side quarter-wave elements between points A and B comprise the J feed arrangement.

At first glance, counterproductive currents in the J section between points A and B may seem a waste of element material, but it is through this arrangement that

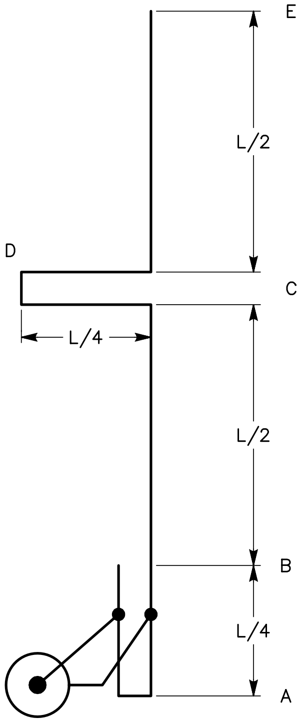

Fig 40—Schematic representation of the Super-J maritime antenna. The radiating section is two half waves in phase.

the antenna is able to perform well in the absence of a good ground. The two halves of the J feed arrangement, side by side, provide a loading mechanism regardless of whether or not a ground plane is present.

The radiation resistance of any antenna fluctuates as a function of height above ground, but the magnitude of this effect is small compared to the wildly changing impedance encountered when the distance from a ground plane element to its counterpoise is varied. Also, the J section adds $^{1}/_{4}$ wavelength of antenna height, reducing the effect of ground height variations even further. Reducing ground-height sensitivity is particularly useful in maritime operation on those days when the water is rough.

The gain afforded by doubling the aperture of a J-pole with the extra half-wave section can be realized only if the added section is excited in phase with the half-wave element B-C. This is accomplished in the Super-J in a conventional manner, through the use of the quarter-wave phasing stub shown between C and D.

Construction and Adjustment

The completed Super-J is shown in **Fig 41**. Details of the individual parts are given in **Fig 42**. The driven element can be liberated from a quarter-wave CB whip antenna and cut to the dimensions shown. All other metal stock can be obtained from metal supply houses or machine shops. Metal may even be scrounged for little or nothing as scraps or remnants, as were the parts for the antenna shown here.

The center insulator and the two J stub spacers are made of $^{1}/_{2}$-inch fiberglass and stainless steel stock, and the end caps are bonded to the insulator sections with epoxy. If you don't have access to a lathe to make the end caps, a simpler one-piece insulator design of wood or fiberglass could be used. However, keep in mind that good electrical connections must be maintained at all joints, and strength is a consideration for the center insulator.

The quarter-wave phasing stub is made of $^{1}/_{8}$-inch stainless steel tubing, **Fig 43**. The line comprising this stub is bent in a semicircular arc to narrow the vertical profile and to keep the weight distribution balanced. This makes for an attractive appearance and keeps the antenna from leaning to one side.

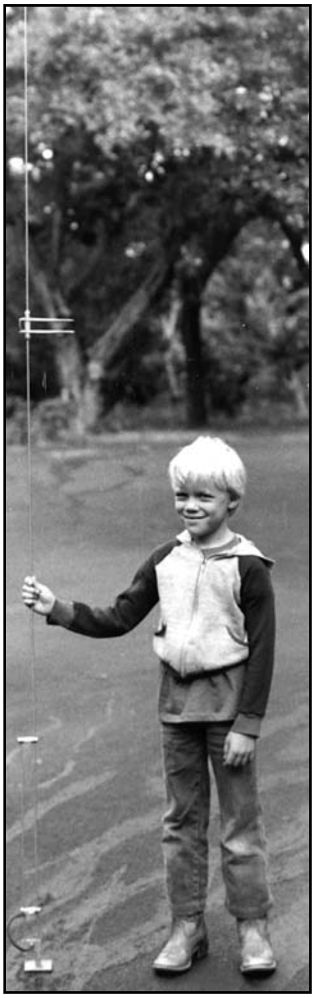

Fig 41—Andy and the assembled Super-J antenna.

The bottom shorting bar and base mounting plate are made of $^{1}/_{4}$-inch stainless steel plate, shown in **Fig 44**. The J stub is made of $^{3}/_{16}$-inch stainless-steel rod stock. The RF connector may be mounted on the shorting bar as shown, and connected to the adjustable slider with a short section of coaxial cable. RTV sealant should be used at the cable ends to keep out moisture. The all-stainless construction looks nice and weathers well in maritime mobile applications.

The antenna should work well over the whole 144-MHz band if cut to the dimensions shown. The only tuning required is adjustment of the sliding feed point for minimum SWR in the center of the band segment you use most. Setting the slider $2^{13}/_{16}$ inch above the top of the shorting bar gave the best match for this antenna and may be used for a starting point. Four turns of coax made into a coil at the feed point or a ferrite-sleeve balun act as a common-mode choke balun to ensure satisfactory performance.

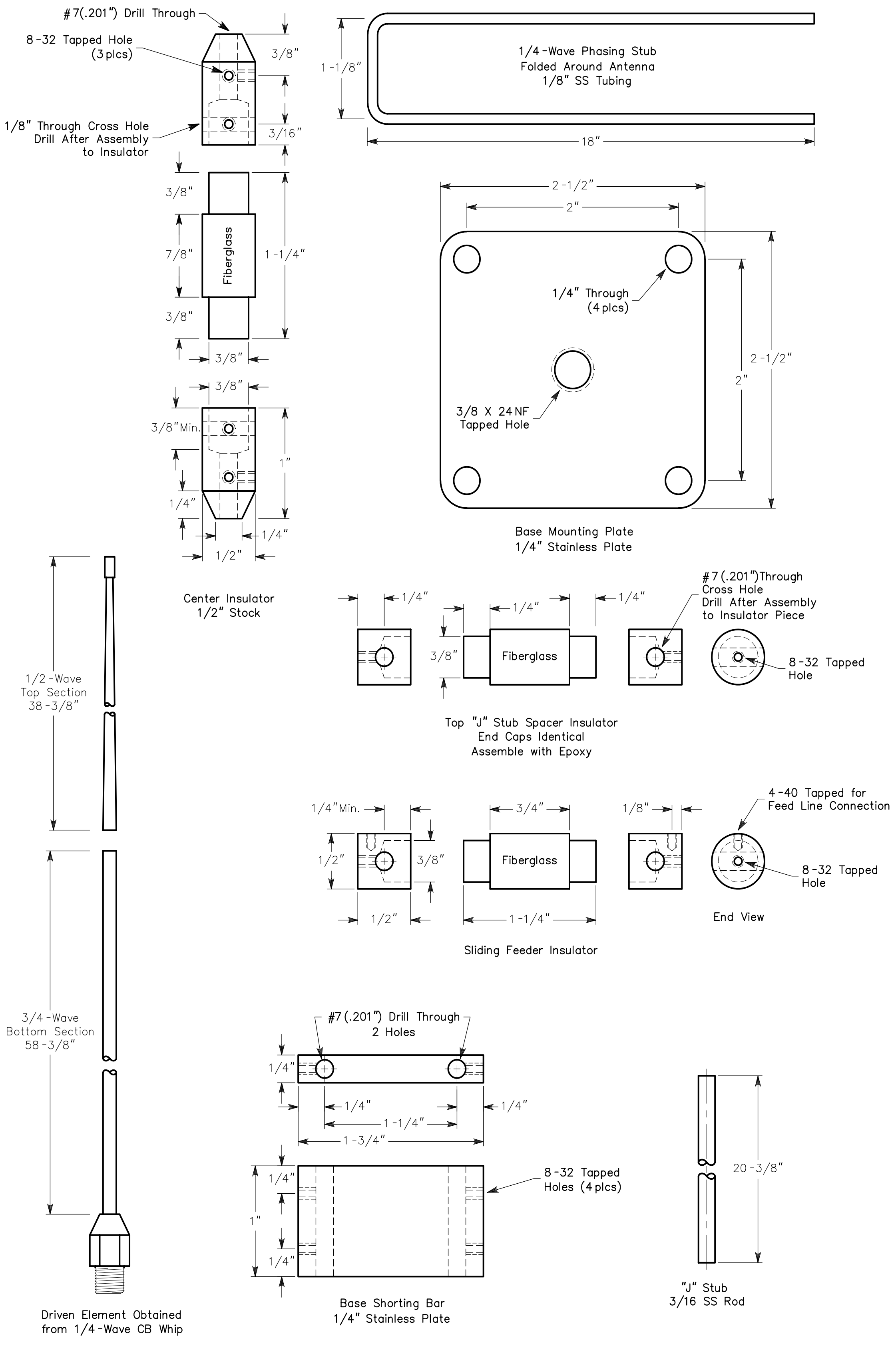
#7(.201") Drill Through
8-32 Tapped Hole (3 plcs)
1/8" Through Cross Hole Drill After Assembly to Insulator
3/8"
3/16"
3/8"
7/8"
Fiberglass
1-1/4"
3/8"
3/8"
3/8"
3/8" Min.
1"
1/4"
1/4"
1/2"
Center Insulator 1/2" Stock
1-1/8"
1/4-Wave Phasing Stub Folded Around Antenna 1/8" SS Tubing
18"
2-1/2"
2"
1/4" Through (4 plcs)
3/8 X 24 NF Tapped Hole
2-1/2"
2"
Base Mounting Plate 1/4" Stainless Plate
1/2-Wave Top Section 38-3/8"
3/4-Wave Bottom Section 58-3/8"
Driven Element Obtained from 1/4-Wave CB Whip
1/4"
1/4"
1/4"
3/8"
Fiberglass
#7 (.201")Through Cross Hole Drill After Assembly to Insulator Piece
8-32 Tapped Hole
Top "J" Stub Spacer Insulator End Caps Identical Assemble with Epoxy
1/4" Min.
3/4"
1/8"
4-40 Tapped for Feed Line Connection
1/2"
3/8"
Fiberglass
8-32 Tapped Hole
1/2"
1-1/4"
End View
Sliding Feeder Insulator
#7 (.201") Drill Through 2 Holes
1/4"
1/4"
1/4"
1-1/4"
1-3/4"
1/4"
1"
1/4"
8-32 Tapped Holes (4 plcs)
Base Shorting Bar 1/4" Stainless Plate
20-3/8"
"J" Stub 3/16 SS Rod

Fig 42—Details of parts used in the construction of the 144-MHz Super-J. Not to scale.

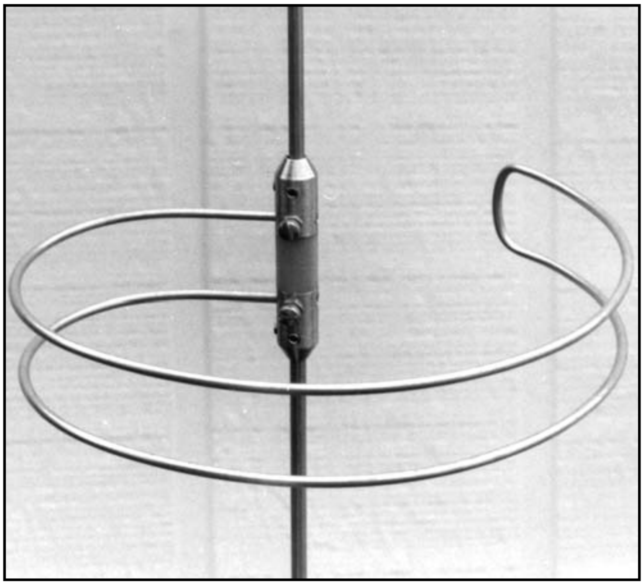

Fig 43—A close-up look at the $^{1}/_{4}$-λ phasing section of the Super-J. The insulator fitting is made of stainless-steel end caps and fiberglass rod.

Performance

Initial tests of the Super-J were performed in portable use and were satisfactory, if not exciting. **Fig 45** shows the Super-J mounted on a wooden mast at a portable site. Simplex communication with a station 40 miles away with a 10-W mobile rig was full quieting both ways. Stations were worked through distant repeaters that were thought inaccessible from this location.

Comparative tests between the Super-J and a commercial $^{5}/_{8}$-wave antenna mounted on the car showed the Super-J to give superior performance, even when the Super-J was lowered to the same height as the car roof. The mast shown in Fig 45 was made from two 8-foot lengths of 1 × 2-inch pine. (The two mast sections and the Super-J can be easily transported in most vehicles.)

The Super-J offers a gain of about 6 dB over a quarter-wave whip and around 3 dB over a $^{5}/_{8}$-wave antenna. Actual performance, especially under less-than-ideal or variable ground conditions, is substantially better than other vertical antennas operated under the same conditions. The freedom from ground-plane radials proves to be a real benefit in maritime mobile operation, especially for those passengers in the back of the boat with sensitive ribs!

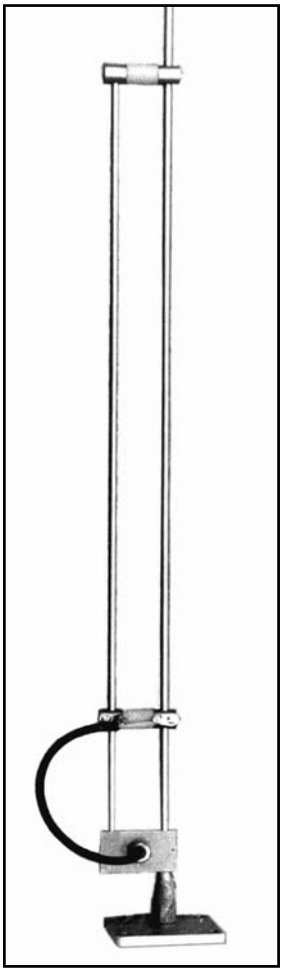

Fig 44—The bottom shorting bar and base mounting plate assembly.

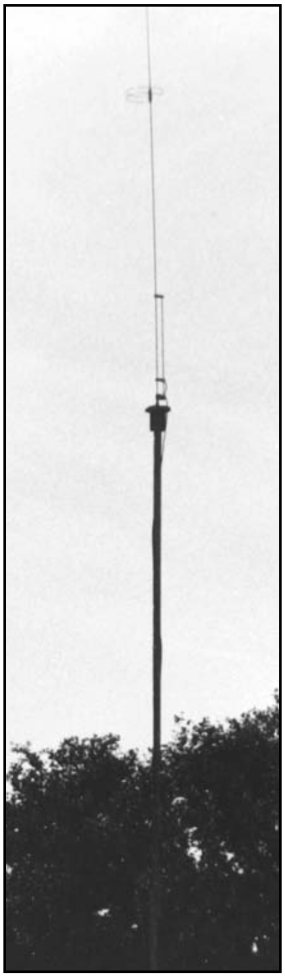

Fig 45—The Super-J in portable use at a field site.

A TOP-LOADED 144-MHz MOBILE ANTENNA

Earlier in this chapter, the merits of various loading schemes for shortened whip antennas were discussed. Quite naturally, one might be considering HF mobile operation for the application of those techniques. But the principles may be applied at any frequency. **Fig 46** shows a 144-MHz antenna that is both top and center loaded. This antenna is suitable for both mobile and portable operation, being intended for use on a handheld transceiver. This antenna was devised by Don Johnson, W6AAQ, and Bruce Brown, W6TWW.

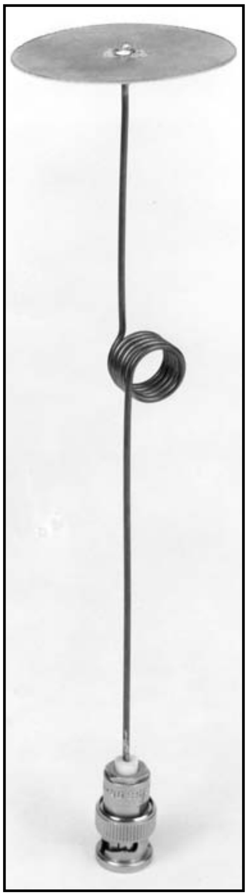

Fig 46—This 144-MHz antenna uses a combination of top and center loading. It offers low construction cost and improved efficiency over continuously loaded rubber-ducky antennas.

A combination of top and center loading offers improved efficiency over continuously loaded antennas such as the “stubby” pictured at the beginning of this chapter. This antenna also offers low construction cost. The only materials needed are a length of stiff wire and a scrap of circuit-board material, in addition to the appropriate connector.

Construction

The entire whip section with above-center loading coil is made of one continuous length of material. An 18-inch length of brazing rod or #14 Copperweld wire is suitable.

In the antenna pictured in Fig 46, the top loading disk was cut from a scrap of circuit-board material, but flashing copper or sheet brass stock could be used instead. Aluminum is not recommended.

The dimensions of the antenna are given in **Fig 47**. First wind the center loading coil. Use a $^1/_2$-inch bolt, wood dowel, or other cylindrical object for a coil form. Begin winding at a point 3 inches from one end of the wire, and wrap the wire tightly around the coil form. Wind $5^1/_2$ turns, with just enough space between turns so they don’t touch.

Remove the coil from the form. Next, determine the length necessary to insert the wire into the connector you’ll be using. Cut the long end of the wire to this length plus 4 inches, measured from the center of the coil. Solder the wire to the center pin and assemble the connector. A tight-fitting sleeve made of Teflon or Plexiglas rod may be used to support and insulate the antenna wire inside the shell. An alternative is to fill the shell with epoxy cement, and allow the cement to set while the wire is held centered in the shell.

The top loading disk may be circular, cut with a hole saw. A circular disk is not required, however—it may be of any shape. Just remember that with a larger disk, less coil inductance will be required, and vice versa. Drill a hole at the center of the disk for mounting it to the wire. For a more rugged antenna, reinforce the hole with a brass eyelet. Solder the disk in place at the top of the antenna, and construction is completed.

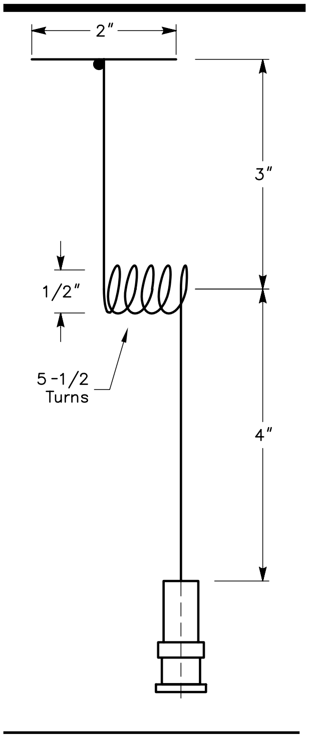

Fig 47—Dimensions for the top-loaded 144-MHz antenna. See text regarding coil length.

Tune-Up

Adjustment consists of spreading the coil turns for the correct amount of inductance. Do this at the center frequency of the range you’ll normally be using. Optimum inductance is determined with the aid of a field strength meter at a distance of 10 or 15 feet.

Attach the antenna to a handheld transceiver operating on low power, and take a field-strength reading. With the transmitter turned off, spread the coil turns slightly, and then take another reading. By experiment, spread or compress the coil turns for the maximum field-strength reading. Very little adjustment should be required. There is one precaution, however. You must keep your body, arms, legs, and head in the same relative position for each field-strength measurement. It is suggested that the transceiver be placed on a nonmetal table and operated at arm’s length for these checks.

Once the maximum field-strength reading is obtained, adjustments are completed. With this antenna in operation, you’ll likely find it possible to access repeaters that are difficult to reach with other shortened antennas. W6AAQ reports that in distant areas his antenna even outperforms a $^5/_8$-λ vertical.

VHF QUARTER-WAVELENGTH VERTICAL

Ideally, a VHF vertical antenna should be installed over a perfectly flat reflector to assure uniform omnidirectional radiation. This suggests that the center of the automobile roof is the best place to mount it for mobile use. Alternatively, the flat portion of the trunk deck can be used, but will result in a directional pattern because of car-body obstruction.

Fig 48 illustrates how a Millen high-voltage connector can be used as a roof mount for a VHF whip. The hole in the roof can be made over the dome light, thus providing accessibility through the upholstery. RG-59 and the $^1/_4$-wave matching section, L (Fig 48C), can be routed between the car roof and the ceiling upholstery and brought into the trunk compartment, or down to the dashboard of the car. Instead of a Millen connector, some operators install an SO-239 coax connector on the roof for mounting the whip. The method is similar to that shown in Fig 48.

It has been established that in general, $^1/_4$-λ vertical antennas for mobile repeater work are not as effective as $^5/_8$-λ verticals are. With a $^5/_8$-λ antenna, more of the transmitted signal is directed at a low wave angle, toward the horizon, offering a gain of about 1 dB over the $^1/_4$-λ vertical. However, in areas where the repeater is located nearby on a very high hill or a mountain top, the $^1/_4$-λ antenna will usually offer more reliable per-

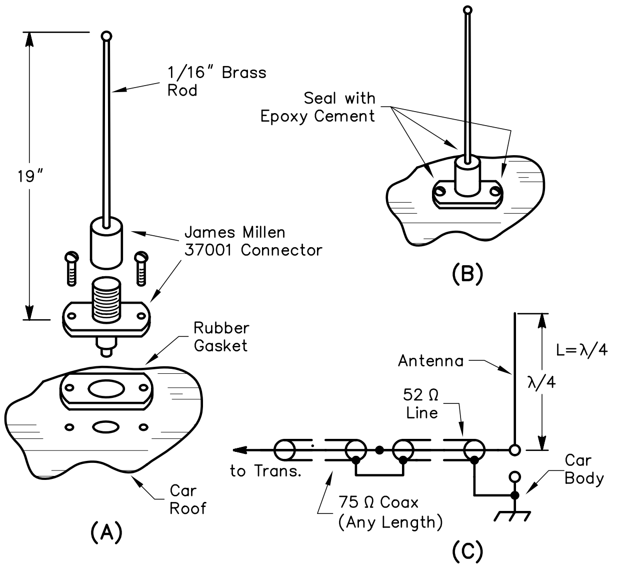

Fig 48—At A and B, an illustration of how a quarter-wavelength vertical antenna can be mounted on a car roof. The whip section should be soldered into the cap portion of the connector and then screwed into the base socket. This arrangement allows for the removal of the antenna when desired. Epoxy cement should be used at the two mounting screws to prevent the entry of moisture through the screw holes. Diagram C is discussed in the text.

formance than a 5/8-λ antenna. This is because there is more power in the lobe of the 1/4-λ vertical at higher angles.

144-MHz 5/8-WAVELENGTH VERTICAL

Perhaps the most popular antenna for 144-MHz FM mobile and fixed-station use is the 5/8-wavelength vertical. As compared to a 1/4-wavelength vertical, it has 1 dB of gain.

This antenna is suitable for mobile or fixed-station use because it is small, omnidirectional, and can be used with radials or a solid-plane ground (such as a car body). If radials are used, they need be only 1/4 wavelength long.

Construction

The antenna shown here is made from low-cost materials. **Fig 49** shows the base coil and aluminum mounting plate. The coil form is a piece of low-loss solid rod, such as Plexiglas or phenolic. The dimensions for this and other parts of the antenna are given in **Fig 50**. A length of brazing rod is used as the whip section.

The whip should be 47 inches long. However, brazing rod comes in standard 36-inch lengths, so if used, it is necessary to solder an 11-inch extension to the top of the whip. A piece of #10 copper wire will suffice. Alternatively, a stainless-steel rod can be purchased to make a 47-inch whip. Shops that sell CB antennas should have such rods for replacement purposes on base-loaded antennas. The limitation one can expect with brazing rod is the relative fragility of the material, especially when the threads are cut for screwing the rod into the base coil form. Excessive stress can cause the rod to break where it enters the form. The problem is compli-

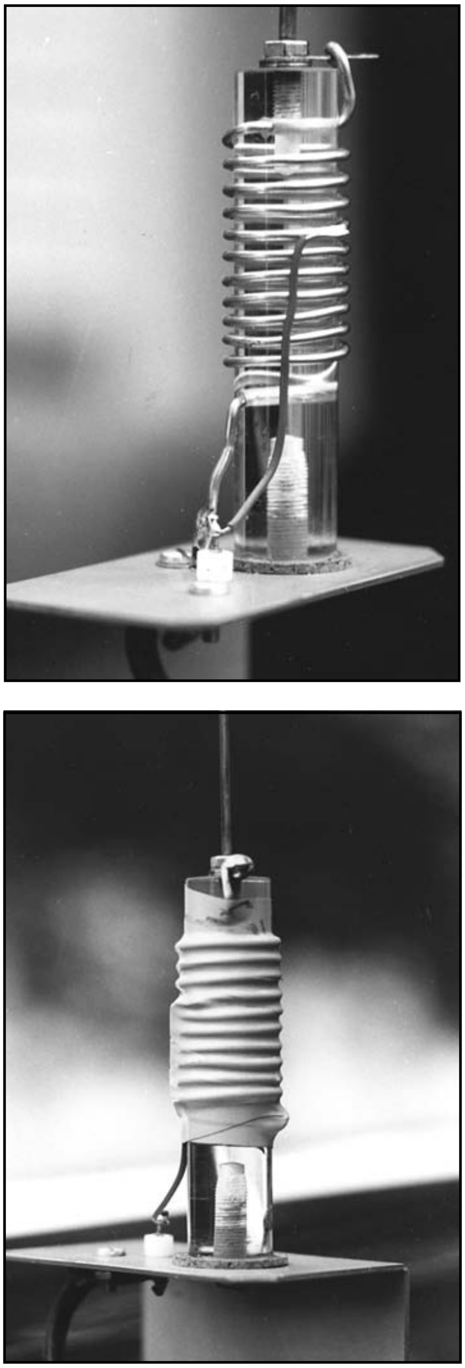

Fig 49—At top, a photograph of the 5/8-λ vertical base section. The matching coil is affixed to an aluminum bracket that screws onto the inner lip of the car trunk. Above, the completed assembly. The coil has been wrapped with vinyl electrical tape to keep out dirt and moisture.

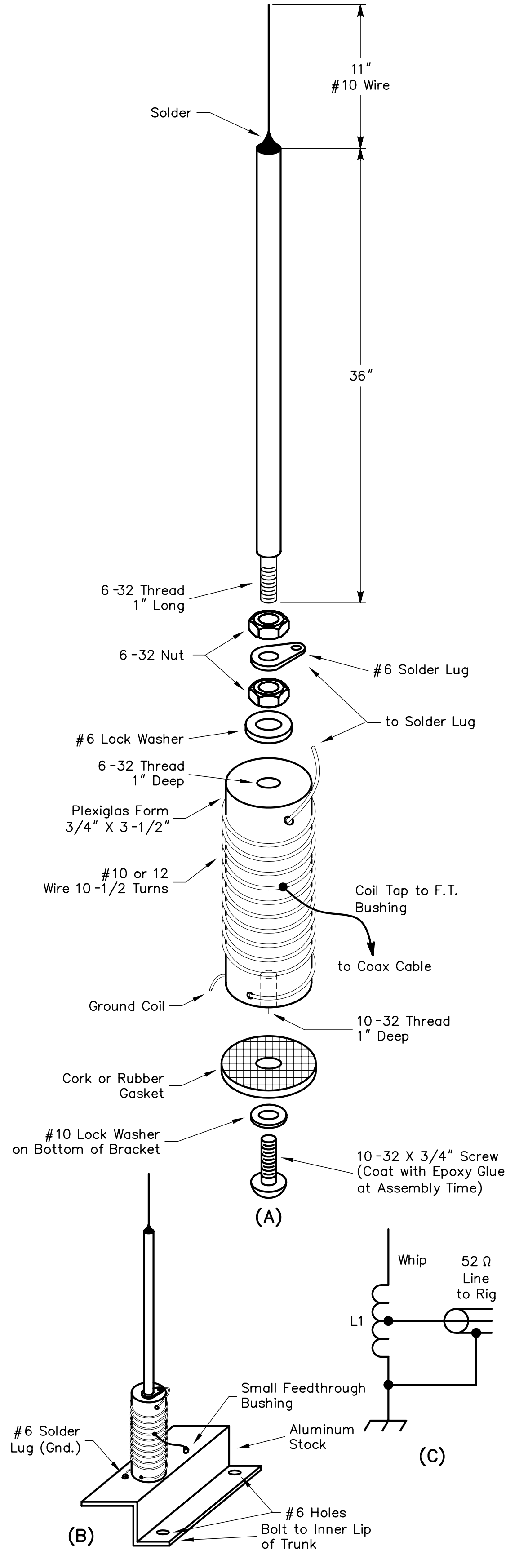
11″
#10 Wire
Solder
36″
6-32 Thread
1″ Long
6-32 Nut
#6 Solder Lug
to Solder Lug
#6 Lock Washer
6-32 Thread
1″ Deep
Plexiglas Form
3/4″ X 3-1/2″
#10 or 12
Wire 10-1/2 Turns
Coil Tap to F.T.
Bushing
to Coax Cable
Ground Coil
10-32 Thread
1″ Deep
Cork or Rubber
Gasket
#10 Lock Washer
on Bottom of Bracket
10-32 X 3/4″ Screw
(Coat with Epoxy Glue
at Assembly Time)
(A)
Whip
52 Ω
Line
to Rig
L1
(C)
Small Feedthrough
Bushing
#6 Solder
Lug (Gnd.)
Aluminum
Stock
#6 Holes
Bolt to Inner Lip
of Trunk
(B)

Fig 50—Structural details for the 2-meter $^5/_8$-λ antenna are provided at A. The mounting bracket is shown at B and the equivalent circuit is given at C.

cated somewhat in this design because a spring is not used at the antenna mounting point. Builders of this antenna can find all kinds of solutions to the problems just outlined by changing the physical design and using different materials when constructing the antenna. The main purpose of this description is to provide dimensions and tune-up information.

The aluminum mounting bracket must be shaped to fit the car with which it will be used. The bracket can be used to effect a no-holes mount with respect to the exterior portion of the car body. The inner lip of the vehicle trunk (or hood) can be the point where the bracket is attached by means of no. 6 or no. 8 sheet-metal screws. The remainder of the bracket is bent so that when the trunk lid or car hood is raised and lowered, there is no contact between the bracket and the moving part. Details of the mounting unit are given in Fig 50B. A 14-gauge metal (or thicker) is recommended for rigidity.

Wind 10$^1/_2$ turns of #10 or #12 copper wire on the $^3/_4$-inch diameter coil form. The tap on L1 is placed approximately four turns below the whip end. A secure solder joint is imperative.

Tune-Up

After the antenna has been mounted on the vehicle, connect an SWR indicator in the 50-Ω transmission line. Key the 144-MHz transmitter and experiment with the coil tap placement. If the whip section is 47 inches long, an SWR of 1:1 can be obtained when the tap is at the right location. As an alternative method of adjustment, place the tap at four turns from the top of L1, make the whip 50 inches long, and trim the whip length until an SWR of 1:1 occurs. Keep the antenna well away from other objects during tune-up, as they may detune the antenna and yield false adjustments for a match.

A 5/8-WAVELENGTH 220-MHz MOBILE ANTENNA

The antenna shown in **Figs 51** and **52** was developed to fill the gap between a homemade $^1/_4$-λ mobile antenna and a commercially made $^5/_8$-λ model. While antennas can be made by modifying CB models, that presents the problem of cost in acquiring the original antenna. The major cost in this setup is the whip portion. This can be any tempered rod that will spring easily.

Construction

The base insulator portion is made of $^1/_2$-inch Plexiglas rod. A few minutes' work on a lathe is sufficient to shape and drill the rod. (The innovative builder can use an electric drill and a file for the lathe work.)

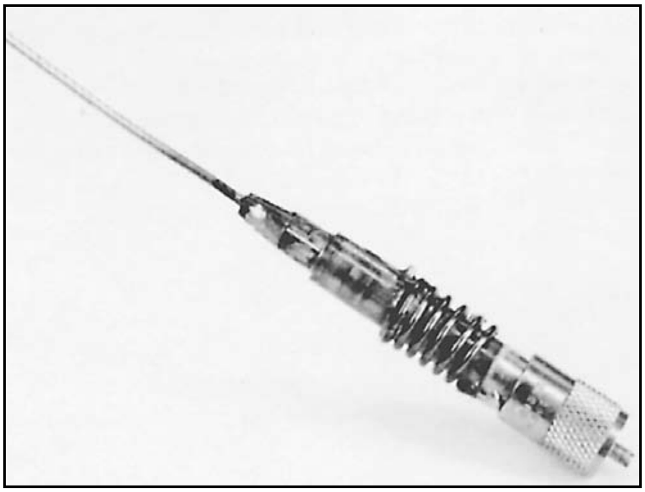

Fig 51—The 220-MHz 5/8-λ mobile antenna. The coil turns are spaced over a distance of 1 inch, and the bottom end of the coil is soldered to the coax connector.

The bottom 1/2 inch of the rod is turned down to a diameter of 3/8 inch. This portion will now fit into a PL-259 UHF connector. A 1/8-inch diameter hole is drilled through the center of the rod. This hole will hold the wires that make the connections between the center conductor of the connector and the coil tap. The connection between the whip and the top of the coil is also run through this opening. A stud is force-fitted into the top of the Plexiglas rod. This allows for removal of the whip from the insulator.

The coil should be initially wound on a form slightly smaller than the base insulator. When the coil is transferred to the Plexiglas rod, it will keep its shape and will not readily move. After the tap point has been determined, a longitudinal hole is drilled into the center of the rod. A #22 wire can then be inserted through the center of the insulator into the connector. This method is also used to attach the whip to the top of the coil. After the whip has been fully assembled, a coating of epoxy cement is applied. This seals the entire assembly and provides some additional strength. During a full winter's use there was no sign of cracking or other mechanical failure. The adjustment procedure is the same as for the 144-MHz version described previously.

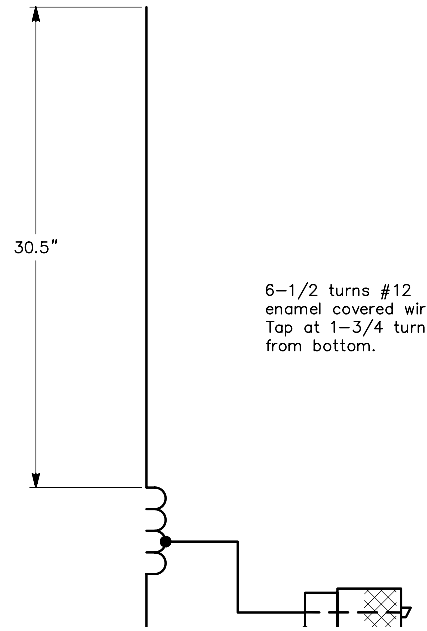

Fig 52—Diagram of the 220-MHz mobile antenna.

HF Antennas For Sailboats

This material was contributed by Rudy Severns, N6LF. Many of the antenna ideas appearing earlier in this chapter can be applied to sailboats. However, the presence of the mast and the rigging, plus the prevalence of non-conducting fiberglass hulls complicates the issue. There are many possibilities for antennas aboard sailboats. This includes both permanently installed antennas and antennas that can be hoisted for temporary use at anchor:

1. ***Permanent***
   Commercial or home-brew automobile-type verticals
   Backstay verticals and slopers
   Shunt feed of uninsulated rigging

2. ***Temporary***
   Sloping dipoles
   Inverted V's
   Yagis

You should remember some basic facts of life on a sailboat:

1) On most boats the spars, standing rigging and some running rigging will be conductors. Stainless steel wire is usually used for the rigging and aluminum for the spars.
2) Topping lifts, running backstays and jackstays all may be made of conducting materials and may often change position while the boat is underway. This changes the configuration of the rigging and may affect radiation patterns and feed-point impedances.
3) Shipboard antennas will always be close to the mast and rigging, in terms of electrical wavelength. Some antennas may in fact be part of the rigging.
4) The feed-point impedance and radiation pattern can be strongly influenced by the presence of the rigging.
5) Because of the close proximity, the rigging is an integral part of the antenna and should be viewed as such.
6) The behavior of a given antenna will depend on the details of the rigging on a particular vessel. The performance of a given antenna can vary widely on different boats, due to differences in dimensions and arrangement of the rigging.
7) Even though you may be floating on a sea of salt water, grounding still requires careful attention!

ANTENNA MODELING

Because of the strong interaction between the rigging and the antenna, accurate prediction of radiation patterns and a reasonable guess at expected feed-point impedance requires that you model both the antenna and the rigging. Unless you do accurately model the system, considerable cut-and-try may be needed. This can be expensive when it has to be done in 1 × 19 stainless steel wire with $300 swaged insulator fittings!

In fact, when your antenna is going to be part of standing rigging, it's a very good idea to try your designs out at the dock. You could temporarily use Copperweld wire and inexpensive insulators in place of the stainless rigging wire and the expensive insulators. This approach can save a good deal of money and aggravation. A wide variety of modeling programs are available and can be very helpful in designing a new antenna but they have to be used with some caution:

1) The rigging will have many small intersection angles and radically different conductor diameters, this can cause problems for *NEC* and *MININEC* programs.
2) You must usually taper the segment lengths near the junctions. This is done automatically in programs like *ELNEC* and *EZNEC*.
3) It is usually necessary to use one wire size for the mast, spars and rigging. Some improvement in accuracy can be obtained by modeling the mast as a cage of 3 or 4 wires.

The predicted radiation patterns will be quite good but the feed-point impedance predictions should be viewed as preliminary. Some final adjustment will usually be required. Because of the wide variation between boats, even those of the same class, each new installation is unique and should be analyzed separately.

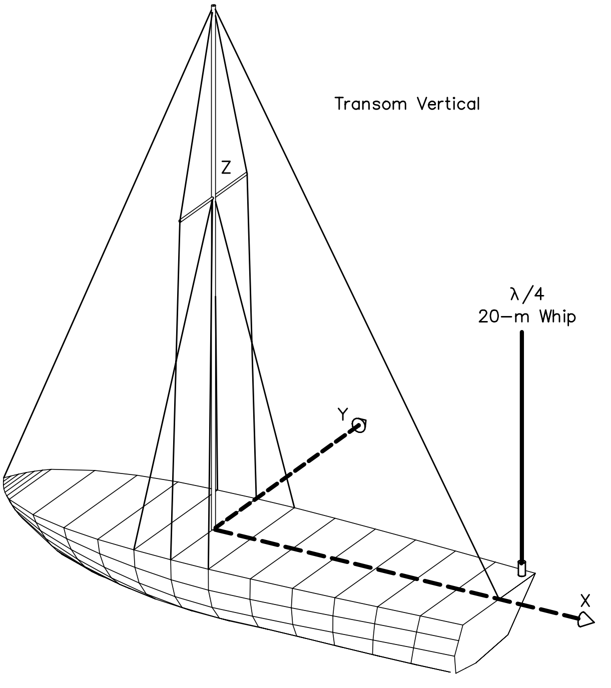

Fig 53—An example of a 20-meter λ/4 whip mounted on the transom. A local ground system must also be provided, as described in the section on grounding.

A SAFETY NOTE

Ungrounded rigging endpoints near deck level can have high RF potentials on them when you transmit. For example, the shrouds on a fiberglass boat connect to chainplates that are bolted to the hull, but are not grounded. These can inflict painful RF burns on the unwary, even while operating at low power! As a general rule all rigging, spars and lifelines near deck level should be grounded. This also makes good sense for lightning protection. For a backstay antenna with its feed point near deck level, a sleeve of heavy wall PVC pipe can be placed over the lower end of the stay as a protective shield.

TRANSOM AND MASTHEAD MOUNTED VERTICALS

A very common antenna for boats is a vertical, either a short mobile antenna or a full λ/4, placed on the transom, as shown in **Fig 53**. Note that in this example the antenna is mounted off to one side—it could also be mounted in the center of the transom. The 20-meter radiation pattern for this antenna as shown in **Fig 54**. Unlike a free standing vertical, this antenna doesn't have an omnidirectional pattern. It is asymmetrical, with a front-to-back ratio of about 13 dB. Further, the angle for maximum gain is offset in the direction the antenna is placed on the transom.

This is a very good example of the profound effect the rigging can have on any antenna used on board a sailboat. Not only is the pattern affected but the feed-point impedance will be reduced from a nominal 36 Ω to 25 to 30 Ω.

The directive gain can be useful—if you point the boat in the right direction! Usually, however, a more uniform omnidirectional pattern is more desirable. It is tempting to suggest putting the vertical at the masthead, perhaps using a 6-foot loaded automobile whip, with the mast and rigging acting as a ground plane. **Fig 55** shows such a system. Unfortunately, this usually doesn't work very well because the overall height of the mast and antenna will very likely be > 5/8 λ. This will result in high-angle lobes, as shown in **Fig 56**. Depending on the mast height, this idea may work reasonably well on 40 or 80 meters, but

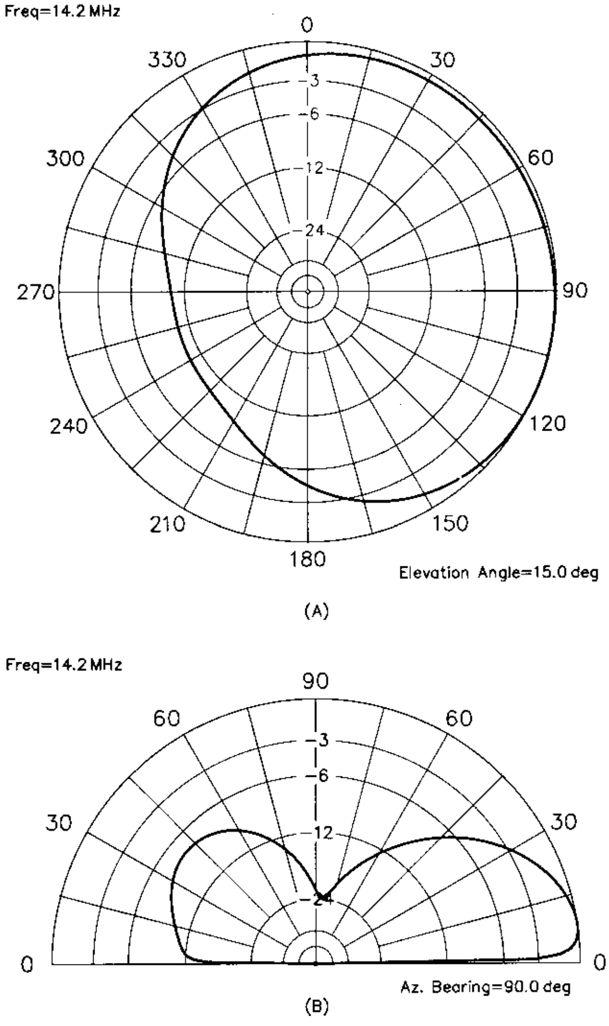

Fig 54—Typical radiation pattern for the λ/4 transom mounted whip in Fig 53.

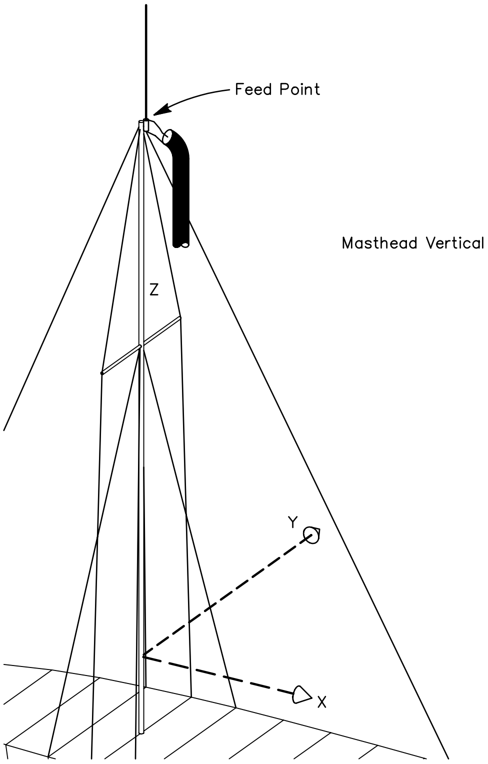

Fig 55—A whip mounted at the masthead. The feed line is fed back down the mast either inside or outside. The base of the mast and the rigging is assumed to be properly grounded.

you will still be faced with severe mechanical stress due to magnified motion at the masthead in rough sea. The masthead is usually reserved to VHF antennas, with their own radial ground plane.

THE BACKSTAY VERTICAL

A portion of the backstay can be insulated and used as a vertical as shown in **Fig 57**. The length of the insulated section will be λ/4 on the lowest band of interest. Typically, due to the loading effect of the rest of the rigging, the resonant length of the insulated section will be shorter than the classic 234/f (MHz) relation, although it can in some case actually be longer. Either modeling or trial adjustment can be used to determine the actual length needed. On a typical 35 to 40-foot sailboat, the lowest band for λ/4 resonance will be 40 meters due to the limited length of the backstay. Examples of the radiation patterns on several bands for such an an-

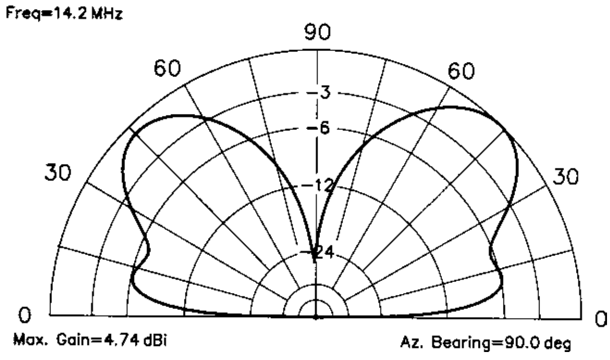

Fig 56—Typical radiation pattern for a masthead-mounted vertical. The multiple vertical lobes are due to the fact the antenna is higher than λ/2.

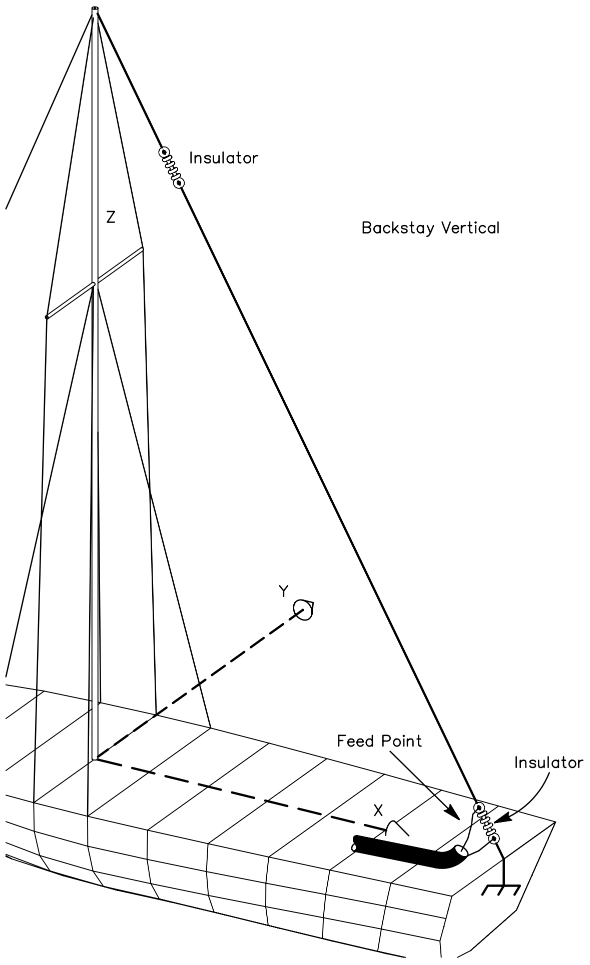

Fig 57—An example of a backstay vertical. A local ground point must be established on the transom next to the base of the backstay.

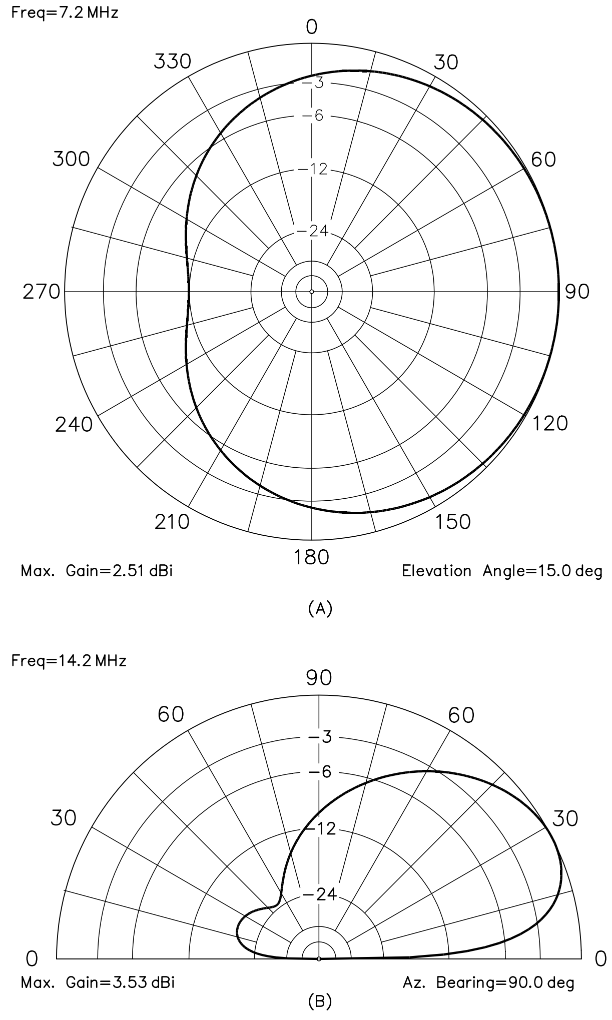

Fig 58—Typical radiation patterns on 40 meters for the backstay vertical in Fig 57.

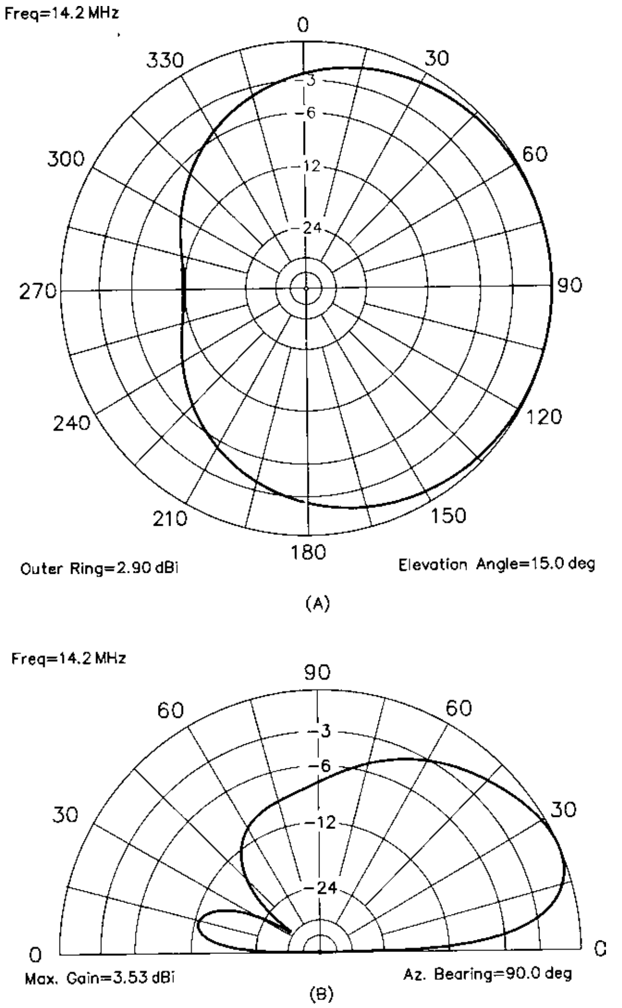

Fig 59—Typical radiation patterns on 20 meters for the backstay vertical in Fig 57.

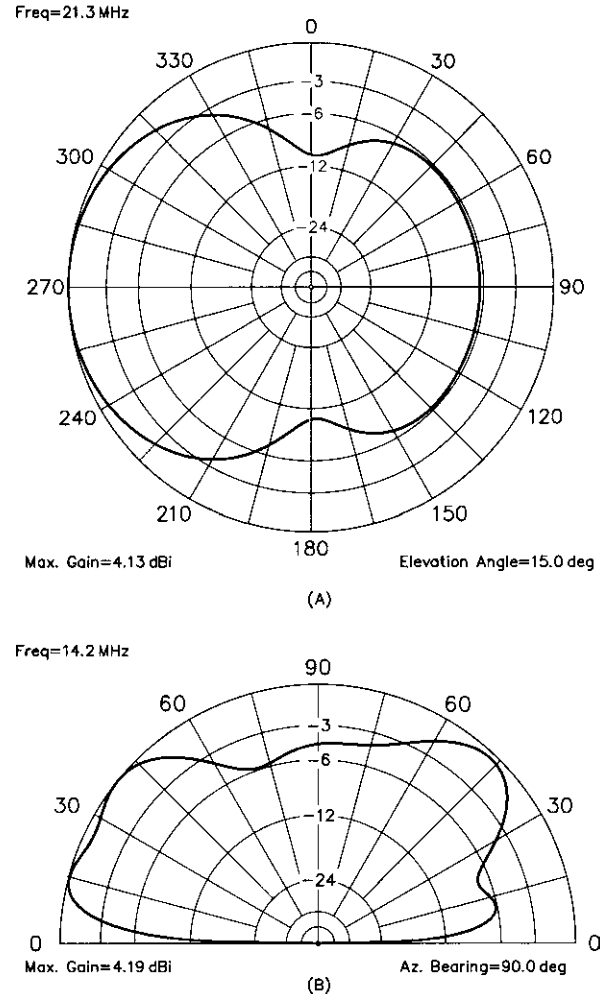

Fig 60—Typical radiation patterns on 15 meters for the backstay vertical in Fig 57.

tenna are given in **Figs 58** through **60**.

The pattern is again quite directional due to the presence of the mast and rigging. On 15 meters, where the antenna is approximately $^3/_4$ λ, higher angle lobes appear. On 40 and 15 meters, the feed point is near a current maximum and is in the range of 30 to 50 Ω. On 20 meters, however, the feed point is a very high impedance because the antenna is near λ/2 resonance. One way to get around this problem for multiband use is to make the antenna longer than λ/4 on the lowest band. If the lowest band is 40 meters then on 20 meters the feed-point impedance will be much lower. This antenna is non-resonant on any of the bands but can be conveniently fed with a tuner because the feed-point impedances are within the range of commonly available commercial tuners. Tuners specifically intended for marine applications frequently can accommodate very high input impedances, but they tend to be quite expensive.

The sensitivity of the radiation pattern to small details of the mast and rigging is illustrated in **Fig 61**. This is the same antenna as shown in Fig 57 with the exception that the forestay is assumed to be ungrounded. In this particular example, ungrounding the forestay drastically increases the front-to-back ratio. With slightly different dimensions, however, the pattern could have changed in other ways.

High-quality insulators for rigging wire can be quite expensive and represent a potential weak point—if they fail the mast may come down. It is not absolutely necessary to use two insulators in a backstay vertical. As shown in **Fig 62**, the upper insulator may be omitted. The radiation patterns are shown in **Figs 63** and **64**. In this case

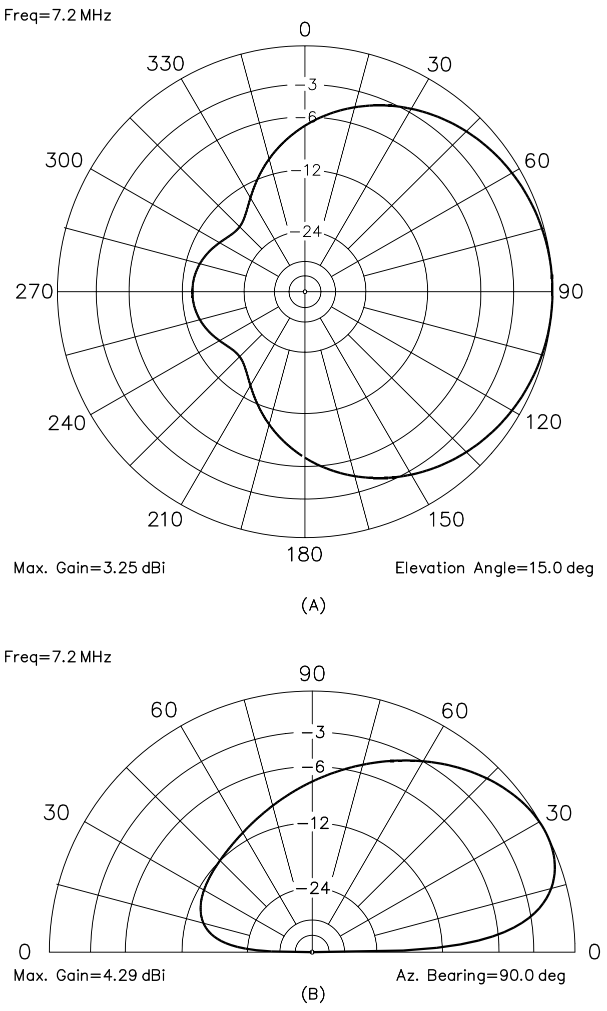

Fig 61—The effect of ungrounding the forestay on the radiation pattern. This is for 40-meter operation.

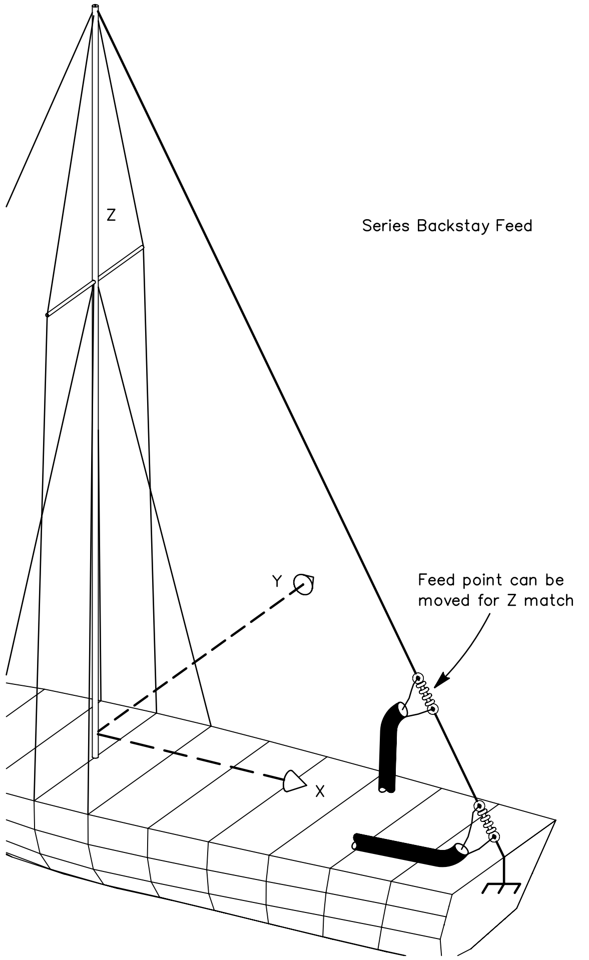

Fig 62—Feeding the backstay without an insulator at the top. The feed point may be moved along the antenna to find a point with a better match on a particular band or to provide a better range of impedances for the tuner to match. The coaxial feed line is taped to the lower portion of the backstay. Again, a good local ground is needed at the base of the backstay.

the pattern is actually more symmetrical than it was with an upper insulator—but this may not hold true for other rigging dimensions. The feed point does not have to be at the bottom of the backstay. As indicated in Fig 62, the feed point can be moved up into the backstay to achieve a better match or a more desirable feed-point impedance variation with frequency. In that case, the center of the coaxial feed line is connected to the upper section and the shield to the lower section. The cable is then taped to the lower portion of the backstay.

If single-band operation is all you want, even the lower insulator can be omitted by using shunt feed. A gamma match would be quite effective for this purpose, as discussed in Chapter 6 when driving a grounded tower.

A 40-METER BACKSTAY HALF SLOPER

A half-sloper antenna can be incorporated into the backstay, as shown in **Fig 65**. This will behave very much the same as the slopers described in Chapter 6. The advantage of this antenna for a sailboat installation is that you don't need to create a good ground connection at the stern, as you would have to do for a transom-mounted vertical or the backstay vertical just described. This may be more convenient. The mast, shrouds and stays must still be grounded for the half-sloper but the arrangement is somewhat simpler.

TEMPORARY ANTENNAS

Not everyone needs permanent antennas. A variety

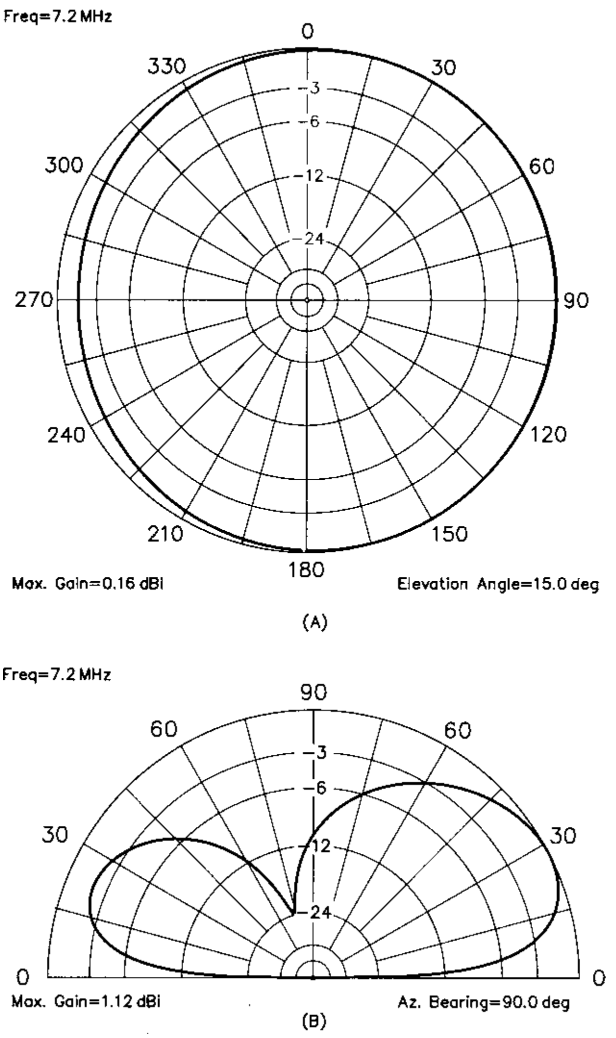

Fig 63—Typical 40-meter radiation pattern in Fig 62.

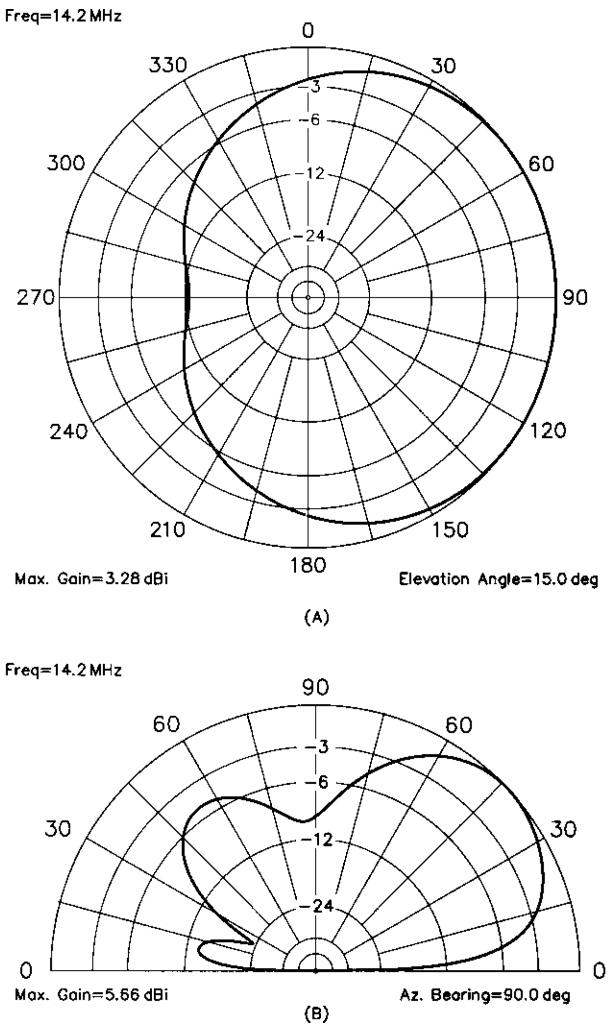

Fig 64—Typical 20-meter radiation pattern in Fig 62.

of temporary antennas can be arranged. A few of these are shown in **Figs 66** through **68**. Of all of these the rigid dipole (Fig 66) will provide the best operation and will have a pattern close to that expected from a freestanding dipole. The other two examples will be strongly affected by their close proximity to the rigging.

GROUNDING SYSTEMS

You may be sitting in the middle of a thousand miles of saltwater. This is great for propagation but you will still have to connect to that ground if you want to use a vertical. There are many possibilities, but the scheme shown in **Fig 69** is representative. First a bonding wire, or better yet a copper strap (it can be very thin!), is connected from bow-to-stern on each side, connecting the forestay, lifeline stanchions, chainplates, bow and stern pulpits and the backstay. Other bonding wires are run from the bow, stern and chainplates on both sides to a common connection at the base of the mast. The fore-and-aft bonding can be attached to the engine and to the keel bolts. The question arises: "What about electrolysis between the keel and propeller if you bond them together?" This has to be dealt with on a case-by-case basis. If your protective zincs are depleting more rapidly after you install a bonding scheme, change it by disconnecting something, the engine-shaft-propeller, for example.

Grounding will vary in every installation and has to be customized to each vessel. However, just as on shore, the better the ground system, the better the performance of the vertical!

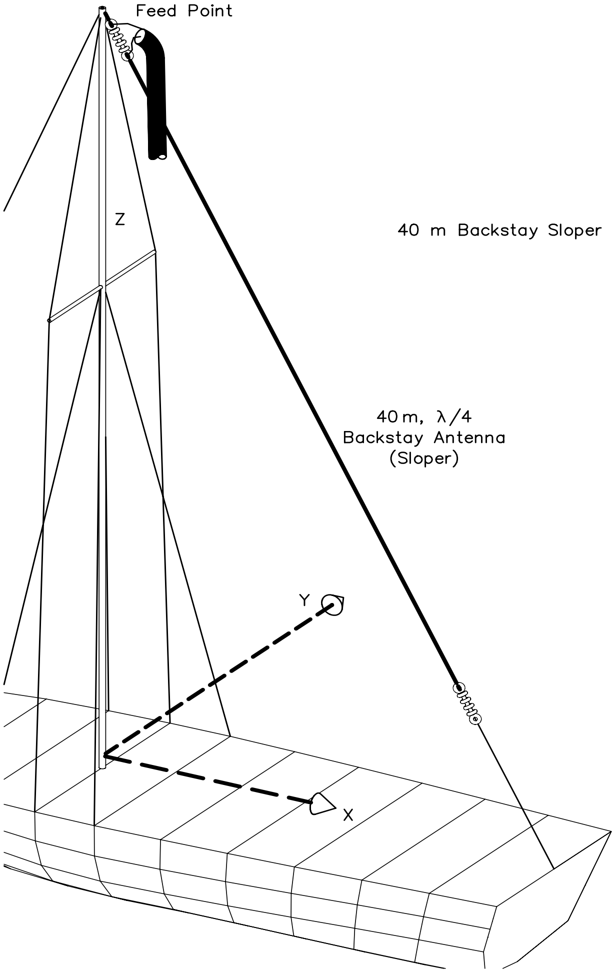
Feed Point
Z
40 m Backstay Sloper
40 m, λ/4
Backstay Antenna
(Sloper)
Y
X

Fig 65—A 40-meter half-sloper fed at the masthead.

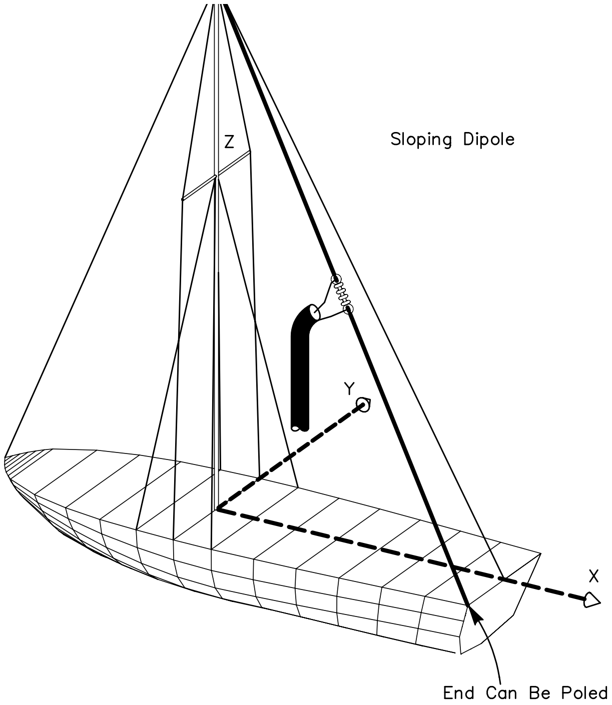
Z
Sloping Dipole
Y
X
End Can Be Poled

Fig 67—One end of dipole can be attached to the main halyard and pulled up to the masthead. The bottom end of the dipole should be pulled out away from the bottom end of the dipole should be pulled out away from the rigging as much as possible to reduce the impact of the rigging on the impedance.

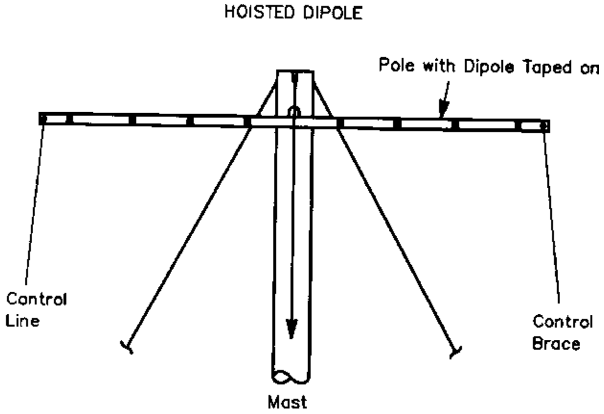
HOISTED DIPOLE
Pole with Dipole Taped on
Control
Line
Control
Brace
Mast

Fig 66—A dipole can be taped to a wood or bamboo pole and hoisted to the masthead with the main halyard while at anchor. It is possible to make this a multiband dipole.

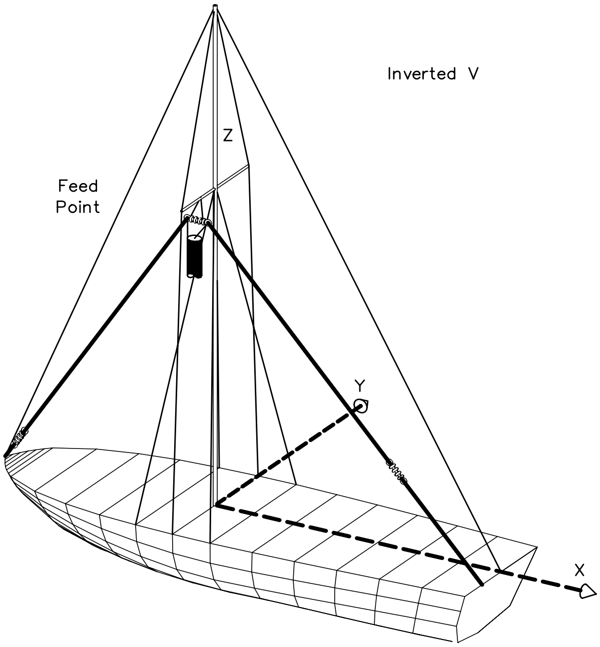
Inverted V
Z
Feed
Point
Y
X

Fig 68—The flag halyard can be used to hoist the center of an inverted V to the spreaders, or alternatively, the main halyard can be used to hoist the center of the antenna to the masthead. Interaction between the rigging and the antenna will be very pronounced and the length of the antenna will have to be adjusted on a cut-and-try basis.

Antennas for Power Boats

Powerboaters are not usually faced with the problems and opportunities created by the mast and rigging on a sailboat. A powerboat may have a small mast, but usually not on the same scale as a sailboat. Antennas for power boats have much more in common with automotive mobile operation, but with some important exceptions:

1) In an automobile, the body is usually metal and it provides a groundplane or counterpoise for a whip antenna. Most modern powerboats, however, have fiberglass hulls. These are basically insulators, and will not work as counterpoises. (On the other hand, metal-hulled power boats can provide nearly ideal grounding!)
2) A height restriction on automotive mobile whips is imposed by clearance limits on highway overpasses and also by the need to sustain wind speeds of up to 80 miles per hour on the highway.
3) In general, powerboats can have much taller antennas that can be lowered for the occasional low bridge.
4) The motion on a powerboat, especially in rough seas, can be quite severe. This places additional mechanical strain on the antennas.
5) On both powerboats and sailboats, operation in a saltwater marine environment is common. This means that a careful choice of materials must be made for the antennas to prevent corrosion and premature failure.

The problem of a ground plane for vertical antennas can be handled in much the same manner as shown in **Fig 69** for sailboats. Since there will most likely not be a large keel structure to connect to and provide a large surface area, additional copper foil can be added inside the hull to increase the counterpoise area. Because of the small area of the propeller, it may be better not to connect to the engine, but to rely instead on increasing the area of the counterpoise and operate it as a true counterpoise—that is, isolated from ground. Sometimes a number of radial wires are used for a vertical, much like that for a ground-plane antenna. This is not a very good idea unless the “wires” are actually wide copper-foil strips that can lower the Q substantially.

The problem is the high voltage present at the ends of normal ground-plane antenna radials. For a boat these radials are likely to be in close proximity to the cabin, which in turn contains both people and electronic equipment. The high potential at the ends of the radials is both a safety hazard and can result in RF coupling back into the equipment, including ham gear, navigational instruments and entertainment devices. The cook is not likely to be happy if he or she gets an RF burn after touching the galley stove! Decoupling the counterpoise from the transmission line, as discussed in Chapter 6, can be very helpful to keep RF out of other equipment.

One way to avoid many of the problems associated with grounding is to use a rigid dipole antenna. On 20 meters and higher, a rigid dipole made up from aluminum tubing, fiberglass poles or some combination of these, can be effective. As shown in **Fig 70A** the halves of the dipole can be slanted upward like rabbit ears to reduce the wingspan and increase the feed-point impedance for a better match to common coax lines. On the lower bands a pair of mobile whips can be used, as shown in Fig 70B. Home-brew coils could also be used.

For short-range communication, a relatively low dipole over saltwater can be effective. However, if long-range communication is needed, then a well-designed vertical, operating over seawater, will work much better. For these to work, of course, you must solve the ground problems associated with a vertical.

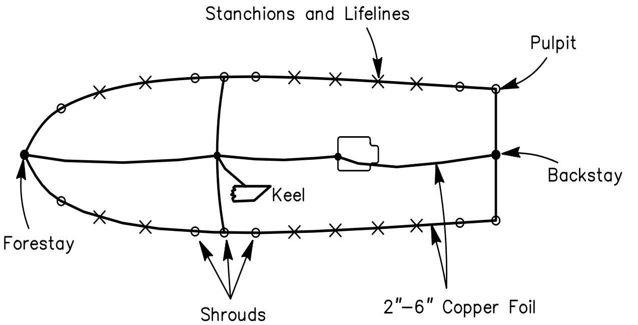

Fig 69—A typical sailboat grounding scheme.

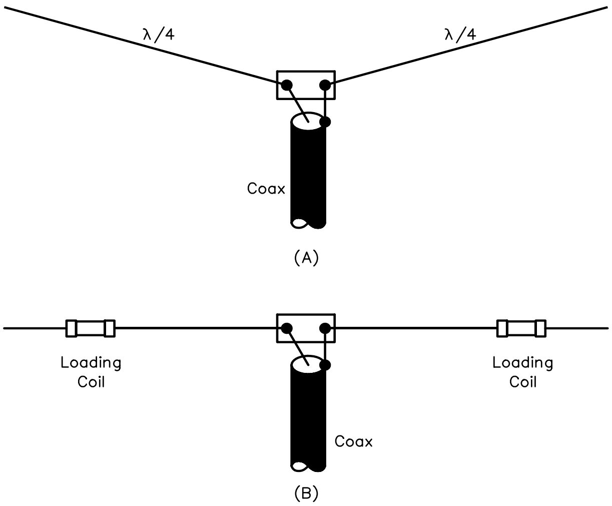

Fig 70—At A, a rigid dipole made from aluminum tubing, fiberglass poles or a combination of these. At B, a pair of mobile whips used as a dipole.

It is not uncommon for large powerboats to have a two or three-element multiband Yagi installed on a short mast. While these can be effective, if they are not mounted high (> λ/2) they may be disappointing for longer-range communication. Over saltwater, vertical polarization is very effective for longer distances. A simpler, but well-designed, vertical system on a boat may outperform a low Yagi.

The combination of a good ground system and one of the high-quality, motor-tuned multiband mobile whips now available commercially can also be very effective.

BIBLIOGRAPHY

Source material and more extended discussions of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

J. S. Belrose, "Short Antennas for Mobile Operation," *QST*, Sep 1953.

J. S. Belrose, "Vertical J Antenna for 2 meters," *The Canadian Amateur*, Jul/Aug 1979, pp 23-26.

J. M. Boyer, "Antenna-Transmission Line Analog," *Ham Radio*, May 1977.

B. F. Brown, "Tennamatic: An Auto-Tuning Mobile Antenna System," *73*, Jul 1979.

B. F. Brown, "Optimum Design of Short Coil-Loaded High-Frequency Mobile Antennas," *ARRL Antenna Compendium Volume 1* (Newington: ARRL, 1985), p 108.

W. B. Freely, "A Two-meter J Antenna," *QST*, Apr 1977, pp 35-36.

E. A. Laport, *Radio Antenna Engineering* (New York: McGraw-Hill Book Co., 1952), p 23.

C. E. Smith and E. M. Johnson, "Performance of Short Antennas," Proceedings of the IRE, Oct 1947.

F. E. Terman, *Radio Engineering Handbook*, 3rd edition (New York: McGraw-Hill Book Co., 1947), p 74.

Chapter 17

Repeater Antenna Systems

There is an old adage in Amateur Radio that goes, "If your antenna did not fall down last winter, it wasn't big enough." This adage might apply to antennas for MF and HF work, but at VHF things are a bit different, at least as far as antenna size is concerned. VHF antennas are smaller than their HF counterparts, but yet the theory is the same—a dipole is a dipole, and a Yagi is a Yagi, regardless of frequency. A 144-MHz Yagi may pass as a TV antenna, but most neighbors can easily detect a radio hobbyist if a 14-MHz Yagi looms over his property.

Repeater antennas are discussed in this chapter. Because the fundamental operation of these antennas is no different than presented in earlier chapters, there is no need to delve into any exotic theory. Certain considerations must be made and certain precautions must be observed, however, since most repeater operations—amateur and commercial—take place at VHF and UHF.

Basic Concepts

The antenna is a vital part of any repeater installation. Because the function of a repeater is to extend the range of communications between mobile and portable stations, the repeater antenna should be installed in the best possible location to provide the desired coverage. This usually means getting the antenna as high above average terrain as possible. In some instances, a repeater may need to have coverage only in a limited area or direction. When this is the case, antenna installation requirements will be completely different, with certain limits being set on height, gain and power.

Horizontal and Vertical Polarization

Until the upsurge in FM repeater activity in the 1970s, most antennas used in amateur VHF work were horizontally polarized. These days, very few repeater groups use horizontal polarization. (One of the major reasons for using horizontal polarization is to allow separate repeaters to share the same input and/or output frequencies with closer-than-normal geographical spacing.) The vast majority of VHF and UHF repeaters use vertically polarized antennas, and all the antennas discussed in this chapter are of that type.

Transmission Lines

Repeaters provide the first venture into VHF and UHF work for many amateurs. The uninitiated may not be aware that the transmission lines used at VHF become very important because feed-line losses increase with frequency.

The characteristics of feed lines commonly used at VHF are discussed in Chapter 24, Transmission Lines. Although information is provided there for small-diameter RG-58 and RG-59 coaxes, these should not be used except for very short feed lines (25 feet or less). These cables are very lossy at VHF. In addition, the losses can be much higher if fittings and connections are not carefully installed.

The differences in loss between solid-polyethylene dielectric types (RG-8 and RG-11) and those using foam polyethylene are significant at VHF and UHF. If you can afford the line with the least loss, buy it.

If you must bury coaxial cable, check with the cable manufacturer before doing so. Many popular varieties of coaxial cable should not be buried, since the dielectric can become contaminated from moisture and soil chemicals. Some coaxial cables are labeled as non-contaminating. Such a label is the best way to be sure your cable can be buried without damage.

Matching

Losses are lowest in transmission lines that are matched to their characteristic impedances. If there is a mismatch at the end of the line, the losses increase. The *only way* to reduce the SWR on a transmission line is by matching the line *at the antenna*. Changing the length of a transmission line does not reduce the SWR. The SWR is established by the impedance of the line and the impedance of the antenna, so matching must be done at the antenna end of the line.

The importance of matching, so far as feed-line losses are concerned, is sometimes overstressed. But under some conditions, it is necessary to minimize feed-line losses related to SWR if repeater performance is to be consistent. It is important to keep in mind that most VHF/UHF equipment is designed to operate into a 50-Ω load. The output circuitry will not be loaded properly if connected to a mismatched line. This leads to a loss of power, and in some cases, damage to the transmitter.

Repeater Antenna System Design

Choosing a repeater or remote-base antenna system is as close as most amateurs come to designing a commercial-grade antenna system. The term *system* is used because most repeaters utilize not only an antenna and a transmission line, but also include duplexers, cavity filters, circulators or isolators in some configuration. Assembling the proper combination of these items in constructing a reliable system is both an art and a science. In this section prepared by Domenic Mallozzi, N1DM, the functions of each component in a repeater antenna system and their successful integration are discussed. While every possible complication in constructing a repeater cannot be foreseen at the outset, this discussion should serve to steer you along the right lines in solving any problems encountered.

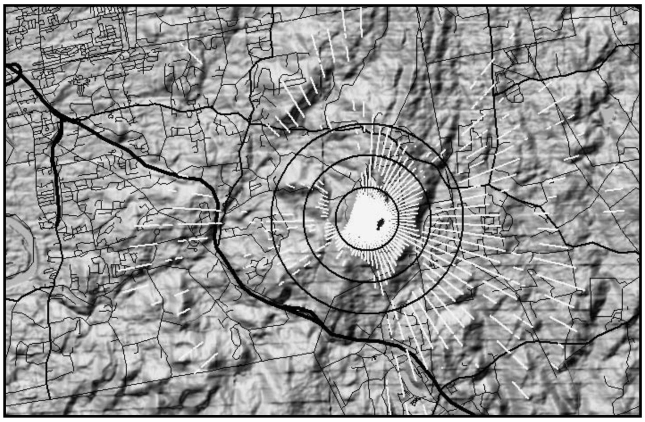

Fig 1—*MicroDEM* topographic map, showing the coverage for a repeater placed on a 30-meter high tower in Glastonbury, CT. The white radial lines indicate the coverage in 5° increments of azimuth around the tower. The range circles are 1000 meters apart.

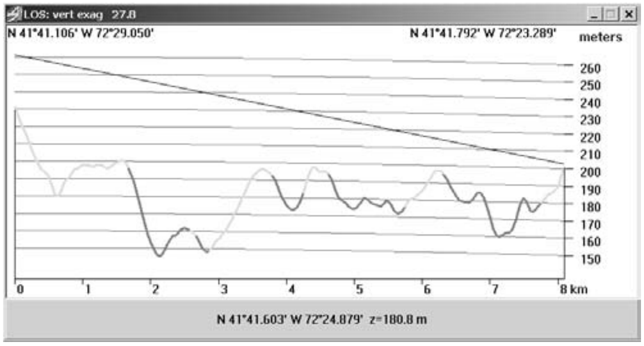
LOS: vert exag 27.8
N 41°41.106' W 72°29.050'
N 41°41.792' W 72°23.289'
meters
260
250
240
230
220
210
200
190
180
170
160
150
0
1
2
3
4
5
6
7
8 km
N 41°41.603' W 72°24.879' z=180.8 m

Fig 2—An "LOS" (line of sight) profile at an azimuth of 80° from the tower in Fig 1. The light-gray portions of the terrain profile are visible from the top of the tower, while the dark portions are blocked by the terrain.

COMPUTING THE COVERAGE AREA FOR A REPEATER

Modern computer programs can show the coverage of a repeater using readily available topographic data from the Internet. In Chapter 3, The Effects of Ground, we described the *MicroDEM* program supplied on the CD-ROM accompanying this book. Dr Peter Guth, the author of *MicroDEM*, built into it the ability to generate terrain profiles that can be used with ARRL's *HFTA* (HF Terrain Assessment) program (also included on the CD-ROM).

MicroDEM has a wide range of capabilities beyond simply making terrain profiles. It can do *LOS* (line of sight) computations, based on visual or radio-horizon considerations. **Fig 1** shows a *MicroDEM* map for the area around Glastonbury, Connecticut. This is somewhat hilly terrain, and as a result the coverage for a repeater placed here on a 30-meter (100-foot) high tower would be somewhat spotty. Fig 1 shows a "Viewshed" on the map, in the form of the white terrain profile strokes in 5° increments around the tower.

Fig 2 shows the LOS for an azimuth of 80°, from a 30-meter high tower out to a distance of 8000 meters. The light-shaded areas on the profile are those that are illuminated directly by the antenna on the tower, while the dark portions of the profile are those that cannot be seen directly from the tower. This profile assumes that the mobile station is 2 meters high—the height of a 6-foot high person with a handheld radio.

The terrain at an 80° azimuth allows direct radio view from the top of the tower out to about 1.8 km. From here, the downslope prevents direct view until about 2.5 km, where the terrain is briefly visible again from several hundred meters, disappearing from radio view until about 2.8 km, after which it becomes visible until about 3.6 km. Note that other than putting the repeater antenna on a higher tower, there is nothing that can be done to improve repeater coverage over this hilly terrain, although knife-edge diffraction off the hill tops will help fill in coverage gaps.

The Repeater Antenna

The most important part of the system is the antenna itself. As with any antenna, it must radiate and collect RF energy as efficiently as possible. Many repeaters use omnidirectional antennas, but this is not always the best choice. For example, suppose a group wishes to set up a repeater to cover towns A and B and the interconnecting state highway shown in **Fig 3**. The available repeater site is marked on the map. No coverage is required to the west or south, or over the ocean. If an omnidirectional antenna is used in this case, a significant amount of the radiated signal goes in undesired directions. By using an antenna with a cardioid pattern, as shown in Fig 1, the coverage is concentrated in the desired directions. The repeater will be more effective in these locations, and signals from low-power portables and mobiles will be more reliable.

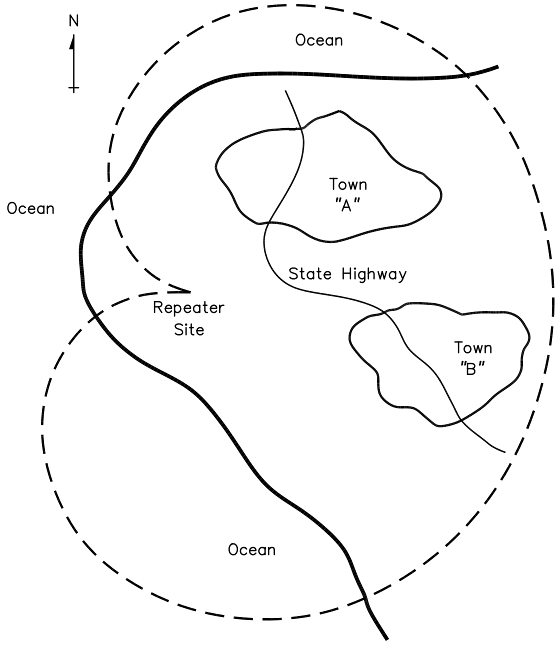

Fig 3—There are many situations where equal repeater coverage is not desired in all directions from the "machine." One such situation is shown here, where the repeater is needed to cover only towns A and B and the interconnecting highway. An omnidirectional antenna would provide coverage in undesired directions, such as over the ocean. The broken line shows the radiation pattern of an antenna that is better suited to this circumstance.

In many cases, antennas with special patterns are more expensive than omnidirectional models. This is an obvious consideration in designing a repeater antenna system. Over terrain where coverage may be difficult in some direction from the repeater site, it may be desirable to skew the antenna pattern in that direction. This can be accomplished by using a phased-vertical array or a com-

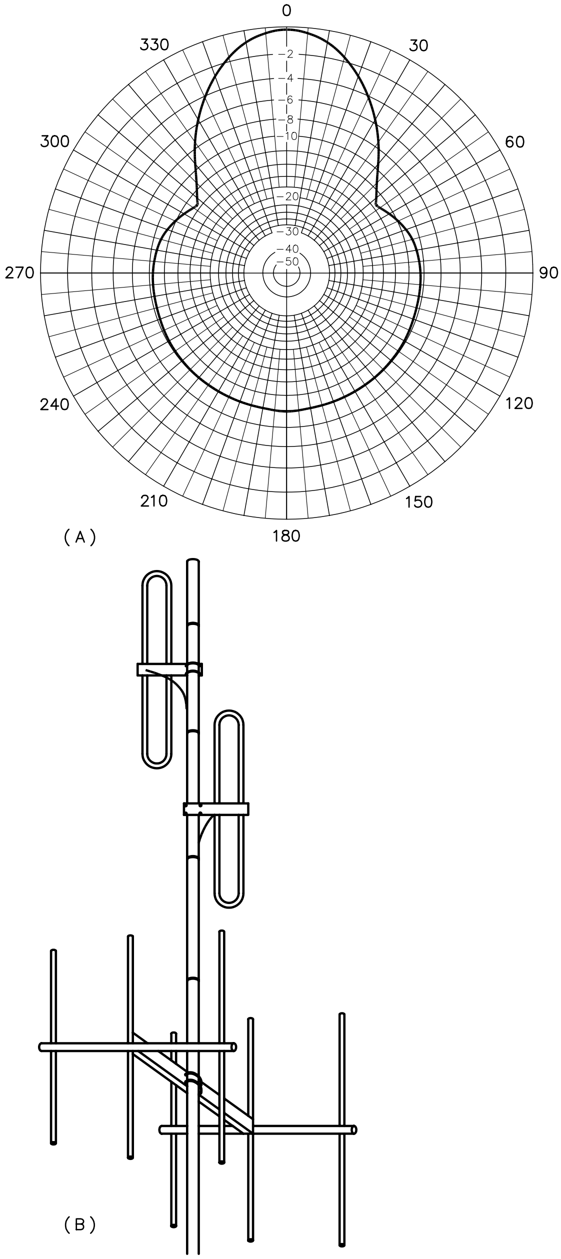

Fig 4—The "keyhole" horizontal radiation pattern at A is generated by the combination of phased Yagis and vertical elements shown at B. Such a pattern is useful in overcoming coverage blockages resulting from local terrain features. (*Based on a design by Decibel Products, Inc.*)

bination of a Yagi and a phased vertical to produce a "keyhole" pattern. See **Fig 4**.

Repeaters are common on 440 MHz and above, and many groups invest in high-gain omnidirectional antennas. A consequence of getting high gain from an omnidirectional antenna is vertical beamwidth reduction. In most cases, these antennas are designed to radiate their peak gain at the horizon, resulting in optimum coverage when the antenna is located at a moderate height over normal terrain. Unfortunately, in cases where the antenna is located at a very high site (overlooking the coverage area) this may not be the most desirable pattern. The vertical pattern of the antenna can be tilted downward, however, to facilitate coverage of the desired area. This is called *vertical-beam downtilt*.

An example of such a situation is shown in **Fig 5**. The repeater site overlooks a town in a valley. A 450-MHz repeater is needed to serve low-power portable and mobile stations. Constraints on the repeater dictate the use of an antenna with a gain of 11 dBi. (An omnidirec-

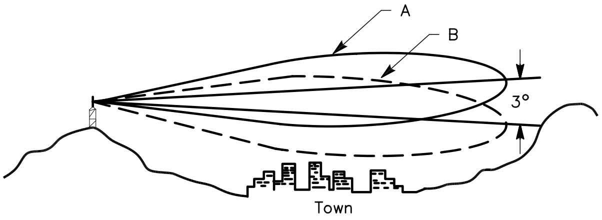

Fig 5—Vertical-beam downtilt is another form of radiation-pattern distortion useful for improving repeater coverage. This technique can be employed in situations where the repeater station is at a greater elevation than the desired coverage area, when a high-gain omnidirectional antenna is used. Pattern A shows the normal vertical-plane radiation pattern of a high-gain omnidirectional antenna with respect to the desired coverage area (the town). Pattern B shows the pattern tilted down, and the coverage improvement is evident.

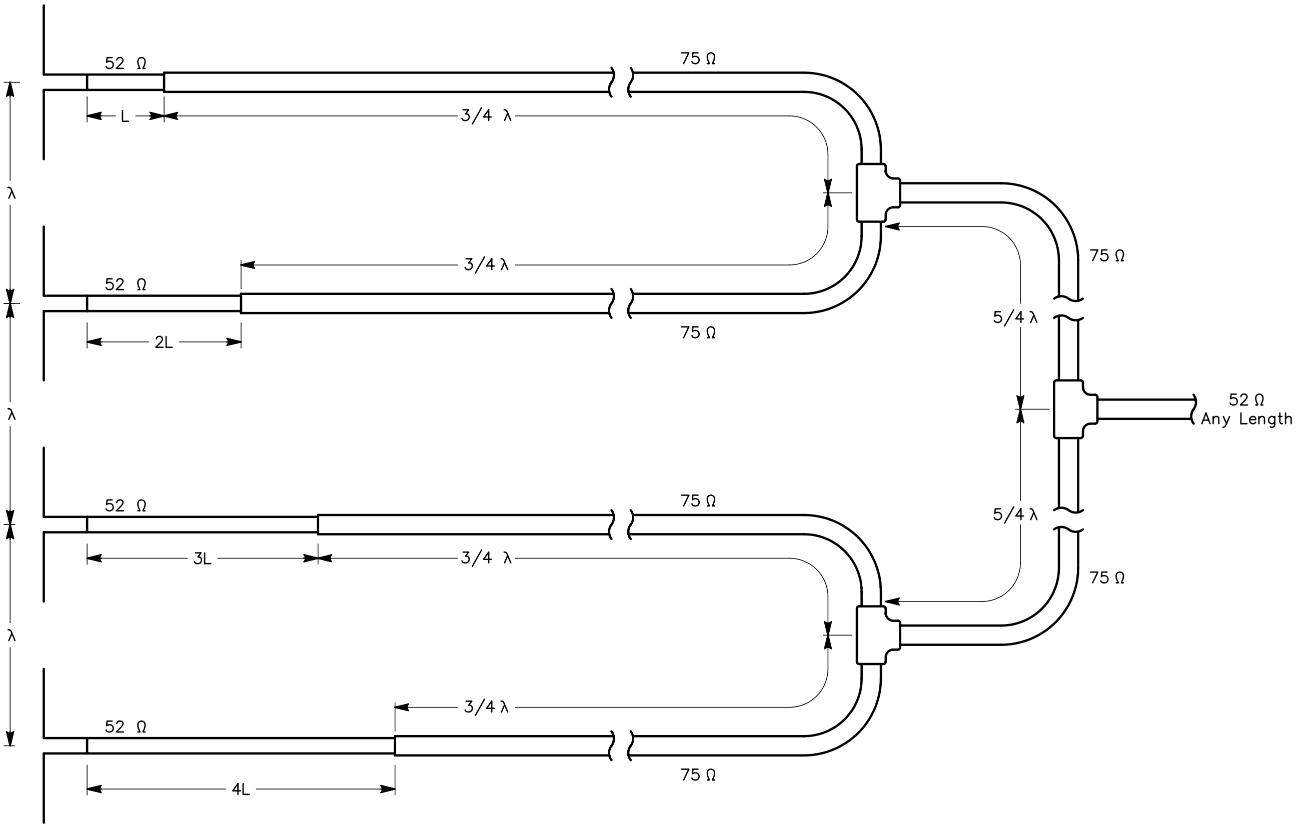

Fig 6—Vertical-beam downtilt can be facilitated by inserting 52-Ω delay lines in series with the 75-Ω feed lines to the collinear elements of an omnidirectional antenna. The delay lines to each element are progressively longer so the phase shift between elements is uniform. Odd $^1/_4$-λ coaxial transformers are used in the main (75-Ω) feed system to match the dipole impedances to the driving point. Tilting the vertical beam in this way often produces minor lobes in the vertical pattern that do not exist when the elements are fed in phase.

tional antenna with this gain has a vertical beamwidth of approximately 6°.) If the repeater antenna has its peak gain at the horizon, a major portion of the transmitted signal and the best area from which to access the repeater exists *above* the town. By tilting the pattern down 3°, the peak radiation will occur in the town.

Vertical-beam downtilt is generally produced by feeding the elements of a collinear vertical array slightly out of phase with each other. Lee Barrett, K7NM, showed such an array in *Ham Radio* magazine. (See the Bibliography at the end of this chapter.) Barrett gives the geometry and design of a four-pole array with progressive phase delay, and a computer program to model it. The technique is shown in **Fig 6**, with a free-space elevation plot showing downtilting in **Fig 7**.

Commercial antennas are sometimes available (at extra cost) with built-in downtilt characteristics. Before ordering such a commercial antenna, make sure that you really require it—they generally are special-order items and are not returnable.

There are disadvantages to improving coverage by means of vertical-beam downtilt. When compared to a standard collinear array, an antenna using vertical-beam downtilt will have somewhat greater extraneous lobes in the vertical pattern, resulting in reduced gain (usually less than 1 dB). Bandwidth is also slightly reduced. The reduction in gain, when combined with the downtilt characteristic, results in a reduction in total coverage area. These trade-offs, as well as the increased cost of a commercial antenna with downtilt, must be compared to the improvement in total performance in a situation where vertical-beam downtilt is contemplated.

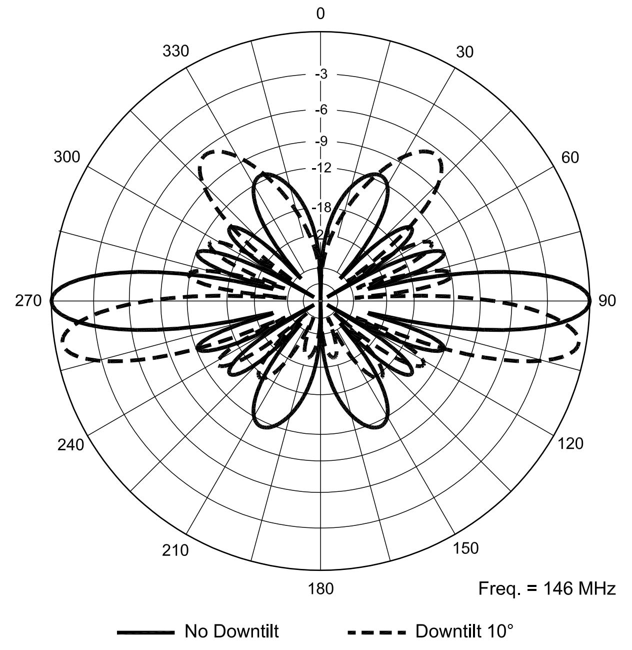

Fig 7—Free-space elevation-plane patterns showing downtilting that results from progressive phase shifts for the feed currents for the dipole in Fig 6.

Top Mounting and Side Mounting

Amateur repeaters often share towers with commercial and public service users. In many of these cases, other antennas are at the top of the tower, so the amateur antenna must be side mounted. A consequence of this arrangement is that the free-space pattern of the repeater antenna is distorted by the tower. This effect is especially noticeable when an omnidirectional antenna is side mounted on a structure.

The effects of supporting structures are most pronounced at close antenna spacings to the tower and with large support dimensions. The result is a measurable increase in gain in one direction and a partial null in the other direction (sometimes 15 dB deep). The shape of the supporting structure also influences pattern distortion. Many antenna manufacturers publish radiation patterns showing the effect of side mounting antennas in their catalogs.

Side mounting is not always a disadvantage. In cases where more (or less) coverage is desired in one direction, the supporting structure can be used to advantage. If pattern distortion is not acceptable, a solution is to mount antennas around the perimeter of the structure and feed them with the proper phasing to synthesize an omnidirectional pattern. Many manufacturers make antennas to accommodate such situations.

The effects of different mounting locations and arrangements can be illustrated with an array of exposed dipoles, **Fig 8**. Such an array is a very versatile antenna because, with simple rearrangement of the elements, it can develop either an omnidirectional pattern or an offset pattern. Drawing A of Fig 8 shows a basic collinear array of four vertical ½-λ elements. The vertical spacing between adjacent elements is 1 λ. All elements are fed in phase. If this array is placed in the clear and supported by a nonconducting mast, the calculated radiation resistance of each dipole element is on the order of 63 Ω. If the feed line is completely decoupled, the resulting azimuth pattern is omnidirectional. The vertical-plane pattern is shown in **Fig 9**.

Fig 8B shows the same array in a side-mounting arrangement, at a spacing of ¼ λ from a conducting mast. In this mounting arrangement, the mast takes on the role of a reflector, producing an F/B on the order of 5.7 dB. The azimuth pattern is shown in **Fig 10**. The vertical pattern is not significantly different from that of Fig 9, except the four small minor lobes (two on either side of the vertical axis) tend to become distorted. They are not as "clean," tending to merge into one minor lobe at some mast heights. This apparently is a function of currents in the supporting mast. The proximity of the mast also alters the feed-point impedance. For elements that are resonant in the configuration of Fig 8A, the calculated impedance in the arrangement of Fig 8B is in the order of 72 + *j* 10 Ω.

If side mounting is the only possibility and an omnidirectional pattern is required, the arrangement of Fig 8C may be used. The calculated azimuth pattern takes on a slight cloverleaf shape, but is within 1.5 dB of being circular. However, gain performance suffers, and the idealized vertical pattern of Fig 9 is not achieved. See **Fig 11**. Spacings other than $^1/_4$-λ from the mast were not investigated.

One very important consideration in side mounting an antenna is mechanical integrity. As with all repeater components, reliability is of great importance. An antenna hanging by the feed line and banging against the tower provides far from optimum performance and reliability. Use a mount that is appropriately secured to the tower and the antenna. Also use good hardware, preferably stainless steel (or bronze). If your local hardware store does not carry stainless steel hardware, try a boating supplier.

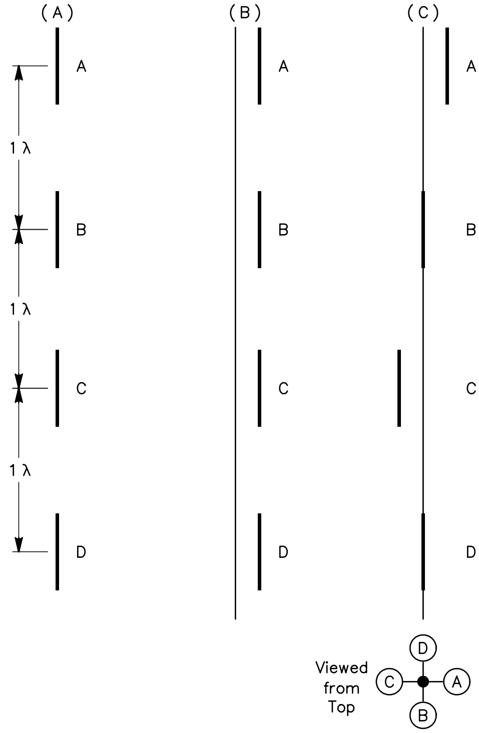

Fig 8—Various arrangements of exposed dipole elements. At A is the basic collinear array of four elements. B shows the same elements mounted on the side of a mast, and C shows the elements in a side-mounted arrangement around the mast for omnidirectional coverage. See text and Figs 9 through 11 for radiation-pattern information.

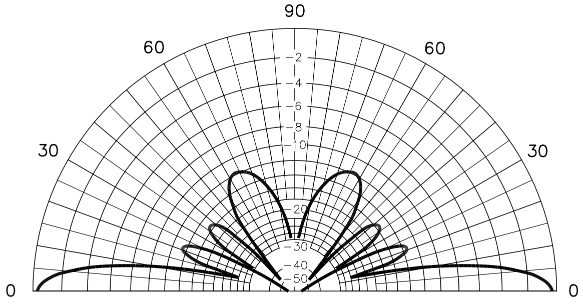

Fig 9—Calculated vertical-plane pattern of the array of Fig 8A, assuming a nonconducting mast support and complete decoupling of the feeder. In azimuth the array is omnidirectional. The calculated gain of the array is 8.6 dBi at 0° elevation; the –3 dB point is at 6.5°.

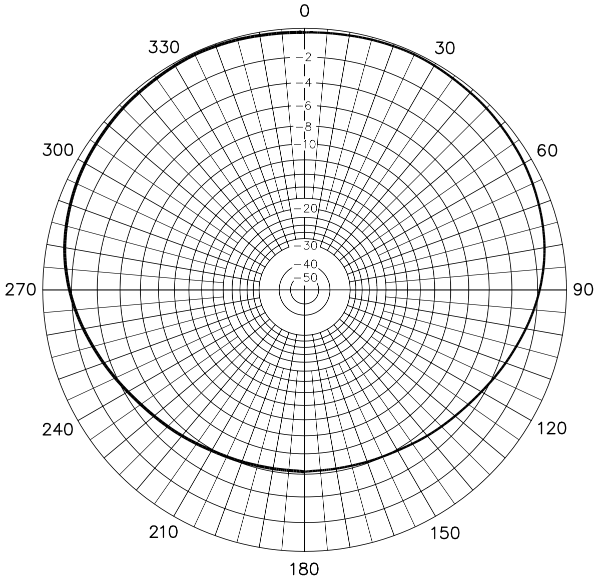

Fig 10—Calculated azimuth pattern of the side-mounted array of Fig 8B, assuming $^1/_4$-λ spacing from a 4-inch mast. The calculated gain in the favored direction, away from the mast and through the elements, is 10.6 dBi.

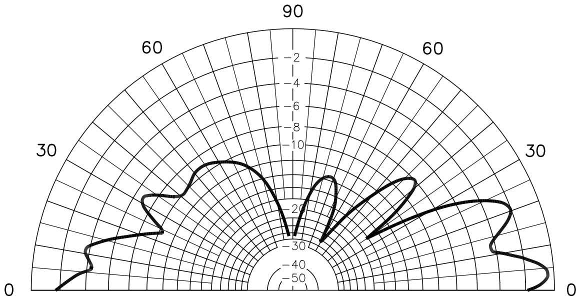

Fig 11—Calculated vertical pattern of the array of Fig 8C, assuming $^1/_4$-λ element spacing from a 4-inch mast. The azimuth pattern is circular within 1.5 dB, and the calculated gain is 4.4 dBi.

Be certain that the feed line is properly supported along its length. Long lengths of cable are subject to contraction and expansion with temperature from season to season, so it is important that the cable not be so tight that contraction causes it to stress the connection at the antenna. This can cause the connection to become intermittent (and noisy) or, at worst, an open circuit. This is far from a pleasant situation if the antenna connection is 300 feet up a tower, and it happens to be the middle of the winter!

Effects of Other Conductors

Feed-line proximity and tower-access ladders or cages also have an effect on the radiation patterns of side-mounted antennas. This subject was studied by Connolly and Blevins, and their findings are given in *IEEE Conference Proceedings* (see the Bibliography at the end of this chapter). Those considering mounting antennas on air-conditioning evaporators or maintenance penthouses on commercial buildings should consult this article. It gives considerable information on the effects of these structures on both unidirectional and omnidirectional antennas.

Metallic guy wires also affect antenna radiation patterns. Yang and Willis studied this and reported the results in *IRE Transactions on Vehicular Communications*. As expected, the closer the antenna is to the guy wires, the greater the effect on the radiation patterns. If the antennas are near the point where the guy wires meet the tower, the effect of the guy wires can be minimized by breaking them up with insulators every 0.75 λ for 2.25 λ to 3.0 λ.

ISOLATION REQUIREMENTS IN REPEATER ANTENNA SYSTEMS

Because repeaters generally operate in full *duplex* (the transmitter and receiver operate simultaneously), the antenna system must act as a filter to keep the transmitter from blocking the receiver. The degree to which the transmitter and receiver must be isolated is a complex problem. It is quite dependent on the equipment used and the difference in transmitter and receiver frequencies (offset). Instead of going into great detail, a simplified example can be used for illustration.

Consider the design of a 144-MHz repeater with a 600-kHz offset. The transmitter has an RF output power of 10 watts, and the receiver has a squelch sensitivity of 0.1 µV. This means there must be at least 1.9×10^{-16} watts at the 52-Ω receiver-antenna terminals to detect a signal. If both the transmitter and receiver were on the same frequency, the isolation (attenuation) required between the transmitter and receiver antenna jacks to keep the transmitter from activating the receiver would be

$$\text{Isolation} = 10 \log \frac{10 \text{ watts}}{1.9 \times 10^{-16} \text{ watts}} = 167 \text{ dB}$$

Obviously there is no need for this much attenuation, because the repeater does not transmit and receive on the same frequency.

If the 10-watt transmitter has noise 600 kHz away from the carrier frequency that is 45 dB below the carrier power, that 45 dB can be subtracted from the isolation requirement. Similarly, if the receiver can detect a 0.1 µV on-frequency signal in the presence of a signal 600 kHz away that is 40 dB greater than 0.1 µV, this 40 dB can also be subtracted from the isolation requirement. Therefore, the isolation requirement is

167 dB – 45 dB – 40 dB = 82 dB

Other factors enter into the isolation requirements as well. For example, if the transmitter power is increased by 10 dB (from 10 to 100 watts), this 10 dB must be added to the isolation requirement. Typical requirements for 144- and 440-MHz repeaters are shown in **Fig 12**.

Obtaining the required isolation is the first problem to be considered in constructing a repeater antenna system. There are three common ways to obtain this isolation:

1) Physically separate the receiving and transmitting antennas so the combination of path loss for the spacing and the antenna radiation patterns results in the required isolation.
2) Use a combination of separate antennas and high-Q filters to develop the required isolation. (The high-Q filters serve to reduce the physical distance required between antennas.)
3) Use a combination filter and combiner system to

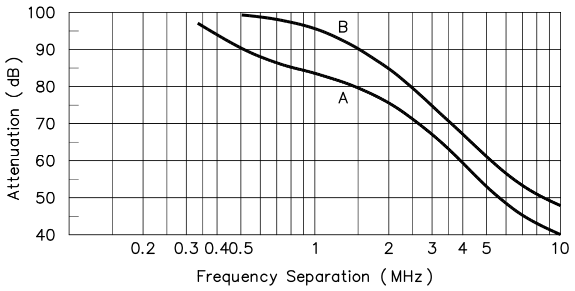

Fig 12—Typical isolation requirements for repeater transmitters and receivers operating in the 132-174 MHz band (Curve A), and the 400-512 MHz band (Curve B). Required isolation in dB is plotted against frequency separation in MHz. These curves were developed for a 100-W transmitter. For other power levels, the isolation requirements will differ by the change in decibels relative to 100 W. Isolation requirements will vary with receiver sensitivity. (The values plotted were calculated for transmitter-carrier and receiver-noise suppression necessary to prevent more than 1 dB degradation in receiver 12-dB SINAD sensitivity.)

allow the transmitter and receiver to share one antenna. Such a filter and combiner is called a *duplexer*.

Repeaters operating on 28 and 50 MHz generally use separate antennas to obtain the required isolation. This is largely because duplexers in this frequency range are both large and very expensive. It is generally less expensive to buy two antennas and link the sites by a committed phone line or an RF link than to purchase a duplexer. At 144 MHz and higher, duplexers are more commonly used. Duplexers are discussed in greater detail in a later section.

Separate Antennas

Receiver desensing (gain limiting caused by the presence of a strong off-frequency signal) can be reduced, and often eliminated, by separation of the transmitting and receiving antennas. Obtaining the 55 to 90 dB of isolation required for a repeater antenna system requires separate antennas to be spaced a considerable distance apart (in wavelengths).

Fig 13 shows the distances required to obtain specific values of isolation for vertical dipoles having horizontal separation (at A) and vertical separation (at B). The isolation gained by using separate antennas is subtracted from the total isolation requirement of the system. For example, if the transmitter and receiver antennas for a 450-MHz repeater are separated horizontally by 400 feet, the total isolation requirement in the system is reduced by about 64 dB.

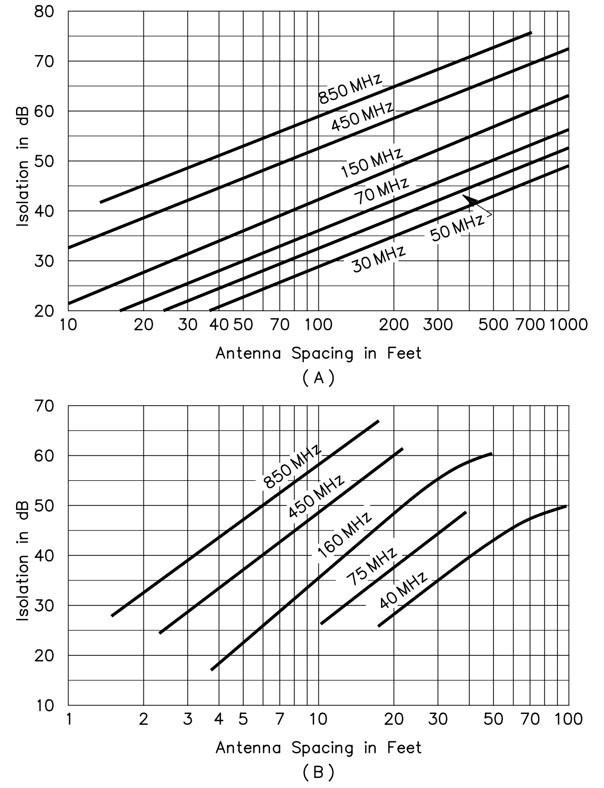

Fig 13—At A, the amount of attenuation (isolation) provided by horizontal separation of vertical dipole antennas. At B, isolation afforded by vertical separation of vertical dipoles.

Note from Fig 13B that a vertical separation of only about 25 feet also provides 64 dB of isolation. Vertical separation yields much more isolation than does horizontal separation. Vertical separation is also more practical than horizontal, since only a single support is required.

An explanation of the significant difference between the two graphs is in order. The vertical spacing requirement for 60 dB attenuation (isolation) at 150 MHz is about 43 feet. The horizontal spacing for the same isolation level is on the order of 700 feet. **Fig 14** shows why this difference exists. The radiation patterns of the antennas at A overlap; each antenna has gain in the direction of the other. The path loss between the antennas is given by

$$\text{Path loss (dB)} = 20 \log \frac{4 \pi d}{\lambda} \qquad \text{(Eq 1)}$$

where

d = distance between antennas
λ = wavelength, in the same units as d.

The isolation between the antennas in Fig 14A is the path loss less the antenna gains. Conversely, the antennas at B share pattern nulls, so the isolation is the path loss added to the depth of these nulls. This significantly reduces the spacing requirement for vertical sepa-

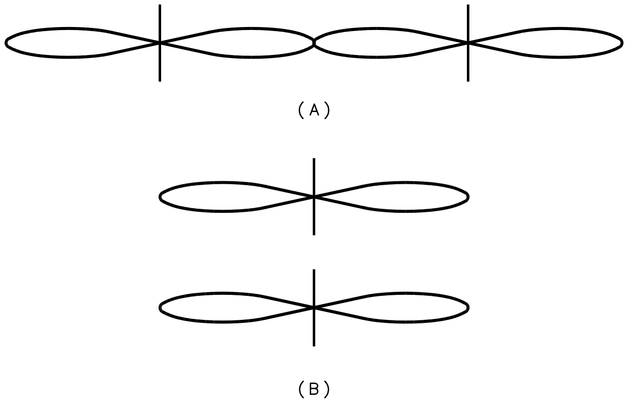

Fig 14—A relative representation of the isolation advantage afforded by separating antennas horizontally (A) and vertically (B) is shown. A great deal of isolation is provided by vertical separation, but horizontal separation requires two supports and much greater distance to be as effective. Separate-site repeaters (those with transmitter and receiver at different locations) benefit much more from horizontal separation than do single-site installations.

ration. Because the depth of the pattern nulls is not infinite, some spacing is required. Combined horizontal and vertical spacing is much more difficult to quantify because the results are dependent on both radiation patterns and the positions of the antennas relative to each other.

Separate antennas have one major disadvantage: They create disparity in transmitter and receiver coverage. For example, say a 50-MHz repeater is installed over average terrain with the transmitter and repeater separated by 2 miles. If both antennas had perfect omnidirectional coverage, the situation depicted in **Fig 15** would exist. In this case, stations able to hear the repeater may not be able to access it, and vice versa. In practice, the situation can be considerably worse. This is especially true if the patterns of both antennas are not omnidirectional. If this disparity in coverage cannot be tolerated, the solution involves skewing the patterns of the antennas until their coverage areas are essentially the same.

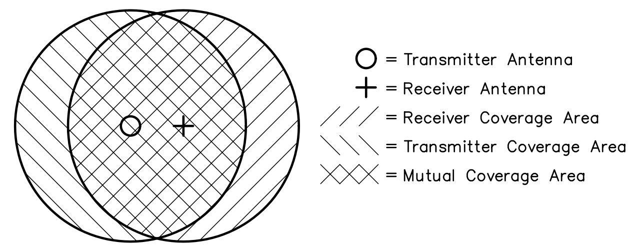
○ = Transmitter Antenna
+ = Receiver Antenna
/// = Receiver Coverage Area
\\\ = Transmitter Coverage Area
XXX = Mutual Coverage Area

Fig 15—Coverage disparity is a major problem for separate-site repeater antennas. The transmitter and receiver coverage areas overlap, but are not entirely mutually inclusive. Solving this problem requires a great deal of experimentation, as many factors are involved. Among these factors are terrain features and distortion of the antenna radiation patterns from supports.

Cavity Resonators

As just discussed, receiver desensing can be reduced by separating the transmitter and receiver antennas. But the amount of transmitted energy that reaches the receiver input must often be decreased even farther. Other nearby transmitters can cause desensing as well. A *cavity resonator* (cavity filter) can be helpful in solving these problems. When properly designed and constructed, this type of resonator has very high Q. A commercially made cavity is shown in **Fig 16**.

A cavity resonator placed in series with a transmission line acts as a band-pass filter. For a resonator to operate in series, it must have input and output coupling loops (or probes). A cavity resonator can also be connected across (in parallel with) a transmission line. The cavity then acts as a band-reject (notch) filter, greatly attenuating energy at the frequency to which it is tuned. Only one coupling loop or probe is required for this method of filtering. This type of cavity could be used in the receiver line to "notch" the transmitter signal. Several cavities can be connected in series or parallel to increase the attenuation in a given configuration. The graphs of **Fig 17** show the attenuation of a single cavity (A) and a pair of cavities (B).

The only situation in which cavity filters would not help is the case where the off-frequency noise of the transmitter was right on the receiver frequency. With cavity resonators, an important point to remember is that addition of a cavity across a transmission line may change the impedance of the system. This change can be compensated by adding tuning stubs along the transmission line.

Duplexers

The material in this section was prepared by Domenic Mallozzi, N1DM. Most amateur repeaters in the 144-, 220- and 440-MHz bands use duplexers to obtain the necessary transmitter to receiver isolation. Duplexers have been commonly used in commercial repeaters for many years. The duplexer consists of two high-Q filters. One filter is used in the feed line from the transmitter to the antenna, and another between the antenna and the receiver. These filters must have low loss at the frequency to which they are tuned while having very high attenuation at the surrounding frequencies. To meet the high attenuation requirements at frequencies within as little as 0.4% of the frequency to which they are tuned, the filters usually take the form of cascaded transmission-line cavity filters. These are either band-pass filters, or band-pass filters with a rejection notch. (The rejection notch is tuned to the center frequency of the other filter.) The number of cascaded filter sections is determined by the frequency separation and the ultimate attenuation requirements.

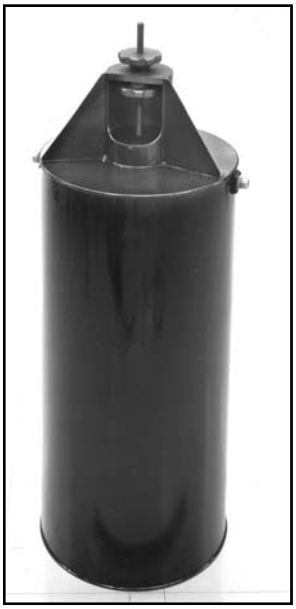

Fig 16—A coaxial cavity filter of the type used in many amateur and commercial repeater installations. Center-conductor length (and thus resonant frequency) is varied by adjustment of the knob (top).

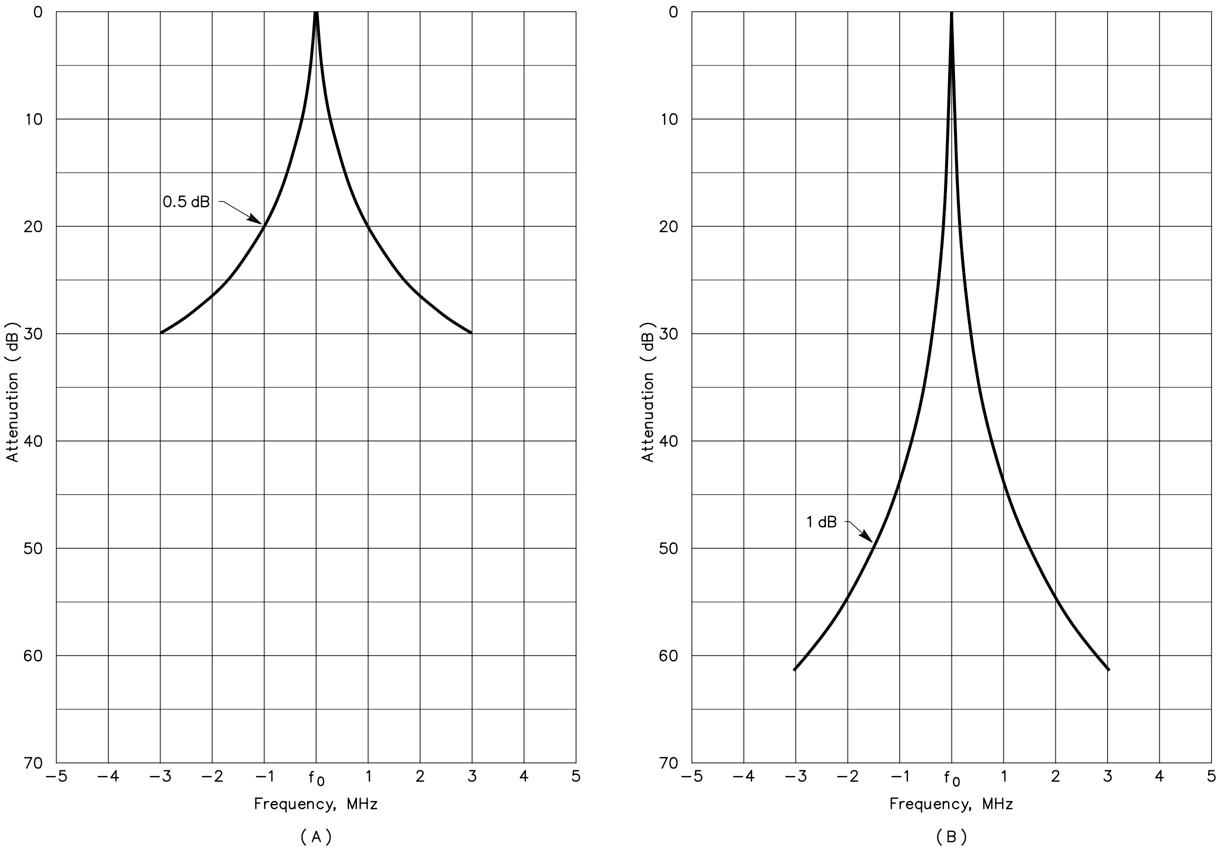

Fig 17—Frequency response curves for a single cavity (A) and two cavities cascaded (B). These curves are for cavities with coupling loops, each having an insertion loss of 0.5 dB. (The total insertion loss is indicated in the body of each graph.) Selectivity will be greater if lighter coupling (greater insertion loss) can be tolerated.

Duplexers for the amateur bands represent a significant technical challenge, because in most cases amateur repeaters operate with significantly less frequency separation than their commercial counterparts. Information on home construction of duplexers is presented in a later section of this chapter. Many manufacturers market high-quality duplexers for the amateur frequencies.

Duplexers consist of very high-Q cavities whose resonant frequencies are determined by mechanical components, in particular the tuning rod. **Fig 18** shows the cutaway view of a typical duplexer cavity. The rod is usually made of a material that has a limited thermal expansion coefficient (such as Invar). Detuning of the cavity by environmental changes introduces unwanted losses in the antenna system. An article by Arnold in *Mobile Radio Technology* considered the causes of drift in the

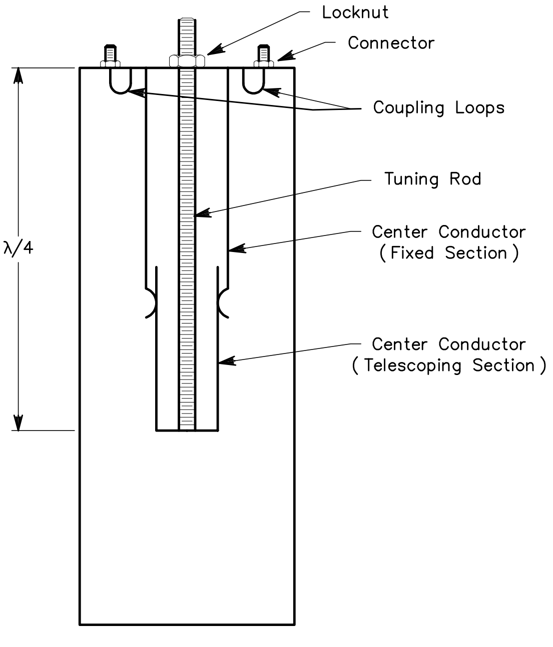

Fig 18—Cutaway view of a typical cavity. Note the relative locations of the coupling loops to each other and to the center conductor of the cavity. A locknut is used to prevent movement of the tuning rod after adjustment.

cavity (see the Bibliography at the end of this chapter). These can be broken into four major categories.

1) Ambient temperature variation (which leads to mechanical variations related to the thermal expansion coefficients of the materials used in the cavity).
2) Humidity (dielectric constant) variation.
3) Localized heating from the power dissipated in the cavity (resulting from its insertion loss).
4) Mechanical variations resulting from other factors (vibration, etc).

In addition, because of the high-Q nature of these cavities, the insertion loss of the duplexer increases when the signal is not at the peak of the filter response. This means, in practical terms, that less power is radiated for a given transmitter output power. Also, the drift in cavities in the receiver line results in increased system noise figure, reducing the sensitivity of the repeater.

As the frequency separation between the receiver and the transmitter decreases, the insertion loss of the duplexer reaches certain practical limits. At 144 MHz, the minimum insertion loss for 600 kHz spacing is 1.5 dB per filter.

Testing and using duplexers requires some special considerations (especially as frequency increases). Because duplexers are very high-Q devices, they are very sensitive to the termination impedances at their ports. A high SWR on any port is a serious problem, because the apparent insertion loss of the duplexer will increase, and the isolation may appear to decrease. Some have found that when duplexers are used at the limits of their isolation capabilities, a small change in antenna SWR is enough to cause receiver desensitization. This occurs most often under ice-loading conditions on antennas with open-wire phasing sections.

The choice of connectors in the duplexer system is important. BNC connectors are good for use below 300 MHz. Above 300 MHz, their use is discouraged because even though many types of BNC connectors work well up to 1 GHz, older style standard BNC connectors are inadequate at UHF and above. Type N connectors should be used above 300 MHz. It is false economy to use marginal quality connectors. Some commercial users have reported deteriorated isolation in commercial UHF repeaters when using such connectors. The location of a bad connector in a system is a complicated and frustrating process. Despite all these considerations, the duplexer is still the best method for obtaining isolation in the 144- to 925 MHz range.

ADVANCED TECHNIQUES

As the number of available antenna sites decreases and the cost of various peripheral items (such as coaxial cable) increases, amateur repeater groups are required to devise advanced techniques if repeaters are to remain effective. Some of the techniques discussed here have been applied in commercial services for many years, but until recently have not been economically justified for amateur use.

One technique worth consideration is the use of *cross-band couplers*. To illustrate a situation where a cross-band coupler would be useful, consider the following example. A repeater group plans to install 144- and 902-MHz repeaters on the same tower. The group intends to erect both antennas on a horizontal cross arm at the 325-foot level. A 325-foot run of $^{7}/_{8}$-inch Heliax costs approximately $2000. If both antennas are to be mounted at the top of the tower, the logical approach would require two separate feed lines. A better solution involves the use of a single feed line for both repeaters, along with a cross-band coupler at each end of the line.

The use of the cross-band coupler is shown in **Fig 19**. As the term implies, the coupler allows two signals on different bands to share a common transmission line. Such couplers cost approximately $300 each. In our hypothetical example, this represents a saving of $1400 over the cost of using separate feed lines. But, as with all compromises, there are disadvantages. Cross-band cou-

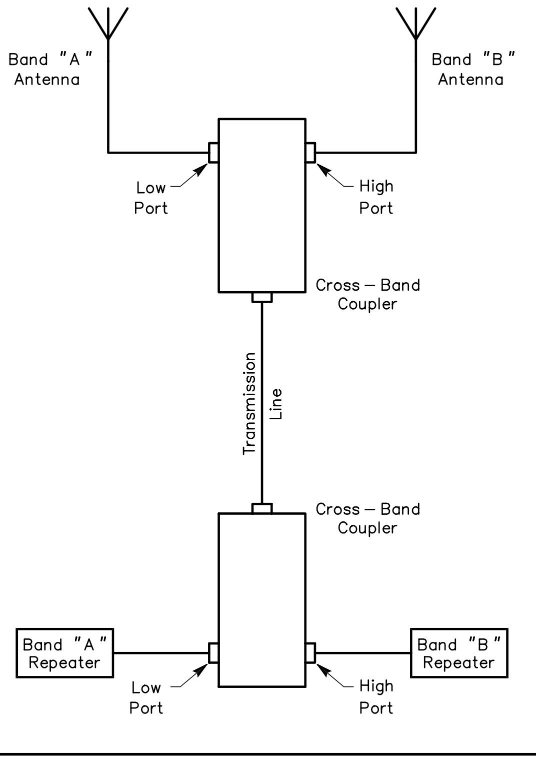

Fig 19—Block diagram of a system using cross-band couplers to allow the use of a single feed line for two repeaters. If the feeder to the antenna location is long (more than 200 feet or so), cross-band couplers may provide a significant saving over separate feed lines, especially at the higher amateur repeater frequencies. Cross-band couplers cannot be used with two repeaters on the same band.

plers have a loss of about 0.5 dB per unit. Therefore, the pair required represents a loss of 1.0 dB in *each* transmission path. If this loss can be tolerated, the cross-band coupler is a good solution.

Cross-band couplers do not allow two repeaters *on the same band* to share a single antenna and feed line. As repeater sites and tower space become more scarce, it may be desirable to have two repeaters on the same band share the same antenna. The solution to this problem is the use of a *transmitter multicoupler*. The multicoupler is related to the duplexers discussed earlier. It is a cavity filter and combiner that allows multiple transmitters and receivers to share the same antenna. This is a common commercial practice. A block diagram of a multicoupler system is shown in **Fig 20**.

The multicoupler, however, is a very expensive device, and has the disadvantage of even greater loss per transmission path than the standard duplexer. For example, a well-designed duplexer for 600 kHz spacing at 146 MHz has a loss per transmission path of approximately 1.5 dB. A four-channel multicoupler (the requirement for two repeaters) has an insertion loss per transmission path on the order of 2.5 dB or more. Another constraint of such a system is that the antenna must present a good match to the transmission line at all frequencies on which it will be used (both transmitting and receiving). This becomes difficult for the system with two repeaters operating at opposite ends of a band.

If you elect to purchase a commercial base-station antenna that requires you to specify a frequency to which the antenna must be tuned, be sure to indicate to the manufacturer the intended use of the antenna and the frequency extremes. In some cases, the only way the manufacturer can accommodate your request is to provide an antenna with some vertical-beam uptilt at one end of the band and some downtilt at the other end of the band. In the case of antennas with very high gain, this in itself may become a serious problem. Careful analysis of the situation is necessary before assembling such a system.

Diversity Techniques for Repeaters

Mobile flutter, "dead spots" and similar problems are a real problem for the mobile operator. The popularity of hand-held transceivers using low power and mediocre antennas causes similar problems. A solution to these difficulties is the use of some form of *diversity reception*. Diversity reception works because signals do not fade at the same rate when received by antennas at different locations (space diversity) or of different polarizations (polarization diversity).

Repeaters with large transmitter coverage areas often have difficulty "hearing" low power stations in peripheral areas or in dead spots. Space diversity is especially useful in such a situation. Space diversity utilizes separate receivers at different locations that are linked to the repeater. The repeater uses a circuit called a *voter* that determines which receiver has the best signal, and then selects the appropriate receiver from which to feed the repeater transmitter. This technique is helpful in urban areas where shadowing from large buildings and bridges causes problems. Space-diversity receiving, when properly executed, can give excellent results. But with the improvement come some disadvantages: added initial cost, maintenance costs, and the possibility of failure created by the extra equipment required. If installed and maintained carefully, problems are generally minimal.

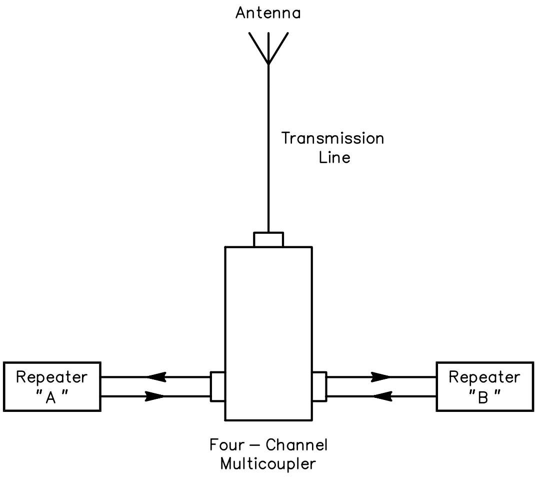

Fig 20—Block diagram of a system using a transmitter multicoupler to allow a single feed line and antenna to be used by two repeaters on one band. The antenna must be designed to operate at all frequencies that the repeaters utilize. More than two repeaters can be operated this way by using a multicoupler with the appropriate number of input ports.

A second improvement technique is the use of *circularly polarized* repeater antennas. This technique has been used in the FM broadcast field for many years, and has been considered for use in the mobile telephone service as well. Some experiments by amateurs have proved very promising, as discussed by Pasternak and Morris (see the Bibliography at the end of this chapter).

The improvement afforded by circular polarization is primarily a reduction in *mobile flutter*. The flutter on a mobile signal is caused by reflections from large buildings (in urban settings) or other terrain features. These reflections cause measurable polarization shifts, sometimes to the point where a vertically polarized signal at the transmitting site may appear to be primarily horizontally polarized after reflection.

A similar situation results from *multipath propagation*, where one or more reflected signals combine with the direct signal at the repeater, having varying effects on the signal. The multipath signal is subjected to large amplitude and phase variations at a relatively rapid rate.

In both of the situations described here, circular polarization can offer considerable improvement. This is because circularly polarized antennas respond equally to all linearly polarized signals, regardless of the plane of polarization. At this writing, there are no known sources of commercial circularly polarized omnidirectional antennas for the amateur bands. Pasternak and Morris describe a circularly polarized antenna made by modifying two commercial four-pole arrays.

EFFECTIVE ISOTROPIC RADIATED POWER (EIRP)

It is useful to know effective isotropic radiated power (EIRP) in calculating the coverage area of a repeater. The FCC formerly required EIRP to be entered in the log of every amateur repeater station. Although logging EIRP is no longer required, it is still useful to have this information on hand for repeater-coordination purposes and so system performance can be monitored periodically.

Calculation of EIRP is straightforward. The PEP output of the transmitter is simply multiplied by the gains and losses in the transmitting antenna system. (These gains and losses are best added or subtracted in decibels and then converted to a multiplying factor.) The following worksheet and example illustrates the calculations.

Feed-line loss ______ dB

Duplexer loss ______ dB

Isolator loss ______ dB

Cross-band coupler loss ______ dB

Cavity filter loss ______ dB

Total losses (L) ______ dB

G (dB) = antenna gain (dBi) –L

where G = antenna system gain. (If antenna gain is specified in dBd, add 2.14 dB to obtain the gain in dBi.)

$M = 10^{G/10}$

where M = multiplying factor

EIRP (watts) = transmitter output (PEP)×M

Example

A repeater transmitter has a power output of 50 W PEP (50-W FM transmitter). The transmission line has 1.8 dB loss. The duplexer used has a loss of 1.5 dB, and a circulator on the transmitter port has a loss of 0.3 dB. There are no cavity filters or cross-band couplers in the system. Antenna gain is 5.6 dBi.

Feed-line loss	1.8 dB
Duplexer loss	1.5 dB
Isolator loss	0.3 dB
Cross-band coupler loss	0 dB
Cavity filter loss	0 dB
Total losses (L)	3.6 dB

Antenna system gain in dB = G = antenna gain (dBi) –L

G = 5.6 dBi – 3.6 dB = 2 dB

Multiplying factor = $M = 10^{G/10}$

$M = 10^{2/10} = 1.585$

EIRP (watts) = transmitter output (PEP) × M

EIRP = 50 W × 1.585 = 79.25 W

If the antenna system is lossier than this example, G may be *negative*, resulting in a multiplying factor less than one. The result is an EIRP that is less than the transmitter output power. This situation can occur in practice, but for obvious reasons is not desirable.

Assembling a Repeater Antenna System

This section will aid you in planning and assembling your repeater antenna system. The material was prepared by Domenic Mallozzi, N1DM. Consult Chapter 23, Radio Wave Propagation, for information on propagation for the band of your interest.

First, a repeater antenna selection checklist such as this will help you in evaluating the antenna system for your needs.

Gain needed ________ dBi
Pattern required ________ Omnidirectional
________ Offset
________ Cardioidal
________ Bidirectional
________ Special pattern (specify)
____
Mounting ________ Top of tower
________ Side of tower

(Determine effects of tower on pattern. Is the result consistent with the pattern required?)

Is downtilt required? ________ Yes
________ No
Type of RF connector ________ UHF
________ N
________ BNC
________ Other (specify)
Size (length) ________
Weight ________
Maximum cost $_______

Table 1 (see next page) has been compiled to provide general information on commercial components available for repeater and remote-base antenna systems. The various components are listed in a matrix format by manufacturer, for equipment designed to operate in the various amateur bands. See Chapter 21, Antenna Products Suppliers, for further information for these components. Although every effort has been made to make this data complete, the ARRL is not responsible for omissions or errors. The listing of a product in Table 1 does not constitute an endorsement by ARRL. Manufacturers are urged to contact the editors with updating information.

Even though almost any antenna can be used for a repeater, the companies indicated in the *Antennas* column in Table 1 are known to have produced heavy-duty antennas to commercial standards for repeater service. Many of these companies offer their antennas with special features for repeater service (such as vertical-beam downtilt). It is best to obtain catalogs of current products from the manufacturers listed, both for general information and to determine which special options are available on their products.

A 144 MHz Duplexer

Obtaining sufficient isolation between the transmitter and receiver of a repeater can be difficult. Many of the solutions to this problem compromise receiver sensitivity or transmitter power output. Other solutions create an imbalance between receiver and transmitter coverage areas. When a duplexer is used, insertion loss is the compromise. But a small amount of insertion loss is more than offset by the use of one antenna for both the transmitter and receiver. Using one antenna assures equal antenna patterns for both transmitting and receiving, and reduces cost, maintenance and mechanical complexity.

As mentioned earlier in this chapter, duplexers may be built in the home workshop. Bob Shriner, WA0UZO, presented a small, mechanically simple duplexer for low-power applications in April 1979 *QST*. Shriner's design is unique, as the duplexer cavities are constructed of circuit-board material. Low cost and simplicity are the result, but with a trade-off in performance. A silver-plated version of Shriner's design has an insertion loss of approximately 5 dB at 146 MHz. The loss is greater if the copper is not plated, and increases as the inner walls of the cavities tarnish.

This duplexer construction project by John Bilodeau, W1GAN, represents an effective duplexer. The information originally appeared in July 1972 *QST*. It is a time-proven project used by many repeater groups, and can be duplicated relatively easily. Its insertion loss is just 1.5 dB.

Fig 21 will help you visualize the requirements for a duplexer, which can be summed up as follows. The duplexer must attenuate the transmitter carrier to avoid overloading the receiver and thereby reducing its sensitivity. It must also attenuate any noise or spurious frequencies from the transmitter on or near the receiver frequency. In addition, a duplexer must provide a proper impedance match between transmitter, antenna, and receiver.

Table 1

Product Matrix Showing Repeater Equipment and Manufacturer by Frequency Band

	Antennas							*Duplexers*				*Cavity Filters*			
Source	*28*	*50*	*144*	*220*	*450*	*902*	*1296*	*144*	*220*	*450*	*902*	*144*	*220*	*450*	*902*
Austin	S	S	S	S	S	S	S								
Celwave	C	C	C	C	C	C		C	C	C	C	C	C	C	C
Comet															
Cushcraft		C	C	C	C										
Dec Prod		C	C	C	C	C		C	C		C	C		C	
MA/COM												C		C	
RF Parts			C		C	C	C								
Sinclair	C	C	C					C		C		C		C	C
TX/RX								C	C	C		C		C	C
Wacom															

	Isolators/Circulators						*Transmitter Combiners*				*Cross-Band Couplers*			
Source	*28*	*50*	*144*	*220*	*450*	*902*	*144*	*220*	*450*	*902*	*0-174* *450-512*	*0-512* *800-960*	*59-174* *806-960*	*406-512* *806-960*
Celwave			C	C	C	C	C	C	C	C		S		
Dec Prod			C		C	C		C	C	C				
Sinclair			C		C	C								
TX/RX			C		C	C	C	C	C	C	C	C	C	C
Wacom			C	C	C	C	C	C	C	C		C		C

Abbreviated names above are for the following manufacturers: Austin Antennas, Celwave RF Inc, Cushcraft Corp, Decibel Products Inc, RF Parts, Sinclair Radio Laboratories Inc, TX/RX Systems Inc and Wacom Inc. A manufacturer's contact list appears in Chapter 21.

Key to codes used:
C = catalog (standard) item
S = special-order item

Note: Coaxial cable is not listed, because most manufacturers sell only to dealers.

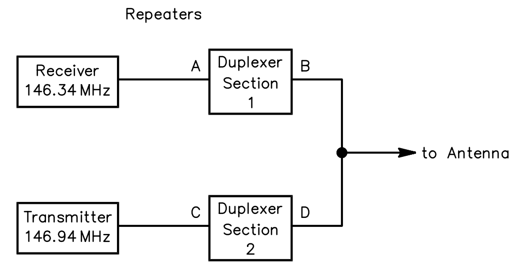

Fig 21—Duplexers permit using one antenna for both transmitting and receiving in a repeater system. Section 1 prevents energy at the transmitter frequency from interfering with the receiver, while section 2 attenuates any off-frequency transmitter energy that is at or near the receiver frequency.

As shown in Fig 21, transmitter output on 146.94 MHz going from point C to D should not be attenuated. However, the transmitter energy should be greatly attenuated between points B and A. Duplexer section 2 should attenuate any noise or signals that are on or near the receiver input frequency of 146.34 MHz. For good reception the noise and spurious signal level must be less than –130 dBm (0 dBm = 1 milliwatt into 50 Ω). Typical transmitter noise 600 kHz away from the carrier frequency is 80 dB below the transmitter power output. For 60 watts of output (+48 dBm), the noise level is –32 dBm. The duplexer must make up the difference between –32 dBm and –130 dBm, or 98 dB.

The received signal must go from point B to A with a minimum of attenuation. Section 1 of the duplexer must also provide enough attenuation of the transmitter energy to prevent receiver overload. For an average receiver, the transmitter signal must be less than –30 dBm to meet this requirement. The difference between the transmitter output of +48 dBm and the receiver overload point of –30 dBm, 78 dB, must be made up by duplexer section 1.

THE CIRCUIT

Fig 22 shows the completed 6-cavity duplexer, and **Fig 23** shows the assembly of an individual cavity. A $^1/_4$-λ resonator was selected for this duplexer design. The length of the center conductor is adjusted by turning a

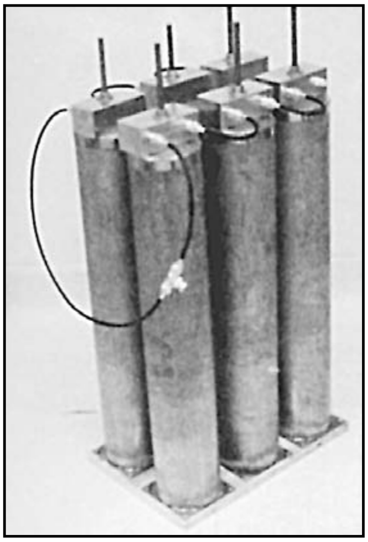

Fig 22—A six-cavity duplexer for use with a 144-MHz repeater. The cavities are fastened to a plywood base for mechanical stability. Short lengths of double-shielded cable are used for connections between individual cavities. An insertion loss of less than 1.5 dB is possible with this design.

threaded rod, which changes the resonant frequency of the cavity. Energy is coupled into and out of the tuned circuit by the coupling loops extending through the top plate.

The cavity functions as a series resonant circuit. When a reactance is connected across a series resonant circuit, an anti-resonant notch is produced, and the resonant frequency is shifted. If a capacitor is added, the notch appears below the resonant frequency. Adding inductance instead of capacitance makes the notch appear above the resonant frequency. The value of the added component determines the spacing between the notch and the resonant frequency of the cavity.

Fig 24 shows the measured band-pass characteristics of the cavity with shunt elements. With the cavity tuned to 146.94 MHz and a shunt capacitor connected from input to output, a 146.34-MHz signal is attenuated

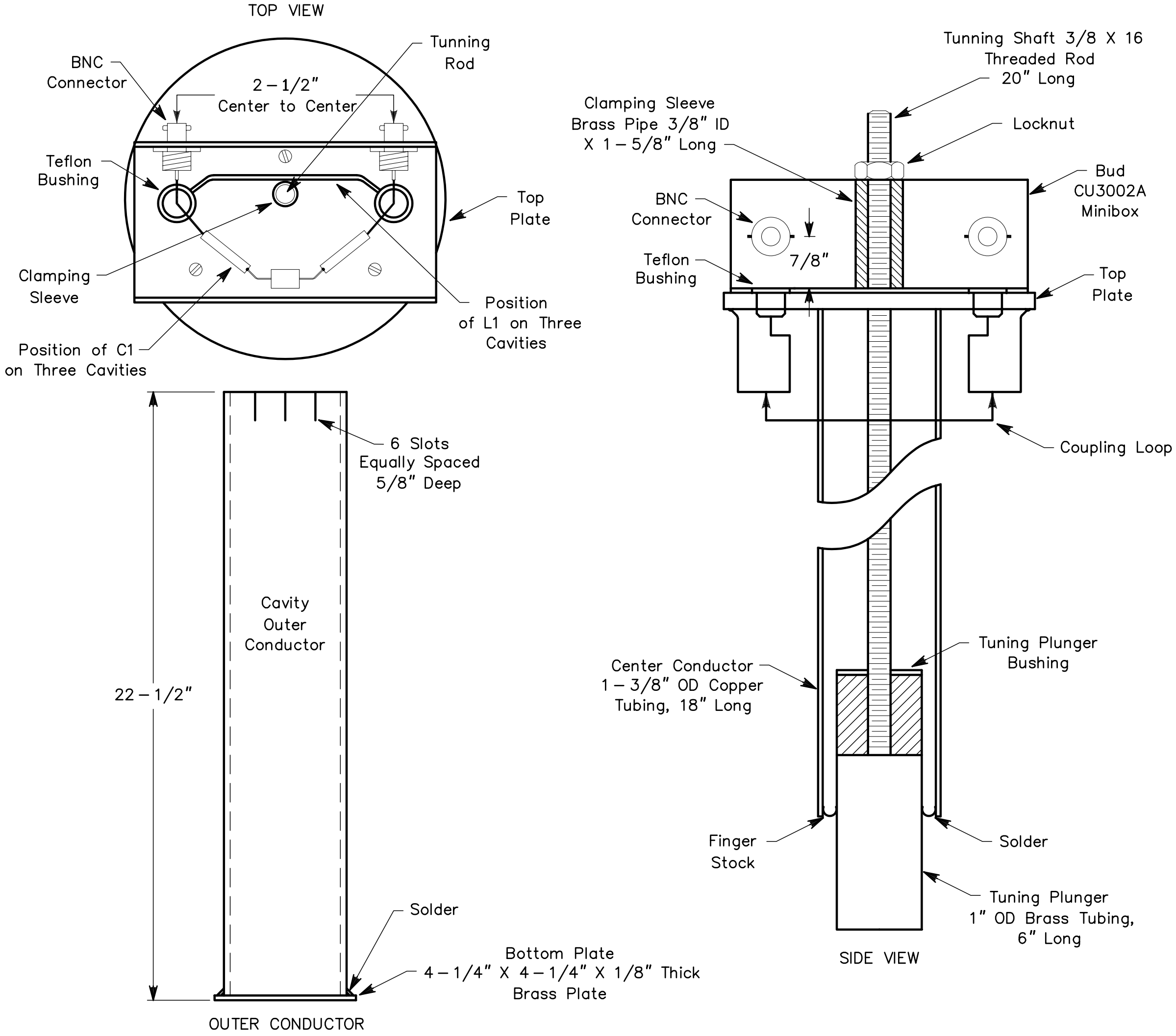
TOP VIEW
BNC Connector
2 – 1/2″ Center to Center
Tunning Rod
Teflon Bushing
Top Plate
Clamping Sleeve
Position of C1 on Three Cavities
Position of L1 on Three Cavities
6 Slots Equally Spaced 5/8″ Deep
Cavity Outer Conductor
22 – 1/2″
Solder
Bottom Plate 4 – 1/4″ X 4 – 1/4″ X 1/8″ Thick Brass Plate
OUTER CONDUCTOR
Tunning Shaft 3/8 X 16 Threaded Rod 20″ Long
Locknut
Clamping Sleeve Brass Pipe 3/8″ ID X 1 – 5/8″ Long
Bud CU3002A Minibox
BNC Connector
7/8″
Teflon Bushing
Top Plate
Coupling Loop
Center Conductor 1 – 3/8″ OD Copper Tubing, 18″ Long
Tuning Plunger Bushing
Finger Stock
Solder
Tuning Plunger 1″ OD Brass Tubing, 6″ Long
SIDE VIEW

Fig 23—The assembly of an individual cavity. A Bud Minibox is mounted on the top plate with three screws. A clamping sleeve made of brass pipe is used to prevent crushing the box when the locknut is tightened on the tuning shaft. Note that the positions of both C1 and L1 are shown, but that three cavities will have C1 installed and three will have L1 in place.

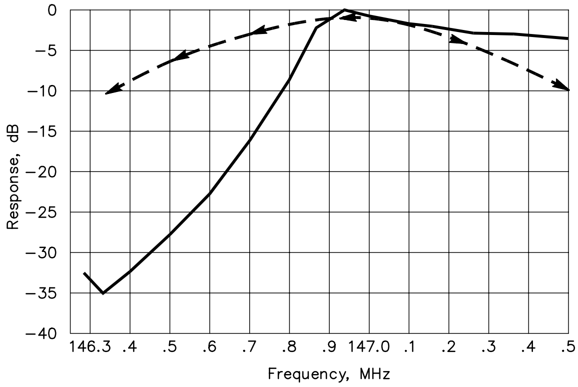

Fig 24—Typical frequency response of a single cavity of the type used in the duplexer. The dotted line represents the passband characteristics of the cavity alone; the solid line for the cavity with a shunt capacitor connected between input and output. An inductance connected in the same manner will cause the rejection notch to be above the frequency to which the cavity is tuned.

by 35 dB. If an inductance is placed across the cavity and the cavity is tuned to 146.34 MHz, the attenuation at 146.94 MHz is 35 dB. Insertion loss in both cases is 0.4 dB. Three cavities with shunt capacitors are tuned to 146.94 MHz and connected together in cascade with short lengths of coaxial cable. The attenuation at 146.34 MHz is more than 100 dB, and insertion loss at 146.94 MHz is 1.5 dB. Response curves for a six-cavity duplexer are given in **Fig 25**.

Construction

The schematic diagram for the duplexer is shown in **Fig 26**. Three parts for the duplexer must be machined; all others can be made with hand tools. A small lathe can be used to machine the brass top plate, the threaded tuning plunger bushing and the Teflon insulator bushing. The dimensions of these parts are given in **Fig 27**.

Type DWV copper tubing is used for the outer conductor of the cavities. The wall thickness is 0.058 inch, with an outside diameter of $4^{1}/_{8}$ inches. You will need a tubing cutter large enough to handle this size (perhaps borrowed or rented). The wheel of the cutter should be tight and sharp. Make slow, careful cuts so the ends will be square. The outer conductor is $22^{1}/_{2}$ inches long.

The inner conductor is made from type M copper tubing having an outside diameter of $1^{3}/_{8}$ inches. A 6-inch length of 1-inch OD brass tubing is used to make the tuning plunger.

The tubing types mentioned above are designations used in the plumbing and steam-fitting industry. Other types may be used in the construction of a duplexer, but you should check the sizes carefully to assure that the parts will fit each other. A greater wall thickness will make the assembly heavier, and the expense will increase accordingly. Soft solder is used throughout the assembly. Unless you have experience with silver solder, do not use it. Eutectic type 157 solder with paste or acid

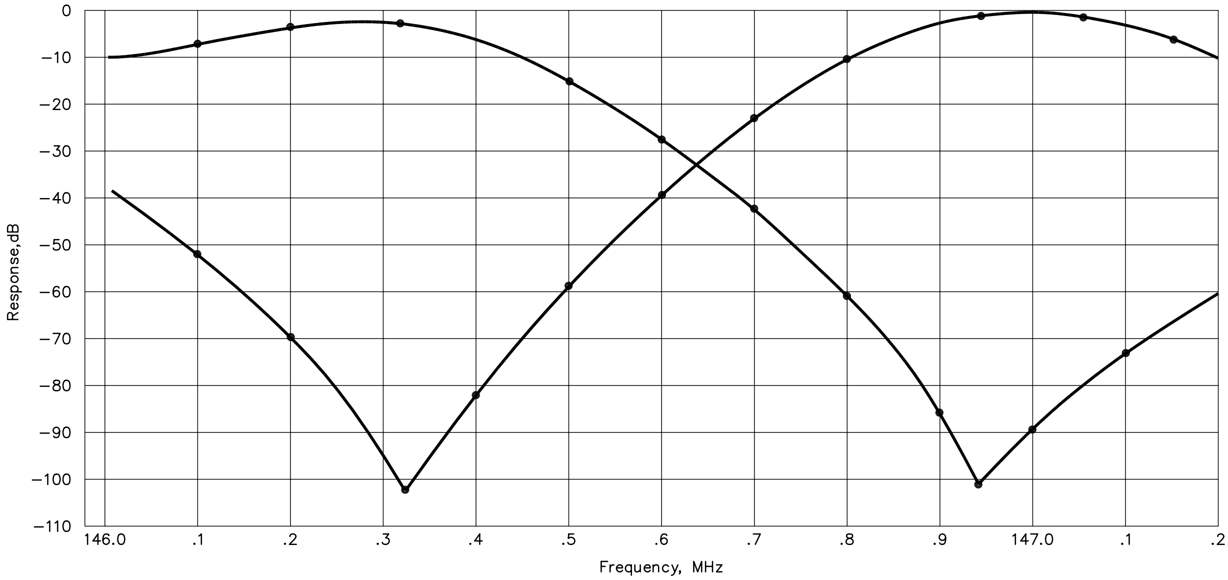

Fig 25—Frequency response of the six-cavity duplexer. One set of three cavities is tuned to pass 146.34 MHz and notch 146.94 MHz (the receiver leg). The remaining set of three cavities is tuned to pass 146.94 MHz and notch 146.34 MHz. This duplexer provides approximately 100 dB of isolation between the transmitter and receiver when properly tuned.

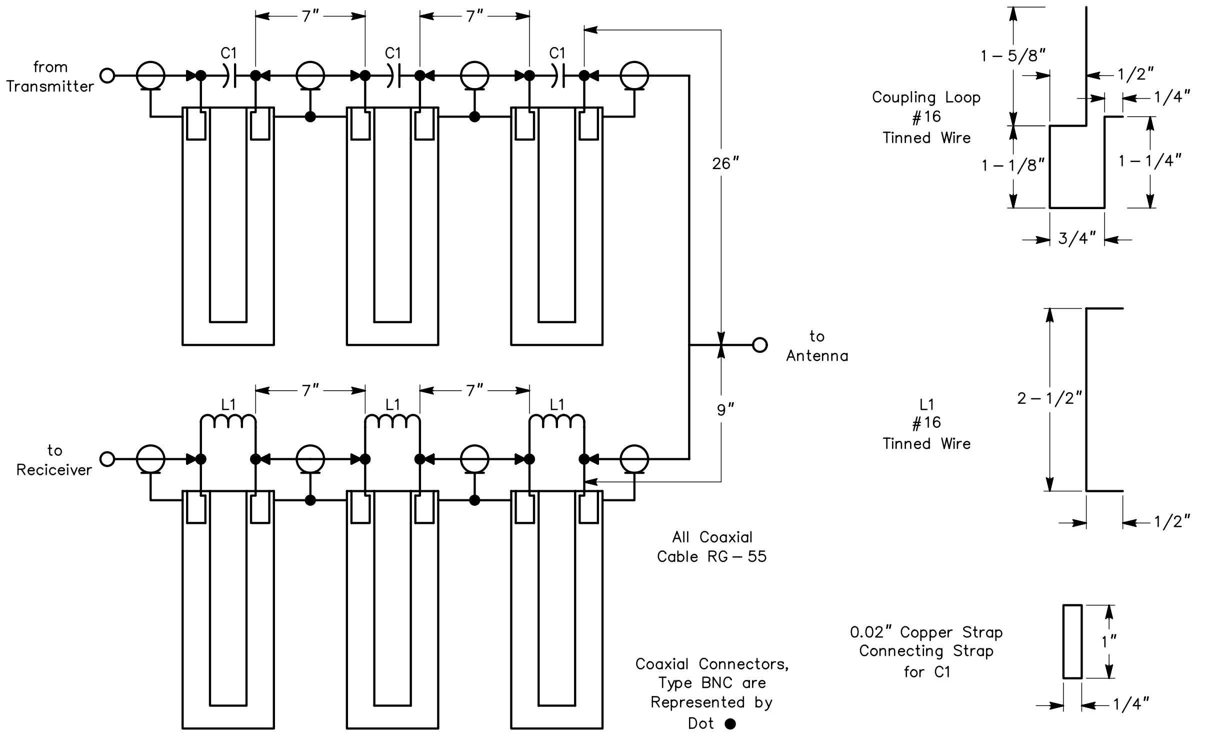
from Transmitter
7"
7"
C1
C1
C1
26"
to Antenna
9"
L1
L1
L1
7"
7"
to Reciceiver
All Coaxial Cable RG – 55
Coaxial Connectors, Type BNC are Represented by Dot ●
Coupling Loop #16 Tinned Wire
1 – 5/8"
1/2"
1/4"
1 – 1/8"
1 – 1/4"
3/4"
L1 #16 Tinned Wire
2 – 1/2"
1/2"
0.02" Copper Strap Connecting Strap for C1
1"
1/4"

Fig 26—Diagram of the six-cavity duplexer. Coaxial cable lengths between cavities are critical and must be followed closely. Double shielded cable and high quality connectors should be used throughout. The sizes and shapes of the coupling loops, L1, and the straps for connecting C1 should be observed.
C1—1.7-11 pF circuit-board mount, E. F. Johnson 189-5-5 or equiv. Set at $^3/_4$ closed for initial alignment.

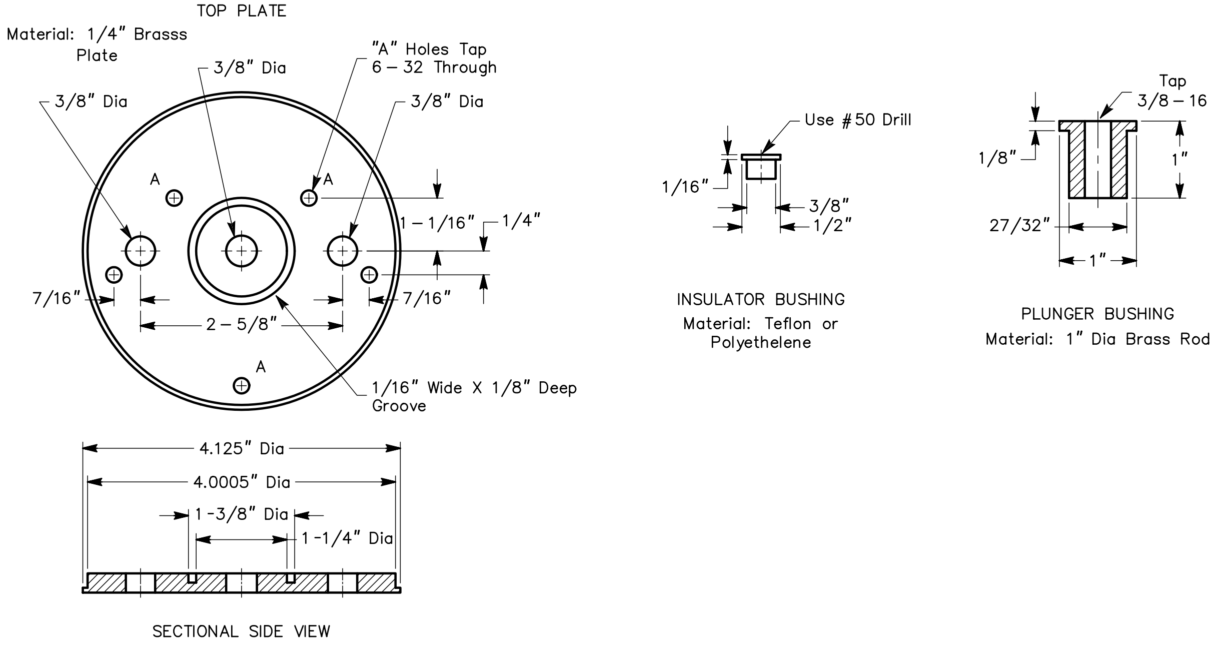
TOP PLATE
Material: 1/4" Brasss Plate
3/8" Dia
"A" Holes Tap 6 – 32 Through
3/8" Dia
3/8" Dia
A
A
A
1 – 1/16"
1/4"
7/16"
7/16"
2 – 5/8"
1/16" Wide X 1/8" Deep Groove
4.125" Dia
4.0005" Dia
1 -3/8" Dia
1 -1/4" Dia
SECTIONAL SIDE VIEW
Use #50 Drill
1/16"
3/8"
1/2"
INSULATOR BUSHING
Material: Teflon or Polyethelene
Tap 3/8 – 16
1/8"
1"
27/32"
1"
PLUNGER BUSHING
Material: 1" Dia Brass Rod

Fig 27—Dimensions for the three parts that require machining. A small metal-working lathe should be used for making these parts.

flux makes very good joints. This type has a slightly higher melting temperature than ordinary tin-lead alloy, but has considerably greater strength.

First solder the inner conductor to the top plate (**Fig 28**). The finger stock can then be soldered inside the lower end of the inner conductor, while temporarily held in place with a plug made of aluminum or stainless steel. While soldering, do not allow the flame from the torch to overheat the finger stock. The plunger bushing is soldered into the tuning plunger and a 20-inch length of threaded rod is soldered into the bushing.

Cut six slots in the top of the outer conductor. They should be $^{5}/_{8}$ inch deep and equally spaced around the tubing. The bottom end of the 4-inch tubing is soldered to the square bottom plate. The bottom plates have holes in the corners so they can be fastened to a plywood base by means of wood screws. Because the center conductor has no support at one end, the cavities must be mounted vertically.

The size and position of the coupling loops are critical. Follow the given dimensions closely. Both loops should be $^{1}/_{8}$ inch away from the center conductor on opposite sides. Connect a solder lug to the ground end of the loop, then fasten the lug to the top plate with a screw. The free end of the loop is insulated by Teflon bushings where it passes through the top plate for connection to the BNC fittings.

Before final assembly of the parts, clean them thoroughly. Soap-filled steel wool pads and hot water work well for this. Be sure the finger stock makes firm contact with the tuning plunger. The top plate should fit snugly in the top of the outer conductor—a large hose clamp tightened around the outer conductor will keep the top plate in place.

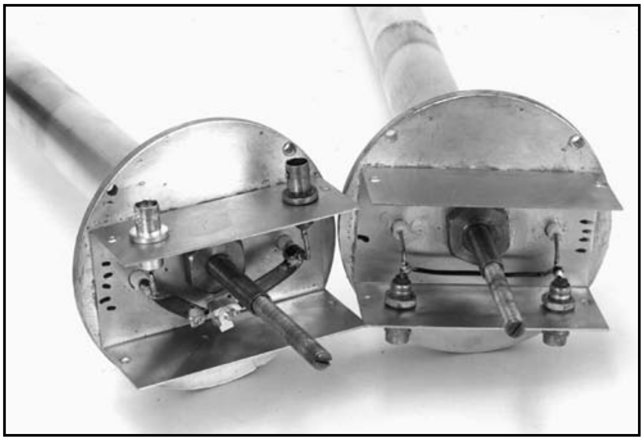

Fig 28—Two of the center conductor and top plate assemblies. In the assembly at the left, C1 is visible just below the tuning shaft, mounted by short straps made from sheet copper. The assembly on the right has L1 in place between the BNC connectors. The Miniboxes are fastened to the top plate by a single large nut in these units. Using screws through the Minibox into the top plate, as described in the text, is preferred.

ADJUSTMENT

After the cavities have been checked for band-pass characteristics and insertion loss, install the anti-resonant elements, C1 and L1. (See Fig 24.) It is preferable to use laboratory test equipment when tuning the duplexer. An option is to use a low-power transmitter with an RF probe and an electronic voltmeter. Both methods are shown in **Fig 29**.

With the test equipment connected as shown in Fig 29A, adjust the signal generator frequency to the desired repeater input frequency. Connect a calibrated step attenuator between points X and Y. With no attenuation, adjust the HP-415 for 0 on the 20-dB scale. You can check the calibration of the 415 by switching in different amounts of attenuation and noting the meter reading. You may note a small error at either high or very low signal levels.

Next, remove the step attenuator and replace it with a cavity that has the shunt inductor, L1, in place. Adjust the tuning screw for maximum reading on the 415 meter. Remove the cavity and connect points X and Y. Set the

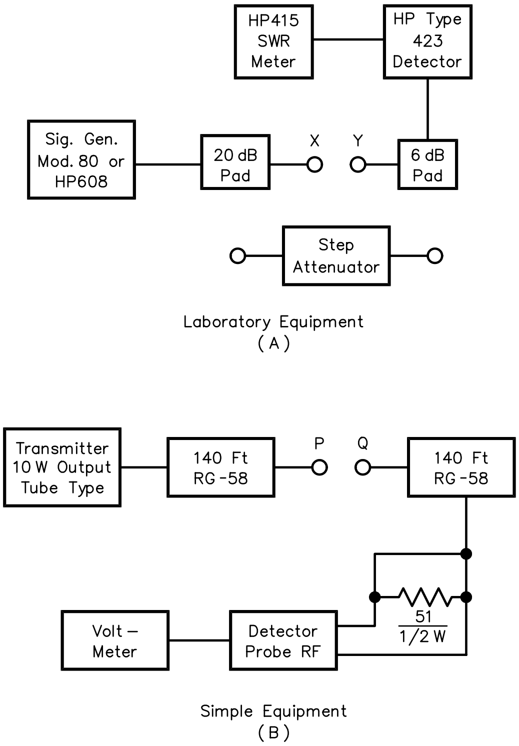

Fig 29—The duplexer can be tuned by either of the two methods shown here, although the method depicted at A is preferred. The signal generator should be modulated by a 1-kHz tone. If the setup shown at B is used, the transmitter should not be modulated, and should have a minimum of noise and spurious signals. The cavities to be aligned are inserted between X and Y in the setup at A, and between P and Q in B.

signal generator to the repeater output frequency and adjust the 415 for a 0 reading on the 20-dB scale.

Reinsert the cavity between X and Y and adjust the cavity tuning for minimum reading on the 415. The notch should be sharp and have a depth of at least 35 dB. It is important to maintain the minimum reading on the meter while tightening the locknut on the tuning shaft.

To check the insertion loss of the cavity, the output from the signal generator should be reduced, and the calibration of the 415 meter checked on the 50-dB expanded scale. Use a fixed 1-dB attenuator to make certain the error is less than 0.1 dB. Replace the attenuator with the cavity and read the loss. The insertion loss should be 0.5 dB or less. The procedure is the same for tuning all six cavities, except that the frequencies are reversed for those having the shunt capacitor installed.

Adjustment with Minimum Equipment

A transmitter with a minimum of spurious output is required. Most modern transmitters meet this requirement. The voltmeter in use should be capable of reading 0.5 volt (or less), full scale. The RF probe used should be rated to 150 MHz or higher. Sections of RG-58 cable are used as attenuators, as shown in Fig 29B. The loss in these 140-foot lengths is nearly 10 dB, and helps to isolate the transmitter in case of mismatch during tuning.

Set the transmitter to the repeater input frequency and connect P and Q. Obtain a reading between 1 and 3 volts on the voltmeter. Insert a cavity with shunt capacitors in place between P and Q and adjust the cavity tuning for a minimum reading on the voltmeter. (This reading should be between 0.01 and 0.05 volt.) The rejection in dB can be calculated by

$$dB = 20 \log (V1/V2)$$

This should be at least 35 dB. Check the insertion loss by putting the receiver on the repeater output frequency and noting the voltmeter reading with the cavity out of the circuit. A 0.5-dB attenuator can be made from a 7-foot length of RG-58. This 7-foot cable can be used to check the calibration of the detector probe and the voltmeter.

Cavities with shunt inductance can be tuned the same way, but with the frequencies reversed. If two or more cavities are tuned while connected together, transmitter noise can cause the rejection readings to be low. In other words, there will be less attenuation.

Results

The duplexer is conservatively rated at 150 watts input, but if constructed carefully should be able to handle as much as 300 watts. Silver plating the interior surfaces of the cavities is recommended if input power is to be greater than 150 watts. A duplexer of this type with silver-plated cavities has an insertion loss of less than 1 dB, and a rejection of more than 100 dB. Unplated cavities should be disassembled at least every two years, cleaned thoroughly, and then retuned.

Miscellaneous Notes

1) Double shielded cable and high quality connectors are *required* throughout the system.
2) The SWR of the antenna should not exceed 1.2:1 for proper duplexer performance.
3) Good shielding of the transmitter and receiver at the repeater is essential.
4) The antenna should have four or more wavelengths of vertical separation from the repeater.
5) Conductors in the near field of the antenna should be well bonded and grounded to eliminate noise.
6) The feed line should be electrically bonded and mechanically secured to the tower or mast.
7) Feed lines and other antennas in the near field of the repeater antenna should be well bonded and as far from the repeater antenna as possible.
8) Individual cavities can be used to improve the performance of separate antenna or separate site repeaters.
9) Individual cavities can be used to help solve intermodulation problems.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references below.

P. Arnold, “Controlling Cavity Drift in Low-Loss Combiners,” *Mobile Radio Technology*, Apr 1986, pp 36-44.

L. Barrett, “Repeater Antenna Beam Tilting,” *Ham Radio*, May 1983, pp 29-35. (See correction, *Ham Radio*, Jul 1983, p 80.)

W. F. Biggerstaff, “Operation of Close Spaced Antennas in Radio Relay Systems,” *IRE Transactions on Vehicular Communications*, Sep 1959, pp 11-15.

J. J. Bilodeau, “A Homemade Duplexer for 2-Meter Repeaters,” *QST*, Jul 1972, pp 22-26, 47.

W. B. Bryson, “Design of High Isolation Duplexers and a New Antenna for Duplex Systems,” *IEEE Transactions on Vehicular Communications*, Mar 1965, pp 134-140.

K. Connolly and P. Blevins, “A Comparison of Horizontal Patterns of Skeletal and Complete Support Structures,” *IEEE 1986 Vehicular Technology Conference Proceedings*, pp 1-7.

S. Kozono, T. Tsuruhara and M. Sakamoto, “Base Station Polarization Diversity Reception for Mobile Radio,” *IEEE Transactions on Vehicular Technology*, Nov 1984, pp 301-306.

J. Kraus, *Antennas*, 2nd ed. (New York: McGraw-Hill Book Co., 1988).

W. Pasternak and M. Morris, *The Practical Handbook of Amateur Radio FM & Repeaters*, (Blue Ridge Summit, PA: Tab Books Inc., 1980), pp 355-363.

M. W. Scheldorf, “Antenna-To-Mast Coupling in Communications," *IRE Transactions on Vehicular Communications*, Apr 1959, pp 5-12.

R. D. Shriner, "A Low Cost PC Board Duplexer," *QST*, Apr 1979, pp 11-14.

W. V. Tilston, "Simultaneous Transmission and Reception with a Common Antenna," *IRE Transactions on Vehicular Communications*, Aug 1962, pp 56-64.

E. P. Tilton, "A Trap-Filter Duplexer for 2-Meter Repeaters," *QST*, Mar 1970, pp 42-46.

R. Wheeler, "Fred's Advice solves Receiver Desense Problem," *Mobile Radio Technology*, Feb 1986, pp 42-44.

R. Yang and F. Willis, "Effects of Tower and Guys on Performance of Side Mounted Vertical Antennas," *IRE Transactions on Vehicular Communications*, Dec 1960, pp 24-31.

Chapter 18

VHF and UHF Antenna Systems

A good antenna system is one of the most valuable assets available to the VHF/UHF enthusiast. Compared to an antenna of lesser quality, an antenna that is well designed, is built of good quality materials, and is well maintained, will increase transmitting range, enhance reception of weak signals and reduce interference problems. The work itself building antennas is by no means the least attractive part of the job. Even with high-gain antennas, experimentation is greatly simplified at VHF and UHF because the antennas are a physically manageable size. Setting up a home antenna range is within the means of most amateurs, and much can be learned about the nature and adjustment of antennas. No large investment in test equipment is necessary.

The Basics

Selecting the best VHF or UHF antenna for a given installation involves much more than scanning gain figures and prices in a manufacturer's catalog. There is no one "best" VHF or UHF antenna design for all purposes. The first step in choosing an antenna is figuring out what you want it to do.

Gain

At VHF and UHF, it is possible to build Yagi antennas with very high gain—15 to 20 dBi—on a physically manageable boom. Such antennas can be combined in arrays of two, four, six, eight or more antennas. These arrays are attractive for EME, tropospheric scatter or other weak-signal communications modes.

Radiation Patterns

Antenna radiation can be made omnidirectional, bidirectional, practically unidirectional, or anything between these conditions. A VHF net operator may find an omnidirectional system almost a necessity, but it may be a poor choice otherwise. Noise pickup and other interference problems are greater with such omnidirectional antennas, and omnidirectional antennas having some gain are especially bad in these respects. Maximum gain and low radiation angle are usually prime interests of the weak-signal DX aspirant. A clean pattern, with lowest possible pickup and radiation off the sides and back, may be important in high-activity areas, or where the noise level is high.

Frequency Response

The ability to work over an entire VHF band may be important in some types of work. Modern Yagis can achieve performance over a remarkably wide frequency range, providing that the boom length is long enough and enough elements are used to populate the boom. Modern Yagi designs in fact are competitive with directly driven collinear arrays of similar size and complexity. The primary performance parameters of gain, front-to-rear ratio and SWR can be optimized over all the VHF or UHF amateur bands readily, with the exception of the full 6meter band from 50.0 to 54.0 MHz, which is an 8% wide bandwidth. A Yagi can be easily designed to cover any 2.0 MHz portion of the 6-meter band with superb performance.

Height Gain

In general, higher is better in VHF and UHF antenna installations. Raising the antenna over nearby obstructions may make dramatic improvements in coverage. Within reason, greater height is almost always worth its cost, but height gain (see Chapter 23, Radio Wave Propagation) must be balanced against increased transmission-line loss. This loss can be considerable, and it increases with frequency. The best available line may not be very good if the run is long in terms of wavelengths. Line loss considerations (see Chapter 24, Transmission Lines) are important in antenna planning.

Physical Size

A given antenna design for 432 MHz has the same gain as the same design for 144 MHz, but being only one-third as large intercepts only one-ninth as much energy in receiving. In other words, the antenna has less pickup efficiency at 432 MHz. To be equal in communication effectiveness, the 432-MHz array should be at least equal in *size* to the 144-MHz antenna, which requires roughly three times as many elements. With all the extra difficulties involved in using the higher frequencies effectively, it is best to keep antennas as large as possible for these bands.

DESIGN FACTORS

With the objectives sorted out in a general way, decisions on specifics, such as polarization, type of transmission line, matching methods and mechanical design must be made.

Polarization

Whether to position antenna elements vertically or horizontally has been widely debated since early VHF pioneering days. Tests have shown little evidence about which polarization sense is most desirable. On long propagation paths there is no consistent advantage either way. Shorter paths tend to yield higher signal levels with horizontally polarized antennas over some kinds of terrain. Man-made noise, especially ignition interference, also tends to be lower with horizontal antennas. These factors make horizontal polarization somewhat more desirable for weak-signal communications. On the other hand, vertically polarized antennas are much simpler to use in omnidirectional systems and in mobile work.

Vertical polarization was widely used in early VHF work, but horizontal polarization gained favor when directional arrays started to become widely used. The major use of FM and repeaters, particularly in the VHF/UHF bands, has tipped the balance in favor of vertical antennas in mobile and repeater use. Horizontal polarization predominates in other communication on 50 MHz and higher frequencies. An additional loss of 20 dB or more can be expected when cross-polarized antennas are used.

TRANSMISSION LINES

Transmission line principles are covered in detail in Chapter 24, Transmission Lines. Techniques that apply to VHF and UHF operation are dealt with in greater detail here. The principles of carrying RF from one location to another via a feed line are the same for all radio frequencies. As at HF, RF is carried principally via open wire lines and coaxial cables at VHF/UHF. Certain aspects of these lines characterize them as good or bad for use above 50 MHz.

Properly built open-wire line can operate with very low loss in VHF and UHF installations. A total line loss under 2 dB per 100 feet at 432 MHz can easily be obtained. A line made of #12 wire, spaced $^3/_4$ inch or more with Teflon spreaders and run essentially straight from antenna to station, can be better than anything but the most expensive coax. Such line can be home made or purchased at a fraction of the cost of coaxial cables, with comparable loss characteristics. Careful attention must be paid to efficient impedance matching if the benefits of this system are to be realized. A similar system for 144 MHz can easily provide a line loss under 1 dB.

Small coax such as RG-58 or RG-59 should never be used in VHF work if the run is more than a few feet. Lines of $^1/_2$-inch diameter (RG-8 or RG-11) work fairly well at 50 MHz, and are acceptable for 144-MHz runs of 50 feet or less. These lines are somewhat better if they employ foam instead of ordinary PE dielectric material. Aluminum-jacket *hardline* coaxial cables with large inner conductors and foam insulation are well worth their cost, and can sometimes be obtained for free from local Cable TV operators as “end runs”—pieces at the end of a roll. The most common CATV cable is $^1/_2$-inch OD 75-Ω hardline. Matched-line loss for this cable is about 1.0 dB/100 feet at 146 MHz and 2.0 dB/100 feet at 432 MHz. Less commonly available from CATV companies is the $^3/_4$-inch 75Ω hardline, sometimes with a black self-healing hard plastic covering. This line has 0.8 dB of loss per 100 feet at 146 MHz, and 1.6 dB loss per 100 feet at 432 MHz. There will be small additional losses for either line if 75-to-50Ω transformers are used at each end.

Commercial connectors for hardline are expensive but provide reliable connections with full waterproofing. Enterprising amateurs have homebrewed low-cost connectors. If they are properly water proofed, connectors and hardline can last almost indefinitely. Hardline must not be bent too sharply, because it will kink. See Chapter 24, Transmission Lines, for details on hardline connectors.

Beware of any “bargains” in coax for VHF or UHF use. Feed-line loss can be compensated to some extent by increasing transmitter power, but once lost, a weak signal can never be recovered in the receiver. Effects of weather on transmission lines should not be ignored. Well-constructed open-wire line works optimally in nearly any weather, and it stands up well. Twin-lead is almost useless in heavy rain, wet snow or icing. The best grades of coax are completely impervious to weather—they can be run underground, fastened to metal towers without insulation and bent into any convenient position with no adverse effects on performance.

WAVEGUIDES

Above 2 GHz, coaxial cable is a losing proposition for communications work. Fortunately, at this frequency the wavelength is short enough to allow practical, efficient energy transfer by an entirely different means. A *waveguide* is a conducting tube through which energy is transmitted in the form of electromagnetic waves. The tube is not considered as carrying a current in the same sense that the wires of a two-conductor line do, but rather as a *boundary* that confines the waves in the enclosed space. Skin effect prevents any electromagnetic effects from being evident outside the guide. The energy is injected at one end, either through capacitive or inductive coupling or by radiation, and is removed from the other end in a like manner. Waveguide merely confines the energy of the fields, which are propagated through it to the receiving end by means of reflections against its inner walls.

Analysis of waveguide operation is based on the assumption that the guide material is a perfect conductor of electricity. Typical distributions of electric and magnetic fields in a rectangular guide are shown in **Fig 1**. The intensity of the electric field is greatest (as indicated by closer spacing of the lines of force) at the center along the X dimension (Fig 1C), diminishing to zero at the end walls. The fields must diminish in this manner, because the existence of any electric field parallel to the walls at the surface would cause an infinite current to flow in a perfect conductor. Waveguides, of course, cannot carry RF in this fashion.

Modes of Propagation

Fig 1 represents the most basic distribution of the electric and magnetic fields in a waveguide. There are an infinite number of ways in which the fields can arrange themselves in a waveguide (for frequencies above the low cutoff frequency of the guide in use). Each of these field configurations is called a *mode*.

The modes may be separated into two general groups. One group, designated *TM* (transverse magnetic), has the magnetic field entirely transverse to the direction of propagation, but has a component of the electric field in that direction. The other type, designated *TE* (transverse electric) has the electric field entirely transverse, but has a component of magnetic field in the direction of propagation. TM waves are sometimes called *E waves*, and TE waves are sometimes called *H waves*, but the TM and TE designations are preferred.

The mode of propagation is identified by the group letters followed by two subscript numerals. For example, TE_{10}, TM_{11}, etc. The number of possible modes increases with frequency for a given size of guide, and there is only one possible mode (called the *dominant mode*) for the lowest frequency that can be transmitted. The dominant mode is the one generally used in amateur work.

Waveguide Dimensions

In rectangular guide the critical dimension is X in Fig 1. This dimension must be more than $\frac{1}{2}\lambda$ at the lowest frequency to be transmitted. In practice, the Y dimension usually is made about equal to $\frac{1}{2}$ X to avoid the possibility of operation in other than the dominant mode.

Cross-sectional shapes other than a rectangle can be used, the most important being the circular pipe. Much the same considerations apply as in the rectangular case.

Wavelength dimensions for rectangular and circular guides are given in **Table 1**, where X is the width of a rectangular guide and r is the radius of a circular guide. All figures apply to the dominant mode.

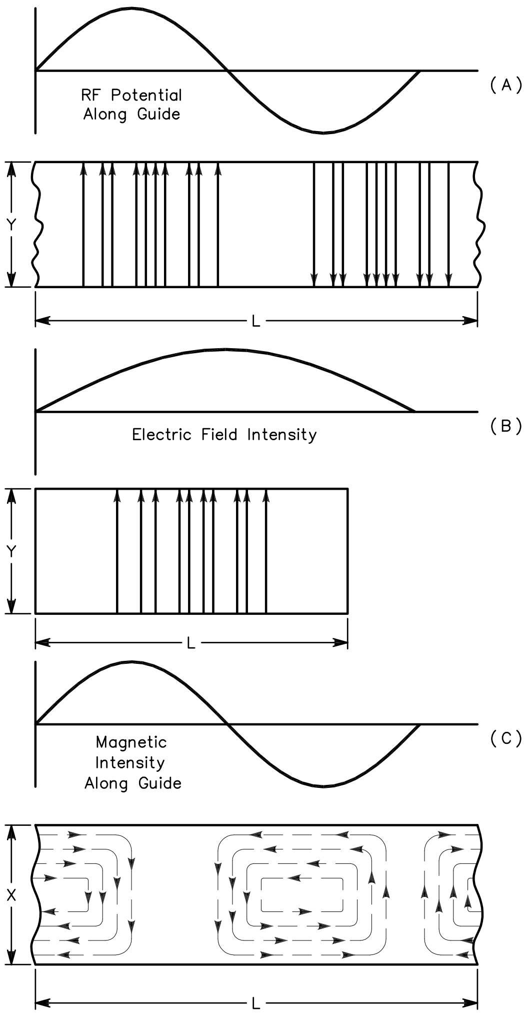

Fig 1—Field distribution in a rectangular waveguide. The TE10 mode of propagation is depicted.

Coupling to Waveguides

Energy may be introduced into or extracted from a waveguide or resonator by means of either the electric or magnetic field. The energy transfer frequently is through a coaxial line. Two methods for coupling to coaxial line are shown in **Fig 2**. The probe shown at A is simply a short extension of the inner conductor of the coaxial line, oriented so that it is parallel to the electric lines of force. The loop shown at B is arranged so that it encloses some of the magnetic lines of force. The point at which maximum coupling is obtained depends upon the mode of propagation in the guide or cavity. Coupling is maximum when the coupling device is in the most intense field.

Coupling can be varied by turning the probe or loop through a 90° angle. When the probe is perpendicular to the electric lines the coupling is minimum. Similarly, when the plane of the loop is parallel to the magnetic lines the coupling is minimum.

If a waveguide is left open at one end it will radiate energy. This radiation can be greatly enhanced by flaring the waveguide to form a pyramidal horn antenna. The horn acts as a transition between the confines of the waveguide and free space. To effect the proper impedance transformation the horn must be at least $^1/_2\ \lambda$ on a side. A horn of this dimension (cutoff) has a unidirectional radiation pattern with a null toward the waveguide transition. The gain at the cutoff frequency is 3 dB, increasing 6 dB with each doubling of frequency. Horns are used extensively in microwave work, both as primary radiators and as feed elements for more elaborate focusing systems. Details for constructing 10-GHz horn antennas are given later in this chapter.

Evolution of a Waveguide

Suppose an open-wire line is used to carry RF energy from a generator to a load. If the line has any appreciable length it must be mechanically supported. The line must be well insulated from the supports if high losses are to be avoided. Because high-quality insulators are difficult to construct at microwave frequencies, the logical alternative is to support the transmission line with $^1/_4\ \lambda$ stubs, shorted at the end opposite the feed line. The open end of such a stub presents an infinite impedance to the transmission line, provided the shorted stub is nonreactive. However, the shorting link has a finite length, and therefore some inductance. The effect of this inductance can be removed by making the RF current flow on the surface of a plate rather than a thin wire. If the plate is large enough, it will prevent the magnetic lines of force from encircling the RF current.

Table 1
Waveguide Dimensions

	Rectangular	*Circular*
Cutoff wavelength	2X	3.41r
Longest wavelength transmitted with little attenuation	1.6X	3.2r
Shortest wavelength before next mode becomes possible	1.1X	2.8r

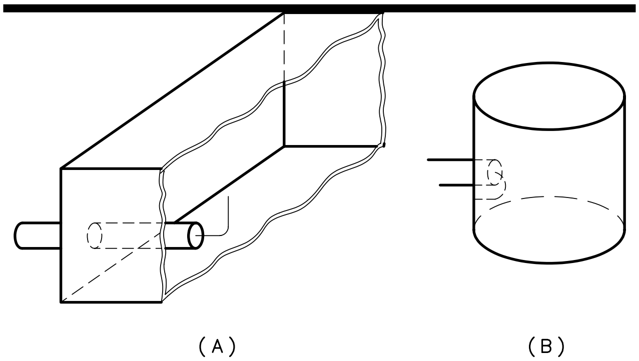

Fig 2—Coupling coaxial line to waveguide and resonators.

An infinite number of these $^1/_4\ \lambda$ stubs may be connected in parallel without affecting the standing waves of voltage and current. The transmission line may be supported from the top as well as the bottom, and when an infinite number of supports are added, they form the walls of a waveguide at its cutoff frequency. **Fig 3** illustrates how a rectangular waveguide evolves from a two-wire parallel transmission line as described. This simplified analysis also shows why the cutoff dimension is $^1/_2\ \lambda$.

While the operation of waveguides is usually described in terms of fields, current does flow on the inside walls, just as on the conductors of a two-wire transmission line. At the waveguide cutoff frequency, the current is concentrated in the center of the walls, and disperses toward the floor and ceiling as the frequency increases.

IMPEDANCE MATCHING

Impedance matching is covered in detail in Chapter 25, Coupling the Transmitter to the Line, and Chapter 26, Coupling the Line to the Antenna. The theory is the same for frequencies above 50 MHz. Practical aspects are similar, but physical size can be a major factor in the choice of methods. Only the matching devices used in practical construction examples later in this chapter are discussed in detail here. This should not rule out consideration of other methods, however, and you should read the relevant portions of both Chapters 25 and 26.

Universal Stub

As its name *universal stub* implies, the double-adjustment stub of **Fig 4A** is useful for many matching purposes. The stub length is varied to resonate the system and the transmission-line attachment point is varied until the transmission line and stub impedances are equal. In practice this involves moving both the sliding short and

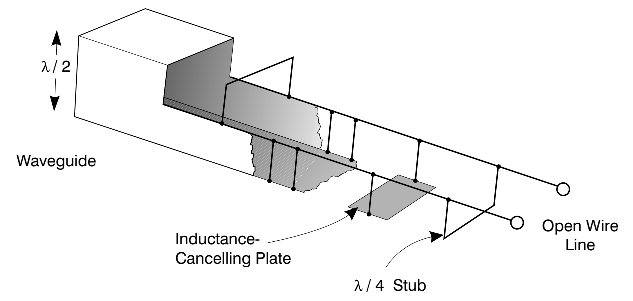

Fig 3—At its cutoff frequency a rectangular waveguide can be thought of as a parallel two-conductor transmission line supported from top and bottom by an infinite number of $^{1}/_{4}$-λ stubs.

the point of line connection for zero reflected power, as indicated on an SWR bridge connected in the line.

The universal stub allows for tuning out any small reactance present in the driven part of the system. It permits matching the antenna to the line without knowledge of the actual impedances involved. The position of the short yielding the best match gives some indication of the amount of reactance present. With little or no reactive component to be tuned out, the stub must be approximately $^{1}/_{2}$ λ from load toward the short.

The stub should be made of stiff bare wire or rod, spaced no more than $^{1}/_{20}$ λ apart. Preferably it should be mounted rigidly, on insulators. Once the position of the short is determined, the center of the short can be grounded, if desired, and the portion of the stub no longer needed can be removed.

It is not necessary that the stub be connected directly to the driven element. It can be made part of an open-wire line as a device to match coaxial cable to the line. The stub can be connected to the lower end of a delta match or placed at the feed point of a phased array. Examples of these uses are given later.

Delta Match

Probably the most basic impedance matching device is the *delta match*, fanned ends of an open-wire line tapped onto a $^{1}/_{2}$ λ antenna at the point of the most-efficient power transfer. This is shown in Fig 4B. Both the side length and the points of connection either side of the center of the element must be adjusted for minimum reflected power on the line, but as with the universal stub, you needn't know the impedances. The delta match makes no provision for tuning out reactance, so the universal stub is often used as a termination for it.

At one time, the delta match was thought to be inferior for VHF applications because of its tendency to radiate if improperly adjusted. The delta has come back into favor now that accurate methods are available for measuring the effects of matching. It is very handy for phasing multiple-bay arrays with open-wire lines, and its dimensions in this use are not particularly critical. It should be checked out carefully in applications like that of Fig 4C, where no tuning device is used.

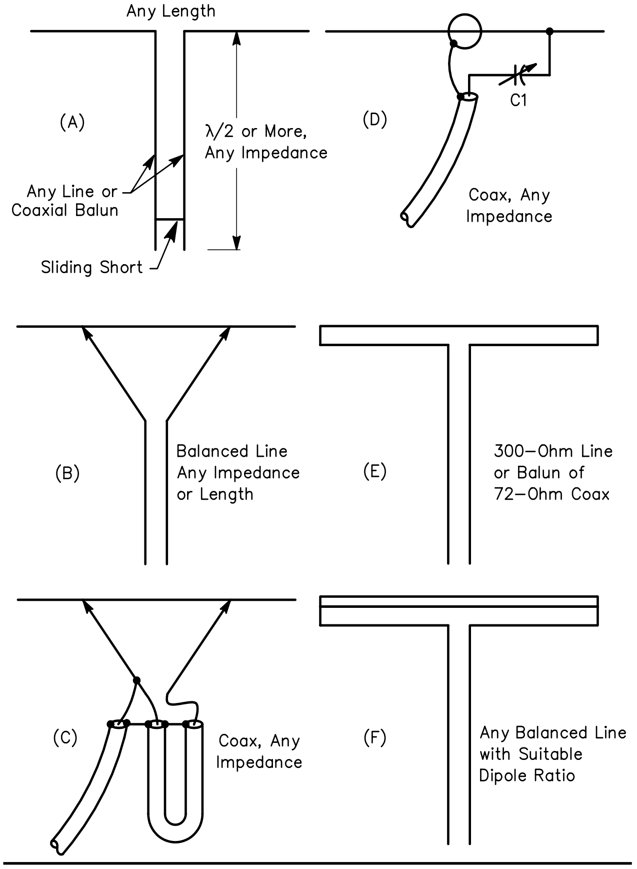
Any Length
(A)
λ/2 or More, Any Impedance
Any Line or Coaxial Balun
Sliding Short
(D)
C1
Coax, Any Impedance
(B)
Balanced Line Any Impedance or Length
(E)
300–Ohm Line or Balun of 72–Ohm Coax
(C)
Coax, Any Impedance
(F)
Any Balanced Line with Suitable Dipole Ratio

Fig 4—Matching methods commonly used at VHF. The universal stub, A, combines tuning and matching. The adjustable short on the stub and the points of connection of the transmission line are adjusted for minimum reflected power on the line. In the delta match, B and C, the line is fanned out and connected to the dipole at the point of optimum impedance match. Impedances need not be known in A, B or C. The gamma match, D, is for direct connection of coax. C1 tunes out inductance in the arm. A folded dipole of uniform conductor size, E, steps up antenna impedance by a factor of four. Using a larger conductor in the unbroken portion of the folded dipole, F, gives higher orders of impedance transformation.

Gamma and T Matches

An application of the same principle allowing direct connection of coax is the *gamma match*, Fig 4D. Because the RF voltage at the center of a $^{1}/_{2}$ λ dipole is zero, the outer conductor of the coax is connected to the element at this point. This may also be the junction with a metallic or wooden boom. The inner conductor, carrying the RF current, is tapped out on the element at the matching point. Inductance of the arm is tuned out by means of C1, resulting in electrical balance. Both the point of contact with the element and the setting of the capacitor are adjusted for zero reflected power, with a bridge connected in the coaxial line.

The capacitance can be varied until the required value is found, and the variable capacitor replaced with a fixed unit of that value. C1 can be mounted in a waterproof box. The maximum required value should be about 100 pF for 50 MHz and 35 to 50 pF for 144 MHz.

The capacitor and arm can be combined in one coaxial assembly, with the arm connected to the driven element by means of a sliding clamp and the inner end of the arm sliding inside a sleeve connected to the center conductor of the coax. An assembly of this type can be constructed from concentric pieces of tubing, insulated by plastic or heat-shrink sleeving. RF voltage across the capacitor is low when the match is adjusted properly, so with a good dielectric, insulation presents no great problem. The initial adjustment should be made with low power. A clean, permanent high-conductivity bond between arm and element is important, since the RF current is high at this point.

Because it is inherently somewhat unbalanced, the gamma match can sometimes introduce pattern distortion, particularly on long-boom, highly directive Yagi arrays. The *T-match*, essentially two gamma matches in series creating a balanced feed system, has become popular for this reason. A coaxial balun like that shown in **Fig 5** is used from the 200 Ω balanced T-match to the unbalanced 50 Ω coaxial line going to the transmitter. See the K1FO Yagi designs later in this chapter for details on practical use of a T-match.

Folded Dipole

The impedance of a $^1/_2$ λ dipole broken at its center is about 70 Ω. If a single conductor of uniform size is folded to make a $^1/_2$ λ dipole as shown in Fig 4E, the impedance is stepped up four times. Such a folded dipole can be fed directly with 300Ω line with no appreciable mismatch. If a 4:1 balun is used, the antenna can be fed with 75Ω coaxial cable. (See balun information presented below.) Higher step-up impedance transformation can be obtained if the unbroken portion is made larger in cross-section than the fed portion, as shown in Fig 4F.

Hairpin Match

The feed-point resistance of most multielement Yagi arrays is less than 50 Ω. If the driven element is split and fed at the center, it may be shortened from its resonant length to add capacitive reactance at the feed point. Then, shunting the feed point with a wire loop resembling a *hairpin* causes a step-up of the feed-point resistance. The hairpin match is used together with a 4:1 coaxial balun in the 50 MHz arrays described later in this chapter. See Chapter 26, Coupling the Line to the Antenna, for details on the hairpin match.

BALUNS AND ANTENNA TUNERS

Conversion from balanced loads to unbalanced lines (or vice versa) can be performed with electrical circuits, or their equivalents made of coaxial cable. A balun made from flexible coax is shown in Fig 5A. The looped portion is an electrical $^1/_2$ λ. The physical length depends on the velocity factor of the line used, so it is important to check its resonant frequency as shown in Fig 5B. The two ends are shorted, and the loop at one end is coupled to a dip meter coil. This type of balun gives an impedance step-up of 4:1 (typically 50 to 200 Ω, or 75 to 300 Ω).

Coaxial baluns that yield 1:1 impedance transformations are shown in **Fig 6**. The coaxial sleeve, open at the top and connected to the outer conductor of the line at the lower end (A) is the preferred type. At B, a conductor of approximately the same size as the line is used with the outer conductor to form a $^1/_4$ λ stub. Another piece of coax, using only the outer conductor, will serve this purpose. Both baluns are intended to present an infi-

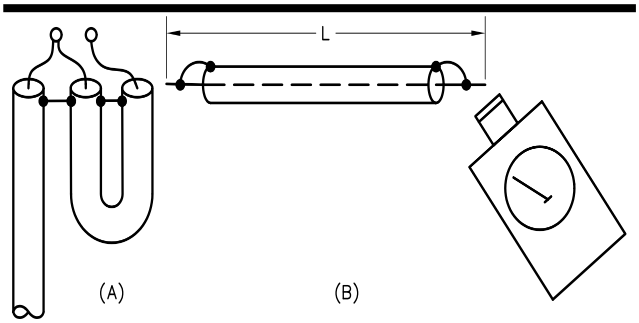

Fig 5—Conversion from unbalanced coax to a balanced load can be done with a ½-λ, coaxial balun at A. Electrical length of the looped section should be checked with a dip meter, with the ends shorted, as at B. The $^1/_2$-λ balun gives a 4:1 impedance step-up.

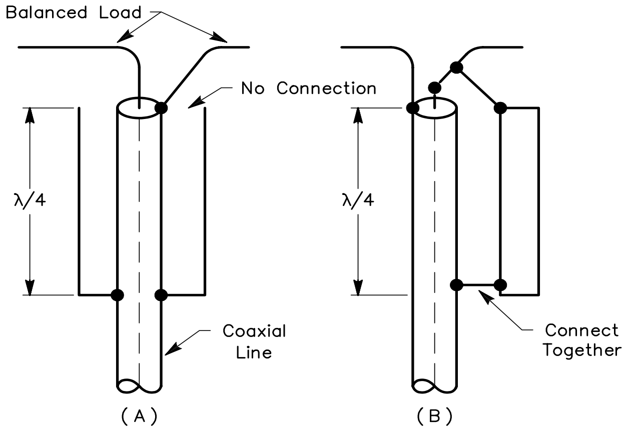

Fig 6—The balun conversion function, with no impedance transformation, can be accomplished with ¼-λ lines, open at the top and connected to the coax outer conductor at the bottom. The coaxial sleeve at A is preferred.

nite impedance to any RF current that might otherwise flow on the outer conductor of the coax.

The functions of the balun and the impedance transformer can be handled by various tuned circuits. Such a device, commonly called an *antenna tuner* or a *Transmatch*, can provide a wide range of impedance transformations. Additional selectivity inherent in the antenna tuner can reduce RFI problems.

THE YAGI AT VHF AND UHF

Without doubt, the Yagi is king of home-station antennas these days. Today's best designs are computer optimized. For years amateurs as well as professionals designed Yagi arrays experimentally. Now we have powerful (and inexpensive) personal computers and sophisticated software for antenna modeling. These have brought us antennas with improved performance, with little or no element pruning required. Chapter 11, HF Yagi Arrays, describes the parameters associated with Yagi-Uda arrays. Except for somewhat tighter dimensional tolerances needed at VHF and UHF, the properties that make a good Yagi at HF also are needed on the higher frequencies. See the end of this chapter for practical Yagi designs.

STACKING YAGIS

Where suitable provision can be made for supporting them, two Yagis mounted one above the other and fed in phase can provide better performance than one long Yagi with the same theoretical or measured gain. The pair occupies a much smaller turning space for the same gain, and their wider elevation coverage can provide excellent results. The wide azimuthal coverage for a vertical stack often results in QSOs that might be missed with a single narrow-beam long-boom Yagi pointed in a different direction. On long ionospheric paths, a stacked pair occasionally may show an *apparent* gain much greater than the measured 2 to 3 dB of stacking gain. (See also the extensive section on stacking Yagis in Chapter 11, HF Yagi Arrays.)

Optimum vertical spacing for Yagis with boom longer than 1 λ or more is about 1 λ (984/50.1 = 19.64 feet), but this may be too much for many builders of 50-MHz antennas to handle. Worthwhile results can be obtained with as little as $^1/_2$ λ (10 feet), but $^5/_8$ λ (12 feet) is markedly better. The difference between 12 and 20 feet, however, may not be worth the added structural problems involved in the wider spacing, at least at 50 MHz. The closer spacings give lower measured gain, but the antenna patterns are cleaner in both azimuth and elevation than with 1 λ spacing. Extra gain with wider spacings is usually the objective on 144 MHz and higher-frequency bands, where the structural problems are not as severe.

Yagis can also be stacked in the same plane (collinear elements) for sharper azimuthal directivity. A spacing of $^5/_8$ λ between the ends of the inner elements yields the maximum gain within the main lobe of the array.

If individual antennas of a stacked array are properly designed, they look like noninductive resistors to the phasing system that connects them. The impedances involved can thus be treated the same as resistances in parallel.

Three sets of stacked dipoles are shown in **Fig 7**. Whether these are merely dipoles or the driven elements of Yagi arrays makes no difference for the purpose of these examples. Two 300 Ω antennas at A are 1 λ apart, resulting in a paralleled feed-point impedance of 150 Ω at the center. (Actually it is slightly less than 150 Ω because of coupling between bays, but this can be neglected for illustrative purposes.) This value remains the same regardless of the impedance of the phasing line. Thus, any convenient line can be used for phasing, as long as the *electrical* length of each line is the same.

The velocity factor of the line must be taken into account as well. As with coax, this is subject to so much variation that it is important to make a resonance check on the actual line used. The method for doing this is shown in Fig 5B. A $^1/_2$ λ line is resonant both open and shorted, but the shorted condition (both ends) is usually the more convenient test condition.

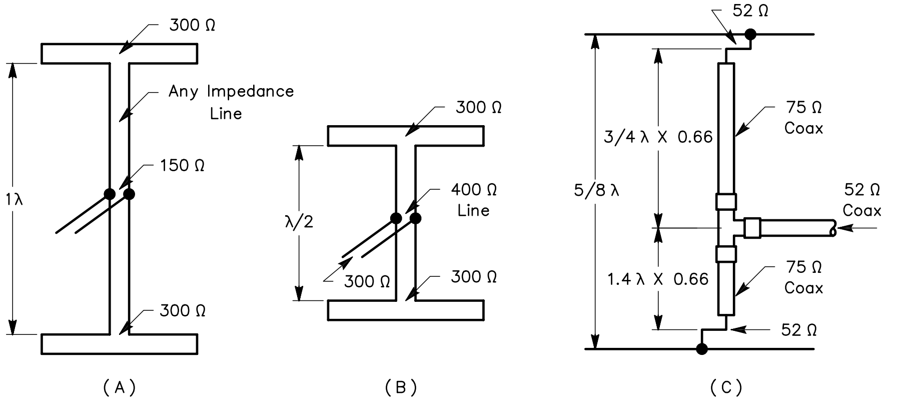

Fig 7—Three methods of feeding stacked VHF arrays. A and B are for bays having balanced driven elements, where a balanced phasing line is desired. Array C has an all-coaxial matching and phasing system. If the lower section is also $^3/_4$ λ no transposition of line connections is needed.

The impedance transforming property of a $^{1}/_{4}$ λ line section can be used in combination matching and phasing lines, as shown in Fig 7B and C. At B, two bays spaced $^{1}/_{2}$ λ apart are phased and matched by a 400-Ω line, acting as a double-Q section, so that a 300-Ω main transmission line is matched to two 300-Ω bays. The two halves of this phasing line could also be $^{3}/_{4}$-λ or $^{5}/_{4}$-λ long, if such lengths serve a useful mechanical purpose. (An example is the stacking of two Yagis where the desirable spacing is more than $^{1}/_{2}$ λ.)

A double-Q section of coaxial line is illustrated in Fig 7C. This is useful for feeding stacked bays that were designed for 50-Ω feed. A spacing of $^{5}/_{8}$ λ is useful for small Yagis, and this is the equivalent of a full electrical wavelength of solid-dielectric coax such as RG-11.

If one phasing line is electrically $^{1}/_{4}$ λ and $^{3}/_{4}$ λ on the other, the connection to one driven element should be reversed with respect to the other to keep the RF currents in the elements in phase—the gamma match is located on opposite sides of the driven elements in Fig 7C. If the number of $^{1}/_{4}$ λ lengths is the same on either side of the feed point, the two connections should be in the same position, and not reversed. Practically speaking however, you can ensure proper phasing by using exactly equal lengths of line from the same roll of coax. This ensures that the velocity factor for each line is identical.

One marked advantage of coaxial phasing lines is that they can be wrapped around the vertical support, taped or grounded to it, or arranged in any way that is mechanically convenient. The spacing between bays can be set at the most desirable value, and the phasing lines placed anywhere necessary.

Stacking Yagis for Different Frequencies

In stacking horizontal Yagis one above the other on a single rotating support, certain considerations apply when the bays are for different bands. As a very general rule of thumb, the minimum desirable spacing is half the boom length of the higher frequency Yagi.

For example, assume the stacked two-band array of **Fig 8A** is for 50 and 144 MHz. This vertical arrangement is commonly referred to as a *Christmas tree*, because it resembles one. The 50MHz Yagi has 5elements on a 12-foot boom. It tends to look like “ground” to the 8-element 144 MHz Yagi on a 12-foot boom directly above it. [The exact Yagi designs for the examples used in this section are located on the CD-ROM accompanying this book. They may be evaluated as monoband Yagis using the *YW* (Yagi for Windows) program also supplied on the CD-ROM. In each case the bottom Yagi in the stack (at the top of the tower) is assumed to be 20 feet high.]

SWR Change in a Multi-Frequency Stack

Earlier editions of *The ARRL Antenna Book* stated that the feed-point impedance of the higher-frequency antenna would likely be affected the most by the proxim-

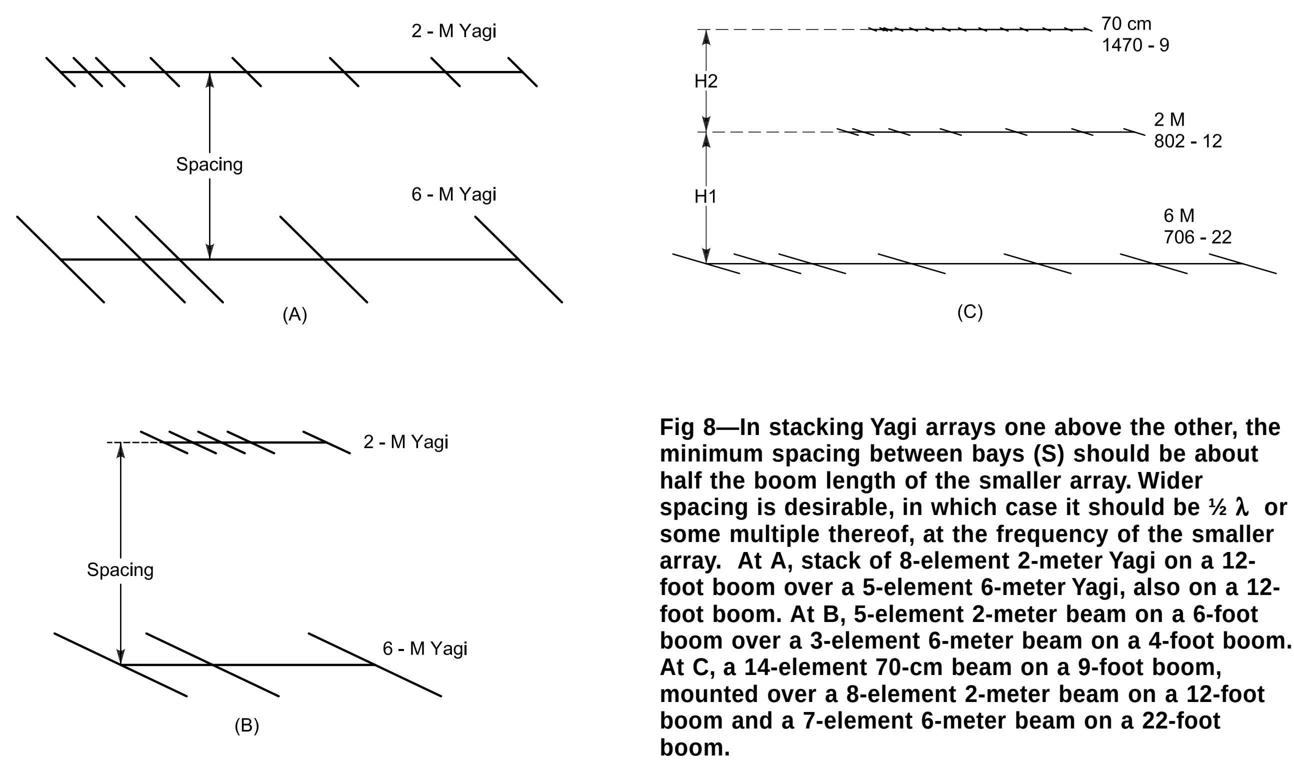

Fig 8—In stacking Yagi arrays one above the other, the minimum spacing between bays (S) should be about half the boom length of the smaller array. Wider spacing is desirable, in which case it should be ½ λ or some multiple thereof, at the frequency of the smaller array. At A, stack of 8-element 2-meter Yagi on a 12-foot boom over a 5-element 6-meter Yagi, also on a 12-foot boom. At B, 5-element 2-meter beam on a 6-foot boom over a 3-element 6-meter beam on a 4-foot boom. At C, a 14-element 70-cm beam on a 9-foot boom, mounted over a 8-element 2-meter beam on a 12-foot boom and a 7-element 6-meter beam on a 22-foot boom.

ity of the lower-frequency Yagi. Modern computer modeling programs reveal that while the feed-point SWR can indeed be affected, by far the greatest degradation is in the forward gain and rearward pattern of the higher-frequency Yagi when the booms are closely spaced. In fact, the SWR curve is usually not affected enough to make it a good diagnostic indicator of interaction between the two Yagis.

Fig 9 shows an overlay of the SWR curves across the 2-meter band for four configurations: an 8-element 2-meter Yagi by itself, and then over a 5-element 6-meter Yagi with spacings between the booms of 1, 2, 4 and 6 feet. The SWR curves are similar—it would be difficult to see any difference between these configurations using typical amateur SWR indicators for anything but the very closest (1-foot) spacing. For example, the SWR curve for the 2-foot spacing case is virtually indistinguishable from that of the Yagi by itself, while the forward gain has dropped more than 0.6 dB because of interactions with the 6-meter Yagi below it.

Gain and Pattern Degradation Due to Stacking

Fig 10 shows four overlaid rectangular plots of the azimuth response from 0° to 180° for the 8-element 2-meter Yagi described above, spaced 1, 2, 4 and 6 feet over a 5-element 6-meter beam. The rectangular presentation gives more detail than a polar plot. The most closely spaced configuration (with 1-foot spacing between the booms) shows the largest degradation in the forward gain, a drop of 1.7 dB. The worst-case front-to-rear ratio for the 6-foot spacing is 29.0 dB, while it is 36.4 dB for the 1-foot spacing—actually better than the F/R for the 8-element 2-meter Yagi by itself. Performance change due to the nearby presence of other Yagis can be enormously complicated (and sometimes is non-intuitive as well).

What happens when a different kind of 6-meter Yagi is mounted below the 8-element 2-meter Yagi? **Fig 11** compares the change in forward gain and the worst-case F/R performance as a function of spacing between the booms for two varieties of 6-meter Yagis: the 5-element design on a 12-foot boom and a 7-element Yagi on a

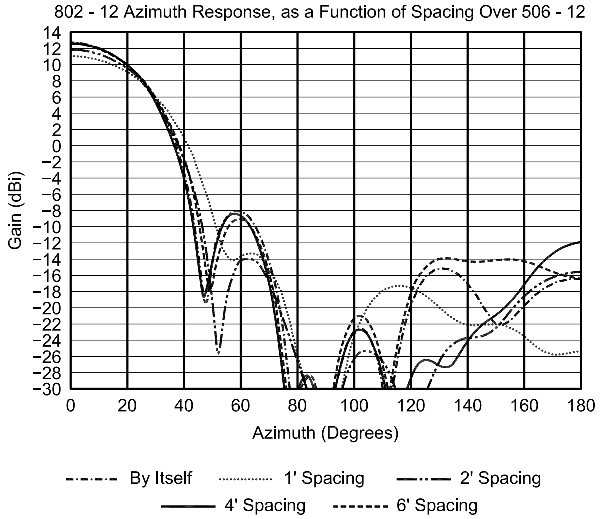

Fig 10—Plots of the 8-element 2-meter Yagi's azimuth response from 0° to 180° for spacing distances from 1 to 6 feet. The sidelobe at about 60° varies about 6 dB over the range of boom spacings, while the shape of worst-case F/R curve varies considerably due to interactions with the lower 6-meter beam. The gain for the 1-foot spacing is degraded by more than 3 dB compared to the 2-meter antenna by itself.

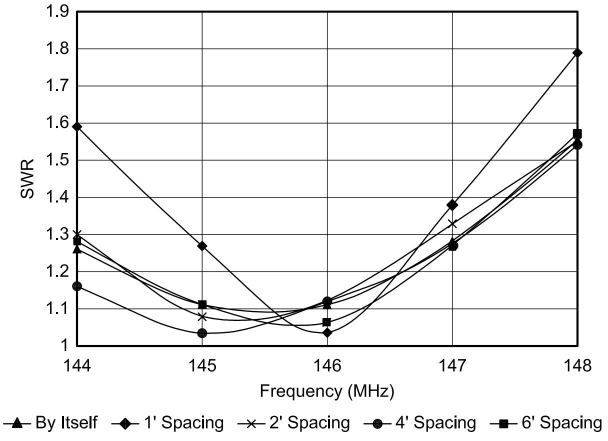

Fig 9—SWR curves for different boom spacing between 8-element 2-meter Yagi on 12-foot boom, over a 5-element 6-meter Yagi on a 12-foot boom. For spacings greater than 1 foot between the booms, differences between the SWR curves are difficult to discern.

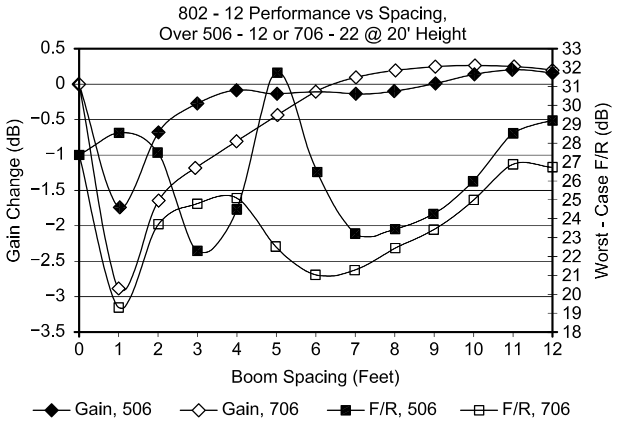

Fig 11—Plot of 8-element 2-meter Yagi's gain and worst-case F/R as a function of distance over two types of 6-meter beams, one on a 12-foot boom and the other on a 22-foot boom. Beyond a spacing of about 5 feet the performance is degraded a minimal amount.

22-foot boom. The spacing of “0 feet” represents the 8-element 2-meter Yagi when it is used alone, with no other antenna nearby. This sets the reference expectations for gain and F/R.

The most severe degradation occurs for the 1-foot spacing, as you might imagine, for both the 12 and 22-foot boom lengths. Over the 5-element 6-meter Yagi, the 2-meter gain doesn’t recover to the reference level of the 8-element 2-meter beam by itself until the spacing is greater than 9 feet. However, the gain is within 0.25 dB of the reference level for spacings of 3 feet or more. Interestingly, the F/R is higher than that of the 2-meter antenna by itself for the 1, 2 and 5-foot spacings and for spacings greater than 11 feet. The 2-meter F/R in the presence of the 12-foot 5-element 6-meter Yagi remains above 20 dB for spacings beyond 1 feet.

Overall, the 2-meter beam performs reasonably well for spacings of 3 feet or more over the 5-element 6-meter Yagi. Put another way, the 2-meter beam’s performance is degraded only slightly for boom spacings greater than 3 feet. A spacing of 3 feet is less than the old rule of thumb that the minimum spacing between booms be greater than one-half the boomlength of the higher-frequency Yagi, which in this case is 6 feet long.

For the 7-element 6-meter Yagi, the 2-meter gain recovers to the reference level for spacings beyond 7 feet, but the F/R is degraded below the reference level for all spacings shown in Fig 11. If we use a gain reduction criterion of less than 0.25 dB and a 20-dB F/R level as the minimum acceptable level, then the spacing must be 5 feet or more over the larger 6-meter Yagi. Again, this is less than the rule of thumb that the minimum spacing between booms be greater than one-half the boomlength of the higher-frequency Yagi.

Now, let’s try a smaller setup of 2- and 6-meter Yagis stacked vertically in a Christmas-tree configuration to see if the rule of thumb for spacing the booms still holds. **Fig 12** shows the performance curves versus boom spacing for a 5-element 2-meter Yagi on a 4-foot boom stacked over a 3-element 6-meter Yagi on a 6-foot boom. Again, the 1-foot spacing produces a substantial gain reduction of about 1.3 dB compared to the reference gain when the 2-meter Yagi is used by itself. Beyond a boom spacing of 3 feet the 2-meter gain drops less than 0.25 dB from the reference level of the 2-meter Yagi by itself and the F/R remains above about 20 dB. In this example, the simple rule of thumb that the minimum spacing between booms be greater than half the boom length (half of 4 feet) of the higher-frequency Yagi does not hold up. However, the same minimum spacing of 3 feet we found for the larger 2-meter Yagi remains true. Three feet spacing is almost 0.5 λ between the booms at the higher frequency.

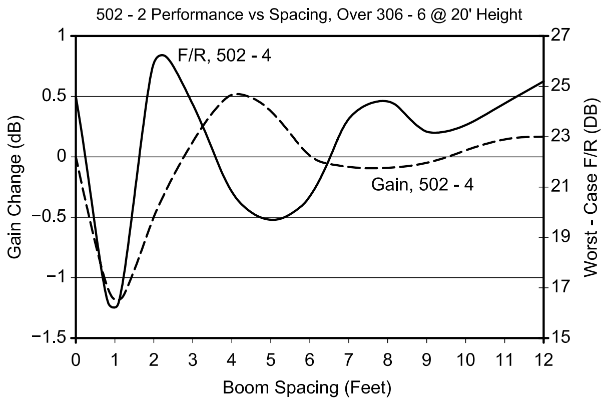

Fig 12— Plot of gain and worst-case F/R of a 5-element 2-meter Yagi on a 4-foot boom as a function of distance over a 3-element 6-meter beam on a 6-foot boom. Beyond a spacing of about 3 feet the performance is degraded a minimal amount.

Adding a 70-cm Yagi to the Christmas Tree

Let’s get more ambitious and set up a larger VHF/UHF Christmas tree, with a 14-element 70-cm Yagi on a 9-foot boom at the top, mounted 5 feet over an 8-element 2-meter Yagi on a 12-foot boom. At the bottom of the stack (at the top of the tower) is either the 5-element 6-meter beam on a 12-foot boom, or a 7-element 6-meter beam on a 22-foot boom. See Fig 8C. As before, we will vary the spacing between the 70-cm Yagi and the 2-meter Yagi below it to assess the interactions that degrade the 70-cm performance.

Fig 13 compares the change in gain and F/R curves as a function of boom spacings between the 70-cm and 2-meter Yagis for the two different 6-meter Yagis (with a fixed distance of 5 feet between the 2-meter and 6-meter Yagis). In this example, the 70-cm Yagi was designed to be an intrinsic 50-Ω feed, where the F/R has been compromised to some extent. Still, the F/R is greater than 20 dB when the 70-cm Yagi is used by itself.

For spacings greater than 4 feet between the 70-cm and 2-meter booms, the 70-cm gain is equal to or even slightly greater than that of the 70-cm antenna by itself. The increase of gain indicates that the elevation pattern of the 70-cm antenna is slightly compressed by the presence of the other Yagis below it. The F/R stays above at 19.5 dB for spacings greater than or equal to 4 feet. This falls just below our desired lower limit of 20 dB, but it is highly doubtful that anyone would notice this 0.5-dB drop in actual operation. A spacing of 4 feet between booms falls under the rule of thumb that the minimum spacing be at least half the boomlength of the higher-frequency Yagi, which in this case is 9 feet.

What should be obvious in this discussion is that you should model the exact configuration you plan to build to avoid unnecessary performance degradation.

Stacking Same-Frequency Yagis

This subject has been examined in some detail in Chapter 11, HF Yagi Arrays. The same basic principles

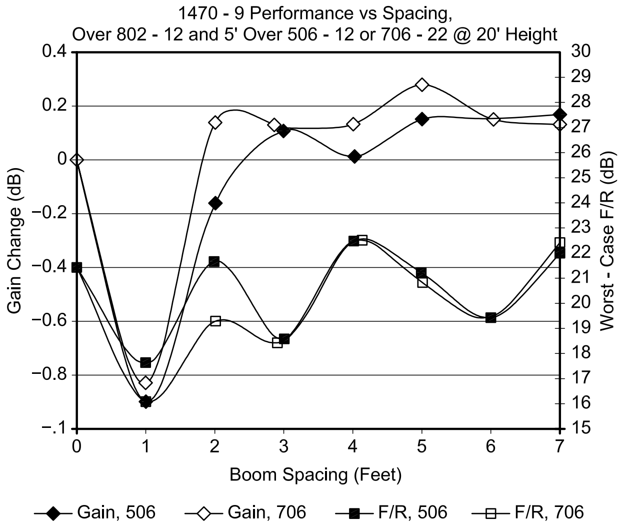

Fig 13—Performance of a 14-element 70-cm Yagi on a 9-foot boom, mounted a variable distance over an 8-element 2-meter Yagi on a 12-foot boom, which is mounted 5 feet above either a 5-element 6-meter Yagi on a 12-foot boom or a 7-element 6-meter Yagi on a 22-foot boom. Beyond a spacing of about 4 feet, the performance of the 70-cm beam is degraded a minimal amount.

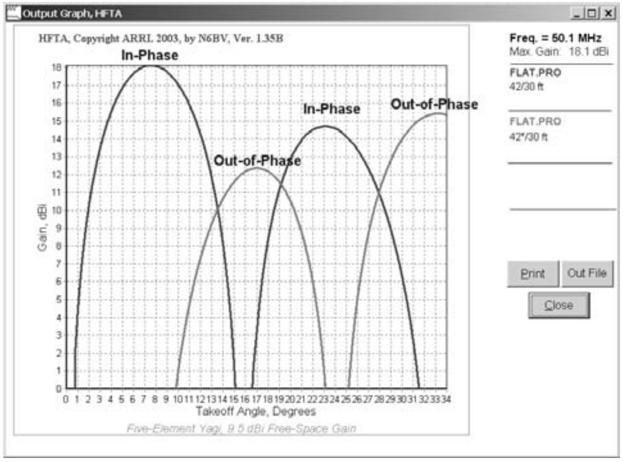

Freq. = 50.1 MHz
Max. Gain: 18.1 dBi
FLAT.PRO
42/30 ft
FLAT.PRO
42°/30 ft

Fig 14—*HFTA* comparison plots of the elevation responses for two 5-element 6-meter Yagis mounted at 42 and 30 feet above flat ground, when they are fed in-phase and out-of-phase. By switching the phasing (adding a half-wavelength of coax to one of the antennas), the elevation angle can be controlled to enhance performance when a Sporadic-E cloud is nearly overhead.

hold at VHF and UHF as they do on HF. That is, the gain increases gradually with increasing spacing between the booms, and then falls off gradually past a certain spacing distance.

At HF, Chapter 11 emphasizes that you should avoid nulls in the antenna's elevation response—so that you can cover all the angles needed for geographic areas of interest. At VHF/UHF, propagation is usually at low elevation angles for most propagation modes, and signals are often extremely weak. Thus, achieving maximum gain is the most common design objective for a VHF/UHF stack. Of secondary importance is the cleanliness of the beam pattern, to discriminate against interference and noise sources.

Six-meter Sporadic-E can sometimes occur at high elevation angles, especially if the E_s cloud is overhead, or nearly overhead. Since Sporadic-E is exactly that, *sporadic*, it's not a good design practice to try to cover a wide range of elevation angles, as you must often do at HF to cover large geographic areas. On 6 meters, you can change to high-angle coverage when necessary. For example, you might switch to a separate Yagi mounted at a low height, or you might provide means to feed stacked antennas out-of-phase. **Fig 14** shows an *HFTA* (HF Terrain Assessment) plot of two 5-element 6-meter Yagis, fed either in-phase or out-of-phase to cover a much wider range of elevation angles than the in-phase stack alone.

Fig 15A shows the change in gain for four 2-meter stacked designs, as a function of the spacing in wavelengths between the booms. The 3-element Yagi is mounted on a 2-foot boom (occupying 0.28 λ of that boom). The 5-element Yagi is on a 4-foot boom (0.51 λ of the boom), while the 8-element Yagi is on a 12-foot boom (1.72 λ of boom). The biggest antenna in the group has 16 elements, on a 27-foot boom (4.0 λ of boom). This range of boom lengths pretty much covers the practical range of antennas used by hams.

The stack of two 3-element Yagis peaks at 3.2 dB of additional gain over a single Yagi for 0.75 λ spacing between the booms. Further increases in spacing see the gain change gradually drop off. Fig 15B shows the worst-case F/R of the four stacks, again as a function of boom length. The F/R of a single 3-element Yagi is just over 24 dB, but in the presence of the second 3-element Yagi in the stack, the F/R of the pair oscillates between 15 to 26 dB, finally remaining consistently over the desired 20-dB level for spacings greater than about 1.7 λ, where the gain has fallen about 0.6 dB from the peak possible gain. A boom spacing of 1.7 λ at 146 MHz is 11.5 feet. Thus you must compromise in choosing the boom spacing between achieving maximum gain and the best pattern.

The increase in gain of the stack of two 5-element Yagis peaks at a spacing of about 1 λ (6.7 feet), where the F/R is an excellent 25 dB. Having more elements on a particular length of boom aids in holding a more consistent F/R in the presence of the second antenna.

The gain increase for the bigger stack of 8-element Yagis peaks at a spacing of about 1.5 λ (10.1 feet), where the F/R is more than 27 dB. The 16-element Yagi's gain

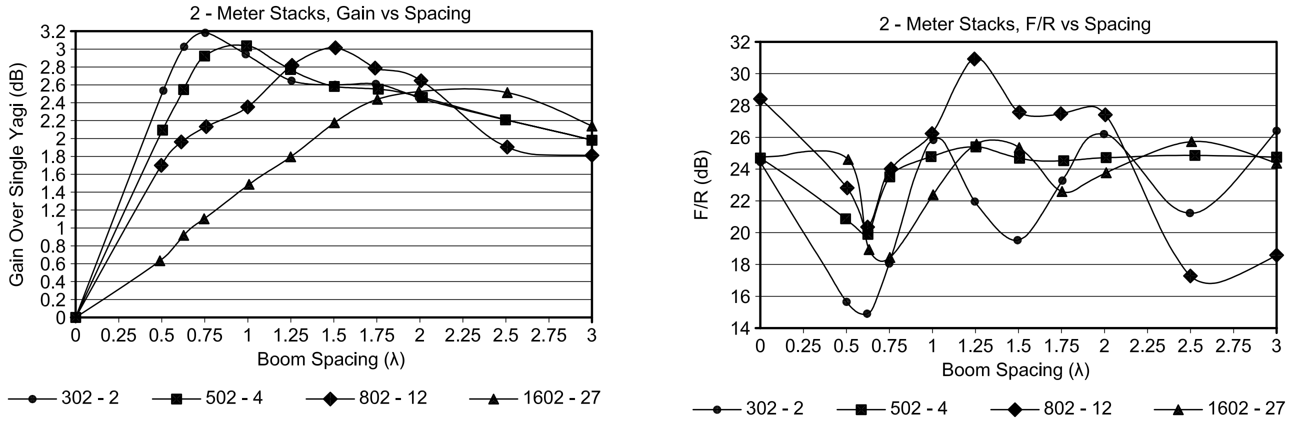

Fig 15—Performance of two different 2-meter Yagis (5-elements on 4-foot boom and 8-elements on 12-foot boom) fed in-phase, as a function of spacing between the booms. Note that the distance is measured in wavelengths.

increase is 2.6 dB for a spacing of about 2.25 λ (15.2 feet), where the F/R remains close to 25 dB. The stacking distance of 15.2 feet for an antenna with a 27-foot long boom may be a real challenge physically, requiring a very sturdy rotating mast to withstand wind pressures without bending.

These examples show that the exact spacing between booms is not overly critical, since the gain varies relatively slowly around the peak. Fig 15A shows that the boom spacing needed to achieve peak gain from a stack increases when higher-gain (longer-boom) individual antennas are used in that stack. It also shows that the increase in maximum gain from stacking decreases for long-boom antennas. Fig 15B shows that beyond boom spacings of about 1 λ, the F/R pattern holds well for Yagi designs with booms longer than about 0.5 λ, which is about 4 feet at 146 MHz.

The plots in Fig 15 are representative of typical modern Yagis. You could simply implement these designs as is, and you'll achieve good results. However, we recommend that you model any specific stack you design, just to make sure. Since the boom spacings are displayed in terms of wavelength, you can extend the results for 2 meters to other bands, provided that you use properly scaled Yagi designs to the other bands too.

You can even tweak the element dimensions and spacings of each Yagi used in a stack to optimize the rearward pattern for a particular stacking distance. This strategy can work out well at VHF/UHF, where stacks are often configured for best gain (and pattern) and are "hard-wired" with fixed lengths of feed lines permanently junctioned together.

This is in contrast to the situation at HF (and even on 6 meters). The HF operator usually wants flexibility to select individual Yagis (or combinations of Yagis) from the stack, to match the array's takeoff angle with ionospheric propagation conditions. See Chapter 11, HF Yagi Arrays. The designer of a flexible HF stack thus usually doesn't try to redo the element lengths and spacings of the Yagis to optimize a particular stack.

Stacking Stacks of Different-Frequency Yagis

The investment in a tower is usually substantial, and most hams want to put as many antennas as possible on a tower, provided that interaction between the antennas can be held to a reasonable level. Really ambitious weak-signal VHF/UHF enthusiasts may want "stacked stacks"—sets of stacked Yagis that cover different bands. For example, a VHF contester might want a stack of two 8-element 2-meter Yagis mounted on the same rotating mast as a stack of two 5-element 6-meter Yagis. Let's assume that the boom length of the 8-element 2-meter Yagis is 12 feet (1.78 λ). We'll assume a boom length of 12 feet (0.61 λ) for the 5-element 6-meter Yagis.

From Fig 15, we find the stacking distance between the 8-element 2-meter beams for peak gain and good pattern is 1.5 λ, or 10 feet, but adequate performance can be had for a boom spacing of 0.75 λ, which is 5 feet on 2 meters.

The boom spacing for two 5-element 6-meter beams is 1 λ for peak stacking gain, but a compromise of 0.625 λ (12 feet) still yields an acceptable gain increase of 2 dB over a single Yagi. The overall height of the rotating mast sticking out of the top of the tower is thus set by the 0.625 λ stacking distance on 6 meters, at 12 feet. In-between the 6-meter Yagis at the bottom and top of the rotating mast we will mount the 2-meter Yagi stack. With only 12 feet available on the mast, the spacing for symmetric placement of the two 2-meter Yagis in-between the 6-meter Yagis dictates a distance of only 4 feet between the 2-meter beams. This is less than optimal.

The performance of the 2-meter stack in this "stack within a stack" is affected by the close spacing, but the interactions are not disastrous. The stacking gain is 1.62 dB more than the gain for a single 8-element 2-meter Yagi and the F/R remains above 20 dB across the 2-meter band.

On 6 meters, the stacking gain for two 5-element 6-meter Yagis spaced 12 feet apart is 2.2 dB more than the gain of a single Yagi, while the F/R pattern remains about 20 dB over the weak-signal portion of the 6-meter band. As described in Chapter 11, HF Yagi Arrays, stacking gives more advantages than merely a gain increase, and 6-meter propagation does require coverage of a range of elevation angles because much of the time ionospheric modes are involved.

Increasing the length of the rotating mast to 18 feet sticking out of the top of the tower will increase performance, particularly on 2 meters. The stacking gain on 6 meters will increase to 2.3 dB while the F/R decreases to 18.5 dB, modest changes both. The 18-foot mast allows the 2-meter Yagis to be spaced 6 feet from each other and 6 feet away from both top and bottom 6-meter antennas. The stacking gain goes to 2.14 dB and the F/R approaches 27 dB in the weak-signal portion of the 2-meter band.

Whether the modest increase in stacking gain is worth the cost and mechanical complexity of stacking two 2-meter Yagis in-between a stack of 6-meter Yagis is a choice left to the operator. Certainly the cost and weight of a rotating mast that is 20 feet long (18 feet out of the top of the tower and 2 feet down inside the tower), a mast that must be sturdy enough to support the antennas in high winds without bending, should give pause to even the most enthusiastic 6-meter weak-signal operator.

QUADS FOR VHF

The quad antenna can be built with inexpensive materials, yet its performance is comparable to other arrays of its size. Adjustment for resonance and impedance matching can be accomplished readily. Quads can be stacked horizontally and vertically to provide high gain, without sharply limiting frequency response. Construction of quad antennas for VHF use is covered later in this chapter.

Stacking Quads

Quads can be mounted side by side or one above the other, or both, in the same general way as other beam antennas. Sets of driven elements can also be mounted in front of a screen reflector. The recommended spacing between adjacent element sides is $^1/_2$ λ. Phasing and feed methods are similar to those employed with other antennas described in this chapter.

Adding Quad Directors

Parasitic elements ahead of the driven element work in a manner similar to those in a Yagi array. Closed loops can be used for directors by making them 5% shorter than the driven element. Spacings are similar to those for conventional Yagis. In an experimental model the reflector was spaced 0.25 λ and the director 0.15 λ. A square array using four 3element bays worked extremely well.

VHF AND UHF QUAGIS

At higher frequencies, especially 420 MHz and above, Yagi arrays using dipole-driven elements can be difficult to feed and match, unless special care is taken to keep the feed-point impedance relatively high by proper element spacing and tuning. The cubical quad described earlier overcomes the feed problems to some extent. When many parasitic elements are used, however, the loops are not nearly as convenient to assemble and tune as are straight cylindrical ones used in conventional Yagis. The *Quagi*, designed and popularized by Wayne Overbeck, N6NB, is an antenna having a full-wave loop driven element and reflector, and Yagi type straight rod directors. Construction details and examples are given in the projects later in this chapter.

COLLINEAR ANTENNAS

The information given earlier in this chapter pertains mainly to parasitic arrays, but the collinear array is worthy of consideration in VHF/UHF operations. This array tends to be tolerant of construction tolerances, making it easy to build and adjust for VHF applications. The use of many collinear driven elements was once popular in very large phased arrays, such as those required in moonbounce (EME) communications, but the advent of computer-optimized Yagis has changed this.

Large Collinear Arrays

Bidirectional curtain arrays of four, six, and eight half waves in phase are shown in **Fig 16**. Usually reflector elements are added, normally at about 0.2 λ behind each driven element, for more gain and a unidirectional pattern. Such parasitic elements are omitted from the sketch in the interest of clarity.

The feed-point impedance of two half waves in phase is high, typically 1000 Ω or more. When they are combined in parallel and parasitic elements are added, the feed impedance is low enough for direct connection to open wire line or twin-lead, connected at the points indicated by black dots. With coaxial line and a balun, it is suggested that the universal stub match, Fig 4A, be used at the feed point. All elements should be mounted at their electrical centers, as indicated by open circles in Fig 16. The framework can be metal or insulating material. The metal supporting structure is entirely behind the plane of the reflector elements. Sheet-metal clamps can be cut from scraps of aluminum for this kind of assembly. Collinear elements of this type should be mounted at their centers (where the RF voltage is zero), rather than at their ends, where the voltage is high and insulation losses and detuning can be harmful.

Collinear arrays of 32, 48, 64 and even 128 elements can give outstanding performance. Any collinear array should be fed at the center of the system, to ensure balanced current distribution. This is very important in large arrays, where sets of six or eight driven elements are treated as "sub arrays," and are fed through a balanced harness. The sections of the harness are resonant lengths, usually of open wire line. The 48-element collinear array for 432 MHz in **Fig 17** illustrates this principle.

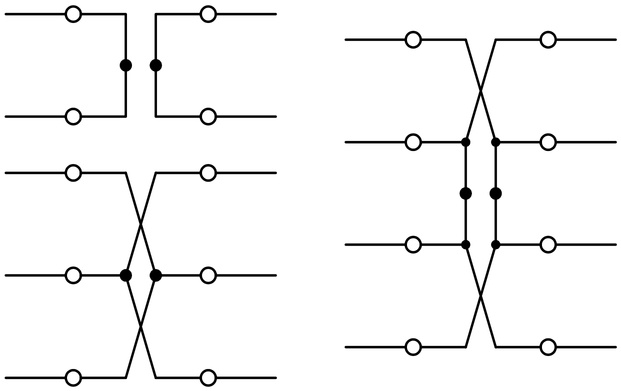

Fig 16—Element arrangements for 8-, 12- and 16-element collinear arrays. Elements are ½ λ long and spaced ½ λ. Parasitic reflectors, omitted here for clarity, are 5% longer and 0.2 λ behind the driven elements. Feed points are indicated by black dots. Open circles show recommended support points. The elements can run through wood or metal booms, without insulation, if supported at their centers in this way. Insulators at the element ends (points of high RF voltage) detune and unbalance the system.

A reflecting plane, which may be sheet metal, wire mesh, or even closely spaced elements of tubing or wire, can be used in place of parasitic reflectors. To be effective, the plane reflector must extend on all sides to at least $^1/_4$ λ beyond the area occupied by the driven elements. The plane reflector provides high F/B ratio, a clean pattern, and somewhat more gain than parasitic elements, but large physical size limits it to use above 420 MHz. An interesting space-saving possibility lies in using a single plane reflector with elements for two different bands mounted on opposite sides. Reflector spacing from the driven element is not critical. About 0.2 λ is common.

THE CORNER REFLECTOR

When a single driven element is used, the reflector screen may be bent to form an angle, giving an improvement in the radiation pattern and gain. At 222 and 420 MHz its size assumes practical proportions, and at 902 MHz and higher, practical reflectors can approach ideal dimensions (very large in terms of wavelengths), resulting in more gain and sharper patterns. The corner reflector can be used at 144 MHz, though usually at much less than optimum size. For a given aperture, the corner reflector does not equal a parabola in gain, but it is simple to construct, broadbanded, and offers gains from about 9 to 14 dBi, depending on the angle and size. This section was written by Paul M. Wilson, W4HHK.

The corner angle can be 90, 60 or 45°, but the side length must be increased as the angle is narrowed. For a 90° corner, the driven element spacing can be anything from 0.25 to 0.7 λ, 0.35 to 0.75 λ for 60°, and 0.5 to

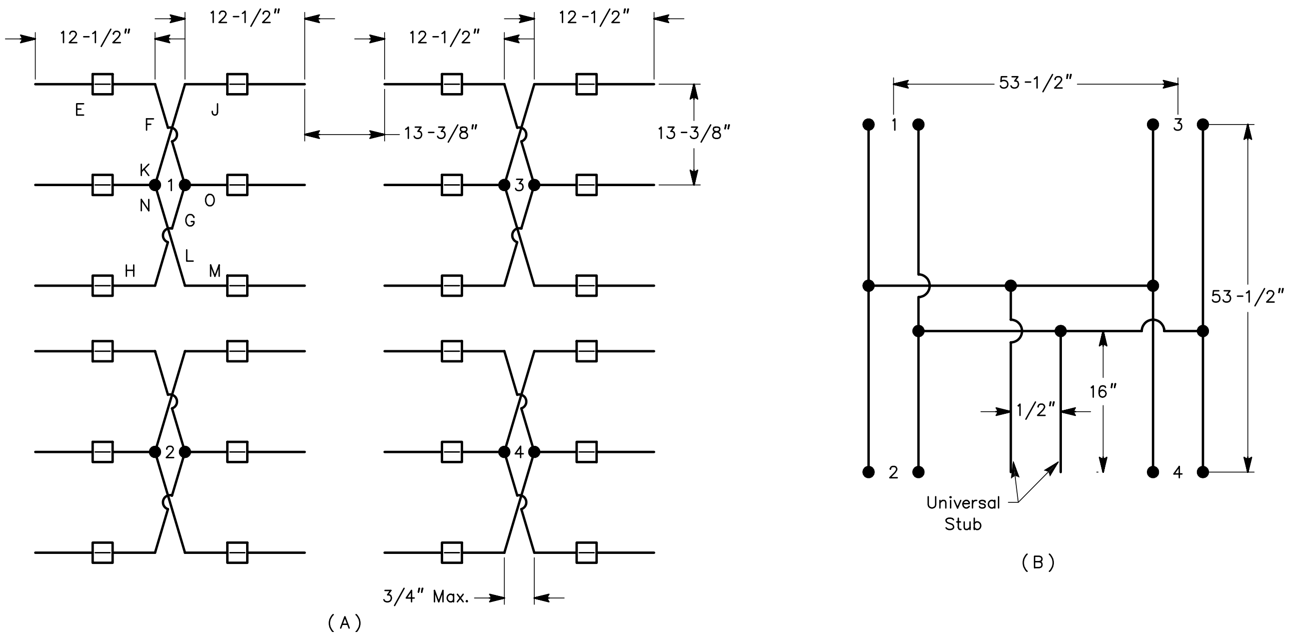

Fig 17—Large collinear arrays should be fed as sets of no more than eight driven elements each, interconnected by phasing lines. This 48-element array for 432 MHz (A) is treated as if it were four 12-element collinear antennas. Reflector elements are omitted for clarity. The phasing harness is shown at B. Squares represent insulators.

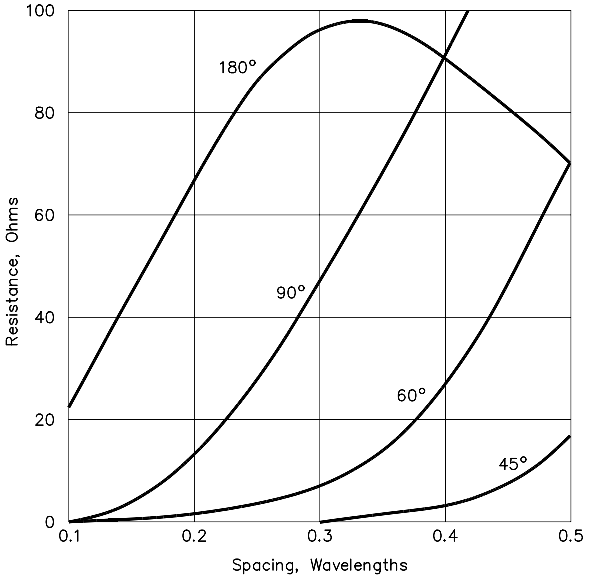

Fig 18—Radiation resistance of the driven element in a corner reflector array for corner angles of 180° (flat sheet), 90°, 60° and 45° as a function of spacing D, as shown in Fig 19.

0.8 λ for 45°. In each case the gain variation over the range of spacings given is about 1.5 dB. Because the spacing is not very critical to gain, it may be varied for impedance-matching purposes. Closer spacings yield lower feed-point impedances, but a folded dipole radiator could be used to raise this to a more convenient level.

Radiation resistance is shown as a function of spacing in **Fig 18**. The maximum gain obtained with minimum spacing is the primary mode (the one generally used at 144, 222 and 432 MHz to maintain reasonable side lengths). A 90° corner, for example, should have a minimum side length (S, **Fig 19**) equal to twice the dipole spacing, or 1 λ long for 0.5-λ spacing. A side length greater than 2 λ is ideal. Gain with a 60° or 90° corner reflector with 1-λ sides is about 10 dB. A 60° corner with 2-λ sides has about 13 dBi gain, and a 45° corner with 3-λ sides has about 14 dBi gain.

Reflector length (L, Fig 19) should be a minimum of 0.6 λ. Less than that spacing causes radiation to increase to the sides and rear, and decreases gain.

Spacing between reflector rods (G, Fig 19) should not exceed 0.06 λ for best results. A spacing of 0.06 λ results in a rear lobe that is about 6% of the forward lobe

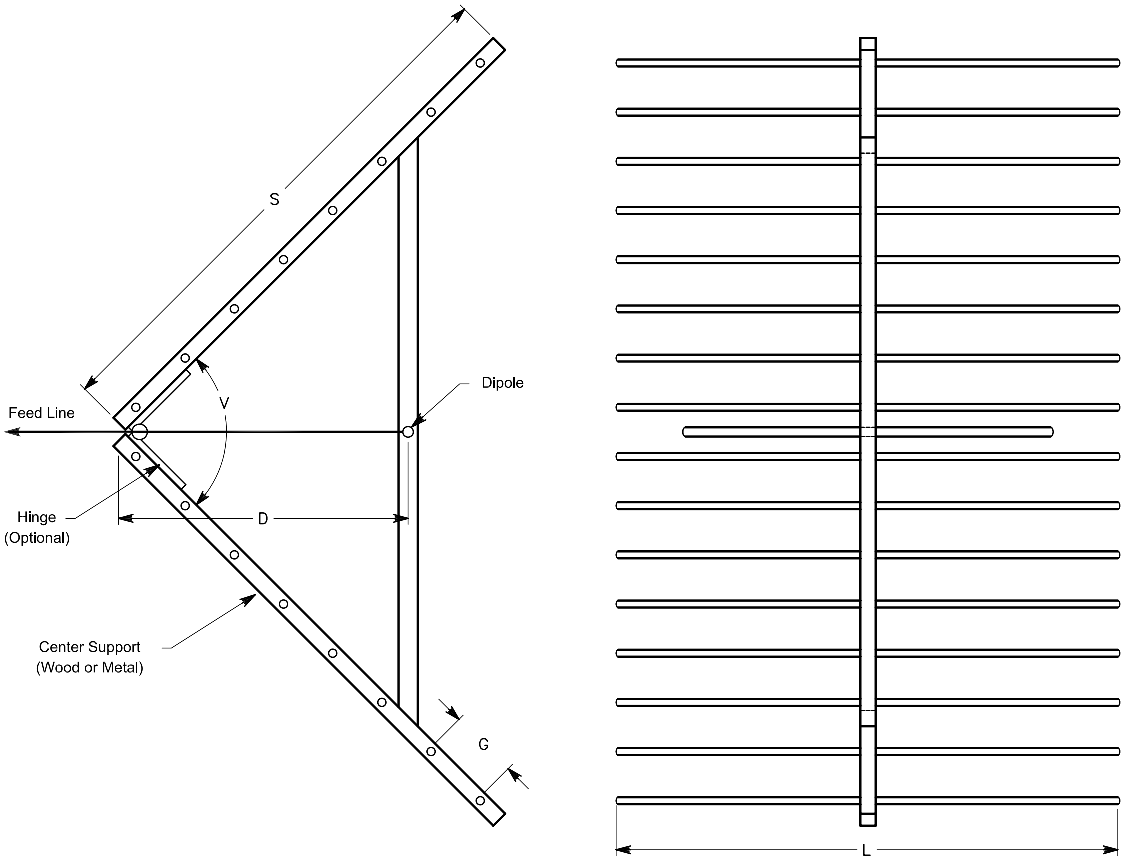

Fig 19—Construction of a corner reflector array. The frame can be wood or metal. Reflector elements are stiff wire or tubing. Dimensions for several bands are given in Table 2. Reflector element spacing, G, is the maximum that should be used for the frequency; closer spacings are optional. The hinge permits folding for portable use.

(down 12 dB). A small mesh screen or solid sheet is preferable at the higher frequencies to obtain maximum efficiency and highest F/B ratio, and to simplify construction. A spacing of 0.06 λ at 1296 MHz, for example, requires mounting reflector rods about every $^{1}/_{2}$ inch along the sides. Rods or spines may be used to reduce wind loading. The support used for mounting the reflector rods may be of insulating or conductive material. Rods or mesh weave should be parallel to the radiator.

A suggested arrangement for a corner reflector is shown in Fig 19. The frame may be made of wood or metal, with a hinge at the corner to facilitate portable work or assembly atop a tower. A hinged corner is also useful in experimenting with different angles. **Table 2** gives the principal dimensions for corner reflector arrays for 144 to 2300 MHz. The arrays for 144, 222 and 420 MHz have side lengths of twice to four times the driven element spacing. The 915 MHz corner reflectors use side lengths of three times the element spacing, 1296 MHz corners use side lengths of four times the spacing, and 2304 MHz corners employ side lengths of six times the spacing. Reflector lengths of 2, 3, and 4 wavelengths are used on the 915, 1296 and 2304 MHz reflectors, respectively. A 4 × 6 λ reflector closely approximates a sheet of infinite dimensions.

A corner reflector may be used for several bands, or for UHF television reception, as well as amateur UHF work. For operation on more than one frequency, side length and reflector length should be selected for the lowest frequency, and reflector spacing for the highest frequency. The type of driven element plays a part in determining bandwidth, as does the spacing to the corner. A fat cylindrical element (small λ/dia ratio) or triangular dipole (bow tie) gives more bandwidth than a thin driven element. Wider spacings between driven element and corner give greater bandwidths. A small increase in gain can be obtained for any corner reflector by mounting collinear elements in a reflector of sufficient size, but

Table 2
Dimensions of Corner Reflector Arrays for VHF and UHF

Freq, MHz	*Side Length S, in.*	*Dipole to Vertex D, in.*	*Reflector Length L, in.*	*Reflector Spacing G, in.*	*Corner Angle, Vo*	*Radiation Resistance, Ω*
144*	65	27½	48	7¾	90	70
144	80	40	48	4	90	150
222*	42	18	30	5	90	70
222	52	25	30	3	90	150
222	100	25	30	Screen	60	70
420	27	8¾	16¼	2$^{5}/_{8}$	90	70
420	54	13½	16¼	Screen	60	70
915	20	6½	25¾	0.65	90	70
915	51	16¾	25¾	Screen	60	65
915	78	25¾	25¾	Screen	45	70
1296	18	4½	27½	½	90	70
1296	48	11¾	27½	Screen	60	65
1296	72	18¼	27½	Screen	45	70
2304	15½	2½	20½	¼	90	70
2304	40	6¾	20½	Screen	60	65
2304	61	10¼	20½	Screen	45	70

*Side length and number of reflector elements somewhat below optimum—slight reduction in gain.

Notes:

915 MHz
Wavelength is 12.9 in.
Side length S is 3 × D, dipole to vertex distance
Reflector length L is 2.0 λ
Reflector spacing G is 0.05 λ

1296 MHz
Wavelength is 9.11 in.
Side length S is 4 × D, dipole to vertex distance
Reflector length L is 3.0 λ
Reflector spacing G is 0.05 λ

2304 MHz
Wavelength is 5.12 in.
Side length S is 6 × D, dipole to vertex distance
Reflector length L is 4.0 λ
Reflector spacing G is 0.05 λ

the simple feed of a dipole is lost if more than two elements are used.

A dipole radiator is usually employed with a corner reflector. This requires a balun between the coaxial line and the balanced feed-point impedance of the antenna. Baluns are easily constructed of coaxial line on the lower VHF bands, but become more difficult at the higher frequencies. This problem may be overcome by using a ground-plane corner reflector, which can be used for vertical polarization. A ground-plane corner with monopole driven element is shown in **Fig 20**. The corner reflector and a ¼ λ radiator are mounted on the ground plane, permitting direct connection to a coaxial line if the proper spacing is used. The effective aperture is reduced, but at the higher frequencies, second- or third-mode radiator spacing and larger reflectors can be employed to obtain more gain and offset the loss in effective aperture. A J antenna could be used to maintain the aperture area and provide a match to a coaxial line.

For vertical polarization work, four 90° corner reflectors built back-to-back (with common reflectors) could be used for scanning 360° of horizon with modest gain. Feed-line switching could be used to select the desired sector.

TROUGH REFLECTORS

To reduce the overall dimensions of a large corner reflector the vertex can be cut off and replaced with a plane reflector. Such an arrangement is known as a *trough reflector*. See **Fig 21**. Performance similar to that of the large corner reflector can thereby be had, provided that the dimensions of S and T as shown in Fig 21 do not exceed the limits indicated in the figure. This antenna provides performance very similar to the corner reflector, and presents fewer mechanical problems because the plane center portion is relatively easy to mount on the mast. The sides are considerably shorter, as well.

The gain of both corner reflectors and trough reflectors may be increased by stacking two or more and arranging them to radiate in phase, or alternatively by

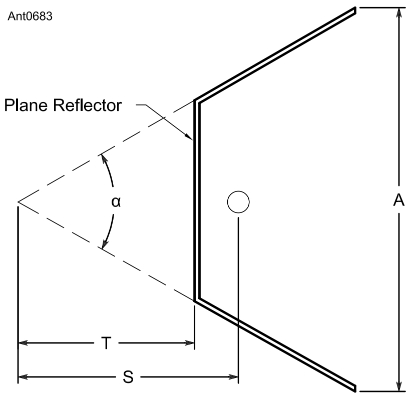

Angle α	*Value of S for maximum gain*	*Gain*	*Value of T*
90°	1.5 λ	13.5 dB	1λ - 1.25 λ
60°	1.75 λ	15 dB	1.0 λ
45°	2.0 λ	16 dB	1.9 λ

Fig 21—The trough reflector. This is a useful modification of the corner reflector. The vertex has been cut off and replaced by a simple plane section. The tabulated data shows the gain obtainable for greater values of S than those covered in Table 2, assuming that the reflector is of adequate size.

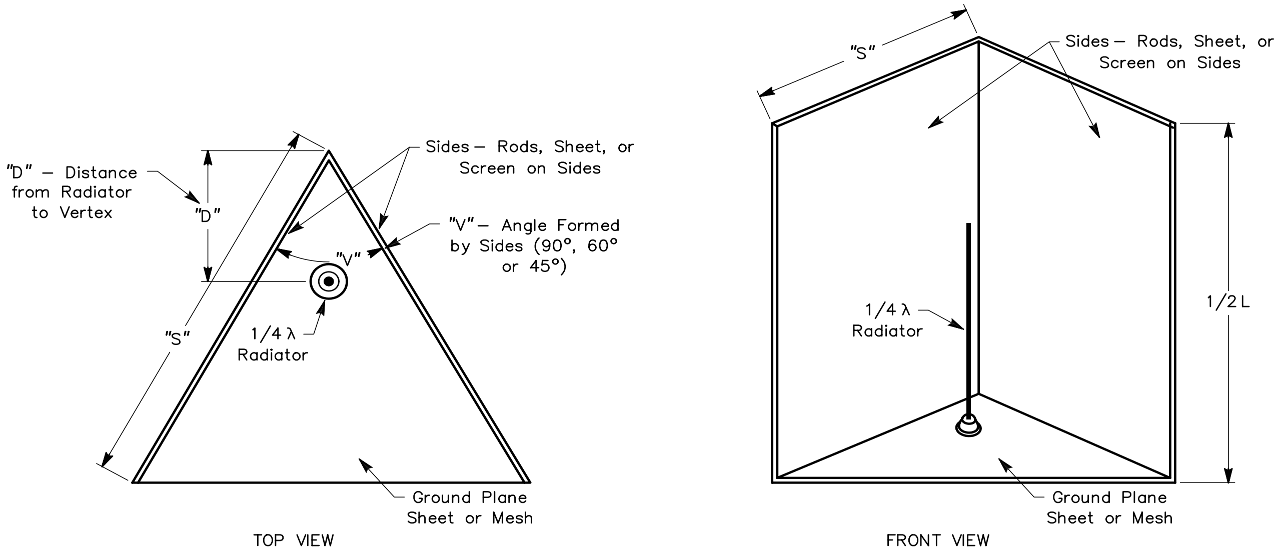
"D" — Distance from Radiator to Vertex
"D"
Sides— Rods, Sheet, or Screen on Sides
"V"— Angle Formed by Sides (90°, 60° or 45°)
"V"
"S"
1/4 λ Radiator
Ground Plane Sheet or Mesh
TOP VIEW
"S"
Sides— Rods, Sheet, or Screen on Sides
1/4 λ Radiator
1/2 L
Ground Plane Sheet or Mesh
FRONT VIEW

Fig 20—A ground-plane corner reflector antenna for vertical polarization, such as FM communications or packet radio. The dimension ½ λ in the front view refers to data in Table 2.

adding further collinear dipoles (fed in phase) within a wider reflector. Not more than two or three radiating units should be used, because the great virtue of the simple feeder arrangement would then be lost.

HORN ANTENNAS FOR THE MICROWAVE BANDS

Horn antennas were briefly introduced in the section on coupling energy into and out of waveguides. For amateur purposes, horns begin to show usable gain with practical dimensions in the 902 MHz band.

It isn't necessary to feed a horn with waveguide. If only two sides of a pyramidal horn are constructed, the antenna may be fed at the apex with a two-conductor transmission line. The impedance of this arrangement is on the order of 300 to 400 Ω. A 60° two-sided pyramidal horn with 18 inch sides is shown in **Fig 22**. This antenna has a theoretical gain of 15 dBi at 1296 MHz, although the feed system detailed in **Fig 23** probably degrades this value somewhat. A $\frac{1}{4}\lambda$, 150-Ω matching section made from two parallel lengths of twin-lead connects to a bazooka balun made from RG-58 cable and a brass tube. This matching system was assembled strictly for the purpose of demonstrating the two-sided horn in a 50-Ω system. In a practical installation the horn would be fed with open-wire line and matched to 50 Ω at the station equipment.

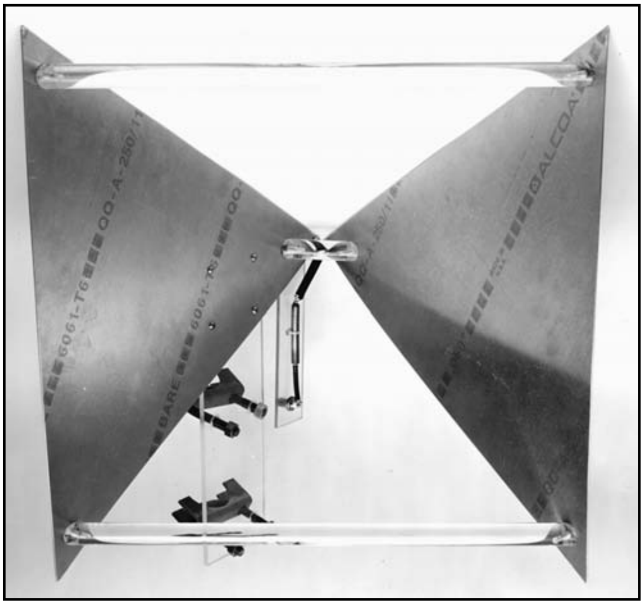

Fig 22—An experimental two-sided pyramidal horn constructed in the ARRL laboratory. A pair of muffler clamps allows mounting the antenna on a mast. This model has sheet-aluminum sides, although window screen would work as well. Temporary elements could be made from cardboard covered with aluminum foil. The horizontal spreaders are Plexiglas rod. Oriented as shown here, the antenna radiates horizontally polarized waves.

PARABOLIC ANTENNAS

When an antenna is located at the focus of a parabolic reflector (dish), it is possible to obtain considerable gain. Furthermore, the beamwidth of the radiated energy will be very narrow, provided all the energy from the driven element is directed toward the reflector. This section was written by Paul M. Wilson, W4HHK.

Gain is a function of parabolic reflector diameter, surface accuracy and proper illumination of the reflector by the feed. Gain may be found from

$$G = 10 \log k \left(\frac{\pi D}{\lambda}\right)^2 \qquad \text{(Eq 1)}$$

where

G = gain over an isotropic antenna, dBi (subtract 2.15 dB for gain over a dipole)

k = efficiency factor, usually about 55%

D = dish diameter in feet

λ = wavelength in feet

See **Table 3** for parabolic antenna gain for the bands 420 MHz through 10 GHz and diameters of 2 to 30 feet.

A close approximation of beamwidth may be found from

$$\psi = \frac{70\lambda}{D} \qquad \text{(Eq 2)}$$

where

ψ = beamwidth in degrees at half-power points (3 dB down)

D = dish diameter in feet

λ = wavelength in feet

At 420 MHz and higher, the parabolic dish becomes a practical antenna. A simple, single feed point eliminates phasing harnesses and balun requirements. Gain is dependent on good surface accuracy, which is more difficult to achieve with increasing frequency. Surface

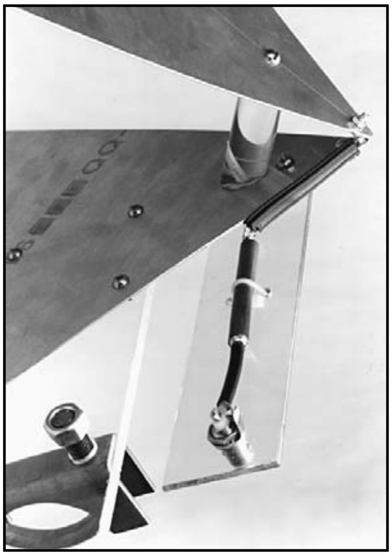

Fig 23—Matching system used to test the horn. Better performance would be realized with open wire line. See text.

Table 3
Gain, Parabolic Antennas*

	Dish Diameter (Feet)						
Frequency	*2*	*4*	*6*	*10*	*15*	*20*	*30*
420 MHz	6.0	12.0	15.5	20.0	23.5	26.0	29.5
902	12.5	18.5	22.0	26.5	30.0	32.5	36.0
1215	15.0	21.0	24.5	29.0	32.5	35.0	38.5
2300	20.5	26.5	30.0	34.5	38.0	40.5	44.0
3300	24.0	30.0	33.5	37.5	41.5	43.5	47.5
5650	28.5	34.5	38.0	42.5	46.0	48.5	52.0
10 GHz	33.5	39.5	43.0	47.5	51.0	53.5	57.0

*Gain over an isotropic antenna (subtract 2.1 dB for gain over a dipole antenna). Reflector efficiency of 55% assumed.

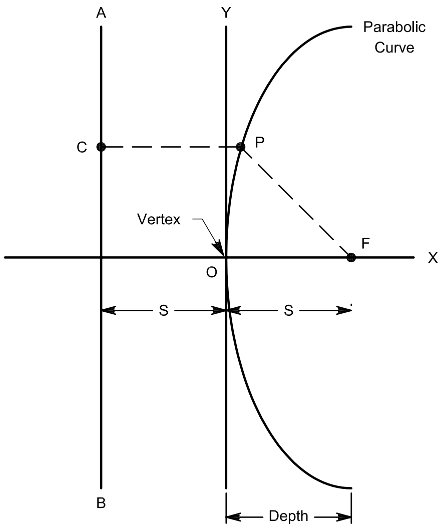

Fig 24—Details of the parabolic curve, $Y^2 = 4SX$. This curve is the locus of points that are equidistant from a fixed point, the focus (F), and a fixed line (AB) that is called the *directrix*. Hence, FP = PC. The focus (F) is located at coordinates S,0.

errors should not exceed $^1/_8\ \lambda$ in amateur work. At 430 MHz $^1/_8\ \lambda$ is 3.4 inches, but at 10 GHz it is 0.1476 inch! Mesh can be used for the reflector surface to reduce weight and wind loading, but hole size should be less than $^1/_{12}\ \lambda$. At 430 MHz the use of 2-inch hole diameter poultry netting (chicken wire) is acceptable. Fine mesh aluminum screening works well as high as 10 GHz.

A support form may be fashioned to provide the proper parabolic shape by plotting a curve (**Fig 24**) from

$$Y^2 = 4SX$$

as shown in the figure.

Optimum illumination occurs when power at the reflector edge is 10 dB less than that at the center. A circular waveguide feed of correct diameter and length for the frequency and correct beamwidth for the dish focal length to diameter (f/D) ratio provides optimum illumination at 902 MHz and higher. This, however, is impractical at 432 MHz, where a dipole and plane reflector are often used. An f/D ratio between 0.4 and 0.6 is considered ideal for maximum gain and simple feeds.

The focal length of a dish may be found from

$$f = \frac{D^2}{16d} \qquad \text{(Eq 3)}$$

where

f = focal length
D = diameter
d = depth distance from plane at mouth of dish to vertex (see Fig 17)

The units of focal length f are the same as those used to measure the depth and diameter. **Table 4** gives the subtended angle at focus for dish f/D ratios from 0.2 to 1.0. A dish, for example, with a typical f/D of 0.4 requires a 10-dB beamwidth of 130°. A circular waveguide feed with a diameter of approximately 0.7 λ provides nearly optimum illumination, but does not uniformly illuminate the reflector in both the magnetic (TM) and electric (TE) planes. **Fig 25** shows data for plotting radiation patterns from circular guides. The waveguide feed aperture can be modified to change the beamwidth.

One approach used successfully by some experimenters is the use of a disc at a short distance behind the aperture as shown in **Fig 26**. As the distance between the aperture and disc is changed, the TM plane patterns become alternately broader and narrower than with an unmodified aperture. A disc about 2 λ in diameter appears to be as effective as a much larger one. Some experimenters have noted a 1 to 2 dB increase in dish gain with this modified feed. Rectangular waveguide feeds can also be used, but dish illumination is not as uniform as with round guide feeds.

The circular feed can be made of copper, brass, aluminum or even tin in the form of a coffee or juice can, but the latter must be painted on the outside to prevent rust or corrosion. The circular feed must be within a proper size (diameter) range for the frequency being used.

This feed operates in the dominant circular waveguide mode known as the TE_{11} mode. The guide must be large enough to pass the TE_{11} mode with no attenuation, but smaller than the diameter that permits the next higher TM_{01} mode to propagate. To support the desirable TE_{11} mode in circular waveguide, the cutoff frequency, F_C, is given by

$$F_C\,(TE_{11}) = \frac{6917.26}{d\,(\text{inches})} \qquad \text{(Eq 4)}$$

where

F_C = cutoff frequency in MHz for TE_{11} mode
d = waveguide inner diameter

Circular waveguide will support the TM_{01} mode having a cutoff frequency

$$F_C\,(TM_{01}) = \frac{9034.85}{d\,(\text{inches})} \qquad \text{(Eq 5)}$$

The wavelength in a waveguide always exceeds the free-space wavelength and is called guide wavelength, λ_g. It is related to the cutoff frequency and operating frequency by the equation

$$\lambda_g = \frac{11802.85}{\sqrt{f_0^{\,2} - f_C^{\,2}}} \qquad \text{(Eq 6)}$$

where

λ_g = guide wavelength, inches
f_0 = operating frequency, MHz
f_C = TE_{11} waveguide cutoff frequency, MHz

An inside diameter range of about 0.66 to 0.76 l is

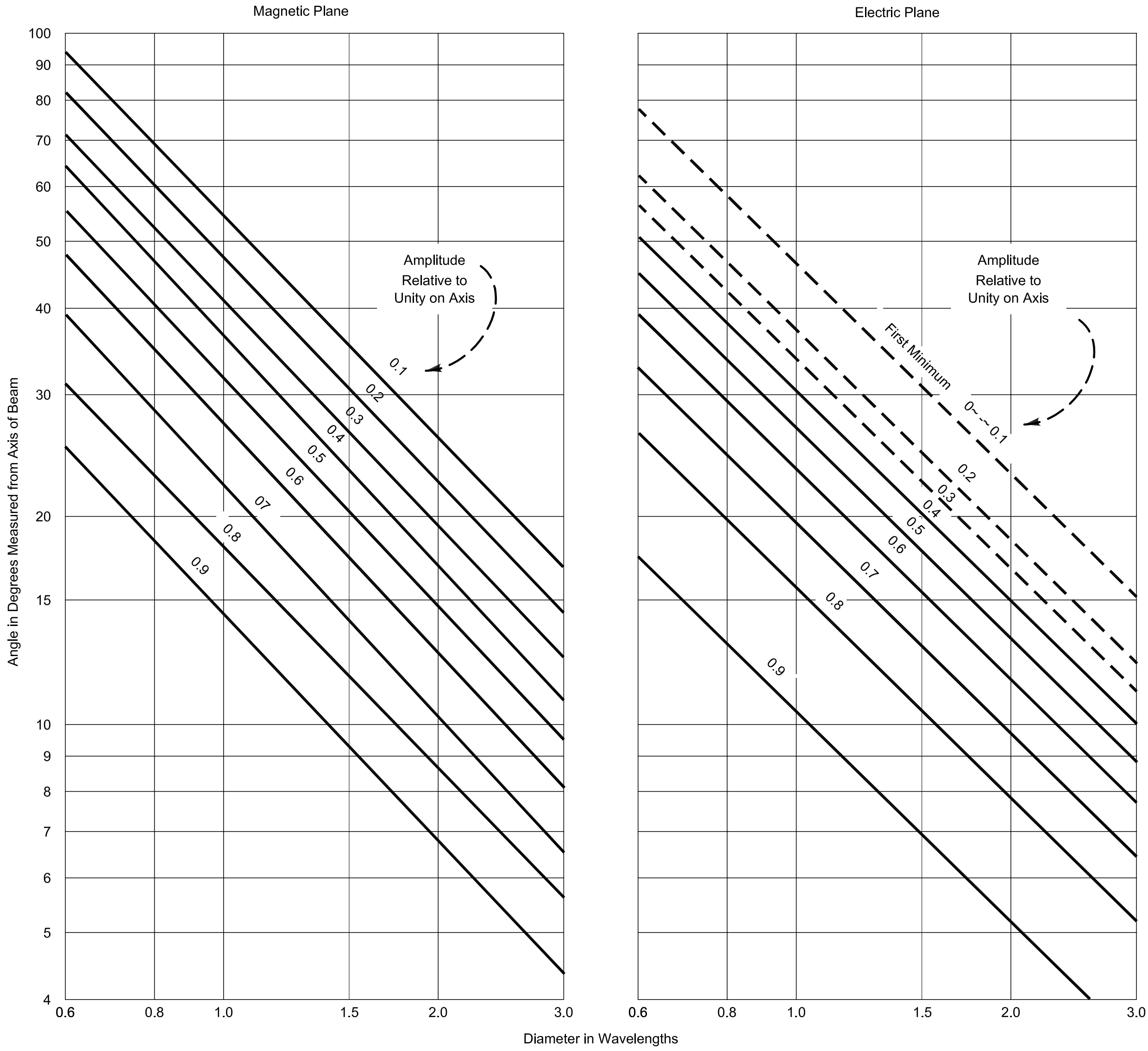

Fig 25—This graph can be used in conjunction with Table 4 for selecting the proper diameter waveguide to illuminate a parabolic reflector.

Table 4

f/D Versus Subtended Angle at Focus of a Parabolic Reflector Antenna

f/D	*Subtended Angle (Deg.)*	*f/D*	*Subtended Angle (Deg.)*
0.20	203	0.65	80
0.25	181	0.70	75
0.30	161	0.75	69
0.35	145	0.80	64
0.40	130	0.85	60
0.45	117	0.90	57
0.50	106	0.95	55
0.55	97	1.00	52
0.60	88		

Taken from graph "f/D vs Subtended Angle at Focus," page 170 of the 1966 *Microwave Engineers' Handbook and Buyers Guide*. Graph courtesy of K. S. Kelleher, Aero Geo Astro Corp, Alexandria, Virginia

Table 5

Circular Waveguide Dish Feeds

Freq. (MHz)	*Inside Diameter Circular Waveguide Range (in.)*
915	8.52-9.84
1296	6.02-6.94
2304	3.39-3.91
3400	2.29-2.65
5800	1.34-1.55
10,250	0.76-0.88

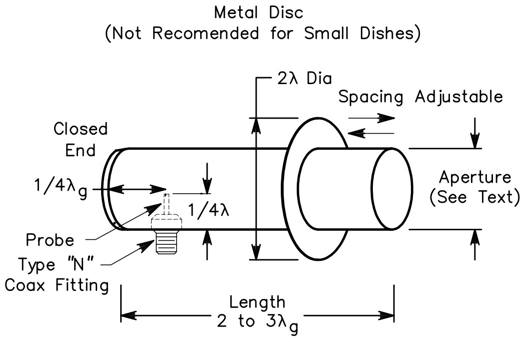

Fig 26—Details of a circular waveguide feed.

suggested. The lower frequency limit (longer dimension) is dictated by proximity to the cutoff frequency. The higher frequency limit (shorter dimension) is dictated by higher order waves. See **Table 5** for recommended inside diameter dimensions for the 902- to 10,000-MHz amateur bands.

The probe that excites the waveguide and makes the transition from coaxial cable to waveguide is $^1/_4\ \lambda$ long and spaced from the closed end of the guide by $^1/_4$ guide wavelength. The length of the feed should be two to three guide wavelengths. The latter is preferred if a second probe is to be mounted for polarization change or for polaplexer work where duplex communication (simultaneous transmission and reception) is possible because of the isolation between two properly located and oriented probes. The second probe for polarization switching or polaplexer work should be spaced $^3/_4$ guide wavelength from the closed end and mounted at right angles to the first probe.

The feed aperture is located at the focal point of the dish and aimed at the center of the reflector. The feed mounts should permit adjustment of the aperture either side of the focal point and should present a minimum of blockage to the reflector. Correct distance to the dish center places the focal point about 1 inch inside the feed aperture. The use of a nonmetallic support minimizes blockage. PVC pipe, fiberglass and Plexiglas are commonly used materials. A simple test by placing a material in a microwave oven reveals if it is satisfactory up to 2450 MHz. PVC pipe has tested satisfactorily and appears to work well at 2300 MHz. A simple, clean looking mount for a 4-foot dish with 18 inches focal length, for example, can be made by mounting a length of 4-inch PVC pipe using a PVC flange at the center of the dish. At 2304 MHz the circular feed is approximately 4 inches ID, making a snug fit with the PVC pipe. Precautions should be taken to keep rain and small birds from entering the feed.

Never look into the open end of a waveguide when power is applied, or stand directly in front of a dish while transmitting. Tests and adjustments in these areas should be done while receiving or at extremely low levels of transmitter power (less than 0.1 watt). The US Government has set a limit of 10 mW/cm^2 averaged over a 6-minute period as the safe maximum. Other authorities believe even lower levels should be used. Destructive thermal heating of body tissue results from excessive exposure. This heating effect is especially dangerous to the eyes. The accepted safe level of 10 mW/cm^2 is reached in the near field of a parabolic antenna if the level at $2D^2/\lambda$ is 0.242 mW/cm^2. The equation for power density at the far-field boundary is

$$\text{Power density} = \frac{137.8\ P}{D^2}\ \text{mW/cm}^2 \qquad \text{(Eq 7)}$$

where

P = average power in kilowatts
D = antenna diameter in feet
λ = wavelength in feet

New commercial dishes are expensive, but surplus ones can often be purchased at low cost. Some amateurs build theirs, while others modify UHF TV dishes or circular metal snow sleds for the amateur bands. **Fig 27** shows a dish using the homemade feed just described.

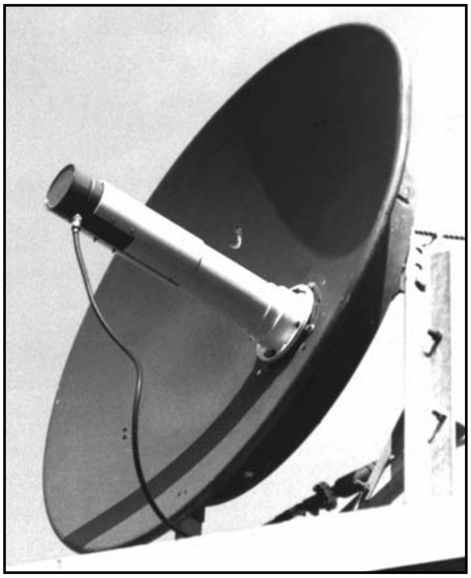

Fig 27—Coffee-can 2304 MHz feed described in text and Fig 26 mounted on a 4-ft dish.

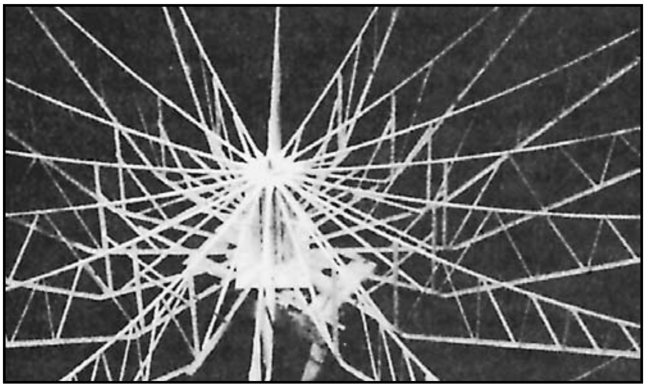

Fig 29—Detailed look at the hub assembly for the ZL1BJQ dish. Most of the structural members are made from ¾-inch T section.

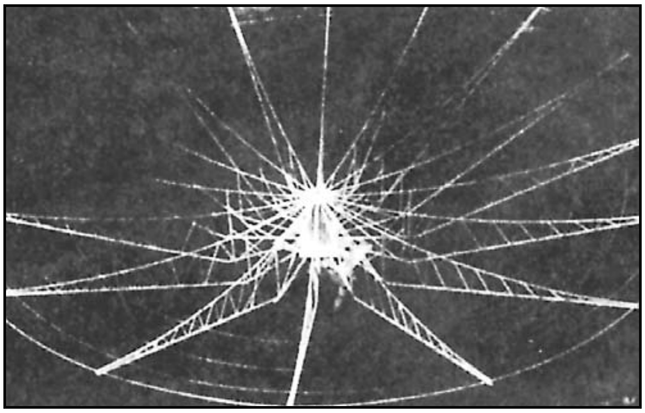

Fig 28—Aluminum framework for a 23-foot dish under construction by ZL1BJQ.

Photos showing a highly ambitious dish project under construction by ZL1BJQ appear in **Figs 28** and **29**. Practical details for constructing this type of antenna are given in Chapter 19. Dick Knadle, K2RIW, described modern UHF antenna test procedures in February 1976 *QST* (see Bibliography). Also see Chapter 19.

OMNIDIRECTIONAL ANTENNAS FOR VHF AND UHF

Local work with mobile stations requires an antenna with wide coverage capabilities. Most mobile work is on FM, and the polarization used with this mode is generally vertical. Some simple vertical systems are described below. Additional material on antennas of this type is presented in Chapter 16, Mobile and Maritime Antennas.

Ground-plane Antennas for 144, 222 and 440 MHz

For the FM operator living in the primary coverage area of a repeater, the ease of construction and low cost of a $^1/_4$ λ ground-plane antenna make it an ideal choice. Three different types of construction are detailed in Figs 30 through 43; the choice of construction method depends upon the materials at hand and the desired style of antenna mounting.

The 144-MHz model shown in **Fig 30** uses a flat piece of sheet aluminum, to which radials are connected with machine screws. A 45° bend is made in each of the radials. This bend can be made with an ordinary bench vise. An SO239 chassis connector is mounted at the center of the aluminum plate with the threaded part of the connector facing down. The vertical portion of the

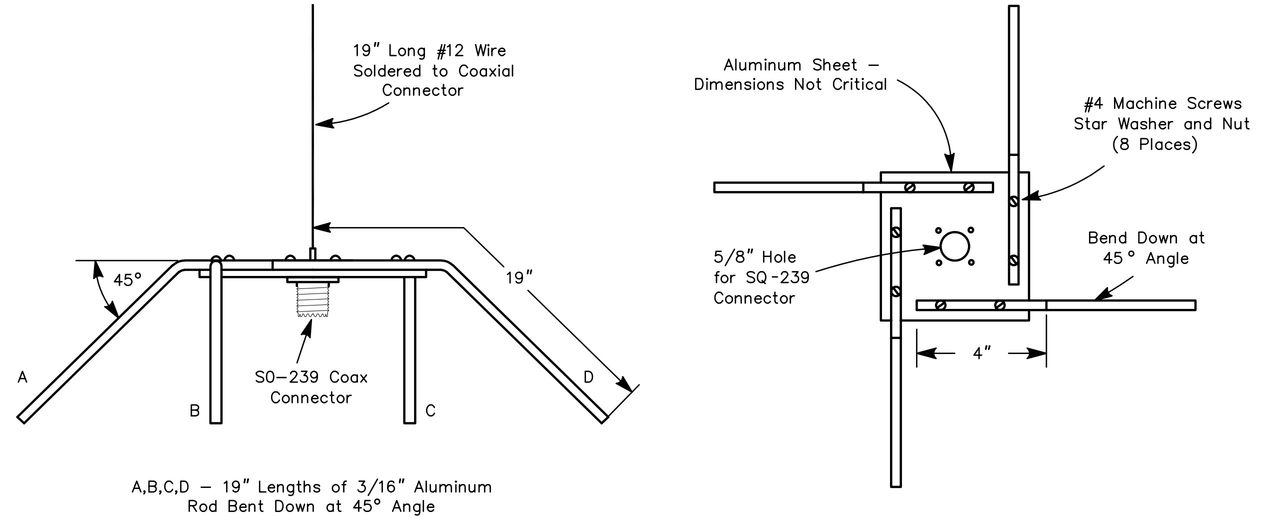
19″ Long #12 Wire
Soldered to Coaxial
Connector
45°
19″
A
B
S0−239 Coax
Connector
C
D
A,B,C,D — 19″ Lengths of 3/16″ Aluminum
Rod Bent Down at 45° Angle
Aluminum Sheet —
Dimensions Not Critical
#4 Machine Screws
Star Washer and Nut
(8 Places)
5/8″ Hole
for SQ-239
Connector
Bend Down at
45° Angle
4″

Fig 30—These drawings illustrate the dimensions for the 144-MHz ground-plane antenna. The radials are bent down at a 45° angle.

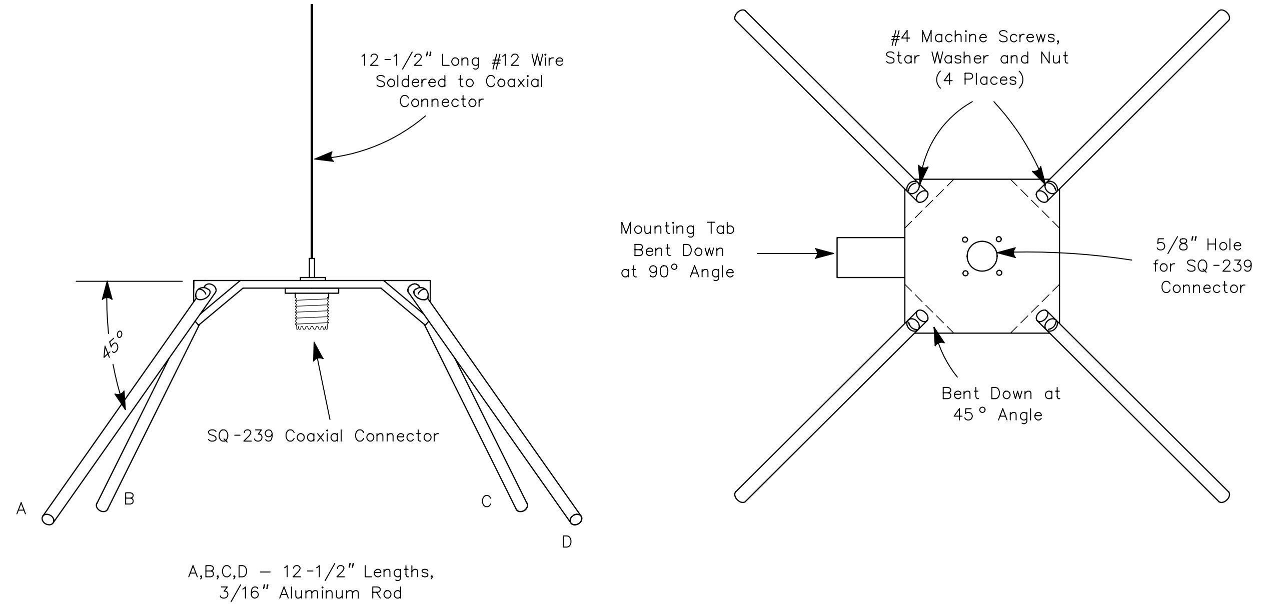
12-1/2″ Long #12 Wire
Soldered to Coaxial
Connector
45°
SQ-239 Coaxial Connector
A
B
C
D
A,B,C,D — 12-1/2″ Lengths,
3/16″ Aluminum Rod
#4 Machine Screws,
Star Washer and Nut
(4 Places)
Mounting Tab
Bent Down
at 90° Angle
5/8″ Hole
for SQ-239
Connector
Bent Down at
45° Angle

Fig 31—Dimensional information for the 222-MHz ground-plane antenna. Lengths for A, B, C and D are the total distances measured from the center of the SO-239 connector. The corners of the aluminum plate are bent down at a 45° angle rather than bending the aluminum rod as in the 144-MHz model. Either method is suitable for these antennas.

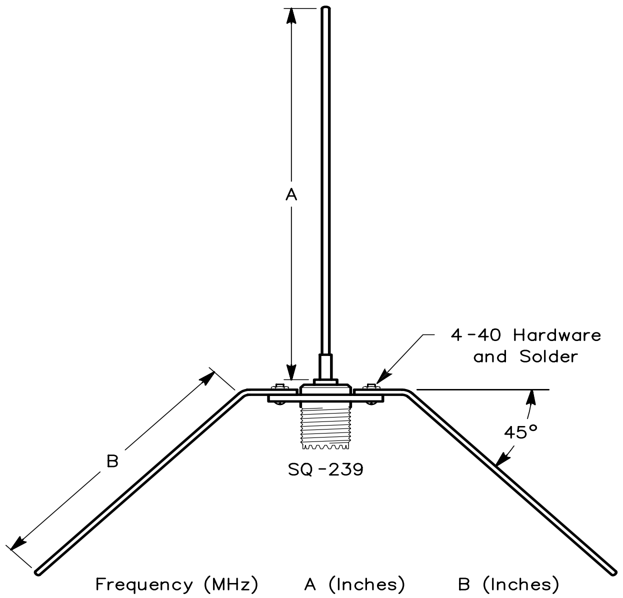

Frequency (MHz)	A (Inches)	B (Inches)
146	19-5/16"	18-11/16"
225	12-5/8"	12"
445	6-3/8"	5-3/4"

Fig 32—Simple ground-plane antenna for the 144-, 222- and 440-MHz bands. The vertical element and radials are $^{3}/_{32}$- or $^{1}/_{16}$-in. brass welding rod. Although $^{3}/_{32}$-in. rod is preferred for the 144-MHz antenna, #10 or #12 copper wire can also be used.

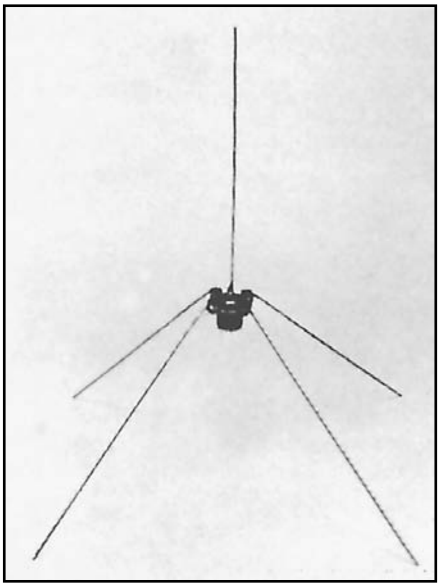

Fig 33—A 440-MHz ground-plane constructed using only an SO-239 connector, no. 4-40 hardware and $^{1}/_{16}$-in. brass welding rod.

antenna is made of #12 copper wire soldered directly to the center pin of the SO-239 connector.

The 222-MHz version, **Fig 31**, uses a slightly different technique for mounting and sloping the radials. In this case the corners of the aluminum plate are bent down at a 45° angle with respect to the remainder of the plate. The four radials are held to the plate with machine screws, lock washers and nuts. A mounting tab is included in the design of this antenna as part of the aluminum base. A compression type of hose clamp could be used to secure the antenna to a mast. As with the 144-MHz version, the vertical portion of the antenna is soldered directly to the SO-239 connector.

A very simple method of construction, shown in **Figs 32** and **33**, requires nothing more than an SO-239 connector and some no. 4-40 hardware. A small loop formed at the inside end of each radial is used to attach the radial directly to the mounting holes of the coaxial connector. After the radial is fastened to the SO-239 with no. 4-40 hardware, a large soldering iron or propane torch is used to solder the radial and the mounting hardware to the coaxial connector. The radials are bent to a 45° angle and the vertical portion is soldered to the center pin to complete the antenna. The antenna can be mounted by passing the feed line through a mast of ¾-inch ID plastic or aluminum tubing. A compression hose clamp can be used to secure the PL-259 connector, attached to the feed line, in the end of the mast. Dimensions for the 144-, 222- and 440-MHz bands are given in Fig 32.

If these antennas are to be mounted outside it is wise to apply a small amount of RTV sealant or similar material around the areas of the center pin of the connector to prevent the entry of water into the connector and coax line.

The J-Pole Antenna

The J-Pole is a half-wave antenna that is end-fed at its bottom. Since the radiator is longer than that of a 1/4-wave ground-plane antenna, the vertical lobe is compressed down toward the horizon and it has about 1.5 dB of gain compared to the ground-plane configuration. The stub-matching section used to transform the high impedance seen looking into a half-wave to 50 Ω coax is shorted at the bottom, making the antenna look like the letter "J," and giving the antenna its name.

Rigid copper tubing, fittings and assorted hardware can be used to make a really rugged J-pole antenna for 2 meters. When copper tubing is used, the entire assembly can be soldered together, ensuring electrical integrity, and making the whole antenna weatherproof. This

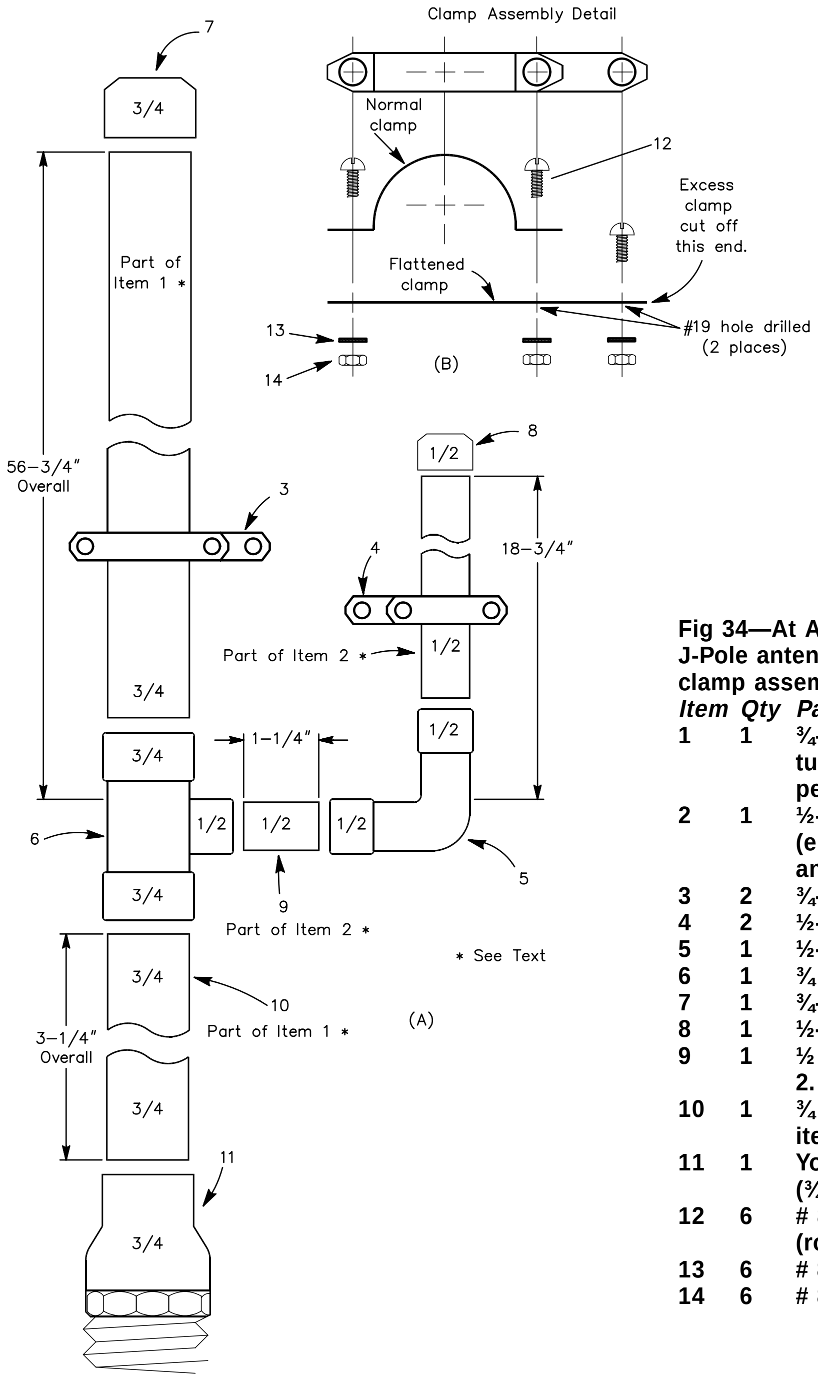
Clamp Assembly Detail
Normal clamp
12
Excess clamp cut off this end.
Flattened clamp
13
14
(B)
#19 hole drilled (2 places)
7
3/4
Part of Item 1 *
56–3/4″ Overall
3
8
1/2
18–3/4″
4
Part of Item 2 *
1/2
3/4
3/4
1–1/4″
1/2
1/2
1/2
1/2
6
5
3/4
9
Part of Item 2 *
* See Text
3/4
10
Part of Item 1 *
(A)
3–1/4″ Overall
3/4
11
3/4

Fig 34—At A, exploded assembly diagram of all-copper J-Pole antenna. At B, detail of clamp assemblies. Both clamp assemblies are the same.

Item	***Qty***	***Part or Material Name***
1	**1**	**¾-inch × 10 foot length of rigid copper tubing (enough for 2 antennas, 60 inches per antenna)**
2	**1**	**½-inch × 10 ft length of rigid copper tubing (enough for 6 antennas, 20 inches per antenna)**
3	**2**	**¾-inch copper pipe clamps**
4	**2**	**½-inch copper pipe clamps**
5	**1**	**½-inch copper elbow**
6	**1**	**¾ × ½-inch copper tee**
7	**1**	**¾-inch copper end cap**
8	**1**	**½-inch copper end cap**
9	**1**	**½ × 1¼-inch copper nipple (Make from item 2. See text)**
10	**1**	**¾ × 3 ¼-inch copper nipple (Make from item 1. See text)**
11	**1**	**Your choice of coupling to mast fitting (¾ × 1 inch NPT used at KD8JB)**
12	**6**	**# 8-32 × ½-inch brass machine screws (round, pan, or binder head)**
13	**6**	**# 8 brass flat washers**
14	**6**	**# 8-32 brass hex nuts**

material came from an article by Michael Hood, KD8JB, in *The ARRL Antenna Compendium,Vol. 4*.

No special hardware or machined parts are used in this antenna, nor are insulating materials needed, since the antenna is always at dc ground. Best of all, even if the parts aren't on sale, the antenna can be built for less than $15. If you only build one antenna, you'll have enough tubing left over to make most of a second antenna.

Construction

Copper and brass is used exclusively in this antenna. These metals get along together, so dissimilar metal corrosion is eliminated. Both metals solder well, too. See **Fig 34**. Cut the copper tubing to the lengths indicated. Item 9 is a 1$^{1}/_{4}$-inch nipple cut from the 20-inch length of $^{1}/_{2}$-inch tubing. This leaves 18$^{3}/_{4}$ inches for the $^{1}/_{4}$-matching stub. Item 10 is a 3$^{1}/_{4}$-inch long nipple cut from the 60-inch length of $^{3}/_{4}$-inch tubing. The $^{3}/_{4}$-wave element should measure 56$^{3}/_{4}$-inches long. Remove burrs from the ends of the tubing after cutting, and clean the mating surfaces with sandpaper, steel wool, or emery cloth.

After cleaning, apply a very thin coat of flux to the mating elements and assemble the tubing, elbow, tee, end caps and stubs. Solder the assembled parts with a propane torch and rosin-core solder. Wipe off excess solder with a damp cloth, being careful not to burn yourself. The copper tubing will hold heat for a long time after you've finished soldering. After soldering, set the assembly aside to cool.

Flatten one each of the $^{1}/_{2}$-inch and $^{3}/_{4}$-inch pipe clamps. Drill a hole in the flattened clamp as shown in Fig 34A. Assemble the clamps and cut off the excess metal from the flattened clamp using the unmodified clamp as a template. Disassemble the clamps.

Assemble the $^{1}/_{2}$-inch clamp around the $^{1}/_{4}$-wave element and secure with two of the screws, washers, and nuts as shown in Fig 34B. Do the same with the $^{3}/_{4}$-inch clamp around the $^{3}/_{4}$-wave element. Set the clamps initially to a spot about 4 inches above the bottom of the "J" on their respective elements. Tighten the clamps only finger tight, since you'll need to move them when tuning.

Tuning

The J-Pole can be fed directly from 50-Ω coax through a choke balun (3 turns of the feed coax rolled into a coil about 8 inches in diameter and held together with electrical tape). Before tuning, mount the antenna vertically, about 5 to 10 feet from the ground. A short TV mast on a tripod works well for this purpose. When tuning VHF antennas, keep in mind that they are sensitive to nearby objects—such as your body. Attach the feed line to the clamps on the antenna, and make sure all the nuts and screws are at least finger tight. It really doesn't matter to which element (¾-wave element or stub) you attach the coaxial center lead. The author has done it both ways with no variation in performance. Tune the antenna by moving the two feed-point clamps equal distances a small amount each time until the SWR is minimum at the desired frequency. The SWR will be close to 1:1.

Final Assembly

The final assembly of the antenna will determine its long-term survivability. Perform the following steps with care. After adjusting the clamps for minimum SWR, mark the clamp positions with a pencil and then remove the feed line and clamps. Apply a very thin coating of flux to the inside of the clamp and the corresponding surface of the antenna element where the clamp attaches. Install the clamps and tighten the clamp screws.

Solder the feed line clamps where they are attached to the antenna elements. Now, apply a small amount of solder around the screw heads and nuts where they contact the clamps. Don't get solder on the screw threads! Clean away excess flux with a non-corrosive solvent. After final assembly and erecting/mounting the antenna in the desired location, attach the feed line and secure with the remaining washer and nut. Weather-seal this joint with RTV. Otherwise, you may find yourself repairing the feed line after a couple years.

On-the-Air Performance

Years ago, prior to building the first J-Pole antenna for this station, the author used a standard $^{1}/_{4}$-wave ground plane vertical antenna. While he had no problem working various repeaters around town with a $^{1}/_{4}$-wave antenna, simplex operation left a lot to be desired. The J-Pole performs just as well as a Ringo Ranger, and significantly better than the $^{1}/_{4}$-wave ground-plane vertical.

Practical 6-Meter Yagis

Boom length often proves to be the deciding factor when one selects a Yagi design. **Table 6** shows three 6-meter Yagis designed for convenient boom lengths (6, 12 and 22 feet). The 3-element, 6-foot boom design has 8.0 dBi gain in free space; the 12 foot boom, 5-element version has 10.1 dBi gain, and the 22-foot, 7 element Yagi has a gain of 11.3 dBi. All antennas exhibit better than 22 dB front-to-rear ratio and cover 50 to 51 MHz with better than 1.7:1 SWR.

Half-element lengths and spacings are given in the table. Elements can be mounted to the boom as shown in **Fig 35**. Two muffler clamps hold each aluminum plate to the boom, and two U bolts fasten each element to the plate, which is 0.25 inches thick and 4 × 4 inches square. Stainless steel is the best choice for hardware, however, galvanized hardware can be substituted. Automotive muffler clamps do not work well in this application, because they are not galvanized and quickly rust once exposed to the weather. Please note that the element lengths shown in Table 6 are half the overall element lengths. See page 20-7 to 20-11 in Chapter 20 for practical details of telescoping aluminum elements.

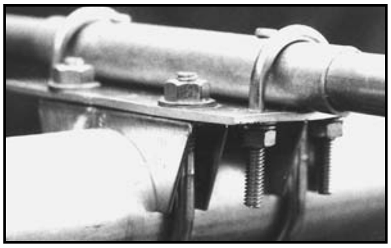

Fig 35—The element to boom clamp. U bolts are used to hold the element to the plate, and 2-in. galvanized muffler clamps hold the plates to the boom.

The driven element is mounted to the boom on a Bakelite or G-10 fiberglass plate of similar dimension to the other mounting plates. A 12-inch piece of Plexiglas rod is inserted into the driven element halves. The Plexiglas allows the use of a single clamp on each side of the element and also seals the center of the elements against moisture. Self-tapping screws are used for electrical connection to the driven element.

Refer to **Fig 36** for driven-element and hairpin match details. A bracket made from a piece of aluminum is used to mount the three SO239 connectors to the driven element plate. A 4:1 transmission-line balun connects the two element halves, transforming the 200 Ω resistance at the hairpin match to 50 Ω at the center connector. Note

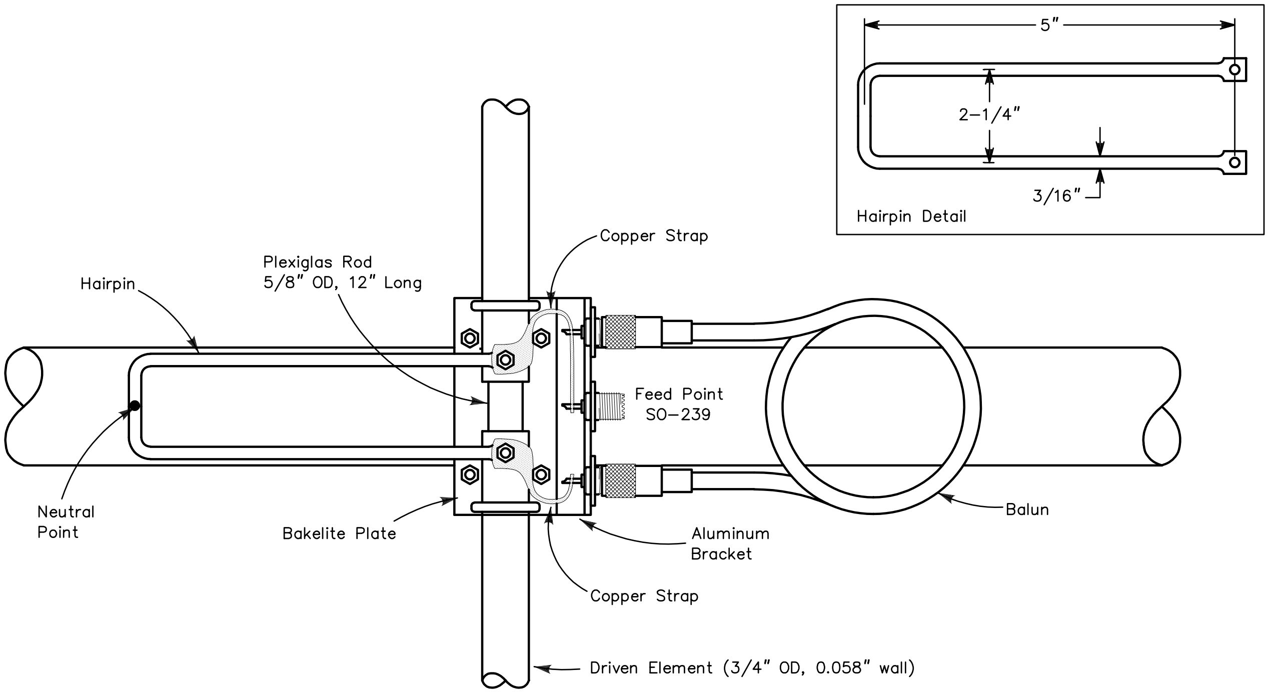

Fig 36—This shows how the driven element and feed system are attached to the boom. The phasing line is coiled and taped to the boom. The center of the hairpin loop may be connected to the boom electrically and mechanically if desired.
Phasing-line lengths:
For cable with 0.80 velocity factor – 7 ft, 10$^{3}/_{8}$ in.
For cable with 0.66 velocity factor – 6 ft, 5$^{3}/_{4}$ in.

Table 6
Optimized 6-Meter Yagi Designs

	Spacing Between Elements inches	*Seg 1 OD* Length inches*	*Seg2 OD* Length inches*	*Midband Gain F/R*
306-06				
OD		0.750	0.625	
Refl.	0	36	23.500	7.9 dBi
D.E.	24	36	16.000	27.2 dB
Dir. 1	66	36	15.500	
506-12				
OD		0.750	0.625	
Refl.	0	36	24.000	10.1 dBi
D.E.	24	36	17.125	24.7 dB
Dir. 1	12	36	19.375	
Dir. 2	44	36	18.250	
Dir. 3	58	36	15.375	
706-22				
OD		0.750	0.625	
Refl.	0	36	25.000	11.3 dBi
D.E.	27	36	17.250	29.9 dB
Dir. 1	16	36	18.500	
Dir. 2	51	36	15.375	
Dir. 3	54	36	15.875	
Dir. 4	53	36	16.500	
Dir. 5	58	36	12.500	

*See pages 20-7 to 20-11 for telescoping aluminum tubing details.

that the electrical length of the balun is λ/2, but the physical length will be shorter due to the velocity factor of the particular coaxial cable used. The hairpin is connected directly across the element halves. The exact center of the hairpin is electrically neutral and should be fastened to the boom. This has the advantage of placing the driven element at dc ground potential.

The hairpin match requires no adjustment as such. However, you may have to change the length of the driven element slightly to obtain the best match in your preferred portion of the band. Changing the driven-element length will not adversely affect antenna performance. *Do not adjust the lengths or spacings of the other elements—they are optimized already*. If you decide to use a gamma match, add 3 inches to each side of the driven element lengths given in the table for all antennas.

High-Performance Yagis for 144, 222 and 432 MHz

This construction information is presented as an introduction to the three high-performance VHF/UHF Yagis that follow. All were designed and built by Steve Powlishen, K1FO. For years the design of long Yagi antennas seemed to be a mystical black art. The problem of simultaneously optimizing 20 or more element spacings and element lengths presented an almost unsolvable set of simultaneous equations. With the unprecedented increase in computer power and widespread availability of antenna analysis software, we are now able to quickly examine many Yagi designs and determine which approaches work and which designs to avoid.

At 144 MHz and above, most operators desire Yagi antennas two or more wavelengths in length. This length (2λ) is where most classical designs start to fall apart in terms of gain per boom length, bandwidth and pattern quality. Extensive computer and antenna range analysis has proven that the best possible design is a Yagi that has both varying element spacings and varying element lengths.

This design approach starts with closely spaced directors. The director spacings gradually increase until a constant spacing of about 0.4 λ is reached. Conversely, the director lengths start out longest with the first director and decrease in length in a decreasing rate of change until they are virtually constant in length. This method of construction results in a wide gain bandwidth. A bandwidth of 7% of the center frequency at the –1 dB forward-gain points is typical for these Yagis even when they are longer than 10 λ. The log-taper design also reduces the rate of change in driven-element impedance vs frequency. This allows the use of simple dipole driven elements while still obtaining acceptable driven-element SWR over a wide frequency range. Another benefit is that the resonant frequency of the Yagi changes very little as the boom length is increased.

The driven-element impedance also changes moderately with boom length. The tapered approach creates a Yagi with a very clean radiation pattern. Typically, first side lobe levels of ~17 dB in the E plane, ~15 dB in the H plane, and all other lobes at ~20 dB or more are possible on designs from 2 λ to more than 14 λ.

The actual rate of change in element lengths is determined by the diameter of the elements (in wavelengths). The spacings can be optimized for an individual boom length or chosen as a best compromise for most boom lengths.

The gain of long Yagis has been the subject of much debate. Recent measurements and computer analysis by both amateurs and professionals indicates that given an optimum design, doubling a Yagi's boom length will result in a maximum theoretical gain increase of about 2.6 dB. In practice, the real gain increase may be less because of escalating resistive losses and the greater possibility of construction error. **Fig 37** shows the maximum possible gain per boom length expressed in decibels, referenced to an isotropic radiator. The actual number of directors does not play an important part in determining the gain vs boom length as long as a reasonable number of directors are used. The use of more directors per boom length will normally give a wider gain bandwidth, however, a point exists where too many directors will adversely affect all performance aspects.

While short antennas (< 1.5 λ) may show increased gain with the use of quad or loop elements, long Yagis (> 2 λ) will not exhibit measurably greater forward gain or pattern integrity with loop-type elements. Similarly, loops used as driven elements and reflectors will not significantly change the properties of a long log-taper Yagi. Multiple-dipole driven-element assemblies will also not result in any significant gain increase per given boom length when compared to single-dipole feeds.

Once a long-Yagi director string is properly tuned, the reflector becomes relatively non critical. Reflector spacings between 0.15 λ and 0.2 λ are preferred. The spacing can be chosen for best pattern and driven element impedance. Multiple-reflector arrangements will not significantly increase the forward gain of a Yagi which has its directors properly optimized for forward gain. Many multiple-reflector schemes such as tri-reflectors and corner reflectors have the disadvantage of lowering the driven element impedance compared to a single optimum-length reflector. The plane or grid reflector, shown in **Fig 38**, may however reduce the intensity of unwanted rear lobes. This can be used to reduce noise pickup on EME or satellite arrays. This type of reflector will usually increase the driven-element impedance compared to a single reflector. This sometimes makes driven-element matching easier. Keep in mind that even for EME, a plane reflector will add considerable wind load and weight for

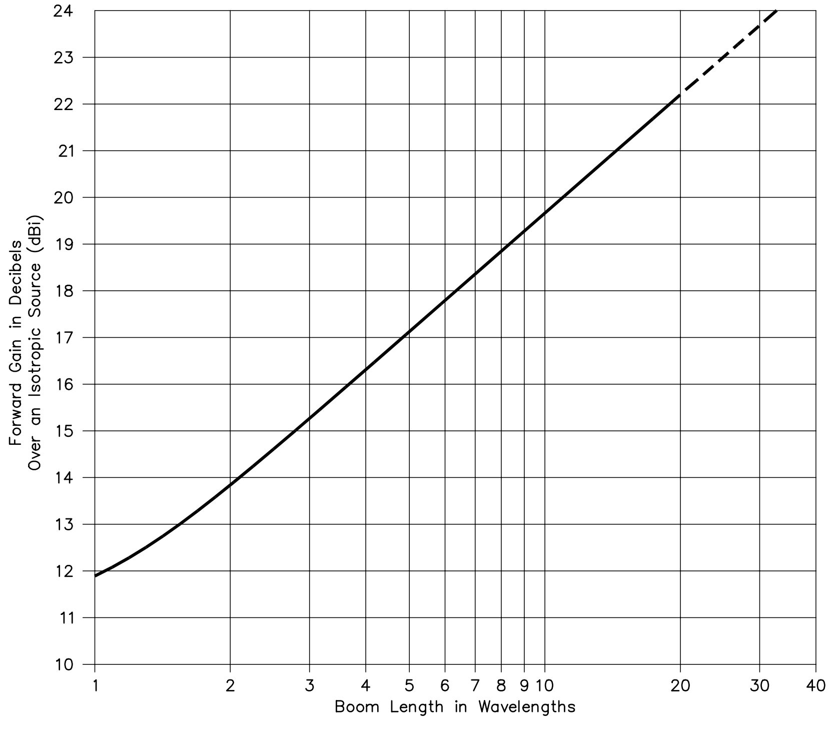

Fig 37—This chart shows maximum gain per boom length for optimally designed long Yagi antennas.

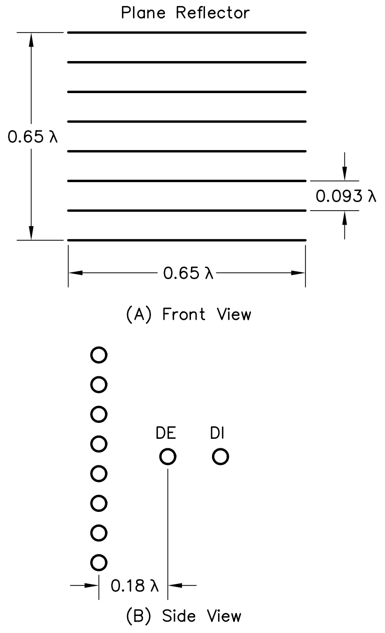

Fig 38—Front and side views of a plane-reflector antenna.

only a few tenths of a decibel of receive signal-to-noise improvement.

Yagi Construction

Normally, aluminum tubing or rod is used for Yagi elements. Hard-drawn enamel-covered copper wire can also be used on Yagis above 420 MHz. Resistive losses are inversely proportional to the square of the element diameter and the square root of its conductivity.

Element diameters of less than $^3/_{16}$ inch or 4 mm should not be used on any band. The size should be chosen for reasonable strength. Half-inch diameter is suitable for 50 MHz, $^3/_{16}$ to $^3/_8$ inch for 144 MHz and $^3/_{16}$ inch is recommended for the higher bands. Steel, including stainless steel and unprotected brass or copper wire, should not be used for elements.

Boom material may be aluminum tubing, either square or round. High-strength aluminum alloys such as 6061-T6 or 6063-T651 offer the best strength-to-weight advantages. Fiberglass poles have been used (where available as surplus). Wood is a popular low-cost boom material. The wood should be well seasoned and free from knots. Clear pine, spruce and Douglas fir are often used. The wood should be well treated to avoid water absorption and warping.

Elements may be mounted insulated or uninsulated, above or through the boom. Mounting uninsulated elements through a metal boom is the least desirable method unless the elements are welded in place. The Yagi elements will oscillate, even in moderate winds. Over several years this element oscillation will work open the boom holes. This will allow the elements to move in the boom. This will create noise (in your receiver) when the wind blows, as the element contact changes. Eventually the element-to-boom junction will corrode (aluminum oxide is a good insulator). This loss of electrical contact between the boom and element will reduce the boom's effect and change the resonant frequency of the Yagi.

Noninsulated elements mounted above the boom will perform fine as long as a good mechanical connection is made. Insulating blocks mounted above the boom will also work, but they require additional fabrication. One of the most popular construction methods is to mount the elements through the boom using insulating shoulder washers. This method is lightweight and durable. Its main disadvantage is difficult disassembly, making this method of limited use for portable arrays.

If a conductive boom is used, element lengths must be corrected for the mounting method used. The amount of correction is dependent upon the boom diameter in

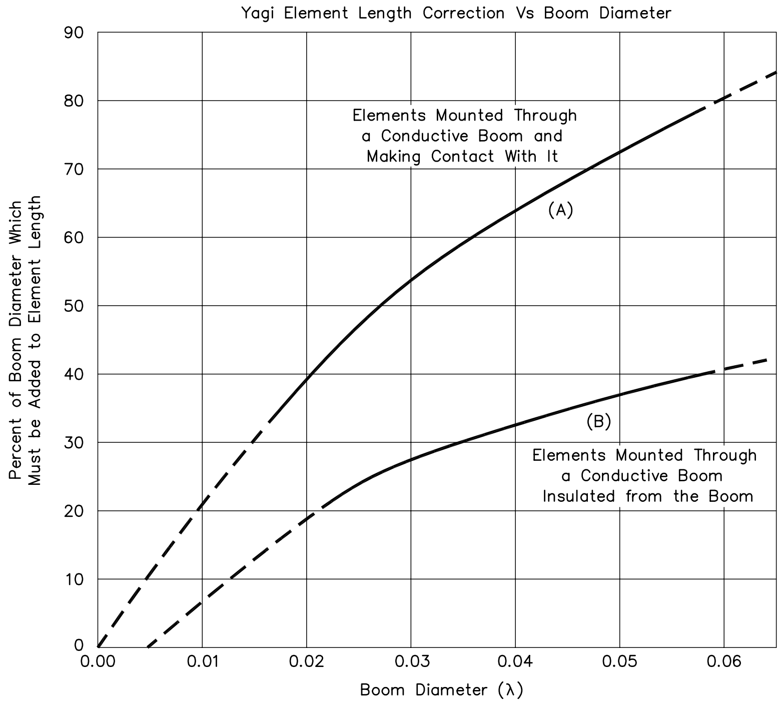

Fig 39—Yagi element correction vs boom diameter. Curve A is for elements mounted through a round or square conductive boom, with the elements in mechanical contact with the boom. Curve B is for insulated elements mounted through a conductive boom, and for elements mounted on top of a conductive boom (elements make electrical contact with the boom). The patterns were corrected to computer simulations to determine Yagi tuning. The amount of element correction is not affected by element diameter.

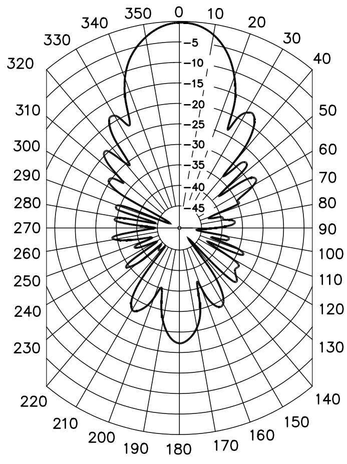

Fig 40—Measured E-plane pattern for the 22-element Yagi. Note: This antenna pattern is drawn on a linear dB grid, rather than on the standard ARRL log-periodic grid, to emphasize low sidelobes.

wavelengths. See **Fig 39**. Elements mounted through the boom and not insulated require the greatest correction. Mounting on top of the boom or through the boom on insulated shoulder washers requires about half of the through-the-boom correction. Insulated elements mounted at least one element diameter above the boom require no correction over the free-space length.

The three following antennas have been optimized for typical boom lengths on each band.

A HIGH-PERFORMANCE 432-MHz YAGI

This 22-element, 6.1-λ, 432-MHz Yagi was originally designed for use in a 12-Yagi EME array built by K1FO. A lengthy evaluation and development process preceded its construction. Many designs were considered and then analyzed on the computer. Next, test models were constructed and evaluated on a home-made antenna range. The resulting design is based on WlEJ's computer-optimized spacings.

The attention paid to the design process has been worth the effort. The 22-element Yagi not only has exceptional forward gain (17.9 dBi), but has an unusually clean radiation pattern. The measured E-plane pattern is shown in **Fig 40**. Note that a 1-dB-per-division axis is used to show pattern detail. A complete description of the design process and construction methods appears in December 1987 and January 1988 *QST*.

Table 7

Specifications for 432-MHz Yagi Family

No. of Ele.	*Boom length(λ)*	*Gain (dBi)**	*F/B ratio (dB)*	*DE imped (Ω)*	*Beamwidth E/H (°)*	*Stacking E/H (inches)*
15	3.4	15.67	21	23	30/32	53/49
16	3.8	16.05	19	23	29/31	55/51
17	4.2	16.45	20	27	28/30	56/53
18	4.6	16.8	25	32	27/29	58/55
19	4.9	17.1	25	30	26/28	61/57
20	5.3	17.4	21	24	25.5/27	62/59
21	5.7	17.65	20	22	25/26.5	63/60
22	6.1	17.9	22	25	24/26	65/62
23	6.5	18.15	27	30	23.5/25	67/64
24	6.9	18.35	29	29	23/24	69/66
25	7.3	18.55	23	25	22.5/23.5	71/68
26	7.7	18.8	22	22	22/23	73/70
27	8.1	19.0	22	21	21.5/22.5	75/72
28	8.5	19.20	25	25	21/22	77/75
29	8.9	19.4	25	25	20.5/21.5	79/77
30	9.3	19.55	26	27	20/21	80/78
31	9.7	19.7	24	25	19.6/20.5	81/79
32	10.2	19.8	23	22	19.3/20	2/80
33	10.6	9.9	23	23	19/19.5	83/81
34	11.0	20.05	25	22	18.8/19.2	84/82
35	11.4	20.2	27	25	18.5/19.0	85/83
36	11.8	20.3	27	26	18.3/18.8	86/84
37	12.2	20.4	26	26	18.1/18.6	87/85
38	12.7	20.5	25	25	18.9/18.4	88/86
39	13.1	20.6	25	23	18.7/18.2	89/87
40	13.5	20.8	26	21	17.5/18	90/88

*Gain is approximate real gain based on gain measurements made on six different-length Yagis.

Like other log-taper Yagi designs, this one can easily be adapted to other boom lengths. Versions of this Yagi have been built by many amateurs. Boom lengths ranged between 5.3 λ (20 elements) and 12.2 λ (37 elements).

The size of the original Yagi (169 inches long, 6.1 λ) was chosen so the antenna could be built from small-diameter boom material ($^{7}/_{8}$-inch and 1 inch round 6061-T6 aluminum) and still survive high winds and ice loading. The 22-element Yagi weighs about 3.5 pounds and has a wind load of approximately 0.8 square feet. This allows a high-gain EME array to be built with manageable wind load and weight. This same low wind load and weight lets the tropo operator add a high-performance 432-MHz array to an existing tower without sacrificing antennas on other bands.

Table 7 lists the gain and stacking specifications for the various length Yagis. The basic Yagi dimensions are shown in **Table 8**. These are free-space element lengths for $^{3}/_{16}$-inch-diameter elements. Boom corrections for the element mounting method must be added in. The element-length correction column gives the length that must be added to keep the Yagi's center frequency optimized for use at 432 MHz. This correction is required to use the same spacing pattern over a wide range of boom lengths. Although any length Yagi will work well, this design is at its best when made with 18 elements or more (4.6 λ). Element material of less than $^{3}/_{16}$-inch diameter is not recommended because resistive losses will reduce the gain by about 0.1 dB, and wet-weather performance will be worse.

Quarter-inch-diameter elements could be used if all elements are shortened by 3 mm. The element lengths are intended for use with a slight chamfer (0.5 mm) cut into the element ends. The gain peak of the array is centered at 437 MHz. This allows acceptable wet-weather performance, while reducing the gain at 432 MHz by only 0.05 dB.

The gain bandwidth of the 22-element Yagi is 31 MHz (at the –1 dB points). The SWR of the Yagi is less than l.4: l between 420 and 440 MHz. **Fig 41** is a network analyzer plot of the driven-element SWR vs frequency. These numbers indicate just how wide the frequency response of a log-taper Yagi can be, even with a simple dipole driven element. In fact, at one antenna gain contest, some ATV operators conducted gain vs frequency measurements from 420 to 440 MHz. The 22-element Yagi beat all entrants including those with so-called broadband feeds.

To peak the Yagi for use on 435 MHz (for satellite use), you may want to shorten all the elements by 2 mm. To peak it for use on 438 MHz (for ATV applications), shorten all elements by 4 mm. If you want to use the Yagi

Table 8

Free-Space Dimensions for 432-MHz Yagi Family

*Element correction is the amount to shorten or lengthen all elements when building a Yagi of that length.
Element lengths are for 3/16-inch diameter material.

Ele. No.	*Element Position (mm from reflector)*	*Element Length (mm)*	*Element Correction**
Refl	0	340	
DE	104	334	
D1	146	315	
D2	224	306	
D3	332	299	
D4	466	295	
D5	622	291	
D6	798	289	
D7	990	287	
D8	1196	285	
D9	1414	283	
D10	1642	281	–2
D11	1879	279	–2
D12	2122	278	–2
D13	2373	277	–2
D14	2629	276	–2
D15	2890	275	–1
D16	3154	274	–1
D17	3422	273	–1
D18	3693	272	0
D19	3967	271	0
D20	4242	270	0
D21	4520	269	0
D22	4798	269	0
D23	5079	268	0
D24	5360	268	+1
D25	5642	267	+1
D26	5925	267	+1
D27	6209	266	+1
D28	6494	266	+1
D29	6779	265	+2
D30	7064	265	+2
D31	7350	264	+2
D32	7636	264	+2
D33	7922	263	+2
D34	8209	263	+2
D35	8496	262	+2
D36	8783	262	+2
D37	9070	261	+3
D38	9359	261	+3

on FM between 440 MHz and 450 MHz, shorten all the elements by 10 mm. This will provide 17.6 dBi gain at 440 MHz, and 18.0 dBi gain at 450 MHz. The driven element may have to be adjusted if the element lengths are shortened.

Although this Yagi design is relatively broadband, it is suggested that close attention be paid to copying the design exactly as built. Metric dimensions are used because they are convenient for a Yagi sized for 432 MHz.

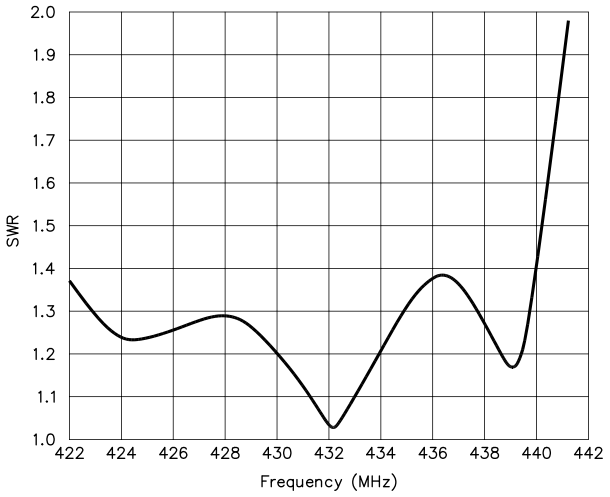

Fig 41—SWR performance of the 22-element Yagi in dry weather.

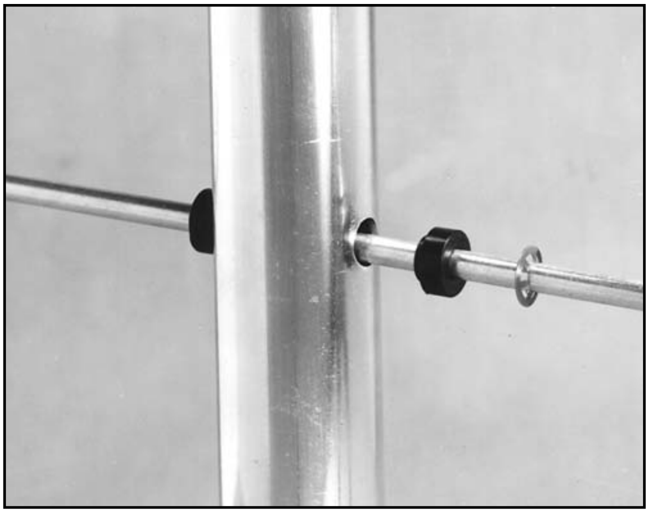

Fig 42—Element-mounting detail. Elements are mounted through the boom using plastic insulators. Stainless steel push-nut retaining rings hold the element in place.

Element holes should be drilled within ±2 mm. Element lengths should be kept within ±0.5 mm. Elements can be accurately constructed if they are first rough cut with a hack saw and then held in a vise and filed to the exact length.

The larger the array, the more attention you should pay to making all Yagis identical. Elements are mounted on shoulder insulators and run through the boom (see **Fig 42**). The element retainers are stainless-steel push nuts. These are made by several companies, including Industrial Retaining Ring Co in Irvington, New Jersey, and AuVeco in Ft Mitchell, Kentucky. Local industrial hardware distributors can usually order them for you. The element insulators

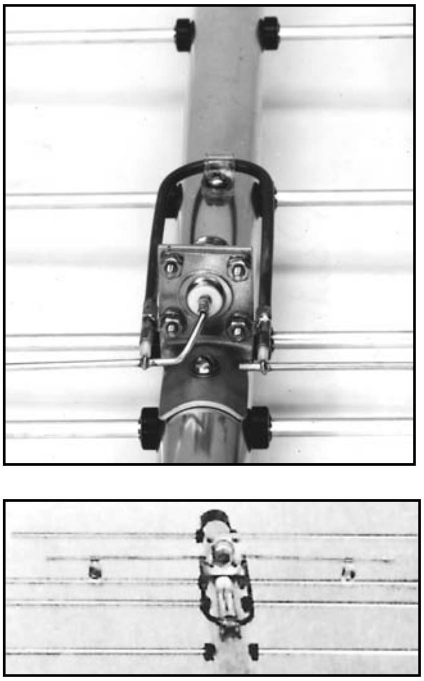

Fig 43—Several views of the driven element and T match.

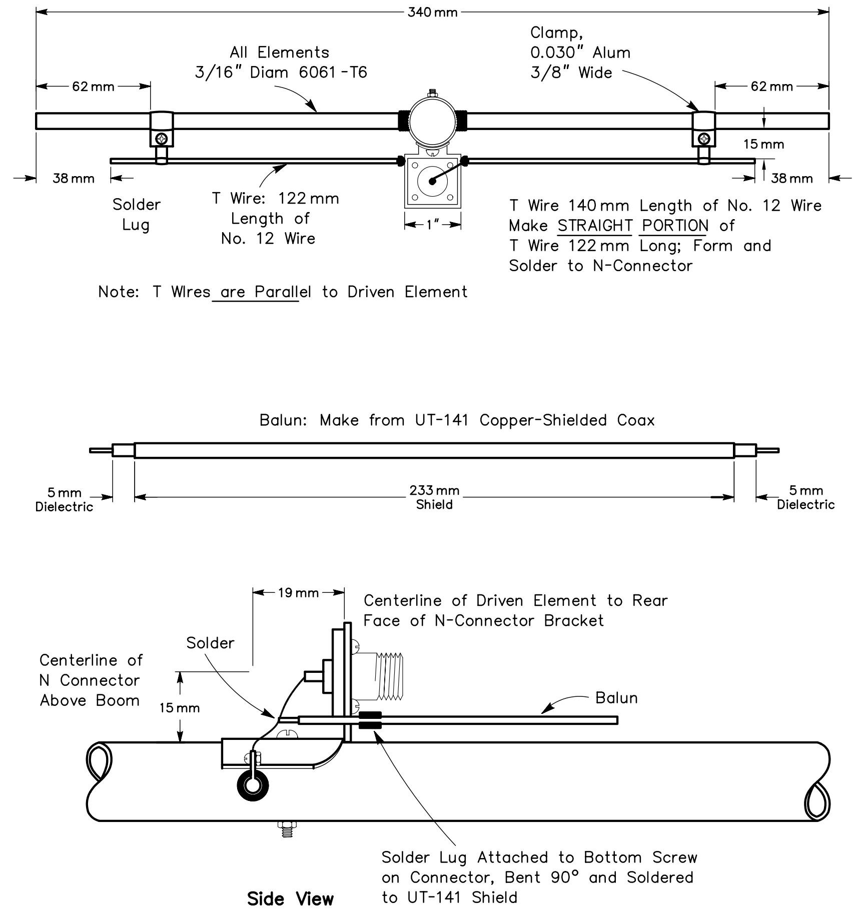
340 mm
All Elements
3/16″ Diam 6061-T6
Clamp,
0.030″ Alum
3/8″ Wide
62 mm
62 mm
15 mm
38 mm
38 mm
Solder
Lug
T Wire: 122 mm
Length of
No. 12 Wire
1″
T Wire 140 mm Length of No. 12 Wire
Make STRAIGHT PORTION of
T Wire 122 mm Long; Form and
Solder to N-Connector
Note: T Wires are Parallel to Driven Element
Balun: Make from UT-141 Copper-Shielded Coax
5 mm
Dielectric
233 mm
Shield
5 mm
Dielectric
19 mm
Centerline of Driven Element to Rear
Face of N-Connector Bracket
Solder
Centerline of
N Connector
Above Boom
15 mm
Balun
Solder Lug Attached to Bottom Screw
on Connector, Bent 90° and Soldered
to UT-141 Shield
Side View

Fig 44—Details of the driven element and T match for the 22-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna. See text.

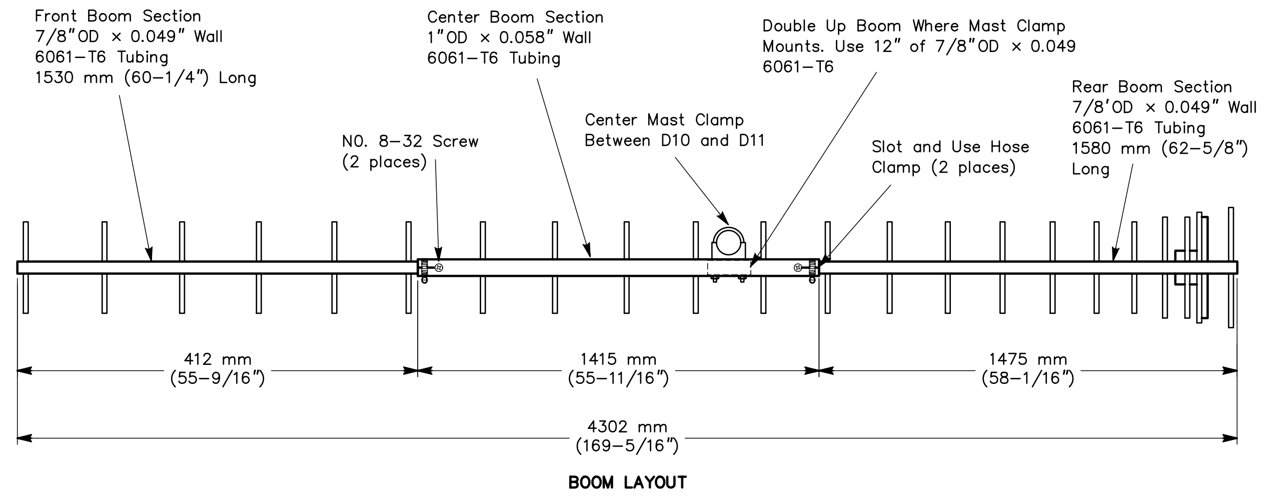
Front Boom Section
7/8″OD × 0.049″ Wall
6061–T6 Tubing
1530 mm (60–1/4″) Long
Center Boom Section
1″OD × 0.058″ Wall
6061–T6 Tubing
Double Up Boom Where Mast Clamp
Mounts. Use 12″ of 7/8″OD × 0.049
6061–T6
Rear Boom Section
7/8′OD × 0.049″ Wall
6061–T6 Tubing
1580 mm (62–5/8″)
Long
Center Mast Clamp
Between D10 and D11
NO. 8–32 Screw
(2 places)
Slot and Use Hose
Clamp (2 places)
412 mm
(55–9/16″)
1415 mm
(55–11/16″)
1475 mm
(58–1/16″)
4302 mm
(169–5/16″)
BOOM LAYOUT

Fig 45—Boom-construction information for the 22-element Yagi Lengths are given in millimeters to allow precise duplication of the antenna. See text.

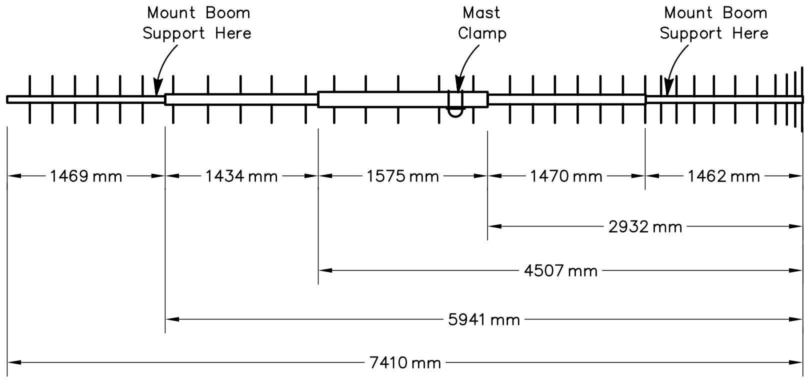

Fig 46—Boom-construction information for the 33-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna.

are not critical. Teflon or black polyethylene are probably the best materials. The Yagi in the photographs is made with black Delryn insulators, available from C3i in Washington, DC.

The driven element uses a UG-58A/U connector mounted on a small bracket. The UG58A/U should be the type with the press-in center pin. UG-58s with center pins held in by "C" clips will usually leak water. Some connectors use steel retaining clips, which will rust and leave a conductive stripe across the insulator. The T-match wires are supported by the UT-141 balun. RG-303/U or RG-142/U Tefloninsulated cable could be used if UT-141 cannot be obtained. **Fig 43A** and Fig 42B show details of the driven-element construction. Driven element dimensions are given in **Fig 44**.

Dimensions for the 22-element Yagi are listed in **Table 9**. **Fig 45** details the Yagi's boom layout. Element material can be either $^3/_{16}$ inch 6061-T6 aluminum rod or hard aluminum welding rod.

A 24-foot-long, 10.6-λ, 33-element Yagi was also built. The construction methods used were the same as the 22-element Yagi. Telescoping round boom sections of 1, $1^1/_8$, and $1^1/_4$ inches in diameter were used. A boom support is required to keep boom sag acceptable. At 432 MHz, if boom sag is much more than two or three inches, H-plane pattern distortion will occur. Greater amounts of boom sag will reduce the gain of a Yagi. **Table 10** lists the proper dimensions for the antenna when built with the previously given boom diameters. The boom layout is shown in **Fig 46**, and the driven element is described in **Fig 47**. The 33-element Yagi exhibits the same clean pattern traits as the 22-element Yagi (see **Fig 48**). Measured gain of the 33-element Yagi is 19.9 dBi at 432 MHz. A measured gain sweep of the 33-element Yagi gave a –1 dB gain bandwidth of 14 MHz with the –1 dB points at 424.5 MHz and 438.5 MHz.

A HIGH-PERFORMANCE 144MHZ YAGI

This 144MHz Yagi design uses the latest log-tapered element spacings and lengths. It offers near theoretical gain per boom length, an extremely clean pattern and wide bandwidth. The design is based upon the spacings used in a 4.5-λ 432-MHz computerdeveloped design by W1EJ. It is quite similar to the 432MHz Yagi described elsewhere in this chapter. Refer to that project for additional construction diagrams and photographs.

Mathematical models do not always directly translate into real working examples. Although the computer design provided a good starting point, the author, Steve Powlishen, K1FO, built several test models before the final working Yagi was obtained. This hands-on tuning included changing the element-taper rate in order to obtain the flexibility that allows the Yagi to be built with different boom lengths.

The design is suitable for use from 1.8 λ (10 elements) to 5.1 λ (19 elements). When elements are added to a Yagi, the center frequency, feed impedance and front-to-back ratio will range up and down. A modern tapered design will minimize this effect and allow the builder to select any desired boom length. This Yagi's design capabilities per boom length are listed in **Table 11**.

The gain of any Yagi built around this design will be within 0.1 to 0.2 dB of the maximum theoretical gain at the design frequency of 144.2 MHz. The design is intentionally peaked high in frequency (calculated gain peak is about 144.7 MHz). It has been found that by doing this, the SWR bandwidth and pattern at 144.0 to 144.3 MHz will be better, the Yagi will be less affected by weather and its performance in arrays will be more predictable. This design starts to drop off in performance if built with fewer than 10 elements. At less than 2 λ, more traditional designs perform well.

Table 12 gives free-space element lengths for $^1/_4$ inch-diameter elements. The use of metric notation allows for much easier dimensional changes during the design stage. Once you become familiar with the metric system, you'll probably find that construction is easier without the burden of cumbersome English fractional units. For $^3/_{16}$ inch-diameter elements, lengthen all parasitic elements by 3 mm. If $^3/_8$ inch diameter elements are used, shorten all of the

Table 9

Dimensions for the 22-Element 432-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl	30	346	
DE	134	340	
D1	176	321	
D2	254	311	7/8
D3	362	305	
D4	496	301	
D5	652	297	
D6	828	295	
D7	1020	293	
D8	1226	291	
D9	1444	289	
D10	1672	288	
D11	1909	286	
D12	2152	285	1
D13	2403	284	
D14	2659	283	
D15	2920	281	
D16	3184	280	
D17	3452	279	7/8
D18	3723	278	
D19	3997	277	
D20	4272	276	

Table 10

Dimensions for the 33-Element 432-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
REF	30	348	
DE	134	342	
D1	176	323	
D2	254	313	
D3	362	307	
D4	496	303	1
D5	652	299	
D6	828	297	
D7	1020	295	
D8	1226	293	
D9	1444	291	
D10	1672	290	
D11	1909	288	
D12	2152	287	1 1/8
D13	2403	286	
D14	2659	285	
D15	2920	284	
D16	3184	284	
D17	3452	283	
D18	3723	282	1 1/4
D19	3997	281	
D20	4272	280	
D21	4550	278	
D22	4828	278	
D23	5109	277	1 1/8
D24	5390	277	
D25	5672	276	
D26	5956	275	
D27	6239	274	
D28	6524	274	1
D29	6809	273	
D30	7094	273	
D31	7380	272	

directors and the reflector by 6 mm. The driven element will have to be adjusted for the individual Yagi if the 12-element design is not adhered to.

For the 12-element Yagi, 1/4-inch diameter elements were selected because smaller-diameter elements become rather flimsy at 2 meters. Other diameter elements can be used as described previously. The 2.5-λ boom was chosen because it has an excellent size and wind load vs gain and pattern trade-off. The size is also convenient; three 6-foot-long pieces of aluminum tubing can be used without any waste. The relatively large-diameter boom sizes (1 1/4 and 1 3/8 inches) were chosen, as they provide an extremely rugged Yagi that does not require a boom support. The 12-element 17-foot-long design has a calculated wind survival of close to 120 mph! The absence of a boom support also makes vertical polarization possible.

Longer versions could be made by telescoping smaller-size boom sections into the last section. Some sort of boom support will be required on versions longer than 22 feet. The elements are mounted on shoulder insulators and mounted through the boom. However, elements may

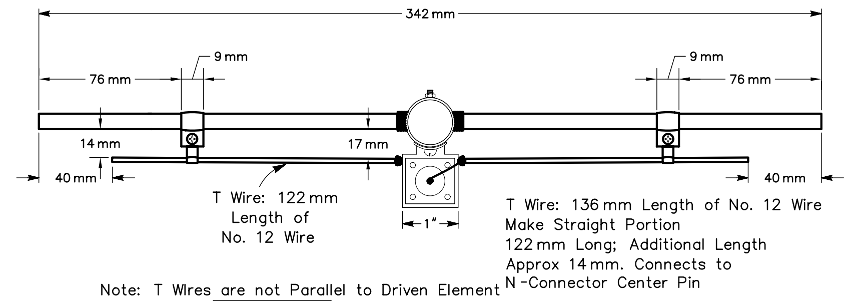
342 mm
9 mm
76 mm
9 mm
76 mm
14 mm
17 mm
40 mm
40 mm
T Wire: 122 mm Length of No. 12 Wire
1"
T Wire: 136 mm Length of No. 12 Wire Make Straight Portion 122 mm Long; Additional Length Approx 14 mm. Connects to N-Connector Center Pin
Note: T Wires are not Parallel to Driven Element

Fig 47—Details of the driven element and T match for the 33-element Yagi. Lengths are given in millimeters to allow precise duplication of the antenna.

Table 11

Specifications for the 144-MHz Yagi Family

No. of Ele.	Boom Length(λ)	Gain (dBd)	DE Imped (Ω)	FB Ratio (dB)	Beamwidth E/H (°)	Stacking E/H (°)
10	1.8	11.4	27	17	39/42	10.2/9.5
11	2.2	12.0	38	19	36/40	11.0/10.0
12	2.5	12.5	28	23	34/37	11.7/10.8
13	2.9	13.0	23	20	32/35	12.5/11.4
14	3.2	13.4	27	18	31/33	12.8/12.0
15	3.6	13.8	35	20	30/32	13.2/12.4
16	4.0	14.2	32	24	29/30	13.7/13.2
17	4.4	14.5	25	23	28/29	14.1/13.6
18	4.8	14.8	25	21	27/28.5	14.6/13.9
19	5.2	15.0	30	22	26/27.5	15.2/14.4

Table 12

Free-Space Dimensions for the 144-MHz Yagi Family

Element diameter is ¼ inch

Element No.	Element Position (mm from reflector)	Element Length
Refl.	0	1038
DE	312	955
D1	447	956
D2	699	932
D3	1050	916
D4	1482	906
D5	1986	897
D6	2553	891
D7	3168	887
D8	3831	883
D9	4527	879
D10	5259	875
D11	6015	870
D12	6786	865
D13	7566	861
D14	8352	857
D15	9144	853
D16	9942	849
D17	10744	845

be mounted, insulated or uninsulated, above or through the boom, as long as appropriate element length corrections are made. Proper tuning can be verified by checking the depth of the nulls between the main lobe and first side lobes. The nulls should be 5 to 10 dB below the first sidelobe level at the primary operating frequency. The boom layout for the 12-element model is shown in **Fig 49**. The actual corrected element dimensions for the 12-element 2.5-λ Yagi are shown in **Table 13**.

The design may also be cut for use at 147 MHz. There is no need to change element spacings. The element lengths should be shortened by 17 mm for best operation between 146 and 148 MHz. Again, the driven element will have to be adjusted as required.

The driven-element size ($^1/_2$-inch diameter) was chosen to allow easy impedance matching. Any reasonably sized driven element could be used, as long as appropriate length and T-match adjustments are made. Different driven-element dimensions are required if you change the boom length. The calculated natural driven-element impedance is given as a guideline. A balanced T-match was chosen because it's easy to adjust for best SWR and provides a balanced radiation pattern. A 4:1 half-wave coaxial balun is used, although impedance-transforming quarter-wave sleeve baluns could also be used. The cal-

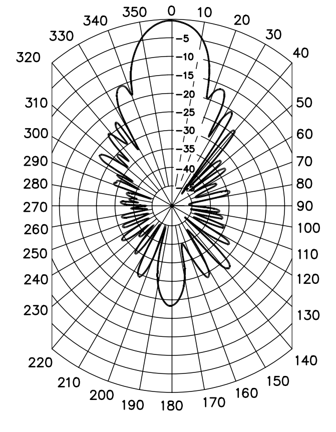

Fig 48—E-plane pattern for the 33-element Yagi. This pattern is drawn on a linear dB grid scale, rather than the standard ARRL log-periodic grid, to emphasize low sidelobes.

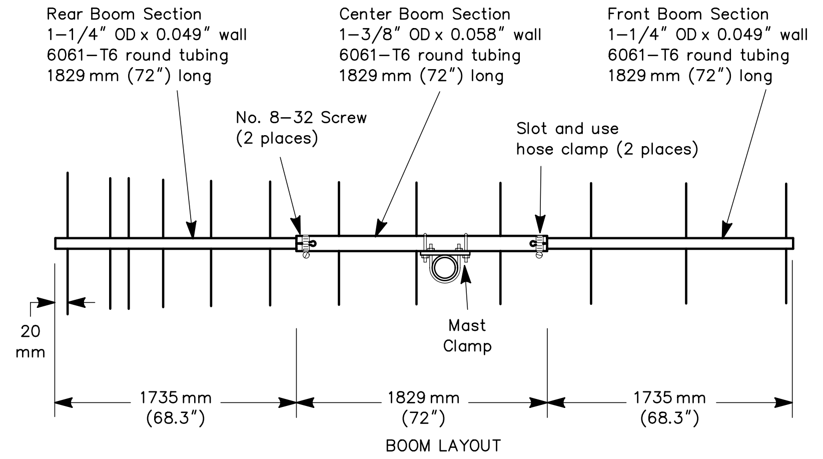
Rear Boom Section
1-1/4″ OD x 0.049″ wall
6061-T6 round tubing
1829 mm (72″) long
Center Boom Section
1-3/8″ OD x 0.058″ wall
6061-T6 round tubing
1829 mm (72″) long
Front Boom Section
1-1/4″ OD x 0.049″ wall
6061-T6 round tubing
1829 mm (72″) long
No. 8-32 Screw
(2 places)
Slot and use
hose clamp (2 places)
20
mm
Mast
Clamp
1735 mm
(68.3″)
1829 mm
(72″)
1735 mm
(68.3″)
BOOM LAYOUT

Fig 49—Boom layout for the 12-element 144-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

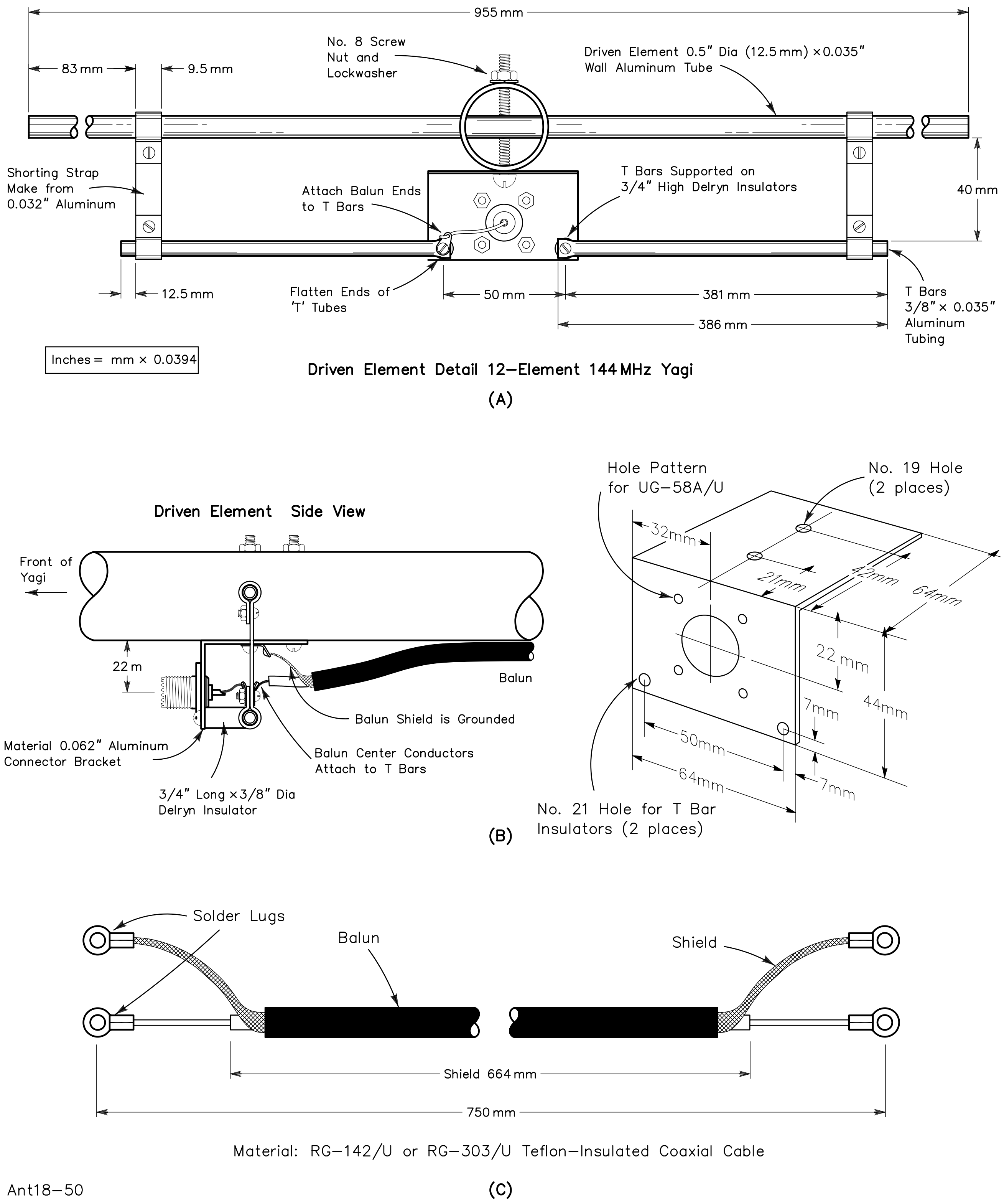
955 mm
83 mm
9.5 mm
No. 8 Screw
Nut and
Lockwasher
Driven Element 0.5″ Dia (12.5 mm) × 0.035″
Wall Aluminum Tube
Shorting Strap
Make from
0.032″ Aluminum
Attach Balun Ends
to T Bars
T Bars Supported on
3/4″ High Delryn Insulators
40 mm
12.5 mm
Flatten Ends of
'T' Tubes
50 mm
381 mm
386 mm
T Bars
3/8″ × 0.035″
Aluminum
Tubing
Inches = mm × 0.0394
Driven Element Detail 12—Element 144 MHz Yagi
(A)
Driven Element Side View
Front of
Yagi
22 m
Balun
Balun Shield is Grounded
Material 0.062″ Aluminum
Connector Bracket
Balun Center Conductors
Attach to T Bars
3/4″ Long × 3/8″ Dia
Delryn Insulator
Hole Pattern
for UG—58A/U
No. 19 Hole
(2 places)
32mm
21mm
42mm
64mm
22 mm
44mm
7mm
50mm
64mm
7mm
No. 21 Hole for T Bar
Insulators (2 places)
(B)
Solder Lugs
Balun
Shield
Shield 664 mm
750 mm
Material: RG—142/U or RG—303/U Teflon—Insulated Coaxial Cable
Ant18—50
(C)

Fig 50—Driven-element detail for the 12-element 144-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

Table 13

Dimensions for the 12-Element 2.5-λ Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl.	0	1044	
DE	312	955	
D1	447	962	$1^{1}/_{4}$
D2	699	938	
D3	1050	922	
D4	1482	912	
D5	1986	904	
D6	2553	898	$1^{3}/_{8}$
D7	3168	894	
D8	3831	889	
D9	4527	885	$1^{1}/_{4}$
D10	5259	882	

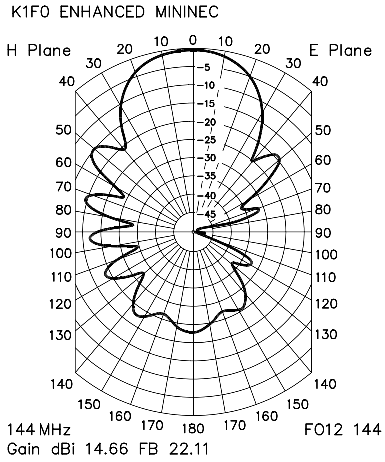

Fig 51—H- and E-plane pattern for the 12-element 144-MHz Yagi.

culated natural impedance will be useful in determining what impedance transformation will be required at the 200-Ω balanced feed point. Chapter 26, Coupling the Line to the Antenna, contains information on calculating folded-dipole and T-match driven-element parameters. A balanced feed is important for best operation on this antenna. Gamma matches can severely distort the pattern balance. Other useful driven-element arrangements are the Delta match and the folded dipole, if you're willing to sacrifice some flexibility. **Fig 50** details the driven-element dimensions.

A noninsulated driven element was chosen for mounting convenience. An insulated driven element may also be used. A grounded driven element may be less affected by static build-up. On the other hand, an insulated driven element allows the operator to easily check his feed lines for water or other contamination by the use of an ohmmeter from the shack.

Table 14

Free-Space Dimensions for the 222-MHz Yagi Family

Element diameter is $^{3}/_{16}$-inch.

Element No.	*Element Position (mm from reflector)*	*Element Length (mm)*
Refl.	0	676
DE	204	647
D1	292	623
D2	450	608
D3	668	594
D4	938	597
D5	1251	581
D6	1602	576
D7	1985	573
D8	2395	569
D9	2829	565
D10	3283	562
D11	3755	558
D12	4243	556
D13	4745	554
D14	5259	553
D15	5783	552
D16	6315	551
D17	6853	550
D18	7395	549
D19	7939	548
D20	8483	547

Fig 51 shows computer-predicted E- and H-plane radiation patterns for the 12-element Yagi. The patterns are plotted on a l-dB-per-division linear scale instead of the usual ARRL polar-plot graph. This expanded scale plot is used to show greater pattern detail. The pattern for the 12-element Yagi is so clean that a plot done in the standard ARRL format would be almost featureless, except for the main lobe and first sidelobes.

The excellent performance of the 12-element Yagi is demonstrated by the reception of Moon echoes from several of the larger 144MHz EME stations with only one 12-element Yagi. Four of the 12-element Yagis will make an excellent starter EME array, capable of working many EME QSOs while being relatively small in size. The advanced antenna builder can use the information in Table 11 to design a dream array of virtually any size.

A HIGH-PERFORMANCE 222-MHz YAGI

Modern tapered Yagi designs are easily applied to 222 MHz. This design uses a spacing progression that is in between the 12-element 144-MHz design, and the 22-element 432-MHz design presented elsewhere in this chapter. The result is a design with maximum gain per boom length, a clean, symmetrical radiation pattern, and wide bandwidth. Although it was designed for weak-signal work (tropospheric scatter and EME), the design is

Table 15
Specifications for the 222-MHz Yagi Family

No of Ele.	*Boom Length(λ)*	*Gain (dBd)*	*FB Ratio (dB)*	*DE Imped (Ω)*	*Beamwidth E/H (°)*	*Stacking E/H (feet)*
12	2.4	12.3	22	23	37/39	7.1/6.7
13	2.8	12.8	19	28	33/36	7.8/7.2
14	3.1	13.2	20	34	32/34	8.1/7.6
15	3.5	13.6	24	30	30/33	8.6/7.8
16	3.9	14.0	23	23	29/31	8.9/8.3
17	4.3	14.35	20	24	28/30.5	9.3/8.5
18	4.6	14.7	20	29	27/29	9.6/8.9
19	5.0	15.0	22	33	26/28	9.9/9.3
20	5.4	15.3	24	29	25/27	10.3/9.6
21	5.8	15.55	23	24	24.5/26.5	10.5/9.8
22	6.2	15.8	21	23	24/26	10.7/10.2

suited to all modes of 222-MHz operation, such as packet radio, FM repeater operation and control links.

The spacings were chosen as the best compromise for a 3.9-λ 16-element Yagi. The 3.9-λ design was chosen, like the 12-element 144-MHz design, because it fits perfectly on a boom made from three 6-foot-long aluminum tubing sections. The design is quite extensible, and models from 12 elements (2.4 λ) to 22 elements (6.2 λ) can be built from the dimensions given in **Table 14**. Note that free-space lengths are given. They must be corrected for the element-mounting method. Specifications for various boom lengths are shown in **Table 15**.

Construction

Large-diameter ($1^{1}/_{4}$- and $1^{3}/_{8}$-inch diameter) boom construction is used, eliminating the need for boom supports. The Yagi can also be used vertically polarized. Three-sixteenths-inch-diameter aluminum elements are used. The exact alloy is not critical; 6061-T6 was used, but hard aluminum welding rod is also suitable. Quarter-inch-diameter elements could also be used if all elements are shortened by 3 mm. Three-eighths-inch-diameter elements would require 10-mm shorter lengths. Elements smaller than $^{3}/_{16}$ inch-diameter are not recommended. The elements are insulated and run through the boom. Plastic shoulder washers and stainless steel retainers are used to hold the elements in place. The various pieces needed to build the Yagi may be obtained from C3i in Washington, DC. **Fig 52** details the boom layout for the 16-element Yagi. **Table 16** gives the dimensions for the 16-element Yagi as built. The driven element is fed with a T match and a 4:1 balun. See **Fig 53** for construction details. See the 432-MHz Yagi project elsewhere in this chapter for additional photographs and construction diagrams.

The Yagi has a relatively broad gain and SWR curve, as is typical of a tapered design, making it usable over a wide frequency range. The example dimensions are intended for use at 222.0 to 222.5 MHz. The 16-element Yagi is quite usable to more than 223 MHz. The best compromise for covering the entire band is to shorten all parasitic elements by 4 mm. The driven element will have to be adjusted in length for best match. The position of the T-wire shorting straps may also have to be moved.

The aluminum boom provides superior strength, is lightweight, and has a low wind-load cross section. Aluminum is doubly attractive, as it will long outlast wood and fiberglass. Using state-of-the-art designs, it is unlikely that significant performance increases will be achieved in the next few years. Therefore, it's in your best interest to build an antenna that will last many years. If suitable wood or fiberglass poles are readily available, they may be used without any performance degradation, at least when the wood is new and dry. Use the free-space element lengths given in Table 16 for insulated-boom construction.

The pattern of the 16-element Yagi is shown in **Fig 54**. Like the 144-MHz Yagi, a l-dB-per-division plot is used to detail the pattern accurately. This 16-element design makes a good building block for EME or tropo DX arrays. Old-style narrow-band Yagis often perform

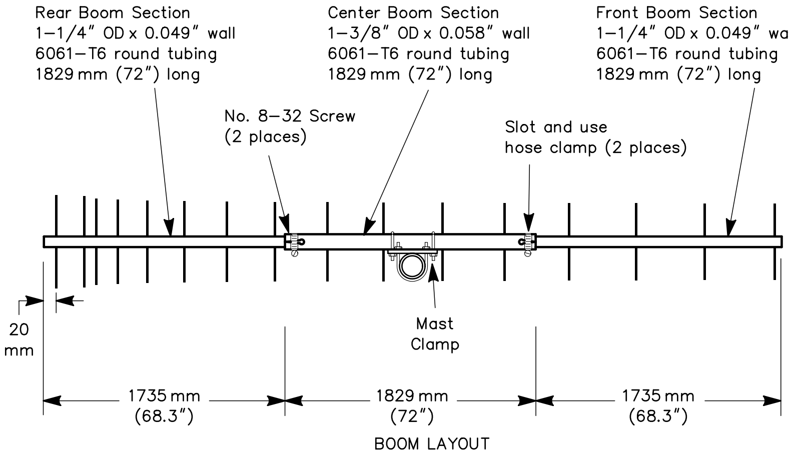

Fig 52—Boom layout for the 16-element 222-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

Table 16

Dimensions for 16-Element 3.9-λ 222-MHz Yagi

Element Number	*Element Position (mm from reflector)*	*Element Length (mm)*	*Boom Diam (in)*
Refl.	0	683	
DE	204	664	
D1	292	630	
D2	450	615	
D3	668	601	1¼
D4	938	594	
D5	1251	588	
D6	1602	583	
D7	1985	580	
D8	2395	576	
D9	2829	572	1⅜
D10	3283	569	
D11	3755	565	
D12	4243	563	
D13	4745	561	1¼
D14	5259	560	

unpredictably when used in arrays. The theoretical 3.0-dB stacking gain is rarely observed. The 16-element Yagi (and other versions of the design) reliably provides stacking gains of nearly 3 dB. (The spacing dimensions listed in Table 15 show just over 2.9 dB stacking gain.) This has been found to be the best compromise between gain, pattern integrity and array size. Any phasing line losses will subtract from the possible stacking gain. Mechanical misalignment will also degrade the performance of an array.

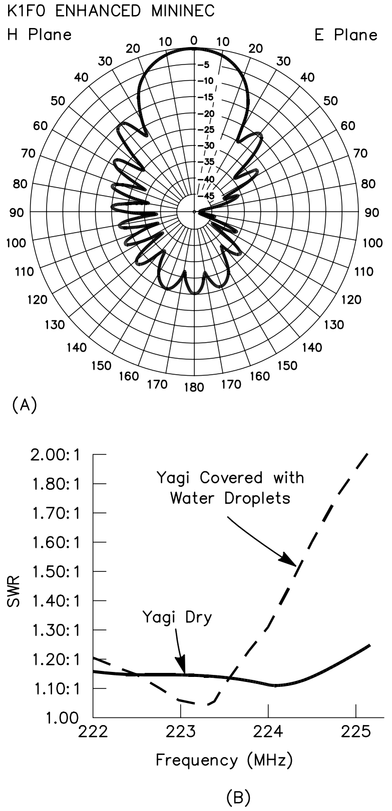

Fig 54—H- and E-plane patterns for the 16-element 222-MHz Yagi at A. The driven-element T-match dimensions were chosen for the best SWR compromise between wet and dry weather conditions. The SWR vs frequency curve shown at B demonstrates the broad frequency response of the Yagi design.

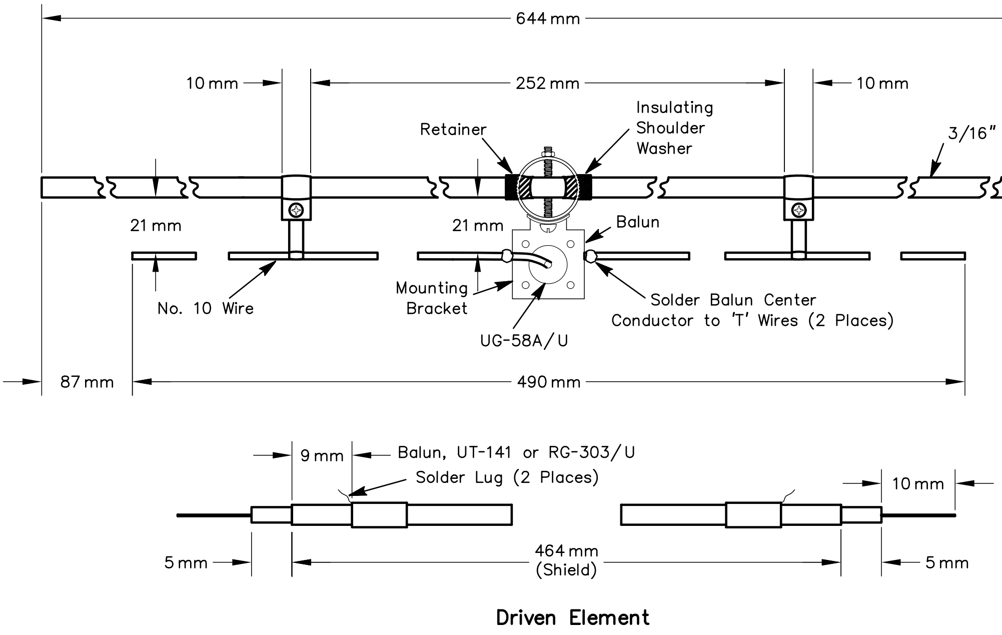
644 mm
10 mm
252 mm
10 mm
Retainer
Insulating Shoulder Washer
3/16″ Diam
21 mm
21 mm
Balun
No. 10 Wire
Mounting Bracket
UG-58A/U
Solder Balun Center Conductor to 'T' Wires (2 Places)
87 mm
490 mm
9 mm
Balun, UT-141 or RG-303/U
Solder Lug (2 Places)
10 mm
5 mm
464 mm (Shield)
5 mm
Driven Element

Fig 53—Driven-element detail for the 16-element 222-MHz Yagi. Lengths are given in millimeters to allow precise duplication.

A 144 MHz 2-Element Quad

The basic 2-element quad array for 144 MHz is shown in **Fig 55**. The supporting frame is 1 × 1-inch wood, of any kind suitable for outdoor use. Elements are #8 aluminum wire. The driven element is 1 λ (83 inches) long, and the reflector five percent longer (87 inches). Dimensions are not critical, as the quad is relatively broad in frequency response.

The driven element is open at the bottom, its ends fastened to a plastic block. The block is mounted at the bottom of the forward vertical support. The top portion of the element runs through the support and is held firmly by a screw running into the wood and then bearing on the aluminum wire. Feed is by means of 50-Ω coax, connected to the driven-element loop.

The reflector is a closed loop, its top and bottom portions running through the rear vertical support. It is held in position with screws at the top and bottom. The loop can be closed by fitting a length of tubing over the element ends, or by hammering them flat and bolting them together as shown in the sketch.

The elements in this model are not adjustable, though this can easily be done by the use of stubs. It would then be desirable to make the loops slightly smaller to compensate for the wire in the adjusting stubs. The driven element stub would be trimmed for length and the point of connection for the coax would be adjustable for best match. The reflector stub can be adjusted for maximum gain or maximum F/B ratio, depending on the builder's requirements.

In the model shown only the spacing is adjusted, and this is not particularly critical. If the wooden supports are made as shown, the spacing between the elements can be adjusted for best match, as indicated by an SWR meter connected in the coaxial line. The spacing has little effect on the gain (from 0.15 to 0.25 λ), so the variation in impedance with spacing can be used for matching. This also permits use of either 50- or 75-Ω coax for the transmission line.

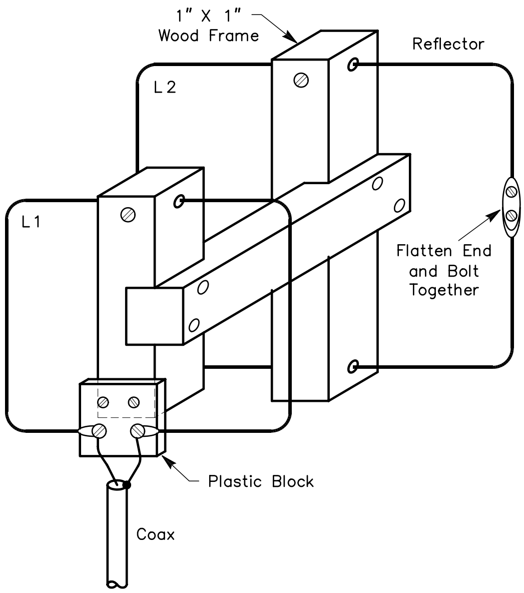

Fig 55—Mechanical details of a 2-element quad for 144 MHz. The driven element, L1, is one wavelength long; reflector L2 is 5% longer. With the transmission line connected as shown here, the resulting radiation is horizontally polarized. Sets of elements of this type can be stacked horizontally and vertically for high gain with broad frequency response. Recommended bay spacing is ½ λ between adjacent element sides. The example shown may be fed directly with 50-Ω coax.

A Portable 144 MHz 4-Element Quad

Element spacing for quad antennas found in the literature ranges from 0.14 λ to 0.25 λ. Factors such as the number of elements in the array and the parameters to be optimized (F/B ratio, forward gain, bandwidth, etc), determine the optimum element spacing within this range. The 4-element quad antenna described here was designed for portable use, so a compromise between these factors was chosen. This antenna, pictured in **Fig 56**, was designed and built by Philip D'Agostino, W1KSC.

Based on several experimentally determined correction factors related to the frequency of operation and the wire size, optimum design dimensions were found to be as follows.

$$\text{Reflector length (ft)} = \frac{1046.8}{f_{\text{MHz}}} \qquad \text{(Eq 8)}$$

$$\text{Driven element (ft)} = \frac{985.5}{f_{\text{MHz}}} \qquad \text{(Eq 9)}$$

$$\text{Directors (ft)} = \frac{937.3}{f_{\text{MHz}}} \qquad \text{(Eq 10)}$$

Cutting the loops for 146 MHz provides satisfactory performance across the entire 144MHz band.

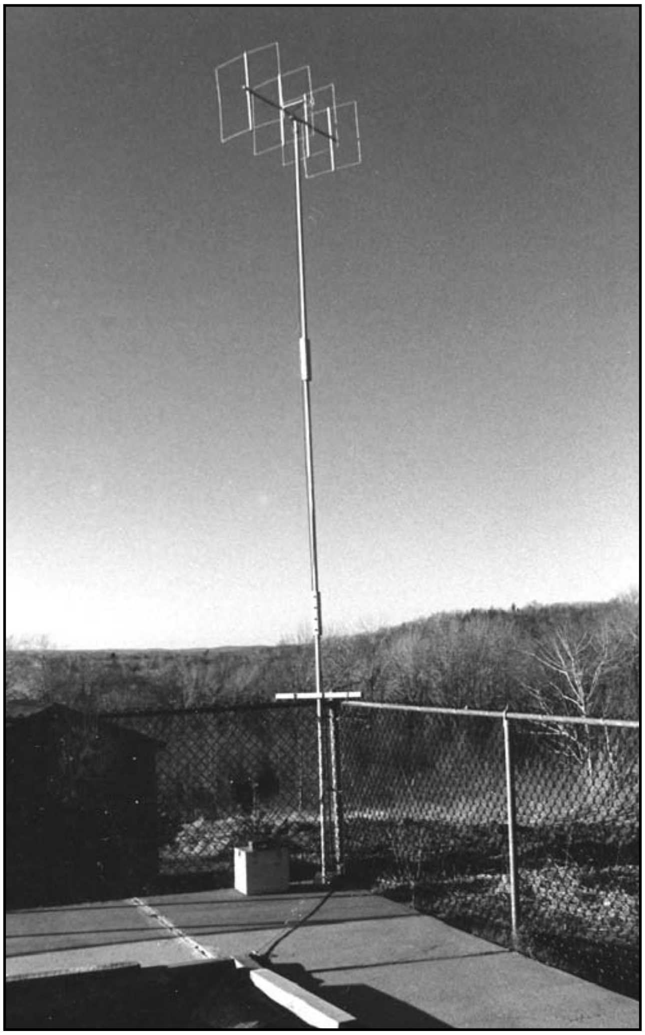

Fig 56—The 4-element 144-MHz portable quad, assembled and ready for operation. Sections of clothes closet poles joined with pine strips make up the mast. (*Photo by Adwin Rusczek, W1MPO)*

Materials

The quad was designed for quick and easy assembly and disassembly, as illustrated in **Fig 57**. Wood (clear trim pine) was chosen as the principal building material because of its light weight, low cost, and ready availability. Pine is used for the boom and element supporting arms. Round wood clothes closet poles comprise the mast material. Strips connecting the mast sections are made of heavier pine trim. Elements are made of no. 8 aluminum wire. Plexiglas is used to support the feed point. **Table 17** lists the hardware and other parts needed to duplicate the quad.

Construction

The elements of the quad are assembled first. The mounting holes in the boom should be drilled to accommodate $1^1/_2$ inch no. 8 hardware. Measure and mark the locations where the holes are to be drilled in the element spreaders, **Fig 58**. Drill the holes in the spreaders just

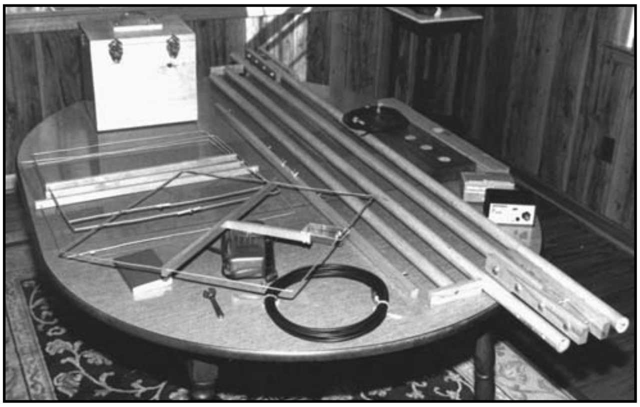

Fig 57—The complete portable quad, broken down for travel. Visible in the foreground is the driven element. The pine box in the background is a carrying case for equipment and accessories. A hole in the lid accepts the mast, so the box doubles as a base for a short mast during portable operation. *(W1MPO photo*)

Table 17

Parts List for the 144 MHz 4-element Quad

Boom: ¾ × ¾ ×48-in. pine

Driven element support (spreader): ½ × ¾ × 21¼ in. pine

Driven element feed point strut: ½ × ¾ × 7½ in. pine

Reflector support (spreader): ½ × ¾ × 22½ in. pine

Director supports (spreaders): ½ × ¾ × 20¼ in. pine, 2 req'd

Mast brackets: ¾ × 1½ × 12 in. heavy pine trim, 4 req'd

Boom to mast bracket: ½ × $1^5/_8$ × 5 in. pine

Element wire: Aluminum ground wire (Radio Shack no. 15-035)

Wire clamps: ¼in. electrician's copper or zinc plated steel clamps, 3 req'd

Boom hardware:
- 6 no. 8-32 × 1½ in. stainless steel machine screws
- 6 no. 8-32 stainless steel wing nuts
- 12 no. 8 stainless steel washers

Mast hardware:
- 8 hex bolts, ¼-20 × 3½ in.
- 8 hex nuts, ¼-20
- 16 flat washers

Mast material: $1^5/_{16}$ in. × 6 ft wood clothes closet poles, 3 req'd

Feed point support plate: 3½ × 2½ in. Plexiglas sheet

Wood preparation materials: Sandpaper, clear polyurethane, wax

Feed line: 52-W RG-8 or RG-58 cable

Feed line terminals: Solder lugs for no. 8 or larger hardware, 2 req'd

Miscellaneous hardware: 4 small machine screws, nuts, washers; 2 flat-head wood screws large enough to accept the #8 wire elements. It is important to drill all the holes straight so the elements line up when the antenna is assembled.

Construction of the wire elements is easiest if the directors are made first. A handy jig for bending the elements can be made from a piece of 2 × 3-inch wood cut to the side length of the directors. It is best to start with about 82 inches of wire for each director. The excess can be cut off when the elements are completed. (The total length of each director is 77 inches.) Two bends should initially be made so the directors can be slipped into the spreaders before the remaining corners are bent. See **Fig 59**. Electrician's copper-wire clamps can be used to join the wires after the final bends are made, and they facilitate adjustment of element length. The reflector is made the same way as the directors, but the total length is 86 inches.

The driven element, total length 81 inches, requires special attention, as the feed attachment point needs to be adequately supported. An extra hole is drilled in the driven element spreader to support the feed-point strut, as shown in **Fig 60**. A Plexiglas plate is used at the feed point to support the feed- point hardware and the feed line. The feed-point support strut should be epoxied to the spreader, and a wood screw used for extra mechanical strength.

For vertical polarization, locate the feed point in the center of one side of the driven element, as shown in Fig 60. Although this arrangement places the spreader supports at voltage maxima points on the four loop conductors, D'Agostino reports no adverse effects during operation. However, if the antenna is to be left exposed to the weather, the builder may wish to modify the design to provide support for the loops at current maxima points, such as shown in Fig 60. (The element of Fig 60 should be rotated 90° for horizontal polarization.)

Orient the driven element spreader so that it mounts properly on the boom when the antenna is assembled. Bend the driven element the same way as the reflector and directors, but do not leave any overlap at the feed point. The ends of the wires should be 3/4 inch apart where they mount on the Plexiglas plate. Leave enough excess that small loops can be bent in the wire for attachment to the coaxial feed line with stainless steel hardware.

Drill the boom as shown in **Fig 61**. It is a good idea to use hardware with wing nuts to secure the element spreaders to the boom. After the boom is drilled, clean all the wood parts with denatured alcohol, sand them, and give them two coats of glossy polyurethane. After the polyurethane dries, wax all the wooden parts.

The boom to mast attachment is made next. Square the ends of a 6-foot section of clothes closet pole (a miter box is useful for this). Drill the center holes in both the boom

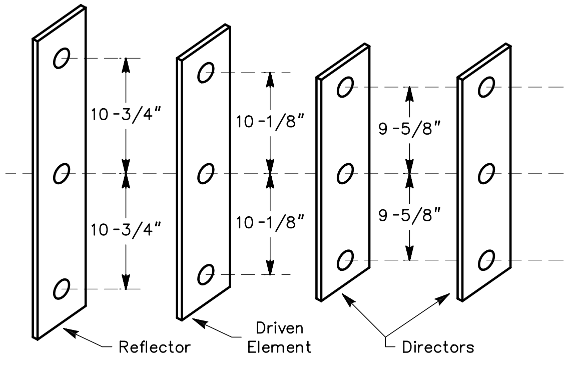

Fig 58—Dimensions for the pine element spreaders for the 144-MHz 4-element quad.

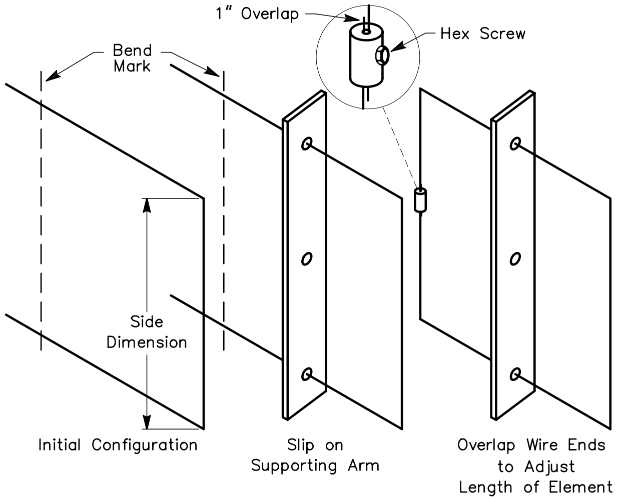

Fig 59—Illustration showing how the aluminum element wires are bent. The adjustment clamp and its location are also shown.

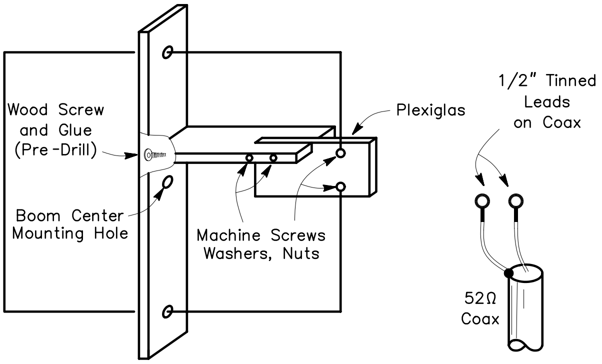
Wood Screw and Glue (Pre-Drill)
Boom Center Mounting Hole
Plexiglas
Machine Screws Washers, Nuts
1/2" Tinned Leads on Coax
52Ω Coax

Fig 60—Layout of the driven element of the 144-MHz quad. The leads of the coaxial cable should be stripped to ½ in. and solder lugs attached for easy connection and disconnection. See text regarding impedance at loop support points.

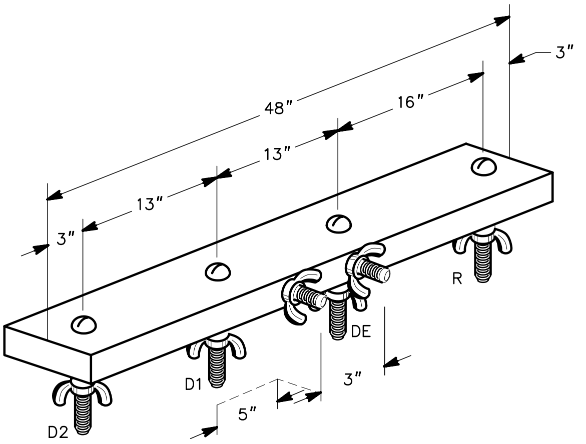

Fig 61—Detail of the boom showing hole center locations and boom to mast connection points.

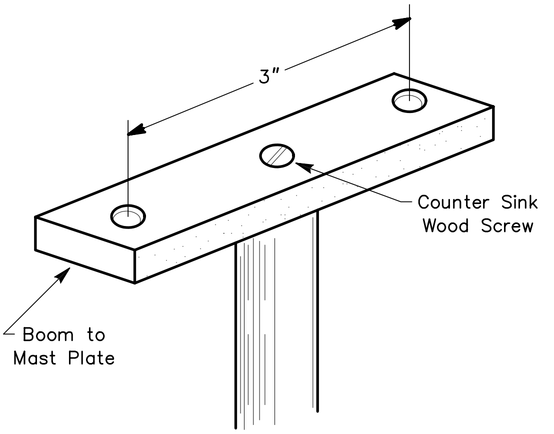

Fig 62—Boom to mast plate for the 144-MHz quad. The screw hole in the center of the plate should be countersunk so the wood screw attaching it to the mast does not interfere with the fit of the boom.

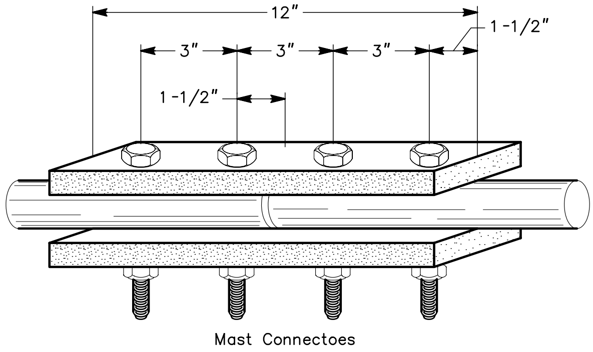

Fig 63—Mast coupling connector details for the portable quad. The plates should be drilled two at a time to ensure the holes line up.

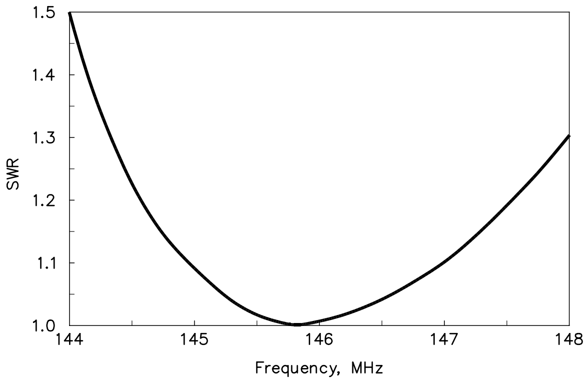

Fig 64—Typical SWR curve for the 144MHz portable quad. The large wire diameter and the quad design provide excellent bandwidth.

attachment piece and one end of the mast section (**Fig 62**). Make certain that the mast hole is smaller than the flat-head screw to be used to ensure a snug fit. Accurately drill the holes for attachment to the boom as shown in Fig 62.

Countersink the hole for the flat-head screw to provide a smooth surface for attachment to the boom. Apply epoxy cement to the surfaces and screw the boom attachment piece securely to the mast section. One 6 foot mast is used for attachment to the other mast sections.

Two additional 6-foot mast sections are prepared next. This brings the total mast height to 18 feet. It is important to square the ends of each pole so the mast stands straight when assembled. Mast-section connectors are made of pine as shown in **Fig 63**. Using $3^1/_2 \times {}^1/_4$-inch hex bolts, washers, and nuts, sections may be attached as needed, for a total height of 6, 12 or 18 feet. Drill the holes in two connectors at a time. This ensures good alignment of the holes. A drill press is ideal for this job, but with care a hand drill can be used if necessary.

Line up two mast sections end to end, being careful that they are perfectly straight. Use the predrilled connectors to maintain pole straightness, and drill through the poles, one at a time. If good alignment is maintained, a straight 18-foot mast section can be made. Label the connectors and poles immediately so they are always assembled in the same order.

When assembling the antenna, install all the elements on the boom before attaching the feed line. Connect the coax to the screw connections on the driven element support plate and run the cable along the strut to the boom. From there, the cable should be routed directly to the mast and down. Assemble the mast sections to the desired height. The antenna provides good performance, and has a reasonable SWR curve over the entire 144 MHz band (**Fig 64**).

Building Quagi Antennas

The Quagi antenna was designed by Wayne Overbeck, N6NB. He first published information on this antenna in 1977 (see Bibliography). There are a few tricks to Quagi building, but nothing very difficult or complicated is involved. In fact, Overbeck mass produced as many as 16 in one day. **Tables 18** and **19** give the dimensions for Quagis for various frequencies up to 446 MHz.

For the designs of Tables 18 and 19, the boom is *wood* or any other nonconductor (such as, fiberglass or Plexiglas). If a metal boom is used, a new design and new element lengths will be required. Many VHF antenna builders go wrong by failing to follow this rule: If the original uses a metal boom, use the same size and shape metal boom when you duplicate it. If it calls for a wood boom, use a nonconductor. Many amateurs dislike wood booms, but in a salt air environment they outlast aluminum (and surely cost less). Varnish the boom for added protection.

The 144-MHz version is usually built on a 14 foot, 1 × 3 inch boom, with the boom tapered to 1 inch at both ends. Clear pine is best because of its light weight, but construction grade Douglas fir works well. At 222 MHz the boom is under 10 feet long, and most builders use 1 × 2 or (preferably) $^{3}/_{4}$ × $1^{1}/_{4}$ inch pine molding stock. At 432 MHz, except for long-boom versions, the boom should be $^{1}/_{2}$ inch thick or less. Most builders use strips of $^{1}/_{2}$-inch exterior plywood for 432 MHz.

The quad elements are supported at the current maxima (the top and bottom, the latter beside the feed point) with Plexiglas or small strips of wood. See **Fig 65**. The quad elements are made of #12 copper wire, commonly used in house wiring. Some builders may elect to use #10 wire on 144 MHz and #14 on 432 MHz, although this changes the resonant frequency slightly. Solder a type N connector (an SO-239 is often used at 144 MHz) at the midpoint of the driven element bottom side, and close the reflector loop.

Table 19
432MHz, 15-Element, Long Boom Quagi Construction Data

Element Lengths, Inches	*Interelement Spacing, Inches*
R—28	R-DE7
DE—$26^{5}/_{8}$	DE-D1—$5^{1}/_{4}$
D1—$11^{3}/_{4}$	D1-D2—11
D21—$^{11}/_{16}$	D2-D3—$5^{7}/_{8}$
D3—$11^{5}/_{8}$	D3-D4—$8^{3}/_{4}$
D4—$11^{9}/_{16}$	D4-D5—$8^{3}/_{4}$
D5—$11^{1}/_{2}$	D5-D6—$8^{3}/_{4}$
D6—$11^{7}/_{16}$	D6-D7—12
D7—$11^{3}/_{8}$	D7-D8—12
D8—$11^{5}/_{16}$	D8-D9—$11^{1}/_{4}$
D9—$11^{5}/_{16}$	D9-D10—$11^{1}/_{2}$
D10—$11^{1}/_{4}$	D10-D11—$9^{3}/_{16}$
D11—$11^{3}/_{16}$	D11-D12—$12^{3}/_{8}$
D12—$11^{1}/_{8}$	D12-D13—1-$^{3}/_{4}$
D13—$11^{1}/_{16}$	

Boom: 1 × 2in. × 12-ft Douglas fir, tapered to $^{5}/_{8}$ in. at both ends.
Driven element: #12 TW copper wire loop in square configuration, fed at bottom center with type N connector and 52-Ω coax.
Reflector: #12 TW copper wire loop, closed at bottom.
Directors: $^{1}/_{8}$ in. rod passing through boom.

Table 18
Dimensions, Eight-Element Quagi

Element			*Frequency*		
Lengths	*144.5 MHz*	*147 MHz*	*222 MHz*	*432 MHz*	*446 MHz*
Reflector[1]	$86^{5}/_{8}$"	85"	$56^{3}/_{8}$"	28"	$27^{1}/_{8}$"
Driven[2]	82"	80"	53½"	$26^{5}/_{8}$"	$25^{7}/_{8}$"
Directors	$35^{15}/_{16}$" to 35" in $^{3}/_{16}$" steps	$35^{5}/_{16}$" to $34^{3}/_{8}$" in $^{3}/_{16}$" steps	$23^{3}/_{8}$" to 23¾" in $^{1}/_{8}$" steps	11¾" to $11^{7}/_{16}$" in $^{1}/_{16}$" steps	$11^{3}/_{8}$" to $11^{1}/_{16}$" in $^{1}/_{16}$" steps
Spacing					
R-DE	21"	20½"	$13^{5}/_{8}$"	7"	6.8"
DE-D1	15¾"	$15^{3}/_{8}$"	10¼"	5¼"	5.1"
D1-D2	33"	32½"	21½"	11"	10.7"
D2-D3	17½"	$17^{1}/_{8}$"	$11^{3}/_{8}$"	5.85"	5.68"
D3-D4	26.1"	$25^{5}/_{8}$"	17"	8.73"	8.46"
D4-D5	26.1"	$25^{5}/_{8}$"	17"	8.73"	8.46"
D5-D6	26.1"	$25^{5}/_{8}$"	17"	8.73"	8.46"
Stacking Distance Between Bays					
	11'	10' 10"	7' 1½"	3'7"	3' $5^{5}/_{8}$"

[1] All #12 TW (electrical) wire, closed loops.
[2] All #12 TW wire loops, fed at bottom.

The directors are mounted through the boom. They can be made of almost any metal rod or wire of about $^{1}/_{8}$-inch diameter. Welding rod or aluminum clothesline wire works well if straight. (The designer uses $^{1}/_{8}$-inch stainless-steel rod obtained from an aircraft surplus store.)

A TV type U bolt mounts the antenna on a mast. A single machine screw, washers and a nut are used to secure the spreaders to the boom so the antenna can be quickly “flattened” for travel. In permanent installations two screws are recommended.

Construction Reminders

Based on the experiences of Quagi builders, the following hints are offered. First, remember that at 432 MHz even a $^{1}/_{8}$-inch measurement error results in performance deterioration. Cut the loops and elements as carefully as possible. No precision tools are needed, but accuracy is necessary. Also make sure to get the elements in the right order. The longest director goes closest to the driven element.

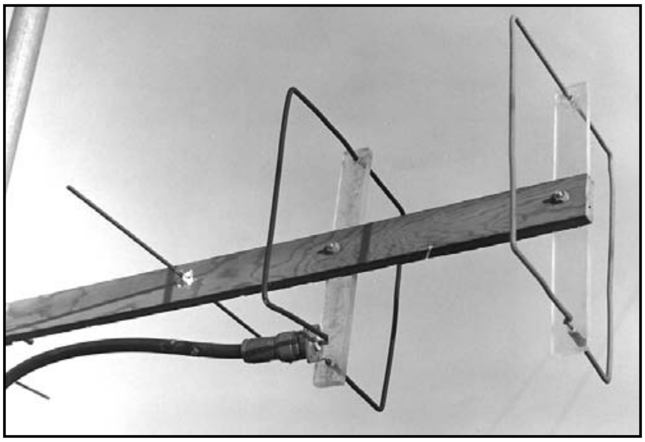

Fig 65—A close-up view of the feed method used on a 432-MHz Quagi. This arrangement produces a low SWR and gain in excess of 13 dBi with a 4-ft 10-in. boom! The same basic arrangement is used on lower frequencies, but wood may be substituted for the Plexiglas spreaders. The boom is ½-in. exterior plywood.

Finally, remember that a balanced antenna is being fed with an unbalanced line. Every balun the designer tried introduced more trouble in terms of losses than the feed imbalance caused. Some builders have tightly coiled several turns of the feed line near the feed point to limit line radiation. In any case, the feed line should be kept at right angles to the antenna. Run it from the driven element directly to the supporting mast and then up or down perpendicularly for best results.

QUAGIS FOR 1296 MHz

The Quagi principle has recently been extended to the 1296-MHz band, where good performance is extremely difficult to obtain from homemade conventional Yagis. **Fig 66** shows the construction and **Table 20** gives the design information for antennas with 10, 15 and 25 elements.

At 1296 MHz, even slight variations in design or building materials can cause substantial changes in performance. The 1296 MHz antennas described here work every time—but only if the same materials are used and the antennas are built *exactly* as described. This is not to discourage experimentation, but if modifications to these 1296-MHz antenna designs are contemplated, consider building one antenna as described here, so a reference is available against which variations can be compared.

The Quagis (and the cubical quad) are built on $^{1}/_{4}$-inch thick Plexiglas booms. The driven element and reflector (and also the directors in the case of the cubical quad) are made of insulated #18 AWG solid copper bell wire, available at hardware and electrical supply stores. Other types and sizes of wire work equally well, but the dimensions vary with the wire diameter. Even removing the insulation usually necessitates changing the loop lengths.

Quad loops are approximately square (**Fig 67**), although the shape is relatively uncritical. The element lengths, however, *are* critical. At 1296 MHz, variations

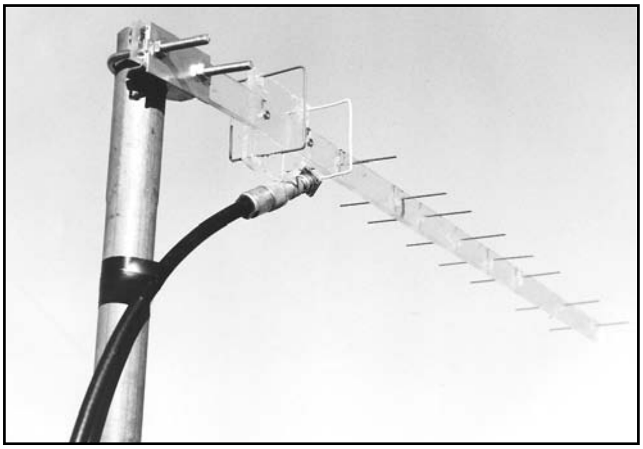

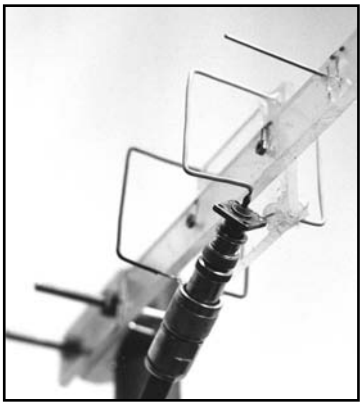

Fig 66—A view of the 10-element version of the 1296-MHz Quagi. It is mounted on a 30-in. Plexiglas boom with a 3 × 3-in. square of Plexiglas to support the driven element and reflector. Note how the driven element is attached to a standard UG-290 BNC connector. The elements are held in place with silicone sealing compound.

of 1/16 inch alter the performance measurably, and a 1/8 inch departure can cost several decibels of gain. The loop lengths given are *gross* lengths. Cut the wire to these lengths and then solder the two ends together. There is a 1/8-inch overlap where the two ends of the reflector (and director) loops are joined, as shown in Fig 67.

The driven element is the most important of all. The #18 wire loop is soldered to a standard UG-290 chassis-mount BNC connector as shown in the photographs. This exact type of connector must be used to ensure uniformity in construction. Any substitution may alter the driven element electrical length. One end of the 9 1/4 inch driven loop is pushed as far as it can go into the center pin, and is soldered in that position. The loop is then shaped and threaded through small holes drilled in the Plexiglas support. Finally, the other end is fed into one of the four mounting holes on the BNC connector and soldered. In most cases, the best SWR is obtained if the end of the wire just passes through the hole so it is flush with the opposite side of the connector flange.

Table 20
Dimensions, 1296-MHz Quagi Antennas

Note: All lengths are gross lengths. See text and photos for construction technique and recommended overlap at loop junctions. All loops are made of #18 AWG solid-covered copper bell wire. The Yagi type directors are 1/16-in. brass brazing rod. See text for a discussion of director taper.

Feed: Direct with 52-Ω coaxial cable to UG-290 connector at driven element; run coax symmetrically to mast at rear of antenna.

Boom: 1 1/4-in. thick Plexiglas, 30 in. long for 10-element quad or Quagi and 48 in. long for 15-element Quagi; 84 in. for 25-element Quagi.

10-Element Quagi for 1296 MHz

Element	*Length, Inches*	*Construction*	*Element*	*Interelement Spacing, In.*
Reflector	9.5625	Loop	R-DE	2.375
Driven	9.25	Loop	DE-D1	2.0
Director 1	3.91	Brass rod	D1-D2	3.67
Director 2	3.88	Brass rod	D2-D3	1.96
Director 3	3.86	Brass rod	D3-D4	2.92
Director 4	3.83	Brass rod	D4-D5	2.92
Director 5	3.80	Brass rod	D5-D6	2.92
Director 6	3.78	Brass rod	D6-D7	4.75
Director 7	3.75	Brass rod	D7-D8	3.94
Director 8	3.72	Brass rod		

15-Element Quagi for 1296 MHz

The first 10 elements are the same lengths as above, but the spacing from D6 to D7 is 4.0 in.; 07 to D8 is also 4.0 in.

Director 9	3.70	D8-D9	3.75
Director 10	3.67	D9-D10	3.83
Director 11	3.64	D10-D11	3.06
Director 12	3.62	D11-D12	4.125
Director 13	3.59	D12-D13	4.58

25-Element Quagi for 1296 MHz

The first 15 elements use the same element lengths and spacings as the 15-element model. The additional directors are evenly spaced at 3.0-in. intervals and taper in length successively by 0.02 in. per element. Thus, D23 is 3.39 in.

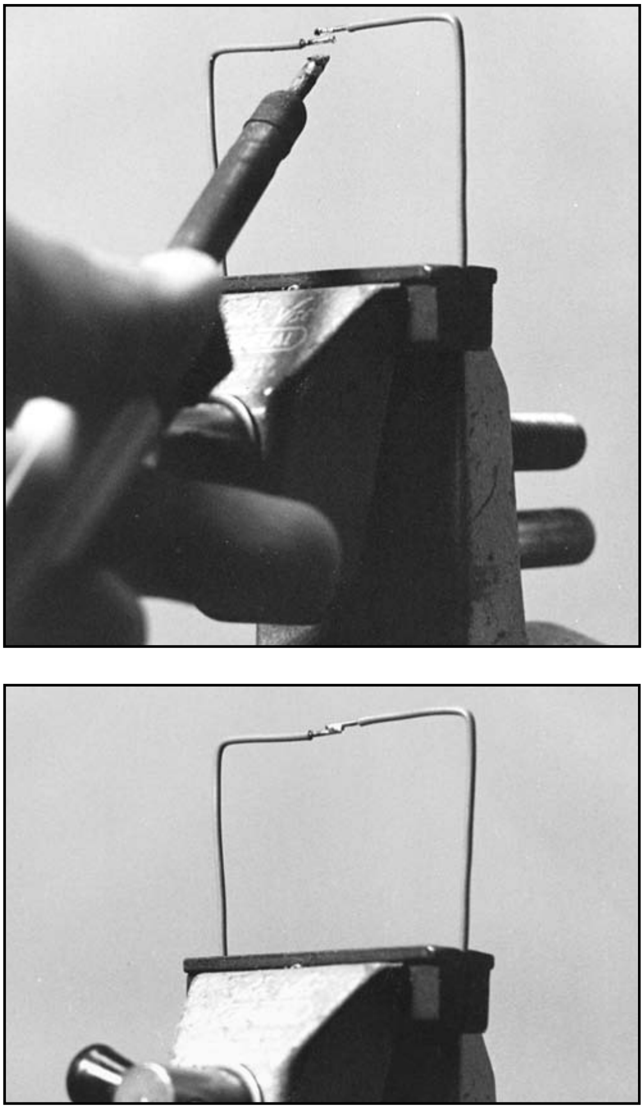

Fig 67—These photos show the construction method used for the 1296-MHz quad type parasitic elements. The two ends of the #18 bell wire are brought together with an overlap of 1/8 in. and soldered.

Loop Yagis for 1296 MHz

Described here are loop Yagis for the 1296-MHz band. The loop Yagi fits into the quad family of antennas, as each element is a closed loop with a length of approximately 1 λ. Several versions are described, so the builder can choose the boom length and frequency coverage desired for the task at hand. Mike Walters, G3JVL, brought the original loop-Yagi design to the amateur community in the 1970s. Since then, many versions have been developed with different loop and boom dimensions. Chip Angle, N6CA, developed the antennas shown here.

Three sets of dimensions are given. Good performance can be expected if the dimensions are carefully followed. Check all dimensions before cutting or drilling anything. The 1296-MHz version is intended for weak-signal operation, while the 1270-MHz version is optimized for FM and mode L satellite work. The 1283-MHz antenna provides acceptable performance from 1280 to 1300 MHz.

These antennas have been built on 6- and 12-foot booms. Results of gain tests at VHF conferences and by individuals around the country show the gain of the 6-foot model to be about 18 dBi, while the 12-foot version provides about 20.5 dBi. Swept measurements indicate that gain is about 2 dB down from maximum gain at ±30 MHz from the design frequency. The SWR, however, deteriorates within a few megahertz on the low side of the design center frequency.

The Boom

The dimensions given here apply only to a ¾-inch OD boom. If a different boom size is used, the dimensions must be scaled accordingly. Many hardware stores

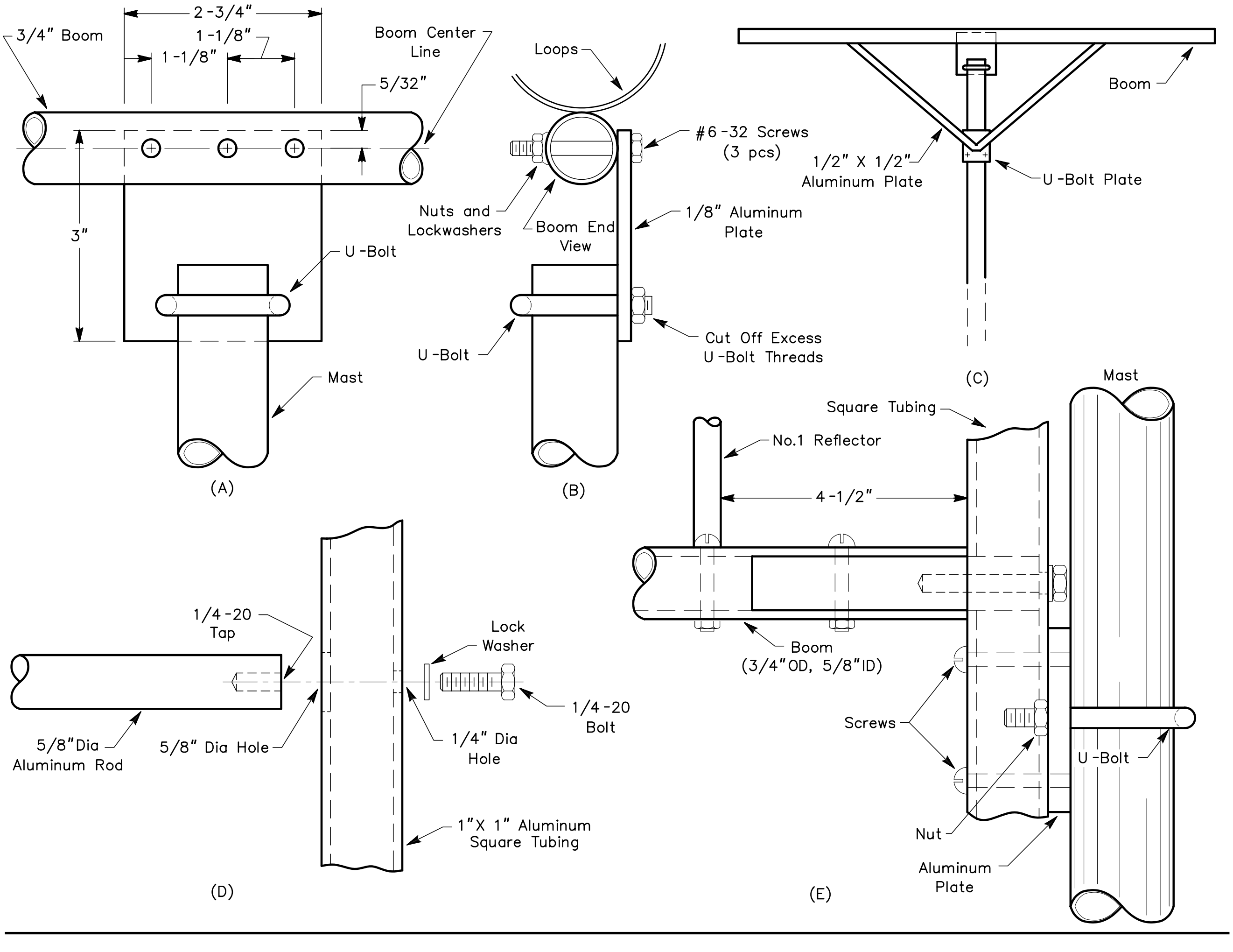
2-3/4"
3/4" Boom
1-1/8"
1-1/8"
Boom Center Line
5/32"
3"
U-Bolt
Mast
(A)
Loops
#6-32 Screws (3 pcs)
Nuts and Lockwashers
Boom End View
1/8" Aluminum Plate
U-Bolt
Cut Off Excess U-Bolt Threads
(B)
Boom
1/2" X 1/2" Aluminum Plate
U-Bolt Plate
(C)
1/4-20 Tap
Lock Washer
1/4-20 Bolt
5/8"Dia Aluminum Rod
5/8" Dia Hole
1/4" Dia Hole
1"X 1" Aluminum Square Tubing
(D)
Mast
Square Tubing
No.1 Reflector
4-1/2"
Boom (3/4"OD, 5/8"ID)
Screws
U-Bolt
Nut
Aluminum Plate
(E)

Fig 68—Loop Yagi boom-to-mast plate details are given at A. At B, the mounting of the antenna to the mast is detailed. A boom support for long antennas is shown at C. The arrangement shown in D and E may be used to rear-mount antennas up to 6 or 7 ft long.

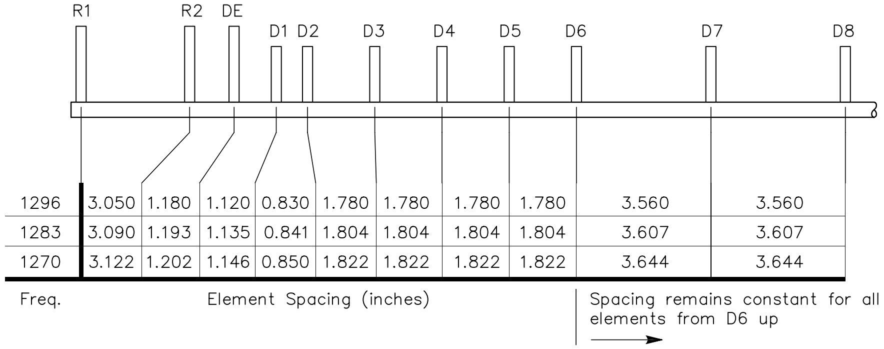
R1
R2
DE
D1
D2
D3
D4
D5
D6
D7
D8
Spacing remains constant for all elements from D6 up

Freq.	Element Spacing (inches)									
1296	3.050	1.180	1.120	0.830	1.780	1.780	1.780	1.780	3.560	3.560
1283	3.090	1.193	1.135	0.841	1.804	1.804	1.804	1.804	3.607	3.607
1270	3.122	1.202	1.146	0.850	1.822	1.822	1.822	1.822	3.644	3.644

Fig 69—Boom drilling dimensions. These dimensions must be carefully followed and the same materials used if performance is to be optimum. Element spacings are the same for all directors after D6—use as many as necessary to fill the boom.

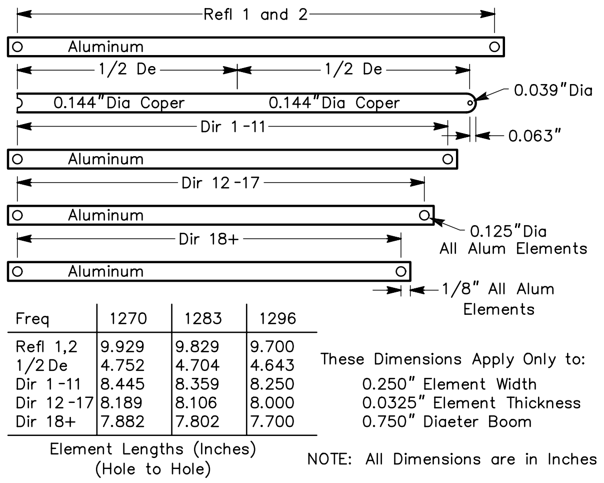
Refl 1 and 2
Aluminum
1/2 De
1/2 De
0.144"Dia Coper
0.144"Dia Coper
0.039"Dia
Dir 1-11
0.063"
Aluminum
Dir 12-17
Aluminum
0.125"Dia
All Alum Elements
Dir 18+
Aluminum
1/8" All Alum
Elements

Freq	1270	1283	1296
Refl 1,2	9.929	9.829	9.700
1/2 De	4.752	4.704	4.643
Dir 1-11	8.445	8.359	8.250
Dir 12-17	8.189	8.106	8.000
Dir 18+	7.882	7.802	7.700

Element Lengths (Inches)
(Hole to Hole)

These Dimensions Apply Only to:
0.250" Element Width
0.0325" Element Thickness
0.750" Diaeter Boom

NOTE: All Dimensions are in Inches

Fig 70—Parasitic elements for the loop Yagi are made from aluminum sheet, the driven element from copper sheet. The dimensions given are for ¼-in. wide by 0.0325-in. thick elements only. Lengths specified are hole to hole distances; the holes are located $^1/_8$ in. from each element end.

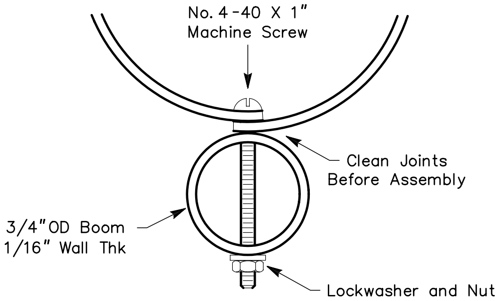
No. 4-40 X 1"
Machine Screw
Clean Joints
Before Assembly
3/4"OD Boom
1/16" Wall Thk
Lockwasher and Nut

Fig 71—Element-to-boom mounting details.

carry aluminum tubing in 6- and 8-foot lengths, and that tubing is suitable for a short Yagi. If a 12-foot antenna is planned, find a piece of more rugged boom material, such as 6061-T6 grade aluminum. Do not use anodized tubing. The 12foot antenna must have additional boom support to minimize boom sag. The 6 foot version can be rear mounted. For rear mounting, allow 4$^1/_2$ inches of boom behind the last reflector to eliminate SWR effects from the support.

The antenna is attached to the mast with a gusset plate. This plate mounts at the boom center. See **Fig 68**. Drill the plate mounting holes perpendicular to the element mounting holes (assuming the antenna polarization is to be horizontal).

Elements are mounted to the boom with no. 4-40 machine screws, so a series of no. 33 (0.113inch) holes must be drilled along the center of the boom to accommodate this hardware. **Fig 69** shows the element spacings for different parts of the band. Dimensions should be followed as closely as possible.

Parasitic Elements

The reflectors and directors are cut from 0.032-inch thick aluminum sheet and are $^1/_4$ inch wide. **Fig 70** indicates the lengths for the various elements. These lengths apply only to elements cut from the specified material. For best results, the element strips should be cut with a shear. If the edges are left sharp, birds won't sit on the elements.

Drill the mounting holes as shown in Fig 70 after carefully marking their locations. After the holes are drilled, form each strap into a circle. This is easily done by wrapping the element around a round form. (A small juice can works well.)

Mount the loops to the boom with no. 4-40 × 1-inch machine screws, lock washers and nuts. See **Fig 71**. It is best to use only stainless steel or plated-brass hardware. Although the initial cost is higher than for ordinary plated-steel hardware, stainless or brass hardware will not rust and need replacement after a few years. Unless the antenna is painted, the hardware will definitely deteriorate.

Driven Element

The driven element is cut from 0.032-inch copper sheet and is $^1/_4$ inch wide. Drill three holes in the strap as detailed in Fig 69. Trim the ends as shown and form the strap into a loop similar to the other elements. This antenna is like a quad; if the loop is fed at the top or bottom, it is horizontally polarized.

Driven element mounting details are shown in **Fig 72**. A mounting fixture is made from a $^1/_4$-20 × 1$^1/_4$ inch brass bolt. File the bolt head to a thickness of $^1/_8$ inch. Bore a 0.144-inch (no. 27 drill) hole lengthwise through the center of the bolt. A piece of 0.141 inch semi-rigid Hardline (UT-141 or equivalent) mounts through this hole and is soldered to the driven loop feed point. The point at which the UT-141 passes through the copper loop and brass mounting fixture should be left unsoldered at this time to allow for matching adjustments when the antenna is completed, although the range of adjustment is not very large.

The UT-141 can be any convenient length. Attach the connector of your choice (preferably type N). Use a short piece of low-loss RG-8 size cable (or $^1/_2$-inch Hardline) for the run down the boom and mast to the main feed line. For best results, the main feed line should be the lowest loss 50-Ω cable obtainable. Good $^7/_8$-inch Hardline has 1.5 dB of loss per 100 feet and virtually eliminates the need for remote mounting of the transmit converter or amplifier.

Tuning the Driven Element

If the antenna is built carefully to the dimensions given, the SWR should be close to 1:1. Just to be sure, check the SWR if you have access to test equipment. Be sure the signal source is clean, however; wattmeters respond to "dirty" signals and can give erroneous readings. If problems are encountered, recheck all dimensions. If they look good, a minor improvement may be realized by changing the shape of the driven element. Slight bending of reflector 2 may also improve the SWR. When the desired match has been obtained, solder the point where the UT-141 jacket passes through the loop and brass bolt.

Tips for 1296-MHz Antenna Installations

Construction practices that are common on lower frequencies cannot be used on 1296 MHz. This is the most important reason why all who venture to these frequencies are not equally successful. First, when a proven design is used, copy it exactly¾don't change *anything*. This is especially true for antennas.

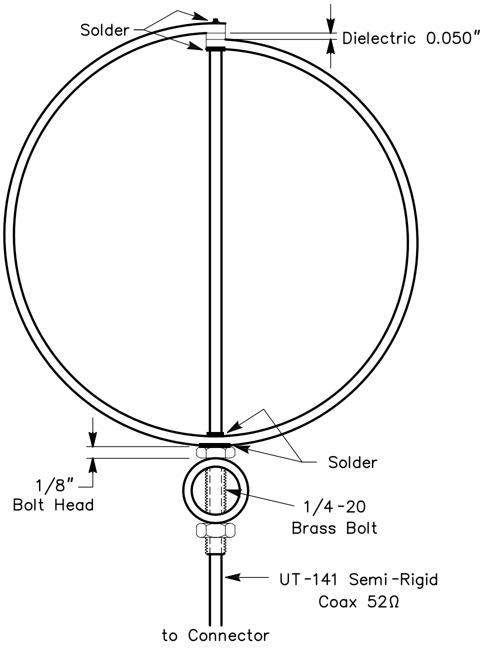

Fig 72—Driven-element details. See Fig 70 and the text for additional information.

Use the best feed line you can get. Here are some realistic measurements of common coaxial cables at 1296 MHz (loss per 100 feet).

RG-8, 213, 214: 11 dB
$^1/_2$ in. foam/copper Hardline: 4 dB
$^7/_8$ in. foam/copper Hardline: 1.5 dB

Mount the antennas to keep feed line losses to an absolute minimum. Antenna height is less important than keeping the line losses low. *Do not* allow the mast to pass through the elements, as is common on antennas for lower frequencies. Cut all U-bolts to the minimum length needed; $^1/_4$ λ at 1296 MHz is only a little over 2 inches. Avoid any unnecessary metal around the antenna.

Trough Reflectors for 432 and 1296 MHz

Dimensions are given in **Fig 73** for 432- and 1296-MHz trough reflectors. The gain to be expected is 16 dBi and 15 dBi, respectively. A very convenient arrangement, especially for portable work, is to use a metal hinge at each angle of the reflector. This permits the reflector to be folded flat for transit. It also permits experiments to be carried out with different apex angles.

A housing is required at the dipole center to prevent the entry of moisture and, in the case of the 432-MHz antenna, to support the dipole elements. The dipole may be moved in and out of the reflector to get either minimum SWR or, if this cannot be measured, maximum gain. If a two-stub tuner or other matching device is used, the dipole may be placed to give optimum gain and the matching device adjusted to give optimum match. In the case of the 1296-MHz antenna, the dipole length can be adjusted by means of the brass screws at the ends of the elements. Locking nuts are essential.

The reflector should be made of sheet aluminum for 1296 MHz, but can be constructed of wire mesh (with twists parallel to the dipole) for 432 MHz. To increase the gain by 3 dB, a pair of these arrays can be stacked so the reflectors are barely separated (to prevent the formation of a slot radiator by the edges). The radiating dipoles must then be fed in phase, and suitable feeding and matching must be arranged. A two-stub tuner can be used for matching either a single- or double-reflector system.

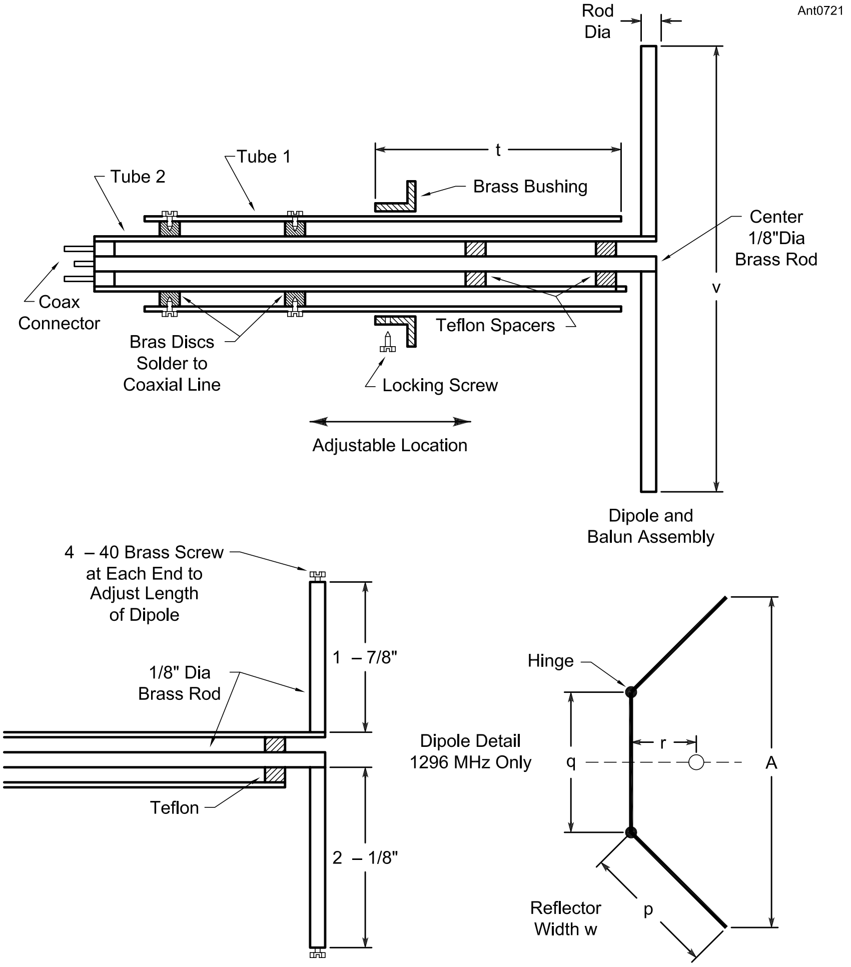
Ant0721
Rod Dia
Tube 1
Tube 2
t
Brass Bushing
Center 1/8"Dia Brass Rod
v
Coax Connector
Bras Discs Solder to Coaxial Line
Teflon Spacers
Locking Screw
Adjustable Location
Dipole and Balun Assembly
4 – 40 Brass Screw at Each End to Adjust Length of Dipole
1/8" Dia Brass Rod
1 – 7/8"
2 – 1/8"
Teflon
Dipole Detail 1296 MHz Only
Hinge
q
r
A
Reflector Width w
p

Dimensions in Inches

Band	A	p	q	r	w	t	u	v	Tube 1 Brass or Copper	Tube 2 Brass or Copper	Gain
432	120	100	50	7 – 1/2	60	6 – 1/2	1/4	12 – 3/4	1 – 1/2" ID 1/16" Wall	5/8" OD 1/16" Wall	16 dBi
1296	45	36	9 – 1/2	2 – 1/2	20	2 – 1/8	1/8	4 – 1/4	3/4" ID	3/8" OD 20 s.w.g.	15 dBi

Fig 73—Practical construction information for trough reflector antennas for 432 and 1296 MHz.

A Horn Antenna for 10 GHz

The horn antenna is the easiest antenna for the beginner on 10 GHz to construct. It can be made out of readily available flat sheet brass. Because it is inherently a broadband structure, minor constructional errors can be tolerated. The one drawback is that horn antennas become physically cumbersome at gains over about 25 dBi, but for most line-of-sight work this much gain is rarely necessary. This antenna was designed by Bob Atkins, KA1GT, and appeared in *QST* for April and May, 1987.

Horn antennas are usually fed by waveguide. When operating in its normal frequency range, waveguide propagation is in the TE_{10} mode. This means that the electric (E) field is across the short dimension of the guide and the magnetic (H) field is across the wide dimension. This is the reason for the E-plane and H-plane terminology shown in **Fig 74**.

There are many varieties of horn antennas. If the waveguide is flared out only in the H-plane, the horn is called an H-plane sectoral horn. Similarly, if the flare is only in the E-plane, an Eplane sectoral horn results. If the flare is in both planes, the antenna is called a pyramidal horn.

For a horn of any given aperture, directivity (gain along the axis) is maximum when the field distribution across the aperture is uniform in magnitude and phase. When the fields are not uniform, side lobes that reduce the directivity of the antenna are formed. To obtain a uniform distribution, the horn should be as long as possible with minimum flare angle. From a practical point of view, however, the horn should be as short as possible, so there is an obvious conflict between performance and convenience.

Fig 75 illustrates this problem. For a given flare angle and a given side length, there is a path-length difference from the apex of the horn to the center of the aperture (L), and from the apex of the horn to the edge of the aperture (L’). This causes a phase difference in the field across the aperture, which in turn causes formation of side lobes, degrading directivity (gain along the axis) of the antenna. If L is large this difference is small, and the field is almost uniform. As L decreases however, the phase difference increases and directivity suffers. An optimum (shortest possible) horn is constructed so that this phase difference is the maximum allowable before side lobes become excessive and axial gain markedly decreases.

The magnitude of this permissible phase difference is different for E-plane and H-plane horns. For the E-plane horn, the field intensity is quite constant across the aperture. For the H-plane horn, the field tapers to zero at the edge. Consequently, the phase difference at the edge of the aperture in the E-plane horn is more critical and should be held to less than 90° ($^1/_4$ λ). In an H-plane horn, the allowable phase difference is 144° (0.4 λ). If the aperture of a pyramidal horn exceeds one wavelength in both planes, the Eplane and Hplane patterns are essentially independent and can be analyzed separately.

The usual direction for orienting the waveguide feed is with the broad face horizontal, giving vertical polarization. If this is the case, the H-plane sectoral horn has a narrow horizontal beamwidth and a very wide vertical beamwidth. This is not a very useful beam pattern for most amateur applications. The E-plane sectoral horn has a narrow vertical beamwidth and a wide horizontal beamwidth. Such a radiation pattern could be useful in a beacon system where wide coverage is desired.

The most useful form of the horn for general applications is the optimum pyramidal horn. In this configuration the two beamwidths are almost the same. The E-plane (vertical) beamwidth is slightly less than the H-plane (horizontal), and also has greater side lobe intensity.

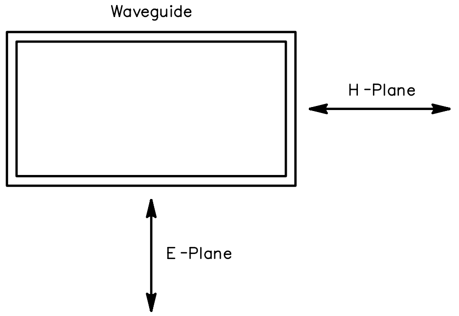

Fig 74—10-GHz antennas are usually fed with waveguide. See text for a discussion of waveguide propagation characteristics.

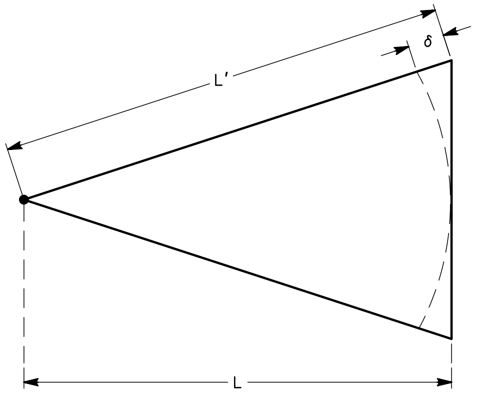

Fig 75—The path-length (phase) difference between the center and edge of a horn antenna is δ.

Building the Antenna

A 10-GHz pyramidal horn with 18.5 dBi gain is shown in **Fig 76**. The first design parameter is usually the required gain, or the maximum antenna size. These are of course related, and the relationships can be approximated by the following:

L = H-plane length (λ) = $0.0654 \times \text{gain}$ (Eq 1)
A = H-plane aperture (λ) = $0.0443 \times \text{gain}$ (Eq 2)
B = E-plane aperture (λ) = $0.81\ A$ (Eq 3)

where
gain is expressed as a *ratio*; 20 dBi gain = 100
L, A and B are dimensions shown in **Fig 77.**

From these equations, the dimensions for a 20-dBi gain horn for 10.368 GHz can be determined. One wavelength at 10.368 GHz is 1.138 inches. The length (L) of such a horn is $0.0654 \times 100 = 6.54\ \lambda$. At 10.368 GHz, this is 7.44 inches. The corresponding H-plane aperture (A) is 4.43 λ (5.04 inches), and the E-plane aperture (B), 4.08 inches.

The easiest way to make such a horn is to cut pieces from brass sheet stock and solder them together. Fig 77 shows the dimensions of the triangular pieces for the sides and a square piece for the waveguide flange. (A standard commercial waveguide flange could also be used.) Because the E-plane and H-plane apertures are different, the horn opening is not square. Sheet thickness is unimportant; 0.02 to 0.03 inch works well. Brass sheet is often available from hardware or hobby shops.

Note that the triangular pieces are trimmed at the apex to fit the waveguide aperture (0.9 × 0.4 inch). This necessitates that the length, from base to apex, of the smaller triangle (side B) is shorter than that of the larger (side A). Note that the length, S, of the two different sides of the horn must be the same if the horn is to fit together! For such a simple looking object, getting the parts to fit together properly requires careful fabrication.

The dimensions of the sides can be calculated with simple geometry, but it is easier to draw out templates on a sheet of cardboard first. The templates can be used to build a mock antenna to make sure everything fits together properly before cutting the sheet brass.

First, mark out the larger triangle (side A) on cardboard. Determine at what point its width is 0.9 inch and draw a line parallel to the base as shown in Fig 77. Measure the length of the side S; this is also the length of the sides of the smaller (side B) pieces.

Mark out the shape of the smaller pieces by first drawing a line of length B and then constructing a second line of length S. One end of line S is an end of line B, and the other is 0.2 inch above a line perpendicular to the center of line B as shown in Fig 76. (This procedure is much more easily followed than described.) These smaller pieces are made slightly oversize (shaded area in Fig 77) so you can construct the horn with solder seams on the outside of the horn during assembly.

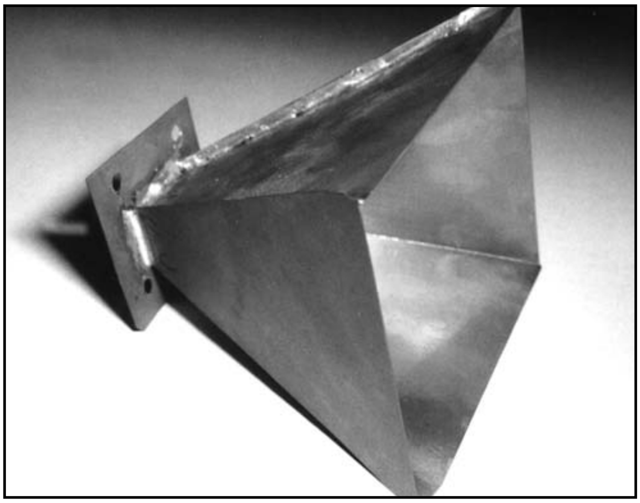

Fig 76—This pyramidal horn has 18.5 dBi gain at 10 GHz. Construction details are given in the text.

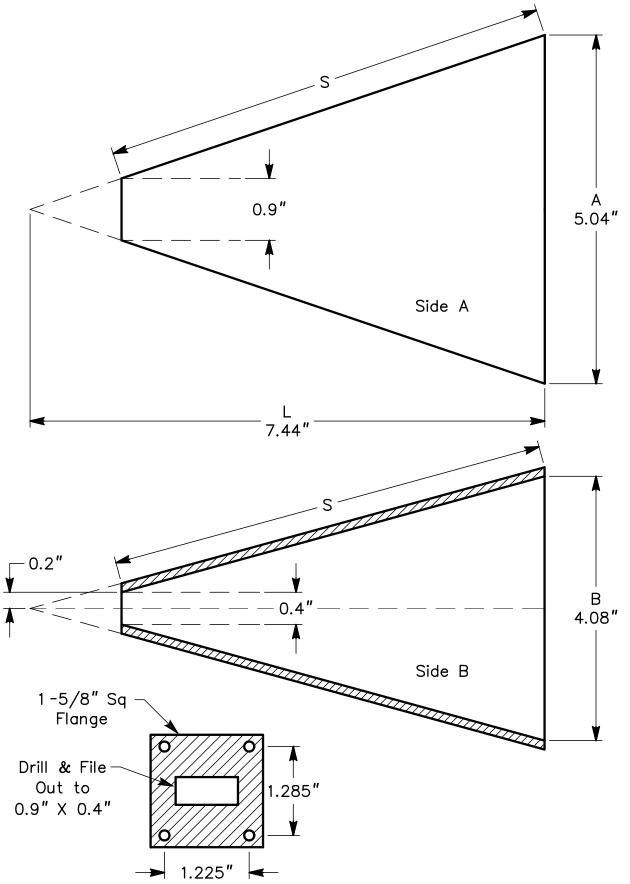

Fig 77—Dimensions of the brass pieces used to make the 10-GHz horn antenna. Construction requires two of each of the triangular pieces (side A and side B).

Cut out two cardboard pieces for side A and two for side B and tape them together in the shape of the horn. The aperture at the waveguide end should measure 0.9 × 0.4 inch and the aperture at the other end should measure 5.04 × 4.08 inches.

If these dimensions are correct, use the cardboard templates to mark out pieces of brass sheet. The brass sheet should be cut with a bench shear if one is available, because scissors type shears tend to bend the metal. Jig the pieces together and solder them on the *outside* of the seams. It is important to keep both solder and rosin from contaminating the inside of the horn; they can absorb RF and reduce gain at these frequencies.

Assembly is shown in **Fig 78**. When the horn is completed, it can be soldered to a standard waveguide flange, or one cut out of sheet metal as shown in Fig 77. The transition between the flange and the horn must be smooth. This antenna provides an excellent performance-to-cost ratio (about 20 dBi gain for about five dollars in parts).

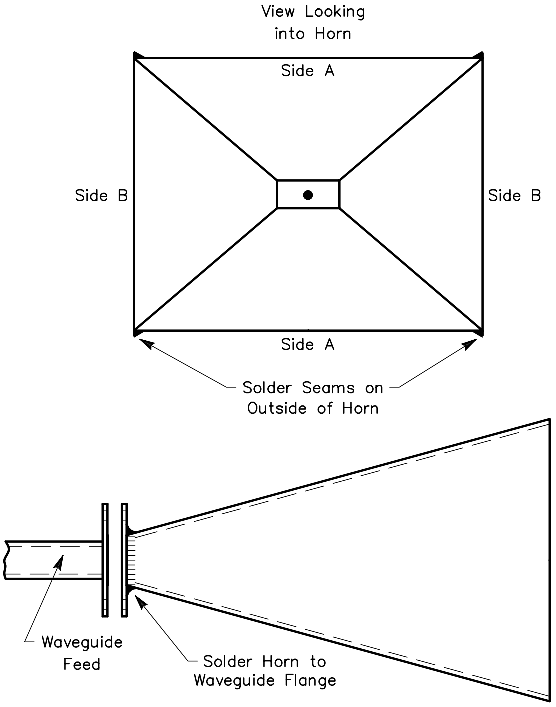
View Looking into Horn
Side A
Side B
Side B
Side A
Solder Seams on Outside of Horn
Waveguide Feed
Solder Horn to Waveguide Flange

Fig 78—Assembly of the 10-GHz horn antenna.

Periscope Antenna Systems

One problem common to all who use microwaves is that of mounting an antenna at the maximum possible height while trying to minimize feed-line losses. The higher the frequency, the more severe this problem becomes, as feeder losses increase with frequency. Because parabolic dish reflectors are most often used on the higher bands, there is also the difficulty of waterproofing feeds (particularly waveguide feeds). Inaccessibility of the dish is also a problem when changing bands. Unless the tower is climbed every time and the feed changed, there must be a feed for each band mounted on the dish. One way around these problems is to use a periscope antenna system (sometimes called a “flyswatter antenna”).

The material in this section was prepared by Bob Atkins, KA1GT, and appeared in *QST* for January and February 1984. **Fig 79** shows a schematic representation of a periscope antenna system. A plane reflector is mounted at the top of a rotating tower at an angle of 45°. This reflector can be elliptical with a major to minor axis ratio of 1.41, or rectangular. At the base of the tower is mounted a dish or other type of antenna such as a Yagi, pointing straight up. The advantage of such a system is that the feed antenna can be changed and worked on easily. Additionally, with a correct choice of reflector size, dish size, and dish to reflector spacing, feed losses can be made small, increasing the effective system gain. In fact, for some particular system configurations, the gain of the overall system can be greater than that of the feed antenna alone.

Gain of a Periscope System

Fig 80 shows the relationship between the effective gain of the antenna system and the distance between the reflector and feed antenna for an elliptical reflector. At first sight, it is not at all obvious how the antenna system can have a higher gain than the feed alone. The reason lies in the fact that, depending on the feed to reflector spacing, the reflector may be in the near field (Fresnel) region of the antenna, the far field (Fraunhöffer) region, or the transition region between the two.

In the far field region, the gain is proportional to the reflector area and inversely proportional to the distance between the feed and reflector. In the near field region, seemingly strange things can happen, such as decreasing gain with decreasing feed to reflector separation. The reason for this gain decrease is that, although the reflec-

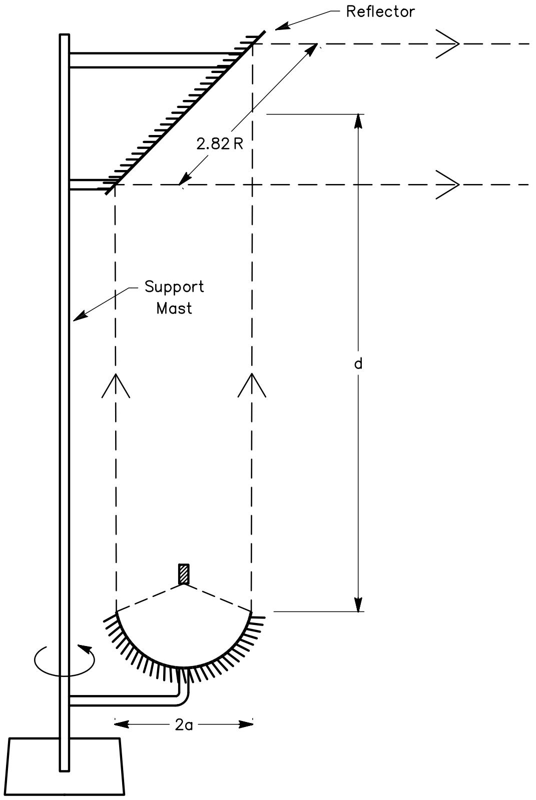

Fig 79—The basic periscope antenna. This design makes it easy to adjust the feed antenna.

tor is intercepting more of the energy radiated by the feed, it does not all contribute in phase at a distant point, and so the gain decreases.

In practice, rectangular reflectors are more common than elliptical. A rectangular reflector with sides equal in length to the major and minor axes of the ellipse will, in fact, normally give a slight gain increase. In the far field region, the gain will be proportional to the area of the reflector. To use Fig 80 with a rectangular reflector, R^2 may be replaced by A / π, where A is the projected area of the reflector. The antenna pattern depends in a complicated way on the system parameters (spacing and size of the elements), but **Table 21** gives an approximation of what to expect. R is the radius of the projected circular area of the elliptical reflector (equal to the minor axis radius), and b is the length of the side of the projected square area of the rectangular reflector (equal to the length of the short side of the rectangle).

For those wishing a rigorous mathematical analysis of this type of antenna system, several references are given in the Bibliography at the end of this chapter.

Mechanical Considerations

There are some problems with the physical construction of a periscope antenna system. Since the antenna gain of a microwave system is high and, hence, its beamwidth narrow, the reflector must be accurately aligned. If the reflector does not produce a beam that is horizontal, the

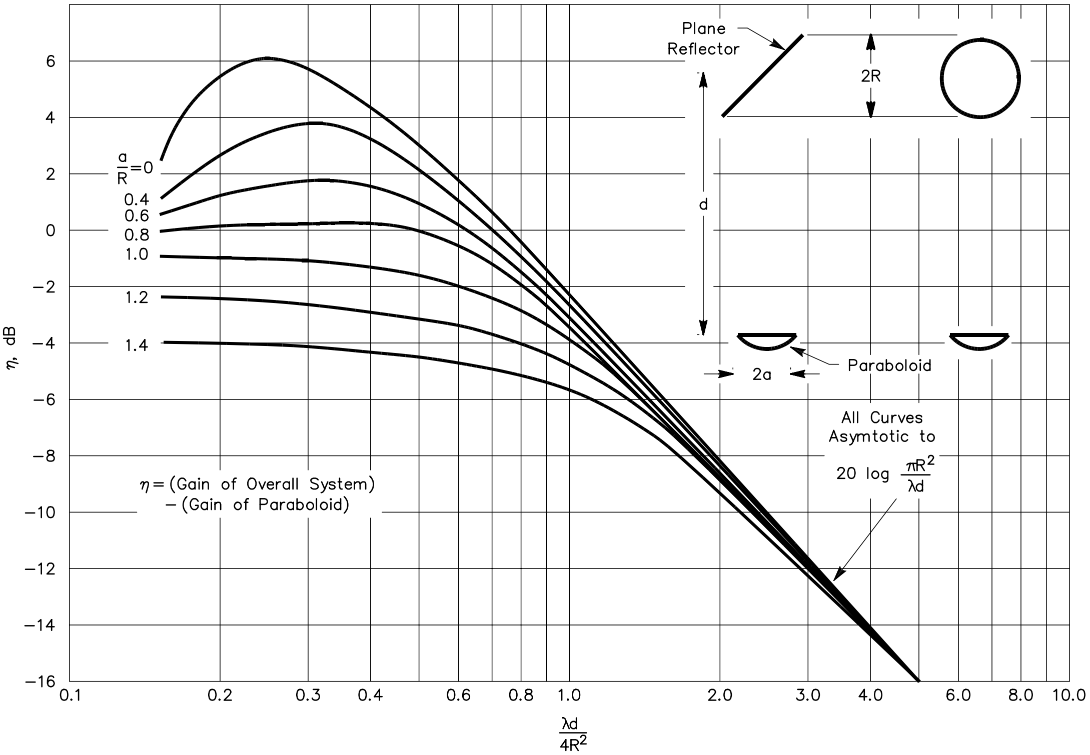

Fig 80—Gain of a periscope antenna using a plane elliptical reflector (after Jasik—see Bibliography).

Table 21
Radiation Patterns of Periscope Antenna Systems

	Elliptical Reflector	*Rectangular Reflector*
3-dB beamwidth, degrees	60 $\lambda/2R$	52 λ/b
6-dB beamwidth, degrees	82 $\lambda/2R$	68 λ/b
First minimum, degrees from axis	73 $\lambda/2R$	58 λ/b
First maximum, degrees from axis	95 $\lambda/2R$	84 λ/b
Second minimum, degrees from axis	130 $\lambda/2R$	116 λ/b
Second maximum, degrees from axis	156 $\lambda/2R$	142 λ/b
Third minimum, degrees from axis	185 $\lambda/2R$	174 λ/b

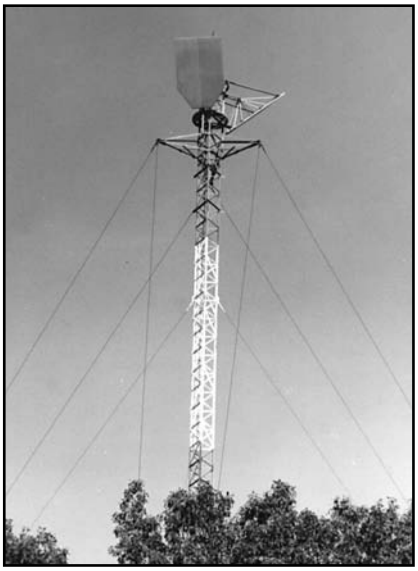

Fig 81—Commercial periscope antennas, such as this one, are often used for point-to-point communication.

useful gain of the system will be reduced. From the geometry of the system, an angular misalignment of the reflector of X degrees in the vertical plane will result in an angular misalignment of 2X degrees in the vertical alignment of the antenna system pattern. Thus, for a dish pointing straight up (the usual case), the reflector must be at an angle of 45° to the vertical and should not fluctuate from factors such as wind loading.

The reflector itself should be flat to better than $^1/_{10}$ λ for the frequency in use. It may be made of mesh, provided that the holes in the mesh are also less than $^1/_{10}$ λ in diameter. A second problem is getting the support mast to rotate about a truly vertical axis. If the mast is not vertical, the resulting beam will swing up and down from the horizontal as the system is rotated, and the effective gain at the horizon will fluctuate. Despite these problems, amateurs have used periscope antennas successfully on the bands through 10 GHz. Periscope antennas are used frequently in commercial service, though usually for point-to-point transmission. Such a commercial system is shown in **Fig 81**.

Circular polarization is not often used for terrestrial work, but if it is used with a periscope system there is an important point to remember. The circularity sense changes when the signal is reflected. Thus, for right hand circularity with a periscope antenna system, the feed arrangement on the ground should produce left hand circularity. It should also be mentioned that it is possible (though more difficult for amateurs) to construct a periscope antenna system using a parabolically curved reflector. The antenna system can then be regarded as an offset fed parabola. More gain is available from such a system at the added complexity of constructing a parabolically curved reflector, accurate to $^1/_{10}$ λ.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2, Antenna Fundamentals.

B. Atkins, "Periscope Antenna Systems," The New Frontier, *QST*, Jan and Feb 1984.

B. Atkins, "Horn Antennas for 10 GHz," The New Frontier, *QST*, Apr and May 1987.

J. Drexler, "An Experimental Study of a Microwave Periscope," *Proc. IRE*, Correspondence, Vol 42, Jun 1954, p 1022.

D. Evans and G. Jessop, *VHF-UHF Manual*, 3rd ed. (London: RSGB), 1976.

N. Foot, "WA9HUV 12foot Dish for 432 and 1296 MHz," The World Above 50 Mc., *QST*, Jun 1971, pp 98-101, 107.

N. Foot, "Cylindrical Feed Horn for Parabolic Reflectors," *Ham Radio*, May 1976.

G. Gobau, "Single-Conductor Surface-Wave Transmission Lines," *Proc. IRE*, Vol 39, Jun 1951, pp 619-624; also see *Journal of Applied Physics*, Vol 21 (1950), pp 1119-1128.

R. E. Greenquist and A. J. Orlando, "An Analysis of Passive Reflector Antenna Systems," *Proc. IRE*, Vol 42, Jul 1954, pp 1173-1178.

G. A. Hatherell, "Putting the G Line to Work," *QST*, Jun 1974, pp 11-15, 152, 154, 156.

D. L. Hilliard, "A 902MHz Loop Yagi Antenna," *QST*, Nov 1985, pp 30-32.

W. C. Jakes, Jr., "A Theoretical Study of an Antenna-Reflector Problem," *Proc. IRE*, Vol 41, Feb 1953, pp 272-274.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

R. T. Knadle, “UHF Antenna Ratiometry,” *QST*, Feb 1976, pp 22-25.

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw- Hill, 1948).

W. Overbeck, “The VHF Quagi,” *QST*, Apr 1977, pp 11-14.

W. Overbeck, “The Long-Boom Quagi,” *QST*, Feb 1978, pp 20-21.

W. Overbeck, “Reproducible Quagi Antennas for 1296 MHz,” *QST*, Aug 1981, pp 11-15.

G. Southworth, *Principles and Applications of Waveguide Transmission* (New York: D. Van Nostrand Co, 1950).

P. P. Viezbicke, “Yagi Antenna Design,” *NBS Technical Note 688* (U. S. Dept. of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

D. Vilardi, “Easily Constructed Antennas for 1296 MHz,” *QST*, Jun 1969.

D. Vilardi, “Simple and Efficient Feed for Parabolic Antennas,” *QST*, Mar 1973.

Radio Communication Handbook, 5th ed. (London: RSGB, 1976).

Chapter 19

Antenna Systems for Space Communications

When we consider amateur space communications, we usually think about two basic modes: satellite and earth-moon-earth (EME—also referred to as *moonbounce*). At their essence, both modes communicate using one of the Earth's satellites—our natural satellite (the Moon) or one of a variety of manmade satellites.

There are two main differences between these satellites. The first is one of distance. The Moon is about 250,000 miles from Earth, while man-made satellites can be as far as 36,000 miles away. This 7:1 difference in distance makes a huge difference in the signals that arrive at the satellite, since transmission loss varies as the square of the distance. In other words, the signal arriving at the Moon is 20 dB weaker than that arriving at a geo-synchronous satellite 25,000 miles high, due to distance alone.

The second difference between the Moon and a man-made satellite is that the Moon is a *passive reflector*—and not a very good one at that, since it has a craggy and rather irregular surface, at least when compared to a flat mirror-like surface that would make an ideal reflector. Signals scattered by the Moon's irregular surface are thus weaker than for better reflecting surfaces. By comparison, a man-made satellite is an *active* system, where the satellite receives the signal coming from Earth, amplifies it and then retransmits the signal (usually at a different frequency) using a high-gain antenna. Think of a satellite as an ideal reflector, with gain.

The net result of these differences between a man-made satellite and the Earth's natural satellite is that moonbounce (EME) operation challenges the station builder considerably more than satellite operation, particularly in the area of antennas. Successful EME requires high transmitting power, superb receiver sensitivity and excellent operators capable of pulling weak signals out of the noise. This chapter will first explore antennas suitable for satellite operations and then describe techniques needed for EME work.

Common Ground

There are areas of commonality between satellite and EME antenna requirements, of course. Both require consideration of the effects of polarization and elevation angle, along with the azimuth directions of transmitted and received signals.

On the HF bands, signal polarization is generally of little concern, since the original polarization sense is lost after the signal passes through the ionosphere. At HF, vertical antennas receive sky-wave signals emanating from horizontal antennas, and vice versa. It is not beneficial to provide a means of varying the polarization at HF. With satellite communications, however, because of polarization changes, a signal that would disappear into the noise on one antenna may be S9 on one that is not sensitive to polarization direction. Elevation angle is also important from the standpoint of tracking and avoiding indiscriminate ground reflections that may cause nulls in signal strength.

Antennas for Satellite Work

We have amateur satellites providing links from 2 meters and up, and these provide opportunities to use antennas of many types—from the very simple to some pretty complex ones. This section was written by Dick Jansson, WD4FAB. It covers descriptions of a wide range of satellite antennas and points operators to source material for construction of many of them.

Antennas for LEO Satellites

Antenna design and construction requirements for use with Amateur satellites vary from low-gain antennas for low-earth-orbit (LEO) satellites to higher-gain antennas for the high-altitude elliptical-orbit satellites. You can operate the FM LEO satellites with a basic dual-band VHF/UHF FM transceiver or even a good FM H-T, as some amateurs have managed. Assuming that the transceiver is reasonably sensitive, you can even use a good "rubber duck" antenna. Some amateurs manage to work the FM birds with H-Ts and a multi-element directional antenna such as the popular Arrow Antenna, **Fig 1**. Of course, this means they must aim their antennas at the satellites, even as they cross overhead.

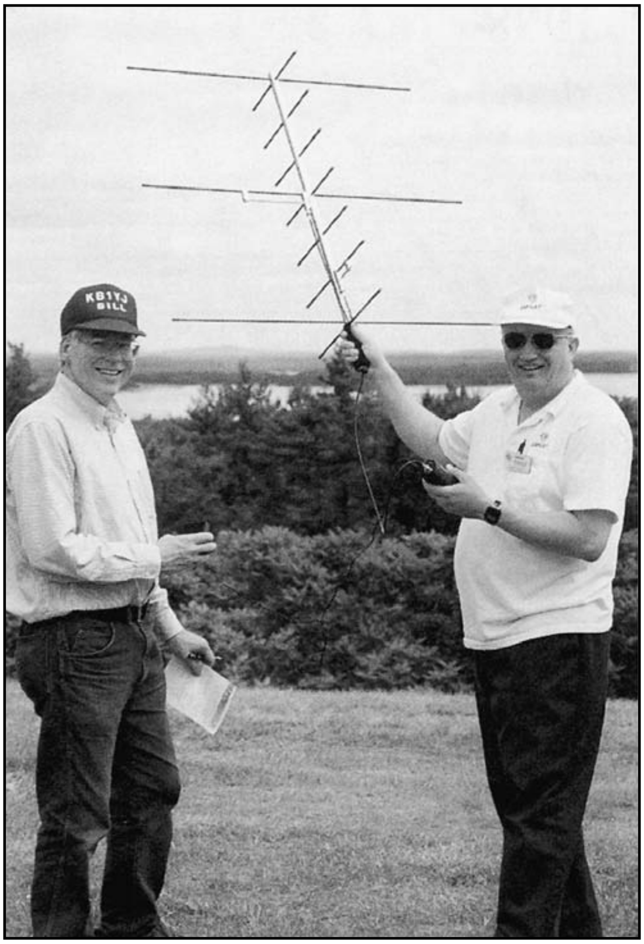

Fig 1—The hand-held "Arrow" gain antenna is popular for LEO FM operations. (*Photo courtesy* The AMSAT Journal, *Sept/Oct 1998.*)

High-quality omnidirectional antennas for LEO service come in quite a number of forms and shapes. M^2 Enterprises has their EB-144 and EB-432 Eggbeater antennas, which have proven to be very useful and do not require any rotators for control. See **Fig 2**. The turnstile-over-reflector antenna has been around for a long time, as shown in **Fig 3**. Other operators have done well using low-gain Yagi antennas, such as those shown in **Fig 4**.

For even better performance, at the modest cost of a single, simple TV antenna rotator, check out the fixed-elevation *Texas Potato Masher* antenna by K5OE, **Fig 5.** This antenna provides a dual-band solution for medium-gain directional antennas for LEO satellites. This is a considerable improvement over omnidirectional antennas and does not require an elevation rotator for good performance.

There are still two LEO satellites that work on the 10-meter band, RS-15 and the newly resurrected AO-7. Both have 10-meter downlinks in the range of 29.3 to 29.5 MHz. Low-gain 10-meter antennas, such as dipoles or long-wire antennas, are used to receive these satellites.

Antennas for High-Altitude Satellites

The high-altitude, Phase-3 satellites, such as AO-10 (and the late AO-13), have been around for quite a number of years. The greater distances to the Phase-3 satellites mean that more transmitted power is needed to access them and weaker signals are received on the ground. Successful stations usually require ground-station antennas with significant gain (12 dBi or more), such as a set of high-gain Yagi antennas. See **Fig 6**. Note the use of two Yagi antennas mounted on each boom to provide *circular polarization*, usually referred to as CP.

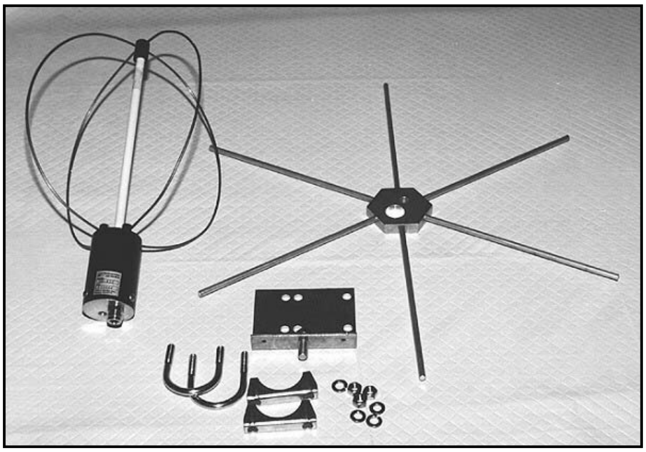

Fig 2—Eggbeater antennas are popular for base station LEO satellite operations. This EB-432 eggbeater antenna for 70 cm is small enough to put in an attic. Antenna gain pattern is helped with the radials placed below the antenna.

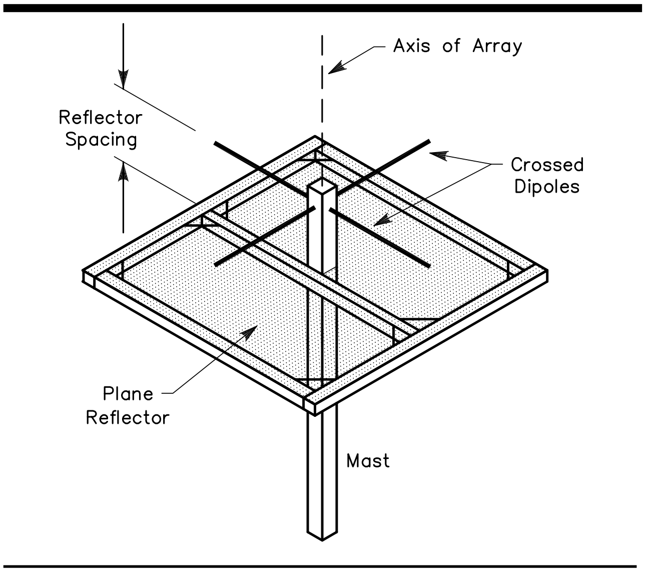

Fig 3—The Turnstile Over Reflector antenna has served well for LEO satellite service for a number of years.

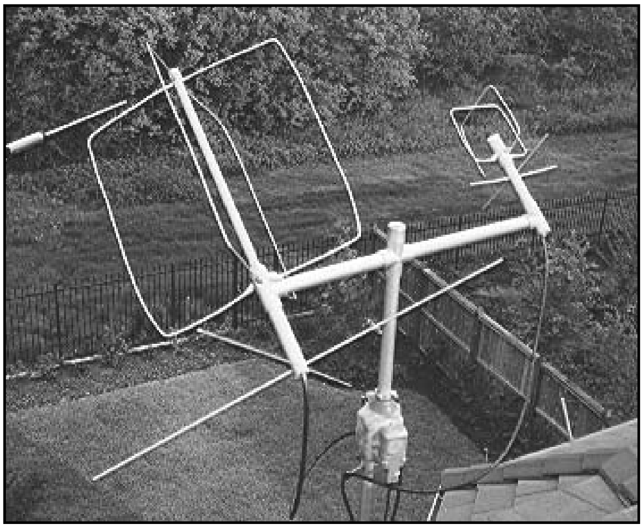

Fig 5—Jerry Brown, K5OE, uses his Texas Potato Masher antennas to work LEO satellites.

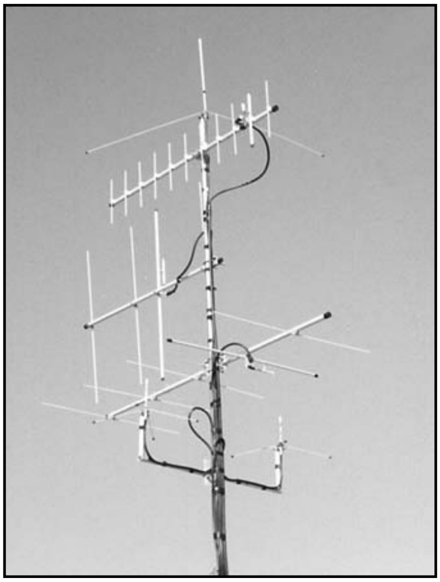

Fig 4—Simple ground plane and Yagi antennas can be used for LEO satellite contacts.

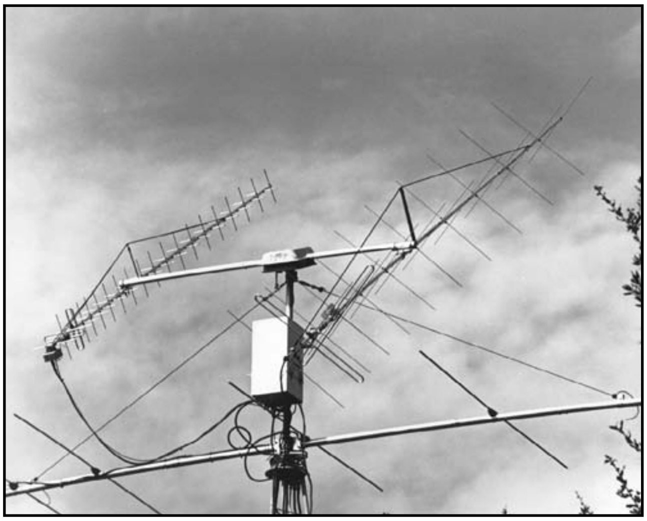

Fig 6—Dick Jansson, WD4FAB, used these 2-meter and 70-cm crossed Yagi's in RHCP for AO-10 and AO-13 operations. The satellite antennas are shown mounted above a 6-meter long-boom Yagi.

CIRCULAR POLARIZATION

Linearly polarized antennas are horizontal or vertical in terms of the antenna's position relative to the surface of the Earth, a reference that loses its meaning in space. The need to use circularly polarized (CP) antennas for space communications is well established. If spacecraft antennas used linear polarization, ground stations would not be able to maintain polarization alignment with the spacecraft because of changing orientations. The ideal antenna for random satellite polarizations is one with a circularly polarized radiation pattern.

There are two commonly used methods for obtaining circular polarization. One is with crossed linear elements such as dipoles or Yagis, as Fig 6 shows. The second popular CP method uses a helical antenna, described below. Other methods also exist, such as with the omnidirectional quadrifilar helix, **Fig 7**.

Polarization *sense* is a critical factor, especially in EME and satellite work. The IEEE standard uses the term "clockwise circular polarization" for a *receding* wave. Amateur technology follows the IEEE standard, calling clockwise polarization for a receding wave as right-hand, or RHCP. Either clockwise or a counter-clockwise (LHCP) sense can be selected by reversing the phasing harness of a crossed-Yagi antenna, see **Fig 8**. The sense of a helical antenna is fixed, determined by its physical construction.

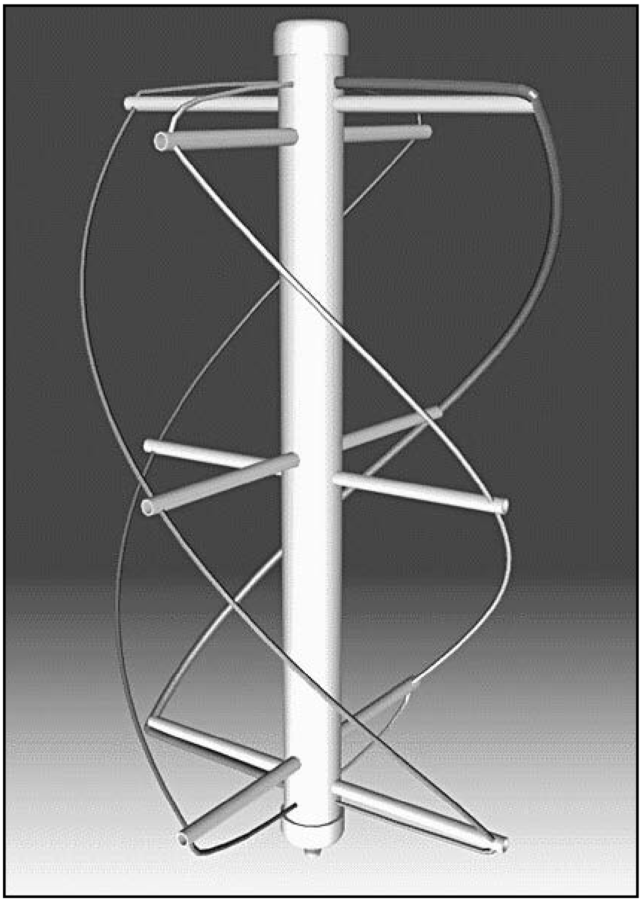

Fig 7—W3KH suggests that quadrifilar antennas can serve well for omnidirectional satellite-station antenna service.

See **Fig 9** for construction details.

In working through a satellite with a circularly polarized antenna, it is often convenient to have the capability of switching polarization sense. This is because the sense of the received signal of some of the LEO satellites reverses when the satellite passes its nearest point to you. If the received signal has right-hand circular polarization as the satellite approaches, it may have left-hand circularity as the satellite recedes. There is a sense reversal in EME work, as well, because of a phase reversal of the signal as it is reflected from the surface of the moon. A signal transmitted with right-hand circularity will be returned to the Earth with left-hand circularity. Similarly, the polarization is reversed as it is reflected from a dish antenna, so that for an overall RHCP performance, the feed antenna for the dish needs to be LHCP.

Crossed Linear Antennas

Dipoles radiate linearly polarized signals, and the polarization direction depends on the orientation of the antenna. It two dipoles are arranged for horizontal and vertical dipoles, and the two outputs are combined with the correct phase difference (90°), a circularly polarized wave results. Because the electric fields are identical in magnitude, the power from the transmitter will be equally divided between the two fields. Another way of looking at this is to consider the power as being divided between the two antennas; hence the gain of each is decreased by 3 dB when taken alone in the plane of its orientation.

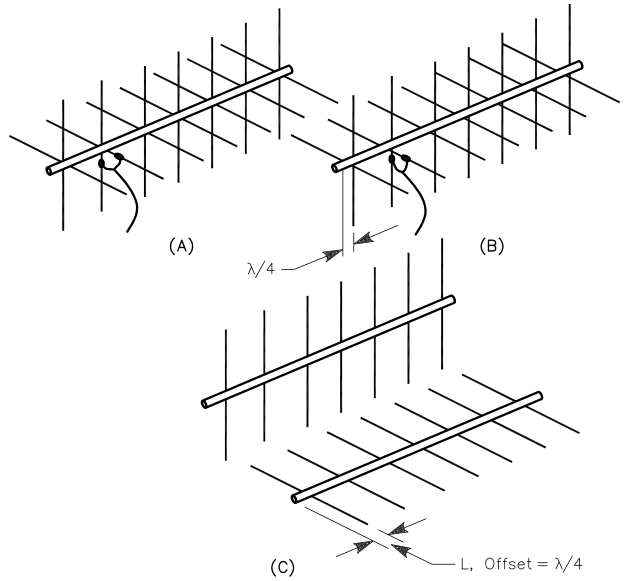

Fig 8—Evolution of the circularly polarized Yagi. The simplest form of crossed Yagi, A, is made to radiate circularly by feeding the two driven elements 90° out of phase. Antenna B has the driven elements fed in phase, but has the elements of one bay mounted $^1/_4$ λ forward from those of the other. Antenna C offers elliptical (circular) polarization using separate booms. The elements in one set are perpendicular to those of the other and are $^1/_4$ λ forward from those of the other.

A 90° phase shift must exist between the two antennas and the simplest way to obtain this shift is to use two feed lines to a coplanar pair of crossed-Yagi antennas. One feed-line section is $^1/_4$ λ longer than the other, as shown in Fig 8A. These separate feed lines are then paralleled to a common transmission line to the transmitter or receiver. Therein lies one of the headaches of this system. Assuming negligible coupling between the crossed antennas, the impedance presented to the common transmission line by the parallel combination is one half that of either section alone. (This is not true when there is mutual coupling between the antennas, as in phased arrays.) A practical construction method for implementing a RHCP/LHCP coplanar switched system is shown in **Fig 10**.

Another example of a coplanar crossed-Yagi antenna is shown in **Fig 11**. With this phasing-line method, any mismatch at one antenna will be magnified by the extra $^1/_4$ λ of transmission line. This upsets the current balance between the two antennas, resulting in a loss of polarization circularity. Another factor to consider is the attenuation of the cables used in the harness, along with the

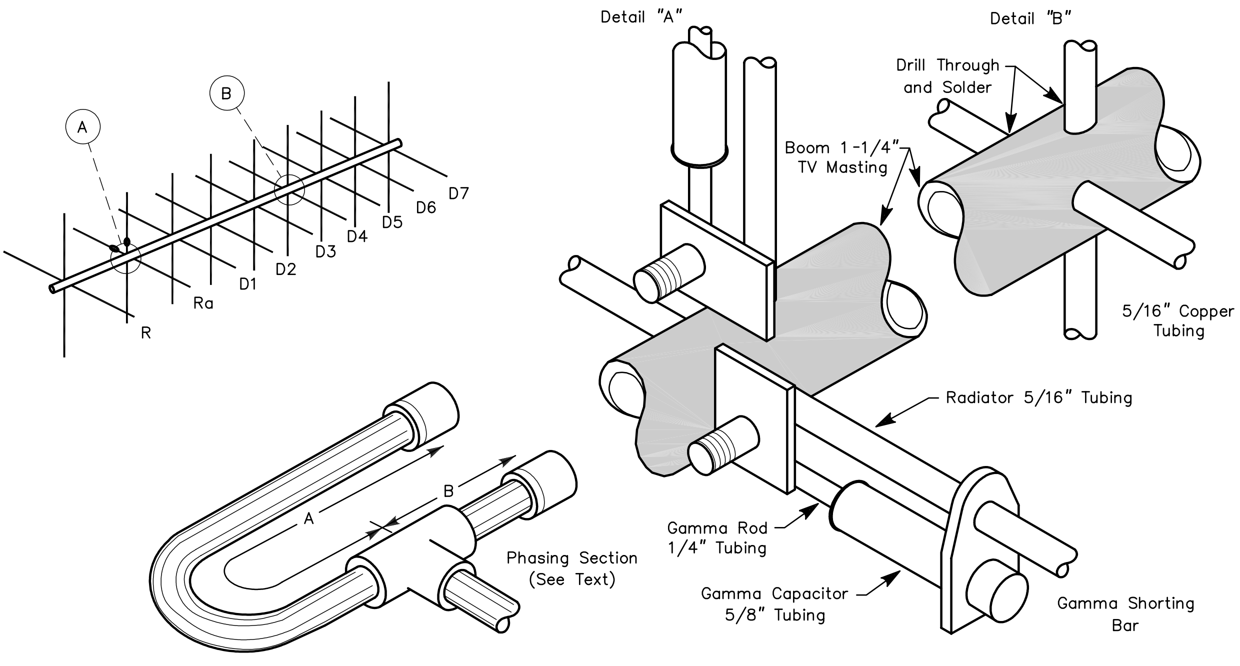
Detail "A"
Detail "B"
Drill Through and Solder
Boom 1-1/4" TV Masting
5/16" Copper Tubing
Radiator 5/16" Tubing
Gamma Rod 1/4" Tubing
Gamma Capacitor 5/8" Tubing
Gamma Shorting Bar
Phasing Section (See Text)
A
B
R
Ra
D1
D2
D3
D4
D5
D6
D7

Fig 9—Construction details of a co-planar crossed-Yagi antenna.

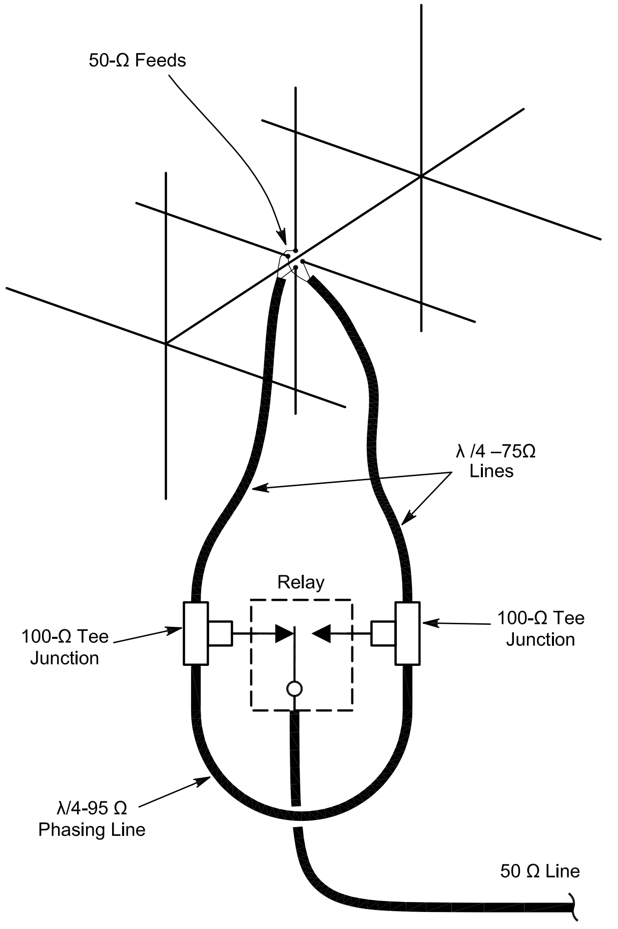
50-Ω Feeds
λ /4 –75Ω Lines
Relay
100-Ω Tee Junction
100-Ω Tee Junction
λ/4-95 Ω Phasing Line
50 Ω Line

Fig 10—Co-planar crossed Yagi, circularly polarized antenna with switchable polarization phasing harness.

connectors. Good low-loss coaxial line should be used. Type-N or BNC connectors are preferable to the UHF variety.

Another method to obtain circular polarization is to use equal-length feed lines and place one antenna $^{1}/_{4}\,\lambda$ ahead of the other. This offset pair of Yagi-crossed antennas is shown in Fig 8B. The advantage of equal-length feed lines is that identical load impedances will be presented to the common feeder, as shown in **Fig 12**, which shows a fixed circularity sense feed. To obtain a switchable sense feed with the offset Yagi pair, you can use a connection like that of **Fig 13**, although you must compensate for the extra phase added by the relay and connectors.

Fig 8C diagrams a popular method of mounting two separate off-the-shelf Yagis at right angles to each other. The two Yagis may be physically offset by $^{1}/_{4}\,\lambda$ and fed in parallel, as shown in Fig 8C, or they may be mounted with no offset and fed 90° out of phase. Neither of these arrangements on two separate booms produces true circular polarization. Instead, *elliptical* polarization results from such a system. **Fig 14** is a photo of this type of mounting of Yagis on two booms for elliptical operation.

Helical Antennas

As mentioned, the second method to create a circularly polarized signal is by means of a helical antenna. The axial-mode helical antenna was introduced by Dr John Kraus, W8JK, in the 1940s. **Fig 15** shows examples of S-band (2400-MHz), V-band (145-MHz), and U-band (435-MHz) helical antennas, all constructed by WD4FAB for satellite service.

This antenna has two characteristics that make it

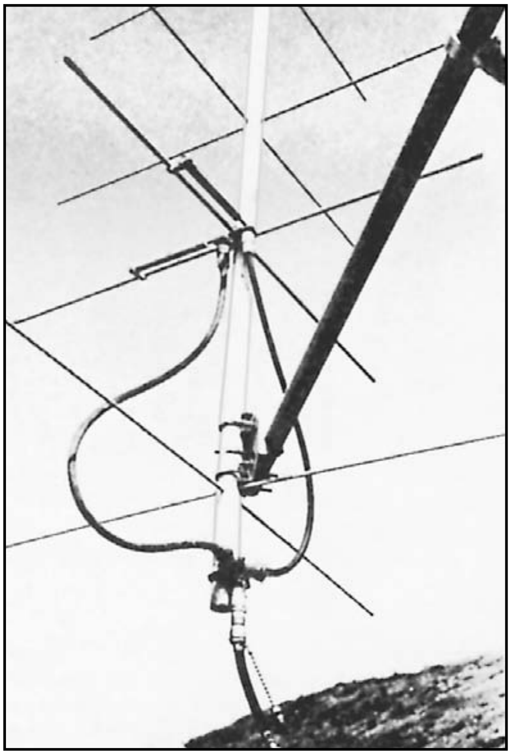

Fig 11—This VHF crossed Yagi design by KH6IJ (Jan 1973 *QST*) illustrates the co-planar, fixed-circularity Yagi.

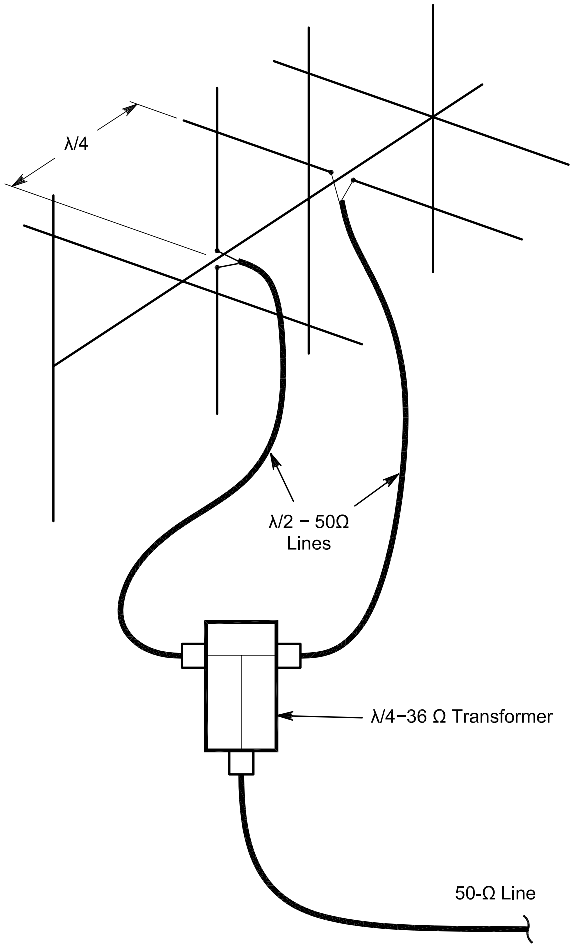

Fig 12—Offset crossed-Yagi circularly polarized antenna-phasing harness with fixed polarization.

especially interesting and useful in many applications. First, the helix is circularly polarized. As discussed earlier, circular polarization is simply linear polarization that continually rotates as it travels through space. In the case of a helical antenna, this rotation is about the axis of the antenna. This can be pictured as the second hand of a watch moving at the same rate as the applied frequency, where the position of the second hand can be thought of as the instantaneous polarization of the signal.

The second interesting property of the helical antenna is its predictable pattern, gain and impedance characteristics over a wide frequency range. This is one of the few antennas that has both broad bandwidth and high gain. The benefit of this property is that, when used for narrow-band applications, the helical antenna is very forgiving of mechanical inaccuracies.

Probably the most common amateur use of the helical antenna is in satellite communications, where the spinning of the satellite antenna system (relative to the earth) and the effects of Faraday rotation cause the polarization of the satellite signal to be unpredictable. Using a linearly polarized antenna in this situation results in deep fading, but with the helical antenna (which responds equally to linearly polarized signals), fading is essentially eliminated.

This same characteristic makes helical antennas useful in polarization-diversity systems. The advantages of circular polarization have been demonstrated on VHF voice schedules over non-optical paths, in cases where linearly polarized beams did not perform satisfactorily.

Another use for the helical antenna is the transmission of color ATV signals. Many beam antennas (when adjusted for maximum gain) have far less bandwidth than the required 6 MHz, or have non-uniform gain over this frequency range. The result is significant distortion of the transmitted and received signals, affecting color reproduction and other features. This problem becomes more aggravated over non-optical paths. The helix exhibits maximum gain (within 1 dB) more than 20 MHz anywhere above 420 MHz.

The helical antenna can be used to advantage with multimode rigs, especially above 420 MHz. Not only does the helix give high gain over an entire amateur band, but it also allows operation on FM, SSB and CW without the need for separate vertically and horizontally polarized antennas.

Helical Antenna Basics

The helical antenna is an unusual specimen in the antenna world, in that its physical configuration gives a

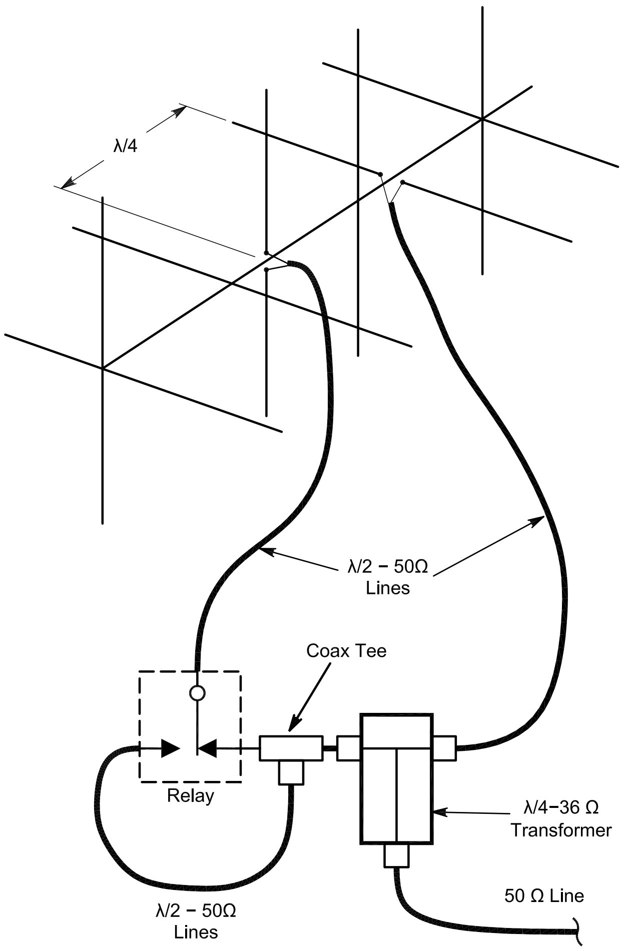

Fig 13—Offset crossed-Yagi circularly polarized antenna-phasing harness with switchable polarization.

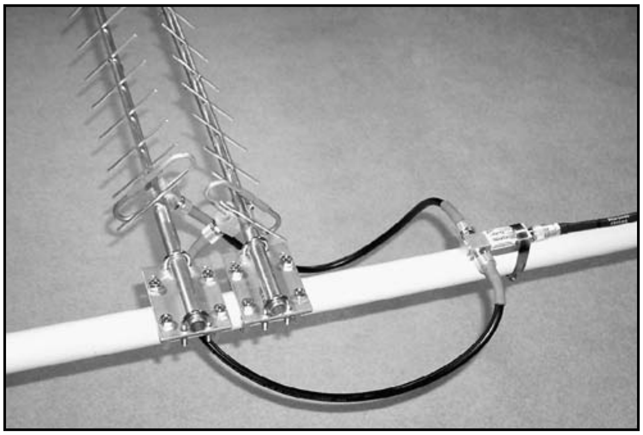

Fig 14—An example of offset crossed-Yagi circularly polarized antennas with fixed polarization. This example is a pair of M^2 23CM22EZA antennas, for L band (1269 MHz), mounted on an elevation boom. (*WD4FAB photo.*)

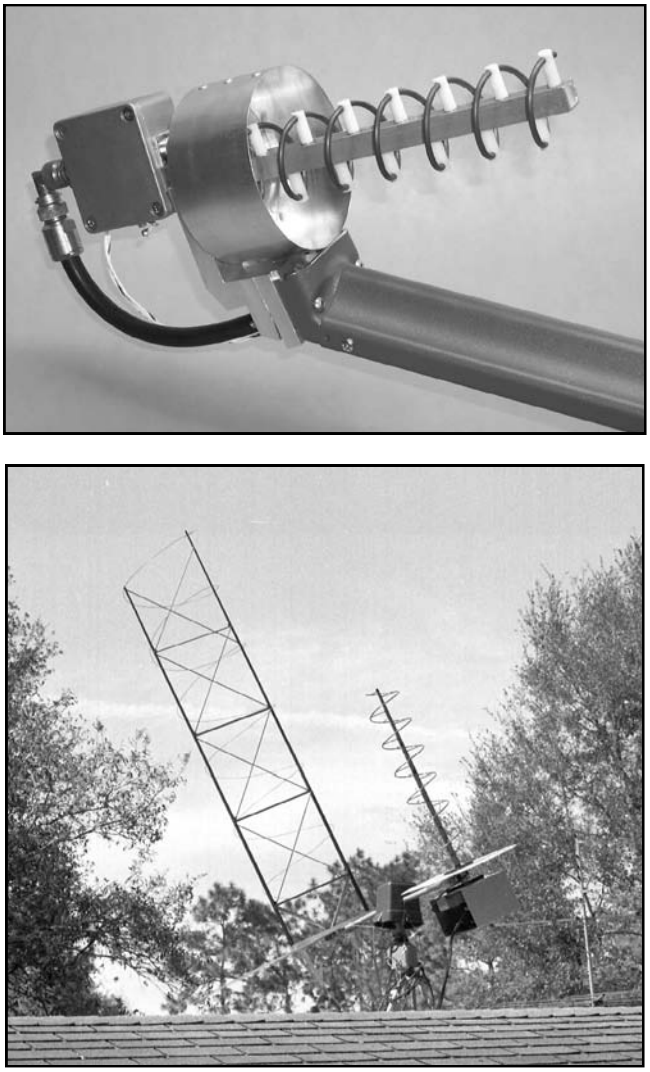

Fig 15—At top, a seven-turn LHCP helical antenna for S-band dish feed for AO-40 service. This helical antenna uses a cupped reflector and has a preamplifier mounted directly to the antenna feed point. At bottom, a pair of helical antennas for AO-10 service on 2 meters and 70 cm. The 2-meter helical antenna is not small! (*WD4FAB photos.*)

hint to its electrical performance. A helix looks like a large air-wound coil with one of its ends fed against a ground plane, as shown in **Fig 16**. The ground plane is a screen of 0.8 λ to 1.1 λ diameter (or on a side for a square ground plane). The circumference (C_λ) of the coil form must be between 0.75 λ and 1.33 λ for the antenna to radiate in the axial mode. The coil should have at least three turns to radiate in this mode. The ratio of the spacing between turns (in wavelengths), S_λ to C_λ, should be in the range of 0.2126 to 0.2867. This ratio range results from the requirement that the pitch angle, α, of the helix be between 12° and 16°, where:

$$\alpha = \arctan \frac{S_\lambda}{C_\lambda} \qquad \text{(Eq 1)}$$

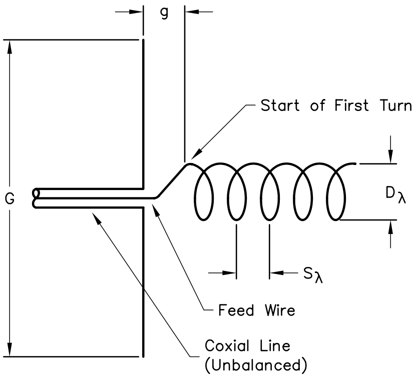

Fig 16—The basic helical antenna and design equations.

These constraints result in a single main lobe along the axis of the coil. This is easily visualized from Fig 15A. The winding of the helix comes away from the cupped reflector with a counterclockwise winding direction for a LHCP. (The winding can also be a clockwise—this results in a RHCP polarization sense.)

A helix with a C_λ of 1λ has a wave propagating from one end of the coil (at the ground plane), corresponding to an instantaneous dipole "across" the helix. The electrical rotation of this dipole produces circularly polarized radiation. Because the wave is moving along the helix conductor at nearly the speed of light, the rotation of the electrical dipole is at a very high rate, and true circular polarization results.

The IEEE definition, in simple terms, is that when viewing the antenna from the feed-point end, a clockwise wind results in right-hand circular polarization (RHCP), and a counterclockwise wind results in left-hand circular polarization (LHCP). This is important, because when two stations use helical antennas over a nonreflective path, both must use antennas with the same polarization sense. If antennas of opposite sense are used, a signal loss of at least 20 dB results from the cross polarization alone.

As mentioned previously, circularly polarized antennas can be used in communications with any linearly polarized antenna (horizontal or vertical), because circularly polarized antennas respond equally to all linearly polarized signals. The gain of a helix is 3 dB less than the theoretical gain in this case, because the linearly polarized antenna does not respond to linear signal components that are orthogonally polarized relative to it.

The response of a helix to all polarizations is indicated by a term called *axial ratio*, also known as *circularity*. Axial ratio is the ratio of amplitude of the polarization that gives maximum response to the amplitude of the polarization that gives minimum response. An ideal circularly polarized antenna has an axial ratio of 1.0. A well-designed practical helix exhibits an axial ratio of 1.0 to 1.1. The axial ratio of a helix is:

$$AR = \frac{2n+1}{2n} \tag{Eq 2}$$

where:

AR = axial ratio
n = the number of turns in the helix

Axial ratio can be measured in two ways. The first is to excite the helix and use a linearly polarized antenna with an amplitude detector to measure the axial ratio directly. This is done by rotating the linearly polarized antenna in a plane perpendicular to the axis of the helix and comparing the maximum and minimum amplitude values. The ratio of maximum to minimum is the axial ratio.

The impedance of the helix is easily predicted. The terminal impedance of a helix is unbalanced, and is defined by:

$$Z = 140 \times C_\lambda \tag{Eq 3}$$

where Z is the impedance of the helix in ohms.

The gain of a helical antenna is determined by its physical characteristics. Gain can be calculated from:

$$\text{Gain}\,(\text{dBi}) = 11.8 + 10\log\left(C_\lambda^2\, nS_\lambda\right) \tag{Eq 4}$$

In practice, helical antennas do not deliver the gain in Eq 4 for antennas with turns count greater than about twelve. There will be more discussions in this area when practical antennas are discussed.

The beamwidth of the helical antenna (in degrees) at the half-power points is:

$$BW = \frac{52}{C_\lambda \sqrt{nS_\lambda}} \tag{Eq 5}$$

The diameter of the helical antenna conductor should be between 0.006 λ and 0.05 λ, but smaller diameters have been used successfully at 144 MHz. The previously noted diameter of the ground plane (0.8 λ to 1.1 λ) should not be exceeded if you desire a clean radiation pattern. As the ground plane size is increased, the sidelobe levels also increase. Cupped ground planes have been used according to Kraus, as in Fig 15. (The ground plane need not be solid; it can be in the form of a spoked wheel or a frame covered with hardware cloth or screen.)

50-Ω Helix Feed

Joe Cadwallader, K6ZMW, presented this feed method in June 1981 *QST*. Terminate the helix in an N connector mounted on the ground screen at the periphery of the helix. See **Fig 17**. Connect the helix conductor to the N connector as close to the ground screen as possible (**Fig 18**). Then adjust the first quarter turn of the helix to a close spacing from the reflector.

This modification goes a long way toward curing a deficiency of the helix—the 140-Ω nominal feed-point

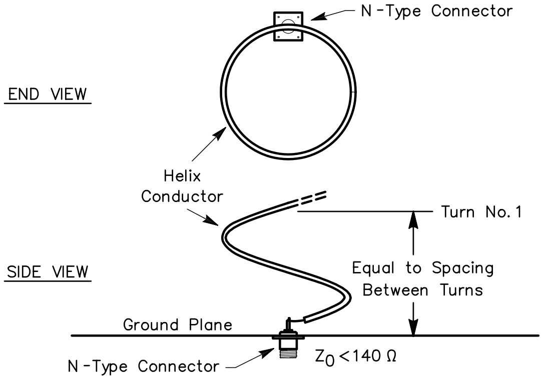

Fig 17—End view and side view of peripherally fed helix.

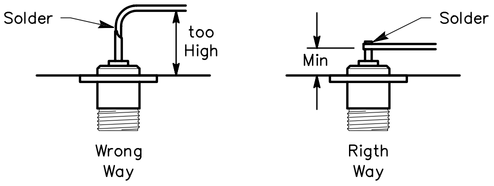

Fig 18—Wrong and right ways to attach helix to a type N connector for 50-Ω feed.

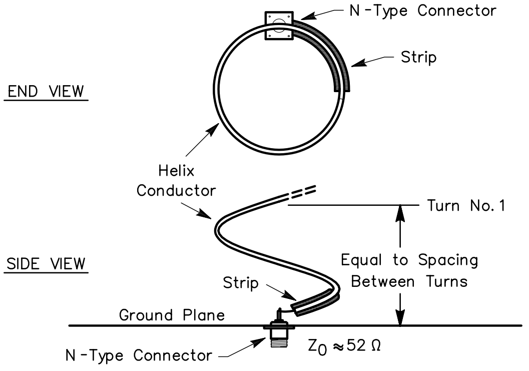

Fig 19—End view and side view of peripherally fed helix with metal strip added to improve transformer action.

impedance. The traditional λ/4 matching section has proved difficult to fabricate and maintain. But if the helix is fed at the periphery, the first quarter turn of the helix conductor (leaving the N connector) acts much like a transmission line—a single conductor over a perfectly conducting ground plane. The impedance of such a transmission line is:

$$Z_0 = 138 \log \frac{4h}{d} \qquad \text{(Eq 6)}$$

where:

Z_0 = line impedance in ohms
h = height of the center of the conductor above the ground plane
d = conductor diameter (in the same units as h).

The impedance of the helix is 140 Ω a turn or two away from the feed point. But as the helix conductor swoops down toward the feed connector (and the ground plane), h gets smaller, so the impedance decreases. The 140-Ω nominal impedance of the helix is transformed to a lower value. For any particular conductor diameter, an optimum height can be found that will produce a feed-point impedance equal to 50 Ω. The height should be kept very small, and the diameter should be large. Apply power to the helix and measure the SWR at the operating frequency. Adjust the height for an optimum match.

Typically, the conductor diameter may not be large enough to yield a 50-Ω match at practical (small) values of h. In this case, a strip of thin brass shim stock or flashing copper can be soldered to the first quarter turn of the helix conductor (**Fig 19**). This effectively increases the conductor diameter, which causes the impedance to decrease further yet. The edges of this strip can be slit every $^1/_2$ inch or so, and the strip bent up or down (toward or away from the ground plane) to tune the line for an optimum match.

This approach yields a perfect match to nearly any coax. The usually wide bandwidth of the helix (70% for less than 2:1 SWR) will be reduced slightly (to about 40%) for the same conditions. This reduction is not enough to be of any consequence for most amateur work. The improvements in performance, ease of assembly and adjustment are well worth the effort in making the helix more practical to build and tune.

ANTENNAS FOR AO-40 OPERATIONS

Antennas for successful operations on AO-40 come in many shapes and sizes. AO-40 has provided amateurs the opportunity to broadly experiment with antennas.

Fig 20 shows the satellite antennas at WD4FAB. The Yagi antennas are used for the U- and L-band AO-40 uplinks and the V-band AO-10 downlink, while the S-band dish antenna is for the AO-40 downlink. These satellite antennas are tower mounted at 63 feet (19 meters) to avoid pointing into the many nearby trees and suffering from the resulting "green attenuation." Of course, satellite antennas do not always need to be mounted high on a tower if dense foliage is not a problem. If satellite antennas are mounted lower down, feed-line length and losses can reduced.

Another benefit, however, to tower mounting of satellite antennas is that they can be used for terrestrial ham communications and contests. The fact that the antennas are set up for CP does not really degrade these other operating activities.

Experience with AO-40 has clearly shown the advan-

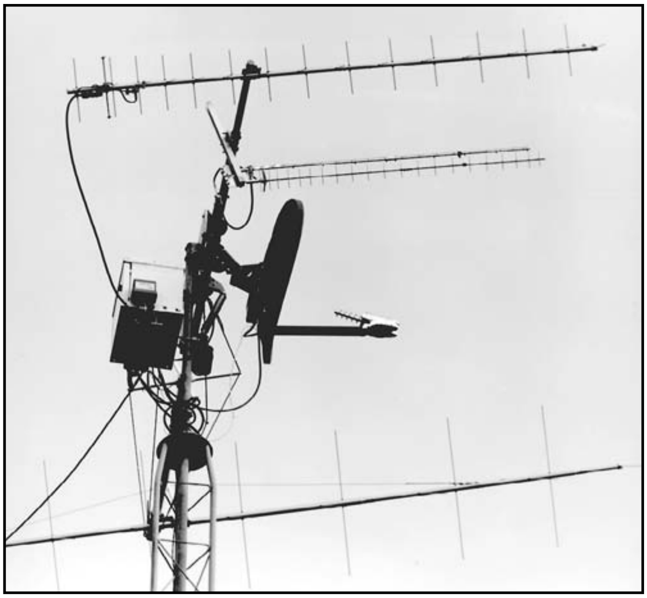

Fig 20—Details of WD4FAB's tower cluster of satellite antennas including a home-brew elevation rotator. Top to bottom: M^2 436-CP30, a CP U-band antenna; two M^2 23CM22EZA antennas in a CP array for L band; "FABStar" dish antenna with helix feed for S band; M^2 2M-CP22, a CP V-band antenna (only partially shown.) To left of dish antenna is a NEMA4 equipment box with an internal 40-W L-band amplifier, and also hosts externally mounted preamplifiers. (*WD4FAB photo*)

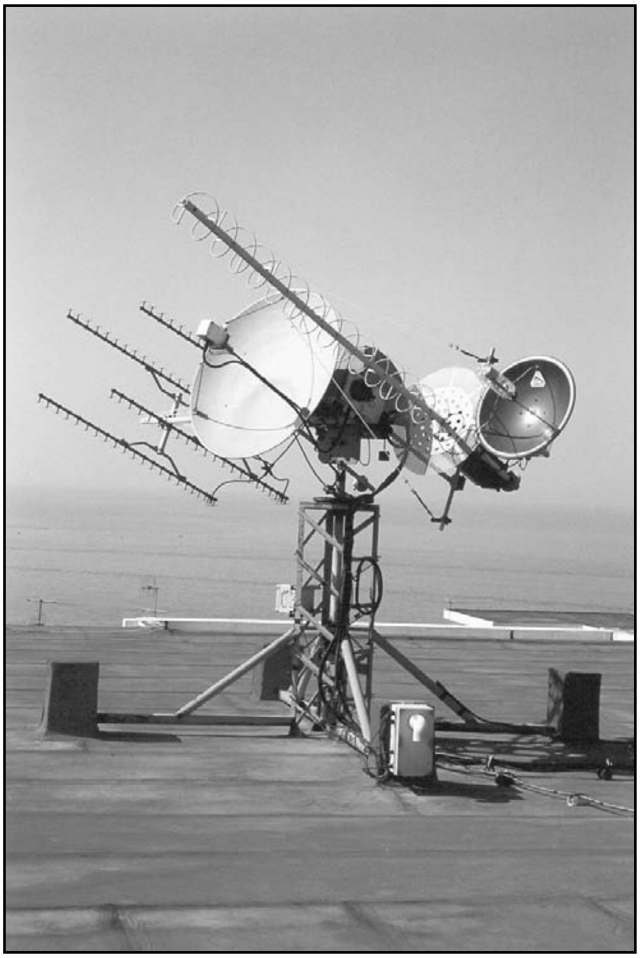

Fig 21—Domenico, I8CVS, has this cluster of satellite antennas for AO-40. Left to right: array of 4 × 23-element Yagi horizontally polarized for L band; 1.2-meter dish with 3-turn helix feed for S band; 15-turn RHCP helical antenna for U band; 60-cm dish for X band. All microwave preamplifiers and power amplifiers are homebrew and are mounted on this antenna cluster. (*I8CVS photo.*)

tages of using RHCP antennas for both the uplink and downlink communications. The antennas shown in Fig 20 are a single-boom RHCP Yagi antenna for U band, a pair of closely spaced Yagi antennas phased for RHCP for L band (see Fig 14), and a helix-fed dish antenna for S band. The antenna gain requirements for U band can easily be met with the gain of a 30-element crossed Yagi. Antennas of this size have boom lengths of 4 to 4$^1/_2$ wavelengths. The enterprising constructor can build a Yagi antenna from one of several references, however most of us prefer to purchase well-tested antennas from commercial sources as M^2 or Hy-Gain. In the past, KLM (now out of business) had offered a 40-element CP Yagi for U-band satellite service, and many of these are still in satisfactory use today.

U-band uplink requirements for AO-40 have clearly demonstrated the need for gain less than 16 to 17 dBic RHCP, with an RF power of less than 50 W PEP at the antenna (≈ 2,500 W-PEP EIRP with a RHCP antenna) depending upon the *squint angle*. (The squint angle is the angle at which the main axis of the satellite is pointed away from your antenna on the ground. If the squint angle is less than half of the half-power beamwidth, the ground station will be within the spacecraft antenna's nominal beam width.)

A gain of 16 to 17 dBic RHCP can be obtained from a 30-element crossed Yagi, AO-13 type antenna, and is good news, considering that the satellite may be over 60,000 km (37,000 miles) from your station. Success on the U-band uplinks to AO-40 is easier than those for L band at wider squint angles more than 20°. At squint angles less than 10°, U-band uplink operation can even be done with 1-5 W power outputs to a RHCP antenna (≈ 200 W-PEP EIRP with RHCP). These lower levels mean that smaller antennas can be used. In practice, these uplinks will produce downlink signals that are 10 to 15 dB above the noise floor, or S7 signals over an S3 noise floor. The beacon will give a downlink S9 signal for these same conditions.

WD4FAB's experience with the AO-40 L-band uplink has demonstrated that 40 W-PEP delivered to an antenna with a gain of ≈ 19dBic (3,000 W-PEP EIRP with RHCP) is needed for operations at the highest altitudes of AO-40 and with squint angles ≤ 15°. This is the pretty compact L-band antenna arrangement with two 22-element antennas in a RHCP array shown in Fig 14 and 20. Other operators have experience that using a 1.2-meter L-band dish antenna and 40 W of RF power (6,100 W-PEP EIRP with RHCP) can also provide a superb uplink for squint angles even up to 25°. A dish antenna can have a practical gain of about 21 to 22 dBic. These uplinks will provide the user a downlink that is 10 to 18 dB above the transponder noise floor. In more practical terms, this is an S7 to 8 signal over a S3 transponder noise floor, a very comfortable armchair copy.

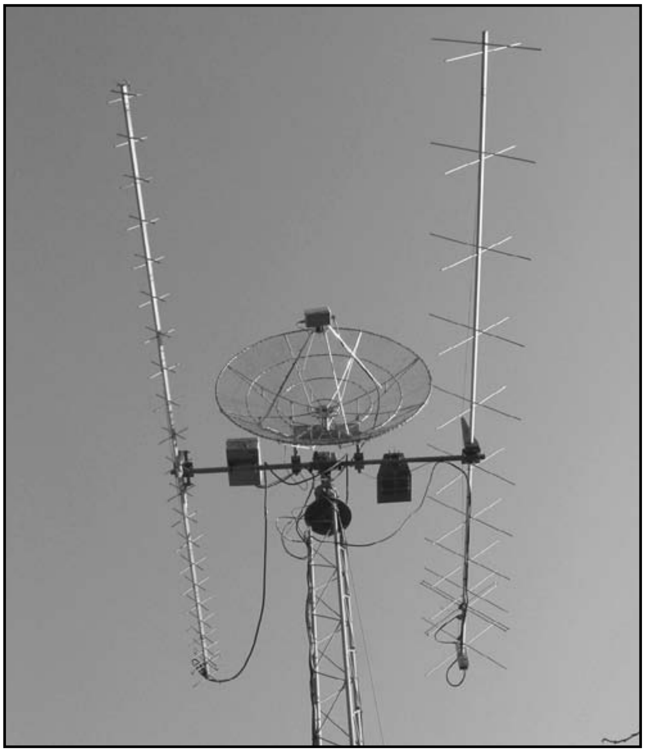

Fig 22—Wilfred Carey, ZS6JT, constructed this cluster of satellite and EME antennas. Left to right: 2 × 23-element offset feed Yagi for U band; 1.64-meter dish with 2$^1/_4$-turn helix feed for S band; 2 × 11-element coplanar feed Yagi for V band. (*ZS6JT photo.*)

Using the L-band uplink for AO-40, instead of the U-band uplink, allows the use of Yagi antennas that more manageable, since their size for a given gain is only one third of those for U-band. With L band there is a narrower difference between using a dish antenna and a Yagi, since a 21- to 22-dBic dish antenna would be only about 1.2 meters (4 feet) in diameter. However, some of us may not have such "real estate" available on our towers and may seek a lower wind-loading solution offered by Yagis. Long-boom rod-element Yagi, or loop-Yagi antennas are commercially offered by M^2 and DEM, although this band is about the highest for practical Yagis. The example shown in Fig 20 is a pair of rod-element Yagi antennas from M^2 in a CP arrangement with a gain of 18 to 19 dBic.

Other amateurs have successful AO-40 operation with different arrangements. **Fig 21** shows I8CVS's 4 × 23-element linear array for a 1270 MHz, a 1.2-meter solid dish for 2400 MHz, a 15-turn helical antenna for 435 MHz, and a 60-cm dish for 10,451 MHz. This arrangement clearly shows the advantage and accessibility of having a roof-mounted antenna.

Fig 22 shows ZS6JT's setup, with a 1.64-meter home-built mesh dish for 2400 MHz and two home-built crossed Yagi antennas, one for 435 MHz and the other for 145 MHz. Note that in these examples, the antennas permit terrestrial communication as well as satellite service. Two of these stations have also maintained the capability to operate the LEO satellites with U- and V-band antennas.

A number of amateurs have taken advantage of the availability of surplus C-band TVRO dishes, since most users of satellite television have moved up to the more

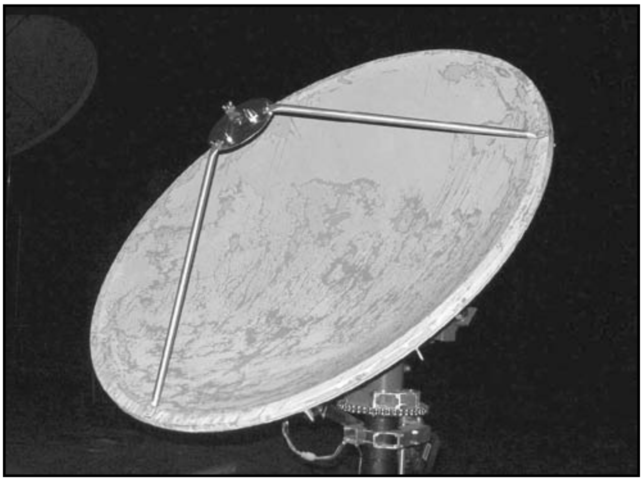

Fig 23—Robert Suding, W0LMD, modified this 4-foot dish antenna with a patch feed for S band and an Az-El mount. (*W0LMD photo.*)

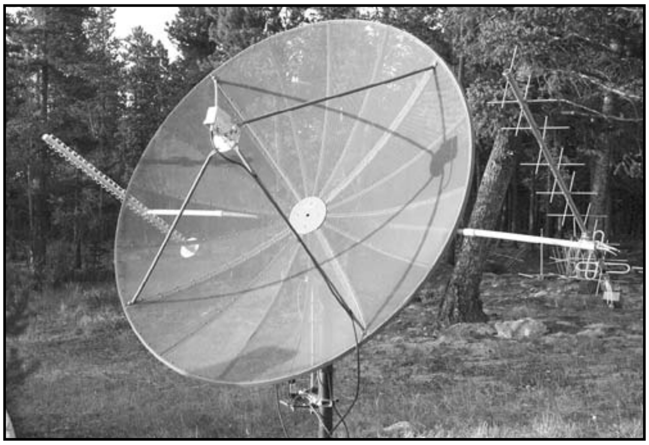

Fig 24—W0LMD graduated to this 8-foot dish with patch feed for S band for AO-40. On the left is a helical antenna for L band and on the right is a 2 × 9-element offset-feed Yagi for U band. A home-brew Az-El mount is provided. (*W0LMD photo.*)

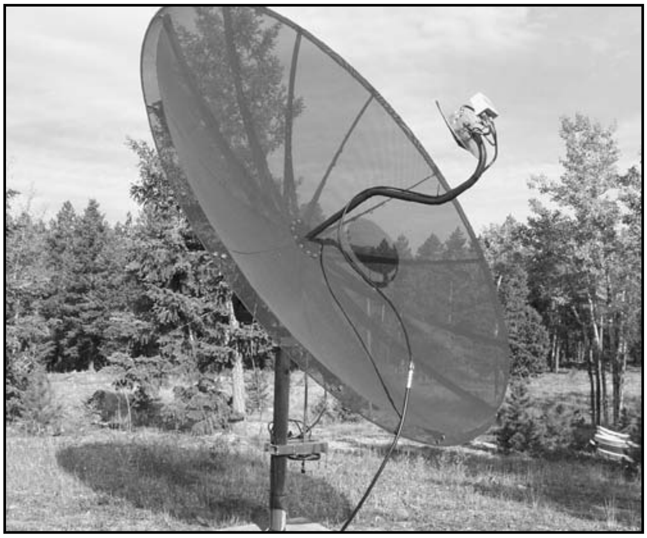

Fig 25—W0LMD increased to this 10-foot dish for AO-40 operations, with a triband patch feed for U, L, and S bands on an Az-El mount. (*W0LMD photo.*)

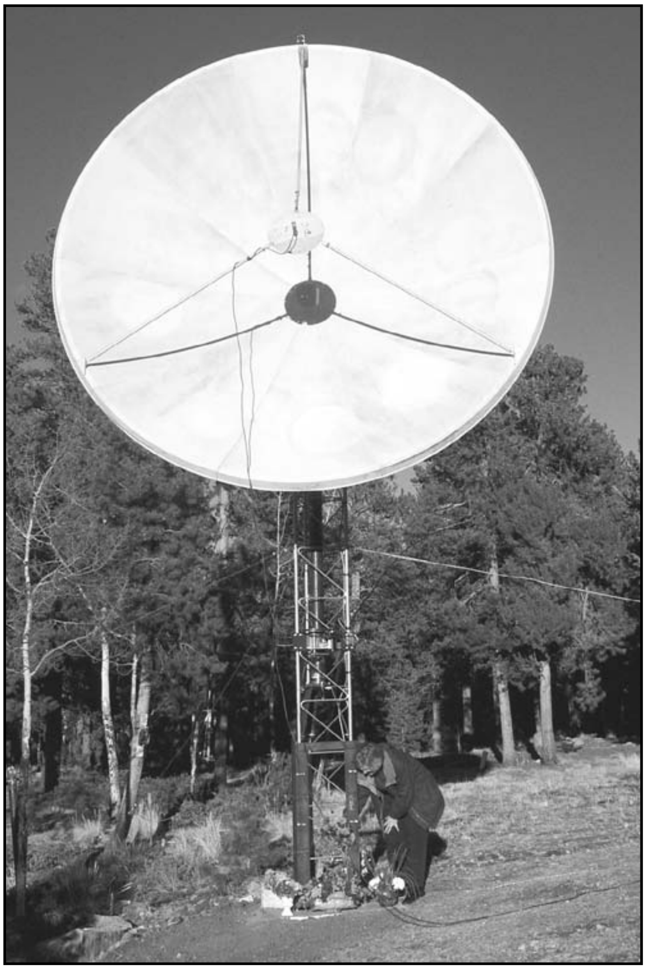

Fig 26—W0LMD found the ultimate in this 14-foot dish for AO-40, with a triband patch feed and Az-El mount. (*W0LMD photo.*)

convenient K band using 0.5-meter dishes. Some examples of these dish conversions for satellite communications are shown in **Fig 23**, a W0LMD 4-foot dish with patch feed and Az-El mount. **Fig 24** shows a W0LMD 8-foot dish with patch feed, Az-El mount, a U-band Yagi, and an L-band helical antenna.

Fig 25 is a W0LMD 10-foot dish with tri-band patch feed and Az-El mount; and **Fig 26** is also a W0LMD

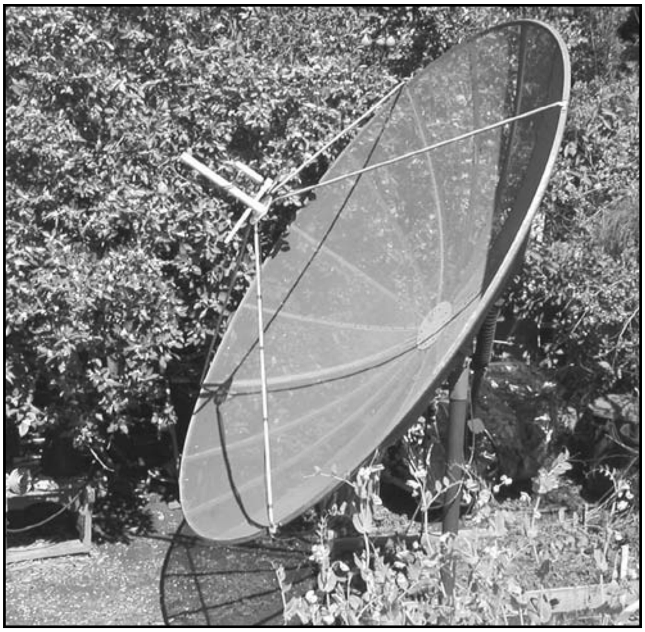

Fig 27—Clair E. Cessna, K6LG, has this 10-foot dish with S-band patch feed. This dish uses the original polar-mounting system and offsets the patch feed to compensate for AO-40's deviation from the Clarke belt. (*K6LG photo.*)

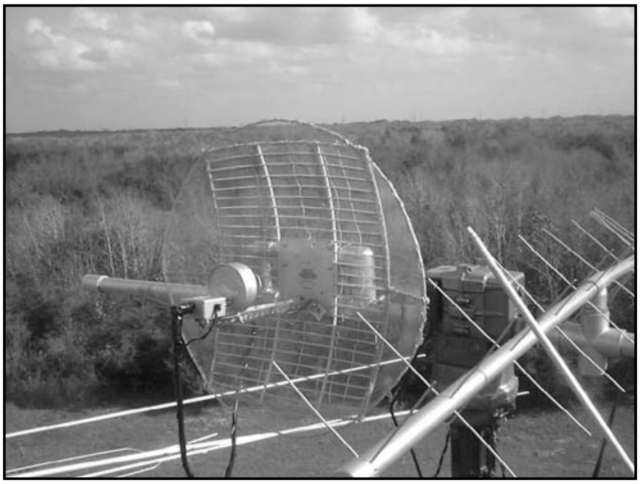

Fig 28—K5GNA's "circularized" mesh modification of an MMDS dish antenna with a helix-CP feed and DEP preamp. The dish modification reduces the spillover loss by making the antenna fully circular. (*K5OE photo.*)

14-foot dish with tri-band patch feed and Az-El mounting. Other operators, like K6LG, have been able to use TVRO dishes, **Fig 27**, with multiband patch feeds and still use, within limits, their polar-mounting system, as will be explained later.

Other hams have taken advantage of other surplus dish situations. **Fig 28** shows modified MMDS dishes, by K5GNA, and **Fig 29**, by K5OE, both using helix feeds. **Fig 30** shows a 75-cm high modified PrimeStar offset feed dish, by WD4FAB, using a longer helical feed antenna because of the higher f/D ratio of this dish configuration. This dish provides 5 dB of Sun noise, which is good performance. These efforts have rewarded their users with superb service on AO-40. Many have experimented with different feed and mounting systems. These experiments will be further illustrated.

One very popular spun-aluminum dish antenna seen in use on AO-40 has been the G3RUH-ON6UG 60-cm unit with its S-band patch feed, **Fig 31**. A kit, complete with a CP-patch feed is available from SSB-USA and has a gain of 21 dBic. It provides a 2.5-dB Sun noise signal. Surplus dishes have not been the only source for antennas for AO-40 operations, since some ingenious operators have even turned to the use of cardboard boxes. See **Figs 32** to **35**.

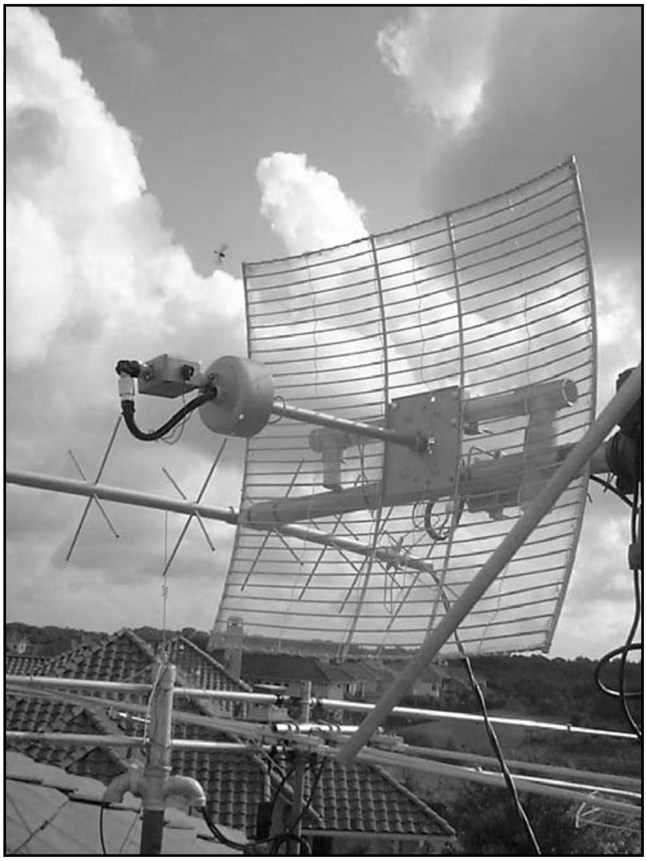

Fig 29—Mesh modification of an MMDS dish antenna by Jerry Brown, K5OE, with a helix-CP feed and DEM preamplifier mounted directly to the helix feed point. (*K5OE photo.*)

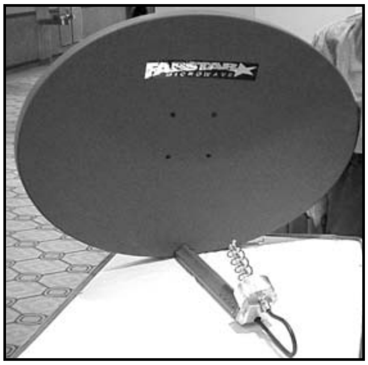

Fig 30—PrimeStar offset-fed dish with WD4FAB's helix-feed antenna. N0NSV was so pleased with the modification that he renamed the dish "FABStar," and made a new label! (*N0NSV photo.*)

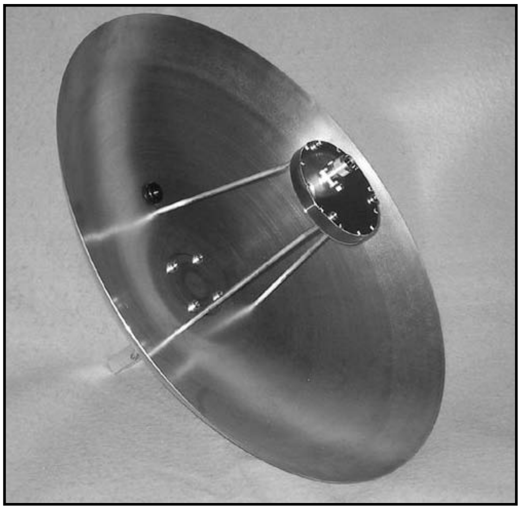

Fig 31—G3RUH's 60-cm spun-aluminum dish with CP-patch feed is available as a kit. This antenna has been popular with many AO-40 operators all over the world.

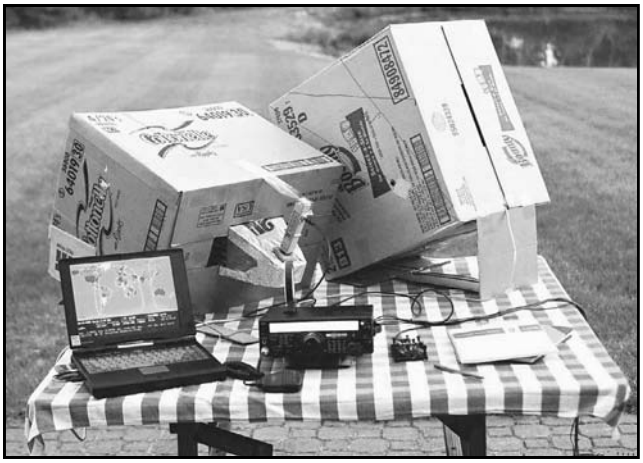

Fig 32—A complete satellite station with the tracking laptop, FT-847 transceiver, and both downlink and uplink cardboard-box antennas.

Parabolic Reflector Antennas

The satellite S-band downlinks have become very popular for a variety of reasons:

- Good performance with physically small downlink antennas
- Availability of good-quality downconverters
- Availability of preamps at reasonable prices.

A number of people advocate S-band operation, including Bill McCaa, K0RZ, who led the team that designed and built the AO-13 S-band transponder and James Miller, G3RUH, who operates one of the AO-40 command stations. Ed Krome, K9EK, and James Miller have published a number of articles detailing construction of preamps, downconverters and antennas for S band.

Some access AO-40's S-band downlink using compact S-band helical antennas. See **Fig 36**. With the demise of AO-40's S1 transmitter and its high-gain downlink antenna, enthusiasts have had to employ high-gain parabolic-dish antennas to use AO-40's S2 downlink, with its lower-gain helical antenna.

W0LMD notes that like a bulb in a flashlight, the

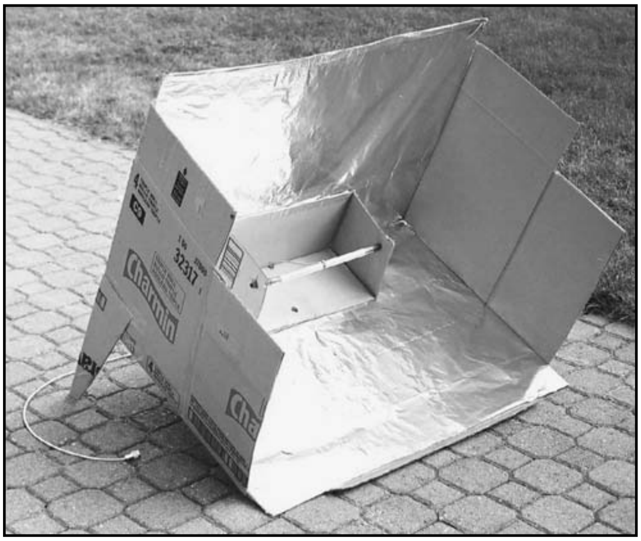

Fig 33—The completed high-performance corner-reflector uplink antenna for U band. Note how the box corners hold the reflectors and dipole feed in place. The rear legs set the antenna elevation to 20°—this gives good coverage at the design latitude but will need modification for other stations.

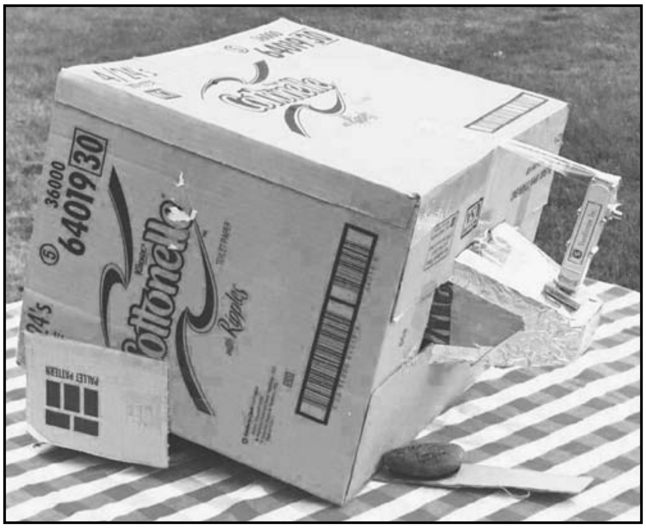

Fig 35—Side view of the downlink pyramidal horn showing the elevation control supports and how the downconverter is attached to the horn. The downconverter is pulled forward by the tape to align the probe wire parallel to the rear surface of the horn.

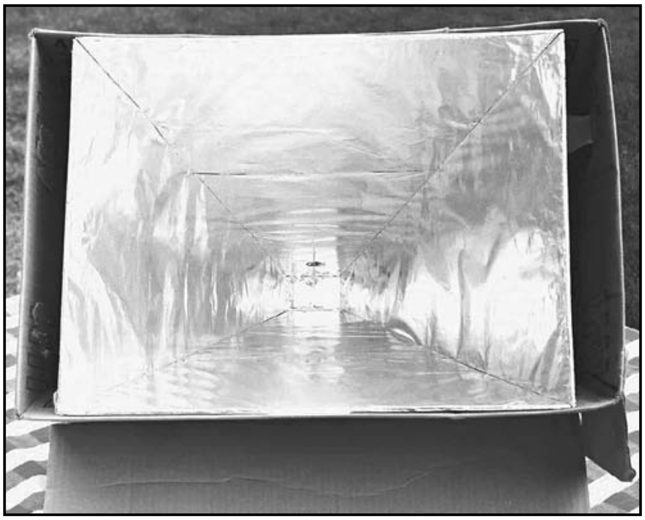

Fig 34—Front view of the downlink pyramidal horn showing how it is mounted in the support carton. Notice the coax probe at the back of the horn.

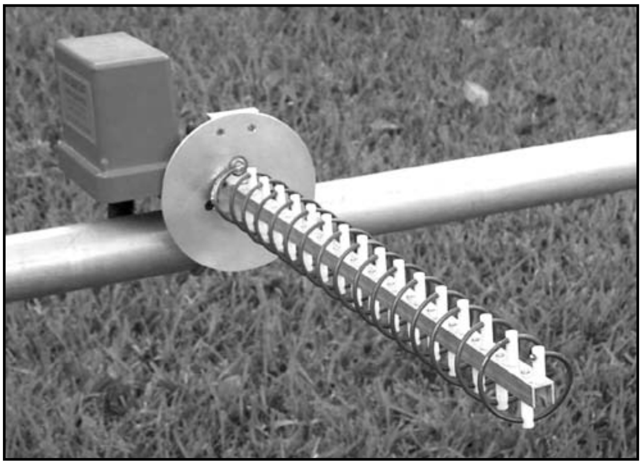

Fig 36—WD4FAB's example of a 16-turn S-band helical antenna for AO-40. This is about the maximum length of any practical helix. Note the SSB UEK2000 downconverter mounted behind the reflector of the antenna. (*WD4FAB photo.*)

parabolic reflector or dish antenna must have a feed source looking into the surface of the dish. Some dishes are designed so that the feed source is mounted directly in front of the dish. This is referred to as a *center-fed dish*. Other dishes are designed so that the feed source is off to one side, referred to as an *off-center-fed dish*, or just offset-fed dish, as shown in Fig 30. The offset-fed dish may be considered a side section of a center-fed dish. The center-fed dish experiences some signal degradation due to blockage of the feed system, but this is usually an insignificantly small amount. The offset-fed dish is initially more difficult to aim, since the direction of reception is not the center axis, as it is for center-fed dishes.

The basic design precepts of parabolic-dish antennas are covered in more detail in the EME Antenna section of this chapter. Dish antenna properties specific to satellite operations are covered here. The dish's parabola can be designed so the focus point is closer to the surface of the dish, referred to a *short-focal-length* dish, or further away from the dish's surface, referred to as a *long-focal-length* dish. To determine the exact focal length, measure the diameter of the dish and the depth of the dish.

$$f = \frac{D^2}{16d} \qquad \text{(Eq 7)}$$

The focal length divided by the diameter of the dish gives the *focal ratio*, commonly shown as f/D. Center-fed dishes usually have short-focal ratios in the range of f/D = 0.3 to 0.45. Offset-fed dishes usually have longer focal lengths, with f/D = 0.45 to 0.80. If you attach two small mirrors to the outer front surface of a dish and then point the dish at the Sun, you can easily find the focus point of the dish. Put the reflector of the patch or helix feed just beyond this point of focus.

An alternate method for finding a dish's focal length is suggested by W1GHZ (ex-N1BWT), who provides a computer program called *HDL_ANT*, available at: **www.w1ghz.org/10g/10g_home.htm**. The method literally measures a solid-surface dish by the dimensions of the bowl of water that it will form when properly positioned. (See: **www.qsl.net/n1bwt/chap5.pdf**.) WD4FAB used this method on the dish of Fig 30, carefully leveling the bowl, plugging bolt holes, and filling it with water to measure the data needed by the W1GHZ Web-site calculation.

While many of us enjoy building our own antennas, surplus-market availability of these small dish antennas makes their construction unproductive. Many AO-40 operators have followed the practices of AO-13 operators using a surplus MMDS linear-screen parabolic reflector antenna, Figs 28 and 29. These grid-dish antennas are often called *barbeque dishes*. K5OE and K5GNA have shown how to greatly improve these linearly polarized reflectors by adapting them for the CP service desired for AO-40. Simple methods can be used to *circularize* a linear dish and to further add to its gain using simple methods to increase the dish area and feed efficiency.

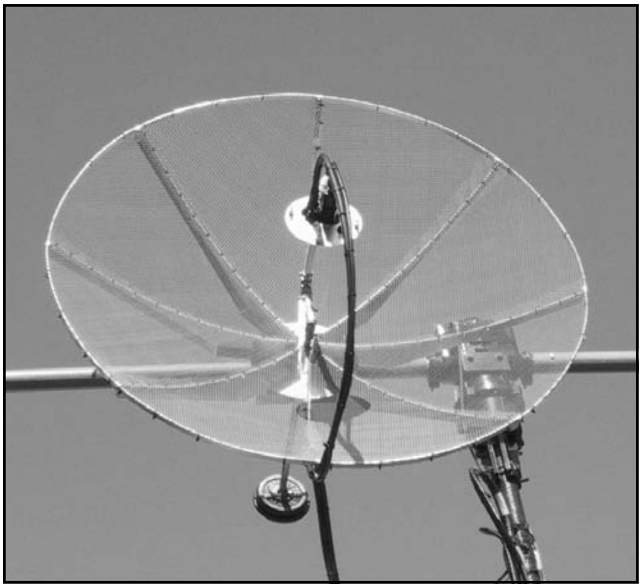

Fig 37—Prototype 1.2-meter dish by Rick Fletcher, KG6IAL, using a dual-band (L and S) patch-feed antenna for AO-40. See text. This kit dish is covered with $^1/_4$-inch mesh. (*KG6IAL photo.*)

Another approach is the construction of a kit-type dish antenna, just becoming available in 1.2-meter and 1.8-meter diameters. This ingenious design by KG6IAL is available from his Web site **www.teksharp.com/**. **Fig 37** shows the prototype of the 1.2-meter dish with an f/D of 0.30. The 1.2-meter dish is fed with a dual-band patch feed for L and S bands. The 1.8-meter dish is designed for up to three bands using a tri-band patch feed for the U, L and S bands. This dish will permit U-band operation. A Central States VHF Society measurement on a similar sized dish (by WØLMD) with a patch feed showed a gain of about 17.1 dBic (actual measurement was 12.0 dBd linearly fed). This performance along with a small V-band (145 MHz) Yagi would permit a very modest satellite antenna assembly for all of the VHF/UHF LEO satellites, as well as AO-40.

The ingenuity of the design of the KG6IAL antenna is that it is constructed of robust, $^1/_8$-inch-thick aluminum sheet that is numerically machined for the parabolic shape of the ribs. The backsides of the ribs are stiffened by a bent flange edge. The panel mesh is attached by using small tie-wraps or small aluminum wire through the mesh and holes provided along the parabolic edge. KG6IAL used $^1/_4$-inch mesh in the prototype antenna to reduce wind loading. A single formed conduit post is provided in the kit for mounting the patch-feed assembly. The post extends rearward to permit the attachment of a counterweight, if needed.

AO-40 has also provided some additional challenges to the ham operator. Besides its well-known S-band downlink, AO-40 also has a K-band downlink in the range of

24.05 GHz. This quite low-powered transmitter has provided a substantial challenge to some operators, such as N1JEZ, K5OE, W5LUA, G3WDG and others. N1JEZ documented his work in *QST* while K5OE shows his K-band work on his Web site and in the *Proceedings of the AMSAT Space Symposium*. See **Fig 38**.

Parabolic Dish Antenna Construction

In the USA large numbers of dishes can be obtained either free or at low cost. But in some parts of the world dishes are not so plentiful, so hams make their own. **Fig 39** shows G3RUH's S-band dish antenna. There are three parts to the dish antenna—the parabolic reflector, the boom and the feed. There are as many ways to construct this as there are constructors. You need not slavishly replicate every nuance of the design. The only critical dimensions occur in the feed system. After construction, you will have a 60-cm diameter S-band RHCP dish antenna with a gain of about 20 dBi and a 3-dB beamwidth of 18°. Coupled with the proper downconverter, performance will be more than adequate for S-band downlink.

The parabolic reflector used for the original antenna was intended to be a lampshade. Several of these aluminum reflectors were located in department-store surplus. The dish is 585 mm in diameter and 110 mm deep, corresponding to an f/D ratio of 585/110/16 = 0.33 and a focal length of 0.33 × 585 = 194 mm. The f/D of 0.33 is a bit too concave for a simple feed to give optimal performance but the price was right, and the under-illumination keeps ground noise pickup to a minimum. The reflector already had a 40-mm hole in the center with three 4-mm holes around it in a 25-mm radius circle.

The boom passes through the center of the reflector and is made from 12.7-mm square aluminum tube. The boom must be long enough to mount to the rotator boom on the backside of the dish. The part of the boom extending through to the front of the dish must be long enough

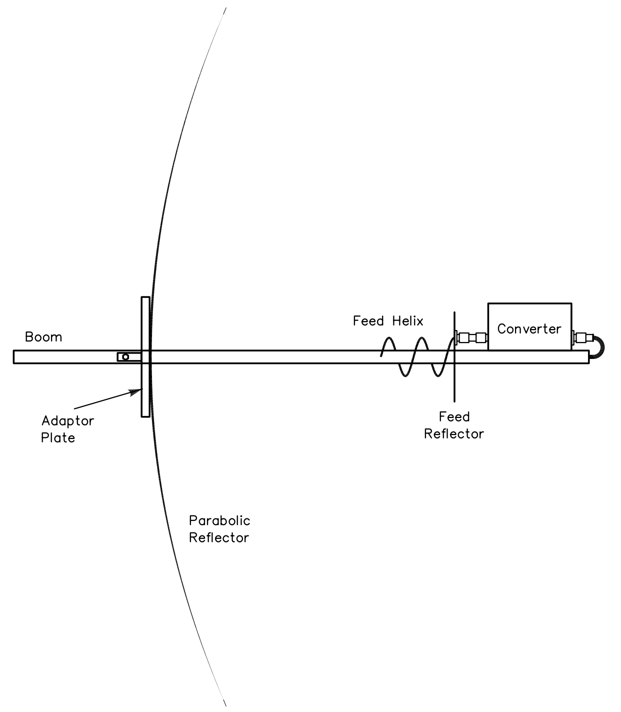

Fig 39—Detail of 60-cm S-band dish antenna with feed.

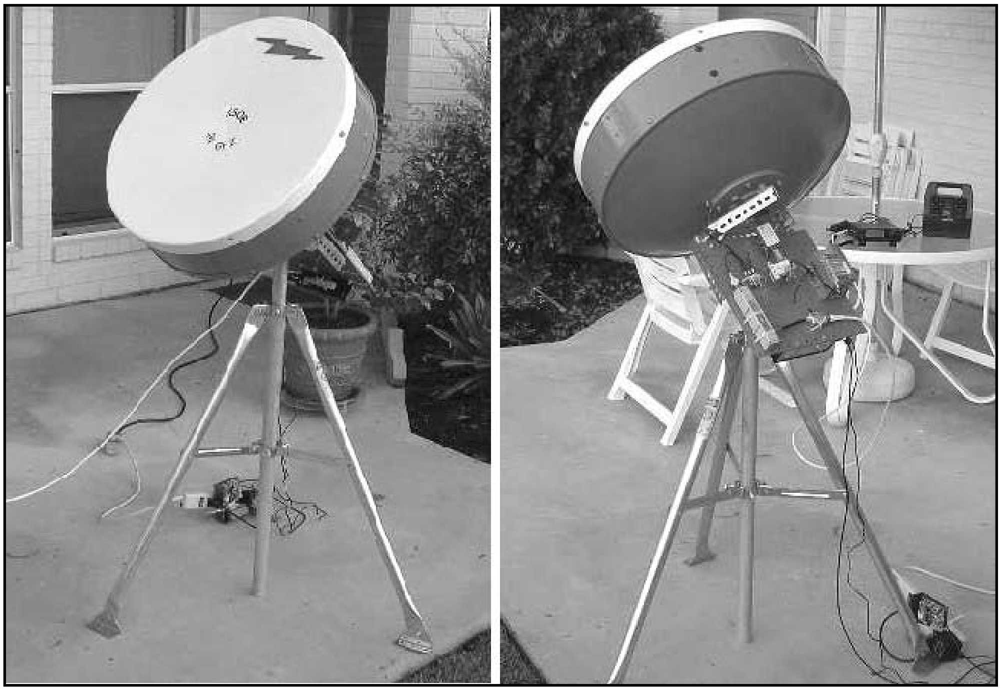

Fig 38—K5OE found this K-band dish on the Web and has set it up for the AO-40 K-band downlink. (*K5OE photo.*)

to mount the feed at the focus. If you choose to mount the downconverter or a preamp near the feed, some additional length will be necessary. Carefully check the requirements for your particular setup.

A 3-mm thick piece of aluminum, 65 mm in diameter, supports the boom at the center of the reflector. Once the center mounting plate is installed, the center boom is attached using four small angle brackets—two on each side of the reflector. See Fig 39 for details of reflector and boom assembly.

A small helix is used for the S-band antenna feed. The reflector for the helix is made from a 125-mm square piece of 1.6-mm thick aluminum. The center of the reflector has a 13-mm hole to accommodate the square center boom described above. The type-N connector is mounted to the reflector about 21.25 mm from the middle. This distance from the middle is, of course, the radius of a helical antenna for S-band. Mount the N connector with spacers so that the back of the connector is flush with the reflector surface. The helix feed assembly is shown in **Fig 40**.

Copper wire, or tubing, about 3.2 mm in diameter is used to form the helix. Wind four turns around a 40-mm diameter form. The turns are wound counterclockwise. This is because the polarization sense is reversed from RHCP when reflected from the dish surface. The wire helix will spring out slightly when winding is complete.

Once the helix is wound, carefully stretch it so that the turns are spaced 28 mm (±1 mm). Make sure the finished spacing of the turns is nice and even. Cut off the first half turn. Carefully bend the first quarter turn about 10° so it will be parallel to the reflector surface once the helix is attached to the N connector. This quarter turn will form part of the matching section.

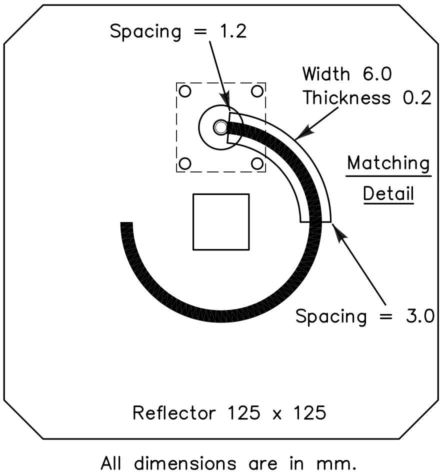

Fig 40—Details of helix feed for S-band dish antennas. The type-N connector is fixed with three screws and is mounted on a 1.6-mm spacer to bring the PTFE molding flush with the reflector. An easier mounting can be using a smaller TNC connector. Reflectors should be 95 to 100 mm in diameter.

Cut a strip of brass, 0.2 mm thick and 6 mm wide, and match the curvature of the first quarter turn of the helix, using a paper pattern. Be careful to get this pattern and subsequent brass cutting done exactly right. Using a large soldering iron and working on a heatproof surface, solder the brass strip to the first $^1/_4$ turn of the helix. Unless you are experienced at this type of soldering, getting the strip attached just right will require some practice. If it doesn't turn out right, just dismantle, wipe clean and try again.

After tack soldering the end of the helix to the type-N connector, the first $^1/_4$ turn, with its brass strip in place, should be 1.2 mm above the reflector at its start (at the N connector) and 3.0 mm at its end. Be sure to line up the helix so its axis is perpendicular to the reflector. Cut off any extra turns to make the finished helix have $2^1/_4$ turns total. Once you are satisfied, apply a generous amount of solder at the point the helix attaches to the N connector. Remember this is all that supports the helix.

Once the feed assembly is completed, pass the boom through the middle hole and complete the mounting by any suitable method. The middle of the helix should be at the geometric focus of the dish. In the figures shown here, the feed is connected directly to the downconverter and then the downconverter is attached to the boom. You may require a slightly different configuration depending on whether you are attaching a downconverter, preamp or just a cable with connector. Angle brackets may be used to secure the feed to the boom in a manner similar to the boom-to-reflector mounting. Be sure to use some method of waterproofing if needed for your preamp and/or downconverter.

Dish Feeds

W0LMD describes in **www.ultimatecharger.com/** that feeding a dish has two major factors that determine the efficiency. Like a flashlight bulb, the feed source should evenly illuminate the entire dish, and none of the feed energy should spillover outside the dish's reflecting surface. No feed system is perfect in illuminating a dish. Losses affect the gain from either under-illuminating or over-illuminating the dish (spillover losses). Typical dish efficiency is 50%. That's 3 dB of lost gain. A great feed system for one dish can be a real lemon on another. A patch feed system is very wide angle, but a helix feed system is narrow angle.

W0LMD has experimented with helical feeds for low f/D antennas ("deep" dishes) shown in **Fig 41**. A short-focal-ratio center-fed dish requires a wide-angle feed system to fully illuminate the dish, making the CP patch the preferred feed system. When used with an offset-fed dish, a patch-type feed system will result in a considerable spillover, or over-illumination loss, with an increased sensitivity to off-axis QRM, due to the higher f/D of this dish. Offset-fed

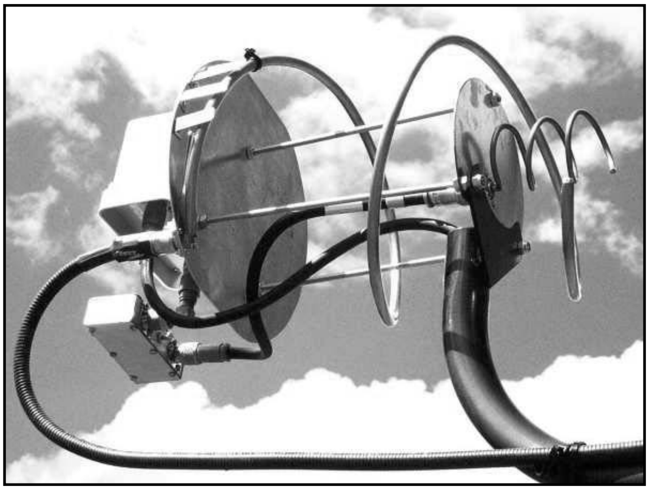

Fig 41—WØLMD's dual helix-dish feed for U and S bands. This early experimental feed was found to be wanting and he then turned to patch feeds for dishes. (*WØLMD photo.*)

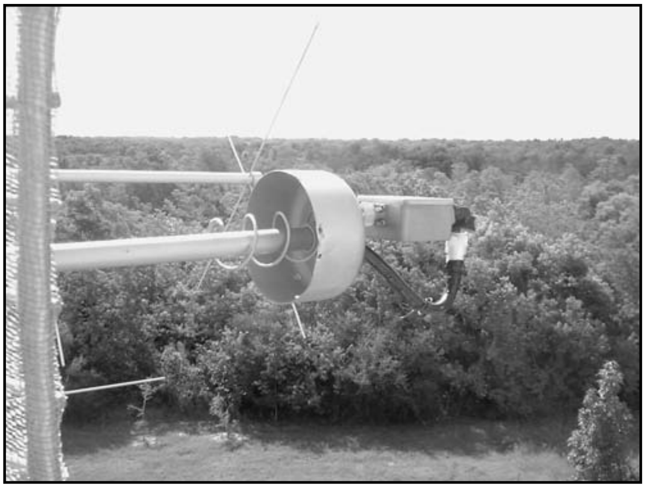

Fig 42—K5OE's helix feed for his MMDS S-band dish antenna. (*K5OE photo.*)

dishes do much better when fed with a helix antenna.

A helix feed is simplicity personified. Mount a type N connector on a flat reflector plate and solder a couple of turns wire to the inner terminal. Designs are anywhere from 2 to 6 turns. The two-turn helices are used for very short-focal-length dishes in the f/D = 0.3 region, and the 6-turn helices are used with longer-focal-length (f/D ~ 0.6) dishes, typically offset-fed dishes. Since AO-40 is right circular and the dish reflection will reverse the polarity, the helix should be wound left circular, looking forward from the connector. Helix feeds work poorly on the short-focal-length dishes but really perform well on the longer-focal-length offset-fed dishes. K5OE shows us the helix feed for his modified MMDS dish in **Fig 42**. This design employs the cupped reflector of W8JK.

A Helix Feed for an Offset-Dish Antenna

This section describes WD4FAB's surplus PrimeStar offset-fed dish antenna with a 7-turn helical feed antenna, shown in Fig 30. This S-band antenna can receive Sun noise 5 dB above sky noise. (Don't try to receive Sun noise with the antenna looking near the horizon, since terrestrial noise will be greater than 5 dB, at least in a big-city environment.) WD4FAB received the dish from NØNSV, who renamed the finished product the "FABStar."

The dish's reflector is a bit out of the ordinary, with the shape of a horizontal ellipse. It is still a single paraboloid, illuminated with an unusual feed horn. At 2401 MHz (S band) we can choose to under-illuminate the sides of the dish while properly feeding the central section, or over-illuminate the center while properly feeding the sides. WD4FAB chose to under-illuminate. The W1GHZ water-bowl measurements showed this to be a dish with a focal point of 500.6 mm and requiring a feed for an f/D = 0.79. The total illumination angle of the feed is 69.8° in the vertical direction and a feed horn with a 3-dB beamwidth of 40.3°. At 50% efficiency this antenna was calculated to provide a gain of 21.9 dBi. A 7-turn helical feed antenna was estimated to provide the needed characteristics for this dish and is shown in **Fig 43**.

The helix is basically constructed as described for the G3RUH parabolic dish above. A matching section for the first λ/4 turn of the helix is spaced from the reflector at 2 mm at the start and 8 mm at the end of that fractional turn. Modifications of the G3RUH design include the addition of a cup reflector, a design feature used by the originator of the helical antenna, John Kraus, W8JK. For the reflector, a 2-mm thick circular plate is cut for a 94 mm (0.75 λ) diameter with a thin aluminum sheet metal cup, formed with a depth of 47 mm. Employment of the cup enhances the performance of the reflector for a dish feed, as shown by K5OE. (See the K5OE material on the CD-ROM accompanying this book.)

The important information for this 7-turn helical antenna is:

- Boom: 12.7-mm square tube or "C" channel.
- Element: $^1/_8$-inch diameter copper wire or tubing.

Close wind the element on a circular 1.50-inch tube or rod; the finished winding is 40 mm in diameter and spaced to a helical angle of 12.3°, or 28 mm spacing. These dimensions work out for an element circumference of 1.0 λ about the center of the wire.

When WD4FAB tackled this antenna, he felt that the small number of helical element supports used by G3RUH would be inadequate, in view of the real-life bird traffic on the antennas at his QTH. He chose to use PTFE (Teflon) support posts every $^1/_2$ turn. This closer spacing of posts permitted a careful control of the helix-winding diameter and spacing and also made the antenna very

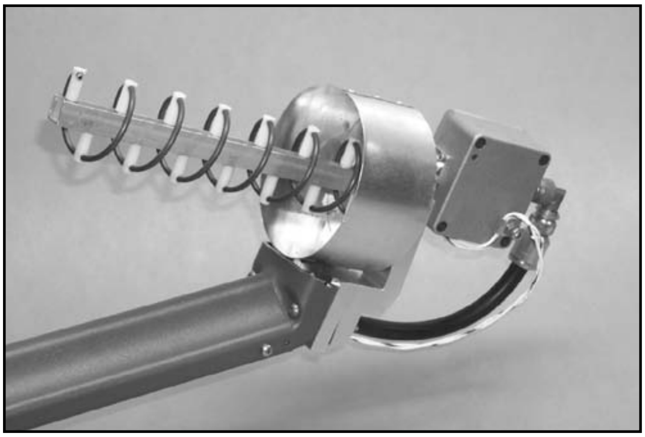

Fig 43—Seven-turn LHCP helix feed for an offset dish, long f/D, antenna, with DEM preamp. (*WD4FAB photo.*)

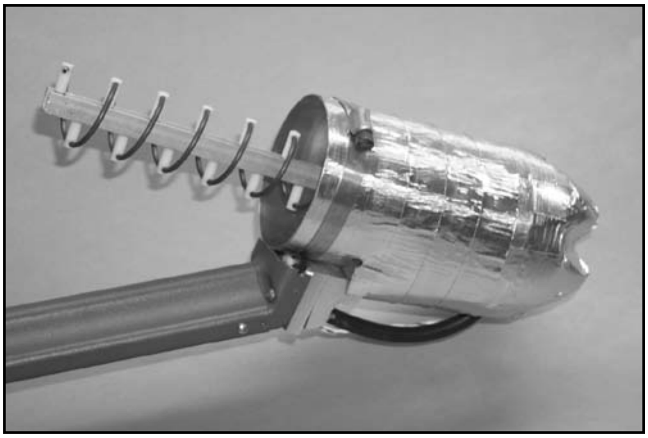

Fig 45—Rain cover for preamp using a two-liter soft-drink bottle with aluminum foil tape for protection from sun damage. (*WD4FAB photo.*)

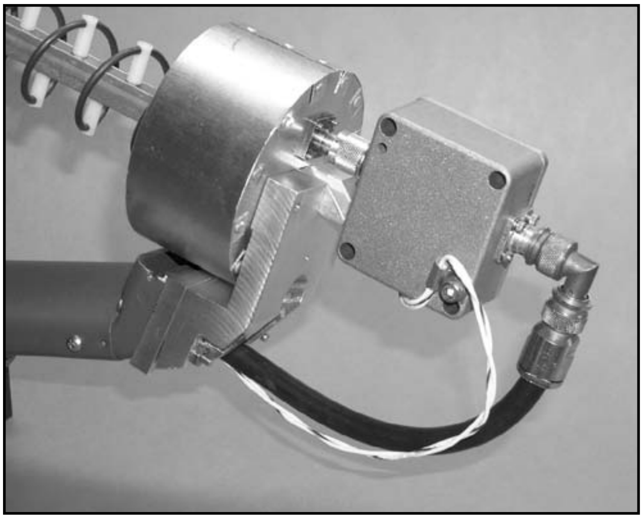

Fig 44—Mounting details of seven-turn helix and preamp. (*WD4FAB photo.*)

robust. He set up a fixture on the drill press to uniformly predrill the holes for the element spacers and boom. Attachment of the reflector is through three very small aluminum angle brackets on the element side of the boom.

Mounting of the helix to the dish requires modification of the dish's receiver-mounting boom. **Fig 44** shows these modifications using a machined mount. NM2A constructed one of these antennas and showed that a machine shop is not needed for this construction. He made a "Z" shaped mount from aluminum-angle plate and then used a spacer from a block of acrylic sheet. The key here is to get the dish focal point at the 1.5-turn point of the feed antenna, which is also at about the lip of the reflector cup.

The W1GHZ data for this focal point is 500.6 mm from the bottom edge of the dish and 744.4 mm from the top edge. A two-string measurement of this point can confirm the focal point, as shown by Wade in his writings. When mounting this feed antenna the constructor must be cautious to aim the feed at the beam-center of the dish, and not the geometric center, as the original microwave horn antenna was constructed. Taking the illumination angle information noted above, the helical feed antenna should be aimed 5.5° down from the geometric center of the dish.

As illustrated in Fig 44, a DEM preamp was directly mounted to the feed helix, using a TNC female connector on the helix, chosen for this case, since N connectors are quite large for this antenna. A male chassis connector should be mounted on the preamp so that the preamp can be directly connected to the antenna without any adaptors. This photo also illustrates how the reflector cup walls were riveted to the reflector plate.

Exposed connectors must be protected from rainwater. Commonly materials such as messy Vinyl Mastic Pads (3M 2200) or Hand Moldable Plastic (Coax Seal) are used. Since this is a tight location for such mastic applications, a rain cover was made instead from a 2-liter soft-drink bottle, **Fig 45**. Properly cutting off the top of the bottle allows it to be slid over the helix reflector cup and secured with a large hose clamp. You must provide UV protection for the plastic bottle and that was done with a wrapping of aluminum foil pressure-sensitive adhesive tape.

There are many methods for mounting this dish antenna to your elevation boom. You must give consideration to the placement of the dish to reduce the wind loading and off-balance to the rotator system. In WD4FAB's FABStar installation, the off-balance issue was not a major factor, as the dish was placed near the center of the elevation boom, between the pillow-block bearing supports. Since there is already a sizeable aluminum plate for these bearings, the dish was located to "cover" part of that plate, so as to not add measurably to the existing wind-loading area of the overall assembly.

A mounting bracket provided with the stock dish clamps to the end of a standard 2-inch pipe stanchion (actual measure: 2.38 inches in diameter). This bracket was turned around on the dish and clamped to the leg of a welded-pipe Tee assembly. See **Fig 46**. Pipe-reducing fittings were machined and fitted in the Tee-top bar, which was sawn in half for clamping over the $1^1/_2$ inch pipe used for the elevation boom. Bolts were installed through drilled holes and used to clamp this assembly.

Patch Feeds for Dish Antennas

Patch feeds are almost as simple as helix feeds. A patch is typically an N connector on a flat reflector plate with a tuned flat-metal plate soldered to the inner terminal. Sometimes the flat plate is square; sometimes it is rectangular; sometimes it is round. It could have two feed points, 90° out of phase for circular polarization, as used in the construction of the AO-40 U-band antennas. Some patches are rectangular with clipped corners to create a circular radiation pattern.

On 2401 MHz, the plate is 57 mm square and spaced 3 mm away from the reflector. The point of attachment is about halfway between the center and the edge. A round patch for 2401 MHz is about 66 mm in diameter. These patches work well on the shorter focal length center-fed MMDS and TVRO dishes. G3RUH made a CP patch feed for these short f/D dishes, shown in Fig 31 and **Fig 47**.

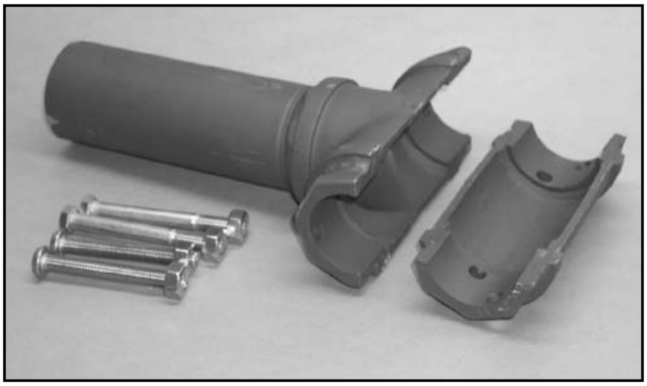

Fig 46—Welded pipefitting mount bracket for FABStar dish antenna. (*WD4FAB photo.*)

Robert, WØLMD, has done a considerable amount of experimenting with patch feeds for his dish antennas. One tri-band feed is shown in **Fig 48**. These are circular patches that have CP properties through the arrangement of the feed point and a small piston-variable capacitor that is offset from the feed point. **Fig 49** shows some of the many patches that Robert has created for his trials.

A No-Tune Dual-Band Feed for Mode L/S

Jerry, K5OE, notes that the AO-40 transponder has two uplink receivers active most of the time for CW/SSB activity. Most operators use U band at 435 MHz (70 cm). Also available, however, are two L-band (23-cm) receivers: L1 at 1269 MHz and L2 at 1268 MHz. The reasons for going to L band can be varied, but there is no arguing the benefits in reduced antenna size and AGC suppression. The types of L-band antennas are varied as well. Many use helices. Others use beams and arrays of beams. Still others use dishes, small and large.

K5OE recently acquired an old UHF TV dish measuring 1.2 meters in diameter. He wanted to use it both to receive on S band at 2401 MHz (13cm) and to transmit on the uplink on L band. He covered it with aluminum mesh and built a dual-helix feed for it, but was unhappy with the L-band performance. It seems the concentric helices interacted with each other substantially. Having had good success with patch feeds on S band, he designed, built and installed a dual-patch feed on a 1.5-meter solid dish for Field Day 2002. This arrangement worked superbly on uplink (with 25 W), but was embarrassingly deaf on receive. This second dual-band feed failure led him to experiment for months with different configurations, leading ultimately to the design presented here. The project goals were:

- Good performance on both S-band receive and L-band uplink.
- An easy-to-produce model using common hardware and simple hand tools.

Patches are better than helices as dish feeds. This revelation came to K5OE while doing investigation and

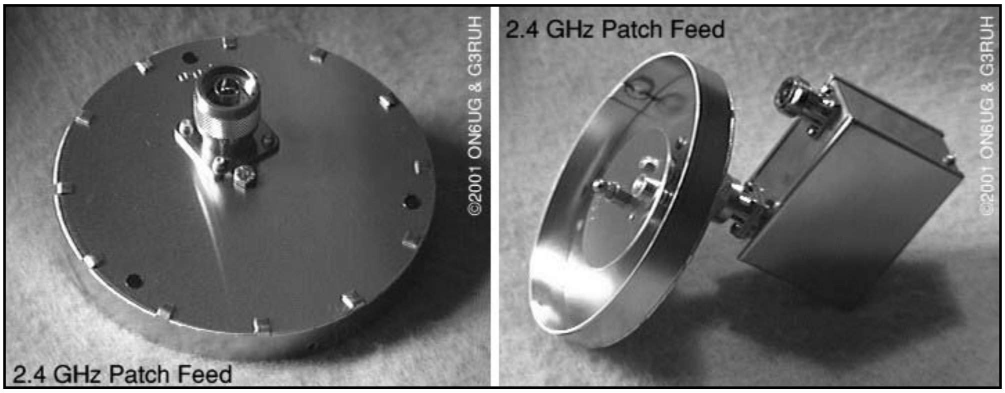

Fig 47—Details of CP-patch feed for short f/D dish antennas by G3RUH and ON6UG.

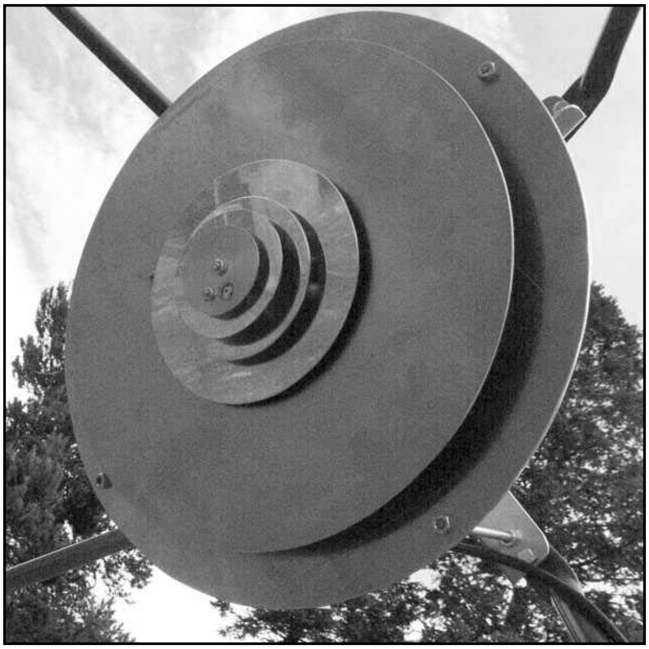

Fig 48—A triband (U, L and S bands) patch-CP feed for large dish antennas for AO-40 service. (*WØLMD photo.*)

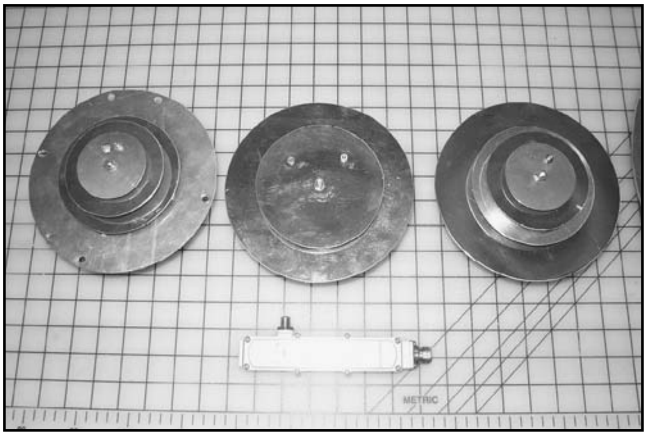

Fig 49—Some of the many experimental CP-patch-feed antennas by WØLMD. (*WØLMD photo.*)

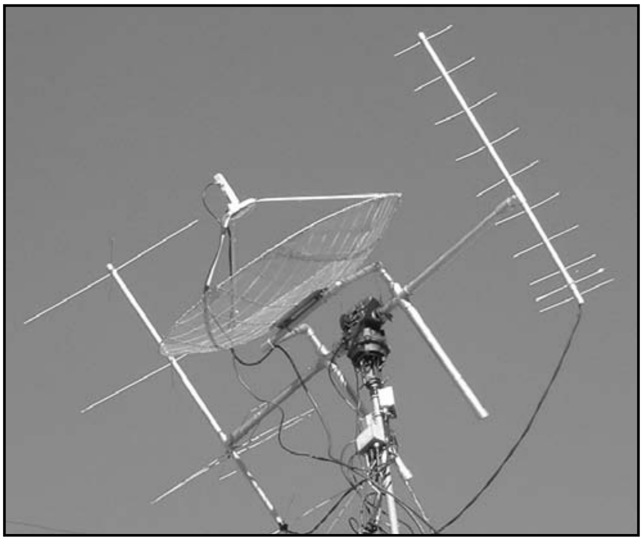

Fig 50—The 1.2-meter dish with dual-band patch feed installed. (*K5OE photo; courtesy of* The AMSAT Journal.)

experimenting with helix antennas. In the middle of this investigative foray, he saw the radiation pattern for the G3RUH patch feed published on James Miller's web site. When he modeled that pattern and input it into the W1GHZ feed pattern program, it produced an amazing 72% efficiency. The best helix he ever modeled has about 60% efficiency. I8CVS recently ran his own antenna range tests of a design similar to the G3RUH patch and produced a similarly impressive pattern.

Then K5OE came across the *truncated corners* square patch design popularized by K3TZ. This AO-40 design here is attributed to 7N1JVW, JF6BCC and JG1IIK. There are references in the literature going back over a decade for this now-common commercial design. The first model K5OE built outperformed his best helix-in-cup design by a full S unit (delta over the noise) on his FT-100 portable setup. Compared to a helix, the patch simply has better illumination efficiency with less spillover from side lobes.

Patch theory is beyond the scope of this article, but can be summarized as building a shape that resonates at the desired frequency, compensated in size by the capacitive inductance between itself and the reflector. A patch can be practically any shape since it basically acts like a parallel-plate transmission line. Current in the patch flows from the feed point to the outer edge(s), where all the radiation occurs. The reputed, but often disputed, circularity of the truncated corner patch is accomplished by effectively designing two antennas into the patch element (of two different diagonal lengths) and feeding them 90º out of phase.

For K5OE's 1.2-meter dish, shown in **Fig 50**, computations predicted 21-dBi gain on L band and almost 27 dBi on S band, with an assumed 50% efficiency:

$$G = 10\log_{10}\left[\eta A\left(\frac{4\pi}{\lambda^2}\right)\right] \qquad \text{(Eq 8)}$$

where

η = efficiency

λ = wavelength in meters.

A = aperture of the dish in meters = $\pi \times r^2$

r = dish radius in meters = diameter/2 in meters = 0.6 meters

At 1269 MHz, λ= 300/1269 = 0.236 meters:

$$G = 10\log_{10}\left[0.50 \times \left(3.14 \times 0.6^2\right) \times \frac{4 \times 3.14}{0.236^2}\right] = 21.1\,\text{dBi}$$

At 2401.5 MHz, λ = 300/2401.5 = 0.125 meters:

$$G = 10\log_{10}\left[0.50 \times \left(3.14 \times 0.6^2\right) \times \frac{4 \times 3.14}{0.125^2}\right] = 26.6\,\text{dB}^2$$

Where does the feed get mounted? The *focal point* is where the parabolic shape of the dish concentrates the reflected signal. In K5OE's case the antenna was placed flat on the garage floor to measure the depth:

$f = D^2 / 16d$ (Eq 9)

where

D = diameter of the dish in inches
d = depth of the dish in inches
$f = 48^2 / (16 \times 7.25) = 19.8$ inches (50.5 cm)

This is just one example of countless combinations of hardware and patch designs. Inherent in this design, however, are five key design and construction features developed from building and empirical testing of a number of patch feeds.

1. The specified dimensions are critical for no-tune operation. **Fig 51** shows the dimensions necessary to build the dual-feed patch. (K5OE recommends you reproduce this sketch accurately on graph paper. When you cut your patches you can lay them on the paper template for checking.) Repeat: These dimensions are critical. Even a 0.5-mm error will throw your resonance off considerably—patches are not broadband.
2. The reflector must be rigid. Spacing between the driven element (patch) and the reflector affects the resonant frequency. K5OE found 0.025-inch aluminum sheet and 26-gauge copper sheet acceptable for a single S-band patch feed, but too flimsy for an L-band reflector. Use more rigid material or provide additional stiffening for the L-band reflector, as shown in Figs **52**, **53** and **54**.
3. The patches must be electrically isolated from each other. A metallic center support works for a single patch but creates harmonic-coupling problems when patches are stacked for multiband use. The use of nylon machine screws and nuts helps solve the vexing problem of the S-band patch coupling to the L-band patch.
4. The "straight corners" of the truncated corner patch must be kept clear of any nearby metal. This includes the edges of the feed support or cup, if used. See Fig 54.
5. Feeding the patches at 90° to each other minimizes the electromagnetic interaction between the two antenna fields.

One final design issue deals with the first harmonic of the L-band antenna. You must significantly reduce the potentially destructive effect from the 1269-MHz signal's second harmonic. Severe desense of your receive signal could occur and potentially even overload and damage the first active device in your system. Sensitive preamps and downconverters without a pre-RF-amplifier filter will need an external filter. K5OE has used a G3WDG stub filter rated at 100-dB rejection with good success ahead of his preamp. His current setup, however, uses a AIDC-3731AA downconverter with its internal combline filter providing adequate filtering. Using the downconverter directly at the

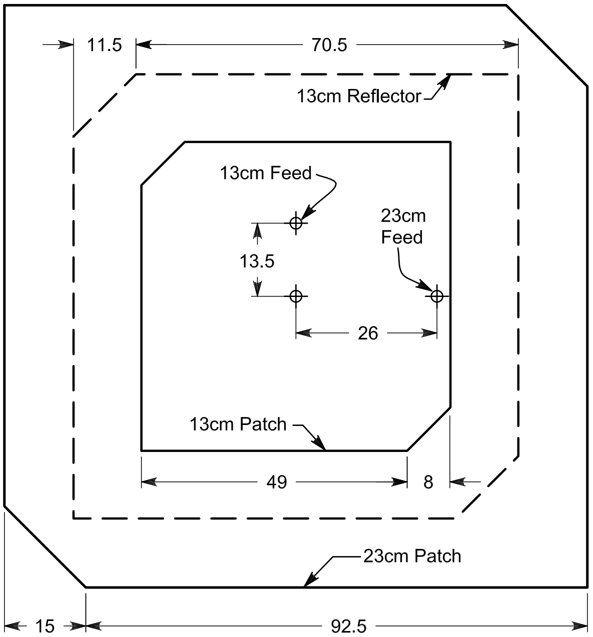

Fig 51—Dual-band patch feed dimensions, in millimeters. (*K5OE diagram; courtesy of* The AMSAT Journal.)

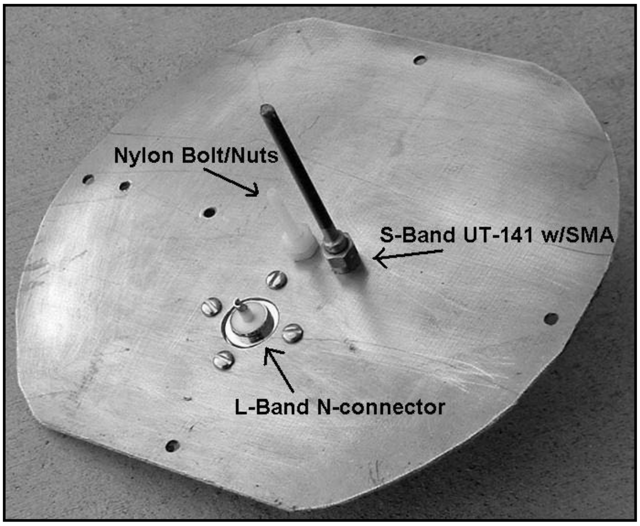

Fig 52—Assembly of the L band reflector. (*K5OE photo; courtesy of* The AMSAT Journal.)

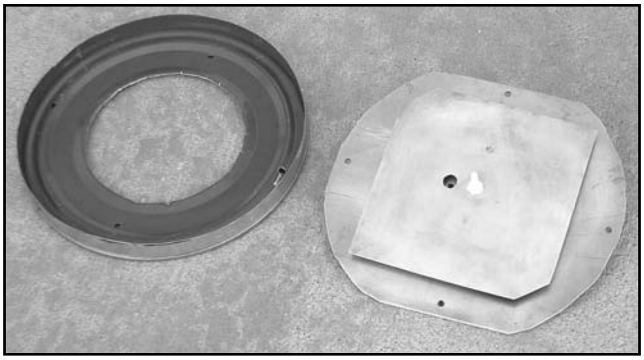

Fig 53—The support, L-band reflector and patch. (*K5OE photo; courtesy of* The AMSAT Journal.)

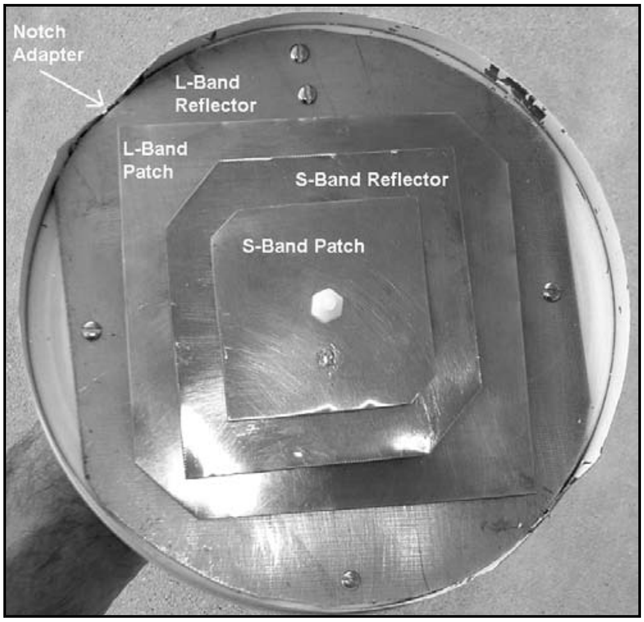

Fig 54—The completed dual-band patch feed. (*K5OE photo; courtesy of* The AMSAT Journal.)

feed point has a noise figure (NF) of 1.0 dB, compared to the cumulative NF of 1.6 dB using a filter and a preamp.

Construction of the feed begins with selection of material for both the electrical parts (the antennas) and the mechanical parts (the support structure). The L-band antenna is constructed using a 6 × 6-inch double-sided circuit board for the reflector and a piece of 26-gauge copper sheet for the driven element (patch). A flanged female type-N connector is used for the feed connection. The S-band antenna is constructed of two pieces of 26-gauge copper sheeting and the feed connection is made with a short piece of UT-141 (0.141-inch copper-clad semirigid coax) terminated in a male SMA fitting. Fig 52 illustrates the assembly of the L-band reflector with the nylon-center support bolt, the L-band N-connector, and the S-band semirigid coax terminated onto an SMA-to-N adapter through the circuit board.

The support structure began life as a paint can, measuring 155 mm in diameter. It was cut down to a 15-mm depth. Cut a hole in the middle of the bottom of the can and trim the PC board to fit inside the can bottom. Use stainless-steel $^{3}/_{8}$-inch 4-40 bolts, washers and nuts to secure the PC board to the can bottom. A 1$^{1}/_{2}$-inch 6-32 nylon bolt is secured through the center of the PC board with two nylon nuts to provide the 6-mm spacing for the L-band patch. Fig 53 shows the L-band patch in position and ready to be soldered to the N-connector. Note the hole through the L-band patch allowing the S-band UT-141 coax to pass (without making contact).

The remainder of the antenna is then assembled in order: First the L-band patch is secured with two nylon nuts and soldered to the N-connector. Then the S-band reflector is secured with one nylon nut to provide 3-mm spacing, and the UT-141 coax shield is soldered to the S-band reflector. Finally, the S-band patch is secured with a single nylon nut (3-mm spacing) and soldered to the center conductor of the UT-141 coax. To summarize the overall order of assembly: L-band reflector, two nylon nuts, L-band patch, two nylon nuts, S-band reflector, one nylon nut, S-band patch, and one nylon nut.

An electrical check with an ohmmeter of the completed feed should show the two reflectors connected, with the patches isolated from the reflectors and from each other. Fig 54 shows the completed feed. Note how the sides of the support are cut out to avoid proximity to the L-band patch and how the L-band and S-band patches are at 90° to each other. **Fig 55** shows the back of the feed, complete with an angle support for the downconverter. The flanged N-connector is for the L-band coax and the male-N adapter is secured from the other side of the feed with the SMA fitting on the UT-141 coax.

For those who are tempted to tune the patch, K5OE recommends doing it with the feed installed on the antenna—since the dish surface affects the feed-point impedance slightly. The feed-point impedance, and thus the resonant frequency, can be changed quite a bit by adjustment of the spacing of just the straight corners. There is no need to change the spacing at the center or the feed—just a slight up or down bending of the straight corners will change the tuning. Do this carefully: a little bit goes a long way. This patch design is very repeatable and will work adequately (an SWR below 1.5:1) with no adjustments.

The antenna performs to the calculated predictions above. On receive, this antenna is 4 S units better than K5OE's 45-cm dish and 3 S units above his 65-cm dish (both other dishes have similar patch feeds and the same downconverter). It also clearly outperforms his previous dual-helix arrangement on the 1.2-meter dish, but he was unable to do a side-by-side comparison.

On transmit, it does equally well, with a decent signal into the satellite with only 10 W measured at the

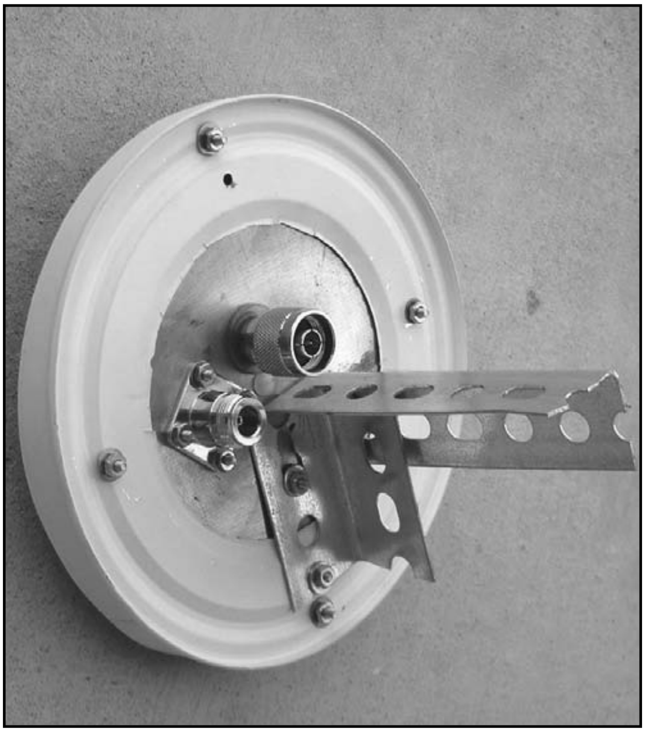

Fig 55—Rear of the completed feed. (K5OE photo; courtesy of *The AMSAT Journal.*)

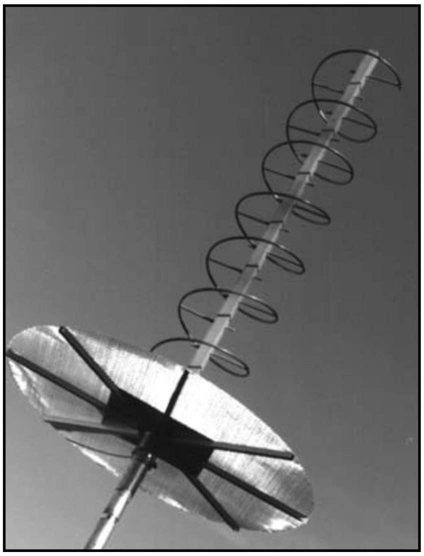

Fig 56—The portable 435-MHz helix assembled and ready for operation. (*WØCY photo.*)

antenna. The L band is noticeably improved over the helix predecessor. At low squint angles K5OE finds the L-band uplink to be about 1 S unit weaker than his U-band uplink. He later added a small plastic hat to extend over the top of the patches to keep the rain and bird droppings off—both detune the patches when built up between the patch and the reflector.

Though simple and effective, this is merely one way to construct a dual feed. Cookie-tin lids also make excellent supports. Tin snips are a good investment and much easier to use than a hacksaw. Use a flat file to remove burrs from the edges of the patches. Use stainless-steel hardware, most notably $^{3}/_{8}$-inch 4-40 machine bolts and nuts for the antenna hardware and $^{1}/_{2}$-inch 6-32 for the support-structure connections to the support arms ($^{1}/_{2}$-inch aluminum tubing). The copper sheet is much easier to solder to than aluminum. Once completed, the feed received a few coats of white enamel paint to protect the copper and to minimize the visual reflections.

This is not the only dual-band antenna on AO-40. There are many varied, innovative designs available, including G6LVB's simple and effective 1.2-meter homebrew stressed *chicken wire* dish with a dual-G3RUH helix feed. G3WDG has a 3-meter dish with L/S-band helices and a K-band (1.3-cm) feed horn, and WØLMD has developed some popular dual- and tri-band "round" patch feeds. (See the Notes and References, as well as the CD-ROM bundled with this book.)

For additional information on constructing antennas, feeds and equipment techniques for use at microwave frequencies, see *The ARRL UHF/Microwave Experimenter's Manual* and *The ARRL UHF/Microwave Projects Manual*. Both of these books have a wealth of information for the experimenter.

PORTABLE HELIX FOR 435 MHZ

Helical antennas for 435 MHz are excellent uplinks for U-band satellite communications. The true circular polarization afforded by the helix minimizes signal *spin fading* that is so predominant in these applications. The antenna shown in **Fig 56** fills the need for an effective portable uplink antenna for OSCAR operation. Speedy assembly and disassembly and light weight are among the benefits of this array. This antenna was designed by Jim McKim, WØCY.

As mentioned previously, the helix is about the most tolerant of any antenna in terms of dimensions. The dimensions given here should be followed as closely as possible, however. Most of the materials specified are available in any well supplied do-it-yourself hardware or building supply store. The materials required to construct the portable helix are listed in **Table 1**.

The portable helix consists of eight turns of $^{1}/_{4}$-inch soft-copper tubing spaced around a 1-inch fiberglass tube or maple dowel rod 4 feet, 7 inches long. Surplus aluminum jacket Hardline can be used instead of the copper tubing if necessary. The turns of the helix are supported by 5-inch lengths of $^{1}/_{4}$-inch maple dowel mounted through the 1-inch rod in the center of the antenna. **Fig 57A** shows the overall dimensions of the antenna. Each of these support dowels has a V-shaped notch in the end to locate the tubing, as shown in Fig 57B.

The rod in the center of the antenna terminates at the feed-point end in a 4-foot piece of 1-inch ID galva-

Table 1
Parts List for the Portable 435-MHz Helix

Qty	*Item*
1	Type N female chassis mount connector
18 feet	1/4-in. soft copper tubing
4 feet	1-inch ID galvanized steel pipe
1	5 feet × 1-inch fiberglass tube or maple dowel
14	5-inch pieces of 1/4-inch maple dowel (6 feet total)
1	1/8-inch aluminum plate, 10 inches diameter
3	2 × 3/4-inch steel angle brackets
1	30 × 30-inch (round or square) aluminum screen or hardware cloth
8 feet	1/2 × 1/2 × 1/2-inch aluminum channel stock or old TV antenna element stock
3	Small scraps of Teflon or polystyrene rod (spacers for first half turn of helix)
1	1/8 × 5 × 5-inch aluminum plate (boom-to-mast plate)
4	1½-inch U bolts (boom-to-mast mounting)
3 feet	#22 bare copper wire (helix turns to maple spacers)

Assorted hardware for mounting connector, aluminum plate and screen, etc.

nized steel pipe. The pipe serves as a counterweight for the heavier end of the antenna. The 1-inch rod material inside the helix must be nonconductive. Near the point where the nonconductive rod and the steel pipe are joined, a piece of aluminum screen or hardware cloth is used as a reflector screen.

If you have trouble locating the 1/4-inch soft copper tubing, try a refrigeration supply house. The perforated aluminum screening can be cut easily with tin snips. This material is usually supplied in 30 × 30-inch sheets, making this size convenient for a reflector screen. Galvanized 1/4-inch hardware cloth or copper screen could also be used for the screen, but aluminum is easier to work with and is lighter.

A 1/8-inch-thick aluminum sheet is used as the support plate for the helix and the reflector screen. Surplus rack panels provide a good source of this material. **Fig 58** shows the layout of this plate.

Fig 59 shows how aluminum channel stock is used to support the reflector screen. (Aluminum tubing also works well for this. Discarded TV antennas provide plenty of this material if the channel stock is not available.) The screen is mounted on the bottom of the 10-inch aluminum center plate. The center plate, reflector screen and channel stock are connected together with plated hardware or pop rivets. This support structure is very sturdy. Fiberglass tubing is the best choice for the center rod material although maple dowel can be used.

Mount the type-N connector on the bottom of the center plate with appropriate hardware. The center pin should be exposed enough to allow a flattened end of the copper tubing to be soldered to it. Tin the end of the tubing after it is flattened so that no moisture can enter it. If the helix is to be removable from the ground-plane screen, do not solder the copper tubing to the connector. Instead, prepare a small block of brass, drilled and tapped at one side for a 6-32 screw. Drill another hole in the brass block to accept the center pin of the type-N connector, and solder this connection. Now the connection to the copper tubing helix can be made in the field with a 6-32 screw instead of with a soldering iron.

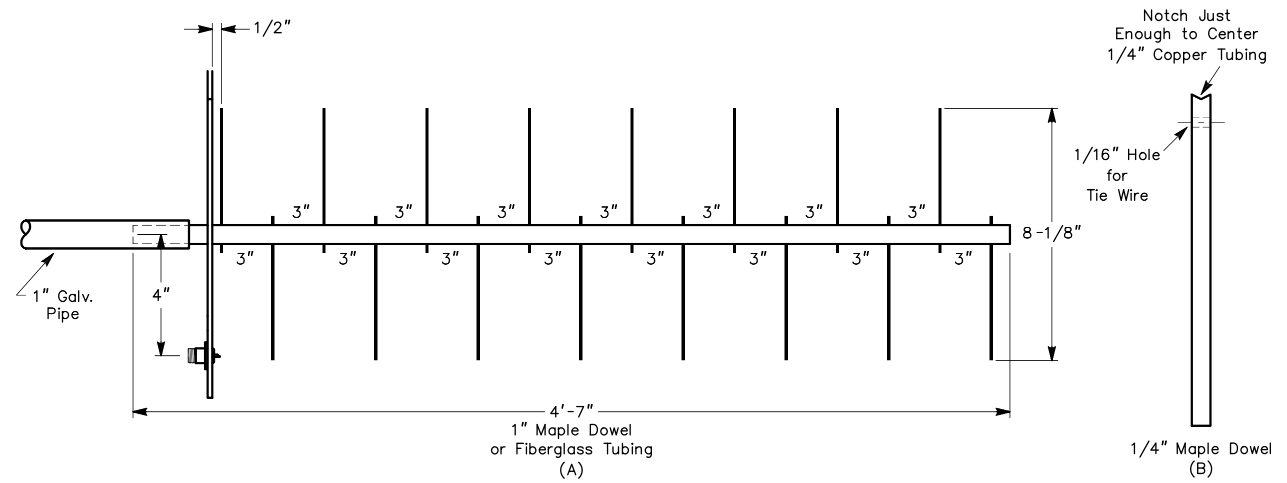

Fig 57—At A, the layout of the portable 435-MHz helix is shown. Spacing between the first 5-inch winding-support dowel and the ground plane is 1/2 inch; all other dowels are spaced 3 inches apart. At B, the detail of notching the winding-support dowels to accept the tubing is shown. As indicated, drill a 1/16-inch hole below the notch for a piece of small wire to hold the tubing in place.

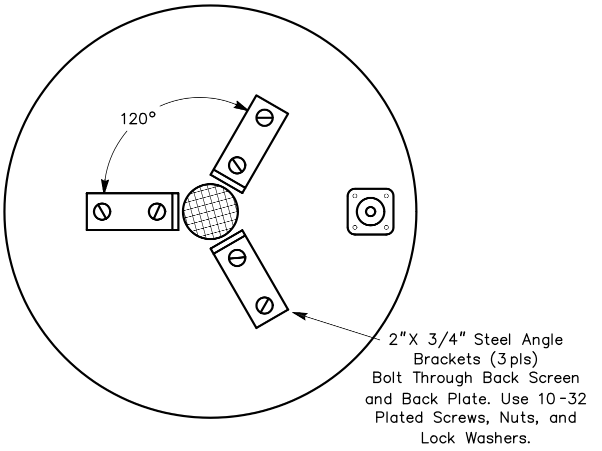
120°
2″X 3/4″ Steel Angle Brackets (3 pls) Bolt Through Back Screen and Back Plate. Use 10-32 Plated Screws, Nuts, and Lock Washers.

Fig 58—The ground plane and feed-point support assembly are shown. The circular piece is a 10-inch diameter, $^{1}/_{8}$-inch thick piece of aluminum sheet. (A square plate may be used instead.) Three 2 × $^{3}/_{4}$-inch angle brackets are bolted through this plate to the backside of the reflector screen to support the screen on the pipe. The type-N female chassis connector is mounted in the plate 4 inches from the 1-inch diameter center hole.

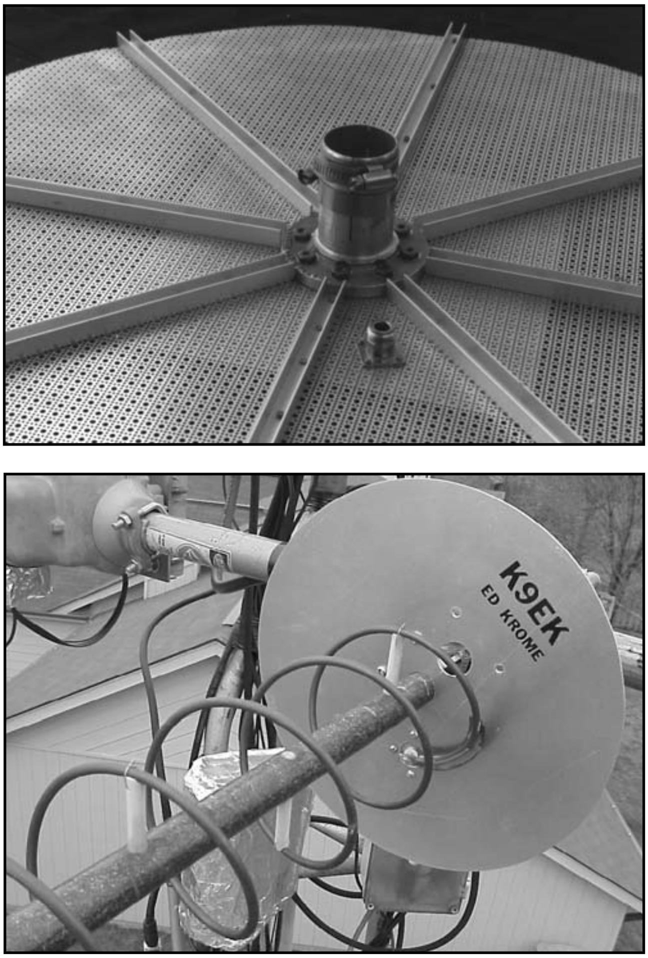

Fig 59—At top, the method of reinforcing the reflector screen with aluminum channel stock is shown. In this version of the antenna, the three angle brackets of Fig 58 have been replaced with a surplus aluminum flange assembly. (*W0CY photo.*) At bottom, this helix view shows the details of a $^{1}/_{4}$-turn matching transformer, as discussed in the text. (*K9EK photo.*)

Refer to Fig 57A. Drill the fiberglass or maple rod at the positions indicated to accept the 5-inch lengths of $^{1}/_{2}$-inch dowel. (If maple doweling is used, the wood must be weatherproofed as described below before drilling.) Drill a $^{1}/_{16}$-inch hole near the notch of each 5-inch dowel to accept a piece of #22 bare copper wire. (The wire is used to keep the copper tubing in place in the notch.) Sand the ends of the 5-inch dowels so the glue will adhere properly, and epoxy them into the main support rod.

Begin winding the tubing in a clockwise direction from the reflector screen end. First drill a hole in the flattened end of the tubing to fit over the center pin of the type-N connector. Solder it to the connector, or put the screw into the brass block described earlier. Carefully proceed to bend the tubing in a circular winding from one support to the next.

See the earlier section entitled “50-Ω Helix Feed” and Figs 19 and 20 to see how the first half-turn of the helix tubing must be positioned close above the reflector assembly. **Fig 59B** shows also an excellent example by K9EK on matching his U-band helical antenna to a 52-Ω feed line. It is important to maintain this spacing, since extra capacitance between the tubing and ground is required for impedance-matching purposes.

Insert a piece of #22 copper wire in the hole in each support as you go. Twist the wire around the tubing and the support dowel. Solder the wire to the tubing and to itself to keep the tubing in the notches. Continue in this way until all eight turns have been wound. After winding the helix, pinch the far end of the tubing together and solder it closed.

Weatherproofing the Wood

A word about preparing the maple doweling is in order. Wood parts must be protected from the weather to ensure long service life. A good way to protect wood is to boil it in paraffin for about half an hour. Any holes to be drilled in the wooden parts should be drilled after the paraffin is applied, since epoxy does not adhere well to wood after it has been coated with paraffin. The small dowels can be boiled in a saucepan. Caution must be exercised here—the wood can be scorched if the paraffin is too hot. Paraffin is sold for canning purposes at most grocery stores. Wood parts can also be protected with three or four coats of spar varnish. Each coat must be allowed to dry fully before another coat is applied.

The fiberglass tube or wood dowel must fit snugly with the steel pipe. The dowel can be sanded or turned down to the appropriate diameter on a lathe. If fiberglass is used, it can be coupled to the pipe with a piece of wood dowel that fits snugly inside the pipe and the tubing. Epoxy the dowel splice into the pipe for a permanent connection.

Drill two holes through the pipe and dowel and bolt them together. The pipe provides a solid mount to the boom of the rotator, as well as most of the weight needed to counterbalance the antenna. More weight can be added to the pipe if the assembly is "front-heavy." (Cut off some of the pipe if the balance is off in the other direction.)

The helix has a nominal impedance of about 105 Ω in this configuration. By varying the spacing of the first half turn of tubing, a good match to 52-Ω coax should be obtainable. When the spacing has been established for the first half turn to provide a good match, add pieces of polystyrene or Teflon rod stock between the tubing and the reflector assembly to maintain the spacing. These can be held in place on the reflector assembly with silicone sealant. Be sure to seal the type-N connector with the same material.

Exposed Antenna Relays and Preamplifiers

For stations using crossed Yagi antennas for CP operation, one feature that has been quite helpful for communicating through most of the LEO satellites, has been the ability to switch polarization from RHCP to LHCP. In some satellite operation this switchable CP ability has been essential. Operation through AO-40 has not shown a great need for such CP agility, since if the satellite is seriously off-pointed the signals are not particularly useable. When AO-40's squint angle is less than 25° the need for LHCP has not been observed. For those using helical antennas or helical-fed dish antennas, we just would not have the choice to switch CP unless an entirely new antenna is added to the cluster for that purpose. Not many of us have the luxury of that kind of space available on our towers.

For stations with switchable-circularity Yagi antennas, experience with exposed circularity switching relays and preamplifiers mounted on antennas have shown that they are prone to failure caused by an elusive mechanism known as *diurnal pumping*. Often these relays are covered with a plastic case, and the seam between the case and PC board is sealed with a silicone sealant. Preamps may also have a gasket seal for the cover, while the connectors can easily leak air. None of these methods create a true hermetic seal and as a result the day/night temperature swings pump air and moisture in and out of the relay or preamp case. Under the right conditions of temperature and moisture content, moisture from the air will condense inside the case when the outside air cools down. Condensed water builds up inside the case, promoting extensive corrosion and unwanted electrical conduction, seriously degrading component performance in a short time.

A solution for those antennas with "sealed" plastic relays, such as the KLM CX series; you can avoid problems by making the modifications shown in **Fig 60**. Relocate the 4:1 balun as shown and place a clear polystyrene plastic refrigerator container over the relay. Notch the container edges for the driven element and the boom so the container will sit down over the relay, sheltering it from the elements. Bond the container in place with a few dabs of RTV adhesive sealant. Position the antenna in an "X" orientation, so neither set of elements is parallel to the ground. The switcher board should now be canted at an angle, and one side of the relay case should be lower than the other. An example for the protective cover for an S-band preamp can be seen in the discussion on feeds for parabolic antennas.

For both the relay and preamp cases, carefully drill a 3/32-inch hole through the low side of the case to provide the needed vent. The added cover keeps rainwater off the

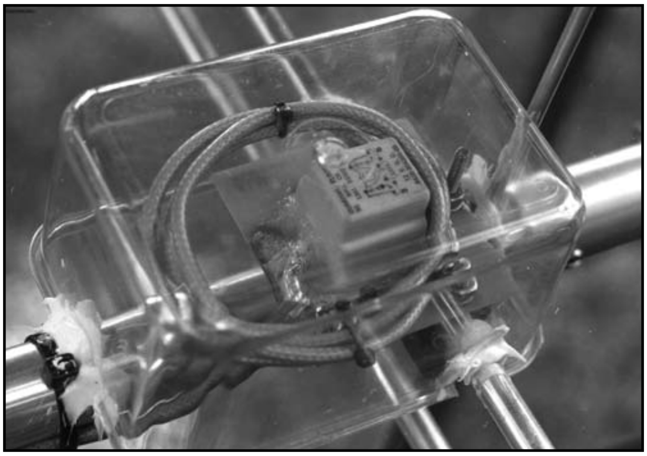

Fig 60—KLM 2M-22C antenna CP switching relay with relocated balun. The protective cover is needed for rain protection, be sure to use a polystyrene kitchen box, see text. (*WD4FAB photo.*)

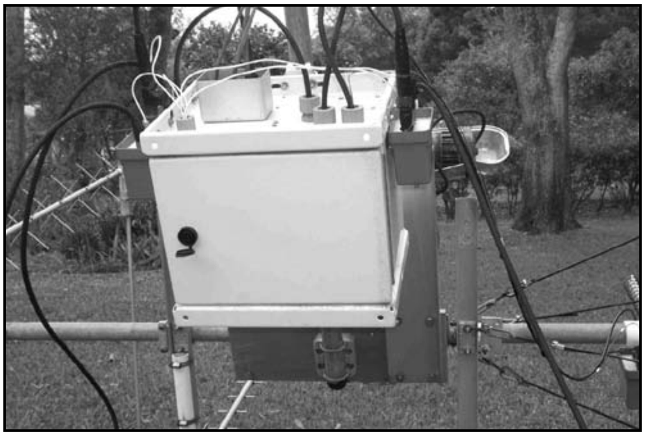

Fig 61—A NEMA4 box is used to shelter the L-band electronics and power supply. The box flanges are convenient for mounting preamplifiers. The box is shown inverted since it is on a tilt-over tower. (*WD4FAB photo.*)

relay and preamp, and the holes will prevent any buildup of condensation inside the relay case. Relays and preamplifiers so treated have remained clean and operational over periods of years without problems.

Another example for the protection of remotely, tower-mounted equipment is shown in Fig 50, illustrating the equipment box and mast-mounted preamplifiers at the top of WD4FAB's tower. The commercial NEMA4-rated equipment box, detailed in **Fig 61** (shown inverted), is used to protect the 23-cm power amplifier and its power supply, as well as a multitude of electrical connections. This steel box is very weather resistant, with an exceptionally good epoxy finish, but it is not sealed and so it will not trap moisture to be condensed with temperature changes. Be sure to use a box with at least a NEMA3 rating for rainwater and dust protection. The NEMA4 rating is just a little better protection than the NEMA3 rating. Using a well-rated equipment box is very well worth the expense of the box. As you can see, the box also provides some pretty good flanges to mount the mast-mounted preamplifiers for three bands. This box is an elegant solution for the simple need of rain shelter for your equipment. See **Fig 62**.

Elevation Control

Satellite antennas need to have elevation control to point up to the sky. This is the "El" part of Az-El control of satellite antennas. Generally, elevation booms for CP satellite antennas need to be non-conducting so that the boom does not affect the radiation pattern of the antenna. In the example shown next, the elevation boom center section is a piece of extra-heavy-wall $1^1/_2$-inch pipe (for greater strength) coupled with a tubular fiberglass-epoxy boom extension on the 70-cm end and a home-brew long extension on the 2-meter end. This uses large PVC pipe reinforced with four braces of Phillystran non-metallic guy cable. (PVC pipe is notoriously flexible, but the Phillystran cables make a quite stiff and strong boom of the PVC pipe.) For smaller installations, a continuous piece of fiberglass-epoxy boom can be placed directly through the elevation rotator.

Elevation boom motion needs to be powered, and one solution by WD4FAB, shown in **Fig 63**, uses a surplus jackscrew drive mechanism. I8CVS has also built his own robust elevation mechanism. See **Fig 64**. Note in each of these applications the methods used to provide bearings for the elevation mechanism. In WD4FAB's case, the elevation axis is a piece of heavy-duty $1^1/_2$-inch pipe,

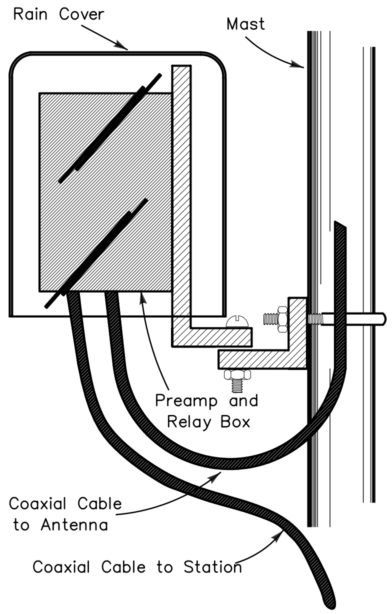

Fig 62—Protection for tower-mounted equipment need not be elaborate. Be sure to dress the cables as shown so that water drips off the cable jacket before it reaches the enclosure. One hazard for such open-bottom enclosures is that of animals liking the cable insulation as a delicacy. Flying insects also like to build their houses in these enclosures.

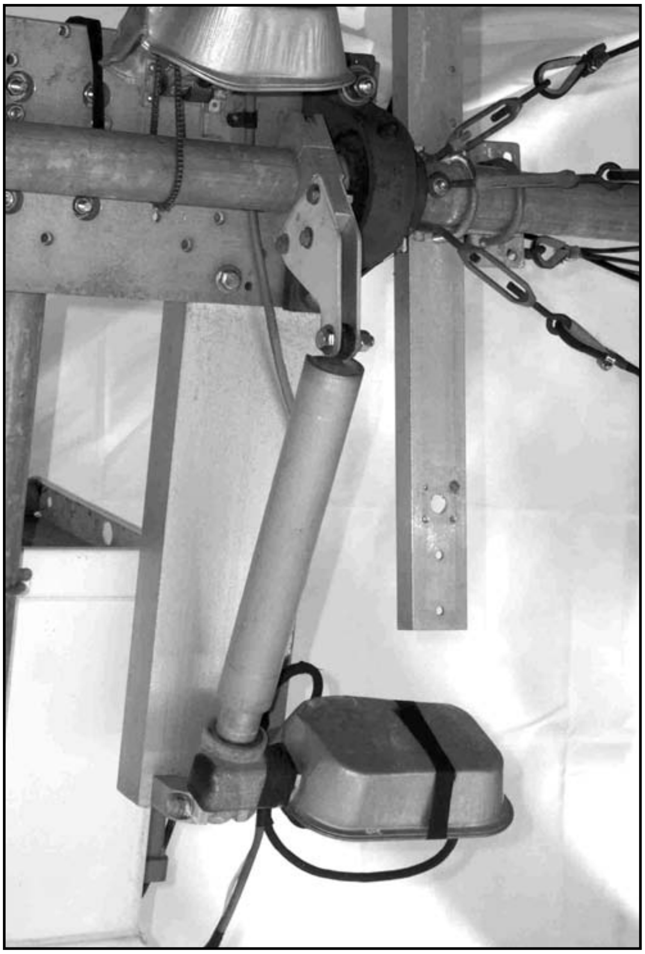

Fig 63—WD4FAB's homebrew elevation rotator drive using a surplus-store drive screw mechanism. Note also the large journal bearing supporting the elevation axis pipe shaft. (*WD4FAB photo.*)

($1^{15}/_{16}$-inch OD) and large 2 inch journal bearings are used for the motion. I8CVS uses a very large hinge to allow his motion.

Robust commercial solutions for Az-El rotators have given operators good service over the years. **See Fig 65**. Manufacturers such as Yaesu and M^2 are among these suppliers. One operator, VE5FP, found a solution for his Az-El needs by using two low-cost, lightweight TV rotators. See Fig 65B.

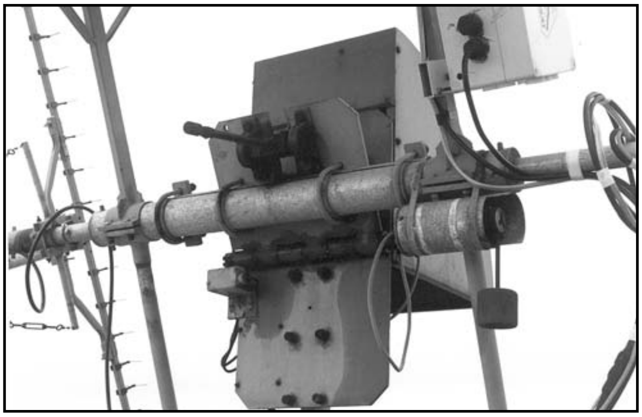

Fig 64—I8CVS's homebrew elevation mechanism using a very large, industrial hinge as the pivot and a jackscrew drive. (*I8CVS photo.*)

CONVERTED C-BAND TVRO DISHES

In working with larger, converted C-band TVRO dishes for AO-40, some operators have used only the polar mount with its jack-screw mechanism. See **Fig 66**. This dish is called *Big Ugly Dish* or just "BUD" by their users. Only using the polar mount mechanism limits the operator in the range of motion, as previously discussed. WØLMD provides for a greater degree of articulation of these dishes through several mechanisms. One of these is a sector-gear elevation drive, shown in **Fig 67**.

For the azimuth motion of our satellite antennas, most use motorized rotator drives, mainly the commercial sources previously mentioned. Most antennas are tower-mounted, allowing the placement of the rotator inside the tower. For the large wind loads of satellite antennas, these commercial rotators become rather expensive.

High loads are also prominent with the use of BUD

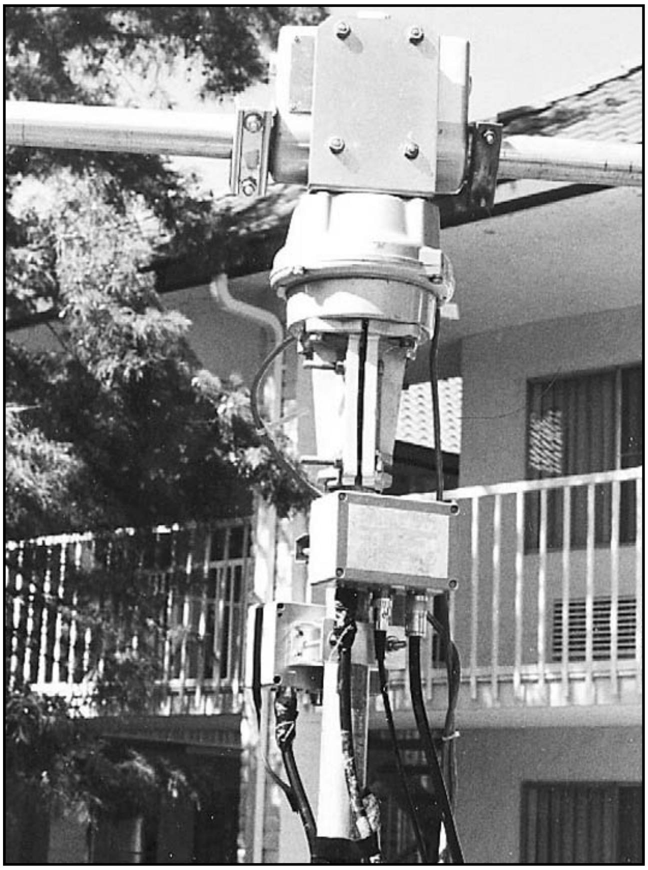

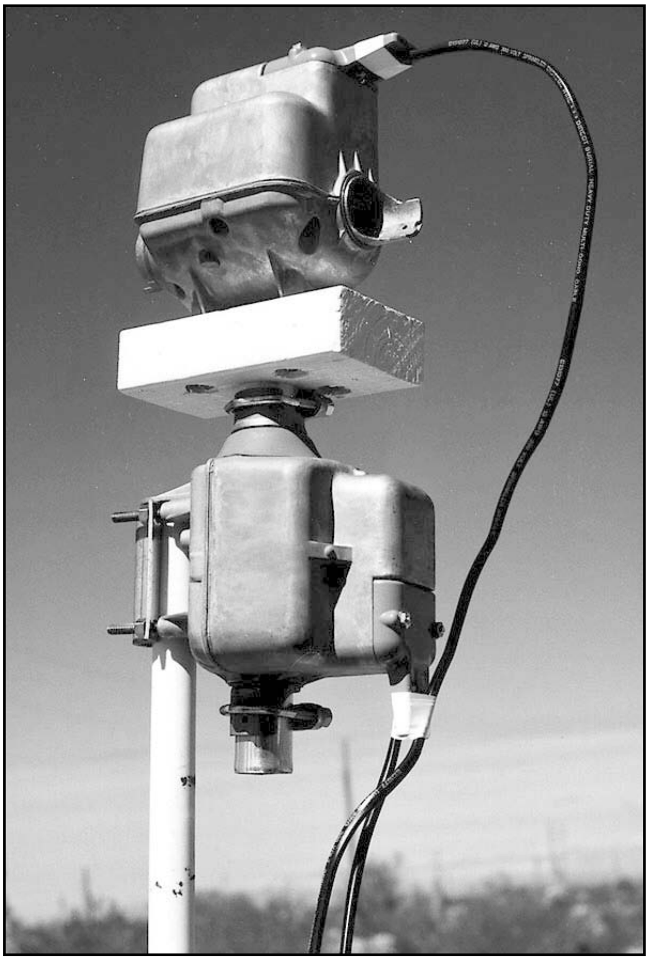

Fig 65—At left, Yaesu Az-El antenna-rotator mounting system is shown. Note that antenna loads must be more carefully balanced on this rotator than in the previously shown systems. At right, VE5FP has a solution for his Az-El rotators by bolting two of them together in his "An Inexpensive Az-El Rotator System", *QST*, December 1998.

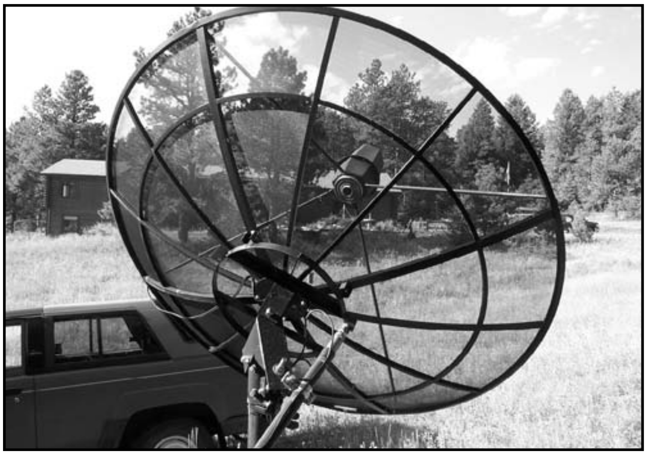

Fig 66—A TVRO dish-drive system is shown on its polar mount, using a protected drive-screw mechanism. (*WØLMD photo.*)

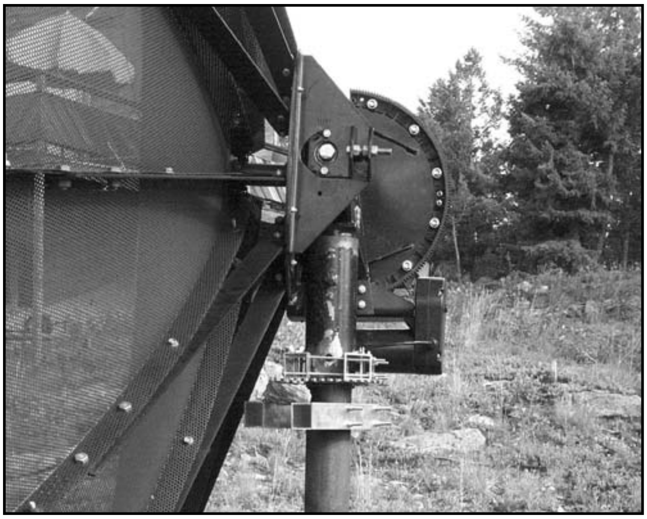

Fig 67—A modified TVRO dish mount is shown using an Az-El mount and a sector-gear drive for the elevation. (*WØLMD photo.*)

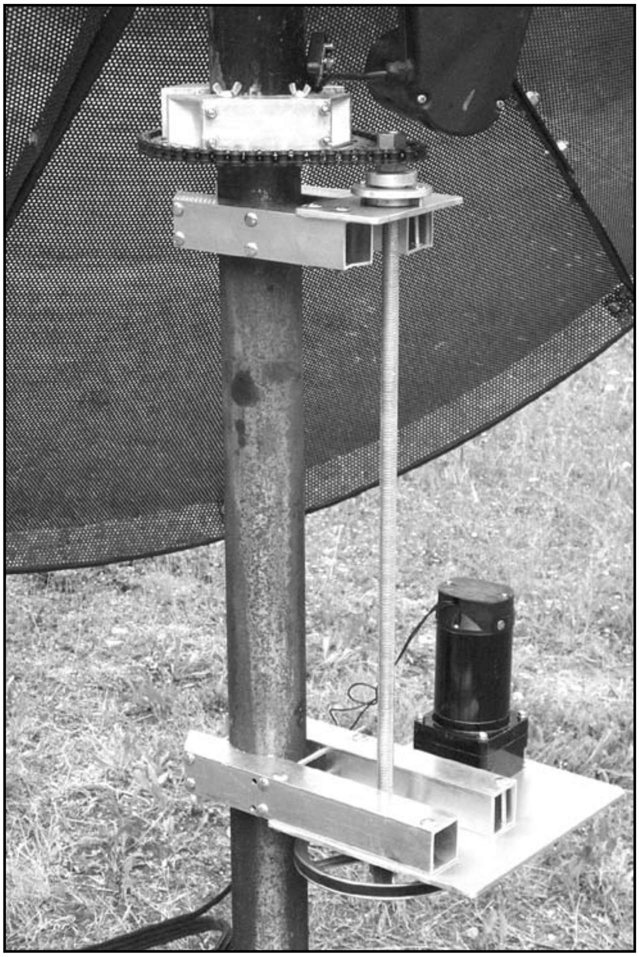

Fig 68—WØLMD constructed a very robust and low-cost Az drive mechanism. (*WØLMD photo.*)

antennas, and WØLMD has again engineered some very robust mechanisms using combinations of motorcycle-chain drives, V-belt drives and gear-head motors, as seen in **Fig 68**. An overall view of one of his BUD antennas is shown in **Fig 69**, showing the Az drive with an El drive that uses a jackscrew mechanism.

Operators through the years have employed many methods for the control of their antenna positions, ranging from true *arm-strong* manual positioning, to manual operation of the powered antenna azimuth and elevation rotators, to fully automated computer control of the rotators. While computer control of the rotators is not essential, life is greatly assisted with their use. For many years, one of the keystone control units for rotators has been the *Kansas City Tracker* (KCT) board installed in your computer. Most satellite-tracking programs can connect to the KCT with ease. One difficulty with the KCT unit is that they are 8-bit digital units, providing positioning precision of 0.35° in elevation and 1.41° in azimuth. For the larger dishes, with their narrow beamwidths, these values of precision are unacceptable. There are other options to replace the KCT unit.

A recent trend for amateur antenna control has been evolving in the form of a standalone controller that translates computer antenna-position information into controller commands with an understanding of antenna-position limits. These boxes, represented by the *EasyTrak* unit, **Fig 70**, from the Tucson Amateur Packet Radio (TAPR) group, have made this capability readily available for many amateurs. This unit is a 10-bit encoder, providing precisions of 0.09° in elevation and 0.35° in azimuth. The computer can also control the operation of your station transceiver through the radio interface provided in *EasyTrak*; you will not need any other radio interface.

Other position readout and control options are available. For many years ham operators have employed synchros, or *selsyns*, for their position readouts. These are

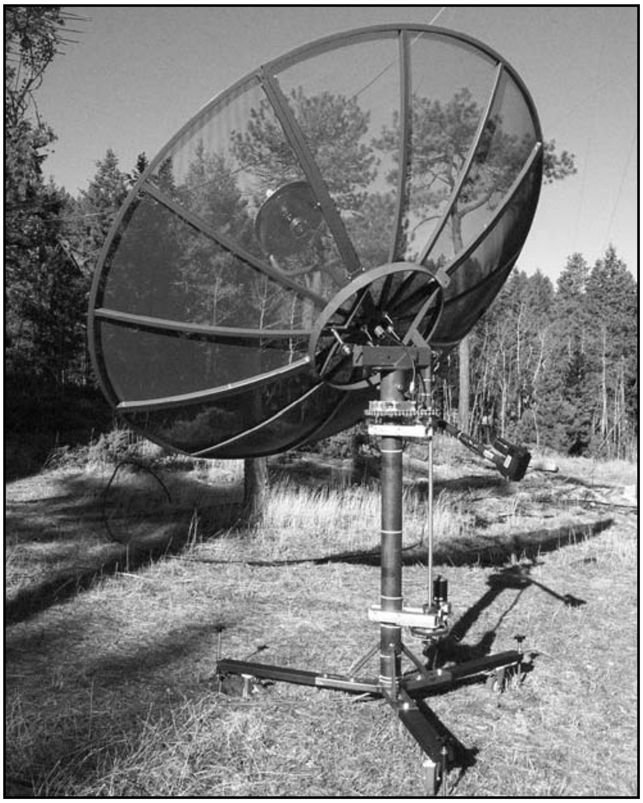

Fig 69—A completed TVRO dish Az-El mounting system is shown, using a jackscrew elevation drive. (*WØLMD photo.*)

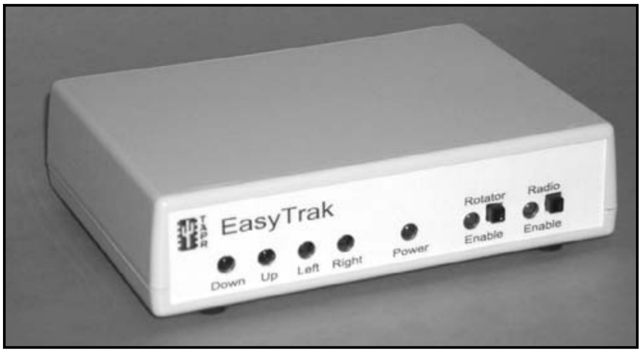

Fig 70—The EasyTrak automated antenna rotator and radio controller by TAPR. (*WD4FAB photo.*)

specialized transformers, using principles developed over sixty years ago and employed in such devices as surplus "radio compass" steering systems for aircraft. While the position readout of these devices can be quite precise, in general they only provide a visual position indication, one that is not easily adapted to computer control. I8CVS employs such a system at his station and his elevation synchro can be seen in Fig 64, using a weighted arm on the synchro to provide a constant reference to the Earth's gravity vector.

The more up-to-date, computer-friendly position readout methods used these days are usually based on precision potentiometers or digital code wheels. **Fig 71** shows such a digital code-wheel system employed by WØLMD. He notes that such systems, while providing a very high precision of angular position, they are not absolute systems and that once calibrated, they must be continually powered so they do not lose their calibration. Precision potentiometers, on the other hand, provide an absolute position reference, but with a precision that is limited to the quality of the potentiometer, typically 0.5% (0.45° in El and 1.80° in Az) to 1.0%. So the choices have their individual limits, unless a lot of money is spent for very precise commercial systems.

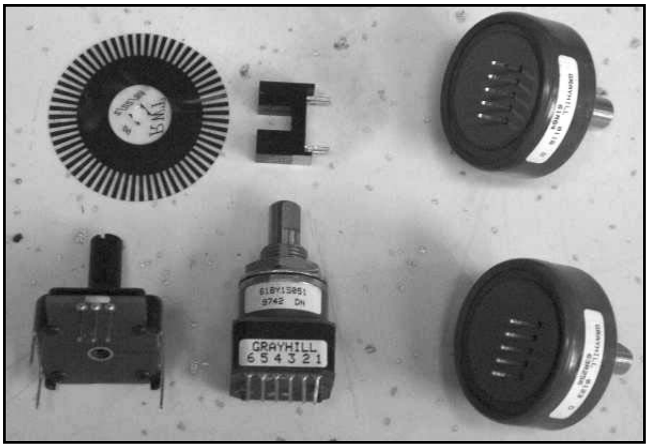

Fig 71—WØLMD has experimented with highly precise optical encoders for his antenna position systems. See text. (*WØLMD photo.*)

NOTES AND REFERENCES, SATELLITE ANTENNAS

AO-40 frequencies, "Official Transponder Frequency Bandplan for P-3D"; **www.amsat-dl.org/p3dqrg.html**.

Brown, Gerald R., K5OE, "A Helix Feed for Surplus MMDS, Antennas" *Proceedings of the 2001AMSAT-NA Symposium*, October 2001, pp 89-94, also: **members.aol.com/k5oe**.

Brown, Jerry, K5OE, "A K-Band Receiver for AO-40", *Proceedings for the AMSAT-NA Space Symposium, 2002*.

Brown, Gerald R., K5OE, "Build This No-Tune Dual-Band Feed for Mode L/S" *The AMSAT Journal*, Vol 26, No 1, Jan/Feb 2003.

Brown, Gerald R., K5OE, "Dual-Band Dish Feeds for 13/23 cm," *Proceedings of the 2002AMSAT-NA Symposium*, October 2002, pp 123-131.

Brown, Jerry, K5OE, **members.aol.com/k5oe/** MMDS Dishes.

Brown, Gerald R., K5OE, "Patch Feeds," **members.aol.com/k5oe**.

Brown, Jerry, K5OE, "The Texas Potato Masher: A Medium-Gain Directional Satellite Antenna For

LEOs," *The AMSAT Journal*, Vol 22, No. 1, Jan/Feb 1999.

Jansson, Dick, WD4FAB, *QST*, Sep 2002, "Product Review", "M^2 23CM22EZA 1.2 GHz Antenna", p 59.

Kingery, Mike, KE4AZN, "Setting Up for AO-40 L-Band Uplink," *The AMSAT Journal*, May/Jun 2002, pp 14-16, also: **web.infoave.net/~mkmk518/**.

Koehler, Jim, VE5FP, "An Inexpensive Az-El Rotator System", *QST*, Dec 1998.

Kraus, John D., "The Helical Antenna", Chapter 7, *Antennas*, McGraw-Hill Book Company, 1988.

Kraus, John D., W8JK, "16-12, Patch or Microstrip Antennas," *Antennas*, 1988, pp 745-749.

Krome, Ed, K9EK, "Development of a Portable Mode S Ground Station." *The AMSAT Journal*, Vol 16, No. 6, Nov/Dec 1993, pp 25-28.

Krome, Ed, K9EK, "S band Reception: Building the DEM Converter and Preamp Kits," *The AMSAT Journal*, Vol 16, No. 2, Mar/Apr 1993, pp 4-6.

Krome, Ed., K9EK, *Mode S—The Book*, p 96, 109.

Krome, Ed, K9EK, "Mode S: Plug and Play!" *The AMSAT Journal*, Vol 14, No. 1, Jan 1991, pp 21-23, 25.

Long, Howard, G6LVB, "My Shack Configuration—Spring 2002": **www.g6lvb.com/g6lvb_shack_spring_2002.htm**.

McCaa, William D., K0RZ, "Hints on Using the AMSAT-OSCAR 13 Mode S Transponder," *The AMSAT Journal*, Vol 13, No. 1, Mar 1990, pp 21-22.

MacAllister, Andy, W5ACM, "Field Day 2002," *73 Amateur Radio Today*, Sept. 2002, pp 48-52.

Miller, James, G3RUH, "Mode S — Tomorrow's Downlink?" *The AMSAT Journal*, Vol 15, No. 4, Sep/Oct 1992, pp 14-15.

Miller, James S., G3RUH, "'Patch' Feed For S-Band Dish Antennas"; **www.jrmiller.demon.co.uk/products/patch.html**.

Miller, James, G3RUH, "A 60-cm S-Band Dish Antenna," *The AMSAT Journal*, Vol 16 No. 2, Mar/Apr 1993, pp 7-9.

Miller, James, G3RUH, "Small is Best," *The AMSAT Journal*, Vol 16, No. 4, Jul/Aug 1993, p 12.

Monteiro, Anthony, "Work OSCAR 40 with Cardboard-box Antennas!," *QST*, Mar 2003, pp 57-62.

Orban Microwave, "The Basics of Patch Antennas: **www.orbanmicrowave.com/antenna_application_notes.htm**.

Seguin, Mike, N1JEZ, "OSCAR 40 on 24GHz", *QST*, Dec 2002, pp 55-56.

Seydler, Robert, K5GNA,"Modifications of the AIDC 3731 Downconverters", **members.aol.com/k5gna/AIDC3731modifications.doc**.

Suckling, Charles, G3WDG, "K-Band Results From AO-40": **www.g3wdg.free-online.co.uk/kband.htm**.

Suckling, Charles, G3WDG "Notch Filters for AO-40 Mode L/S": **www.g3wdg.free-online.co.uk.notch.htm**.

Suding, Dr. Robert, WØLMD, "Converting TVRO Dish & Dishes For Amateurs" **www.ultimatecharger.com/dish.html**.

Thiel, David & Smith, Staphani, *Switched Parasitic Antennas for Cellular Communications*, Artech House, 2002, "Chapter 3, Patch Antennas," pp 79-96.

Wade, Paul, N1BWT (W1GHZ), *Online Microwave Antenna Handbook*, "Chapter 4, Parabolic Dish Antennas," **www.w1ghz.cx/antbook/chap4.pdf**, Mar. 98. (See CD-ROM bundled with this book.)

Zibrat,Timothy S, K3TZ, "2.4 GHz Patch Design": **www.qsl.net/k3tz/**.

Antenna Systems for EME Communications

This section was updated by David Hallidy, K2DH. As mentioned earlier, the tremendous path loss incurred over an EME circuit places stringent requirements on Earth-station performance. Low-noise receiving equipment, maximum available power and high-gain antenna arrays are required for successful EME operation. Although it is possible to copy some of the better-equipped stations with a low-gain antenna, it is unlikely that such an antenna can provide reliable two-way communications. Antenna gain of at least 20 dBi is required for reasonable EME success. Generally speaking, more antenna gain yields the most noticeable improvement in station performance, since the increased gain improves both the received and transmitted signals.

VHF/UHF EME ANTENNAS

Several types of antennas for 2 meters and 70 cm are popular among EME enthusiasts. Perhaps the most popular antenna for 144-MHz work is an array of either 4 or 8 long-boom (14 to 15 dBi gain) Yagis. The 4-Yagi array provides approximately 20 dB gain, and an 8-Yagi array gives an approximate 3 dB increase over the 4-antenna array. **Fig 72** shows the computed response at a 30° tilt above the horizon for a stack of four 14-element 2-meter Yagis, each with a boomlength of 3.1 λ (22 feet).

At 432 MHz, EME enthusiasts often use 8 or 16 long-boom Yagis in an array. Such Yagis are commercially available or they can be constructed from readily available materials. Chapter 18, VHF and UHF Antenna Systems, has details on some popular Yagi designs.

The main disadvantage of Yagi arrays is that the polarization plane of the individual Yagis cannot be conveniently changed. One way around this is to use cross-polarized Yagis and a relay switching system to select the desired polarization, as described in the previous section. This represents a considerable increase in system complexity to select the desired polarization. Some amateurs have gone so far as to build complicated mechanical systems to allow constant polarization adjustment of all the Yagis in a large array. **Fig 73** shows the K1FO 70-cm EME 16-Yagi array with full polarization control, described in *The ARRL Antenna Compendium, Vol 3*. This 432-MHz EME array uses open-wire phasing lines to minimize feed-line losses. **Fig 74** shows the computed response for this array, which employs rugged but lightweight 14-element Yagis on 3.1 λ (7.1 foot) booms. Feed-line losses are not explicitly accounted for in the *EZNEC Professional* computer model, but are estimated to be less than 0.25 dB.

Polarization shift of EME signals at 144 MHz is fairly rapid, and the added complexity of a relay-controlled cross-polarized antenna system or a mechanical polarization adjustment scheme is probably not worth the effort. At 432 MHz, however, where the polarization shifts at a much slower rate, an adjustable polarization system does offer a definite advantage over a fixed one.

The Yagi antenna system used by Ed Stallman, N5BLZ, is shown in **Fig 75**. His system employs twelve 144-MHz long-boom 17-element Yagi antennas. The monster 48-Yagi 2-meter array of Gerald Williamson, K5GW, is shown in **Fig 76**, and the huge 48-Yagi 70-cm EME array of Frank Potts, NC1I, is shown in **Fig 77**.

Although not as popular as Yagis, Quagi antennas (made from both quad and Yagi elements) are sometimes used for EME work. Slightly more gain per unit boom length is possible as compared to the conventional Yagi, at the expense of some robustness. Additional information on the Quagi is presented in Chapter 18, VHF and UHF Antenna Systems.

The collinear array is an older type of antenna for EME work. A 40-element collinear array has approximately

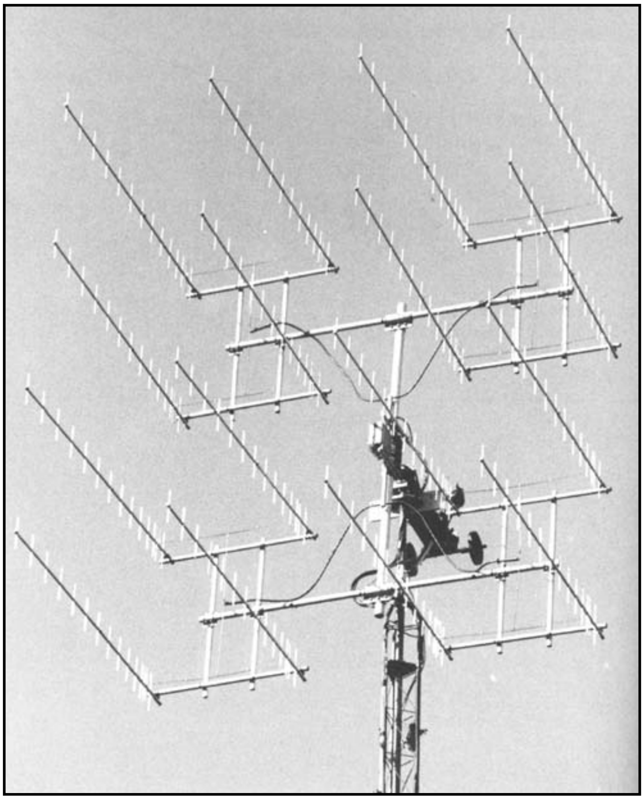

Fig 73—K1FO's variable polarization 16 × 14-element (3.6-λ boom lengths) 432-MHz EME array shown at 2° elevation and vertical polarization. (See *The ARRL Antenna Compendium, Vol 3*.)

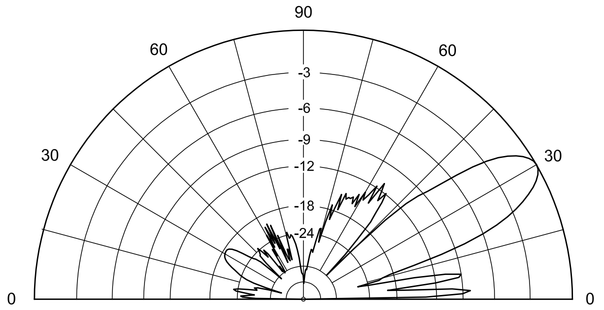

Fig 72—*EZNEC Pro* elevation pattern for four 14-element 2-meter Yagis (3.6-λ boom lengths) at an elevation angle of 30° above the horizon. The computed system gain is 21.5 dBi, suitable for 2-meter EME. This assumes that the phasing system is made of open-wire transmission lines so that feed-line losses can be kept below 0.25 dB.

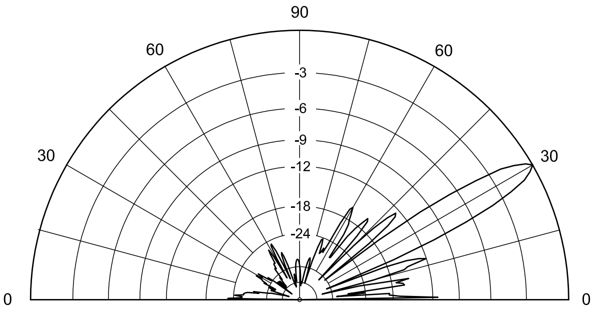

Fig 74—Computed elevation response for K1FO 16-Yagi 432-MHz array shown in Fig 73. (The *EZNEC Pro* model required 2464 segments!) With assumed phasing harness feed-line losses of 0.25 dB, the overall gain exceeds 27.5 dBi.

the same frontal area as an array of four Yagis, but produces approximately 1 to 2 dB less gain. One attraction to a collinear array is that the depth dimension is considerably less than the long-boom Yagis. An 80-element collinear is marginal for EME communications, providing approximately 19 dB gain. As with Yagi and Quagi antennas, the collinear cannot be adjusted easily for polarity changes. From a construction standpoint, there is little difference in complexity and material costs between the collinear and Yagi arrays.

DISH ANTENNAS FOR EME

On 2 meters the minimum antenna gain for reliable EME communications is about 20 dBi. While a few amateurs have had access to parabolic dishes large enough for EME work at 144 and 222 MHz, at those frequencies an array of four long Yagis is equal in gain to a dish 24 feet in diameter! To achieve truly high-gain performance from a dish on 2 meters would require a reflector diameter of nearly 96 feet (providing 32 dBi gain). Such undertakings are generally beyond amateur means, so there has been little work done with dishes at low frequencies, except for the occasional expedition to one of the large radio telescopes that have accommodated amateur EME work.

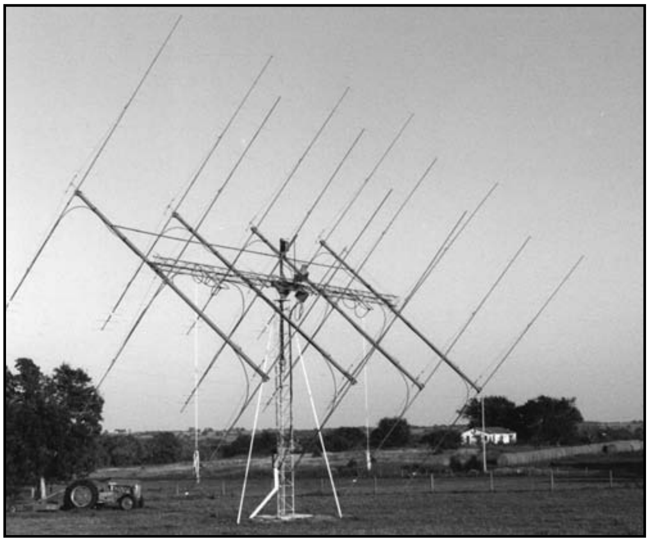

Fig 75—The EME array used at N5BLZ consists of twelve long-boom 144-MHz Yagis. The tractor, lower left, really puts this array into perspective! (*Photo courtesy N5BLZ.*)

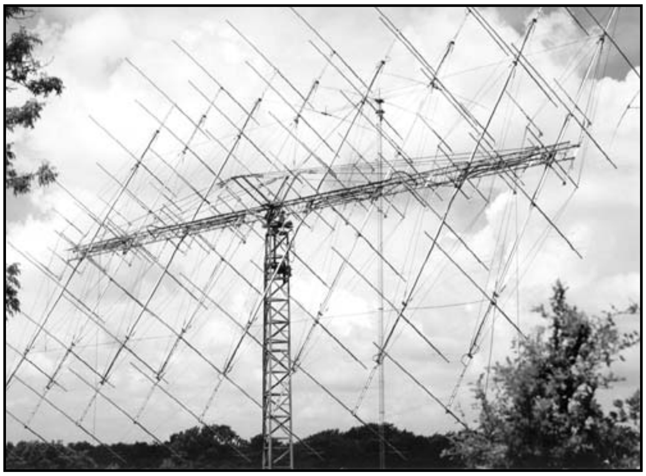

Fig 76—K5GW's huge 48-Yagi 2-meter EME array. (*Photo courtesy K5GW.*)

Microwave Parabolic Dish Antennas

The major problems associated with parabolic dish antennas are mechanical ones. A dish of about 16 feet in diameter is the minimum size required for successful EME operation on 432 MHz. With wind and ice loading, structures of this size place a real strain on the mounting and positioning system. Extremely rugged mounts are required for large dish antennas, especially when used in windy locations. **Fig 78** shows the impressive 7-meter diameter

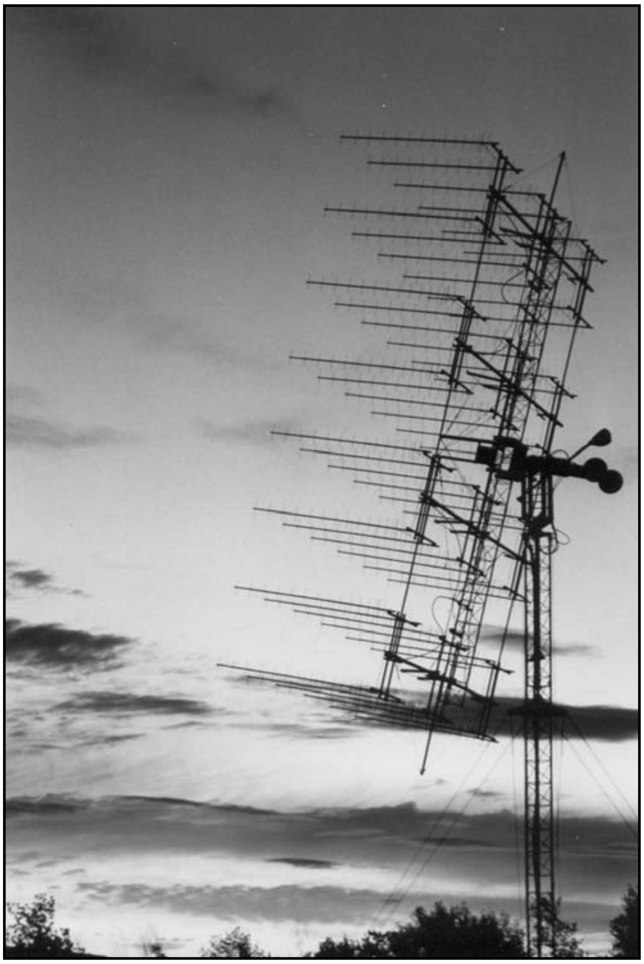

Fig 77—NC1I's magnificent 48-Yagi 70-cm EME array. (*Photo courtesy NC1I.*)

dish built by David Wardley, ZL1BJQ.

Several aspects of parabolic dish antennas make the extra mechanical problems worth the trouble, however. For example, the dish antenna is inherently broadband, and may be used on several different amateur bands by simply changing the feed. An antenna that is suitable for 432 MHz work will most likely be usable on several of the higher amateur bands too. Increased gain is available as the frequency of operation is increased.

Another advantage of a dish is the flexibility of the feed system. The polarization of the feed, and therefore the polarization of the antenna, can be changed with little difficulty. It is a relatively easy matter to devise a system to rotate the feed remotely from the shack to change polarization. Because polarization changes can account for as much as 30 dB of signal attenuation, the rotatable feed can make the difference between consistent communications and no communications at all. Further information on Parabolic Antennas can be found in Chapter 18, VHF and UHF Antenna Systems as well as in the section below.

A 12-FOOT STRESSED HOMEBREW PARABOLIC DISH

Very few antennas evoke as much interest among UHF amateurs as the parabolic dish, and for good reason. First, the parabola and its cousins—Cassegrain, hog horn and Gregorian—are probably the ultimate in high-gain antennas. One of the highest-gain antennas in the world (148 dB) is a parabola. This is the 200-inch Mt. Palomar telescope. (The very short wavelength of light rays causes such a high gain to be realizable.)

Second, the efficiency of the parabola does not change as size increases. With Yagis and collinear arrays, the losses in the phasing harness increase as the array size increases. The corresponding component of the parabola is lossless air between the feed horn and the reflecting surface. If there are a few surface errors, the efficiency of the system stays constant regardless of antenna size. This project was presented by Richard Knadle, K2RIW, in August 1972 *QST*.

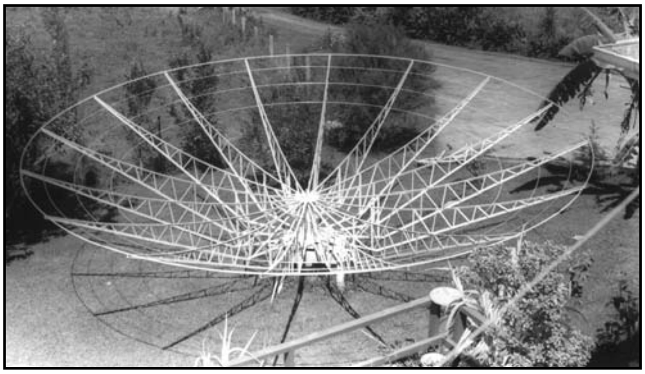

Fig 78—ZL1BJQ's homemade 7-meter (23-foot) parabolic dish, just prior to adding $^1/_2$-inch wire mesh. (*Photo courtesy ZL1BJQ*.)

Some amateurs reject parabolic antennas because of the belief that they are all heavy, hard-to-construct, have large wind-loading surfaces and require precise surface accuracy. However, with modern construction techniques, a prudent choice of materials and an understanding of accuracy requirements, these disadvantages can be largely overcome. A parabola may be constructed with a 0.6 f/D (focal length/diameter) ratio, producing a rather flat dish, which makes it easy to surface and allows the use of recent advances in high-efficiency feed horns. This results in greater gain for a given dish size over conventional designs.

Such an antenna is shown in **Fig 79**. This parabolic dish is lightweight, portable, easy to build, and can be used for 432 and 1296-MHz mountain topping, as well as on 2304, 3456 and 5760 MHz. Disassembled, it fits into the trunk of a car, and can be assembled in 45 minutes.

The usually heavy structure that supports the surface of most parabolic dish antennas has been replaced in this design by aluminum spokes bent into a near parabolic shape by string. These strings serve the triple function of guying the focal point, bending the spokes and reducing the error at the dish perimeter (as well as at the center) to nearly zero. By contrast, in conventional designs, the dish perimeter (which has a greater surface area than the center) is farthest from the supporting center hub. For these reasons, it often has the greatest error. This error becomes more severe when the wind blows.

Here, each of the spokes is basically a cantilevered beam with end loading. The equations of beam bending

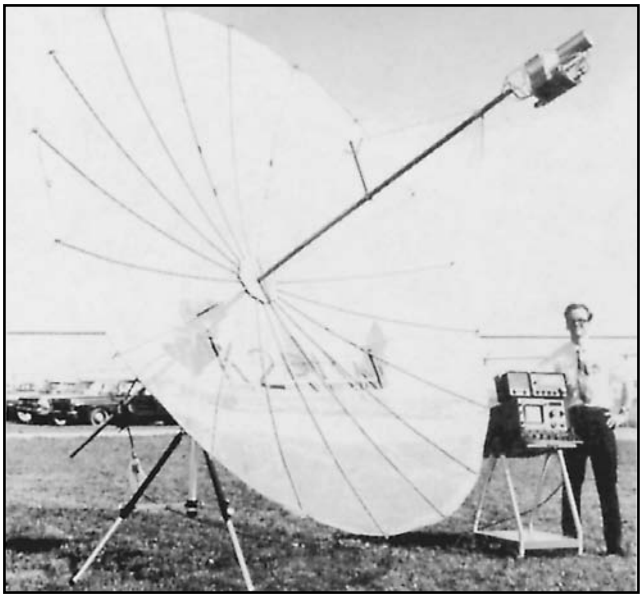

Fig 79—A 12-foot stressed parabolic dish set up for satellite signals near 2280 MHz. A preamplifier is shown taped below the feed horn. The dish was designed by K2RIW, standing at the right. From *QST*, August 1972.

predict a near-perfect parabolic curve for extremely small deflections. Unfortunately the deflections in this dish are not that small and the loading is not perpendicular. For these reasons, mathematical prediction of the resultant curve is quite difficult. A much better solution is to measure the surface error with a template and make the necessary correction by bending each of the spokes to fit. This procedure is discussed later.

The uncorrected surface is accurate enough for 432 and 1296-MHz use. Trophies taken by this parabola in antenna-gain contests were won using a completely natural surface with no error correction. By placing the transmission line inside the central pipe that supports the feed horn, the area of the shadows or blockages on the reflector surface is much smaller than in other feeding and supporting systems, thus increasing gain. For 1296 MHz, a backfire feed horn may be constructed to take full advantage of this feature. At 432 MHz, a dipole and reflector assembly produces 1.5 dB additional gain over a corner-reflector feed system. Because the preamplifier is located right at the horn on 2300 MHz, a conventional feed horn may be used.

Construction

Table 2 is a list of materials required for construction. Care must be exercised when drilling holes in the connecting center plates so assembly problems will not be experienced later. See **Fig 80**. A notch in each plate allows them to be assembled in the same relative positions. The two plates should be clamped together and drilled at the same time. Each of the $18^{1}/_{2}$-inch diameter aluminum spokes has two no. 28 holes drilled at the base to accept no. 6-32 machine screws that go through the center plates. The 6-foot long spokes are cut from standard 12-foot lengths of tubing. A fixture built from a block of aluminum assures that the holes are drilled in exactly the same position in each spoke. The front and back center plates constitute an I-beam type of structure that gives the dish center considerable rigidity.

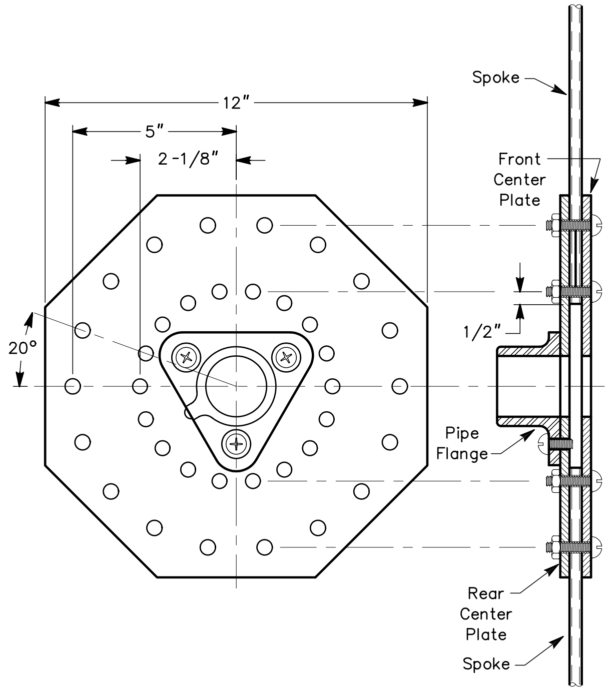

Fig 80—Center plate details. Two center plates are bolted together to hold the spokes in place.

A side view of the complete antenna is shown in **Fig 81**. Aluminum alloy (6061-T6) is used for the spokes, while 2024-T3 aluminum alloy sheet, $^{1}/_{8}$ inch thick, is used for the center plates. (Aluminum has approximately three times the strength-to-weight ratio of wood, and aluminum cannot warp or become water logged.) The end of each of the 18 spokes has an eyebolt facing the dish focal point, which serves a dual purpose:

1) To accept the #9 galvanized fence wire that is routed through the screw eyes to define the dish perimeter, and
2) To facilitate rapid assembly by accepting the S hooks which are tied to the end of each of the lengths of 130-pound test Dacron fishing string.

The string bends the spokes into a parabolic curve; the dish may be adapted for many focal lengths by tightening or slackening the strings. Dacron was chosen because it has the same chemical formula as Mylar. This is a low-stretch material that keeps the dish from changing shape. The galvanized perimeter wire has a 5-inch overlap area that is bound together with baling wire after the spokes have been hooked to the strings.

The aluminum window screening is bent over the perimeter wire to hold it in place on the back of the spokes. Originally, there was concern that the surface perturbations (the spokes) in front of the screening might decrease the gain. The total spoke area is so small, however, that

Table 2

Materials List for the 12-Foot Stressed Parabolic Dish

1) Aluminum tubing, 12 ft × $^{1}/_{2}$ in. OD × 0.049-in. wall, 6061-T6 alloy, 9 required to make 18 spokes.
2) Octagonal mounting plates 12 × 12 × $^{1}/_{8}$ in., 2024-T3 alloy, 2 required.
3) $1^{1}/_{4}$ in. ID pipe flange with setscrews.
4) $1^{1}/_{4}$ in. × 8 ft TV mast tubing, 2 required.
5) Aluminum window screening, 4 × 50 ft.
6) 130-pound test Dacron trolling line.
7) 38 ft #9 galvanized fence wire (perimeter).
8) Two hose clamps, $1^{1}/_{2}$ in.; two U bolts; $^{1}/_{2}$ × 14 in. Bakelite rod or dowel; water-pipe grounding clamp; 18 eye bolts; 18 S hooks.

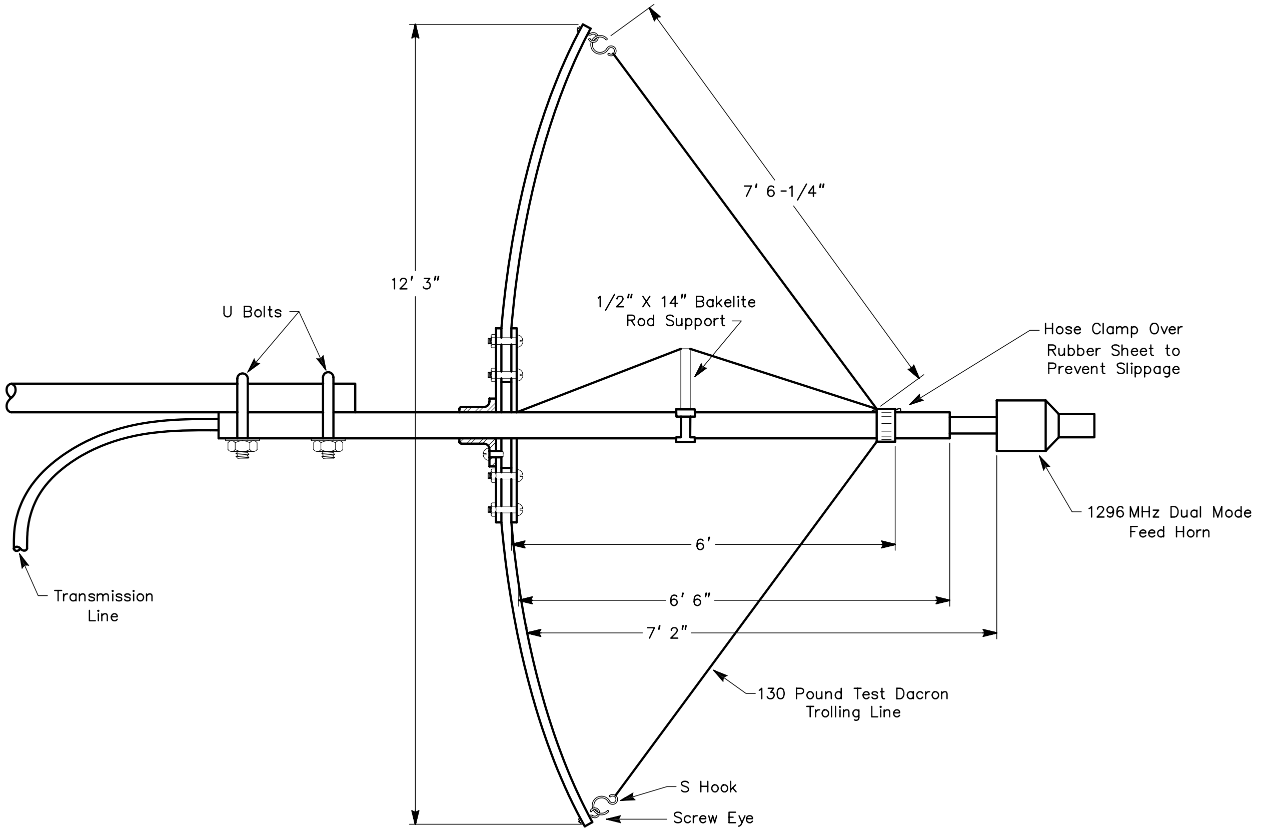
7′ 6-1/4″
12′ 3″
U Bolts
1/2″ X 14″ Bakelite Rod Support
Hose Clamp Over Rubber Sheet to Prevent Slippage
1296 MHz Dual Mode Feed Horn
6′
6′ 6″
7′ 2″
Transmission Line
130 Pound Test Dacron Trolling Line
S Hook
Screw Eye

Fig 81—Side view of the stressed parabolic dish.

this fear proved unfounded.

Placing the aluminum screening in front of the spokes requires the use of 200 pieces of baling wire to hold the screening in place. This would increase the assembly time by at least an hour. For contest and mountaintop operation (when the screening is on the back of the spokes) no fastening technique is required other than bending the screen to overlap the wire perimeter.

The Parabolic Surface

A 4-foot wide roll of aluminum screening 50 feet long is cut into appropriate lengths and laid parallel, with a 3-inch overlap between the top of the unbent spokes and hub assembly. The overlap seams are sewn together on one half of the dish using heavy Dacron thread and a sailmaker's curved needle. Every seam is sewn twice; once on each edge of the overlapped area. The seams on the other half are left open to accommodate the increased overlap that occurs when the spokes are bent into a parabola. The perimeter of the screening is then trimmed. Notches are cut in the 3-inch overlap to accept the screw eyes and S hooks.

The first time the dish is assembled, the screening strips are anchored to the inside surface of the dish and the seams sewn in this position. It is easier to fabricate the surface by placing the screen on the back of the dish frame with the structure inverted. The spokes are sufficiently strong to support the complete weight of the dish when the perimeter is resting on the ground.

The 4-foot wide strips of aluminum screening conform to the compound bend of the parabolic shape very easily. If the seams are placed parallel to the E-field polarization of the feed horn, minimum feedthrough will occur. This feedthrough, even if the seams are placed perpendicular to the E field, is so small that it is negligible. Some constructors may be tempted to cut the screening into pie-shaped sections. This procedure will increase the seam area and construction time considerably. The dish surface appears most pleasing from the front when the screening perimeter is slipped between the spokes and the perimeter wire, and is then folded back over the perimeter wire. In disassembly, the screening is removed in one piece, folded in half, and rolled.

The Horn and Support Structure

The feed horn is supported by $1^{1}/_{4}$-inch aluminum television mast. The Hardline that is inserted into this tubing is connected first to the front of the feed horn, which then slides back into the tubing for support. A setscrew assures that no further movement of the feed horn occurs. During antenna-gain competition the setscrew is omit-

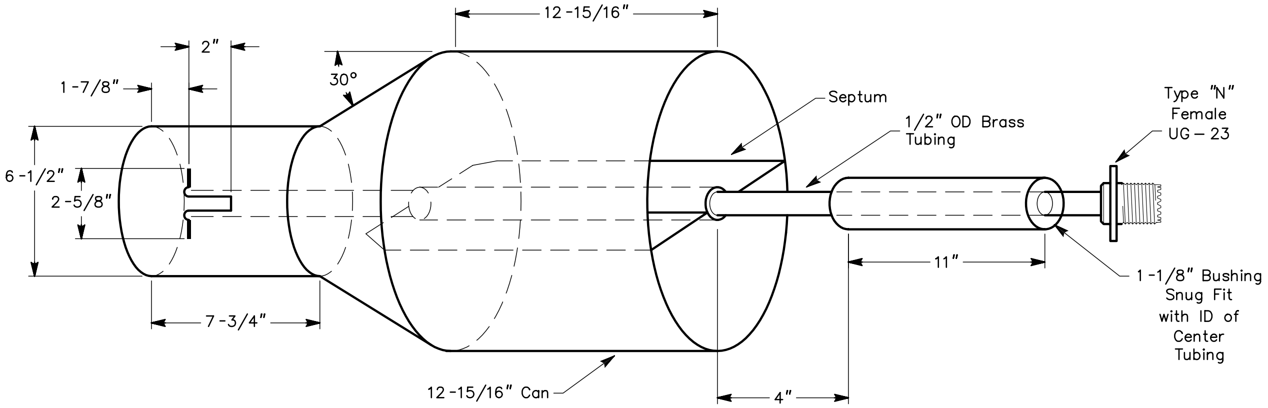

Fig 82—Backfire type 1296-MHz feed horn, linear polarization only. The small can is a Quaker State oil container; the large can is a 50-pound shortening container (obtained from a restaurant, Gold Crisp brand). Brass tubing, $^{1}/_{2}$-inch OD extends from UG-23 connector to dipole. Center conductor and dielectric are obtained from $^{3}/_{8}$-inch Alumafoam coaxial cable. The dipole is made from 3/32-inch copper rod. The septum and 30° section are made from galvanized sheet metal. Styrofoam is used to hold the septum in position. The primary gain is 12.2 dBi.

ted, allowing the $^{1}/_{2}$-inch semirigid CATV transmission line to move in or out while adjusting the focal length for maximum gain. The TV mast is held firmly at the center plates by two setscrews in the pipe flange that is mounted on the rear plate. At 2300 MHz, the dish is focused for best gain by loosening these setscrews on the pipe flange and sliding the dish along the TV mast tubing. (The dish is moved instead of the feed horn.)

The fishing strings are held in place by attaching them to a hose clamp that is permanently connected to the TV tubing. A piece of rubber sheet under the hose clamp prevents slippage and keeps the hose clamp from cutting the fishing string. A second hose clamp is mounted below the first as extra protection against slippage.

The high-efficiency 1296-MHz dual mode feed horn, detailed in **Fig 82**, weighs $5^{3}/_{4}$ pounds. This weight causes some bending of the mast tubing, but this is corrected by a $^{1}/_{2}$-inch diameter bakelite support, as shown in Fig 81. This support is mounted to a pipe grounding clamp with a no. 8-32 screw inserted in the end of the rod. The bakelite rod and grounding clamp are mounted midway between the hose clamp and the center plates on the mast. A double run of fishing string slipped over the notched upper end of the bakelite rod counteracts bending.

The success of high-efficiency parabolic antennas is primarily determined by feed horn effectiveness. The multiple diameter of this feed horn may seem unusual. This patented dual-mode feed, designed by Dick Turrin, W2IMU, achieves efficiency by launching two different kinds of waveguide modes simultaneously. This causes the dish illumination to be more constant than conventional designs.

Illumination drops off rapidly at the perimeter, reducing spillover. The feed backlobes are reduced by at least 35 dB because the current at the feed perimeter is almost zero; the phase center of the feed system stays constant across the angles of the dish reflector. The larger diameter section is a phase corrector and should not be changed in length. In theory, almost no increase in dish efficiency can be achieved without increasing the feed size in a way that would increase complexity, as well as blockage.

The feed is optimized for a 0.6 f/D dish. The dimensions of the feeds are slightly modified from the original design in order to accommodate the cans. Either feed type can be constructed for other frequencies by changing the scale of all dimensions.

Multiband Use

Many amateurs construct multiband antenna arrays by putting two dishes back to back on the same tower. This is cost inefficient. The parabolic reflector is a completely frequency independent surface, and studies have shown that a 0.6 f/D surface can be steered seven beamwidths by moving the feed horn from side to side before the gain diminishes by 1 dB. Therefore, the best dual-band antenna can be built by mounting separate horns side by side. At worst, the antenna may have to be moved a few degrees (usually less than a beamwidth) when switching between horns, and the unused horn increases the shadow area only slightly. In fact, the same surface can function simultaneously on multiple frequencies, making crossband duplex operation possible with the same dish.

Order of Assembly

1) A single spoke is held upright behind the rear center plate with the screw eye facing forward. Two no. 6-

32 machine screws are pushed through the holes in the rear center plate, through the two holes of the spoke, and into the corresponding holes of the front center plate. Lock washers and nuts are placed on the machine screws and hand tightened.

2) The remaining spokes are placed between the machine screw holes. Make sure that each screw eye faces forward. Machine screws, lock washers and nuts are used to mount all 18 spokes.
3) The no. 6-32 nuts are tightened using a nut driver.
4) The mast tubing is attached to the spoke assembly, positioned properly, and locked down with the setscrews on the pipe flange at the rear center plate. The S hooks of the 18 Dacron strings are attached to the screw eyes of the spokes.
5) The ends of two pieces of fishing string (which go over the bakelite rod support) are tied to a screw eye at the forward center plate.
6) The dish is laid on the ground in an upright position and #9 galvanized wire is threaded through the eyebolts. The overlapping ends are lashed together with baling wire.
7) The dish is placed on the ground in an inverted position with the focus downward. The screening is placed on the back of the dish and the screening perimeter is fastened as previously described.
8) The extension mast tubing (with counterweight) is connected to the center plate with U bolts.
9) The dish is mounted on a support and the transmission line is routed through the tubing and attached to the horn.

Parabola Gain Versus Errors

How accurate must a parabolic surface be? This is a frequently asked question. According to the Rayleigh limit for telescopes, little gain increase is realized by making the mirror accuracy greater than ± $^1/_8$ λ peak error. John Ruze of the MIT Lincoln Laboratory, among others, has derived an equation for parabolic antennas and built models to verify it. The tests show that the tolerance loss can be predicted within a fraction of a decibel, and less than 1 dB of gain is sacrificed with a surface error of ±$^1/_8$ λ. (A $^1/_8$ λ is 3.4 inches at 432 MHz, 1.1 inches at 1296 MHz and 0.64 inch at 2300 MHz.)

Some confusion about requirements of greater than $^1/_8$-λ accuracy may be the result of technical literature describing highly accurate surfaces. Low sidelobe levels are the primary interest in such designs. Forward gain is a much greater concern than low sidelobe levels in amateur work; therefore, these stringent requirements do not apply.

When a template is held up against a surface, positive and negative (±) peak errors can be measured. The graphs of dish accuracy requirements are frequently plotted in terms of RMS error, which is a mathematically derived function much smaller than ± peak error (typically $^1/_3$). These small RMS accuracy requirements have discouraged many constructors who confuse them with ± peak errors.

Fig 83 may be used to predict the resultant gain of various dish sizes with typical errors. There are a couple of surprises, as shown in **Fig 84**. As the frequency is increased for a given dish, the gain increases 6 dB per octave until the tolerance errors become significant. Gain deterioration then increases rapidly. Maximum gain is realized at the frequency where the tolerance loss is 4.3 dB. Notice that at 2304 MHz, a 24-foot dish with ±2-inch peak errors has the same gain as a 6-foot dish with ±1-inch peak errors. This is quite startling, when it is realized that a 24-foot dish has 16 times the area of a 6-foot dish. Each time the diameter or frequency is doubled or halved, the gain changes by 6 dB. Each time all the errors are halved, the frequency of maximum gain is doubled. With this information, the gain of other dish sizes with other tolerances can be predicted.

These curves are adequate for predicting gain, assuming a high-efficiency feed horn is used (as described earlier), which realizes 60% aperture efficiency. At frequencies below 1296 MHz where the horn is large and causes considerable blockage, the curves are somewhat optimistic. A properly built dipole and splasher feed will have about 1.5 dB less gain when used with a 0.6 f/D dish than the dual-mode feed system described.

The worst kind of surface distortion is where the surface curve in the radial direction is not parabolic but gradually departs in a smooth manner from a perfect parabola. The decrease in gain can be severe, because a large area is involved. If the surface is checked with a template, and if reasonable construction techniques are employed, deviations are controlled and the curves represent an upper limit to the gain that can be realized.

If a 24-foot dish with ±2-inch peak errors is being used with 432 and 1296-MHz multiple feed horns, the constructor might be discouraged from trying a 2300-MHz feed because there is 15 dB of gain degradation. The dish will still have 29 dBi of gain on 2300 MHz, however, making it worthy of consideration.

The near-field range of this 12-foot stressed dish (actually 12 feet 3 inches) is 703 feet at 2300 MHz. By using the sun as a noise source and observing receiver noise power, it was found that the antenna had two main lobes about 4° apart. The template showed a surface error (insufficient spoke bending at $^3/_4$ radius), and a correction was made. A recheck showed one main lobe, and the solar noise was almost 3 dB stronger.

Other Surfacing Materials

The choice of surface materials is a compromise between RF reflecting properties and wind loading. Aluminum screening, with its very fine mesh (and weight of 4.3 pounds per 100 square feet) is useful beyond 10 GHz because of its very close spacing. This screening is easy to roll up and is therefore ideal for a portable dish. This close spacing causes the screen to be a 34% filled aperture, bring-

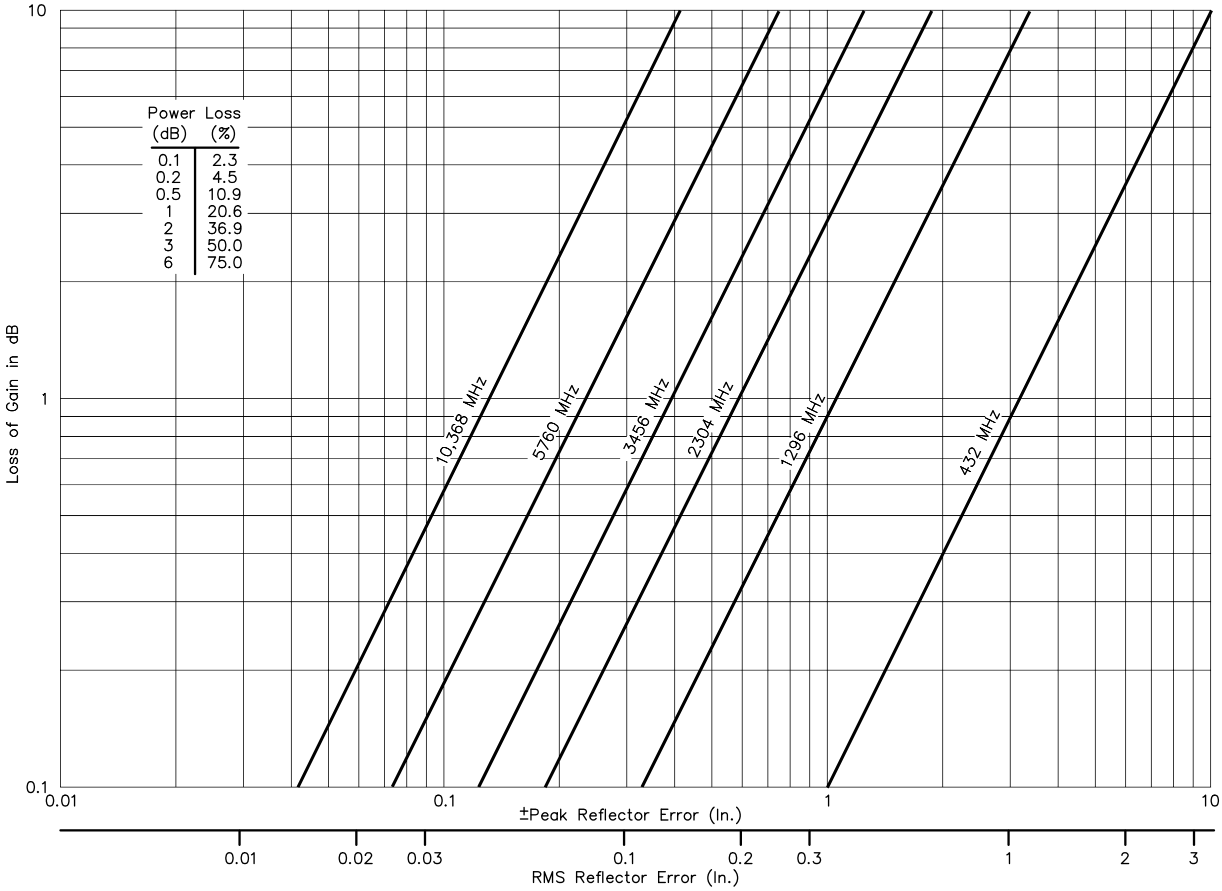

Power Loss (dB)	(%)
0.1	2.3
0.2	4.5
0.5	10.9
1	20.6
2	36.9
3	50.0
6	75.0

Fig 83—Gain deterioration versus reflector error. By Richard Knadle, K2RIW.

ing the wind force at 60 mph to more than 400 *pounds* on this 12-foot dish. Those considering a permanent installation of this dish should investigate other surfacing materials.

Hexagonal 1-inch poultry netting (chicken wire), which is an 8% filled aperture, is nearly ideal for 432-MHz operation. It weighs 10 pounds per 100 square feet, and exhibits only 81 pounds of force with 60 mph winds. Measurement on a large piece reveals 6 dB of feedthrough at 1296 MHz, however. Therefore, on 1296 MHz, one fourth of the power will feed through the surface material. This will cause a loss of only 1.3 dB of forward gain. Since the low-wind loading material will provide a 30-dBi gain potential, it is still a very good tradeoff.

Poultry netting is very poor material for 2300 MHz and above, because the hole dimensions approach $^1/_2$ λ. As with all surfacing materials, minimum feedthrough occurs when the E-field polarization is parallel to the longest dimension of the surfacing holes.

Hardware cloth with $^1/_2$-inch mesh weighs 20 pounds per 100 square feet and has a wind loading characteristic of 162 pounds with 60 mph winds. The filled aperture is 16%, and this material is useful to 2300 MHz.

A rather interesting material worthy of investigation is $^1/_4$-inch reinforced plastic. It weighs only 4 pounds per 100 square feet. The plastic melts with many universal solvents such as lacquer thinner. If a careful plastic-melting job is done, what remains is the $^1/_4$-inch spaced aluminum wires with a small blob of plastic at each junction to hold the matrix together.

There are some general considerations to be made in selecting surface materials:

1) Joints of screening do not have to make electrical contact. The horizontal wires reflect the horizontal wave. Skew polarizations are merely a combination of horizontal and vertical components which are thus reflected by the corresponding wires of the screening. To a horizontally polarized wave, the spacing and diameter of only the horizontal wires determine the reflection coefficient (see **Fig 85**). Many amateurs have the mistaken impression that screening materials that do not make electrical contact at their junctions are poor reflectors.
2) By measuring wire diameter and spacings between the wires, a calculation of percentage of aperture that is filled can be made. This will be one of the major determining factors of wind pressure when the surfacing material is dry.

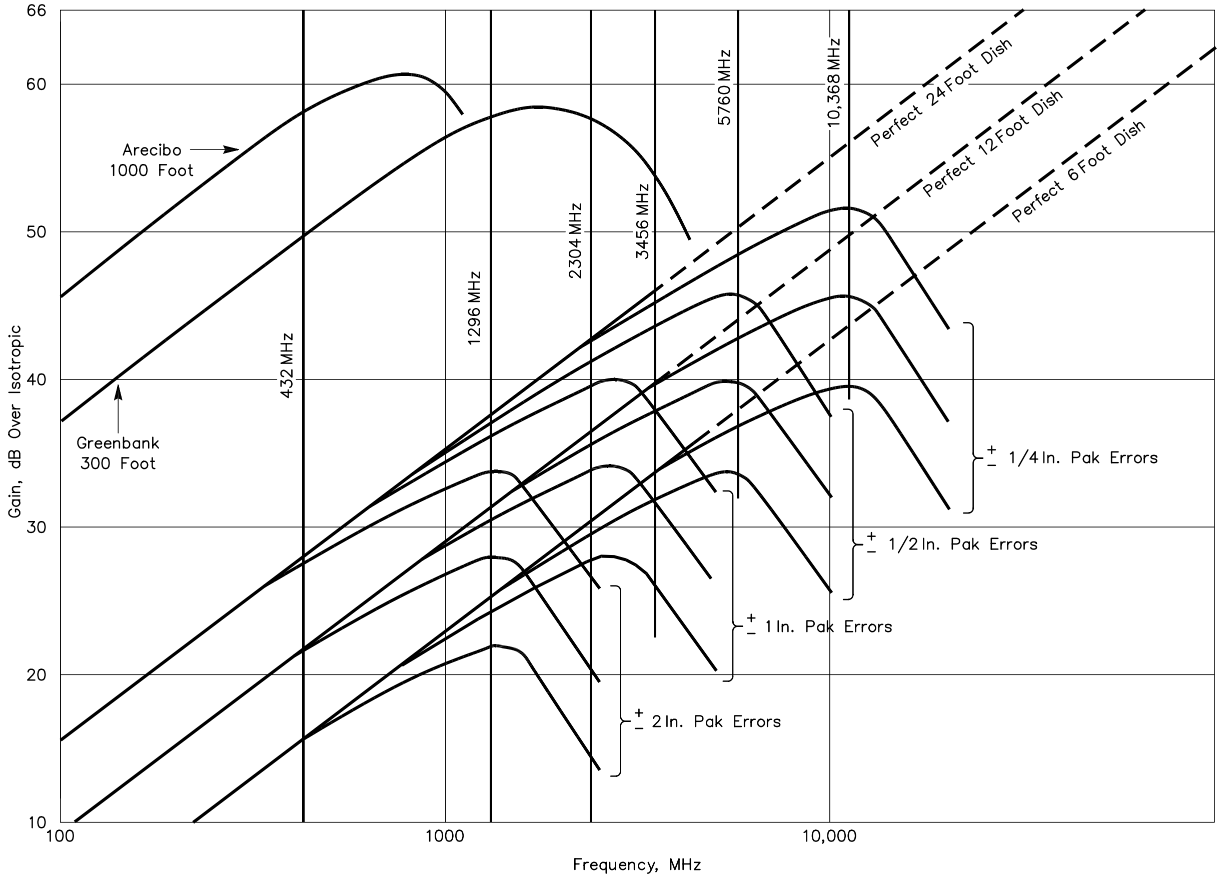

Fig 84—Parabolic-antenna gain versus size, frequency and surface errors. All curves assume 60% aperture efficiency and 10-dB power taper. Graph by K2RIW for ham bands, using display technique of J. Ruze, British IEE.

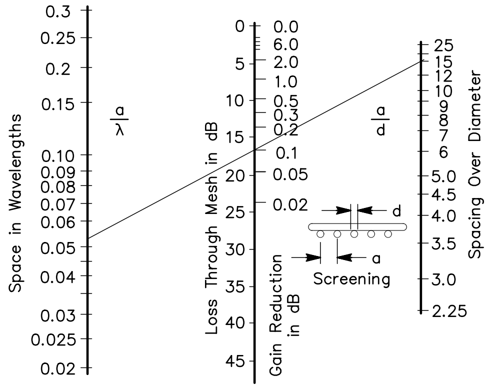

Fig 85—Surfacing material quality.

A Parabolic Template

At and above 2300 MHz (where high surface accuracy is required), a parabolic template should be constructed to measure surface errors. A simple template may be constructed (see **Fig 86**) by taking a 12-foot 3-inch length of 4-foot wide tar paper and drawing a parabolic shape on it with chalk. The points for the parabolic shape are calculated at 6-inch intervals and these points are connected with a smooth curve. For those who wish to use the template with the surface material installed, the template should be cut along the chalk line and stiffened by cardboard or a wood lattice frame. Surface error measurements should take place with all spokes installed and deflected by the fishing lines, as some bending of the center plates does take place. **Fig 87** shows the 12-foot stressed dish built by Franco Marcelo, N2UO.

Variations

All the possibilities of the stressed parabolic antenna have not been explored. For instance, a set of fishing lines or guy wires can be set up behind the dish for error correction, as long as this does not cause permanent bending of the aluminum spokes. This technique also protects the dish against wind loading from the rear. An extended piece of TV mast is an ideal place to hang a counterweight and attach the rear guys. This strengthens the structure considerably.

EME USING SURPLUS TVRO DISH ANTENNAS

Since the 1990s, there has been a significant change

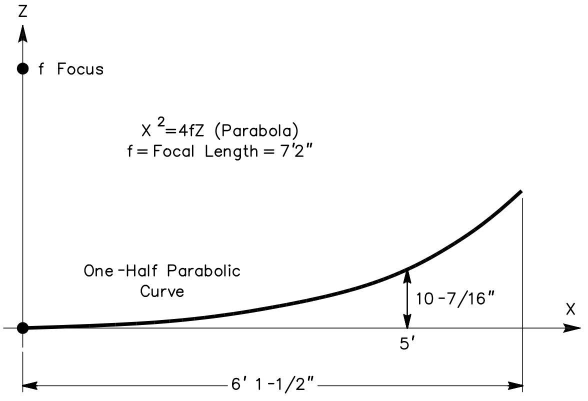

X	Z (to Nearest 1/16″)	X	Z
0.5′	1/8″	3.5′	5-1/8″
1.0′	7/16″	4.0′	6-11/16″
1.5′	15/16″	4.5′	8-1/2″
2.0′	1-11/16″	5.0′	10-7/16″
2.5′	2-5/8″	5.5′	12-11/16″
3.0′	3-3/4″	6.0′	15-1/16″

Fig 86—Parabolic template for 12-foot, 3-inch dish.

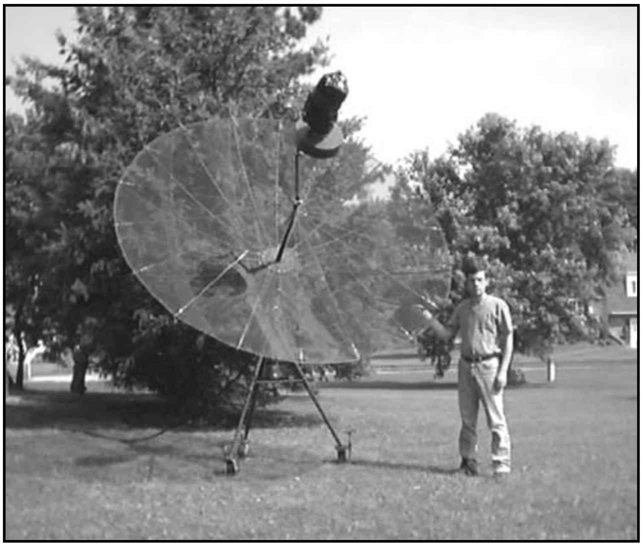

Fig 87—N2UO's homemade 12-foot stressed dish. (*Photo courtesy N2UO.*)

in the systems people use to watch satellite TV broadcasts. Formerly, C-band satellite receivers were used, along with parabolic dish antennas in the 3- to 5-meter diameter range. Now, Ku-band (12-GHz) receivers are the norm, with their associated small (usually 18-inch) dish antennas. This has provided a large body of surplus C-band dishes, which can be used for EME—certainly on the bands at 33 cm and above, and for the larger dishes (5 meters), even at 70 cm. Many times, these dishes and their mounts can be had for the asking, so they truly become an inexpensive way to build a multi-band EME antenna.

This updated article, first presented by David Hallidy, K2DH (ex-KD5RO) in the ARRL *UHF/Microwave Projects Manual*, describes the use of a 3-meter (10-foot) TVRO antenna in such an application. (Also see earlier in this chapter the section describing converted C-Band TVRO Dishes for satellite work.)

Background

Calculations show that a 3-meter dish will have about 30 dBi gain at 1296 MHz. With a state-of-the-art LNA (Low-Noise Amplifier or preamp) at the feed, an efficient feed horn illuminating the dish surface, and 200 W at 1296 MHz, lunar echoes should be easily detected and many stations can be worked. The biggest challenges to such a system are assembling the dish to its mount and steering it to track the Moon. As much as possible, the KISS ("Keep It Simple, Stupid") principle was used to accomplish this task.

In 1987, WA5TNY, KD5RO, KA5JPD, and W7CNK proved that such an EME system could work, even as high as 3.4 and 5.7 GHz, to provide the first EME contacts on those bands. An additional advantage to this (or any) small dish is its ability to be mounted to a trailer and taken out on EME expeditions. It can also be easily disassembled and stored, if necessary.

As can be seen from **Fig 88**, the entire setup is very simple, using a standard amateur tower as the main support for the dish.

Azimuth Drive

In azimuth, direct drive of the main rotating shaft was selected, and a small prop-pitch motor was used. These motors, while not as plentiful as they were some years ago, still turn up with some regularity at flea markets for very little money. The beauty of the prop-pitch motor is that it turns slowly, is reversible, provides very high torque, and requires no braking system (the gear reduction, on the order of 4000:1, provides the necessary braking). Prop-pitch motors are dc motors, and were designed to vary the pitch of propeller blades at start-up, take-off and landing of older large airplanes. Thus, they can be run at different speeds merely by varying the dc voltage to the motor, and can be reversed by reversing the polarity of the dc voltage. By mounting a thrust bearing of the appropriate size at the top of the tower, and mounting the motor directly below it at the end of the rotating shaft that turns the antenna, a simple direct-drive system can be constructed.

The dc power supply and control relays are located in a weatherproof box on the side of the tower, next to the motor. This system requires only 9 V dc at about 5 A to adequately start, turn and stop the prop-pitch motor, and this voltage turns the antenna through 360° of rotation in about $2^1/_2$ minutes. Azimuth position sensing is also a simple task. See **Fig 89**. A linear multi-turn potentiometer is driven by the rotating shaft, using a simple friction drive.

A strip of rubber is attached to the rotating shaft and a wheel is connected to the shaft of the pot. The pot is then mounted so that it presses against the rubber strip, and as the shaft turns so does the pot. If a ten-turn pot is used, and the system is aligned such that the pot is at the center of its rotation when the antenna is pointed approximately south, the pot will not rotate past the end at either extreme of the antenna's rotation (CW/CCW north), and absolute alignment is a simple task of calibrating the change in resistance (change in voltage, when the pot is fed from a constant voltage source) with degrees of rotation (see the discussion on Position Readout for details).

Elevation Drive

The elevation drive is also very simple. Most (nearly all) TVRO setups have a means of moving the dish across the sky to align it with various satellites. To do this, most companies use a device called a *Linear Actuator*. This is a dc motor to which is attached a long lead screw that pulls (or pushes) the outer shell of the actuator in or out to make it longer or shorter. The movable end of the actuator is attached to the dish and the motor end is fixed to the mount. The dish rests on pivots, which allow it to move as the actuator extends/retracts. To convert this type of mount (called a *Polar Mount*) to an Az/El mount is usually very simple.

Fig 90 shows how this can be done. Simply breaking the welds that held the mount in a polar fashion allows the mount to be turned on its side and used to pivot the dish vertically with the linear actuator. Another feature of linear actuators is that they also have some means of feeding their relative position to the satellite receiver. This is usually just a multi-turn potentiometer geared to the lead screw. All we have to do is connect this pot to a readout system, and we can calibrate the lift of the actuator in degrees. We thus have a simple means of rotating the dish and elevating it—but how do we know that it's pointed at the Moon?

Position Readout

Readout of the position of the antenna, in both azimuth and elevation is also a relatively simple task. On the

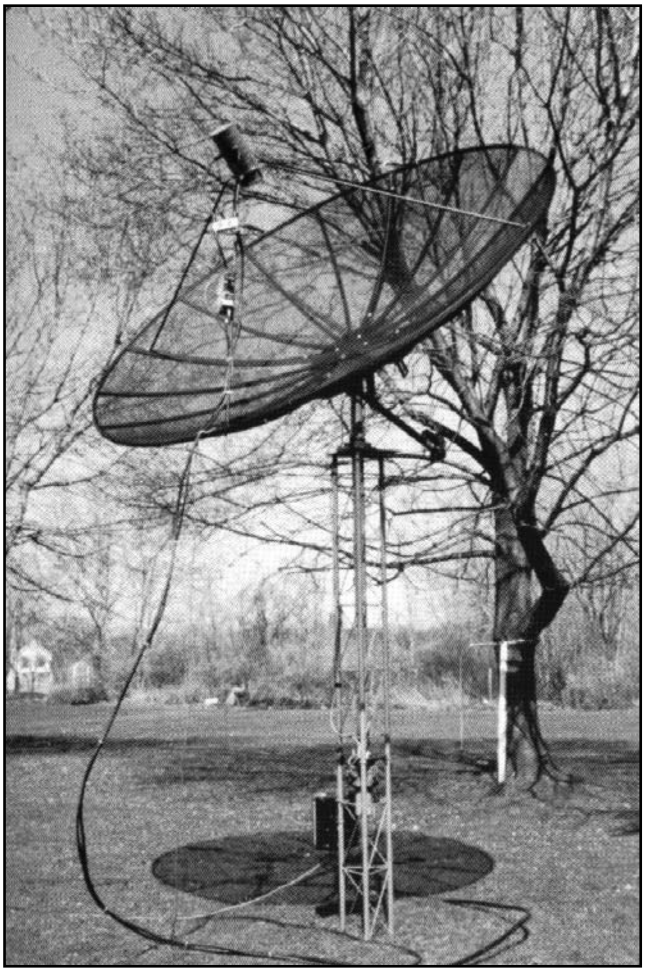

Fig 88—View of K2DH's (ex-KD5RO) complete TVRO antenna installation. (*K2DH photo.*)

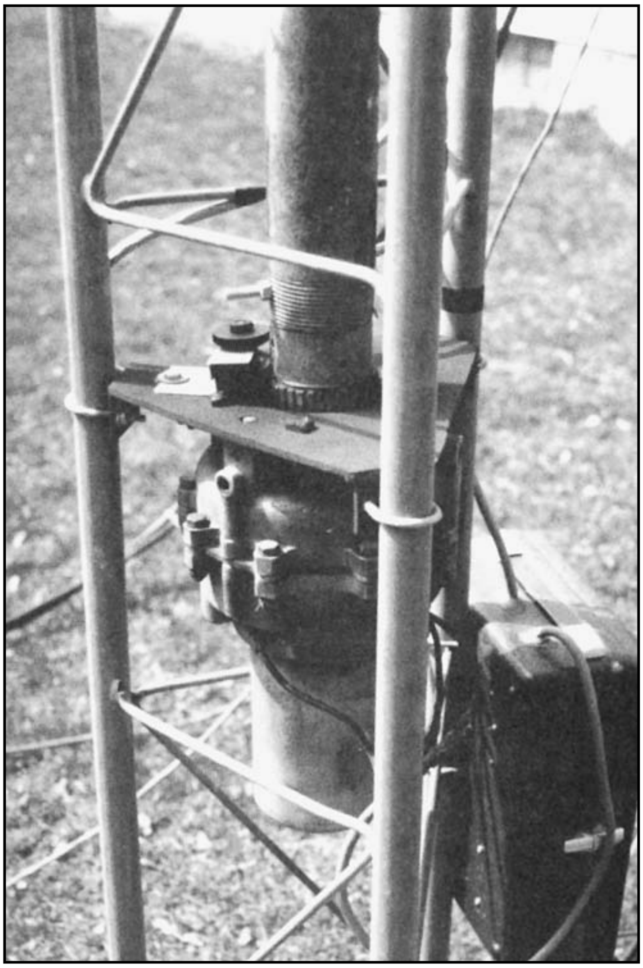

Fig 89—Azimuth rotation systems, showing prop-pitch motor and position sensor.

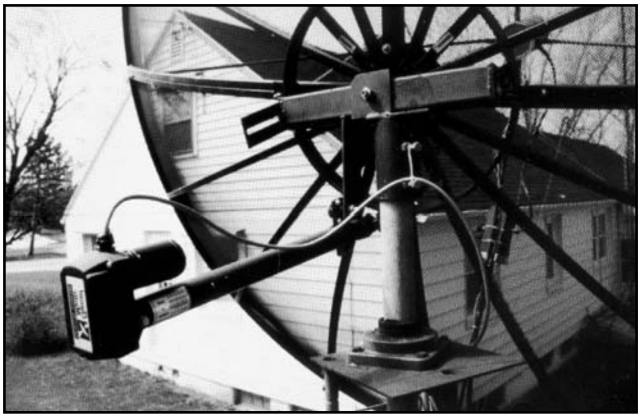

Fig 90—Elevation system, showing modified TVRO mount.

surplus market there are available Digital Volt Meters (DVMs) using LED or LCD displays that can do this job nicely, and that have more precision than is probably necessary for a dish (or Yagi array) of small size. As mentioned earlier, a multi-turn potentiometer on the elevation-drive mechanism can be used to readout elevation, and the same technique can be used for azimuth readout—a potentiometer coupled to the main rotating shaft that turns the antenna.

When using a pot for readout, the most important thing to know is how many degrees of antenna position change occur (in Az or El) for each turn of the pot. This then can be used to calibrate a voltmeter to read volts directly as degrees—for example, 3.60 V could correspond to 360° azimuth (Clockwise North), and 9.0 V could correspond

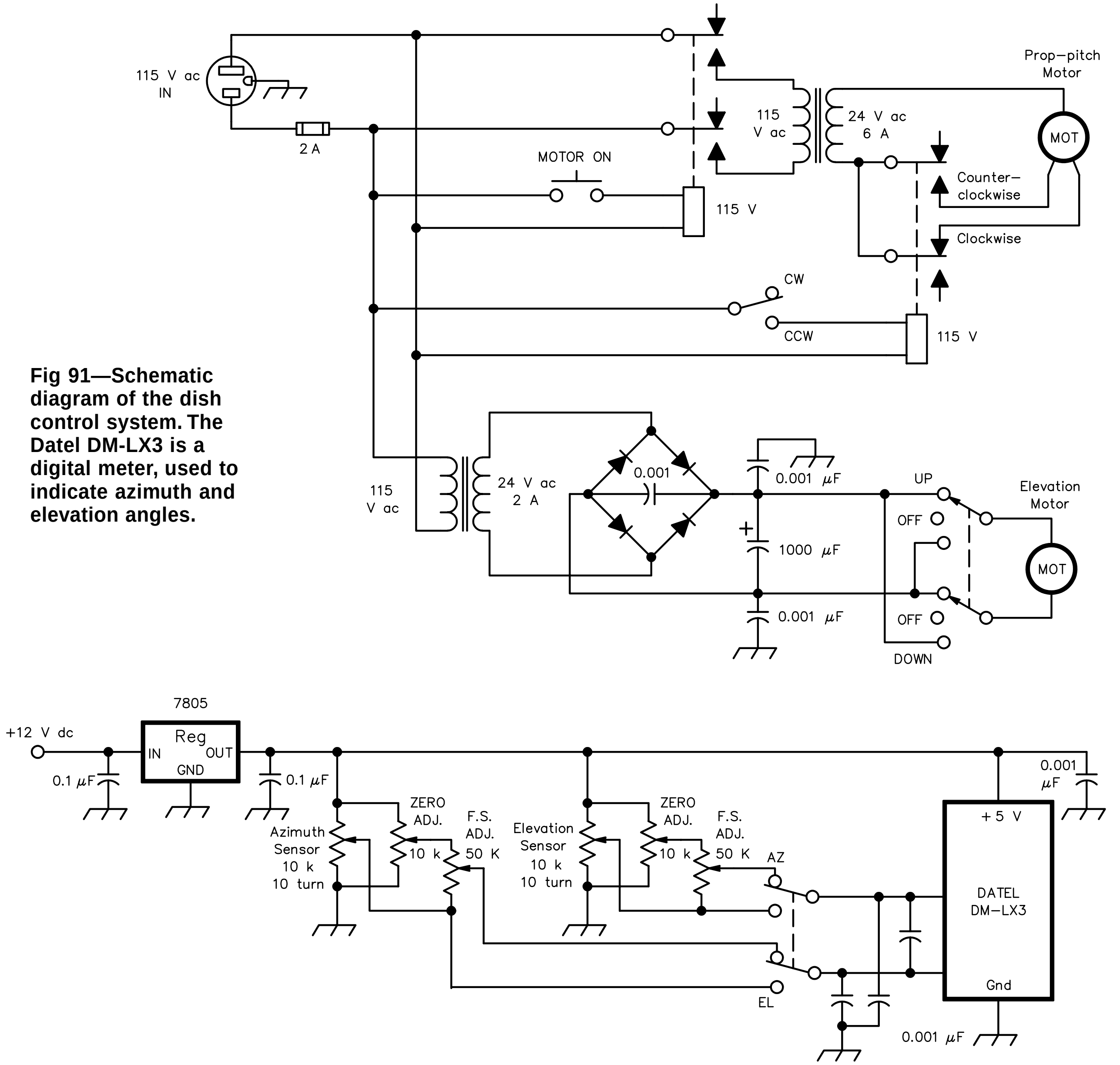
115 V ac
IN
2 A
MOTOR ON
115 V
115
V ac
24 V ac
6 A
Prop-pitch
Motor
MOT
Counter-
clockwise
Clockwise
CW
CCW
115 V
115
V ac
24 V ac
2 A
0.001
0.001 µF
1000 µF
0.001 µF
UP
OFF
OFF
DOWN
Elevation
Motor
MOT
7805
+12 V dc
Reg
IN
OUT
GND
0.1 µF
0.1 µF
ZERO
ADJ.
F.S.
ADJ.
Azimuth
Sensor
10 k
10 turn
10 k
50 K
Elevation
Sensor
10 k
10 turn
ZERO
ADJ.
F.S.
ADJ.
10 k
50 K
AZ
EL
0.001
µF
+5 V
DATEL
DM-LX3
Gnd
0.001 µF

Fig 91—Schematic diagram of the dish control system. The Datel DM-LX3 is a digital meter, used to indicate azimuth and elevation angles.

to 90° elevation (straight up).

A resistance bridge circuit is best used in this application, since it is less sensitive to changes in the supply voltage. The only thing to be careful about is that the DVM must have both the positive (high) and negative (low) inputs isolated from ground (assuming the power supply used to power the DVM is grounded). You could also use a pair of small, cheap Digital Multi-Meters (DMMs), which can sometimes be found for under $10. Because they are battery powered, the isolation issue just discussed is eliminated.

Please see **Fig 91** for a complete schematic of the azimuth, elevation and readout electronics for this antenna-drive system. Also note that while this discussion is geared towards the use of a small dish, the same positioning and readout systems could be used in a Yagi array for 2 meters or 70 cm.

Now that we know where the dish is pointed, how do we know where the Moon is? There are several software programs available to the Amateur for tracking celestial bodies such as the Moon, the Sun, certain stars (usable as noise sources), and even Amateur Satellites. Programs by W9IP, VK3UM, F1EHN and others can be obtained very reasonably and these work well to provide highly accurate position information for tracking.

Feeding the Surplus TVRO Dish

An area that needs particular attention when attempting EME with a small dish is an efficient feed system. An efficient feed system can be a real challenge with TVRO dishes, because many are "deep"—that is, their f/D (focal length to Diameter ratio) is small.

The satellite TV industry used deep dishes because they tend to be quieter, picking up less Earth noise due to spillover effects. A deep dish has a short focal length, and therefore, the feed is relatively close to the surface of the dish. To properly illuminate the reflector out to its edges, a feed horn of relatively wide beamwidth must be used. The feeds designed several years ago by Barry Malowanchuk, VE4MA, are intended for use with just such dishes, and have the advantage of being adjustable to optimize their pattern to the dish in use.

The feed that was used with this dish was modeled after VE4MA's 1296-MHz feed, and a version was even scaled for use at 2304 MHz that worked as well as the original. See **Fig 92** and also see the Notes and References section at the end of this chapter. (Also see the earlier section in the satellite portion of this chapter describing patch feeds for small dishes.)

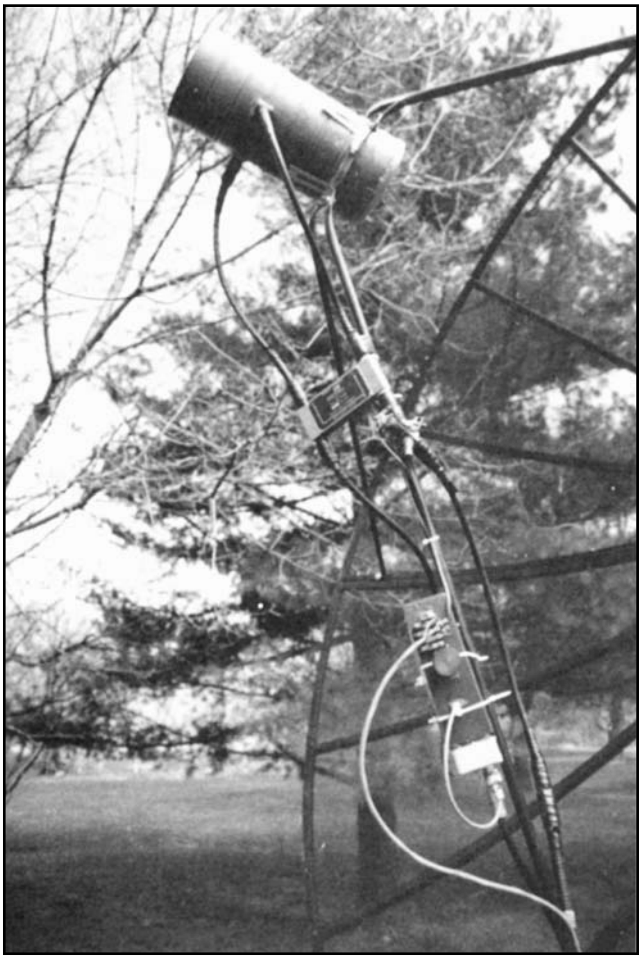

Fig 92—View of feed, showing coffee-can feed horn and hybrid coupler.

SHF EME CHALLENGES

The challenges met when successfully building a station for EME at 900 MHz to 5.7 GHz only become more significant on the SHF bands at 10 GHz and above. Absolute attention to detail is the primary requirement, and this extends to every aspect of the EME antenna system. The dish surface is probably the most difficult problem to solve. As was discussed earlier in this chapter, the shape and accuracy of the reflector contribute directly to the overall gain of the antenna.

But where slight errors in construction can be tolerated at the lower frequencies, the same cannot be said at millimetric wavelengths. Those who have attempted EME on 10 and 24 GHz have discovered that the weight of the dish reflector itself will distort its shape enough to lower the gain to the point where echoes are degraded. Stiffening structures at the back of such dishes are often found necessary. **Fig 93** illustrates the back struts added by Al Ward, W5LUA, to strengthen his dish.

Pointing accuracy is also paramount—a 16-foot dish at 10 GHz has a beamwidth about equal to the diameter of the Moon—0.5°. This means that the echo degradation due to the Moon's movement away from where the dish is pointed is almost immediate, and autotracking systems become more of a necessity than a luxury. At these frequencies, most amateurs actually peak their antennas on

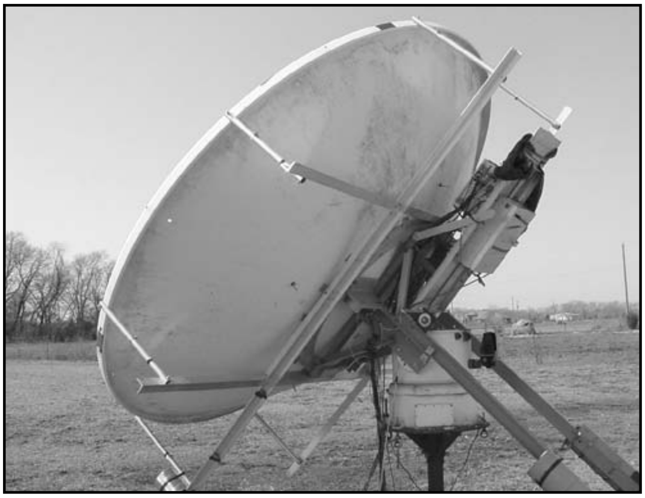

Fig 93—Strengthening struts W5LUA added to the back of his dish to hold down distortion. (*Photo courtesy W5LUA.*)

Moon Noise—the black-body Radiation from the Moon that becomes the dominant source of noise in space.

At these frequencies, the elevation of the Moon above the horizon also plays a role in the ability to communicate, since tropospheric absorption due to water vapor is greatest at low elevation angles (the signal must pass through a greater portion of the troposphere than when the Moon is highly elevated). It is beyond the abilities of most Amateurs to construct their own dishes for these frequencies, so surplus dishes for Ku-band satellite TV (typically 3 meters in diameter) are usually employed, as have high-performance dishes designed for millimetric radar and point-to-point communications at 23 and 38 GHz.

NOTES AND REFERENCES, EME ANTENNAS

W. Allen, "A Mode J Helix," *AMSAT Newsletter*, Jun 1979, pp 30-31.

D. J. Angelakos and D. Kajfez, "Modifications on the Axial-Mode Helical Antenna," *Proc. IEEE*, Apr 1967, pp 558-559.

A. L. Bridges, "Really Zap OSCAR With this Helical Antenna," *73*, in three parts, Jul, Aug and Sep 1975.

K. R. Carver, "The Helicone—A Circularly Polarized Antenna with Low Sidelobe Level," *Proc. IEEE*, Apr 1967, p 559.

M. R. Davidoff, "A Simple 146-MHz Antenna for Oscar Ground Stations," *QST*, Sep 1974, pp 11-13.

M. R. Davidoff, *The Radio Amateur's Satellite Handbook* (Newington: ARRL, 1998-2000).

D. DeMaw, "The Basic Helical Beam," *QST*, Nov 1965, pp 20-25, 170.

D. Evans and G. Jessop, *VHF-UHF Manual, 3rd ed.* (London: RSGB), 1976.

N. Foot, "WA9HUV 12-Foot Dish for 432 and 1296 MHz," The World Above 50 Mc., *QST*, Jun 1971, pp 98-101, 107.

N. Foot, "Cylindrical Feed horn for Parabolic Reflectors," *Ham Radio*, May 1976.

O. J. Glasser and J. D. Kraus, "Measured Impedances of Helical Beam Antennas," *Journal of Applied Physics*, Feb 1948, pp 193-197.

D. Hallidy, "Microwave EME Using A Ten-Foot TVRO Antenna," *The ARRL UHF/Microwave Projects Manual*, 1994 pp 10-9 to 10-13.

G. R. Isely and W. G. Smith, "A Helical Antenna for Space Shuttle Communications," *QST*, Dec 1984, pp 14-18.

H. Jasik, *Antenna Engineering Handbook*, 1st ed. (New York: McGraw-Hill, 1961).

R. T. Knadle, "A Twelve-Foot Stressed Parabolic Dish," *QST*, Aug 1972, pp 16-22.

J. D. Kraus, "Helical Beam Antenna," *Electronics*, Apr 1947, pp 109-111.

J. D. Kraus and J. C. Williamson, "Characteristics of Helical Antennas Radiating in the Axial Mode," *Journal of Applied Physics*, Jan 1948, pp 87-96.

J. D. Kraus, "Helical Beam Antenna for Wide Band Applications," *Proc of the IRE*, Oct 1948, pp 1236-1242.

J. D. Kraus, *Antennas* (New York: McGraw-Hill Book Co., 1950).

J. D. Kraus, "A 50-Ohm Input Impedance for Helical Beam Antenna," *IEEE Transactions on Antennas and Propagation*, Nov 1977, p 913.

J. D. Kraus, *Big Ear* (Powell, OH: Cygnus-Quasar Books, 1976).

T. S. M. MacLean, "Measurements on High-Gain Helical Aerial and on Helicals of Triangular Section," *Proc. of IEE* (London), Jul 1964, pp 1267-1270.

D. M. Mallozzi, "The Tailored Helical," *AMSAT Newsletter*, Mar 1978, pp 8-9.

B. Malowanchuk, "Use of Small TVRO Dishes for EME," *Proceedings of the 21st Conference of The Central States VHF Society*, 1987 pp 68-77.

B. Malowanchuk, "Selection of An Optimum Dish Feed," *Proceedings of the 23rd Conference of The Central States VHF Society*, 1989 pp 35-43.

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington, CT: ARRL, 1990) [out of print].

T. Moreno, *Microwave Transmission Design Data* (New York: McGraw-Hill, 1948).

K. Nose, "Crossed Yagi Antennas for Circular Polarization," *QST*, Jan 1973, pp 11-12.

K. Nose, "A Simple Az-El Antenna System for Oscar," *QST*, Jun 1973, pp 11-12.

S. Powlishen, K1FO, "432-MHz EME 1990s Style," Parts 1 and 2, *Communications Quarterly*, Oct 1990, pp 29-39; Oct 1991, pp 33-48.

S. Powlishen, K1FO, "Rear-Mount Yagi Arrays for 432-MHz EME: Solving the EME Polarization Problem," *The ARRL Antenna Compendium, Vol 3* (Newington:

ARRL, 1992), pp 79-98.

C. Richards, "The 10 Turn Chopstick Helical (Mk2) for OSCAR 10 432 MHz Uplink," *Radio Communication*, Oct 1984, pp 844-845.

S. Sander and D. K. Cheng, "Phase Center of Helical Beam Antennas," *IRE National Convention Record Part 6*, 1958, pp 152-157.

E. A. Scott and H. E. Banta, "Using the Helical Antenna at 1215 Mc.," *QST*, Jul 1962, pp 14-16.

R. Soifer, W2RS, "More 'QRP' EME on 144 MHz," *QST*, Oct 1990, pp 36-38.

G. Southworth, *Principles and Applications of Waveguide Transmission* (New York: D. Van Nostrand Co., 1950).

B. Sykes, "Circular Polarization and Crossed-Yagi Antennas," Technical Topics, *Radio Communication*, Feb 1985, p 114.

H. E. Taylor and D. Fowler, "A V-H-F Helical Beam Antenna," *CQ*, Apr 1949, pp 13-16.

G. Tillitson, "The Polarization Diplexer—A Polaplexer," *Ham Radio*, Mar 1977.

P. P. Viezbicke, "Yagi Antenna Design," *NBS Technical Note 688* (U. S. Dept. of Commerce/National Bureau of Standards, Boulder, CO), Dec 1976.

D. Vilardi, "Easily Constructed Antennas for 1296 MHz," *QST*, Jun 1969.

D. Vilardi, "Simple and Efficient Feed for Parabolic Antennas," *QST*, Mar 1973.

J. L. Wong and H. E. King, "Broadband Quasi-Tapered Helical Antenna," *IEEE Transactions on Antennas and Propagation*, Jan 1979, pp 72-78.

Chapter 20

Antenna Materials and Accessories

This chapter contains information on materials amateurs use to construct antennas—what types of material to look for in a particular application, tips on working with and using various materials. Chapter 21 contains information on where to purchase these materials.

Basically, antennas for MF, HF, VHF and the lower UHF range consist simply of one or more conductors that radiate (or receive) electromagnetic waves. However, an antenna system must also include some means to support those conductors and maintain their relative positions—the boom for a Yagi antenna and the halyards for a wire dipole, for example. In this chapter we'll look at materials for those applications, too. Structural supports such as towers, masts, poles, etc. are discussed in Chapter 22.

There are two main types of material used for antenna conductors, wire and tubing. Wire antennas are generally simple and therefore easier to construct, although some arrays of wire elements can become rather complex. When tubing is required, aluminum tubing is used most often because of its light weight. Aluminum tubing is discussed in a subsequent section of this chapter.

Wire Antennas

Although wire antennas are relatively simple, they can constitute a potential hazard unless properly constructed. Antennas should never be run under or over public utility (telephone or power) lines. Several amateurs have lost their lives by failing to observe this precaution.

The National Electric Code® of the National Fire Protection Association contains a section on amateur stations in which a number of recommendations are made concerning minimum size of antenna wire and the manner of bringing the transmission line into the station. Chapter 1 contains more information about this code. The code in itself does not have the force of law, but it is frequently made a part of local building regulations, which are enforceable. The provisions of the code may also be written into, or referred to, in fire and liability insurance documents.

The RF resistance of copper wire increases as the size of the wire decreases. However, in most types of antennas that are commonly constructed of wire (even quite small wire), the radiation resistance will be much higher than the RF resistance, and the efficiency of the antenna will still be adequate. Wire sizes as small as #30, or even smaller, have been used quite successfully in the construction of "invisible" antennas in areas where more conventional antennas cannot be erected. In most cases, the selection of wire for an antenna will be based primarily on the physical properties of the wire, since the suspension of wire from elevated supports places a strain on the wire.

WIRE TYPES

Wire having an enamel coating is preferable to bare wire, since the coating resists oxidation and corrosion. Several types of wire having this type of coating are available, depending on the strength needed. "Soft-drawn" or annealed copper wire is easiest to handle; unfortunately, it stretches considerably under stress. Soft-drawn wire should be avoided, except for applications where the wire will be under little or no tension, or where some change in length can be tolerated. (For example, the length of a horizontal antenna fed at the center with open-wire line is not critical, although a change in length may require some readjustment of coupling to the transmitter.)

"Hard-drawn" copper wire or copper-clad steel wire (also known as Copperweld™) is harder to handle, because it has a tendency to spiral when it is unrolled. These types of wire are ideal for applications where significant stretch cannot be tolerated. Care should be exercised in using this wire to make sure that kinks do not develop—the wire will have a far greater tendency to break at a kink. After the coil has been unwound, suspend the wire a few feet above ground for a day or two before using it. The wire should not be recoiled before it is installed.

Several factors influence the choice of wire type and size. Most important to consider are the length of the unsupported span, the amount of sag that can be tolerated, the stability of the supports under wind pressure, and whether or not an unsupported transmission line is to be suspended from the span. **Table 1** shows the wire diam, current-carrying capacity and resistance of various sizes of copper wire. **Table 2** shows the maximum rated working tensions of hard-drawn and copper-clad steel wire of various sizes. These two tables can be used to select the appropriate wire size for an antenna.

Table 1
Copper-Wire Table

Wire Size AWG (B&S)	*Dia in Mils*[1]	*Dia in mm*	*Turns per Linear Inch Enamel*	*Feet per Pound Bare*	*Ohms per 1000 ft 25°C*	*Cont.-duty current*[2] *Single Wire in Open Air*
1	289.3	7.348	—	3.947	0.1264	—
2	257.6	6.544	—	4.977	0.1593	—
3	229.4	5.827	—	6.276	0.2009	—
4	204.3	5.189	—	7.914	0.2533	—
5	181.9	4.621	—	9.980	0.3195	—
6	162.0	4.115	—	12.58	0.4028	—
7	144.3	3.665	—	15.87	0.5080	—
8	128.5	3.264	7.6	20.01	0.6405	73
9	114.4	2.906	8.6	25.23	0.8077	—
10	101.9	2.588	9.6	31.82	1.018	55
11	90.7	2.305	10.7	40.12	1.284	—
12	80.8	2.053	12.0	50.59	1.619	41
13	72.0	1.828	13.5	63.80	2.042	—
14	64.1	1.628	15.0	80.44	2.575	32
15	57.1	1.450	16.8	101.4	3.247	—
16	50.8	1.291	18.9	127.9	4.094	22
17	45.3	1.150	21.2	161.3	5.163	—
18	40.3	1.024	23.6	203.4	6.510	16
19	35.9	0.912	26.4	256.5	8.210	—
20	32.0	0.812	29.4	323.4	10.35	11
21	28.5	0.723	33.1	407.8	13.05	—
22	25.3	0.644	37.0	514.2	16.46	—
23	22.6	0.573	41.3	648.4	20.76	—
24	20.1	0.511	46.3	817.7	26.17	—
25	17.9	0.455	51.7	1031	33.00	—
26	15.9	0.405	58.0	1300	41.62	—
27	14.2	0.361	64.9	1639	52.48	—
28	12.6	0.321	72.7	2067	66.17	—
29	11.3	0.286	81.6	2607	83.44	—
30	10.0	0.255	90.5	3287	105.2	—
31	8.9	0.227	101	4145	132.7	—
32	8.0	0.202	113	5227	167.3	—
33	7.1	0.180	127	6591	211.0	—
34	6.3	0.160	143	8310	266.0	—
35	5.6	0.143	158	10480	335	—
36	5.0	0.127	175	13210	423	—
37	4.5	0.113	198	16660	533	—
38	4.0	0.101	224	21010	673	—
39	3.5	0.090	248	26500	848	—
40	3.1	0.080	282	33410	1070	—

[1] A mil is 0.001 inch.

[2] Max wire temp of 212° F and max ambient temp of 135° F.

Table 2
Stressed Antenna Wire

American Wire Gauge	***Recommended Tension[1] (pounds)***		***Weight (pounds per 1000 feet)***	
	Copper-clad steel[2]	*Hard-drawn copper*	*Copper-clad steel[2]*	*Hard-drawn copper*
4	495	214	115.8	126.0
6	310	130	72.9	79.5
8	195	84	45.5	50.0
10	120	52	28.8	31.4
12	75	32	18.1	19.8
14	50	20	11.4	12.4
16	31	13	7.1	7.8
18	19	8	4.5	4.9
20	12	5	2.8	3.1

[1] Approximately one-tenth the breaking load. Might be increased 50% if end supports are firm and there is no danger of ice loading.
[2] Copperweld,™ 40% copper.

Wire Tension

If the tension on a wire can be adjusted to a known value, the expected sag of the wire (**Fig 1**) may be determined before installation using Table 2 and the nomograph of **Fig 2**. Even though there may be no convenient method to determine the tension in pounds, calculation of the expected sag for practicable working tensions is often desirable. If the calculated sag is greater than allowable it may be reduced by any one or a combination of the following:

1) Providing additional supports, thereby decreasing the span
2) Increasing the tension in the wire if less than recommended
3) Decreasing the size of the wire

Instructions for Using the Nomograph

1) From Table 2, find the weight (pounds/1000 feet) for the particular wire size and material to be used.
2) Draw a line from the value obtained above, plotted on the weight axis, to the desired span (feet) on the span axis, Fig 2. Note in Fig 1 that the span is one half the distance between the supports.
3) Choose an operating tension level (in pounds) consistent with the values presented in Table 2 (preferably less than the recommended wire tension).
4) Draw a line from the tension value chosen (plotted on the tension axis) through the point where the work axis crosses the original line constructed in step 2, and continue this new line to the sag axis.
5) Read the sag in feet on the sag axis.

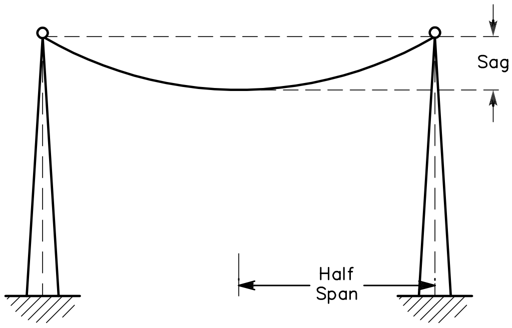

Fig 1—The half span and sag of a long-wire antenna.

Example:

Weight = 11 pounds/1000 feet
Span = 210 feet
Tension = 50 pounds
Answer: Sag = 4.7 feet

These calculations do not take into account the weight of a feed line supported by the antenna wire.

Wire Splicing

Wire antennas should preferably be made with unbroken lengths of wire. In instances where this is not feasible, wire sections should be spliced as shown in **Fig 3**. The enamel insulation should be removed for a distance of about 6 inches from the end of each section by scraping with a knife or rubbing with sandpaper until the copper underneath is bright. The turns of wire should be brought up tight around the standing part of the wire by twisting with broad-nose pliers.

The crevices formed by the wire should be completely filled with rosin-core solder. An ordinary soldering iron or gun may not provide sufficient heat to melt solder outdoors; a propane torch is desirable. The joint should be heated sufficiently so the solder flows freely into the joint when the source of heat is removed momentarily. After the joint has cooled completely, it should be wiped clean with a cloth, and then sprayed generously with acrylic to prevent corrosion.

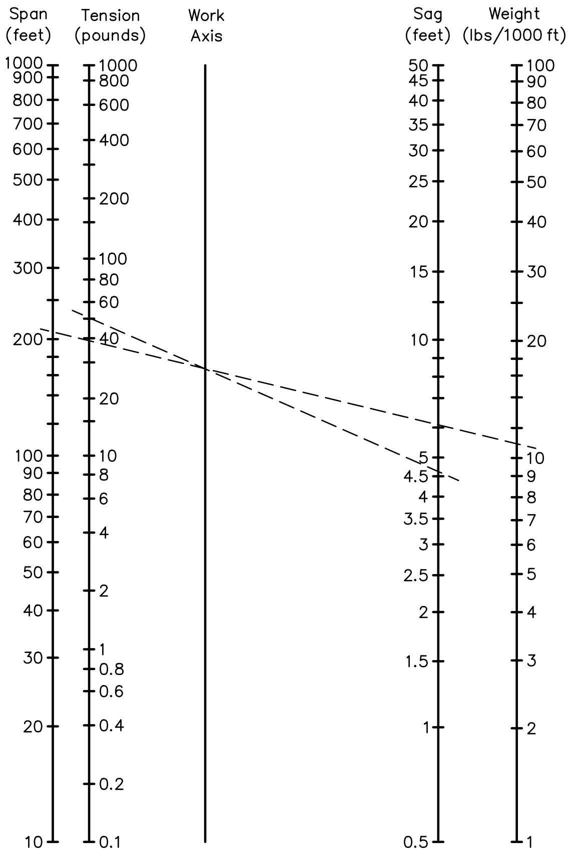

Fig 2—Nomograph for determining wire sag. (John Elengo, Jr, K1AFR)

ANTENNA INSULATION

To prevent loss of RF power, the antenna should be well insulated from ground, unless of course it is a shunt-fed system. This is particularly important at the outer end or ends of wire antennas, since these points are always at a comparatively high RF potential. If an antenna is to be installed indoors (in an attic, for instance) the antenna may be suspended directly from the wood rafters without additional insulation, if the wood is permanently dry. Much greater care should be given to the selection of proper insulators when the antenna is located outside where it is exposed to wet weather.

Insulator Leakage

Antenna insulators should be made of material that will not absorb moisture. The best insulators for antenna use are made of glass or glazed porcelain. Depending on the type of material, plastic insulators may be suitable. The length of an insulator relative to its surface area is indicative of its comparative insulating ability. A long thin insulator will have less leakage than a short thick insulator. Some antenna insulators are deeply ribbed to increase the surface leakage path without increasing the physical length of the insulator. Shorter insulators can be used at low-potential points, such as at the center of a dipole. If such an antenna is to be fed with open-wire line and used on several bands, however, the center insulator should be the same as those used at the ends, because high RF potential may exist across the center insulator on some bands.

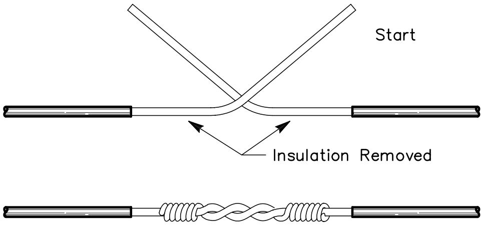

Fig 3—Correct method of splicing antenna wire. Solder should be flowed into the wraps after the connection is completed. After cooling, the joint should be sprayed with acrylic to prevent oxidation and corrosion.

Insulator Stress

As with the antenna wire, the insulator must have sufficient physical strength to support the stress of the antenna without danger of breakage. Long elastic bands or lengths of nylon fishing line provide long leakage paths and make satisfactory insulators within their limits to resist mechanical strain. They are often used in antennas of the "invisible" type mentioned earlier.

For low-power work with short antennas not subject to appreciable stress, almost any small glass or glazed-porcelain insulator will do. Homemade insulators of Lucite rod or sheet will also be satisfactory. More care is required in the selection of insulators for longer spans and higher transmitter power.

For a given material, the breaking tension of an insulator will be proportional to its cross-sectional area. It should be remembered, however, that the wire hole at the end of the insulator decreases the effective cross-sectional area. For this reason, insulators designed to carry heavy strains are fitted with heavy metal end caps, the eyes being formed in the metal cap, rather than in the insulating material itself. The following stress ratings of antenna insulators are typical:

$^{5}/_{8}$ in. square by 4 in. long—400 lb
1 in. diameter by 7 or 12 in. long—800 lb
$1^{1}/_{2}$ in. diameter by 8, 12 or 20 in. long, with special metal end caps—5000 lb

These are rated breaking tensions. The actual working tensions should be limited to not more than 25% of the breaking rating.

The antenna wire should be attached to the insulators as shown in **Fig 4**. Care should be taken to avoid sharp angular bends in the wire when it is looped through the insulator eye. The loop should be generous enough in size that it will not bind the end of the insulator tightly. If the length of the antenna is critical, the length should be measured to the outward end of the loop, where it passes through the eye of the insulator. The soldering should be done as described earlier for the wire splice.

Strain Insulators

Strain insulators have their holes at right angles, since they are designed to be connected as shown in **Fig 5**. It can be seen that this arrangement places the insulating material under compression, rather than tension. An insulator connected this way can withstand much greater stress. Furthermore, the wire will not collapse if the insulator breaks, since the two wire loops are inter-locked. Because the wire is wrapped around the insulator, however, the leakage path is reduced drastically, and the capacitance between the wire loops provides an additional leakage path. For this reason, the use of the strain insulator is usually confined to such applications as breaking up resonances in guy wires, where high levels of stress prevail, and where the RF insulation is of less importance. Such insulators might be suitable for use at low-potential points on an antenna, such as at the center of a dipole. These insulators may also be fastened in the conventional manner if the wire will not be under sufficient tension to break out the eyes.

Insulators for Ribbon-Line Antennas

Fig 6A shows the sketch of an insulator designed to be used at the ends of a folded dipole or a multiple dipole made of ribbon line. It should be made approximately as shown, out of Lucite or bakelite material about 1/4 inch thick. The advantage of this arrangement is that the strain of the antenna is shared by the conductors and the plastic webbing of the ribbon, which adds considerable strength. After soldering, the screw should be sprayed with acrylic.

Fig 6B shows a similar arrangement for suspending one dipole from another in a stagger-tuned dipole system. If better insulation is desired, these insulators can be wired to a conventional insulator.

PULLEYS AND HALYARDS

Pulleys and halyards commonly used to raise and lower a wire antenna must also be capable of taking the

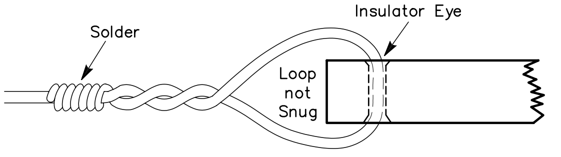

Fig 4—When fastening antenna wire to an insulator, do not make the wire loop too snug. After the connection is complete, flow solder into the turns. Then when the joint has cooled completely, spray it with acrylic.

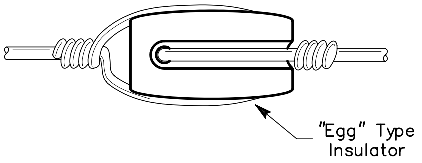

Fig 5—Conventional manner of fastening wire to a strain insulator. This method decreases the leakage path and increases capacitance, as discussed in the text.

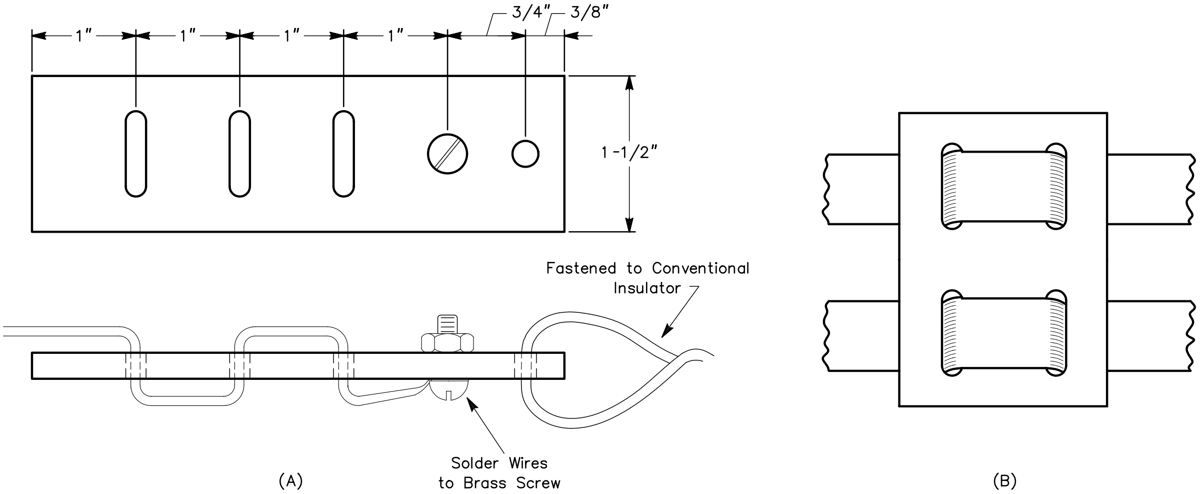

Fig 6—At A, an insulator for the ends of folded dipoles, or multiple dipoles made of 300-ohm ribbon. At B, a method of suspending one ribbon dipole from another in a multiband dipole system.

same strain as the antenna wire and insulators. Unfortunately, little specific information on the stress ratings of most pulleys is available. Several types of pulleys are readily available at almost any hardware store. Among these are small galvanized pulleys designed for awnings and several styles and sizes of clothesline pulleys. Heavier and stronger pulleys are those used in marine work. The factors that determine how much stress a pulley will handle include the diameter of the shaft, how securely the shaft is fitted into the sheath and the size and material of the frame.

Another important factor to be considered in the selection of a pulley is its ability to resist corrosion. Galvanized awning pulleys are probably the most susceptible to corrosion. While the frame or sheath usually stands up well, these pulleys usually fail at the shaft. The shaft rusts out, allowing the grooved wheel to break away under tension.

Most good-quality clothesline pulleys are made of alloys which do not corrode readily. Since they are designed to carry at least 50 feet of line loaded with wet clothing in stiff winds, they should be adequate for normal spans of 100 to 150 feet between stable supports. One type of clothesline pulley has a 4-inch diameter plastic wheel with a ¼-inch shaft running in bronze bearings. The sheath is made of cast or forged corrosion-proof alloy. Some look-alike low-cost pulleys of this type have an aluminum shaft with no bearings. For antenna work, these cheap pulleys are of little long-term value.

Marine pulleys have good weather-resisting qualities, since they are usually made of bronze, but they are comparatively expensive and are not designed to carry heavy loads. For extremely long spans, the wood-sheathed pulleys used in "block and tackle" devices and for sail hoisting should work well.

Halyards

Table 3 shows the recommended maximum tensions for various sizes and types of line and rope suitable for hoisting halyards. Probably the best type for general amateur use for spans up to 150 or 200 feet is ¼-inch nylon rope. Nylon is somewhat more expensive than ordinary rope of the same size, but it weathers much better. Nylon also has a certain amount of elasticity to accommodate gusts of wind, and is particularly recommended for antennas using trees as supports. A disadvantage of new nylon rope is that it stretches by a significant percentage. After an installation with new rope, it will be necessary to repeatedly take up the slack created by stretching. This process will continue over a period of several weeks, at which time most of the stretching will have taken place. Even a year after installation, however, some slack may still arise from stretching.

Most types of synthetic rope are slippery, and some types of knots ordinarily used for rope will not hold well. **Fig 7** shows a knot that should hold well, even in nylon rope or plastic line.

For exceptionally long spans, stranded galvanized steel sash cord makes a suitable support. Cable advertised as "wire rope" usually does not weather well. A boat winch, sold at marinas and at Sears, is a great convenience in antenna hoisting (and usually a necessity with metal halyards).

Table 3
Approximate Safe Working Tension for Various Halyard Materials

Material	*Dia, In.*	*Tension, Lb*
Manila hemp rope	1/4	120
	3/8	270
	1/2	530
	5/8	800
Polypropylene rope	1/4	270
	3/8	530
	1/2	840
Nylon rope	1/4	300
	3/8	660
	1/2	1140
7×11 galvanized sash cord	1/16	30
	1/8	125
	3/16	250
	1/4	450
High-strength stranded galvanized steel guy wire	1/8	400
	3/16	700
	1/4	1200
Rayon-filled plastic clothesline	7/32	60 to 70

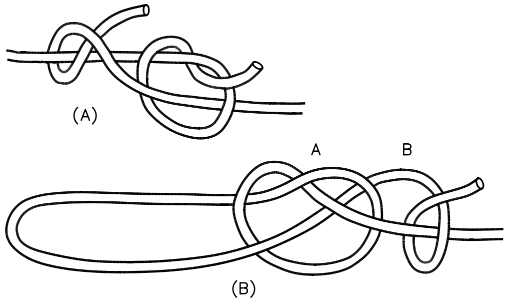

Fig 7—This is one type of knot that will hold with smooth rope, such as nylon. Shown at A, the knot for splicing two ends. B shows the use of a similar knot in forming a loop, as might be needed for attaching an insulator to a halyard. Knot A is first formed loosely 10 or 12 in. from the end of the rope; then the end is passed through the eye of the insulator and knot A. Knot B is then formed and both knots pulled tight. (Richard Carruthers, K7HDB)

Antennas of Aluminum Tubing

Aluminum is a malleable, ductile metal with a mass density of 2.70 grams per cubic centimeter. The density of aluminum is approximately 35% that of iron and 30% that of copper. Aluminum can be polished to a high brightness, and it will retain this polish in dry air. In the presence of moisture, aluminum forms an oxide coating (Al_2O_3) that protects the metal from further corrosion. Direct contact with certain metals, however (especially ferrous metals such as iron or steel), in an outdoor environment can bring about galvanic corrosion of aluminum and its alloys. Some protective coating should be applied to any point of contact between two dissimilar metals. Much of this information about aluminum and aluminum tubing was prepared by Ralph Shaw, K5CAV.

Aluminum is non-toxic; it is used in cooking utensils and to hold and cover "TV dinners" and other frozen foods, so it is certainly safe to work with. The ease with which it can be drilled or sawed makes it a pleasure to work with. Aluminum products lend themselves to many and varied applications.

Aluminum alloys can be used to build amateur antennas, as well as for towers and supports. Light weight and high conductivity make aluminum ideal for these applications. Alloying lowers the conductivity ratings, but the tensile strength can be increased by alloying aluminum with one or more metals such as manganese, silicon, copper, magnesium or zinc. Cold rolling can be employed to further increase the strength.

A four-digit system is used to identify aluminum alloys, such as 6061. Aluminum alloys starting with a 6 contain di-magnesium silicide (Mg_2Si). The second digit indicates modifications of the original alloy or impurity limits. The last two digits designate different aluminum alloys within the category indicated by the first digit.

In the 6000 series, the 6061 and 6063 alloys are a commonly used for antenna applications. Both types have good resistance to corrosion and medium strength. A further designation like T-6 denotes thermal treatment (heat tempering). More information on the available aluminum alloys can be found in **Table 4**.

SELECTING ALUMINUM TUBING

Table 5 shows the standard sizes of aluminum tubing that are stocked by most aluminum suppliers or distributors in the United States and Canada. Note that all tubing comes in 12-foot lengths (local hardware stores sometimes stock 6- and 8-foot lengths) and larger-diameter sizes may be available in lengths up to 24 feet. Note also that any diameter tubing will fit snugly into the next larger size, if the larger size has a 0.058-inch wall thickness. For example, $^{5}/_{8}$-inch tubing has an outside diameter of 0.625 inch. This will fit into $^{3}/_{4}$-inch tubing with a 0.058-inch wall, which has an inside diameter of 0.634 inch. A clearance of 0.009 inch is just right for a slip fit or for slotting the tubing and then using hose clamps. Always get the next larger size and specify a 0.058-inch wall to obtain the 0.009-inch clearance.

Table 4
Aluminum Numbers for Amateur Use

	Common Alloy Numbers
Type	*Characteristic*
2024	Good formability, high strength
5052	Excellent surface finish, excellent corrosion resistance, normally not heat treatable for high strength
6061	Good machinability, good weldability
6063	Good machinability, good weldability
7075	Good formability, high strength
	Common Tempers
Type	*Characteristics*
T0	Special soft condition
T3	Hard
T6	Hardest, possibly brittle
TXXX	Three digit tempers—usually specialized high strength heat treatments, similar to T6
	General Uses
Type	*Uses*
2024-T3	Chassis boxes, antennas, anything that will be bent or
7075-T3	Flexed repeatedly
6061-T6	Tubing and pipe; angle channel and bar stock
6063-T832	Tubing and pipe; angle channel and bar stock

A little figuring with Table 5 will give you all the information you need to build a beam, including what the antenna will weigh. The 6061-T6 type of aluminum has a relatively high strength and has good workability. It is highly resistant to corrosion and will bend without taking a "set."

SOURCES FOR ALUMINUM

Aluminum can be purchased new, and suppliers are listed in Chapter 21. But don't overlook the local metal scrap yard. The price varies, but between 35 and 60 cents per pound is typical for scrap aluminum. Some aluminum items to look for include aluminum vaulting poles, tent poles, tubing and fittings from scrapped citizen's band antennas, and aluminum angle stock. The scrap yard may even have a section or two of triangular aluminum tower.

Aluminum vaulting poles are 12 or 14 feet long and range in diameter from 1$^{1}/_{2}$ to 1$^{3}/_{4}$ inches. These poles are suitable for the center-element sections of large 14-MHz beams or as booms for smaller antennas. Tent poles range in length from 2½ to 4 feet. The tent poles are

Table 5

Aluminum Tubing Sizes

6061-T6 (61S-T6) Round Aluminum Tube In 12-Foot Lengths

	Wall Thickness			Approximate Weight	
Tubing Diameter	*Inches*	*Stubs Ga.*	*ID, Inches*	*Pounds Per Foot*	*Pounds Per Length*
3/16 in.	0.035	(#20)	0.117	0.019	0.228
	0.049	(#18)	0.089	0.025	0.330
1/4 in.	0.035	(#20)	0.180	0.027	0.324
	0.049	(#18)	0.152	0.036	0.432
	0.058	(#17)	0.134	0.041	0.492
5/16 in.	0.035	(#20)	0.242	0.036	0.432
	0.049	(#18)	0.214	0.047	0.564
	0.058	(#17)	0.196	0.055	0.660
3/8 in.	0.035	(#20)	0.305	0.043	0.516
	0.049	(#18)	0.277	0.060	0.720
	0.058	(#17)	0.259	0.068	0.816
	0.065	(#16)	0.245	0.074	0.888
7/16 in.	0.035	(#20)	0.367	0.051	0.612
	0.049	(#18)	0.339	0.070	0.840
	0.065	(#16)	0.307	0.089	1.068
1/2 in.	0.028	(#22)	0.444	0.049	0.588
	0.035	(#20)	0.430	0.059	0.708
	0.049	(#18)	0.402	0.082	0.984
	0.058	(#17)	0.384	0.095	1.040
	0.065	(#16)	0.370	0.107	1.284
5/8 in.	0.028	(#22)	0.569	0.061	0.732
	0.035	(#20)	0.555	0.075	0.900
	0.049	(#18)	0.527	0.106	1.272
	0.058	(#17)	0.509	0.121	1.452
	0.065	(#16)	0.495	0.137	1.644
3/4 in.	0.035	(#20)	0.680	0.091	1.092
	0.049	(#18)	0.652	0.125	1.500
	0.058	(#17)	0.634	0.148	1.776
	0.065	(#16)	0.620	0.160	1.920
	0.083	(#14)	0.584	0.204	2.448
7/8 in.	0.035	(#20)	0.805	0.108	1.308
	0.049	(#18)	0.777	0.151	1.810
	0.058	(#17)	0.759	0.175	2.100
	0.065	(#16)	0.745	0.199	2.399
1 in.	0.035	(#20)	0.930	0.123	1.476
	0.049	(#18)	0.902	0.170	2.040
	0.058	(#17)	0.884	0.202	2.424
	0.065	(#16)	0.870	0.220	2.640
	0.083	(#14)	0.834	0.281	3.372
1 1/8 in.	0.035	(#20)	1.055	0.139	1.668
	0.058	(#17)	1.009	0.228	2.736
1 1/4 in.	0.035	(#20)	1.180	0.155	1.860
	0.049	(#18)	1.152	0.210	2.520
	0.058	(#17)	1.134	0.256	3.072
	0.065	(#16)	1.120	0.284	3.408
	0.083	(#14)	1.084	0.357	4.284
1 3/8 in.	0.035	(#20)	1.305	0.173	2.076
	0.058	(#17)	1.259	0.282	3.384
1 1/2 in.	0.035	(#20)	1.430	0.180	2.160
	0.049	(#18)	1.402	0.260	3.120
	0.058	(#17)	1.384	0.309	3.708
	0.065	(#16)	1.370	0.344	4.128
	0.083	(#14)	1.334	0.434	5.208
	*0.125	1/8 in.	1.250	0.630	7.416
	*0.250	1/4 in.	1.000	1.150	14.832
1 5/8 in.	0.035	(#20)	1.555	0.206	2.472
	0.058	(#17)	1.509	0.336	4.032
1¾ in.	0.058	(#17)	1.634	0.363	4.356
	0.083	(#14)	1.584	0.510	6.120
1 7/8 in.	0.058	(#17)	1.759	0.389	4.668
2 in.	0.049	(#18)	1.902	0.350	4.200
	0.065	(#16)	1.870	0.450	5.400
	0.083	(#14)	1.834	0.590	7.080
	*0.125	1/8 in.	1.750	0.870	9.960
	*0.250	1/4 in.	1.500	1.620	19.920
2 1/4 in.	0.049	(#18)	2.152	0.398	4.776
	0.065	(#16)	2.120	0.520	6.240
	0.083	(#14)	2.084	0.660	7.920
2 1/2 in.	0.065	(#16)	2.370	0.587	7.044
	0.083	(#14)	2.334	0.740	8.880
	*0.125	1/8 in.	2.250	1.100	12.720
	*0.250	1/4 in.	2.000	2.080	25.440
3 in.	0.065	(#16)	2.870	0.710	8.520
	*0.125	1/8 in.	2.700	1.330	15.600
	*0.250	¼ in.	2.500	2.540	31.200

*These sizes are extruded. All other sizes are drawn tubes.

usually tapered; they can be split on the larger end and then mated with the smaller end of another pole of the same diameter. A small stainless-steel hose clamp (sometimes also available at scrap yards!) can be used to fasten the poles at this junction. A 14- or 21-MHz element can be constructed from several tent poles in this fashion. If a longer continuous piece of tubing is available, it can be used for the center section to decrease the number of junctions and clamps.

Other aluminum scrap is sometimes available, such as US Army aluminum mast sections designated AB-85/GRA-4 (J&H Smith Mfg). These are 3 foot sections with a 1 5/8 inch diameter. The ends are swaged so they can be assembled one into another. These are ideal for making a portable mast for a 144-MHz beam or for Field Day applications.

CONSTRUCTION WITH ALUMINUM TUBING

Most antennas built for frequencies of 14 MHz and above are made to be rotated. Constructing a rotatable antenna requires materials that are strong, lightweight and easy to obtain. The materials required to build a suitable antenna will vary, depending on many factors. Perhaps the most important factor that determines the type of hardware needed is the weather conditions normally encountered.

High winds usually don't cause as much damage to an antenna as does ice, especially ice along with high winds. Aluminum element and boom sizes should be selected so the various sections of tubing will telescope to provide the necessary total length.

The boom size for a rotatable Yagi or quad should be selected to provide stability to the entire system. The best diameter for the boom depends on several factors; most important are the element weight, number of elements and overall length. Tubing of 1-¼-inch diameter can easily support three-element 28-MHz arrays and perhaps a two-element 21-MHz system. A 2-inch diameter boom will be adequate for larger 28-MHz antennas or for harsh weather conditions, and for antennas up to three elements on 14 MHz or four elements on 21 MHz. It is not recommended that 2-inch diameter booms be made any longer than 24 feet unless additional support is given to reduce both vertical and horizontal bending forces. Suitable rein-forcement for a long 2-inch boom can consist of a truss or a truss and lateral support, as shown in **Fig 8**.

A boom length of 24 feet is about the point where a 3-inch diameter begins to be very worthwhile. This dimension provides a considerable improvement in overall mechanical stability as well as increased clamping surface area for element hardware. Clamping surface area is extremely important if heavy icing is common and rotation of elements around the boom is to be avoided. Pinning an element to the boom with a large bolt helps in this regard. On smaller diameter booms, however, the elements sometimes work loose and tend to elongate the pinning holes in both the element and the boom. After some time the elements shift their positions slightly (sometimes from day to day!) and give a rather ragged appearance to the system, even though this doesn't generally harm the electrical performance.

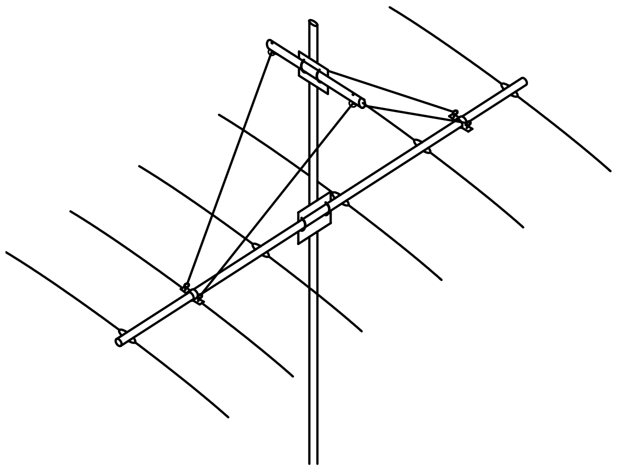

Fig 8—A long boom needs both vertical and horizontal support. The cross bar mounted above the boom can support a double truss to help keep the antenna in position.

A 3-inch diameter boom with a wall thickness of 0.065 inch is satisfactory for antennas up to about a five-element, 14-MHz array that is spaced on a 40-foot long boom. A truss is recommended for any boom longer than 24 feet.

There is no RF voltage at the center of a parasitic element, so no insulation is required in mounting elements that are centered on the boom (driven elements excepted). This is true whether the boom is metal or a nonconducting material. Metal booms have a small "shortening effect" on elements that run through them. With materials sizes commonly employed, this is not more than one percent of the element length, and may not be noticeable in many applications. It is just perceptible with $^1/_2$-inch tubing booms used on 432 MHz, for example. Design-formula lengths can be used as given, if the matching is adjusted in the frequency range one expects to use. The center frequency of an all-metal array will tend to be 0.5 to 1 percent higher than a similar system built of wooden supporting members.

Element Assembly

While the maximum safe length of an antenna element depends to some extent on its diameter, the only laws that specify the minimum diameter of an element are the laws of nature. That is, the element must be rugged enough to survive whatever weather conditions it will encounter.

Fig 9 shows tapered Yagi element designs that will survive winds in excess of 80 mi/h. With a $^1/_4$-inch thickness of radial ice, these designs will withstand winds up to approximately 60 mi/h. (Ice increases the wind area but does not increase the strength of the element.) More rugged designs are shown in **Fig 10**. With no ice loading, these elements will survive in 120-mi/h winds, and in winds exceeding 85 mi/h with $^1/_4$ inch of radial ice. If you lose an antenna made with elements like these, you'll have plenty of company among your neighbors with commercially made antennas!

Figs 9 and 10 show only half elements. When the element is assembled, the largest size tubing for each element should be double the length shown in the drawing, with its center being the point of attachment to the boom. These designs are somewhat conservative, in that they are self-resonant slightly below the frequency indicated for each design. Telescoping the outside end sections to shorter lengths for resonance will increase the survival wind speeds. Conversely, lengthening the outside end sections will reduce the survival wind speeds. [See Bibliography listing for David Leeson (W6NL, ex-W6QHS) at the end of this chapter.]

Fig 11 shows several methods of fastening antenna element sections together. The slot and hose clamp method shown in Fig 11A is probably the best for joints where adjustments are required. Generally, one adjustable joint per element half is sufficient to tune the antenna. Stainless-steel hose clamps (beware—some "stainless

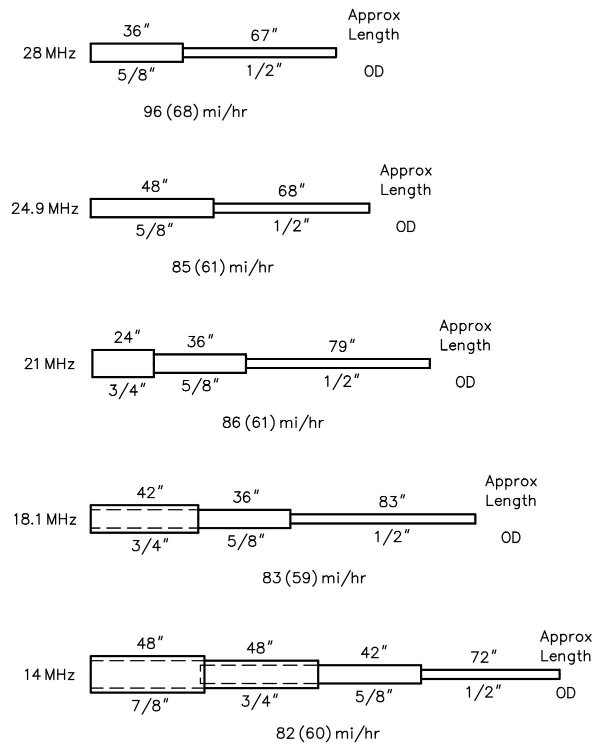
28 MHz
36″
67″
Approx Length
5/8″
1/2″
OD
96 (68) mi/hr
24.9 MHz
48″
68″
Approx Length
5/8″
1/2″
OD
85 (61) mi/hr
21 MHz
24″
36″
79″
Approx Length
3/4″
5/8″
1/2″
OD
86 (61) mi/hr
18.1 MHz
42″
36″
83″
Approx Length
3/4″
5/8″
1/2″
OD
83 (59) mi/hr
14 MHz
48″
48″
42″
72″
Approx Length
7/8″
3/4″
5/8″
1/2″
OD
82 (60) mi/hr

Fig 9—Half-element designs for Yagi antennas. The other side of the element is identical, and the center section should be a single piece twice as long as the length shown here for the largest diameter section. Use 0.058-in.-wall aluminum tubing throughout. Broken lines indicate double tubing thickness, where one tube is inserted into another. The overlap insertion depth into a tube two sizes larger, where shown, should be at least two inches. Maximum survival wind speeds without ice are shown adjacent to each design; values enclosed in parentheses are survival speeds for ¼ inch of radial ice.

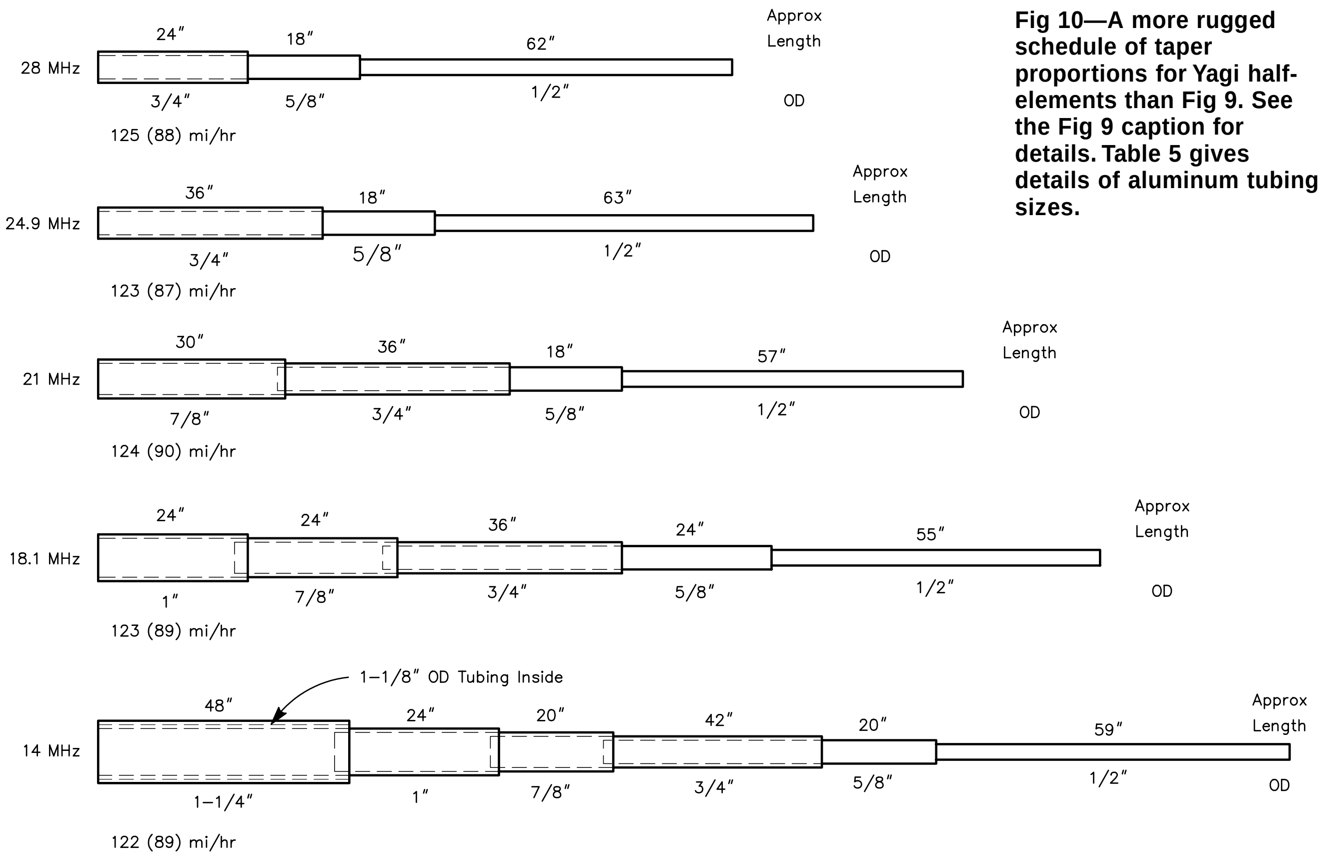
28 MHz
24″
18″
62″
Approx Length
3/4″
5/8″
1/2″
OD
125 (88) mi/hr
24.9 MHz
36″
18″
63″
Approx Length
3/4″
5/8″
1/2″
OD
123 (87) mi/hr
21 MHz
30″
36″
18″
57″
Approx Length
7/8″
3/4″
5/8″
1/2″
OD
124 (90) mi/hr
18.1 MHz
24″
24″
36″
24″
55″
Approx Length
1″
7/8″
3/4″
5/8″
1/2″
OD
123 (89) mi/hr
1–1/8″ OD Tubing Inside
14 MHz
48″
24″
20″
42″
20″
59″
Approx Length
1–1/4″
1″
7/8″
3/4″
5/8″
1/2″
OD
122 (89) mi/hr

Fig 10—A more rugged schedule of taper proportions for Yagi half-elements than Fig 9. See the Fig 9 caption for details. Table 5 gives details of aluminum tubing sizes.

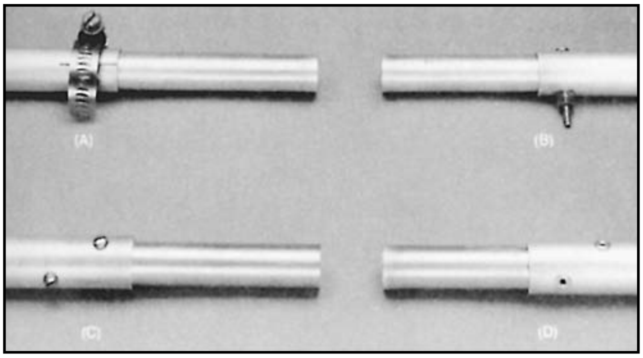

Fig 11—Methods of connecting telescoping tubing sections to build beam elements. See text for a discussion of each method.

Table 6
Hose-Clamp Diameters

	Clamp Diameter (In.)	
Size No.	*Min*	*Max*
06	7/16	7/8
08	7/16	1
10	1/2	1 1/8
12	5/8	1 1/4
16	3/4	1 1/2
20	7/8	1 3/4
24	1 1/8	2
28	1 3/8	2 1/4
32	1 5/8	2 1/2
36	1 7/8	2 3/4
40	2 1/8	3
44	2 5/16	3 1/4
48	2 5/8	3 1/2
52	2 7/8	3 3/4
56	3 1/8	4
64	3 1/2	4 1/2
72	4	5
80	4 1/2	5 1/2
88	5 1/8	6
96	5 5/8	6 1/2
104	6 1/8	7

steel" models do not have a stainless screw and will rust) are recommended for longest antenna life. **Table 6** shows available hose-clamp sizes.

Figs 11B, 11C and 11D show possible fastening methods for joints that do not require adjustment. At B, machine screws and nuts hold the elements in place. At C, sheet metal screws are used. At D, rivets secure the tubing. If the antenna is to be assembled permanently, rivets are the best choice. Once in place, they are permanent. They will never work free, regardless of vibration or wind. If aluminum rivets with aluminum mandrels are used, they will never rust. In addition, there is no danger of dissimilar-metal corrosion with aluminum rivets and aluminum antenna elements. If the antenna is to be disassembled and moved periodically, either B or C will work. If machine screws are used, however, take all possible precautions to keep the nuts from vibrating free. Use lock washers, lock nuts and flexible sealant such as silicone bathtub sealant to keep the hardware in place.

Very strong elements can be made by using a double thickness of tubing, made by telescoping one size inside another for the total length. This is usually done at the center of an element where more element strength is desired at at the boom support point, as in the 14-MHz element in Fig 10. Other materials can be used as well, such as wood dowels, fiberglass rods, and so forth.

In each case where a smaller diameter length of tubing is telescoped inside a larger diameter one, it's a good idea to coat the inside of the joint with Penetrox or a similar substance to ensure a good electrical bond. Antenna elements have a tendency to vibrate when they are mounted on a tower, and one way to dampen the vibrations is by running a piece of clothesline rope through the length of the element. Cap or tape the end of the element to secure the clothesline. If mechanical requirements dictate (a U-bolt going through the center of the element, for instance), the clothesline may be cut into two pieces.

Antennas for 50 MHz need not have elements larger than 1/2-inch diameter, although up to 1 inch is used occasionally. At 144 and 220 MHz the elements are usually 1/8 to 1/4 inch in diameter. For 420 MHz, elements as small as 1/16 inch diameter work well, if made of stiff rod. Aluminum welding rod of 3/32 to 1/8 inch diameter is fine for 420-MHz arrays, and 1/8 inch or larger is good for the 220-MHz band. Aluminum rod or hard-drawn wire works well at 144 MHz.

Tubing sizes recommended in the paragraph above are usable with most formula dimensions for VHF/UHF antennas. Larger diameters broaden the frequency response; smaller ones sharpen it. Much smaller diameters than those recommended will require longer elements, especially in 50-MHz arrays.

Element Taper and Electrical Length

The builder should be aware of one important aspect of telescoping or tapered elements. When the element diameters are tapered, as shown in Figs 9 and 10, the electrical length is not the same as it would be for a cylindrical element of the same total length. Length corrections for tapered elements are discussed in Chapter 2.

Other Materials for Antenna Construction

Wood is very useful in antenna work. It is available in a great variety of shapes and sizes. Rug poles of wood or bamboo make fine booms. Bamboo is quite satisfactory for spreaders in quad antennas.

Round wood stock (doweling) is found in many hardware stores in sizes suitable for small arrays. Wood is good for the framework of multibay arrays for the higher bands, as it keeps down the amount of metal in the active area of the array. Square or rectangular boom and frame materials can be cut to order in most lumber yards if they are not available from the racks in suitable sizes.

Wood used for antenna construction should be well seasoned and free of knots or damage. Available materials vary, depending on local sources. Your lumber dealer can help you better than anyone else in choosing suitable materials. Joining wood members at right angles can be done with gusset plates, as shown in **Fig 12**. These can be made of thin outdoor-grade plywood or Masonite. Round materials can be handled in ways similar to those used with metal components, with U clamps and with other hardware.

In the early days of Amateur Radio, hardwood was used as insulating material for antennas, such as at the center and ends of dipoles, or for the center insulator of a driven element made of tubing. Wood dowels cut to length were the most common source. To drive out moisture and prevent the subsequent absorption of moisture into the wood, it was treated before use by boiling it in paraffin. Of course today's technology has produced superior materials for insulators in terms of both strength and insulating qualities. However, the technique is worth consideration in an emergency situation or if low cost is a prime requirement. "Baking" the wood in an oven for a short period at 200° F should drive out any moisture. Then treatment as described in the next paragraph should prevent moisture absorption. The use of wood insulators should be avoided at high-voltage points if high power is being used.

All wood used in outdoor installations should be protected from the weather with varnish or paint. A good grade of marine spar varnish or polyurethane varnish will offer protection for years in mild climates, and one or more seasons in harsh climates. Epoxy-based paints also offer good protection.

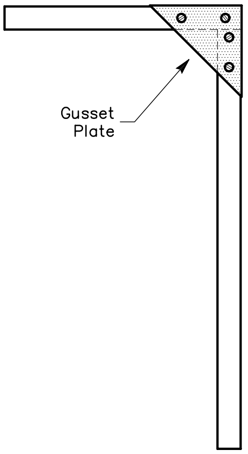

Fig 12—Wood members can be joined at right angles using gusset plates.

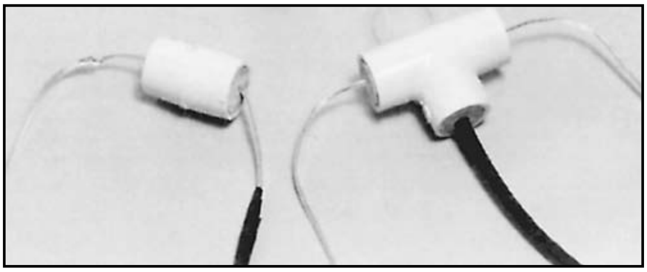

Fig 13—Plastic plumbing parts can be used as antenna center and end insulators.

Plastics

Plastic tubing and rods of various sizes are available from many building-supplies stores. The uses for the available plastic materials are limited only by your imagination. Some amateurs have built beam antennas for VHF using wire elements run inside thin PVC plumbing pipe. The pipe gives the elements a certain amount of physical strength. Other hams have built temporary antennas by wrapping plastic pipe with aluminum foil or other conductive material. Plastic plumbing pipe fittings can also be used to enclose baluns and as the center insulator or end insulators of a dipole, as shown in **Fig 13**. Plastic or Teflon rod can be used as the core of a loading coil for a mobile antenna (**Fig 14**) but the material for this use should be selected carefully. Some plastics become quite warm in the presence of a strong RF field, and the loading-coil core might melt or catch fire!

Fiberglass

Fiberglass poles are the preferred material for spreaders for quad antennas. They are lightweight, they withstand harsh weather well, and their insulating qualities are excellent. One disadvantage of fiberglass poles is that they may be crushed rather easily. Fracturing occurs at the point where the pole is crushed, causing it to lose its strength. A crushed pole is next to worthless. Some amateurs have repaired crushed poles with fiberglass cloth and epoxy, but the original strength is

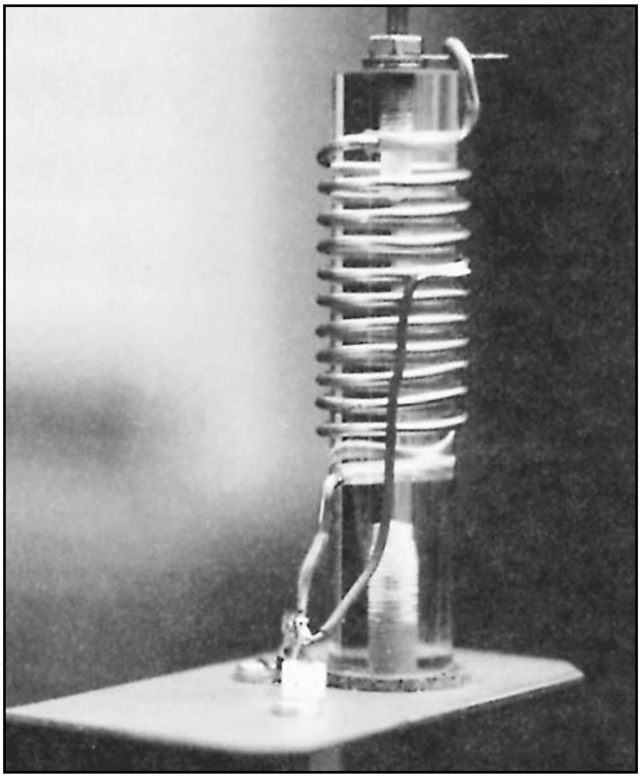

Fig 14—A mobile-antenna loading coil wound on a polystyrene rod.

nearly impossible to regain.

Fiberglass poles can also be used to construct other types of antennas. Examples are helically wound Yagi elements or verticals, where a wire is wound around the pole.

CONCLUSION

The antenna should be put together with good quality hardware. Stainless steel is best for long life. Rust will quickly attack plated steel hardware, making nuts difficult, if not impossible, to remove. If stainless-steel muffler clamps and hose clamps are not available, the next best thing is to have them plated. If you can't have them plated, at least paint them with a good zinc-chromate primer and a finish coat or two.

Galvanized steel generally has a longer life than plated steel, but this depends on the thickness of the galvanizing coat. Even so, in harsh climates rust will usually develop on galvanized fittings in a few years. For the ultimate in long-term protection, galvanized steel should be further protected with zinc-chromate primer and then paint or enamel before exposing it to the weather.

Good quality hardware is expensive initially, but if you do it right the first time, you won't have to take the antenna down in a few years and replace the hardware. When the time does come to repair or modify the antenna, nothing is more frustrating than fighting rusty hardware at the top of the tower.

Basically any conductive material can be used as the radiating element of an antenna. Almost any insulating material can be used as an antenna insulator. The materials used for antenna construction are limited mainly by physical considerations (required strength and resistance to outdoor exposure) and by the availability of materials. Don't be afraid to experiment with radiating materials and insulators.

BIBLIOGRAPHY

Source Material and more extended discussion of topics covered in this chapter can be found in the reference given below.

J. J. Elengo, Jr., "Predicting Sag in Long Wire Antennas," *QST*, Jan 1966, pp 57-58.

D. B. Leeson, *Physical Design of Yagi Antennas* (Newington, CT: ARRL).

Chapter 21

Antenna Products Suppliers

Antenna Manufacturers Products

Finding parts can be the most difficult aspect of an antenna project. Suppliers of aluminum exist in most major metropolitan areas. They can be found in the Yellow Pages of the phone book. Some careful searching of the Yellow Pages may also reveal sources of other materials and accessories. If you live away from a metropolitan area, try using telephone books for the nearest large metropolitan area; they may be available in the reference section of your local library.

Many dealers and distributors will ship their products by freight or by mail. Tables 1 through 7 list several categories of antenna products and some suppliers of them. Company names have been abbreviated where necessary. Table 8 is an address list arranged alphabetically by company name.

Product lines change often; we recommend that you

Table 1
VHF/UHF/Microwave Antenna Suppliers

Manufacturer	Yagi	Quad	Loop Yagi	Vertical	Mobile	HT	Microwave	Helical	Satellite
Antennaco	✓								
Ant Specialists				✓	✓				
Austin Antenna				✓	✓	✓			
Butternut				✓					
Centurion						✓			
Comet				✓	✓	✓			
Create Design	✓								
Cubex		✓							
Cushcraft	✓		✓	✓	✓				✓
Diamond	✓		✓	✓	✓				
Down East	✓		✓				✓		✓
Eur-AM				✓	✓	✓			
Force 12	✓			✓					
Hustler	✓			✓	✓				
Hy-Gain (MFJ)	✓			✓					✓
Kilo-Tec				✓					
Lakeview				✓	✓	✓			
Larsen				✓	✓				
M^2	✓	✓		✓					
Maldol					✓				
MFJ	✓			✓	✓	✓			
Mosley	✓			✓					
Pro-Am Valor					✓				
RadioShack				✓	✓	✓			
Radio Works					✓				
Sommer	✓								
Spectrum Intl	✓		✓				✓	✓	✓
Tennadyne	LPDA								
Tonna	✓						✓		

request current catalogs from those manufacturers who interest you. In addition, all indications of sales policies and prices for catalogs are given for general information only and are subject to change without notice.

Antenna products for repeaters are listed separately, in Chapter 17.

Table 2
HF Antenna Suppliers

Manufacturer	Yagi	Quad	LPDA	Vertical	Dipole	Mobile	Small Xmtg	Active (RX only)
Alpha Delta				✓	✓	✓		
B & W				✓	✓			
Bilal							✓	
Butternut/Bencher	✓			✓			✓	
Create Design	✓		✓	✓	✓		✓	
Cubex		✓						
Cushcraft	✓		✓	✓	✓			
Flytecraft				✓		✓		
Force 12	✓		✓	✓	✓	✓	✓	
GAP				✓				
Gem Quad		✓						
Grove					✓	✓		✓
High Sierra				✓	✓	✓		
Hustler				✓		✓		
Hy-Gain (MFJ)	✓			✓	✓			
Lakeview/Hamstick						✓		
Lightning Bolt		✓						
M^2	✓				✓			
MFJ	✓	✓	✓	✓	✓	✓	✓	✓
Mosley	✓			✓	✓			
Palomar								✓
Pro-Am Valor						✓		
Roadrunner						✓		
Sommer	✓			✓				
Spi-Ro				✓	✓			
Tennadyne			✓					
Texas Radio					✓	✓		
Radio Works				✓	✓			
Van Gorden				✓	✓			
W1JC					✓		✓	
W9INN					✓			

Table 3
Antenna Parts

Manufacturer	Hardware	Insulators	Traps	Aluminum Tubing	Wire	Transmission Lines
Alexanderson AeroPlane Co				✓		
Barker & Williamson		✓	✓			
Cable X-Perts	✓	✓			✓	✓
Hy-Gain (MFJ)	✓	✓				
Metal & Cable, Inc	✓			✓	✓	
Ocean State Electronics	✓	✓			✓	✓
Spi-Ro		✓	✓		✓	✓
Texas Towers				✓	✓	✓
Radio Works	✓	✓			✓	✓
Van Gorden	✓	✓			✓	✓
The Wireman	✓	✓			✓	✓
W1JC		✓			✓	
W9INN					✓	

Table 4
Suppliers of Quad Antenna Parts

Company	*Material Size*
Cubex Co., Inc.	$1\frac{1}{4}$-$\frac{1}{2}$ in. × 13 ft; $1\frac{1}{2}$ in. × 13 ft cast spiders, boom-to-mast mounts.
Lightning Bolt	Custom-made fiberglass spreaders; any length.
Max-Gain Systems, Inc.	Quad spreaders; fiberglass insulators.

Table 6
Transmission Lines

Source	Coax	Hardline	Ladder Line
Belden	✓	✓	
Cable X-Perts	✓	✓	✓
International Wire & Cable	✓		
Nemal	✓	✓	
The Radio Works	✓	✓	✓
W1JC			✓
W9INN			✓
The Wireman	✓	✓	✓

Table 5
Towers, Masts and Accessories

Towers
Aluma
Champion Radio
Create Design
Force 12
Glen Martin Engineering
Heights Tower
Hy-Gain (MFJ)
National Tower
RadioShack (masts only)
Rohn
Rotating Tower Systems
Tashjian Towers Corporation
Texas Tower
Trylon
Universal Manufacturing
US Tower

Climbing and Safety Equipment
Champion Radio Products
ONV
Texas Towers

Rotators
Create Design
Hy-Gain (MFJ)
M^2
RadioShack
The Rotor Doctor
Yaesu

Stacking Frames
(Unless otherwise noted these frames are for use in stacking the manufacturer's own antennas in pairs or in quads. These stacking kits are for VHF or UHF antennas only.)
C3i
Cushcraft
Down East
IIX
Spectrum International

Combiners, Power Dividers and Phasing Harnesses
Byers
C3i
Down East
Spectrum International
Tonna

Table 7
Transmission Line Instruments and Accessories

Matching Networks
Ameritron
Barker & Williamson
Cubex
ICOM (mobile & fixed)
Kenwood
MFJ
Nye
SGC
Ten-Tec
Texas Radio (mobile only)
Vectronics
XMatch

Ferrite Cores and Rods
Amidon
Palomar
The Wireman

Filters—TVI (Low Pass & High Pass)
K-Com
MFJ

Lightning Arresters
Alpha Delta
Ameritron
Comet
Cushcraft
Hy-Gain (MFJ)
Polyphaser
Radioware
Rohn
The Wireman
Zero Surge Inc

Switches (Manual, Coax)
Alpha Delta
Barker & Williamson
MFJ

Switches (Remote, Coax)
Ameritron
MFJ

SWR and Wattmeters
Autek
Bird
Coaxial Dynamics
MFJ
Nye
Palomar
RF Parts
Texas Radio

Table 8
Suppliers Addresses

We have made every effort to ensure that this list is complete and accurate as of mid 2003. The ARRL takes no responsibility for errors or omissions. Similarly, a listing here does not represent an endorsement of a manufacturer or products by the ARRL. Refer to the product reviews in *QST* for descriptions of particular products that interest you. To the best of our knowledge the suppliers listed are willing to sell products to amateurs by mail unless indicated otherwise. This listing will be updated with each edition of *The Antenna Book* and *The ARRL Handbook* in the TISFIND manufacturer database. Check ads in *QST* and other Amateur Radio publications for any changes to this information. Suppliers who wish to be listed or update their information are urged to contact the editors.

Advanced Composites
1154 S. 300 W.
Salt Lake City, UT 84101
801-467-1204
fax 801-467-4367
e-mail
info@advancedcomposites.com

Advanced Specialties
114 Essex Street
Lodi, NJ 07644
USA
Tel: 800-926-9426
Email: advanspec@aol.com
http://advancedspec.freeyellow.com/

AEA—Div. Tempo Research Corporation
1221 Liberty Way
Vista, CA 92083
760-598-9677
fax 760-598-4898
e-mail tempo@inetworld.com
web www.aea-wireless.com

AFT
132 Boulevard Dauphinot
F-51100 Reims, France
+33 326 070 047
fax +33 326 023 654
e-mail Antennes-ft@wanadoo.fr
web www.f9ft.com

Alexander Aeroplane Co
PO Box 909
Griffin, GA 30224
800-831-2949; 404-229-2329

Alan Broadband Company
(Zap Checker)
93 Arch Street
Redwood City, CA 94062
USA
Tel: 888-369-9627 (orders)
650-369-9627
Fax: 650-369-3788
Email: ABCom@prodigy.net
http://www.alanbroadband.com/

Alpha Delta Communications
PO Box 620
Manchester, KY 40962
606-598-2029
fax 606-598-4413

Aluma Tower Co, Inc
PO Box 2806-AL
Vero Beach, FL 32961-2806
772-567-3423
fax 772-567-3432
e-mail atc-t@alumatower.com
web www.alumatower.com/index.html

Amateur Electronic Supply
5710 W Good Hope Rd
Milwaukee, WI 53223
800-558-0411
web www.aesham.com/

Ameritron Division
116 Willow Road
Starkville, MS 39759
662-323-8211 (Tech)
fax 662-323-6551
e-mail
ameritron@ameritron.com
web www.ameritron.com/ameritron

Amidon Associates, Inc.
1510 E. Edinger Avenue, Unit B
Santa Ana CA 92705
800-898-1883; 714-547-4494
fax 714-547-4433
e-mail sales@amidon-inductive.com
web www.amidon-inductive.com/

Antennaco Inc
102 Armory Road
PO Box 218
Milford, NH 03055-0218
603-673-3153
fax 603-673-4347

Allen Telecom Group
Antenna Specialists Mobile Division
30500 Bruce Industrial Parkway
Cleveland, OH 44139-3996
USA
440-349-8400
fax 440-349-8404
web www.antenna.com/

Array Solutions
350 Gloria Rd
Sunnyvale, Tx 75182
USA
Tel: 972-203-8810
Fax: 972-203-8811
Email: wx0b@arrysolutions.com
http://www.arraysolutions.com

ASA Antenna Sales
PO Box 3461
Myrtle Beach, SC 29578
800-772-2681

Associated Radio Communications
8012 Conser
Overland Park, KS 66204
913-381-5900; 800-497-1457
fax 913-648-3020
e-mail assocrad@tfs.net
web www.associatedradio.com

Austin Amateur Radio Supply
5310 Cammeron Road
Austin, TX 78723
800-423-2604; 512-454-2994
fax 512-454-3069

Austin Antenna, Ltd
10 Main St
Gonic, NH 03839
603-336-6339
fax 603-335-1756

Autek Research
PO Box 7556
Wesley Chapel, FL 33544
813-994-2199
e-mail
Mailbox@autekresearch.com
web www.autekresearch.com/
index.htm

Barker & Williamson Co
603 Cidco Rd
Cocoa, FL 32926
321-639-2545
fax 321-639-2545
e-mail
custsrvc@bwantennas.com
web www.bwantennas.com

Barry Electronics Corp
540 Broadway
New York, NY 10012
212-925-7000
fax 212-925-7001

Belden Wire & Cable
PO Box 1980
Richmond, IN 47374
317-983-5257
fax 317-983-5257
web www.belden.com

Bencher Inc.
831 North Central Avenue
Wood Dale, IL 60191
630-238-1183
fax 630-238-1186
e-mail bencher@bencher.com
web www.bencher.com/

Bilal Company
137 Manchester Dr
Florissant, CO 81816
719-687-0650
web www.isotronantennas.com/

Bird Electronics Corporation
30303 Aurora Rd.
Solon, OH 44139
866-695-4569
e-mail sales@bird-technologies.com
web www.bird-electronic.com/

Brian Beezley, K6STI
3532 Linda Vista Dr
San Marcos, CA 92069
619-599-4962
e-mail k6sti@n2.net

Burghardt Amateur Center, Inc.
710 10th Street SW
PO Box 73
Watertown, SD 57201
800-927-4261
605-886-7314 (Service)
605-886-6914 (Fax-back product info)
fax 605-886-3444
e-mail burghart@daknet.com
web www.burghardt-amateur.com/

Butternut Electronics Co
See Bencher Inc.

Byers Chassis Kits
5120 Harmony Grove Road
Dover, PA 17315
717-292-4901
(6p-9p EST M-F, 8a-4p Sat)
717-292-4901 (24 hrs)

Cable X-Perts Inc.
225 Larkin Drive Suite #6
Wheeling, IL 60090-7209
800-828-3340 (orders only)
847-520-3003 (Tech. Info.)
fax 847-520-3444
e-mail cxp@cablexperts.com
web www.cablexperts.com/

CAL-AV Labs, Inc.
1802 w. Grant road, Ste. 116
Tucson, AZ 85745
USA
Tel: 520-624-1300
888-815-0400 (orders only)
Fax: 520-624-1311
Email: info@cal-av.com
http://www.cal-av.com/

Cardwell Condenser Corp
80 East Montauk Hwy
Lindenhurst, NY 11757
516-957-7200
fax 516-957-7203
web www.cardwellcondenser.com

C.A.T.S.
Formerly known as: Rotor Doctor
7368 SR 105
Pemberville, OH 43450
419-353-2287
fax 419-354-7746
e-mail mailto:craig@rotor-doc.com
web www.rotordoc.com

C-Comm
6115 15th NW
Seattle, WA 98107
206-784-7337; 800-426-6528
fax 206-784-0541

C3i
1401 K Street NW Suite 900
Washington DC 20005
(800) 224-5137
fax (202) 362-8481
e-mail info@C3iusa.com
web www.c3iusa.com/
corpover.html

Centurion International
PO Box 82846
Lincoln, NE 68501-2846
402-467-4491
fax 800-848-3825

Champion Radio Products
Box 572
Woodinville, WA 98072
425-485-7913
206-890-4188 Cell Phone
fax 360-668-1447
e-mail championradio@aol.com
web www.championradio.com/

Coaxial Dynamics
15210 Industrial Parkway
Cleveland, Ohio 44135-3308
800-262-9425
fax 216-267-3142
e-mail cdi-sales@coaxial.com
web www.coaxial.com/
home_frames.htm

Comet North America Inc.
394 Wards Corner Road
Suite 130
Cincinnati, Ohio, 45140
513-831-5000
fax 513-831-7889
e-mail
info@CometNorthAmerica.com
web www.cometnorthamerica.com/

Comet Antenna
See NCG

Comm-Pute
7946 State Street
Midvale, UT 84047
801-567-9944
fax 801-567-9494
e-mail bobwood@xmission.com
web www.comm-pute.com/

Communication Headquarters Inc.
3832 Oleander Drive
Wilmington, NC 28403
910-791-8885
800-688-0073 (orders)
fax 910-452-3891 (orders)
web www.chq-inc.com/

Communications Data Corp
1051 Main St
St Joseph, MI 49085
269-982-0404
fax 269-982-0433
e-mail mailto:did@gtm.net

Comtelco Industries Inc
501 Mitchell Rd
Glendale Heights, IL 60139
800-634-4622; 708-7790-9894
fax 708-798-9799

Create
4-8 Asano-Cho
Kawasaki-Ku
Kawasaki-City, Japan
044 (333) 6681
fax 044 (333) 6598
e-mail email@cd-corp.com

Cubex Co.
228 Hibiscus St, #9
Jupiter, FL 33458
561-748-2830
fax 561-748-2831
e-mail CubexCo@aol.com
web www.cubex.com/

Cushcraft Corp
48 Perimeter Rd.
Manchester, NH 03108
603-627-7877
e-mail sales@cushcraft.com
web www.cushcraft.com/

Dave's Hobby Shop
600 Main St.
Van Buren, AK 72956
USA
Tel: 479-471-0750
Web www.daveswebshop.com

Davis RF Co
See Radioware (distributor)
PO Box 730
Carlisle, MA 01741
978-371-1356
978-369-1738
800-328-4773 (Orders only)
fax 978-369-3484
e-mail davisrfinc@aol.com
web www.davisRF.com

Diamond Antennas
See RF Parts

Down East Microwave
954 Rt 519
Frenchtown, NJ 08825
908-996-3584
fax 908-996-3702
web downeastmicrowave.com/

Dressler Hochfrequenztechnik
GMBH
Werther Strasse 14-16
W-5190 Stolberg
Germany

DX Engineering
POB 440
Peninsula, Ohio 44264
USA
Fax: 330-657-2168
Email:
dxengineering@dxengineering.com
http://www.dxengineering.com/

EDCO - Electronic Distributors Company
325 Mill Street
Vienna, VA 22180
703-938-8105
fax 703-938-6911
e-mail web www.elecdist.com/
index.html

EUR-AM Antennas
PO Box 225
Moultonboro, NH 03254-0225
603-476-5113
fax 603-476-5113
e-mail info@eur-am.com
web www.eur-am.com/

EZ Hang, Inc.
8645 Tower Dr.
Code C
Laurel, MD 20723
Tel: 540-286-0176
Fax: 202-260-3797
http://www.ezhang.com

Fluidmotion Incorporated
14135 233rd Place SE
Issaquah, WA 98027
425-456-0200; 800-885-8700
fax 425-391-6031
e-mail sales@steppir.com
web www.steppir.com/

Flytecraft
PO Box 3141
Simi Valley, CA 93093
805-583-8173

Force 12
PO Box 1349
Paso Robles, CA 93447
800-248-1985
805-227-1680 (Tech)
fax 805-227-1684
e-mail
mailto:force12e@lightlink.com
web force12inc.com/

GAP Antenna Products
99 North Willow St
Fellsmere, FL 32948
772-571-9922
Email: contact@gapantenna.com
web www.gapantenna.com

Gem Quad Products Ltd
PO Box 291
Boissevain, MB R0K 0E0
Canada
204-534-6184
fax 204-534-6492
e-mail
mailto:gemquad@escape.ca
web www.escape.ca/~gemquad/

GLA Systems
PO Box 425
Caddo Mills, TX 75135
903-527-4163
800-588-2841
fax 214-381-2895
web www.texasbugcatcher.com/

Grove Enterprises Inc.
PO Box 98
Brasstown, NC 28902
800-438-8155 (orders)
704-837-9200 (BBS)
fax 704-837-2216
e-mail nada@grove.net
web www.grove-ent.com/

Ham Radio Outlet
1702 W. Camelback Rd
Phoenix, AZ 85015
web www.hamradio.com/
800-444-4799 Mid Atlantic
800-444-9476 Mountain
800-444-0047 New England
800-644-4476 Northeast
800-444-7927 Southeast
800-854-6046 West

The Ham Station
220 North Fulton Avenue
PO Box 6522
Evansville, IN 47719-0522
800-729-4373 (orders)
812-422-0231 (Tech)
812-422-0252 (Service)
fax 812-422-4253
e-mail sales@hamstation.com
web www.hamstation.com/

Hamtronics, Inc
65-Q Moul Rd
Hilton, NY 14468
716-392-9430
fax 716-392-9420
web www.hamtronics.com

Hamware.de
Int'l: Klaus Bemmerer
Tel: +49 4371 869145
e-mail service@hamware.de
Web www.hamware.de
US: Dillon RF Systems
Email dillonel@mtaonline.net

Heights Tower Systems
1529 Gulf Beach Hwy
Pensacola, FL 32507
850-455-1210
fax 850-455-4355
e-mail info@heightstowers.com
web www.heightstowers.com/

Hi-Q Antennas
21085 Cielo Vista Way
Wildomar, CA 92595
909-674-4862
fax 909-245-2031
e-mail sales@hiqantennas.com
web www.hiqantennas.com/

High Sierra Antennas
Box 2389
Nevada City, CA 95959
888-273-3415; 530-273-3415
fax 530-273-7561
e-mail heath@hsantennas.com
web www.hsantennas.com

Hustler Antennas
see New-Tronics Antenna Corp.

Hy-Gain (also see MFJ)
308 Industrial Park Road
Starkville, MS 39759
800-647-1800 (Sales)
662-323-9358 (Tech)
fax 662-323-6551
web www.hy-gain.com/

ICOM America, Inc.
2380 116th Ave, NE, Suite S
Bellevue, WA 98004
USA
Tel: 800 872-4266
Fax: 425 454-1509
Web www.icomamerica.com

Industrial Communications
Engineers (ICE)
Array Solutions
350 Gloria Rd.
Sunnyvale, TX 75182
972-203-2008; 800-423-2666
fax 972-203-8811
e-mail wx0b@arraysolutions.com
web www.iceradioproducts.com/

IIX Equipment, Ltd
PO Box 9
Oak Lawn, IL 60454
708-423-0605
fax 708-423-1691
web www.w9iix.com/

International Radio
13620 Tyee Road
Umpqua, OR 97486
541-459-5623 9AM-1PM PDT,
Tues-Sat
fax 541-459-5632
e-mail inrad@rosenet.net
web www.qth.com/inrad/

Kenwood USA Corporation
3975 Johns Creek Court,
Suite 300
Suwanee, GA 30024
USA
Tel: 678 474-4700
Fax: 678 474-4730
Web www.kenwoodusa.com

K0XG Systems
1117 Highland Park Drive
Bettendorf, IA 52722
USA
Tel: 563 355-7451
Email K0XG@ K0XG.com
Web www.k0xg.com

Kilo-Tec
PO Box 10
Oakview, CA 93022
805-646-9645

K-Com
PO Box 82
Randolph OH 44265
330-325-2110
fax 330-325-2525
e-mail k-com@worldnet.att.net
web www.k-comfilters.com/
howtoget.asp

Lakeview Co Inc
3620-9A Whitehall Rd
Anderson, SC 29626
(864) 226-6990
fax (864) 225-4565
web www.hamstick.com

Larsen Electronics, Inc
See Radial/Larsen

LDG Electronics
1445 Parran Road
St Leonard, MD 20685
877-890-3003 Toll-Free Orders
410-586-2177 Tech Support
fax 410-586-8475
e-mail ldg@ldgelectronics.com
web www.ldgelectronics.com

Lentini Communications
21 Garfield St
Newington, CT 06111
860-666-6227
800-666-0908 (outside CT)
fax 860-667-3561
e-mail radio@lentinicomm.com
web www.lentinicomm.com

Roy Lewallen, W7EL
PO Box 6658
Beaverton, OR 97007
503-646-2885
fax 503-671-9046
e-mail w7el@eznec.com
web www.eznec.com

Lightning Bolt Antennas
RD #2, RT 19
Volant, PA 16156
724-530-7396
fax 724-530-6796
e-mail lbaquads@ztrain.com
web
www.lightningboltantennas.com/

M^2 Antenna Systems Inc.
4202 N. Selland
Fresno CA, 93722
559-432-8873
fax 559-432-3059
e-mail wyatt@m2inc.com
web www.m2inc.com/

M/A-COM, Inc
(an AMP Company)
1011 Pawtucket Blvd
PO Box 3295
Lowell, MA 01853-3295
508-442-4500
fax 508-442-4436
e-mail sales@macom.com
web www.amp.com

Maldol USA
4711 NE 50th St
Seattle, WA 98105
Answerfax 206-525-1896
fax 206-524-7826
e-mail transtec@cyberquest.com

Glen Martin Engineering
13620 Old Highway 40
Boonville, MO 65233
660-882-2734; 800-486-1223
fax 660-882-7200
web www.glenmartin.com

The Mast Company
Henry Pollock - K4TMC
P.O. Box 1932
Raleigh, NC 27602
Email: k4tmc@aol.com
http://www.geocities.com/tmastco/

MCM Electronics
650 Congress Park Dr
Centerville, OH 45459-4072
800-543-4330
fax 800-765-6960
web www.i-mcm.com

Memphis Amateur Electronics
1465 Wells Station Rd
Memphis, TN 38108
901-683-9125; 800-238-6168
fax 901-682-7165

Metal and Cable Corp, Inc
9241 Ravenna Road, Unit C-10
PO Box 117
Twinsburg, OH 44087
330-425-8455
fax 330-425-3504
web www.metal-cable.com/

MFJ Enterprises
300 Industrial Park Rd
Starkville, MS 39759
662-323-5869
Email
mfjcustserv@mfjenterprises.com
web www.mfjenterprises.com/

MGS (Max-Gain Systems, Inc.)
221 Greencrest Court
Marietta, GA 30068-3825
770-973-6251 (before 9 PM
eastern)
fax 815) 461-7730
e-mail info@mgs4u.com
web www.mgs4u.com/

Michigan Radio
26014 Groesbeck Hwy
Warren, MI 48089
810-771-4711
800-TRU-HAMM (800-878-4266)
Orders only
810-771-4712 (Service)
fax 810-771-6546
web www.michiganradio.com/

Mirage Communications
116 Willow Road
Starkville, MS 39759
662-323-8287
fax 662-323-6551
web www.mirageamp.com/

James Millen Electronics
PO Box 4215BV
Andover, MA 01810-4215
978-975-2711
fax 978-474-8949
e-mail info@jamesmillenco.com
web www.jamesmillenco.com/
millen/millmain.htm

Mosley Electronics
1325 Style Master Drive
Union, MO 63084
800-325-4016 (Orders)
800-9MOSLEY (966-7539)
314-994-7872 (Technical)
fax 314-994-7873
web www.mosley-electronics.com/

Multi-band Antennas (originally
Spider Antennas)
7131 Ownesmouth Avenue -
Suite 563C
Canoga Park, CA 91303
818-341-5460
web www.spiderantenna.com/

Multi-tech Industries, Inc. (MTI)
(formally Radio Switch)
64 South Main Street
PO Box 159
Marlboro, NJ 07746-0159
800-431-3223; 732-431-0550
732-462-6100
fax 732-409-6695
e-mail multitech@sprynet.com
web www.multi-tech-industries.com

National RF, Inc.
7969 Engineer Road, Suite 102
San Diego, CA 92111
858-565-1319
fax 858-571-5909
web www.nationalrf.com/

National Tower Co
PO Box 15417
Shawnee Mission, KS 66285
800-762-5049; 913-888-8864

NCG Companies—Comet Antenna
1275 North Grove St
Anaheim, CA 92806
714-630-4541; 800-962-2611
fax 714-630-7024
web www.cometantenna.com
web www.natcommgroup.com/

Nemal Electronics Inc.
12240 NE 14th Avenue
North Miami, FL 33161
305-893-3924; 305-849-0900
800-522-2253
fax 305-895-8178
e-mail info@nemal.com
web www.nemal.com/

New-Tronics Antenna Corp. - Hustler
1 Newtronics Place
Mineral Wells, TX 76067-9563
877-994-9499 Sales
940-325-1386 Service
fax 940-328-1409
e-mail sales@new-tronics.com
web www.new-tronics.com/main/index.html

Nittany Scientific
1733 West 12600 South
Suite 420
Riverton, UT 84065
Phone/fax (801) 446-1426
e-mail sales@nittany-scientific.com
web www.nittany-scientific.com/

William M. Nye Co
PO Box 1877
Priest River, ID 83856
208-448-1762
fax 208-448-1832

Ocean State Electronics
PO Box 1458
6 Industrial Dr
Westerly, RI 02891
401-596-3080; 800-866-6626
fax 401-596-3590
web www.oselectronics.com

ONV Safety Belt Co.
Box 404
Ramsey, NJ 07446
800-345-5634; 201-327-2462
fax 201-327-2462

OptiBeam
Thomas Schmenger DF2BO
Rastatter Straße 37
D-75179 Pforzheim
Tel./Fax: (0049) 07231 / 45 31 53
Email: Info@optibeam.de

OptiBeam NorthAmerica
Array Solutions
350 Gloria Rd
Sunnyvale, Tx 75182 USA
Tel: 972-203-8810
Fax: 972-203-8811
Email: wx0b@arrysolutions.com
http://www.arraysolutions.com

Orion now sold by M^2

Outbacker Antenna Sales
330 Cedar Glen Circle
Chattanooga, TN 37412
615-899-3390
fax 615-899-6536

Palomar Engineers
PO Box 462222
Escondido, CA 92046
619-747-3343
fax 619-747-3346
e-mail palomar@compuserve.com
web www.palomar-engineers.com/

Palstar, Inc.
9676 N Looney Rd
P O Box 1136-1482
Piqua, Ohio 45356
USA
Tel: 937-773-6255
Fax: 937-773-8003
Email: Palstar@erinet.com
http://www.palstarinc.com/

Phillystran, Inc
151 Commerce Dr
Montgomeryville, PA 18936
215-368-6611
fax 215-362-7956
web www.phillystran.com/

PolyPhaser Corp
2225 Park Place
PO Box 9000
Minden, NV 89423-9000
702-782-2511
fax 702-782-4476
e-mail info@polyphaser.com
web www.polyphaser.com

Quicksilver Radio Products
30 Tremont St.
Meriden, CT 06450
USA
Tel: 203-440-4468
http://www.qsradio.com

Radiall/Larsen Antenna Technologies
3611 NE 112th Avenue
Vancouver, WA 98682
800-ANTENNA (800-268-3662)
360-944-7551
FAX 360-944-7556
fax 800-525-6749
e-mail info@radialllarsen.com
web www.radialllarsen.com/default.htm

Radio Bookstore (Radioware)
PO Box 209
Rindge, NH 03461-0209
603-899-6957
800-457-7373 (10am-6pm EST)
fax 603-899-6826
e-mail radware@radio-ware.com
web www.radiobooks.com/

Radio Center USA
1242 Howell
N. Kansas City, MO 64116
816-459-8832; 800-821-7323

Radio City
2663 County Rd I
Mounds View, MN 55112
612-786-4475; 800-426-2891
fax 612-786-6513
web www.radioinc.com

Radio Engineers/Technitron (see National RF, Inc.)

Radio Shack
(Contact your local store)

Radio Switch Corp
See Multi-tech Industries

Radio Works
PO Box 6159
Portsmouth, VA 23703
757-484-0140
800-280-8327 (Orders)
fax 757-483-1873
e-mail jim@radioworks.com
web radioworks.com/

Raibeam Antennas, Int'l
5638 West Alice Avenue
Glendale, AZ 85302
800-530-1913 (orders)
602-931-9135
e-mail wa7rai@raibeam.com
web www.raibeam.com/

RF Parts Co
435 South Pacific St
San Marcos, CA 92069
760-744-0700; 800-737-2787
760-744-0750 (Tech)
fax 760-744-1943; 888-744-1943
e-mail rfp@rfparts.com
web www.rfparts.com

RF TEC Manufacturing
256 Commerce Rd, Suite 517
Peachtree, GA 30269
770-251-2235
fax 770-502-9827
web www.rftec.com/

Roadrunner Resonator
1850 Swanson, #A20
Lake Havasu, AZ 86403
520-453-7211

ROHN Industries, Inc.
PO Box 2000
Peoria, IL 61656
309-697-4400
fax 309-697-5612
e-mail mail@rohnnet.com
web www.rohnnet.com/Index.htm

Ross Distributing Co
78 South State St
Preston, ID 83263
208-852-0830
fax 208-852-0833
web www.rossdist.com

Rotating Tower Sytems, Inc.
Box 44
Prosper, TX 75058
214-347-2560

Rotor Doctor
See C.A.T.S.

SGC Inc
13737 SE 26th St
Bellevue, WA 98005
425-746-6310
fax 425-746-6384
e-mail sgc@sgcworld.com
web www.sgcworld.com/

Shoestring Antennas
PO Box 425
Keyport, WA 98345
360-697-5399
fax 360-697-8416
e-mail jacksa@juno.com
web www.qth.com/shoestring/

Sky-Pole Manufacturing, Inc
1922 Placentia Ave
Costa Mesa, CA 92627
949-548-5596
fax 949-548-5598

Sommer Antennas
PO Box 710
Geneva, FL 32732
407-349-9114
fax 407-349-2485
e-mail sommer1@ix.netcom.com
web www.sommerantennas.com

Spectrum International, Inc
PO Box 1084
Concord, MA 01742
508-263-2145
fax 508-263-7008

Spi-Ro Manufacturing, Inc.
P.O. Box 189
Jonesborough, TN 37659
800-728-7594
423-913-1615 Tech Support
fax 423-913-2131
e-mail mailto:contact@spiromfg.com
web www.spiromfg.com/

Surplus Sales of Nebraska
1502 Jones St
Omaha, NE 68102-3112
402-346-4750
800-244-4567 (Orders only)
fax 402-346-2939
e-mail grinnell@surplussales.com
web www.surplussales.com/

Tarheel Antennas
913 Old Honeycutt Rd
Fuquay-Varina, NC 27256
919-552-8788
fax 919-552-4970
e-mail tarheelantennas@aol.com
web www.tarheelantennas.com/

Tashjian Towers Corporation
(used to be Tri-Ex Towers)
2183 S Highland Ave.
Sanger, CA 93657
559-495-0307
559-284-9707 Mobile
fax 559-495-0557
e-mail tashjian@MSN.com
web www.karltashjian.com/

Telex Communications, Inc
Hy-Gain (see MFJ)

Ten-Tec, Inc
1185 Dolly Parton Parkway
Sevierville, TN 37862
865-453-7172
865-428-0364 (Repairs)
fax 865-428-4483
web www.tentec.com/

Tennadyne Corp
PO Box 34202
San Antonio, TX 78265-4202
210-599-9064
fax 210-599-9064
(same as voice)
e-mail tennadyn@satx.net
web www.tennadyne.com/

Teri Software
PO Box 277
Lincoln, TX 78948
979-542-7952
e-mail
support@antennamodel.com
web www.antennamodel.com

Texas Radio Products
5 E Upshaw
Temple, TX 76501
817-771-1188

Texas Towers
1108 Summit Ave, Suite 4
Plano, TX 75074
800-272-3467
972-422-7306 (Tech)
fax 972-881-0776
e-mail sales@texastowers.com
web www.texastowers.com/

T.G.M. Communications
121 Devon Street
Stratford, Ontario, N5A 2Z8
Canada
519-271-5928
e-mail tgmc@sympatico.ca
web www3.sympatico.ca/tgmc/
index.html

TIC General
PO Box 1 - 1110 Airport Rd
Thief River Falls, MN 56701
218-681-1119
fax 218-681-8509

Timewave Technology, Inc.
1025 Selby Ave, Suite 101
St Paul, MN 55104
USA
Tel: 651-489-5080
Email: sales@timewave.com
http://www.timewave.com

Traffie Technology
421 Jones Hill Road
Ashby, MA 01431-1801
888-599-BEAM (888-599-2326)
978-386-7900
fax 978-386-7900 (same as voice)

Trylon Manufacturing Co, Ltd
21 South Field Dr
PO Box 186
Elmira, Ontario
N3B 2Z6 Canada
519-669-5421
fax 519-669-8912
e-mail info@trylon.com
web www.trylon.com/home.asp

Tucker Electronics and Computers
1717 Reserve St.
PO Box 551419
Garland, TX 75355-1419
800-527-4642 (Test equipment
and catalog requests)
214-348-8800 (x332)
fax 214-340-5460
web www.tucker.com/

Unadilla Antenna Manufacturing
PO Box 4215
Andover, MA 01810-4215
978-975-2711
978-475-7831 (Office)
fax 978-474-8949
e-mail info@unadilla.com
web www.unadilla.com/

Universal Manufacturing Co
43900 Groesbeck Hwy
Clinton Township, MI 48036
810-463-2560
fax 810-463-2964

Universal Radio
6830 Americana Parkway
Reynoldsburg, OH 43068-4113
USA
Tel: 800-431-3939 (Orders)
614-866-4267 (Technical Info)
Fax: 614-866-2339
Email: dx@universal-radio.com
http://www.universal-radio.com/

US Tower Corp
1220 Marcin St
Visalia, CA 93291
559-733-2438
fax 559-733-7194
web www.ustower.com

Valor Enterprises
1711 Commerce Drive
PO Box 601
Piqua, OH 45356-0601
513-778-0074; 800-543-2197
fax 513-778-0259
web www.valorantenna.com/

Van Gorden Engineering
PO Box 21305
South Euclid, OH 44121

Van Valzah Co
38 W 111 Horseshoe Dr
Batavia, IL 60515-9730
708-406-9210

Vectronics
300 Industrial Park Road
Starkville, MS 39759
800-363-2922; 662-323-5800
fax 662-323-6551
e-mail jshurden@vectronics.com
web www.vectronics.com

XMatch
Paul Schrader, N4XM
7001 Briscoe Lane
Louisville, KY 40228
web http://n4xm.myiglou.com/

XX Towers
814 Hurricane Hill Road
Mason, NH 03048
603-878-1102
fax 603-878-4200
e-mail kc1xx@kc1xx.mv.com
web www.xxtowers.com/

W1JC—TV Evans
113 Stratton Brook Rd
Simsbury, CT 06070
860-658-5579
web pages.prodigy.com/w1jc

W3FF Antennas
2390 Templeton Dr.
Redding, CA 96002
USA
Tel: 530-226-8446
Email: sales@buddipole.com
http://www.buddipole.com

W7FG Vintage Manuals
3300 Wayside Drive
Bartlesville, OK 74006
USA
Tel: 800-807-6146 (orders)
918-333-3754
Email: w7fg@eigen.net
http://www.w7fg.com/

W9INN Antennas
PO Box 393
Mt Prospect, IL 60056
847-394-3414

Wacom Products
PO Box 21145
Waco, TX 76702
716-549-4700
e-mail sales@txrx.com

WB0W, Inc
PO Box 8547
Saint Joseph, MO 64508
USA
Tel: 816-364-2691
Email: WB0W@WB0W.com

The Wireman Inc.
261 Pittman Road
Landrum, SC 29356-9544
800-727-WIRE (800-727-9473)
Orders only
864-895-4195 Technical
fax 803-895-5811
e-mail N8UG@earthlink.net
web thewireman.com/

Yaesu U.S.A. (Vertex Standard)
10900 Walker St.
Cypress, CA 90630
714-827-7600
e-mail
amateurtech@vxstdusa.com
web www.yaesu.com/amateur/
amateur.html

YagiStress
Kurt Andress, K7NV
2835 Wade Street
Minden, NV 89423
702-267-5290
e-mail K7NV@contesting.com
web yagistress.freeyellow.com/

Zero Surge Inc
944 State Rt 12
Frenchtown, NJ 08825
908-996-7700
fax 908-996-7773
e-mail zerosurge1@aol.com
web www.zerosurge.com/

Chapter 22

Antenna Supports

A prime consideration in the selection of a support for an antenna is that of structural safety. Building regulations in many localities require that a permit be obtained in advance of the erection of certain structures, often including antenna poles or towers. In general, localities having such requirements also have building safety codes that must be observed. Such regulations may govern the method and materials used in construction of, for example, a self-supporting tower. Checking with your local government building department before putting up a tower may save a good deal of difficulty later, because a tower would have to be taken down or modified if not approved by the building inspector on safety grounds.

Municipalities have the right and duty to enforce any reasonable regulations having to do with the safety of life or property. The courts generally have recognized, however, that municipal authority does not extend to aesthetic questions. The fact that someone may object to the mere presence of a pole, tower or other antenna structure because in his opinion it detracts from the beauty of the neighborhood is not grounds for refusing to issue a permit for a safe structure to be erected. Since the introduction of PRB-1 (federal preemption of unnecessarily restrictive antenna ordinances), this principle has been borne out in many courts. Permission for erecting amateur towers is more easily obtained than in the recent past because of this legislation.

Even where local regulations do not exist or are not enforced, the amateur should be careful to select a location and a type of support that contribute as much safety as possible to the installation. If collapse occurs, the chances of personal injury or property damage should be minimized by careful choice of design and erection methods. A single injury can be far more costly than the price of a more rugged support, in terms of both monetary loss and damage to the public respect for amateur radio.

This chapter has been reviewed and rewritten by Kurt Andress, K7NV.

TREES AS ANTENNA SUPPORTS

From the beginning of Amateur Radio, trees have been used widely for supporting wire antennas. Trees cost noth-

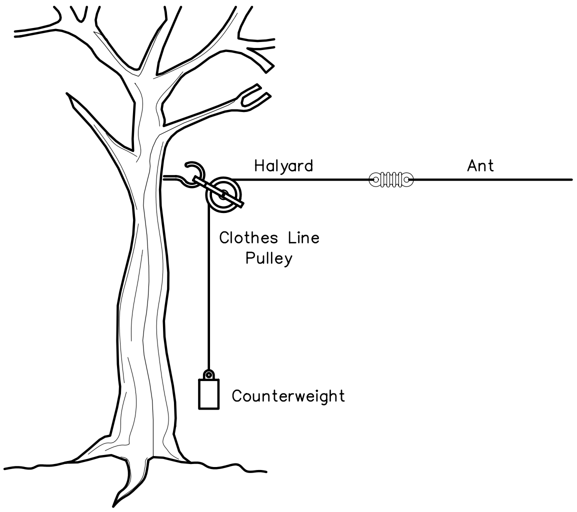

Fig 1—A method of counter weighting to minimize antenna movement and avoid its breaking from tree movement in the wind. The antenna may be lowered without climbing the tree by removing the counterweight and tying additional rope at the bottom end of the halyard. Excess rope may be left at the counterweight for this purpose, as the knot at the lower end of the halyard will not pass through the pulley.

ing to use, and often provide a means of supporting a wire antenna at considerable height. As antenna supports, trees are unstable in the presence of wind, except in the case of very large trees used to support antennas well down from the top branches. As a result, tree-supported antennas must be constructed much more sturdily than is necessary with stable supports. Even with rugged construction, it is unlikely that an antenna suspended from a tree, or between trees, will stand up indefinitely. Occasional repair or replacement usually must be expected.

There are two general methods of securing a pulley to a tree. If the tree can be climbed safely to the desired level, a pulley can be attached to the trunk of the tree, as shown in **Fig 1**. To clear the branches of the tree, the antenna end of the halyard can be tied temporarily to the tree at the pulley level. Then the remainder of the halyard is coiled up, and the coil thrown out horizontally from this level, in the direction in which the antenna runs. It may help to have the antenna end of the halyard weighted.

After attaching the antenna to the halyard, the other end is untied from the tree, passed through the pulley, and brought to ground along the tree trunk in as straight a line as possible. The halyard need only be long enough to reach the ground after the antenna has been hauled up. (Additional rope can be tied to the halyard when it becomes necessary to lower the antenna.)

The other method consists of passing a line over the tree from ground level, and using this line to haul a pulley up into the tree and hold it there. Several ingenious methods have been used to accomplish this. The simplest method employs a weighted pilot line, such as fishing line or mason's chalk line. By grasping the line about two feet from the weight, the weight is swung back and forth, pendulum style, and then heaved with an underhand motion in the direction of the treetop.

Several trials may be necessary to determine the optimum size of the weight for the line selected, the distance between the weight and the hand before throwing, and the point in the arc of the swing where the line released. The weight, however, must be sufficiently large to carry the pilot line back to ground after passing over the tree. Flipping the end of the line up and down so as to put a traveling wave on the line often helps to induce the weight to drop down if the weight is marginal. The higher the tree, the lighter the weight and the pilot line must be. A glove should be worn on the throwing hand, because a line running swiftly through the bare hand can cause a severe burn.

If there is a clear line of sight between ground and a particularly desirable crotch in the tree, it may eventually be possible to hit the crotch after a sufficient number of tries. Otherwise, it is best to try to heave the pilot line completely over the tree, as close to the centerline of the tree as possible. If it is necessary to retrieve the line and start over again, the line should be drawn back very slowly; otherwise the swinging weight may wrap the line around a small limb, making retrieval impossible.

Stretching the line out straight on the ground before throwing may help to keep the line from snarling, but it places extra drag on the line, and the line may snag on obstructions overhanging the line when it is thrown. Another method is to make a stationary reel by driving eight nails, arranged in a circle, through a 1-inch board. After winding the line around the circle formed by the nails, the line should reel off readily when the weighted end of the line is thrown. The board should be tilted at approximately right angles to the path of the throw.

Other devices that have been used successfully to pass a pilot line over a tree are a bow and arrow with heavy thread tied to the arrow, and a short casting rod and spinning reel used by fishermen. The Wrist Rocket slingshot made from surgical rubber tubing and a metal frame has proved highly effective as an antenna-launching device. Still another method that has been used where sufficient space is available is flying a kite to sufficient altitude, walking around the tree until the kite string lines up with the center of the tree, and paying out string until the kite falls to the earth. This method can be used to pass a line over a patch of woods between two higher supports, which may be impossible using any other method.

The pilot line can be used to pull successively heavier lines over the tree until one of adequate size to take the strain of the antenna has been reached. This line is then used to haul a pulley up into the tree after the antenna halyard has been threaded through the pulley. The line that holds the pulley must be capable of withstanding considerable chafing where it passes through the crotch, and at points where lower branches may rub against the standing part. For this reason, it may be advisable to use galvanized sash cord or stranded guy wire for raising the pulley.

Larger lines or cables require special attention when they must be spliced to smaller lines. A splice that minimizes the chances of coming undone when coaxed through the tree crotch must be used. One type of splice is shown in **Fig 2**.

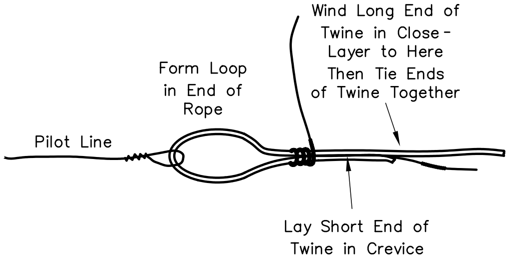

Fig 2—In connecting the halyard to the pilot line, a large knot that might snag in the crotch of a tree should be avoided, as shown.

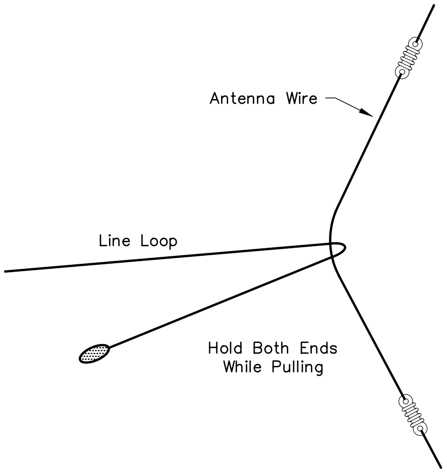

Fig 3—A weighted line thrown over the antenna can be used to pull the antenna to one side of overhanging obstructions, such as tree branches, as the antenna is pulled up. When the obstruction has been cleared, the line can be removed by releasing one end.

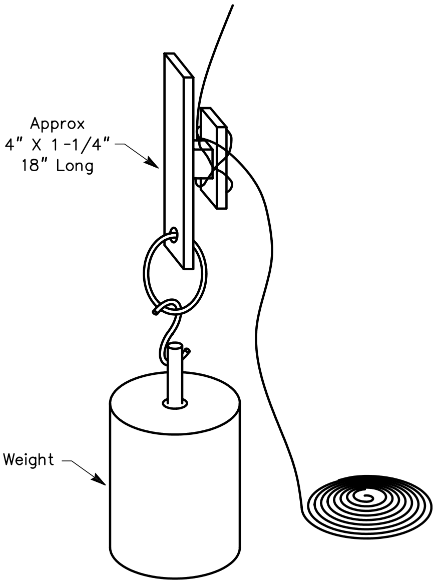

Fig 4—The cleat eliminates the need to untie a knot that may be weather hardened.

The crotch in which the line first comes to rest may not be sufficiently strong to stand up under the tension of the antenna. If, however, the line has been passed over (or close to) the center line of the tree, it will usually break through the lighter crotches and come to rest in a stronger one lower in the tree.

Needless to say, any of the suggested methods should be used with due respect to persons or property in the immediate vicinity. A child's sponge-rubber ball (baseball size) makes a safe weight for heaving a heavy thread line or fishing line.

If the antenna wire snags in the lower branches of the tree when the wire is pulled up, or if other trees interfere with raising the antenna, a weighted line thrown over the antenna and slid to the appropriate point is often helpful in pulling the antenna wire to one side to clear the interference as the antenna is being raised. This is shown in **Fig 3**.

Wind Compensation

The movement of an antenna suspended between supports that are not stable in the wind can be reduced by the use of heavy springs, such as screen-door springs under tension, or by a counterweight at the end of one halyard. This is shown in Fig 1. The weight, which may be made up of junkyard metal, window sash weights, or a galvanized pail filled with sand or stone, should be adjusted experimentally for best results under existing conditions. **Fig 4** shows a convenient way of fastening the counterweight to the halyard. It eliminates the necessity for untying a knot in the halyard, which may have hardened under tension and exposure to the weather.

TREES AS SUPPORTS FOR VERTICAL WIRE ANTENNAS

Trees can often be used to support vertical as well as horizontal antennas. If the tree is tall and has overhanging branches, the scheme of **Fig 5** may be used. The top end of the antenna is secured to a halyard passed over the limb, brought back to ground level, and fastened to the trunk of the tree.

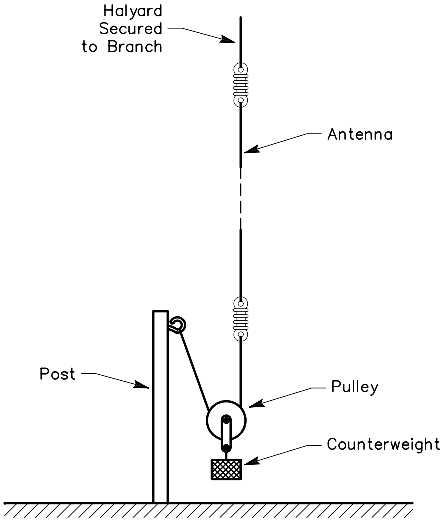

Fig 5—Counterweight for a vertical antenna suspended from an overhanging tree branch.

MAST MATERIALS

Where suitable trees are not available, or a more stable support is desired, light-duty guyed masts are suitable for wire antennas of reasonable span length. At one time, most amateur masts were constructed of lumber, but the TV industry has brought out metal masts that are inexpensive and much more durable than wood. However, there are some applications where wood is necessary or desirable.

A Ladder Mast

A temporary antenna support is sometimes needed for an antenna system for antenna testing, site selection, emergency exercises or Field Day. Ordinary aluminum extension ladders are ideal candidates for this service. They are strong, light, extendable, weatherproof and easily transported. Additionally, they are readily available and can be returned to normal use once the project is concluded. A ladder tower will support a lightweight triband beam and rotator.

With patience and ingenuity one person can erect this assembly. One of the biggest problems is holding the base down while "walking" the ladder to a vertical position. The ladder can be guyed with $^1/_4$-inch polypropylene rope. Rope guys are arranged in the standard fashion with three at each level. If help is available, the ladder can be walked up in its retracted position and extended after the antenna and rotator are attached. The lightweight pulley system on most extension ladders is not strong enough to lift the ladder extension. This mechanism must be replaced (or augmented) with a heavy-duty pulley and rope. Make sure when attaching the guy ropes that they do not foul the operation of the sliding upper section of the ladder.

There is one hazard in this system that must be avoided: Do not climb or stand on the ladder when it is being extended—even as much as one rung. Never stand on the ladder and attempt to raise or lower the upper section. Do all the extending and retracting with the heavy-duty rope and pulley!

If the ladder is to be raised by one person, use the following guidelines. First, make sure the rung-latching mechanism operates properly before beginning. The base must be hinged so that it does not slip along the ground during erection. The guy ropes should be tied and positioned in such a way that they serve as safety constraints in the event that control of the assembly is lost. Have available a device (such as another ladder) for supporting the ladder during rest periods. (See **Fig 6**.)

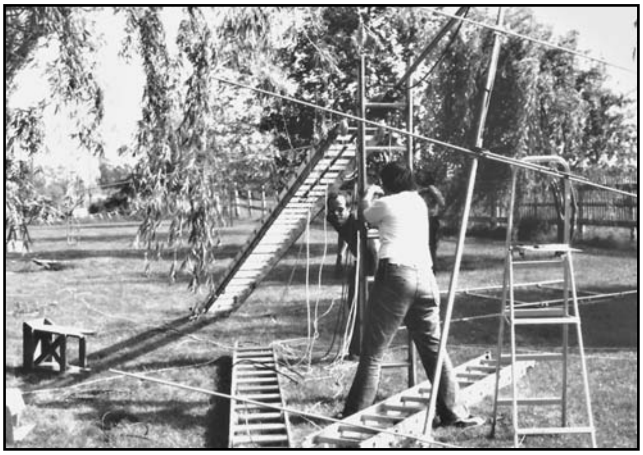

Fig 6—Walking the ladder up to its vertical position. Keith, VE2AQU, supports the mast with a second ladder while Chris, VE2FRJ, checks the ropes. (*Photo by Keith Baker, VE2XL.*)

After the ladder is erect and the lower section guys tied and tightened, raise the upper portion one rung at a time. Do not raise the upper section higher than it is designed to go; safety is far more important than a few extra feet of height.

For a temporary installation, finding suitable guy anchors can be an exercise in creativity. Fence posts, trees, and heavy pipes are all possibilities. If nothing of sufficient strength is available, anchor posts or pipes can be driven into the soil. Sandy soil is the most difficult to work with because it does a very poor job of holding anchors. A discarded car axle can be driven into the ground as an anchor, as its mass and strength are substantial. A chain and car-bumper jack can be used to remove the axle when the operation is done.

Above all else, keep the tower and antenna away from power lines. Make sure that nothing can touch the lines if the assembly falls. Disassemble by reversing the process. Ladder towers are handy for "quickie" antenna supports, but as with any improvisation of support materials, care must be taken to ensure safe construction.

The A-Frame Mast

A light and relatively inexpensive mast is shown in **Fig 7**. In lengths up to 40 feet it is very easy to erect and will stand the pull of ordinary wire antenna systems. The lumber used is 2 × 2-inch straight-grained pine (which many lumber yards know as hemlock) or even fir stock. The uprights can be as long as 22 feet each (for a mast slightly over 40 feet high) and the cross pieces are cut to fit. Four pieces of 2 × 2 lumber, each 22 feet long, provides more than enough. The only other materials required are five $^1/_4$-inch carriage bolts $5^1/_2$ inches long, a few spikes, about 300 feet of stranded or solid galvanized wire for guying, enough glazed porcelain compression ("egg") insulators to break up the guys into sections, and the usual pulley and halyard rope. If the strain insulators are put in every 20 feet, approximately 15 of them will be enough.

After selecting and purchasing the lumber—which should be straight-grained and knot-free—sawhorses or boxes should be set up and the mast assembled as shown in **Fig 8**. At this stage it is wise to give the mast a coat of primer and a coat of outside white latex paint.

After the coat of paint is dry, attach the guys and rig the pulley for the antenna halyard. The pulley anchor should be at the point where the top stays are attached so the backstay will assume the greater part of the load tension. It is better to use wire wrapped around the mast with a small through-bolt to prevent sliding down than to use eyebolts.

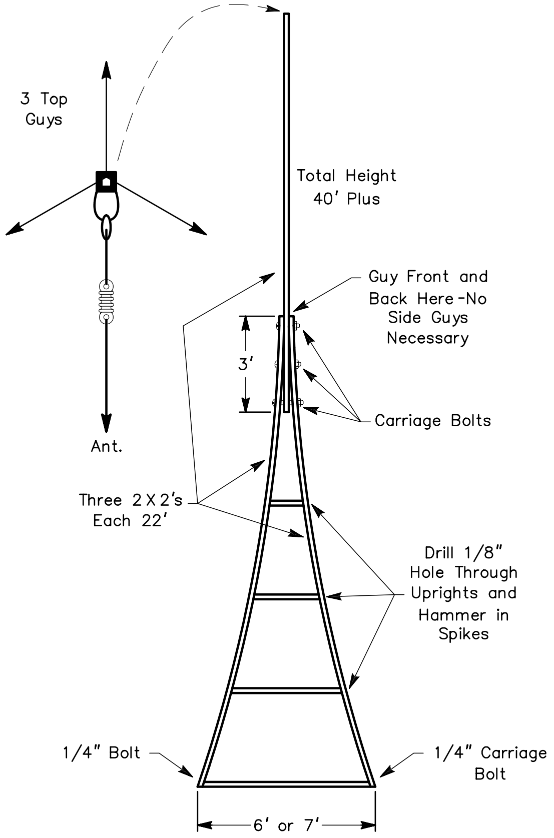
3 Top Guys
Total Height 40′ Plus
Guy Front and Back Here -No Side Guys Necessary
3′
Carriage Bolts
Ant.
Three 2 X 2′s Each 22′
Drill 1/8″ Hole Through Uprights and Hammer in Spikes
1/4″ Bolt
1/4″ Carriage Bolt
6′ or 7′

Fig 7—The A-frame mast is lightweight and easily constructed and erected.

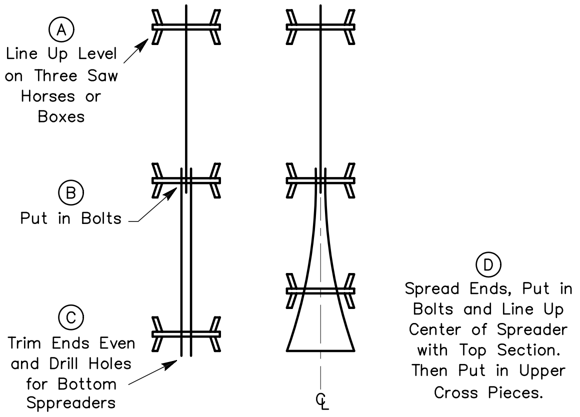
A
Line Up Level on Three Saw Horses or Boxes
B
Put in Bolts
C
Trim Ends Even and Drill Holes for Bottom Sppreaders
D
Spread Ends, Put in Bolts and Line Up Center of Spreader with Top Section. Then Put in Upper Cross Pieces.
℄

Fig 8—Method of assembling the A-frame mast on sawhorses.

If the mast is to stand on the ground, a couple of stakes should be driven to keep the bottom from slipping. At this point the mast may be "walked up" by a helper. If it is to go on a roof, first stand it up against the side of the building and then hoist it, from the roof, keeping it vertical. The whole assembly is light enough for two men to perform the complete operation—lifting the mast, carrying it to its permanent berth, and fastening the guys with the mast vertical. It is entirely practical to put up such a mast on a flat area of roof that would be too small to erect a regular tower installation, one that had to be raised vertically on the same spot.

TV Mast Material

TV mast is available in 5- and 10-foot lengths, $1^1/_4$ inches diameter, in both steel and aluminum. These sections are crimped at one end to permit sections to be joined together. A form that is usually more convenient is the telescoping mast available from many electronic supply houses. The masts may be obtained with three, four or five 10-foot sections, and come complete with guying rings and a means of locking the sections in place after they have been extended. These masts are inherently more suitable for guyed mast installations than the non-telescoping type because the diameters of the sections increase toward the bottom of the mast. For instance, the top section of a 50-foot mast is $1^1/_4$ inches diameter, and the bottom section is $2^1/_2$ inches diameter.

Guy rings are provided at 10-foot intervals, but guys may not be required at every point. Guying is essential at the top and at least one other place near the center of the mast. If the mast has any tendency to whip in the wind, or to bow under the load of a horizontal wire antenna, additional guys should be added at the appropriate points.

MAST GUYING

Three guy wires in each set are usually adequate for a mast. These should be spaced equally round the mast. The required number of sets of guys depends on the height of the mast, its natural sturdiness (or stiffness), and the required antenna tension. A 30-foot-high mast usually requires two sets of guys, and a 50-foot mast needs at least three sets. One guy of the top set should be anchored to a point directly opposing the force exerted by the wire antenna. The other two guys of the same set should be spaced 120° with respect to the first, as shown in the inset in Fig 7.

Generally, the top guys should be anchored at distances from the base of the mast at least 60% of the mast height. The distance of the guy anchors from the mast determines the guy loads and the vertical load compressing the mast. At a 60% distance, the load on the guy wire opposite the wire antenna is approximately twice the antenna tension. The compression in the mast will be 1.66 times the antenna tension. With the anchors out 80% of the mast height, the guy tension will be 1.6 times larger than the antenna load and the mast compression will be 1.25 times larger.

Whenever possible, the largest available anchor spacing should be used. The additional compression on the mast, due to closer anchor spacing, increases the tendency of the mast to buckle. Buckling occurs when the compression on the unsupported spans between guys become too great for the unsupported length. The section then bows out laterally and will usually fold over, collapsing the mast. Additional sets of guys reduce the tendency for the mast to buckle under the compression by decreasing the unsupported span lengths and stabilizing the mast, keeping it in a straight line.

A natural phenomenon, called *vortex shedding,* can occur when the wind passes over the sections of a guyed mast. For every section size, shape, and length, there is a wind speed that can cause the sections to oscillate mechanically. When all the sections of an antenna support mast are close to the same size and length, it is possible for all of the mast sections to vibrate simultaneously between the guys. To reduce the potential for this, you can place the guys at locations along the mast that will result in different span lengths. This creates different mechanical resonant frequencies for each span, eliminating the possibility of all sections oscillating at the same time.

When determining the guy locations along the mast to treat this problem, you also need to consider the mast buckling requirements. Since the compression in the mast is greatest in the bottom span, and the least in the top span, the guys should be placed to make the bottom span the shortest and the top span the longest. A general guide for determining the different span lengths is to make the unguyed lengths change by 10 to 20%.

Example: For a 30-foot high mast with three guy sets, the equal-guy locations would be every 10 feet. We can make the center span, 10 feet long, and then make the lower span 15% shorter and the top span 15% longer. While this is not an exact technical method to determine the best solution, the approach will create different mechanical resonant frequencies for the spans, with the span lengths approximately adjusted for the varying buckling requirements.

You can eliminate electrical resonance from conductive guy materials that might cause distortion of the antenna radiation pattern by breaking each guy into nonresonant lengths using strain insulators (see **Figs 9** and **10**). This subject is covered in detail later in this chapter.

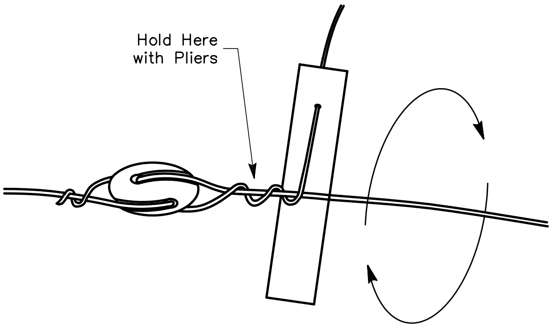

Fig 9—Simple lever for twisting solid guy wires when attaching strain insulators.

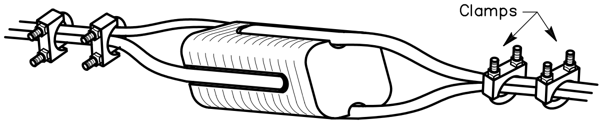

Fig 10—Stranded guy wire should be attached to strain insulators by means of standard cable clamps made to fit the size of wire used.

Guy Material

When used within their safe load ratings, you may use any of the halyard materials listed in Chapter 20 for the mast guys. Nonmetallic materials have the advantage that there is no need to break them up into sections to avoid unwanted resonant interactions. All of these materials are subject to *stretching,* however, which causes mechanical problems in permanent installations. At rated working loads, dry manila rope stretches about 5%, while nylon rope stretches about 20%. Usually, after a period of wind load and wet/dry cycles, the lines will become fairly stable and require less frequent adjustment.

Solid galvanized steel wire is also widely used for guying. This wire has approximately twice the load ratings of similar sizes of copper-clad wire, but it is more susceptible to corrosion. Stranded galvanized wire sold for guying TV masts is also suitable for light-duty applications, but is also susceptible to corrosion. It is prudent to inspect the guys every six months for signs of deterioration or damage.

Guy Anchors

Figs 11 and **12** show two different kinds of guy anchors. In Fig 11, one or more pipes are driven into the ground at right angles to the guy wire. If a single pipe proves to be inadequate, another pipe can be added in tandem, as shown, and connected with a galvanized steel cable. Heavy-gauge galvanized pipe is preferred for corrosion resistance. Steel fence posts may be used in the same manner. Fig 12 shows a *dead-man* type of

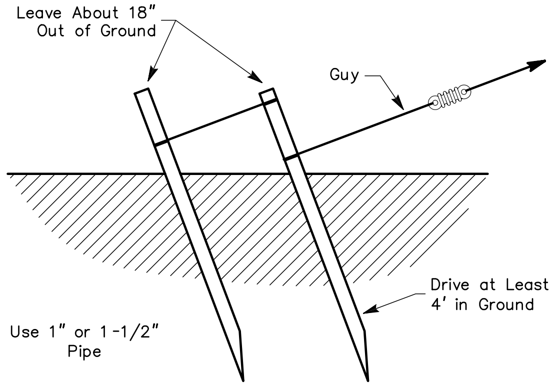

Fig 11—Driven guy anchors. One pipe is usually sufficient for a small mast. For added strength, a second pipe may be added, as shown.

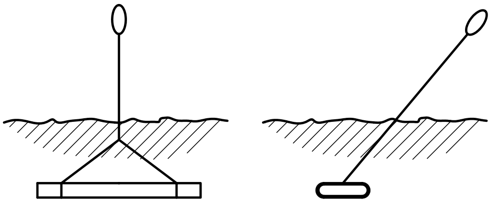

Fig 12—Buried *dead-man* guy anchor (see text).

anchor. The buried anchor may consist of one or more pipes 5 or 6 feet long, or scrap automobile parts, such as bumpers or wheels. The anchors should be buried 3 or 4 feet in the ground. The cable connecting the dead-man to the guys should be galvanized wire rope, like EHS guy cable. You should coat the buried part of the cable with roofing tar, and thoroughly dry it prior to burial to enhance resistance to corrosion.

Also available are some heavy auger-type anchors that screw into the earth. These anchors are usually heavier than required for guying a mast, although they may be more convenient to install. You should conduct annual inspections of the anchors by digging several inches below grade around the anchor to inspect for corrosion.

Trees and buildings may also be used as guy anchors if they are located appropriately. Care should be exercised, however, to make sure that the tree is of adequate size and that any fastening to a building can be made sufficiently secure.

Guy Tension

Many troubles encountered in mast guying are a result of pulling the guy wires too tight. Guy-wire tension should never be more than necessary to correct for obvious bowing or movement under wind pressure. Approximately 10% to 15% of the working load is sufficient. In most cases, the tension needed does not require the use of turnbuckles, with the possible exception of the guy opposite a wire antenna. If any great difficulty is experienced in eliminating bowing from the mast, the guy tension should be reduced or additional sets of guys are required. The mast should be checked periodically, especially after large wind events, to ensure the guys and anchors have not stretched or moved, allowing the mast to get away from the required straight alignment.

ERECTING A MAST OR OTHER SUPPORT

Masts less than 30 feet high usually can be simply walked up after blocking the bottom end securely. Blocking must be done so that the base can neither slip along the ground nor upend when the mast is raised. An assistant should be stationed at each guy wire, and may help by pulling the proper guy wire as the mast nears the vertical position. Halyards can be used in the same manner.

As the mast is raised, it may be helpful to follow the underside of the mast with a scissors rest (**Fig 13**), should a pause in the hoisting become necessary. The rest may also be used to assist in the raising if an assistant mans each leg.

As the mast nears the vertical position, those holding the guy wires should be ready to temporarily fasten the guys to prevent the mast from falling. The guys can then be adjusted until the mast is perfectly straight.

For masts over 30 feet long, a *gin pole* of some form may be required, as shown in Fig 13. Several turns of

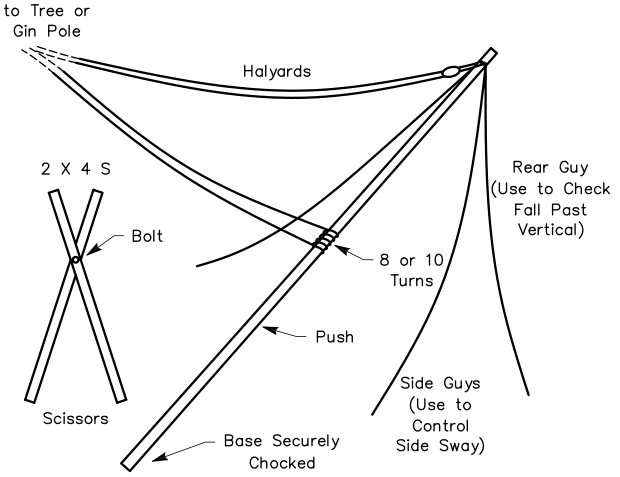

Fig 13—Pulling on a gin line fastened slightly above the center point of the mast and on the halyards can assist in erecting a tall mast. The tensions should be just enough to keep the mast in as straight a line as possible. The "scissors" may be used to push on the under side and to serve as a rest if a pause in raising becomes necessary.

rope are wound around a point on the mast above center. The ends of the rope are then brought together and passed over a tree limb. The rope should be pulled as the mast is walked up to keep the mast from bending at the center. If a tree is not available, a post, such as a 2 × 4, temporarily erected and guyed, can be used. After the mast has been erected, the assisting rope can be removed by walking one end around the mast (inside the guy wires).

Telephone poles and towers are much sturdier supports. Such supports may require no guying, but they are not often used solely for the support of wire antennas because of their relatively high cost. For antenna heights in excess of 50 feet, however, they are usually a most practical form of support.

Tower And Antenna Selection and Installation

The selection of a tower, its height, and the type of antennas and rotator is probably one of the more complex issues faced by station builders. All aspects of the tower, antenna, and rotator system are interrelated, and you should consider the overall system before making any decisions regarding specific system components.

Perhaps the most important consideration for many amateurs is the effect of the antenna system on the surrounding environment. If plenty of space is available for a tower installation and if there is little chance of causing esthetic distress on the part of family members or the neighbors, the amateur is indeed fortunate. Often, the primary considerations are purely financial. For most, however, the size of the property, the effect of the system on others, local ordinances, and the proximity of power lines and poles influence the selection of the tower/antenna system considerably.

The amateur must consider the practical limitations for installation. Some points for consideration are given below:

1) A tower should not be installed in a position where it could fall onto a neighbor's property.
2) The antenna must be located in such a position that it cannot possibly tangle with power lines, either during normal operation or if the structure should fall.
3) Sufficient yard space must be available to position a guyed tower properly. The guy anchors should be between 60% and 80% of the tower height in distance from the base of the tower on level ground—sloping terrain may require larger areas.
4) Provisions must be made to keep children from climbing the support. (Poultry netting around the tower base will serve this need.)
5) Local ordinances should be checked to determine if any legal restrictions affect the proposed installation.

Other important considerations are (1) the total dollar amount to be invested, (2) the size and weight of the antenna desired, (3) the climate, and (4) the ability of the owner to climb a fixed tower.

Most tower manufacturers provide catalogues or data packages that represent engineered tower configurations. These are provided as a convenience for users to help determine the most suitable tower configurations. The most commonly used design specifications for towers are EIA (Electronic Industries Assoc.) RS-222 and UBC (Uniform Building Code). These specifications define how the tower, antenna, and guy loads are determined and applied to the system, and establish general design criteria for the analysis of the tower. Local authorities often require the review and approval of the installation by a state licensed Professional Engineer (P.E.) to obtain building permits. All local authorities in the United States do not subscribe to the same design standards, so often the manufacturers' general-purpose engineering is not applicable.

One of the first things you need to determine in the tower selection process is the type of specification required by the local authorities, if any. Then, you must determine the *Basic Wind Speed* appropriate for the site. The *Basic Wind Speed* used in most specifications is the average wind speed for one mile of wind passing across the structure. It will be a lower value than the peak readings from an anemometer (wind gauge) installed at the site. For example, a Basic Wind Speed of 70 mph could have a maximum value of 80 mph and a minimum of 60 mph, equally distributed during the passage of the mile of wind. Basic wind speeds can be found in tables or maps contained in the appropriate specifications. Often, the basic wind speed used for the location may be obtained from the local permit authority. Check out the Web site at **www.championradio.com**, which contains EIA basic wind speed tables for every county in the USA. UBC speeds are available at almost every local library.

Antenna manufacturers also provide antenna data to assist in the selection process. Unfortunately, antenna mechanical designs do not always follow the same design standards used for towers. Proper antenna selection often means that you must determine the antenna surface areas yourself to avoid overloading the tower. More discussion about this follows later in this chapter.

It is often very helpful to the novice tower installer to visit other local amateurs who have installed towers. Look over their hardware and ask questions. If possible, have a few local experienced amateurs look over your plans—before you commit yourself. They may be able to offer a great deal of help. If someone in your area is planning to install a tower and antenna system, be sure to offer your assistance. There is no substitute for experience when it comes to tower work, and your experience there may prove invaluable to you later.

THE TOWER

Towers for supporting antennas come in a variety of different types. Each type has its own set of benefits and limitations, or conditions and requirements. Often, you can choose a particular tower type by considering issues other than pure mechanical performance. Understanding how each type of tower functions, and what its respective requirements are, are the first steps in making the best tower selection for your own situation.

Guyed Towers

The most common variety of tower is the *guyed tower* made of identical stacked sections, supported by guy cables attached to ground anchors placed symmetrically around the tower. These towers are the most economical, in terms of feet per dollar investment, and are more efficient for carrying antenna loads than non-guyed towers.

The guys resist the lateral loads on the system created by the wind. Since the guys slope down to the ground, horizontal loads due to the wind result in vertical loads applied to the tower at each tower/guy connection. The tower becomes a compression member, trying to resist the column compression generated by the guy reactions. A tower in compression can buckle, so the distance between guy connections along the tower is important.

Tower Bases for Guyed Towers

Another important phenomenon in a guyed tower is stretching of the guy cables. All guys stretch under load and when the wind blows the elongated guys allow the tower to lean over somewhat. If the tower base is buried in the concrete footing—as is commonly done in amateur installations—the bending stress at the tower base can become a significant factor. Towers that have been installed with tapered pier-pin bases much more freely absorb tower leaning, and they are far less sensitive to guy-elongation problems.

The tapered pier-pin tower installation is not without some drawbacks. These installations often require torque-arm guy brackets or six-guy torque-arm assemblies to control tower rotation due to antenna torque. They also require temporary guys when they are being installed to hold the base steady until the permanent guys are mounted. Some climbers also don't like the flexing when they start to climb these types of towers.

On the positive side, pier-pin base towers have all structural members above the concrete footing, eliminating concerns about hidden corrosion that can occur with buried towers. Most decisions regarding the type of base installation are made according to the preference of the tower builder/maintainer. While either type of base configuration can be successfully used, you would be wise to do the stress calculations (or have a professional engineer do them) to ensure safety, particularly when large antenna loads are contemplated and particularly if guys that can easily stretch are used, such as Phillystran guys.

The configuration shown in **Fig 14A** is taken from an older (1983) Unarco-Rohn catalog. This configuration has the top set of guys placed at the top of the tower with the lower set halfway up the tower. This configuration is best for most amateur installations, which usually have the antennas mounted on a rotatable mast extending

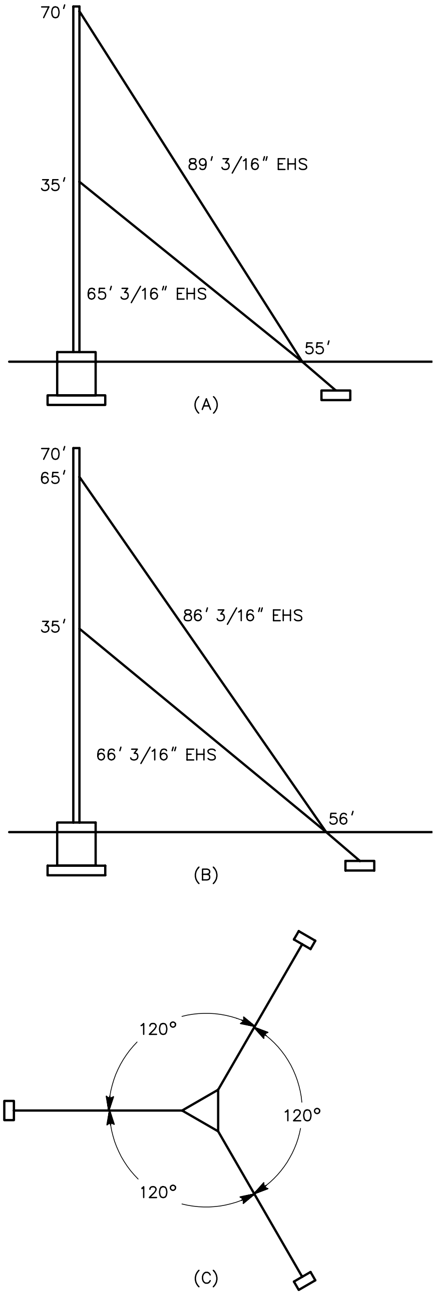

Fig 14—The proper method of installation of a guyed tower. At A, the method recommended for most amateur installations. At B, the method shown in later Rohn catalogs. This places considerable strain on the top section of the tower when large antennas are mounted on the tower.

out the top of the tower—thereby placing the maximum lateral loads when the wind blows at the top of the tower (and the bottom of the rotating mast).

The configuration shown in Fig 14B is from a newer (1998) Rohn catalog. It has 5 feet of unsupported tower extending above the top guy set. The lower guy set is approximately halfway between the top guys and the base. The newer configurations are tailored for commercial users who populate the top region of the tower with fixed arrays and/or dishes. The installation in Fig 14B cannot safely withstand the same amount of horizontal top load as can the configuration shown in Fig 14A, simply because the guys start farther down from the top of the tower.

An overhead view of a guyed tower is given in Fig 14C. Common practice is to use equal angular spacings of 120° between guy wires. If you must deviate from this spacing, the engineering staff of the tower manufacturer or a civil engineer should be contacted for advice.

Amateurs should understand that most catalogs show generic examples of tower configurations that work within the cited design specifications. They are by no means the only solution for any specific tower/antenna configuration. You can usually substantially change the load capability of any given tower by varying the size and number of guys. Station builders are encouraged to utilize the services of professional engineers to get the most out of their guyed towers. Those interested in more generic information about guyed tower behavior can find it at **www.freeyellow.com/members3/yagistress/**.

Unguyed Towers

Another commonly used type of tower is not normally guyed—these are usually referred to as *freestanding* or *self-supporting* towers. Unguyed towers come in three different styles.

One style is comprised of stacked lengths of identical tower sections, just like those used for guyed towers. The only difference is that no guys are used. Manufacturers provide the recommended configurations and allowable loads for this type of installation in their catalogs. Unguyed towers are vastly less capable of supporting antenna loads than their guyed counterparts, but have great utility for light-duty applications—when configured within their capabilities.

The second style utilizes different tower section sizes, varying from large sections at the base and tapering down to smaller sections at the top. This style is more much more efficient for freestanding applications, because the tower is sized for the varying bending loads along the tower length, and is shown in **Fig 15**.

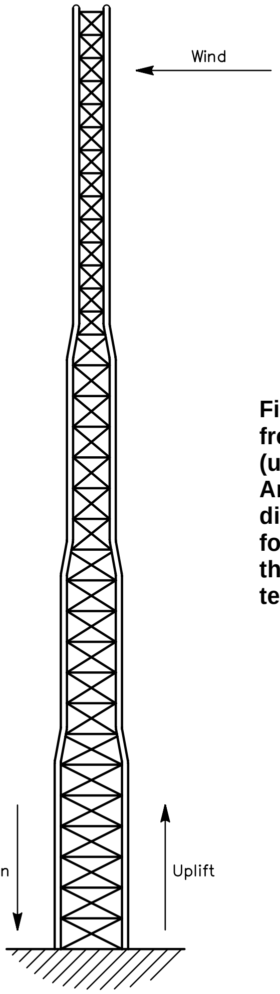

Fig 15—Typical freestanding (unguyed) tower. Arrows indicate the directions of the forces acting upon the structure. See text for discussion.

The third style of unguyed towers is commonly called a *crank-up tower*. It is a freestanding tower with telescoping sections that can be extended or retracted with a winch, cable, and pulley mechanism. This allows the tower to be raised and lowered for maintenance and antenna work. It is usually necessary to retract such towers for moderate to heavy winds. Some consider this a disadvantage because they can't operate their antennas at full height when it is windy. Two different forms of the crank-up style, freestanding tower are shown in **Fig 16**. Fig 16A shows the tubular version; Fig 16B shows the triangular space-frame version.

Some crank-up towers are used with guys and are only retracted for maintenance and antenna work. These towers are specially designed with locking mechanisms between the tower sections to carry the vertical compression created by the guys. *Do not* use guys with normal crank-up towers (those that have no locking devices between sections)! The increased tower compression will be carried by the hoisting cable, which will eventually cause it to fail.

Never climb a crank-up tower unless it is properly nested, with all load removed from the hoisting cable. For general antenna work, this can be accomplished by completely retracting it until the cable becomes loose. When servicing the rotator, the tower must be left partially extended. In this case every tower section must be

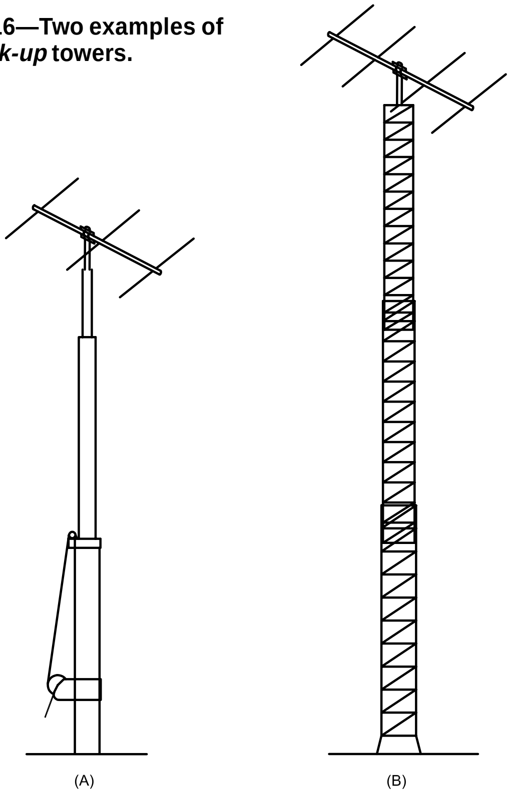

Fig 16—Two examples of *crank-up* towers.

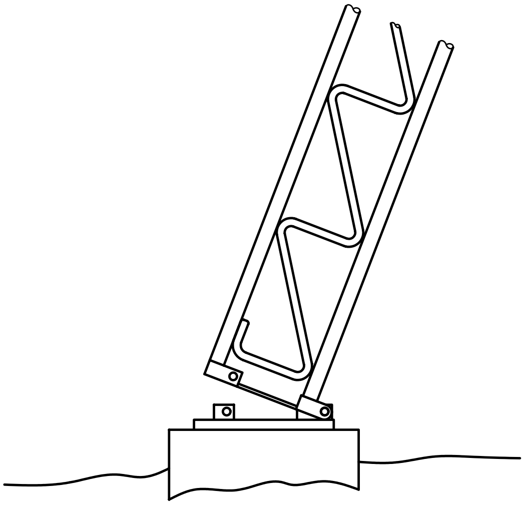

Fig 17—Fold-over or tilting base. There are several different kinds of hinged sections permitting different types of installation. Great care should be exercised when raising or lowering a tilting tower.

blocked with heavy timber or thick-wall tubes, installed through the tower bracing, until all sections are resting on the blocks and the hoisting cable becomes slack. Safely installing the blocks in an extended crank-up tower can be challenging. The object is to get all the blocks installed without a climber having to scale the unblocked tower, risking loss of limbs should the hoisting cable fail. An extension ladder, capable of reaching the required block elevations is the safest approach. If the necessary equipment or expertise is not available, the tower can be retracted, antennas removed, and leaned over, with the base tilt-over assembly, before extending it to access the rotator. Failure to properly block the tower before climbing can result in serious injury should the cable slip or break!

All freestanding towers share some unique characteristics. Each must support antenna and tower loads only by virtue of the bending strength of the tower sections and the tower footing connection to earth. Because of the large overturning moment at the tower base, freestanding towers require larger concrete footings than guyed towers. They are usually more expensive for the same load capability compared to guyed towers, simply because they require larger heavier tower sections and a larger footing to get the job done. The telescoping mechanisms in crank-up tower require more maintenance too.

Freestanding towers are quite popular, and are often the best solutions for sites with limited space and ascetic concerns. When cranked down, a telescoping tower can maintain a low-profile system, out of sight of the neighbors and family.

Tilt-Over Towers

Some towers have another convenience feature—a hinged section that permits the owner to fold over all or a portion of the tower. The primary benefit is in allowing antenna work to be done close to ground level, without the necessity of removing the antenna and lowering it for service. **Fig 17** shows a hinged base used with stacked, guyed tower sections. The hinged section can be designed for portions of the tower above the base. These are usually referred to as *guyed tilt-over towers*, where a conventional guyed tower can be tilted over for installing and servicing antennas. Many crank-up towers come with optional tilt-over base fixtures that are equipped with a winch and cable system for tilting the fully nested tower from horizontal to vertical positions.

Misuse of hinged sections during tower erection is a dangerously common practice among radio amateurs. Unfortunately, these episodes can end in accidents. If you do not have a good grasp of the fundamentals of physics, it might be wise to avoid hinged towers or to consult an expert if there are any questions about safely installing and using such a tower. It is often far easier (and safer) to erect a regular guyed tower or self-supporting tower with gin pole and climbing belt than it is to try to walk up an unwieldy hinged tower.

The AB-577 Military Surplus Tower

Another light duty tower has found acceptance among many amateurs. Available from assorted military surplus dealers is the AB-577 system. This was designed to be a portable, rapid deployment antenna support for field communications. It is a guyed mast that goes up somewhat like a crank-up. The system consists of several short sections of aluminum tubing, with special end connections for joining them. These can be erected from the base fixture, which has a crank-up type winch-driven elevator platform. The tubing sections are installed in the base fixture and connected to the section above it with an over-center locking *Marmon*-style clamp. Then, the elevator platform is raised with the winch and the new tube is locked in place, high on the base fixture. Then the elevator is lowered to accept the next section. While the tower is extended, the supporting guys are adjusted via the unique *snubber* assemblies at the anchor connection. One person can erect this system, even in windy conditions, when special care is given to keeping the guys properly adjusted during each extension.

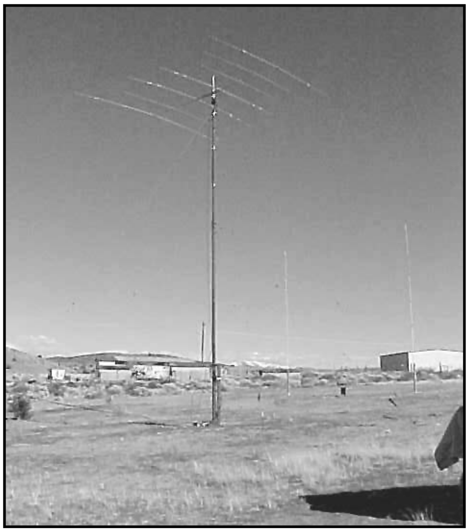

Fig 18—Installation of surplus AB-577 tower with tribander at 45 feet at K7NV. (*Photo by Kurt Andress, K7NV.*)

The standard AB-577 system, with 3 sets of guys, will support a modest triband Yagi at 45 feet. **Fig 18A** shows an installation with a Hy-Gain TH7DX at 45 feet.

TOWER BASES

Tower manufacturers can provide customers with detailed plans for properly constructing tower bases. **Fig 19** is an example of one such plan. This plan calls for a hole that is $3^1/_2 \times 3^1/_2 \times 6$ feet. Steel reinforcement bars are lashed together and placed in the hole. The bars are positioned so that they will be completely embedded in the concrete, yet will not contact any metallic object in the base itself. This is done to minimize the possibility of a direct discharge path for lightning through the base. Should such a lightning discharge occur, the concrete base could be damaged.

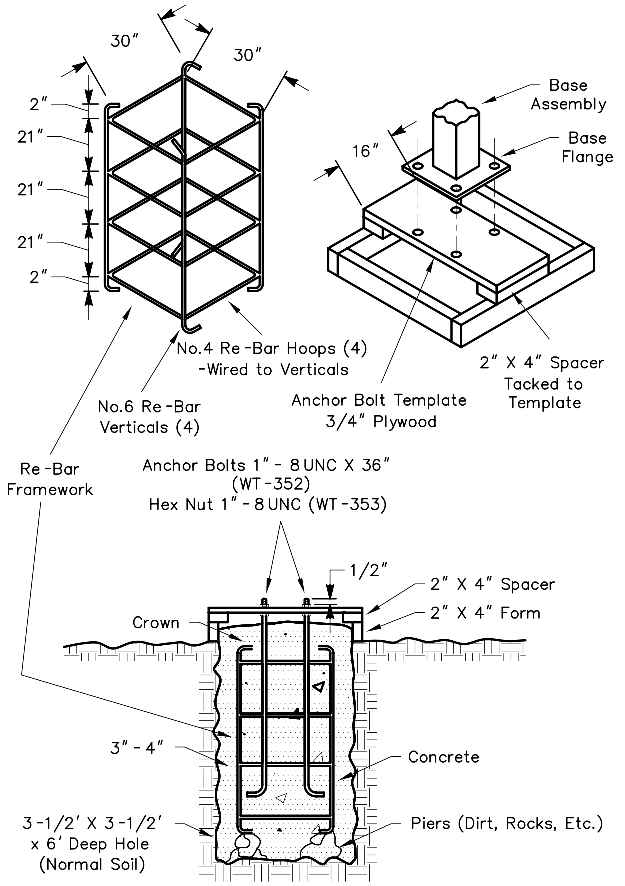
30"
30"
2"
21"
21"
21"
2"
Base Assembly
Base Flange
16"
No.4 Re-Bar Hoops (4) -Wired to Verticals
No.6 Re-Bar Verticals (4)
Anchor Bolt Template 3/4" Plywood
2" X 4" Spacer Tacked to Template
Re-Bar Framework
Anchor Bolts 1" - 8 UNC X 36" (WT-352)
Hex Nut 1" - 8 UNC (WT-353)
1/2"
2" X 4" Spacer
2" X 4" Form
Crown
3" - 4"
Concrete
3-1/2' X 3-1/2' x 6' Deep Hole (Normal Soil)
Piers (Dirt, Rocks, Etc.)

Fig 19—Plans for installing concrete base for Wilson ST-77B tower. Although the instructions and dimensions vary from tower to tower, this is representative of the type of concrete base specified by most manufacturers.

Providing suitable paths for the discharge of lightning energy safely for towers is a complex subject. Several companies offer products and guidance. The basic requirements for providing controlled discharge paths for lightning-induced current is to supply a low-impedance grid of conductors from the tower and feed lines to a field of interconnected ground rods around the base of the tower. Generally, the tower, station, and electrical service grounds need to be connected to prevent damaging potential differences from developing between the various components in the system.

A strong wooden form is constructed around the top of the hole. The hole and the wooden form are filled with concrete so that the resultant block will be 4 inches above grade. The anchor bolts are embedded in the concrete, and aligned with the plywood template, before it hardens. The template serves to align the anchor bolts to properly mate with the tower itself. Once the concrete has

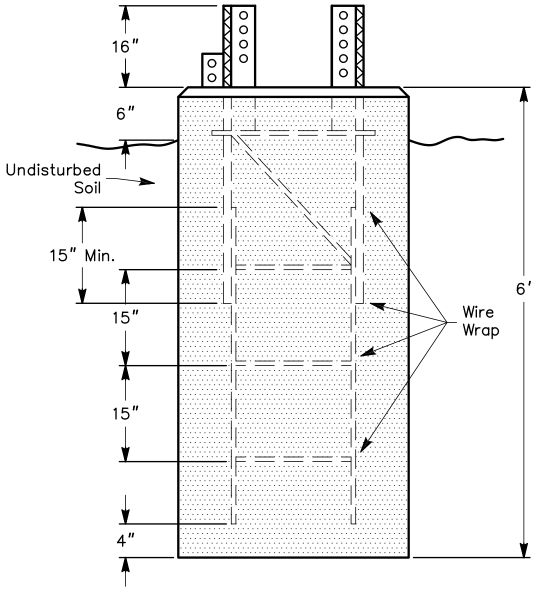

Fig 20—Another example of a concrete base (Tri-Ex LM-470).

cured, the tower base is installed on the anchor bolts and the base connection is adjusted to bring the tower into vertical alignment.

For a tower that bolts to a flat base plate mounted to the footing bolts (as shown in Fig 19), you can bolt the first tower section on the base plate to ensure that the base is level and properly aligned. Use temporary guys to hold things exactly vertical while the concrete cures. (The use of such temporary guys also works well when you place the first tower section in the base hole and plumb it vertically before pouring in the concrete.) Manufacturers can provide specific, detailed instructions for the proper mounting procedure. **Fig 20** shows a slightly different design for a tower base.

The one assumption so far is that *normal* soil is predominant in the area in which the tower is to be installed. Normal soil is a mixture of clay, loam, sand and small rocks. More conservative design parameters for the tower base should be adopted (usually, using more concrete) if the soil is sandy, swampy or extremely rocky. If there are any doubts about the soil, the local agricultural extension office can usually provide specific technical information about the soil in a given area. When this information is in hand, contact the engineering department of the tower manufacturer or a civil engineer for specific recommendations with regard to compensating for any special soil characteristics.

TOWER INSTALLATION

The installation of a tower is not difficult when the proper techniques are used. A guyed tower, in particular, is not hard to erect, because each of the individual sections is relatively lightweight and can be handled with only a few helpers and some good quality rope.

The Gin Pole and Tips on Tower Building

An essential piece of hardware for working on towers is a *gin pole*. This section came from the ARRL book *Simple and Fun Antennas for Hams*. The dictionary describes a gin pole as “a device for moving heavy objects.” See **Fig 21**, which shows a drawing of the Rohn “Erection Fixture” EF2545. This gin pole was designed to work with the nominal 10-foot long sections of Rohn 25 or 45 towers.

We’re going to assume in the following discussion that you are installing Rohn 45, which weighs about 70 pounds. This is a lot of weight, and you must refrain from adding to that during installation. That means, for example, that you *do not* attempt to lift a 10-foot section with the guy wires attached! Neither should you attempt to lift the top section with the rotator and rotor shelf installed. The gin pole (and your ground crew) will not appreciate all that strain.

The main working part of the gin pole is the pulley

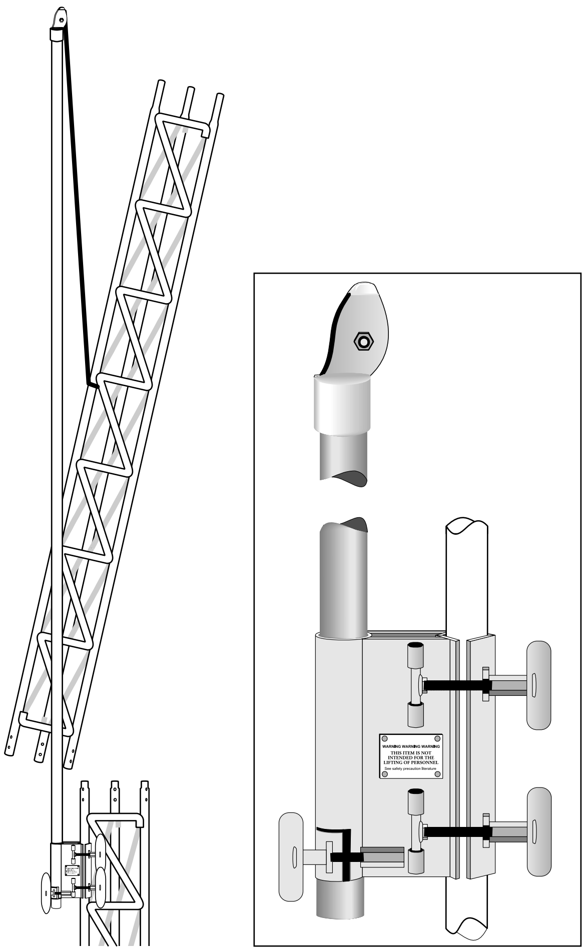

Fig 21—Drawing of Rohn “Erection Fixture” EF2545, also known commonly as a “gin pole.”

mounted at the top of the 12-foot long heavy-wall aluminum tubing. This pulley has a rope going down to the ground crew through the center of the aluminum tube. At the base of the tower, the pull rope should be run through a *snatch block* attached to the tower just above ground level. This block allows the pull rope to be pulled out horizontally away from the tower base. That helps protect ground crew should a tool be dropped by the people on the tower.

An adjustable, sliding clamp towards the bottom of the aluminum tubing is clamped to the tower using a swinging L-bracket-type clamp with two clamping bolts. These have T-bar handles that can be tightened by hand. In fact, this gin pole can be moved and deployed without any tools. The clamp is positioned on the top of the tower section onto which the next tower section is to be installed. Once clamped to the top of the tower, you would loosen the T-bar handle that tightens the clamp against the sliding aluminum tube and slide the tubing up to its maximum extent.

In practice, the following steps are taken as each 10-foot section of tower is installed, one-by-one. We're assuming here that the gin pole starts out on the ground, with at least one person belted in at the top of the tower. We're also assuming that the pull rope has been threaded through the aluminum tube and the top pulley, with a knot tied to prevent it from falling back down the tube.

1. The clamp holding the aluminum tubing is loosened so that the pulley on the tube can be lowered to where it is just above the bottom clamp. Then the T-bar handle for the tube clamp is tightened.
2. The climber lowers a tag rope for the ground crew to tie to the gin-pole pull rope. (This tag rope has been looped through a temporary pulley clipped to the top of the tower. It is also used to pull up tools and other materials.) The ground crew then pulls the gin pole up to the climber, using the tag line rope. Friction of the rope against the top of the pulley-head assembly will prevent the gin-pole assembly from slipping down. [Note that some climbers prefer to "walk" the gin-pole up the tower rather than having it pulled up from the ground below. They free up their hands for climbing and temporarily holding the gin pole by using their belt lanyard looped around the tower as they climb.]
3. Once the gin-pole head reaches the top of the tower, the climber clamps the gin pole clamp securely to the top of the tower.
4. The T-bar handle for the tube clamp is loosened, and the aluminum tube is extended to its maximum height, as shown in Fig 21. Make sure you have tied the free end of the rope coming through the top pulley temporarily to the top of the tower, or else you'll have to lower the gin pole and go through this step again.
5. The free end of the pulley rope is then dropped to the ground, often using a weight such as a medium Crescent wrench or perhaps a hammer to keep the rope from waving about as it dangles down the tower, tangling with every imaginable thing as it proceeds downwards. It's amazing how even a tiny breeze can make an unweighted rope dance like that.
6. The ground crew then ties the free end of the rope *above* the balance point of the tower. For Rohn 25 or 45 there are eight horizontal cross braces per section. You want the crew to tie the rope to the fifth horizontal brace from the bottom. Please remember that you want the bottoms of the tower sections' legs to be pointed downwards, not flipped over, when the bottom of these legs approach you at the top of the tower section the climber is standing on.
7. Once the bottom of the rising tower section is just above the top of the legs of the bottom tower section, the climber guides the tower down onto the top of the three legs, while calling out to the ground crew instructions about *slowly lowering* the new section down onto the legs. See **Fig 22**, which shows the climber guiding the new section of Rohn HBX tower onto the previous section's legs. This process is considerably easier to accomplish if each section of tower has been put together on the ground to make sure that the legs fit together easily. There's nothing more frustrating that trying to manually force-fit tower sections together at the top of a tower.

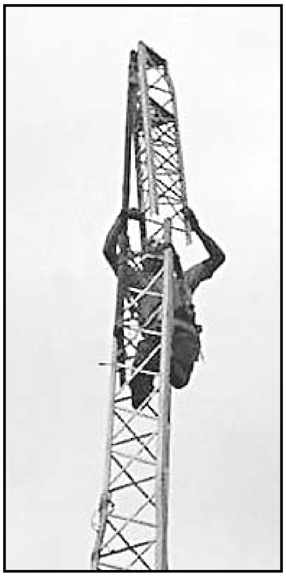
Fig 22—Climber is guiding the new section onto the top of the existing one. The gin pole attached to the left leg is bearing the weight, as the climber gives verbal instructions to the ground crew pulling on the gin-pole rope. (*Photo courtesy Mike Hammer, N2VR.*)

It seems that freight companies don't always handle heavy tower sections very gently and legs easily get bent out of alignment. A careful installer numbers tower sections in the order they've been test-fitted together on the ground, marking them with a laundry marker pen. You should also spray a small amount of WD-40 up inside mating tower legs after test-assembling them to help prevent galling and to ease fitting sections together. [Don't do this to excess—WD-40 is slippery and messy when it runs out of the bottom of tower legs.]

Another caution: Make sure before you start installing any tower that the correct ends of the bottom section's legs, "male" rather than "female," are pointed upwards. A prominent amateur (who will go unnamed) had to have Rohn make and send him a special "gender-bender" flange to turn females into males, since he had installed the base upside-down in the concrete base. You don't want to do that.

8. Once the new tower section has been guided down onto the male ends, the six pinning bolts are inserted and tightened with nuts. Note that Rohn uses two different sized bolts, with the larger diameter one on the bottom.
9. If this section of tower is one where guy wires are to be placed, they can be brought up using the gin pole rope and positioned on the tower. The maximum spacing for Rohn 25 is 30 feet between guy-wire sets, and 40 feet for Rohn 45. Thirty feet of unguyed Rohn 25 tower is wobbly, though safe. Many installers prefer to come down off the tower when setting guy-wire tension, since they do not like to be on a wobbly tower when the ground crew is moving around yanking on guy wires. Many also greatly prefer working on Rohn 45 tower, which is substantially more secure feeling and easier to stand on, with its legs 18 inches apart, while Rohn 25 legs are only 12 inches apart.
10. Finally, you reposition the gin pole for the next section of tower. The T-bar at the clamp is loosened, the tube is dropped down to the level of the clamp, and the climber walks the gin pole up to the top of the section just installed and clamps it there, ready to pull up the next tower section.

Tower Safety

One of the most important aspects of any tower installation project is the safety of all persons involved. See Chapter 1 for details on important safety issues. The use of hard hats is highly recommended for all assistants helping from the ground. Helpers should always stand clear of the tower base to prevent being hit by a dropped tool or hardware. Each person working on the tower must use a good climber's safety belt.

When climbing the tower, if more than one person is involved, one should climb into position before the other begins climbing. The same procedure is required for climbing down a tower after the job is completed. The purpose is to have the non-climbing person stand still so as not to drop any tools or objects on the climbing person, or unintentionally obstruct his movements. When two persons are working on top of a tower, only one should change position (unbelt and move) at a time.

For most installations, a good-quality $^1/_2$-inch diameter Manila hemp rope can adequately handle the workload for the hoisting tasks. The rope must be periodically inspected to assure that no tearing or chafing has developed, and if the rope should get wet from rain, it should be hung out to dry at the first opportunity. The knots used for connecting hoisting lines and hardware are critical to executing any safe installation, and special attention should be given to this detail for any work party.

Here is an important point regarding safety—the person who climbs the tower should be in charge of what happens with the ground crew. Not only does the person on the tower have a better overall view of the situation below, but also any confusion on the ground can result in serious injury to the climber.

GUY WIRES

In typical guyed tower installations, guy wires may experience loads in excess of 1000 pounds. Since the guys are the primary means of carrying the horizontal wind loads, great care should be taken in their selection and installation.

Guys come in a variety of materials and constructions. Normally, the tower manufacturer or professional engineer will specify the size and type of cable to be used. The most common type of cable used for tower guying is the EHS (Extra High Strength) galvanized steel cable. The EHS cables are very stiff and are the highest strength

Table 1
Guy Cable Comparisons

Cable	*Nominal Dia. Inches*	*Breaking Strength Lbs*	*Weight Lbs/100'*	*Elongation Inches/100'*	*%Elongation*
$^3/_{16}$" 1 × 7 EHS	0.188	3990	7.3	6.77	0.56%
$^1/_4$" 1 × 7 EHS	0.250	6700	12.1	3.81	0.32%
HPTG6700	0.220	6700	3.1	13.20	1.10%
HPTG8000	0.290	8000	3.5	8.90	0.74%
$^5/_{16}$" 1 × 7 EHS	0.313	11200	20.5	2.44	0.20%
HPTG11200	0.320	11200	5.5	5.45	0.45%
$^3/_8$" Fiberglass Rod	0.375	13000	9.7	5.43	0.45%

EHS steel cable information is taken from ASTM A 475-89, the industry standard specification for steel wire rope. The HPTG listings are for *Phillystran* aramid cables, and are based on the manufacturers' data sheets. The elongation (stretch) values are for 100 feet of cable with a 3000-pound load.

cables in the wire rope family. Other steel cables are made to be more flexible for running around pulleys. While these are easier to work with when assembling, they are not as strong as the EHS type, and should be avoided for tower guying. Non-conductive guys, such as *Phillystran* or *pultruded* fiberglass rod have become popular for eliminating resonant interaction with antennas.

Do not attempt to use cheaper cables that don't meet or exceed the criteria for those specified for your installation. Using the wrong cable, or failing to install the cable properly can have disastrous results! **Table 1** shows data for several cables commonly used for tower guying. It is important to note that the minimum breaking strength of the various cables are independent of their elongation (stretch) under load.

Guy Cable Installation

Figs 23 and **24** show methods for tensioning and safety wiring guy-wire turnbuckles. **Fig 25** shows the traditional method for fixing the end of a steel guy wire. A thimble is used to prevent the wire from breaking because of a sharp bend at the point of intersection. Conventional wisdom strongly recommends the use of thimbles that are at least one wire size larger than the cable to provide a more gentle wire bend radius. Three cable clamps follow to hold the wire securely. Be sure to follow the note in Fig 25 for which part of the clip bears against the live (loaded) cable. As a final backup measure, the individual strands of the free end are unraveled and wrapped around the guy wire. It is a lot of work, but it is necessary to ensure a safe and permanent connection.

Fig 26 shows the use of a device that replaces the clamps and twisted strands of wire. These devices are known as *dead ends*, *preformed guy grips,* or *Big Grips* and are commonly used on electrical power poles. They are far more convenient to use than are clamps, and are the recommended method for terminating Phillystran and

Fig 23—Proper tension can be placed on the guy wires with the aid of a block-and-tackle system. (*Photo by K1WA*)

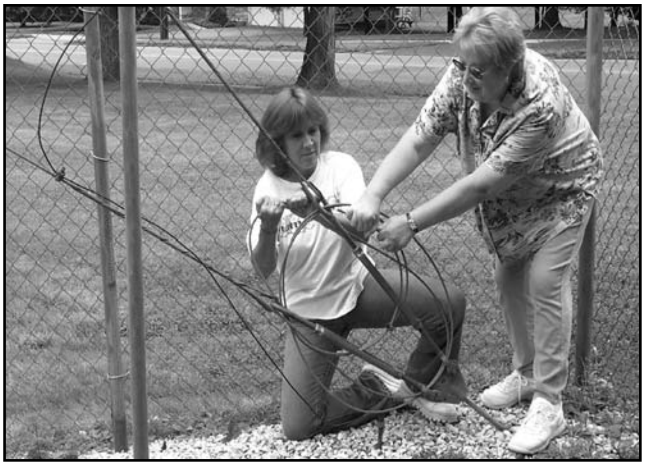

Fig 24—A length of guy cable is used to assure that the turnbuckles remain in place after they are tightened. This procedure is an absolute requirement in guyed tower systems. (*Photo by N4QX*)

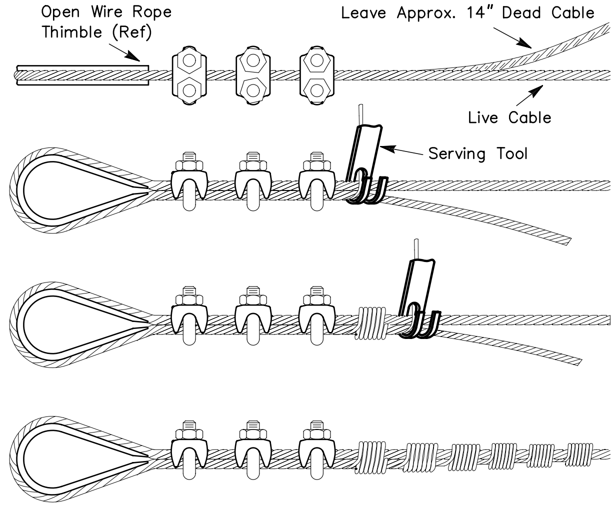

Fig 25—Traditional method for securing the end of a guy wire.

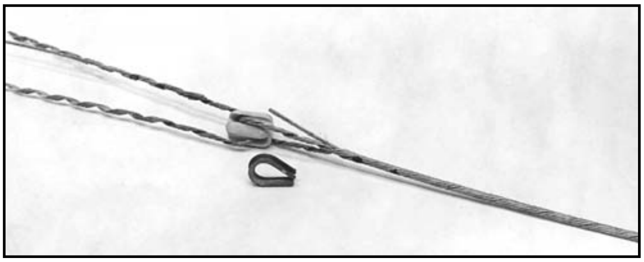

Fig 26—Alternative method for attaching guy wires using *dead ends.* The dead end on the right is completely assembled (the end of the guy wire extends beyond the grip for illustrative purposes). On the left, one side of the dead end is partially attached to the guy wire. In front, a thimble is used where a sharp bend might cause the guy wire or dead end to break.

fiberglass-rod guys. When using the guy grips, it is imperative that the recommended end sleeves are installed over the free end of the grip to prevent ice and falling hardware from sliding down the guy and unraveling the grip connection to the guy. The guy wires must be cut to the proper length. The dead end of each wire is installed into the object to which the guy wire is being attached (use a thimble, if needed to eliminate sharp cable bends). One side of the dead end is then wrapped around the guy wire. The other side of the dead end follows. Using dead ends saves time and trouble, more than making up for their slightly higher cost.

When using the non-conductive guy materials, it is highly recommended that a 25-foot length of EHS steel cable be used at the bottom for connection to the anchor. This serves a valuable purpose. The steel cable is more resistant to damage from ground activity and brush fires, and it is the preferred material for measuring cable pre-tension with commonly available devices.

Fig 27 shows two different methods for attaching guy wires to towers. At Fig 27A, the guy wire is simply looped around the tower leg and terminated in the usual manner. At Fig 27B, a *guy bracket*, with *torque arms* has been added. Even if the torque arms are not required, it is preferred to use the guy bracket to distribute the load from the tower/guy connection to all three tower legs, instead of just one. The torque bracket is more effective resisting torsional loads on the tower than the simpler installation. Rohn offers another guy attachment bracket, called a *Torque Arm Assembly*, that allows six guys to be connected between the bracket and anchors. This is by far the best method of stabilizing a tower against high torque loads, and is recommended for installations with large antennas.

There are two types of commonly used guy anchors. **Fig 28A** depicts an *earth screw*. These are usually 4 to 6 feet long. The screw blade at the bottom typically measures 6 to 8 inches diameter. Fig 28B illustrates two people installing the anchor. The shaft is tilted so that it will be in line with the mean angle of all the guys connecting to the anchor. Earth screws are suitable for use in normal soil where permitted by local building codes. Information about *screw anchors* is available from the manufacturers of these devices. Information from a supplier specializing in this type of anchor can be found at **www.abchance.com**.

The alternative to earth screws is the concrete block anchor. Fig 28C shows the installation of this type of anchor; it is suitable for any soil condition, with the possible exception of a bed of lava rock or coral. Consult the instructions from the manufacturer, or your tower designer, for the precise anchor configuration.

Turnbuckles and associated hardware are used to attach guy wires to anchors and to provide a convenient method for adjusting tension. **Fig 29A** shows a turnbuckle with a single guy wire attached to the eye of the anchor. Turnbuckles are usually fitted with either two eyes, or one eye and one jaw. The eyes are the oval ends, while the jaws are U-shaped with a bolt through each tip.

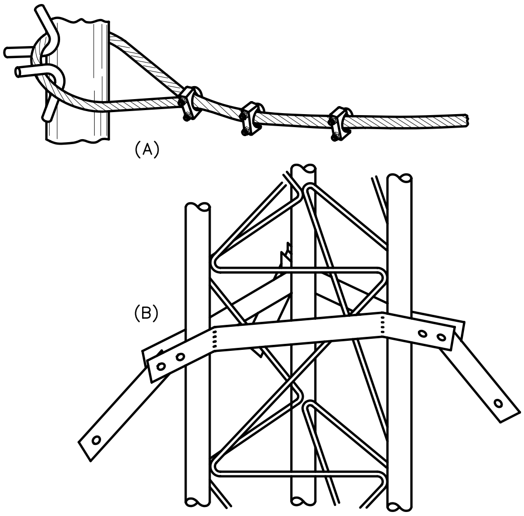

Fig 27—Two methods of attaching guy wires to tower. See text for discussion.

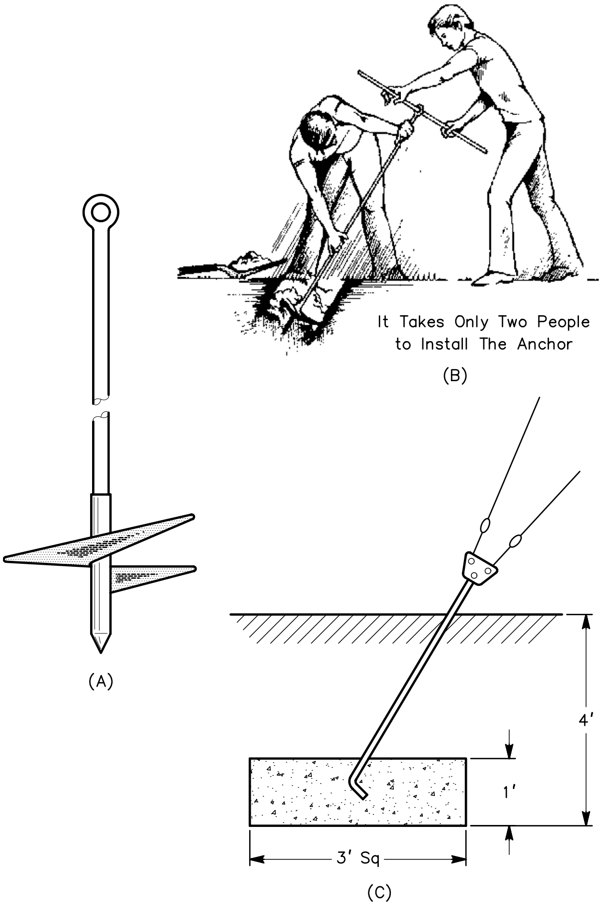

Fig 28—Two standard types of guy anchors. The earth screw shown at A is easy to install and widely available, but may not be suitable for use in certain soils. The concrete anchor is more difficult to install properly, but it is suitable for use with a wide variety of soil conditions and will satisfy most building code requirements.

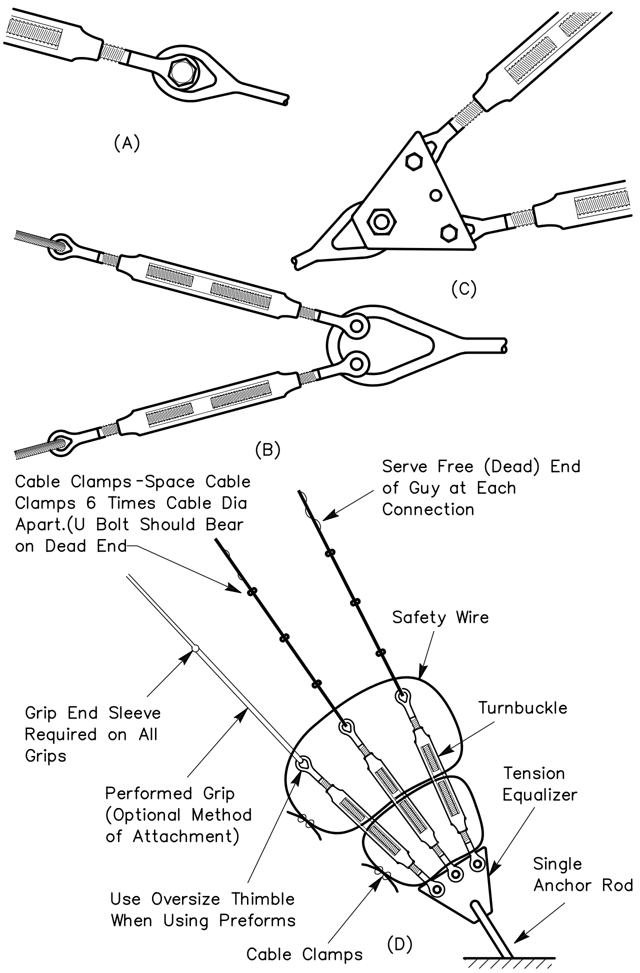
(A)
(C)
(B)
Cable Clamps -Space Cable Clamps 6 Times Cable Dia Apart.(U Bolt Should Bear on Dead End
Serve Free (Dead) End of Guy at Each Connection
Safety Wire
Turnbuckle
Grip End Sleeve Required on All Grips
Performed Grip (Optional Method of Attachment)
Tension Equalizer
Single Anchor Rod
Use Oversize Thimble When Using Preforms
Cable Clamps
(D)

Fig 29—Variety of means available for attaching guy wires and turnbuckles to anchors.

Fig 29B shows two turnbuckles attached to the eye of an anchor. The procedure for installation is to remove the bolt from the jaw, pass the jaw over the eye of the anchor and reinstall the bolt through the jaw, through the eye of the anchor and through the other side of the jaw.

If two or more guy wires are attached to one anchor, *equalizer plates* should be installed (Fig 29C). In addition to providing a convenient point to attach the turnbuckles, the plates pivot slightly to equalize the various guy loads and produce a single load applied to the anchor. Once the installation is complete, a safety wire should be passed through the turnbuckles in a figure-eight fashion to prevent the turnbuckles from turning and getting out of adjustment (Fig 29D).

All guyed towers require the guys to be installed with a certain amount of pre-tension. The tower manufacturer or designer specifies the required pre-tension values, which are usually 10% of the cable breaking strength. Pre-tension is necessary to eliminate looseness in the cable caused by the spiral wire construction and to eliminate excessive dynamic guy and tower motion under wind loading. The recommended method for adjusting the guys is to use a cable tension-measuring device such as the popular *Loos Guy Wire Tensioner*. The guy is gripped with a special clamp, such as the *Klein Cable Grip*, which is connected to the anchor below the eye (or equalizer plate) with a block and tackle arrangement (Fig 23) or a ratcheting come-along. Then the turnbuckle is adjusted to take up the load, the cable grip is released and the final guy tension is adjusted and checked.

When you adjust the guys at each level, you should check the tower for vertical alignment and straightness. This is often done with a transit from two ground points located 90° from each other.

Resonance in Guy Wires

If guy wires are resonant at or near the operating frequency, they can receive and reradiate RF energy. By behaving as parasitic elements, the guy wires may alter and thereby distort the radiation pattern of a nearby antenna. For low frequencies where a dipole or other simple antenna is used, this is generally of little or no consequence. But at the higher frequencies where a unidirectional antenna is installed, it is desirable to avoid pattern distortion if at all possible. The symptoms of re-radiating guy wires are usually a lower front to back ratio and a lower front to side ratio than the antenna is capable of producing. The gain of the antenna and the feed-point impedance will usually not be significantly affected, although sometimes changes in SWR can be noted as the antenna is rotated. (Of course other conductors in the vicinity of the antenna can also produce these same symptoms.)

The amount of re-radiation from a guy wire depends on two factors—its resonant frequency, and the degree of coupling to the antenna. Resonant guy wires near the antenna will have a greater effect on performance than those that are farther away. Therefore, the upper portion of the top level of guy wires should warrant the most attention with horizontally polarized arrays. The lower guy wires are usually closer to horizontal than the top level, but by virtue of their increased distance from the antenna, are not coupled as tightly to the antenna.

To avoid resonance, the guys should be broken up by means of egg or strain insulators. **Fig 30** shows wire lengths that fall within 10% of $^{1}/_{2}$-λ resonance (or a multiple of $^{1}/_{2}$ λ) for all the HF amateur bands. Unfortunately, no single length greater than about 14 feet avoids resonance in all bands. If you operate just a few bands, you can locate greater lengths from Fig 30 that will avoid resonance. For example, if you operate only the 14-, 21- and 24-MHz bands, guy wire lengths of 27 feet or 51 feet would be suitable, along with any length less than 16 feet.

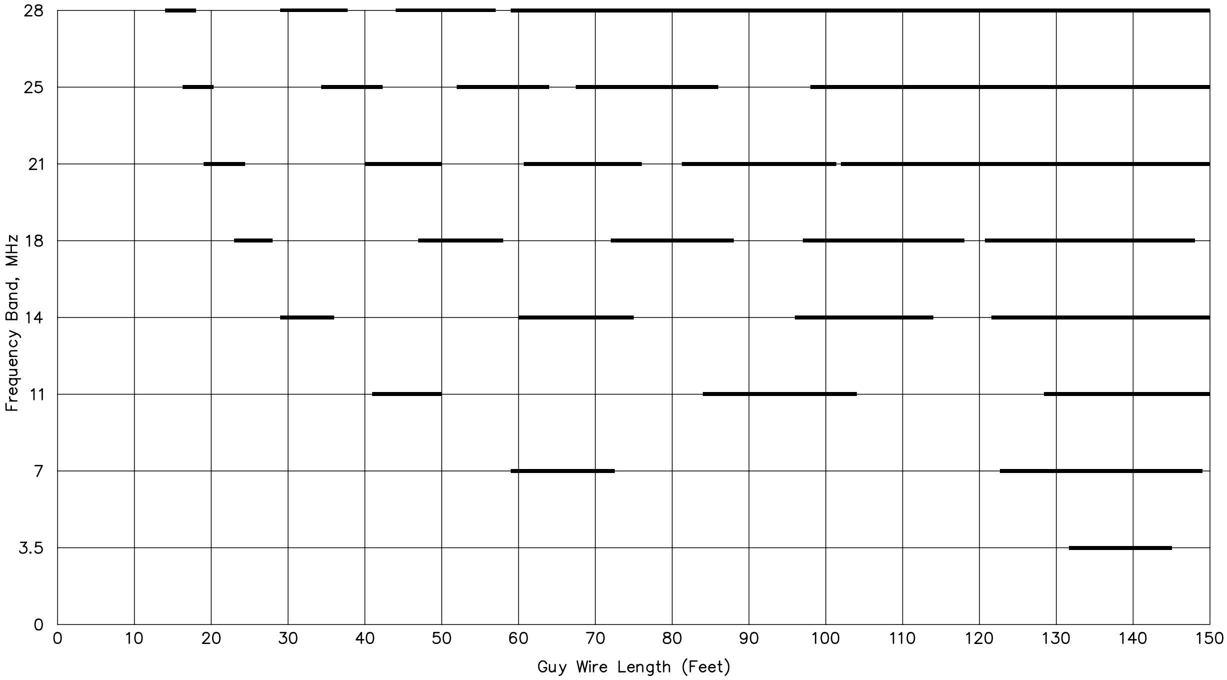

Fig 30—The black bars indicate ungrounded guy wire lengths to avoid for the eight HF amateur bands. This chart is based on resonance within 10% of any frequency in the band. Grounded wires will exhibit resonance at odd multiples of a quarter wavelength. (*By Jerry Hall, K1TD.*)

THE RIGHT TOWER FOR YOUR ANTENNA

Most manufacturers rate their towers in terms of the maximum allowable antenna load that can safely be carried at a specific wind speed. Ensuring that the specific antennas you plan to install meet the tower's design criteria, however, may not always be a straightforward task.

For most towers, the manufacturer assumes that the allowable antenna load is a horizontal force applied at the top of the tower. The allowable load represents a defined amount of exposed antenna area, at a specified wind velocity. Most tower manufacturers rate the load in terms of *Flat Projected Area* (FPA). This is simply the equivalent area of a flat rectangular surface at right angles to the wind. The FPA is not related to the actual shape of the antenna itself, only its rectangular projected area. Some manufacturers provide separate FPAs for antennas made from cylindrical sections and those made from rectangular sections.

In the realm of antenna manufacturers, however, you may encounter another wind load rating called the *Effective Projected Area* (EPA). This attempts to take into account the actual shape of antenna elements. The problem is that there is no agreed-upon standard for the conversion from EPA to load numbers. Different manufacturers may use different conversion factors.

Since most tower manufacturers have provided FPA figures for their towers—allowing us in effect to ignore design-specification details—it would be easiest for us to work only with FPA values for our antennas. This would be fine, if indeed we had good FPA figures for the specific antennas we plan to use! Unfortunately, FPAs are rarely specified for commercially built amateur antennas. Instead, most antenna manufacturers provide effective areas in their specification sheets. You may need to contact the antenna manufacturer directly for the FPA antenna area or for the antenna dimensions so that you can do your own FPA calculations.

Determining Antenna Areas

The method for determining the flat projected area of an antenna is quite simple. We'll use a Yagi antenna as an example. There are two worst-case areas that should be considered here. The first is the FPA of all the elements when the wind blows in the direction along the boom; that is, at right angles to the elements. The second FPA for a Yagi is when the wind is at right angles to the boom. One of these two orientations produces the worst-case exposed antenna area—all other wind angles present lower exposed areas. The idea is to take the highest of the FPAs for these two wind directions and call that the FPA of the antenna structure. See **Fig 31A**.

The element FPA is calculated by multiplying each element's dimension of length by its diameter and then

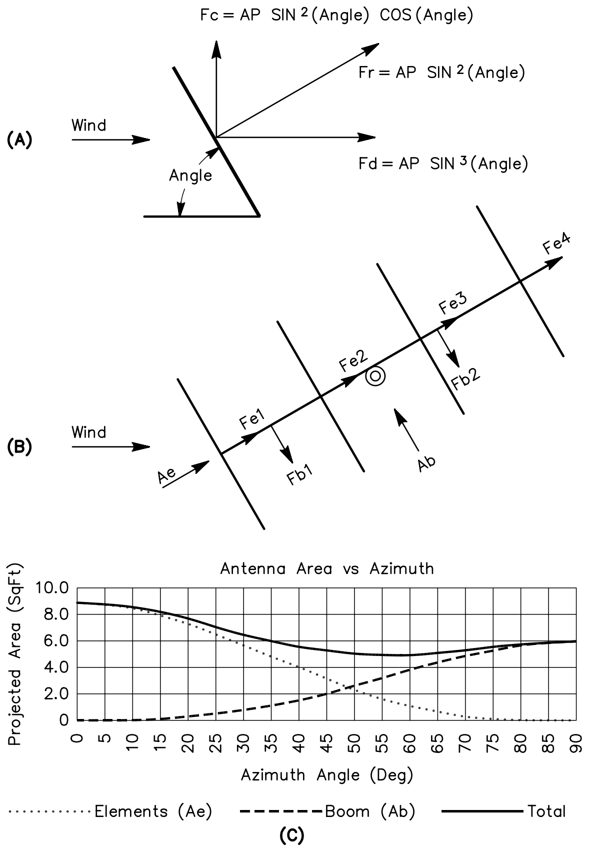

Fig 31—Description of how loads are developed on a Yagi. At A, Fr is the resultant force from the wind load on a generalized member. Fd is the load acting downwind (drag) that creates the load on the tower. Fc is the lateral component of the wind load. The term A is the flat projected area (FPA), which is the broadside area normal to the wind. The term P is the wind pressure. At B, Ae is the total element area, while Ab is the total boom area. All the loads due to the wind act normal to the antenna sections—the force on element #1 (Fe1) acts along the axis of the boom, for example. At C, a plot of the *effective FPA* as a function of the azimuthal wind direction for a Yagi, ignoring drag coefficients. The Yagi in this example has 9.0 square feet of element FPA and 6.0 square feet of boom FPA. The worst-case FPAs occur with the beam pointed in the wind and with the boom broadside to the wind. To determine the actual tower loading, the actual drag coefficients and wind pressures must be used.

summing the FPAs for all elements. The boom's FPA is computed by multiplying the boom's length by its diameter.

The reason for considering two potential peak-load orientations becomes clear when different frequency antennas are stacked on a mast or tower. Some antennas produce peak loads when the elements are broadside to the wind. This is typical of low-frequency Yagis, where the elements are long lengths of aluminum tubing. On the other hand, the boom can dominate the surface area computations in higher-frequency Yagis.

The fundamentals responsible for the need to examine both potential FPAs for Yagis relates to how wind flows over a structure and develops loads. Called *The Cross-Flow Principle*, this was introduced to the communications industry by Dick Weber, K5IU, in 1993. The principle is based on the fact that the loads created by wind flowing across an antenna member only produce forces that are normal to (or perpendicular to) the major axis of the member. The resultant and component load calculations for this method are shown in Fig 31A.

For a Yagi, this means that wind forces on the elements act in-line with the boom, while forces on the boom act in-line with the elements. Fig 31B shows a force diagram for a typical Yagi. Fig 31C shows the FPA for a Yagi rotated through 90° of azimuth.

Antenna Placement on the Mast/Tower

Another important consideration is where the antenna(s) will be placed on the tower. As mentioned before, most generic tower specifications assume that the entire antenna load is applied at the top of the tower. Most amateur installations have a tubular mast extending above the tower top, turned by a rotator mounted down inside the tower. Multiple Yagi antennas are often placed on the mast above the tower top, and you must make sure that both the tower and the mast can withstand the wind forces on the antennas.

For freestanding towers, you can determine how a proposed antenna configuration compares to the tower manufacturer's rating by using an *Equivalent Moment* method. The method computes the bending moment generated at the base of the tower by wind loads on the tower's rated antenna area located right at the top of the tower and compares that to the case when the antenna is mounted on a mast sticking out of the top of the tower.

The exact value of wind pressure is not important, so long as it is the same for both comparisons. The wind load on the tower itself can be ignored because it is the same in both comparisons and the drag coefficients for the antennas can also be ignored if all calculations are performed using flat projected antenna areas, as we've recommended previously.

Keep in mind that this approach does not calculate *actual* loads and moments relevant to any specific tower design standard, but it does allow equivalent comparisons when the wind pressure is constant and all the antenna areas are of the same type. An example is in order.

Fig 32A shows a generic tower configuration, with a concentrated antenna load at the top of the tower. We'll assume that the tower manufacturer rates this tower at 20 square feet of flat projected antenna area. Fig 32B shows a typical amateur installation with a rotating mast and an antenna mounted 7 feet above the top of the tower. To make the calculations easy, we select a wind pressure

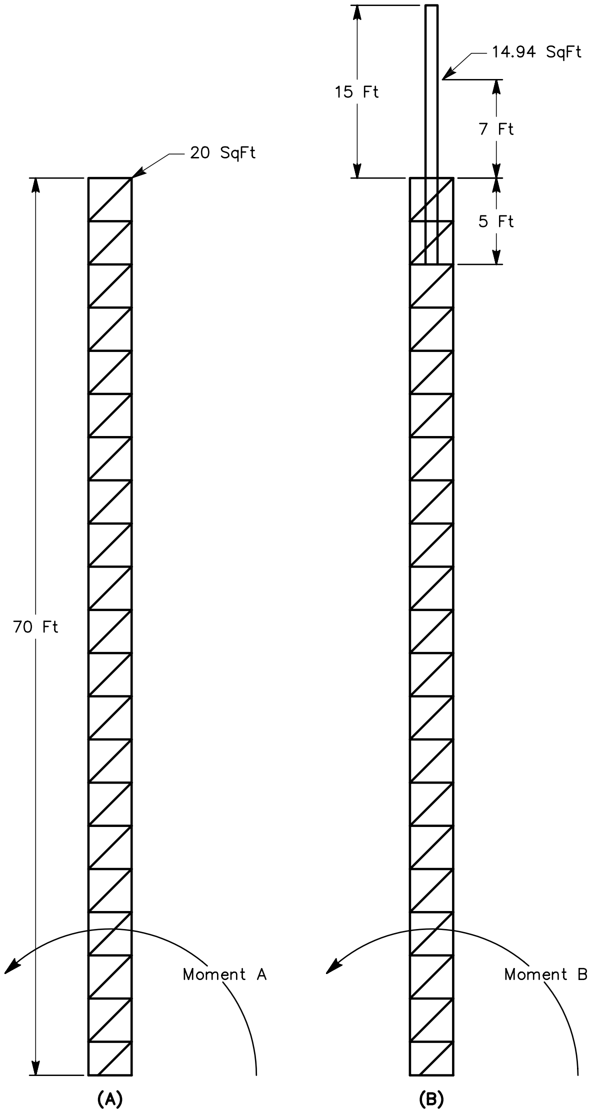

Fig 32—At A, a 70-foot tower rated for 20 square feet of antenna load at the top. At B, the same tower with a 2-inch OD × 20-foot long mast, with an antenna mounted 7 feet above the top of the tower. Both configurations produce the same tower load.

of 1 pound per square foot (1 psf). This makes the tower base moment calculation for Fig 32A:

Antenna load = 20 feet^2 × 1 psf = 20 pounds
Base moment = 70 feet × 20 pounds =
1400 foot-pounds.

This is the target value for the comparison. An equivalent configuration would produce the same base moment. For the configuration in Fig 32B, we assume a tubular 2-inch diameter mast that is 20 feet long, mounted 5 feet down inside the tower. Note that the lattice structure of the tower allows the wind to "see" the whole length of the mast and that we can consider the force distributed along the mast as being a single force concentrated at the mast's center. The flat projected area of the mast by itself, without the antenna, is:

Mast area = 20 feet × 2 inches / 12 inches/foot =
3.33 square feet

The center of the mast is located at a height of 75 feet. Using the same 1-psf-wind load, the base bending moment due to the mast alone is:

Base moment (due to mast) = 3.33 feet^2 × 1 psf ×
75 feet = 249.75 foot-pounds

Including the mast in the configuration reduces the allowable antenna load. The remaining target base moment left for the antenna is found by subtracting the moment due to the mast from the original target value:

New base target moment = 1400 - 249.75 foot-pounds
= 1150.25 foot-pounds.

The antenna in Fig 32B is located at a height of 77 feet. To obtain the allowable antenna area at this elevation we divide the new base target moment by the antenna height, yielding an allowable antenna load of:

1150.25 foot-pounds / 77 feet = 14.94 pounds.

Since we chose a wind load of 1 psf, the allowable antenna FPA has been reduced to 14.94 square feet from 20 square feet. If the projected area of the antenna we are planning to mount in the new configuration is less than or equal to this value, we have satisfied the requirements of the original design. You can use this equivalent-moment method to evaluate different configurations, even ones involving multiple antennas on the mast or situations with additional antennas placed along the tower below the tower top.

For guyed towers, the analyses become much more rigorous to solve. Because the guys and their behaviors are such a significant portion of the tower support mechanism, these designs can become very sensitive to antenna load placements. A general rule of thumb for guyed towers is never to exceed the original tower-top load rating, regardless of distributed loads along its length. Once you redistribute the antenna load placements along a guyed tower, you should do a fresh analysis, just to be sure.

You can run evaluations using the above method for antennas placed on the mast above a guyed tower top. The use of the Equivalent-Moment method for antennas mounted below the top of a guyed tower, however, can become quite suspect, since many generic tower designs have their intermediate guys sized for zero antenna loads lower down the tower. The proper approach in this case is to have a qualified mechanical engineer check the configuration, to see if guy placement and strength is adequate for the additional antennas down the tower.

Mounting the mast and antenna as shown in Fig 32B increases tower loads in the region of the mast. You should investigate these loads to ensure that the tower bracing in that area is sufficient. Now we will consider the problem of bending the rotating mast.

Mast Strength

When you mount antennas on a mast above the tower top, you should examine the bending loads on the mast to ensure that it will be strong enough. This section explains how to perform mast stress calculations for a single sustained wind speed. This procedure does not include height, exposure and gust-response factors found in most tower design standards.

Here are some fundamental formulas and values used to calculate the bending stress in a mast mounted in the top of a tower. The basic formula for wind pressure is:

$$P = .00256\ V^2 \qquad \text{(Eq 1)}$$

where

P is the wind pressure is in pounds per square foot (psf)

V = wind speed in miles per hour (mph)

This assumes an air density for standard temperature and atmospheric pressure at sea level. The wind speed is not the Basic Wind Speed discussed in other sections of this chapter. It is simply a steady state (static) wind velocity.

The formula for calculating the force created by the wind on a structure is:

$$F = P \times A \times C_d \qquad \text{(Eq 2)}$$

where

P = the wind pressure from Eq (1)

A = the flat projected area of the structure (square feet)

C_d = drag coefficient for the shape of the structure's members.

The commonly accepted *drag coefficient* for long cylindrical members like the tubing used for the mast and antenna is 1.20. The coefficient for a flat plate is 2.0.

The formula used to find the *bending stress* in a simple beam like our mast is:

$$\sigma = \frac{M \times c}{I} \qquad \text{(Eq 3)}$$

where

σ = the stress in pounds per square inch (psi)

M = *bending moment* at the base of the mast (inch-pounds)

c = ½ of the mast outside diameter (inches)

I = *moment of inertia* of the mast section (inches^4)

In this equation you must make sure that all values are in the same units. To arrive at the mast stress in pounds per square inch (psi), the other values need to be in inches and pounds also. The equation used to find the moment of inertia for the round tubing mast section is:

$$I = \frac{\pi}{4}(R^4 - r^4) \qquad \text{(Eq 4)}$$

where

I = Moment of Inertia of the section (inches^4)

R = Radius of tube outside diameter (inches)

r = Radius of tube inside diameter (inches)

This value describes the distribution of material about the mast *centroid*, which determines how it behaves under load. The equation used to compute the *bending moment* at the base of the mast (where it is supported by the tower) is:

$$M = (F_M \times L_M) + (F_A \times L_M) \qquad \text{(Eq 5)}$$

where

F_M = wind force from the mast (pounds)

L_M = Distance from tower top to center of mast (inches)

F_A = Wind force from the antenna (pounds)

L_A = Distance from tower top to antenna attachment (inches)

L_M is the distance to the center of the portion of the mast extending above the tower top. Additional antennas can be added to this formula by including their F × L. In the installation shown in Fig 32B, a wind speed of 90 mph, and a mast that is 2 inches OD, with a 0.250-inch wall thickness, the steps for calculating the mast stress are:

1. Calculate the wind pressure for 90 mph, from Eq 1:

$P = .00256\ V^2 = .00256 \times (90)^2 = 20.736$ psf

2. Determine the flat projected area of the mast. The portion of the mast above the tower is 15 feet long and has an outside diameter of 2 inches, which is 2/12 feet.

Mast FPA, A_M = 15 feet × (2 inches / 12 inches/feet) = 2.50 square feet.

3. Calculate the wind load on the mast, from Eq 2:

Mast Force, $F_M = P \times A \times C_d =$ 20.736 psf × 2.50 feet^2 × 1.20 = 62.21 pounds

4. Calculate the wind load on the antenna: From Eq 2

Antenna Force, $F_A = P \times A \times C_d$ = 20.736 psf × 14.94 feet^2 × 1.20 = 371.76 pounds

5. Calculate the mast *Bending Moment*, from Eq 5:

$M = (F_M \times L_M) + (F_A \times L_A)$ = (62.21 pounds × 90 inches) + (371.76 pounds × 84 inches) = 36827 inch-pounds

where L_M = 7.5 feet × 12 inches/foot = 90 inches and L_A = 7.0 feet × 12 inches/foot = 84 inches.

6. Calculate the mast *Moment of Inertia,* from Eq 4:

$$I = \frac{\pi}{4}(R^4 - r^4) = \frac{\pi}{4}(1.0^4 - 0.75^4) = 0.5369 \text{ inches}^4$$

where, for a 2.0-inch OD and 0.250-inch wall thickness tube, R= 1.0 and r = 0.75.

7. Calculate the mast *Bending Stress,* from Eq 3:

$$\sigma = \frac{M \times c}{I} = \frac{36827 \text{ inch-pounds} \times 1.0 \text{ inches}}{0.5369} = 68592 \text{ psi}$$

If the yield strength of the mast material is greater than the calculated bending stress, the mast is considered safe for this configuration and wind speed. If the calculated stress is higher than the mast yield strength, a stronger alloy, or a larger mast, or one with a thicker wall is required.

There are many different materials and manufacturing processes for tubing that may be used for a mast. Yield strengths range from 25,000 psi to nearly 100,000 psi. Knowing the minimum yield strength of the material used for a mast is an important part of determining if it will be safe. Using unknown materials renders efforts from the preceding calculations useless!

When evaluating a mast with multiple antennas attached to it, special care should be given to finding the worst-case condition (wind direction) for the system. What may appear to be the worst load case, by virtue of the combined flat projected antenna areas, may not always be the exposure that creates the largest mast bending moment. Masts with multiple stacked antennas should always be examined to find the exposure that produces the largest mast bending moment. The antenna flat projected areas at 0° and 90° azimuths are particularly useful for this evaluation.

ANTENNA INSTALLATION

All antenna installations are different in some respects. Therefore, thorough planning is the most important first step in installing any antenna. Before anyone climbs the tower, the whole process should be discussed to be sure each crewmember understands what is to be done. And remember that the person on the tower is in charge! Coordinate beforehand what signals and commands are used: “Up” or “Up Slowly” for raising something from the ground; “Down” or “Down Slowly” for the opposite.

“Watch Out!” or ‘Watch Out Below!” works for dropped hardware or tools to alert the ground crew below. Remember, once someone is on the tower, no one should be allowed to stand near the base of the tower!

Consider what tools and parts must be assembled and what items must be taken up the tower, and plan alternative actions for possible trouble spots. Extra trips up and down the tower can be avoided by careful planning.

If done properly, the actual work of getting the antenna into position can be executed quite easily with only one person at the top of the tower. The ground crew should do all the heavy work and leave the person on the tower free to guide the antenna into position. Because the ground crew does all the lifting, a large pulley, preferably on a gin pole placed at the top of the tower, is essential. Local radio clubs often have gin poles available for use by their members. Stores that sell tower materials frequently rent gin poles as well.

A gin pole should be placed along the side of the tower so the pulley is no more than 2 feet above the top of the tower (or the point at which the antenna is to be placed). Normally this height is sufficient to allow the antenna to be positioned easily. An important reason that the pulley is placed at this level is that there can be considerable strain on the gin pole when the antenna is pulled away from the tower to maneuver past guy wires.

Sometime the mast to which the antenna will be mounted is used as a place to hang the pulley. You should take care that you don’t end up bending the mast by placing the pulley too high on the mast. It may be necessary to back-guy the mast on the opposite side of the tower from which the antenna is raised.

The rope (halyard) through the pulley must be somewhat longer than twice the tower height so that the ground crew can raise the antenna from ground level. The rope should be ½ or ⁵/₈ inch diameter for both strength and ease of handling. Smaller diameter rope is less easily manipulated; it has a tendency to jump out of the pulley track and foul pulley operation.

The first person to climb the tower should carry an end of the halyard so that the gin pole can be lifted and secured to the tower. Those climbing the tower must have safety belts. Belts provide safety and convenience; it is simply impossible to work effectively while hanging onto the tower with one hand.

Once positioned, the gin pole and pulley allow parts and tools to be sent up the tower. A useful trick for sending up small items like bolts and pliers is for a ground crew member to slide them through the rope strands where they are held by the rope for the trip to the top of the tower. Items that might be dislodged by contact with the tower should either be taped or tied to the halyard.
Ever present is the hazard of falling tools or hardware. It is foolish to stand near a tower when someone is working above. Ground crew member should wear hard-hats as extra insurance.

Raising the Antenna Alongside the Tower

A technique that can save much effort in raising the antenna is outlined here. First, the halyard is passed through the gin-pole pulley or the pulley mounted to the mast, and the leading end of the rope is returned to the ground crew, where it is tied to the antenna. The assembled antenna should be placed in a clear area of the yard (or the roof) so the boom points toward the tower. The halyard is then passed *under* the front elements of the beam to a position past the midpoint of the antenna, where it is securely tied to the boom (**Fig 33A**).

Note that once the antenna is installed, the tower worker must be able to reach and untie the halyard from the boom; the rope must be tied less than an arm's length along the boom from the mounting point. If necessary, a large loop may be placed around the first element located beyond the midpoint of the boom, with the knot tied near the center of the antenna. The rope may then be untied easily after completion of the installation. The halyard should be tied temporarily to the boom at the front of the antenna by means of a short piece of light rope or twine.

While the antenna is being raised, the ground crew does all the pulling. As soon as the front of the antenna reaches the top of the mast, the person atop the tower unties the light rope and prevents the front of the antenna from falling, as the ground crew continues to lift the antenna (Fig 33B). When the center of the antenna is even with the top of the tower, the tower worker puts one bolt through the mast and the antenna-mounting bracket on the boom. The single bolt acts as a pivot point and the ground crew continues to lift the back of the antenna with the halyard (Fig 33C). After the antenna is horizontal, the tower worker secures the rest of the mounting bolts and unties the halyard. By using this technique, the tower worker performs no heavy lifting.

Avoiding Guy Wires

Although the same basic methods of installing a Yagi apply to any tower, guyed towers pose a special problem. Steps must be taken to avoid snagging the antenna on the guy wires. With proper precautions, however, even large antennas can be pulled to the top of a tower, even if the

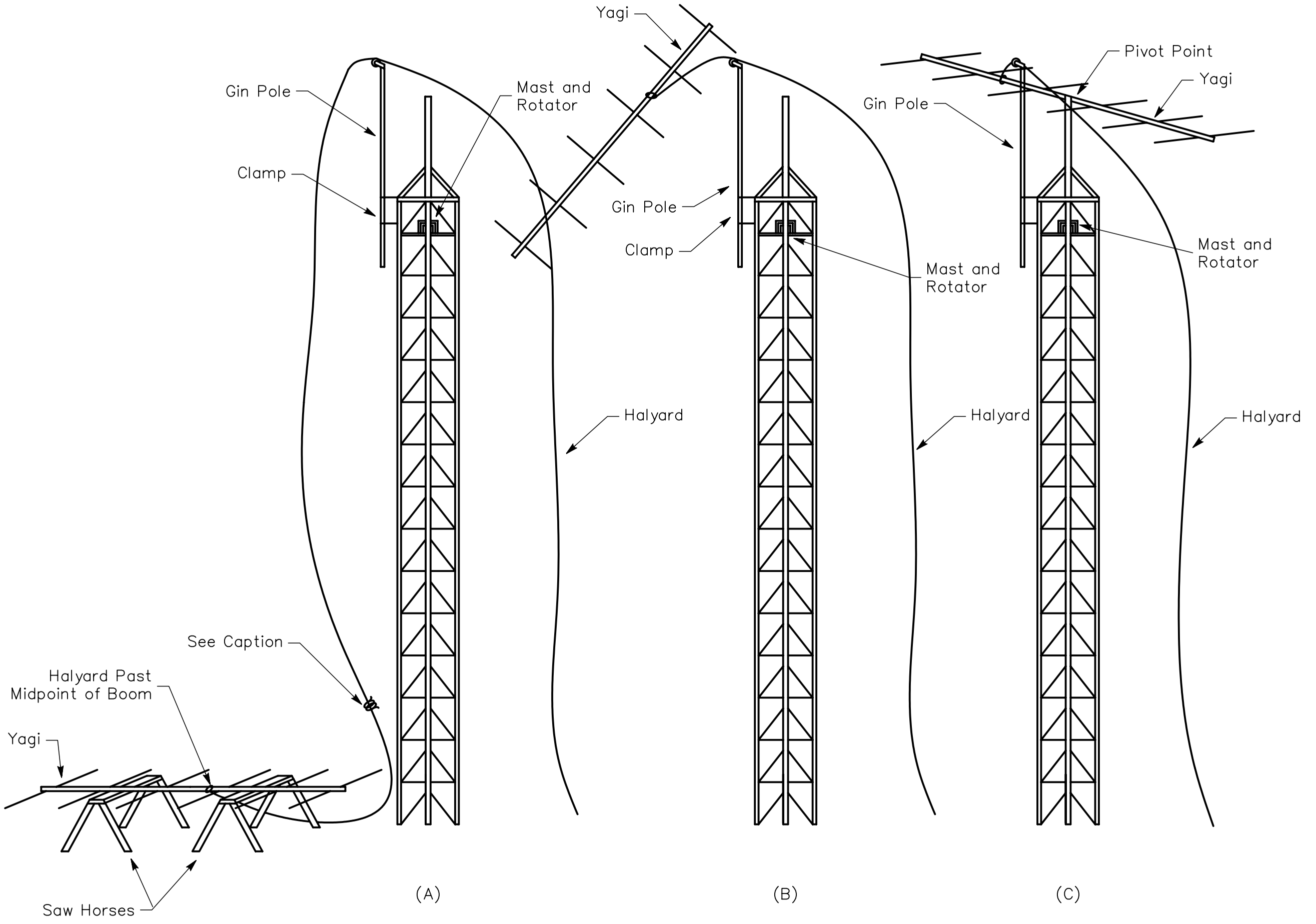

Fig 33—Raising a Yagi antenna alongside the tower. At A the Yagi is placed in a clear area, with the boom pointing toward the tower. The halyard is passed under the elements, then is secured to the boom beyond the midpoint. B shows the antenna approaching the top of the mast. The person on the tower guides it after the lifting rope has been untied from the front of the antenna. At C the antenna is pulled into a horizontal position by the ground crew. The tower worker inserts the pivot bolt and secures it. Note: A short piece of rope is tied around the halyard and the boom at the front of the antenna to stabilize the beam as it is being raised. The tower worker removes it when the boom reaches him at the top of the tower.

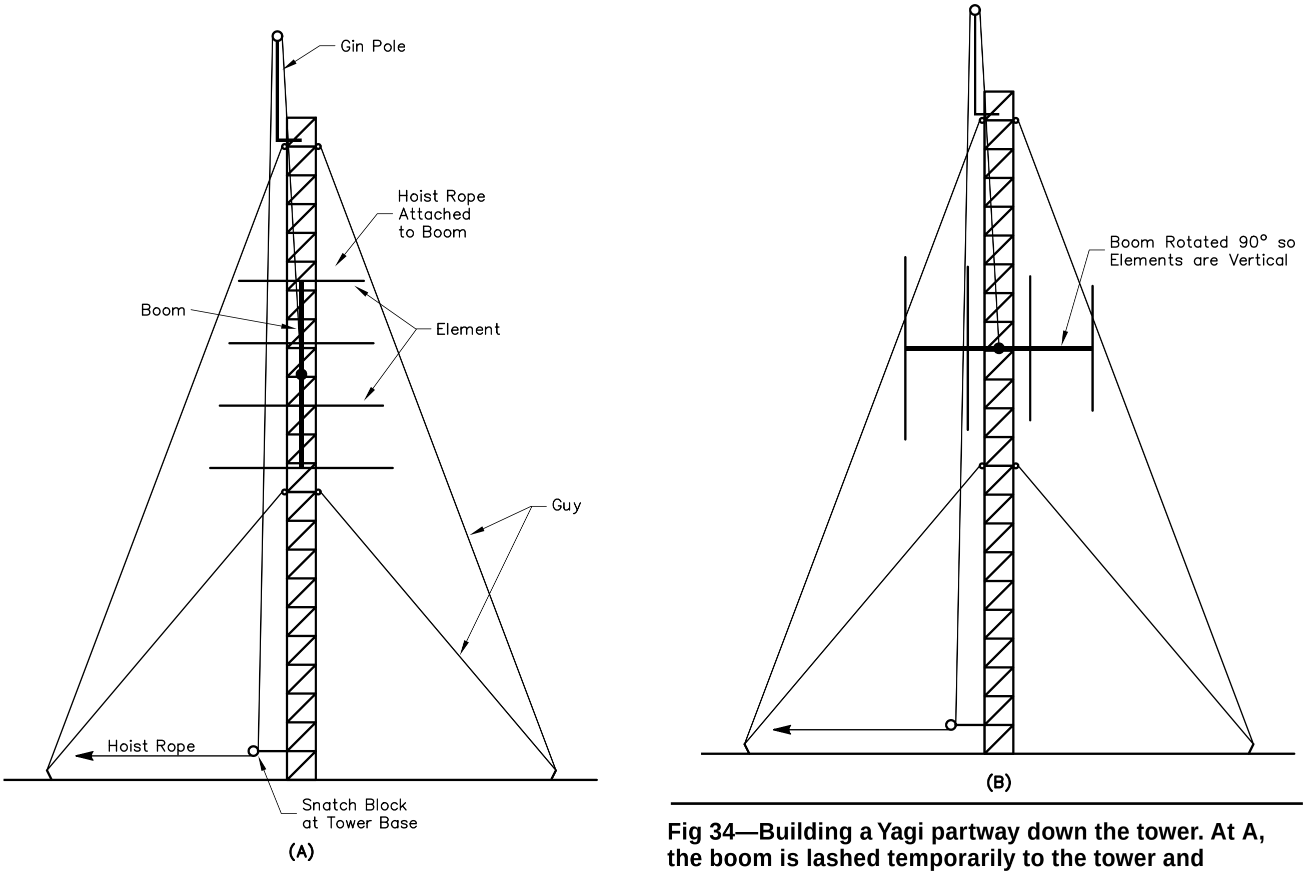
Gin Pole
Hoist Rope Attached to Boom
Boom
Element
Guy
Hoist Rope
Snatch Block at Tower Base
(A)
Boom Rotated 90° so Elements are Vertical
(B)

Fig 34—Building a Yagi partway down the tower. At A, the boom is lashed temporarily to the tower and elements are added, starting at the bottom. At B, the temporary rope securing the boom to the tower is removed and the boom is rotated 90° so that the elements are vertical. At C, the boom is rotated another 90°, "weaving through" guy wires if necessary, until the elements are parallel with the ground, whereupon the boom is secured to the tower.

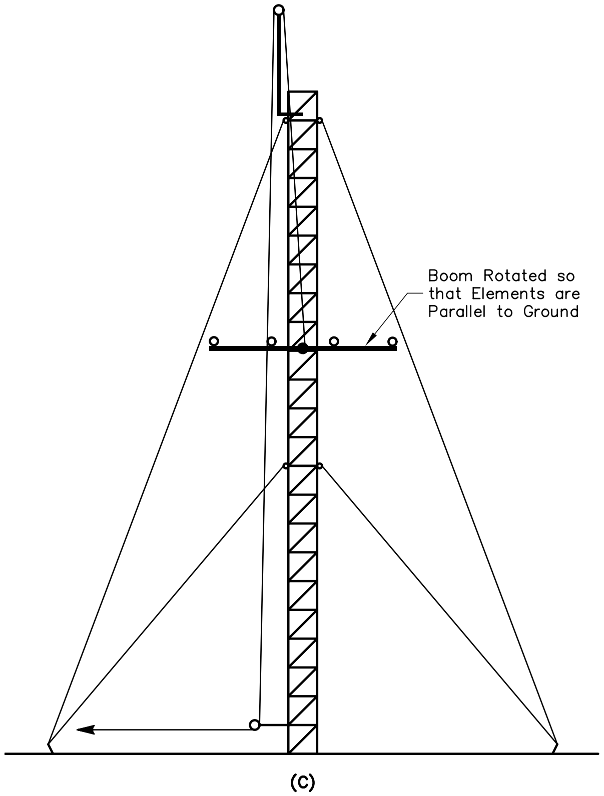
Boom Rotated so that Elements are Parallel to Ground
(C)

mast is guyed at several levels.

Sometimes one of the top guys can provide a *track* to support the antenna as it is pulled upward. Insulators in the guys, however, may obstruct the movement of the antenna. A better track made with rope is an alternative. One end of the rope is secured outside the guy anchors. The other end is passed over the top of the tower and back down to an anchor near the first anchor. So arranged, the rope forms a narrow V-track strung outside the guy wires. Once the V-track is secured, the antenna may simply be pulled up, resting on the track.

Another method is to tie a rope to the back of the antenna (but within reach of the center). The ground crews then pull the antenna out away from the guys as the antenna is raised. With this method, some crewmembers are pulling up the antenna to raise it while others are pulling down and out to keep the beam clear of the guys. Obviously, the opposing crews must act in coordination to avoid damaging the antenna. The beam is especially vulnerable when it begins to tip into the horizontal position. If the crew continues to pull out and down against the antenna, the boom can be broken. Another problem with this approach is that the antenna may rotate on the axis of the boom as it is raised. To prevent such rotation, long lengths of twine may be tied to outer elements, one piece on each side of the boom. Ground personnel may then use these *tag lines* to stabilize the antenna. Where this is done, provisions must be made for untying the twine once the antenna is in place.

A third method is to tie the halyard to the center of the antenna. A crewmember, wearing a safety belt, walks the antenna up the tower as the crew on the ground raises it. Because the halyard is tied at the balance point, the tower worker can rotate the elements around the guys. A tag line can be tied to the bottom end of the boom so that a ground worker can help move the antenna around the guys. The tag line must be removed while the antenna is still vertical.

A fourth method is to build the antenna on the tower and then swing it into position. (See also the section below on the PVC Mount.) Building the Yagi on the tower works particularly well for Yagis mounted partway up the tower, as you might do in a stacked array. The technique works best when the vertical spacing between the guys is greater than the length of the Yagi boom.

Fig 34 illustrates the steps involved. A pull rope through a gin-pole or tower-mounted pulley is secured to the boom at the final balance point and the ground crew raises the boom in a vertical position up the tower. A tie rope is used to temporarily secure the upper end of the boom to keep it stable while the boom is being raised. The tower person removes the tie-rope once the boom is raised to the right level and has been temporarily secured to the tower.

The elements are then brought up one at a time and mounted to the boom. It helps if you have a 2- or 3-foot long *spotting mast* temporarily attached to the boom to form a 90° frame of reference. This allows the ground crew to spot from below so that the elements are all lined up in the same plane. After all the elements are mounted and aligned properly, the temporary rope securing the boom to the tower is released, suspending the antenna on the pull rope. The tower person then rotates the boom 90° so that the elements are vertical. Next the elements are rotated 90° into the tower so that they are parallel to the ground. The ground crew then moves the boom up or down using the pull rope to the final point where it is mounted to the tower.

A modification of this technique also works for building a medium-sized Yagi on the top of the tower. This technique will work if the length of the gin pole at maximum safe extension is long enough. See **Fig 35**.

As usual, the gin-pole pull rope is attached to the balance point of the boom and the boom is pulled up the tower in the vertical position, using a rope to temporarily tie the pull rope to the top end of the boom for stability. The boom is temporarily secured to the tower with rope in the vertical position so that the top end is just higher than the top of the tower. In order to clear the gin pole when the elements are mounted and the boom is raised higher to mount the next element, you must tilt the boom slightly so that the element mounted to the top end of the boom will be *behind* the mast. This is very important!

The elements are first mounted to the bottom side of the boom to provide weight down below for stability. Then the top-most element is mounted to the boom. The tower person removes the temporary rope securing the boom to the tower and the ground crew uses the pull rope

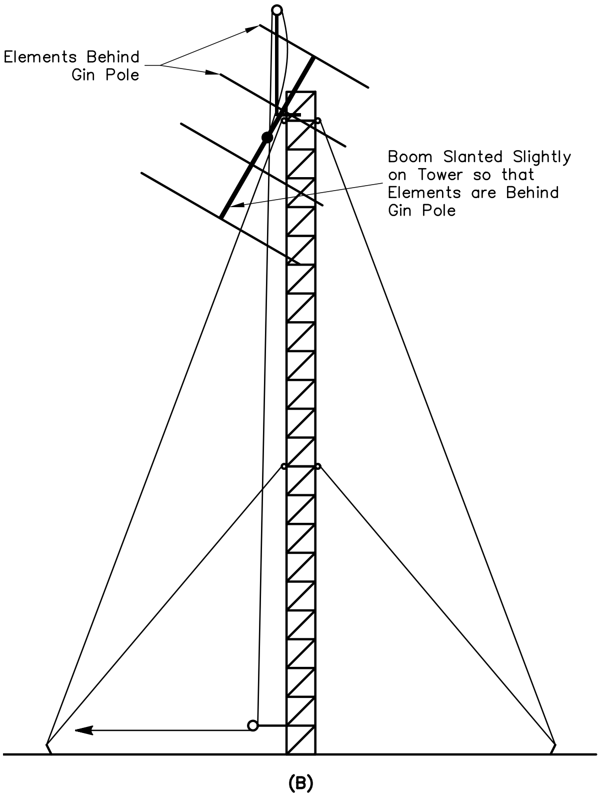
Elements Behind Gin Pole
Boom Slanted Slightly on Tower so that Elements are Behind Gin Pole
(B)

Fig 35—Building a Yagi at the top of the tower. The length of the gin pole must be longer than 1/2 the boom so that the boom can be hoisted upwards to the place where it is mounted to the mast. Usually the boom is initially lashed to the tower slanted slightly from vertical so that the top element ends up *behind* the gin pole. The elements are mounted at the bottom end of the boom first to provide stability. Then the element at the top of the boom is mounted and the boom is moved upwards using the gin-pole hoist rope so that the next-to-top element may be mounted, again behind the gin pole. This process is repeated until all elements are mounted (save possibly the middle element if it can be reached easily from the tower once the beam has been mounted to the mast). Then the boom is tilted to the final position, weaving the elements to clear guy wires if necessary.

to move the mast vertically upwards to the point where the next element from the top can be mounted. Once all the elements are mounted and aligned in the same plane (with perhaps the center element closest to the mast-to-boom bracket left on the ground until later), the temporary securing rope is removed. The boom is now swung so that the elements can be maneuvered to clear the top guy wires. Once the elements are horizontal the boom is secured to the mast and the center element is mounted.

Using a Tram

Another method to get a large Yagi to the top of the tower safely is a *tram*. A tram supports the antenna *under* the tram wire, using a pulley riding on the tram wire. The antenna can thus move more freely—without the friction it would have riding on top of a track rope, as described previously. This puts considerably less strain on the tram wire itself and on the mast to which it is tied on the tower. Some installers prefer to use a wire-rope tramline for its reduced sag.

The tram method uses an easily constructed fixture mounted to the boom of the Yagi to stabilize it from rotating away from the desired attitude as the antenna is raised. A guy-wire cable or heavy rope is fixed to the mast about two feet above the point where the antenna will mount to the mast. A come-along is often used at the ground end to tension the tram wire properly. It is often necessary to back-guy the mast to make sure it doesn't get bent, since the horizontal forces acting on the mast can be considerable in any tram (or track) operation. (Note that the tram technique works well for side-mounted antennas also, where back-guying is not necessary if you are reasonably close to a guy set, as is usually the case.)

Fig 36 is a photograph of a tram fixture built by Kurt Andress, K7NV. This consists of a pulley riding on top of the tramline. This pulley is attached using a caribiner or shackle to two equal-length wires connected to the boom to make an inverted-V shaped *sling*. The two sling wires are secured to the boom using angle irons and muffler clamps so that the antenna is perfectly balanced in the horizontal plane. Balance is very important to make sure the antenna rises properly on the tram without having the boom rotate downward on one end or the other.

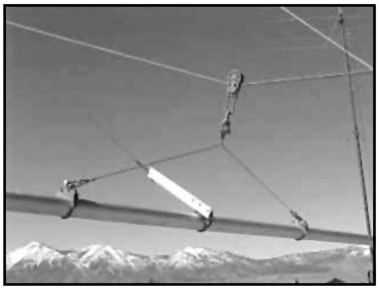

Fig 36—Photo of the tram system used by Kurt Andress, K7NV. (*Photo by K7NV.*)

The hoisting rope running through the tower-mounted gin-pole pulley (and used by the ground crew to pull the antenna up to the tower on the tram line) is attached to a 2-foot piece of angle iron. This is attached to the boom with a muffler clamp. Note that the angle iron is rotated slightly from horizontal so that the plane of the elements is tilted upwards—this allows the elements to clear the guy wires as the antenna is raised. (While the antenna is close to the ground, the angle iron is adjusted so that the elements remain horizontal. Once they are clear of the ground, the angle iron is readjusted to align the elements in the proper direction to clear the guy wires on the tower.) The force of the pull rope along the angle iron also stabilizes the antenna from yawing from side-to-side. Note in Fig 36 that the angle iron is mounted just off-center from where the boom-to-mast plate will be attached so that it clears the mast as the antenna nears the top of the tower. **Fig 37** diagrams how the tram and hoisting line are rigged to the mast.

The tower person directs the activity of the ground crew below and guides the antenna to the mast. Once the end of the pull rope reaches the mast, the tower man ties the boom temporarily to the mast so that he can undo the pull rope from the tram-fixture angle iron and retie it around the boom. Then the antenna can be raised to the point where it can be mounted to the mast.

This technique has been employed to raise Yagis with booms as long as 42 feet at the N6RO contest station on towers as high as 130 feet. As with the track, the tram system requires a good deal of open real estate. While it sounds complicated to set up, you can raise some rather large antennas in less than an hour, once you get the hang of the operation.

THE PVRC MOUNT

The methods described above for hoisting antennas are sometimes not satisfactory for really large, heavy arrays. The best way to handle large Yagis is to assemble them on top of the tower. One way to do this easily is by using the *PVRC Mount*. Many members of the Potomac Valley Radio Club have successfully used this method to install large antennas. Simple and ingenious, the idea involves offsetting the boom from the mast to permit the boom to tilt 360° and rotate axially 360°. This permits the entire length of the boom to be brought alongside the tower, allowing the elements to be attached one by one. (It also allows any part of the antenna to be brought alongside the tower for antenna maintenance.)

See **Figs 38** through **42**. The mount itself consists of a short length of pipe of the same diameter as the rotating mast (or greater), a steel plate, eight U bolts and four pinning bolts. The steel plate is the larger, horizontal one shown in Fig 38. Four U bolts attach the plate to the rotating mast, and four attach the horizontal pipe to the plate. The horizontal pipe provides the offset between the antenna boom and the tower. The antenna boom-to-mast plate is mounted at the outer end of the short pipe. Four bolts are used to ensure that the antenna ends up parallel to the ground, two pinning each plate to the short pipe. When the mast plate pinning bolts are removed and the

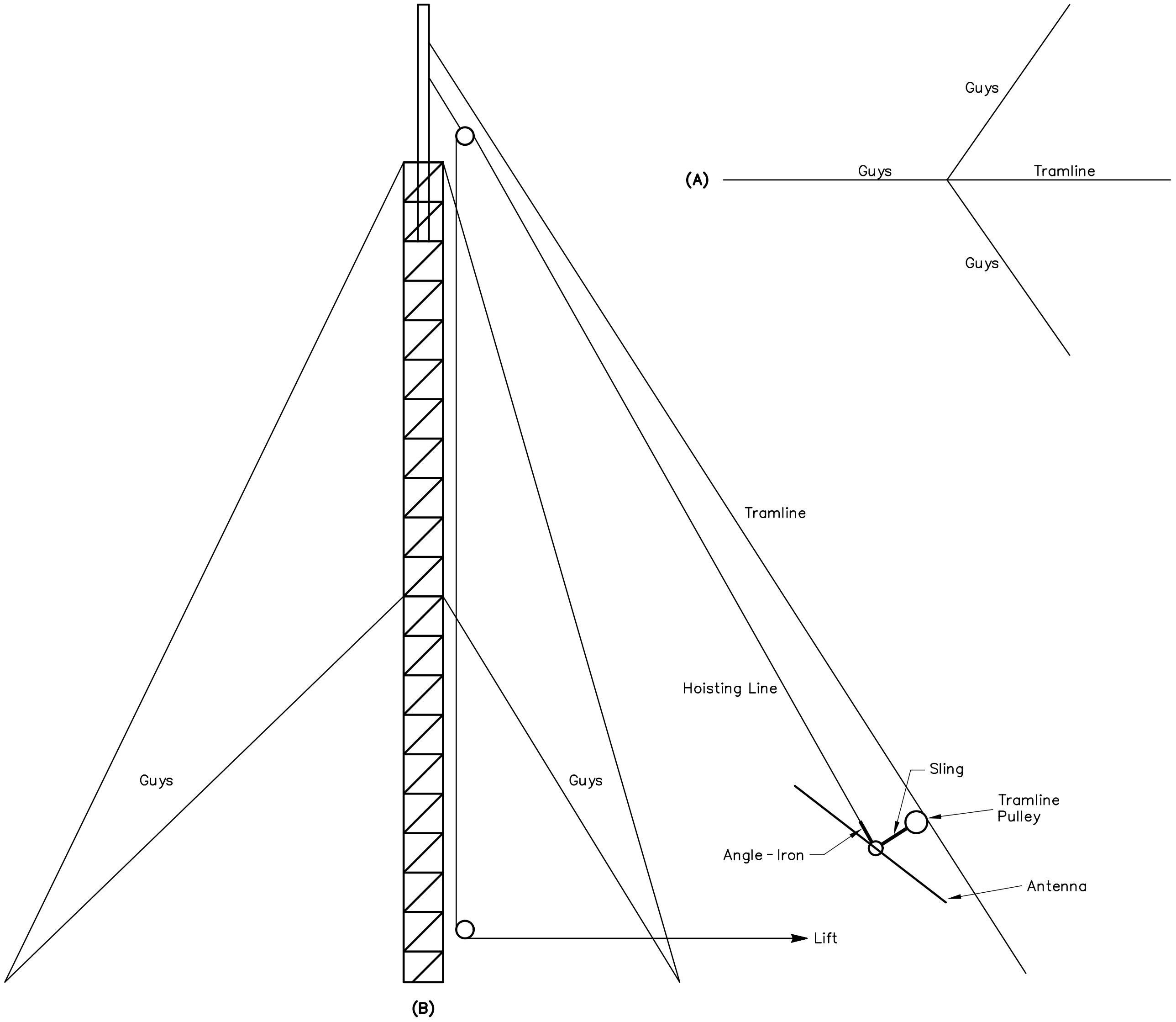

Fig 37—At A, bird's-eye view of tram system used to bring large Yagi antennas from the ground to the top of the tower. At B, side view of rigging used for tramline and hoisting line, along with the sling and tram fixture used to hold the Yagi on the tramline.

four U bolts loosened, the short pipe and boom plate can be rotated through 360°, allowing either half of the boom to come alongside the tower.

First assemble the antenna on the ground. Carefully mark all critical dimensions, and then remove the antenna elements from the boom. Once the rotator and mast have been installed on the tower, a gin pole is used to bring the mast plate and short pipe to the top of the tower. There, the top crew unpins the horizontal pipe and tilts the antenna boom plate to place it in the vertical plane. The boom is attached to the boom plate at the final balance point of the assembled antenna. It is important that the boom be rotated axially so the bottom side of the boom is closest to the tower. This will allow the boom to be tilted without the elements striking the tower.

During installation it may be necessary to loosen one guy wire temporarily to allow for tilting of the boom. As a safety precaution, a temporary guy should be attached to the same leg of the tower just low enough so the assembled antenna will clear it.

The elements are assembled on the boom, starting with those closest to the center of the boom, working out alternately to the farthest director and reflector. This procedure must be followed. If all the elements are put first on one half of the boom, it will be dangerous (if not impossible) to put on the remaining elements. By start-

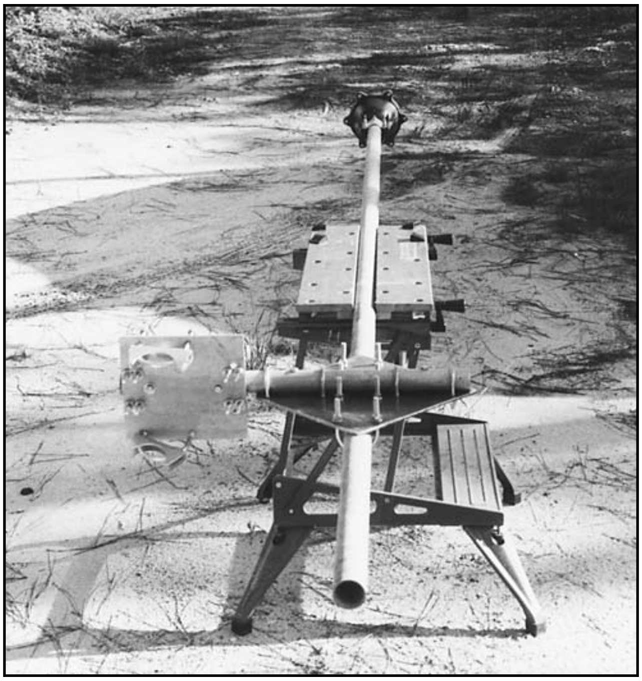

Fig 38—The PVRC mount, boom plate, mast and rotator ready to go. The mast and rotator are installed on the tower first.

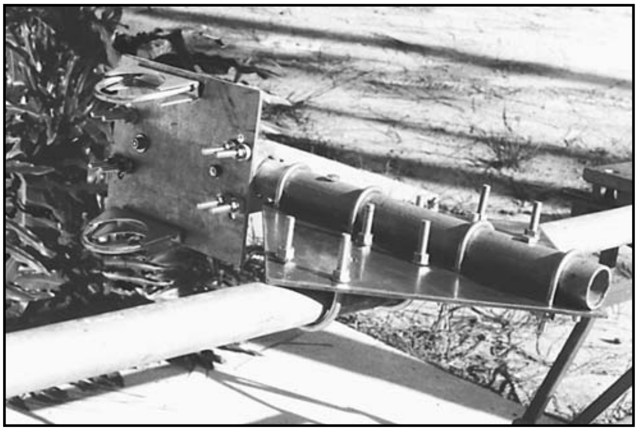

Fig 39—Close-up of the PVRC mount. The long pipe (horizontal in this photo) is the rotating mast. The U bolts in the vertical plate at the left are ready to accept the antenna boom. The heads of two locking pins (bolts) are visible at the midline of the boom plate. The other two pins help secure the horizontal pipe to the large steel mast plate. (The head of the bolt nearest the camera blends in with the right hand leg of the U bolt behind it.)

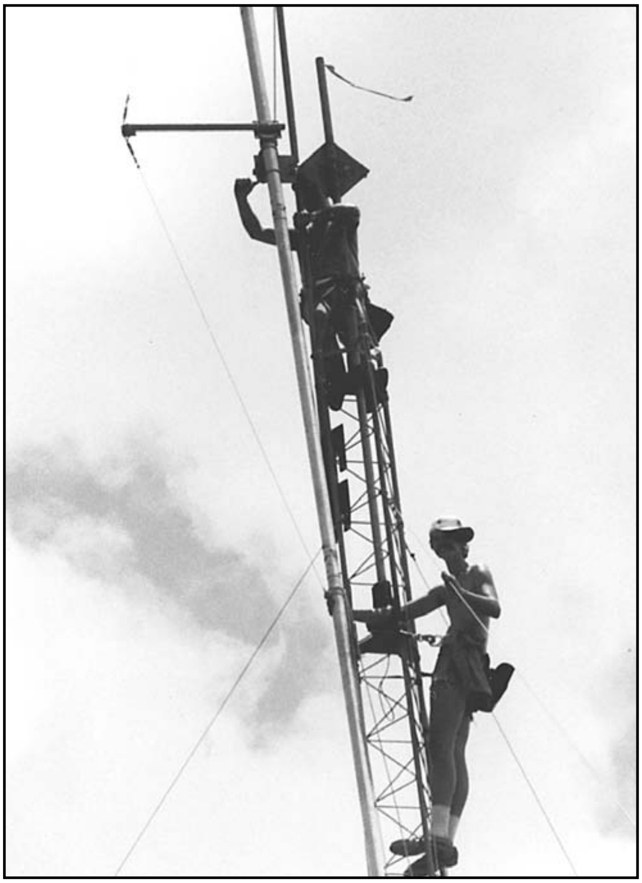

Fig 40—Working at the 70-foot level. A gin pole makes pulling up and mounting the boom to the boom plate a safe and easy procedure.

ing at the middle and working outward, the balance point of the partly assembled antenna will never be so far removed from the tower that tilting of the boom becomes impossible.

When the last element is attached, the boom is brought parallel to the ground, the horizontal pipe is pinned to the mast plate, and the mast plate U bolts tightened. At this point, all the antenna elements will be positioned vertically. Next, loosen the U bolts that hold the boom and rotate the boom axially 90°, bringing the elements parallel to the ground. Tighten the boom bolts and double check all the hardware.

Many long-boom Yagis employ a truss to prevent boom sag. With the PVRC mount, the truss must be attached to a pipe that is independent of the rotating mast. A short length of pipe is attached to the boom as close as possible to the balance point. The truss then moves with the boom whenever the boom is tilted or twisted.

A precaution: Unless you have a really strong rotator, you should consider using this mount mainly for assembling the antenna on the tower. The offset between the boom and the mast with this assembly can generate high torque loads on the rotator. Mounting the boom as close as possible to the mast will minimize the torque when the antenna is pointed into the wind.

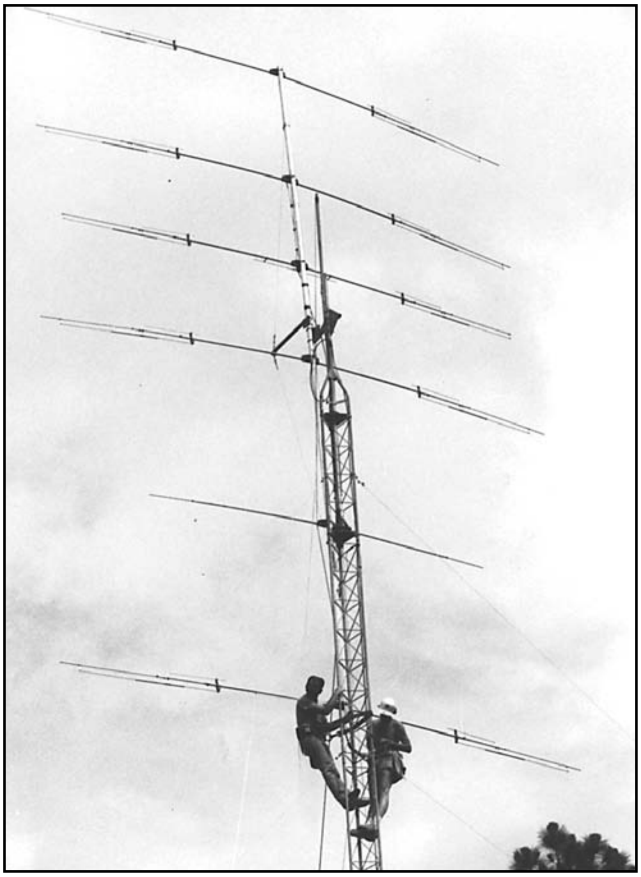

Fig 41—Mounting the last element prior to positioning the boom in a horizontal plane.

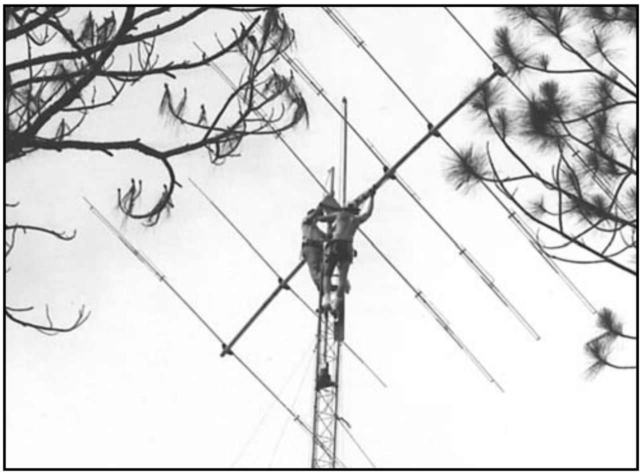

Fig 42—The U bolts securing the short pipe to the mast plate are loosened and the boom is turned to a horizontal position. This puts the elements in a vertical plane. Then the pipe U bolts are tightened and pinning bolts secured. The boom U bolts are then loosened and the boom turned axially 90°.

THE TOWER ALTERNATIVE

A cost saving alternative to the ground-mounted tower is the roof-mounted tripod. Units suitable for small HF or VHF antennas are commercially available. Perhaps the biggest problem with a tripod is determining how to fasten it securely to the roof.

One method of mounting a tripod on a roof is to nail 2 × 6 boards to the undersides of the rafters. Bolts can be extended from the leg mounts through the roof and the 2 × 6s. To avoid exerting too much pressure on the area of the roof between rafters, place another set of 2 × 6s on top of the roof (a mirror image of the ones in the attic). Installation details are shown in **Figs 43** through **46**.

The 2 × 6s are cut 4 inches longer than the outside distance between two rafters. Bolts are cut from a length of ¼-inch-threaded rod. Nails are used to hold the boards in place during installation, and roofing tar is used to seal the area to prevent leaks.

Find a location on the roof that will allow the antenna to turn without obstruction from such things as trees, TV antennas and chimneys. Determine the rafter locations. (Chimneys and vent pipes make good reference points.) Now the tower is set in place atop three 2 × 6s. A plumb line run from the top center of the tower can be used to center it on the peak of the roof. Holes for the mounting bolts can now be drilled through the roof.

Before proceeding, the bottom of the 2 × 6s and the area of the roof under them should be given a coat of roofing tar. Leave about $^1/_8$ inch of clear area around the holes to ensure easy passage of the bolts. Put the tower back in place and insert the bolts and tighten them. Apply tar to the bottom of the legs and the wooden supports, including the bolts. For added security the tripod can be guyed. Guys should be anchored to the frame of the house.

If a rotator is to be mounted above the tripod, pressure will be applied to the bearings. Wind load on the antenna will be translated into a "pinching" of one side of the bearings. Make sure that the rotator is capable of handling this additional stress.

ROTATOR SYSTEMS

There are not that many choices when it comes to antenna rotators for the amateur antenna system. Making the correct decision as to how much capacity the rotator must have is very important to ensure trouble-free operation. Manufacturers generally provide an antenna surface-area rating to help the purchase choose a suitable rotator. The maximum antenna area is linked to the rotator's torque capability.

Some rotator manufacturers provide additional information to help you select the right size of rotator for the antennas you plan to use. Hy-Gain provides an *Effective Moment* value. Yaesu calls theirs a *K-Factor*. Both of these ratings are torque values in foot-pounds. You can compute the effective moment of your antenna by multiplying the antenna turning radius by its weight. So long

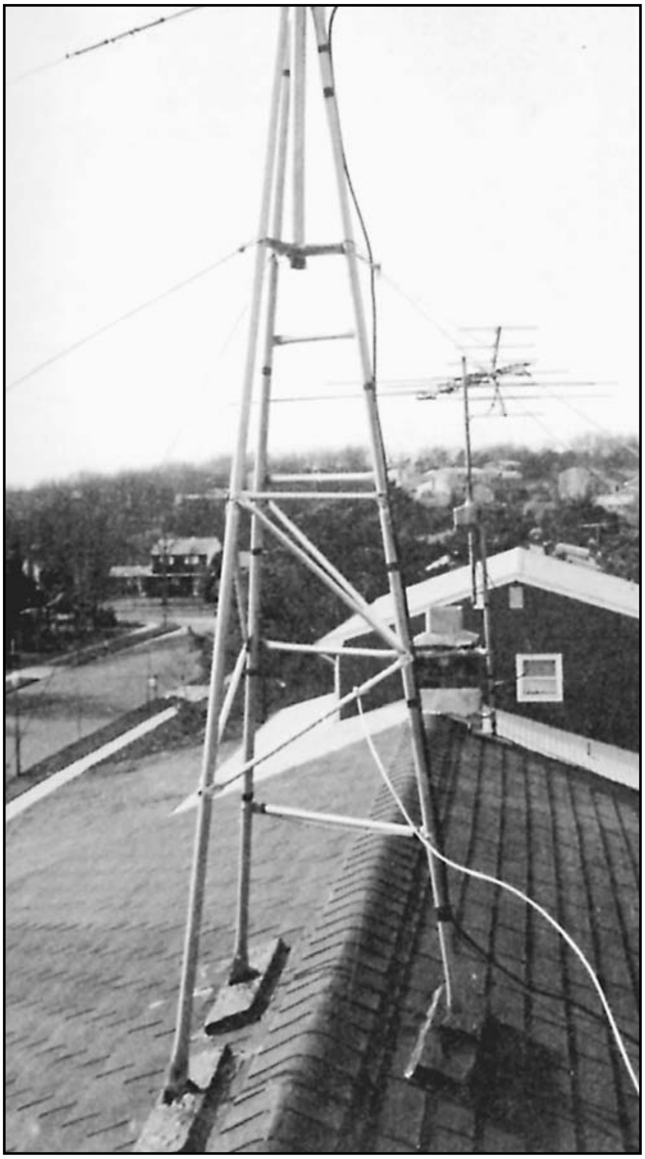

Fig 43—This tripod tower supports a rotary beam antenna. In addition to saving yard space, a roof-mounted tower can be more economical than a ground-mounted tower. A ground lead fastened to the lower part of the frame is for lightning protection. The rotator control cable and the coaxial line are dressed along two of the legs. (*Photo courtesy of Jane Wolfert.*)

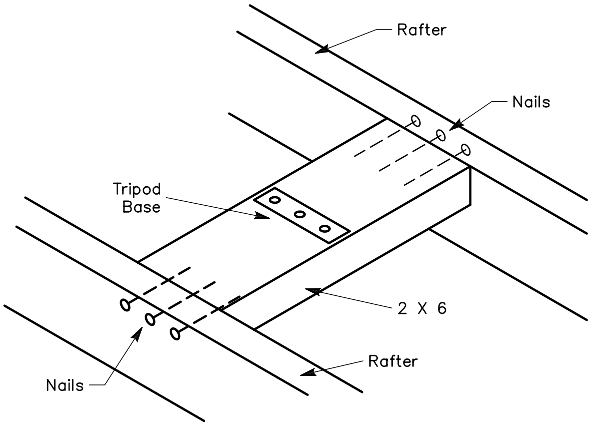

Fig 44—This cutaway view illustrates how the tripod tower is secured to the roof rafters. The leg to be secured to the crosspiece is placed on the outside of the roof. Another cross member is fastened to the underside of the rafters. Bolts, inserted through the roof and the two cross pieces, hold the inner cross member in place because of pressure applied. The inner crosspiece can be nailed to the rafter for added strength.

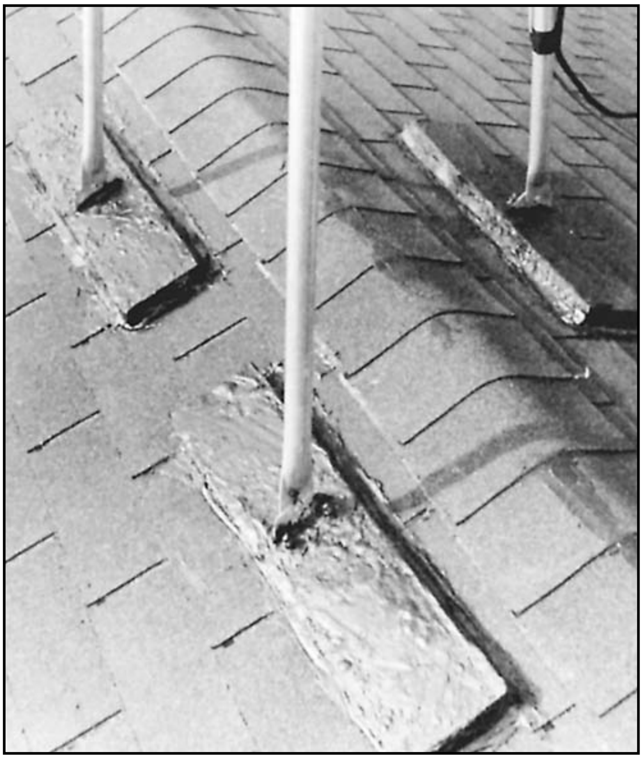

Fig 45—Three lengths of 2 × 6 wood mounted on the outside of the roof and reinforced under the roof by three identical lengths provide a durable means for anchoring the tripod. A thick coat of roofing tar guards against weathering and leaks.

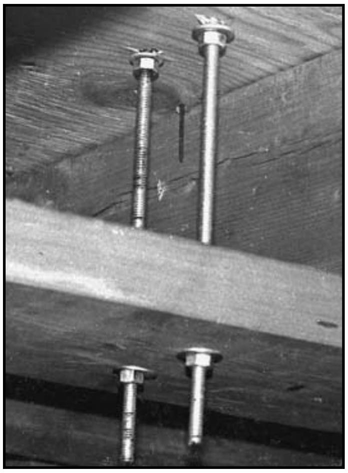

Fig 46—The strengthened anchoring for the tripod. Bolts are placed through two 2 × 6s on the underside of the roof and through the 2 × 6 on the top of the roof, as shown in Fig 43.

as the effective moment rating of the rotator is greater than or equal to the antenna value, the rotator can be expected to provide a useful service life.

There are basically four grades of rotators available to the amateur. The lightest-duty rotator is the type typically used to turn TV antennas. Without much difficulty, these rotators will handle a small 3-element tribander array (14, 21 and 28 MHz) or a single 21- or 28-MHz monoband three-element antenna. The important consideration with a TV rotator is that it lacks braking or holding capability. High winds turn the rotator motor via the gear train in a reverse fashion. Broken gears sometimes result.

The next grade up from the TV class of rotator usually includes a braking arrangement, whereby the antenna is held in place when power is not applied to the rotator. Generally speaking, the brake prevents gear damage on windy days. If adequate precautions are taken, this group of rotators is capable of holding and turning stacked monoband arrays, or up to a five-element 14-MHz system. The next step up in rotator strength is more expensive. This class of rotator will turn just about anything the most demanding amateur might want to install.

A description of antenna rotators would not be complete without the mention of the *prop pitch* class. The prop pitch rotator system consists of a surplus aircraft propeller blade pitch motor coupled to an indicator system and a power supply. There are mechanical problems of installation, however, resulting mostly from the size and weight of these motors. It has been said that a prop pitch rotator system, properly installed, is capable of turning a house. Perhaps in the same class as the prop pitch motor (but with somewhat less capability) is the electric motor of the type used for opening garage doors. These have been used successfully in turning large arrays.

Proper installation of the antenna rotator can provide many years of trouble-free service; sloppy installation can cause problems such as a burned out motor, slippage, binding and casting breakage. Most rotators are capable of accepting mast sizes of different diameters, and suitable precautions must be taken to shim an undersized mast to ensure dead-center rotation. It is very desirable to mount the rotator inside and as far below the top of the tower as possible. The mast absorbs the torsion developed by the antenna during high winds, as well as during starting and stopping.

Some amateurs have used a long mast from the top to the base of the tower. Rotator installation and service can be accomplished at ground level. A mast length of 10 feet or more between the rotator and the antenna will add greatly to the longevity of the entire system by allowing the mast to act as a torsion shock absorber. Another benefit of mounting the rotator 10 feet or more below the antenna is that any misalignment among the rotator, mast and the top of the tower is less significant. A tube at the top of the tower (a sleeve bearing) through which the mast protrudes almost completely eliminates any lateral forces on the rotator casing. All the rotator must do is support the downward weight of the antenna system and turn the array.

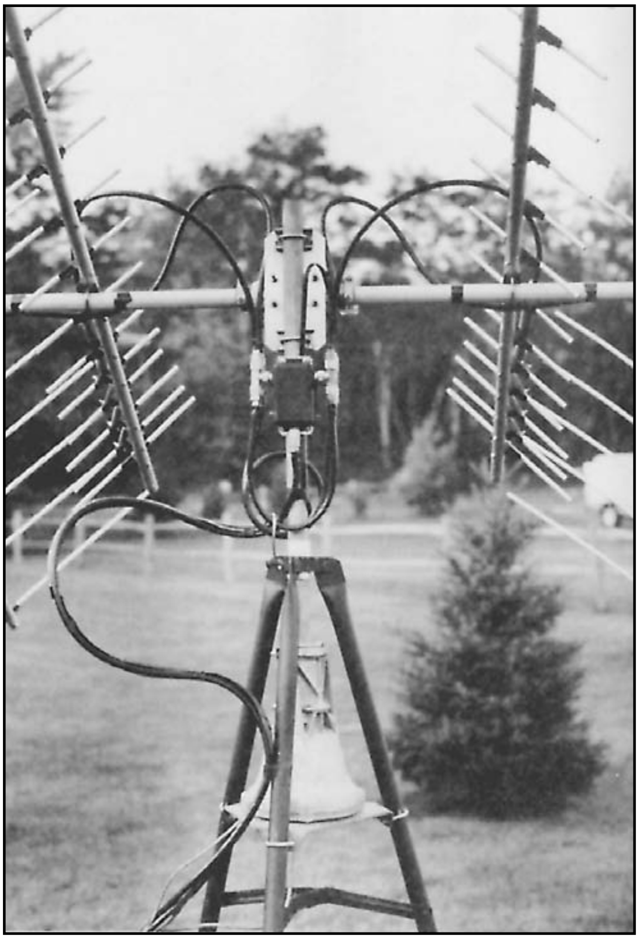

Fig 47—Rotator loop for UHF Yagi array. The feed coax is bundled with a control cable for the polarization relay. The coax/control-cable loop is taped to the rotating mast and to the top of the tower with vinyl electric tape to allow the array to be rotated.

While the normal weight of the antenna and the mast is usually not more than a couple of hundred pounds, even with a large system, one can ease this strain on the rotator by installing a thrust bearing at the top of the tower. The bearing is then the component that holds the weight of the antenna system, and the rotator need perform only the rotating task.

Don't forget to provide a loop of coax to allow your beam to rotate properly. Make sure you position the rotator loop so that it doesn't snag on anything. **Fig 47** shows a rotator loop for an elliptically polarized UHF Yagi array. Note that the coax loop is taped to the rotating mast above the top of the tower and to the tower itself.

Indicator Alignment

A problem often encountered in amateur installations is that of misalignment between the direction indicator in the rotator control box and the heading of the antenna. With a light duty rotator, this happens frequently when the wind blows the antenna to a different heading. With no brake, the force of the wind can move the gear train and motor of the rotator, while the indicator remains fixed. Such rotator systems have a mechanical stop to prevent continuous rotation during operation, and provision is usually included to realign the indicator against the mechanical stop from inside the shack. During installation, the antenna must be oriented correctly for the mechanical stop position, which is usually north.

In larger rotator systems with an adequate brake, indicator misalignment is caused by mechanical slippage in the antenna boom-to-mast hardware. Many texts suggest that the boom be pinned to the mast with a heavy-duty bolt and the rotator be similarly pinned to the mast. There is a trade-off here. If there is sufficient wind to cause slippage in the couplings without pins, with pins the wind could break a rotator casting. The slippage will act as a clutch release, which may prevent serious damage to the rotator. On the other hand, the amateur might not like to climb the tower and realign the system after each heavy windstorm.

BIBLIOGRAPHY

Source material and more extended discussions of the topics covered in this chapter can be found in the references listed below and in the texts listed at the end of Chapter 2.

L. H. Abraham, “Guys for Guys Who Have To Guy,” *QST*, Jun 1955.

K. Baker, “A Ladder Mast,” *QST*, Jun 1981, p 24.

W. R. Gary, “Toward Safer Antenna Installations,” *QST*, Jan 1980, p 56.

S. F. Hoerner, “Fluid Dynamic Drag,” published by author, Bricktown, New Jersey, 1965, pp 1-10.

C. L. Hutchinson, “A Tree-Mounted 30-Meter Ground-Plane Antenna,” *QST*, Sep 1984, pp 16-18.

C. L. Hutchinson, R. D. Straw, *Simple and Fun Antennas for Hams* (Newington: ARRL, 2002).

M. P. Keown and L. L. Lamb, “A Simple Technique for Tower-Section Separation,” *QST*, Sep 1979, pp 37-38.

P. O’Dell, “The Ups and Downs of Towers,” *QST*, Jul 1981, p 35.

S. Phillabaum, “Installation Techniques for Medium and Large Yagis,” *QST*, Jun 1979, p 23.

C. J. Richards, “Mechanical Engineering in Radar and Communications” (London: Van Nostrand Reinhold Co., 1969) pp 162-165.

D. Weber “Determination of Yagi Wind Loads Using the Cross-Flow Principle,” *Communications Quarterly*, Spring 1993.

A. B. White, “A Delayed Brake Release for the Ham-II,” *QST*, Aug 1977, p 14.

B. White, E. White and J. White, “Assembling Big Antennas on Fixed Towers,” *QST*, Mar 1982, pp 28-29

L. Wolfert, “The Tower Alternative,” *QST*, Nov 1980, p 36.

W. C. Young., “Roark’s Formulas for Stress & Strain” (New York: McGraw-Hill Co., 1989), pp 67, 96

Structural Standards for Steel Antenna Towers and Antenna Supporting Structures, TIA/EIA Standard TIA/EIA-222-F, Electronic Industries Association, Mar 1996. May be purchased from Telecommunications Industry Association (TIA), Standards and Technology Department, 2500 Wilson Boulevard, Arlington VA, 22201 or Global Engineering Documents, 15 Inverness Way East, Englewood, CO, 80112-5704, 1-800-854-7179.

Chapter 23

Radio Wave Propagation

Because radio communication is carried on by means of electromagnetic waves traveling through the Earth's atmosphere, it is important to understand the nature of these waves and their behavior in the propagation medium. Most antennas will radiate the power applied to them efficiently, but no antenna can do all things equally well, under all circumstances. Whether you design and build your own antennas, or buy them and have them put up by a professional, you'll need propagation know-how for best results, both during the planning stages and while operating your station.

For station planning, this chapter contains detailed new information on elevation angles from transmitting locations throughout the world to important areas throughout the world. With this information in hand, you can design your own antenna installation for optimum capabilities possible within your budget. See the CD-ROM in back of this book.

The Nature of Radio Waves

You probably have some familiarity with the concept of electric and magnetic fields. A radio wave is a combination of both, with the energy divided equally between them. If the wave could originate at a point source in free space, it would spread out in an ever-growing sphere, with the source at the center. No antenna can be designed to do this, but the theoretical *isotropic antenna* is useful in explaining and measuring the performance of practical antennas we *can* build. It is, in fact, the basis for any discussion or evaluation of antenna performance.

Our theoretical spheres of radiated energy would expand very rapidly at the same speed as the propagation of light, approximately 186,000 miles or 300,000,000 meters per second. These values are close enough for practical purposes, and are used elsewhere in this book. If one wishes to be more precise, light propagates in a vacuum at the speed of 299.7925 meters per microsecond, and slightly slower in air.

The path of a *ray* traced from its source to any point on a spherical surface is considered to be a straight line—a radius of the sphere. An observer on the surface of the sphere would think of it as being flat, just as the Earth seems flat to us. A radio wave far enough from its source to appear flat is called a *plane wave*. From here on, we will be discussing primarily plane waves.

It helps to understand the radiation of electromagnetic energy if we visualize a plane wave as being made up of electric and magnetic forces, as shown in **Fig 1**. The nature of wave propagation is such that the electric and magnetic lines of force are always perpendicular. The plane containing the sets of crossed lines represents the wave front. The direction of travel is always perpendicular to the wave front; *forward* or *backward* is determined by the relative directions of the electric and magnetic forces.

The speed of travel of a wave through anything but a vacuum is always less than 300,000,000 meters per second. How much less depends on the medium. If it is air, the reduction in propagation speed can be ignored in most discussions of propagation at frequencies below 30 MHz. In the VHF range and higher, temperature and moisture content of the medium have increasing effects on the communication range, as will be discussed later. In solid insulating materials the speed is considerably less. In distilled water (a good insulator) the speed is $^1/_9$ that in free space. In good conductors the speed is so low that the opposing fields set up by the wave front occupy practically the same space as the wave itself, and thus cancel it out. This is the reason for "skin effect" in conductors at high frequencies, making metal enclosures good shields for electrical circuits working at radio frequencies.

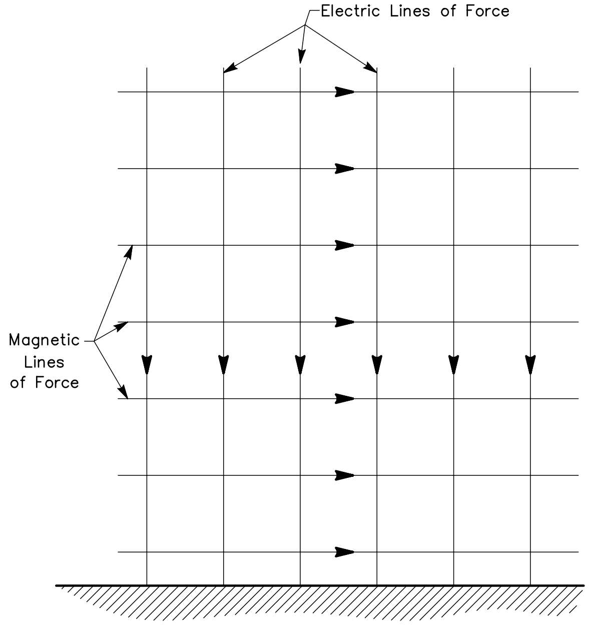

Fig 1—Representation of the magnetic and electric fields of a vertically polarized plane wave traveling along the ground. The arrows indicate instantaneous directions of the fields for a wave traveling perpendicularly out of the page toward the reader. Reversal of the direction of one set of lines reverses the direction of travel. There is no change in direction when both sets are reversed. Such a dual reversal occurs in fact once each half cycle.

Phase and Wavelength

Because the velocity of wave propagation is so great, we tend to ignore it. Only $^1/_7$ of a second is needed for a radio wave to travel around the world—but in working with antennas the *time* factor is extremely important. The wave concept evolved because an alternating current flowing in a wire (antenna) sets up moving electric and magnetic fields. We can hardly discuss antenna theory or performance at all without involving travel time, consciously or otherwise.

Waves used in radio communication may have frequencies from about 10,000 to several billion Hz. Suppose the frequency is 30 MHz. One cycle, or period, is completed in 1/30,000,000 second. The wave is traveling at 300,000,000 meters per second, so it will move only 10 meters during the time that the current is going through one complete period of alternation. The electromagnetic field 10 meters away from the antenna is caused by the current that was flowing one period earlier in time. The field 20 meters away is caused by the current that was flowing two periods earlier, and so on.

If each period of the current is simply a repetition of the one before it, the currents at corresponding instants in each period will be identical. The fields caused by those currents will also be identical. As the fields move outward from the antenna they become more thinly spread over larger and larger surfaces. Their amplitudes decrease with distance from the antenna but they do not lose their identity with respect to the instant of the period at which they were generated. They are, and they remain, *in phase*. In the example above, at intervals of 10 meters measured outward from the antenna, the phase of the waves at any given instant is identical.

From this information we can define both *wave front* and *wavelength*. Consider the wave front as an imaginary surface. On every part of this surface, the wave is in the same phase. The wavelength is the distance between two wave fronts having the same phase at any given instant. This distance must be measured perpendicular to the wave fronts along the line that represents the direction of travel. The abbreviation for wavelength is the Greek letter lambda, λ, which is used throughout this book.

The wavelength will be in the same length units as the velocity when the frequency is expressed in the same time units as the velocity. For waves traveling in free space (and near enough for waves traveling through air) the wavelength is

$$\lambda_{\text{meters}} = \frac{299.7925}{F(\text{MHz})} \qquad \text{(Eq 1)}$$

There will be few pages in this book where phase, wavelength and frequency do not come into the discussion. It is essential to have a clear understanding of their meaning in order to understand the design, installation, adjustment or use of antennas, matching systems or transmission lines in detail. In essence, *phase* means *time*. When something goes through periodic variations, as an alternating current does, corresponding instants in succeeding periods are *in phase*.

The points A, B and C in **Fig 2** are all in phase. They are corresponding instants in the current flow, at 1-λ intervals. This is a conventional view of a sine-wave alternating current, with time progressing to the right. It also represents a *snapshot* of the intensity of the traveling fields, if distance is substituted for time in the horizontal axis. The distance between A and B or between B and C is one wavelength. The field-intensity distribution follows the sine curve, in both amplitude and polarity, corresponding exactly to the time variations in the current that produced the fields. Remember that this is an *instantaneous* picture—the wave moves outward, much as a wave created by a rock thrown into water does.

Polarization

A wave like that in Fig 1 is said to be *polarized* in the direction of the electric lines of force. The polarization here is vertical, because the electric lines are perpendicular to the surface of the Earth. It is one of the laws of electromagnetics that electric lines touching the surface of a perfect conductor must do so perpendicularly, or else they would have to generate infinite currents in the conductor, an obvious impossibility. Most ground is a rather good conductor at frequencies below about 10 MHz, so waves at these

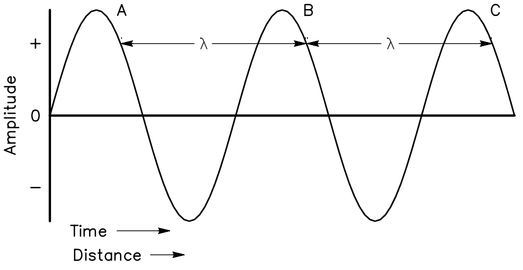

Fig 2—The instantaneous amplitude of both fields (electric and magnetic) varies sinusoidally with time as shown in this graph. Since the fields travel at constant velocity, the graph also represents the instantaneous distribution of field intensity along the wave path. The distance between two points of equal phase such as A-B and B-C is the length of the wave.

frequencies, traveling close to good ground, are mainly vertically polarized. Over partially conducting ground there may be a forward tilt to the wave front; the tilt in the electric lines of force increases as the energy loss in the ground becomes greater.

Waves traveling in contact with the surface of the Earth, called *surface waves*, are of little practical use in amateur communication. This is because as the frequency is raised, the distance over which they will travel without excessive energy loss becomes smaller and smaller. The surface wave is most useful at low frequencies and through the standard AM broadcast band. The surface wave will be covered later. At high frequencies a wave reaching a receiving antenna has had little contact with the ground, and its polarization is not necessarily vertical.

If the electric lines of force are horizontal, the wave is said to be *horizontally polarized*. Horizontally and vertically polarized waves may be classified generally under *linear polarization*. Linear polarization can be anything between horizontal and vertical. In free space, "horizontal" and "vertical" have no meaning, since the reference of the seemingly horizontal surface of the Earth has been lost.

In many cases the polarization of waves is not fixed, but rotates continually, somewhat at random. When this occurs the wave is said to be *elliptically polarized*. A gradual shift in polarization in a medium is known as *Faraday rotation*. For space communication, circular polarization is commonly used to overcome the effects of Faraday rotation. A circularly polarized wave rotates its polarization through 360° as it travels a distance of one wavelength in the propagation medium. The direction of rotation as viewed from the transmitting antenna defines the direction of circularity—right-hand (clockwise) or left-hand (counterclockwise). Linear and circular polarization may be considered as special cases of elliptical polarization.

Field Intensity

The energy from a propagated wave decreases with distance from the source. This decrease in strength is caused by the spreading of the wave energy over ever-larger spheres as the distance from the source increases.

A measurement of the strength of the wave at a distance from the transmitting antenna is its *field intensity*, which is synonymous with *field strength*. The strength of a wave is measured as the voltage between two points lying on an electric line of force in the plane of the wave front. The standard of measure for field intensity is the voltage developed in a wire that is 1 meter long, expressed as volts per meter. (If the wire were 2 meters long, the voltage developed would be divided by two to determine the field strength in volts per meter.)

The voltage in a wave is usually low, so the measurement is made in millivolts or microvolts per meter. The voltage goes through time variations like those of the current that caused the wave. It is measured like any other ac voltage—in terms of the effective value or, sometimes, the peak value. It is fortunate that in amateur work it is not necessary to measure actual field strength, as the equipment required is elaborate. We need to know only if an adjustment has been beneficial, so relative measurements are satisfactory. These can be made easily with home-built equipment.

Wave Attenuation

In free space, the field intensity of the wave varies inversely with the distance from the source, once you are in the radiating far field of the antenna. If the field strength at 1 mile from the source is 100 millivolts per meter, it will be 50 millivolts per meter at 2 miles, and so on. The relationship between field intensity and power density is similar to that for voltage and power in ordinary circuits. They are related by the impedance of free space, which is approximately 377 Ω. A field intensity of 1 volt per meter is therefore equivalent to a power density of

$$P = \frac{E^2}{Z} = \frac{1\,(\text{volt/m})^2}{377\,\Omega} = 2.65\,\text{mW/m}^2 \qquad \text{(Eq 2)}$$

Because of the relationship between voltage and power, the power density therefore varies with the square root of the field intensity, or inversely with the *square* of the distance. If the power density at 1 mile is 4 mW per square meter, then at a distance of 2 miles it will be 1 mW per square meter.

It is important to remember this so-called *spreading loss* when antenna performance is being considered. Gain can come only from narrowing the radiation pattern of an antenna, which concentrates the radiated energy in the desired direction. There is no "antenna magic" by which the total energy radiated can be increased.

In practice, attenuation of the wave energy may be much greater than the inverse-distance law would indicate. The wave does not travel in a vacuum, and the receiving antenna seldom is situated so there is a clear line of sight.

The Earth is spherical and the waves do not penetrate its surface appreciably, so communication beyond visual distances must be by some means that will bend the waves around the curvature of the Earth. These means involve additional energy losses that increase the path attenuation with distance, above that for the theoretical spreading loss in a vacuum.

Bending of Radio Waves

Radio waves and light waves are both propagated as electromagnetic energy. Their major difference is in wavelength, since radio-reflecting surfaces are usually much smaller in terms of wavelength than those for light. In material of a given electrical conductivity, long waves penetrate deeper than short ones, and so require a thicker mass for good reflection. Thin metal however is a good reflector of even long-wavelength radio waves. With poorer conductors, such as the Earth's crust, long waves may penetrate quite a few feet below the surface.

Reflection occurs at any boundary between materials of differing dielectric constant. Familiar examples with light are reflections from water surfaces and window panes. Both water and glass are transparent for light, but their dielectric constants are very different from that of air. Light waves, being very short, seem to bounce off both surfaces. Radio waves, being much longer, are practically unaffected by glass, but their behavior upon encountering water may vary, depending on the purity of that medium. Distilled water is a good insulator; salt water is a relatively good conductor.

Depending on their wavelength (and thus their frequency), radio waves may be reflected by buildings, trees, vehicles, the ground, water, ionized layers in the upper atmosphere, or at boundaries between air masses having different temperatures and moisture content. Ionospheric and atmospheric conditions are important in practically all communication beyond purely local ranges.

Refraction is the bending of a ray as it passes from one medium to another at an angle. The appearance of bending of a straight stick, where it enters water at an angle, is an example of light refraction known to us all. The degree of bending of radio waves at boundaries between air masses increases with the radio frequency. There is slight atmospheric bending in our HF bands. It becomes noticeable at 28 MHz, more so at 50 MHz, and it is much more of a factor in the higher VHF range and in UHF and microwave propagation.

Diffraction of light over a solid wall prevents total darkness on the far side from the light source. This is caused largely by the spreading of waves around the top of the wall, due to the interference of one part of the beam with another. The dielectric constant of the surface of the obstruction may affect what happens to our radio waves when they encounter terrestrial obstructions—but the radio *shadow area* is never totally dark. See Chapter 3, The Effects of Ground, for more information on diffraction.

The three terms, reflection, refraction and diffraction, were in use long before the radio age began. Radio propagation is nearly always a mix of these phenomena, and it may not be easy to identify or separate them while they are happening when we are on the air. This book tends to rely on the words *bending* and *scattering* in its discussions, with appropriate modifiers as needed. The important thing to remember is that any alteration of the path taken by energy as it is radiated from an antenna is almost certain to affect on-the-air results—which is why this chapter on propagation is included in an antenna book.

GROUND WAVES

As we have already seen, radio waves are affected in many ways by the media through which they travel. This has led to some confusion of terms in earlier literature concerning wave propagation. Waves travel close to the ground in several ways, some of which involve relatively little contact with the ground itself. The term *ground wave* has had several meanings in antenna literature, but it has come to be applied to any wave that stays close to the Earth, reaching the receiving point without leaving the Earth's lower atmosphere. This distinguishes the ground wave from a *sky wave*, which utilizes the ionosphere for propagation between the transmitting and receiving antennas.

The ground wave could be traveling in actual contact with the ground, as in Fig 1, where it is called the *surface wave*. Or it could travel directly between the transmitting and receiving antennas, when they are high enough so they can "see" each other—this is commonly called the *direct wave*. The ground wave also travels between the transmitting and receiving antennas by reflections or diffractions off intervening terrain between them. The ground-influenced wave may interact with the direct wave to create a vector-summed resultant at the receiver antenna.

In the generic term ground wave, we also will include ones that are made to follow the Earth's curvature by bending in the Earth's lower atmosphere, or *troposphere*, usually no more than a few miles above the ground. Often called *tropospheric bending*, this propagation mode is a major factor in amateur communications above 50 MHz.

THE SURFACE WAVE

The surface wave travels in contact with the Earth's surface. It can provide coverage up to about 100 miles in the standard AM broadcast band during the daytime, but attenuation is high. As can be seen from **Fig 3**, the attenuation increases with frequency. The surface wave is of little value in amateur communication, except possibly at 1.8 MHz. Vertically polarized antennas must be used, which tends to limit amateur surface-wave communication to where large vertical systems can be erected.

THE SPACE WAVE

Propagation between two antennas situated within line of sight of each other is shown in **Fig 4**. Energy traveling directly between the antennas is attenuated to about the same degree as in free space. Unless the antennas are very high or

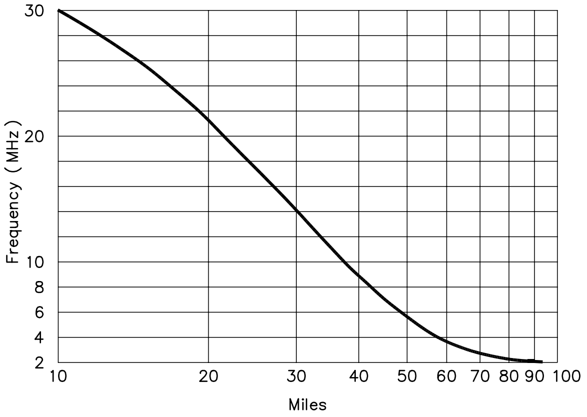

Fig 3—Typical HF ground-wave range as a function of frequency.

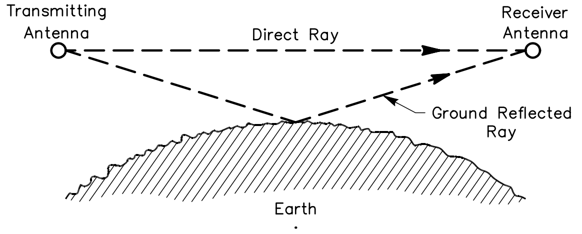

Fig 4—The ray traveling directly from the transmitting antenna to the receiving antenna combines with a ray reflected from the ground to form the space wave. For a horizontally polarized signal a reflection as shown here reverses the phase of the ground-reflected ray.

quite close together, an appreciable portion of the energy is reflected from the ground. This reflected wave combines with direct radiation to affect the actual signal received.

In most communication between two stations on the ground, the angle at which the wave strikes the ground will be small. For a horizontally polarized signal, such a reflection reverses the phase of the wave. If the distances traveled by both parts of the wave were the same, the two parts would arrive out of phase, and would therefore cancel each other. The ground-reflected ray in Fig 4 must travel a little further, so the phase difference between the two depends on the lengths of the paths, measured in wavelengths. The wavelength in use is important in determining the useful signal strength in this type of communication.

If the difference in path length is 3 meters, the phase difference with 160-meter waves would be only 360° × 3/160 = 6.8°. This is a negligible difference from the 180° shift caused by the reflection, so the effective signal strength over the path would still be very small because of cancellation of the two waves. But with 6-meter radio waves the phase length would be 360° × 3/6 = 180°. With the additional 180° shift on reflection, the two rays would add. Thus, the space wave is a negligible factor at low frequencies, but it can be increasingly useful as the frequency is raised. It is a dominant factor in local amateur communication at 50 MHz and higher.

Interaction between the direct and reflected waves is the principle cause of *mobile flutter* observed in local VHF communication between fixed and mobile stations. The flutter effect decreases once the stations are separated enough so that the reflected ray becomes inconsequential. The reflected energy can also confuse the results of field-strength measurements during tests on VHF antennas.

As with most propagation explanations, the space-wave picture presented here is simplified, and practical considerations dictate modifications. There is always some energy loss when the wave is reflected from the ground. Further, the phase of the ground-reflected wave is not shifted exactly 180°, so the waves never cancel completely. At UHF, ground-reflection losses can be greatly reduced or eliminated by using highly directive antennas. By confining the antenna pattern to something approaching a flashlight beam, nearly all the energy is in the direct wave. The resulting energy loss is low enough that microwave relays, for example, can operate with moderate power levels over hundreds or even thousands of miles. Thus we see that, while the space wave is inconsequential below about 20 MHz, it can be a prime asset in the VHF realm and higher.

VHF/UHF PROPAGATION BEYOND LINE OF SIGHT

From Fig 4 it appears that use of the space wave depends on direct line of sight between the antennas of the communicating stations. This is not literally true, although that belief was common in the early days of amateur communication on frequencies above 30 MHz. When equipment became available that operated more efficiently and after antenna techniques were improved, it soon became clear that VHF waves were actually being bent or scattered in several ways, permitting reliable communication beyond visual distances between the two stations. This was found true even with low power and simple antennas. The average communication range can be approximated by assuming the waves travel in straight lines, but with the Earth's radius increased by one-third. The distance to the *radio horizon* is then given as

$$D_{miles} = 1.415\sqrt{H_{feet}} \qquad \text{(Eq 3)}$$

or

$$D_{km} = 4.124\sqrt{H_{meters}} \qquad \text{(Eq 4)}$$

where H is the height of the transmitting antenna, as shown in **Fig 5**. The formula assumes that the Earth is smooth out to the horizon, so any obstructions along the path must be taken into consideration. For an elevated receiving antenna the communication distance is equal to D + D1, that is, the sum of the distances to the horizon of both

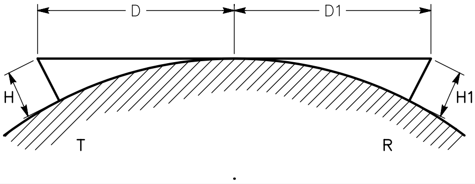

Fig 5—The distance D to the horizon from an antenna of height H is given by equations in the text. The maximum line-of-sight distance between two elevated antennas is equal to the sum of their distances to the horizon as indicated here.

antennas. Radio horizon distances are given in graphic form in **Fig 6**. Two stations on a flat plain, one with its antenna 60 feet above ground and the other 40 feet, could be up to about 20 miles apart for strong-signal line-of-sight communication (11 + 9 mi). The terrain is almost never completely flat, however, and variations along the way may add to or subtract from the distance for reliable communication. Remember that energy is absorbed, reflected or scattered in many ways in nearly all communication situations. The formula or the chart will be a good guide for estimating the potential radius of coverage for a VHF FM repeater, assuming the users are mobile or portable with simple, omnidirectional antennas. Coverage with optimum home-station equipment, high-gain directional arrays, and SSB or CW is quite a different matter. A much more detailed method for estimating coverage on frequencies above 50 MHz is given later in this chapter.

For maximum use of the ordinary space wave it is important to have the antenna as high as possible above nearby buildings, trees, wires and surrounding terrain. A hill that rises above the rest of the countryside is a good location for an amateur station of any kind, and particularly so for extensive coverage on the frequencies above 50 MHz. The highest point on such an eminence is not necessarily the best location for the antenna. In the example shown in **Fig 7**, the hilltop would be a good site in all directions. But if maximum performance to the right is the objective, a point just below the crest might do better. This would involve a trade-off with reduced coverage in the opposite direction. Conversely, an antenna situated on the left side, lower down the hill, might do well to the left, but almost certainly would be inferior in performance to the right.

Selection of a home site for its radio potential is a complex business, at best. A VHF enthusiast dreams of the highest hill. The DX-minded HF ham may be more attracted by a dry spot near a salt marsh. A wide saltwater horizon, especially from a high cliff, just smells of DX. In shopping for ham radio real estate, a mobile or portable rig for the frequencies you're most interested in can provide useful clues.

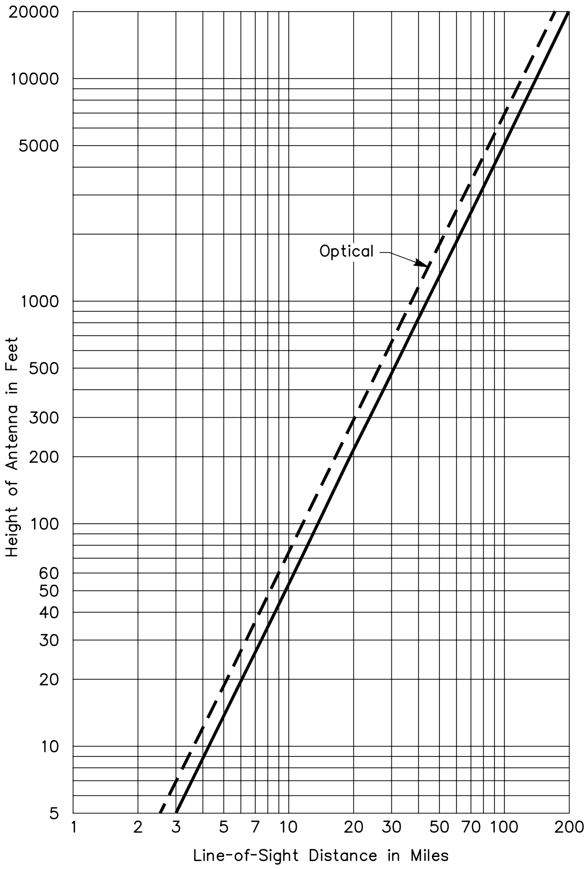

Fig 6—Distance to the horizon from an antenna of given height. The solid curve includes the effect of atmospheric retraction. The optical line-of-sight distance is given by the broken curve.

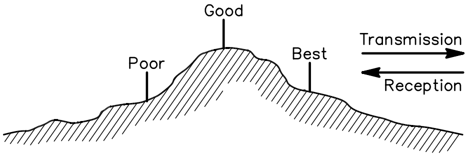

Fig 7—Propagation conditions are generally best when the antenna is located slightly below the top of a hill on the side facing the distant station. Communication is poor when there is a sharp rise immediately in front of the antenna in the direction of communication.

ANTENNA POLARIZATION

If effective communication over long distances were the only consideration, we might be concerned mainly with radiation of energy at the lowest possible angle above the horizon. However, being engaged in a residential avocation often imposes practical restrictions on our antenna projects. As an example, our 1.8 and 3.5-MHz bands are used primarily for short-distance communication because they serve that purpose with antennas that are not difficult or expensive to put up. Out to a few hundred miles, simple wire antennas for these bands do well, even though their radiation is mostly at high angles above the horizon. Vertical systems might be better for long-distance use, but they require extensive ground systems for good performance.

Horizontal antennas that radiate well at low angles are most easily erected for 7 MHz and higher frequencies—horizontal wires and arrays are almost standard practice for work on 7 through 29.7 MHz. Vertical antennas, such as a single omnidirectional antenna of multiband design, are also used in this frequency range. An antenna of this type may be a good solution to the space problem for a city dweller on a small lot, or even for the resident of an apartment building.

High-gain antennas are almost always used at 50 MHz and higher frequencies, and most of them are horizontal. The principal exception is mobile communication with FM through repeaters, discussed in Chapter 17, Repeater Antenna Systems. The height question is answered easily for VHF enthusiasts—the higher the better.

The theoretical and practical effects of height above ground at HF are treated in detail in Chapter 3, The Effects of Ground. Note that it is the height in *wavelengths* that is important—a good reason to think in the metric system, rather than in feet and inches.

In working locally on any amateur frequency band, best results will be obtained with the same polarization at both stations, except on rare occasions when polarization shift is caused by terrain obstructions or reflections from buildings. Where such a shift is observed, mostly above 100 MHz or so, horizontal polarization tends to work better than vertical. This condition is found primarily on short paths, so it is not too important. Polarization shift may occur on long paths where tropospheric bending is a factor, but here the effect tends to be random. Long-distance communication by way of the ionosphere produces random polarization effects, so polarization matching is of little or no importance. This is fortunate for the HF mobile enthusiast, who will find that even his short, inductively loaded whips work very well at all distances other than local.

Because it responds to all plane polarizations equally, circular polarization may pay off on circuits where the arriving polarization is random, but it exacts a 3-dB penalty when used with a single-plane polarization of any kind. Circular systems find greatest use in work with orbiting satellites. It should be remembered that “horizontal” and “vertical” are meaningless terms in space, where the plane-Earth reference is lost.

Polarization Factors Above 50 MHz

In most VHF communication over short distances, the polarization of the space wave tends to remain constant. Polarization discrimination is high, usually in excess of 20 dB, so the same polarization should be used at both ends of the circuit. Horizontal, vertical and circular polarization all have certain advantages above 50 MHz, so there has never been complete standardization on any one of them.

Horizontal systems are popular, in part because they tend to reject man-made noise, much of which is vertically polarized. There is some evidence that vertical polarization shifts to horizontal in hilly terrain, more readily than horizontal shifts to vertical. With large arrays, horizontal systems may be easier to erect, and they tend to give higher signal strengths over irregular terrain, if any difference is observed.

Practically all work with VHF mobiles is now handled with vertical systems. For use in a VHF repeater system, the vertical antenna can be designed to have gain without losing the desired omnidirectional quality. In the mobile station a small vertical whip has obvious aesthetic advantages. Often a telescoping whip used for broadcast reception can be pressed into service for the 144-MHz FM rig. A car-top mount is preferable, but the broadcast whip is a practical compromise. Tests with at least one experimental repeater have shown that horizontal polarization can give a slightly larger service area, but mechanical advantages of vertical systems have made them the almost unanimous choice in VHF FM communication. Except for the repeater field, horizontal is the standard VHF system almost everywhere.

In communication over the Earth-Moon-Earth (EME) route the polarization picture is blurred, as might be expected with such a diverse medium. If the moon were a flat target, we could expect a 180° phase shift from the moon reflection process. But it is not flat. This plus the moon’s *libration* (its slow oscillation, as viewed from the Earth), and the fact that waves must travel both ways through the Earth’s entire atmosphere and magnetic field, provide other variables that confuse the phase and polarization issue. Building a huge array that will track the moon, and give gains in excess of 20 dB, is enough of a task that most EME enthusiasts tend to take their chances with phase and polarization problems. Where rotation of the element plane has been tried it has helped to stabilize signal levels, but it is not widely employed.

PROPAGATION OF VHF WAVES

The wave energy of VHF stations does not simply disappear once it reaches the radio horizon. It is scattered, but it can be heard to some degree for hundreds of miles, well beyond line-of-sight range. Everything on Earth, and in the regions of space up to at least 100 miles, is a potential forward-scattering agent.

Tropospheric scatter is always with us. Its effects are often hidden, masked by more effective propagation modes on the lower frequencies. But beginning in the VHF range, scatter from the lower atmosphere extends the reliable range markedly if we make use of it. Called *troposcatter*, this is what produces that nearly flat portion of the curves that will be described later (in the section where you can compute reliable VHF coverage range). With a decent station, you can consistently make troposcatter contacts out to 300 miles out on the VHF and even UHF bands, especially if you don’t mind weak signals and something less than 99% reliability. As long ago as the early 1950s, VHF enthusiasts found that VHF contests could be won with high power, big antennas and a good ear for signals deep in the noise. They still can.

Ionospheric scatter works much the same as the tropo version, except that the scattering medium is higher up, mainly the E region of the ionosphere but with some help from the D and F layers too. Ionospheric scatter is useful mainly above the MUF, so its useful frequency range depends on geography, time of day, season, and the state of the Sun. With near maximum legal power, good antennas and quiet locations, ionospheric scatter can fill in the skip zone with marginally readable signals scattered from ionized trails of meteors, small areas of random ionization, cosmic dust, satellites and whatever may come into the antenna patterns at 50 to 150 miles or so above the Earth. It's mostly an E-layer business, so it works all E-layer distances. Good antennas and keen ears help.

Transequatorial propagation (TE) was an amateur 50-MHz discovery in the years 1946-1947. Amateurs of all continents observed it almost simultaneously on three separate north-south paths. These amateurs tried to communicate at 50 MHz, even though the predicted MUF was around 40 MHz for the favorable daylight hours. The first success came at night, when the MUF was thought to be even lower. A remarkable research program inaugurated by amateurs in Europe, Cyprus, Zimbabwe and South Africa eventually provided technically sound theories to explain the then-unknown mode.

It has been known for years that the MUF is higher and less seasonally variable on transequatorial circuits, but the full extent of the difference was not learned until amateur work brought it to light. As will be explained in a later section in more detail, the ionosphere over equatorial regions is higher, thicker and more dense than elsewhere. Because of its more constant exposure to solar radiation, the equatorial belt has high nighttime-MUF possibilities. TE can often work marginally at 144 MHz, and even at 432 MHz on occasion. The potential MUF varies with solar activity, but not to the extent that conventional F-layer propagation does. It is a late-in-the-day mode, taking over about when normal F-layer propagation goes out.

The TE range is usually within about 4000 km (2500 miles) either side of the geomagnetic equator. The Earth's magnetic axis is tilted with respect to the geographical axis, so the TE belt appears as a curving band on conventional flat maps of the world. See **Fig 8**. As a result, TE has a different latitude coverage in the Americas from that from Europe to Africa. The TE belt just reaches into the southern continental US. Stations in Puerto Rico, Mexico and even the northern parts of South America encounter the mode more often than those in favorable US areas. It is no accident that TE was discovered as a result of 50-MHz work in Mexico City and Buenos Aires.

Within its optimum regions of the world, the TE mode extends the usefulness of the 50-MHz band far beyond that of conventional F-layer propagation, since the practical TE MUF runs around 1.5 times that of normal F_2. Both its seasonal and diurnal characteristics are extensions of what is considered normal for 50-MHz propagation. In that part of the Americas south of about 20° North latitude, the existence of TE affects the whole character of band usage, especially in years of high solar activity.

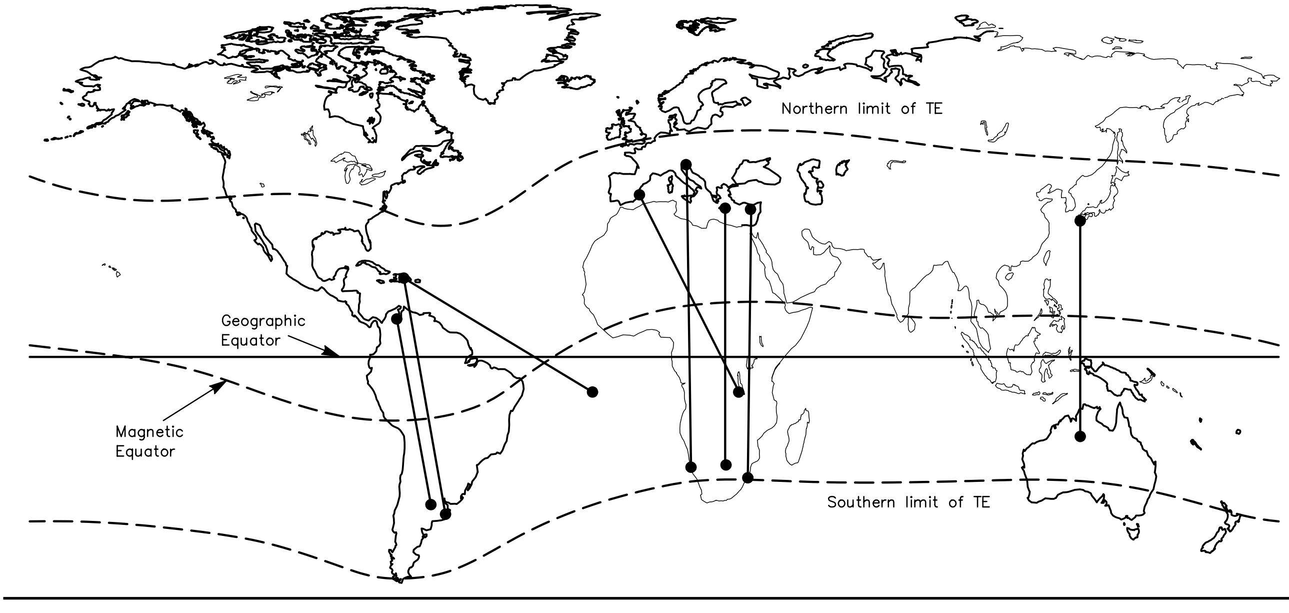

Fig 8—Transequatorial spread-F propagation takes place between stations equidistant across the geomagnetic equator. Distances up to 8000 km (5000 miles) are possible on 28 through 432 MHz. Note that the geomagnetic equator is considerably south of the geographic equator in the Western Hemisphere. (*Figure courtesy of The ARRL Handbook.*)

Weather Effects on VHF/UHF Tropospheric Propagation

Changes in the dielectric constant of the medium can affect propagation. Varied weather patterns over most of the Earth's surface can give rise to boundaries between air masses of very different temperature and humidity characteristics. These boundaries can be anything from local anomalies to air-circulation patterns of continental proportions.

Under stable weather conditions, large air masses can retain their characteristics for hours or even days at a time. See **Fig 9**. Stratified warm dry air over cool moist air, flowing slowly across the Great Lakes region to the Atlantic Seaboard, can provide the medium for east-west communication on 144 MHz and higher amateur frequencies over as much as 1200 miles. More common, however, are communication distances of 400 to 600 miles under such conditions.

A similar inversion along the Atlantic Seaboard as a result of a tropical storm air-circulation pattern may bring VHF and UHF openings extending from the Maritime Provinces of Canada to the Carolinas. Propagation across the Gulf of Mexico, sometimes with very high signal levels, enlivens the VHF scene in coastal areas from Florida to Texas. The California coast, from below the San Francisco Bay Area to Mexico, is blessed with a similar propagation aid during the warmer months. Tropical storms moving west, across the Pacific below the Hawaiian Islands, may provide a transpacific long-distance VHF medium. Amateurs first exploited this on 144, 220 and 432 MHz, in 1957. It has been used fairly often in the summer months since, although not yearly.

The examples of long-haul work cited above may occur infrequently, but lesser extensions of the minimum operating range are available almost daily. Under minimum conditions there may be little more than increased signal strength over paths that are workable at any time.

There is a diurnal effect in temperate climates. At sunrise the air aloft is warmed more rapidly than that near the Earth's surface, and as the Sun goes lower late in the day the upper air is kept warm, while the ground cools. In fair, calm weather such sunrise and sunset temperature inversions can improve signal strength over paths beyond line of sight as much as 20 dB over levels prevailing during the hours of high sun. The diurnal inversion may also extend the operating range for a given strength by some 20 to 50%. If you would be happy with a new VHF antenna, try it first around sunrise!

There are other short-range effects of local atmospheric and topographical conditions. Known as *subsidence*, the flow of cool air down into the bottom of a valley, leaving warm air aloft, is a familiar summer-evening pleasure. The daily inshore-offshore wind shift along a seacoast in summer sets up daily inversions that make coastal areas highly favored as VHF sites. Ask any jealous 144-MHz operator who lives more than a few miles inland.

Tropospheric effects can show up at any time, in any season. Late spring and early fall are the most favored periods, although a winter warming trend can produce strong and stable inversions that work VHF magic almost equal to that of the more familiar spring and fall events.

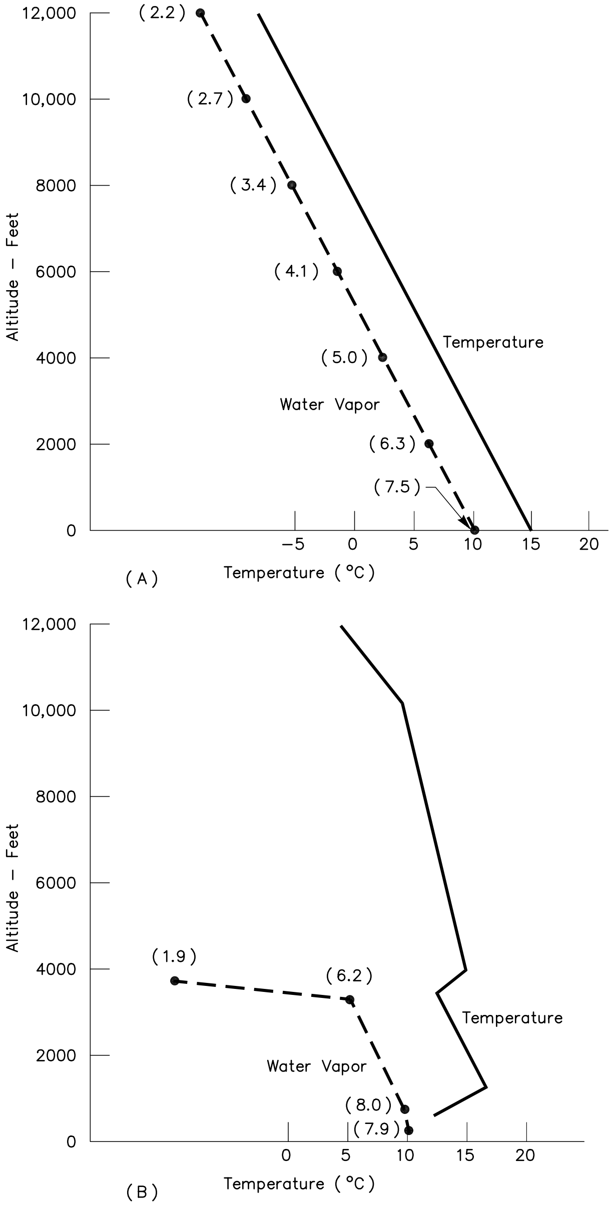

Fig 9—Upper air conditions that produce extended-range communication on the VHF bands. At the top is shown the US Standard Atmosphere temperature curve. The humidity curve (dotted) is what would result if the relative humidity were 70%, from ground level to 12,000 feet elevation. There is only slight refraction under this standard condition. At the bottom is shown a sounding that is typical of marked refraction of VHF waves. Figures in parentheses are the "mixing ratio" —grams of water vapor per kilogram of dry air. Note the sharp break in both curves at about 3500 feet.

Regions where the climate is influenced by large bodies of water enjoy the greatest degree of tropospheric bending. Hot, dry desert areas see little of it, at least in the forms described above.

Tropospheric Ducting

Tropospheric propagation of VHF and UHF waves can influence signal levels at all distances from purely local to something beyond 4000 km (2500 miles). The outer limits are not well known. At the risk of over simplification, we will divide the modes into two classes—extended local and long distance. This concept must be modified depending on the frequency under consideration, but in the VHF range the extended-local effect gives way to a form of propagation much like that of microwaves in a waveguide, called *ducting*. The transition distance is ordinarily somewhere around 200 miles. The difference lies in whether the atmospheric condition producing the bending is localized or continental in scope. Remember, we're concerned here with frequencies in the VHF range, and perhaps up to 500 MHz. At 10 GHz, for example, the scale is much smaller.

In VHF propagation beyond a few hundred miles, more than one weather front is probably involved, but the wave is propagated between the inversion layers and ground, in the main. On long paths over the ocean (two notable examples are California to Hawaii and Ascension Island to Brazil), propagation is likely to be between two atmospheric layers. On such circuits the communicating station antennas must be in the duct, or capable of propagating strongly into it. Here again, we see that the positions and radiation angles of the antennas are important. As with microwaves in a waveguide, the low-frequency limit for the duct is critical. In long-distance ducting it is also very variable. Airborne equipment has shown that duct capability exists well down into the HF region in the stable atmosphere west of Ascension Island. Some contacts between Hawaii and Southern California on 50 MHz are believed to have been by way of tropospheric ducts. Probably all contact over these paths on 144 MHz and higher bands is because of duct propagation.

Amateurs have played a major part in the discovery and eventual explanation of tropospheric propagation. In recent years they have shown that, contrary to beliefs widely held in earlier times, long-distance communication using tropospheric modes is possible to some degree on all amateur frequencies from 50 to at least 10,000 MHz.

RELIABLE VHF COVERAGE

In the preceding sections we discussed means by which amateur bands above 50 MHz may be used intermittently for communication far beyond the visual horizon. In emphasizing distance we should not neglect a prime asset of the VHF band: reliable communication over relatively short distances. The VHF region is far less subject to disruption of local communication than are frequencies below 30 MHz. Since much amateur communication is essentially local in nature, our VHF assignments can carry a great load, and such use of the VHF bands helps solve interference problems on lower frequencies.

Because of age-old ideas, misconceptions about the coverage obtainable in our VHF bands persist. This reflects the thoughts that VHF waves travel only in straight lines, except when the DX modes described above happen to be present. However, let us survey the picture in the light of modern wave-propagation knowledge and see what the bands above 50 MHz are good for on a day-to-day basis, ignoring the anomalies that may result in extensions of normal coverage.

It is possible to predict with fair accuracy how far you should be able to work consistently on any VHF or UHF band, provided a few simple facts are known. The factors affecting operating range can be reduced to graph form, as described in this section. The information was originally published in November 1961 *QST* by D. W. Bray, K2LMG (see the Bibliography at the end of this chapter).

To estimate your station's capabilities, two basic numbers must be determined: station gain and path loss. Station gain is made up of seven factors: receiver sensitivity, transmitted power, receiving antenna gain, receiving antenna height gain, transmitting antenna gain, transmitting antenna height gain and required signal-to-noise ratio. This looks complicated but it really boils down to an easily made evaluation of receiver, transmitter, and antenna performance. The other number, path loss, is readily determined from the nomogram, **Fig 10**. This gives path loss over smooth Earth, for 99% reliability.

For 50 MHz, lay a straightedge from the distance between stations (left side) to the appropriate distance at the right side. For 1296 MHz, use the full scale, right center. For 144, 222 and 432, use the dot in the circle, square or triangle, respectively. Example: At 300 miles the path loss for 144 MHz is 214 dB.

To be meaningful, the losses determined from this nomograph are necessarily greater than simple free-space path losses. As described in an earlier section, communication beyond line-of-sight distances involves propagation modes that increase the path attenuation with distance.

VHF/UHF Station Gain

The largest of the eight factors involved in station design is receiver sensitivity. This is obtainable from **Fig 11**, if you know the approximate receiver noise figure and transmission-line loss. If you can't measure noise figure, assume 3 dB for 50 MHz, 5 for 144 or 222, 8 for 432 and 10 for 1296 MHz, if you know your equipment is working moderately well. These noise figures are well on the conservative side for modern solid-state receivers.

Line loss can be taken from information in Chapter 24 for the line in use, if the antenna system is fed properly. Lay a straightedge between the appropriate points at either side of Fig 11, to find effective receiver sensitivity in decibels below 1 watt (dBW). Use the narrowest bandwidth that is practical for the emission intended, with the receiver you will be using. For CW, an average value for effective work

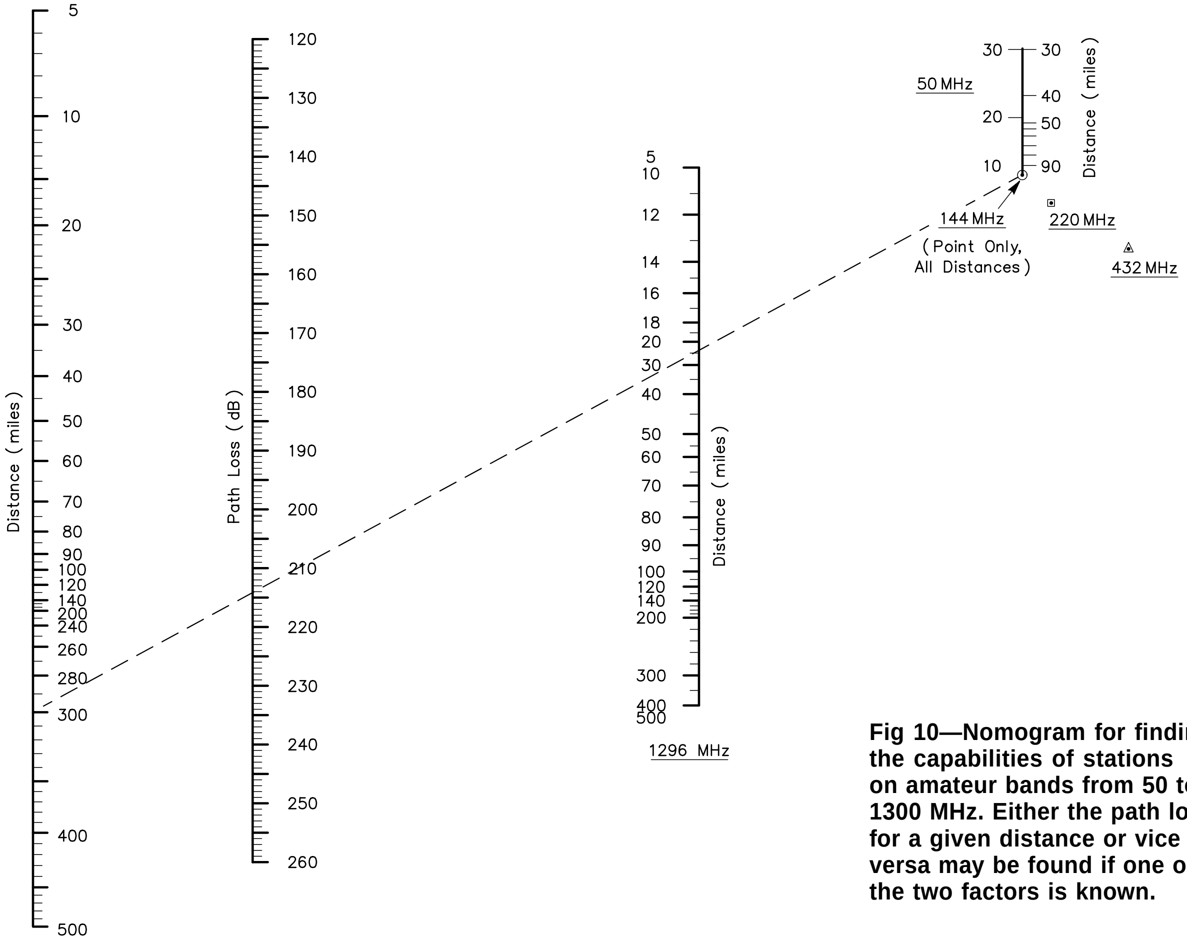

Fig 10—Nomogram for finding the capabilities of stations on amateur bands from 50 to 1300 MHz. Either the path loss for a given distance or vice versa may be found if one of the two factors is known.

is about 500 Hz. Phone bandwidth can be taken from the receiver instruction manual, but it usually falls between 2.1 to 2.7 kHz.

Antenna gain is next in importance. Gains of amateur antennas are often exaggerated. For well-designed Yagis the gain (over isotropic) run close to 10 times the boom length in wavelengths. (Example: A 24-foot Yagi on 144 MHz is 3.6 wavelengths long; 3.6 x 10 = 36, and 10 $\log_{10}$ 36 = 15.5 dBi in free space.) Add 3 dB for stacking, where used properly. Add 4 dB more for ground reflection gain. This varies in amateur work, but averages out near this figure.

We have one more plus factor—antenna height gain, obtained from **Fig 12**. Note that this is greatest for short distances. The left edge of the horizontal center scale is for 0 to 10 miles, the right edge for 100 to 500 miles. Height gain for 10 to 30 feet is assumed to be zero. For 50 feet the height gain is 4 dB at 10 miles, 3 dB at 50 miles, and 2 dB at 100 miles. At 80 feet the height gains are roughly 8, 6 and 4 dB for these distances. Beyond 100 miles the height gain is nearly uniform for a given height, regardless of distance.

Transmitter power output must be stated in decibels above 1 watt. If you have 500 W output, add 10 log (500/1), or 27 dB, to your station gain. The transmission-line loss must be subtracted from the station gain. So must the required signal-to-noise ratio. The information is based on CW work, so the additional signal needed for other modes must be subtracted. Use a figure of 3 dB for SSB. Fading losses must be accounted for also. It has been shown that for distances beyond 100 miles, the signal will vary plus or minus about 7 dB from the average level, so 7 dB must be subtracted from the station gain for high reliability. For distances under 100 miles, fading diminishes almost linearly with distance. For 50 miles, use –3.5 dB for fading.

What It All Means

Add all the plus and minus factors to get the station gain. Use the final value to find the distance over which you can expect to work reliably from the nomogram, Fig 10. Or work it the other way around: Find the path loss for the distance you want to cover from the nomogram and then figure out what station changes will be needed to overcome it.

The significance of all this becomes more obvious when we see path loss plotted against frequency for the various bands, as in **Fig 13**. At the left this is done for 50% reliability. At the right is the same information for 99% reliability.

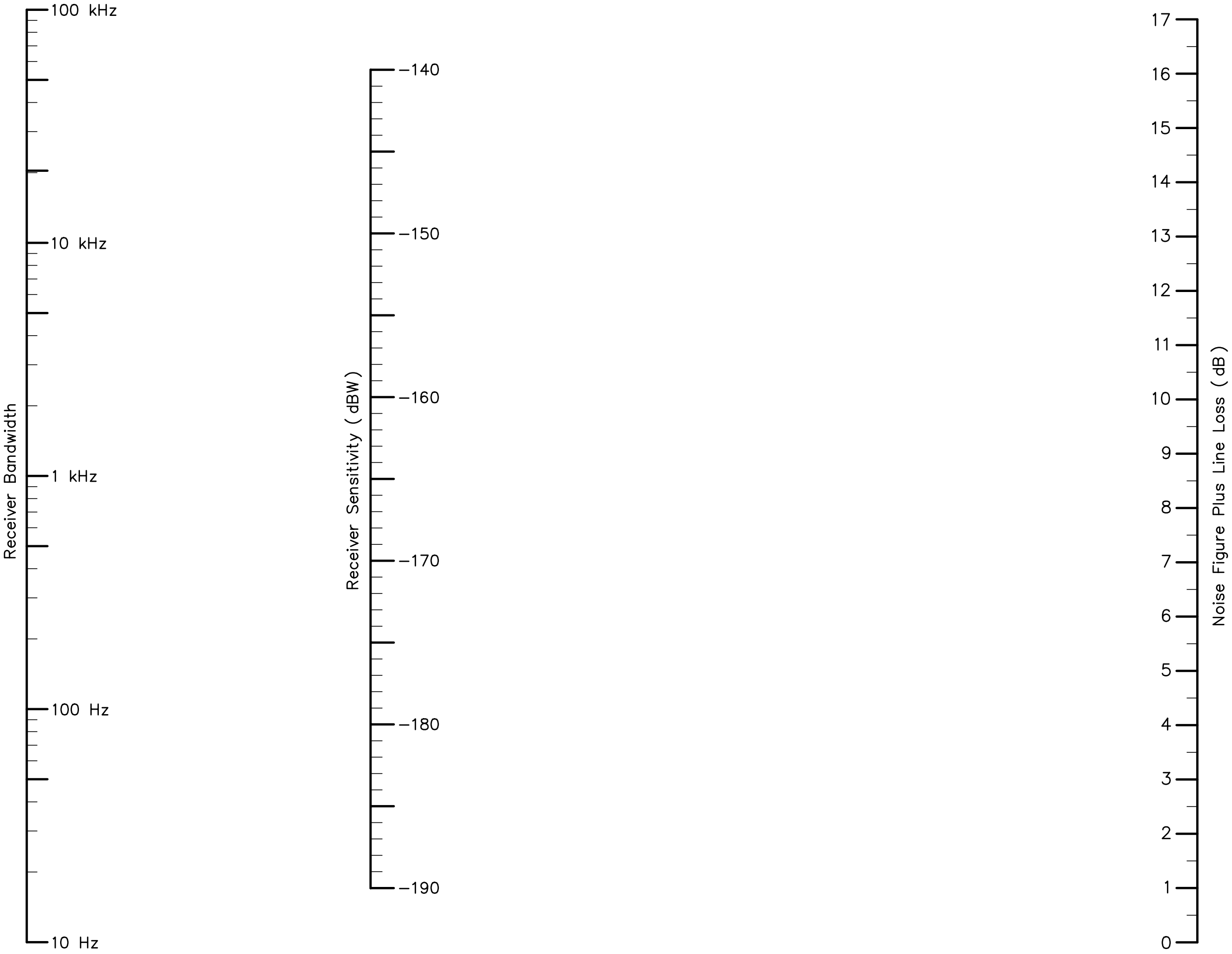

Fig 11—Nomogram for finding effective receiver sensitivity.

For near-perfect reliability, a path loss of 195 dB (easily encountered at 50 or 144 MHz) is involved in 100-mile communication. But look at the 50% reliability curve: The same path loss takes us out to well over 250 miles. Few amateurs demand near-perfect reliability. By choosing our times, and by accepting the necessity for some repeats or occasional loss of signal, we can maintain communication out to distances far beyond those usually covered by VHF stations.

Working out a few typical amateur VHF station setups with these curves will show why an understanding of these factors is important to any user of the VHF spectrum. Note that path loss rises very steeply in the first 100 miles or so. This is no news to VHF operators; locals are very strong, but stations 50 or 75 miles away are much weaker. What happens beyond 100 miles is not so well known to many of us.

From the curves of Fig 13, we see that path loss levels off markedly at what is the approximate limit of working range for average VHF stations using wideband modulation modes. Work out the station gain for a 50-W station with an average receiver and antenna, and you'll find that it comes out around 180 dB. This means you'd have about a 100-mile working radius in average terrain, for good but not perfect reliability. Another 10 dB may extend the range to as much as 250 miles. Just changing from AM phone to SSB and CW makes a major improvement in daily coverage on the VHF bands.

A bigger antenna, a higher one if your present beam is not at least 50 feet up, an increase in power to 500 W from 50 W, an improvement in receiver noise figure if it is presently poor—any of these things can make a big improvement in reliable coverage. Achieve all of them, and you will have very likely tripled your sphere of influence, thanks to that hump in the path-loss curves. This goes a long way toward explaining why using a 10-W packaged station with a small antenna, fun though it may be, does not begin to show what the VHF bands are really good for.

Terrain at VHF/UHF

The coverage figures derived from the above procedure are for average terrain. What of stations in mountainous country? Although an open horizon is generally desirable for the VHF station site, mountain country should not be

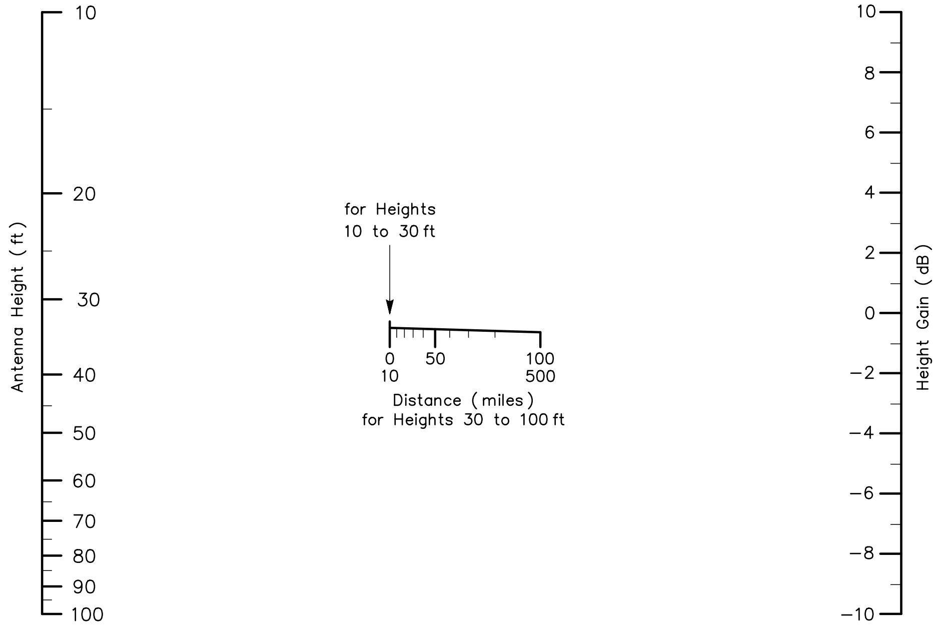

Fig 12—Nomogram for determining antenna-height gain.

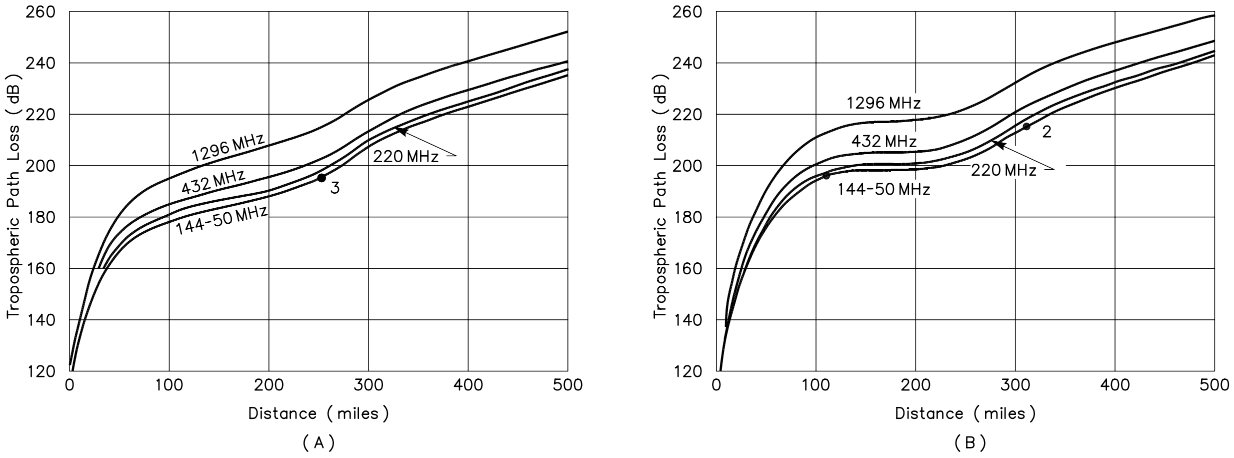

Fig 13—Path loss versus distance for amateur frequencies above 50 MHz. At A are curves for 50% of the time; at B, for 99%. The curves at A are more representative of Amateur Radio requirements.

considered hopeless. Help for the valley dweller often lies in the optical phenomenon known as *knife-edge diffraction*. A flashlight beam pointed at the edge of a partition does not cut off sharply at the partition edge, but is diffracted around it, partially illuminating the shadow area. A similar effect is observed with VHF waves passing over ridges; there is a shadow effect, but not a complete blackout. If the signal is strong where it strikes the mountain range, it will be heard well in the bottom of a valley on the far side. (See Chapter 3, The Effects of Ground, for a more thorough discussion of the theory of diffraction.)

This is familiar to all users of VHF communications equipment who operate in hilly terrain. Where only one ridge lies in the way, signals on the far side may be almost as good as on the near side. Under ideal conditions (a very high and sharp-edged obstruction near the midpoint of a long-enough path so that signals would be weak over average terrain), knife-edge diffraction may yield signals even stronger than would be possible with an open path.

The obstruction must project into the radiation patterns of the antennas used. Often mountains that look formidable to the viewer are not high enough to have an appreciable effect, one way or the other. Since the normal radiation pattern from a VHF array is several degrees above the horizontal, mountains that are less than about three degrees above the horizon, as seen from the antenna, are missed by the radiation from the array. Moving the mountains out of the way would have substantially no effect on VHF signal strength in such cases.

Rolling terrain, where obstructions are not sharp enough to produce knife-edge diffraction, still does not exhibit a complete shadow effect. There is no complete barrier to VHF propagation—only attenuation, which varies widely as the result of many factors. Thus, even valley locations are usable for VHF communication. Good antenna systems, preferably as high as possible, the best available equipment, and above all, the willingness and ability to work with weak signals may make outstanding VHF work possible, even in sites that show little promise by casual inspection.

AURORAL PROPAGATION

The Earth has a *magnetosphere* or magnetic field surrounding it. NASA scientists have described the magnetosphere as a sort of protective "bubble" around the Earth that shields us from the solar wind. Under normal circumstances, there are lots of electrons and protons moving in our magnetosphere, traveling along magnetic lines of force that trap them and keep them in place, neither bombarding the earth nor escaping into outer space.

Sudden bursts of activity on the Sun are sometimes accompanied by the ejection of charged particles, often from so-called Coronal Mass Ejections (CME) because they originate from the Sun's outer coronal region. These charged particles can interact with the magnetosphere, compressing and distorting it. If the orientation of the magnetic field contained in a large blast of solar wind or in a CME is aligned opposite to that of the Earth's magnetic field, the magnetic bubble can partially collapse and the particles normally trapped there can be deposited into the Earth's atmosphere along magnetic lines near the North or South poles. This produces a visible or radio *aurora*. An aurora is visible if the time of entry is after dark.

The visible aurora is, in effect, fluorescence at E-layer height—a curtain of ions capable of refracting radio waves in the frequency range above about 20 MHz. D-region absorption increases on lower frequencies during auroras. The exact frequency ranges depend on many factors: time, season, position with relation to the Earth's auroral regions, and the level of solar activity at the time, to name a few.

The auroral effect on VHF waves is another amateur discovery, this one dating back to the 1930s. The discovery came coincidentally with improved transmitting and receiving techniques then. The returning signal is diffused in frequency by the diversity of the auroral curtain as a refracting (scattering) medium. The result is a modulation of a CW signal, from just a slight burbling sound to what is best described as a "keyed roar." Before SSB took over in VHF work, voice was all but useless for auroral paths. A sideband signal suffers, too, but its narrower bandwidth helps to retain some degree of understandability. Distortion induced by a given set of auroral conditions increases with the frequency in use. 50-MHz signals are much more intelligible than those on 144 MHz on the same path at the same time. On 144 MHz, CW is almost mandatory for effective auroral communication.

The number of auroras that can be expected per year varies with the geomagnetic latitude. Drawn with respect to the Earth's magnetic poles instead of the geographical ones, these latitude lines in the US tilt upward to the northwest. For example, Portland, Oregon, is 2° farther north (geographic latitude) than Portland, Maine. The Maine city's geomagnetic latitude line crosses the Canadian border before it gets as far west as its Oregon namesake. In terms of auroras intense enough to produce VHF propagation results, Portland, Maine, is likely to see about 10 times as many per year. Oregon's auroral prospects are more like those of southern New Jersey or central Pennsylvania.

The antenna requirements for auroral work are mixed. High gain helps, but the area of the aurora yielding the best returns sometimes varies rapidly, so sharp directivity can be a disadvantage. So could a very low radiation angle, or a beam pattern very sharp in the vertical plane. Experience indicates that few amateur antennas are sharp enough in either plane to present a real handicap. The beam heading for maximum signal can change, however, so a bit of scanning in azimuth may turn up some interesting results. A very large array, such as is commonly used for moonbounce (with azimuth-elevation control), should be worthwhile.

The incidence of auroras, their average intensity, and their geographical distribution as to visual sightings and VHF propagation effects all vary to some extent with solar activity. There is some indication that the peak period for auroras lags the sunspot-cycle peak by a year or two. Like sporadic E, an unusual auroral opening can come at any season. There is a marked diurnal swing in the number of auroras. Favored times are late afternoon and early evening, late evening through early morning, and early afternoon, in about that order. Major auroras often start in early afternoon and carry through to early morning the next day.

HF Sky-Wave Propagation

As described earlier, the term *ground wave* is commonly applied to propagation that is confined to the Earth's lower atmosphere. Now we will use the term *sky wave* to describe modes of propagation that use the Earth's ionosphere. First, however, we must examine how the Earth's ionosphere is affected by the Sun.

THE ROLE OF THE SUN

Everything that happens in radio propagation, as with all life on Earth, is the result of radiation from the Sun. The variable nature of radio propagation here on Earth reflects the ever-changing intensity of ultraviolet and X-ray radiation, the primary ionizing agents in solar energy. Every day, solar nuclear reactions are turning hydrogen into helium, releasing an unimaginable blast of energy into space in the process. The total power radiated by the Sun is estimated at 4×10^{23} kW—that is, the number four followed by 23 zeroes. At its surface, the Sun creates about 60 *megawatts* per square meter. That is a very potent transmitter!

The Solar Wind

The Sun is constantly ejecting material from its surface in all directions into space, making up the so-called *solar wind*. Under relatively quiet solar conditions the solar wind blows around 200 miles per second—675,000 miles per hour—taking away about two million tons of solar material each second from the Sun. You needn't worry—the Sun is not going to shrivel up anytime soon. It's big enough that it will take many billions of years before that happens.

A 675,000 mile/hour wind sounds like a pretty stiff breeze, doesn't it? Lucky for us, the density of the material in the solar wind is very small by the time it has been spread out into interplanetary space. Scientists calculate that the density of the particles in the solar wind is less than that of the best vacuum they've ever achieved on Earth. Despite the low density of the material in the solar wind, the effect on the Earth, especially its magnetic field, is very significant.

Before the advent of sophisticated satellite sensors, the Earth's magnetic field was considered to be fairly simple, modeled as if the Earth were a large bar magnet. The axis of this hypothetical bar magnet is oriented about 11° away from the geographic north-south pole. We now know that the solar wind alters the shape of the Earth's magnetic field significantly, compressing it on the side facing the Sun and elongating it on the other side—in the same manner as the tail of a comet is stretched out radially in its orientation from the Sun. In fact, the solar wind is also responsible for the shape of a comet's tail.

Partly because of the very nature of the nuclear reactions going on at the Sun itself, but also because of variations in the speed and direction of the solar wind, the interactions between the Sun and our Earth are incredibly complex. Even scientists who have studied the subject for years do not completely understand everything that happens on the Sun. Later in this chapter, we'll investigate the effects of the solar wind when conditions on the Sun are *not* "quiet." As far as amateur HF skywave propagation is concerned, the results of disturbed conditions on the Sun are not generally beneficial.

Sunspots

The most readily observed characteristic of the Sun, other than its blinding brilliance, is its tendency to have grayish black blemishes, seemingly at random times and at random places, on its fiery surface. (See **Fig 14**.) There are written records of naked-eye sightings of *sunspots* in the Orient back to more than 2000 years ago. As far as is known, the first indication that sunspots were recognized as part of the Sun was the result of observations by Galileo in the early 1600s, not long after he developed one of the first practical telescopes.

Galileo also developed the projection method for observing the Sun safely, but probably not before he had suffered severe eye damage by trying to look at the Sun di-

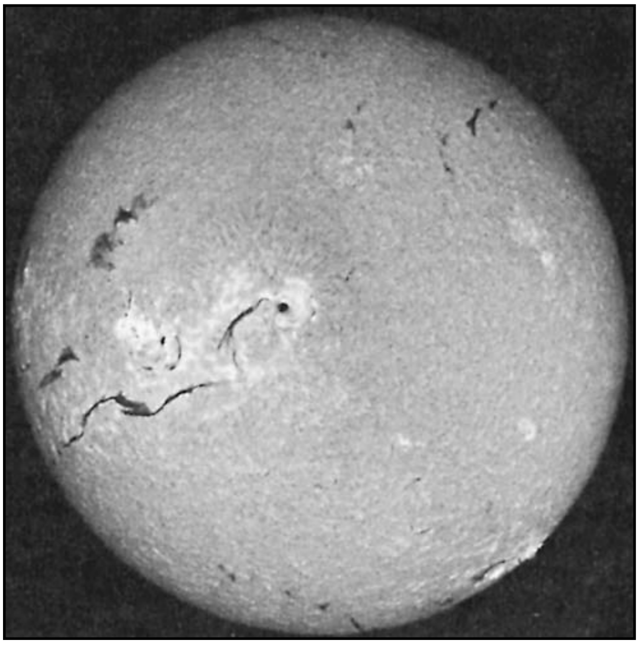

Fig 14—Much more than sunspots can be seen when the sun is viewed through selective optical filters. This photo was taken through a hydrogen-alpha filter that passes a narrow light segment at 6562 angstroms. The bright patches are active areas around and often between sunspots. Dark irregular lines are filaments of activity having no central core. Faint magnetic field lines are visible around a large sunspot group near the disc center. (*Photo courtesy of Sacramento Peak Observatory, Sunspot, New Mexico*).

rectly. (He was blind in his last years.) His drawings of sunspots, indicating their variable nature and position, are the earliest such record known to have been made. His reward for this brilliant work was immediate condemnation by church authorities of the time, which probably set back progress in learning more about the Sun for generations.

The systematic study of solar activity began about 1750, so a fairly reliable record of sunspot numbers goes back that far. (There are some gaps in the early data.) The record shows clearly that the Sun is always in a state of change. It never looks exactly the same from one day to the next. The most obvious daily change is the movement of visible activity centers (sunspots or groups thereof) across the solar disc, from east to west, at a constant rate. This movement was soon found to be the result of the rotation of the Sun, at a rate of approximately four weeks for a complete round. The average is about 27.5 days, the Sun's *synodic* rotation speed, viewed from the perspective of the Earth, which is also moving around the Sun in the same direction as the Sun's rotation.

Sunspot Numbers

Since the earliest days of systematic observation, our traditional measure of solar activity has been based on a count of sunspots. In these hundreds of years we have learned that the average number of spots goes up and down in cycles very roughly approximating a sine wave. In 1848, a method was introduced for the daily measurement of sunspot numbers. That method, which is still used today, was devised by the Swiss astronomer Johann Rudolph Wolf. The observer counts the total number of spots visible on the face of the Sun and the number of groups into which they are clustered, because neither quantity alone provides a satisfactory measure of sunspot activity. The observer's sunspot number for that day is computed by multiplying the number of groups he sees by 10, and then adding to this value the number of individual spots. Where possible, sunspot data collected prior to 1848 have been converted to this system.

As can readily be understood, results from one observer to another can vary greatly, since measurement depends on the capability of the equipment in use and on the stability of the Earth's atmosphere at the time of observation, as well as on the experience of the observer. A number of observatories around the world cooperate in measuring solar activity. A weighted average of the data is used to determine the *International Sunspot Number* or ISN for each day. (Amateur astronomers can approximate the determination of ISN values by multiplying their values by a correction factor determined empirically.)

A major step forward was made with the development of various methods for observing narrow portions of the Sun's spectrum. Narrowband light filters that can be used with any good telescope perform a visual function very similar to the aural function of a sharp filter added to a communications receiver. This enables the observer to see the actual area of the Sun doing the radiating of the ionizing energy, in addition to the sunspots, which are more a by-product than a cause. The photo of Fig 14 was made through such a filter. Studies of the ionosphere with instrumented probes, and later with satellites, manned and unmanned, have added greatly to our knowledge of the effects of the Sun on radio communication.

Daily sunspot counts are recorded, and monthly and yearly averages determined. The averages are used to see trends and observe patterns. Sunspot records were formerly kept in Zurich, Switzerland, and the values were known as *Zurich Sunspot Numbers*. They were also known as Wolf sunspot numbers. The official international sunspot numbers are now compiled at the Sunspot Index Data Center in Bruxelles, Belgium.

The yearly means (averages) of sunspot numbers from 1700 through 2002 are plotted in **Fig 15**. The cyclic nature of solar activity becomes readily apparent from this graph. The duration of the cycles varies from 9.0 to 12.7 years, but averages approximately 11.1 years, usually referred to as the 11-year solar cycle. The first complete cycle to be observed systematically began in 1755, and is numbered Cycle 1. Solar cycle numbers thereafter are consecutive. Cycle 23 began in October, 1996.

The "Quiet" Sun

For more than 60 years it has been well known that radio propagation phenomena vary with the number and size of sunspots, and also with the position of sunspots on the surface of the Sun. There are daily and seasonal variations in the Earth's ionized layers resulting from changes in the amount of ultraviolet light received from the Sun. The 11-year sunspot cycle affects propagation conditions because there is a direct correlation between sunspot activity and ionization.

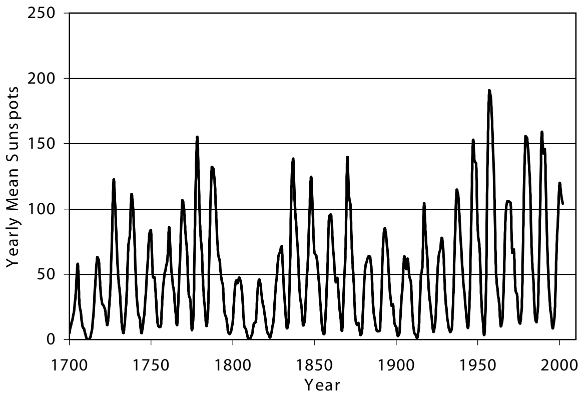

Fig 15—Yearly means of smoothed sunspot numbers from data for 1700 through 2002. This plot clearly shows that sunspot activity takes place in cycles of approximately 11 years duration. There is also a longer-term periodicity in this plot, the Gleissberg 88-year cycle. Cycle 1, the first complete cycle to be examined by systematic observation, began in 1755.

Activity on the surface of the Sun is changing continually. In this section we want to describe the activity of the so-called *quiet Sun*, meaning those times when the Sun is not doing anything more spectacular than acting like a "normal" thermonuclear ball of flaming gases. The Sun and its effects on Earthly propagation can be described in *statistical* terms—that's what the 11-year solar cycle does. You may experience vastly different conditions on any particular day compared to what a long-term average would suggest.

An analogy may be in order here. Have you ever gazed into a relatively calm campfire and been surprised when suddenly a flaming ember or a large spark was ejected in your direction? The Sun can also do unexpected and sometimes very dramatic things. Disturbances of propagation conditions here on Earth are caused by disturbed conditions on the Sun. More on this later.

Individual sunspots may vary in size and appearance, or even disappear totally, within a single day. In general, larger active areas persist through several rotations of the Sun. Some active areas have been identified over periods up to about a year. Because of these continual changes in solar activity, there are continual changes in the state of the Earth's ionosphere and resulting changes in propagation conditions. A short-term burst of solar activity may trigger unusual propagation conditions here on Earth lasting for less than an hour.

Smoothed Sunspot Numbers (SSN)

Sunspot data are averaged or smoothed to remove the effects of short-term changes. The sunspot values used most often for correlating propagation conditions are *Smoothed Sunspot Numbers* (SSN), often called 12-month running average values. Data for 13 consecutive months are required to determine a smoothed sunspot number.

Long-time users have found that the upper HF bands are reliably open for propagation only when the average number of sunspots is above certain minimum levels. For example, between mid 1988 to mid 1992 during Cycle 22, the SSN stayed higher than 100. The 10-meter band was open then almost all day, every day, to some part of the world. However, by mid 1996, few if any sunspots showed up on the Sun and the 10-meter band consequently was rarely open. Even 15 meters, normally a workhorse DX band when solar activity is high, was closed most of the time during the low point in Cycle 22. So far as propagation on the upper HF bands is concerned, the higher the sunspot number, the better the conditions.

Each smoothed number is an average of 13 monthly means, centered on the month of concern. The 1st and 13th months are given a weight of 0.5. A monthly mean is simply the sum of the daily ISN values for a calendar month, divided by the number of days in that month. We would commonly call this value a monthly average.

This may all sound very complicated, but an example should clarify the procedure. Suppose we wished to calculate the smoothed sunspot number for June 1986. We would require monthly mean values for six months prior and six months after this month, or from December 1985 through December 1986. The monthly mean ISN values for these months are

Dec	85	17.3	Jul	86	18.1
Jan	86	2.5	Aug	86	7.4
Feb	86	23.2	Sep	86	3.8
Mar	86	15.1	Oct	86	35.4
Apr	86	18.5	Nov	86	15.2
May	86	13.7	Dec	86	6.8
Jun	86	1.1			

First we find the sum of the values, but using only one-half the amounts indicated for the first and 13th months in the listing. This value is 166.05. Then we determine the smoothed value by dividing the sum by 12: 166.05/12 = 13.8. (Values beyond the first decimal place are not warranted.) Thus, 13.8 is the smoothed sunspot number for June 1986. From this example, you can see that the smoothed sunspot number for a particular month cannot be determined until six months afterwards.

Generally the plots we see of sunspot numbers are averaged data. As already mentioned, smoothed numbers make it easier to observe trends and see patterns, but sometimes this data can be misleading. The plots tend to imply that solar activity varies smoothly, indicating, for example, that at the onset of a new cycle the activity just gradually increases. But this is definitely not so! On any one day, significant changes in solar activity can take place within hours, causing sudden band openings at frequencies well above the MUF values predicted from smoothed sunspot number curves. The durations of such openings may be brief, or they may recur for several days running, depending on the nature of the solar activity.

Solar Flux

Since the late 1940s an additional method of determining solar activity has been put to use—the measurement of *solar radio flux*. The quiet Sun emits radio energy across a broad frequency spectrum, with a slowly varying intensity. Solar flux is a measure of energy received per unit time, per unit area, per unit frequency interval. These radio fluxes, which originate from atmospheric layers high in the Sun's chromosphere and low in its corona, change gradually from day to day, in response to the activity causing sunspots. Thus, there is a degree of correlation between solar flux values and sunspot numbers.

One solar flux unit equals 10^{-22} joules per second per square meter per hertz. Solar flux values are measured daily at 2800 MHz (10.7 cm) at The Dominion Radio Astrophysical Observatory, Penticton, British Columbia, where daily data have been collected since 1991. (Prior to June 1991, the Algonquin Radio Observatory, Ontario, made the measurements.) Measurements are also made at other observatories around the world, at several frequencies. With some variation, the daily measured flux values increase with increasing frequency of measurement, to at least 15.4 GHz. The daily 2800 MHz Penticton value is sent to Boulder, Colorado, where it is incorporated into WWV propagation bulletins (see later section). Daily solar flux information can be of some value in determining current propagation conditions, as sunspot numbers on a given day do not relate directly to maximum usable frequency. Solar flux values are much more reliable for this purpose, when it is averaged over time, as will be discussed later in the section on computer-prediction programs.

Correlating Sunspot Numbers and Solar Flux Values

Based on historical data, an exact mathematical relationship does not exist to correlate sunspot data and solar flux values. Comparing daily values yields almost no correlation. Comparing monthly mean values (often called monthly averages) produces a degree of correlation, but the spread in data is still significant. This is indicated in **Fig 16**, a scatter diagram plot of monthly mean sunspot numbers versus the monthly means of solar flux values adjusted to one astronomical unit. (This adjustment applies a correction for differences in distance between the Sun and the Earth at different times of the year.)

A closer correlation exists when smoothed (12-month running average) sunspot numbers are compared with smoothed (12-month running average) solar flux values adjusted to one astronomical unit. A scatter diagram for smoothed data appears in **Fig 17**. Note how the plot points establish a better defined pattern in Fig 17. The correlation is still no better than a few percent, for records indicate a given smoothed sunspot number does not always correspond with the same smoothed solar flux value, and vice versa.

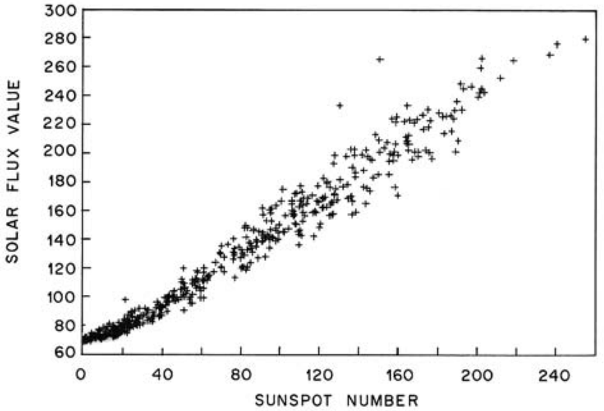

Fig 16—Scatter diagram or X-Y plot of monthly mean sunspot numbers and monthly mean 2800MHz solar flux values. Data values are from February 1947 through February 1987. Each "+" mark represents the intersection of data for a given month. If the correlation between sunspot number and flux values were consistent, all the marks would align to form a smooth curve.

Table 1 illustrates some of the inconsistencies that exist in the historical data. Smoothed or 12-month running average values are shown.

Even though there is no precise mathematical relationship between sunspot numbers and solar flux values, it is helpful to have some way to convert from one to the other. The primary reason is that sunspot numbers are valuable as a long-term link with the past, but the great usefulness of solar flux values are their immediacy, and their direct bearing on our field of interest. (Remember, a smoothed sunspot number will not be calculated until six months after the fact.)

The following mathematical approximation has been derived to convert a smoothed sunspot number to a solar flux value.

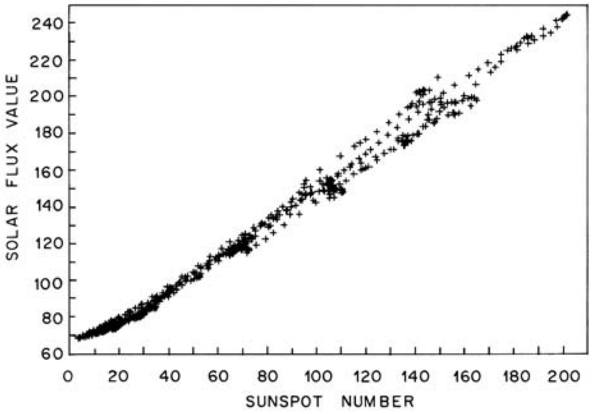

Fig 17—Scatter diagram of smoothed, or 12-month running averages, sunspot numbers versus 2800MHz solar flux values. The correlation of smoothed values is better than for monthly means, shown in Fig 16.

Table 1

Selected Historical Data Showing Inconsistent Correlation Between Sunspot Number and Solar Flux

Month	*Smoothed Sunspot Number*	*Smoothed Solar Flux Value*
May 1953	17.4	75.6
Sept 1965	17.4	78.5
Jul 1985	17.4	74.7
Jun 1969	106.1	151.4
Jul 1969	105.9	151.4
Dec 1982	94.6	151.4
Aug 1948	141.1	180.5
Oct 1959	141.1	192.3
Apr 1979	141.1	180.4
Aug 1981	141.1	203.3

$$F = 63.75 + 0.728\,S + 0.00089\,S^2 \qquad \text{(Eq 5)}$$

where

F = solar flux number

S = smoothed sunspot number

A graphic representation of this equation is given in **Fig 18**. Use this chart to make conversions graphically, rather than by calculations. With the graph, solar flux and sunspot number conversions can be made either way. The equation has been found to yield errors as great as 10% when historical data was examined. (Look at the August 1981 data in Table 1.) Therefore, conversions should be rounded to the nearest whole number, as additional decimal places are unwarranted. To make conversions from flux to sunspot number, the following approximation may be used.

$$S = 33.52\sqrt{85.12 + F} - 408.99 \qquad \text{(Eq 6)}$$

THE IONOSPHERE

There will be inevitable "gray areas" in our discussion of the Earth's atmosphere and the changes wrought in it by the Sun and by associated changes in the Earth's magnetic field. This is not a story that can be told in neat equations, or values carried out to a satisfying number of decimal places. The story must be told, and understood—with its well-known limitations—if we are to put up good antennas and make them serve us well.

Thus far in this chapter we have been concerned with what might be called our "above-ground living space"—that portion of the total atmosphere wherein we can survive without artificial breathing aids, or up to about 6 km (4 miles). The boundary area is a broad one, but life (and radio propagation) undergo basic changes beyond this zone. Somewhat farther out, but still technically within the Earth's atmosphere, the role of the Sun in the wave-propagation picture is a dominant one.

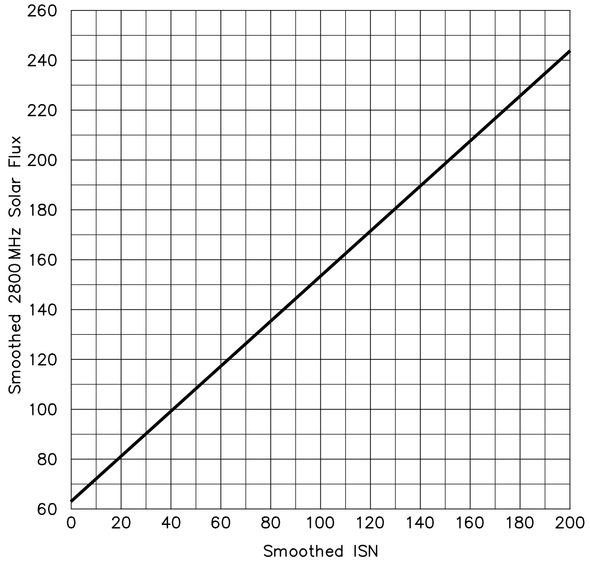

Fig 18—Chart for conversions between smoothed International Sunspot Numbers and smoothed 2800MHz solar flux. This curve is based on the mathematical approximation given in the text.

This is the *ionosphere*—a region where the air pressure is so low that free electrons and ions can move about for some time without getting close enough to recombine into neutral atoms. A radio wave entering this rarefied atmosphere, a region of relatively many free electrons, is affected in the same way as in entering a medium of different dielectric constant—its direction of travel is altered.

Ultraviolet (UV) radiation from the Sun is the primary cause of ionization in the outer regions of the atmosphere, the ones most important for HF propagation. However, there are other forms of solar radiation as well, including both hard and soft x-rays, gamma rays and extreme ultraviolet (EUV). The radiated energy breaks up, or *photoionizes*, atoms and molecules of atmospheric gases into electrons and positively charged ions. The degree of ionization does not increase uniformly with distance from the Earth's surface. Instead there are relatively dense regions (layers) of ionization, each quite thick and more or less parallel to the Earth's surface, at fairly well-defined intervals outward from about 40 to 300 km (25 to 200 miles). These distinct layers are formed due to complex photochemical reactions of the various types of solar radiation with oxygen, ozone, nitrogen and nitrous oxide in the rarefied upper atmosphere.

Ionization is not constant within each layer, but tapers off gradually on either side of the maximum at the center of the layer. The total ionizing energy from the Sun reaching a given point, at a given time, is never constant, so the height and intensity of the ionization in the various regions will also vary. Thus, the practical effect on long-distance communication is an almost continuous variation in signal level, related to the time of day, the season of the year, the distance between the Earth and the Sun, and both short-term and long-term variations in solar activity. It would seem from all this that only the very wise or the very foolish would attempt to predict radio propagation conditions, but it is now possible to do so with a fair chance of success. It is possible to plan antenna designs, particularly the choosing of antenna heights, to exploit known propagation characteristics.

Ionospheric Layer Characteristics

The lowest known ionized region, called the *D layer* (or the *D region*), lies between 60 and 92 km (37 to 57 miles) above the Earth. In this relatively low and dense part of the atmosphere, atoms broken up into ions by sunlight recombine quickly, so the ionization level is directly related to sunlight. It begins at sunrise, peaks at local noon and disappears at sundown. When electrons in this dense medium are set in motion by a passing wave, collisions between particles are so frequent that a major portion of their energy may be used up as heat, as the electrons and disassociated ions recombine.

The probability of collisions depends on the distance an electron travels under the influence of the wave—in other words, on the wavelength. Thus, our 1.8- and 3.5-MHz bands, having the longest wavelengths, suffer the highest daytime absorption loss as they travel through the D layer, particularly for waves that enter the medium at the lowest angles. At times of high solar activity (peak years of the solar cycle) even waves entering the D layer vertically suffer almost total energy absorption around midday, making these bands almost useless for communication over appreciable distances during the hours of high sun. They "go dead" quickly in the morning, but come alive again the same way in late afternoon. The diurnal (daytime) D-layer effect is less at 7 MHz (though still marked), slight at 14 MHz and inconsequential on higher amateur frequencies.

The D region is ineffective in bending HF waves back to Earth, so its role in long-distance communication by amateurs is largely a negative one. It is the principal reason why our frequencies up through the 7-MHz band are useful mainly for short-distance communication during the high-sun hours.

The lowest portion of the ionosphere useful for long-distance communication by amateurs is the *E layer* (also known as the *E region*) about 100 to 115 km (62 to 71 miles) above the Earth. In the E layer, at intermediate atmospheric density, ionization varies with the Sun angle above the horizon, but solar ultraviolet radiation is not the sole ionizing agent. Solar X-rays and meteors entering this portion of the Earth's atmosphere also play a part. Ionization increases rapidly after sunrise, reaches maximum around noon local time, and drops off quickly after sundown. The minimum is after midnight, local time. As with the D layer, the E layer absorbs wave energy in the lower-frequency amateur bands when the Sun angle is high, around mid-day. The other varied effects of E-region ionization will be discussed later.

Most of our long-distance communication capability stems from the tenuous outer reaches of the Earth's atmosphere known as the *F layer*. At heights above 100 miles, ions and electrons recombine more slowly, so the observable effects of the Sun develop more slowly. Also, the region holds its ability to reflect wave energy back to Earth well into the night. The *maximum usable frequency* (MUF) for F-layer propagation on east-west paths thus peaks just after noon at the midpoint, and the minimum occurs after midnight. We'll examine the subject of MUF in more detail later.

Judging what the F layer is doing is by no means that simple, however. The layer height may be from 160 to more than 500 km (100 to over 310 miles), depending on the season of the year, the latitudes, the time of day and, most capricious of all, what the Sun has been doing in the last few minutes and in perhaps the last three days before the attempt is made. The MUF between Eastern US and Europe, for example, has been anything from 7 to 70 MHz, depending on the conditions mentioned above, plus the point in the long-term solar-activity cycle at which the check is made.

During a summer day the F layer may split into two layers. The lower and weaker F_1 *layer*, about 160 km (100 miles) up, has only a minor role, acting more like the E than the F_2 *layer*. At night the F_1 region disappears and the F_2 region height drops somewhat.

Propagation information tailored to amateur needs is transmitted in all information bulletin periods by the ARRL Headquarters station, W1AW. Finally, solar and geomagnetic field data, transmitted hourly and updated eight times daily, are given in brief bulletins carried by the US Time Standard stations, WWV and WWVH, and also on Internet Web sites. But more on these services later.

Bending in the Ionosphere

The degree of bending of a wave path in an ionized layer depends on the density of the ionization and the length of the wave (inversely related to its frequency). The bending at any given frequency or wavelength will increase with increased ionization density and will bend away from the region of most-intense ionization. For a given ionization density, bending increases with wavelength (that is, it decreases with frequency).

Two extremes are thus possible. If the intensity of the ionization is sufficient and the frequency is low enough, even a wave entering the layer perpendicularly will be reflected back to Earth. Conversely, if the frequency is high enough or the ionization decreases to a low-enough density, a condition is reached where the wave angle is not affected enough by the ionosphere to cause a useful portion of the wave energy to return to the Earth. The frequency at which this occurs is called the vertical-incidence *critical frequency.* Each region in the ionosphere has a critical frequency associated with it, and this critical frequency will change depending on the date, time and state of the 11-year solar cycle.

Fig 19 shows a simplified graph of the electron density (in electrons per cubic meter) versus height in the ionosphere (in km) for a particular set of daytime and nighttime conditions. Free electrons are what return the signals you launch into the ionosphere back down to the Earth at some distance from your transmitter—The more free electrons in the ionosphere, the better propagation will be, particularly at higher frequencies.

Electron-density profiles are extremely complicated and vary greatly from one location to the next, depending on a bewildering variety of factors. Of course, this sheer variability makes it all the more interesting and challenging for hams to work each other on ionospheric HF paths!

The following discussion about sounding the ionosphere provides some background information about the scientific instruments used to decipher the highly intricate mechanisms behind ionospheric HF propagation.

SOUNDING THE IONOSPHERE

For many years scientists have *sounded* the ionosphere to determine its communication potential at various elevation angles and frequencies. The word "sound" stems from

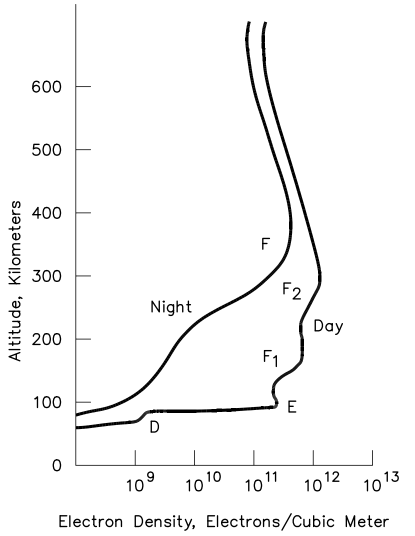

Fig 19—Typical electron densities for nighttime and daytime conditions in the various ionospheric regions.

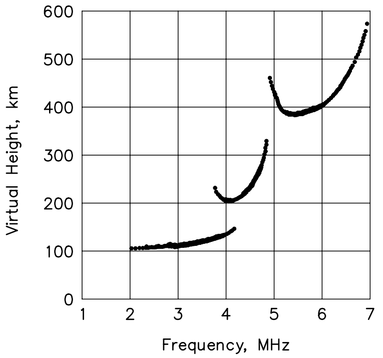

Fig 20—Very simplified ionogram from a vertical-incidence sounder. The lowest trace is for the E region; the middle for the F_1 and the upper trace for the F_2 region.

an old idea—one that has nothing to do with the audio waves that we can hear as "sounds." Long ago, sailors sounded the depths beneath their boats by dropping weighted ropes, calibrated in fathoms, into the water. In a similar fashion, the instrument used to probe the height of the ionosphere is called an *ionosonde*, or ionospheric sounder. It measures distances to various layers by launching a calibrated electronic signal directly up into the ionosphere.

Radar uses the same techniques as ionospheric sounding to detect targets such as airplanes. An ionosonde sends precisely timed pulses into the ionosphere over a range of MF and HF frequencies. The time of reception of an echo reflected from a region in the ionosphere is compared to the time of transmission. The time difference is multiplied by the speed of light to give the apparent distance that the wave has traveled from the transmitter to the ionosphere and back to the receiver. (It is an *apparent* or *virtual distance* because the speed of a wave slows very slightly in the ionosphere, just as the speed of propagation through any medium other than a vacuum slows down because of that medium.)

Another type of ionosonde sweeps the frequency of transmission, from low to high. This is called an "FM-CW," or more colorfully, a "chirp" sounder. Since a received echo takes time to travel from the transmitter up to the reflection point and then back again to the receiver, the echo will be at a lower frequency than the still-moving frequency of the transmitter. The frequency difference is an indication of the height of the echo's reflection off the various ionospheric layers.

Vertical-Incidence Sounders

Most ionosondes are *vertical-incidence sounders*, bouncing their signals perpendicularly off the various ionized regions above it by launching signals straight up into the ionosphere. The ionosonde frequency is swept upwards until echos from the various ionospheric layers disappear, meaning that the critical frequencies for those layers have been exceeded, causing the waves to disappear into space.

Fig 20 shows a highly simplified ionogram for a typical vertical-incidence sounder. The echos at the lowest height at the left-hand side of the plot show that the E region is about 100 km high. The F_1 region shown in the middle of the plot varies from about 200 to 330 km in this example, and the F_2 region ranges from just under 400 km to almost 600 km in height. You can see that the F_1 and F_2 ionospheric regions take a "U" shape, indicating that the electron density varies throughout the layer. In this example, the peak in electron density is at a virtual height of the F_2 region of about 390 km, the lowest point in the F_2 curve.

Scientists can derive a lot of information from a vertical-incidence ionogram, including the critical frequencies for each region, where raising the frequency any higher causes the signals to disappear into space. In Fig 20, the E-region critical frequency (abbreviated f_oE) is about 4.1 MHz. The F_1-region critical frequency (abbreviated f_oF_1) is 4.8 MHz. The F2-region critical frequency (abbreviated f_oF_2) is this simplified diagram is 6.8 MHz.

The observant reader may well be wondering what the subscripted "o" in the abbreviations f_oE, f_oF_1 and f_oF_2 mean. The abbreviation "o" means "ordinary." When an electromagnetic wave is launched into the ionosphere, the Earth's magnetic field splits the wave into two independent waves—the "ordinary" (o) and the "extraordinary" (x) components. The ordinary wave reaches the same height in the ionosphere whether the Earth's magnetic field is present or not, and hence is called "ordinary." The extraordinary wave, however, is greatly affected by the presence of the Earth's magnetic field, in a very complex fashion.

Fig 21 shows an example of an actual ionogram from the vertical Lowell Digisonde at Millstone Hill in Massa-

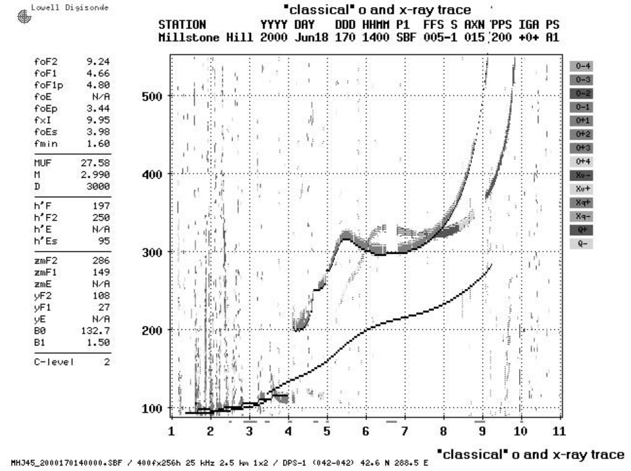

```
STATION            YYYY DAY   DDD HHMM P1  FFS S AXN PPS IGA PS
Millstone Hill 2000 Jun18 170 1400 SBF 005-1 015 200 +0+ A1

foF2      9.24
foF1      4.66
foF1p     4.80
foE        N/A
foEp      3.44
fxI       9.95
foEs      3.98
fmin      1.60

MUF      27.58
M        2.990
D         3000

h'F        197
h'F2       250
h'E        N/A
h'Es        95

zmF2       286
zmF1       149
zmE        N/A
yF2        108
yF1         27
yE         N/A
B0       132.7
B1        1.50

C-level      2

MHJ45_200017014000.SBF / 400fx256h 25 kHz 2.5 km 1x2 / DPS-1 (042-042) 42.6 N 288.5 E
```

Fig 21—Actual vertical-incidence ionogram from the Lowell Digisonde, owned and operated at Millstone Hill in Massachusetts by MIT. The ordinary (o) and extraordinary (x) traces are shown for heights greater than about 300 km. At the upper left are listed the computer-determined ionospheric parameters, such as f_oF_2 of 9.24 MHz and f_oF_1 at 4.66 MHz.

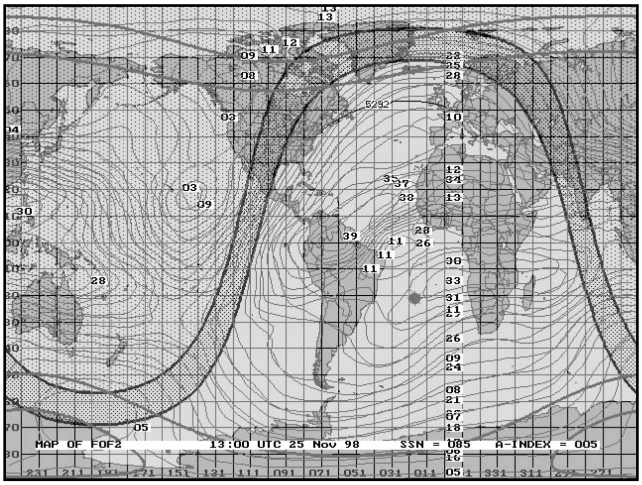

Fig 22—Computer simulation of the f_oF_2 contours for 25 November 1998, for an SSN of 85 and a quiet planetary A_p index of 5. Note the two regions of high f_oF_2 values off the upper and lower west coast of Africa. These are the "equatorial anomalies," regions of high electronic density in the F_2 region that often allow chordal-hop north-south propagation. See also Fig 8. (*PropLab Pro simulation, courtesy of Solar Terrestrial Dispatch.*)

chusetts, owned and operated by the Massachusetts Institute of Technology. This ionogram was made on June 18, 2000, and shows the conditions during a period of very high solar activity. The black-and-white rendition in Fig 21 of the actual color ionogram unfortunately loses some information. However, you can still see that a real ionogram is a lot more complicated looking than the simple simulated one in Fig 20.

The effects of noise and interference from other stations are shown by the many speckled dots appearing in the ionogram. The critical frequencies for various ionospheric layers are listed numerically at the left-hand side of the plot and the signal amplitudes are color-coded by the color bars at the right-hand side of the plot. The x-axis is the frequency, ranging from 1 to 11 MHz.

Compared to the simplified ionogram in Fig 20, Fig 21 shows another trace that appears on the plot from about 5.3 to 9.8 MHz, a trace shifted to the right of the darker ordinary trace. This second trace is the extraordinary (x) wave mentioned above. Since the x and o waves are created by the Earth's magnetic field, the difference in the ordinary and extraordinary traces is about 1/2 the *gyro frequency*, the frequency at which an electron will spiral down a particular magnetic field line. The electron gyro frequency is different at various places around the Earth, being related to the Earth's complicated and changing magnetic field. The extraordinary trace always has a higher critical frequency than the ordinary trace on a vertical-incidence ionogram, and it is considerably weaker than the ordinary trace, especially at frequencies below about 4 MHz because of heavy absorption.

The Big Picture Overhead

There are about 150 vertical-incidence ionosondes around the world. Ionosondes are located on land, even on a number of islands. There are gaps in sounder coverage, however, mainly over large expanses of open ocean. The compilation of all available vertical-incidence data from the worldwide network of ionospheric sounders results in global f_oF_2 maps, such as the map shown in **Fig 22**, a simulation from the highly sophisticated *PropLab Pro* computer program.

This simulation is for 1300 UTC, several hours after East Coast sunrise on Nov 25, 1998, with a high level of solar activity of 85 and a planetary A_p index of 5, indicating calm geomagnetic conditions. The contours of f_oF_2 peak over the ocean off the west coast of Africa at 38 MHz. Over the southern part of Africa, f_oF_2 peaks at 33 MHz.

These two "humps" in f_oF_2 form what is known as the "equatorial anomaly" and are caused by upwelling "fountains" of high electron concentration located in daylight areas about ±20° from the Earth's magnetic dip equator. The equatorial anomaly is important in transequatorial propagation. Those LU stations in Argentina that you can hear on 28 MHz from the US in the late afternoon, even during low portions of the solar cycle when other stations to the south are not coming through, are benefiting from transequatorial propagation, sometimes called "chordal hop" propagation, because signals going through this area remain in the ionosphere without lossy intermediate hops to the ground.

From records of f_oF_2 profiles, the underlying electron densities along a path can be computed. And from the electron density profiles computerized "ray tracing" may be done throughout the ionosphere to determine how a wave propagates from a transmitter to a particular receiver location. *PropLab Pro* can do complex ray tracings that explicitly include the effect of the Earth's magnetic field, even taking into effect ionospheric stormy conditions.

Oblique-Angle Ionospheric Sounding

A more elaborate form of ionospheric sounder is the *oblique ionosonde*. Unlike a vertical-incidence ionospheric sounder, which sends its signals directly overhead, an oblique sounder transmits its pulses obliquely through the ionosphere, recording echos at a receiver located some distance from the transmitter. The transmitter and distant receiver are precisely coordinated in GPS-derived time in modern oblique sounders.

Interpretation of ionograms produced by oblique sounders is considerably more difficult than for vertically incident ones. An oblique ionosonde purposely transmits over a continuous range of elevation angles simultaneously and hence cannot give explicit information about each elevation angle it launches. **Fig 23** shows a typical HF oblique-sounder ionogram for the path from Hawaii to California in March of 1973, during a period of medium-level sunspot activity. The y-axis is calibrated in time delay, in milliseconds. Longer distances involve longer time delays between the start of a transmitted pulse and the reception of the echo. The x-axis in this ionogram is the frequency, just like a vertical-incidence ionogram. Note that the frequency range for this plot extends to 32 MHz, while vertical-incidence ionograms usually don't sweep higher than about 12 MHz.

Six possible modes are shown in this ionogram: $1F_2$, $2F_2$, $3F_2$, $4F_2$ and $5F_2$. These involve multiple modes of propagation (commonly called *hops*) between the ionosphere and reflections from the Earth. For example, at an operating frequency of 14 MHz, there are three modes open during the mid-morning: $2F_2$, $3F_2$ and $4F_2$. We'll discuss multiple hops later in more detail.

The lowest mode, $1F_2$ in Fig 23, employs a single F_2 hop to cover the 3900-km long path from Hawaii to California, but it is only open on 28-MHz. (Note that 3900 km is close to the maximum possible single-hop length for the F_2 region. We'll look at this in more detail later too.) In general, each mode that involves more than a single hop is weaker than a single hop. For example, you can see that the received $5F_2$ echo is weak and broken up because of the accumulation of losses at each ground-level reflection in its five hops, with absorption in the ionosphere all along its complicated path to the receiver.

The trace labeled "FOT" is the *frequency of optimum traffic*, considered the most reliable frequency for communications on this particular circuit and date/time. In this example, the FOT would be near the 21-MHz amateur band.

Another interesting point in Fig 23 is labeled "High Angle Ray." This refers to the Pedersen ray. Before we go into more details about the Pedersen high-angle wave, we need to examine how launch angles affect the way waves are propagated through the ionosphere.

Fig 24 shows a highly simplified situation, with a single ionospheric layer and a smooth Earth. This illustrates several important facts about antenna design for long-distance communication. In Fig 24, Wave #1 is launched at the lowest elevation angle (that is, most nearly horizontal to the horizon). Wave #1 manages to travel from the transmitter to the receiving location at point C in a single hop.

Wave #2 is launched at a higher elevation angle than Wave #1, and penetrates further into the ionospheric layer before it is refracted enough to return to Earth. The ground distance covered from the transmitter to point B is less for

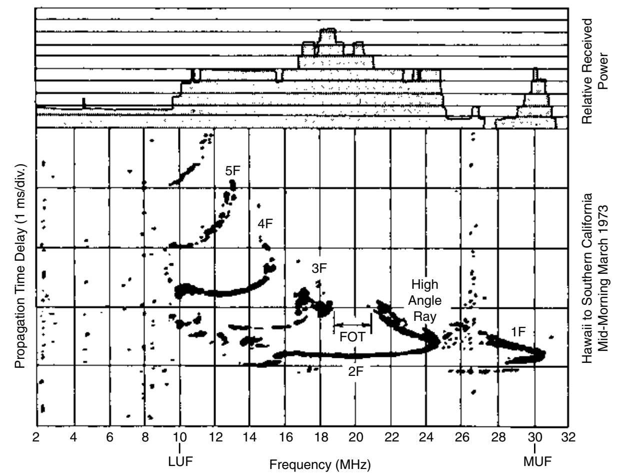

Fig 23—HF oblique-sounder ionogram. This is a typical chirpsounder measurement on a 2500-mile path from Hawaii to southern California during midmorning in March at a medium level of solar activity. Six possible modes (hops) are shown. The "FOT" is the frequency of optimum traffic, considered most reliable for this path/ time.

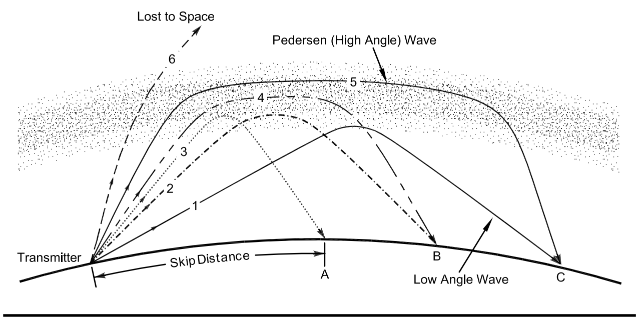

Fig 24—Very simplified smooth-Earth/ionosphere diagram showing how the ground range from transmitter to receiver can vary as the elevation angle is gradually raised. The Pedersen wave, launched at a relatively high angle, has the same ground range as the low-angle wave #1, but is weaker because it travels for a long distance in the ionosphere.

Wave #2 than for lower-angle Wave #1. Wave #3 is launched at a still-higher elevation angle. Like Wave #2 before it, Wave #3 penetrates further into the ionosphere and covers less ground downrange than #2.

Now, we see something very interesting happening for Wave #4, whose launch elevation angle is still higher than #3. Wave #4 penetrates even higher into the ionosphere than #3, reaching the highest level of ionization in our theoretical ionospheric layer, where it is finally refracted sufficiently to bend down to Earth. Wave #4 manages to arrive at the same point B as Wave #2, which was launched at a much lower elevation angle.

In other words, in the sequence from #1 to #3 we have been continually *increasing* the elevation launch angle and the ground range covered from the transmitter to the return of the signal back to Earth has been continually *decreasing*. However, starting with Wave #4, the ground range starts to *increase* with increased elevation angle. A further increase in the elevation angle causes Wave #5 to travel for an even longer distance through the ionosphere, exiting finally at point C, the same ground distance as lowest-angle Wave #1.

Finally, increasing the elevation angle even further results in Wave #6 being lost to outer space because the ionization in the layer is insufficient to bend the wave back to Earth. In other words, Wave #6 has exceeded the *critical angle* for this hypothetical ionospheric layer and this frequency of operation.

Both Waves #4 and #5 in Fig 24 are called "high-angle" or Pedersen waves. Because Wave #5 has traveled a greater distance through the ionosphere, it is always weaker than Wave #1, the one launched at the lowest elevation angle. Pedersen waves are usually not very stable, since small changes in elevation angle can result in large changes in the ground range that these high-angle waves cover.

SKIP DISTANCE

Fig 24 shows that we can communicate with the point on the Earth labeled "A" (where Wave #3 arrives), but not any closer to our transmitter site. When the critical angle is less than 90° (that is, directly overhead) there will always be a region around the transmitting site where an ionospherically propagated signal cannot be heard, or is heard weakly. This area lies between the outer limit of the ground-wave range and the inner edge of energy return from the ionosphere. It is called the *skip zone*, and the distance between the originating site and the beginning of the ionospheric return is called the skip distance. This terminology should not to be confused with ham jargon such as "the skip is in," referring to the fact that a band is open for sky-wave propagation.

The signal may often be heard to some extent within the skip zone, through various forms of scattering (discussed in detail later), but it will ordinarily be marginal in strength. When the skip distance is short, both ground-wave and sky-wave signals may be received near the transmitter. In such instances the sky wave frequently is stronger than the ground wave, even as close as a few miles from the transmitter. The ionosphere is an efficient communication medium under favorable conditions. Comparatively, the ground wave is not.

If the radio wave leaves the Earth at a radiation angle of zero degrees, just at the horizon, the maximum distance that may be reached under usual ionospheric conditions in the F_2 region is about 4000 km (2500 miles).

MULTI-HOP PROPAGATION

As mentioned previously in the discussion about Fig 24, the Earth itself can act as a reflector for radio waves, resulting a multiple hops. Thus, a radio signal can be reflected from the reception point on the Earth back into the ionosphere, reaching the Earth a second time at a still more-distant point. This effect is illustrated in **Fig 25**, where a single ionospheric layer is depicted, although this time we show both the layer and the Earth beneath it as curved rather than flat. The wave identified as "Critical Angle" travels from the transmitter via the ionosphere to point A, in the center of the drawing, where it is reflected upwards and travels through the ionosphere to point B, at the right. This shows a two-hop signal.

As in the simplified case in Fig 24, the distance at which a ray eventually reaches the Earth depends on the launch elevation angle at which it left the transmitting antenna. You have some control of the launch angle by adjusting the height of the antennas you use, as described in Chapter 3, The Effects of Ground.

The information in Fig 25 is greatly simplified. On actual communication paths the picture is complicated by many factors. One is that the transmitted energy spreads over a considerable area after it leaves the antenna. Even with an antenna array having the sharpest practical beam pattern, there is what might be described as a *cone of radiation* centered on the wave lines (rays) shown in the drawing. The reflection/refraction in the ionosphere is also highly variable, and is the cause of considerable spreading and scattering.

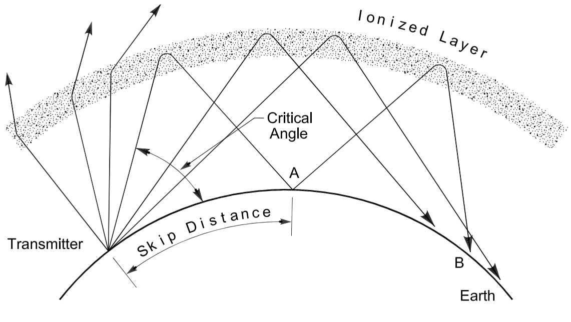

Fig 25—Behavior of waves encountering a simple curved ionospheric layer over a curved Earth. Rays entering the ionized region at angles above the critical angle are not bent enough to be returned to Earth, and are lost to space. Waves entering at angles below the critical angle reach the Earth at increasingly greater distances as the launch angle approaches the horizontal. The maximum distance that may normally be covered in a single hop is 4000 km. Greater distances are covered with multiple hops.

Under some conditions it is possible for as many as four or five signal hops to occur over a radio path, as illustrated by the oblique ionogram in Fig 23. But no more than two or three hops is the norm. In this way, HF communication can be conducted over thousands of miles.

An important point should be recognized with regard to signal hopping. A significant loss of signal occurs with each hop. The D and E layers of the ionosphere absorb energy from signals as they pass through, and the ionosphere tends to scatter the radio energy in various directions, rather than confining it in a tight bundle. The roughness of the Earth's surface also scatters the energy at a reflection point.

Assuming that both waves do reach point B in Fig 25, the low-angle wave will contain more energy at point B. This wave passes through the lower layers just twice, compared to the higher-angle route, which must pass through these layers four times, plus encountering an Earth reflection. Measurements indicate that although there can be great variation in the relative strengths of the two signals—the one-hop signal will generally be from 7 to 10 dB stronger. The nature of the terrain at the mid-path reflection point for the two-hop wave, the angle at which the wave is reflected from the Earth, and the condition of the ionosphere in the vicinity of all the refraction points are the primary factors in determining the signal-strength ratio.

The loss per hop becomes significant at greater distances. It is because of these losses that no more than four or five propagation hops are useful; the received signal becomes too weak to be usable over more hops. Although modes other than signal hopping also account for the propagation of radio waves over thousands of miles, backscatter studies of actual radio propagation have displayed signals with as many as five hops. So the hopping mode is arguably the most prevalent method for long-distance communication.

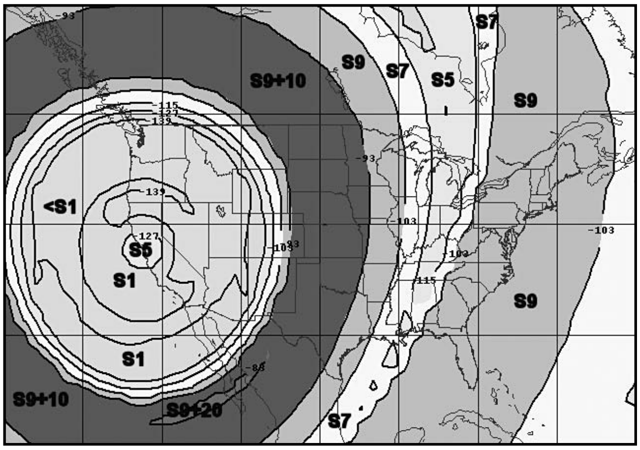

Fig 26—Modified *VOAAREA* plot for 21.2 MHz from San Francisco to the rest of the US, annotated with signal levels in S units, as well as signal contours in dBW (dB below a watt). Antennas are assumed to be 3-element Yagis at 55 feet above flat ground; the transmitter power is 1500 W; the month is November with SSN = 50, a moderate level of solar activity, at 22 UTC. The most obvious feature is the large "skip zone" centered on the transmitter in San Francisco, extending almost a 1/3 of the distance across the US.

Fig 26 shows another way of looking at propagation—a *geographic area* look. Fig 26 shows 15-meter signal levels across the US as they propagate from a transmitting station in San Francisco. This simulation of propagation conditions is for the month of November, with a medium level of solar activity (SSN = 50) at 22 UTC. Fig 26 was created using the *VOAAREA* software program, part of the *VOACAP* software suite. Transmitter power is assumed to be 1500 W, with 3-element Yagis, 55 feet high, at the transmitter and at each receiving location.

From the transmitter out to about 50 miles, signals are moderate, at about S5 on an S meter. Beyond that coverage area to almost 1/3 of the way across the country (to Colorado), there is a large and distinctive skip zone, where only very weak signals return to Earth (S1 or less). Beyond Colorado, signals rapidly build up to S9+10 dB across the middle of the US, falling to S9 and then to S7 in the vicinity of Chicago, Illinois. Beyond Chicago, the signals drop to S5 in a swath from Michigan and part of Ohio down to Alabama. All along the US East Coast, signals come back strong at S9.

The reason why the signals in Fig 26 drop down to S5 in the Midwest is that the necessary elevation angles to cover this region in a single F2 hop are extremely low even at a moderate level of solar activity. To achieve launch angles as low as 1° requires either very high antenna heights or a high mountaintop location. Beyond the Midwest, out to the US East Coast, two F2 hops are required, with higher elevation angles and hence greater antenna gain for moderate antenna heights.

Non-Hopping Propagation Modes

Present propagation theory holds that for communication distances of many thousands of kilometers, signals do not always hop in relatively short increments from ionosphere-to-Earth-to-ionosphere and so forth along the entire path. Instead, the wave is thought to propagate *inside* the ionosphere throughout some portion of the path length, tending to be *ducted* in the ionized layer.

As was shown in Fig 24, the high-angle Pedersen ray can also penetrate an ionospheric layer farther than lower-angle rays. In the less-densely ionized upper edge of the layer, the amount of refraction is less, nearly equaling the curvature of the layer itself as it encircles the Earth.

Non-hopping theory of long-distance propagation is further supported by studies of travel times for signals that go completely around the world. The time required is significantly less than would be necessary to hop between the Earth and the ionosphere 10 or more times while circling the Earth.

Propagation between two points thousands of kilometers apart may in fact consist of a combination of ducting and hopping. It may involve combinations of refractions from the E layer and the F layer. Despite all the complex factors involved, most long-distance propagation can be seen to follow certain general rules. Thus, much commercial and military point-to-point communication over long distances employs antennas designed to make maximum use of known radiation angles and layer heights, even on paths where multihop propagation is assumed.

In amateur work, however, we usually try for the lowest practical radiation angle, hoping to keep reflection losses to a minimum. Years of amateur experience have shown this to be a decided advantage under all usual conditions.

The geometry of propagation by means of the F_2 layer limits our maximum distance along the Earth's surface to about 4000 km (2500 miles) for a single hop. For higher radiation angles, this same distance may require two or more hops (with higher reflection loss). And fewer hops are better, in most cases. If you have a nearby neighbor who consistently outperforms you on the longer paths, a radiation angle difference in his favor is probably the reason.

MAXIMUM USABLE FREQUENCY

The vertical-incidence critical frequency is the *maximum usable frequency* for local sky-wave high-angle communication. It is also useful in the selection of optimum working frequencies and the determination of the maximum usable frequency for distant points at a given time. The abbreviation "MUF" for maximum usable frequency will be used hereafter.

In geographic middle latitudes, the vertical-incident critical frequency ranges between about 1 and 4 MHz for the E layer, and between 2 and 13 MHz for the F_2 layer. The lowest figures are for nighttime conditions in the lowest years of the solar cycle. The highest are for the daytime hours in the years of high solar activity. These are average figures. Critical frequencies have reached as high as 20 MHz briefly during exceptionally high solar activity in the middle latitudes. As was pointed out earlier in Fig 22, $f_o F_2$ levels approaching 40 MHz are possible at low latitudes.

While vertical-incidence critical frequencies are interesting from a scientific point of view, hams are far more concerned about how we can exploit propagation conditions to communicate, preferably at long distances. The MUF for a 4000-km (2500 miles) distance is about 3.5 times the vertical-incidence critical $f_o F_2$ frequency existing at the path midpoint. For one-hop signals, if a uniform ionosphere is assumed, the MUF decreases with shorter distances along the path. This is true because the higher-frequency waves must be launched at higher elevation angles for shorter ranges, and at these launch angles they are not bent sufficiently to reach the Earth. Thus, a lower frequency (where more bending occurs) must be used.

Precisely speaking, a maximum usable frequency or MUF is defined for communication between two specific points on the Earth's surface, for the conditions existing at the time, including the minimum elevation angle that the station can launch at the frequency in use. (This practical form of MUF is sometimes called the *operational MUF*). At the same time and for the same conditions, the MUF from either of these two points to a third point may be different.

Table 2

Time and Frequency Stations Useful for Propagation Monitoring

Call	*Frequency (MHz)*	*Location*
WWV	2.5, 5, 10, 15, 20	Ft Collins, Colorado
WWVH	Same as WWV but no 20	Kekaha, Kauai, Hawaii
CHU	3.330, 7.335, 14.670	Ottawa, Ontario, Canada
RID	5.004, 10.004, 15.004	Irkutsk, USSR*
RWM	4.996, 9.996, 14.996	Novosibirsk, USSR
VNG	2.5, 5, 8.634, 12.984, 16	Lyndhurst, Australia
BPM	5, 5.43, 9.351, 10, 15	Xiang, China
BSF	15	Taoyuan, Taiwan
JJY	2.5, 5, 8, 10, 15	Tokyo, Japan
LOL	5, 10, 15	Buenos Aires, Argentina

*The call, taken from an international table, may not be the one used during actual transmission. Locations and frequencies appear to be accurate as provided.

Therefore, the MUF cannot be expressed broadly as a single frequency, even for any given location at a particular time. The ionosphere is never uniform, and in fact at a given time and for a fixed distance, the MUF changes significantly with changes in compass direction for almost any point on the Earth. Under usual conditions, the MUF will always be highest in the direction toward the Sun—to the east in the morning, to the south at noon (from northern latitudes), and to the west in the afternoon and evening.

For the strongest signals at the greatest distance, especially where the limited power levels of the Amateur Radio Service are concerned, it is important to work fairly near the MUF. It is at these frequencies where signals suffer the least loss. The MUFs can be estimated with sufficient accuracy by using the prediction charts that appear on the ARRL Web site (**www.arrl.org/qst/propcharts/**) or by using a computer prediction program. The CD-ROM bundled in the back of this book contains detailed and summary tables for more than 175 transmitting locations around the world. (See section on "What HF Bands Are Open—Where and When?" later in this chapter.)

MUFs can also be observed, with the use of a continuous coverage communications receiver. Frequencies up to the MUFs are in round-the-clock use today. When you "run out of signals" while tuning upward in frequency from your favorite ham band, you have a pretty good clue as to which band is going to work well, right then. Of course, it helps to know the direction to the transmitters whose signals you are hearing. Shortwave broadcasters know what frequencies to use, and you can hear them anywhere, if conditions are good. Time-and-frequency stations are also excellent indicators, since they operate around the clock. See **Table 2**. WWV is also a reliable source of propagation data, hourly, as discussed in more detail later in this chapter.

The value of working near the MUF is two-fold. Under undisturbed conditions, the absorption loss decreases proportional to the square of a change in frequency. For example, the absorption loss is four times higher at 14 MHz than it is at 28 MHz. Perhaps more important, the hop distance is considerably greater as the MUF is approached. A transcontinental contact is thus much more likely to be made on a single hop on 28 MHz than on 14 MHz, so the higher frequency will give the stronger signal most of the time. The strong-signal reputation of the 28-MHz band is founded on this fact.

LOWEST USABLE FREQUENCY

There is also a lower limit to the range of frequencies that provide useful communication between two given points by way of the ionosphere. *Lowest usable frequency* is abbreviated LUF. If it were possible to start near the MUF and work gradually lower in frequency, the signal would decrease in strength and eventually would disappear into the ever-present “background noise.” This happens because signal absorption increases proportional to the square of the lowering of the frequency. The frequency nearest the point where reception became unusable would be the LUF. It is not likely that you would want to work at the LUF, although reception could be improved if the station could increase power by a considerable amount, or if larger antennas could be used at both ends of the path.

For example, when solar activity is very high at the peak of a solar cycle, the LUF often rises higher than 14 MHz on the morning Eastern US-to-Europe path on 20 meters. Just before sunrise in the US, the 20-meter band will be first to open to Europe, followed shortly by 15 meters, and then 10 meters as the Sun rises further. By mid-morning, however, when 10 and 15 meters are both wide open, 20 meters will become very marginal to Europe, even when both sides are running maximum legal power levels. By contrast, stations on 10 meters can be worked readily with a transmitter power of only 1 or 2 watts, indicating the wide range between the LUF and the MUF.

Frequently, the *window* between the LUF and the MUF for two fixed points is very narrow, and there may be no amateur frequencies available inside the window. On occasion the LUF may be higher than the MUF between two points. This means that, for the highest possible frequency that will propagate through the ionosphere for that path, the absorption is so great as to make even that frequency unusable. Under these conditions it is impossible to establish amateur sky-wave communication between those two points, no matter what frequency is used. (It would normally be possible, however, to communicate between either point and other points on some frequency under the existing conditions.) Conditions when amateur sky-wave communication is impossible between two fixed points occur commonly for long distances where the total path is in darkness, and for very great distances in the daytime during periods of low solar activity.

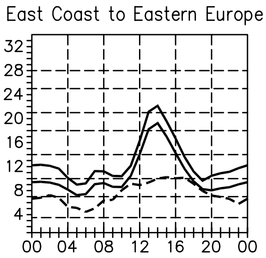

Fig 27—Propagation prediction chart for East Coast of US to Europe. This appeared in December 1994 *QST*, where an average 2800-MHz (10.7-cm) solar flux of 83 was assumed for the mid-December to mid-January period. On 10% of these days, the highest frequency propagated was predicted at least as high as the uppermost curve (the Highest Possible Frequency, or HPF, approximately 21 MHz), and for 50% of the days as high as the middle curve, the MUF. The broken lines show the Lowest Usable Frequency (LUF) for a 1500-W CW transmitter.

Fig 27 shows a typical propagation prediction from the ARRLWeb members-only site (**www.arrl.org/qst/propcharts/**, previously from the “How’s DX” column in *QST*). In this instance, the MUF and the LUF lines are blurred together at about 10 UTC, meaning that the statistical likelihood of any amateur frequency being open for that particular path at that particular time was not very good. Later on, after about 11 UTC, the gap between the MUF and LUF increased, indicating that the higher bands would be open on that path.

DISTURBED IONOSPHERIC CONDITIONS

So far, we have discussed the Earth’s ionosphere when conditions at the Sun are undisturbed. There are three general types of major disturbances on the Sun that can affect radio propagation here on the Earth. On the air, you may hear people grousing about *Solar Flares*, *Coronal Holes* or *Sudden Disappearing Filaments*, especially when propagation conditions are not good. Each of these disturbances causes both electromagnetic radiation and ejection of material from the Sun.

Solar Flares

Solar flares are cataclysmic eruptions that suddenly release huge amounts of energy, including sustained, high-energy bursts of radiation from VLF to X-ray frequencies and vast amounts of solar material. Most solar flares occur around the peak of the 11-year solar cycle.

The first Earthly indication of a huge flare is often a visible brightness near a sunspot group, along with increases in UV, X-ray radiation and VHF radio noise. If the geometry between the Sun and Earth is right, intense X-ray radiation takes eight minutes, traveling the 93 million miles to Earth at the speed of light. The sudden increase in X-ray energy can immediately increase RF absorption in the Earth's lowest ionospheric layers, causing a phenomenon known as a *Sudden Ionospheric Disturbance* (SID).

An SID affects all HF communications on the sunlit side of the Earth. Signals in the 2 to 30-MHz range may disappear entirely, and even most background noise may cease in extreme cases. When you experience a big SID, your first inclination may be to look outside to see if your antenna fell down! SIDs may last up to an hour before ionospheric conditions temporarily return to normal.

Between 45 minutes and 2 hours after an SID begins, particles from the mass eruption on the Sun may begin to arrive. These high-energy particles are mainly protons and they can penetrate the ionosphere at the Earth's magnetic poles, where intense ionization can occur, with attendant absorption of HF signals propagating through the polar regions. This is called a *Polar Cap Absorption* (PCA) event and it may last for several days. A PCA results in spectacular auroral displays at high latitudes.

Coronal Holes

As described earlier in the section dealing with auroral propagation at VHF, a second major solar disturbance is a so-called "coronal hole" in the Sun's outer layer (the *corona*). Temperatures in the corona can be more than four million °C over an active sunspot region but more typically are about two million °C. A coronal hole is an area of somewhat lower temperature. Solar-terrestrial scientists have a number of competing theories about how coronal holes are formed.

Matter ejected through this "hole" takes the form of a *plasma*, a highly ionized gas made up of electrons, protons and neutral particles, traveling at speeds up to 1,000 km per second (2 million miles per hour). The plasma becomes part of the solar wind and can affect the Earth's magnetic field, but only if the Sun-Earth geometry is right. A plasma has a very interesting and somewhat bizarre ability. It can lock-in the orientation of the magnetic field where it originates and carry it outwards into space. However, unless the locked-in magnetic field orientation is aligned properly with the Earth's magnetic field, even a large plasma mass may not severely disrupt our magnetosphere, and thence our ionosphere.

Presently, we don't have the ability to predict very long in advance when the Sun might erupt in a disturbance that results in Earthly propagation problems. The SOHO satellite can help determine whether a mass ejection is heading towards Earth, and the ACE satellite about 1 million miles away from Earth can give about an hour's warning whether the imbedded magnetic field in a mass ejection from the Sun might impact the Earth's magnetosphere, causing propagation problems for hams.

Statistically, coronal holes tend to occur most often during the declining phase of the 11-year solar cycle and they can last for a number of solar rotations. This means that a coronal hole can be a "recurring coronal hole," disrupting communications for several days about the same time each month for as long as a year, or even more.

Sudden Disappearing Filaments

A sudden disappearing filament (SDF) is the third major category of solar disturbance that can affect propagation. SDFs take their names from the manner in which they suddenly arch upward from the Sun's surface, spewing huge amounts of matter as plasma out into space in the solar wind. They tend to occur mostly during the rising phase of the 11-year solar cycle.

IONOSPHERIC STORMS

When the conditions are right, a flare, coronal hole or an SDF can launch a plasma cloud into the solar wind, resulting in an *ionospheric storm* here on Earth. Unlike a hurricane or a winter Nor'easter storm in New England, an ionospheric storm is not something we can see with our eyes or feel on our skins. We can't easily measure things occurring in the ionosphere some 200 miles overhead. However, we can see the indirect effects of an ionospheric storm on magnetic instruments located on the Earth's surface, because disturbances in the ionosphere are closely related to the Earth's magnetic field. The term *Geomagnetic Storm* ("Geo" means "Earth" in Greek) is used almost synonymously with ionospheric storm.

During a ionospheric storm, we may experience extraordinary radio noise and interference, especially at HF. You may hear solar radio emissions as increases of noise at VHF. A geomagnetic storm generally adds noise and weakens or disrupts ionospheric propagation for several days. Transpolar signals at 14 MHz or higher may be particularly weak, with a peculiar hollow sound or flutter—even more than normal for transpolar signals.

Depending on the severity of the disturbance to the Earth's geomagnetic field and the consequent disturbance of the ionosphere, propagation may be disrupted completely or it might be at least degraded for a period of time that ranges from a day to three or four days before returning to normal propagation conditions.

What can we do about the solar disturbances and related disturbed ionospheric propagation on Earth? The truth is that we are powerless faced with the truly awesome forces of solar disturbances like flares, coronal holes or sudden disappearing filaments. Perhaps there is some comfort, however, in understanding what has happened to cause our HF bands to be so poor. And as a definite consolation, conditions on the VHF bands are often exceptionally good just when HF propagation is remarkably poor due to solar disturbances. VHF operators enthusiastically look forward to conditions when they can engage in auroral communications—exactly the kind of conditions that have HF operators scratching their heads, wondering where the ionosphere went.

ELEVATION ANGLES FOR HF COMMUNICATION

It was shown in connection with Fig 25 that the distance at which a ray returns to Earth depends on the eleva-

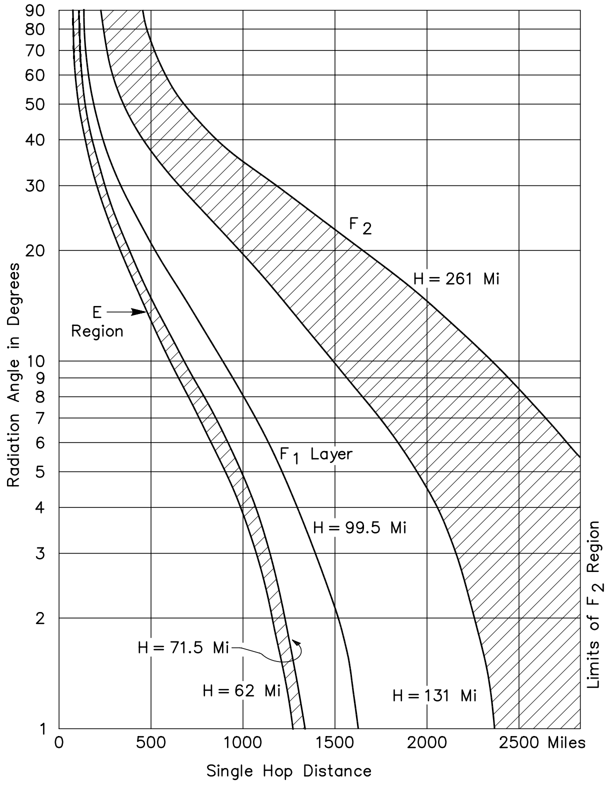

Fig 28—Distance plotted against wave angle (one-hop transmission) for the nominal range of heights for the E and F2 layers, and for the F1 layer.

tion angle at which it left the Earth (also known by other names: takeoff, launch or wave angle). Chapter 3, The Effects of Ground, in this book deals with the effects of local terrain, describing how the elevation angle of a horizontally polarized antenna is determined mainly by its height above the ground.

Although it is not shown specifically in Fig 25, propagation distance also depends on the layer height at the time, as well as the elevation angle. As you can probably imagine, the layer height is a very complex function of the state of the ionosphere and the Earth's geomagnetic field. There is a large difference in the distance covered in a single hop, depending on the height of the E or the F_2 layer. The maximum single-hop distance by the E layer is about 2000 km (1250 miles) or about half the maximum distance via the F_2 layer. Practical communicating distances for single-hop E or F layer work at various wave angles are shown in graphic form in **Fig 28**.

Actual communication experience usually does not fit the simple patterns shown in Fig 25. Propagation by means of the ionosphere is an enormously complicated business (which makes it all the more intriguing and challenging to radio amateurs, of course), even when the Sun is not in a disturbed state. Until the appearance of sophisticated computer models of the ionosphere, there was little definitive information available to guide the radio amateur in the design of his antenna systems for optimal performance over all portions of the 11-year solar cycle. Elevation angle information that had appeared for many years in *The ARRL Antenna Book* was measured for only one transmitting path, during the lowest portion of Solar Cycle 17 in 1934.

The *IONCAP* Computer Propagation Model

Since the 1960s several agencies of the US government have been working on a detailed computer program that models the complex workings of the ionosphere. The program has been dubbed *IONCAP*, short for "Ionospheric Communications Analysis and Prediction Program." *IONCAP* was originally written for a mainframe computer, but later versions have been rewritten to allow them to be run by high-performance personal computers. *IONCAP* incorporates a detailed database covering almost three complete solar cycles. The program allows the operator to specify a wide range of parameters, including detailed antenna models for multiple frequency ranges, noise models tailored to specific local environments (from low-noise rural to noisy residential QTHs), minimum elevation angles suitable for a particular location and antenna system, different months and UTC times, maximum levels of multipath distortion, and finally solar activity levels, to name the most significant of a bewildering array of options.

While *IONCAP* has a well-justified reputation for being very *unfriendly* to use, due to its mainframe, non-interactive background, it is also the one ionospheric model most highly regarded for its accuracy and flexibility, both by amateurs and professionals alike. It is the program used for many years to produce the long-term MUF charts formerly included in the "How's DX" monthly column of *QST* and now available on the Members Only ARRLWeb page.

IONCAP is not well suited for short-term forecasts of propagation conditions based on the latest solar indices received from WWV. It is an excellent tool, however, for long-range, detailed planning of antenna systems and shortwave transmitter installations, such as that for the Voice of America, or for radio amateurs. See the section later in this chapter describing other computer programs that can be used for short-term, interactive propagation predictions.

IONCAP/VOACAP Parameters

The elevation-angle statistical information contained in this section was compiled from thousands of *VOACAP* runs (an improved version of *IONCAP* developed by scientists from VOA, the Voice of America). These runs were done for a number of different transmitting locations throughout the world to important DX locations throughout the world.

Some assumptions were needed for setting *VOACAP* parameters. The transmitting and receiving sites were all assumed to be located on flat ground, with "average" ground conductivity and dielectric constant. Each site was assumed to have a clear shot to the horizon, with a minimum elevation angle less than or equal to 1°. Electrical noise at each

Table 3

Boston, Massachusetts, to All of Europe

E*lev*	*80 m*	*40 m*	*30 m*	*20 m*	*17 m*	*15 m*	*12 m*	*10 m*
1	4.1	9.6	4.6	1.7	2.1	4.4	5.5	7.2
2	0.8	2.3	7.2	1.4	2.8	2.8	3.7	5.3
3	0.3	0.7	4.3	3.1	2.4	2.2	4.4	7.9
4	0.5	4.1	8.7	11.6	12.2	9.4	8.1	3.9
5	4.6	4.8	7.5	12.7	14.3	13.1	9.2	11.2
6	7.1	8.9	5.5	9.2	9.6	12.2	9.2	7.2
7	8.5	6.9	7.2	4.6	7.9	7.4	10.0	5.9
8	5.1	7.0	5.4	3.2	5.9	7.4	4.8	6.6
9	3.3	5.6	3.2	3.1	2.1	3.9	8.1	9.2
10	1.0	4.0	7.9	6.3	5.1	3.7	11.1	6.6
11	1.9	3.8	9.7	10.2	7.2	5.4	3.7	7.9
12	5.6	3.4	4.8	8.5	6.9	7.4	4.8	6.6
13	11.0	3.0	2.4	4.1	5.9	4.6	3.3	2.6
14	7.6	4.8	2.0	2.7	3.8	3.9	6.3	5.9
15	5.3	7.9	2.0	1.5	2.4	1.7	1.5	2.0
16	2.8	6.4	3.8	2.9	1.5	1.3	2.6	2.6
17	5.0	3.4	4.5	3.1	1.0	1.5	0.0	0.0
18	4.2	2.0	3.1	3.1	2.0	2.2	1.8	1.3
19	5.7	1.4	1.4	2.3	1.3	0.7	0.0	0.0
20	6.6	1.4	1.2	1.8	1.1	1.3	0.7	0.0
21	4.4	1.4	0.5	0.8	0.7	0.7	0.4	0.0
22	2.3	2.4	1.0	1.1	0.6	1.3	0.7	0.0
23	1.3	1.8	0.1	0.3	0.1	0.0	0.0	0.0
24	0.6	1.0	0.5	0.5	0.4	0.7	0.0	0.0
25	0.3	0.8	0.3	0.1	0.4	0.0	0.0	0.0
26	0.0	0.5	0.7	0.2	0.1	0.4	0.0	0.0
27	0.1	0.1	0.1	0.2	0.1	0.2	0.0	0.0
28	0.0	0.3	0.1	0.2	0.0	0.2	0.0	0.0
29	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.0
30	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
31	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
33	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Percentage of time a particular frequency band is open on this specific propagation path.

receiving location was also assumed to be very low.

Transmitting and receiving antennas for the 3.5 to 30-MHz frequency range were specified to be isotropic-type antennas, but with +6 dBi gain, representing a good amateur antenna on each frequency band. These theoretical antennas radiate uniformly from the horizon, up to 90° directly overhead. With response patterns like this, these are obviously not real-world antennas. They do, however, allow the computer program to explore all possible modes and elevation angles.

Looking at the Elevation-Angle Statistical Data

Table 3 shows detailed statistical elevation informa-

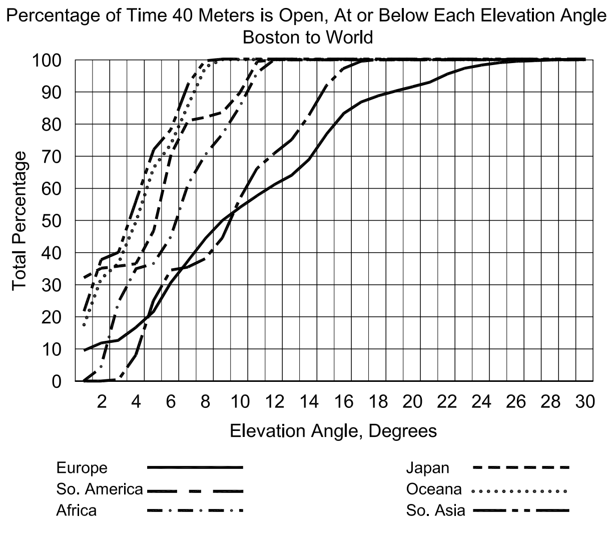

Fig 29—The cumulative distribution function showing the total percentage of time that 40 meters is open, at or below each elevation angle, from Boston to the world. For example, 50% of the time the band is open to Europe from Boston, it is at 10° or less. The angles for DX work are indeed low.

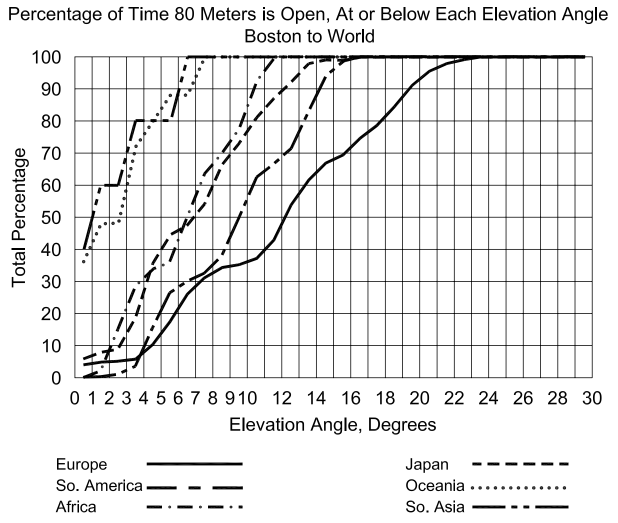

Fig 30—The cumulative distribution function showing the total percentage of time that 80 meters is open, at or below each elevation angle, from Boston to the world. For example, 50% of the time the band is open to Europe from Boston, it is at 13° or less.

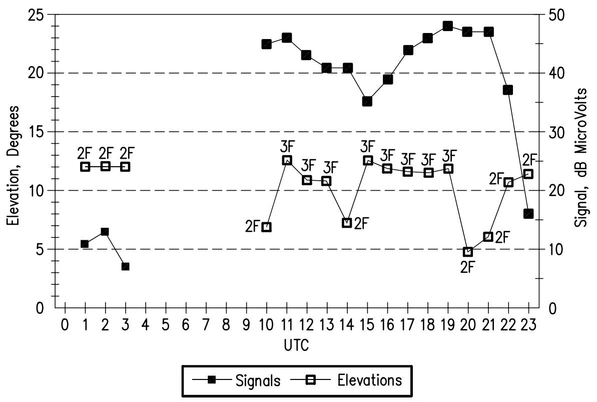

Fig 31—Overlay of signals and elevation angles, together with hop-mode information. This is for one month, October, at one level of solar activity, SSN=70, for the path from Newington, CT, to London, England. The mode of propagation does not closely follow the elevation angle. From 15 to 19 UTC the mode is 3F2 hops, and the elevation angle is approximately 12°. The same elevation angle is required from 23 to 03 UTC, but here the mode is 2F2 hops.

tion for the path from Boston, Massachusetts, near ARRL HQ in Newington, Connecticut, to all of Europe. The data incorporated into Table 3 shows the percentage of time versus elevation angle for all HF bands from 80 meters to 10 meters, over all portions of the 11-year solar cycle. The CD-ROM accompanying this book contains more tables such as this for more than 150 transmitting sites around the world. These tables are used by the *HFTA* program (and earlier *YT* program) and can also be imported into many programs, such as word processors or spreadsheets. Six important areas throughout the world are covered, one per table: all of Europe (from London, England, to Kiev, Ukraine), the Far East (centered on Japan), South America (Paraguay), Oceania (Melbourne, Australia), Southern Africa (Zambia) and South Asia (New Delhi, India).

You may be surprised to see in Table 3 that angles lower than 10° dominate the possible range of incoming angles for this moderate-distance path from New England to Europe. In fact, 1.7% of all the times when the 20-meter band is open to Europe, the takeoff angle is as low as 1°. You should recognize that very few real-world 20-meter antennas achieve much gain at such an extremely low angle—unless they just happen to be mounted about 400 feet high over flat ground or else are located on the top of a tall, steep mountain.

The situation is even more dramatic on 40 and 80 meters. **Fig 29** shows the “cumulative distribution function” of the total percentage of time (derived from Table 3) when 40 meters is open from Boston to the rest of the world, plotted against the elevation angle. For example, into Europe from Boston, 50% of the time when the band is open, it is at 10° or less. Into Japan from Boston, the statistics are even more revealing: 50% of the time when the band is open, the angle is 6° or less, and 90% of the time the angle is 13° or less!

Fig 30 shows the same sort of information for 80 meters from Boston to the world. For 50% of the time from Boston to Europe the elevation angle is 13° or less; at the 90% level the angle is 20° or less. For the path to Japan on 80 meters from Boston, 50% of the time the angle is 8° or less; at the 90% level, the angle is 13° or less. Now, to achieve peak gain on 80 meters at an elevation angle of 8° over flat land, a horizontally polarized antenna must be 500 feet high. You can begin to see why verticals can do very well on long-distance contacts on 80 meters, even when they are mounted over poorly conducting, rocky ground. Clearly, low angles are very important for successful DXing.

The Ionosphere Controls Propagation

You should always remember that it is the *ionosphere* that controls the elevation angles, *not* the transmitting antenna. The elevation response of a particular antenna only determines how strong or weak a signal is, at whatever angle (or angles) the ionosphere is supporting at that particular instant, for that propagation path and for that frequency.

If only one propagation mode is possible at a particular time, and if the elevation angle for that one mode happens to be 5°, then your antenna will have to work satisfactorily at that very low angle or else you won’t be able to communicate. For example, if your low dipole has a gain of –10 dBi at 5°, compared to your friend’s Yagi on a mountain top with +10 dBi gain at 5°, then you will be down 20 dB compared to his signal. It’s not that the elevation angle is somehow *too low*—the real problem here is that you don’t have *enough gain* at that particular angle where the ionosphere is supporting propagation. Many “flatlanders” can vividly recall the times when their mountain-top friends could easily work DX stations, while they couldn't even hear a whisper.

Looking at the Data—Further Cautions

A single propagation mode is quite common at the opening and the closing of daytime bands like 20, 15 or 10 meters, when the elevation angle is often lower (but not always) than when the band is wide open. The lower-frequency bands tend to support multiple propagation modes simultaneously. For example, **Fig 31** plots the signal strength (in dBμV) and the elevation angle for the dominant mode (with the strongest signal) over a 24-hour period from Newington to London in October, for a medium-level SSN = 70. The morning opening at 10 UTC starts out with a two-hop $2F_2$ mode (labeled 2F) at an elevation angle of 6°. By 11 UTC the mode has changed to a three-hop $3F_2$ (labeled 3F) at a 12° elevation angle. The band starts to close down with weaker signals after about 23 UTC. Note that this path actually supports both $2F_2$ and $3F_2$ modes most of the time. Either mode may be stronger than the other, depending on the particular time of day.

It is tempting to think that two-hop signals always occur at lower elevation launch angles, while three-hop signals require higher elevation angles. In reality, the detailed workings of the ionosphere are enormously complicated. From 22 UTC to 03 UTC, the elevation angles are higher than 11° for $2F_2$ hops. During much of the morning and early afternoon in Newington (from 11 to 13 UTC, and from 15 to 19 UTC), the angles are also higher than 11°. However, $3F_2$ hops are involved during these periods of time. The number of hops is not directly related to the elevation angles needed—changing layer heights account for this.

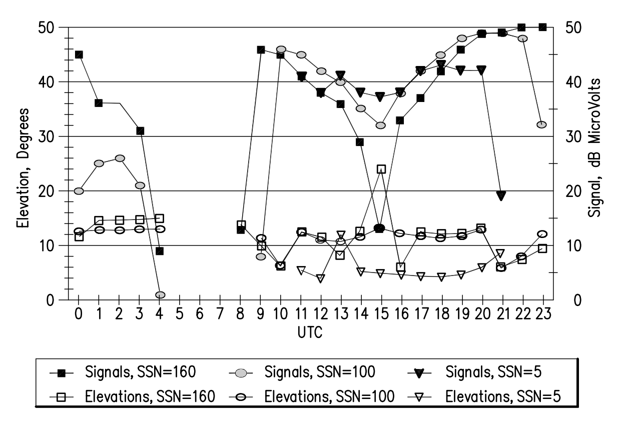

Fig 32—October 20-meter signals and elevation angles for the full range of solar activity, from W1 to England. The elevation angle does not closely follow the level of solar activity. What is important in designing a station capable of covering all levels of solar activity is to have flexibility in antenna elevation pattern response — to cover a wide range of possible angles.

Note that starting around 15 UTC, the mid-morning 20-meter "slump" (down some 10 dB from peak signal level) is caused by higher levels of mainly E-layer absorption when the Sun is high overhead. This condition favors higher elevation angles, since signals launched at lower angles must travel for a longer time through the lossy lower layer.

How does the situation change with different levels of solar activity? **Fig 32** overlays predicted signals and elevation angles for three levels of solar activity in October, again for the Newington-London path. Fig 32 shows the mid-morning slump dramatically when the solar activity is at

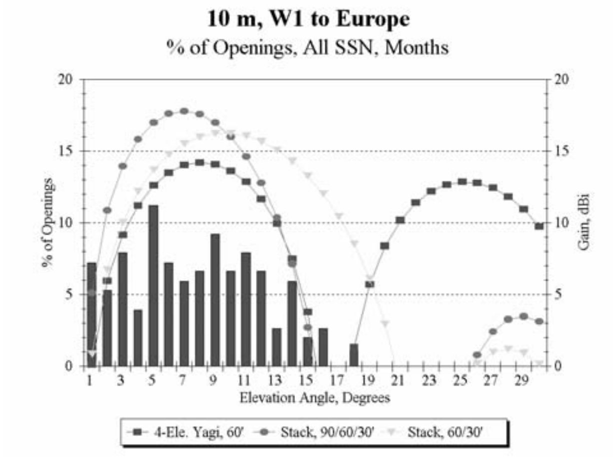

Fig 33—10-meter graph of the percentage of all openings versus elevation angles, together with overlay of elevation patterns over flat ground for three 10-meter antenna systems. Stacked antennas have wider "footprints" in elevation angle coverage for this example from New England to Europe.

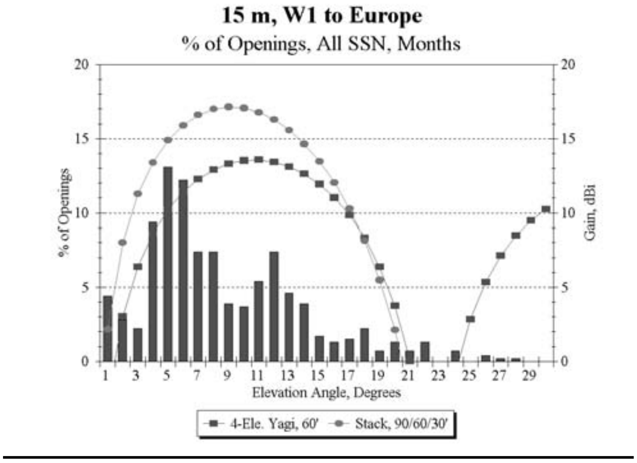

Fig 34—15-meter graph of the percentage of all openings versus elevation angles, together with overlay of elevation patterns over flat ground for two 15-meter antenna systems. Like 10 meters, 15-meter stacked antennas have wider footprints in elevation angle coverage for this example from New England to Europe.

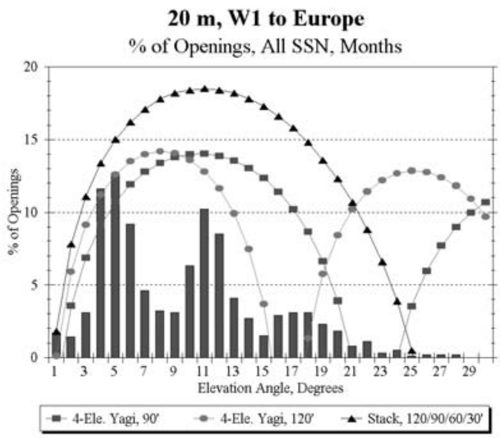

Fig 35—20-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for three 20-meter antenna systems.

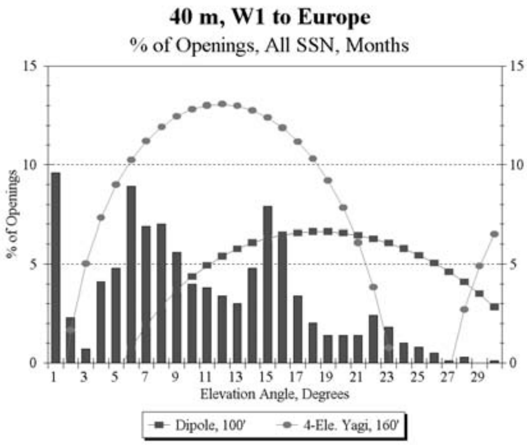

Fig 36—40-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlays of elevation patterns over flat ground for a 100-foot high dipole and a large 4-element Yagi at 160 feet. Achieving gain at very low elevation angles requires very high heights above ground.

a very high level, represented by SSN = 160. At 15 UTC, the signal level drops 35 dB from peak level, and the elevation angle rises all the way to 24°. By the way, as a percentage of all possible openings, the 24° angle occurs only rarely, 0.5% of the time. It barely shows up as a blip in Table 3. Elevation angles are *not* closely related to the level of solar activity.

IONCAP/VOACAP demonstrates that elevation angles do not follow neat, easily identified patterns, even over a 24-hour period—much less over all portions of the solar cycle. Merely looking at the percentage of all openings versus elevation angle, as shown in Table 3, does not tell the whole story, although it is probably the most statistically valid approach to station design, and possibly the most emotionally satisfying approach too. Neither is the whole story revealed by looking only at a snapshot of elevation angles versus time for one particular month, or for one solar activity level.

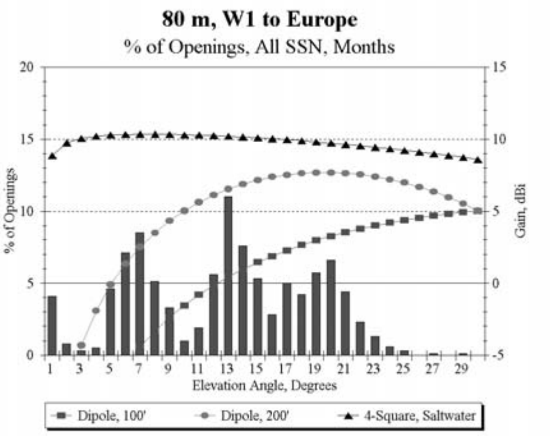

Fig 37—80-meter graph of the percentage of all openings from New England to Europe versus elevation angles, together with overlay of elevation patterns over flat ground for dipoles at two different heights. The 200-foot-high dipole clearly covers the necessary elevation angles better than does the 100-foot-high dipole, although a Four Square vertical array located over saltwater is even better for all angles needed.

What is important to recognize is that the most effective antenna system will be one that can cover the *full range* of elevation angles, over the whole spectrum of solar activity, even if the actual angle in use at any one moment in time may not be easy to determine. For this particular path, from New England to all of Europe, an ideal antenna would have equal response over the full range of angles from 1° to 28°. Unfortunately, real-world antennas have a tough time covering such a wide range of elevation angles equally well.

Antenna Elevation Patterns

Figs 33 through **37** show overlays of the same sort of elevation angle information listed in Table 3, together with the elevation response patterns for typical antennas for the HF amateur bands 80, 40, 20, 15 and 10 meters. For example, Fig 34 shows an overlay for 20 meters, with three different types of 20-meter antennas. These are a 4-element Yagi at 90 feet, a 4-element Yagi at 120 feet and a large stack of four 4-element Yagis located at 120, 90, 60 and 30 feet. Each antenna is assumed to be mounted over flat ground. Placement on a hill with a long slope in the direction of interest would lower the required elevation angle by the amount of the hill's slope. For example, if a 10° launch angle is desired, and the antenna is placed on a hill with a slope of –5°, the antenna itself should be designed for a height that would optimize the response at 15° over flat ground—one wavelength high.

In Fig 35, the large stack of four 20-meter Yagis over flat ground comes closest to being ideal, but even this large array will not work well for that very small percentage of time when the angle needed is higher than about 20°. Some hams might conclude that the tiny percentage of time when the angles are very high doesn't justify an antenna tailored for that response. However, when that new DX country pops up on a band, or when a rare multiplier shows up in a contest, doesn't it always seem that the desired signal only comes in at some angle your antenna doesn't cover well? What do you do then, if your only antenna happens to be a large stack?

The answer to this, perhaps unique, high-angle problem lies in switching to using only the top antenna in the stack. In this example, the second elevation lobe of the 120-foot high antenna would cover the angles from 20° to 30° well, much better than the stack does. Note that the top antenna by itself would not be ideal for all conditions. It is simply too high much of the time when the elevation angles are higher than about 12°. The experience of many amateurs on the US East Coast with high 20-meter antennas bears this out—they find that 60 to 90-foot high antennas are far more consistent performers into Europe.

ONE-WAY PROPAGATION

On occasion a signal may be started on the way back toward the Earth by reflection from the F region, only to come down onto the top of the E region and be reflected back up again. This set of conditions is one possible explanation for the often-reported phenomenon called *one-way skip*. The reverse path may not necessarily have the same multilayer characteristic. The effect is more often a difference in the signal strengths, rather than a complete lack of signal in one direction, and many times there may be local noises that mask signals at one end of the path. It is important to remember these sorts of possibilities when a long-distance test with a new antenna system yields apparently conflicting results. Even many tests, on paths of different lengths and headings, may provide data that are difficult to understand. Communication by way of the ionosphere is not always a source of consistent answers to antenna questions.

Fig 37 shows the 80-meter path from New England to Europe with three different antennas. A really high dipole at a height of 200 feet above flat ground would certainly be an impressive antenna. But it would still be overshadowed dramatically by a Four-Square vertical array, at least at the low elevation angles needed often on this path. This is predicated on the Four Square being located over salt water, which provides a virtually perfect RF ground. At an elevation angle of 7°, the Four Square has 7 dB more gain than the 200-foot high dipole.

SHORT OR LONG PATH?

Propagation between any two points on the Earth's surface is usually by the shortest direct route—the *great-circle path* found by stretching a string tightly between the two points on a globe. If an elastic band going completely around the globe in a straight line is substituted for the string, it will show another great-circle path, going "the long way around." The long path may serve for communication over the desired circuit when conditions are favorable along the longer route. There may be times when communication is possible over the long path but not possible at all over the short path. Especially if there is knowledge of this potential at both ends of the circuit, long-path communication may work very well. Cooperation is almost essential, because both the aiming of directional antennas and the timing of the attempts must be right for any worthwhile result. The *IONCAP/VOACAP* computations in the preceding tables were made for short-path azimuths only.

Sunlight is a required element in long-haul communication via the F layer above about 10 MHz. This fact tends to define long-path timing and antenna aiming. Both are essentially the reverse of the "normal" for a given circuit. We know also that salt-water paths work better than overland ones. This can be significant in long-path work.

We can better understand several aspects of long-path propagation if you become accustomed to thinking of the Earth as a ball. This is easy if you use a globe frequently. A

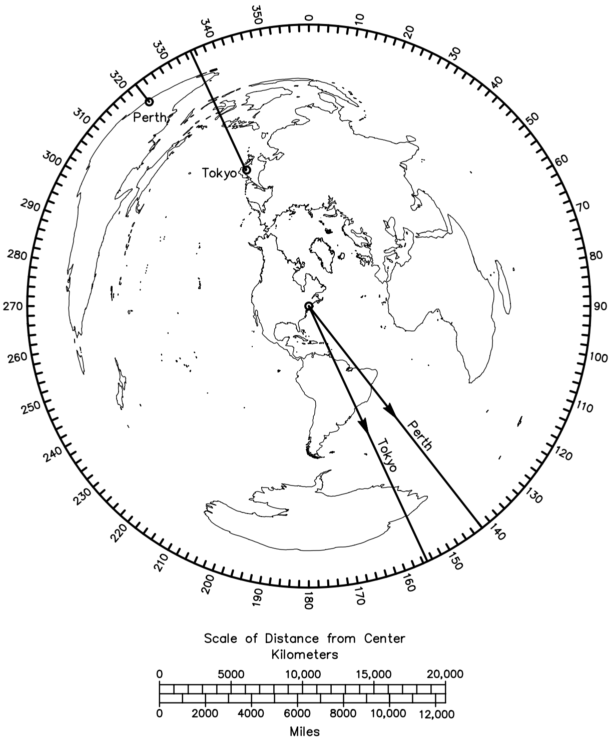

Fig 38—K5ZI's computer-generated azimuthal-equidistant projection centered on Newington, Connecticut. (See Bibliography for ordering information.) Land masses and information showing long paths to Perth and Tokyo have been added. Notice that the paths in both cases lie almost entirely over water, rather than over land masses.

flat map of the world, of the azimuthal-equidistant projection type, is a useful substitute. The ARRL World Map is one, centered on Wichita, Kansas. A similar world map prepared by K5ZI and centered on Newington, Connecticut, is shown in **Fig 38**. These help to clarify paths involving those areas of the world.

Long-Path Examples

There are numerous long-path routes well known to DX-minded amateurs. Two long paths that work frequently and well when 28 MHz is open from the northeastern US are New England to Perth, Western Australia, and New England to Tokyo. Although they represent different beam headings and distances, they share some favorable conditions. By the long path, Perth is close to halfway around the world; Tokyo is about three-quarters of the way. On 28 MHz, both areas come through in the early daylight hours, Eastern Time, but not necessarily on the same days. Both paths are at their best around the equinoxes. (The sunlight is more uniformly distributed over transequatorial paths at these times.) Probably the factor that most favors both is the nature of the first part of the trip at the US end. To work Perth by way of long path, northeastern US antennas are aimed southeast, out over salt water for thousands of miles—the best low-loss start a signal could have. It is salt water essentially all the way, and the distance, about 13,000 miles, is not too much greater than the “short” path.

The long path to Japan is more toward the south, but still with no major land mass at the early reflection points. It is much longer, however, than that to Western Australia. Japanese signals are more limited in number on the long path than on the short, and signals on the average somewhat weaker, probably because of the greater distance.

On the short path, an amateur in the Perth area is looking at the worst conditions—away from the ocean, and out across the huge land mass of North America, unlikely to provide strong ground reflections. The short paths to both Japan and Western Australia, from most of the eastern half of North America, are hardly favorable. The first hop comes down in various western areas likely to be desert or mountains, or both, and not favored as reflection points.

A word of caution: Don’t count on the long-path signals always coming in on the same beam heading. There can be notable differences in the line of propagation via the ionosphere on even relatively short distances. There can be more variations on long path, especially on circuits close to halfway around the world. Remember, for a point exactly halfway around, all directions of the compass represent great-circle paths.

FADING

When all the variable factors in long-distance HF communication are taken in account, it is not surprising that signals vary in strength during almost every contact beyond the local range. In VHF communication we can also encounter some fading at distances greater than just to the visible horizon. These are mainly the result of changes in the temperature and moisture content of the air in the first few thousand feet above the ground.

On paths covered by HF ionospheric modes, the causes of fading are very complex—constantly changing layer height and density, random polarization shift, portions of the signal arriving out of phase, and so on. The energy arriving at the receiving antenna has components that have been acted upon differently by the ionosphere. Often the fading is very different for small changes in frequency. With a signal of a wideband nature, such as high-quality FM, or even double-sideband AM, the sidebands may have different fading rates from each other, or from the carrier. This causes severe distortion, resulting in what is termed *selective fading*. The effects are greatly reduced (but still present to some extent) when single-sideband (SSB) is used. Some immunity from fading during reception (but not to the distortion induced by selective fading) can be had by using two or more receivers on separate antennas, preferably with different polarizations, and combining the receiver outputs in what is known as a *diversity* receiving system.

OTHER PROPAGATION MODES

In propagation literature there is a tendency to treat the various propagation modes as if they were separate and distinct phenomena. This they may be at times, but often there is a shifting from one to another, or a mixture of two or more kinds of propagation affecting communication at one time. In the upper part of the usual frequency range for F-region work, for example, there may be enough tropospheric bending at one end (or both ends) to have an appreciable effect on the usable path length. There is the frequent combination of E and F-region propagation in long-distance work. And in the case of the E region, there are various causes of ionization that have very different effects on communication. Finally, there are weak-signal variations of both tropospheric and ionospheric modes, lumped under the term “scatter.” We look at these phenomena separately here, but in practice we have to deal with them in combination, more often than not.

Sporadic E (E_s)

First, note that this is *E-subscript-s,* a usefully descriptive term, wrongly written “Es” so often that it is sometimes called “ease,” which is certainly not descriptive. *Sporadic E* is ionization at E-layer height, but of different origin and communication potential from the E layer that affects mainly our lower amateur frequencies.

The formative mechanism for sporadic E is believed to be wind shear. This explains ambient ionization being distributed and compressed into a ledge of high density, without the need for production of extra ionization. Neutral winds of high velocity, flowing in opposite directions at slightly different altitudes, produce shears. In the presence of the Earth’s magnetic field, the ions are collected at a particular altitude, forming a thin, overdense layer. Data from rockets entering E_s regions confirm the electron density, wind velocities and height parameters.

The ionization is formed in clouds of high density, lasting only a few hours at a time and distributed randomly. They vary in density and, in the middle latitudes in the Northern Hemisphere, move rapidly from southeast to northwest. Although E_s can develop at any time, it is most prevalent in the Northern Hemisphere between May and August, with a minor season about half as long beginning in December (the summer and winter solstices). The seasons and distribution in the Southern Hemisphere are not so well known. Australia and New Zealand seem to have conditions much like those in the US, but with the length of the seasons reversed, of course. Much of what is known about E_s came as the result of amateur pioneering in the VHF range.

Correlation of E_s openings with observed natural phenomena, including sunspot activity, is not readily apparent, although there is a meteorological tie-in with high-altitude winds. There is also a form of E_s, mainly in the northern part of the north temperate zone, that is associated with auroral phenomena.

At the peak of the long E_s season, most commonly in late June and early July, ionization becomes extremely dense and widespread. This extends the usable range from the more common "single-hop" maximum of about 1400 miles to "double-hop" distances, mostly 1400 to 2500 miles. With 50-MHz techniques and interest improving in recent years, it has been shown that distances considerably beyond 2500 miles can be covered. There is also an E_s "link-up" possibility with other modes, believed to be involved in some 50-MHz work between antipodal points, or even long-path communication beyond 12,500 miles.

When E_s is particularly strong and widespread, even the HF bands can suddenly go *short skip* producing exceptionally strong signals from distances that would normally be in the no-signal "skip zone." Editor N6BV distinctly remembers a spectacular 20-meter E_s opening in September 1994, during the "Hiram Percy Maxim/125" anniversary celebration, when he was living in New Hampshire. Signals on 20 meters were 30 to 40 dB over S9 from all along the Eastern Seaboard, from W2 to W4. One exasperated W3 complained that he had been calling in the huge pileup for 20 minutes. N6BV glanced at the S meter and saw that the W3 was 20 dB over S9, normally a very strong 20-meter SSB signal, but not when almost everybody else was 40 dB over S9!

Such short-skip conditions caused by Sporadic E are more common on 10 meters than they are on 15 or 20 meters. They can result in excellent transAtlantic 10-meter openings during the summer months—when 10 meters is not normally open for F_2 ionospheric propagation.

The MUF for E_s is not known precisely. It was long thought to be around 100 MHz, but in the last 25 years or so there have been thousands of 144-MHz contacts during the summer E_s season. Presumably, the possibility also exists at 222 MHz. The skip distance at 144 MHz does average much longer than at 50 MHz, and the openings are usually brief and extremely variable.

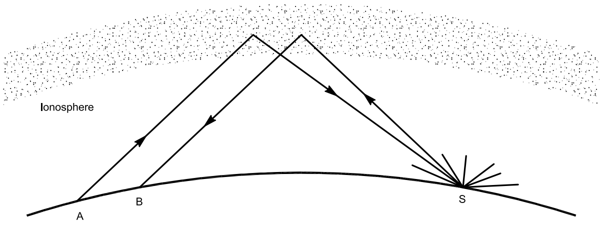

Fig 39—Schematic of a simple backscatter path. Stations A and B are too close to make contact via normal F-layer ionospheric refraction. Signals scattered back from a distant point on the Earth's surface (S), often the ocean, may be accessible to both A and B to create a backscatter circuit. (*Courtesy of The ARRL Handbook.*)

The terms "single" and "double" hop may not be accurate technically, since it is likely that cloud-to-cloud paths are involved. There may also be "no-hop" E_s. At times the very high ionization density produces critical frequencies up to the 50-MHz region, with no skip distance at all. It is often said that the E_s mode is a great equalizer. With the reflecting region practically overhead, even a simple dipole close to the ground may do as well over a few hundred miles as a large stacked antenna array designed for low-angle radiation. It's a great mode for low power and simple antennas on 28 and 50 MHz.

HF Scatter Modes

The term "skip zone" (where no signals are heard) should not be taken too literally. Two stations communicating over a single ionospheric hop can be heard to some degree by other stations at almost any point along the way, unless the two are running low power and using simple antennas. Some of the wave energy is *scattered* in all directions, including back to the starting point and farther.

Backscatter functions like a sort of HF ionospheric radar. **Fig 39** shows a schematic for a simple backscatter path. The signal launched from point A travels through the ionosphere back to earth at Point S, the scattering point. Here, the rough terrain of the land scatters signals in many directions, one of which propagates a weak signal back through the ionosphere to land at point B. Point B would normally be in the no-signal skip zone between A and S. Because backscatter signals arrive from multiple directions, through various paths through the ionosphere, they have a characteristic "hollow" sound, much like you get when you talk into a paper tube with its many internal reflections.

Because backscatter involves mainly scattering from the Earth at the point where the strong ionospherically propagated signal comes down, it is a part of HF over-the-horizon radar techniques. (The infamous 1970s-era "woodpecker" was an over-the-horizon HF radar.) Amateurs using sounding techniques have shown that you can tell to what part of the world a band is usable (single-hop F) by probing the backscatter with a directive antenna and high transmitter power, even when the Earth contact point is open ocean. In fact, that's where the mode is at its best, because ocean waves can be efficient backscatter reflectors.

Backscatter is very useful on 28 MHz, particularly when that band seems dead simply because nobody is active in the right places. The mode keeps the 10-meter band lively in the low years of the solar cycle, thanks to the never-say-die attitude of some users. The mode is also an invaluable tool of 50-MHz DX aspirants, in the high years of the sunspot cycle, for the same reasons. On a high-MUF morning, hundreds of 6-meter beams may zero in on a hot spot somewhere in the Caribbean or South Atlantic, where there is no land, let alone other 6-meter stations—keeping in contact while they wait for the band to open to a place where there is somebody.

Sidescatter is similar to backscatter, except the ground scatter zone is off the direct line between participants. A typical example, often observed during the lowest years of the solar cycle, is communication on 28 MHz between the eastern US (and adjacent areas of Canada) and much of the European continent. Often, this may start as "backscatter chatter" between Europeans whose antennas are turned toward the Azores. Then suddenly the North Americans join the fun, perhaps for only a few minutes, but sometimes much longer, with beams also pointed toward the Azores. Duration of the game can be extended, at times, by careful reorientation of antennas at both ends, as with backscatter. The secret, of course, is to keep hitting the highest-MUF area of the ionosphere and the most favorable ground-reflection points.

The favorable route is usually, but not always, south of the great-circle heading (for stations in the Northern Hemisphere). There can also be sidescatter from the auroral regions. Sidescatter signals are stronger than backscatter signals using the same general area of ground scattering.

Sidescatter signals have been observed frequently on the l4-MHz band, and can take place on any band where there is a large window between the MUF and the LUF. For sidescatter communications to occur, the thing to look for is a common area to which the band is open from both ends of the path (the Azores, in the above example), when there is no direct-path opening. It helps if the common area is in the open ocean, where there is less scattering loss than over land.

GRAY-LINE PROPAGATION

The *gray line*, sometimes called the *twilight zone*, is a band around the Earth between the Sunlit portion and darkness. Astronomers call this the *terminator*. The terminator is a somewhat diffused region because the Earth's atmosphere tends to scatter the light into the darkness. **Fig 40** illustrates the gray line. Notice that on one side of the Earth, the gray line is coming into daylight (sunrise), and on the other side it is coming into darkness (sunset).

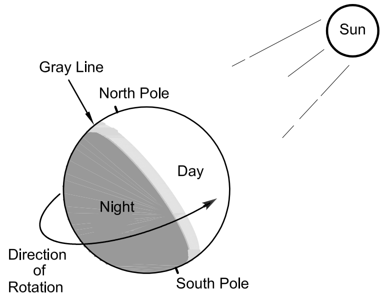

Fig 40—The gray line or terminator is a transition region between daylight and darkness. One side of the Earth is coming into sunrise, and the other is just past sunset.

Propagation along the gray line is very efficient, particularly on the lower bands, especially on 80 or 160 meters, so greater distances can be covered than might be expected for the frequency in use. One major reason for this is that the D layer, which absorbs HF signals, disappears rapidly on the sunset side of the gray line, and has not yet built up on the sunrise side.

The gray line runs generally north and south, but varies as much as 23° either side of the north-south line. This variation is caused by the tilt of the Earth's axis relative to its orbital plane around the Sun. The gray line will be exactly north and south at the equinoxes (March 21 and September 21). On the first day of Northern Hemisphere summer, June 21, it is tilted to the maximum of 23° one way, and on December 21, the first day of winter, it is tilted 23° the other way.

To an observer on the Earth, the direction of the terminator is always at right angles to the direction of the Sun at sunrise or sunset. It is important to note that, except at the equinoxes, the gray-line direction will be different at sunrise from that at sunset. This means you can work different areas of the world in the evening than you worked in the morning.

It isn't necessary to be located inside the twilight zone in order to take advantage of gray-line propagation. The effects can be used to advantage before sunrise and after sunset. This is because the Sun "rises" earlier and "sets" later on the ionospheric layers than it does on the Earth below.

What HF Bands Are Open—Where and When?

The CD-ROM included at the back of this book includes summary and detailed propagation predictions for more than 150 transmitting locations around the world. This propagation data was calculated using *CapMAN*, an upgraded variety of the mainframe propagation program *IONCAP*. The predictions were done for default antennas and powers that are representative of a "big-gun" station. Of course, not everyone has a big-gun station in his/her backyard, but this represents what the ultimate possibilities are, statistically speaking. After all, if the bands aren't open for the big guns, they are unlikely to be open for the "little pistols" too.

Let's see how propagation is affected if the smoothed sunspot number is 0 (corresponding to a smoothed solar flux of about 65), which is classified as a "Very Low" level of solar activity. And we'll examine the situation for a sunspot number of 100 (a smoothed solar flux of 150), which is typical of a "Very High" portion of the solar cycle.

Five-Band Summary Predictions

Tables 4 and **5** are Summary tables showing the predicted signal levels (in S units) from Boston, Massachusetts, to the rest of the world for the month of January. The Boston transmitting site is representative of the entire New England area of the USA. The target geographic receiving regions for the major HF bands from 80 through 10 meters are tabulated versus UTC (Universal Coordinated Time) in hours. Table 4 represents a Very Low level of solar activity, while Table 5 is for a Very High level of solar activity.

Each transmitting location is organized by six levels of solar activity over the whole 11-year solar cycle:

- VL (Very Low: SSN between 0 to 20)
- LO (Low: SSN between 20 to 40)
- ME (Medium: SSN between 40 to 60)
- HI (High: SSN between 60 to 100)
- VH (Very High: SSN between 100 to 150)
- UH (Ultra High: SSN greater than 150)

The receiving geographic regions for each frequency band are abbreviated:

Table 4

Printout of summary propagation table for Boston to the rest of the world, for a Very Low level of solar activity in the month of January. The abbreviations for the target geographic areas are: EU = Europe, FE = Far East, SA = South America, AF = Africa, AS = south Asia, OC = Oceania, and NA = North America.

```
Jan., MA (Boston), for SSN = Very Low, Sigs in S-Units. By N6BV, ARRL.
           80 Meters              40 Meters              20 Meters              15 Meters              10 Meters
UTC  EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA    UTC
 0   9  -  9+ 9  9  -  9+   9  8  9+ 9+ 9  2  9+   -  8  9+ 7  4  8  9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      0
 1   9  -  9+ 9  9  -  9+   9  6  9+ 9+ 9+ 6  9+   -  4  9  4  2  6  9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      1
 2   9  -  9+ 9+ 8  1  9+   9  6  9+ 9+ 9  8  9+   -  1  8  1  2  3  9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      2
 3   9  -  9+ 9+ 8  6  9+   9  6  9+ 9+ 9  8  9+   -  -  8  2  2  -  9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      3
 4   9  -  9+ 9+ 1  8  9+   9  8  9+ 9+ 9  9  9+   -  1  8  7  2  -  9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      4
 5   9  -  9+ 9+ -  9  9+   9  8  9+ 9+ 8  9  9+   -  1  9  8  2  -  9    -  -  -  -  -  -  1    -  -  -  -  -  -  2      5
 6   9+ -  9+ 9+ -  9  9+   7  8  9+ 9+ 8  9  9+   -  1  9+ 8  -  -  9    -  -  -  -  -  -  1    -  -  -  -  -  -  2      6
 7   9  7  9+ 9  -  9  9+   7  8  9+ 9  8  9  9+   -  1  9+ 1  -  1  8    -  -  -  -  -  -  1    -  -  -  -  -  -  2      7
 8   9  8  9+ 9  -  9  9+   8  9  9+ 9  8  9+ 9+   -  1  9+ -  -  5  9    -  -  -  -  -  -  1    -  -  -  -  -  -  2      8
 9   8  8  9+ 7  6  9  9+   8  9  9+ 9  9  9+ 9+   -  -  9  1  -  7  9    -  -  -  -  -  -  1    -  -  -  -  -  -  2      9
10   5  8  9+ 4  6  9  9+   9  9  9+ 9  9  9+ 9+   -  3  9  5  -  6  9    -  -  -  -  -  -  1    -  -  -  -  -  -  2     10
11   3  8  9+ -  5  9  9+   8  9  9+ 7  9  9+ 9+   5  -  9+ 9  5  1* 8    -  -  -  -  -  -  1    -  -  -  -  -  -  2     11
12   1  8  9  -  4  9  9+   7  9  9+ 4  8  9  9+   9  5  9+ 9+ 9  2* 8    -  -  5  6  -  -  1    -  -  -  -  -  -  2     12
13   -  6  1  -  -  7  9+   6  8  9+ 1  8  9  9+   9+ 9  9+ 9  9  7  8    4  -  9+ 9  7  -  1    -  -  -  -  -  -  1     13
14   -  -  -  -  -  1  9+   5  7  8  -  8  8  9+   9+ 9  9+ 9  9  9  9+   7  2* 9+ 9  9  -  8    -  -  5  -  -  -  1     14
15   -  -  -  -  -  -  9+   4  6  5  -  6  7  9+   9+ 9  9+ 9  9  9  9+   7  5  9+ 9  2  2  5    -  -  5  -  -  -  -     15
16   -  -  -  -  -  -  9+   5  6  4  2  5  4  9+   9+ 8  9+ 9+ 9  9  9+   5  1  9+ 8  2* 2  9    -  -  5  -  -  -  1     16
17   -  -  -  -  -  -  9+   6  5  5  5  6  1  9+   9+ 5  9+ 9+ 3  9  9+   -  -  9+ 9  -  3  9+   -  -  5  -  -  -  1     17
18   1  -  -  -  -  -  9+   8  6  6  7  6  -  9+   9+ 6  9+ 9+ 4  9  9+   -  -  9+ 9  -  7  9+   -  -  5  -  -  -  1     18
19   3  -  -  2  -  -  9+   9  7  8  8  8  -  9+   6  6  9+ 9+ 6  9  9+   -  -  9+ 9  -  9  9+   -  -  2  -  -  -  1     19
20   5  -  7  5  -  -  9+   9  8  9+ 9  8  4  9+   1  7  9+ 9+ 8  9  9+   -  -  9+ 4  -  9  9    -  -  -  -  -  -  1     20
21   8  3  9  8  6  -  9+   9  8  9+ 9+ 9  7  9+   -  8  9+ 9  8  9  9+   -  -  9+ -  -  9  6    -  -  -  -  -  -  1     21
22   9  3  9+ 9  8  -  9+   9  8  9+ 9+ 9  5  9+   -  9+ 9+ 9  8  9  9+   -  -  9  -  -  7  1    -  -  -  -  -  -  1     22
23   9  2  9+ 9  9  -  9+   9  8  9+ 9+ 9  4  9+   -  9+ 9+ 9  5  9  9+   -  1  6  -  -  2  3    -  -  -  -  -  -  2     23
     EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA
```

Table 5

Printout of summary propagation table for Boston to the rest of the world, for a Very High level of solar activity in the month of January.

```
Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.
          80 Meters              40 Meters              20 Meters              15 Meters              10 Meters
UTC  EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   UTC
  0  9+ -  9+ 9+ 8  -  9+   9+ 5  9+ 9+ 9  -  9+   1  9+ 9+ 9+ 9+ 9  9+   -  9  9+ 2  2  9+ 9+   -  1  8  -  -  8  9+     0
  1  9+ -  9+ 9+ 8  -  9+   9+ 4  9+ 9+ 9  2  9+   1  9  9+ 8  9+ 9+ 9+   -  3  9  -  7  9+ 9    -  -  -  -  -  4  2      1
  2  9+ -  9+ 9+ 7  -  9+   9+ 4  9+ 9+ 9  7  9+   1  9  9+ 8  9  9+ 9+   -  -  3  -  -  7  9    -  -  -  -  -  -  2      2
  3  9+ -  9+ 9+ 1  2  9+   9+ 4  9+ 9+ 9  9  9+   -  7  9+ 7  8  9+ 9    -  -  -  -  -  -  -    -  -  -  -  -  -  2      3
  4  9+ -  9+ 9+ -  7  9+   9+ 5  9+ 9+ 8  9  9+   -  5  9+ 9  9  9  9+   -  -  1  -  -  -  -    -  -  -  -  -  -  2      4
  5  9+ -  9+ 9+ -  8  9+   9+ 6  9+ 9+ 7  9  9+   -  5  9+ 9  9  5  9+   -  -  -  -  -  -  -    -  -  -  -  -  -  2      5
  6  9+ -  9+ 9+ -  8  9+   9+ 7  9+ 9+ 7  9  9+   -  8  9+ 8  9  5  9+   -  -  -  -  -  -  -    -  -  -  -  -  -  2      6
  7  9+ -  9+ 9+ -  8  9+   9  8  9+ 9+ 7  9+ 9+   -  9  9+ -  7  9  9+   -  -  1  -  -  -  -    -  -  -  -  -  -  2      7
  8  9  7  9+ 9  -  8  9+   9  8  9+ 9+ 8  9+ 9+   -  9  9+ -  4  9+ 9+   -  -  1  -  -  -  2    -  -  -  -  -  -  2      8
  9  8  7  9+ 7  -  8  9+   9  9  9+ 9  8  9+ 9+   -  6  9+ -  1  9+ 9+   -  -  -  -  -  -  1    -  -  -  -  -  -  2      9
 10  5  8  9+ 2  3  8  9+   9  9  9+ 8  8  9  9+   4  -  9+ 9+ 1  5  9    -  -  -  -  -  -  -    -  -  -  -  -  -  2     10
 11  1  8  9+ -  4  9  9+   8  9  9+ 5  8  9  9+   9+ 4* 9+ 9+ 7  -  8    -  -  9  9  -  -  -    -  -  -  -  -  -  2     11
 12  -  7  8  -  1  9  9+   6  9  9+ 1  8  9  9+   9+ 9  9+ 9  9  1* 9+   9  8* 9+ 9+ 9  5* -    -  2* 9  9  1  1* 2     12
 13  -  -  -  -  -  2  9+   4  8  8  -  7  9  9+   9+ 9  9+ 9  9  9+ 9+   9+ 7  9+ 9+ 9+ 3* 9    9  5* 9+ 9+ 9  6* 2     13
 14  -  -  -  -  -  -  9+   2  7  4  -  5  8  9+   9+ 9  9+ 8  9  9  9+   9+ 9  9+ 9+ 9+ 9  9+   9  6* 9+ 9+ 9  1* 1     14
 15  -  -  -  -  -  -  9    1  5  -  -  4  5  9+   9+ 9  9+ 9  9  9  9+   9+ 9+ 9+ 9+ 9+ 9  9+   9  5  9+ 9+ 6  6  8     15
 16  -  -  -  -  -  -  8    3  4  -  -  3  1  9+   9+ 8  9  9  9  9  9+   9+ 9+ 9+ 9+ 9  9+ 9+   9  8  9+ 9+ -  8  9     16
 17  -  -  -  -  -  -  8    5  3  -  2  4  -  9+   9+ 8  9+ 9+ 9  9  9+   9+ 9  9+ 9+ 1* 9+ 9+   -  8  9+ 9+ -  8  9+    17
 18  -  -  -  -  -  -  9    7  4  2  5  5  -  9+   9+ 9  9+ 9+ 9  9  9+   9+ 9  9+ 9+ 1  9+ 9+   -  7  9+ 9+ -  9+ 9+    18
 19  1  -  -  1  -  -  9+   8  5  6  8  7  -  9+   9+ 9  9+ 9+ 9  9  9+   -  9+ 9+ 9+ 2  9  9+   -  6  9+ 9+ -  9+ 9+    19
 20  4  -  2  5  -  -  9+   9  6  9  9  8  -  9+   9+ 9  9+ 9+ 9  9  9+   -  8  9+ 9+ 3  9  9+   -  1  9+ 9  -  9  9+    20
 21  7  -  8  7  1  -  9+   9+ 7  9+ 9+ 8  1  9+   8  9  9+ 9+ 9  9  9+   -  6  9+ 9+ 3  9  9+   -  -  9+ 5* -  9+ 9+    21
 22  9  2  9+ 9  8  -  9+   9+ 7  9+ 9+ 9  4  9+   2  9+ 9+ 9+ 9  9  9+   -  9+ 9+ 9  1  9+ 9+   -  5  9+ 4* -  9  6     22
 23  9  -  9+ 9  8  -  9+   9+ 7  9+ 9+ 9  -  9+   1  9+ 9+ 9+ 9  9  9+   -  9+ 9+ 6  -  9  9+   -  7  9+ 2* -  9  2     23
     EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA   EU FE SA AF AS OC NA
```

EU	All of Europe
FE	The Far East, centered on Japan
SA	South America, centered on Paraguay
AF	All of Africa, centered on Zambia
AS	South Asia, centered on India
OC	Oceania, centered on Sydney, Australia
NA	North America, all across the USA

These propagation files show the highest predicted signal strength (in S-units) throughout the generalized receiving area, for a 1500-W transmitter and rather good antennas on both sides of the circuit. The standard antennas are:

- 100-foot high inverted-V dipoles for 80 and 40 meters
- 3-element Yagi at 100 feet for 20 meters
- 4-element Yagi at 60 feet for 15 and 10 meters.

For example, Summary Table 4 shows that in January during a period of Very Low solar activity, 15 meters is open to somewhere in Europe from Boston for only 4 hours, from 13 to 16 UTC, with a peak signal level between S4 and S7. Now look at Table 5, where 15 meters is predicted to be open to Europe during a period of Very High solar activity for 7 hours, from 12 to 18 UTC, with peak signals ranging from S9 to S9+.

Both Tables 4 and 5 represent *snapshots* of predicted signal levels to generalized receiving locations—that is, they are computed for a particular month, from a particular transmitting location, and for a particular level of solar activity. These tables provide summary information that is particularly valuable for someone planning for an operating event such as a DXpedition or a contest.

What happens if you don't have a big-gun station with high antennas or the 1500-W power assumed in the analyses above? You can discount the S-Meter readings to reflect a smaller station:

- Subtract 2 S units for a dipole instead of a Yagi at same height on 20/15/10 meters.
- Subtract 3 S units for a dipole at 50 feet instead of a Yagi at 100 feet on 20 meters.
- Subtract 1 S unit for a dipole at 50 feet rather than a dipole at 100 feet on 40/80 meters.
- Subtract 3 S units for 100 W rather than 1500 W.
- Subtract 6 S units for 5 W (QRP) rather than 1500 W.

For example, Table 4 predicts an S7 signal into Boston from Europe on 15 meters at 14 UTC. If a European station is using a dipole at 50 feet, with 100 W of power, what

20 Meters: Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.

UTC -->

Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
KL7 = 01	9+	9+	9+	7	-	-	-	-	-	-	-	-	-	-	-	3	9+	9+	9+	9+	9+	9+	9+	9+
VO2 = 02	9+	9	9	9	9	9	8	7	5	3	2	1	5	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9+
W6 = 03	9+	9+	9+	7	7	1	1	5	8	8	3	-	-	1	9	9+	9+	9+	9+	9	9	9+	9+	9+
W0 = 04	9+	9+	9+	8	5	5	5	5	3	2	1	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+
W3 = 05	4	2	2	2	2	2	2	3	3	3	3	2	1	1	8	9+	9+	9+	9+	9+	9+	9+	9+	9
XE1 = 06	9+	9+	7	9	9+	9+	9+	9+	9+	9+	9	8	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+
TI = 07	9+	9+	8	9	9	9	9	9	9	9+	9	9+	9+	9+	9+	9+	9	8	9	9	9+	9+	9+	9+
VP2 = 08	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9	9+	9+	9+	9+	9+	9	9+	9+	9+	9+	9+	9+	9+
P4 = 09	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+	9+
HC = 10	9+	8	9+	9	9	9	9	9	7	3	1	7	9+	9+	9	5	5	5	7	8	9+	9+	9+	9+
PY1 = 11	9+	9+	9+	9	9	9+	9+	9	8	6	9	9+	8	2	1	-	-	1	4	8	9	9+	9+	9+
CE = 12	9+	9+	9+	9+	9+	9+	9+	9+	9	8	8	9+	9	8	2	1	1	-	1	3	7	9	9+	9+
LU = 13	9+	9+	9+	9+	9+	9+	9+	9+	8	8	8	9+	8	4	2	1	-	-	1	4	8	9	9+	9+
G = 14	-	-	-	-	-	-	-	-	-	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	2	-
I = 15	-	-	-	-	-	-	-	-	-	-	4	9	9	9	9	9	9	9+	9+	9	8	2	-	-
UA3 = 16	1	1	1	-	-	-	-	-	-	-	-	8	9	9+	9+	9+	9	8	5	-	-	-	-	1
UN = 17	1	-	-	8	7	7	7	1	-	-	-	2	9	9	9	6	-	-	2	4	8	9	5	4
UA9 = 18	6	7	6	6	9	9	9	7	4	1	-	-	8	8	6	6	5	6	7	8	9	9	8	7
UA0 = 19	9+	9	9	6	5	5	8	8	8	4	-	-	2	6	8	8	8	7	4	4	7	9	9+	9+
4X = 20	8	6	3	1	-	3	4	-	-	-	1	8	8	8	8	8	9	9	9+	9	9	8	7	7
HZ = 21	9+	9	4	3	8	8	2	-	-	-	1	7	8	9	8	8	9	9	9	9	9	9	9	9
VU = 22	7	5	8	7	6	7	5	-	-	-	-	6	9	9	9	9	3	2	2	2	8	8	9	8
JT = 23	9	9+	9	5	7	8	8	6	3	-	-	2*	8	8	5	6	8	8	8	8	9	7	5	6
VS6 = 24	9	9	9	5	4	5	7	8	6	1	-	1*	5	7	1	1	1	1	4	2	-	-	-	9
JA1 = 25	9	9	8	7	5	5	8	9	9	6	-	1	1	2	7	7	6	2	-	-	7	9	9+	9
HS = 26	9	9	6	4	2	-	-	2	1	-	-	2*	9	9	9	9	8	7	5	4	5	-	1*	1
DU = 27	9	8	7	-	-	-	5	7	7	1	-	-	1*	9	9	7	6	4	5	3	1*	1*	8	9
YB = 28	9	8	1	-	-	-	-	-	-	-	-	4*	8	9	9	9	8	8	9	9	9	9	9+	9+
VK6 = 29	3*	4*	-	-	-	-	-	-	5	3	-	-	-	5	9	9	9	8	9	9	9	9	9	8
VK3 = 30	1*	-	-	-	-	-	1	3	9	9	4	-	-	9+	9	8	2	1	-	-	1	2*	5*	4*
KH6 = 31	9	9+	9+	9+	8	2	2	6	4	-	-	-	-	-	-	-	9	9	8	7	6	4	6	7
KH8 = 32	-	2	9	9	9	5	5	9	9+	9+	5	-	-	9+	9	9	8	5	3	1	-	-	-	-
CN = 33	-	-	-	-	-	-	-	-	-	-	9	9+	9	9	8	9	9	9+	9+	9+	9+	9+	9+	7
SU = 34	9	8	3	3	-	1	4	-	-	-	2	7	8	8	8	8	9	9	9+	9+	9+	9+	8	8
6W = 35	9+	8	-	-	2	7	5	-	-	-	9+	9+	8	5	4	3	7	9	9+	9+	9+	9+	9+	9+
D2 = 36	9+	9+	5	3	9	9	8	-	-	-	3	-	-	-	-	4	4	7	8	9	9+	9+	9+	9+
5Z = 37	9+	9	2	4	8	8	1	-	-	-	2	-	-	3	5	5	7	8	9	9	9+	9+	9+	9+
ZS6 = 38	9+	9+	8	7	8	9	6	-	-	-	-	-	-	-	1*	1	2	6	8	9	9+	9+	9+	9+
FR = 39	9+	8	2	1	4	1	-	-	-	-	-	-	-	2*	3*	1*	1	3	8	9	9+	9+	9+	9+
FJL = 40	9+	9+	7	4	7	8	7	1	-	-	-	1*	8	9	9	9	9	9	9	9	9+	9+	9+	9+
Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23

UTC -->

Expected signal levels using 1500 W and 3-element Yagis at 100 feet at each station.

Fig 41—The 20-meter page from Detailed propagation-prediction for the month of January, during Very High solar conditions, from Boston to 40 CQ Zones throughout the world. There are similar pages for each month/SSN level for 160, 80, 40, 20, 15 and 10 meters. These Detailed tables are very useful for planning DX work.

would this do to the predicted signal level in Boston? You would compute: S7 – 2 S units (for a dipole instead of Yagi) – 3 S units (100 W rather than 1500 W) = an S2 signal in Boston. A QRP station with a 4-element 15-meter Yagi at 60 feet would yield: S7 – 6 S units = an S1 signal in Boston.

More Detailed Predictions

Let's now look at table in **Fig 41**, which is the Detailed 20-meter page for the same conditions in Table 5: January at a Very High level of solar activity from Boston to the world. There are six such pages per month/SSN level, covering 160, 80, 40, 20, 15 and 10 meters.

In a Detailed prediction table, the world is divided into the 40 CQ Zones, with a particular sample location in each zone. For example, Zone 14 in Western Europe is represented by a location in London, England (call sign G), while Zone 25 is represented by a location in Tokyo, Japan (call sign JA1). Note that Zones with large ham populations are highlighted with dark shadowing for easy identification. For example, Zones 3, 4 and 5 cover the USA, while Zones 14, 15 and 16 cover the majority of Europe. Zone 25 covers the big ham population in Japan.

Let's revisit the example above for computing the signal strength for a station in London, but this time on 20 meters. Again, we'll assume that the G station has a dipole at 50 feet and 100 W of transmitter power. At 14 UTC in Zone 14, the table in Fig 40 predicts a very healthy signal for the reference big-gun station, at S9+. This is a signal at least S9 + 10 dB. Here, we're going to round off the plus 10 dB to 2 S units, giving a fictional 11 S units to start. We discount this for the smaller station: S11 – 3 S units (for a dipole at 50 feet instead of a 3-element Yagi at 100 feet) – 3 S units (100 W rather than 1500 W) = S5 signal in Boston. This is a respectable signal and will probably get through, in the absence of stronger signals calling the Boston station at the same time, of course.

Here's another example of how to use the Detailed propagation-prediction tables. Let's say that at 1230 UTC in January you work a VU2 station in New Delhi on 15 meters from Boston, where the local time is 7:30 AM.

You need a 20-meter contact also for the 5-Band DXCC award, so you quickly check the table in Fig 40 for Zone 22 (VU) and find that the predicted signal strength is S9. Your new VU2 friend is willing to jump to 20 meters and so you QSY to make the contact.

But perhaps you are late leaving for work and so you ask your new VU2 friend to make a schedule with you later that evening. Again, you consult the Detailed prediction table for 20 meters and find that signals are predicted to be S8 or stronger from 20 to 23 UTC, dropping to S7 at 00 UTC. You quickly ask your new friend whether he minds waking up at 4:30 AM his time to make a schedule with you at 2300 UTC, because New Delhi is 5½ hours ahead of UTC. You determined this using the program *GeoClock*, which is included with the software on the CD-ROM in the back of this book and which you run in the background on *Windows*. Luckily, he's a very gracious fellow and agrees to meet you on a specific frequency at that time.

The Detailed propagation-prediction tables give you all the information needed to plan your operations to maximize your enjoyment chasing DX. You can use these tables to plan a 48-hour contest next month, or next year—or you can use them to plan a schedule with your ham cousin on the West Coast on Saturday afternoon.

THE PROPAGATION BIG PICTURE

A newcomer to the HF bands could easily be overwhelmed with the sheer amount of data available in the Summary (and particularly the Detailed) prediction tables on the CD-ROM included with this book. So here's a long-term, "big-picture" view of HF propagation that might help answer some common questions. For example, what month really is the best for working DX around the clock? Or what level of solar activity is necessary to provide an opening between your QTH and somewhere in the South Pacific?

Table 6 is a table showing the number of hours in a day during each month when each major HF band is open to the same receiving areas shown in Tables 4 and 5. The listing is for New England, for three levels of solar activity: Very Low, Medium and Very High. The number of hours are separated in Table 6 by slashes. (Versions of Table 6 for other areas around the US are on the CD-ROM that accompanies this book in **Fig6Tab.PDF**.)

Let's examine the conditions for New England to Europe on 15 meters for October. The entry shows "7/11/17," meaning that for a Very Low level of solar activity, 15 meters is open for 7 hours; for a Medium level, it is open for 11 hours and for a Very High level of solar activity it is open for 17 hours a day.

Even for a Very Low level of solar activity, the month with the most hours available per day from Boston to somewhere in Europe is October, with 7 hours, followed by the next largest month of March, with 6 hours. For a Very High level of solar activity, however, the 15-meter band is open to Europe for 18 hours in April, followed by 17 hours availability in September and October. Arguably, the CQ World Wide Contest Committee picked the very best month for higher-frequency propagation when they chose October for the Phone portion of that contest.

You can easily see that even at a Very High level of solar activity, the summer months are not very good to work DX, particularly on east-west paths. For example, the 10-meter band is very rarely open from New England to Europe after the month of April, even when solar activity is at the highest levels possible. Things pick up after September, even for a Medium level of solar activity. Again, October looks like the most fruitful month in terms of the number of hours 10 meters is open to Europe under all levels of solar conditions.

Ten meters is open more regularly on north-south paths, such as from New England to South America or to southern Africa. It is open as much as 10 hours a day during March and October to deep South America, and 7 hours a day in October to Africa—even during the lowest parts of the solar cycle. (Together with the sporadic-E propagation that 10 meters enjoys during the summer, this band can often be a lot of fun even during the sunspot doldrums. You just have to *be operating* on the band, rather than avoiding it because you know the sunspots are "spotty!")

Now, look at the 20-meter band in Table 6. From New England, twenty is open to somewhere in South America for 24 hours a day, no matter the level of solar activity. Note that Table 6 doesn't predict the level of signals available; it just shows that the band is open with a signal strength greater than 0 on the S meter. Look back at Summary Table 4 for the predicted signal strengths in January at a Very Low level of solar activity. There, you can see that the signal strength from New England into deep South America is always S8 or greater for a big gun station. A lot of the time during the night the band sounds dead, simply because everyone is either asleep or operating on a lower frequency.

For the 40-meter band in Table 6, during the month of January the band is open to Europe for 24 hours a day, whatever the level of solar activity is. Look now at Table 4, and you'll see that the predicted level for Very Low solar activity varies from S4 to S9. Local QRM or QRN would probably disrupt communications on 40 meters in Europe for stateside signals weaker than perhaps S3 or S4. Even though you might well be able to hear Europeans from New England during the day, they probably won't hear you because of local conditions, including local S9+ European stations and atmospheric noise from nearby thunderstorms. New England stations with big antennas can often hear Europeans on 40 meters as early as noontime, but must wait until the late afternoon before the Europeans can hear them above their local noise and QRM.

Let's say that you want to boost your country total on 80 meters by concentrating on stations in the South Pacific. The best months would be from November to

Table 6

The number of hours per day when a particular band is open to the target geographic areas in Table 4, as related to the level of solar activity (Very Low, Medium and Very High). This table is customized for Boston to the rest of the world. Some paths are open 24 hours a day, plus or minus QRM and local QRN, no matter what the level of solar activity is. See CD-ROM for other transmitting locations.

```
MA (Boston)
Hours Open to Each Region for Very-Low/Medium/Very-High SSNs

80 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan   17/17/16   5/ 4/ 3  17/17/16  16/16/15   8/ 7/ 5  11/10/ 9  24/24/24
Feb   17/16/15   3/ 3/ 2  17/16/16  15/15/14   6/ 4/ 4  10/ 9/ 9  24/24/24
Mar   15/15/14   3/ 2/ 1  16/16/15  15/13/13   4/ 4/ 3   9/ 8/ 7  24/24/24
Apr   13/13/12   1/ 0/ 0  16/16/14  13/13/13   3/ 3/ 1   9/ 8/ 7  24/24/24
May   12/11/10   0/ 0/ 0  16/15/14  12/11/10   2/ 1/ 1   7/ 6/ 6  24/24/24
Jun   10/ 9/ 8   0/ 0/ 0  14/14/14  11/10/10   1/ 1/ 0   6/ 5/ 5  24/24/24
Jul   11/11/ 9   0/ 0/ 0  15/14/14  11/11/11   2/ 1/ 1   7/ 6/ 5  24/24/24
Aug   13/11/11   0/ 0/ 0  16/16/14  13/12/11   3/ 2/ 1   7/ 7/ 6  24/24/24
Sep   14/13/11   2/ 1/ 0  17/16/14  13/13/12   4/ 4/ 2   9/ 8/ 8  24/24/24
Oct   15/15/13   3/ 2/ 1  17/17/16  14/14/13   5/ 4/ 4   9/ 9/ 7  24/24/24
Nov   17/17/15   4/ 4/ 2  17/17/16  16/15/14   8/ 7/ 4  11/10/ 9  24/24/24
Dec   19/18/17   7/ 6/ 4  18/18/17  16/16/16  11/ 9/ 7  12/11/11  24/24/24

40 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan   24/24/24  15/16/15  24/24/21  21/20/19  21/21/19  19/18/15  24/24/24
Feb   24/24/21  13/11/11  24/23/20  20/19/18  19/19/17  16/15/14  24/24/24
Mar   23/22/19  10/ 9/ 7  24/21/18  19/17/17  17/17/13  13/13/13  24/24/24
Apr   21/19/18   8/ 6/ 4  22/20/18  17/16/15  16/11/ 8  13/13/11  24/24/24
May   19/17/17   5/ 4/ 3  22/18/17  17/16/14   9/ 8/ 5  12/11/10  24/24/24
Jun   17/15/13   4/ 2/ 2  22/18/16  16/15/14   7/ 5/ 5  11/10/ 9  24/24/24
Jul   18/16/15   5/ 4/ 2  24/18/17  17/15/14   8/ 7/ 5  12/11/10  24/24/24
Aug   19/17/16   7/ 5/ 4  24/19/18  18/16/15  11/10/ 6  13/12/11  24/24/24
Sep   22/21/17   9/ 8/ 5  23/20/18  18/17/16  14/11/ 7  13/13/12  24/24/24
Oct   24/23/20  12/11/ 8  24/23/19  20/18/17  17/16/14  16/13/13  24/24/24
Nov   24/24/22  14/13/12  24/24/20  21/19/18  21/20/17  17/17/13  24/24/24
Dec   24/24/24  18/19/22  24/24/21  23/21/19  24/23/22  21/19/18  24/24/24
```

February in terms of the number of hours per day when the 80-meter band is open to Oceania. You can see by reading across the line for each month that the level of solar activity is not hugely important on 80 meters to any location. Common experience (backed by the statistical information in Table 6) is that the 80-meter band is open only marginally longer when sunspots are low.

This is true to a greater extent on 40 meters. Thus you may hear the generalization that the low bands tend to be better during periods of low solar activity, while the upper HF bands (above 10 MHz) tend to be better when the sun is more active.

Table 6 can give you a good handle on what months are the most productive for DXing and contesting. It should be no surprise to most veteran operators that the fall and winter months are the best times to work DX.

```
20 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan   13/16/22  15/22/22  24/24/24  20/21/21  18/20/22  18/23/22  24/24/24
Feb   12/18/23  13/21/24  24/24/24  22/22/24  15/21/24  18/23/24  24/24/24
Mar   15/18/24  17/20/24  24/24/24  22/24/24  18/21/24  16/24/24  24/24/24
Apr   15/20/24  19/22/24  24/24/24  21/24/24  19/22/24  18/24/24  24/24/24
May   19/23/24  22/24/24  24/24/24  23/24/24  23/24/24  21/24/24  24/24/24
Jun   22/24/24  24/24/24  24/24/24  24/24/24  24/24/24  24/24/24  24/24/24
Jul   19/24/24  24/24/24  24/24/24  21/24/24  24/24/24  23/24/24  24/24/24
Aug   15/20/24  20/24/24  24/24/24  20/24/24  20/24/24  19/24/24  24/24/24
Sep   16/19/24  17/21/24  24/24/24  21/24/24  18/21/24  17/24/24  24/24/24
Oct   15/21/24  16/20/24  24/24/24  22/24/24  19/22/24  17/24/24  24/24/24
Nov   14/20/23  14/22/24  24/24/24  20/24/24  17/21/24  19/23/24  24/24/24
Dec   11/17/24  13/22/24  24/24/24  17/23/24  12/22/24  16/24/24  24/24/24

15 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan    4/ 6/ 7   2/ 9/13  12/15/16   9/13/13   3/ 4/ 7   9/12/13  24/15/16
Feb    4/ 7/12   4/10/14  13/18/23  11/13/16   3/ 7/13   8/13/15  22/16/19
Mar    6/ 9/14   2/13/15  14/21/24  13/17/22   5/11/17  10/14/17  15/16/23
Apr    0/10/18   3/13/18  15/23/24  15/18/24   9/15/19  11/15/21  16/16/24
May    1/13/16   6/10/19  17/20/24  14/18/24  13/17/18  10/16/19  20/19/24
Jun    0/ 2/16   0/ 9/15  16/21/24  14/18/24   5/15/18  10/12/20  24/22/22
Jul    0/ 2/16   0/ 5/18  15/19/24  12/18/24   0/12/18   4/12/20  24/22/21
Aug    0/ 2/14   0/ 8/17  14/18/22  13/16/22   0/12/17   6/10/19  22/19/21
Sep    1/10/17   6/13/17  14/16/24  13/17/22   9/14/17   9/14/17  16/16/22
Oct    7/11/17  10/13/17  12/16/22  12/15/22   7/12/17  12/13/15  18/15/22
Nov    5/ 8/14   8/11/14  12/16/22  11/14/17   3/ 7/16  10/13/15  20/16/21
Dec    3/ 6/ 9   2/10/13  12/15/23   8/13/15   2/ 4/12   9/12/14  24/15/18

10 Meters:
Month  Europe   Far East  So. Amer.  Africa   So. Asia  Oceania   No. Amer.
Jan    0/ 1/ 4   0/ 1/ 8   6/11/13   0/ 7/10   0/ 1/ 3   0/ 3/11  23/24/24
Feb    0/ 2/ 7   0/ 2/10   8/12/14   0/ 9/13   0/ 3/ 5   0/ 7/13  24/24/24
Mar    0/ 0/ 8   0/ 1/10  10/14/20   1/11/14   0/ 0/ 8   0/ 7/13  23/24/24
Apr    0/ 0/ 8   0/ 0/ 8   7/14/21   0/12/17   0/ 0/13   0/ 5/11  18/24/24
May    0/ 0/ 0   0/ 0/ 1   7/12/20   1/10/17   0/ 1/12   0/ 2/11  17/20/22
Jun    0/ 0/ 0   0/ 0/ 0   7/11/18   0/ 3/17   0/ 0/ 0   0/ 0/ 2  21/19/23
Jul    0/ 0/ 0   0/ 0/ 0   2/ 9/19   0/ 2/18   0/ 0/ 7   0/ 0/ 6  16/16/24
Aug    0/ 0/ 0   0/ 0/ 0   2/10/17   0/ 1/16   0/ 0/10   0/ 0/ 8  17/17/24
Sep    0/ 0/ 8   0/ 1/10   7/13/18   0/11/16   0/ 0/10   0/ 2/ 9  19/24/24
Oct    0/ 5/ 9   0/ 2/11  10/12/16   7/12/14   0/ 5/ 9   0/ 8/12  24/24/24
Nov    0/ 4/ 8   0/ 3/11   9/12/15   5/10/13   0/ 3/ 6   4/10/12  24/24/24
Dec    0/ 3/ 6   0/ 1/ 8   8/11/13   1/ 8/12   0/ 1/ 4   2/ 7/12  23/23/24
```

Do-It-Yourself Propagation Prediction

Very reliable methods of determining the MUF for any given radio path have been developed over the last 50 years. As discussed previously, these methods are all based on the smoothed sunspot number (SSN) as the measure of solar activity. It is for this reason that smoothed sunspot numbers hold so much meaning for radio amateurs and others concerned with radio-wave propagation—they are the link to past (and future) propagation conditions.

Early on, the prediction of propagation conditions required tedious work with numerous graphs, along with charts of frequency contours overlaid, or overprinted, on world maps. The basic materials were available from an agency of the US government. Monthly publications provided the frequency-contour data a few months in advance. Only rarely did amateurs try their hand at predicting propagation conditions using these hard-to-use methods.

Today's powerful PCs have given the amateur wonderful tools to make quick-and-easy HF propagation predictions, whether for a contest or a DXpedition. The summary and detailed prediction tables described earlier in this chapter were generated using *CAPMan*, a modernized version of the mainframe *IONCAP* program, on a PC.

While tremendously useful to setting up schedules and for planning strategy for contests, both the Summary and Detailed prediction tables located on the CD-ROM accompanying this book show signal strength. They do not show

Table 7
Features and Attributes of Propagation Prediction Programs

	ASAPS V. 4	*CAPMan*	*VOACAP Windows*	*ACE-HF*	*W6ELProp V. 2.70*	*WinCAP Wizard 2*	*PropLab Pro*
User Friendliness	Good	Good	Good	Excellent	Good	Good	Poor
Operating System	Windows	DOS	Windows	Windows	Windows	Windows	DOS
Uses k or A index	No	Yes	No	No	Yes	Yes	Yes
User library of QTHs	Yes	Yes	Yes	Yes-RX	Yes	Yes	No
Bearings, distances	Yes	Yes	Yes	Yes	Yes	Yes	Yes
MUF calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LUF calculation	Yes	Yes	Yes	Yes	No	Yes	Yes
Wave angle calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Vary minimum wave angle	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Path regions and hops	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Multipath effects	Yes	Yes	Yes	Yes	No	Yes	Yes
Path probability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Signal strengths	Yes	Yes	Yes	Yes	Yes	Yes	Yes
S/N ratios	Yes	Yes	Yes	Yes	No	Yes	Yes
Long-path calculation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Antenna selection	Yes	Yes	Yes	Yes	Indirectly	Isotropic	Yes
Vary antenna height	Yes	Yes	Yes	Yes	Indirectly	No	Yes
Vary ground characteristics	Yes	Yes	Yes	Yes	No	No	No
Vary transmit power	Yes	Yes	Yes	Yes	Indirectly	Yes	Yes
Graphic displays	Yes	Yes	Yes	Yes	Yes	Yes	2D/3D
UT-day graphs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Area Mapping	No	Yes	Yes	Yes	Yes	No	Yes
Documentation	Yes	Yes	On-line	Yes	Yes	Yes	Yes
Price class	$275+	$89	free†	$99	free§	$29.95+	$150††

Price classes are for late 2003 and subject to change.
†Available on the World Wide Web: **elbert.its.bldrdoc.gov/hf.html**
§Available on the World Wide Web at: **www.qsl.net/w6elprop/**
+ Shipping and handling extra.
††Available on the World Wide Web: **www.spacew.com/www/proplab.html**

other information that is also in the underlying databases used to generate them. They don't, for example, show the dominant elevation angles and neither do they show reliability statistics. You may want to run propagation-prediction software yourself to get into the really "nitty-gritty" details.

Modern programs are designed for quick-and-easy predictions of propagation parameters. See **Table 7** for a listing of a number of popular programs. The basic input information required is the smoothed sunspot number (SSN) or smoothed solar flux, the date (month and day), and the latitudes and longitudes at the two ends of the radio path. The latitude and longitude, of course, are used to determine the great-circle radio path. Most commercial programs tailored for ham use allow you to specify locations by the call sign. The date is used to determine the latitude of the Sun, and this, with the sunspot number, is used to determine the properties of the ionosphere at critical points on the path.

Of course, just because a computer program predicts that a band will be open on a particular path, it doesn't follow that the Sun and the ionosphere will always cooperate! A sudden solar flare can result in a major geomagnetic storm, taking out HF communication anywhere from hours to days. There is still art, as well as a lot of science, in predicting propagation. In times of quiet geomagnetic activity, however, the prediction programs are good at forecasting band openings and closings.

Obtaining Sunspot Number/Solar Flux Data

After you have chosen and then set up a computer program for evaluation of a particular path, you will still need the sunspot number or solar flux level for the period in question. A caution must be stated here—for best accuracy and consistency, use the average of solar flux values taken from actual observations, perhaps from WWV/WWVH, over the previous three or four days. Many amateur packet systems archive WWV flux numbers (plus planetary k_p and A_p indices). Solar flux numbers can vary dramatically from day to day, but the Earth's ionosphere is relatively slow to respond to instantaneous changes in solar radiation. This caveat also

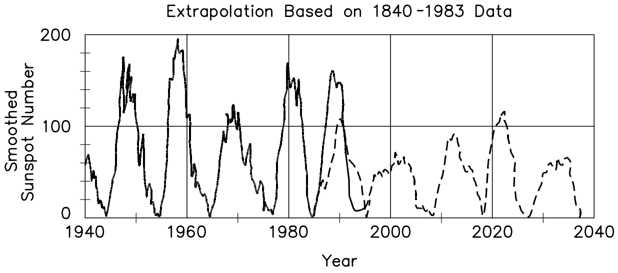

Fig 42—Smoothed sunspot number, with predictions, from 1940 to 2040. This was extrapolated based on data from 1840 to 1983. Cycle 22 actually peaked in Nov 1989, at a monthly smoothed sunspot number of 158. Propagation on the higher frequencies throughout the peak of Cycle 22 was good to excellent, since the monthly smoothed sunspot number stayed at 100 or above from July 1988 through May 1992. (*Courtesy of Naval Ocean Systems Center,* San Diego.*)*

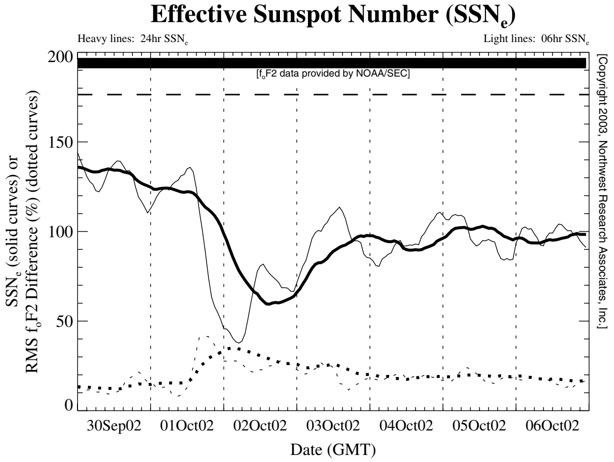

Fig 43—Effective Sunspot Number (SSN_e) produced by NWRA. Note large drop in effective SSN due to a geomagnetic storm commencing Oct 1, 2002. (*Courtesy of Northwest Research Associates.*)

holds for sunspot numbers derived, using Fig 18, from WWV/WWVH solar flux numbers.

Fig 42 shows a graph produced in the early 1980s of smoothed sunspot numbers for Solar Cycles 17 through 21, with predictions for Cycles 22 through 26. The graph covers a period of 100 years, from 1940 to 2040, and may be used for making long-term or historical calculations. Just remember that the graph shows smoothed numbers. The solar activity at any given time can be significantly lower or significantly higher than the graph indicates. In fact, Cycle 22 peaked at the end of 1989, as predicted, but with a monthly smoothed sunspot level of 158, quite a bit higher than predicted. Cycle 23 peaked in early 2002 at a level of about 115, a considerably higher level than the predicted value.

WWV PROPAGATION DATA

For the most current data on what the Sun is doing, National Institute of Standards and Technology stations WWV and WWVH broadcast information on solar activity at 18 and 45 minutes past each hour, respectively. These propagation bulletins give the solar flux, geomagnetic A-Index, Boulder K-Index, and a brief statement of solar and geomagnetic activity in the past and coming 24-hour periods, in that order. The solar flux and A-Index are changed daily with the 2118 UT bulletin, the rest every three hours—0018, 0318, 0618 UT and so on. On the Web, up-to-date WWV information can be found at: **ftp://ftp.sel.noaa.gov/pub/latest/wwv.txt** or on the NOAA Web page **www.sec.noaa.gov/**.

Some other useful Web sites are: **dx.qsl.net/propagation/**, **www.dxlc.com/solar**, **hfradio.org/propagation.html**. The Solar Terrestrial Dispatch page contains a wealth of propagation-related information: **www.spacew.com/**. You may also access propagation information on your local PacketCluster. Use the command SH/WWV/*n*, where *n* is the number of spots you wish to see (five is the default).

Another excellent method for obtaining an "equivalent sunspot number" (SSN_e) is to go to the Space Weather site of Northwest Research Services: **www.nwra-az.com/spawx/ssne24.html**. NWRA compares real-time ionospheric sounder data around the world with predictions using various levels of SSN looking for the best match. They thus "back into" the actual effective sunspot number. **Fig 43** is a typical NWRA graph, which covers the week ending 6 October 2002. Note the sudden decrease in SSN_e after a geomagnetic storm depressed SSN_e by more than 50%.

The A-Index

The WWV/WWVH A-Index is a daily figure for the state of activity of the Earth's magnetic field. It is updated with the 2118/2145 UT bulletin. The A-Index tells you mainly how yesterday was, but it is very revealing when charted regularly, because geomagnetic disturbances nearly always recur at four-week intervals.

The K-Index

The K-Index (new every three hours) reflects Boulder readings of the Earth's geomagnetic field in the hours just preceding the bulletin data changes. It is the nearest thing to current data on radio propagation available. With new data every three hours, K-Index trend is important. Rising is bad news; falling is good, especially related to propagation on paths involving latitudes above 30° north. Because this is a Boulder, Colorado, reading of geomagnetic activity, it may not correlate closely with conditions in other areas.

The K-Index is also a timely clue to aurora possibilities. Values of 4, and rising, warn that conditions associated with

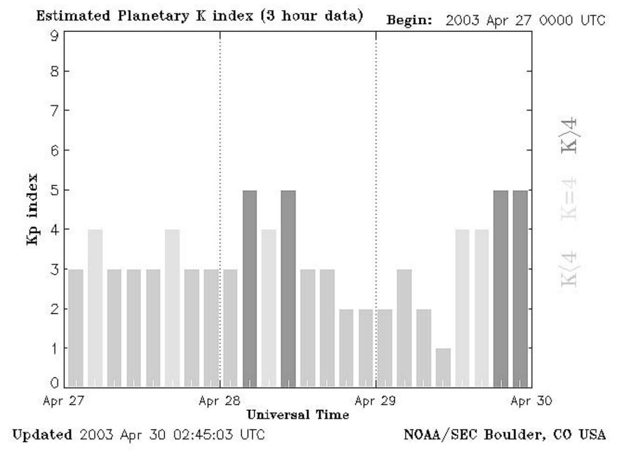

Fig 44—Plot of the estimated planetary k index, K_p, for the last four days. (*Courtesy NOAA/SEC.*)

auroras and degraded HF propagation are present in the Boulder area at the time of the bulletin's preparation. A NOAA Web site that carries up-to-date planetary K_p data is: **www.sel.noaa.gov/ftpmenu/plots/2003_plots/kp.html**. **Fig 44** is a graph from this Web site for four days starting 27 April 2003 to 30 April 2003. This was a period of significant geomagnetic activity, indicated by K_p indices of 5 and 4.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

E. V. Appleton and W.R. Piggott, "Ionospheric Absorption Measurements during a Sunspot Cycle," *J. Atmos. Terr. Phys.*, Vol 3, p 141, 1954.

D. Bilitiza, *International Reference Ionosphere (IRI 90)*, National Space Science Data Center, Greenbelt, MD, 1990.

D. Bray, "Method of Determining VHF/HF Station Capabilities," *QST*, Nov 1961.

A. Brekke, "Physics of the Upper Polar Atmosphere," (John Wiley and Sons: New York, 1997).

R. R. Brown, "Demography, DXpeditions and Magneto-Ionic Theory," *The DX Magazine*, Vol. X, No. 2, p 44, Mar/Apr 1998.

R. R. Brown, *The Little Pistol's Guide to HF Communication* (Sacramento: WorldRadio Books, 1996).

R. R. Brown, "Signal Ducting on the 160 Meter Band," *Communications Quarterly*, p 65, Spring 1998.

R. R. Brown, "Unusual Low-Frequency Signal Propagation at Sunrise," *Communications Quarterly*, p 67, Fall 1998.

R. R. Brown,, "Atmospheric Ozone, a Meteorological Factor in Low-Frequency and 160 Meter Propagation," *Communications Quarterly*, Spring 1999.

K. Davies, *Ionospheric Radio* (London: Peter Peregrinus Ltd, 1990). Excellent technical reference.

R. Garcia, S. Solomon, S. Avery, G. C. Reid, "Transport of Nitric Oxide and the D-Region Winter Anomaly," *J. Geophys. Res.*, Vol 92, p 977, 1987.

J. Hall, "Propagation Predictions and Personal Computers," Technical Correspondence, *QST*, Dec 1990, pp 58-59 (description of *IONCAP* as used for ARRL. publications).

E . Harper, *Rhombic Antenna Design* (New York: D. Van Nostrand Co, Inc, 1941).

H. Hertz, *Electric Waves*, translated by D. E. Jones (London: MacMillan, 1893).

R. D. Hunsucker, *Radio Techniques for Probing the Terrestrial Ionosphere* (New York: Springer-Verlag).

R. D. Hunsucker, J. K. Hargreaves, *The High Latitude Ionosphere and Its Effects on Radio Propagation* (Cambridge: Cambridge University Press, 2003).

W. D. Johnston, Computer-calculated and computer-drawn great-circle maps are offered. An 11 × 14-inch map is custom made for your location. Write to K5ZI, PO Box 640, Organ, NM 88052, tele 505-382-7804.

T.L. Killeen, R.M. Johnsson, "Upper Atmospheric Waves, Turbulence, and Winds: Importance for Mesospheric and Thermospheric Studies," **earth.agu.org/revgeophys/killee00/killee00.html**.

J. L. Lynch, "The Maunder Minimum," *QST*, Jul 1976, pp 24-26.

J. C. Maxwell, *A Treatise on Electricity and Magnetism*, Vols I and II (Oxford: Oxford University Press, 1873).

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington, CT: ARRL, 1990) [out of print].

M. W. Maxwell, *Reflections II—Transmission Lines and Antennas* (Sacramento, CA: 2001).

L. F. McNamara, *The Ionosphere: Communications, Surveillance, and Direction Finding* (Malabar, FL: Krieger Publishing Company, 1991). Another excellent technical reference on propagation.

L. F. McNamara, *Radio Amateur's Guide to the Ionosphere* (Malabar, FL: Krieger Publishing Company, 1994). Excellent, quite-readable text on HF propagation.

A. K. Paul, "Medium Scale Structure of the F Region," *Radio Science*, Volume 24, No. 3, p. 301, 1989.

W. R. Piggott, K. Rawer, "URSI Handbook of Ionogram Interpretation and Reduction," Report UAG-50. World Data Center A for Solar-Terrestrial Physics, Boulder, CO, 1975.

E. Pocock, "Sporadic-E Propagation at VHF: A Review of Progress and Prospects," *QST*, Apr 1988, pp 33-39.

E. Pocock, "Auroral-E Propagation at 144 MHz," *QST*, Dec 1989, pp 28-32.

E. Pocock, "Propagation Forecasting During Solar Cycle 22," *QST*, Jun 1989, pp 18-20.

G. C. Reid, "Ion Chemistry of the D-region," *Advances in Atomic and Molecular Physics*, Vol 12, Academic Press, 1976.

R. B. Rose, "MINIMUF: A Simplified MUF-Prediction Program for Microcomputers," *QST*, Dec 1982, pp 36-38

M. L. Salby, "Fundamentals of Atmospheric Physics," (Academic Press: Boulder, CO, 1996).

S. C. Shallon, W6EL: *MINIPROP PLUS*, a commercially prepared program written for Amateur Radio users; 11058 Queensland St, Los Angeles, CA 90034-3029.

R. D. Straw, *All the Right Angles* (New Bedford, PA: LTA, 1993). Out of print.

R. D. Straw, "*ASAPS* and *CAPMAN*: HF Propagation-Prediction Software for the IBM PC," *QST*, Dec 1994, pp 79-81.

R. D. Straw, "Heavy-Duty HF Propagation-Prediction/Analysis Software," Part 1: *QST*, Sep 1996, pp 28-32; Part 2: *QST*, Oct 1996, pp 28-30.

Chapter 27

Antenna and Transmission-Line Measurements

The principal quantities measured on transmission lines are line current or voltage, and standing-wave ratio (SWR). You make measurements of current or voltage to determine the power input to the line. SWR measurements are useful in connection with the design of coupling circuits and the adjustment of the match between the antenna and transmission line, as well as in the adjustment of these matching circuits.

For most practical purposes a relative measurement is sufficient. An uncalibrated indicator that shows when the largest possible amount of power is being put into the line is just as useful, in most cases, as an instrument that measures the power accurately. It is seldom necessary to know the actual number of watts going into the line unless the overall efficiency of the system is being investigated. An instrument that shows when the SWR is close to 1:1 is all you need for most impedance-matching adjustments. Accurate measurement of SWR is necessary only in studies of antenna characteristics such as bandwidth, or for the design of some types of matching systems, such as a stub match.

Quantitative measurements of reasonable accuracy demand good design and careful construction in the measuring instruments. They also require intelligent use of the equipment, including a knowledge not only of its limitations but also of stray effects that often lead to false results. Until you know the complete conditions of the measurements, a certain amount of skepticism regarding numerical data resulting from amateur measurements with simple equipment is justified. On the other hand, purely qualitative or relative measurements are easy to make and are reliable for the purposes mentioned above.

LINE CURRENT AND VOLTAGE

A current or voltage indicator that can be used with coaxial line is a useful piece of equipment. It need not be elaborate or expensive. Its principal function is to show when the maximum power is being taken from the transmitter; for any given set of line conditions (length, SWR, etc). This will occur when you adjust the transmitter coupling for maximum current or voltage into the transmission line. Although the final-amplifier plate or collector current meter is frequently used for this purpose, it is not always a reliable indicator. In many cases, particularly with a screen-grid tube in the final stage, minimum loaded plate current does not occur simultaneously with maximum power output.

RF VOLTMETER

You can put together a germanium diode in conjunc-

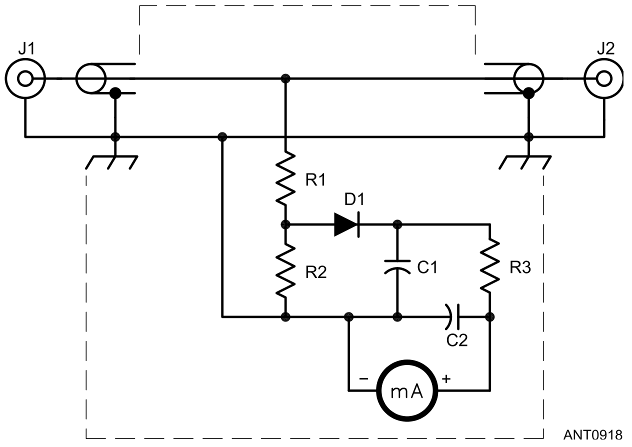

Fig 1—RF voltmeter for coaxial line.

C1, C2—0.005- or 0.01-µF ceramic.

D1—Germanium diode, 1N34A.

J1, J2—Coaxial fittings, chassis-mounting type.

M1—0-1 milliammeter (more sensitive meter may be used if desired; see text).

R1—6.8 kΩ, composition, 1 W for each 100 W of RF power.

R2—680 Ω, 1/2 or 1 W composition.

R3—10 kΩ, 1/2 W (see text).

tion with a low-range milliammeter and a few resistors to form an RF voltmeter suitable for connecting across the two conductors of a coaxial line, as shown in **Fig 1**. It consists of a voltage divider, R1-R2, having a total resistance about 100 times the Z_0 of the line (so the power consumed will be negligible) with a diode rectifier and milliammeter connected across part of the divider to read relative RF voltage. The purpose of R3 is to make the meter readings directly proportional to the applied voltage, as nearly as possible, by swamping the resistance of D1, since the diode resistance will vary with the amplitude of the current through the diode.

You may construct the voltmeter in a small metal box, indicated by the dashed line in the drawing, and fitted with coax receptacles. R1 and R2 should be carbon-composition resistors. The power rating for R1 should be 1 W for each 100 W of carrier power in the matched line; separate 1- or 2-W resistors should be used to make up the total power rating required, to the total resistance as given. Any type of resistor can be used for R3; the total resistance should be such that about 10 V dc will be developed across it at full scale. For example, a 0-1 milliammeter would require 10 kΩ, a 0-500 microammeter would take 20 kΩ, and so on. For comparative measurements only, R3 may be a variable resistor so the sensitivity can be adjusted for various power levels.

In constructing such a voltmeter, you should exercise care to prevent inductive coupling between R1 and the loop formed by R2, D1 and C1, and between the same loop and the line conductors in the assembly. With the lower end of R1 disconnected from R2 and grounded to the enclosure, but without changing its position with respect to the loop, there should be no meter indication when full power is going through the line.

If more than one resistor is used for R1, the units should be arranged end-to-end with very short leads. R1 and R2 should be kept 1/2 inch or more from metal surfaces parallel to the body of the resistor. If you observe these precautions the voltmeter will give consistent readings at frequencies up to 30 MHz. Stray capacitance and stray coupling limit the accuracy at higher frequencies but do not affect the utility of the instrument for comparative measurements.

Calibration

You may calibrate the meter for RF voltage by comparison with a standard such as an RF ammeter. This requires that the line be well matched so the impedance at the point of measurement is equal to the actual Z_0 of the line. Since in that case $P = I^2 Z_0$, the power can be calculated from the current. Then $E = \sqrt{PZ_0}$. By making current and voltage measurements at a number of different power levels, you can obtain enough points to draw a calibration curve for your particular setup.

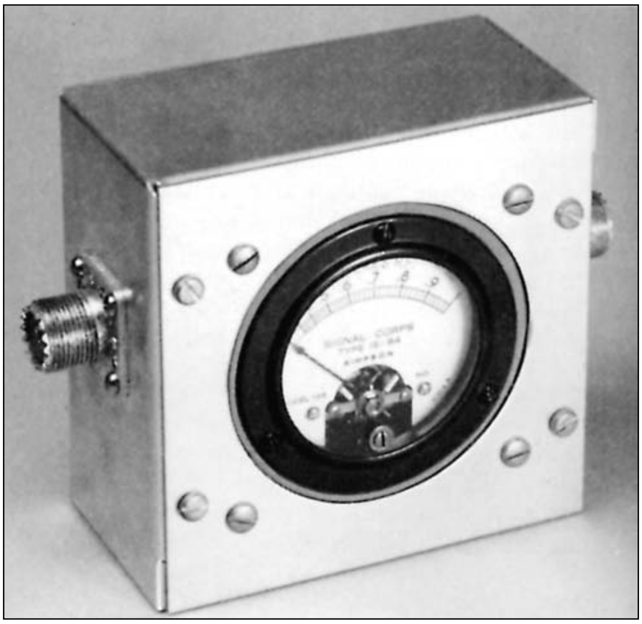

Fig 2—A convenient method of mounting an RF ammeter for use in a coaxial line. This is a metal-case instrument mounted on a thin bakelite panel. The cutout in the metal clears the edge of the meter by about 1/8 inch.

RF AMMETERS

Although they are not as widely available as they used to be, if you can find one on the surplus market or at a hamfest, an RF ammeter is a good way to gauge output power. You can mount an RF ammeter in any convenient location at the input end of the transmission line, the principal precaution being that the capacitance to ground, chassis, and nearby conductors should be low. A bakelite-case instrument can be mounted on a metal panel without introducing enough shunt capacitance to ground to cause serious error up to 30 MHz. When installing a metal-case instrument on a metal panel, you should mount it on a separate sheet of insulating material so that there is 1/8 inch or more separation between the edge of the case and the metal.

A 2-inch instrument can be mounted in a 2 × 4 × 4-inch metal box, as shown in **Fig 2**. This is a convenient arrangement for use with coaxial line. Installed this way, a good quality RF ammeter will measure current with an accuracy that is entirely adequate for calculating power in the line. As discussed above in connection with calibrating RF voltmeters, the line must be closely matched by its load so the actual impedance is resistive and equal to Z_0. The scales of such instruments are cramped at the low end, however, which limits the range of power that can be measured by a single meter. The useful current range is about 3 to 1, corresponding to a power range of about 9 to 1.

SWR Measurements

On parallel-conductor lines it is possible to measure the standing-wave ratio by moving a current (or voltage) indicator along the line, noting the maximum and minimum values of current (or voltage) and then computing the SWR from these measured values. This cannot be done with coaxial line since it is not possible to make measurements of this type inside the cable. The technique is, in fact, seldom used with open lines because it is not only inconvenient but sometimes impossible to reach all parts of the line conductors. Also, the method is subject to considerable error from antenna currents flowing on the line.

Present-day SWR measurements made by amateurs practically always use some form of *directional coupler* or RF-bridge circuit. The indicator circuits themselves are fundamentally simple, but they require considerable care in construction to ensure accurate measurements. The requirements for indicators used only for the adjustment of impedance-matching circuits, rather than actual SWR measurement, are not so stringent, and you can easily make an instrument for this purpose.

BRIDGE CIRCUITS

Two commonly used bridge circuits are shown in **Fig 3**. The bridges consist essentially of two voltage dividers in parallel, with a voltmeter connected between the junctions of each pair of *arms*, as the individual elements are called. When the equations shown to the right of each circuit are satisfied there is no potential difference between the two junctions, and the voltmeter indicates zero voltage. The bridge is then said to be in *balance*.

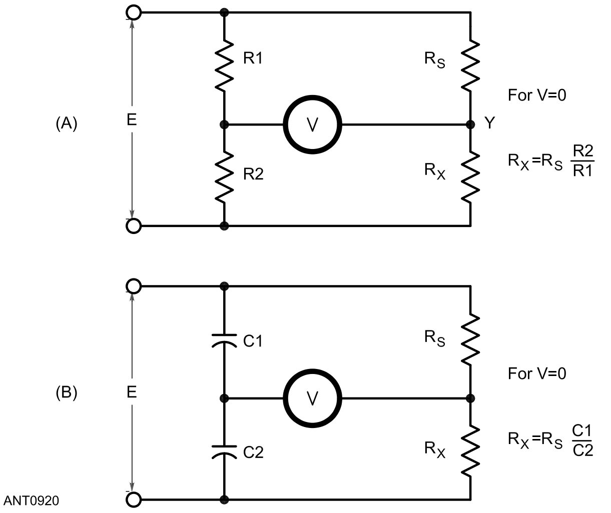

Fig 3—Bridge circuits suitable for SWR measurement. At A, Wheatstone type using resistance arms. At B, capacitance-resistance bridge ("Micromatch"). Conditions for balance are independent of frequency in both types.

Taking Fig 3A as an illustration, if R1 = R2, half the applied voltage, E, will appear across each resistor. Then if $R_S = R_X$, $\frac{1}{2}$E will appear across each of these resistors and the voltmeter reading will be zero. Remember that a matched transmission line has essentially a purely resistive input impedance. Suppose that the input terminals of such a line are substituted for R_X. Then if R_S is a resistor equal to the Z_0 of the line, the bridge will be balanced.

If the line is not perfectly matched, its input impedance will not equal Z_0 and hence will not equal R_S, since you chose the latter to be equal to Z_0. There will then be a difference in potential between points X and Y, and the voltmeter will show a reading. Such a bridge therefore can be used to show the presence of standing waves on the line, because the line input impedance will be equal to Z_0 only when there are no standing waves.

Considering the nature of the incident and reflected components of voltage that make up the actual voltage at the input terminals of the line, as discussed in Chapter 24, it should be clear that when $R_S = Z_0$, the bridge is always in balance for the incident component. Thus the voltmeter does not respond to the incident component at any time but reads only the reflected component (assuming that R2 is very small compared with the voltmeter impedance). The incident component can be measured across either R1 or R2, if they are equal resistances. The standing-wave ratio is then

$$\text{SWR} = \frac{E1+E2}{E1-E2} \qquad \text{(Eq 1)}$$

where E1 is the incident voltage and E2 is the reflected voltage. It is often simpler to normalize the voltages by expressing E2 as a fraction of E1, in which case the formula becomes

$$\text{SWR} = \frac{1+k}{1-k} \qquad \text{(Eq 2)}$$

where k = E2/E1.

The operation of the circuit in Fig 3B is essentially the same, although this circuit has arms containing reactance as well as resistance.

It is not necessary that R1 = R2 in Fig 3A; the bridge can be balanced, in theory, with any ratio of these two resistances provided R_S is changed accordingly. In practice, however, the accuracy is highest when the two are equal; this circuit is most commonly used.

A number of types of bridge circuits appear in **Fig 4**, many of which have been used in amateur products or amateur construction projects. All except that at G can have the generator and load at a common potential. At G, the generator and detector are at a common potential. You may interchange the positions of the detector and trans-

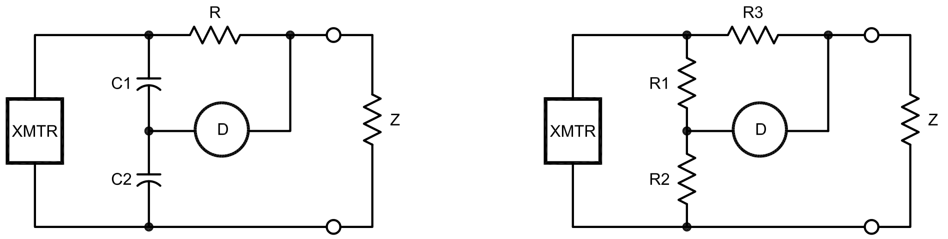
R
C1
XMTR
D
Z
C2
R3
R1
XMTR
D
Z
R2

De Sauty/Wein (Micromatch)
(A)

Balance $Z=\frac{RC1}{C2}$

Christie/Wheatstone (Antenna – Scope)
(B)

Balance $Z=\frac{R2\ R3}{R1}$

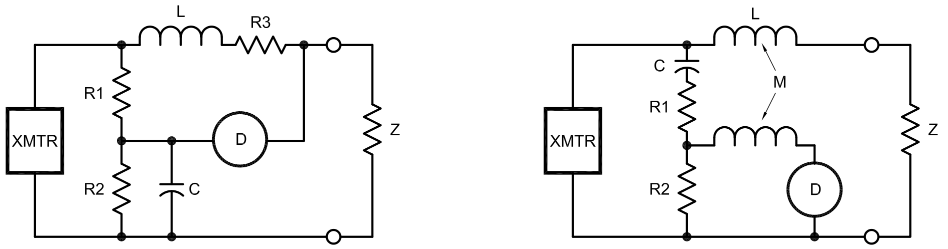
L
R3
R1
XMTR
D
Z
R2
C
L
C
M
R1
XMTR
Z
R2
D

Maxwell (Universal)
(C)

Balance R1 Z=R2 R3=L/C
No Discontinuity: R2→∞,
R3→0, R1=Z

Carey – Foster
(Twin – Lamp, Monomatch Mickey – Match)
(D)

Balance M=C R2 Z
L=M(1+R1/R2)

No Discontinuity: R1+ R2 = Z = $\sqrt{L/C}$

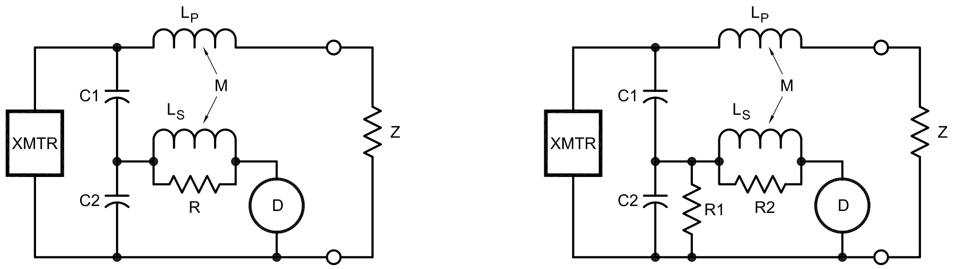
L_P
M
C1
L_S
XMTR
Z
C2
R
D
L_P
M
C1
L_S
XMTR
Z
C2
R1
R2
D

Bruene (Collins Radio)
(E)
Balance (Approx.) Z C1 L_S=M R (C1+C2)
($2\pi f L_S >> R$)
(L_P=M Approx)

Max. 1% Error: $2\pi f L_S \geq 7R$

Phase – Compensated
(F)
Balance: Z R1 C1=M=L_P
L_S=R1 R2 (C1+C2)

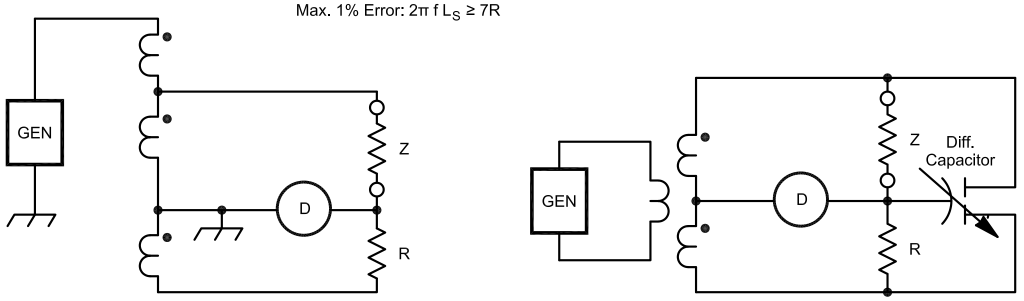
GEN
Z
D
R
GEN
Z
Diff.
Capacitor
D
R

(G)

(H)

Starr's "Hybrid Coil"

Balance: R=Z
(The Differential Capacitor Can Balance Parallel Reactance.)

ANT0921

Fig 4—Various types of SWR indicator circuits and commonly known names of bridge circuits or devices in that they have been used. Detectors (D) are usually semiconductor diodes with meters, isolated with RF chokes and capacitors. However, the detector may be a radio receiver. In each circuit, Z represents the load being measured. (*This information provided by David Geiser, WA2ANU*)

mitter (or generator) in the bridge, and this may be advantageous in some applications.

The bridges shown at D, E, F and H may have one terminal of the generator, detector and load common. Bridges at A, B, E, F, G and H have constant sensitivity over a wide frequency range. Bridges at B, C, D and H may be designed to show no discontinuity (impedance lump) with a matched line, as shown in the drawing. Discontinuities with A, E and F may be small.

Bridges are usually most sensitive when the detector bridges the midpoint of the generator voltage, as in G or H, or in B when each resistor equals the load impedance. Sensitivity also increases when the currents in each leg are equal.

Resistance Bridge

The basic bridge configuration shown in Fig 3B may be home constructed and is reasonably accurate for SWR measurement. A practical circuit for such a bridge is given in **Fig 5** and a representative layout is shown in **Fig 6**. Properly built, a bridge of this design can be used for measurement of SWRs up to about 15:1 with good accuracy.

You should observe these important construction points:

1) Keep leads in the RF circuit short, to reduce stray inductance.
2) Mount resistors two or three times their body diameter away from metal parts, to reduce stray capacitance.
3) Place the RF components so there is as little inductive and capacitive coupling as possible between the bridge arms.

In the instrument shown in Fig 6, the input and line connectors, J1 and J2, are mounted fairly close together so the standard resistor, R_S, can be supported with short leads directly between the center terminals of the connectors. R2 is mounted at right angles to R_S, and a shield partition is used between these two components and the others.

The two 47-kΩ resistors, R5 and R6 in Fig 5, are voltmeter multipliers for the 0-100 microammeter used as an indicator. This is sufficient resistance to make the voltmeter linear (that is, the meter reading is directly proportional to the RF voltage) and no voltage calibration curve is needed. D1 is the rectifier for the reflected volt-

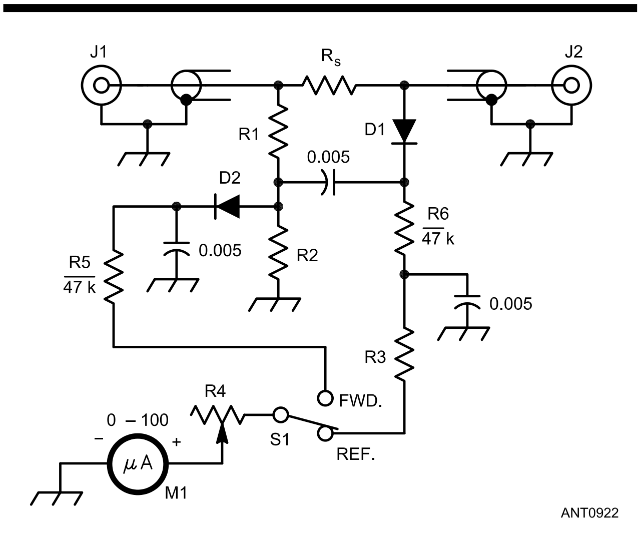

Fig 5—Resistance bridge for SWR measurement. Capacitors are disc ceramic. Resistors are $^1/_2$-watt composition except as noted below.
D1, D2—Germanium diode, high back resistance type (1N34A, 1N270, etc).
J1, J2—Coaxial connectors, chassis-mounting type.
M1—0-100 dc microammeter.
R1, R2—47 Ω, $^1/_2$-W composition (see text).
R3—See text.
R4—50-kΩ volume control.
R_S—Resistance equal to line Z_0 ($^1/_2$ or 1 W composition).
S1—SPDT toggle.

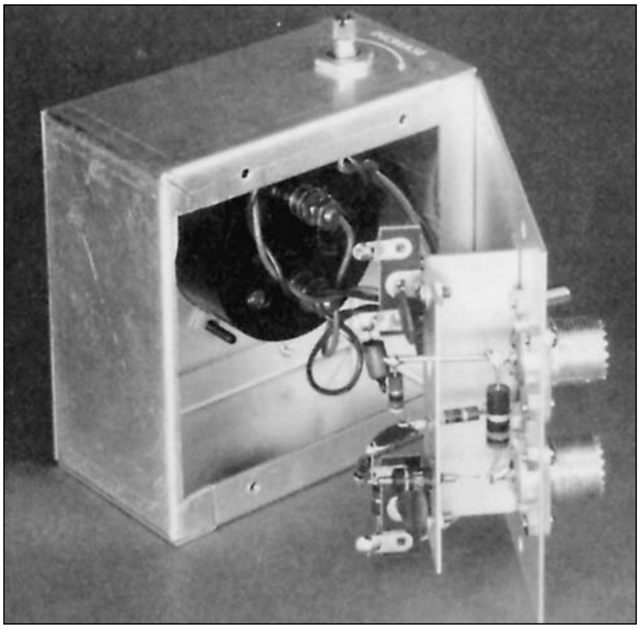

Fig 6—A 2 × 4 × 4-inch aluminum box is used to house this SWR bridge, which uses the circuit of Fig 5. The variable resistor, R4, is mounted on the side. The bridge components are mounted on one side plate of the box and a subchassis formed from a piece of aluminum. The input connector is at the top in this view. R_S is connected directly between the two center posts of the connectors. R2 is visible behind it and perpendicular to it. One terminal of D1 projects through a hole in the chassis so the lead can be connected to J2. R1 is mounted vertically to the left of the chassis in this view, with D2 connected between the junction of R1-R2 and a tie point.

age and D2 is for the incident voltage. Because of manufacturing variations in resistors and diodes, the readings may differ slightly with two multipliers of the same nominal resistance value, so a correction resistor, R3, is included in the circuit. You should select its value so that the meter reading is the same with S1 in either position, when RF is applied to the bridge with the line connection open. In the instrument shown, a value of 1000 Ω was required in series with the multiplier for reflected voltage; in other cases different values probably would be needed and R3 might have to be put in series with the multiplier for the incident voltage. You can determine this by experiment.

The value used for R1 and R2 is not critical, but you should match the two resistors within 1% or 2% if possible. Keep the resistance of R_S as close as possible to the actual Z_0 of the line you use (generally 50 or 75 Ω). Select the resistor by actual measurement with an accurate resistance bridge, if you have one available.

R4 is for adjusting the incident-voltage reading to full scale in the measurement procedure described below. Its use is not essential, but it offers a convenient alternative to exact adjustment of the RF input voltage.

Testing

Measure R1, R2 and R_S with a reliable digital ohmmeter or resistance bridge after completing the wiring. This will ensure that their values have not changed from the heat of soldering. Disconnect one side of the microammeter and leave the input and output terminals of the unit open during such measurements to avoid stray shunt paths through the rectifiers.

Check the two voltmeter circuits as described above, applying enough RF (about 10 V) to the input terminals to give a full-scale reading with the line terminals open. If necessary, try different values for R3 until the reading is the same with S1 in either position.

With J2 open, adjust the RF input voltage and R4 for full-scale reading with S1 in the incident-voltage position. Then switch S1 to the reflected-voltage position. The reading should remain at full scale. Next, short-circuit J2 by touching a screwdriver between the center terminal and the frame of the connector to make a low-inductance short. Switch S1 to the incident-voltage position and readjust R4 for full scale, if necessary. Then throw S1 to the reflected-voltage position, keeping J2 shorted, and the reading should be full scale as before. If the readings differ, R1 and R2 are not the same value, or there is stray coupling between the arms of the bridge. You must read the reflected voltage at full scale with J2 either open or shorted, when the incident voltage is set to full scale in each case, to make accurate SWR measurements.

The circuit should pass these tests at all frequencies at which it is to be used. It is sufficient to test at the lowest and highest frequencies, usually 1.8 or 3.5 and 28 or 50 MHz. If R1 and R2 are poorly matched but the bridge construction is otherwise good, discrepancies in the readings will be substantially the same at all frequencies. A difference in behavior at the low and high ends of the frequency range can be attributed to stray coupling between bridge arms, or stray inductance or capacitance in the arms.

To check the bridge for balance, apply RF and adjust R4 for full scale with J2 open. Then connect a resistor identical with R_S (the resistance should match within 1% or 2%) to the line terminals, using the shortest possible leads. It is convenient to mount the test resistor inside a cable connector (PL-259), a method of mounting that also minimizes lead inductance. When you connect the test resistor the reflected-voltage reading should drop to zero. The incident voltage should be reset to full scale by means of R4, if necessary. The reflected reading should be zero at any frequency in the range to be used. If a good null is obtained at low frequencies but some residual current shows at the high end, the trouble may be the inductance of the test resistor leads, although it may also be caused by stray coupling between the arms of the bridge itself.

If there is a constant low (but not zero) reading at all frequencies the problem is poor matching of the resistance values. Both effects can be present simultaneously. You should make sure you obtain a good null at all frequencies before using your bridge.

Bridge Operation

You must limit the RF power input to a bridge of this type to a few watts at most, because of the power-dissipation ratings of the resistors. If the transmitter has no provision for reducing power output to a very low value—less than 5 W—a simple power-absorber circuit can be made up, as shown in **Fig 7**. Lamp DS1 tends to maintain constant current through the resistor over a fairly

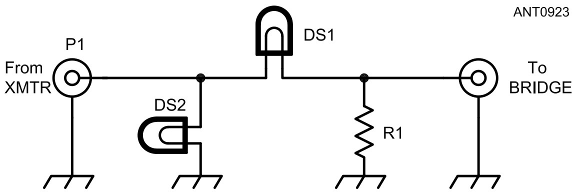

Fig 7—Power-absorber circuit for use with resistance-type SWR bridges when the transmitter has no special provisions for power reduction. For RF powers up to 50 W, DS1 is a 117-V 40-W incandescent lamp and DS2 is not used. For higher powers, use sufficient additional lamp capacity at DS2 to load the transmitter to about normal output; for example, for 250 W output DS2 may consist of two 100-W lamps in parallel. R1 is made from three 1-W 68-Ω resistors connected in parallel. P1 and P2 are cable-mounting coaxial connectors. Leads in the circuit formed by the lamps and R1 should be kept short, but convenient lengths of cable may be used between this assembly and the connectors.

wide power range, so the voltage drop across the resistor also tends to be constant. This voltage is applied to the bridge, and with the constants given is in the right range for resistance-type bridges.

To make a measurement, connect the unknown load to J2 and apply sufficient RF voltage to J1 to give a full-scale incident-voltage reading. Use R4 to set the indicator to exactly full scale. Then throw S1 to the reflected voltage position and note the meter reading. The SWR is then found by using these readings in Eq 1.

For example, if the full-scale calibration of the dc instrument is 100 µA and the reading with S2 in the reflected-voltage position is 40 µA, the SWR is

$$\mathrm{SWR} = \frac{100+40}{100-40} = \frac{140}{60} = 2.33:1$$

Instead of calculating the SWR value, you could use the voltage curve in **Fig 8**. In this example the ratio of reflected to forward voltage is 40/100 = 0.4, and from Fig 8 the SWR value is about 2.3:1.

You may calibrate the meter scale in any arbitrary units, so long as the scale has equal divisions. It is the ratios of the voltages, and not the actual values, that determine the SWR.

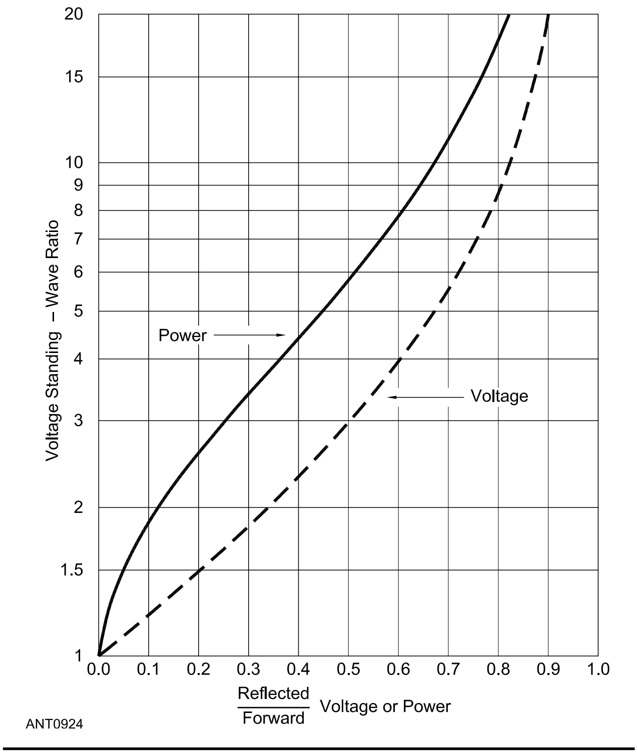

Fig 8—Chart for finding voltage standing-wave ratio when the ratio of reflected-to-forward voltage or reflected-to-forward power is known.

AVOIDING ERRORS IN SWR MEASUREMENTS

The principal causes of inaccuracies within the bridge are differences in the resistances of R1 and R2, stray inductance and capacitance in the bridge arms, and stray coupling between arms. If the checkout procedure described above is followed carefully, the bridge in Fig 5 should be amply accurate for practical use. The accuracy is highest for low standing-wave ratios because of the nature of the SWR calculation; at high ratios the divisor in the equation above represents the difference between two nearly equal quantities, so a small error in voltage measurement may mean a considerable difference in the calculated SWR.

The standard resistor R_S must equal the actual Z_0 of the line. The actual Z_0 of a sample of line may differ by a few percent from the nominal figure because of manufacturing variations, but this has to be tolerated. In the 50- to 75-Ω range, the RF resistance of a composition resistor of $^1/_2$- or 1-W rating is essentially identical with its dc resistance.

Common-Mode Currents

As explained in Chapter 26, there are two ways in which unwanted *common-mode* (sometimes called *antenna*) currents can flow on the outside of a coaxial line—currents radiated onto the line because of its spatial relationship to the antenna and currents that result from the direct connection between the coax outer conductor and (usually) one side of the antenna. The radiated current usually will not be troublesome if the bridge and the transmitter (or other source of RF power for operating the bridge) are shielded so that any RF currents flowing on the outside of the line cannot find their way into the bridge. This point can be checked by inserting an additional section of line ($^1/_8$ to $^1/_4$ electrical wavelength preferably) of the same Z_0. The SWR indicated by the bridge should not change except for a slight decrease because of the additional line loss. If there is a marked change, you may need better shielding.

Parallel-type currents caused by the connection to the antenna without using a common-mode choke balun will change the SWR with variations in line length, even though the bridge and transmitter are well-shielded and the shielding is maintained throughout the system by the use of coaxial fittings. Often, merely moving the transmission line around will cause the indicated SWR to change. This is because the outside of the coax becomes part of the antenna system—being connected to the antenna at the feed point. The outside shield of the line thus constitutes a load, along with the desired load represented by the antenna itself. The SWR on the line then is determined by the composite load of the antenna and the outside of the coax. Since changing the line length (or position) changes one component of this composite load, the SWR changes too.

The remedy for such a situation is to use a good balun or to detune the outside of the line by proper choice of length. Note that this is not a *measurement error*, since what the instrument reads is the actual SWR on the line. However, it is an undesirable condition since the line is usually operating at a higher SWR than it should—and would if the parallel-type current on the outside of the coax were eliminated.

Spurious Frequencies

Off-frequency components in the RF voltage applied to the bridge may cause considerable error. The principal components of this type are harmonics and low-frequency subharmonics that may be fed through the final stage of the transmitter driving the bridge. The antenna is almost always a fairly selective circuit, and even though the system may be operating with a very low SWR at the desired frequency, it is almost always mismatched at harmonic and subharmonic frequencies. If such spurious frequencies are applied to the bridge in appreciable amplitude, the SWR indication will be very much in error. In particular, it may not be possible to obtain a null on the bridge with any set of adjustments of the matching circuit. The only remedy is to filter out the unwanted components by increasing the selectivity of the circuits between the transmitter final amplifier and the bridge.

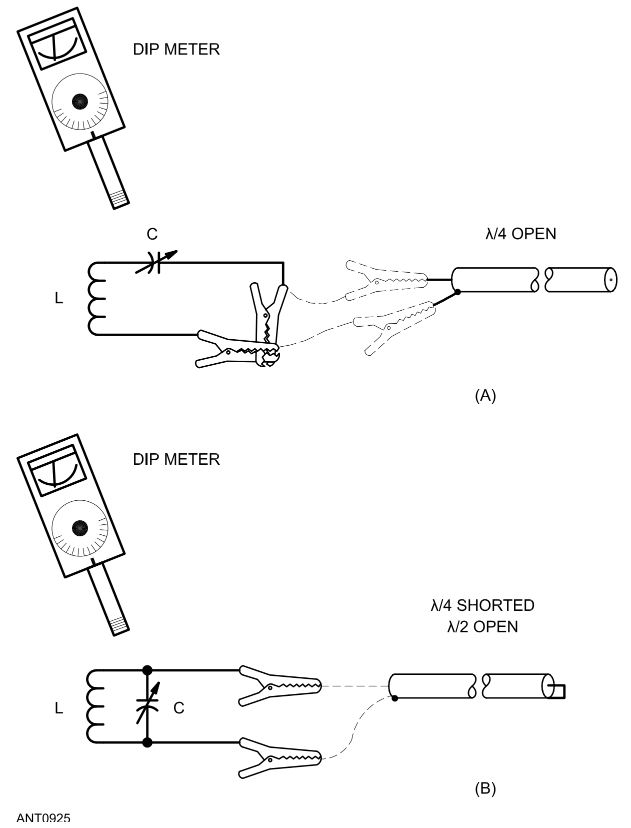

Fig 9—Methods of determining $^1/_4$ and $^1/_2$-λ line lengths. At A, $^1/_4$-λ open-circuited line; at B, $^1/_4$-λ shorted and $^1/_2$-λ open-circuited line.

MEASURING LINE LENGTH

The following material is taken from information in September 1985 *QST* by Charlie Michaels, W7XC (see Bibliography).

There is a popular myth that one may prepare an open quarter-wave line by connecting a loop of wire to one end and trimming the line to resonance (as indicated by a dip meter). This actually yields a line with capacitive reactance equal to the inductive reactance of the loop: a 4-inch wire loop yields a line 82.8° line at 18 MHz; a 2-inch loop yields an 86° line. As the loop size is reduced, line length approaches—but never equals—90°.

To make a quarter-wave open line, parallel connect a coil and capacitor that resonate at the required frequency (see **Fig 9A**). After adjusting the network to resonance, do not make further network adjustments. Open the connection between the coil and capacitor and series connect the line to the pair. Start with a line somewhat longer than required, and trim it until the circuit again resonates at the desired frequency. For a shorted quarter-wave line or an open half-wave line, connect the line in parallel with the coil and capacitor (see Fig 9B).

Another method to accurately measure a coaxial transmission line length uses one of the popular “SWR analyzers,” portable hand-held instruments with a tunable low-power signal generator and an SWR bridge. While an SWR analyzer cannot compute the very high values of SWR at the input of a shorted quarter-wave line at the fundamental frequency, most include another readout showing the magnitude of the impedance. This is very handy for finding a low-impedance dip, rather than a high-impedance peak.

At the operating frequency, a shorted quarter-wave line results in a high-impedance open-circuit at the input to that line. At twice the frequency, where the line is now one-half wave long electrically, the instrument shows a low-impedance short-circuit. However, when you are pruning a line to length by cutting off short pieces at the end, it is inconvenient to have to install a short before measuring the response. It is far easier to look for the dip in impedance when a quarter-wave line is terminated in an open circuit.

Again, the strategy is to start with a line physically a little longer than a quarter-wave length. A good rule of thumb is to cut the line 5% longer to take into account the variability in the velocity factor of a typical coax cable. Compute this using:

$$\text{Length (feet)} = 0.25 \times 1.05 \times \text{VF} \times 984/\text{Freq} = \text{VF}/\text{Freq}$$

where

Freq is in MHz

VF is the velocity factor in %.

Plug the coax connector installed at one end of the line into the SWR analyzer and find the frequency for the impedance dip. Prune the line by snipping off short pieces at the end. Once you've pruned the line to the desired frequency, connect the short at the end of the line and recheck for a short circuit at twice the fundamental frequency. Seal the shorted end of the coax and you're done.

REFLECTOMETERS

Low-cost *reflectometers* that do not have a guaranteed wattmeter calibration are not ordinarily reliable for accurate numerical measurement of standing-wave ratio. They are, however, very useful as aids in the adjustment of matching networks, since the objective in such adjustment is to reduce the reflected voltage or power to zero. Relatively inexpensive devices can be used for this, since only good bridge balance is required, not actual calibration of SWR. Bridges of this type are usually frequency-sensitive that is, the meter response increase with increasing frequency for the same applied voltage. When matching and line monitoring, rather than SWR measurement, is the principal use of the device, this is not a serious handicap.

Various simple reflectometers, useful for matching and monitoring, have been described from time to time in *QST* and in *The ARRL Handbook*. Because most of these are frequency sensitive, it is difficult to calibrate them accurately for power measurement, but their low cost and suitability for use at moderate power levels, combined with the ability to show accurately when a matching circuit has been properly adjusted, make them a worthwhile addition to the amateur station.

The Tandem Match—An Accurate Directional Wattmeter

Most SWR meters are not very accurate at low power levels because the detector diodes do not respond to low voltage in a linear fashion. This design uses a compensating circuit to cancel diode nonlinearity. It also provides peak detection for SSB operation and direct SWR readout that does not vary with power level. The following information is condensed from an article by John Grebenkemper, KI6WX, in January 1987 *QST.*

DESIGN PRINCIPLES

Directional wattmeters for Amateur Radio use consist of three basic elements: a directional coupler, a detector and a signal-processing and display circuit. A directional coupler samples forward and reflected-power components on a transmission line. An ideal directional coupler would provide signals proportional to the forward and reflected voltages (independent of frequency), which could then be used to measure forward and reflected power over a wide frequency range. The best contemporary designs work over two decades of frequency.

The detector circuit provides a dc output voltage proportional to the ac input voltage. Most directional wattmeters use a single germanium diode as the detector element. A germanium, rather than silicon, diode is used to minimize diode nonlinearity at low power levels. Diode non-linearity still causes SWR measurement errors unless it is compensated ahead of the display circuit. Most directional wattmeters do not work well at low power levels because of diode nonlinearity.

The signal-processing and display circuits compute and display the SWR. There are a number of ways to perform this function. Meters that display only the forward and reflected power require the operator to compute the SWR manually. Many instruments require that the operator adjust the meter to a reference level while measuring forward power, then switch to measure reflected power on a special scale that indicates SWR. Meters that directly compute the SWR using analog signal-processing circuits have been described by Fayman, Perras,

Table 1
Performance Specifications for the Tandem Match

Power range: 1.5 to 1500 W
Frequency range: 1.8 to 54 MHz
Power accuracy: Better than ± 10% (± 0.4 dB)
SWR accuracy: Better than ± 5%
Minimum SWR: Less than 1.05:1
Power display: Linear, suitable for use with either analog or digital meters
Calibration: Requires only an accurate voltmeter

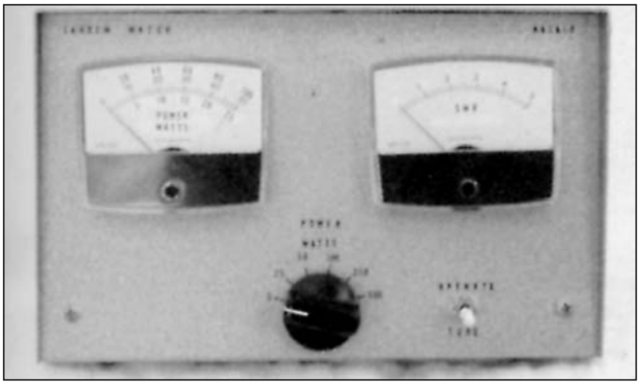

Fig 10—The Tandem Match uses a pair of meters to display net forward power and true SWR simultaneously.

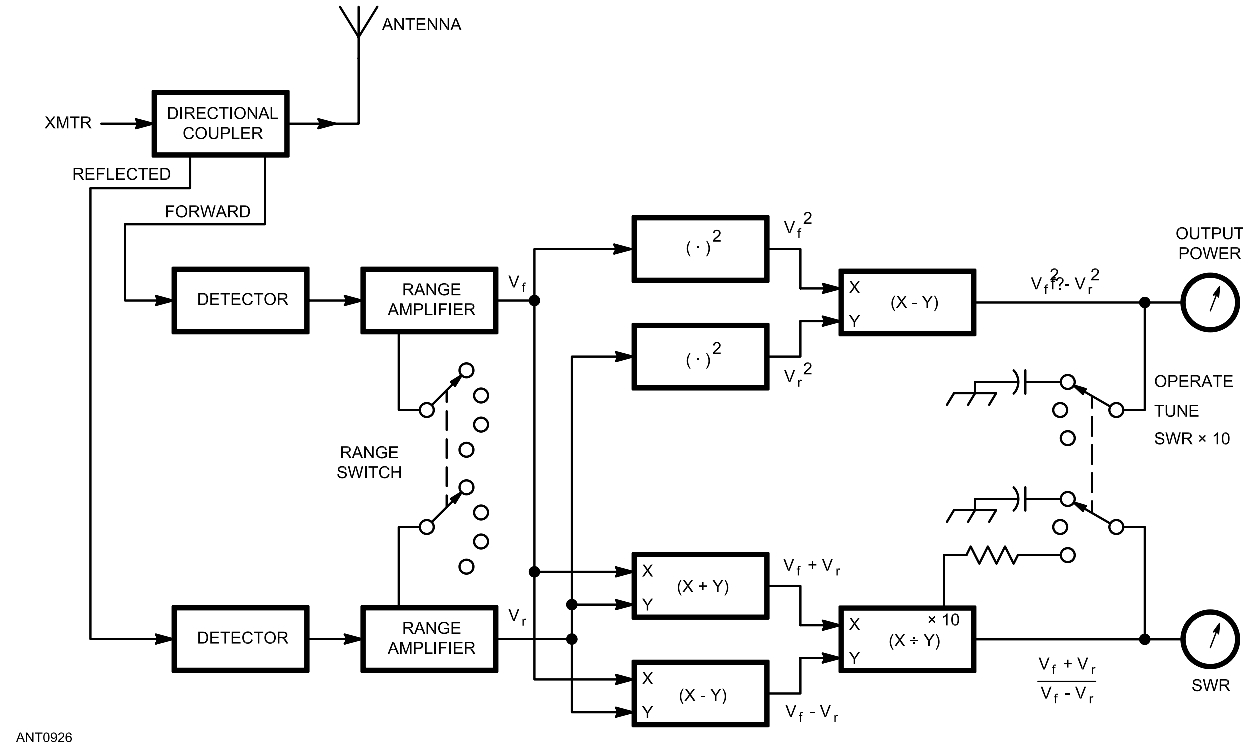

Fig 11—Block diagram of the Tandem Match.

Leenerts and Bailey (see the Bibliography at the end of this chapter).

The next section takes a brief look at several popular circuits that accomplish the functions above and compares them to the circuits used in the Tandem Match. The design specifications of the Tandem Match are shown in **Table l**, and a block diagram is shown in **Fig 11**.

CIRCUIT DESCRIPTION

A directional coupler consists of an input port, an output port and a coupled port. The device takes a portion of the power flowing from the input port to the output port and directs it to the coupled port, but *none* of the power flowing from the output port to the input port is directed to the coupled port.

There are several terms that define the performance of a directional coupler:

1) *Insertion loss* is the amount of power that is lost as the signal flows from the input port to the output port. Insertion loss should be minimized so the coupler doesn't dissipate a significant amount of the transmitted power.

2) *Coupling factor* is the amount of power (or voltage) that appears at the coupled port relative to the amount of power (or voltage) transferred from the input port to the output port. The "flatness" (with frequency) of the coupling factor determines how accurately the directional wattmeter can determine forward and reflected power over a range of frequencies.

3) *Isolation* is the amount of power (or voltage) that appears at the coupled port relative to the amount of power (or voltage) transferred from the output port to the input port.

4) *Directivity* is the isolation less the coupling factor. Directivity dictates the minimum measurable SWR. A directional coupler with 20 dB of directivity measures a 1:1 SWR as l.22:l, but one with 30 dB measures a 1:1 SWR as 1.07:1.

The directional coupler most commonly used in amateur radio was first described in 1959 by Bruene in *QST* (see Bibliography). The coupling factor was fairly flat (± l dB), and the directivity was about 20 dB for a Bruene coupler measured from 3 to 30 MHz. Both factors limit the accuracy of the Bruene coupler for measuring low values of power and SWR. It is a simple directional coupler, however, and it works well over a wide frequency range if great precision is not required.

The coupler used in the Tandem Match (see **Fig 12**) consists of a pair of toroidal transformers connected in tandem. The configuration was patented by Carl G. Sontheimer and Raymond E. Fredrick (US Patent no. 3,426,298, issued February 4, 1969). It has been described by Perras, Spaulding (see Bibliography) and others. With coupling factors of 20 dB or greater, this coupler is suitable for sampling both forward and reflected power.

The configuration used in the Tandem Match works well over the frequency range of 1.8 to 54 MHz, with a

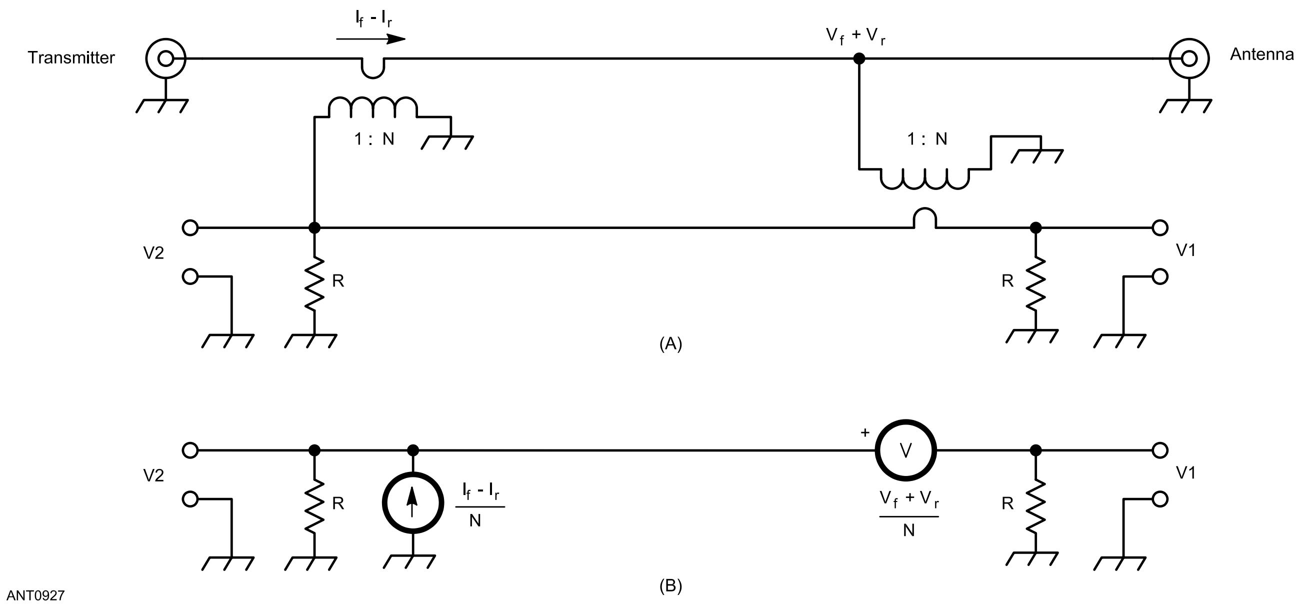
Transmitter
$I_f - I_r$
$V_f + V_r$
Antenna
1 : N
1 : N
V2
R
R
V1
(A)
V2
R
$\frac{I_f - I_r}{N}$
+
V
$\frac{V_f + V_r}{N}$
R
V1
ANT0927
(B)

Fig 12—Simplified diagram of the Tandem Match directional coupler. At A, a schematic of the two transformers. At B, an equivalent circuit.

nominal coupling factor of 30 dB. Over this range, insertion loss is less than 0.1 dB. The coupling factor is flat to within ± 0.1 dB from 1.8 to 30 MHz, and increases to only ±0.3 dB at 50 MHz. Directivity exceeds 35 dB from 1.8 to 30 MHz and exceeds 26 dB at 50 MHz.

The low-frequency limit of this directional coupler is determined by the inductance of the transformer secondary windings. The inductive reactance should be greater than 150 Ω (three times the line characteristic impedance) to reduce insertion loss. The high-frequency limit of this directional coupler is determined by the length of the transformer windings. When the winding length approaches a significant fraction of a wavelength, coupler performance deteriorates.

The coupler described here may overheat at 1500 W on 160 meters (because of the high circulating current in the secondary of T2). The problem could be corrected by using a larger core or one with greater permeability. A larger core would require longer windings; that option would decrease the high-frequency limit.

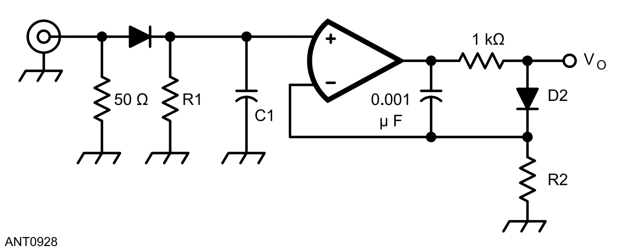
50 Ω
R1
C1
+
–
0.001
μ F
1 kΩ
V_O
D2
R2
ANT0928

Fig 13—Simplified diagram of the detector circuit used in the Tandem Match. The output voltage, V_O is approximately equal to the input voltage. D1 and D2 must be a matched pair (see text). The op amp should have a low offset voltage (less than 1 mV), a low leakage current (less than 1 nA), and be stable over time and temperature. The resistor and capacitor in the feedback path assure that the op amp will be stable.

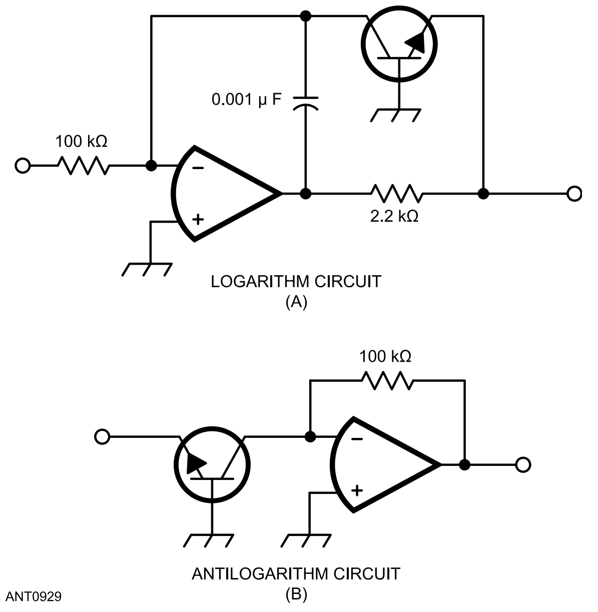
0.001 μ F
100 kΩ
–
+
2.2 kΩ
LOGARITHM CIRCUIT
(A)
100 kΩ
–
+
ANTILOGARITHM CIRCUIT
(B)
ANT0929

Fig 14—Simplified diagrams of the log circuit at A and the antilog circuit at B.

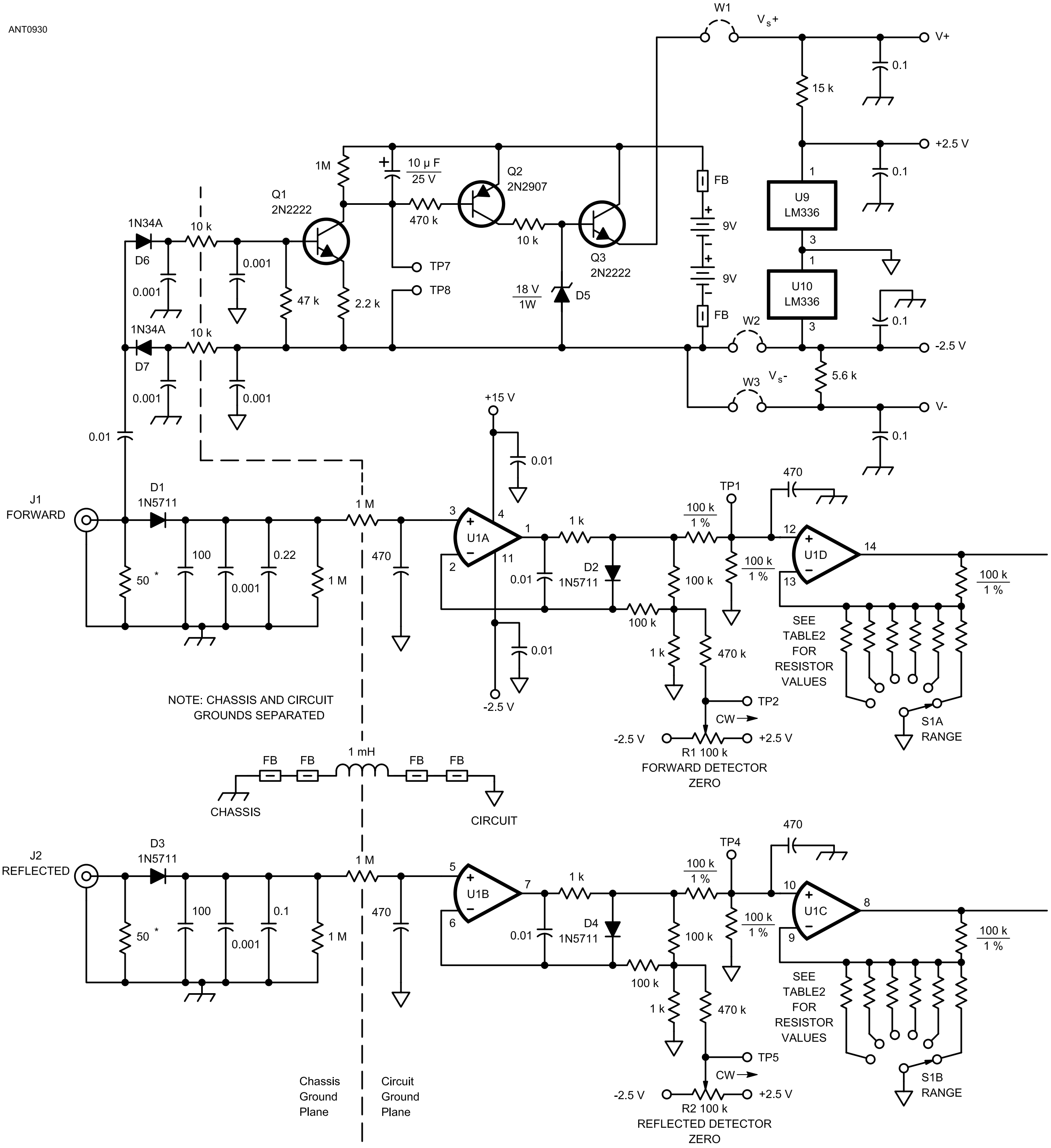
ANT0930
W1
V_s+
V+
0.1
15 k
+2.5 V
0.1
1M
10 μ F
25 V
Q2
2N2907
Q1
2N2222
470 k
FB
1
U9
LM336
9V
1N34A
10 k
D6
0.001
0.001
47 k
2.2 k
TP7
TP8
10 k
Q3
2N2222
18 V
1W
D5
3
1
U10
LM336
9V
FB
W2
3
0.1
-2.5 V
1N34A
10 k
D7
0.001
0.001
W3
V_s-
5.6 k
V-
0.1
+15 V
0.01
0.01
J1
FORWARD
D1
1N5711
1 M
3
4
U1A
1
2
11
100
0.22
0.001
50 *
1 M
470
1 k
0.01
D2
1N5711
100 k
1 %
TP1
470
12
U1D
14
13
100 k
100 k
1 %
100 k
1 %
100 k
1 k
470 k
0.01
-2.5 V
SEE
TABLE2
FOR
RESISTOR
VALUES
TP2
CW
-2.5 V
+2.5 V
R1 100 k
FORWARD DETECTOR
ZERO
S1A
RANGE
NOTE: CHASSIS AND CIRCUIT
GROUNDS SEPARATED
FB
FB
1 mH
FB
FB
CHASSIS
CIRCUIT
J2
REFLECTED
D3
1N5711
1 M
5
U1B
7
6
100
0.1
0.001
50 *
1 M
470
1 k
0.01
D4
1N5711
100 k
1 %
TP4
470
10
U1C
8
9
100 k
100 k
1 %
100 k
1 %
100 k
1 k
470 k
SEE
TABLE2
FOR
RESISTOR
VALUES
TP5
CW
-2.5 V
+2.5 V
R2 100 k
REFLECTED DETECTOR
ZERO
S1B
RANGE
Chassis
Ground
Plane
Circuit
Ground
Plane

Fig 15—Schematic diagram for the Tandem Match directional wattmeter. Parts identified as RS are from RadioShack. For other parts sources, see Table 3. See Fig 17 for construction of 50-Ω loads at J1 and J2.

D1, D2—Matched pair 1N5711, or equivalent.
D3, D4—Matched pair 1N5711, or equivalent.
D6, D7—1N34A.
D8-D14—1N914.
FB—Ferrite bead, Amidon FB-73-101 or equiv.
J1, J2—SO-239 connector.
J3, J4—Open-circuit jack.
M1, M2—50 μA panel meter, RS 270-1751.
Q1, Q3, Q4—2N2222 or equiv.
Q2—2N2907 or equiv.

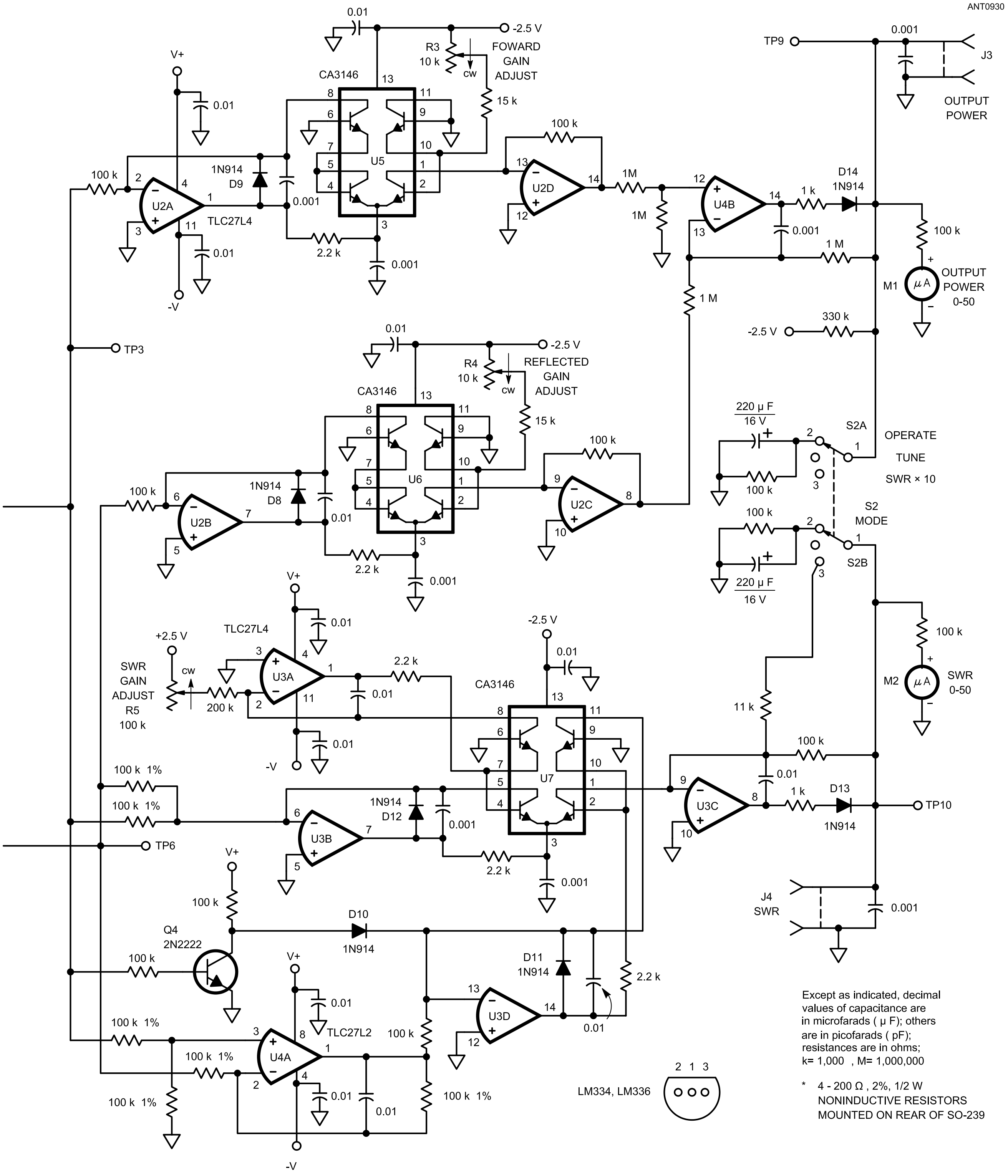
ANT0930
0.01
-2.5 V
R3
10 k
CW
FOWARD
GAIN
ADJUST
15 k
CA3146
U5
13
8
6
7
5
4
11
9
10
1
2
3
V+
0.01
100 k
2
U2A
4
1
3
11
TLC27L4
1N914
D9
0.001
2.2 k
0.001
0.01
-V
100 k
U2D
13
12
14
1M
1M
U4B
12
13
14
1 k
D14
1N914
0.001
1 M
1 M
TP9
0.001
J3
OUTPUT
POWER
100 k
M1
μA
OUTPUT
POWER
0-50
-2.5 V
330 k
TP3
0.01
-2.5 V
R4
10 k
CW
REFLECTED
GAIN
ADJUST
15 k
CA3146
U6
100 k
U2B
6
5
7
1N914
D8
0.01
2.2 k
0.001
100 k
U2C
9
10
8
220 μ F
16 V
S2A
OPERATE
TUNE
SWR × 10
100 k
S2
MODE
S2B
100 k
220 μ F
16 V
V+
0.01
+2.5 V
TLC27L4
U3A
SWR
GAIN
ADJUST
R5
100 k
CW
200 k
2.2 k
0.01
0.01
-V
-2.5 V
0.01
CA3146
U7
100 k
M2
μA
SWR
0-50
11 k
100 k
U3C
9
10
8
0.01
1 k
D13
1N914
TP10
100 k 1%
100 k 1%
U3B
6
5
7
1N914
D12
0.001
2.2 k
0.001
TP6
J4
SWR
0.001
V+
100 k
Q4
2N2222
100 k
D10
1N914
D11
1N914
2.2 k
U3D
13
12
14
0.01
V+
0.01
TLC27L2
100 k 1%
U4A
3
8
2
4
1
100 k 1%
100 k 1%
100 k
100 k 1%
0.01
0.01
-V
LM334, LM336
2 1 3

Except as indicated, decimal values of capacitance are in microfarads (μ F); others are in picofarads (pF); resistances are in ohms; k= 1,000 , M= 1,000,000

* 4 - 200 Ω , 2%, 1/2 W NONINDUCTIVE RESISTORS MOUNTED ON REAR OF SO-239

R1, R2, R5—100 kΩ, 10-turn, cermet Trimpot.
R3, R4—10 kΩ, 10-turn, cermet Trimpot.
U1-U3—TLC27L4 or TLC27M4 quad op amp (Texas Instruments).
U4—TLC27L2 or TLC27M2 dual op amp.
U5-U7—CA3146 quad transistor array.
U8—LM334 adjustable current source.
U9-U10—LM336 2.5-V reference diode. See text.

Detector Circuits

Most amateur directional wattmeters use a germanium diode detector to minimize the forward voltage drop. Detector voltage drop is still significant, however, and an uncompensated diode detector does not respond to small signals in a linear fashion. Many directional wattmeters compensate for diode nonlinearity by adjusting the meter scale.

The effect of underestimating detected power worsens at low power levels. Under these conditions, the ratio of the forward power to the reflected power is overestimated because the reflected power is always less than the forward power. This results in an instrument that underestimates SWR, particularly as power is reduced. A directional wattmeter can be checked for this effect by measuring SWR at several power levels: the SWR should be independent of power level.

The Tandem Match uses a feedback circuit to compensate for diode nonlinearity. A simplified diagram of the compensated detector is shown in **Fig 13**. When used with the 30-dB directional coupler, the output voltage of this circuit tracks the square root of power over a range from 10 mW to 1.5 kW. The compensated diode detector tracks the peak input voltage down to 30 mV, while an uncompensated germanium-diode detector shows significant errors at peak inputs of 1 V and less. More information about compensated detectors appears in Grebenkemper's *QEX* article, "Calibrating Diode Detectors" (see Bibliography).

The compensation circuit uses the voltage across a feedback diode, D2, to compensate for the voltage drop across the detector diode, D1. (The diodes must be a matched pair.) The average current through D1 is determined by the detector diode load resistor, R1. The peak current through this diode is several times larger than the average current; therefore, the current through D2 must be several times larger than the average current through D1 to compensate adequately for the peal voltage drop across D1. This is accomplished by making the feedback-diode load resistor, R2, several times smaller than R1. The voltage at the output of the compensated detector approximates the peak RF voltage at the input. For Schottky barrier diodes and a 1 MΩ detector-diode load resistor, a 5:1 ratio of R1 to R2 is nearly optimal.

Signal-Processing and Display Circuits

The signal-processing circuitry calculates and displays transmission-line power and SWR. When measuring forward power, most directional wattmeters display the actual forward power present in the transmission line, which is the sum of forward and reflected power if a match exists at the input end of the line. Transmission-line forward power is very close to the net forward power (the actual power delivered to the line) so long as the SWR is low. As the SWR increases, however, forward power becomes an increasingly poor measure of the power delivered to the load. At an SWR of 3:1, a forward power reading of 100 W implies that only 75 W is delivered to the load (the reflected power is 25 W), assuming the transmission-line loss is zero.

The Tandem Match differs from most wattmeters in that *it displays the net forward power, rather than the sum of forward and reflected power.* This is the quantity that must be optimized to result in maximum radiated power (and which concerns the FCC).

The Tandem Match directly computes and displays the transmission-line SWR on a linear scale. As the displayed SWR is not affected by changes in transmitter power, a matching network can be simply adjusted to minimize SWR. Transmatch adjustment requires only a few watts.

The heart of the Tandem Match signal-processing circuit is the analog logarithm and antilogarithm circuitry shown in **Fig 14**. The circuit is based on the fact that collector current in a silicon transistor is proportional to the exponential (antilog) of its base-emitter voltage over a range of collector currents from a few nanoamperes to a few milliamperes when the collector-base voltage is zero (see Gibbons and Horn reference in the Bibliography). Variations of this circuit are used in the squaring circuits to convert voltage to power and in the divider circuit used to compute the SWR. With good op amps, this circuit will work well for input voltages from less than 100 mV to greater than 10 V.

(For the Tandem Match, "good" op amps are quad-packaged, low-power-consumption, unity-gain-stable parts with input bias less than 1 nA and offset voltage less than 5 mV. Op amps that consume more power than those shown may require changes to the power supply.)

CONSTRUCTION

The schematic diagram for the Tandem Match is shown in **Fig 15** (see pages 14 and 15). The circuit is designed to operate from batteries and draw very little power. Much of the circuitry is of high impedance, so take care to isolate it from RF fields. House it in a metal case. Most problems in the prototype were caused by stray RF in the op-amp circuitry.

Directional Coupler

The directional coupler is constructed in its own small ($2^3/_4 \times 2^3/_4 \times 2^1/_4$-inch) aluminum box (see **Fig 16**). Two pairs of S0-239 connectors are mounted on opposite sides of the box. A piece of PC board is run diagonally across the box to improve coupler directivity. The pieces of RG-8X coaxial cable pass through holes in the PC board.

(Note: Some brands of "mini 8" cable have extremely low breakdown voltage ratings and are unsuitable to carry even 100 W when the SWR exceeds 1:1. See the subsequent section, "High-Power Operation," for details of a coupler made with RG-8 cable.)

Begin by constructing T1 and T2, which are identi-

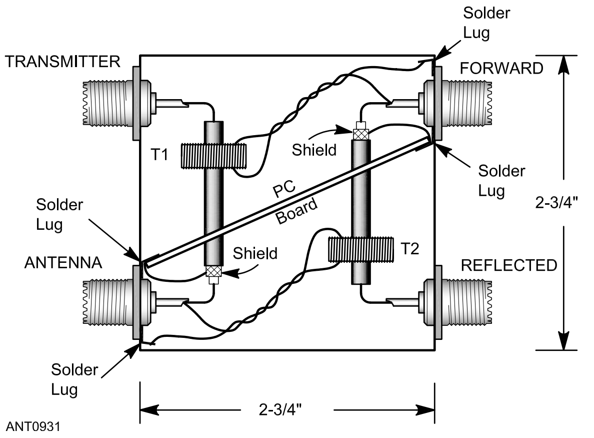

Fig 16—Construction details for the directional coupler.

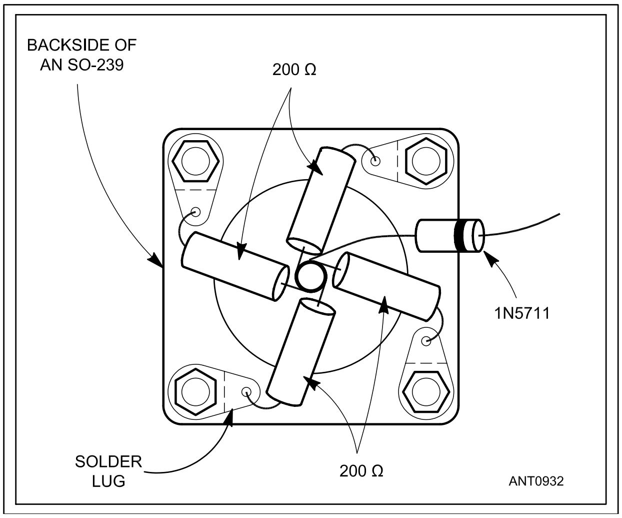

Fig 17—The parallel load resistors mounted on an SO-239 connector. Four 200-Ω, 2%, $^1/_2$-W resistors are mounted in parallel to provide a 50-Ω detector load.

cal except for their end connections. Refer to Fig 16. The primary for each transformer is the center conductor of a length of RG-8X coaxial cable. Cut two cable lengths sufficient for mounting as shown in the figure. Strip the cable jacket, braid and dielectric as shown. The cable braid is used as a Faraday shield between the transformer windings, so it is only grounded at one end. *Important—connect the braid only at one end or the directional-coupler circuit will not work properly!* Wind two transformer secondaries, each 31 turns of #24 enameled wire on an Amidon T50-3 or equivalent powdered-iron core.

Slip each core over one of the prepared cable pieces (including both the shield and the outer insulation). Mount and connect the transformers as shown in Fig 16, with the wire running through separate holes in the copper-clad PC board. The directional coupler can be mounted separately from the rest of the circuitry if desired. If so, use two coaxial cables to carry the forward and reflected-power signals from the directional coupler to the detector inputs. Be aware, however that any losses in the cables will affect power readings.

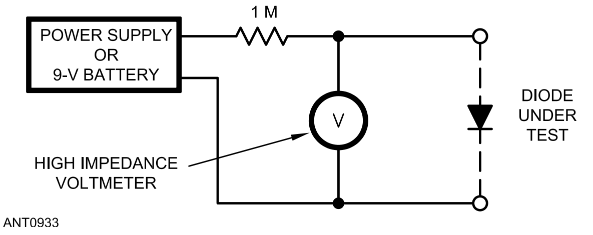

Fig 18—Diode matching test setup.

This directional coupler has not been used at power levels in excess of 100 W. For more information about using the Tandem Match at high power levels, see the section, “High-Power Operation.”

Detector and Signal-Processing Circuits

The detector and signal-processing circuits were constructed on a perforated, copper-clad circuit board. These circuits use two separate grounds—*it is extremely important that the grounds be isolated as shown in the circuit diagram.* Failure to do so may result in faulty circuit operation. Separate grounds prevent RF currents on the cable braid from affecting the op-amp circuitry.

The directional coupler requires good 50-Ω loads. They are constructed on the back of female UHF chassis connectors where the cables from the directional coupler enter the wattmeter housing. Each load consists of four 200-Ω resistors connected from the center conductor of the UHF connector to the four holes on the mounting flange, as shown in **Fig 17**. The detector diode is then run from the center conductor of the connector to the 100-pF and 1000-pF bypass capacitors, which are mounted next to the connector. The response of this load and detector combination measures flat to beyond 500 MHz.

Schottky-barrier diodes (type 1N5711) were used in this design because they were readily available. Any RF-detector diode with a low forward voltage drop (less than 300 mV) and reverse break-down voltage greater than 30 V could be used. (Germanium diodes could be used in this circuit, but performance will suffer. If germanium diodes are used, reduce the resistance values for the detector-diode and feedback-diode load resistors by a factor of 10.)

The detector diodes must be matched. This can be done with dc, using the circuit shown in **Fig 18**. Use a high-impedance voltmeter (10 MΩ or greater). For this project, diodes are matched when their forward voltage drops are equal (within a few millivolts). Diodes from the same batch will probably be sufficiently matched.

Table 2
Range-Switch Resistor Values

Full-Scale Power Level (W)	*Range Resistor (1% Precision) (kΩ)*
1	2.32
2	3.24
3	4.02
5	5.23
10	7.68
15	9.53
20	11.0
25	12.7
30	14.0
50	18.7
100	28.7
150	37.4
200	46.4
250	54 9
300	63.4
500	100.0
1000	237.0
1500	649.0
2000	Open

The rest of the circuit layout is not critical, but keep the lead lengths of the 0.001 and 0.01-pF bypass capacitors short. The capacitors provide additional bypassing for the op-amp circuitry. D6 and D7 form a voltage doubler to detect the presence of a carrier. When the forward power exceeds 1.5 W, Q3 switches on and stays on until about 10 seconds after the carrier drops. (A connection from TP7 to TP9 forces the unit on, even with no carrier present.) The regulated references of +2.5 V and –2.5 V generated by the LM334 and two LM336s are critical. Zener-diode substitutes would significantly degrade performance.

The four op amps in U1 compensate for the nonlinearity of the detector diodes. D1-D2 and D3-D4 are the matched diode pairs discussed above. A RANGE switch selects the meter range. (A six-position switch was used here because it was handy.) The resistor values for the RANGE switch are shown in **Table 2**. Full-scale input power gives an output at U1C or U1D of 7.07 V. The forward and reflected-power detectors are zeroed with R1 and R2.

The forward and reflected-detector voltages are squared by U2, U5 and U6 so that the output voltages are proportional to forward and reflected power. The gain constants are adjusted using R3 and R4 so that an input of 7.07 V to the squaring circuit gives an output of 5 V. The difference between these two voltages is used by U4B to yield an output that is proportional to the power delivered to the transmission line. This voltage is peak detected (by an RC circuit connected to the OPERATE position of the MODE switch) to hold and indicate the maximum power during CW or SSB transmissions. SWR is computed from the forward and reflected voltages by U3, U4 and U7. When no carrier is present, Q4 forces the SWR reading to be zero (that is, when the forward power is less than 2% of the full-scale setting of the RANGE switch). The SWR computation circuit gain is adjusted by R5. The output is peak detected in the OPERATE mode to steady the SWR reading during CW or SSB transmissions.

Transistor arrays (U5, U6 and U7) are used for the log and antilog circuits to guarantee that the transistors will be well matched. Discrete transistors may be used, but accuracy may suffer. A three-position toggle switch selects the three operating modes. In the OPERATE mode, the power and SWR outputs are peak detected and held for a few seconds to allow meter reading during actual transmissions. In the TUNE mode, the meters display instantaneous output power and SWR.

A digital voltmeter is used to obtain more precise readings than are possible with analog meters. The output power range is 0 to 5 V (0 V = 0 W and 5 V = full scale). SWR output varies from 1 V (SWR = 1:1) to 5 V (SWR = 5:1). Voltages above 5 V are unreliable because of voltage limiting in some of the op amp circuits.

Calibration

The directional wattmeter can be calibrated with an accurate voltmeter. All calibration is done with dc voltages. The directional-coupler and detector circuits are inherently accurate if correctly built. To calibrate the wattmeter, use the following procedure:

1) Set the MODE switch to TUNE and the RANGE switch to 100 W or less.
2) Jumper TP7 to TP8. This turns the unit on.
3) Jumper TP1 to TP2. Adjust R1 for 0 V at TP3.
4) Jumper TP4 to TP5. Adjust R2 for 0 V at TP6.
5) Adjust R1 for 7.07 V at TP3.
6) Adjust R3 for 5.00 V at TP9, or a full-scale reading on M1.
7) Adjust R2 for 7.07 V at TP6.
8) Adjust R4 for 0 V at TP9, or a zero reading on M1.
9) Adjust R2 for 4.71 V at TP6.
10) Adjust R5 for 5.00 V at TP10, or a full-scale reading on M2.
11) Set the RANGE switch to its most sensitive scale.
12) Remove the jumpers from TP1 to TP2 and TP4 to TP5.
13) Adjust R1 for 0 V at TP3.
14) Adjust R2 for 0 V at TP6.
15) Remove the jumper from TP7 to TP8.

This completes the calibration procedure. This procedure has been found to equal calibration with expensive laboratory equipment. The directional wattmeter should now be ready for use.

ACCURACY

Performance of the Tandem Match has been compared to other well-known directional couplers and laboratory test equipment, and it equals any amateur directional wattmeter tested. Power measurement accuracy compares well to a Hewlett-Packard HP-436A power meter. The HP meter has a specified measurement error of less than ± 0.05 dB. The Tandem Match tracked the HP436A within +0.5 dB from 10 mW to 100 W, and within ± 0.1 dB from 1 W to 100 W. The unit was not tested above 100 W because a transmitter with a higher power rating was not available.

SWR performance was equally good when compared to the SWR calculated from measurements made with the HP436A and a calibrated directional coupler. The Tandem Match tracked the calculated SWR within ± 5% for SWR values from 1:1 to 5:1. SWR measurements were made at 8 W and 100 W.

OPERATION

Connect the Tandem Match in the 50-Ω line between the transmitter and the antenna matching network (or antenna if no matching network is used). Set the RANGE switch to a range greater than the transmitter output rating and the MODE switch to TUNE. When the transmitter is keyed, the Tandem Match automatically switches on and indicates both power delivered to the antenna and SWR on the transmission line. When no carrier is present, the OUTPUT POWER and SWR meters indicate zero.

The OPERATE mode includes RC circuitry to momentarily hold the peak-power and SWR readings during CW or SSB transmissions. The peak detectors are not ideal, so there could be about 10% variation from the actual power peaks and the SWR reading. The SWR ×10 mode increases the maximum readable SWR to 50:1. This range should be sufficient to cover any SWR value that occurs in amateur use. (A 50-foot open stub of RG-8 yields a measured SWR of only 43:1, or less, at 2.4 MHz because of cable loss. Higher frequencies and longer cables exhibit a lesser maximum SWR.)

It is easy to use the Tandem Match to adjust an antenna matching network: Adjust the transmitter for minimum output power (at least 1.5 W). With the carrier on and the MODE switch set to TUNE or SWR ×10, adjust the matching network for minimum SWR. Once the minimum SWR is obtained, set the transmitter to the proper operating mode and output power. Place the Tandem Match in the OPERATE mode.

DESIGN VARIATIONS

There are several ways in which this design could be enhanced. The most important is to add UHF capability. This would require a new directional-coupler design for the band of interest. (The existing detector circuit should work to at least 500 MHz.)

Those who desire a low-power directional wattmeter can build a directional coupler with a 20-dB coupling factor by decreasing the transformer turns ratio to 10:1. That version should be capable of measuring output power from 1 mW to about 150 W (and it should switch on at about 150 mW).

This change should also increase the maximum operating frequency to about 150 MHz (by virtue of the shorter transformer windings). If you desire 1.8-MHz operation, it may be necessary to change the toroidal core material for sufficient reactance (low insertion loss).

The Tandem Match circuit can accommodate coaxial cable with a characteristic impedance other than 50 Ω. The detector terminating resistors, transformer secondaries and range resistors must change to match the new design impedance.

The detector circuitry can be used (without the directional coupler) to measure low-level RF power in 50-Ω circuits. RF is fed directly to the forward detector (J1, Fig 15), and power is read from the output power meter. The detector is quite linear from 10 µW to 1.5 W.

HIGH-POWER OPERATION

This material was condensed from information by Frank Van Zant, KL7IBA, in July 1989 *QST.* In April 1988, Zack Lau, W1VT, described a directional-coupler circuit (based on the same principle as Grebenkemper's circuit) for a QRP transceiver (see the Bibliography at the end of this chapter). The main advantage of Lau's circuit is a very low parts count.

Grebenkemper used complex log-antilog amplifiers to provide good measurement accuracy. This application gets away from complex circuitry, but retains reasonable measurement accuracy over the 1 to 1500-W range. It also forfeits the SWR-computation feature. Lau's coupler uses ferrite toroids. It works well at low power levels, but the ferrite toroids heat excessively with high power, causing erratic meter readings and the potential for burned parts.

The Revised Design

Powdered-iron toroids are used for the transformers in this version of Lau's basic circuit. The number of turns on the secondaries was increased to compensate for the lower permeability of powdered iron.

Two meters display reflected and forward power (see **Fig 19**). The germanium detector diodes (D1 and D2—1N34) provide fairly accurate meter readings, particularly if the meter is calibrated (using R3, R4 and R5) to place the normal transmitter output at mid scale. If the winding sense of the transformers is reversed, the meters are transposed (the forward-power meter becomes the reflected-power meter, and vice versa).

Construction

Fig 20 shows the physical layout of the coupler. The pickup unit is mounted in a $3^1/_2 \times 3^1/_2 \times$ 4-inch box. The meters, PC-mount potentiometers and HIGH/LOW power

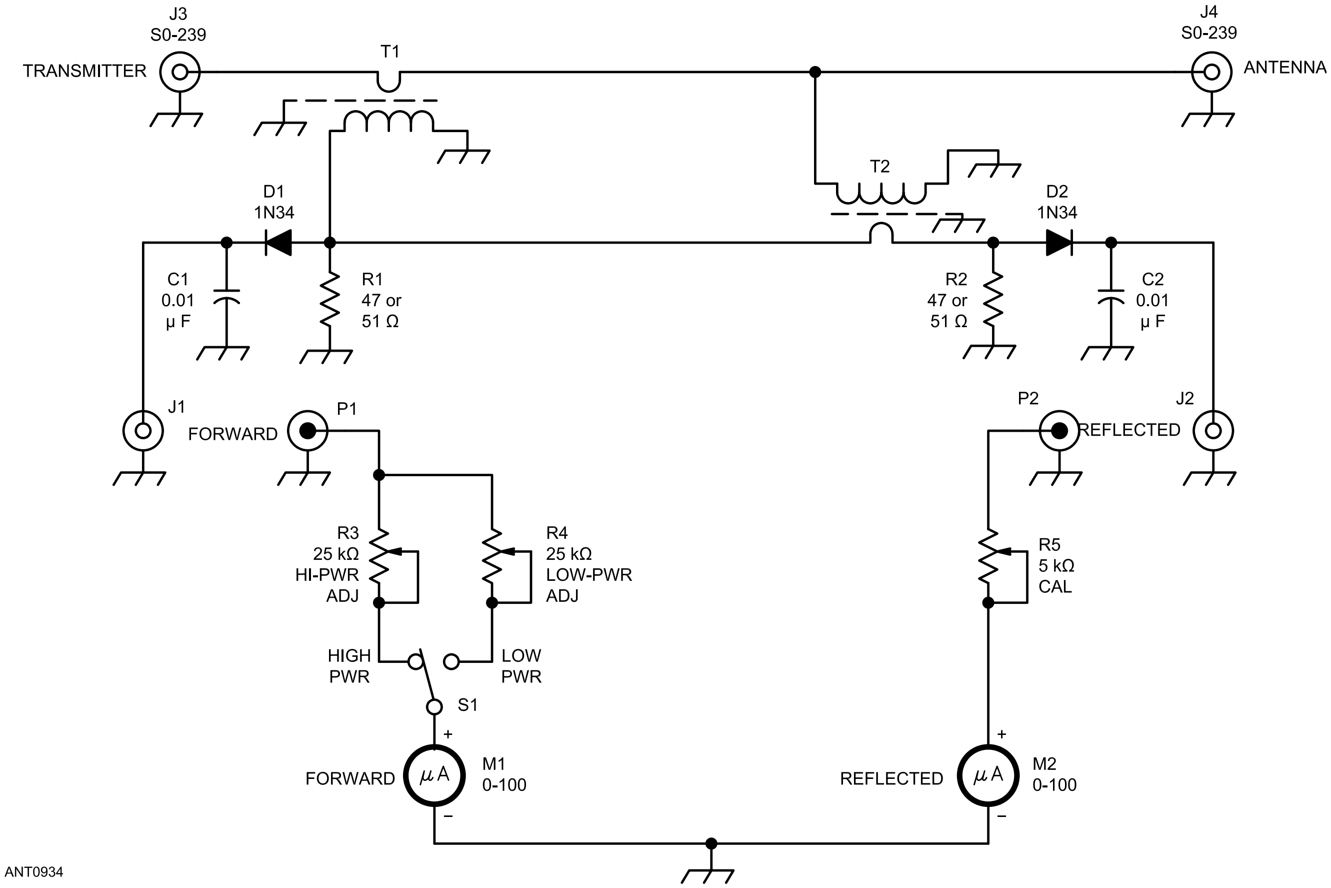
J3
S0-239
TRANSMITTER
T1
J4
S0-239
ANTENNA
T2
D1
1N34
D2
1N34
C1
0.01
μ F
R1
47 or
51 Ω
R2
47 or
51 Ω
C2
0.01
μ F
J1
FORWARD
P1
P2
REFLECTED
J2
R3
25 kΩ
HI-PWR
ADJ
R4
25 kΩ
LOW-PWR
ADJ
R5
5 kΩ
CAL
HIGH
PWR
LOW
PWR
S1
+
FORWARD
μA
M1
0-100
−
+
REFLECTED
μA
M2
0-100
−
ANT0934

Fig 19—Schematic diagram of the high-power directional coupler. D1 and D2 are germanium diodes (1N34 or equiv). R1 and R2 are 47 or 51-Ω, $^1/_2$-W resistors. C1 and C2 have 500-V ratings. The secondary windings of T1 and T2 each consist of 40 turns of #26 to #30 enameled wire on T-68-2 powdered-iron toroid cores. If the coupler is built into an existing antenna tuner, the primary of T1 can be part of the tuner coaxial output line. The remotely located meters (M1 and M2) are connected to the coupler box at J1 and J2 via P1 and P2.

switch are mounted in a separate box or a compartment in an antenna tuner. Parts for this project are available from the suppliers listed in **Table 3**.

The primary windings of Tl and T2 are constructed much as Grebenkemper described, but use RG-8 with its jacket removed so that the core and secondary winding may fit over the cable. The braid is wrapped with fiberglass tape to insulate it from the secondary winding. An excellent alternative to fiberglass tape—with even higher RF voltage-breakdown characteristics—is ordinary plumber's Teflon pipe tape, available at most hardware stores.

The transformer secondaries are wound on T-68-2 powdered-iron toroid cores. They are 40 turns of #26 to #30 enameled wire spread evenly around each core. By using #26 to #30 wire on the cores, the cores slip over the tape-wrapped RG-8 lines. With #26 wire on the toroids, a single layer of tape (slightly more with Teflon tape) over the braid provides an extremely snug fit for the core. Use care when fitting the cores onto the RG-8 assemblies.

After the toroids are mounted on the RG-8 sections,

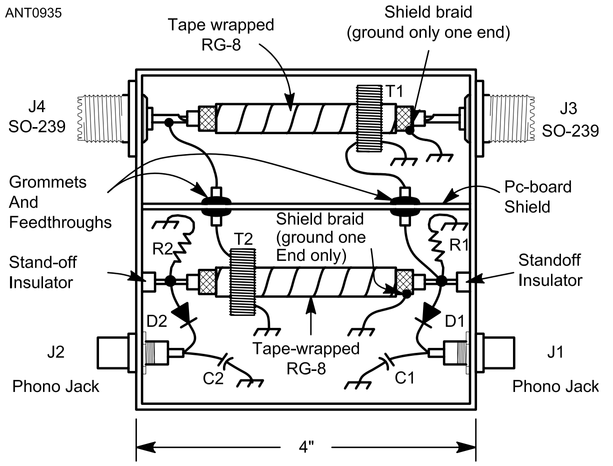
ANT0935
Tape wrapped
RG-8
Shield braid
(ground only one end)
T1
J4
SO-239
J3
SO-239
Grommets
And
Feedthroughs
Pc-board
Shield
Shield braid
(ground one
End only)
R2
T2
R1
Stand-off
Insulator
Standoff
Insulator
D2
D1
J2
Phono Jack
Tape-wrapped
RG-8
C2
C1
J1
Phono Jack
4"

Fig 20—Directional-coupler construction details. Grommets or feedthrough insulators can be used to route the secondary winding of T1 and T2 through the PC board shield. A 3 $^1/_2 \times 3\ ^1/_2 \times$ 4-inch box serves as the enclosure.

Table 3
Parts Sources
(Also see Chapter 21)

Components	*Source*
TLC-series and CA3146 ICs	Newark Electronics 4801 N Ravenswood St Chicago, IL 60640 773-784-5100
LM334, LM336, 1% resistors, trimmer potentiometers	Digi-Key Corporation 701 Brooks Ave S PO Box 677 Thief River Falls, MN 56701 800-344-4539
Toroid cores, Fiberglass tape	Amidon Associates 240 & 250 Briggs Ave Costa Mesa, CA 92626 714-850-4660
Meters	Fair Radio Sales PO Box 1105 Lima, OH 45804 419-227-6573
Toroid cores	Palomar Engineers PO Box 462222 Escondido, CA 92046 760-747-3343
0-150/1500-W-scale meters, A&M model no. 255-138, 1N5711 diodes	Surplus Sales of Nebraska 1502 Jones St Omaha, NE 68102 402-346-4750

coat the assembly with General Cement Corp Polystyrene Q Dope, or use a spot or two of RTV sealant to hold the windings in place and fix the transformers on the RG-8 primary windings.

Mount a PC-board shield in the center of the box, between T1 and T2, to minimize coupling between the transformers. Suspend T1 between the SO-239 connectors and T2 between two standoff insulators. The detector circuits (C1, C2, D1, D2, R1 and R2) are mounted inside the coupler box as shown.

Calibration, Tune Up and Operation

The coupler has excellent directivity. Calibrate the meters for various power levels with an RF ammeter and a 50-Ω dummy load. Calculate I^2R for each power level, and mark the meter faces accordingly. Use R3, R4 and R5 to adjust the meter readings within the ranges. Diode nonlinearities are thus taken into account, and Grebenkemper’s signal-processing circuits are not needed for relatively accurate power readings. Start the tune-up process using about 10 W, adjust the antenna tuner for minimum reflected power, and increase power while adjusting the tuner to minimize reflected power.

This circuit has been built into several antenna tuners with good success. The instrument works well at 1.5-kW output on 1.8 MHz. It also works well from 3.5 to 30 MHz with 1.2 and 1.5-kW output.

The antenna is easily tuned for a 1:1 SWR using the null indication provided. Amplifier settings for a matched antenna, as indicated with the wattmeter, closely agreed with those for a 50-Ω dummy load. Checks with a Palomar noise bridge and a Heath Antenna Scope also verified these findings. This circuit should handle more than 1.5 kW, as long as the SWR on the feed line through the wattmeter is kept at or near 1:1. (On one occasion high power was applied while the antenna tuner was not coupled to a load. Naturally the SWR was extremely high, and the output transformer secondary winding opened like a fuse. This resulted from the excessively high voltage across the secondary. The damage was easily and quickly repaired.)

An Inexpensive VHF Directional Coupler

Precision in-line metering devices capable of reading forward and reflected power over a wide range of frequencies are very useful in amateur VHF and UHF work, but their rather high cost puts them out of the reach of many VHF enthusiasts. The device shown in Figs 14 through 16 is an inexpensive adaptation of their basic principles. It can be made for the cost of a meter, a few small parts, and bits of copper pipe and fittings that can be found in the plumbing stocks at many hardware stores.

Construction

The sampler consists of a short section of handmade coaxial line, in this instance, of 50 Ω impedance, with a reversible probe coupled to it. A small pickup loop built into the probe is terminated with a resistor at one end and a diode at the other. The resistor matches the impedance of the loop, not the impedance of the line section. Energy picked up by the loop is rectified by the diode, and the resultant current is fed to a meter equipped with a calibration control.

The principal metal parts of the device are a brass plumbing T, a pipe cap, short pieces of $^{3}/_{4}$-inch ID and $^{5}/_{16}$-inch OD copper pipe, and two coaxial fittings. Other available tubing combinations for 50-Ω line may be usable. The ratio of outer conductor ID to inner conductor OD should be 2.4/1. For a sampler to be used with other

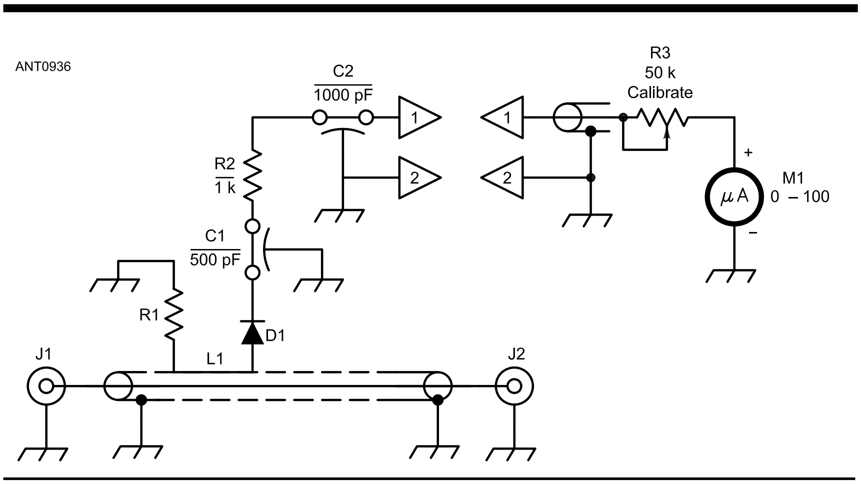

Fig 21—Circuit diagram for the line sampler.

C1—500-pF feedthrough capacitor, solder-in type.
C2—1000-pF feedthrough capacitor, threaded type.
D1—Germanium diode 1N34, 1N60, 1N270, 1N295, or similar.
J1, J2—Coaxial connector, type N (UG-58A).
L1—Pickup loop, copper strap 1-inch long × $^{3}/_{16}$-inch wide. Bend into "C" shape with flat portion $^{5}/_{8}$-inch long.
M1—0-100 μA meter.
R1—Composition resistor, 82 to 100 Ω. See text.
R3—50-kΩ composition control, linear taper.

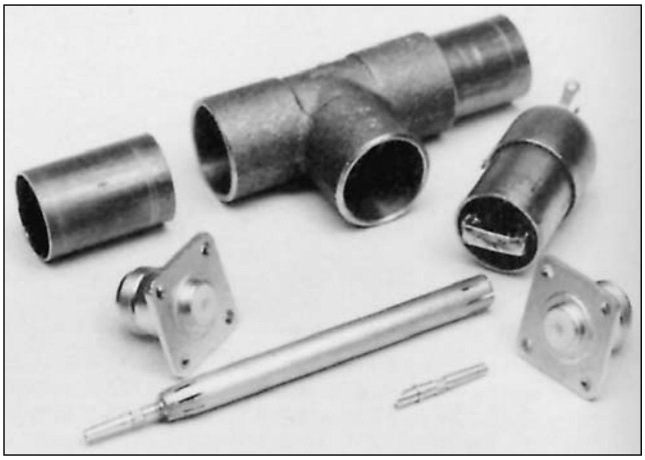

Fig 22—Major components of the line sampler. The brass T and two end sections are at the upper left in this picture. A completed probe assembly is at the right. The N connectors have their center pins removed. The pins are shown with one inserted in the left end of the inner conductor and the other lying in the right foreground.

impedances of transmission line, see Chapter 24 for suitable ratios of conductor sizes. The photographs and **Fig 21** show construction details.

Soldering of the large parts can be done with a 300-watt iron or a small torch. A neat job can be done if the inside of the T and the outside of the pipe are tinned before assembling. When the pieces are reheated and pushed together, a good mechanical and electrical bond will result. If a torch is used, go easy with the heat, as an overheated and discolored fitting will not accept solder well.

Coaxial connectors with Teflon or other heat-resistant insulation are recommended. Type N, with split-ring retainers for the center conductors, are preferred. Pry the split-ring washers out with a knife point or small screwdriver. Don't lose them, as they'll be needed in the final assembly.

The inner conductor is prepared by making eight radial cuts in one end, using a coping saw with a fine-toothed blade, to a depth of $^{1}/_{2}$ inch. The fingers so made are then bent together, forming a tapered end, as shown in **Figs 22** and **23**. Solder the center pin of a coaxial fitting into this, again being careful not to overheat the work.

In preparation for soldering the body of the coax connector to the copper pipe, it is convenient to use a similar fitting clamped into a vise as a holding fixture. Rest the T assembly on top, held in place by its own weight. Use the partially prepared center conductor to assure that the coax connector is concentric with the outer conductor. After being sure that the ends of the pipe are cut exactly perpendicular to the axis, apply heat to the coax fitting, using just enough so a smooth fillet of solder can be formed where the flange and pipe meet.

Before completing the center conductor, check its length. It should clear the inner surface of the connector by the thickness of the split ring on the center pin. File to length; if necessary, slot as with the other end, and solder the center pin in place. The fitting can now be soldered onto the pipe, to complete the 50-Ω line section.

The probe assembly is made from a 1$^{1}/_{2}$ inch length of the copper pipe, with a pipe cap on the top to support the upper feedthrough capacitor, C2. The coupling loop is mounted by means of small Teflon standoffs on a copper disc, cut to fit inside the pipe. The disc has four small tabs around the edge for soldering inside the pipe. The diode, D1, is connected between one end of the loop and

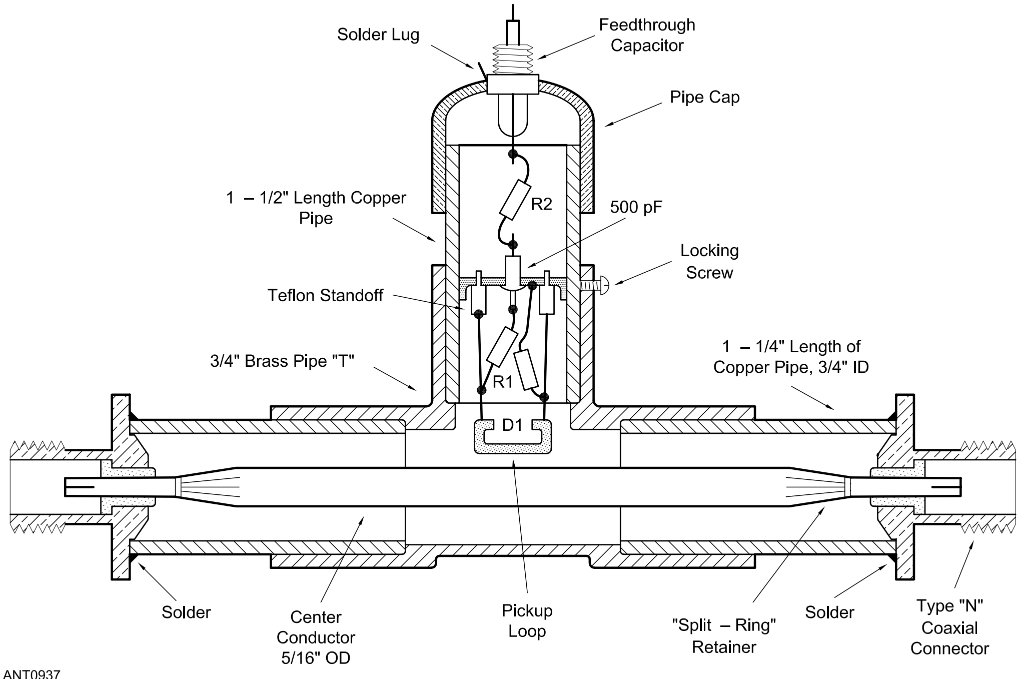
Solder Lug
Feedthrough Capacitor
Pipe Cap
1 – 1/2" Length Copper Pipe
R2
500 pF
Locking Screw
Teflon Standoff
3/4" Brass Pipe "T"
R1
1 – 1/4" Length of Copper Pipe, 3/4" ID
D1
Solder
Center Conductor 5/16" OD
Pickup Loop
"Split – Ring" Retainer
Solder
Type "N" Coaxial Connector
ANT0937

Fig 23—Cross-section view of the line sampler. The pickup loop is supported by two Teflon standoff insulators. The probe body is secured in place with one or more locking screws through holes in the brass T.

a 500-pF feedthrough capacitor, C1, soldered into the disc. The terminating resistor, R1, is connected between the other end of the loop and ground, as directly as possible.

When the disc assembly is completed, insert it into the pipe, apply heat to the outside, and solder the tabs in place by melting solder into the assembly at the tabs. The position of the loop with respect to the end of the pipe will determine the sensitivity of a given probe. For power levels up to 200 watts the loop should extend beyond the face of the pipe about $^{5}/_{32}$ inch. For use at higher power levels the loop should protrude only $^{3}/_{32}$ inch. For operation with very low power levels the best probe position can be determined by experiment.

The decoupling resistor, R2, and feedthrough capacitor, C2, can be connected, and the pipe cap put in place. The threaded portion of the capacitor extends through the cap. Put a solder lug over it before tightening its nut in place. Fasten the cap with two small screws that go into threaded holes in the pipe.

Calibration

The sampler is very useful for many jobs even if it is not accurately calibrated, although it is desirable to calibrate it against a wattmeter of known accuracy. A good 50-Ω VHF dummy load is required.

The first step is to adjust the inductance of the loop, or the value of the terminating resistor, for lowest reflected power reading. The loop is the easier to change. Filing it to reduce its width will increase its impedance. Increasing the cross-section of the loop will lower the impedance, and this can be done by coating it with solder. When the reflected power reading is reduced as far as possible, reverse the probe and calibrate for forward power by increasing the transmitter power output in steps and making a graph of the meter readings obtained. Use the calibration control, R3, to set the maximum reading.

Variations

Rather than to use one sampler for monitoring both forward and reflected power by repeatedly reversing the

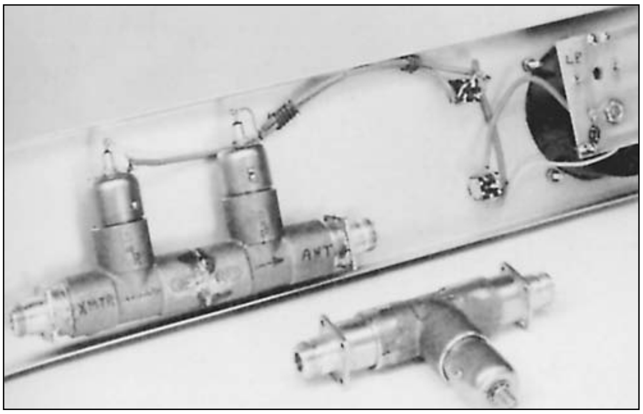

Fig 24—Two versions of the line sampler. The single unit described in detail here is in the foreground. Two sections in a single assembly provide for monitoring forward and reflected power without probe reversal.

probe, it is better to make two assemblies by mounting two T fittings end-to-end, using one for forward and one for reflected power. The meter can be switched between the probes, or two meters can be used.

The sampler described was calibrated at 146 MHz, as it was intended for repeater use. On higher bands the meter reading will be higher for a given power level, and it will be lower for lower frequency bands. Calibration for two or three adjacent bands can be achieved by making the probe depth adjustable, with stops or marks to aid in resetting for a given band. Of course more probes can be made, with each probe calibrated for a given band, as is done in some of the commercially available units.

Other sizes of pipe and fittings can be used by making use of information given in Chapter 24 to select conductor sizes required for the desired impedances. (Since it is occasionally possible to pick up good bargains in 75-Ω line, a sampler for this impedance might be desirable.)

Type-N fittings were used because of their constant impedance and their ease of assembly. Most have the split-ring retainer, which is simple to use in this application. Some have a crimping method, as do apparently all BNC connectors. If a fitting must be used and cannot be taken apart, drill a hole large enough to clear a soldering-iron tip in the copper-pipe outer conductor. A hole of up to $^3/_8$-inch diameter will have very little effect on the operation of the sampler.

A Calorimeter For VHF And UHF Power Measurements

A quart of water in a Styrofoam ice bucket, a roll of small coaxial cable and a thermometer are all the necessary ingredients for an accurate RF wattmeter. Its calibration is independent of frequency. The wattmeter works on the calorimeter principle: A given amount of RF energy is equivalent to an amount of heat, which can be determined by measuring the temperature rise of a known quantity of thermally insulated material. This principle is used in many of the more accurate high-power wattmeters. This procedure was developed by James Bowen, WA4ZRP, and was first described in December 1975 *QST*.

The roll of coaxial cable serves as a dummy load to convert the RF power into heat. RG-174 cable was chosen for use as the dummy load in this calorimeter because of its high loss factor, small size, and low cost. It is a standard 50-Ω cable of approximately 0.11 inch diameter. A prepackaged roll marked as 60 feet long, but measured to be 68 feet, was purchased at a local electronics store. A plot of measured RG-174 loss factor as a function of frequency is shown in **Fig 25**.

In use, the end of the cable not connected to the transmitter is left open-circuited. Thus, at 50 MHz, the reflected wave returning to the transmitter (after making a round trip of 136 feet through the cable) is 6.7 dB × 1.36 = 9.11 dB below the forward wave. A reflected wave 9.11 dB down represents an SWR to the transmitter of 2.08:1. While this value seems larger than would be desired, keep in mind that most 50-MHz transmitters can be tuned to match into an SWR of this magnitude efficiently. To assure accurate results, merely tune the transmitter for maximum power into the load before making

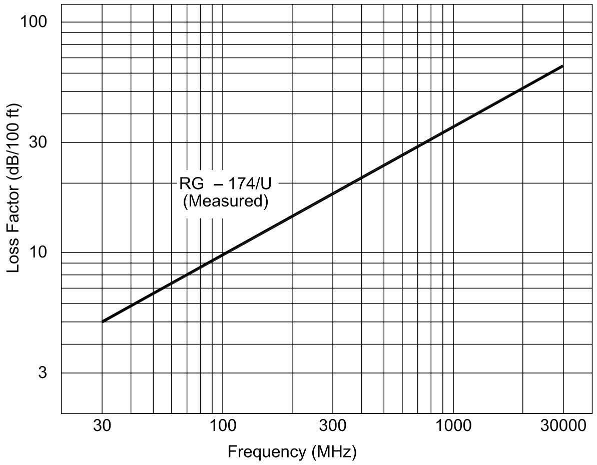

Fig 25—Loss factor of RG-174 coax used in the calorimeter.

Table 4

Calculated Input SWR for 68 Feet of Unterminated RG-174 Cable

Freq. (MHz)	*SWR*
50	2.08
144	1.35
220	1.20
432	1.06
1296	1.003
2304	1.0003

the measurement. At higher frequencies the cable loss increases so the SWR goes down. **Table 4** presents the calculated input SWR values at several frequencies for 68 feet of RG-174. At 1000 MHz and above, the SWR caused by the cable connector will undoubtedly exceed the very low cable SWR listed for these frequencies.

In operation, the cable is submerged in a quart of water and dissipated heat energy flows from the cable into the water, raising the water temperature. See **Fig 26**. The calibration of the wattmeter is based on the physical fact that one calorie of heat energy will raise one gram of liquid water 1° Celsius. Since one quart of water contains 946.3 grams, the transmitter must deliver 946.3 calories of heat energy to the water to raise its temperature 1° C. One calorie of energy is equivalent to 4.186 joules and a joule is equal to 1 W for 1 second. Thus, the heat capacitance of 1 quart of water expressed in joules is 946.3 × 4.186 = 3961 joules/° C.

The heat capacitance of the cable is small with respect to that of the water, but nevertheless its effect should be included for best accuracy. The heat capacitance of the cable was determined in the manner described below. The 68-foot roll of RG-174 cable was raised to a uniform temperature of 100° C by immersing it in a pan of boiling water for several minutes. A quart of tap water was poured into the Styrofoam ice bucket and its temperature was measured at 28.7° C. the cable was then transferred quickly from the boiling water to the water in the ice bucket. After the water temperature in the ice bucket had ceased to rise, it measured 33.0° C. Since the total heat gained by the quart of water was equal to the total heat lost from the cable, we can write the following equation:

$$(\Delta T_{WATER})(C_{WATER}) = -(\Delta T_{CABLE})(C_{CABLE})$$

where

ΔT_{WATER} = the change in water temperature

C_{WATER} = the water heat capacitance

Δ_{CABLE} = the change in cable temperature

C_{CABLE} = the cable heat capacitance

Substituting and solving:

$$(33.0 - 28.7)(3961) = -(33.0 - 100)(C_{CABLE})$$

Thus, the total heat capacitance of the water and cable in the calorimeter is 3961 + 254 = 4215 joules/º C. Since 1º F = 5/9º C, the total heat capacitance can also be expressed as 4215 × 5/9 = 2342 joules/º F.

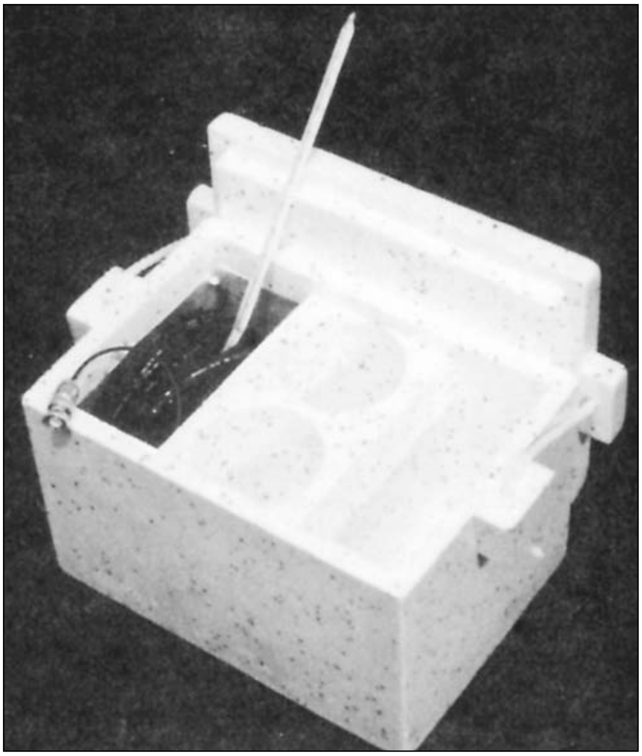

Fig 26—The calorimeter ready for use. The roll of coaxial cable is immersed in one quart of water in the left-hand compartment of the Styrofoam container. Also shown is the thermometer, which doubles as a stirring rod.

Materials and Construction

The quart of water and cable must be thermally insulated to assure that no heat is gained from or lost to the surroundings. A Styrofoam container is ideal for this purpose since Styrofoam has a very low thermal conductivity and a very low thermal capacitance. A local variety store was the source of a small Styrofoam cold chest with compartments for carrying sandwiches and drink cans. The rectangular compartment for sandwiches was found to be just the right size for holding the quart of water and coax.

The thermometer can be either a Celsius or Fahrenheit type, but try to choose one that has divisions for each degree spaced wide enough so that the temperature can be estimated readily to one-tenth degree. Photographic supply stores carry darkroom thermometers, which are ideal for this purpose. In general, glass bulb thermometers are more accurate than mechanical dial-pointer types.

The RF connector on the end of the cable should be a constant-impedance type. A BNC type connector especially designed for use on 0.11-inch diameter cable was located through surplus channels. If you cannot locate one of these, wrap plastic electrical tape around the cable near its end until the diameter of the tape wrap is the same as that of RG-58. Then connect a standard BNC connector for RG-58 in the normal fashion. Carefully seal the opposite open end of the cable with plastic tape or silicone caulking compound so no water can leak into the cable at this point.

Procedure for Use

Pour 1 quart of water (4 measuring cups) into the Styrofoam container. As long as the water temperature is not very hot or very cold, it is unnecessary to cover the top of the Styrofoam container during measurements. Since the transmitter will eventually heat the water several degrees, water initially a few degrees cooler than air temperature is ideal because the average water temperature will very nearly equal the air temperature and heat transfer to the air will be minimized.

Connect the RG-174 dummy load to the transmitter through the shortest possible length of lower loss cable such as RG-8. Tape the connectors and adapter at the RG-8 to RG-174 joint carefully with plastic tape to prevent water from leaking into the connectors and cable at this point. Roll the RG-174 into a loose coil and submerge it in the water. Do not bind the turns of the coil together in any way, as the water must be able to freely circulate among the coaxial cable turns. All the RG-174 cable must be submerged in the water to ensure sufficient cooling. Also submerge part of the taped connector attached to the RG-174 as an added precaution.

Upon completing the above steps, quickly tune up the transmitter for maximum power output into the load. Cease transmitting and stir the water slowly for a minute or so until its temperature has stabilized. Then measure the water temperature as precisely as possible. After the initial temperature has been determined, begin the test transmission, measuring the total number of seconds of key-down time accurately. Stir the water slowly with the thermometer and continue transmitting until there is a significant rise in the water temperature, say 5° to 10°. The test may be broken up into a series of short periods, as long as you keep track of the total key-down time. When the test is completed, continue to stir the water slowly and monitor its temperature. When the temperature ceases to rise, note the final indication as precisely as possible.

To compute the transmitter power output, multiply the calorimeter heat capacitance (4215 for C or 2342 for F) by the difference in initial and final water temperature. Then divide by the total number of seconds of key-down time. The resultant is the transmitter power in watts. A nomogram that can also be used to find transmitter power output is given in **Fig 27**. With a straight line, connect the total number of key-down seconds in the time column to the number of degrees change (F or C) in the temperature rise column, and read off the transmitter power output at the point where the straight line crosses the power-output column.

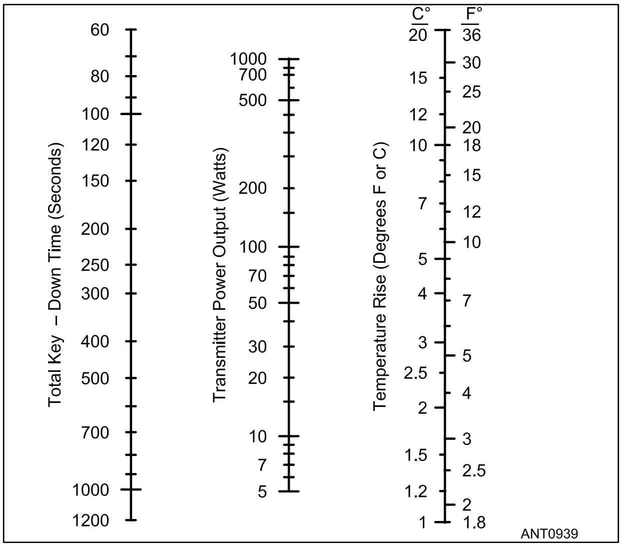

Fig 27—Nomogram for finding transmitter power output for the calorimeter.

Power Limitation

The maximum power handling capability of the calorimeter is limited by the following. At very high powers the dielectric material in the coaxial line will melt because of excessive heating or the cable will arc over from excessive voltage. As the transmitter frequency gets higher, the excessive-heating problem is accentuated, as more of the power is dissipated in the first several feet of cable. For instance, at 1296 MHz, approximately 10% of the transmitting power is dissipated in the first foot of cable. Overheating can be prevented when working with high power by using a low duty cycle to reduce the average dissipated power. Use a series of short transmissions, such as two seconds on, ten seconds off. Keep count of the total key-down time for power calculation purposes. If the cable arcs over, use a larger-diameter cable, such as RG-58, in place of the RG-174. The cable should be long enough to assure that the reflected wave will be down 10 dB or more at the input. It may be necessary to use more than one quart of water in order to submerge all the cable conveniently. If so, be sure to calculate the new value of heat capacitance for the larger quantity of water. Also you should measure the new coaxial cable heat capacitance using the method previously described.

A Noise Bridge For 1.8 Through 30 MHz

The noise bridge, sometimes referred to as an antenna (RX) noise bridge, is an instrument for measuring the impedance of an antenna or other electrical circuits. The unit shown here in **Fig 28**, designed for use in the 1.8 through 30-MHz range, provides adequate accuracy for most measurements. Battery operation and small physical size make this unit ideal for remote-location use. Tone modulation is applied to the wide-band noise generator as an aid for obtaining a null indication. A detector, such as the station receiver, is required for operation.

The noise bridge consists of two parts—the noise generator and the bridge circuitry. See **Fig 29**. A 6.8-V

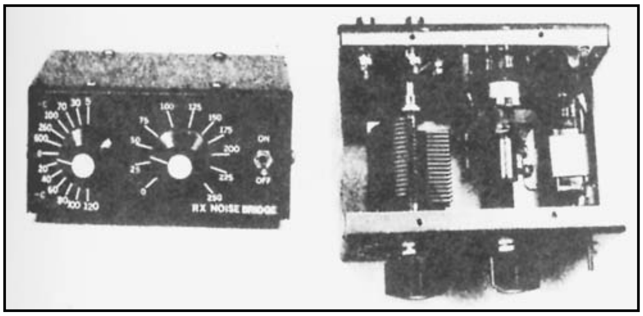
RX NOISE BRIDGE
ON
OFF

Fig 28—Exterior and interior views of the noise bridge. The unit is finished in red enamel. Press-on lettering is used for the calibration marks. Note that the potentiometer must be isolated from ground.

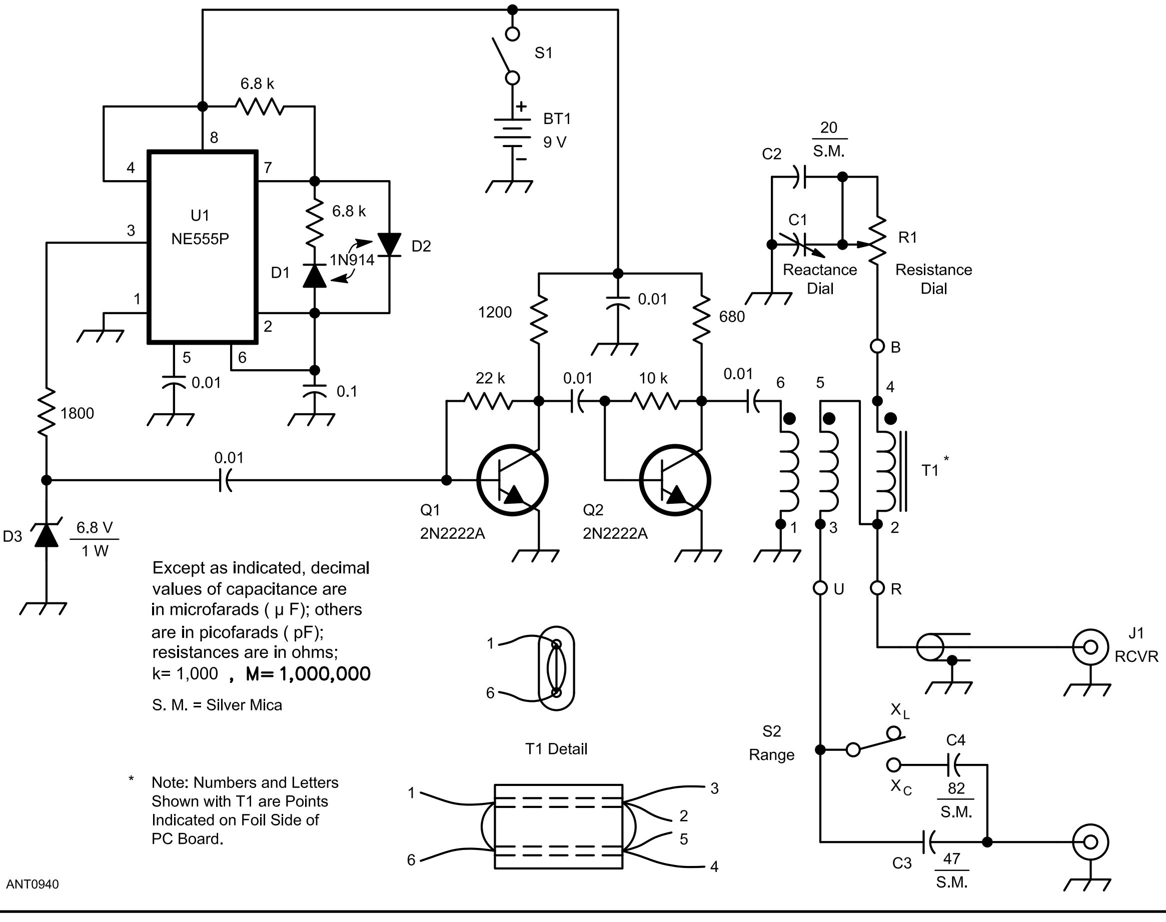
S1
BT1
9 V
6.8 k
8
4
7
3
U1
NE555P
6.8 k
1N914
D1
D2
1
2
5
6
0.01
0.1
1800
0.01
D3
6.8 V
1 W
20
S.M.
C2
C1
R1
Reactance Dial
Resistance Dial
1200
0.01
680
22 k
0.01
10 k
0.01
6
5
4
B
T1 *
Q1
2N2222A
Q2
2N2222A
1
3
2
U
R
J1
RCVR
Except as indicated, decimal values of capacitance are in microfarads (µ F); others are in picofarads (pF); resistances are in ohms; k= 1,000 , M= 1,000,000
S. M. = Silver Mica
1
6
T1 Detail
S2
Range
X_L
X_C
C4
82
S.M.
C3
47
S.M.
* Note: Numbers and Letters Shown with T1 are Points Indicated on Foil Side of PC Board.
1
6
3
2
5
4
ANT0940

Fig 29—Schematic diagram of the noise bridge. Use $^1/_4$ W composition resistors. Capacitors are miniature ceramic units unless indicated otherwise. Component designations indicated in the schematic but not called out in the parts list are for text and parts-placement reference only.

BT1—9-V battery, NEDA 1604A or equiv.
C1—15- to 150-pF variable
C2—20-pF mica.
C3—47-pF mica.
C4—82-pF mica.
J1, J2—Coaxial connector.
R1—Linear, 250 Ω, AB type. Use a good grade of resistor.
S1, S2—Toggle, SPST.
T1—Transformer; 3 windings on an Amidon BLN-43-2402 ferrite binocular core. Each winding is three turns of #30 enameled wire. One turn is equal to the wire passing once through both holes in the core. The primary winding starts on one side of the transformer, and the secondary and tertiary windings start on the opposite side.
U1—Timer, NE555 or equiv.

Zener diode serves as the noise source. U1 generates an approximate 50% duty cycle, 1000-Hz square wave signal which is applied to the cathode of the Zener diode. The 1000-Hz modulation appears on the noise signal and provides a useful null detection enhancement effect. The broadband-noise signal is amplified by Q1, Q2 and associated components to a level that produces an approximate S9 signal in the receiver. Slightly more noise is available at the lower end of the frequency range, as no frequency compensation is applied to the amplifier. Roughly 20 mA of current is drawn from the 9-V battery, thus ensuring long battery life—providing the power is switched off after use!

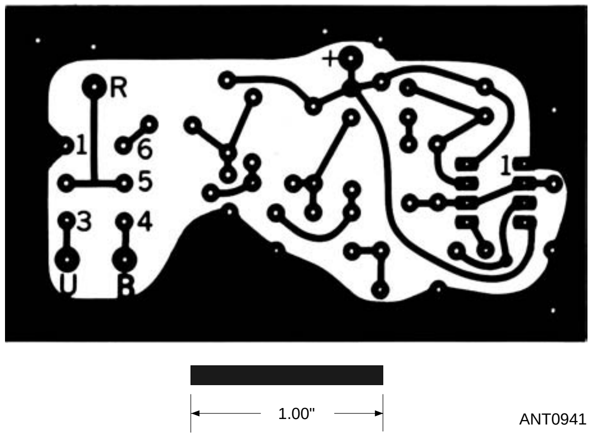

Fig 30—Etching pattern for the noise bridge PC board, at actual size. Black represents copper. This is the pattern for the bottom side of the board. The top side of the board is a complete ground plane with a small amount of copper removed from around the component holes.

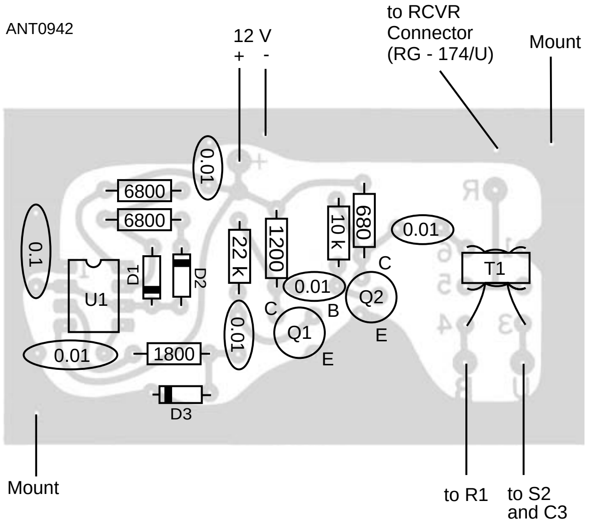

Fig 31—Parts-placement guide for the noise bridge as viewed from the component or top side of the board. Mounting holes are located in two corners of the board, as shown.

The bridge portion of the circuit consists of T1, C1, C2 and R1. T1 is a trifilar wound transformer with one of the windings used to couple noise energy into the bridge circuit. The remaining two windings are arranged so that each one is in an arm of the bridge. C1 and R1 complete one arm and the UNKNOWN circuit, along with C2, comprise the remainder of the bridge. The terminal labeled RCVR is for connection to the detector.

The reactance range of a noise bridge is dependent on several factors, including operating frequency, value of the series capacitor (C3 or C3 plus C4 in Fig 29) and the range of the variable capacitor (C1 in Fig 29). The RANGE switch selects reactance measurements weighted toward either capacitance or inductance by placing C4 in parallel with C3. The zero-reactance point occurs when C1 is either nearly fully meshed or fully unmeshed. The RANGE switch nearly doubles the resolution of the reactance readings.

CONSTRUCTION

The noise bridge is contained in a homemade aluminum enclosure that measures $5 \times 2\frac{3}{8} \times 3\frac{3}{4}$ inches. Many of the circuit components are mounted on a circuit board that is fastened to the rear wall of the cabinet. The circuit-board layout is such that the lead lengths to the board from the bridge and coaxial connectors are at a minimum. An etching pattern and a parts-placement guide for the circuit board are shown in **Figs 30** and **31**.

Care must be taken when mounting the potentiometer, R1. For accurate readings the potentiometer must be well insulated from ground. In the unit shown this was accomplished by mounting the control on a piece of plexiglass, which in turn was fastened to the chassis with a piece of aluminum angle stock.

Additionally, a $\frac{1}{4}$-inch control-shaft coupling and a length of phenolic rod were used to further isolate the control from ground where the shaft passes through the front panel. A high-quality potentiometer is required if good measurement results are to be obtained.

There is no such problem when mounting the variable capacitor because the rotor is grounded. Use a high-quality capacitor; do not try to save money on that component. Two RF connectors on the rear panel are connected to a detector (receiver) and to the UNKNOWN circuit. Do not use plastic-insulated phono connectors (they might influence bridge accuracy at higher frequencies). Use miniature coaxial cable (RG-174) between the RCVR connector and circuit board. Attach one end of C3 to the circuit board and the other directly to the UNKNOWN circuit connector.

Bridge Compensation

Stray capacitance and inductance in the bridge circuit can affect impedance readings. If a very accurate bridge is required, use the following steps to counter the effects of stray reactance. Because the physical location

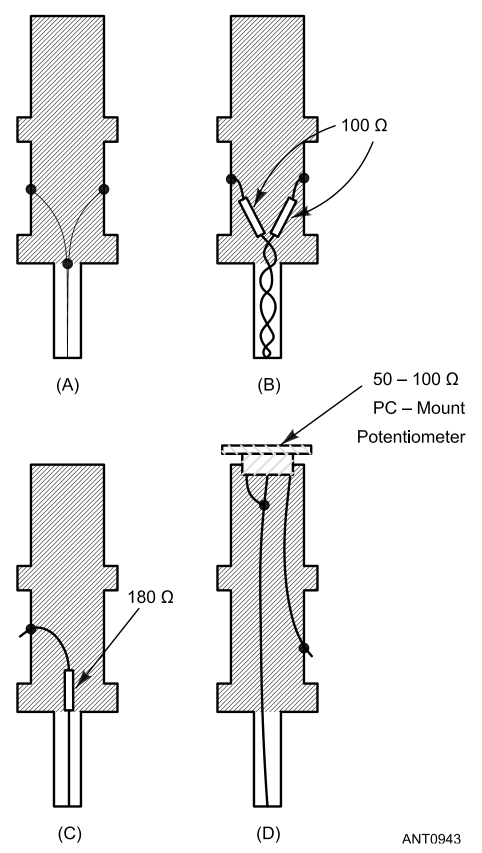

Fig 32—Construction details of the resistive loads used to check and calibrate the noise bridge. Each of the loads is constructed inside a coaxial connector that matches those on the bridge. (Views shown are cross-sections of PL-259 bodies; the sleeves are not shown.) Leads should be kept as short as possible to minimize parasitic inductance. A is a 0-Ω load; B depicts a 50-Ω load; C is a 180-Ω load; D shows a variable-resistance load used to determine the loss in a coaxial cable.

of the board, connectors and controls in the cabinet determine where compensation is needed, there is no provision for the compensation components on the printed circuit board.

Good calibration loads are necessary to check the accuracy of the noise bridge. Four are needed here: a 0-Ω (short-circuit) load, a 50-Ω load, a 180-Ω load, and a variable-resistance load. The short-circuit and fixed-resistance loads are used to check the accuracy of the noise bridge; the variable-resistance load is used when measuring coaxial-cable loss.

Construction details of the loads are shown in **Fig 32**. Each load is constructed inside a connector. When building the loads, keep leads as short as possible to minimize parasitic effects. The resistors must be noninductive (*not* wirewound).

Quarter-watt, carbon-composition resistors should work fine. The potentiometer in the variable-resistance load is a miniature PC-mount unit with a maximum resistance of 100 Ω or less. The potentiometer wiper and one of the end leads are connected to the center pin of the connector; the other lead is connected to ground.

Stray Capacitance

Stray capacitance on the variable-resistor side of the bridge tends to be higher than that on the unknown side. This is so because the parasitic capacitance in the variable resistor, R1, is comparatively high.

The effect of parasitic capacitance is most easily detected using the 180-Ω load. Measure and record the actual resistance of the load, R_L. Connect the load to the UNKNOWN connector, place S2 in the X_L position, tune the receiver to 1.8 MHz, and null the bridge. (See the section, "Finding the Null" for tips.) Use an ohmmeter across R1 to measure its dc resistance. The magnitude of the stray capacitance can be calculated by

$$C_P = C3\left(\sqrt{\frac{R1}{R2}}-1\right) \qquad \text{(Eq 4)}$$

where

R_L = load resistance (as measured)
R1 = resistance of the variable resistor
C3 = series capacitance.

You can compensate for C_P by placing a variable capacitor, C_C, in the side of the bridge with lesser stray capacitance. If R1 is greater than R_L, stray capacitance is greater on the variable resistor side of the bridge: Place C_C between point U (on the circuit board) and ground. If R1 is less than R_L, stray capacitance is greater on the unknown side: Place C_C between point B and ground. If the required compensating capacitance is only a few picofarads, you can use a gimmick capacitor (made by twisting two short pieces of insulated, solid wire together) for C_C. A gimmick capacitor is adjusted by trimming its length.

Stray Inductance

Parasitic inductance, if present, should be only a few tens of nanohenries. This represents a few ohms of inductive reactance at 30 MHz. The effect is best observed by reading the reactance of the 0-Ω test load at 1.8 and 30 MHz; the indicated reactance should be the same at both frequencies.

If the reactance reading decreases as frequency is increased, parasitic inductance is greater in the known arm, and compensating inductance is needed between point U and C3. If the reactance increases with frequency, the unknown-arm inductance is greater, and compensating inductance should be placed between point B and R1.

Compensate for stray inductance by placing a single-turn coil, made from a 1 to 2-inch length of solid wire, in the appropriate arm of the bridge. Adjust the size of this coil until the reactance reading remains constant from 1.8 to 30 MHz.

Calibration

Good calibration accuracy is necessary for accurate noise-bridge measurements. Calibration of the resistance scale is straightforward. To do this, tune the receiver to a frequency near 10 MHz. Attach the 0-Ω load to the UNKNOWN connector and null the bridge. This is the zero-resistance point; mark it on the front-panel resistance scale. The rest of the resistance range is calibrated by adjusting R1, measuring R1 with an accurate ohmmeter, calculating the increase from the zero point and marking the increase on the front panel.

Most bridges have the reactance scale marked in capacitance because capacitance does not vary with frequency. Unfortunately, that requires calibration curves or non-trivial calculations to arrive at the load reactance. An alternative method is to mark the reactance scale in *ohms* at a reference frequency of 10 MHz. This method calibrates the bridge near the center of its range and displays reactance directly, but it requires a simple calculation to scale the reactance reading for frequencies other than 10 MHz. The scaling equation is:

$$X_{u(f)} = X_{u(10)}\frac{10}{f} \qquad \text{(Eq 5)}$$

where

f = frequency in MHz

$X_{u(10)}$ = reactance of the unknown load at 10 MHz

$X_{u(f)}$ = reactance of the unknown load at f.

A shorted piece of coaxial cable serves as a reactance source. (The reactance of a shorted, low-loss coaxial cable is dependent only on the cable length, the measurement frequency and the cable characteristic impedance.) Radio Shack RG-8M is used here because it is readily available, has relatively low loss and has an almost purely resistive characteristic impedance.

Prepare the calibration cable as follows:

1) Cut a length of coaxial cable that is slightly longer than $^1/_4$ λ at 10 MHz (about 20 feet for RG-8M). Attach a suitable connector to one end of the cable; leave the other end open-circuited.
2) Connect the 0-Ω load to the noise bridge UNKNOWN connector and set the receiver frequency to 10 MHz. Adjust the noise bridge for a null. Do not adjust the reactance control after the null is found.
3) Connect the calibration cable to the bridge UNKNOWN terminal. Null the bridge by adjusting *only* the variable resistor and the receiver frequency. The receiver frequency should be less than 10 MHz; if it is above 10 MHz, the cable is too short, and you need to prepare a longer one.
4) Gradually cut short lengths from the end of the coaxial cable until you obtain a null at 10 MHz by adjusting only the resistance control. Then connect the cable center and shield conductors at the open end with a short length of braid. Verify that the bridge nulls with zero reactance at 20 MHz.
5) The reactance of the coaxial cable (normalized to 10 MHz) can be calculated from:

$$X_{i(10)} = R_0\frac{f}{10}\tan\left(2\pi\frac{f}{40}\right) \qquad \text{(Eq 6)}$$

where

$X_{i(10)}$ = cable reactance at 10 MHz

R_0 = characteristic resistance of the coaxial cable (52.5 Ω for Radio Shack RG-8M)

f = frequency in MHz

The results of Eq 6 have less than 5% error for reactances less than 500 Ω, so long as the test-cable loss is less than 0.2 dB. This error becomes significantly less at lower reactances (2% error at 300 Ω for a 0.2-dB-loss cable). The loss in 18 feet of RG-8M is 0.13 dB at 10 MHz. Reactance data for Radio Shack RG-8M is given in **Table 5**.

Table 5

Noise Bridge Calibration Data: Coaxial-Cable Method

This data is for Radio Shack RG-8M cable (R_0 =52.5 Ω) cut to exactly $^1/_4$ λ at 10 MHz; the reactances and capacitances shown correspond to this frequency.

Reactance				**Capacitance**	
X_i	*f(MHz)*	*X_i*	*f(MHz)*	*C(pF)*	*f(MHz)*
10	3.318	–10	19.376	10	9.798
20	4.484	–20	18.722	20	9.612
30	5.262	–30	18.048	30	9.440
40	5.838	–40	17.368	40	9.280
50	6.286	–50	16.701	50	9.130
60	6.647	–60	16.062	60	8.990
70	6.943	–70	15.471	70	8.859
80	7.191	–80	14.936	80	8.735
90	7.404	–90	14.462	90	8.618
100	7.586	–100	14.044	100	8.508
110	7.747	–110	13.682	110	8.403
120	7.884	–120	13.369	120	8.304
130	8.009	–130	13.097	130	8.209
140	8.119	–140	12.861	140	8.119
150	8.217	–150	12.654		
160	8.306	–160	12.473		
170	8.387	–170	12.313		
180	8.460	–180	12.172		
190	8.527	–190	12.045		
200	8.588	–200	11.932	*C(pF)*	*f(MHz)*
210	8.645	–210	11.831	–10	10.219
220	8.697	–220	11.739	–20	10.459
230	8.746	–230	11.655	–30	10.721
240	8.791	–240	11.579	–40	11.010
250	8.832	–250	11.510	–50	11.328
260	8.872	–260	11.446	–60	11.679
270	8.908	–270	11.387	–70	12.064
280	8.942	–280	11.333	–80	12.484
290	8.975	–290	11.283	–90	12.935
300	9.005	–300	11.236	–100	13.407
350	9.133	–350	11.045	–110	13.887
400	9.232	–400	10.905	–120	14.357
450	9.311	–450	10.798	–130	14.801
500	9.375	–500	10.713	–140	15.211

With the prepared cable and calibration values on hand, proceed to calibrate the reactance scale. Tune the receiver to the appropriate frequency for the desired reactance (given in Table 5, or found using Eq 6). Adjust the resistance and reactance controls to null the bridge. Mark the reactance reading on the front panel. Repeat this process until all desired reactance values have been marked. The resistance values needed to null the bridge during this calibration procedure may be significant (more than 100 Ω) at the higher reactances.

This calibration method is much more accurate than using fixed capacitors across the UNKNOWN connector. Also, you can calibrate a noise bridge in less than an hour using this method.

Finding the Null

In use, a receiver is attached to the RCVR connector and some load of unknown value is connected to the UNKNOWN terminal. The receiver allows us to hear the noise present across the bridge arms at the frequency of the receiver passband. The strength of the noise signal depends on the strength of the noise-bridge battery, the receiver bandwidth/sensitivity and the impedance difference between the known and unknown bridge arms. The noise is stronger and the null more obvious with wide receiver passbands. Set the receiver to the widest bandwidth AM mode available.

The noise-bridge output is heard as a 1000-Hz tone. When the impedances of the known and unknown bridge arms are equal, the voltage across the receiver is minimized; this is a null. In use, the null may be difficult to find because it appears only when both bridge controls approach the values needed to balance the bridge.

To find the null, set C1 to mid-scale, sweep R1 slowly through its range and listen for a reduction in noise (it's also helpful to watch the S meter). If no reduction is heard, set R1 to mid-range and sweep C1. If there is still no reduction, begin at one end of the C1 range and sweep R1. Increment C1 about 10% and sweep R1 with each increment until some noise reduction appears. Once noise reduction begins, adjust C1 and R1 alternately for minimum signal.

MEASURING COAXIAL-CABLE PARAMETERS WITH A NOISE BRIDGE

Coaxial cables have a number of properties that affect the transmission of signals through them. Generally, radio amateurs are concerned with cable attenuation and characteristic impedance. If you plan to use a noise bridge or SWR analyzer to make antenna-impedance measurements, however, you need to accurately determine not just cable impedance and attenuation, but also electrical length. Fortunately, all of these parameters are easy to measure with an accurate noise bridge or SWR analyzer.

Cable Electrical Length

With a noise bridge and a general-coverage receiver, you can easily locate frequencies at which the line in question is a multiple of $^1/_2\ \lambda$, because a shorted $^1/_2\ \lambda$ line has a 0-Ω impedance (neglecting line loss). By locating two adjacent null frequencies, you can solve for the length of line in terms of $^1/_2\ \lambda$ at one of the frequencies and calculate the line length (overall accuracy is limited by bridge accuracy and line loss, which broadens the nulls). As an interim variable, you can express cable length as the frequency at which a cable is 1 λ long. This length will be represented by f λ. Follow these steps to determine f λ for a coaxial cable.

1) Tune the receiver to the frequency range of interest. Attach the short-circuit load to the noise bridge UNKNOWN connector and null the bridge.
2) Disconnect the far end of the coaxial cable from its load (the antenna) and connect it to the 0-Ω test load. Connect the near end of the cable to the bridge UNKNOWN connector.
3) Adjust the receiver frequency and the noise-bridge resistance control for a null. *Do not change the noise bridge reactance-control setting during this procedure*. Note the frequency at which the null is found; call this frequency f_n. The noise-bridge resistance at the null should be relatively small (less than 20 Ω).
4) Tune the receiver upward in frequency until the next null is found. Adjust the resistance control, if necessary, to improve the null, *but do not adjust the reactance control*. Note the frequency at which this second null is found; this is f_{n+2} .
5) Solve Eq 7 for n and the electrical length of the cable.

$$n = \frac{2f_n}{f_{n+2} - f_n} \qquad \text{(Eq 7)}$$

$$f_\lambda = \frac{4f_n}{n} \qquad \text{(Eq 8)}$$

where

n = cable electrical length in quarter waves, at f_n

f_l = frequency at which the cable is 1λ

ℓ = cable electrical length, in λ

For example, consider a 74-foot length of Columbia 1188 foam-dielectric cable (velocity factor = 0.78) to be used on the 10-meter band. Based on the manufacturer's specification, the cable is 2.796 λ at 29 MHz. Nulls were found at 24.412 (f_n) and 29.353 (f_{n+2}) MHz. Eq 7 yields n = 9.88, which produces 9.883 MHz from Eq 8 and 2.934 l for Eq 9. If the manufacturer's specification is correct, the measured length is off by less than 5%, which is very reasonable. Ideally, n would yield an integer. The difference between n and the closest integer indicates that there is some error.

This procedure also works for lines with an open circuit as the termination (n will be close to an odd number). End effects from the PL-259 increase the effective length of the coaxial cable; however, this decreases the calculated f_λ.

Cable Characteristic Impedance

The characteristic impedance of the coaxial cable is found by measuring its input impedance at two frequencies separated by $^1/_4\ f_\lambda$. This must be done when the cable is terminated in a resistive load.

Characteristic impedance changes slowly as a function of frequency, so this measurement must be done near the frequency of interest. The measurement procedure is as follows.

1) Place the 50-Ω load on the far end of the coaxial cable and connect the near end to the UNKNOWN connector of the noise bridge. (Measurement error is minimized when the load resistance is close to the characteristic impedance of the cable. This is the reason for using the 50-Ω load.)
2) Tune the receiver approximately $^1/_8$ f below the frequency of interest. Adjust the bridge resistance and reactance controls to obtain a null, and note their readings as R_{f1} and X_{f1}. Remember, the reactance reading must be scaled to the measurement frequency.
3) Increase the receiver frequency by exactly $^1/_4\ f\lambda$. Null the bridge again, and note the readings as R_{f2} and X_{f2}.
4) Calculate the characteristic impedance of the coaxial cable using Eqs 10 through 15. A scientific calculator is helpful for this.

$$\ell = \frac{f_0}{f_\lambda} \qquad \text{(Eq 9)}$$

$$R = R_{f1} \times F_{f2} - X_{f1} \times X_{f2} \qquad \text{(Eq 10)}$$

$$X = R_{f1} \times X_{f2} + X_{f1} \times R_{f2} \qquad \text{(Eq 11)}$$

$$Z = \sqrt{R^2 + X^2} \qquad \text{(Eq 12)}$$

$$R_0 = \sqrt{Z} \cos\left[\frac{1}{2} \tan^{-1}\left(\frac{X}{R}\right)\right] \qquad \text{(Eq 13)}$$

$$X_0 = \sqrt{Z} \sin\left[\frac{1}{2} \tan_{-1}\left(\frac{X}{R}\right)\right] \qquad \text{(Eq 14)}$$

$$Z_0 = R_0 + jX_0 \qquad \text{(Eq 15)}$$

Let's continue with the example used earlier for cable length. The measurements are:

f1 = 29.000 – (9.883 / 8) = 27.765 MHz
R_{f1} = 64 Ω
X_{f1} = –22 Ω × (10 / 27.765) = –7.9 Ω
f2 = 27.765 + (9.883/4) = 30.236 MHz
R_{f2} = 50 Ω
X_{f2} = –24 Ω × (10 / 30.236) = –7.9 Ω

When used in Eqs 10 through 15, these data yield:

R = 3137.59
X = –900.60
Z = 3264.28
R_0 = 56.58 Ω
X_0 = –7.96 Ω

Cable Attenuation

Cable loss can be measured once the cable electrical length and characteristic resistance are known. The measurement must be made at a frequency where the cable presents no reactance. Reactance is zero when the cable electrical length is an integer multiple of λ/4. You can easily meet that condition by making the measurement frequency an integer multiple of $^1/_4\ f\lambda$. Loss at other frequencies can be interpolated with reasonable accuracy. This procedure employs a resistor-substitution method that provides much greater accuracy than is achieved by directly reading the resistance from the noise-bridge scale.

1) Determine the approximate frequency at which you want to make the loss measurement by using

$$n = \frac{4f_0}{f_\lambda} \qquad \text{(Eq 16A)}$$

Round n to the nearest integer, then

$$f1 = \frac{n}{4} f_\lambda \qquad \text{(Eq 16B)}$$

2) If n is odd, leave the far end of the cable open; if n is even, connect the 0-Ω load to the far end of the cable. Attach the near end of the cable to the UNKNOWN connector on the noise bridge.
3) Set the noise bridge to zero reactance and the receiver to f1. Fine tune the receiver frequency and the noise-bridge resistance to find the null.
4) Disconnect the cable from the UNKNOWN terminal, and connect the variable-resistance calibration load in its place. Without changing the resistance setting on the bridge, adjust the load resistor and the bridge reactance to obtain a null.
5) Remove the variable-resistance load from the bridge UNKNOWN terminal and measure the load resistance using an ohmmeter that's accurate at low resistance levels. Refer to this resistance as R_i.
6) Calculate the cable loss in decibels using

$$a\ell = 8.69 \frac{R_i}{R_0} \qquad \text{(Eq 17)}$$

To continue this example, Eq 16A gives n = 11.74, so measure the attenuation at n = 12. From Eq 16B, f1 = 29.649 MHz. The input resistance of the cable measures 12.1 Ω with 0-Ω load on the far end of the cable; this corresponds to a loss of 1.86 dB.

USING A BRIDGE TO MEASURE THE IMPEDANCE OF AN ANTENNA

The impedance at the end of a transmission line can be easily measured using a noise bridge or SWR analyzer. In many cases, however, you really want to measure the impedance of an antenna—that is, the impedance of the load at the far end of the line. There are several ways to handle this.

1) Measurements can be made with the bridge at the antenna. This is usually not practical because the antenna must be in its final position for the measurement to be accurate. Even if it can be done, making such a measurement is certainly not very convenient.
2) Measurements can be made at the source end of a coaxial cable—if the cable length is an exact integer multiple of $^1/_2$ λ. This effectively restricts measurements to a single frequency.
3) Measurements can be made at the source end of a coaxial cable and corrected using a Smith Chart as shown in Chapter 28. This graphic method can result in reasonable estimates of antenna impedance—as long as the SWR is not too high and the cable is not too lossy. However, it doesn't compensate for the complex impedance characteristics of real-world coaxial cables. Also, compensation for cable loss can be tricky to apply. These problems, too, can lead to significant errors.
4) Last, measurements can be corrected using the transmission-line equation. The *TLW* program included on the CD-ROM in the back of this book, can do these complicated computations for you. This is the best method for calculating antenna impedances from measured parameters, but it requires that you measure the feed-line characteristics beforehand—measurements for which you need access to both ends of the feed line.

The procedure for determining antenna impedance is to first measure the electrical length, characteristic impedance, and attenuation of the coaxial cable connected to the antenna. After making these measurements, connect the antenna to the coaxial cable and measure the input impedance of the cable at a number of frequencies. Then use these measurements in the transmission-line equation to determine the actual antenna impedance at each frequency.

Table 6 and **Fig 33** give an example of such a calculation. The antenna used for this example is a 10-meter inverted V about 30 feet above the ground. The arms of the antenna are separated by a 120° angle. Each arm is exactly 8 feet long, and the antenna is made of #14 wire. The feed line is the 74-foot length of Columbia 1188 characterized earlier.

See Fig 33A. From this plot of impedance measurements, it is very difficult to determine anything about the antenna. Resistance and reactance vary substantially over this frequency range, and the antenna appears to be resonant at 27.7, 29.0 and 29.8 MHz.

The plot in Fig 33B shows the true antenna imped-

Table 6
Impedance Data for Inverted-V Antenna

Freq (MHz)	*R_u (Ω)*	*X_u @ 10 MHz (Ω)*	*X_u (Ω)*	*R_L (Ω)*	*X_L (Ω)*
27.0	44	85	31.5	24	–65
27.2	60	95	34.9	26	–56
27.4	75	85	31.0	30	–51
27.6	90	40	14.5	32	–42
27.8	90	–20	–7.2	35	–34
28.0	75	–58	–20.7	38	–24
28.2	65	–65	–23.0	40	–19
28.4	56	–52	–18.3	44	–12
28.6	50	–40	–14.0	44	–6
28.8	48	–20	–6.9	47	1
29.0	50	0	0.0	52	8
29.2	55	20	6.8	57	15
29.4	64	30	10.2	63	21
29.6	78	20	6.8	75	26
29.8	85	0	0	78	30
30.0	90	–50	–16.7	89	33

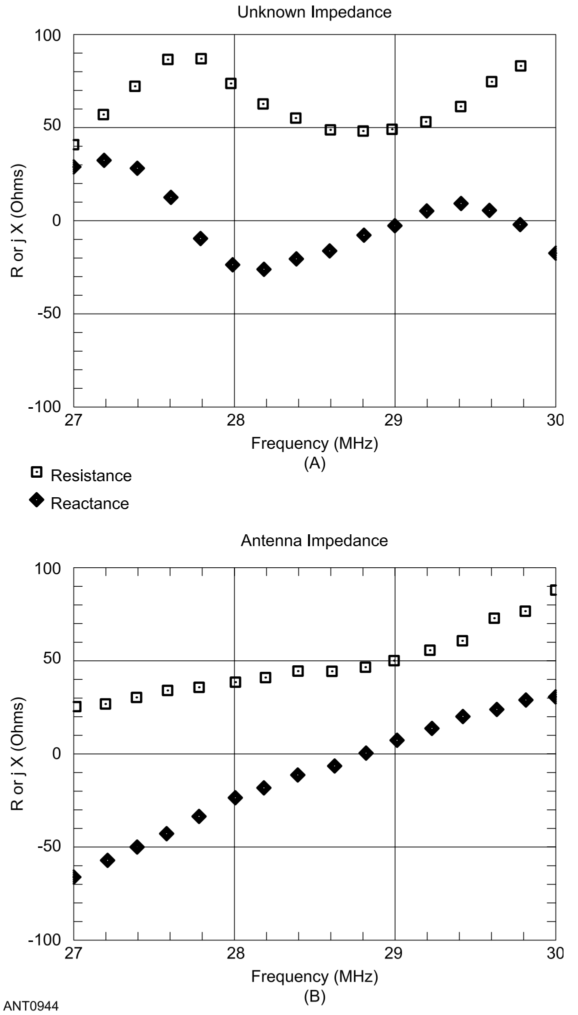

Fig 33–Impedance plot of an inverted-V antenna cut for 29 MHz. At A, a plot of resistances and reactances, measured using the noise bridge, at the end of a 74-foot length of Columbia 1188 coaxial cable. At B, the actual antenna-impedance plot (found using the transmission-line equation to remove the effects of the transmission line).

ance. This plot has been corrected for the effects of the cable using the transmission-line equation. The true antenna resistance and reactance both increase smoothly with frequency. The antenna is resonant at 28.8 MHz, with a radiation resistance at resonance of 47 Ω. This is normal for an inverted V.

When doing the conversions, be careful not to make measurement errors. Such errors introduce more errors into the corrected data. This problem is most significant when the transmission line is near an odd multiple of a $^1/_4$ λ and the line SWR and/or attenuation is high. Measurement errors are probably present if small changes in the input impedance or transmission-line characteristics appear as large changes in antenna impedance. If this effect is present, it can be minimized by making the measurements with a transmission line that is approximately an integer multiple of $^1/_2$ λ.

A Practical Time-Domain Reflectometer

A time-domain reflectometer (TDR) is a simple but powerful tool used to evaluate transmission lines. When used with an oscilloscope, a TDR displays impedance "bumps" (open and short circuits, kinks and so on) in transmission lines. Commercially produced TDRs cost from hundreds to thousands of dollars each, but you can add the TDR described here to your shack for much less. This material is based on a *QST* article by Tom King, KD5HM (see Bibliography), and supplemented with information from the references.

How a TDR Works

A simple TDR consists of a square-wave generator and an oscilloscope. See **Fig 34**. The generator sends a train of dc pulses down a transmission line, and the oscilloscope lets you observe the incident and reflected waves from the pulses (when the scope is synchronized to the pulses).

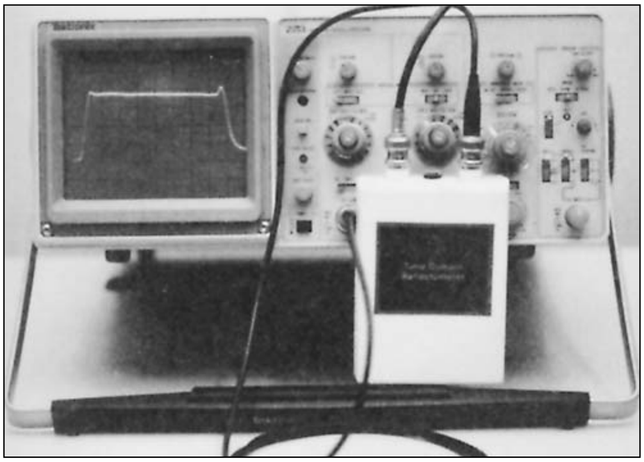

Fig 34—The time-domain reflectometer shown here is attached to a small portable oscilloscope.

A little analysis of the scope display tells the nature and location of any impedance changes along the line. The nature of an impedance disturbance is identified by comparing its pattern to those in **Fig 35**. The patterns are based on the fact that the reflected wave from a disturbance is determined by the incident-wave magnitude and the reflection coefficient of the disturbance. (The patterns shown neglect losses; actual patterns may vary somewhat from those shown.)

The location of a disturbance is calculated with a simple proportional method: The round-trip time (to the disturbance) can be read from the oscilloscope screen (graticule). Thus, you need only read the time, multiply it by the velocity of the radio wave (the speed of light adjusted by the velocity factor of the transmission line) and divide by two. The distance to a disturbance is given by:

$$\ell = \frac{983.6 \times \mathrm{VF} \times \mathrm{t}}{2} \qquad \text{(Eq 18)}$$

where

ℓ = line length in feet

VF = velocity factor of the transmission line (from 0 to 1.0)

t = time delay in microseconds (μs).

The Circuit

The time-domain reflectometer circuit in **Fig 36** consists of a CMOS 555 timer configured as an astable multivibrator, followed by an MPS3646 transistor acting as a 15-ns-risetime buffer. The timer provides a 71-kHz square wave. This is applied to the 50-Ω transmission line under test (connected at J2). The oscilloscope is connected to the circuit at J1.

Construction

An etching pattern for the TDR is shown in **Fig 37**. **Fig 38** is the part-placement diagram. The TDR is

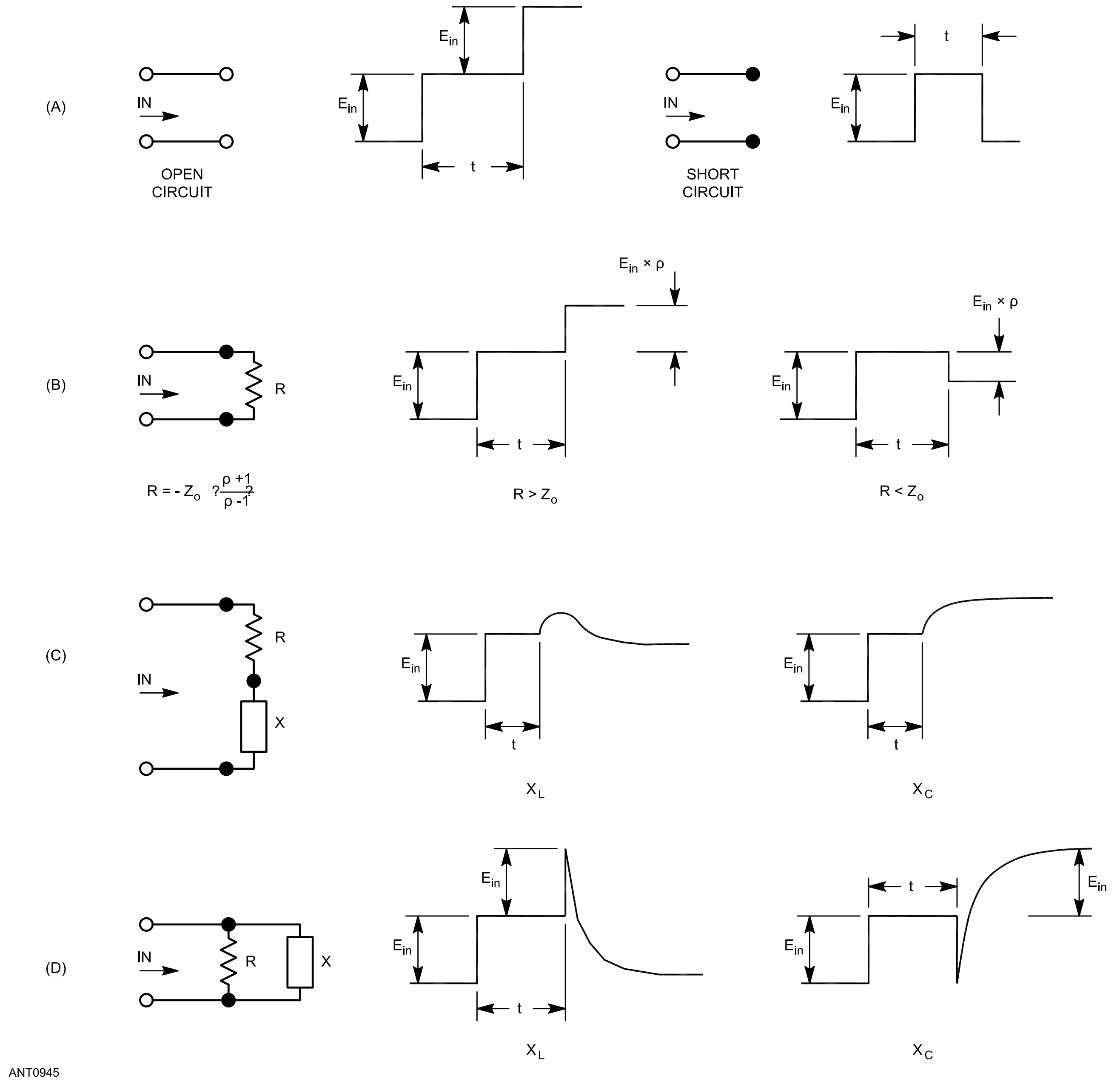

Fig 35—Characteristic TDR patterns for various loads. The location of the load can be calculated from the transit time, t, which is read from the oscilloscope (see text). R values can be calculated as shown (for purely resistive loads only—ρ < 0 when R < Z_0; ρ < 0 when R > Z_0). Values for reactive loads cannot be calculated simply.

designed for a 4 × 3 × 1-inch enclosure (including the batteries). Sl, Jl and J2 are right-angle-mounted components. Two aspects of construction are critical. First use *only* an MPS3646 for Q1. This type was chosen for its good performance in this circuit. If you substitute another transistor, the circuit may not perform properly.

Second, for the TDR to provide accurate measurements, the cable connected to J1 (between the TDR and the oscilloscope) must not introduce impedance mismatches in the circuit. *Do not make this cable from ordinary coaxial cable.* Oscilloscope-probe cable is the best thing to use for this connection.

(It took the author about a week and several phone calls to determine that scope-probe cable isn't "plain old coax." Probe cable has special characteristics that prevent undesired ringing and other problems.)

Mount a binding post at J1 and connect a scope probe to the binding post when testing cables with the TDR. R5 and C2 form a compensation network—much like the networks in oscilloscope probes—to adjust for effects of

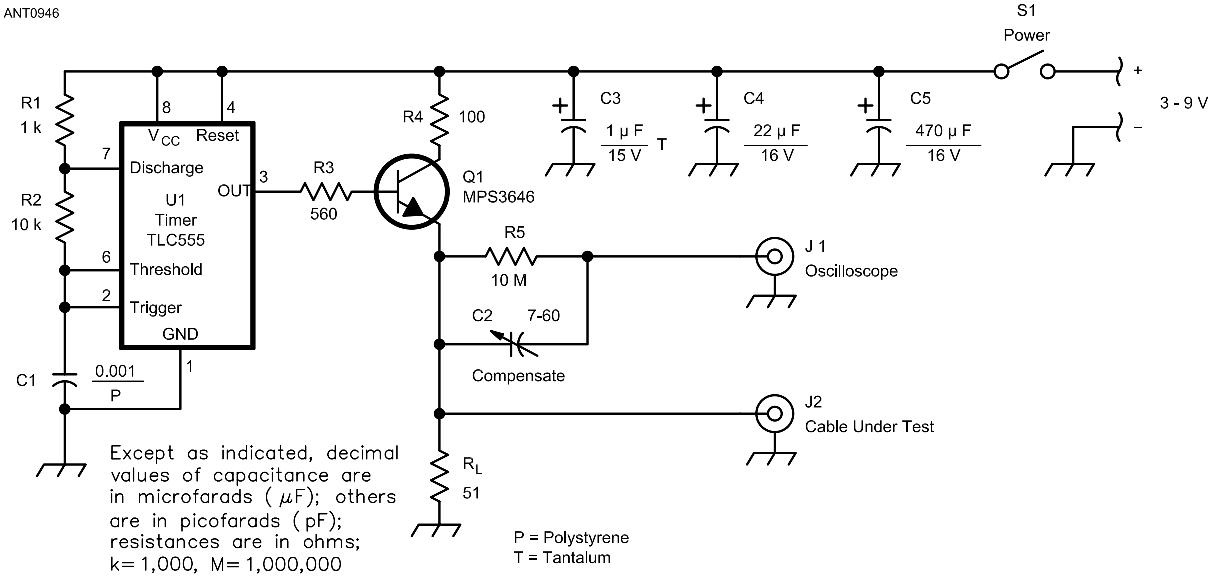
ANT0946
S1
Power
+
3 - 9 V
−
R1
1 k
8
4
VCC
Reset
7
Discharge
R2
10 k
U1
Timer
TLC555
OUT
3
R3
560
R4
100
Q1
MPS3646
C3
1 μ F
15 V
T
C4
22 μ F
16 V
C5
470 μ F
16 V
R5
10 M
J 1
Oscilloscope
6
Threshold
2
Trigger
GND
1
C2
7-60
Compensate
C1
0.001
P
J2
Cable Under Test
RL
51
P = Polystyrene
T = Tantalum

Except as indicated, decimal values of capacitance are in microfarads (μF); others are in picofarads (pF); resistances are in ohms; k= 1,000, M= 1,000,000

Fig 36—Schematic diagram of the time-domain reflectometer. All resistors are 1/4-W, 5% tolerance. U1 is a CMOS 555 timer. Circuit current drain is 10 to 25 mA. When building the TDR, observe the construction cautions discussed in the text. C2 is available from Mouser Electronics, part no. ME242-8050. Right-angle BNC connectors for use at J1 and J2 can be obtained from Newark Electronics, part no. 89N1578. S1 can be obtained from All Electronics, part no. NISW-1. An SPST toggle switch can also be used at S1.

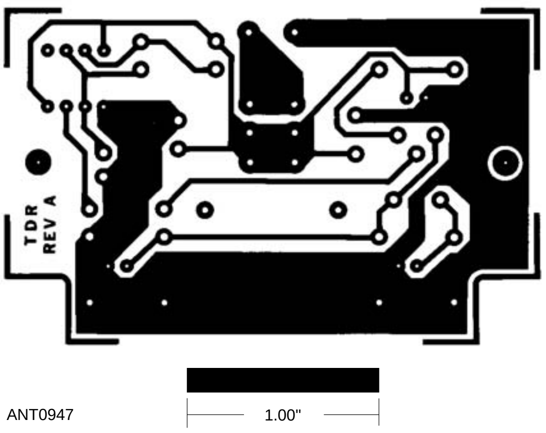
TDR
REV A
ANT0947
1.00"

Fig 37—Full-size PC-board etching pattern for the TDR. Black areas represent unetched copper foil.

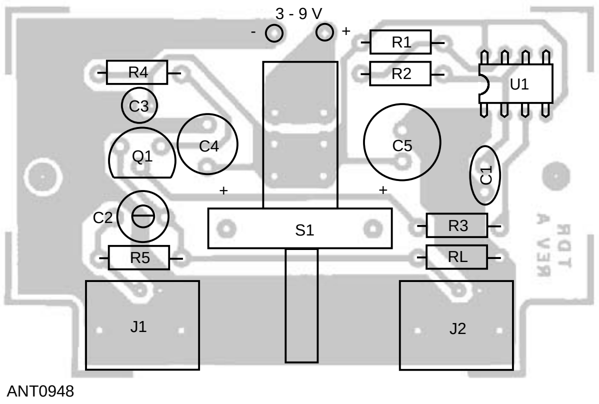
3 - 9 V
-
+
R1
R4
R2
U1
C3
C4
C5
Q1
C1
+
+
C2
S1
R3
R5
RL
J1
J2
TDR
REV A
ANT0948

Fig 38—Part-placement diagram for the TDR. Parts are mounted on the nonfoil side of the board; the shaded area represents an X-ray view of the copper pattern. Be sure to observe the polarity markings of C3, C4 and C5.

the probe wire.

The TDR is designed to operate from dc between 3 and 9 V. Two C cells (in series—3 V) supply operating voltage in this version. The circuit draws only 10 to 25 mA, so the cells should last a long time (about 200 hours of operation). U1 can function with supply voltages as low as 2.25 to 2.5.

If you want to use the TDR in transmission-line systems with characteristic impedances other than 50 Ω, change the value of R_L to match the system impedance as closely as possible.

Calibrating and Using the TDR

Just about any scope with a bandwidth of at least 10 MHz should work fine with the TDR, but for tests in short-length cables, a 50-MHz scope provides for much more accurate measurements. To calibrate the TDR, terminate CABLE UNDER TEST connector, J2, with a 51-Ω

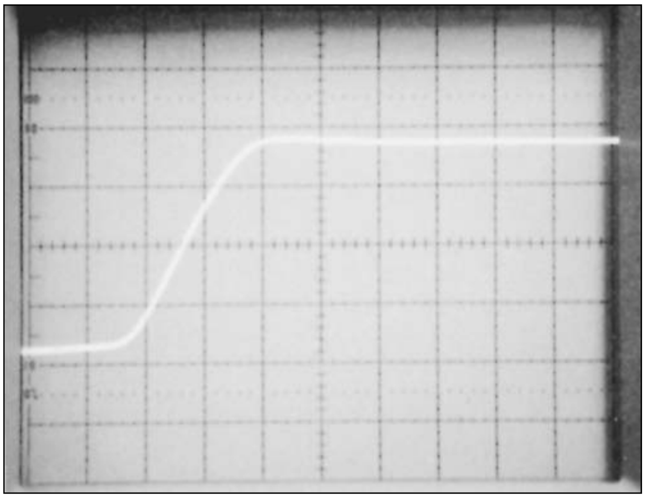

Fig 39—TDR calibration trace as shown on an oscilloscope. Adjust C2 (See Figs 36 and 38) for maximum deflection and sharpest waveform corners during calibration. See text.

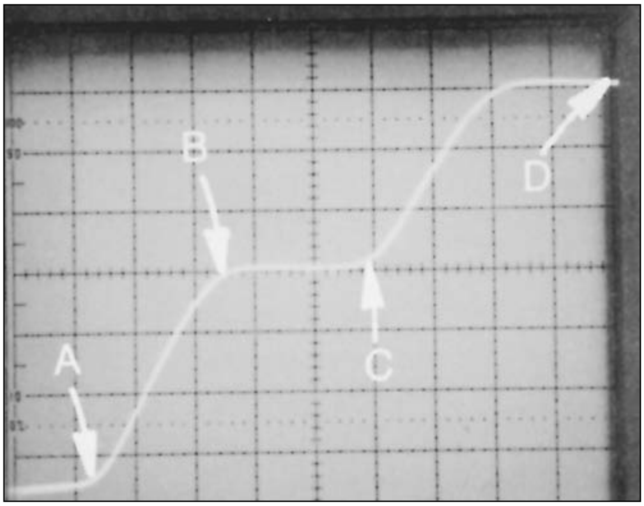

Fig 40—Open-circuited test cable. The scope is set for 0.01 ms per division. See text for interpretation of the waveform.

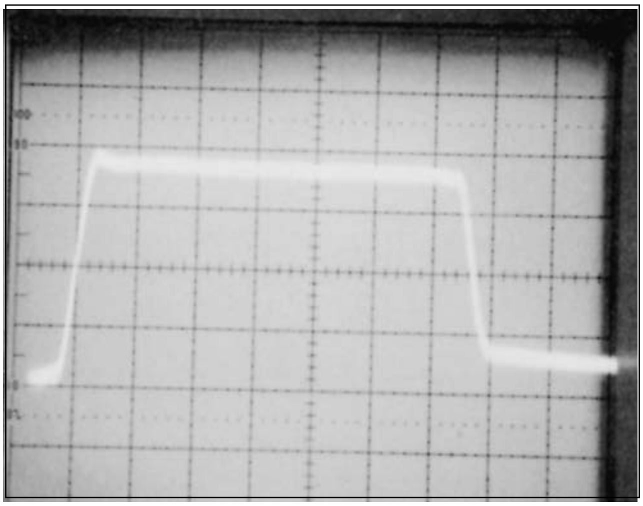

Fig 41—TDR display of the impedance characteristics of the 142-foot Hardline run to the 432-MHz antenna at KD5HM. The scope is set for 0.05 ms per division. See text for discussion.

resistor. Connect the scope vertical input to J1. Turn on the TDR, and adjust the scope timebase so that one square-wave cycle from the TDR fills as much of the scope display as possible (without uncalibrating the timebase). The waveform should resemble **Fig 39**. Adjust C2 to obtain maximum amplitude and sharpest corners on the observed waveform. That's all there is to the calibration process!

To use the TDR, connect the cable under test to J2, and connect the scope vertical input to J1. If the waveform you observe is different from the one you observed during calibration, there are impedance variations in the load you're testing. See **Fig 40**, showing an unterminated test cable connected to the TDR. The beginning of the cable is shown at point A. (AB represents the TDR output-pulse rise time.)

Segment AC shows the portion of the transmission line that has a 50-Ω impedance. Between points C and D, there is a mismatch in the line. Because the scope trace is higher than the 50-Ω trace, the impedance of this part of the line is higher than 50 Ω—in this case, an open circuit.

To determine the length of this cable, read the length of time over which the 50-Ω trace is displayed. The scope is set for 0.01 μs per division, so the time delay for the 50-Ω section is (0.01 μs × 4.6 divisions) = 0.046 μs. The manufacturer's specified velocity factor (VF) of the cable is 0.8. Eq 1 tells us that the 50-Ω section of the cable is

$$\ell = \frac{983.6 \times 0.8 \times 0.046\,\mu\text{s}}{2} = 18.1\,\text{feet}$$

The TDR provides reasonable agreement with the actual cable length—in this case, the cable is really 16.5 feet long. (Variations in TDR-derived calculations and actual cable lengths can occur as a result of cable VFs that can vary considerably from published values. Many cables vary as much as 10% from the specified values.)

A second example is shown in **Fig 41**, where a length of $^{3}/_{4}$-inch Hardline is being tested. The line feeds a 432-MHz vertical antenna at the top of a tower. Fig 41 shows that the 50-Ω line section has a delay of (6.6 divisions × 0.05 μs) = 0.33 μs. Because the trace is straight and level at the 50-Ω level, the line is in good shape. The trailing edge at the right-hand end shows where the antenna is connected to the feed line.

To determine the actual length of the line, use the same procedure as before: Using the published VF for the Hardline (0.88) in Eq 1, the line length is

$$\ell = \frac{983.6 \times 0.88 \times 0.33\ \mu\text{s}}{2} = 142.8\ \text{feet}$$

Again, the TDR-derived measurement is in close agreement with the actual cable length (142 feet).

Final Notes

The time-domain reflectometer described here is not frequency specific; its measurements are not made at the frequency at which a system is designed to be used. Because of this, the TDR cannot be used to verify the impedance of an antenna, nor can it be used to measure cable loss at a specific frequency. Just the same, in two years of use, it has never failed to help locate a transmission-line problem. The vast majority of transmission-line problems result from improper cable installation or connector weathering.

Limitations

Certain limitations are characteristic of TDRs because the signal used to test the line differs from the system operating frequency and because an oscilloscope is a broadband device. In the instrument described here, measurements are made with a 71-kHz square wave. That wave contains components at 71 kHz and odd harmonics thereof, with the majority of the energy coming from the lower frequencies. The leading edge of the trace indicates that the response drops quickly above 6 MHz. (The leading edge in Fig 40 is 0.042 µs, corresponding to a period of 0.168 µs and a frequency of 5.95 MHz.) The result is dc pulses of approximately 7 µs duration. The scope display combines the circuit responses to all of those frequencies. Hence, it may be difficult to interpret any disturbance which is narrowband in nature (affecting only a small range of frequencies, and thus a small portion of the total power), or for which the travel time plus pattern duration exceeds 7 µs. The 432-MHz vertical antenna in Fig 41 illustrates a display error resulting from narrow-band response.

The antenna shows as a major impedance disturbance because it is mismatched at the low frequencies that dominate the TDR display, yet it is matched at 432 MHz. For an event that exceeds the observation window, consider a 1-µF capacitor across a 50-Ω line. You would see only part of the pattern shown in Fig 35C because the time constant ($1 \times 10^{-6} \times 50 = 50$ ms) is much larger than the 7-µs window.

In addition, TDRs are unsuitable for measurements where there are major impedance changes inside the line section to be tested. Such major changes mask reflections from additional changes farther down the line.

Because of these limitations, TDRs are best suited for spotting faults in dc-continuous systems that maintain a constant impedance from the generator to the load. Happily, most amateur stations would be ideal subjects for TDR analysis, which can conveniently check antenna cables and connectors for short and open-circuit conditions and locate the position of such faults with fair accuracy.

Ground Parameters for Antenna Analysis

This section is taken from an article in *The ARRL Antenna Compendium, Vol 5* by R. P. Haviland, W4MB. In the past, amateurs paid very little attention to the characteristics of the earth (ground) associated with their antennas. There are two reasons for this. First, these characteristics are not easy to measure—even with the best equipment, extreme care is needed. Second, almost all hams have to put up with what they have—there are very few who can afford to move because their location has poor ground conditions! Further, the ground is not a dominant factor in the most popular antennas—a tri-band Yagi at 40 feet or higher, or a 2-meter vertical at roof height, for example.

Even so, there has been a desire and even a need for ground data and for ways to use it. It is very important for vertically polarized antennas. Ground data is useful for antennas mounted at low heights generally, and for such specialized ones as Beverages. The performance of such antennas change a lot as the ground changes.

Importance of Ground Conditions

To see why ground conditions can be important, let us look at some values. For a frequency of 10 MHz, *CCIR Recommendation 368*, gives the distance at which the signal is calculated to drop 10 dB below its free-space level as:

Conductivity (mS/meter)	*Distance for 10 dB Drop (km)*
5000	100
30	15
3	0.3

The high-conductivity condition is for sea-water. Inter-island work in the Caribbean on 40 and 80 meters is easy, whereas 40-meter ground-wave contact is difficult for much of the USA, because of much lower ground conductivity. On the other hand, the Beverage works because of poor ground conductivity.

Fig 42 shows a typical set of expected propagation curves for a range of frequencies. This data is also from *CCIR Recommendation 368* for relatively poor ground, with a dielectric constant of 4 and a conductivity of 3 mS/m (one milliSiemens/meter is 0.001 mho/meter). The same data is available in the *Radio Propagation Handbook*. There are equivalent FCC curves, found in the book *Reference Data for Radio Engineers*, but only the ones near 160 meters are useful. In Florida the author has difficulty hearing stations across town on ground

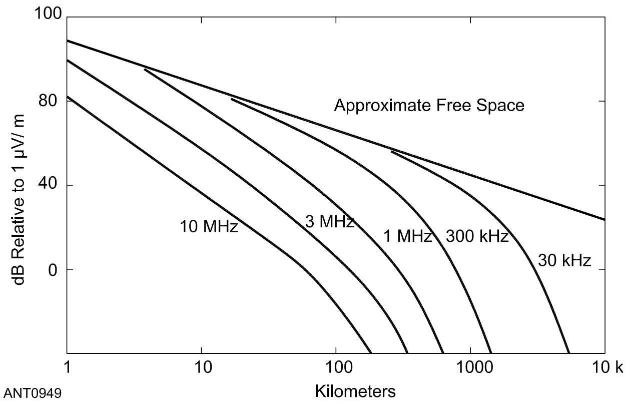

Fig 42—Variation of field strength with distance. Typical field strengths for several frequencies are shown. This is from CCIR data for fairly poor soil, with dielectric constant of 4 and conductivity of 3 mS/m. The curves for good soil are closer to the free-space line, and those for sea water are much closer to the free-space line.

wave, an indication of the poor soil conditions—reflected sky-wave signals are often stronger.

Securing Ground Data

There are only two basic ways to approach this matter of ground data. One is to use generic ground data typical to the area. The second is to make measurements, which haven't really gotten easier. For most amateurs, the best approach seems to be a combination of these—use some simple measurements, and then use the generic data to make a better estimate. Because of equipment costs and measurement difficulties, none of these will be highly accurate for most hams. But they will be much better than simply taking some condition preset into an analysis program. Having a good set of values to plug into an analysis can help you evaluate the true worth of a new antenna project.

Generic Data

In connection with its licensing procedure for broadcast stations, the FCC has published generic data for the

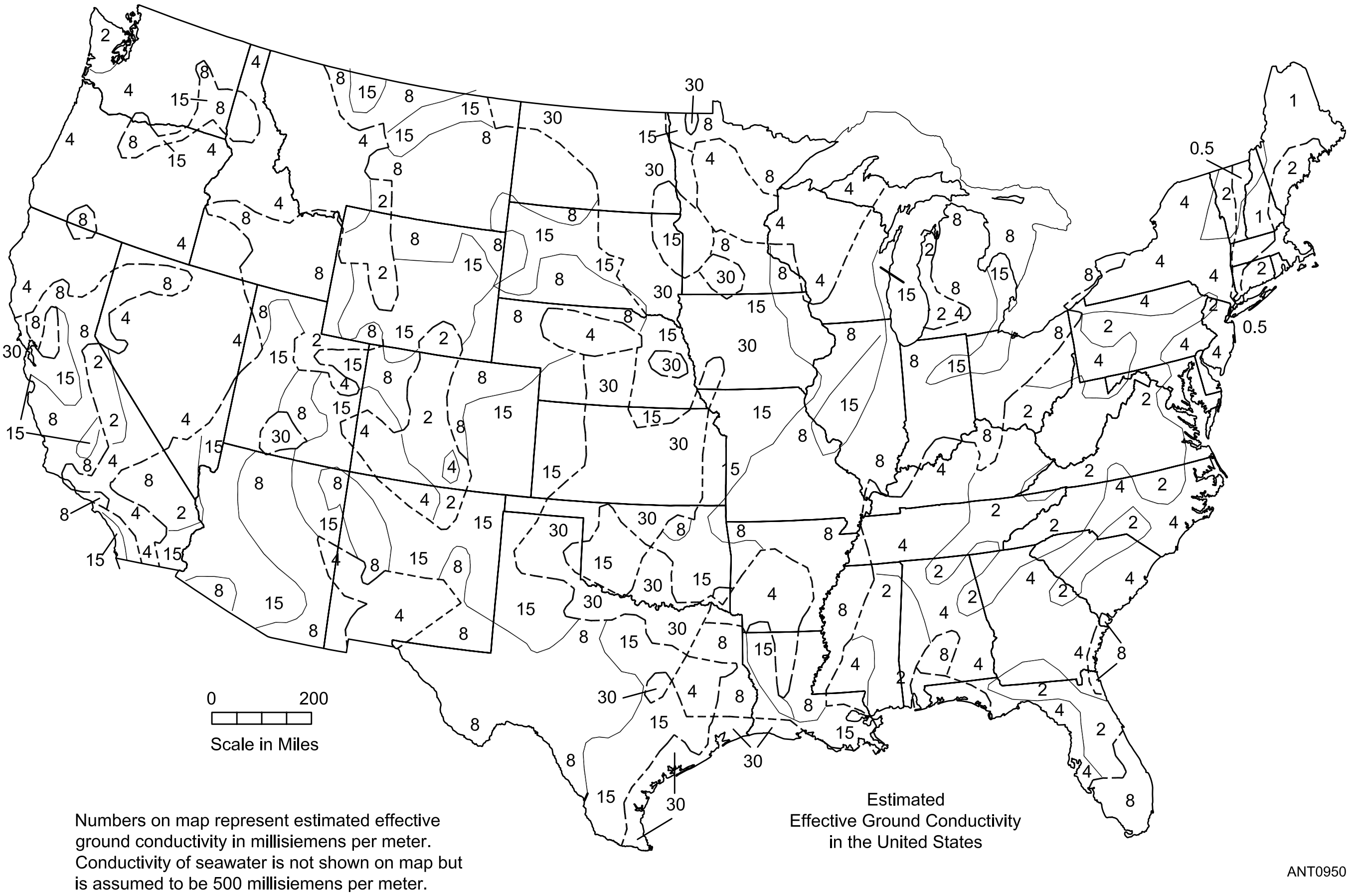

Numbers on map represent estimated effective ground conductivity in millisiemens per meter. Conductivity of seawater is not shown on map but is assumed to be 500 millisiemens per meter.

Fig 43—Estimated effective ground conductivity in the United States. FCC map prepared for the Broadcast Service, showing typical conductivity for continental USA. Values are for the band 500 to 1500 KHz. Values are for flat, open spaces and often will not hold for other types of commonly found terrain, such as seashores, river beds, etc.

entire country. This is reproduced in **Fig 43**, a chart showing the "estimated effective ground conductivity in the United States." A range of 30:1 is shown, from 1 to 30 mS/m. An equivalent chart for Canada has been prepared, originally by DOT, now DOC.

Of course, some judgment is needed when trying to use this data for your location. Broadcast stations are likely to be in open areas, so the data should not be assumed to apply to the center city. And a low site near the sea is likely to have better conductivity than the generic chart for, say, the coast of Oregon. Other than such factors, this chart gives a good first value, and a useful cross-check if some other method is used.

Still another FCC-induced data source is the license application of your local broadcast station. This includes calculated and measured coverage data. This may include specific ground data, or comparison of the coverage curves with the CCIR or FCC data to give the estimated ground conductivity. Another set of curves for ground conditions are those prepared by SRI. These give the conductivity and dielectric constant versus frequency for typical terrain conditions. These are reproduced as **Fig 44** and **Fig 45**. By inspecting your own site, you may select the curve most appropriate to your terrain. The curves are based on measurements at a number of sites across the USA, and are averages of the measured values.

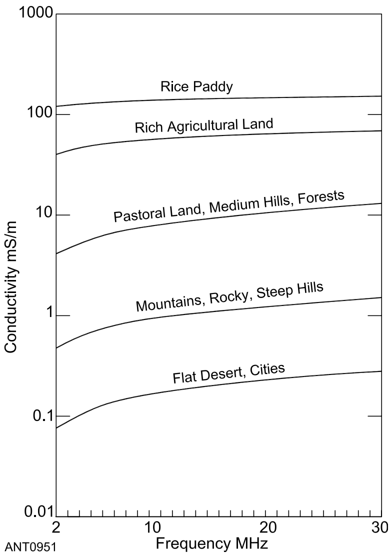

Fig 44—Typical terrain conductivities versus frequency for 5 types of soils. This was measured by SRI. Units are mS/m. Conductivity of seawater is usually taken as 5000 mS/m. Conductivity of fresh water depends on the impurities present, and may be very low. To extrapolate conductivity values (for 500 to 1500 KHz) shown in Fig 43 for a particular geographic area to a different frequency, move from the conductivity at the left edge of Fig 44 to the desired frequency. For example, in rocky New Hampshire, with a conductivity of 1 mS/m at BC frequencies, the effective conductivity at 14 MHz would be approximately 4 mS/m.

Figs 46 through 48 are data derived from these measurements. **Fig 46** gives the ground-dissipation factor. Sea water has low loss (a high dissipation factor), while soil in the desert or in the city is very lossy, with a low dissipation factor. **Fig 47** gives the skin depth, the distance for the signal to decease to 63% of its value at the surface. Penetration is low in high-conductivity areas and deep in low-conductivity soil. Finally, **Fig 48** shows the wavelength in the earth. For example, at 10 meters (30 MHz), the wavelength in sea water is less than 0.3 meters. Even in the desert, the wavelength has been reduced to about 6 meters at this frequency. This is one reason why buried antennas have peculiar properties. Lacking other data, it is suggested that the values of Figs 44 and 45 be used in computer antenna modeling programs.

Measuring Ground Conditions

W2FNQ developed a simple technique to measure low-frequency earth conductivity, which has been used by W2FMI. The test setup is drawn in **Fig 49**, and uses a very old technique of 4-terminal resistivity measurements. For probes of $^{9}/_{16}$-inch diameter, spaced 18 inches and pene-trating 12 inches into the earth, the conductivity is:

$$C = 21\ V_1/V_2 \quad \text{mS/m} \qquad \text{(Eq 19)}$$

The voltages are conveniently measured by a digital voltmeter, to an accuracy of about 2%. In soil suitable for farming, the probes can be copper or aluminum. The strength of iron or copperweld may be needed in hard soils. A piece of 2 × 4 or 4 × 4 with guide holes drilled through it will help maintain proper spacing and vertical alignment of the probes. Use care when measuring—there is a shock hazard. An isolating transformer with a 24-V secondary instead of 115 V will reduce the danger.

Ground conditions vary quite widely over even small

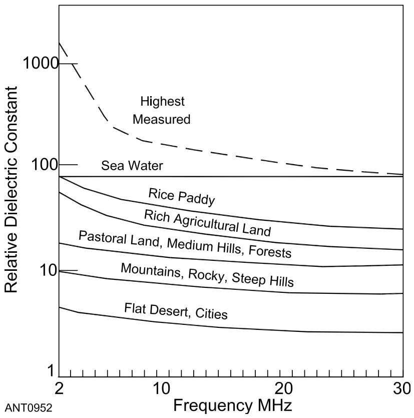

Fig 45—Typical terrain relative dielectric constant for the 5 soil types of Fig 44, plus sea water. The dashed curve shows the highest measured values reported, and usually indicates mineralization.

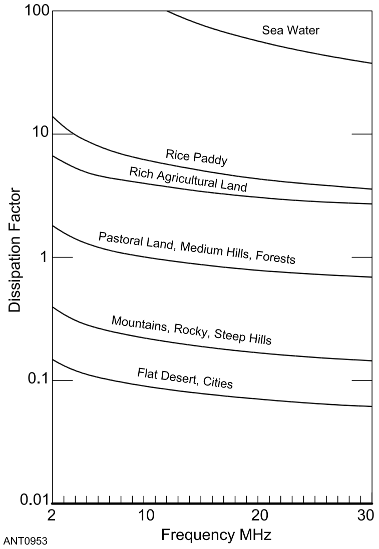

Fig 46—Typical values of dissipation factor. The soil behaves as a leaky dielectric. These curves showing the dimensionless dissipation factor versus frequency for various types of soils and for sea water. The dissipation factor is inversely related to soil conductivity. Among other things, a high dissipation factor indicates that a signal penetrating the soil or water will decrease in strength rapidly with depth.

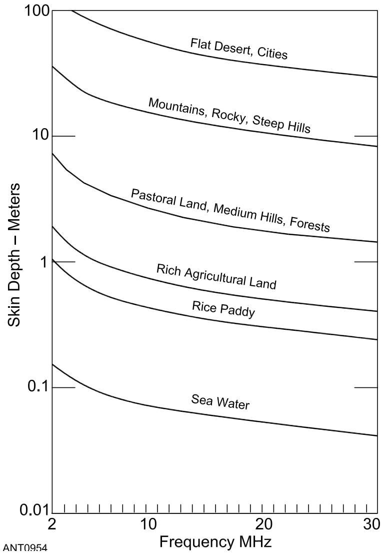

Fig 47—Typical values of skin depth. The skin depth is the depth at which a signal will have decreased to 1/e of its value at the surface (to about 30%). The effective height above ground is essentially the same as the physical height for sea water, but may be much greater for the desert. For practical antennas, this may increase low-angle radiation, but at the same time will increase ground losses.

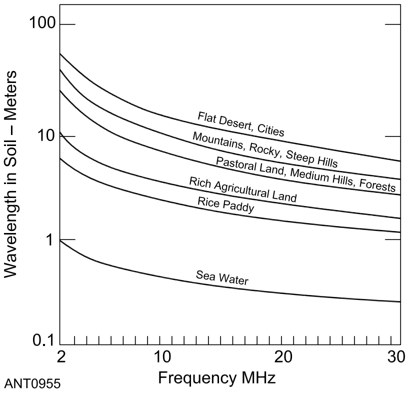

Fig 48—Typical values of wavelength in soil. Because of its dielectric constant, the wavelength in soils and water will be shorter than that for a wave traveling in air. This can be important, since in a Method of Moment the accuracy is affected by the number of analysis segments per wavelength. Depending on the program being used, adjust the number of segments for antennas wholly or partly in the earth, for ground rods, and for antennas very close to earth.

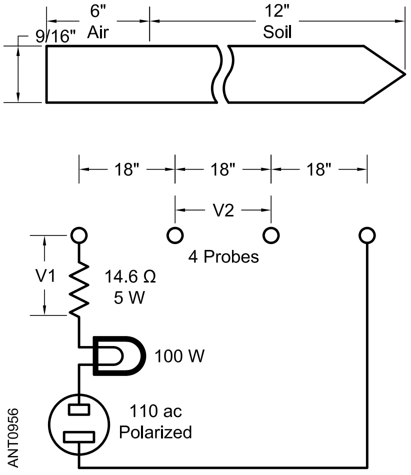

Fig 49—Low-frequency conductivity measurement system. A 60-Hz measuring system devised by W2FNQ and used by W2FMI. The basic system is widely used in geophysics. Use care to be certain that the plug connection is correct. A better system would use a lower voltage and an isolation transformer. Measure the value of V_2 with no power applied—there may be stray ground currents present, especially if there is a power station or an electric railway close.

areas. It is best to make a number of measurements around the area of the antenna, and average the measured values.

While this measurement gives only the low-frequency conductivity, it can be used to select curves in Fig 44 to give an estimate of the conductivity for the common ham bands. Assume that the 60 Hz value is valid at 2 MHz, and find the correct value on the left axis. Move parallel to the curves on the figure to develop the estimated curve for other soil conditions.

A small additional refinement is possible. If the dielectric constant from Fig 45 is plotted against the conductivity from Fig 44 for a given frequency, a scatter plot develops, showing a trend to higher dielectric constant as conductivity increases. At 14 MHz, the relation is:

$$k = \sqrt{1000/C} \qquad \text{(Eq 20)}$$

where k is the dielectric constant and C is the measured conductivity. Using these values in *MININEC* or *NEC* calculations should give better estimates than countrywide average values.

Direct Measurement of Ground Properties

For really good values, both the conductivity and dielectric constant should be measured at the operating frequency. One way of doing this is the two-probe technique described in George Hagn's article (see Bibliography). This was the technique used to secure the data for Figs 44 through 48. The principle is sketched in **Fig 50**. In essence, the two probes form a short, open-circuited, two-wire transmission line. As shown by the equations for such lines, the input impedance is a function of the conductivity and dielectric constant of the medium. A single measurement is difficult to calculate, since the end effect of the two probes must be determined, a complex task if they are pointed for easy driving. The calculation is greatly simplified if a set of measurements is made with several sets of probes that vary in length by a fixed ratio, since the measured difference is largely due to the increased two-wire length, with some change due to the change in soil moisture with depth.

The impedance to be measured is high because of the short line length, so impedance bridges are not really suitable. An RF vector impedance meter, such as the HP-4193A, is probably the best instrument to use, with a RF susceptance bridge, such as the GR-821A, next best. With care, a Q-meter can be substituted. Because of the rarity of these instruments among amateurs, this method of measurement is not explored further here.

Indirect Measurement

Since the terminal impedance and resonant frequency of an antenna change as the antenna approaches earth, measurement of an antenna at one or more heights permits an analysis of the ground characteristics. The technique is to calculate the antenna drive impedance for an assumed ground condition, and compare this with measured values. If not the same, another set of ground conditions is assumed, and the process is repeated. It is best to have a plan to guide the assumptions.

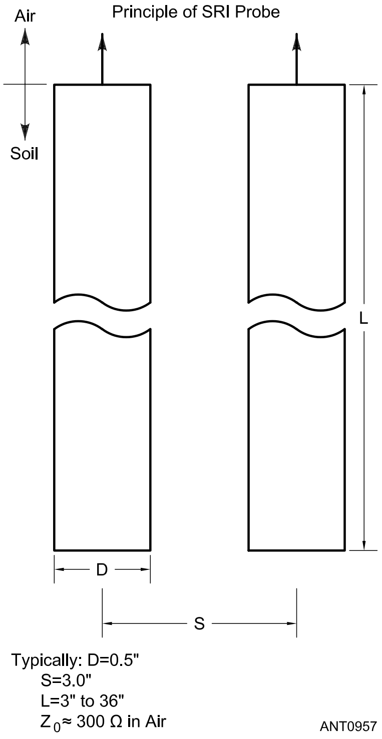

Fig 50—High-frequency conductivity/dielectric constant measurement system. System for measuring ground conditions at frequencies up to about 100 MHz, devised by SRI and used to obtain the data in Figs 44 through 48. Basically, this is a section of transmission line with soil as the dielectric. Requires measurement of high impedances to good accuracy.

In connection with his studies of transmission lines, Walt Maxwell, W2DU, made such measurements on 20, 40 and 80 meters. Some of the data was included in his book *Reflections*. The following example is based on his 80-meter data. Data came from his Table 20-1, for a 66-foot, 2-inch dipole of #14 wire at 40 feet above ground. His table gives an antenna impedance of 72.59 + *j* 1.38 Ω at 7.15 MHz.

Table 7 shows calculated antenna impedances for ground conductivities of three different ground conductivities: 10, 1 and 0.1 mS/m, and for dielectric constants of 3, 15 and 80. The nearest value to the measured drive impedance is for a conductivity of 0.1 mS/m and a dielectric constant of 3. Figs 44 and 45 indicate that these are typical of flat desert and city land. The effect on antenna performance is shown in **Fig 51**. The maximum lobe gain for soil typical of a city is over 2 dB lower than that for the high-conductivity, high-dielectric constant value. Note that the maximum lobe occurs for a radiation angle that is directly overhead.

Table 7

Calculated values of drive resistance, in ohms, for an 40 meter dipole at 40 feet elevation versus conductivity and dielectric constant.

Conductivity	*Dielectric Constant*		
(mS/m)	*3*	*15*	*50*
10	89.78 – *j* 12.12	88.53 – *j* 10.69	88.38 – *j* 7.59
1	80.05 – *j* 17.54	83.72 – *j* 10.23	87.33 – *j* 6.98
0.1	76.44 – *j* 15.69	83.18 – *j* 9.85	97.30 – *j* 6.46

The value measured by W2DU was 72.59 – *j* 1.28 Ω, and compares closest to the poor soil condition of dielectric constant of 3 and conductivity of 0.1 mS/m.

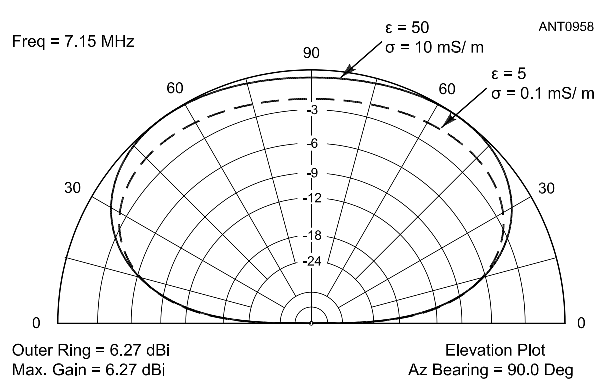

Fig 51—Plot showing computed elevation patterns for 40-foot high, 40-meter dipole for two different ground conditions: poor ground, with dielectric constant of 3 and conductivity of 0.1 mS/m, and good ground, with dielectric constant of 50 and conductivity of 10 mS/m. Note that for a low horizontal antenna, high-angle radiation is most affected by poor ground, with low-angle radiation least affected by ground characteristics.

The ground at the W2DU QTH is a suburban Florida lot, covered with low, native vegetation. The ground is very sandy (a fossil sand dune), and is some 60-70 feet above sea-level. Measurements were made near the end of the Florida dry season. The water table is estimated to be 20 to 30 feet below the surface. Thus the calculated and measured values are reasonably consistent.

In principle, a further analysis, using values around 0.1 mS/m conductivity and 3 for dielectric constant, will give a better ground parameter estimate. However, the results should be taken with a grain of salt, because the opportunities for error in the computer modeling must be considered. The antenna should have no sag, and its length and height should be accurate. The measurement must be with accurate equipment, free from strays, such as current on the outer conductor of the coax. The feed-point gap effect must be estimated. Further, the ground itself under the antenna must be flat and have constant characteristics for modeling to be completely accurate.

Finally, the feed-line length and velocity constant of the transmission line must be accurately measured for transfer of the measured values at the feeding end of the transmission line to the antenna itself. Because of all the possibilities for error, most attempts at precision should be based on measured values at two or three frequencies, and preferably at two or three heights. Orienting the antenna to right angles for another set of measurements may be useful. Obviously, this can involve a lot of detailed work.

The author was not been able to find any guidelines for the best height or frequency. The data in the book *Exact Image Method for Impedance Computation of Antennas Above the Ground* suggests that a height of 0.3 λ will give good sensitivity to ground conditions. Very low heights may give confusing results, since several combinations of ground parameters can give nearly the same drive impedance. Both this data and experience suggest that sensitivity to ground for heights above 0.75 λ is small or negligible.

If an overall conclusion about ground characteristics is needed, we can just restate from the first paragraph—it is not greatly important for the most common horizontally polarized antenna installations. But it's worth taking a look when you need to depart from typical situations, or when the performance of a vertically polarized antenna is contemplated. Then the techniques outlined here can be helpful.

A Switchable RF Attenuator

A switchable RF attenuator is helpful for making antenna-gain comparisons or for plotting antenna radiation patterns. You may switch attenuation in or out of the line leading to the receiver to obtain an initial reference reading on a signal strength meter. Some form of attenuator is also helpful for locating hidden transmitters, where the real trick is pinpointing the signal source from within a few hundred feet. At such a close distance, strong signals may overload the front end of the receiver, making it impossible to obtain any indication of a bearing.

The attenuator of **Figs 52** and **53** is designed for low power levels, not exceeding $^{1}/_{4}$ watt. If for some reason the attenuator will be connected to a transceiver, a means of bypassing the unit during transmit periods must be devised. An attenuator of this type is commonly called a *step attenuator*, because any amount of attenuation from 0 dB to the maximum available (81 dB for this particular instrument) may be obtained in steps of 1 dB. As each switch is successively thrown from the OUT to the IN position, the attenuation sections add in cascade to yield the total of the attenuator steps switched in. The maximum attenuation of any single section is limited to 20 dB because leak-through would probably degrade the accuracy of higher values. The tolerance of resistor values also becomes more significant regarding accuracy at higher attenuation values.

A good quality commercially made attenuator will cost upwards from $150, but for less than $25 in parts and a few hours of work, you can build an attenuator at home. It will be suitable for frequencies up to 450 MHz. Double-sided pc board is used for the enclosure. The version of the attenuator shown in Fig 52 has identification lettering etched into the top surface (or front panel) of the unit. This adds a nice touch and is a permanent means of labeling. Of course rub-on transfers or Dymo tape labels could be used as well.

Female BNC single-hole, chassis-mount connectors are used at each end of the enclosure. These connectors

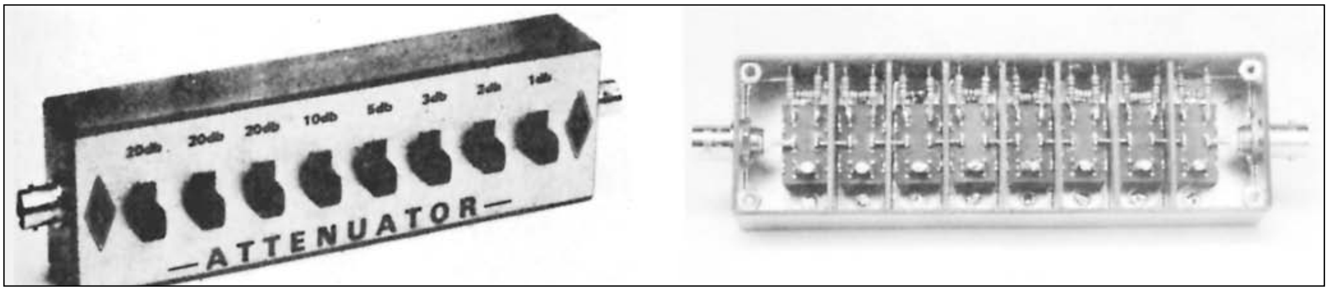

Fig 52—A construction method for a step attenuator. Double-sided circuit-board material, unetched (except for panel identification), is cut to the desired size and soldered in place. Flashing copper may also be used, although it is not as sturdy. Shielding partitions between sections are necessary to reduce signal leakage. Brass nuts soldered at each of the four corners allow machine screws to secure the bottom cover. The practical limit for total attenuation is 80 or 90 dB, as signal leakage around the outside of the attenuator will defeat attempts to obtain much greater amounts.

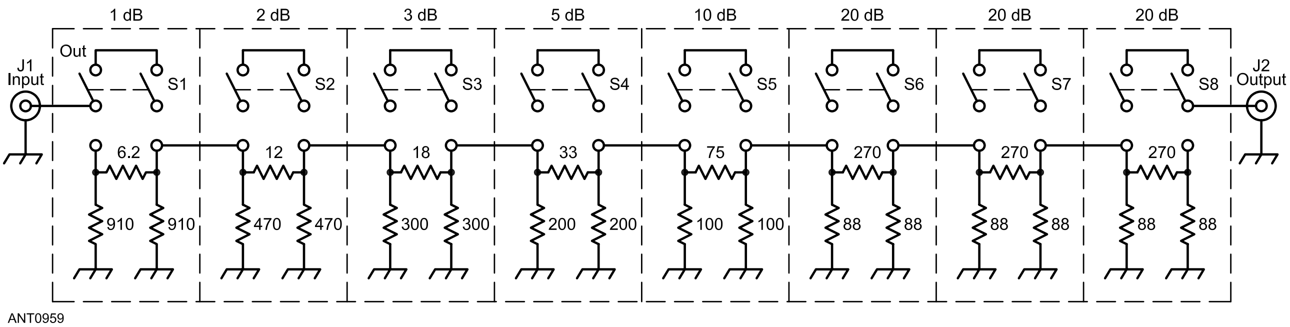

Fig 53—Schematic diagram of the step attenuator, designed for a nominal impedance of 50 Ω. Resistance values are in ohms. Resistors are $^{1}/_{4}$-watt, carbon-composition types, 5% tolerance. Broken lines indicate walls of circuit-board material. A small hole is drilled through each partition wall to route bus wire. Keep all leads as short as possible. The attenuator is bilateral; that is, the input and output ends may be reversed.

J1, J2—Female BNC connectors, Radio Shack 278-105 or equiv.

S1-S8, incl.—DPDT slide switches, standard size. (Avoid subminiature or toggle switches.) Stackpole S-5022CD03-0 switches are used here.

provide a means of easily connecting and disconnecting the attenuator.

Construction

After all the box parts are cut to size and the necessary holes made, scribe light lines to locate the inner partitions. Carefully tack-solder all partitions in position. A 25-W pencil type of iron should provide sufficient heat. Dress any pc board parts that do not fit squarely. Once everything is in proper position, run a solder bead all the way around the joints. Caution! Do not use excessive amounts of solder, as the switches must later be fit flat inside the sections. Complete the top, sides, ends and partitions. Dress the outside of the box to suit your taste. For instance, you might wish to bevel the box edges. Buff the copper with steel wool, add lettering, and finish off the work with a coat of clear lacquer or polyurethane varnish.

Using a little lacquer thinner, soak the switches to remove the grease that was added during their manufacture. When they dry, spray the inside of the switches lightly with a TV tuner cleaner/lubricant. Use a sharp drill bit (about $^3/_{16}$ inch will do), and countersink the mounting holes on the actuator side of the switch mounting plate. This ensures that the switches will fit flush against the top plate. At one end of each switch, bend the two lugs over and solder them together. Cut off the upper halves of the remaining switch lugs. (A close look at Fig 52 will help clarify these steps.)

Solder the series-arm resistors between the appropriate switch lugs. Keep the lead lengths as short as possible and do not overheat the resistors. Now solder the switches in place to the top section of the enclosure by flowing solder through the mounting holes and onto the circuit-board material. Be certain that you place the switches in their proper positions; correlate the resistor values with the degree of attenuation. Otherwise, you may wind up with the 1-dB step at the wrong end of the box how embarrassing!

Once the switches are installed, thread a piece of #18 bare copper wire through the center lugs of all the switches, passing it through the holes in the partitions. Solder the wire at each switch terminal. Cut the wire between the poles of each individual switch, leaving the wire connecting one switch pole to that of the neighboring one on the other side of the partition, as shown in Fig 52. At each of the two end switch terminals, leave a wire length of approximately $^1/_8$ inch. Install the BNC connectors and solder the wire pieces to the connector center conductors.

Now install the shunt-arm resistors of each section. Use short lead lengths. Do not use excessive amounts of heat when soldering. Solder a no. 4-40 brass nut at each inside corner of the enclosure. Recess the nuts approximately $^1/_{16}$-inch from the bottom edge of the box to allow sufficient room for the bottom panel to fit flush. Secure the bottom panel with four no. 4-40, $^1/_4$-inch machine screws and the project is completed. Remember to use caution, always, when your test setup provides the possibility of transmitting power into the attenuator.

A Portable Field Strength Meter

Few amateur stations, fixed or mobile, are without need of a field-strength meter. An instrument of this type serves many useful purposes during antenna experiments and adjustments. When work is to be done from many wavelengths away, a simple wavemeter lacks the necessary sensitivity. Further, such a device has a serious fault because its linearity leaves much to be desired. The information in this section is based on a January 1973 *QST* article by Lew McCoy, W1ICP.

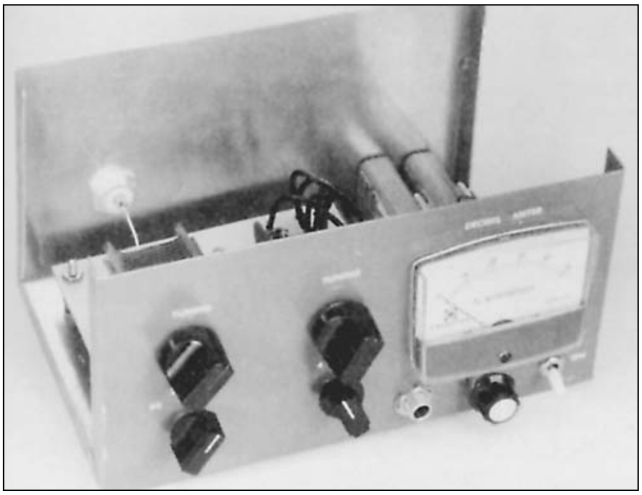

Fig 54—The linear field strength meter. The control at the upper left is for C1 and the one to the right for C2. At the lower left is the band switch, and to its right the sensitivity switch. The zero-set control for M1 is located directly below the meter.

The field-strength meter described here takes care of these problems. Additionally, it is small, measuring only 4 × 5 × 8 inches. The power supply consists of two 9-volt batteries. Sensitivity can be set for practically any amount desired. However, from a usefulness standpoint, the circuit should not be too sensitive or it will respond to unwanted signals. This unit also has excellent linearity with regard to field strength. (The field strength of a received signal varies inversely with the distance from the source, all other things being equal.) The frequency range includes all amateur bands from 3.5 through

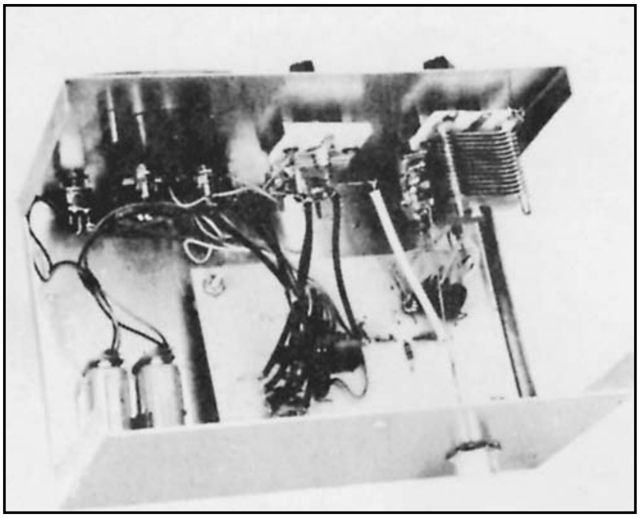

Fig 55—Inside view of the field-strength meter. At the upper right is C1 and to the left, C2. The dark leads from the circuit board to the front panel are the shielded leads described in the text.

148 MHz, with band-switched circuits, thus avoiding the use of plug-in inductors. All in all, it is a quite useful instrument.

The unit is pictured in **Figs 54** and **55**, and the schematic diagram is shown in **Fig 56**. A type 741 op-amp IC is the heart of the unit. The antenna is connected to J1, and a tuned circuit is used ahead of a diode detector. The rectified signal is coupled as dc and amplified in the op amp. Sensitivity of the op amp is controlled by inserting resistors R3 through R6 in the circuit by means of S2.

With the circuit shown, and in its most sensitive setting, M1 will detect a signal from the antenna on the order of 100 µV. Linearity is poor for approximately the first $^{1}/_{5}$ of the meter range, but then is almost straight-line from there to full-scale deflection. The reason for the poor linearity at the start of the readings is because of nonlinearity of the diodes at the point of first conduction. However, if gain measurements are being made this is of no real importance, as accurate gain measurements can be made in the linear portion of the readings.

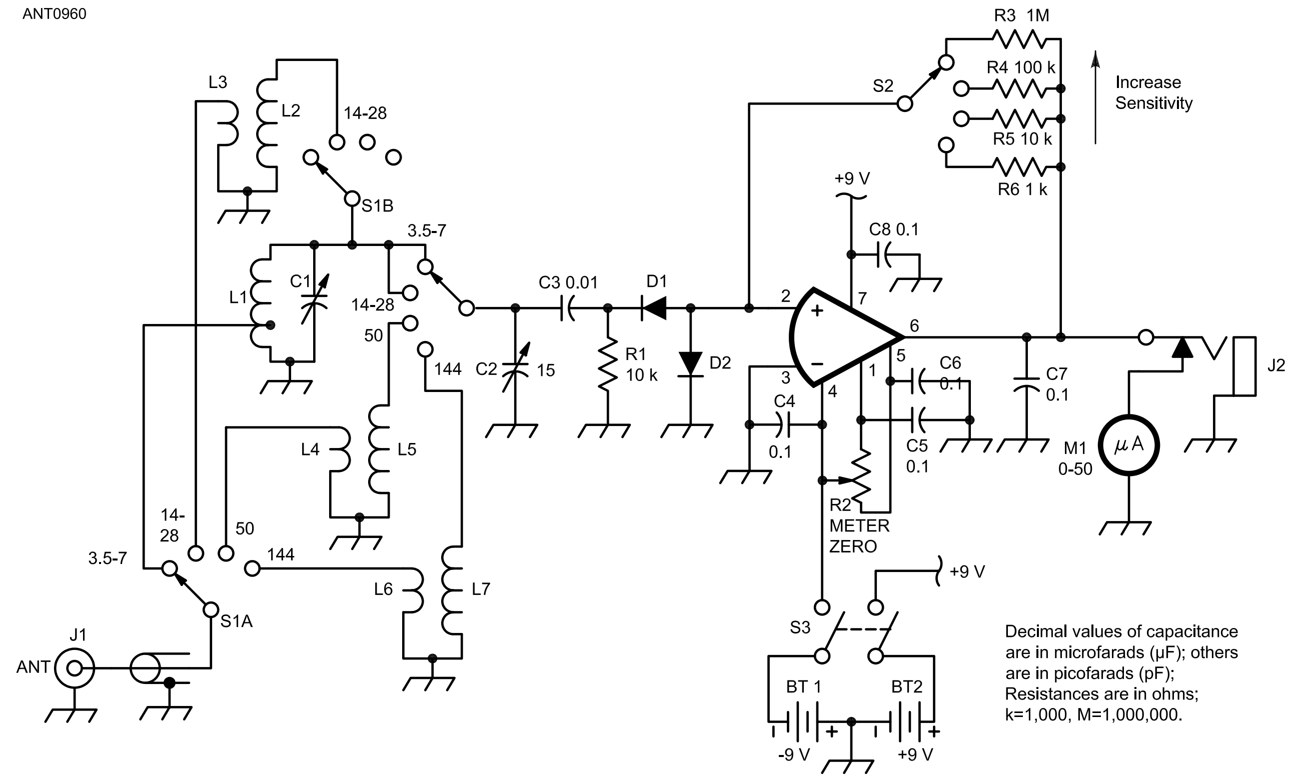
ANT0960
R3 1M
R4 100 k
R5 10 k
R6 1 k
S2
Increase Sensitivity
L3
L2
14-28
S1B
3.5-7
C1
L1
14-28
50
144
C2
15
C3 0.01
D1
R1 10 k
D2
+9 V
C8 0.1
2
7
6
3
4
1
5
C6 0.1
C4 0.1
C5 0.1
C7 0.1
R2 METER ZERO
M1 0-50
µA
J2
L4
L5
L6
L7
14-
28
50
144
3.5-7
S1A
J1
ANT
S3
+9 V
BT 1
BT2
-9 V
+9 V
Decimal values of capacitance are in microfarads (µF); others are in picofarads (pF); Resistances are in ohms; k=1,000, M=1,000,000.

Fig 56—Circuit diagram of the linear field strength meter. All resistors are $^{1}/_{4}$- or $^{1}/_{2}$-W composition types.

C1 — 140 pF variable.
C2 — 15-pF variable
D1, D2 — 1N914 or equiv.
L1 — 34 turns #24 enam. wire wound on an Amidon T-68-2 core, tapped 4 turns from ground end.
L2 — 12 turns #24 enam. wire wound on T-68-2 core.
L3 — 2 turns #24 enam. wire wound at ground end of L2.
L4 — 1 turn #26 enam. wire wound at ground end of L5.
L5 — 12 turns #26 enam. wire wound on T-25-12 core.
L6 — 1 turn #26 enam. wire wound at ground end of L7.
L7 — 1 turn #18 enam. wire wound on T-25-12 core.
M1 — 50 or 100 µA dc.
R2 — 10-kΩ control, linear taper.
S1 — Rotary switch, 3 poles, 5 positions, 3 sections.
S2 — Rotary switch, 1 pole, 4 positions.
S3 — DPST toggle.
U1 — Type 741 op amp. Pin numbers shown are for a 14-pin package.

The 741 op amp requires both a positive and a negative voltage source. This is obtained by connecting two 9-volt batteries in series and grounding the center. One other feature of the instrument is that it can be used remotely by connecting an external meter at J2. This is handy if you want to adjust an antenna and observe the results without having to leave the antenna site.

L1 is the 3.5/7 MHz coil and is tuned by C1. The coil is wound on a toroid form. For 14, 21 or 28 MHz, L2 is switched in parallel with L1 to cover the three bands. L5 and C2 cover approximately 40 to 60 MHz, and L7 and C2 from 130 MHz to approximately 180 MHz. The two VHF coils are also wound on toroid forms.

Construction Notes

The majority of the components may be mounted on an etched circuit board. A shielded lead should be used between pin 4 of the IC and S2. The same is true for the leads from R3 through R6 to the switch. Otherwise, parasitic oscillations may occur in the IC because of its very high gain.

In order for the unit to cover the 144-MHz band, L6 and L7 should be mounted directly across the appropriate terminals of S1, rather than on a circuit board. The extra lead length adds too much stray capacitance to the circuit. It isn't necessary to use toroid forms for the 50- and 144-MHz coils. They were used in the version described here simply because they were available. You may substitute air-wound coils of the appropriate inductance.

Calibration

The field strength meter can be used *as is* for a relative-reading device. A linear indicator scale will serve admirably. However, it will be a much more useful instrument for antenna work if it is calibrated in decibels, enabling the user to check relative gain and front-to-back ratios. If you have access to a calibrated signal generator, connect it to the field-strength meter and use different signal levels fed to the device to make a calibration chart. Convert signal-generator voltage ratios to decibels by using the equation

$$\text{dB} = 20 \log (V1/V2) \qquad \text{(Eq 21)}$$

where

V1/V2 is the ratio of the two voltages

log is the common logarithm (base 10)

Let's assume that M1 is calibrated evenly from 0 to 10. Next, assume we set the signal generator to provide a reading of 1 on M1, and that the generator is feeding a 100-µV signal into the instrument. Now we increase the generator output to 200 µV, giving us a voltage ratio of 2:1. Also let's assume M1 reads 5 with the 200-µV input. From the equation above, we find that the voltage ratio of 2 equals 6.02 dB between 1 and 5 on the meter scale. M1 can be calibrated more accurately between 1 and 5 on its scale by adjusting the generator and figuring the ratio. For example, a ratio of 126 µV to 100 µV is 1.26, corresponding to 2.0 dB. By using this method, all of the settings of S2 can be calibrated. In the instrument shown here, the most sensitive setting of S2 with R3, 1 MΩ, provides a range of approximately 6 dB for M1. Keep in mind that the meter scale for each setting of S1 must be calibrated similarly for each band. The degree of coupling of the tuned circuits for the different bands will vary, so each band must be calibrated separately.

Another method for calibrating the instrument is using a transmitter and measuring its output power with an RF wattmeter. In this case we are dealing with power rather than voltage ratios, so this equation applies:

$$\text{dB} = 10 \log (P1/P2) \qquad \text{(Eq 22)}$$

where P1/P2 is the power ratio.

With most transmitters the power output can be varied, so calibration of the test instrument is rather easy. Attach a pickup antenna to the field-strength meter (a short wire a foot or so long will do) and position the device in the transmitter antenna field. Let's assume we set the transmitter output for 10 W and get a reading on M1. We note the reading and then increase the output to 20 W, a power ratio of 2. Note the reading on M1 and then use Eq 2. A power ratio of 2 is 3.01 dB. By using this method the instrument can be calibrated on all bands and ranges.

With the tuned circuits and coupling links specified in Fig 56, this instrument has an average range on the various bands of 6 dB for the two most sensitive positions of S2, and 15 dB and 30 dB for the next two successive ranges. The 30-dB scale is handy for making front-to-back antenna measurements without having to switch S2.

An RF Current Probe

The RF current probe of **Figs 57** through **59** operates on the magnetic component of the electromagnetic field, rather than the electric field. Since the two fields are precisely related, as discussed in Chapter 23, the relative field strength measurements are completely equivalent. The use of the magnetic field offers certain advantages, however. The instrument may be made more compact for the same sensitivity, but its principal advantage is that it may be used near a conductor to measure the current flow without cutting the conductor.

In the average amateur location there may be substantial currents flowing in guy wires, masts and towers, coaxial-cable braids, gutters and leaders, water and gas pipes, and perhaps even drainage pipes. Current may be flowing in telephone and power lines as well. All of these RF currents may have an influence on antenna patterns or can be of significance in the case of RFI.

The circuit diagram of the current probe appears in Fig 58, and construction is shown in the photo, Fig 59. The winding data given here apply only to a ferrite rod of the particular dimensions and material specified. Almost any microammeter can be used, but it is usually convenient to use a rather sensitive meter and provide a series resistor to swamp out nonlinearity arising from diode conduction characteristics. A control is also used to adjust instrument sensitivity as required during operation. The tuning capacitor may be almost anything that will cover the desired range.

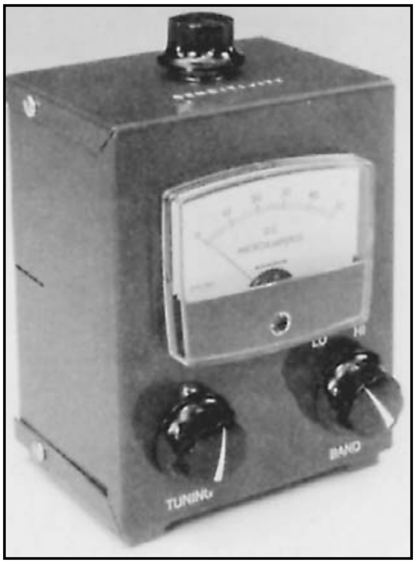

Fig 57—The RF current probe. The sensitivity control is mounted at the top of the instrument, with the tuning and band switches on the lower portion of the front panel. Frequency calibration of the tuning control was not considered necessary for the intended use of this particular instrument, but marks identifying the various amateur bands would be helpful. If the unit is provided with a calibrated dial, it can also be used as an absorption wavemeter.

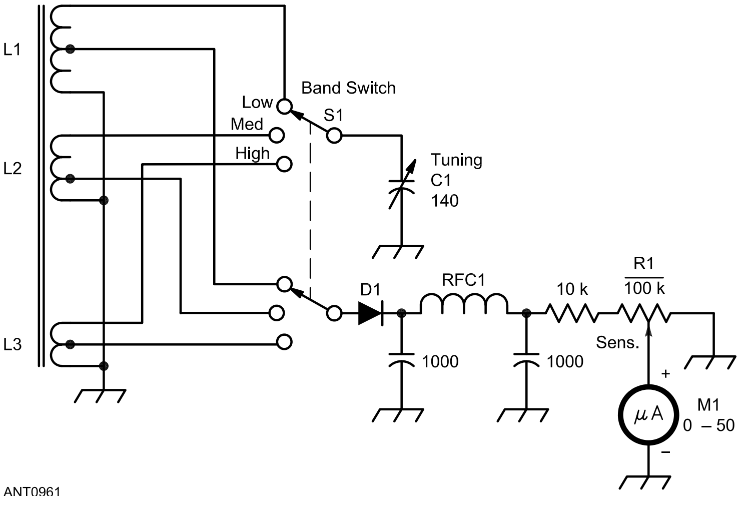

Fig 58—Schematic diagram of the RF current probe. Resistances are in ohms; k = 1000. Capacitances are in picofarads; fixed capacitors are silver mica. Be sure to ground the rotor of C1, rather than the stator, to avoid hand capacitance. L1, L2 and L3 are each close-wound with #22 enameled wire on a single ferrite rod, 4 inch long and 1/2 inch diameter, with µ = 125 (Amidon R61-50-400). Windings are spaced approximately 1/4 inch apart.

C1—Air variable, 6-140 pF; Hammarlund HF140 or equiv.
D1—Germanium diode; 1N34A, 1N270 or equiv.
L1—1.6-5 MHz; 30 turns, tapped at 3 turns from grounded end.
L2—5-20 MHz; 8 turns, tapped at 2 turns from grounded end.
L3—17-39 MHz; 2 turns, tapped at 1 turn.
M1—Any microammeter may be used. The one pictured is a Micronta meter, RadioShack no. 270-1751.
R1—Linear taper.
RFC1—1 mH; Miller no. 4642 or equiv. Value is not critical.
S1—Ceramic rotary switch, 1 section, 2 poles, 2 to 6 positions; Centralab PA2002 or PA2003 or equiv.

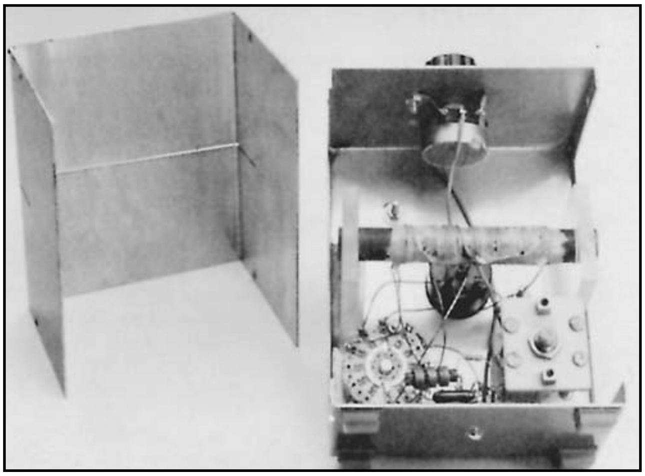

Fig 59—The current probe just before final assembly. Note that all parts except the ferrite rod are mounted on a single half of the 3 × 4 × 5-inch Minibox (Bud CU-2105B or equiv.). Rubber grommets are fitted in holes at the ends of the slot to accept the rod during assembly of the enclosure. Leads in the RF section should be kept as short as possible, although those from the rod windings must necessarily be left somewhat long to facilitate final assembly.

As shown in the photos, the circuit is constructed in a metal box. This enclosure shields the detector circuit from the electric field of the radio wave. A slot must be cut with a hacksaw across the back of the box, and a thin file may be used to smooth the cut. This slot is necessary to prevent the box from acting as a shorted turn.

Using the Probe

In measuring the current in a conductor, the ferrite rod should be kept at right angles to the conductor, and at a constant distance from it. In its upright or vertical position, this instrument is oriented for taking measurements in vertical conductors. It must be laid horizontal to measure current in horizontal conductors.

Numerous uses for the instrument are suggested in an earlier paragraph. In addition, the probe is an ideal instrument for checking the current distribution in antenna elements. It is also useful for measuring RF ground currents in radial systems. A buried radial may be located easily by sweeping the ground. Current division at junctions may be investigated. *Hot spots* usually indicate areas where additional radials would be effective.

Stray currents in conductors not intended to be part of the antenna system may often be eliminated by bonding or by changing the physical lengths involved. Guy wires and other unwanted parasitic elements will often give a tilt to the plane of polarization and can make a marked difference in front-to-back ratios. When the ferrite rod is oriented parallel to the electric field lines, there will be a sharp null reading that may be used to locate the plane of polarization quite accurately. When using the meter, remember that the magnetic field is at right angles to the electric field.

You may also use the current probe as a relative signal strength meter. When making measurements on a vertical antenna, locate the meter at least two wavelengths away, with the rod in a horizontal position. For horizontal antennas, hold the instrument at approximately the same height as the antenna, with the rod vertical.

Antenna Measurements

Of all the measurements made in Amateur Radio systems, perhaps the most difficult and least understood are various measurements of antennas. For example, it is relatively easy to measure the frequency and CW power output of a transmitter, the response of a filter, or the gain of an amplifier. These are all what might be called *bench measurements* because, when performed properly, all the factors that influence the accuracy and success of the measurement are under control. In making antenna measurements, however, the "bench" is probably your backyard. In other words, the environment surrounding the antenna can affect the results of the measurement.

Control of the environment is not at all as simple as it was for the bench measurement, because now the work area may be rather spacious. This section describes antenna measurement techniques that are closely allied to those used in an antenna measuring event or contest. With these procedures you can make measurements successfully and with meaningful results. These techniques should provide a better understanding of the measurement problems, resulting in a more accurate and less difficult task. The information in this section was provided by Dick Turrin, W2IMU, and was originally published in November 1974 *QST*.

SOME BASIC IDEAS

An antenna is simply a transducer or coupler between a suitable feed line and the environment surrounding it. In addition to the efficient transfer of power from feed line to environment, an antenna at VHF or UHF is most frequently required to concentrate the radiated power into a particular region of the environment.

To be consistent while comparing different antennas, you must standardize the environment surrounding the antenna. Ideally, you want to make measurements with the measured antenna so far removed from any objects causing environmental effects that it is literally in outer space—a very impractical situation. The purpose of the measurement techniques is therefore to simulate, under practical conditions, a *controlled environment*. At VHF and UHF, and with practical-size antennas, the environment can be controlled so that successful and accurate measurements can be made in a reasonable amount of space.

The electrical characteristics of an antenna that are most desirable to obtain by direct measurement are: (1) gain (relative to an isotropic source, which by definition has a gain of unity); (2) space-radiation pattern; (3) feed-point impedance (mismatch) and (4) polarization.

Polarization

In general the polarization can be assumed from the geometry of the radiating elements. That is to say, if the antenna consists of a number of linear elements (straight lengths of rod or wire that are resonant and connected to the feed point) the polarization of the electric field will be linear and polarized parallel to the elements. If the elements are not consistently parallel with each other, then the polarization cannot easily be assumed. The following techniques are directed to antennas having polarization that is essentially linear (in one plane), although the method can be extended to include all forms of elliptic (or mixed) polarization.

Feed-Point Mismatch

The feed-point mismatch, although affected to some degree by the immediate environment of the antenna, does not affect the gain or radiation characteristics of an antenna. If the immediate environment of the antenna does not affect the feed-point impedance, then any mismatch intrinsic to the antenna tuning reflects a portion of the incident power back to the source. In a receiving antenna this reflected power is reradiated back into the environment, and can be lost entirely.

In a transmitting antenna, the reflected power travels back down the feed line to the transmitter, where it changes the load impedance presented to that transmitter. The amplifier output controls are customarily altered during the normal tuning procedure to obtain maximum power transfer to the antenna. You can still use a mismatched antenna to its full gain potential, provided the mismatch is not so severe as to cause heating losses in the system, especially the feed line and matching devices. (See also the discussion of additional loss caused by SWR in Chapter 24.)

Similarly, a mismatched receiving antenna may be matched into the receiver front end for maximum power transfer. In any case you should clearly keep in mind that the feed-point mismatch does not affect the radiation characteristics of an antenna. It can only affect the system efficiency when heating losses are considered.

Why then do we include feed-point mismatch as part of the antenna characteristics? The reason is that for efficient system performance, most antennas are resonant transducers and present a reasonable match over a relatively narrow frequency range. It is therefore desirable to design an antenna, whether it be a simple dipole or an array of Yagis, such that the final single feed-point impedance is essentially resistive and matched to the feed line. Furthermore, in order to make accurate, absolute gain measurements, it is vital that the antenna under test accept all the power from a matched-source generator, or that the reflected power caused by the mismatch be measured and a suitable error correction for heating losses be included in the gain calculations. Heating losses may be determined from information contained in Chapter 24.

While on the subject of feed-point impedance, mention should be made of the use of *baluns* in antennas. A balun is simply a device that permits a lossless transition between a balanced system feed line or antenna and an unbalanced feed line or system. If the feed point of an antenna is symmetric, such as with a dipole, and it is desired to feed this antenna with an unbalanced feed line such as coax, you should provide a balun between the line and the feed point. Without the balun, current will be allowed to flow on the outside of the coax. The current on the outside of the feed line will cause radiation, and thus the feed line will become part of the antenna radiation system. In the case of beam antennas, where it is desired to concentrate the radiated energy is a specific direction, this extra radiation from the feed line will be detrimental, causing distortion of the expected antenna pattern. See Chapter 26 for additional details on this problem.

ANTENNA TEST SITE SET-UP AND EVALUATION

Since an antenna is a reciprocal device, measurements of gain and radiation patterns can be made with the test antenna used either as a transmitting or as a receiving antenna. In general and for practical reasons, the test antenna is used in the receiving mode, and the source or transmitting antenna is located at a specified fixed remote site and unattended. In other words the source antenna, energized by a suitable transmitter, is simply required to illuminate or flood the receiving site in a controlled and constant manner.

As mentioned earlier, antenna measurements ideally should be made under free-space conditions. A further restriction is that the illumination from the source antenna be a *plane wave* over the effective aperture (capture area) of the test antenna. A plane wave by definition is one in which the magnitude and phase of the fields are uniform, and in the test-antenna situation, *uniform over the effective area plane of the test antenna*. Since it is the nature of all radiation to expand in a spherical manner at great distance from the source, it would seem to be most desirable to locate the source antenna as far from the test site as possible. However, since for practical reasons the test site and source location will have to be near the earth and not in outer space, the environment must include the effects of the ground surface and other obstacles in the vicinity of both antennas. These effects almost always dictate that the test range (spacing between source and test antennas) be as short as possible consistent with maintaining a nearly error-free plane wave illuminating the test *aperture*.

A nearly error-free plane wave can be specified as one in which the phase and amplitude, from center to edge of the illuminating field over the test aperture, do not deviate by more than about 30° and 1 decibel, respectively. These conditions will result in a gain-measurement error of no more than a few percent less than the true gain. Based on the 30º phase error alone, it can be shown that the minimum range distance is approximately

$$S_{min} = 2\frac{D^2}{\lambda} \qquad \text{(Eq 23)}$$

where D is the largest aperture dimension and λ is the free-space wavelength in the same units as D. The phase error over the aperture D for this condition is $^1/_{16}$ wavelength.

Since aperture size and gain are related by

$$\text{Gain} = \frac{4\pi A_e}{\lambda^2} \qquad \text{(Eq 24)}$$

where A_e is the effective aperture area, the dimension D may be obtained for simple aperture configurations. For a square aperture

$$D_2 = G\frac{\lambda^2}{4\pi} \qquad \text{(Eq 25)}$$

that results in a minimum range distance for a square aperture of

$$S_{min} = G\frac{\lambda}{2\pi} \qquad \text{(Eq 26)}$$

and for a circular aperture of

$$S_{min} = G\frac{2\lambda}{\pi^2} \qquad \text{(Eq 27)}$$

For apertures with a physical area that is not well defined or is much larger in one dimension that in other directions, such as a long thin array for maximum directivity in one plane, it is advisable to use the maximum estimate of D from either the expected gain or physical aperture dimensions.

Up to this point in the range development, only the conditions for minimum range length, S_{min}, have been established, as though the ground surface were not present. This minimum S is therefore a necessary condition even under free-space environment. The presence of the ground further complicates the range selection, not in the determination of S but in the exact location of the source and test antennas above the earth.

It is always advisable to select a range whose intervening terrain is essentially flat, clear of obstructions, and of uniform surface conditions, such as all grass or all pavement. The extent of the range is determined by the illumination of the source antenna, usually a Yagi, whose gain is no greater than the highest gain antenna to be measured. For gain measurements the range consists essentially of the region in the beam of the test antenna. For radiation-pattern measurements, the range is considerably larger and consists of all that area illuminated by the source antenna, especially around and behind the test site. Ideally you should choose a site where the test-antenna location is near the center of a large open area and the source antenna is located near the edge where most of the obstacles (trees, poles, fences, etc.) lie.

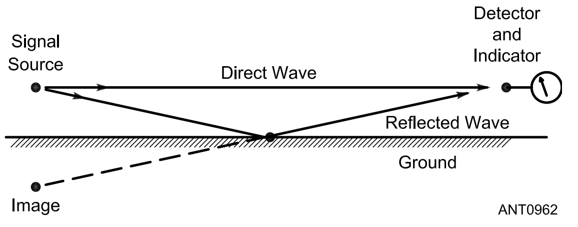

Fig 60—On an antenna test range, energy reaching the receiving equipment may arrive after being reflected from the surface of the ground, as well as by the direct path. The two waves may tend to cancel each other, or may reinforce one another, depending on their phase relationship at the receiving point.

The primary effect of the range surface is that some of the energy from the source antenna will be reflected into the test antenna, while other energy will arrive on a direct line-of-sight path. This is illustrated in **Fig 60**. The use of a flat, uniform ground surface assures that there will be essentially a mirror reflection, even though the reflected energy may be slightly weakened (absorbed) by the surface material (ground). In order to perform an analysis you should realize that horizontally polarized waves undergo a 180º phase reversal upon reflection from the earth. The resulting illumination amplitude at any point in the test aperture is the vector sum of the electric fields arriving from the two directions, the direct path and the reflected path.

If a perfect mirror reflection is assumed from the ground (it is nearly that for practical ground conditions at VHF/UHF) and the source antenna is isotropic, radiating equally in all directions, then a simple geometric analysis of the two path lengths will show that at various point in the vertical plane at the test-antenna site the waves will combine in different phase relationships. At some points the arriving waves will be in phase, and at other points they will be 180º out of phase. Since the field amplitudes are nearly equal, the resulting phase change caused by path length difference will produce an ampli-

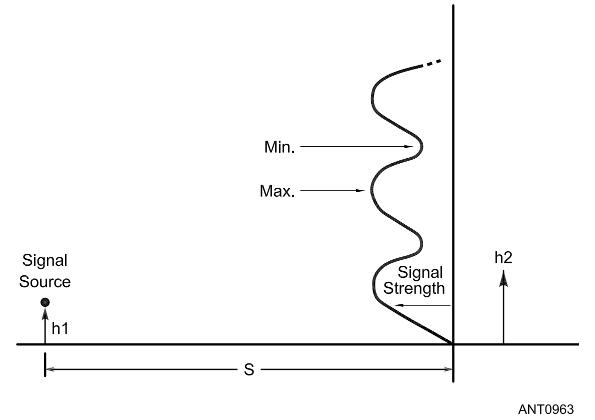

Fig 61—The vertical profile, or plot of signal strength versus test-antenna height, for a fixed height of the signal source above ground and at a fixed distance. See text for definitions of symbols.

tude variation in the vertical test site direction similar to a standing wave, as shown in **Fig 61**.

The simplified formula relating the location of h2 for maximum and minimum values of the two-path summation in terms of h1 and S is

$$h2 = n\frac{\lambda}{4}\times\frac{S}{h1} \qquad \text{(Eq 28)}$$

with n = 0, 2, 4, . . . for minimums and n = 1, 3, 5, . . . for maximums, and S is much larger than either h1 or h2.

The significance of this simple ground reflection formula is that it permits you to determine the approximate location of the source antenna to achieve a nearly plane-wave amplitude distribution *in the vertical direction* over a particular test *aperture size*. It should be clear from examination of the height formula that as h1 is decreased, the vertical distribution pattern of signal at the test site, h2, expands. Also note that the signal level for h2 equal to zero is always zero on the ground regardless of the height of h1.

The objective in using the height formula then is, given an effective antenna aperture to be illuminated from which a minimum S (range length) is determined and a suitable range site chosen, to find a value for h1 (source antenna height). The required value is such that the *first* maximum of vertical distribution at the test site, h2, is at a practical distance above the ground, and at the same time the signal amplitude over the aperture in the vertical direction does not vary more than about 1 dB. This last condition is not sacred but is closely related to the particular antenna under test.

In practice these formulas are useful only to initialize the range setup. A final check of the vertical distribution at the test site must be made by direct measurement. This measurement should be conducted with a small low-gain but unidirectional probe antenna such as a corner reflector or 2-element Yagi that you move along a vertical line over the intended aperture site. Care should be exercised to minimize the effects of local environment around the probe antenna and that the beam of the probe be directed at the source antenna at all times for maximum signal. A simple dipole is undesirable as a probe antenna because it is susceptible to local environmental effects.

The most practical way to instrument the vertical distribution measurement is to construct some kind of vertical track, preferably of wood, with a sliding carriage or platform that may be used to support and move the probe antenna. It is assumed of course that a stable source transmitter and calibrated receiver or detector are available so variations of the order of $^1/_2$ dB can be clearly distinguished.

Once you conduct these initial range measurements successfully, the range is now ready to accommodate any aperture size less in vertical extent than the largest for which S_{min} and the vertical field distribution were selected. Place the test antenna with the center of its aperture at the height h2 where maximum signal was found. Tilt the test antenna tilted so that its main beam is pointed in the direction of the source antenna. The final tilt is found by observing the receiver output for maximum signal. This last process must be done empirically since the apparent location of the source is somewhere between the actual source and its image, below the ground.

An example will illustrate the procedure. Assume that we wish to measure a 7-foot diameter parabolic reflector antenna at 1296 MHz (λ = 0.75 foot). The minimum range distance, S_{min}, can be readily computed from the formula for a circular aperture.

$$S_{min} = 2\frac{D^2}{\lambda} = 2\times\frac{49}{0.75} = 131 \text{ feet}$$

Now a suitable site is selected based on the qualitative discussion given before.

Next determine the source height, h1. The procedure is to choose a height h1 such that the first minimum above ground (n = 2 in formula) is at least two or three times the aperture size, or about 20 feet.

$$h1 = n\frac{\lambda\ S}{4\ h1} = 2\times\frac{0.75}{4}\times\frac{131}{20} = 2.5 \text{ feet}$$

Place the source antenna at this height and probe the vertical distribution over the 7-foot aperture location, which will be about 10 feet off the ground.

$$h2 = n\frac{\lambda\ S}{4\ h1} = 1\times\frac{0.75}{4}\times\frac{131}{2.5} = 9.8 \text{ feet}$$

Plot the measured profile of vertical signal level versus height. From this plot, empirically determine whether the 7-foot aperture can be fitted in this profile such that the 1-dB variation is not exceeded. If the variation exceeds

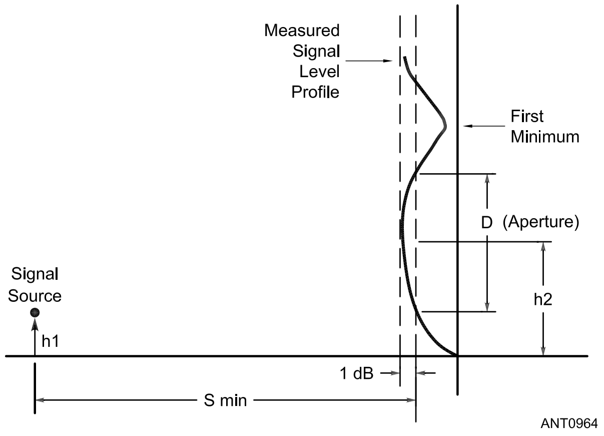

Fig 62—Sample plot of a measured vertical profile.

1 dB over the 7-foot aperture, the source antenna should be lowered and h2 raised. Small changes in h1 can quickly alter the distribution at the test site. **Fig 62** illustrates the points of the previous discussion.

The same set-up procedure applies for either horizontal or vertical linear polarization. However, it is advisable to check by direct measurement at the site for each polarization to be sure that the vertical distribution is satisfactory. Distribution probing in the horizontal plane is unnecessary as little or no variation in amplitude should be found, since the reflection geometry is constant. Because of this, antennas with apertures that are long and thin, such as a stacked collinear vertical, should be measured with the long dimension parallel to the ground.

A particularly difficult range problem occurs in measurements of antennas that have depth as well as cross-sectional aperture area. Long end-fire antennas such as long Yagis, rhombics, V-beams, or arrays of these antennas, radiate as *volumetric arrays* and it is therefore even more essential that the illuminating field from the source antenna be reasonably uniform in depth as well as plane wave in cross section. For measuring these types of antennas it is advisable to make several vertical profile measurements that cover the depth of the array. A qualitative check on the integrity of the illumination for long end-fire antennas can be made by moving the array or antenna axially (forward and backward) and noting the change in received signal level. If the signal level varies less than 1 or 2 dB for an axial movement of several wavelengths then the field can be considered satisfactory for most demands on accuracy. Large variations indicate that the illuminating field is badly distorted over the array depth and subsequent measurements are questionable. It is interesting to note in connection with gain measurements that any illuminating field distortion will always result in measurements that are lower than true values.

ABSOLUTE GAIN MEASUREMENT

Having established a suitable range, the measure-

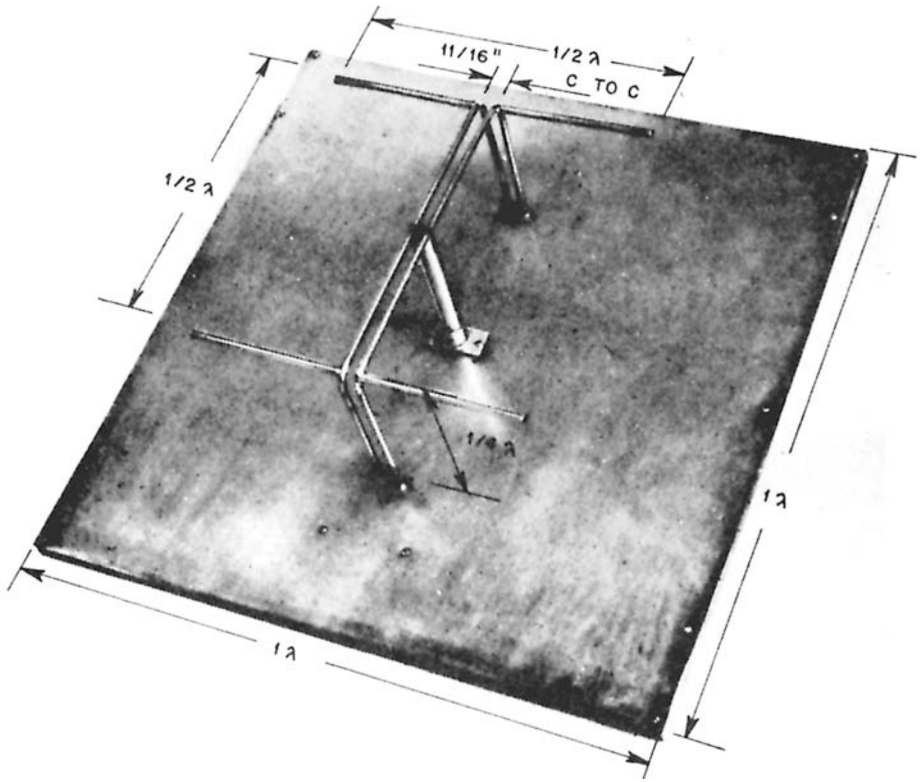

Fig 63—Standard-gain antenna. When accurately constructed for the desired frequency, this antenna will exhibit a gain of 7.7 dB over a dipole radiator, plus or minus 0.25 dB. In this model, constructed for 432 MHz, the elements are $^{3}/_{8}$-inch diameter tubing. The phasing and support lines are of $^{5}/_{16}$-inch diameter tubing or rod.

ment of gain relative to an isotropic (point source) radiator is almost always accomplished by direct comparison with a calibrated standard-gain antenna. That is, the signal level with the test antenna in its optimum location is noted. Then you remove the test antenna and place the standard-gain antenna with its aperture at the center of location where the test antenna was located. Measure the difference in signal level between the standard and the test antennas and add to or subtract from the gain of the standard-gain antenna to obtain the absolute gain of the test antenna. Here, *absolute* means with respect to a point source with a gain of unity, by definition. The reason for using this reference rather than a dipole, for instance, is that it is more useful and convenient for system engineering. We assume that both standard and test antennas have been carefully matched to the appropriate impedance and an accurately calibrated and matched detecting device is being used.

A standard-gain antenna may be any type of unidirectional, preferably planar-aperture, antenna, which has been calibrated either by direct measurement or in special cases by accurate construction according to computed dimensions. A standard-gain antenna has been suggested by Richard F. H. Yang (see Bibliography). Shown in **Fig 63**, it consists of two in-phase dipoles $^1/_2$ λ apart and backed up with a ground plane 1 λ square.

In Yang's original design, the stub at the center is a balun formed by cutting two longitudinal slots of $^1/_8$-inch width, diametrically opposite, on a $^1/_4$-λ section of $^7/_8$-inch rigid 50-Ω coax. An alternative method of feeding is to feed RG-8 or RG-213 coax through slotted $^7/_8$-inch copper tubing. Be sure to leave the outer jacket on the coax to insulate it from the copper-tubing balun section. When constructed accurately to scale for the frequency of interest, this type of standard will have an absolute gain of 9.85 dBi (7.7 dBd gain over a dipole in free space) with an accuracy of ± 0.25 dB.

RADIATION-PATTERN MEASUREMENTS

Of all antenna measurements, the radiation pattern is the most demanding in measurement and the most difficult to interpret. Any antenna radiates to some degree in all directions into the space surrounding it. Therefore, the radiation pattern of an antenna is a three-dimensional representation of the magnitude, phase and polarization. In general, and in practical cases for Amateur Radio communications, the polarization is well defined and only the magnitude of radiation is important.

Furthermore, in many of these cases the radiation in one particular plane is of primary interest, usually the plane corresponding to that of the Earth's surface, regardless of polarization. Because of the nature of the range setup, measurement of radiation pattern can be successfully made only in a plane nearly parallel to the earth's surface. With beam antennas it is advisable and usually sufficient to take two radiation pattern measurements, one in the polarization plane and one at right angles to the plane of polarization. These radiation patterns are referred to in antenna literature as the principal E-plane and H-plane patterns, respectively. *E-plane* means parallel to the electric field that is the polarization plane and *H-plane* means parallel to the magnetic field in free space. The electric field and magnetic field are always perpendicular to each other in a plane wave as it propagates through space.

When the antenna is located over real earth, the terms *Azimuth* and *elevation* planes are commonly used, since the frame of reference is the Earth itself, rather than the electric and magnetic fields in free space. For a horizontally polarized antenna such as a Yagi mounted with its elements parallel to the ground, the azimuth plane is the E-plane and the elevation plane is the H-plane.

The technique to obtain these patterns is simple in procedure but requires more equipment and patience than does making a gain measurement. First, a suitable mount is required that can be rotated in the azimuth plane (horizontal) with some degree of accuracy in terms of azimuth-angle positioning. Second, a signal-level indicator calibrated over at least a 20-dB dynamic range with a readout resolution of at least 2 dB is required. A dynamic range of up to about 40 dB would be desirable but does not add greatly to the measurement significance.

With this much equipment, the procedure is to locate first the area of maximum radiation from the beam antenna by carefully adjusting the azimuth and elevation positioning. These settings are then arbitrarily assigned an azimuth angle of zero degrees and a signal level of zero decibels. Next, without changing the elevation setting (tilt of the rotating axis), the antenna is carefully rotated in azimuth in small steps that permit signal-level readout of 2 or 3 dB per step. These points of signal level corresponding with an azimuth angle are recorded and plotted on polar coordinate paper. A sample of the results is shown on ARRL coordinate paper in **Fig 64**.

On the sample radiation pattern the measured points are marked with an X and a continuous line is drawn in, since the pattern is a continuous curve. Radiation patterns should preferably be plotted on a logarithmic radial scale, rather than a voltage or power scale. The reason is that the log scale approximates the response of the ear to signals in the audio range. Also many receivers have AGC systems that are somewhat logarithmic in response; therefore the log scale is more representative of actual system operation.

Having completed a set of radiation-pattern measurements, one is prompted to ask, "Of what use are they?" The primary answer is as a diagnostic tool to determine if the antenna is functioning as it was intended to. A second answer is to know how the antenna will discriminate against interfering signals from various directions.

Consider now the diagnostic use of the radiation patterns. If the radiation beam is well defined, then there is an approximate formula relating the antenna gain to the

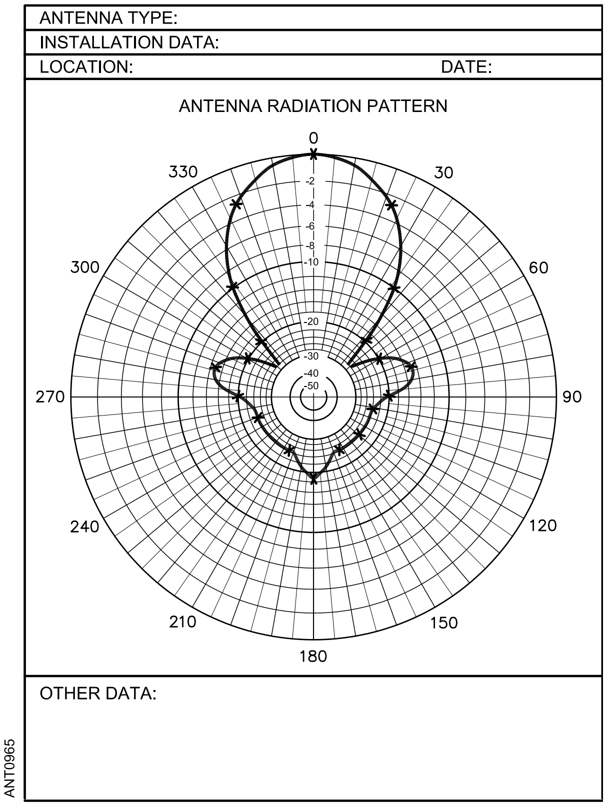

For more information see QST, July 1980, p.26 Copyright 1980, ARRL Inc.

Fig 64—Sample plot of a measured radiation pattern, using techniques described in the text. The plot is on coordinate paper available from ARRL HQ. The form provides space for recording significant data and remarks.

measured half-power beamwidth of the E- and H-plane radiation patterns. The half-power beamwidth is indicated on the polar plot where the radiation level falls to 3 dB below the main beam 0-dB reference on either side. The formula is

$$\text{Gain}\left(\text{dBi}\right) \cong \frac{41{,}253}{\theta_E \phi_H} \qquad \text{(Eq 29)}$$

where θ_E and ϕ_H are the half-power beamwidths in degrees of the E- and H-plane patterns, respectively. This equation assumes a lossless antenna system, where any side-lobes are well suppressed.

To illustrate the use of this equation, assume that we have a Yagi antenna with a boom length of two wavelengths. From known relations (described in Chapter 11) the expected free-space gain of a Yagi with a boom length of 2 λ is about 13 dBi; its gain, G, equals 20. Using the above relationship, the product of $\theta_E \times \phi_H \approx 2062$ square degrees. Since a Yagi produces a nearly symmetric beam shape in cross section, $\theta_E = \phi_H = 45°$. Now if the measured values of θ_E and ϕ_H are much larger than 45°, then the gain will be much lower than the expected 13 dBi.

As another example, suppose that the same antenna (a 2-wavelength-boom Yagi) gives a measured gain of 9 dBi but the radiation pattern half power beamwidths are approximately 45°. This situation indicates that although the radiation patterns seem to be correct, the low gain shows inefficiency somewhere in the antenna, such as lossy materials or poor connections.

Large broadside collinear antennas can be checked for excessive phasing-line losses by comparing the gain computed from the radiation patterns with the direct-measured gain. It seems paradoxical, but it is indeed possible to build a large array with a very narrow beamwidth indicating high gain, but actually having very low gain because of losses in the feed distribution system.

In general, and for most VHF/UHF Amateur Radio communications, gain is the primary attribute of an antenna. However, radiation in other directions than the main beam, called *sidelobe radiation*, should be examined by measurement of radiation patterns for effects such as nonsymmetry on either side of the main beam or excessive magnitude of sidelobes. (Any sidelobe that is less than 10 dB below the main beam reference level of 0 dB should be considered excessive.) These effects are usually attributable to incorrect phasing of the radiating elements or radiation from other parts of the antenna that was not intended, such as the support structure or feed line.

The interpretation of radiation patterns is intimately related to the particular type of antenna under measurement. Reference data should be consulted for the antenna type of interest, to verify that the measured results are in agreement with expected results.

To summarize the use of pattern measurements, if a beam antenna is first checked for gain (the easier measurement to make) and it is as expected, then pattern measurements may be academic. However, if the gain is lower than expected it is advisable to make pattern measurements to help determine the possible causes for low gain.

Regarding radiation pattern measurements, remember that the results measured under proper range facilities will not necessarily be the same as observed for the same antenna at a home-station installation. The reasons may be obvious now in view of the preceding information on the range setup, ground reflections, and the vertical-field distribution profiles. For long paths over rough terrain where many large obstacles may exist, the effects of ground reflection tend to become diffused, although they still can cause unexpected results. For these reasons it is usually unjust to compare VHF/UHF antennas over long paths.

Vector Network Analyzers

The process of building and properly tuning a phased array often involves making a number of different measurements to achieve a desired level of performance, as was pointed out in Chapter 8, Multielement Arrays. This section was written by Rudy Severns, N6LF.

After erecting an array we would like to measure the resonant frequency of each element, the self-impedances of each element and the mutual impedances between the elements. We will also want to know these impedances over the whole operating band to help design a feed network. When building the feed network, we may need to check the values and Qs of the network elements and we will want to determine the electrical lengths of transmission lines.

Final tuning of the array requires that the relative current amplitudes and phases in each element be measured and adjusted, if necessary. We also will want to determine the SWR at the feed point. Doing all of this even moderately well can require quite a bit of equipment, some of which is heavy and requires ac line power. This can be a nuisance in the field, especially if the weather is not cooperating.

Professionals make these measurements by employing a *vector network analyzer* (VNA) or the somewhat simpler, *reflection-transmission test set*. These instruments can make all the necessary measurements quickly and with great accuracy. However, in the past VNAs have been very expensive, out of reach for general amateur use. But thanks to modern digital technology VNAs that work with a laptop computer are now becoming available at prices an amateur might consider. It's even possible to homebrew a VNA[1] with performance that approaches a professional instrument. Considering the cost of even a simple array, investment in a VNA makes sense.

VNAs are based on reflection and transmission measurements. To use a VNA it is very helpful to have a basic understanding of *Scattering Parameters* (S-parameters). Microwave engineers have long used these because they have to work with circuits that are large in terms of wavelength, where measurements of forward and reflected power are easy.

HF arrays are also large in terms of wavelength. The techniques for measuring forward and reverse powers work well even at 160 meters. For example, even though the array elements may be 100 feet apart, you can place your instruments in a central location and run cables out to each element. The effect of the cables from the VNA to the elements can be absorbed in the initial calibration procedure so the measurements read out at the VNA are effectively those at each element. In other words, the measurement reference points can be placed electrically at the base of the element, regardless of the physical location of the instrumentation and the interconnecting cables.

In a completed array with its feed network, the network can be excited by the VNA at the feed point and the relative current amplitudes and phases at each element can be measured over a frequency band. Then, adjustments can be made as needed. When the final values for the current amplitudes and phases are known, these values can be put back into an array model in a program like *EZNEC* to determine the pattern of the array across the whole frequency band.

S-PARAMETERS

In Chapter 24, Transmission Lines, the reflection coefficient rho (ρ) is defined as the ratio of the reflected voltage (V_r) to the incident voltage (V_i):

$$\rho = \frac{V_r}{V_i} \qquad \text{(Eq 30)}$$

If we know the load impedance (Z_L) and the transmission line impedance (Z_0) we can calculate ρ from:

$$\rho = \frac{Z_L - Z_0}{Z_L + Z_0} \qquad \text{(Eq 31)}$$

Keep in mind that ρ is a complex number (a vector), which we represent by either amplitude and phase (|Z|, θ) or by real and imaginary parts ($R \pm j\,X$). The two representations are equivalent. From ρ we can then calculate SWR. That's very handy, but here we want to do something different. If we have an instrument that measures ρ and we know Z_0 then we can determine Z_L from:

$$Z_L = Z_0 \left(\frac{1+\rho}{1-\rho} \right) \qquad \text{(Eq 32)}$$

Measuring ρ is one of the things that VNAs do very well. With a VNA, the measurement can be made at one end of a long transmission line with the load at the other end. The effect of the line can be calibrated out, as mentioned above, so that we are in effect measuring right at the load.

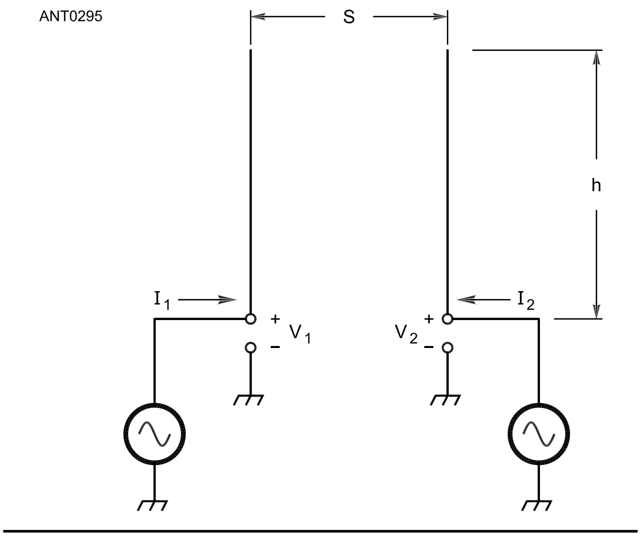

Fig 65—A 2-element array, where h is the element height and S is the spacing between the elements.

This approach can be used directly to measure the impedance and resonant frequency of a single element. By open and short circuiting elements in an array we can determine the mutual as well as self impedances for, and between, all the elements. We can also use this approach to measure component values, inductor Qs, etc.

This is an example of a *one-port* measurement; that is, a load at the end of a transmission line. However, a multi-element array actually behaves as a multi-port network, so to get the most out of a VNA, you need to generalize the above procedure. This is where S-parameters come into play. To illustrate the principles we will use a simple 2-element array like that shown in **Fig 65**.

To design a feed network to drive this array we need to know the input impedance of each element (Z_1 and Z_2) as a function of the drive currents (I_1 and I_2). The input impedances will depend on the self impedance of each element, the coupling between them (the mutual impedance) and the drive currents in each element. To manage this problem we can represent a 2-element array as a two-port network, as shown in **Fig 66**. And we can relate the port voltages, currents and impedances with Eq 33:

$$\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2 \\ V_2 &= Z_{21}I_1 + Z_{22}I_2 \end{aligned} \qquad \text{(Eq 33)}$$

Normally we know I_1 and I_2 from the design of the array, but we need to determine the resulting element impedances. That's the challenge. Fortunately, an array is a linear network, so $Z_{12} = Z_{21}$, which means we need only determine three variables: the self impedances Z_{11} and Z_{22} and the mutual impedance, Z_{12}.

Once we know Z_{11}, Z_{12}, Z_{22} and are given I_1 and I_2, we can determine the feed-point impedances at each element from:

$$\begin{aligned} Z_1 &= Z_{11} + \left(\frac{I_2}{I_1}\right) Z_{12} \\ Z_2 &= Z_{22} + \left(\frac{I_1}{I_2}\right) Z_{12} \end{aligned} \qquad \text{(Eq 34)}$$

This is the conventional approach. However, there are some problems here. We have to be able to accurately measure either voltages and currents or impedances in multiple elements that may be separated by large fractions of a wavelength. In addition, accurate measurements of current, voltage and impedance become increasingly more difficult as we go up in frequency.

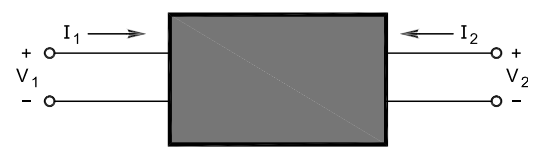

Fig 66—Two-port representation of currents and voltages in the 2-element array in Fig 65.

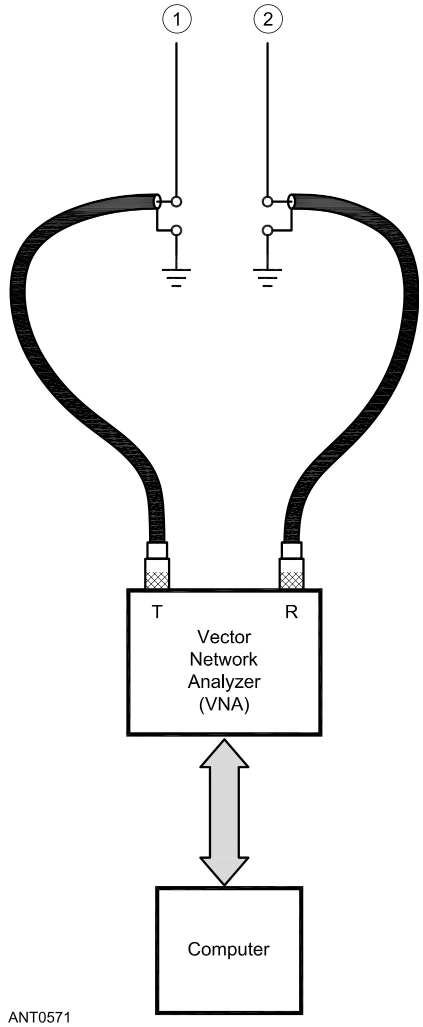

Fig 67—Test setup to measure a 2-element array using a VNA.

It turns out that we can get the information more easily by measuring incident and reflected voltages at the ports and from those measurements determine the feed-point impedances. A VNA is an instrument for measuring these voltages. It turns out to be easier to measure the ratios of two voltages rather than their absolute values.

The measurement setup using a VNA for a 2-element array is shown in **Fig 67**. VNAs usually have at least two RF connections: the transmit port (T) and the receive port (R). Professional units may have more RF connections. The T output provides an signal from a 50-Ω source and the R port is a detector with a 50-Ω input impedance. Basically we have a transmitter and a receiver. The transmit port uses a directional coupler to provide measurements of the forward and reflected signals at that output. The receive port measures the signal transmitted through the network. Transmission lines usually have $Z_0 = 50\ \Omega$ and may be of any length required by the size of the array.

Using incident and reflected voltages, the two-port network representation is now changed, as shown in **Fig 68**, where:

V_{1i} = incident voltage at port 1
V_{1r} = reflected voltage at port 1
V_{2i} = incident voltage at port 2
V_{2r} = reflected voltage at port 2

In a manner analogous to Eq 33, we can write an expression in terms of the incident and reflected voltages:

$$\begin{aligned} b_1 &= S_{11}a_1 + S_{12}a_2 \\ b_2 &= S_{21}a_1 + S_{22}a_2 \end{aligned} \qquad \text{(Eq 35)}$$

where:

$$\begin{aligned} &a_1 \frac{V_{1i}}{\sqrt{Z_0}} \qquad b_1 \frac{V_{1r}}{\sqrt{Z_0}} \\ &a_2 = \frac{V_{2i}}{\sqrt{Z_0}} \qquad b_2 = \frac{V_{2r}}{\sqrt{Z_0}} \end{aligned} \qquad \text{(Eq 36)}$$

We see that the a_n and b_n are simply the incident and reflected voltages at the two ports divided by $\sqrt{Z_o}$. Because this is a linear network, $S_{21} = S_{12}$.

What are the S_{ij} quantities? These are called the *S-parameters*, which are defined by:

$$\begin{aligned} S_{11} &\equiv \left.\frac{b_1}{a_1}\right|_{a_2=0} = \left.\frac{V_{1r}}{V_{1i}}\right|_{V_{2i}=0} \\ S_{21} &\equiv \left.\frac{b_2}{a_1}\right|_{a_2=0} = \left.\frac{V_{2r}}{V_{1i}}\right|_{V_{2i}=0} \\ S_{12} &\equiv \left.\frac{b_1}{a_2}\right|_{a_1=0} = \left.\frac{V_{1r}}{V_{2i}}\right|_{V_{1i}=0} \\ S_{22} &\equiv \left.\frac{b_2}{a_2}\right|_{a_1=0} = \left.\frac{V_{2r}}{V_{2i}}\right|_{V_{1i}=0} \end{aligned} \qquad \text{(Eq 37)}$$

Note that the S_{ij} parameters are all ratios of reflected and incident voltages, and they are usually complex numbers. The condition that $a_2 = 0 = V_{2i}$ is the same as saying that port 2 is terminated in a load equal to Z_0 and the network is excited at port 1. This means there is no reflection from the load on port 2, which makes $V_{2i} = 0$. Similarly, if we terminate port 1 with Z_0 and excite port 2, then $V_{1i} = 0 = a_1$.

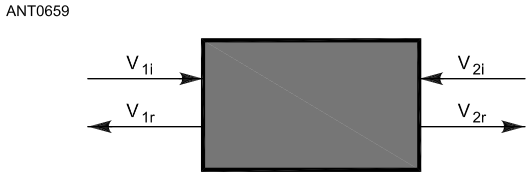

Fig 68—Two-port network with incident and reflected waves.

If we compare Eq 30 to the first line of Eq 37 we see that $S_{11} = \rho_1$, the reflection coefficient at port 1. We can now restate Eq 32 in terms of S_{11}:

$$Z = Z_0\left(\frac{1+S_{11}}{1-S_{11}}\right) \qquad \text{(Eq 38)}$$

where Z is the impedance looking into port 1 with port 2 terminated in Z_0. In the case where port 2 does not exist—that is, you are measuring a single element (for example, measuring element 1 with element 2 open-circuited) or a component, then Z is simply the self impedance (Z_{11} in Eq 33). Since S_{11} is a standard measurement for VNAs you can calculate Z using Eq 38. In many cases the VNA software will do this calculation for you automatically. You can also measure element 2 with element 1 open and determine Z_{22}.

S_{21} represents the ratio of the signal coming out of port 2 (V_{2r}) to the input signal on port 1 (V_{1i}) and is another standard VNA measurement. S_{21} is a measurement of the signal transmission between the ports through the network, or in the case of an array, the signal transmission due to the coupling between the elements. Again, port 2 is terminated in Z_0.

A full-feature VNA will measure all the S_{ij} parameters at once, but most of the lower-cost units of interest to amateurs are what we call *reflection-transmission test sets*. What this means is that they only measure S_{11} and S_{21}. To obtain S_{22} and S_{12} we have to interchange the test cables at the array elements (see Fig 67) and run the measurements again. Normally the software will accommodate this as a second entry and we end up with the full set of S_{ij} parameters.

If we do run a full set of S_{ij} parameters then we can transform these to Z_{ij} (Eq 33) using the following expressions, assuming that $S_{21} = S_{12}$:

$$\begin{aligned} Z_{11} &= \frac{(1+S_{11})(1-S_{22})+S_{12}^{\,2}}{(1-S_{11})(1-S_{22})-S_{12}^{\,2}} \\ Z_{22} &= \frac{(1-S_{11})(1+S_{22})+S_{12}^{\,2}}{(1-S_{11})(1-S_{22})-S_{12}^{\,2}} \\ Z_{12} &= \frac{(2S_{12})}{(1-S_{11})(1-S_{22})-S_{12}^{\,2}} \end{aligned} \qquad \text{(Eq 39)}$$

The example to this point has been for a 2-element array. The S-parameters can be determined for an array with any number of elements. In an n-port S-parameter measurement, all ports are terminated in Z_0 at the same time. Measurements are made between one set of ports at a time and repeated until all pairs of ports are measured.

ARRAY MEASUREMENT EXAMPLE

A good way to illustrate the use of a VNA for array measurements is to work through an example with a real array. **Fig 69** is a picture of a 2-element 20-meter phased array built by Mark Perrin, N7MQ.

Each element is λ/4 (self resonant at 14.150 MHz) and spaced λ/4 (17 feet 5 inches). In the ideal case, both elements would have the same current amplitude with a 90º phase difference. This gives the cardioid pattern shown in Chapter 8, Multielement Arrays. There are many schemes for correctly feeding such an array. The one used in this example uses two different 75-Ω transmission lines (one λ/4 and the other λ/2, electrically), as described by Roy Lewallen, W7EL.[2,3]

The first task is to resonate the elements individually. With the VNA set to measure S_{11} phase, we will get a graph like that shown in **Fig 70**.

At the λ/4 resonant frequency (f_r) we will see a sharp phase transition as we go from –180° to +180°. This is typical of any series resonant circuit. The length of each element is adjusted until the desired f_r is achieved. This is a very sensitive measurement. You can see the shift in f_r due to the wind blowing, the length of the element changing as it heats up in the sun or any interactions between the feed line and the antenna as you move the feed line around. In fact this is very good point in the process to make sure everything is mechanically stable and free of unexpected couplings. Usually you will find it necessary place choke baluns on each element to reduce stray coupling.

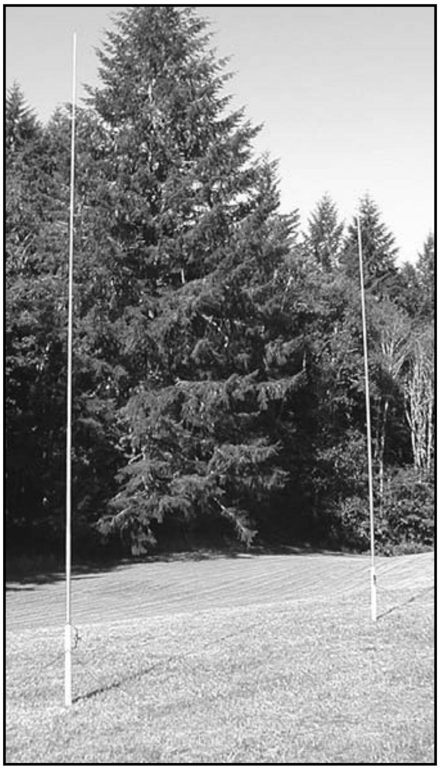

Fig 69—2-element 20-meter phased array (*Photo courtesy N7MQ*).

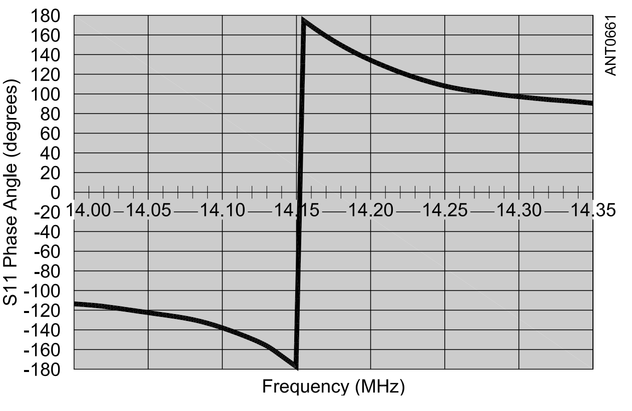

Fig 70—S_{11} phase plot for an individual element.

The next step is to determine the self (Z_{11} and Z_{22}) and mutual (Z_{12}) impedances from which the actual driving point impedances present when the array is excited can be determined. See Chapter 8, Multielement Arrays. There are two ways to go.

First, we can simply use the VNA as an impedance bridge—ie, make two S_{11} measurements at one element, first with the other element open (Z_{11} or Z_{22}) and then with it shorted (Z_1 or Z_2). We can convert the S_{11} measurements to impedances using Eq 38. The value for Z_{12} can be obtained from Eq 40:

$$\begin{aligned} Z_{12} &= \pm\sqrt{Z_{11}\left(Z_{11} - Z_1\right)} \\ Z_{12} &= \pm\sqrt{Z_{22}\left(Z_{22} - Z_2\right)} \end{aligned} \qquad \text{(Eq 40)}$$

The second approach is to do a full two-port S-parameter measurement (S_{11}, S_{21}, S_{12} and S_{22}) and derive the impedances using Eq 38. Both approaches will work but the second approach has the advantage that the ± ambiguity in Eq 40 is eliminated.

For this example, the impedance values from the measurements at 14.150 MHz, turn out to be:

$$\begin{aligned} Z_{11} &= 51.4 + j\,0.35 \\ Z_{22} &= 50.3 + j\,0.299 \\ Z_{12} &= 15.06 - j\,19.26 \end{aligned} \qquad \text{(Eq 41)}$$

With these values we can now determine the feed-point impedances from:

$$\begin{aligned} Z_1{}' &= Z_{11} + \frac{I_2}{I_1} Z_{12} \\ Z_2{}' &= Z_{22} + \frac{I_1}{I_2} Z_{12} \\ \frac{I_1}{I_2} &= -j \end{aligned} \qquad \text{(Eq 42)}$$

Note that $-j$ represents the 90° phase shift between the currents. Substituting the values from Eq 41 into Eq 42:

$$\begin{aligned} Z_1{}' &= 32.09 - j\,14.7 \\ Z_2{}' &= 69.6 + j\,15.32 \end{aligned} \qquad \text{(Eq 43)}$$

With these impedances in hand we can now design the feed network. In this particular example however, we have decided to use the λ/4 and λ/2 cables as described by Lewallen[2,3] and accept the results. So we now proceed to cut and trim the two cables to length.

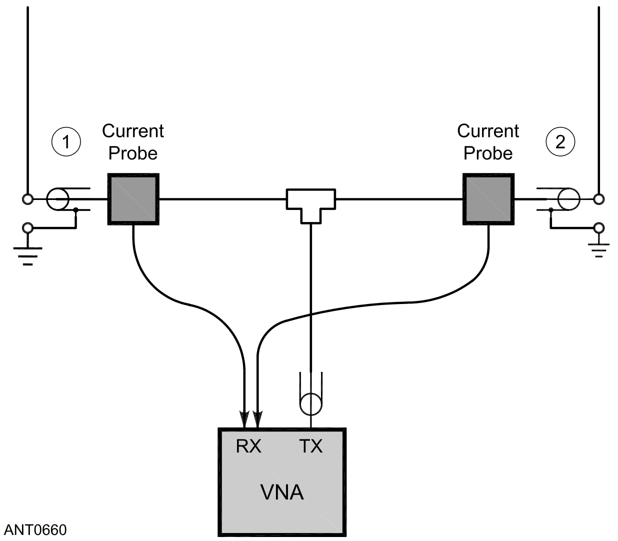

Fig 71—Current phase and amplitude ratio test setup.

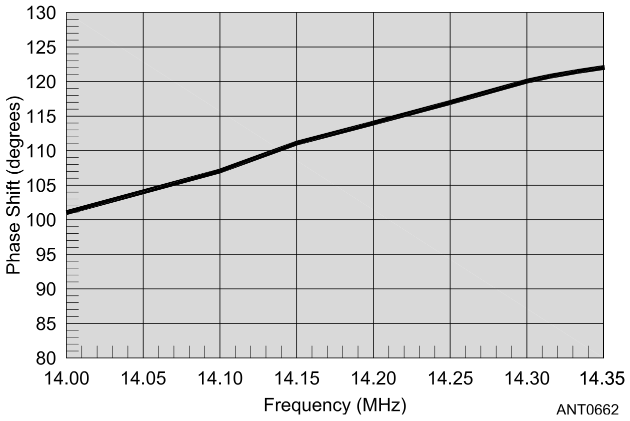

Fig 73—Measured relative current phase shift over the 20-meter band.

Again, there are two ways to go. First we can determine the frequency at which each cable is λ/4 long. At this point the input impedance of the cable will be equivalent to a series-resonant circuit and we can simply measure the phase of S_{11} as we did earlier for f_r and get a plot like that shown in Fig 70. In this example the λ/4 resonant frequencies of the two cables are 7.075 MHz and 14.150 MHz.

The second approach would be to measure S_{21} for each cable at 14.150 MHz. The phase shift in S_{21} tells you how long the cable is, in degrees, at a given frequency. Because there is a small variation in cable characteristics with frequency (dispersion) this approach is slightly more accurate since it is done at the desired operating frequency. But this is not very large effect at HF.

This brings us down to the final measurements, which are to check that the relative current amplitudes and phases between the two elements are correct. We can then determine the feed-point SWR. The phase and amplitude ratios are made using the S_{12} capability of the VNA and the test setup shown in **Fig 71**.

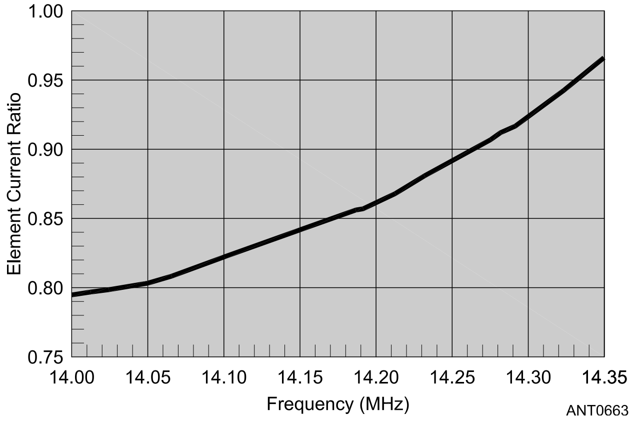

Fig 72—Measured element current ratio over the 20-meter band.

The VNA transmit port is connected to the normal feed point. A current sensor (see Chapter 8, Multielement Arrays, for a discussion of current sensors) is inserted at the base of element 1 and the output of the sensor is returned to the detector or receive port of the VNA. A calibration run is then made to normalize this path. That makes it the reference.

Next, the current sensor is shifted to element 2. The amplitude and phase plots for S_{12} obtained at this point will be the desired relative phase shift and amplitude ratio between the currents in the array when driven at the normal feed point. **Figs 72** and **73** show the behavior of the example array over the 20-meter band. Note that the amplitude ratio has been converted from dB. We can now use these values in a *EZNEC* model of the array to determine the actual radiation pattern.

Obviously the W7EL feed scheme is not perfect, but it has a definite advantage of simplicity. If better performance is desired we can use the values of Z_1' and Z_2' determined earlier to design and fabricate a new feed network and then proceed to evaluate its performance in the same way.

The final measurement is to connect the transmit port of the VNA to the feed point and measure S_{11}. From this we can calculate the SWR:

$$\mathrm{SWR} = \frac{1+|S_{11}|}{1-|S_{11}|} \qquad \text{(Eq 44)}$$

In this example, the return loss, $|S_{11}|$, is about –19 dB over the entire 20-meter band. This corresponds to SWR= 1.25:1.

Notes

[1] Paul Kiciak, N2PK, **http://n2pk.com**.

[2] Roy Lewallen, W7EL, *QST*, Aug 1979, Technical Correspondence, pp 42-43.

[3] Orr and Cowan, *Vertical Antennas*, Radio Amateur Call Book, 1986, pp 148-150.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below.

A. Bailey, "The Antenna Lab, Parts 1 and 2," *Radio Communication*, Aug and Sep 1983.

L. Blake, *Transmission Lines and Waveguides* (New York: John Wiley & Sons, 1969), pp 244-251.

J. H. Bowen, "A Calorimeter for VHF and UHF Power Measurements, *QST*, Dec 1975, pp 11-13.

W. Bruene, "An Inside Picture of Directional Wattmeters," *QST*, Apr 1959.

CCIR Recommendation 368, Documents of the CCIR XII Plenary assembly, ITU, Geneva, 1967.

J. Carr, "Find Fault with Your Coax," *73*, Oct 1984, pp 10-14.

D. DeMaw, "In-Line RF Power Metering," *QST*, Dec 1969.

D. Fayman, "A Simple Computing SWR Meter," *QST*, Jul 1973, pp 23-33.

J. Gibbons and H. Horn, "A Circuit With Logarithmic Response Over Nine Decades," *IEEE Transactions on Circuit Theory*, Vol CT-11, No. 3, Sep 1964, pp 378-384.

J. Grebenkemper, "Calibrating Diode Detectors," *QEX*, Aug 1990.

J. Grebenkemper, "The Tandem Match—An Accurate Directional Wattmeter," *QST*, Jan 1987.

J. Grebenkemper, "Improving and Using R-X Noise Bridges," *QST*, Aug 1989; Feedback, *QST*, Jan 1990, p 27.

G. H. Hagn, SRI, "HF Ground Measurements at the Lawrence Livermore National Laboratory (LLNL) Field Site," *Applied Computational Electromagnetics Society Journal and Newsletter*, Vol 3, Number 2, Fall 1988.

T. King, "A Practical Time-Domain Reflectometer," *QST*, May 1989, pp 22-24.

Z. Lau and C. Hutchinson, "Improving the HW-9 Transceiver," *QST*, Apr 1988, pp 26-29.

V. G. Leenerts, "Automatic VSWR and Power Meter," *Ham Radio*, May 1980, pp 34-43.

J. Lenk, *Handbook of Oscilloscopes* (Englewood Cliffs, NJ: Prentice-Hall, 1982), pp 288-292.

I. Lindell, E. Alanen, K. Mannerslo, "Exact Image Method for Impedance Computation of Antennas Above the Ground," *IEEE Trans. On Antennas and Propagation*, AP-33, Sep 1985.

L. McCoy, "A Linear Field-Strength Meter," *QST*, Jan 1973.

T. McMullen, "The Line Sampler, an RF Power Monitor for VHF and UHF," *QST*, Apr 1972.

M. Walter Maxwell, W2DU, *Reflections* (Newington: ARRL, 1990), p 20-3. [Out of print.]

C. Michaels, "Determining Line Lengths," Technical Correspondence, *QST*, Sep 1985, pp 43-44.

H. Perras, "Broadband Power-Tracking VSWR Bridge," *Ham Radio*, Aug 1979, pp 72-75.

S. Ramo, J. Whinnery and T. Van Duzer, *Fields and Waves in Communication Electronics* (New York: John Wiley & Sons, 1967), Chap 1.

Reference Data for Radio Engineers, 5th edition (Indianapolis: Howard W. Sams, 1968), Chapter 28.

P. N. Saveskie, *Radio Propagation Handbook* (Blue Ridge Summit, PA: TAB Books, 1960).

J. Sevick, "Short Ground-Radial Systems for Short Verticals," *QST*, Apr 1978, pp 30-33.

J. Sevick, "Measuring Soil Conductivity," *QST*, Mar 1981, pp 38- 39.

W. Spaulding, "A Broadband Two-Port S-Parameter Test Set," *Hewlett-Packard Journal*, Nov 1984.

D. Turrin, "Antenna Performance Measurements," *QST*, Nov 1974, pp 35-41.

F. Van Zant, "High-Power Operation with the Tandem Match Directional Coupler," Technical Correspondence, *QST*, Jul 1989, pp 42-43.

R. F. H. Yang, "A Proposed Gain Standard for VHF Antennas," *IEEE Transactions on Antennas and Propagation*, Nov 1966.

Chapter 28

Smith Chart Calculations

The Smith Chart is a sophisticated graphic tool for solving transmission line problems. One of the simpler applications is to determine the feed-point impedance of an antenna, based on an impedance measurement at the input of a random length of transmission line. By using the Smith Chart, the impedance measurement can be made with the antenna in place atop a tower or mast, and there is no need to cut the line to an exact multiple of half wavelengths. The Smith Chart may be used for other purposes, too, such as the design of impedance-matching networks. These matching networks can take on any of several forms, such as L and pi networks, a stub matching system, a series-section match, and more. With a knowledge of the Smith Chart, the amateur can eliminate much “cut and try” work.

Named after its inventor, Phillip H. Smith, the Smith Chart was originally described in *Electronics* for January 1939. Smith Charts may be obtained at most university book stores. Smith Charts are also available from ARRL HQ. (See the caption for Fig 3.)

It is stated in Chapter 24 that the input impedance, or the impedance seen when “looking into” a length of line, is dependent upon the SWR, the length of the line, and the Z_0 of the line. The SWR, in turn, is dependent upon the load which terminates the line. There are complex mathematical relationships which may be used to calculate the various values of impedances, voltages, currents, and SWR values that exist in the operation of a particular transmission line. These equations can be solved with a personal computer and suitable software, or the parameters may be determined with the Smith Chart. Even if a computer is used, a fundamental knowledge of the Smith Chart will promote a better understanding of the problem being solved. And such an understanding might lead to a quicker or simpler solution than otherwise. If the terminating impedance is known, it is a simple matter to determine the input impedance of the line for any length by means of the chart. Conversely, as indicated above, with a given line length and a known (or measured) input impedance, the load impedance may be determined by means of the chart—a convenient method of remotely determining an antenna impedance, for example.

Although its appearance may at first seem somewhat formidable, the Smith Chart is really nothing more than a specialized type of graph. Consider it as having curved, rather than rectangular, coordinate lines. The coordinate system consists simply of two families of circles—the resistance family, and the reactance family. The resistance circles, **Fig 1**, are centered on the resistance axis (the only straight line on the chart), and are tangent to the outer circle at the right of the chart. Each circle is assigned a value of resistance, which is indicated at the point where the circle crosses the resistance axis. All points along any one circle have the same resistance value.

The values assigned to these circles vary from zero at the left of the chart to infinity at the right, and actually represent a *ratio* with respect to the impedance value assigned to the center point of the chart, indicated 1.0. This center point is called prime center. If prime center is assigned a value of 100 Ω, then 200 Ω resistance is represented by the 2.0 circle, 50 Ω by the 0.5 circle, 20 Ω by the 0.2 circle, and so on. If, instead, a value of 50 is assigned to prime center,

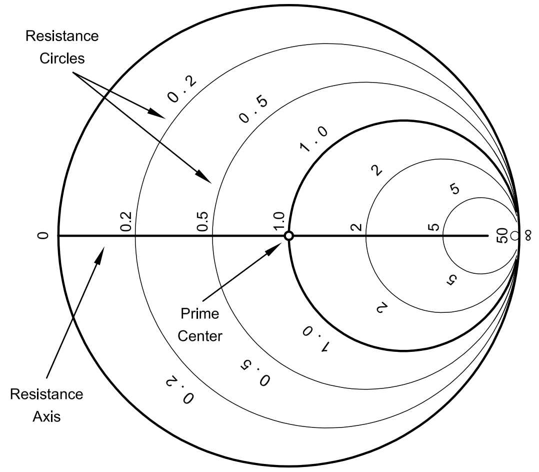

Fig 1—Resistance circles of the Smith Chart coordinate system.

the 2.0 circle now represents 100 Ω, the 0.5 circle 25 Ω, and the 0.2 circle 10 Ω. In each case, it may be seen that the value on the chart is determined by dividing the actual resistance by the number assigned to prime center. This process is called normalizing.

Conversely, values from the chart are converted back to actual resistance values by multiplying the chart value times the value assigned to prime center. This feature permits the use of the Smith Chart for any impedance values, and therefore with any type of uniform transmission line, whatever its impedance may be. As mentioned above, specialized versions of the Smith Chart may be obtained with a value of 50 Ω at prime center. These are intended for use with 50-Ω lines.

Now consider the reactance circles, **Fig 2**, which appear as curved lines on the chart because only segments of the complete circles are drawn. These circles are tangent to the resistance axis, which itself is a member of the reactance family (with a radius of infinity). The centers are displaced to the top or bottom on a line tangent to the right of the chart. The large outer circle bounding the coordinate portion of the chart is the reactance axis.

Each reactance circle segment is assigned a value of reactance, indicated near the point where the circle touches the reactance axis. All points along any one segment have the same reactance value. As with the resistance circles, the values assigned to each reactance circle are normalized with respect to the value assigned to prime center. Values to the top of the resistance axis are positive (inductive), and those to the bottom of the resistance axis are negative (capacitive).

When the resistance family and the reactance family of circles are combined, the coordinate system of the Smith Chart results, as shown in **Fig 3**. Complex impedances (R + *j*X) can be plotted on this coordinate system.

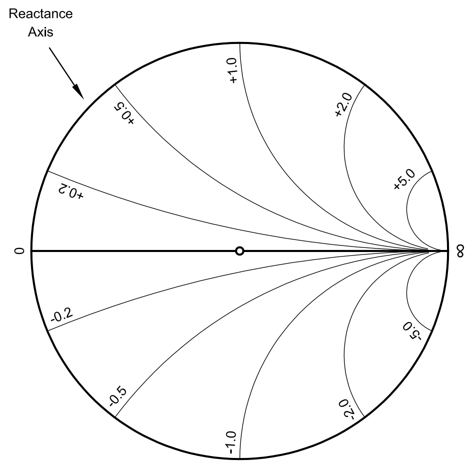

Fig 2—Reactance circles (segments) of the Smith Chart coordinate system.

IMPEDANCE PLOTTING

Suppose we have an impedance consisting of 50 Ω resistance and 100 Ω inductive reactance (Z = 50 + *j* 100). If we assign a value of 100 Ω to prime center, we normalize the above impedance by dividing each component of the impedance by 100. The normalized impedance is then 50/100 + *j* (100/100) = 0.5 + *j* 1.0. This impedance is plotted on the Smith Chart at the intersection of the 0.5 resistance circle and the +1.0 reactance circle, as indicated in Fig 3. Calculations may now be made from this plotted value.

Now say that instead of assigning 100 Ω to prime center, we assign a value of 50 Ω. With this assignment, the 50 + *j* 100 Ω impedance is plotted at the intersection of the 50/50 = 1.0 resistance circle, and the 100/50 = 2.0 positive reactance circle. This value, 1 + *j* 2, is also indicated in Fig 3. But now we have *two* points plotted in Fig 3 to represent the same impedance value, 50 + *j* 100 Ω. How can this be?

These examples show that the same impedance may be plotted at different points on the chart, depending upon the value assigned to prime center. But two plotted points cannot represent the same impedance *at the same time*! It is customary when solving transmission-line problems to assign to prime center a value equal to the characteristic impedance, or Z_0, of the line being used. This value should always be recorded at the start of calculations, to avoid possible confusion later. (In using the specialized charts with the value of 50 at prime center, it is, of course, not necessary to normalize impedances when working with 50-Ω line. The resistance and reactance values may be read directly from the chart coordinate system.)

Prime center is a point of special significance. As

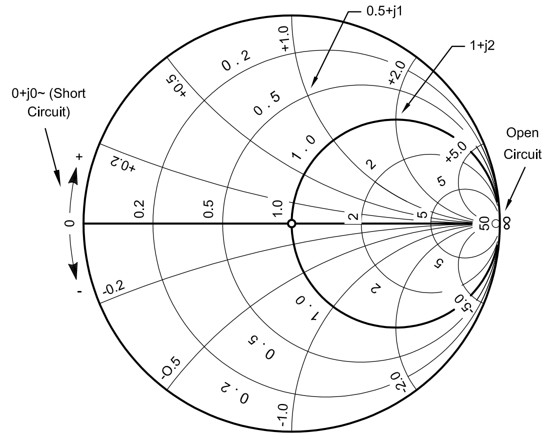

Fig 3—The complete coordinate system of the Smith Chart. For simplicity, only a few divisions are shown for the resistance and reactance values. Various types of Smith Chart forms are available from ARRL HQ. At the time of this writing, five 8½ × 11 inch Smith Chart forms are available for $2.

just mentioned, is is customary when solving problems to assign the Z_0 value of the line to this point on the chart—50 Ω for a 50-Ω line, for example. What this means is that the center point of the chart now represents 50 +*j* 0 ohms–a pure resistance equal to the characteristic impedance of the line. If this were a load on the line, we recognize from transmission-line theory that it represents a perfect match, with no reflected power and with a 1.0 to 1 SWR. Thus, prime center also represents the 1.0 SWR circle (with a radius of zero). SWR circles are also discussed in a later section.

Short and Open Circuits

On the subject of plotting impedances, two special cases deserve consideration. These are short circuits and open circuits. A true short circuit has zero resistance and zero reactance, or 0 + *j* 0). This impedance is plotted at the left of the chart, at the intersection of the resistance and the reactance axes. By contrast, an open circuit has infinite resistance, and therefore is plotted at the right of the chart, at the intersection of the resistance and reactance axes. These two special cases are sometimes used in matching stubs, described later.

Standing-Wave-Ratio Circles

Members of a third family of circles, which are not printed on the chart but which are added during the process of solving problems, are standing-wave-ratio or SWR circles. See **Fig 4**. This family is centered on prime center, and appears as concentric circles inside the reactance axis. During calculations, one or more of these circles may be added with a drawing compass. Each circle represents a value of SWR, with every point on a given circle representing the same SWR. The SWR value for a given circle may be determined directly from the chart coordinate system, by reading the resistance value where the SWR circle crosses the resistance axis to the right of prime center. (The reading where the circle crosses the resistance axis to the left of prime center indicates the inverse ratio.)

Consider the situation where a load mismatch in a length of line causes a 3-to-1 SWR ratio to exist. If we temporarily disregard line losses, we may state that the SWR remains constant throughout the entire length of this line. This is represented on the Smith Chart by drawing a 3:1 constant SWR circle (a circle with a radius of 3 on the resistance axis), as in **Fig 5**. The design of the chart is such that any impedance encountered *anywhere* along the length of this mismatched line will fall on the SWR circle. The impedances may be read from the coordinate system merely by the progressing around the SWR circle by an amount corresponding to the length of the line involved.

This brings into use the wavelength scales, which appear in Fig 5 near the perimeter of the Smith Chart. These scales are calibrated in terms of portions of an electrical wavelength along a transmission line. Both scales start from 0 at the left of the chart. One scale, running

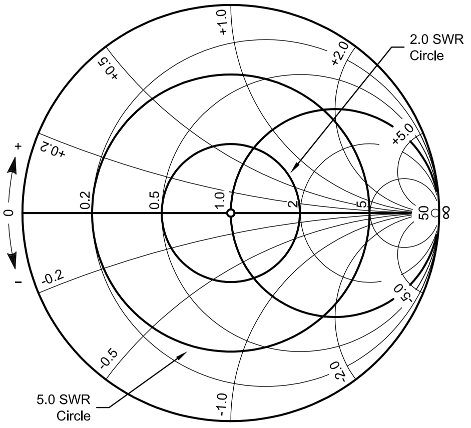

Fig 4—Smith Chart with SWR circles added.

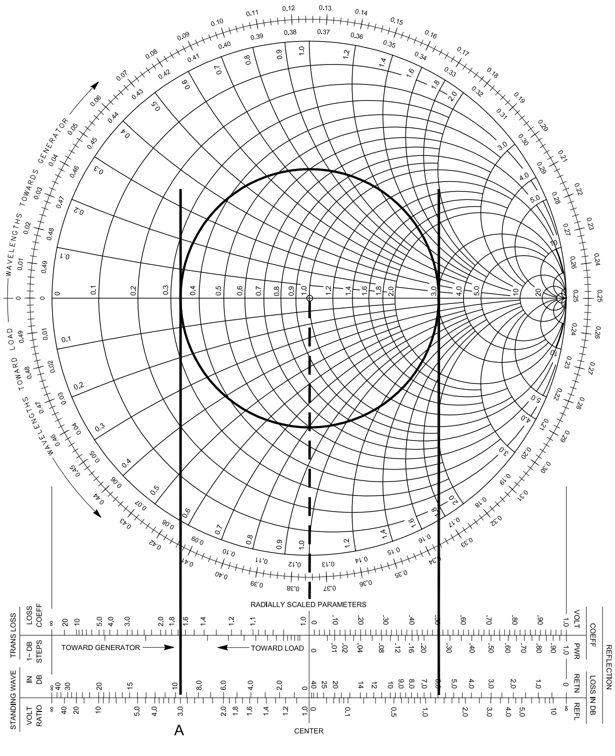

Fig 5—Example discussed in text.

counterclockwise, starts at the generator or input end of the line and progresses toward the load. The other scale starts at the load and proceeds toward the generator in a clockwise direction. The complete circle around the edge of the chart represents $^{1}/_{2}$ λ. Progressing once around the perimeter of these scales corresponds to progressing along a transmission line for $^{1}/_{2}$ λ. Because impedances repeat themselves every $^{1}/_{2}$ λ along a piece of line, the chart may be used for any length of line by disregarding or subtracting from the line's total length an integral, or whole number, of half wavelengths.

Also shown in Fig 5 is a means of transferring the radius of the SWR circle to the external scales of the chart, by drawing lines tangent to the circle. Another simple way to obtain information from these external scales is to transfer the radius of the SWR circle to the external scale with a drawing compass. Place the point of a drawing compass at the center or 0 line, and inscribe a short arc across the appropriate scale. It will be noted that when this is done in Fig 5, the external STANDING-WAVE VOLTAGE-RATIO scale indicates the SWR to be 3.0 (at A)—our condition for initially drawing the circle on the chart (and the same as the SWR reading on the resistance axis).

SOLVING PROBLEMS WITH THE SMITH CHART

Suppose we have a transmission line with a characteristic impedance of 50 Ω and an electrical length of 0.3 λ. Also, suppose we terminate this line with an impedance having a resistive component of 25 Ω and an inductive reactance of 25 Ω (Z = 25 + *j* 25). What is the input impedance to the line?

The characteristic impedance of the line is 50 Ω, so we begin by assigning this value to prime center. Because the line is not terminated in its characteristic impedance, we know that standing waves will exist on the line, and that, therefore, the input impedance to the line will not be exactly 50 Ω. We proceed as follows. First, normalize the load impedance by dividing both the resistive and reactive components by 50 (Z_0 of the line being used). The normalized impedance in this case is 0.5 + *j* 0.5. This is plotted on the chart at the intersection of the 0.5 resistance and the +0.5 reactance circles, as in **Fig 6**. Then draw a constant SWR circle passing through this point. Transfer the radius of this circle to the external scales with the drawing compass. From the external STANDING-WAVE VOLTAGE-RATIO scale, it may be seen (at A) that the voltage ratio of 2.62 exists for this radius, indicating that our line is operating with an SWR of 2.62 to 1. This figure is converted to decibels in the adjacent scale, where 8.4 dB may be read (at B), indicating that the ratio of the voltage maximum to the voltage minimum along the line is 8.4 dB. (This is mathematically equivalent to 20 times the log of the SWR value.)

Next, with a straightedge, draw a radial line from prime center through the plotted point to intersect the wavelengths scale. At this intersection, point C in Fig 6, read a value from the wavelengths scale. Because we are starting from the load, we use the TOWARD GENERATOR or outermost calibration, and read 0.088 λ.

To obtain the line input impedance, we merely find the point on the SWR circle that is 0.3 λ toward the generator from the plotted load impedance. This is accomplished by adding 0.3 (the length of the line in wavelengths) to the reference or starting point, 0.088; 0.3 + 0.088 = 0.388. Locate 0.388 on the TOWARD GENERATOR scale (at D). Draw a second radial line from this point to prime center. The intersection of the new radial line with the SWR circle represents the normalized line input impedance, in this case 0.6 – *j* 0.66.

To find the unnormalized line impedance, multiply by 50, the value assigned to prime center. The resulting value is 30 – *j* 33, or 30 Ω resistance and 33 Ω capacitive reactance. *This* is the impedance that a transmitter must match if such a system were a combination of antenna

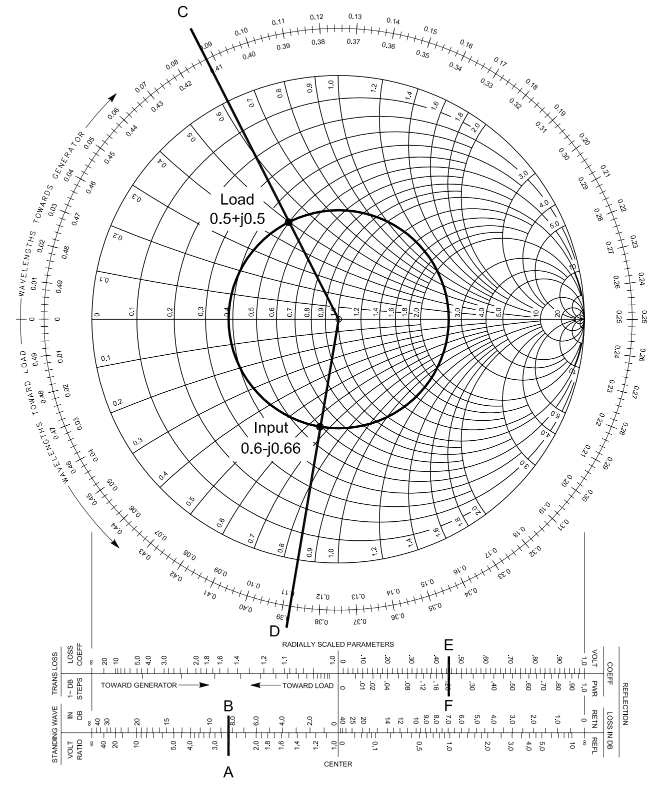

Fig 6—Example discussed in text.

and transmission line. This is also the impedance that would be measured on an impedance bridge if the measurement were taken at the line input.

In addition to the line input impedance and the SWR, the chart reveals several other operating characteristics of the above system of line and load, if a closer look is desired. For example, the voltage reflection coefficient, both magnitude and phase angle, for this particular load is given. The phase angle is read under the radial line drawn through the plot of the load impedance, where the line intersects the ANGLE OF REFLECTION COEFFICIENT scale. This scale is not included in Fig 6, but will be found on the Smith Chart just inside the wavelengths scales. In this example, the reading is 116.6 degrees. This indicates the angle by which the reflected voltage wave leads the incident wave at the load. It will be noted that angles on the bottom half, or capacitive-reactance half, of the chart are negative angles, a "negative" lead indicating that the reflected voltage wave actually lags the incident wave.

The magnitude of the voltage-reflection-coefficient may be read from the external REFLECTION COEFFICIENT VOLTAGE scale, and is seen to be approximately 0.45 (at E) for this example. This means that 45 percent of the incident voltage is reflected. Adjacent to this scale on the POWER calibration, it is noted (at F) that the power reflection coefficient is 0.20, indicating that 20 percent of the incident power is reflected. (The amount of reflected power is proportional to the square of the reflected voltage.)

ADMITTANCE COORDINATES

Quite often it is desirable to convert impedance information to admittance data—conductance and susceptance. Working with admittances greatly simplifies determining the resultant when two complex impedances are connected in parallel, as in stub matching. The conductance values may be added directly, as may be the susceptance values, to arrive at the overall admittance for the parallel combination. This admittance may then be converted back to impedance data, if desired.

On the Smith Chart, the necessary conversion may be made very simply. The equivalent admittance of a plotted impedance value lies diametrically opposite the impedance point on the chart. In other words, an impedance plot and its corresponding admittance plot will lie on a straight line that passes through prime center, and each point will be the same distance from prime center (on the same SWR circle). In the above example, where the normalized line input impedance is 0.6 – *j* 0.66, the equivalent admittance lies at the intersection of the SWR circle and the extension of the straight line passing from point D though prime center. Although not shown in Fig 6, the normalized admittance value may be read as 0.76 + *j* 0.84 if the line starting at D is extended.

In making impedance-admittance conversions, remember that capacitance is considered to be a positive susceptance and inductance a negative susceptance. This corresponds to the scale identification printed on the chart. The admittance in siemens is determined by dividing the normalized values by the Z_0 of the line. For this example the admittance is 0.76/50 + *j* 0.84/50 = 0.0152 + *j* 0.0168 siemen. Of course admittance coordinates may be converted to impedance coordinates just as easily—by locating the point on the Smith Chart that is diametrically opposite that representing the admittance coordinates, on the same SWR circle.

DETERMINING ANTENNA IMPEDANCES

To determine an antenna impedance from the Smith Chart, the procedure is similar to the previous example. The electrical length of the feed line must be known and the impedance value at the input end of the line must be determined through measurement, such as with an impedance-measuring or a good quality noise bridge. In this case, the antenna is connected to the far end of the line and becomes the load for the line. Whether the

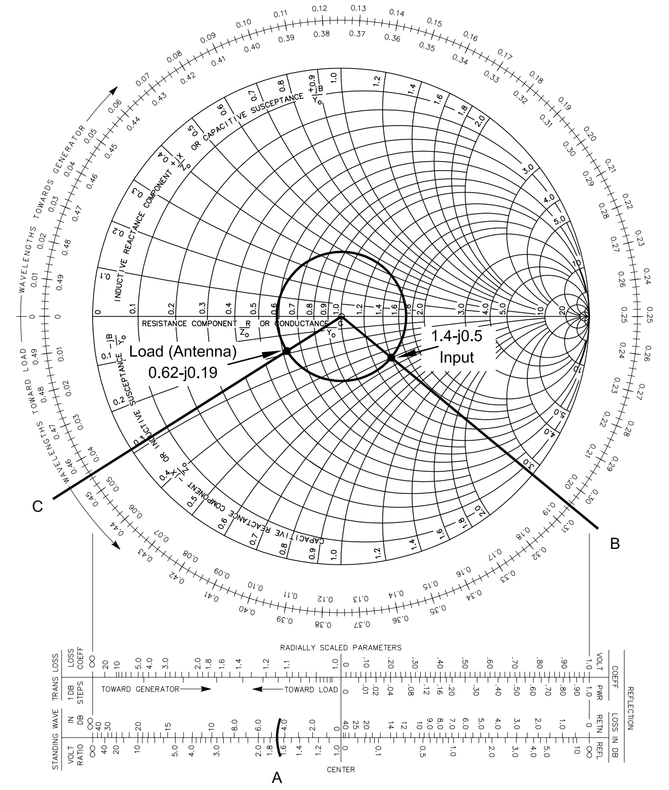

Fig 7—Example discussed in text.

antenna is intended purely for transmission of energy, or purely for reception makes no difference; the antenna is still the terminating or load impedance on the line as far as these measurements are concerned. The input or generator end of the line is that end connected to the device for measurement of the impedance. In this type of problem, the measured impedance is plotted on the chart, and the TOWARD LOAD wavelengths scale is used in conjunction with the electrical line length to determine the actual antenna impedance.

For example, assume we have a measured input impedance to a 50-Ω line of 70 – *j* 25 Ω. The line is 2.35 λ long, and is terminated in an antenna. What is the antenna feed impedance? Normalize the input impedance with respect to 50 Ω, which comes out 1.4 – *j* 0.5, and plot this value on the chart. See **Fig 7**. Draw a constant SWR circle through the point, and transfer the radius to the external scales. The SWR of 1.7 may be read from the VOLTAGE RATIO scale (at A). Now draw a radial line from prime center through this plotted point to the wavelengths scale, and read a reference value (at B). For this case the value is 0.195, on the TOWARD LOAD scale. Remember, we are starting at the generator end of the transmission line.

To locate the load impedance on the SWR circle, add the line length, 2.35 λ, to the reference value from the wavelengths scale; 2.35 + 0.195 = 2.545. Locate the new value on the TOWARD LOAD scale. But because the calibrations extend only from 0 to 0.5, we must first subtract a number of half wavelengths from this value and use only the remaining value. In this situation, the largest integral number of half wavelengths that can be subtracted with a positive result is 5, or 2.5 λ. Thus, 2.545 – 2.5 = 0.045. Locate the 0.045 value on the TOWARD LOAD scale (at C). Draw a radial line from this value to prime center. Now, the coordinates at the intersection of the second radial line and the SWR circle represent the load impedance. To read this value closely, some interpolation between the printed coordinate lines must be made, and the value of 0.62 – *j* 0.19 is read. Multiplying by 50, we get the actual load or antenna impedance as 31 – *j* 9.5 Ω, or 31 Ω resistance with 9.5 Ω capacitive reactance.

Problems may be entered on the chart in yet another manner. Suppose we have a length of 50-Ω line feeding a base-loaded resonant vertical ground-plane antenna which is shorter than $^1/_4$ λ. Further, suppose we have an SWR monitor in the line, and that it indicates an SWR of 1.7 to 1. The line is known to be 0.95 λ long. We want to know both the input and the antenna impedances.

From the information available, we have no impedances to enter into the chart. We may, however, draw a circle representing the 1.7 SWR. We also know, from the definition of resonance, that the antenna presents a purely resistive load to the line, that is, no reactive component. Thus, the antenna impedance must lie on the resistance axis. If we were to draw such an SWR circle and observe the chart with only the circle drawn, we would see two points which satisfy the resonance requirement for the load. These points are 0.59 + *j* 0 and 1.7 + *j* 0. Multiplying by 50, we see that these values represent 29.5 and 85 Ω resistance. This may sound familiar, because, as was discussed in Chapter 24, when a line is terminated in a pure resistance, the SWR in the line equals Z_R/Z_0 or Z_0/Z_R, where Z_R=load resistance and Z_0=line impedance.

If we consider antenna fundamentals described in Chapter 2, we know that the theoretical impedance of a $^1/_4$-λ ground-plane antenna is approximately 36 Ω. We therefore can quite logically discard the 85-Ω impedance figure in favor of the 29.5-Ω value. This is then taken as the load impedance value for the Smith Chart calculations. To find the line input impedance, we subtract 0.5 λ from the line length, 0.95, and find 0.45 λ on the TOWARD GENERATOR scale. (The wavelength-scale starting point in this case is 0.) The line input impedance is found to be 0.63 – *j* 0.20, or 31.5 – *j* 10 Ω.

DETERMINATION OF LINE LENGTH

In the example problems given so far in this chapter, the line length has conveniently been stated in wavelengths. The electrical length of a piece of line depends upon its physical length, the radio frequency under consideration, and the velocity of propagation in the line. If an impedance-measurement bridge is capable of quite reliable readings at high SWR values, the line length may be determined through line input-impedance measurements with short- or open-circuit line terminations. Information on the procedure is given later in this chapter. A more direct method is to measure the physical length of the line and calculate its electrical length from

$$N = \frac{Lf}{984VF} \qquad \text{(Eq 1)}$$

where

N = number of electrical wavelengths in the line
L = line length in feet
f = frequency, MHz
VF = velocity or propagation factor of the line

The velocity factor may be obtained from transmission-line data tables in Chapter 24.

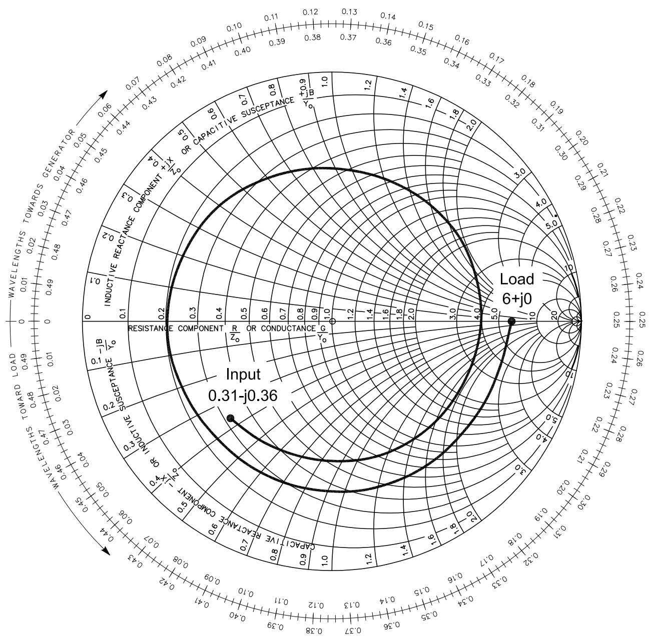

Fig 8—This spiral is the actual "SWR circle" when line losses are taken into account. It is based on calculations for a 16-ft length of RG-174 coax feeding a resonant 28-MHz 300-Ω antenna (50-Ω coax, velocity factor = 66%, attenuation = 6.2 dB per 100 ft). The SWR at the load is 6:1, while it is 3.6:1 at the line input. When solving problems involving attenuation, two constant SWR circles are drawn instead of a spiral, one for the line input SWR and one for the load SWR.

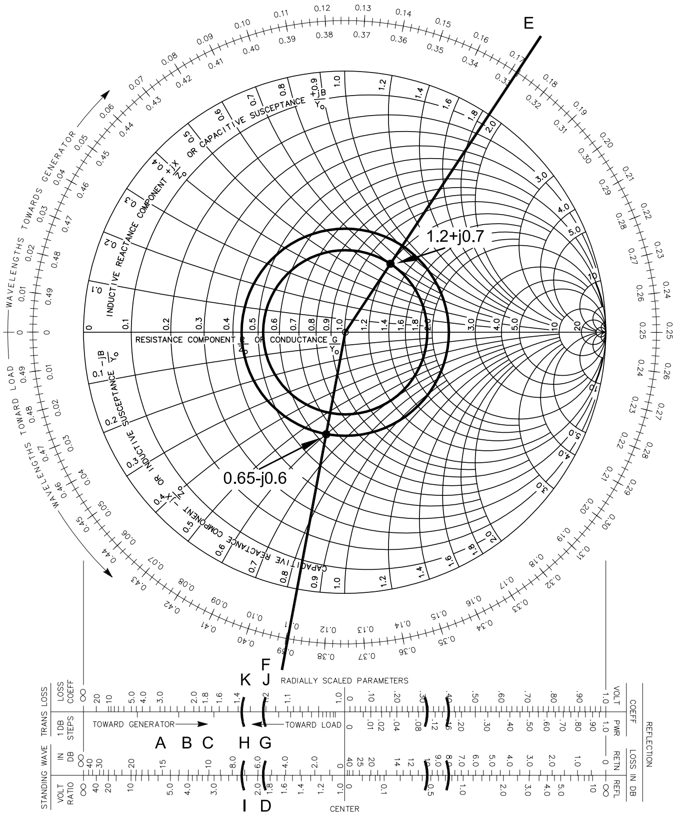

Fig 9—Example of Smith Chart calculations taking line losses into account.

LINE-LOSS CONSIDERATIONS WITH THE SMITH CHART

The example Smith Chart problems presented in the previous section ignored attenuation, or line losses. Quite frequently it is not even necessary to consider losses when making calculations; any difference in readings obtained are often imperceptible on the chart. However, when the line losses become appreciable, such as for high-loss lines, long lines, or at VHF and UHF, loss considerations may become significant in making Smith Chart calculations. This involves only one simple step, in addition to the procedures previously presented.

Because of line losses, as discussed in Chapter 24 the SWR does not remain constant throughout the length of the line. As a result, there is a decrease in SWR as one progresses away from the load. To truly present this situation on the Smith Chart, instead of drawing a constant SWR circle, it would be necessary to draw a spiral inward and clockwise from the load impedance toward the generator, as shown in **Fig 8**. The rate at which the curve spirals toward prime center is related to the attenuation in the line. Rather than drawing spiral curves, a simpler method is used in solving line-loss problems, by means of the external scale TRANSMISSION LOSS 1-DB STEPS. This scale may be seen in **Fig 9**. Because this is only a relative scale, the decibel steps are not numbered.

If we start at the left end of this external scale and proceed in the direction indicated TOWARD GENERATOR, the first dB step is seen to occur at a radius from center corresponding to an SWR of about 9 (at A); the second dB step falls at an SWR of about 4.5 (at B), the third at 3.0 (at C), and so forth, until the

15th dB step falls at an SWR of about 1.05 to 1. This means that a line terminated in a short or open circuit (infinite SWR), and having an attenuation of 15 dB, would exhibit an SWR of only 1.05 at its input. It will be noted that the dB steps near the right end of the scale are very close together, and a line attenuation of 1 or 2 dB in this area will have only slight effect on the SWR. But near the left end of the scale, corresponding to high SWR values, a 1 or 2 dB loss has considerable effect on the SWR.

Using a Second SWR Circle

In solving a problem using line-loss information, it is necessary only to modify the radius of the SWR circle by an amount indicated on the TRANSMISSION-LOSS 1-DB STEPS scale. This is accomplished by drawing a second SWR circle, either smaller or larger than the first, depending on whether you are working toward the load or toward the generator.

For example, assume that we have a 50-Ω line that is 0.282 λ long, with 1-dB inherent attenuation. The line input impedance is measured as 60 + *j* 35 Ω. We desire to know the SWR at the input and at the load, and the load impedance. As before, we normalize the 60 + *j* 35-Ω impedance, plot it on the chart, and draw a constant SWR circle and a radial line through the point. In this case, the normalized impedance is 1.2 + *j* 0.7. From Fig 9, the SWR at the line input is seen to be 1.9 (at D), and the radial line is seen to cross the TOWARD LOAD scale, first subtract 0.500, and locate 0.110 (at F); then draw a radial line from this point to prime center.

To account for line losses, transfer the radius of the SWR circle to the external 1-DB STEPS scale. This radius crosses the external scale at G, the fifth decibel mark from the left. Since the line loss was given as 1 dB, we strike a new radius (at H), one "tick mark" to the left (toward load) on the same scale. (This will be the fourth decibel tick mark from the left of the scale.) Now transfer this new radius back to the main chart, and scribe a new SWR circle of this radius. This new radius represents the SWR at the load, and is read as 2.3 on the external VOLTAGE RATIO scale. At the intersection of the new circle and the load radial line, we read 0.65 – *j* 0.6. This is the normalized load impedance. Multiplying by 50, we obtain the actual load impedance as 32.5 – *j* 30 Ω. The SWR in this problem was seen to increase from 1.9 at the line input to 2.3 (at I) at the load, with the 1-dB line loss taken into consideration.

In the example above, values were chosen to fall conveniently on or very near the "tick marks" on the 1-dB scale. Actually, it is a simple matter to interpolate between these marks when making a radius correction. When this is necessary, the relative distance between marks for each decibel step should be maintained while counting off the proper number of steps.

Adjacent to the 1-DB STEPS scale lies a LOSS COEFFICIENT scale. This scale provides a factor by which the matched-line loss in decibels should be multiplied to account for the increased losses in the line when standing waves are present. These added losses do not affect the SWR or impedance calculations; they are merely the additional dielectric copper losses of the line caused by the fact that the line conducts more average voltage in the presence of standing waves. For the above example, from Fig 9, the loss coefficient at the input end is seen to be 1.21 (at J), and 1.39 (at K) at the load. As a good approximation, the loss coefficient may be averaged over the length of line under consideration; in this case, the average is 1.3. This means that the total losses in the line are 1.3 times the matched loss of the line (1 dB), or 1.3 dB. This is the same result that may be obtained from procedures given in Chapter 24 for this data.

Smith Chart Procedure Summary

To summarize briefly, any calculations made on the Smith Chart are performed in four basic steps, although not necessarily in the order listed.

1) Normalize and plot a line input (or load) impedance, and construct a constant SWR circle.
2) Apply the line length to the wavelengths scales.
3) Determine attenuation or loss, if required, by means of a second SWR circle.
4) Read normalized load (or input) impedance, and convert to impedance in ohms.

The Smith Chart may be used for many types of problems other than those presented as examples here. The transformer action of a length of line—to transform a high impedance (with perhaps high reactance) to a purely resistive impedance of low value—was not mentioned. This is known as "tuning the line," for which the chart is very helpful, eliminating the need for "cut and try" procedures. The chart may also be used to calculate lengths for shorted or open matching stubs in a system, described later in this chapter. In fact, in any application where a transmission line is not perfectly matched, the Smith Chart can be of value.

ATTENUATION AND Z_0 FROM IMPEDANCE MEASUREMENTS

If an impedance bridge is available to make accurate measurements in the presence of very high SWR values, the attenuation, characteristic impedance and velocity factor of any random length of coaxial transmission line can be determined. This section was written by Jerry Hall, K1TD.

Homemade impedance bridges and noise bridges will seldom offer the degree of accuracy required to use this technique, but sometimes laboratory bridges can be found as industrial surplus at a reasonable price. It may also be possible for an amateur to borrow a laboratory type of bridge for the purpose of making some weekend measurements. Making these determinations is not difficult, but the procedure is not commonly known among amateurs. One equation treating complex numbers is used, but the math can be handled with a calculator supporting trig functions. Full details are given in the paragraphs that follow.

For each frequency of interest, two measurements are required to determine the line impedance. Just one measurement is used to determine the line attenuation and velocity factor. As an example, assume we have a 100-foot length of unidentified line with foamed dielectric, and wish to know its characteristics. We make our measurements at 7.15 MHz. The procedure is as follows.

1) Terminate the line in an open circuit. The best "open circuit" is one that minimizes the capacitance between the center conductor and the shield. If the cable has a PL-259 connector, unscrew the shell and slide it back down the coax for a few inches. If the jacket and insulation have been removed from the end, fold the braid back along the outside of the line, away from the center conductor.
2) Measure and record the impedance at the input end of the line. If the bridge measures admittance, convert the measured values to resistance and reactance. Label the values as $R_{oc} + j\,X_{oc}$. For our example, assume we measure 85 + *j*179 Ω. (If the reactance term is capacitive, record it as negative.)
3) Now terminate the line in a short circuit. If a connector exists at the far end of the line, a simple short is a mating connector with a very short piece of heavy wire soldered between the center pin and the body. If the coax has no connector, removing the jacket and center insulation from a half inch or so at the end will allow you to tightly twist the braid around the center conductor. A small clamp or alligator clip around the outer braid at the twist will keep it tight.
4) Again measure and record the impedance at the input end of the line. This time label the values as $R_{sc} \pm j\,X$. Assume the measured value now is $4.8 - j\,11.2\ \Omega$.

This completes the measurements. Now we reach for the calculator.

As amateurs we normally assume that the characteristic impedance of a line is purely resistive, but it can (and does) have a small capacitive reactance component. Thus, the Z_0 of a line actually consists of $R_0 + j\,X_0$. The basic equation for calculating the characteristic impedance is

$$Z_0 = \sqrt{Z_{oc} \times Z_{sc}} \qquad \text{(Eq 2)}$$

where

$Z_{oc} = R_{oc} + jX_{oc}$

$Z_{sc} = R_{sc} + jX_{sc}$

From Eq 2 the following working equation may be derived.

$$Z_0 = \sqrt{\left(R_{oc}R_{sc} - X_{oc}X_{sc}\right) + j\left(R_{oc}X_{sc} + R_{sc}X_{oc}\right)} \qquad \text{(Eq 3)}$$

The expression under the radical sign in Eq 3 is in the form of R + *j* X. By substituting the values from our example into Eq 3, the R term becomes 85 × 4.8 – 179 × (–11.2) = 2412.8, and the X term becomes 85 × (–11.2) + 4.8 × 179 = –92.8. So far, we have determined that

$$Z_0 = \sqrt{2412.8 - j\,92.8}$$

The quantity under the radical sign is in rectangular form. Extracting the square root of a complex term is handled easily if it is in polar form, a vector value and its angle. The vector value is simply the square root of the sum of the squares, which in this case is

$$\sqrt{2412.8^2 + 92.8^2} = \sqrt{2414.58}$$

The tangent of the vector angle we are seeking is the value of the reactance term divided by the value of the resistance term. For our example this is arctan –92.8/2412.8 = arctan –0.03846. The angle is thus found to be –2.20°. From all of this we have determined that

$$Z_0 = \sqrt{2414.58\,\angle{-2.20^\circ}}$$

Extracting the square root is now simply a matter of finding

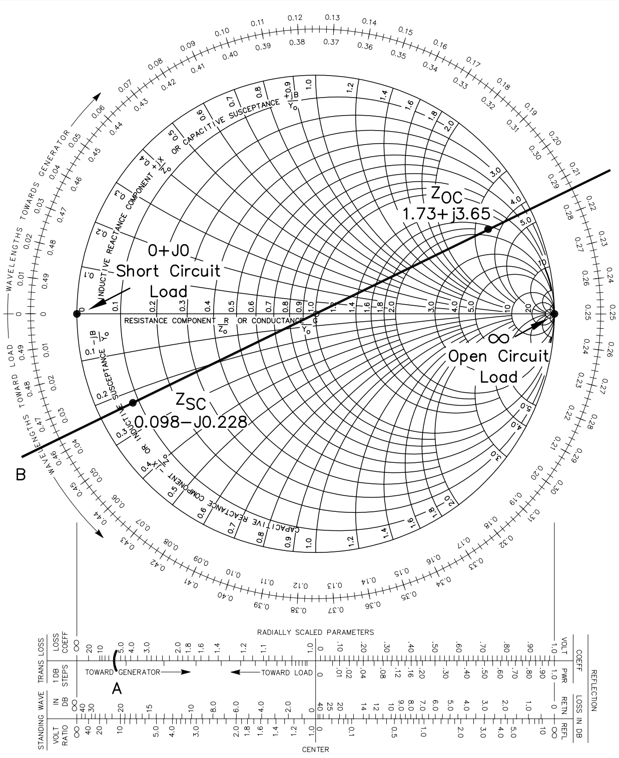

Fig 10—Determining the line loss and velocity factor with the Smith Chart from input measurements taken with open-circuit and short-circuit terminations.

the square root of the vector value, and taking half the angle. (The angle is treated mathematically as an exponent.)

Our result for this example is $Z_0 = 49.1\underline{/-1.1°}$. The small negative angle may be ignored, and we now know that we have coax with a nominal 50-Ω impedance. (Departures of as much as 6 to 8% from the nominal value are not uncommon.) If the negative angle is large, or if the angle is positive, you should recheck your calculations and perhaps even recheck the original measurements. You can get an idea of the validity of the measurements by normalizing the measured values to the calculated impedance and plotting them on a Smith Chart as shown in **Fig 10** for this example. Ideally, the two points should be diametrically opposite, but in practice they will be not quite 180° apart and not quite the same distance from prime center. Careful measurements will yield plotted points that are close to ideal. Significant departures from the ideal indicates sloppy measurements, or perhaps an impedance bridge that is not up to the task.

Determining Line Attenuation

The short circuit measurement may be used to determine the line attenuation. This reading is more reliable than the open circuit measurement because a good short circuit is a short, while a good open circuit is hard to find. (It is impossible to escape some amount of capacitance between conductors with an "open" circuit, and that capacitance presents a path for current to flow at the RF measurement frequency.)

Use the Smith Chart and the 1-DB STEPS external scale to find line attenuation. First normalize the short circuit impedance reading to the calculated Z_0, and plot this point on the chart. See Fig 10. For our example, the normalized impedance is 4.8/49.1 – *j* 11.2 / 49.1 or 0.098 – *j* 0.228. After plotting the point, transfer the radius to the 1-DB STEPS scale. This is shown at A of Fig 10.

Remember from discussions earlier in this chapter that the impedance for plotting a short circuit is 0 + *j* 0, at the left edge of the chart on the resistance axis. On the 1-DB STEPS scale this is also at the left edge. The total attenuation in the line is represented by the number of dB steps from the left edge to the radius mark we have just transferred. For this example it is 0.8 dB. Some estimation may be required in interpolating between the 1-dB step marks.

Determining Velocity Factor

The velocity factor is determined by using the TOWARD GENERATOR wavelength scale of the Smith Chart. With a straightedge, draw a line from prime center through the point representing the short-circuit reading, until it intersects the wavelengths scale. In Fig 10 this point is labeled B. Consider that during our measurement, the short circuit was the load at the end of the line. Imagine a spiral curve progressing from 0 + *j* 0 clockwise and inward to our plotted measurement point. The wavelength scale, at B, indicates this line length is 0.464 λ. By rearranging the terms of Eq 1 given early in this chapter, we arrive at an equation for calculating the velocity factor.

$$VF = \frac{Lf}{984N} \qquad \text{(Eq 4)}$$

where

VF = velocity factor
L = line length, feet
f = frequency, MHz
N = number of electrical wavelengths in the line

Inserting the example values into Eq 4 yields VF = 100 × 7.15/(984 × 0.464) = 1.566, or 156.6%. Of course, this value is an impossible number—the velocity factor in coax cannot be greater than 100%. But remember, the Smith Chart can be used for lengths greater than $^1/_2$ λ. Therefore, that 0.464 value could rightly be 0.964, 1.464, 1.964, and so on. When using 0.964 λ, Eq 4 yields a velocity factor of 0.753, or 75.3%. Trying successively greater values for the wavelength results in velocity factors of 49.6 and 37.0%. Because the cable we measured had foamed dielectric, 75.3% is the probable velocity factor. This corresponds to an electrical length of 0.964 λ. Therefore, we have determined from the measurements and

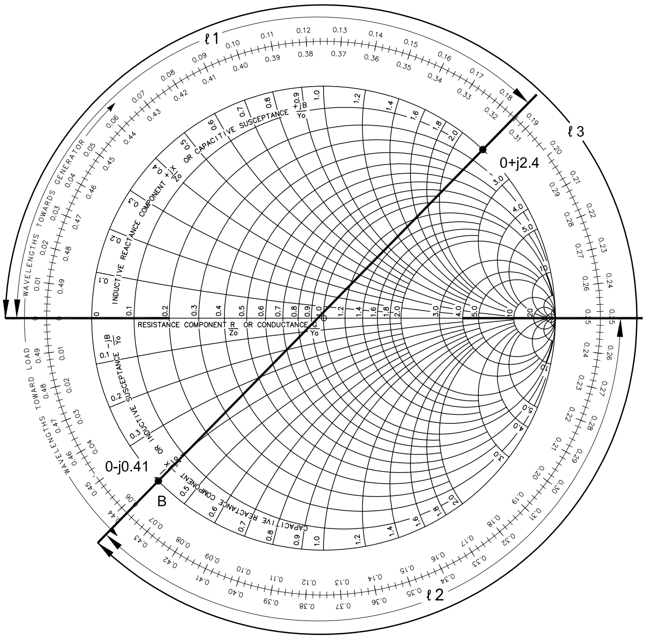
ℓ1
ℓ3
0+j2.4
0-j0.41
B
ℓ2
WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
RESISTANCE COMPONENT (R/Zo) OR CONDUCTANCE (G/Yo)
INDUCTIVE REACTANCE COMPONENT (+jX/Zo) OR CAPACITIVE SUSCEPTANCE (+jB/Yo)
CAPACITIVE REACTANCE COMPONENT (-jX/Zo) OR INDUCTIVE SUSCEPTANCE (-jB/Yo)

Fig 11—Smith Chart determination of input impedances for short- and open-circuited line sections, disregarding line losses.

calculations that our unmarked coax has a nominal 50-Ω impedance, an attenuation of 0.8 dB per hundred feet at 7.15 MHz, and a velocity factor of 75.3%.

It is difficult to use this procedure with short lengths of coax, just a few feet. The reason is that the SWR at the line input is too high to permit accurate measurements with most impedance bridges. In the example above, the SWR at the line input is approximately 12:1.

The procedure described above may also be used for determining the characteristics of balanced lines. However, impedance bridges are generally unbalanced devices, and the procedure for measuring a balanced impedance accurately with an unbalanced bridge is complicated.

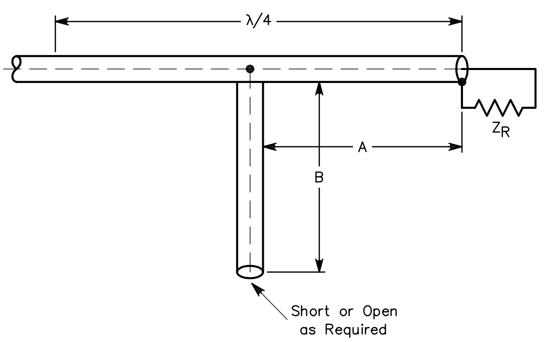

Fig 12—The method of stub matching a mismatched load on coaxial lines.

LINES AS CIRCUIT ELEMENTS

Information is presented in Chapter 24 on the use of transmission-line sections as circuit elements. For example, it is possible to substitute transmission lines of the proper length and termination for coils or capacitors in ordinary circuits. While there is seldom a practical need for that application, lines are frequently used in antenna systems in place of lumped components to tune or resonate elements. Probably the most common use of such a line is in the hairpin match, where a short section of stiff open-wire line acts as a lumped inductor.

The equivalent "lumped" value for any "inductor" or "capacitor" may be determined with the aid of the Smith Chart. Line losses may be taken into account if desired, as explained earlier. See **Fig 11**. Remember that the top half of the Smith Chart coordinate system is used for impedances containing inductive reactances, and the bottom half for capacitive reactances. For example, a section of 600-Ω line $^3/_{16}$-λ long (0.1875 λ) and short-circuited at the far end is represented by ℓ1, drawn around a portion of the perimeter of the chart. The "load" is a short-circuit, 0 + *j* 0 Ω, and the TOWARD GENERATOR wavelengths scale is used for marking off the line length. At A in Fig 11 may be read the normalized impedance as seen looking into the length of line, 0 + *j* 2.4. The reactance is therefore inductive, equal to 600 × 2.4 = 1440 Ω. The same line when open-circuited (termination impedance = ∞, the point at the right of the chart) is represented by ℓ2 in Fig 11. At B the normalized line-input impedance may be read as 0 – *j* 0.41; the reactance in this case is capacitive, 600 × 0.41 = 246 Ω. (Line losses are disregarded in these examples.) From Fig 11 it is easy to visualize that if ℓ1 were to be extended by $^1/_4$ λ, the total length represented by ℓ3, the line-input impedance would be identical to that obtained in the case represented by ℓ2 alone. In the case of ℓ2, the line is open-circuited at the far end, but in the case of ℓ3 the line is terminated in a short. The added section of line for ℓ3 provides the "transformer action" for which the $^1/_4$-λ line is noted.

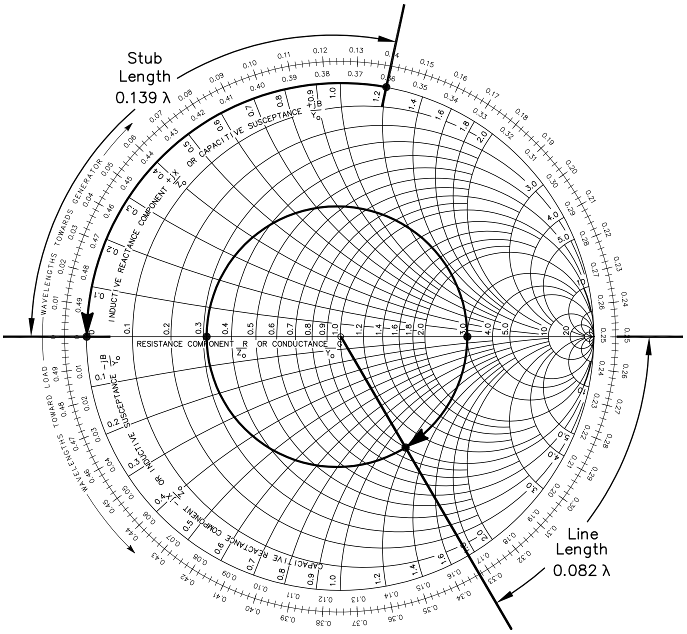

Fig 13—Smith Chart method of determining the dimensions for stub matching.

The equivalent inductance and capacitance as determined above can be found by substituting these values in the equations relating inductance and capacitance to reactance, or by using the various charts and calculators available. The frequency corresponding to the line length in degrees must be used, of course. In this example, if the frequency is 14 MHz the equivalent inductance and capacitance in the two cases are 16.4 µH and 46.2 pF, respectively. Note that when the line length is 45° (0.125 λ), the reactance in either case is numerically equal to the characteristic impedance of the line. In using the Smith Chart it should be kept in mind that the electrical length of a line section depends on the frequency and velocity of propagation, as well as on the actual physical length.

At lengths of line that are exact multiples of $^1/_4$ λ, such lines have the properties of resonant circuits. At lengths where the input reactance passes through zero at the left of the Smith Chart, the line acts as a series-resonant circuit. At lengths for which the reactances theoretically pass from "positive" to "negative" infinity at the right of the Smith Chart, the line simulates a parallel-resonant circuit.

Designing Stub Matches with the Smith Chart

The design of stub matches is covered in detail in Chapter 26. Equations are presented there to calculate the electrical lengths of the main line and the stub, based on a purely resistive load and on the stub being the same type of line as the main line. The Smith Chart may also be used to determine these lengths, without the requirements that the load be purely resistive and that the line types be identical.

Fig 12 shows the stub matching arrangement in coaxial line. As an example, suppose that the load is an antenna, a close-spaced array fed with a 52-Ω line. Further suppose that the SWR has been measured as 3.1:1. From this information, a constant SWR circle may be drawn on the Smith Chart. Its radius is such that it intersects the right portion of the resistance axis at the SWR value, 3.1, as shown at point B in **Fig 13**.

Since the stub of Fig 12 is connected in parallel with the transmission line, determining the design of the matching arrangement is simplified if Smith Chart values are dealt with as admittances, rather than impedances. (An admittance is simply the reciprocal of the associated impedance. Plotted on the Smith Chart, the two associated points are on the same SWR circle, but diametrically opposite each other.) Using admittances leaves less chance for errors in making calculations, by eliminating the need for making series-equivalent to parallel-equivalent circuit conversions and back, or else for using complicated equations for determining the resultant value of two complex impedances connected in parallel.

A complex impedance, Z, is equal to R + *j* X, as described in Chapter 24. The equivalent admittance, Y, is equal to G – *j* B, where G is the conductive component and B the susceptance. (Inductance is taken as negative susceptance, and capacitance as positive.) Conductance and susceptance values are plotted and handled on the Smith Chart in the same manner as resistance and reactance.

Assuming that the close-spaced array of our example has been resonated at the operating frequency, it will present a purely resistive termination for the load end of the 52-Ω line. From information in Chapter 24, it is known that the impedance of the antenna equals Z_0/SWR = 52/3.1 = 16.8 Ω. (We can logically discard the possibility that the antenna impedance is SWR × Z_0, or 0.06 Ω.) If this 16.8-Ω value were to be plotted as an impedance on the Smith Chart, it would first be normalized (16.8/52 = 0.32) and then plotted as 0.32 + *j* 0. Although not necessary for the solution of this example, this value is plotted at point A in Fig 13. What is necessary is a plot of the admittance for the antenna as a load. This is the reciprocal of the impedance; 1/16.8 Ω equals 0.060 siemen. To plot this point it is first normalized by multiplying the conductance and susceptance values by the Z_0 of the line. Thus, (0.060 + *j* 0) × 52 = 3.1 + *j* 0. This admittance value is shown plotted at point B in Fig 13. It may be seen that points A and B are diametrically opposite each other on the chart. Actually, for the solution of this example, it wasn't necessary to compute the values for either point A or point B as in the above paragraph, for they were both determined from the known SWR value of 3.1. As may be seen in Fig 13, the points are located on the constant SWR circle which was already drawn, at the two places where it intersects the resistance axis. The plotted value for point A, 0.32, is simply the reciprocal of the value for point B, 3.1. However, an understanding of the relationship between impedance and admittance is easier to gain with simple examples such as this.

In stub matching, the stub is to be connected at a point in the line where the conductive component equals the Z_0 of the line. Point B represents the admittance of the load, which is the antenna. Various admittances will be encountered along the line, when moving in a direction indicated by the TOWARD GENERATOR wavelengths scale, but all admittance plots must fall on the constant SWR circle. Moving clockwise around the SWR circle from point B, it is seen that the line input conductance will be 1.0 (normalized Z_0 of the line) at point C, 0.082 λ toward the transmitter from the antenna. Thus, the stub should be connected at this location on the line.

The normalized admittance at point C, the point representing the location of the stub, is 1 – *j* 1.2 siemens, having an inductive susceptance component. A capacitive susceptance having a normalized value of + *j* 1.2 siemens is required across the line at the point of stub connection, to cancel the inductance. This capacitance is to be obtained from the stub section itself; the problem now is to determine its type of termination (open or shorted), and how long the stub should be. This is done by first plotting the susceptance required for cancellation, 0 + *j* 1.2, on the chart (point D in Fig 13). This point represents the input admittance as seen looking into the stub. The "load" or termination for the stub section is found by moving in the TOWARD LOAD direction around the chart, and will appear at the closest point on the resistance/conductance axis, either at the left or the right of the chart. Moving counterclockwise from point D, this is located at E, at the left of the chart, 0.139 λ away. From this we know the required stub length. The "load" at the far end of the stub, as represented on the Smith Chart, has a normalized admittance of $0 + j\,0$ siemen, which is equivalent to an open circuit.

When the stub, having an input admittance of $0 + j\,1.2$ siemens, is connected in parallel with the line at a point 0.082 λ from the load, where the line input admittance is $1.0 - j\,1.2$, the resultant admittance is the sum of the individual admittances. The conductance components are added directly, as are the susceptance components. In this case, $1.0 - j\,1.2 + j\,1.2 = 1.0 + j\,0$ siemen. Thus, the line from the point of stub connection to the transmitter will be terminated in a load which offers a perfect match. When determining the physical line lengths for stub matching, it is important to remember that the velocity factor for the type of line in use must be considered.

MATCHING WITH LUMPED CONSTANTS

It was pointed out earlier that the purpose of a matching stub is to cancel the reactive component of line impedance at the point of connection. In other words, the stub is simply a reactance of the proper kind and value shunted across the line. It does not matter what physical shape this reactance takes. It can be a section of transmission line or a "lumped" inductance or capacitance, as desired. In the above example with the Smith Chart solution, a capacitive reactance was required. A capacitor having the same value of reactance can be used just as well. There are cases where, from an installation standpoint, it may be considerably more convenient to connect a capacitor in place of a stub. This is particularly true when open-wire feeders are used. If a variable capacitor is used, it becomes possible to adjust the capacitance to the exact value required.

The proper value of reactance may be determined from Smith Chart information. In the previous example, the required susceptance, normalized, was $+j\,1.2$ siemens. This is converted into actual siemens by dividing by the line Z_0; 1.2/52 = 0.023 siemen, capacitance. The required capacitive reactance is the reciprocal of this latter value, 1/0.023 = 43.5 Ω. If the frequency is 14.2 MHz, for instance, 43.5 Ω corresponds to a capacitance of 258 pF. A 325-pF variable capacitor connected across the line 0.082 λ from the antenna terminals would provide ample adjustment range. The RMS voltage across the capacitor is

$$E = \sqrt{P \times Z_0}$$

For 500 W, for example, E = the square root of 500 × 52 = 161 V. The peak voltage is 1.41 times the RMS value, or 227 V.

The Series-Section Transformer

The series-section transformer is described in Chapter 26, and equations are given there for its design. The transformer can be designed graphically with the aid of a Smith Chart. This information is based on a *QST* article by Frank A. Regier, OD5CG. Using the Smith Chart to design a series-section match requires the use of the chart in its less familiar off-center mode. This mode is described in the next two paragraphs.

Fig 14 shows the Smith Chart used in its familiar centered mode, with all impedances normalized to that of the transmission line, in this case 75 Ω, and all constant SWR circles concentric with the normalized value r = 1 at the chart center. An actual impedance is recovered by multiplying a chart reading by the normalizing impedance of 75 Ω. If the actual (unnormalized) impedances represented by a constant SWR circle in Fig 14 are instead divided by a normalizing impedance of 300 Ω, a different picture results. A Smith Chart shows all possible impedances, and so a closed path such as a constant SWR circle in Fig 14 must again be represented by a closed path. In fact, it can be shown that the path remains a circle, but that the constant SWR circles are no longer concentric. **Fig 15** shows the circles that result when the impedances along a mismatched 75-Ω line are normalized by dividing by 300 Ω instead of 75. The constant SWR circles still surround the point corresponding to the characteristic impedance of the line (r = 0.25) but are no longer concentric with it. Note that the normalized impedances read from corresponding points on Figs 14 and 15 are different but that the actual, unnormalized, impedances are exactly the same.

An Example

Now turn to the example shown in **Fig 16**. A complex load of $Z_L = 600 + j\,900\ \Omega$ is to be fed with 300-Ω line, and a 75-Ω series section is to be used. These characteristic impedances agree with those used in Fig 15, and thus Fig 15 can be used to find the impedance variation along the 75-Ω series section. In particular, the constant SWR circle which passes through the Fig 15 chart center, SWR = 4 in this case, passes through all the impedances (normalized to 300 Ω) which the 75-Ω series section is able to match to the 300-Ω main line. The length $\ell 1$ of 300-Ω line has the job of transforming the load impedance to some impedance on this matching circle.

Fig 17 shows the whole process more clearly, with all

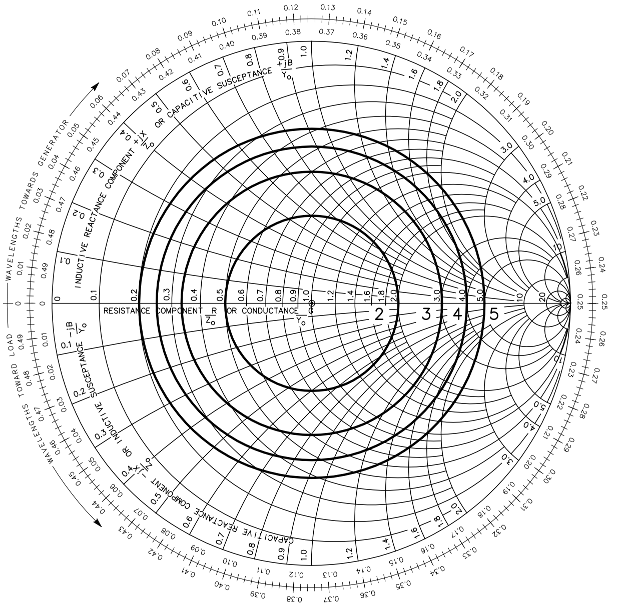

Fig 14—Constant SWR circles for SWR = 2, 3, 4 and 5, showing impedance variation along 75-Ω line, normalized to 75 Ω. The actual impedance is obtained by multiplying the chart reading by 75 Ω.

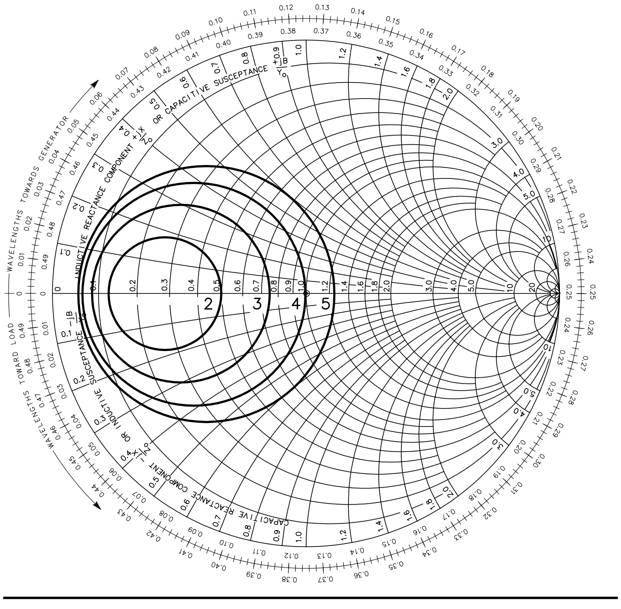

Fig 15—Paths of constant SWR for SWR = 2, 3, 4 and 5, showing impedance variation along 75-Ω line, normalized to 300 Ω. Normalized impedances differ from those in Fig 14, but actual impedances are obtained by multiplying chart readings by 300 Ω and are the same as those corresponding in Fig 14. Paths remain circles but are no longer concentric. One, the matching circle, SWR = 4 in this case, passes through the chart center and is thus the locus of all impedances which can be matched to a 300-Ω line.

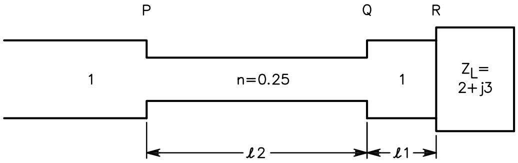

Fig 16—Example for solution by Smith Chart. All impedances are normalized to 300 Ω.

impedances normalized to 300 Ω. Here the normalized load impedance $Z_L = 2 + j\,3$ is shown at R, and the matching circle appears centered on the resistance axis and passing through the points r = 1 and $r = n^2 = (75/300)^2 = 0.0625$. A constant SWR circle is drawn from R to an intersection with the matching circle at Q or Q′ and the corresponding length $\ell1$ (or $\ell1'$) can be read directly from the Smith Chart. The clockwise distance around the matching circle represents the length of the matching line, from either Q′ to P or from Q to P. Because in this example the distance QP is the shorter of the two for the matching section, we choose the length $\ell1$ as shown. By using values from the TOWARD GENERATOR scale, this length is found as 0.045 – 0.213, and adding 0.5 to obtain a positive result yields a value of 0.332 λ.

Although the impedance locus from Q to P is shown in Fig 17, the length $\ell2$ cannot be determined directly from this chart. This is because the matching circle is not concentric with the chart center, so the wavelength scales do not apply to this circle. This problem is overcome by forming **Fig 18**, which is the same as Fig 17 except that all normalized impedances have been divided by n = 0.25, resulting in a Smith Chart normalized to 75 Ω instead of 300. The matching circle and the chart center are now concentric, and the series-section length $\ell2$, the distance between Q and P, can be taken directly from the chart. By again using the TOWARD GENERATOR scale, this length is found as 0.250 – 0.148 = 0.102 λ.

In fact it is not necessary to construct the entire impedance locus shown in Fig 18. It is sufficient to plot Z_Q/n (Z_Q is read from Fig 17) and $Z_p/n = 1/n$, connect them by a circular arc centered on the chart center, and to determine the arc length $\ell2$ from the Smith Chart.

Procedure Summary

The steps necessary to design a series-section transformer by means of the Smith Chart can now be listed:

1) Normalize all impedances by dividing by the characteristic impedance of the main line.
2) On a Smith Chart, plot the normalized load impedance Z_L at R and construct the matching circle

so that its center is on the resistance axis and it passes through the points r = 1 and r = n^2.

3) Construct a constant SWR circle centered on the chart center through point R. This circle should intersect the matching circle at two points. One of these points, normally the one resulting in the shorter clockwise distance along the matching circle to the chart center, is chosen as point Q, and the clockwise distance from R to Q is read from the chart and taken to be $\ell 1$.
4) Read the impedance Z_Q from the chart, calculate Z_Q/n and plot it as point Q on a second Smith Chart. Also plot r = 1/n as point P.
5) On this second chart construct a circular arc, centered on the chart center, clockwise from Q to P. The length of this arc, read from the chart, represents $\ell 2$. The design of the transformer is now complete, and the necessary physical line lengths may be determined.

The Smith Chart construction shows that two design solutions are usually possible, corresponding to the two intersections of the constant SWR circle (for the load) and the matching circle. These two values correspond to positive and negative values of the square-root radical in the equation for a mathematical solution of the problem. It may happen, however, that the load circle misses the matching circle completely, in which case no solution is possible. The cure is to enlarge the matching circle by choosing a series section whose impedance departs more from that of the main line.

A final possibility is that, rather than intersecting the matching circle, the load circle is tangent to it. There is then but one solution—that of the ¼-λ transformer.

BIBLIOGRAPHY

Source material and more extended discussion of topics covered in this chapter can be found in the references given below and in the textbooks listed at the end of Chapter 2.

W. N. Caron, *Antenna Impedance Matching* (Newington: ARRL, 1989).

C. MacKeand, “The Smith Chart in BASIC,” *QST*, Nov 1984, pp 28-31.

M. W. Maxwell, *Reflections—Transmission Lines and Antennas* (Newington: ARRL, 1990).

F. A. Regier, Series-Section Transmission-Line Impedance Matching,” *QST*, Jul 1978, pp 14-16.

P. H. Smith, *Electronic Applications of the Smith Chart*, reprint ed. (Malabar, FL: Krieger Pub Co, Inc, 1983).

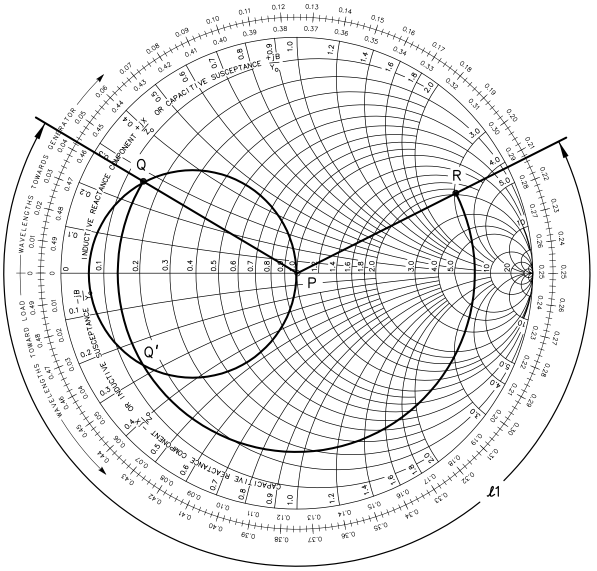
Q
R
P
Q′
ℓ1
WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT
OR CAPACITIVE SUSCEPTANCE
CAPACITIVE REACTANCE COMPONENT
OR INDUCTIVE SUSCEPTANCE

Fig 17—Smith Chart representation of the example shown in Fig 16. The impedance locus always takes a clockwise direction from the load to the generator. This path is first along the constant SWR circle from the load at R to an intersection with the matching circle at Q or Q′, and then along the matching circle to the chart center at P. Length ℓ1 can be determined directly from the chart, and in this example is 0.332 λ.

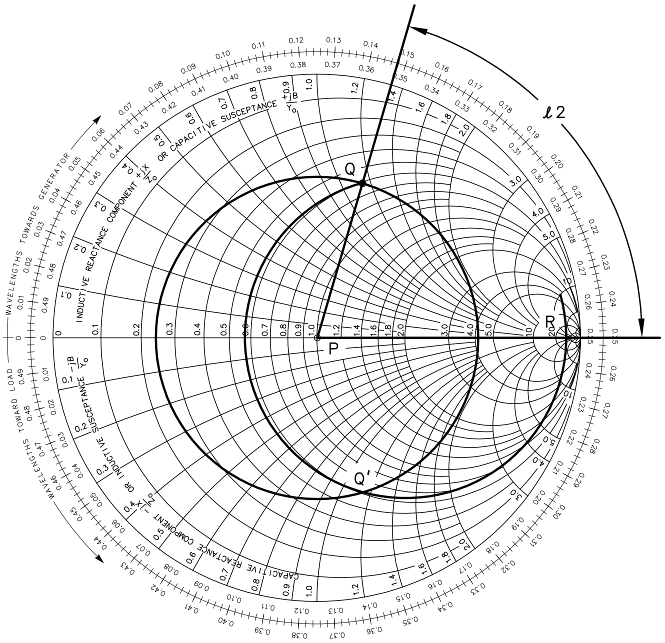
ℓ2
Q
R
P
Q′
WAVELENGTHS TOWARDS GENERATOR
WAVELENGTHS TOWARD LOAD
INDUCTIVE REACTANCE COMPONENT
OR CAPACITIVE SUSCEPTANCE
CAPACITIVE REACTANCE COMPONENT
OR INDUCTIVE SUSCEPTANCE

Fig 18—The same impedance locus as shown in Fig 17 except normalized to 75 Ω instead of 300. The matching circle is now concentric with the chart center, and ℓ2 can be determined directly from the chart, 0.102 λ in this case.

Appendix

This appendix contains a glossary of terms, a list of common abbreviations, length conversion information (feet and inches), metric equivalents and antenna-gain-reference data.

Glossary of Terms

This glossary provides a handy list of terms that are used frequently in Amateur Radio conversation and literature about antennas. With each item is a brief definition of the term. Most terms given here are discussed more thoroughly in the text of this book, and may be located by using the index.

Actual ground—The point within the earth's surface where effective ground conductivity exists. The depth for this point varies with frequency and the condition of the soil.

Antenna—An electrical conductor or array of conductors that radiates signal energy (transmitting) or collects signal energy (receiving).

Antenna tuner—A device containing variable reactances (and perhaps a balun). It is connected between the transmitter and the feed point of an antenna system, and adjusted to "tune" or resonate the system to the operating frequency.

Aperture, effective—An area enclosing an antenna, on which it is convenient to make calculations of field strength and antenna gain. Sometimes referred to as the "capture area."

Apex—The feed-point region of a V type of antenna.

Apex angle—The included angle between the wires of a V, an inverted-V dipole, and similar antennas, or the included angle between the two imaginary lines touching the element tips of a log periodic array.

Balanced line—A symmetrical two-conductor feed line that has uniform voltage and current distribution along its length.

Balun—A device for feeding a balanced load with an unbalanced line, or vice versa. May be a form of choke, or a transformer that provides a specific impedance transformation (including 1:1). Often used in antenna systems to interface a coaxial transmission line to the feed point of a balanced antenna, such as a dipole.

Base loading—A lumped reactance that is inserted at the base (ground end) of a vertical antenna to resonate the antenna.

Bazooka—A transmission-line balancer. It is a quarter-wave conductive sleeve (tubing or flexible shielding) placed at the feed point of a center-fed element and grounded to the shield braid of the coaxial feed line at the end of the sleeve farthest from the feed point. It permits the use of unbalanced feed line with balanced feed antennas.

Beamwidth—Related to directive antennas. The width, in degrees, of the major lobe between the two directions at which the relative radiated power is equal to one half its value at the peak of the lobe (half power = –3 dB).

Beta match—A form of hairpin match. The two conductors straddle the boom of the antenna being matched, and the closed end of the matching-section conductors is strapped to the boom.

Bridge—A circuit with two or more ports that is used in measurements of impedance, resistance or standing waves in an antenna system. When the bridge is adjusted for a balanced condition, the unknown factor can be determined by reading its value on a calibrated scale or meter.

Capacitance hat—A conductor of large surface area that is connected at the high-impedance end of an antenna to effectively increase the electrical length. It is sometimes mounted directly above a loading coil to reduce the required inductance for establishing resonance. It usually takes the form of a series of wheel spokes or a solid circular disc. Sometimes referred to as a "top hat."

Capture area—See aperture.

Center fed—Transmission-line connection at the electrical center of an antenna radiator.

Center loading—A scheme for inserting inductive reactance (coil) at or near the center of an antenna element for the purpose of lowering its resonant frequency. Used with elements that are less than $^1/_4$ wavelength at the operating frequency.

Coax—See coaxial cable.

Coaxial cable—Any of the coaxial transmission lines that have the outer shield (solid or braided) on the same axis as the inner or center conductor. The insulating material can be air, helium or solid-dielectric compounds.

Collinear array—A linear array of radiating elements (usually dipoles) with their axes arranged in a straight line. Popular at VHF and above.

Conductor—A metal body such as tubing, rod or wire that permits current to travel continuously along its length.

Counterpoise—A wire or group of wires mounted close to ground, but insulated from ground, to form a low-impedance, high-capacitance path to ground. Used at MF and HF to provide an RF ground for an antenna. Also see ground plane.

Current loop—A point of current maxima (antipode) on an antenna.

Current node—A point of current minima on an antenna.

Decibel—A logarithmic power ratio, abbreviated dB. May also represent a voltage or current ratio if the voltages or currents are measured across (or through) identical impedances. Suffixes to the abbreviation indicate references: dBi, isotropic radiator; dBic, isotropic radiator circular; dBm, milliwatt; dBW, watt.

Delta loop—A full-wave loop shaped like a triangle or delta.

Delta match—Center-feed technique used with radiators that are not split at the center. The feed line is fanned near the radiator center and connected to the radiator symmetrically. The fanned area is delta shaped.

Dielectrics—Various insulating materials used in antenna systems, such as found in insulators and transmission lines.

Dipole—An antenna that is split at the exact center for connection to a feed line, usually a half wavelength long. Also called a "doublet."

Direct ray—Transmitted signal energy that arrives at the receiving antenna directly rather than being reflected by any object or medium.

Directivity—The property of an antenna that concentrates the radiated energy to form one or more major lobes.

Director—A conductor placed in front of a driven element to cause directivity. Frequently used singly or in multiples with Yagi or cubical-quad beam antennas.

Doublet—See dipole.

Driven array—An array of antenna elements which are all driven or excited by means of a transmission line, usually to achieve directivity.

Driven element—A radiator element of an antenna system to which the transmission line is connected.

Dummy load—Synonymous with dummy antenna. A nonradiating substitute for an antenna.

E layer—The ionospheric layer nearest earth from which radio signals can be reflected to a distant point, generally a maximum of 2000 km (1250 ml).

E plane—Related to a linearly polarized antenna, the plane containing the electric field vector of the antenna and its direction of maximum radiation. For terrestrial antenna systems, the direction of the E plane is also taken as the polarization of the antenna. The E plane is at right angles to the H plane.

Efficiency—The ratio of useful output power to input power, determined in antenna systems by losses in the system, including in nearby objects.

EIRP—Effective isotropic radiated power. The power radiated by an antenna in its favored direction, taking the gain of the antenna into account as referenced to isotropic.

Elements—The conductive parts of an antenna system that determine the antenna characteristics. For example, the reflector, driven element and directors of a Yagi antenna.

End effect—A condition caused by capacitance at the ends of an antenna element. Insulators and related support wires contribute to this capacitance and lower the resonant frequency of the antenna. The effect increases with conductor diameter and must be considered when cutting an antenna element to length.

End fed—An end-fed antenna is one to which power is applied at one end, rather than at some point between the ends.

F layer—The ionospheric layer that lies above the E layer. Radio waves can be refracted from it to provide communications distances of several thousand miles by means of single- or double-hop skip.

Feed line—See feeders.

Feeders—Transmission lines of assorted types that are used to route RF power from a transmitter to an antenna, or from an antenna to a receiver.

Field strength—The intensity of a radio wave as measured at a point some distance from the antenna. This measurement is usually made in microvolts per meter.

Front to back—The ratio of the radiated power off the front and back of a directive antenna. For example, a dipole would have a ratio of 1, which is equivalent to 0 dB.

Front to rear—Worst-case rearward lobe in the 180°-wide sector behind an antenna's main lobe, in dB.

Front to side—The ratio of radiated power between the major lobe and that 90° off the front of a directive antenna.

Gain—The increase in effective radiated power in the desired direction of the major lobe.

Gamma match—A matching system used with driven antenna elements to effect a match between the transmission line and the feed point of the antenna. It consists of a series capacitor and an arm that is mounted close to the driven element and in parallel with it near the feed point.

Ground plane—A system of conductors placed beneath an elevated antenna to serve as an earth ground. Also see counterpoise.

Ground screen—A wire mesh counterpoise.

Ground wave—Radio waves that travel along the earth's surface.

H plane—Related to a linearly polarized antenna. The plane containing the magnetic field vector of an antenna and its direction of maximum radiation. The H plane is at right angles to the E plane.

HAAT—Height above average terrain. A term used mainly in connection with repeater antennas in determining coverage area.

Hairpin match—A U-shaped conductor that is connected to the two inner ends of a split dipole for the purpose of creating an impedance match to a balanced feeder.

Harmonic antenna—An antenna that will operate on its fundamental frequency and the harmonics of the fundamental frequency for which it is designed. An end-fed half-wave antenna is one example.

Helical—A helically wound antenna, one that consists of a spiral conductor. If it has a very large winding length to diameter ratio it provides broadside radiation. If the length-to-diameter ratio is small, it will operate in the axial mode and radiate off the end opposite the feed point. The polarization will be circular for the axial mode, with left or right circularity, depending on whether the helix is wound clockwise or counterclockwise.

Helical hairpin—"Hairpin" match with a lumped inductor, rather than parallel-conductor line.

Image antenna—The imaginary counterpart of an actual antenna. It is assumed for mathematical purposes to be located below the earth's surface beneath the antenna, and is considered symmetrical with the antenna above ground.

Impedance—The ohmic value of an antenna feed point, matching section or transmission line. An impedance may contain a reactance as well as a resistance component.

Inverted V—A misnomer, as the antenna being referenced does not have the characteristics of a V antenna. See inverted-V dipole.

Inverted-V dipole—A half-wavelength dipole erected in the form of an upside-down V, with the feed point at the apex. Its radiation pattern is similar to that of a horizontal dipole.

Isotropic—An imaginary or hypothetical point-source antenna that radiates equal power in all directions. It is used as a reference for the directive characteristics of actual antennas.

Lambda—Greek symbol (λ) used to represent a wavelength with reference to electrical dimensions in antenna work.

Line loss—The power lost in a transmission line, usually expressed in decibels.

Line of sight—Transmission path of a wave that travels directly from the transmitting antenna to the receiving antenna.

Litz wire—Stranded wire with individual strands insulated; small wire provides a large surface area for current flow, so losses are reduced for the wire size.

Load—The electrical entity to which power is delivered. The antenna system is a load for the transmitter.

Loading—The process of a transferring power from its source to a load. The effect a load has on a power source.

Lobe—A defined field of energy that radiates from a directive antenna.

Log periodic antenna—A broadband directive antenna that has a structural format causing its impedance and radiation characteristics to repeat periodically as the logarithm of frequency.

Long wire—A wire antenna that is one wavelength or greater in electrical length. When two or more wavelengths long it provides gain and a multilobe radiation pattern. When terminated at one end it becomes essentially unidirectional off that end.

Marconi antenna—A shunt-fed monopole operated against ground or a radial system. In modern jargon, the term refers loosely to any type of vertical antenna.

Matching—The process of effecting an impedance match between two electrical circuits of unlike impedance. One example is matching a transmission line to the feed point of an antenna. Maximum power transfer to the load (antenna system) will occur when a matched condition exists.

Monopole—Literally, one pole, such as a vertical radiator operated against the earth or a counterpoise.

Nichrome wire—An alloy of nickel and chromium; not a good conductor; resistance wire. Used in the heating elements of electrical appliances; also as conductors in transmission lines or circuits where attenuation is desired.

Null—A condition during which an electrical unit is at a minimum. A null in an antenna radiation pattern is a point in the 360-degree pattern where a minima in field intensity is observed. An impedance bridge is said to be "pulled" when it has been brought into balance, with a null in the current flowing through the bridge arm.

Octave—A musical term. As related to RF, frequencies having a 2:1 harmonic relationship.

Open-wire line—A type of transmission line that resembles a ladder, sometimes called "ladder line." Consists of parallel, symmetrical wires with insulating spacers at regular intervals to maintain the line spacing. The dielectric is principally air, making it a low-loss type of line.

Parabolic reflector—An antenna reflector that is a portion of a parabolic revolution or curve. Used mainly at UHF and higher to obtain high gain and a relatively narrow beamwidth when excited by one of a variety of driven elements placed in the plane of and perpendicular to the axis of the parabola.

Parasitic array—A directive antenna that has a driven element and at least one independent director or reflector, or a combination of both. The directors and reflectors are not connected to the feed line. Except for VHF and UHF arrays with long booms (electrically), more than one reflector is seldom used. A Yagi antenna is one example of a parasitic array.

Phasing lines—Sections of transmission line that are used to ensure the correct phase relationship between the elements of a driven array, or between bays of an array of antennas. Also used to effect impedance transformations while maintaining the desired phase.

Polarization—The sense of the wave radiated by an antenna. This can be horizontal, vertical, elliptical or circular (left or right hand circularity), depending on the design and application. (See H plane.)

Q section—Term used in reference to transmission-line matching transformers and phasing lines.

Quad—A parasitic array using rectangular or diamond shaped full-wave wire loop elements. Often called the "cubical quad." Another version uses delta-shaped elements, and is called a delta loop beam.

Radiation pattern—The radiation characteristics of an antenna as a function of space coordinates. Normally, the pattern is measured in the far-field region and is represented graphically.

Radiation resistance—The ratio of the power radiated by an antenna to the square of the RMS antenna current, referred to a specific point and assuming no losses. The effective resistance at the antenna feed point.

Radiator—A discrete conductor that radiates RF energy in an antenna system.

Random wire—A random length of wire used as an antenna and fed at one end by means of an antenna tuner. Seldom operates as a resonant antenna unless the length happens to be correct.

Reflected ray—A radio wave that is reflected from the earth, ionosphere or a man-made medium, such as a passive reflector.

Reflector—A parasitic antenna element or a metal assembly that is located behind the driven element to enhance forward directivity. Hillsides and large man-made structures such as buildings and towers may act as reflectors.

Refraction—Process by which a radio wave is bent and returned to earth from an ionospheric layer or other medium after striking the medium.

Resonator—In antenna terminology, a loading assembly consisting of a coil and a short radiator section. Used to lower the resonant frequency of an antenna, usually a vertical or a mobile whip.

Rhombic—A rhomboid or diamond-shaped antenna consisting of sides (legs) that are each one or more wavelengths long. The antenna is usually erected parallel to the ground. A rhombic antenna is bidirectional unless terminated by a resistance, which makes it unidirectional. The greater the electrical leg length, the greater the gain, assuming the tilt angle is optimized.

Shunt feed—A method of feeding an antenna driven element with a parallel conductor mounted adjacent to a low-impedance point on the radiator. Frequently used with grounded quarter-wave vertical antennas to provide an impedance match to the feeder. Series feed is used when the base of the vertical is insulated from ground.

Stacking—The process of placing similar directive antennas atop or beside one another, forming a "stacked array." Stacking provides more gain or directivity than a single antenna.

Stub—A section of transmission line used to tune an antenna element to resonance or to aid in obtaining an impedance match.

SWR—Standing-wave ratio on a transmission line in an antenna system. More correctly, VSWR, or voltage standing-wave ratio. The ratio of the forward to reflected voltage on the line, and not a power ratio. A VSWR of 1:1 occurs when all parts of the antenna system are matched correctly to one another.

T match—Method for matching a transmission-line to an unbroken driven element. Attached at the electrical center of the driven element in a T-shaped manner. In effect it is a double gamma match.

Tilt angle—Half the angle included between the wires at the sides of a rhombic antenna.

Top hat—See capacitance hat.

Top loading—Addition of a reactance (usually a capacitance hat) at the end of an antenna element opposite the feed point to increase the electrical length of the radiator.

Transmatch—An antenna tuner.

Trap—Parallel L-C network inserted in an antenna element to provide multiband operation with a single conductor.

Unipole—See monopole.

Velocity factor—The ratio of the velocity of radio wave propagation in a dielectric medium to that in free space. When cutting a transmission line to a specific electrical length, the velocity factor of the particular line must be taken into account.

VSWR—Voltage standing-wave ratio. See SWR.

Wave—A disturbance or variation that is a function of time or space, or both, transferring energy progressively from point to point. A radio wave, for example.

Wave angle—The angle above the horizon of a radio wave as it is launched from or received by an antenna. Also called elevation angle.

Wave front—A surface that is a locus of all the points having the same phase at a given instant in time.

Yagi—A directive, gain type of antenna that utilizes a number of parasitic directors and a reflector. Named after one of the two Japanese inventors (Yagi and Uda).

Zepp antenna—A half-wave wire antenna that operates on its fundamental and harmonics. It is fed at one end by means of open-wire feeders. The name evolved from its popularity as an antenna on Zeppelins. In modern jargon the term refers loosely to any horizontal antenna.

Abbreviations

Abbreviations and acronyms that are commonly used throughout this book are defined in the list below. Periods are not part of an abbreviation unless the abbreviation otherwise forms a common English word. When appropriate, abbreviations as shown are used in either singular or plural construction.

-A-
A—ampere
ac—alternating current
AF—audio frequency
AFSK—audio frequency-shift keying
AGC—automatic gain control
AM—amplitude modulation
ANT—antenna
ARRL—American Radio Relay League
ATV—amateur television
AWG—American wire gauge
az-el—azimuth-elevation

-B-
balun—balanced to unbalanced
BC—broadcast
BCI—broadcast interference
BW—bandwidth

-C-
ccw—counterclockwise
cm—centimeter
coax—coaxial cable
CT—center tap
cw—clockwise
CW—continuous wave

-D-
D—diode
dB—decibel
dBd—decibels referenced to a dipole
dBi—decibels referenced to isotropic
dBic—decibels referenced to isotropic, circular
dBm—decibels referenced to one milliwatt
dBW—decibels referenced to one watt
dc—direct current
deg—degree
DF—direction finding
dia—diameter
DPDT—double pole, double throw
DPST—double pole, single throw
DVM—digital voltmeter
DX—long distance communication

-E-
E—ionospheric layer, electric field
ed.—edition
Ed.—editor
EIRP—effective isotropic radiated power
ELF—extremely low frequency
EMC—electromagnetic compatibility
EME—earth-moon-earth
EMF—electromotive force
ERP—effective radiated power
E_S—ionospheric layer (sporadic E)

-F-
f—frequency
F—ionospheric layer, farad
F/B—front to back (ratio)
F/R—worst-case front to rear (ratio)
FM—frequency modulation
FOT—frequency of optimum transmission
ft—foot or feet (unit of length)
F_1—ionospheric layer
F_2—ionospheric layer

-G-
GDO—grid- or gate-dip oscillator
GHz—gigahertz
GND—ground

-H-
H—magnetic field, henry
HAAT—height above average terrain
HF—high frequency (3-30 MHz)
Hz—hertz (unit of frequency)

-I-
I—current
ID—inside diameter
IEEE—Institute of Electrical and Electronic Engineers
in.—inch
IRE—Institute of Radio Engineers (now IEEE)

-J-
j—vector notation

-K-
kHz—kilohertz
km—kilometer
kW—kilowatt
kW—kilohm

-L-
L—inductance
lb—pound (unit of mass)
LF—low frequency (30-300 kHz)
LHCP—left-hand circular polarization
ln—natural logarithm
log—common logarithm
LP—log periodic
LPDA—log periodic dipole array
LPVA—log periodic V array
LUF—lowest usable frequency

-M-
m—meter (unit of length)
m/s—meters per second
mA—milliampere
max—maximum
MF—medium frequency (0.3-3 MHz)
mH—millihenry
MHz—megahertz
mi—mile
min—minute
mm—millimeter
ms—millisecond
mS—millisiemen
MS—meteor scatter
MUF—maximum usable frequency
mW—milliwatt
MW—megohm

-N-
NC—no connection, normally closed
NiCd—nickel cadmium
NIST—National Institute of Standards and Technology
NO—normally open
no.—number

-O-
OD—outside diameter

-P-
p—page (bibliography reference)
P-P—peak to peak
PC—printed circuit
PEP—peak envelope power
pF—picofarad
pot—potentiometer
pp—pages (bibliography reference)
Proc—Proceedings

-Q-
Q—figure of merit

-R-
R—resistance, resistor
RF—radio frequency
RFC—radio frequency choke
RFI—radio frequency interference
RHCP—right-hand circular polarization
RLC—resistance-inductance-capacitance
r/min—revolutions per minute
RMS—root mean square
r/s—revolutions per second
RSGB—Radio Society of Great Britain
RX—receiver

-S-
s—second
S—siemen
S/NR—signal-to-noise ratio
SASE—self-addressed stamped envelope
SINAD—signal-to-noise and distortion
SPDT—single pole, double throw
SPST—single pole, single throw
SWR—standing wave ratio
sync—synchronous

-T-
tpi—turns per inch
TR—transmit-receive
TVI—television interference
TX—transmitter

-U-
UHF—ultra-high frequency (300-3000 MHz)
US—United States
UTC—Universal Time, Coordinated

-V-
V—volt
VF—velocity factor
VHF—very-high frequency (30-300 MHz)
VLF—very-low frequency (3-30 kHz)
Vol—volume (bibliography reference)
VOM—volt-ohm meter
VSWR—voltage standing-wave ratio
VTVM—vacuum-tube voltmeter

-W-
W—watt
WPM—words per minute
WRC—World Radio Conference
WVDC—working voltage, direct current

-X-
X—reactance
XCVR—transceiver
XFMR—transformer
XMTR—transmitter

-Z-
Z—impedance

-Other symbols and Greek letters-
°—degrees
λ—wavelength
λ/dia—wavelength to diameter (ratio)
µ—permeability
µF—microfarad
µH—microhenry
µV—microvolt
Ω—ohm
ϕ—angles
π—3.14159
θ—angles

Length Conversions

Throughout this book, equations may be found for determining the design length and spacing of antenna elements. For convenience, the equations are written to yield a result in feet. (The answer may be converted to meters simply by multiplying the result by 0.3048.) If the result in feet is not an integral number, however, it is necessary to make a conversion from a decimal fraction of a foot to inches and fractions before the physical distance can be determined with a conventional tape measure. Table 1 may be used for this conversion, showing inches and fractions for increments of 0.01 foot. The table deals with only the fractional portion of a foot. The integral number of feet remains the same.

For example, say a calculation yields a result of 11.63 feet, and we wish to convert this to a length we can find on a tape measure. For the moment, consider only the fractional part of the number, 0.63 foot. In Table 1 1ocate the line with “0.6” appearing in the left column. (This is the 7th line down in the body of the table.) Then while staying on that line, move over to the column headed “0.03.” Note here that the sum of the column and line heads, 0.6 + 0.03, equals the value of 0.63 that we want to convert. In the body of the table for this column and line we read the equivalent fraction for 0.63 foot, $7^{9}/_{16}$ inches. To that value, add the number of whole feet from the value being converted, 11 in this case. The total length equivalent of 11.63 feet is thus 11 feet $7^{9}/_{16}$ inches.

Similarly, Table 2 may be used to make the conversion from inches and fractions to decimal fractions of a foot. This table is convenient for using measured distances in equations. For example, say we wish to convert a length of 19 feet $7^{3}/_{4}$ inches to a decimal fraction. Considering only the fractional part of this value, $7^{3}/_{4}$ inches, locate the decimal value on the line identified as “7-” and in the column headed “$^{3}/_{4}$,” where we read 0.646. This decimal value is equivalent to $7 + ^{3}/_{4} = 7^{3}/_{4}$ inches. To this value add the whole number of feet from the value being converted for the final result, 19 in this case. In this way, 19 feet $7^{3}/_{4}$ inches converts to 19 + 0.646 = 19.646 feet.

Table 2
Conversion, Inches and Fractions to Decimal Feet

	Fractional Increments							
	0	*$^{1}/_{8}$*	*$^{1}/_{4}$*	*$^{3}/_{8}$*	*$^{1}/_{2}$*	*$^{5}/_{8}$*	*$^{3}/_{4}$*	*$^{7}/_{8}$*
0-	0.000	0.010	0.021	0.031	0.042	0.052	0.063	0.073
1-	0.083	0.094	0.104	0.115	0.125	0.135	0.146	0.156
2-	0.167	0.177	0.188	0.198	0.208	0.219	0.229	0.240
3-	0.250	0.260	0.271	0.281	0.292	0.302	0.313	0.323
4-	0.333	0.344	0.354	0.365	0.375	0.385	0.396	0.406
5-	0.417	0.427	0.438	0.448	0.458	0.469	0.479	0.490
6-	0.500	0.510	0.521	0.531	0.542	0.552	0.563	0.573
7-	0.583	0.594	0.604	0.615	0.625	0.635	0.646	0.656
8-	0.667	0.677	0.688	0.698	0.708	0.719	0.729	0.740
9-	0.750	0.760	0.771	0.781	0.792	0.802	0.813	0.823
10-	0.833	0.844	0.854	0.865	0.875	0.885	0.896	0.906
11-	0.917	0.927	0.938	0.948	0.958	0.969	0.979	0.990

Table 1
Conversion, Decimal Feet to Inches (Nearest 16th)

	Decimal Increments									
	0.00	*0.01*	*0.02*	*0.03*	*0.04*	*0.05*	*0.06*	*0.07*	*0.08*	*0.09*
0.0	0-0	$0^{1}/_{8}$	$0^{1}/_{4}$	$0^{3}/_{8}$	$0^{1}/_{2}$	$0^{5}/_{8}$	$0^{3}/_{4}$	$0^{13}/_{16}$	$0^{15}/_{16}$	$1^{1}/_{16}$
0.1	$1^{3}/_{16}$	$1^{5}/_{16}$	$1^{7}/_{16}$	$1^{9}/_{16}$	$1^{11}/_{16}$	$1^{13}/_{16}$	$1^{15}/_{16}$	$2^{1}/_{16}$	$2^{3}/_{16}$	$2^{1}/_{4}$
0.2	$2^{3}/_{8}$	$2^{1}/_{2}$	$2^{5}/_{8}$	$2^{3}/_{4}$	$2^{7}/_{8}$	3-0	$3^{1}/_{8}$	$3^{1}/_{4}$	$3^{3}/_{8}$	$3^{1}/_{2}$
0.3	$3^{5}/_{8}$	$3^{3}/_{4}$	$3^{13}/_{16}$	$3^{15}/_{16}$	$4^{1}/_{16}$	$4^{3}/_{16}$	$4^{5}/_{16}$	$4^{7}/_{16}$	$4^{9}/_{16}$	$4^{11}/_{16}$
0.4	$4^{13}/_{16}$	$4^{15}/_{16}$	$5^{1}/_{16}$	$5^{3}/_{16}$	$5^{1}/_{4}$	$5^{3}/_{8}$	$5^{1}/_{2}$	$5^{5}/_{8}$	$5^{3}/_{4}$	$5^{7}/_{8}$
0.5	6-0	$6^{1}/_{8}$	$6^{1}/_{4}$	$6^{3}/_{8}$	$6^{1}/_{2}$	$6^{5}/_{8}$	$6^{3}/_{4}$	$6^{13}/_{16}$	$6^{15}/_{16}$	$7^{1}/_{16}$
0.6	$7^{3}/_{16}$	$7^{5}/_{16}$	$7^{7}/_{16}$	$7^{9}/_{16}$	$7^{11}/_{16}$	$7^{13}/_{16}$	$7^{15}/_{16}$	$8^{1}/_{16}$	$8^{3}/_{16}$	$8^{1}/_{4}$
0.7	$8^{3}/_{8}$	$8^{1}/_{2}$	$8^{5}/_{8}$	$8^{3}/_{4}$	$8^{7}/_{8}$	9-0	$9^{1}/_{8}$	$9^{1}/_{4}$	$9^{3}/_{8}$	$9^{1}/_{2}$
0.8	$9^{5}/_{8}$	$9^{3}/_{4}$	$9^{13}/_{16}$	$9^{15}/_{16}$	$10^{1}/_{16}$	$10^{3}/_{16}$	$10^{5}/_{16}$	$10^{7}/_{16}$	$10^{9}/_{16}$	$10^{11}/_{16}$
0.9	$10^{13}/_{16}$	$10^{15}/_{16}$	$11^{1}/_{16}$	$11^{3}/_{16}$	$11^{1}/_{4}$	$11^{3}/_{8}$	$11^{1}/_{2}$	$11^{5}/_{8}$	$11^{3}/_{4}$	$11^{7}/_{8}$

Metric Equivalents

Throughout this book, distances and dimensions are usually expressed in English units—the mile, the foot, and the inch. Conversions to metric units may be made by using the following equations:

km = mi × 1.609

m = ft (′) × 0.3048

mm = in. (″) × 25.4

An inch is $^1/_{12}$ of a foot. Tables in the previous section provide information for accurately converting inches and fractions to decimal feet, and vice versa, without the need for a calculator.

Gain Reference

Throughout this book, gain is referenced to an isotropic radiator (dBi) or to an isotropic radiator with circular polarization (dBic).

Advertisers Index

Advertising Department Staff
Debra Jahnke, K1DAJ, *Business Services Manager*
Janet Rocco, W1JLR, *Business Services Associate*
Lisa Tardette, KB1MOI, *Business Services Associate*
Diane Szlachetka, KB1OKV, *Advertising Graphic Design*

1-800-243-7768

Direct Line: 860-594-0207 ▪ Fax: 860-594-4285 ▪ e-mail: ads@arrl.org ▪ Web: www.arrl.org/ads

Advanced Specialties AB-18
Array Solutions.. AB-13
Cable X-PERTS, Inc. AB-6
Dave's Hobby Shop AB12
EZ Hang AB-17
Gap Antenna Products, Inc. AB-11
Ham Radio Outlet .... COVER II, COVER III
Hamware.de AB-18
High Sierra Antennas AB-7
Hy-Gain AB-4, AB-5
IIX Equipment, Ltd. AB-12
K0XG Systems AB-19
MFJ Enterprises AB-2, AB-3
NCG Company AB-15
Quicksilver Radio Products AB-12
Rohn Products/ Div.of Radian
Communication Services, Inc. AB-16
TGM Communications AB-21
Teri Software AB-19
Texas Towers AB-22
Timewave Technology, Inc. AB-8
Unadilla Antenna Mfg. Co. AB-17
Universal Radio AB-14
W3FF Antennas AB-10
WB0W, Inc. AB-9

If your company provides products or services of interest to our readers,
please contact the ARRL Advertising Department today for information on building your business.

MFJ TUNERS

MFJ *IntelliTuner*™ Automatic Tuners

MFJ-993B $\$259^{95}$

The MFJ-993B *IntelliTuner*™ lets you tune any antenna *automatically -- ultra fast.*

It's a comprehensive automatic *antenna tuning center* complete with SWR/Wattmeter, antenna switch for two antennas and 4:1 current balun for balanced lines.

MFJ's exclusive *IntelliTuner*™, *Adaptive Search*™ and *Instant Recall*™ algorithms give you *ultra fast* automatic tuning with over 20,000 *VirtualAntenna*™ Memories.

You get a highly efficient L-network, 6-1600 ohm matching at 300 Watts SSB/CW or *extra-wide* 6-3200 Ohm matching at 150 Watts SSB/CW, 1.8-30 MHz coverage, Cross-Needle *and* digital meters, audio SWR meter, *backlit* LCD, remote control port, radio interface, heavy-duty 16 amp/1000V relays.

The MFJ-993B automatically tunes for minimum SWR and remembers your frequency and tuner settings. The next time you operate on that frequency and antenna, these tuner settings are instantly restored and *you're ready to operate in milliseconds!* 10W x2¾ Hx9D". Use 12-15 VDC/1 amp or 110 VAC with **MFJ-1316, $21.95**. Radio interface cables, remote control available. See ***www.mfjenterprises.com***

600 Watt MFJ Automatic Tuner

MFJ-994B $\$359^{95}$

Like MFJ-993B but handles 600 Watts SSB /CW, matches 12-800 Ohms. 10,000 memories. Does not have LCD display, antenna switch, 4:1 current balun, audio SWR meter/feedback. 10Wx2¾Hx9D in.

***No Matter What*™ Warranty**

Every MFJ tuner is protected by MFJ's famous one year *No Matter What*™ limited warranty. We will repair or replace your MFJ tuner (at our option) for a full year.

More hams use MFJ tuners than all other tuners in the world!

MFJ-9982 *2500 Watt* Tuner

MFJ-9982 $\$699^{95}$

The MFJ-9982 *ContinuousCarrier*™ antenna tuner handles 2500 Watts *continuous carrier* output on all modes and all HF bands into most unbalanced antennas -- even 160 Meters. 6-position antenna switch, 4-core balun, 1.5kW dummy load, true *peak*/average SWR/Wattmeter, 13¾Wx7Hx6¼D".

MFJ-989D *Legal Limit* Tuner

MFJ-989D $\$369^{95}$

New, improved MFJ-989D *legal limit* antenna tuner gives you better efficiency, lower losses and a new *true* peak reading meter. Easily handles *full* 1500 Watts SSB/CW, 1.8-30 MHz, including MARS/WARC bands. Six position antenna switch, dummy load. *New* 500 pF air variable capacitors. *New* improved *AirCore*™ Roller Inductor. *New* high voltage current balun. *New* crank knob. 12⅞Wx6Hx11⅝D".

MFJ-962D *compact* kW Tuner

MFJ-962D $\$279^{95}$

A few more dollars steps you up to a KW tuner for an amp later. Handles 1.5 KW PEP SSB amplifier input power (800W output). Ideal for Ameritron's AL-811H! *AirCore*™ roller inductor, gear-driven turns counter, pk/avg lighted Cross-Needle SWR/Wattmeter, antenna switch, balun, Lexan front, 1.8-30MHz. 10¾Wx4½H x10⅞D in.

MFJ-969 300W *Roller Inductor* Tuner

MFJ-969 $\$199^{95}$

Superb *AirCore*™ Roller Inductor tuning. Covers *6 Meters* thru 160 Meters! 300 Watts PEP SSB. *Active* true peak reading lighted Cross-Needle SWR Wattmeter, *QRM-Free PreTune*™, antenna switch, dummy load, 4:1 balun, Lexan front panel. 10½Wx3½Hx9½D inches.

MFJ-949E *deluxe* 300 Watt Tuner

MFJ-949E $\$169^{95}$

More *hams use MFJ-949s than any other antenna tuner in the world!* Handles 300 Watts. Full 1.8 to 30 MHz coverage, custom inductor switch, 1000 Volt tuning capacitors, *full size* peak/average lighted Cross-Needle SWR/ Wattmeter, 8 position antenna switch, dummy load, *QRM-Free PreTune*™, scratch proof Lexan front panel. 10⅝Wx3½Hx7D inches.

MFJ-948, $139.95. Economy version of MFJ-949E, less dummy load, Lexan front panel.

MFJ-941E *super value* Tuner

MFJ-941E $\$129^{95}$

The *most for your money!* Handles 300 Watts PEP, covers 1.8-30 MHz, *lighted* Cross-Needle SWR/ Wattmeter, 8 position antenna switch, 4:1 balun, 1000 volt capacitors, Lexan front panel. Sleek 10½Wx2½Hx7D".

MFJ-945E HF/6M *mobile* Tuner

MFJ-945E $\$119^{95}$

Extends your mobile antenna bandwidth so you don't have to stop, go outside and adjust your antenna. Tiny 8Wx2Hx6D in. *Lighted* Cross-Needle SWR/Wattmeter. Lamp and bypass switches. Covers 1.8-30 MHz *and* 6 Meters. 300 Watts PEP. **MFJ-20, $6.95,** mobile mount.

MFJ-974HB *Balanced Line* Tuner

MFJ-974HB $\$199^{95}$

The MFJ-974HB *true* fully balanced antenna tuner tunes any balanced lines. Matches 12-2000 Ohms. Covers 1.8-54 MHz *continuously* including all WARC bands. 300 Watts SSB/150 Watts CW. Lighted Cross-Needle SWR/Wattmeter. 7½Wx6Hx8D in.

MFJ-976, $499.95. 1500 Watt *fully balanced* antenna tuner. 1-30 MHz.

MFJ-971 *portable/QRP* Tuner

MFJ-971 $\$109^{95}$

Tunes coax, balanced lines, random wire 1.8-30 MHz. Cross-Needle Meter. SWR, 30/300 or 6 Watt QRP ranges. Matches popular MFJ transceivers. 6Wx6½Hx2½D in.

MFJ-902 *Tiny Travel* Tuner

MFJ-902 $\$89^{95}$

Tiny 4½Wx2¼Hx3D", full 150 Watts, 80-10 Meters, has tuner bypass switch, for coax/random wire.

MFJ-904H, $139.95. Same but adds Cross-needle SWR/Wattmeter and 4:1 balun for balanced lines. 7¼Wx2¼Hx2¾D in.

MFJ-16010 *random wire* Tuner

MFJ-16010 $\$59^{95}$

Operate all bands anywhere with MFJ's reversible L-network. Turns random wire into powerful transmitting antenna. 1.8-30 MHz. 200 Watts PEP. Tiny 2Wx3Hx4D inches.

Dealer/Catalog/Manuals

Visit: http://www.mfjenterprises.com
or call toll-free 800-647-1800

• 1 Year *No Matter What*™ warranty • 30 day money back guarantee (less s/h) on orders direct from MFJ

MFJ
MFJ ENTERPRISES, INC.
300 Industrial Pk Rd, Starkville,
MS 39759 **PH:** (662) 323-5869
Tech Help: (662) 323-0549
FAX:(662)323-6551 8-4:30 CST, Mon.-Fri. ***Add shipping.***

Prices and specifications subject to change. (c) *2007 MFJ Enterprises, Inc.*

MFJ-259B *Antenna Analyzer*

The world's most popular antenna analyzer gives you a complete picture of your antenna performance 1.8 to 170 MHz. Super easy-to-use -- makes tuning your antenna quick and easy. Read antenna SWR, complex impedance, return loss, reflection coefficient. Determine velocity factor, coax cable loss in dB, length of coax and distance to a short or open in feet. Read inductance in uH and capacitance in pF at RF frequencies. Large easy-to-read two line LCD screen and side-by-side meters clearly display your information. Built-in frequency counter, Ni-Cad charger circuit, battery saver, low battery warning and smooth reduction drive tuning.

MFJ-259B
$279.95

MFJ-99B, $88.80. Accessory Package for MFJ-259B. Includes MFJ-29C custom foam-filled carrying case, MFJ-66 dip meter adapters, MFJ-1312D 110VAC adapter and 10-pack high capacity Ni-MH MFJ *SuperCell*™ batteries. ***Save $7!***

MFJ
POWER
97.158 MHz 1.2
R= 39 X= 0 SWR
MFJ HF/VHF SWR ANALYZER
MODEL MFJ-259B
SWR
IMPEDANCE
GATE
MODE
TUNE
FREQUENCY MHz

1.8-170 MHz *plus* 415-470 MHz *Antenna Analyzer*

All-in-one handheld antenna test lab lets you quickly check/tune HF, VHF, ***UHF*** antennas anywhere. Measures: SWR, Return Loss, Reflection Coefficient, R, X, Z, Phase Angle, Coax cable loss, Coax cable length, Distance to short/open in coax, Inductance, Capacitance, Resonant Frequency, Bandwidth, Q, Velocity Factor, Attenuation, more!

MFJ-269
$379.95

MFJ *Weather-Proof* Antenna *Feedthrough* Panel

Bring three coax-fed HF/VHF/UHF antennas, balanced line, random wire and ground into your hamshack without drilling through walls . . . ***New!*** MFJ-4602 $59.95

MFJ's Weather-proof *Antenna Feedthrough Panel* mounts in your window sill. Lets you feed three coax-fed antennas, balanced line, random wire and ground without drilling through walls.

Simply place in window sill and close window. One cut customizes it for any window up to 48 inches. Use horizontally or vertically. High-quality pressure-treated wood with excellent 3/4 inch thick insulating properties is painted with heavy coat of white outdoor enamel paint. Edges sealed by weather-stripping. Seals and insulates against all weather conditions. Gives years of trouble-free service. ¾Dx3½Hx48W in.

Inside/outside stainless steel plates bond all coax shields to ground. Stainless steel ground post brings outside ground connection inside. 3 *Teflon*(R) SO-239 coax connectors, *ceramic* balanced line/randomwire feedthru insulators.

Indoor View

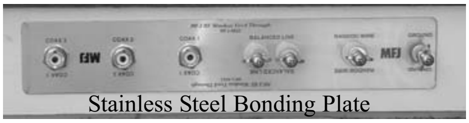

Stainless Steel Bonding Plate

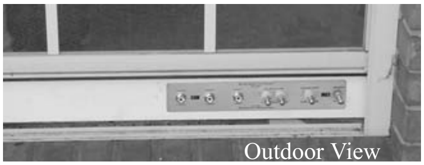
Outdoor View

Current Balun/Center Insulator

MFJ-918
$24.95

True *1:1 Current Balun/Center Insulator* forces equal antenna currents in dipoles. *Improves all dipoles/ doublets!* Reduces coax feedline radiation, field pattern distortion, TVI, RFI, RF hot spots. SO-239. 1.8-30 MHz, 1.5kW. Stainless steel hardware. *Direct* wire connection to antenna. 5x2 inches.

MFJ-915 RF Isolator

MFJ
MFJ-915
RF Isolator

MFJ-915
$29.95

MFJ-915 *RF Isolator* prevents unwanted RF from traveling on the outside of your coax shield into your transceiver. Clears up RFI. Eliminates painful RF "bites" on your rig or microphone. Keeps your transceiver and vehicle displays and settings from locking up. Plugs between coax and transceiver. 5x2 inches. 1.8-30 MHz, 1.5 kW.

MFJ W9INN Balun Box

W9INN BALUN BOX
MODEL MFJ-912
TRANSMITTER
GND
MFJ

MFJ-912
$59.95

Converts unbalanced coax line to balanced ladder transmission line. *Giant* 2-core 4:1 balun wound with *Teflon*(R) wire. 3¼x2¼x7 inches. Heavy duty ceramic feed-through insulators. 1500 Watts full legal limit. SO-239. Ground post, mounting holes.

Antenna Switches

MFJ-1702C 2 position antenna coax switch. Exclusive lightning arrestor and center pos. ground, lightning arrestor. 2.5 kW PEP, 1 kW CW. Loss below .2 dB. 50 dB isolation @ 450 MHz. 50 Ohms. 3x2x3". **MFJ-1704.** Like MFJ-1702C but has 4 positions. 6¼x4¼ x1¼ in. 50 load. SWR 1.1: 1 to 30 MHz; 1.5:1, 30-650 MHz. SO-239.

MFJ-1702C
$34.95

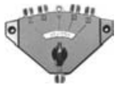

MFJ-1704
$74.95

All Band G5RV Antennas

160-10 Meters with tuner. 102 foot.1.5 kW. Use as Marconi on 160 Meters with tuner/ ground. Ladder line/SO-239. **MFJ-1778M, $34.95.** *Half-size*, 52 ft. G5RV-Jr. 60-10 Meters with tuner. 1.5 kW.

MFJ-1778
$39.95

33 Foot *fiberglass* Mast

MFJ-1910
$79.95

33 foot super strong fiberglass *telescoping* mast collapses to 3.8 feet, weighs 3.3 lbs. Huge, strong 1¾ inch bottom section. Flexes to resist breaking. Resists UV. Put up *full size* inverted Vee dipole/vertical antenna in minutes and get *full size performance!*

6-Band *Rotatable* mini-Dipole

MFJ-1775
$239.95

Low-profile mini 14 foot rotatable dipole has tiny 7-foot turning radius. Fits on smallest roof -- perfect for town houses, apartments and condos.

Easily turned by TV rotator.

It's no Wimp! Good *directivity* reduces QRM/noise and lets you *focus* your signal in the direction that you want -- so you can work some *real* DX.

Operate 40, 20, 15, 10, 6, 2 meters. Run *full 1500 Watts* SSB/CW on all HF bands!

Ham radio's most versatile rotatable dipole features automatic band switching. Uses highly efficient end-loading with its *entire length* always radiating.

No lossy traps! Has separate, efficient end-loading *Teflon*™ coils on fiberglass forms and capacitance hats for each HF band. 6/2 Meters are *full-length* dipoles.

Built-to-last -- super strong solid fiberglass center insulator, 6063 T-6 aircraft strength aluminum tubing. Assembles in an afternoon.

MFJ-1775W, $239.95. WARC band version for 12, 17, 30, 60 Meters only.

MFJ Field Strength Meters

Huge 3" meter. Telescoping ground-independent *dipole.* .1-500 MHz. Sensitivity control. Remote sensor, MFJ-802R, $24.95. **MFJ-801, $24.95.** *Monopole* field strength meter, 1¼" meter.

MFJ-802
$44.95

Clamp-On RF Ammeter

MFJ-854
$109.95

Clamp-On RF Ammeter snaps over cables up to ½ inch diameter. Measure RF currents in antenna elements, radials, ground wires, coax shield. Five *calibrated* ranges to 3 Amps. 1-30 MHz. 2½Wx5½Hx1½D inches.

Free MFJ Catalog

and Nearest Dealer . . . 800-647-1800

http://www.mfjenterprises.com

• 1 Year *No Matter What*™ warranty • 30 day money back guarantee (less s/h) on orders direct from MFJ

MFJ
MFJ ENTERPRISES, INC.
300 Industrial Pk Rd, Starkville,
MS 39759 **PH:** (662) 323-5869
Tech Help: (662) 323-0549
FAX:(662)323-6551 8-4:30 CST, Mon.-Fri. ***Add shipping.***

Prices and specifications subject to change. (c) 2007 MFJ Enterprises, Inc.

hy-gain® HF VERTICALS

Self-supporting -- no guys required . . . Remarkable DX performance -- low angle radiation, omnidirectional . . . Handles 1500 Watts . . . Low SWR . . . Automatic band switching . . . Aircraft quality aluminum tubing . . . Stainless steel hardware . . . Recessed SO-239 connector . . . Two year limited Warranty . . .

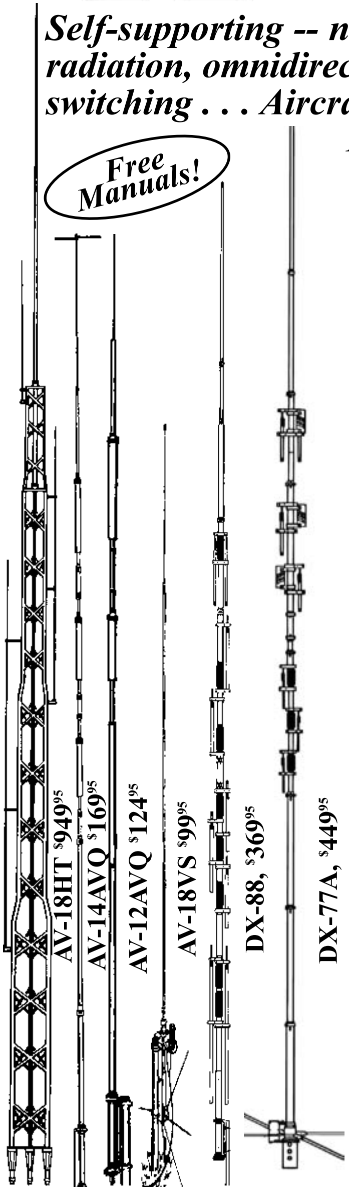

hy-gain(R)
Classics

All hy-gain multi-band vertical antennas are entirely self supporting -- no guys required.

They offer remarkable DX performance with their extremely low angle of radiation and omni-directional pattern.

All handle 1500 Watts PEP SSB, have low SWR, automatic bandswitching (except AV-18VS) and include a 12-inch heavy duty mast support bracket (except AV-18HT).

Heavy duty, slotted, tapered swaged, aircraft quality aluminum tubing with full circumference compression clamps is used for radiators.

Includes all stainless steel hardware. Recessed SO-239 prevents moisture damage.

Hy-gain verticals go up easily with just hand tools and their cost is surprisingly low.

Two year limited warranty.

AV-18HT, $949.95. (10,12,15,20,40,80 M, 160, 17 Meters optional). 53 ft., 114 lbs.

Standing 53 feet tall, the famous *Hy-Gain HyTower* is the world's best performing vertical! The AV-18HT features automatic band selection achieved through a unique stub-decoupling system which effectively isolates various sections of the antenna so that an electrical 1/4 wavelength (or odd multiple of a 1/4 wavelength) exists on all bands. Approximately 250 kHz bandwidth at 2:1 VSWR on 80 Meters. The addition of a base loading coil (**LC-160Q, $109.95**), provides exceptional 160 Meter performance. **MK-17, $89.95**. Add-on 17 Meter kit. 24 foot tower is all rugged, hot-dip galvanized steel and all hardware is iridited for corrosion resistance. Special tilt-over hinged base for easy raising & lowering.

AV-14AVQ, $169.95. (10,15,20,40 Meters). 18 ft., 9 lbs. The Hy-Gain AV-14AVQ uses the same trap design as the famous Hy-Gain Thunderbird beams. Three separate air dielectric Hy-Q traps with oversize coils give superb stability and 1/4 wave resonance on all bands. Roof mount with Hy-Gain AV-14RMQ kit, $89.95.

AV-12AVQ, $124.95. (10, 15, 20 Meters). 13 ft., 9 lbs. AV-12AVQ also uses Thunderbird beam design air dielectric traps for extremely Hy-Q performance. This is the way to go for inexpensive tri-band performance in limited space. Roof mount with AV-14RMQ kit, $89.95.

AV-18VS, $99.95. (10,12,15,17,20,30,40,80 Meters). 18 ft., 4 lbs. High quality construction and low cost make the AV-18VS an exceptional value. Easily tuned to any band by adjusting feed point at the base loading coil. Roof mount with Hy-Gain AV-14RMQ kit, $89.95.

DX-88, $369.95. (10, 12, 15,17,20,30,40,80 Meters, 160 Meters optional). 25 ft., 18 lbs.

All bands are *easily* tuned with the DX-88's *exclusive* adjustable capacitors. 80 and 40 Meters can even be tuned from the ground without having to lower the antenna. Super heavy-duty construction. DX-88 OPTIONS: 160 Meter add-on kit, KIT-160-88, $199.95. Ground Radial System, GRK-88, $99.95. Roof Radial System, RRK-88, $99.95.

DX-77A, $449.95. (10, 12, 15, 17, 20, 30, 40 Meters). 29 ft., 25 lbs.

No ground radials required! Off-center-fed Windom has 55% greater bandwidth than competitive verticals. Heavy-duty tiltable base. Each band independently tunable.

Model #	Price	Bands	Max Power	Height	Weight	Wind Surv.	Rec. Mast
AV-18HT	$949.95	10,15,20,40,80	1500 W PEP	53 feet	114 pounds	75 MPH	------
AV-14AVQ	**$169.95**	**10,15,20,40**	**1500 W PEP**	**18 feet**	**9 pounds**	**80 MPH**	**1.5-1.625"**
AV-12AVQ	$134.95	10/15/20 M	1500 W PEP	13 feet	9 pounds	80 MPH	1.5-1.625"
AV-18VS	**$99.95**	**10 - 80 M**	**1500 W PEP**	**18 feet**	**4 pounds**	**80 MPH**	**1.5-1.625"**
DX-88	$369.95	10 - 40 M	1500 W PEP	25 feet	18 pounds	75 mph no guy	1.5-1.625"
DX-77A	**$449.95**	**10 - 80 M**	**1500 W PEP**	**29 feet**	**25 pounds**	**60 mph no guy**	**1.5-1.625"**

*hy-gain^R^*
PATRIOT

Hy-Gain's new PATRIOT HF verticals are the best built, best performing and best priced multiband verticals available today. For exciting DX make full use of your sunspot cycle with the PATRIOT's *low 17 degree angle* signal.

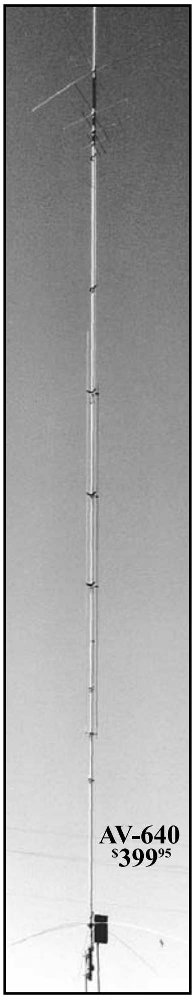

No ground or radials needed

Effective counterpoise replaces radials and ground.

Automatic bandswitching

Single coax cable feed. Each band is individually tunable. Extra wide VSWR bandwidth. End fed with broadband matching unit.

Sleek and low-profile

Low 2.5 sq. ft. wind surface area. Small area required for mounting. Mounts easily on decks, roofs and patios.

Full legal limit

Handles 1500 Watts key down continuous for two minutes.

Built-to-last

High wind survival of 80 mph. Broadband matching unit made from all *Teflon*^R^ insulated wire. Aircraft quality aluminum tubing, stainless steel hardware.

hy-gain^R^ warranty

Two year limited warranty. All replacement parts in stock.

AV-640, $399.95. (6,10,12, 15,17,20,30,40 Meters). 25.5 ft., 17.5 lbs. The AV-640 uses quarter wave stubs on 6, 10, 12 and 17 meters and efficient end loading coil and capacity hats on 15, 20, 30 and 40 meters -- no traps. Resonators are placed in parallel not in series. End loading of the lower HF bands allows efficient operation with a manageable antenna height.

AV-620, $299.95. (6,10,12,15,17,20 Meters). 22.5 ft., 10.5 lbs. The AV-620 covers all bands 6 through 20 Meters with no traps, no coils, no radials yielding an uncompromised signal across all bands.

Free Hy-Gain Catalog
and Nearest Dealer . . . 800-973-6572
Call your dealer for your best price!

hy-gain®
Antennas, Rotators & Towers
308 Industrial Park Road, Starkville, MS 39759 USA
Toll-free Customer Sales Hotline: 800-973-6572
• TECH: 662-323-9538 **• FAX:** 662-323-6551
http://www.hy-gain.com

Prices and specifications subject to change without notice or obligation. (c) Hy-Gain^R^, 2006.

Antennas *built-to-last . . . Use hy-gain parts!*

FAMOUS Heavy Duty *hy-gain* BALUNS

BN-86, $79.95. Broadband 50 Ohm ferrite balun. Useable from 3 to 30 MHz. Recommended for all HF Yagi antennas. It provides improved front-to-back ratio. Comes with clamp to bolt to boom, SO-239 connector.

BN-4000B, $109.95. **B**roadband 50 Ohm Yagi current balun is potted to seal out moisture. Rated at 4000 Watts for increased power handling capabilities. 1 to 54 MHz. SO-239 connector. **BN-4000D, $109.95.** dipoles. **BN-4000N, $119.95.** "N" connector.

hy-gain Center Dipole Insulators

C-1, $29.95. Center Insulator for multiband dipole antenna.

C-1C, $29.95. Similar to C-1, but has SO-239 antenna connector.

E-1, $6.95. 6 inch super high-power, high-voltage *end insulator* for Multi-band doublet antenna. Set of 2.

Arrowpoint Anchor

APA-3, $14.95. Up to 2000 lbs. pull-out resistance to secure antennas, towers. Use one per guy wire. **DR-3, $29.95**. Drive Rod.

Glazed Ceramic Insulator

E-2 **79¢** each

E-26 **$\$4^{56}$** 6-Pack

Extra-strong -- will not break, arc over or melt even under full legal power. Molded ridges give extra-long path to prevent high-voltage breakdown.

Ground Rod

GR-1 **$\$6^{95}$**

Durable four foot by 3/8 inch diameter *copper bonded* steel earth ground rod. Sharp point makes it easy to insert into ground. Bleed off static/RFI. Prevent RF hot spots. Use several to form a good RF ground.

hy-gain Rotator

HAM-IV **$\$559^{95}$**

World's most rugged rotator! Many in use after 25 years. Highly resistant to lightning, static, RFI. Bell design gives total weather protection. 15 sq. ft. windload, 5-sec. brake delay, 5000 in-lb steel wedge brake, ring gear handles 100,000 PSI, dual 98 ball bearing race, -30° F operation. Max mast 2 1/16".

ATM-65 **$\$199^{95}$**

hy-gain Aluminum Tubing

6063 T832 round, drawn aluminum tubing sold in 2 styles: ***6' standard*** or ***6' slitted*** *one end, (.083" gaps, 1-2" slit lengths*).

Tubing are sized to allow telescoping into one another from .625 OD to 2 inches OD. Select slitted tubing for telescoping. Use Hy-gain Stainless Steel hose clamps to secure.

OD x Wall-Th x ID	6'stnrd; p/n, price/ea	6'slitd; p/n, price/ea.	Ad'l/ft.
.625"x.035"x.555"	810-0625C-6U, $5.70	810-0625C-6S, $6.90	$1.00
.750"x.058"x.634"	810-0750E-6U, $10.55	810-0750E-6S, $12.66	$1.85
.875"x.058"x.759"	810-0875E-6U, $8.72	810-0875E-6S, $9.92	$1.53
1.000"x.058"x.884"	810-1000E-6U, $11.46	810-1000E-6S, $12.66	$2.01
1.125"x.058"x1.009"	810-1125E-6U, $10.03	810-1125E-6S, $11.23	$1.76
1.250"x.058"x1.134"	810-1250E-6U, $11.57	810-1250E-6S, $12.77	$2.03
1.375"x.058"x1.259"	810-1375E-6U, $13.17	810-1375E-6S, $14.37	$2.31
1.500"x.058"x1.384"	810-1500E-6U, $17.95	810-1500E-6S, $19.16	$3.15
1.625"x.058"x1.509"	810-1625E-6U, $21.32	810-1625E-6S, $22.52	$3.74
1.750"x.058"x1.634"	810-1750G-6U, $27.86	810-1750G-6S, $29.62	$4.88
2.000'x.120"x1.760"	810-2000I-6U, $38.93	810-2000I-6S, $41.03	$6.83

hy-gain Stainless Steel Hose Clamps

Part No.	Price	Description
745-3106S	$1.00	#6 Clamp for .75" OD to .875" OD tubing
745-3110S	$1.00	#10 Clamp for 1" OD to 1.125" OD tubing
745-3116S	$1.25	#16 Clamp for 1.25" OD to 1.50" OD tubing
745-3128S	$2.84	#28 Clamp for 1.75" OD to 2" OD tubing

← *hy-gain* Telescopic *Aluminum Mast*

ATM-65, $199.95. Pictured left. 11-section *multi-purpose* aluminum telescopic pole can be used as a vertical antenna or as a mast. Great for portable, temporary use and permanent use -- traveling, camping, etc. Includes one size of each slitted tube listed above and ten hose clamps. 65 ft. fully extended, 6 ft. collapsed without clamps. Guying required if mast is extended beyond 30 feet. Top section is .625" OD. Bottom section is 2 inches OD x .120 wall.

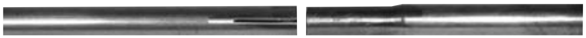

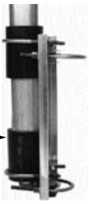

810-2000I-6W, $44.95 each. Unique tubing has one end slitted and one end swaged to make it stackable. Slide one end into another tubing for longer poles. Secure with standard host clamp. Strong, 2"OD x .120" wall. Good for base sections of vertical antennas. Standard 6-ft. length, longer sections, to 20 ft. available. Add $6.83 each add'l foot. **6** ft. lengths ship by UPS, longer units must ship UPS *oversize*/truck.

hy-gain **Base Mount Assembly Kit** →

877200, $69.95. ***New!*** Complete ATM-65 base assembly kit with base mounting plate, insulators, hardware.

hy-gain Custom *Precision* Antenna Hardware

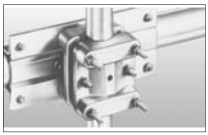

BMC-249, $79.95. ***Boom to Mast Clamp*** Heavy duty 1 1/4" OD thick aluminum die casted boom to mast clamp designed for heavy antennas. Supports 2-inch boom antennas, such as the TH-11DX, Long John Monobanders and other beams with a 2" boom. The BMC-249 allows mast to extend past the top of the clamp. This feature allows stacking of several antennas on one mast. Clamp is adjustable for mast sizes from 1 1/4" to 2 3/4" OD. The tilt-over feature permits easy assembly of the elements and simplifies maintenance of the antenna. 4Wx4Hx1 1/4" thick. Kit comes complete with stainless steel hardware.

Element Insulators These element insulators are made to fit the element bracket (right). Precision injection molded with black UV resisted ABS plastic.

Part No.	Price	Fits these Element sizes	Insulator Size	Fits Element to boom clamp
465216	$3.42	7/16" OD	7/8"OD x 2 3/4"	163409
461057	$4.40	5/8" OD	1 1/4" OD x 3"	165919
463767	$4.40	7/8" OD	1 1/4" OD x 3"	165919
465833	$4.80	1 1/4" OD	1 1/2" OD x 4"	165920

hy-gain TH-series HF Beam Trap Caps

A. 464723, $.75 ea. Trap caps, for large Hy-gain TH-series traps.

B. 466221, $1.59 ea. Trap caps for Hy-gain TH-3Jr. traps.

Element to Boom Clamps

Made from high tensile strength aluminum alloy. Custom die is used to stamp these clamps. No miss-drilled holes. They fit perfect everytime. Required 2 clamps per element. Element Insulators available (left).

Part#	Price	Boom OD"	Element OD w/o insulator	Element OD w/insulator	Description	Insulator
161409	$2.71	2"	5/8"	---	2 1/2x7"x.08 Thick, #12	None
163764	$3.80	2"	7/8"	7/16"	2 3/8x7 1/8"x.08 Thick, #4	465216
165919	$5.32	2"	1 1/4"	5/8", 7/8"	3 1/4x7 1/4"x.125 Thick, #13	463767, 461057
165920	$6.28	2"	1 1/2"	1 1/4"	3 1/4x9 1/2"x.125 Thick, #14	465833
160055	$1.22	1 1/4"	7/16"	---	3/4x5"x.063 Thick, #0	None
161422	$3.63	1 1/4"	5/8"	---	2 1/2x6 1/2"x.08 Thick, #11	None
165141	$3.76	1 1/4"	7/8"	7/16"	2 5/16x6 5/8"x.08 Thick, #12	465216
165138	$5.20	1 1/4"	1 1/4"	5/8", 7/8"	2 3/4x6 5/8"x.08 Thick, #1	463767, 461057
872223	$6.60	Stainless steel screws, nuts and washers for 1 set of element-to-boom clamp				

hy-gain Vertical Antenna Hardware

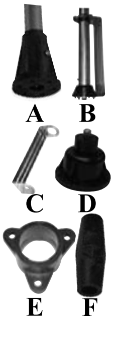

A. 463011, $14.00, Base tubing insulator, fits 1 1/4" OD tube. Six 5/16" mounting holes. Super strong fiberglass filled ABS. Used in Hy-tower to isolate tower legs from ground. Tubing not included.

B. 877130, $28.64, Vertical antenna bracket assembly. Includes bracket (C), base cover insulator (D), support insulator (E) and SO-239. Fits 1 1/4" tubing. Tubing not included.

C. 160043-1, $11.84, Base mount bracket. Custom-tooled 1/8" thick x 12 inch length aluminum bracket with vertical ridges for super strength. Pre-punched holes for SO-239 or N-connector.

D. 863427, $6.90, Base tubing insulator. Molded pin connects base tubing to SO-239 or N-connector.

E. 463056, $3.12, Base antenna support insulator. Made to fit top support of antenna bracket (C). Hole fits 1 1/4" OD tubing.

F. 460316, $12.96, Tubing coupler/insulator. Used to couple/ insulate 1" tube to 7/8" tubing. Made of fiberglass filled ABS, 5" long.

hy-Gain Tower Bases

Specially designed for *hy-gain* crank-up towers. Complete with base hinge and foundation welded steel-cage. Specify tower model. *Shipped by truck, call for quote.

Model#	Tower	Price	Weight
TB-37	HG-37SS	$299.95	50 lbs.*
TB-52	HG-52SS	$299.95	80 lbs.*
TB-54	HG-54HD	$299.95	90 lbs.*
TB-70	HG-70HD	$349.95	102 lbs.*

Free *hy-gain* Catalog

and Nearest Dealer . . . 800-973-6572

Call your dealer for your best price!

hy-gain®

Antennas, Rotators & Towers

308 Industrial Park Road, Starkville, MS 39759 USA
Toll-free Customer Sales Hotline: 800-973-6572
• **TECH:** 662-323-9538 • **FAX:** 662-323-6551

http://www.hy-gain.com

Prices and specifications subject to change without notice or obligation. © Hy-Gain®, 2007.

Andrew Cinta® Cable Assemblies

From

CABLE X-PERTS, INC.
Connecting You to the World...

ANDREW® Cinta CNT-600

CABLE X-PERTS, INC
800 - 828 - 3340
www.cablexperts.com

HALF INCH SIZE
SHOWN

Connector: **N, PL259, TNC & 7/16,** Burial: **Yes**, UV Resistant: **Yes**.
Shields: **2** (100% bonded foil +90% TC Braid) **VP 87%,** Attenuation 3.9dB @ 2 GHz at 100ft.
Usage 450 MHz and Higher.

PL259
Connector

PL259 CONNECTORS EACH END

Part #	Footage	Price
600C125	125 ft	$235.95
600C100	100 ft	$205.95
600C75	75 ft	$175.95
600C50	50 ft	$145.95
600C25	25 ft	$115.95
600C3	3 ft	$ 88.95

PL259(UHF) TO "N" MALE CONNECTOR

Part #	Footage	Price
600CN100	100 ft	$187.95

"N" MALE CONNECTORS EACH END

Part #	Footage	Price
600N100	100 ft	$168.95
600N75	75 ft	$139.95
600N50	50 ft	$108.95
600N25	25 ft	$ 77.95
600N15	15 ft	$ 66.49
600N6	6 ft	$ 55.49

N
Connector

**Or CREATE YOUR OWN
CUSTOM LENGTH ON-LINE.**

ANDREW® Cinta CNT-400

CABLE X-PERTS, INC
1-800-828-3340
WWW.CABLEXPERTS.COM

RG8U SIZE
SHOWN

Connector: **N, PL259, TNC, SMA, BNC & QMA,** Burial: **Yes**, UV Resistant: **Yes**.
Shields: **2** (100% bonded foil +90% TC Braid) **VP 85%,** Attenuation 6.0dB @ 2 GHz at 100ft.
Usage 450 MHz and Higher.

PL259
Connector

PL259 CONNECTORS EACH END

Part #	Footage	Price
400C100	100 ft	$ 82.49
400C75	75 ft	$ 63.95
400C50	50 ft	$ 45.49
400C25	25 ft	$ 27.49
400C6	6 ft	$ 14.95
400C3	3 ft	$ 11.95

"N" MALE CONNECTORS EACH END

Part #	Footage	Price
400N50	50 ft	$ 54.95
400N6	6 ft	$ 22.95
400N3	3 ft	$ 20.95

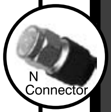
N
Connector

**Or CREATE YOUR OWN
CUSTOM LENGTH ON-LINE.**

ANDREW® Cinta CNT-240

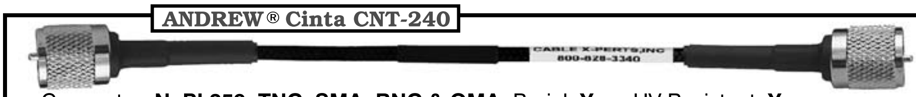

RG8X SIZE
SHOWN

Connector: **N, PL259, TNC, SMA, BNC & QMA,** Burial: **Yes**, UV Resistant: **Yes**.
Shields: **2** (100% bonded foil +90% TC Braid) **VP 84%,** Attenuation 3.0dB @ 150 MHz at 100ft.
Usage 1 MHz and Higher.

PL259
Connector

PL259 CONNECTORS EACH END

Part #	Footage	Price
240C6	6 ft	$ 14.75
240C3	3 ft	$ 12.95

**Or CREATE YOUR OWN
CUSTOM LENGTH ON-LINE.**

Prices subject to change without notice or obligation.

What makes us better than the competition?

- Quality products at competitive prices.
- Prompt service and customer attention.
- State-of-the-art cable assembly manufacturing facility.
- Stringent quality control to ensure top-quality finished products.
- A large inventory of raw materials.
- Short lead times on custom cable assemblies.
- Experienced engineering and knowledgeable sales staff.
- Centrally located in the Chicago area.

All assemblies are tested to ensure optimum performance.

Many more cables listed on-line
www.cablexperts.com
or call: 800-828-3340

CABLE X-PERTS, INC.
Connecting You to the World...

www.cq73.com

High Sierra AntennAs

Professional Antennas for the Amateur Since 1993

HF Transceiver Speaker

HF Antenna Tripod Deluxe

HF Antenna
Stake Hole Mount

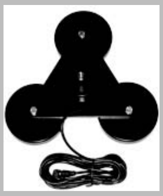

HF Antenna
Magnet Mount

HF, WX, AM, FM, TV
Solar, Battery, AC
Crank Powered
Radio

Explore *All of HF* with Your Sidekick Antenna

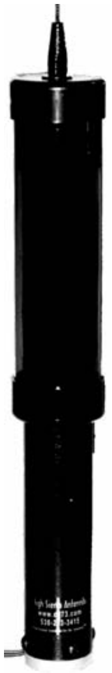

Motor Driven Remotely Tunes 75 to 6 Meters
160 Meters with an 8 Foot Whip & Six Shooter Capacity Hat
1.5:1 or Less Typical SWR
Rated at 200 Watts
Includes 3 Foot Stainless Black Whip & Standard Controller
Can Be Used Just About Anywhere!
No Hazardous Beryllium
Weighs Less Than 3 Pounds
EZ-Tune Controllers Available
See Our Website www.cq73.com for Package Deals

High Sierra AntennAs

530-273-3415 www.cq73.com

See Our New Website with Lots of HF Products

TIMEWAVE
TECHNOLOGY INC.

Introducing the... AntennaSmith™!

AntennaSmith™
TIMEWAVE
POWER
LOAD
TUNE

Patent Pending

TZ-900 *Antenna Impedance Analyzer*

2 Sec Sweeps, Sweep Memories, 1 Hz steps, Manual & Computer Control w/software, USB, low power. Rugged Extruded Aluminum Housing - ***Take it up the tower!***

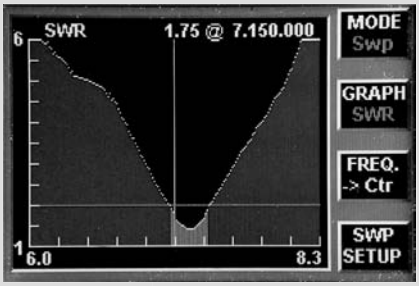
SWR
1.75 @ 7.150.000
6
1
6.0
8.3
MODE Swp
GRAPH SWR
FREQ. -> Ctr
SWP SETUP

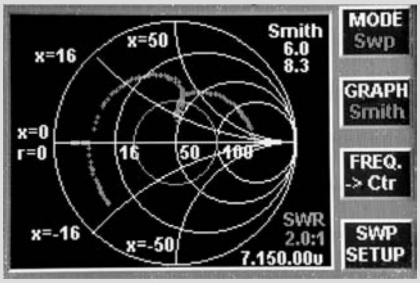
Smith
6.0
8.3
x=16
x=50
x=0
r=0
16
50
100
x=-16
x=-50
SWR
2.0:1
7.150.000
MODE Swp
GRAPH Smith
FREQ. -> Ctr
SWP SETUP

- **Full Color TFT LCD Graphic Display**
- **Visible in Full Sunlight**
- **0.2 - 55 MHz**
- **SWR**
- **Impedance (Z)**
- **Reactance (r+jx)**
- **Reflection coefficient (ρ, θ)**
- **Smith Chart**

NEW!

Rev. 2
Now Shipping!

Check Your Antennas and Transmisson Lines
Once you use the TZ-900 -
you'll never want to use any other!

and... our *new* line-up

Data & Rig Control:

- **HamLinkUSB™ Rig Control Plus**
  TTL serial interface with PTT
- **U232™ RS-232-to-USB Adapter**
  Universal Conversion Module
  replaces PCB-mount DB-9 & DB-25
- **PK-232/USB Multimode Data Controller**
  (upgrades available)
- **PK-96/USB TNC**
  (upgrades available)
- **HamLinkBT-Lite™ Bluetooth Rig Control**
  Coming soon for your radio!
- **HamLinkBT™ Bluetooth Wireless Remote Control Head**
  Coming soon for Kenwood, Icom & Yaesu!

Noise & QRM Control:

- **DSP-599zx Audio Signal Processor**
- **ANC-4 Antenna Noise Canceller**

From the Timewave Fountain of Youth - **Upgrades for many of our DSP & PK products. Call Us Now!**

Timewave Technology Inc
1025 Selby Ave., Suite 101
St. Paul, MN 55104 USA

www.timewave.com
sales@timewave.com

651-489-5080
Fax 651-489-5066

HEIL • GAP • DIAMOND • SPI-RO • HUSTLER • MFJ • COMET

• WEST MOUNTAIN RADIO PRODUCTS • ANDREWS • AMERITRON • CUSHCRAFT • VECTRONICS •

• TARHEEL ANTENNAS • TUBRO TUNER • HY-GAIN • LARSEN • ISOTRON • BELDEN • LMR

WB0W, INC.

800-626-0834

816-364-2692 • Fax 816-364-2619
P.O. Box 8547, St. Joseph, MO 64508

WWW.WB0W.COM

E-mail WB0W@WB0W.COM

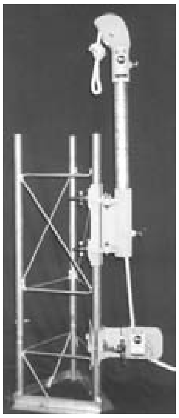

Ginpole for Rohn 25-55 and round leg towers up to 2" O.D. Shown here with optional base pulley. Also ginpoles for BX type Towers.

Tower Insulators for Rohn 25G
$399.00

BX Tower & Accessories

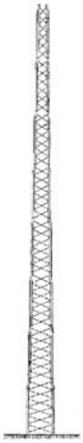

ROHN/Radian Tower & Accessories

25G/45G/55G $99/199/259

KENDOO Batteries

Nifty! Reference Guides
Free Shipping in Continental U.S.

RF Connectors

ARRL Books, Cd's, & Study Guides

25% off if you mention ARRL Antenna Handbook Ad!

TURBO TUNER $239.00
Automatic Motorized
Antenna Controller

WORLD'S SMALLEST MOTORIZED ANTENNA

Model 100
10-80 mtrs.
$349

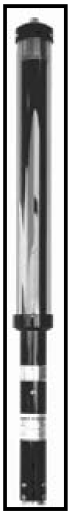

Model 200
6-80 mtrs.
$369

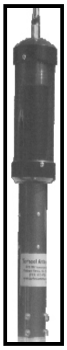

Model 75
"Stubby"
10-80 mtrs.
$349

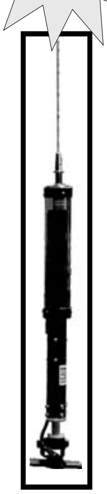

Little
Tarheel II
6-80 mtrs.
$349

TarheelAntennas

Get the quality that only CNC machined blueprinted components in either aircraft aluminum or stainless steel can offer.

ANTENNA SPECIFICATIONS

Little Tarheel II * 3.5 Mhz-54.0 MHz * 200 watts PEP * 1-1/4" coil 16" base * 32" whip * 1.9 lbs * 48" @ 54.0 MHz * 54" @ 3.5 MHz

Model 75 "Stubby" * 3.7 Mhz - 34.0 MHz * 400 watts PEP * 2" coil 16" base * 6' whip * 5.5 lbs * 7' 6" @ 54 MHz * 8' @ 3.5 MHz

Model 100 * 3.4 Mhz - 30.0 * 1.5 Kw PEP * 2" coil * 3' base * 6' whip 8 lbs * 9' 4" tall @ 30.0 MHz * 12' tall @ 3.5 MHz

Model 200 * 3.4 Mhz - 28.0 MHz * 1.5 Kw PEP * 2" coil * 4' base 6' whip * 10' 4" tall @ 28.0 MHz * 12' tall @ 3.5 MHz

Model 300 * 1.8 Mhz - 30.0 MHz * 400 watts PEP * 2" coil * 3' base 6' whip * 8.5 lbs * 9' 4" tall @ 30.0 MHz * 11' tall @ 1.8 MHz

Model 400 * 1.8 Mhz - 28.0 MHz * 400 watts PEP * 2" coil * 4' base 6' whip * 9 lbs * 10' 4" tall @ 28.0 MHz * 12' tall @ 1.8 MHz

All Prices Subject to Change without Notice

TE SYSTEMS AMPS • LDG • ROHN • BX TOWERS • GINPOLES•

BUDDIPOLE

BUDDIPOLE

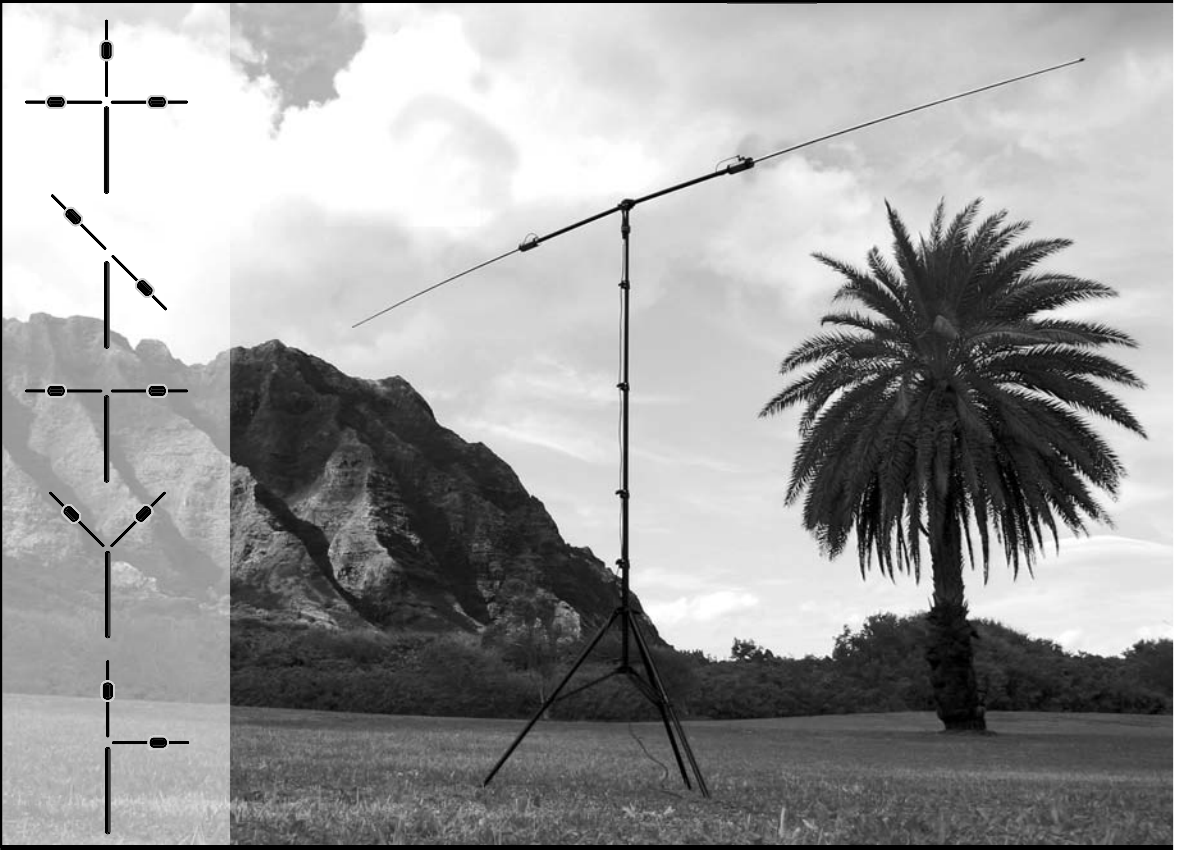

www.buddipole.com

The Buddipole™ Portable Dipole fits in your travel bag and assembles in minutes. The Buddipole is more than an antenna, it's a versatile system for launching your signal. Optimized for transmit power and proven for DX work, the Buddipole is the secret weapon used by HF portable operators all over the world.

What is the Buddipole?

- Multi-band design works 9 bands – 40 meters thru 2 meters with one set of adjustable coils!
- Rated from QRP to 250 watts PEP
- Modular Design – create dozens of different antennas with interchangeable parts
- Rotatatable/Directional
- Lightweight, rugged components
- Optional Rotating Arm Kit allows users to instantly change antenna configurations
- Used by the U.S. Military Special Forces and Emergency Services Groups throughout the world

Antenna Systems starting at $199!

Call or write for free brochure!

2390 Templeton Dr.
Redding, CA 96002

tel: (530) 226 8446
fax: (530) 232 2753
info@buddipole.com

THE PERFECT ANTENNA

For almost 2 decades GAP has stated there is no perfect antenna. No singular antenna can do it all. That is why we designed a family of antennas that utilize the patented (# 5.592.183), GAP Technology, to give you options. Within this family is the antenna or antennas right for you. If deed restrictions or visibility are a problem, take a look at the Eagle DX. If you are looking for a simple general coverage antenna with capture area, consider the Challenger DX. What about a low band antenna with great 40m and 80m performance? Then the Voyager DX is worth your attention. If you seek the most popular multiband antenna, then the Titan DX will end your search.

The GAP Technology eliminates traps, electromechanical devices and coils, which can increase loss and decrease reliability. Simplicity in design, results in a very user friendly assembly process. Slide one tube over another until the big hole lines up with the little and insert a screw. A GAP is pretuned, will handle 1500 watts PEP and provides instantaneous coverage with a VSWR under 2:1 across the entire band.*

CATCH THE WAVE

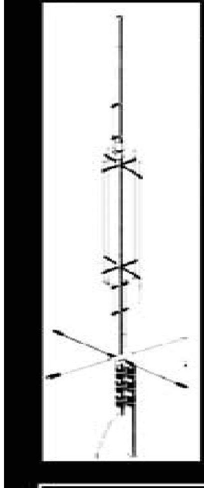

TITAN DX

10m, 12m, 15m, 17m, 20m, 30m, 40m, & 80m

25' tall 25 lbs

Mounts on 1-1/4" OD mast

No radials required.

*VSWR under 2:1 on 80m 100 KHz

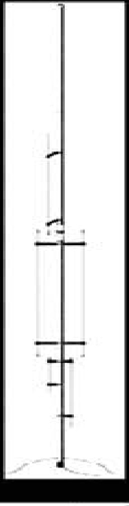

CHALLENGER DX

2m, 6m, 10m, 12m, 15m, 20m, 40m, & 80m

31.5' tall 21 lbs

Drop in ground mount supplied.

3 counterpoises required 25' long.

*VSWR under 2:1 on 80m 130 KHz

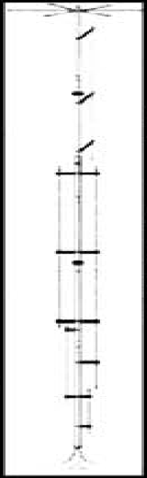

VOYAGER DX

20m, 40m, 80m, & 160m

45' tall 39 lbs

Supllied with hinged ground mount.

3 counterpoises required 57' long.

*VSWR under 2:1 on 160m 90 KHz

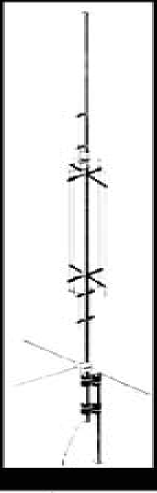

EAGLE DX

10m, 12m, 15m, 17m, 20m & 40m

21' tall 19 lbs

Mounts on 1-1/4" OD mast

No radials required.

*40m limited to 300 watts

QUICK TILT

The easy way to raise and lower your antenna, support mast or flagpole. Works with most antennas.

GROUND MOUNT

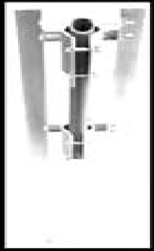

Accomdates
1 1/4" 2" OD Mast

LADDER MOUNT

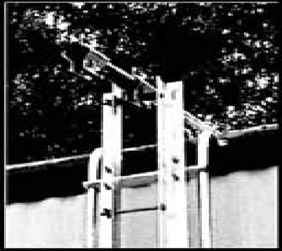

STAINLESS STEEL & ALUMINUM CONSTRUCTION

ANTENNA ACCESSORIES FOR ALL

GAP
ANTENNA PRODUCTS

(772) 571-9922
Visit us at:
gapantenna.com

GUY KITS

Insure your antenna or support mast against the elements with a Guy Kit.

STAINLESS STEEL & ALUMINUM CLAMP
3/32" UV Protected Black Double Braided Polyester Rope

Your Powerpole Headquarters

More Powerpole Stuff Than *Anybody*

A huge variety of ready-made cables with Anderson Powerpoles for ICOM (***including the new IC-7000***), Kenwood (***including the TS-480***), Yaesu, Alinco, and most other HF, mobile, and handheld rigs. Power supply cables for almost any source — home, mobile, or portable. Accessories including panel mounts, mounting plates, crimpers, and wire.

See us at Hamfests from Maine to Maryland, or visit our Web Site for all of your Powerpole needs

Of course, we also stock a full selection of our famous PaneRelief™ coax feed-through panels, and TuneRite™ tuning modules for ICOM, Yaesu, and Kenwood.

Quicksilver Radio Products • Meriden, Conn.

www.qsradio.com

IIX EQUIPMENT LTD.

▪ Custom Fabrication ▪ Hot Dip Galvanizing ▪

MM-Series Mobile Mounts

Rotating Standoff

Standoffs and Gin Poles

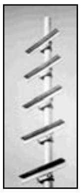

Climbing Steps

Gin Pole KIts

Cell: 708-337-8172
4421 West 87th Street, Hometown, IL 60456
Email: iix@w9iix.com
708-423-0605 ▪ Fax: 708-423-1691
www.w9iix.com

VISA MasterCard DISCOVER PayPal

Electronic Parts

1000s of Component Parts
– No Minimum Order –

– Shortwave – Ham Gear – Scanners –
– AM/FM Radios – Antennas – Coax –
– Accessories – Books – Computer Cables –

Visa/ MC/AMEX/Discover

Dave's Hobby Shop

600 Main Street, Van Buren, Arkansas
479-471-0750
www.daveswebshop.com

Antennas to Build— Ideas to Experiment With

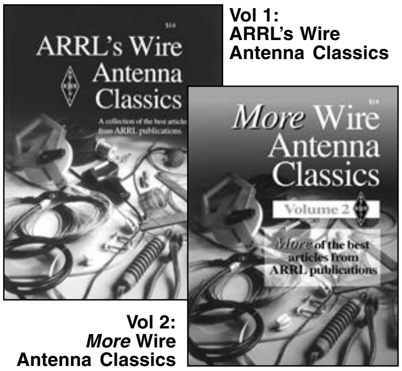

Vol 1: ARRL's Wire Antenna Classics

Vol 2: *More* Wire Antenna Classics

Hang 'em high or hang 'em low, wires are the best antennas to use to get on the air simply and inexpensively. Experience the satisfaction of building your own wire antennas.

Now you can enjoy a collection of the best antenna articles from ARRL publications. Inside these books, you'll find more than just creative ideas. Each is filled with innovative designs that work!

- Dipoles
- Loops
- Collinears
- Wire Beams
- Vertically Polarized
- Multiband Antennas

and more…

– 2 Volumes! –

Vol 1: ARRL's Wire Antenna Classics
ARRL Order No. 7075

Vol 2: *More* Wire Antenna Classics
ARRL Order No. 7709

Available from ARRL and dealers everywhere!

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL ***The national association for AMATEUR RADIO***

225 Main Street
Newington, CT 06111-1494 USA

AB 3/2007

Big Winners from Array Solutions

F V 1.00 1127W
POWER MASTER
PowerMaster HF Coupler
LOAD
SOURCE
COUPLER CALIBRATION
ARRAY SOLUTIONS
www.arraysolutions.com
MADE IN U.S.A.
TO DISPLAY
VSWR
POWER
ALARM
MENU
MODE SELECT

AIM 4170
POWER
ON-OFF
RF
ARRAY SOLUTIONS
RUN

PowerMaster
Watt/VSWR Meter

- Sets the leading edge for all watt/vswr meters to follow
- Unheard of accuracy for the price
- Fast Bright reading meter
- Application software included
- Upgradeable via Internet

$430

AIM 4170
Antenna Analyzer

- Most advanced vector impedance analyzer at a fraction of the cost
- Accurate and easy to use
- Application software included
- Lab instrument quality
- Upgradeable via Internet

$400

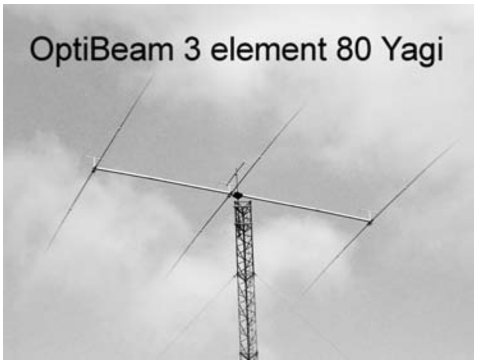
OptiBeam 3 element 80 Yagi

OB3-80+

OptiBeam Antennas
German Engineering means High Performance and Reliability

Prosistel Rotators
Strongest Rotators
on the Market
Prosistel
PST 71 DC

TERMINATION
DIRECTION
1
2
3
4
5
6
7
8
NW
NE
SW
SE
ON
OFF
POWER
ARRAY SOLUTIONS
K9AY Loop System
PREAMP

AS-AYL-4
NEW 4 Direction K9AY Loop Antenna
Hear What You've Been Missing on the Low Bands

www.arraysolutions.com

Phone 972-203-2008
sales@arraysolutions.com
Fax 972-203-8811

We've got your stuff!

WX0B
ARRAY SOLUTIONS

Quality Radio Equipment Since 1942

AMATEUR TRANSCEIVERS

Universal Radio is your authorized dealer for all major amateur radio equipment lines including: Alinco, Icom, Kenwood and Yaesu.

SHORTWAVE RECEIVERS

Whether your budget is \$50, \$500 or \$5000, Universal can recommend a shortwave receiver to meet your needs. Universal is the *only* North American authorized dealer for *all* major shortwave lines including: Elad, Icom, Yaesu, Japan Radio Company, Ten-Tec, Grundig, etón, Sony, Sangean, Palstar, RFspace and Kaito.

WIDEBAND RECEIVERS

Universal Radio carries a broad selection of wideband and scanner receivers. Alinco, AOR, Icom, Yaesu and Uniden models are offered at competitive prices. Portable and tabletop models are available.

ANTENNAS AND ACCESSORIES

Your antenna is just as important as your radio. Universal carries antennas for every need, space requirement and budget. Please visit our website or request our catalog to learn more about antennas.

BOOKS

Whether your interest is amateur, shortwave or collecting; Universal has the very best selection of radio books and study materials.

■ USED EQUIPMENT

Universal carries an extensive selection of used amateur and shortwave equipment. All items have been tested by our service department and carry a 60 day warranty unless stated otherwise. Request our monthly printed used list or visit our website:

www.universal-radio.com

◆ VISIT OUR WEBSITE

Guaranteed lowest prices on the web? Not always. But we ***do*** guarantee that you will find the Universal website to be the **most** informative.

www.universal-radio.com

◆ VISIT OUR SHOWROOM

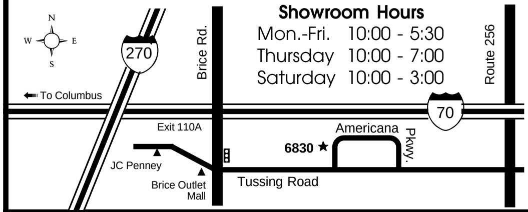

universal radio inc.

FREE 112 PAGE CATALOG

Our informative print catalog covers everything for the amateur, shortwave and scanner enthusiast. All items are also viewable at:

www.universal-radio.com

Universal Radio, Inc.
6830 Americana Pkwy.
Reynoldsburg, OH 43068
➤ 800 431-3939 Orders
➤ 614 866-4267 Information
➤ 614 866-2339 Fax
➤ dx@universal-radio.com

universal radio inc.

- ✓ Established in 1942.
- ✓ We ship worldwide.
- ✓ Visa, Master and Discover cards.
- ✓ Used equipment list available.
- ✓ Returns subject to 15% restocking fee.

The ARRL Antenna Compendium

Volume 7

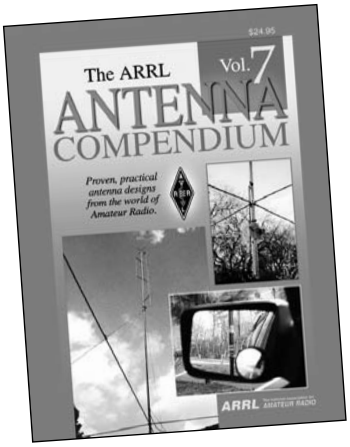

$24.95
The ARRL
Vol. 7
ANTENNA
COMPENDIUM
Proven, practical antenna designs from the world of Amateur Radio.
ARRL

Hooray for antennas! This is the seventh in the very popular **ARRL Antenna Compendium** series. Inside, you'll find articles covering a very wide range of antenna-related topics:

- 30, 40, 80 and 160-Meter Antennas
- Measurements and Computations
- Mobile Antennas
- Multiband Antennas
- Practical Tips
- Propagation and Ground Effects
- Quad Antennas
- Special Antennas
- Stealth Antennas
- Tuners and Transmission Lines
- Vertical Antennas
- VHF/UHF Antennas
- Wire Antennas
- Yagi Antennas

And, this volume includes even more articles on low-band antennas and operating, and great designs for operating on the road-from cars, vans or motor homes.

ARRL Order No. 8608
Only $24.95*
*shipping $8 US (ground)/ $13.00 International

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL *The national association for* ***AMATEUR RADIO***

AB 3/2007

Life is a JOURNEY.
Enjoy the ride!

COMET and Maldol
Base Antennas

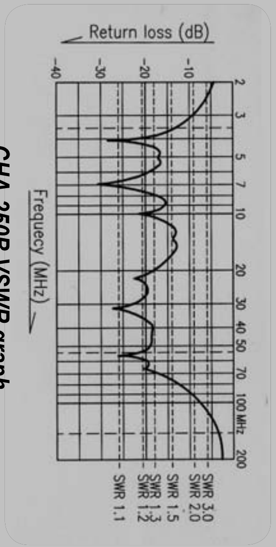

CHA-250B VSWR graph

"One person can effortlessly raise the antenna at night when no one can spot it, and take it down before daybreak. This antenna is also a great choice for portable operations, such as quick and easy mini-DXpedition to a campground or a nice tropical island! In short, the Comet CHA-250B is simple to assemble, painless to elevate and is easy on the eyes, while at the same time getting you on 6 meters thru 80 meters without the requirement of an antenna tuner and ground radials. You'll even be able to work some DX while you're at it!" – Dan Dankert N6PEQ

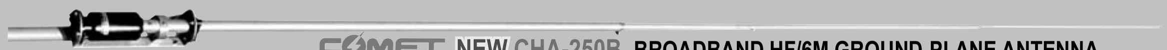

COMET NEW CHA-250B BROADBAND HF/6M GROUND-PLANE ANTENNA
A newly designed broadband vertical with NO GROUND RADIALS. EXTREMELY easy to assemble, requires no tuning or adjustments and VSWRis under 1.5:1 from 3.5-57MHz! • TX: 3.5MHz – 57MHz • RX: 2.0 – 90MHz • VSWR is 1.5:1 or less, continuous • Max Power: 250W SSB/125W FM• Impedance: 50 Ohm • Length: 23' 5" • Weight: 7 lbs. 1 oz. • Conn: SO-239 • Mast Req'd: 1" – 2" dia. • Max wind speed: 67MPH

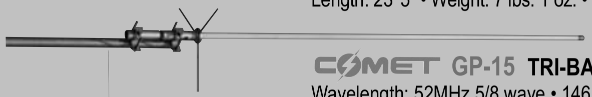

COMET GP-15 TRI-BAND 52/146/446MHZ BASE REPEATER ANTENNA
Wavelength: 52MHz 5/8 wave • 146MHz 5/8 wave x 2 • 446MHz 5/8 wave x 4 • Max Pwr: 150W • Length: 7'11" • Weight: 3lbs. 1oz. • Conn: Gold-plated SO-239• 2MHz band-width after tuning (6M) • Construction: Single-piece fiberglass

COMET CX-333 TRI-BAND 146/220/446MHZ BASE REPEATER ANTENNA
Wavelength: 146MHz 5/8 wave x 2 • 220MHz 5/8 wave x 3• 446MHz 5/8 wave x 5 • Max Pwr: 120W • Length: 10'2" • Weight: 3lbs. 1oz.• Conn: Gold-plated SO-239 • Construction: Fiberglass, 2 Sections

COMET GP-3 DUAL-BAND 146/446MHZ BASE REPEATER ANTENNA
Wavelength: 146MHz 6/8 wave • 446MHz 5/8 wave x 3 • Max Pwr: 200W • Length: 5'11"• Weight: 2lbs. 9ozs. • Conn: Gold-plated SO-239 • Construction: Single-piece fiberglass

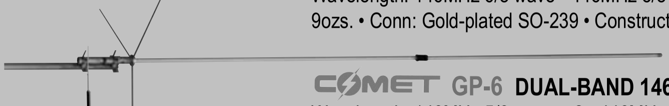

COMET GP-6 DUAL-BAND 146/446MHZ BASE REPEATER ANTENNA
Wavelength: 146MHz 5/8 wave x 2 • 446MHz 5/8 wave x 5 • Max Pwr: 200W • Length: 10'2"• Weight: 3lbs. 8ozs. • Conn: Gold-plated SO-239 • Construction: Fiberglass, 2 Sections

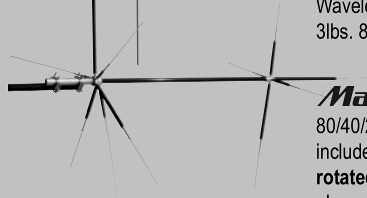

Maldol HVU-8 ULTRA-COMPACT 8 BAND HF/VHF/UHF VERTICAL ANTENNA
80/40/20/15/10/6/2M/70cm Only 1/2 the traditional size and weight of vertical HF antennas, and it includes 2M/70cm! Unique radial system rotates for balcony installations, **the radials can all be rotated to one side.** • Wavelength: HF and 6M: 1/4 wave • 2M: 1/2 wave • 70cm: Two 5/8waves in phase • Impedance: 50 Ohm • Max Power: HF 200W SSB • 6M–70cm: 150W FM• Conn: SO-239 • Height: Only 8'6" • Weight: 5lbs. 7ozs.

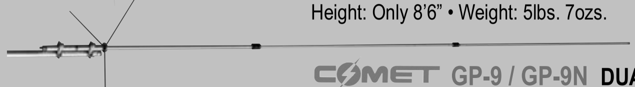

COMET GP-9 / GP-9N DUAL-BAND 146/446MHZ BASE REPEATER ANTENNA
BEST SELLER!• Wavelength: 146MHz 5/8 wave x 3 • 446MHz 5/8 wave x 8 • Max Pwr: 200W• Length: 16' 9"" • Weight: 5lbs. 11ozs. • Conn: GP-9 Gold-plated SO-239 • GP-9N Gold-plated N-type female • Construction: Fiberglass, 3 Sections

WI-FI/802.11B/LAN/HSMM
ACCESS POINT ANTENNAS

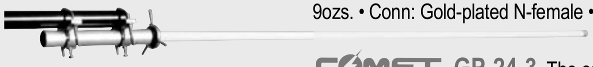

COMET GP-24 2.4GHZ ACCESS POINT ANTENNA The standard high gain 2.4GHz access point antenna for maximum range/signal strength. • 2400-2500MHz • Length: 5'10" • Weight: 2lbs. 9ozs. • Conn: Gold-plated N-female • Construction: Heavy-duty white fiberglass radome

COMET GP-24-3 The same physical high gain 2.4GHz antenna as the GP-24,but with a 3-degree electrical downtilt to concentrate the signal to the horizon and below

COMET GP-24-S MEDIUM GAIN ACCESS POINT ANTENNA For applications where maximum distance is not as important as a solid signal around and directly below the antenna. • Length:3'9" • Weight: 1lb. 14ozs. • Conn: Gold-plated N-female • Construction: Heavy-duty white fiberglass radome

COMET and Maldol Antennas

For a complete catalog, call or visit your local dealer.
Or contact NCG Company.1275 N. Grove Street, Anaheim, CA 92806
714-630-4541 • 800-962-2611 • FAX 714-630-7024 • www.natcommgroup.com

Alphabet soup got you confused?

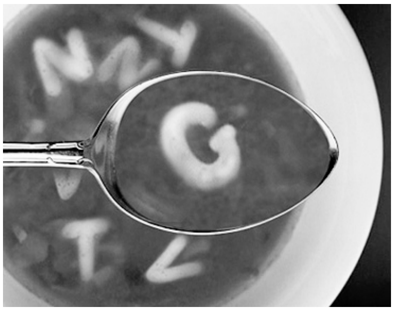

There's only one letter you need . . . ***it's the G*.**

We're the name you've known and trusted for over 56 years!

Our name stands for . . .

Quality

Integrity

Enduring Strength

Choose the original Rohn G Series.
Because everyone knows "G" comes before all the rest!

ROHN™
Products

6718 W. Plank Road ~ Peoria, IL 61604
Ph. 309-697-4400 ~ Fax 309-697-5612
www.radiancorp.com

Division of Radian Communication Services, Inc.

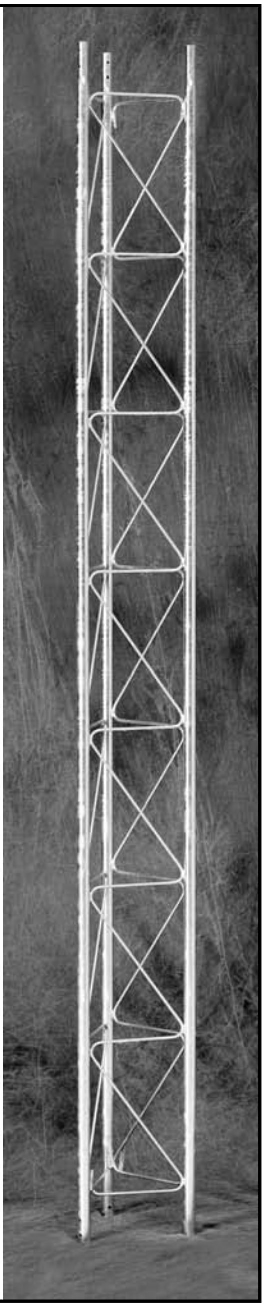

ARRL's Yagi Antenna Classics

—Yagis, Quads, Loops, and other Beam Antennas

Contents:

- **Monobanders:** Beams for your favorite band
- **Multibanders:** Beams that cover two or more bands
- **HF, VHF and UHF Beams:** From 80 meters to 2304 MHz
- **Computer Modeling:** Optimize your beam's performance
- **Towers, Masts and Guys:** Your beam needs solid support
- **The "WOW" Factor:** Can you believe this?

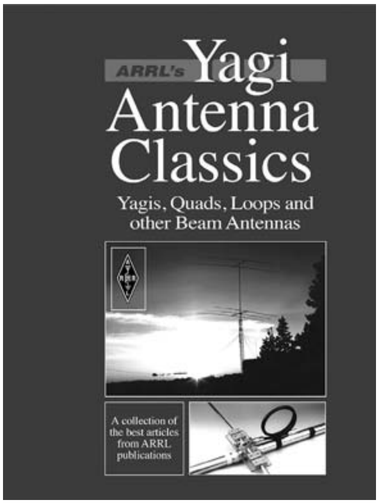

ARRL's Yagi Antenna Classics
Yagis, Quads, Loops and other Beam Antennas

A collection of the best articles from ARRL publications

ARRL *The national association for AMATEUR RADIO*

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL Order No. 8187
Only $17.95*
*shipping $7 US (ground)/ $12.00 International

AB 3/2007

"Being the Original has its advantages."

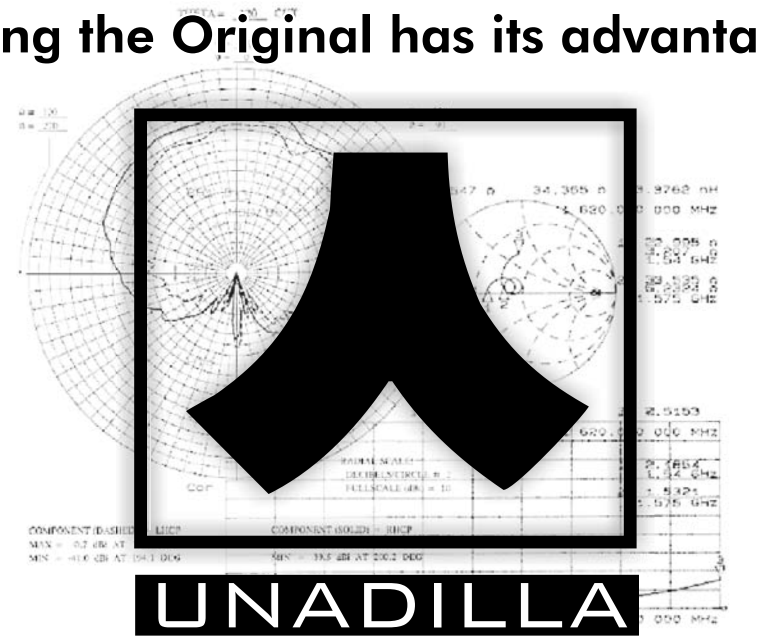

UNADILLA

ANTENNA MANUFACTURING COMPANY

VHF - UHF - HF Applications

Special Models for Beams, Dipoles and Line Isolation

Manufacturing Baluns and Accessories for over 40 years

Ham - Commercial - Industrial - Government - Military

PO BOX 4215-BV
ANDOVER, MA 01810
PLANT 978 975 2711

www.unadilla.com

No Antenna No Ham Radio!

Antenna Zoning for the Radio Amateur
by Fred Hopengarten, K1VR

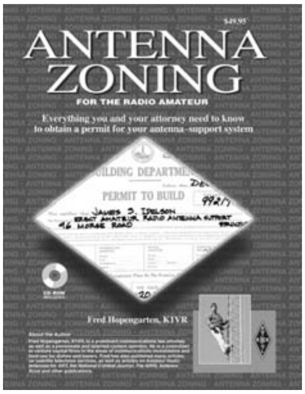

ANTENNA ZONING FOR THE RADIO AMATEUR

Everything you and your attorney need to know to obtain a permit for your antenna-support system

Fred Hopengarten, K1VR

Here's help navigating the thicket of ordinances and bylaws to obtain a building permit for your antenna system. Don't get caught in the crazy quilt of regulations.

Everything you and your attorney need to know to obtain a permit for your antenna-support system

CD-ROM included containing detailed case law, customizable form letters and other invaluable aids.

Providing your attorney with this information can save many hours—and many dollars!

Antenna Zoning for the Radio Amateur
ARRL Order No. 8217
Only $49.95*
*shipping $10 US (ground)/ $15.00 International

ARRL *The national association for AMATEUR RADIO*

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

AB 3/2007

Hang your
next wire antenna with the

EZ HANG KIT!

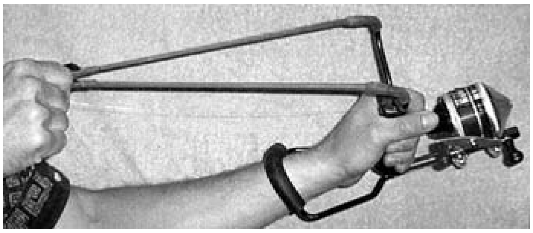

Comes with everything you need:

- The original EZ Hang
- The EZ Winder
- A spare set of Bands and six weights and clips.

EZ HANG

Code AB
2217 Princess Anne St., Suite 105-6
Fredericksburg, VA 22401 ▪ Phone: 540-286-0176
Check Our Website for New Products
www.ezhang.com

High-Performance Remote Automatic Antenna Tuners

Places the Tuner Circuitry at the Antenna Feedpoint for

Maximum Efficiency!

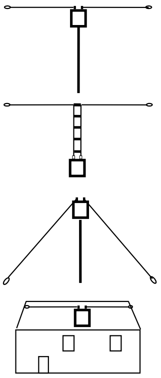

1.8 to 30 MHz Balanced π-Network Antenna Tuners:

- **AT-502** – 200W SSB/CW
  2 x 85 Memories $699.95
- **AT-515** – 1.5kW SSB/CW
  85 Memories $1,699.95

Desktop Control Unit with LCD; Weather-Protected Remote Tuner Unit (to IP65/66 standards).

www.hamware.de

Int'l: Klaus Bemmerer ▪ **service@hamware.de**
Tel: +49 4371 869145
US: Dillon RF Systems ▪ **dillonel@mtaonline.net**

Backyard Antennas

Simple techniques for building high performance antennas

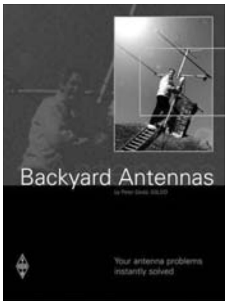

- End-fed and Center-fed Antennas
- Rotary Beams
- Loops
- Tuning Units
- VHF/UHF Antennas
- Antenna and Mast Construction
- Transmission Lines
- Estimating and Measuring the Performance of your Antenna

ARRL *The national association for* **AMATEUR RADIO**

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

Backyard Antennas
ARRL Order No. RBYA
Only $34.95*

*shipping $9 US (ground)/ $14.00 International

AB 3/2007

ADVANCED SPECIALTIES INC.

New Jersey's Communications Store

VX-7R
Quadband
Submersible HT

YAESU *Choice of the World's top DX'ers*

ALINCO *AMATEUR RADIO'S VALUE LEADER*

DR-635T
Hi Tech FM Mobile
Dual Band

Authorized Dealer

ALINCO * LARSEN * COMET * MALDOL * MFJ * UNIDEN
LDG * ANLI * RANGER * YAESU * TRAM * PROCOMM * PYRAMID

FT-857D
HF, 6M, 2M, 70 CM

AMATEUR RADIO – SCANNERS – BOOKS
ANTENNAS – MOUNTS – FILTERS
ACCESSORIES & MORE

FT-2800M
65W 2M Mobile

Uniden
BCD-396T
Digital Trunk
Tracking
Scanner

Closed Sunday & Monday
Orders/Quotes 1-800-926-9HAM
(201)-VHF-2067
114 Essex Street, Lodi, NJ 07644

FT-8900R
10M, 6M, 2M,
70 CM FM Mobile

Big Online Catalog at
www.advancedspecialties.net

Simple and Fun Antennas for Hams

Lots and lots of real world, practical antennas you can BUILD YOURSELF!

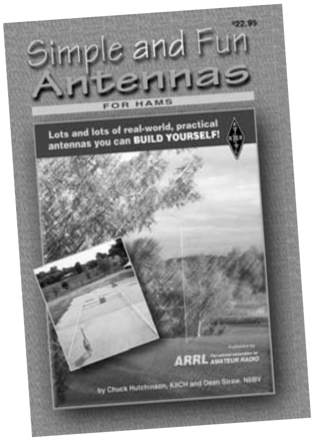
$22.95
Simple and Fun Antennas
FOR HAMS
Lots and lots of real-world, practical antennas you can BUILD YOURSELF!
ARRL
by Chuck Hutchinson, K8CH and Dean Straw, N6BV

Contents:

- Your First VHF Antenna
- Your First HF Antenna
- Facts About Transmission Lines
- Antenna Masts and Supports
- HF Verticals
- More Simple and Fun Antennas for VHF and UHF
- More HF Dipoles
- Dual-Band VHF/UHF Antennas
- An HF Vertical That Needs No Radials—Try the HVD
- Yet More HF Dipoles
- More Facts About Feed Lines
- A Potpourri of Antenna Ideas
- VHF Beam Antennas
- Towers
- HF Beam Antennas
- Getting the Most Out of Your Antenna

Simple and Fun Antennas for Hams
ARRL Order No. 8624
Only $22.95*

*shipping $8 US (ground)/ $13.00 International

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL *The national association for* **AMATEUR RADIO**

AB 3/2007

Vertical Antennas are Everywhere

On cell phones, broadcast towers, portable radios and vehicles. And for good reason! Here are the best articles on vertical antennas from *QST* and ARRL publications.

- Theory and Modeling
- VHF and UHF
- Directional Arrays
- Radials
- Single-Band and Multi-Band HF Vertical Antennas
- Ground Systems
- Portable Antennas
- Phased Vertical Arrays

and more….

The Sky's the Limit!

$12
VERTICAL ANTENNA CLASSICS
The best articles from ARRL publications
More VERTICAL ANTENNA CLASSICS
The best articles from ARRL Publications
ARRL

Vertical Antenna Classics
ARRL Order No. 5218
Only $12*

More Vertical Antenna Classics Second Edition
ARRL Order No. 9795
Only $17.95*

*shipping $7 US (ground)/ $12.00 International

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL *The national association for* ***AMATEUR RADIO***

AB 3/2007

K0XG Systems

Antenna Rotation Equipment

Orbital Ring Rotor

Unit shown is the 45/55g Orbital Ring rotor with MonstIR Yagi installed.

Models available for any size tower (including cell towers). For any size boom, or tower boom.

1117 Highland Park Dr., Bettendorf, IA 52722
email: k0xg@k0xg.com
(563) 340-5111

Visit our Web Site at:
www.K0XG.com

A picture is worth a thousand words...

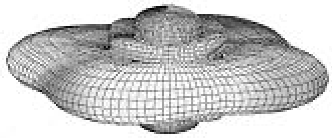

With the

ANTENNA MODEL™

wire antenna analysis program for Windows® you get true 3D far field patterns that are far more informative than conventional 2D patterns or wire-frame pseudo-3D patterns.

Now you can have all the advantages of the MININEC code and Sommerfeld-Norton ground too, down to 0.02 wavelengths above ground. No split load, split source, or equivalent monotaper element approximations are needed. Enjoy superior graphing and 2D far field plot comparison capability.

Describe the antenna to the program in an easy-to-use spreadsheet-style format, and then with one mouse-click the program shows you the antenna pattern, front to back ratio, front to rear ratio, input impedance, efficiency, SWR, and more.

An optional **Symbols** window with formula evaluation capability can do your calculations for you. A **Match Wizard** designs Gamma, T, or Hairpin matches for Yagi antennas. A **Clamp Wizard** calculates the equivalent diameter of Yagi element clamps. **Yagi Optimization** finds Yagi dimensions that satisfy performance objectives you specify. Major antenna properties can be graphed as a function of frequency.

There is **no built-in segment limit**. Your models can be as large and complicated as your system permits.

ANTENNA MODEL is only $90. This includes a Web site download **and** a permanent backup copy on CD-ROM. Visit our Web site for more information about **ANTENNA MODEL**.

Teri Software
P.O. Box 277
Lincoln, TX 78948

www.antennamodel.com
e-mail **sales@antennamodel.com**
phone **979-542-7952**

Fourth Edition

ON4UN's Low-Band DXing

Antennas, Equipment and Techniques for DXcitement on 160, 80 and 40 Meters

by John Devoldere, ON4UN

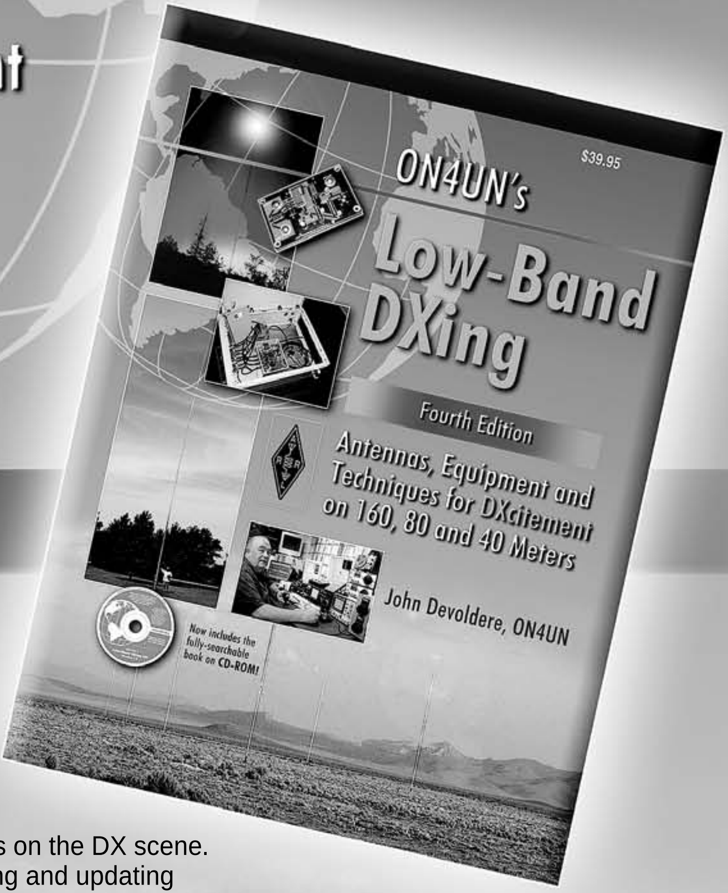
$39.95

ON4UN's

Low-Band DXing

Fourth Edition

Antennas, Equipment and Techniques for DXcitement on 160, 80 and 40 Meters

John Devoldere, ON4UN

Now includes the fully-searchable book on CD-ROM!

Thoroughly updated with entirely new material on low-band antennas and high-gain transmitting arrays. Includes new insights and new design techniques for receiving antennas and vertical arrays.

"John is one of the most recognizable figures on the DX scene. He has spent more than 1,500 hours rewriting and updating **ON4UN's Low-Band DXing**, known far and wide as the 'low-bander's bible.'"—R. Dean Straw, N6BV, Editor

CD-ROM included! This edition is bundled with the fully searchable and complete book on **CD-ROM** for *Windows*® and Macintosh® systems. (Also contains additional **ON4UN software** and over 2000 quality pictures.)

ARRL Order No. 9140—**$39.95***
*shipping: $9 US (ground)/$14 International

Shipping and Handling charges apply. Sales Tax is required for orders shipped to CA, CT, VA, and Canada.

Prices and product availability are subject to change without notice.

ARRL *The national association for AMATEUR RADIO*

SHOP DIRECT *or call for a dealer near you.*
ORDER TOLL-FREE 888/277-5289 (US)
ONLINE WWW.ARRL.ORG/SHOP

AB 3/2007

Basic Radio

Understanding the
Key Building Blocks

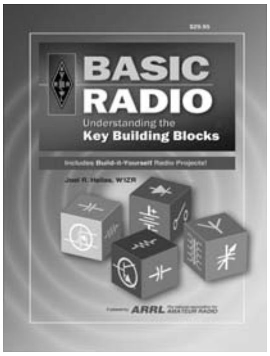
$29.95
BASIC RADIO
Understanding the Key Building Blocks
Includes Build-it-Yourself Radio Projects!
Joel R. Hallas, W1ZR
ARRL

Finally — an introduction to radio FOR EVERYONE! — what it does and how it does it.

Basic Radio reveals the key building blocks of radio: **receivers, transmitters, antennas,** and **propagation.** Includes simple, build-it-yourself projects to turn theory into practice.

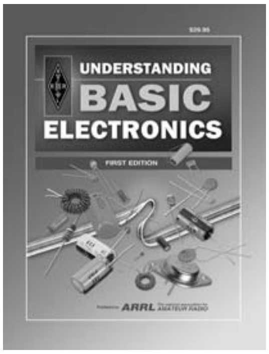
$29.95
UNDERSTANDING BASIC ELECTRONICS
FIRST EDITION
ARRL

Basic Electronics

The foundations taught in this book will prepare you for more advanced concepts in electronics. The text is written in an easy-to-understand style that nontechnical readers will enjoy. You don't have to be an engineer or a math whiz to enjoy the thrill of experimenting with electronics circuits!

Basic Radio
ARRL Order No. 9558
Only $29.95*

Understanding Basic Electronics
ARRL Order No. 3983
Only $29.95*
*shipping $9 US (ground)/ $14.00 International

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

ARRL *The national association for* ***AMATEUR RADIO***

AB 3/2007

HYBRID-QUAD ANTENNA

MQ-24SR

MINI
HF BEAM

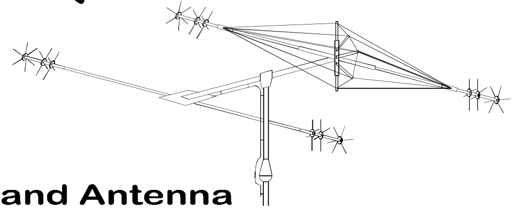

MQ-1 Four-Band Antenna
6,10,15,20 Meters, Two element

MQ-24SR Four-Band Antenna
6,10,15,20 Meters, Enhanced reflector

MQ-34SR Four-Band Antenna
6,10,15,20 Meters, Three element, E/Ref.

MQ-26 Six-Band Antenna
6,10,12,15,17,20 Meters, Two element

MQ-26SR Six-Band Antenna
6,10,12,15,17,20 Meters, Enhanced Refl.

MQ-36SR Six-Band Antenna
6,10,12,15,17,20 Meters,Three element,E/R

T.G.M. Communications

121 Devon St. Stratford,
ON Canada N5A 2Z8
Tel. & Fax (519) 271-5928
www3.sympatico.ca/tgmc

ARRL's VHF/UHF Antenna Classics

Build portable, mobile, and fixed antenna designs

Ground planes, J-poles, mobile antennas, Yagis, and more. Build a better antenna for your hand-held radio. Construct a 2-meter Yagi that will rival similar commercial antennas. Build a dual-band vertical for 146 and 445-MHz. The results will be rewarding! Includes projects gathered from the pages of *QST*.

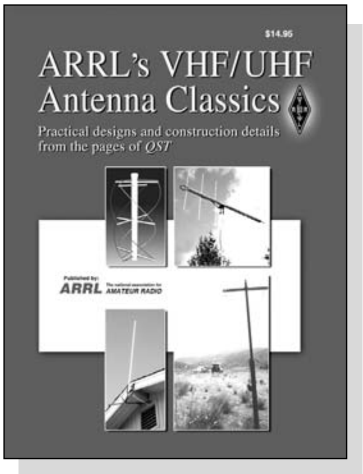
$14.95
ARRL's VHF/UHF Antenna Classics
Practical designs and construction details from the pages of QST
Published by: ARRL The national association for AMATEUR RADIO

ARRL Order No. 9078
Only $14.95*
*shipping $7 US (ground)/$12.00 International

ARRL *The national association for* ***AMATEUR RADIO***

SHOP DIRECT *or call for a dealer near you.*
ONLINE WWW.ARRL.ORG/SHOP
ORDER TOLL-FREE 888/277-5289 (US)

AB 3/2007

MORE DEALS THAN YOU CAN SHAKE A STICK AT!

- Great Gear
- Great Deals
- Great Service
- Free UPS S/H!*

*On all radio orders shipped within the contiguous USA.

TEXAS TOWERS

Savings As Big As Texas!

Alfa-Spid, Alpha-Delta • Ameritron • ARRL • Astron • Bencher • Butternut • Cal-Av • Comet • Cushcraft • Daiwa • Diamond • Force 12 • Gap • Glen Martin • Hustler • Hygain • Icom • Kantronics • Kenwood • Lakeview • Larsen • LDG Electronics • M2 • MFJ • Mirage • Palstar • Phillystran • Rohn • Tex-Com • Times Microwave • Unadilla • Universal Tower • US Tower • Van Gorden • Vectronics • Vibroplex • Yaesu

KENWOOD

KENWOOD TS-2000X

Huge Band Coverage, All Mode HF/6m/2m/70cm/23cm, Auto Tuner, CW Memory Keyer, Dual RX, Dual DSP, Built-in 1200/9600 bps TNC, and More!

CALL FOR YOUR BEST PRICE!

KENWOOD TS-480SAT

Mobile Performance, All Mode HF/6m, Auto Tuner, Separate Front Control Panel, 16-Bit DSP, CTCSS Encode/Decode, More!

CALL FOR YOUR BEST PRICE!

KENWOOD TM-D700A

Mobile 2m/70cm FM XCVR With TNC, Separate Front Control Panel, CTCSS Encode/Decode, 200 Memories, and More!

CALL FOR YOUR BEST PRICE!

butternut

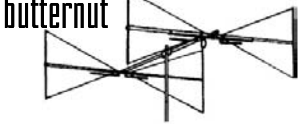

BUTTERNUT HF-5B

Lightweight five band beam, small enough to turn with a TV rotator—perfect for portable or limited space operation or for deed restricted stations.

CALL FOR MORE BUTTERNUT!

cushcraft

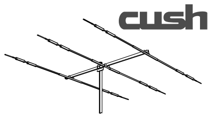

CUSHCRAFT A3S

Three element triband beam with optional 30/40m kit. Compact yet rugged design with all stainless steel hardware. Perfect for limited space application.

CALL FOR MORE CUSHCRAFT!

hy-gain

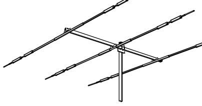

HYGAIN TH-3MK4

Three element triband beam designed for broadband performance. Rugged construction and all stainless steel hardware for long term reliability.

CALL FOR MORE HYGAIN!

M2

M2 KT-34M2

Four element, short boom triband beam is long on performance! Rugged, with all stainless hardware. Broad gain, high F/B, flat match. Balun supplied.

CALL FOR M2 ANTENNAS!

Yaesu G-1000DXA

ANTENNA ROTATORS

Hygain, AR-40	Call
Hygain, CD-45II	Call
Hygain, Ham-IV	Call
Hygain, Ham-V	Call
Hygain, T2X	Call
Hygain, T2X Digital	Call
Hygain, HDR-300A	Call
M2, OR-2800PDX	Call
Yaesu, G-450A	Call
Yaesu G-550	Call
Yaesu, G-800SA	Call
Yaesu, G-800DXA	Call
Yaesu, G-1000DXA	Call
Yaesu, G-2800DXA	Call
Yaesu G-5500	Call

ROTOR CABLE IN STOCK, CALL!

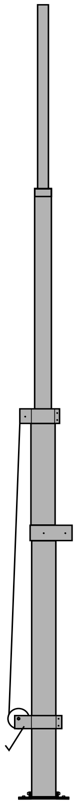

US TOWER CORPORATION

MA SERIES

Neighbor Friendly Design offers a flag-pole-like appearance, while supporting antenna loads up to 22 square feet. Models range from 40–85 feet in height and come with a hand winch and house bracket. Optional self supporting bases are also available.

TOWER EXPERTS

We Ain't Braggin.' But we've helped so many Hams order US Towers over the years that we've become the US Tower experts. Please call for help selecting the perfect US Tower for your QTH!

Universal

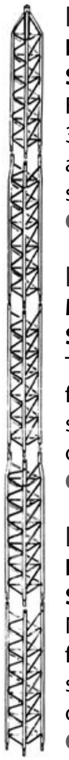

B-18 SERIES

Light Duty Aluminum Self Supporting Towers. Five models range from 30–50 feet in height, and support up to 12 sq. ft. of antenna load.

CALL FOR MORE INFO!

B-26 SERIES

Medium Duty Aluminum Self Supporting Towers. Thirteen models range from 30–90 feet and support up to 34.5 sq. ft. of antenna wind load.

CALL FOR MORE INFO!

B-30 SERIES

Heavy Duty Aluminum Self Supporting Towers. Nineteen models range from 40–100 feet, and support up to 34.5 sq. ft. of antenna wind load.

CALL FOR MORE INFO!

TEXAS TOWERS

TOLL FREE (800) 272–3467

1108 Summit Avenue, #4 • Plano, TX 75074
Hours: M-F 9 AM–5 PM Central Time
Email: sales@texastowers.com

MASTERCARD
VISA • DISCOVER

Proudly Serving Ham Operators Since 1978!
Visit Our Website for More Great Deals:
http://www.texastowers.com

INDEX

A

A-frame mast: 22-4, 22-5
A-Index: 23-45
AA6GL: 6-29, 15-8
AA6ZM: 6-5
AAT: 25-9ff
AB-577 military tower: 22-12
Absorber circuit, power: 27-6
Accuracy tests: 4-15ff
Acute-angle junction: 4-8
Adcock
  Antenna: 14-5ff
  Pattern: 14-6
Adcock, F.: 14-5
Adding parasitic director to LPDA: 10-8
Additional loss due to SWR: 24-10ff
Addresses of suppliers: 21-4ff
Adjustable L-network feed system: 8-19ff
Adler, Dick, K3CXZ: 3-24
Admittance: 28-5
AI1H: 26-22
AL7KK: 7-22
Algorithm for ray tracing: 3-31
Aluminum tubing: 4-7
AM broadcast stations: 2-8
American National Standards Institute: 1-19ff
American Physical Society: 1-19
Ammeters, RF: 27-2
Amplitude: 3-2, 3-12, 4-14
  Reflected signals: 3-2
Andress, Kurt, K7NV: 22-1, 22-27
Angle:
  Azimuth: 2-8, 2-15, 19-1
  Critical: 23-24
  Elevation: 2-8, 2-15, 19-1
  Near overhead: 11-37
  Zenith: 2-9, 2-15
Annealed copper wire: 20-1
ANSI RF exposure guidelines: 1-19
Antenna Model: 4-2
Antenna Modeling course: 4-3
Antenna tuner: 2-2, 4-6, 4-23, 7-3, 7-5, 7-7, 26-1ff
Antennas:
  Addresses of suppliers: 21-4ff
  Analysis by computer: 4-1ff
  Antenna terms, definitions of: 8-2ff
  AO-40: 19-9ff
  Apartment antennas: 4-24ff
  Aperture: 27-48
  Base loading: 16-1

Antennas:
  Broadband: 9-1ff
  Building on the tower: 22-25ff
  Capacitance: 16-3
  Center loading: 16-1
  Center-fed: 7-3
  Clothesline: 4-25
  Co-planar crossed Yagi: 19-5
  Coaxial sleeve: 7-23
  Construction details and practical considerations: 4-31
  Continuously loaded: 16-13ff, 6-29
  Crossed linear: 19-4ff
  Directive: 2-2
  Directivity and gain: 2-6
  Effect: 14-2
  Electrical loading: 16-1
  EME: 19-32ff
  End effect: 4-31
  End-fed: 7-2ff
  Essential characteristics: 2-1ff
  Flattop: 7-3
  Folded monopole: 6-19
  Folded, efficiency: 8-11ff
  for repeaters: 17-1ff
  G5RV & antenna tuner: 7-5
  G5RV multiband: 7-5ff
  Ground plane: 2-17, 4-8ff, 6-18ff, 7-16
  Half-sloper: 6-44
  Half-square: 6-24ff
  Height: 2-17
  Helical: 19-3, 19-5ff
  High horizontal: 6-10
  Horizontal: 6-10
  Image: 3-11
  Indoor: 4-27ff
  Installation: 1-1, 22-23ff
  Inverted L: 6-41ff
  Inverted V: 7-3ff
  Invisible: 4-24ff
  Isolation, for repeaters: 17-7ff
  Isotropic: 11-2
  K-factor: 2-5
  LC-matching networks: 9-9ff
  Limited space: 4-24ff
  Long-wire and traveling-wave: 13-1ff
  Losses: 3-2
  Low frequency: 6-1ff
  Magnetic core: 5-6ff
  Manufacturers product list: 21-1

- Antennas:
  - Matching network: 9-2
  - Matching network design: 9-6ff
  - Materials and accessories: 20-1ff
  - Mobile: 16-1ff
  - Mobile & marine: 16-1ff
  - Mobile, transient protection: 1-15
  - Modeling programs: 4-1ff
  - Multi-dipole: 7-7ff
  - Multiband: 7-1ff
  - Near resonance: 9-2ff
  - Off-center-fed dipole: 7-8ff
  - Omnidirectional, VHF and UHF: 18-22
  - Open-Sleeve: 7-18ff
  - Other characteristics: 2-15ff
  - Parabolic reflector: 19-14ff
  - Pattern measurements: 2-7
  - Pattern planes: 2-8
  - Polarization: 2-8, 27-48
  - Portable 2-element 6-meter quad: 15-11ff
  - Portable 3-element portable 6-meter Yagi: 15-13ff
  - Practical aspects: 4-19ff
  - PVRC mount: 22-27ff
  - Q: 9-2, 9-3
  - Quad arrays: 12-1ff
  - Quadrifilar: 19-4
  - Raising alongside the tower: 22-23ff
  - Range setup: 27-47ff
  - Re-radiation: 4-21
  - Receiving: 2-6
  - Receiving wave antennas: 13-15ff
  - Reciprocity: 2-7
  - Repeater systems design: 17-1ff
  - Resonant rhombic: 13-9ff
  - Rotators: 22-30ff
  - Satellite: 19-2ff
  - Selection criteria: 18-1
  - Separate, for repeaters: 17-8ff
  - Short antennas: 6-29ff
  - Simple wire: 7-1ff
  - Site planning: 4-19ff
  - Slopers: 6-44ff
  - Space communications: 19-1ff
  - Support: 22-1ff
  - System: 2-2, 4-19
  - System compromises: 4-22
  - Terminated rhombic: 13-11ff
  - Texas potato masher: 19-2
  - Textbooks: 2-19
  - Top and side mounting: 17-5ff
  - Trap: 7-9ff, 25-2
  - Turnstile over reflector: 19-3
  - Vertical: 6-15ff
  - W3DZZ: 7-10ff
  - Windom: 7-6
  - with parallel-tuned circuits: 9-6ff
  - Zip-cord: 15-2ff
- AO-10: 19-2, 19-7
- AO-13: 19-2
- AO-40: 19-7, 19-10, 19-12, 19-14, 19-24, 19-29
- AO-7: 19-2
- Aperture, antenna: 27-48
- Appendix A: 8-12, 8-56
- Appendix B: 8-17, 8-19, 8-37
- Arm-strong manual positioning: 19-30
- Army loop: 5-10
- Array Solutions: 11-42
- *Arrayfeed1* program: 8-18ff, 8-23ff
- Arrays:
  - Adjusting feed systems: 8-32ff
  - Baluns: 8-21
  - Bi-square: 8-48ff
  - Bidirectional: 8-2
  - Binomial current grading: 8-43
  - Bobtail curtain: 8-44ff
  - Broadside: 8-2, 8-5, 8-37ff
  - Bruce: 8-45ff
  - Close-spaced driven: 8-13
  - Collinear 2 elements: 8-37ff
  - Common phased-array feed systems: 8-14ff
  - Current distribution: 8-7
  - Design examples: 8-23ff
  - Directional switching: 8-33ff
  - Driven: 8-2
  - Driven combinations: 8-53ff
  - End-fire: 8-2, 8-5, 8-50ff
  - Feeding, shunt components: 8-21
  - Forward gain: 8-26
  - Four-Square: 8-18, 8-27
  - Four-Square, L-network feed: 8-29
  - Four-Square, simplest feed: 8-28ff
  - Ground losses: 3-10
  - L-network feed system: 8-25ff
  - Large, feed systems: 8-16
  - Lazy-H: 8-48
  - Loss resistance & pattern: 8-26
  - Measuring currents: 8-32ff
  - Mutual coupling: 8-26
  - Other broadside: 8-42ff
  - Parallel broadside: 8-41ff
  - Parasitic: 8-2, 8-6, 8-13
  - Pattern null degradation: 8-21
  - Phased: 4-14
  - Phased arrays, feeding: 8-13ff
  - Phased horizontal: 6-12ff
  - Phased, element feed-point impedance: 8-10
  - Phased, feeding: 8-10
  - Phased, patterns vs spacing: 8-8, 8-9
  - Phased, techniques: 8-10ff
  - Phasing arrows: 8-55
  - Practical aspects: 8-31ff
  - Quad: 12-1ff
  - Receiving: 8-15, 8-22
  - Recommended feed methods: 8-16ff

- Arrays:
  - Shunt- or gamma-fed feeding: 8-20ff
  - Simplest feed system: 8-18ff
  - Sterba: 8-41
  - Unidirectional: 8-2
  - W8JK: 8-51ff
  - 3-element binomial broadside: 8-27
  - 4-element driven: 8-53
  - 4-element rectangular array: 8-29ff
  - 4-element, feeding: 8-20ff
  - 8-element driven: 8-54
  - 120 deg. fed, 60 deg. spaced: 8-30ff
- ARRL log coordinate system: 2-12
- Arrow antenna: 19-2
- ASTM A 475-89: 22-15
- Asymmetrical routing of dipole feed line: 26-19ff
- Atchley, Dana, W1CF: 6-47
- Atkins, Bob, KA1GT: 18-52, 18-54
- Attenuation:
  - From impedance measurements: 27-30, 28-8ff
  - Line, in Smith Chart calculations: 28-6ff
- Attenuator pad, switchable: 27-42ff
- Aurora: 23-36
- Autek Research RF-1, VA1: 9-2
- Automatic tuner: 5-26
- Average gain test: 4-14, 4-16ff
- Az-El: 19-28ff
- Az-el mount: 19-11, 19-12
  - angle: 2-8, 2-15, 4-1, 19-1
  - drive: 19-42
  - pattern: 2-9, 11-3

B

- Backscatter: 23-36
- Balanced electrostatically: 14-2
- Balanced feeder: 4-23
- Balloons demonstration: 2-10
- Baluns: 6-19, 18-6, 26-21ff
  - And off-center-fed dipole: 7-9
  - Arrays: 8-21
  - Choke: 6-19, 6-33, 6-45, 7-7, 7-9, 8-21, 11-42
  - Coax line: 18-6
  - Common-mode choke: 11-11, 16-24, 26-21ff
  - Common-mode current: 11-27
  - Current: 6-15, 7-9
  - Guanella: 6-13
  - Impedance step-up: 18-6
  - Impedance step-up/step-down: 26-27
  - Input of tuner: 25-17
  - Voltage: 7-7
  - W2DU: 8-21
- Bamboo spreaders: 12-4
- Bandwidth: 18-1
- Bar clamp (carpenter's clamp): 15-4
- Barkley, H. B.: 7-20
- Barrett, Lee, K7NM: 17-5
- Base loading: 16-1, 16-4ff
- Base spring: 16-1
- *BASIC*: 4-1
- Basic wind speed: 22-8
- Beamwidth: 2-11, 8-2
  - Definition of: 8-3
  - Versus gain: 8-3, 27-53
- Bedspring antenna: 4-27
- Bell labs: 3-21, 4-17
- Belrose, Dr. John S., VE2CV: 16-22
- Belts, safety: 1-2, 1-5
- Bending of radio waves: 23-3ff, 23-4
- Bending stress: 22-22
- Bent dipoles: 4-28, 6-39
- Beta match: 26-10ff
- Beverage: 13-16
  - Characteristic impedance: 13-19
  - Characteristic impedance, 2-wire: 13-21
  - Cross-fire feed system: 13-22
  - Feed-point transformers: 13-19, 13-20
  - In echelon: 13-21
  - Optimum length: 13-19
  - Practical considerations: 13-22ff
  - Signal-to-noise ratio: 13-22
  - Two-wire: 13-20ff
- Beverage, Harold, W2BML: 13-16
- Bi-square: 8-48ff
- Bidirectional: 8-3
- Big grips: 22-16
- Big ugly dish (BUD): 19-29
- Bilodeau, John, W1GAN: 17-14
- BIP/BOP: 11-43
- Bloodworth, Max, KO4TV: 16-16
- Bloom, Alan, N1AL: 9-9
- BNC connectors: 24-24
- Bobtail curtain: 8-44ff
- Bohrer, Paul, W9DUU: 14-23
- Boom length: 11-2
- Bowl-shaped SWR curve: 9-2
- Boyer, Joseph M., W6UYH: 7-27
- Braskamp, Leon, AA6GL: 15-8
- Breakall, Jim, WA3FET: 3-24, 3-32
- Breed, Gary, K9AY: 7-22
- Brewster angle: 3-12
- Bridge:
  - Circuits: 27-3ff
  - Noise, for HF: 27-25ff
  - Resistance type: 27-3ff
  - Sensitivity: 27-5
- Broadband:
  - Antenna matching: 9-1ff
  - Bifilar transformer: 26-16
  - Matching transformers: 26-16
- Broadcast approach, feeds: 8-16

- Broadside arrays: 3-10, 8-37ff, 8-41ff
- Brown, Bruce, W6TWW: 16-5, 16-26
- Brown, Fred, W6HPH: 4-30
- Brown, Jerry, K5OE: 19-3
- Brown, Lewis and Epstein: 3-2
- Bruce array: 8-45ff
- Brute force feeding: 11-38
- Buchanan, Chester, W3DZZ: 7-10, 11-26
- Building:
  - Department: 22-1
  - Regulations: 22-1
  - Safety codes: 22-1
- Bumper-mounted HF mobile antenna: 16-2
- Bunny hunting: 14-1
- Buried radials: 6-22
- Buried-radial ground system: 6-16

C

- C-band TVRO dish: 19-11
- Cable Ties, black: 4-33
- Cage:
  - Dipole: 2-4, 9-3ff
  - Vertical: 6-21
- Calorimeter, for power measurements: 27-22ff
- Capacitance:
  - Hat: 4-14, 6-23, 16-4, 16-15
  - Short verticals: 16-3
- Capacitive stubs: 7-12
- Capacitive top loading: 6-29ff
- Capacitor, coax: 6-13
- Capacitors, lines as: 28-10ff
- *CapMan*: 23-38, 23-43
- Capon, Robert, WA3ULH: 5-17
- Cardboard box antenna: 19-13
- Cardiac pacemakers and RF safety: 1-22ff
- Cardioid pattern: 14-2, 14-11, 14-15
- Carey, Wilfred, ZS6JT: 19-11
- Carolina Windom: 7-6ff, 7-9
- Cartesian coordinates: 4-4ff
- Casual antenna evaluation: 4-27
- Cataract formation: 1-17
- Caustic problem: 3-25
- Cavity:
  - Filters: 17-9ff
  - Resonators: 17-9ff
- Cebik, LB, W4RNL: 4-3, 10-1, 11-46
- Center:
  - Feed: 7-3
  - Insulator: 4-6
  - Loading: 16-1, 16-4ff
  - Prime (Smith Chart): 28-1
- Center-fed dipole: 2-3
- Central States VHF Society: 19-15
- Cerwin, Steve, WA5FRF: 16-23
- Cessna, Clair, K6LG: 19-12
- Characteristic impedance (also see Z_0): 24-4, 28-4
- Chebyshev matching: 9-9
- Chicken-wire screen: 3-2
- Chirp sounder: 23-21
- Choke balun: 6-19, 6-45, 7-7
- Choosing a transmission line: 26-1ff
- Christman, Al, K3LC (ex-KB8I): 6-16
- Christmas tree: 4-9
- Christmas tree stacks: 11-27
- Circadian rhythms: 1-18
- Circles:
  - Resistance (Smith Chart): 28-1
  - SWR (Smith Chart): 28-3
- Circuits, short and open, as line terminations: 28-3
- Circular polarization (CP): 19-2ff, 23-6
- Circular waveguide dish feeds: 18-19ff
- Circularize: 19-15
- Circularizing tau: 10-7, 10-8
- Clamps, hose: 20-9
- Climber’s safety belt: 22-15
- Climbing tips: 1-4ff
- Closely spaced wires: 4-17
- Clothing, towers: 1-2
- Cloud warming: 6-4
- Co-planar crossed Yagi: 19-5
- Coaxial:
  - Sleeve antenna: 7-23
  - Switch: 4-27
- Coaxial cable:
  - Braid and water: 4-32
  - Direct burial: 4-33
- Coaxial-resonator match: 9-13
- Code, National Electrical: 1-8, 20-1
- Coefficient:
  - Horizontal reflection: 3-16ff
  - Vertical reflection: 3-13ff
- Coffee can 2304-MHz feed: 18-22
- Coil, loading: 4-14
- Coiled-coax balun: 26-22
- Collinear: 8-2
  - Arrays: 8-37ff, 18-13ff
  - Elements: 8-7
  - Extended double Zepp: 8-40ff
  - Gain and directivity: 8-37
  - Vertical: 17-6
  - 2 elements: 8-37ff
  - 3 and 4 elements: 8-39ff
- Combinations of long wires: 13-5ff
- Combined balun and matching stub: 26-26ff
- Commercial implementations, *NEC-2*: 4-2
- Common logarithm: 2-9
- Common-mode:
  - Choke balun: 6-15, 6-33, 26-21ff
  - Currents: 8-21, 12-7
  - Effects with directional antennas: 26-20ff
  - Transmission-line currents: 26-17ff

Compact vertical dipole (CVD): 6-31ff
Comparing antennas on fading signals: 4-27
Comparisons:
  Between top and base loading: 6-29ff
  Quad vs Yagi: 12-1ff
Compass needle: 3-3
Complementary RLC network: 9-3
Conductance: 28-5
Conduction current: 3-3ff
Conductivity: 3-11ff, 3-20, 4-1, 4-11
  Map: 3-14
  Soil: 3-1
Conductivity, soil:
  Measurement of: 27-36ff
Conductors:
  Copper clad steel: 20-3
  Copper wire data table: 20-2
  Fat: 4-7
  Losses: 24-6
  Tension and sag on wires: 20-2ff
  Wire types: 20-1
Cone of radiation: 23-24
Conservation of energy: 3-28, 8-12
Continental US: 3-14
Continuous:
  Loaded antennas: 16-13ff
  Loaded verticals (helical): 6-37ff
  Loading: 6-29
Controlled environments: 1-20ff
Convergence test: 4-15ff
Cookie-can lid: 19-24
Cookie-sheet & picture-frame-glass capacitor: 5-16ff
Coordinate scales, radiation patterns: 2-11
Coordinate shortcuts, *EZNEC:* 4-7
Copper wire, data table: 20-2
Copper-clad steel conductors: 7-8
Copperweld: 20-1
Corner reflector: 18-9ff
  Design table: 18-16
Corona: 6-13, 12-1
Coronal holes: 23-28
Counterpoise: 2-17, 3-2, 6-16ff, 16-38
  Elevated: 6-46
Counterweighting: 22-2
Coupled-resonator:
  Bandwidth: 7-25
  Complexity: 7-25
  Equations: 7-24
  Frequency independence: 7-25
  30/17/12-meter dipole: 7-26
Coupled-resonator dipole: 7-22ff
Couplers:
  Cross band: 17-11
  Directional, VHF: 27-19ff
  Hybrid: 8-16
Coupling:
  Coupling the line to the antenna: 26-1ff
  Coupling the transmitter to the line: 25-1ff
  Loops: 17-16
  Mutual: 4-14, 8-4, 8-13, 11-33
Covenants: 4-19
Coverage, repeaters: 17-2ff
Cox, Roger, WB0DGF: 10-3
CP: 19-2ff
CP patch: 19-17
Crank-up tower: 4-22, 22-10ff
Create 714X-3 triband Yagi: 11-38
Critical:
  Angle: 23-24
  Frequency, vertical incidence: 11-36, 23-20
Cross-fire feed system: 13-22
Cross-flow principle: 22-20
Crossed double bazooka: 9-6, 9-7
Crossed linear antennas: 19-3ff
Cumulative distribution function: 6-1, 23-30, 23-31
Cupped reflector, S-band: 19-7
Current:
  Balun: 6-15, 26-22ff
  Balun model: 26-22
  Common-mode, on lines: 27-7ff
  Conduction: 3-3ff
  Displacement: 3-3ff, 6-17
  Distribution: 8-11, 8-14
  Distribution on open sleeve antennas: 7-20ff
  Feed-point vs element: 8-14
  Induced in conductors: 11-1
  Induced, amplitude and phase: 8-4
  Loop: 7-2, 7-8
  Measurement: 27-1
  Measuring phased arrays: 8-32ff
  Parallel: 27-7
  Parasitic conductor: 4-30
  Probe: 8-33
  RF current penetration: 3-17ff
  RF measurement probes: 27-47ff
  Source: 4-14
  Total: 3-1
  Transmission lines: 24-15
  Zone: 3-5, 3-8
Current forcing: 8-17ff, 8-19, 8-26, 8-28, 8-29
CVD: 6-31ff
  80-meter: 6-33ff

D

D'Agostino, Philip, W1KSC: 18-41
D-layer (region): 23-19ff
D-Ring: 1-5
Daytime band: 6-7
dBd: 11-2

dBi: 8-12, 11-2ff
Dead-man anchor: 22-6ff
Decibel: 2-9
Decouple antenna from feed line: 6-19
Deed restrictions: 4-19
Delay lines: 17-4
Delta match: 18-5, 26-7ff
DeMaw, Doug, W1FB: 6-43, 6-47, 14-9, 14-11
DeMaw, Jean, W1CKK: 5-23
Depth of rf current penetration: 3-17ff
Detuning sleeve balun: 26-25ff
Dielectric constant: 3-1ff, 3-11ff, 3-20, 4-1, 4-11
Dielectric, losses: 24-6
Diffraction: 3-23, 3-24, 3-24ff, 23-4
  Diffraction model: 3-24ff
  Incident-shadow boundary: 3-25
  Reflection-shadow boundary: 3-25
  Shadow boundaries: 3-25
  Wedge: 3-24
Digital elevation model (DEM): 3-30ff
Dipole:
  Arrays, log periodic: 10-1ff
  Bent: 4-28, 6-39
  Cage: 2-4, 9-3ff
  Center-fed: 2-3
  Coupled-resonator: 7-22ff
  Coupled-resonator 30/17/12-meter: 7-26
  Flattop: 4-4
  Folded: 6-10ff
  Folded, matching: 18-6
  G5RV multiband: 7-5ff, 26-1
  Halfwave center-fed: 2-3, 6-10ff
  Horizontal: 6-10ff
  Inverted-V: 4-7, 6-11ff
  Multiwire: 6-11
  Off-center fed (OCF): 2-3, 7-8ff
  Off-center-loaded dipoles: 6-39ff
  Orientation: 4-28
  Paralleled: 7-7ff
  Stagger tuned: 7-7, 9-4, 9-5
  Trap: 7-9ff
  Trap, 3.5 to 28 MHz, 5 bands: 7-10ff
  Trap, 80, 40, 20, 15 and 10-meter: 7-11ff
  Trap, 80, 40, 20, 17 and 12 meters: 7-12ff
  Twin-lead folded: 15-1, 15-2
  Vertical half-wave: 4-8
  W8NX trap: 7-11ff
  135-foot, 80 to 10-meter: 7-3ff
Direct matching to the antenna: 26-3ff
Direct ray: 3-11, 3-12, 3-16, 3-19
Direct wave: 23-4
Direction finding antennas: 14-1ff
Directional coupler: 27-3
  VHF: 27-19ff
Directional switching, arrays: 8-33ff
Directive antenna: 2-2
Directivity: 4-28
  And gain: 2-10
  And radiation pattern: 2-6
  Antenna: 2-6
  RDF: 14-1
Director: 6-13, 11-1ff
  Retrograde: 11-28
Disastrous nulls: 3-22, 11-33
Discone: 7-26
  HF, for construction: 7-26ff
  Patterns: 7-32ff
Dish:
  Barbeque: 19-15
  Center-fed: 19-15
  Circularize: 19-15
  Deep: 19-17
  EME antennas: 19-34ff
  Feeds: 19-17ff
  Focal point: 19-22
  Focal ratio (f/D): 19-15
  Gain vs errors: 19-39
  K-band: 19-16
  Long focal length: 19-15
  MMDS, S-band: 19-18
  Off-center fed: 19-15
  Offset-fed: 19-17
  Parabolic 12-foot stressed: 19-35ff
  Patch feed: 19-20ff
  Patch, no-tune dual-band: 19-20ff
  Patch, round feed: 19-24
  Patch, truncated corners: 19-21
  Short focal length: 19-15
  60-cm S-band: 19-16
Displacement current: 3-3ff, 6-17
Distributed material loads: 4-15
Diurnal pumping: 19-27
Diversity:
  Effect: 13-1
  Receiving system: 23-35
  Space: 11-38
  Techniques, repeaters: 17-12ff
Do-it-yourself propagation prediction: 23-43ff
Dorr, John, K1AR: 11-32
Double bazooka: 9-6
Double-ducky direction finder: 14-18ff
Downtilt, vertical beam: 17-4ff
Drag coefficient: 22-22
Dream station: 4-19, 4-20
Drip (loop): 4-33
Driven arrays: 8-6ff
Driven element impedance: 11-4
Drooping dipole: 7-3
Ducting: 23-25ff
Dummy element: 11-8
Dump power: 6-9

Duplexers: 17-9ff
  Adjustment: 17-19ff
  144 MHz, for construction: 17-14ff

E

E and H fields: 3-3ff
E-field: 3-3
  Ground loss: 3-6, 3-7
  Intensity: 3-6, 4-18
  Losses: 3-6, 3-7
E-layer (region): 23-20
E-plane pattern: 2-8
Earth screw: 22-17
Earth-moon-earth (EME): 19-1
EasyTrak: 19-30,1931
EB-144: 19-2
EB-432: 19-2
Echelon, Beverage antennas: 13-21
Eckols, Ansyl, YV5DLT: 10-17
Effective:
  Electrical height: 3-5
  Ground resistance: 3-7
  Height: 5-3
  Projected area (EPA): 22-20
Effects:
  of ground: 3-1ff
  of ground in far field: 3-11ff
  of irregular terrain in the far field: 3-21ff
  of other conductors: 17-7
Efficiency: 3-5
  Antenna matching network: 9-3
  Antenna traps: 7-10
  Radiation: 8-11ff
  Radiation, short verticals: 16-9ff
EFFLEN.FOR program: 2-16
Eggbeater antenna: 19-2
EIA RS-222: 22-8
8-element driven arrays: 8-54
EIRP: 17-13
Electric field: 2-7
Electrical loading: 16-1
Electrician's knot: 15-2
Electromagnetic:
  Pulse protection: 1-11ff
  Radiation hazards: 1-17ff
  Waves: 2-1
Electron density profiles: 23-20
Electronic:
  Antenna rotation: 14-7
  Beam forming: 14-7
Electrostatic shield: 5-5, 5-8
Elements:
  Assembly: 20-9
  Collinear: 8-2
  Currents: 8-14
Elements:
  Driven: 8-2
  Parallel: 8-2
  Parasitic: 8-2
  Self impedance: 11-9
  Tapered: 4-17ff
  Telescoping aluminum tubing: 11-28
Elengo, John, Jr, K1AFR: 20-3
Elevated:
  Counterpoise: 6-46
  Ground-plane antennas: 6-23ff
  Radials: 6-16, 6-20
Elevation:
  Angle: 2-8, 2-15, 3-12ff, 4-1, 19-1
  Angle statistical data: 6-1, 23-30ff
  Angles for HF communications: 23-28ff
  Another way of looking at: 6-2
  Control: 19-28
  Drive: 19-42
  Footprint: 11-30
  Moderate distances on 40 meters: 6-9ff
  Moderate distances on 75/80 meters: 6-9
  Pattern: 2-10, 11-3, 23-33ff
  Peak statistical: 11-30
  Wide footprint: 11-33
ELF (60 Hz) electromagnetic fields: 1-18
Eliminating common-mode effects—the balun: 26-21ff
Elliptical: 2-15
Elliptical polarization: 19-5, 23-3
ELNEC: 16-31
EME: 11-32
  Dish: 19-34ff
EMP protection: 1-11ff
Empiricism: 4-27
End:
  Effect: 4-31ff, 16-6
  Fed Zepp: 7-2ff
  Load: 7-1
  Loading: 6-31
End-fire:
  Arrays: 3-10
  2-element: 8-50
  Unidirectional: 8-50ff
  3-element: 8-51
  4-element collinear: 8-53
Entrance panel: 1-14
Epidemiological studies: 1-18ff
Equalizer plate: 22-18
Equations:
  Antenna test range setup: 27-49ff
  Broadband matching network efficiency: 9-3
  Broadband matching network loss: 9-3
  Calculating long-wire length: 13-4
  Calorimeter calibration: 27-23
  Capacitance, short verticals: 16-3
  Capacity hat: 6-31

- Equations:
  - Coaxial cable: 24-21
  - Decibels: 27-53
  - Depth of RF current penetration: 3-17
  - Dipole with end effect: 4-31
  - Distance to a disturbance, time domain reflectometry: 27-32ff
  - EIRP calculation: 17-13
  - Ferrite-rod antenna design: 5-6ff
  - Frequency for measuring cable losses: 27-30
  - Gain: 27-53
  - Gain & beamwidth: 8-3
  - Horizontal reflection coefficient: 3-16
  - Horn antenna design: 18-53
  - Impedance, characteristic, open-sleeve elements: 7-18
  - Impedance, characteristic, transmission lines: 27-30
  - Isolation requirement, repeater antennas: 17-7ff
  - L-network: 25-7
  - L-network matching: 6-39
  - Length, line, electrical: 27-29, 28-6
  - Loading inductance, short verticals: 16-8
  - Loop ac resistance: 5-6
  - Loop antennas, design: 5-3ff
  - Loop antennas, effective height: 5-3
  - Loop antennas, voltage: 5-3
  - Loop distributed capacitance: 5-5
  - Loop inductance: 5-4
  - Loop Q: 5-6
  - Loop radiation efficiency: 5-11
  - Loop radiation resistance: 5-10
  - Loop sensitivity: 5-8
  - Loopstick antenna design: 5-6ff
  - Loss,path: 17-8
  - Losses, line: 27-30
  - Noise bridge compensation and calibration: 27-27ff
  - Off-center loading coils: 6-41
  - Open sleeve elements, characteristic impedance: 7-18
  - PBA: 3-15
  - Pi-network: 25-7
  - Quarter-wave matching transformer: 6-24
  - Quarter-wave sloper: 6-45
  - Quarter-wave vertical: 6-16
  - Radiation resistance, short vertical: 16-3, 16-5
  - Reflection coefficient: 9-1
  - Return loss in dB and SWR: 7-20
  - S-parameters: 27-54ff
  - Series equivalent of unlike parallel reactances: 7-10
  - SWR: 9-1, 27-3
  - Transient clamping voltage: 1-15
  - Transmission line equation: 24-12
  - Transmitting loops, design: 5-10
  - Velocity factor, line: 28-10
  - Velocity of propagation: 24-3
  - Vertical reflection coefficient: 3-14
  - Wavelength: 23-2
- Equations:
  - Z_0, line, from impedance measurements: 28-8ff
  - 1-l loop circumference: 5-2
- Equatorial anomaly: 23-22
- Equivalent:
  - Circuit, resonant antenna: 9-2
  - Loading for Yagis: 6-23ff
  - Moment method: 22-20ff
- Erecting a mast: 22-7ff
- Eskenazi, Danny, K7SS: 11-40
- Euclid: 3-24
- Evaluate property: 4-19
- Extended double Zepp: 8-40ff
  - Modified: 6-13ff
- Extraordinary wave: 23-21,2322
- Extreme ultraviolet (EUV): 23-19
- *EZNEC*: 3-22, 4-2, 7-24, 9-2, 10-3, 11-28, 11-40, 12-4, 16-31
  - Example, nulls: 8-56
  - Example, phasing-line feed: 8-56ff
  - *EZNEC M Pro*: 4-2, 19-33
  - *EZNEC-ARRL*: 3-22, 8-10, 8-12ff, 8-16ff, 8-23, 11-44
  - *EZNEC/4*: 4-2

F

- F-layer (region): 23-20
- F/B: 6-13
- f/D: 19-15
- F/R: 11-35
- F1 layer: 23-20
- F2 layer: 23-20
- FABStar: 19-10
- Fading: 4-28, 23-35
  - Frequency selective: 11-38
  - Spin: 19-24
- Fall arrest safety harness: 1-3
- Fano, R. M.: 9-6
- Far field: 3-1
  - Elevation pattern: 3-24
- Faraday rotation: 19-6, 23-3
- Farmer, Ed, AA6ZM: 6-5
- Fat conductors: 4-7
  - Connected to skinny wires: 4-17
  - Elements: 8-14
- FCC: 2-8
  - Regulations: 4-18
  - RF-exposure regulations: 1-20ff
- Federal preemption (PRB-1), antennas: 22-1
- Feed line: 24-1ff
  - Radiation: 7-7
- Feed methods:
  - Antenna impedance mismatching: 9-5ff
  - Center: 7-3
  - Delay lines: 17-4
  - Direct: 7-2

Feed methods:
  Dish: 19-17ff
  Dual-helix: 19-18
  End: 7-2ff, 7-2ff
  Feeding a multiband non-resonant antenna: 26-2
  Feeding a multiband resonant antenna: 26-2
  Feeding a single-band antenna: 26-1ff
  Helix: 19-17ff
  Parallel fed dipoles: 7-7ff
  Patch: 19-17, 19-20ff
  Patch, no-tune dual-band: 19-20ff
  Phased arrays: 8-14ff
  Quagis: 18-47
  Stacked Yagis: 18-7ff
Feed point: 4-4
  Impedance: 2-2, 3-1, 3-5, 4-10
  Resistance: 9-2
  vs element current: 8-14
Feeder (also see Lines): 24-1
Fenwick, Richard, K5RR: 11-27
Ferrite rod antennas: 5-6ff, 14-3
Ferrite-core:
  Balun: 26-23ff
  Loops: 14-11ff
Fiberglass:
  for antenna construction: 20-12
  Poles: 20-12ff
  Spreaders: 12-4
Field Day: 22-3
Field intensity: 3-3, 23-3
Field strength: 8-10, 23-3
  Measurement of: 27-43ff
  Meter: 2-7
  RF power density: 1-20
Fields: 2-7
  Reactive near: 2-2
  Reinforcement & cancellation: 8-12ff
Figure of merit: 11-46
Filters, cavity type: 17-9ff
Finding capacitance hat size: 6-31ff
Finger stock: 16-16,1619
Fire extinguisher: 1-8
Fire hazards: 4-30
Fire-escape antenna: 4-26
Fisher, Reed, W2CQH: 9-15
Flagpole: 6-15
  Antenna: 4-26
  Bracket: 15-4
Flashlight analogy: 2-6
Flat earth: 3-18
  Ground models: 3-24
  Reflections: 3-15ff
Flat lines: 24-9ff
Flat projected area (FPA): 22-20
Flattop dipole: 4-4
Fletcher, Rick, KG6IAL: 19-15
Floating point number: 4-13
Focal point, dish: 19-22
Focal ratio (f/D): 19-15
foE: 23-21
foF1: 23-21
foF2: 23-21
Folded dipole: 6-10ff, 26-8
Folded monopole antenna: 6-19
Force 12 C3: 11-27ff
Forcing currents: 8-17ff
Ford, Jim, N6JF: 15-8
Formula Translation (Fortran): 2-16, 3-25, 4-3
Forward stagger: 11-27ff
FOT: 23-23
Four Square: 8-18, 8-27
Four Square switching: 8-35
Four-element driven arrays: 8-53
Four-way mobile DF system: 14-24ff
Four-wire lines: 24-16
Fox hunting: 14-1
  DF twin ’tenna: 14-20ff
Fractional wavelength: 2-5
Frame antenna: 14-3
Frame loops: 14-9ff
Francisco, Al, K7NHV: 4-24
Fraunhöfer: 2-7
Free space: 2-2, 2-3, 2-10
Freely, W. B., K6HMS: 16-23
Freestanding tower: 4-22
Frequencies, spurious, and RF measurements: 27-8
Frequency:
  Critical: 11-36
  E-region critical: 23-21
  F-region critical: 23-21
  Gyro: 23-22
  Optimum traffic (FOT): 23-23
  Range: 4-12
  Scaling: 2-16, 4-13
  Spot: 4-12
Fresh water: 3-17
Fresnel: 2-7
  Integral: 3-25
Front-to-back ratio: 4-10, 4-12, 4-16, 6-13, 6-47, 8-3, 11-3
Front-to-rear ratio: 4-10, 4-16, 11-35
Frost line: 4-33
Full duplex: 17-7
Full-wave resonant frequency: 2-4, 2-11, 2-13

G

G-line: 18-2
G3RUH: 19-13, 19-14, 19-16, 19-18, 19-20, 19-24
G3VJL: 18-48
G3WDG: 19-16, 19-22

G5RV:
  Dipole: 7-5ff, 26-11
  Inverted V: 7-6
  Lobes: 7-6
  Losses in coax: 7-5
  Multiband: 7-5ff
G6LVB: 19-24
G6XN: 6-31, 6-32, 11-46
Gain: 2-6, 8-3ff, 8-12, 18-1
  and feed point mismatch: 27-48
  Array dimensions: 8-6
  Average gain test: 4-14
  Long wire antennas: 13-1
  Measurement: 27-51ff
  Pattern degradation due to stacking: 18-9
  Relative to single element: 8-12
  Versus beamwidth: 8-3, 27-53
  Versus front-to-rear ratio, Yagi array: 11-2ff
  Versus height: 18-1
  Versus mutual impedance: 8-5ff
Galvanized steel: 20-13
Gamma match: 9-13, 18-5, 26-9ff
Gamma-match capacitor: 6-22
Geiser, David, WA2ANU: 14-18, 14-24, 14-23
Generators, emergency power: 1-12ff
Gentle slope: 3-26
Geomagnetic storm: 23-28
Geometric optics: 3-24, 3-25
Geometric theory of diffraction (GTD): 3-25
Geometry, wires: 4-4ff
Gilbert, Ed, K2SQ: 26-22
Gillette, R.F., W9PE: 14-20
Gin pole: 1-3ff, 22-7, 22-13, 22-23
G*NEC*: 4-2, 10-3
Going around the horn: 4-11
Gold, Dr Robert E, WB0KIZ: 1-17
Goniometer: 14-6ff
GPS receiver: 3-31
Graphs: 4-3
  Origin: 4-4
  Polar: 4-3
  RF currents: 4-3
  SWR: 4-3
  3D Wire-frame: 4-3
Gray-line propagation: 23-37
Greenburg, A., W2LH: 7-11
Grimaldi: 3-24
Ground: 2-2
  Array losses: 3-10
  Buried radials: 6-22
  Cities: 3-15
  Conductivity: 3-1, 4-1, 4-11
  Dielectric constant: 3-1, 4-1, 4-11
  Effect in far field: 3-11ff
  Effects, multielement arrays: 8-4
  Electrical: 1-8ff
Ground:
  Fresh water: 3-17
  Loops: 1-10
  Loss: 16-9, 16-10ff
  Lossy: 3-2
  Parameters, measuring: 27-36ff
  Perfect: 3-2, 4-2
  Plane: 2-17, 6-16
  Practical suggestions: 3-9
  Reflected ray: 23-5
  Reflection factor: 3-20
  Reflections: 3-11, 3-19, 4-1
  Resistance: 8-24
  Return loss: 3-11
  RF alternative: 1-10
  Salt water: 3-17
  Screen: 3-2, 3-7
  Strap: 1-9
  Water pipe: 7-2
  Wave: 11-37, 23-4ff
  Wave range: 23-5
Ground-plane antennas: 3-2, 4-8ff, 6-18ff, 6-18ff, 7-16
  for 144, 222 and 440 MHz: 18-23ff
Group of dipoles: 8-1
Grover, F. W.: 5-4
Guanella balun: 6-13
Gusset plates: 20-12
Guth, Dr Peter: 3-30, 17-2
Guy wires:
  Anchors: 22-6ff
  Avoiding, during installation: 22-25ff
  Earth screw: 22-17
  EHS: 22-15
  Elongation: 22-9
  Equalizer plate: 22-18
  Guy bracket: 22-17
  Klein cable grip: 22-18
  Loos guy wire tensioner: 22-18
  Material: 22-6ff
  Phillystran: 22-15ff
  Preformed guy grips: 22-16
  Resonance: 22-18ff
  Snubber assemblies: 22-12
  Tension: 22-7
  Thimbles: 22-16
  Torque arm assembly: 22-17, 20-09
  Tower: 4-22
  Track, for raising antennas: 22-25
  Turnbuckles: 22-16
Gyro frequency: 23-22

H

H field: 3-3
  Intensity: 3-6, 4-18
  Losses: 3-7

H-Adcock: 14-6
H-plane pattern: 2-8
Hairpin match: 18-6, 26-11ff
Half elements: 11-11
Half sloper: 6-23ff, 6-44
Half-power points: 2-10, 2-11, 8-2, 8-3, 11-3
Half-square antenna: 6-24ff, 8-44ff
  Changing the shape: 6-26ff
  Feed-point impedance: 6-27
  Patterns with frequency: 6-27ff
  Voltage feed: 6-27ff
Half-wave (full) sloper: 6-44
Half-wave resonance: 2-4, 2-4
Half-wave vertical dipole, feeding: 6-15
Hall, Jerry, K1TD: 6-40, 10-22, 15-1, 15-2, 22-19
Hallidy, David, K2DH: 19-32, 19-43
Halyards: 20-5ff, 22-1
  Safe working tension: 20-6
Hansen, Markus, VE7CA: 10-21, 15-11
Hard hat: 1-2, 1-5, 22-15
Hard-drawn copper wire: 20-1
Hart, Ted, W5QJR: 5-10, 5-12
Hat, capacitance: 4-14
Haviland, Bob, W4MB: 4-1, 27-36
Hazards:
  Electrical shock: 1-8ff
  Health: 1-17ff
  RF awareness guidelines: 1-23
  RF radiation: 1-17ff
HCJB, Missionary radio station: 12-1
HDL_ANT program: 19-15
Healy, Rus, K2UA (ex-NJ2L): 13-16
Heat loss: 2-2
Height: 7-1
  Effective electrical: 3-5
  Gain: 18-1
  Vertical antennas: 2-17
Helical antenna: 19-3, 19-5ff
  Axial ratio: 19-8
  Basics: 19-6ff
  Beamwidth: 19-8
  Broad bandwidth: 19-6
  Diameter: 19-8
  Gain: 19-8
  Impedance: 19-8
  Polarization diversity: 19-6
  S-band: 19-14
  Satellite: 19-6
  50-W feed: 19-8ff
Helically wound:
  Radiator: 16-1
  Verticals: 6-37ff
Helix:
  CP feed: 19-12
  Feed: 19-17ff
  Feed for offset-dish: 19-18ff
Helix:
  For Sailboats: 16-31ff
  Yagi arrays: 11-1ff
HFTA (HF Terrain Assessment):3-26ff, 4-1, 4-19, 6-1, 11-30, 11-44, 11-45, 17-2, 23-31
  and canyons: 3-30
  PDF: 3-30, 3-32
High accuracy ground: 4-11
High:
  Efficiency: 7-1
  Frequency asymptotic solution: 3-32
  Impedance point: 4-14
  Pass T-network: 25-2
  Power ARRL antenna tuner: 25-15ff
  Q circuit: 2-6
Hill-ahead: 3-26
Hill-valley: 3-26
History of antenna modeling: 4-1ff
Homemade dipole center insulators: 4-31
Hooke: 3-24
Hop, extra: 11-36
Hopengarten, Fred, K1VR: 11-37
Hops: 23-23
Horizonal:
  Polarization: 3-1, 3-2
  Polarized antennas: 3-15ff
  Reflection coefficient: 3-16ff
  Yagi model: 4-9
Horn antennas: 18-13ff, 18-18, 18-52ff
  Feed: 19-37ff
  for 10 GHz: 18-52ff
  Pyramidal: 18-52, 19-14
  Sectoral: 18-52
Hose clamps, data table: 20-11
Hutchinson, Chuck, K8CH: 4-19, 11-11, 15-4
Huygens: 3-24
HVD (Halfwave Vertical Dipole): 6-15, 6-32
  Feeding: 6-15
Hybrid coupler: 8-10, 8-16

I

I8CVS: 19-10, 19-21, 19-28, 19-29, 19-31
IEEE standard: 19-3
Igniting grass: 3-6
Image antenna: 2-17, 3-11
Impedance:
  Antenna: 27-48
  Antenna, measuring: 27-30ff, 28-5ff
  Antenna, versus element coupling: 8-5
  Feed-point: 2-2, 3-5, 4-10, 8-13
  Input: 28-1
  Load: 28-1, 28-5ff
  Measurement: 27-30ff
  Mutual: 2-2, 3-10, 4-1, 8-4ff
  Plotting (Smith Chart): 28-2ff

- Impedance:
  - Self: 2-2
  - Source: 4-7
- Importance of low angles for low-band DXing: 6-1ff
- Improved:
  - Crossed-double bazooka: 9-15
  - Current distribution: 16-4
  - Telerana: 10-21
- In phase: 8-1
- Incident power: 24-5
- Indoor stealth loop: 5-26ff
- Induced current: 8-4
- Induced voltage: 3-1
- Induction field: 2-7
- Inductors:
  - Lines as: 28-10ff
- Installation:
  - Antennas: 1-1, 1-3
  - Clotheslines: 4-25ff
- Insulators: 4-31ff
  - Center: 4-6
  - for antennas: 20-3ff
  - for ribbon-line antennas: 20-4, 20-5
  - Polystyrene sheet: 7-8
  - Strain type: 20-4
  - Stress on: 20-4
- Inter-penetration: 4-8
- Interference: 2-6, 3-24
- Interferometer: 14-18
- Interlaced elements: 11-26
- International sunspot number (ISN): 23-16
- Inverse-distance law: 23-3
- Inverse-fed ground plane: 9-5
- Inversion: 23-9
- Inverted-L antennas: 6-41ff
  - Single elevated radial: 6-42
- Inverted-V dipole: 4-7, 6-11ff, 7-3ff
- Invisible long wire: 4-25
- *IONCAP*: 3-21, 23-29, 23-38, 23-43
- Ionization: 23-19
- Ionogram: 23-21
- Ionosonde: 23-21
  - Oblique angle: 23-23
- Ionosphere: 19-1, 23-19ff
  - Backscatter: 23-36ff
  - Bending in the: 23-20ff
  - Controls propagation: 23-31
  - D-layer: 23-19ff
  - Disturbed conditions: 23-27ff
  - E-region: 23-20
  - F-layer: 23-20
  - F-region: 23-20
  - Hop: 23-24
  - Layer characteristics: 23-19ff
  - Scatter: 23-36
  - Sidescatter: 23-37
- Ionosphere:
  - Sounding: 23-20ff
  - Storms: 23-28
  - Transequatorial scatter (TE): 23-22
- Irregular local terrain: 3-21ff, 4-1, 13-3
- Irrigation tubing: 6-20
- Isaacs, John, W6PZV: 14-15
- Isbell, D. E.: 10-1
- Isolate antenna from feed line: 6-19
- Isolation, repeater antennas: 17-7ff
- Isotropic antenna: 8-3, 8-12, 11-2, 23-1
- Ives, Dick, W7ISV: 6-14

J

- J-Pole, 144 MHz antenna: 18-25ff
- Jackscrew drive: 19-28ff
- Jammers of repeaters, finding: 14-1
- Jansson, Dick, WD4FAB: 19-2, 19-3
- JF6BCC: 19-21
- JG1IIK: 19-21
- Johns, Robert, W3JIP: 15-4, 15-9
- Johnson, Don, W6AAQ: 16-16, 16-26
- Jones, Bill, KD7S: 5-15

K

- K-band downlink: 19-15ff
- K-factor: 2-5
- K-Index: 23-45
- K0RZ: 19-14
- K1AFR: 20-3
- K1AR: 11-32
- K1EA: 11-31
- K1FO: 18-28ff, 19-33
- K1FO 144 MHz Yagi, 12 elements: 18-34ff
- K1FO 222 MHz Yagi, 16 elements: 18-38ff
- K1FO 432 MHz Yagi, 22 elements: 18-31ff
- K1KI: 11-46
- K1TD: 6-40, 10-22, 15-1, 15-2, 22-19
- K1VR: 11-27, 11-37, 11-40ff
- K1WA: 6-48
- K2DH: 19-43
- K2RIW: 19-35, 19-40
- K2SQ: 26-22
- K3LC: 6-16
- K3OQF: 6-31
- K3TZ: 19-21
- K4ERO: 3-9
- K4EWG: 10-25
- K5GNA: 19-12, 19-15
- K5GW: 19-33
- K5IU: 22-20
- K5OE: 19-2, 19-3, 19-12, 19-13, 19-15, 19-16, 19-18, 19-20, 19-22
- K5RR: 11-27

- K5ZI: 23-34
- K6HMS: 16-23
- K6LG: 19-12, 19-13
- K6SE: 6-21
- K6STI: 7-33
- K7NHV: 4-24
- K7NM: 17-5
- K7NV: 22-1, 22-27
- K7SS: 11-40
- K8CH: 1-14, 4-19, 6-31, 6-32, 11-11, 15-4
  - CVD, 30 meters: 6-31ff
- K9AY: 7-22, 15-17
- K9EK: 19-26
- KA1GT: 18-52, 18-54
- KA5FSB: 4-28
- KA5JPD: 19-42
- Kandonian, Armig G.: 7-27
- Kansas City Tracker (KCT): 19-30
- KB8I: 6-16
- KD5RO: 19-43
- KD7S: 5-15
- Keller, J. B.: 3-24
- Keyhole pattern: 17-3
- KG6IAL: 19-15
- KH6IJ: 19-6
- Kite: 22-2
- Klein cable grip: 22-18
- Knadle, Dick, K2RIW: 18-22, 19-35, 19-40
- Knife-edge diffraction: 23-13
- Knob and tube: 4-33
- Knots, for halyards: 20-5
- KO4TV: 16-16
- Kouyoumjian, R. G. and Pathak, P. H.: 3-25
- Kraus, John, W8JK: 4-1, 8-51, 19-5, 19-18
- Krome, Ed, K9EK: 19-14
- Krupp, Daniel A., W8NWF: 7-26
- KT-34XA tribander: 11-40

L

- L-network: 25-6ff
  - Adjustable feed system: 8-19ff
  - Feed system: 8-25ff
- L/C ratio: 7-9
- Ladder mast: 22-4
- Ladder-line:
  - Feeder: 4-33ff
  - Linear-loaded dipole: 6-34ff
- Lahlum, Robye, W1MK: 8-20
- Lambda: 2-2
- Lambert, Edgar, WA4LVB: 7-6
- Laplace loads: 4-15
- Lattin, William J., W4JRW: 7-11
- Law of reflection: 3-24
- Lawrence Livermore National Laboratories: 4-2
- Lawson, Jim, W2PV: 11-1
- Lazy-H: 8-48
- LC-matching networks, antenna: 9-9ff
- Lead weight: 11-9
- Leakage flux: 5-7
- Leeson corrections: 4-17
- Leeson, Dave, W6NL (ex-W6QHS): 2-16, 4-17, 11-27, 26-10
- Leggio, Joe, WB2HOL: 14-32
- Length-to-wire-diameter ratio: 4-6
- LEO satellite: 19-2, 19-3, 19-15
- Leslie, Sam, W4PK: 9-5
- Lewallen, Roy, W7EL: 4-2, 4-3, 8-10, 8-16, 8-23, 8-31, 27-57
- LHCP: 19-3, 19-19
- Libration: 23-7
- Light-duty guyed mast: 22-3
- Lightning:
  - Arrester: 1-9ff, 4-34ff
  - Grounds: 4-35
  - Protection: 1-11ff, 4-34ff
- Linear:
  - Coordinate systems: 2-12
  - Loading: 6-34ff
  - Polarization: 23-3
- Line: 18-2
  - as circuit elements: 28-10ff
  - Characteristic impedance: 24-4
  - Common-mode currents on: 27-7ff
  - Coupling to antennas: 26-1ff
  - Flattener: 26-1
  - for repeaters: 17-1
  - G-line: 18-2
  - Impedance transformation with: 28-13ff
  - Isolator: 7-7
  - Length, electrical, determining: 27-7, 27-29, 28-6
  - Loss: 25-1
  - Losses: 24-6ff
  - Matched-line losses: 24-6ff
  - Measurement, coaxial cable parameters: 27-29ff
  - Measurement, electrical length: 27-29
  - of force: 3-3
  - Printer: 4-3, 4-5
  - Radiation from: 7-3
  - Reactance & circuit Q: 25-4
  - Sampler, VHF: 27-19ff
  - Single-wire: 24-26
  - Smith Chart calculations: 28-1ff
  - Transient protection: 1-12
  - Velocity factor, from impedance measurements: 28-10
  - Z_0, from impedance measurements: 28-8ff
- Load: 4-14ff
  - Coil: 3-5, 4-14
  - Coil design: 16-3
  - Coil resistance: 16-9
  - Distributed material: 4-15

- Load:
  - Laplace: 4-15
  - Mathematical: 4-15
  - SWR: 9-1
  - Up into itself: 25-9
- Loading:
  - End: 6-31
  - Top: 3-5
- Lobe: 2-10, 3-11, 7-4
  - Higher angle: 11-36
  - Low sidelobe: 11-33
  - Major: 3-19, 8-3
  - Minor: 8-2
  - Second: 3-22
  - Worst-case: 11-35
- Log periodic antennas: 10-1ff
  - Log-cell Yagi: 10-1, 10-10ff
- Logarithm, common: 2-9
- Logarithmic: 2-3
  - Coordinate system: 2-12
- Long boom, support: 20-9
- Long path: 23-34ff
- Long-distance communication: 4-20
- Long-wire and traveling-wave antennas: 13-1ff
  - Calculating length: 13-4
  - Combinations: 13-5ff
  - Directivity: 13-3ff
  - Diversity effect: 13-1
  - Feeding: 13-5
  - Feeding the V beam: 13-9
  - Gain: 13-1
  - Gain vs wire length: 13-2
  - Other V combinations: 13-8, 13-9
  - Parallel wires: 13-6
  - Resonant rhombic: 13-9ff
  - Terminated: 13-10ff
  - Tilted wires: 13-5
  - V-beam: 13-6ff
- Loop antennas: 5-1ff, 14-2
  - and environment: 5-8ff
  - and propagation effects: 5-8
  - Antenna effect: 5-5
  - Aperiodic arrays: 5-9
  - Arrays of: 5-9
  - Balance: 5-5
  - Cookie-sheet variable capacitor: 5-16ff
  - Current: 7-2, 7-8
  - Design data tables: 5-13
  - Distributed capacitance: 5-4
  - Electrostatically shielded: 5-5
  - Half wave: 5-1
  - Horizontal: 5-19ff
  - Inductive loading: 5-2
  - Large: 5-1
  - Loopstick: 5-2
  - Losses in: 5-14
- Loop antennas:
  - One wavelength: 5-1
  - Pattern comparisons: 5-12
  - Pattern distortion: 5-5
  - Pattern nulls: 5-4
  - Patterns: 5-3
  - Practical transmitting loops: 5-12ff
  - Q of: 5-6
  - Shielded: 5-5
  - Skywire: 5-19ff
  - Small: 5-1, 5-2ff
  - Small, transmitting: 5-10, 5-10ff
  - Stealth: 5-26ff
  - Transmitting, construction and tune-up: 5-15
  - Trombone variable capacitor: 5-15ff
  - Tuned: 5-3ff
  - Tuning: 5-3
  - Tuning capacitor: 5-14ff
  - Typical construction: 5-17ff
  - Untuned: 5-2
  - Vertical: 5-22
  - vs phased arrays: 14-7ff
  - Voltage: 7-2, 7-8
  - Yagis, for 1296 MHz: 18-48ff
- Loopstick: 5-6ff, 14-3
  - for 3.5 MHz: 14-14ff
- Loos guy wire tensioner: 22-18
- Lord Rayleigh: 3-24
- LOS (line of sight): 17-2
- Loss:
  - Additional loss due to SWR: 24-10ff
  - and Smith Chart calculations: 28-6ff
  - and SWR: 28-6ff
  - Antenna broadband matching network: 9-3
  - E-field: 3-6
  - Earth: 3-2
  - Ground: 16-9ff
  - Heat: 2-2
  - In loop antennas: 5-14
  - In trap antennas: 7-10
  - Line: 27-30, 28-6ff
  - Measuring: 27-30, 28-8ff
  - Near-field earth: 7-16
  - Path: 17-8
  - Radiation from transmission line: 24-1
  - Resistance: 3-4ff, 9-2
  - Resistive: 2-11, 8-10
  - Return (dB), and SWR: 7-20
  - Spillover: 19-17ff
  - Transmission: 19-1
- Loudspeaker: 2-1
- Low:
  - Earth orbit (LEO): 19-2
  - Frequency antennas: 6-1ff
  - Frequency fields: 1-17ff
  - Impedance point: 4-13

- Low:
  - Power link-coupled antenna tuner: 25-13ff, 25-13ff
  - Q circuit: 2-6
- Lowest usable frequency (LUF): 23-27
- *LPCAD30*: 10-3, 10-5
- LPDAs: 4-17, 10-1ff, 26-2
  - Adding parasitic director: 10-8
  - Alpha: 10-1ff
  - Basic design considerations: 10-1ff
  - Behavior: 10-4
  - Circularizing tau: 10-7, 10-8
  - Design and computers: 10-3ff
  - Design procedure: 10-9ff
  - Equalizing upper-frequency response: 10-8
  - Feeding and constructing: 10-6ff
  - Log-cell Yagis: 10-10ff
  - Optimal sigma: 10-2
  - Performance weakness: 10-4
  - Phase line: 10-1
  - Relative current magnitude: 10-2, 10-5
  - Sigma: 10-1ff
  - Special correction: 10-8
  - Special design considerations: 10-7ff
  - Tau: 10-1ff
  - Wire LPDA for 3.5 or 7 MHz: 10-11ff
  - 5-band LPDA: 10-15ff

M

- Maer, Claude, W0IC: 14-14
- Magic bullet: 8-16
- Magnetic field: 2-7, 3-3
  - Health hazards: 1-18ff
  - 60 Hz, near equipment and appliances: 1-23ff
- Main lobe: 2-10, 11-3
- Mainframe: 4-3
- Major lobes: 3-19, 8-2, 13-3
- Mallete, Malcolm C., WA9BVS: 14-24
- Mallozzi, Domenic, N1DM: 5-2, 16-21, 17-3
- Malowanchuk, Barry, VE4MA: 19-45
- Manufacturers, antenna products: 21-1
- Maps:
  - Site planning: 4-19
  - Soil conductivities for continental USA: 3-14
- Marconi: 3-21
- Maritime antennas: 16-1ff
- Marmon-style clamp: 22-12
- Massachusetts Institute of Technology: 23-22
- Mast (also see Towers):
  - Centroid: 22-22
  - Guying: 22-5
  - Stout: 4-20
  - Strength: 22-22ff
- Masthead: 16-32
- Matched: 2-15
  - Lines: 24-4, 24-5, 26-3ff
  - -line loss: 24-10
- Matching: 17-1ff
  - Circuit: 6-11
  - Deliberate network mismatching: 9-6
  - Delta match: 18-5
  - Double adjustment stub: 18-5
  - Gamma: 9-13
  - Gamma match: 18-5
  - *GAMMA* program: 26-10
  - Hairpin match: 11-11
  - Impedance, short verticals: 16-12ff
  - Matching stubs: 26-12ff
  - Network loss: 9-8
  - of folded dipoles: 18-6
  - Optimum network matching: 9-6ff
  - Series section: 28-13ff
  - Stub match: 28-11ff
  - the line to the transmitter: 25-1ff
  - to the transmitter, short verticals: 16-13ff
  - Unit, Carolina Windom: 7-7
  - Universal stub: 18-5
  - with inductive coupling: 25-3ff
  - with lumped constants: 28-12ff
- Mathematical loads: 4-15
- Maximum:
  - Forward gain: 6-13
  - Line-of-sight distance: 23-6
  - Permissible exposure (MPE): 1-18ff, 4-18
  - Usable frequency (MUF): 23-20, 23-26ff
- Maxwell, James: 3-24
- Maxwell, M. Walter, W2DU: 26-22, 27-40
- McCaa, Bill, K0RZ: 19-14
- McCoy, Lew, W1ICP: 5-10
- McKim, Jim, W0CY: 19-24
- Measurements:
  - and spurious frequencies: 27-8
  - Antenna and transmission lines: 27-1ff
  - Current and voltage: 27-1
  - Current probe, RF: 8-33ff, 27-47ff
  - Feed-line electrical length: 8-36ff
  - Field strength: 27-43ff
  - Gain, antennas: 27-51ff
  - General information, antennas: 27-48ff
  - Impedance, antennas: 27-30ff, 28-5ff
  - Line length, electrical: 8-36ff, 27-8ff, 27-29
  - Noise bridge, for HF: 27-25ff
  - Radiation pattern, antennas: 27-52ff
  - Range testing, antennas: 27-47ff
  - Soil conductivity: 27-36ff
  - SWR: 27-1, 27-3ff
  - SWR, errors: 27-6
  - Transmission line parameters: 27-29ff
  - Wattmeter, calorimeter type: 27-22ff
  - Wattmeter, RF directional in-line: 27-9ff
- Mechanical strength: 4-32
- Medium: 2-1
- Metal fence: 4-26
- Metal-oxide varistor (MOV): 1-11

Method of moments: 4-1
MFJ-259B: 9-2
Michaels, Charles, W7XC: 3-12, 27-8
MicroDEM: 3-30ff, 17-2
Microphone: 2-1
MicroSmith: 4-13
Miller, James, G3RUH: 19-14
Millstone Hill: 23-21
Minimum SWR: 9-7
MININEC: 1-24, 2-15, 3-23, 4-1, 4-2, 7-22, 7-25, 8-21, 8-26, 10-3, 16-31
  Ground: 4-11
Minor lobes: 8-2ff, 8-7, 13-3
Mirror: 3-11
Mismatched lines: 24-5
  Deliberate: 9-6
  Results of: 27-48
Missionary radio station, HCJB: 12-1
Mitchell, Bob, N5RM: 11-40
MMDS dish: 19-12, 19-13, 19-18
Mobile and maritime antennas: 16-1ff
Mobile antennas:
  Base loading: 16-1
  Center loading: 16-1
  Continuously loaded: 16-13ff
  Equivalent circuit: 16-2ff
  Flutter: 23-5
  Mobile J antenna for 144 MHz: 16-21ff
  Optimum design: 16-5ff
  Top-loaded 144-MHz mobile: 16-26, 16-26
  VHF quarter-wavelength vertical: 16-27, 16-27
  5/8-wavelength 220-MHz mobile antenna: 16-28ff
  144-MHz 5/8-wavelength vertical: 16-28ff
Modeling:
  Environment, frequency: 4-12
  Environment, ground: 4-11
  History: 4-1ff
  Interactions: 4-20ff
  Limitations: 4-17ff
  Loads: 4-14ff
  Programs: 3-23, 4-1ff
  Tools: 4-19
Modifying Hy-Gain Yagis: 11-25ff
Moment of inertia: 22-22
  Antennas, stacking of: 18-7ff
  Yagi performance optimization: 11-5ff
  2-element quad model: 4-10
Monopole:
  Vertical: 2-17
  Vertical efficiency: 2-18
Moon: 19-1
Moonbounce: 19-1
Moore, Clarence, W9LZX: 12-1
Mountain tops: 6-3
Moxon rectangle beams: 11-46
Moxon, Les, G6XN: 6-31, 11-46
MUF: 23-20, 23-26ff
Multi:
  Hop propagation: 23-24ff
  Dipole antennas: 7-7ff
  Port network: 27-55
Multiband antennas: 7-1ff
  Ground systems: 3-10
  Harmonic radiation from: 7-34
  Random-length wires: 7-1ff
  Yagis: 11-26ff
Multicouplers, transmitter: 17-12
Multielement arrays: 8-1ff
Multiple quarter-wave sections: 26-7
Multiturn loops: 14-2
Mutual coupling: 4-14, 8-13, 8-56, 11-33
  and loss: 8-22
  Arrays: 8-26
  HFTA: 3-32
Mutual impedance: 2-2, 3-1, 3-10, 4-1
  and gain: 8-5ff

N

Trough reflector: 18-51
N0NSV: 19-18
N1DM: 5-2, 16-21, 17-2
N1JEZ: 19-16
N4KG: 6-23
N5BLZ: 19-33
N5RM: 11-40
N6BV: 3-21, 3-28, 11-27, 11-32, 11-37, 23-36, 25-15
N6BV/1: 6-9, 11-38ff, 11-45
N6JF: 15-8
N6LF: 6-13, 6-24, 7-22, 16-31, 27-54
  Extended Double Zepp: 6-13
N6NB: 18-13, 18-45
N6RO: 22-27
N7MQ: 27-57
National Academy of Sciences: 1-19
National Cancer Institute: 1-19
National Electrical Code: 1-8, 20-1
Natural disasters: 6-3ff
Natural low-frequency EM fields: 1-18
NC1I: 19-33
NCRP: 1-22
Near vertical incidence skywave (NVIS): 6-3ff
Near-field: 2-7, 3-1
  Earth loss: 7-16
Near-overhead angles: 11-37
Nearby communications: 6-3
Nearly resonant: 2-2
NEC: 1-24, 2-15, 3-23, 7-22, 7-25, 7-25, 10-3, 11-40, 12-4, 16-31
  NEC-2: 4-1, 4-3
  NEC-4: 4-2, 4-11, 4-17
  NEC-BSC: 3-24
NEC-4.1: 2-3, 8-21, 10-3
NEC-Win Plus: 3-22, 4-2, 10-3

- *NEC-Win Pro*: 4-2
- *NEC/Wires*: 7-33
- Negative feed-point resistance: 8-15
- Nelson, R. A., WB0IKN: 26-10
- NEMA4: 19-10, 19-27
- Newfoundland: 3-21
- Nighttime band: 6-7
- NM2A: 19-19
- NN0F: 6-34
- NOALOX: 15-9
- Noise bridge, for HF: 27-25ff
- Non reciprocal: 3-29
- Norton source: 4-14
- Nulls: 2-10, 3-11, 3-22, 8-7, 8-12, 11-30, 13-3
  - Array: 8-21
  - Disastrous: 11-33
  - Mislocated: 8-10
- NVIS: 6-3ff
  - Antenna height: 6-8ff
  - Choosing the right frequency: 6-7
  - Geographic coverage: 6-4ff
  - Low antennas and powerline noise: 6-9
  - Strategy: 6-7ff
  - Summary: 6-10

O

- O'Dell, Pete, KB1N: 14-16
- Oblique angle sounding: 23-23
- OCF dipoles: 2-3, 7-8ff
  - Even harmonics: 7-8
  - Odd harmonics: 7-8
- OD5CG: 26-5
- Oerstad: 3-3
- OET Bulletin 65: 1-20
- Off-center-loaded dipoles: 6-39ff, 7-8ff
- Offset crossed Yagi: 19-7
- Ohio State University: 3-25
- Ohm's law: 2-2
- Omega match: 6-23, 26-11
- Omnidirectional: 2-17
- ON4UF: 7-8
- ON4UN: 13-23
- *ON4UN's Low-Band DXing:* 8-20
- ON5UG: 19-13
- One:
  - Element Yagi: 11-8
  - Port measurement: 27-55
  - Way propagation: 23-34
  - Way skip: 23-34
- Open circuit, line termination: 28-3
- Open sleeve antennas: 7-18ff, 7-23, 11-28
  - Antenna mode: 7-18ff
  - Bandwidth: 7-20
  - Current distribution on: 7-20ff
  - Gain: 7-21ff
  - Impedance: 7-18ff
- Open sleeve antennas:
  - Monopole: 7-18
  - Patterns: 7-24
  - Practical construction: 7-24
  - Transmission-line mode: 7-18ff
- Open-wire lines: 2-2, 4-6, 26-1
- Operator skill: 4-19
- Optimized:
  - 10-meter Yagis: 11-12ff
  - 12-meter Yagis: 11-14ff
  - 15-meter Yagis: 11-17ff
  - 17-meter Yagis: 11-19ff
  - 20-meter Yagis: 11-20ff
  - 30-meter Yagis: 11-22ff
  - 40-meter Yagis: 11-23ff
- Optimizing over local terrain: 11-45
- Optimum ground-system configurations: 3-3
- Optimum traffic frequency: 23-23
- Ordinary wave: 23-21, 23-22
- Origin: 4-4
- Other antenna characteristics: 2-15ff
- Over-illuminating, dish: 19-17ff
- Overbeck, Wayne, N6NB: 18-13, 18-45
- Overblown claims: 4-1
- Overlay plots: 4-12

P

- Pad, switchable attenuator: 27-42ff
- Painter, J. R., W4BBP: 10-25
- Parabolas: 18-18ff
  - Dish: 19-15
  - EME: 19-34ff
  - Feed methods: 18-18ff
  - Illuminating: 18-13ff
- Paraffin, applying to wood: 20-12
- Parallel:
  - Equivalent circuit: 24-12
  - Conductor lines: 26-1
  - Tuned: 7-2
  - Tuned circuits, at antenna terminals: 9-6ff
- Parallel-wire transmission lines: 4-17
  - Arrays: 8-6, 11-1ff
  - Conductors: 4-29
  - Effects: 2-2
  - Element: 6-13, 6-13
- Passive reflector, moon: 19-1
- Patch:
  - CP: 19-17
  - Feeds: 19-20ff
  - Round feed: 19-24
  - S-band: 19-13
  - Truncated corners: 19-21
- Path loss: 17-8, 23-10,2313
- Pattern: 18-1
  - Bidirectional: 8-3
  - Dipoles at different frequencies: 2-11

Pattern:
  Distortion, arrays: ... 8-10
  Distortion, dipole: ... 26-18
  E-plane: ... 2-8
  Elevation: ... 2-10
  Factor: ... 3-17ff
  Ground-reflection: ... 11-33
  H-plane: ... 2-8
  Half-power points: ... 8-3
  Measurement of: ... 27-52
  Planes: ... 2-8
  Shape: ... 2-8
  Unidirectional: ... 8-3
  Worksheets for: ... 27-52
PBA: ... 3-13
  and earth quality: ... 3-15ff
  and Salt water: ... 3-15
Pedersen wave: ... 23-23, 23-24
Penetrox: ... 7-28
Perfect:
  Conducting ground: ... 3-18
  Earth: ... 3-12
  Ground: ... 3-2, 4-2
  Transmission line: ... 4-17
Periscope antennas: ... 18-45ff
  Gain: ... 18-55
  Patterns: ... 18-56
Perrin, Mark, N7MQ: ... 27-57
Phase: ... 3-2, 3-12
  and wavelength: ... 23-2
  Definition of: ... 8-3
  Reflected signals: ... 3-2
  Relationships: ... 8-4
  Shift: ... 4-14
Phased arrays: ... 4-14, 6-12ff
  Adjusting feed systems: ... 8-32ff
  Baluns: ... 8-21
  Bi-square: ... 8-48ff
  Bidirectional: ... 8-2
  Binomial current grading: ... 8-43
  Bobtail curtain: ... 8-44ff
  Broadside: ... 8-2, 8-5, 8-37ff
  Bruce: ... 8-45ff
  Collinear 2 element: ... 8-37ff
  Common feed systems: ... 8-14ff
  Current distribution: ... 8-7
  Design examples: ... 8-23ff
  Directional switching: ... 8-33ff
  Driven: ... 8-2
  Driven combinations: ... 8-53ff
  Element feed-point impedances: ... 8-10
  End-fire: ... 8-2, 8-5, 8-50ff
  Feeding: ... 8-10, 8-14ff
  Feeding, shunt components: ... 8-21
  Forward gain: ... 8-26
  Four Square: ... 8-18, 8-27
  Four Square, L-network feed: ... 8-29
Phased arrays:
  Four Square, simplest feed: ... 8-28ff
  L-network feed system: ... 8-25ff
  Large, feed systems: ... 8-16
  Lazy-H: ... 8-48
  Loss resistance & pattern: ... 8-26
  Measuring currents: ... 8-32ff
  Mutual coupling: ... 8-26
  Other broadside: ... 8-42ff
  Parallel broadside: ... 8-41ff
  Pattern null degradation: ... 8-21
  Phasing arrows: ... 8-55
  Practical aspects: ... 8-31ff
  RDF: ... 14-5ff
  Receiving: ... 8-15, 8-22
  Recommended feed methods: ... 8-16ff
  Shunt or gamma-fed feeding: ... 8-20ff
  Simplest feed system: ... 8-18ff
  Sterba: ... 8-41
  Techniques: ... 8-10ff
  Unidirectional: ... 8-2
  Verticals: ... 6-24ff
  W8JK: ... 8-51ff
  3-element binomial broadside: ... 8-27
  4-element rectangular array: ... 8-29ff
  4-element, feeding: ... 8-20ff
  120 deg. fed, 60 deg. spaced: ... 8-30ff
Phasing lines: ... 8-10, 8-14ff
Phelps, Ted, W8TP: ... 5-26
Phi: ... 2-15
Photoionizes: ... 23-19
Physical Design of Yagi Antennas: ... 2-16
Pi-network: ... 25-1, 25-7ff
Pietraszewski, David, K1WA: ... 6-48
PL-259: ... 4-32
  Assembly: ... 24-22ff
Plane:
  Diagrams: ... 2-10
  Wave: ... 23-1
Plasencia, Richard, W0RPV: ... 5-16
Plastics, for antenna construction: ... 20-12
PlastiDip: ... 4-32
Plot:
  Polar: ... 11-33
  Rectangular: ... 11-33
Plumber's delight: ... 26-10
Point source, *HFTA:* ... 3-31
Polar coordinates: ... 2-10, 2-15
Polarization: ... 2-8, 2-13, 3-11, 11-2, 19-1, 23-2ff
  Antennas: ... 17-1, 27-48
  Circular: ... 19-2ff
  Considerations: ... 18-2
  Diversity: ... 17-12
  Elliptical: ... 2-15, 2-16, 19-5, 23-3
  EME: ... 19-4
  Horizontal: ... 2-13
    Over flat ground: ... 3-15ff
    Polarized: ... 3-1ff

Polarization:
  Linear: 2-14, 23-3
  Reflection coefficients: 3-16
  Sense: 19-3
  Switching sense: 19-4
  Through ionosphere: 2-16
  Verical: 2-13
   Antennas: 3-20ff
   Polarized: 3-1ff
Pole, gin: 1-3ff
Pole-vault arm spreaders: 12-4, 12-5
Polygons: 4-5
Polyphaser Corp: 1-14
Portable:
  Antennas: 15-1ff
  Dipole for 80 to 2 meters: 15-4ff
  2-element 6-meter quad: 15-11ff
  3-element portable 6-meter Yagi: 15-13ff
Position readout: 19-42
Potts, Frank, NC1I: 19-33
Power:
  Density: 1-22ff, 2-10
  Effective isotropic radiated: 17-13
  Gain: 8-1
  Generators, emergency: 1-12
  Loss, traps: 7-14ff
  Ratio: 2-9
  RF, absorber circuit: 27-6
  Supplies, transient protection: 1-12
Powerboats:
  Antennas: 16-38ff
  Counterpoise: 16-38
Powlishen, Steve, K1FO: 18-28
Practical 6-meter Yagis: 18-27ff
Practical aspects, array design: 8-31ff
PRB-1 (federal preemption): 22-1
Prime center: 28-1, 28-2
Probe, for RF current measurement: 8-33, 27-46ff
Proceedings of the AMSAT Space Symposium: 19-16
Programs:
  AAT: 25-9ff
  GAMMA: 26-10
  HFTA (HF Terrain Assessment): 3-26ff
  HFTA.PDF: 3-29
  LPCAD30: 10-3
  MOBILE: 6-29, 15-8, 15-9, 16-6
  TLW: 24-12, 25-7, 26-13
  WinSmith: 24-12
  YO (Yagi Optimizer): 3-23
  YT (Yagi Terrain analyzer): 3-26
  YW (Yagi for Windows): 3-23
Prop-pitch rotator: 22-32
Propagation:
  Auroral: 23-36, 23-36
  Bending: 23-4
  Beyond line-of-sight, VHF: 23-5ff
  Diffraction: 23-4
  Do-it-yourself prediction: 23-43ff
Propagation:
  Ducting: 23-25ff
  Fading: 23-35
  Geographic area: 23-25
  Gray line: 23-37
  Ground wave: 23-4ff
  Ground-wave range: 23-5
  Long path: 23-34ff
  Maximum line-of-sight distance: 23-6
  Multi-hop: 23-24ff
  One-way: 23-34
  Prediction programs: 23-44
  Radio wave propagation: 23-1ff
  Reflection: 23-4
  Refraction: 23-4
  Scatter modes: 23-36ff
  Sky wave: 23-4, 23-15ff
  Space wave: 23-4ff
  What HF bands are open?: 23-38ff
Proplab Pro program: 23-22
Protection:
  Against electromagnetic pulse: 1-11ff
  Against lightning: 1-11ff
  Circuitry: 25-1
  Coaxial-line: 1-13, 1-13ff
  Transient, devices for: 1-11ff, 1-15
  Transient, for feed lines: 1-13
Proximity effect: 5-11
Pseudo-Brewster angle: 3-12
PTFE: 19-18
Ptolemy: 3-24
Pulley: 1-3, 20-5ff, 22-1
PVC pipe: 4-33, 7-27
PVRC mount: 22-27ff
Pyramidal horn: 19-14

Q

Q factor, antenna traps: 7-10
QTH for DXing: 3-21
Quad antennas: 4-18, 12-1ff
  Arrays: 12-1ff
  at low height: 12-3
  Constructing: 12-4ff
  Diamond or square: 12-5
  Loop Yagis, for 1296 MHz: 18-48ff
  Making it sturdy: 12-4ff
  Monoband 2-element: 4-10
  Multiband: 12-3ff
  Stacking: 18-13
  Swiss: 4-18
  VHF: 18-13
  vs Yagi: 12-1ff, 12-1ff, 12-1ff
  2-element, 8-foot boom pentaband: 12-9ff
  5-element, 26-foot boom triband: 12-5ff
  144 MHz, 2 element: 18-41
  144 MHz, 4 element portable: 18-41ff
Quadrifilar antenna: 19-4

- Quagis:
  - Construction: 18-46ff
  - Design tables: 18-45ff, 18-45ff
  - Feed method: 18-47
  - VHF and UHF: 18-13
  - 1296 MHz; 10, 15 or 25 elements: 18-47
- Quarter-wave:
  - Half-sloper: 6-44
  - Transformers: 26-4
- Quartering lobes: 12-2
- Quiet sun: 23-16ff

R

- Radar: 23-21
- Radials: 2-17
  - Close to ground: 4-12
  - Elevated: 6-20
  - Quarter-wave, proximity to ground: 7-16
  - Resonating the antenna: 6-17
  - Wire systems: 3-3, 3-7ff
  - Wires: 6-16
- Radiated: 2-1
- Radiating far field: 2-7, 2-8, 3-1
- Radiating near field: 2-7, 2-8
- Radiation:
  - Efficiency: 8-11ff, 16-2, 16-9
  - Feed line & directive array: 7-3
  - Harmonic, from multiband antennas: 7-34
  - Hazards: 1-17ff
  - Nonionizing: 1-17ff
  - Pattern: 2-6
  - Pattern, coordinate scales: 2-12
  - Pattern, dipoles at different frequencies: 2-11
  - Pattern, of corner reflector: 18-15
  - Resistance: 2-2, 3-2, 8-1, 8-10, 9-2, 16-3, 16-5ff
- Radio waves:
  - Bending: 23-3ff
  - Horizon: 23-5
  - Nature of: 23-1ff
  - Path loss: 17-8
  - Propagation: 23-1ff
- Rain seepage: 4-34
- Rain-gutter antenna: 4-25
- Random length wire: 4-26
- Range:
  - of elevation angles needed: 3-21ff
  - of frequencies: 4-12
  - Test site for antenna measurements: 27-47ff
- Rauch, Tom, W8JI: 13-22
- Ray:
  - Reflected: 3-11ff
  - Technique: 3-24
  - Tracing: 3-24ff, 23-1, 23-22
- RDF antennas:
  - Adcock: 14-5ff
- RDF antennas:
  - ARDF (Amateur Radio direction finding): 14-1ff
  - by triangulation: 14-1
  - Calibration and use: 14-8ff
  - Electronic antenna rotation: 14-7
  - Ferrite rod: 14-3
  - Ferrite-core: 14-11ff
  - Four-way mobile DF system: 14-24ff
  - Fox-hunting DF twin ’tenna: 14-20ff
  - Interferometer: 14-18
  - Loop antennas: 14-2
  - Loopstick for 3.5 MHz: 14-15ff
  - Multiturn loops: 14-2
  - Phased arrays: 14-4ff
  - Sensing: 14-2
  - Shielded frame: 14-10ff
  - Shielded loop: 14-2
  - Shielded loop with sensing antenna for 28 MHz: 14-13ff
  - Snoop Loop for close-range HF RDF: 14-13ff
  - TDOA: 14-18
  - Wullenweber: 14-7
  - 144-MHz cardioid-pattern RDF antenna: 14-16ff
- Re-radiation: 4-21
- Reactance circles (Smith Chart): 28-2
- Reactive:
  - Fields: 2-8
  - Near field: 2-2, 2-7, 3-1, 3-2
- Real-world terrain and stacks: 11-45ff
- Rear quartering lobes: 12-2
- Receiving antenna: 2-6
  - Arrays: 8-15
  - Wave antennas: 13-16ff
- Reciprocity: 2-1, 2-15
- Recommended feed methods: 8-16ff
- Recreational vehicles (RV): 16-1
- Reduction in propagation speed: 23-2
- Reflected:
  - Power: 24-5
  - Ray: 3-11
  - Waves: 3-11ff
- Reflection: 23-4
  - Coefficient: 9-1, 24-7, 28-4
  - Factor: 3-18ff
- Reflectometers: 27-9ff
  - Time domain: 27-32ff
- Reflector: 6-13
  - Corner: 18-14ff
  - Cupped S-band: 19-7
  - Parasitic: 11-1ff
  - Retrograde: 11-28
  - Trough: 18-17ff, 18-51, 18-51
- Refraction: 23-4
- Regier, Frank, OD5CG: 26-5
- Regulatory restrictions: 4-19
- Relative values: 2-8

- Relay:
  - Antenna & preamplifier: 19-27
  - Remote switching: 8-35
- Repeater antennas: 17-1ff
  - Cardioid: 17-3
  - Downtilt, vertical beam: 17-4ff
  - Isolation: 17-7ff
  - Matching: 17-1ff
  - Omnidirectional: 17-3
  - Polarization: 17-1
  - Top and side mounting: 17-5ff
- Repeaters:
  - Coverage: 17-2ff
  - Diversity techniques: 17-12ff
  - Duplexers: 17-9ff, 17-9ff
  - Equipment manufacturers: 17-15
  - Isolation: 17-7ff
  - System assembly: 17-14
  - Transmission lines: 17-1
  - 144 MHz duplexer: 17-14ff
- Resistance:
  - Bridge: 27-3ff
  - Circles (Smith Chart): 28-1, 28-2
  - Effective ground: 3-7
  - Feed point, versus element coupling: 8-5
  - Ground system, arrays: 8-24
  - Loss: 8-10
  - Loss, mutual coupling & gain: 8-13ff
  - Radiation: 2-2, 3-2, 8-10
  - RF loss: 2-3
  - Swamping: 27-2
  - Virtual: 2-2
- Resonance:
  - Full-wave: 2-4
  - Half-wave: 2-4
  - in guy wires: 22-18ff
- Resonant: 2-2
  - Antennas: 9-2ff
  - Breaker: 4-30
  - Rhombic antenna: 13-9ff
- Resonators, cavity type: 17-9
- Return loss (dB), and SWR: 7-20
- Return path: 6-17
- RF:
  - ANSI exposure guidelines: 1-21
  - Athermal effects: 1-18
  - Awareness guidelines: 1-23
  - Chokes: 4-30
  - Current probe: 4-30
  - Decoupling: 6-46
  - Feedback: 7-2
  - Ground, alternative: 1-10
  - in the shack: 7-7
  - Loss resistance: 2-3
  - Rectification: 4-26
  - Safe exposure levels: 1-19
  - Thermal effects: 1-18
- RFI: 4-29
- RHCP: 19-3, 19-10
- Rhodes, Peter D., K4EWG: 10-15, 10-25
- Rhombic: 3-31, 13-1, 13-9ff
  - Front-to-back ratio: 13-15
  - Methods of feed: 13-15
  - Multiwire: 13-14, 13-15
  - Resonant: 13-9ff
  - Terminated: 13-11ff
  - Terminated, multiband design: 13-12
  - Terminated, tilt angle: 13-12
  - Termination: 13-13
- Richard, Louis, ON4UF: 7-8
- RingRotor: 11-41
- Roadway illumination standards: 6-20
- Rohn: 22-9
- Role of the Sun: 23-15ff
- Rope: 1-2ff
- Rotators: 22-30ff
  - Heavy duty: 4-20
  - Transient protection: 1-14ff
- RS-15: 19-2
- Rubber duck: 19-2
- Rusgrove, Jay, W1VD: 15-1
- Russell, Thomas, N4KG: 6-23

S

- S-band dish feed: 19-7
- S-parameters: 27-54ff
- Safety: 1-1ff, 4-22
  - Belts: 1-2, 1-5
  - D-ring: 1-5
  - Electrical: 1-8ff
  - Fall arrest: 1-3
  - Gloves: 1-5
  - Hard hats: 1-2, 1-5
  - Knots: 1-6
  - Passers-by: 1-1
  - RF burns: 1-1
  - RF radiation hazards: 1-17ff
  - Safety belt: 22-15
  - Slingshot: 1-1
  - Structural: 22-1
  - Switch box: 1-4
  - Working on towers: 1-1ff
- Sag, in antenna wires: 20-1, 20-3ff
- Sailboats:
  - Antenna modeling: 16-31ff
  - Flag halyard: 16-37
  - Grounding systems: 16-36ff
  - HF antennas: 16-31ff
  - Patterns, backsay vertical: 16-33
  - Patterns, masthead vertical: 16-33
  - Patterns, transom vertical: 16-32
  - Rigging: 16-31
  - Temporary antennas: 16-35ff

Sailboats:
  Transom and masthead mounted verticals: 16-32ff
  40-meter backstay half sloper: 16-35
Saltwater: 3-17, 3-20, 6-16, 6-50, 7-7
  Swamp on a hill: 3-20
Sampler, line, VHF: 27-19ff
Satellites: 19-1
  Geo-synchronous: 19-1
  High-altitude: 19-2ff
  Phase-3: 19-2ff
SCALE program: 2-16
Scaling:
  Boom diameter: 2-16
  Element diameters: 2-16
  Element lengths: 2-16
  Element spacings: 2-16
  Frequency: 4-13
Scatter modes: 23-36ff
Scattering: 23-4
  Parameters: 27-54ff
Schedule, taper, for antenna elements: 20-9, 20-10
Schelkunoff: 4-17
Schmidt, Kevin, W9CF: 8-16
Schulz, Walter, K3OQF: 6-31
Schuster, Jack, W1WEF: 16-1
Scissors rest: 22-7
Screwdriver:
  Coil cover: 16-20
  Loading coil: 16-16ff
  Mobile antenna: 16-16ff
Seawater: 3-12, 3-17, 3-20, 6-16, 6-50, 7-7
Second lobe: 3-22
Segments: 4-1ff
  Density: 4-6
  Length-to-wire-diameter ratio: 4-6
  Mutually coupled: 4-1
  Source: 4-5ff
Selective fading: 23-35
Self impedance: 2-2, 8-4, 11-9
Selsyn: 19-30
Sensing antennas: 5-8, 14-3ff
Sensitivity, bridge measurements: 27-5
Series:
  Equivalent circuit: 24-12
  Resonant circuit: 2-5, 2-6
  Section transformers: 26-5ff, 28-13ff
Severns, Rudy, N6LF: 6-13, 6-24, 7-22, 16-31, 27-54
Sevick, Jerry, W2FMI: 6-36, 7-16, 8-11
Seybold, Mack, W2RYI: 7-27
Shield, anti-climbing, for towers: 1-6ff
Shielded:
  Frame loops: 14-10ff
  Loop: 14-2
  Loop with sensing antenna for 28 MHz: 14-12
  Parallel lines: 24-23
Shock, electrical, and ground systems: 1-8ff
Short:
  Antennas: 6-29ff
  Circuit, line termination: 28-2
  Fat wires: 4-8
  Path: 23-34
  Range communication: 4-20
Shortened:
  Dipoles: 6-39ff
  Radials: 6-66ff
Shorty forty: 6-41ff
Shriner, Bob, WA0UZO: 17-14
Shunt:
  Components, array feeding: 8-21
  Feed, installing: 6-22
Sidelobes: 2-11, 27-53
  Quartering: 12-2
Sidescatter: 23-36
Signal-to-interference ratio: 6-5
Signal-to-noise ratio: 2-6
Silicone sealant: 6-13
Simple and Fun Antennas for Hams: 11-11
Simple:
  Broadband dipole for 80 meters: 9-16, 9-17
  Broadband matching techniques: 9-3ff
  Twin-lead antenna for HF portable operation: 15-1, 15-2
  Wire antennas: 7-1ff
Simplest feed system: 8-18ff, 8-25
Simulations of reality: 3-26ff
Site map: 4-19
Skeleton disc: 6-31
Skin depth: 3-17
Skin effect: 23-2
Skip:
  Distance: 23-24
  Zone: 23-24
Sky wave: 3-32, 23-4, 23-15ff
Sleeve, open: 11-28
Slide switch: 4-28
Sling: 22-27
Slingshot: 22-2
Slopers:
  Antennas: 6-44ff
  Front-to-back ratio: 6-49
  7-MHz sloper system: 6-48ff
Sloping the radials of a ground plane: 6-18
SMA: 19-23
Smith chart: 4-3, 26-5, 28-1ff
  Designing series-section transformers: 28-13ff
  Designing stub matches: 28-11ff
  Off-center mode: 28-14
  Procedure summary: 28-8
Smith, George, W4AEO: 10-17
Smith, Phillip H.: 28-1
Smoothed sunspot numbers (SSN): 23-17ff
Snell's law: 3-24

Snyder dipole: 9-15
SO-239: 4-31
Sobel, Jack, WØSVM: 6-41
Soft-drawn copper wire: 20-1
Solar:
  Flares: 23-27, 23-28
  Flux: 23-17ff
  Wind: 23-15
Solid radiation pattern: 2-8
Sommerfeld: 3-24, 8-23
Sommerfeld-Norton: 4-11
Sound waves: 2-1
Sounder:
  Chirp: 23-21
  Vertical incidence: 23-21
  Ionospheric: 23-20ff
Source: 4-4
  Current: 4-14, 4-22
  Impedances: 4-7
  Segment: 4-5ff
  Sensitivity to placement: 4-13
  Specification: 4-13
  Split: 4-13
  Voltage: 4-14, 4-22
Space:
  Communications: 19-1ff
  Diversity: 11-38, 17-12
  Wave: 23-4ff
Spare me the nulls: 11-32
Specific absorption rate (SAR): 1-18ff
Specific monoband Yagi design: 11-11ff
Spiders, quad: 12-7
Spillover loss: 19-17ff
Spin fading: 19-24
Spinning reel: 22-2
Splattering: 25-1
Splicing: 20-4, 22-2
Split source: 4-13
Sporadic E: 23-35
Spot frequency: 4-12
Spotting mast: 22-23
Spreading loss: 23-3
Spreadsheet: 4-13
Squint angle: 19-10
SSN: 23-17ff
Stacking: 4-9
  and azimuthal diversity: 11-38
  and fading: 11-37ff
  and gain: 11-29
  and precipitation static: 11-38
  and wide elevation footprint: 11-29ff
  Disimilar Yagis: 11-44
  Distances between Yagis: 11-32ff
  Distances for multiband Yagis: 11-37
  Main antenna: 11-39
  Minimum stacking distance: 4-22
  Monoband antennas: 18-7ff
Stacking:
  Multiplier antenna: 11-39
  Quads: 18-13
  Stacks of stacks: 18-12ff
  VHF/UHF Yagis: 18-7ff
  Yagis: 3-22, 3-27, 11-29ff
  Yagis at different frequencies: 18-8ff
  Yagis at same frequencies: 18-10ff
  10-meter example: 11-30ff
  15-meter example: 11-31
  20-meter example: 11-31ff
Stackmatch: 11-42
Stacks:
  Boom length & spacing: 11-34
  Christmas tree: 11-27
  Distance and Lobes at HF: 11-33ff
  Spacing: 11-34
  Switching out Yagis: 11-33
  Vertical spacing not critical: 11-35
Stagger tuned: 7-7
  Dipoles: 9-4, 9-5
  Radials: 9-5
Stallman, Ed, N5BLZ: 19-33
Standing:
  Wave antennas: 13-2
  Waves: 24-8ff
  Wave ratio (SWR): 24-9
Stanford, John, NNØF: 18-51, 6-34
Stanley, John, K4ERO: 3-9
Statistical entities: 3-22
Statistics, elevation angle: 6-1
Stepped diameter correction: 4-18
Sterba curtain: 8-41
Stork, Rudy, KA5FSB: 4-28
Straw, R. Dean, N6BV: 3-21, 11-37, 25-15
Structural safety: 22-1
Stubby antenna: 16-1
Stubs:
  Capacitive: 7-12
  Filter: 19-22
  Harmonics: 25-3
  Matching: 28-11ff
  on coax lines: 26-14
Subsidence: 23-9
Sudden disappearing filaments: 23-28
Suding, Robert, WØLMD: 19-11
Summer solstice: 6-6
Sun noise: 19-13
Sunspots: 23-15, 23-16
Super-J Maritime Antenna: 16-23ff
Suppliers:
  Addresses: 21-4ff
  HF antennas: 21-2
  Quad parts: 21-3
  Towers: 21-3
Supports:
  Belt hooks: 1-5

Supports:
  for antennas: 22-1ff
  for open-wire line: 4-33ff
  Ladder mast: 22-3
  Trees for verticals: 22-3
  Truss: 22-1ff
Surface acoustic wave (SAW): 14-8
Surface wave: 23-3ff
Surge impedance: 24-4
Susceptance: 28-5
Swinging link: 25-5
SWR: 24-9
  Along a line: 28-6ff
  and line losses: 28-6ff
  and the transmission line: 26-1
  Bowl-shaped curve: 9-2
  Change with common-mode current: 26-18
  Circles (Smith Chart): 28-3
  Load: 9-1
  Measurement: 27-1, 27-3ff
  Measurement, errors: 27-6
  Return loss (dB), and SWR: 7-20
Symmetry, RDF: 14-2ff
Synchros: 19-30
System approach: 4-1

T

T-match: 26-8ff
T-network: 25-8ff
  high pass: 25-2
Tag lines: 22-26
Tandem-Match, in-line directional RF wattmeter: 27-9ff
  High-power operation: 27-17ff
Tape measure beam for RDFing: 14-32ff
Taper, in antenna elements: 2-16, 4-17ff, 20-9
  and electrical length: 20-11
Tapered:
  Element, spacing: 11-10
  Lines: 26-6
  Pier-pin base: 22-9
Tapped-coil matching network: 16-15ff
TAPR: 19-30
TDR (Time Domain Reflectometer): 11-42
Teflon: 19-18
Telephone pole: 22-8
Telerana: 10-17ff
Telerana, improving the: 10-21
Telescoping aluminum tubing: 2-16, 4-17ff
Tennis ball shock absorber: 16-1
Tension, on antenna wires: 20-2
Terleski, Jay,: 11-42
Terminated:
  Long-wire antennas: 13-10ff
  Rhombic antenna: 13-11ff
Terminations on lines:
  Short and open circuits: 28-3
Terminator: 23-37
Terrain:
  Data from the Internet: 3-30ff
  Optimizing over local: 11-45
  Profile: 3-29, 17-2
Test site, antenna: 27-47ff
Texas potato masher antenna: 19-2
Textbooks, Chap 2: 2-19
TH6DXX: 11-37, 11-41
TH7DX: 3-31, 11-32, 11-37, 11-39, 11-45
The AMSAT Journal: 19-2, 19-21
The ARRL Antenna Compendium series: 9-1
The ARRL Antenna Compendium, Vol 1: 16-5
The ARRL Antenna Compendium, Vol 2: 8-18, 9-13, 9-15
The ARRL Antenna Compendium, Vol 3: 3-12, 19-33
The ARRL Antenna Compendium, Vol 4: 4-1, 6-13, 9-7, 9-17, 10-21
The ARRL Antenna Compendium, Vol 5: 6-24, 6-34, 7-22, 15-8, 15-11
The ARRL Antenna Compendium, Vol 6: 9-5, 11-46
The ARRL Antenna Compendium, Vol 7: 5-26, 16-16
The ARRL Handbook, 2006 Ed.: 6-31
The ARRL UHF/Microwave Experimenter's Manual: 19-24
The ARRL UHF/Microwave Projects Manual: 19-24, 19-42
The Pounder: 10-22ff
The Snoop Loop for close-range HF RDF: 14-13ff
Theta: 2-15
Thevenin source: 4-14
Thin-wire approximation: 4-7
Three-dimensional: 4-10
Time and frequency stations: 23-26
Time domain reflectometry: 27-32ff
Time-Difference-of-Arrival (TDOA): 14-18
TLW (Transmission Line for Windows: 8-24, 11-44, 25-7
Tolles, H. F., W7ITB: 26-10
Top (elevated) counterpoise: 6-45
Top hat: 3-5
  Construction: 6-37ff
  Multiwire: 6-30
Top loading: 3-5, 3-6, 6-18
Top-loaded 144-MHz mobile antenna: 16-26, 16-26
Top-loading capacitance: 16-14
Topography: 3-29
Torque arm assembly: 22-17,2209
Total current: 3-1
Towers:
  and lightning: 22-12
  Antenna installation: 22-23ff
  Antenna placement: 22-20ff
  Antenna selection and installation: 22-8ff
  Anti-climbing shield: 1-6ff

- Towers:
  - Bases: 22-9, 22-12ff
  - Building antennas on the tower: 22-25ff
  - Canvas bucket: 1-2
  - Climbing equipment: 1-4ff, 1-5ff
  - Climbing tips: 1-4ff
  - Clothing: 1-2
  - Crank-up: 4-22, 4-23, 22-10ff
  - Fold-over: 4-23
  - Freestanding: 4-22, 22-10ff
  - Ground captain: 1-6
  - Guyed: 4-22, 22-8ff
  - Hard hats: 1-2
  - Installation: 22-13ff
  - Installing antennas on: 1-4ff
  - Notebook: 1-5
  - PVRC mount: 22-27ff
  - Raising antennas alongside: 22-23ff
  - Rest periods: 1-6
  - Rohn: 22-9
  - Roof-mounted tripod: 22-30ff
  - Rope and pulley: 1-2
  - Safety belt: 1-1ff
  - Suppliers: 21-3
  - Tapered pier-pin base: 22-9
  - Tie-strings: 1-6
  - Tilt-over: 22-11
  - Tram, sling: 22-27
  - Tram, using a: 22-27ff
  - Unguyed: 22-10ff
  - Wet tower: 1-5
  - Working on: 1-1ff
- Tram, using a: 22-27ff
- Transducers: 2-1, 2-6
- Transformers: 2-4, 2-6
  - Quarter wave: 26-4
  - Series section: 28-13ff
- Transition function: 3-25
- Transmatch: 18-6ff, 25-1
- Transmission line: 24-1ff
  - Air-insulated coaxial: 24-17ff
  - Air-insulated lines: 24-15ff
  - as circuit elements: 24-13, 24-14
  - Cable capacitance: 24-21
  - Characteristic impedance: 24-4
  - Characteristics, table: 24-18
  - Coaxial cables: 24-19ff
  - Coaxial fittings: 24-22ff
  - Construction: 24-15ff
  - Equation: 24-11
  - Flexible lines: 24-17ff
  - Half-wavelength line: 24-13
  - Highly reactive loads: 24-14
  - Impedance transformer: 24-11ff
  - Input impedance: 24-11ff
  - Installation: 24-22ff
- Transmission line:
  - Installing: 4-31ff
  - Line voltages and currents: 24-11ff
  - Losses: 24-6ff, 25-1
  - Losses and deterioration: 24-20
  - Matched lines: 24-4, 24-5
  - Matched-line attenuation, chart: 24-20
  - Matched-line losses: 24-6ff
  - Mismatched lines: 24-5
  - Nonresistive terminations: 24-6
  - Open-wire lines: 2-2, 24-2
  - Parallel-conductor lines: 24-18ff
  - Practical: 24-7ff
  - Quarter-wavelength line: 24-13
  - Reflection coefficient: 24-7
  - Running the feed line: 4-33ff
  - Simple lightning arrester: 4-35
  - Single-wire: 24-22
  - Standing-wave ratio: 24-9
  - Support for open-wire line: 4-33ff
  - Surge impedance: 24-4
  - Terminated: 24-4ff
  - Testing: 24-27
  - Tuned feeders: 7-1
  - Twisted: 4-33
  - Two-wire lines: 24-2
  - Velocity of propagation: 24-3
  - Voltage & current: 24-15
  - Voltage and power ratings: 24-21
  - Waveguides: 24-2
  - Window lead-in panel: 4-34
  - Window line: 7-4, 24-17
  - Z_0: 24-4
  - Zip-cord: 15-2, 15-3
- Transmission loss: 19-1
- Transmission-line resonators (TLR): 9-10ff
  - as part of antenna: 9-13ff
  - as part of feed line: 9-16ff
  - Transformer: 9-11ff
- Transmitter chassis: 7-2
- Transom: 16-23
- Transom and masthead mounted verticals: 16-32ff
- Transponder noise floor: 19-11
- Transportable antennas: 15-1
- Trap Antennas: 7-9ff, 25-2
  - Dipole, 3.5 to 28 MHz, 5 bands: 7-10ff
  - Dipole, 80, 40, 17 and 12 meters: 7-12ff
  - Dipole, 80, 40, 20, 15 and 10 meters: 7-11ff
  - Trapped dipole: 26-2
  - Trapped tribanders: 11-26ff
  - 21 and 28 MHz vertical: 7-17ff
- Traps: 4-14
  - Coax cable: 7-11ff
  - Coax capacitor: 7-17
  - Coaxial: 7-11ff, 7-13
  - Construction: 7-11, 7-17ff

- Traps:
  - Doorknob capacitor: 7-11
  - Electrical loading: 7-9
  - Fundamentals: 7-9
  - Impedance equation: 7-10
  - L/C ratio: 7-9
  - Losses: 7-10
  - Losses & power rating: 7-14ff
  - Non-resonant: 7-10
  - Parallel resonant: 7-9
  - Power loss: 7-14ff
  - Power rating: 7-14ff
  - PVC pipe: 7-12
  - RTV sealant: 7-12
- Tree-mounted HF ground-plane: 15-3, 15-4
- Trees, as antenna supports: 22-1ff
  - Vertical antennas: 4-24
- Triband Yagi: 4-20, 25-2, 26-2
- Tropospheric:
  - Bending: 23-4
  - Ducting: 23-10ff
  - Propagation of VHF: 23-4, 23-7ff
  - Scatter: 23-35
- Trough reflector: 18-51
- Trough reflectors: 18-17ff, 18-51
  - for 432 and 1296 MHz: 18-51
- True north: 2-9
- Tubing, aluminum: 4-7, 20-6ff
  - Alloy numbers: 20-6
  - Construction with: 20-8ff
  - Data table: 20-7
  - Sources for: 20-7
  - Telescoping: 4-17ff
- Tucson Amateur Packet Radio (TAPR): 19-30
- Tuned feeders: 7-1
- TUNER.SUM: 25-15
- Tuners: 25-1
  - Antenna: 4-6, 7-3, 7-7
  - Balanced: 8-22
  - Circuit balance: 25-6
  - Circuit Q: 25-3
  - Harmonic attenuation: 25-2ff
  - High-power ARRL antenna tuner: 25-15ff
  - L-network: 25-6ff
  - Low-power link-coupled: 25-13ff, 25-13ff
  - Measurement of line input current: 25-6
  - Pi-network: 25-7ff
  - Series and parallel coupling: 25-3, 25-4
  - T-network: 25-8ff
  - Variable coupling: 25-5
  - with inductive coupling: 25-3ff
- Turnstile over reflector antenna: 19-3
- TV antenna: 4-25
- TV mast material: 22-5
- TV-type standoff insulators: 4-33
- TVRO:
  - C-band: 19-11
  - Converted C-band: 19-29ff
  - EME, surplus: 19-41ff
  - Feeding: 19-45
  - Modified elevation mount: 19-44
- Twilight zone: 23-37
- Twin-lead folded dipole: 15-1, 15-2
- Type-N connectors: 24-25

U

- U-Adcock: 14-6
- UBC (Uniform Building Code): 22-8
- UHF Quagis: 18-45ff
- UHF-VHF antenna systems: 18-1ff
- UL 1449: 1-12
- Ultraviolet (UV): 23-19
- Umbrella shape loading: 6-32
- Unbalanced coax feeding a balanced dipole: 26-17ff
- Uncontrolled environments: 1-20ff
- Under-illuminating (dish): 19-17ff
- Underwriters Laboratories: 1-12
- Undisturbed ionospheric conditions: 6-3
- Uneven local terrain: 3-24ff
- Uniform theory of diffraction (UTD): 3-24ff
- Unmatched system: 26-2
- Unterminated long-wire antennas: 13-2
- US Geological Survey: 3-29
- US Naval Academy: 3-30, 17-2
- USGS map: 3-29
- Using capacitive end hats: 6-36
- UT-141 coax: 19-23
- UTD: 3-29
- UV: 11-11
  - Protection: 19-19
  - Radiation: 4-33

V

- V-Beam antenna: 13-1, 13-6ff
  - Feeding: 13-9
  - Other combinations: 13-8, 13-9
- Vacuum-tube amplifiers: 25-1
- Vaisala Lightning Explorer: 6-5
- Variable coupling: 25-5
- Varney, Louis, G5RV: 7-5
- VE2CV: 16-22
- VE4MA: 19-45
- VE7CA: 10-21, 15-11
- VE7CA 2-element portable HF triband Yagis: 15-15ff
- Vector Network Analyzer (VNA): 27-54ff
- Velocity factor:
  - of line, from impedance measurements: 28-8ff
- Velocity of propagation: 24-3

Vertical antennas: 6-15ff
  Buried radials: 6-22
  Buried-radial ground system: 6-16
  Continuously loaded verticals: 6-37ff
  Counterpoises: 6-16ff
  Elevated radials: 6-16ff
  Equivalent loading: 6-23ff
  Examples: 6-20ff
  Examples of short verticals: 6-36ff
  Far-field ground reflections: 3-11ff
  Feed-point impedance and ground system: 6-17ff
  Flagpoles: 6-20
  Folded monopole: 6-19
  Ground conductivity: 3-11
  Ground dielectric constant: 3-11ff
  Ground plane: 6-18ff
  Ground systems for vertical monopoles: 3-2ff
  Half-wave dipole: 4-8
  Half-wave vertical dipole (HVD): 6-15
  Irrigation tubing: 6-20
  Less-than-ideal ground systems: 3-2ff
  Monopoles: 3-2ff
  Multiband: 7-16ff
  Omega match: 6-23
  Reflection coefficient: 3-13ff
  Roadway illumination standards: 6-20
  Short: 7-16ff
  Shortening the radials: 6-36ff
  Shunt-fed: 6-21ff
  Shunt-fed gamma: 6-22
  Sloping the radials of a ground plane: 6-18
  Standoff insulators: 6-22
  Top loading: 6-18
  Trap: 7-17ff, 7-17ff
  Tree mounted: 4-4
  Umbrella shape loading: 6-32
  1.8-3.5 MHz using existing tower: 6-21ff
  6-Foot high 7-MHz antenna: 6-36ff
Vertical incidence critical frequency: 23-26
Vertical incidence sounder: 23-21
Vertical monopole: 2-17
Vertical monopole, efficiency: 2-18
VHF:
  Quads: 18-13, 18-41ff
  Quagis: 18-13, 18-45ff
  Quarter-wavelength vertical: 16-27
VHF/UHF:
  Antenna systems: 18-1ff
  Antenna-height gain, nomograph: 23-13
  Capabilities, nomograph: 23-11
  Effective receiver sensitivity, nomograph: 23-12
  Reliable coverage: 23-10ff
  Station gain: 23-10ff
  Terrain: 23-12
View antenna window (*EZNEC*): 4-8
Viewshed: 3-31, 17-2
Vinyl electrical tape: 4-32
Virtual resistance: 2-2
VNA: 27-54ff
  Array example: 27-56
  Multi-port network: 27-55
VOAAREA: 6-5, 23-25
VOACAP: 6-5, 6-6, 11-36, 23-29, 23-33
  Parameters: 23-29
Voltage:
  Balun: 26-27
  Clamping, protection: 1-15
  Induced: 3-1
  Loop: 7-2, 7-8
  Ratio: 2-9
  Reflection coefficient: 28-4
  RF, measurement: 27-1
  Source: 4-14
  Transmission lines: 24-15
Voltmeter, RF: 27-1
Vortex shedding: 22-6

W

W-shaped SWR curve: 9-7
W0CY: 19-24, 19-26
W0LMD: 19-12, 19-15, 19-18, 19-20, 19-21, 19-24, 19-29, 19-31
W0RPV: 5-16
W0SVM: 6-41
W1AW: 23-20
W1CF: 6-47
W1EYI: 16-21
W1FB: 6-43, 6-47
W1GAN: 17-14
W1GHZ (ex-N1BWT): 19-15, 19-18, 19-19, 19-21
W1ICP: 5-10
W1KSC: 18-41
W1MK: 8-20
W1VD: 15-1
W1WEF: 6-9, 16-1, 16-2
W2BML: 13-16
W2DU: 8-21, 26-22, 27-40
W2DU balun: 26-24
W2FMI: 6-36, 7-16
W2PV: 11-1, 11-8
W3DZZ: 7-10, 7-12
W3JIP: 15-4, 15-9
W3KH: 19-4
W4AEO: 10-17
W4BBP: 10-25
W4HHK: 18-13, 18-14
W4JRW: 7-11
W4MB: 27-36
W4RNL: 10-1, 11-46
W5LUA: 19-16, 19-45
W5QJR: 5-10, 5-12

W5RTQ (now K6SE): 6-21
W6AAQ: 16-28
W6HPH: 4-30
W6NL: 4-17, 11-27, 26-10
W6NL (ex-W6QHS): 2-16
W6OWQ: 11-40
W6TWW: 16-5, 16-28
W6UYH: 7-27
W7CNK: 19-42
W7EL: 8-10, 8-16, 8-23, 8-31, 27-57
W7ITB: 26-10
W7XC: 3-12, 27-8
W8GZ: 7-6
W8JI: 13-23
W8JK: 8-13, 8-51, 19-5, 19-18
W8JK array: 6-13, 8-51ff
W8NWF: 7-26
W8NX: 7-11, 7-13
W8NX trap dipoles: 7-11ff
W8TP: 5-26
W9CF: 8-16
W9DUU: 14-23
W9LZX: 12-1
W9PE: 14-20
WA0UZO: 17-14
WA2ANU: 14-18
WA3FET: 3-24
WA3ULH: 5-17
WA4JVE: 10-15
WA5FRF: 16-23
WA5TNY: 19-42
WA9BVS: 14-24
Walters, Mike, G3JVL: 18-48
Ward, Al, W5LUA: 19-45
Wardley, David, ZL1BJQ: 19-35
Water pipe: 7-2
Wattmeters:
  A directional in-line RF: 27-9ff
  Calorimeter type, for VHF and UHF: 27-22ff
Wave:
  Angle: 3-13ff
  Attenuation: 23-3
  Extraordinary: 23-21, 23-22
  Front: 23-1, 23-2
  Ordinary: 23-21, 23-22
  Theory: 3-24
Waveguides: 18-3ff, 24-2
  Coupling to: 18-4
  Dimensions: 18-3
  Evolution of: 18-3
  Modes of propagation in: 18-3
Wavelength: 2-1
Wavelength-to-diameter ratio: 2-3
WB0DGF: 10-3
WB0IKN: 26-10
WB0KIZ: 1-17
WD4FAB: 1-99, 19-10, 19-2, 19-3, 19-5, 19-14, 19-18, 19-19, 19-27
Weatherproofing, wood: 19-26
Weber, Dick, K5IU: 22-20
White, Allan, W1EYI: 16-21
Why tribanders?: 11-40
Wilkie, Jim, WY4R: 7-6
Wilkinson divider: 8-10, 8-16
Williamson, Gerald, K5GW: 19-33
Wilson, Paul M., W4HHK: 18-14, 18-18
Wilson, Robert, AL7KK: 7-22
Wind compensation: 22-3
Wind loading: 11-8
Windom:
  Antenna: 7-6
  Carolina: 7-6ff
  Carolina, matching unit: 7-7
Windom, Loren G, W8GZ: 7-6
Window 450-Ω ladder line: 7-5ff
Winter solstice: 6-6
Wire: 4-7
  Antenna, splicing: 20-4
  as antenna conductors: 20-1
  Closely spaced: 4-17
  Coordinate shortcuts: 4-7
  Copper clad steel: 20-2
  Copper, data table: 20-2
  Crossing: 4-17
  Fat wires connected to skinny wires: 4-17
  Geometry: 4-4ff
  Grid: 2-8
  Loss: 4-15
  Mesh: 6-17
  Radial systems: 3-7ff
  Random length: 4-26
  Short fat and acute-angle junction: 4-8
  Straight: 4-5
  Stressed, data table: 20-3
  Tension and sag: 20-2ff
  Types of: 20-1
Wire LPDA for 3.5 or 7 MHz: 10-11ff
Wiring, electrical, and safety: 1-8ff
Witt, Frank, AI1H: 9-1, 9-6, 9-9
Wolf, Johann Rudolph: 23-16
Wolff, Ken, K1EA: 11-31
Wood:
  for antenna construction: 20-12
  Weatherproofing: 20-12
Worst-case front-to-rear: 8-3
Worst-case front-to-rear ratio: 11-3
Wright, Joe, W4UEB: 7-6
Wullenweber antenna: 14-7
WWV/WWVH: 23-44
WX0B: 11-42
WX0B Approach to stack matching and feeding: 11-42ff

Y

Yagi: 2-16, 11-1ff
  Bandwidth: 11-4ff
  Co-planar crossed: 19-5
  Comparisons, 3-element designs: 11-6
  Comparisons, 5-element designs: 11-6
  Comparisons, 6-element designs: 11-6, 11-8
  Comparisons, 8-element designs: 11-6, 11-9
  Construction with aluminum tubing: 11-10
  Drive impedance and SWR: 11-4ff
  Element tuning: 11-9
  F/R: 11-5
  F-R ratio: 11-2ff
  Feeding: 11-5
  Gain: 11-2ff
  Gain & pattern degradation due to stacking: 18-9
  Gain and boom length: 11-6ff, 11-10
  Gain versus F/R ratio: 11-2ff
  Horizontal model: 4-9
  Input impedance: 11-4ff
  Interlaced elements: 11-26
  Loop, for 1296 MHz: 18-48ff
  Mechanical strength: 11-2
  Modifying Hy-Gain Yagis: 11-25ff, 11-25ff
  Monoband performance optimization: 11-5ff
  Multiband: 11-26ff
  Multiband stack spacing: 11-37
  Offset crossed: 19-7
  Optimized 10-meter Yagis: 11-12ff
  Optimized 12-meter Yagis: 11-14ff
  Optimized 15-meter Yagis: 11-18ff
  Optimized 17-meter Yagis: 11-19ff
  Optimized 20-meter Yagis: 11-20ff
  Optimized 30-meter Yagis: 11-22ff
  Optimized 40-meter Yagis: 11-23ff
  Optimum design parameters: 11-5ff
  Optimum designs and element spacing: 11-8ff
  Patterns: 11-2ff, 11-2ff
  Performance optimization: 11-5ff
  Performance parameters: 11-2ff
  Specific designs: 11-12ff
  Stacked Yagis: 11-29ff
  Stacking: 18-7ff
  Stacking different frequencies: 18-8ff
  Stacking same frequencies: 18-10ff
  Stacking stacks of different frequencies: 18-12ff
  Stacking, feed method: 18-7ff
  Trapped tribanders: 11-26ff
  Two elements: 11-2ff
  VHF and UHF: 18-7ff
  2-element designs: 11-8
  6-meter: 18-27ff
  15 m Yagis: 11-17ff
Yagi vs quad: 12-1ff
Yagi-Uda: 11-1
YagiMax: 3-23
Yankee Clipper Contest Club: 11-31
Yield strength: 22-23
YO (Yagi Optimizer): 3-23
YT (Yagi Terrain analyzer): 3-26, 23-31
YV5DLT: 10-17
YW (Yagi for Windows): 3-23

Z

Z_0 of lines: 24-4
  From impedance measurements: 27-30, 28-8
Zenith angle: 2-9, 2-15
Zepp, end-fed: 7-2
Zepp, extended double: 8-40ff
Zeppelin: 7-2
Zinc-chromate primer: 20-13
Zip-cord antennas: 15-2ff
ZL1BJQ: 18-22, 19-35
Zone current: 3-5, 3-8
ZS6JT: 19-11
Zurich sunspot numbers: 23-16
1-800-MAPS-USA: 3-29
1.8 MHz antennas, receiving loop: 5-23ff
1.8-3.5 MHz vertical using existing tower: 6-21ff
1.8-MHz antennas using towers: 6-47ff
1.8-MHz inverted-L: 6-42ff
2-element broadside/end-fire switching: 8-34
2-element quad: 12-1
2-element Yagis: 11-8
2-element, 8-foot boom pentaband quad: 12-9ff
2M-22C: 19-27
2M-CP22: 19-10
3-element binomial broadside array: 8-27
3.5 MHz antennas, horizontal loop: 5-22ff
3.5 to 28 MHz trap dipole, 5 bands: 7-10ff
3.5 to 30 MHz small transmitting loops: 5-13ff
3D analysis: 4-16
4-element arrays, feeding: 8-20ff
4-element rectangular array: 8-29ff
4-element rectangular switching: 8-34
5-band LPDA: 10-15ff
5-element, 26-foot boom triband quad: 12-5ff
5/8-wavelength 220-MHz mobile antenna: 16-29ff
6-Foot high 7-MHz vertical antenna: 6-36ff
7 MHz antennas:
  Horizontal loop: 5-22ff
  Vertical loop: 5-22
7-MHz sloper system: 6-48ff
7N1JVW: 19-21
10 GHz pyramidal horn: 18-52ff
14-14.35 MHz Log-Yag: 10-25ff
20/15/10-meter triband 2-element Yagi: 15-15ff
21 and 28 MHz trap vertical: 7-17ff
23CM22EZA: 19-7
30/17/12-meter triband 2-element Yagi: 15-17ff
40-meter backstay half sloper: 16-35ff

40-meter wire Yagi: 15-19ff
80, 40, 17 and 12-meter trap dipole: 7-12ff
80, 40, 20, 15 and 10-meter trap dipole: 7-11ff
80-meter dipole with TLR transformer: 9-17
80-meter DX special: 9-13ff
120 deg. fed, 60 deg. spaced array: 8-30ff
135-foot, 80 to 10-meter dipole: 7-3ff
144 MHz antennas:
  Cardioid-pattern RDF antenna: 14-16ff
  Ground-plane: 18-23, 18-23ff
  Quad, 2 elements: 18-41
  Quad, 4 element portable: 18-41ff
  Quagi, 15 elements: 18-45ff
  Yagi, 12 elements: 18-34ff
  5/8-wavelength vertical: 16-28ff
144 MHz duplexer: 17-14ff
222 MHz antennas:
  Ground-plane: 18-23ff
  Quagi, 15 elements: 18-45ff
  Yagi, 16 elements: 18-38ff
432 MHz antennas:
  Ground-plane: 18-23ff
  Quagi, 15 elements: 18-45ff
  Quagi, 8 elements: 18-45ff
  Trough reflector: 18-42
  Yagi, 22 elements: 18-31ff
436-CP30: 19-10
1296 MHz antennas:
  Quagi, 10 elements: 18-46ff
  Quagi, 15 elements: 18-46ff
  Quagi, 25 elements: 18-47ff

Notes

Notes

Notes

Notes

Notes

Notes

Notes

FEEDBACK

Please use this form to give us your comments on this book and what you'd like to see in future editions, or e-mail us at **pubsfdbk@arrl.org** (publications feedback). If you use e-mail, please include your name, call, e-mail address and the book title, edition and printing in the body of your message. Also indicate whether or not you are an ARRL member.

Where did you purchase this book?
☐ From ARRL directly ☐ From an ARRL dealer

Is there a dealer who carries ARRL publications within:
☐ 5 miles ☐ 15 miles ☐ 30 miles of your location? ☐ Not sure.

License class:
☐ Novice ☐ Technician ☐ Technician with code ☐ General ☐ Advanced ☐ Amateur Extra

---

Name ______________________ ARRL member? ☐ Yes ☐ No
Call Sign ______________________
Daytime Phone ( ) ______________________ Age ______________________
Address ______________________
City, State/Province, ZIP/Postal Code ______________________ E-mail: ______________________
If licensed, how long? ______________________
Other hobbies ______________________
____
Occupation ______________________

For ARRL use only **ANT BK**
Edition 21 22 23 24 25 26
Printing 1 2 3 4 5 6 7 8 9 10 11 12

From ______________________________

____

____

Please affix postage. Post Office will not deliver without postage.

EDITOR, ARRL ANTENNA BOOK
ARRL—THE NATIONAL ASSOCIATION FOR
AMATEUR RADIO
225 MAIN STREET
NEWINGTON CT 06111-1494

please fold and tape

Introduction to the CD-ROM Edition
The ARRL Antenna Book, 21st Edition

“In recent years it has been borne home on us most forcibly that there is greater room for increased performance through superior antenna systems than in any other part of the equipment.”

—from the Foreword to the First Edition,

Much has changed since 1939, including the possible ways of delivering information—this CD-ROM is evidence of that! One thing that hasn’t changed is that antennas and antenna systems are still the make-or-break component of any amateur station. We are pleased to bring you this 21st edition of *The ARRL Antenna Book* on CD-ROM. The CD-ROM book contains all of the text, drawings and photos contained in the printed 21st edition. And exclusive to the CD-ROM are over 70,000 pages of propagation tables that can help you determine what HF propagation to expect throughout the world, throughout the year and throughout the sunspot cycle.

Make sure you try out the software included for the PC. Follow the instructions for installation of the software and data files.

Using this CD-ROM

This CD-ROM is viewed using Adobe’s *Reader* software. Version 7.0 of the software is included on the CD-ROM. The book and the companion files include hyperlinks. These links will appear in blue or green text. Clicking on the text of a hyperlink will cause *Reader* to display another, related part of the book. (See the *Reade*r Help documentation for information on configuring this feature.)

The version of *Acrobat Reader* used with this CD-ROM includes Acrobat Search capability, which allows rapid full-text search of the entire book. This functions as an instant index for every chapter and word in the book. We strongly recommend that you take a few minutes to view the on-line documentation available from *Acrobat Reader*’s Help menu.

Note: Adobe *Reader* version 6.0 or higher is required to use the search index included on this CD-ROM. Older versions of *Acrobat Reader* will not work with the search index, but are otherwise functional. Again, Adobe *Reader* 7.0 is included on the CD-ROM, if you need to install this version.

Installing Adobe *Acrobat Reader*

You may not have *Acrobat Reader* installed on your computer or you may have an older version installed. Installing Adobe *Acrobat Reader Ver.7* is optional during the main Setup installation, and you can do it later, after installing the *Antenna Book* software too.

To install *Acrobat Reader* for *Windows*:
Select **Run** from the *Windows* **Start** menu.

1. Type or Browse to **d:\AdbeRdr709_en_US.exe** (where d: is the drive letter of your CD-ROM drive; if the CD-ROM is a different drive on your system, use the appropriate letter) and press **Enter**.
2. Follow the instructions that appear on your screen.

To install *Acrobat Reader* for the Macintosh:

1. From the top-level folder of the CD, double click on the file “Adobe_Reader_MAC_708.dmg” to open up the disk image. Then double click the Adobe Reader Installer icon to launch the installer.

AB21 Intro to CD.DOC HE:RDS 3/7/2007 8:14 AM Page 3

Choices, Summary and Detailed Propagation Tables

USA

W1B	Boston, MA
W2A	Albany, NY
W2N	NYC, NY
W3D	Washington, DC
W4A	Montgomery, AL
W4F	Miami, FL
W4G	Atlanta, GA
W4K	Louisville, KY
W4N	Raleigh, NC
W4T	Memphis, TN
W5A	Little Rock, AR
W5H	Houston, TX
W5L	New Orleans, LA
W5M	Jackson, MS
W5N	Albuquerque, NM
W5O	Oklahoma City, OK
W5T	Dallas, TX
W6L	Los Angeles, CA
W6S	San Francisco, CA
W7A	Phoenix, AZ
W7I	Boise, ID
W7M	Helena, MT
W7N	Las Vegas, NV
W7O	Portland, OR
W7U	Salt Lake City, UT
W7W	Seattle, WA
W7Y	Cheyenne, WY
W8M	Detroit, MI
W8O	Cincinnati, OH
W8W	Charleston, WV
W9C	Chicago, IL
W9I	Indianapolis, IN
W9W	Milwaukee, WI
W0C	Denver, CO
W0D	Bismarck, ND
W0I	Kansas City, MO
W0K	Middle of US, KS
W0M	St. Louis, MO
W0N	Omaha, NE
W0S	Pierre, SD

Other, North America

6Y	Kingston, Jamaica
8P	Bridgetown, Barbados
HP	Panama City, Panama
KL7	Anchorage, Alaska
KP2	Virgin Islands
TI	San Jose, Costa Rica
V3	Belmopan, Belize
VE1	Halifax, Nova Scotia
VE2	Montreal, Quebec
VE3	Toronto, Ontario
VE4	Winnipeg, Manitoba
VE5	Regina, Saskatchewan
VE6	Edmonton, Alberta
VE7	Vancouver, BC
VE8	Yellowknife, NWT
VO1	St. John’s, NFL
VP2	Anguilla
VP5	Turks & Caicos
XE1	Mexico City, Mexico

Europe

CT	Lisbon, Portugal
DL	Bonn, Germany
EA	Madrid, Spain
EI	Dublin, Ireland
ER	Kishinev, Moldava
F	Paris, France
G	London, England
I	Rome, Italy
JW	Svalbard
OH	Helsinki, Finland
OK	Prague, Czech Republic
ON	Brussels, Belgium
OZ	Copenhagen, Denmark
SV	Athens, Greece
TF	Reykjavik, Iceland
UA3	Moscow, Russia
UA6	Rostov, Russia
UR	Kiev, Ukraine
YO	Bucharest, Romania
YU	Belgrade, Yugoslavia

South America

CE	Santiago, Chile
CP	La Paz, Bolivia
FY	Cayenne, French Guiana
HC	Quito, Ecuador
HC8	Galapagos Islands
HK	Bogota, Columbia
LU	Buenos Aires, Argentina
OA	Lima, Peru
P4	Aruba
PY1	Rio de Janeiro, Brazil
PY0	Fernando de Noronha
YV	Caracas, Venezuela
YV0	Aves Island
ZP	Asuncion, Paraguay

Asia

1S	Spratly Islands
3W	Ho Chi Minh City, Vietnam
4J	Baku, Azerbaijan
4S	Columbo, Sri Lanka
4X	Jerusalem, Israel
9N	Katmandu, Nepal
A6	Dubai, UAE
AP	Karachi, Pakistan
BY1	Beijing, China
BY4	Shanghai, China
BY0	Lhasa, China
HS	Bangkok, Thailand
HZ	Riyadh, Saudi Arabia
JA1	Tokyo, Japan
JA3	Osaka, Japan
JA8	Sapporo, Japan
JT	Ulan Bator, Mongolia
TA	Ankara, Turkey
UA9	Perm, Russia
UA0	Khabarovsk, Russia
UN	Alma-Ata, Kazakh
VR2	Hong Kong
VU	New Delhi, India
VU7	Andaman Islands
XZ	Rangoon, Myanmar

Oceania

3D2	Fiji Islands
DU	Manila, Philippines
FO	Tahiti
H4	Honiara, Solomon Islands
JD1	Ogasawara Island
KH0	Saipan, Mariana Islands
KH5K	Kingman Reef
KH6	Honolulu, Hawaii
KH8	American Samoa
V7	Kwajalein, Marshall Islands
VK2	Sydney, Australia
VK6	Perth, Australia
VK8	Darwin, Australia
YB	Jakarta, Indonesia
ZL1	Aukland, New Zealand
ZL3	Christchurch, New Zealand

Africa

3B9	Rodrigues
3C	Bata, Equatorial Guinea
3V	Tunis, Tunisia
5N	Lagos, Nigeria
5R	Antananarivo, Madagascar
5U	Niamey, Niger Republic
5Z	Nairobi, Kenya
6W	Dakar, Senegal
7Q	Lolongwe, Malawi
7X	Algiers, Algeria
9J	Lusaka, Zambia
9L	Freetown, Sierra Leone
9X	Kigali, Rwanda
C9	Maputo, Mozambique
CN	Casablanca, Morroco
CT3	Madeira Islands
D2	Luanda, Angola
EA8	Canary Islands
IG9	Lampedusa, Italy
J2	Djibouti
ST	Khartoum, Sudan
SU	Cairo, Egypt
VQ9	Chagos, Diego Garcia
XT	Burkina Faso
ZS1	Capetown, So. Africa
ZS6	Johannesburg, So. Africa

These PDF files contain propagation prediction tables valid from the transmitting site indicated in the filename to seven generalized receiving locations throughout the world in the Summary Tables and for the 40 CQ Zones in the Detailed Tables. The user selects a single transmitting site closest to his/her location. You can access this data by opening *Adobe Acrobat Reader* and selecting **Prop Index.pdf**. Or you can operate from the main table of contents in the left pane of the opening window.

Each transmitting location is organized by five levels of solar activity over the whole 11-year solar cycle:

VL (Very Low: SSN between 0 to 20)
LO (Low: SSN between 20 to 40)
ME (Medium: SSN between 40 to 60)
HI (High: SSN between 60 to 100)
VH (Very High: SSN between 100 to 150)
UH (Ultra High: SSN greater than 150)

The seven generalized locations throughout the world for the Summary Tables are:

EU = Europe (all of Europe)
FE = Far East (centered on Tokyo, Japan)
SA = South America (centered on Asuncion, Paraguay)
AF = Africa (centered on Lusaka, Zambia)
AS = southern Asia (centered on New Delhi, India)
OC = Oceania (centered on Sydney, Australia)
NA = North America (all of USA).

Both types of propagation files show the highest predicted signal strength (in S-units) throughout the generalized receiving area, for a 1500-W transmitter and rather good antennas on both sides of the circuit. The standard antennas are 100-foot high inverted-V dipoles for 80 and 40 meters, a 3-element Yagi at 100 feet for 20 meters, and a 4-element Yagi at 60 feet for 15 and 10 meters. Discount the S-Meter readings in the tables to represent a smaller station:

Subtract 2 S units for a dipole instead of a Yagi
Subtract 3 S units for a dipole at 50 feet instead of a Yagi at 100 feet
Subtract 1 S unit for a dipole at 50 feet rather than a dipole at 100 feet
Subtract 3 S units for 100 W rather than 1500 W.
Subtract 6 S units for 5 W rather than 1500 W.

Shown below is an image of a Summary Table printout from Boston to the rest of the world, for Very High solar activity in January. This table could be used, for example, to help plan which bands to operate when on a DXpedition to some exotic location.

The Detailed Table printout from Boston to the rest of the world on 20 meters for January from Boston during a Very High level of the solar cycle is shown on the following page. It shows the predicted signal strength in each of the 40 CQ Zones around the world. Note that long-path openings are predicted by an asterisk appended to the end of the predicted signal strength.

Also located on the CD-ROM in the \Propagation subdirectory is the **Fig6Tab.pdf** file described in Chapter 3 of the printed book. This set of tables shows the hours open to each of 10 regions throughout the USA for Very-Low/Medium/Very-High levels of SSN.

Enjoy the software. We would appreciate any feedback or bug reports you might have.

73,

R. Dean Straw, N6BV
Editor, *The ARRL Antenna Book*
email: **n6bv@arrl.org**

Sample Summary Propagation-Prediction Table, January from Boston to the World

Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.

	80 Meters							40 Meters							20 Meters							15 Meters							10 Meters							
UTC	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	UTC
0	9+	-	9+	9+	8	-	9+	9+	5	9+	9+	9	-	9+	1	9+	9+	9+	9+	9	9+	-	9	9+	2	2	9+	9+	-	1	8	-	-	8	9+	0
1	9+	-	9+	9+	8	-	9+	9+	4	9+	9+	9	2	9+	1	9	9+	8	9+	9+	9+	-	3	9	-	7	9+	9	-	-	-	-	-	4	2	1
2	9+	-	9+	9+	7	-	9+	9+	4	9+	9+	9	7	9+	1	9	9+	8	9	9+	9+	-	-	3	-	-	7	9	-	-	-	-	-	-	2	2
3	9+	-	9+	9+	1	2	9+	9+	4	9+	9+	9	9	9+	-	7	9+	7	8	9+	9	-	-	-	-	-	-	-	-	-	-	-	-	-	2	3
4	9+	-	9+	9+	-	7	9+	9+	5	9+	9+	8	9	9+	-	5	9+	9	9	9	9+	-	-	1	-	-	-	-	-	-	-	-	-	-	2	4
5	9+	-	9+	9+	-	8	9+	9+	6	9+	9+	7	9	9+	-	5	9+	9	9	5	9+	-	-	-	-	-	-	-	-	-	-	-	-	-	2	5
6	9+	-	9+	9+	-	8	9+	9+	7	9+	9+	7	9	9+	-	8	9+	8	9	5	9+	-	-	-	-	-	-	-	-	-	-	-	-	-	2	6
7	9+	-	9+	9+	-	8	9+	9	8	9+	9+	7	9+	9+	-	9	9+	-	7	9	9+	-	-	1	-	-	-	-	-	-	-	-	-	-	2	7
8	9	7	9+	9	-	8	9+	9	8	9+	9+	8	9+	9+	-	9	9+	-	4	9+	9+	-	-	1	-	-	-	2	-	-	-	-	-	-	2	8
9	8	7	9+	7	-	8	9+	9	9	9+	9	8	9+	9+	-	6	9+	-	1	9+	9+	-	-	-	-	-	-	1	-	-	-	-	-	-	2	9
10	5	8	9+	2	3	8	9+	9	9	9+	8	8	9	9+	4	-	9+	9+	1	5	9	-	-	-	-	-	-	-	-	-	-	-	-	-	2	10
11	1	8	9+	-	4	9	9+	8	9	9+	5	8	9	9+	9+	4*	9+	9+	7	-	8	-	-	9	9	-	-	-	-	-	-	-	-	-	2	11
12	-	7	8	-	1	9	9+	6	9	9+	1	8	9	9+	9+	9	9+	9	9	1*	9+	9	8*	9+	9+	9	5*	-	-	2*	9	9	1	1*	2	12
13	-	-	-	-	-	2	9+	4	8	8	-	7	9	9+	9+	9	9+	9	9	9+	9+	9+	7	9+	9+	9+	3*	9	9	5*	9+	9+	9	6*	2	13
14	-	-	-	-	-	-	9+	2	7	4	-	5	8	9+	9+	9	9+	8	9	9	9+	9+	9	9+	9+	9+	9	9+	9	6*	9+	9+	9	1*	1	14
15	-	-	-	-	-	-	9	1	5	-	-	4	5	9+	9+	9	9+	9	9	9	9+	9+	9+	9+	9+	9+	9	9+	9	5	9+	9+	6	6	8	15
16	-	-	-	-	-	-	8	3	4	-	-	3	1	9+	9+	8	9	9	9	9	9+	9+	9+	9+	9+	9	9+	9+	9	8	9+	9+	-	8	9	16
17	-	-	-	-	-	-	8	5	3	-	2	4	-	9+	9+	8	9+	9+	9	9	9+	9+	9	9+	9+	1*	9+	9+	-	8	9+	9+	-	8	9+	17
18	-	-	-	-	-	-	9	7	4	2	5	5	-	9+	9+	9	9+	9+	9	9	9+	9+	9	9+	9+	1	9+	9+	-	7	9+	9+	-	9+	9+	18
19	1	-	-	1	-	-	9+	8	5	6	8	7	-	9+	9+	9	9+	9+	9	9	9+	-	9+	9+	9+	2	9	9+	-	6	9+	9+	-	9+	9+	19
20	4	-	2	5	-	-	9+	9	6	9	9	8	-	9+	9+	9	9+	9+	9	9	9+	-	8	9+	9+	3	9	9+	-	1	9+	9	-	9	9+	20
21	7	-	8	7	1	-	9+	9+	7	9+	9+	8	1	9+	8	9	9+	9+	9	9	9+	-	6	9+	9+	3	9	9+	-	-	9+	5*	-	9+	9+	21
22	9	2	9+	9	8	-	9+	9+	7	9+	9+	9	4	9+	2	9+	9+	9+	9	9	9+	-	9+	9+	9	1	9+	9+	-	5	9+	4*	-	9	6	22
23	9	-	9+	9	8	-	9+	9+	7	9+	9+	9	-	9+	1	9+	9+	9+	9	9	9+	-	9+	9+	6	-	9	9+	-	7	9+	2*	-	9	2	23
	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	EU	FE	SA	AF	AS	OC	NA	

Sample Detailed Propagation Table for 20 Meters, January, Boston to World for Very High SSN

20 Meters: Jan., MA (Boston), for SSN = Very High, Sigs in S-Units. By N6BV, ARRL.

UTC -->

Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
KL7 = 01	9+	9+	9+	7	-	-	-	-	-	-	-	-	-	-	-	3	9+	9+	9+	9+	9+	9+	9+	9+
VO2 = 02	9+	9	9	9	9	9	8	7	5	3	2	1	5	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9+
W6 = 03	9+	9+	9+	7	7	1	1	5	8	8	3	-	-	1	9	9+	9+	9+	9+	9	9	9+	9+	9+
W0 = 04	9+	9+	9+	8	5	5	5	5	3	2	1	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+
W3 = 05	4	2	2	2	2	2	2	3	3	3	3	2	1	1	8	9+	9+	9+	9+	9+	9+	9+	9+	9
XE1 = 06	9+	9+	7	9	9+	9+	9+	9+	9+	9+	9	8	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+
TI = 07	9+	9+	8	9	9	9	9	9	9	9+	9	9+	9+	9+	9+	9+	9	8	9	9	9+	9+	9+	9+
VP2 = 08	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	9	9+	9+	9+	9+	9+	9	9+	9+	9+	9+	9+	9+	9+
P4 = 09	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	9	9	9	9	9+	9+	9+	9+	9+
HC = 10	9+	8	9+	9	9	9	9	9	7	3	1	7	9+	9+	9	5	5	5	7	8	9+	9+	9+	9+
PY1 = 11	9+	9+	9+	9	9	9+	9+	9	8	6	9	9+	8	2	1	-	-	1	4	8	9	9+	9+	9+
CE = 12	9+	9+	9+	9+	9+	9+	9+	9+	9	8	8	9+	9	8	2	1	1	-	1	3	7	9	9+	9+
LU = 13	9+	9+	9+	9+	9+	9+	9+	9+	8	8	8	9+	8	4	2	1	-	-	1	4	8	9	9+	9+
G = 14	-	-	-	-	-	-	-	-	-	-	-	9+	9+	9+	9+	9+	9+	9+	9+	9+	9+	8	2	-
I = 15	-	-	-	-	-	-	-	-	-	-	4	9	9	9	9	9	9	9+	9+	9	8	2	-	-
UA3 = 16	1	1	1	-	-	-	-	-	-	-	-	8	9	9+	9+	9+	9	8	5	-	-	-	-	1
UN = 17	1	-	-	8	7	7	7	1	-	-	-	2	9	9	9	6	-	-	2	4	8	9	5	4
UA9 = 18	6	7	6	6	9	9	9	7	4	1	-	-	8	8	6	6	5	6	7	8	9	9	8	7
UA0 = 19	9+	9	9	6	5	5	8	8	8	4	-	-	2	6	8	8	8	7	4	4	7	9	9+	9+
4X = 20	8	6	3	1	-	3	4	-	-	-	1	8	8	8	8	8	9	9	9+	9	9	8	7	7
HZ = 21	9+	9	4	3	8	8	2	-	-	-	1	7	8	9	8	8	9	9	9	9	9	9	9	9
VU = 22	7	5	8	7	6	7	5	-	-	-	-	6	9	9	9	9	3	2	2	2	8	8	9	8
JT = 23	9	9+	9	5	7	8	8	6	3	-	-	2*	8	8	5	6	8	8	8	8	9	7	5	6
VS6 = 24	9	9	9	5	4	5	7	8	6	1	-	1*	5	7	1	1	1	1	4	2	-	-	-	9
JA1 = 25	9	9	8	7	5	5	8	9	9	6	-	1	1	2	7	7	6	2	-	-	7	9	9+	9
HS = 26	9	9	6	4	2	-	-	2	1	-	-	2*	9	9	9	9	8	7	5	4	5	-	1*	1
DU = 27	9	8	7	-	-	-	5	7	7	1	-	-	1*	9	9	7	6	4	5	3	1*	1*	8	9
YB = 28	9	8	1	-	-	-	-	-	-	-	-	4*	8	9	9	9	8	8	9	9	9	9	9+	9+
VK6 = 29	3*	4*	-	-	-	-	-	-	5	3	-	-	-	5	9	9	9	8	9	9	9	9	9	8
VK3 = 30	1*	-	-	-	-	-	1	3	9	9	4	-	-	9+	9	8	2	1	-	-	1	2*	5*	4*
KH6 = 31	9	9+	9+	9+	8	2	2	6	4	-	-	-	-	-	-	-	9	9	8	7	6	4	6	7
KH8 = 32	-	2	9	9	9	5	5	9	9+	9+	5	-	-	9+	9	9	8	5	3	1	-	-	-	-
CN = 33	-	-	-	-	-	-	-	-	-	-	9	9+	9	9	8	9	9	9+	9+	9+	9+	9+	9+	7
SU = 34	9	8	3	3	-	1	4	-	-	-	2	7	8	8	8	8	9	9	9+	9+	9+	9+	8	8
6W = 35	9+	8	-	-	2	7	5	-	-	-	9+	9+	8	5	4	3	7	9	9+	9+	9+	9+	9+	9+
D2 = 36	9+	9+	5	3	9	9	8	-	-	-	3	-	-	-	-	4	4	7	8	9	9+	9+	9+	9+
5Z = 37	9+	9	2	4	8	8	1	-	-	-	2	-	-	3	5	5	7	8	9	9	9+	9+	9+	9+
ZS6 = 38	9+	9+	8	7	8	9	6	-	-	-	-	-	-	-	1*	1	2	6	8	9	9+	9+	9+	9+
FR = 39	9+	8	2	1	4	1	-	-	-	-	-	-	-	2*	3*	1*	1	3	8	9	9+	9+	9+	9+
FJL = 40	9+	9+	7	4	7	8	7	1	-	-	-	1*	8	9	9	9	9	9	9	9	9+	9+	9+	9+
Zone	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23

UTC -->

Expected signal levels using 1500 W and 3-element Yagis at 100 feet at each station.

Electronic Projects for Beginners

F. G. RAYER, T.Eng. (CEI), Assoc. IERE

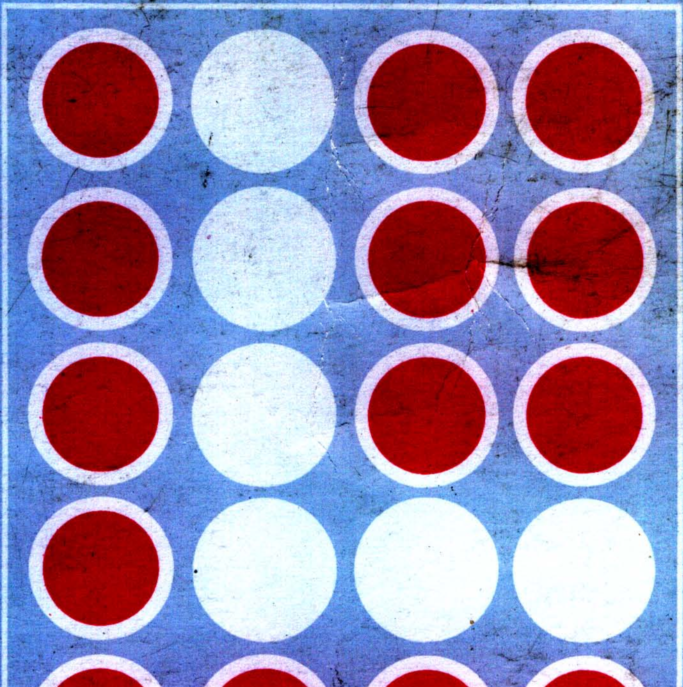

ELECTRONIC PROJECTS FOR BEGINNERS

ALSO BY THE SAME AUTHOR

No. BP32	How to Build Your Own Metal and Treasure Locators
No. BP37	50 Projects Using Relays, SCRs & Triacs
No. BP39	50 (FET) Field Effect Transistor Projects
No. BP43	How to Make Walkie-Talkies
No. BP57	How to Build Your Own Solid State Oscilloscope
No. BP67	Counter Driver and Numeral Display Projects
No. BP75	Electronic Test Equipment Construction
No. BP79	Radio Control for Beginners
No. BP84	Digital IC Projects
No. BP90	Audio Projects
No. BP93	Electronic Timer Projects
No. BP97	IC Projects for Beginners

ELECTRONIC PROJECTS FOR BEGINNERS

by

F.G. RAYER, T.Eng.(CEI), Assoc. IERE

BERNARD BABANI (publishing) LTD
THE GRAMPIANS
SHEPHERDS BUSH ROAD
LONDON W6 7NF
ENGLAND

PLEASE NOTE

Although every care has been taken with the production of this book to ensure that any projects, designs, modifications and/or programs etc. contained herein, operate in a correct and safe manner and also that any components specified are normally available in Great Britain, the Publishers do not accept responsibility in any way for the failure, including fault in design, of any project, design, modification or program to work correctly or to cause damage to any other equipment that it may be connected to or used in conjunction with, or in respect of any other damage or injury that may be so caused, nor do the Publishers accept responsibility in any way for the failure to obtain specified components.

Notice is also given that if equipment that is still under warranty is modified in any way or used or connected with home-built equipment then that warranty may be void.

©1978 BERNARD BABANI (publishing) LTD

ISBN 0 85934 054 6

First Published – August 1978
Reprinted – September 1980
Reprinted – June 1982
Reprinted – January 1984
Reprinted – April 1986
Reprinted – February 1988
Reprinted – April 1989
Reprinted – August 1991

Printed and bound in Great Britain by Mayhew McCrimmon Printers Ltd.

CONTENTS

Page
INTRODUCTION 7
  Conversion Table 8

SECTION 1: "NO SOLDERING" PROJECTS 9
  Theoretical Circuits 11
  Signal Tracer 13
  Multivibrator 14
  "C" Substitution Box 16
  "R" Substitution Box 18
  SCR Controlled Light 18
  One Transistor Intercom 21
  Tuned Ferrite Aerial 23
  Diode Radio 25
  One Transistor Amplifier 31
  Two Transistor Amplifier 32
  Push-Pull Output Stage 34
  Boost Speaker 37

SECTION 2: MISCELLANEOUS DEVICES 39
  Soldering 39
  One Transistor Audio Oscillator 40
  Multivibrator 41
  Glissade 43
  Morse Oscillator for Phones 45
  Two Door Siren 46
  UJT Oscillator 48
  UJT Continuity Tester 49
  UJT High Voltage Producer 50
  Signal Blinker 51
  Mains Wiring 52
  Dimmer/Heat Control 53
  Two Power 55
  Sensitive Touch Switch 56
  Water Level Indicator 58
  Optical Isolator 60
  LED Indicator 61
  "Magic Candle" 62
  "Nerve Tester" 63

Page
Burglar Alarm 64
Transistor Tester 65
Treasure Locator 67

SECTION 3: RADIO AND AUDIO FREQUENCY 71
MW Booster 71
Midget MW Diode Radio/Tuner 74
One Transistor VHF Receiver/Converter 75
Morse and Sideband Resolver 78
Easy Preamplifier 81
741 Preamplifier 83
Tone Control for Your Discs Etc. 85
Two Channel Mixer 86
¼ Watt IC Amplifier 88
Class A Amplifier 90
Two Stage Amplifier 92

SECTION 4: POWER SUPPLIES 95
Easy PSU 95
Easy Regulated PSU 97
Multi-Voltage Regulated PSU 98
One Transistor Adjustable PSU 99
Zener Regulated Supply 102
Motor PSU with Reverse 103
Wrong Polarity Protector 104
Transistor Lead-Out Positions 105

EQUIVALENT SEMICONDUCTORS 107

Page 7 & 8 missing --------->

SECTION I
"NO SOLDERING" PROJECTS

With these, connections are generally made with small terminals, or nuts and bolts. By this means components are connected together, and fixed to an insulated board or panel.

A project can be built using one size of bolt and nut throughout, but it will be found convenient to have two or three sizes. Small nuts and bolts can then be used for connections between components, while larger bolts with extra nuts, or terminals, can be used to connect external items, such as headphones.

For one size throughout, 6ba screws or bolts will be satisfactory, about ½ inch long. Only a few are needed for each project, but they are quite inexpensive in 50 or 100 packs. Some extra nuts will prove useful, and also a quantity of 6ba washers. If the insulated board or panel is not too thick, the ½ inch bolts with an extra nut each will do as terminals for connecting external items.

A 4ba bolt is *larger* than 6ba, and a few of these, ½ inch or more long, also with washers and nuts, can be used for terminals, if preferred. It is also possible to obtain terminal heads, so that bolts with nuts can be used as terminals.

The 8ba size is *smaller* than 6ba. These can be of advantage in a project where size is to be kept down, or for positions where several leads come close together (as at a transistor). Bolts, nuts and washers (8ba throughout) can be obtained together.

Naturally all sorts of nuts and bolts, to hand or obtained from constructional toys or old electronic equipment, can be used. Results are not going to depend on the sizes used, which are more a matter of convenience and appearance

There is also considerable latitude in the insulated panels or boards used to carry the components. Some projects need only a few holes in a small board. Here, it is convenient to use 1/16 inch thick or similar paxolin sheet, cutting a piece of

the wanted size and drilling holes for the bolts or terminals. A small hand drill is satisfactory for this purpose. The drills used can be 3/32 inch for 8ba bolts, 1/8 inch for 6ba, and 5/32 inch for 4ba, or any similar sizes giving clearance.

Where drilling is inconvenient, a ready-perforated board must be used. For projects where size is not important, perforated 1/8 inch thick hardboard is suitable – holes are generally at 1 inch intervals, but this is not important. For this type of board, the size of the holes will generally mean that 4ba nuts and bolts will be beeded.

Plain perforated board (that is, without any conductor foil) can be obtained with holes at 0.1 inch or 0.15 inch spacing but these holes are too small and too close together. However, such boards can be used instead of plain unperforated paxolin sheet by drilling the holes larger in just those places where bolts must be placed.

There is also a type of eyelet board, with holes at 5mm spacing and just large enough to take 8ba bolts. This is handy for small projects.

It is also relatively easy to prepare a fully perforated board or panel, which can be used for a number of projects in turn.

For a board up to about 4 x 6 inch paxolin 1/16 inch thick is adequate, with 1/8 inch sheet for larger boards. It is also possible to use 1/8 inch or similar PVC or other transparent sheet, which allows connections to be seen on both sides. Plain hardboard, or thin plywood, is also satisfactory. Hardboard or wood should be given a coat of shellac or similar varnish, before use, to keep out damp.

Where size is not important, the board can be drilled with holes spaced 1 inch from each other throughout. Do *not* mark off a grid with pencil lines, as this may cause leakage between adjacent terminals. An easy method is to mark such a grid on paper, secure this over the board or panel with tape, and drill all the holes. Make these of an appropriate size for the terminals or bolts.

For small items, the board will need to be prepared to suit.

But where a board is going to be a general purpose support for several projects in turn, it can be much larger. It is also handy to support it on two strips of wood, about 1 inch high, so that there is clearance underneath for bolts and connections.

It will be found that neat and efficient projects can be assembled on such boards, and that it is easy to check connections or components, or to make changes for experimental or other purposes.

Theoretical Circuits

A theoretical circuits shows an electronic device in simple form. Capacitors, resistors, transistors and other items are shown by their own symbols. Lines representing connections are drawn from symbol to symbol. The layout of actual components in the device may not be the same as the positions of the symbols in the theoretical circuit.

It will be found that it soon becomes easy to follow a theoretical circuit. An understanding of such circuits develops when they are compared with an actual layout which shows components and the connections.

The Signal Tracer in Fig. 1 is an example of this. The main item here is the transistor, which has emitter (E), base (B) and collector (C) wires. For the transistors shown, these wires come from the base as in the diagram, which is of the underside of the transistor. (Emitter, base and collector wires may be in different positions in other transistors.) It will be seen that the emitter E is connected to negative at the battery, and to a clip on a lentth of flexible wire.

The headphones – or single earpiece – go between collector C and battery positive. A headphone jack, marked P, is used to take a plug for the phones. Generally, an on-off switch would be present in one battery connection, but this is not necessary here as the circuit can be interrupted (to switch "off") by unplugging the phones. Exactly the same results would be obtained if two separate sockets were used, to take two plugs from the headphones.

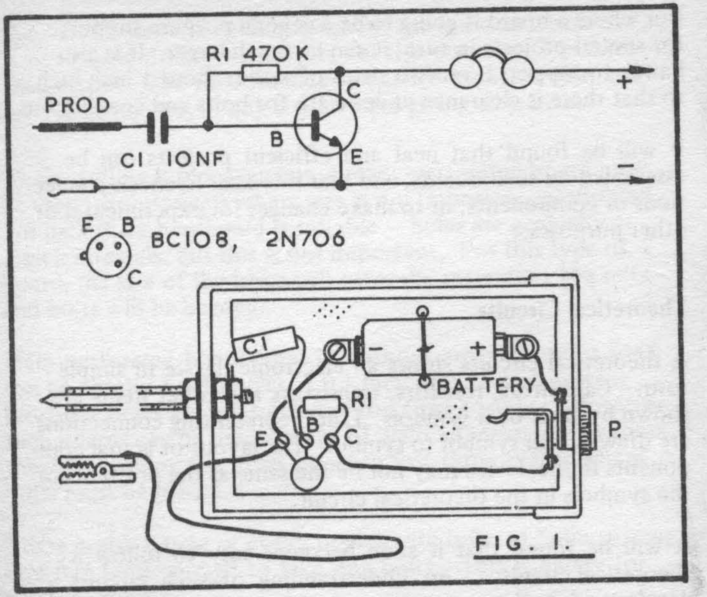

FIG. 1.

Resistor R1 is connected from collector C to base B, and it supplies a base current for the transistor, so that the latter can operate as an amplifier.

The signal to be amplified is obtained through capacitor C1. A capacitor has plates or foils separated from each other by an insulator. There is no direct current circuit through the capacitor, so placing the test prod on various points of a circuit being tested does not upset direct current working conditions.

It is clear, from Fig. 1, how the theoretical circuit is made up as a practical piece of equipment, with C1, R1, transistor, phones and battery connected together in the way required.

With later projects, the theoretical circuit and practical wiring diagram may be shown together in this way, or may appear as separate diagrams when they are larger. If the theoretical circuit and wiring diagram are compared, so that capacitors, resistors and other items on one are also found on the other, it will soon be found that a theoretical circuit is readily followed and understood. It will be noted that the theoretical circuit could be built up in numerous ways, all equally satisfactory.

Signal Tracer

This locates and amplifies audio (sound) signals which may be present in a circuit. Construction is most easily carried out in a small plastic case (Fig. 1). An insulated board about 1¼ x 2 inch (or to suit the case) has three small bolts to secure emitter, base and collector leads of the transistor. Bolts also hold two small brackets, which press on the negative and positive ends of the 9v battery. String or elastic through holes in the board helps to secure the battery.

The prod is a long 6ba or 4ba bolt, or length of screwed rod. Lock one end of C1 lead with nuts, or loop it under the bolt head. Thread the clip flex through a hole in the case, and pass the prod through another hole, as shown. A washer and nut then fix the prod, with the board inside the case. The case lid can then be fitted, and the securing nut can be screwed on the phone outlet. Connections here can be tightly twisted round the outlet tags, and compressed with pliers. A piece of insulating sleeving is put on the test prod, and a clip with fixing screw is attached to the flexible lead.

Making Connections

For all the projects in this book, any thin insulated connecting wire can be used; or bare wire with insulated sleeving. It is handy to have two or more colours – such as red for positive, and black for negative circuits.

A loop is made at the bare end of the wire, and is put under the screw head, or under a washer. Where two or more leads go to the same bolt, it is convenient in some cases to have a washer or two extra, to go between loops. Each connection must be reasonably tight, and short circuits must not be allowed between wires or loops which are shown separated from each other.

Battery and external leads are generally of flexible wire. Positive and negative battery clips, with leads already attached, are available for transistor radio and similar batteries. Elsewhere it may be convenient to arrange contacts as in Fig. 1, or to twist wires to the battery tags.

The battery polarity (positive and negative) *must* always be correct, as shown.

When dealing with transistors, it will be found that some types have long wire leads, more than adequate for easy screwed connections. Others, such as the BC108 and similar types, have rather short wires, but can be held satisfactorily with 8ba bolts fairly near eath other. (With other transistor types not used in this section, soldering may be essential for proper connections to be made.)

Circuit Tracing

To use the signal tracer, connect the clip to the earth return line of the receiver or amplifier. The prod is then touched one by one on points where the audio signal should be present, working systematically through the audio amplifier, or from the first point where the signal can be heard. Take in each capacitor, joint, or stage in turn. When signals cease to be heard, the fault lies between this point and the previous point, where signals were heard.

Do not test any mains operated or high voltage equipment, as this is a job for an engineer who will follow the necessary and essential precautions for safety.

Medium or high impedance phones are most suitable for this tracer. A 600 ohm DC headset will be ideal. The tracer can be utilised as an amplifier for a crystal diode radio receiver

Multivibrator

You will be able to use this as an audio oscillator for numerous test purposes, or for Morse Code practice, or as a signal device in place of a bell or buzzer.

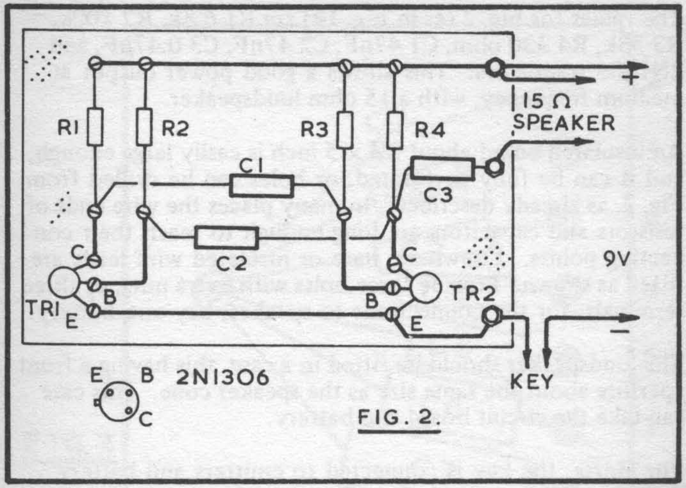

FIG. 2.

Fig. 2 shows the practical assembly of the device on an insulated board. A theoretical circuit is given in Fig. 14, and is not repeated here, so Figs. 2 and 14 should be looked at together.

R1 is the collector load resistor for transistor TR1. Collector current passes through this resistor, and an audio signal voltage is developed across it. Resistor R2 supplies base current for TR1.

TR2 has a similar collector resistor R4, and base resistor R3. Capacitor C1 couples the signal at the collector C of TR1 to the base B of TR2, while C2 similarly couples TR2 collector to TR1 base. Amplification in the transistors and feedback by the capacitors in this way result in the circuit oscillating. That is, each transistor goes rapidly in and out of conduction in turn, so that a voltage at audio frequency is produced across the collector load resistors. C3 is a coupling capacitor, allowing the signal voltage at the collector of TR2 to be taken to a loudspeaker or headphones.

With a circuit of this type, the component values, transistors, and battery voltage will not be critical. However, resistor values, capacitor values, battery voltage, transistors, and the loudspeaker or other sound-reproducing device will all have an influence on the frequency of oscillation, or tone produced.

The values for Fig. 2 (as in Fig. 14) are R1 6.8k, R2 100k, R3 33k, R4 330 ohm, C1 47nF, C2 47nF, C3 0.47μF, and 2N1306 transistors. This allows a good power output at medium frequency, with a 15 ohm loudspeaker.

An insulated board about 2½ x 5 inch is easily large enough, and it can be fully perforated, or holes can be drilled from Fig. 2, as already described. In many places the wire ends of resistors and capacitors are long enough to reach their connecting points. Elsewhere, bare or insulated wire leads are fitted as shown. Provide three bolts with extra nuts, or three terminals, for the connections to speaker, key and battery.

The loudspeaker should be fitted in a case, this having a front aperture about the same size as the speaker cone. This case can take the circuit board and battery.

For Morse, the key is connected to emitters and battery negative as shown, with flexible leads of convenient length. If the multivibrator is to be used as a warning device then these leads can instead go to a door bell push, or to a relay or SCR, used to complete the circuit, as with some later projects. In this way the multivibrator, with its speaker, can replace a buzzer or bell.

If the frequency of the multivibrator is to be changed, this can readily be done by altering the values of capacitors C1 and C2 in particular. The two capacitors need not be of the same value. Larger values reduce the frequency (lower the tone) while smaller capacitors increase the frequency. C3 also has some effect, due to loading from the speaker.

Such a circuit can be used with a battery voltage as low as 1.5v, for headphones. Transistor types are very uncritical. But if PNP transistors are fitted, the battery polarity must be reversed. The transistors shown are NPN.

"C" Substitution Box

When experimenting with circuits or checking a suspected capacitor, a box bringing into circuit a range of capacitor values may be used. In many circuit positions an exactly correct value is not required for testing by substitute, so that a very large number of capacitors need not be provided.

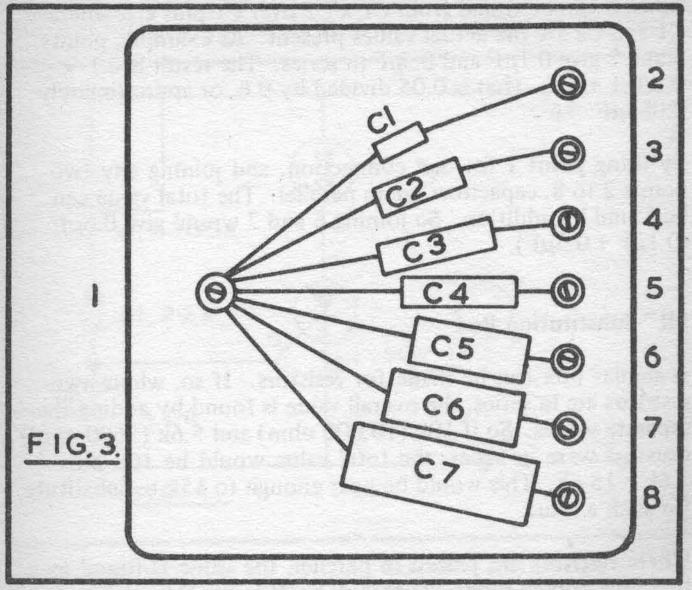

FIG. 3.

Such a box is easily made as in Fig. 3, using a plastic or other insulated case. Here, point 1 is common to all the capacitors, and any capacitor can be selected by connecting to points 2 to 8. Connections can be taken to 6ba bolts, using extra nuts. An alternative is to have eight sockets, with a pair of test leads equipped with plugs.

Suggested values are C1 100pF, C2 470pF, C2 1nF (1000pF or 0.001μF), C4 10nF (0.01μF), C5 0.1μF, C6 0.5μF, and C7 8μF. As the last capacitor will be polarised, mark its terminal to agree.

These values will be found sufficiently near for use in many coupling, decoupling, by-pass and similar circuit positions. The substitution box can be used in place of C1 or C2 in Fig. 2. But where a defective capacitor is found in a tuned circuit, it should be replaced by the exact value.

By ignoring terminal 1 and connecting leads to any pair of the terminals 2 to 8, two capacitors can be used in series. For those who like to work out the value of such combina-

tions, it can be found from C1 x C2 over C1 plus C2, where C1 and C2 are the actual values present. As example, points 6 and 7 give 0.1μF and 0.5μF in series. The result is 0.1 x 0.5/0.1 + 0.5. That is 0.05 divided by 0.6, or approximately 0.083μF.

By using point 1 for one connection, and joining any two points 2 to 8, capacitors are in parallel. The total value can be found by addition. So joining 6 and 7 would give 0.6μF (0.1μF + 0.5μF).

"R" Substitution Box

A similar box can be made for resistors. If so, where two resistors are in series, the overall value is found by adding the separate values. So if 10k (10,000 ohm) and 5.6k (5600 ohm) resistors were in series, the total value would be 10k plus 5.6k = 15.6k. This would be near enough to 15k to substitute for such a value.

Where resistors are placed in parallel, the value is found as for capacitors in series. That is, R1 x R2 over R1 + R2. So for the 10k and 5.6k resistors, this would be approximately 3.6k.

Naturally each resistor may be selected alone, as for the capacitors. If sufficient terminals are provided, a large number of resistors could be fitted. For most transistor circuits, values from about 150 ohm to 270k will fill most needs.

SCR Controlled Light

This device will switch on a light, when the level of general illumination falls below that for which triggering is set. It can thus be used for an automatic night light. Operation can be from a battery; or from a mains power supply, as described later.

Fig. 4 is the circuit, and introduces two new devices. One is the light-dependent resistor LDR. The resistance of this component depends on the strength of the light falling on it. The

resistance is low (some hundreds of ohms) with bright light, but rises to many thousands of ohms in dim light, or darkness.

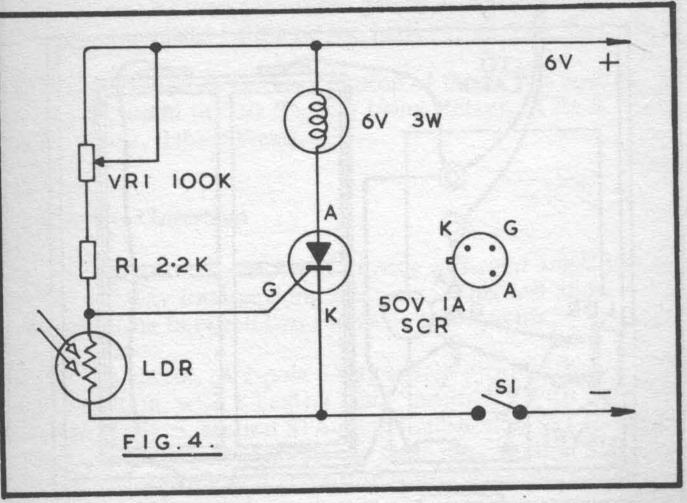

FIG. 4.

The second device is the silicon controlled rectifier SCR. Normally, this does not allow current to pass from its cathode K to its anode A. However, a small current applied to its gate G will trigger it so that it conducts from cathode to anode. It remains in this state even when gate current is removed. It is restored to the non-conducting state by removing the voltage from its cathode, by opening switch S1.

In this device the SCR is controlled by the LDR. With illumination present the resistance from gate to cathode is low, so that when S1 is closed, current does not flow from the anode circuit through the 6 volt 3 watt lamp. As illumination falls, the resistance of the LDR rises. This continues until a sufficient gate potential is present to operate the SCR, which moves into its avalanche or conducting state, lighting the lamp. The potentiometer VR1 enables the resistance from positive to gate to be adjusted, to set the illumination level at which the circuit operates.

A 50v 1 amp SCR is more than adequate for this circuit, and it resembles a large transistor externally, and can be mounted by its wire ends.

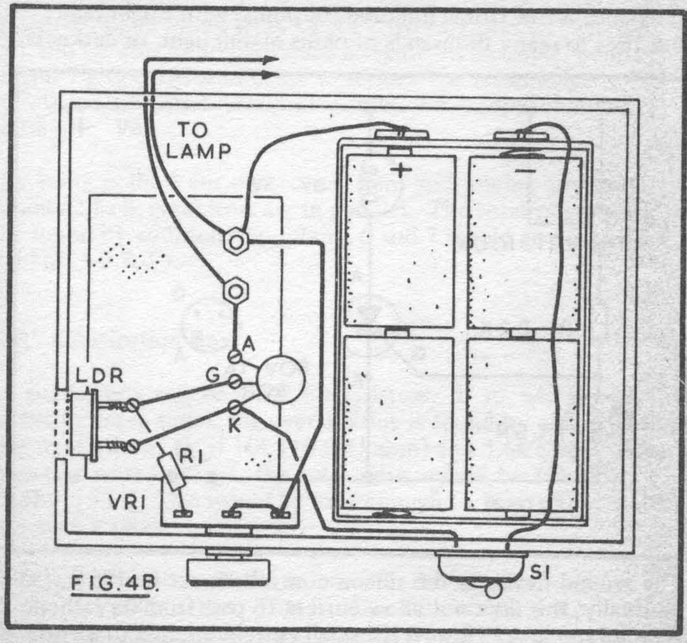

FIG.4B.

Fig. 4B shows constructional and wiring details. The SCR is mounted on a small insulated board. Clear bared wires are twisted tightly to the LDR, so that this also can be fitted by small bolts and nuts, as shown. The connections to VR1 are also tightly twisted and pressed together with pliers.

A battery holder for four 1.5v cells is shown, and connections to this can be by means of clips which have leads attached. The switch S1 is the small, low-voltage type having small screws, as used for toy lighting circuits and similar purposes.

A hole is cut in the box level with the LDR, which may be directed towards the window of the room, or positioned as found to give the results wanted. The bulb, in holder, can be positioned on top of the case. If light from it reaches the LDR this will not stop the circuit working, as the SCR can only be returned to the non-conducting state by opening the switch S1,

For an automatic night light or similar device, a case can be made from thin wood, with opening back or bottom. The device can also be used for various games, where momentarily interrupting the light falling on the LDR switches on the bulb.

If a detailed reference to the operation of the SCR is wanted, this can be found in "50 Projects Using Relays, SCRs & Triacs" (BP37, Babani Press).

One Transistor Intercom

A single high gain transistor will provide sufficient amplification for a 2-way loudspeaking intercom. This will allow communication between two rooms or other points.

Fig. 5 is the circuit. A 2-pole 3-way switch provides 2-way communication, with a central "Off" position. With the switch as in Fig. 5, section S1A takes loudspeaker SP1 to the base coupling capacitor C1. At the same time, section S1B connects the battery and transistor collector circuit to the

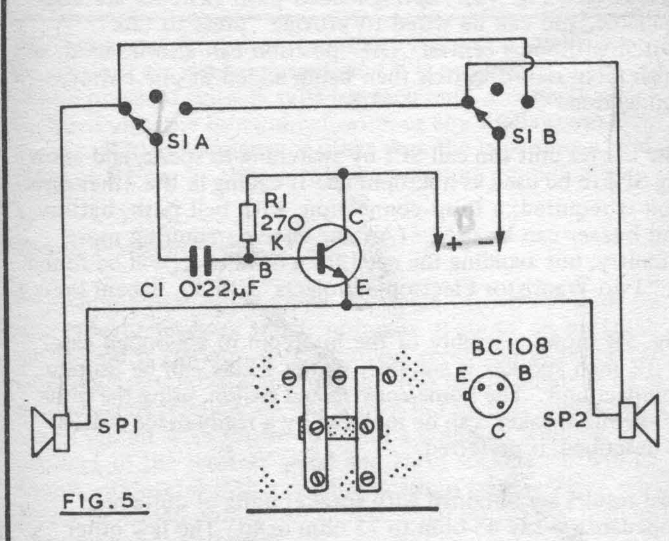

FIG. 5.

speaker SP2. SP1 then acts as a microphone, and sounds are reproduced in SP2.

With the switch in the opposite position, S1A selects SP2 as microphone, and amplified signals pass via S1B to SP1, so that speech can be in the other direction.

The switch, amplifier, battery and SP2 are assembled as the master control unit, and twin leads run to SP1. The person at SP1 should say "Over" so that the other person can operate the switch, and he in turn should say "Over" to signify that SP1 is again acting as microphone.

A simplified home-made change over switch is shown in Fig. 5. This is two strips of metal, pivoted so that they can be swung from side to side, and bearing on two of the three round-headed screws. A strip of insulating material is fitted so that both strips move together. In Fig. 5, this switch is in its "Off" position.

Ready made 2-pole 2-way switches can be obtained and used instead (see Fig. 49). Spring-loaded push switches are also available, and can be wired to provide "press to talk". A switch without a central "Off" position can also be used, a single pole on-off switch then being added in one battery connection.

The master unit can call SP1 by switching to speak, and allowing SP2 to be used as microphone. If calling in the other direction is required, a third connection, with bell push, battery and buzzer can be used. (An alternative, requiring more circuitry, but avoiding the need for a third wire, will be found in "Two Transistor Electronic Projects", (BP30, Babani Press.)

Fig. 5B shows assembly of the intercom in a wooden case. A 3½ inch speaker is suitable, so the whole can be quite a compact unit. The home-constructed switch, using the bolts below the speaker, can be replaced by a ready-made switch, as described, if preferred.

Best results are obtained with speaker units of quite high impedance – say 45 ohm to 75 ohm or so. The few other components are assembled on a small insulated board, as for earlier circuits.

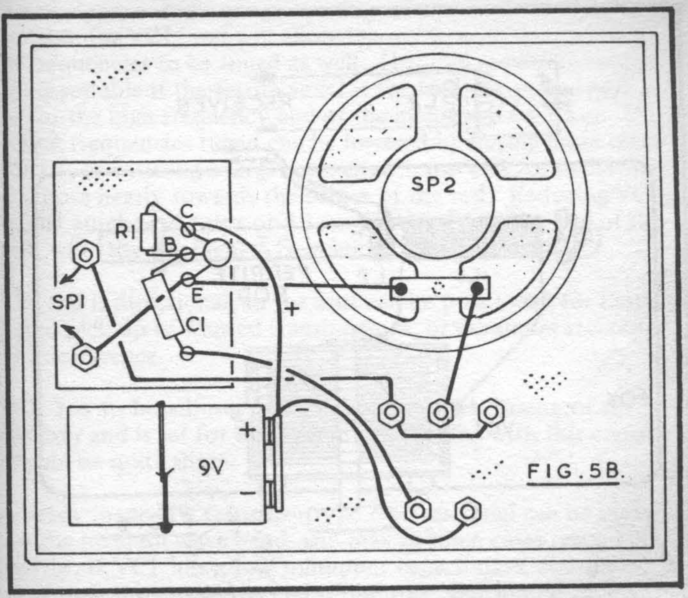

FIG. 5B.

The speaker at the remote point also requires a small case. Dimensions are not important, and it is possible to utilise plastic boxes of suitable type for each unit. Connections between units are by twin bell wire, or any similar twin conductor. With a transistor of other than the type shown, the value of R1 can be chosen to give maximum amplification.

Tuned Ferrite Aerial

This simple device will allow "portable" reception with the old type of receiver which normally requires an external aerial, and it can also give improved reception in terms of freedom from interference on the 160 metre amateur band and adjacent frequencies. As shown in Fig. 6, it consists of a tuned circuit with ferrite rod, and a coupling winding to connect to the receiver aerial and earth sockets. It will be found very useful with a communications type receiver, where a long extended outdoor or similar aerial results in a high level of noise and interference.

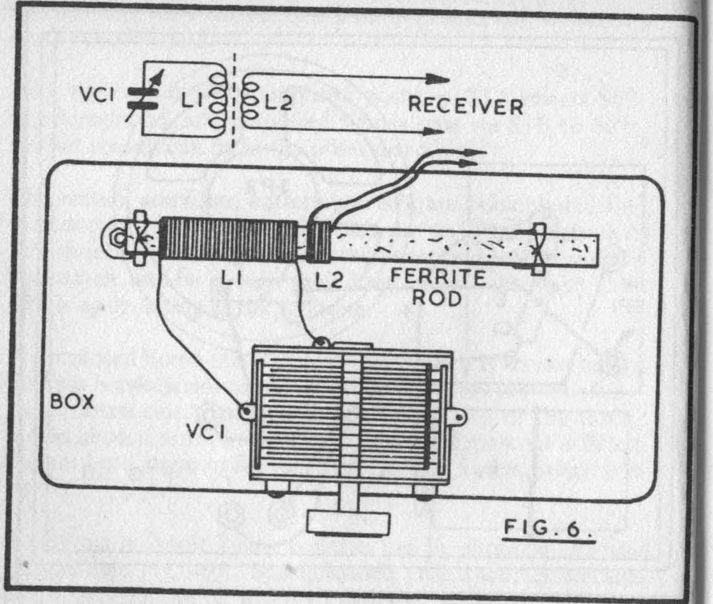

FIG. 6.

The device is constructed in a plastic lunch box about 6 x 4 x 4 inch or other *insulated* case. Some boxes of this kind, though strong, are of brittle material. These must be drilled without too much pressure, or they may crack. Drill points should also be sharp.

The rod is 6 inch long and 3/8 inch in diameter, and it is supported by two strips of insulating material, which are fixed with brackets. Each strip has a V-shaped notch, so that the rod can be held with thin string. L1 is the tuned winding, and it had sixty turns of 24 swg enamelled wire, side by side. Connections can be made to the frame (moving plates) and fixed plates of the capacitor VC1 by means of small bolts, or may be lightly twisted.

L2 is the coupling winding. Many communications type receivers have a low impedance input circuit, and twelve turns will be adequate for L2. The same wire can be used as for L1. With other receivers it may be worth while to experiment with the number of turns on L2, for best reception.

An air-spaced single gang capacitor of about 315pF is most suitable for VC1, and will allow the usual medium wave band of frequencies to be tuned as well. However, a smaller capacitor is suitable if the ferrite aerial is for 160m reception only, or for the high frequency end of the medium wave band. The actual frequencies tuned can be lowered by having more turns for L1, or by using a larger capacitance for VC1, or by having L1 more nearly towards the centre of the rod. Reducing VC1, or the number of turns on L1, or moving L1 to the end of the rod, will raise the highest frequency tuned.

The rod is directional, so the unit can be positioned for best signal pick up of wanted transmissions, or minimum reception of interference.

VC1 has to be adjusted in conjunction with tuning of the receiver and is set for best reception. Tuning with this control should be quite sharp.

A ready made MW transistor type receiver aerial can be useful for the medium wave band, and may in some cases reach 160 metres, if VC1 has a low minimum capacitance, and the winding is quite near the end of the rod. But should tuning not reach a high enough frequency, some turns will have to be removed.

Diode Radio

A lot of interest can be obtained from a diode radio receiver. It requires few components, no battery or other supply, and can give good headphone reception of a few of the more powerful transmitters. It can also be used as a tuner, as shown later, and will provide excellent audio quality for local station reception.

The circuit in Fig. 7 is intended for medium wave reception. That is, from about 200 metres to 500 metres, or 1500 kilohertz to 600 kilohertz.

The coil L1 is tuned by the variable capacitor VC1. At resonance with the wanted transmission, a signal voltage is developed across the coil. Detection or demodulation by the diode D1 allows the audio component of the signal to be

recovered, and taken to the phones or to an amplifier. Such "crystal sets" were used in very many homes, in the early days of radio.

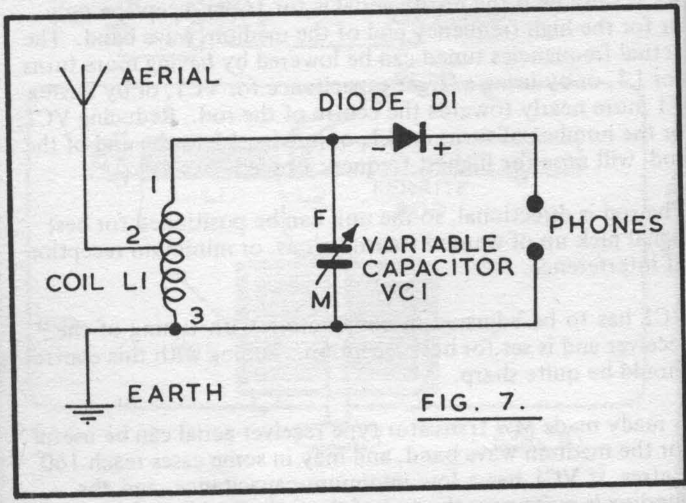

FIG. 7.

Fig. 7B shows the practical layout. It is convenient to use an insulated board larger than necessary, so that a transistor amplifier can be added later. A board about 5 x 8 inch is suggested, and it can be supported on 1 inch high runners (strips of wood) so that there is clearance for screws or terminals.

Coil L1 is wound with 26swg or similar wire, and has seventy turns, side by side. Secure the wire by passing it through two small holes, wind on about thirty five turns, and form a loop. Continue to wind in the same direction for a further thirty five turns, cut off the wire, and secure the end through two holes. The tube used can be 1½ inch to 2 inch in diameter, and can be paxolin, or can be made by winding thin card round a suitable object, smearing a little adhesive on to form a tube.

Point 1, the beginning of the coil, is taken to the fixed plates tag F of the variable capacitor, which is also connected to diode D1. The tapping 2 goes to the aerial terminal. End 3 is taken to the moving plates connection or metal frame of the capacitor. Ends 1 and 2, and loop 3, are left long enough to reach these points, and the insulation is carefully scraped off the wire where connections are to be made. The capacitor VC1 can best be about 500pF, but this exact value is not essential.

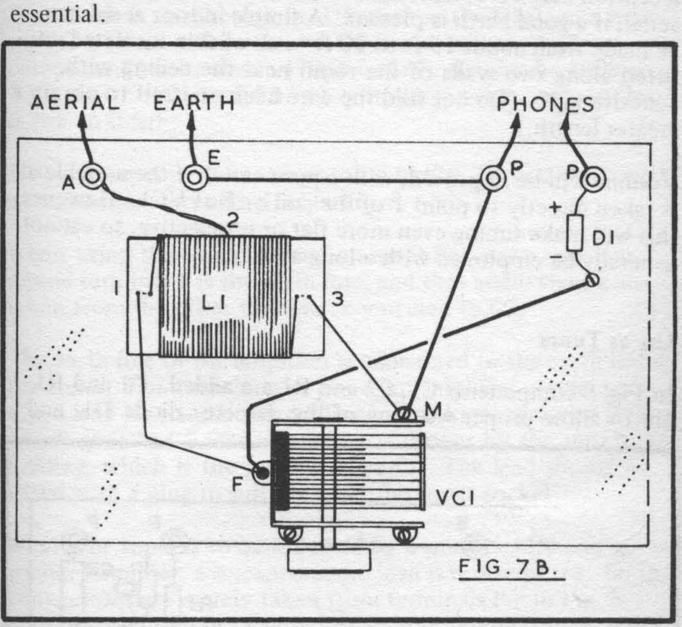

FIG.7B.

Four terminals, for Aerial, Earth, and Phones, will complete the receiver. Best results will be had by connecting a good headset, of about 500 ohm to 2,000 ohm. Very low impedance earpieces, such as are employed with some transistor receivers, are not suitable here.

With a simple receiver of this kind it is necessary to employ a reasonably effective aerial and earth. Only in some circumstances will volume be great enough with no earth, or with a short aerial. The earth connection can be taken to a metal spike in the ground, or to a metal object buried in soil, or to any other available earth other than a mains circuit earth. (No connection must be made to any mains plug or other mains circuit.) A cold water pipe may provide a good earth. Gas pipes should not be used.

For the aerial, a length of aerial wire some 30 ft to 75 ft or so in all may form the aerial itself and lead-in. If this can be suspended quite high, and clear of buildings or earthed objects, signal strength will be improved. However, satisfactory reception may be obtained from an indoor aerial, or shorter aerial, if a good earth is present. A simple indoor aerial can be made from about 15 ft to 20 ft or so of thin insulated wire, fitted along two walls of the room near the ceiling with adhesive tape. (Do not fold the wire back on itself to obtain a greater length.)

Volume will be improved, with a poor aerial, if the aerial lead is taken directly to point 1 of the coil or F at VC1. However, this will make tuning even more flat or unselective, so cannot generally be employed with a long aerial.

Use as Tuner

In Fig. 8 components C1, C2 and R1 are added. C1 and R1 are to allow proper working of the detector diode D1, and may be considered as replacing the headphones. C2 is a coupling capacitor passing audio signals to the following amplifier.

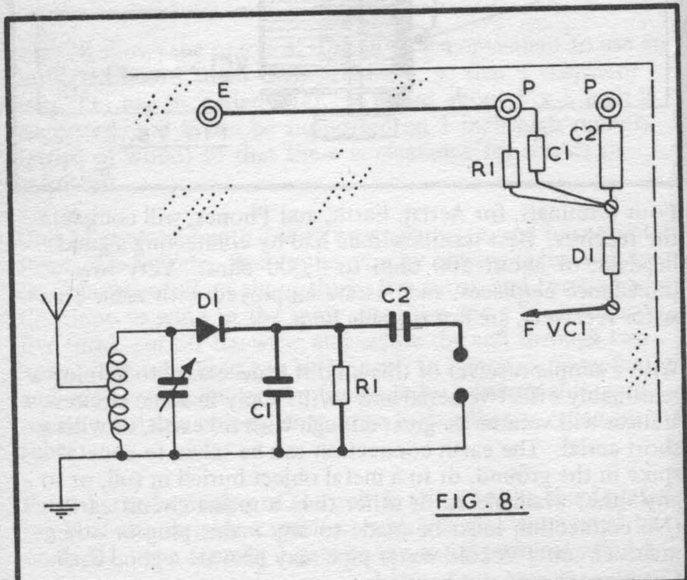

FIG. 8.

As with the previous circuit, any general purpose detection type diode is satisfactory for D1, such as the OA91 and many similar devices. The value of C1 is not critical, and this component can be about 470pF or 1nF. R1 can be 100k to 220k or so. C2 should be of quite large value, and can be $0.1\mu F$ to $0.5\mu F$.

D1 can be moved from the position in Fig. 7B, and the few extra components added, as in Fig. 8.

When using the tuner to feed an amplifier, note that one phone terminal P is the earth line, and that audio signals are taken from the other terminal, connected to C2.

The earth line of the amplifier is connected to the earth line of the tuner. For a sensitive or powerful amplifier, a screened audio lead is recommended. This has an inner screened lead, which goes to C2, and a conductor former by the outer brading, which is the ground or earth. The lead should be fitted with a plug to suit the amplifier input socket.

Where the tuner is to be followed by a small headphone or similar amplifier, a screened audio lead is not required. So then connections are merely taken from terminals P-P in Fig. 8, remembering to join amplifier earth and tuner earth circuits as explained.

Medium and Long Wave

The addition of long wave coverage is worthwhile in some areas, and the coil can be modified as in Fig. 9 so that both MW and LW can be tuned.

A tube about 3 inch long and approximately 1½ inch in diameter will be necessary. The section L1 is wound as before. L2 consists of about two hundred turns of 34swg or 36swg wire, in two compact piles of one hundred turns each. All turns throughout are in the same direction. The junction of L1 and L2 goes to the waveband switch S1. The second

switch terminal goes to 4 and earth line. MW tuning is obtained with S1 closed, and LW tuning with S1 open.

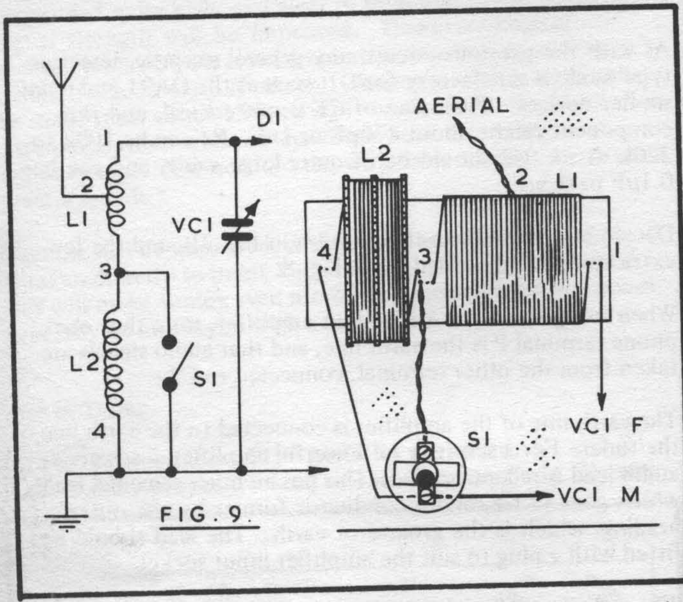

FIG. 9.

For accurate coverage of the usual MW and LW bands, and using a 1¼ inch diameter tube, L1 can be eighty-four turns of 32swg enamelled wire, with tapping 2 sixty turns from point 1. Leave a space of 5/8 inch and glue on a card disc cut to fit the tube. Also glue on two further discs, leaving spaces of ¼ inch between them. L2 is wound with 32swg double cotton covered wire, with one hundred and twenty turns in each slot (making 240 in all for L2). The coil can be fixed with two screws or small bolts, with spacers or extra nuts to hold it a little clear of the board. The windings should not be varnished, covered with adhesive, or wax.

With circuits of this kind there is no ganged tuning, or need for exact tuning of a particular frequency band. This means that coil winding details (wire gauge and covering, and diameter, or number of turns) can be varied to suit materials to hand. Ready-made coils can also be tried. Some commercially made coils with cores can give a slightly better performance than the air-cored home made type of coil. Coils can also be wound in a similar way to that in Fig. 6.

One Transistor Amplifier

Headphone volume obtained from the crystal diode receiver can be considerably boosted by adding a single transistor audio amplifier. Space for this can be found on the circuit board.

In Fig. 10, resistor R2 and the transistor have been added, with a switch S2 for on-off switching of the battery. R1, C1, D1 and C2 are already present, as shown in Fig. 8.

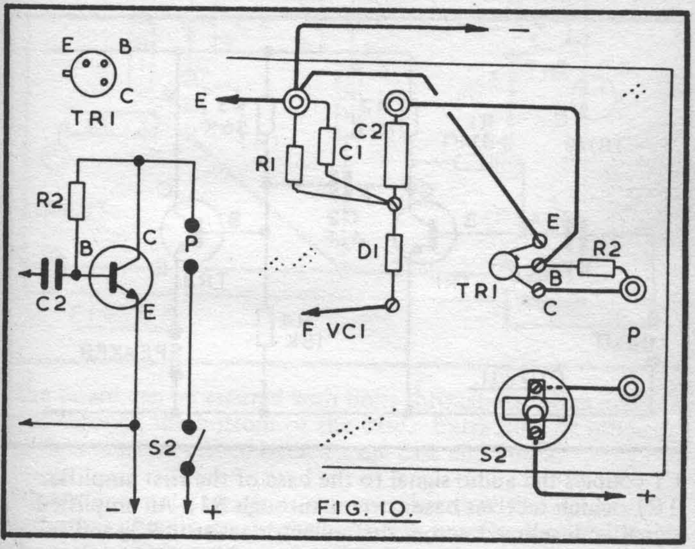

FIG. 10.

TR1, R2 and S2 are connected as shown. For TR1, a BC108, 2N706, or similar general purpose NPN transistor will be satisfactory. R2 can generally be 1 megohm with satisfactory results, but it can be worth while trying values from about 100k to 2.2 megohm here, as performance depends on the transistor, phones, and battery voltage.

Medium and high impedance phones, such as a 600 ohm DC headset, will be most suitable. The battery can be 4.5v to 9v.

If the phones are first connected to terminals P-P in Fig. 8, then taken to terminals P in Fig. 10, a very considerable increase in volume should be obtained.

If a PNP transistor is to be tried in this circuit, the polarity of the battery *must* be reversed.

Two Transistor Amplifier

Fig. 11 is the circuit for an easily made 2-transistor amplifier which can provide reasonable loudspeaker volume. The audio signal is taken to the input terminals, battery negative being the grounded side of this circuit.

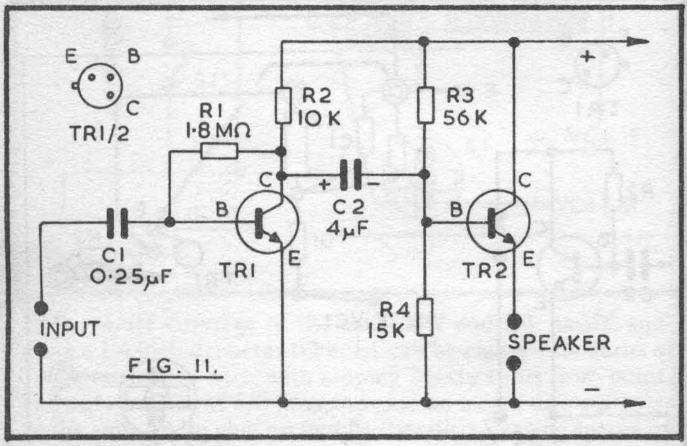

FIG. 11.

C1 couples the audio signal to the base of the first amplifier TR1, which receives base current through R1. An amplified signal is developed across the collector resistor R2, and is coupled to the base of the following amplifier TR2 by means of C2.

The two resistors R3 and R4 set the base operating conditions for TR2, and the output is taken from TR2 emitter.

Fig. 11B shows assembly of the components on a small insulated board. If preferred, connections between components can be on the underside of the board, in the way shown for later circuits. Fit red and black flexible leads for battery connections. The amplifier, with its battery, can be fixed in an insulated box or metal case. Sockets can then be used for input and output circuits, and an on-off switch can be placed in one battery lead.

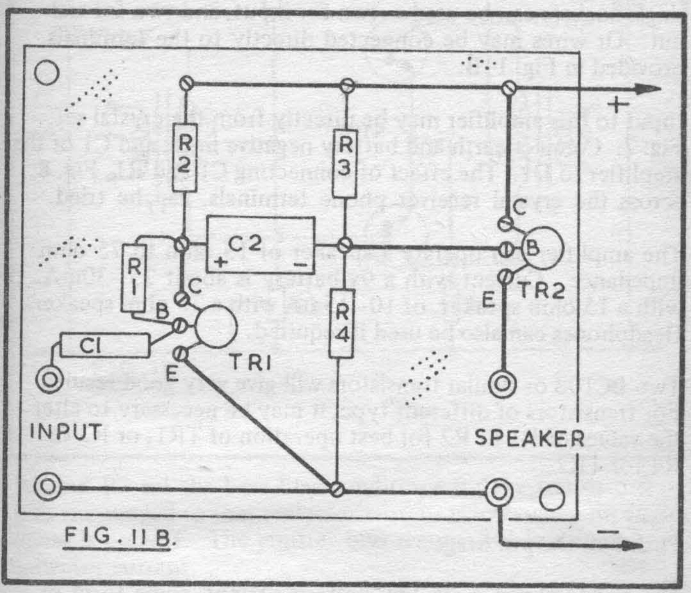

FIG. 11 B.

The board can be secured with bolts through its fixing holes, and through the bottom of the case. Extra nuts or other spacers must be placed between the case and board.

Whenever a metal box is used, remember that all connections and other circuits except the earth line must be insulated from the metal. The metal case itself is normally connected to the earth line, and a lead can be put under one of the fixing screws for this purpose. In Fig. 11, both input and speaker are returned to the earth line. So 3.5mm jack sockets of the type in which the outer or sleeve contact is common to the fixing bush could be used, as these points will be connected to battery negative and the earth line and metal box. But with some circuits it is necessary to use an insulated jack socket, or to insulate this from the metal. An example of this would

arise if the speaker were placed in the collector circuit of TR2, as the outer or sleeve circuit would then need to go to battery positive.

Instead of the type of socket which takes a jack plug, ordinary single sockets can be used – two for input, and two for output. Or wires may be connected directly to the terminals provided in Fig. 11B.

Input to this amplifier may be directly from the crystal set, Fig. 7. Connect earth and battery negative lines; and C1 of the amplifier to D1. The effect of connecting C1 and R1, Fig. 8, across the crystal receiver phone terminals, can be tried.

The amplifier can operate a speaker of 15 ohm to 75 ohm impedance. Current with a 9v battery is about 25–30mA with a 15 ohm speaker, or 10–15mA with a 75 ohm speaker. Headphones can also be used if required.

Two BC108 or similar transistors will give very good results. For transistors of different type, it may be necessary to alter the values of R1 or R2 for best operation of TR1, or R3 or R4 for TR2.

Push-Pull Output Stage

For good volume, with low battery current, some form of push-pull output stage is generally used. This has two transistors, and each amplifies one half of the audio signal. Base bias conditions of the two transistors are so arranged that only a low current flows during silent intervals. Overall battery drain is thus much lower than if a single transistor were used to obtain a similar level of volume.

Fig. 12 is a typical push-pull output stage. T1 is the input or driver transformer, and an audio signal is taken to its primary P. The secondary S is centre tapped. So for one half of the audio cycle one transistor base is driven positive and the other transistor base is driven negative, this being reversed for the remaining half cycle. Thus each transistor conducts alternately, and the centre-tapped output transformer T2 combines the collector currents, to operate the speaker connected to the secondary S.

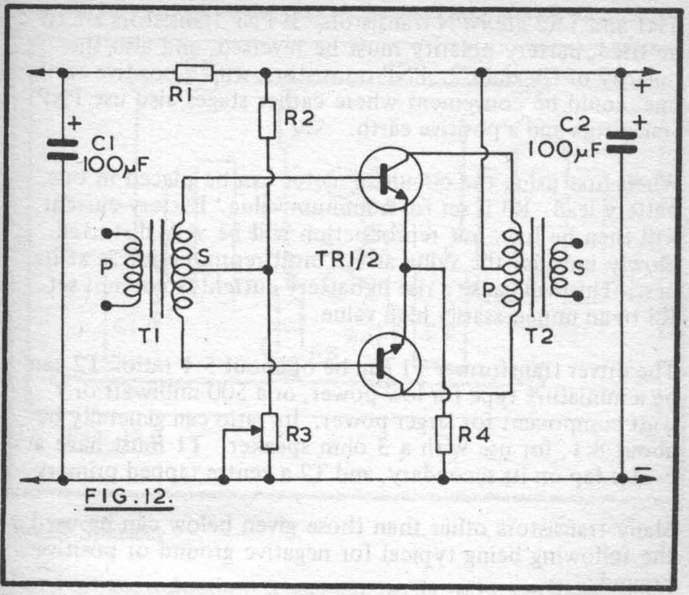

FIG. 12.

R2 and R3 set the base bias conditions for the transistors. Bias is arranged so that collector current is low, when no audio signal is present. The emitter bias resistor R4 limits peak collector current.

If R3 is a pre-set resistor, as in Fig. 12, operating conditions can be arranged to suit the transistors fitted. For best results TR1 and TR2 should be a matched pair. This means that they will have similar gain and current characteristics. However, two transistors of the same type number will generally be satisfactory.

It is assumed that the same battery will supply earlier stages of the amplifier, and current for this purpose is taken from C1 and R1. R1 and C1 prevent audio signals present in the positive line passing back to earlier amplifier stages. The by-pass capacitor C2 also helps in this respect. With simple circuits not giving much overall amplification, or small power outputs, R1, C1 and C2 might all be omitted.

TR1 and TR2 are NPN transistors. If PNP transistors are to be tried, battery polarity must be reversed, and also the polarity of C1 and C2. PNP transistors, with a positive earth line, could be convenient where earlier stages also use PNP transistors and a positive earth.

When first using the circuit, a meter can be placed in one battery lead. R3 is set for minimum value. Battery current will then be low, but reproduction will be very distorted. Slowly increase the value at R3 until reproduction is at its best. This will cause a rise in battery current, so do not set R3 to an unnecessarily high value.

The driver transformer T1 can be of about 5:1 ratio. T2 can be a miniature type for low power, or a 500 milliwatt or 1 watt component for larger power. Its ratio can generally be about 8:1, for use with a 3 ohm speaker. T1 must have a centre tap on its secondary, and T2 a centre tapped primary.

Many transistors other than those given below can be used, the following being typical for negative ground or positive ground.

Negative earth (NPN transistors). 2 x BC108. R1 1.2k. R3 100 ohm fixed or 250 ohm pre-set. R4 4.7 ohm.

Positive earth (PNP transistors). 2 x AC128 or 2 x AC142. R2 4.7k. R3 82 ohm or 250 ohm pre-set. R4 4.7 ohm.

R1 will generally be 1k to 2.2k, depending on the earlier amplifier. The primary of T1 can be connected to terminals P in Fig. 10; or alternatively to any audio source where a reasonably strong signal is present.

Where signals are too strong, an audio gain control needs to be fitted. In Fig. 8 this can be done by omitting R1, and disconnecting C2 from D1. Connect the outer tags of a 100k volume control or log pot from D1 to earth line. Take C2 to the centre tag or slider of the potentiometer. Any required audio level can then be taken off by adjusting the control knob.

Fig. 12B shows the board layout.

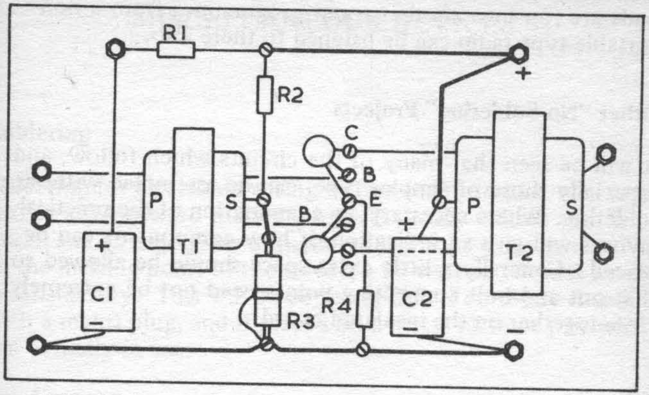

FIG. 12 B.

Boost Speaker

For best reproduction, a speaker needs to be enclosed in a cabinet or fitted to a baffle board. A cabinet can be made from wood, and can be about 6 x 10 inch and 4 inch deep, or as required for the speaker unit. Cut an aperture in the cabinet front to match the cone diameter, and glue thin fabric behind this opening. The speaker can be fixed with nuts and bolts, or short woodscrews driven in from behind.

Connect flexible leads, with separate plugs or a jack plug, to suit the equipment.

Small pocket receivers, which have a miniature speaker, can give much improved volume and reproduction when used with a larger speaker. If a headphone socket, operated from the receiver output stage, is present, it is only necessary to plug the speaker into this. A fairly large speaker, of sensitive type, will give best results. A unit 5 inch to 6 inch in diameter, or about 4 x 7 inch, will be convenient. The speaker has to be of approximately the correct impedance, which should be found in the receiver operating instructions. Where the speaker impedance is wrong, a matching transformer has to be employed between receiver and speaker.

Extension speakers are less frequently used today, but if two leads are run into another room programmes from a non-portable type radio can be listened to there also.

Other "No Soldering" Projects

It will be seen that many of the circuits which follow, and especially those of simpler type, can be assembled without soldering. Where necessary, an examination of the practical layouts will give an indication of how components can be placed. Generally a little extra space should be allowed so that nut and bolt connecting points need not be extremely close together on the insulated board.

SECTION 2
MISCELLANEOUS DEVICES

Soldering

Soldering for electronic wiring will be found to be a very easy and straightforward matter. An electrically heated iron is usual, and as large areas of metal do not have to be raised to the melting point of the solder, a 20 watt to 25 watt iron is satisfactory. This will be perfectly safe if correctly fitted with a mains plug, and if placed on a metal plate or tray when not actually in use.

It is best to use cored solder, as made for electronic wiring.

Leads, tags and other items to be joined should be clean and bright. They have to be raised in temperature to the melting point of the solder at the actual points where a joint is to be made, and the active core of the solder, or flux, must be present here when the joint is being made.

The iron should be tinned – that is, have melted solder on its tip – and have reached its correct temperature. Leads can be hooked or looped into position. The joint is then heated with the iron, and the solder is simultaneously applied to the joint, so that it melts and flows over the surfaces to be soldered. This usually takes place in a second or two. The solder and iron should be removed at once, as excessive heating may damage some components, and especially transistors or other semiconductors. The joint cools in a few seconds, and should be rigid, and solder having flown onto the items to be joined.

A probable cause of poor joints is insufficient heating of items to be soldered, or carrying solder to the joint on the iron, so that the activity of the flux is wasted.

Electronic items such as tags and the wire ends of components are generally designed for easy soldering. They will be copper, tinned copper, and similar metals and alloys which solder readily. For this reason, difficulties are unusual.

Some metals do not solder readily, or need special methods. Aluminium is one of these, and it is not used where popular cored solders intended for copper or general wiring are employed.

Make sure that heat from the iron does not damage furniture or other items, and keep the flexible cord away from the heated tip. Always switch the iron off at once when the work in progress is to be left.

One Transistor Audio Oscillator

This can be used instead of a bell or buzzer warning, or for code practice and similar purposes. It needs very few components, Fig. 13, and can be made to work satisfactorily with a wide range of transistors.

In Fig. 13, an NPN transistor is shown in the theoretical circuit, so battery polarity is as indicated, with negative to emitter. If a PNP transistor is fitted, battery polarity must be reversed.

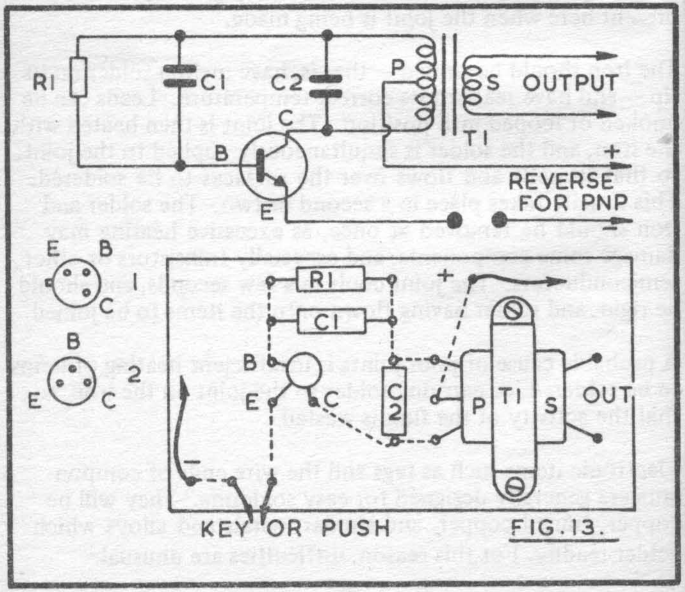

FIG.13.

(1) shows leads from a typical NPN transistor, such as the BC108, 2N706, and similar types. (2) is the base of the PNP AC128 transistor.

T1 can be a small speaker matching transformer, as used for push-pull stages. A centre tapped primary is necessary. The secondary S can be taken to a loudspeaker or headphones.

The values of R1, C1 and C2 can be varied considerably. C1 can be 0.1μF or so. R1 determines to some extent the power obtained, and can be about 22k, increased to 47k or 100k for lower power. The value of C2 has considerable influence on the note, or frequency of oscillation. Increasing the value of C2 lowers the pitch. C2 can be about 47nF to 4.7nF or thereabouts. A 9v battery can be used. Changes to the supply voltage will influence the note produced.

Fig. 13 also shows a layout for components, using either a fully perforated board, or a panel drilled to suit. Transformer T1 is fixed by screws. Small transformers of this type are usually wire ended, or have pins which pass down through holes in the board.

The wire ends of the resistors and capacitors are passed down through suitable holes, and bent over. The board is then turned so that the underside wiring can be carried out. In most places the wire ends of the components will be long enough to reach connecting points. Elsewhere, 24swg tinned copper wire, or similar connecting wire, will be used. Make the soldered joints as described, and snip off unwanted ends of wire. It is as well to leave the transistor to last, as this avoids heating it unnecessarily when soldering other items. Lengths of flex, anchored at holes in the board, are used for battery, key (or push) and speaker connections.

Multivibrator

Fig. 14 is the circuit of the multivibrator assembled in Fig. 2. Reference to Fig. 14B will show how this can also be prepared on a tag board.

The tag board has two rows of tags fitted to an insulated

board, and it is five tags long in Fig. 14B. Boards of this type can be obtained in many sizes. It is also easy to cut a section from a long tag board, to suit the needs of the item to be assembled. Such boards are very handy indeed for electronic assemblies, as the tags support resistors, capacitors and other items, and a neat layout is readily achieved.

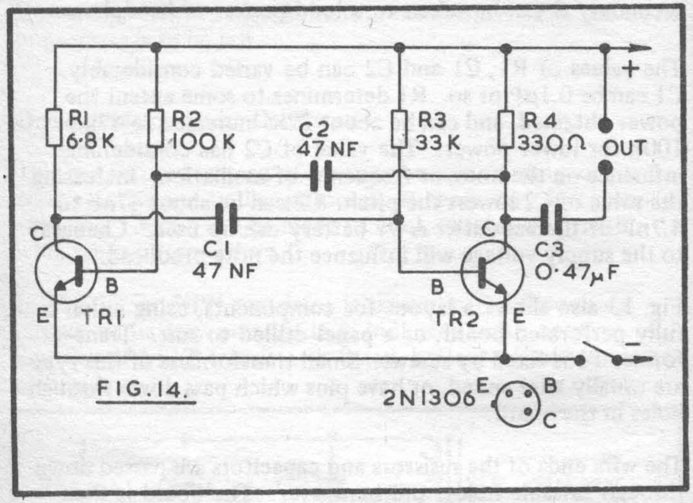

FIG. 14.

Tag strips are also useful for projects, and may have any number of tags required. Long tag strips can be cut down as necessary. Short tag strips have one tag for mounting the strip, longer strips having two tags formed as brackets for this purpose.

Place the components approximately as in Fig. 14B, and solder the wires to the tags. If any component values may be changed, do not twist the leads tightly round the tags before soldering, or removal will be difficult.

Wires which cross can be kept clear of each other, and then no insulated sleeving will be necessary. The transistors can be soldered on last, and can readily be changed without disturbing other parts.

The same layout can be used for the next project, though the component values are different.

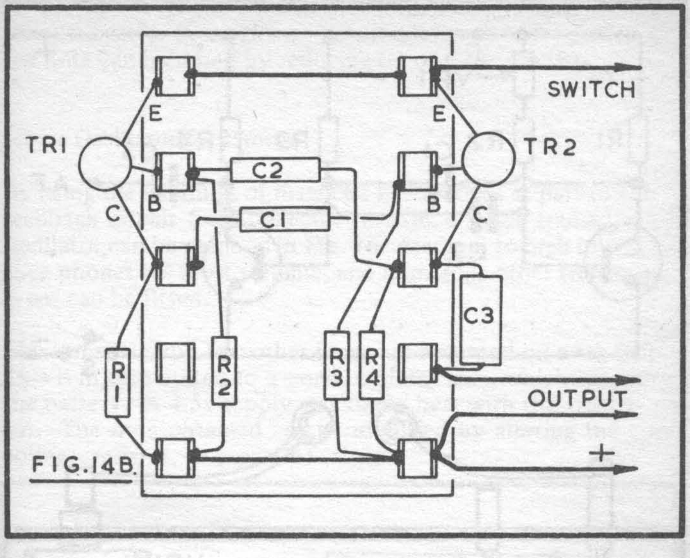

FIG. 14B.

Glissade

This device will produce swooping, rising and falling tones, for "effects" purposes, or can with additional skill be used to obtain musical notes through several octaves. The circuit is shown in Fig. 15, and reference can be made to Fig. 14B for a practical layout on a tag board.

In Fig. 15, suitable component values are:

R1	10k	C1	22nF
R2	12k	C2	22nF
R3	100k	C3	22nF
R4	10k	VR1	100k linear pot.

Transistors: 2 x BC108 or similar types.

The circuit is a wide frequency audio oscillator, in which the pitch can be controlled by VR1, and "on" switching is obtained by S. The oscillator is placed in a box, so that switch S and VR1 can be mounted on the top. The switch is a strip of springy metal, fitted with a spacer and bolt, so that it is just clear of a further bolt. A small downwards movement of the arm or knob completes the circuit.

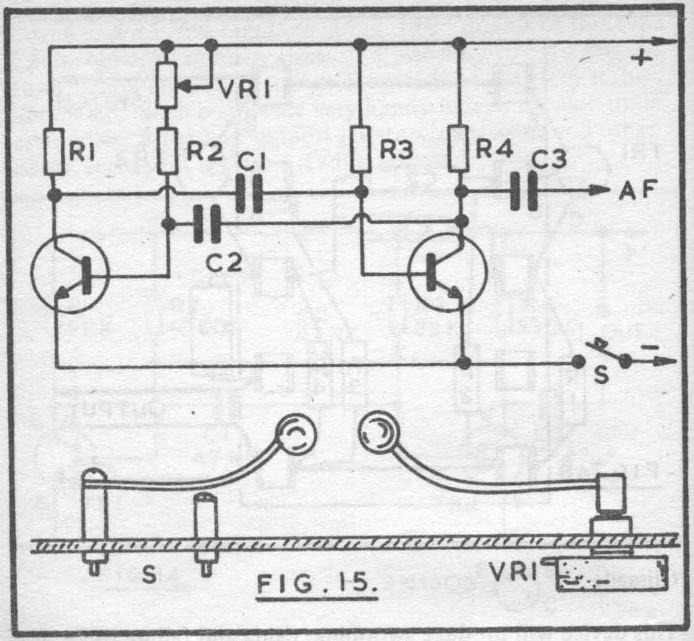

FIG. 15.

VR1 has a ¼ inch shaft coupler, into which is screwed a length of 4ba threaded rod. A knob is also fitted to the end of this rod. It can be swung backwards or forwards, to change the resistance of VR1. Connect VR1 so that movement of the knob towards the right raises the pitch.

Output from C3 can go to high resistance headphones, for amusement, but should be taken to an audio amplifier. Gain of the amplifier will have to be turned back somewhat, as the output from the glissade is at a quite high level.

Swooping tones are produced by holding down the switch S and swinging the control knob attached to VR1 from side to side. For notes which do not swoop to frequency, switch S must be released while VR1 is rotated. A slight quivering motion of the hand controlling VR1 will produce a vibrate effect. It is useful to have an ordinary switch in parallel with the knob control S, and to have the amplifier audio gain control near the left hand, so that fading effects can be produced as well.

About 4.5v to 9v can be used, this influencing volume. Should other transistor types give an unsuitable range of frequencies, the note can be raised by reducing C1 or C2 (or both).

Morse Oscillator for Phones

By using the windings of magnetic headphones as part of the feedback circuit from collector to base, a single transistor oscillator can be made as in Fig. 16. Medium to high impedance phones are most suitable, and numerous other transistor types can be fitted.

The transistor and few other items are mounted on a tag strip. This is in turn bolted to a small insulated box, which can carry the battery. A 4.5v supply was found best with this oscillator. The note obtained can be modified by altering the voltage, resistor, or capacitor values.

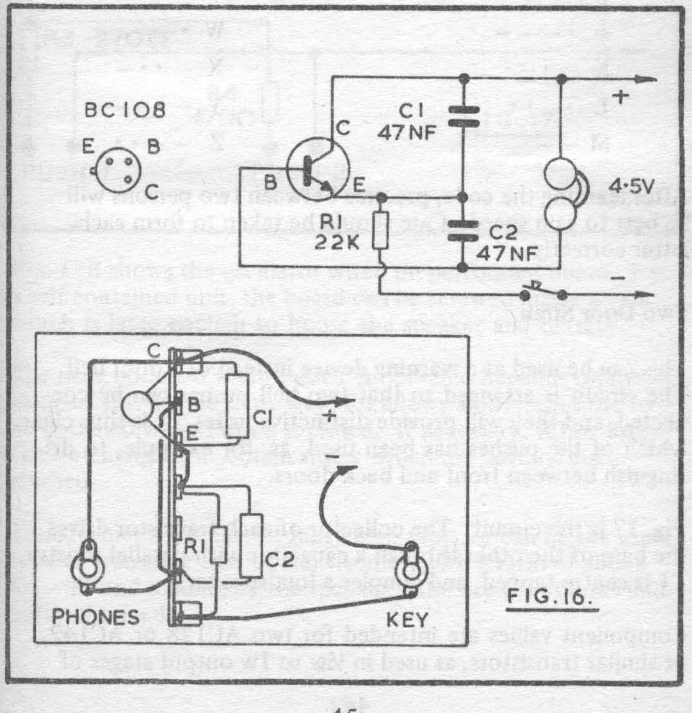

FIG. 16.

For serious code practice a good Morse key is required. It is adjusted so that only a small movement of the wrist is needed to operate it. Each letter consists of short and long sounds, shown as (•) and (–). The (–) has three times the length of the (•)

A •–
B –•••
C –•–•
D –••
E •
F ••–•
G ––•
H ••••
I ••
J •–––
K –•–
L •–••
M ––
N –•
O –––
P •––•
Q ––•–
R •–•
S •••
T –
U ••–
V •••–
W •––
X –••–
Y –•––
Z ––••

After learning the code, practice between two persons will be best to gain speed. Care should be taken to form each letter correctly.

Two Door Siren

This can be used as a warning device instead of a door bell. The circuit is arranged so that two bell pushes can be connected, and they will provide distinctive notes. It is thus clear which of the pushes has been used, as, for example, to distinguish between front and back doors.

Fig. 17 is the circuit. The collector of each transistor drives the base of the other through a capacitor with parallel resistor. T1 is centre tapped, and couples a loudspeaker.

Component values are intended for two AC128 or AC142, or similar transistors, as used in ½w to 1w output stages of

receivers and small amplifiers. Other transistors will of course operate in this circuit. The pitch obtained can be modified by changing C3. The circuit shown is for PNP transistors, and battery polarity has to be reversed for NPN transistors.

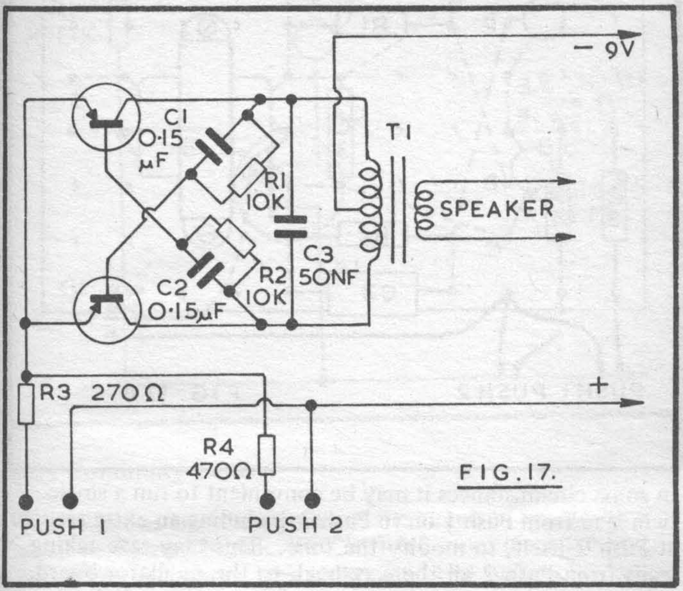

FIG.17.

Fig. 17B shows the oscillator wired on perforated board. For a self contained unit, the board can be screwed inside a case which is large enough to house the speaker and battery.

The note obtained when Push 1 is operated depends on the value of R3. In the same way, R4 allows a different note to be obtained when Push 2 is closed. If necessary, R3 or R4 can be changed, to obtain notes which are readily distinguished.

Transformer T1 is best of fairly high ratio, or can be the usual type of transistor output transformer, with a ratio of about 8:1. T1 and loading by the speaker influence the volume and notes obtained.

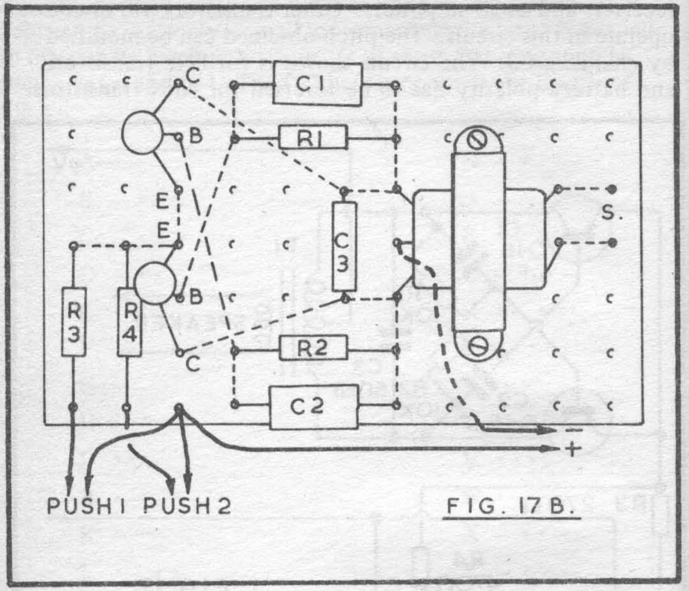

FIG. 17B.

In some circumstances it may be convenient to run a single twin lead from Push 1 on to Push 2, including an extra resistor at Push 2 itself, to modify the tone. This may save taking leads from Push 2 all the way back to the oscillator board.

UJT Oscillator

The uni-junction transistor is a device having emitter, Base 1, and Base 2, Fig. 18. B2 receives positive bias, in this case through the headphones. When forward bias on the emitter E is reached the UJT conducts.

In Fig. 18, the emitter potential rises as C1 charges through the resistor R1. When the UJT conducts, C1 is discharged, and the process is repeated. This can take place over a very wide range of frequencies, depending on the values used for R1 and C1. Here, these are chosen to produce an audio tone.

A battery supply of 3v to 9v will be convenient. For use of the UJT oscillator during code practice, place the key in one battery lead.

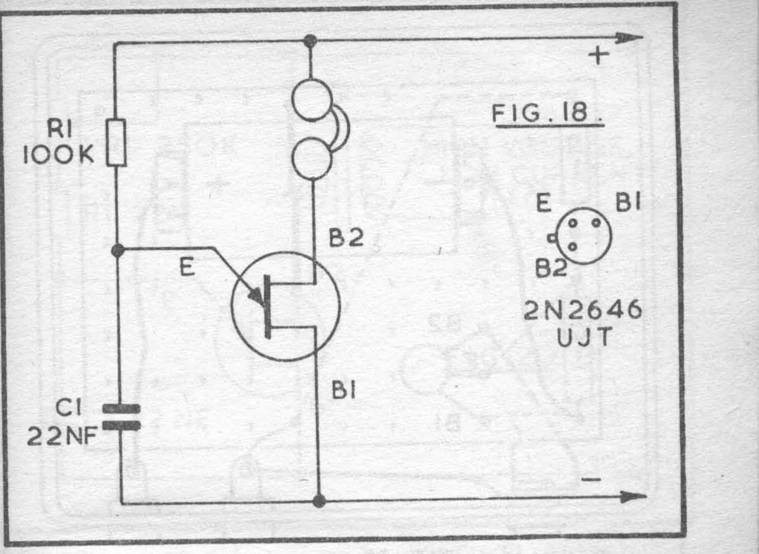

FIG. 18.

UJT Continuity Tester

Fig. 19 shows actual assembly of the circuit in Fig. 18 on a small perforated board. This is fitted in an insulated box, which carries a phone jack socket, and two sockets for Morse key or test leads. A 9v battery is used, secured with string or elastic.

For continuity tests, take prods or clips on the test leads to the suspected item. If the circuit is in order, the UJT oscillator will produce an audio tone. Oscillation will be obtained with a series resistance between the prods of up to 20k or more, so various components can also be tested, and some idea of the resistance in circuit can be gained from the sound obtained. The tester can also be used to check bulbs, lamps, fuses, extension and other leads, and similar items – these must of course be removed or disconnected from the equipment, before being tested.

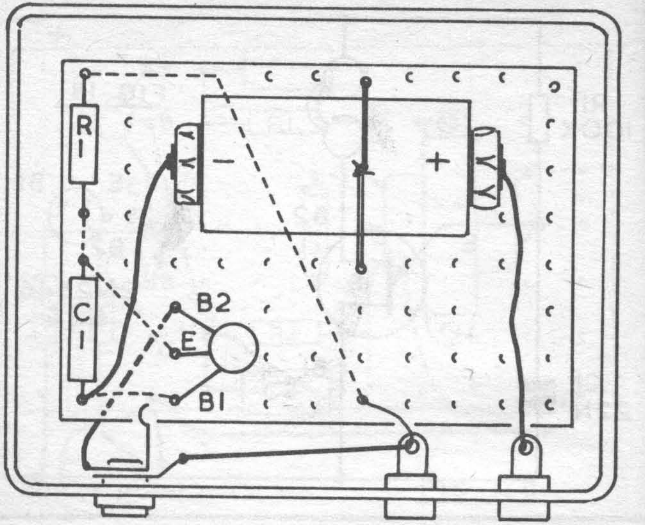

FIG. 19.

UJT High Voltage Producer

The UJT oscillator can be used to produce a high voltage, at very low current, by connecting a transformer as in Fig. 20. A loudspeaker transformer, as would be used in valve equipment, is suitable, and will have a ratio of about 45:1 or so. The usual secondary, or low-resistance winding, is connected to B2 and positive, so that a step-up of voltage is obtained. This should be found to light a small neon lamp.

VR1 is fitted to allow some adjustment, to find the most suitable frequency of oscillation. For maximum possible high voltage output, if required, operating voltage and current should be raised. No rating of the UJT must be exceeded. For the 2N2646, the supply should not exceed 30v, emitter current should not be over 50mA, and dissipation of power in the device should not be more than 300 milliwatts. (Dissipation can be found from Voltage x Current.)

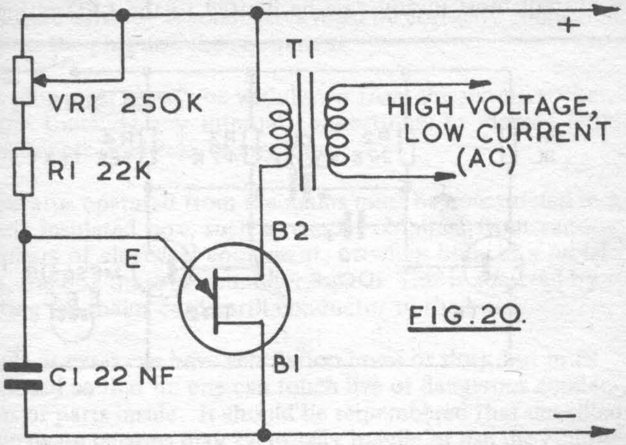

FIG.20.

A small mains transformer can be used for T1, with a low voltage winding, such as that for a 6.3v heater circuit, as primary, and the 200/250v winding as secondary. The high voltage output is alternating current.

Signal Blinker

Fig. 21 is a multivibrator in which the feedback capacitors C1 and C2 are of high value, so that conduction alternates between TR1 and TR2 at a low rate. A light emitting diode is placed in the collector circuit of TR2, and so flashes on and off.

This device can be used for various instruments and other equipment, and a flashing indicator draws attention to a greater extent than does one continuously illuminated. The LED can be located at any required point, with leads running to it.

Transistors other than the types shown can be utilised. R4 can usually be about 150 ohm for a 3v supply, 470 ohm to 1.2k for 6v, or 680 ohm to 1.5k for 9v. This depends to some

extent on the amount of light required when the indicator is operating. Lower resistor values at R4 will increase light, but current must in any case be limited to the LED rating.

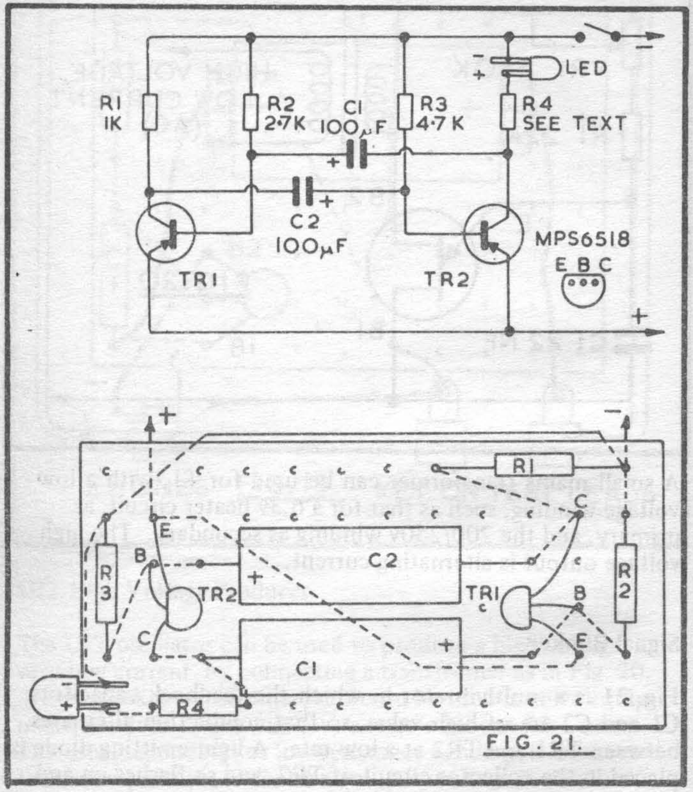

FIG. 21.

C1 and C2 need only be 6.4v working, and the device can be assembled on a very small perforated board. See p.61 for LED polarity.

Mains Wiring

All the devices described so far have operated from low voltage batteries, so than no shock hazard exists with them. However, it is necessary to employ mains power for some units, and proper care must then be taken to avoid any danger.

Power should be drawn from a 3-pin plug, and this should have a 3A or other low rating fuse installed. The live or L, neutral or N, and earth or E conductors must be correctly connected, both at the plug and the equipment.

The plug must always be withdrawn from the supply socket, before touching any internal connections, or making any adjustments or checks of any kind.

Apparatus operated from the mains must be constructed in a strong insulated box, such as can be obtained from various suppliers of electrical equipment; or when built in a metal box, the box must be soundly earthed. This is achieved by bolting the mains cord earth conductor to the box.

Boxes or cases can have ventilation holes or slots, but must be closed so that no one can touch live or dangerous connections or parts inside. It should be remembered that unskilled or unaware persons may eventually handle or use the equipment, so it must be arranged to present no unexpected hazards such as live terminals or parts, or openings for the fingers.

If these points are remembered, mains operated equipment can be safe to use and handle. No switch or other means of interrupting the circuit is put in the earth conductor, and any switch or equivalent device is placed in the live or L conductor. This leaves both earth and neutral permanently connected, for maximum safety.

Dimmer/Heat Control

It is often convenient to dim a filament lamp, or reduce the heat of a soldering iron to be left on for long periods, or the power of a low wattage heater. The silicon controlled rectifier circuit in Fig. 22 can be used for these purposes.

When the switch marked "Full" is closed, current for the load passes through this, so that full power is obtained.

When this switch is opened, current passes through the SCR. This device conducts only when the gate has been triggered, and the instantaneous point in the alternating current cycle

where this arises is determined by the setting of the potentiometer VR1. Triggering is advanced as the wiper of VR1 is moved towards R1. The average power supplied to the load can thus be adjusted between wide limits. Since the device is operating as a half wave controller, full power is not obtained with VR1 at maximum, so the switch is closed as described.

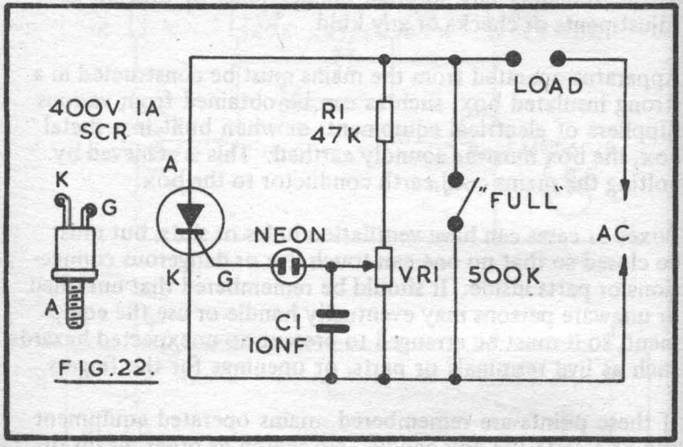

FIG.22.

Fig. 22B shows construction of the SCR controller in a metal box. The front of this carries a socket outlet, and the mains cord has a 3-pin plug, wired as explained earlier. It is thus easy to plug the controller into an outlet, and insert the plug of the lamp or other item to be used in the new outlet.

The mains cord is secured with a clamp, and passes through an insulated grommet. The tag strip supporting the small items is mounted by an unused tag. A strip of metal is cut to clamp round the body of the SCR, and is soldered to one tag, as shown. Run earth and neutral leads from the cord up through a hole, and connect them to the E and N terminals of the socket outlet. Also earth the case, at E. A well insulated wire runs from the L terminal of the outlet, to the "Full" switch. The outlet can be secured with bolts and nuts.

R1 should be 1 watt, and C1 is a 500v capacitor. The neon is a small type as used for indicators, etc.

After checking that the controller operates correctly, the back

of the box should be screwed in place. A suitable size is about 4 x 5 x 2 inch deep.

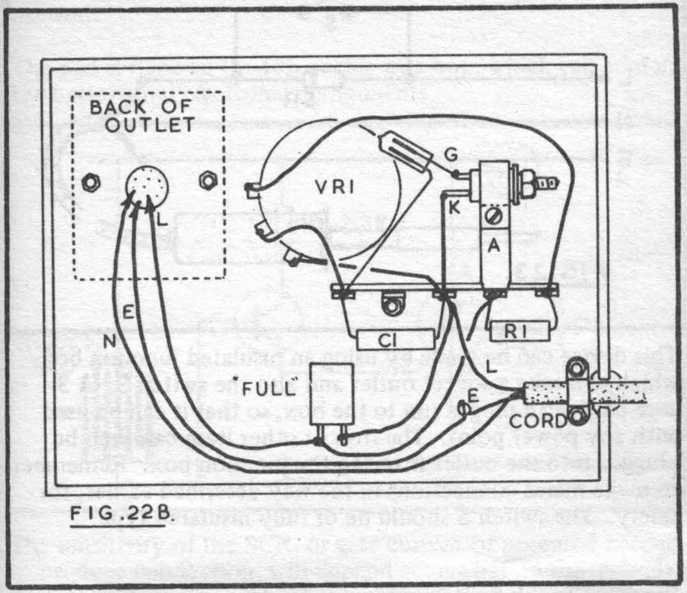

FIG.22B.

If a fully insulated box is used instead, then the bush of VR1 and of the switch should both be earthed by wires, except when these items are constructed wholly with insulated material. (This precaution is necessary to avoid any chance of metal parts outside the box becoming "live".)

Two Power

A circuit which will allow half or full power can be useful to reduce the heat of a soldering iron, or for dimming a filament lamp.

Fig. 23 is an easy circuit of this type. Full power is obtained when the switch S is closed. With S open, the silicon rectifier allows current to flow in one direction only, so that power is approximately halved. A 600v 1A or 1000v 1A rectifier is suitable.

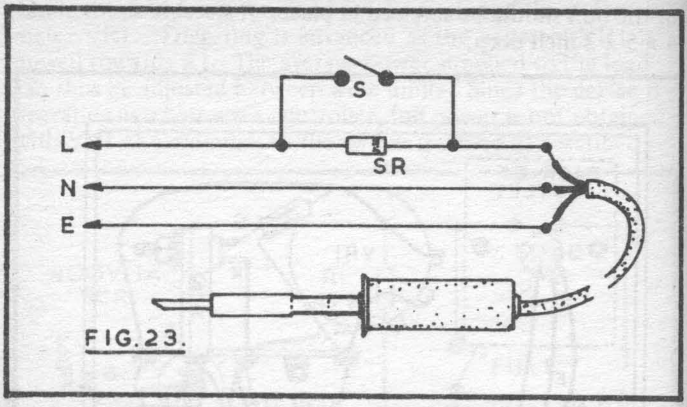

FIG.23.

This device can be made by using an insulated junction box which will take a socket outlet and also the switch S. A 3-core cord with plug is run to the box, so that it can be used with any power point. The iron or other item can then be plugged into the outlet fitted to the junction box. Remember to make mains connections in the way described earlier, for safety. The switch S should be of fully insulated type.

Sensitive Touch Switch

In Fig. 24, finger contact with the touch pad switches on a low voltage bulb. This circuit is suitable for a battery operated portable bedside light, and for some other devices.

Normally the silicon controlled rectifier is not conducting, so the lamp is extinguished. Finger contact with the pad provides positive base bias for the transistor so that it conducts. This moves the transistor emitter and SCR gate positive, triggering the SCR. Current then flows from cathode K to anode A, and through the bulb. The SCR remains in the conducting condition until the battery circuit is interrupted by opening the switch. A push switch, normally closed, is most suitable here. Pressing it then switches off the bulb.

This touch pad is easily made from a piece of circuit board with foil strips. It can be about 2 x 1 inch, or of any convenient size. Join together alternate foils – e.g., 1, 3, 5 and 7 – by soldering a connecting wire through the holes. This will provide one connection to the pad. The other foils – 2, 4 and 6 – are similarly joined together for the other pad connection.

The pad is fixed to the top or side of a box, which can contain the battery and electronic components.

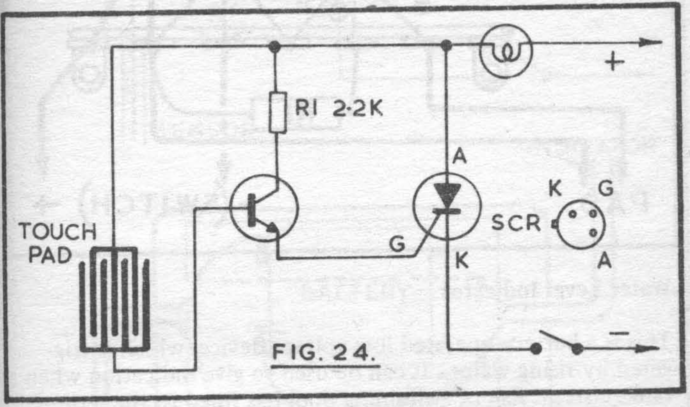

FIG. 24.

The sensitivity of the SCR, or gate current or potential needed to produce conduction, will depend somewhat on the individual device, while the gain of individual transistors will also be different. Numerous NPN types are suitable. If it is found that the switch is too sensitive, this can be corrected by placing a resistor from transistor base to SCR cathode or negative line. This resistor can be about 470k to 1 megohm or so. Excessive sensitivity will be apparent if the circuit is switched on by dust, damp air, or similar effects. The pad should be kept clean by wiping it with a dry cloth.

Fig. 24B shows assembly of the components on a tag strip. A small 50v 1A SCR is more than adequate. The bulb can be a screw in type, and can be 6.3v 0.15A, or 6.3v 0.3A. A 6v 3W (0.5A) bulb is also suitable, provided the battery is fairly large.

It is convenient to make a wooden box which will take the battery, and leave a little space for the tag strip assembly.

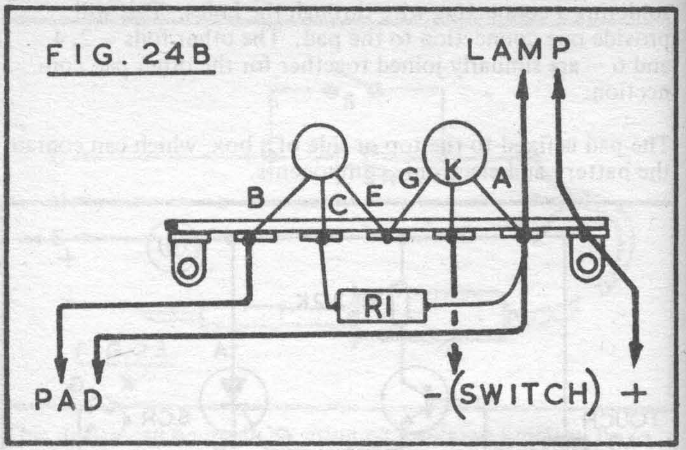

FIG. 24B

Water Level Indicator

This is a battery operated low voltage device, which is triggered by rising water. It can be used to give indication when a tank, cistern, fish or swimming pool has filled to the required level.

Fig. 25 shows the circuit, and also assembly on an insulated board. This also carries the battery, and it is fitted in a metal or plastic box.

In use, the base normally has no positive bias, as the sensor passes no current. When water reaches the sensor, it conducts, and the transistor passes collector current, to operate the indicator. R1 limits base current.

This circuit will operate satisfactorily with a wide range of PNP or NPN transistors, but battery polarity must be reversed for the PNP types.

The indicating or warning device can be one of the multivibrator or other audio oscillators shown earlier. These will give an audible warning, and will draw only a small current, so that a large transistor is not required.

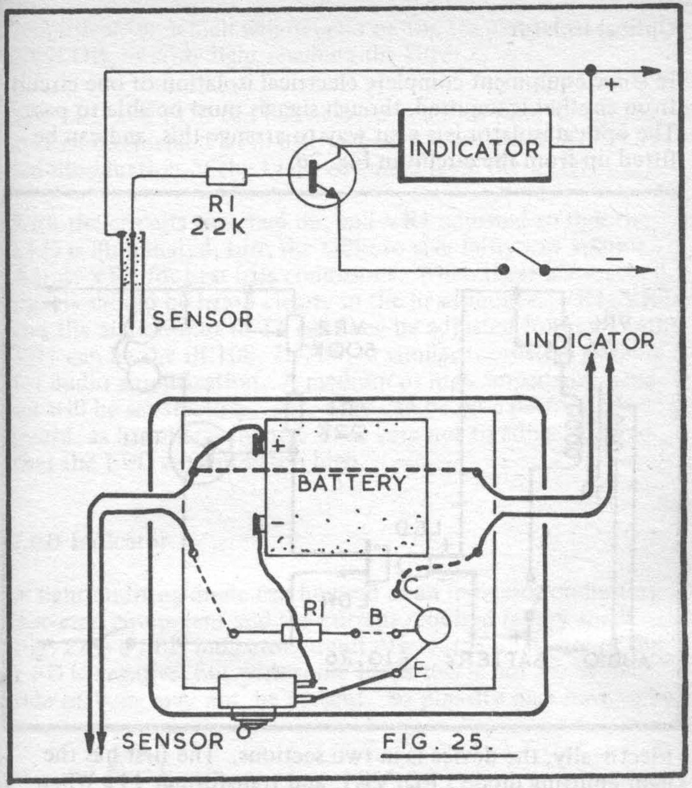

FIG. 25.

For indoor use or in a location protected from rain, the sensor can be two bare wires, about ¼ inch apart, bolted to a strip of insulated material. Where rain could operate the device, it needs to be covered. Alternatively, it can be made from a piece of metal tube, with a central wire, as in Fig. 25. The wire is bound with insulating tape for about half its length, to fit tightly in the tube. Flexible leads are soldered on, and the joints covered with tape. The sensor can be suspended inside the tank or pool, with its bottom at the required water level.

If the transistor is a high gain one, so that damp operates the indicator, sensitivity can be reduced by connecting a 100k resistor from base to emitter.

Optical Isolator

In some equipment complete electrical isolation of one circuit from another is required, though signals must be able to pass. The optical isolator is a neat way to arrange this, and can be fitted up from the circuit in Fig. 26.

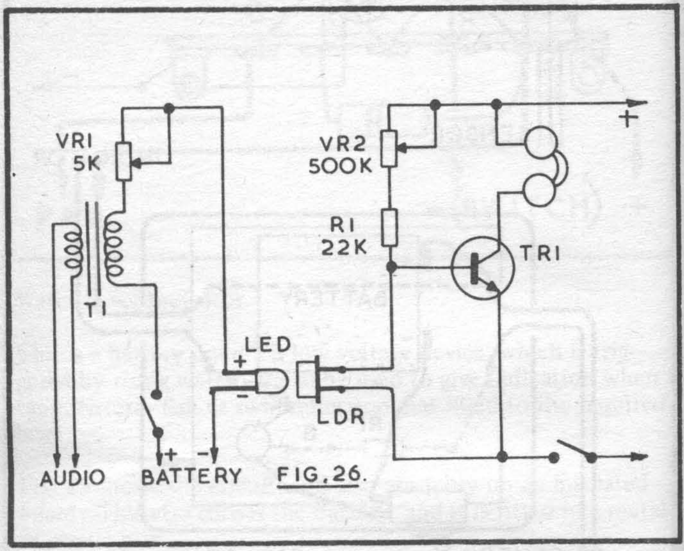

FIG. 26.

Electrically, the device is in two sections. The first has the light emitting diode LED, VR1, and transformer T1. When the battery is switched on, current flowing through the LED is adjusted by VR1, so that the LED glows dimly. A signal voltage in the secondary of T1 then modulates the LED light intensity.

To operate this circuit from a radio having a 2/3 ohm or similar speaker, a speaker matching transformer with a ratio of about 45:1 can be used. (The actual ratio is not very important.) Connect the usual transformer secondary to the radio speaker, so that T1 provides a step up effect.

The second part of the equipment has the light dependent resistor LDR, forming part of the network providing TR1 with base bias. VR2 allows adjustment of conditions for best amplification, which will depend on the transistor, and also the LDR, or stray light reaching the latter.

The LED and LDR can be enclosed in a card tube, or should at least be shielded from light, which will swamp the modulated illumination of the LDR, obtained from the LED.

With the circuits switched on, and VR1 adjusted so that the LED is illuminated, turn the radio to give fairly low volume. Adjust VR2 for best bias conditions. When these are reached, signals should be heard clearly in the headphones. VR1, VR2 and the audio input to T1 can then be adjusted for best results. TR1 can be the BC108, 2N706, or similar transistors suitable for audio amplification. A medium or high impedance headset will be satisfactory. Assembly can be on a perforated board, as for other circuits. Take care not to adjust VR1 so that the LED current is too high.

LED Indicator

A light emitting diode can be used as an indicator on battery powered equipment and the current required is very small. Fig. 27 is a LED indicator circuit. Here, the long wire of the LED is negative, but with some LEDs this is not so. A flat side may, or may not, be present. So polarity may have to be

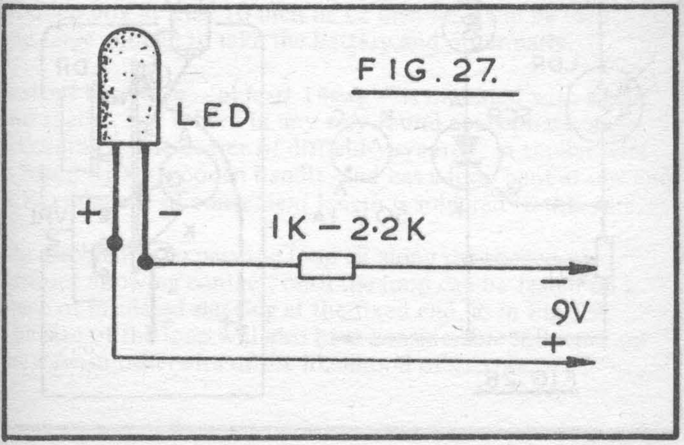

FIG. 27.

found by trial – the correct connections are those allowing the LED to light. The series resistor may be about 150 ohm to 1k or more, depending on the actual voltage, and brightness required. Current drawn varies with different LEDs, and with voltage and series resistance, but can be about 5mA or less, without reducing the brightness too far.

The LED indicator with resistor is connected to the circuit where voltage is present when the equipment is on. LEDs are available in red, yellow, green and other colours, with grommets or clips for panel mounting. Different circuits may be identified by a range of colours.

"Magic Candle"

The triggering of a silicon controlled rectifier can be used for slave photoflash and similar purposes. In Fig. 28 it is employed to switch on a light when a match or lighter is brought near. Sensitivity is best in somewhat subdued light, and it will then be found that striking a match at some distance can bring on the light.

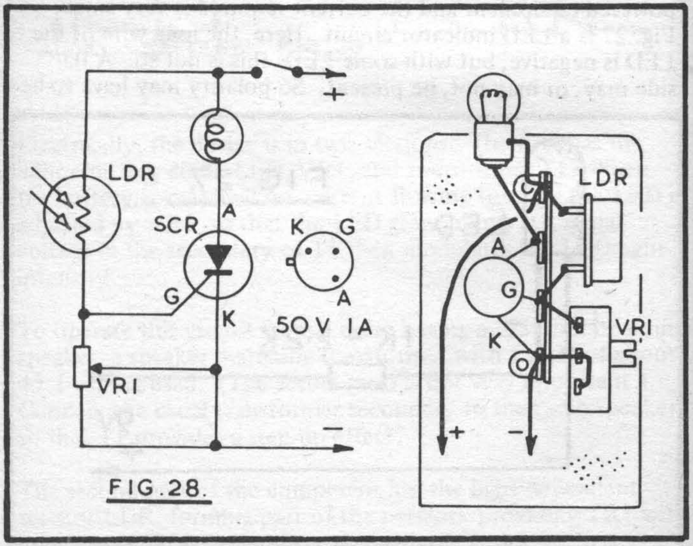

FIG.28.

The LDR has a high resistance when in darkness or subdued light, so that the gate potential of the SCR is not sufficiently high to trigger this device. When light reaches the LDR, its resistance falls, and the SCR gate voltage rises. The SCR then moves into its conducting condition, so that current flows through the bulb. The sensitivity of the circuit, or level of illumination needed to trigger the SCR, is set by VR1.

Components can be assembled on a small tag strip. This, with a battery, can occupy a tubular case, with the lamp on top. An opening is necessary opposite the LDR, and a hole is also made so that VR1 can be adjusted with a screwdriver. Here, a miniature pre-set 50k or 100k linear potentiometer is most suitable. It is necessary to have a small push switch to interrupt the battery current, as the SCR remains in the conducting condition until the circuit is momentarily broken. A 6V battery and bulb can be used. SCRs may fail to operate at all if attempts are made to use them with extremely low voltages, such as a single 1.5v cell.

"Nerve Tester"

This device tests the nerve, concentration, and steadiness of hand, and always provides amusement.

Connections and construction will be seen from Fig. 29. A wooden box at least 10 inch or 12 inch long will be used, and large enough to take the battery and other parts.

A stout bare wire – at least 14swg – is mounted with a bolt and spacer, and is bent in any way found appropriate, and according to the degree of difficulty wanted. A similar wire is fitted with a wooden handle, and has a loop bent at one end. A flexible lead of convenient length is soldered to this wire.

The problem is to pass the loop all along the shaped wire, without allowing contact, until the loop can be rested on a piece of insulated sleeving at the fixed end, as in Fig. 29. The size of the loop will also have considerable influence on the ease or otherwise of the likelihood of success.

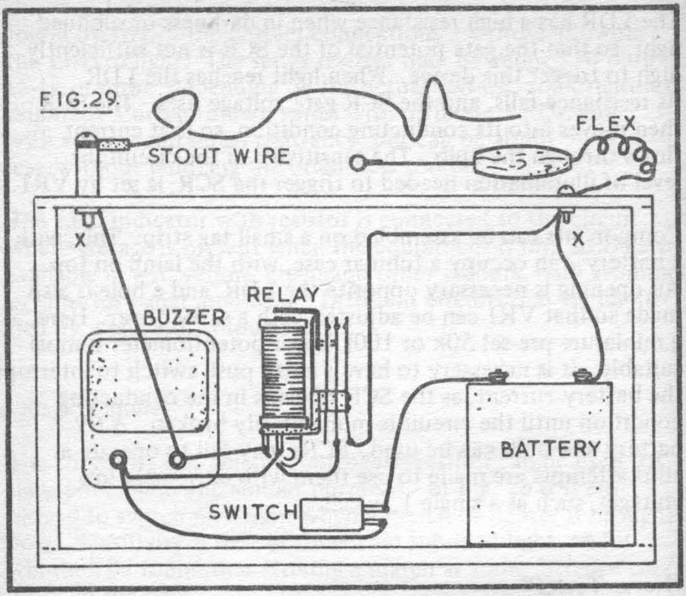
FIG.29.

If loop and wire touch, current flows from the battery to the buzzer, which sounds. Current also flows to the relay winding, so that the relay contacts close. Even a momentary contact between loop and wire will thus cause the buzzer to sound until the switch is used to break the circuit.

Any 4.5v, 6v or similar battery and buzzer will be satisfactory, with a relay which operates on a similar voltage. Connections are to the relay contacts which remain open until the relay is energised.

In use, the device can stand upright. A back is attached with small screws.

Burglar Alarm

The same circuit arrangement as in Fig. 29 will provide a burglar alarm operated by pressure mats. The latter are available in various sizes, and consist of a thin pad which can be placed under a carpet or ordinary mat.

Leads from the mat or mats will run to the terminals at X–X. The buzzer should be changed for a bell, which need not be fitted in the box, but can be situated where it is best heard. Momentary weight on any of the mats will then switch on the bell, which will be locked on by the relay as described, and will ring until the unit is switched off.

Transistor Tester

A very useful check of transistors can be made with the aid of a multi-range test meter, with a selection of base resistors to provide base current, as in Fig. 30. This method does not give an exact indication of transistor gain, but it does allow defective transistors to be eliminated, and provides a good idea of the performance to be obtained from surplus or other unknown transistors.

A 6-way switch allows the base to be open circuit, or to receive current through one of the resistors R1 to R5. These values are R1 330k, R2 110k, R3 33k, R4 11k and R5 3.3k. After allowing for the base-emitter potential, switch positions and corresponding base currents are near the following: 1, zero; 2, 10μA; 3, 30μA; 4, 100μA; 5, 0.3mA, and 6, 1mA.

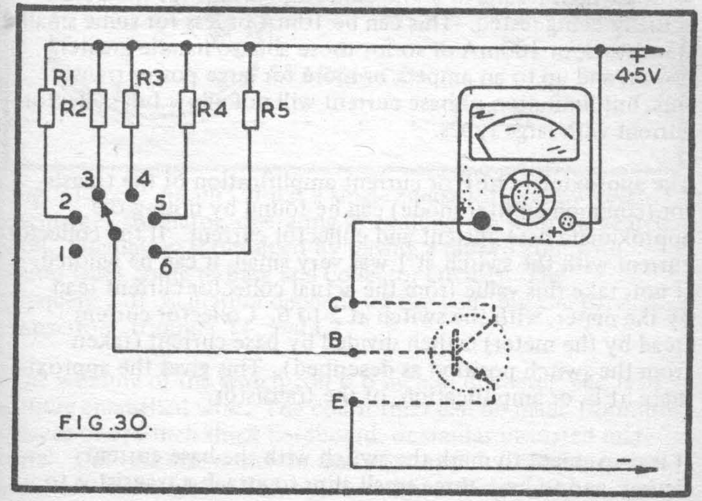
FIG.30.

Polarity is shown for NPN transistors. Reverse polarity of both meter and battery for PNP transistors.

For all tests, initially set the switch to 1, and the meter to its highest range (say 1 ampere). Afterwards, switch the meter to a lower range, after adjustment of the 6-way switch, as may prove to be necessary to obtain an easy reading.

With the switch at 1, a high reading on the meter shows that the transistor has a short circuit from emitter to collector, and is defective.

With germanium transistors, collector current (with the switch at 1) may be 0.1mA or so up to perhaps 0.5mA for small types, and up to some milliamperes for larger types. This small leakage current does not show that the transistor is defective.

In the case of silicon transistors, the leakage current (with switch at 1) will be zero or extremely small, if the device is in good condition.

The switch is then turned to one of the positions 2 to 6. If collector current does not rise, the transistor is defective. If current rises at each step, do not go beyond the switch setting which results in a safe collector current for the device actually being tested. This can be 10mA or less for some small transistors, or 100mA or so for those able to handle greater power, and up to an ampere or more for large power transistors, but limitation of base current will not allow full collector current with large types.

The approximate hFE or current amplification of the transistor (common emitter mode) can be found by noting the approximate base current and collector current. If the collector current with the switch at 1 was very small, it can be ignored. If not, take this value from the actual collector current read by the meter, with the switch at 2 to 6. Collector current (read by the meter) is then divided by base current (taken from the switch position as described). This gives the approximate hFE, or amplification, of the transistor.

It is convenient to mark the switch with the base current figures, and to have three small clips to attach a transistor to the circuit. Switch, clips, and battery can all be fitted to a small box. A development of the circuit would be to include a current meter, with switch and shunts, to replace the external general purpose test instrument.

Treasure Locator

This locator is among the simplest that can be constructed, as it uses an ordinary small portable radio as part of the detection circuitry. No modifications or connections to the radio are needed.

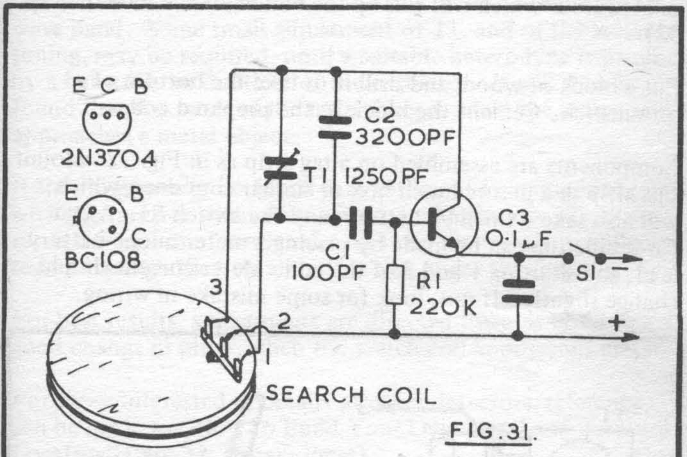

FIG.31.

Fig. 31 is the circuit. It has a single transistor as a low radio-frequency oscillator, operating in conjunction with the search coil. Oscillation is obtained by having a tapped coil, with outer leads going to base and collector circuits. The frequency of oscillation depends on the coil, and the parallel capacitors, trimmer T1, and C2.

The winding of the search coil is 6 inch in diameter, and is of 34swg enamelled wire. The coil former can be made from three pieces of 1/8 inch thick hardboard, or similar insulated material. Cut one disc 6 inch in diameter, and two others about 6½ inch in diameter. Smooth the edges with glasspaper, and

cement them securely together, with the small disc between the larger discs. This provides a former with a channel to take the winding. The adhesive should be allowed to dry with the discs under pressure, so that there is no gap into which turns can fall. The former should also be varnished, and allowed to dry, before winding.

The wire is passed up through a small hole, this being point 1. Wind on fifteen turns. Make a loop and pass this up through a hole, for point 2. Wind a further twenty-five turns in the same direction, finishing off at point 3. These leads are scraped and soldered to a small tag strip. This is an anchor point for insulated leads which will run up the handle of the locator (Fig. 32).

Cut a block of wood, and drill it to take the bottom of a broomstick. Cement the block to the prepared coil.

Components are assembled on a tag strip as in Fig. 32. Mount this strip in a plastic lunch box or similar container, which will also take a small 9v battery, and the switch S1. A test for oscillation can be made by placing a meter in one battery lead, and shorting 1 and 3 of the coil. Meter current should change slightly. If not, look for some mistake in wiring.

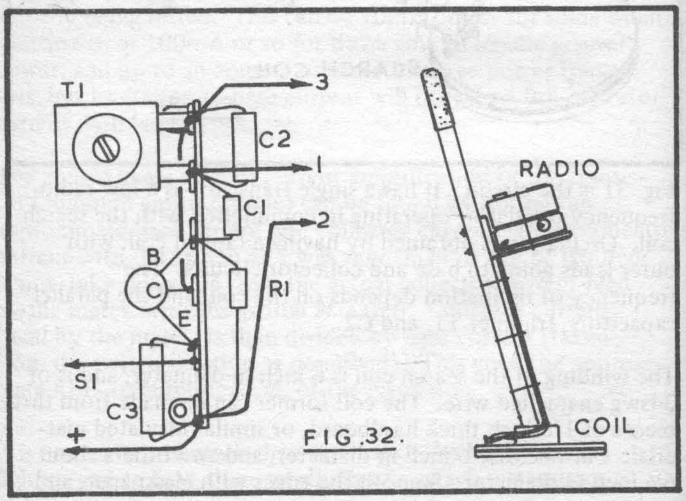

FIG.32.

A shelf is bolted to the handle, and the detector rests on this, with connections passing down to the search coil. A few loops of tape or string will secure these to the handle.

Operation in Great Britain can be on frequencies under 100kHz. Initially tune the radio receiver to 200kHz on long wave, and place it near the coil. Adjust T1 until a strong whistle is heard in the receiver speaker. This is the 2nd harmonic of the oscillator. T1 is then screwed down very slightly from this position.

The small portable radio is placed on top of the locator box, and can be held here with elastic. It is tuned to the medium wave band. Some small adjustment of T1, and of the receiver tuning, may be required, until a suitable heterodyne is produced by a harmonic or multiple of the oscillator. It will then be found that the pitch of the note changes when the search coil approaches a metal object.

It should generally be found that sufficient coupling is obtained with the receiver when they are situated as in Fig. 32. If not, run a few inches of insulated wire from the transistor collector tag up through the box and place this under the receiver.

For best results, adjustments are directed towards obtaining a good change in pitch, when the search coil approaches metal.

For those interested in details of other detectors, reference can be made to "How to Build Your Own Metal and Treasure Locators" (No. 32, Babani Press).

SECTION 3
RADIO AND AUDIO FREQUENCY

MW Booster

This unit can be employed in conjunction with any ordinary portable having a ferrite rod aerial, and will provide a considerable increase in range and sensitivity to weak signals. It has two functions – to allow the use of an external aerial and earth, and to amplify and couple signals to the portable receiver.

Tuning coverage is for medium waves. That is, about 200–550 metres, or 1500–550kHz. Many overseas and other stations can be picked up on this band, and distant reception can be particularly common on the higher frequencies, during the hours of darkness.

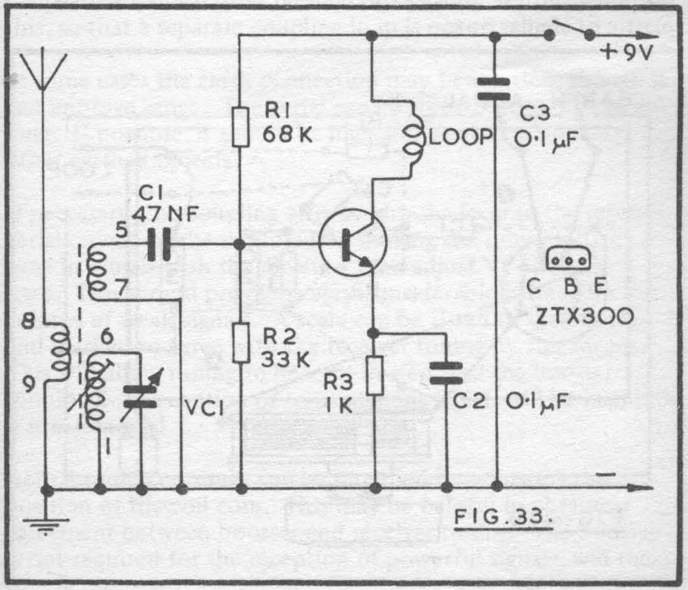

FIG.33.

Fig. 33 is the circuit, and it uses a cored tuning coil with aerial coupling winding 8–9, tuned winding 6–1, and base coupling coil 5–7. These numbers are for the pins of the Denco (Clacton) "Blue" Range 2 component. It will be apparent that other coils, or surplus or home-wound inductors, can be substituted.

For the coil mentioned, VC1 is 315pF. However, this value is not too critical, provided the capacitor fitted allows tuning the wanted frequencies. A solid dielectric capacitor can be used, though the air spaced type is likely to be a little more efficient.

Coupling to the base is by C1, and resistors R1 and R2 set the base bias conditions, with R3 for emitter bias. These values will suit numerous RF transistors.

Fig. 33B shows assembly of the components on an insulated board. The coil is mounted by its threaded bush. When the board has been wired, it can be mounted by means of bolts with extra nuts. The case is preferably metal, for screening purposes, but the booster can be used satisfactorily with a plastic or similar box.

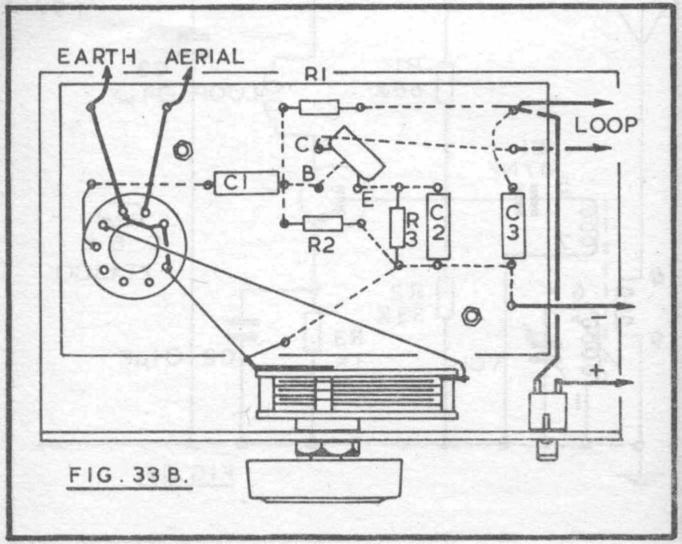

FIG. 33 B.

Fit a small slide switch to the box front, and aerial and earth sockets to the back, near the coil. A 9v battery is accommodated in the case.

Coupling to the transistor receiver is by means of a loop which is placed on the receiver ferrite rod, or near the rod. The loop can be about twenty turns of thin insulated wire, about ½ inch in diameter, and bound with cotton to keep the turns together The ends can be left long enough to reach the two connecting points in Fig. 33B.

If the case is metal, the booster and receiver can be near each other, so the loop leads will only need to be a few inches long. But when an insulated case is used, coupling from the receiver ferrite aerial or other circuits, to the booster coil, may cause instability. This will become apparent as whistles accompanying most signals. If this arises, the booster and receiver should be moved some inches apart. The aerial lead must not run near the loop connections.

Some receivers have an internal winding which is intended to couple an external aerial, and it may be possible to connect to this, so that a separate coupling loop is not required.

In some cases the earth connection may be omitted, though it can improve range. The aerial can be some 20ft to 40ft or so long. If possible, it should be high and clear of walls and other earthed objects.

If necessary, the coupling afforded by the loop to the receiver aerial circuit can be modified by moving the loop. Initially, tune in signals with the receiver, then adjust VC1 for resonance. This should provide a very considerable increase in the volume of weak signals. A scale can be fitted to the booster, and marked to agree with the receiver tuning by this means. This will allow tuning to be kept in step, and the booster should allow reception of transmissions which will be inaudible if it is removed.

Actual tuning coverage can be modified by adjusting the position of the coil core. This may be helpful in obtaining agreement between booster and receiver tuning. The booster is not required for the reception of powerful signals, and the

receiver automatic gain control circuit will tend to mask any increase in signal strength obtained with such transmissions.

Midget MW Diode Radio/Tuner

A small, compact tuner can be of utility in allowing radio reception through an amplifier, and a diode receiver can give excellent quality of reception. Such a unit can also be employed alone, with headphones, for personal listening.

Fig. 34 is a receiver/tuner which can be built in a very small plastic or other insulated box. VC1 is a compression type trimmer, fitted with a small knob for adjustment by hand. To do this, it is necessary to remove the short screw (generally 6ba) and replace this with a longer one, which projects to take the knob. The latter can be a terminal head, or the cap from an empty tube of adhesive, secured with lock nuts.

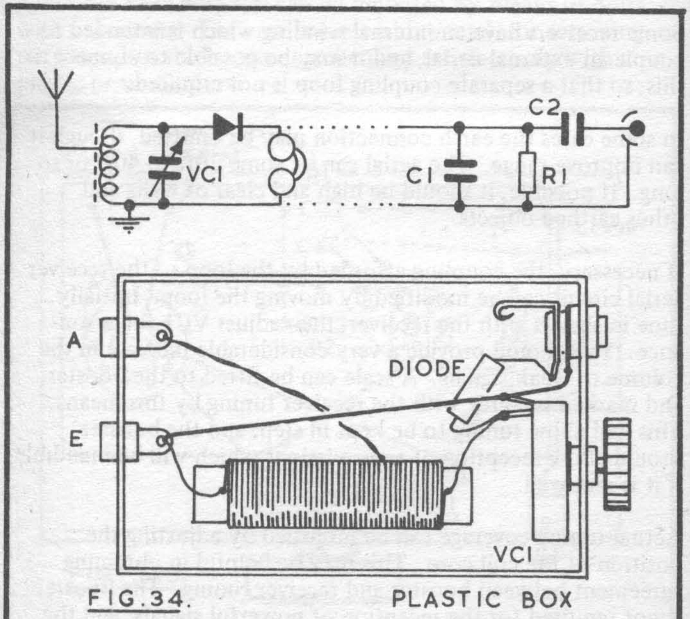

FIG.34.

For best efficiency and small size, the tuning coil is wound on a piece of ferrite rod. It consists of seventy turns of 24swg or similar enamelled wire, on a rod about 1½ inch long and 3/8 inch in diameter. Actual details will not be critical. A piece can be obtained from a longer rod by cutting a notch round the rod with a file and snapping at this point with the fingers. A few touches of adhesive will hold the ends of the winding secure. The loop for the aerial connection is about one-third the total number of turns from the diode end of the coil. (Moving this tapping towards the earthed end of the coil will increase selectivity or sharpness of tuning, but reduce volume.)

Aerial and earth sockets, and a 3.5mm jack outlet, complete the receiver. Headphones can be plugged directly in here.

The aerial and earth can be provided as explained for the diode radio in Fig. 7.

To use the radio as a tuner, C1, R1 and C2 may be added. Suitable values are C1 470pF, R1 220k, and C2 0.25μF. A screened lead with jack plug should be made up to connect the tuner to the amplifier input socket.

One Transistor VHF Receiver/Converter

The super-regenerative receiver is well known for its great sensitivity, which equals that of a multi-stage superhet. It has no alignment difficulties, and few components, so is an interesting project. Its main disadvantages are its background hiss, but this ceases when a signal is tuned in, and its lack of selectivity. The latter is however relatively unimportant on the VHF bands. The circuit here is particularly intended for 2m reception, but is easily modified. It is not intended for FM signals.

In Fig. 35 the detector is a VHF FET, and the field effect transistor will be found very satisfactory in circuits of this kind. T1 is the pre-set tuning capacitor, which allows adjustment over a fairly wide band of frequencies. Tuning is by VC1, which does not require a reduction drive.

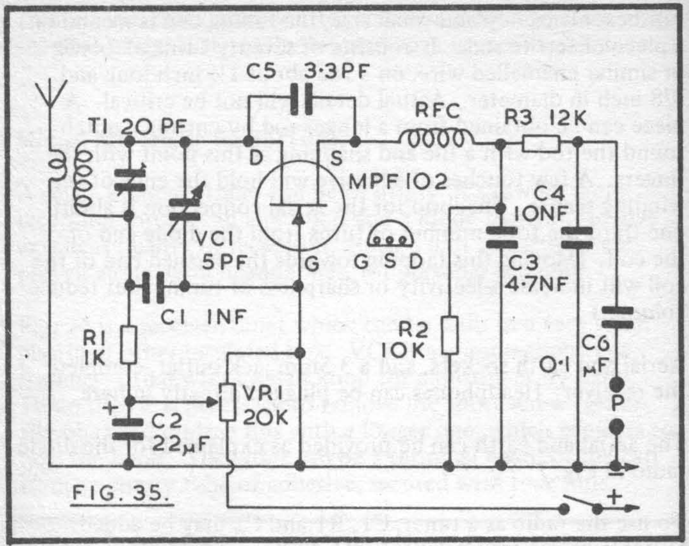

FIG. 35.

Control of super-regeneration is by means of the potentiometer VR1, which adjusts the FET drain potential. The value of VR1 is not very important, but a linear potentiometer is necessary.

C5 provides feedback from source to drain, for regeneration. R3 and the associated components form the quench network, with C6 feeding headphones, or an audio amplifier.

For best results, VR1 should enable the transistor to go smoothly into super-regeneration, which is shown by the hiss mentioned. Achieving this depends somewhat on the transistor, value of C5, adjustment of T1, and aerial loading. However, matters are not too critical.

Transistors unsuitable for VHF cannot be used. Individual FETs of the same type will also differ somewhat, so that in some cases it might be necessary to omit C5, or change the value. Alternatively, a small pre-set (not over 10pF) may be fitted. It will be found that a very small capacitance is required.

Another cause of failing to obtain super-regeneration is too tight aerial coupling, and this is easily corrected by moving the aerial coupling winding a little further from the larger winding.

The RF choke can be a VHF type, or can consist of seventy five turns of 36swg enamelled wire, side by side on a ¼ inch diameter insulated rod. Touches of adhesive may hold the end turns only.

Short leads are essential, and assembly on a tag board is shown in Fig. 35B. The variable capacitor is fitted first, and it can later attach the finished board to a panel, by means of an extra nut.

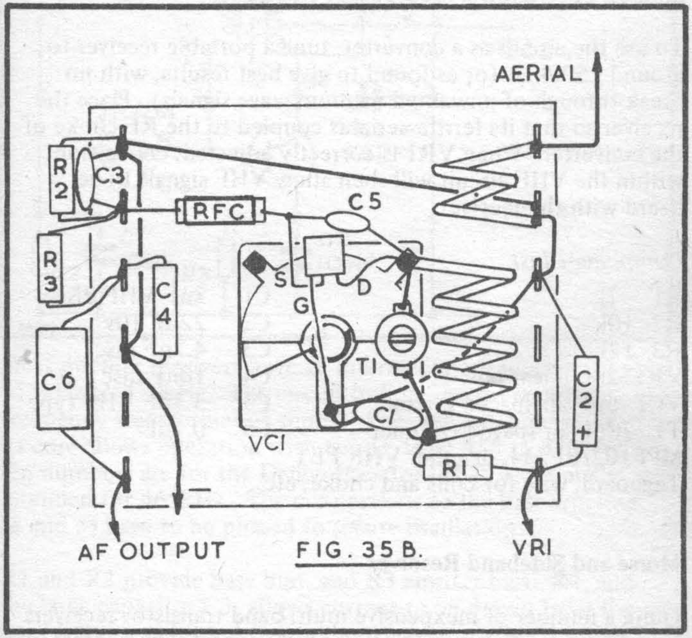

FIG. 35 B.

The larger coil has five turns of 18swg wire, and is ½ inch in diameter and ¾ inch long. The smaller coil has two turns. Wiring to VC1, T1, C5, the coil and C1 must be as short as can be arranged. Connections elsewhere are reasonably direct. A ceramic disc capacitor may be used at C1. T1 should be a low loss trimmer.

The board can be fitted in a case which will carry VR1, and large enough for the battery and on-off switch. The case can be used to support a telescopic aerial, some 30 inch or so when extended.

Frequencies tuned by VC1 can be raised by unscrewing T1, or by stretching the larger coil to separate the turns more, or by using a coil with fewer turns, or of smaller diameter. The receiver is most suitable for use over the 70MHz to 200MHz range.

For the single transistor alone, a sensitive pair of medium or high resistance phones will be needed. Audio ouptut from C6 can be taken to an amplifier, for more powerful reception with phones, or to allow a loudspeaker to be used.

To use the circuit as a converter, tune a portable receiver to around 1500kHz (or as found to give best results, with no break through of unwanted medium wave signals). Place the receiver so that its ferrite aerial is coupled to the RF choke of the converter. When VR1 is correctly adjusted, conversion within the VHF circuit will then allow VHF signals to be heard with the receiver.

Component List

R1	1k	C1	1nF VHF disc
R2	10k	C2	22μF 10v
R3	12k	C3	4.7nF disc
VR1	20k linear pot.	C4	10nF disc
VC1	5pF air-spaced variable	C5	3.3pF VHF type
T1	20pF air spaced trimmer	C5	0.1μF

MPF102/BF244, or other VHF FET
Tagboard, wire for coils and choke, etc.

Morse and Sideband Resolver

Quite a number of inexpensive multi-band transistor receivers have frequency coverage which includes one or more Amateur bands. That is, 1.8, 3.5, 7, 14, 21, or 28MHz. On these bands transmissions are often in Morse code or CW, or single sideband. CW or SSB can be resolved by the specialised Amateur receivers and transceivers, but not with the general purpose home receiver. To receive CW or SSB with the latter type of radio, it is necessary to add an oscillator. This either generates a heterodyne or beat note for code reception, or replaces the carrier which was eliminated from the single sideband transmission.

The circuit in Fig. 36 may be used for these purposes. It generates a carrier of about the same frequency as that used in the intermediate frequency stages of the receiver, exact adjustment of frequency being by VC1,

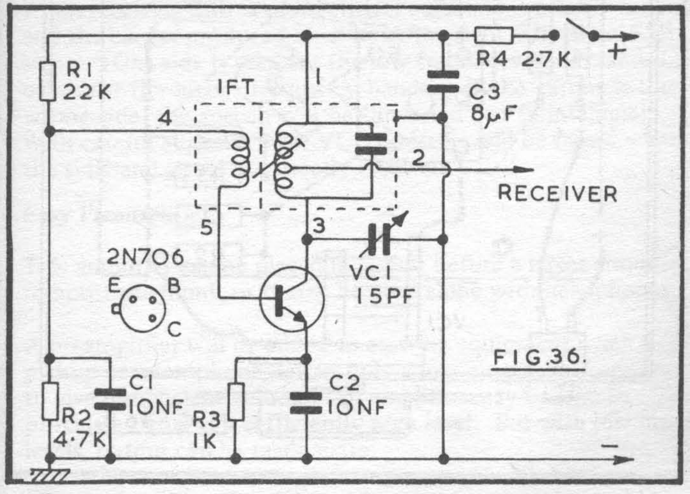

FIG.36.

Most popular receivers have an intermediate frequency of about 455kHz to 470kHz. The circuit in Fig. 36 uses an intermediate frequency transformer of similar type, and adjustment of its core allows operation with any IF from 455 to 470kHz. Pin numbers are for the Denco (Clacton) IFT13, normally intended for 465kHz. The connections to the base winding (4 and 5) have to be phased to secure oscillation.

R1 and R2 provide base bias, and R3 emitter bias. R4, and the large capacitor C3, allows current to be drawn from the same battery as is used to run the receiver, if wished. Alternatively, a separate 9v battery may be employed here.

Fig. 36B shows the practical layout of components, and wiring to them, using a perforated board. Connections and joints are kept close against the underside of the board. When wiring is finished, the board is mounted in a metal box, using ½ inch bolts with extra nuts for clearance.

VC1 can best have a dial with a central zero mark, this showing when the plates are half meshed.

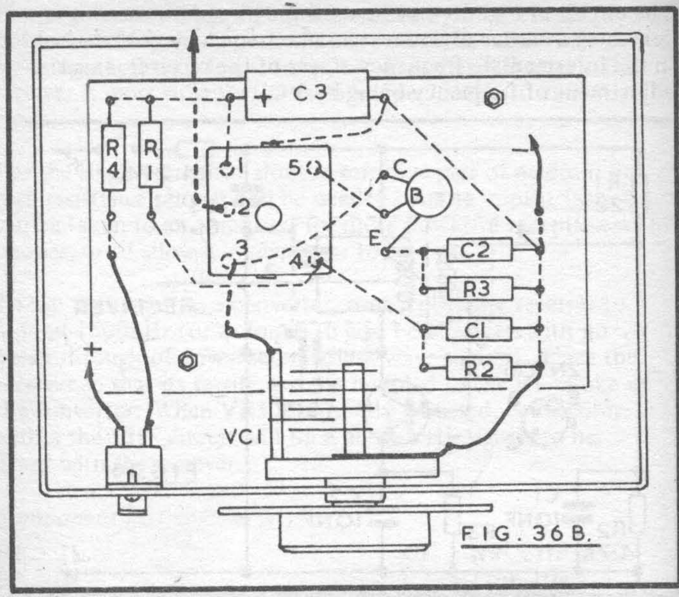

FIG. 36B.

An insulated lead is run from 2 on the IFT, and is placed near the receiver. This lead may be so positioned as to give best coupling into the IF stages of the receiver. If coupling is too weak, the signal from the CW/SSB Resolver will be inadequate. On the other hand, tight coupling will allow the carrier to swamp weak signals tuned in on the receiver. Fortunately the degree of coupling is not at all critical.

To set up the oscillator, first of all tune in any normal broadcast with the receiver. Set VC1 half closed and switch on. The core of the IFT should now be slowly adjusted until a loud heterodyne or beat note is heard accompanying the signal received. Set the core so that this note falls to about zero, or to a low pitch. It will then be possible to rotate VC1 either way from this central zero beat position, in each case producing a note which rises in frequency as VC1 is rotated.

To receive CW or Morse code, it is only necessary to tune in the code signal with the receiver, and then to adjust VC1 to

obtain a note of the most suitable pitch. VC1 can be rotated either way from its central or zero beat position, to obtain best freedom from interference.

When receiving SSB, a more critical adjustment of VC1 is needed and the carrier produced must be at the right side of the SSB signal. (One side is used for the low frequency bands, and the other for the higher frequency bands.) If the carrier is the wrong side, the speech will be "inverted" or unintelligible. With careful adjustment of VC1, a setting will be found where the sideband signal is correctly resolved.

Easy Preamplifier

This amplifier can be placed in circuit before a larger amplifier, to boost the input, or it may be used alone with headphones.

A preamplifier will be added to existing equipment when a pickup or microphone having only a low output has been found to give insufficient volume. An amplifier may be able to provide a signal at a sufficiently high level. But with low input levels, output can be inadequate.

The stage in Fig. 37 can provide considerable gain. Audio input is to C1, and R1 provides base current for the transistor. Amplified signals are obtained across R2, and pass to the main amplifier (or phones) via the isolating capacitor C2.

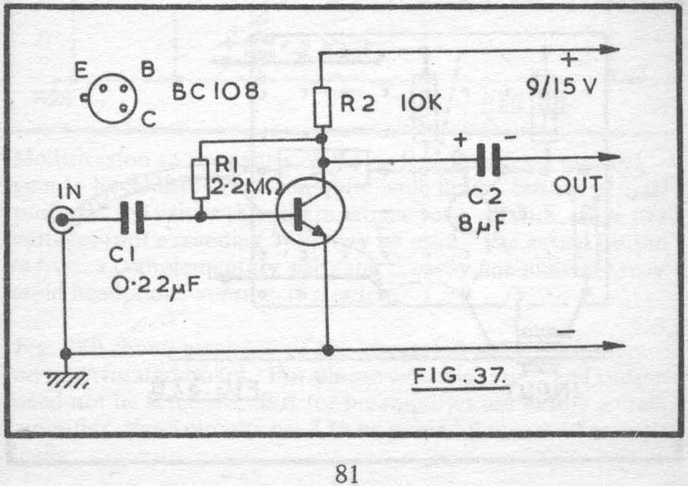

FIG. 37.

With a BC108 or similar transistor, the supply can be about 9v to 15v. Numerous NPN transistors, intended for preamplifier use and of low noise type, will perform well. Current may be from a separate battery. Or it may in many cases be drawn from the main amplifier. To avoid instability or similar troubles, the positive supply should be from a well decoupled and smoothed point in the main amplifier. If this is not available, connect a 220μF capacitor from positive to negative in Fig. 37, and draw the positive supply through a 1k resistor. Negative is returned to the negative line of the main amplifier.

Should the main amplifier have a positive earth line, then the transistor can be a PNP type, R1 and R2 being changed to suit.

Fig. 37B shows assembly on a perforated board. When the circuit provides input to a main amplifier, it should generally be in a metal case. Input is by a screened lead, and a screened lead is also fitted for the audio circuit to the main amplifier. The case should be earthed to the negative line. This can be by means of one of the bolts securing the board.

Screening to avoid hum or instability will generally prove to be unnecessary when using the stage to operate headphones. So two sockets, or a jack socket, will be more convenient for the phone circuit.

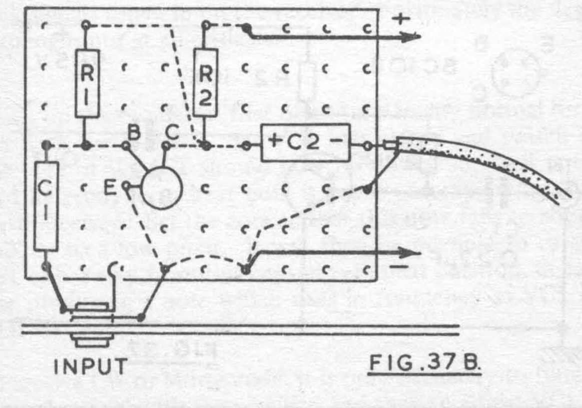

FIG.37B.

741 Preamplifier

Integrated circuits have several stages of amplification, formed as a single chip and enclosed in a small case. The 741 is an easily obtained audio IC which can be used in many circuits. The uA741CE or metal can 741 is convenient, as its leads may be spaced to suit any ready perforated board. The dual in line 741 is electrically similar.

Fig. 38 shows a preamplifier using the metal can 741. Input is via C1 to 3. As shown, 7 and 4 are positive and negative, while output is from 6. Negative feedback is applied to 2, via R4. C3 is an isolating capacitor.

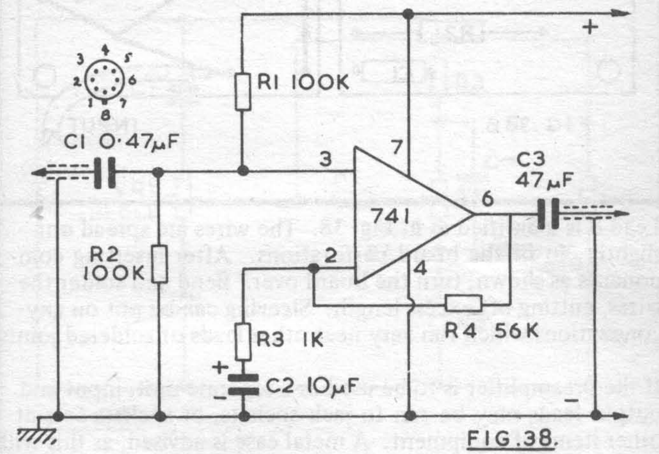

FIG.38.

Modification to the values of R3 or R4 allow feedback and gain to be adjusted between quite wide limits, but for general purposes the values shown are satisfactory. A wide range of voltages (not exceeding 36v) may be used. The actual output is from a complementary pair, and is easily adequate for very good headphone volume, if required.

Fig. 38B shows assembly of the integrated circuit amplifier on a perforated board. For phone working, input and output need not be screened. But for preamplifier use before a main amplifier, these circuits need to be made by means of screened leads.

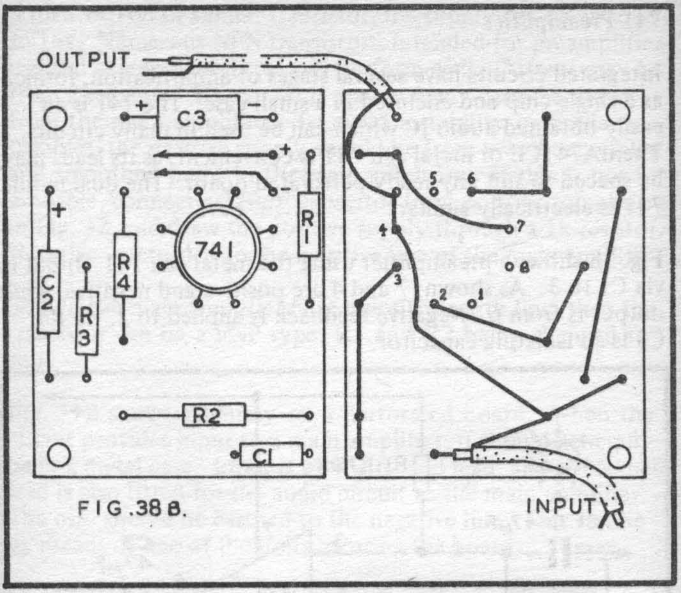

FIG.38B.

Lead 8 is identified as in Fig. 38. The wires are spread out slightly, to fit the board perforations. After inserting components as shown, turn the board over. Bend and solder the wires, cutting off excess length. Sleeving can be put on any connections which run very near other leads or soldered joints.

If the preamplifier is to be used as a separate unit, input and output leads may be run to jack sockets, or sockets to suit other items of equipment. A metal case is advised, as this will help to avoid any pick up of hum, or other troubles due to coupling to the board wiring. The board can be secured to the metal box or case with ½ inch 6ba bolts, and extra nuts to give a little clearance.

The amplifier may be run from its own battery; or from a supply used with a main amplifier. For the latter, it is wise to provide a decoupled point, in a similar manner to that described for the transistor preamplifier. This consists of a quite large capacitor (say 220μF to 470μF) from positive to negative across the preamplifier supply circuit, and a resistor of 1k to 2.2k or so from preamplifier positive to positive on the main amplifier.

Tone Control for Your Discs Etc.

The preamplifiers shown have a substantially flat response. That is, they do not purposely emphasize bass or treble, or have similar effects in modifying the audio frequencies present. In some circumstances it is necessary to boost treble or bass, or reduce the emphasis of these frequencies, in order to obtain the most satisfactory reproduction.

This may be achieved by making a fully adjustable tone control, and Fig. 39 is a suitable circuit. A volume control is also incorporated, as this has not been present in the preamplifiers.

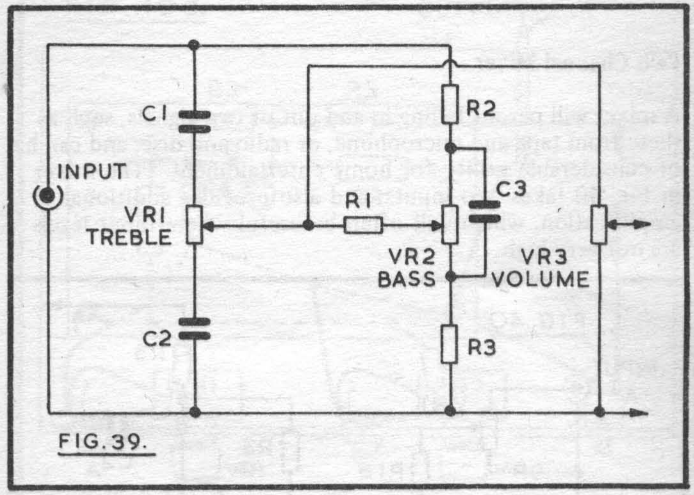

FIG.39.

The tone control is a passive circuit, needing no battery, and it can be placed between the source of audio signals (generally a microphone, pick-up, or radio tuner) and the amplifier or preamplifier. It may also be located between preamplifier and main amplifier, if wished.

Results are substantially flat with the treble control VR1 and base control VR2 in their central positions. Adjustment of VR1 and VR2 will provide separate boost or cut of treble, or boost or cut of bass frequencies. VR3 is the volume control.

The degree of control afforded by circuits of this type may be adjusted by alterations to the component values. In practice, it will be found that there is little point in having extreme levels of adjustment, which will never be required. The following values may be used: C1 and C2, 270pF; R1, 390k; R2 and R3, 100k; C3 680pF; VR1 and VR2, 2 megohm linear potentiometers; VR3, 2 megohm log potentiometer.

Components should be assembled in a metal case, for screening. The case is grounded to the earth line. Input and output sockets can be near the ends of the case, with VR1, VR2 and VR3 between. Components can be soldered directly to the potentiometer tags.

Two Channel Mixer

A mixer will permit fading in and out of two signals, such as these from tape and microphone, or radio and disc, and can be of considerable utility for home entertainment. The mixer in Fig. 40 takes two inputs, and also provides additional amplification, which will often be useful where input levels are not very high.

FIG.40.

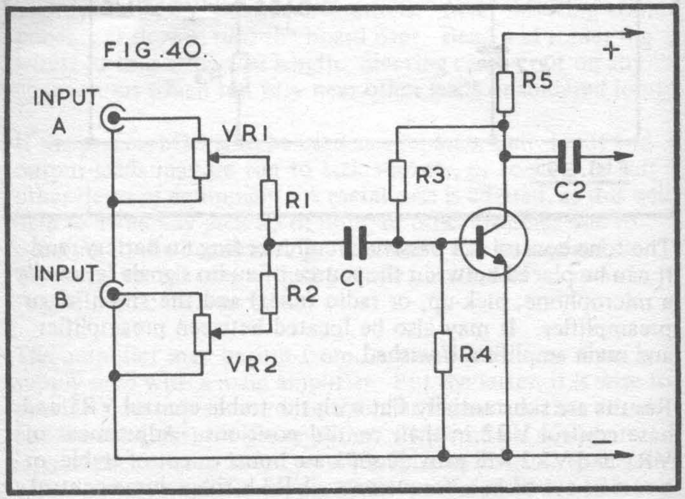

Input A has its individual volume control VR1, and Input B is similarly provided with VR2. Resistors R1 and R2 prevent short circuiting one input when the other potentiometer is set at zero. Coupling to the transistor base is by C1. R3 and R4 set base bias conditions, with R5 as the collector load and audio output from C2.

Component values are not very critical, though R3, R4 and R5 need to be selected to suit the transistor fitted. The mixer is connected and used in a similar way to that explained for the preamplifiers.

FIG.40B.

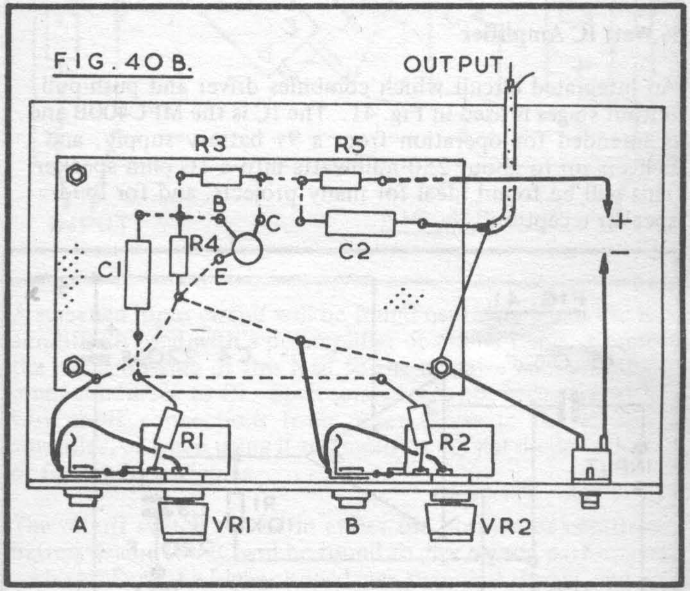

Fig. 40B is the layout, using a metal case or box. A screened output lead is provided. Input sockets A and B may be chosen to suit the type of plugs on existing equipment. The on-off switch is separate from the potentiometers, so that these can be left at pre-set levels if wished.

Component values may be chosen to suit low or high impedance inputs, if wished. Or the following may be used for general purposes:

Components List:

VR1, VR2, 1 megohm log potentiometers.
R1, R2, 200k

C1 0.5μF	C2 0.5μF
R3 330k	BC109
R4 22k	
R5 10k	

This is when using a 9v supply.

¼ Watt IC Amplifier

An integrated circuit which combines driver and push-pull output stages is used in Fig. 41. The IC is the MFC400B and is intended for operation from a 9v battery supply, and delivers up to about 250 milliwatts into a 16 ohm speaker. This will be found ideal for many projects, and for loudspeaker reception.

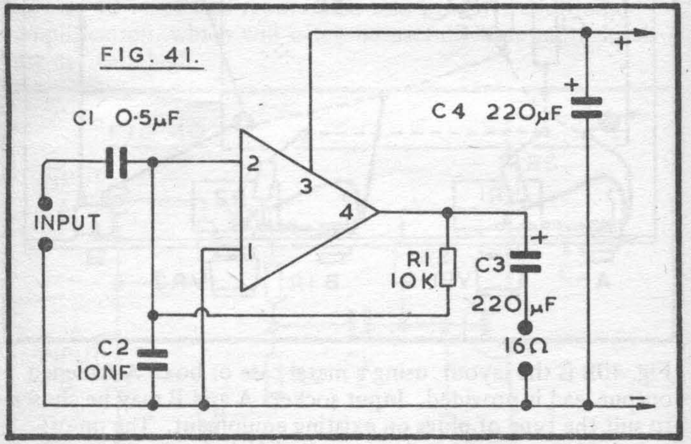
FIG. 41.

Audio input is to 2 of the IC, from the isolating capacitor C1. C2 is a by-pass capacitor to maintain stability. The IC has only four connections, 1 going to the negative line, and 3 to the positive supply. Audio output is from 4, and to the loudspeaker via C3. R1 provides negative feedback and bias to point 2.

The assembly of this amplifier on 0.15 inch matrix board is shown in Fig. 41B. Pin numbering is when looking at the top of the IC and the component side of the board.

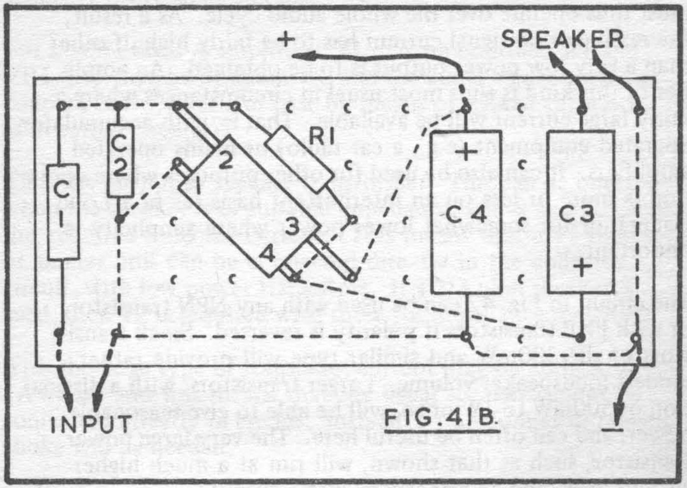
FIG.41B.

A screened input circuit will be found necessary when the IC amplifier is used with a preamplifier or similar items. Connect the outer brading of this lead to the negative line, and the inner conductor to C1. Such screening is not necessary with very short connections from other stages to the IC amplifier, or when using it to amplify a crystal diode radio, or for similar purposes.

The on-off switch can be in either the positive or negative battery lead. The IC will be found to give a good performance with speakers of a higher impedance than that shown, and up to 75 ohm may be used (with some small loss in output). Units of lower impedance than 16 ohms should not be fitted.

The small size of this IC and the few extra components needed make is very suitable for miniature apparatus, probably with a 2½ inch speaker. It will also operate a large speaker well (see "Boost Speaker") and at a volume level adequate for many purposes.

Class A Amplifier

A Class A amplifier uses a single transistor, Fig. 42, which must thus operate over the whole audio cycle. As a result, the resting or no-signal current has to be fairly high, if other than a very low power output is to be obtained. An amplifier of this kind is thus most usual in circumstances where a fairly large current will be available. That is, with accumulator operated equipment (e.g., a car radio) or mains operated amplifiers. It can also be used for other purposes where operation is more or less on an intermittent basis (as in a Loud Hailer), or for somewhat lower power where simplicity is important.

The circuit in Fig. 42 can be used with any NPN transistor, or with PNP transistors if polarity is reversed. Small transistors of the 300mW and similar type will provide rather modest loudspeaker volume. Larger transistors, with a dissipation of 500mW to 1W or so, will be able to give reasonable power, and can often be useful here. The very large power transistor, such as that shown, will run at a much higher current level, and deliver several watts output.

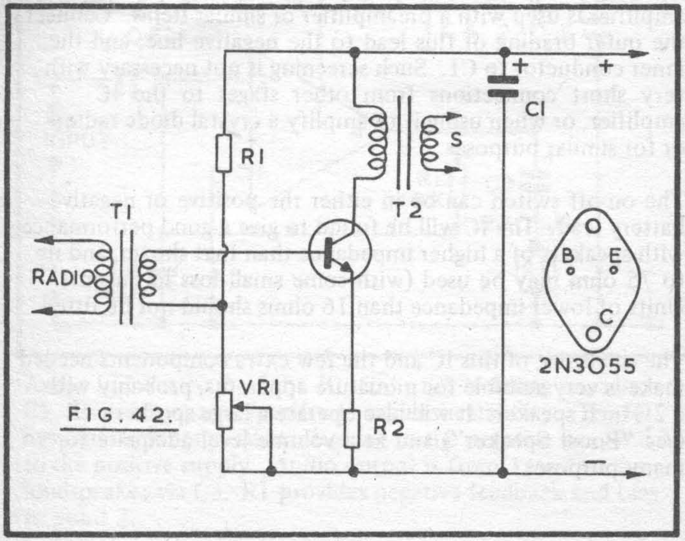

FIG. 42.

T1 is the coupling or driver transformer, which may receive audio signals from a radio receiver, smaller amplifier, or other equipment. A step down ratio of about 5:1 to 10:1 will generally prove to be suitable.

For smaller Class A amplifiers, T1 may be omitted. R1 and VR1 may then be of about 1k, and the audio signal can be coupled to the transistor base by means of a fixed capacitor, as in earlier circuits.

T2 is the output or speaker transformer. This can be omitted when the speaker is of suitable impedance, and is able to carry the required collector current. This means that a 75 ohm or similar unit can be connected directly in the collector circuit, with low power transistors. If a 2/3 ohm speaker is used, then T2 is required with these for impedance matching.

With the large type of transistor, current may be 500mA to 1A or so, and this means that the usual speaker cannot be connected directly in circuit. Instead, a transformer or tapped choke will be needed.

The circuit is often operated from 12v (a vehicle supply) but can be used with 9v, or more than 12v.

R1 and VR1 set bias conditions, and R2 limits peak collector current. For the 2N3055, R2 may be 1 ohm. R1 can be 100 ohm, and VR1 25 ohm. T2 can be 2:1 for a 3.5 ohm speaker. C1 is 3200μF. Begin with VR1 at minimum value, slowly increasing this until reproduction is satisfactory. (This will set collector current at about ½ ampere.)

With other transistors, adjustment can be made in the same way, by slowly increasing the base bias until best results are secured. Raising the value of VR1 in Fig. 42, or reducing the value of R1, will increase collector current, and thus raise the power which can be handled. However, current is limited by the power handling capacity of the transistor, so a meter should be put in one battery lead, and a check made that the power dissipation of the transistor is not exceeded. With small transistors, current can be limited to 50mA or so.

Fig. 42B shows assembly of the circuit in Fig. 42. The power

transistor is mounted on an aluminium plate, which acts as a heat sink to carry away heat from the transistor itself. Note that this plate is common to the collector. Base and Emitter pins pass through clearance holes. Smooth off any burr round the fixing and clearance holes, so that the transistor can rest flat on the metal. Where it is awkward to have the plate or case common to the collector, the usual insulation set can be fitted. This consists of bushes for the fixing bolts, and a thin mica washer to go under the transistor, so that electrical isolation is obtained.

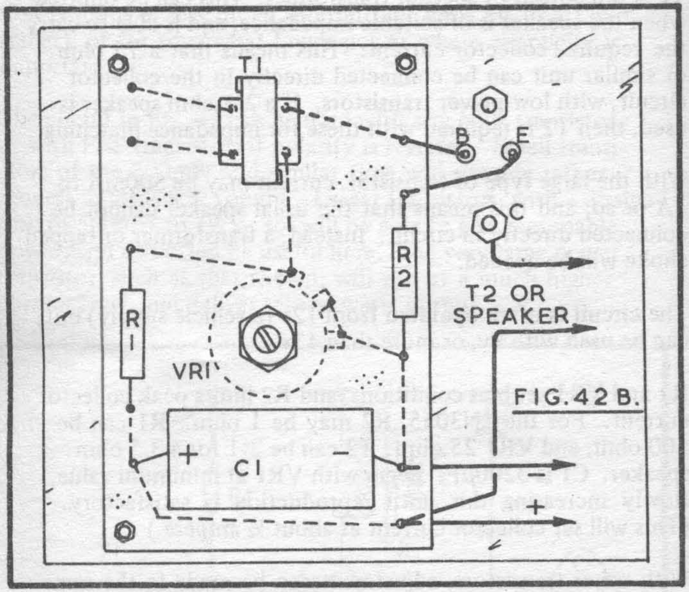

FIG. 42 B.

The insulated board carrying the other components is mounted by long bolts with extra nuts. A large Class A stage transistor of this type will grow very warm to the touch, if run at anything like maximum power rating. This does not indicate any defect.

Two Stage Amplifier

The circuit of a 2-stage amplifier is shown in Fig. 43, and this can be used for loudspeaker reproduction. Input to the driver stage is by the isolating capacitor C1. Operating conditions in this stage depend largely on the values of R2, R3, and the emitter bias resistor R5.

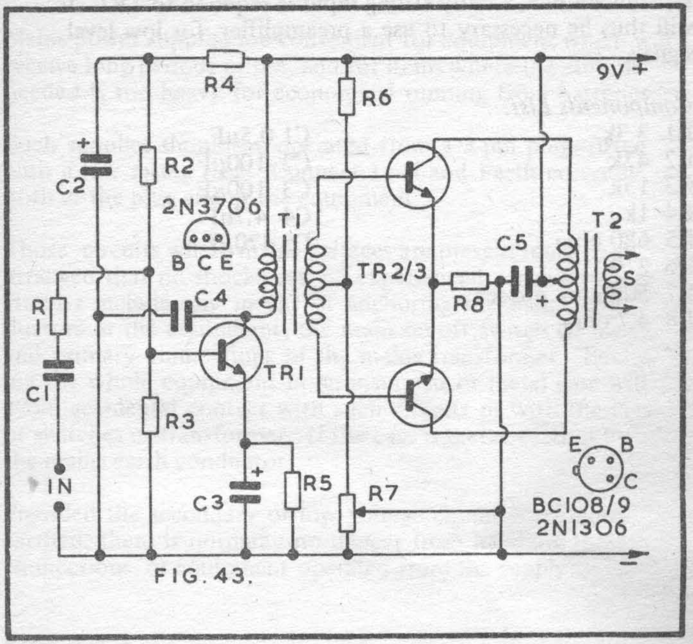

FIG. 43.

T1 is the driver transformer, with centre tapped secondary, so that TR2/3 may be operated in push-pull (see Fig. 12). T2 couples output to a 2/3 ohm speaker.

With the values shown, a BC109 will give very good results in the TR1 position, with 2 x BC108s for the output stage. R7 is set so that battery current is around 10mA, with no signal present. Current rises from 20mA to 60mA or so, according to volume, when the amplifier is driven.

A pair of 2N1306 transistors may be used for the output stage. Other transistors are also possible here, as operating conditions can be adjusted by R7.

The 2N3706 will operate in the TR1 position, though with

lower gain than the BC109. Numerous other small audio transistors can also be used here.

For full volume, a fairly strong input is required to TR1. It will thus be necessary to use a preamplifier, for low level signals.

Components List:

R1 3.3k
R2 47k
R3 15k
R4 1k
R5 680 ohm
R6 2.7k
R7 500 ohm pre-set
R8 4.7 ohm

C1 0.5μF
C2 100μF
C3 100μF
C4 4.7nF
C5 470μF

SECTION 4
POWER SUPPLIES

Mains power supplies are convenient for equipment which will receive long periods of use, and for items where the current needed is too heavy for economical running from batteries.

Such supplies should be operated from a 3-pin plug, fitted with a low rating fuse. Connect L, N and Earth correctly, both at the plug and at the equipment.

Those circuits where mains voltages are present must be so arranged that no shocks can be experienced. Mains voltage circuits include any means of anchoring the mains conductors in the equipment, the main on-off switch (if used), and primary connections of the mains transformer. Enclosing the whole equipment in an insulated or metal case will avoid accidental contact with such circuits or with the tags of switches or transformer. If the case is metal, earth it by the mains earth conductor.

Provided the secondary or low voltage circuit is correctly earthed, there is normally no danger from handling these connections, or equipment operated from the supply.

Easy PSU

A useful power supply suitable for many purposes can use the circuit in Fig. 44. Mains L (live) and N (neutral) conductors go to the primary P of the transforer. A switch is placed in the L circuit.

The secondary S can delivery any wanted low voltage, and also give isolation from the mains.

Almost all equipment will need direct current. The four separate rectifiers, or an equivalent single device, produce pulsating DC. The reservoir capacitor C1 smoothes this. Resistor R1, with the second capacitor C2, gives additional smoothing. Output from the supply is then substantially smooth DC, as would be obtained from a battery.

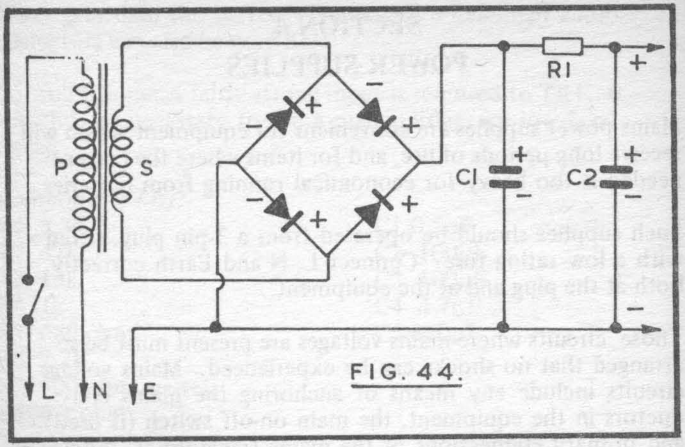

FIG. 44.

Sometimes R1 and C2 are omitted. Should hum accompany reproduction from an amplifier, this shows that smoothing is probably inadequate, and if so the cure is to add R1 and C2.

Any wanted voltage can be obtained by using a transformer with an appropriate secondary, with rectifiers and capacitors to suit.

C1 may charge up to approximately 1.4 times the RMS or usual voltage rating of the secondary S. Thus, if S were a 6.3v winding, about 8.8v would be expected across C1. But should R1 be present, a voltage drop will arise in it, the figure depending on the resistance value of R1, and current flowing.

For these reasons, the actual output may not be exactly the figure expected. This will often be unimportant, and such power supplies are largely used for all sorts of equipment.

For apparatus generally operated from a 9v or similar battery, a bell transformer with 3v, 5v and 8v output tappings is suitable. If R1 is 22 ohm, this will give around 9v at 100mA, with the 8v tap in use.

It is convenient to use 50v or 100v rectifiers, and the rating can be 1 ampere, with a 1 ampere transformer secondary. Four individual wire ended rectifiers can be mounted on a tag strip, and connected as in Fig. 44. A combined rectifier will have AC, positive, and negative tags (see Fig. 45) and can be fitted instead.

The capacitors need to be of large value, and have a voltage rating higher than the actual voltage which will be present. The 12v capacitors could be fitted in the circuit described. C1 can be 2500µF or larger, and C2 can be of similar value.

Easy Regulated PSU

With the circuit in Fig. 44, the exact voltage will depend on the current drawn, rising when current is low. To avoid this, some means of regulation may be employed.

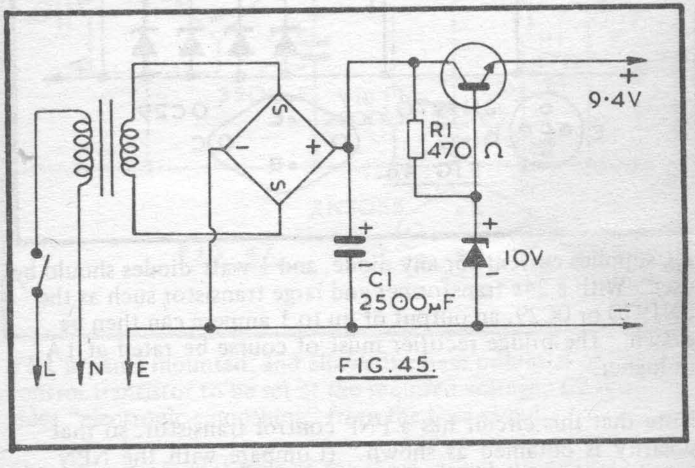

FIG.45.

Fig. 45 shows a typical regulation circuit, giving an output of approximately 9.4v. The output from the rectifier needs to be of somewhat higher voltage than required at the output, and a transformer with a 10v secondary is suitable. Current through R1 supplies the Zener diode, which holds the base of the transistor at virtually the same voltage for all current loads. Some voltage is lost in the transistor, so the output is nearly the same, for any current load. For up to about 200mA, a BFY51 or equivalent transistor is adequate, so this will do well for many purposes.

Multi-Voltage Regulated PSU

Fig. 46 is a somewhat similar circuit, but it has a switch, to select any one of the four Zener diodes. These may be chosen so that 6v, 9v, 12v and 15v will be available.

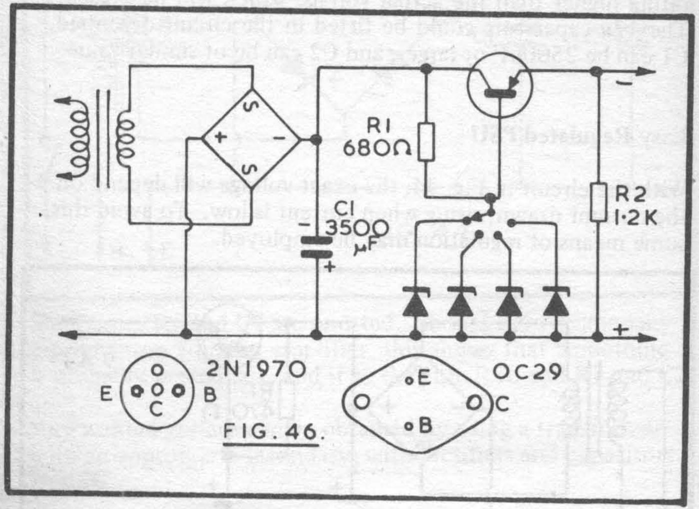

FIG. 46.

R1 supplies current for any diode, and 1 watt diodes should be used. With a 24v transformer and large transistor such as the 2N1970 or 0C29, an output of up to 1 ampere can then be drawn. The bridge rectifier must of course be rated at 1 A or higher.

Note that this circuit has a PNP control transistor, so that polarity is obtained as shown. (Compare with the NPN transistor in the previous circuit.) However, either positive or negative can be grounded, and a separate earth terminal can be provided to allow this. A flying lead is placed between earth terminal and positive, or earth and negative, as required. The permanent negative earth connection shown in Fig. 45 must of course then be omitted.

More details of actual construction will be found later.

One Transistor Adjustable PSU

Fig. 47 is the circuit of a simple power supply unit which can be set to provide any output from zero to 12 volts. The transformer has a 17v secondary, rated at 1 ampere or higher. After rectification, DC is obtained from the reservoir capacitor C1.

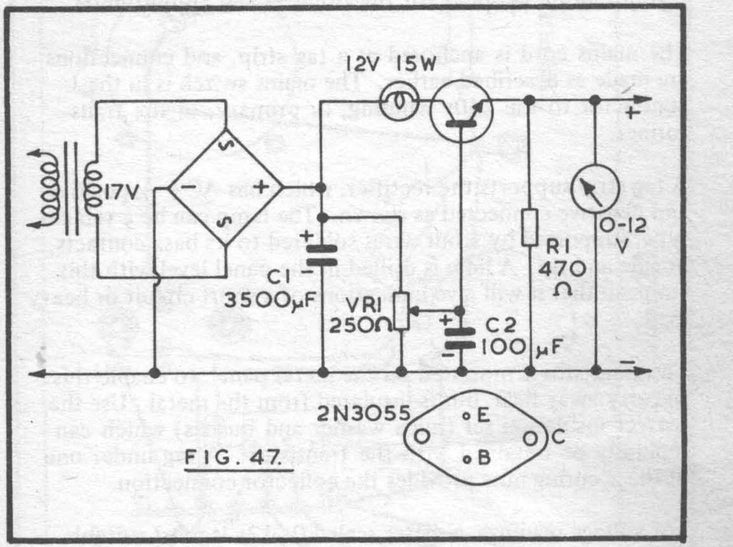

FIG. 47.

VR1 is panel mounted, and allows the base potential of the control transistor to be set at the required voltage. C2 provides "electronic smoothing" from the base circuit.

The 12 volt 15 watt bulb is for short circuit protection. With relatively low currents, little voltage is lost here. But if a short circuit arises in the output circuit, current is limited to a little over 1 A by the bulb. This reduces the chances of any damage due to temporary shorts in leads or equipment.

R1 is a bleeder resistor, drawing some current even when no external load is connected. The meter is the usual high resistance type, scaled from zero to 12v. VR1 is adjusted until this instrument shows the wanted voltage.

The regulation of this supply is not so good as that of supplies using a number of transistors. But it will be found suitable for all ordinary purposes, and to supply power for circuits in this book, as only a small change in voltage arises, with changes to the load, or current drawn.

Fig. 47B shows practical construction of this PSU. A similar layout can be adopted for the other power supply units.

The mains cord is anchored at a tag strip, and connections are made as described earlier. The mains switch is in the L conductor to the 240v winding, or primary, of the transformer.

A tag strip supports the rectifier, which has AC (~), positive and negative connected as shown. The lamp can be a vehicle type, supported by stout wires soldered to its base contacts, or side and tip. A hole is drilled in the panel level with this lamp, so that it will give indications of a short circuit or heavy load.

The transistor is mounted on the metal panel, to enable this to carry away heat, but is insulated from the metal. Use the correct insulation set (mica washer and bushes) which can generally be obtained with the transistor. A tag under one of the securing nuts provides the collector connection.

For voltage readings, a meter scaled 0–12v is most suitable. Or a 1mA meter, with 12k 1% series resistor, may be used. If so, a new scale, marked 0–12, can be drawn for it.

A 3 watt linear potentiometer is used. Earth and negative are connected to the metal panel, for equipment with a negative ground line. The positive output terminal or socket is insulated from the metal.

The case can be metal, or wood. It should have a few ventilation holes. The transistor can be expected to run quite warm, at maximum current. Heating here rises as the output voltage is reduced by VR1, as in these circumstances more voltage is dropped in the transistor.

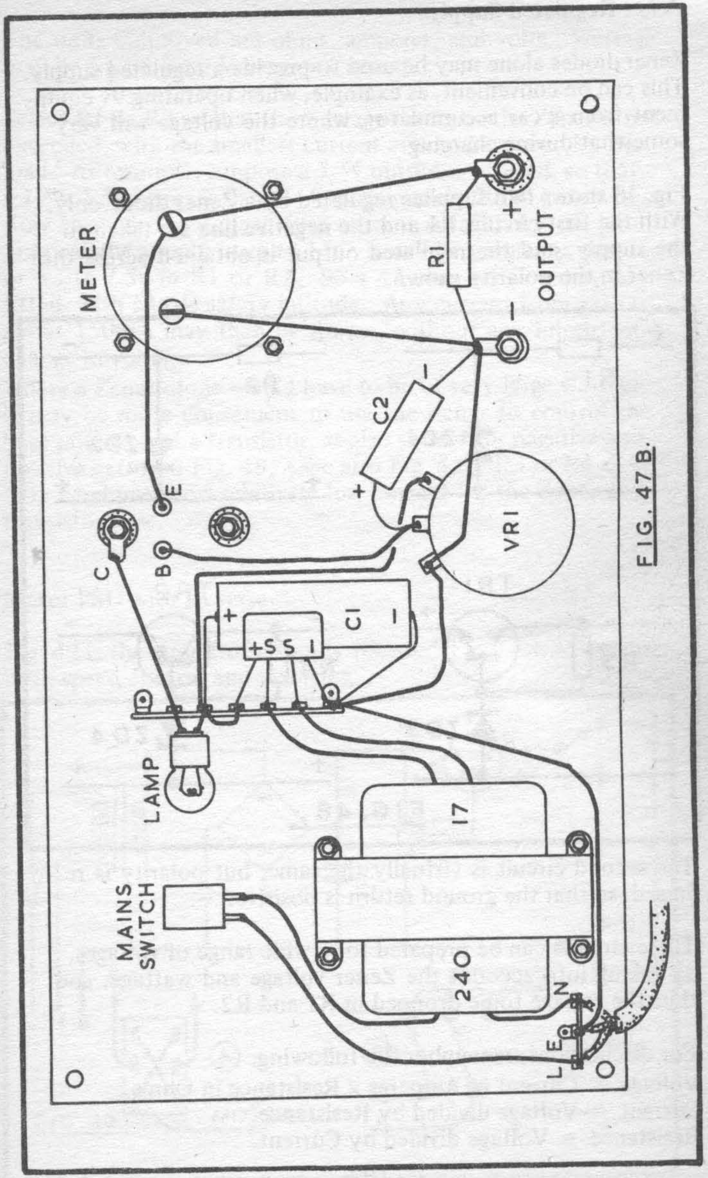

FIG. 47B.

Zener Regulated Supply

Zener diodes alone may be used to provide a regulated supply. This can be convenient, as example, when operating 9v equipment from a car accumulator, where the voltage will vary somewhat during charging.

Fig. 48 shows two supplies regulated by a Zener diode only. With the first circuit, R1 and the negative line are taken to the supply, and the regulated output is obtained across the Zener in the polarity shown.

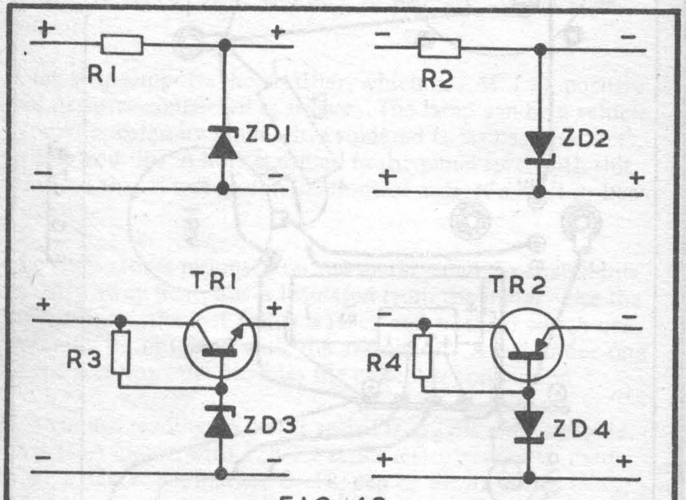

FIG. 48.

The second circuit is virtually the same, but polarity is reversed, so that the ground return is positive.

These circuits can be prepared for a wide range of voltages, by taking into account the Zener voltage and wattage, and thus the voltage to be dropped in R1 and R2.

For calculations, remember the following:
Voltage = Current in Amperes x Resistance in Ohms.
Current = Voltage divided by Resistance.
Resistance = Voltage divided by Current.

The units employed are ohms, amperes, and volts. Wattage may be found from Voltage x Current.

Select R1 or R2 so that the wattage rating of the Zener is not exceeded, with the smallest current ever to be drawn by the load. As example, suppose a 7.5v output is wanted, so that a 1½w 7.5v Zener is fitted. Approximately 200mA may thus flow through the Zener, if no current is taken by the load. Suppose the supply is up to 15v. It will be necessary to drop or waste 7.5v in R1 or R2. So a 42 ohm resistor may be fitted, with a little safety latitude. Any current from zero to about 150mA may then be drawn, without any important change in voltage.

Where a Zener diode would have to be of very large wattage, it may be more convenient to use the Zener to control the base potential of a transistor, as also shown for negative and positive earths in Fig. 48. (See also Fig. 45.) R3 or R4 then only need provide a relatively low current for the Zener and transistor base.

Motor PSU with Reverse

Fig. 49 is the circuit of a supply for model motors, and giving both speed control and reversing.

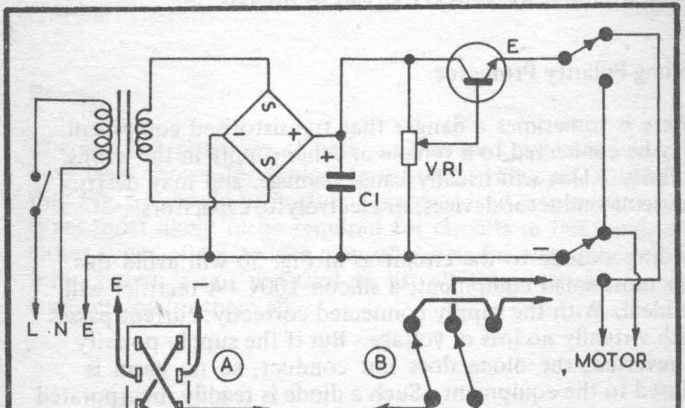

FIG. 49.

A transformer with a 12v secondary will be suitable for the usual model requiring up to 12v. The circuit is similar to that of the supply in Fig. 47. The potentiometer VR1 gives speed control. C1 can be around 2,000μF.

"A" is one type of ready-made switch, and it can be obtained with a central off position. "B" shows a home-built switch, similar to that described in Fig. 5. The switch must not short circuit the supply when changing over polarity, and one with central off is recommended.

Motors reversed in this way must be of the permanent magnet type, in which a change of polarity causes the armature to run in the opposite direction. Train and many other model motors are of this construction. Motors with a wound field will not be reversed by changing the polarity of supply.

A controller of this kind should be constructed as shown earlier, so that it is safe to use, and so that no joints in the mains circuits can be touched.

Earthing can only be omitted when a double-insulated transformer, of the type specifically made to avoid any need for earthing, has been used. The whole controller should be so constructed that, even if a fault develops, mains voltages cannot reach the secondary circuit or model.

Wrong Polarity Protector

There is sometimes a danger that transistorised equipment may be connected to a vehicle or other supply in the wrong polarity. This will usually cause damage, and may destroy the semiconductor devices, or electrolytic capacitors.

Adding a diode to the circuit as in Fig. 50 will avoid this. For most small equipment, a silicon 100v 1A rectifier will be ideal. With the supply connected correctly, current passes with virtually no loss of voltage. But if the supply polarity is reversed, the diode does not conduct, so no harm is caused to the equipment. Such a diode is readily incorporated in almost any item which may from time to time be run from an external accumulator or dry battery supply where the polarity is not known.

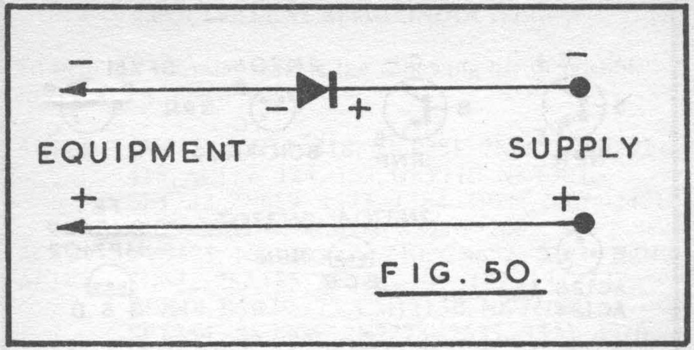

FIG. 50.

Component Tolerances

For circuits in this book, exact component values are generally not too important. Capacitor values such as 47nF (0.047μF) and 0.05μF may be regarded as interchangeable. In the same way, a 125μF capacitor may replace 100μF, or 200μF can be used instead of 220μF, and so on.

Resistors will generally be the standard 5% tolerance types. In some circuits exact resistor values are not at all critical. In others, and especially base, emitter and collector circuits, the values which provide best results can have much less latitude.

Bases

The lead-out positions of base, emitter and collector wires have been shown for typical transistors. However, it is felt that Fig. 51 should be useful. This gives lead information for types most likely to be required for circuits in this book. Other types of transistors may of course be used successfully, but it is not practical to list all these, or show connections for them all.

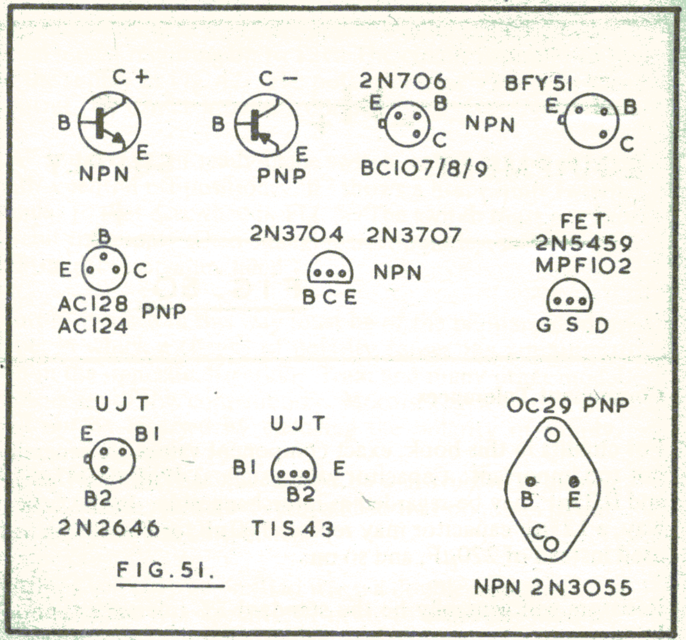
C +
B
E
NPN
C −
B
E
PNP
2N706
E
B
C
NPN
BC107/8/9
BFY51
E
B
C
B
E
C
AC128
AC124
PNP
2N3704
2N3707
B C E
NPN
FET
2N5459
MPF102
G S D
UJT
E
B1
B2
2N2646
UJT
B1
E
B2
TIS43
OC29 PNP
B
E
C
NPN 2N3055

FIG. 51.

EQUIVALENT SEMICONDUCTORS

To help our overseas readers the following list of possible equivalents is shown below: –

AC128	CV9778, OC74–318, NKT281, 2N467, 2SB222–415, AC117–124–153, GET110, NKT302, SFT232, 2N659–1373–1384–1926–2001–2431–4106, 2SB370, SK3004, RS276–2005.
ACY39	CV9475, NKT243, ACZ10, 2N652–2303–2801.
AF117	2N642, 2SA155, AF102–105–127–133–137, BFX48, BSW19–73, GM1213B, NKT603F, SFT354, SK3006, 2N2273–2635–3324–4916–5354, RS276–2003.
BC107	CV9780, AM251, BC147–167–207–317, MPS6566, SK3020–3122, ZTX107, TT107, RS276–2009.
BC108	CV10541, AM252, BC148–168–208–318, MPS6520, SK3020, ZTX108, TT108, RS276–2009.
BC109	CV10769–10806, AM253, BC149–169–209–319, MPS6521, SK3020, ZTX109, TT109, RS276–2009.
BC147	BC171–182/L–207–317, MPS6566, SK3020, RS276–2013,
BC148	BC172–183/L–208–318, MPS6520, SK3020, BS276–2009.
BC149	BC184/L–209–173–319, MPS6521, SK3020, RS276–2009.
BF200	BF162–173–225–271–311–314, SE5020/1/2/3/4, SK3018, 2N616–2708–3288–3693, RS276–2011.
BFY51	CV7723–7726–10827–11238, BFX51–68A, BFR19, BSX45/X, BSY46–85, SK3024, 40347, 2SC708, ME6102 2N2193/4–2218–2410–3053–3252–4046, RS276–2009.
BSY95A	CV7648–9379, BC108A, BSX51–73–91, BFX96, BSY20–58–62, ME9022, SK3122, 2SC321, 2N744–914–1708–2218–2369–5187, RS276–2011.
MPF102	E305, SK3116.
OA81	AA117–118–132–144, AAZ15, OA91–95–161, SD38, 1N38–56–476–618, 1S33.
OC26	CV8560, NKT452, 2N257–297–458–553–1021, 2SB3, AD130–131–138–140–149, OC28, SK3009, TF80, T1156–3027, 2N174–1073–2870–RS276–2006.

OC29	CV7083–8356, NKT402, 2N457, 2SB86–425, AD150, ASZ16, AUY21, OC28–35–36, SK3009, T13028, 2N174–1073–2870, RS276–2006.
OC44	CV5710–7003, OC170–171–410–613, 2N1303, 2SA15, AF101–117–126, ASY55, AC191, BC126, NKT211, SKT358, SK3005, 2N36–1191/2–1352–1373–1384, RS276–2003.
OC70	CV5457, OC303–602, NKT214, 2N279–1305, 2SB75, AC122–125–151–163, ACY27–34, ASY27, BC213–206A, BSX36, SK3003, 2N1193–1352–1375–2429, RS276–2004.
OC71	CV5712–7005–8344, OC70–304–604, NKT214, 2N280–1305, 2SB77, AC122–125–151–163, ACY35, BSX36, OC3041, SK3003, 2N1193–1352–1375–1384–2429, RS276–2004.
OC72	AC122/3–125/6–131–151 IV–152–162–191, BC126–213–206A, NKT211, SFT353, 2N282–1190–1352–1371–1384–1991–2431, CV5713–7006–8440, RS276–2005.
OC76	CV8314, OC307–6025, 2N284–1305, NKT212, 2SB89, AC128–131–151–152, ACY28, ASY58–76, BC126–213–260A, SK3004, 2N394–1191–1352–1371–1384–1991, RS276–2004.
ORP12	RS276–116, LDR03, RPY25.
1N4001	BY127, EM502, RS276–1101/1135, 1S100.
1N4002	BY127, EM502, RS276–1102/1136, 1S100.
2N1306	CV7350–10686, NKT736, 2N634–635–636–1891–1993–1994–1995, ASY26–29, BSX19–20, RS276–2001.
2N1613	CV7440–8843, BFY33–34–67, BSY44, 2N698–699–717–1711–1889–1890–1893, BFY41, BSX45/6, BSY53, MPS6530, 2SC708, 2N2101.

N.B.: Although equivalent semiconductors may have similar electrical properties, physical dimensions may be different and this must be born in mind if space is tight and for mounting details.
Remember polarities if replacing PNP with NPN types and vice versa. If in doubt always be advised by your dealer for suitable equivalent semiconductors.

Notes

Please note overleaf is a list of other titles that are available in our range of Radio and Electronic Books.

These should be available from most good Booksellers, Radio Component Dealers and Mail Order Companies.

However, should you experience difficulty in obtaining any title in your area, then please write directly to the publishers enclosing payment to cover the cost of the book plus adequate postage.

If you would like a copy of our latest catalogue of Radio & Electronics Books then please send a Stamped Addressed Envelope.

BERNARD BABANI (publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

160	Coil Design and Construction Manual	£1.95
202	Handbook of Integrated Circuits (IC's) Equivalents & Substitutes	£1.95
205	First Book of Hi-Fi Loudspeaker Enclosures	£0.95
208	Practical Stereo and Quadrophony Handbook	£0.75
214	Audio Enthusiasts Handbook	£0.85
219	Solid State Novelty Projects	£0.85
220	Build Your Own Solid State Hi-Fi and Audio Accessories	£0.85
221	28 Tested Transistor Projects	£1.25
222	Solid State Short Wave Receivers for Beginners	£1.95
223	50 Projects Using IC CA3130	£1.25
224	50 CMOS IC Projects	£1.35
225	A Practical Introduction to Digital IC's	£1.75
226	How to Build Advanced Short Wave Receivers	£1.95
227	Beginners Guide to Building Electronic Projects	£1.95
228	Essential Theory for the Electronics Hobbyist	£1.95
RCC	Resistor Colour Code Disc	£0.20
BP1	First Book of Transistor Equivalents and Substitutes	£1.50
BP2	Handbook of Radio, TV, Ind & Transmitting Tube & Valve Equivalents	£0.60
BP6	Engineers and Machinists Reference Tables	£0.75
BP7	Radio and Electronic Colour Codes and Data Chart	£0.40
BP14	Second Book of Transistor Equivalents & Substitutes	£1.75
BP24	52 Projects Using IC741	£1.75
BP27	Chart of Radio, Electronic, Semi-conductor and Logic Symbols	£0.50
BP28	Resistor Selection Handbook	£0.60
BP29	Major Solid State Audio Hi-Fi Construction Projects	£0.85
BP32	How to Build Your Own Metal and Treasure Locators	£1.95
BP33	Electronic Calculator Users Handbook	£1.50
BP34	Practical Repair and Renovation of Colour TVs	£1.95
BP36	50 Circuits Using Germanium, Silicon and Zener Diodes	£1.50
BP37	50 Projects Using Relays, SCR's and TRIACS	£1.95
BP39	50 (FET) Field Effect Transistor Projects	£1.75
BP42	50 Simple L.E.D. Circuits	£1.50
BP43	How to Make Walkie-Talkies	£1.95
BP44	IC 555 Projects	£1.95
BP45	Projects in Opto-Electronics	£1.95
BP47	Mobile Discotheque Handbook	£1.35
BP48	Electronic Projects for Beginners	£1.95
BP49	Popular Electronic Projects	£1.95
BP50	IC LM3900 Projects	£1.35
BP51	Electronic Music and Creative Tape Recording	£1.95
BP52	Long Distance Television Reception (TV-DX) for the Enthusiast	£1.95
BP53	Practical Electronics Calculations and Formulae	£2.95
BP54	Your Electronic Calculator and Your Money	£1.35
BP55	Radio Stations Guide	£1.75
BP56	Electronic Security Devices	£1.95
BP57	How to Build Your Own Solid State Oscilloscope	£1.95
BP58	50 Circuits Using 7400 Series IC's	£1.75
BP59	Second Book of CMOS IC Projects	£1.95
BP60	Practical Construction of Pre-amps, Tone Controls, Filters & Attenuators	£1.95
BP61	Beginners Guide To Digital Techniques	£0.95
BP62	Elements of Electronics – Book 1	£2.95
BP63	Elements of Electronics – Book 2	£2.25
BP64	Elements of Electronics – Book 3	£2.25
BP65	Single IC Projects	£1.50
BP66	Beginners Guide to Microprocessors and Computing	£1.95
BP67	Counter, Driver and Numeral Display Projects	£1.75
BP68	Choosing and Using Your Hi-Fi	£1.65
BP69	Electronic Games	£1.75
BP70	Transistor Radio Fault-Finding Chart	£0.50
BP71	Electronic Household Projects	£1.75
BP72	A Microprocessor Primer	£1.75
BP73	Remote Control Projects	£1.95
BP74	Electronic Music Projects	£1.75
BP75	Electronic Test Equipment Construction	£1.75
BP76	Power Supply Projects	£1.95
BP77	Elements of Electronics – Book 4	£2.95

BP78	Practical Computer Experiments	£1.75
BP79	Radio Control for Beginners	£1.75
BP80	Popular Electronic Circuits – Book 1	£1.95
BP81	Electronic Synthesiser Projects	£1.75
BP82	Electronic Projects Using Solar Cells	£1.95
BP83	VMOS Projects	£1.95
BP84	Digital IC Projects	£1.95
BP85	International Transistor Equivalents Guide	£2.95
BP86	An Introduction to BASIC Programming Techniques	£1.95
BP87	Simple L.E.D. Circuits – Book 2	£1.35
BP88	How to Use Op-Amps	£2.25
BP89	Elements of Electronics – Book 5	£2.95
BP90	Audio Projects	£1.95
BP91	An Introduction to Radio DXing	£1.95
BP92	Easy Electronics – Crystal Set Construction	£1.75
BP93	Electronic Timer Projects	£1.95
BP94	Electronic Projects for Cars and Boats	£1.95
BP95	Model Railway Projects	£1.95
BP96	C B Projects	£1.95
BP97	IC Projects for Beginners	£1.95
BP98	Popular Electronic Circuits – Book 2	£2.25
BP99	Mini-Matrix Board Projects	£1.95
BP100	An Introduction to Video	£1.95
BP101	How to Identify Unmarked IC's	£0.65
BP102	The 6809 Companion	£1.95
BP103	Multi-Circuit Board Projects	£1.95
BP104	Electronic Science Projects	£2.25
BP105	Aerial Projects	£1.95
BP106	Modern Op-Amp Projects	£1.95
BP107	30 Solderless Breadboard Projects – Book 1	£2.25
BP108	International Diode Equivalents Guide	£2.25
BP109	The Art of Programming the 1K ZX81	£1.95
BP110	How to Get Your Electronic Projects Working	£1.95
BP111	Elements of Electronics – Book 6	£3.50
BP112	A Z-80 Workshop Manual	£2.75
BP113	30 Solderless Breadboard Projects – Book 2	£2.25
BP114	The Art of Programming the 16K ZX81	£2.50
BP115	The Pre-Computer Book	£1.95
BP116	Electronic Toys Games & Puzzles	£2.25
BP117	Practical Electronic Building Blocks – Book 1	£1.95
BP118	Practical Electronic Building Blocks – Book 2	£1.95
BP119	The Art of Programming the ZX Spectrum	£2.50
BP120	Audio Amplifier Fault-Finding Chart	£0.65
BP121	How to Design and Make your Own P.C.B.s	£1.95
BP122	Audio Amplifier Construction	£2.25
BP123	A Practical Introduction to Microprocessors	£2.25
BP124	Easy Add-on Projects for Spectrum ZX81 & Ace	£2.75
BP125	25 Simple Amateur Band Aerials	£1.95
BP126	BASIC & PASCAL in Parallel	£1.50
BP127	How to Design Electronic Projects	£2.25
BP128	20 Programs for the ZX Spectrum & 16K ZX81	£1.95
BP129	An Introduction to Programming the ORIC-1	£1.95
BP130	Micro Interfacing Circuits – Book 1	£2.25
BP131	Micro Interfacing Circuits – Book 2	£2.25
BP132	25 Simple Shortwave Broadcast Band Aerials	£1.95
BP133	An Introduction to Programming the Dragon 32	£1.95
BP134	Easy Add-on Projects for Commodore 64 & Vic-20	£2.50
BP135	The Secrets of the Commodore 64	£2.50
BP136	25 Simple Indoor and Window Aerials	£1.95
BP137	BASIC & FORTRAN in Parallel	£1.95
BP138	BASIC & FORTH in Parallel	£1.95
BP139	An Introduction to Programming the BBC Model B Micro	£2.50
BP140	Digital IC Equivalents and Pin Connections	£3.95
BP141	Linear IC Equivalents and Pin Connections	£3.95
BP142	An Introduction to Programming the Acorn Electron	£1.95
BP143	An Introduction to Programming the Atari 600XL and 800XL	£2.50
BP144	Further Practical Electronics Calculations and Formulae	£3.75
BP145	25 Simple Tropical and M.W. Band Aerials	£1.95

BERNARD BABANI BP48

Electronic Projects for Beginners

- In this book the newcomer to electronics will find a wide range of easily made projects, many complete with actual component and wiring layouts. Furthermore, a number of projects have been arranged so that they can be constructed without any need for soldering and, thus, avoid the need for a soldering iron.

- This book which is written by the very experienced author Mr. F.G. Rayer is divided into four sections:–
  1. "No Soldering" Projects
  2. Miscellaneous Devices
  3. Radio and Audio Frequency
  4. Power Supplies

- An absolute 'must' for all beginners in electronics.

ISBN 0-85934-054-6

00195

£1.95

9 780859 340540

Popular Mechanics

COMPLETE CAR CARE MANUAL

Updated & Expanded

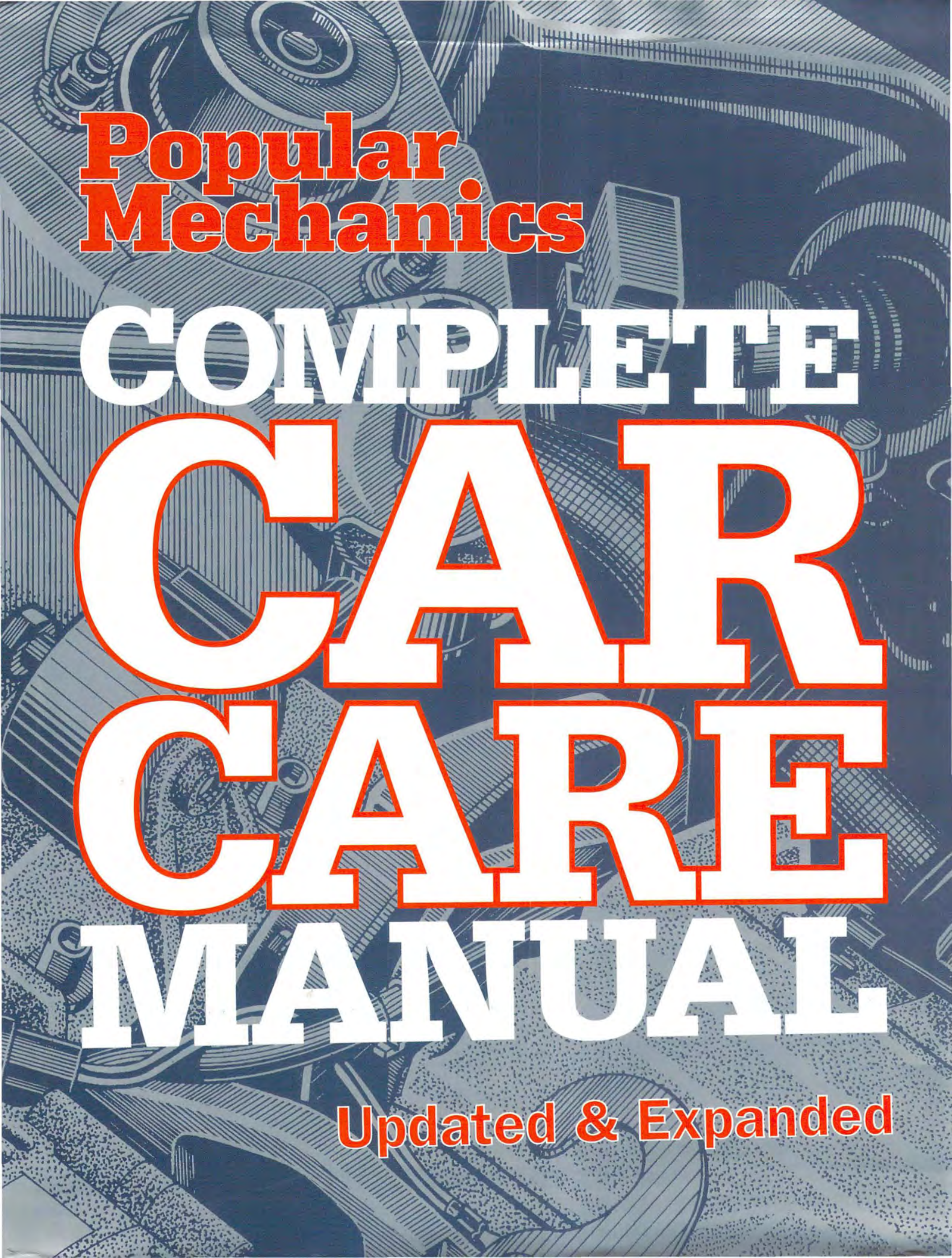

$29.95
Can. $42.95

Use this revised, expanded edition of the classic *Popular Mechanics* automotive guide to keep your car in peak condition and save money.

For more than 100 years, the editors of *Popular Mechanics* have been providing car enthusiasts with the skills and confidence they need to keep their vehicles running right and looking great. By performing any of the more than 90 maintenance and repair projects in these pages, you can avoid and fix minor problems and keep your car in perfect running order on your own.

Become familiar with the basic components of your car, from the engine, drivetrain, and electrical and electronic systems to the chassis and the interior and exterior. Learn how they work, what can go wrong, and how you can get them back in smooth working order.

Ready to try your hand at some more advanced auto care? Tune-ups aren't a thing of the past on today's computerized

(continued on the back flap)

Popular Mechanics

COMPLETE CAR CARE MANUAL

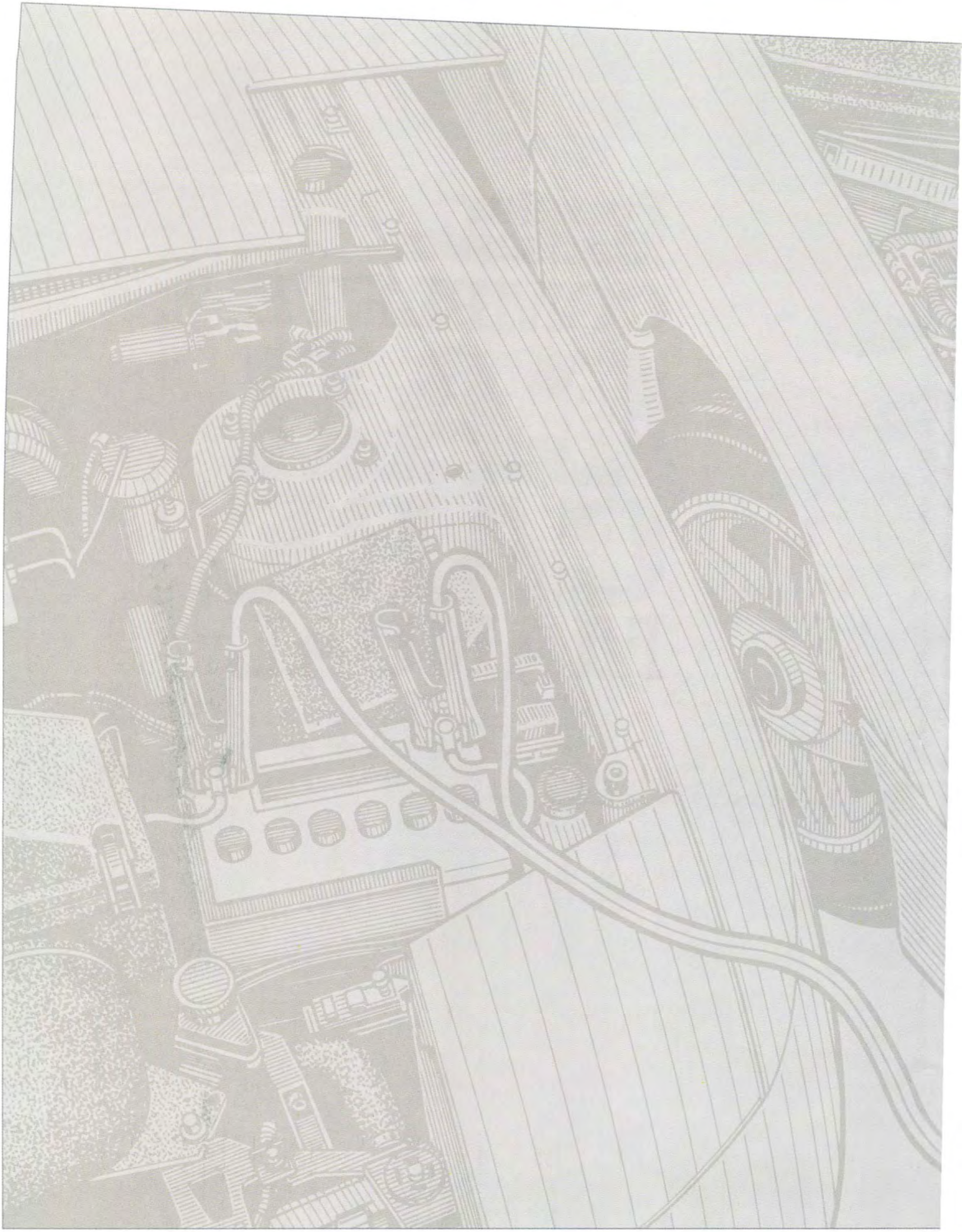

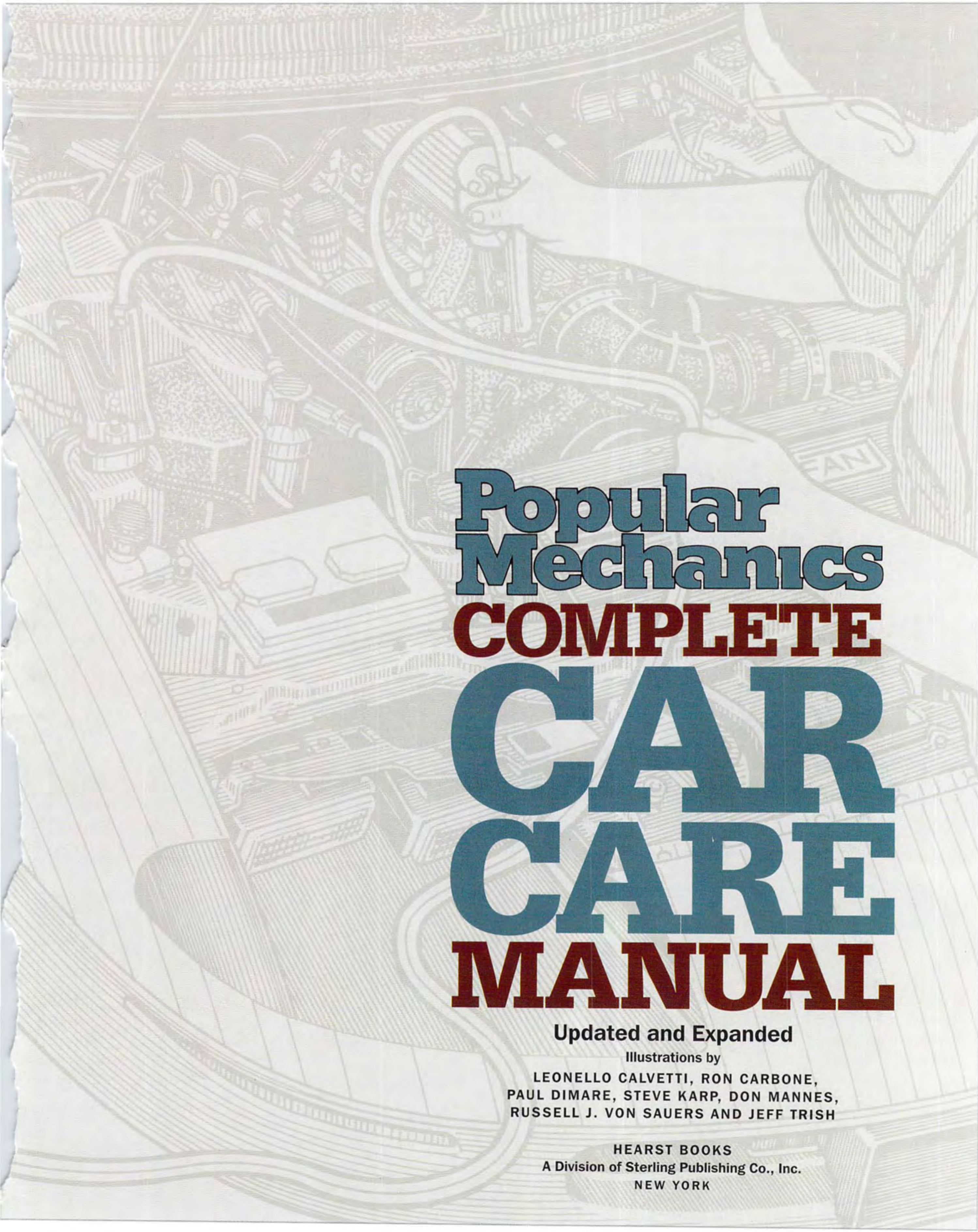

Popular Mechanics COMPLETE CAR CARE MANUAL

Updated and Expanded

Illustrations by
LEONELLO CALVETTI, RON CARBONE,
PAUL DIMARE, STEVE KARP, DON MANNES,
RUSSELL J. VON SAUERS AND JEFF TRISH

HEARST BOOKS
A Division of Sterling Publishing Co., Inc.
NEW YORK

COMPLETE CAR CARE MANUAL
Updated & Expanded

Copyright © 2003, 2005 by
Hearst Communications, Inc.

All rights reserved. The written instructions, illustrations, and photographs in this volume are intended for the personal use of the reader and may be reproduced for that purpose only. Any other use, especially commercial use, is forbidden under law without the written permission of the copyright holder.

Every effort has been made to ensure that all the information in this book is accurate. However, due to differing conditions, tools, and individual skills, the publisher cannot be responsible for any injuries, losses, and/or other damages that may result from the use of the information in this book.

Produced by Bishop Books
777 Westchester Avenue, Suite 100
White Plains, New York 10604

The Library of Congress has cataloged the first hardcover edition as follows:
Popular mechanics complete car care manual /
Illustrations by Russel J. Von Sauers,
Ron Carbone, and Don Mannes.
p. cm.
ISBN 1-58816-260-5
1. Automobiles—Maintenance and repair—
Amateurs' manuals. I. Popular mechanics.
TL152.P625 2003
629.28'72—dc21 2003050841

10 9 8 7 6 5 4 3 2 1

Published by Hearst Books
A Division of Sterling Publishing Co., Inc.
387 Park Avenue South, New York, NY 10016

Popular Mechanics is a trademark owned by Hearst Magazines Property, Inc., in USA, and Hearst Communications, Inc., in Canada. Hearst Books is a trademark owned by Hearst Communications, Inc.

www.popularmechanics.com

For information about custom editions, special sales, premium and corporate purchases, please contact Sterling Special Sales Department at 800-805-5489 or specialsales@sterlingpub.com.

Distributed in Canada by Sterling Publishing
c/o Canadian Manda Group, 165 Dufferin Street
Toronto, Ontario, Canada M6K 3H6

Distributed in Australia by
Capricorn Link (Australia) Pty. Ltd.
P.O. Box 704, Windsor, NSW 2756 Australia

Manufactured in China

ISBN 1-58816-439-X

Contents

Maintenance Basics 2

Engine 88

Drivetrain 138

Electrical and Electronic Systems 158

Chassis 214

Interior and Exterior 254

Appendix 314

Foreword

For almost 100 years, the editors of *Popular Mechanics* have made car care and maintenance a central component of the magazine. From the very beginning, we understood just how passionate readers were about their cars, and just how far they would go to protect, repair, and maintain their beloved vehicles. How else to explain an executive trading in his workday suit for a pair of overalls and spending his Saturday afternoons hunkered down under the hood of his car? For some of our readers, saving money is no doubt a major motivation. But for most, working on their cars is a labor of love, pure and simple. That, and they want to make sure the repair is done right. The cars may have changed over the years, but the dedication of readers has been a constant. And *Popular Mechanics* has been a partner all the way, delivering the hands-on information car enthusiasts need to get the job done right.

We understand that the key to any successful piece of advice is clarity. So we strive to make sure our instructions and our illustrations are legible and to the point. We hope you'll find those qualities in evidence in the pages that follow. You'll find chapters on all the basic car components—engine, chassis, electrical system, drivetrain, and interior/exterior—as well as an opening chapter on the essential maintenance information that every owner should know. Each project is labeled "easy," "moderate," or "difficult" to allow you to make your own decision as to whether you have the requisite expertise, not to mention the time and energy, to undertake a given procedure. So, whether you're breaking out the brand-new wrenches from your first tool set, or sliding under a car on your creeper for the umpteenth time, you'll find this book to be an indispensable reference source, packed with easy-to-follow, step-by-step instructions that take the mystery out of auto maintenance and save you money. And, should you decide that a given repair is a bit too tricky for you, the book will arm you with the knowledge necessary to discuss the work intelligently with your professional mechanic.

Our goal is a simple one: We want to provide you with the information you need to keep that treasured car of yours in peak condition regardless of its make, model, or age. We know how much you love your car and how dedicated you are to maintaining it. We're hoping our book provides you with a way to put that dedication into action.

The Editors
Popular Mechanics

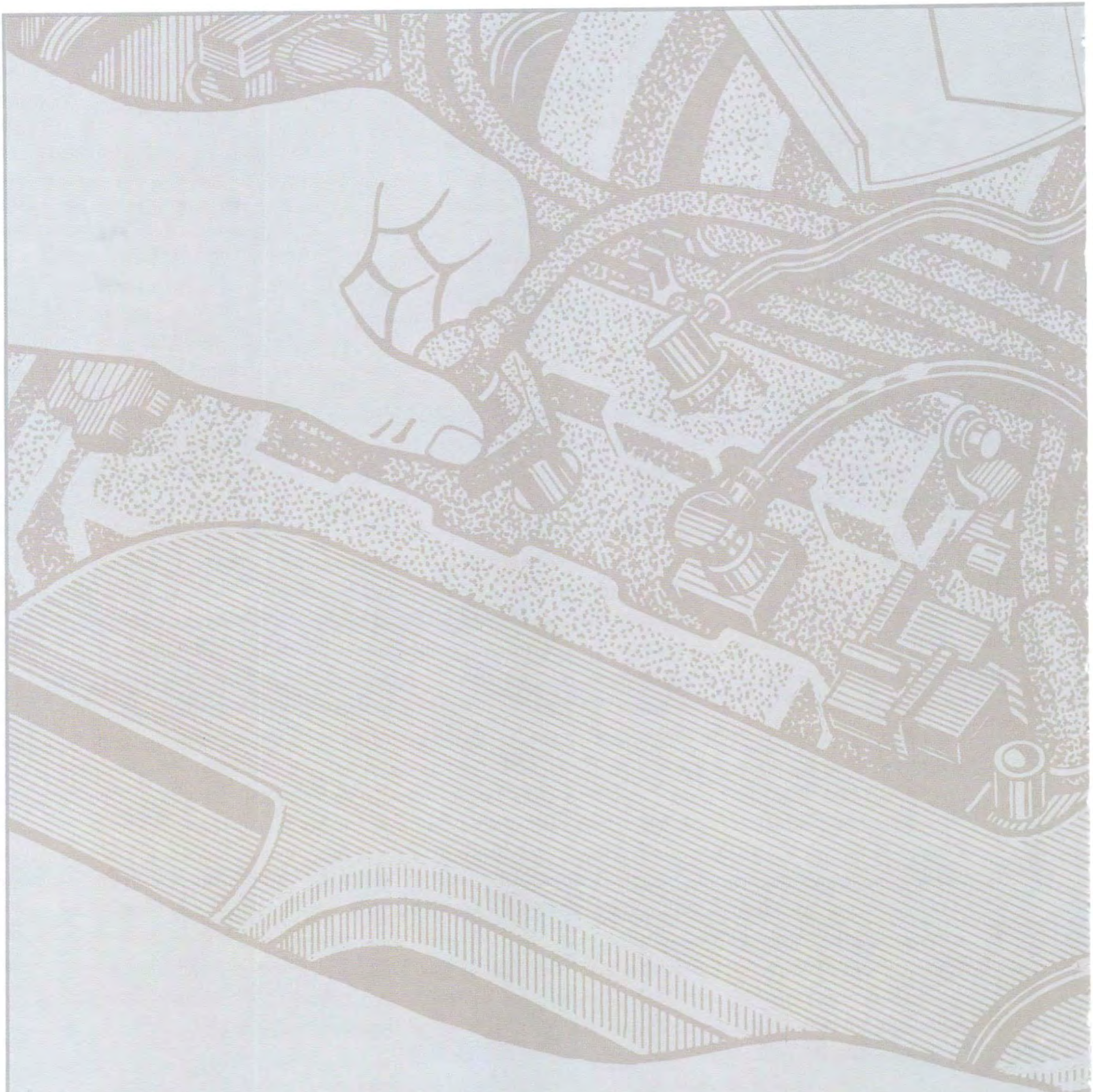

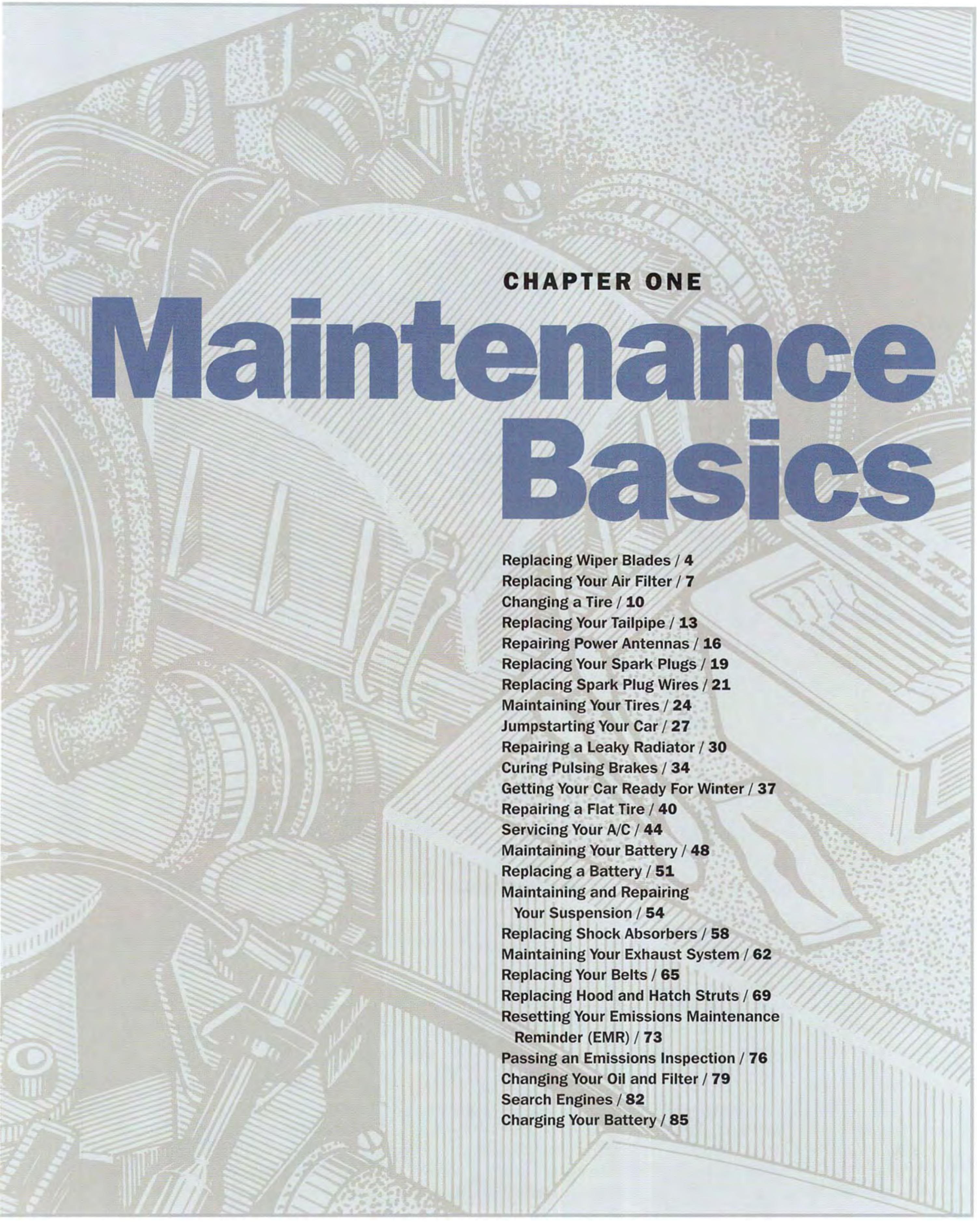

CHAPTER ONE

Maintenance Basics

Replacing Wiper Blades / **4**
Replacing Your Air Filter / **7**
Changing a Tire / **10**
Replacing Your Tailpipe / **13**
Repairing Power Antennas / **16**
Replacing Your Spark Plugs / **19**
Replacing Spark Plug Wires / **21**
Maintaining Your Tires / **24**
Jumpstarting Your Car / **27**
Repairing a Leaky Radiator / **30**
Curing Pulsing Brakes / **34**
Getting Your Car Ready For Winter / **37**
Repairing a Flat Tire / **40**
Servicing Your A/C / **44**
Maintaining Your Battery / **48**
Replacing a Battery / **51**
Maintaining and Repairing Your Suspension / **54**
Replacing Shock Absorbers / **58**
Maintaining Your Exhaust System / **62**
Replacing Your Belts / **65**
Replacing Hood and Hatch Struts / **69**
Resetting Your Emissions Maintenance Reminder (EMR) / **73**
Passing an Emissions Inspection / **76**
Changing Your Oil and Filter / **79**
Search Engines / **82**
Charging Your Battery / **85**

Replacing Wiper Blades

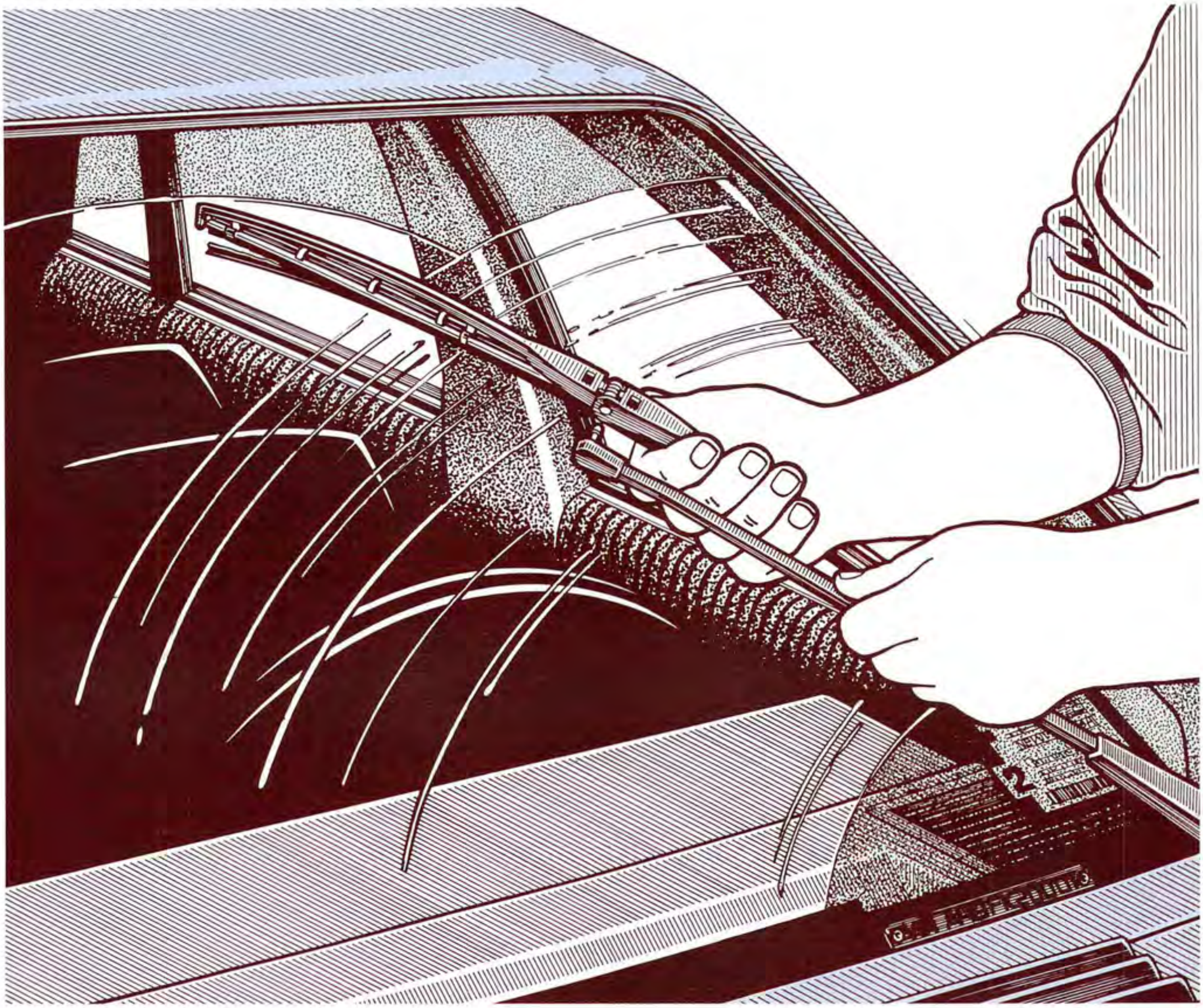

(Fig. 1) In 10 minutes, you can replace your streaky view of the world through a windshield with fresh wiper blades.

Through the monsoonlike rain, you see some sort of flashing lights ahead of you. But despite the best efforts of your windshield wipers, all you can see are blurry blobs of yellow, amber and red. Slowing to half speed, you continue to press forward through the deluge. Suddenly you realize that those flashing blobs are the warning lights of a stopped school bus, and children are running across the road in front of you. Fortunately, your brakes work infinitely better than your windshield wipers.

Once it's dry and sunny again, grab a tape measure and head for the garage. Lift one of the wiper arms (it's usually easier to grab the driver's-side arm) off the windshield against its spring tension and keep lifting until the pivot point locks the arm upright **(Fig. 1)**.

Look for contaminated rubber inserts, which can be caused by road film or car-wash chemical adhesion. Inspect the wiping edges for "park set rubber," the term used to describe hardened finely-cracked inserts that have been exposed to the sun too long in a parked position. This will cause chattering and skipping. Check for a rubber insert that has been partially torn away from its metal support. It will slap the windshield with each wiping pass.

Plain old dried out, cracked rubber inserts mean they've seen better days. You also may find rubber inserts that are abrasion-worn from winter conditions, infrequent car washes and/or a

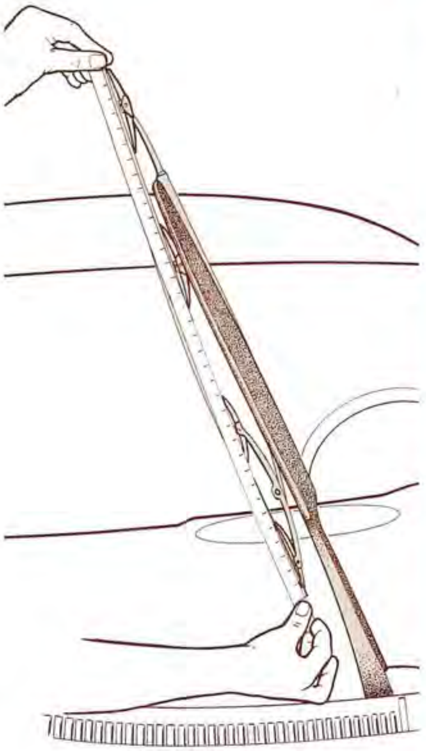

(Fig. 2) Measure your old blades before heading down to the parts store.

lack of preventive maintenance. And rear-window wiper blades are subjected to a lot more road grit than front blades. In the rear, you get aerodynamic backwash, with the rear tires kicking up all sorts of debris. And don't rule out an improperly installed refill. Maybe someone else had tried to replace the inserts before you got to them. Buy a pair of full-blade assemblies to get a factory fit. Also check the wiper arms. If they're okay, proceed to replace the blades.

Blade Runner

But first, measure the length of the wiper blade to the nearest inch (Fig. 2). Most vehicles today use blades anywhere from about 16 to 21 in. in length. However, some cars or minivans with very large windshields may exceed that by a few inches.

Parts catalogs, even those for simple items like wiper blades, are full of all sorts of extraneous information. You may find three different types of refills and three different types of blades for almost every vehicle listed.

HOW IT WORKS

Windshield Wipers

As the rubber insert of a wiper blade sweeps across your wet windshield, a wedge of water builds up in front of the rubber—a squeegee effect—and the wiper displaces that water elsewhere. What's left is clean, clear glass that allows you to continue on your way safely. That's the way it's supposed to work. The illustrations to the right show some of the reasons why it may not.

Other reasons may involve components that work behind the scenes—actually, behind the cowl or dashboard of your car. They include the splined shafts that the wiper arms pivot on, the mechanism that creates the back-and-forth pivoting motion and the electric wiper motor that drives the whole system.

A small nut under a cosmetic plastic cover jams the wiper arm onto the splines of its shaft. The splines keep the shaft from slipping in the arm as it turns in one direction and then the other, over and over and over. This reciprocating motion is created by a metal crank-type linkage assembly that's attached to each wiper arm's splined shaft, and to the wiper motor with another splined shaft. Picture the chugging action of a steam locomotive's piston and drivewheels and you'll get the idea.

Rear wipers—found on minivans, SUVs and some sport coupes—work the same way, except the reciprocating mechanism is much smaller and built into the motor.

Wiper speed depends on the amount of voltage that's sent to the motor from the wiper switch—low voltage equals low speed, high voltage equals high speed. Intermittent wiper action is created by a separate electronic module wired between the switch on your dashboard or steering column stalk and the motor. Some cars even have electronically controlled road-speed-sensitive wipers that wipe faster the quicker you go.

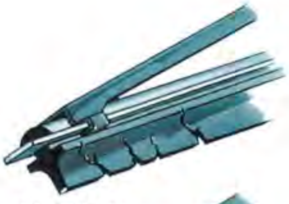

CRACKED RUBBER

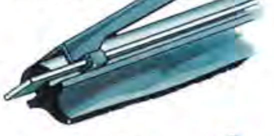

CONTAMINATED RUBBER

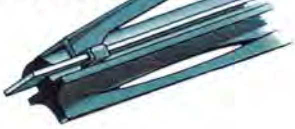

TORN RUBBER

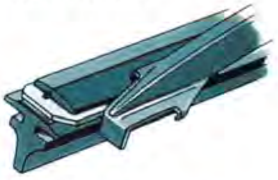

IMPROPERLY INSTALLED REFILL

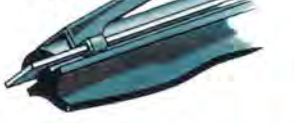

PARK SET RUBBER

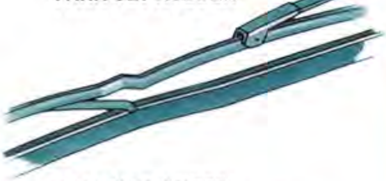

DAMAGED SUPER-STRUCTURE

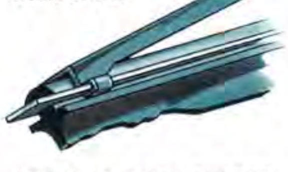

ABRASION-WORN RUBBER

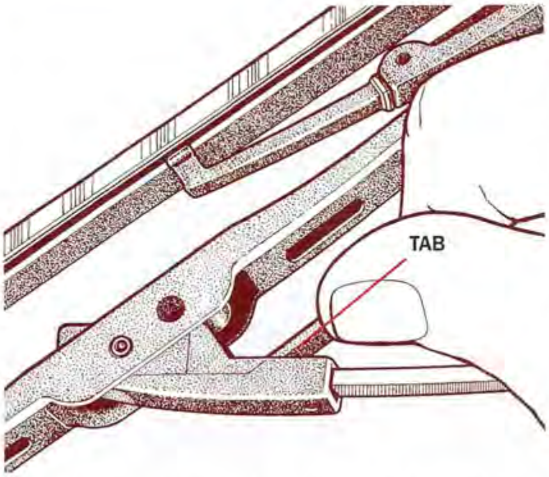

(Fig. 3) Release the tab to disengage the blade from the straightened arm.

If you don't know what brand of blade your car has, don't buy rubber insert refills. The refills may not fit the blades properly, and you'll get lousy wiper performance.

If you do know what brand of blade is currently on the car, refills are okay as long as you buy the same brand and type. Experience has shown that mixing blade and refill brands doesn't lead to optimum wiper performance. But realistically, you're still better off just buying a pair of blade assemblies no matter what the brand. The few extra bucks get you new, matched components that mount quickly and easily.

Fresh Rubber

Now that you've got new blades in hand, it's time to determine what method to use to mount the wiper blade to the wiper arm. There are many variations: hook-slot connector, pin-type arm, wide-straight end, narrow-straight end, side saddle, pin-type blade, narrow-dead locker, flat hook and rock to lock.

At the very least, you'll need a small screwdriver to gently pry the blade off the arm or to unlock a tab of some sort. Sometimes a pair of needle-nose pliers helps too, but the pliers tend to mar the wiper-arm finish more than a screwdriver.

Bear in mind that most wipers will stay raised in an upright position away from the windshield so you can work on them. Others will not. So if this is the case, you'll have to hold the blade off the glass while you work. If your car has wipers that park behind a cowl or the lip of the hood, turn on the ignition, turn on the wipers and then shut off the ignition when the wipers are in midstroke so you can have access to them.

Always remember to lay a clean rag on top of the windshield to protect the glass in case the bare-metal wiper arm suddenly snaps down on the windshield's glass.

For the hook-slot connector, you generally need to raise the arm off the windshield to a working height. Then, swing the blade perpendicular to the arm so that you have more light on the connector. You'll find some sort of tab at the connector that either needs to be lifted or pushed to release the lock that holds the blade pivot in the hook slot. Once the tab is released, a firm shove toward the base of the arm will slide the blade right out of the hook slot (**Fig. 3**). Then, simply line up the new blade's mount (there may be an adapter in the box to create this mount) with the hook slot and slide it in until it clicks. This means it's locked in place.

For the pin-type arm, raise the arm to a working height and take a close look at the type of locking tab that holds the blade onto the pin. You'll either have to push the tab up from underneath the blade or lift the tab from the top with a screwdriver (**Fig. 4**). Once the tab "uncollars" the pin, pull the blade assembly with a sideways motion away from the arm. (Try to imagine pulling the blade up the windshield, to picture the motion.) Depending on the blade manufacturer, the new unit may simply lock onto the pin. Sometimes, the box will contain an adapter that locks into the blade opening and accepts pins in different sizes.

The straight-end connector can be tricky. Sometimes you need to lift a tab on top of the existing blade with a screwdriver to clear the locking nub on the arm. You may also need to cock the old blade a few degrees out-of-parallel with the arm to slide it off the arm's end.

Once the old blade is off, applying a little antiseize compound to the arm's end couldn't hurt. The new blade may come with an adapter to fit the arm's end or it may just slide on. Again, give a firm shove until the blade clicks into place.

If you're replacing a rubber insert, just grab it with needle-nose pliers at the open end. The other end generally has some sort of deadstop that prevents the insert from sliding out. You may need to unlock a tab or squeeze together the locking tangs of the insert before you can begin pulling. Then, just hold the blade firmly while you pull the insert out like a piece of spaghetti. Grab the new insert—make sure that you've got the right end if they're different—line it up with the first set of guides on the blade, then feed it in carefully. Make sure to engage each set of guides as you go. Sometimes there are as many as eight guides. Once the insert is all the way in, make sure it is locked in place.

(Fig. 4) With pin-type wiper blades, push down on the pin lock to release the wiper blade from the wiper arm with a sideways motion.

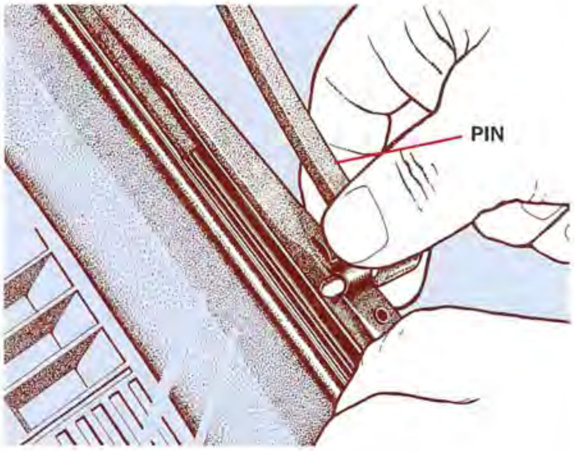

Replacing Your Air Filter

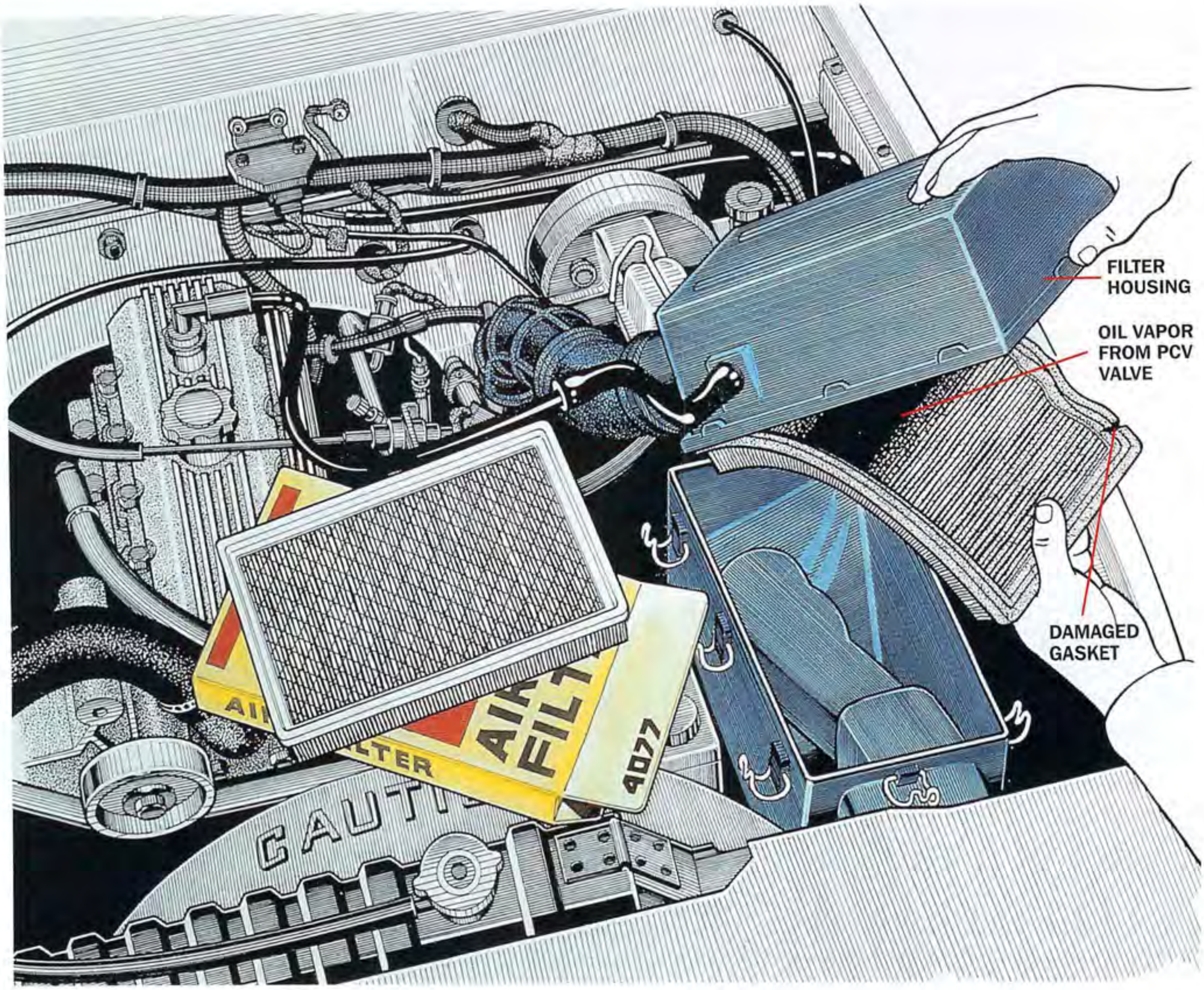

(Fig. 1) Check your old filter element for dirt, oil or moisture, cracks and proper sealing at the gasket.

Your mileage has been dropping steadily for months. A tuneup is in order, so a fine Saturday morning finds you on a deserted stretch of road. You figure a few miles of spirited driving should blow out the cobwebs and restore that ponycar performance. But what happens instead is a lot of noise and not much performance at all. Time to check the air cleaner—which is plugged with dirt, bugs and leaves.

Every engine has an air filter, and although there have been many types of air-filtering materials tried over the years—including oil-wetted plastic foams and wire mesh—dry, pleated "paper" (actually resin-impregnated, heat-cured fiber) continues to be the material of choice for filter elements in everything from cars to heavy-duty trucks and off-road equipment.

Housing Project

Check your air filter at least once a year. All air-filter housings are somewhere under the hood (**Fig. 1**). The housing cover may be held by wingnuts or spring clips—pretty straightforward stuff. Separate the cover from the housing and you can lift the filter out for inspection. Many filters, particularly the narrow cylinders used on older rear-drive cars and even on many of today's trucks, can be rotated. If there's a dirty area near the air intake, tap the filter against a table to remove loose dirt, then turn it 180° and expose a fresh surface to

(Fig. 2) If a conical filter has a dirt buildup on only half the circumference, it can be rotated 180° to equalize dirt pickup, extending service life. Clean the housing's interior before assembly.

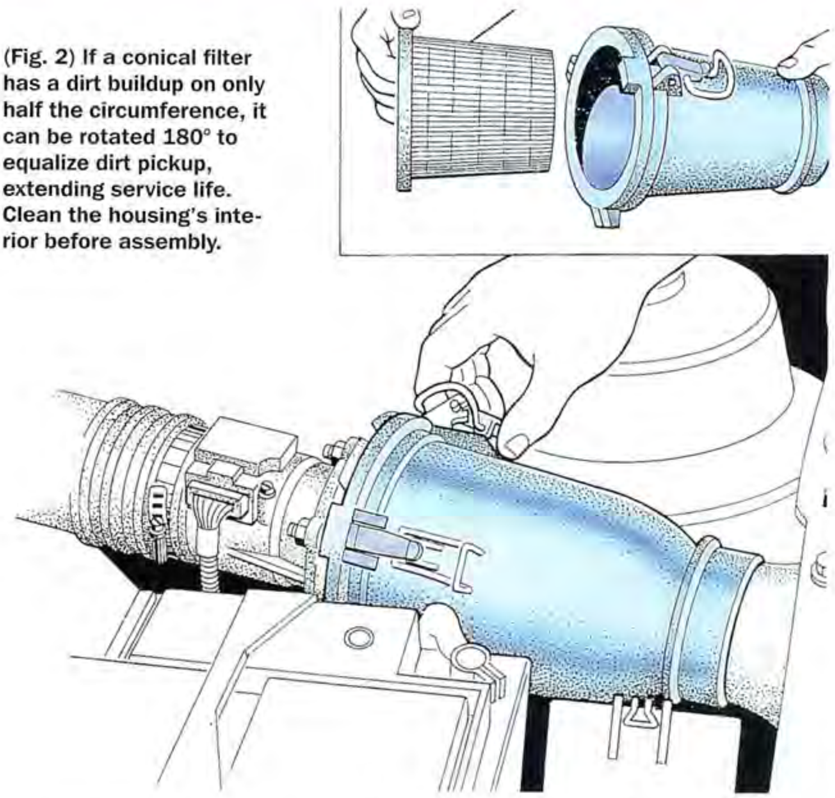

the air intake (**Fig. 2**). Some conical filters also can be rotated, but others can't. Never try to clean a "paper" automotive filter in detergent and water. Some heavy-duty filters are designed to be cleaned several times, but automotive filters are not.

Before you reuse a filter, however, inspect it for cracked pleats. The best way is to hold the dirty side of the filter in front of your eyes with good lighting on the opposite side (with a cylindrical filter, hold a trouble light in the middle of the opening). Of course, you should see some translucence (if you don't, the filter is plugged). But if you see even a single line of light, which indicates a hairline crack, discard the filter. A lot of unfiltered air can get through a crack, and it's the abrasive dust in the air that wears engines. Sure, that wear takes place over time, but if you're in a dusty area that time can be a lot shorter than you'd like.

Buy a quality replacement filter. The top brands are different from the no-name cheapies you may see. There are many grades of filter "paper" and the more expensive ones used by the top brands have more consistent pores, so they do a better job of trapping dirt. In addition, their pleats are more precisely spaced so the filter has the maximum surface area for the size of the housing.

Metering the Air

Just downstream of the filter housing, perhaps connected by a duct, will be the engine's airflow meter, or sensor. (Some engines have no airflow meter. Instead, their computer calculates air intake using rpm, intake vacuum and barometric pressure.) Contamination of the airflow meter will give your engine's computer inaccurate information about the amount (mass) of air the engine is burning. That could mean poor driveability, poor mileage and hard starting.

Don't, however, try to clean the airflow sensor, save for blowing loose dirt and cobwebs with your breath. The calibration of these instruments is fragile, and you stand a good chance of mucking your sensor up irreparably.

Duct Hunting

Before you install the new filter, inspect the duct from the throttle body to the filter housing. Spread the flutes of the duct, looking for cracks. If the duct is cracked, it may be closed at idle but spread on acceleration, when the engine torques. That not only brings in unfiltered air, but on today's engines with mass airflow sensors, the dirty air also is "unmeasured." The amount of fuel sprayed by the injectors is determined by how much air is flowing in, so if some of that airflow isn't measured by the sensor, not enough fuel will be injected. The engine will run lean and hesitate and/or stumble. The same thing can occur if the duct isn't properly fitted and clamped at each end.

Find a crack? Well, it's a duct, so you can repair it with duct tape (**Fig. 3**). That's a move that no professional would make, because in time the tape will lose its grip in the heat of the engine compartment, as the duct flexes. But it's your car, and if you are willing to recheck periodically and retape as necessary, you can save the cost of the new part, or at least postpone the expense.

Fit Is Important

Carefully check the fit between the lips of the housing and the rubber gasket rimming the filter. This gasket not only seals the filter housing's two halves together, but also provides an airtight and dirt-tight seal between the filter and the clean side of the housing. Look carefully at the old filter's gasket. It should show a smooth, clean line 360° around at the sealing face on both sides. The rim of the filter housing will have compressed the rubber gasket if everything is correct. If not, a little detective work is in order. Is the filter the correct part? If not, is your replacement filter the correct part? Don't rely on just a visual comparison—look up the correct part number and cross-reference. Ad it's possible another brand will fit better.

Are all of the spring clamps properly latched? Some air cleaners use screws to hold them shut, while others use a large wingnut. All of the fasteners need to be tight, even that one

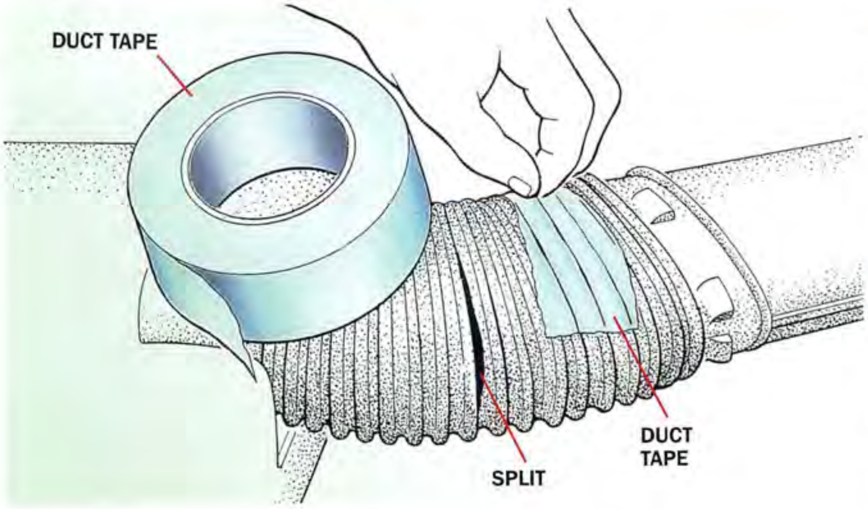

(Fig. 3) Check the air cleaner's ductwork for leaks and splits, especially if you're having driveability problems. Temporary repairs can be made with ordinary duct tape.

you can hardly see or reach next to the battery. On a few vehicles, the housing is attached by a bolt or spring clamp that can be removed to allow the housing to be displaced an inch or two, permitting big hands to fit into a narrow space.

Ready to install the filter element? Clean out the housing first, wiping the inner surface with a cloth lightly moistened with solvent. Insert the element so it seats properly. A filter has a specific side up or toward the throttle body, and you have to install it correctly so the end seal lines up with the shape of the housing or its cover. If it doesn't, you won't get a good end seal, and incoming air may bypass the filter and allow dirt into the engine.

Buttoning Up

It's time to finish up. Install the cover or the halves of the filter housing. In addition to the spring clips and wingnuts that hold the parts together, there also may be alignment tabs in areas you can't reach with your fingers. Be sure the tabs are engaged as you fit the halves together, so the two parts of the housing form an airtight joint when you turn the wingnuts or engage the spring clips.

Be sure that both ends of any ductwork are properly sealed to both the air cleaner body and the airflow meter or throttle body, including hose clamps or retaining clips.

HOW IT WORKS

Your Air Filter

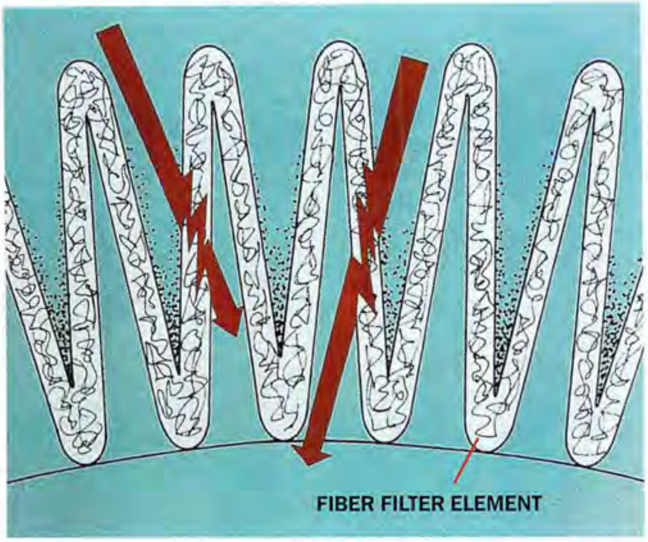

The resin-impregnated, heat-cured "paper" element looks like a pleated sheet, and most of the dirt particles are stopped at the outside surface. A light coating of dirt actually improves the filtering ability of the paper. But soon the dirt builds up and even starts filling the bottoms of the pleats, restricting airflow. The paper also has multiple layers, and if you looked at it under a microscope, it would look like a forest with an irregular crisscross of vines and limbs—the fiber strands of the "paper."

It all looks random, but the premium-priced filter "papers" are more consistent in the sizes of the pores than it seems, and a quality material is a mathematically predictable barrier to dirt particles. In actuality, the layers of fibers form a maze and some of the dirt also is trapped within the layers of the paper. The inrushing air produces some static electricity, which also causes some smaller dirt particles to adhere to the surfaces of the paper fibers.

A point comes when not enough voids remain to pass the air needed to supply your engine. The result is inadequate oxygen to burn the fuel at wide-open throttle. With carbureted engines, this used to mean rich running. Modern fuel-injected engines meter the air admitted to the intake and add appropriate fuel—making for a clean-running engine that eventually won't get out of its own way. At least carbureted cars coughed and sputtered to let you know that you had a problem.

But with today's engines, a filter has to be really bad to restrict acceleration. A bigger hazard is an overloaded filter rupturing and dumping 20 miles of bad road into your engine.

Changing a Tire

Flat tires are never a fun thing—but this one takes the cake. It's late, of course, and in the middle of a frog-choking rain. You're stuck in the ankle-deep mud on the shoulder of a deserted road. How deserted? Out-of-cell-phone-coverage deserted, or you'd be sitting in the cab of a service truck while somebody else gets drenched. That's how deserted.

Yes, you could drive along the shoulder on the rim for a few miles to civilization, but insurance won't cover the damage to your expensive alloy rim. It's time to knuckle down and put on the spare.

A pressure can of flat-fix foam can get you home if the problem is a simple puncture, and that may be a viable option, especially for smaller individuals or the elderly who would have a tough time changing a tire. Just remember two things: This stuff is a temporary solution, and the flat will need to be attended to by a tire technician at the earliest opportunity. Be sure to warn the technician that you've used this stuff. The propellant is flammable, and unless he's warned, he stands the chance of causing a nasty explosion. Aerosol flat-fix, however, is no help if the problem is a tire unseated from the wheel rim after an encounter with a pothole.

Be Prepared

First of all, remember to check the pressure in your neglected spare tire (Fig. 1) whenever you check the pressure in the other four, which you do faithfully every month or so, right? A flat spare is no help. And if you're like most people who save a plugged or nearly worn tire for the spare, it's likely that the spare has a slow leak, which would leave you stranded.

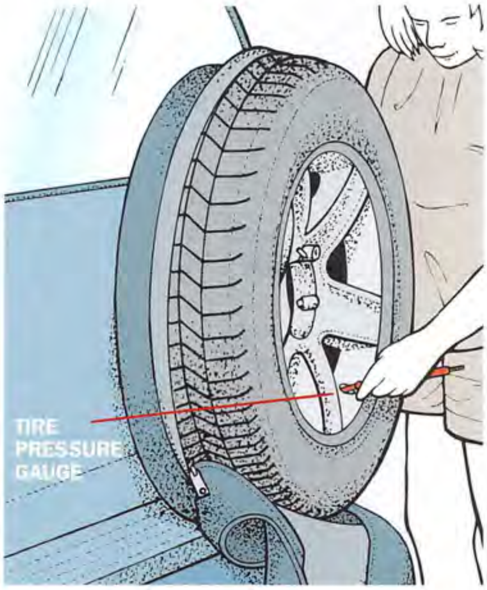

(Fig. 1) Be sure to check the air pressure in your spare tire periodically.

Furthermore, you'll never be able to change a tire if you don't have the basics—a jack and a lug wrench (Fig. 2). Go back into their hidy-hole and confirm their existence (Fig. 3). While you're there and you have enough light to read the owner's manual, figure out how to unship them and make sure the jack isn't rusted into immobility. Check the manual and find the vehicle's jacking points. You'll probably need to lie on the ground to find them, but this will be a lot more palatable now in your driveway than later on the shoulder of some mud bog.

Grunting Helps

The most common difficulty in changing a tire is lugs that are tightened far too tightly. A casual rattle with a mechanic's air wrench can deliver a tightening torque that only a pro wrestler could remove with the stock lug wrench. Wheel lugs need to be tightened to no more than the manufacturer's recommended torque to ensure they won't loosen. Check your

(Fig. 2) Loosening lug nuts calls for a lot of upper-body strength. An inexpensive 4-way lug wrench can substantially increase your leverage and break loose overly tightened lugs.

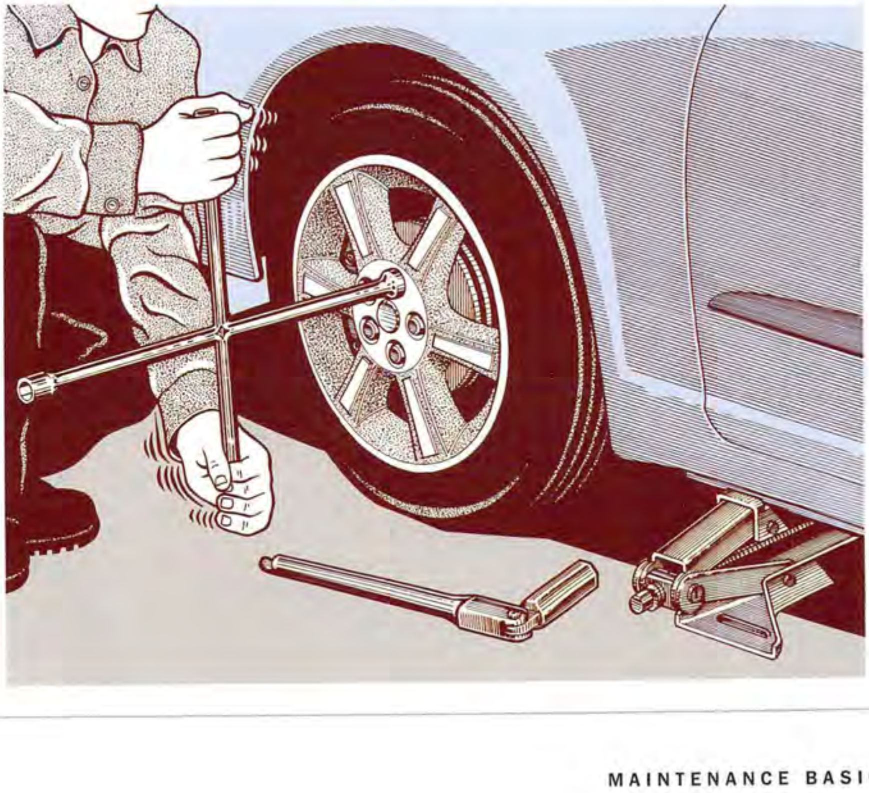

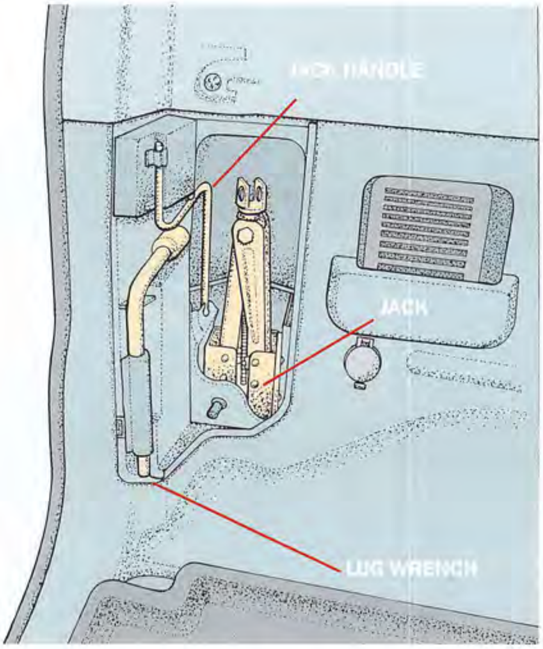

(Fig. 3) Locate the jack and other tire tools and make sure you know how to use them.

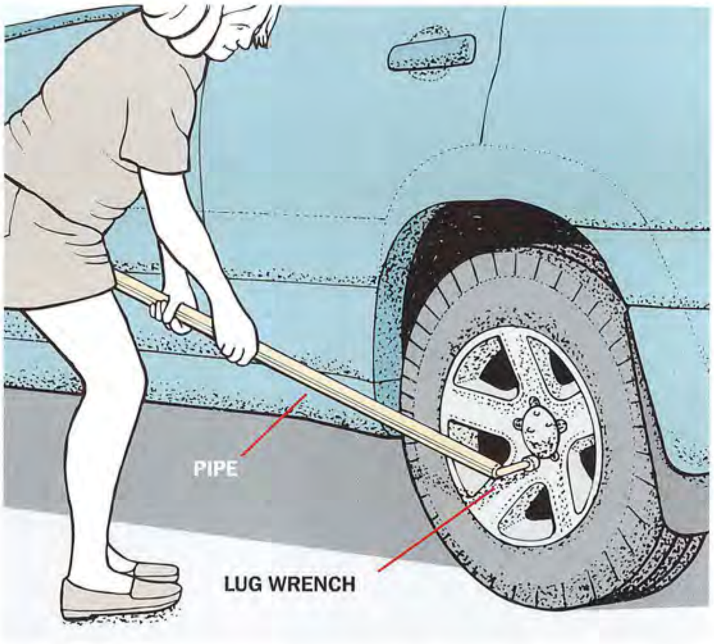

(Fig. 4) A cheater bar helps loosen lugs. Never use it to tighten them.

owner's manual, but the figure will be 75 to 100 ft.-lb. of torque. Do the math—that means a 200-pound adult should be able to tighten the lugs by placing all his weight on top of the wrench only 6 in. away from the fastener. A few drops of engine oil or grease on the threads and the lug chamfer (where the lug touches the wheel, not the threads) will prevent galling and seizing.

When removing a wheel, first loosen all of the lugs in a crisscross pattern a half-turn or so. It may be necessary to use the mechanic's favorite cheater bar (Fig. 4)—a piece of water pipe or thick wall pipe about 4 ft. long—to add enough leverage to break the lugs loose. Remove them one at a time and lube them if they squeak. Retighten them in three stages, again in a crisscross pattern.

Black And Round

Many carmakers, in an attempt to reduce vibration by making sure the wheels are more concentric with the hub, use a protruding lip that mates closely with the centerhole on the rim. This works well for a couple of years, but eventually corrosion from road grime can make it impossible to budge the wheels, even after you've loosened all the lugs. Correct this now, and you won't need to try it in the field. Jack up the corner of the car to take the load off the wheel and kick the wheel, alternating sides until it pops loose. No joy, and now your feet hurt? Be sure the lugs are only a single turn from tight, lower the car and move it a foot forward and back, rapping the brakes smartly to break the wheel loose. A shot of penetrating oil may help in an hour or two. Clean up the corrosion with emery paper (Fig. 5) and coat all the surfaces lightly with wheel bearing grease, Vaseline or, best of all, antiseize compound.

You'll need a few things besides the on-board tools for your emergency tire change. Pack a flashlight with good batteries or, better yet, a cigarette-lighter-powered trouble light, an emergency triangle, a couple of road flares, some gloves and a poncho or ground cloth. Toss in three pieces of scrap 2 x 4 as well, each about a foot long, and a piece of thick plywood 2 x 2 feet.

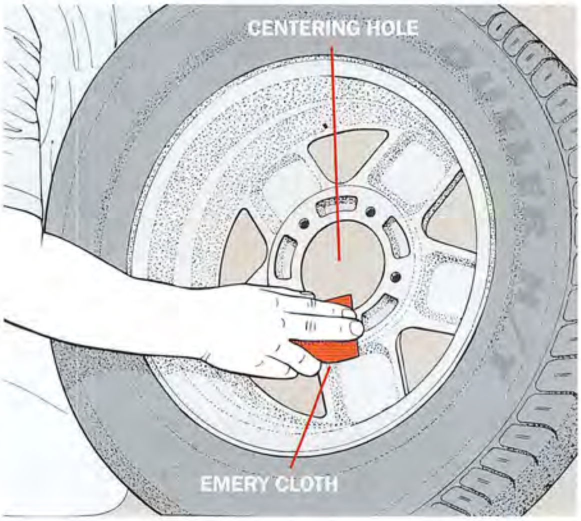

(Fig. 5) Remove corrosion from the centering hole with emery cloth or a wire brush.

That Fateful Day

When changing a tire the first, and most important, thing to do is to get the vehicle to a safe area, far enough from the road to save you from becoming road pizza—particularly if the flat is on the left side and your back side will be poking out into traffic while you work. Set your triangle or flares 100 ft. or so upstream.

Leave the vehicle in Park and set the handbrake. Block the wheel diagonally opposite the flat with two pieces of wood (Fig. 6). Loosen all the lugs on the flat a full turn. Take the spare out of the trunk and put it halfway under the car near the jack. In the unlikely event the car falls off the jack, it will only fall onto the spare—not your foot or head—and will leave you a fighting chance of raising the car and continuing. If the car falls to the ground, you'll have no way to raise it.

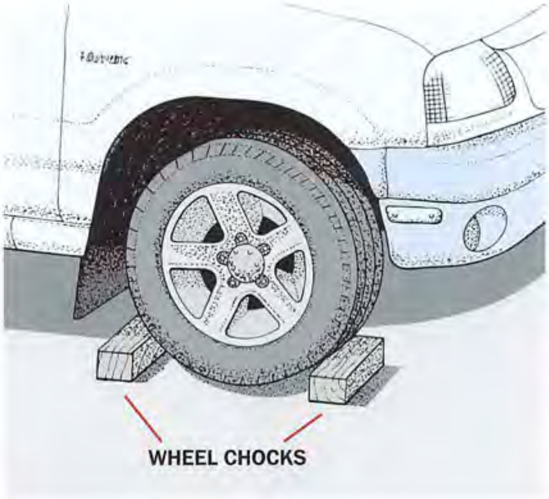

(Fig. 6) Block the opposite corner of the car to keep it from rolling off the jack.

Raise the jack from its stowed position to nearly high enough to contact the bottom of the car. If you've got a different style of jack than the one pictured here (Fig. 7), check your owner's manual for specifics.

If the ground isn't firm, put the 2 x 2 piece of plywood under the jack point, and the jack on top of the wood. Be sure everything is level. Jack the car up until the flat clears the ground by several inches, because the spare isn't flat and will need more clearance.

Remove the lug nuts, and put them inside the hubcap or in some other place where they won't get lost in the dark, or accidentally scattered into the weeds by your feet.

Pull the flat off and put it halfway under the car. Hang the spare on the studs. No studs? You've got lug bolts (common on many European cars) and you'll need to juggle the wheel while you get the top one started. The easy way is to sit down on the ground facing the hub and balance the spare on your legs while you start the first lug bolt. If you begin with the top bolt, the wheel will hang gracefully from it and you can start the rest. Finger-tighten all the lugs and then lightly tighten them with the lug wrench, again in a crisscross pattern.

Be sure you don't have anything (like a stone or mud) trapped between the rim and hub, or the rim will wobble. In fact, if the mud compresses later, the lugs could lose their torque and the wheel could fly off. If you need to, remove the wheel again in the morning to remove the debris, corrosion and rust from all the mating faces and between the wheel and hub, and then lightly lube. This will have the added benefit of making the wheel easier to remove the next time. Hint: There should be some grease to be found on the jackscrew of your jack, and there's probably enough to put a smidgen on the lugs with your finger. Be sure to get it on the mating chamfer as well as on the threads.

Lower the vehicle and pull the jack out. Now you can tighten the lugs to their correct torque. Check the owner's manual for the torque specification. Measure carefully. If the lug is dry and unlubricated, it may take a lot more force to tighten the lugs.

The only accurate way to torque the lugs is to use a mechanic's torque wrench. These can be purchased for 20 bucks or so. If you have expensive alloy wheels, you may want to buy one and keep it in the trunk.

As you're putting away your tools and jack, be sure you haven't left them covered with mud or moisture, which might cause them to rust while stored. If they're a mess, clean them and relubricate at your earliest opportunity. Don't forget to pick up your safety triangle.

Next Morning

Take your flat to a competent tire technician for repair ("Repairing A Flat Tire," page 40). And don't let him overtighten your lugs with his fancy-dan air wrench.

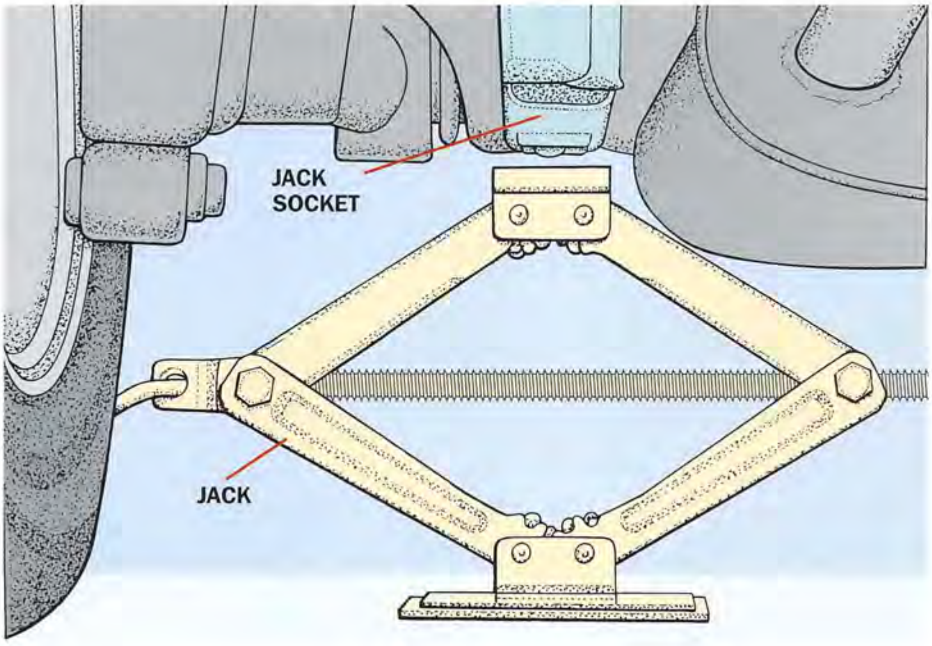

(Fig. 7) The correct jack point on the frame of your vehicle may be difficult to reach.

Replacing Your Tailpipe

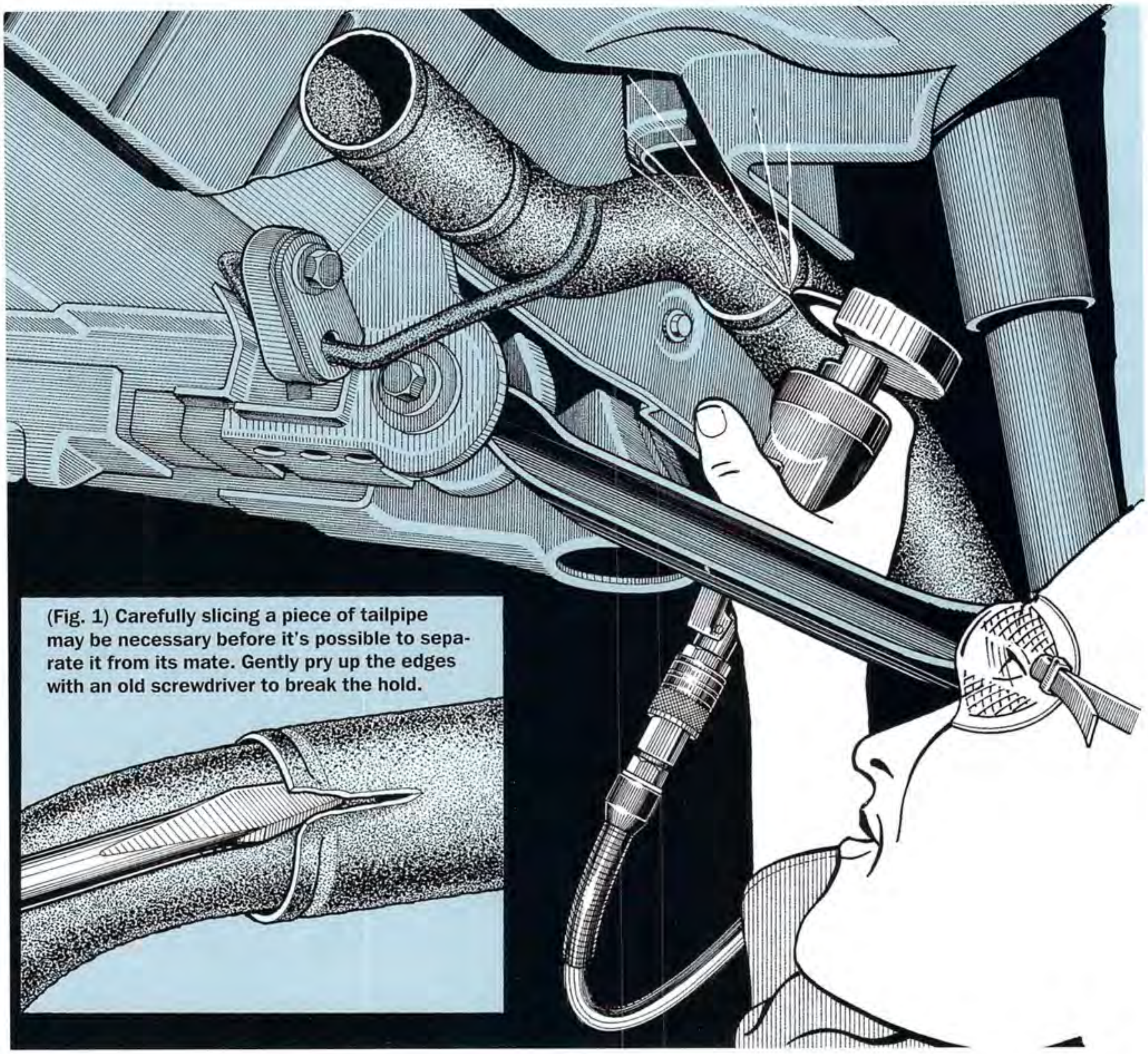

(Fig. 1) Carefully slicing a piece of tailpipe may be necessary before it's possible to separate it from its mate. Gently pry up the edges with an old screwdriver to break the hold.

Brakelights flicker suddenly in front of you, followed rapidly by the screeching of tires. Cars before you swerve to the left and right as you slow as safely and rapidly as practical. Despite your best efforts, the source of the pandemonium, what appears to be half of a cement block, disappears directly under your front bumper at a high speed. Thump. And as it exits from your rear bumper, you follow it in your rearview mirror for a few seconds. It has company. Your muffler—or at least most of your muffler—has apparently developed a lasting relationship with the aforementioned cement block. They're eloping at high speed toward the curb, inextricably intertwined.

The next thing you notice is the sound of what seems to be a B-29 landing in the next lane as you press back down on the gas. Actually, it's your engine—minus its muffler—waking up again as you try to negotiate through the traffic, back home to safety, family and something tall and cold that will make your hands stop shaking.

To add insult to injury, you replaced your exhaust system from the catalytic converter back only last spring. Inspection of the underside of your car reveals good, solid pipe most of the length of the vehicle, terminating in shards of muffler. Here's your chance to replace a few feet of pipe and the missing muffler yourself.

Now That You're Home

Actually, it may not be road debris that removes your muffler and tailpipe. Rust never sleeps, and cars in short-trip service may rust out the farthest, coolest portions of the exhaust system in as little as 18 months.

The following scenario is typical of vehicles that have had part of their exhaust system replaced at least once. The saga begins with a section of the original exhaust plumbing behind the catalytic converter rusting out and making a ruckus. Because the original exhaust pipes are welded together, however, all the pipes from the cat back need to be replaced. There's no practical way to remove just the rotted section because of all the welds.

If you had the wherewithal, you could probably cut out just the rusted portion and have a pipe custom-fitted. But that would be extremely labor-intensive and not particularly economical over the long haul. Also, exhaust work just isn't done that way. So, generally, two or three aftermarket pipes and the muffler are clamped together behind the converter to repair the exhaust.

Then you drive around for another year or two in peace and quiet—unaware that the replacement pipes hanging under you tend to rust at different rates. Sometimes, the middle pipe of the three you installed goes first. Other times, the muffler at the end falls off before the other two. Still other times, the flanged pipe bolted to the cat is the first to go bad. In all cases, though, you want to replace only the bad pipe. Problem is, the pipes it's connected to at both ends have been crimped by clamps. How do you remove the Swiss-cheese pipe without ruining the other two? By using a power cutting wheel and a very steady hand (Fig. 1). The preferred tool is an air-powered die grinder with a cutoff wheel, but a high-speed electric drill will work, albeit more slowly.

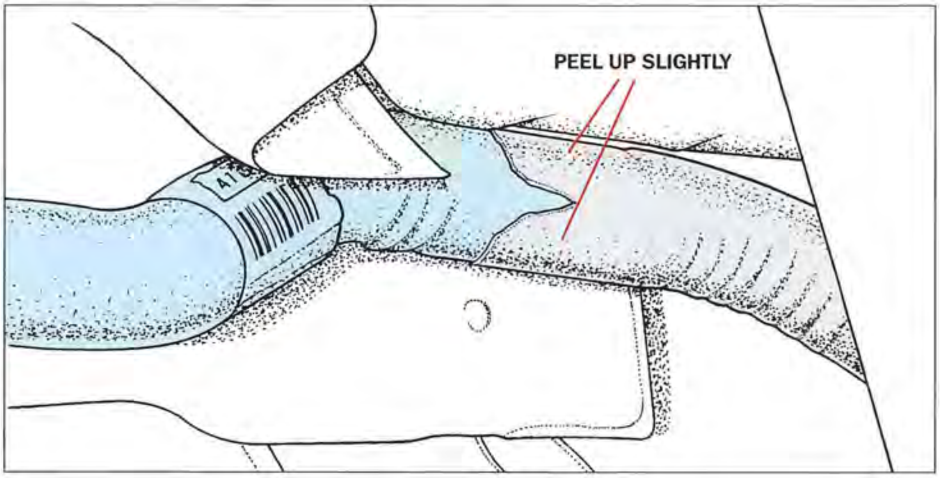

(Fig. 2) Peel up the corners of the old pipe to slide in the new piece.

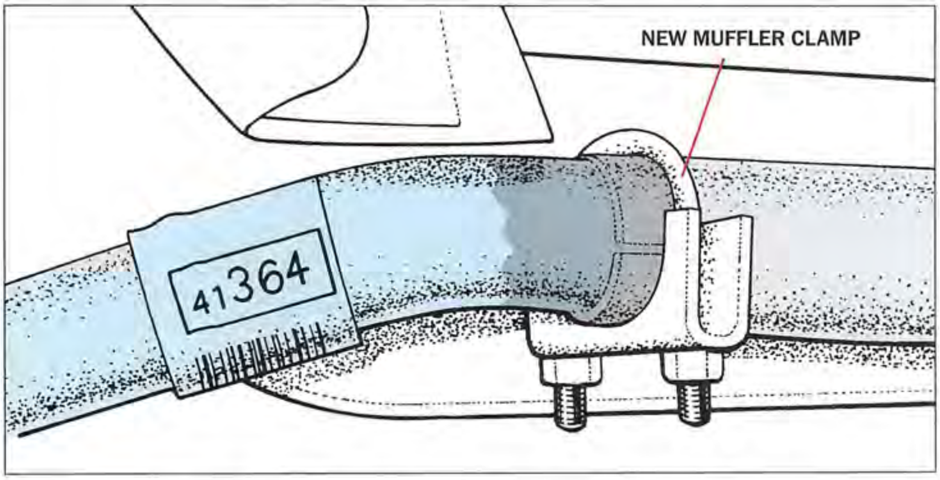

(Fig. 3) After installing all the parts, adjust for fit and tighten the clamps.

Sizing It Up

Get the back of the car up on jackstands, with chocks securing the front wheels. Be sure to wear safety goggles, as exhaust systems tend to drop razor-sharp flakes of rust that are seemingly attracted to your eyes.

Get under the vehicle to see what kind of couplings you're dealing with. The flanged type is easy. Just spray the rusty bolts with penetrant oil, so you can get them out, pry the flange from its rusty partner and secure the new flange with a new gasket and fresh bolts. Use antiseize compound on everything to ease removal the next time.

The other type of coupling is made up of a larger pipe slipped over a smaller one or a smaller pipe slipped into a larger one, depending on which way the coupling is facing. Both may be present in the same exhaust system. The important thing to remember is that you may need to save the inner pipe, so use a gentle touch with the cutting wheel. If the outer pipe needs to be saved, there's nothing you can do to keep from cutting it. But that's why you're going to use a cutting wheel instead of an impact chisel. It makes for a neater, smoother cut.

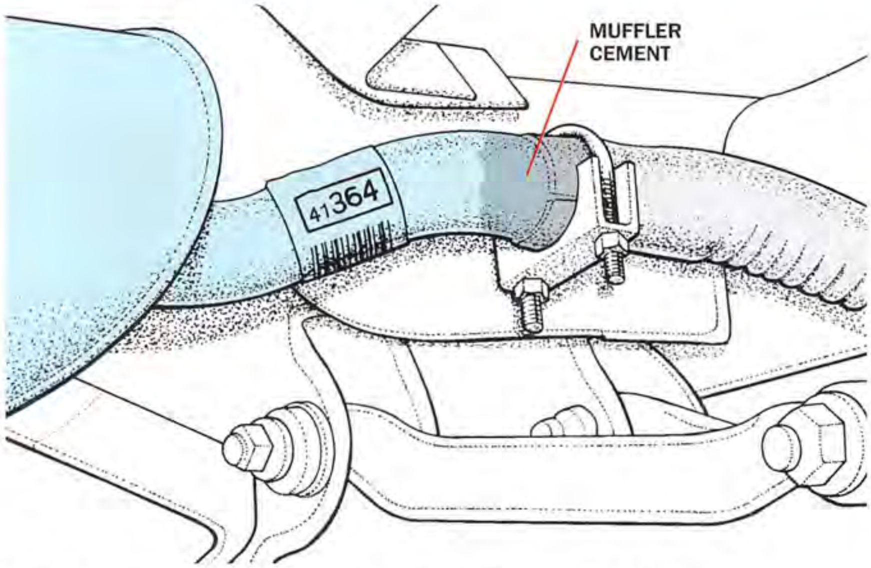

(Fig. 4) Smear muffler cement over the joint after you install the clamp.

Slice, Don't Dice

To begin, remove the bad pipe's clamps. Then, make a lengthwise mark 1 to 2 in. long where you want to slice open the outer pipe. Put on a pair of goggles, get a good double-handed grip on the cutter tool and start it up. Slowly raise the spinning wheel to the mark on the pipe. With gentle pressure, score the mark.

If you're saving the inner pipe, repeat this step until you're through the outer pipe. It's okay for the inner pipe to get a little score groove in it—just as long as you don't cut it all the way through. If you're saving the outer pipe, the second cut can take you all the way through the inner pipe.

With the cut made, use a screwdriver or a chisel to pry open the outer pipe like a tin can. That will give you enough play at the crimped area to wiggle the pipes apart.

Spread 'em

Fitting a new pipe over an old one is no problem. Fitting a new pipe into an old one sometimes is, because of the crimp from the old clamp. If you're having trouble, use an exhaust-pipe expander to gain some extra clearance. Then, simply fit the pipes together as if they were both new after smearing a little muffler cement on the joint.

If you saved the outer pipe, it now has a 1- to 2-in. slice at the end. Make sure to slide the inner pipe in past the end of the cut, so there's no hole between the pipes (Fig. 2). Install a new muffler clamp, but don't tighten it until the rest of the system is hung from the hangers and aligned properly (Fig. 3).

Once everything is back together and clamped down tight, a little muffler cement strategically applied with your thumb over the cut will take care of any tiny residual gaps (Fig. 4). Lower your car and enjoy the silence.

MODERATE

Repairing Power Antennas

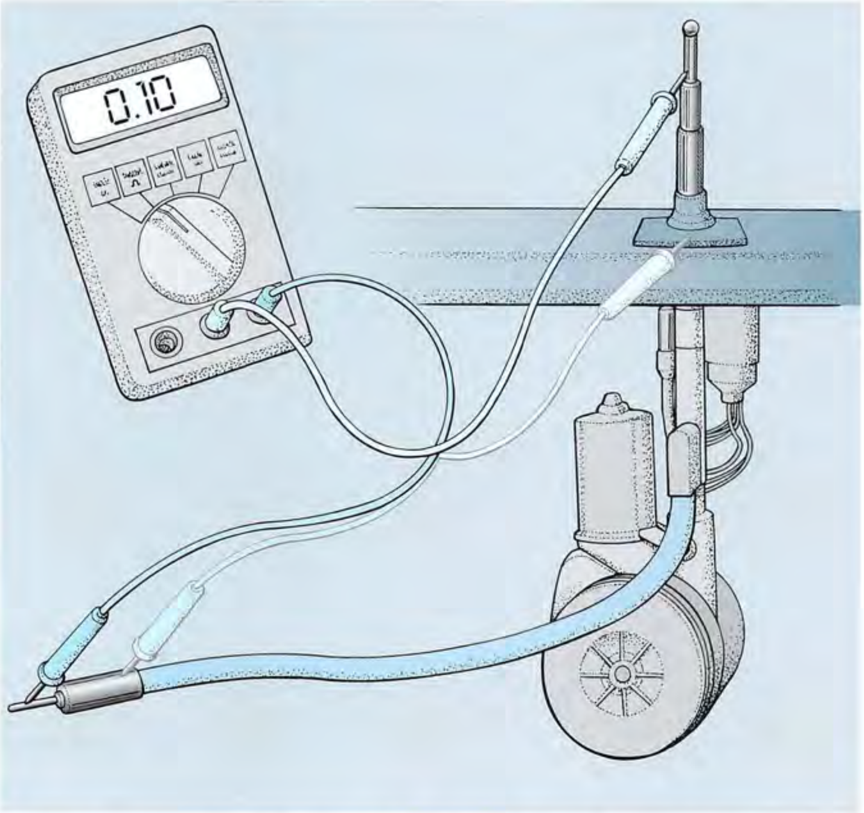

(Fig. 1) Scratchy radio reception might be caused by poor ground or antenna connections. Use an ohmmeter to chase high resistance points.

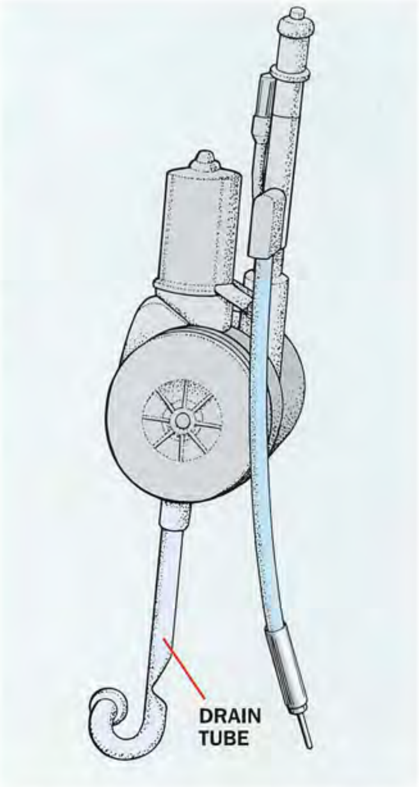

(Fig. 2) If water that leaks down the mast can't escape, the mechanism may freeze up.

It's time to kick out some jams, so you reach into the console for your favorite MC5 CD. Popping open the jewel box, you reach for the CD, only to find out that Junior has liberally lubricated it with peanut butter and jelly. After making a mental note to give Junior a good talking-to, you decide to surf the airwaves for some good oldtime rock-and-roll. Punching the FM button, you hope that at least one of the radio presets has some Bob Seger tunes hiding behind it. But no, there's nothing to be heard on any channel except a great rushing noise, not even so much as CONELRAD. Time to come back to this millennium and find out what's wrong with your antenna.

Manual Or Power?

Fixed antennas have only a couple of failure modes, and generally they involve mechanical damage to the antenna mast or the cable. If part of the mast is still there, you'll generally get some sort of signal. A poor connection between the antenna base and the fender also could be a problem. Unscrew the antenna mounting nut and check for corrosion. A cleanup with a wire brush and reassembly might re-establish a good ground.

Power antennas are more coy. If they're too shy to come out of the fender, you're listening to static.

No-Show

The issue could be either electrical or purely mechanical. Go back to your antenna and try to pull the mast out with your fingers. If it moves out readily, pull it out all the way, and then cycle the radio power with the key on. If the mast moves even the slightest amount, or you can hear the motor running at all, the problem is mechanical.

If the motor seems deceased, go back to basics—check the fuse. Determining which fuse protects the antenna motor may require some detective work. It may be the radio fuse, or it may be a separate fuse, perhaps shared with the rear-window defroster grid or a trunk-mounted CD changer. (Finding something else that doesn't work often is a good clue.)

Fuse is okay? Check the antenna's mounting, as described previously, for

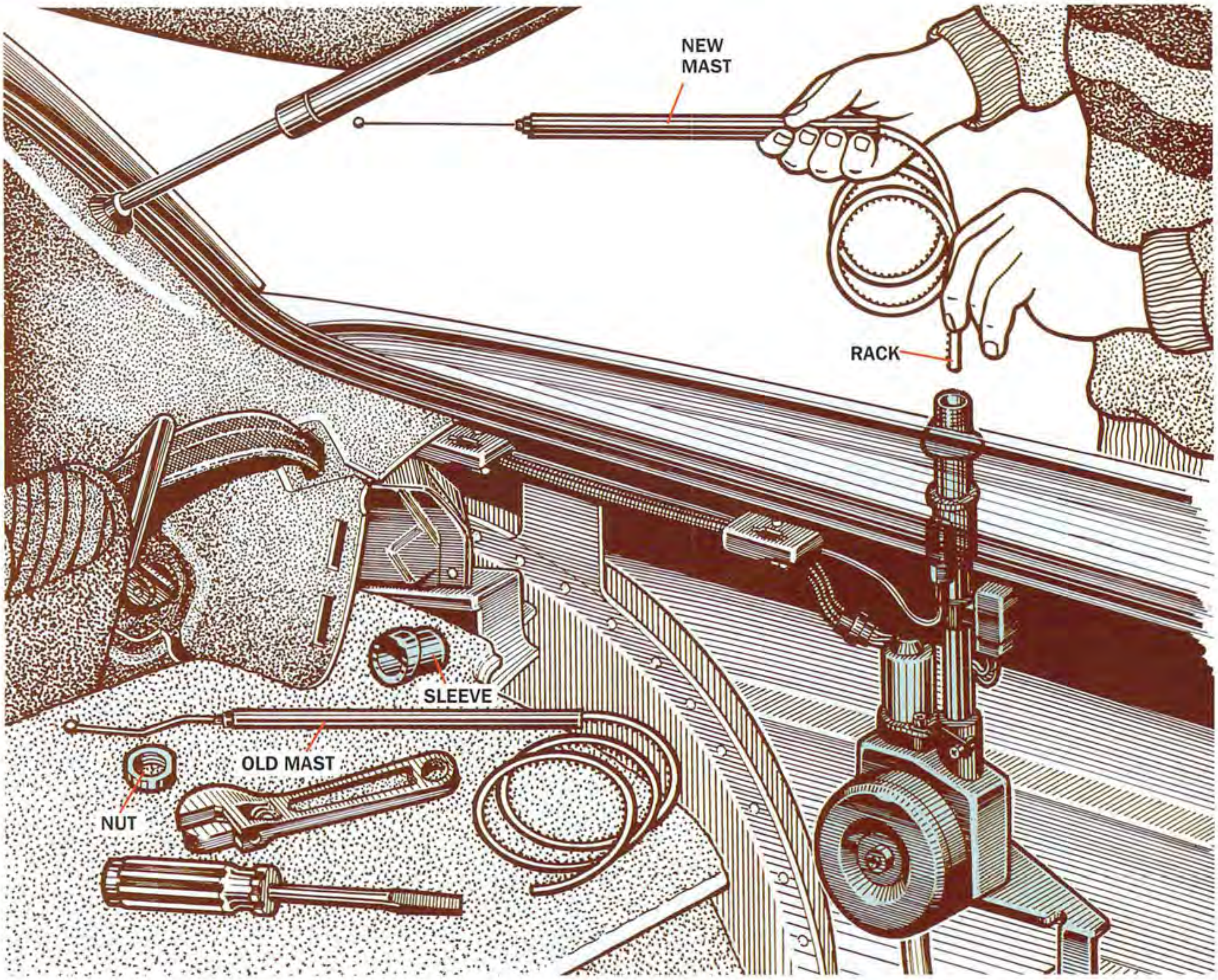

(Fig. 3) A sticky or damaged antenna mast can often be replaced without accessing or disassembling the mechanism. As we show here, this can be done without removing the interior trim.

a good ground pathway, although there may be a separate ground wire to a specific chassis ground point. You may want to use an ohmmeter (Fig. 1) to hunt for resistance. Incidentally, if you try to measure the resistance of the antenna cable's RF lead between the mast and the radio connector, it may check out as infinite. Some antenna assemblies use a capacitor in series with the RF lead, some don't. The ground shell of the antenna cable should have a low resistance to chassis ground, normally 5 ohms or less.

You'll need to find a schematic of the antenna wiring to troubleshoot any deeper. But with the radio and key on, the harness to the antenna assembly should have one hot wire on, and a different wire hot when the key or radio is off. Good hunting.

Stubborn

Does the antenna motor run briefly but not actually move the mast up and down? Or does it move a few inches and then grind to a halt? One problem we've seen is a kinked or pinched drain tube. (Fig. 2) The mechanism can fill with water, which then freezes solid in cold weather, or corrodes the works and causes a jam. Check the tube, especially if you can pull the mast out manually and it's wet. If the tube is compromised and the works are full of water, you may need to disassemble the thing, dry it out, and put it back together with fresh lubricant.

Can you help the mast along in and out? If so, the sliding sections of the mast may be corroded or bent. Careful bending by hand may restore movement, and polishing with 600-grit sandpaper or rubbing compound can smooth the action. At the very least, run the mast completely out and clean it thoroughly with a soft cloth, relubricating with a sparse amount of silicone grease or paste wax.

What's That Funny Noise?

Modern power antennas use a simple, flexible rack-and-pinion mechanism. It's pretty reliable, but abuse and the ravages of time and weather can literally strip the gears. So if you hear noises that sound like interrupted gear meshing, there's probably some section of the rack or pinion that's in need of dental work. It's possible to fix this without replacing the antenna assembly outright (Fig. 3). (You may not even have to access the mechanism.)

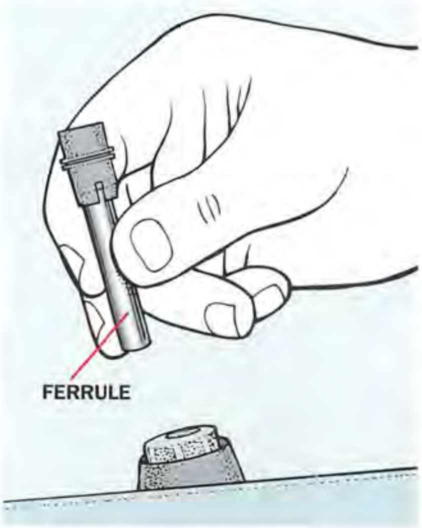

(Fig. 4) Remove the outer nut and collar as a first step to removing or replacing the mast.

Visit your dealership's parts department. GM, for one, offers a repair kit for power antennas. This consists of a new mast and rack assembly, ready to install. You may need to order it, as the two dealers we tried didn't have it in stock. There are aftermarket antenna parts to be had, but finding the right part is going to be difficult unless you have better luck than we did interrogating the staff behind the parts counter at the local warehouse distributor.

Here's how the system works: The limit switches that stop the motor at both ends of the antenna mast's travel are built into the mechanism, and you can consider them unserviceable. The limit switches rely on the antenna bottoming out or topping out to rotate a switch drum, at which point the current to the motor turns off.

At the top of the antenna mechanism is a large nut that holds the whole works onto the fender. Remove this nut. This isn't as easy as it sounds if it doesn't have flats that you can turn with a wrench. You may need to use a spanner with pins—although a pair of snap ring pliers may work if the nut isn't wickedly tight. At the potential expense of the chrome finish, you may need to fall back on the mechanic's friend, locking pliers. Under the nut and perhaps a plastic spacer or two there will be a ferrule (Fig. 4), which is the stepped sleeve that the mast actually bumps into at the end of its travel upward. Pull out this ferrule. Now have someone else turn on the radio. The mast should elevate itself completely out and flop over, so you have to be there to catch it. Carefully notice which way the teeth of the rack point as the mast clears the fender. Note the state of the teeth. Missing teeth probably mean you should disassemble the housing and clear all the bits out, but teeth that are simply worn should be okay.

The new mast's rack has been curled up like a pig's tail in the package, and it will be difficult getting it to mesh with the gears. With your fingers or pliers, bend the bottom 2 or 3 in. backward to remove the curl. The end should be straight now. After making sure that the radio is turned off, take a look at the last tooth—you may need to remove a small amount of casting flash. A pocketknife does that job well.

Take the new mast, with the teeth of its rack appropriately oriented, and insert it into the hole until it bottoms. Rotate the mast a little to the left and right to get the teeth to mesh with the pinion gear (Fig. 5). Have your helper turn the radio on and off while you push the new mast firmly down. This will cause the limit switch to cycle inside the mechanism. Now have your helper turn the radio on, and if all is well, the mast will suck itself very neatly in until it bottoms out. All you need to do now is reinstall the ferrule, spacers and nut. Lightly lube the mast sections with silicone grease or paste wax.

Outer Limits

If the limit switch mechanism stops working, you may be able to disassemble it, clean up any foreign matter or corrosion, and get it running again. The switch consists of a plastic drum with wiper fingers and electrical traces, so it's vulnerable to moisture and dried-out lubricant. If you can't fix it, or the motor itself is toasted, you'll need to replace the whole shootin' match. An aftermarket antenna will set you back about $60 to $75 at the local auto parts store, or more than a hundred at the dealer—maybe more for a luxury import. The OEM parts will, of course, drop in. Aftermarket pieces may require a certain amount of adaptation to mount properly, and perhaps even some creative wiring to make them work properly. Don't forget to check for used parts at the local scrapyard or auto recycler.

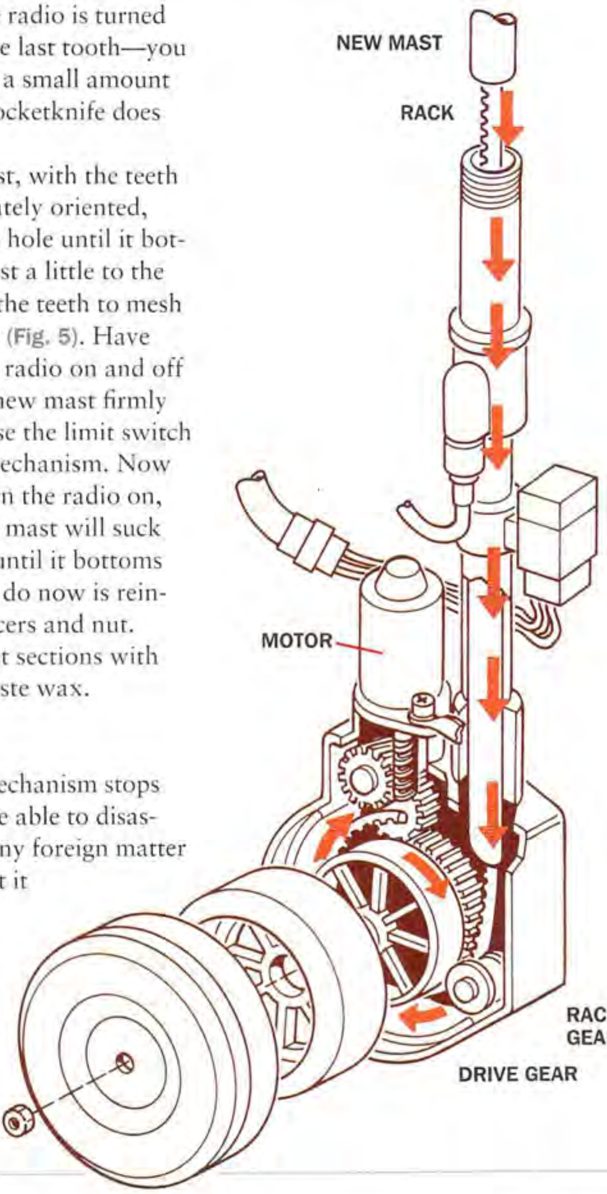

(Fig. 5) It may take a little fumbling to get the rack to mesh with the drive pinion, but when it does the mast will run completely home by itself.

Replacing Your Spark Plugs

Despite the heavy traffic—periods of slow-go interspersed with longer periods of no-go—things are okay. You've got that new Yanni disc in the player and your java is still piping hot. However, as your drive wears on, the CD starts skipping and the coffee in your cup starts spilling over the top. Suddenly you realize that your engine is no longer idling with its accustomed silky smoothness. In fact, it's getting as lumpy as cold oatmeal. The engine's developed a bad misfire.

As you sit and jiggle along in the rough-running car, you figure that it can't be the spark plugs. After all, with today's unleaded gasoline they can last twice as long as they used to. And the engine's high-output electronic ignition system produces enough voltage to fire even a worn plug.

Well, think again.

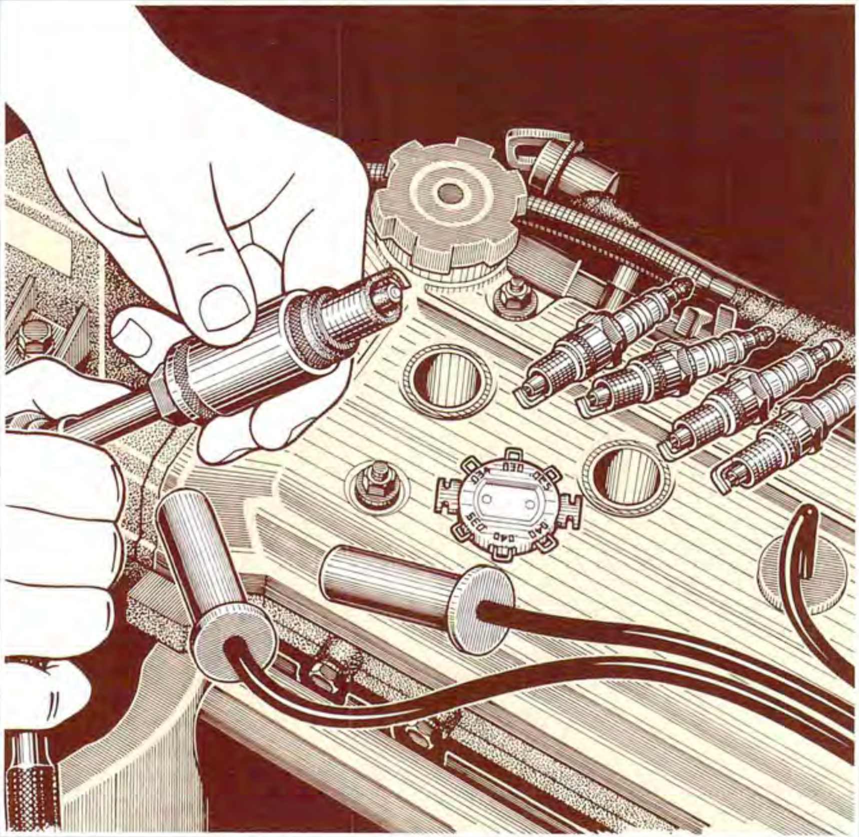

CHANGING PLUGS IS EASY on this overhead-cam engine, but might call for just the right combination of swivels and sockets for your ratchet on V-engines. Always check a new plug's gap with a wire-style gap gauge before installing.

Fair is foul

Although plugs won't lead-foul with today's unleaded gasoline, they can oil-foul, carbon-foul or even be fouled by some fuel additives. So if you've got a misfire, there's a fair chance a plug is the cause. In fact, it's more likely a plug than the fuel injector. So, back home, it's time to take the plugs out. Look for oily-black or sooty carbon-black deposits. The former usually are from defective intake valve stem seals, the latter from an overly rich fuel mixture, or a lazy or defective oxygen (O_2) sensor. The oxygen sensor is what signals the computer to adjust the fuel mixture.

It can be tough to hear a misfire, which is one reason why emissions-control standards require the engine computer's on-board diagnostics (OBD) to log trouble codes for each of the cylinders. However, the first cars equipped with that setup (called OBD II) weren't produced until 1994, and most cars didn't have it until 1996.

Even with an earlier model, you should check for trouble codes indicating a problem with the engine, particularly those involving the oxygen sensor. But if engine performance is down and you're not seeing trouble codes, and the plugs have gone a couple of years or about 30,000 miles, it's worth pulling them for a look. But not so fast—this may not be as easy as in your last car.

First, it may be harder to disconnect the plug wires. If they're recessed, you can't just pull on the wire. Worse, most of the simple spark plug pullers we've tried won't work on a stuck boot. You need pliers that grip the end of the boot and give you something solid to hold. These pullers cost $20 or so. In any case, twisting the boot to break the heat seal often is necessary, even with pliers. Never try to pull on the plug wire itself, or you'll probably cause a separation in the resistance wire.

A second thing to watch for is whether your spark-plug socket really fits. To meet the latest standard of the Society of Automotive Engineers, many new plugs are about ⅛-in. longer than you're probably used to. An older, marginal socket may not fit the hex properly and could slip off and break the plug. It also helps to use a socket with a rubber insert to securely hold the plug.

Finally, the elbow grease

Start with a warm engine. Clean the plug hole recesses with compressed air, crack all the plugs loose one-quarter turn, and let the engine cool. Removing plugs from a hot cylinder head can damage the threads.

Once you have the plugs out, check them carefully before you decide they've got a lot of life left. Use a magnifying glass and feeler gauge to be sure. If the gap is worn .005 in. over the specs, and you can see rounding off of the center electrode and wear on the side electrode, install new plugs. Forget trying to file the plugs.

If you're reinstalling the old plugs, smear a film of antiseize on the threads near the tip. New plugs have a coating that will lubricate the threads on the way in and prevent the plug from seizing on their next removal, but it's only good for one stab.

If any spark plugs are dry-soot black, that's from carbon—a rich-fuel mixture problem. If there's a trouble code indicating the oxygen sensor is incorrectly signaling lean, the sensor could be responsible. If it's correctly indicating a rich mixture, then you have to look for a cause, such as leaking fuel injectors or engine misfire, perhaps caused by defective plug wires. It would seem that misfire would throw a lot of raw gas into the exhaust, which it does. But it also throws in a lot of oxygen, and that's what the sensor sees, and in this case it just tells the computer to add more fuel.

Not sure or no code? Then check the oxygen sensor, which resembles a spark plug threaded into the exhaust manifold. If there's an O_2 trouble code and no apparent other cause, you could just replace the sensor. Or, before you invest in a new one, test the sensor you've got. You can do this using one of the inexpensive testers made for Saturday mechanics. With a warmed-up engine (necessary to get the O_2 sensor working) and the throttle cracked open (so the engine is running at about 2000 rpm), the tester's indicator lights typically will flash on and off (or will both glow steadily) if the O_2 sensor is working properly.

If you don't get the lights, or as an extra check, let the engine cool. Remove the sensor with a tight-fitting wrench and bench-test it. The typical sensor has a closed but slotted end and is a heat battery of sorts. If you envelop the sensor end with the flame of a propane torch (so it gets hot but not cherry red), it should produce over 0.4v with a high-impedance digital voltmeter hooked up, or turn on the tester lights. If the sensor is good, the voltage should drop or the lights go out within 3 seconds after you pull away the flame.

HOW IT WORKS

The Right Plug

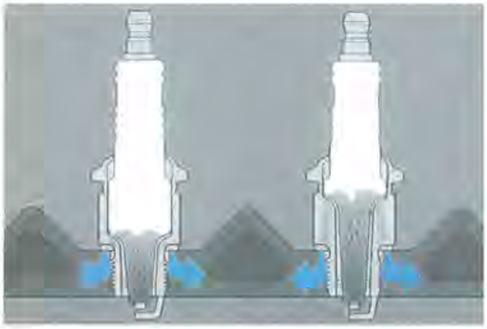

The correct plug for your engine is, of course, the exact part number that the manufacturer installed at the factory. When it's time to retire them, you'll be forced to choose from several manufacturers' offerings, and they're all pretty much alike, right? Wrong. There are subtle differences in a plug's ability to dissipate heat that belie what would otherwise be an exact physical match. A "hot" plug has a long path through the ceramic to the (relatively) cool cylinder head. A "cold" plug has a shorter heat-conduction path and tends to run cooler.

Plug manufacturers cross-reference their product lines to other manufacturers' lines, and you should be able to find a plug to fit almost anything in the cross-reference chart. More important than heat range are things like thread diameter and length, and the type of tip.

Installing the wrong plug can ruin your engine instantly if the plug is too long and whacks the top of the piston, especially in today's aluminum heads.

Too short a plug not only won't fire, carbon will quickly fill the threads on the cylinder's head, preventing you from installing the correct plug.

Occasionally a vehicle has a different type of sensor, and in some designs it has an open tip. This is a varying-resistance device that doesn't respond to these voltage-generating tests, so you'll have to rely on the on-board diagnostic system for openers, and a shop manual checkout procedure to find out if it's bad. This type (widely used on new cars) also was used on 1987-90 Jeeps, an occasional Toyota in the late 1980s and a lot of Nissans. These days you should have access to factory service information (or a good aftermarket equivalent) for your car, so refer to it.

Whether you install a new sensor or refit the one you've got, coat the threads of the O_2 sensor with antiseize compound.

Replacing Spark Plug Wires

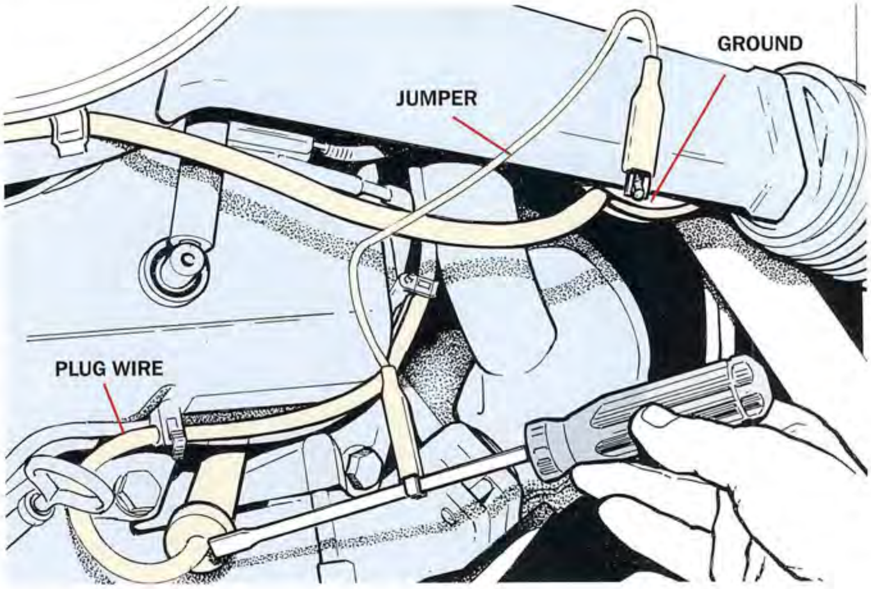

(Fig. 1) Dampen the wires with a spray bottle and clear water, and look for arcs with a grounded, insulated screwdriver.

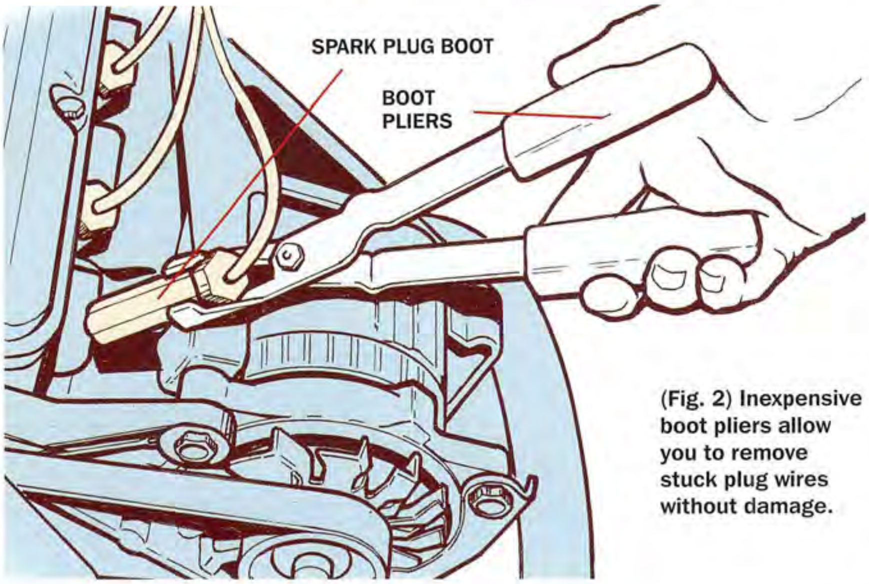

(Fig. 2) Inexpensive boot pliers allow you to remove stuck plug wires without damage.

Before easing out of the fast-food's parking lot, you grab a sip of the hot coffee you picked up at the drive-thru. You take your time as you place the cup in its holder in the center console. No hurry. The approaching traffic may be moving fast, but it's far enough away for you to accelerate up to cruising speed. Good thing, because at about half-throttle you feel the engine misfire. The closing traffic has time to back off and let you sputter up to road speed. But now your engine sputters like a Harley idling at the curb, even at normal traffic speeds. And there's coffee all over the console from the vibration.

What's Up?

A brief underhood inspection confirms that all the vacuum hoses are on and that none of them have been burned open by leaning against the exhaust manifold. You've touched every sensor connection to see if it's on tight, you've even checked for trouble codes.

However, not everything calls for high-tech diagnostics and the latest scan tester. Engine misfire often is caused by leaking spark plug wires, so visual inspection could reveal what's wrong. The jackets may be damaged from the outside by abrasion caused by engine vibration (particularly at the plug-end boots), a hot engine compartment, spilled fluids or battery acid, multiple disconnections and reconnections for other service, or even a family of nesting rodents.

Run the engine in near darkness, and then look and listen. If you see little electrical arcs, or hear a snap-crack, there is high-voltage electrical leakage. Of course, running the engine at idle doesn't exercise the ignition system very hard, so you can add something to the test by spraying the plug wire ends with clean water from a household spray bottle.

Another test: With a jumper wire, ground the metal shank of a screwdriver that has a well-insulated handle. Then, run the tip of the screwdriver along the length of each wire and all around at the coil and plug boots. This will often produce an arc from the wire to the screwdriver (Fig. 1).

Now look at the wires under good lighting. If they are damaged by abrasion, oil-soaked, cut, burned from contact with the exhaust, or have a dried-out look with heat cracks, it definitely shouldn't be a surprise if you see arcs in the dark under some conditions. However, if they look bad but there are no arcs, do a resistance check.

A conventional plug wire has a resistance of 10,000 to 15,000 ohms per foot of length—if it's measurably higher, the wire probably is bad. An absolutely failed wire will have a hairline break somewhere, and the resistance will be infinity.

Out With The Old, In With The New

Once you've found a bad wire, the solution should be simple: Replace the plug wires.

First, you have to decide whether you want to buy from a car dealer at list price or get a quality aftermarket brand, if you have that choice. Prices for 1980s models may be under $35 for a popular make at the parts store. For a late-model V8, be prepared to pay the dealer up to $200 or more, even for a popular car.

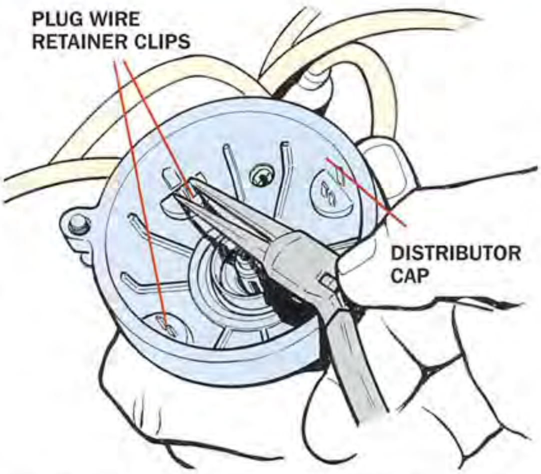

(Fig. 3) Some wires are secured inside the cap by spring clips. Use needle-nose pliers here.

A word of caution: You may not want "high-performance" wires, even if they fit your car (and they may be cheaper than a conventional type). Some high-performance wires are not the resistance type, and although they may do well on racing cars, they can affect the operation of underhood electronics by not containing radio waves.

The actual method of removing the old wire is not always "just pull it off." Pulling on the wire itself is sure to cause an internal separation (see p. 19). Of course, if you're replacing a bad wire, it may not matter, but if any one of the old wires is still good, it's best to keep it as a spare.

Grasp the plug wire by the boot at the plug end. If it's in a recess or is difficult to access for some other reason, use a spark plug wire tool. There are special pliers of all sorts designed to reach in and grasp the boot (Fig. 2). They're not expensive, and having them could prove useful if you ever have to disconnect wires for other underhood service. Twist the boot if necessary to break the heat seal to the plug, then pull. As you do this

HOW IT WORKS

Spark Plug Wires

The spark plug wire has a seemingly simple job: Carry the high-voltage electricity produced by the ignition coil to the terminal of the spark plug. Once at the plug, the electricity travels to the other end of the plug, and jumps a gap between electrodes to produce the "spark" that ignites the fuel mixture. Because high-voltage electricity looks for an easier path to jump across than a pair of electrodes in a spark plug, containing the electricity within the wire takes a thick jacket of insulation, and some wires have an outer jacket to resist cuts, high underhood temperatures, etc. If that outer jacket is damaged, the electricity may leak out to follow the easier path. In addition, high-voltage electricity produces radio waves, which can cause interference with all types of on-car electronic devices, from sensors and computers to radios and other entertainment systems. So, a simple length of solid wire can create problems.

The electrical conductor typically has a carbon-impregnated core of a suitable fiber, and everything from nylon to Kevlar has been used by different makers. That conductor (often covered by a second layer of a nonconductive material such as synthetic rubber) has enough electrical resistance to suppress radio interference without unduly weakening the spark. Some premium wires may have copper or stainless steel wire wound around the carbon core to reduce resistance. In some special applications in which minimum resistance is needed for engine performance, a solid metal wire has been used as a conductor, but over the inner layer of insulation is a layer of wound wire as a shield. On other types of wires, the outer jacket of the wire may have a metallic shield to prevent interference.

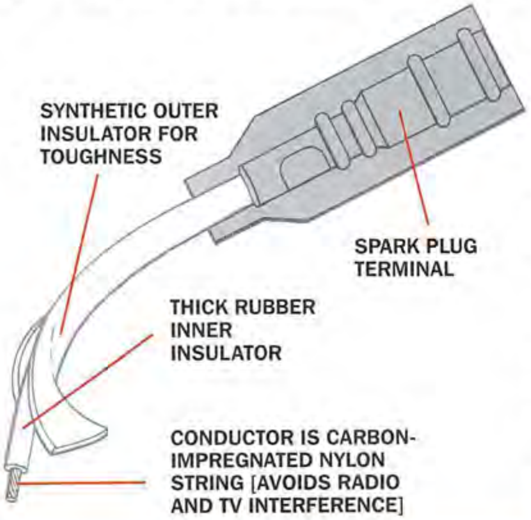
SYNTHETIC OUTER INSULATOR FOR TOUGHNESS
SPARK PLUG TERMINAL
THICK RUBBER INNER INSULATOR
CONDUCTOR IS CARBON-IMPREGNATED NYLON STRING [AVOIDS RADIO AND TV INTERFERENCE]

(Fig. 4) Replace plug wires one at a time to maintain the exact same routing as original. Use silicone grease on the boots to allow easy removal in the future.

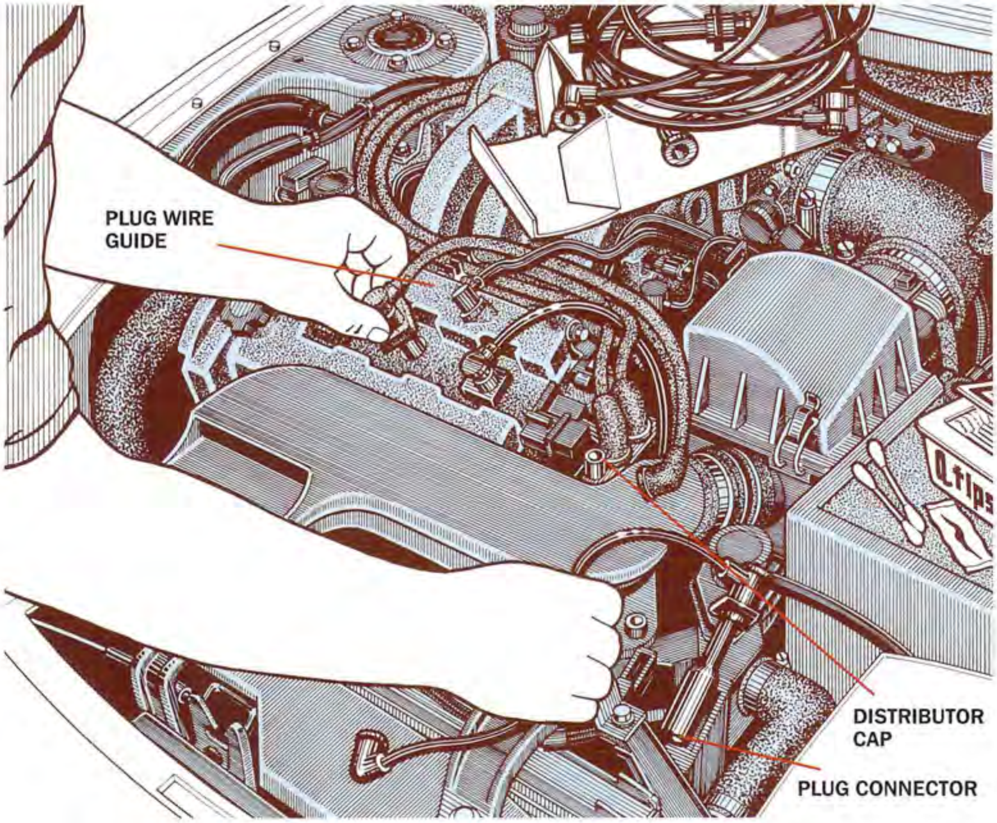

take note if you feel some looseness at the connection (it may have been caused by engine vibration).

The plug wire usually is more accessible at the coil end, but it may not be a simple push-on. Since the early 1980s, Chrysler plug wires on engines with distributors have been held in the cap by spring clips. Compress the tangs of the clip with pliers and push out to disengage the wire (Fig. 3). Some Japanese makes have thread-in "boots."

Follow the routing of a spark plug wire as if it were the road to success, because when it comes to engine operation, it really is. Every manufacturer includes little plastic guides, and although they may cost pennies to make and install, they're used to locate each wire so it doesn't cause crossfire (a transfer of high-voltage electricity from one cylinder's wire to the one that's next in the firing order) (Fig. 4). If you see one wire crossing over or under another at nearly right angles, that's an example of the routing strategy used to avoid crossfire.

Of course, on some engines with coil-on-plug ignition (COP), there is no wire and the boot is out of sight, out of mind. Although this type is immune to hungry rodents, a boot may suffer internal damage from high-voltage electricity or cracking from engine heat. It's even possible for spilled oil to flow under the cover and get down onto the boots on some cars, so have a rag handy when you're adding engine oil. It's impractical to check these boots with the engine running, so you've got to remove the ignition cover and lift the COP modules to make a physical inspection. There are some "hybrid" designs, with one coil on one plug to serve two cylinders (Fig. 5). The coil directly feeds a plug in a boot underneath, and there's a plug wire (replaceable) to a second plug.

The COP boots alone are usually under $25 a set, and are strictly a dealer part. Just pull the coil/electronics module off the plug and then pull the boot off the module. If the boot is integral to the electronics the price could be $100 each.

(Fig. 5) Many newer cars have individual coils mounted on the plugs. Short wires connect opposite-firing cylinders.

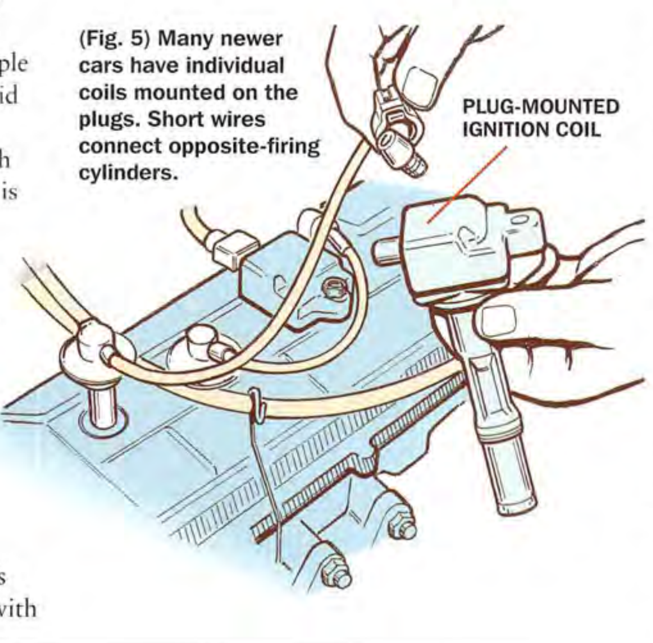

There's no maintenance you can perform to extend the life of plug wires. However, before you install the boot on the plug, coat the inside with boot release lube (a silicone grease) to make it easier to remove next time. And whenever you work under the hood, avoid nicking the wires, spilling solvent on them, or mishandling if you have to set them aside while you work on something else. And if you see a wire dangling very close to the exhaust manifold, reposition it in its guide to gain some clearance. A penny of prevention can be worth a couple of hundred dollars of cure.

Maintaining Your Tires

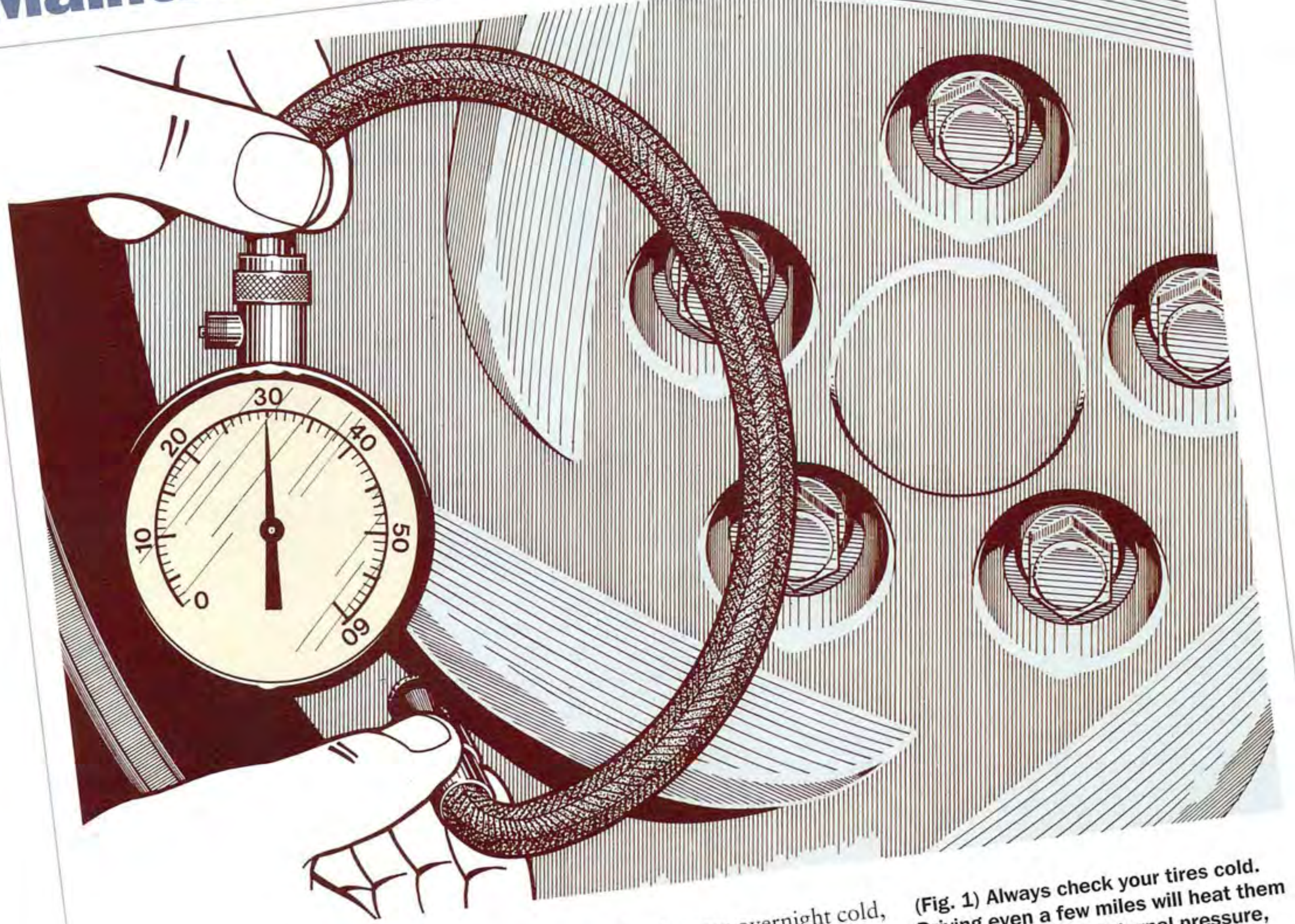

(Fig. 1) Always check your tires cold. Driving even a few miles will heat them up and change the internal pressure, possibly masking underinflation.

Any kind of tire failure—even a simple flat—is a huge hassle. But this can be avoided with careful tire maintenance. If the recent tire recalls and reports of tread separation have alerted you to the importance of regular tire inspection and care, you're ready to go beyond just a quick look to make sure the tires aren't flat. Most failures are caused by underinflation, overloading your vehicle, or damage from debris, curbs and potholes. These are things you can check yourself.

Preflight Inspection

At least once a month, check your tire pressures—in your driveway, preferably when the tires are overnight cold, or when it has been at least 6 hours since the vehicle was driven (Fig. 1). Don't try to "eyeball" a tire for normal inflation. You can't tell the difference between a properly inflated tire and one that's even 10 psi under. Some tires always look underinflated. Some have stiff sidewalls and always look normally inflated.

Read the tire specification label, which usually is on the driver's front doorjamb or the matching surface on the pillar (Fig. 2). These pressure specs usually are lower than the maximum pressure allowed on the tire's sidewall, but they're based on each particular vehicle and its rated load, not what the tires are physically capable of withstanding.

Although tire pressure specifications usually peak at 32 to 36 psi, buy a tire pressure gauge that reads to at least 60 psi. That's a typical pressure specification for a compact spare. If the vehicle manufacturer specifies a pressure range, such as a minimum and a maximum, always use the maximum. The higher the pressure, the greater the load-carrying capacity of the tire, the more stable the vehicle's

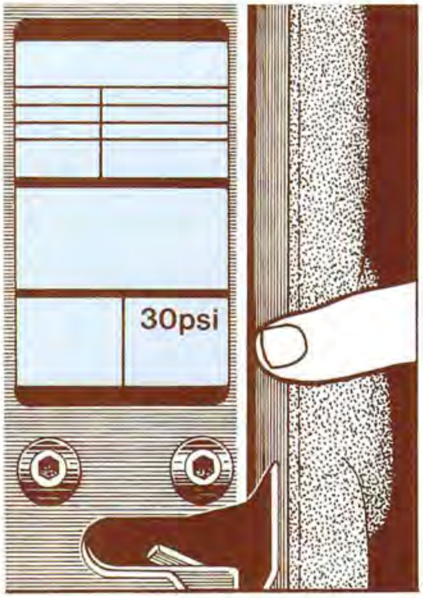

(Fig. 2) Appropriate tire pressures are listed on a sticker on the doorframe or the matching door pillar.

handling will be, and the cooler the tire will run at speed. All tires leak air pressure over time. The closer the tires are to the recommended maximum pressure, the greater the safety margin for all operating parameters. A normal tire leaks about .5 to 1 psi per month.

In addition, as ambient temperatures drop with the changes of seasons, so do the tires' air pressure—one psi per 10°F. It might seem that the temperature-related drop is matched by lower operating temperatures of the tires in cold weather. But the key to tire safety is adequate pressure to carry your vehicle's load. Although higher tire pressures stiffen the ride somewhat, it's a small price to pay for the extra safety and the ability of the vehicle to accommodate greater loads. Maintaining proper tire pressure also improves fuel economy, although not by much. Caution: Overinflation increases center tread wear.

Don't forget the spare. And note our earlier warning about the higher specified pressure for a compact spare. If your spare is carried in the underbody, as it is on many minivans, you'll need to crawl underneath with a pressure gauge and air hose.

Inspecting Tires

If you hit a road hazard such as a sharp rock, of course you're going to look as soon as possible for damage. But remember, even if the exterior looks fine, the tire may be damaged inside. Although it may not slow-leak, it could fail suddenly. You'd be wise to get the car to a tire shop so the tire can be demounted and given an internal inspection. The few dollars is a great investment in peace of mind, and if there is internal damage, forget about a patch. Invest in new rubber.

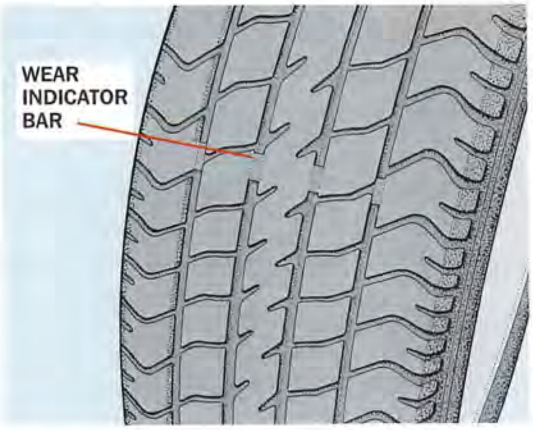

(Fig.3) Check the tread for cuts, foreign objects and abnormal wear. The wear bars will indicate when the tire is worn out.

Wheel alignment also plays an important role in tire performance. If the wheels are misaligned, they don't roll true down the road. The side slippage produces friction, which raises tire temperatures and not only increases tread wear but causes the wear to be uneven.

So inspect the tire treads. As you do, pry out pebbles from the grooves. They reduce traction and can damage the tread area.

The treads should be deepest at midpoint—at least $\frac{1}{8}$ in. thick. That's $\frac{1}{16}$ in. above the tread bars that are the official "replace them" indicators (Fig. 3). However, you need tread above those bars or the tire will do a poor job of shedding any water and slush it runs into on the road. The wear pattern should be relatively even at each side, although it might be somewhat greater in the middle. Tread wear that is "feathered"—worn to a sharp edge at one or both sides—or much greater on one side is a sign of misalignment.

If you see cuplike wear in the treads, typically along one side, the possible causes are wheels that are out of balance, worn-out shocks or struts, and loose suspension components. The classic sign of unbalanced wheels is high-speed (50 mph and up) vibration, and it usually surfaces before the cuplike wear becomes noticeable. Suspension problems usually produce shake at lower speeds.

Look for any cuts on the surface of the tire that expose the steel belt or fabric cord. This is grounds for immediate replacement.

Tires with unevenly worn treads should be replaced, unless the problem is caught early and there's plenty of tread depth left. In that case, they could provide a moderate amount of life on the rear wheels, particularly on a front-drive car. However, if you have an all-wheel drive that you push pretty hard, invest in an entire set of new tires.

Wheel Balancing

Most shops have off-car electronic balancers that allow the wheels to be rotated to any position without affecting their balance. The disadvantage is, if there's any unbalance in the wheel hub, it won't be corrected. If this doesn't cure the vibration, it's worth finding a shop with on-car equipment before you try something else. In addition, make sure the shop has weights designed for your wheels. It takes at least a half-dozen differently shaped weights to fit properly on the rims of all the popular wheels. There are several so-called "universal" weights, but they may not fit your rim, and could pop off or cause rim damage.

If the wheels are balanced and you have wheel shake at medium speeds on up, or the tires have cup wear, check the shocks and struts.

HOW IT WORKS

Reading A Passenger Tire Sidewall

A "P" or "LT" stands for passenger car or light truck tire. The 3-digit number (215 on this diagram) is the tire cross-section width in millimeters. The 2-digit number that follows (65) is the aspect ratio—the height of the sidewall relative to the cross-section. A lower number indicates a wide tread area relative to the sidewall—today's sporty look. The letter "R" means radial tire construction. The next 2-digit number (15) is the wheel rim diameter in inches. A 2-digit number following the size information is a load-versus-pressure rating. A letter that follows that 2-digit rating is a speed rating. Example: "S" is for a tire rated for speeds up to 112 mph. You'll also find maximum pressure (in psi and kilopascals), and maximum load (in kilograms and pounds). In addition, there are numerical or letter grades for relative tread life, traction and temperature (UTQG). The layout of the information may vary according to make.

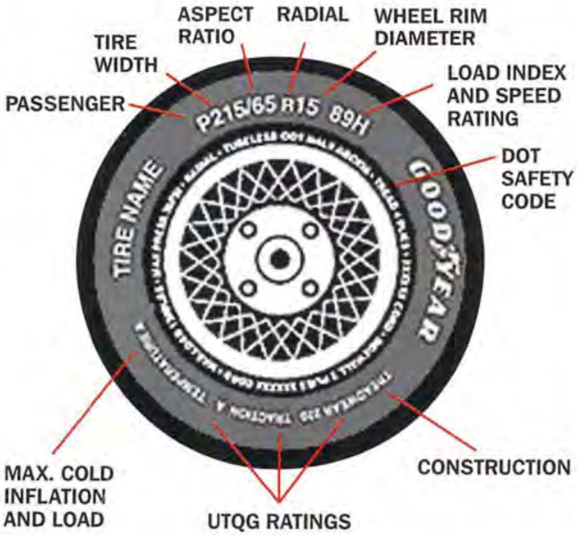

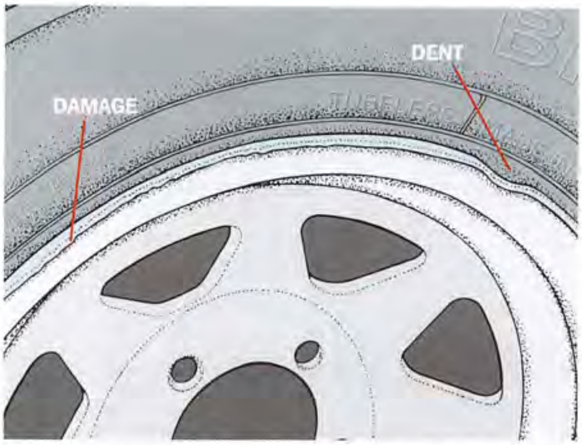

(Fig. 4) Minor scraping on the rim's lip is okay, but a bent bead surface may cause a leak or vibration.

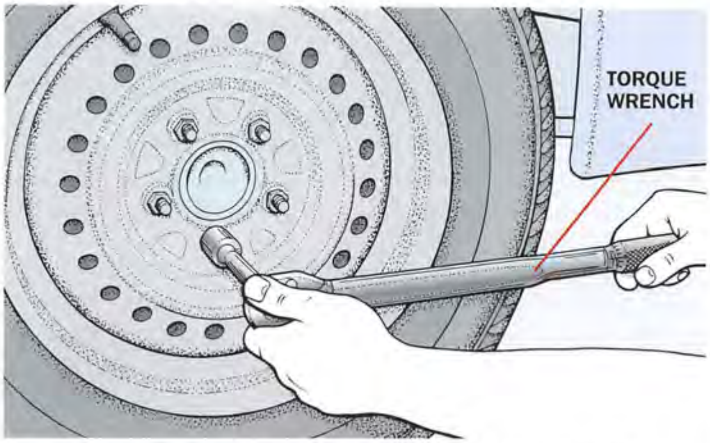

(Fig. 5) To avoid warping brake discs, tighten the lugs in a crisscross pattern with a torque wrench.

Sidewall Check

Using a tire and wheel cleaning product will give you a clear look at both the tires and wheels. You should be able to find cracks in the wheel, damage beyond surface nicks to the tire sidewalls, and damage to the bead area that could be responsible for pressure leakage (Fig. 4).

Tire Rotation

Every owner's manual has a tire rotation diagram, and tire rotation may seem like motherhood and apple pie—always good. However, this subject is not so simple. Cost of rotation versus longer tire life is not a precise equation. It depends on the tires, the driving, wheel alignment and suspension condition. Further, professional service is not always expertly done. Unless the lug nuts are tightened to specifications in three even stages, using a crisscross pattern (Fig. 5), the rotors may become warped, which adds to maintenance costs. If you let the mileage stretch out a bit, such as to 10,000 miles or more, tires may develop almost imperceptible wear patterns that will affect ride when they're moved to a new position on the car. If you can't rotate the tires often, you may be better off leaving them in place and accepting the somewhat shorter tread life.

Some tire treads are directional. They should rotate in only one direction and should not be rotated except by a professional who can demount them.

How can you tell if you have this type of tire? Look for a directional arrow on the sidewall.

Jumpstarting Your Car

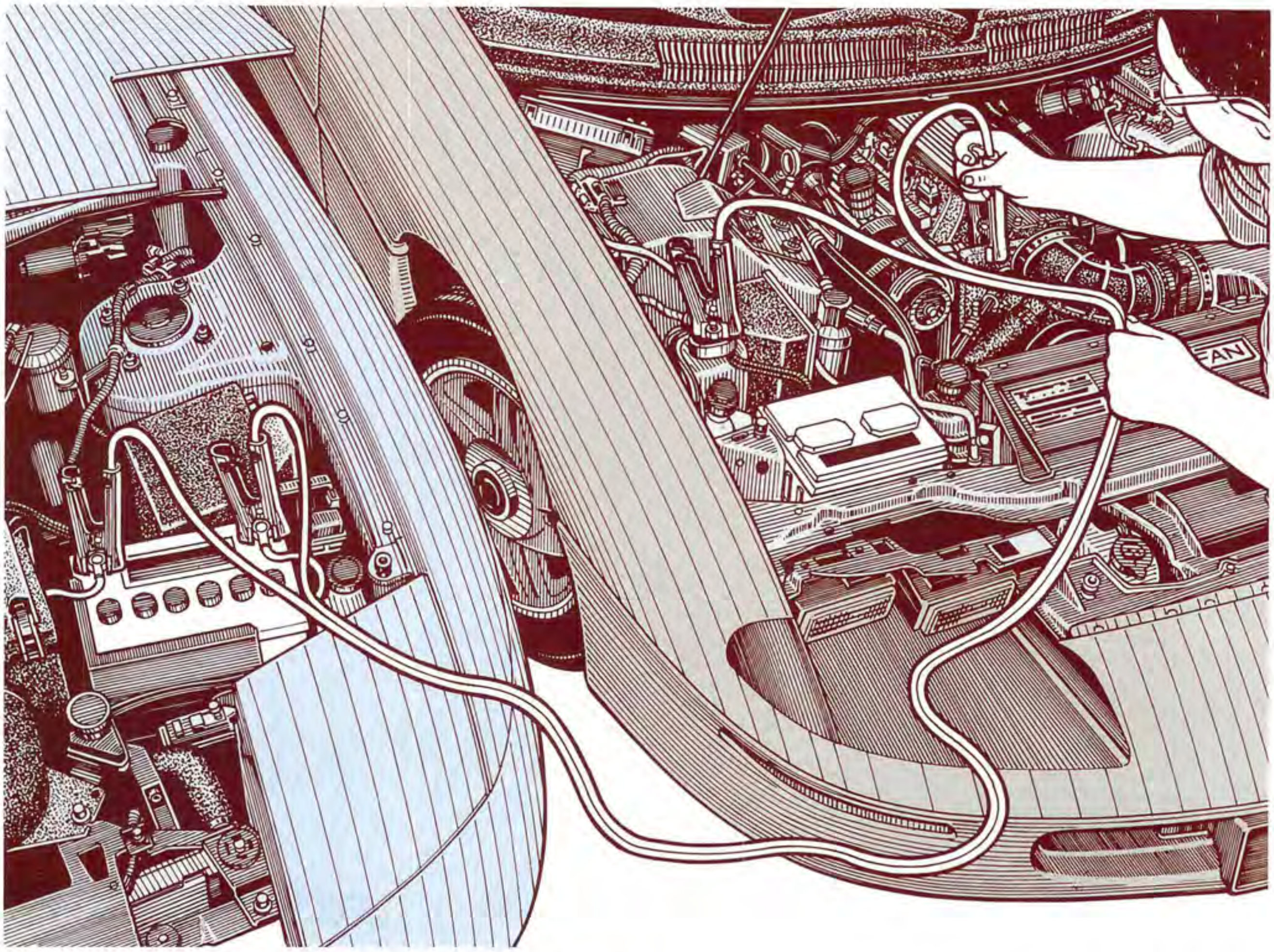

(Fig. 1) Make the last connection to a ground point not on the battery to prevent creating a spark near any venting, explosive hydrogen gas.

It hits you like an electrical shock just as the airliner door slams and the flight attendant oh-so-politely instructs you to turn off your cellphone. Less than an hour ago, you parked your car in the middle of a 10-acre airport parking lot, expecting the vehicle to take you home when you return in five days. Except you realize that you forgot to turn off the headlights after your predawn departure from home and the warning buzzer hasn't worked for eons, if your car has one. And if it doesn't, there's no surprise—you've done this before. Great. Now when you get home you'll have to call the airport-approved towing service and drop close to a hundred bucks for a jump.

Neglect Is Abuse

Lead-acid automotive batteries are actually remarkable at delivering extremely high current for the few seconds it takes to start your engine, even in extremes of heat and cold. What they are not good at is being able to recover after delivering more modest amounts of current for a long time. Irreversible chemical changes take place. Specifically, sulphate needles that bridge the separator between the positive and negative plates form when a battery is deeply discharged. These needles not only internally short the plates, causing a high self-discharge rate, they also coat the plates and interfere with the normal lead/sulfuric acid reaction that makes electricity to spin your starter motor and run your fuel injection and ignition long enough to coax your engine into life.

Before you do anything else, check to see that the dead battery isn't frozen. While a fully charged battery is almost freeze-proof, a highly discharged battery can freeze when temps hit the low 20s. If the battery has them, remove the filler caps and look for ice crystals in the electrolyte. Don't try to charge a frozen battery. It won't work, and will damage the battery further. If you need a boost start to get home, well . . . life's like that.

Safety First

Your first task is to get jumper cables. Preferably nice fat ones with a quality, heavy-duty set of clamps. This is more than just tool envy—there's a lot of current passing along those wires, more than an arc welder, at least for a few seconds. Resistive losses in the cable can reduce the voltage available to your stalled car's electrical system to the point at which it will still be difficult to start, even with a healthy donor battery and alternator adding their all to the mix. It's important to use fat cables and to have good electrical connections to reduce these losses to a minimum.

A set of 6-ft. cables won't do you much good if you're parked nose-in to a parking space—unless you feel up to pushing a 3-ton SUV back a truck length to make the engine compartment available. Twelve or 15 ft. is better, which makes using heavy-duty cables more important because resistance losses are proportional to the length of the cable.

Keep your cables clean and dry to prevent corrosion from becoming a high-resistance factor.

Doing The Deed

It's nitty-gritty time. You've got cables, you've got a healthy donor car available. Open the hoods and position the donor car nearby, so that your cables will reach.

Some cars, like the one illustrated on the previous page **(Fig. 1)**, have a remote positive terminal someplace in the engine compartment. The battery is mounted in an inaccessible area or with its terminals inaccessible even in the engine compartment. As for where to attach the jumper cable, this junction will be clearly marked and covered in a red plastic sheath. If in doubt, consult your owner's manual.

Wear eye protection, even if it's only a pair of sunglasses. Once in a blue moon, a battery will explode

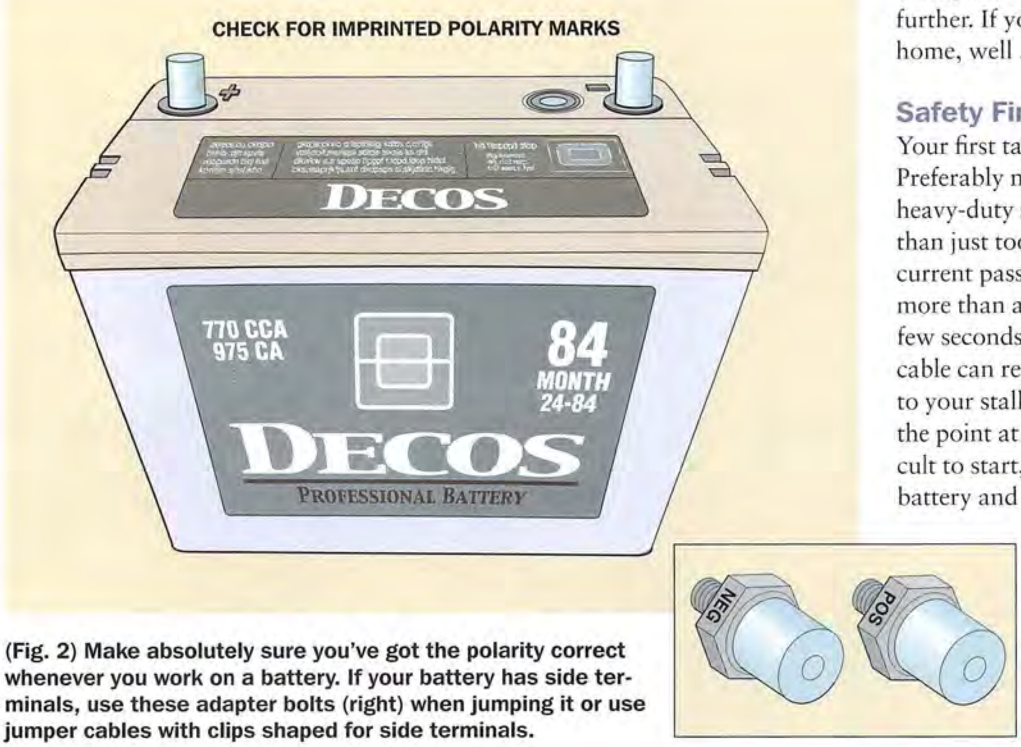

(Fig. 2) Make absolutely sure you've got the polarity correct whenever you work on a battery. If your battery has side terminals, use these adapter bolts (right) when jumping it or use jumper cables with clips shaped for side terminals.

HOW IT WORKS

Deeply Discharged Batteries

A lead-acid car battery, like any battery, has an internal resistance, normally a few tenths of an ohm. This means that a charging voltage only a little higher than the battery's normal 12.6 volts will provide enough current to actually charge the battery at some decent rate, say, 10 to 12 amps. But as the battery discharges, its internal resistance goes up. And the curve is steep. When the battery is almost totally discharged, internal resistance can be high enough to prevent the 13.5 to 14 volts your alternator puts out from doing any significant charging. Charging current in this instance can be as little as only a few hundred milliamps until the battery's resistance goes down—and that might take many hours, or even days of charging at normal voltages.

The answer is a high-voltage, high-rate professional charger that can supply as much as 25 to 30 volts for a brief period to give the battery a kick in the pants. This will generate a lot of heat, enough to cook a battery within a few minutes. Leave this business to a professional, and follow up with a normal 6- to 10-amp charge.

(Fig. 3) If your car or RV is in storage and chronically needs help starting, a trickle charger will keep it ready and improve your battery life as well.

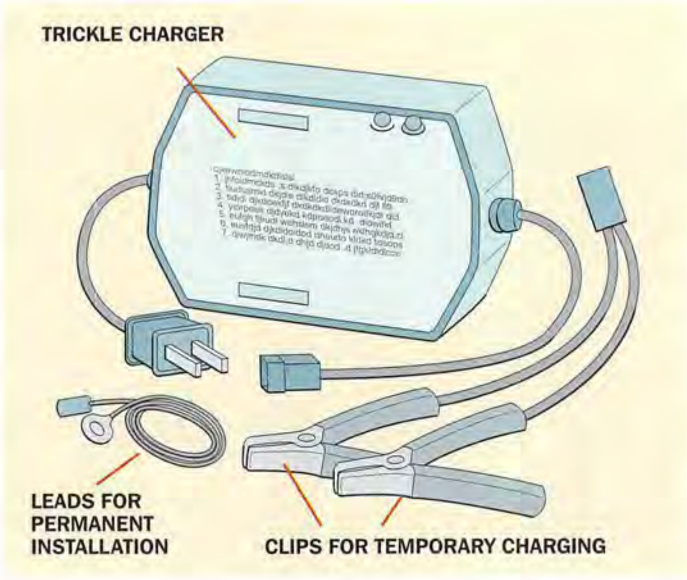
TRICKLE CHARGER
LEADS FOR PERMANENT INSTALLATION
CLIPS FOR TEMPORARY CHARGING

when you try to jump it. Explode? Yes, explode. It's caused by hydrogen gas, which is normally vented by a battery that's being charged or discharged at a high rate, say, when you're trying to start a car, the battery runs down and you need a jumpstart. Hydrogen is explosive, and a spark from making a connection can ignite it. It won't be a big explosion, but it can certainly blow the top of the plastic battery case off and spray acid into your eyes.

Connect the red clamp on one end of the jumpers to the positive terminal on the dead car. Verify the polarity of the terminals by the plus symbol molded into the battery case (Fig. 2). Don't just use the red terminal—someone may have installed an incorrect, red-colored terminal onto the negative pole of the battery. Do the same on the positive terminal of the donor car. Start the donor car and let it idle. Lights, heaters, stereos and rear-window defrosters—all electrical drains—should be off.

If possible, cover the dead battery with a shop towel or a sheet of cardboard. Any acid that manages to bubble out of the vents will wind up on the cloth instead of on your clothes or the paint on the fender.

Connect to the negative terminal of the donor car's battery with the black clamp. Verify the polarity. Now connect the remaining black clamp to the dead car's engine block, an accessory mounting bracket or a protruding ear on a manifold. Use the battery's negative terminal as a last resort. This procedure will generate any sparks far from any hydrogen gas venting from the battery and reduce the risk of explosion.

Now wait. This will let the dead battery recharge slightly. It will charge more when the dead car starts, but it will help the donor car's battery start your engine a little if you give it a quick shot of charge. If the dead car's battery had enough charge left to make the solenoid click and run the interior and instrument lights, then a minute or so is enough. If the battery was dead, dead, dead, give it 5 minutes or so.

Crank, Zoom

Now you can actually try to start your dead car. When the car starts, let it run at fast idle for another few minutes still connected to the donor car to continue charging by both alternators. Idling provides only a modest charge rate, so after a few minutes the best thing is to drive the car at normal speeds for 30 to 60 minutes.

Remove the jumper cables in the reverse order. Wash your hands if possible to prevent any battery acid from bleaching your clothes or getting into your eyes.

Options

If your RV or second car is in storage for extended periods of time, the battery will self-discharge. The simplest solution is to use a trickle, or maintenance, charger (Fig. 3). These chargers will automatically adjust their charge rate to a safe level, low enough to keep the battery at 100 percent charge without cooking it. They're not recommended for charging dead batteries—just maintaining them.

If you're not in a crashing hurry, you might find cigar-lighter cables will get you started (Fig. 4). These gadgets just plug into the two vehicles' lighter plugs, providing a modest level of charging, but not carrying enough current to start a car with a truly flat battery. But if the engine will almost start, a 10- or 20-minute charge will get you on your way, out of the weather and with clean hands. Remember that the lighter socket on the dead car must be electrically hot with the key off, or you'll need to turn on the key in the dead car to complete the circuit.

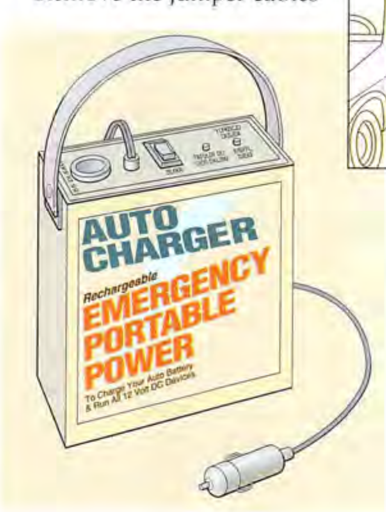

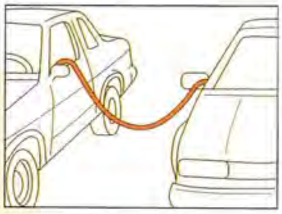

(Fig. 4) Minor cases of won't-quite-start will answer to simple-to-use lighter-plug starters and jumper cables.

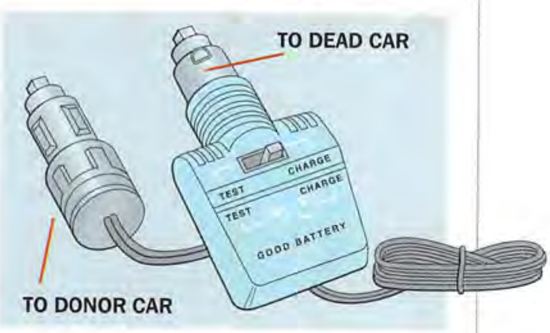

MODERATE

Repairing a Leaky Radiator

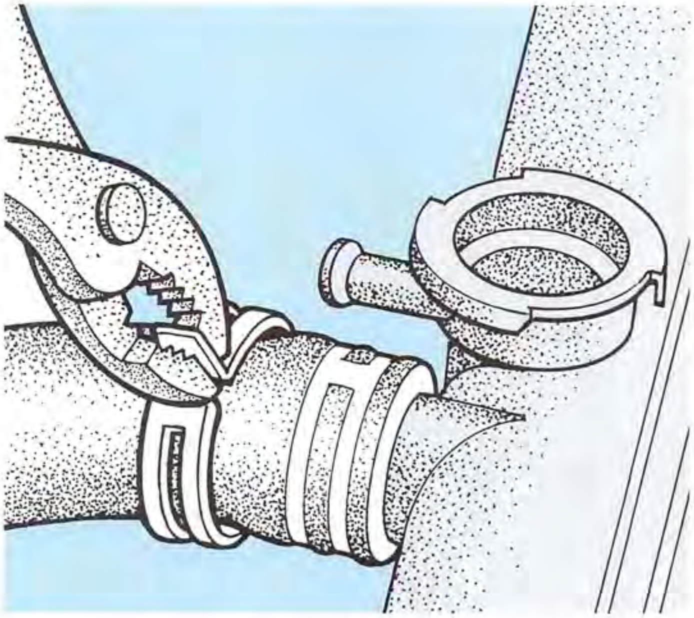

(Fig. 1) Loosen and slide the hose clamps on the hoses clear of the radiator necks. Then, disconnect the hoses.

Your radiator is leaking. Well, it happens. If you haven't changed the coolant on a reasonable schedule, it can become corrosive and eat through the radiator. Even normal shaking and vibration will eventually cause separations between the parts that are connected by solder, epoxy or mechanical crimps. Thermal forces from engines heating up and cooling down can cause cracks in stressed areas. Sealing gaskets can compress and allow leakage. And if you have a rear-drive car with a mechanical fan and weak engine mounts, the engine can lurch forward under braking far enough for the engine-mounted fan to "kiss" the radiator. Or it could be just plain bad luck—a rock gets thrown up and punches a hole in a tube.

Whatever its root cause, a leak is generally obvious. But what to do about the radiator? Replace it? After all, today's radiators are supposed to be unrepairable. This is not true at all, and the best thing you can do is take out the radiator and bring it to a radiator shop for a free consultation. If it can be repaired, you'll not only save money, but the repaired original-equipment (OE) radiator probably will be a lot better than a new discount part. It's no different from any rebuilt automobile component, except that the radiator has no moving parts.

Everything starts with removing the leaking radiator. Although doing so takes some care, it's a straightforward job. In the easier cases, you can lift it out with shroud and fans attached, then complete the strip-down on the garage floor. And even in a really tight engine compartment you can usually get it out without touching the a/c lines, or at worst you would only have to unbolt the radiator from the same module that holds the a/c condenser.

Begin by disconnecting the battery ground strap (even if the car has belt-driven, rather than electrical, fans). The next step normally is to drain the radiator, but before you do, inspect both radiator tanks carefully. If you find a heater hose connection, as on a few cars, pinch off the hose just a few inches from the neck, using locking pliers. (If necessary, use a pair of tongue depressors or popsicle sticks between the jaws to get a good squeeze.) With the heater hose pinched, you won't lose coolant from the heater during the drain. It can be tough to get coolant to completely fill the heater on some cars, so unless you're prepared to do a complete drain-and-refill, this preliminary step can help prevent trapping a lot of air in the system. Air bubbles lead to engine hot spots and ping in summer operation, and poor heater output in the winter.

If there's an underbody plastic cover at the front, remove it, even if it's not necessary to do so to drain the coolant. With the cover off, you'll surely be able to collect the draining coolant with less splash. You'll also be able to get to the lower radiator hose more easily, and you can inspect the lower end of the radiator as well. You might see a bolt-on connection to the condenser or the support member that you wouldn't otherwise notice. And with most Japanese cars, which have downflow radiators (tanks at the top

(Fig. 2) If you have an electrical cooling fan, you'll need to disconnect the wiring to it.

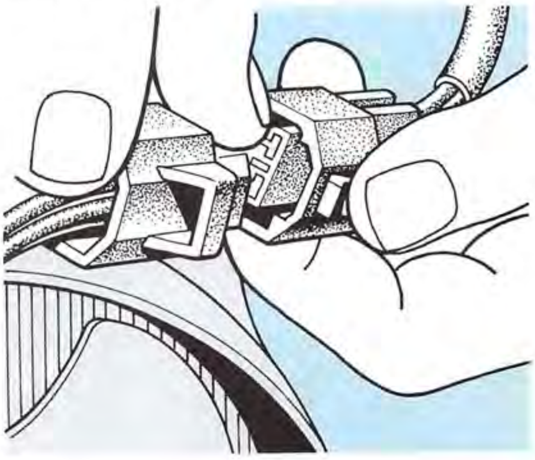

and bottom), you'll be able to reach any hoses or wires that you must disconnect from the bottom tank.

Ready? Now remove the pressure cap, even if it's on a separate tank or on the engine. Remove the radiator drain plug (or open the drain cock) and let the radiator drain. Move the clamps off the necks toward the midpoints of the radiator hoses (**Fig. 1**), then disconnect the hoses—including the heater hose if there's a connection at the radiator. If the hoses are heat-sealed to the radiator necks but in good condition, slip a thin screwdriver between the hose and neck, and work it around to free the hose, then twist the hose and work it off. Don't pry with a large screwdriver or you could distort a metal neck or even crack a plastic one. You may find it necessary to carefully slit the hose with a sharp knife to get it to let go. Don't scratch the neck, or it'll leak in the future.

As you disconnect the lower hose, a fair amount of engine coolant is likely to flow out, so make sure you have a large catch pan underneath.

Evaluate the radiator installation and begin the wrench work to free it up. A typical approach is to remove the upper mount/retainer on each side, or the single support member across the front. If there's enough clearance for the radiator to come up and out with its shroud and electric fans intact, just unplug the electrical connectors for each fan (**Fig. 2**). If not, you'll often be able to just unbolt the shroud and fans from the radiator and tilt them back to provide clearance to lift out the radiator.

PULLING OUT THE RADIATOR is easy once it has been drained. You may need to disconnect or remove the fan shroud.

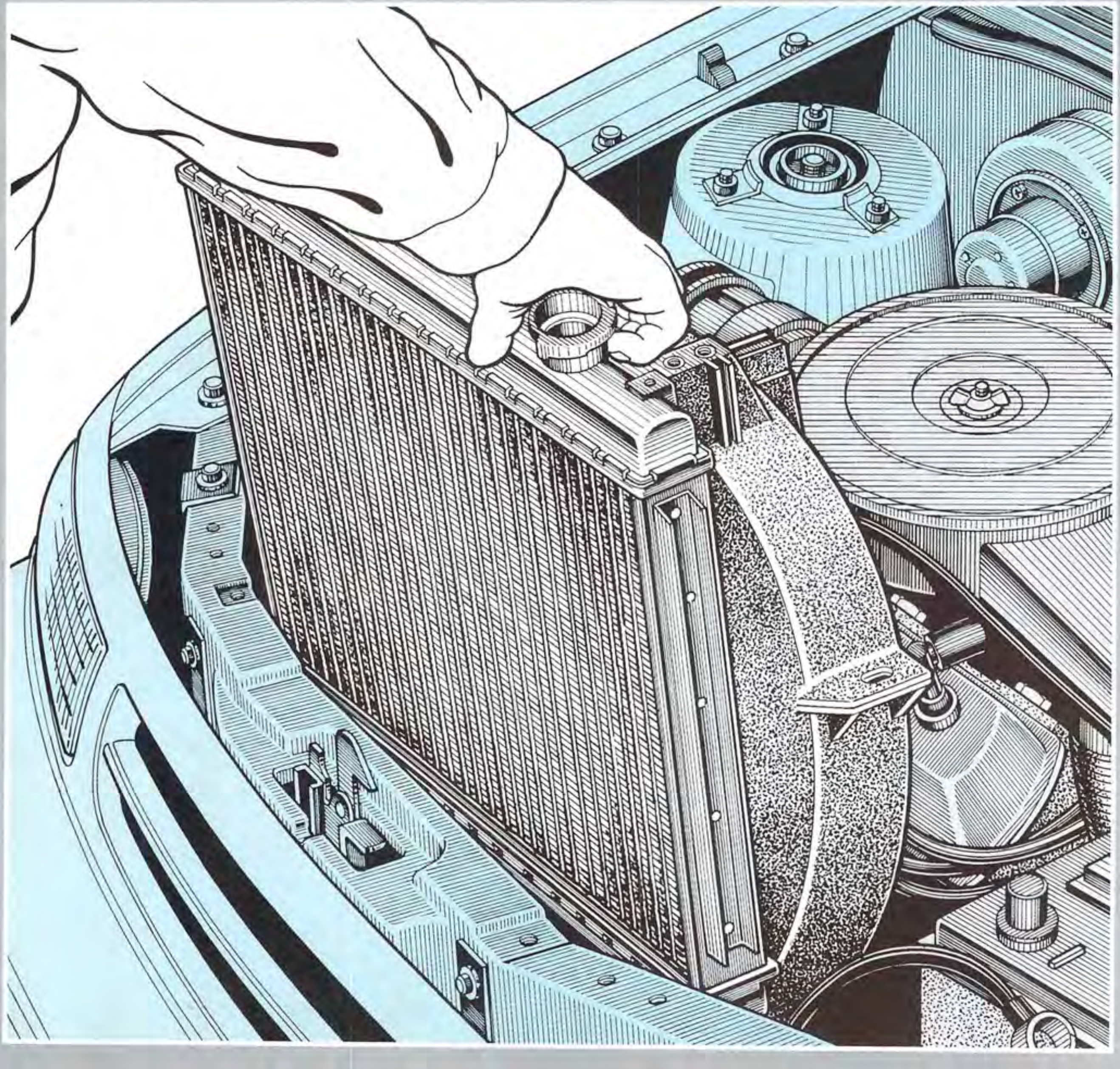

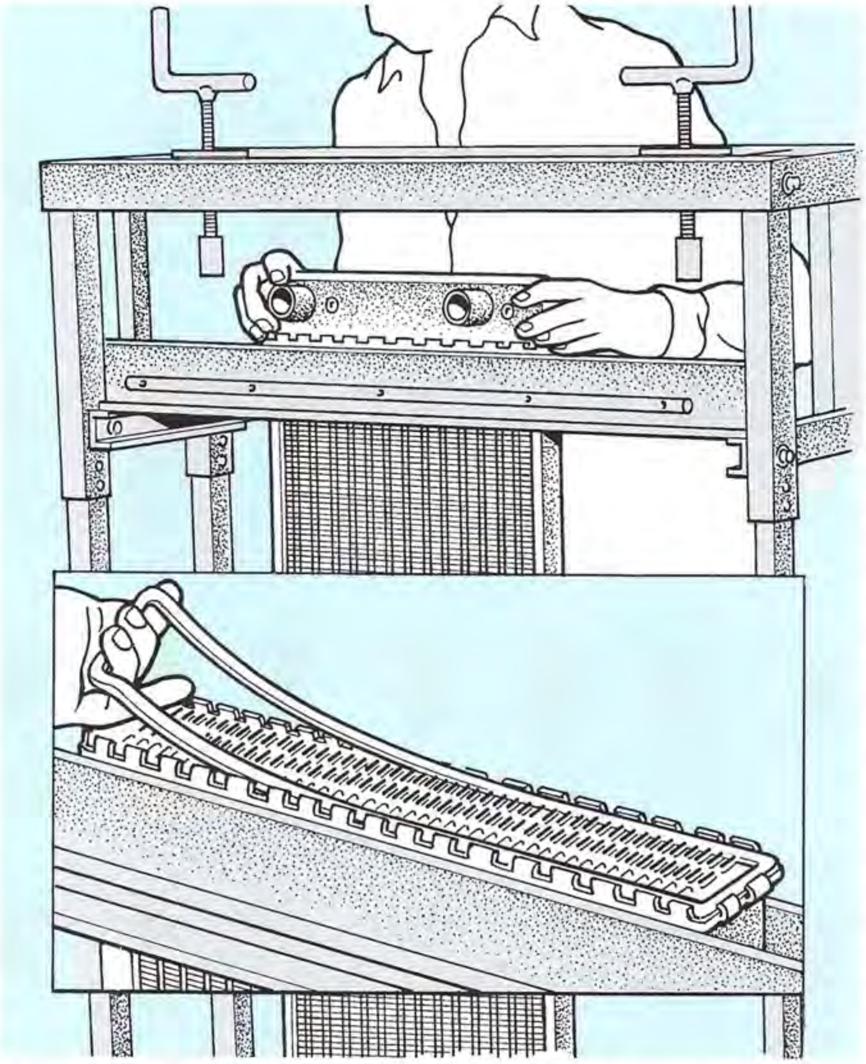

(Fig. 3) A radiator shop can replace a leaky O-ring seal, one or both end tanks or the core relatively inexpensively.

Disconnect the overflow hose and if the reservoir is in the way, take it out. This can give you your first close look at the reservoir. Many reservoirs develop hairline cracks that allow hard-to-trace leaks, so if you have a leak here, you just may have saved a bundle. If not, keep working.

If you have an automatic transmission, the radiator usually contains an automatic transmission cooler. Disconnect the cooler lines, a job that isn't always a simple wrench operation. In the easiest cases, there's a hose connection and a clamp. In the most difficult, there's a quick-disconnect fitting that requires a special (but inexpensive) tool to separate. If you have a quick-disconnect, as on most Ford products, and you can't get the tool, you'll find that the radiator end of the quick-disconnect has fittings that thread into the cooler. So a fitting can be unscrewed, but be careful. Use a tight-fitting wrench and unthread slowly, so you don't damage the quick-disconnect coupling.

Whatever type the cooler connections are, be prepared to plug them. Yes, you want to avoid transmission-fluid leakage. But more critically, if any amount of coolant gets into the cooler or its lines, it can be sayonara for the automatic transmission. You can plug a hose end with a bolt. Plug a fitting with a rubber eraser, vacuum hose plug or some other clean, soft, flexible material that will stay in place (but also come out easily when you're ready to reinstall the radiator).

All clear? Lift out the radiator and strip off any parts you were able to leave in place. Virtually all late-model vehicles have radiators with plastic tanks crimped onto the core. Your next step is to take the radiator to a radiator shop that actually repairs plastic-tank radiators. The National Automotive Radiator Service Association (NARSA) has been giving its members intensive training in the repair of these radiators, whether with copper/brass or aluminum cores, for several years (215/541-4500). So a radiator shop should know how to fix it, and not simply try to unbox a cheap new radiator and sell it to you. The cheap radiators just don't transfer heat as well, and unless you can get a spectacular price on a well-made replacement, you're usually better off with a repair.

Can it be repaired?

Almost all car, light-truck and sport/ute radiators have plastic tanks, either with aluminum or copper/brass core sections. They're precisely engineered to cool. And given today's tiny grille openings or no-grille underbody breathers, a cheap radiator is a bad gamble. It may look shiny new, but it could corrode and fail a lot faster than a properly repaired unit.

When a radiator leaks, the most common causes usually are readily repairable at far less than the cost of a new quality radiator. It certainly has been true with the copper/brass radiators largely out of use. It's even more likely with today's designs, particularly aluminum-core radiators, which can run $200 to $400 for a first-rate replacement. Here are some repair examples:

- Leaking gasket between plastic tank and core: The shop uncrimps the tank, installs a new gasket and recrimps the tank (Fig. 3). It should cost less than $75.
- Damaged tube (perhaps from a stone): These can be fixed for an extremely modest labor charge with special epoxy or solder. Even aluminum radiator tubes can be soldered.

- Cracked plastic tank (perhaps from thermal stresses): A new tank today is relatively inexpensive for most makes, so the job should cost less than $150.
- Leaking header: The headers are the metal plates at each end of the tubes. The radiator tanks attach to the headers, either by crimping a gasket joint or by soldering. Older Ford radiators have epoxy sealing between the headers and the tubes. If leaks occur from cracked epoxy, a radiator shop can remove the old epoxy and install a new sealing film using specially shaped guns and other tools. The cost is typically about $175 to $200 for both ends of the radiator, about half the price of a new Ford radiator.
- Fan "kiss" or corroded radiator: the core is smashed in or leaking from corrosion pinholes. If the tanks are good, the shop can install a new core—perhaps even an OE one—and reuse the old tanks, saving about 25 to 30% on the cost of a new radiator. And the result should be equal to a new OE radiator. If the OE radiator is an inexpensive design, then there won't be much of a difference (perhaps less than $75), and you might as well buy new if it's in stock. However, one core fits many different sets of tanks, so stocking cores is simpler for a shop. Result: The choice may come down to a new core that's available now versus waiting for a new radiator and having to spend up to $100 more besides.

Putting It All Together

When you're ready to install, inspect the rubber support bushings for the radiator. Many of today's radiators are held by mounting systems that allow them to shake at a frequency that absorbs engine vibration. Replace heat-cracked or oil-soaked bushings. Otherwise, engine idle will be rough and the radiator may shake itself into a leak. Clean the threads of oil-cooler quick-disconnect fittings and apply Teflon pipe sealer before reinstallation.

If your drained coolant was relatively fresh, it's okay to put it back in if it's not contaminated with dirt, oil or transmission fluid. Otherwise, use a 50/50 coolant/water mix. If you've taken this opportunity to flush the system with water, then add half the system's capacity of pure coolant and top off with water.

Always check the carmaker's instructions for filling a radiator—and then go beyond that and add a little more. Today's cooling systems are notorious for convoluted passages that balance heat transfer if the system is full, but trap air bubbles if not completely filled. If there are air bleeds, they must be open during the fill-up. Jack up the front of the car to raise the pressure-cap fill neck before you pour in coolant. Even if the fill neck isn't on the radiator itself, the higher it is, the better the system will burp air as you pour in coolant. Make sure the reservoir level is correct and that the air-sealing gasket in the pressure cap is in good condition, so the system won't draw in air as the coolant cools.

And "thermocycle" the system, which simply means to warm it up, then let it cool down. Check the radiator coolant level when the system is cool, and top it off if necessary.

Once is not enough. Keep checking every day or so for a few weeks. Some air purging will continue, and the system should draw in additional coolant from the reservoir. The reservoir level drops modestly? Hey, that's the way it's supposed to work. Top it off when necessary and you can conclude that all has gone well.

MODERATE

Curing Pulsing Brakes

The driver in the left lane suddenly realizes he's about to miss his exit, so he cuts in front of you. You mash the brake pedal, and it vibrates like a running chain saw as the antilock braking system (ABS) keeps you out of a skid. Thankfully, you don't rear-end the jerk's car, as he also brakes to slow down. He's on the off ramp and gone, so you don't get the chance to pull alongside and express your feelings with universal sign language. But that's probably a good thing.

While it can be reassuring to feel that ABS-connected brake pedal pulsating beneath your firmly planted foot, it's reason to suspect a problem if you get the same pedal pulsation with a light to medium braking application.

ABS-esque

If the brake feels like the ABS is cycling, but you know that it's not, check the individual axle trigger wheels in front (and on some cars also in back) and the adjacent wheel speed sensors. If you see a damaged wiring connector, bent sensor mounting bracket or stone-damaged trigger wheel, that's a likely cause. But on almost all cars, the cause normally is something in the service brake itself, and the primary problems are ones you usually can check and correct yourself.

Before you check even the most likely problems on a late model, make absolutely sure there isn't a factory problem with a specific fix. Sure, this is low on the "likely" list, but it's a lot easier to look for a bulletin than to pull wheels, etc. One possibility: the AllData Web site, a leading supplier of information for professional mechanics (www.alldata.com). If you work your way through its consumer information section and technical service bulletins (TSBs), you can check the titles of all bulletins for your specific make and model car.

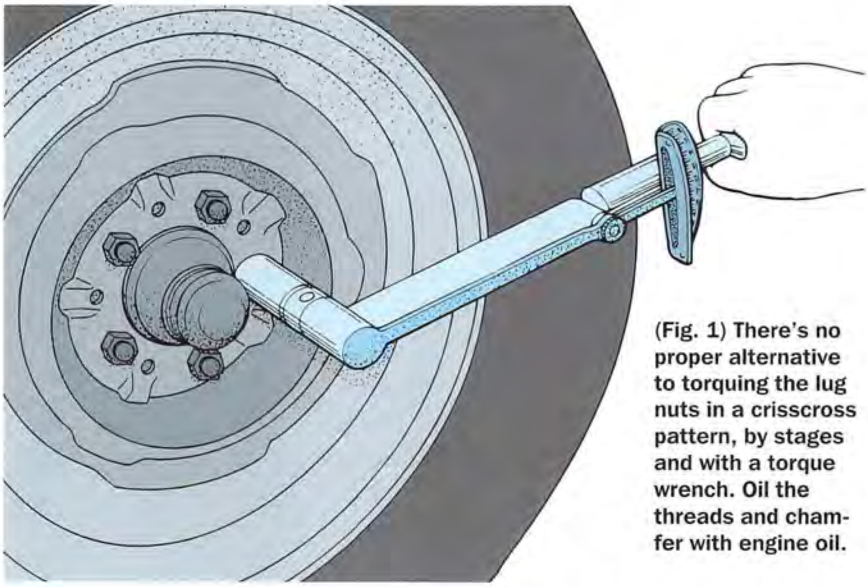

(Fig. 1) There's no proper alternative to torquing the lug nuts in a crisscross pattern, by stages and with a torque wrench. Oil the threads and chamfer with engine oil.

Lug Nut Torque

The No. 1 cause of brake pedal pulsation is uneven lug nut torque. The usual reason is that some mechanic overtightened the wheel lug nuts with an impact wrench, distorting the brake disc.

How do you know if your torque wrench is accurate? Well, if it's by a name-brand tool company and hasn't been thrown around or had a cement block dropped on it, it should be okay. Check the torque at each wheel and compare the reading with manufacturer's specifications. If there's a difference of 20% between any two lug nuts, that's too much. Try loosening all the lug nuts, cleaning the stud threads with

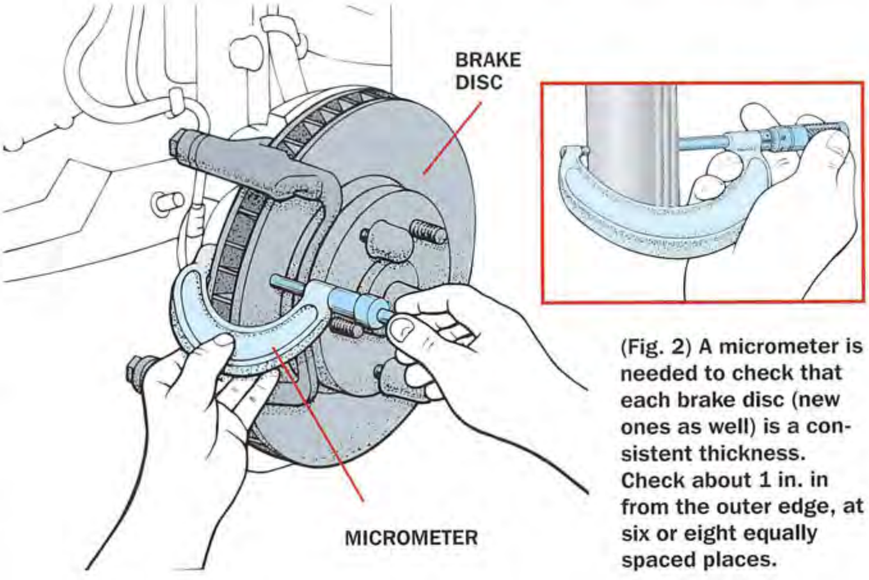

(Fig. 2) A micrometer is needed to check that each brake disc (new ones as well) is a consistent thickness. Check about 1 in. in from the outer edge, at six or eight equally spaced places.

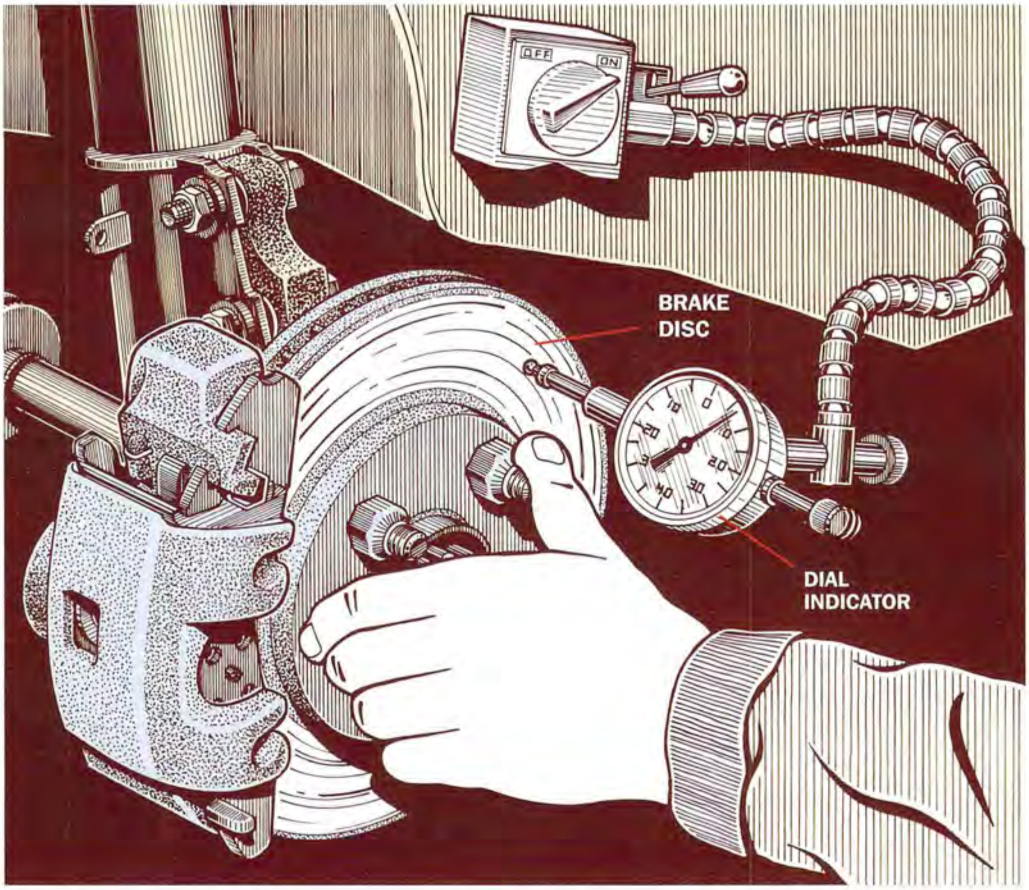

(Fig. 3) Borrow or rent a dial indicator to measure brake disc runout. Even a scant thousandth of an inch of wobble will make your right foot dance on the pedal.

a wire brush, lightly coating them and the nut chamfer with clean engine oil, and then reinstalling the nuts fingertight.

Next, using a crisscross pattern, (Fig. 1) tighten all of them to about one-third the specified reading, then to two-thirds and finally to the specified torque.

Basic Diagnosis

If simple retightening doesn't fix the problem, particularly if all wheels were off torque specs, you'll have to pull the wheels for a closer inspection. First, isolate the problem to the front or rear wheels. Try driving the car and then slowing or stopping with the parking brake lever lightly pulled up and your finger on the ratchet's release button. If the car has a pedal parking brake, try applying that lightly, but be sure to pick a safe, deserted parking lot with smooth pavement. If the car slows down smoothly to a stop, the brake problem is in the front wheels. If the car decelerates in surges, one or more of the rear brakes may be out-of-true.

Full Inspection

Actually inspecting the wheel and brake is the next step. Make alignment marks for the wheel and a lug stud before removal so you're able to index the rim to the same stud. With the wheel off, inspect the mating surfaces of the wheel and disc hub surface. If they're packed with dirt or badly rusted, clean them thoroughly (use 100-grit sandpaper or emery cloth). Remove sanding residue with a cloth and brake solvent. Lube sparingly with antiseize.

Look closely at both sides of the brake lining contact surfaces of the disc. Moderate scores (too small to snag your fingernail on) are not normally a concern, but rust or other materials are (typically brake lining transfer). Rust often occurs on cars that sit for extended periods in damp climates with not-so-clean air. Use a finer abrasive (200 grit) on the disc contact surfaces.

'Miking' A Disc

Check the disc for uneven thickness, using a micrometer (Fig. 2) at six to eight evenly spaced locations on the disc, and compare the readings with manufacturer's specs. As little as .0005 in. (that's five ten-thousandths!) is the maximum, and you can believe that you'll feel a lot of disc pulsation at .0008 in. Should you get the disc machined? A lot of evidence says unless a shop does a good job of maintaining its brake lathe, the results often are poor. And if the disc is cut too thin, it will warp easily in service. A brand-new disc is the best choice, but even that should be checked for even thickness.

Measuring Disc Runout

Measure lateral runout with a dial indicator (Fig. 3) on the lining contact surfaces, on both sides, with the lug nuts reinstalled and tightened to approximately 30 ft.-lb. The maximum amount of runout should be .003 in. If it's more, index-mark the caliper to the hub, and remove the caliper and then the disc. Inspect the mating surfaces of the disc and hub and if there's rust, remove it with gentle use of 200-grit paper. Also remove any debris and clean the surfaces with brake solvent.

Rotate the disc clockwise one stud, reinstall, tighten the lug nuts and recheck. Do this a couple of times to see if you can bring runout down to an acceptable level (the problem may have been produced by

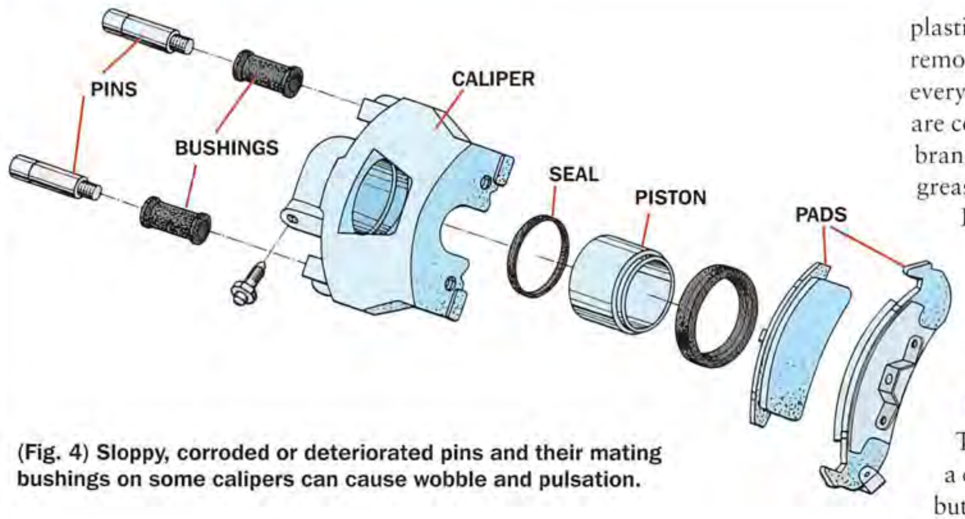

(Fig. 4) Sloppy, corroded or deteriorated pins and their mating bushings on some calipers can cause wobble and pulsation.

failure to make and follow alignment marks after some earlier service procedure).

If you can't bring runout down enough, check the hub for runout—.002 in. is the maximum. The thought of replacing the hub may not be appealing, but unless runout is far beyond reason and pulsation is intolerable, keep it in mind if nothing else surfaces as a possibility. A professional shop with an on-car lathe may (we repeat, may) be able to machine a disc so that it's referenced to the hub, to compensate for a small amount of excessive hub runout.

Brakes Not Releasing

When there's an uneven thickness problem on the discs at relatively low mileage (under 50,000), accompanied by short lining life, it's possible the brakes aren't releasing fully, allowing the shoes to stay in contact with the discs. A defective caliper piston (sticking when you try to push it back), a bad piston seal (not retracting the piston) and a sticking caliper are all possibilities.

When the caliper doesn't slide smoothly back and forth, that's trouble, and short lining life from failure to retract the shoes completely is just one consequence. The brake shoes may also slap unevenly against the disc, another cause of pulsation.

Most calipers slide along guide bolts or pins, through bushings with plastic or metal sleeves. When you remove the caliper (Fig. 4), inspect everything. If the guide bolts or pins are corroded, replace them with brand-new ones, lubed with silicone grease (don't try cleaning them).

Ditto (including silicone lube) for the sleeves if they're corroded or cracked, and the bushings if they are deteriorated.

Rear Drum Brakes

The rear drums are not as frequent a cause of pulsation as rear discs, but if they're warped, or if the linings are not making good contact, they certainly can be responsible. Unless a drum contact surface is badly grooved (grooves here are not as acceptable as on a disc), you'll need a drum gauge to check for taper or out-of-round. Or ask a machine shop to measure it for you, as the typical shop will do it for free or a nominal charge. Linings that are worn very unevenly are a tip-off to a drum that should be measured. Also check for broken springs, or springs that show obvious signs of weakness by allowing shoes to move easily.

Invest in a good torque wrench, keep it in the trunk and use it to tighten the lug nuts instead of the lug wrench when you have a flat. You should enjoy smooth moderate stops for the life of the car.

HOW IT WORKS

Tire Imbalance

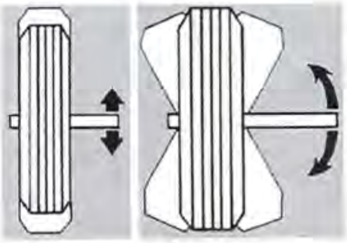

Tire imbalance, per se, shouldn't cause a pulsating brake pedal. Brake pedal vibrations are almost always caused by an out-of-true condition somewhere in the braking system. But on rare occasions it can happen. Normally, each tire has to be balanced statically first—i.e., so no part of the tire is heavier and always sinks to the bottom. Then they all need to be balanced dynamically, so the rim doesn't wobble at higher speeds. Any imbalances from either of these conditions normally will manifest themselves as steering-wheel vibration, not brake pedal vibration. Even a tire with a high or low spot will normally be felt through the wheel.

But once in a great while you'll find a tire that has inconsistent stiffness. Some parts of the tire's sidewall will be more or less easy to deflect under load than others. This won't show up on the wheel balancer, but will make a brake pedal pulsate at a rate proportional to your road speed. The diagnosis is to swap for a different tire and see if the problem goes away.

MODERATE

Getting Your Car Ready For Winter

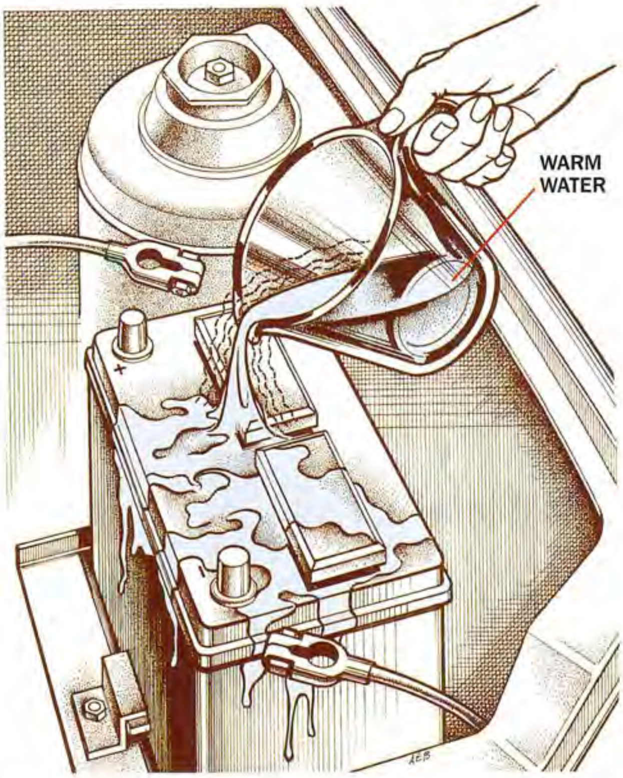

(Fig. 1) Use warm water to clean the battery posts and clamps. Use baking soda for stubborn sulfate deposits.

Snow covers your car. After 10 minutes of shoveling snow and ice from the roof and windows (not to mention 30 minutes of shoveling the driveway), you're ready to go to work. Twisting the key languidly brings the engine to life, much to your relief. But ... the windshield wipers leave streaks, the washers drizzle fluid inches short of the glass, the battery light flickers and the heater fan blows little air of any temperature. Not a particularly good day for even a late-model car that's been neglected.

Winter Takes Its Toll

The traditional tuneup is gone. Newer cars are largely self-tuning, and don't have ignition points or carburetors to tweak. Factory-fill radiator coolant in some new cars is rated for five years or far longer than many people keep a car in this age of low-monthly-payment leases.

Cars are still imperfect, mechanical contrivances, however. Although many maintenance requirements are reduced, there's still plenty of stuff that can break or fall off. An afternoon of preventive maintenance will greatly reduce the possibility of bad things happening. It's autumn—the leaves are falling, but the weather is still relatively warm. A couple of hours of maintenance will be a lot easier to take now than a couple of hours of repairs when you have to shovel your way an extra 50 ft. to work in an unheated garage.

The Easy Stuff

Clean. Clean out the interior and the trunk. Shampoo the carpets and the seats, because a winter of slush and melted snow will infiltrate dirt deep into the fibers, there to remain forever. Dump the ashtray. Clean the wheel well arches and undercarriage of mud and road dirt, so that salty slush doesn't soak into them, providing a perfect environment for rust.

Electrical

While you're cleaning, clean the battery with warm water. Remove the terminals and wirebrush them and the battery posts with warm water and baking soda (Fig. 1). Reattach the terminals and coat all exposed metal with petroleum jelly. You're not done yet—using some sort of household cleaner, clean all traces of dirt and oil from the battery's top and sides, particularly near the terminals. Traces of dirt can trap moisture, especially during damp winter days. This moisture acts as a conductor along a path that normally doesn't conduct electricity at all, and can leak enough current between the battery posts to prematurely drain your battery. In fact, consider replacing your battery if it's more than 4 years old. Newer cars tend to use smaller batteries to reduce weight and improve mileage, and, combined with high underhood temperatures, that spells an earlier demise than you may have gotten 10 to 15 years ago.

While you're at it, check all the electrical connections for looseness and corrosion. That may mean getting underneath the car to see the ground strap and solenoid/starter motor connections.

Check all light bulbs and replace any that aren't working. The days are short during the winter, and you'll depend on these bulbs for visibility a much larger proportion of the day.

Underhood in General

Still in an electrical mode, check the alternator connections and mounts for looseness and evidence of overheating. High electrical demands for lights, heater fans and rear-window defrosters, as well as recharging a battery abused by coaxing a reluctant engine to life, can make a marginal connection overheat.

Check accessory belt condition and tension, because that extra electrical

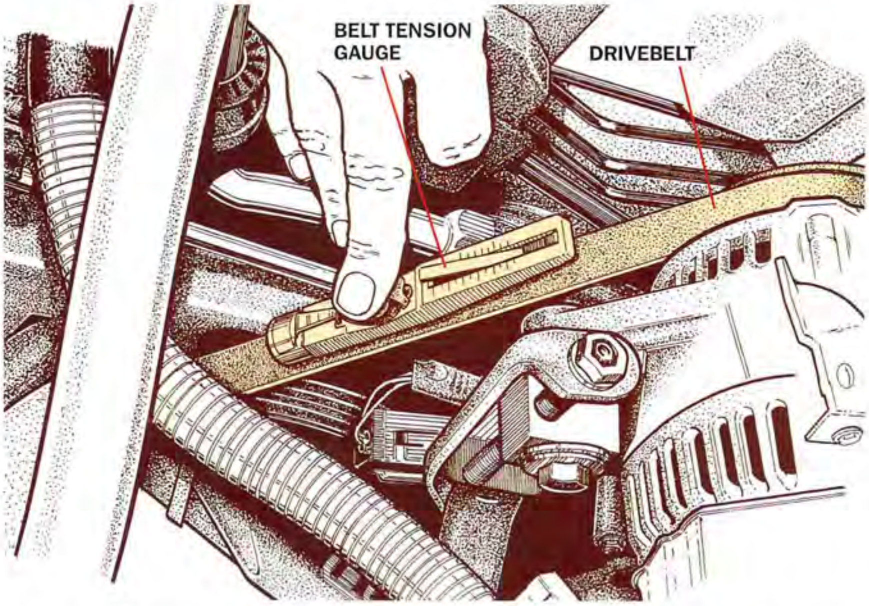

(Fig. 2) Check belts for proper tension using a tension gauge. Serpentine belts can also be checked with a gauge, although they aren't as likely to be loose.

demand will strain a marginal belt (Fig. 2). Many modern cars have a single serpentine, automatically tensioned belt, but it still can fray and come off.

Check all rubber hoses. Five years, in these times of air pollution and high underhood temperatures, is a legitimate life span for a radiator hose. If one seems squishy, brittle or just suspect, now would be a good time to replace them all. Check the date on the coolant. (Remember many new cars have extended drain intervals—but not permanent coolant. Read the owner's manual.) If it's due to be replaced, use a 50-50 mix of fresh coolant and water (Fig. 3).

Consider changing your transmission fluid, differential lubricant and brake fluid if your car is more than 3 years old. Lubricants break down, and brake fluid attracts moisture and deteriorates. At any rate, check the levels, and don't forget the power steering reservoir.

Fill the windshield washer tank with washer fluid. Check the pump and nozzles for a healthy, well-aimed spray pattern. Many nozzles can be re-aimed by inserting a pin into the nozzle to use as a handle.

Windshield wipers are essential for winter driving, and after a summer of sunshine, the rubber squeegees are probably in sorry shape. Play it safe and replace them. Be sure the wiper arms and springs are in good shape,

HOW IT WORKS

Oil Viscosity

Fall is a good time to change oil and filters. Check your owner's manual for the rating and viscosity, but most cars nowadays use an SL-service rated oil. This rating is some measure of the oil's longevity and resistance to oxidation and evaporation. SL oil is fine for older engines that specify an SG-, SH-, or SJ-rated oil, because the rating always improves when a new one is introduced. The SAE viscosity rating is a measure of how thick the oil is. All modern cars should use a multigrade oil, probably a 5W-30 or 10W-30. The W in these ratings stands for a winter viscosity. A multigrade oil will be as thin as the thinner (5W- or 10W- rating) when the engine is first started. This will allow oil to flow more rapidly and lubricate parts that have had all the oil drain off of them overnight. A thicker oil might not be pumped to remote parts of the engine rapidly enough. On the other hand, a thicker oil (the second number in the rating) will resist becoming too thin when the engine reaches operating temperature. The American Petroleum Institute is an industry watchdog and oils bearing the API emblem can be expected to meet their specifications.

In spite of what your brother-in-law the shade-tree mechanic tells you, don't use a heavier grade of oil just because you used to in your older car. Modern cars were engineered to use 5W-20, 5W-30 or 10W-30 oil year-round. Check the owner's manual to be sure. The lighter oils are also better on fuel consumption and can significantly improve mileage.

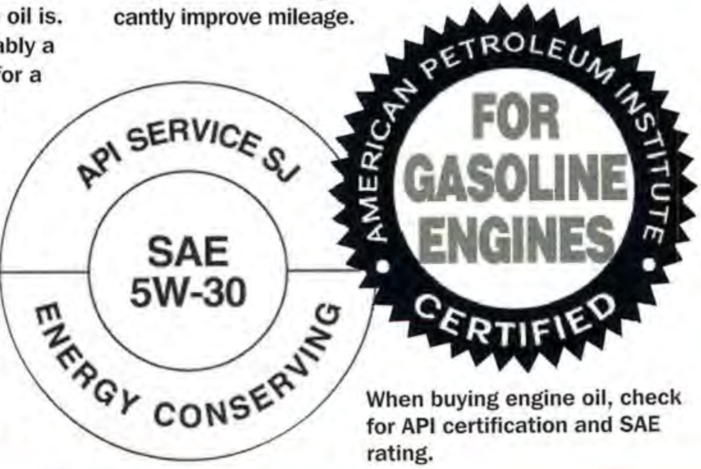

When buying engine oil, check for API certification and SAE rating.

and that the blade is held square to the windshield surface. You may need to bend it slightly to keep everything square.

Check the air cleaner, and consider replacing it and the fuel filter (if your vehicle has a replacement type, not the permanent one in the gas tank). The fuel filter will tend to trap water, and once it traps enough, it will fail and may dump some accumulated dirt into your fuel injection system. And winter, of course, is when you tend to find the most water in gasoline at the pump.

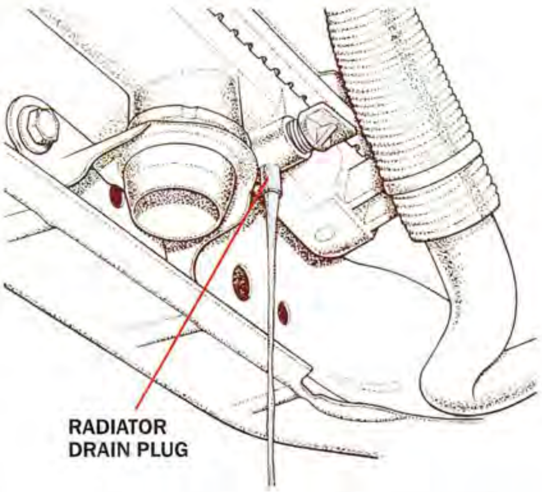

(Fig. 3) In spite of long-life coolant in many new cars, eventually you'll need to drain and refill with fresh coolant mix.

Where The Rubber Meets The Road

Driving in snow demands good tires. Be sure you have adequate tread depth, and consider changing to snow tires if you live in a heavy snow area. Go ahead and pick up a set of inexpensive steel wheels at the local salvage yard for your snows instead of having your tires remounted and balanced every fall and spring. And while the car is up on the safety stands, inspect the suspension bushings, control arms, ball joints and tie rods, and the brakes. Now would be a good time to replace the brake pads if they are more than 60 percent worn. Even if they are not, check carefully for corrosion around the calipers and sliding pins. While the wheel is off, pull the pads out of the calipers and be sure everything is sliding freely. Wirebrush sliding-key ways and pins that let the pads pull back from the disc when you remove your foot from the brake. Exercise the pistons by pushing them back into the calipers and then pressing on the brake pedal once or twice to break up corrosion between the piston and the caliper.

Again, it's a lot easier to look at this stuff on a crisp autumn day than it is to try to fix it some subzero morning when all of your wrenches stick to your fingers like the proverbial dared 6-year-old's tongue sticks to a frigid flagpole.

Rust Never Sleeps

Before the salt trucks come out is a good time to get out the touchup paint. Use it to cover all the bare metal at the bottom of any stone chips and parking lot dings. Clean the chipped area thoroughly, and use a toothpick or a match to apply a touch of zinc-rich primer to the bare metal. Allow this to dry, and chase any overlap back to the lip of the scratch with lacquer thinner. Then carefully fill the chip with touchup paint. It'll shrink, so you may need several applications to build up the level of paint to flush.

On the Road

Even a perfectly running car may wind up stuck in a snowbank. So put together a kit of essentials for your trunk. This should include flares, a flashlight, jumper cables (if only to help some unfortunate soul whose car wasn't winterized like yours), a folding camp shovel, kitty litter for traction, a can of gasoline anti-freeze, and tire chains if they're legal in your state. If you ever travel outside of urban areas, a couple of blankets and a cellphone or CB radio ought to be on board as well.

Repairing a Flat Tire

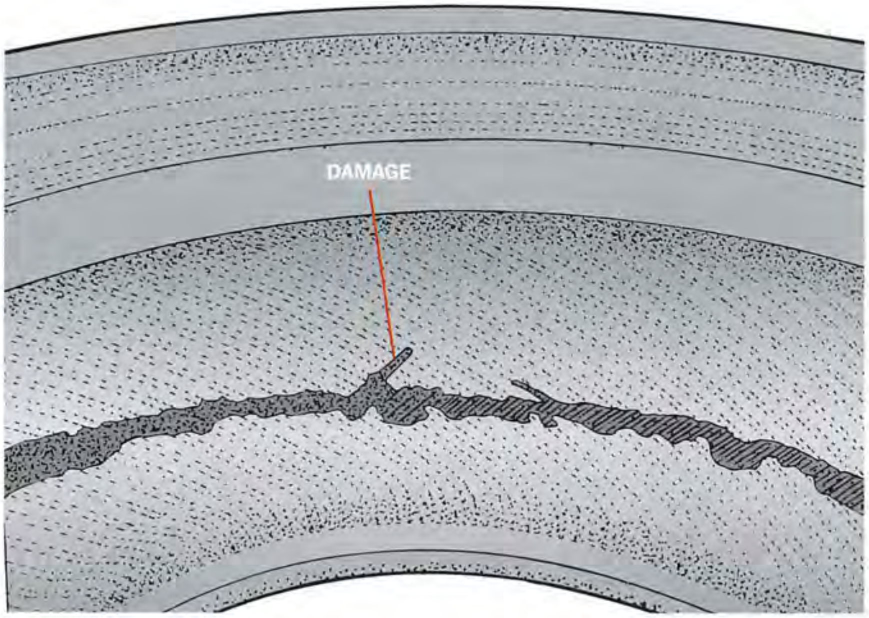

(Fig. 1) This damaged area inside the sidewall was caused by driving the tire when it was flat.

There's rarely a more welcome sight than the inside of a warm garage late at night, as your door rolls open to welcome you and your family home after a trip out of town. Particularly tonight, because you're far later than expected. Flat tires are generally no more than a nuisance—but this time you found your spare tire to be as flat as yesterday's beer. Fortunately, a nice young gentleman working in a gas station only a half mile up the road managed to plug your tire within a few minutes. You watch the inflation pressure of the tire for a week or so, and everything seems fine.

Until a few weeks later. Then the same tire disintegrates suddenly on the interstate in heavy traffic. You manage to maintain control of your SUV by lifting partway off the throttle and gingerly steering over to the shoulder, ignoring the cacophony of the failed tire as it flaps violently in the wheel well, shedding parts as it tears into pieces. It's not just flat, it's a smoking ruin, and there's substantial damage to your wheel well's sheetmetal to boot. Obviously, there was more damage to your tire than met the eye—and the nice fellow who repaired it for you didn't necessarily do you any favors.

The Right Way—And The Wrong Way

The tire industry has a rule about externally plugging a tire with a worm. Don't. It's a qualified exclusion, and we agree with it, for good reasons. An externally applied worm might be an acceptable emergency get-home-tonight quick fix, but a properly trained tire technician must later complete the repair by removing the tire from the rim and doing a complete inspection of the tire for additional punctures and damage. Damage can come from not only the nail or whatever caused the air loss, but also from running the tire at low pressure or flat. Driving more than a few feet with the tire at zero inflation pressure may cause the rim to gouge the inner liner, damaging the body cord. Driving at more than a walk with the tire at very low pressure may build up enough heat to damage the sidewall beyond serviceability (Fig. 1). The portion of the nail or wire that entered the tire's interior may have flailed around and ripped the inner liner or plies (Fig. 2). After inspecting the tire, the technician may apply a proper patch, remount the tire and reinflate it.

Here's How

Okay, we doubt many readers will ever repair a flat themselves because it involves some pretty expensive tools. However, we've done it with nothing more than hand tools and a 12-volt portable compressor, and more than once, too. Once, it was late at night and we didn't want to wait for the tire shop to open. Another time, it was because we were in a very remote area, and a trip to town was nearly a two-day hike.

Busted

Wheel/tire assembly off the car? (We've done it with the rim still bolted to the axle of a trailer but it's tough work.) Start by inflating the tire and submerging it in water to look for bubbles. Now roll the tire along the ground in a good light to look for the puncture. Chalk any suspicious holes in the tread, even if there was no stream of tiny little bubbles associated with it. You'd be amazed at how many times tires are taken in for repair with multiple punctures. Make a chalkmark on the tire sidewall next to the valve stem, because the tire should be reinstalled in the same index. Now you can remove the valve core to completely

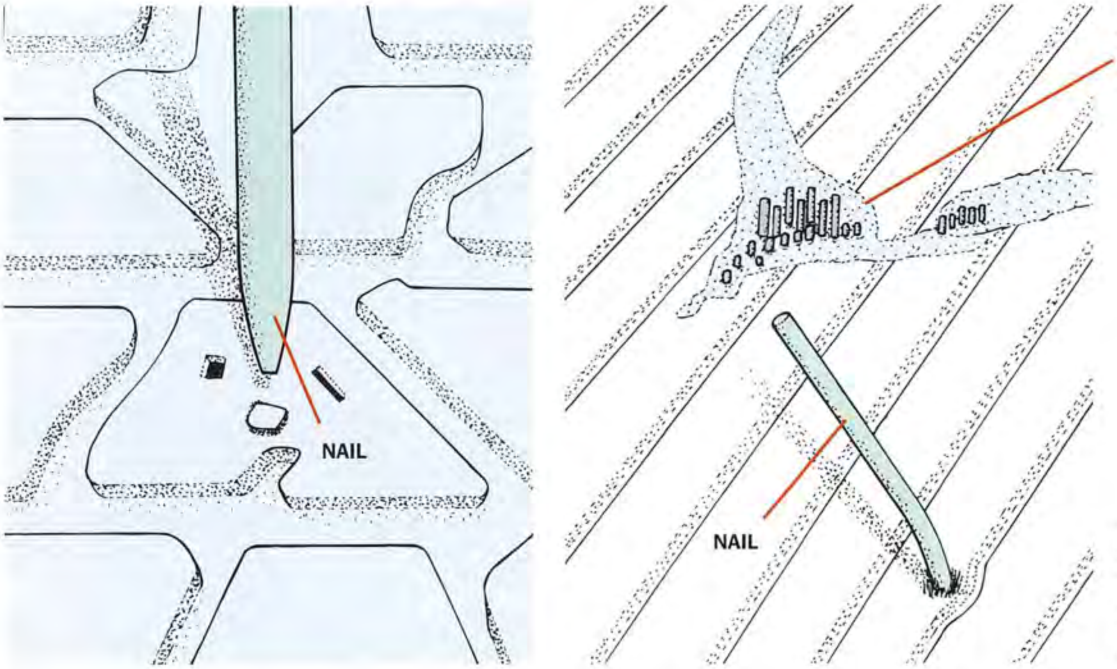

(Fig. 2) An innocuous-looking nail through the tread may have damaged the inside of the tire.

deflate the tire and demount it. Next, with a bright light, carefully inspect every square inch of the inner liner. Pay particular attention to the areas inboard of the chalkmarks you've made. Look for damage caused by the foreign object. Such a hole may be patched if the object hasn't damaged any cord, although any exposed cord usually means the tire should be scrapped.

Also look for peripheral damage caused by the tire being pinched between the rim and the pavement. Any damage to the inner liner that came about as a result of friction by being pinched and then driven makes the tire a likely candidate for immediate replacement.

Any puncture through the sidewall of the tire automatically means the tire needs to be scrapped.

Remove any foreign objects in the tread by grasping them with pliers and yanking them out. Make note of the angle the object made to the surface of the tread. In fact, you should take a thin probe, like an awl, and probe the hole straight through into the inside of the tire. Look for evidence of cut cords or separated plies, because they'll be sticking out of the hole.

HOW IT WORKS

Run-Flat Tire

A few late-model cars come equipped with tires that can be run without any air at all, at least far enough to get to a place where they can be repaired. These tires, available from several different manufacturers, can be retrofitted to any car that uses an appropriate size tire. The sidewalls of these tires have an extra piece of molded rubber near the bead. This reinforcement is stiff enough to keep the rim from contacting the inside of the tire, even under cornering and braking. The tire and auto manufacturers specify an upper limit on speed—usually 50 or so miles per hour—and a maximum range—often 50 or 100 miles—when the tire is deflated. Because the performance of these tires with zero air pressure is so good, tire manufacturers require any car fitted with them to have a tire-pressure monitoring system that will alert the driver to the fact that he's driving on a flat.

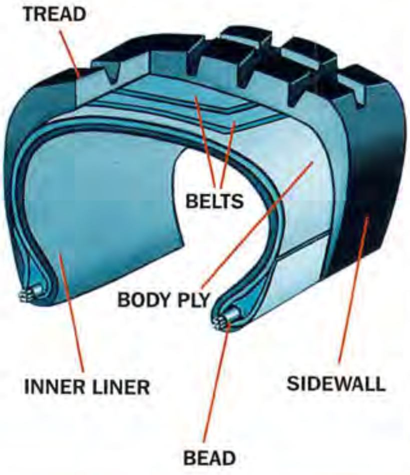

Run-flat tires' bead reinforcement provides stiffness to keep the bead seated on the rim and prevent sidewall damage.

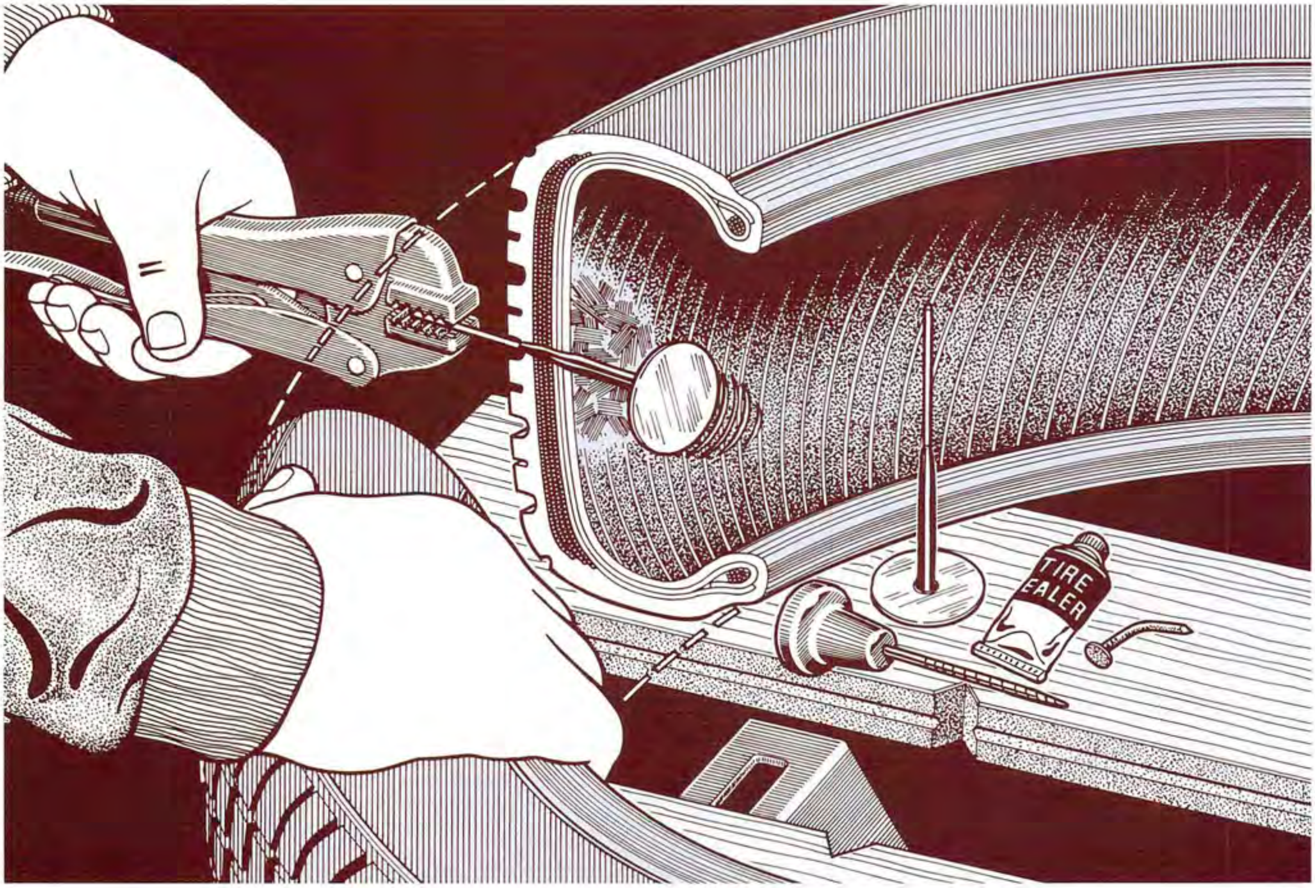

(Fig. 3) A tubeless tire puncture can be properly repaired only by breaking the tire off the rim and pulling a stem-type patch through the tread from the inside.

Straight Holes

If the hole makes a 25° or smaller angle to the tread, the correct patch to use is a "mushroom" patch **(Fig. 3)**. The head of the mushroom provides a good adhesive seal to the tire's inner liner for proper air retention. The stem of the patch serves several purposes. It plugs the hole to anchor the patch and, in concert with the chemical vulcanizing cement, prevents water from entering the hole and working its way into the tire's tread.

Using a proper tire reamer, clean out the hole from the inside out **(Fig. 4)**. Again, look carefully for evidence of wire or tire cord. The reamer will remove any road dirt or oil and any damaged rubber from the hole, leaving a fresh rubber surface for bonding the cement.

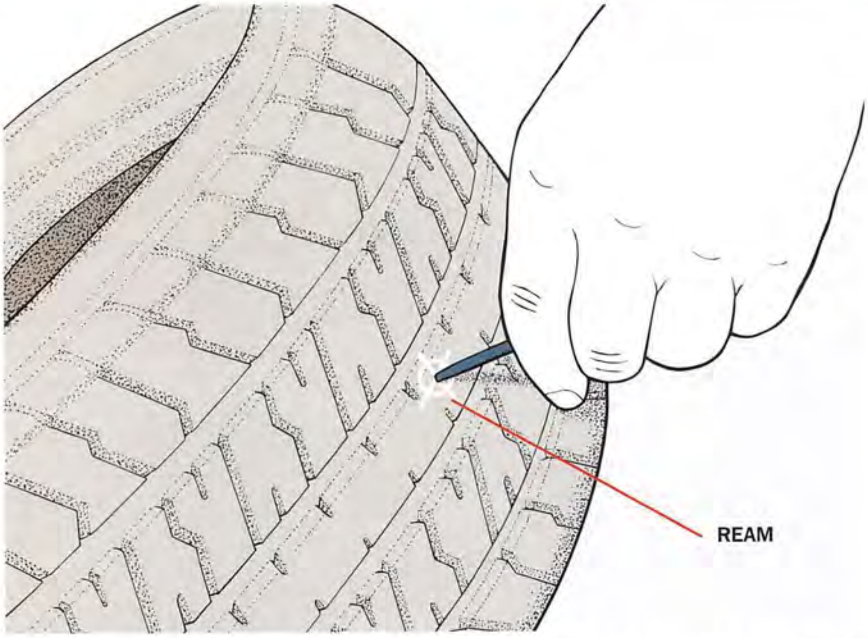

(Fig. 4) Thoroughly ream the puncture to clean the rubber and provide a good bond.

Center the patch over the hole inside the tire, and mark a half-inch around its periphery with chalk. Use a tire buffing tool to remove the surface of the rubber and leave a smooth texture. A power buffing tool is best, but a hand scraper will work. Don't buff through the liner into the ply rubber or the cord. Brush, blow or vacuum the rubber dust away from the buffed area. Don't touch this area with your fingers or a rag.

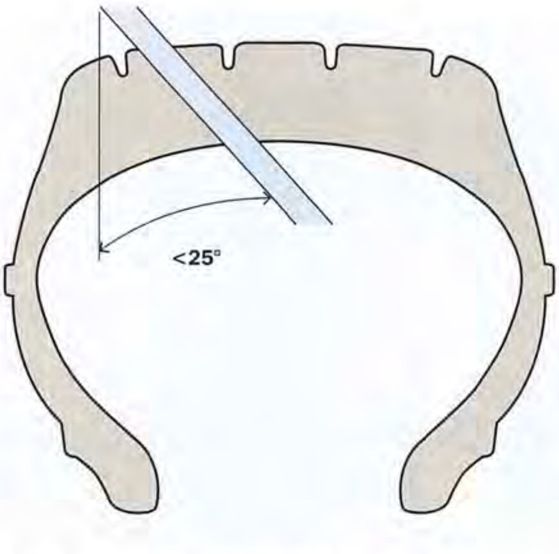

(Fig. 5) Oblique punctures require a different plug.

Coat your awl liberally with cement and run it in and out of the hole several times to carry the cement throughout the hole. Leave the awl in the hole until just before installing the patch so the cement stays liquid. Remove the peel-off tab from the patch, and put a single coat of vulcanizing cement on the head of the patch and the buffed area. Allow this to dry thoroughly, and don't touch the cemented areas because the moisture and oil in your fingerprints will prevent good adhesion.

Lightly coat the stem of the patch with cement. Pull out the awl, and pull the stem of the patch through the hole. Stitch the patch (using a tire patch stitching tool) to the tire from the middle out to prevent any air bubbles from being trapped under the patch. Now cut the protruding stem off nearly flush with the surface of tread rubber.

Oblique Holes

If the hole isn't within 25° of vertical (Fig. 5), you'll need to use a two-piece patch. This involves reaming the hole and installing a rubber stem with cement. The stem is then trimmed flush with the inner liner, the area buffed and a patch is cemented over the top of the stem.

Back On The Road

Before remounting the tire on the rim (a job for a professional with special equipment), take a hard look at the rubber valve stem. If it's weather-beaten, damaged or just more than about 5 years old, pop it out and pull in a new one. Use the correct diameter and length, and a valve core that matches. Valve stem caps aren't there to simply backstop the valve core's air retention—they keep moisture and dirt out of the stem and out of your tire.

Reinflate the tire. Use soapy water to check the beads, valve stem and the repair for leaks. And keep a close eye on your tire for a few weeks, both by monitoring the air pressure and looking for evidence of delamination or a belt failure. A failed belt usually will cause an out-of-round condition and the ride will often indicate it.

Can you adequately repair speed-rated tires? Some tire manufacturers disagree, but at least one (Goodyear) says you can. They do specify that there can be only a single repair, less than one-quarter inch in diameter, per tire.

MODERATE

Servicing Your A/C

It's hot, so hot your car's interior shimmers. Getting into the car is clearly out of the question, at least until the air conditioning hauls out a few million BTUs. You lean in and twist the key, being careful not to raise blisters by touching the steering wheel. The engine fires and idles smoothly, then you punch the max cold button on the dash—nothing happens.

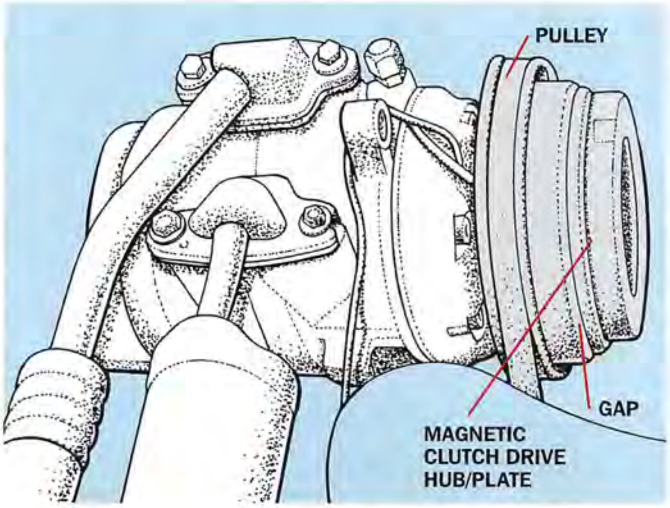

(Fig. 1) Gap between hub/plate and pulley should close when a/c is turned on.

Great, you think, at least in the old days, ordinary folks used to be able to purchase R-12 refrigerant to punch up a tepid air conditioner. Not anymore. And it won't be too long before R-12 is off the market completely—a victim of federal regulations that restrict the manufacture of ozone-depleting chlorofluorocarbons.

Fortunately, it's still possible for the average Saturday mechanic to fix some of what ails his air conditioner without having to open up the system. To keep you in the service picture, let's consider just a few of the common causes of poor—or zero—cooling that you can diagnose and cure. Start by acquiring the a/c service manual and electrical schematic for your vehicle. You'll also need a 12-volt test light. A volt-ohmmeter comes in handy as well.

No Cooling

If the compressor doesn't engage, there will be zero cooling. An electromagnetic clutch should lock the drive hub/plate on the compressor (**Fig. 1**), with the belt-driven pulley to spin the compressor shaft.

The clutch circuit is fused, so check the fuse. The fuse is good? With the engine and the a/c on, check to see if juice is getting to the clutch by probing its wiring connector with a 12-volt test light (**Fig. 2**). If the light doesn't go on, there's a break in the circuit. If it does go on brightly, attach a jumper wire to the ground side of the clutch circuit. If the clutch still won't engage, it's bad.

It takes special tools to change a compressor clutch. Unless you can rent them, leave this job to a pro.

If the clutch engages when you jump it, there's a problem with the ground circuit, which may be wired through to the engine computer. Both circuit problems require straight electrical diagnosis and a good wiring diagram.

However, the refrigeration system may also be involved. There's a low-pressure switch in the current-feed circuit that is designed to keep the compressor from engaging when pressure is low.

See if there's enough refrigerant pressure to close that switch, which requires pressure gauges, and hoses that attach to your system. When the system is running, one half operates at low pressure and the other at high. However, after the system has been shut off for a few minutes, pressures equalize, so you can check pressure at both of the service valve ports (one low-pressure, one high-pressure). They should be the same—roughly equal to the ambient temperature.

If pressure is less, the refrigeration system is low on refrigerant, and that could account for the failure of the compressor to engage. Have a professional trace the leak and then recharge the system with refrigerant. Incidentally, if the service valve caps are missing, they could be the cause of your loss, as they provide the real seal, not the valve itself. Warning: Do not try to seal leaks witha/c sealer products—period.

Important: Use only R-12 in an R-12 system, R-134a in an R-134a system; never mix the two! Right now, there's still some R-12 around, so you might want to keep the system on R-12. However, if you ever have a compressor failure, the replacement will likely be compatible with R-134a, which would make changing over to the new refrigerant simpler. But this is not do-it-yourself work. It takes special equipment.

The a/c pressure is normal? Apparently, the problem is purely electrical. And in this case the first step is to check for a computer trouble code. The engine computer? You bet. On most American cars and some imports, turning on the air conditioning at the dashboard is just a request to the engine computer to provide the electrical ground to complete the compressor clutch circuit. If the engine computer sees a reason not to, the a/c won't come on. Possible reasons, in

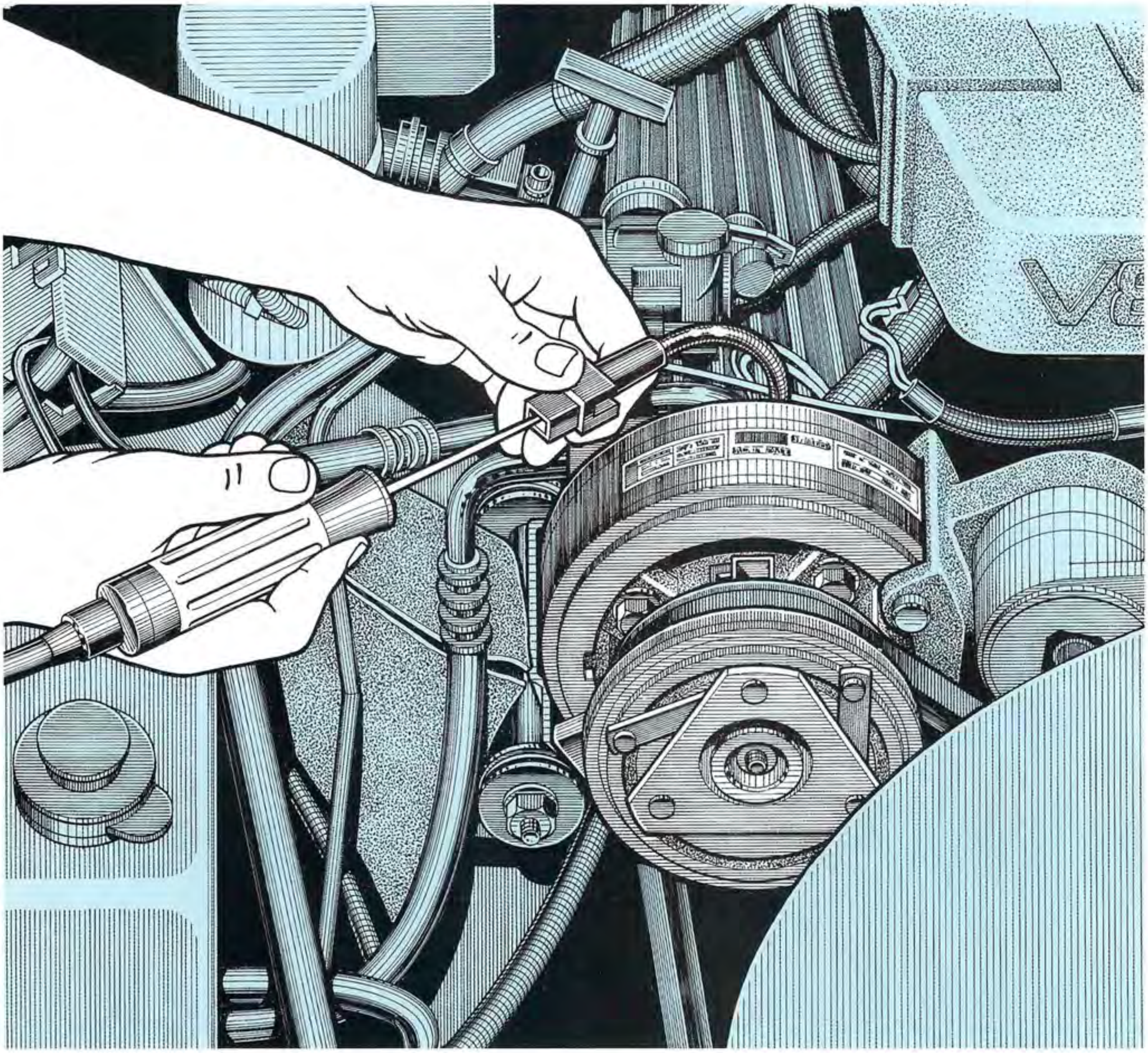

addition to low refrigerant, include warning signals from certain engine sensors. The coolant-temperature sensor may be signaling a very high coolant temperature, and allowing the a/c to come on could cause overheating. Or a power-steering switch may be signaling high pressure (as in a parking maneuver), and allowing the a/c to come on could add to the engine load from the belt-driven accessories and stall the engine. Of course, some of these signals may be wrong, and you'll need to investigate the solution to these engine problems to get the a/c online again.

No computer trouble codes? Check for a bad a/c clutch relay. Also, note that some of the computer sensors are not covered by trouble codes—the power-steering switch is a common example. If your car has one, bypass it for testing. If the compressor now comes on, replace the switch.

No clues in the computer circuit? Refer to an a/c circuit diagram to see what additional switches or relays are in the compressor-clutch circuit. On many cars of all makes (particularly R-12 systems), there's a switch that cycles the compressor clutch on and off to control performance. You can find out if the switch is bad by unplugging it and attaching a jumper wire across its connector. If the compressor now comes on, the switch apparently isn't closing. Confirm this with an ohmmeter across the switch terminals. On GM and Ford vehicles, the switch usually is on a large cylindrical can called the accumulator. On many Chrysler products, it's in a block at the firewall.

(Fig. 2) Use a 12-volt test light to probe the a/c compressor's clutch wiring. If the clutch isn't pulling in, there's no cooling.

Bad switch? On late-models, the switch is mounted on a Schrader valve. As you unthread the switch, the valve closes so you won't lose any refrigerant. Grease the new O-ring with mineral oil, gently position it on

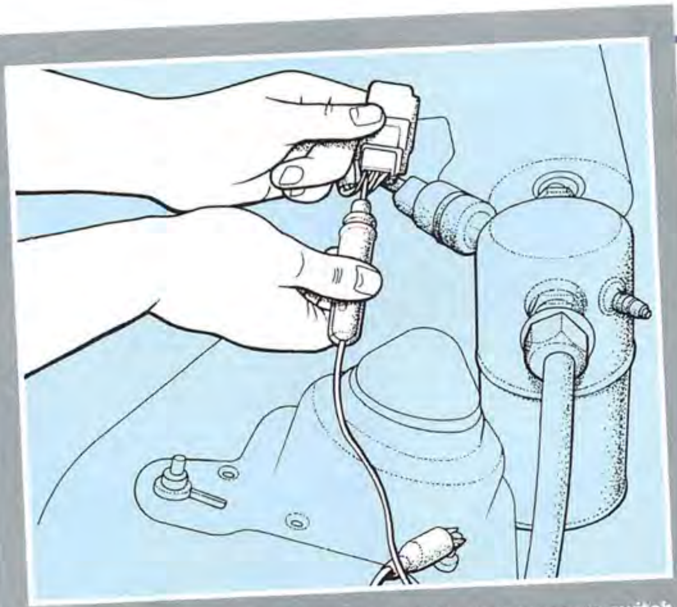

USE A TEST LIGHT to check for voltage at the low-pressure switch. Back up that reading by checking the switch with an ohmmeter.

the port and then thread the new switch into place. If the switch has a plastic body, a firm hand-tightening is enough. With metal, a gentle nudge with a wrench does it, because the O-ring makes the seal.

Note: Your circuit diagram may show other pressure switches. The variety is almost endless, depending on the type of vehicle, so without a diagram, you're in Guess City. Among the most common variety are a high-pressure cutout and a dedicated low-pressure cutout (often in addition to the cycling clutch switch). If the basic pressure check showed nothing awry, unplug and bypass these switches for testing. In most cases, the bypass is achieved with a jumper wire across the wiring connector terminals—but read the wiring diagram details, as some switches are normally closed and open up as a signal. If so, simply unplug these to bypass. Double-check each switch with an ohmmeter.

Many cars—particularly Japanese models and many late-model Chrysler products—use a temperature sensor on the evaporator (the in-car heat exchanger that cools the interior air) in conjunction with an electronic module to cycle the clutch. The shop manual should provide specifications for a resistance check of the temperature sensor and where to connect a voltmeter to test for an output signal to operate the clutch. Make these checks under the dashboard.

Poor Cooling

The temperature of the air coming out of the registers will not necessarily be very low on a humid, hot day. The a/c may expend all its effort just wringing humidity out of the air, and the air blowing out of the registers easily could be as high as 60° to 65°F. If there is some cooling, but clearly not a normal amount, make this simple check: Is the front of the condenser (the heat exchanger in front of, or alongside, the radiator) free of bugs, leaves, road film and other debris? If it isn't, airflow is restricted. Clean the condenser with a soft brush and a detergent-and-water solution (Fig. 3).

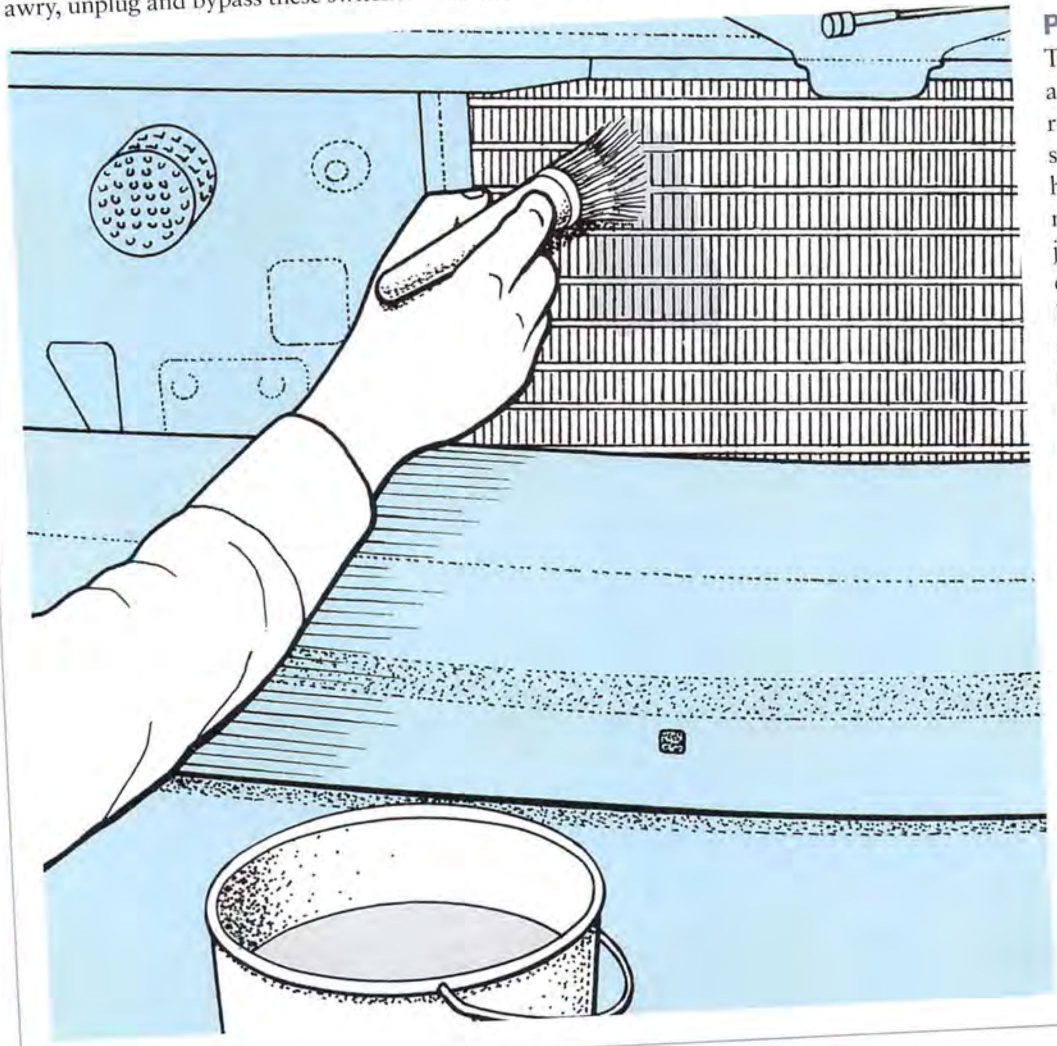

(Fig. 3) Use a soft brush and detergent to clean leaves and dead bugs from the condenser.

If you've got a car that has either no grille or a tiny one, you've got an "underbody breather." Inspect any underbody airdams and covers. If they're damaged or missing, they could be responsible for reduced airflow and poor cooling. And on almost all cars, when the air conditioning is turned on, a radiator/condenser electric fan should also come on. If it doesn't, that also reduces airflow and a/c performance, particularly in slow-driving conditions.

Of course, if the airflow is normal, you'll want to make sure a partial loss of coolant is not responsible for a drop in performance. Some older Japanese-made and some Chrysler systems (with R-12 refrigerant) have a sight glass—a tiny window into the refrigeration circuit. Check it to see if there are any bubbles when the clutch is engaged (bubbling when it's disengaged or in the first few seconds after engagement is normal). Bubbles with the clutch engaged may—repeat, may—indicate a low charge (Fig. 4). Also, check operating pressures with a pressure gauge and compare them with factory specifications. You must use the specs because pressures vary widely with ambient temperatures.

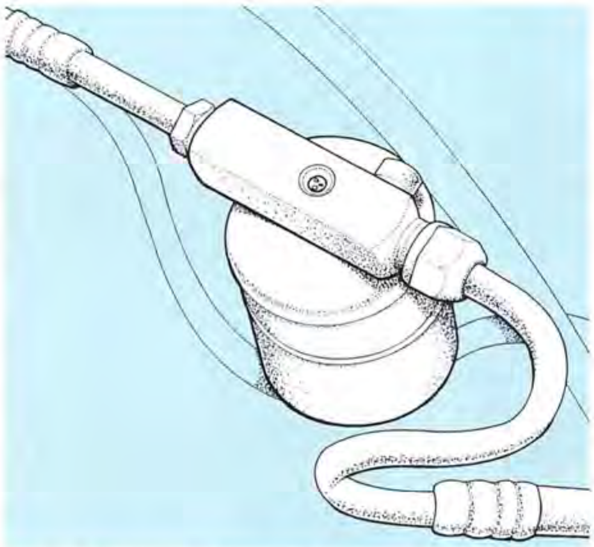

(Fig. 4) Bubbles in the sight glass may mean the air conditioner is undercharged. Have a professional (using special equipment) recharge.

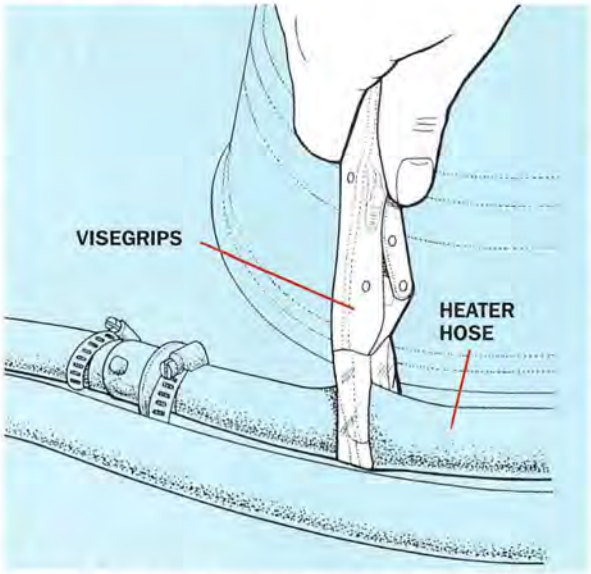

(Fig. 5) Clamp off the heater hose with padded Vise-Grip pliers to check for a leaky flap.

As explained earlier, a power-steering switch may misbehave, signaling "high pressure" with just a slight rise in pressure, thus killing the a/c. If your system runs fine with the car parked, but stops as soon as you pull away, either a bad power-steering switch or a misbehaving throttle-position sensor (TPS) is possible. The TPS is supposed to signal "kill the a/c" on full throttle to provide improved passing performance, but the sensor may come in with this signal when you're just barely touching the gas pedal. Or if the air-conditioning system has an evaporator-temperature sensor, take a careful ohmmeter reading and convert it to degrees by consulting the table in the manual. You may find that the sensor thinks it's cold when it's really not.

If the air blowing in your face is cool, but there isn't enough air blowing, operate the blower switch and see if the airflow picks up. A non-running blower fan and a slow-running blower motor are purely electrical problems, and are as likely to be responsible for poor system performance as anything. If the blower simply won't reach high speed, check the circuit diagram for a high-speed blower relay. You might be able to fix a bad connection by simply plugging in the wiring connector properly. Or you may find a broken wire by probing the wiring connector with a test light.

Another common cause of poor performance: the heater core's high temperature is bleeding into what should be purely chilled air from the evaporator. Some heat transfer is normal if you move the temperature lever or turn the knob from the max cold position. However, at max cold, there should be no hot-air bleed.

Clamp off a heater hose to block the hot-coolant flow (Fig. 5). If a/c performance improves considerably, one of the following could be occurring:

- A heater coolant-flow valve (used on many but not all vehicles) is stuck in the open position.
- The flap door that's supposed to regulate the airflow through the heater or evaporator isn't operating properly. If it's a manual, cable-controlled type with a lever, try slamming the lever against the max cold stop. If you don't hear a flap door hit a stop inside the underdash case, the cable may need adjustment. With an electric motor-driven design, check the case for a bad wiring connection at the motor.

And, of course, there may be cold air, but in the wrong place. If the cold air is going to your feet or out of the defrost ducts—but not to your face—that's an air-distribution problem. Air distribution is controlled by operating flap doors inside the heater/air-conditioning case, and it's done either by vacuum or electricity. If it's vacuum-operated, look for a leak.

Maintaining Your Battery

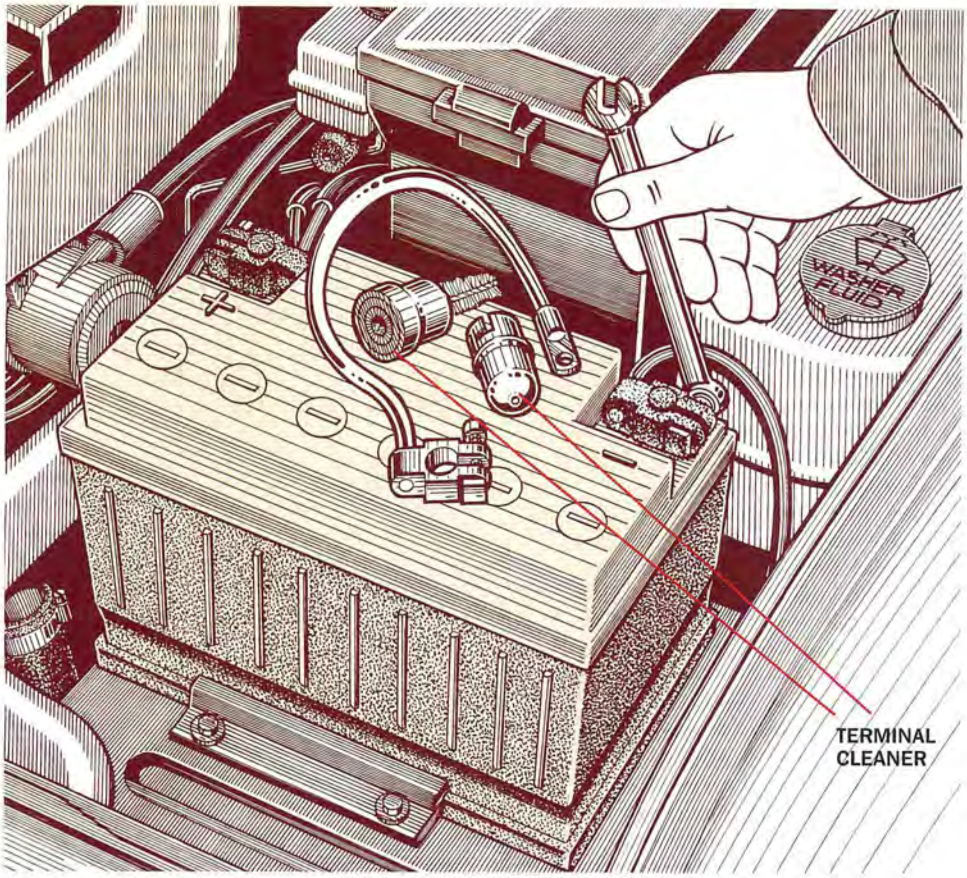

(Fig. 1) Always remove the negative connection from a battery first and reconnect it last to prevent shorting the wrench out on nearby metal objects.

It's the second time this week that you've had to jumpstart your car. Bad enough first thing in the morning, when you can use the battery in your RV to give you a start—but standing in a rainy Wal-Mart parking lot with your hood open and jumper cables in your dripping hands is just plain humiliating, not to mention uncomfortable and inconvenient.

The standard solution: Install a new battery—which means coughing up enough money to buy a new one and have it installed.

A few minutes of attention once or twice a year can perhaps double the life of your battery, saving you cold, hard cash. And, just as importantly, keep rain from running down your neck in shopping mall parking lots on stormy nights.

Would You Like Fries With That, Sir?

First, you need to determine if your battery is indeed fried or not. Other reasons for a no-start or barely start condition include, for openers, a marginal starter motor and high-resistance wiring in the primary (starter motor) circuit.

Open the hood. If your top-post battery terminals are covered with green fur that looks like it belongs in a bad sci-fi movie, you'll need to clean that up. Start by pouring some warm water over the terminals to dissolve the sulfation. This accretion is normal, but it can prevent good contact between the battery posts and the clamp, which in turn keeps your battery from providing enough voltage to start or prevents it from charging properly. Remove the terminals, and brighten up all the metal with a wire brush-style battery post terminal cleaner and a baking soda paste. Don't splash liquid around the engine compartment or onto the paint, and rinse everything off thoroughly. Side-terminal batteries rarely look worse than dull, but that's enough to cause hard starting. Wire-brush the contact surfaces to brighten them. (See page 80.)

Dirt on the top and sides of the battery, even greasy dirt, can hold enough moisture to create a current leak from the positive terminal directly to the negative terminal, making your battery self-discharge more rapidly. Leave the battery clean and dry.

Now take a close look at the terminal clamps on the cables. If they're eroded too badly to provide good mechanical contact, replace them.

Warning: Whenever you work on a car battery, always disconnect the negative terminal first—not the positive (**Fig. 1**). Why, you may ask? Simple: If you accidentally brush up against the fenderwell or any other grounded metal part of the car while you're unhooking the ground side, nothing will happen. But if the ground is still connected, and you're wrenching on the positive side and happen to touch any metal, you'll be holding onto a wrench that's suddenly conducting several hundred amperes. That's as much current as

HOW IT WORKS

Maintenance-Free Batteries

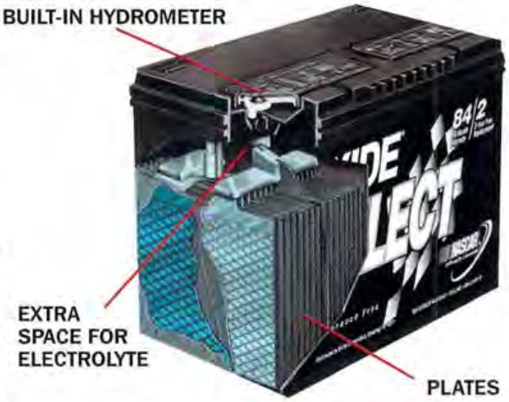

Maintenance-free batteries often can have their lives extended by proper attention to electrolyte level and good connections.

Like many things in life, the term "maintenance-free" is only partially true. Lead-acid batteries normally consume some of the water in their dilute sulfuric acid electrolyte during a normal charge-discharge cycle. It actually electrolyzes into hydrogen and oxygen and escapes as gas. So adding water periodically is necessary to keep the plates flooded, although it can't be done on some batteries. (See page 50) Maintenance-free batteries use a calcium alloy of lead instead of an antimony alloy, which reduces the amount of electrolysis. In addition, the amount of free-standing electrolyte above the plates is designed to be much higher in a new maintenance-free battery. This means that there's enough electrolyte to keep the plates covered even after a few seasons of normal use. So, during the battery's normal service life there should be no need to add water.

Any abnormal electrical system condition or high ambient temperatures may boil off more than the normal amount of water, however. Adding water where possible may extend the service life of these supposedly maintenance-free batteries.

an arc welder—which means a large spark, lots of heat and enough energy released to turn your wrench cherry red within a few seconds.

Avoid quickie aftermarket universal clamps that simply let you saw off the old clamp and clamp on the new one. While these have their uses (keep one in your on-car toolbox as a get-home-tonight expedient) they'll develop corrosion between the clamp and the cable in a few months and leave you stranded, again. If you feel up to it and have a soldering iron (probably propane-fired) capable of generating enough heat, you can solder a new clamp onto the existing cable. Use rosin-core electrical solder, not plumbing solder.

Don't have the Great Mother of All Soldering Irons? The simple way out is to replace the entire cable. If you're lucky, you can pick up an appropriate cable at the auto parts counter. Generally, all they stock are simple one-lead cables, in positive and ground. Yes, there is a significant difference between the terminals on the battery, and you really ought to use a black-insulated cable on the ground and a red-insulated cable on the positive or the post will fit the clamp poorly, which is how you got into this mess in the first place.

Unfortunately, it's not always that simple. Some battery clamps, both positive and negative, are soldered to more than a single wire. You have a choice here. The best solution is to go to the dealer and get the correct piece. That may be inconvenient, and will certainly be expensive. Also (see page 89), you may be able to get (at an auto parts store) a "cable repair section" (terminal attached to a section of cable). It's designed to attach to cable ends from which the battery connections and extra wires have been amputated. Or you can solder lugs of the appropriate size and gauge to the extra wires, and then attach the lugs to the clamp either by means of the clamp bolt or a large sheetmetal screw into the clamp body. Look for a clamp with an extra terminal if this is necessary, although if you've got side-terminal batteries this might be impossible. As a last resort, splice and solder the cables together ahead of the terminal. Be sure to use shrink tubing to cover the splice, especially if you're splicing the positive side.

When you reinstall the clamps, cover all exposed metal with petroleum jelly, white grease or my own personal favorite, Cop-Graf or Nevr-Seez brand antiseize compound. This will delay the inevitable growth of sulfur salts on the terminals by keeping moisture from collecting on the bare metal.

All Hooked Up

You've eliminated battery cables and clamps as the cause of your no-start or almost-start problem. And you figure

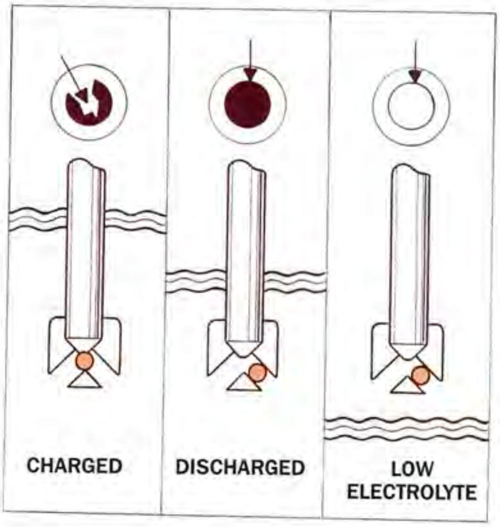

(Fig. 2) A built-in hydrometer eye is a good check, but won't necessarily find a bad battery.

it's not a problem with the starter motor, because a jumpstart instantly brings the engine whirring to life. So far, so good. Maybe it really is the battery.

Does the battery have a built-in hydrometer eye (**Fig. 2**)? This eye has a small plastic ball suspended at the bottom of a clear plastic window. The ball floats when the battery is charged, turning the window green. When the battery is discharged, the ball sinks and turns the window yellow. If the level of electrolyte is below the hydrometer, then the battery is past its useful life.

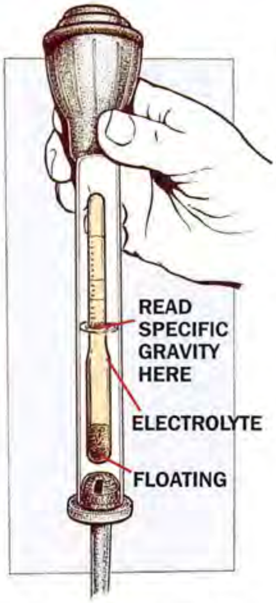

(Fig. 3) A proper battery hydrometer can detect an undercharged battery, as well as pinpoint a single bad cell.

One danger of these built-in hydrometers: They check only one of the six cells. If you've got a bad cell (the normal failure mode for batteries), there's only one chance in six you'll know it from the hydrometer.

A better check is to use a handheld hydrometer (**Fig. 3**). First you'll need to remove the filler caps. Battery clean? Cover the battery with a rag and pry the caps up. Warning! Battery acid will remove paint, corrode electrical connections and potentially destroy your eyesight. Wear eye protection.

First hitch: no battery caps to remove, it's a sealed, maintenance-free battery. Maybe so, but there's access to the individual cells, perhaps disguised under a label (**Fig. 4**).

Open it up and look inside. The electrolyte should come up to the bottom of the filler cap, or at least near it. If the level is down, especially if the plates are exposed, add water. Always use distilled water in a battery. There is never any need to add additional electrolyte. Batteries will normally consume water, but the sulfuric acid component of the electrolyte remains behind and never needs to be replenished.

Before you add any water, suck enough electrolyte out of one cell into the hydrometer to float the bulb. Write down the reading. Squirt the liquid back and repeat with the next cell. All six cells should have readings that agree within 0.050, or you've got a bad cell. These readings are temperature sensitive, so if it's very cold or hot out, you may need to correct the readings. The hydrometer scale equates the specific gravity of the electrolyte with the state of charge of the battery. If all the readings agree, and the gravity is low, the battery needs to be charged.

(Fig. 4) Battery filler caps may be concealed under a tab, a pry-up cover, or even under a large vinyl label.

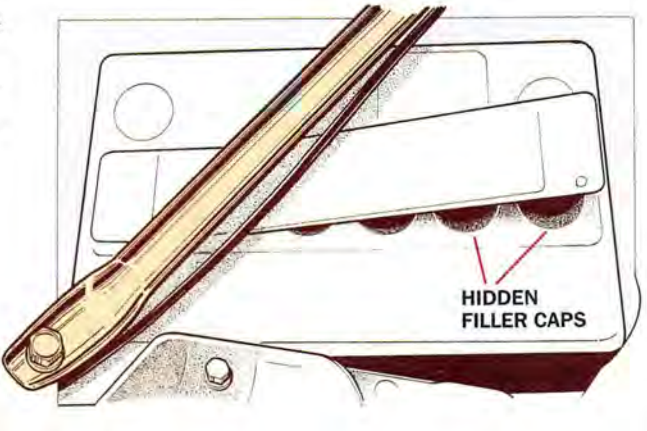

Charge

An alternative to the hydrometer test (and one used by professionals): With the battery charged, open circuit voltage should be 12.6 to 12.8 volts with the engine off and all loads off. With the car running and all loads off, the voltage should be between 13.6 and 14.5 volts.

Disable the engine by pulling the fuel pump fuse. Measure the battery voltage while cranking the engine. It should read at least 9.5 to 10 volts while cranking.

If the charging voltage is low, suspect some charging system problem. But if the charging voltage is correct and the cranking voltage is low, then the battery is suspect, as you've already troubleshot the wiring from the battery to the starter and ground.

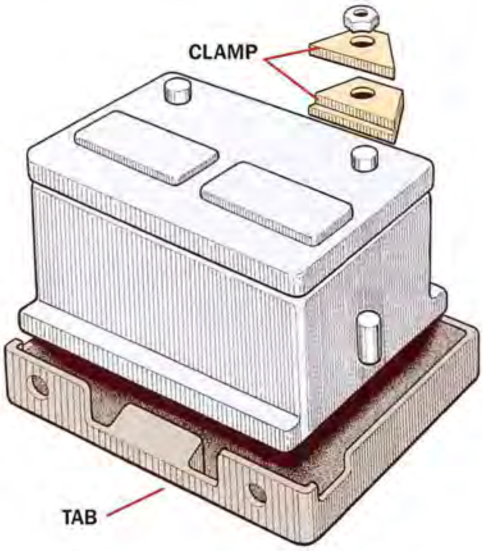

(Fig. 5) Battery hold-down clamps are essential to prevent a battery from becoming ballistic.

Strapping In

Replacing the battery is straightforward. It's vital that the hold-down clamps be reinstalled properly (**Fig. 5**). Some batteries use a wedge/lip arrangement at the bottom of the battery, while others (usually in older cars) use a pair of long bolts and a frame or lip to hold the battery down to the tray. A battery is perfectly capable of holing a radiator or smashing expensive fuel-injection components if it doesn't stay put.

Replacing a Battery

"Honey! There's a woodpecker in the car!" she says. It's too early in the morning for this, and far too cold for woodpeckers anyway. You politely suggest to your wife that she do two things immediately: Let you go back to sleep, and let the woodpecker fly away in his own good time. No luck, the kids have to go to hockey practice—now.

As you stumble outside, you realize the absurdity of it all. Now, you know that there is no woodpecker in your car, or anywhere nearby. The pecking sound your significant other has identified as avian in origin is really just your starter solenoid clacking away merrily as she tries to start the car. Sleepily, you jumpstart the recalcitrant conveyance, and then try to grab another 40 winks. After hockey practice, you'll have to find out why your night's rest was suddenly abbreviated.

Time Is On Our Side

The useful life span of an average automotive battery is four to seven years. With the auto companies' eternal quest to make cars and trucks more efficient, two things have conspired to shorten this life span: smaller batteries and higher underhood temperatures. Higher temperatures are so damaging that many cars have insulating blankets for the battery, or the battery is mounted in the trunk.

A single instance of deep-discharging a battery can reduce its life span by half, so avoid running a battery into deep discharge. If you do, then charge it properly as soon as possible before the plates inside build up a layer of sulfates. A properly charged battery won't freeze, but a discharged battery will freeze at temperatures well above zero. Any battery that's been frozen is finished.

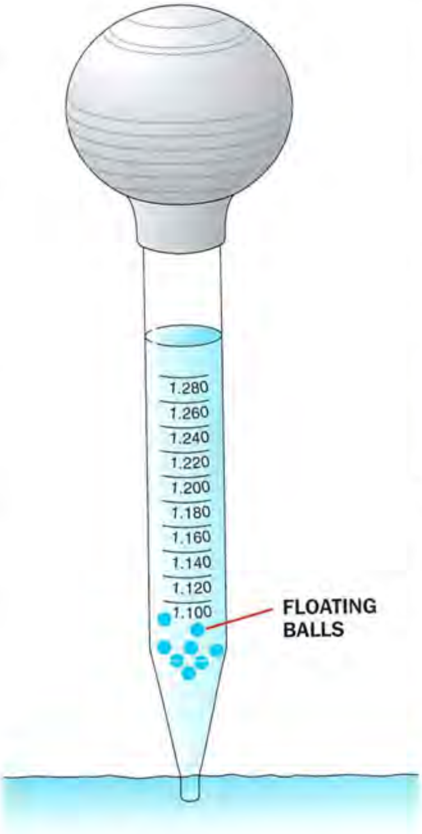

(Fig. 1) This battery hydrometer shows a fully discharged cell.

How Can You Tell?

Is one cell of the battery visibly different than the others? Look through the side of a translucent-cased battery. If one cell has substantially less electrolyte than the rest, or if one cell's lead plates are much lighter in color, then that cell has gone bad and the battery is junk. Charge the battery for an appropriate time (depending on how powerful your battery charger is) and then use a hydrometer (Fig 1) to measure the specific gravity of the electrolyte. The more charge on the battery, the more sulfuric acid there is in the electrolyte and the higher the hydrometer float will be on the scale. A battery that shows a low state of charge on the hydrometer after a long charge is dead. Don't rush to judgment, however, because a deeply discharged battery will take a charge very slowly at first.

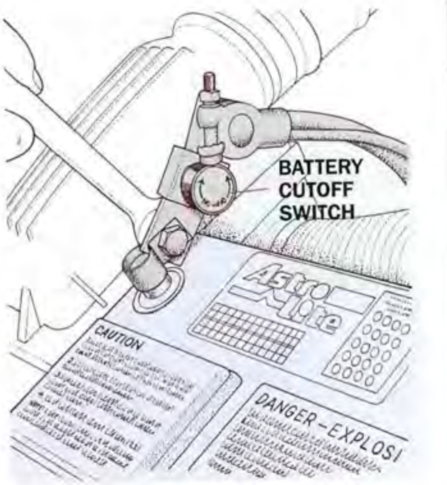

(Fig 2) Use a battery cutoff switch to diagnose parasitic current drains.

Is there some parasitic drain in your vehicle that's discharging your battery? An easy way to check is to temporarily install a battery cutoff switch (Fig 2) in the ground side. With everything in the car shut off, put an ammeter across the switch, turn the switch off, wait at least one hour, and measure the current drain. Up to 40 milliamperes (ma) is normal to keep the vehicle computers, radio presets and clock alive. Anything more than 200 ma will run your battery down in a few days.

Transplant

Once you've determined the battery is bad and the charging system and the rest of the electrical system is good, it's time to install a new battery. Start by rinsing the bad battery and the surrounding area with plenty of water. This will wash away any splattered battery acid, which otherwise would eat through sheetmetal, electrical wiring and the seat of your britches. Try not to splash it around.

Next, disconnect the battery's negative terminal. If you accidentally touch your wrench to anything nearby, no big deal. Then, you can unhook the positive side. If you try to unhook the positive side first, any random contact with nearby metal, such as the inner fender, will result in a welder-like shower of sparks and heat. Removing the ground side first breaks the circuit and prevents this.

Side-terminal battery hardware is notorious for not wanting to turn. Try tapping lightly with a hammer, and rotating the wire terminal slightly as you turn the wrench. Top-post battery terminals are sometimes very difficult to release. Light tapping may help, but you stand the chance of breaking the post off if you apply any substantial force. You may need to use a battery post tool to elevate the clamp loose from the post.

Nurse! Clamp!

Now's the time to seriously evaluate all of the wiring and clamps. This hardware carries substantial amounts of current—hundreds of amps for short periods when starting the car, and dozens for longer periods when the engine is running. It's imperative that it all be in good shape. The clamps should not be deformed. Light corrosion can be wire-brushed off, and a paste of baking soda will help neutralize any acid. Rinse thoroughly. If the attachment point of the wire to the metal clamp is good, fine. If the clamp or the connection is nasty-looking, you should replace the cable. Reserve the clamp-on, universal cable-end terminals for emergencies—they won't last as long as a proper connection and may contribute to yet another premature battery failure. Remember that the positive and negative posts on top-post batteries are physically, as well as electrically, different. Don't interchange. You may be able to rehabilitate an otherwise good terminal clamp by replacing the bolt—but be sure to use a lead-plated battery terminal bolt, not just something left over from installing your screen door.

How Dry I Am

Most conventional batteries are shipped without any electrolyte in them. If you have a good auto parts store, they'll keep a supply of ready-to-use batteries in assorted sizes on hand. If not, you'll have to have one filled and charged. Here's the procedure: Fill the dry battery with electrolyte of the correct specific gravity. Any place that can sell you a battery will have this. Fill to the bottom of the filler neck, no higher. Now wait several hours while the electrolyte seeps into the spongy lead plates. The level will go down, but as long as it covers all of the plates inside you can proceed to charging the battery. At this point the battery probably will start the car—but it's important

(Fig 3) A proper battery holddown is necessary for safety and long battery life.

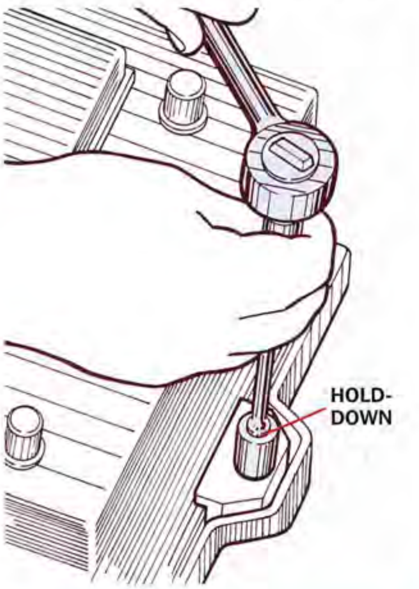

HOW IT WORKS

Built-In Battery Test Windows

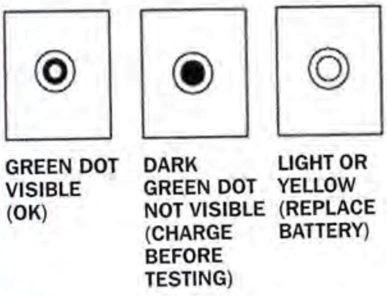
GREEN DOT VISIBLE (OK)
DARK GREEN DOT NOT VISIBLE (CHARGE BEFORE TESTING)
LIGHT OR YELLOW (REPLACE BATTERY)

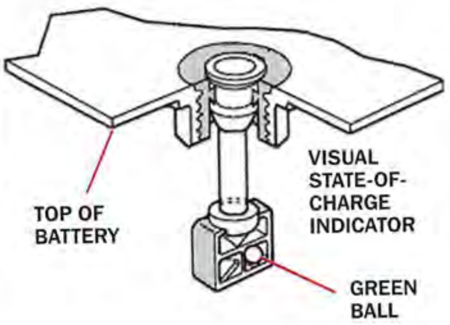

Many maintenance-free batteries have no access to the electrolyte to check the level or state of charge. But many of them have a built-in tester. This is a Plexiglas wand that sticks down into the battery just below the normal level of the electrolyte. At the bottom of this is a small chamber containing a loose plastic ball that floats in the dense electrolyte when the battery is charged. When it's floating, you can see the green ball at the top window through the Plexiglas. When the battery is discharged, the ball sinks and all you can see is the dark electrolyte. But when the electrolyte is too low, the plastic wand's bottom isn't submerged in it, and you'll notice a light color through the window. This means it's time to replace the battery. Remember, this tester sees into only one cell—there are five others, any one of which could be faulty.

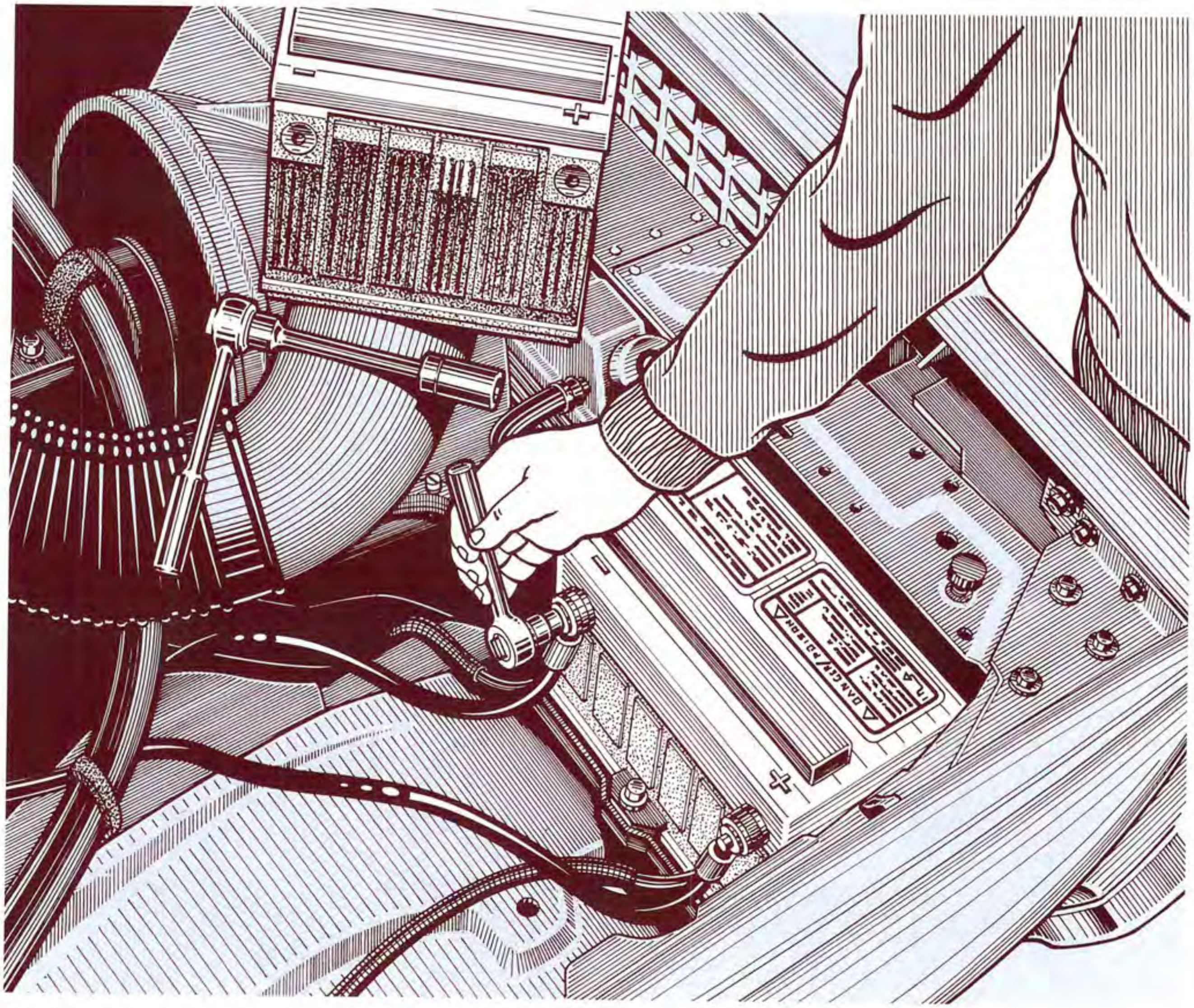

(Fig 4) Always reattach the battery's negative (–) terminal last.

that it be fast-charged for an hour or so. This means being charged at a higher current than your 4-amp household charger can muster, so it must be done with at least 1½ hours of super highway cruising, or by a shop with a fast charger capable of putting out 20 amps or so. Failing to do so will shorten the battery's life span and reduce its capacity. Now top up the electrolyte, and rinse well.

Dropping It In

You've got the correct battery, right? Check that the positive and negative terminals are on the proper sides, and that the holddown clamp is the correct style. You may as well fasten the

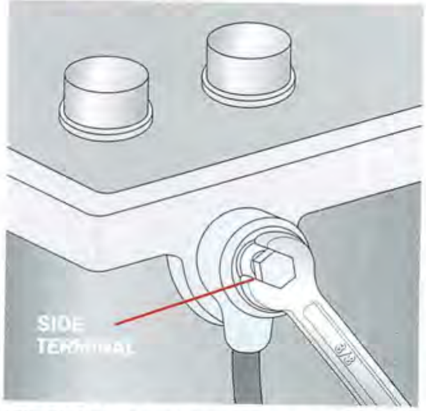

(Fig 5) Side-terminal battery posts are easy to overtighten—be careful.

holddown first (Fig 3). Use some anti-seize compound on the hardware in case you ever need to remove the battery in the future. Don't skip this step because vibration will shorten the battery's life by cracking the delicate lead plates away from the internal bus bars. The worst case would be the battery shooting out of the tray as you bottom out in some pothole.

Now hook up the positive terminal. When it's tight, put on the red rubber cover. Now hook up the negative (ground) side (Fig 4). Side-terminal batteries (Fig 5) have a very small bolt head for a reason. It's to keep overenthusiastic mechanics from overtightening them and stripping the threads out of the soft lead terminal. Use a short wrench to keep torque down.

Now cover the terminals with petroleum jelly. This will substantially slow down the buildup of corrosion on the terminals. Finally, be sure to replace the battery heat insulator if there was one installed.

MODERATE

Maintaining and Repairing Your Suspension

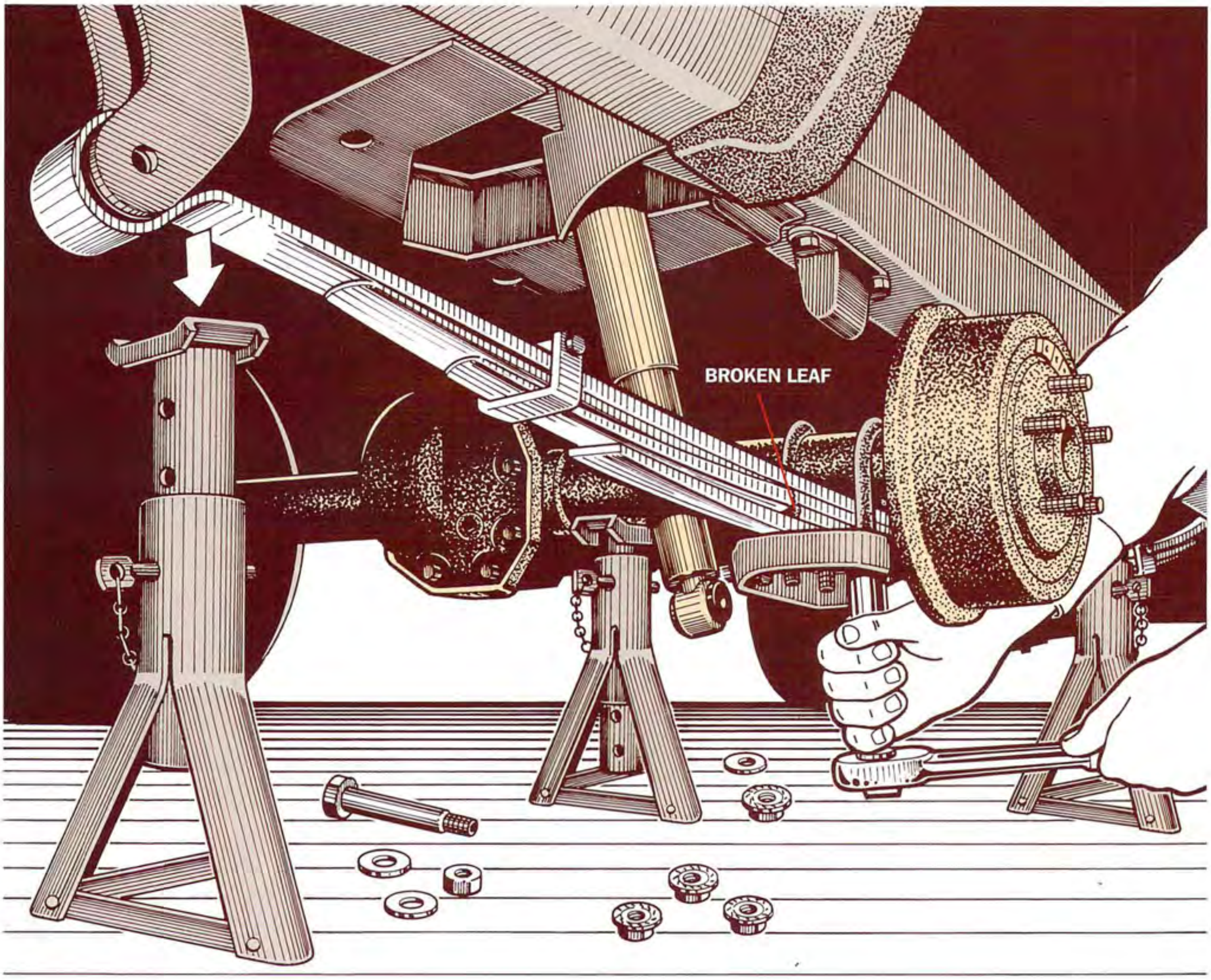

(Fig. 1) **Raise the vehicle on safety stands. Then support the axle itself to get the tension off the spring. Remove the center U-bolt nuts, washers, plate and then the U-bolts.**

"It feels like driving a truck," complains your significant other. Okay, she was the inspiration for "The Princess and the Pea," and you point out, "Well, it is a truck." But you know the truth: It used to ride and drive a lot more like a car, and that's why you bought it.

The carlike feel of today's pickups and sport utility vehicles is there when they're new. But most pickups and some SUVs have solid axles (at the rear and even the front of some four-wheel-drive models), and rear suspensions with leaf springs. The tuning is carefully balanced between an acceptable ride at no load (besides passengers) and something tolerable when the pickup bed is stacked high with 2 x 4s, paneling and decorative brick.

So it doesn't take a lot of deterioration to make the ride/drive experience something that even you have to admit is somewhat harsh.

You can get all kinds of assist springs for the rear axle, including auxiliary leaves—even air assists that you can deflate or inflate. But the object of these is to increase the load-carrying capacity of the vehicle.

What you want to do is restore the ride/handling of everyday or weekend trips after 20,000 to 30,000 miles or so have precipitated a harsher ride and vaguer handling. There are aftermarket shock absorbers made for pickups and SUVs used primarily as transportation, and they will help. If you want more, there are additional steps you can take to further restore the ride and handling.

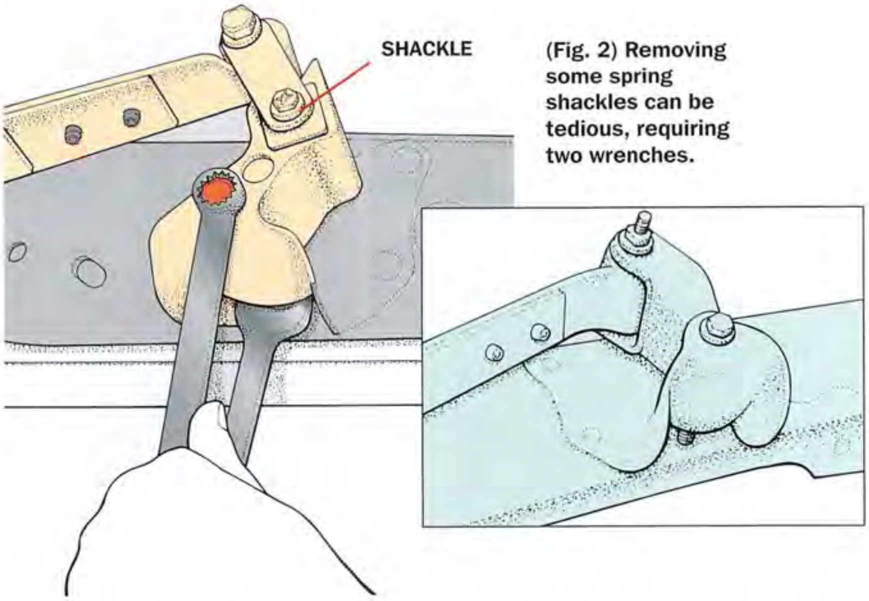

(Fig. 2) Removing some spring shackles can be tedious, requiring two wrenches.

Springs and Shackles

Look at the leaf springs—at the rear on each side, even at the front if your vehicle also has them there. Loose center U-bolts are an occasional problem, and retightening with a torque wrench could be helpful. But when the ride quality is down significantly, look closely for a cracked leaf (Fig. 1), and if you find one, change the spring, plus the mounting bushings and shackle in back. In fact, even with the leaves intact, there can be enough deterioration in the rear shackles to justify replacing both for a real ride improvement.

Take out the bolts that hold the spring assembly at front and rear, which can be easier said than done in many cases. Turning the bolts while holding the nuts with a second wrench is straightforward stuff (Fig. 2), but the rear bolts can almost "mate" with the rubber bushings inside the shackle, and getting them out can be a tedious job if that happens. You may have to slowly unthread the bolt, while at the same time prying under the bolt head to keep outward pressure on the bolt. Or you may be able to tap it out with a hammer and punch. Just keep clear of the top of the end of the leaf spring, which could snap up (against the frame) if there's still some tension on the spring.

On some trucks, the shackle's upper

HOW IT WORKS

The Leaf Spring Suspension

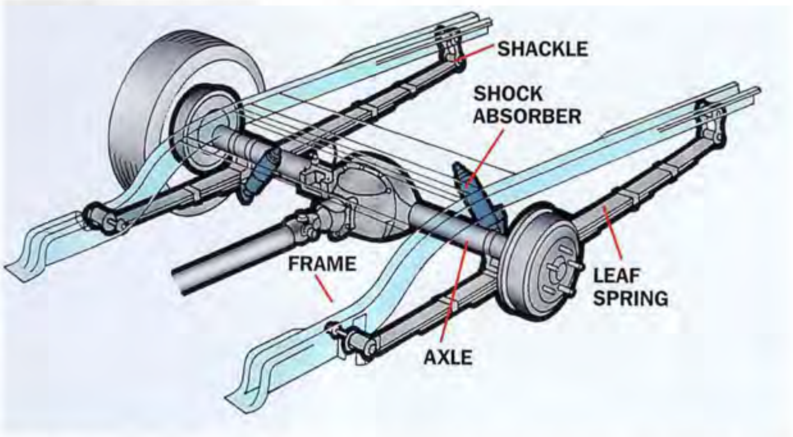

The leaf spring suspension is a popular choice for the rear of trucks and some sport utility vehicles. It has been used for some heavy-duty truck front suspensions and the rear of passenger cars, but the leading application is the truck/SUV with the nonindependent rear axle. The leaf spring has several leaves—simply adding leaves increases the load-carrying ability of the suspension. The top leaf typically is the longest and each end of that leaf is formed into an eye, into which a rubber bushing is installed. The spring eyes are bolted to the chassis in front and attached at the rear through a hinge joint called a shackle. The shackle permits the spring to effectively change its length as it flexes to absorb impacts. The leaf spring also attaches (through U-bolts) to the solid rear axle, so it locates the axle without the use of arms, an important function. This permits a simple suspension design with obvious packaging benefits. However, these advantages are offset in passenger cars and some SUVs by the superior ride qualities of the coil spring, which merely supports the vehicle and simply compresses and expands as it absorbs the impacts.

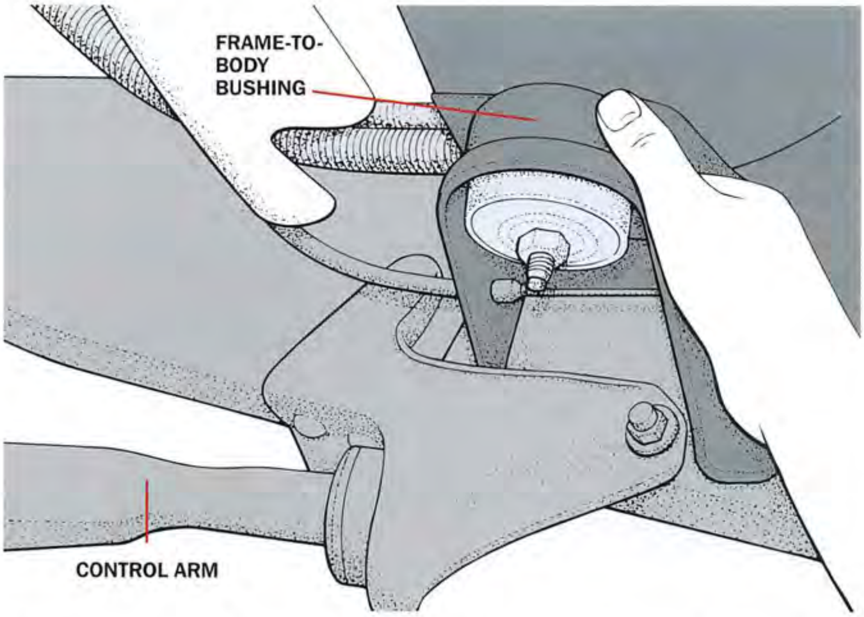

(Fig. 3) Check frame-to-body and suspension bushings. Look for cracks or overly brittle rubber as well as missing chunks.

bolt is installed on the inboard side and it won't clear the frame to come out. Jack up the rear of the vehicle until the shackle end of the leaf spring comes down far enough below the frame.

The original equipment bushings with those metal shackles are part of the reason for loss of ride quality. To get a long-term smoother ride, install aftermarket shackles that include small rubber springs with two metal sleeve inserts and metal arms—almost a reverse of the original equipment design. The Dana Velvet-Ride series that we installed is an example. You attach the metal arms to one metal sleeve of the rubber shackle, insert and tighten the through bolt and nut, and then install the new shackle in place of the original. The metal arms connect to the chassis and the rubber spring, and the rubber spring is bolted (through its second metal sleeve) to the leaf spring.

Inspect the underbody frame-to-body bushings (Fig. 3) and the radius rod bushings, and apply rubber lubricant as routine maintenance. If they're badly cracked, perhaps missing chunks of rubber, installing replacements will help eliminate body shake. Occasionally, the metal floor pan of the vehicle is the problem—it's rusted and weak. A piece of galvanized flat sheetmetal, thick enough for support, should be welded in place.

Even if you just have to replace bad bushings, don't simply loosen the bolts and pry to get clearance to install replacements, or you could cause damage you won't see. This work normally requires loosening radiator mounts, the steering gear and column, and other parts, so check and follow the factory-prescribed procedure for your vehicle. Or better still, lube periodically and the bushings will be fine.

Steering Stabilizers

Is a lot of vibration coming through the steering wheel? The problem could be wheel balance or alignment, even the steering gear's free play adjustment. But if it's a low-speed problem on secondary roads, look for something worn out.

If your operation has pounded the tightness out of steering parts, you can feel the looseness when you grab

(Fig. 4) Shimmy can be caused by a worn or loose steering stabilizer. Check the mounts, and check the damper for a bent rod or leakage.

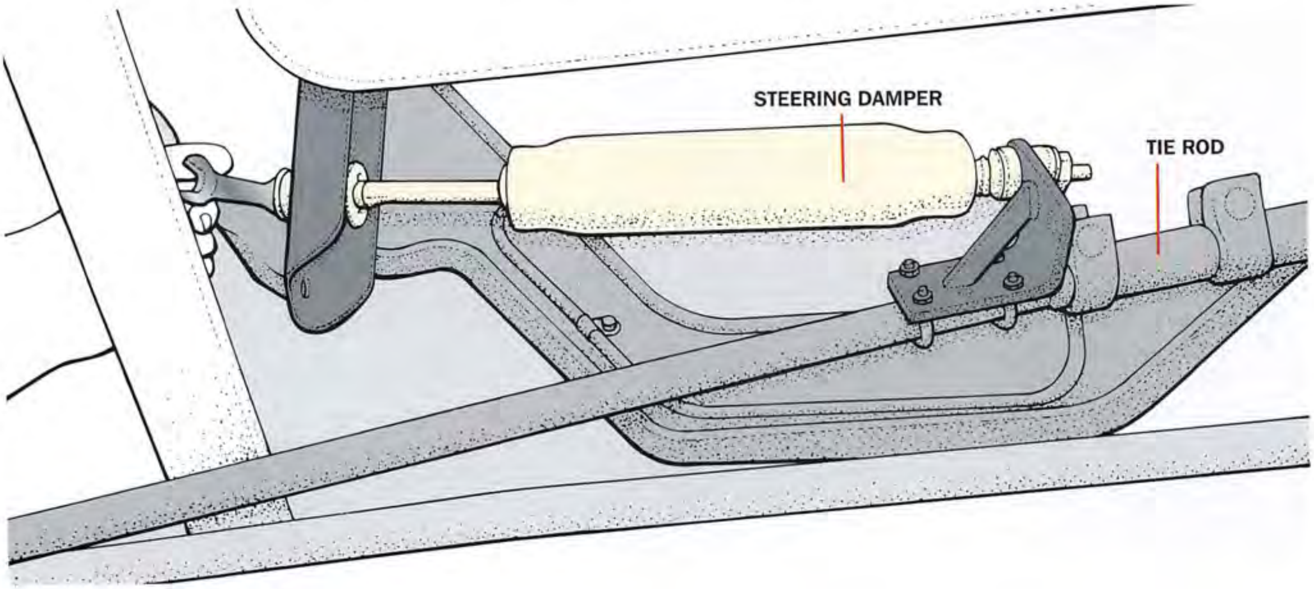

the tie rod and idler arms and find you can flex them too easily. If you want to replace them just this once, get heavy-duty parts specifically designed for pickups and SUVs. They'll not only be labeled for heavy-duty use, but they'll have grease fittings, which the original equipment may not.

Many vehicles have steering stabilizers. If the vehicle is suffering from vibration and shimmy, look for worn-out stabilizers, basically one or two horizontally installed "dampers" (shock absorbers) that connect the steering linkage to the frame or axle (**Fig. 4**).

The dampers mount to the steering linkage with a bracket (held by U-bolts) and to the frame or front axle with another bracket, and there are washers and rubber bushings at each end. If they're loose but the bushings look good, a simple tightening may remedy the problem. When they have significant mileage on them, replace the dampers along with mounting hardware and bushings.

If you don't have stabilizers, but use the vehicle on bad roads a lot and want to get rid of the shimmy, there are kits you can install, complete with frame brackets and U-bolts. These kits have been a factory fix for otherwise unfixable shimmy for years, and aftermarket kits for trucks and SUVs are available.

Jounce Bumpers

Have the rear axle jounce bumpers been taking a beating? If they have, you may see that the ends are scuffed. However, you probably know without even looking whether or not the vehicle has been bottoming. If you get stiffer shocks, your normal ride will suffer. A better approach is to replace the jounce bumpers with ones that really are small rubber assist springs, such as Monroe's Muscle LSE. They look like a short stack of thick pancakes, and when you're driving on bad roads they make contact with the jounce pad on the axle and prevent bottoming out. You could perceive this as causing a slight stiffening of the ride, but the bottoms of the "short stacks" have a couple of inches of clearance to the pads, so they're only in the picture on bumps. And because they're springs, they absorb impact so the overall ride quality should be better.

On many trucks, you can just unbolt the bracket that holds the original jounce bumper, bolt on a universal bracket (with elongated holes) and attach the "short stack." If the jounce bumper bracket is riveted to the frame, use a chisel or hand grinder to cut it loose.

After installing the bracket and rubber spring, tighten the bracket bolts lightly, then lower the vehicle to the ground, adjust the bracket height to the recommended clearance with a normal load in the truck bed, and do a final tightening.

CHECK SHOCK MOUNTS and bushings before condemning any shock absorbers.

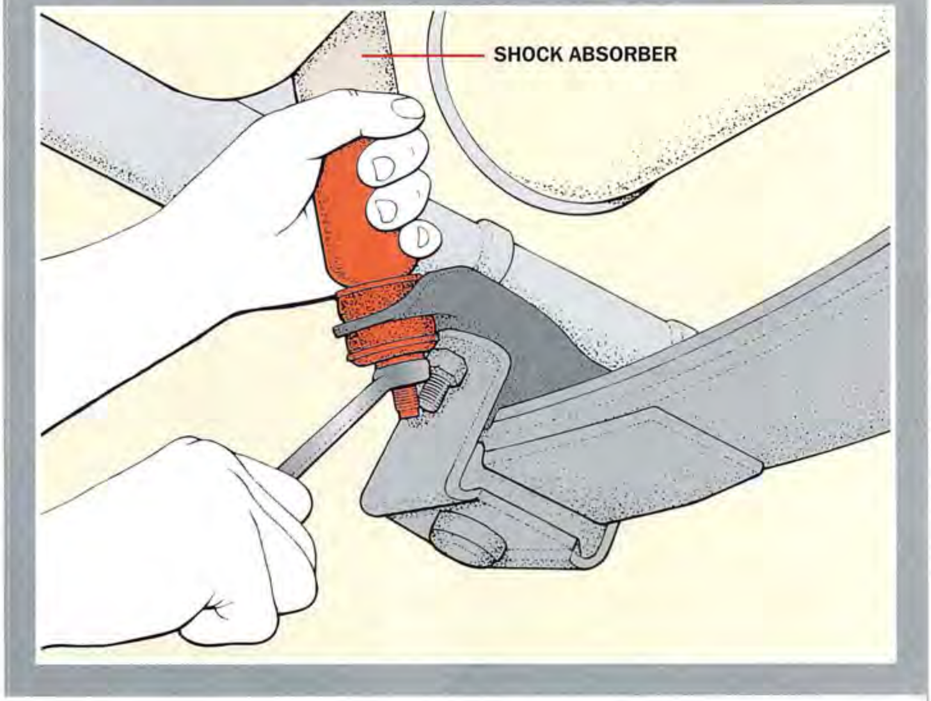

Replacing Shock Absorbers

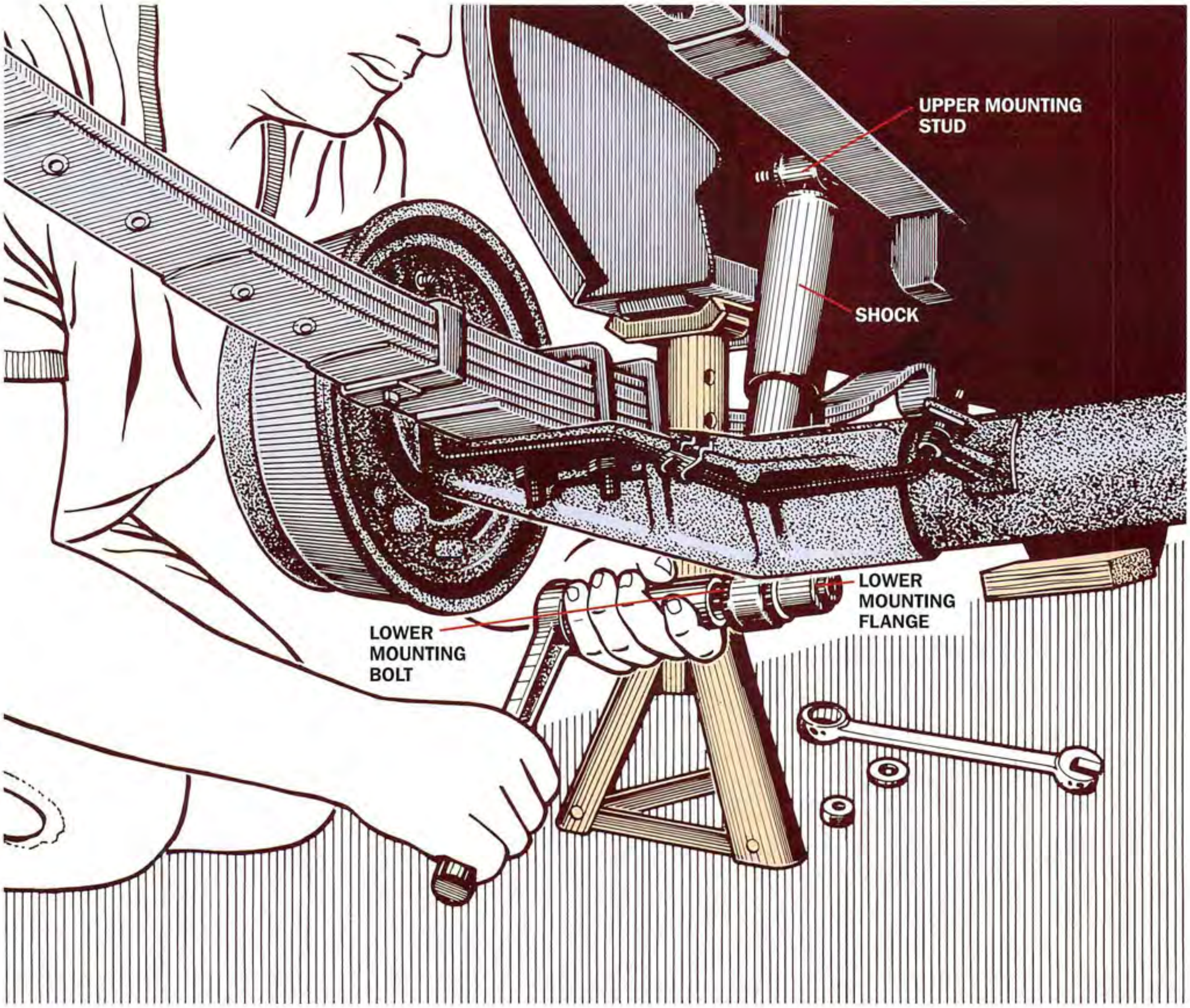

(Fig. 1) Removing shocks is usually straightforward, but getting under the vehicle and applying sufficient torque to remove large-diameter rusted-on fasteners can be daunting.

Your wife has started to sound like a detergent commercial, bragging about the spilled coffee stains she's gotten out of your shirts. The kids are making excuses about why they'd rather ride their bikes to the mall than get a lift in your car, and even the dog has started to get carsick. Wake up, Bunky, and realize that your vehicle's ride has gone, literally and figuratively, downhill. Hey, there are over 60,000 miles on the odometer, and the tires are starting to develop little concave "cups," sure signs it's time for new shocks.

Of course, you could get the jouncy ride of worn shocks at much lower mileage, perhaps because you drive briskly on bad roads and accelerate the wearing-out process. Or maybe the shock mountings have loosened. Occasionally, a shock produces a visible leak of fluid, but some hint of weep is normal, and most shocks we've seen that were well worn out, also looked bone dry.

That's right, we're talking about shock absorbers. The MacPherson strut is in wide use, but conventional shocks are on lots of cars, plus almost all those SUVs, pickups and at least the rear suspension of minivans. Isn't replacing them just like changing struts, just easier because you don't have to compress and remove the coil spring? No, although some things are similar (a strut is really a type of shock absorber), the mountings—and therefore the replacement procedure—are different.

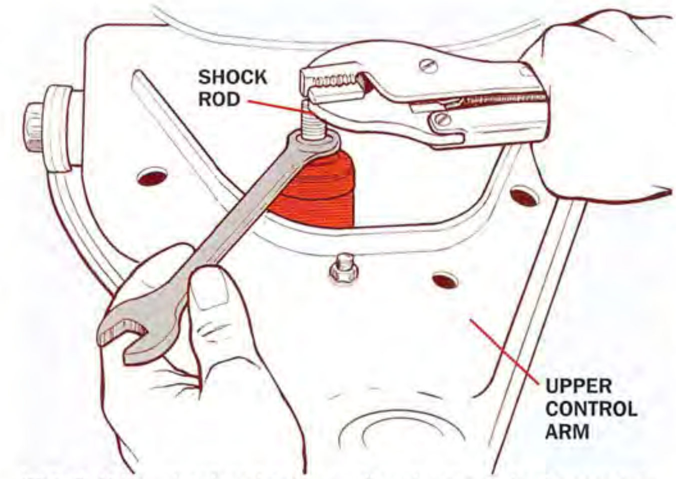

(Fig. 2) If the shock rod spins endlessly as you try to remove the upper mount, try Vise-Grips on the flatted end of the rod.

Those Mountings

Both shock mountings on most late-model vehicles are in the underbody, so if you support the chassis on safety stands, you have access to the top and bottom. One problem is reaching them. Another is getting off the hardware (**Fig. 1**).

The first thing to do is see if the mountings are loose. If you can turn the mounting nut or bolt at the top or bottom without much effort, the mounting is loose. On some shocks there's a stud in a rubber bushing, and if it tears loose, you'll see and feel it turn as you try to tighten the nut. There's typically a hex on the stud, so you can hold the stud while you tighten the nut.

When a nut and bolt loosen, you may be able to just retighten. However, it will loosen again if you don't apply some thread-locking compound to the threads.

If the shock bushings' rubber is obviously cracked or badly distorted, you'll usually have to replace the shocks to get the new bushings. That's okay, because the shocks themselves don't figure to be in much better shape.

Replacement Shocks

The aftermarket manufacturers produce replacement shocks for just about everything, except some electronically controlled designs. Only a handful of electronic systems' shocks are not available from third parties, either because they're an oddball size or because the tie-in to the vehicle electronics would result in a ride/handling glitch if you installed nonelectronic replacements. It might be as little as a trouble code or warning light, but if there's major integration,

HOW IT WORKS

Shock Absorbers Don't Absorb Shocks

A shock absorber doesn't absorb shock (the spring does that, by flexing to absorb the energy of an impact). The shock absorber actually stops the vehicle from bouncing up and down on the flexing spring. It does this by transferring the spring-flexing energy to a piston in an oil-filled chamber, which dissipates it in the form of heat. Most shocks have twin-tube (chamber) arrangements, an inner chamber with the piston and a calibrated valve at the bottom, which has the entry passage to an outer chamber—the reservoir. When the piston comes down on compression, it forces fluid through the calibrated valve into the reservoir. When the piston moves up as the shock absorber extends, oil is drawn from the reservoir into the main chamber, and some also flows through orifices in the piston, as part of the ride control calibration. A charge of low-pressure gas in the reservoir reduces oil foaming, which would affect ride control. Many high-performance shocks, including some for heavy loads such as motorhomes, are of single-tube construction, with a high-pressure gas charge in the base. Fluid flow is between the areas below and above the piston, which has a sophisticated valve assembly. The higher pressure of the single-tube design is even more effective against fluid aeration, but at the expense of ride comfort. Strut replacement is sometimes a job for a pro with the right tools and coil spring compressors.

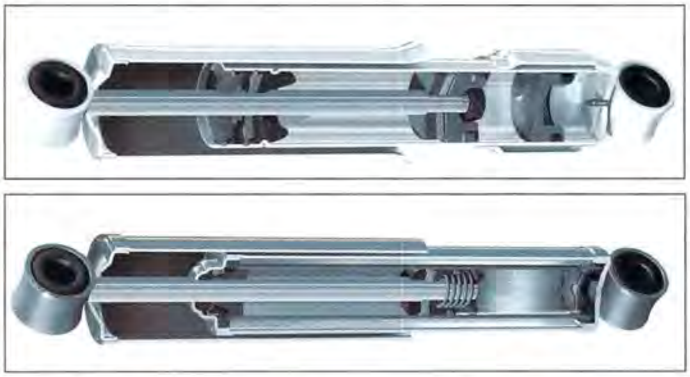

Single- and double-tube shocks are used on either the front or rear end of most vehicles. Struts are simply a shock that includes the spring perch and a bearing that replaces the upper ball joint.

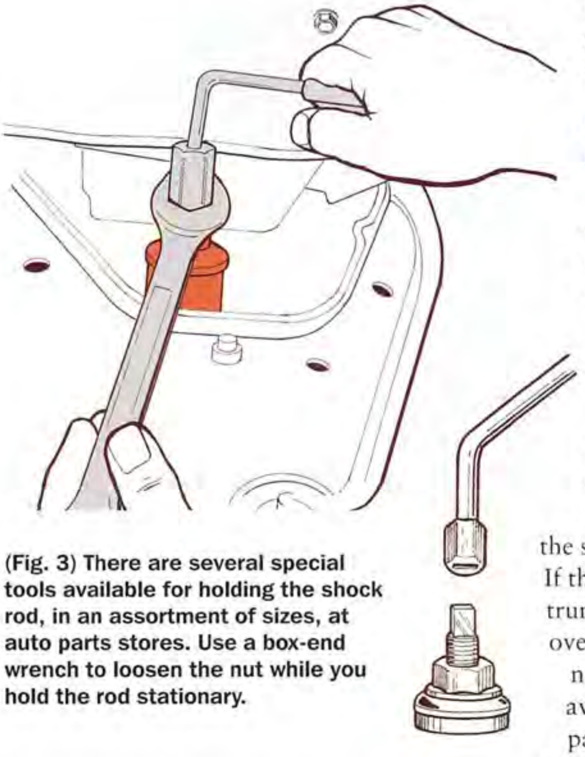

(Fig. 3) **There are several special tools available for holding the shock rod, in an assortment of sizes, at auto parts stores. Use a box-end wrench to loosen the nut while you hold the rod stationary.**

it could be more than that.

Just check the manufacturer's catalog, and if there are shock absorbers listed for your exact year, make and model, and the listing says it covers the electronic control system type, you're cool. You will be giving up the electronic control, but there isn't a system that's all that sophisticated on anything but a few premium cars, such as Cadillac, Mercedes, Corvette and some other high-end sports cars.

If the aftermarket catalog doesn't list a replacement, you're stuck with ordering replacement shocks from a dealer. On some Ford products there's an external motor drive, and you can unplug it from the old shock and install it on the new one.

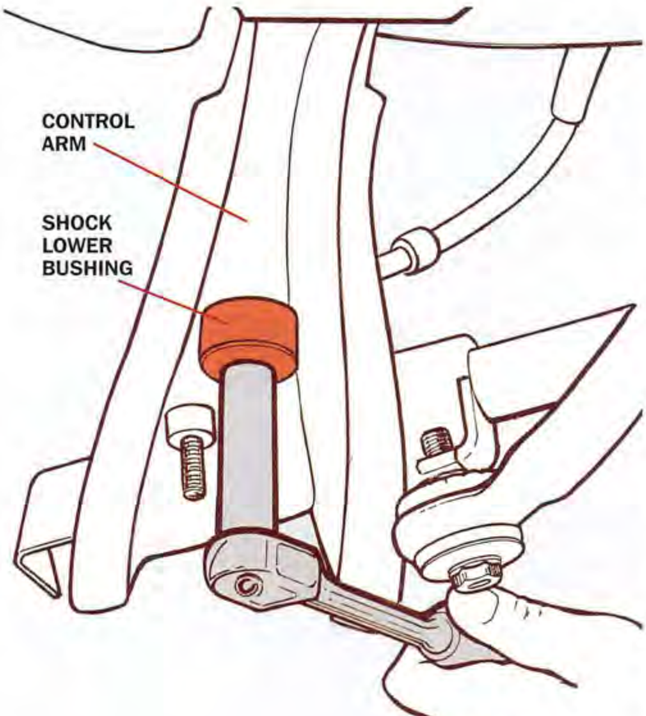

(Fig. 4) **Generally, a socket will spin off the mounting hardware. New hardware is usually included with the shock.**

In most other cases, you get a choice of shock absorber designs.

Getting the Shock Off

If there's an electrical connector, start by unplugging it. No electronics problem with installing conventional shocks? You could just cut the wiring, and tape securely. But if there's any chance you or a subsequent owner will want to re-establish the system, that won't work. If the connector is in the trunk, that's easy (just tape over it). But if it's underneath, find a safe location away from exhaust heat, pack it with pieces of plastic foam stuffed against the terminals, put it into a heavy plastic pouch, seal and hold it in place with a cable tie.

Loosening a rust-frozen nut and bolt is one of the major headaches in shock replacement. If a nut is exposed, you should be able to loosen it with a nut splitter, a clamp-like tool you often can rent from an auto parts store. No space for the splitter? Penetrating solvent, followed by heat from a torch, is another possibility. The most difficult is the type where the piston rod turns when you try to loosen the nut. One approach is to clamp locking pliers onto the "double-D" end of the piston rod, hold it from turning with the pliers and loosen the nut with a conventional wrench (Fig. 2). Something better (particularly for recessed installations) is a special toolkit with a hollow hex that fits over the rod onto the nut, followed by a wrench (made to fit the double-D) that goes through the hollow hex onto the double-D end (Fig. 3). Inexpensive kits, with three hollow hexes for most U.S. makes, or even some for specific models, are readily available in auto parts stores. It comes with three hollow hexes for most U.S. makes. There's also one specifically for some Ford products. If the shock piston end (or stud end) has a conventional hex, you can use an ordinary socket, of course (Fig. 4).

Many shocks are mounted on studs, and we've even seen those where the shock upper mount has a retaining bracket held by an additional stud and bolt. It's not a major difficulty, just extra hardware to remove. When you get the nuts off a stud mounting, you still have to pry off the shock, working evenly at top and bottom so it doesn't cock.

If you're not sure of the condition of a bone-dry shock, just disconnect it from the bottom mount if possible (or remove it completely if it's on studs). Then slowly compress the shock (even low-pressure gas shocks will be easily

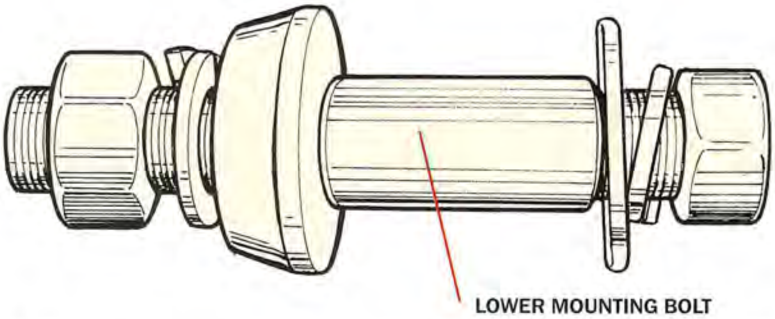

(Fig. 5) Lower mounting bolts are usually, but not always, included in the new shock's box.

compressed). If you feel any lost motion ("looseness" or unevenness), the shock is worn out.

Installing New Shocks

If you're on a budget and installing nongas shocks, it's a good idea to make sure they don't have air inside and bleed it out if they do. (Don't try this on old shocks. If they have air inside, it's from internal wear, and bleeding is not a cure.) With new nongas shocks, bleed by holding each shock upright (installed position): Extend it, and then turn it upside down and compress. Repeat the procedure a few times and the new shock should operate smoothly, with somewhat greater resistance on extension.

Using adjustable shocks? Some have soft, normal and firm positions, and although you can make adjustments after installation, start with soft or normal (firm is best for "ride restoration" after tens of thousands of miles of use).

The mounts for replacement shocks often are very different from those for the originals. That's okay, so long as you follow the instructions carefully. A common example is the aftermarket hardware for the "stud-in-bushing" design used on many General Motors cars. There's a specially shaped steel replacement stud assembly that goes into the lower mounting, and the bushing that's in the replacement shock just slides onto it (**Fig. 5**).

If the shock mounts on studs and the nut is tough to thread on, clean the threads with a wire brush and apply penetrating oil/solvent. If the nut almost spun off, apply a film of thread-locking compound to the stud. If you can tighten the shock hardware to specs without distorting the bushing, that's ideal. If the nut hits the end of the threads and the bushings are loose, you've got the wrong ones—or you're missing some hardware.

Maintaining Your Exhaust System

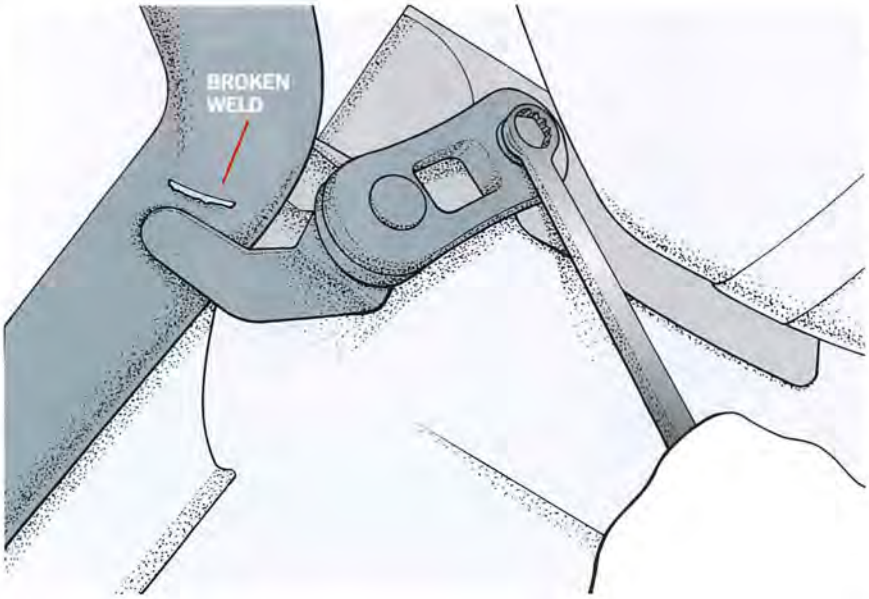

(Fig. 1) Tightening a loose clamp won't help if the bracket has broken loose at the weld line.

The sign says "DIP." And it's not advertising a samba school—there's a dip in the pavement large enough to lose a first-grader in. You touch the brakes lightly to keep from spilling your java as you traverse it. But there's a sudden graunching noise from under your car, followed by the scream as about a gallon of fast-food coffee flenses the skin from your thigh. You pull over to inspect your leg and the underside of your car, in that order. There are big, hot chunks of metal hanging loose under your chassis, and the exhaust system is nearly dragging.

Today's exhaust systems consist of a lot of expensive parts. There are one or more catalysts, an oxygen sensor, maybe even a 2-stage muffler. There's piping that's not only shaped to clear underbody lines, but also may be made of premium metals, in some cases dual-wall tubing. There are sheetmetal underbody heat shields to prevent the exhaust from igniting dry grass.

What's holding everything together? A few clamp joints and some welds. What's holding everything up? A few pieces of rubber with some brackets. What's keeping everything aligned? Those same few pieces of rubber and brackets. Failures are common and they range from cracked rubber to failed welds.

Before you suffer the same expensive fate we described earlier, take time to perform a regular underbody inspection—it's easy. To get started, jack up the car and support it on safety stands front and rear. Now you're ready for an in-depth look.

Hangers

If any hanger is broken, it has to be replaced, without delay—even if the exhaust system seems to be hanging level. One broken hanger means that when the system moves up and down as you drive, a whiplash effect goes through the clamped and welded joints. Eventually clamps loosen and welds crack.

Replacing a broken hanger used to be simple, because most companies used some type that bolted to the body and clamped to some part of the exhaust system. Today, the most popular design uses rods welded to the pipes and, often, to the vehicle underbody (**Fig. 1**). These hangers also have a thick rubber section that resembles a flattened football or beach ball, with holes for a rod at each apex of the rubber. Welding positively locates all the parts, and although you may have to tug and pry, replacing the rubber piece when it becomes cracked is very straightforward.

You can install a brand-new pipe if the rod comes off the pipe, but if the pipe itself is good, that's an unnecessarily tough and expensive replacement. The alternative is to install a universal hanger, though this may require a bit of jury-rigging. First, remove the broken hanger. Then look for a nearby hole, perhaps even from the old hanger. Take a universal hanger that can twist and tilt and has an adjustable length setup, and attach it to the underbody with a bolt (and if it isn't threaded, also with a nut) through that hole. Next, make the connection around the pipe—a simple clamp and U-bolt usually works (**Fig. 2**). Don't be surprised if a part of the welded-on bracket (from the broken hanger) has to be cut or ground off to provide space for the clamp that will attach to the universal hanger. You should try to make attachments very close to the locations of the original equipment setup. This will maintain system alignment and the balanced hang of the entire system.

If there is noise from exhaust system contact with the underbody or an underbody part, you should check for damage underneath. Exhaust system clearances are limited, and if you see the need to increase one, it may be possible to bend or shim an original-equipment hanger. If a rod-type hanger is welded in place, see if you can bend it with a piece of pipe over the rod's open end. If you need more than ¼ in. or so, look for an alternative. If the hanger is a bolt-on to the body, for example, you may be able to install washers as shims.

Some older imports have rubber hangers that resemble thick rubber bands. If the old band has broken off, you can get a replacement, but installing it is not a matter of simply stretching it over the retaining tabs. The band is so thick and stiff you probably will have to pry pretty hard to get it on.

Many exhaust systems have a bracket to provide firm support close to the transmission. If the bracket cracks—or if it was removed for service and never reinstalled—that can account for a lot of exhaust system flex and eventual cracking. If you can't get a replacement bracket, you may be able to get the old cracked one rewelded. But replacement is the best choice, and if the dealer doesn't have it, the wrecking yard may.

Shields

Exhaust system heat shields are prone to damage from driving on rough roads. To avoid the possibility of a grass fire, replace any that are missing or barely hanging on. If a shield is dented, unbolt it, and check the exhaust system for damage.

Exhaust Joints

If there has been a lot of exhaust system up-down travel because of a broken hanger, the clamps and joint may be damaged. If a welded joint is cracked at a muffler or resonator joint, you can replace the parts. Or to save money, cut away the piping and part of the neck joint, then install a short connector pipe and secure it with clamps.

If a clamped joint isn't cracked, but is leaking exhaust gas, remove the old clamp. If the pipe ends are badly distorted, you'll have to rent a pipe expander to reshape them. Otherwise, separate the pipes, sand them to remove any rust, then apply a coat of

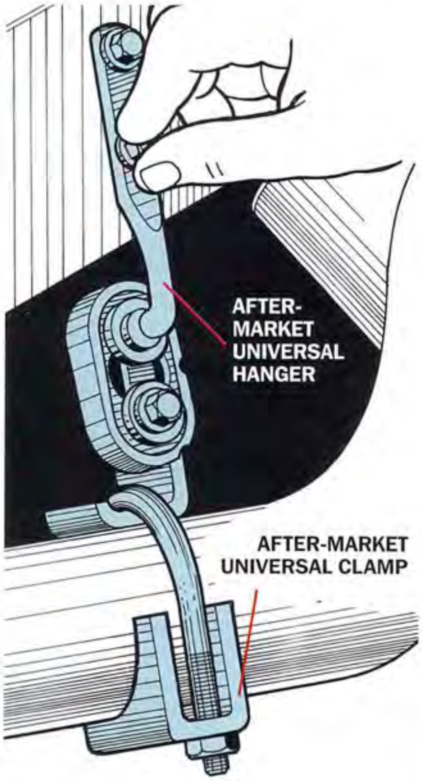

(Fig. 2) Sometimes a universal aftermarket clamp will fill the bill and let you reattach the OEM hanger.

HOW IT WORKS

Muffler

A muffler is a canister with internal tubular passages, shaped in part by baffles, designed to gradually reduce the pulsations of flowing exhaust gas. The slowdown creates back pressure in the exhaust system, all the way back to the engine itself. This slows down the exit of exhaust gas from the engine, reducing its performance. To get around this problem, many high-performance engines increase the exhaust flow with larger mufflers and accept some increase in exhaust system noise as a result. With careful design, however, including the use of silencing materials, the systems can be made to meet legal restrictions. Some back pressure is beneficial at low speed, because slowing down the flow pattern of intake and exhaust gases improves performance somewhat. Many engines have 2-stage mufflers with an internal spring-loaded valve. These mufflers maintain a specified amount of back pressure for good low-speed performance. But when the engine is at higher speed and there's more exhaust gas to flow through the muffler, back pressure builds up. Before it can become excessive, it pushes open the spring-loaded valve, and the gases take a far less restrictive flow through the muffler. This "second stage" permits the engine to develop considerably greater horsepower at higher rpm.

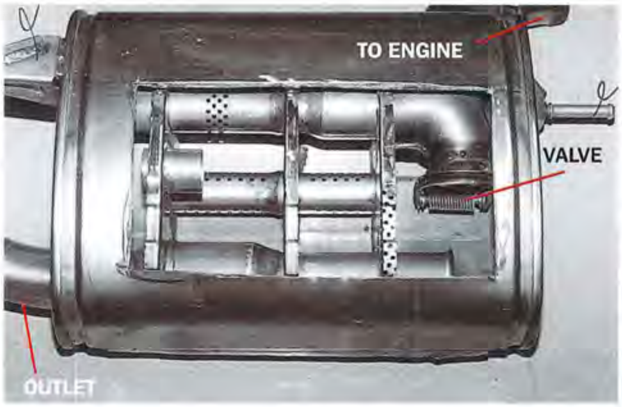

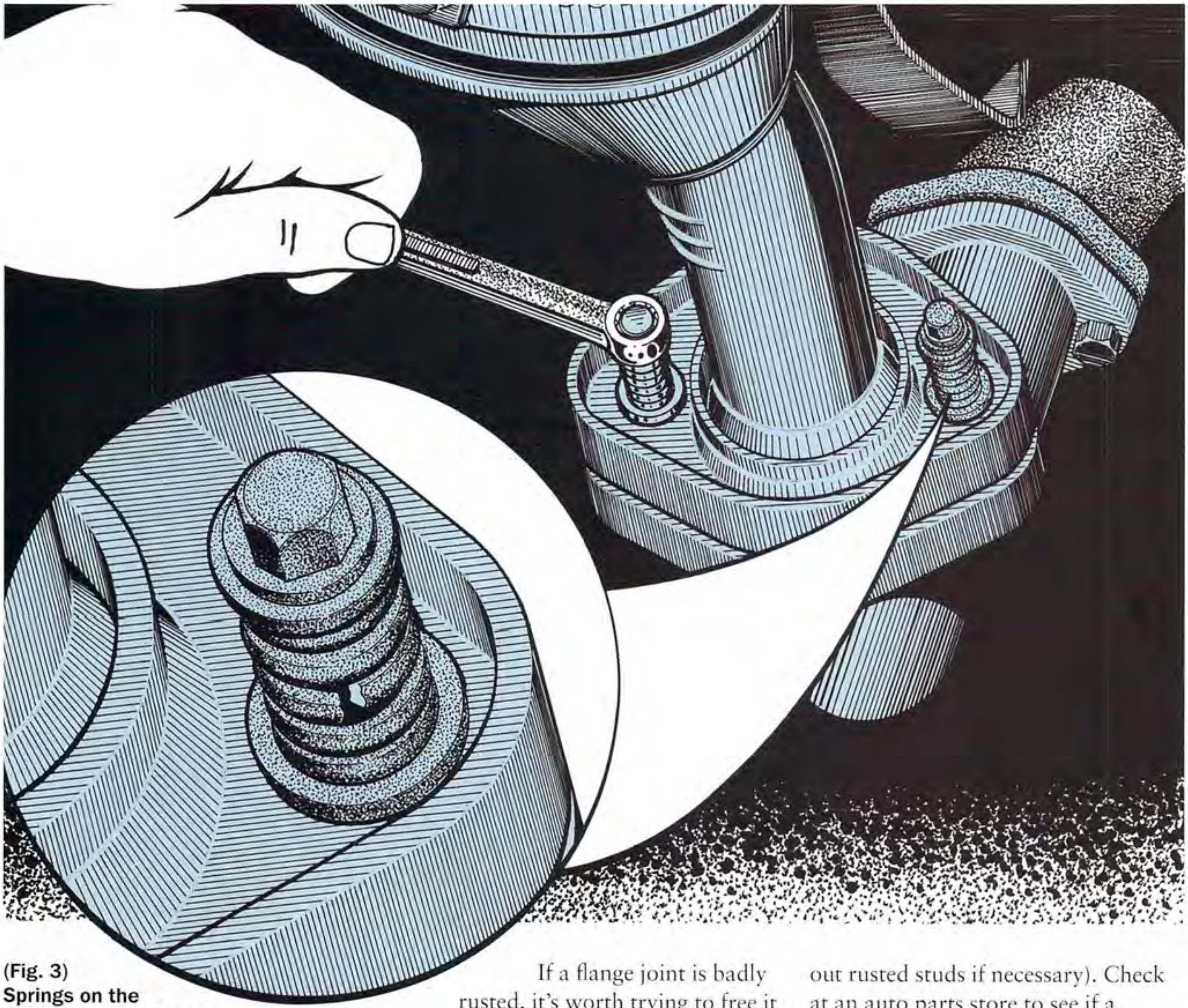

(Fig. 3) **Springs on the bolts at the joint at the base of the headpipe allow the exhaust system to flex slightly as the engine moves on its mounts. Replace broken or sacked-out springs to prevent a leak.**

exhaust pipe sealer. Reassemble and install a heavy-duty clamp, which can tolerate much more torque, to provide a leak-free joint.

Many systems have a flange joint at the exhaust manifold or at the manifold's Y-pipe. That joint is held together with spring-loaded bolts (Fig. 3). The heat may weaken the springs, and the joint will loosen. Try simply tightening the bolts, but if the springs have cracked or are weak, replace the bolts. They're sold, with new springs, in auto parts stores.

If a flange joint is badly rusted, it's worth trying to free it up with penetrating solvent. Then, take it apart and see if it can be salvaged by sanding it clean, installing a new gasket with sealer and reassembling with new bolts and nuts (drill out rusted studs if necessary). Check at an auto parts store to see if a clamp-over repair fitting is available.

Damper Weights

Some exhaust systems have vibration dampers (Fig. 4), which are weights that are attached to a pipe or are part of a flange connection. If a damper weight breaks off or is bent, the exhaust system will vibrate noticeably. This not only makes for an uncomfortable ride, but it can affect the life of welded and clamped joints.

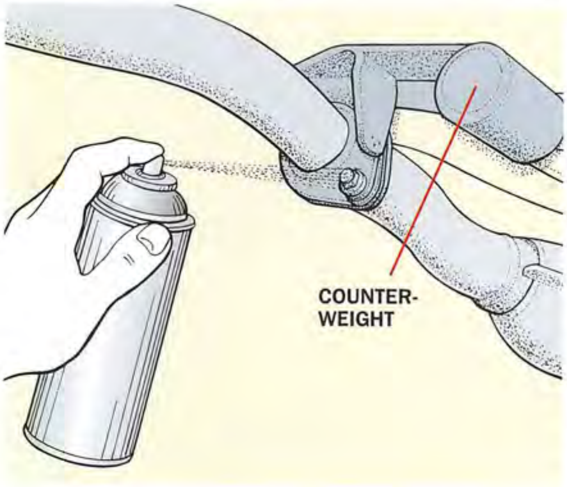

(Fig. 4) **When using replacement parts, be sure to reinstall any original vibration-damping counterweights.**

Replacing Your Belts

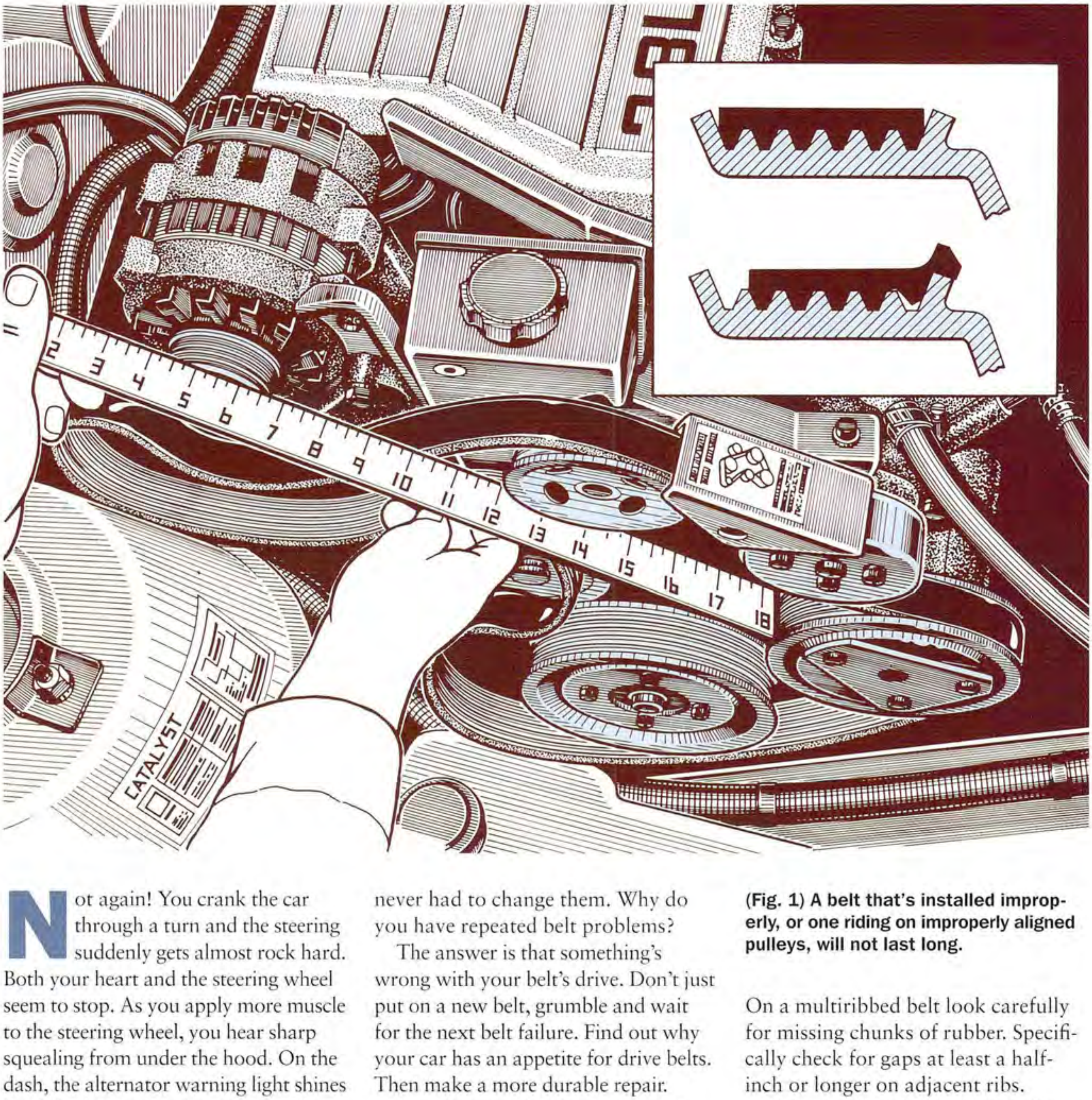

(Fig. 1) A belt that's installed improperly, or one riding on improperly aligned pulleys, will not last long.

Not again! You crank the car through a turn and the steering suddenly gets almost rock hard. Both your heart and the steering wheel seem to stop. As you apply more muscle to the steering wheel, you hear sharp squealing from under the hood. On the dash, the alternator warning light shines brightly. As your vital signs normalize, you remember that you've had this happen before: The accessory drive belt for the power-steering pump either is slipping badly or has popped off.

Today's drive belts, particularly the ribbed belts, are supposed to last for years. You even know people who've never had to change them. Why do you have repeated belt problems?

The answer is that something's wrong with your belt's drive. Don't just put on a new belt, grumble and wait for the next belt failure. Find out why your car has an appetite for drive belts. Then make a more durable repair.

First, inspect all of the belts and each of the pulleys **(Fig. 1)**. Twist over each belt and check it thoroughly. On a simple V-belt, look at the sidewalls, and if they're glazed, that's a sign of slippage—typically caused by improper adjustment, but possibly a sign of a bad pulley, too. On a multiribbed belt look carefully for missing chunks of rubber. Specifically check for gaps at least a half-inch or longer on adjacent ribs.

If the belt hasn't lasted long, the reason will be found during inspection of the pulleys. There are lots of reasons why some cars eat belts.

Feel The Tension

Most cars and trucks use a single, serpentine belt (see "How it Works,"

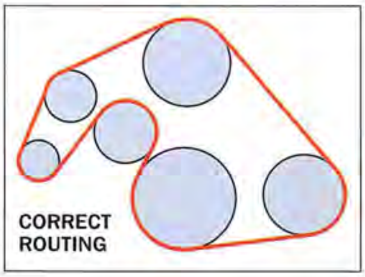

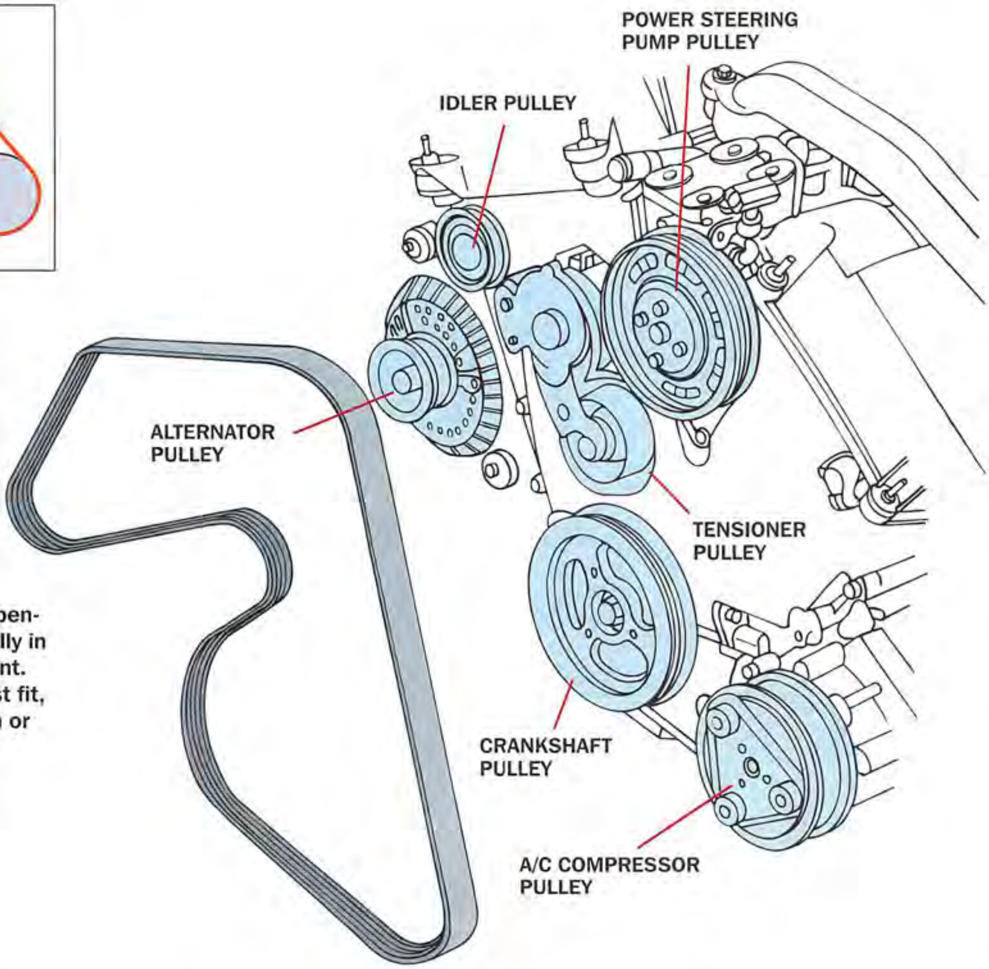

(Fig. 2) Correct routing of serpentine belts is not easy, especially in a crowded engine compartment. Incorrect routings may almost fit, so check the routing diagram or the owner's manual.

page 68) to drive a lot of accessories. These belts typically are tensioned by an idler pulley with an automatic tensioner. Just because there's an automatic tensioner doesn't mean it's working right. Any of the root causes discussed below can result in a belt getting chewed up fast, or popping off the pulleys.

There's a tension indicator on the idler and it should be reasonably close to midpoint. It should not be resting on the stop tab at the "loose" end. If it is at the stop—or very close—the belt has stretched and should be replaced before it pops off. You just put in a new belt, you say? Maybe you routed the belt incorrectly. Yes, it is possible to get the belt wrapped around all the pulleys in what seems to be the right way but have it wrong. Check the belt-routing diagram **(Fig. 2)**, which on many cars is on an underhood label.

Are you sure you have the right belt? Just because the tension indicator is in the right place doesn't mean you have the correct belt. It's hard to eyeball the difference between 7-rib and 8-rib belts, and if the pulleys have more or fewer grooves, the belt won't sit right. Even if the number of grooves matches, it's possible to install the belt too far inboard. While you're looking at the tensioner, also check for cracks in the housing which would allow it to flex enough for trouble.

Nothing obviously wrong yet? If the belt is still on the pulleys, run the engine and eyeball the accessory drive. If the tensioner is vibrating a lot but the pulleys are running smoothly, the damper bushing is bad and you'll have to install a new tensioner.

Put a wrench on the spring tensioner **(Fig. 3)**. Usually there's a square hole for a ratchet or breaker bar drive, and occasionally you can use one of those. If the tensioner on your transverse-engine car is buried so deep down between the pulleys and the suspension tower that you can barely see it, check out the assortment of special long, thin wrenches designed for those installations at your local auto parts store. The leading makers of these specialty tools will have theirs on display (Lisle Tools, K-D/Dannaher and Schley/SP). With the wrench on the tensioner, apply some light torque in the disconnect direction, and if the wrench moves very easily a short distance and then gets stiff, the tensioner housing spring either is out of position or broken. This condition often causes the tensioner and its idler pulley to vibrate.

Not today? Keep applying effort with the wrench, and you should feel uniform spring tension all the way to the belt-off position. If you feel binding or lost motion at any point, replace the tensioner. Never "let go" of a tensioner. Aside from the possible physical danger, the sudden release can cause internal parts of the tensioner to snap (including the spring). Release it very gradually. If you do need a new tensioner, don't think you have to get it from a dealer. Reputable aftermarket tensioners are available from automotive parts stores.

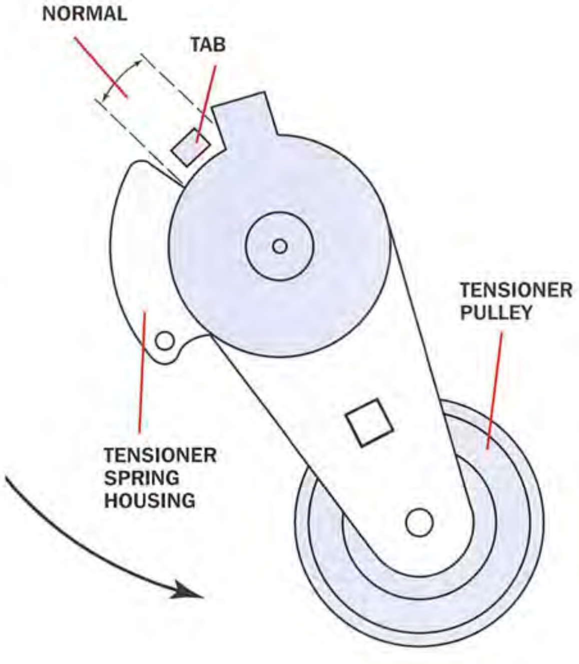

(Fig. 3) Most automatic tensioners provide a square hole for a ½-in. ratchet.

Playing With Pulleys

If there's wobble in the idler or any one of the other accessory pulleys, the pulley could be cracked or worn loose. Just one damaged pulley anywhere in the accessory drive also can affect the operating appearance of the tensioner. Replace the wobbling pulley—easier said than done on some accessories. Besides the problems of near-impossible access on a front-drive's engine, you may need to rent a special pulley-puller tool if you don't own one.

If the pulleys aren't wobbling but the belt isn't tracking well—seems to twist slightly or move in and out as it spins the accessories—there may be pulley misalignment. That is, one pulley is just too far forward or behind another. This happens on new cars when manufacturing tolerances are exceeded, but it doesn't just "develop" unless you did some front-of-engine work and failed to tighten all the accessory mountings carefully. It also could happen if an accessory recently was changed, and that accessory's pulley doesn't line up. (Did you just put in a new alternator? Power steering pump? Air-conditioning compressor? Water pump?)

If there's enough access room, lay a stiff mechanic's straightedge or a thin steel rod across the pulleys to help you decide if the pulley on that new accessory is misaligned. Unless you want to take back that accessory and check out another one, however, the only fix is to shim out the accessory with washers, which is a bit of cut-and-try.

ACCESSORY DRIVES wrap a long way around to provide adequate friction to prevent slippage.

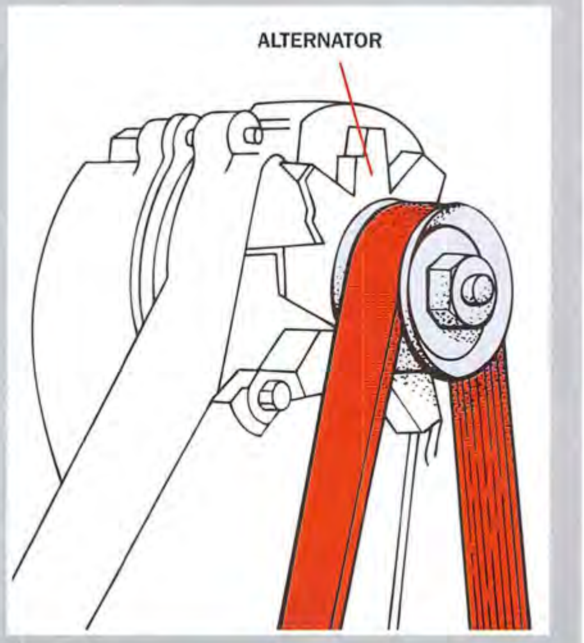

Do some touchy-feely on the idler pulley. Try to turn it, and if it's seized, the tensioner has to go. If it turns, feel for roughness (which indicates a bad bearing) and if you come up with a dab of grease from the bearing area, that indicates a grease leak past a bad seal. Try to rock the pulley: In-and-out movement may be a sign of a loose bolt, but more likely a damaged pulley. If a pulley is steel and it's clearly rusted, that not only indicates the coating has worn off (normal, in time), but that it is damaging the underside of the belt and you have to replace it.

Pulley alignment okay and mountings are solid? No matter how badly they're buried, inspect all the pulleys. Look for chips and other physically damaged sections that also could come in contact with the belt. If any of the grooved pulleys are packed with road film, belt debris, etc., wire-brush them clean. If the smooth pulleys (idler and water pump) are obviously worn on the belt-tracking area, you'll have to install new ones or a new tensioner assembly.

Doing It By Hand

A lot of older engines—particularly on Japanese cars—still have simple V-belts. Many also have ribbed belts with manual adjusters. The typical adjuster is a jackscrew; just loosen the lock bolt and turn the jackscrew, counterclockwise to loosen the belt, clockwise to tighten.

Even if the pulley is for a simple V-belt, it can accumulate debris in the groove. So clean it out with a wire brush, just as you would a multigroove. The days of getting an accurate sense of belt tension by pressing down on a belt with your thumb are long gone. It takes a tension gauge **(Fig. 4)**.

Although there are expensive professional gauges, we've found the Gate Krik-It gauges do a good job and fit into really tight places. Look for the longest belt span you can reach, preferably at least a foot. Just lay the Krik-It gauge at midspan, press down slowly on the center until you hear a cricketlike click, then stop pressing. The gauge bar rises as it encounters belt tension, and stops when you stop pressing. You can lift the gauge away and read the point of alignment of the edge of the bar with the linear dial on the gauge housing. Just turn down the jackscrew to get the specified tension, then tighten the lock bolt.

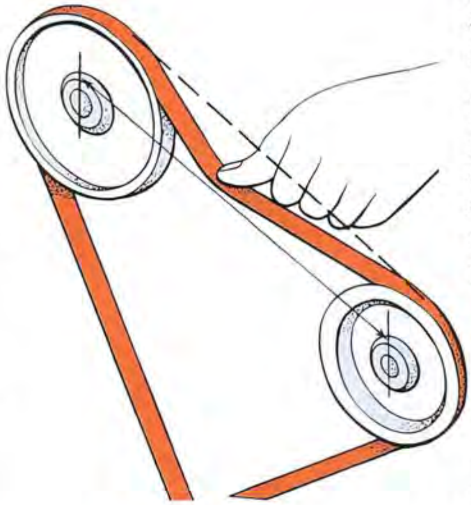

(Fig. 4) Don't check for correct tension with the time-honored one-thumb method. Use a gauge.

You're probably familiar with "belt dressing," an aerosol spray for belt squeal. Use it only to make a quick check of the belt as a source of noise. Repeated use ruins most belts.

HOW IT WORKS

Serpentine Belt

Gates Micro-V
Gates Micro-V
Gates Micro-V
Gates Micro-V

The serpentine belt is so named because its routing resembles the shape of a serpent. Carmakers developed the system so one belt would drive several, if not all, underhood accessories. A single belt, no matter how long, saves space under the hood since all the pulleys and accessories are on the same plane and there's no need to stack or stagger them. The ribbed side of the belt fits into grooved pulleys (each one a "mini" V), and the smooth back side of the belt often wraps around the water pump pulley and one or more idler pulleys. An idler is used to route the belt so there's plenty of "wrap" around each accessory pulley for good power transfer. To maintain adequate tension for good power transfer, and still eliminate the need for periodic adjustment, an automatic tensioner is usually used.

There's a nearly infinite variety of tensioner designs, but they all have a powerful coil spring inside. The poly-V belt design used on serpentine belts is capable of transmitting more power with less frictional loss than traditional V-belts. A properly installed and automatically tensioned serpentine belt system could remain troublefree for four to five years or 100,000 miles.

Replacing Hood and Hatch Struts

While topping off your transmission fluid one Saturday morning—hunched over the front of the engine, dog at your side—you feel the touch of cold steel on the back of your neck. You've received a weak "stab" from the car's hood latch. Time, weight and wear have conspired to force an unexpected surrender of the gas-charged hood support struts.

Unfortunately, gas struts do wear out—and sometimes at the most inopportune moments. Weak ones are more likely to fail in cold weather, when the gas pressure inside drops a little. Consequently, you'll find that these same weak struts work perfectly well on a hot summer day. But they should still be replaced for safety's sake. The good news is that one manufacturer recently introduced a temperature-compensating gas strut that should eliminate seasonal performance variations.

Sizing up the job

Large, modern auto parts stores often have parts catalogs dangling at the end of each aisle. So you may be able to look up strut applications on your own. If not, just go to the parts counter and ask for help. Be careful either way. Hood and hatch struts come in many different sizes and with several different mounting ends.

Some employ a ball-joint mount at each end. Others make do with any combination of the following at either end: a ball joint, a nonarticulating hard mount with an integral hinge or a flattened piston-rod end tab that mates with a stationary mounting peg. Some hood struts even

AFTER SECURELY PROPPING your hatch, examine the support struts' mounting ends to determine what socket or driver you'll need to begin removing the mounting hardware.

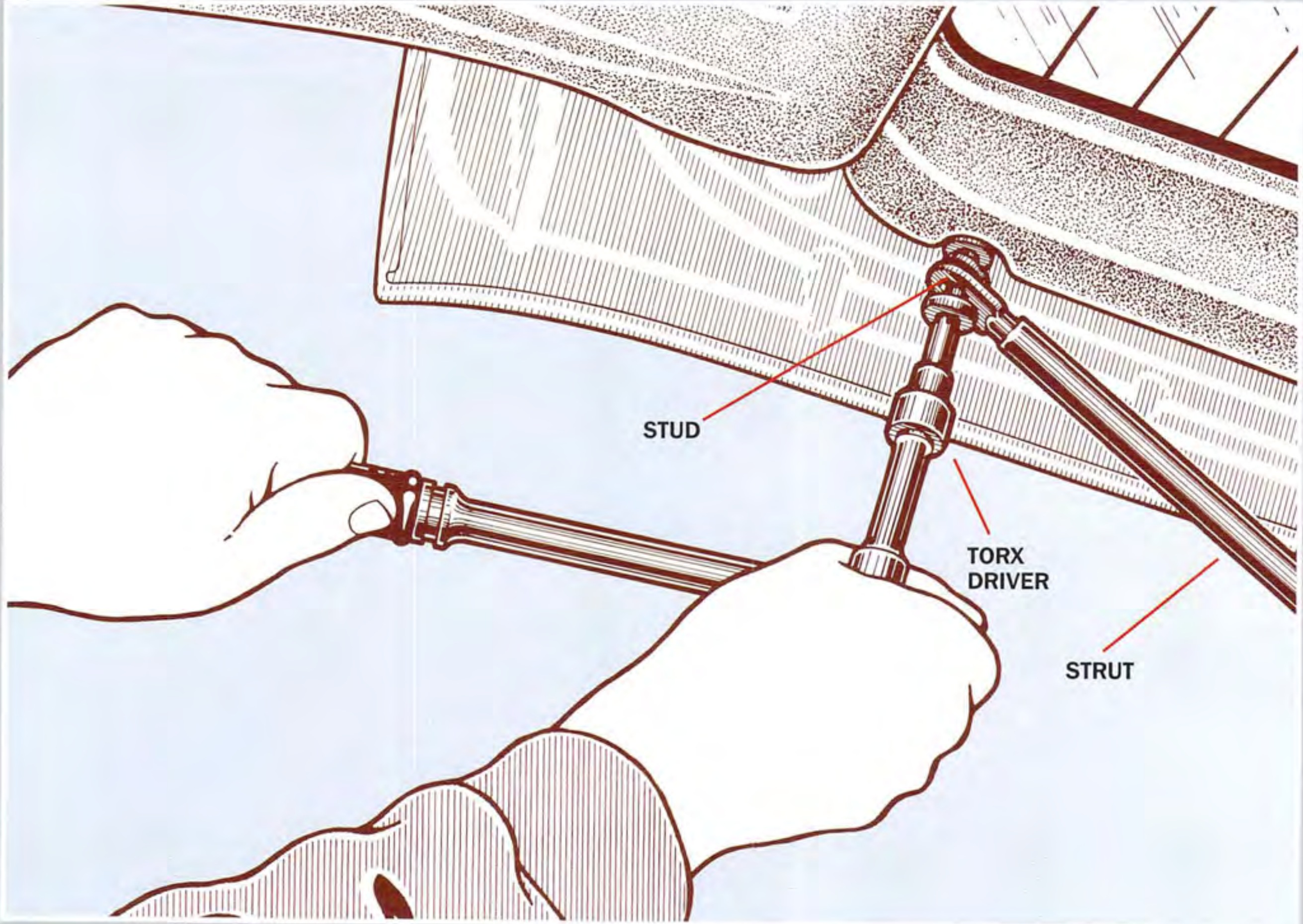

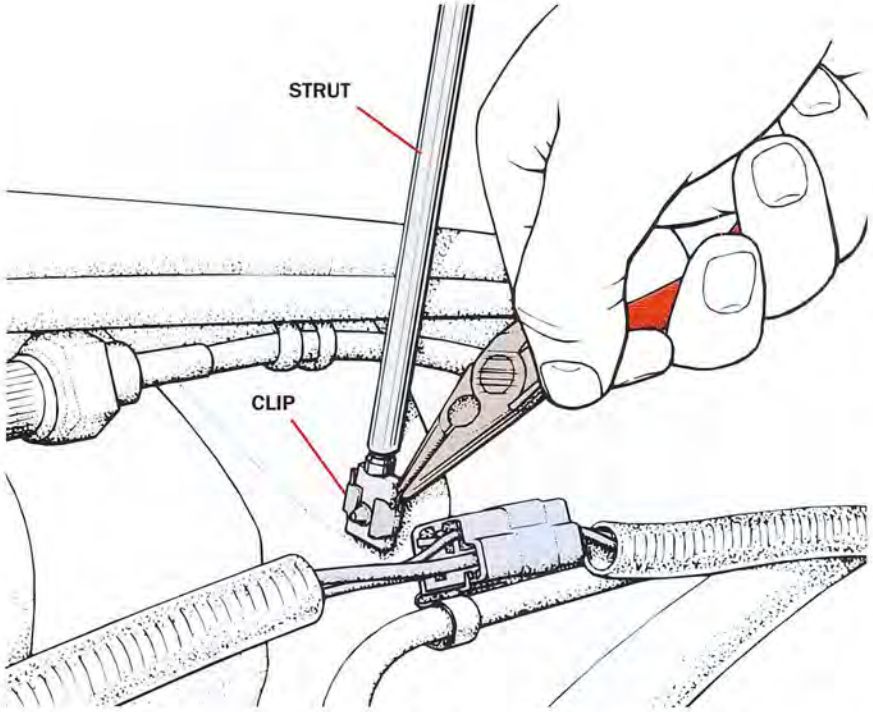

(Fig. 1) Slide the locking clip away from the mounting leg to allow strut removal. You might have to push or pull the clip to release it.

have two stages. On the Volvo 960, for example, unlocking a plastic clip on each strut allows the hood to continue opening past vertical for easier engine service.

Measure the full-open length of your existing struts before heading out to the parts store. Then, check the length of the new ones before going home. Or, take one of the old struts with you for comparison. It may save you a return trip. Be aware that some support struts for older cars may no longer be available in the traditional aftermarket. (The 1981 Datsun 310GX is one example.) If that's the case, you'll have to visit a dealer. Or maybe you'll get lucky and find the struts you need in the J.C. Whitney catalog.

With the correct unit(s) in hand, open the hood or hatch and prop it up with a stout piece of wood. On a 1-strut setup, the wood will be the only thing keeping the lid from meeting your head with a thud. On a 2-strut deal, the wood allows you to work on either side without worrying that the other strut is going to let go.

Doing the double-joint strut

With all the prep work done, replacing the old struts should take all of about 10 minutes. Those ball-joint ends are usually held together with some sort of clip: either a flat C-clip that mounts through the back of the socket and grabs the underside of the ball, or a wire C-clip that slides through the bottom of the socket and under the ball to prevent release.

Wedging a big screwdriver between the sheetmetal and the ball joint for leverage is not the best way to pop the strut. A little finesse, please. Use either needle-nose pliers or the smallest screwdriver you own to gently pull or pry clip away from the socket (**Fig. 1 & 2**). Then, simply pull the socket off the ball. Do the same thing at the other end. Save the clips if they're not broken. Sometimes an extra one comes in handy.

Be careful when working on hatch struts mounted near the rear-

(Fig. 2) This clip holds the socket onto the ball stud. Carefully remove the clip to release the sockets so you can pop off the struts.

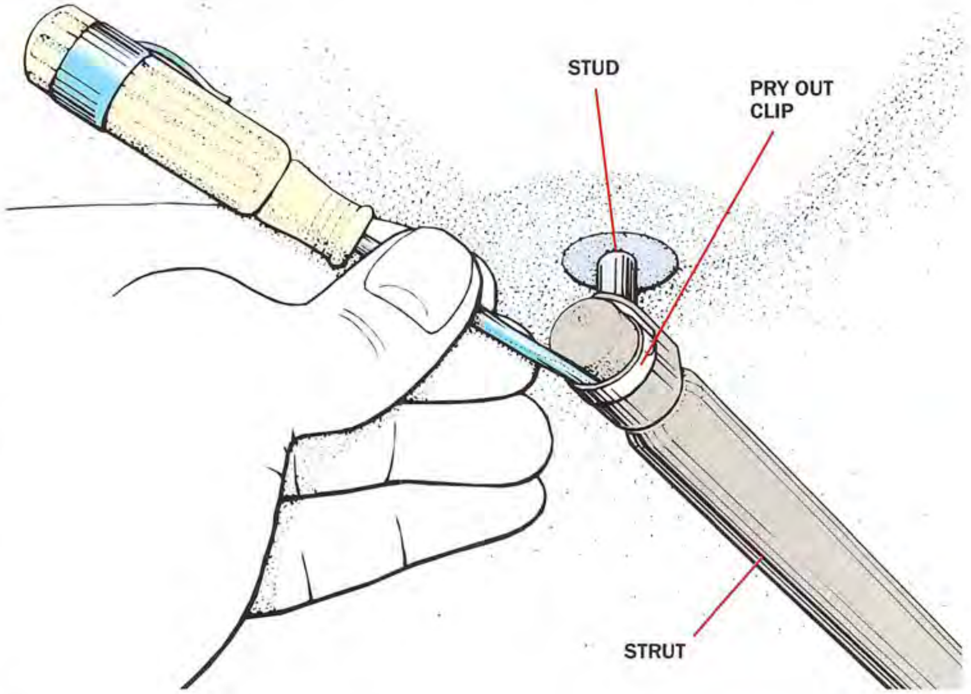

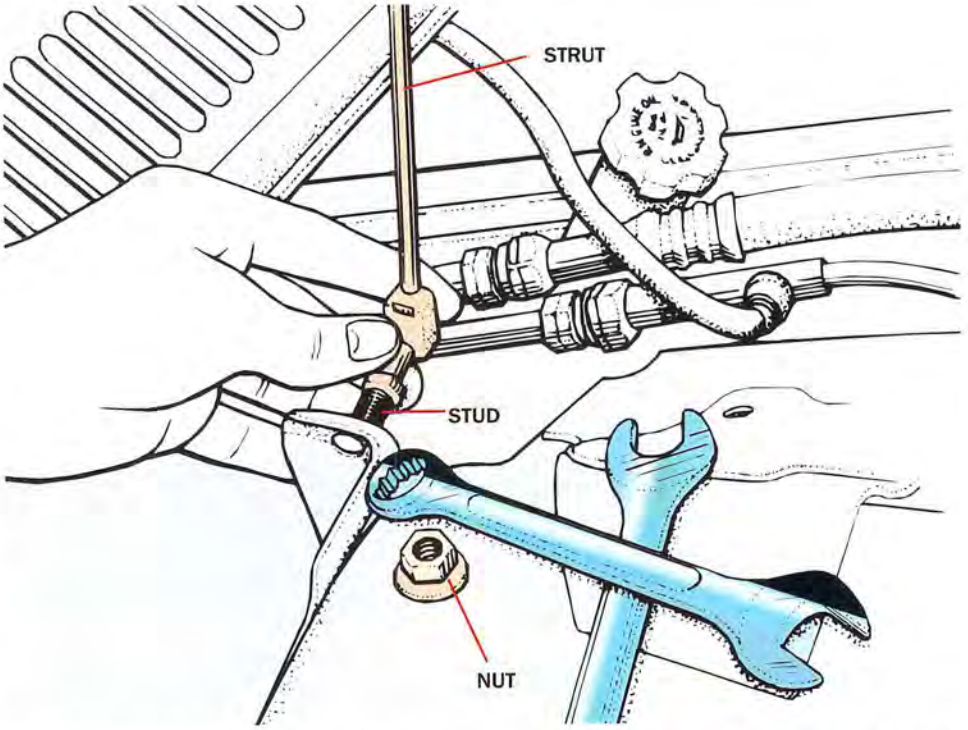

(Fig. 3) Insert the new strut's integral ball stud into its mounting hole and fasten it with the nut.

window defogger grid. If you slip with the screwdriver or pliers, you could ruin the defogger element.

Pop 'em in

To ease installation of the new unit, mount the clip(s) in a partially open position on the new socket, and put lube on the inside of the socket with a little silicone grease. Stick the socket over the ball, then snap the clip home. Repeat this for the other end (and the other side, if applicable) and you're done.

If you come across a ball joint that's not held together with a clip, see if the ball end of the joint is screwed into the sheetmetal (**Fig. 3**). If so, it's very likely that a hex is included for removing the old ball while it's still attached to the strut and screwing in the new strut's ball with an open-end wrench. Look at the new strut for guidance.

For the hard-mount type of strut end, just undo the bolt, screw or nut that holds the old strut in place, remove it and attach the new piece (**Fig. 4**). If you're dealing with a peg-style mount, you should find an E-clip of some sort or a cotter pin preventing the strut end from sliding off the peg.

E-clips generally need to be pulled or pushed in one direction or another in order to release them. Sometimes just a gentle nudge with the tip of a screwdriver is enough to pop them off. Otherwise, you may need to grab one of the clip's end tabs with needle-nose pliers and pull the clip out of position. Then, remove the old strut, mount the new one and slide the E-clip back home until you feel it click or can see that it is locked in.

Although the strut's piston rod has a hardened corrosionproof and weatherproof finish, you should still take care not to nick or scratch the polished surface. This would result in the marred metal ruining the strut's main seal. If the seal is damaged, the nitrogen gas would leak out and your brand-new strut would be trash.

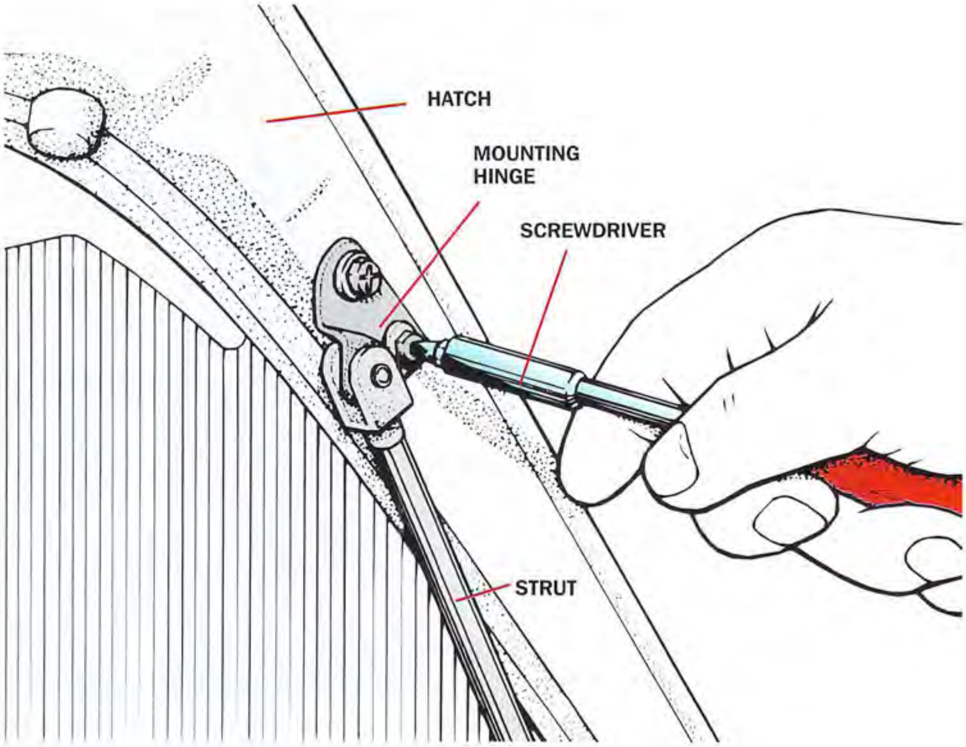

(Fig. 4) Unscrew the hinge from its mounting point on either end to remove this type of strut. You may even be able to use a ratchet wrench and socket.

Also, don't inadvertently mount the new unit piston-side up, especially if the old one was piston-side down. Most applications—those that contain some hydraulic oil as well as nitrogen gas—perform more smoothly if oriented with the rod end down. This allows the more effective oil compression damping to occur later in the strut's stroke for nifty end cushioning upon opening.

With the installation complete, pull out the wood prop and enjoy the convenience and safety of your hoods and hatches staying open on their own.

HOW IT WORKS

Support Struts

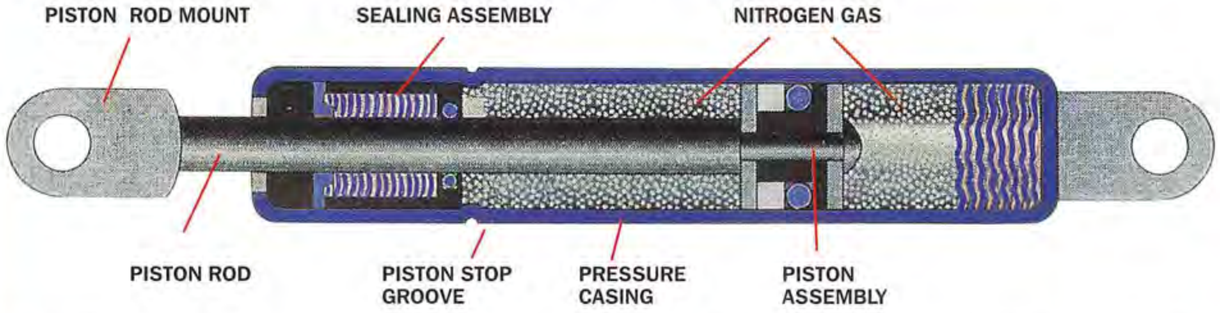

In the old days, your heavy metal hood and trunklid required beefy, space-robbing, spring-loaded hinges to stay open. Then the carmakers tried flimsy prop-rods for holding up the hood. But they invariably get in the way during service and tend to wear out and break at the pivot point on the radiator support.

So, as a result, gas-charged hood and hatch support struts, or gas springs, have steadily infiltrated the market. They're compact, light, durable and relatively inexpensive—key for auto manufacturing in the '90s.

One of the slickest applications around is on the Chrysler Cirrus/Dodge Stratus trunklid. No bulky suitcase-crushing hinges here, thanks to two stubby gas struts that open the lid past vertical to allow for easier trunk access.

How do they work? Like tiny shock absorbers really. But instead of being filled primarily with fluid for 2-way damping, they're charged with nitrogen gas for 1-way expansion. One end is fastened to the inner fender or cowl, the other to the hood or hatch. When you lift the lid a little bit, it extends the strut's piston rod enough for the gas to push on the piston and keep expanding the strut. The lid has nowhere to go but up.

On closing, you're actually pushing against the gas pressure to compress the strut back to its original size and position. A typical gas support strut is designed to last for at least 50,000 operations. They're maintenance-free throughout their life cycle but must be replaced when worn out.

MODERATE

Resetting Your Emissions Maintenance Reminder (EMR)

Your older minivan has been a trusted member of the family. It has helped you move your sister across town, kept your kid's Little League van pool rolling and faithfully transported the beloved dog to the vet. But it sure caused havoc last weekend. Leaving at 5 am with your buddies and a week's worth of fishing gear, it sputtered, spewed black smoke from the exhaust and wouldn't even idle. Walleye washout!

It's bad enough your fishing pals won't talk to you, but now the truck's got to go in for service. And the only way to get to work is by municipal bus because your wife needs the Bimmer to make all her sales calls. Peachy.

Your Dash Is Talking To You

Hold on. Have you been ignoring that warning light on the dash because your "car guy" brother-in-law said it's not important? Well, maybe it is. You see, that warning light, widely used on

(Fig. 1) Press button on EMR relay to reset reminder light for another 30,000 miles. This is an older Nissan engine.

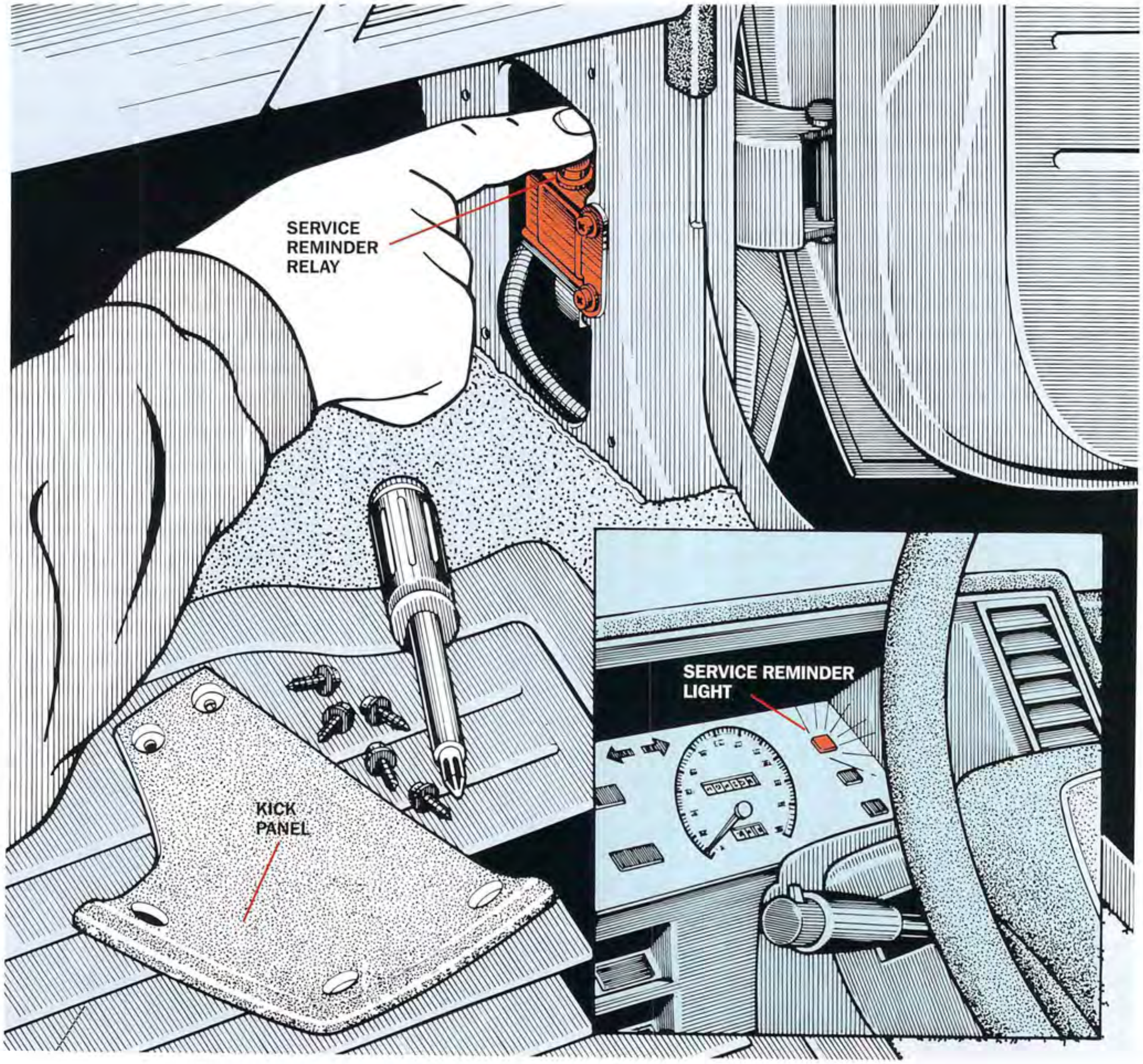

pre-OBDII vehicles (before 1996), is designed to come on at about 30,000-mile intervals, depending on the vehicle. (Check your owner's manual for the specifics.) It's called an emissions maintenance reminder, or EMR, and it definitely shouldn't be ignored.

When the EMR indicator glows, it doesn't necessarily mean trouble. But each time the light comes on, the emissions system should be checked. The idea is to work on the engine before the engine works you over. After giving your okay, the light has to be reset so the system starts counting down for the next EMR cycle.

There's a different reset procedure for just about every EMR system. Some indicators, such as those on most Chrysler light trucks since 1989, can be reset only by tapping a scan tool into the on-board computer controls. If you don't have a scanner, it's off to the repair shop or the dealer for assistance. Some setups are reset by replacing a module. After a certain number of miles, the module is simply left disconnected. There's no resetting provision at all on some vehicles. Just disconnect the plug or connector to shut off the light.

Still other EMR lights, such as the sensor warning on most Nissans (**Fig. 1**), require nothing more than a screwdriver and about 5 minutes of your time to either push a button on a relay or pull apart a connector—that is, after you've given the emissions controls a once-over.

Only under the following three conditions should you consider resetting the EMR light without inspecting the emissions system: 1. You've had no driveability trouble and no CHECK ENGINE light, 2. You've noticed no appreciable change in fuel economy, and 3. Your vehicle always passes the state emissions test with flying colors.

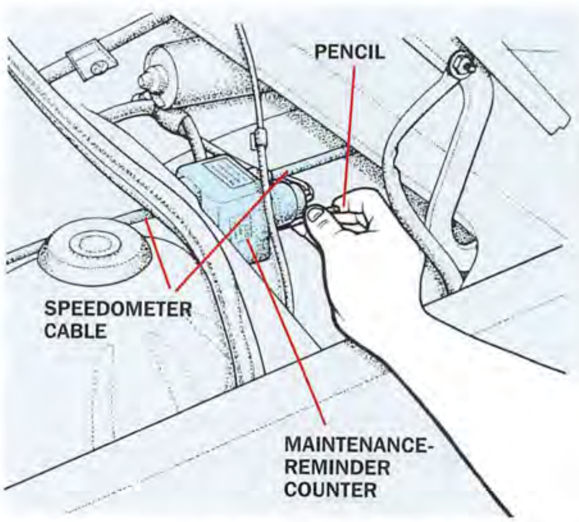

(Fig. 2) Stick a pen or pencil in the back of the mileage counter in the cowl tray on Audis and Volkswagens to reset EMR system.

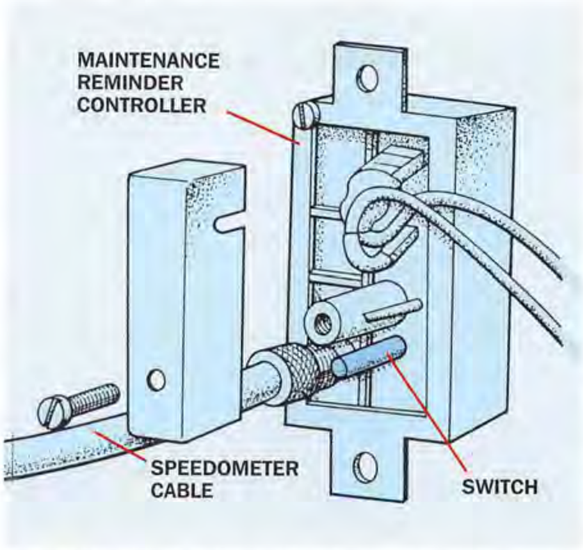

(Fig. 3) You'll find Volvo's mileage counter under dash. Press reset switch to turn off light.

Checking Emissions System

The key element in your emissions system is the oxygen sensor that's mounted in the exhaust manifold. It senses if the engine is running too rich or too lean. When the O_2 sensor gets lazy and stops reporting accurate information to the electronic engine controls, driveability trouble can begin—and fishing trips can be ruined. So check out the O_2 operation first.

Next on the list is the exhaust gas recirculation (EGR) valve, if your engine has one. Apply a vacuum to the valve's nipple and make sure the valve seat opens and closes smoothly. On late-model computer-controlled EGR systems, there may be no way to check the solenoid/valve mechanism without special tools. If you haven't heard any engine ping or knock lately, idle quality is okay, and if the CHECK ENGINE light hasn't come on, the EGR system is probably fine.

Other operations you can check include the smog pump (or pulse-air system) on engines that use them. Examine the pump belt for cracks, wear and proper tension. Follow the tubes from the pump to the diverter valve and down to the exhaust manifold and catalytic converter. Make sure the rubber hoses are intact and the metal pipes and check valves aren't rusted out. You might even be able to reach under the vehicle and rap the catalytic converter with a mallet or hammer. A rattling noise means the cat has broken down internally—and its better days are behind it.

If you can see the fuel-system charcoal-vapor canister (sometimes it's buried in a front fender), inspect the integrity of its rubber lines and fittings. And while you're at it, look at the air filter, air-box breather element and the engine's vacuum lines. Make sure the positive crankcase ventilation valve still rattles when you shake it. Check the spark plugs and wires and the distributor cap and rotor (if applicable) for wear. Ultimately, all powertrain components make up the emissions system, so check whatever you can get to.

Finding the Reset Location

Although it varies, reset-switch locations can generally be broken down into three areas: the speedometer cable/firewall area, the kick panel/under-dash area and the instrument panel area. The positions often

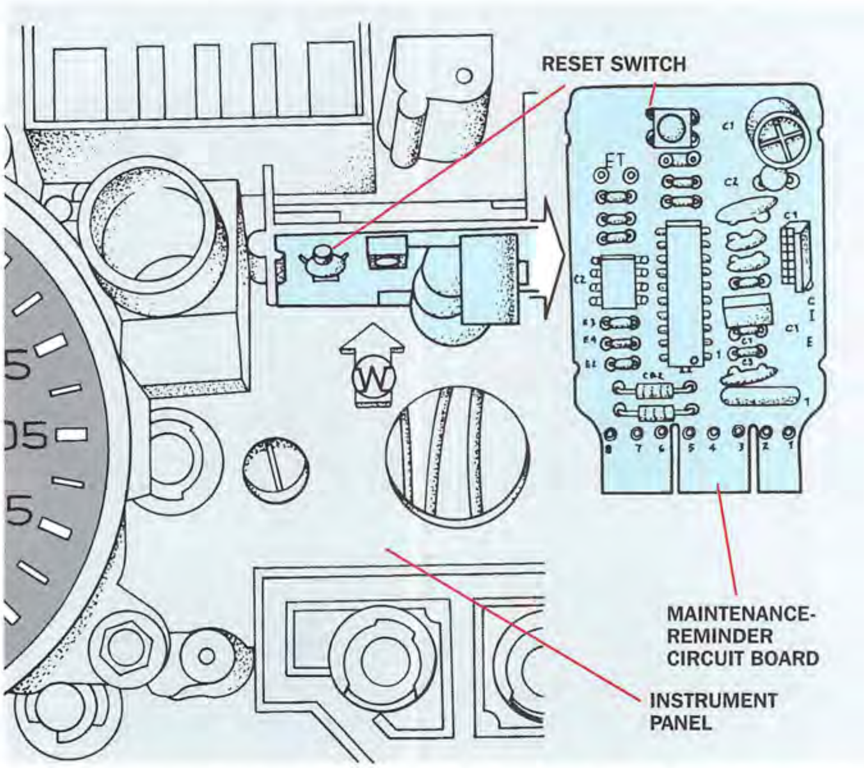

(Fig. 4) On Chrysler minivans, you must remove the fuel gauge or tachometer to get to the emissions-maintenance-reminder reset switch.

change from model year to model year across the same vehicle make, so always approach your search with an open mind. And check your owner's manual.

The speedo cable resets are usually the easiest because the cable sticks out like a sore thumb, and there's usually plenty of room to work. Vehicles in this category include many models from the early 1980s.

Audis and Volkswagens are a breeze. You'll find the mileage counter in the cowl tray at the base of the windshield (the speedo cable leads right to it). There's a hole on the back of the counter. Stick a pen or pencil in the hole and press in the reset button (Fig. 2). You'll feel the button lock in place. That's it. On Volvos, the mileage counter is also in line with the speedo cable. Once you've found the counter, press its reset button with your finger (Fig. 3).

Still others are done from the inside. Drop the lower steering-column cover. Feel for the sensor reset cable to the left of the speedo cluster. Yank on the cable until you see the flag reset itself on the left of the odometer.

Hunting for Switches

Vehicles that fall into the second category number too many to call out separately, but they all have some sort of EMR module, relay, connector or pair of connectors.

On some models, you'll find a cancellation switch on the right side of the fuse panel. Press the switch to the OFF position. Older Fords hide EMR modules behind the passenger kick panel on their trucks. Stick a small rod in the hole marked RESET. While holding the button down with the rod, turn the key to RUN and continue to depress the button for 5 seconds. On older Jaguars, the mileage counter is behind the trim in the trunk. Press the white button with the key on to reset the light.

Pulling the Instrument Panel

This third category can get complicated because you usually have to dismantle part of the dash to remove the instrument cluster or a gauge to get in behind the cluster, as on Chrysler's minivans (Fig. 4). According to a diagnostic hotline we are friendly with, these vans are the cause of the most EMR reset calls. The circuit-board-mounted switch is tricky to find if you don't know what to look for. It's mounted behind the fuel gauge on the standard cluster or behind the tachometer on the optional cluster. Sometimes you'll find a 9-volt battery in there, too. It must be replaced with an alkaline battery. Some Chrysler captive import light trucks have a reset switch on the back of the cluster near the speedo junction or on the lower righthand corner of the cluster, behind the face panel.

On early-model GM vehicles, pull the instrument cluster lens and use a small tool to push on the stem of the orange plastic flag that pops up near the odometer. An alignment mark should appear in the left center of the odometer window when the flag is properly reset. Late-model GMs use no specific EMR indicator. The CHECK ENGINE light illuminates if there's a problem.

Early-model non-turbo Audis have a cancel switch labeled OXS on the back of the instrument cluster that must be removed from the pod to gain access. Late-model turbos hide a mileage counter with a reset button under the rear seat. Some late-model BMWs use a resettable mileage counter. On 1989-93 Geo Trackers, locate the 3-wire cancellation switch in the main wiring harness behind the instrument cluster or in the steering-column lower-access panel.

You'll find the reset on the back of the speedo on Isuzu Troopers. Pull the tape from Hole B and swap the screw from Hole A to Hole B. Tape over Hole A. Mazda MPVs are similar. Swap the screw from Hole NO to Hole NC. Late-model Mercedes-Benz cars use no EMR indicator. However, 1980-85s, other than the 280, require bulb removal. On VW Rabbits and Pickups, remove the instrument panel cover trim plate to get to the mileage counter release arms at the top left of the speedo housing. Good luck!

Passing An Emissions Inspection

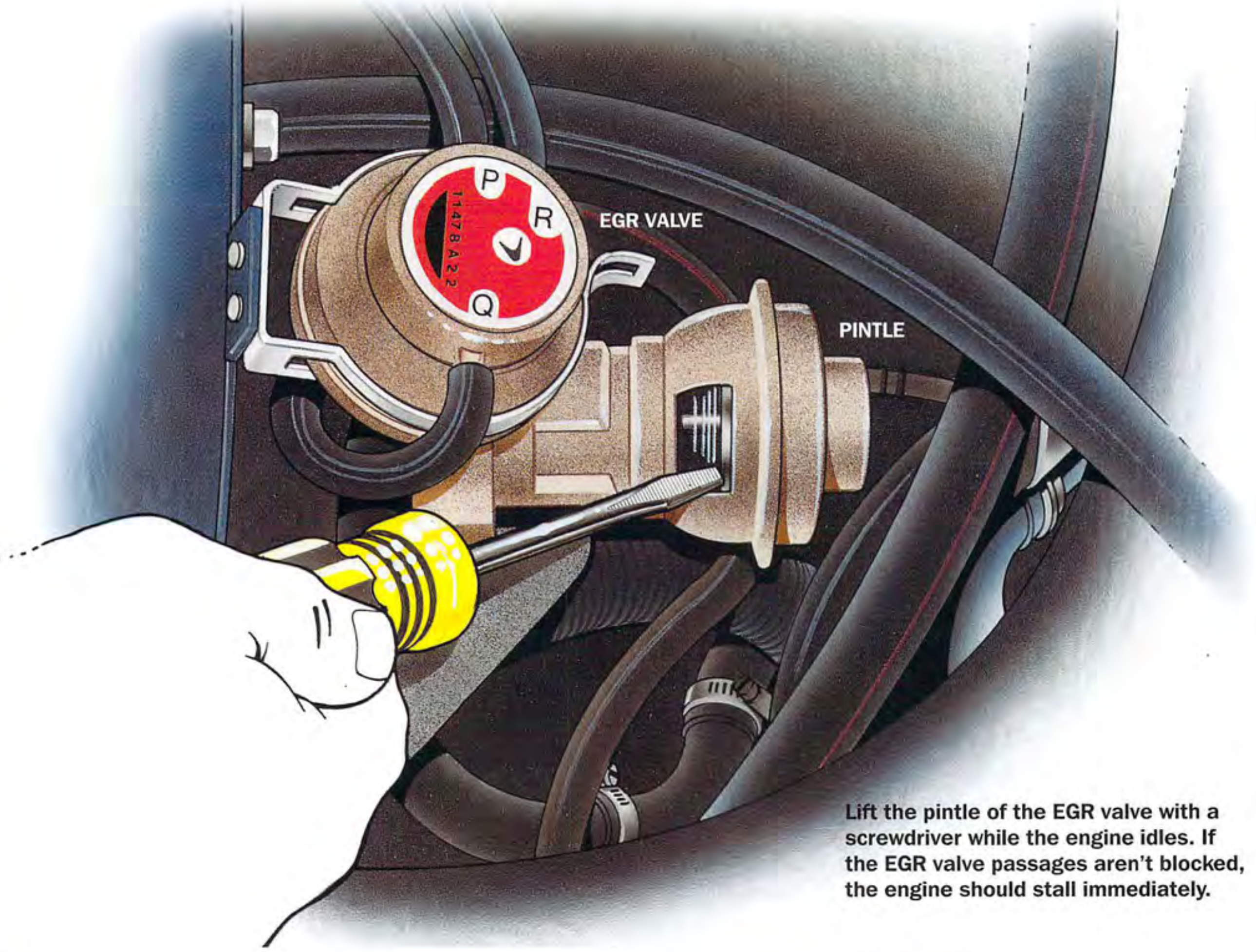

Lift the pintle of the EGR valve with a screwdriver while the engine idles. If the EGR valve passages aren't blocked, the engine should stall immediately.

Your car is running beautifully, and all seems right with the world. You have no trepidation whatsoever when the time comes to take it in for your state or local emissions test. After an interminable wait in line at the inspection station, a smiling, snappily dressed functionary shoves a probe up your tailpipe and asks you to gun it. Moments later, the inspector smiles wearily as he hands you an envelope—not a new inspection sticker for your license plate, but an envelope that says "FAIL" in bright red letters. You have only a month to do your homework and pass inspection. Heck, getting into college wasn't this hard.

Wasn't your car designed to meet these emissions standards? Haven't you done your best to maintain it with regular oil changes and tuneups? What's gone wrong? And, most important, how can you make it pass?

Take a minute to read the rejection slip carefully. Typically, it will give you the test results for your car along with the acceptable limits. Note which of the "terrible trio"—hydrocarbons (HC), carbon monoxide (CO) or oxides of nitrogen (NOx)—is out of bounds.

Raw Gas

Hydrocarbon emissions are simply unburned fuel being pumped raw into the exhaust system. Misfiring is the most likely culprit, and that can come from an ignition problem, or an internal engine failure that reduces compression. Another possibility is a mixture that's too lean to catch fire dependably because of either a fuel system malfunction or a vacuum leak. If a high HC number isn't from a miss, it's probably due to an overly rich mix.

So, worn or fouled spark plugs are the first things to suspect. If you can

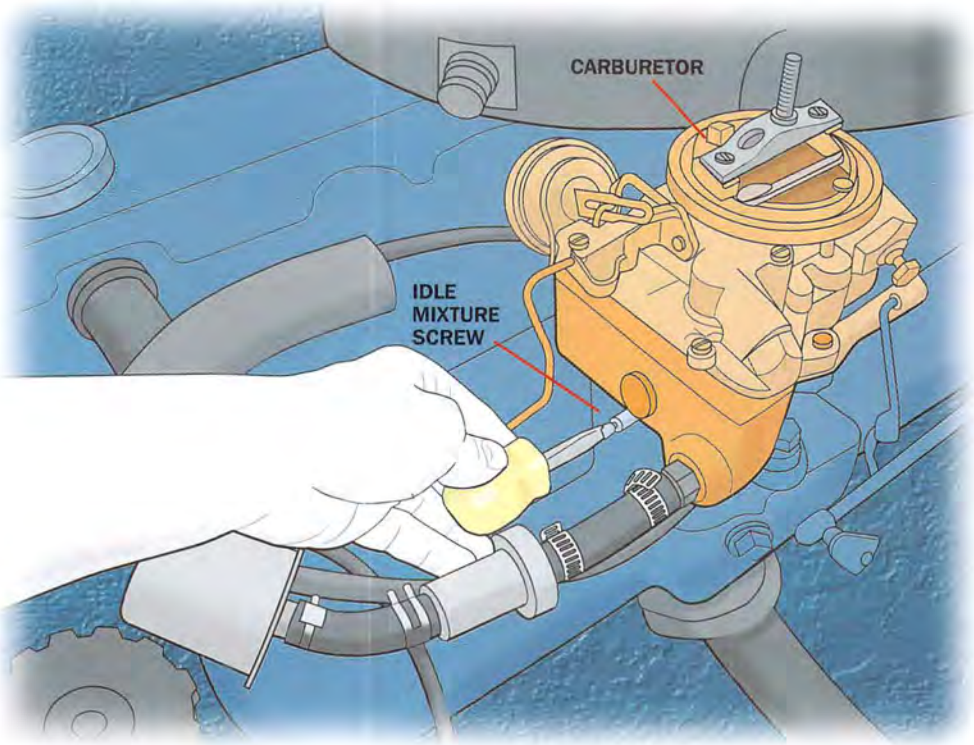

Older cars might benefit from tweaking the idle mixture screw no more than a half turn in either direction to get the idle mixture setting correct.

feel a dead miss at idle, use insulated pliers to pull plug wires one at a time to see if one in particular does not have any effect on smoothness (don't do this on '96 and newer cars with OBD II, as it will set a misfire code and turn on the Malfunction Indicator Lamp on the dash). Remove and inspect that spark plug. If it's okay, use an ohmmeter to find out if the plug wire is open—you should meter 5000 to 20,000 ohms, then examine the distributor cap for heavy corrosion in the sockets, cracks, etc.

With the plug out, do a compression test—anything much less than 100 psi should be suspect (a valve may be burned). Also, look over the intake manifold to see if there's a vacuum tap in the runner leading to the problem cylinder. A leak here can lean out the mix to the point of misfire at idle.

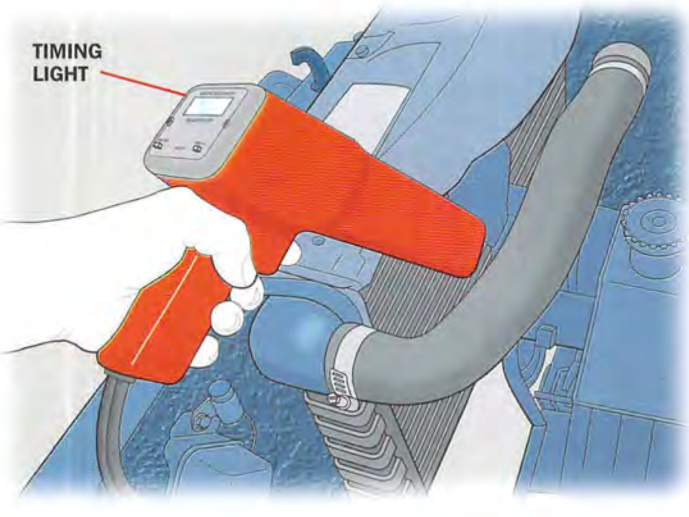

Many cars no longer have adjustable ignition timing, but you should check it with a timing light nonetheless.

There are a couple of other things that can supply enough extra gasoline to put hydrocarbons above the limit. One is the evaporative-emissions control system that traps gasoline vapors that would otherwise escape into the atmosphere. It then meters the vapors into the intake during off-idle modes to be burned. If it flows at idle, the mix will be too rich.

Another potential culprit is engine oil that's diluted with gasoline. Fuel vapors are drawn through the Positive Crankcase Ventilation system and into the intake manifold. Smelling the dipstick may tip you off to this, and an oil change will correct the problem—temporarily.

Killer Compound

Carbon monoxide is produced by incomplete combustion—the air/fuel charge is ignited, but isn't burning properly or entirely. In most cases, this is due to fuel-system troubles that make the engine run rich, but intake air restrictions could affect it too. It's important to note that some combustion is required for CO to be created. So, you won't see a high CO reading if there's no spark.

With a carburetor, richness might be due to an idle mixture screw that's open too much, a heavy float (the foam type may absorb gasoline, and the brass pontoon type can develop holes and fill up with fuel), a choke that stays on after the engine has warmed up, or a power valve that sticks in the open position. With fuel injection, perhaps the fuel pressure in the rail is too high due to a faulty fuel-pressure regulator, or there may be a cracked vacuum hose to the regulator.

Of course, just about any car built within the last two decades will have an electronic engine management system that controls the air/fuel mixture (among other things), so consider the sensors that feed information

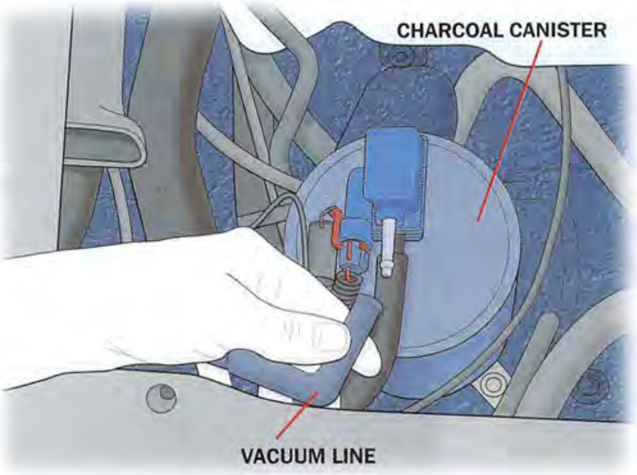

A cracked vacuum line can cause the charcoal canister to malfunction, increasing emissions.

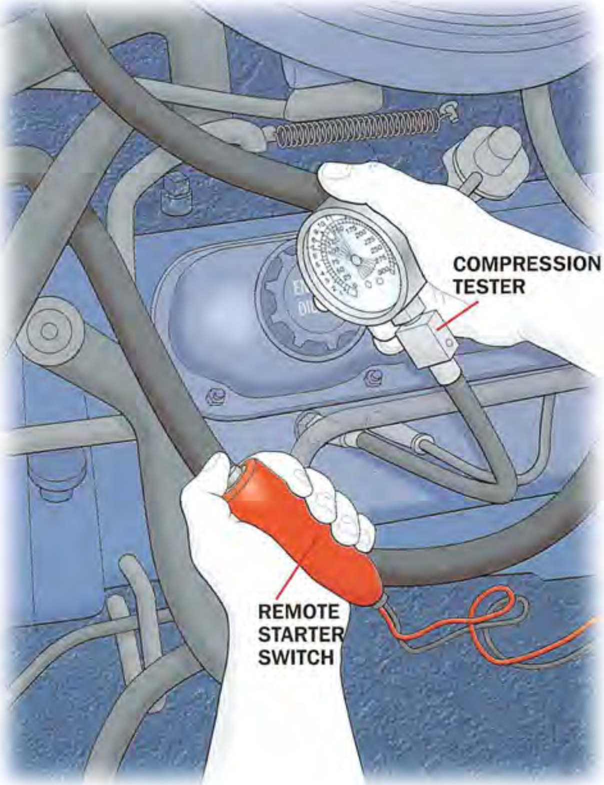

A misfiring cylinder may have bad rings or valves, which can be spotted with a compression test.

to the computer. This, however, graduates into professional-level troubleshooting.

On a precomputer carbureted car, one way to get under the HC and CO limits is to retard the ignition timing. This will slow the idle and provide a longer, hotter burn. Then, bring the idle speed back up to specs with the throttle stop screw or idle stop solenoid, which will admit more air. Performance will be off, but you'll probably pass the test.

Hard NOx

NOx is a tough one. It forms when nitrogen (78 percent of the air we breathe) combines with oxygen (21 percent of the atmosphere), which can happen only at very high combustion temperatures. The trouble is, many of the things that were initially done to cut HC and CO involved dramatically increasing the heat inside a car's cylinders. That heat created plenty of NOx. Fortunately, it was found that only the peak temps (over 2500°F) produced this smog-promoting gas, not the long, hot burn that oxidized the other two pollutants. Introducing Exhaust Gas Recirculation (EGR) knocked the top off the temperature curve, thus putting NOx within bounds. One of the main reasons electronic engine management systems were developed in the late 1970s was to allow a 3-way catalyst, which adds rhodium to the 2-way cat's platinum and palladium to further diminish NOx.

The first things to suspect are an inoperative EGR valve (perhaps the vacuum line is cracked, or the internal diaphragm has ruptured) and carbon-clogged EGR passages. Try applying vacuum to the valve at idle. If the engine starts to run roughly or stalls, the valve is okay and the passages are at least partially open. Remove the valve and dig out all the deposits you can.

There's a good chance that the catalytic converter is no longer capable of reducing NOx, but determining that requires high-tech diagnosis best left to a professional. There's another distinct possibility that you can deal with, however: heavy carbon deposits in the combustion chamber and on the backs of the intake valves. These can raise the compression ratio enough to cause the hot flash that forms NOx, and will also interfere with the carefully engineered swirl pattern of the incoming mixture that makes it burn evenly.

Numerous treatments have been devised to avoid the huge job of removing the head and manually cleaning the combustion chambers, piston tops and valves, but the only one suitable for the do-it-yourselfer is chemical treatment. The most basic approach is to pour a suitable additive into the gas tank. A more thorough job can be achieved by adding detergent- or solvent-based cleaners directly into the intake stream.

Too much heat is another thing to think about. Is the engine running hot? Is the intake air overheated due to a jammed Early Fuel Evaporation valve, or is a Thermostatic Air Cleaner door in the snorkel stuck closed?

Ignition timing that's too far advanced can result in detonation that causes NOx to rise. A switch to premium fuel may reduce it. A final note: Moisture-laden air can lower NOx output by as much as 30 percent, so you might consider visiting the emissions station on a rainy day.

Changing Your Oil And Filter

The line at the quick-lube goes around the building and nearly to the street. But the auto parts store parking lot is empty—now's your chance to show the wife that it really is cheaper and easier to change your own oil. Especially after watching the grease-stained quick-lube mechanic through the bay doors for a couple of oil changes. Air wrenches on the oil-drain plug? Even though you realize that cars have changed since you got out of the habit of changing your own, they haven't changed that much.

All Jacked Up

Jacking up the front of the car and putting it on safety stands is still the first step, and placing the drain pan under the oil-drain plug is the second. Okay, you're ready to loosen the plug, but it's really tight. That mechanic with his air wrench apparently was the last one to tighten it. Drain plugs typically have a soft metal washer or a sealing washer with a rubber insert. Use a properly fitting box wrench or a 6-point socket wrench to remove the plug. Be careful not to rock the socket to the side as this could damage the flats on the bolt.

If putting the car up on stands doesn't leave you enough underbody space to get adequate leverage, you can use a 4-way lug wrench provided it has the proper end. This will allow you to twist the bolt without adding any side thrust, and you'll be able to

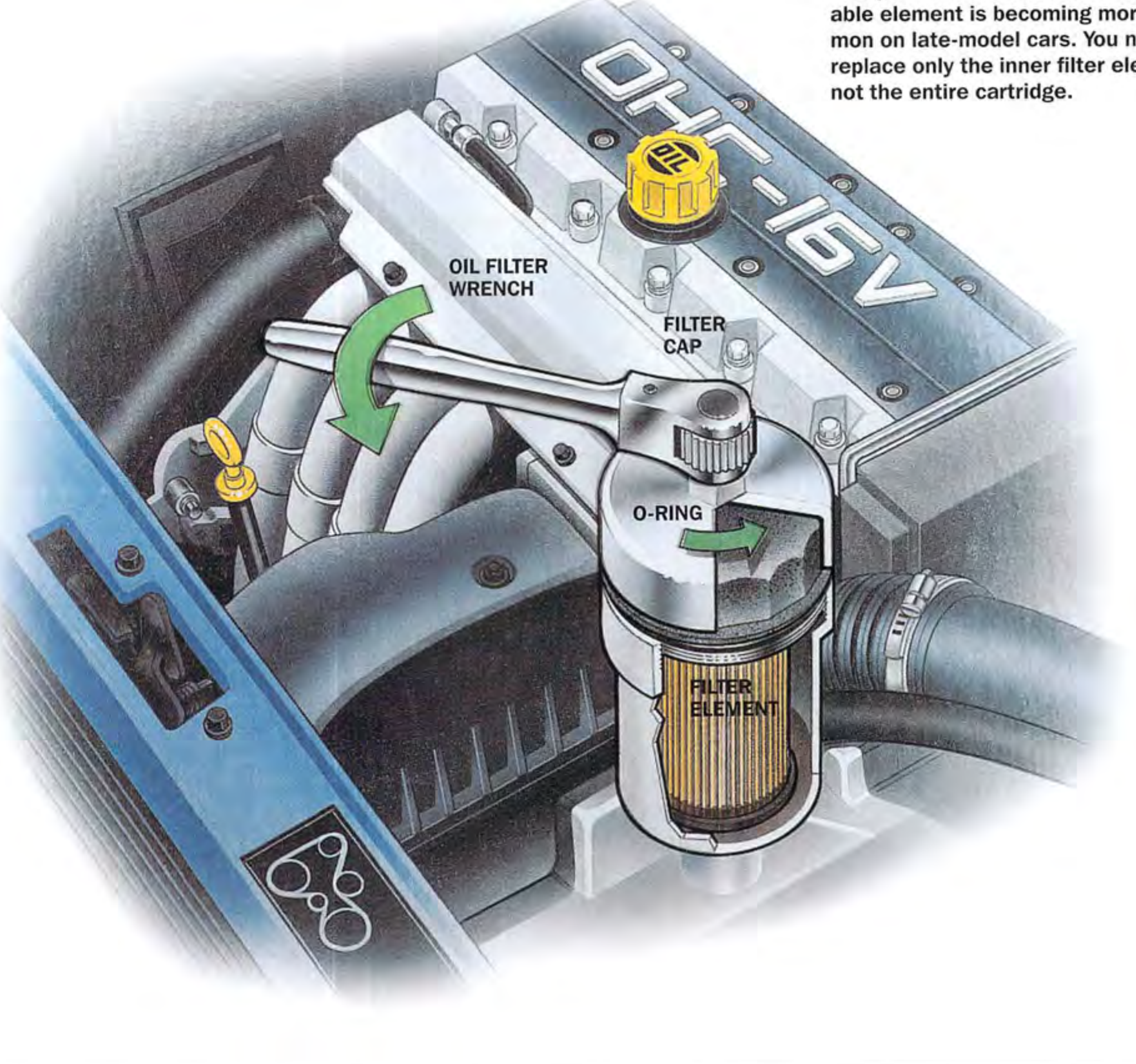

This permanent filter with a replaceable element is becoming more common on late-model cars. You need to replace only the inner filter element, not the entire cartridge.

use both hands and all of your upper-body strength. Remain calm, have patience and you'll get it loose.

But of course, life isn't easy. The threads in the pan and on the plug are damaged, perhaps from partial cross-threading. What to do now? Forget the drain plug and get a repair kit. A typical kit has a replacement fitting that cuts deeper threads, and when tightened, it seals against a washer and remains in place. Some kits cut fresh, deeper threads for a new plug. A hexhead brass cap with an O-ring seal threads onto the end. When it's time to change the oil, you unthread the cap. Others (for odd-size, severely damaged holes) are fat, cone-shaped synthetic rubber plugs larger than the drain-plug hole. You force a special rod into an opening in the cone, which temporarily stretches it and reduces its diameter, allowing it to fit in the hole. Withdraw the rod, and the cone relaxes and seals the hole—the cone won't come out until you force in that rod to stretch it.

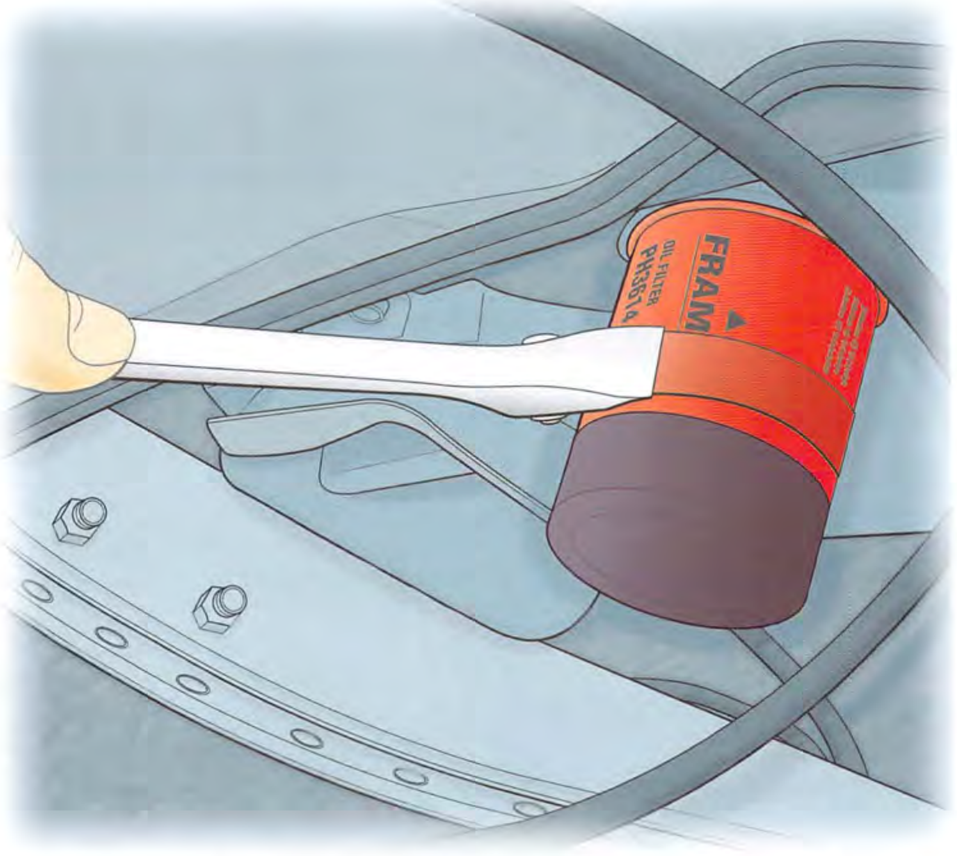

A simple band wrench may be all that's necessary to remove a canister filter. Make sure you have enough room to swing the wrench.

If the drain plug looks marginal, consider installing a Fram oil-drain valve kit. These are available for the most common types of drain-plug holes. Thread a spring-loaded valve assembly with a copper washer into the hole and tighten. The valve is the primary oil seal, and a knurled cap threads on fingertight against an O-ring—this keeps out dirt.

When it's time to drain oil, unthread the cap and thread on a fitting with a drain hose, which you can aim right into the pan (no splatter and no hot oil running down your arm). The hose fitting has an internal tip that pushes open the spring-loaded valve, and the oil drains out. When the pan is drained, unthread the hose fitting (the valve springs shut), reinstall the fingertight cap and you're good to go (after changing the filter and putting fresh oil in the engine, of course). If the drain plug is okay and you want to reuse it, replace the washer and then tighten the plug to specifications—20 ft.-lb. to perhaps 35 ft.-lb.—depending on the size of the plug.

Empty the drained oil into a suitable container. (In my area, the county gives away flat jugs with a giant built-in funnel.) When the jug is full, take the oil to the store where you bought the new oil to be emptied into an oil-recycling tank. Many states require service stations to take small quantities of old oil. Dumping oil in that low spot behind the shed or into the sink is not accept-

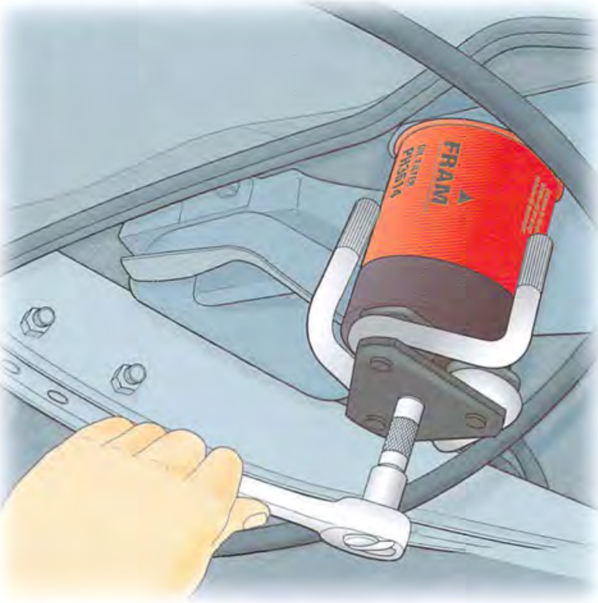

A spider-style wrench may be more suitable for filters that are not easily accessible.

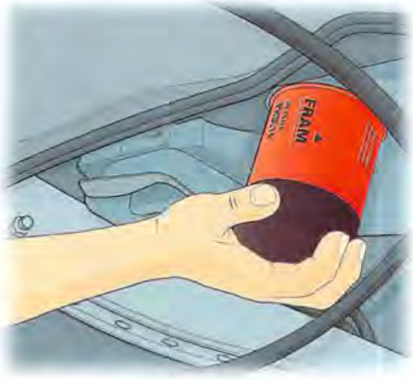

This filter has a rubberized area to provide for tightening by hand.

able—and probably is illegal where you live.

Off With The Filter

When removing an oil filter, what you need more than anything is the appropriate wrench. There are many sizes, and perhaps the most common answer is the cap wrench, which fits on the end of the filter. Only problem: The wrench fits against a fluted pattern and there are countless fluted patterns. So not only do you need the right size, but the right internal shape to fit the flutes. The one that fits the filter on your car now, we should warn you, may not fit the replacement filter you buy.

Sometimes the end of a filter has a hexnut in the center, so you could use a conventional wrench. If you can't find the right size and shape cap wrench, try one of the following:

- Nylon band wrench. This universal tool wraps a band around the filter tightly.
- Coil-spring wrench. It fits over the end of the conventional spin-on filter, and the band coil extends to near the filter's base. Turn the end with a wrench and the coils tighten around the entire filter. This prevents damage and separation from the base.
- Spring-band wrench. This wrench fits over the end of the filter and extends just past the flutes, so it grips the full circle of the filter body.
- A "spider" has three fluted legs that clamp against the spin-on cartridge and "dig in," preventing slippage. It has an end plate for a ratchet, and because the spider legs extend 2½ in., they grip well past the outer end of the filter. This design provides well-distributed gripping power.

It is possible to loosen a badly stuck filter with a good filter wrench.

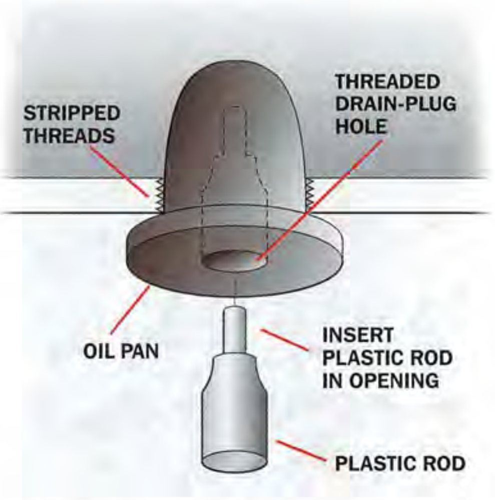

This simple rubber plug can rescue a stripped-out oil-drain-plug hole. Insert the rod to install

How about driving a big screwdriver through the spin-on cartridge and using that to loosen a stuck filter? You're more likely to destroy the cartridge and still not loosen the filter.

Have the drain pan in place, sitting on a spread of newspapers or an oil-absorbent pad—just in case. Once the filter is loosened, oil may start flowing to the ground.

Unthread the filter and carefully empty it into the drain pan.

Refilling

Apply a film of clean engine oil to the gasket of the new filter, then thread the filter on by hand. Some filters have a rubberized surface to make it easy to turn. Every reputable oil filter is designed to seal for tens of thousands of miles with no more than a good hand-tightening. You don't need a wrench unless you have one of those deeply recessed filters with no space around it for your hands. (If that's the case, the only choice is a cap wrench.) Turn the filter until you feel the filter base just make contact with its mounting plate. Then use the wrench to tighten a half-turn more.

If you haven't bought engine oil in a while, you may be confused by the choices. For a little friendly advice, see "What's That Starburst?" left.

WHAT'S THAT STARBURST?

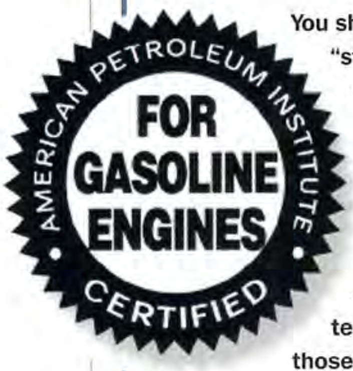

You should see two labels on a container of oil. One is a "starburst" with the words American Petroleum Institute Certified printed on it. This means that the oil has passed tests for the recommended service in which it's to be used—that includes a test by ISLAC (International Lubricant Standardization and Approval Committee). Passing the ISLAC oil test enables carmakers to add to their fuel economy numbers. Oils such as 0W (a synthetic grade) and 5W (thin) pass this test, and so do some 10W-30 oils. Heavier oils, including those formulated for older cars, do not.

On the second label you'll see the words "Energy Conserving." Real world: Oils labeled this way won't make a noticeable difference in your car's gas mileage. This label also lists the service category. On late models you want an oil designated For Service SL. You may also see some SL/SJ-rated oil on shelves. Read your owner's manual.

DIFFICULT

Search Engines

Your temperature gauge inches up toward the danger point. Simultaneously, traffic slows ever closer to 0 mph—without ever quite reaching it—as the rush hour deepens. There's no air moving through your radiator, and any that does is too damned hot. You start to smell that syrupy-sweet pancake-breakfast-in-a-Turkish-bath odor of near-boiling antifreeze. You turn off the a/c to reduce the load, but the needle doesn't drop. Time to upgrade the cooling system.

After spending a full Saturday and $450, you've installed a high-performance radiator and the problem disappears. Then, your friend with the same make of car has the same problem. So you tell him the good—if expensive—news. "I also got the problem fixed," he tells you, "but it cost me only $75 at the car dealer—something about reprogramming the computer. The car was out of the place in a half-hour."

Reprogramming toggled the cooling fan on and up to high speed much earlier, using a revised fan control algorithm. If only you had known.

Look It Up

Now you can know. You can do more—and better—work on your car with the special manuals, bulletins and some of the special tools the dealer shops have. In fact, you even can reprogram your car's computer using your PC and a generic device that works on all makes. You thought all that was secret stuff or incredibly expensive? Not anymore—it's available to everyone from factory tech-service Web sites. Aside from a shop uniform, anything the dealer techs get, you can too, including once-secret memos. Only a few low-volume cars, such as Ferrari, Rolls-Royce, Aston Martin, etc., are exempt.

Virtually every Web site meets or beats a cost of $20 per day for all information (except BMW at $25 per day, Porsche and VW/Audi with a per-document price). Hyundai and Kia sites are free, and others offer some freebies even if you don't buy anything.

Web site information is "professional," intended for service technicians with vehicle-specific training programs under their belt—so you may need to go slowly to understand it. However, most sites have tutorials, even online auto tech courses you can take, as well as technical manuals to read. They're designed for basic training, so they won't lose a Saturday mechanic.

Most of the Web sites have pretty straightforward navigation, but it's still not like opening a book. Always read the system requirements and fre-

Reprogramming your car's engine management computer is only slightly more complicated than downloading music onto your MP3 player. The pass-through device connects the car's OBD II diagnostic connector to your computer.

quently asked questions (FAQ) first. High-speed Internet service may be just "recommended," but without it you're in for real frustration, both for downloading special software and using certain features. Some downloads are over 15MB, and your actual transfer speed may be far less than the connection rate—meaning literally hours of downloading. Throughput rates might be better in the wee hours when there's less traffic on the Internet. After your days-long subscription to the Web site expires, you'll still want access to the data you've paid for. So print out the service data as you download it, or at least keep it on your hard drive and backed up to some other place.

Advice For The Slow Lane

If you must use a dial-up modem, understand that only straightforward wiring diagrams (not interactive types) and non-PDF repair instructions will open within tolerable times. Also, you should ask someone with a broadband connection to download needed software and store it on a CD. Installing from a CD saves lots of time.

An alternative that works reasonably well with dial-up is Alldatadiy.com, the Saturday mechanic's site from the supplier of professional service information. For $25 (one car, $15 each additional), you get a one-year subscription to all the information Alldata provides to independent garages as far back as 1982, including continuous updating of Technical Service Bulletin (TSB) listings. It's easy to navigate, although service manual information lags model introductions by one to two years. Information is edited and reformatted from factory manuals, although some specialty information is excluded, such as automatic transmission overhaul. All leading makes are covered except Honda and BMW (which are only in Alldata's professional system). TSB postings run only slightly behind the carmakers' own listings. If you have an older car, Alldatadiy.com may be your only choice, as most factory Web sites (except for service bulletins) cover largely late-model stuff.

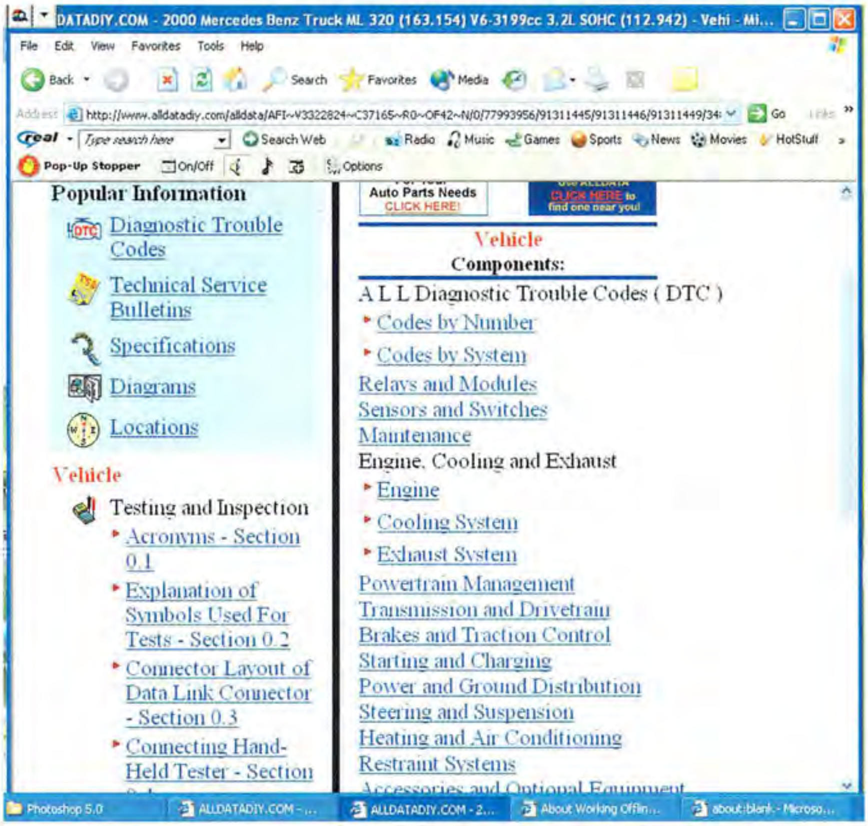
DATADIY.COM - 2000 Mercedes Benz Truck ML 320 (163.154) V6-3199cc 3.2L SOHC (112.942) - Vehi - Mi...
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://www.alldatadiy.com/alldata/AFI~V3322824~C37165~R0~OF42~N/0/77993956/91311445/91311446/91311449/34
Go
Type search here
Search Web
Radio
Music
Games
Sports
News
Movies
HotStuff
Pop-Up Stopper
On/Off
Options
Auto Parts Needs
CLICK HERE!
CLICK HERE to find one near you!
Photoshop 5.0
ALLDATADIY.COM - ...
ALLDATADIY.COM - 2...
About Working Offlin...
about:blank - Microso...

Popular Information

- Diagnostic Trouble Codes
- Technical Service Bulletins
- Specifications
- Diagrams
- Locations

Vehicle

Testing and Inspection
- Acronyms - Section 0.1
- Explanation of Symbols Used For Tests - Section 0.2
- Connector Layout of Data Link Connector - Section 0.3
- Connecting Hand-Held Tester - Section

Vehicle
Components:

A L L Diagnostic Trouble Codes (DTC)
- Codes by Number
- Codes by System

Relays and Modules
Sensors and Switches
Maintenance
Engine, Cooling and Exhaust
- Engine
- Cooling System
- Exhaust System

Powertrain Management
Transmission and Drivetrain
Brakes and Traction Control
Starting and Charging
Power and Ground Distribution
Steering and Suspension
Heating and Air Conditioning
Restraint Systems
Accessories and Optional Equipment

IT'S NOT A HEAP: Alldatadiy.com provides tech info for cars as far back as 1982.

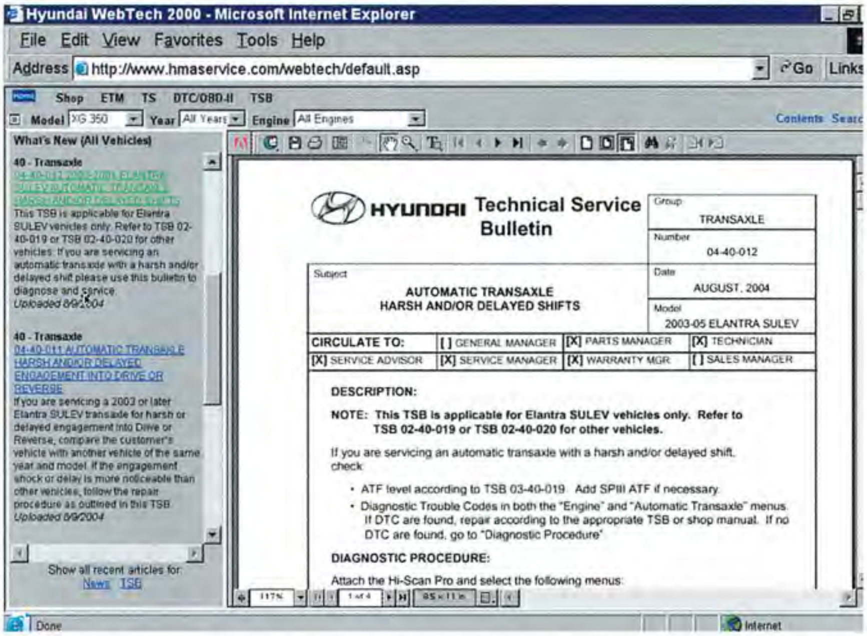
Hyundai WebTech 2000 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
http://www.hmaservice.com/webtech/default.asp
Go
Links
Shop
ETM
TS
DTC/OBD-II
TSB
Model XG 350
Year All Years
Engine All Engines
Contents
Searc
117%
1 of 4
8.5 x 11 in
Done
Internet

What's New (All Vehicles)

40 - Transaxle
04-40-012 2003-2005 ELANTRA SULEV AUTOMATIC TRANSAXLE HARSH AND/OR DELAYED SHIFTS
This TSB is applicable for Elantra SULEV vehicles only. Refer to TSB 02-40-019 or TSB 02-40-020 for other vehicles. If you are servicing an automatic transaxle with a harsh and/or delayed shift please use this bulletin to diagnose and service.
Uploaded 8/9/2004

40 - Transaxle
04-40-011 AUTOMATIC TRANSAXLE HARSH AND/OR DELAYED ENGAGEMENT INTO DRIVE OR REVERSE
If you are servicing a 2003 or later Elantra SULEV transaxle for harsh or delayed engagement into Drive or Reverse, compare the customer's vehicle with another vehicle of the same year and model. If the engagement shock or delay is more noticeable than other vehicles, follow the repair procedure as outlined in this TSB.
Uploaded 8/9/2004

Show all recent articles for:
News TSB

HYUNDAI Technical Service Bulletin

Group	TRANSAXLE
Number	04-40-012
Subject: AUTOMATIC TRANSAXLE HARSH AND/OR DELAYED SHIFTS	Date: AUGUST, 2004
	Model: 2003-05 ELANTRA SULEV

CIRCULATE TO:	[ ] GENERAL MANAGER	[X] PARTS MANAGER	[X] TECHNICIAN
[X] SERVICE ADVISOR	[X] SERVICE MANAGER	[X] WARRANTY MGR	[ ] SALES MANAGER

DESCRIPTION:

NOTE: **This TSB is applicable for Elantra SULEV vehicles only. Refer to TSB 02-40-019 or TSB 02-40-020 for other vehicles.**

If you are servicing an automatic transaxle with a harsh and/or delayed shift, check:

- ATF level according to TSB 03-40-019. Add SPIII ATF if necessary.
- Diagnostic Trouble Codes in both the "Engine" and "Automatic Transaxle" menus. If DTC are found, repair according to the appropriate TSB or shop manual. If no DTC are found, go to "Diagnostic Procedure".

DIAGNOSTIC PROCEDURE:

Attach the Hi-Scan Pro and select the following menus:

THE PRICE IS RIGHT: Hyundai's Web site provides service data free of charge.

Programming Your Car's Computer

As more repairs are done by reprogramming computers, the Web sites will become your source for programs you can download into your PC, then transfer through a generic device called a pass-through to the data bus in your car.

"Me? Reprogram my car's computer? If it's anything like installing software, I'll never get to work." The idea is the same. And there are safeguards to keep you from leaving your car dead in the water. Ford, Chrysler Group, BMW and Mazda already offer reprogramming services on their Web sites. You download a program that checks your car's powertrain computer. For other on-board computers, you have to check calibration charts on the Web site to determine what is available. Because of the delays in finishing SAE standard J2534, low-cost, mass-production pass-throughs are not yet available. When GM and the leading Japanese makes offer reprogramming software over the Internet, you can expect pass-throughs covering most makes to cost $100 to $300—versus $900 to $1600 now.

Web site service information and reprogramming are leveling the playing field for you in your home garage, and for the independent garage you may patronize.

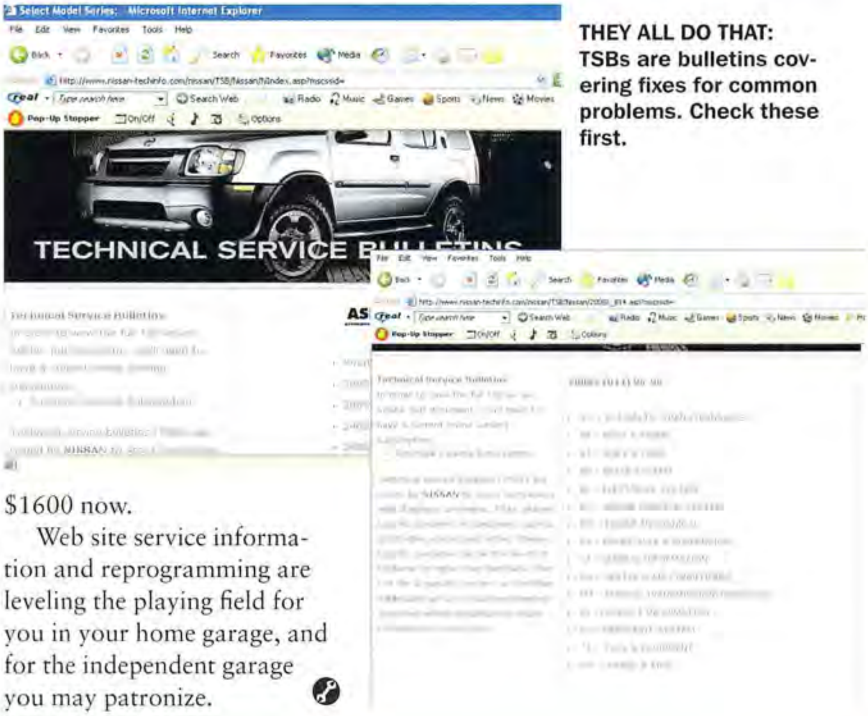
Select Model Series: Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
TECHNICAL SERVICE BULLETINS

THEY ALL DO THAT: TSBs are bulletins covering fixes for common problems. Check these first.

Tech Web Sites

Most tech Web sites (take note of the exceptions below) take the form: www.nameofvehicletechinfo.com. Example: www.landrovertechinfo.com.

- Alldata Do-It-Yourself: www.alldatadiy.com (excludes Honda/Acura and BMW coverage).
- Chrysler Group: www.techauthority.com $20 for a day of TSBs dating to 1992, and all service/diagnostic information to 1996 (a few '95s). Good selection of Chrysler training materials for subscribers.
- Ford (and Lincoln Mercury): www.motorcraftservice.com Free selection of OBD II training books, owner's manuals and towing/camping books. It's $10 per day for TSBs from 1988 on and $20 per day for all repair information on a single model/year/car, which includes service/diagnostic information and an e-mail link to OASIS, the Ford factory-tech help line.
- General Motors: www.acdelcotechconnect.com Free access to "Tech Link," a monthly professional tech service magazine that covers a wide range of GM service issues, with archives back to 2000. Pay $20 per day for bulletins (from 1980) and all service/diagnostic information (1998 on, with powertrain manuals back to '96).
- Honda and Acura: www.serviceexpress.honda.com Buy access separately for $20 for three days. Selecting a subject (such as Fuel & Emissions) gets you information from the service/diagnostic manual, applicable service bulletins and anything on the subject from "Honda Service News," a dealer technician publication.
- Hyundai and Kia: www.hmaservice.com Entirely free, includes good training you can download, as well as service/diagnostic information.
- Mercedes-Benz: www.startekinfo.com A $20-per-day site loaded with technical information going back to the early 1990s.
- Nissan and Infiniti: www.nissan-techinfo.com and www.infinititechinfo.com It's $20 per day for access to both sites (bulletins back to 1987, service/diagnostic manuals since 1994). Manuals come up by subject and read like a book.
- Porsche: http://techinfo.porsche.com (no www)
- Subaru: www.subaru.com (then click on home, and then technical information).
- Toyota and Lexus: http://techinfo.lexus.com and http://techinfo.toyota.com (no www) Get both easy-to-use, info-packed Web sites for $10 per day. Included are diagnostic/repair manuals, wiring diagrams and bulletins since 1990.
- Volkswagen and Audi: https://erwin.audi.de (no www) and www.erwin.vw.com Separate Web sites are run by the same office in Germany. You're charged by the document, so click repeatedly on Search, which gets you to the price and Contents.
- Volvo: www.volvotechinfo.com Sells service/diagnostic documents, wiring diagrams and two types of service bulletins for three-day periods. Manuals are $10, diagrams $5 and the bulletins $13. The sales are only for specific models, except wiring diagrams.

Charging Your Battery

Spring has come and gone. With the yard finally whipped into shape, you've managed to schedule a long weekend up in the mountains—just you and your family and a few days of a lot of nothing. The motorhome starts with the aid of jumper cables, and a few hours of freeway driving should fully charge the battery. The sad part is, when you finally reach the campsite, the kids can't watch TV, the microwave won't work and, horrors, the blender is incapable of making your favorite frozen concoction. The battery hasn't charged much at all. You could run your generator or the motor-home engine for power, but you hate to disturb the untrammeled wilderness by doing that. And besides, the campground has rules against generators after sunset.

Too Close To Home?

The problem you have is common to many RVs, and there are many possible reasons for your predicament. The RV's deep-cycle house battery was left uncharged over the winter, froze and is now useless except as an anchor for your boat. Or, you simply forgot to turn on the battery switch. Or, the alternator on the RV was maxed out by running the a/c fan, GameBoy and stereo and trying to recharge the battery all at the same time.

How Batteries Charge

A conventional lead-acid car battery is filled with interleaved plates of spongy lead and lead dioxide soaking in sulfuric acid and water electrolyte. The sulfuric acid is diluted to a specific gravity of about 1.26 to 1.30, depending on the type of battery. Charging the battery combines some of the sulfuric acid and lead, reducing the concentration of sulfuric acid

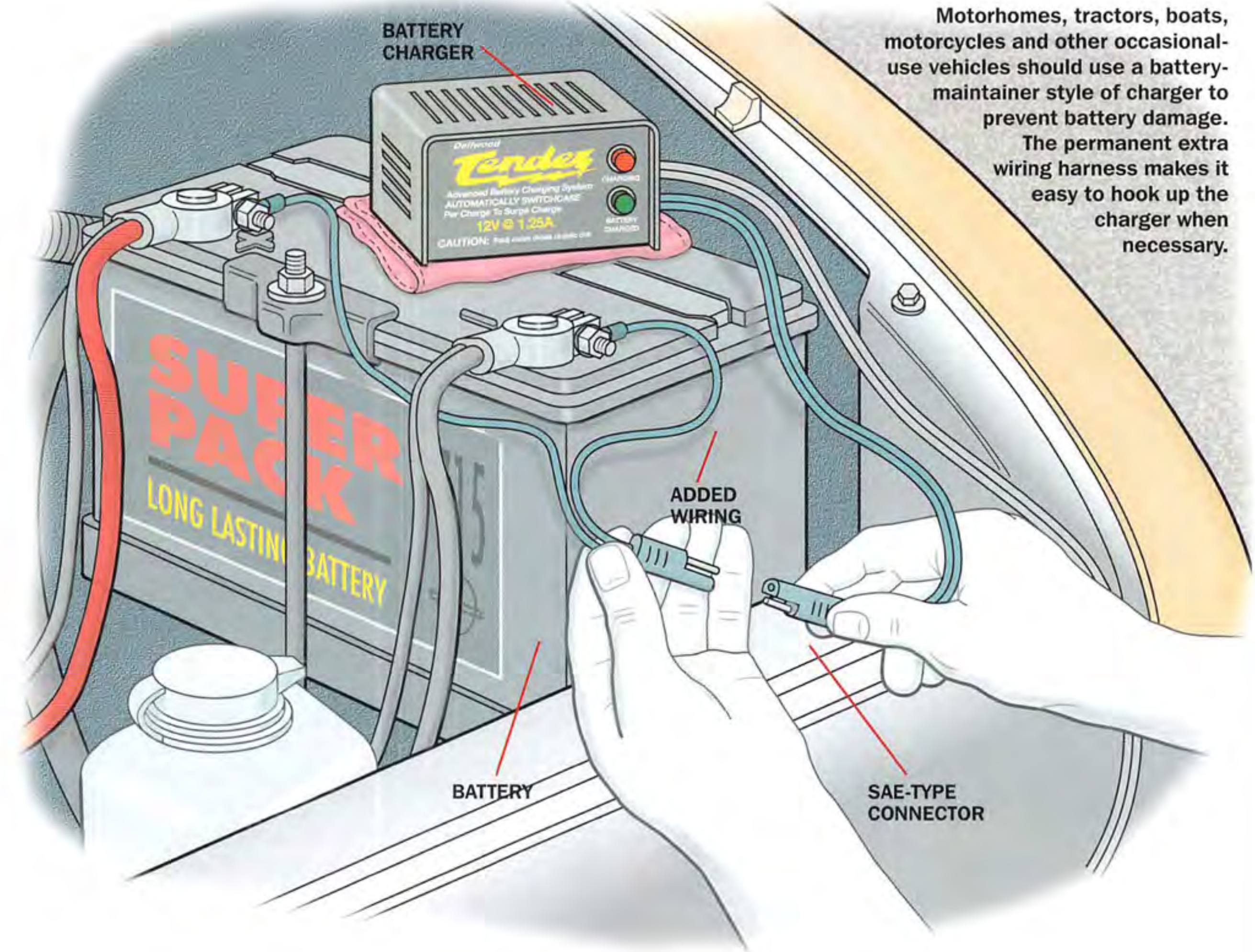

Motorhomes, tractors, boats, motorcycles and other occasional-use vehicles should use a battery-maintainer style of charger to prevent battery damage. The permanent extra wiring harness makes it easy to hook up the charger when necessary.

With hard-to-access batteries, look for a remote jumper cable post.

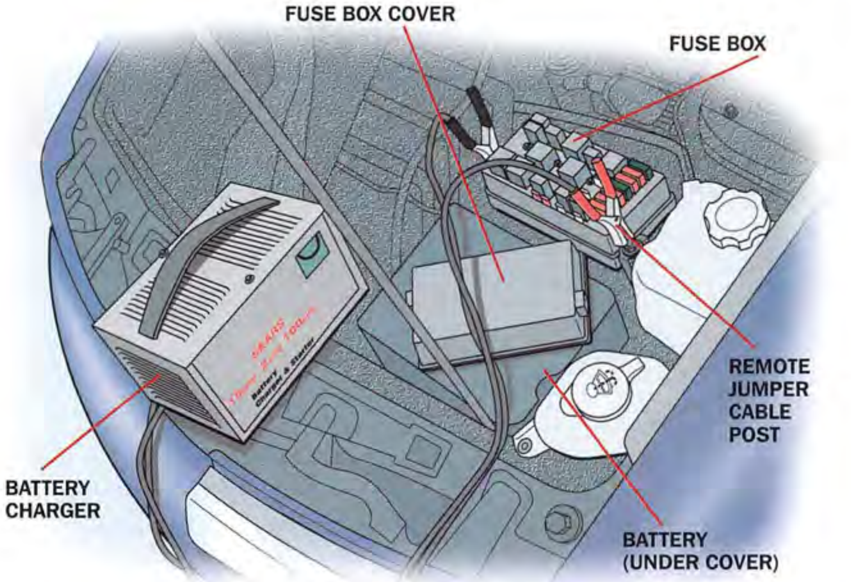

down to about 1.12 and coating the plates with a layer of spongy lead sulfate. Reversing this process gives you back most of that electricity. This means that checking the specific gravity of the electrolyte is a very good way of finding out the actual state of charge of a battery.

It's also fussy and potentially dangerous if the battery electrolyte ends up someplace it's not supposed to be, like in your eye or on the leather upholstery of your wife's vehicle. Also, some maintenance-free batteries have no filler caps and there's no way to check or add electrolyte.

So, we're left to estimate the battery's state of charge by reading its voltage. A fully charged 12-volt battery should have an open-circuit voltage of 12.6 volts. A discharged battery will have a voltage of about 11.2 volts—but these numbers need to be checked with no load and with several hours of rest after charging. And this is at a normal battery temperature of 80°F—colder batteries have a slightly lower voltage. A battery left discharged for even a few hours will convert some of its soft lead sulfate deposits—the kind that readily change back into lead oxide and sulfuric acid—into hard permanent deposits. These deposits interfere with normal battery operation, and can damage the plates.

Be careful when storing batteries in lower temps. A discharged battery contains mostly water and can freeze solid, fracturing the lead plates. Always store a battery fully charged and in an area that stays above freezing.

Uphill, Downhill

As a battery is charged or discharged, electrolyte needs to diffuse from one plate to another. At high or low discharge or charge rates, the acid in the electrolyte develops a gradient, becoming more concentrated next to one plate and less concentrated at the other. Only time will even this out, and that's why you can't quick-charge a battery to any more than about 50 percent of its capacity.

An RV parked outdoors is a good candidate for a solar battery maintainer, such as this type.

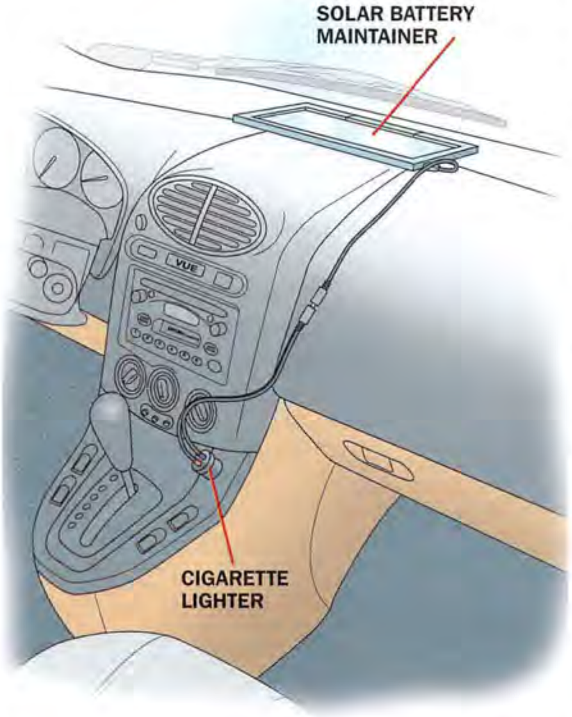

To charge a battery, a charger or your car's alternator needs to raise the voltage across the battery posts. Normal charging voltage needs to remain below about 14.4 volts, which is the level at which gassing of hydrogen and oxygen starts. Your car alternator was never intended to charge a totally discharged battery. Here's what happened on that fateful trip to the mountains. The battery was totally discharged after being left unattended for months. A totally discharged battery needs very high voltage (over 16 volts) or a long charging time at 14 volts to even begin to accept a charge. Once that battery started to accept a charge (but in this case it did not), it would have made the alternator work very hard to pump current into the battery, shortening the alternator's life.

This type of situation calls for an external battery charger. The traditional charger is a simple transformer and diode affair that charges a battery to a constant voltage just below the gassing voltage of 14.4 volts. It will start out charging at a good clip, and then taper off, taking as much as a full day to fully charge a large, dead battery. It's not a candidate for keeping a battery charged, because the voltage it generates will still

HOW IT WORKS

There Are Batteries, And Then There Are Batteries.

We commonly think of car batteries as lead-acid flooded wet-cell batteries that haven't changed substantially since the earliest electric cars used them (the early batteries' cases were made of glass, not plastic). The lead in these is alloyed with antimony to make it hard enough to withstand the vibration it receives in a moving vehicle. In the last few generations, maintenance-free batteries have become the standard. These use an alloy of lead and calcium. In contrast, the antimony used in older batteries promotes the electrolysis of water, lowering the electrolyte level and requiring replenishment. Maintenance-free batteries generally have a higher level of electrolyte covering the plates, and they commonly lack filler caps. Totally discharging one of these types of batteries even once can permanently compromise its ability to take a full charge. Deep-cycle batteries use a purer lead in thicker plates, and are capable of many cycles of near-total discharge. They need periodic replenishment of the water in their electrolyte.

The newest lead-acid battery technology uses a gel electrolyte that won't run out if the battery or vehicle is upended. Another type becoming popular is the absorbed glass mat (AGM) cell. The electrolyte in this battery is held in a spongy mat and can't move around. AGM and gel-cell batteries internally recombine the hydrogen and oxygen from electrolysis so you never have to add water to them.

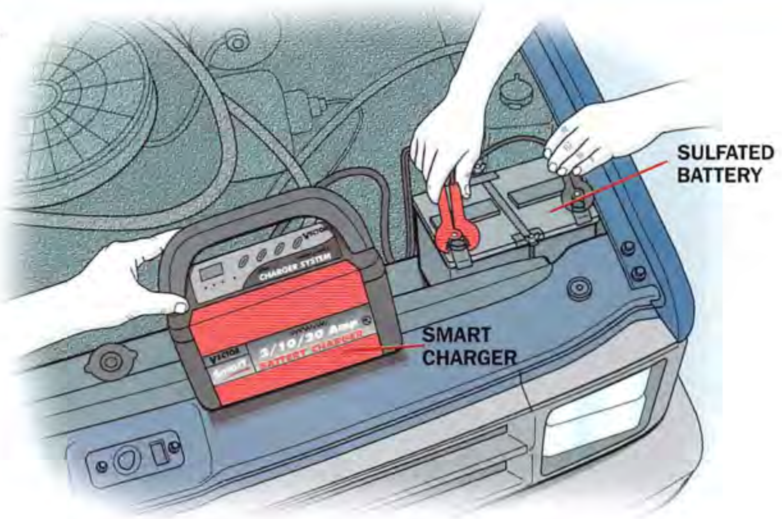

A smart charger will rapidly recharge a dead battery without boiling off electrolyte.

make the electrolyte disappear as gases—an especially bad idea for a maintenance-free battery that can't be replenished. Even a small trickle charger will damage a battery over a few months if left on.

The more modern battery chargers are smart. They will raise the current flow high enough to bulk charge the battery rapidly, then taper off until the battery is fully charged. This reduction is necessary to keep below the 14.4-volt gassing point. The chargers then throttle back to a somewhat lower voltage that won't boil the electrolyte, and periodically check the voltage, occasionally raising the voltage for a few hours to maintain a full charge.

The latest crop of smart chargers even have a battery-reconditioning function that is supposed to break up hard sulfate deposits by blasting them with pulses of high-frequency voltage spikes, returning a badly sulfated battery to near-new condition.1spikes, returning a badly sulfated battery to near-new condition.

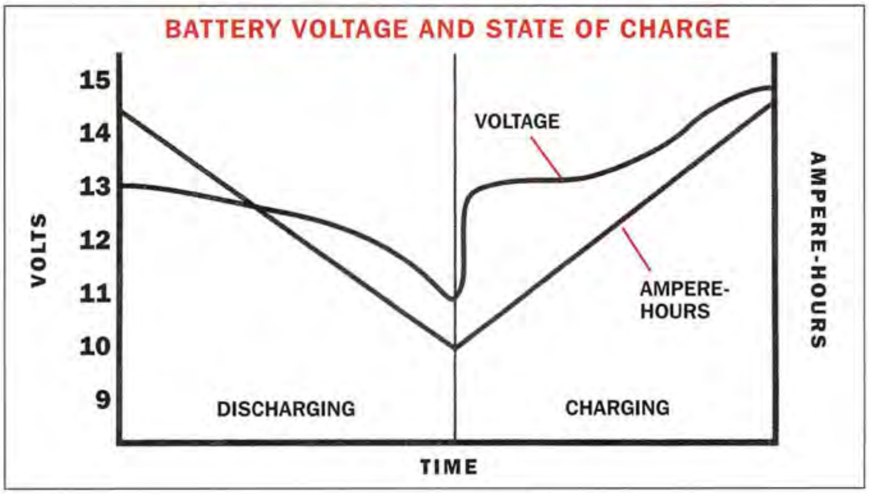

Voltage is a good indicator of a battery's state of discharge, but not during the charging phase.

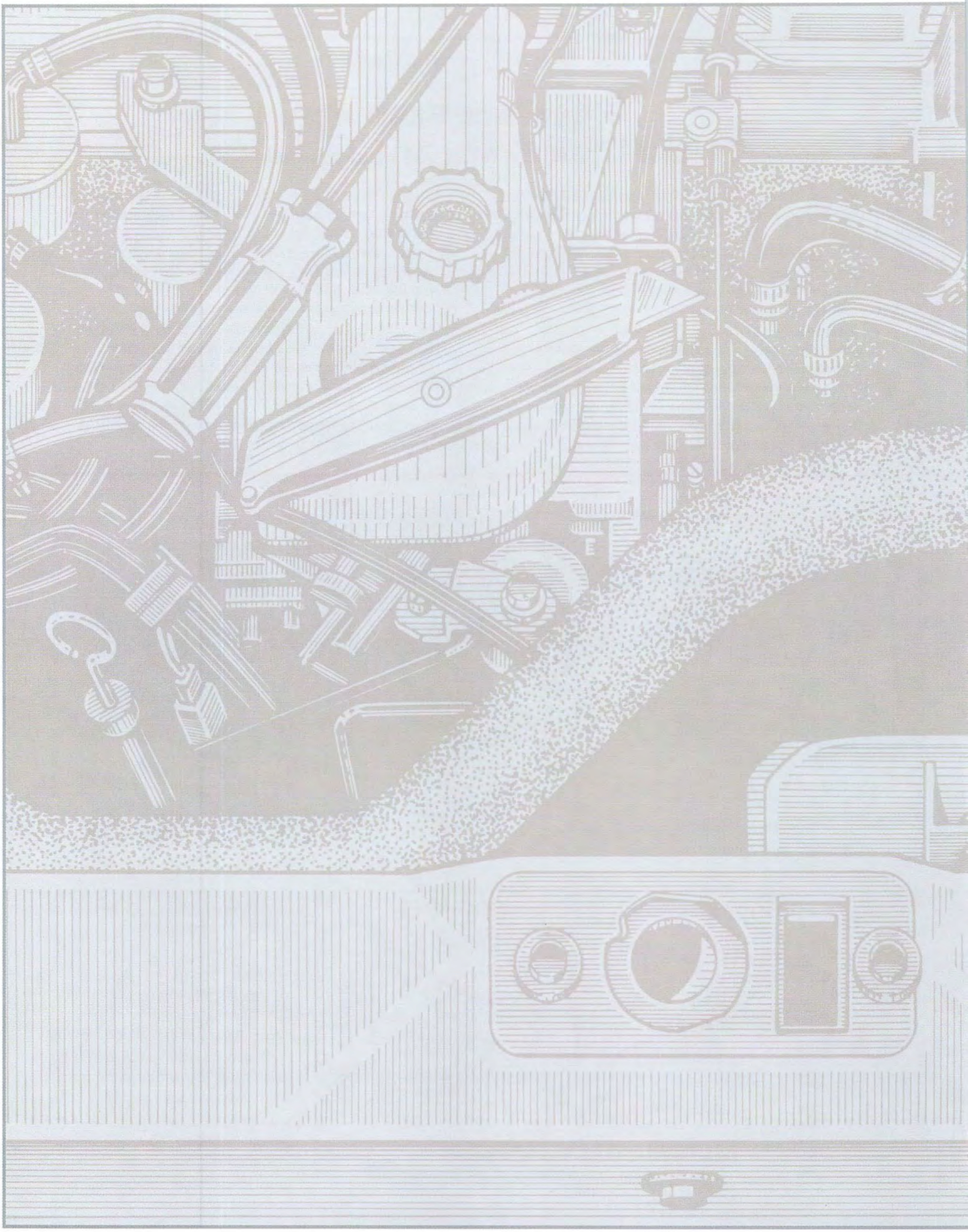

CHAPTER TWO

Engine

Today's Tuneup / **90**
Smoothing Out a Lumpy Idle / **93**
Replacing Your Water Pump / **96**
Curing a Clicking Starter / **99**
Repairing Cooling System Leaks / **102**
Curing Slow Cranking / **105**
Finding Oil Leaks / **109**
Flushing Your Cooling System / **112**
Checking Your Ignition Timing / **116**
Replacing Engine Mounts / **119**
Cleaning Throttle Bodies / **122**
Electric Cooling Fan Problems / **125**
Finding Engine Knock / **128**
Plugging Leaks / **132**
Turning Up Your Heater / **135**

Today's Tuneup

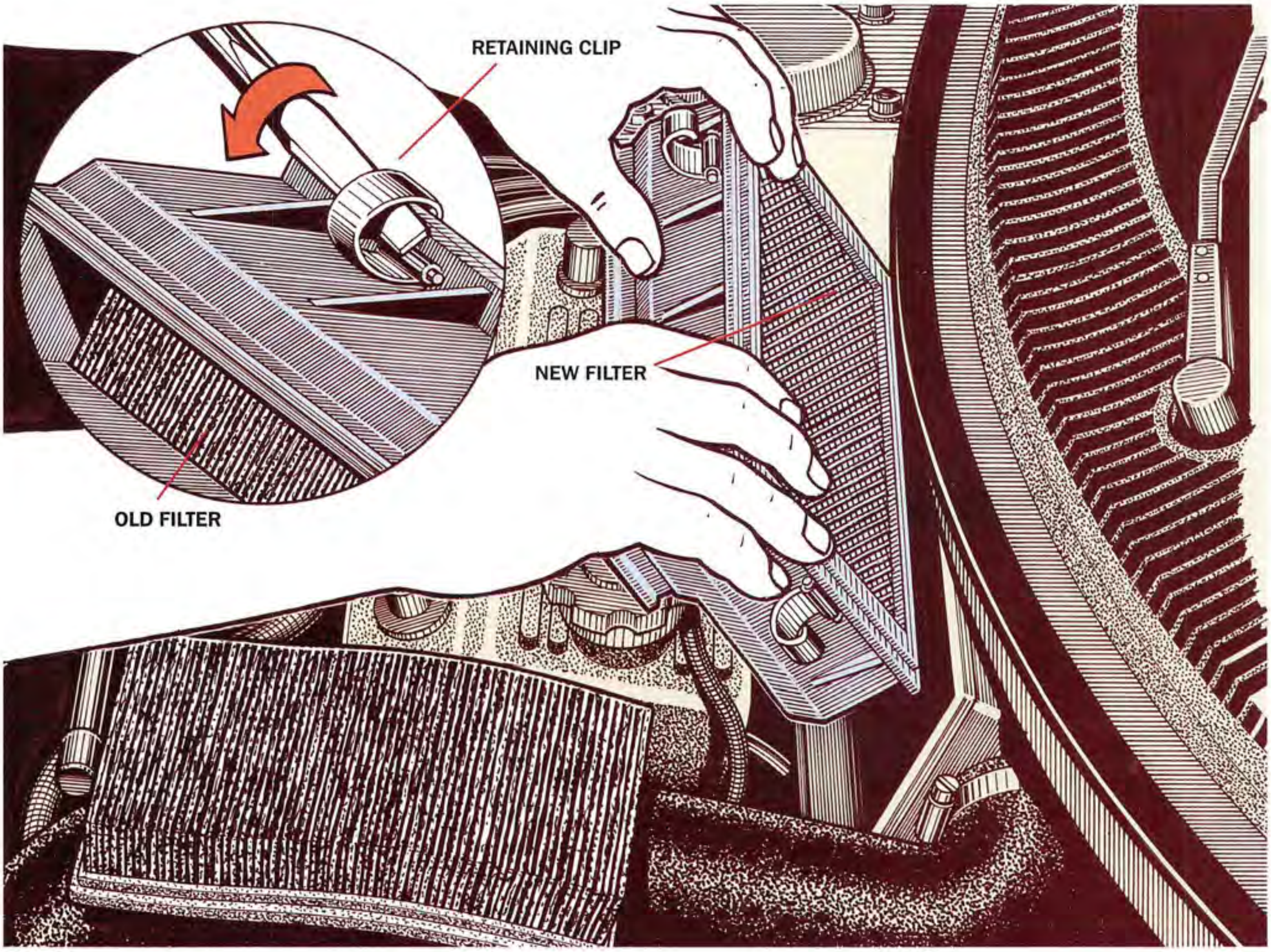

(Fig. 1) Replacing the air filter will require popping off a couple of clips. Slide the new filter into place, be sure it seals properly, and replace the cover.

There's the thruway exit, but you're stuck in the left lane. Hey—a break in the traffic. You flip up the turn signal, floor the gas pedal and try to edge into the right lane. Whoops! Your V8 is embarrassed by a 4-cylinder econobox with the same objective. It was behind you, but it apparently has better acceleration. You hit the brakes, crank the wheel to get back into the left lane, cruise to the next exit miles and miles ahead, and make the time-wasting U-turn. The Wife suggests a tuneup. But we don't tune up today's engines, right? They're computerized and there's nothing to adjust.

Wrong. It's true we stopped replacing ignition points over 20 years ago. If it's really a late-model vehicle, it may not even have a replaceable fuel filter under the hood. The battery is a fill-free design. And there's no ignition timing to adjust. However, while it has changed significantly—even in the last seven or eight years—the concept of a tuneup itself is anything but gone.

Emissions regulations may have tightened, but ambient air is still dirty. So the air filter still plugs up, and the engine gasps for breath until a clean new filter is installed **(Fig. 1)**. Gasoline is unleaded, so spark plugs don't lead-foul anymore, and the new precious-metal designs last much longer. But they aren't guaranteed to last a lifetime, and engine performance often can improve when plugs are replaced well before the owner's manual dictates.

Underhood Checks

A good way to start today's tuneup is to look for trouble codes, using a scan tool. The only codes that turn on a Check Engine light are those that directly impact emissions, provided the computer can detect them, to keep people from getting nervous. Most

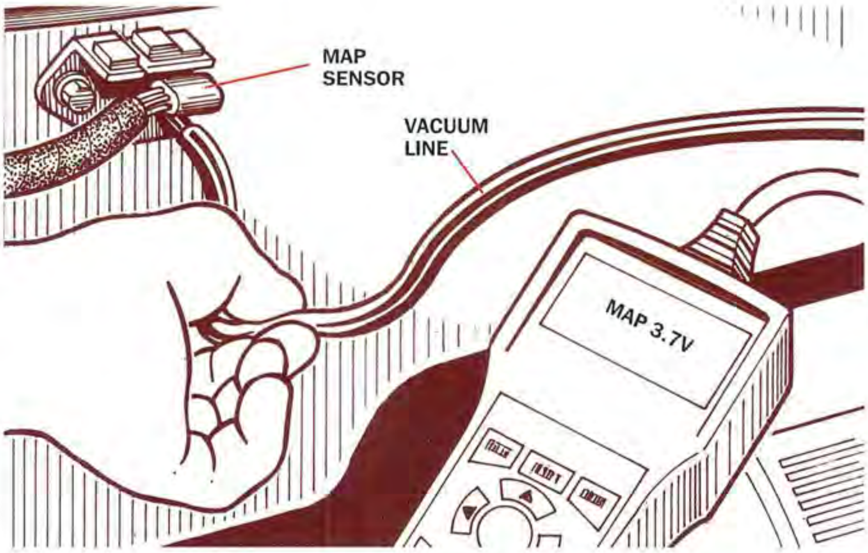

(Fig. 2) Check the MAP sensor with a scan tool by pinching its hose.

codes show up only with a scan tool. On the other hand, "no code" driveability problems are also extremely common.

If you find a trouble code, you should trace the circuit to pinpoint the problem. Sure, it could be just a bad sensor, but it may be a bad wiring connection, chafed wire or damaged hose. You'll need the factory service information to be able to do this.

No trouble code? Inspect under the hood. Look for damaged vacuum hoses that should be replaced, loose connections at any of the sensors and solenoids, tears in the air cleaner ductwork, a disconnected duct or poor spark plug wiring connections.

Test the behavior of the engine controls and their sensors. There still are timing marks on many late models and if your vehicle is one of them, you can check basic timing (at idle) with a timing light. No marks? You still can check ignition timing on the scan tool by reading the ignition advance data item. Look for a steady increase in ignition timing as the throttle is gradually opened from idle to a midthrottle position.

Even if the timing is all right, check the throttle position sensor and intake airflow (mass airflow) sensor readings, which also should show gradual increases as the accelerator pedal is depressed. At the same time, tap on the mass airflow sensor with a screwdriver handle and if the engine hiccups or the scan tool reading spikes, the sensor is defective.

A coolant temperature sensor should show a continuous increase in the reading until the engine is fully warm (195°F to 230°F). A MAP (manifold absolute pressure) engine vacuum sensor should show changes when a pinched-closed hose supplying vacuum to it is released (Fig. 2). The engine should read about 750 to 850 rpm at idle with the engine warmed, and increase gradually as the gas pedal is depressed.

Still no reason for a driveability glitch? You need to access the powertrain computer. It may be located under the hood or under the dash. The most common under-dash location is behind the passenger-side kickpad. Remove the kickpad and, while a helper is slowly accelerating the engine, flex the computer wiring connectors and tap on the computer housing (Fig. 3). If the engine hiccups, there's a bad connection or possibly cracked solder joints on the circuit board. You may need a replacement.

Replacing Spark Plugs

Time was, spark plugs were always visible, even if it took a struggle to reach them. Now they're often recessed into the head, and in many cases there are no plug wires. A mini ignition module, perhaps with the coil built into it (or into the plug boot), is used instead. You may have to remove a cover that holds the ignition modules and plug boots for access. Whatever the design, very carefully make any necessary electrical disconnection at the plug prior to plug removal. If there are plug wires, grasp and lift by the plug boots. Never pull on a plug wire or it may separate internally. Look inside the boot and if it appears soaked in oil or cracked by heat, replace it. If the ignition coil is built in, expect it to be pricey.

Inspect the plugs. If they all have a coating of black carbon, that indicates a rich fuel mixture and likely a bad oxygen sensor. A scan tool should be able to pinpoint a malfunctioning oxygen sensor even if it didn't log a trouble code. If there's just a single carbon-blackened plug, you probably have a leaking fuel injector.

When you're ready to install the new plug, apply a thin film of antiseize compound so the new plug doesn't heat-seal in place. Use a torque wrench to tighten the plug.

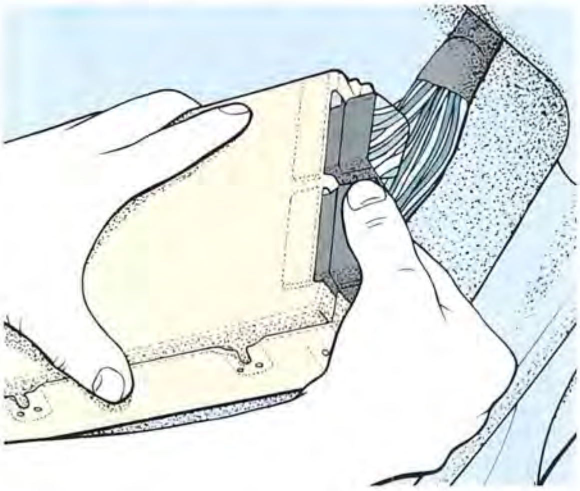

(Fig. 3) Wiggle the wiring connectors while the engine is running to check for misfires.

(Fig. 4.) You may not be able to clean the pintle on a recessed EGR valve like this one.

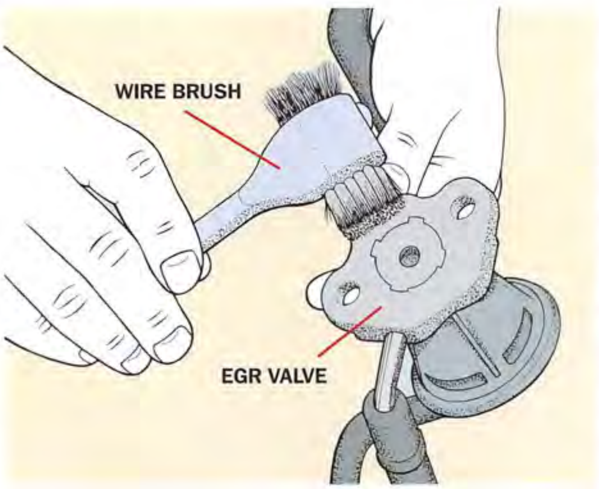

Exhaust Gas Recirculation

Exhaust gas recirculation (EGR) valves may stick open, fail to close completely or just not operate smoothly. If so, the engine will hesitate and may stall, fuel economy may drop, and emissions may be affected—even if you pass the state inspection test. The EGR valve meters some exhaust gas back into the cylinders to lower peak temperatures of the air-fuel charge during combustion. This not only reduces a key pollutant (oxides of nitrogen) but often improves fuel economy. The EGR valve, typically a diaphragm-actuated device with a pin-type valve, is pulled open by engine vacuum, but that vacuum must be precisely regulated. Too much vacuum results in too much exhaust gas flow, which can cause the engine to lose power, even stall.

Late models with OBD II computers have sophisticated strategies to detect severely malfunctioning EGR. However, this basic check will work on all vacuum-controlled EGR systems. Locate the EGR valve and if it has a vacuum hose connection, unplug it. Run the engine at idle, and connect a spare hose from a source of engine vacuum or use a manual vacuum pump and apply vacuum to the hose neck. The engine should slow significantly, probably even stall, if you apply full engine vacuum (17 to 21 in.).

If there's no significant change in engine idle—in fact, if the engine has been idling rough—the valve may be sticking open. Remove it and if you can see a heavy accumulation of deposits in the port, and if it's an exposed pin-type valve, clean it with a wire brush **(Fig. 4)**. If the end of the pin is recessed in the port, it can't be cleaned. Replace it.

Battery Terminals

Clean the battery terminals, then reinstall and tighten. Today's batteries—top-post and side-terminal—are subject to continuous drain with the engine off, to keep alive the memories of many computers (from powertrain to car radio). Just a slight coat of corrosion—perhaps too subtle to be visible—can reduce battery charge. Disconnect the cable terminals. On a top-post battery, brush around the post and inside the cable terminal **(Fig. 5)**. If the post or cable terminal is badly corroded, replace it with a premium terminal, the kind that often includes a cable section. Forget the cheap screw-together terminals—they'll become severely corroded and cause more problems than they solve. With a side-terminal battery, brush both contact faces even if they look clean. Make sure the bolt threads in without a feeling of looseness, and if there is any, don't try to muscle the bolt supertight with a wrench. You're more likely to make things worse.

Where To Get Service Information

Diagnosing failures that produce trouble codes, and finding out normal readings for engine sensors, is not subject to rule of thumb procedures. You need the latest factory diagnostic sequences and specifications. Vehicle makers have Web sites with this service information, and you'll be able to access the data for a day at a time. Many sites charge for this information—you can pay using a credit card. At present, General Motors information is available at www.acdelco.com (click on ACDelco TechConnect). Or you can get a low-cost subscription from www.alldata.com (a leading source for professionals, with information on all makes). In addition, AutoZone offers free scan tool connections and readouts at its stores.

(Fig. 5) Clean the battery posts and cable clamps regularly.

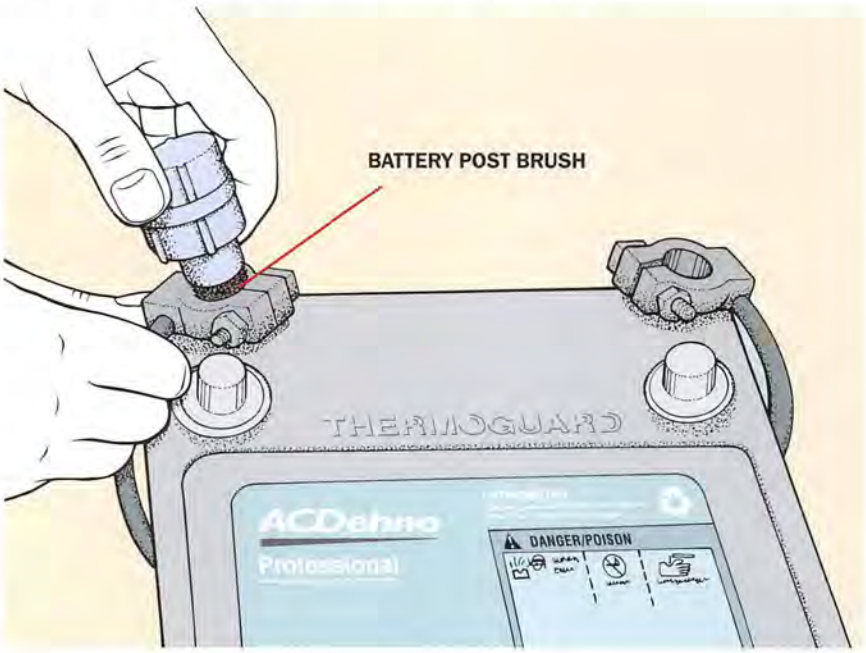

Smoothing Out a Lumpy Idle

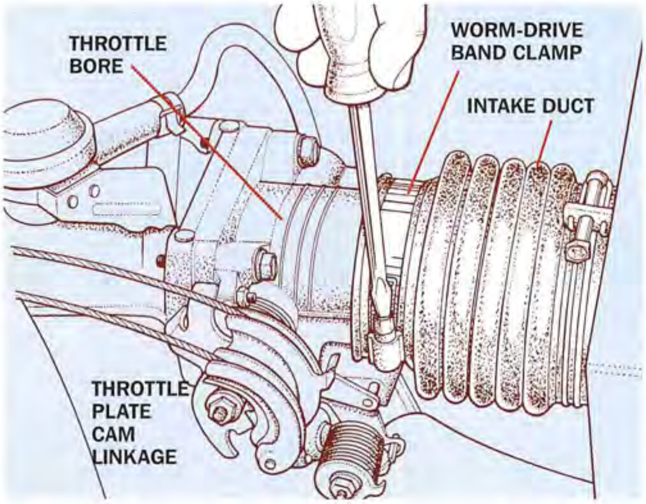

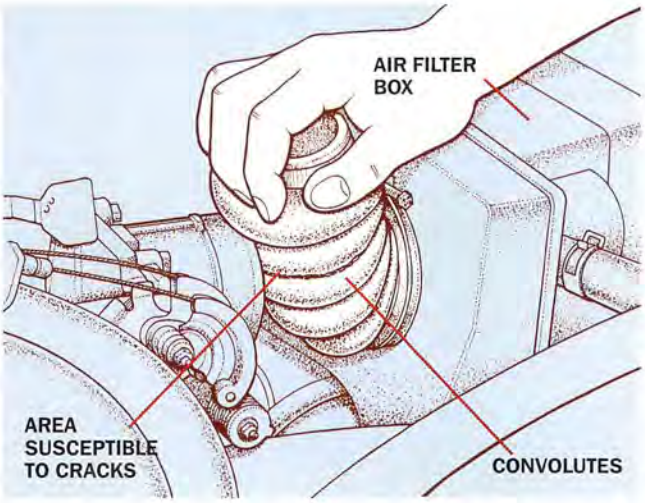

(Fig. 1, left) Make sure all the clamps sealing the intake duct are tight.
(Fig. 2, right) Pull back the intake duct to check for hidden cracks underneath.

Fearful pedestrians glare at you as they pass in front of your bumper at crosswalks. Out of embarrassment, you try to ignore them by glancing around nonchalantly, stroking your hair as you look in the rearview mirror or tuning the radio. Meanwhile, the size of your right thigh is growing to Dan Jansen proportions from the force being applied to the brake pedal. Your engine's idle speed is surging up and down, creating crescendos that would make Pavarotti proud: WAAAAaaaah, WAAAAaaaah, WAAAAaaaah! It's all you can do to keep your maniacal car from lurching forward into a lawsuit.

There are a number of generic steps you can take to see what's ailing your wheels. And if there ever was a time to step back and remember the basics, this is it.

Pop quiz: What is an internal-combustion engine?

It's basically little more than an air pump. The more air that gets in, the faster the engine runs. And airflow controls idle. By allowing a certain volume of air to bypass the closed throttle plate(s), idle rpm can be maintained at a healthy level. Even to accelerate, we don't "step on the gas," we "step on the air" by opening the throttle plate farther. Fuel is added a nanosecond later, in response to the greater intake airflow.

Check Your Intake Tract

If your engine uses a remotely mounted air filter in an air box, inspect everything from there back to the throttle plate(s). At the air box, check all hose connections and make sure the clamps are tight **(Fig. 1)**. Replace the air filter if it's so dirty that light from a 100-watt bulb doesn't pass through the element. Make sure the new one seats properly in the air box and that the cover sits flush and clamps down evenly. Follow the intake tract toward the throttle plate, tightening all the clamps as you go. If there's an inline mass airflow meter in the tract, take extra care to examine its connections for leaks.

Big, convoluted rubber tubing-style intake tracts are susceptible to developing cracks between the convolutes on the underside of the tube. They're generally not visible unless you remove one end of the tube and bend it back to get a good look below **(Fig. 2)**. If the engine controls measure airflow by means of a manifold absolute pressure (MAP) sensor, this type of leak shouldn't affect idle quality. But the hole still needs to be sealed, or else dirt and debris will find its way in. If the engine uses an airflow sensor, this type of leak tends to audibly reveal itself when the engine torques forward on its mounts and opens wide the convolute crack.

Then the engine gets an unmetered gulp of air and revs up on its own. The idle speed control may try to catch the speed burst by closing down the bypass channel. Then, when the engine returns to its normal position and the convolute crack closes, the idle speed is too low. In response, the idle controls may open up the channel again to raise the idle. This scenario can turn into a cycling condition that produces a lumpy, rolling idle.

If you don't mind getting your engine bay a little messy, another quick way to go about checking for intake tract and runner leaks is to simply spray carburetor cleaner around connections and the intake manifold while the engine is running **(Fig. 3)**. If you get an rpm change when you spray, it means the vapors are getting in somehow. So you have to play spy to find the leak. Don't spray near the distributor—if you have one—because there's a chance the solvent will ignite and leave you minus your eyebrows and nose hair.

Keep Your Bore Clean

When you get to the last clamp at the throttle bore, unscrew it and remove the intake tract. Take a look inside the bore with a flashlight. If the muck is so thick that it's a wonder the throttle plate can move at all, you've found a major cause of your lumpy idle. In addition to the idle air-bypass channel, a small amount of air must be able to pass around the throttle plate itself. When blowby vapors from the PCV system (and EGR exhaust gases) sludge up the bore over time, the idle air-bypass function is seriously affected.

Steal an old toothbrush from your kids (so your wife doesn't blow a gasket) and pick up some noncaustic fuel-injection intake cleaner at an auto parts store. With the engine off, spray some cleaner in the bore and start scrubbing with the toothbrush. Pay particular attention to the circumferential area where the throttle plate sits when closed. Also clean both sides of the plate and its edges.

If the externally mounted idle air-bypass valve is easily removed and its channel easily accessed, try to get cleaner to pass through the channel into the bore. Be sure to clean the valve's pintle tip, too **(Fig. 4)**. (Also see page 122.)

Wash Down the Residue

With the bore and channel clean, stick the intake tract back on and start the engine. If the engine doesn't use an airflow meter, you can pull the tract off with the engine running and spray some more cleaner in the bore to wash down the residue. Goose the throttle a couple of times. Then tighten the clamp and let the motor idle so that the engine-management system can relearn the parameters necessary for increased throttle plate air bypass.

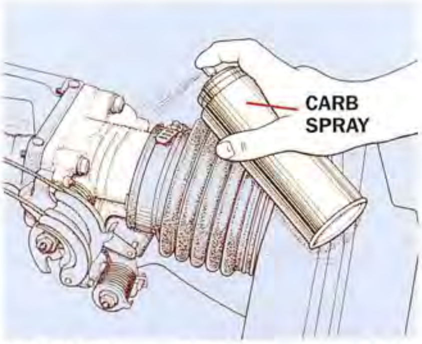

(Fig. 3) Spray carb cleaner around the throttle bore to find air leaks.

If your engine does use an airflow meter, it'll probably stall when you pull the intake tract off the bore. Get it started again and just pull the end of the rubber tract back a bit with one or two fingers to spray some more cleaner down in the bore. The engine will stumble for a second, but that's okay. Whatever you do, don't spray the cleaner into the tract before the airflow meter. You could damage the meter.

Single- and dual-point injection throttle bodies typically don't sludge up much because they're up top on

HOW IT WORKS

Engine Management

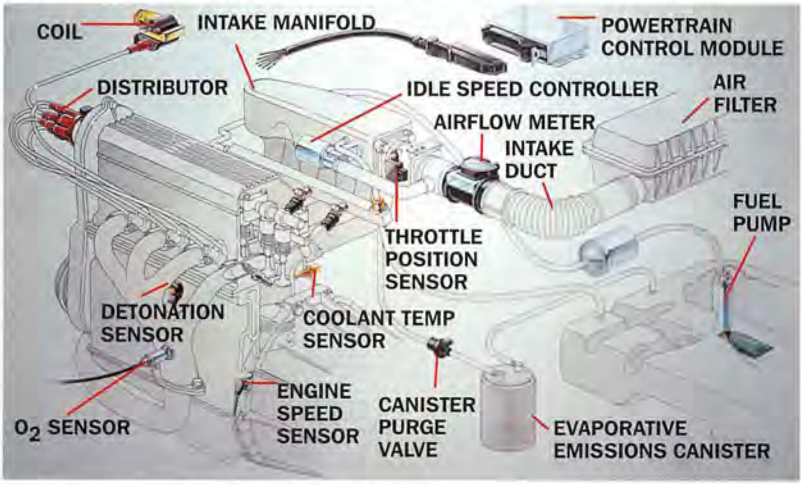

An automotive computerized engine-management system works like any computer as it controls idle speed. The central processing unit relies on various inputs to calculate necessary outputs. On modern motors, however, the inputs to the processor, or powertrain control module (PCM), are called sensors. The outputs are called actuators. And the PCM is programmed to control the actuators under any condition that the sensors deem necessary.

Key sensory inputs of most engine-management systems include engine speed, coolant temperature, crankshaft position, intake airflow, manifold vacuum, throttle position and exhaust oxygen content. Many systems go further, factoring in such inputs as camshaft position, barometric pressure, intake air temperature, detonation detection, EGR valve position, misfire detection, engine-oil temperature, power-steering pressure, air-conditioning pressures, gear-lever position, vehicle speed, automatic-transmission-fluid temperature, catalytic converter efficiency, system voltage and others.

Key actuators on many systems include the fuel injectors, idle speed control motor, EGR valve, evaporative canister purge, ignition coil timing and dwell (saturation time), torque converter clutch, smog pump diverter valve, cooling fan, alternator output and fuel pump.

the engine and PCV vapors flow in below them. However, if the plate(s) looks really dirty, it's okay to hit it with some cleaner. Just be careful not to drown the fuel injector sitting directly above the plate.

Assuming no one has ever played with the base idle settings (the screw may be sealed) and there's nothing wrong with the powertrain control module's programming (the malfunction indicator lamp has never illuminated), that covers the basic idle controls. Unless the manufacturer has issued a software update, today's engine-management systems are smart enough to take it from here and continue to provide a smooth, carefree idle until the next time that throttle bore has janitorial needs.

Loaded to the Hilt

When you consider the number of belt-driven accessories hanging off the typical engine today, in addition to all the mechanical and electrical loads, it's a wonder the pistons can continue pumping at all. That's why some vehicles today have 120-amp alternators and 140-amp fuses.

In rainy, cold weather at night, your alternator is working overtime to make enough amperage to power everything. Problem is, it's also trying to stop the engine from turning. Here's where your idle speed control really shines. As soon as alternator output drops to a certain level, the voltage regulator does its thing and the engine-management system sees the need for increased idle rpm to keep charging-system voltage between 13½ and 15 volts.

(Fig. 4) Pull the idle speed controller out of its bore to inspect the pintle tip, spring and air channel.

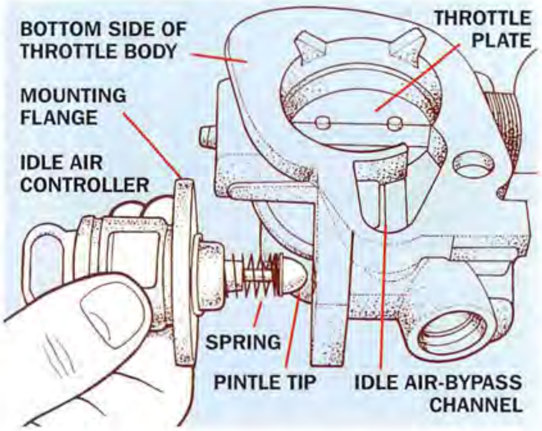

If a signal gets lost through a short or an open circuit or an intermittent connection, however, you're likely to get a lumpy or just plain low idle. Here's where you really need the specific service manual and wiring diagrams for the vehicle, because there are too many variations in form and function to discuss these systems in general terms. Some simple things you can do, however, include visually checking for corroded—and even burned—connections at the alternator and battery terminals. Eyeball the condition of all the fusible links around the battery, and check for fouled spark plugs (Fig. 5). Make sure the accessory belt(s) is tight and unfrayed.

Every time the a/c compressor kicks in at idle, engine speed would drop a couple hundred rpm were it not for the idle speed controls keeping things at an even keel. Here again, intermittent connections and a faulty pressure switch or two could cause the climate control to cycle in and out. A low or contaminated refrigerant charge might do the same thing.

Sometimes there's a pressure switch monitoring your power steering system, especially on 4-cylinder engines. During a tight parking-lot maneuver, when power-steering pressure skyrockets, the engine controls take the reins and bump open the idle speed control motor so your engine doesn't stall or bog down to the point of misfiring. A bad connection or leak at the switch would affect this system and possibly lead to a hunting idle. Periodic power-steering-system flushes go a long way toward preventing clogged switches and orifices.

An engine with high mileage that's tired and worn out may not pull a healthy vacuum of 18 to 20 in. Hg at idle (closed throttle plate) anymore. That means the MAP sensor will always read the engine as under load (low vacuum=high voltage) and—just doing its job—inform the powertrain computer to add more fuel. When the oxygen sensor picks up the rich mixture in the exhaust stream, it will call for a leaner mixture. Common strategy is to open the idle air-bypass valve to let some more air in. But an engine that's just on the borderline of wheezing may intermittently "loosen up," leading to a roller coaster idle. It may have trouble breathing when cold, but once warm—with expanded piston rings, gaskets and the like—show a perfectly healthy intake vacuum.

What it all boils down to is that there's a whole lot more than just a carburetor's throttle stop screw controlling your engine's idle speed these days. Simply understanding the system, however, is half the battle toward finding and fixing the offending troublemaker.

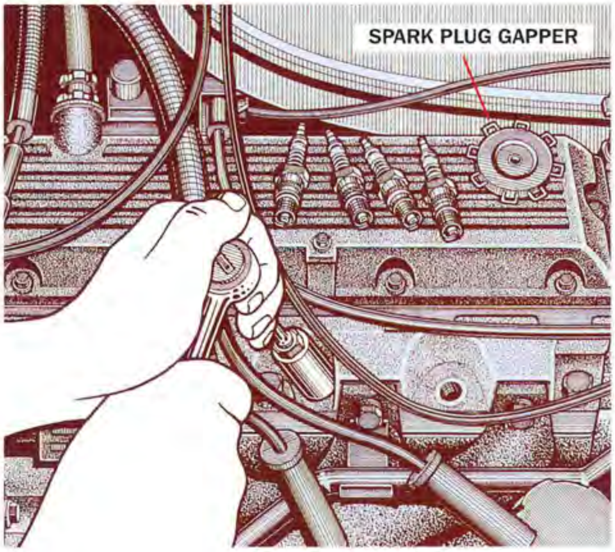

(Fig. 5) Fouled spark plugs can lead to misfiring at idle. They can be cleaned and regapped or simply replaced.

Replacing Your Water Pump

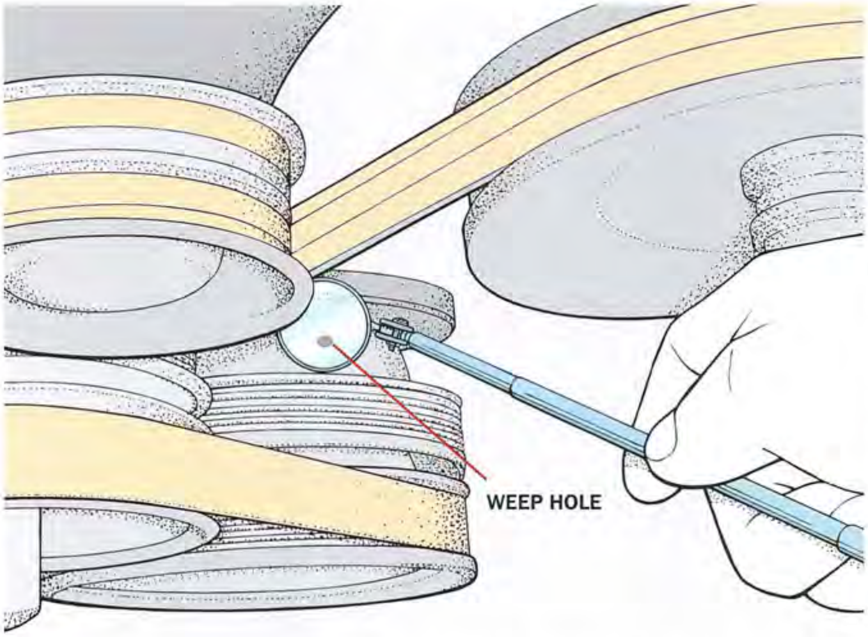

(Fig. 1) Wet liquid seeping from the weep hole signifies the end of the service life of the seals in your pump.

Bombing down the interstate, you glance at the odometer and realize you've just reached an automotive milestone: 100,000 miles with no major repairs. Time was a car that had passed its belly over that much real estate was considered pretty much used up, but yours still runs great and looks practically new. Amazing how technology has advanced.

Then you notice the temperature gauge. Holy smokes—pegged! You put it in neutral, coast off onto the shoulder and shut her down. There's that maple-syrup-spilled-on-the-radiator odor again—you've smelled it before, but filed it under "things to think about later." Sooner would've been better—now you're walking.

Seal Deal

The biggest change in water pump design occurred decades ago when the spring-loaded mechanical seal was adopted. However, its rubber parts may disintegrate if the engine overheats, and its polished sealing faces can wear and warp if the engine is run dry. Typically, pumps will start leaking catastrophically shortly after a boil-over.

This kind of failure can be worse than it sounds. Besides the vastly expensive internal engine damage that running without coolant may cause, a leaking seal can wash away the shaft bearing's lubricant, perhaps resulting in a seized shaft, and a flying fan or belt pulley can destroy the radiator or even dent the hood.

So, leakage is the No. 1 failure. Noise is second, and is always indicative of a terminal condition. While service literature on water pumps often will show a picture of a badly eroded impeller that contributes to overheating, technicians say that's not as common as it once was. Another possible problem with the same consequences is an impeller that's come loose from its shaft. Erosion of the inside surfaces of the pump chamber caused by cavitation (a weak cap or a chronically low level, perhaps?) can open up the working space and reduce flow, as can corrosion from weak antifreeze in the coolant mix.

Evidence Gathering

If you start to smell the distinctive odor of engine coolant, or you notice that the level in the overflow bottle is dropping rapidly, it's time for a careful exam. (If you're lucky, it may be just a leaky hose connection, but look the radiator over, too. If the smell is strong in the passenger compartment and the windshield tends to steam up, think about the heater core. In cases where none of the above is the culprit, better check out the water pump.)

First, use an inspection mirror and a good light to view the vent hole that's at the bottom of the pump casting's nose **(Fig. 1)**. Or, support the front of the car safely on jackstands and look up from underneath. All seals are supposed to weep slightly (a little coolant is needed to lube the faces), but drips mean you'd better go shopping for an estimate.

Next, grasp the fan or water pump pulley and see if it rocks from side to side. If there's anything but slight movement, the bearing is on its way out. Also, you can remove the belt and see how the bearing feels as you rotate the pump shaft **(Fig. 2)**. Roughness isn't acceptable.

A low-flow situation that results in hot running can be hard to diagnose. Drain the level down to the top of the radiator tubes, get the engine hot, and then shut it off for 10 minutes and let it heat soak to make sure the thermostat is wide open. Now, fire it up again and run it at 3000 rpm. On Japanese cars using the radiator fill neck on the top tank, look down into it with a flashlight and you should see strong circulation **(Fig. 3)**. Another possibility is to squeeze the upper hose to feel

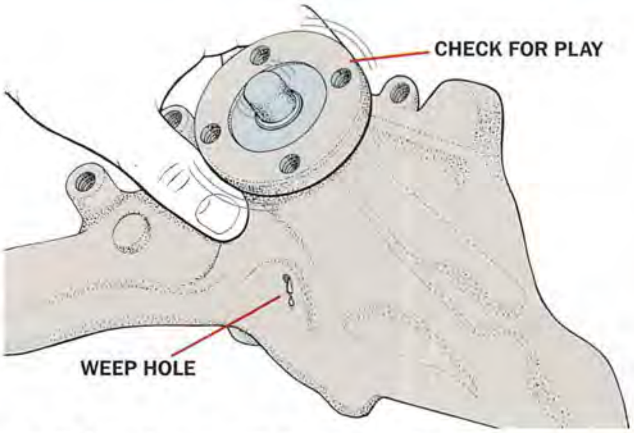

(Fig. 2) Almost any radial play in the water pump shaft is grounds for immediate pump replacement before something fails.

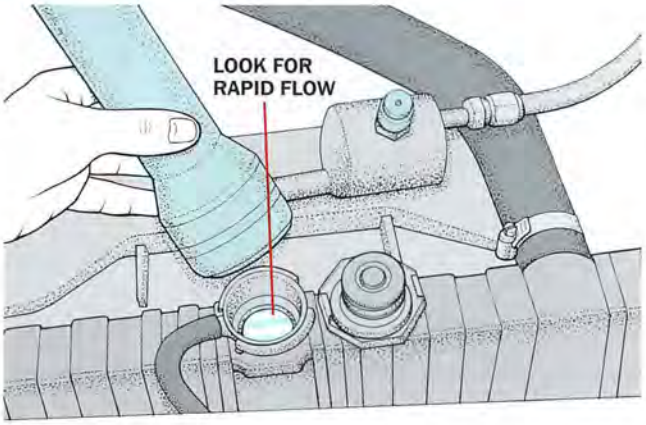

(Fig. 3) With the engine warm enough to open the thermostat, but not hot enough to pressurize the system, look for vigorous flow at 3000 rpm.

for flow, but that's pretty subjective. Unfortunately, there's no good way for the do-it-yourselfer to differentiate a weak pump from a clogged radiator.

Major Undertaking?

On some front-wheel drives the job is a horror story. Make sure you resist the temptation to get into this repair if you're not prepared to invest the time required. Find out the flat-rate hours by asking a local garage owner or dealer service manager. It's a job that could take several hours, depending on what you drive, particularly if the pump is recessed into the engine block.

The first steps in removing any water pump are to let the engine cool off completely, and then drain the cooling system by either opening the radiator petcock or disconnecting the bottom radiator hose and removing engine coolant drain plugs. Next, do whatever is necessary to remove the accessory drivebelts. If the car has a longitudinally mounted engine as found in every rwd and some front-

HOW IT WORKS

Water Pump Shaft Seals

Two perfectly flat rings, one stationary and the other rotating with the pump shaft, are pressed together by means of a coil spring. The rings may be made of carbon or may be ceramic, phenolic, porous bronze, cast iron, etc., in any combination. This allows only enough seepage to keep the elements lubricated. It has a weak link, however, in the form of its rubber parts—the bellows that seals the spring and the rubber seat cup between the rotating element and the shaft. If the engine is ever run dry, the temperature of the pump is apt to rise far beyond what the rubber can survive, and a leak occurs. Another possibility is warpage of the sealing elements, also from overheating.

There's conflicting evidence on whether silicates and phosphates from antifreeze, or other hard particles such as casting core sand, can actually damage the seal faces. Engineers have told us the running clearance is way too small to admit solids of any appreciable diameter.

Since cars are lasting longer than they used to, and since we've become such a litigious society, carmakers are working harder than ever to make water pumps last. After all, when a pump goes you've got to get out and walk, which exposes you to dangers that horrify auto company lawyers.

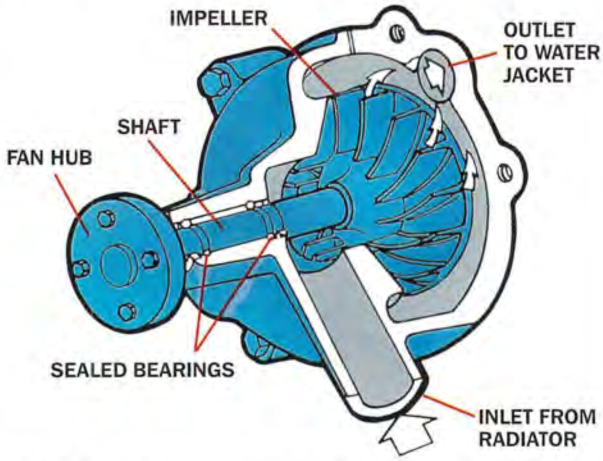

So, there has been a push for water pump seals that will rarely, if ever, fail. One design that looks promising is from Michael Ostrowski and John Crane International, the leading supplier of conventional water pump seals. Called the Advanced Metal Diaphragm seal, it uses a flexible stainless steel diaphragm that acts as both the spring and the bellows, and incorporates an improved method of mounting the seal faces.

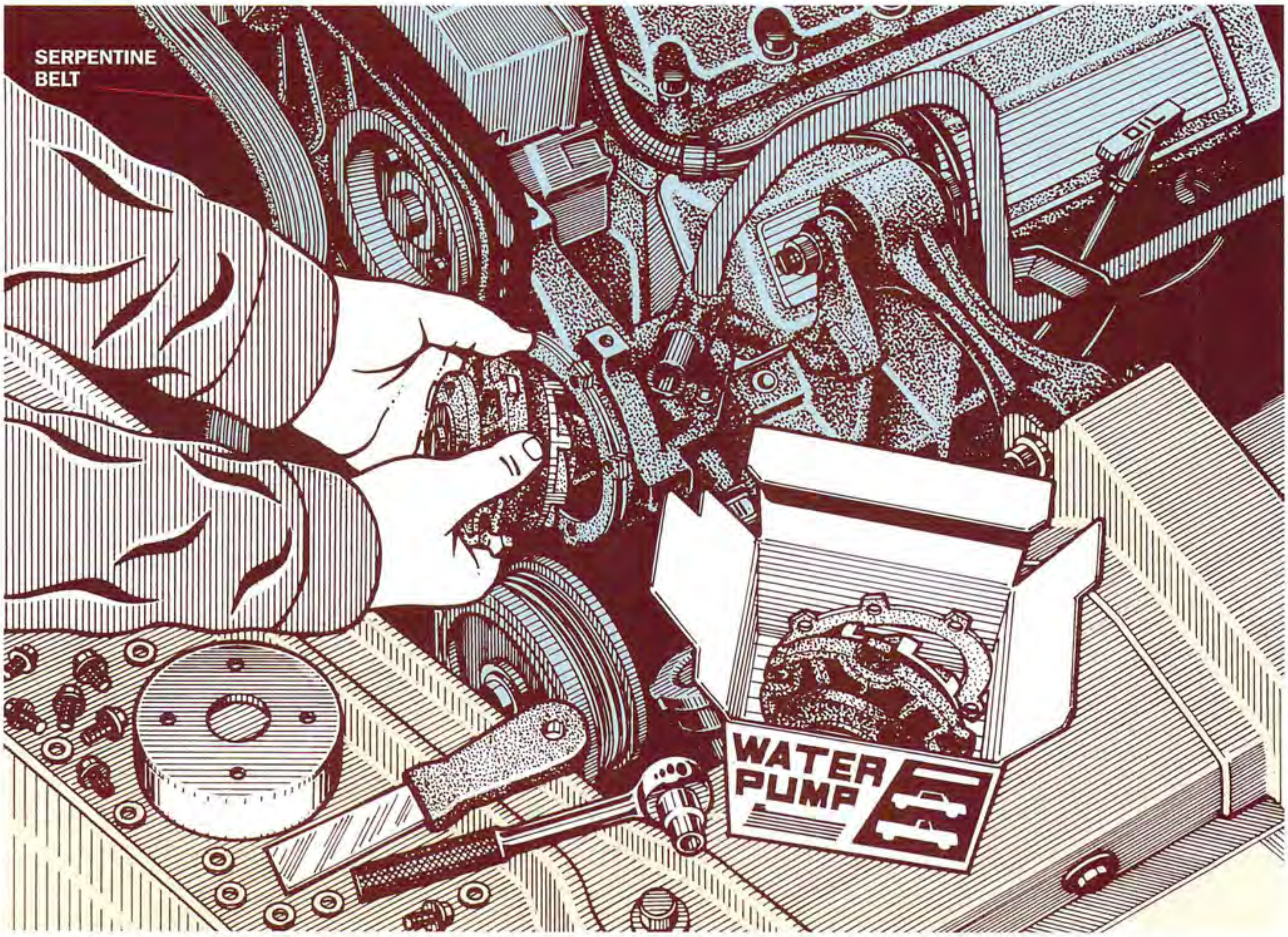

(Fig. 4) This is an easy one, because it's near the top. Your job would be a lot tougher if the pump were near the bottom.

drivers, remove the air shrouds (in some cases, the radiator, too), and then the fan and its clutch, which bolt to the front of the water pump shaft.

Now you can start on the pump by disconnecting the hoses. Those for the heater will probably be stubborn, so you may need to split their ends with a utility knife (if they're long enough, you might be able to trim them off square and reuse them).

Extract the pump-to-engine bolts and keep them and any brackets they retain in strict order or you'll regret it at assembly. There may be hidden bolts, so take a look at a diagram.

The pump should come off with a good tug **(Fig. 4)**. If not, make sure you didn't miss any bolts, and then tap on the inlet or outlet neck with a mallet or a block of wood and a hammer. Don't use a screwdriver to pry the seam open or you'll make a nick that the gasket may not seal.

Pumps seal against the engine or backing plate with a gasket, an O-ring or RTV silicone. Get the mating surface clean with a scraper.

If you must use silicone sealer, use one rated for automotive use. It should also be labeled low-volatility. Outgassing from the curing sealant can poison oxygen sensors—so leave the bathtub caulk with your plumbing tools.

We should mention a potential problem. Suppose you're installing a water pump on a car with a serpentine belt. Engines on many older models were equipped with a regular ribbed or V-belt, but the water pump for the old model may fit physically on the new one, if you change over the pulley. Can you guess the possible mistake? Think about direction: The serpentine belt might drive the impeller in the opposite direction from that of the V-belt, so you could end up installing a pump that runs backward, causing a seemingly incurable overheating situation. Make sure you get exactly the right part, and compare the impellers. Also, heed that routing diagram **(Fig. 5)**.

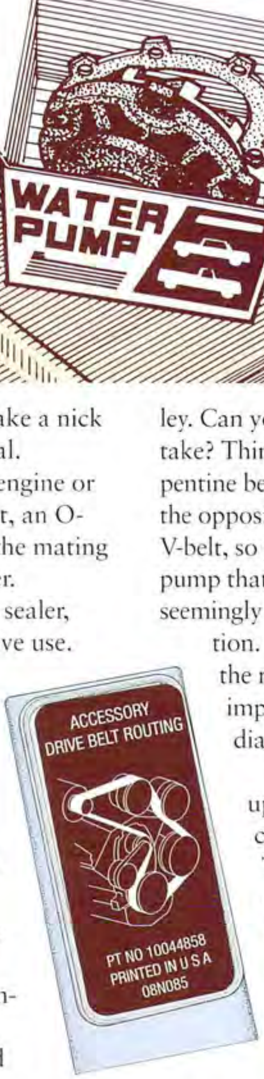

(Fig. 5) Vehicles with serpentine belts have a belt routing diagram under the hood somewhere. Read and obey.

After everything is buttoned up, figure out how much coolant your system holds. This should be in your owner's manual, or in the shop manual. Using the appropriate manufacturer's coolant, add half that amount and then top off with water. This will give you a 50-50 mix. Be sure to follow the service procedure for bleeding the air bubbles out of the system.

MODERATE

Curing a Clicking Starter

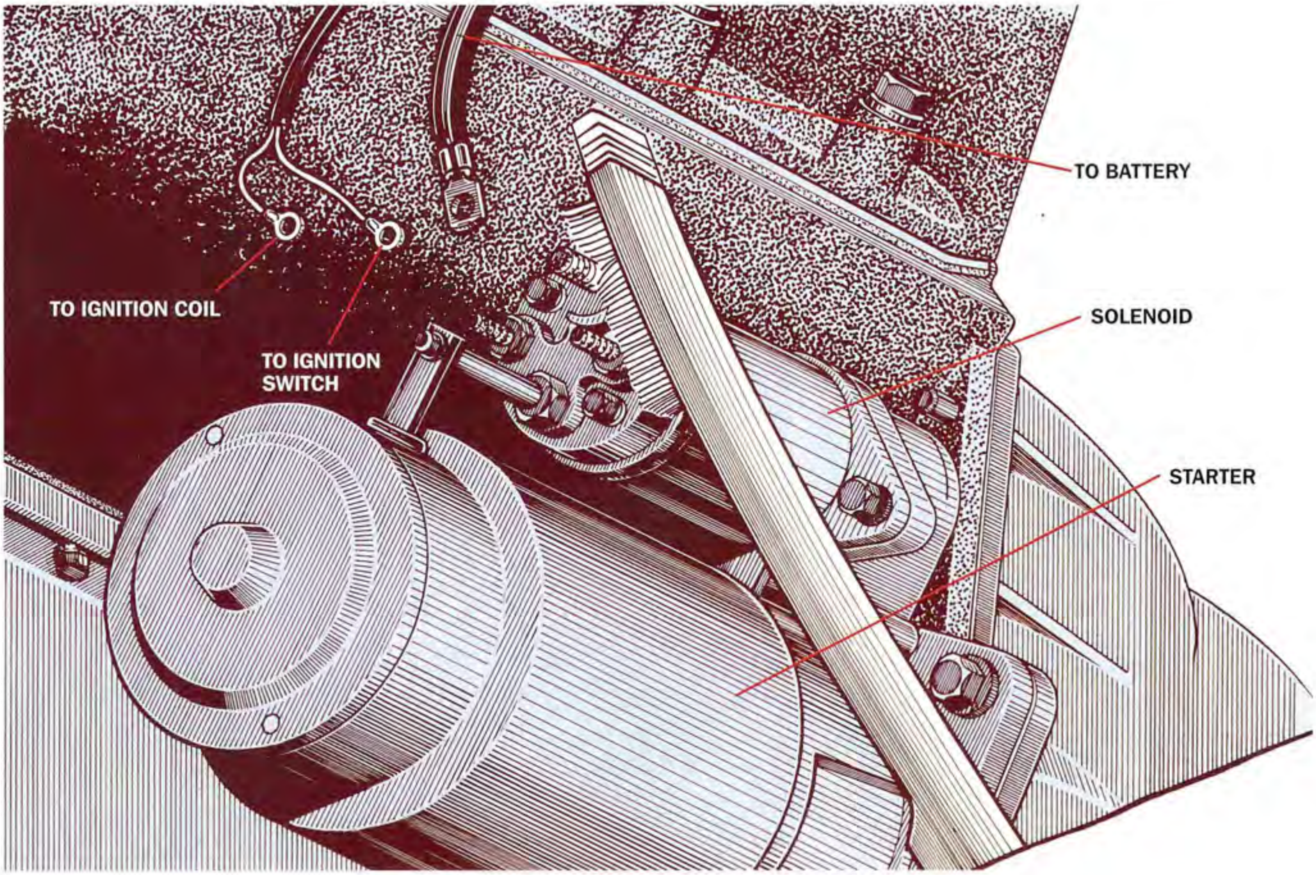

(Fig. 1) Loose or corroded starter cables can cause enough voltage drop to cause a clicking, intermittent connection. Clean with a wire brush and then rinse with warm water.

You twist the ignition key to start your engine, and instead of the syncopated whirling of a cranking starter and crankshaft, you hear a click or a series of chatterlike clicks. The engine that started every day for so long is not going to start today. Time to hitch a ride to work. If it's a weekend, you get the chance to find out why your vehicle has died and fix the problem before Monday morning.

That clicking noise is from either the starter solenoid or the relay. The solenoid is part of the starter. Typically it has a terminal for a thick power feed wire from the battery (Fig. 1), and a thinner terminal for the current supply wire to a switching mechanism in the solenoid. The relay is a remotely mounted switching device between starter and battery that controls either the thick power feed or the thinner electrical feed to the solenoid's switch terminal.

The first step (although at this point it doesn't tell you the root cause) is to find out where the click is coming from. If you have a helper turn the ignition key to start, you'll be able to trace the underhood click. If it comes from the starter, your problem is in the solenoid.

Ford and Chrysler products may have a relay in the circuit. If the solenoid isn't the source of the click, tracing the sound should take you to the relay on those models.

Even after you know what part is clicking, begin your real diagnosis at the battery. If the battery top has an "eye" indicator (actually a battery hydrometer that indicates the state of charge), recharge if the indicator is black. If it's green, it's got a normal charge. If it's yellow, get a new battery because the electrolyte is too low.

There's no indicator eye? Connect a voltmeter across the battery terminals (positive lead to positive terminal, negative lead to negative terminal). If the meter reads under 12.4 to 12.5 volts, it's borderline or undercharged (depending on the design of the battery). Recharge it for the day. Batteries do run down as a result of a temporary series of operating condi-

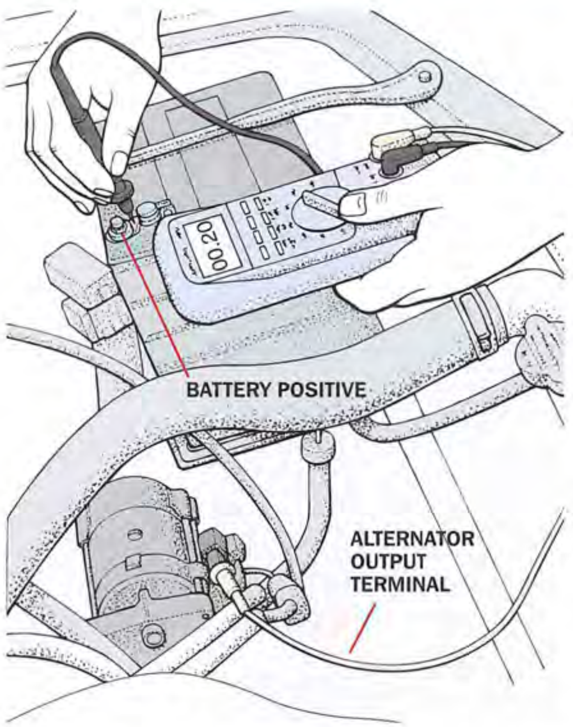

(Fig. 2) Check for voltage drop between battery post and starter terminal while a helper cranks the engine. Up to 0.50 volt is acceptable.

tions (lots of short trips, for example), but if the problem recurs, you'll have to check charging system output and the possibility of a short circuit.

If the reading is 12.7 to 12.9 volts, that's a good starting point. After a recharge, operate the headlamps for 15 seconds to remove what is called the "surface charge." The meter reading should not drop more than about 0.2 volt.

The reading is okay? Have a helper turn the ignition key to start, and in 15 seconds, read the meter. If it's below 9.5 volts, the battery may not be strong enough. Professionals have battery load testers to make sure. Your alternative: If the battery voltage was normal when you started, but is low during the attempt to crank, try a jumpstart. If the engine cranks normally with a boost, the battery probably is bad.

Corroded, Distorted Terminals

If the engine still won't crank, next inspect the cables and their connections at both ends. If you see corrosion or a possibly poor connection, make a voltage drop test **(Fig. 2)**. First, connect the voltmeter negative lead to the battery ground terminal and the positive lead to the engine block, close to the starter. With the key held in the start position, the voltmeter should read under 0.5 volt. If it's 0.5 volt or higher, the drop is excessive. In fact, if it's above 0.2 volt, that's really too high and could be a contributor to the problem if the battery is marginal. Perform the same check with the power feed side of the circuit (in this case, connect the voltmeter's positive lead to the battery, and the negative lead to the starter's battery cable terminal).

Get a high reading? Repeat the test, taking care to make contact at the battery post or side-terminal bolt, not the cable end. If the voltage drop now is within reason, the cable terminals are the problem. A simple cleaning may be all that's needed, but if a cable's

HOW IT WORKS

Starter Solenoids

A solenoid is an electromagnetic device that is capable of doing work, and in some starters, it does two jobs. 1) It moves a plunger that makes electrical contact between terminals for the battery and the starter motor, so the motor turns. If that's all it does, it's really just a switching device. 2) In some starters, the movement of the plunger also pushes a linkage that moves the starter's drive gear into mesh with the flywheel ring gear.

The solenoid has two wire coils. One is large, draws a lot of current and produces a strong magnetic field. That's enough to move the plunger. Once the plunger is in position (having completed the circuit and moved the starter drive), the large coil is disconnected and the circuit for a small coil is completed. The small coil draws a small amount of current and produces a weaker magnetic field—just enough to hold the coil in position. This saves battery energy for the big job of cranking the engine. If there isn't quite enough battery electrical pressure (voltage) to provide the current flow, however, the plunger won't lock into position so the small coil can take over. If this happens, all you hear is a solenoid click, and the plunger springs back. Some solenoids have an extra small-gauge terminal. This bypasses the ballast resistor, ensuring a hot spark while cranking.

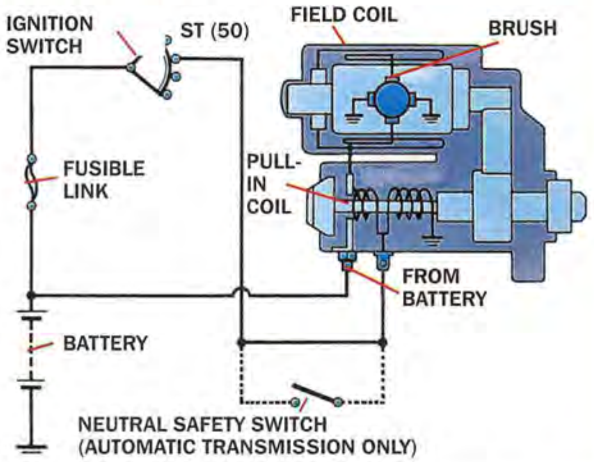

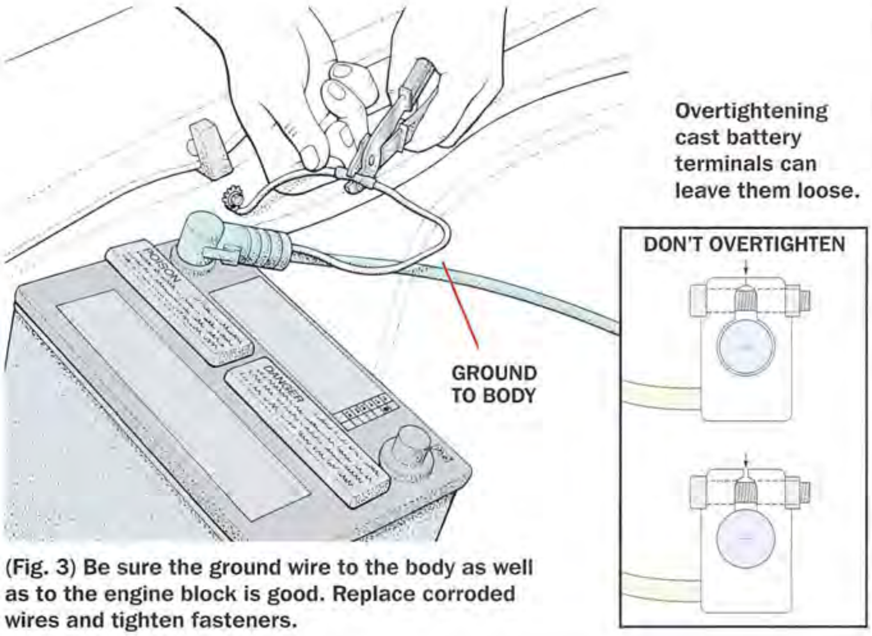

Overtightening cast battery terminals can leave them loose.

(Fig. 3) Be sure the ground wire to the body as well as to the engine block is good. Replace corroded wires and tighten fasteners.

post terminal is distorted, the jaws may be tight even though the inside surface is not making good contact.

If the cable end is distorted or corroded, replace it (**Fig. 3**). Get a quality cable end, which includes a section of cable with a protective sheath, not just a terminal. Cut the corroded cable back to where the copper wire is absolutely free of any corrosion. Install the new cable end and join it to the remaining cable (some repair cable sections have heat-shrink insulation, and others have screw retainers or crimp on).

No sign of corrosion on a ground cable? Remove the grounding bolt at the engine, clean the cable end and bolt, reinstall and tighten.

Checking the Starter Terminals

At the starter, inspect the terminals for both the battery (thicker wire) and solenoid switch for corrosion and physical damage. If the corrosion is minor, you may be able to remove the retaining nut and battery cable and wire-brush corrosion away. If the corrosion is so severe that cleaning it off leaves the threads damaged, install a repair stud, which cuts new threads onto the damaged studs (**Fig. 4**).

Now try to crank the engine. No improvement? With good connections at both ends (battery, ground and starter), try direct wiring with jumper cables, and if cranking still is weak, the starter apparently is bad.

Don't Have a Good Meter?

If you don't have an accurate voltmeter, you still can eyeball and make hands-on inspections for tightness of the cable connections at both ends. Clean and tighten the cables and see if the engine will crank.

Still no success? Disconnect both battery cables and make direct connections with booster cables, one from the battery's negative post to engine ground, one from the battery positive terminal to the starter solenoid terminal. Do a follow-up test with the booster to the starter battery terminal, plus a jumper to the solenoid terminal. These test procedures can be physically difficult to perform on many vehicles, particularly those with a side terminal battery, but they usually can be done. Just take the time to make good connections with the jumpers. If you can get the engine to crank this way, the problem obviously is in the cable connection.

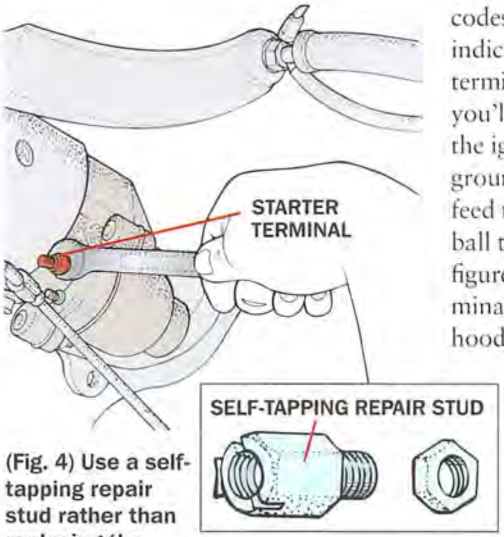

(Fig. 4) Use a self-tapping repair stud rather than replacing the whole starter if the terminal threads are damaged or stripped.

If a Ford or Chrysler product's relay is the source of the click, it may not be operating properly. If running jumpers to the solenoid or starter (bypassing the relay) gets the engine to crank, test the relay.

On Ford products, connect a booster cable across the thick-wire terminals of the relay, and if the engine now cranks, replace the relay. On Chrysler vehicles, find the power feed (it's the wire terminal that turns on a grounded 12-volt test light). Connect a jumper wire from that terminal to the one for the power output wire (usually red) that goes to the starter solenoid. Needless to say, be sure the vehicle is out of gear and the wheels are blocked before making any attempt to turn the engine over.

On models with plug-in relays, you may have trouble finding the color codes, but the wiring diagram should indicate the power feed and output terminals and their numbers, which you'll find on the relay itself. Turn on the ignition. You now can use a grounded test light to find the power feed terminal, but you'll have to eyeball the relay's terminal numbers to figure out which is the output wire terminal (to the solenoid) in the underhood center itself. Connect a jumper from the power feed to the output, and if the engine cranks, replace the relay.

Make cleaning and tightening all the connections a spring and fall ritual and your engine will make reliable cranking a year-round habit.

Repairing Cooling System Leaks

Okay, you replaced your lower radiator hose last week after it burst on the freeway, and the syrupy, turkish bath odor of boiling glycol coolant hitting a red-hot exhaust manifold is something you can live the rest of your life without ever smelling again. Lying face-up in a spreading pool of cooling coolant to change the hose is pretty low on the list, too. So it's a bad omen when that smell hits you at a tollbooth a week later—and a worse omen when you open the hood and realize the new hose is leaking from both ends. What gives?

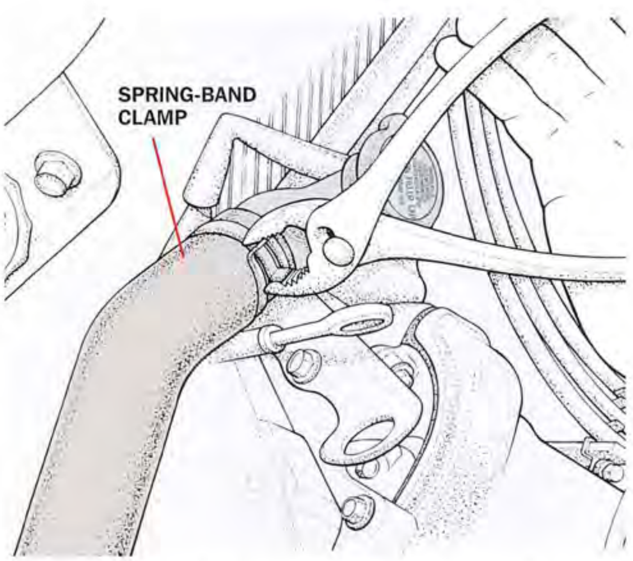

(Fig. 1) Spring-band clamps can usually be removed or replaced with pliers or locking pliers.

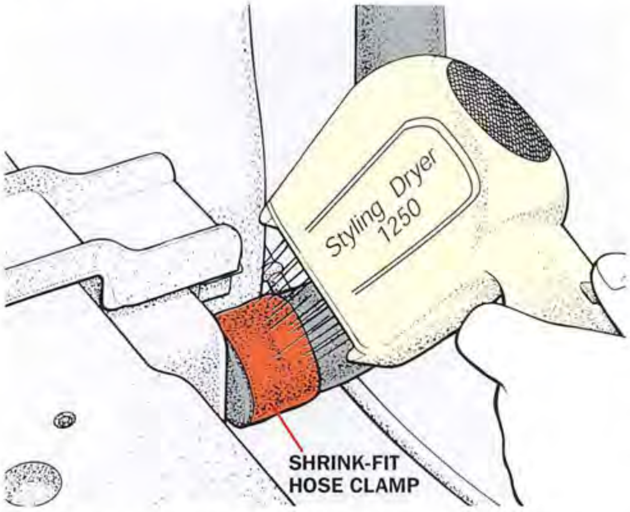

(Fig. 2) Shrink-fit hose clamps will continue to shrink as engine warms up, ensuring a tight seal.

Drip Patrol

There are about two dozen coolant hose connections underhood today, and it's a constant effort to find and fix the loose ones that leak coolant. Ingesting air is a routine issue. It used to be simple: Look for an antifreeze stain, then just tighten the hose clamp, right? Sorry, but that's not always true anymore.

First, the powertrain compartment is so tight that you can hardly spot a leak without a dedicated inspection. You're more likely to look closely at a hose connection when you have to disconnect a hose to reach something else. In either case, when you do look at the coolant hose connections under your hood, you may see very few of the type you loosen and tighten with a screwdriver.

Today's engines have complex coolant flow patterns and the compartments are so tight that the engineers have to use special hose designs to provide safe routing. Some of those hoses have plastic fittings, called quick-connects, to help an assembly line worker make errorfree connections.

Depending on the age of your car and whether or not clamps were replaced, you could have a variety of hose clamps. You can tighten some of these, but not all. (See page 121.) These are the ones you can tighten:

SCREW-TOWER—The screw is perpendicular to the band, and turning it down tightens the band. It's been around forever, it's cheap and it rust-freezes in place, so tightening an old one usually is impossible. To get it off, spray it generously with penetrating solvent, loosen the tower screw and slip in a slim screwdriver if necessary to pry it open. Or just cut it off.

DOUBLE-WIRE OR BAND WITH RETAINING SCREW AND NUT—A double-winding of wire or a band is held together by a tangential screw at one end that fits into a nut at the other. When it's overtightened, the wire type digs into the hose and may cut through.

WORM-DRIVE CLAMP—This has been the longtime favorite, and even was considered a premium design. Some worm-drives are, but most aren't. One reason it's popular is that it can be opened up and taken off without disconnecting the hose, although that feature has limited utility.

The quality worm-drives have such features as: rust-resistant plating; rolled up edges so the band doesn't dig into the hose if overtightened; offset teeth that keep the band from twisting when tightened; and even "teeth" that aren't cut through the band, so the hose rubber doesn't extrude into the slots.

CONSTANT-TENSION WORM-DRIVE—The

best ones (by Oetiker, a leading European maker with extensive U.S. marketing) have an internal band that glides through a slot inside the main band, bridging the joint of the worm-drive. Result: The clamp provides true 360° clamping. That clamp also has a coil spring to provide constant tension even if the hose underneath takes a compression set.

Spring-Band Clamp

Today most carmakers are using a clamp that you can't tighten, so it also never can be retightened. It's the spring-band (**Fig. 1**), an inexpensive form of constant-tension clamp. It may not be everywhere under the hood, but it usually is in a lot of places. Because it can't be pretightened to any spec, it's sized so even if the hose takes a set underneath, it maintains some tension, hopefully adequate to prevent a leak, but only if the hose neck is in perfect shape.

Shrink-Band Clamp

Would you like a low-cost non-adjustable clamp that not only maintains tension but seals well even if the hose neck is far from perfect? It's here, and it also can help with other problem clamping situations: the plastic shrink-band clamp.

You must buy a shrink band that's sized for your particular hose diameter. The band comes on a thick cardboard roll so it doesn't shrink in storage. Just crush the roll, remove the band, lube the hose neck with antifreeze, slip the band onto the hose and the hose onto the neck. Unlike a conventional clamp, the shrink band should be positioned so it extends onto a bead on a hose neck.

Apply heat with a hair dryer (**Fig. 2**) or heat gun (from within a couple of inches or so), and in a couple of minutes the band conforms completely to the neck and bead, increasing leak resistance. And the heat from the coolant will cause it to continue to shrink in service to compensate for any compression set in the hose.

How do you get it off? If you're replacing the hose, just cut it with a single-edge razor blade. If you're planning to reuse the hose, you have these choices: 1) Cut the band itself with a soldering iron, but be careful; 2) Force a feeler gauge between band and hose, and run the razor blade through the band just over the feeler; 3) If you plan to use these shrink bands everywhere on your cars, get a band-slitter, a tool that does basically the same thing but is easier to use.

HOW IT WORKS

How Coolant Hose Leaks Occur

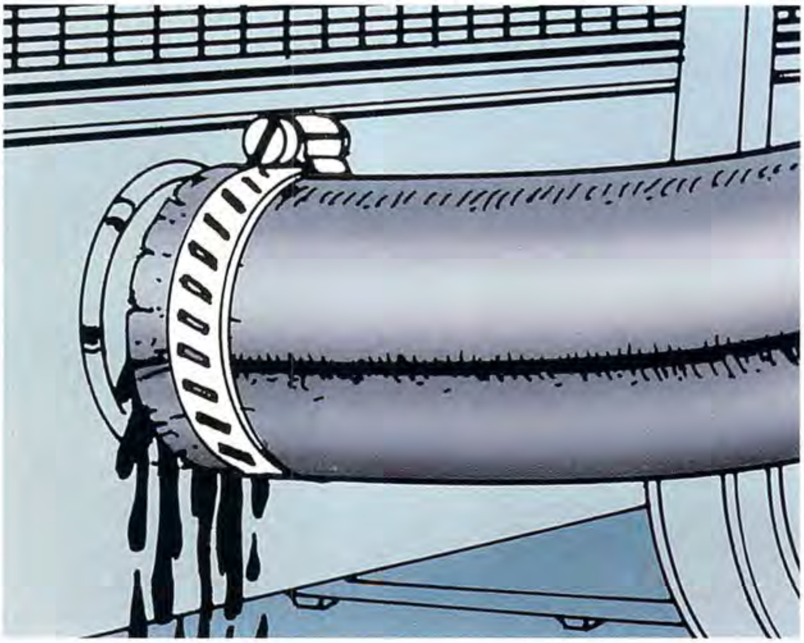

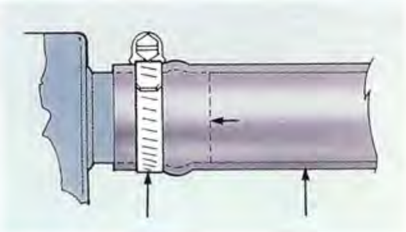

Why does a coolant hose connection leak after you've tightened a conventional clamp? After all, the clamp itself doesn't loosen. What happens is this: Both the hose neck and the hose expand when the coolant warms up. The clamp, however, is relatively unchanged, so it squeezes the rubber underneath even more, and this causes the rubber to become permanently compressed, which is called a set. When the engine cools, the neck contracts more than the hose. Many hose materials become virtually glued to the neck, so a seal is maintained. Others do not. In fact, silicone is almost immune to heat-sealing. That makes the silicone hose easy to replace, but it is the most prone to cold coolant leakage. Always install the clamp next to, but not overlapping, the raised bead on the fitting to keep from trapping a bubble of coolant in the void space inboard of the bead.

Those Quick-Connects

In many cases, today's cars and trucks use quick-connects instead of clamps for many heater circuit hoses and also for some radiator hoses. The quick-connect is a fitting with an O-ring seal, and if it ever leaks, you have something else—one or two O-rings—to check and replace.

Every quick-connect comes off a bit differently, but it's usually obvious. On GM pickups you unthread a retainer, then turn a metal tab that provides a secondary hold and pull the hose. The hose end is a plastic fitting with O-ring seals. Just peel them off and install new ones. Clean out any debris from inside the female metal fitting on the engine.

So if the leak is from the quick-connect, get a replacement from the car dealer, then just pry open and remove the old quick-connect's permanent clamp and install the new quick-connect on the hose.

If you can't get a new quick-connect, you may be able to cut through a metal section in a long underbody coolant line, install a short hose section (secure it with some form of constant-tension clamps), and that should give you the extra hose length at the end that has the leaking quick-connect. Then you should be able to remove the quick-connect and make a hose-to-metal fitting clamp joint.

The Hoses

Many of today's hoses, particularly for the heater circuit, have crimped-on sections, and crimps are known to leak. A brand-new hose is a simple but expensive solution. An alternative is to grind or saw into the crimp, just enough to be able to break it apart.

Even without crimp sections, the coolant hoses themselves are anything but simple, flexible lines of rubber. On cars with pressurized coolant reservoirs (where the cap is on the reservoir, not the radiator or engine), the upper radiator hose typically has a tee fitting with a secondary hose to the reservoir (and the lower hose may have one to the heater circuit). So unless a repair tee is available, don't be surprised if the hose prices out at $100. Other hoses (also not cheap) are permanently crimped to metal lines, much like an air conditioning line or power steering hose. And even where there is a simple-looking hose, it may be a molded design, so that it fits into a very tight area (perhaps so it can connect to a metal line) without the possibility of kinking.

THEY MAKE SPECIAL PLIERS for wire-band constant-tension hose clamps, but ordinary pliers will usually get them free. It's a lot tougher in constricted areas, though.

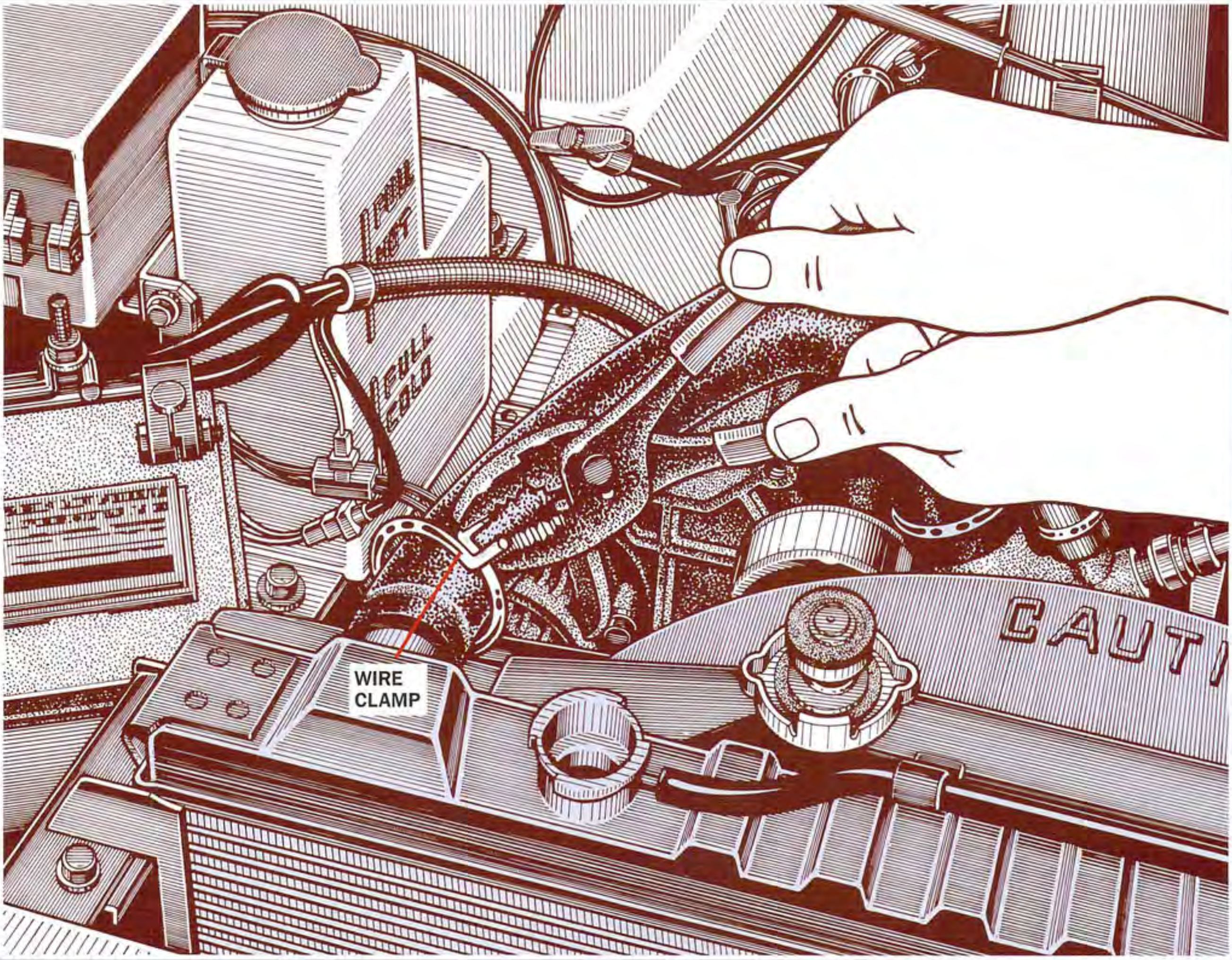

Curing Slow Cranking

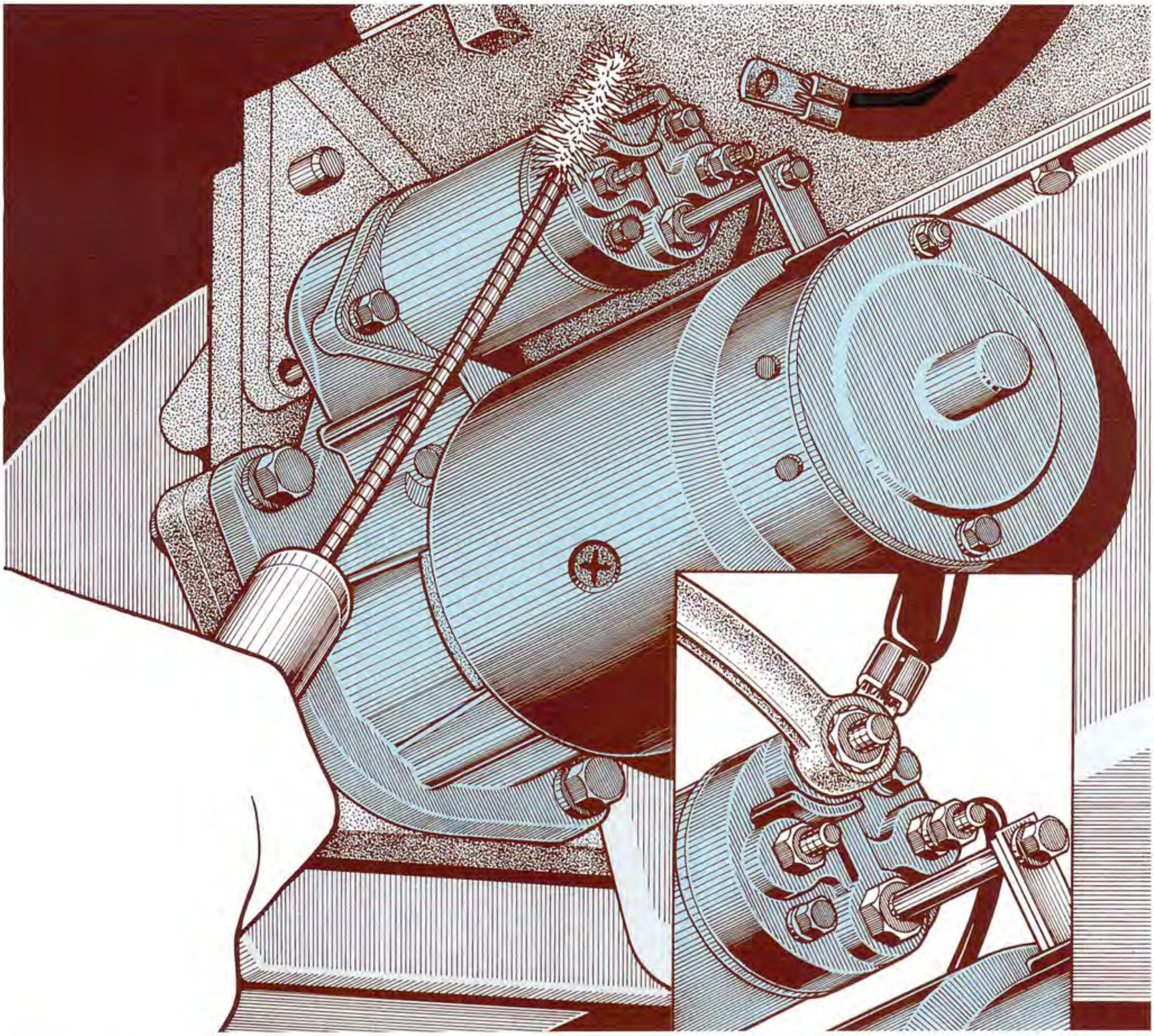

(Fig. 1) **Remove the wires from the starter motor or solenoid and clean off any corrosion with a wire brush and warm water. Rinse thoroughly and reassemble. Smear protective grease on the terminals to keep them clean.**

You've just treated your date and yourself to a postprandial double-caf no-fat cappuccino, and the caffeine courses through your veins like lava. You're pumped for a night of dancing as you head for the nightclub. The world seems to be in slow motion as the caffeine hits your brain—especially when you twist the key and your car engine grinds over very slowly, far too slowly for the engine to start.

The ugh-ugh-ugh of an engine that's cranking too slow to start is a sound that everyone can identify. And the immediate—if temporary—cure is to call for a battery boost. The long-term fix, however, requires careful examination to make a diagnosis.

Thinking Positively

If the battery positive-cable terminal on the starter is easily visible, begin by inspecting it for white/green corrosion and feeling the cable end for a loose connection. A wire-brushing to remove corrosion from the stud, nut and cable terminal, or just a tightening, may be the simple fix **(Fig. 1)**. Perhaps the cable terminal has become loose because at some time it had been reinstalled on damaged starter-stud threads, or the nut had been cross-threaded when installed. Poor-quality rebuilts, used starters, goofs during cable or starter replacement are some of the usual suspects. In fact, these problems are so common that you can buy an inexpensive (less than $5) starter-terminal repair stud kit at auto parts stores. This cuts its own threads as it is threaded onto a stripped stud (example: Thexton Nos. 530 and 531). See page 89.

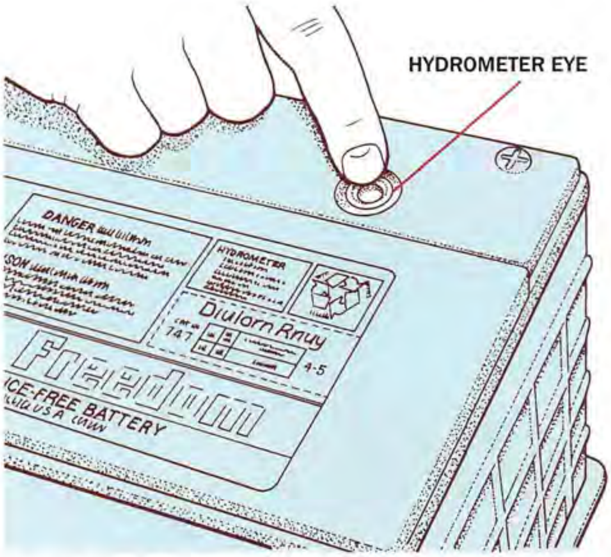

(Fig. 2) Check the battery's hydrometer eye for enough electrolyte or low charge.

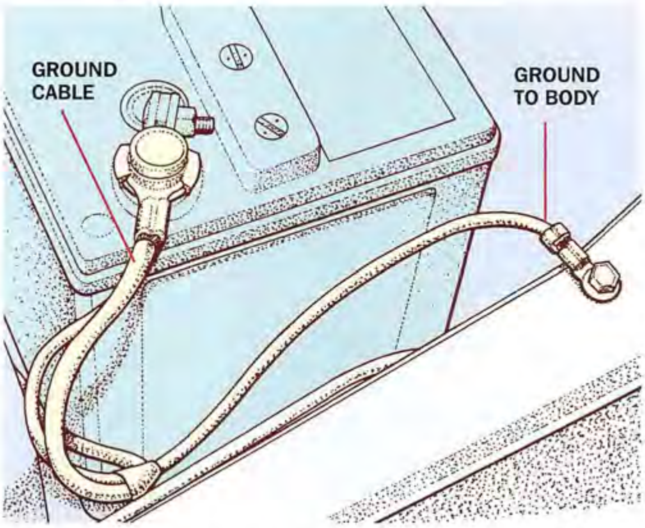

(Fig. 3) Check the ground connection to the vehicle's body for good contact.

Having a Heat Wave?

When the slow cranking occurs only on a hot restart in warm weather, you have to consider the possibility of high internal resistance in the starter, which is often caused by underhood heat. Some starters are factory-equipped with heat shields, particularly those in recreational vehicles, and the shield may have been bent or taken off and never reinstalled. Or it could just be that the starter wasn't designed to handle the heat well—a problem both with some original-equipment starters and with rebuilts that may fit physically but aren't designed to take the heat of a heavy-duty application.

If the engine starts easily after it cools off, the starter is one possibility. An easy check is to spray the starter with cold water if you get slow cranking during a hot restart. Now it cranks well? If there's no missing heat shield, check with a dealer to find out if a heavy-duty starter with better hot-starting ability is available.

If there's no sign of a problem at the starter, or if it's buried and you want to begin somewhere else, then move on to the battery. But we're not really giving the starter a passing grade just yet.

Batteries Not Included

Discard the old advice that says a small, cheap battery is adequate for a 4-cylinder, and only a big V8 needs a battery with a high energy rating. When you crank an engine, the electronics all wake up and the air starts flowing through the throttle body into the cylinders. It takes much higher cranking speeds to produce the necessary airflow through the complex flow paths to start today's Fours. And with the power demands of the electronics and the fuel system, a light-duty battery just can't hack it.

So if the old battery died and you installed the lowest-price replacement you could get, that could be the reason behind the slow cranking. In the days of carburetors, the battery-rating rule was one cold-cranking amp for each cubic inch of engine displacement. Well, we talk about liters for displacement today, but even if you look for 61 cold-cranking amps per liter, that's just not enough. There's hardly any late-model engine—or any engine—that will start reliably in all conditions with a battery rated at under 500 cold-cranking amps.

Also, you can't just look at the battery to determine its condition, even a battery with the circular window ("eye") in the top (Fig. 2). That eye is over just one of the six cells. That cell might be okay—and you'll see a dark green, the indication of health—while there's weakness in another cell.

Maybe you're ready to blame the battery because you can't see the "healthy" color. Try tapping on the eye before you start shopping for a new battery. The color may change.

The only reliable check is a load test, which requires a special tester. However, at least you can check the battery voltage with a meter connected across the battery terminals. But don't just look for something over 12 volts because it's a 12-volt system. If it's under 12.7 to 12.8 volts, the battery is undercharged, and at 12 volts it's 75% discharged. Yes, discharged.

If the battery is undercharged, you have to find out why. With the engine running at fast idle, you should check charging voltage across the battery terminals. If the voltage isn't well over 13, something's wrong. Unless there's a loose or glazed belt you can plainly see, inspect the battery-cable connections to make sure they're clean and tight. You can't see what the contact surfaces of side terminals look like without taking them apart. An almost-invisible coating of corrosion can constitute a charging barrier, so wire-brush them even if they appear good. Then, recheck the charging voltage.

If charging voltage still is low, inspect the other end of the ground (negative) cable. If the ground cable is a split design, the cable to the vehicle's body or chassis may be the easiest to see. But it has less to do with charging and starting, so more carefully check—and tighten if necessary—the cable to the engine, which may not be quite as accessible. Warning: Some cars have just a single ground cable from the battery to the body (Fig. 3). But don't stop there. Somewhere else on the car there will be a second strap, from the body to the engine or transmission (Fig. 4). It may even be buried underneath the car, so look around.

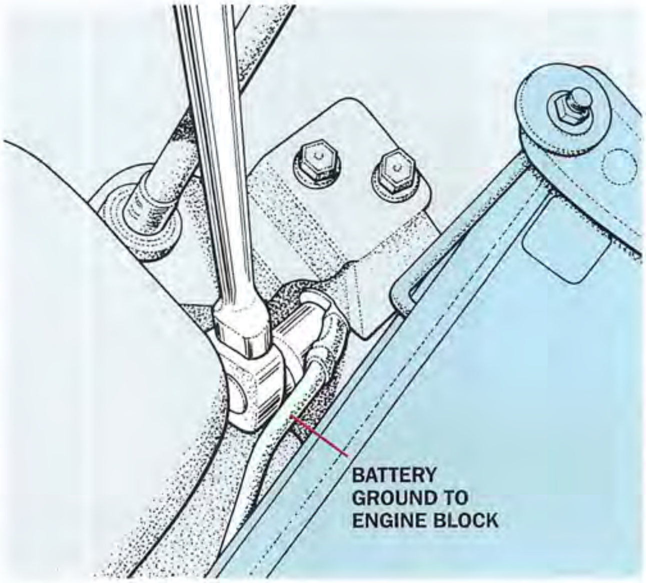

(Fig. 4) Check the ground connection to the vehicle's engine for contact as well.

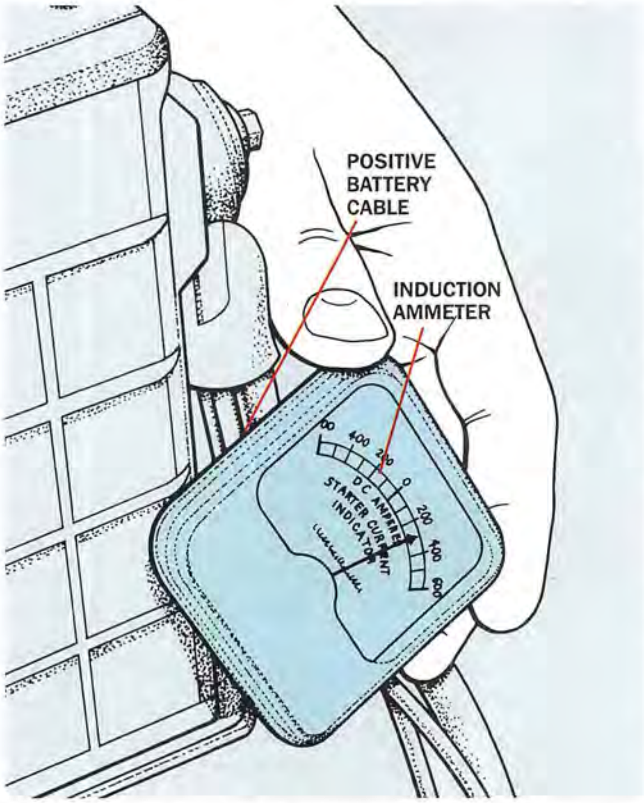

(Fig. 5) You can monitor the cranking current with an inexpensive induction ammeter.

With a fully charged battery, you're ready to find out if the starter circuit is pulling normal current. A low-cost induction ammeter (Fig. 5), which clips onto the battery cable, may not be a precision instrument, but it's close enough to pinpoint excessive current draw. Compare the reading with factory specs, if they exist. If they don't, here's a general guideline: Average starter draw is 150 to 200 amps, with 250 at the high end of normal.

Oil Is Thicker Than Water

When all the battery connections are good, there are two common causes of high draw: engine/transmission friction and a bad starter.

In cold weather, the friction may be caused by engine oil that's too thick. If you were using a few cans of oil-thickening additive during the summer to try to cure an oil-pressure problem, poor compression and so on, and haven't changed the oil since, you should do it before the cold weather hits.

If the vehicle has a manual transmission but can normally be cranked in Neutral without having the clutch pedal floored, a thick gear oil may be slowing down engine cranking. The thick gear oil may silence gear noise during warm weather, but by winter you may need something much lighter, both for easier starting and better shift quality. Some carmakers recommend automatic transmission fluid in a manual transmission oil for cold weather. Throw in the clutch pedal and if the engine now cranks normally, that's the tipoff. Check the owner's manual for acceptable lubes.

In hot weather, engine friction has the opposite cause: engine oil that's too thin. If you don't change the oil frequently enough, the additives in it will wear out and the oil won't provide the thin-film lubrication the cylinder walls need. Synthetic oils are better at maintaining their lubricity, but they're also slightly more expensive.

In addition, if your engine is running on the hot side—not overheating, just running a bit hotter than it should—that thins out the oil a bit more. The extra engine heat also may heat up the starter, which then will draw more current than normal. If the coolant temperature gauge is reading on the high side, get a precise reading, and if it's pushing 245°F or even 250°F, that's just too high.

When everything else seems all right, make the following checks:

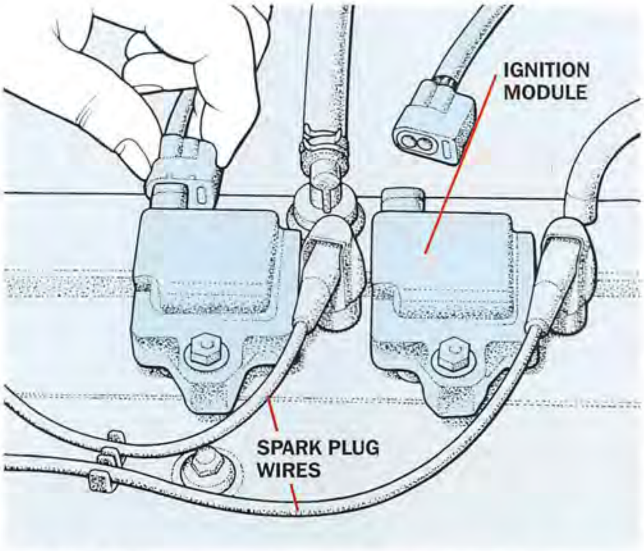

(Fig. 6) Pull the primary wiring to the ignition module or coil to check for misfires during engine cranking.

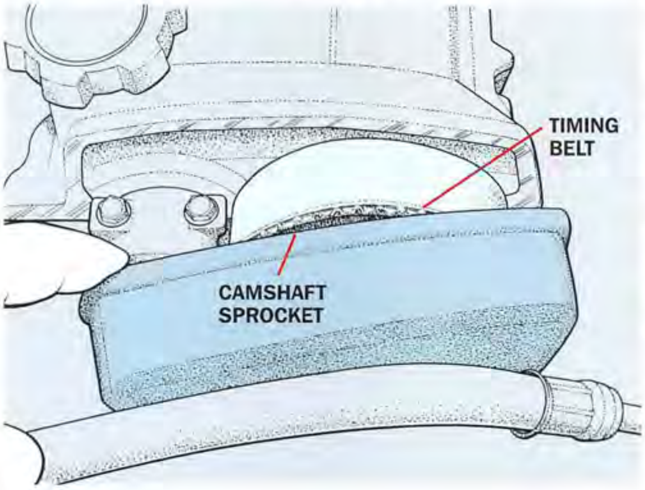

(Fig. 7) A camshaft drive belt can jump a tooth or two and affect cam timing and cause slow cranking.

- Is the ignition timing being upset by electromagnetic interference from the battery cables during engine cranking? This is most likely to happen if the battery or ignition wiring has been moved, even slightly, during service. Perhaps a cable wasn't reinserted in a clip or guide. A simple check is to disconnect the ignition primary wiring, such as from a distributor or distributorless ignition module (**Fig. 6**). If the engine now cranks faster, look for a mislocated cable. If the engine has a conventional distributor, also lift the cap to see if it's cracked or has carbon tracks between terminals. Either could disturb spark-plug firing enough to affect cranking.
- If the battery and ignition pass inspection, and the slow-cranking problem came on suddenly, check to see if the camshaft-timing belt jumped a tooth or two (**Fig. 7**). Unless they snap, timing chains are almost-forever items. However, a belt can jump a couple of teeth, and more than half of overhead-camshaft engines have belts, not chains. When a belt jumps time to advance valve timing, this can keep the engine from cranking and starting smoothly. If the cranking sounds unevenly slow and the engine seems to be trying to start but won't, valve timing definitely moves to the top of the list of possibilities.

Access to the belt on a transverse engine may not be easy, and checking the cam timing rarely is a 5-minute operation. Even the old Chrysler 2.2/2.5-liter 4-cylinder, which was one of the easiest to get to (thanks to an access window in the timing belt's upper cover), takes a bit longer than that.

On most cars, it is fairly easy to take off or at least substantially loosen the upper-half belt cover, gently flex it away and inspect the belt. If you see damaged belt teeth, that's a strong indication that belt timing may be off, and it's worth the extra effort of a timing check and a more careful belt inspection. Few late-model overhead-cam engines are "free-wheeling" (capable of tolerating a snapped belt without major engine damage). If you find the belt is near the end of its life and has jumped time, too, you not only could cure slow cranking, but save the engine or at least a cylinder head. And if the belt hasn't jumped but is well-worn, replacing it is great insurance, so you really haven't lost anything with the effort.

Finding Oil Leaks

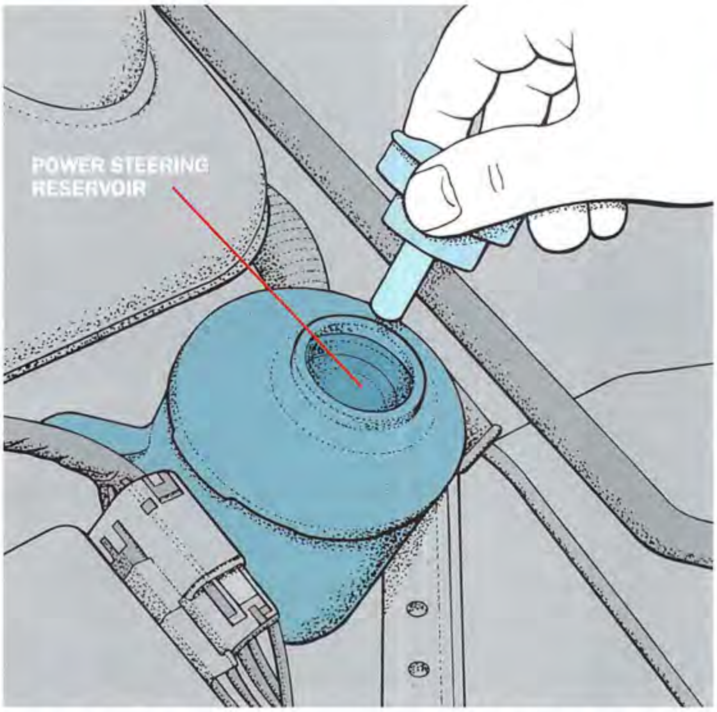

(Fig. 1) Check the dipsticks to see where your leak is starting.

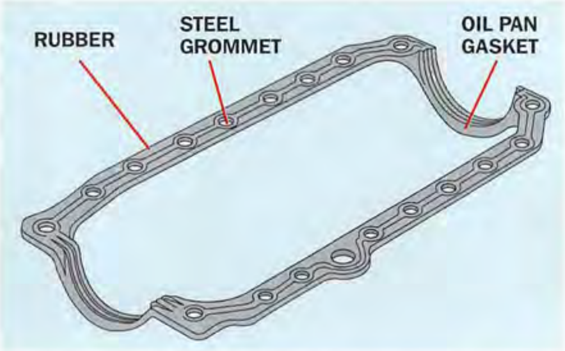

(Fig. 2) Oil pan gaskets may be a simple cork cutout or a pricey engineered rubber molding with metal inserts.

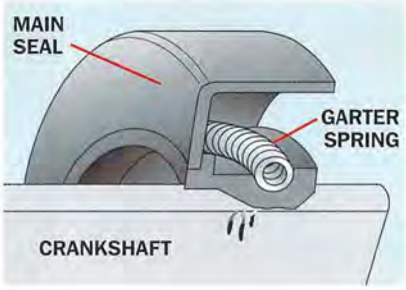

(Fig. 3) A leaky main seal will spray oil all over, and will need to be replaced.

Your tabby cat crawls into your lap, purring. Nice, but there's a huge patch of smelly, slimy, oily goo covering most of its back. Ick. Backtracking oily little catprints leads you into the garage, where a pool of oleaginous fluid has mysteriously stained the concrete floor. What is it and exactly what part of your car is oozing it? Time to put the cat down and get under the car. But there's leakage all over the engine compartment, and even a distinct pattern of windblown drops freckling the trunklid. Where's it coming from?

Begin by finding out what kind of oil is leaking. You can usually determine the color by putting a few drops on a sheet of white paper. Normally, engine oil turns black. Automatic transmission and power steering fluid is red but may discolor to brown or even be so dark that you can't tell it from engine oil. Washer fluid is blue and antifreeze is, well, it could be green, gold, orange, brown or blue, depending on the supplier. Feel the fluid. If it's very oily, it's lubricant. Antifreeze may have a light oiliness.

Your initial analysis points to oil, but you're not sure about the color or where it's coming from. Start pulling dipsticks. The power steering reservoir is a good place to start if the oil seems reddish (**Fig. 1**). If the reservoir is topped up, and the leak is at the front, check the automatic transmission cooler lines, particularly if they have sections of rubber hose with clamps.

If the oil is definitely black and the drops are directly under the engine, it's engine oil, which is the most common leak. But, once again, you're faced with that same question: Where is it coming from? A slow leak follows ribs on the engine block, is blown along the top of gasket joints, and oil gets everywhere. Put a lot of light in the engine compartment and take a look. You might get lucky and see the source. But unless you're sure, don't replace anything yet. A lot of gaskets are not only tough to replace, but they have sophisticated designs that are anything but cheap (**Fig. 2**). It would be a shame to waste a day replacing a costly gasket and still have the leak.

Oil also may be seeping past a worn crankshaft or camshaft seal (**Fig. 3**). The rubber lip that seals to the rotating shaft will eventually wear to the point at which the tension in the garter spring won't keep oil from leaking. This type of seal will only leak when the engine is running—and when it does oil will spray everywhere from the spinning shaft.

Because pinpointing an engine oil leak can be difficult, you should get all the help you can and, fortunately, a method preferred by the professionals is within easy reach: trace dyes

HOW IT WORKS

Gaskets

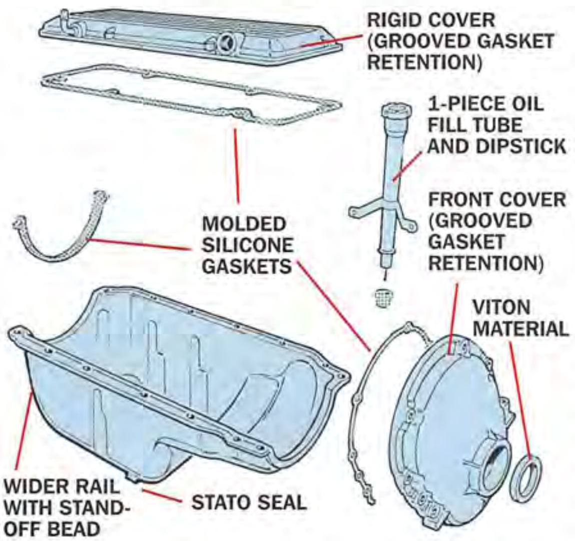
RIGID COVER (GROOVED GASKET RETENTION)
1-PIECE OIL FILL TUBE AND DIPSTICK
FRONT COVER (GROOVED GASKET RETENTION)
MOLDED SILICONE GASKETS
VITON MATERIAL
WIDER RAIL WITH STAND-OFF BEAD
STATO SEAL

When you look at a pair of machined surfaces, or even today's well-finished engine surfaces, you may wonder why a gasket, an O-ring or other type of seal is needed to prevent an oil leak. Except for seals around rotating parts, wouldn't clamping the parts together tightly be enough to prevent oil leakage?

The answer: close, but no. Sorry, even mating joints bolted together look a lot more precise than they really are, and fluids can seep past them. It takes a flexible material in the joint to compensate for any unevenness and looseness to prevent, or at least minimize, leakage. That flexible material is a gasket. The design of a gasket itself is a complex art. Simple materials such as cork compress nicely and compensate for a fair amount of unevenness, but under the compression of a line of bolts, cork gaskets soon take a "heat set," also called a "compression set." Retightening gaskets helps and may work for a considerable time. But when it doesn't, the gasket must be replaced. Today's oil-sealing gaskets (like all automotive gaskets) are made of high-temperature synthetic materials, in combination with natural fibers, that are more resistant to a compression set. They have engineered shapes that, when compressed, provide a more effective seal. There may be metal grommets around bolt holes to prevent overtightening. The gaskets often are made with raised rubber "beads," in some cases a single broad bead, in others, two or more riblike layouts, which set up a series of barriers to oil leakage. Gaskets often are shaped to fit into grooves in the mating metal surfaces. Or a gasket may look almost like a large rounded band or square-cut rail and fit into a groove in each gasket surface.

with ultraviolet (UV) light (**Fig. 4**). These can be used for all fluids—oils, fuel, coolant, even a/c refrigerants. The trace dye is fluorescent, so under UV light (so-called black light), it produces an unmistakable yellow/green glow. Aim the light, and a small dye stain may show you the source of the leak. The newest trace dyes were formulated in response to complaints from mechanics about low visibility of the dyes and the difficulty of positioning the large UV lamps. Now you can find kits with compact, flexible lamps, improved trace dyes and coated yellow glasses that enhance the appearance of the dye.

The kits typically include two bottles of dye, one for oils and another for antifreeze/coolant. The trace dyes for a/c refrigerants are very specific formulations, and they require special injectors. They are not part of the general purpose kits, but are sold in specific a/c kits.

We used the Tracer Products Leak-Finder Kit, a product that won a *Popular Mechanics* Editor's Choice Award at the 2000 Automotive Aftermarket Industry Week trade show. It's under $60 and includes a compact UV lamp with a flexible head, so it can be aimed into all sorts of underhood nooks and crannies. It also comes with a 10-ft. cord that has alligator clips to connect to the battery terminals, yellow glasses, and 1-ounce bottles of trace dyes for both oils and antifreeze/ coolant. You can buy individual bottles of any trace dye you use up.

Start by mixing a dose of oil-leak trace dye (½ ounce, which is half the amount in the see-through bottle) with as much engine oil as you have left in your top-up bottle, then pour that into the engine. You could just pour the 1/2-ounce dose into the engine, but if there's a leak you're probably down on oil anyway. If you just pour in the trace dye dose, it will coat the oil filler neck, and take a lot longer to be washed away by engine oil and mixed thoroughly into the oil supply. And it will take longer for the leak to show up.

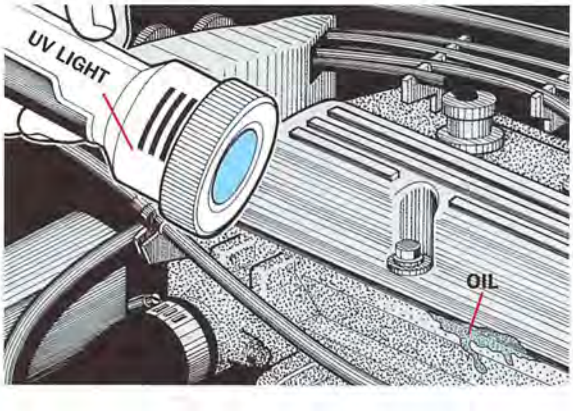

(Fig. 4) Ultraviolet light will make trace dyes glow brightly.

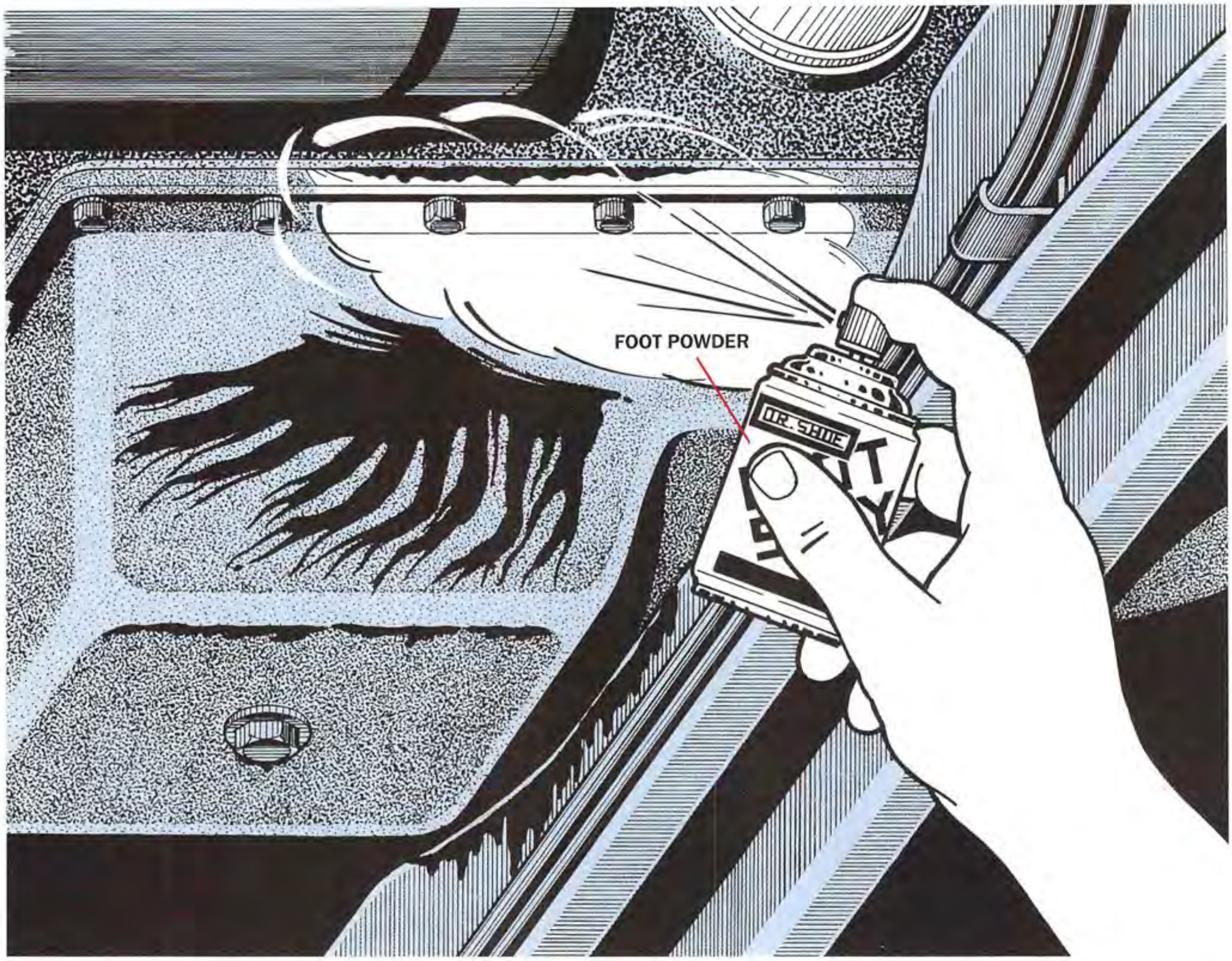

(Fig. 5) Clean the suspected area thoroughly and dry it. Coat liberally with powder to locate seepage.

After driving the car long enough to allow oil leakage, park the car over newspaper to catch the leaking oil. Check the drops on the paper with the UV lamp. If the drops glow, you're ready to look for the leak. Jack up the vehicle, support it on safety stands and connect the lamp clips to the battery. Then, put on the yellow glasses, aim the light up from underneath and press the switch.

Because leaking oil may follow a twisty path, look for the highest point of any oil trace, and that should lead you to the source. In our case, the path started at the oil dipstick tube, which had a leaking O-ring seal. In most cases, you'll find a loose gasket joint, which you may be able to tighten. However, in many cases, the gasket will have taken a severe "set" in the joint, and retightening won't stop the leak.

If you don't want to make the investment in a trace dye kit, there is an alternative method for finding a leak. You can use a couple of products you may already have at home—one in the garage, one in the medicine cabinet.

The garage item is aerosol engine degreaser/cleaner. Use the aerosol cleaner to loosen road film, then remove the film from the engine, transmission and adjacent underbody with a water hose. Drive the vehicle to dry it off—or when you jack it up and support it on safety stands, wipe the area clean and dry with a cloth. That's what you may have to do anyway to remove a lot of the road film. The objective is to clean the underbody well enough so road film and leaking black oil aren't confused.

From the medicine cabinet, get aerosol powder, such as that used to treat athlete's foot, and spray the underside of the general area of the leak, going as high up the block as possible **(Fig. 5)**. The powder will adhere to and coat the metal, producing a white haze. Then, drive the vehicle until the oil leaks—be careful to avoid wet, muddy roads in the process. The hope here is that the leak will take a single large, reasonably direct downward path (even if there are some streaks from airflow). If it does, it will show up as a primary "stream" down the engine. You may have to perform a similar treatment closer to the top of the engine to really pinpoint some leaks, particularly those from intake manifold gaskets. However, if the oil stream is blown through a complex path along the engine before a drop hits the ground, using a trace dye and UV light is a surer way.

Flushing Your Cooling System

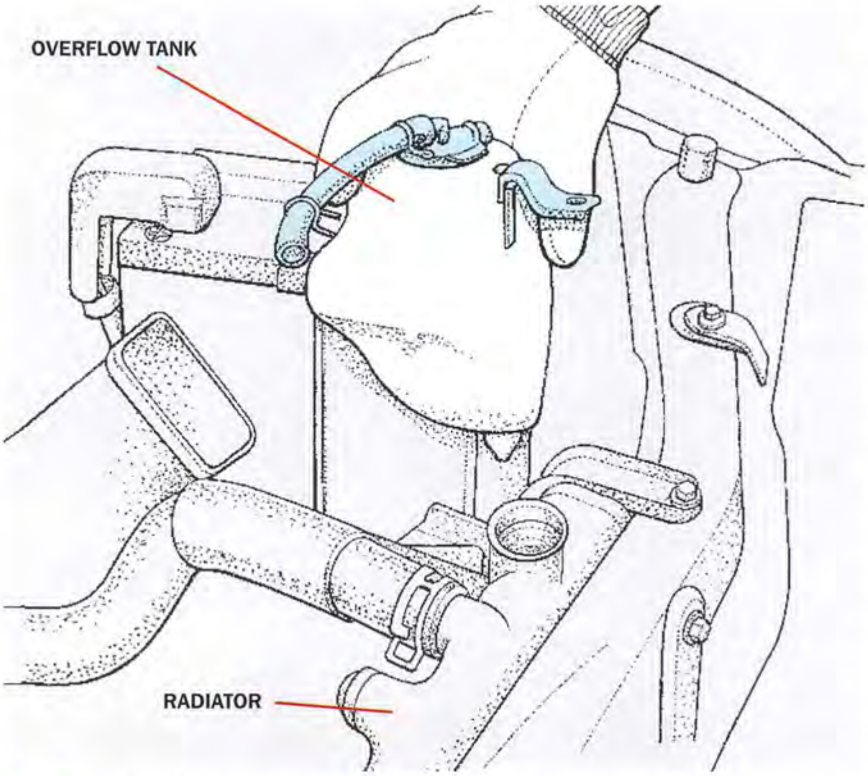

(Fig. 1) Remove the overflow tank, and carefully pour the coolant in it into your drain pan.

You take off the cap and look at the coolant. It's a nice shade of green, or maybe red, or maybe even orange. It looks good. Should you leave it in? Unless it's recommended by the vehicle or coolant maker for extended life, the answer is no, especially if it's been two years or more since the last time you drained it.

Today's engines are loaded with aluminum components: cylinder heads, water pumps, manifolds, even engine blocks. And the two primary heat exchangers—radiator and heater—are also aluminum. Aluminum needs great corrosion protection to survive, and the corrosion protection in conventional antifreeze is used up in about two years. If your car came with conventional, you can't switch to extended life without a fair amount of preparation. And if your car is much more than four years old, a switch is not likely to yield long-term coolant life—you'd still face the usual 2-year drain interval.

Glug, glug

First, let's do a proper job of draining the coolant. Start by checking the specs to see how much is in the system. This is important, because capacities vary all over the lot. Some Toyota Fours and V6s, for example, hold only about 5½ quarts. Other systems hold 14 to 18 quarts. This way, you'll know what percentage of the coolant drains out.

Start with a cool engine. If the pressure cap is on the engine or radiator, look at the overflow reservoir (Fig. 1), and if it's easy to disconnect and empty, go ahead. Then, remove the radiator cap and open the radiator drain cock. If the drain cock is in tight quarters, use a special socket available at most auto parts stores (Fig. 2). Let the coolant drain into a pan. Unless your town has a coolant collection setup, pour the old antifreeze into a household drain, clothes-

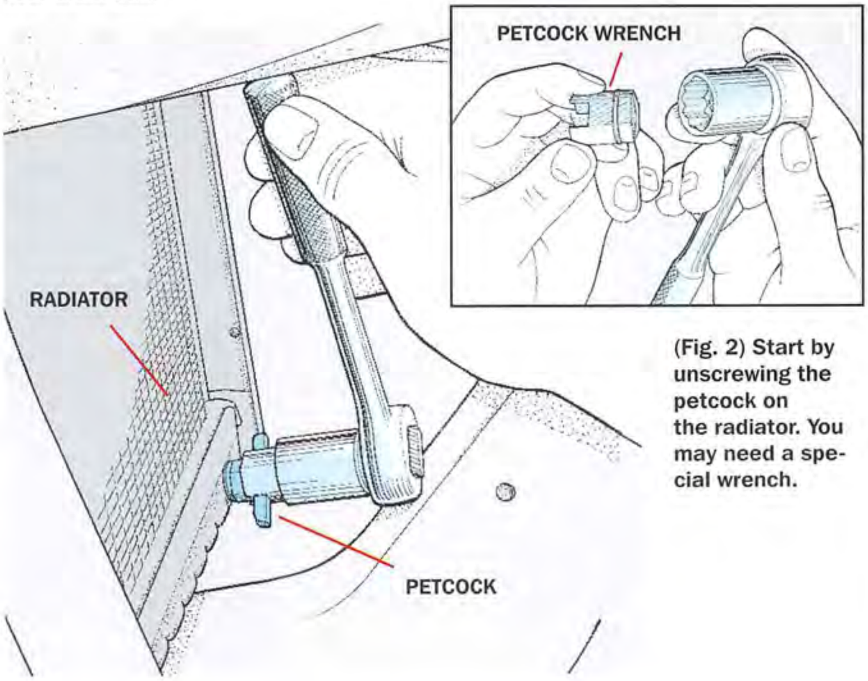

(Fig. 2) Start by unscrewing the petcock on the radiator. You may need a special wrench.

(Fig. 3) It may be necessary to remove the radiator hoses to completely drain the system. Remove the clamps and twist the hoses off.

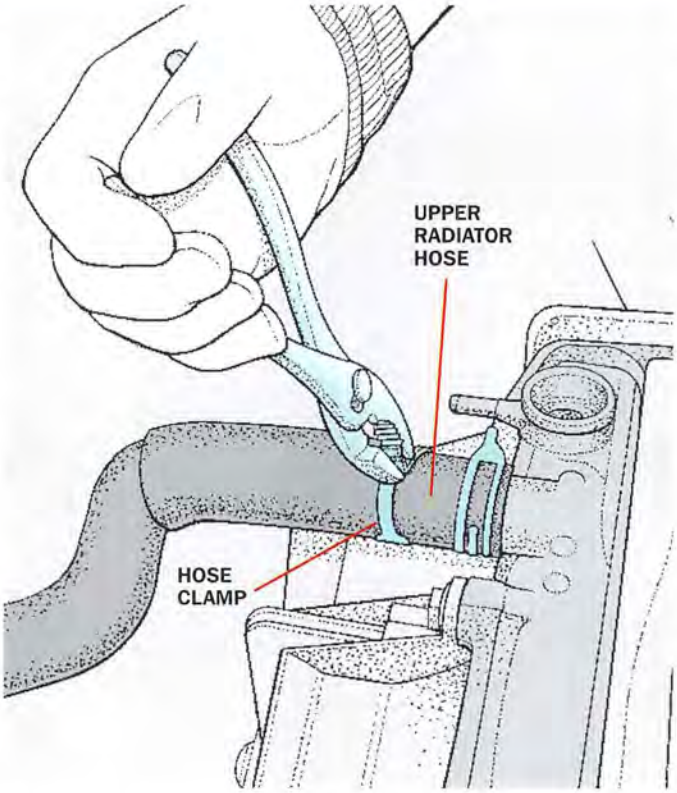

washer pipe or a toilet. That's an environmentally safe approach. Don't pour it on the ground or into a storm sewer. If your car has a copper radiator or heater core, the coolant may be contaminated with lead solder. Many municipalities have hazardous-waste disposal facilities that will take it. Also, in most of the United States it's illegal for professional mechanics to dump used coolant, so you may be able to take it to a shop or the municipality for recycling.

Next, move the dashboard temperature lever to HOT, so if your car happens to have a heater coolant control valve, it will open. If the pressure cap is on the plastic reservoir, remove the cap, then open the drain cock. No radiator drain cock? Disconnect the lower radiator hose from the radiator (Fig. 3). Move the hose clamp back from the radiator neck, slip a thin screwdriver between the hose end and radiator neck to free up the hose, then twist slightly to disconnect the hose. Draining the radiator alone normally should remove 40 to 45% of the coolant. After the first drain, fill the system with water, then warm up the engine and let it cool. Drain the radiator again and fill it once more with water. Repeat. Or remove engine coolant drain plugs if they're not corroded in place.

Let it bleed

Now comes the hard part—filling the system. If the system holds 12 quarts, you want to install 6 quarts of undiluted antifreeze, or exactly half of the cooling system's capacity (Fig. 4).

The cooling system has lots of nooks and crannies that trap air, making it difficult to fill the system with coolant. The fill cap and neck are supposed to be at the high point of the system to help air bleed out, but often they aren't. And even if they are, you need all the natural help you can get. So jack up the front of the car, which gets the coolant fill neck as high as possible.

Check for air bleeds on the engine. Sometimes you'll see an obvious air bleed, such as a boltlike item threaded into a hose. If there's an air bleed, open it. If there are several, open them all. If you have access to factory or after-market service information for your car, check it for a coolant fill procedure.

Pour in the required amount of antifreeze slowly until you see coolant oozing out of the open air bleeds. Then close the bleeds and top off the system with the remaining antifreeze and then plain water.

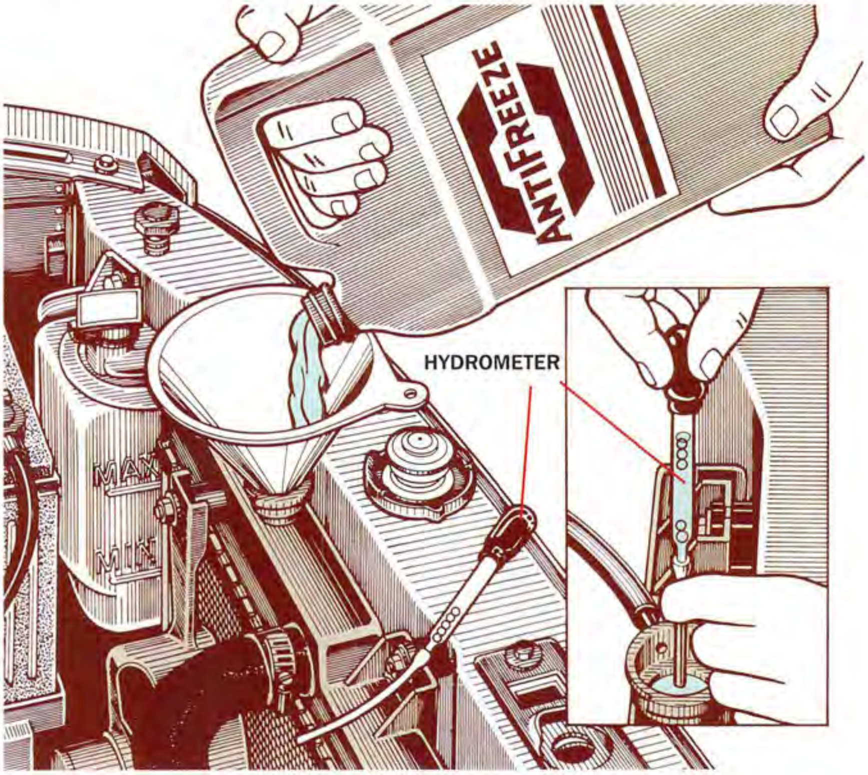

(Fig. 4) Refill the system to half its capacity with coolant. Then add water to fill, achieving a 50 percent concentration of coolant. Check the concentration with a hydrometer.

HOW IT WORKS

Cooling Your Engine

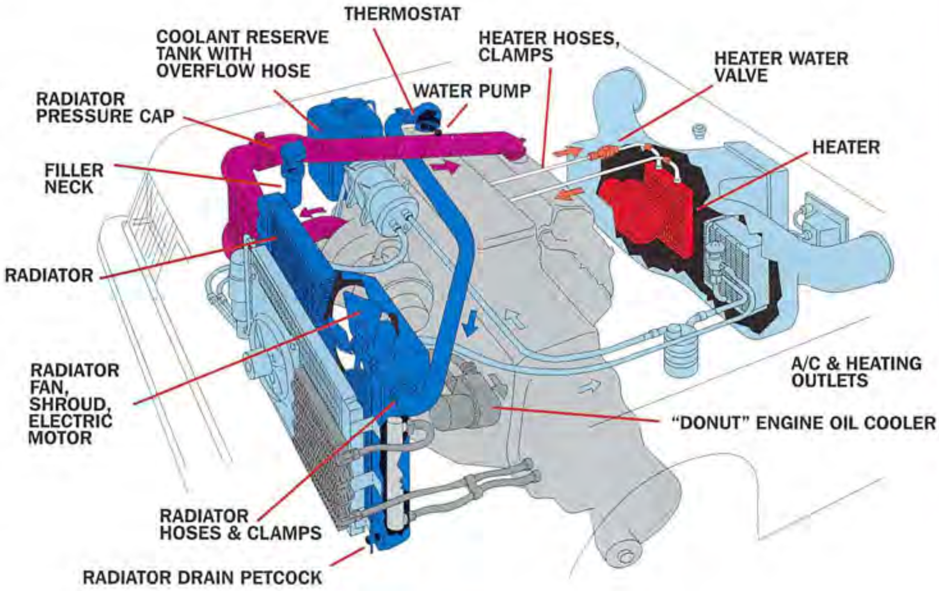

A coolant pump, or water pump, circulates the antifreeze and water mixture between the engine and the radiator. After the coolant circulates through the engine, the pump pushes it out the upper radiator hose into the radiator, a heat exchanger made of metal tubes (aluminum on today's cars) to which fins are attached. The fins draw away heat and dissipate it to the air that is drawn through the radiator by fans and the forward motion of the car. The cooled coolant is drawn from the radiator through the lower radiator hose and back into the engine by the pump, and the cycle starts all over again.

When the engine is cold, coolant circulates only within the engine, so engine heat warms it up faster. At about 195° F, the coolant heats a temperature-sensitive valve (the thermostat) that opens to allow the coolant to flow through the radiator. The thermostat may be located at the engine outlet, in line with the upper radiator hose, or at the inlet to the water pump (the preferred location on today's cars).

The coolant also flows through hoses into and out of the heater, which, like a miniature radiator, gives up its heat to the surrounding air. In this case, however, the heated air is blown into the passenger compartment.

Raising the cooling system's pressure also raises the coolant's boiling point, so the radiator cap (which also could be on the engine or on the separate reservoir) has a pressure valve to raise the pressure in the cooling system by about 15 psi. This increases the boiling point of the coolant by about 40° F. So, the boiling point of a 50/50 mix of antifreeze and water in a properly functioning system is about 265° F or higher.

If the system has a heater coolant valve, close it by moving the temperature control lever or knob to COLD. With the engine running at fast idle and warmed up, have a helper move the lever or knob to HOT while you listen at the coolant valve. If after the first rush of coolant you hear a continuous gurgling noise, there's still air in the coolant, and you should be prepared to watch the coolant level in the reservoir over the next few weeks.

Picking the right antifreeze

Most antifreeze is made with a base chemical called ethylene glycol. Green, yellow, red, orange, pink, and blue dye is used. The dye is not an indicator of the formula. Extended-life antifreezes, rated for four years/60,000 miles, were on the market until the mid-1990s. But the newest entries are superlong-life antifreezes with totally new rust/corrosion inhibitors, in some cases developed originally for diesels. The original, from Texaco (used as original equipment by GM), is called Dex-Cool. Other Dex-Cool antifreezes also are available from Prestone and Zerex, perhaps others. Dex-Cool is recomended for five years or 150,000 miles. Ford and Chrysler Group use a different antifreeze called a "hybrid," recommended for five years or 100,000 miles.

The rust/corrosion inhibitors vary, but if antifreeze is a major brand that is not extended life, but is in a car with a lot of aluminum components, change it in two years or 30,000 miles—whichever comes first. You can push that to a third year if the engine is all cast-iron. "Toyota Red" is a specific formula, but if you drain it, you can replace it with any name-brand American formula. Here again, the replacement interval is two years or 30,000 miles.

If you have at least 5000 miles on the antifreeze, the chemical bond with the aluminum components is "solid." So if you want to get extended life with a coolant installa-

THIS TIME-HONORED method of flushing your radiator is acceptable, but dispose of the used coolant into a sanitary sewer.

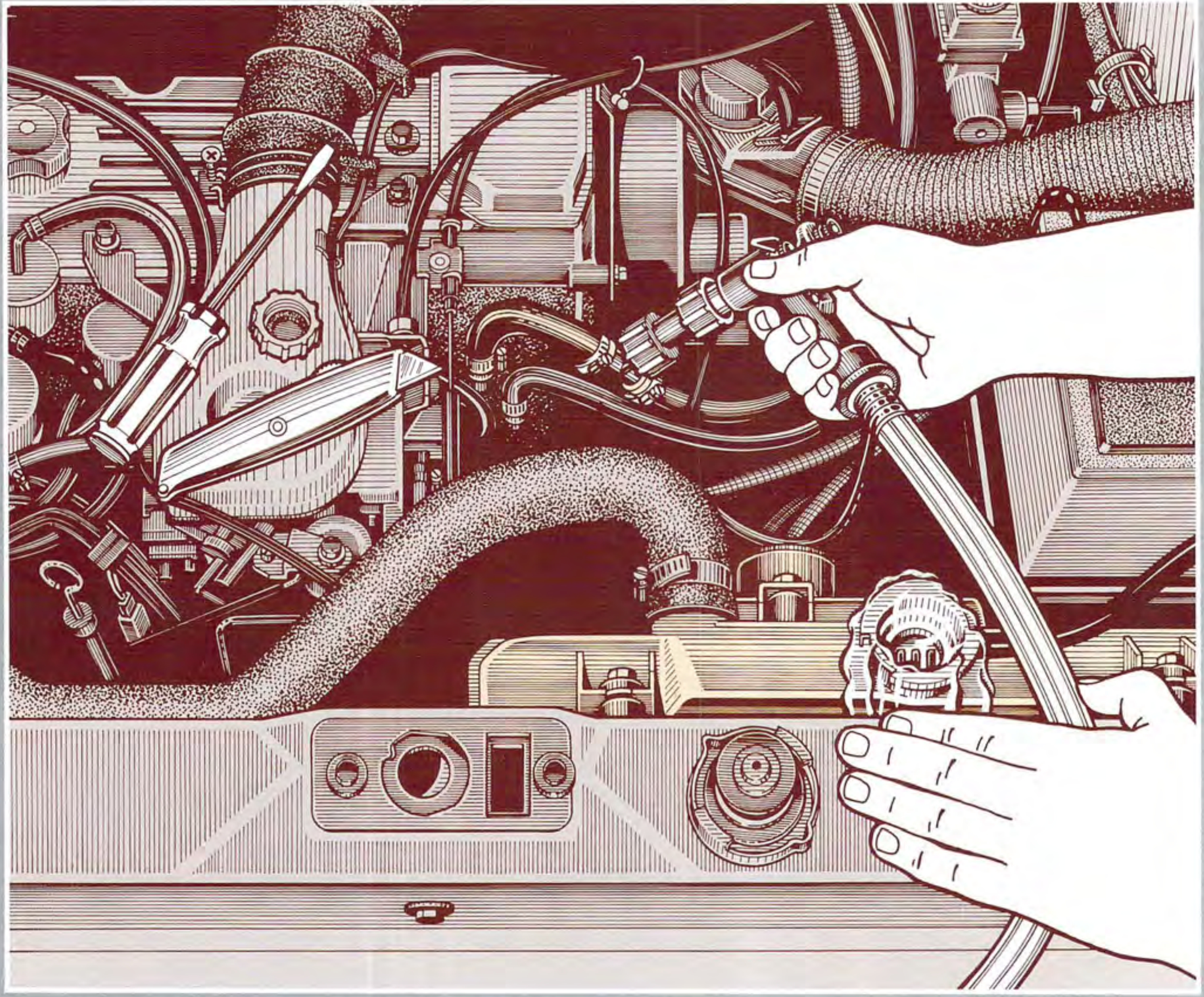

tion, just do a thorough drain-by-dilution, at least three times. The coolant you drain out should be virtually clear, like the color of water. If it's still colored, you have to repeat the process until it's all out.

With the radiator and reservoir drained, pour in the amount of antifreeze necessary—there should be plenty of room—and then top up with water. Follow the procedures we've discussed to ensure a full system.

Remember: You should do this only as a maintenance procedure, not as a final step when replacing a part, such as the water pump and radiator.

What about "pet-friendly" or "safer" antifreezes made with a base of propylene glycol? The name brands (Sierra and Prestone Lo-Tox) will do the same job as ethylene glycol. But they cost a little more and actually require a greater quantity to provide the same freeze protection, and in truth, they're only a bit less hazardous.

They aren't sweet, and consequently aren't as likely to be consumed by toddlers or pets. Don't rely on this, however. Store all unused coolant, low-tox or not, safely. Dispose of all drained coolant in a sanitary sewer or in accordance with local regulations.

Checking Your Ignition Timing

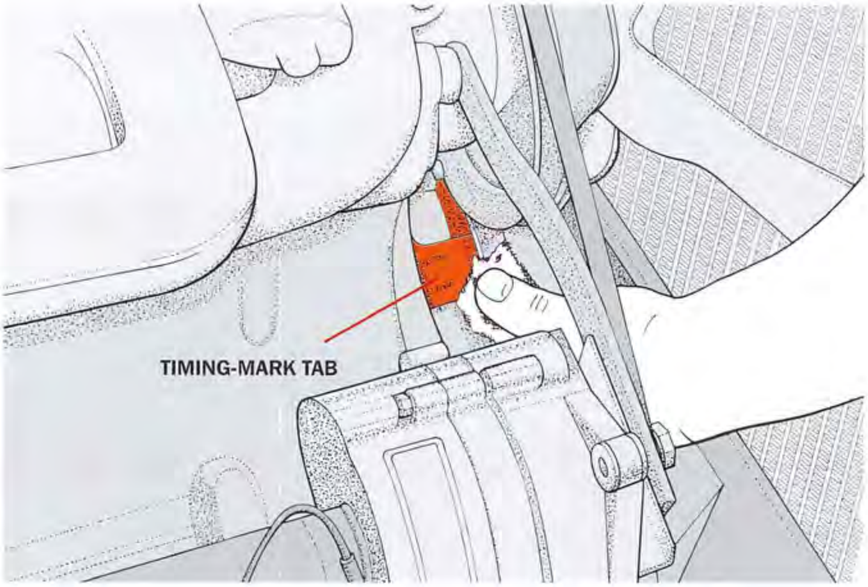

(Fig. 1) It's deep and dark in the engine compartment, and filthy to boot. On a cold engine, clean the timing tab and scale for better visibility.

You floor the accelerator to pull around a line of exit traffic, and you realize that your car is barely responding. A taxicab is camped half an inch from your rear bumper, and the cabby is alternating between using one of his fingers to semaphore his displeasure with your forward progress and leaning on the horn button. Your car's acceleration, nonetheless, remains a fraction of its former self. What could be wrong? The engine runs smoothly, but simply lacks power. There's no Check Engine light or trouble codes.

Pardon our asking, but when was the last time you checked the ignition timing? Ignition timing? Hey, you're driving a '97 model with a big, fast engine computer. All of those electronic gizmos handle that stuff, right?

Actually, you may need to dig out your old timing light and put it to use. Checking and, if necessary, adjusting ignition timing may have become lost skills, but a majority of cars and trucks on the road still have timing marks—and they weren't put there just for show.

If the sparks arrive at the plugs too early or too late, engine performance drops and the vehicle is likely to fail an emissions test. Sure, changes in ignition timing have been controlled by the engine computer on cars and light trucks for many years. However, the basic timing still has to be right.

There is a movement toward "coil on plug" ignition, eliminating spark plug wires (and the convenient attachment for the timing light) and the timing marks, too. If your engine has no marks, you'll have to check timing with a scan tester, which reads the sensor data processed by the computer. However, if there are marks, you can make the most meaningful check—the actual timing rather than a computer-processed signal.

If you have a car with a distributor, the timing probably is adjustable. It may have very little range for adjustment, but if you see any sign of a slot in the distributor's holes for the retaining bolts or nuts, timing can be reset.

Even if there is no distributor, there is a base timing specification, and if it's incorrect, the engine won't run right. The problem could be caused by a bad signal from a sensor or even by a bad computer, but unless you check the timing, you won't know.

Most pre-'96 models were equipped with an emissions diagnostic system called OBD (On-Board Diagnostics). A large number of these models, including two best-selling cars, Honda Accord and Toyota Camry, have adjustable ignition timing. In 1996, a far more sensitive, complex system called OBD II came into use. With either system, there's no guarantee that an ignition timing error will produce a Check Engine light or even a trouble code.

Checking ignition timing is a straightforward procedure—if you have a good timing light that's right for your car. The type you need depends on the type of timing marks on your engine. Get the timing spec from a shop manual or an after-market service information system.

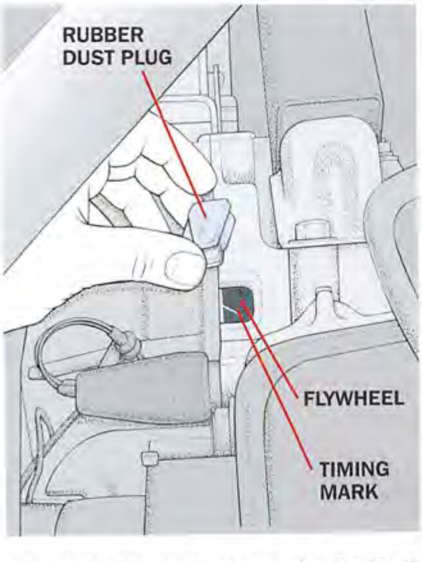

(Fig. 2) The timing mark on the flywheel may be under a dust cover or plug.

Finding Timing Marks

Locating the marks may not be easy if they're on the engine's front and are covered with dirt (Fig. 1). The marks may be found on an opening in the transmission bellhousing. (A rubber plug may be used to keep road film out of the bellhousing. Remove the plug for the timing check (Fig. 2).) Clean off the marks if necessary, and highlight the specified timing with a piece of chalk or white nail polish (Fig.3). Keep the highlight thin since a wide mark can span several degrees. You want the timing to be right on.

When a timing light is aimed at the timing marks with the engine running, a fixed mark should align with the illuminated mark on the spinning part, be it the crankshaft pulley or damper, or the flywheel. If the marks don't align, make the necessary adjustment of the distributor or go through a shop manual diagnostic procedure to find out if a sensor or the computer is bad.

If there's just a fixed point and a single line on the crankshaft damper or pulley, you probably can't check timing with a simple timing light. When those two marks are aligned, the timing is zero unless otherwise marked. Zero means the spark would arrive when the piston is at the very top of its stroke—"top dead center." The actual timing is typically some number of degrees before (or even after) top dead center. If there's no mark for that specification, all you'll see when the timing light flashes is that the two marks are some space apart. Is it the correct number of degrees? You need an adjustable timing light. With it, you can change the timing of the strobe light so that the light flashes earlier or later. You then read the dial or digital display to determine how many degrees you had to change the flashing of the timing light before the timing marks

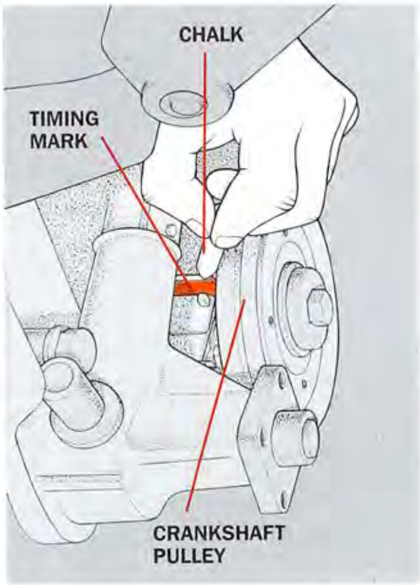

(Fig. 3) Chalk or light-colored touchup paint will make it easier to see the timing mark.

HOW IT WORKS

Distributorless Ignition

An electronic module with a double-ended ignition coil for each pair of companion cylinders is the most common type of distributorless ignition. The companions are cylinders with pistons that are in the same position in their respective cylinders at the same time—one on compression stroke ready for spark ignition, the other on exhaust. The coil module also contains electronic circuitry that communicates with the engine computer and executes its commands for ignition. There are three key inputs to the computer: the crankshaft and camshaft position sensors and a detonation (knock) sensor. The first two generally are electromagnetic devices that react to precisely located gaps in trigger wheels on the shafts. Changes in voltage caused by changes in electromagnetism provide signals from these sensors that tell the computer when each piston is rising on compression stroke. So, the computer can tell the coil module exactly when to discharge an ignition coil to provide a spark. The computer then can keep track of the sequence of cylinders for the electronic coil module to fire (so long as the engine is running). All the system continues to need are the crank sensor signals. If the weather, fuel quality or operating conditions produce engine knock, the detonation sensor sends a signal to the computer, which retards ignition timing. Each ignition coil discharges a spark from both ends, but only the one in the cylinder on compression has a fuel mixture to ignite. The other spark dissipates harmlessly in the exhaust gas of the companion cylinder. Ignition timing, along with fuel injection, is computer-adjusted primarily according to changes in throttle position, rpm, altitude and vehicle speed.

COMPUTER-CONTROLLED COIL IGNITION

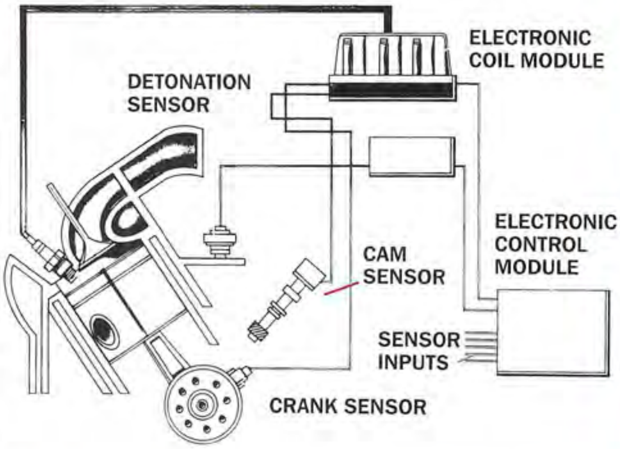

This ignition uses a separate coil for every pair of cylinders, usually in a pack of three or four coils mounted together.

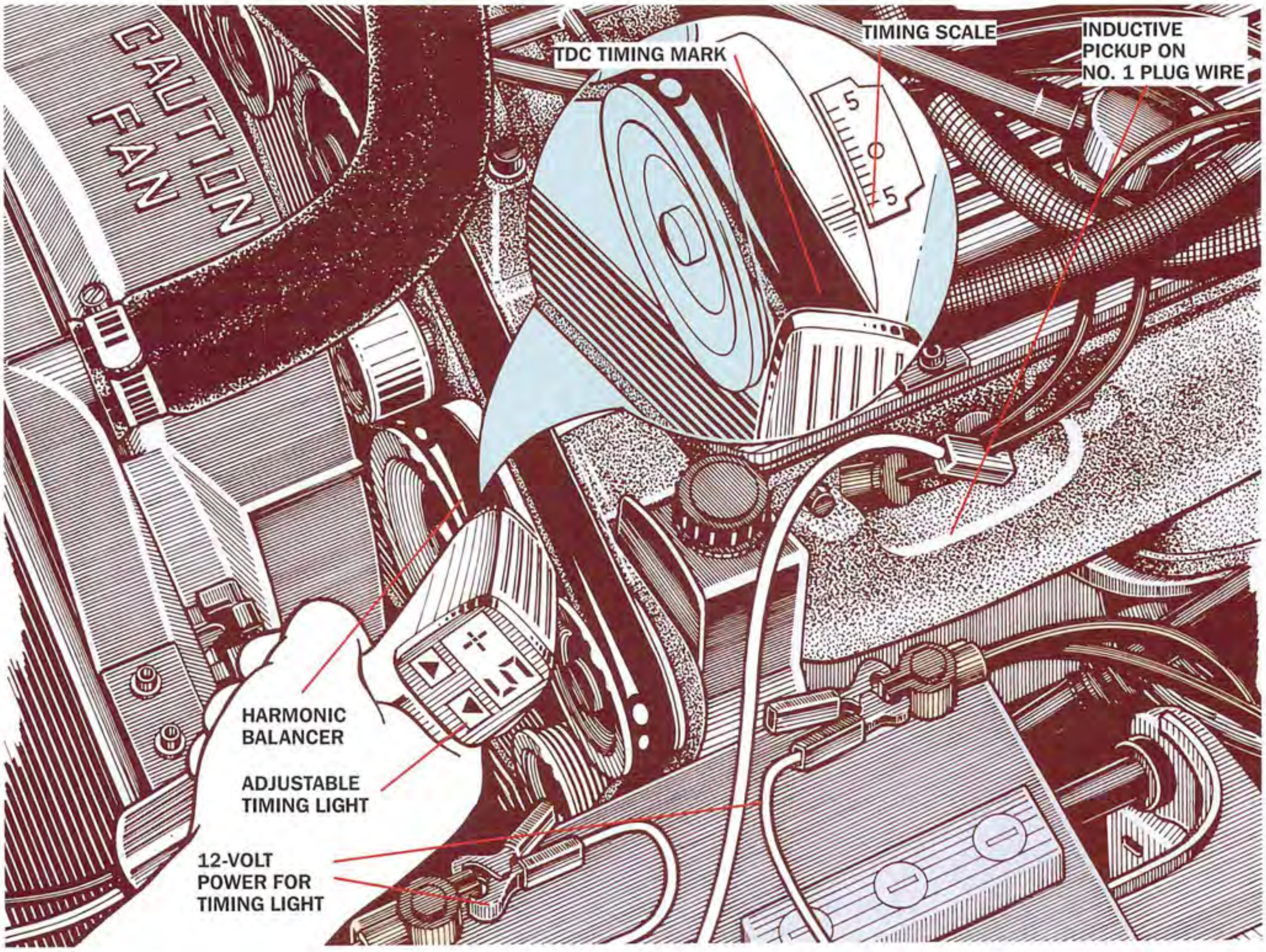

(Fig. 4) Modern timing lights often have an adjustable delay/advance function that allows you to set ignition timing at a value that's not printed on the timing tab.

appeared to align. If you changed it 16° in the retard direction, the timing really is advanced 16°. If the factory spec is 10° advance, the timing is too far ahead. If you changed it 6° but the timing specification is 10° advance, the timing is retarded. In either case, an adjustment is needed.

Okay, that's the way it should work. Let's go through a real timing check to demonstrate. First, the timing marks have to be clean and visible. Visibility often is tough, because of engine-compartment crowding. You may have to unbolt and hang a fluid reservoir out of the way to get a straight-on viewing angle. That's important, because if your viewing angle is bad, you get what's called "parallax error," and it can mislead you by several degrees.

Connect the timing light. The power and ground lead go to the battery, the lead with the plastic clip (called the pickup) goes to the No. 1 spark plug **(Fig. 4)**. The pickup must be squarely positioned on the plug wire. On an inline four-cylinder or six-cylinder engine, the plug nearest the belt-and-pulleys end of the engine is virtually always No. 1. If you have a V-type engine, you'll have to check the manual, since No. 1 could be the first plug on either side.

After connecting the timing light, start the engine and warm it up, so it runs at idle speed. It's rare, but some manufacturers specify checking timing at a speed other than idle. If the timing isn't correct, you've got to find out why. The usual reasons are a defective crankshaft sensor (the part that reads crankshaft position for timing adjustment by the computer) or throttle-position sensor. There may or may not be a trouble code to guide you, but with an ignition-timing check, you've made a good start. Scan-tool and multimeter tests of the sensors can take you the rest of the way.

THIS FORD REQUIRES its connector to be disconnected before you check the base timing.

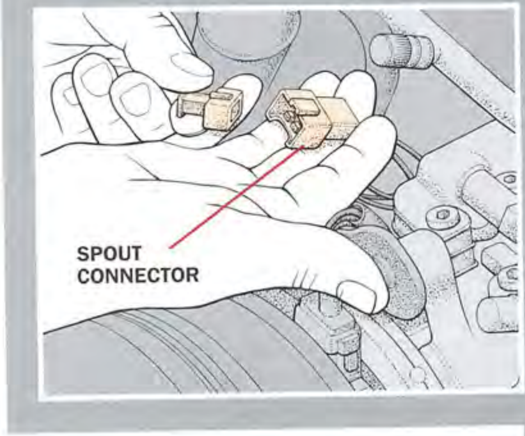

MODERATE

Replacing Engine Mounts

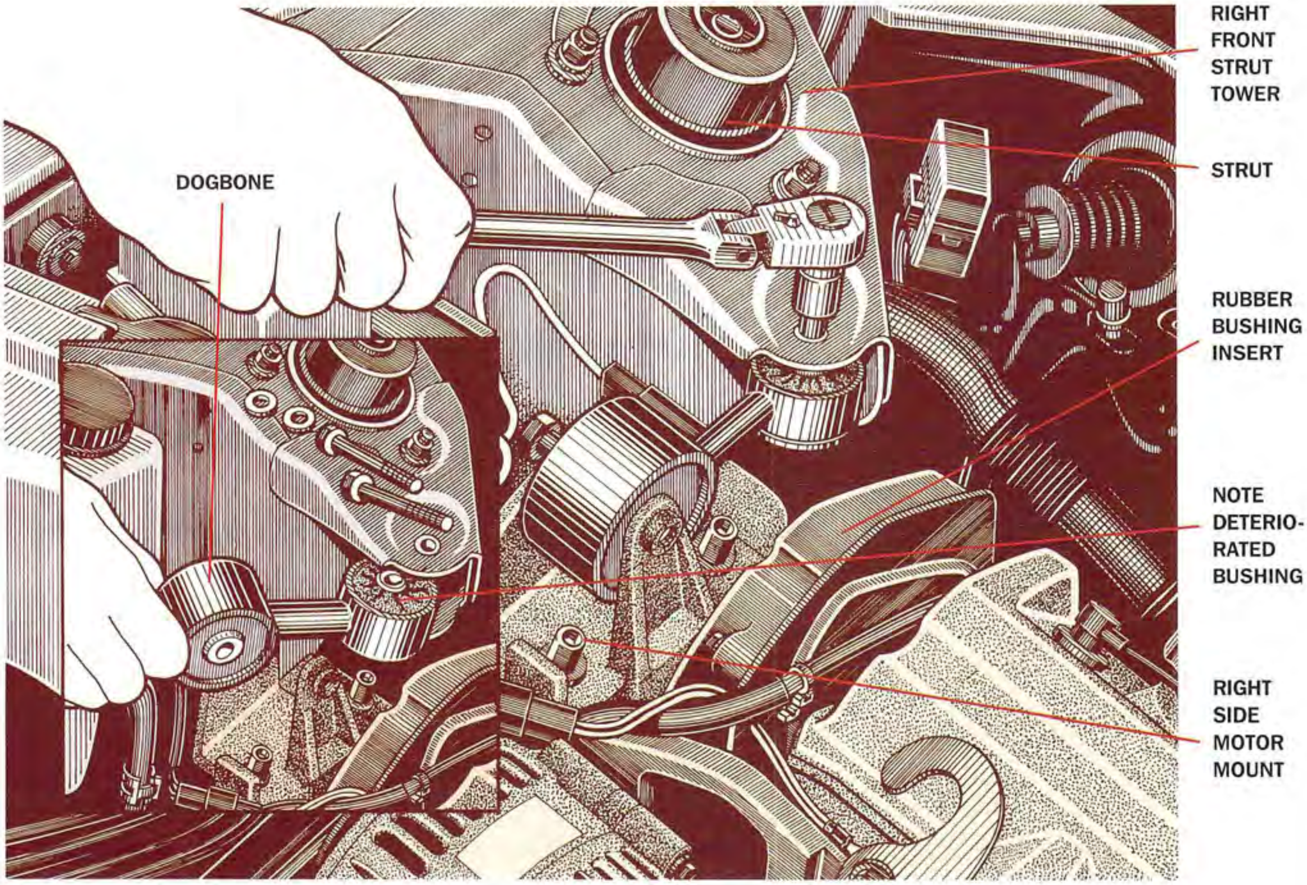

(Fig. 1) Check for deteriorated rubber bushings in the dogbone (shown) and other engine mounts that may permit the drivetrain to move too far and shake at idle. Replacement is usually straightforward.

You pull up to a traffic light and once you've stopped, you can feel the engine shaking. Okay, it's not so bad while you're stopped, but as you accelerate you can feel the engine clunk and shudder. Hey, you just did a tuneup last month, and you know it should last for a couple of years. You start to think, "I must have left a plug wire loose or something like that."

Lift the hood and check every wiring connection in sight, every air-duct clamp, every vacuum hose. Nothing looks out of place. But the engine is shaking. If you have a rear-wheel drive with an engine-driven fan, you may see the hint of a fan "scallop" on the rear of the radiator core, a clear-cut sign that the engine is tipping forward on acceleration. It's time to check the engine mounts and "dog bone" struts before there's a major failure and possible sticking of the gas pedal (**Fig. 1**).

Today's engine mounts are anything but simple two-piece metal brackets with slabs of rubber bonded between them. They have engineered shapes, specific durometer measurements (hardness) and often have strangely shaped cutout areas called "voids"—air gaps that tune the mounts' operation. In some cases they have hydraulic chambers that are filled with silicone fluid, and also in limited use are ones with electronic controls.

Okay, your car isn't brand-new and that's why the mounts are likely to be worn. The rubber may have been soaked with road film and oil (from a leaking valve-cover gasket). But it's more likely, particularly at 50,000 to 60,000 miles and up, that the mounts have cracks and distorted shape from the high temperatures common in today's engine compartments. If the weatherstripping along the cowl is missing, for example, that breaks the seal for the back of the powertrain compartment, allowing air to flow over the top instead of circulating through the compartment to cool the mounts (and other components).

The mount at the drivebelt (front) end of the engine has the biggest job in today's transverse-engined models. It reacts to both vertical and fore-aft shake, so check it with a pry bar (**Fig. 2**) and look at the rubber. If it cracks open at the internal edges of any voided area, or if it exhibits virtually no resistance, it is not providing much control.

Check fore-aft mounts, also by prying. In these cases, however, you'll probably have to force the bar against the metal edges of the mounts to be able to pry hard enough.

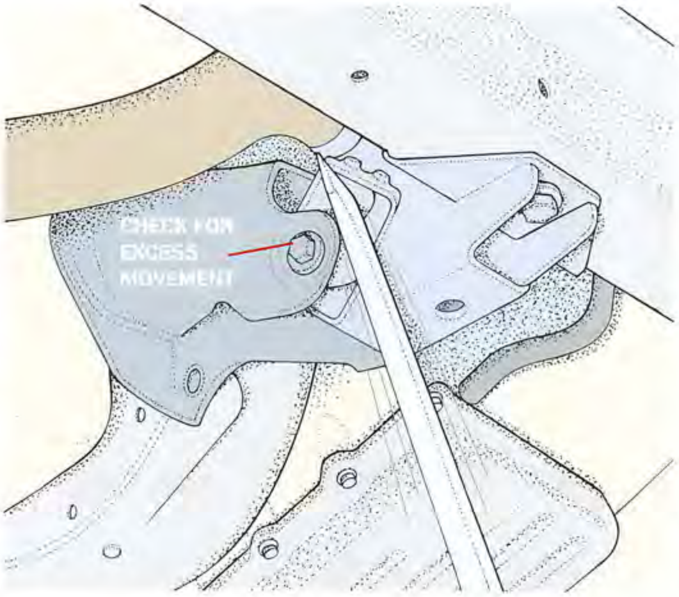

(Fig. 2) It might be necessary to use a pry bar to look for extra play in engine mounts.

What? Leaking Silicone?

Mounts with hydraulic chambers are circular and have a metal shell as part of their structure. It is possible for them to leak internally, but you're more likely to encounter an external leak. Surface cracks in the rubber generally are harmless, but run your fingers around them and if you pick up some thick, black goo that's not road film, it's the silicone oil leaking out. As with an all-rubber design, the only fix is replacement. You often can obtain aftermarket rubber/metal mounts to replace hydraulics at a significant savings. Are they likely to be as smooth? Probably not. Are they likely to satisfy you after experiencing the shake of a leaking hydraulic mount? Probably, particularly if the car is not destined for many, many years of additional use.

Replacing Engine Mounts

Although the powertrain end mounts on a transverse-engined car can be replaced individually if the one at the opposite end is good, it's better to replace both. When a fore or aft mount is bad on a transverse powertrain, always replace both. Ditto for the mounts on the driver and passenger sides of a north-south powertrain.

Support the engine with a hydraulic floor jack. If you have to place the jack on the oil pan, put a block of wood on the jack lifting pad to protect

HOW IT WORKS

Electronic-Hydraulic Engine Mounts

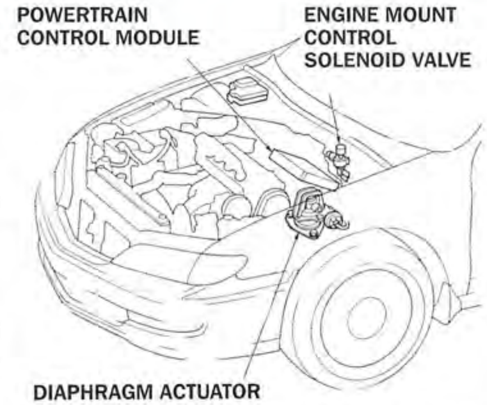
POWERTRAIN CONTROL MODULE
ENGINE MOUNT CONTROL SOLENOID VALVE
DIAPHRAGM ACTUATOR

The Honda electronically controlled mount is typical of high-tech mounts used in a variety of cars to dampen the last little bit of shake produced by the powertrain. It's a hydraulic engine mount with two fluid chambers connected by a large orifice (kept closed by a rotary valve) and a small orifice (always open). The small orifice allows the mount to operate much like a simple, fairly stiff shock absorber. At idle, however, a vacuum diaphragm unit operates the rotary valve to open the large orifice, permitting more fluid flow between the chambers. It's like a shock that suddenly becomes softer, in this case absorbing the idle shake of the engine. The vacuum (from the engine) is fed through a solenoid valve that is controlled by the powertrain computer. This system operates when the engine is idling, perhaps triggered by turning on the a/c (which produces a load that affects engine idle). The computer opens the solenoid valve, and the engine vacuum flows to the vacuum diaphragm unit at the engine mount. When the solenoid valve closes above idle speed, the vacuum is vented and the engine mount goes back to single-orifice operation. If the vacuum reading fails to hold steady, there's a vacuum leak in the mount and it should be replaced. The next stage in electronic control of engine mounts already is in limited use (on some Lexus cars): A diaphragm pulsates in the hydraulic mount. The diaphragm creates a proportional counter-vibration to the frequency of the engine shake in the 600-to-900-rpm range, canceling the shake so the engine idles smoothly. If one of these mounts ceases to work, you may need to troubleshoot the wiring and check the mount itself.

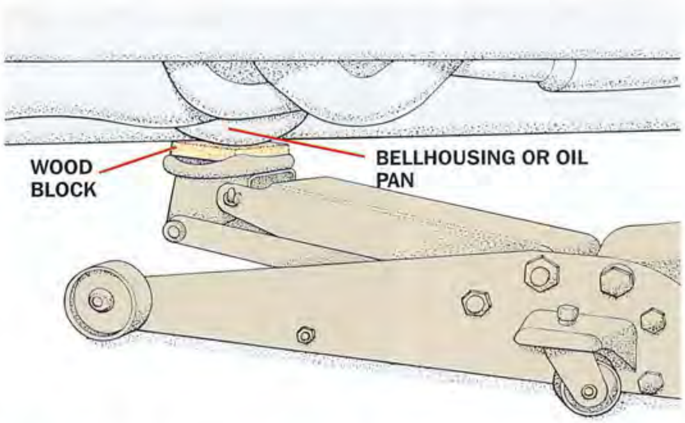

(Fig.4) Pad the nose of a floor jack with a block of wood to support the engine while you replace the mounts.

THIS MOUNT USES a soft rubber bushing. Install it with equal gaps on both sides of the bushing.

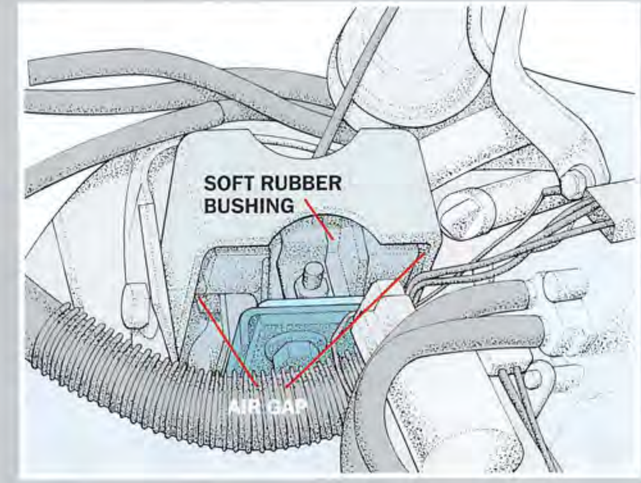

the pan (Fig. 3). If the mount's nuts or bolts are frozen, spray them with penetrating solvent so you can loosen them with a wrench and not have to resort to a chisel and hammer. In most cases, removing the fasteners and jacking the engine up slightly is all that's necessary to slide out a defective mount and insert the replacement. If you have to jack up a lot, there's a chance you could stress another mount unless you loosen it. Check before you pump the jack handle.

If there's a stud in the mount you're replacing, make sure you have it aligned precisely with its hole before you lower the engine. On many vehicles, you have to work both from underneath and above to insert and tighten nuts and bolts, and you occasionally will need a helper to pry on the engine while you get a bolt started through its mounting hole. If the mounting holes are slotted, run the engine and let it idle for a minute or two so the engine settles in the mount. Then tighten the nuts and bolts. Do the final tightening with a torque wrench.

Shock Absorber and "Dog Bone" Struts

Some engines have a shock absorber for vertical shake control. When used, it typically is bolted to the engine at the top end, and a chassis crossmember at the bottom. Disconnect the shock at either end and operate it by hand, just as you would to check a shock at a wheel. If you feel looseness or roughness with lost motion, that's a sure indicator of a worn-out shock.

Is there a dog bone strut or two? The dog bone usually is at the top of the engine compartment, at the fore side of a transverse powertrain. However, usually does not mean always. It may be installed at the top rear side (typically at the accessory-belt end, where it can directly assist the engine mount), or even at the bottom center on the aft side of a transverse engine. So if you don't see one at first glance, don't assume it isn't there.

(Fig. 5) This inexpensive vacuum pump will make it easier to diagnose bad actuators on some hydraulic motor mounts.

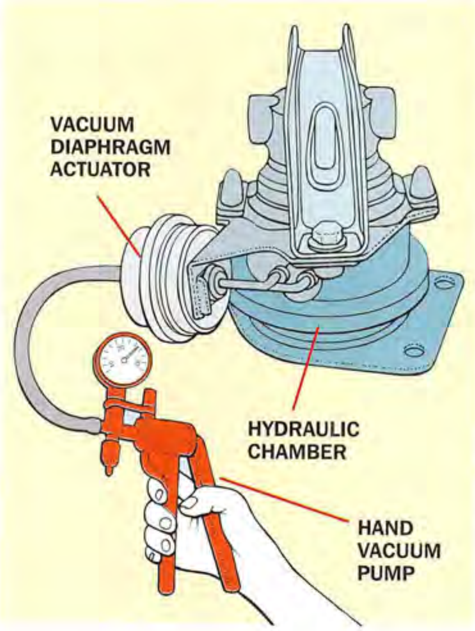

The ends of the dog bone have rubber sleeve bushings (Fig. 4), and if the rubber is cracked or distorted, replace the entire strut (the bushings are not replaceable). It's a straightforward job. Just unbolt it at each end, and install the new one. The engine will shift slightly when the old dog bone is removed, in which case you'll have to pry it forward into alignment with the bolt holes. You also may need a helper for this operation.

Some premium aftermarket dog bone struts include hydraulic chambers, much like a shock absorber—so the damping is done hydraulically instead of by squeezing rubber bushings. There's no noticeable difference in engine smoothness, but the hydraulic type reportedly has better long-term performance. The electronic-hydraulic engine mount (see page 108) may look complex, but it's easy even for a less experienced Saturday mechanic to check. Turn on the air conditioning with the engine idling, and then disconnect the wiring at the solenoid valve. If the engine shakes noticeably more, the system is working. If there's no difference, first disconnect the vacuum hose at the engine mount to check for vacuum flow through the solenoid valve. If there's vacuum flow, turn off the engine and apply vacuum to the diaphragm unit neck with a manual vacuum pump (Fig. 4).

MODERATE

Cleaning Throttle Bodies

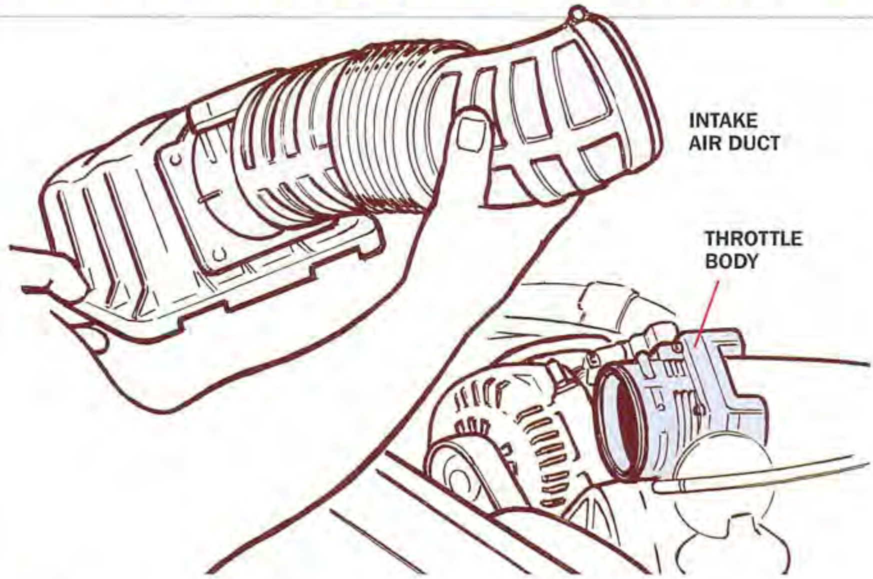

(Fig. 1) Remove the air duct and inspect it carefully for leaks or tears that might permit uncleaned air into the engine.

You slip behind the wheel for the morning commute. The engine starts easily, runs fine while fast-idling down the driveway and down the street for a few miles, but then stalls. You know instantly when the stall occurs because the air horn in the tractor-trailer rig immediately behind you curdles the fluid in your inner ear's semicircular canals when forward motion fails to proceed at an orderly pace. You restart, your car accelerates with a slight stumble, and when you coast to a stop at a traffic light, it stalls again. The stumble and possible stall problem doesn't go away when the engine is warm, so the common cold-engine poor-running possibilities are scratched from your priority checklist. This is more than annoying and as soon as the weekend arrives, the hood goes up.

Diagnosis: Deposits

All sensor wires are connected. There's no Check Engine light and there are no trouble codes. What do you look for now? A disconnected or damaged vacuum hose? Sure, that's a good bet, because many vacuum hoses can be pinched and damaged where it's really difficult to see. But take the time to check them all, making sure they're not only on tight, but not cracked or burned through due to contact with an exhaust pipe or exhaust gas recirculation tube. Don't be surprised if the last hose you check is the one that's damaged. If all the hoses are intact and on tight, there's another possibility that's just as likely: dirt and gum accumulation in the throttle body.

Get Dirty

To check for this, you've got to remove the intake air duct between the air cleaner housing and the throttle body (Fig. 1). First, disconnect all hoses and unplug any sensor wiring connectors (Fig. 2). If there's a chance you could confuse a hose connection or wiring connector, put a piece of masking tape on the hose or connector, another on the hose neck or sensor, and mark each with the same letter.

Next, loosen any clamps, work the duct off the throttle body and set it aside. With the wiring and hoses disconnected, you should not run the engine. Even if it starts and runs, it will log trouble codes, possibly triggering the Check Engine light. Then you'll have to go through a code-clearing procedure—extra work you don't want. Moreover, the computer may have to relearn some driveability trim settings, which may leave you with a marginally running engine for a while.

On some vehicles, it's possible to leave hoses and sensors connected and still move the intake air duct safely out of the way. In that case, it would be possible to spray a solvent into the throttle body opening with the engine running. However, there's no great advantage to this and, as you'll see, there are reasons why it's not such a great idea.

Take a good look inside the throttle body using a flashlight. Operate the throttle linkage to open the throttle plate so you can see past the outside surface of the throttle body. If you see a coating of dirt and oily film on the inside wall of the throttle body or the edge of the throttle plate, you've likely found the problem. The coating both upsets and restricts the airflow when the throttle is closed or slightly open.

Where do these deposits come from? Some are in airborne dirt that gets through the air filter or past a crack in the intake air duct. So be sure to check the intake air duct for cracks, particularly in the "accordion" area where they're not as obvious. Most deposits, however, are from oil gases transferred from the positive crankcase ventilation system and pushed forward by normal engine pulsation as the intake valves close and open, and from exhaust gases that seep in.

Some throttle body bores have a coating to reduce the buildup, but, in time, bores with protective coats can be affected. There are several ways to

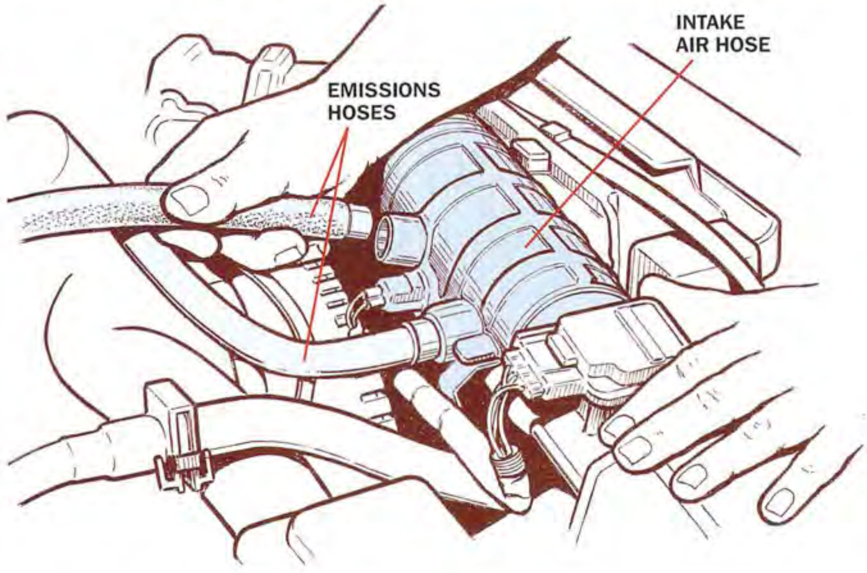

(Fig. 2) Remove any hoses or wires on the air duct connecting the throttle body to the air cleaner housing. Mark which hose or wire goes where to avoid confusion later.

clean out the area. The best is with a professional tool called the Intake Snake, which comes with an effective yet safe solvent, and the simplest is with an old worn-out toothbrush with soft bristles and a mild solvent.

There are three reasons why you have to be careful, both in choice of solvent and in application:

First, if the throttle body has a protective coating (as on many Ford products) to reduce the buildup, a strong solvent and hard brushing will remove it, so you'd have to do the job more often. If you see a warning label on a Ford product, that's why.

Second, there may be a sensor tip projecting into a small opening in the area of the throttle plate, and a strong solvent or hard brushing might damage it. In addition, a sensor O-ring seal may be damaged by a strong solvent.

Third, the throttle plate shaft is sealed at the mounting holes in the throttle body to prevent entry of unmeasured air (which would upset the fuel mixture). A strong solvent (and hard brushing) could damage the seals.

Stronger Than Dirt

Most aerosol carburetor and choke cleaners are pretty strong. Ditto for those aerosol cleaners that can be sprayed into the air intake with the throttle open for combustion-chamber cleaning. These aerosols have to be very strong to clean surfaces without the mechanical advantage

HOW IT WORKS

How A Throttle Body Works

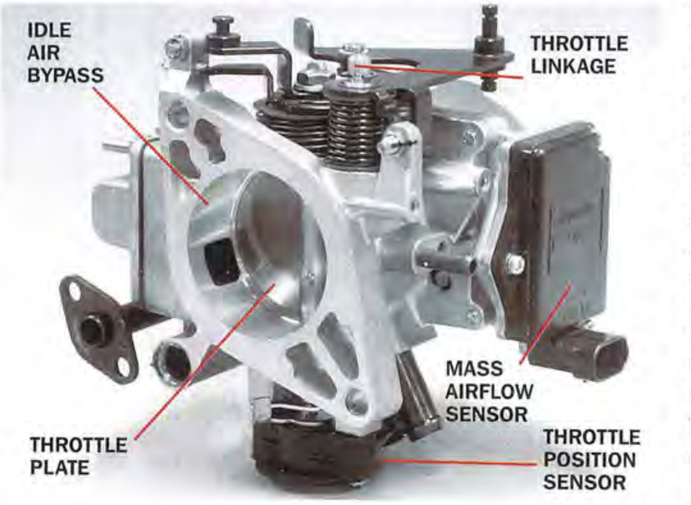

An engine runs primarily on air (about 15 parts by weight, to one of fuel), and the throttle body in a modern fuel-injected engine is the device that controls the airflow through its round opening. When the intake valve for a particular cylinder is open, the air flows through the throttle body, which is mounted on the intake manifold. The airflow continues into the intake manifold, through a chamber, and then through the open intake valve into the cylinder. At the same time the fuel injector sprays in fuel, the inrushing air and fuel mix, the intake valve closes and a spark from the plug ignites the mixture. The intake airflow control is provided by the throttle plate, a pivoting plate on a spring-loaded shaft that goes through the center of the throttle body's round opening. The throttle plate is a type of air valve, often called a "butterfly." When your foot is off the gas pedal, the throttle plate springs closed, and only a minimum amount of air, enough for engine idle, goes through a bypass around the throttle plate. One end of the throttle plate shaft also holds the movable contact arm of the throttle position sensor, a variable resistor-type sensor that tells the engine computer whether the vehicle is idling, accelerating or holding a throttle position. On many new engines, the gas pedal is just a variable resistor, sending a signal to the engine computer. The computer controls the throttle plate to produce the desired opening. This is called "drive by wire."

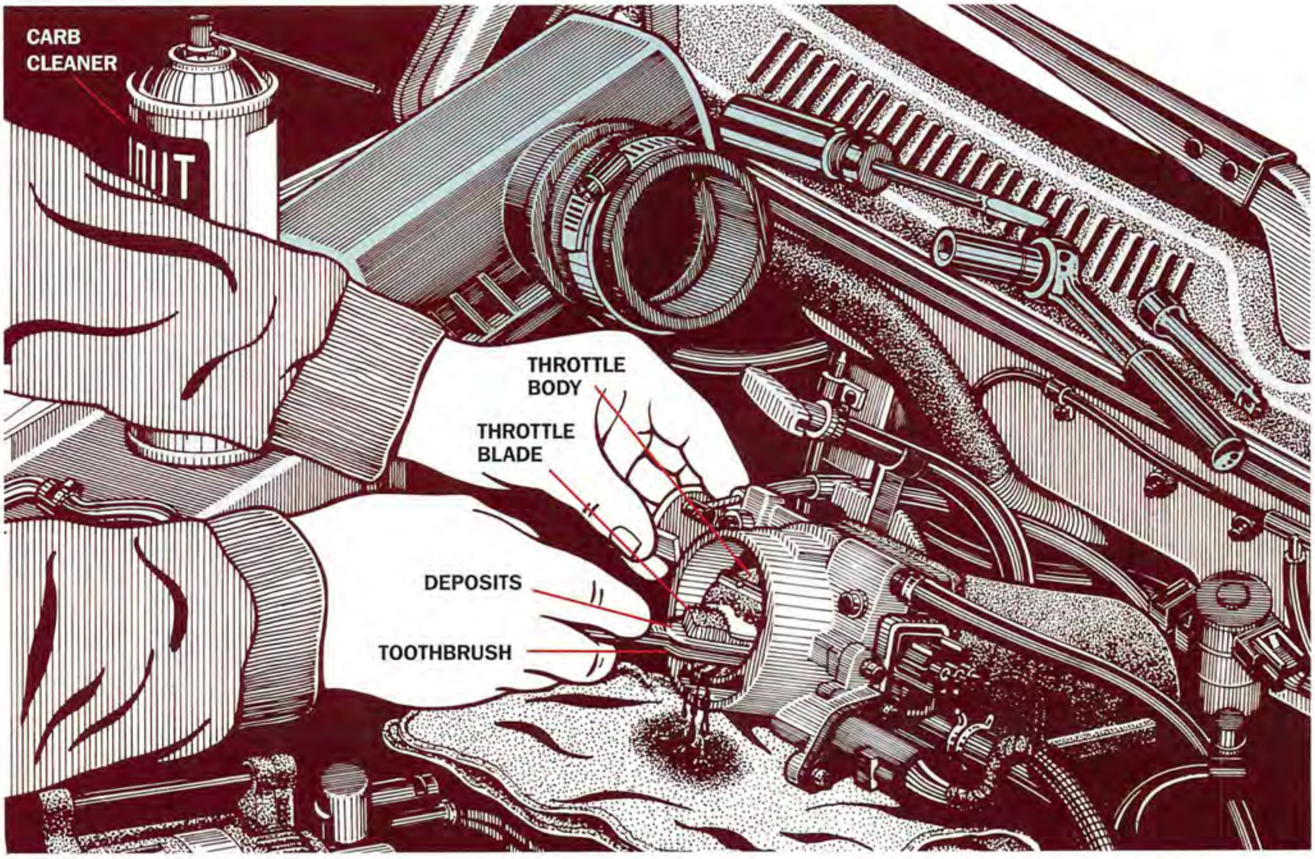

(Fig. 3) Spray-can carb cleaner may be a little too aggressive for safety—dilute it with gasoline to prevent damage to seals and sensors. Always remove the negative connection from a battery first and reconnect it last to prevent shorting the wrench out on nearby metal objects.

of a rubbing brush. A strong solvent can cause deterioration of not only sensor seals, but also throttle shaft seals. Further, you can't see which deposits have been removed and which have not. The aerosol spray will not clean all areas, particularly the back side of the throttle plate.

A tuneup/injector cleaner mixed with gasoline (1:4 or 1:5 ratio) should be safe (**Fig. 3**). You won't be using much, and you can pour the rest into the gas tank after you're done.

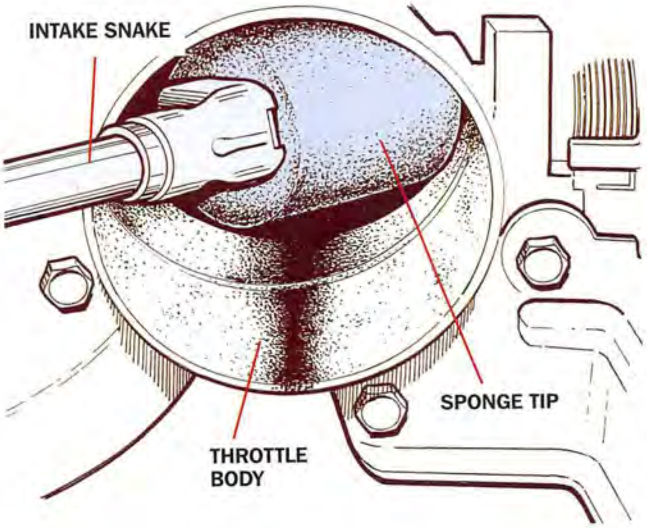

(Fig. 4) The Intake Snake is one way to safely clean throttle blades and throttle body castings of deposits.

Start by cleaning the exterior of the throttle plate, and then hold it open using a wire tied to the throttle linkage so you can clean the inside perimeter of the plate. All you're trying to do with brushing is to loosen all the deposits.

Next, clean the throttle body wall, being careful to work around any electronic sensors, and around the throttle plate shaft holes. When you're done, remove the deposits with solvent and a clean rag.

Because this service is something your vehicle may need every few years, the professional kit (www.intakesnake.com) is a worthwhile investment at under $60. It includes a container of two cleaning tips (containers of two tips are available separately for under $15). The "snake" is made of a smooth plastic and won't scratch anything. It's 14 in. long and moderately flexible, so you can reach deep into the throttle body and follow any contours (**Fig. 4**). The tool holds the spongelike work tips that are coated with a hypoallergenic solvent that's also safe for sensors and throttle plate seals. Because the dirty film is held by the work tip, there's little or no wiping needed when you're done. Press a release tab and the dirty tip falls off the tool (into a garbage can).

Electric Cooling Fan Problems

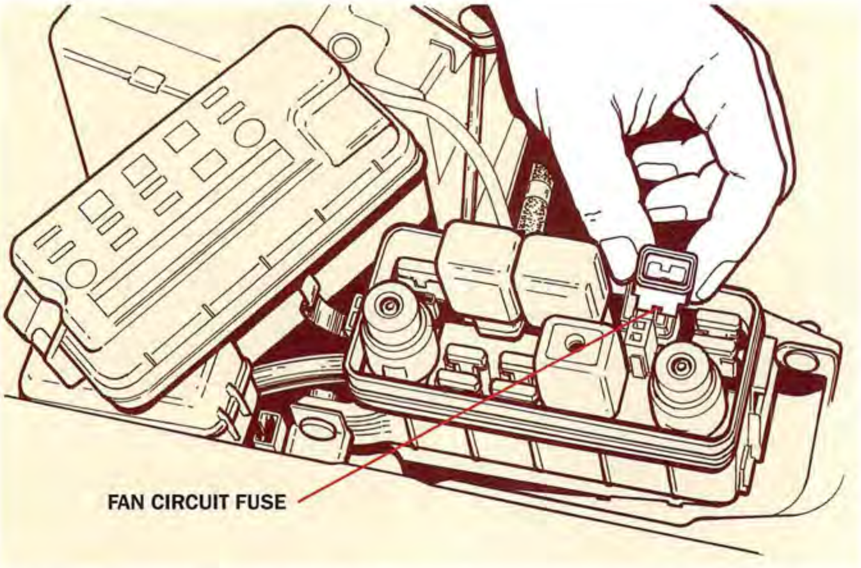

(Fig. 1) Check the high-amperage fan fuse in the underhood electrical box.

It's a really, really warm day, but the temperature gauge reads normal and the a/c is blowing out a lot of cold air, so who cares? Then the traffic suddenly slows to a crawl and doesn't get any better. Gradually the temperature gauge needle creeps toward the Hot end of the dial, and the air blowing out of the a/c registers is anything but cold. In fact, to forestall overheating, you shut it off. Even worse, you soon have to turn on the heat to try to prevent a boil-over, and open the windows to try to get a breeze.

If you're lucky, the traffic tie-up breaks, and you're cruising again. The gauge needle starts dropping, and you turn off the heater. Still sweaty, you turn on the a/c again, and even if the air isn't frigid, it's cold enough to make you feel better.

Later, you lift the hood with the engine running near an overheat and the a/c on—the radiator electric fans aren't running. But why? The fan may be under the complex control of a powertrain computer, but the basic circuit is pretty straightforward.

On almost all cars the fans should go on if the a/c is turned on, triggered by the powertrain computer or by an a/c pressure switch. The fans may not go on instantly, but certainly within a couple of minutes. If they don't, proceed to a circuit diagnosis.

The simplest first circuit check you can make is the current feed, starting at the fuse, which on late models is in an underhood fuse box **(Fig. 1)**. Just pull it out and look. Replace it if it's blown. On older models it's often a fusible link—a short section of wire typically four gauge numbers higher than the wire of which it's part. You may see it's singed. Even if it isn't, pull on the ends of the wrap, and if it stretches, the wire underneath is broken.

Checking The Fan Motors

Fuse okay? The most practical approach is to go next to the end of the circuit: the fans themselves. Most cars have two fans. Unplug one fan's wiring connector and (referring to the wiring diagram color codes) hot-wire the fan motor's half of the connector. Run one jumper wire from the current feed terminal of the connector half to the battery's positive terminal. Attach a second jumper wire to the other terminal of the connector and run it to an electrical ground **(Fig. 2)**. The fan should run. If it doesn't, the fan motor apparently is defective.

Some cars have a single fan that runs at two speeds. In these cases, the connector may have:

- Two terminals (one from a dual power feed—through a dropping resistor for low speed and a resistor-bypass circuit for high speed—plus the second one for the ground).

There's a splice or connector for the dual power feed earlier in the circuit. If you check by hot-wiring at the fan motor itself, it will run at high speed.

- Three terminals (one for low-speed feed, one for high-speed and a common ground). Check each power feed terminal separately, using the common ground in each case.
- Four terminals (high, low and two grounds).

(Fig. 2) Try jumping the battery direct to the fan to see if it spins.

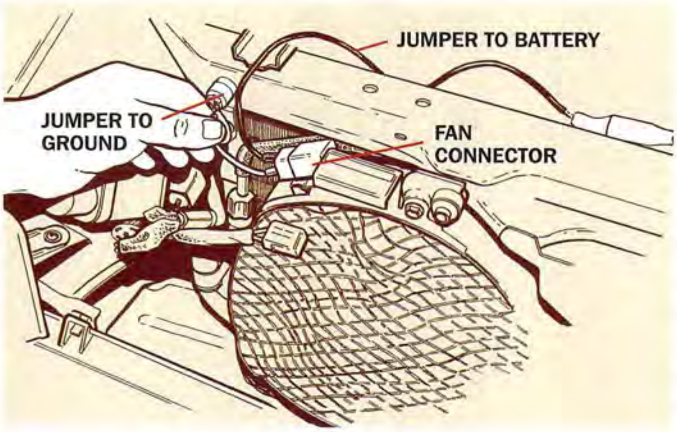

THERE MAY BE SEVERAL temperature sensors and switches on the engine. Find the correct one and check it with a test light or an ohmmeter.

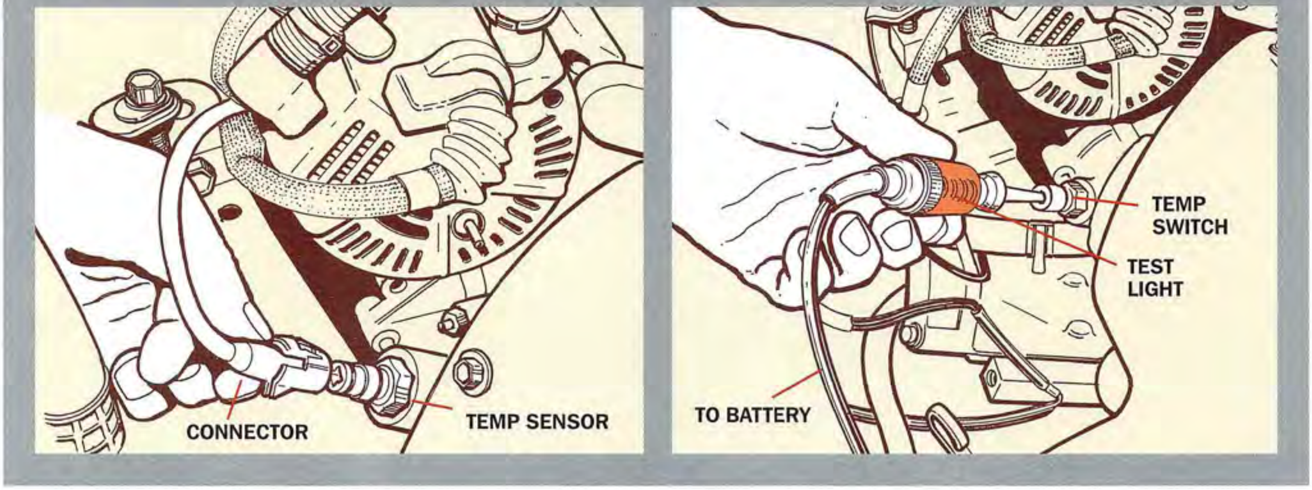

Hot-wire each pair separately with your jumper wires.

Testing Coolant Switch Or Sensor

If both fans run when they're hot-wired, the problem is somewhere between the current feed at the fuse and the fan motors. On older systems where the trigger for the fans is a coolant temperature switch, unplug the switch connector. Be sure you have the right one, as some vehicles have as many as three: one for a dashboard warning light, one for an overhead console and one for the powertrain computer.

With the engine running and with the coolant switch unplugged, the fan should come on because the powertrain computer logs a trouble code for a failed switch and turns on the fans in most systems. If it doesn't, and the switch is a typical domestic design, it has a normally open type that stays open until coolant temperatures are high, when it closes to turn on the fan. To double-check, just ground the unplugged connector's single wire terminal, using a jumper wire. The fan should spin. Many older Japanese cars have circuits with normally closed switches. They open with high coolant temperatures to turn on the fan, so a simple unplugging (with the ignition on) always should get the fans to operate with this setup.

Most late models have a coolant temperature sensor. It isn't a switch, but a temperature-sensitive resistor (thermistor) that sends a signal to the powertrain computer. The computer, which also has some control over a/c operation, then decides when the fans should operate and at what speed. You may not be able to get a response by simply unplugging the sensor connector or even by grounding it. Instead, check its calibration by connecting an ohmmeter (Fig. 3). The thermistor develops lower resistance as the temperature goes up, higher resistance as it

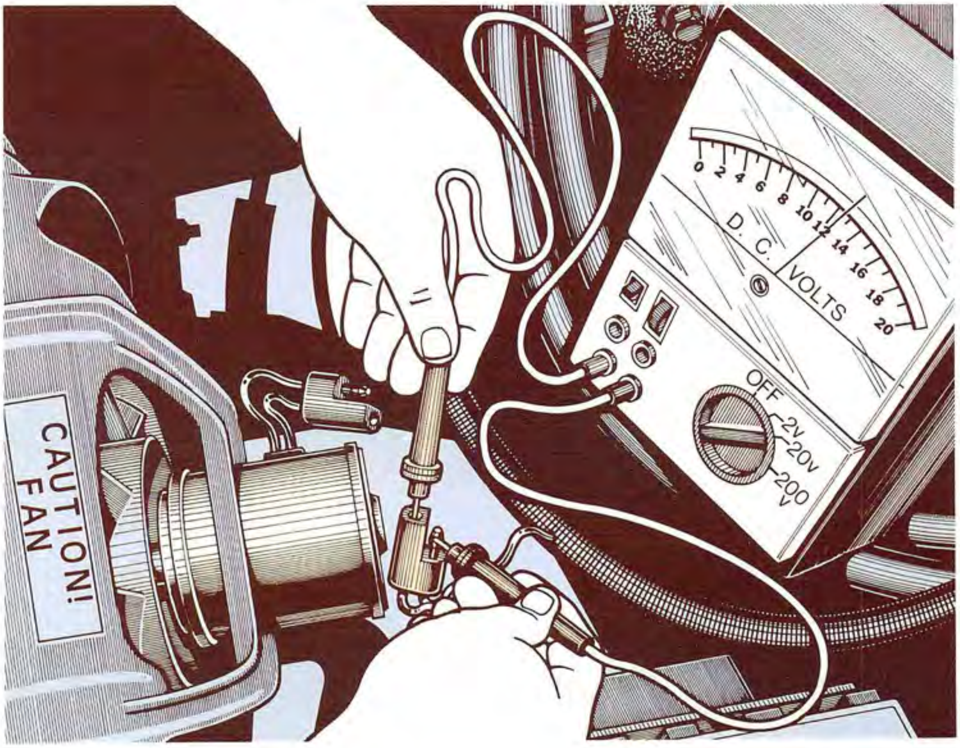

(Fig. 3) Connecting an ohmmeter is one reliable way to check the thermistor.

drops. If the sensor reads a thousand ohms or more (which signals low temperature) when the engine coolant is warmed up, it's out of calibration. Replace it. Even easier is to use a scan tool which will give you a reading in degrees (C. or F.).

Relay Diagnosis

Head for the fan relay (or relays—some systems have low- and high-speed fan relays) (Fig. 4). Turn on the ignition, and locate the relay terminal for the wire from the switch. When you ground its terminal with a jumper wire, the relay should click and the fan should run. If it does, the relay and its circuit to the fans are good. Any problem is in the wiring between the coolant temperature switch and the relay.

If the fan still won't spin, continue at the relay, which has two current feeds, one to the electromagnetic coil, one to the electrical contacts. Probe the two current feed terminals of its wiring connector with a grounded test light. With the ignition on, it should go on in both cases. No light in one? There's a break in the wiring from the fuse to that relay terminal.

On systems with computer control via a sensor signal, make a similar test of the relay current feed terminals with a grounded (computer-safe) test light and the ignition on.

If the current feeds are good, ground the relay's switch terminal (the one with the wire that goes to the coolant switch). If there's a sensor and the switch terminal wire goes to the powertrain computer, unplug the wire before grounding the terminal. The relay should click and operate the fans. If it doesn't, replace the relay.

You may have trouble doing this with relays that plug into an underhood relay "center." Unplug the relay, turn on the ignition and with your test light probe the two current feed terminals in the relay center. If they pass (turning on the test light), make up short jumper wires to connect to the unplugged relay and one long jumper (that you run to an electrical ground) for the switch terminal. If the relay still doesn't work (ignition on), replace the relay. If the relay does click, probe the output terminal to the fan motor with a grounded test light. Light goes on? The problem is in the wiring from relay to fan motor.

Your diagnosis may point to the powertrain computer. It's rare, but a computer failure can be responsible—without a Check Engine light or trouble code. Perhaps just a single driver has blown, so the computer itself seems to be performing normally. In addition, in a number of cars, particularly some GM models, the powertrain computer turns off the fans if the vehicle is cruising 40 to 45 mph or higher. This strategy relies on the vehicle speed sensor, which may be misbehaving. If your speedometer is way off, or not working at all, it's something to consider if the relay works when grounded. Want to work through one of these problems? Open your factory service manual, plug in a Saturday Mechanic-level scan tool and go through the trouble-tree diagnostics to find the answer.

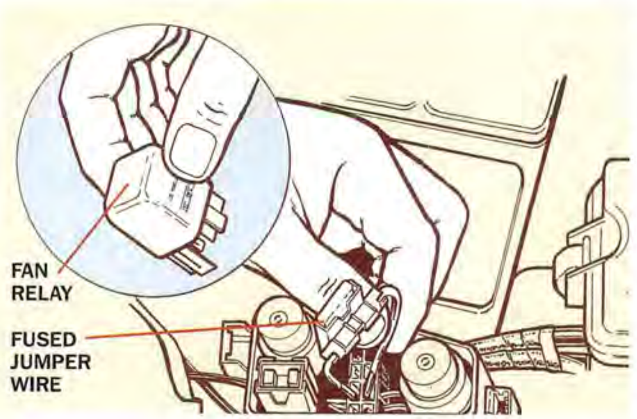

(Fig. 4) You can jump across the fan relay to confirm the relay's failure.

HOW IT WORKS

The Electric Fan

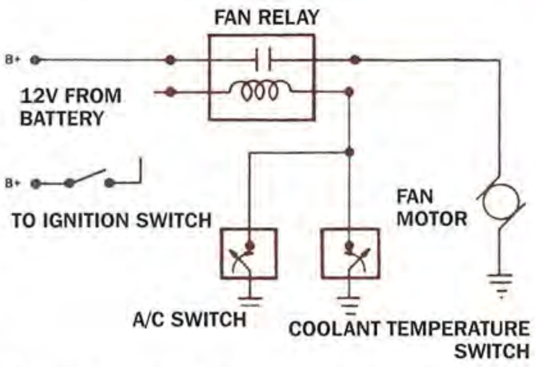

The fan is powered by an electric motor, which is wired to a relay. The relay is a magnetically controlled switch. It typically has four terminals, two of which are current feeds from a fused source of 12-volt electricity. At the relay, there is one current feed into an electromagnetic coil, and one feed into a set of contacts for the switch, which is open. The other terminal for that open switch is connected to the electric motor for the fan. The fourth terminal, also for the coil, is wired to the powertrain computer, which may look at various signals to determine whether or not to provide an electrical ground for that coil. If, as in this circuit example, it gets a high-temperature signal from the coolant temperature sensor, also wired to it, the powertrain computer can choose to ground the coil terminal. This energizes the coil, creating a magnetic field that pulls the arm on the power feed contact, so that it touches the terminal for the wire to the electric motor. That closes the switch of the relay. Current flows through the closed contacts of this switch to the electric motor, which spins to operate the fan.

MODERATE

Finding Engine Knock

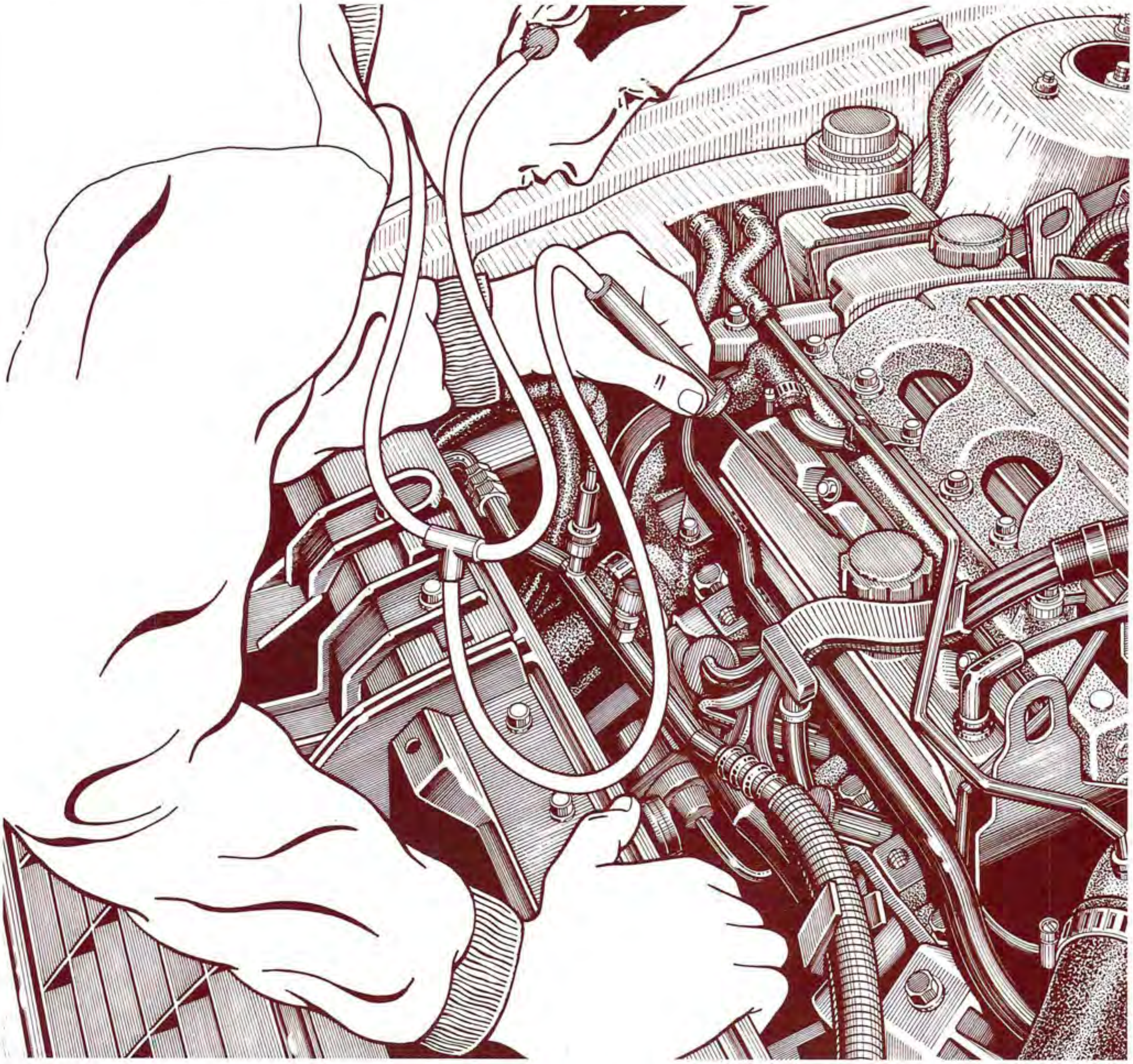

Fig. 1. Try to isolate engine noises with a mechanic's stethoscope. You can usually determine which cylinder is at fault. Try pulling one plug wire or injector harness at a time to see if the noise changes.

You're waiting in the drive-up line at the bank, all the better to redeem some of your hard-earned cash from the vault. But as you inch your way to the teller window, your engine misses so badly that it either stalls or has to be held at high idle with your right foot. The teller sends you a note through the pneumatic tube, "Sir, you forgot to endorse this check, and please get your car fixed—I can't hear a thing." It's off to the garage instead of the beach.

Days later, you're at the end of your rope. Above idle, the engine purrs like a kitten. But below 2000 rpm, there's a steady miss in the No. 4 cylinder. You've checked everything you can think of. Spark, compression and the fuel injector to that cylinder are working just fine. Yet the miss remains.

Complete Engine Testing

Many people—even mechanics—think that the beginning and end of internal engine diagnosis lie in using a compression tester. But the truth is that many internal engine problems (such as a weak valve spring, for instance) are not revealed by compression testing alone.

"A weak valve spring is easy to spot, if you use a vacuum gauge," said Dave Hakim, an engine and fuel system specialist with Federal Mogul Corp. "Connecting the gauge to the intake vacuum would reveal normal vacuum at idle. But when a valve spring is bad, the vacuum needle would fluctuate rapidly between 12 and 24 in. when you raise the rpm."

You can diagnose most internal engine problems using a vacuum gauge, a compression gauge and a leakage tester. And you can usually get to the root of mysterious internal engine noises by being a good listener.

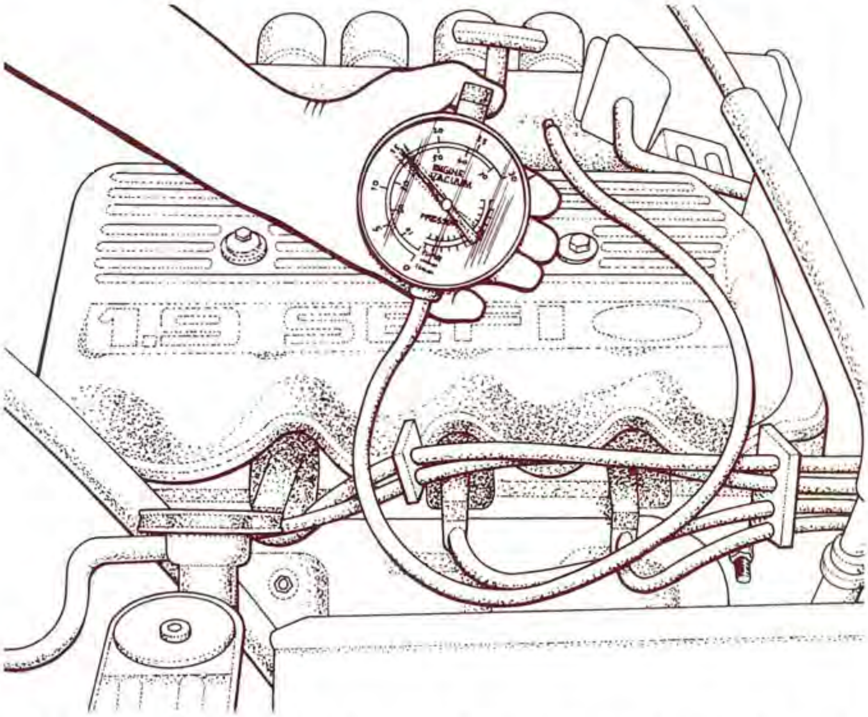

Fig. 2. With low, rapidly pulsating vacuum, this engine probably has valve guides that are worn out.

Engine Noises

The most common source of engine noise comes from the valvetrain. The noise—usually a clicking or tapping at half the engine's rotating speed—is caused by excessive clearance between components or a bad hydraulic lifter. On camshaft-in-block engines, you can isolate valvetrain noises by briefly running the engine with the valve cover off and pressing on each valve rocker with the end of a hammer handle. When the pressure from the handle removes the clearance, the noise disappears and you've found the problem.

Diagnosing valvetrain noises is more difficult on overhead-cam engines because you can't run them with the valve cover off.

Running the probe of a mechanic's stethoscope along the length of the valve cover with the engine running should allow you to isolate the noisy valvetrain component (**Fig. 1**). If you don't have a stethoscope, use a long screwdriver as a probe and listen by placing your ear against the handle.

Other internal noises include main- and rod-bearing knocks, wrist-pin noise and piston slap. A main-bearing knock has a deep metallic sound and is usually loudest when the engine is under load. If the knock is because of a failed bearing, low oil pressure may accompany the sound. Connecting-rod noises have a lighter rapping sound under light engine load. Disconnecting the spark plug to the offending cylinder usually results in a marked decrease in the noise. Piston slap, caused by excessive piston-to-cylinder wall clearance, causes a dull rattle. Disconnecting the spark plug makes it go away. Wrist-pin noise usually gets louder when the spark plug is disconnected.

Keep in mind that there are many external engine noises that can mimic internal noises, such as a loose flywheel or broken flexplate. Drivebelt pulleys can also cause noise, which can be isolated by running the engine briefly with the drivebelts removed.

Vacuum Testing

A good manifold vacuum gauge costs only about $35, yet it can quickly tell you a lot about what is going on in an engine. A vacuum gauge can reveal how well the valves are working, whether the piston rings are sealing adequately or whether camshaft timing is retarded.

To perform a manifold vacuum test, begin by connecting the vacuum gauge to an adequate source of intake-manifold vacuum.

On most engines it's easiest to tee the gauge hose into a vacuum hose such as the hose that runs from the intake manifold to the vacuum brake booster. With the gauge connected, start the engine and let it warm up to operating temperature. On most properly running engines, the gauge will hold steady between 15 and 20 in. If the needle fluctuates at idle, it indicates that one or more valves are leaking. Doing a compression test or an engine leak-down test will reveal the leaking valve or valves.

A slightly low vacuum-gauge reading, along with a slight flutter of the gauge, may indicate that the valve guides are worn (**Fig. 2**). On engines without overhead camshafts, you can verify guide and seal conditions by removing the valve cover(s) and squirting engine oil on the tops of the valve guides while the engine is running. If the vacuum-gauge reading increases and blue smoke starts to appear from the tailpipe, it's a good indication that the valve guides are worn.

If the vacuum reading fluctuates between 5 and 7 in., it indicates that valve timing is excessively late. A skipped tooth or excessively worn timing chain or belt are all likely causes.

To check for weak or broken valve springs, increase engine speed to 2000 rpm. If the gauge needle fluctuates between 12 and 24 in. and the fluctuations increase as engine speed increases, the valve springs are weak. If the fluctuation is irregular, it's likely that one valve spring is weak or broken.

If intake vacuum appears normal in the above tests and the engine oil is in good condition, you can do a quick test to determine piston-ring sealing. To do this, bring the engine speed up to 2000 rpm, then let the throttle snap closed. If the vacuum reading jumps 5 in. or more over the normal reading, it indicates that the piston rings are in good condition.

Compression Testing

A compression test can reveal a lot about an engine's internal condition, but only if it's performed properly. Many people mistakenly do compression tests with old diluted oil, the engine cold or with the throttle closed.

Begin by warming the engine to normal operating temperature. Then, use compressed air to blow all the spark plug wells clean before removing the spark plugs.

Disable the ignition, then connect the compression gauge to the No. 1 cylinder, according to the instructions supplied with the gauge. Have an assistant hold the throttle wide open while cranking the engine over until five complete compression strokes have been completed (as indicated by pulses on the gauge). Record the compression readings on the first and fifth strokes. Repeat the process for each cylinder.

A cylinder has good compression if it achieves two-thirds of its total compression on the first stroke and reaches the minimum factory compression specification by the fifth stroke. If the compression reading is low on the first stroke and builds with following strokes but doesn't reach minimum compression, the piston rings are probably not sealing well. If the gauge reading remains low on all five strokes, it usually indicates a leaking valve. Low compression in two adjacent cylinders indicates a blown head gasket.

Fig. 3. A leakage tester can pinpoint which cylinder is bad and where the leakage is.

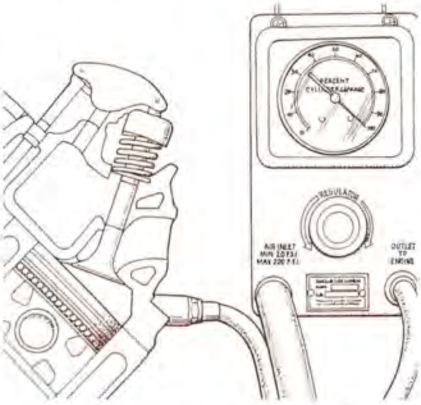

If all engine cylinders achieve proper compression, then you're done. If any of the cylinders had low compression, however, you'll have to confirm the cause by doing a "wet compression test." Squirting oil into combustion chambers with low readings temporarily improves sealing of the piston rings. If compression goes up 10% or more when you repeat the compression test, then the rings are bad. If compression stays the same, then the compression leakage is most likely from the valves.

Keep in mind that a wet test won't work on horizontally opposed Volkswagen, Porsche and Subaru engines.

Leakage Testing

Long favored by professional racing mechanics as the most accurate method for testing an engine's combustion-chamber sealing, a cylinder leakage tester works by pressurizing the combustion chamber with compressed air and then measuring total leakage as a percentage.

A leakage tester, which costs about $75, will not only tell you the condition of the rings and valves, it can tell you if the leak is coming from an intake or exhaust valve **(Fig. 3)**.

To do a leakage test, warm the

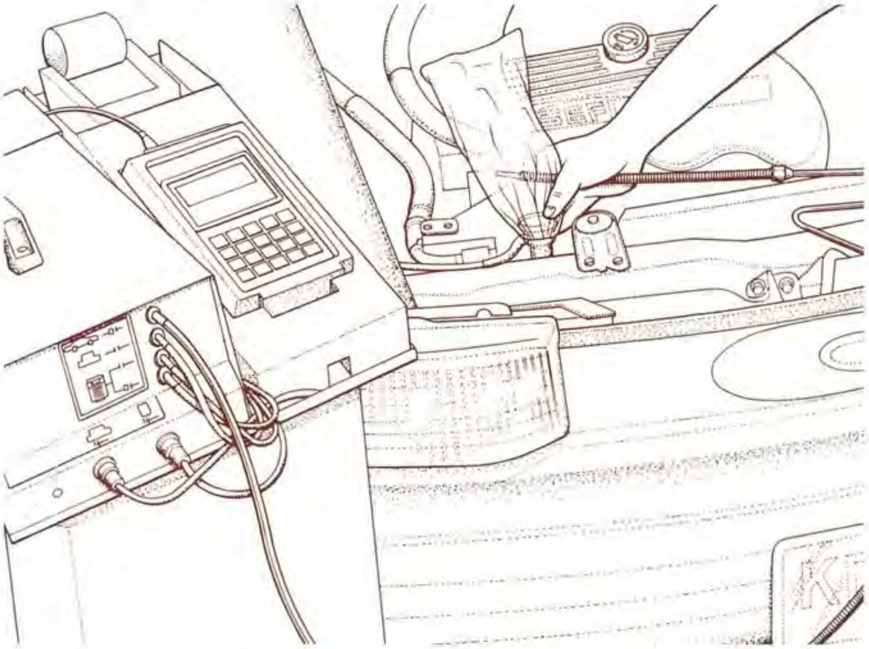

Fig. 4. To look for combustion leakage into the cooling system, hold a bag over the radiator filler. The emissions probe should show no CO_2 when the engine is running.

engine and remove the spark plugs. Remove the air cleaner and block the throttle wide open. Rotate the engine so the No. 1 cylinder is at the exact top dead center (TDC) on the compression stroke. Install the fitting from the leakage tester in the No. 1 spark plug hole, then connect an air-compressor line to the tester. Pressurize the cylinder and note the percentage of leakage on the gauge.

AN ENGINE WITH GOOD compression will show readings on the gauge within a 20-pound spread or less.

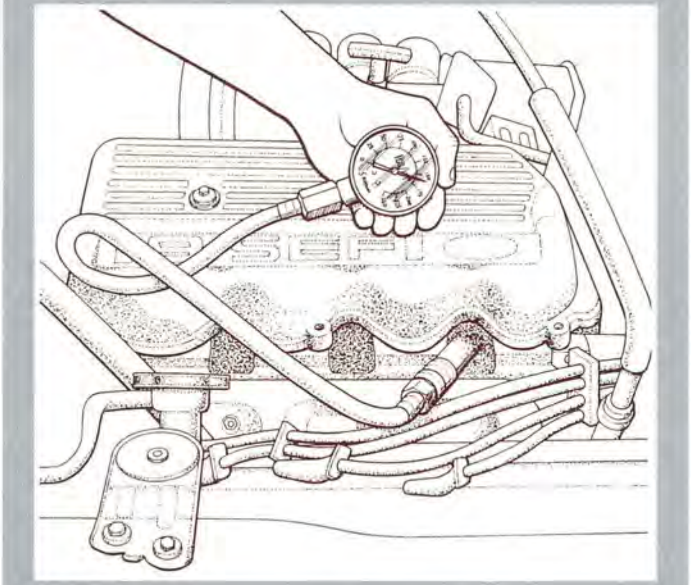

An engine in great shape will have cylinders with 5% to 10% leakage. An engine with 20% leakage is still in pretty good condition. If the engine has a cylinder with more than 30% leakage, however, you should isolate the cause.

You can do that by listening for escaping air while the combustion chamber is pressurized. If the sound comes from the intake manifold, the problem is a leaking intake valve. From the exhaust manifold, it's a leaking exhaust valve. If you hear air escaping from the oil/filler cap opening, the problem is with piston sealing. You can even look for bubbles in the engine coolant at the radiator cap to check for a blown head gasket. And, like the compression test, high leakage readings between cylinders equals a head gasket leak.

It's also possible to find head gasket leaks at the local emissions inspection station. Warm the engine and ask the mechanic to place the emission tester's exhaust probe in a plastic bag. By placing the plastic bag over the radiator opening while the engine is running, the probe will be able to pick up any exhaust gases that are leaking from the combustion chamber into the cooling system **(Fig. 4)**.

Plugging Leaks

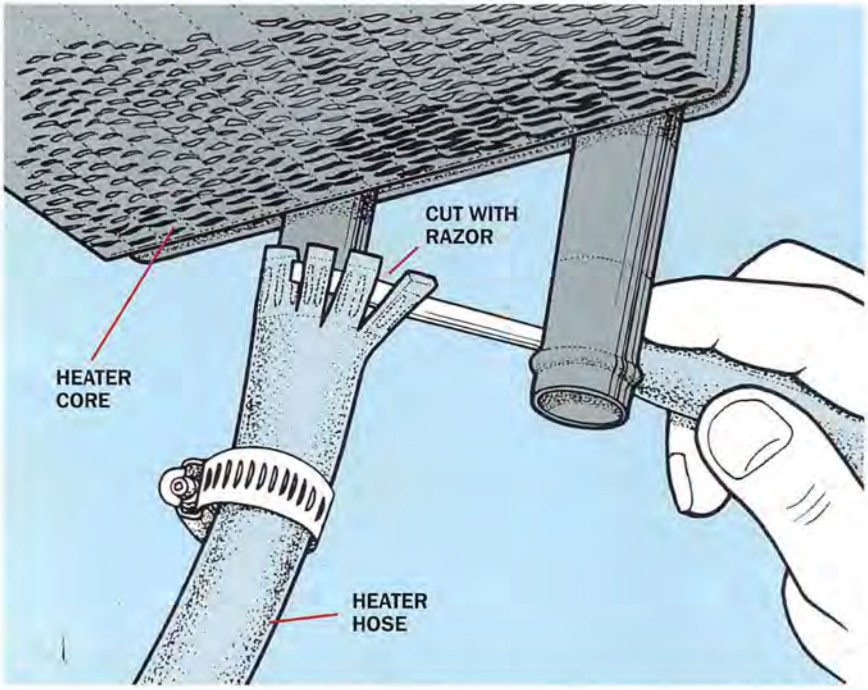

Fig. 1. Remove a stubborn hose by carefully slicing it. Don't nick the neck it fits over.

The sweat pours down your brow as you lift the hood while gasoline gurgles into your tank in the self-serve lane. Through the shimmer of heat that escapes from underhood you lean over to check the level of bright green coolant in the translucent overflow tank. Empty, again.

You've topped it up a few times—but the last few tankfuls it's been dry inside. In fact, coolant seems to be disappearing at an accelerating rate. The weather is getting hotter, and you're about to take a vacation. It's time for some action.

It should be easy to prevent most major coolant leaks since they usually occur when a hose fails. And you should be able to spot the bad hose before it bursts by those telltale signs of heat cracking and other deterioration, right? Wrong. While coolant hoses near the end of their useful lives used to show the classic warning signs, they no longer necessarily do. Today's hoses are so much better you can't tell only by looking.

You should also be able to easily spot a coolant leak, even a small one. After all, the antifreeze contains a green dye that leaves a stain. Well, the truth of the situation is that your current model engine may have more than a dozen hose connections—and other joints sealed by gaskets and O-rings—and many of them are buried. To top off the problem, there are ways that coolant can escape from the system without leaving so much as a trace.

Even when you find the leak, it may take some special steps to fix it. Result: Finding, fixing and preventing future coolant leaks involves a lot more than it used to. But if you take the time to do the job right, you'll be able to avoid the repeated need to top up the system, and the disaster that would occur if you forget.

Fixing Obvious Leaks

If a leak is obvious, such as a leaking radiator, cracked hose or badly leaking joint, that's great—a solid starting point. If it's an O-ring type at a coolant pipe, clean the groove if necessary with ultrafine sandpaper or steel wool. Wipe clean, and then replace the O-ring, lubricating it with antifreeze. Leaky radiator? Check with a radiator shop, and you may get away with an inexpensive repair.

It might be possible to resolder a split seam or a leaking joint at the neck/tank junction on an old copper-brass radiator without even removing the radiator from the car. Even leaky plastic-tank radiators can sometimes be repaired by replacing the O-ring between the end tank and the aluminum core. This is definitely a task to leave to a professional, however. It's a hose? If the clamp is not tight, and the leak is from the hose end, perhaps the hose has taken a compression set under the clamp. Try a simple retightening—but don't overdo it, or you'll distort the hose and neck, and just end up with a worse leak than you had.

Tightening the clamp doesn't help? Replace the hose. Uh-oh, it's stuck. Grab it with pliers and twist? Not unless you want to crush the hose neck and make it impossible to seal properly. Better to cut it off instead.

With a razor blade, slit the leaking end of the hose at several points around the circumference, gently pry up each slice, then peel the rest of the hose end. Don't nick the neck on the radiator or engine, or it'll never seal again. Clean the neck—on the radiator, engine, heater core, connector pipe—with a wire brush and wipe clean. Then install the replacement hose, using fresh antifreeze as a hose lubricant.

Radiator necks are generally pretty sturdy, but heater-hose pipes are usually very fragile. Twisting and prying

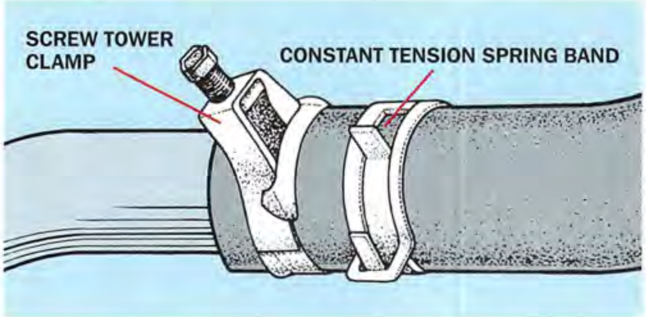

Fig. 2. OEM screw-tower clamps don't seal nearly as well as spring clamps. A regular worm-drive clamp provides a tight seal, too.

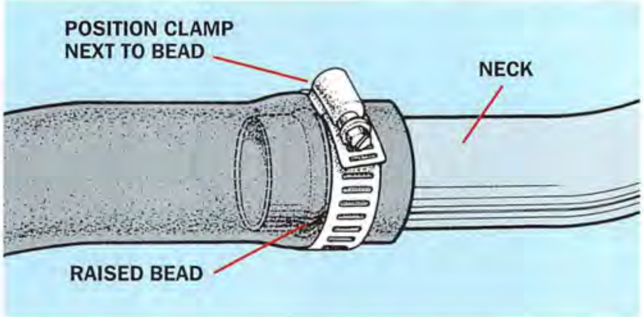

Fig. 3. Don't give coolant a place to collect between the clamp and the sealing bead, or you'll give corrosion a place to start on copper-brass radiators.

to remove a hose—or replace it later—can easily break the seal between the pipe and the heater core's tank. And removing a heater core to repair or replace it can involve removing half of the underside of the dashboard—an operation that can consume a dozen hours or more. Use caution.

Unless there's lots of slack in the hose, don't try to trim the cut end and reuse it. What's left probably will kink and ultimately restrict coolant flow.

Use a new clamp and position it correctly for a leak-free connection, being extra careful if there is any damage or distotion to the hose neck. The two best types of clamps are the constant-tension spring band and the 360° adjustable double-band. A decent-quality, worm-drive tangential-force clamp, with a slotted band that won't dig deep into the hose rubber or cut it when you tighten, also does a good job. The screw-tower clamp, widely used as original equipment, is tolerable for a brand-new system, but is about the least effective for resealing.

Install the hose as far as possible on the neck, and if there's a stop, right up to it. Position the clamp carefully so that it's about midway between the end of the hose and the end of the neck, typically about one-half inch or more from the end of the hose. If there's a raised sealing bead on the neck, position the clamp flush against the back side of it. If it's any distance away, it won't seal as well and it will force a crevice between the hose and neck. Coolant will collect and stagnate in the crevice, accelerating corrosion (and erosion) of the neck on copper-brass radiators, eventually resulting in a significant leak. If a hose neck is badly distorted or eroded, get it reshaped or repaired by a radiator shop if you expect it to seal.

Finding Hidden Leaks

Sure, you have to look hard to inspect all the hose connections, O-rings and gasket joints, but you can get some help if you first pressurize the system with a cooling-system pressure tester. Pump up the system to its rated pressure plus a couple of psi. If the gauge reading holds for 2 minutes or so, the cooling system isn't leaking—at the moment. However, you may still be losing coolant.

Does the gauge's reading start dropping almost immediately? That tells you the system isn't holding pressure, and if you start looking you should see leakage from the offending joint or possibly the water pump's weep hole. All water pumps seep a minor amount of coolant (there's no such thing as a perfect seal), but if you see steady dripping or more, the pump seal is bad. Replace the pump.

Okay, the pressure gauge reading hardly budged, and you're still losing coolant. Now the diagnosis gets a bit tougher, because there are these possibilities

- Although cold engine leaks are more common, it's also possible the leakage only occurs with a hot system. Run the engine to warm up the system and retest.
- There is minor seepage from many joints and it all adds up. If your engine's cooling system has a lot of connections, inspect them all for even the slightest trace of coolant staining. Simple retightening of all the hose clamps may do the job.
- The leakage is from the

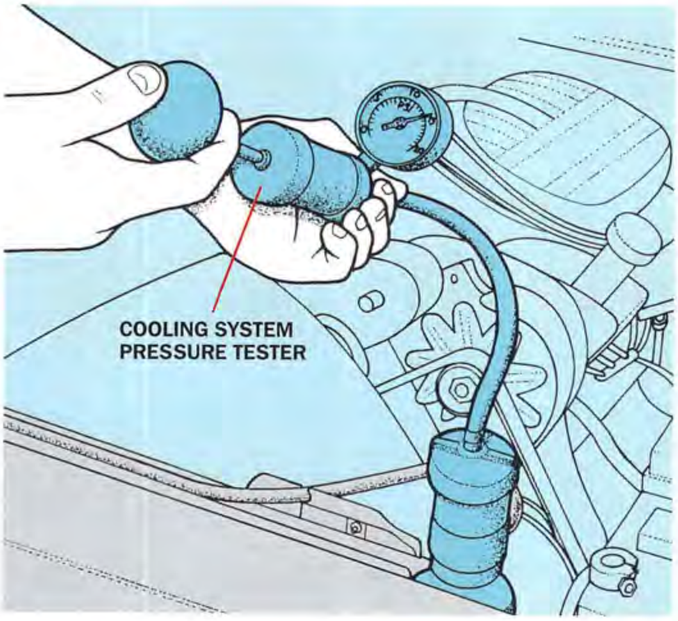

Fig. 4. Pump your system up to a couple of pounds higher than the rating on the radiator cap to check for leaks. You can rent the tool for an afternoon.

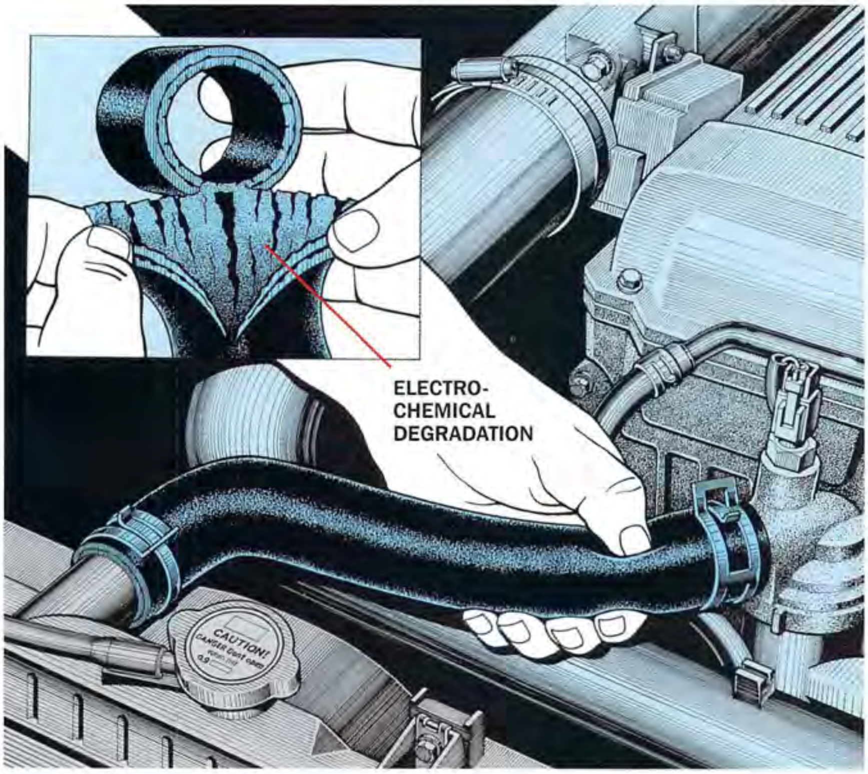

Fig. 5. Check your hoses for electrochemical degradation by pinching and rolling the hose between your fingers and thumb. Feel for cracks in the inner liner.

overflow reservoir, which on many cars is at atmospheric pressure and isolated from the pressurized part of the cooling system. Hairline cracks in reservoirs often result in leaks that travel along body seams and are detectable only if you are really looking for them.

- The engine periodically runs so hot that coolant fills the atmospheric overflow reservoir and overflows it. Or the pressure cap valve is weak, so that coolant flows from the system into the reservoir even under normal operating temperatures. Check the coolant level in the reservoir immediately after getting stuck in heavy traffic. If coolant is ready to pour out of the reservoir vent, pressure-test the cap to see if it's holding the specified pressure.

On systems where the reservoir is pressurized, inspect the radiator-cap valve seal for deterioration. A bad seal can allow coolant to be lost by evaporation.

If the cap is good, next check the antifreeze concentration with a hydrometer. If the percentage of antifreeze is low (below 50%), the coolant will boil at a lower-than-normal temperature and contribute to the problem. If the problem is the coolant running too hot, a poor cooling system condition may be responsible. This would include low coolant level, an electric fan coming on too late (or a mechanical clutch fan slipping), a defective water pump, a plugged radiator or even something as simple as a bug-and-road film coating on the front of the condenser or radiator. Such a coating blocks the cooling airflow.

Preventing Future Leaks

Today's hoses fail more often from the inside, because of a problem called electrochemical degradation, the catchy ECD for short. The metals of the cooling system and the hose material, combined with the coolant, form an electrical cell that produces a low DC voltage, particularly at high coolant temperatures. Over a period of time, the electrical flow is enough to cause fine cracks to develop inside the hose, starting at both hose ends and working along the length of the hose. Eventually, the hose develops a pinhole leak or may even burst.

You can't see this damage, but you can feel it. Wrap your thumb and two or three fingers (not your whole hand) around the hose at the end and feel with your fingertips for soft "channels." Run your fingers around the hose to develop a mental picture of what the inside circumference of the hose must look like. Work your way from each hose end toward the center of the hose, and if the hose feels more solid as you approach the center, it's probably suffering from ECD. According to studies by Gates Rubber Co., most upper radiator hoses, for example, are ready to leak (if they haven't already) after about four years of service.

So if any of your cooling system hoses fails this fingertips test, drain the system, cut the hose and peel it off. And install new premium hoses designed to resist ECD. You can get silicone hoses, but newer designs of another synthetic, EPDM, are less prone to tearing and are designed to resist ECD. Secure them with good-quality clamps and your engine should keep its cool.

If you're topping up your system after a leak, all you really need to do is add a 50/50 mixture of coolant and water to the radiator. But if you've been topping up with water, you have no real idea of the relative concentration of the two fluids. Perhaps you've been using water from some roadside ditch or rain puddle—a better proposition than walking or melting an engine for sure, but one that requires you to drain the cooling system completely and flush it.

If you've flushed the system for whatever reason, don't just start pouring 50/50 coolant into the radiator. There still will be a lot of water in the system that won't drain from the radiator (see page 101) and they'll leave you with a low concentration of glycol. Instead, look up the capacity of your cooling system in the owner's manual or the shop manual. Add half that amount of undiluted coolant, and top it all off with pure water.

You did collect the drained coolant for proper recycling or disposal, didn't you?

Turning Up Your Heater

It's colder than a witch's belt buckle outside, so you gulp down some hot coffee, dash to the car, start the engine and run inside. Although you know engine idling isn't a great way to warm up an engine, comfort counts most on a cold morning. Once embarked on your morning commute, you turn the heater control up to Hot and flip the blower switch. The effect of the hot coffee has worn off, but you're not getting enough heat from the floor registers to make up for it. In fact, the windows start to frost over.

Your heater system is under-performing. You plan to tear into it as soon as you can get the car into a heated garage.

First step: Confirm the problem by driving long enough to warm up the engine. Turn the heater control to Hot, select the floor mode and set the blower to medium/high speed. Make sure there's strong airflow from the floor registers. Check the temperature with a thermometer inserted into the driver's and passenger's registers. Although many vehicles deliver a toasty 135° to 155°, 115° to 120° is acceptable. If the air temperature is okay but airflow is weak, the problem is the blower or a dashboard ductwork problem.

Heater register temperatures not even close to 115° to 120°? Check factory service bulletins to see if there's a known problem (at the vehicle maker's technical service Web site or at www.alldatadiy.com). If nothing applies, do an underhood inspection.

- The coolant level in the reservoir (and radiator if it has the pressure cap) should be at specs. The coolant should look clean and have been regularly changed within the specified interval. If it's been neglected, all bets are off.
- Cabin heat comes from hot coolant, so drive long enough to warm up the engine. The coolant temperature should hit at least 160°, preferably 180° to 220°.
- The air conditioner compressor should be disengaged with the climate control selector out of the defrost position. If it's engaged, there's an a/c circuit problem.
- If coolant temperature is acceptable, feel both heater hoses, which should be hot. They are? Check the temperature-control flap-door operation in the under-dash heater-a/c case.

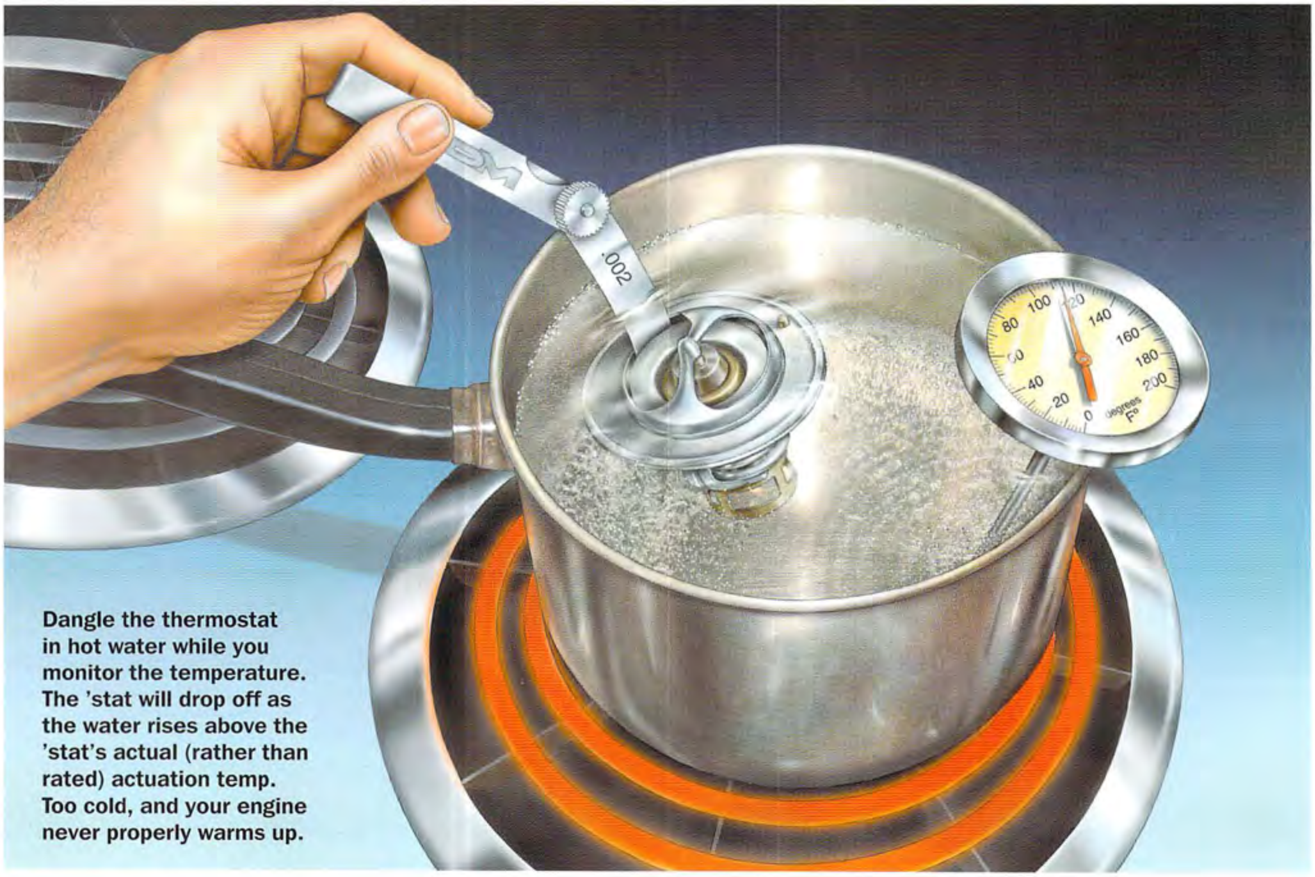

Dangle the thermostat in hot water while you monitor the temperature. The 'stat will drop off as the water rises above the 'stat's actual (rather than rated) actuation temp. Too cold, and your engine never properly warms up.

If you have automatic temperature control, there may be built-in diagnostics you can access for trouble codes covering the temperature control door. (Refer to a service manual, factory tech service Web site or www.alldatadiy.com.) On many cars you can watch the temperature-door motor and shaft operation by removing an under-dash panel and/or lowering the glovebox. (Here again, you'll have to check a manual or Web site for instructions.) If you don't see and/or feel the motor or flap-door shaft operate as you change the temperature reading, you've pinpointed the problem. In many cases, the flap-door shaft breaks and the assembly must be replaced, but the motor is good and can be reused.

On econocars with manual a/c, the temperature flap may be controlled by a cable from the dashboard lever or knob. Operate it quickly in each direction and listen for a soft slam against a stop. If you feel binding, the cable is kinked or the flap is jammed. Total free play? The cable is loose or disconnected, or the flap is broken.

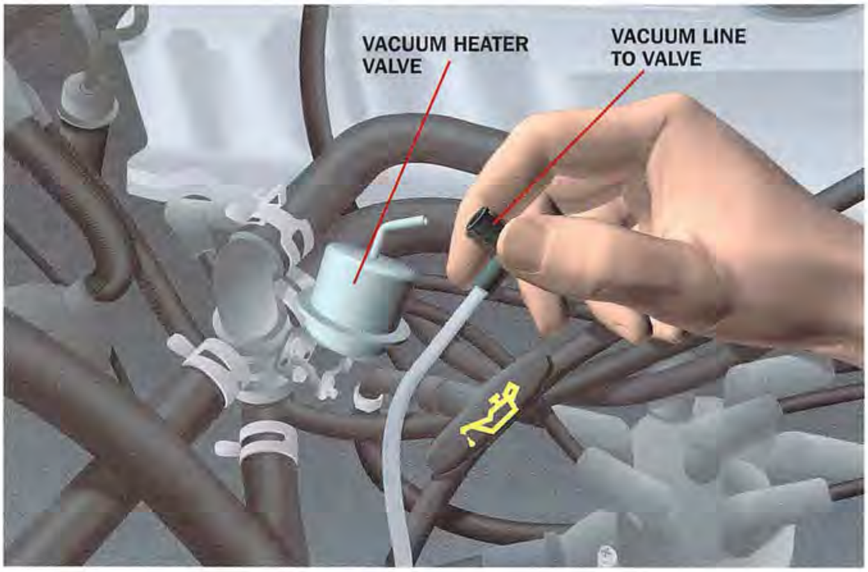

Check for good vacuum supply to the heater's control valve. If it's good, check the valve itself for proper operation.

Remove or open any air bleed plugs on the block or heads while refilling the engine to purge any air in the system.

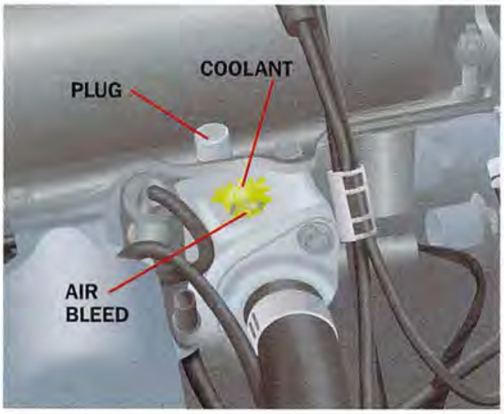

Simply blocking the radiator can make your car warm up faster in extremely cold weather.

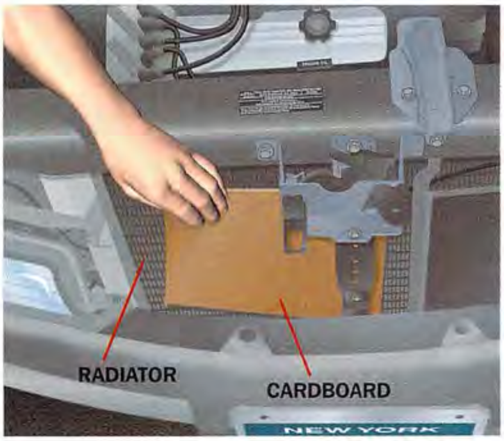

Nothing's obviously wrong? There may be air pockets in the cooling system—these commonly develop over time. If there's a pressure cap on the radiator or engine and a low coolant level in the radiator, but the level in the reservoir is normal, a defective gasket in the cap is admitting air. Replace the cap.

Top up the system and purge air, following the vehicle maker's specific procedure.

As a generic approach, start with the engine cold, all bleed valves open, and the side of the vehicle with the coolant neck jacked up as high as possible. This raises the fill neck to the high point, so air bubbles out. Slowly add a 50/50 antifreeze/ water mixture to the fill neck. You'll see some coolant (and perhaps air bubbles) oozing out of the bleeders. Close the bleeders, refit the pressure cap, and run the engine at about 2500 rpm until it's warmed up (or as warm as it will get). Allow the engine to cool down, then repeat the top-up procedure. You may have to repeat this every few days over a week to get most of the air out.

Coolant temperature too low? On vehicles with a clutch-type fan on the engine, make sure the clutch is not locked up. Spin it by hand (with the engine off but warmed up). If there's virtually no slippage, the clutch is defective. If it spins two to three times, it's probably okay.

Could someone have removed the thermostat, probably trying to remedy an overheating situation? That's a very bad idea, not to mention the fact that it is illegal to tamper with emissions controls.

No fan operation problem and coolant temperature is very low? If you've neglected your system, the thermostat might be stuck open from debris. If it's really buried, pour in a cooling system cleaner and run the car per the instructions. Then, do a triple drain-and-refill of the radiator and engine block, and go through the air-bleeding top-up process. If coolant temperature remains low, bite the bullet and remove the thermostat for a test.

If the access to the 'stat isn't bad, hold off on the drain-and-refill, and

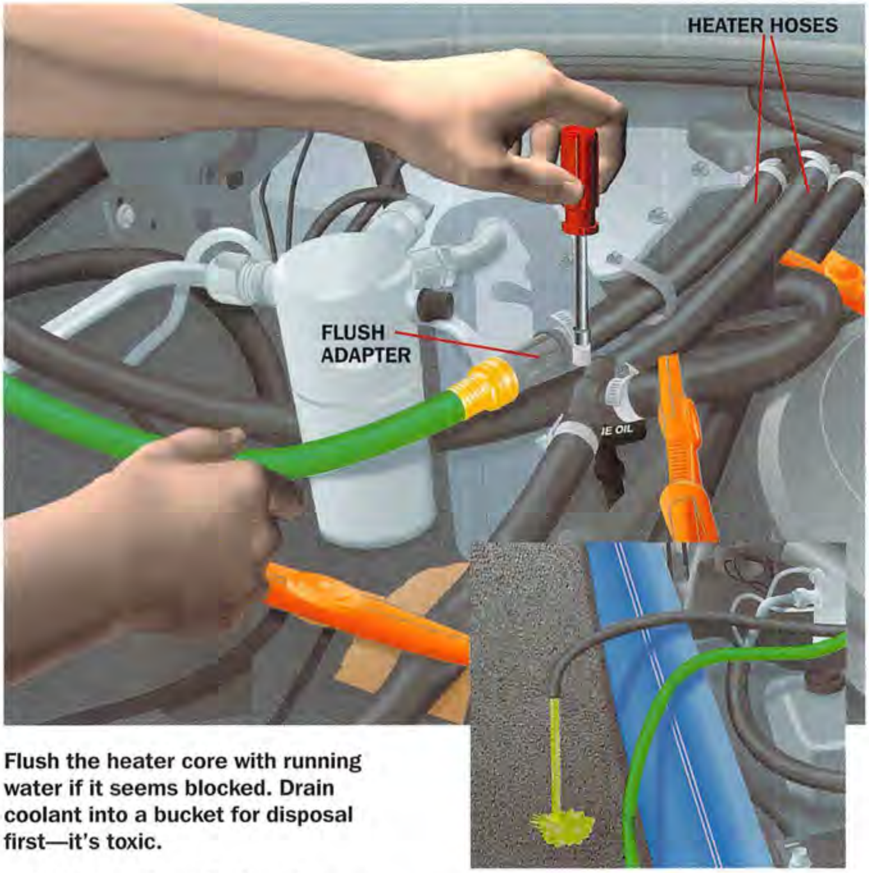

Flush the heater core with running water if it seems blocked. Drain coolant into a bucket for disposal first—it's toxic.

pull the 'stat for a look and a test. Clean if necessary, but before putting it back refit the 'stat housing, then run a cleaner through the system. Then do the triple drain-and-refill, and test the thermostat before you decide to reinstall.

Testing: Wedge a 0.002-in. feeler between the closed 'stat valve and its seat, then submerge the 'stat in a pot of water on a stove. Rest a radiator thermometer in the water, heat the water to boiling, and note the temperature at which the thermostat cracks open—it will drop off the feeler at its rated temperature.

Some engines, particularly transverse four-cylinders, radiate so much heat so fast that there's no internal fix for extremely cold weather. It may not be an elegant solution, but put a piece of cardboard against the front of the radiator, blocking about half of the surface area. Trim it back if the engine starts running hot. Remember to remove the cardboard in the spring.

Coolant temperature reading is normal, but one or both hoses are not hot? Refer to a service manual or Web site to identify the heater inlet versus outlet hose.

Neither hose is warm? Many cars have a heater coolant shutoff valve spliced into the heater inlet valve for better a/c performance in summer. If the shutoff valve fails to open fully, the flow to the heater will be restricted. Vacuum operated? Disconnect the hose. If you feel vacuum (with some valves you'll also see the linkage move), the valve is being held open by a defective vacuum control. Leave the hose disconnected (plug the end with a golf tee) and see if that cures the problem. If it doesn't, the shutoff valve is stuck. Replace it. If the valve is cable controlled, you should see the valve shaft move continuously as the dashboard temperature lever is operated. If it doesn't or if it moves just slightly, the cable needs adjustment. Note: The VW Golf and Jetta have a cable-type valve, but it opens proportionally to the flow of hot coolant as needed—there are no flaps in the dash.

On some other European cars and on the Lincoln LS, the valve works similarly to the VW's, but it's electronically controlled. Quick test: Unplugging the electrically controlled type should force the system to full heat.

If the heater inlet hose is noticeably warmer than the outlet, the heater core is restricted from debris in old coolant or from excessive use of sealer. First, try to reverse-flush the heater core. Do this by either disconnecting both heater hoses from the heater necks on the cowl and attaching spare hoses, or by disconnecting both heater hoses at the engine. (Clamp off the heater hoses if you're using spare hoses; plug the engine hose necks if you're disconnecting there.) Force a garden hose into the heater outlet hose, aim the inlet hose away from the vehicle, and open the bibcock. If the debris flows out and you eventually see a healthy flow of coolant, you may have caught a break. Otherwise, you've got to replace the heater core.

If heater performance during highway operation is all right but it really drops off in low-speed driving, there's a cooling system flow problem. Some systems have a hose restrictor in the inlet hose, and if it gets plugged with debris, that lowers the flow. You should find this by feeling along the inlet hose with the engine idling. (You'll reach a section where the temperature significantly drops.) A few engines (on the Dodge Durango, Lincoln LS and a number of German cars) have auxiliary water pumps to improve coolant flow at low speed/idle. If the pump fails, heater output drops during that operation. Of course, if the cooling system was long neglected, it's possible the engine water pump has eroded internally, reducing coolant flow.

In most cases, poor heater operation can be cured by following the factory top-up procedure with the air bleeds open. Do this several times over the course of a week. Is there a faster fix? A pro with a coolant changer that purges air as it fills the system can give you faster heat.

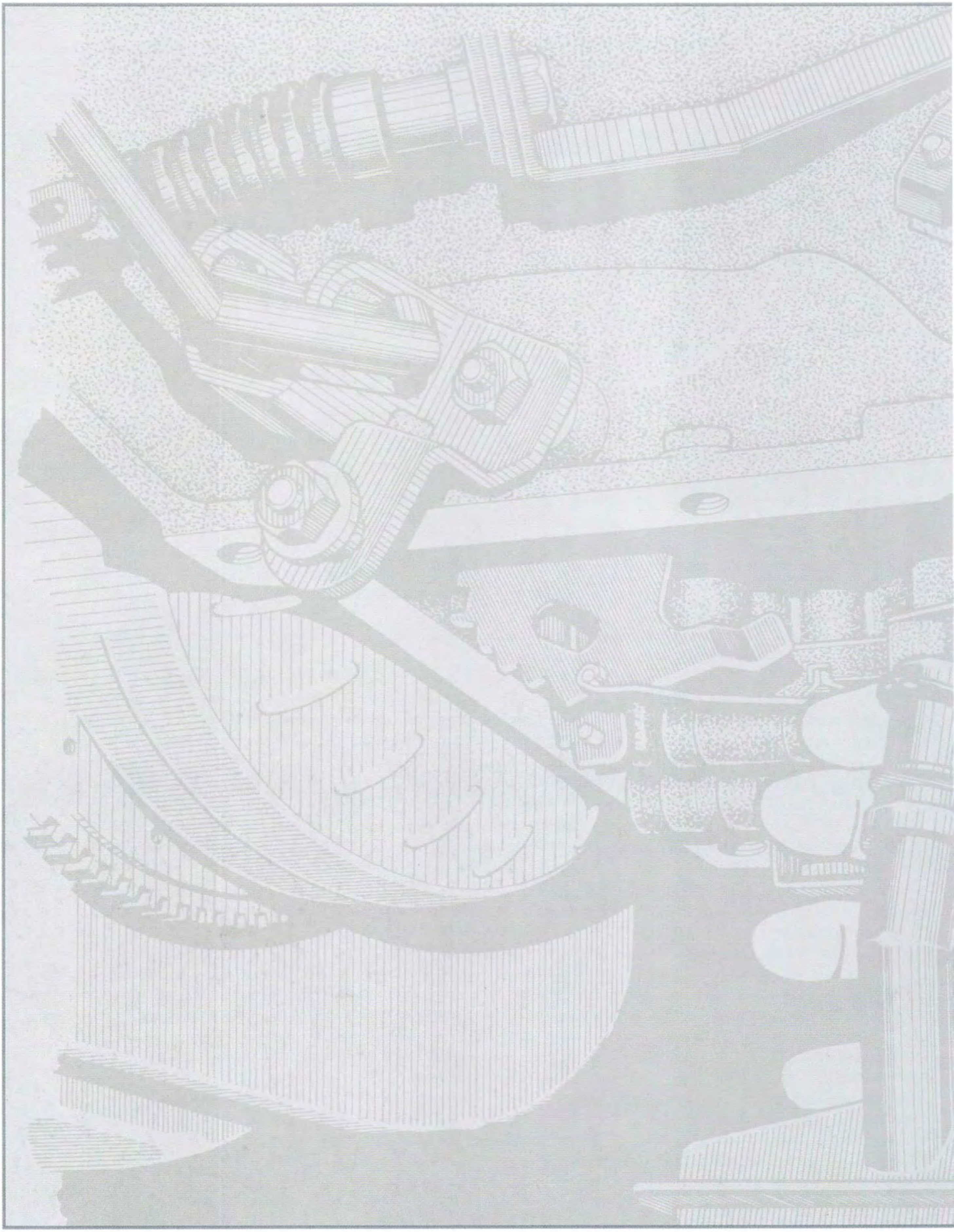

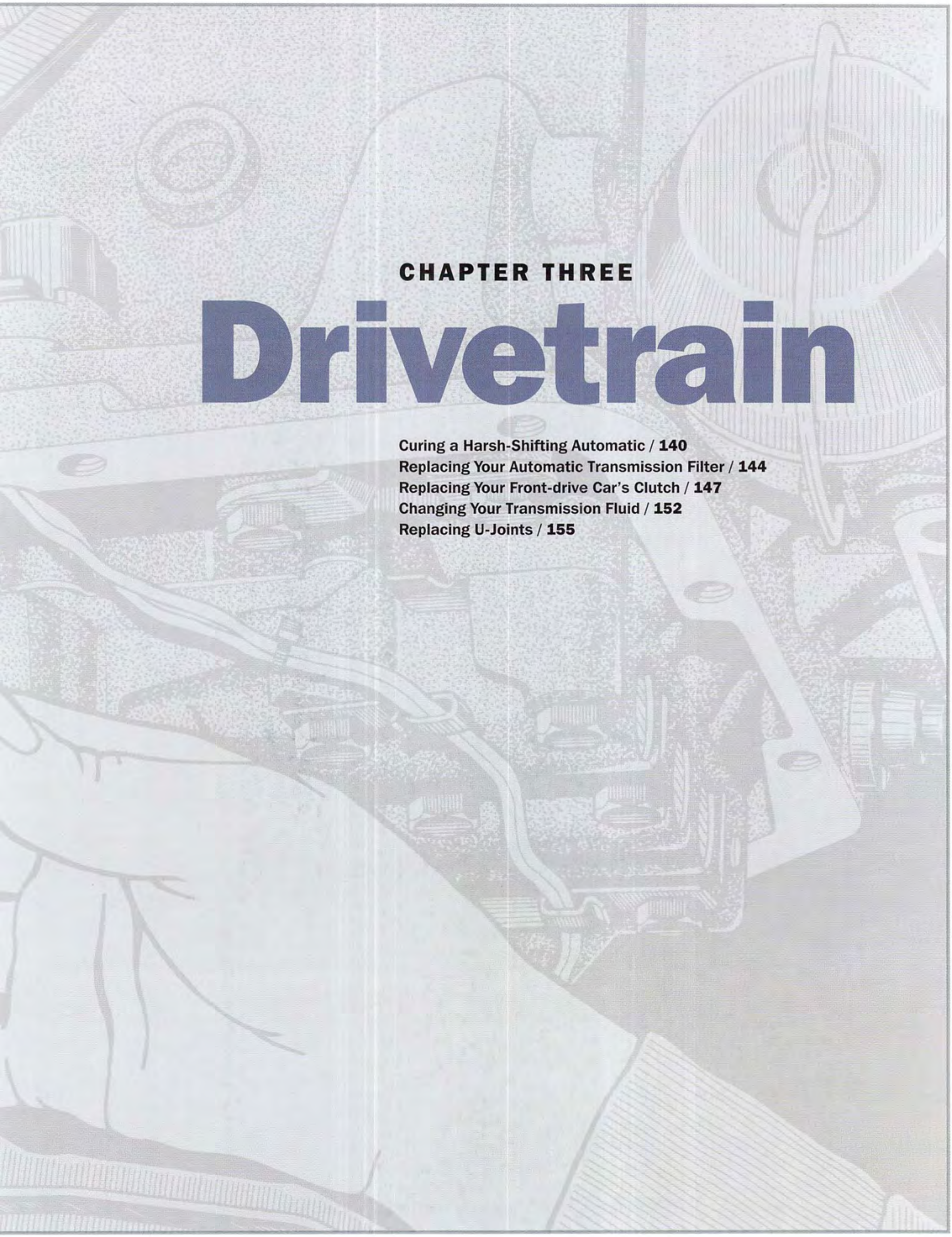

CHAPTER THREE

Drivetrain

Curing a Harsh-Shifting Automatic / **140**
Replacing Your Automatic Transmission Filter / **144**
Replacing Your Front-drive Car's Clutch / **147**
Changing Your Transmission Fluid / **152**
Replacing U-Joints / **155**

MODERATE

Curing a Harsh-Shifting Automatic

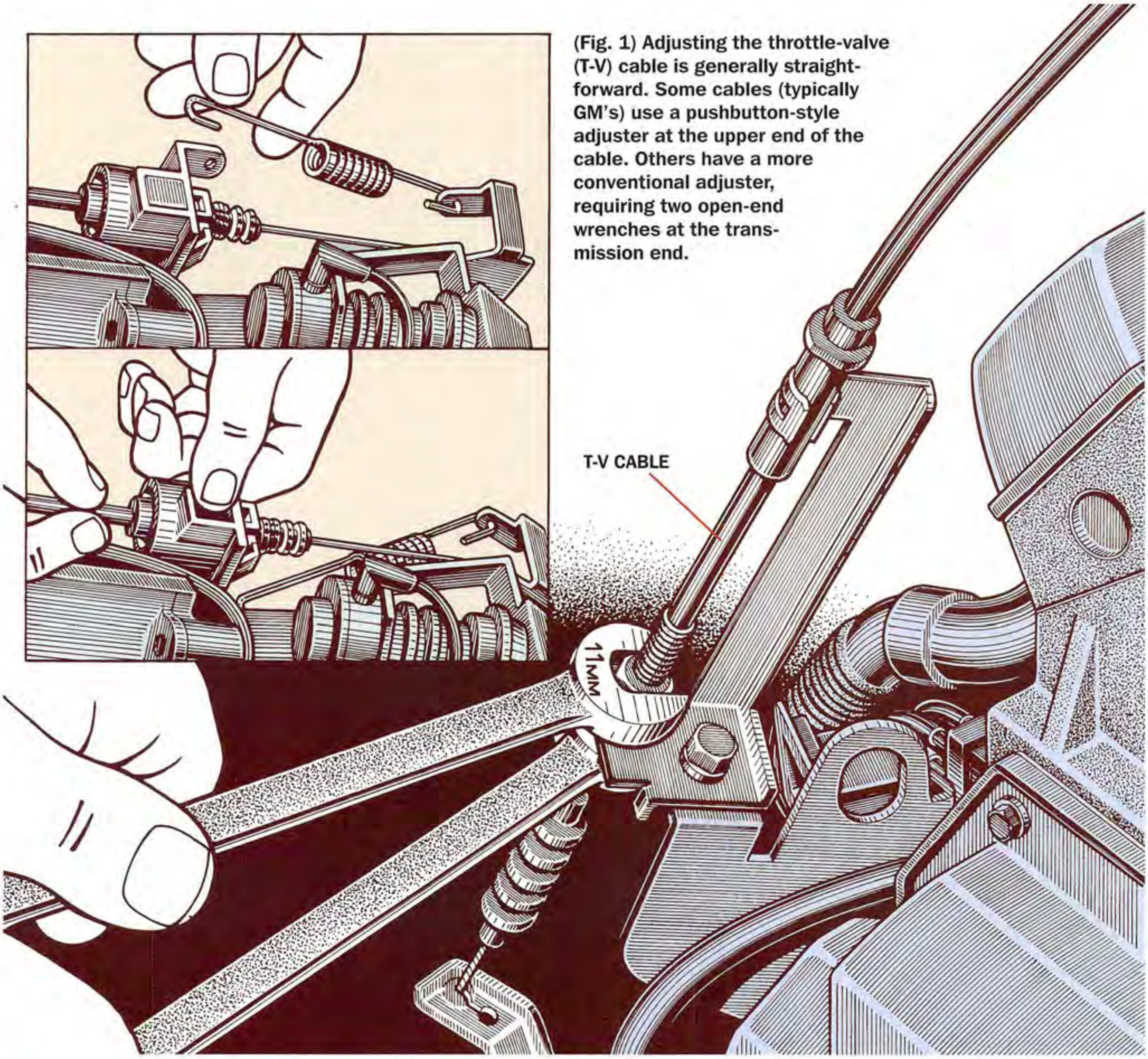

(Fig. 1) Adjusting the throttle-valve (T-V) cable is generally straightforward. Some cables (typically GM's) use a pushbutton-style adjuster at the upper end of the cable. Others have a more conventional adjuster, requiring two open-end wrenches at the transmission end.

There's coffee all over your shirt, the dog refuses to ride in the car and your teenagers have threatened you with a class-action whiplash lawsuit. The fact is, when you were a teenager, you would have paid plenty for a car that got rubber in Second—but when it happens leaving every traffic light, it stops being cool.

An automatic that clunks into gear and suffers from irritating vibrations in the process is not something you have to live with. Nor is it a problem you have to drop a few weeks' pay to correct. Most of the causes are things you can fix or prevent, and although you'll need a good service manual for adjustment procedures, specs and parts locations, you won't need a collection of special tools or instruments.

There are two basic types of automatics: the fully hydraulic and electronic hydraulic. In a fully hydraulic automatic, found primarily on older vehicles, all the shifting is controlled by piston-type valves that slide in a valve body and direct the pressurized fluid to the appropriate bands and clutches. The road-speed signal comes from a centrifugal hydraulic device called a governor and the engine load (acceleration) signal comes from a cable or linkage between the trans-

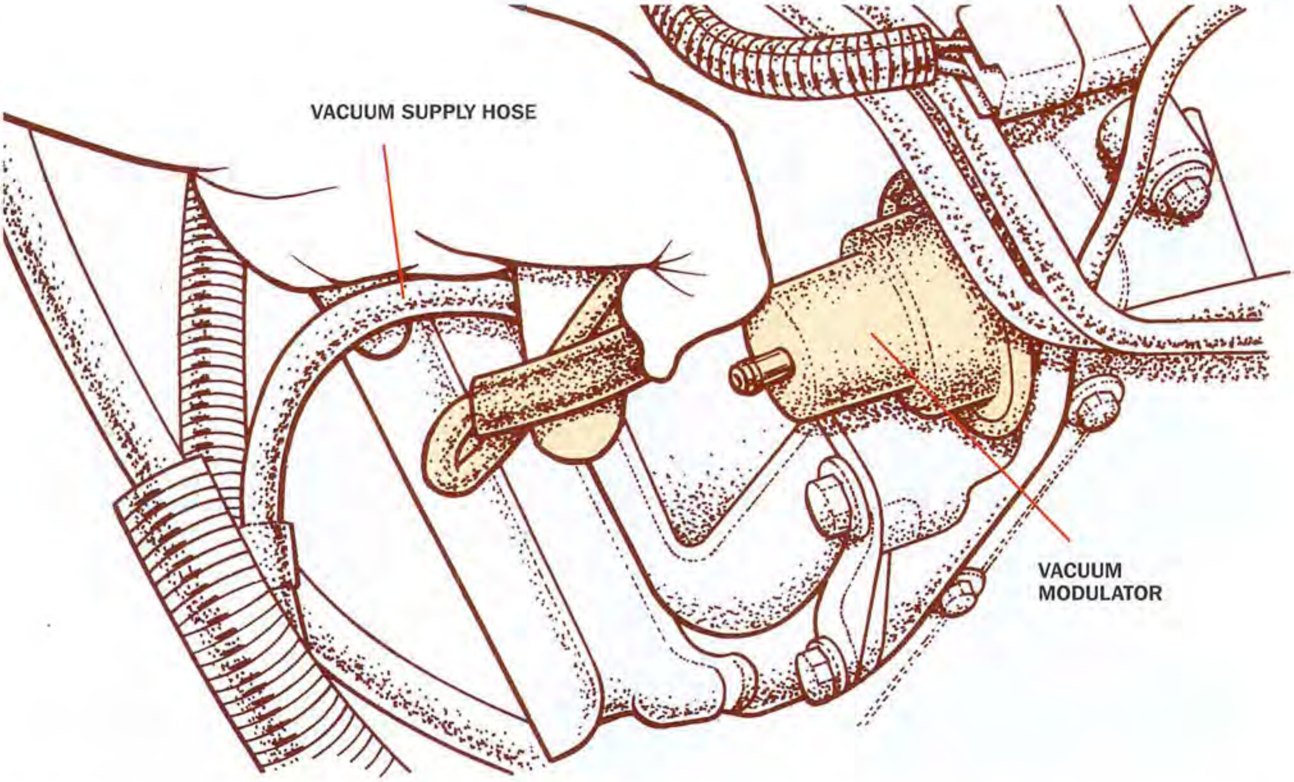

(Fig. 2) The vacuum line at the modulator should exhibit healthy suction when disconnected and the engine is at idle. Look for ATF inside the modulator.

mission and the throttle body. Shift quality and engine load also may be signaled by a vacuum modulator, a diaphragm unit connected by a hose to engine vacuum, and by a link to a piston-type valve in the valve body.

The electronic hydraulic may have a hydraulic valve body with electronically controlled solenoids that provide signals for directing fluid flow. Or it may have a set of solenoids that directly control fluid flow. The road-speed signal comes from a vehicle-speed sensor and the load signal from the throttle body's throttle-position sensor.

Garage Shifts

If the Park or Neutral to Drive shift—called the "garage shift"—is rough, but other shifts are normal and the idle speed is high, then high idle is very likely to be the reason. There are many possible causes of high idle speed on cars with computer-controlled idle, including vacuum leaks, a maladjusted throttle-position sensor and leaking O-rings around fuel injectors.

Idle speed normal? If you'd been pulling a trailer, particularly in hot weather, the automatic transmission fluid may be oxidized. This is a common cause of harsh shifts, particularly with the garage shift. You won't always get a bad smell from oxidized fluid, but if you've done the kind of high-load operation that could overheat it, a fluid and filter change might help.

All Shifts

If all the shifts are harsh, the mileage is high, the fluid has never been changed and the service has been severe (such as hot-weather trailer towing), it's possible that all the control hydraulics are sludged or sticky.

You could drop the oil pan for an inspection and if that's what you find, a straightforward fix is to bolt in a replacement valve body. That's still just a moderate investment. However, before you go this far, check out all the other possibilities that can be easily fixed.

For example, when the harsh shifting is in every gear, the hydraulic-pressure feed, or line pressure, is likely to be too high. Also, on some transmissions the design is such that harshness could result from pressure that is too low, so don't forget a fluid-level check with the dipstick.

Note: Some late model automatics may suffer poor shift quality if they were drained and filled with other than a specified fluid. A fluid designed for Ford Type F applications will cause problems in a Ford transmission for which Type H is specified, for example (see page 132).

Before you go further on a car or truck with an electronically controlled automatic, watch the speedometer. If the readings bounce around a lot, perhaps from zero to 2 or 3 mph when the car is stopped and idling, the vehicle-speed sensor apparently is misbehaving and a bad road-speed signal may be what's upsetting the shift quality. Note: The sensor could be misbehaving without the speedometer effect. The best check is with a scan tool.

On vehicles with fully hydraulic transmissions, the hydraulic governors that supply the road-speed signal may cause late or mushy shifts, but are not likely to cause harshness.

Road Load Signals

The transmission's hydraulic pressure is modulated by a throttle valve,

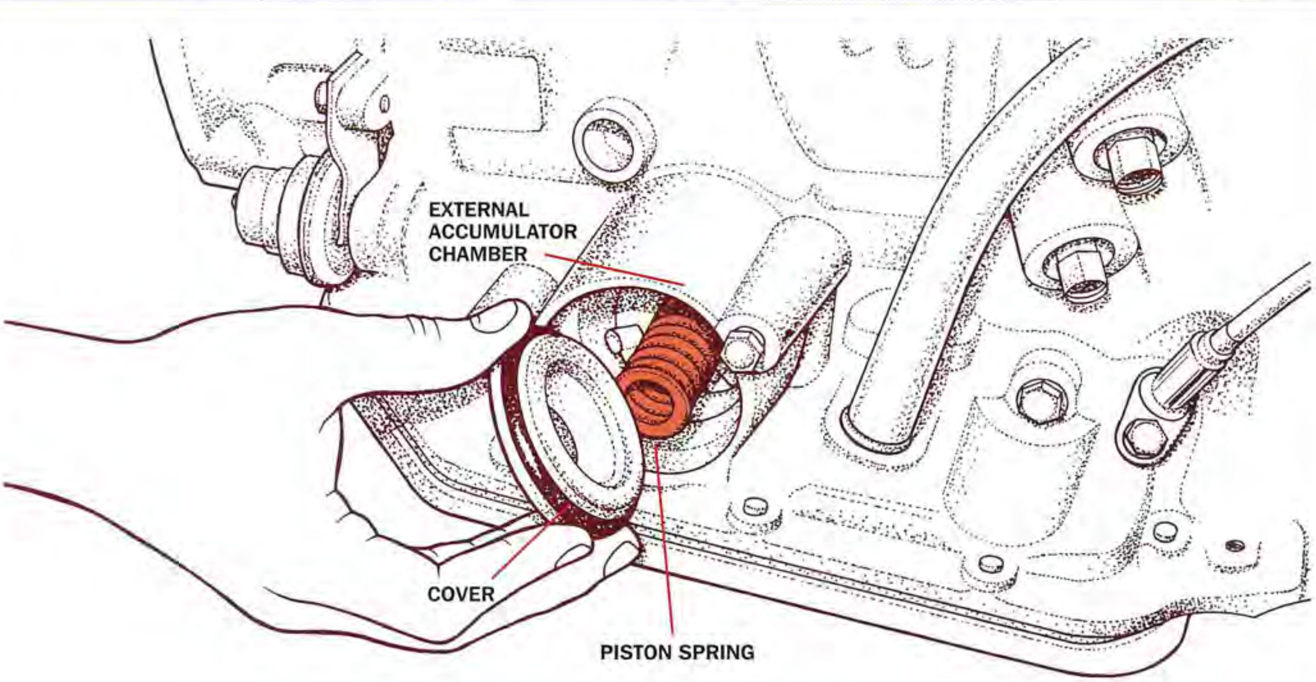

(Fig. 3) Check for sludge or heavy varnish inside the accumulator, which could also indicate sludge in the valve body. Check the spring as well.

which responds to how hard you're accelerating. The valve is in the transmission's valve body, the assembly of hydraulic valves that routes the pressurized oil to make the shifts in most transmissions. And with an automatic transmission that is controlled hydraulically instead of electronically, the valve is connected to external linkage and a cable that goes to the engine's throttle body to provide that throttle-position signal.

However, almost all 1980s and up to mid-1990s cars and trucks use a cable from the transmission to the throttle-body linkage, and that cable has an adjuster at one end. On most General Motors and Ford products, it's at or close to the throttle body. On late-model Chrysler products with the 3-speed (basically hydraulic) transaxle and many imports, including Honda and Toyota, the adjuster is at the transmission end (Fig. 1).

The adjuster has some provision for varying the length of the cable assembly, and it has a spring that takes up backlash. The details of adjustment vary (which is why you need a service manual), but the principle—including inspection—is similar among most makes.

Inspect the spring to see if it's broken or weak (so it doesn't compensate for free play in the cable). Replace a bad spring. Check the cable-retaining brackets to ensure they're not loose or bent, and that they hold the cable housing properly. Look for a bent link at either end of the cable. Next, operate the throttle linkage and watch the cable, which should function smoothly and progressively at both ends. If it doesn't, make the adjustment.

On General Motors vehicles, a typical procedure would be to press the release button on the adjuster and pull the cable slider through the adjuster with the throttle linkage held closed, leaving just a hair of free play so there's no binding.

On a Honda Accord, the usual method begins with running the engine at idle, pulling on the cable housing toward the throttle body to remove all play and checking for some free play between the bracket and the locknut on the throttle-body side (about .040 in., which you can measure with a feeler gauge) (Fig. 2). If there's not enough or too much free play, hold the locknut on the transmission side (the lower locknut) and turn the locknut on the throttle-body side (the upper one).

Things vary somewhat for electronically controlled transmissions. Because the throttle-position sensor provides the upshift/downshift signals, it's the primary item to check. If the throttle-position signal to that throttle-valve assembly is wrong, the shift quality goes down. To see if the sensor is operating within factory specs, perform a scan tool test. Or probe the signal terminal of the sensor with a high-impedance (minimum 10 megohms) digital voltmeter, following the service manual procedure.

Both fully hydraulic and electronically controlled transmissions may use a vacuum modulator. And the modulator is a major factor in shift quality in either case. The modulator takes the engine vacuum signal and uses it to control the shift characteristics, so if vacuum is low (because the engine is out of tune or because there's a vacuum leak), the transmission will behave as if the car is being accelerated and "stiffens" the shifts. With the engine idling, disconnect the vacuum hose and feel the hose end for a strong vacuum (at least 13 in. if you want to check with a gauge). Also test for a leaking diaphragm by inserting a cotton swab into the neck. If there is any transmission oil on the swab, the diaphragm is leaking. Replace the diaphragm unit.

High altitude and low vacuum can affect shift quality, so if you've moved up to the mountains, ask the local

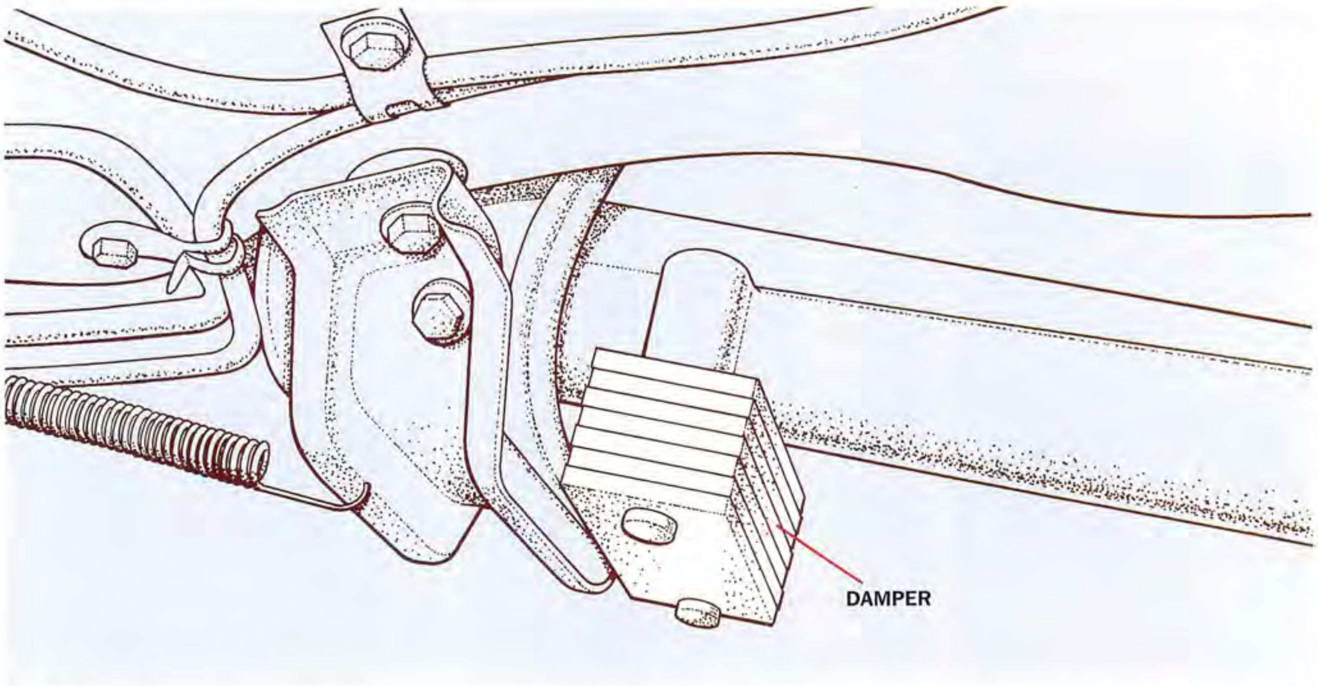

(Fig. 4) Some transmissions use an external damper bobweight at the end of the tailcone to control vibration. If it's missing or loose, you could have poor shift quality.

dealer if there's a special modulator with an altitude compensator.

Only One Shift

Does the harsh shifting occur only in one gear change, such as First to Second or Third to Fourth, for example? The accumulator is something to suspect on fully hydraulic and partly electronic automatics. Most hydraulic clutch/band circuits have an accumulator, which is a spring-loaded piston assembly in a chamber, to prevent shift shock. If the chamber is filled with sludge or the piston spring is broken, it could be responsible (Fig. 3). If the accumulator is external, you can remove the cover and look inside.

Of course, a sludge-filled or varnished accumulator chamber could indicate a varnished valve body under the oil pan. So one harsh shift now could become all harsh shifts soon. However, just cleaning out the one external-access accumulator now could be smooth-shift news for a long time.

If the accumulators are on the valve body, the job is more difficult because you'll have to remove the oil pan. But at least you get a look at the valve body too. Refer to a service manual to see what's involved in an inspection.

Shudder and/or Vibration

The automatic transmission lockup clutch, used on every modern automatic to help the carmakers meet fuel-economy standards, is responsible for all sorts of shudders, particularly if it engages at the wrong time. A simple way to determine if it's the reason for poor shift quality is to unplug the wiring connector on the transmission. Do this only for a brief test, as many transmissions will overheat fluid if they don't get clutch engagement. And be sure to unplug the right connector, as many automatics have more than one, particularly if they have electronic controls.

If the transmission becomes silky smooth with the clutch circuit disabled, check with the dealer to find out if there's a factory fix. On newer cars there may be a reprogrammed computer or a new computer module to raise the shift speed. Or there may be a recalibrated pressure switch (to accomplish the same thing) in a hydraulic passage for the transmission-speed governor.

Another cause of shudder or a vibration is a loose, broken or missing vibration damper that may be attached to your transmission (Fig. 4). Not a lot of cars have a damper, and the damper alone doesn't directly affect shift quality. But if it's not intact and tightly mounted, the overall response of the transmission is affected, and that may worsen the feel of the shifts.

Whatever you do, don't try to cure shift problems by pouring some additive, particularly a limited-slip differential oil, into the transmission. Limited-slip oil contains a large dose of special friction modifiers that cures harsh shifts and shudder. But it also causes so much slippage that the clutches and bands will burn out fast.

What about harsh downshifts on acceleration? Well, the typical downshift, made at full throttle or close to it, generally is harsh. If it's the only harsh shift and it suddenly becomes much harsher, check the fluid level. If the fluid level is normal, and you have a fully hydraulic automatic, open the manual and look for a downshift cable adjustment. It often is similar to the throttle-valve cable adjustment, but made at wide-open throttle. Check the manual.

The odds are very good that you'll find the reason for your car's harsh shifts somewhere among the basic items we've covered here. Then, you can send those nasty little shifts back to charm school.

EASY

Replacing Your Automatic Transmission Filter

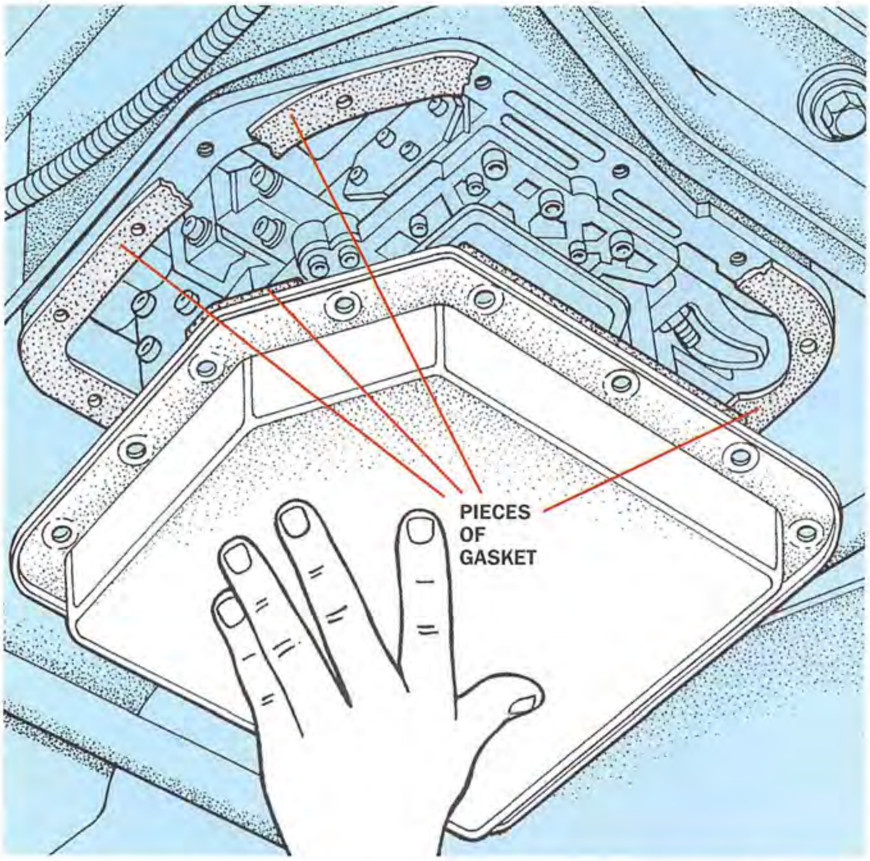

(Fig. 1) Jack up the car and get a good-size drain pan. Loosen and remove the bolts on the transmission pan from rear to front, then drain the oil.

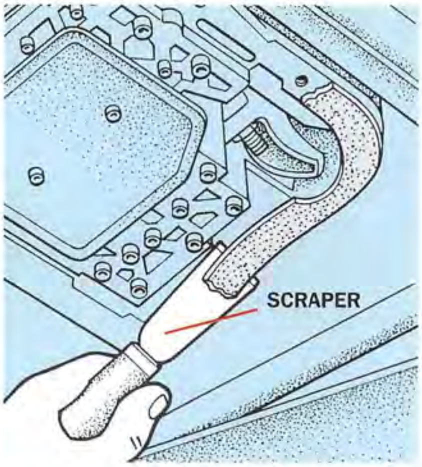

(Fig. 2) Use a scraper to remove the remains of the old gasket, being careful not to scratch the aluminum transmission casting.

You've just spent the hottest 4 hours of your life creeping through a colossal traffic jam. You pull into the driveway and collapse into your favorite comfy chair in your air-conditioned den. Later, in the cool of the evening—before you begin packing the family car for a getaway—you check the car's vital fluids. The oil and coolant are okay, but the automatic transmission dipstick is still too hot to touch—the fluid on it is as black as squid ink and smells like the bottom of a barbecue grill after a picnic. Time to change the transmission oil and filter.

The oil in your automatic transmission can get very hot—often approaching 300° F—when you're pulling a heavy load or moving slowly in hot weather. But it also can get every bit as torrid when the wheels are spinning and slipping on icy roads in winter, so this is a year-round issue. And the oil becomes contaminated with particles that have worn off the transmission's clutches and bands, even metal shavings from the pump and other moving parts.

Don't wait until the oil sample on the automatic transmission dipstick has a burned odor to change both the oil and the filter. Unlike engine oil, the automatic transmission oil not only has to lubricate, it also has to provide hydraulic pressure to operate the clutches and bands that produce the gear-ratio changes.

Factory recommendations for transmission oil and filter changes range widely. GM, for example, says 100,000 miles, but that's under ideal conditions. And the recommendation comes down to as low as every 15,000 miles given one or more of the following: you drive in a lot of traffic and high temperatures, periodically pull a trailer or other heavy load, let the car sit idling for long periods or subject it to other operation similar to taxi, police or delivery services. Mom's taxi service surely falls in the latter category, so a transmission oil and filter change every two years or every 30,000 miles is good practice.

Getting Greasy

Jack up the front of the car, support it on safety stands, put a large catch pan underneath the transmission and loosen the transmission-pan bolts (**Fig. 1**). Remove the bolts at the rear, then loosen the others, so the oil drains out one side into your catch pan. On the typical car, several quarts will come out, which is less than half of the oil inside—and there's no practical way to get out any more oil. Consider that

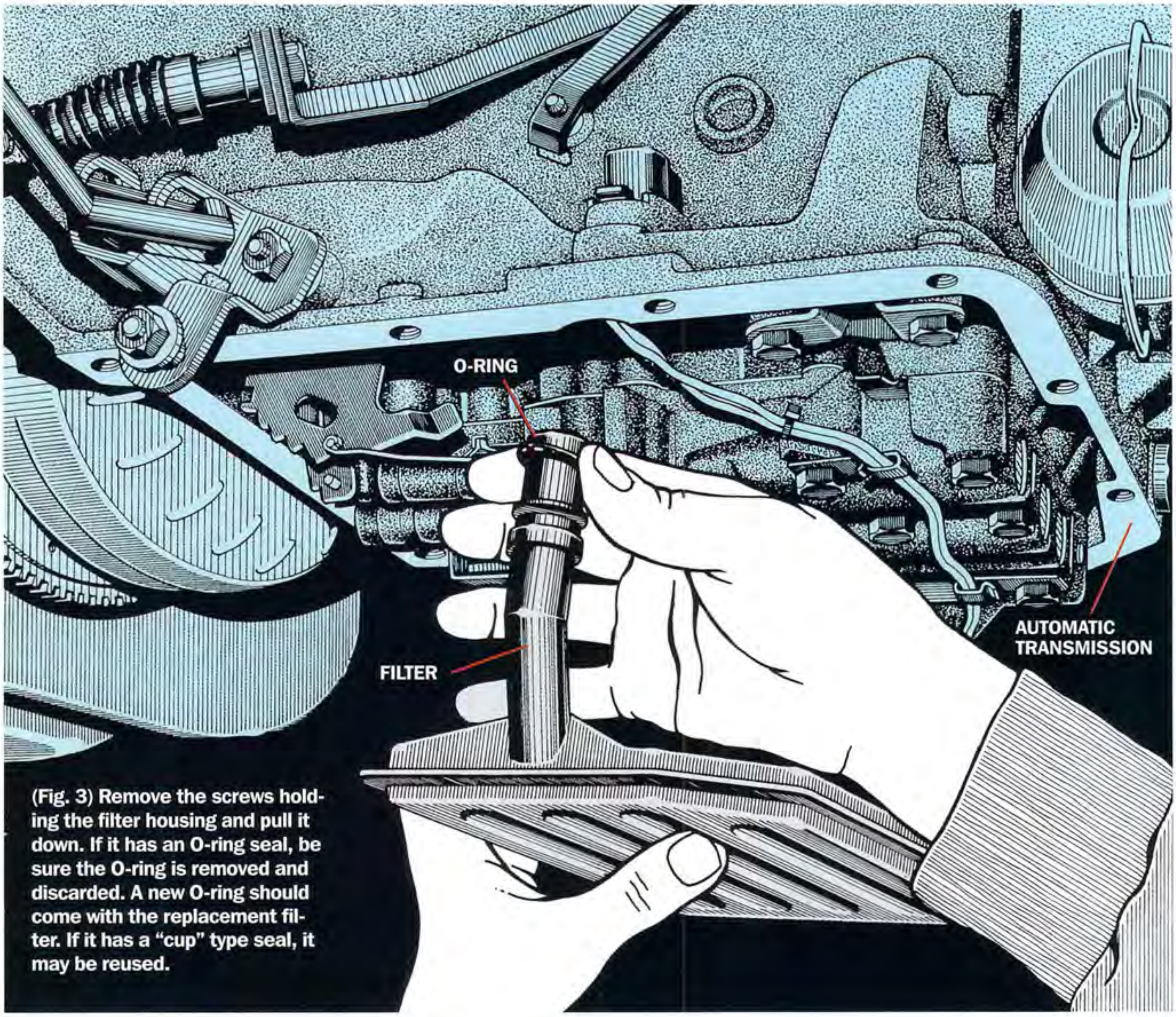

(Fig. 3) Remove the screws holding the filter housing and pull it down. If it has an O-ring seal, be sure the O-ring is removed and discarded. A new O-ring should come with the replacement filter. If it has a "cup" type seal, it may be reused.

the next time you want to postpone a transmission service. Finally, remove all the bolts and lower the pan. (And no, most transmissions don't have a drain plug.)

Scrape the gasket from the pan, taking care not to gouge the flange surface **(Fig. 2)**. Wire-brush any gasket residue from the pan. If any piece of gasket sticks to the underside of the transmission, gently remove it with a scraper—not a screwdriver, as the surface normally is soft aluminum and prone to being gouged. A gouged surface may not seal, even with a new gasket. If the gasket is rubber with a raised center section, it's reusable if it is still in perfect condition. Just clean and dry the gasket surfaces on the pan and transmission. If it's damaged, replace it with one of the same type.

Clean the pan thoroughly, as there will probably be a fair amount of silt in the corners, from sludge, worn friction-clutch material and tiny bits of metal. Big flakes of metal might indicate a problem, but expect a certain amount of shiny stuff.

If the oil-pan bolt holes are distorted by previous overtightening, place the edge of the pan on a heavy metal surface (your workbench, for example) and with a ball-peen hammer, tap them down flat to ensure a good seal.

This Is a filter?

It probably looks more like a fine-mesh window screen, but this is what passes for a filter. It's adequate to screen out chunks, but very small abrasive particles pass right through, another compelling reason to service the tranny regularly.

Remove the old filter, which typically is held by screws. Pull it down slowly, as it may have a neck with an O-ring, and you should remember where that ring is **(Fig. 3)**. Some O-rings sit on the neck, but are prone to sticking inside the case when you remove the filter. If you don't remove the old O-ring and add a new one with the new filter, it won't seat properly. Most General Motors cars have an O-ring, but there's a cup seal that goes in the case on some models since 1994. The filter neck is inserted into it, and there's no O-ring at all. Consider retrofitting this new filter

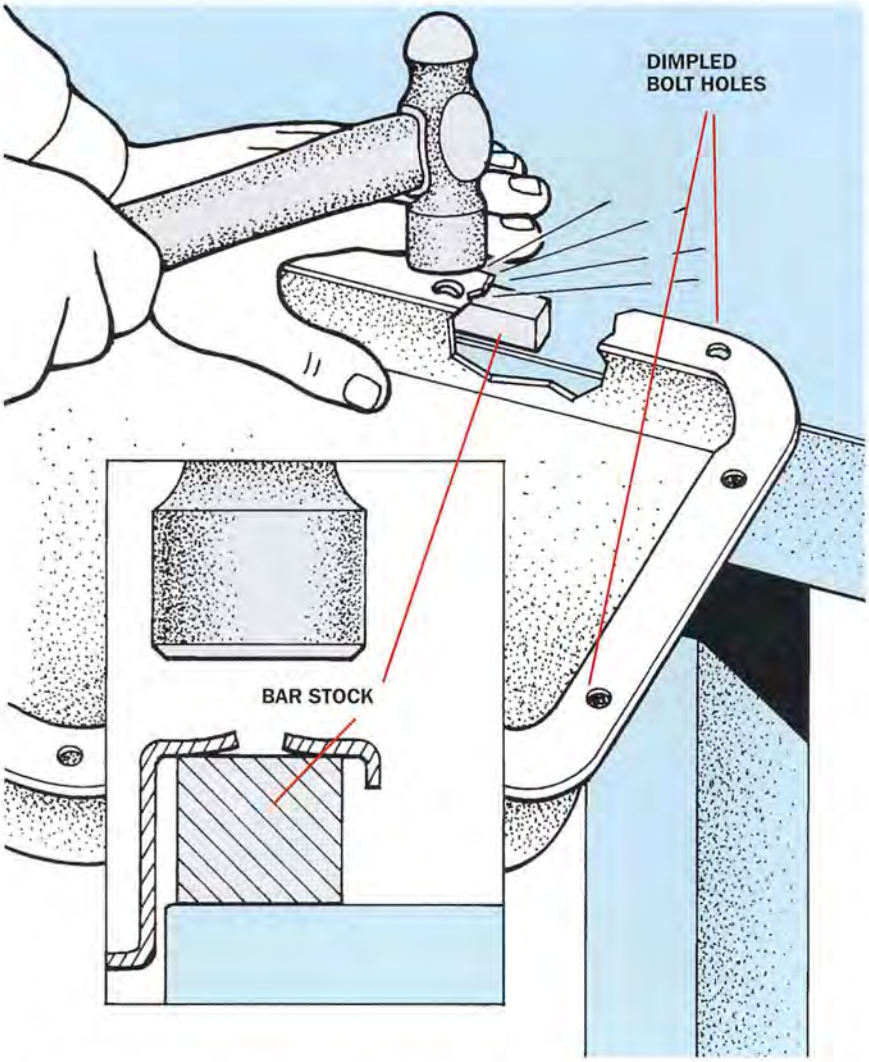

(Fig. 4) Overtightened bolts can distort the pan's sealing surface. Flatten the dimples with a hammer against a hard surface.

and cup seal to the earlier-type tranny.

Apply a thin film of nonhardening gasket sealer to the oil pan, lay the new gasket on it and install a bolt through each corner to hold the gasket in place.

Hand-start each bolt. Don't tighten any bolts more than two or three threads until they're all in place (Fig. 4). Then run them down with a ratchet until the gasket is in contact all the way around. Tighten the bolts to factory specifications with a torque wrench, using a spiral pattern starting from the center of the pan and working alternately to the ends. Go in steps, a few foot-pounds at a time, up to the torque spec.

Glug, Glug

Finally, add fresh oil. Not all cars use the same type. Most late models take Dexron IIE or Dexron III, although Ford recommends an oil that meets its Mercon standard (including some older cars for which a "Type H" fluid was prescribed). Check your owner's manual. Major brands of oil are formulated to meet all these carmaker requirements, and you should use one of them. A few European cars and older Fords (some through 1981) may require a Type F fluid, which is nothing like the Dexron and Mercon formulas. If Type F (or the virtually identical Euro-spec Type G) is listed in the owner's manual, that's what you should pour in. New Nissan products may call for Type J or Type K fluid, which is not a Dextron type.

Oil is generally added to the trans through the dipstick tube and you'll need a long skinny funnel or a funnel and a hose to get it in. Some GM automatics have no dipstick and removal of a fill plug is necessary.

Whatever type of oil you need, follow the carmaker's precise recommendations for checking fluid level. Typically, you have to cycle the shift lever from P to L and back a few times and add the oil with the engine running. Do it slowly, recheck repeatedly, as less than a pint can be the difference. Overfilling causes as many performance problems as a low fluid level, so if there's any transmission oil left in the quart bottle, don't pour it in. Put the cap back on and save it. If someday down the road you find the dipstick reads low, what's left in the bottle should be just the right amount to top it off.

DIFFICULT

Replacing Your Front-drive Car's Clutch

The problem was almost imperceptible at first: Going up a slight hill at highway speeds, you noticed that the engine was revving higher even though your car wasn't going any faster. Backing off the gas or dropping into a lower gear made the problem go away each time.

But that was a month ago. Now you spend your days looking for routes that have no hills, because even a slight incline causes the engine to rev wildly while the car slows down. There's also the smell of burnt buttered popcorn in the air as you downshift in a feeble attempt (the semi driver on your bumper is blasting his air horn) to reduce the engine load and get the car going again. Your biggest fear is that you'll end up backing down the hill. The problem isn't in the engine or transaxle. All that slipping is caused by a worn-out or misadjusted clutch.

Compared to other automotive devices, the clutch **(Fig. 1)** is decidedly low-tech (see How It Works, following page). And while replacing and adjusting the clutch doesn't require exceptional skill or high-tech equipment, there are some things you should look out for.

Please Release Me

It couldn't be simpler: Push the clutch pedal in and the clutch should disengage. However, if the clutch doesn't fully disengage, you'll hear a lot of grinding during shifts or when you try to put the car in gear. "A lot of people immediately think there's a problem with the clutch disc or pressure plate when the clutch doesn't disengage," says Gabe Vajda, a technician with LuK, a maker of clutch parts.

"Any play in the clutch-pedal parts soaks up travel, so the clutch won't fully release," Vajda cautions. He says that before you remove the transaxle

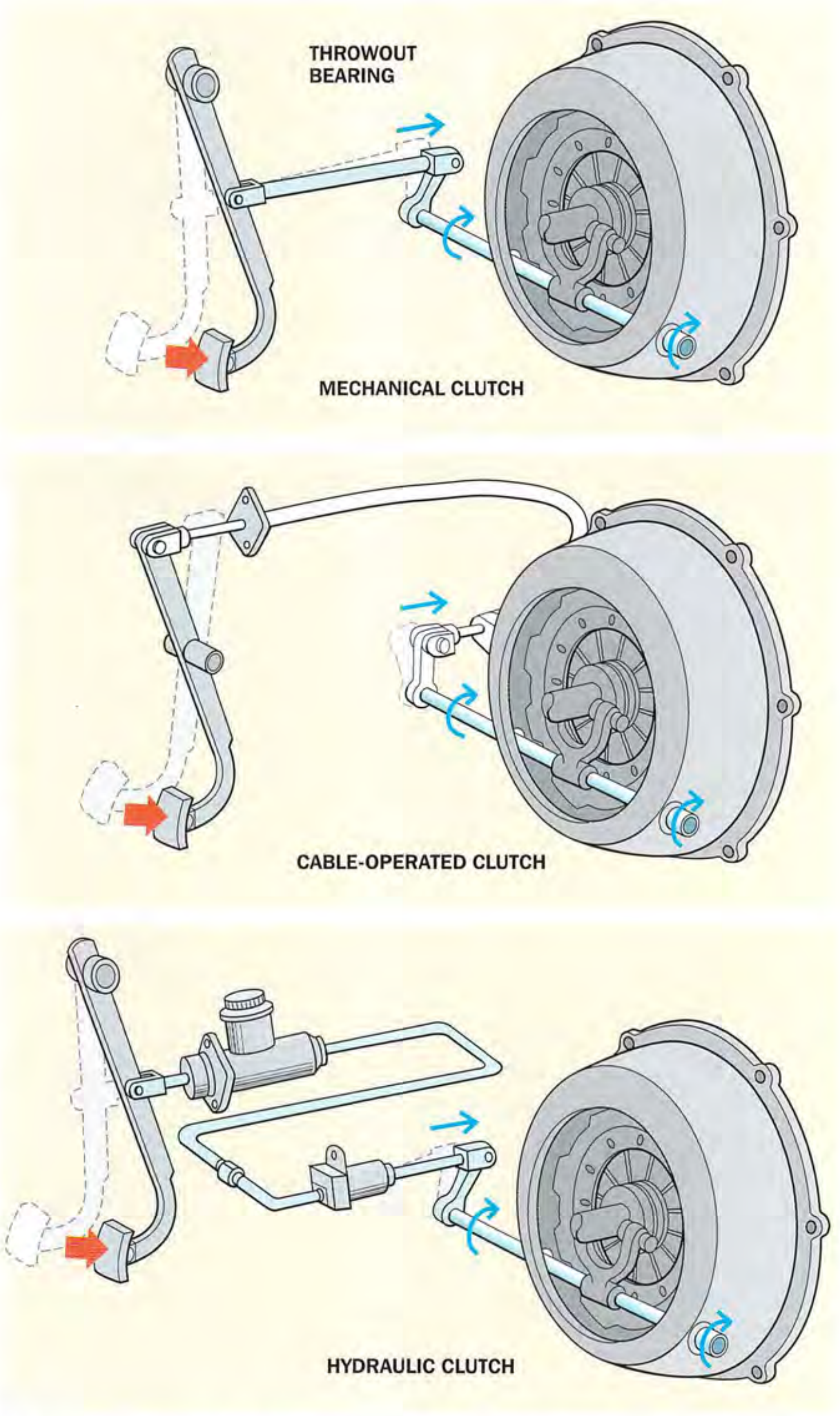

(Fig. 1) Regardless of method, depressing the clutch pedal will push the throwout bearing against the pressure plate. There should be some adjustment mechanism as well.

and clutch, it's wise to check the pedal-arm bushings and the actuating cable, or, on cars with hydraulic systems, the slave and master cylinders. Also, a nonreleasing clutch could even be caused by something as simple as carpeting under the pedal that's too thick, which has been the case with some 1988-90 Chevy Beretta and Corsica models. Trimming away insulation under the carpet will restore full pedal travel.

If a clutch isn't releasing, start by manually prying the release arm at the bellhousing with a screwdriver. If the arm moves freely and the clutch releases there, chances are good that the problem lies outside the bell-

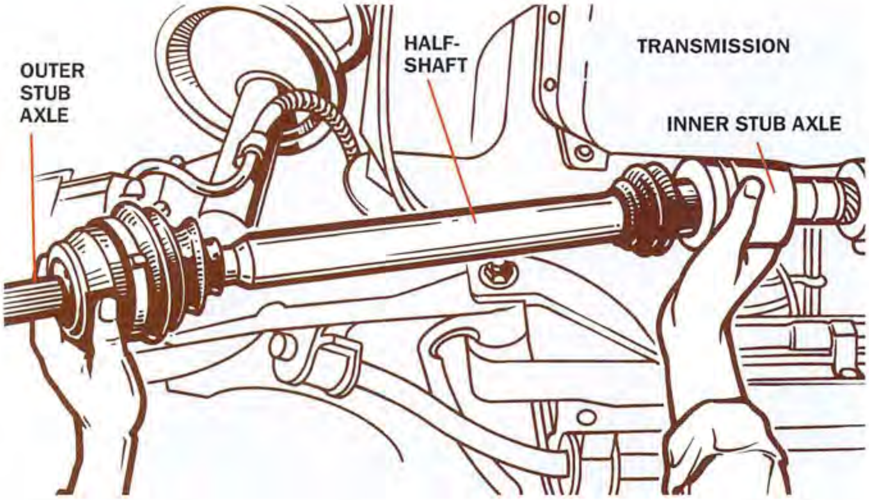

(Fig.2) Remove the outer stub axle from the hub, and then pull the entire axle free of the transaxle. You may need to plug the inner socket.

HOW IT WORKS

Clutch Play

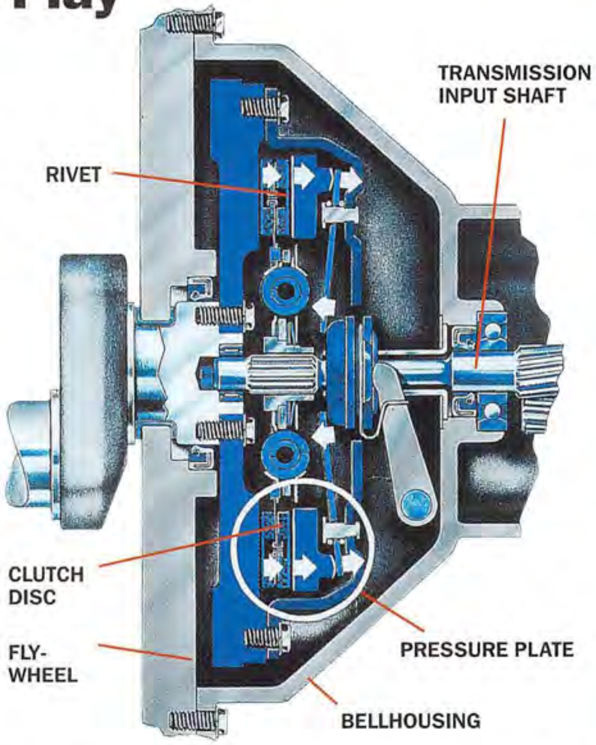

At the heart of the clutch assembly is the clutch disc. It has friction material, similar to brake-lining material, on each side and a splined hub that slides back and forth on the transaxle's splined input shaft. When the clutch pedal is not depressed, the clutch disc is tightly clamped between the flywheel and the pressure plate to firmly connect the engine to the transaxle. Depressing the clutch pedal allows the clutch disc to spin freely between the flywheel and pressure plate, allowing the engine, flywheel and pressure plate to spin on their own, while the splined input shaft and transaxle freewheel with the disc.

The clamping force that engages the clutch comes from a large spring, or series of springs, located inside the pressure plate called a diaphragm. Depressing the clutch pedal presses on the diaphragm spring, causing the pressure plate to move away from the clutch disc, and the clamping force diminishes until the engine can spin independently of the transaxle.

To get the clutch moving, the clutch pedal is connected to a sheathed cable or a hydraulic master cylinder and slave cylinder. The cable system works much like a heavy-duty version of the brake-caliper cable that can be found on a 10-speed bicycle—depressing the pedal pulls on the cable to release the clutch. With the hydraulic system, pressing on the clutch pedal moves a piston in the master cylinder, which displaces fluid, causing the piston in the slave cylinder to move. This, in turn, operates the clutch. Some older cars and trucks use a simple mechanical bellcrank linkage instead of a cable or slave cylinder, but the mechanical principle is the same: Depress the clutch pedal and the clutch is released. How?

Whether the clutch-pedal system is operated hydraulically or by cable, it moves a release fork that pivots on the bellhousing. The pivoting action moves the end of the fork toward the pressure plate. A throwout bearing (also called a clutch-release bearing) at the end of the release fork applies even pressure to the diaphragm spring and the pressure plate moves back, releasing the clutch disc. Let the clutch out, and the pressure on the spring releases and the clamping force on the disc returns.

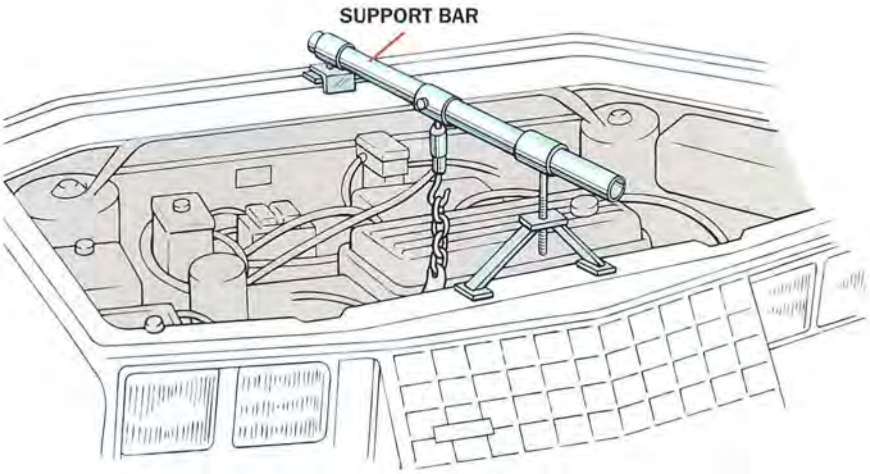

(Fig.3) A support bar or a stout piece of timber will be necessary to support the engine while the tranny is out.

housing. On cable-operated clutches, there could be too much play in the cable, which can reduce travel. To reduce cable play, find the adjusting nut at the top or bottom of the cable and adjust the cable to the proper specifications. On hydraulically operated clutches, air in the lines, or a worn master or slave cylinder, can result in reduced travel. To bleed the air, top off the master cylinder reservoir with brake fluid. Place a piece of hose on the slave cylinder's hydraulic vent, then place the other end of the hose into a partially filled glass bottle. Open the vent, then have an assistant repeatedly depress the clutch pedal slowly until any bubbles coming from the end of the vent line disappear. After tightening the vent, clutch travel should return to normal. If it doesn't, then the slave and master cylinders should be replaced.

Removal and Inspection

Barring adjustment, the only way to fix a slipping clutch is to replace it. Installing a new clutch is very straightforward. But you have to make sure the work is done right, since fixing any mistakes usually means removing the transaxle—something you don't want to do more than once. It's best to consult a service manual for instructions specific to your vehicle. The following is a general procedure for replacing a clutch.

Start by disconnecting the battery positive cable. Then, while working under the hood, prepare the transaxle for removal by disconnecting the clutch cable or hydraulic slave cylinder. Also, remove or disconnect items that prevent the transaxle from being removed, such as backup-light wires, air-cleaner ducting, exhaust pipes, the starter motor, speedometer cable and other items. If you are in doubt as to exactly what needs to be removed, consult a manual.

Next, chock the rear wheels and jack up the front end of the car and support it with stands. Working from under the car, remove any additional parts that you couldn't reach from under the hood, including the bolts that hold the axle shafts to the transaxle on most vehicles (Fig. 2).

To remove the transaxle, you normally have to remove one or more engine mounts. But before you can start removing them, you'll have to support the engine. With many cars, it's possible to support the engine from underneath by placing a jack under the oil pan (with a third piece of wood on the lift pad to prevent damage to the pan). But on other cars, a special engine-support rod that allows the engine to hang while the transaxle is out must be installed under the hood. (Fig. 3) Again, consult a manual for the proper method.

To disconnect the transaxle from the engine, support the transaxle with a jack and remove the bolts from around the flywheel bellhousing. Next, slide the transaxle away from the engine until the transaxle input shaft clears the pressure plate, then lower the transaxle to the ground and roll it from under the car and out of the way.

Once you have removed the trans-

(Fig.4) The clutch is readily accessible—but only after removing the transaxle from the engine. You'll need to remove or unhook the axles as well.

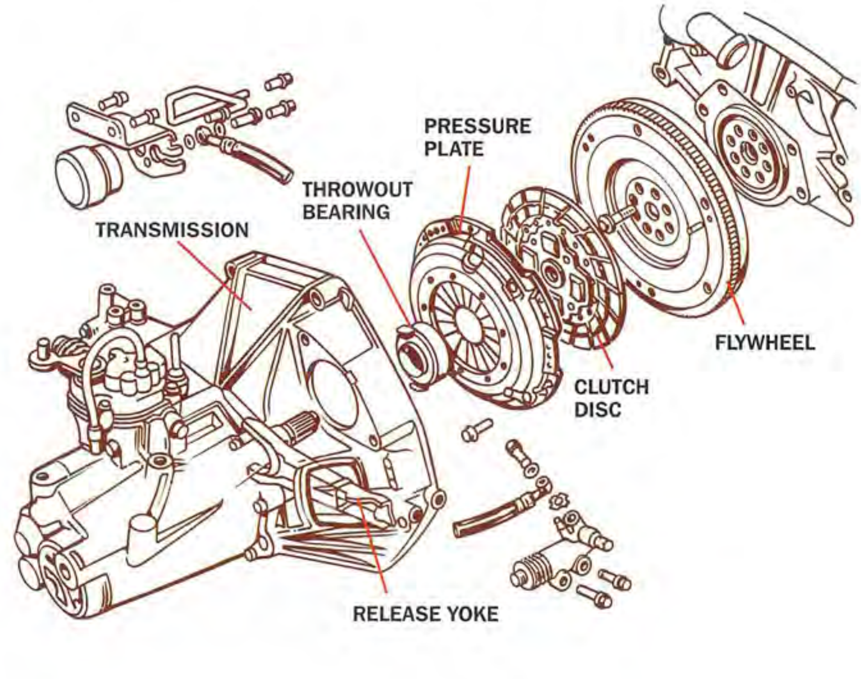

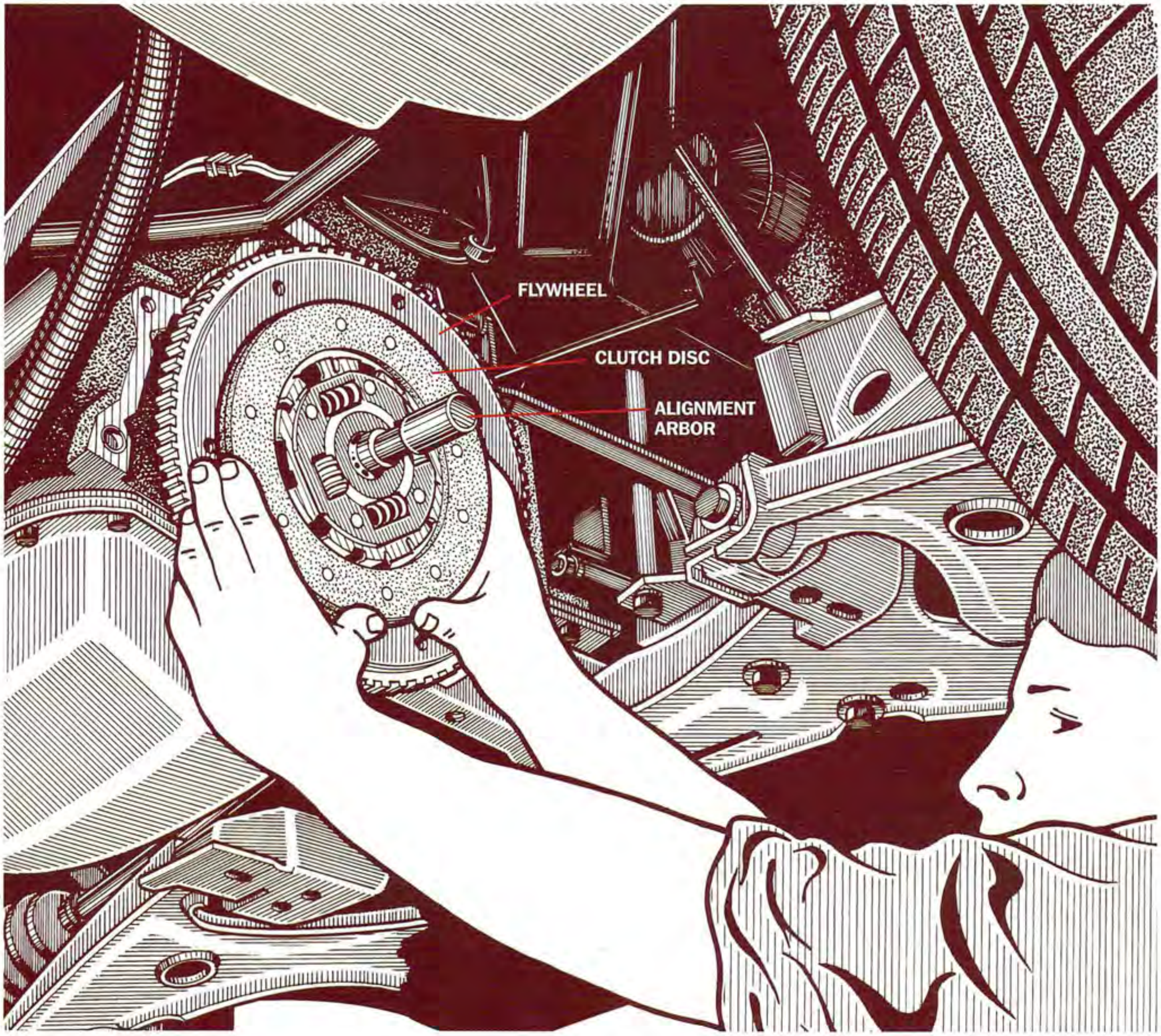

axle, you will then be able to access the pressure plate and clutch disc (Fig. 4). Remove the bolts from around the pressure plate, then remove the plate and clutch disc. Inspect the friction surface of the flywheel. If it is scored, checked or shows signs of hot spots, it must be removed, machined at a machine shop and then reinstalled. Keep in mind that it's good practice to have the flywheel machined so the clutch disc has a proper surface to break in on. If you don't have the flywheel machined, sand it lightly with medium-grit sandpaper wrapped around a small block of wood, then wipe clean with a damp cloth.

When inspecting the flywheel, also inspect the pilot bearing or bushing at the center of the flywheel. The needle bearings (if applicable) should be lubricated and show no signs of galling. If there's any doubt about the quality of the bearing, remove it with the appropriate tool and install a new one. You should also inspect the back side of the flywheel for signs of oil leakage at the engine's rear main seal. If oil reaches the surface of the clutch, chattering and grabbing will result. Some cars have "stepped" flywheels, which is when the clutch disc rides on one surface and the pressure plate bolts to a surface that's stepped above the clutch surface. It's important that both surfaces be machined the same amount, or slipping or failure to release could result.

(Fig.5) You'll need an old tranny input shaft, this special arbor, or a wooden dowel to center clutch disc on flywheel as the pressure plate is installed and torqued down.

If the flywheel is machined too much, it moves the clutch and pressure plate away from the release bearing. This can keep the clutch from fully releasing. You might be tempted to use shims between the crankshaft and flywheel to move the clutch assembly toward the release bearing. However, it's possible for a thin flywheel to burst at high rpm. The only safe alternative is to replace the worn flywheel with a new one.

Inspect the transaxle's input shaft

seal to be sure it isn't leaking. (There's no need to inspect the clutch disc, pressure plate or release bearing, since you'll be replacing those.) If the seal is leaking, replace it. Next, look for leakage at the engine's rear main seal. If there's more than a hint of seepage there, this may well be the reason for the early demise of the clutch disc. Mark the flywheel so you can reinstall it back on the crankshaft in the correct index. Remove the flywheel and remove the old seal. Carefully install the new seal without gasket sealer by tapping it in with a seal installation tool, or at least a large socket or a piece of wood. Check the crankshaft's sealing surface for a depression where the seal has eroded the metal (this is particularly common in dusty areas). There may be a repair sleeve available for your engine, so check with your machine shop if you think you have a problem.

Installation

Before you install a new or freshly machined flywheel, make sure the crankshaft flange is clean. After placing the flywheel on the flange, tighten the bolts in a star pattern to the proper torque specification. There are two tricks to installing the clutch disc and pressure plate. First, the disc is designed to go on in one direction. If you look at the center hub, you'll notice a series of damper springs. On one side of the clutch, this damper hub sticks out from the friction surface. That side always goes toward the transaxle. Put it in the other way, and you'll soon be taking the transaxle out to do it all over again. Second, you'll need a "clutch pilot tool" to install the clutch disc so it aligns with the pilot bushing in the crankshaft (Fig. 5). To install the disc, slide it onto the pilot tool, then securely stick the center of the tool into the pilot bearing. Install the pressure plate over the clutch disc, torque all bolts to specifications using a star pattern, then remove the pilot tool.

Before installing the transaxle, install the new release bearing onto the release fork, then make sure that the fork operates freely. Jack the transaxle into position and slide it forward so that the input shaft slides into the splined hole in the clutch disc. You may have to turn either the input shaft or the engine's crankshaft (and the flywheel and clutch along with it) to get the splines to align. Don't force the input shaft into its splined hole. When everything's properly aligned, the shaft will slide right in. If it doesn't slide right in, don't force it. Back it out and try again.

Never let the transaxle hang unsupported from the engine, since this can lead to damage to the pilot bearing, input shaft and front input shaft bearing.

After installing and torquing all the bellhousing bolts to spec, you can reinstall any motor mounts that you had removed, the drive axles, as well as any other components under the car. Once the underside work is done, remove the jack holding the transaxle and lower the car to the ground. Reinstall any underhood components.

You don't really have to replace the clutch cable at this time, but because it costs less than $75, replacing it is good insurance against trouble later on. In any case, adjust the cable to the proper amount of free-play.

Most hydraulic clutch actuators are self-adjusting and, therefore, have no free-play. Still, you'll want to make sure the self-adjusting action is working properly. To do that, simply push the slave cylinder piston back into its bore. If it moves, the system is working properly. If it doesn't move, it means the slave cylinder's pressure-relief orifice is clogged and the slave cylinder should be replaced.

MODERATE

Changing Your Transmission Fluid

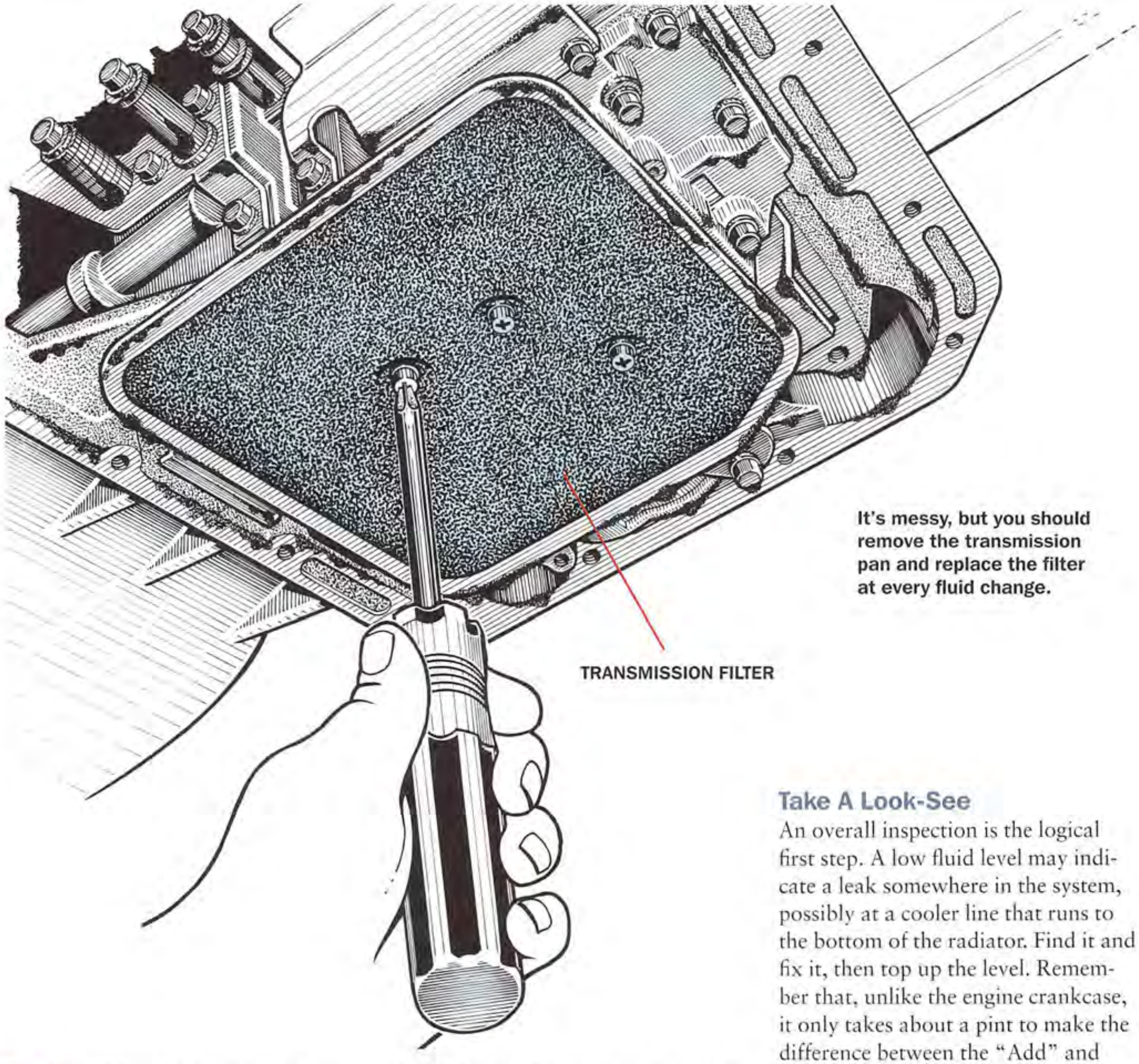

It's messy, but you should remove the transmission pan and replace the filter at every fluid change.

You drive the same route to work every day—same open stretches, same intersections, same stop-and-go. But today there's a subtle discord in the usual harmony, a blip on your vehicular radar, a bad vibe in your mechanical karma. Shifts seem oddly late and soft. Later, as you pull into the drive, you sense something peculiar. Letting your car idle, you pull the dipstick out of your auto transmission. Fresh automatic transmission fluid (ATF) is bright red and has a distinct petroleum smell. Your dipstick shows a low level, is the color of institutional linoleum and smells like the bottom of a barbecue pit after a biker wedding. Your transmission fluid is badly in need of changing, and the tranny may already be damaged.

Take A Look-See

An overall inspection is the logical first step. A low fluid level may indicate a leak somewhere in the system, possibly at a cooler line that runs to the bottom of the radiator. Find it and fix it, then top up the level. Remember that, unlike the engine crankcase, it only takes about a pint to make the difference between the "Add" and "Full" marks. Also, make sure you use the correct ATF, which we'll discuss later. If you're lucky, the lag or shifting problem may just disappear after you add ATF.

The fluid should be bright red, clear and "sweet" smelling. If it's a smoky dark color, or has a burned odor, a complete change is needed, but the damage may already have been done.

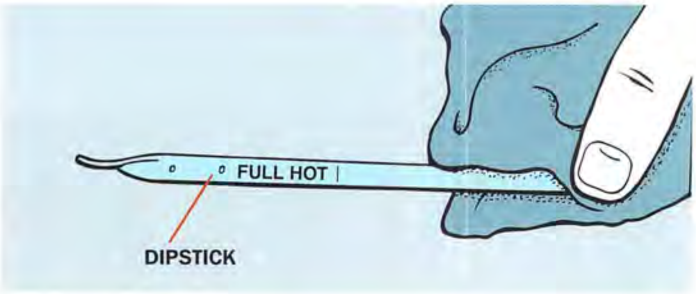

Check the dipstick in Neutral while the engine is warm and idling.

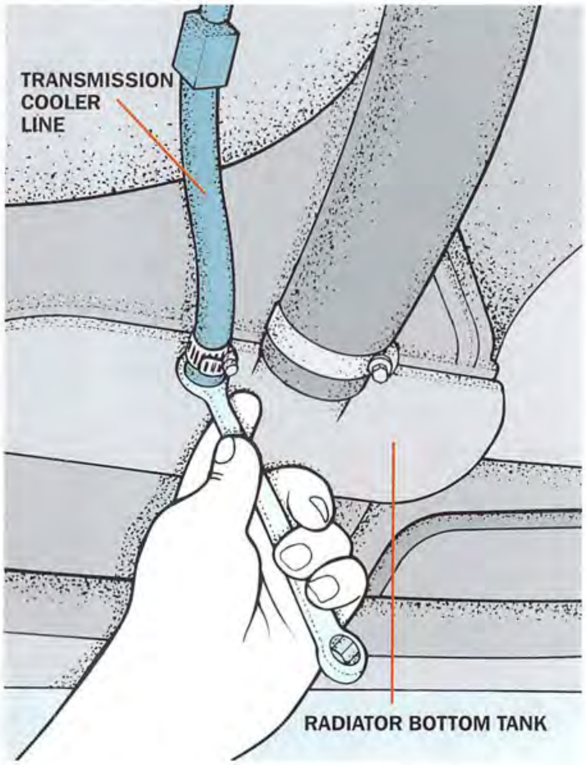

Remove trans cooler line at the radiator to flush out the last of the ATF.

Lockup?

All modern automatics (except for the continuously variable transmissions—CVTs—found on a few late-model Asian cars) have locking torque converters to eliminate slip at cruising speeds, thus saving fuel. These are controlled by the powertrain control module (the engine and transmission management computer) on the basis of speed, temperature, throttle position, etc. If the engine is running at a higher rpm on the highway than usual—300 to 500 more—to maintain the same speed, it's possible that lockup isn't occurring. Besides reducing fuel economy, this can have the much more disastrous effect of causing the transmission to overheat.

Check that the transmission converter clutch wire that runs from the harness to the transmission is connected and intact.

Fresh Fluid

The single most important thing you can do to head off big-bucks transmission repairs is to change the ATF on a regular basis.

Some carmakers have backed down from the unrealistic 100,000-mile trans fluid change interval recommended in the past. Every 30,000 miles is much more reasonable. If you tow heavy loads in hot weather, you might even think about annual changes.

Going through the messy operation of dropping the transmission pan and replacing the filter is fine as far as it goes. The trouble is, it doesn't go far enough. At least half of the old, burned-up ATF and its contaminants remain in the torque converter (the days of those convenient converter drain plugs are long gone), clutch drums, valve body and elsewhere. If you want to get the full benefit from this maintenance service, you've got to work a little harder.

Regardless of how far you're willing to go here, you still must take the transmission pan off, and there are a couple of ways of making this job a little neater. Start by putting the car on sturdy jackstands or, better yet, ramps. Block the rear wheels. If you have a gravel driveway, toss a 4 x 8 sheet of plywood down first to prevent the stands from tunneling into the ground while you're under the vehicle.

If you just remove the pan (leave a few bolts along one side partway in), ATF will flow out in a wave all around the seam, probably splashing outside the radius of your catch pan. If you've got a suitable pump, you can run the pickup hose down into the dipstick tube until it bottoms out, then pump until you stop getting fluid. This will vastly reduce spillage.

To extract as much of the old ATF as possible, leave the pan on, remove a trans cooler line at the radiator, put a drain pan under it, then start the engine for a few seconds to

Pour ATF back into the dipstick tube with a funnel.

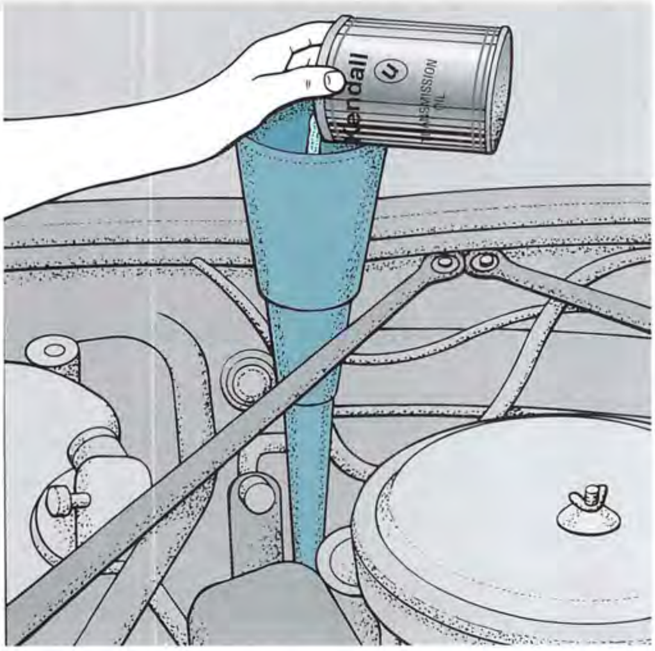

find out which way the fluid is flowing. It doesn't matter whether you use the inlet or outlet line except that you have to attach a small hose either to the line connector or the radiator outlet in order to collect the ATF. Put the hose into the largest jug you can find, and let the engine idle until air starts spurting. Many professionals enhance this procedure by pouring a few quarts of fresh fluid into the dipstick tube at roughly the same rate that the old fluid is coming out, thus adding flushing action.

Now you can remove the pan. This is not only necessary for changing the filter, but also allows deposits and sediment to be washed out of the pan. There's another important consideration: This operation provides the opportunity to find out if failure is impending. Judging this is somewhat subjective, so we asked an ASE Master Automotive Technician his opinion.

"You should see next to no swarf or debris, and then only on the first change," he says. "Subsequent changes should be nearly dead clean. If a newer gearbox is making junk, it's in trouble. You might find just a trace of aluminum shavings, or other very minor debris, but the assembly process is so clean, and the newer gearboxes so unforgiving of dirt, that any real accumulation generally means a problem is in development."

Now's the time to replace the filter and its seal, which probably can be purchased in the same kit as the transmission pan gasket. When reinstalling the pan, start every pan bolt by hand for at least two threads before tightening any of them.

If the last person to install your pan got overly enthusiastic with the wrench, you may find the pan rail has dimples around the bolts. Use a hammer and dolly to flatten them out. Otherwise, the pan gasket will leak. A cork gasket often can benefit from a thin layer of gasket sealer or adhesive, especially to keep it in place while you're trying to start those first few bolts. Don't use a thick bead of silicone sealant, as it will squish out between the mating surfaces into little silicone worms, which will eventually break off and clog the pump intake.

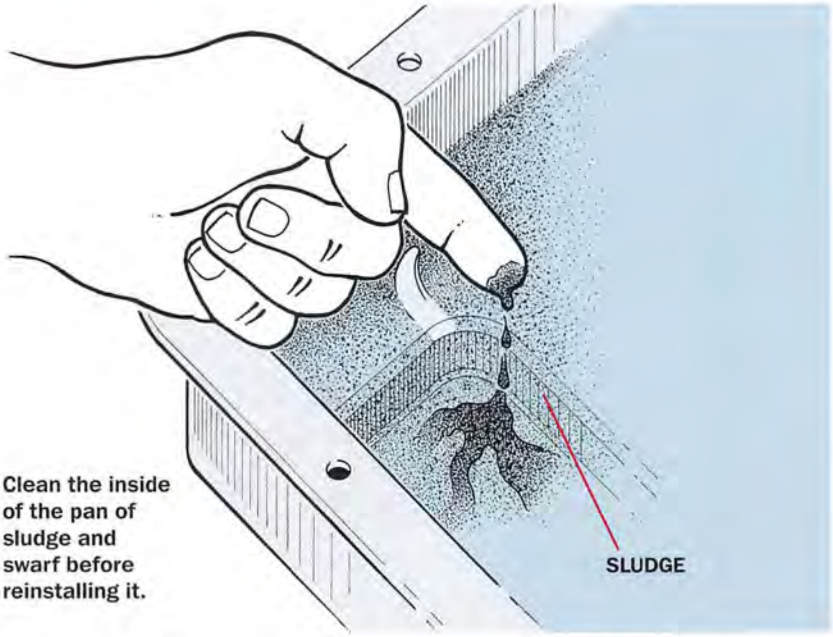
Clean the inside of the pan of sludge and swarf before reinstalling it.

Delegate

Of course, you can go to your favorite auto service facility and have a trans flush and refill done. Many shops today have a machine for this purpose, but you've got to be sure of what you're getting. Some quick-lube places will just attach the machine to a cooler line, exchange the fluid, and call it done. We beg to differ. The pan should be removed for cleaning.

You can try cleaning the filter, but it's usually better to replace it.

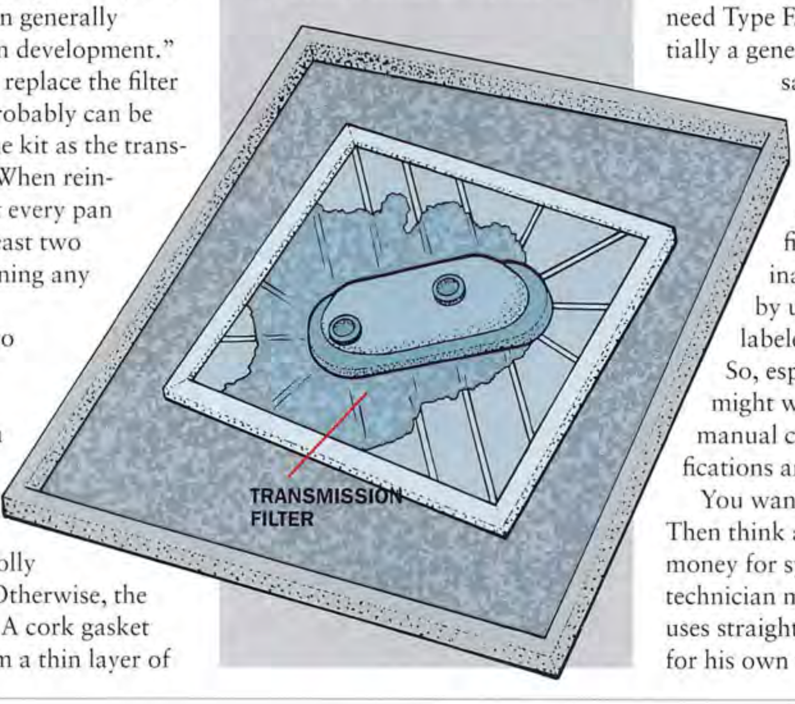

The Right Stuff

Most of the automatic transmissions on the road will work just fine on Dexron III/Mercon ATF, except for '92 and earlier domestic Fords, which need Type F. But the Dexron is essentially a generic fluid, and some experts say they've cured shifting problems simply by replacing it with the exact O.E.-specified stuff. They have also confided in us that they've inadvertently caused trouble by using bulk ATF that was labeled, "Will also work in ..." So, especially with imports, you might want to read your owner's manual carefully where fluid specifications are listed.

You want real peace of mind? Then think about spending the extra money for synthetic ATF. The master technician mentioned earlier always uses straight $5-per-quart synthetic for his own vehicles.

MODERATE

Replacing U-Joints

That vibration you've managed to ignore for weeks has finally gotten to the point at which it's hard to keep from spilling your morning coffee on your pant leg. You resolve to make two stops on the way home from work this afternoon—one at the dry cleaners to pick up clean pants, and a second at the mechanic's to investigate the vibration. As you accelerate away from the dry cleaners, there's a big change in the vibration. It's followed immediately by a lack of power and big clanging noises, the latter due to your driveshaft falling to the pavement. You've had a universal-joint failure.

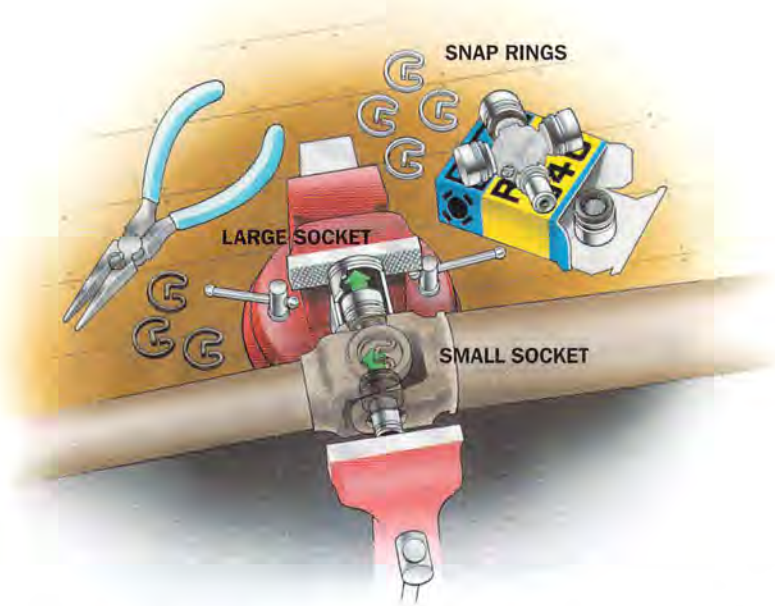

A bench vise and a pair of sockets make it simple to remove a worn-out U-joint. Use a small socket to press the cup into a socket bigger than the cup's outside diameter.

Lack Of Lube

Most of the time, loss of lubrication is the reason U-joints fail. Original-equipment joints typically have no grease fittings, and even replacement units often have zerks that you can't reach with your grease gun. The tiny needle bearings inside the cups over the ends of the trunnions rust and eventually crumble, which results in clearance where there should be none at all.

Another destroyer of U-joints is an excessive angle of operation. This may be due to an overloading condition, which makes the tail end sag, or a set of helper springs or air shocks that puts the rear of the car up into the stratosphere. In either case, leveling the vehicle is the only way to increase the universal joints' life span.

The driveshaft itself is a simple-enough component. Unless it's bent from contact with, say, a railroad track or is out of balance due to the loss of a weight, it'll keep spinning just about forever.

Rules Of Thumb

Vibration caused by a U-joint or driveshaft usually occurs at 25 to 35 mph, about 60 mph, or while braking at low speeds. That "body boom" that hurts your ears probably is caused by a bent or unbalanced driveshaft. A clunk when the transmission is put into gear, or when "floating" at about 10 mph, is either the result of damaged U-joints or a differential problem.

To find out if a vibration is coming from the driveline or the engine, attach a tachometer if you don't have one on the dash, run up to the speed at which the roughness is felt, and note the rpm. Then, shift into a lower gear, go back up to the same rpm and see if the vibration has diminished. No? Then you've found engine trouble. If, on the other hand, the vibes change considerably at the original engine speed, the U-joints or driveshaft are probably at fault.

Next, raise the vehicle and support it under the rear axle or suspension control arms so that the springs are compressed with a normal amount of weight and the driveshaft is at the proper angle. Grab the shaft near either end and force it up and down, then twist it hard back and forth. Any play you can see or feel is reason for replacement.

Brutal R&R

Odds are, you'll need to raise the vehicle off the ground (unless you're driving a giant-wheeled 4x4). Ramps are great for this, but a floor jack and a pair of jackstands will do the job fine. To replace a U-joint, unscrew the nuts of the U-bolts that clamp the universal's cups to the rear axle's yoke. Mark the parts so that you can put them back the way they were. Using a big screwdriver, pry the shaft forward until the joint is free of the yoke, then lower it and pull it out of the transmission tailshaft housing.

While a hydraulic press is ideal for driving the cups out of the yoke ears, you might be able to manage with a

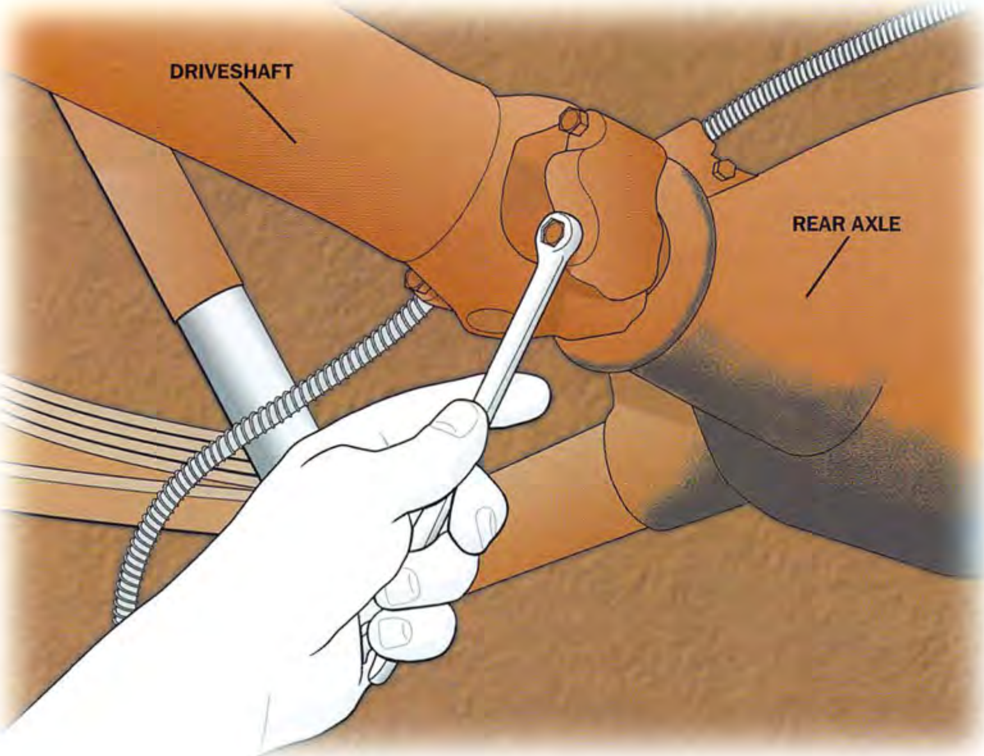

Remove the nuts on the U-bolts at the rear axle to disassemble the driveshaft.

husky bench vise. Or, you can drive the cups out with a drift and a heavy hammer. If these are the originals, they're probably retained by injected nylon, which will simply be destroyed in the process. Another option is to see if there are C-clips in grooves in the inner or outer sides of the ears. The inner type can be forced out easily with a screwdriver and hammer, but you may need snap-ring pliers for the outer variety.

Whether you're using a vise or a drift and hammer, place a socket that's bigger than the outside diameter of the cup against one ear to accept the cup. Then, with a vise, put a smaller socket on the other cup to push it through the hole in the ear. Now, turn the vise screw as necessary to force the cups through the yoke. If you're doing the pounding routine, position the large socket under one ear and use a big drift against the opposite cup. This probably will take an extra pair of hands.

When it comes to installing the new joint, first make sure it was packed with grease at the factory. The cups should be filled to about one-half of the needle length with the proper grease (usually SAE 140). While forcing the cups into place, you must be careful not to dislodge the needle bearings or jam them at an angle.

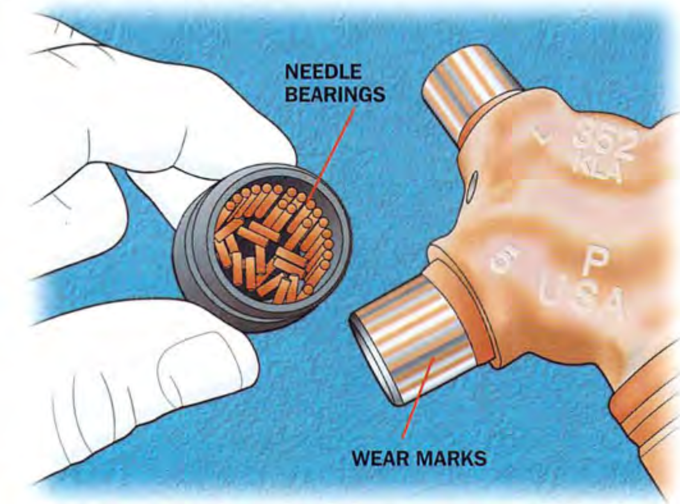

Dry U-joints will have rust inside, not grease. Do not try to save one that's rusty. Replace it.

Shaft Shivers

If you believe the driveshaft is out of balance, first make sure that the problem isn't due to a defunct U-joint, an excessively worn slip yoke or transmission tailshaft housing, wheel imbalance, or excessive runout of the shaft itself (generally, this shouldn't be more than about 0.040 in. as measured with a dial indicator).

Next, give the shaft a thorough visual inspection. Make sure it's clean—in high gear, it rotates at the same speed as the engine, so a lump of undercoating would be enough to give a car the shakes—then look for evidence of a thrown balance weight. Also, see if it's bent.

If you still haven't found the trouble, disconnect the shaft from the rear, rotate it 180°, bolt it back up, and road test. If the vibration is still present, you can resort to the old hose clamp trick:

1. Raise and support the car as safely and solidly as you can with the rear wheels off the ground.

2. Start it up, put it in gear, and have a helper step on the gas and rev it up to 40 to 50 mph.

3. Tape a piece of chalk or a crayon to a broomstick, hold it firmly with both hands and slide it under the car. Slowly move it so that it has the least possible contact with the shaft just ahead of the rear joint, then just behind the one at the front. The heavy side will be making a larger

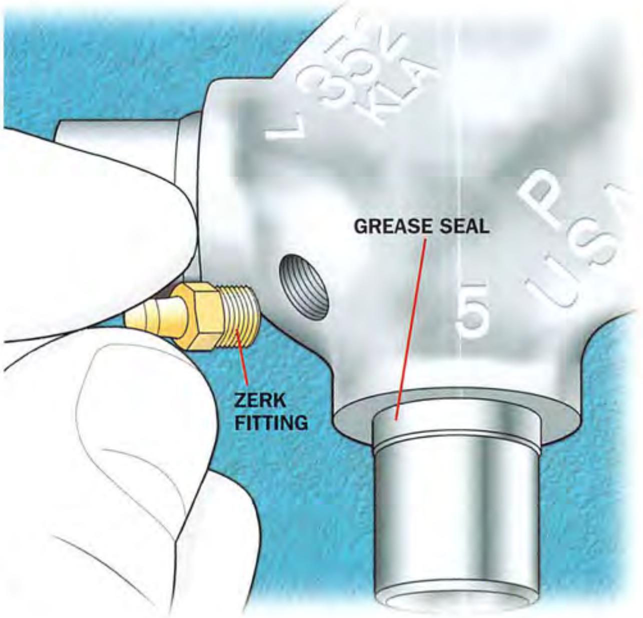

A new U-joint should have a zerk fitting even if the original did not. Fill the joint with grease until it runs out of the seals.

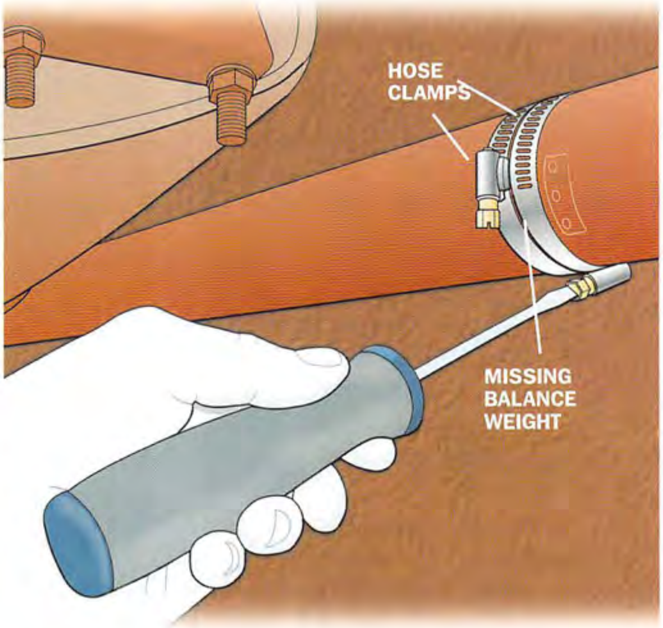

Hose clamps are an acceptable field fix for a missing balance weight. Move the heads apart to adjust the weight.

circle than the light one, so the mark should appear where the excess weight is.

4. Shut the engine down and install two worm-gear hose clamps around the shaft with their heads 180° from the mark you just made.

5. Rev it up to 65 to 70 mph and see if the vibration has disappeared. If it has, give it a road test to be sure. If it hasn't, rotate the clamp heads about 45° away from each other and try again. Getting better? Then keep moving the clamp heads apart until you get it right.

As you probably know, the basic function of a U-joint is to allow two shafts to operate at an angle to each other. But you may not realize that while the driving yoke rotates at a constant speed, the driven one speeds up and slows down twice during each revolution. At an angle of 4°, the variation in speed is a negligible 0.5 percent. At 10°, however, the change is 3 percent, and at 30° with the yoke going 1000 rpm, its partner goes from 866 to 1155 rpm each quarter of a revolution. You can imagine the surging vibration such an arrangement would set up.

Fortunately, this condition is eliminated by the use of two joints with the driving yokes rotated 90° from each other. As long as the angle of each joint is about the same, the acceleration and deceleration of one is canceled out by the deceleration and acceleration of the other, and a smooth transmission of power is accomplished.

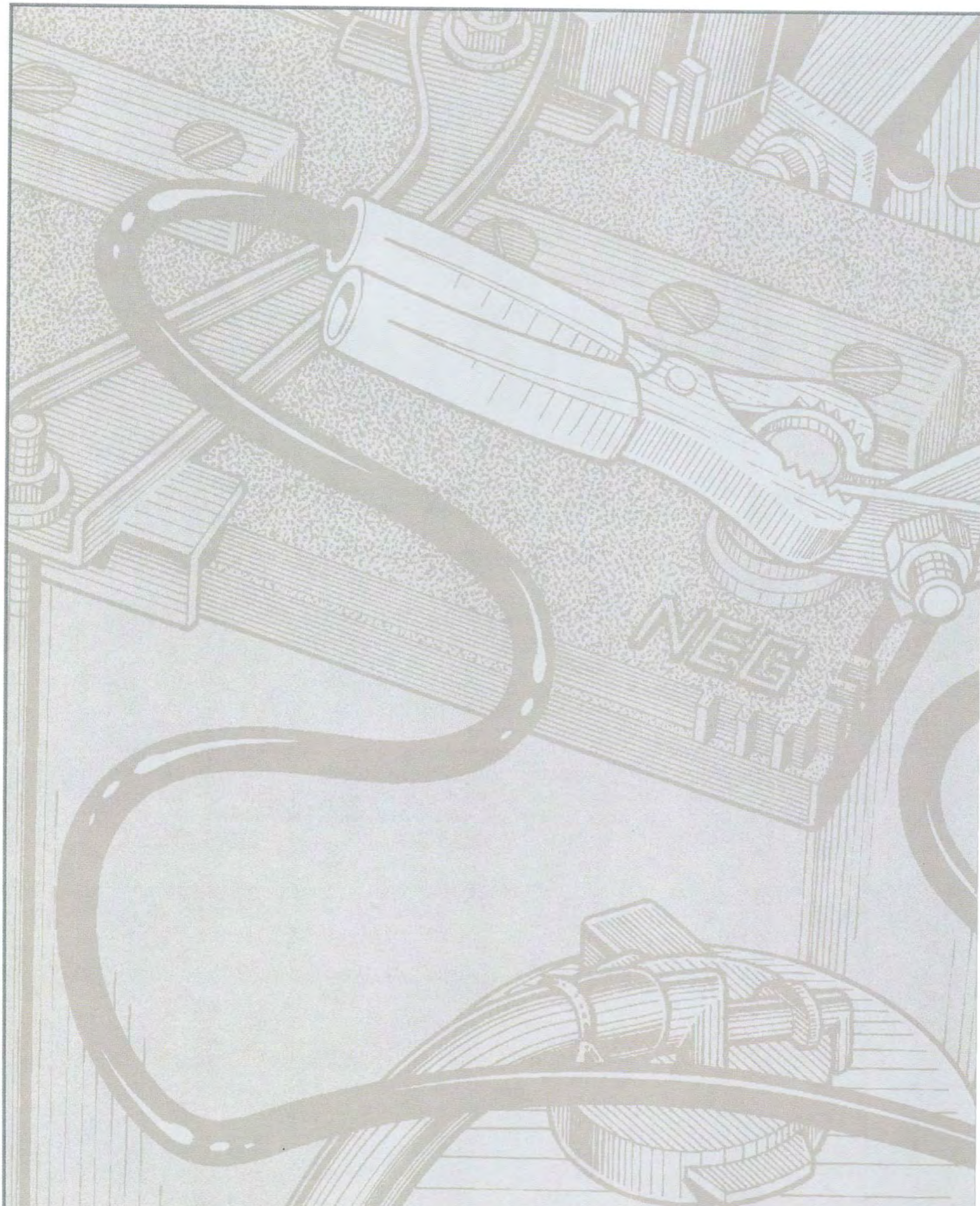

NEG

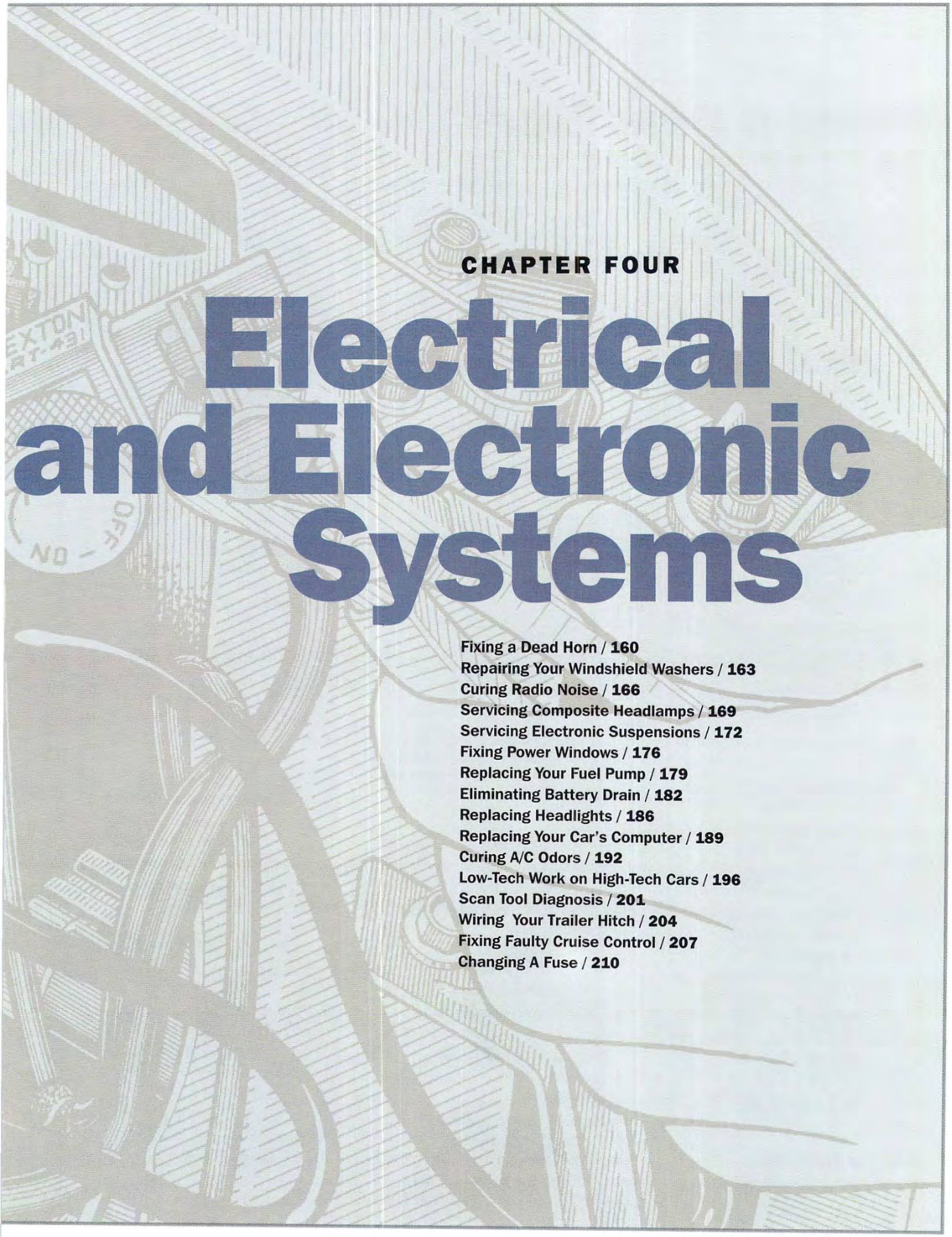

CHAPTER FOUR

Electrical and Electronic Systems

Fixing a Dead Horn / **160**
Repairing Your Windshield Washers / **163**
Curing Radio Noise / **166**
Servicing Composite Headlamps / **169**
Servicing Electronic Suspensions / **172**
Fixing Power Windows / **176**
Replacing Your Fuel Pump / **179**
Eliminating Battery Drain / **182**
Replacing Headlights / **186**
Replacing Your Car's Computer / **189**
Curing A/C Odors / **192**
Low-Tech Work on High-Tech Cars / **196**
Scan Tool Diagnosis / **201**
Wiring Your Trailer Hitch / **204**
Fixing Faulty Cruise Control / **207**
Changing A Fuse / **210**

MODERATE

Fixing a Dead Horn

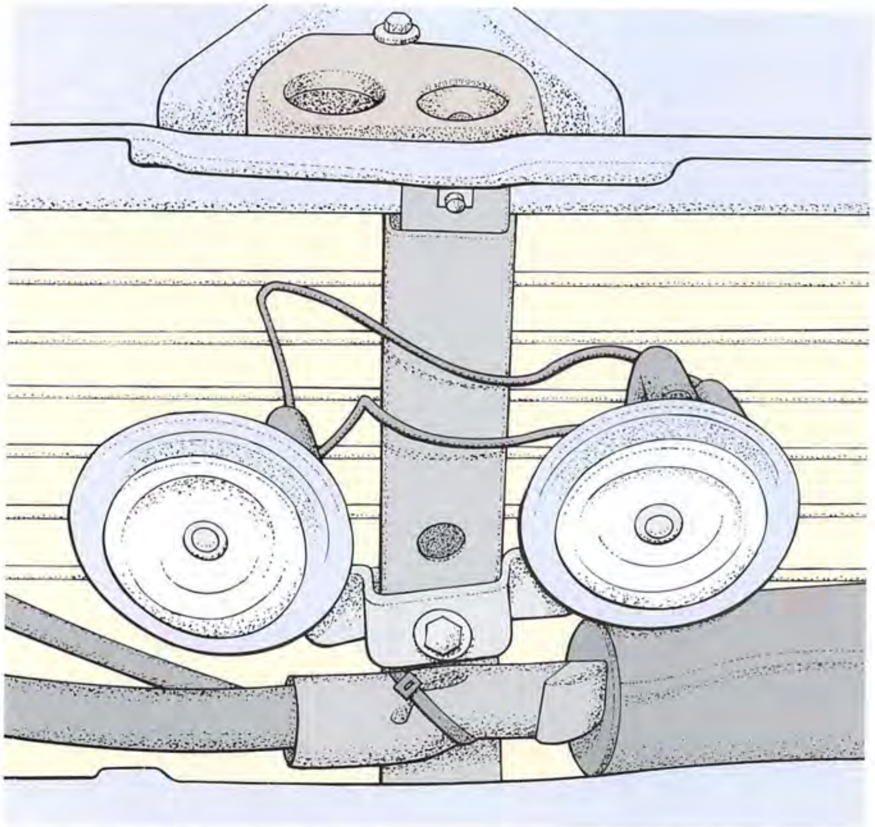

(Fig. 1) The horns are usually mounted behind the grille and in front of the radiator, although they may be to one side, higher or lower.

A flash of yellow whips through your peripheral field of vision, as a taxicab screeches into a downshift in front of you. The driver flips you a hand signal rooted in antiquity as he blithely makes a right turn from the third lane, forcing you to slam on the brakes. There's only one appropriate response—you slam your palm against the center of the steering wheel to audibly signal your displeasure with the hackie's driving, personal hygiene and relationship with his mother.

But your horn doesn't blow. A sound more akin to the utterings of a dyspeptic goat emanates from your grille for a few seconds—then silence. Your horn is broken. The moment is gone. And your mood sours even further.

This Has Got to Stop

Actually, blowing your horn as a sign of sheer displeasure is totally inappropriate. The cab driver will undoubtedly continue to drive dangerously despite any message you might have delivered. Some municipalities prohibit blowing a horn within the city limits unless there's a traffic situation that genuinely requires it—if not, you may get a ticket. But horns do have more purpose than letting your teenage son signal to his date that he's waiting in the driveway. A brief tootle can gain the attention of a pedestrian about to step off a curb without looking, and may be the best way to get the attention of that soccer mom driving a 7000-pound SUV while talking on her cellphone and looking in the rearview mirror at her whining kids.

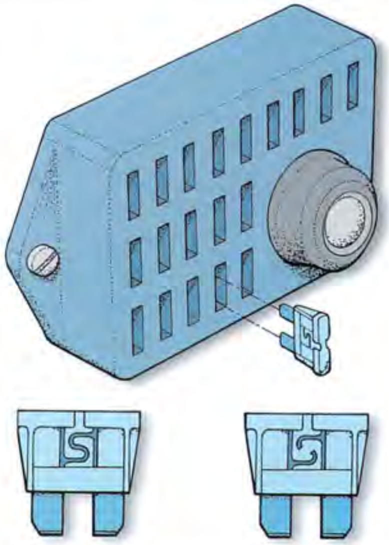

(Fig. 2) See the panel cover or manual to know which fuse protects the horn's circuits. If the S-strip is broken (fuse at right), replace the fuse.

The horn is required equipment by law—you'll need to get it fixed or you may not be able to pass inspection. A zealous police officer may even find it appropriate to write you a citation if your horn isn't functioning properly.

Less-Than-Adequate Honking

In our scenario, the horn blew for a few seconds at reduced volume. This means that the circuitry that feeds the horn with 12 volts is alive and well. Most vehicles actually have two horns, wired to sound at the same time. They usually will be of a different pitch, sounding a chord instead of a single note. Some luxury cars use three horns for a more complex, melodious sound. The horns generally are mounted to the radiator core support or somewhere behind the vehicle's grille (Fig. 1). A cursory rooting around in the engine compartment should find them.

If one or more horns are inoperative, the sound will be slightly reduced and less pleasant. If you sus-

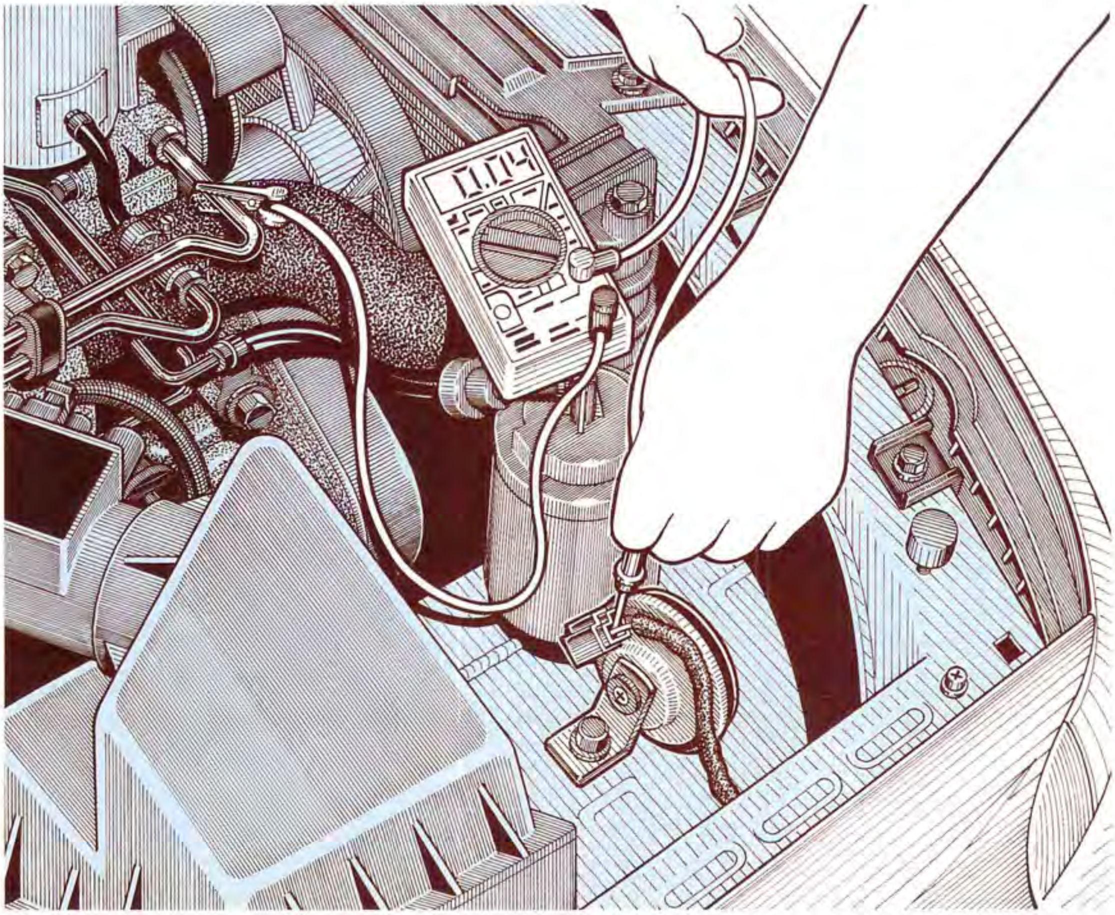

(Fig. 3) Using a simple voltmeter, back-probe the horn's connector to see if 12 volts is present when the horn button is depressed and the key is on.

pect one of your horns isn't sounding, have a helper depress the horn button briefly (key on) while you touch the suspected horn with your fingers. (We recommend earplugs, for obvious reasons.) Sound vibrations? Remove the wiring connector, clean up the spade lugs and then reinstall the connector. Still nothing? If the horn has only one wire running to it, it relies on its mounting bolt to complete the circuit to ground. Unbolt the horn, clean up any corrosion and reinstall it.

If there's still no sound, make sure that a bad horn is indeed the problem by running a jumper wire directly to the battery terminals. If one horn is defunct, replace it. Generally, the horn will have an imprint or a sticker identifying it as a High or Low tone. To preserve the OEM horn character, you'll need an exact replacement. This will ensure that the horn's frequency (pitch) will be the same. If you're not fussy, any universal replacement will work, although the mounting may not be the same and you'll have to adapt it.

Sound Off

Not so much as a squeak from either or any of the horns? Then the problem lies elsewhere in the wiring. Now's the time to check the fuse, which probably is buried in the bottom of the dashboard, the driver's kick panel or under the hood—or almost anywhere on the front half of the vehicle if it's not in one of those obvious places. Check the owner's manual for the location of the fuse box and of the specific fuse (Fig. 2). If it looks like the metal strip inside the fuse has failed, replace it with one of the same amperage rating. Once in a blue moon, fuses will seem to fail for no apparent reason. If the fuse is bad, odds are you've got a problem that will make it blow again, sooner or later.

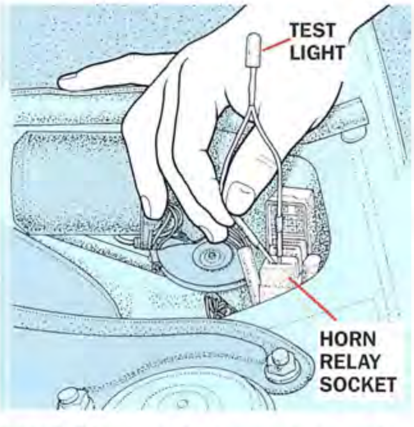

(Fig. 4) Check at the horn relay for 12 volts when the key is turned on.

The Fuse Passed Inspection

Now you'll need a 12-volt test light or a voltmeter. You also should find a shop manual or, at the very least, a schematic diagram of the horn circuit.

Start by testing for 12 volts at the connector to the horn (Fig. 3). One side of the circuit is carried by the car's metal frame, so you'll need to check between a good ground point and the wiring connector for the horn. If there's no voltage at the connector, try checking the horn relay (Fig. 4).

The horn relay switches a large current to the horns at a signal from the low-current horn button in the steering wheel (Fig. 5). It's a simple, inexpensive single-pole single-throw (SPST) relay, packaged in a small metal or plastic box with five spade-lug connectors. If you're unlucky, your vehicle uses a relay that's integrated into a larger box of sparks that also controls several other functions on your car, such as the headlights or the turn signals. This part is considerably more expensive. Sorry. Consult the shop manual for the location of the relay, as it, like the fuse box, may be hidden almost anywhere underhood or under the dash—and not necessarily near the fuse box. A simple SPST relay will have a constant supply of 12 volts to it, a lead that runs through the harness to the horn, another lead that runs to the horn switch in the wheel, and a ground. Check that voltage is coming into the relay at two terminals and leaving it at two other terminals when the switch is pushed. If it's not, try grounding the lead that goes to the horn button to make the relay pull in. If the horn sounds, the relay is good but the horn button or its wiring is bad. You may be able to simply replace the relay if it's the problem. Otherwise, you'll need to do some serious detective work to chase down the open circuit.

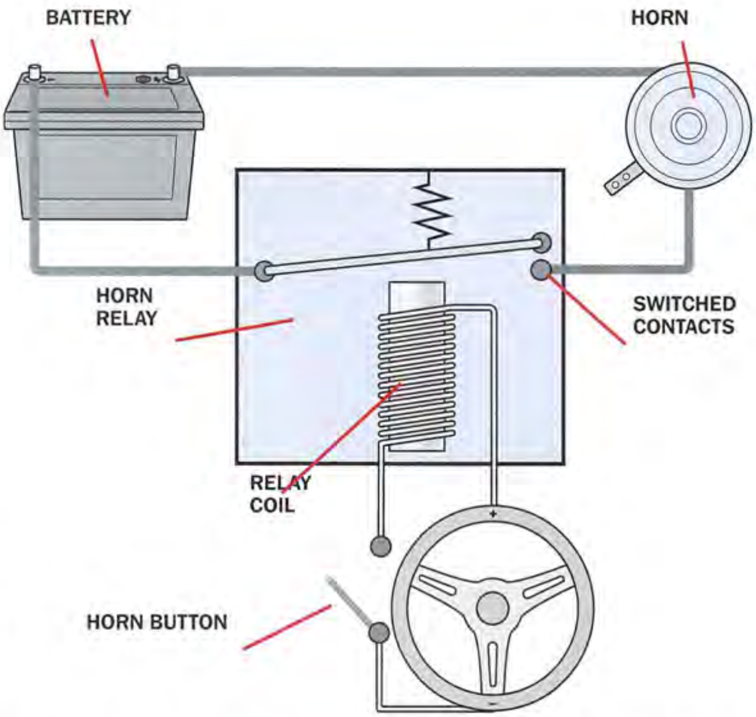

(Fig. 5) The horn relay steps up the low current from the horn button to the higher current needed by the horns.

Button, Button, Who's Got The Button?

If you have a "panic" button on your keyfob/remote, you can check to determine if the horn circuit works. If it sounds the horn, but the horn doesn't work when you press the button, the problem is in the horn button or its wiring to the relay. Every late-model car has an airbag camped in the middle of the steering wheel. This airbag, if improperly handled, can expand with lethal force. Even the static from dragging your clothing across the seat can set it off. And the horn button (or buttons) is generally attached to the top of the airbag shroud. That means you should get a trained technician to remove the airbag for you and then reinstall it after your horn has been fixed. He can test the airbag system to be sure it will go off when it's supposed to and not go off when it's not. The airbag must be removed to access the slip rings or clock spring that carries voltage up the rotating steering column from the relay to the horn button (Fig. 6). If you need to access this, we recommend you leave the job to a service technician who has training and service manuals detailing the correct procedures. Incidentally, a horn that intermittently blows on its own generally is caused by a bad clock spring that shorts out to ground randomly. Older car "budget" alternative: Clip the horn button wire near the relay, splice an auxiliary wire to it (and wrap the splice with electrical tape). Run the auxiliary wire to an aftermarket universal horn button you can mount on the dash or console. It's not eye-appealing and you do have to "learn" the location, but it costs just a few dollars.

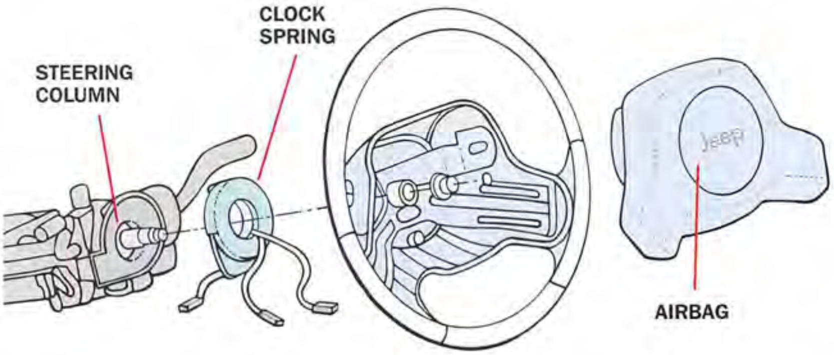

(Fig. 6) The clock spring allows the wheel to turn in circles while still making an electrical connection to the car's wiring system.

EASY

Repairing Your Windshield Washers

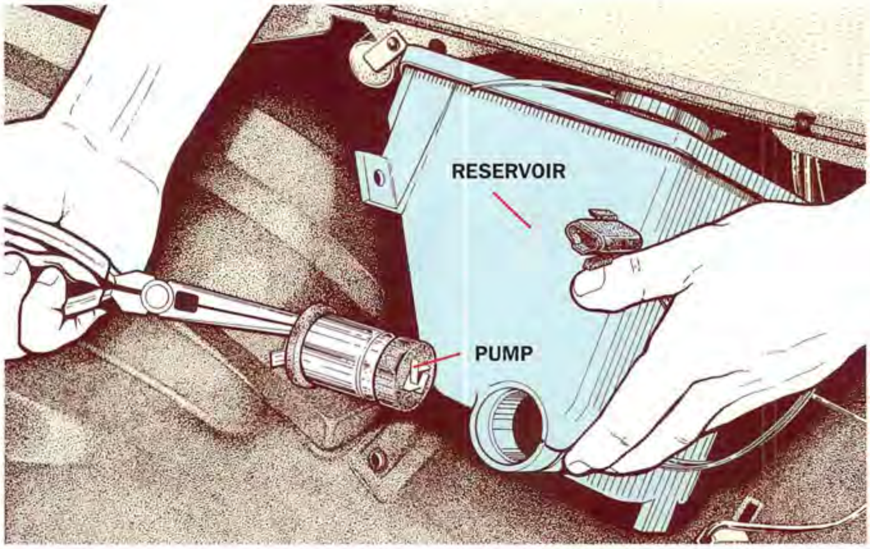

(Fig. 1) Replacing a nonworking pump is simple once the reservoir is empty.

It's a light drizzle—barely enough to wet the windshield. So you press the washer button and turn on the wipers. There's nothing—or maybe just a weak squirt—out of the washer nozzles, and the wipers quickly slow down and just streak the windshield to cover up whatever clear areas there had been. Traffic becomes a blur, and you're forced to slow down, pick your way through a slalom of cars and trucks to a side street, and just shake for a few moments until you have the presence of mind to get out and clean the window by hand.

Where's the Liquid?

You just filled the reservoir last week and haven't used the washers since. So the reservoir should have plenty of fluid, right? It was a washer antifreeze solution? Okay, is the reservoir still full? It's empty? There must be a leak. Even if that solution contains a dye, you still are likely to have a tough time finding a cracked seam in the reservoir, particularly if the reservoir is under the battery or in a fender well, the late-model location of choice for space-deprived powertrain compartment engineers.

Fill the reservoir and run your fingers along any seam, and also around the pump seal, which may be a simple rubber grommet. If you're losing fluid at a noticeable rate, you should be able to find a wet spot pretty fast. If it's a seam, empty the reservoir, dry the exterior and apply a bead of windshield sealer or epoxy. If it's the pump grommet, pull the pump out (just work it side-to-side or up-and-down with an outward tugging motion). Replacement grommets are not necessarily easy to buy, so if you can't get one, it's worth trying to apply beads of RTV silicone sealer around the pump opening in the reservoir and around the barbs of the pump. Allow adequate time for the RTV silicone to "skin" (partly cure) so it doesn't extrude out of position when you reinstall the grommet and pump.

The reservoir is full? Next question: Is the pump working? Have a helper press the washer button while you listen at the pump. If you hear it working, disconnect a hose that runs from the pump, aim it into a container and repeat the test. If there's a solid stream, repeat the test at the hose connection to a spray nozzle. If a car has two nozzles, it's unlikely that both nozzles would plug up simultaneously, but it is possible that the hose from the reservoir is kinked, that a check valve (installed in some lines) has closed up or an inline filter is plugged. So if you get a solid stream from the pump, but nothing at a hose connection, work your way back to the pump. If all you can find is a check valve or an inline filter, remove it and make a temporary splice with a plastic hose connector. You may find the washer system operates well enough to make that your permanent repair.

If the pump isn't working, either it has failed or there's a break in the circuit leading to it. Unplug the wiring connector and again have your helper press the washer button while you probe the terminals with a grounded 12-volt test light. It should go on with the probe in one of the two terminals, and when you identify that one (the current feed), try to jump across the two terminals with the test light. Again it should go on. If you can't get the light to go on in the first test, there's a break in the current-feed side of the circuit from the washer button, and you'll have to trace it from there. If you can't get the test light to go on when you jump across the two terminals of the connector, there's a bad ground.

The connector passes both tests? The pump is the problem. On most late models, you can just pull it out (the rubber grommet may come out with it).

On many older cars, where little more than the end of the pump sticks out, look for a serrated retainer you can pry out with a thin-blade screwdriver. Then pull the pump itself with slip-joint pliers on an edge of the electrical terminal (**Fig. 1**).

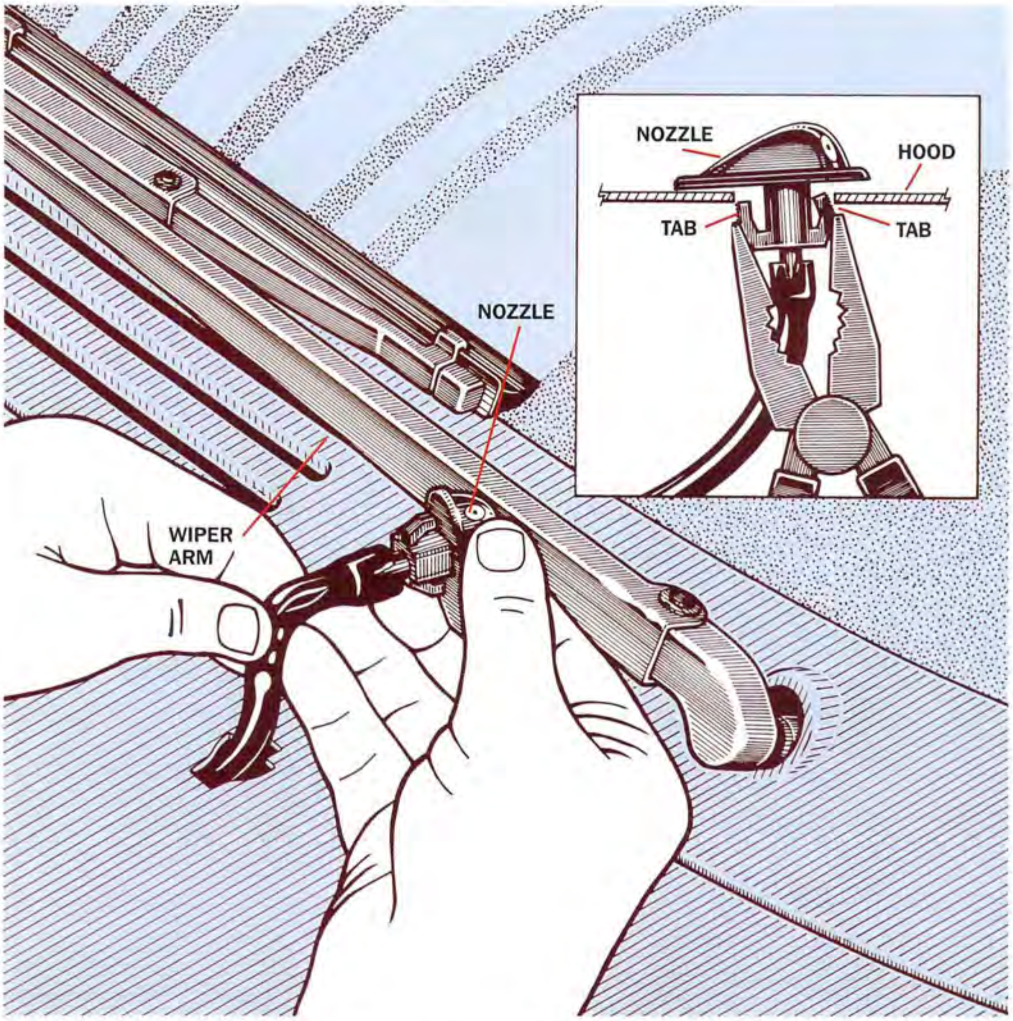

(Fig. 2) Replacing a damaged hose or nozzle is simple: Squeeze the plastic tabs together with pliers and the nozzle should pop out of the hood.

Just a weak squirt from the nozzles? If there's a solid stream from the hose connections, the nozzles apparently are plugged. The easiest approach is to just replace them, if you can get new ones. Most nozzles are held by retaining screws or built-in squeeze clips. Some nozzles have just a single-side clip (press toward the nozzle and push the nozzle out). Others have clips on both sides of the nozzle (squeeze together and push out the nozzle). If the opening is large enough, you may reach in with fingers to squeeze the clip(s). Otherwise, use needle-nose pliers **(Fig. 2)**.

Can't get new nozzles? On some old cars you could work in a needle to clear out the passage, but today, few nozzles have passages straight enough to do that. And if the car has a single, centrally located nozzle, it's probably a "fluidic nozzle." It seems to apply a super-wide-angle spray to the windshield, but actually it produces an oscillating stream that is so fast it deceives your eye **(Fig. 3)**. There are no moving parts within the fluidic nozzle, but it does contain a complex set of passages that produces the oscillation that swings the spray from one side to the other.

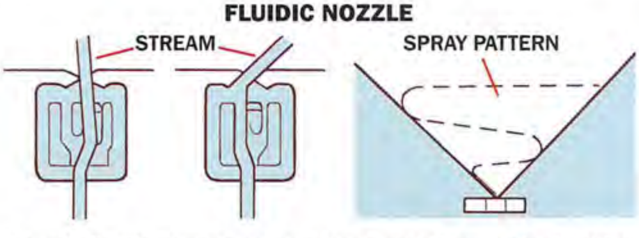

(Fig. 3) Fluidic nozzles cover a large area from a single point.

Soak the clogged nozzles in vinegar, at least overnight, to dissolve deposits, then blow compressed air through the nozzles, and that should restore the spray pattern.

Nozzle Aim

If the spray doesn't go where you want it to, it may be possible to make an adjustment. Typically, if the mounting pad is round, and the nozzle head has two or three spray holes, it can be adjusted side-to-side, up-or-down or both, by inserting a thin probe tool into a spray hole **(Fig. 4)**. Some nozzles have a rectangular hole between two spray holes, meant specifically to accommodate a probe for adjustment **(Fig. 5)**.

On older cars, the nozzle may be a simple piece of metal tubing (with a calibrated orifice inside), held by a retaining screw. With a pair of needle-nose pliers, make a very tiny bend, at a joint if there is one, near the base if there isn't.

You can't reshape plastic, and most nozzles are rectangular and fit into a position that defies change. If it's held by a screw, you may be able to slip a thin piece of plastic or aluminum under the fore or aft end of the nozzle to lower or raise the spray tip, so the spray hits the windshield higher or lower. However, if it's retained by plastic clips, you can't do this, or the clips won't seat. If a nozzle starts to spray off the usual pattern, that's usually the result of front-end bodywork that included hood adjustment, and it may take some readjustment to correct.

Nozzles in Wiper Blades

When the washer nozzles are in the wiper blades, a damaged hose can allow leakage that weakens the spray. Or the hose may have become kinked from careless cleaning of the windshield. The nozzles are never adjustable, so if you can't clean them, you'll have to pop in replacements.

Rear Systems

Some vehicles, particularly wagons, minivans and sport utilities, are equipped with rear washers, which often means a second pump and dual-chamber reservoir, plus a long hose to the rear to check out. If there are dual sections, most are designed so the rear drains into the front when the front is nearly empty. So the fact that you haven't used the rear, and it now doesn't work, could mean a problem with the front system, particularly a leak from the front chamber or a front circuit hose.

There's also a second washer pump, so be sure to check the right one. Washer system service isn't complex, but it may require working in tight, dark quarters. So use a bright droplight and spend the extra minutes to do it right.

Inside a Fluidic Nozzle

The single fluidic nozzle seems to apply a wide-angle fan spray across the windshield, but actually it's an oscillating stream. Fluidics is the name of the operating principle. The internal passages of the nozzle are shaped so the incoming fluid goes through a main passage, sort of "leaning" against one wall (a deflector in the main passage "biases" the nozzle to start the flow to one side). As it flows, some of the fluid is diverted through an opening on that wall into a feedback passage that produces a second fluid flow against the incoming flow at its entry point. The pressure of the second flow deflects the incoming flow to the opposite side wall of the main passage. There's also a feedback passage on this side, and again the second fluid flow deflects the main flow back to the opposite wall. This goes on so fast that it appears to be a spray. But because it's an oscillating stream, the flow from one centrally located nozzle covers the entire windshield.

Fluidics works on airflows as well, and the same principle was used in the defroster nozzles of the 1990-96 General Motors minivans, where two standard-size defroster nozzles with normal airflow could cover the large windshield.

HOW IT WORKS

A Washer System

When you operate the washer button, you complete a circuit to the system's electric pump, typically mounted in the bottom of the reservoir.

The pump draws fluid from the reservoir and pushes it through a hose to one or two nozzles in front of the windshield, typically in the hood or the cowl area just to the rear of the hood. On vehicles with a rear window washer, there's a second pump with its own hose, and generally a single nozzle at the rear. Some rear washers have their own reservoirs as well.

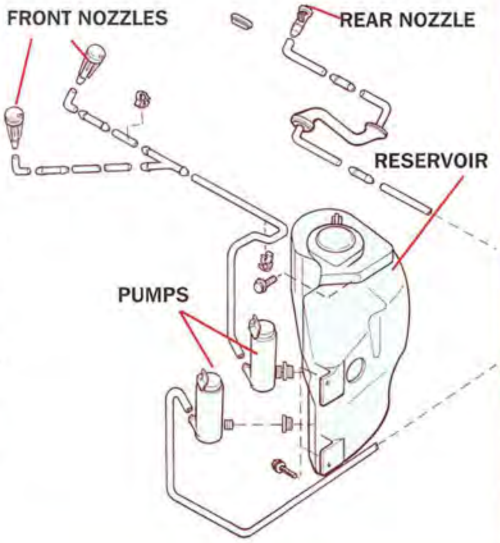

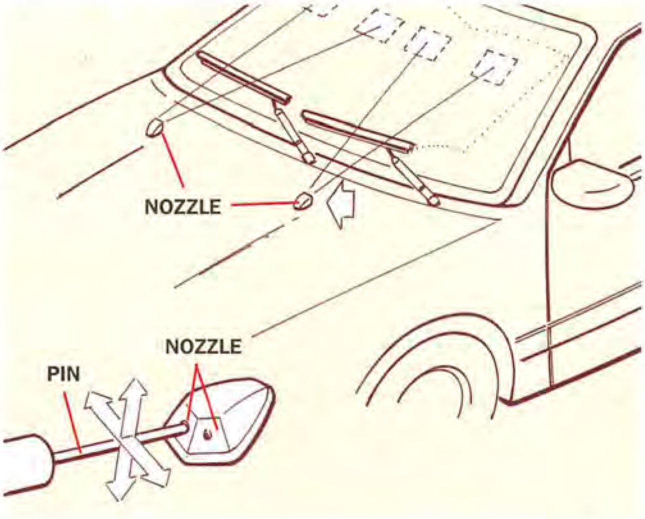

(Fig. 4) Some nozzles can be aimed with a straight pin inserted in the hole.

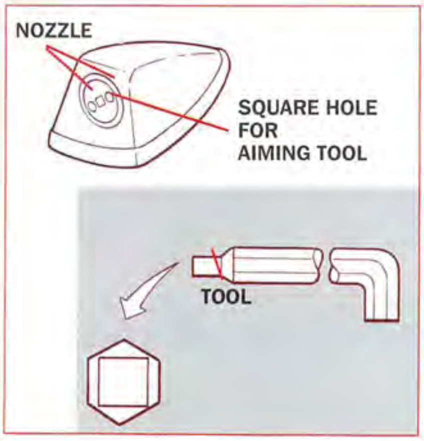

(Fig. 5) Other styles of nozzles provide a receptacle for a special aiming tool.

Curing Radio Noise

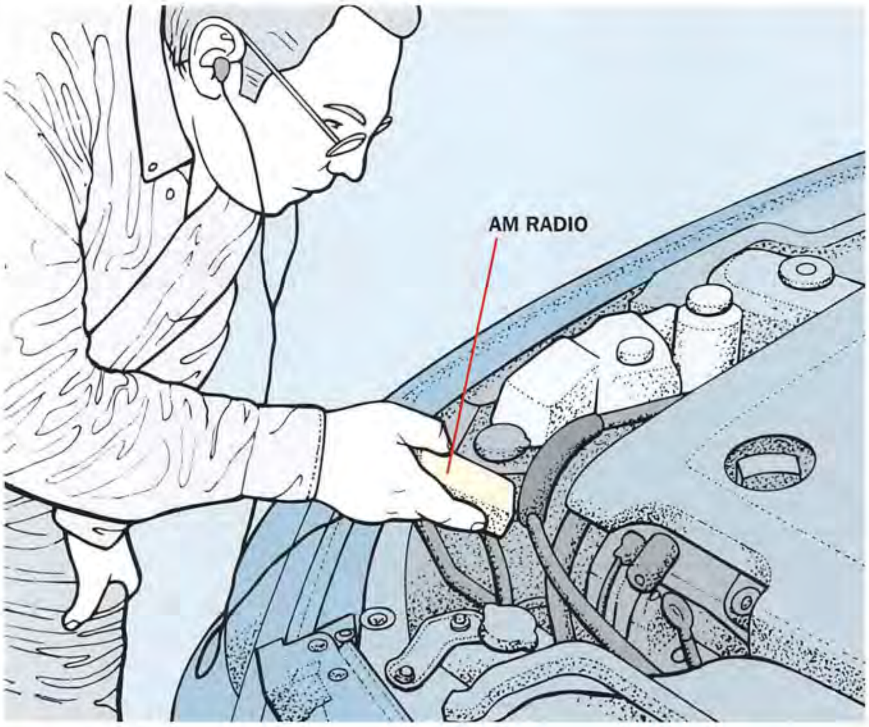

(Fig. 1) A cheap AM radio makes a good Geiger counter-style device for hunting down RFI. Here, we're looking for an alternator with a bad diode. Mind your fingers.

You're looking forward to the chance to listen to the ballgame, having just dropped your purple-haired teenage daughter off at her first semester at college. A lot of ballgames, actually, now that you don't have to listen to her headbanger music, with its shrieking vocals and droning guitars. But instead of the first inning, you hear what sounds like some awful guitar note—and it changes pitch as you speed up and slow down, turning to a maniacal rhythmic clicking as you wait for a traffic light to change. There's no tape in the player. What you have here is radio frequency interference (RFI).

Finding the Source

There are three classes of RF noise—constant, intermittent and engine-speed following. Our hapless father's problem was in the last category—his noise varied up and down in pitch and volume as his engine sped up and slowed down. This type of noise is caused by something that varies its speed with the engine. Likely culprits include the ignition, the alternator or even a fuel injector. Constant-speed noises are usually caused by an electric motor—most likely the electric fuel pump found in the tank of most modern vehicles which runs at a constant speed anytime the engine is running. An electric fan motor will also run at a constant speed—until you change the fan setting or turn it off. Intermittent noises are easier to associate with a source, such as an electric seat adjuster or a power-window motor. In other words, even though the noise comes from a radio speaker, it may be caused by any manner of device anywhere on your vehicle.

Constants

The one motor that's guaranteed to be on whenever the car is running is the fuel pump's, and unfortunately it's usually buried inside the fuel tank. Here's one sure way to tell. Turn the key to the Run position without starting the car. The pump should run for 2 or 3 seconds. Then, when the computer senses that the engine isn't running, it will shut off the pump to prevent spilling fuel and draining the battery. Other almost-constants are heater fans, wipers and electric radiator-cooling fans.

Engine-Speed Noises

Candidates here include the alternator/voltage regulator, a faulty fuel injector and the ignition system.

Geiger-Counter Test

Here's our favorite low-tech tool for hunting down and killing wild noises—a cheap AM radio (Fig. 1). Tune it to an empty channel around 1400 kHz, crank up the volume in your headset, and use the radio to sniff out the noise. These cheapo radios use a ferrite-bar antenna that has good reception along its side, but poor reception along its length. Once

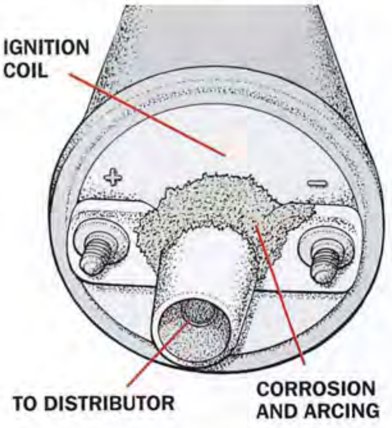

(Fig. 2) Carbon tracking in the ignition system is a noise source.

(Fig. 3) Be sure both ends of the antenna are properly grounded and corrosion-free.

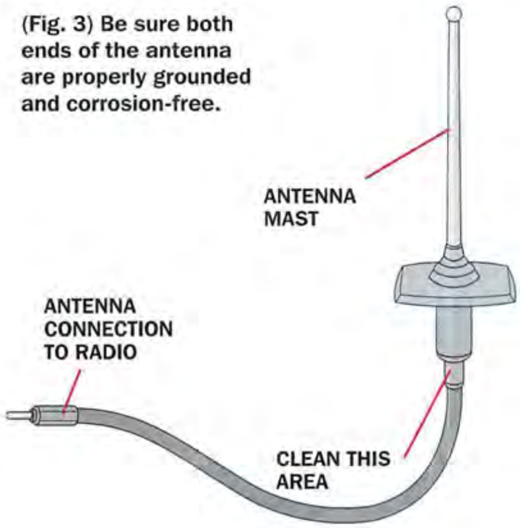

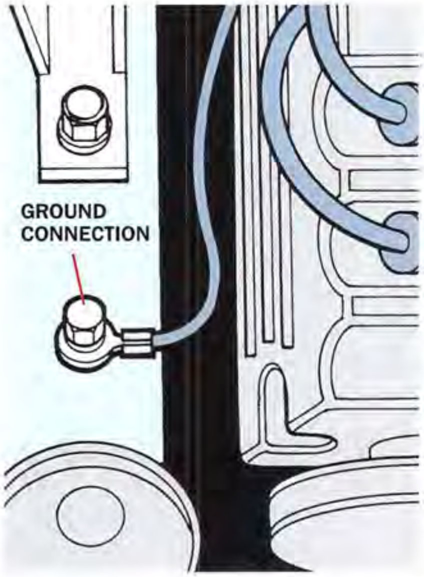

(Fig. 4) Check grounding points by loosening, cleaning and tightening.

you find the noise within a few feet, turn the radio 90° to minimize the noise. The top of the radio will point at the source like a gunsight.

Making Sure

Find some way of disconnecting the source of the RFI and check to see if the noise ceases. This will be difficult in the case of the fuel pump or ignition, but you can pull the belt off the alternator. (Don't disconnect the alternator electrically—the back EMF [voltage] may smoke the diodes.) If you think it's a fuel injector, try disconnecting it from the harness.

Plugged Up

Virtually all cars today use resistor-type plug cables if they use spark plug wires at all. If your vehicle is more than a few years old, degraded cables may be the source of the problem. Remove and replace the cables one at a time, clean them of grease and dirt with mild detergent, and check the connections to the ends. Now get out your ohmmeter and measure the cables' resistance along their length—they should measure approximately 10,000 ohms per foot. Resistance in the megohms or single digits may well be the source of not only your RFI, but a nagging misfire. Replace any suspect wires with factory or high-quality aftermarket wires. Check the plugs, coil and any distributor for evidence of carbon tracking or arcing as well (Fig. 2).

Intermittents

These noises are easy to associate with the source. Any RFI that sounds only when one window is going up or down will be easy to blame on the window motor.

Which Way Home?

Now that you've pinpointed the noise, what can be done about it? It depends on whether your RFI source is broadcasting or cabling the noise to your radio. This matters because the cure is different. Try pulling your radio's antenna lead out. If the noise goes away or is substantially quieter, it's coming in through the antenna. If the noise stays the same or gets louder, it's coming in along the 12-volt power cables.

Try removing the antenna from the fender and cleaning the fender sheetmetal and the antenna mount. Clean down to the shiny metal, using sandpaper to remove any corrosion (Fig. 3). Smear the area lightly with Vaseline and reattach. This will provide a proper ground at the antenna. Be sure the radio chassis is properly grounded to the car body (Fig. 4). Aftermarket installations are more likely to have a radio that's grounded only by the shield in the antenna co-ax. A simple wire added between the component's metal case and the car's sheetmetal will often eliminate any RFI.

There are two ways to reduce noise: by using an inductance in a power cable to keep the high-frequency noise from traveling, or by using a capacitor to shunt it off to ground harmlessly. Sometimes both ploys are necessary. In fact, most of the electrical motors in your vehicle use some sort of capacitor for noise suppression. Any good car stereo shop or Radio Shack will have the parts you need. The noise filter we show below is one example (Fig. 5). It has a large inductance in series with the 12-volt power cable to a buzzing aftermarket stereo, as well as a couple of small capacitors in parallel. The inductance prevents noise from entering the amplifier through the power leads, and the capacitors bypass any leftover noise.

Electric motors, as we mentioned, often have a capacitor in parallel with the armature for the purpose of reducing RFI. If the motor's brushes are worn and sparking, the noise may overpower the capacitor's filtering. Most automotive electric motors are not serviceable, and they'll have to be

(Fig. 5) This RFI filter goes in line with both power and ground connections.

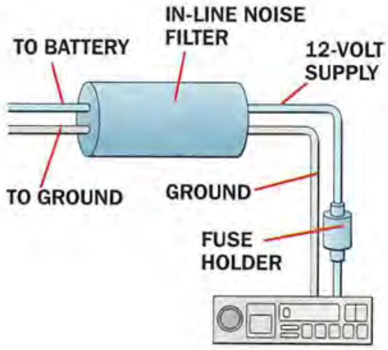

replaced if a simple filter doesn't quiet them.

As we said earlier, fuel pumps are difficult to get to—in most modern cars they are mounted inside the fuel tank. To access the pump, or even the wiring that connects to the pump, it's necessary to drop the tank out from under the car, which is a lengthy, messy and potentially dangerous job. Be careful if you attempt to add a filter to the external tank wiring, as the filter itself is large and bulky. You'll need to securely mount it to the top of the tank to keep it from breaking loose over potholes and bumps.

Soldering

One of the most common sources of RFI is poor connections. If you find a poor, corroded connection or a loose electrical joint, don't just crimp the connector tighter onto the wire. Remove the connector and clean the wire. It may be necessary to trim the wire back a few inches to get past any corrosion inside the insulation. There's only one acceptable technique for splicing wires if you're having RFI issues—and the common automotive-style crimp connector isn't it. You'll need a good, clean, high-wattage soldering pencil or gun, rosin-core 60-40 solder and PVC shrink tube.

Start by using proper, stranded automotive-grade wire. Unstranded household wire will fracture and eventually break. Start with wire of at least the same diameter as the wire you're adding to. If you're adding a power cable, a ground connection or a filter, use 12-ga. wire for ultralow resistance. Strip and pre-tin both wires to allow you to twist them together for a sound mechanical connection. Slip a length of PVC shrink tube over one wire, twist together, and solder. Use enough heat and a sparing amount of solder to make a shiny, wet-looking solder joint. Allow the joint to cool without disturbing it. This will prevent the liquid solder from crystallizing as it cools.

Cover the joint with the PVC tubing and shrink the tubing with a heat gun, or carefully with your lighter if you must. If the solder joint you're making will be exposed to the elements, use shrink tube that has waterproofing adhesive inside it to keep corrosion from creeping into your fresh joint. Silicone seal or liquid electrical tape is a good alternative.

Dress any new wires carefully to prevent chafing on corners. Support any components you've added. Remember that even a foot or two of heavy-gauge wire can flex itself to the breaking point if it's unsupported.

THIS STEREO POWER amp is getting an in-line noise filter added to its power lead. Solder and PVC shrink tube make for a noise-free, high-current splice.

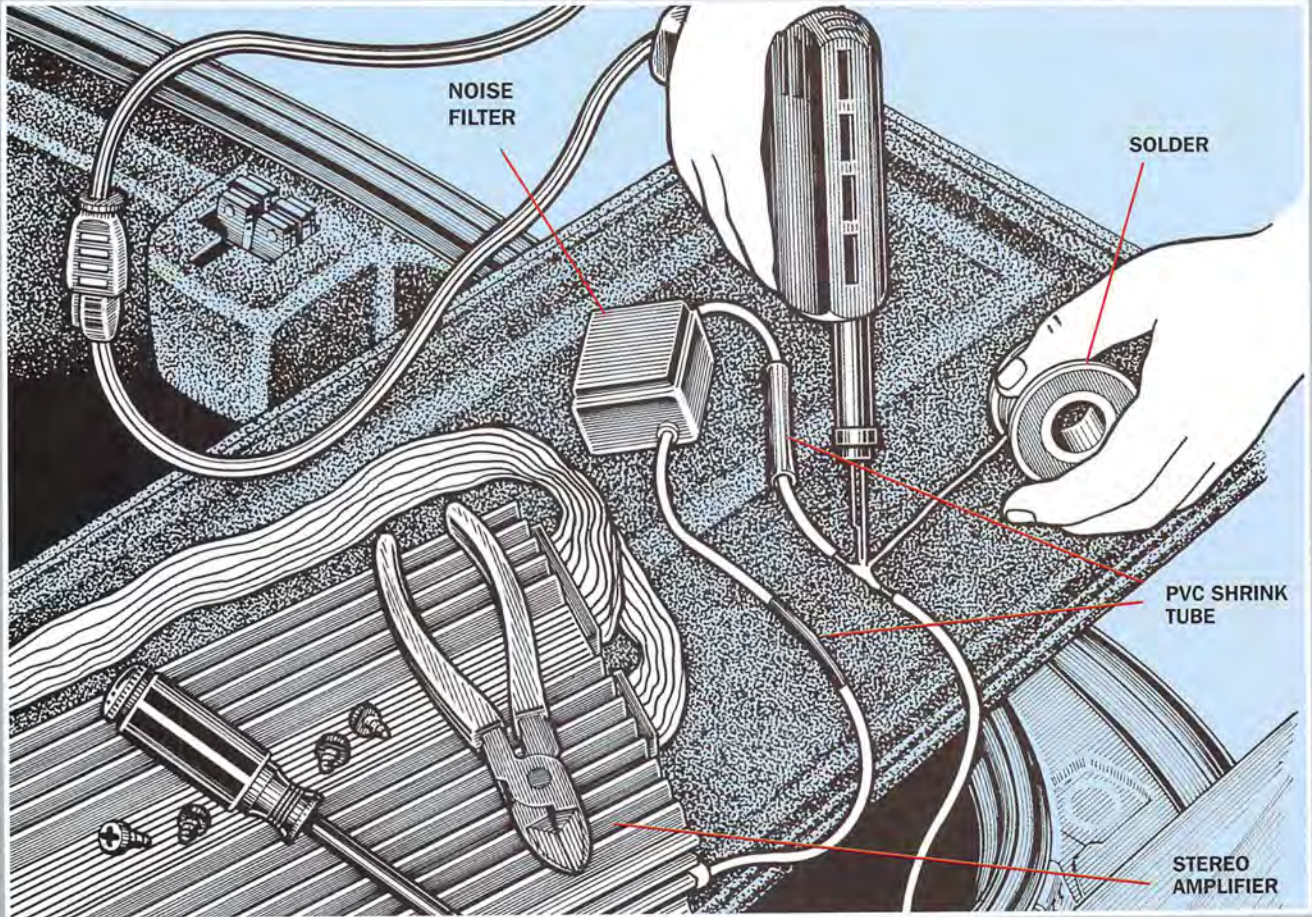

Servicing Composite Headlamps

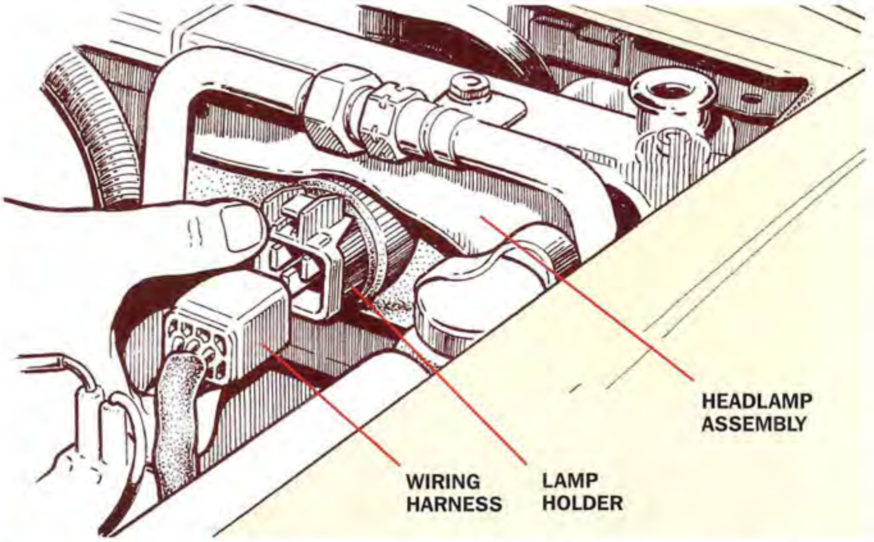

(Fig. 1) The first step to replacing a lamp is to disconnect the wiring harness from the lamp holder. Never touch the glass on the replacement bulb with your fingers.

It's your first chance in months to get out of Dodge. It'll be great to pack up some cold ones, a couple of fishing rods and some old clothes and escape to the country for a day or two. But as you leave the mercury-vapor-illuminated metroplex, you realize what you forgot—carrots, lots and lots of carrots, because you can barely see the road in front of you in the starlight of the countryside. Relax, it's not your failing eyesight. It's a burned-out headlamp.

Older Ways

For generations, American cars, and any car sold in the United States, had the same kind of headlight—a sealed-beam, either in a single or a quad arrangement. This fragile blown-glass envelope was filled with an inert gas and worked pretty well until it burned out. It had only modest performance, but the Department of Transportation mandated its use.

Most modern cars use what's called a composite headlamp—a plastic reflector bonded to a plastic or glass lens and fitted with a bulb. The bulb is of a quartz-halogen design. The "glass" bulb is actually made of silica quartz, which is highly resistant to heat. The filament is engineered to run at a much higher temperature, producing more light and heat. The silica envelope is filled with a mixture of halogen gases (iodine or bromine) to scavenge evaporated tungsten filament from the inside of the quartz, keeping each bulb's brightness constant until it fails.

Simply replacing a bulb is easy. A socket holds the bulb in place at the back of the reflector assembly. Unscrew it and pluck the bulb out. New quartz bulbs are always packaged in a bag or sleeve. This is to prevent the oils on your fingers from contacting the quartz. Temperatures at the surface of the bulb are high enough—several hundred degrees—that the oils will carbonize onto the surface, creating a localized hot spot. This hot spot will overstress the silica, resulting in a crack that lets air into the bulb, blowing the filament out within minutes. Never allow your fingers, or anything that isn't squeaky clean and dry, to touch the bulb **(Fig. 1)**. If you do, clean the bulb with alcohol and a clean cloth. The easiest way to avoid problems is to leave the protective sleeve on until the bulb is in the socket and ready to reinstall. When reinstalling the lamp holder, a small smear of silicone grease over the O-rings will help it slide back in and keep moisture away. For more detail on this work, read "Replacing Headlights" on page 166.

Foggy, Foggy Nights

Both of your headlights are lighting up but you still can't see? Are the sur-

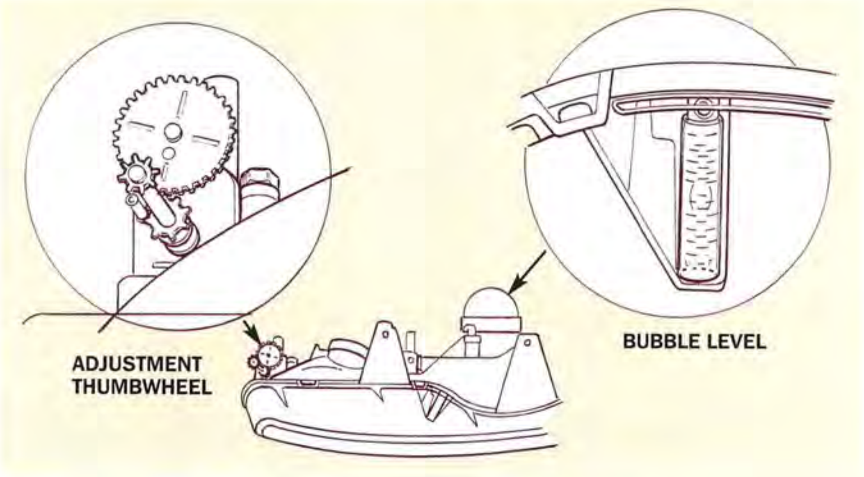

(Fig. 2) Many modern headlamp assemblies have a bubble level to assist in aiming the light beam. You'll still need a marked wall to see the beam pattern.

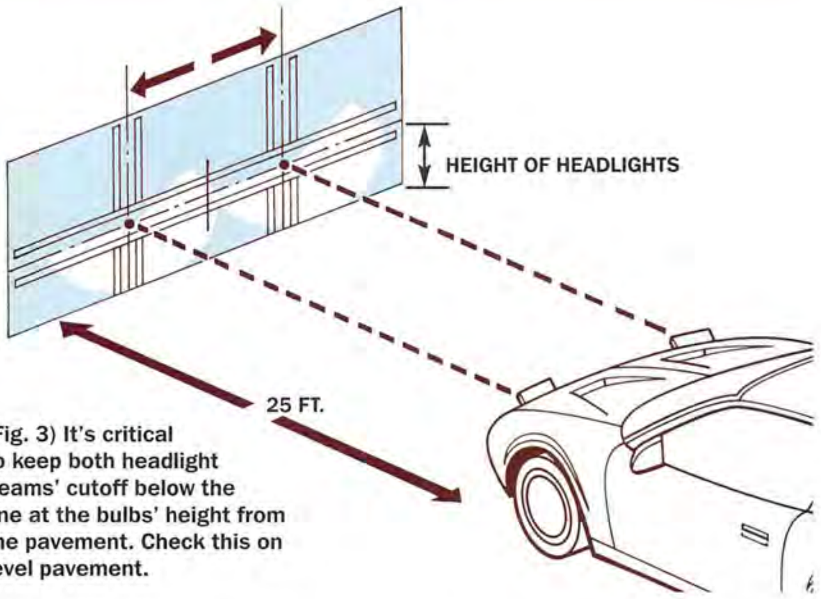

(Fig. 3) It's critical to keep both headlight beams' cutoff below the line at the bulbs' height from the pavement. Check this on level pavement.

faces of your headlamps fogged? Plastic lenses are covered with a special UV-resistant coating. After years of exposure to pollution and UV-containing sunlight, it can fog. Using rubbing compound to remove the haze is a short-term solution. With the coating polished off, the lens will yellow and haze. Your only solution is to replace the entire assembly.

If moisture has crept into the assembly and fogged the interior, you may have a problem with the housing's vent system. Look for collapsed vent hoses, or hoses plugged with mud, insects or rustproofing.

If the vehicle has been immersed in muddy water, you'll need to remove the entire housing and flush it out. Dry thoroughly inside and out before reassembly. Corroded plating on the inside of the reflector is grounds for replacement.

Aim High

Be sure to set the aim correctly. Old-style sealed-beams were not fussy about aiming, but the modern quartz lamp in composite headlamps has a very sharp horizontal cutoff to keep light out of the face of oncoming traffic. Consequently, the aiming of the beams is critical.

Sealed-beam lamps used a relatively crude aiming mechanism that required a Phillips-head screwdriver to turn adjusting screws, which invariably got more difficult to turn as the socket aged and corroded. Sooner or later, the heads of the screws stripped, and you aimed your headlights poorly or not at all.

Some headlamp assemblies are adjustable by means of a large Phillips-head screw accessible from the front of the vehicle through holes

HOW IT WORKS

High Intensity Discharge Lighting

Some high-end vehicles are available with an extremely bright, tightly focused type of lighting known as High Intensity Discharge (HID). Unlike conventional lighting, there is no filament to burn out, as the light is generated by incandescent gases in a quartz tube. How hot is it? Hot enough to create a plasma of the molecules by stripping the electrons away from their nuclei. This requires, at least initially, nearly 20,000 volts to discharge across electrodes in the bulb. The plasma envelope's shape is easier to focus than a springy tungsten wire, so less stray light goes into oncoming traffic's windshields and more, far more, goes onto the verge of the roadway. There's nothing to burn out, so the lamp should outlast the vehicle. Aftermarket retrofit kits are available to upgrade your vehicle. We've installed a Xenarc low-beam kit from Sylvania on one of our vehicles, and it has given us a new appreciation for the number of deer browsing near the side of the road at night.

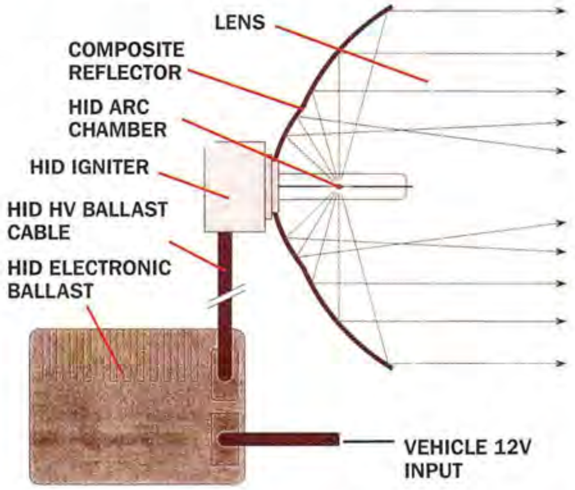

Just remember that correctly aiming these types of lights is far more critical than conventional lamps, because their intense light can potentially blind oncoming traffic if they're aimed too high. Warning: Before you replace a bulb with this system, be sure the switch is off, battery is disconnected, and ballast connectors are unplugged. High voltage is hazardous.

in the grille. Others may have thumbwheels that you simply crank up and down and left and right without any tools **(Fig. 2)**. Vehicles vary, so you'll need to check your owner's manual or the shop manual for specifics—but here are the generic instructions.

Park your car in front of a light-colored wall in a dark spot indoors, or outdoors at night. Position the car so that the headlights are 25 ft. from the wall, and be sure the car is parked at right angles to the wall. Measure the height of the center of the headlights from the ground. Make a line on the wall with a marker or masking tape at this same height **(Fig. 3)**. Now mark a pair of vertical lines directly in front of each headlight. With the adjustment thumbwheel, you can now slue the beam up or down and left or right. If you have trouble visualizing one beam because the other obscures its light, try pulling that lamp's connector, or just covering it with your jacket.

The specific numbers vary with each vehicle manufacturer, but at the very least, the left and right beams should be the same with respect to their individual centerlines. The cutoff line should be just below the line at the headlamps' center. The kickup to the right of the beam should start just to the right of the centerline. Do all the adjusting with a trunkful of luggage, a tankful of gas and a warm body in the driver's seat.

Many late-model cars incorporate a small bubble level directly into the headlamp housing **(Fig. 2)**. Observing the level will help you to make adjustments when you are aiming the light beam. Remember that this is only an initial adjustment. You'll still need to visualize the beam on a wall to trim out the correct alignment. This is because manufacturing tolerances don't always place the filament in the lamp in exactly the same position relative to the lamp's metal base—which can make the beam's alignment quite different when the lamp is replaced. The level will let you make headlamp adjustments when you have to drive a heavily loaded vehicle. Check the settings on the bubble levels, load up the trunk, reset the bubble levels and go. Don't forget to raise the beam after you unload.

REPLACING A BROKEN HEADLAMP assembly is straightforward. Most of the fasteners and mounting hardware will have to be transferred to the new housing.

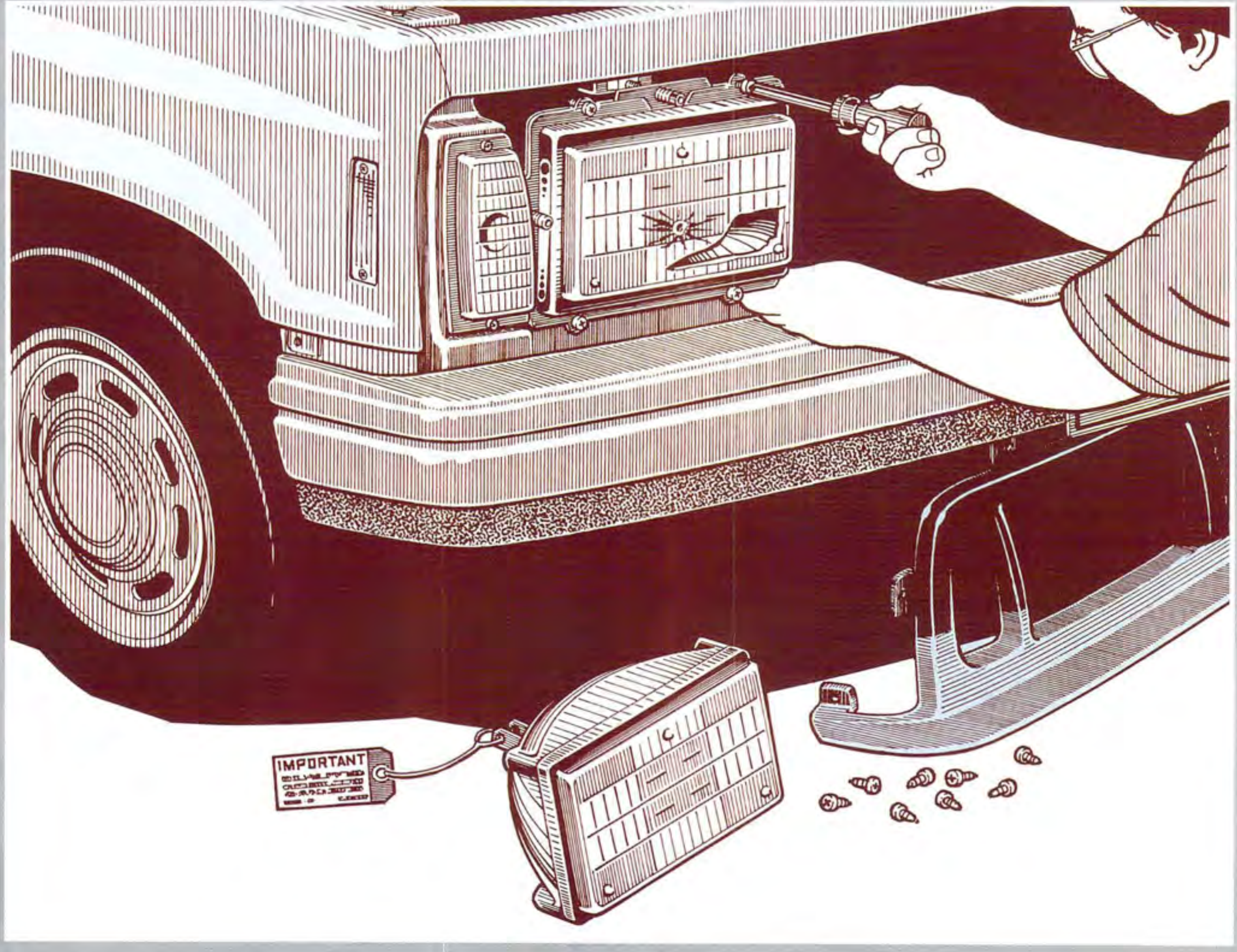

Servicing Electronic Suspensions

You and your Corvette have been terrorizing the same back road on the way home for months. A warning lamp that says SERVICE RIDE CONTROL winks at you from the dash. No problem, you say—I'll take it to the dealer for service next week.

That sensuous, sweeping switchback turn is disappearing—rapidly, very rapidly—under your front tires. You've been through this turn a hundred times before, but this time something is different. A small pothole suddenly makes your outside front tire start to skitter across the asphalt as you mow down a solid hundred feet of weeds between the verge and the ditch before regaining control.

Suddenly the need to see the dealer has become much more urgent. Then the dealer says it may cost as much as (gulp!) two grand to fix. Time for the Saturday mechanic to get busy.

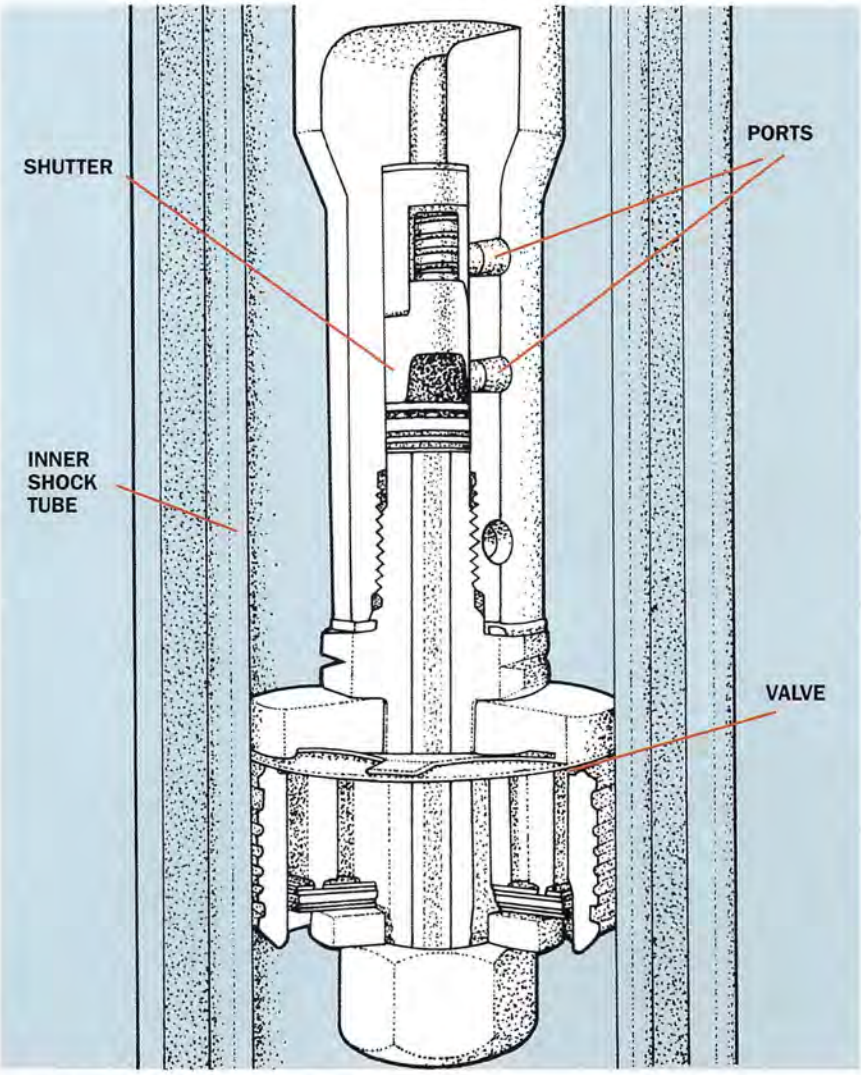

Fig. 1. Lincoln Continental's air-strut damper shutter varies oil flow through ports. Other systems are similar.

Finding the Problem

Consider the fact that the warning light mysteriously came on after the car was jostled around on some broken pavement. That means there's a good chance you're dealing with just a bad electrical connection. Or maybe one of the Bilstein shock absorbers was severely worn and just needed that last pounding to jam the adjustment mechanism. In retrospect, the right front corner has felt a bit mushy lately, hasn't it? Luckily the Selective Ride Control (SRC) system in the Corvette is one of the simplest setups, if not *the* simplest, of its kind. OE replacement parts aren't cheap (the shocks cost about $300 apiece), but at least you don't have to pay labor when doing the job yourself.

Pity the guy who has problems with an air-strut setup, such as those found in Lexus LS 400s, some Mitsubishis and the Lincoln Continental (Fig. 1). Except for the Infiniti's Q45 active suspension, these are the most complicated systems on the road. The Lexuses and Mitsus adjust damping rate, ride height and spring rate automatically through a computer. The Continental changes just damping and ride height. Unlike the Corvette and some other ride-control systems, the Continental's air struts can't be converted to a conventional suspension. Tough luck.

Check the Basics First

Roughly half of all the mainstream carmakers offer some type of electronic suspension system on some models. In general, they all make computer-controlled adjustments on response to one or more of the following inputs: road conditions, vehicle speed and weight/load, brake

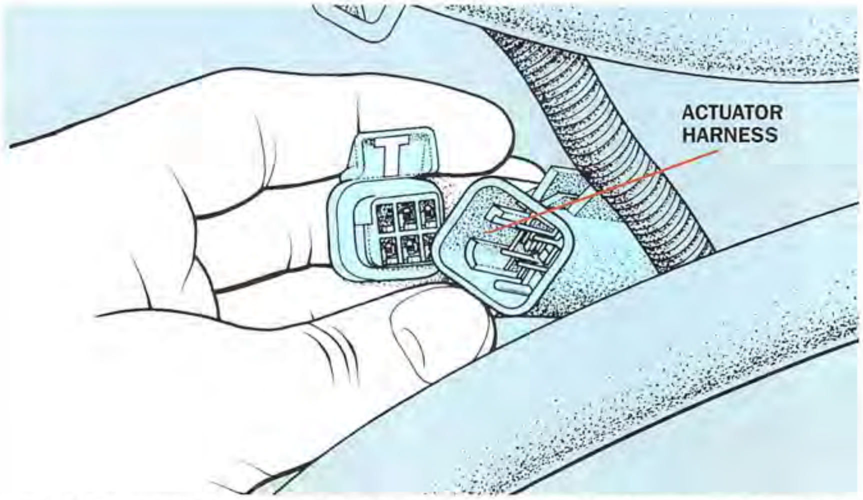

Fig. 2. With no trouble codes stored in a ride-control system's computer, begin the diagnosis by visual inspection of the parts and wiring.

application, steering-wheel angle, power-steering pressure, throttle position, lateral/longitudinal acceleration and the position of a dash- or console-mounted driver-selectable switch.

But don't let this list overwhelm you. The key to any modern automotive system is to keep it simple and check the basics first, then move on. One prerequisite for working on almost any modern car is adequate documentation. Beg, borrow or steal the manual for your car, which may involve ordering it from the dealer. Without some sort of manual, you'll be guessing, and probably not only won't be able to make the repair but you might inadvertently damage some pretty expensive parts.

In the case of your Corvette, there should be at least one trouble code stored in the computer. So you would jump right to "Corvette diagnostics" on the page 155, because it would probably lead you to the problem a lot faster.

If the computer hadn't turned on the warning message but you felt something wrong with the suspension (or you don't have the self-diagnostic information for the system you're dealing with), you would start by taking a close look at the adjustable shocks for some signs of oil leakage as if dealing with plain ol' shocks. If you see any stains or drips, the first piece of business is to replace the bad unit and hope the system is fixed. The same logic applies to all electronic suspensions, be they air, hydraulic or both. No matter how complex the system is, the damping portion of the shock/strut is still filled with hydraulic oil and needs to be sealed.

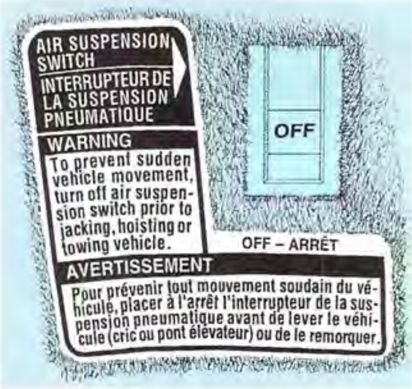
AIR SUSPENSION SWITCH
INTERRUPTEUR DE LA SUSPENSION PNEUMATIQUE
WARNING
To prevent sudden vehicle movement, turn off air suspension switch prior to jacking, hoisting or towing vehicle.
OFF
OFF – ARRÊT
AVERTISSEMENT
Pour prévenir tout mouvement soudain du véhicule, placer à l'arrêt l'interrupteur de la suspension pneumatique avant de lever le véhicule (cric ou pont élévateur) ou de le remorquer.

Fig. 3. Air suspension must be off before jacking up Lincoln Town Cars or Mark VII/VIIIs.

Next, do a visual inspection for any air/fluid lines that have burst and for damaged or misadjusted height sensors on those setups that use them. Look around for any obvious harness connector trouble and check the shock/strut actuators and their connectors (Fig. 2). If you hear the air compressor running constantly, it usually means a height sensor is out of whack, there's a small air leak or a relay is stuck in the ON position. If the compressor never turns on, you can suspect a bum height sensor, a stuck-open relay, a clogged air bleed-off valve, a blown fuse or even the compressor itself.

Ford issued a technical service bulletin on this exact subject back in the mid-90s. It seems that on 1990-93 Crown Victoria/Grand Marquis and Town Car models, the rear air suspension may not pump up when loaded because the air-supply line at the Y connector near the No. 4 crossmember tends to rub on the underside of the car body. The constant rubbing eventually wears through the line and creates a leak. An improved air line with a protective shield was issued to remedy the problem. The OE part numbers are F2AZ-5A897-B (Crown Vic/Grand Marquis) and F2VY-5A897-A (Town Car).

Corvette Nitty-Gritty

Understanding how the SRC system works will help to diagnose and fix it and will make other setups easier to comprehend. SRC consists of four adjustable gas-charged shocks, an actuator motor on top of each shock and a computer to drive the actuators. SRC can adjust the damping rate only—the ride height never changes. Inputs to the computer include a 3-position driver-selectable switch, the vehicle speed sensor and signals from each actuator's Hall-effect position sensor. Based on the car's speed and where the driver has the switch, the computer powers the actuator motors so that they rotate a certain number of degrees. Every time the key is turned on, the computer rotates the actuators back to the 0° stop.

Inside the conventional-looking shock shaft is a thin rod, called the damping rod. It extends from the top of the shaft down to the piston. On top of the damping rod is a gear. When the actuator is mounted, its gear fits over the gear on the rod, which enables the actuator to rotate the rod up to 160° inside the larger shaft. Because the bottom of the rod is beveled, as it turns clockwise it gradually closes off the bypass orifice machined in the side of the larger shaft. The smaller the orifice

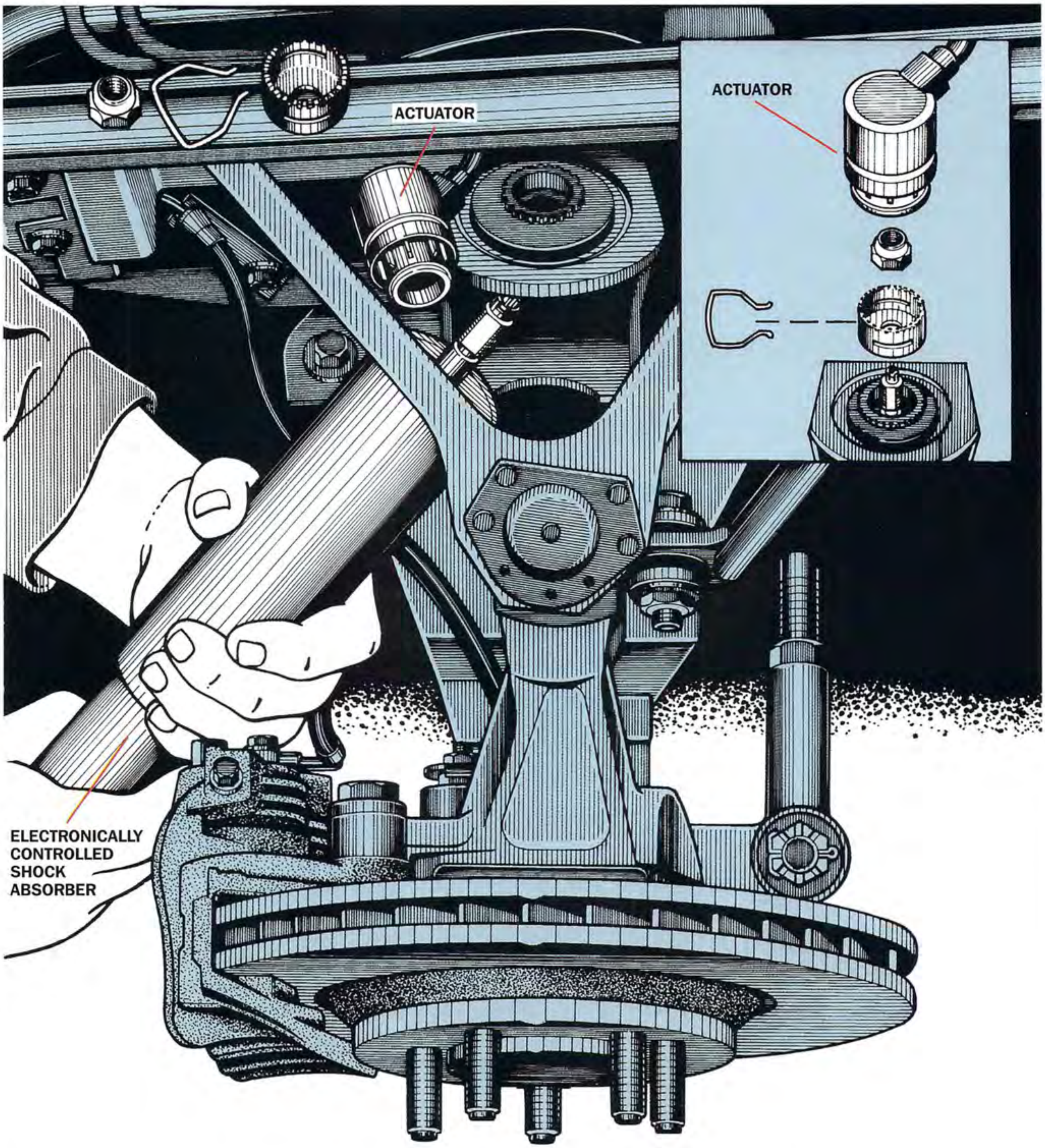

gets, the stiffer the shock gets. Mechanically it's simple. As long as the electronics don't screw up, only normal wear should affect the system.

Other electronic dampers work in basically the same way but they vary in design. For instance, the Continental's air struts use a shutter-type mechanism with different size orifices to vary the oil flow.

Replacing the Shocks

If you find a leaker or one shock that's clearly worn out, replacing it is about as straightforward as any conventional shock. Start by loosening the lug nuts of the wheel that has to come off and jack up the appropriate end. Remove the wheel. (On some air systems, such as the Lincoln Town Car and Mark series cars, you must flip off the air suspension switch in the trunk before jacking up the vehicle—don't say we didn't warn you (Fig. 3). Next, pull the actuator's locking clip

Fig. 4. The Corvette SRC actuator must mesh cleanly with its shock's damping rod gear or the computer will set a trouble code. The locking clip should slide easilty into place.

and pull the actuator straight up off the shock to expose the damping rod's little gear. Lay the actuator gently on a frame rail if possible. Don't let it hang on the harness. If there's no place to set the actuator down, either tie it out of the way or pull its connector from the main wiring harness and remove the actuator and its harness completely.

Now, you just need to remove the upper and lower mounting bolts and pull the shock. You may need to pump up like Arnold Schwarzenegger for a few seconds to compress the stiff gas shock enough to slip it out between the control arms. The same goes for slipping the new unit through the arms into place. Be careful not to damage the new damping rod gear. The actuator must remesh smoothly with the new gear. Anything that prevents the actuator from doing its job will set a trouble code in the computer (**Fig. 4**).

Actuator Precautions

Don't hit the Corvette's actuator with anything during installation—press it on gently by hand. Never whale on the upper retaining nut with an impact gun or you'll jam the damping rod. The nut must be torqued to 33 ft.-lb. Hold things steady with a wrench on the flats of the retainer insulator while tightening the nut. Make sure there's a good mesh between the damping rod gear and the actuator gear, and a good fit for the locking clip (**Fig. 5**).

Once mounted, make sure the actuator harness pigtail points directly rearward for a front shock and directly forward for a rear shock. This prevents the harnesses from rubbing against the fender splash shields and frame rails.

The actuators on other electronic shocks/struts are usually held in place by two screws, so they can be installed in only one position. This means the computer doesn't need to take a zero reading with each key cycle.

There's no way to install an SRC shock backward because the actuator position sensor keeps tabs on the damping rod angle. However, this doesn't necessarily apply to all electronic suspensions. Be sure to check the manual for the latest information.

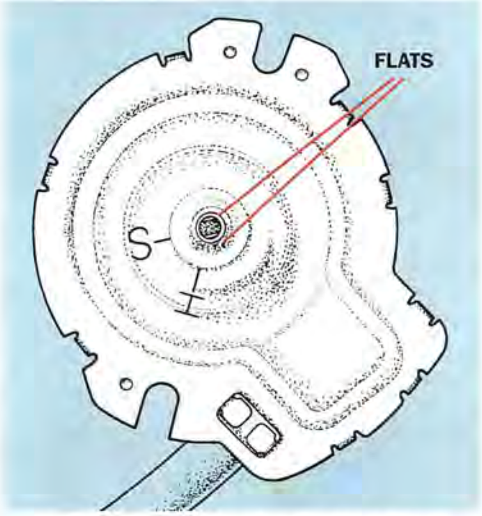

Fig. 5. Actuator must align with top of damping rod to maintain proper soft/hard positions.

Corvette Diagnostics

Remember that SERVICE RIDE CONTROL warning? When it's on, it means the system is in a fail-safe mode and the computer has locked all the actuators in the 60° position until the problem is fixed. With the car stopped and the ignition on, jump terminals A and B (1989 only) or A and C (later models) of the ALDL (assembly line diagnostic link) located under the knee bolster with a paper clip or a piece of jumper wire to make the SERVICE RIDE CONTROL warning message flash out 2-digit codes. There are 13 possible codes that can be spit out, but you'll need a service manual to decipher them and to do the proper tests.

Four things can prevent the warning light from flashing codes or from coming on with the key for a bulb check: a burned-out bulb, a blown fuse, a bad harness connection at the computer or a bad computer. The electronic brain for this system is found in the bin behind the driver's seat in the Corvette. In other cars, it's usually mounted behind the rear seat in the trunk area. Only two problems won't set codes in the SRC system: a leaking or worn shock and system voltage out of the 10- to 16-volt range. To clear codes after the repair has been completed, jump ALDL terminals A and B (1989) or A and C (later models) three times 2 seconds each.

The Real Time Damping system on late-model Corvettes (mid-late 1990s) also permits on-board trouble code diagnostics without a scan tool, through the Central Control Module. Codes are displayed on the Driver Information Center. Refer to the factory service manual (or the service information website—www.acdelco.com).

DIFFICULT

Fixing Power Windows

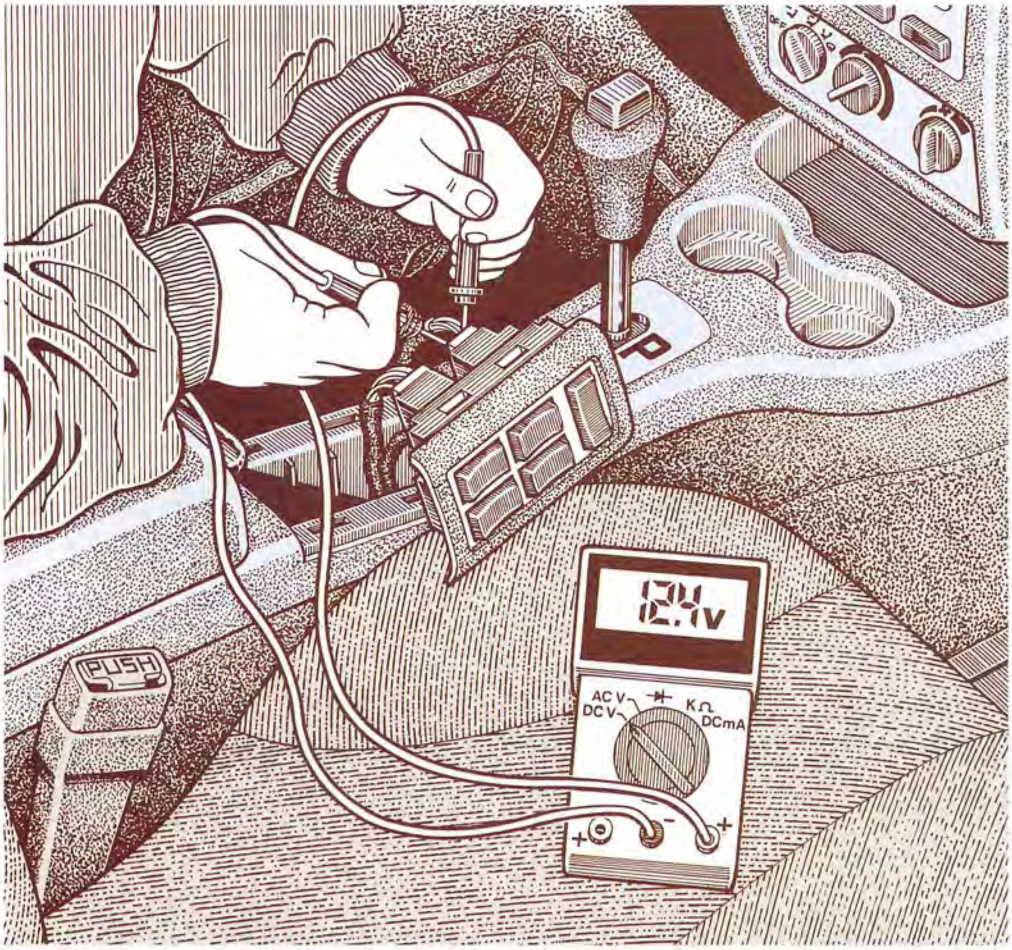

(Fig. 1) Carefully backprobe the window switches to isolate any electrical faults in the switches, connectors or wiring.

Another toll-booth, another mile—or at least it seems as if the tollbooths come every mile on this road, with a half mile of traffic idling its leisurely way up to the token monster. Within an arm's length of the bin, you toggle the power window switch with one hand while the other hand fingers a token, preparing to whip it into the basket just as you floor the throttle. All goes as planned—except the window doesn't move, the token bounces back into your face, and you have to jam on the brakes, crack the door and pitch a second token backhanded to keep from getting a ticket as a toll evader, all to the tune of horns blaring from the cars behind you.

Fortunately, power windows are usually one of the more reliable systems on a late-model car. And diagnosis and repair are usually pretty straightforward.

What's Up?

The most common power window mechanism is pretty basic. There's a simple regulator mechanism, usually similar to the mechanism used on garden-variety hand-cranked windows. It comes in several varieties—rack, sector and cable drive. Troubleshooting is pretty straightforward, once you get the door panels off—but your problem may be terribly simple and may not require removing any trim at all.

First: Are all of the windows on the fritz? Or just one? If you can't move any of the windows, the first place to look is at the fuse. Window regulators are high-current devices, and the fuse is sized to just barely be able to open all four windows together. Age and a few sticky window channels can pop a fuse.

Turn the key to the Run position, but don't start the car. If the fuse is blown, pushing a window button will do nothing at all: The motor won't groan and the glass won't quiver. If the fuse is good and you can hear the motor, or the glass acts like it wants to move, then you've got some sort of mechanical problem. If not, check the fuse. If the fuse box isn't labeled, check the owner's manual to see which fuse is the culprit. Don't go yanking fuses willy-nilly looking for a bad one—you might interrupt the power to the engine management computer, causing poor driveability for 30 minutes or so—or you might reset all the buttons on your car radio to that undersea-alien rock-gospel station.

Fuse okay, but the window still won't budge? Again, are all the windows dormant? Or just one? If it's just one, you still may get an opportunity to go spelunking inside the door. If it's all four, maybe it's something simpler you can troubleshoot under the dash.

At this point, if you've narrowed the fault down to some electrical problem that's not as simple as a blown fuse you need to round up a schematic of your car's electrical system and a voltmeter or 12v test light. All that's necessary now is to start at the fuse panel and follow the wiring to the switch, and from there move on to the motor, testing along the way for 12 volts. Somewhere, you'll find a loose or corroded connector interrupting the voltage to the motor. Or, the switch itself might be bad. If the driver's door switch won't open the right rear door, but the switch in the door will, look for either a bad switch in the driver's door or a fault in the intervening wiring.

Open Sesame

At this point, you probably need to be able to access the inside of the switch panel. On some vehicles, like the one in our lead illustration, you can simply pry the panel up with your fingers and backprobe the connectors (Fig. 1). Other vehicles may require that you remove the panel.

Door panels are held on with a bewildering variety of fasteners. Start by pulling off all of the door pulls and handles. The perimeter of the panel is customarily held on fragile plastic studs intended for one-time use. Pry them up carefully, and you should be able to reuse them.

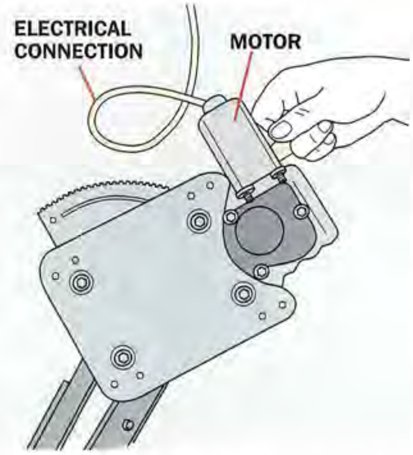

(Fig. 2) It may be possible to replace a bad motor, or you may need the entire mechanism.

Once you've got the door panel off, carefully remove the weather sheeting. You'll need to replace this later, and you may need fresh contact cement to do so.

Warning: You now have the ability to put your fingers into places where fingers normally don't go. As our mechanic pal Lefty points out, "A power window motor has enough torque to put a serious hurtin' on ya if it's actuated while errant digits are in the gears."

Proof Positive

As an absolute proof that the problem is electrical, try running a jumper wire direct from the battery positive terminal to the positive side of the motor to see if it comes alive.

Be aware that a few window regulator systems supply 12v constantly, and switch the ground side of the circuit. Check the schematic. Also, most vehicles have the ability to lock—and deactivate—the rear windows. Check this switch if only the rears are balky. Occasionally, the true problem is a bad motor (Fig. 2). You'll have to replace it. Otherwise, you can simply trace the wires until you find the problem.

Sticky

All windows have gaskets and seals to keep wind noise and rain out. If the window has a slow spot or won't open or close properly, check the gaskets. A gasket that's misplaced or torn can prevent proper operation. If the gasket is loose, or even torn, you may be able to repair it. If the gasket is simply loose, get some 3M Super Weatherstrip Adhesive at the auto parts store. Clean off the old adhesive with lacquer thinner and reglue the gasket into place. Allow this to dry overnight with the window closed, and be certain you're not gluing the window to the gasket.

If the gasket is torn, you might be able to use a super glue to simply repair the tear. You may be able to judiciously trim a loose corner of gasket away with a single-edge razor blade. Be particularly careful about doing this on the part of the gasket that sits outside of the glass, because it may admit rain and salt spray to the inside of the door in quantities too large for the door's internal drainage system to cope with.

HOW IT WORKS

Automatic Windows

Some late-model high-end cars have frameless windows that automatically crank themselves open a quarter-inch or so as the doors are opened. It happens so fast that you may not notice it. The window opens rapidly, clearing the seal before the door latch clears. It then closes automatically about a second after the door latch latches. There are two advantages to this. First, the slightly open window vents interior air, which can actually make doors on tightly sealed cars hard to open by springing the door back open against air pressure. It also lets the manufacturer use a vastly different style of seal on the top of the window. The seal can more closely resemble a sedan door seal, with a small lip protruding over the top of the glass. This type of seal won't work on frameless windows because the glass has to clear the seal as it opens and closes. This type of seal allows less water and noise intrusion. A similar automatic system is used on those high-end convertibles with fully automatic tops, to provide clearance for the moving parts of the top. The downside with any automatic window system is with the logic control module needed to achieve this. Repairs will probably need a factory shop manual and, potentially, some expensive parts.

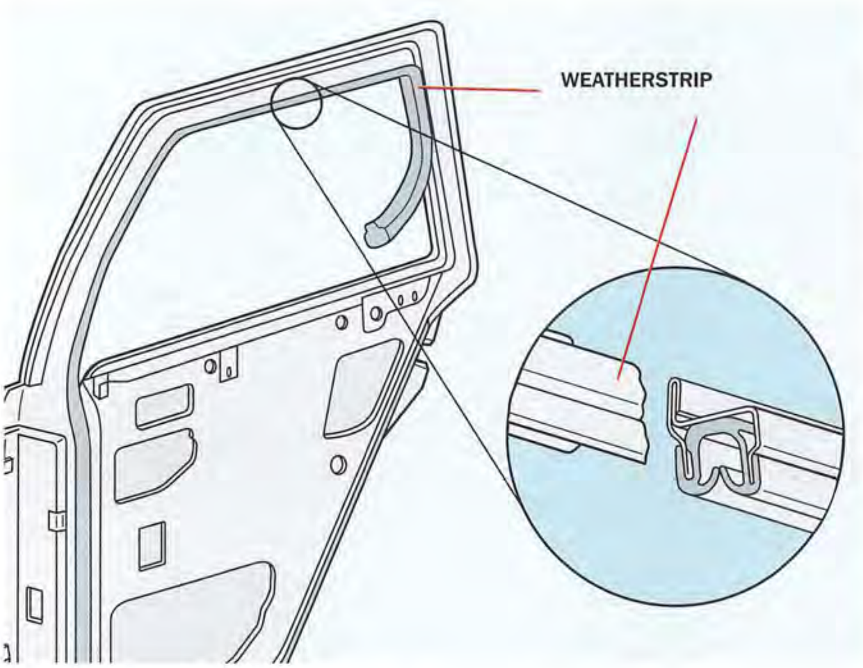

(Fig. 3) Check the weatherstripping and window channel for torn, loose or folded rubber parts, or foreign objects in the way.

Replacing a gasket or seal with a new part is generally straightforward. If it's not obvious that the gasket is astray, inspect the entire gasket and channel carefully. Look for damage, but also look for such things as pine sap, fossilized Froot Loops or other foreign objects that might make the window stick or bind. Clean the surface of the gasket and window with lacquer thinner to remove oxidized rubber and scum (Fig. 3).

There's a fair amount of friction between the gasket and the window glass. Almost any misalignment can dramatically increase the friction to the point where the motor no longer has enough torque to move the glass properly.

Lubricate the entire channel with silicone spray or protectant, because the reduced friction just might get your window working again.

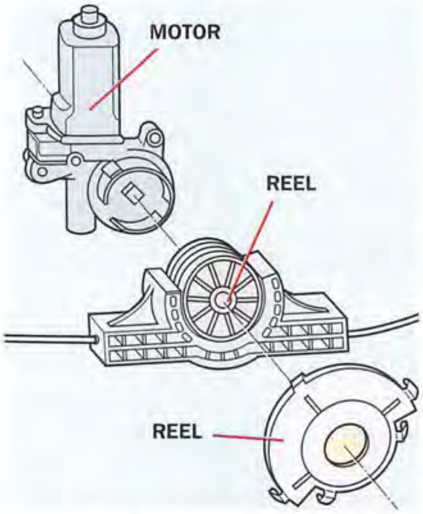

(Fig. 4) Reel and cable window regulators are simple mechanisms, but can be fussy about cable routing and may snag if jammed.

It's also possible that the problem is deeper inside the door. If so, you'll need to pull the door panel and go poking around. Remember to pull the fuse to prevent amputating your fingers. You can use either a rubber wedge doorstop or a couple of feet of duct tape to anchor the glass up while you work.

Sometimes the problem is nothing more than a loose bolt allowing the door's inner structure to move around, misaligning the window track. Many doors have slotted holes for the attachment points for internal parts, so careful consideration of the misalignment will sometimes let you simply slide one adjustment a quarter-inch or so and straighten it all out. All bets are off if the door has been damaged in a crash. It may take a long time to get everything working right.

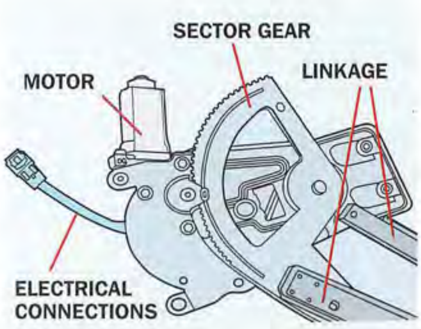

(Fig. 5) Severe misalignment caused by loose fasteners can jam gear-type regulators.

Lastly, the mechanism that runs the window up and down may be faulty. Whether it's a gear-and-sector, scissors lift or cable-operated mechanism (Fig. 4), you'll need to watch it moving up and down a few times. Again, keep your fingers out of the works. Sometimes the problem will be a loose fastener or rivet, sometimes a broken or missing bushing (Fig. 5). Cables can bind on the drum or become sticky. Lube all the friction points with white grease. Don't forget there are gaskets in the window track down below the top of the door, and you may need to reglue, repair or lubricate them.

DIFFICULT

Replacing Your Fuel Pump

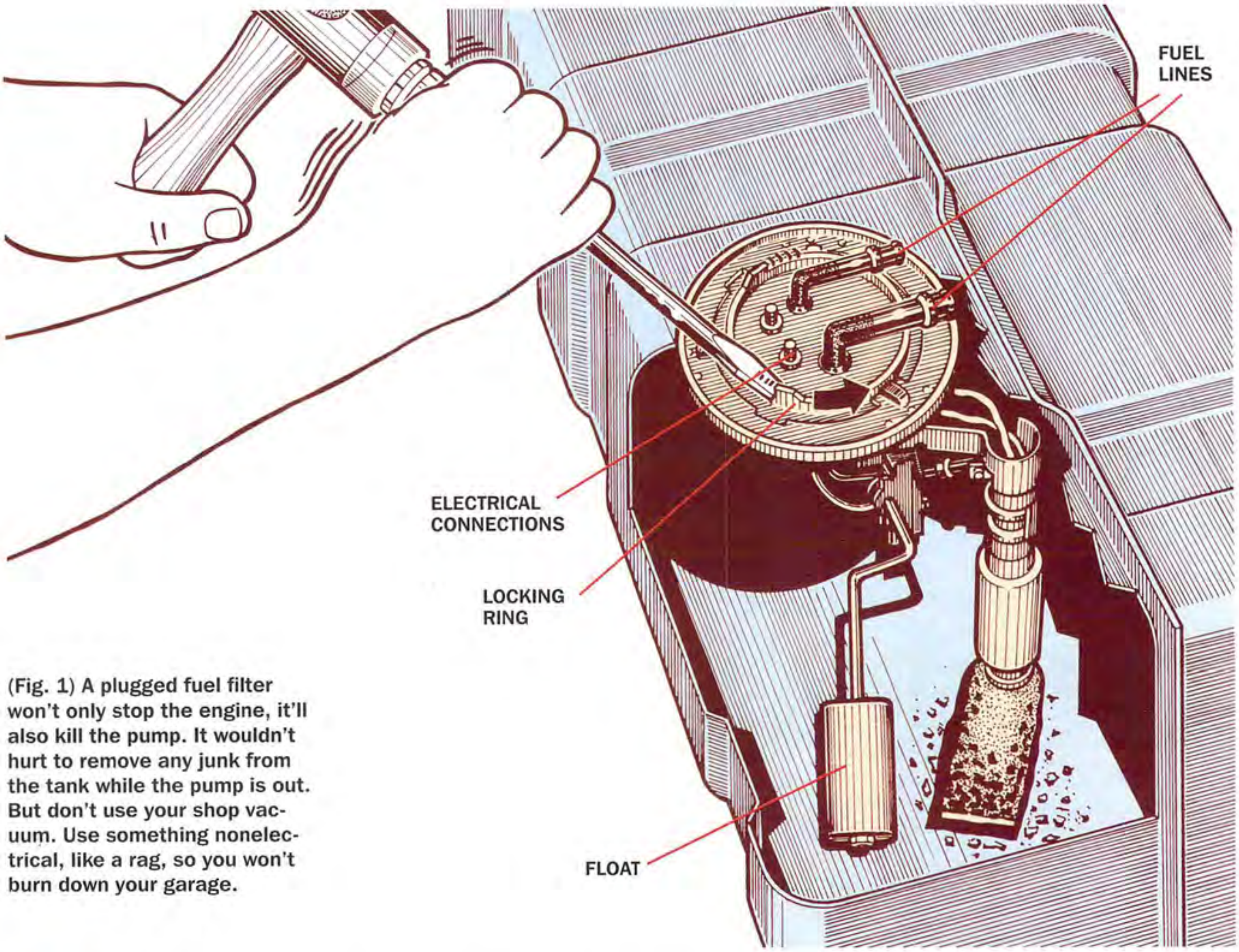

(Fig. 1) A plugged fuel filter won't only stop the engine, it'll also kill the pump. It wouldn't hurt to remove any junk from the tank while the pump is out. But don't use your shop vacuum. Use something nonelectrical, like a rag, so you won't burn down your garage.

Ten minutes late. Gulp half your coffee, sprint out to the car. Twist the key and she cranks. And cranks. And cranks. No fire, and your ulcer potential increases along with your blood pressure. It always seems to happen when you're behind schedule, as if the gods of internal combustion are eternally against you. You should've paid attention to that occasional bucking on the highway.

If there's anything over 60,000 miles on the odometer, but more commonly 100,000 plus, there's a good chance that the cause of this distressing no-start condition is an electric fuel pump that's no longer capable of forcing fuel forward with sufficient pressure. Even if the condition hasn't reached the point of grounding you, pump inadequacies can cause numerous driveability problems, such as momentary cutting out, hesitation, low power and stalling at inopportune moments (typically, it'll start again after it's cooled off).

For years, the car companies have been looking for a pump life of 10,000 hours (say, 400K miles) as a bogey. Now, they're asking their O.E. suppliers for "life-of-car" (that many miles isn't a car's life span anymore?). To reach this goal, pump makers have lowered amperage draw and balanced armatures more precisely. That's all very admirable, but ask any service technician and he'll tell you that they simply don't last anywhere near that long in most cases. What's happening?

In a word, crud. Many (most?) of the fuel filters replaced are so jammed up you can't blow through them (Fig. 1). If you had checked the pump's amperage draw before removing the old filter, you might have seen up to twice the expected number—perhaps eight to 10 instead of four to six. It shouldn't take much of an intellectual leap to realize that a plugged fuel filter at the outlet site will make a pump work harder, and all that extra current will wipe out the brushes and groove the commutator, killing the pump prematurely. Also, the pump's already digested whatever's causing the restriction.

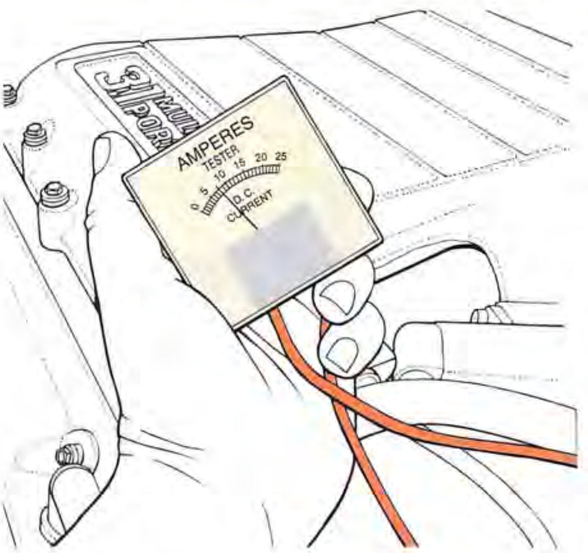

(Fig. 2) Find out how many amps the pump is drawing with a current draw test. Just bypass the wiring and run jumpers from the battery.

While on the subject of pump fatalities, another reason pumps burn out is because immersion in gasoline is necessary for cooling, yet people run around on "E." So, try to keep your tank at least half full.

Two-Legged Tripod?

For a century, mechanics have been taught that spark, compression and fuel are the legs of the tripod that support an engine's ability to run, and that's still true. So, before you jump to any unfortunate conclusions about your fuel pump, do a spark check (a story for another day) and make sure the camshaft is still connected to the crankshaft (remove the oil filler cap and watch the cam or rockers while you have a helper crank the engine). If these two essentials are okay, you can start thinking gasoline.

Turn the key on and listen carefully. In most cars, you'll hear the pump run for a few seconds. No? Then check the fuse. If it's blown, and the car starts after you replace it, you should still find out how many amps it's being forced to carry. If pump electrical draw is too high, a fix now will head off a future breakdown.

You'll need either a low-reading analog ammeter (Fig. 2) or a DMM (Digital Multi-Meter) with sufficient current-carrying capacity. Amperage testing is done either by hooking up the meter in series with the load, or with an inductive pickup that you clamp around the wire. The latter works best with big current flow, such as you'd find in the starter circuit, so we prefer the former for diagnosing fuel pumps.

That means you've got to break into the circuit. Connect one of the ammeter's leads to the positive battery post and the other to the pump's hot wire. Look this up in a service manual. Ditto for the specs, but we will say that if you see anything over 5 amps with a low- to mid-pressure system (13 to 45 psi), or 7 amps with a high-pressure version (60 psi) you've got a problem.

Will on-board diagnostics help you here? Not very often, although a problem in the pump relay circuit will set trouble code No. 42 on a typical Chrysler product, or code No. 87 on a garden-variety Ford.

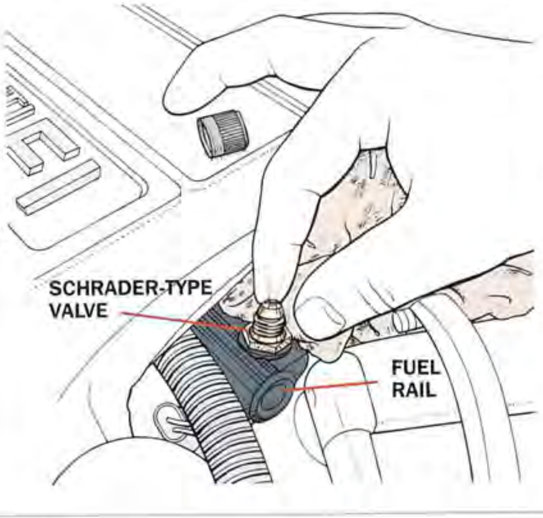

(Fig. 3) Before you open any connections, relieve system pressure.

Excluder and PSI

Have you replaced the fuel filter at the recommended intervals? How about ever? If you've been lax in this regard, it's certainly one of the first things to check. Even if it's not the reason the engine won't fire up, it's critical maintenance that you should be doing anyway, so you're not wasting your time.

Important: Filter removal will be a lot less dangerous and messy if you relieve system pressure first. Wrap a rag around the Schrader-type valve on the fuel rail and press down on the valve's pin (Fig. 3). Or, remove the fuel pump fuse, and crank for 20 seconds.

Checking fuel pressure is one of the basics of troubleshooting, and reasonably priced gauges are available. Most screw onto the Schrader-type valve, but with some imports you'll need a special banjo connector or a special fitting.

Turn the ignition on and read the gauge. No or low pressure may mean the pump's electrical circuit is faulty. This could be because of a defective control relay, but check for bad connections before you think about a new relay. Examine all the connectors for evidence of corrosion or looseness. Test for the presence of battery voltage at the pump's positive terminal or wire, and check for a bad ground as well.

By the way, an inertia-type safety switch in the pump's circuit, as found on many Fords, can be tripped by a minor impact, such as bumping into a snowbank.

Voltage-drop testing is the best way to locate high resistance. With the circuit powered up, use an accurate voltmeter to see if you get a reading across connections and lengths of wire. Anything more than about 0.2 volt is too much.

Other possible causes of too few psi to start the car are the already-mentioned clogged filter, a pick-up sock in the tank that's blocked, a crimped line or, of course, a weak pump.

If key-on/engine-off pressure was within specs, use a test light at the

HOW IT WORKS

Fuel Pumps

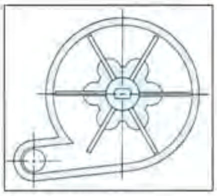

TURBINE

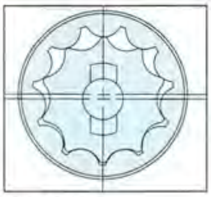

GEROTOR

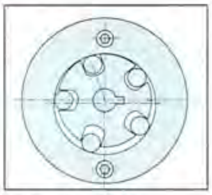

ROLLER CELL

Unlike mechanical fuel pumps, which are necessarily attached to the engine, the electric variety is mounted back in or by the tank. There are three types: turbine, gerotor and roller cell. The first uses rollers in a notched rotor to catch gasoline and force it into a small-volume area of the housing. It has lots of moving parts and can be noisy. Gerotors are similar to some oil pumps—they squeeze liquid by means of the eccentric action between a star-shaped rotor and a matching element that surrounds it. Today, however, the O.E. trend is toward the turbine type.

With throttle-body fuel injection (also called "wet throttle plate"), electric pumps typically put out 15 to 20 psi, whereas port injection requires 45 to 60 psi.

If you're still wondering why mechanical pumps aren't used with fuel-injected engines, think pressure and volume. The former has to be higher than the 3 to 5 psi a typical mechanical unit supplies. Otherwise, you won't get that nicely vaporized plume that burns so well. Plenty of fuel has to be flowing all the time, too. Only a small percentage of what's available is actually used even at full throttle. The rest serves to keep those injectors and rails cool as it recirculates to the tank. That, plus the location of the pump way back in or by the tank, pretty much eliminates vapor lock problems.

appropriate injector wire (or, pull an injector connector and plug one of those inexpensive "noid" lights into the harness). No flashing? Then the injectors aren't firing, which means troubleshooting will have to graduate to a whole other level of complexity as you look into the electronic engine management and EFI systems.

It Starts, But ...

In cases where the engine runs, but has a driveability problem, concentrate on your gauge readings. You should see violent needle swings between dead-head and running pressures (Fig. 4). A slow rise means trouble. Pull the pressure regulator's vacuum hose and you should see an increase. Too many psi may be due to a defective regulator or a restricted return line.

Just because you've got specified pressure doesn't mean there's sufficient flow. Total system volume at the Schrader-type valve of 1 pint in 20 seconds will run any car. (While a typical pump may flow 30 gal. per hour, less than a tenth of that is needed to run the engine. The rest is returned to the tank, which helps keep everything cool.)

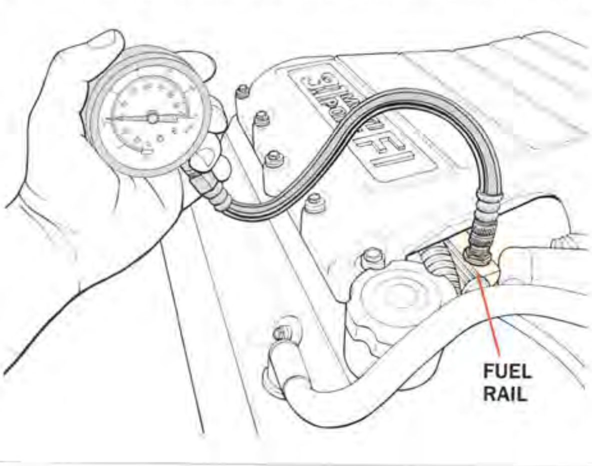

(Fig. 4) A fuel pressure gauge can tell you plenty. Take readings dead-head (key on, engine off) and running.

Kaput

Pump replacement may be easy or difficult, depending on the design. On vehicles with in-tank pumps, access is sometimes just a matter of removing a seat, or folding back the trunk mat and removing a panel. Others require that you drop the tank, which can be quite a ponderous undertaking, especially if you've just filled up (gasoline weighs 6 pounds per gallon). We use the term "drop" in a mechanic's sense here—we don't really mean to drop it. You can siphon it down—dump what you get into your other car, or give it to a neighbor. If you have a good floor jack, put a big square of plywood between it and the tank. Remember that there's still a certain amount of fuel in the tank after you've drained it. No smoking, and dry up any spilled fuel right away. It'd be a really good idea to work outdoors, too. Remember that gasoline vapor is heavier than air, and will spill down the basement stairs until it finds the pilot light for the water heater. *Poof!*

Speaking of flammable, don't even think about working on an electric fuel pump until you've disconnected the negative battery cable. On some cars, the tank is above a rear-suspension subframe, so access requires unbolting it.

Once you've got access, the actual removal and installation is pretty straightforward. Special spanners are often recommended for unscrewing the pump assembly locking ring, but we'll bet you'll figure something out. Just make sure you get that sock or strainer on there right.

Eliminating Battery Drain

It's a familiar feeling, twisting the key and being rewarded with little more than a groan from the starter. Your mechanic diagnoses it as nothing more than a tired battery, and a simple transplant cures the problem. It's *dejà vu*, as you've heard the same thing every winter—your car's battery needs replacement on an annual basis. It's time to get to the bottom of the problem.

Given normal circumstances and use, a car battery should last four to five years or more. Sure, there's a warranty, but it's typically pro rata, which means that unless the battery dies an almost instant death, all you will get is a partial credit toward the price of a new one.

There are a lot of reasons why a car eats batteries. If you've had hard luck with the ones you've been buying, the problem almost surely is not the fault of the battery manufacturers. Oh, there's an occasional dud, but most of the time the reason is in the car itself. Deep-discharging a few times will damage a battery, and so will constant over- or undercharging.

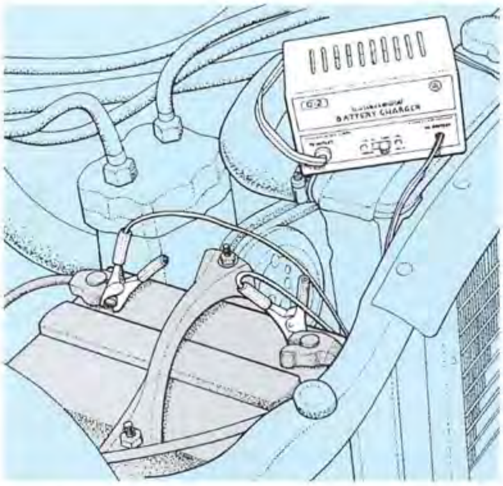

(Fig. 1) Inexpensive low-current trickle chargers may not revive a completely discharged battery.

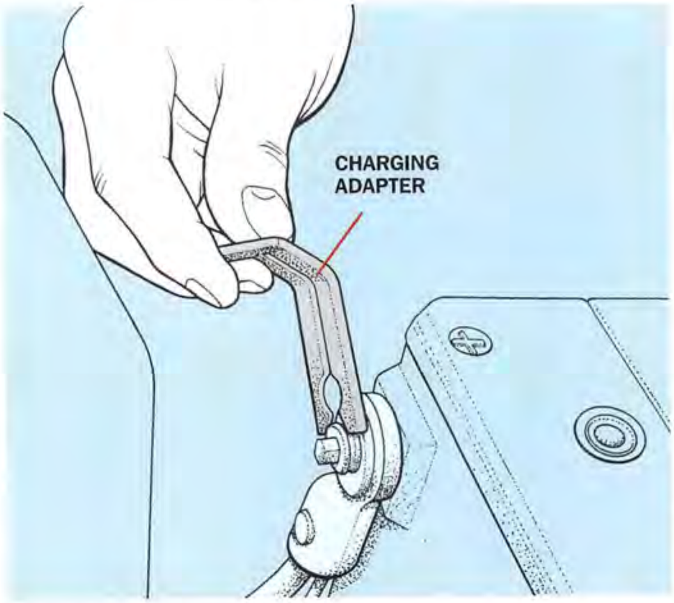

(Fig. 2) It may be necessary to use a battery-charging adapter to let the jumper cables get a good bite on side-terminal batteries.

All Charged Up

First, make sure the battery really is damaged before you spring for a new one. The simplest test is to fully charge it and see if it takes that charge. If it takes a charge, it's still good—but if it doesn't, it must be bad.

Here's where we get to the procedure that convinces many Saturday mechanics that they need a new battery. They connect a trickle charger overnight, and then conclude the next day that the battery can't take the charge. But there's a flaw in this approach: If a modern low-maintenance battery goes stone dead for some reason, it may need a real jolt to accept a charge. The discharged battery's internal resistance is very high, and many smaller chargers won't supply enough voltage to get over the hump (Fig. 1). There are professional chargers that start out with well over 12 volts (typically 18 to 24 volts) to break through, and your trickle charger (or even a reasonably decent homeowner-model charger) just won't kick out that much. You'll have to take the battery to a service station that has a premium charger. If you insist on trying to use your own small charger, it may take several days. Remember that the dashboard gauge really doesn't tell you much. Check the voltage across the battery terminals when you're done (charger off and headlights on for 15 seconds to remove a "surface charge"). It should be at least 12.7 volts, preferably about 12.8 volts.

Even if your battery isn't so rundown that it needs breakthrough voltage to start taking a charge, it does need good connections from the charger. Unfortunately, the side-terminal battery has a small contact area for the cable's alligator clip. Unless you have a specially shaped clip, you really need a charging adapter, made of copper, that clamps all the way around the side-terminal contact area (Fig. 2).

If the battery takes a charge but runs down in use, there's a problem that has to be traced. Otherwise, the car will seem to eat the next battery, too, even if it's a higher-capacity model.

Now that the battery is charged, you can check the charging voltage. With

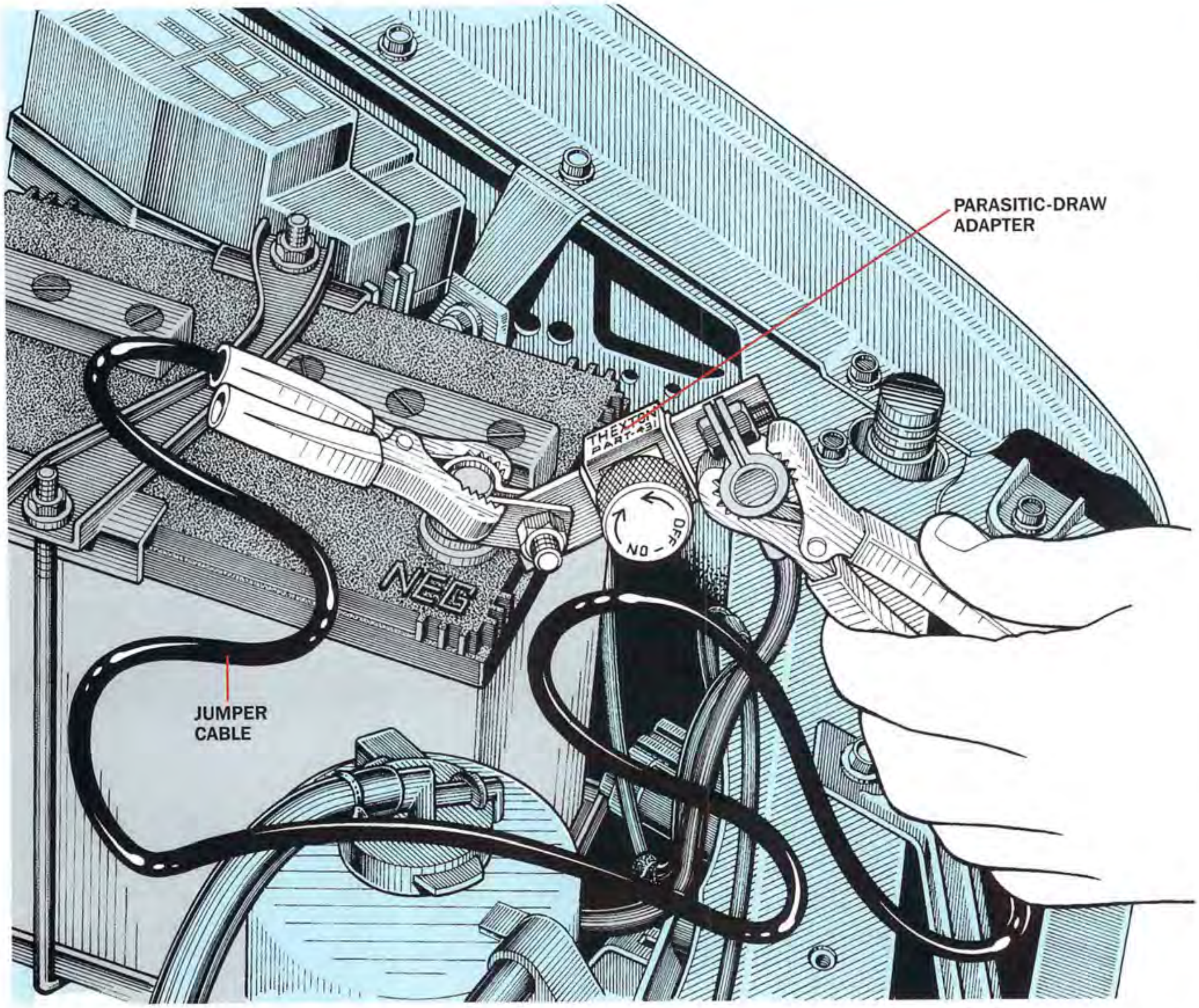

(Fig. 3) An inexpensive parasitic-draw adapter can simplify the diagnosis of key-off current-drain problems. Connect it in series with the battery's ground lead.

the engine running and warm and all the power-consuming accessories switched off, the voltage across the battery should read at least 13.8 volts. Next, turn on the headlights and rear-window defroster, and set the heater/air-conditioning fan on high. The voltage should still be above 13.5 volts. Anything less means you may have a malfunctioning alternator.

When demand exceeds supply

Let's see if you're pulling more juice than the charging system can replace during your normal driving cycle. If you have a high-performance sound system and a cellular phone, and run other accessories, this kind of usage really sucks amps. But you still should be okay—if you drive the car far enough. A half-mile in the morning with the a/c on and a half-mile back home at night with the a/c and lights on, perhaps with your favorite CD blasting through the speakers, is a great recipe for battery rundown. Drive your car long enough to pay back the battery for what you've borrowed, plus interest, at least once a week.

In addition, today's cars pull a fair amount of current even when everything seems to be off, just to keep alive the memories for all the electronics: the engine computer, keyless-entry receiver, the radio presets and the alarm system, to name the most common.

To keep the draw from depleting the battery, the electronics are supposed to go into a deep sleep, during which they require only a few milliamps. If they don't take this snooze, they can add to the current loss. To check, disconnect the battery ground cable. Install a parasitic-draw adapter (Fig. 3) in series with the battery and cable terminals, then connect an ammeter to the terminals. Open the adapter to see if the draw is less than 50 milliamps. Normally, you have to wait until the electronics go into snooze mode, which can take 60 to 70 minutes on some cars. Until then, the current draw may be sev-

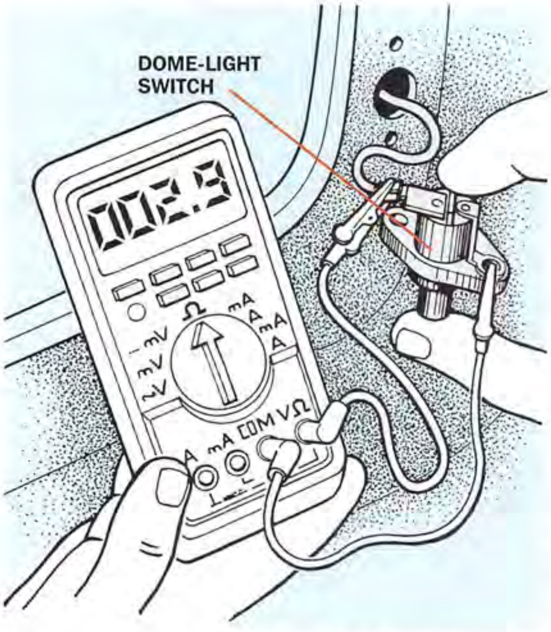

(Fig. 4) Check the dome-light switch for intermittent or abnormal resistance with an ohmmeter.

eral hundred milliamps, or even as much as several amps.

However, even if the draw is near 50 milliamps or so when you actually take your reading, it might be far too high before the snooze, or perhaps the electronics are awake for far too long. So, next, wake up the electronics and see what the readings are.

To do that, disconnect the ammeter from the parasitic-draw adapter. Connect a battery booster cable from the battery terminal to the cable terminal, bypassing the adapter. Turn the ignition key to the run (not to start) position. Wait a few seconds, then turn the key back to lock and take it out. Reconnect the ammeter, disconnect the booster cable and take the current-draw reading, which should be the awake number.

If the awake number is acceptable, make sure the electronics go to sleep within a reasonable amount of time. If the electronics still are awake and pulling hundreds of milliamps—or even an amp or more—hours later, something is very wrong. You can isolate a high current draw to a specific circuit by pulling one fuse at a time and rechecking. Just make sure the ignition is off.

A high-current-drawing short will cause a battery to go dead in hours. More likely, there's a small-to-medium draw that's bleeding the system, or a charging system that's not performing up to speed. Sometimes, you can spot the problem if you look hard. Many cars have interior lighting delays, but occasionally a key part of the delay system doesn't shut down. It might not be the in-plain-view dome light, but perhaps the dashboard lights or some floor lighting. If they're still on 15 minutes after the dome light is off, you've found a circuit glitch. For instance, 1990 to '91 Chevy Luminas had a problem caused by an electrical feedback from the radio. Even though the electronic timer said "Lights out!" and other bulbs were extinguished, the instruments continued to glow.

The lights for the glovebox, luggage compartment and hood are out of sight unless those areas are open, in which case the lights are supposed to be on. Peek inside as you just lift the cover, decklid or hood slightly to see if the light already is on or just coming on. If it's constantly on, then you've found your culprit.

A bad doorjamb switch can cause a lot of problems. If it's making intermittent contact, it's repeatedly waking up the electronics and turning on the interior lights.

The simplest overall check is to remove the fuse or fuses for the courtesy lamps to see if the battery stops running down. If that seems to be a possible cure, check out each doorjamb switch. Disconnect the battery ground strap, then remove the switch from the door opening and connect an ohmmeter (Fig. 4) from the switch terminal to its metal bracket, which provides the ground. The ohmmeter should read close to zero (a couple of ohms is okay). Repeatedly press the switch plunger part of the way and see if the ohmmeter reading goes to infinity (fully off position). If it takes a lot of plunger travel, or if the plunger response is inconsistent, the switch is probably bad.

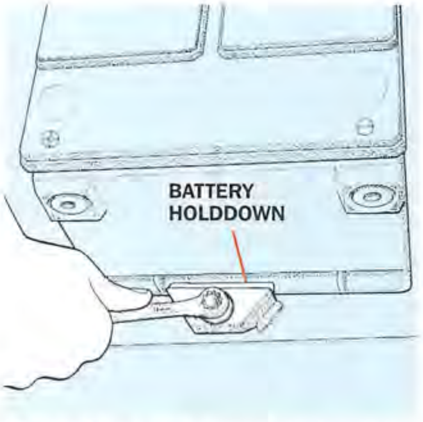

(Fig. 5) Check and tighten the holddown to keep the battery from vibrating.

Also check for a sticky radiator/condenser fan relay that keeps the fan going with the ignition off for a longer period of time than any computer strategy dictates, until it finally pops open. In fact, the relay could keep the fan on until the battery is dead. Look and listen for this one.

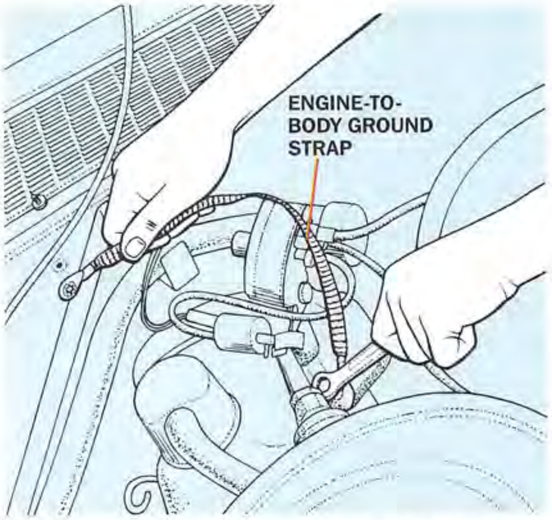

(Fig. 6) Make sure the ground strap is connected and in good shape. It's often hidden and neglected.

Rocking the Boat

If the current draw is acceptable and the issue seems to be that the battery just suffered from short life, make sure that the battery case is in good condition and the battery holddown is tight (Fig. 5). Even a slightly loose battery can vibrate enough to suffer a short life. Just because a top holddown looks tight doesn't mean the battery actually is secure. See if it rocks, even a little. Many holddowns are against a bottom lip on the battery, almost out of sight, so if the battery mounting isn't rock solid, check a down-low holddown.

Are the cables in good shape and securely tightened? The alternator may be pumping, but if the cable connections are bad, the juice won't actually get to the battery. The battery ground usually is the big offender.

In some cases, there's a split ground cable, and you can see one of the cable ends, in plain view and clearly tight. But the other part of the cable (usually the one to the engine) is buried deep, and what's out of sight shouldn't be out of mind. Find it, and put a wrench on it to see if it's snug (Fig. 6).

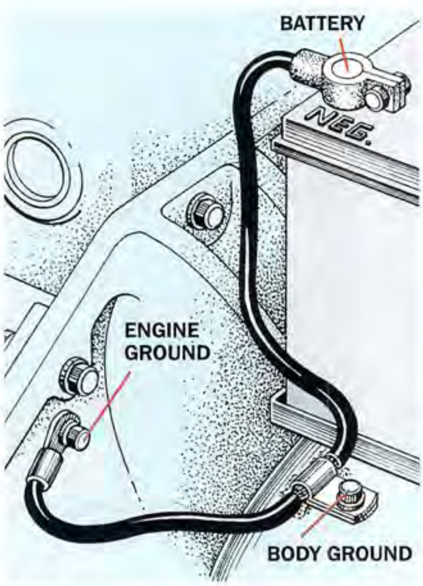

(Fig. 7) Check all three ground points on this type of ground strap for corrosion or looseness.

Many cars have a separate ground strap between body and engine, and it's usually nowhere near the battery (Fig. 7). In fact, if it's underhood, it's often disconnected for service and never reconnected. So the car's electrical system "forces" the juice through less efficient routes, and the battery is one of those that suffers. We've even seen engine-to-body ground straps that actually ran from the transmission to some part of the underbody. These ground straps fray, come apart and are far from any place you normally would look.

So if you don't see a clear-cut route from the battery negative (ground) terminal to both the engine and the body, look again—everywhere. Some older Ford products have a single ground cable to the engine, but with the insulation stripped at midpoint, and the body ground made by a clip from the stripped section to the fender-well sheetmetal.

Okay, the battery takes a charge and the cable connections are good. Well, perhaps the accessories you use are simply pulling more juice than the car can provide in your short-trip operation. A regular diet of trickle charging every night should provide the vitamins your battery needs for a normal life span.

EASY

Replacing Headlights

You're cruising down a 2-lane road late at night. Oncoming cars flash their lights and toot their horns at you. Or they come straight at you and suddenly veer away. No, they're not partied out. Unless you're blinding them with your high beams, one of your headlamps probably is out.

Replacing burned-out headlamps is a straightforward procedure, if the problem is just the headlamp. Although most are about $15, some headlamps are special numbers, available only from a dealer and can cost up to $45 apiece, even for an econobox. And there are even higher-tech (and higher-priced) systems, such as High Intensity Discharge (HID) headlamps, in some luxury models. They may look like ordinary headlamps—just as halogen sealed beams resemble the conventional sealed beams used years ago—but the manufacturing processes, the coatings, the materials and the gases used inside the bulbs are totally different and produce totally different lighting.

Don't leave in a headlamp with a cracked lens just because the light hasn't gone out. Even if only a small part of the lens is missing, dirt gets in and coats the bulb inside, which reduces light output. And if there's a chunk missing from the lens, the light beam isn't diffused as it should be and can blind oncoming drivers. With normal vehicle vibration, the lamp also may not hold its aim and you won't get the best view of the road. Also read "Servicing Composite Headlamps on page 149.

Circuit Check

If it seems that a headlamp is burned out, make a simple circuit check. With just one of two or four on the blink, the odds are that the sealed beam is the problem. It's unlikely that two headlamps will go out simultaneously.

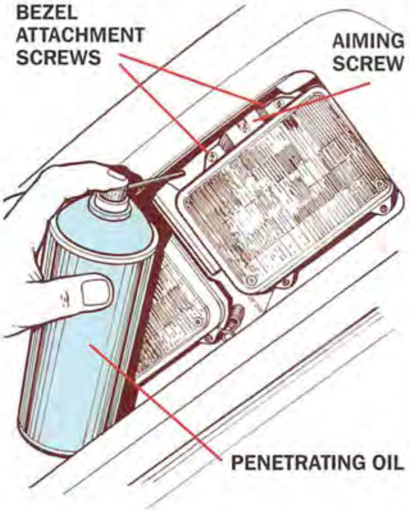

(Fig. 1) If the trim bezel screws are hard to turn, use penetrant to loosen the threads.

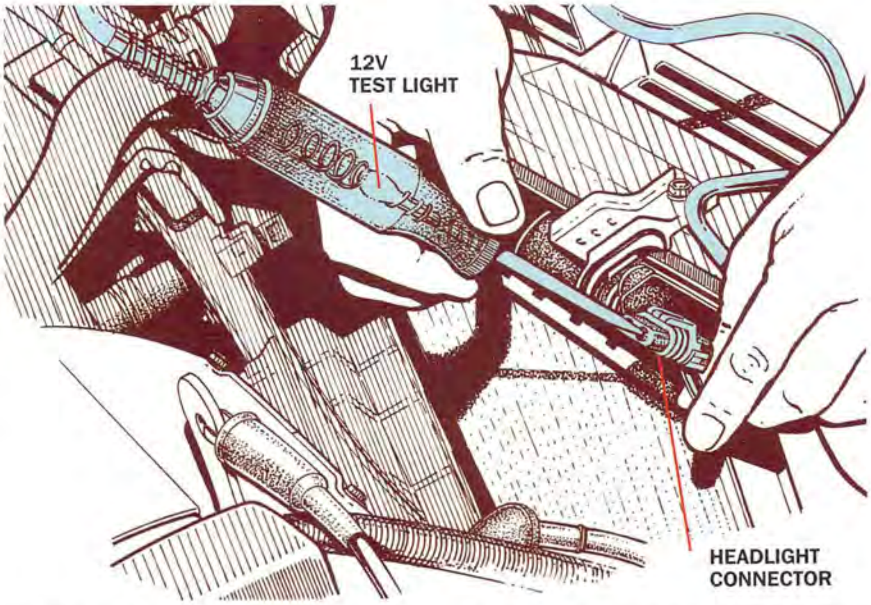

(Fig. 2) Use a simple 12v test light to confirm operating power and ground circuits before replacing an expensive bulb or sealed beam.

Of course, if the low and the high beams are in a single bulb, it is possible, but it's much less likely when they're separate lights.

Checking or changing a burned-out headlamp begins with gaining access. If it's a sealed beam, there's probably a trim plate in front, which comes off first. Most trim plates are held by either Phillips or Torx external screws. A few may be on studs that project through to the back of the lamp assembly—these are accessible from under the hood.

Next, look for and remove the four screws holding the headlamp rim or bezel. Be careful not to confuse them with the screws that control headlamp aim—the ones with coil springs under the screwheads. Are the screws rust-frozen in place? Don't try too hard to turn them, or you may strip out the screwheads. Spray them with penetrating solvent and allow it to soak in (Fig. 1). The attachment screws usually have a Phillips head, but—like the trim screws—they may be Torx.

Once the rim is out, pull the sealed beam forward and unplug the wiring connector from the back. You'll see two or three terminals on the lamp and connector. A 2-headlamp system always has three terminals. A 4-eye lamp may have two or three, depending on how the entire circuit is wired.

A simple check of the circuit is to see if there's power at the connector and a complete electrical ground. Take an ordinary 12v test light (**Fig. 2**), connect the alligator clip end to a ground and check each connector terminal with the probe. The headlamps should be turned on. The test light should go on in at least one terminal. If the sealed beam is a 2-eye with three terminals, flick on the high-beam switch and the test light should go on with the probe in a second one of the three terminals.

Next, unground the test light and gently press the alligator clip end (with a nail inserted, if necessary) to make contact with the other (ground wire) terminal—ground is usually a black wire. The test light should go on again.

If the test light goes on during these two checks, install the new sealed beam. If it doesn't, forget the installation until the circuit problem is repaired. If the test light went on when the alligator clip was connected to ground, but not when you probed the other connector terminal for a complete ground, the problem is a defective ground connection—a common problem.

Check a factory manual or aftermarket information system, such as www.alldata.com, for the location of your car's ground. If the connection is corroded, wire-brush it clean. If it's loose, retighten it.

No power to the headlamp connector? That can be more difficult to trace, and you'll need a wiring diagram and parts locator again.

A new circuit complication is on the scene: daytime running lights, or DRLs. These may be the low-beam or high-beam headlamps, or other external lights such as turn signals.

Headlamps are wired in what is called a parallel circuit—each lamp receives full battery voltage. However, if your car is a newer-than-'96 model

HOW IT WORKS

Quartz-Halogen Sealed Beam

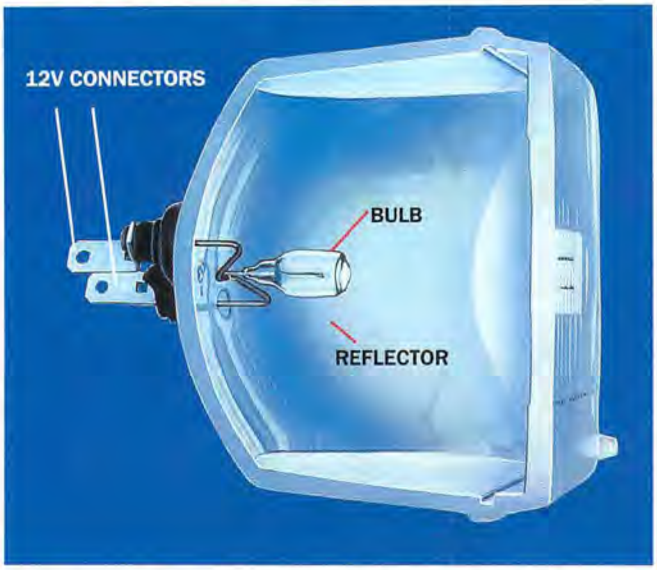

The conventional sealed beam of years ago had a tungsten filament inside a lens assembly with a reflector. Today's quartz-halogen beam is a separate bulb inside that lens assembly. It produces a brighter, whiter light—at least 25% more down-the-road visibility, with less power draw.

The bulb envelope is a high-temperature glass—the "quartz." The filament material is the familiar tungsten, but the bulb is filled with a special gas from a "family" called halogens, such as iodine or bromine gas (that family also includes a/c refrigerants, by the way). In a conventional sealed beam, the high temperature of the tungsten filament causes particles of the metal to vaporize from it. The particles deposit on the wall of the conventional sealed beam (darkening it), and eventually the filament burns out.

The quartz lens allows the bulb to operate at a much higher temperature to help produce that brighter, whiter light. The halogen gas inside the quartz envelope causes the "evaporating" tungsten particles to continuously redeposit on the filament, so it enjoys long life and consistent brightness virtually until it fails.

Headlamp technology is fast-moving. The High Intensity Discharge lamp (see page 150) is just one of the new approaches, but it's already installed as a low beam in some BMWs and Mercedes-Benzes and the Lincoln Mark VIII. This type has no filament, but instead it has two electrodes in an "atmosphere" of xenon gas. The current jumps the gap between the electrodes, producing a comparatively brighter light. You may have already seen these lights coming at you out on the highway. They're so white that they appear a little blue. But because there's no filament to burn out, the hour life of the xenon bulb is expected to go into six figures.

There are several other new headlamp designs on the market, including one on the '97 Lexus ES 300 that uses multiple reflectors. All the new lamps—including the high-performance halogens—are capable of producing a lot of brightness with reasonable power consumption. And introduced in some 2003 models are "smart" headlight systems that use signals from sensors to pivot the headlamps side to side for the road conditions and the way you're driving the car. These combination bright/smart lights are in part a technological fix for the declining night vision of a graying population.

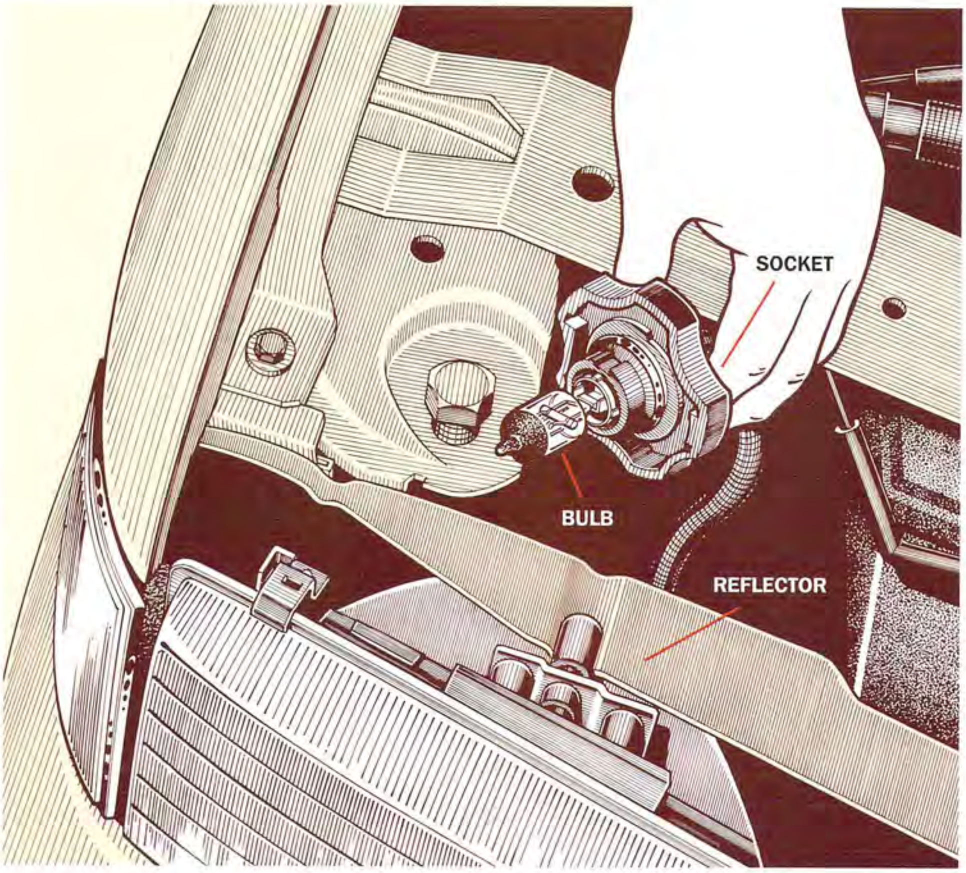

(Fig. 3) Newer cars use bulbs that can be replaced without using tools. Unscrew the socket from the reflector and change the bad bulb, then replace the socket.

with daytime running lights and the headlamp high beams are used, there's a second circuit controlling them through an electronic module. The high beams could be wired to the module so that they can be fed current for DRLs in a series circuit—the current flows through both, so that one acts as a resistor for the other. The result is less brightness, but also less current flow for reduced power consumption. If you have a problem with the high beams, a defective DRL module is one of the new possibilities of which you should be aware, both for sealed beams and replaceable bulb designs. Some cars have DRLs that are also designed to shut off when the parking brake is applied. Read those circuit descriptions carefully!

Replaceable Bulbs

Many late-model cars have a replaceable bulb plugged into the back of the lens assembly (Fig. 3) instead of the sealed beam. (European cars have always used this type.) Typically, the bulb has a wiring connector held by a plastic locktab, and a plastic lockring to hold the bulb.

Pry up the tab (with a thin screwdriver, if necessary) and hold it up as you pull the harness connector off the terminals on the bulb. You can probe the harness connector the same way as on the sealed-beam design to make sure the bulb is at fault. An alternative: If the bulb is easy to remove, take it out for inspection. Turn the plastic lockring, which may come out or stay loosely in place. Then grasp the bulb by its connector and gently rock it out (it seals with an O-ring to keep out dirt and moisture). If the bulb filament is burned out, that tells you it's most likely the problem. Warning: Replaceable bulbs are usually quartz-halogen—and touching the bulb with your fingers will leave fingerprints that will crack the quartz glass within a few hours. The new bulb will be wrapped in cardboard or foam, so don't remove the wrapping until last and avoid touching the bulb. Too late? Clean the bulb with alcohol and a clean cloth.

When you install the new bulb, work it in very carefully, then push smoothly to seat it in the lens housing. This ensures that the O-ring seals properly instead of being twisted or dislodged.

A bulb or sealed-beam harness connector may have a ribbed-rubber moisture seal if the location poses a problem with road splash. When you see this setup, make sure it enters the bulb connector uniformly and seats evenly.

Replace it with the same type of bulb or sealed beam as the one that came out. You may see high-performance versions of the type you removed on the auto parts store shelf. If you want to install one of these instead, replace the one on the other side too, for even lighting.

SEALED-BEAM LAMPS generally have this simple spade-lug triple-blade connector.

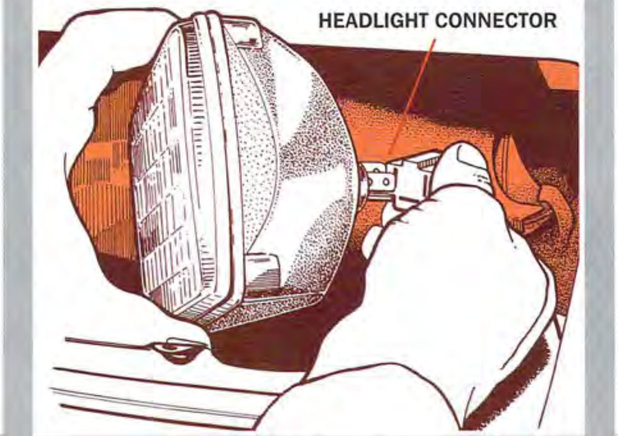

Replacing Your Car's Computer

Experiencing bad engine performance? Your vehicle may need brain surgery, and believe it or not, it's something an experienced Saturday mechanic can do. Yes, we're talking about the powertrain computer—the "brain" behind the operation of your engine, transmission, even the air conditioning.

Today, you can order a computer from your neighborhood parts store, which can choose one from a long list of reliable aftermarket suppliers. The computers are reconditioned, and in many cases you change over a program module, which holds the specific calibration instructions for the powertrain, such as ignition timing. If you're careful, you can do the job. In fact, in many cases you'll find you don't have to replace the computer, but can perform a lower-cost or even a no-cost "minor operation."

Diagnosis

Before you do actual brain surgery on your car, you have to find out if the engine is performing poorly because the brain is bad or if it's just responding to a "bad case of nerves" (bad signals from electronic sensors).

It could simply be a blown engine-computer circuit fuse, but if not, it's time for diagnosis. For this, refer to your vehicle's service manual—either a factory manual or an aftermarket manual with a complete engine-computer diagnostic section.

Check for diagnostic trouble codes. Although the procedure varies from car to car, on most pre-1996 vehicles there's a way that doesn't involve special equipment. Typically, you make a simple jumper-wire connection at the computer's diagnostic plug, which is under the dash or under the hood. Check your service manual. On later models meeting OBD II onboard diagnostic II regulations, you need a scan tool. More on OBD II later in this section.

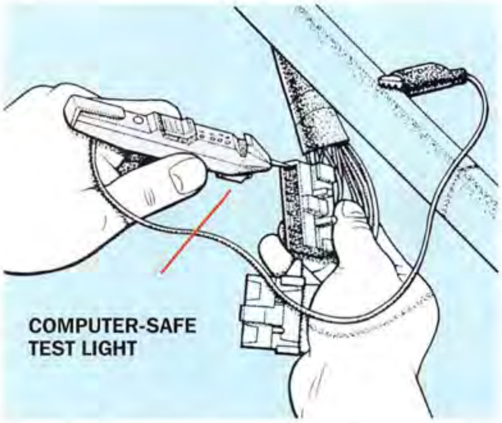

(Fig. 1) Check the feed with a high-impedance voltmeter or a computer-safe test light.

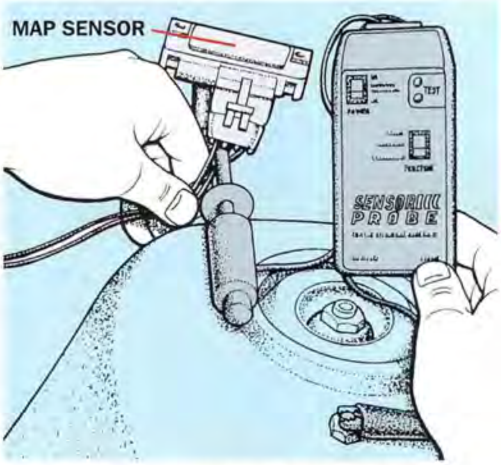

(Fig. 2) Check sensors with tester or voltmeter to rule out noncomputer problems.

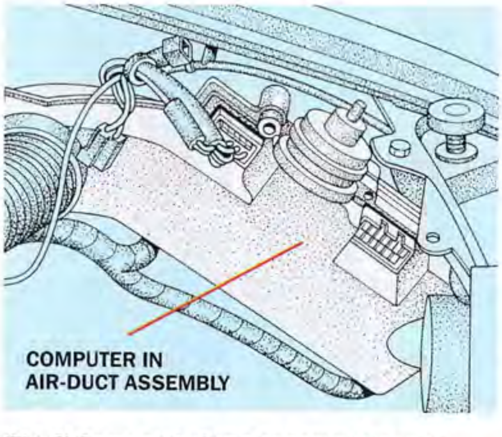

(Fig. 3) Some early Chrysler computers were housed in an air duct under the hood for cooling.

Just a single code? Check that particular item, whether it's a sensor, switch, motor or solenoid. The manual will give you a "diagnostic tree," a series of steps to follow, typically using a jumper wire and a computer-safe (high-impedance) multimeter to isolate the problem **(Fig. 1)**, which could be in the part itself or somewhere in its wiring circuit. Or you can use a sensor tester **(Fig. 2)**, which comes with adapters and/or special tips so you can probe wiring connectors. Correcting a problem in any circuit tied to the computer is an essential first step in every case, as it can be responsible for a computer failure.

In many instances, the diagnostic tree will conclude with REPLACE COMPUTER. You also could be led to the computer if there's a flickering CHECK ENGINE light, or no light at all (but the bulb is good).

Still another possibility: a laundry list of trouble codes. In many cases, the problem is in what's called a driver, a type of electronic switching device that is in the computer box. However, it isn't replaceable, so if it's bad, the only cure is a new computer.

Minor Surgery

Computers typically have been located under the dashboard, accessible after removing a trim pad or the glovebox on the passenger's

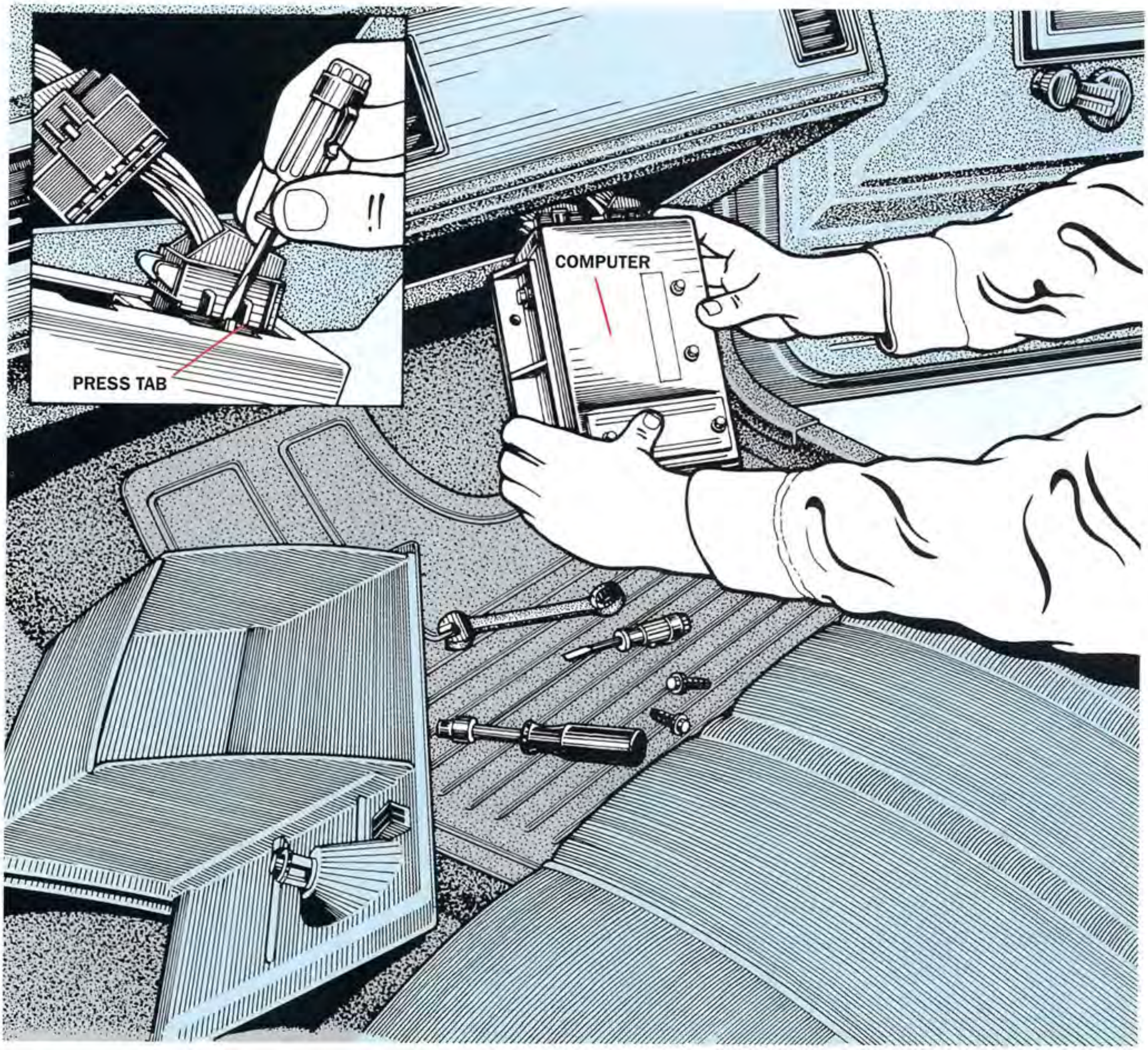

side. Older Chrysler computers usually are under the hood, integrated into an engine air-intake duct, but they were an exception for the 1980s (Fig. 3). Today, and as the electronics have become more heat-resistant, many computers are located under the hood.

Wherever the computer is located, the first steps are to turn off the ignition and disconnect the battery ground cable.

Next, unplug the computer's wiring connector. It often has plastic latches or locking tabs, so you should look carefully before you try to pry anything off and press tabs or latches wherever necessary (Fig. 4).

Eyeball the terminals in the computer and the wiring connector, looking for both corrosion and signs of bending. You can try to straighten a bent pin—but no rough stuff. And don't try to sand or scrape off corrosion, or you'll ruin the terminal. Brace the terminal with your fingers and clean it with a pink pencil eraser.

If everything looks good, start putting the connector back on by working it back and forth a few times, then pushing it all the way home. This may clean off a small amount of virtually invisible corrosion (enough to create havoc with electronics and produce multiple trouble codes) and get the computer system working again. If the problem comes back, well, at least you tried.

(Fig. 4) Once you've determined that the computer is really fried, replacing it is simple. It's often situated under the dash or behind the passenger's-side kick panel. There may be a locking tab holding the computer's umbilical connector in place.

Back to disconnecting the computer. With the wiring connector off the computer, probe the connector's terminal for the ground wire with a computer-safe test light or multimeter (set to volts), connected to a 12-volt source (either at the battery or at the fuse box). The test light should go on to full brightness, or the meter should read battery voltage. Also check the 12-volt power feed (or feeds) in the connector with a grounded test light or voltmeter. The results should be the same. If you get a low-voltage reading or a weak light, there's a bad connection. Whatever the result, always double-check all electronic grounds.

Brain Transplant

Is it a clear-cut case of a bad computer? Order the replacement and make the module changeover, if necessary (Fig. 5). General Motors vehicles and most European cars have computers with a plug-in program module, which doesn't come with the replacement computers. In fact, some have two modules, and you have to change over both.

Remove the module cover, then take out the module. Some of them are held by retainers—push them back and simultaneously lift the module straight up and out. With most others, you have to pry it out, and the safest way is with thumb and forefinger at the ends, applying an ever-so-slight rocking upward motion. The typical module is in a plastic carrier (don't try to separate the pieces). If space is too tight for your fingers, use two small screwdrivers—one at each end—to rock the module assembly up as evenly as possible. Do not pry one side up or you may bend the terminals.

Note the way the module comes out, as it has to go into the new computer the same way. It's possible to install it backward on some cars, which could destroy the new computer. Position the module on the mounting area of the replacement computer, push slowly (so you know that the terminals are lined up for engagement), applying pressure only on the ends of the module, until it seats. Lock the tabs.

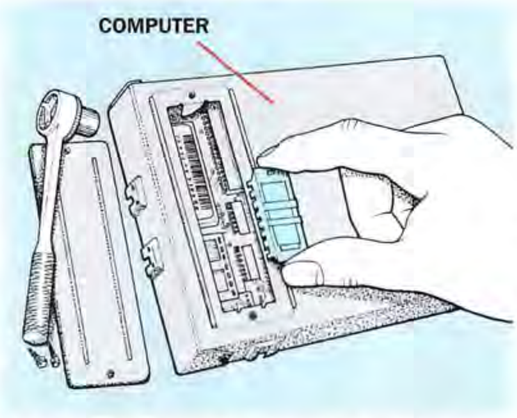

(Fig. 5) You may need to remove your old calibration module and install it in the new computer.

Reconnect the wiring harness to the computer. If you're working under the dash, you may build up a lot of static electricity while sliding into working position. As you make the harness connection, you could zap the computer with enough voltage from the static electricity to fry it. Simply ground part of your body to an underdash ground, such as by keeping a bare elbow against a metal brace.

Make sure you not only seat the wiring connector, but that you close any latches that hold it in position. Reinstall the computer, engaging all clips and tightening all bolts, screws and nuts so the computer is held securely in position.

Reconnect the car battery. The installation instructions with the replacement computer will provide a basic test of the system (similar to checking for trouble codes), and once that's done, you should be ready to drive. But all modern engine computers must go through a period of "learning" how you drive and making compensating adjustments (primarily to fuel-system calibrations). So be prepared for perhaps 100 miles or 4 to 5 hours of so-so performance as the new computer (even with an old program module) learns how to do its job.

This is something to keep in mind if you are doing a test run to see if one of the minor operations (cleaning corrosion, tightening a ground and so on) fixed a problem. Disconnecting the battery cable erases the trouble codes. However, if the faults remain, most trouble codes show up again within a half-hour of vehicle operation, even though it takes much longer for overall engine performance to return to normal after a successful fix.

All models since '96 have computers that conform to On-Board Diagnostics II (OBD II), an enhanced computer system with some standardization of trouble codes and diagnostic routines. One scan tester (computer-system data analyzer) will fit the underdash diagnostic plug on all OBD II cars, and there are models for Saturday mechanics for under $200 (see page 181).

The OBD II computers on many vehicles are reprogrammable through the diagnostic plug. With previous systems, carmakers offered factory fixes by having the dealer replace the computer or the program module. With the specific OBD II system, they now have the dealer simply reprogram the computer. A few cars have had reprogrammable computers in past (pre-OBD II) years, but with OBD II, it's basically across the board on U.S. nameplates and most imports.

EASY

Curing A/C Odors

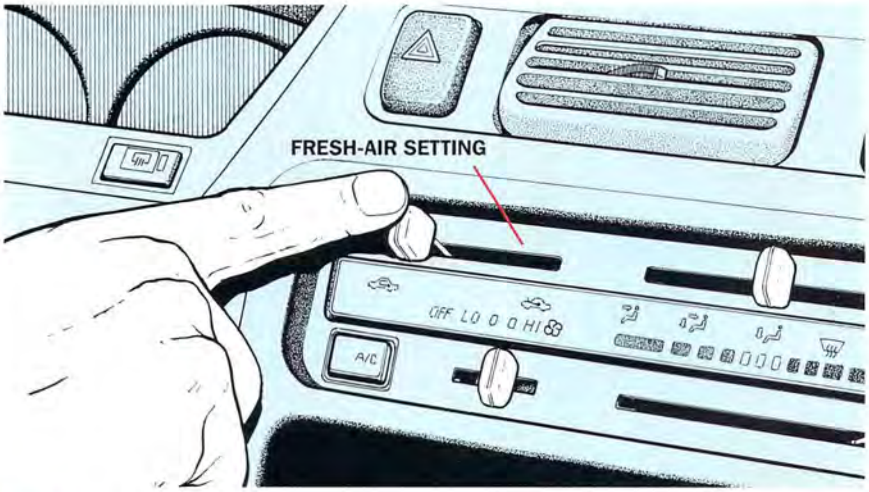

(Fig 1) The simplest cure for odor is to run the air conditioning in fresh-air mode to lower humidity in the evaporator plenum.

It's a sweltering day. You get behind the wheel, start the engine and hit the max a/c button. But instead of refreshingly cool air hitting you, there's a blast that smells like a basketball team's locker room after a hard game. Your car's air conditioning is suffering from what could be called "jock socks odor." And you're not alone—it's one of the most common complaints that *Popular Mechanics* receives from readers.

If you take the car in to a dealer, he'll tell you it's a fungus growth in the under-dash a/c case that holds the evaporator. That's the heat exchanger that pulls the BTUs out of the passenger compartment. His solution is an expensive cleanout of the a/c with disinfectant. And after you've spent a lot of money, the smell is gone—for a while. It may be back before the air conditioning season is over.

Is there a less costly and more durable way to fix it yourself? Can you prevent the problem from occurring in the first place? The answer to both questions is yes.

The price of prevention is free, and the most important step doesn't even require a tool. Just don't use max a/c. If you have a car with an air-recirculation switch or lever, just move it to the outside-air position **(Fig. 1)**. The evaporator case is a humid environment because it's where the air is dehumidified (as well as cooled), and when you shut off the engine, heat builds up inside the evaporator housing. A hot, humid environment with stagnant air is the perfect setting for fungus growth.

The recirculation switch closes off outside air, and causes the a/c to merely recirculate the air in the passenger compartment. With little or no entry of hot outside air, the interior gets cooler faster and remains much colder. It's okay to use recirculation to help with the initial cool-down on a sweltering day, but after a few minutes, you should shift out of recirculation to bring in fresh air.

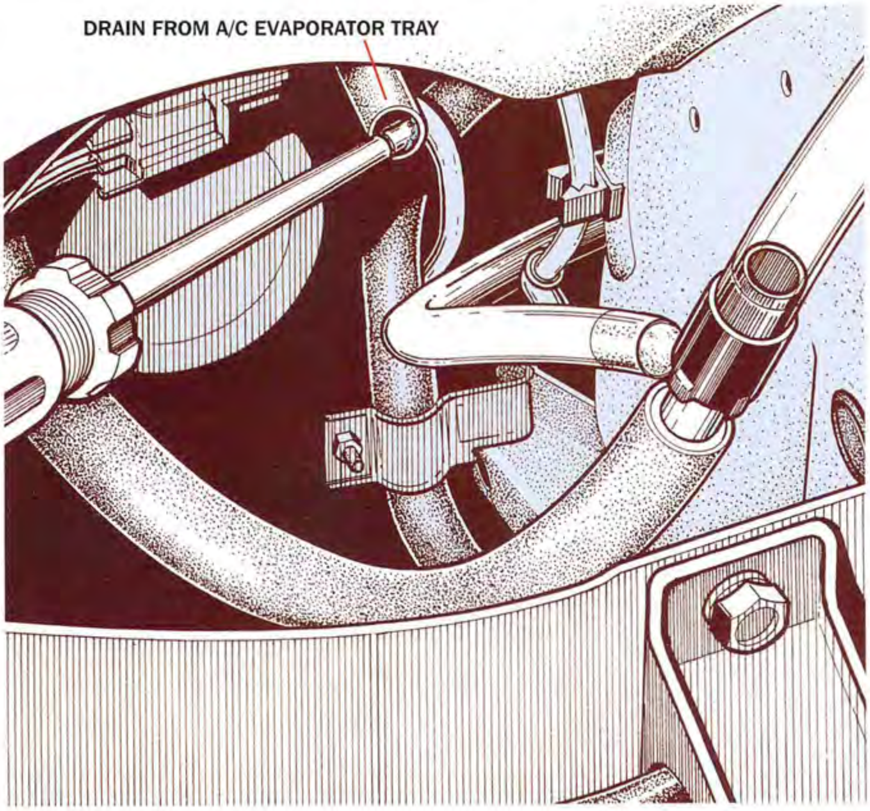
DRAIN FROM A/C EVAPORATOR TRAY

(Fig 2) An evaporator-tray drain that's plugged with dirt and debris might keep moisture from draining onto the ground.

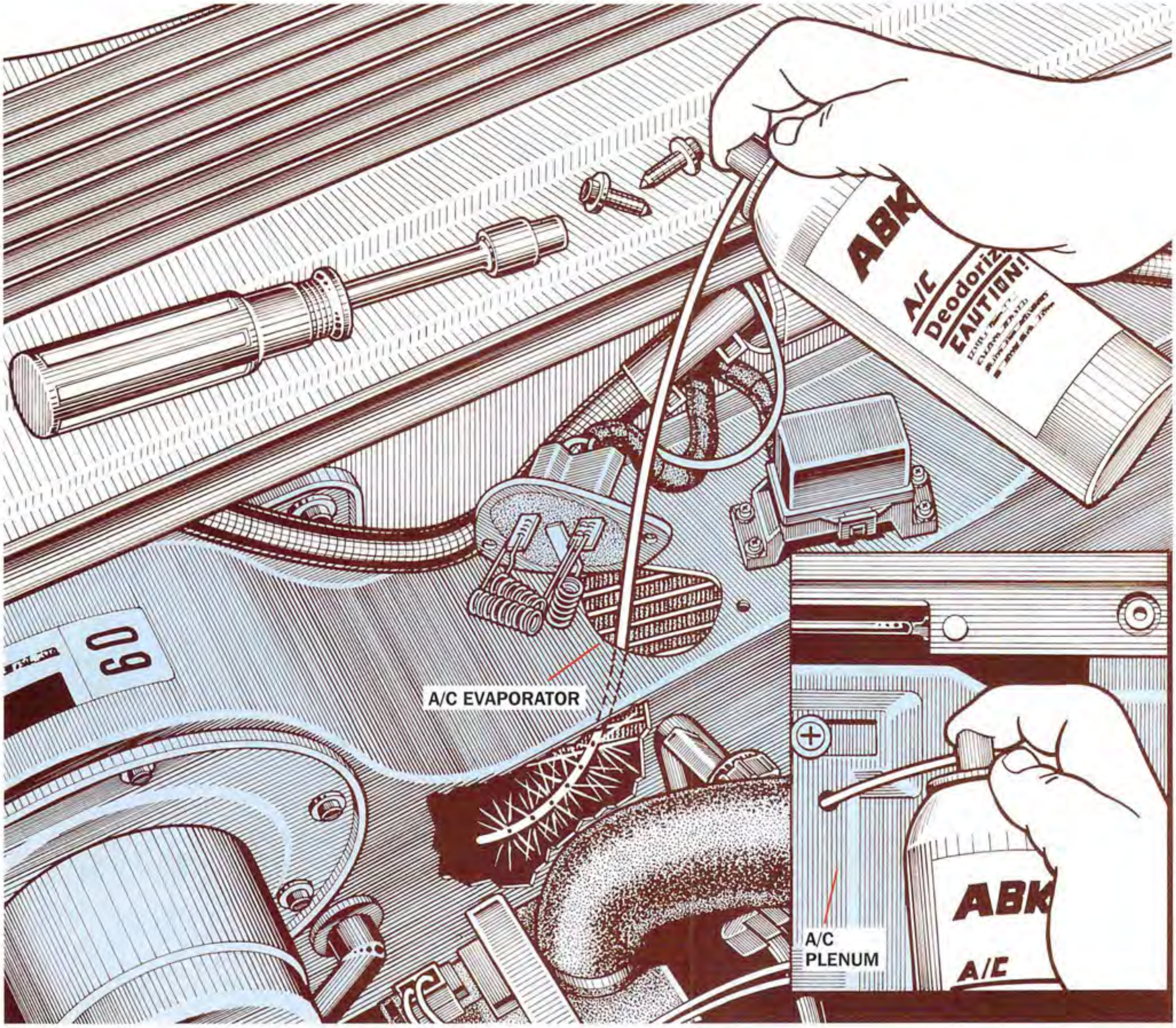

(Fig. 3) There are effective and ineffective cures for a/c odor. This aftermarket disinfectant works when sprayed directly onto the evaporator core. Before spraying, tape over the blower resistor opening (illustrated), making a small opening for the straw.

Many manual and automatic temperature-control systems are in recirculation when in max a/c or at the lowest temperature on the automatic system's dial. In this instance, you should also shift to a slightly warmer setting.

Garbage Collection

The next step is to lift the hood and inspect the air intake at the cowl. If it's packed with leaves and other debris, you may have a miniature compost pile forming. The odors from it are drawn into the passenger compartment by the blower fan. These odors are equally unpleasant, and you will need to dig out the debris or clean it up with a shop vacuum and wash out the area to get rid of them.

After you've cleaned it, remove the air-intake grille, insert a fine window-screen mesh underneath, seal the perimeter with plumber's putty, then refit the screen. Periodically check the screened area and remove debris to ensure a fresh air supply. Most late-models have an intake air filter that removes smoke and pollen from the outside air. These filters also keep out debris and may have a replacement interval of typically one year or 15,000-20,000 miles.

When you're doing an oil change or other regular maintenance under the hood, inspect the evaporator-tray drains **(Fig. 2)**, which are one or two hoses on most vehicles. If the hose ends are plugged or badly damaged, the moisture that the evaporator wrings from the interior air will not drain properly, leaving a stagnant pool for fungi. Dig out any road film with a thin screwdriver. Many hose ends have built-in check valves, which should be carefully inspected for damage. The ends are specially shaped, so don't trim them flat with a razor blade. Open the check valves gently, and if the ends are not damaged or plugged, leave them alone. If the ends are damaged, however, the best route is to replace the hose.

The next step is at the evaporator case. But before you do any work, check with the car dealer to determine if there's a specific factory solution that goes beyond disinfecting and other steps we'll be covering. For example, many 1985-94 Cadillacs have evaporators with such tight spacing between the fins that the evaporator holds too much moisture, increasing fungi growth. General Motors has replacement evaporators that may decrease cooling a few degrees, but drain moisture more readily.

Late-model Chrysler products since 1992, particularly minivans, have been known to suffer from air-conditioning odor. It doesn't smell like dirty socks, but it is offensive, and the cause is an overly thick coating of a corrosion treatment. Here again, the only fix is a new evaporator. See your car dealer.

For the overwhelming majority of cars, the treatment is to disinfect the evaporator case. Some shops use expensive professional equipment that takes several hours to use correctly. It's the best treatment, but it's fairly costly. You can save some money and get very good results with a quality handheld disinfectant spray if you work carefully **(Fig. 3)**. A household disinfectant such as Lysol can be used, but it isn't made for this problem, and any results will be too short-lived to justify the effort. Listerine won't work either, despite the fact that it's often recommended by people who don't know any better.

There are several products you'll find in auto accessory stores, and

A FEW SIMPLE SPLICES are all that's necessary to graft the afterblow electronics into your car's harness. Newer kits come with twist-lock connectors.

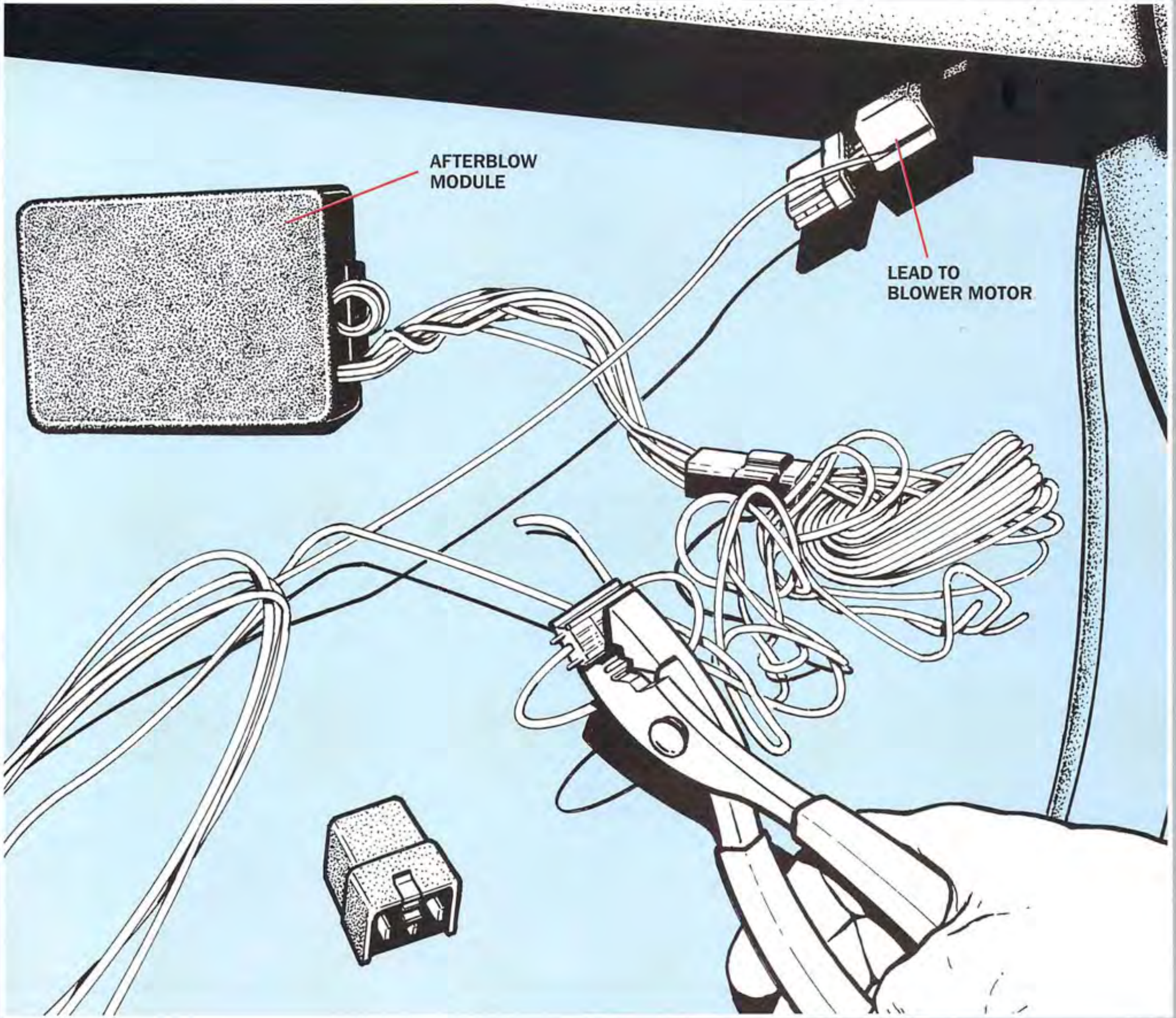

those with which we're familiar seem to work to some degree. The only one we know that has a successful track record is an expensive aerosol (about $25-$35) called ATP Clean 'N Coat. This is sold under other labels, and nationwide by GM as AC-Delco air-conditioning deodorizer (Part No. 15-102). The aerosol includes a special straw with strategically placed pinholes that produce a fogging spray, which thoroughly saturates the evaporator housing and all of the fins.

The basic approach to using the disinfectant begins by disconnecting the air-conditioner compressor. Just remove the wiring connector (typically two wires) between the clutch and the compressor body.

Next, run the system on high blower speed, in maximum heat and in recirculation (outside air shut off). If you have an automatic temperature-control system that seems to refuse to do this, override the automatic functions and pick a manual mode. Let the blower run on high speed for at least 20 minutes. This should dry out the evaporator case.

Find a spot to insert the disinfectant spray can's straw between the blower fan and the evaporator. Don't try to spray into the cowl air intake and rely on the blower to blow the disinfectant to the evaporator. If you do, virtually all the disinfectant will stick to the blower wheel.

On most cars, you can remove the blower resistor. If you look inside, you should see the evaporator, which vaguely resembles a radiator. Tape over the resistor opening and make a small hole for the straw. If you can't get to the blower resistor, find a suitable location between the blower motor and the evaporator (you may have to check a service manual to see the layout of the HVAC case). Drill a 1/8-in. hole in a safe place between them. Be careful not to drill in the wrong place or too deep, or you risk the danger of drilling into the evaporator or even the heater core.

Attach the straw to the can, insert the straw and hold the can relatively upright. Restart the engine, switching the a/c to outside-air mode, maximum heat and the highest blower speed. Spray in one direction, then the other. Just keep turning the nozzle and straw back and forth, spraying in each direction until the can is empty. Then stop the engine, let the disinfectant soak in for 30 minutes, then restart and run the blower for another 20 minutes, fan on high, temperature on maximum heat, system in recirculation. Finish up by shutting off the engine and reconnecting the wiring to the compressor clutch, then reinstalling the blower resistor or sealing the 1/8-in. hole with plumber's putty. This disinfectant is toxic, so don't leave the cat, the kids or delicate houseplants in the car while you're doing this. Opening the garage door or working outdoors is a good idea, too.

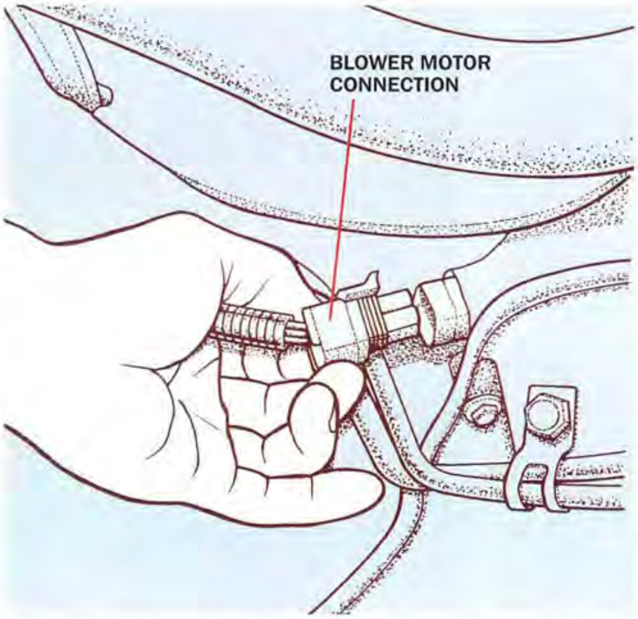

(Fig. 4) Start installation of the afterblow module by finding the power connector for the air conditioner's blower motor.

The Afterblower Solution

Now let's talk about a long-term fix: the afterblow electronic module. Its strategy is to run the blower at high speed for about 5 minutes an hour after the ignition is turned off—if the a/c has been on for at least 4 minutes. General Motors developed it as a factory solution for Cadillac and found that it worked so well at drying out the evaporator case that the electronic feature was incorporated into the a/c computer software on some Cadillacs in 1994. The kit was redesigned and released for other GM models. If you find a GM kit for your GM car or truck, it comes with connectors and a wiring schematic customized for your specific model vehicle.

Then, someone got the bright idea to use the module in a generic kit for any make. So Air Sept, AC Delco and others are selling it with an all-makes approach to the wiring and a batch of screw-lock connectors. You can install it in almost any post-1997 car.

Wiring

Just locate the blower motor's wiring connector (**Fig. 4**) and, with a grounded test light, isolate the current-feed terminal if there are two wires. The module's wiring is very straightforward, and when it's installed, the module turns on the blower at the appropriate time in high ambient temperatures when a/c operation normally would occur. It really is easier than it sounds. The most difficult part will probably be finding a good place to sinstall the module, which is held by a Velcro strip.

And when you're done, you'll not only enjoy cool fresh air, but the sweet smell of success, too.

Low-Tech Work on High-Tech Cars

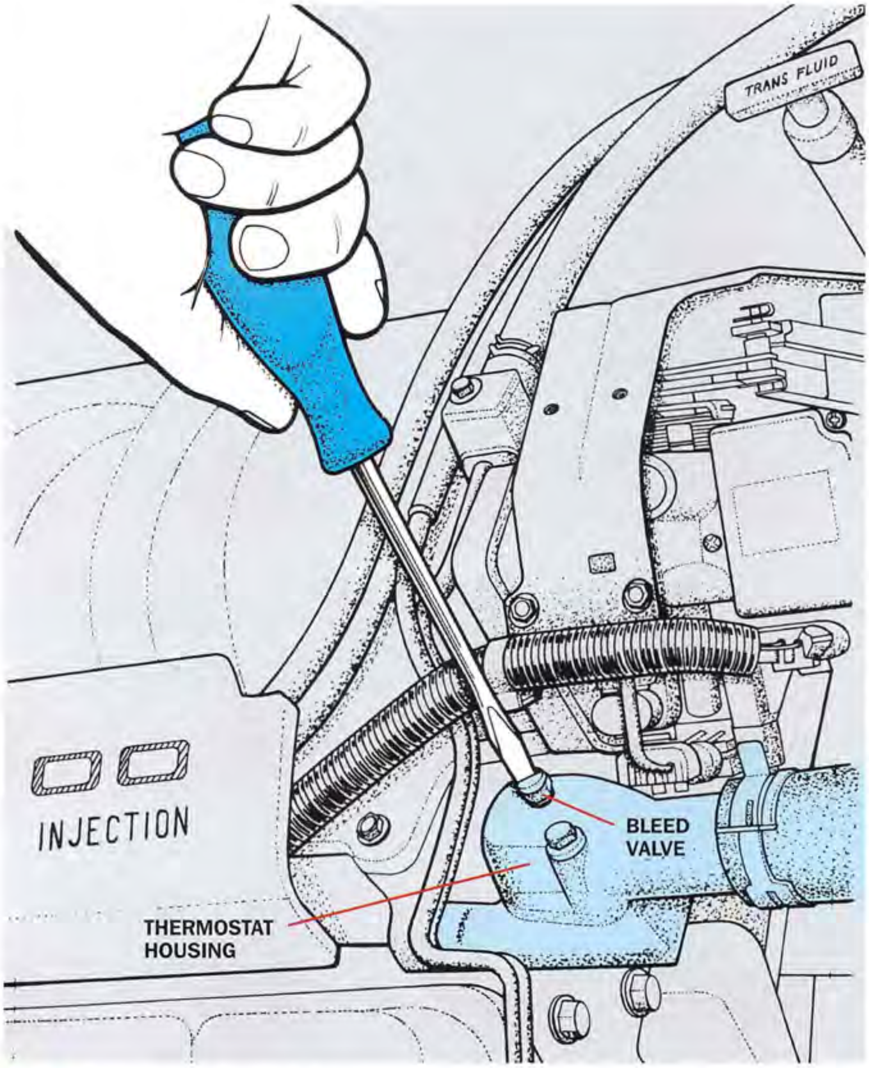

(Fig. 1) Many late-model cars with low hoodlines require bleeding the cooling system, not simple topping up.

You've finally gotten the new car through the warranty period, and the dealer's ministrations aren't necessary—or affordable—any longer. Time to start performing some of that maintenance yourself, a daunting proposition on most computer-controlled technorides. Without returning to college for degrees in engineering and computer science, how's a Saturday mechanic supposed to stay ahead of the maintenance curve? Not to mention having to buy a shop full of tools and test equipment.

Fortunately, most of the work necessary on your car is "low tech." It's routine maintenance such as replacement of lubricants and parts, basic electrical diagnosis, brakes, shocks and struts, spark plugs, cooling systems and belts, and gaskets. It's all the kind of work you can do without a major investment in equipment and shop manual "book learning." However, there's so much in today's cars that you can't avoid it, even if you stick to the basics.

Gaining respect for the sensitive items and learning how to avoid mistakes that produce problems is a learning experience (although less involved than the high-tech work itself). Here are the important beginning steps.

Strapping In

Draining the cooling system and refilling it with a fresh mixture of antifreeze and water is the No. 1 do-it-yourself job, and it's worth doing right. That means opening the engine block drain (if there is one) as well as the one on the radiator and getting out all the old coolant. You then face the problem of completely filling the cooling system. Because the radiator or surge tank neck on many cars isn't the high point of the system, air is trapped elsewhere. Even if it is the high point, there are nooks and crannies in modern systems, many of which were developed with sophisticated computer programs for an efficient flow pattern.

Start by jacking up the front of the car, to help expel air. Then fill the system as well as you can and start the engine. You have to either crack open the little bleed valves **(Fig. 1)**—and some systems have several—or you may have to remove a plug. Even with these precautions, you should check the coolant level repeatedly over the next few weeks and keep topping up as necessary. If you don't, the engine will feel the difference in summer and you'll feel it in winter, when the heater output is low. If there's a really big bubble trapped in the system, the engine could overheat. Check your service manual for specific bleeding instructions for your car.

Need to replace a drive belt? The new multi-V ribbed type is almost universal. If you get the right belt (not just one that's close), follow the

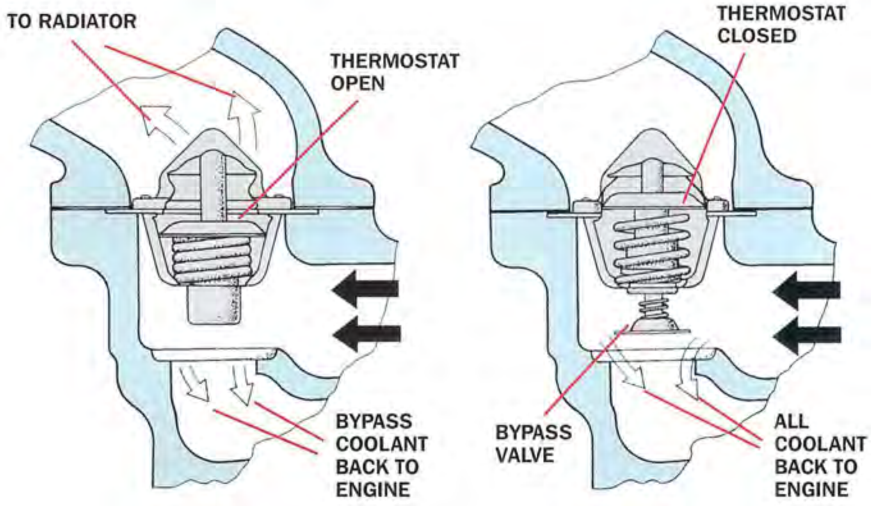

(Fig. 2) Bypass-style thermostats force all the cold water back into the block during warmup and all of the hot water into the radiator at running temperatures. If you use one that doesn't have the needed bypass valve (as at left), coolant flow will not go to the radiator for proper cooling.

underhood decal's routing diagram and make sure the belt is properly aligned in all of its grooves, and it will transfer power smoothly.

If you have to change a thermostat, don't just get one that fits. Many late models have a stat with a bypass valve, intended to make *all* the hot water—not just most of it—flow through the radiator (Fig. 2). If you replace the thermostat with one that doesn't have a bypass valve, the engine will warm up slowly, and it may run hotter in summer and heat poorly in winter. A thermostat also may have a jiggle valve, a pinlike device that balances pressures in the system and helps bleed air.

Even if you get the right thermostat, you have to install it correctly so it operates normally. On many Japanese cars with thermostat jiggle valves, for example, there's a specified position for the valve, such as being aligned with a tab at 12 o'clock.

The gaskets that seal the thermostat housing and water pump also are very different from the fiber/paper type you may be familiar with on older engines. In fact, new gasket designs are all over the powertrain.

Gaskets

The simple precut gasket is still around, and it may even be a good choice on older cars for replacing formed-in-place silicone, which can smear during installation and therefore seal poorly.

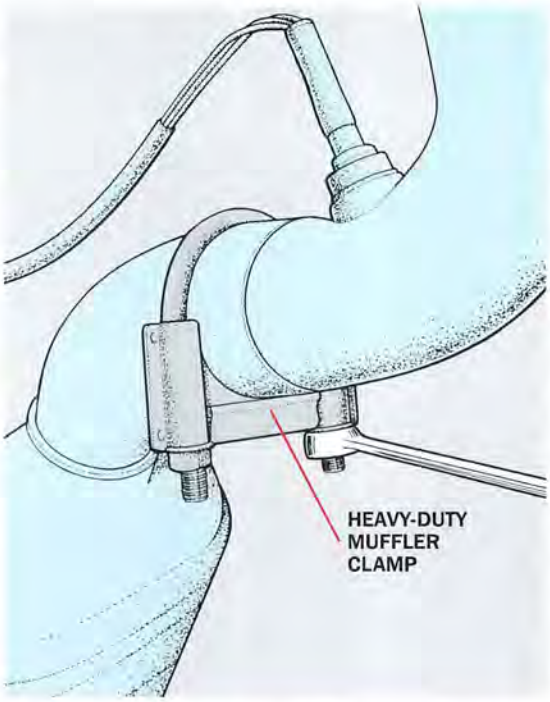

(Fig. 3) Stainless-steel exhaust systems necessitate heavy-duty clamps to properly seal pipes.

However, more of the gaskets used today are high-tech designs that resemble giant O-rings. Some of them have a round cross section, others are oval or even an indescribable engineered shape that is designed to seal when clamped down a precise amount. Still others, which resemble conventional gaskets, have torque-limit buttons, which are metal or hard plastic rings for the bolt holes, to resist over-tightening. Others use a hard-plastic carrier to carry an engineered-shape rubber sealing ring, as shown in the illustration on the following page. Intake manifold gaskets are traditionally difficult to seal on V-shaped engines, especially when there's a combination of aluminum cylinder heads and a cast-iron block.

You could use a conventional gasket in place of these high-tech wonders, and it will cost less—but it had better be made by a reputable gasket company specifically listed for your car's exact year and application.

Usually the torque required to seal a gasket properly is a lot less than you would think feels right, and if you overtighten, you distort the O-ring and it will leak. You have to use a torque wrench to be sure.

Exhaust System

Today's high-temperature stainless-steel exhaust systems are anything but cheap, and they're either welded together or held by superduty clamps (Fig. 3). If you have to cut a weld to replace a pipe or muffler, get a high-quality clamp to ensure it holds the joint on today's heavier pipes. You may have to shop around, or even place a special order, because many parts stores stock only the inexpensive clamps with bent metal saddles. The premium clamps have welded or double-crimped saddles and are probably two or three times the price of the more common variety, but they are the only kind that can take the 60 ft.-lb, or more of torque needed to clamp the thicker pipes.

Brakes

If your car has antilock brakes, be careful not to damage the wheel sensor when you pull a caliper for brake lining work. In fact, it's a good time to inspect the sensor connector for physical damage and the trigger wheel for badly chipped teeth. You may find a reason for misbehavior of your high-tech brake system that's repairable with some pretty common do-it-yourself parts replacements.

Bleeding brakes at the wheels after you replace a caliper should not be a problem with antilock brakes—just stuff a tapered plastic plug into the brake hose end when you disconnect the caliper, to help keep out air.

However, replacing the master cylinder might be a special problem. If it's physically separate from the antilock hydraulic components, you can bleed a master cylinder very well on the bench (or in some cases on the car, just before you reconnect the brake lines)—there are special kits available in auto parts stores for this. And when you install the master cylinder on the car or reconnect the brake lines, the pedal should be firm.

THE HIGH-TECH ENGINEERED SHAPE of this intake-manifold seal does the sealing, while the plastic carrier plate holds the seal in place and ensures that it's properly crushed. The complex shape of the end seal is the result of similar high-tech engineering.

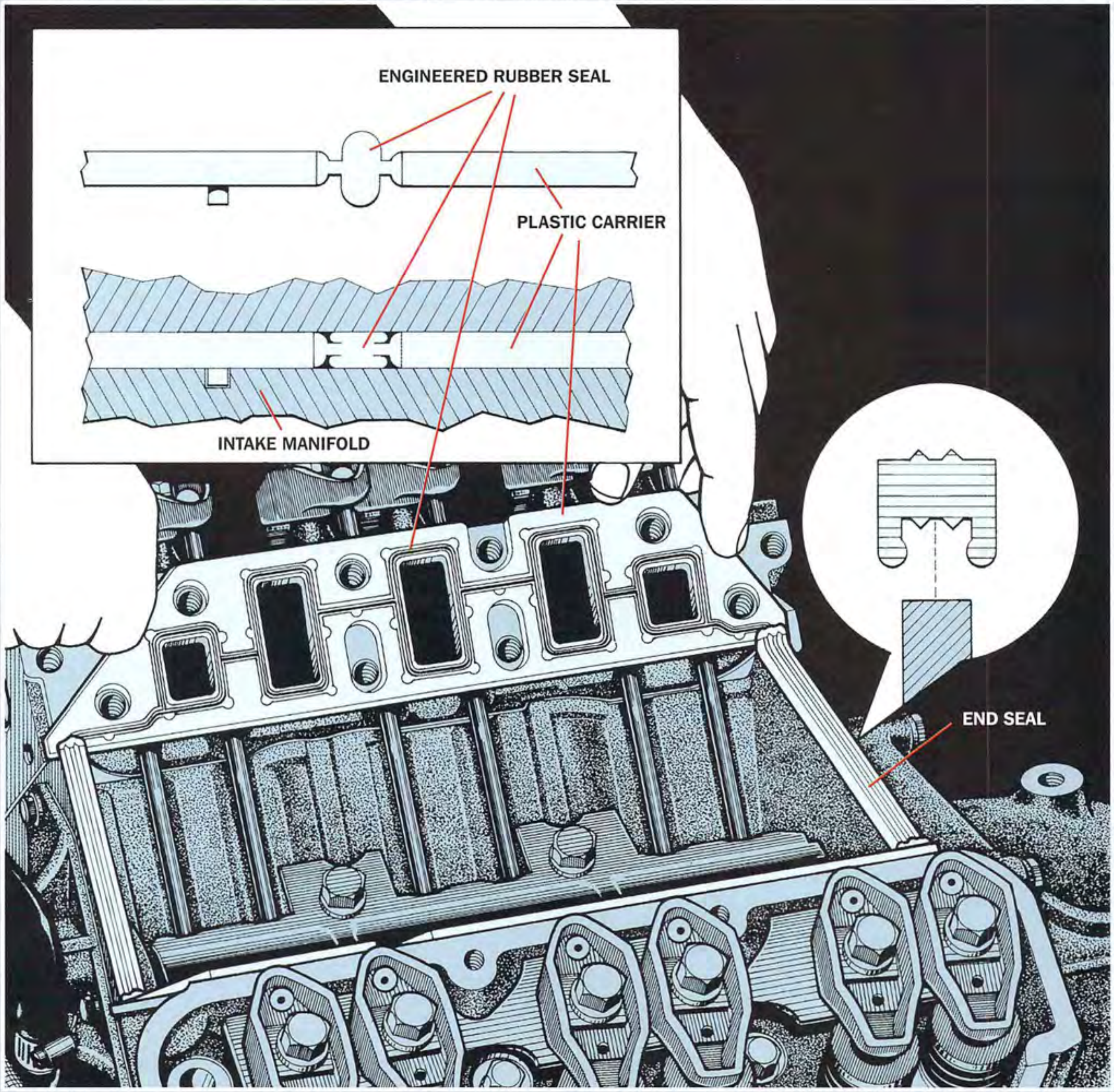

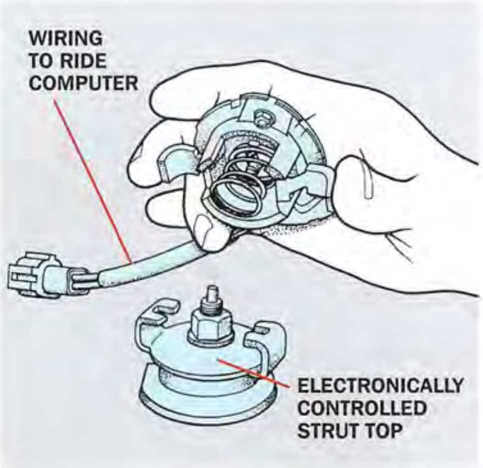

(Fig. 4) Replacing ride-control shocks doesn't necessarily mean replacing the actuator.

Just plug any hydraulic lines you disconnect to keep out as much air as possible when you start the job.

If the master cylinder and antilock system are a single assembly, you may have to follow a specific shop manual procedure to bleed it. Sometimes this is d-i-y doable, and sometimes it isn't. For example: The master cylinder/antilock assembly on many Fords requires a special procedure. But, the late-model Taurus/Sable also calls for a special electronic switching device and a breakout box (a special adapter for testing circuits) to bleed the single assembly which (for now) makes it a job for a dealer.

Shocks and Struts

Those computer-controlled shocks may have seemed like a neat feature when the car was new, but they may cost hundreds of dollars a pair at replacement time. If you want to maintain the electronic control, on many cars you can transfer the electric motors that operate the originals to a pair of new shocks and save some money (Fig. 4). Even with switching over the motors, you can be in for major expense with electronically controlled shocks, and it may not seem worth it on an older car.

Fortunately, on most cars you can unplug the computer at a connector in the trunk and then install conventional shocks. There are some cases where doing this confuses the car's electronics. This problem afflicts late-model Thunderbirds and Cougars with Programmed Ride Control, early Mazda 626s with Auto Adjust Suspension (but not the later models), Nissan 300ZX Turbos and more exceptions to come. We suspect, however, that there will be computer bypass kits available eventually.

Tuneup and General Electrical

Tuneup and general electrical are areas where you have to be careful, but if you are, you should not let fear of smoking some expensive component keep you from doing work. Here are a few guidelines to follow:

- When you have to disconnect a fuel line on most Ford and Chrysler products, you'll generally find it doesn't have a threaded fitting, but a quick-connect type. It takes a special but inexpensive tool, available from auto parts stores (Fig. 5). Don't try to pry the fuel fitting apart, or you'll damage it. These fittings often have O-ring seals, which should be replaced. Be sure to get these in a packet for the exact application. Ford, for example, uses spring-lock couplings on both fuel and a/c lines. The O-rings are not only a slightly different size, but a totally different rubber. Fuel-line O-rings won't live in a/c and vice versa.
- If you have to disconnect a wiring terminal from a plastic connector, such as for an electrical test, you can't just shove in a small screwdriver blade to disengage it, or you could damage the terminal and create a problem where none existed. Virtually all late-model cars have special terminals and plastic connectors for positive retention, and these usually require special tools. The pick-type tools for General Motors and many Ford cars are available from auto parts stores, but you could have difficulty with other makes. In '93, Chrysler switched to a new line of connectors that require a half-dozen special tools that we still haven't seen in auto parts stores.
- If you have to disconnect a battery cable, consider the use of a computer memory saver. It not only will keep the time on the electric clock and hold the preset radio stations, it will maintain the long-term calibrations of the powertrain computer. If you lose these, the car's driveability may be poor until the computer "relearns," which could take 100 or more miles of driving.

There are two basic types of memory savers. The 9-volt battery with one type is okay, but just a small voltage drop could kill the memory. So you must keep all electrical accessories and lights off (that means door domelight switches, too, so either keep the doors closed or install hold-closed clips on the switches). An alternative is a saver that hooks up to a 12-volt battery. If your battery is okay, just carefully make the connections as you dis-

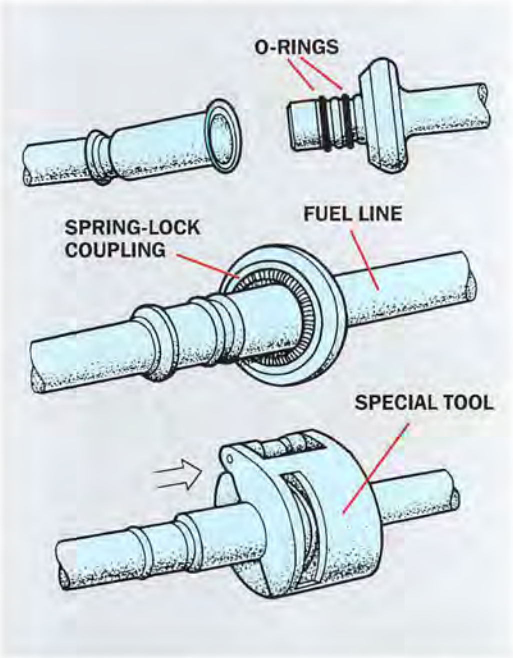

(Fig. 5) Spring-lock fuel and a/c connectors need a special inexpensive tool to open them.

connect the cable. Or use the battery from another car or a lantern battery.

- Use a computer-safe test light. A standard 12-volt light may allow enough current draw through a circuit with electronics to measurably shorten the life (or even fry) some solid-state components. Computer-safe test lights start at under $15, and even the best—which plug into a cigarette lighter socket and identify both power feed and ground—are under $60, including one that has a built-in flashlight (for probing under the dash) and a buzzer if you're probing a power feed in a blind spot.
- The test meter should be a "high-impedance" type (10 megohms or greater), for the same aspects of safety for electronic circuits as the test light. Name-brand multimeters may go over $500, but there are suitable volt-ohmmeters that start at under $50.

If you have to probe a circuit and want to do it safely, one way is to buy a "bed of nails" test lead that makes mini holes in a wire, and when you disconnect the lead, they self-heal. You can pierce the insulation with a dressmaker's hemming pin, but it may not make good enough contact with the copper wiring inside.

- Don't reposition wires underhood. This applies not only to spark-plug wires, but to any harness, including a battery cable (particularly if you get a replacement that's a different length from the original, or is outside a harness and is not easy to route the same way). The reason is the high-voltage electricity from spark-plug wires may produce electrical interference in adjacent low-voltage circuits, particularly if they contain electronic components. Because many electronic components operate on 5-8 volts, even the sudden flow of 12-volt high-amperage electricity through a battery cable during cranking can cause big trouble.
- Never disconnect or connect a circuit, or an electronic module, with the ignition key in the ON position. In fact, to avoid static electricity damage, unplug an electronic module with one hand held to a body ground. Static electricity that discharges from your hand may not totally trash an electronic module, but it can shorten its life from years down to weeks or even days. Professionals who face this problem all the time may wear a grounding strap on one of their wrists.

As you get familiar with your car and read the manuals, you will be able to move into higher-technology service. You'll discover that it takes careful attention to the diagnostic procedures in the manuals, but it isn't rocket science, and it doesn't take a degree in computer science, either. The high tech you've already learned in doing the low-tech service will get you started.

Scan Tool Diagnosis

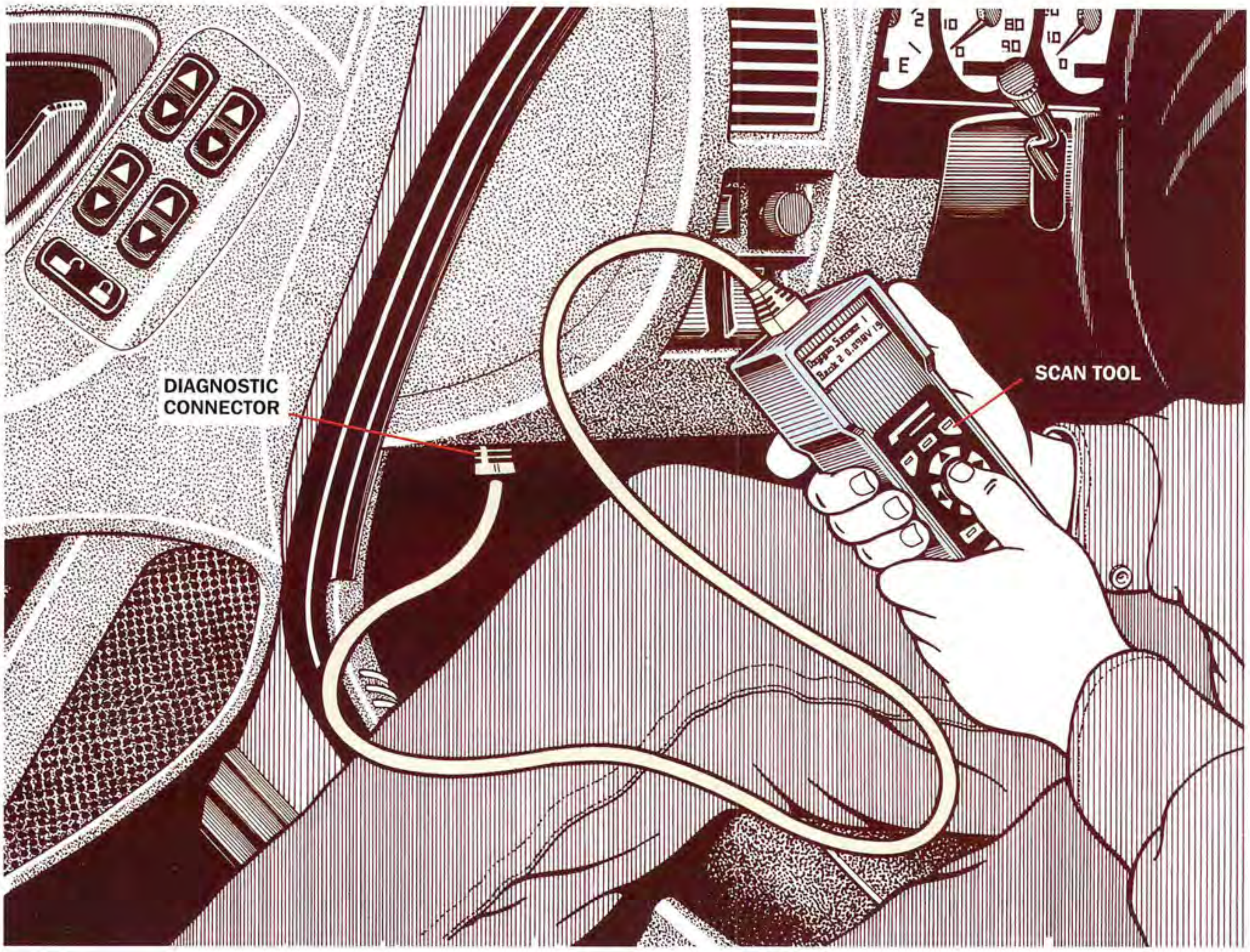

(Fig. 1) **OBD II scan tools hook to the car's engine-management computer via a connector under the steering wheel.**

It's state inspection time and you're ready to leave bright and early for the inspection station with the family's '98 minivan. You're starting the day in a less than jovial mood because your daughter said "the light came on" last night on the way home. She said she threw a few dollars of gas into the tank to make it home. Turns out your annoyance is misplaced. Sure enough, the light's on all right, but it's not the fuel warning, it's the CHECK ENGINE warning. You stop for gas anyway, and then you pull into the inspection lane.

Well, full tank or not, the minivan does not pass muster.

Failure Is Not an Option

This is a late-model vehicle with what's known as OBD II, the second-generation on-board diagnostic system that replaced OBD I starting in 1994. It's industry-wide and federally mandated. One of the problems with OBD II for the do-it-yourselfer is that you can't get trouble codes by counting the blinks on the CHECK ENGINE light like you can on earlier computer-controlled vehicles. You could take your minivan into a high-tech shop, where the minimum charge for diagnosis and inspection could run you somewhere into three figures. Or you can learn about OBD II yourself, but you will need a piece of diagnostic equipment that you probably don't have—an OBD II scan tool **(Fig. 1)**.

If the Check Engine light came on, it should be no surprise that the vehicle failed an emissions test. With OBD II, that light comes on only if there's a failure that significantly affects emissions. That makes the scan tool even more important since it will reveal a lot of the problems that do not cause the warning light to come on.

With many problems, the light

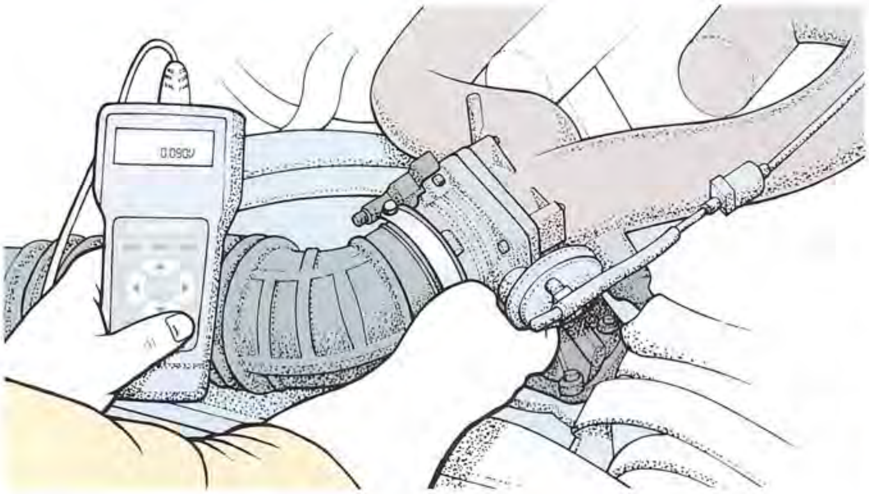

(Fig. 2) Datastream scan tools will let you investigate the functioning of most sensors, like the throttle position sensor being checked here.

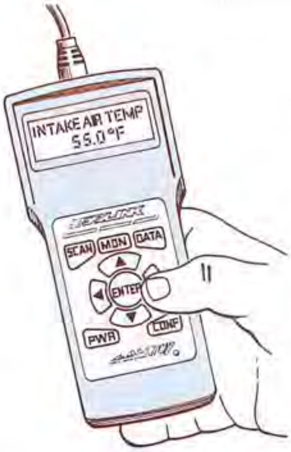

(Fig. 3) Units with more features will let you snapshot and store data while the car is being driven.

stays on after the repair is made and the code remains in the computer memory for a certain number of ignition on/off cycles. Your daughter didn't tighten the gas cap correctly, causing a code for the evaporative emissions system to set. Eventually the light will go out and the code will self-erase, perhaps after the next time you start and stop the car. You also can use the scan tool to erase it immediately.

With many other problems, however, the only way to turn off the light and erase the code is with the scan tool. Just a warning: If you erase trouble codes with a scan tool or disconnect the battery for any reason, you also erase the computer's continuous monitoring system. So if you take your car in for a state inspection before enough normal driving, the computer might not have completed all its tests, and your car will fail inspection for that reason.

The OBD II scan tester not only will enable you to find answers to the simpler problems, but will tell you into what areas the seemingly more complicated ones fall. Then you will have a better understanding of what the technician is (or should be) looking for.

Cracking the Code

With an OBD II scan tool you also can read a certain amount of engine operating data: typically rpm, ignition timing, fuel-injection calibrations, readings from a variety of sensors (such as the oxygen, throttle position, barometric and mass airflow sensors), a "calculated load" value and sometimes switch position signals (Fig. 2). OBD II also includes a "capture" mode, in which you can use the scan tool to take a "snapshot" of what the sensors were reading at the exact instant a driveability hiccup occurred (Fig. 3).

Sensor Scan

With enhanced scan tool capability, you can discover problems that do not trigger the engine-warning light. For example, we recently turned up a generic PO713 code on a late-model car. This is listed in the shop manual as "transmission fluid temperature sensor circuit—high input." If the transmission fluid is running very hot, the transmission could completely fail, and quickly.

A scan tool could save a lot of worry and effort if it has enhanced diagnostics. You can do much of the troubleshooting from the driver's seat the way we did. First we cleared the code with a simple push of a button on the tester. Then we rechecked: The code came back quickly. We cleared it again and looked at the reading that was coming from the transmission temperature sensor to the powertrain computer. It was 131° F, nothing abnormal.

The problem apparently was in the sensor or the circuit, possibly a bad connection. The wiring was good and when we started poking around, we found physical damage that clearly indicated the problem was a defective sensor.

So Many Choices

If you have a late-model vehicle, you have OBD II. However, just because it's generic, and the wiring connector from any OBD II scan tool will plug into your vehicle, doesn't mean any OBD II scan tool will work on your car. The Europeans are the problem, as the ones after 1998 require a software upgrade. Korean cars are also problematic, and how well they work with any given scan tool needs to be investigated on a case-by-case basis.

Further, new cars are being converted (on a phase-in basis starting with Saturn and Saab in 2003) to "CAN" (Controller Area Network). This is a new diagnostic "protocol" and most scan tools will need an adapter module, or complete replacement.

The OTC Mind Reader for OBD I can be upgraded with an additional chip to read generic OBD II in domestic, Japanese and earlier European models (but not the later Europeans).

The Actron ScanTool for OBD I can be updated to the same level of OBD II as the OTC Mind Reader with a plug-in cartridge (or you can buy an OBD II-only model).

AutoXray produces a programmable scan tool. Although it doesn't have the OBD I Chrysler command tests of the Mind Reader, it is the one home mechanic's scan tool we've tested to date that covers all generic OBD II models (including the Europeans) with more on the way and covers the new CAN diagnostic with just a software

THIS TESTER WILL let you check sensor outputs directly without using the vehicle's on-board computer.

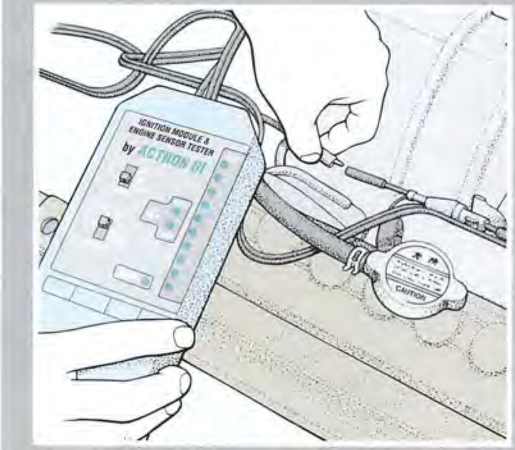
IGNITION MODULE & ENGINE SENSOR TESTER
CAUTION

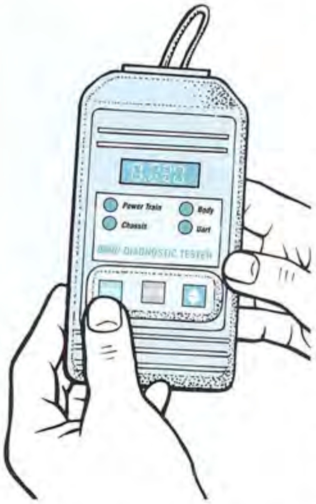

(Fig. 4) Simple code readers/ resetters like this one do not give you sensor data.

update. The software is sold over the Internet. You'll be able to store it on your personal computer and then download it to your scan tool via a cable available from the manufacturer.

Any AutoXray scan tool is designed to be upgraded electronically, from one-make OBD I coverage through the latest models. Although the professionals have had this software (and a lot more) in their scan tools, you have to wait for it in the consumer market. Other scan tools may be upgraded to enhanced status and beyond with new cartridges, CD-ROMs or via the Internet. Ask about upgrades before you buy.

Although Actron has a line of OBD II scan tools, its top tester for car owners is the Actron CP9087, a simple code reader with read-codes and code-erase buttons. You get no sensor readings or other data items (Fig. 4). It's a low-cost device (under $200) that comes with a good assortment of wire leads for making test connections—including a backprobe adapter that has a thin, curved metal terminal. This terminal lets the probe slip through a water-sealed connector to reach a wiring terminal for a test connection.

OBD II has been in use on all vehicles since 1998, so the earliest vehicles that have this system are well off warranty. OBD II is complex and we've given you just a basic introduction. The OBD II powertrain computer is getting a lot better at finding problems and logging codes. But, the computer won't tell you a thing unless you hook up a scan tool.

HOW IT WORKS

How OBD II Transmits Data

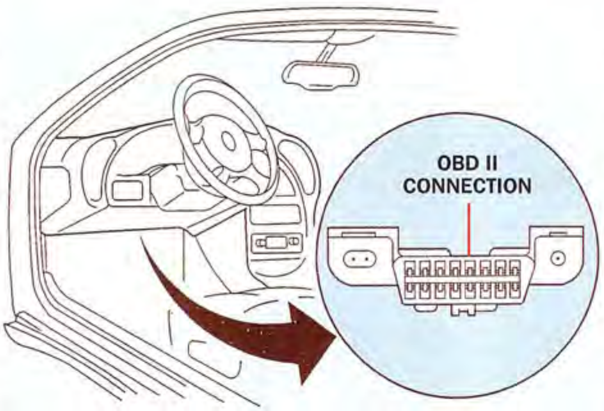

Today's powertrain computers are at the heart of a vehicle's communications network, circulating information from switches and sensors to the other computers that control antilock brakes, air conditioning, transmission, suspension and safety systems. The powertrain computer is also in charge of systems that affect engine emissions, so the information it processes has to be available for evaluation by a technician. That information travels along a wire to a standard 16-terminal diagnostic connector (although, generally, fewer than a half-dozen terminals are live in any given vehicle). Because manufacturers do not all use the same data transmission protocols, the scan tool must be programmed to recognize which one is being used. Fortunately, there has been some standardization, but we're not down to one transmission protocol for all, hence the problem with late-model European cars (including the Cadillac Catera). There are four different so-called standard data pins in the standard connector, and at least four different types of data transmissions that could be used, plus enhanced versions of the standard stuff (other pins are available for additional manufacturer-specific data, such as vehicle diagnostic systems for a/c and ABS).

As anyone with a household PC knows, "plug and play" doesn't always work, and OBD II scan tools can encounter compatibility problems. Some European and Korean cars don't always work when they're suppossed to. How can you tell? Checking a scan tool manufacturer's Web site for updates is the way to keep your tool current.

MODERATE

Wiring Your Trailer Hitch

The rental yard is busy. The utility trailer you want is inexpensive, and the yard help assists you in hooking up the trailer hitch and safety chains. Now to drive home and load the trailer with the entirety of a teenager's life, then ship her off to college several states away.

But not until you hook up the lights on the trailer. The trailer dealership won't let you leave until the lights work. And the trailer plug on the back of your truck doesn't remotely match the one on the trailer. The friendly trailer-hitch counterman points to the display of electrical connectors and suggests that if you can't get them working on your own, he can have a mechanic do it for you, first thing Monday morning.

Desperate Times, Desperate Measures

Trailer wiring may be easy. Many vehicles, particularly pickups and SUVs, come prewired for trailers. If there's a preinstalled hitch, the connector might already be in place behind a convenient spring-loaded cap. Or, if you've bought a new vehicle, you may find the harness needed to install the connector inside the glovebox. Simply unplug the connector to the rear lights on your truck, plug in the trailer harness between the male and female ends, and you're ready.

But wiring a trailer may not be easy. If you're towing just for the day and don't ever plan to tow again, you can pick up some adapters that replace the taillight bulbs in your vehicle and have pigtailed wires hanging from them. How you route the wires out of the light sockets and down to the hitch can be problematic. And what if you have a trailer with simple 1-lamp lights and a car with multiple taillights, like most European cars? At the very least, odds are the trailer and vehicle use different connectors, as the trailering industry uses at least four different styles.

Red To Green To White

Let's start with the basics. Most trailers

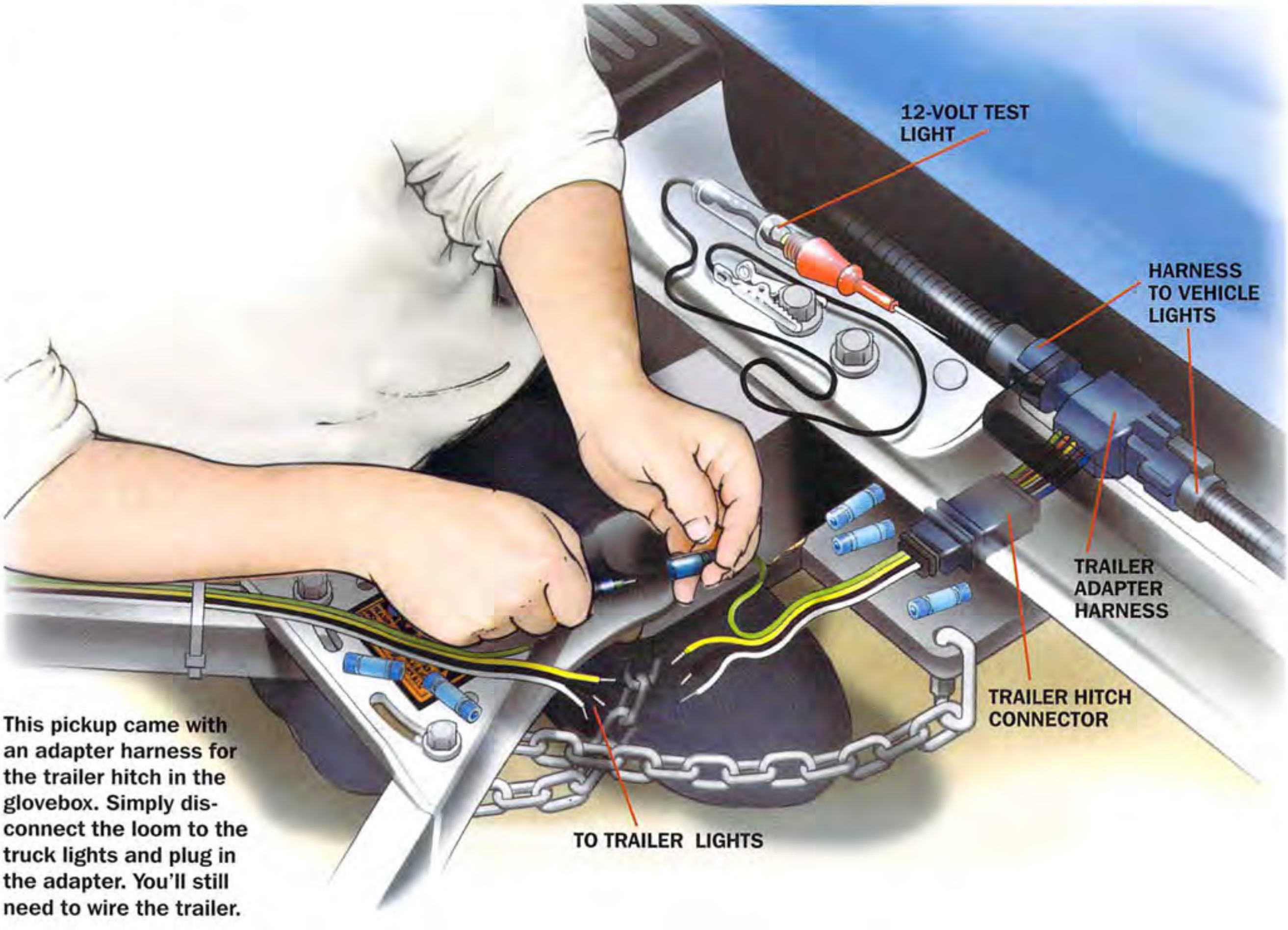

This pickup came with an adapter harness for the trailer hitch in the glovebox. Simply disconnect the loom to the truck lights and plug in the adapter. You'll still need to wire the trailer.

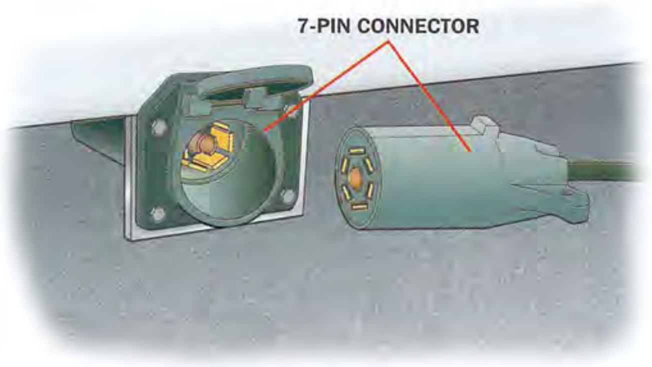

This standard RV-style connector has high-amperage-capacity flat pins and room for six circuits.

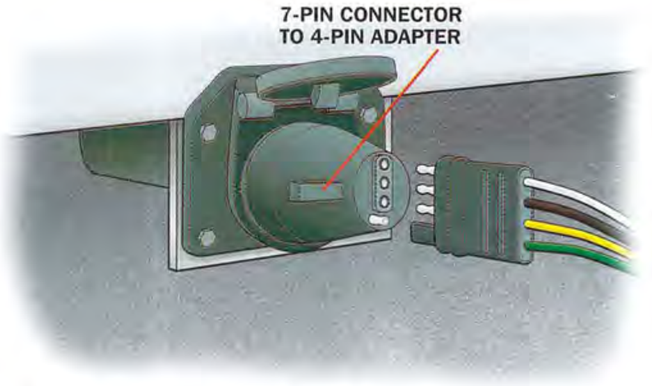

A simple adapter can make 5-, 6- and 7-pin connectors work with other styles of connectors.

have three circuits—for running lights, and left and right brake lights. The brake lights will flash for the turn-signal function—it's up to the flasher relay in the car to know when the brake lights are on and to flash the appropriate filament. So technically, you need only three wires to the trailer plus a ground wire.

Always run a ground wire from the frame of the vehicle (not from the hitch or bumper) to the trailer. The metal-to-metal contact in the hitch is not reliable enough for a consistent connection.

This means you'll need a connector with at least four contacts. Not surprisingly, the standard flat-style connector for trailers uses four pins, with one unshrouded male pin on the vehicle end for the ground. Other common styles of connectors use five, six or seven pins, and these are the styles you'll see installed by the manufacturers on new vehicles. Why more than four? Larger trailers sometimes use a separate circuit for running lights on the sides and front of the trailer. Some trailers use electrically actuated brakes. It's not uncommon for enclosed trailers to have interior lights or accessories powered by the vehicle battery, so that means another circuit.

Do not freak. As long as you concern yourself with only one circuit at a time, this job will not require you to read those schematic diagrams that have tiny little wires labeled with type too small for even a lawyer to read.

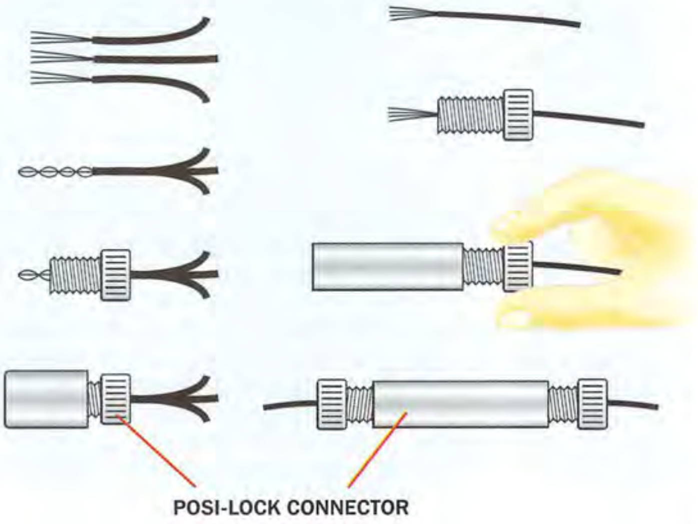

Posi-Lock or crimp connectors can be used without soldering, but aren't weathertight for long-term use.

Over-The-Counter

Got a 7-pin connector on the truck and an inline 4-pin connector on the trailer? We found an adapter at the auto parts store that makes this plug-and-play. Other permutations of connectors may be adaptable with off-the-shelf parts.

Splice It

Worst case—you have a car or truck with no provision for trailer lights, and a trailer with a harness that ends with chopped-off wires. You'll need nothing more complicated than a 12-volt test light or volt-meter, and some splicing capability. Start by turning on the vehicle's running lamps. Now probe the vehicle wiring to find out which wire leading to the taillights is hot. Splice into this wire and bring it to the vehicle side of the trailer connector. Which pin? The connector should come with a list specifying which pin handles which circuit. Follow it—but don't for a minute trust that someone else who wired your vehicle before you followed anything except his own whim. Tag each wire with its function as you identify it. Now you can turn off the running lamps and activate the left turn signal. Identify which circuit is the blinker, and repeat for the right turn signal. Wire all to the connector.

Separate But Not Equal

A few vehicles use separate lamps for the brake and the turn-signal functions

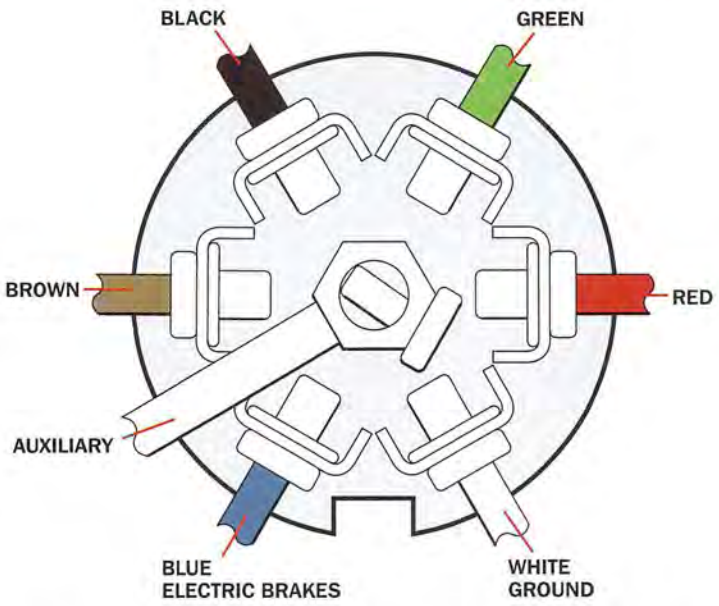
RV-STYLE 7-WAY CAR END

This wiring color code is supposed to be the standard, but vehicle manufacturers don't always follow it. Check with a test light or you risk blowing fuses.

of the rear lights. That's fine—until you realize the trailer uses a single 2-filament bulb for all three functions. Adding another taillamp housing to the trailer is one option, but a simpler solution exists. Your trailer-rental place, and most auto parts stores, sell simple electronic adapters that will combine the brake-lamp and turn-signal functions. It installs in the harness just before the trailer connector.

Other Circuits

Some larger trailers have a separate circuit for clearance or marker lights, which are separate from the running lights in the taillamp assemblies. You may need to add another circuit to handle them. If you try to wire them in parallel with the customary pair of running lamps, you may overload the vehicle's headlamp switch—so you'll need to add a relay to handle the extra current.

There also may be a 12-volt circuit for running lights inside the trailer or for charging the trailer battery in the case of a travel trailer. This circuit probably will be hot whenever the engine is running, but it may be controlled by a switch on the dash. Be sure the wires are of a gauge large enough to handle the load (a fully discharged deep-cycle trailer storage battery may draw 20 or more amps when you start the engine). The use of a battery isolator is strongly recommended. This will prevent discharging the vehicle battery from the trailer, and prevent surges of current from overcooking the wiring or blowing a fuse.

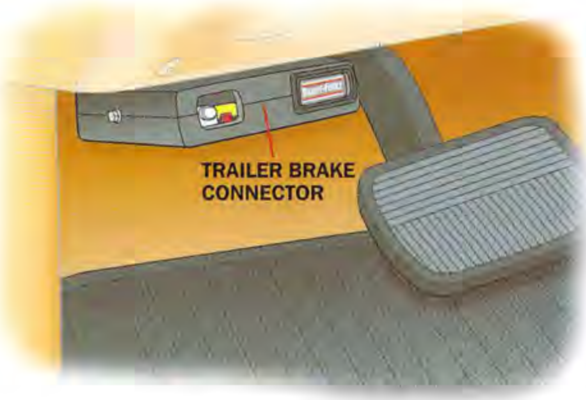

This controller allows the use of electric trailer brakes. Many vehicles are prewired to accept it.

Whoa, Nellie

This leads to the concept of electric trailer brakes. These are actuated by a controller in the vehicle that increases the voltage to the brakes proportional to the vehicle's deceleration. This requires a wire from the controller to the trailer brake. (This is why some connectors have as many as seven pins, to make provisions for the multiple circuits.) Use 12-ga. wire for this circuit. If you have electric trailer brakes, there must be an auxiliary battery and a breakaway switch in the circuit to slow the trailer if it comes loose from the hitch.

Keeping It All Together

How to make the connections? We've seen plenty of trailers lumbering down the road with blinking or dim lights. Odds are, if you looked at the wiring on one of these rigs, you'd see wires twisted together and insulated with electrical tape. Almost as bad are the automotive-style crimp connectors, which can't handle the vibration and moisture. Any crimp connector exposed to the elements will have a short life span. Don't even think of using household-style wire nuts—they'll unscrew themselves within a few hundred miles. We use the screw-in Posi-Lock connectors illustrated here for quick jobs. Our own trailer has every connection made with solder and PVC shrink tube.

As you button up, smear a film of dielectric grease on the hitch connectors to prevent moisture from corroding the pins.

DIFFICULT

Fixing Faulty Cruise Control

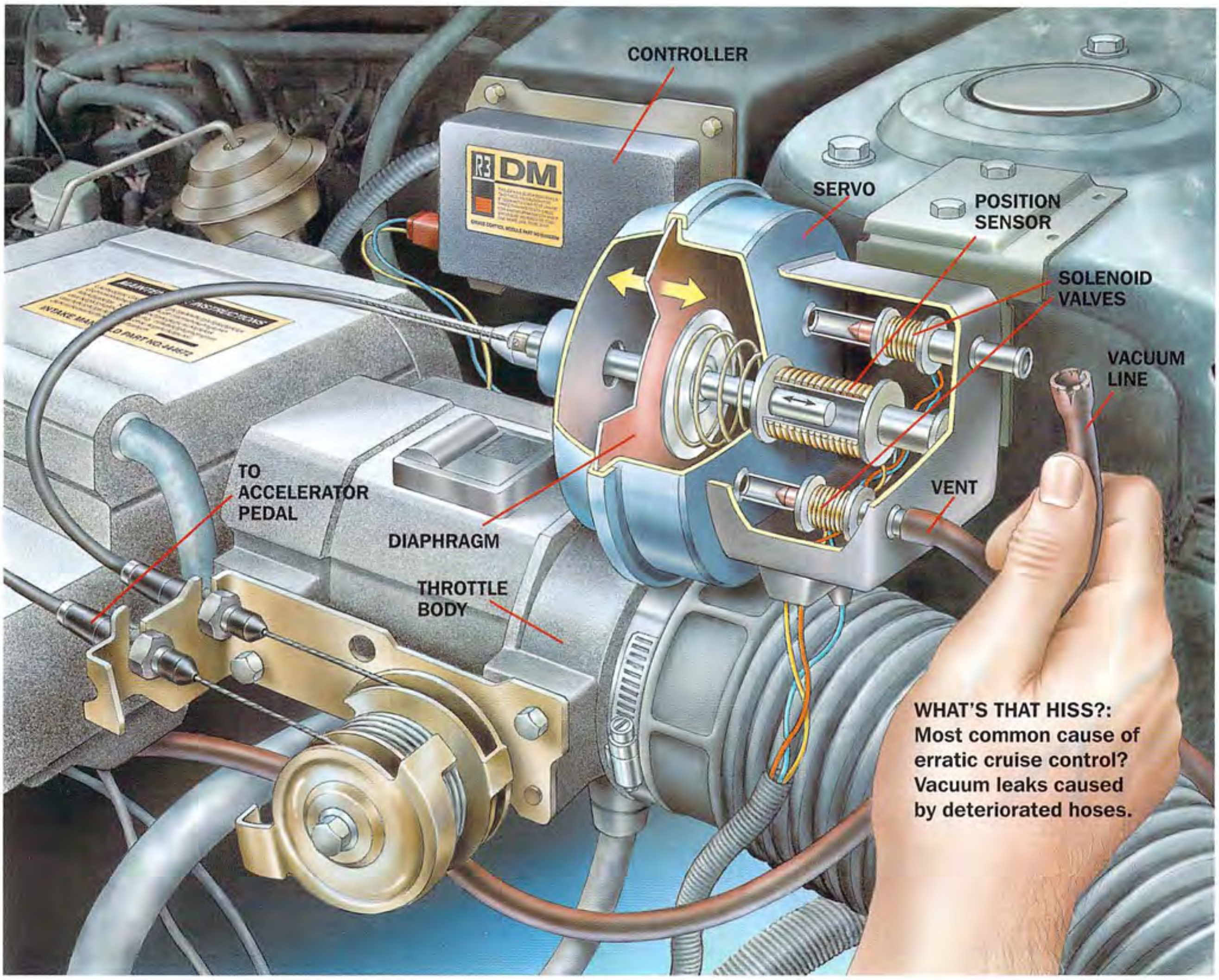

WHAT'S THAT HISS?: Most common cause of erratic cruise control? Vacuum leaks caused by deteriorated hoses.

Alarmingly, the speedometer needle steadily winds down from 75 mph toward 50. Just as you uncurl your feet and try to accelerate back to traffic speeds, the vehicle downshifts with a lurch and abruptly climbs back to over 80 mph. So you tap the brakes and disengage the cruise control to avoid a conversation with one of the many law enforcement officers lurking behind every other billboard. Toggling the Resume switch settles things down, holding to a legal speed on both the uphill and downhill sections of the interstate. The kids in the back seat have stopped threatening to throw up, too. Then you look in your mirror 20 miles later and see the lights. Red and blue flashing lights. You're doing over 85 mph and, odds are, Smokey isn't going to believe you have the cruise set to 70. Time to find out why your cruise control has a mind of its own.

It's not a bug, it's a feature

Does your cruise control fall out of engagement partway up steep hills? Actually, it will normally drop out if the engine has to work too hard, mainly because after a while there isn't enough vacuum left to pull in the servo after sustained near-wide-open-throttle. You'll just have to put your foot into it. Downshifting helps.

Does your Japanese car not remember the set speed after tollbooths? Unlike most American and European cars, some Toyotas and Hondas are designed not to remember their set speed if the vehicle speed falls below 25 mph, and you're supposed to accelerate the vehicle to your set speed and hit the Set button again. Annoying, but that's the way they were designed.

Do you have to ride the brakes on

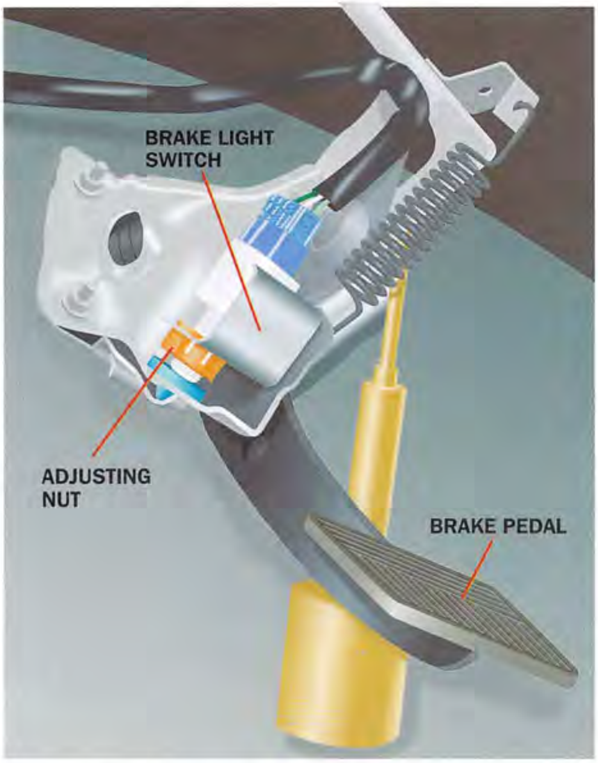

WHOA, NELLIE!: If your brake light switch is incorrectly adjusted, your cruise control may not lock in. A burned-out brake light can do the same.

longer downhills to keep from building up excess speed? That's normal too. The cruise control only has authority to reduce engine speed to idle. It doesn't activate the brakes. Modern cars, in an attempt to improve mileage, have very tall gear ratios, low-friction engine designs, low-rolling-resistance tires and optimized aerodynamics. That long downgrade outside of town may have accelerated your '60s-era Pontiac to only a couple of miles per hour above legal. But, it may well propel your new economy car to blatantly illegal velocities unless you intervene by braking or downshifting.

Smooth and steady

Cruise control on your vehicle is a terrific device. It substantially reduces the driver's workload on longer trips, and can save substantial amounts of fuel and expense over the life of the vehicle—until it stops working.

We'll get to the scenario of not holding a steady speed later, but here are a couple of things to check immediately if the cruise control is on strike.

Does the Cruise icon on the dash light up when you turn the switch on? Duh. Check the fuse. You may need to look in the owner's manual to see which one if it's not tagged on the fuse box cover. An aftermarket cruise may have an inline fuse holder in the wiring to the controller.

If there is power to the system, the next check is the brake lights. Brake lights? Yup, cruise controls have a switch to toggle them off when you touch the brake pedal, and many use the same switch as the brake lights. If one of the brake lights has failed, the cruise control thinks the brakes are on all the time and won't come on. Same result if the switch is incorrectly adjusted or broken or jammed. Wait, there's more—if your vehicle has a manual transmission, there's a similar switch on the clutch pedal. You may need to break out a test light or multimeter to verify the function of this array of switches. These switches usually are normally closed switches, and close their contacts when the pedal is depressed. We've seen several cases of intermittent cruise control dropout caused by a brake light switch that was adjusted very tight. Any small bump would jiggle the brake pedal down far enough to toggle the brake lights on for a brief instant—long enough to shut down the cruise. Adjusting the switch to specs (usually so the brake lights come on after the pedal travels ½ in.) fixed it.

Once you verify all of these things, it's time to look for more subtle causes. Check underhood. Inspect the linkage between the cruise control actuator and the throttle. It may be as simple as a toilet-tank bead chain or a separate throttle cable with its own actuator cam to the throttle body. Look for disconnected or damaged wiring to the controller or actuator, particularly if it's an aftermarket system, which may have been installed by someone with poor mechanical skills.

Another obvious failure point is the vacuum line to the actuator. If it has fallen off, there's no vacuum to the actuator and no force can be applied to the throttle. We'll get ahead of ourselves here and mention that a vacuum line that looks fine may

WHAT'S THIS THING?: The cruise control servo motor could be anywhere in the engine compartment.

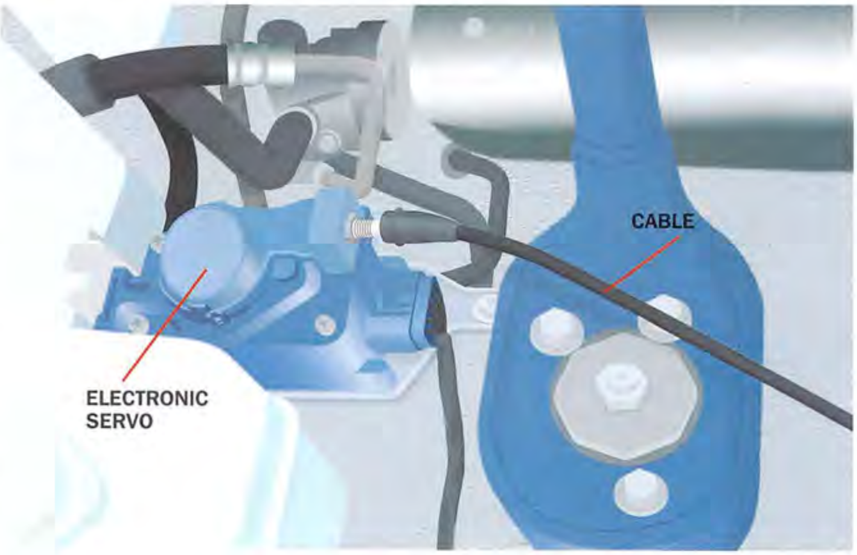

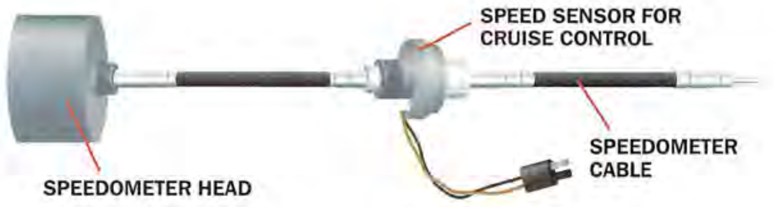

WHERE'S THIS THING?: Speed sensor may be buried under the dash near speedo head.

have a subtle leak caused by aging, brittle rubber or connections that don't seal well. A lot of the cruise controls we've seen with issues in regulating their speed correctly are fixed by simply replacing all the vacuum lines between the servo, vacuum reservoir (if there is one) and the intake manifold vacuum supply.

Control issue with your servo

There are two main components of the cruise control system: the controller and the servo.

The controller integrates all the inputs and tells the servo how far to actuate the throttle plates. Normally, there is no way you can service this item, so if you trace a problem to the controller, you'll have to buy a new one. A bad controller probably will require a service manual for specific diagnostics. You'll also need a scan tool to access any computer trouble codes to boot.

The servo does the work of moving the throttle blades to speed up the vehicle. Conventional systems are actuated by manifold vacuum. Check the linkage from the servo to the throttle to ensure it's properly hooked up and not binding or sticky. Inside the servo is a diaphragm that moves to pull on the linkage. Some older systems, used on cars or trucks that don't have throttle-position sensors, may have a rod that moves in and out of a magnetic coil to tell the controller the throttle position. There are also two electrically operated solenoid valves. One valve admits vacuum to the diaphragm chamber to add more throttle. The other bleeds air back into the chamber to reduce throttle. Normally, they will never be open at the same time, so if one is sticky or leaking, cruise control operation will be erratic at best. Check for leakage with a handheld vacuum pump. A leaky valve may benefit from a quick shot of silicone spray.

Diesels and some late models may use a servo that is completely electrical. These usually are mounted on the fenderwell or firewall, and are connected to the throttle by a second throttle cable. Other than keeping the cable lubricated and properly adjusted, there's not much to fiddle with. As always, check the shop manual for specifics on your vehicle.

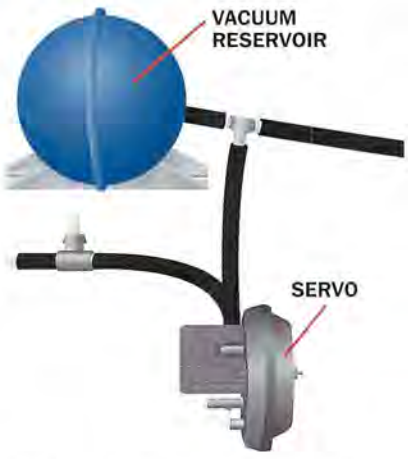

SUCK IT UP: The vacuum reservoir provides an added source of vacuum for those long uphill grades.

Many late-model cars have throttle plates connected directly to an actuator controlled by the engine management computer. There is no physical connection between your right foot and the throttle blades (Scary, eh?), and there may or may not be a separate cruise control servo or controller.

How fast was I going, officer?

If you still have no engagement, there may be an issue with the cruise control's speed input. Modern cars use information from the ABS's vehicle speed sensor (VSS) because the engine-management computer needs to know road speed. If your Check Engine light is on, the cruise may not work, especially if the VSS is malfunctioning. Older vehicles may use speed information from the speedometer. Aftermarket systems may resort to a ring of magnets clamped to a driveshaft or axle shaft, with a magnetic coil positioned nearby. If the magnets have fallen off or the coil has been smashed by a rock or road debris, there's no speed data for the controller.

Urban Legend

The legend has been related to us by all manner of people, including a couple of state police officers. Using the cruise control in rainy or slick conditions will make your car speed up uncontrollably, until you lose control and crash. It's a myth. Engaging the cruise will not make the car speed up. The cruise will attempt to maintain a steady speed. If the wheels lose traction and the car starts to slow down, the cruise will speed up the engine to attempt to accelerate back to the set speed. This will make the drivewheels spin more briskly. The vehicle speed will go slower, regardless of how furiously the wheels spin. If you have a rear-wheel-drive car, the effect is to destabilize the vehicle, and you probably will spin out unless the vehicle regains traction in a reasonable length of time. A front-wheel-drive car with spinning tires will attempt to continue in a more or less straight line—which will make steering control dodgy at best.

Heavy rain or snow makes use of the cruise control inadvisable, but such conditions certainly won't make your car accelerate to ludicrous speeds while you hang on for dear life. If the cruise is engaged and you perceive a loss of traction, the best bet is to tap the brake to disengage the system, and then add just enough throttle to maintain steering until the vehicle slows down to a more appropriate speed.

Changing A Fuse

It's a dark and stormy night. Made even darker because the headlights on your car have suddenly quit. The car runs fine—but you have to have it towed to a garage. Tow: $85. Labor to replace headlight fuse: $25. One 10-amp blade-style fuse: 65 cents. The realization the next morning that there was a spare fuse already in the fuse box that you could have swapped in yourself for free in a few seconds: Priceless.

Better To Light A Single Candle

Modern cars and trucks are heavily dependent on their electrical systems. The one in your car probably generates more power than was available in your great-grandfather's entire house. You might find three different fuse boxes on your vehicle, and dozens of fuses. Unlike domestic (house) electrical panels, which rely on reusable, resettable circuit breakers, most automotive circuits use disposable, one-time-only fuses. The only automotive applications that we're aware of that use circuit breakers use self-resetting breakers, which will cycle automatically every few seconds.

Fuse basics

Let's start with the basics. Any electrical component uses a certain amount of power to accomplish its task. Take, for example, a 21-watt brake-light bulb. Two of these add up to 42 watts, which at your car's nominal

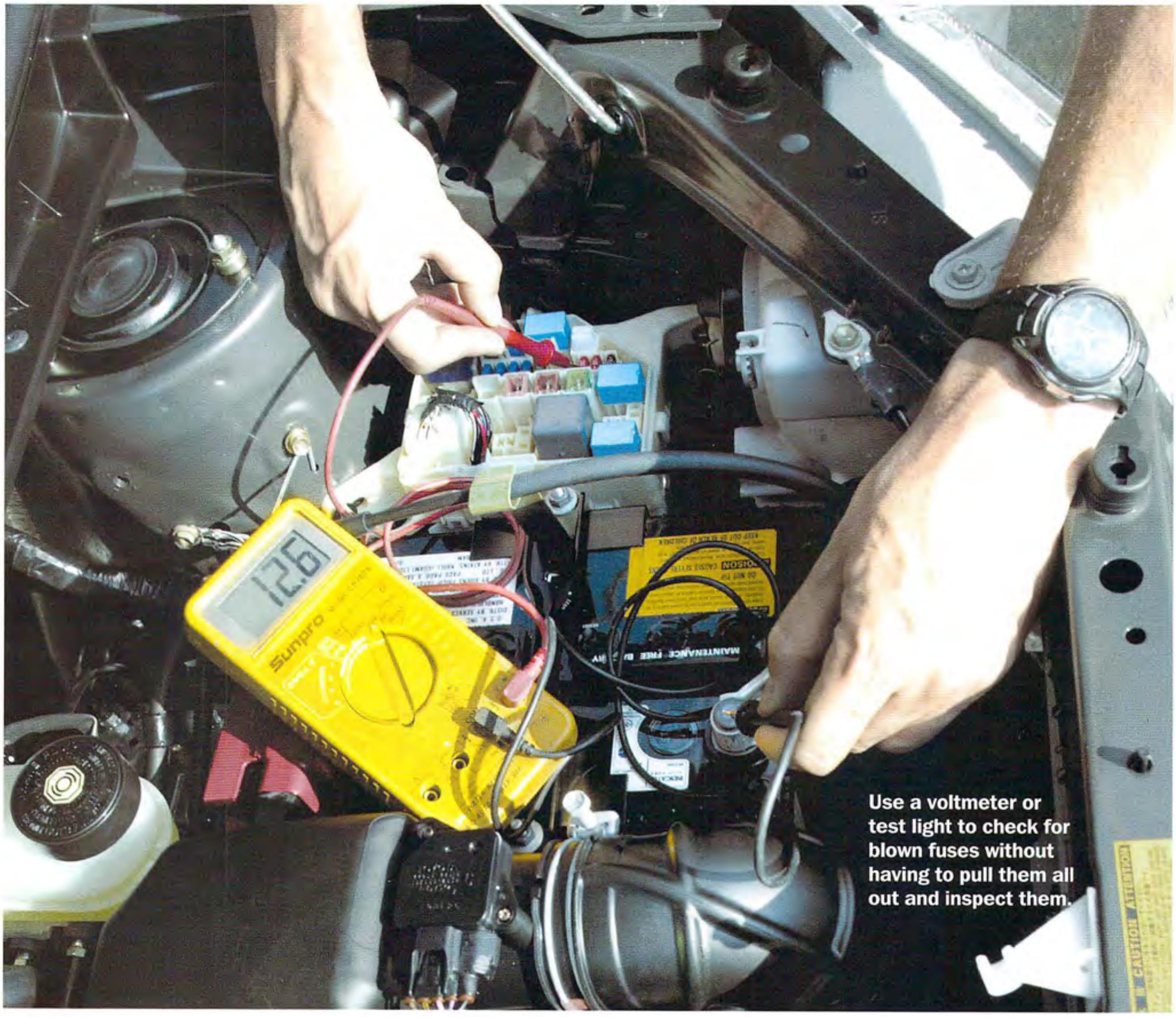

Use a voltmeter or test light to check for blown fuses without having to pull them all out and inspect them.

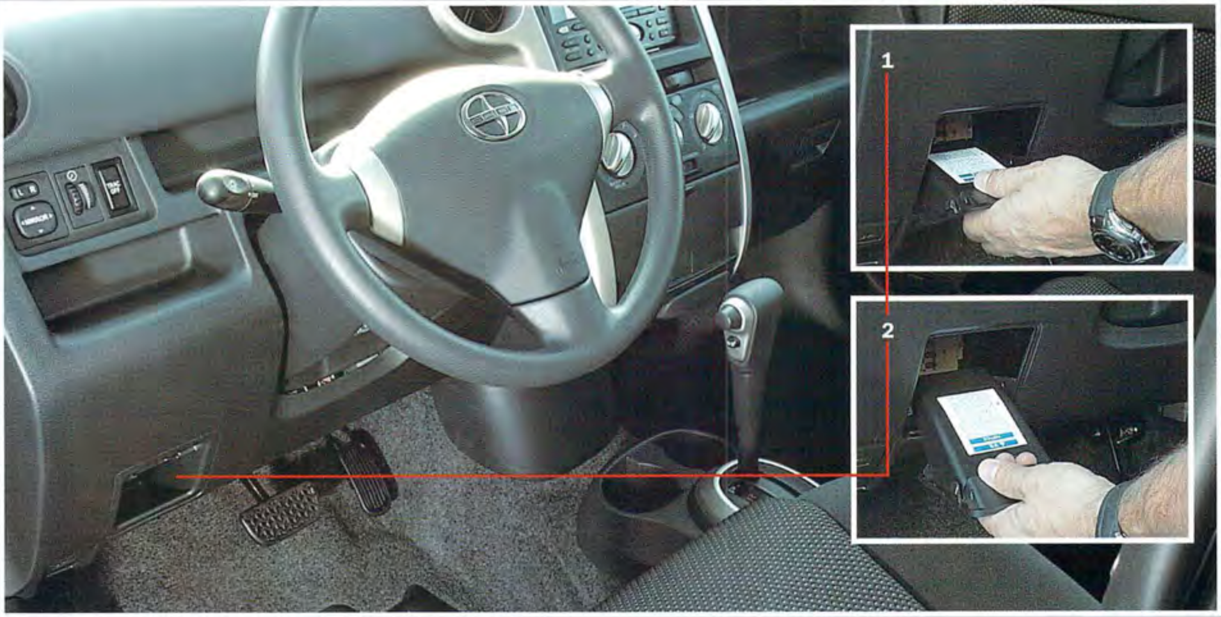

FINDING THE FUSE BOX

Which fuse box? Many vehicles have more than one. Generally, the primary box is housed under the dash. Look above the driver's feet (1), behind an access panel at the left side of the dash panel or near the bottom of the glovebox. If it's not obvious, then all you need to do is look in the owner's manual, which will tell you how to find it. The fuse-box cover should have a chart on it (2). There may be a secondary panel underhood, which probably controls electrical circuits for underhood items like the air conditioning or the ABS controller. Again, the owner's manual will tell you where to find this panel. Aftermarket items (like stereos) sometimes have a fuse in line with their power cord. It may be hidden under the dash and you'll have to run the whole wire with your fingers to find it.

operating voltage of 14 volts is about 3 amps of current. Suppose the wires to the lamp socket were chafed and the bare wire touched some exposed metal, shorting the circuit to ground. The current draw could instantly rise to dozens of amps—and in a few seconds the wiring harness would become a smoking, charred ruin—if it weren't for the fuse. The fuse is designed to carry a precise amount of current and no more. Ask it to do more, and a small wire or strip of metal inside it melts rapidly, sparing the rest of the wiring. And then you can find the short, repair it and replace the fuse for a few cents.

Fuses sometimes fail for what seems to be no good reason. Go figure. One common scenario is when a marker lamp or headlamp burns out. The melting filament in the bulb can momentarily short the contacts inside the bulb just as it vaporizes, sucking enough current to pop the fuse. You'll need to replace the bulb and the fuse.

You'll need pliers or some sort of tool to pull the fuse out of its spot.

Gone Fishing

Not sure which fuse is popped because you don't have a chart? You could always pull all of them, one at a time, and inspect them. A faster way is to use a voltmeter or test light. The conductors inside blade-type fuses are exposed at the upper surface. Start by turning on the ignition and the circuit you're trying to troubleshoot. Connect the ground wire of your meter/test light to a good chassis ground point—one with bare metal. Touch the probe to the fuse's exposed conductor. A good fuse will show voltage on both sides. A blown one will be electrically dead on one side.

What Amps?

Fuses are carefully manufactured to permit only a calibrated amount of current to flow. Replace a blown fuse with one with a higher amperage rating at your own risk. You could cause a meltdown in the wiring harness—which will in turn cause more short

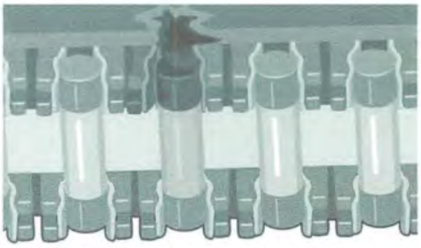

This glass fuse fuse-box housing melted after a loose clamp overheated.

circuits, some of which may not be protected by other fuses. You may need to replace the entire wiring harness, or you may start an electrical fire that destroys your entire car or truck.

Now that I've got your attention, I'll backpedal and say that in an emergency I once had to replace a blown 15-amp with a 20-amp because that was the only spare in the fuse box. But first thing the next morning I replaced the fuse with the correct one—which promptly blew. The fuel pump was drawing too much current because the fuel filter was dirty, but that's a different Saturday Mechanic.

There are three sizes of blade-style fuses—mini, normal and maxi. And they are all color-coded: normal-size 10-amp fuses are red and 15-amp fuses are blue, and so on. Just to confuse the issue, the color-coding is different for the maxi sizes. Don't make a mistake and assume a red fuse is always 10 amps. Additional warning: Don't assume the last person to replace a fuse did so with one of the correct rating, even if he was a professional. As always, you can look in the owner's manual—and probably on the fuse chart on the fuse box cover—for the correct rating.

No Spares?

Some vehicle manufacturers do not provide spare fuses or even a place in the fuse panel to store them. Even if there are spares in your fuse box, there are probably only a couple, and not one of every value. Go to the auto parts department or auto parts store and buy a fuse kit. For less than 10 bucks you can get a box containing an assortment of fuses with a fuse puller inside. Leave this in your glovebox to keep it dry and clean.

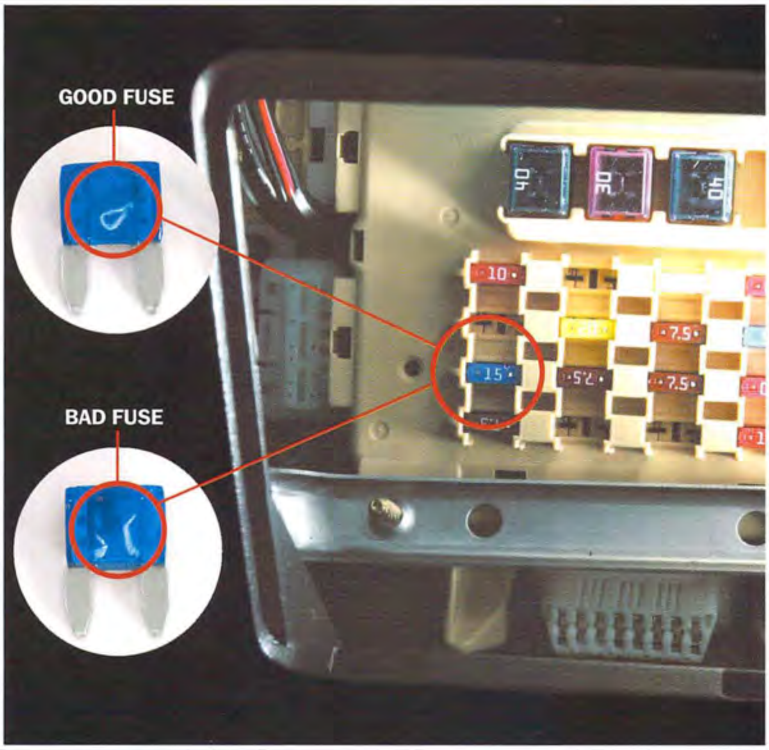

THE RIGHT FUSE

Okay, your headlights are out. Or your dome light, or the power socket won't charge your phone anymore. You've checked the lamp and your phone charger, so you're pretty sure the fuse is blown. But which fuse? There are a couple of dozen, of all different colors. The fuse box cover panel almost always has an index printed on its interior as to which fuse controls which circuit. If not, the owner's manual has an index.

Removing the suspect fuse is simple—yank it out. If you have a car made in the late '80s or afterward, it probably uses blade-type fuses, which can be simply pulled straight out of their sockets. Of course, it's not likely that you can get them out with your bare fingers. You'll need to use a tool of some sort. Some thoughtful car manufacturers provide a fuse puller, neatly installed in the fuse box. It's simply a small pair of tweezers with jaws perfect for yanking fuses. Grab, squeeze and yank.

If your vehicle uses old-style glass fuses, you may need to pry them out of their clamp. Warning: Glass fuses will break if you pry on them hard enough. Instead of levering on the glass, pry from one end of the fuse. If you must pry from the side (because the fuse box is in an inaccessible area), pry the metal cap. And remember that you need to replace that fuse, so don't muck up the fuseholder blades—they need to make good, tight electrical contact. No fuse puller? A small pair of pliers will work. Also, you can try prying the fuse from the panel with a screwdriver—there's a small lip on the top of the fuse body to grab. Just be careful where you stick the screwdriver so you don't short something else out. To replace the fuse, just push it in with your finger.

Older, European-style fuses use a simple piece of sheetmetal wrapped around a ceramic fuse body. They're pretty stout, but they're prone to corrosion between the end of the fuse and the contact; try spinning the fuse in place to re-establish contact if you've got a problem and the fuse looks good. This style of fuse isn't color-coded—read the fine print for the rating.

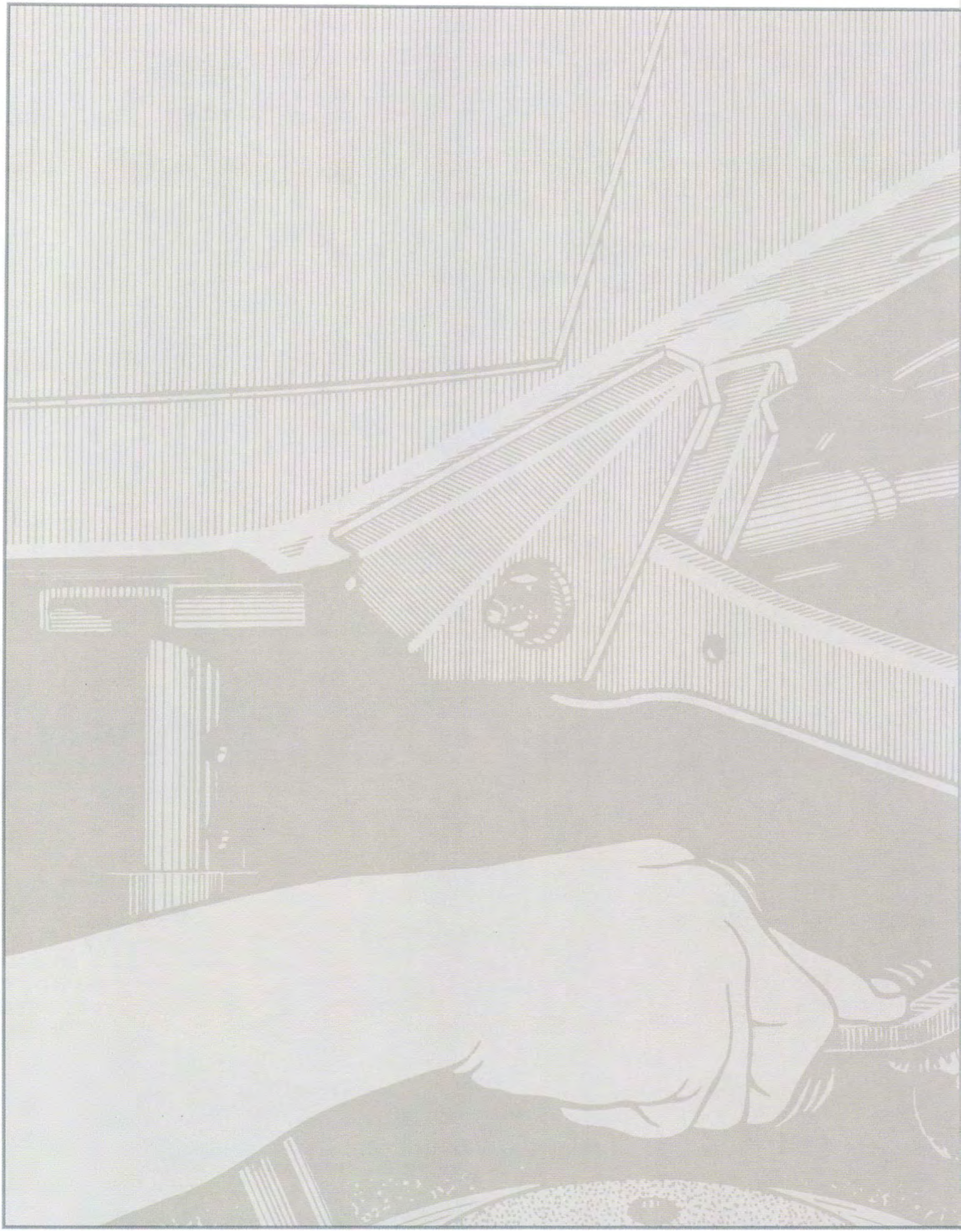

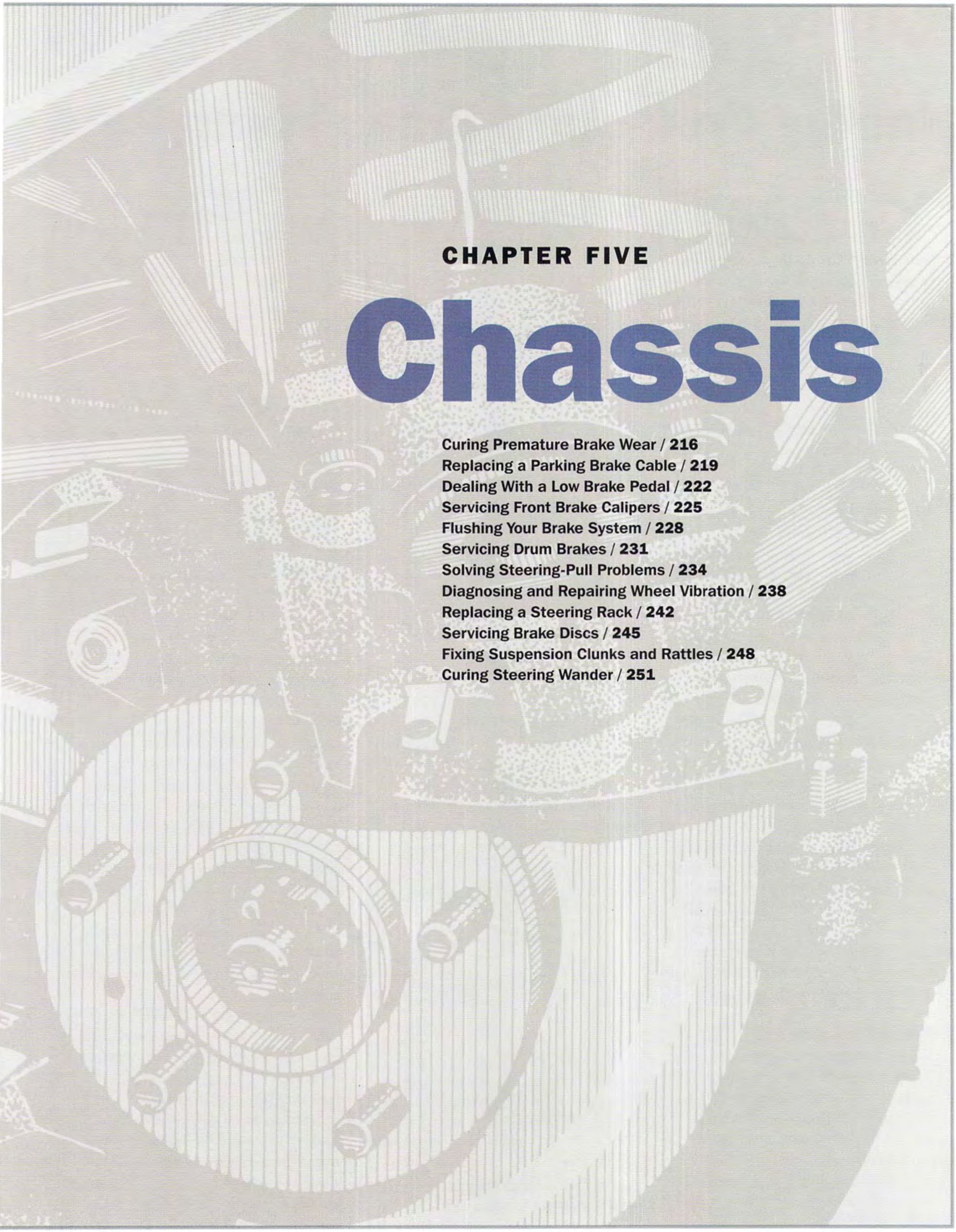

CHAPTER FIVE

Chassis

Curing Premature Brake Wear / **216**
Replacing a Parking Brake Cable / **219**
Dealing With a Low Brake Pedal / **222**
Servicing Front Brake Calipers / **225**
Flushing Your Brake System / **228**
Servicing Drum Brakes / **231**
Solving Steering-Pull Problems / **234**
Diagnosing and Repairing Wheel Vibration / **238**
Replacing a Steering Rack / **242**
Servicing Brake Discs / **245**
Fixing Suspension Clunks and Rattles / **248**
Curing Steering Wander / **251**

Curing Premature Brake Wear

You're taking a pleasant ride in the country, enjoying your car's performance. A stop sign. You step unhurriedly on the brake pedal—what the heck? You let off and try again. There's no mistake. You hear a high-pitched squeal that you recognize as a pad wear indicator sounding off. Those front linings are going, going, gone.

That wouldn't be so troubling except that you distinctly remember doing a brake job less than a year ago. Hey, what's going on?

Low-Life Linings

Thirty years ago when disc brakes were becoming common, we were impressed by how long linings lasted. Not anymore. Many late model cars and trucks eat their pads quickly. The tendency is to blame new friction material formulas for accelerated pad wear, but the real reason is usually something else altogether.

Now, you've got overdrive, a low-friction V6 and an aerodynamic design that not only lets the car coast farther, but also cuts airflow to the brakes.

Another factor in the short-lived pad scenario is the SUV phenomenon. A typical Blazer or Explorer might weigh 4800 pounds, but the brakes are not enough better than those designed for the 3000-pound pickups most of these vehicles evolved from. It doesn't take much of an intellectual leap to see that even normal driving amounts to heavy-duty service.

Take a Proactive Role

What can you do to extend the life of those brakes? First off, buy the highest-quality brand-name linings you can find. These may be "full-ceramic," have ceramics added to the semimetallic mix, or have a sacrificial titanium coating that speeds break-in and improves initial feel.

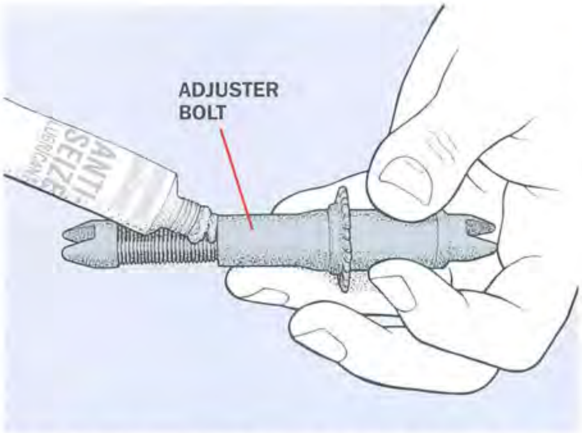

(Fig. 1) Coat those star-wheel adjuster threads with grease or antiseize compound.

That's not, however, all there is to the job by a long shot. You've got to look at the brake system as a whole.

Many motorists just don't use the parking brake. Some may think it's an emergency stopping system only. But with many common rear-disc designs and some rear drums, lining-to-rotor/drum adjustment simply doesn't occur unless the parking brake is applied (a symptom is a low pedal). So, adopt the sensible habit of engaging this device every time you park instead of just depending on Park to keep the car from rolling away.

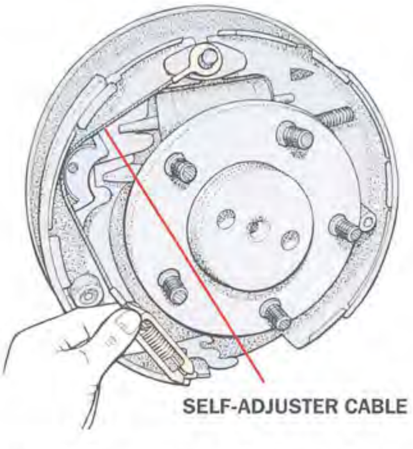

(Fig. 2) Drum brake self-adjuster hardware is frequently faulty.

Innards and Fresh Juice

Even the most diligent person in this regard is still going to have problems if the self adjusters of the rear discs aren't operating. Typically, corrosion and contamination jam the piston and immobilize the screw mechanism.

Some professionals have switched from overhauling rear calipers to installing quality remanufactured units, saying that it wasn't worth the labor and headache to pull them apart and put in a kit—it's hard to get the adjusting mechanism to work properly. Great, except calipers can be expensive in some cases. That's the price you pay for driving a modern car and neglecting maintenance.

What maintenance is that? Brake fluid changes, of course. With that parking brake/self-adjustment mechanism present and immersed, internal corrosion is to be strenuously avoided. Ergo, periodic fluid changes are even more important than with disc/drum systems. This has been advised by many import manufacturers for decades, and is now starting to show up in the service recommendations found in the owner's manuals of domestics.

This car-care item can save you serious money in the long run. Besides disc brake internals, consider ABS. Corrosion and debris in the control unit can result in a disastrously expensive repair—$1500 isn't unusual! Use a turkey baster to get most of the old stuff out of the master cylinder reservoir, refill it with fresh, then use a clear bottle and hose setup

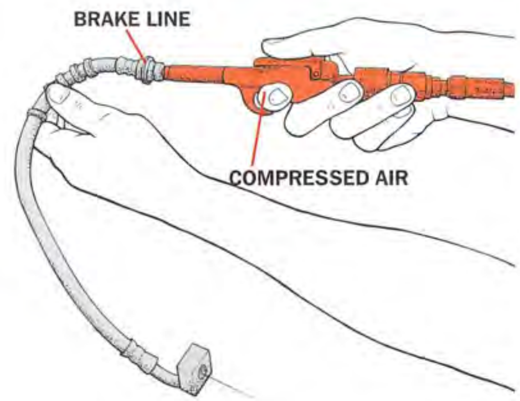

(Fig. 3) A partially clogged brake line can let pressure build up in the caliper.

at each of the bleeders and pump until you see a nice, clean liquid appear. Do this every other year.

Beaten Drums

Rear drum brakes often shirk their duty, too, resulting in burned-up front pads. Seized star-wheel screws (Fig. 1) and otherwise inoperative self adjusters (Fig. 2) are practically an epidemic, so you're asking for accelerated front lining wear if you don't inspect and lubricate the hardware involved, and replace any items you're dubious about (frayed or fatigued cables, for instance).

We fixed a car recently that had fried pads and a badly grooved rotor on one side. It turns out the hose was plugged—it allowed pressure to gradually build up in the caliper, and then the piston couldn't retract (Fig. 3). Another possible cause of drag is overfilling the master reservoir, which can apply the brakes as the fluid expands.

Broken In or Just Broken?

We know people who still believe that the right way to seat new linings is to really stand on the brakes a few times. That's an anachronistic idea left over from the days when linings were supplied "green." Panic stops would indeed get that friction material hot enough to cure it. But you don't get uncooked pads and shoes from manufacturers anymore, so this whole idea belongs to a bygone era, a time of ignition points and bias ply tires.

There's no way to overemphasize the importance of proper lining break-in (some authorities say overlooking this procedure is the No. 1 cause of noise and hard pedal complaints). The ideal way to start new linings off is to make 30 slow stops (spaced 2 minutes apart) from about 30 mph using light to moderate pressure. You may not be willing to take that much time, but you'd be foolish not to make at least 10 moderate stops at 30-second intervals (you

HOW IT WORKS

Pad Wear Indicators

Mechanical pad wear indicators, which have been around since the early '70s, are so simple and can head off so much damage that it's surprising they didn't show up even before then. All that's required to warn the driver that he's almost down to the rivets is the addition of a light steel tab that contacts the rotor when the linings are getting too thin for comfort. This produces a high-pitched squeal that appears suddenly and is unmistakable. The tab is too flimsy to cause any damage to the disc, unlike the rivets, which will grind away at that nice stopping surface, leaving wide, deep grooves. With bonded linings, the steel pad plates will do the same, and have been called "full-mets" by brake guys with a twisted sense of humor.

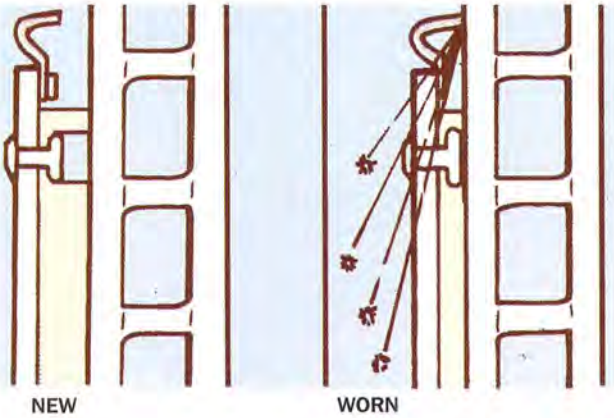

By the time you hear and feel either, considerable rotor damage has already begun.

Unfortunately, many of the cars out there still aren't so equipped, and often replacement pads for cars that had wear indicators as O.E. don't include this helpful device.

A slightly higher-tech variation on this theme is the electrical pad wear warning system, which first appeared a couple of decades ago on such cars as Toyotas and Mercedes. Here, the pad is drilled for an electrical contact that stands slightly proud of the lining rivets. When wear reaches the point that this touches the rotor, the ground circuit to a Brakes warning light on the dash is completed, and the lamp winks on. Again, unfortunately, some companies that manufacture replacements for these applications don't bother to include the contact, with the result being that the wire ends up tied off in the suspension somewhere.

should be able to feel the action smooth out), then avoid heavy braking for the first 200 miles.

What if one pad is worn-out, but the other still has lots of friction material left? Think about the way a single-piston caliper operates. The piston pushes one side directly against the disc, and this action pulls the other side into contact. If anything interferes with the sliding or floating movement that transfers and divides the force, only the pad that the piston actually bears on will wear. So, every time you put in a new lining, clean and lubricate the machined grooves of sliding calipers, using special brake grease **(Fig. 4)**. Ditto for the bushings and pins or bolts of the floating variety, and make sure any rubber parts are in good shape.

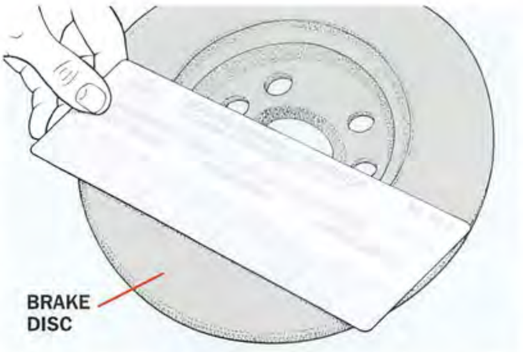

(Fig. 5) Discs need to be refinished to a finer finish for today's pad material.

Voice of Experience

When lubrication is called for on an internal mechanism, use silicone grease. Anything else will attack the rubber parts.

Don't force caliper pistons back for pad replacement without opening the bleeder. The line comes in near the bottom of the cylinder where the sediment is, and this forces debris up into the ABS unit—often resulting in a glowing antilock warning light. In fact, many brake experts say just opening the bleeder isn't good enough. They want you to clamp that hose (use a special tool or pad the jaws of locking pliers with heater hose).

Don't believe you can get away with the rough directional disc finish that worked fine with asbestos. Smoothness is the rule today. If you're familiar with roughness scales, where 80-100 RMS or 73-91 RA was once considered correct, now 40-60 RMS or 36-55 RA is recommended (you can check this with a surface comparitor gauge). In other words, finish that rotor up with 120- to 150-grit paper instead of the traditional 80-grit **(Fig. 5)**.

Regardless of whether the rotors have been refinished, you absolutely must take the time to clean them. Otherwise, hard particles will become embedded in the new linings and you'll get noise and scoring. But brake cleaner isn't the answer because it won't float those iron filings away. Use detergent and water, and then dry with a paper towel.

(Fig. 4) If the caliper won't slide on its grooves or pins, brakes will wear fast. Clean, and then lube sparingly.

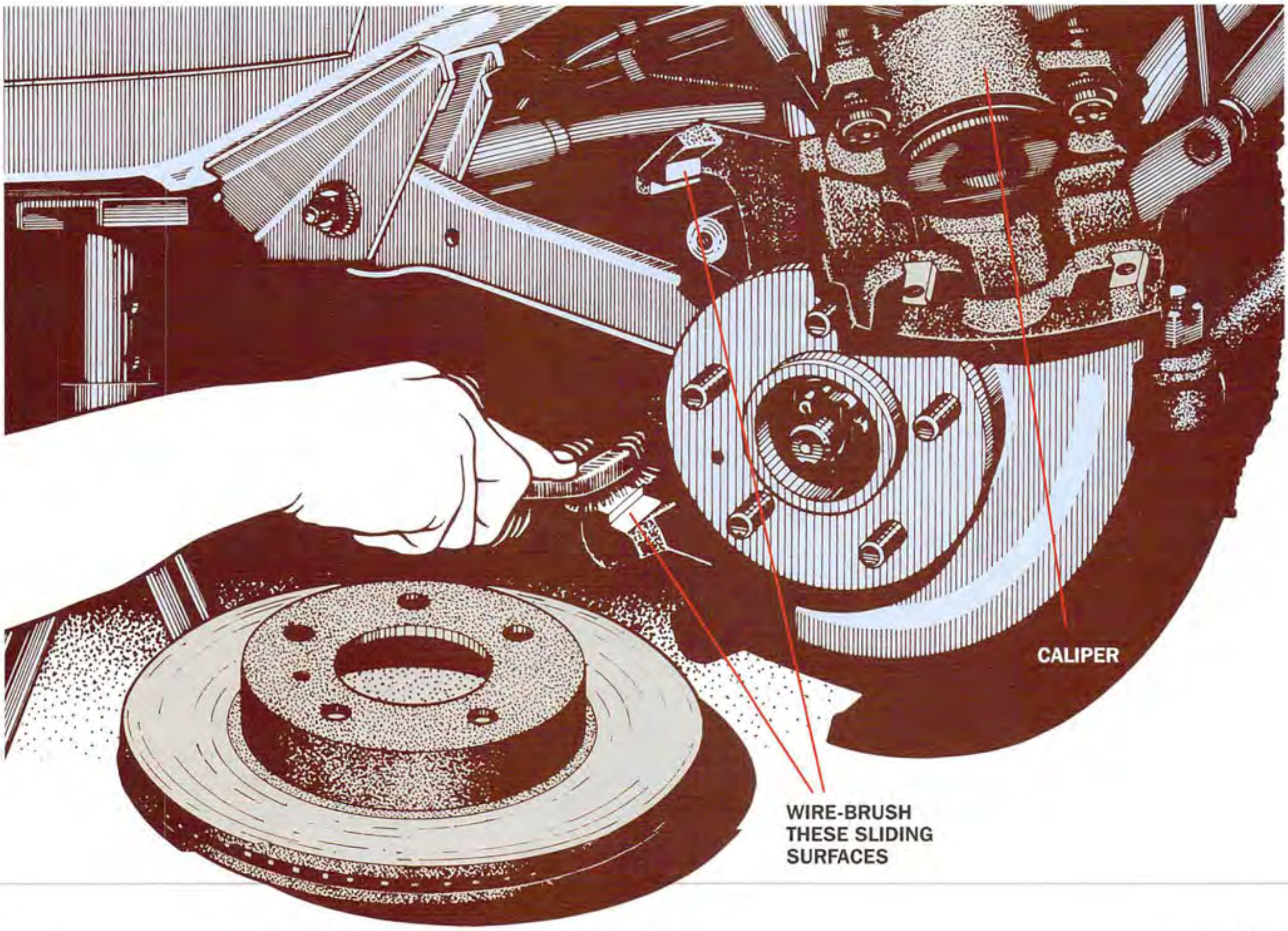

Replacing a Parking Brake Cable

You've just parallel-parked your manual-transmission car on a steep side street, carefully setting the parking brake by pulling it up farther than you've ever pulled before. So far that it actually pops a little as it edges up that last click. Exiting the vehicle, you turn to lock the door—just as your car slides forward 3 ft. and nudges the luxury car in front of you in the bumper. Setting off the alarm. Hurriedly, you restart your car and back up to your original position, determined to pull the brake up far enough to anchor the *Titanic* this time. Unfortunately, this time the handbrake offers no resistance whatsoever until the mechanism tops out—and you have no parking brake at all, forcing you to find a flat parking space three streets away and walk through the rain to your job interview.

Not to mention having to replace the parted handbrake cable.

(Fig. 1) Unlike the two separate cables shown on page 190, this vehicle uses an equalizer bar to distribute the pull to both rear wheels equally. This means three potential adjustments.

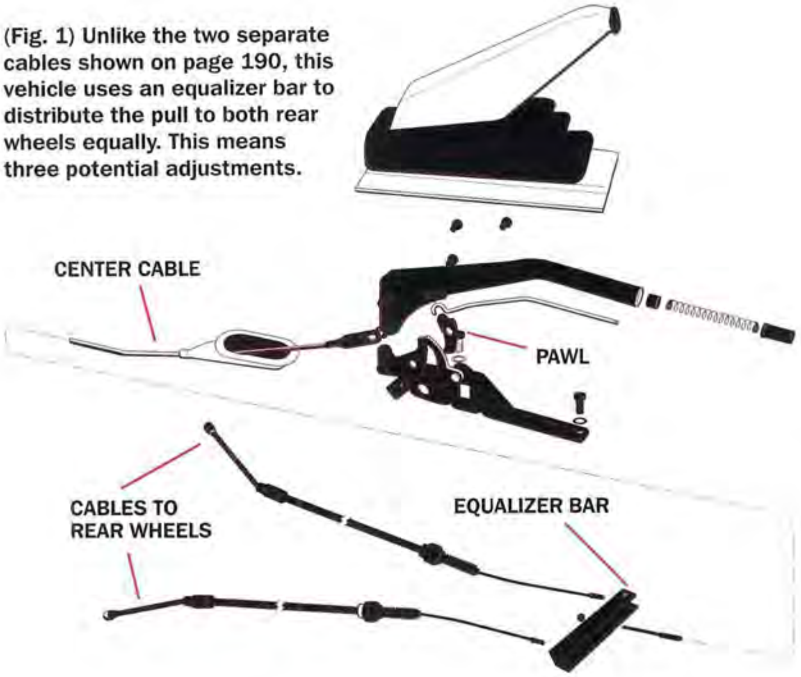

Why It Failed

Parking brakes are operated by a long, steel cable that runs between the handle in the cockpit and the rear wheels. It has crimped-on ends that can potentially slip off (unusual but not unknown). More commonly, the rubber-covered outer cable that runs between the car and the rear brakes tears or splits, allowing moisture and road salt to corrode through the inner steel cable. Sooner or later, the corroded cable fails under tension.

Live in the desert where it's flat, and never use your parking brake? Or maybe you have an automatic transmission and just put your car in Park. Then you face another issue—the cable needs to be exercised regularly or it will seize up. So you should use the parking brake regularly. Ever have trouble getting your transmission out of Park when parking on a grade because the parking pawl in the tranny is jammed against its gear? Setting the parking brake before putting the trans in Park and letting the car roll forward will prevent this.

Getting Dirty

First, you need to determine which cable has parted. Remove the boot over the mechanism between the seats and see which cable is slack. If there's only a single cable to the handle, you'll often need to pull up the rear seat and look for the equalizer bar (Fig. 1). One cable to the bar will be easy to pull straight out. Warning! Parking brake cables are generally lubricated with thick black grease that will get everywhere, and generally destroy your car interior and your clothes. Be ready with rags to contain the mess.

Now, you need to check the outer cable that leads to the rear wheel. If it's damaged, you'll need to replace that as well. You may have no option if the inner and outer cable are sold as an assembly, which they often are. To actually look at the outer cable, you'll need to get under the car. This means either backing onto ramps or jacking up the car and resting it on jackstands. If you're using jackstands, loosen the lug nuts first. (Never get under a car that's lifted on a jack, even a floor jack.) Be sure to block the wheels.

Check the outer cable, which leads from the backing plate of the brakes over to the car body. Don't confuse it with the brake hydraulic line, which is a steel tube connected to the body at some point with a short rubber hose. The brake cable will be much thicker and rubber-covered along its whole length. Look for cracks, abrasions and deteriorated rubber, and don't forget both sides—if the other side is damaged you may as well replace both cables now.

Inside the Brake Drum

Once you've determined which cable(s) to replace, get the car up in

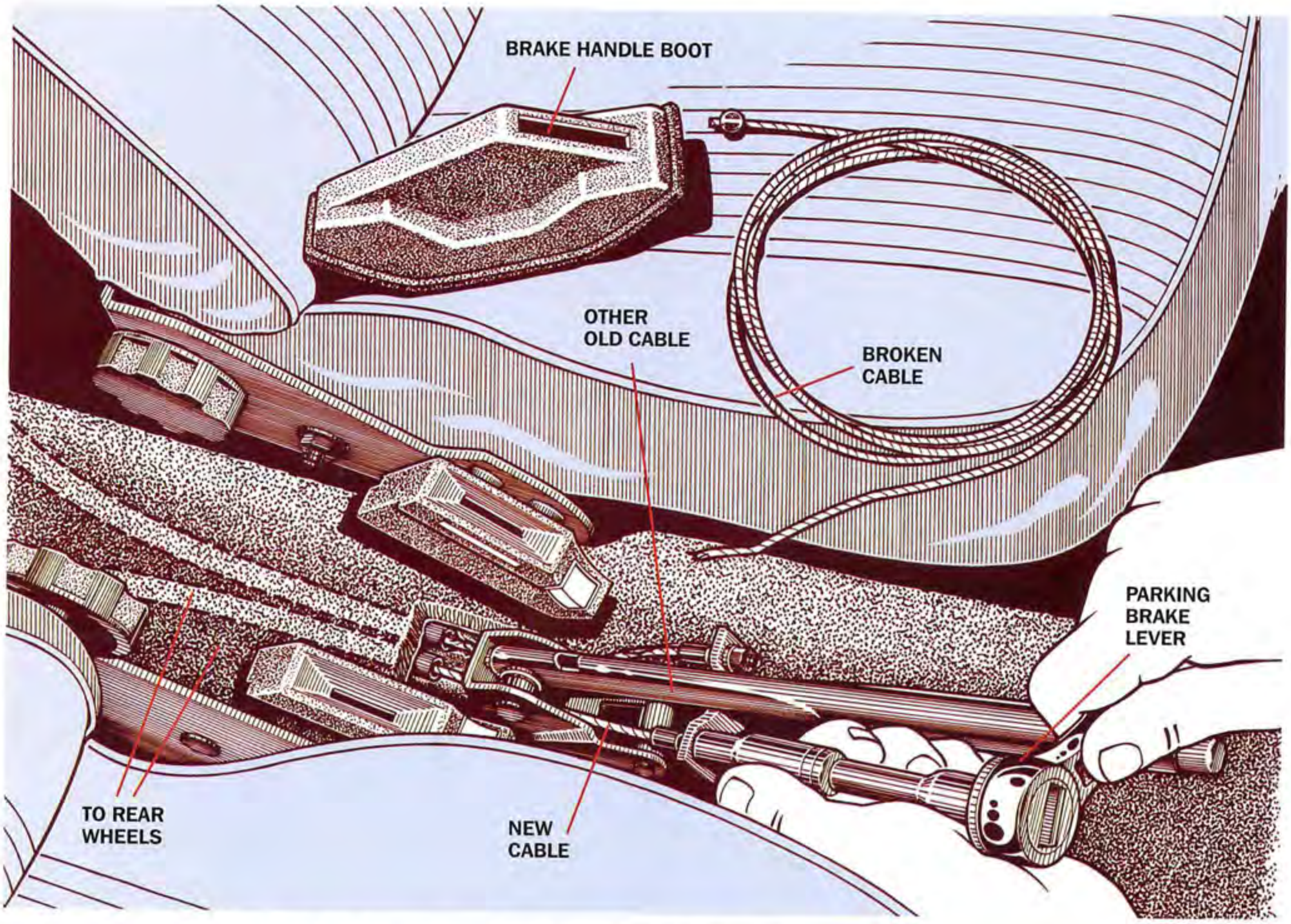

(Fig. 2) The final adjustment of the new cable is usually done at the handle inside the car. You may need to hold the new cable with locking pliers to keep it from turning as you use a socket on the adjuster nut.

the air and the wheel or wheels off. Now you need to remove the brake drum. Generally, the drum is simply sandwiched between the axle flange and the wheel, retained by a small sheetmetal clip or Phillips setscrew to keep it from landing on your foot when you remove the wheel. Occasionally, they will become intimately attached to the hub and need persuading. Judicious application of penetrating oil, light hammer taps near the hub, profanity and moderate heat will see them loosen. You may need to back off the adjuster mechanism to get the shoes out of the way if the drum has developed a wear ridge on the inner lip.

Now you can see the brake shoes and mechanism. (This might be a good time to replace the shoes if they're worn down to anywhere near the rivets). If you're replacing only the inner cable, simply disassemble the mechanism far enough to pull out the old cable and thread the new one into place. If you need to replace the outer cable as well, it's probably pretty well rusted onto the backing plate. Spray it with penetrating oil, and twist with pliers until it lets loose. Some cables are retained by snap rings or a small collet-style clip—check the new cable for its attachment style.

Clean It Up

Now's your chance to wire brush rust and dirt from all the hardware and mating surfaces on the brake mechanism, and sparingly lubricate all the moving parts with antiseize compound or high-temp grease. Replace any suspicious hardware like springs and clips. Be sure to clean the mating surface of the axle flange and brake drum, to prevent foreign matter from keeping the drum from seating properly on the flange. Lightly lube the mating surfaces with antiseize so you won't have to cuss the next time you need to take it apart. Reinstall the brakes and drum, hooking up the new cable to the actuating arm. Adjust the brake shoes properly if you needed to back off the adjuster.

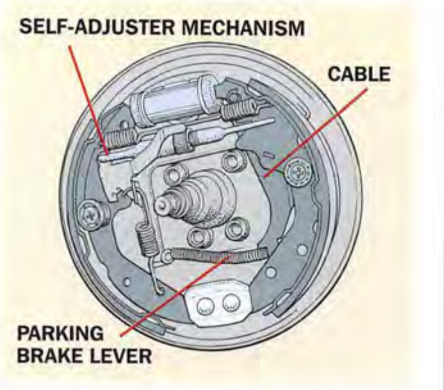

(Fig. 3) Be sure the rear brakes are properly adjusted before trying to set the parking brake adjustment.

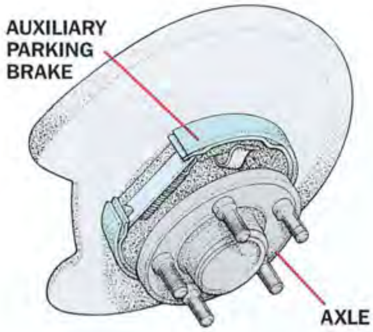

(Fig. 4) Some cars with rear discs use a small set of brake shoes dedicated to the parking brake.

Back Inside

Now it's time to hook up the inner cable's forward end **(Fig. 2)**. If it didn't come with a new nut, you'll have to rescue the old one. It may be a locknut, or there may simply be two nuts jammed together. Thread the threaded end of the cable into the equalizer bar or the handle, and lightly run the nut down. Pull the handle up (or depress the pedal) three clicks, and then tighten the adjusting nut until the wheel lightly scrapes as you rotate it. This assumes you had the brake shoes adjusted correctly and they don't scrape with the handle down.

If you don't have the brake shoes correctly adjusted, and then you adjust the handbrake, bad things can happen. As the brake shoes' automatic adjuster mechanism **(Fig. 3)** kicks in over the next few days, the slack in the handbrake cable will gradually disappear, and your brakes will start to drag. Be careful for the first few days that the handbrake retains two or three clicks of free play.

Is the equalizer bar straight? Not cocked over at some wacky angle, but reasonably close to perpendicular to the cables? If not, adjust the nuts appropriately. If your brakes use two separate cables with no equalizer bar, adjust the individual cables so both rear wheels have equal tension when the brakes are actuated.

Four-wheel Discs

If your car has four-wheel disc brakes, your job is simpler, because the attachment of the handbrake cable to the caliper is generally much simpler to get at than with drum brakes, requiring less disassembly and readjustment. Some rear disc brakes actually have a small drum brake system just for the parking brakes **(Fig. 4)**, but the cable arrangements are pretty easy to deal with. You generally can replace the cable without removing the disc or caliper.

PEDAL-OPERATED BRAKES have a longer cable that runs up under the dashboard.

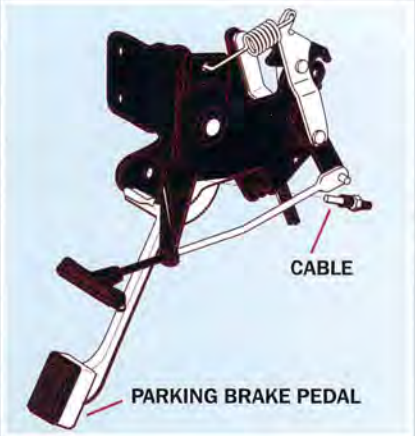

One last word of advice: Regardless of the state of health of your brakes, always turn your front wheels into the curb when parking on a hill.

HOW IT WORKS

Self-adjusting Brakes

Self-adjusting brakes are universal now, except for some elderly Volkswagens still on the road. The service brakes are actuated by a hydraulic cylinder that forces the two shoes apart, jamming them into the inner diameter of the brake drum. But as the lining material wears, a mechanical escapement mechanism is necessary to keep the shoes close to the drum. Otherwise, the pedal would gradually get closer and closer to the floorboards as the friction material wore down. This self-adjust mechanism is actuated by the slight vertical movement of the trailing brake shoe that occurs whenever the brakes are applied in reverse. Normal driving will keep the clearances appropriate. If you have a circular driveway, never back out of parking spots or generally back up like Granny, the mechanism may not be doing its job. Try a few smart applications of the brakes while backing up (needless to say, this needs to happen where you won't hit anything). You don't need to go fast—just walking speed. Hit the brake pedal good enough to stop rapidly, but not fast enough to chirp the tires. You may be surprised—your wimpy handbrake may regain some authority, and your braking may improve. Repeat as necessary. If you've been rocking about with the manual adjuster, you may need to do this a number of times. You'll need to do it if you've disassembled the rear brakes as well.

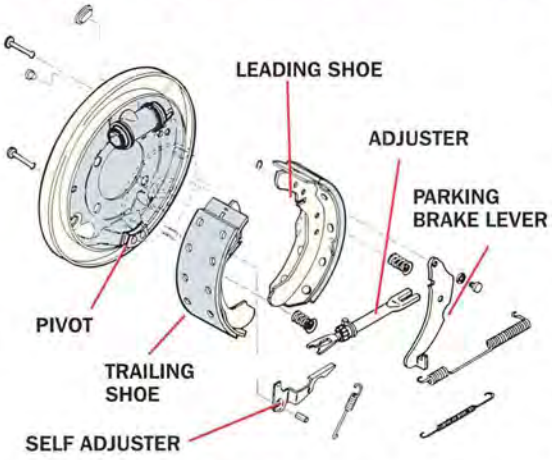

DIFFICULT

Dealing With a Low Brake Pedal

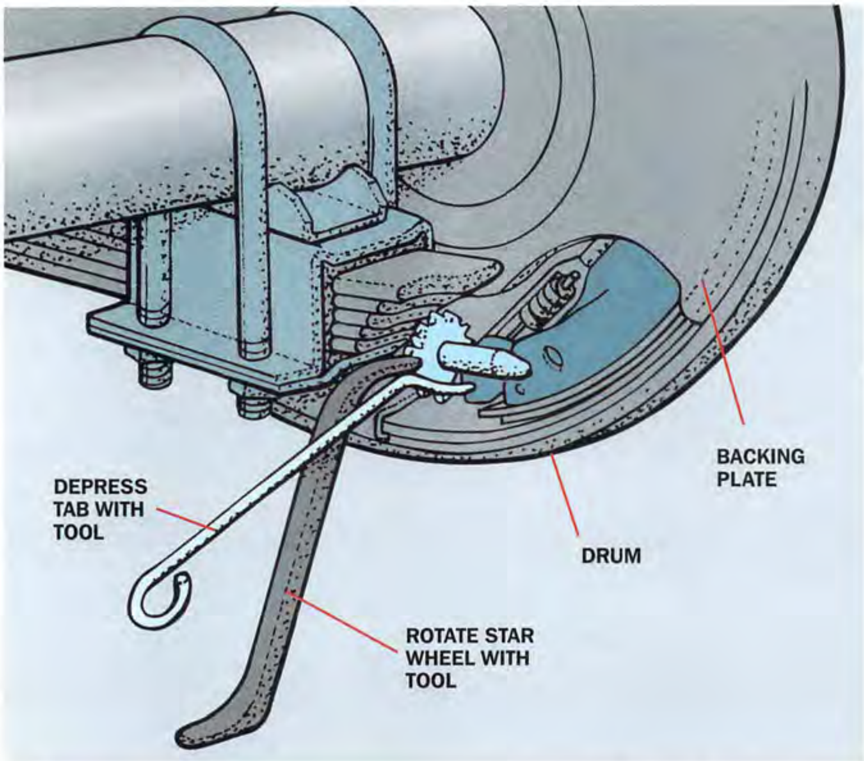

(Fig. 1) Depress the tab while you rotate the star wheel to close up the clearance. When the wheel scrapes lightly, go back one click.

(Fig. 2) Brake drums will be marked clearly as to how far they can be machined safely to remove out of round.

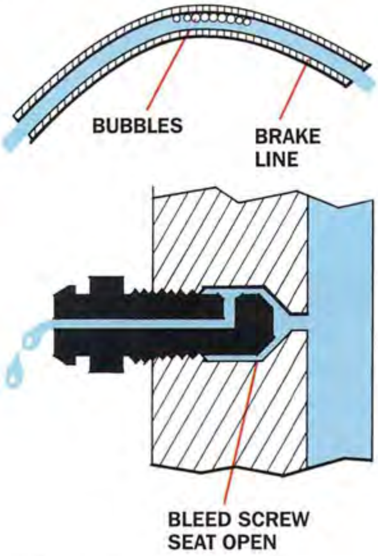

(Fig. 3) Bubbles collecting in high spots in the brake system need to be removed by opening the bleeder valves to flush them out.

"Honey, can I borrow your car?" No problem. She tosses you the keys. But at the first stop sign you draw a startled breath—the pedal's going, going, almost gone! There's even a little dent in the carpet under the pedal. Didn't she notice? Well, no she didn't. Typically, low-pedal trouble develops so gradually that people don't realize it.

Hydraulic brakes have been around since Duesenberg introduced them in 1921, but apparently a long history is no defense against troubles. And professionals and do-it-yourselfers alike are often guilty of misdiagnosis—they blame the master cylinder, though it is seldom the culprit.

There are only two plausible reasons for a low pedal: air in the system; and excessive movement between linings and rotors or drums (due to lack of adjustment, an out-of-round drum, or a wobbly disc that's knocking the pistons back so that there's extra space to take up before braking action begins).

Isolation

You can find out all you need to know about the master cylinder by removing the lines, screwing brass or plastic plugs into the outlets, and then applying the brakes. If the pedal's high and hard now, the master has been properly bled and its seals are okay. The pedal would sink gradually if it were bypassing—that is, if fluid were finding its way around the sliding seals. You've also confirmed that the booster is okay. Reattach the lines.

Continue the process of elimination by clamping hoses to isolate each wheel. Use a suitable rounded-jaw tool, either the locking-pliers type or one of those inexpensive J-hooks with a knurled screw. Releasing one at a time should locate the problem.

Use That Parking Brake

If you never engage the parking brake, self-adjustment of the pads and rotor simply won't occur, and that means a low pedal. Another impediment to adjustment is corrosion and contamination of the piston, cylinder and self-adjustment hardware. So, change your habits and start using the parking brake every time you leave the car, and replace those calipers if they're not just right. If the parking brake isn't used regularly, one of these days a parking lot attendant will apply it and your car will be immobilized until those corroded cables and other seized parts are replaced.

Beat the Drums

Rear drum brakes can cause a low pedal, too. Seized star-wheel screws and otherwise inoperative self-adjusters are all too common, and you're risking trouble if you don't replace the hardware when replacing shoes (Fig. 1). At the very least, clean the star-wheel threads and treat them to a coating of antiseize compound.

There's another factor that's usually not recognized: drivers who never stop aggressively enough in Reverse to ratchet the self-adjusters. It's a good idea to stomp on the brake pedal every week or so while backing up—preferably in a deserted lot or other safe place.

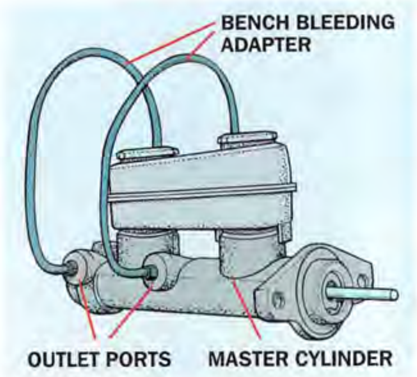

(Fig. 4) Bench bleed a master cylinder to get air out before installing it into the vehicle.

What about the drums themselves? They're frequently out of round (Fig. 2), leaving excess shoe-to-drum clearance and, of course, causing pulsation.

Bubble Trouble

For all practical purposes, brake fluid is incompressible. Air, on the other hand, can be squeezed down into a smaller-than-natural volume, and its presence will disrupt the operation of any hydraulic system. It promotes internal corrosion, too. Ergo, it must be expelled (Fig. 3).

One possible cause of pedal problems is failure to bench bleed a new master cylinder (Fig. 4). Screw the supplied fittings into the outlets and place the tips of the tubes in the fluid in the reservoir. Clamp one of the master's mounting ears in a vise—don't grip around the cylinder—so the unit is as level as possible. Use a rod or drift to stroke the piston slowly. Wait at least 15 seconds

HOW IT WORKS

The Dual Master Cylinder

Whether you call it the dual, split or tandem master cylinder, it has been used on every car sold in this country since 1967, although Cadillac had it in '62. Even so, most people don't understand its construction and operation. A typical modern specimen is of the composite variety—aluminum with a plastic reservoir—but iron 1-piece units are still around on older vehicles. Two pistons ride in the bore, and here's where we encounter some potentially confusing terminology. The rear piston is the primary, the one in the front is the secondary. This apparent misnaming resulted because the rear piston is the first to receive the signal from the brake pedal, so it does make a certain amount of sense. Kind of. Each piston has a primary cup seal at its front and a secondary at its rear. In normal braking, the pushrod from the booster forces the primary piston forward. No pressure is created until the primary seal covers the compensating or vent port from the reservoir, but once it does fluid is trapped in the chamber between the pistons and it becomes a solid column. Pressure is routed from this chamber to two wheels. A combination of the trapped fluid and the primary piston coil spring bears on the secondary piston, to which the line to the other two wheels is attached. The replenishing ports allow fluid to move freely between the chambers behind both pistons' primary cups and the reservoir, determined by demand and expansion and contraction from temperature changes. If a hose lets go or a saboteur has sawed through one of the brake lines, the other half will still provide a means of decelerating the vehicle, albeit with a lower pedal and reduced stopping power. This protective function is, of course, the dual master's reason for being.

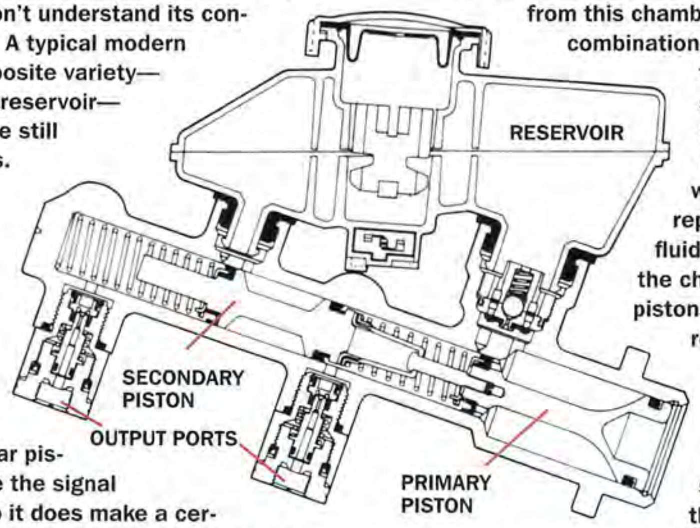

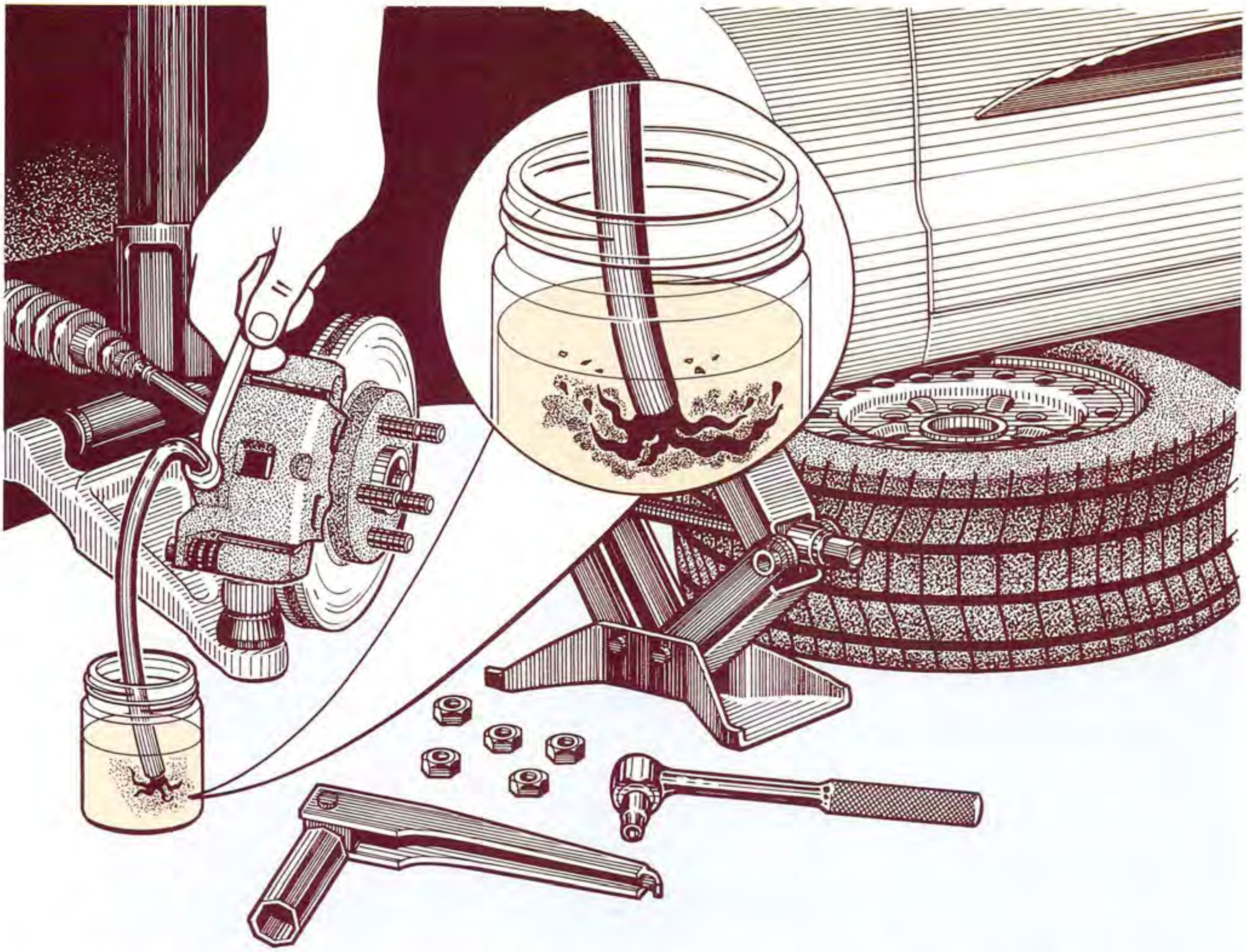

between strokes to allow the low-pressure chamber to release all its bubbles and fill completely. Keep stroking until there's no more evidence of air at the ports and tube tips.

If the car has a replacement cylinder that somebody didn't bench bleed, you might be able to do it with the master in place, provided you can jack the rear of the vehicle high enough to get the cylinder to be level. Again, pump slowly and allow time between strokes.

An important precaution to observe during any bleeding procedure that involves pumping the pedal is to limit pedal travel. You don't want the delicate lips of the master cylinder's piston seals to ride so deep in the bore that they encounter rough corrosion or deposits, which can scratch them. Just throw a chunk of 2 x 4 on the floor under the pedal.

When it comes to the bleeders at the wheels, most people just open them and let the fluid squirt. Not only will this result in slippery puddles on the floor, the fluid can shoot farther than you might expect—think about the 2500-plus psi of line pressure on some ABS-equipped cars. Brake fluid is a pretty effective paint remover, and it really burns when you get it in your eye. Wear eye protection.

One convenient setup is a tube and transparent bottle kept half full of fresh fluid **(Fig. 5)**. There are also inexpensive 1-man bleeder hoses that contain a 1-way valve to eliminate the possibility of air being drawn back in when you release the pedal.

The bleeder cups and hoses that are often included in manual vacuum pump kits, such as those from Mighty Vac, work well. Once again, you can see what you're getting, and you don't have to keep climbing into the seat to pump the pedal.

(Fig. 5) The old-fashioned, low-tech way to bleed brakes is to use a jelly jar half full of brake fluid, a short piece of hose, and a patient helper to depress the brake pedal.

You should also be aware of special procedures. For example, on Teves Mark II ABS systems, you can't get fluid to the rear brakes unless you turn the key on and then apply the pedal slightly. Be sure to check the shop manual if your vehicle has an antilock braking system.

Finally, there's the bleeding sequence. Since you're supposed to do the longest line in the circuit first, the traditional order is right rear, left rear, right front and left front. But with the diagonally split systems you'll find mostly on fwd cars, the order is right rear, left front, left rear then right front. ABS-equipped cars may have special procedures to follow.

Servicing Front Brake Calipers

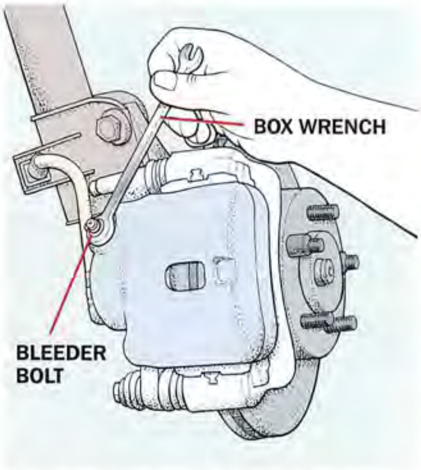

(Fig. 1) Before you go too far, be sure the bleeder isn't frozen. It may need to be retapped.

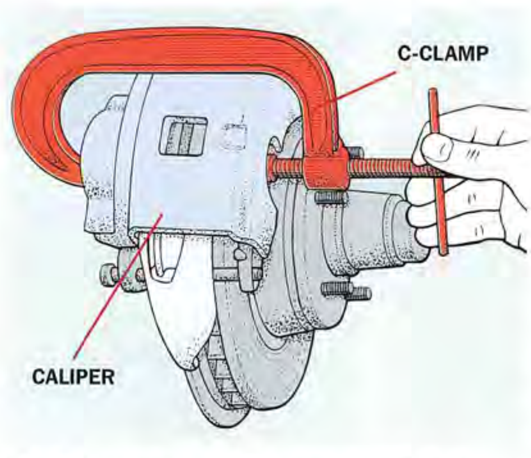

(Fig. 2) Some types of sliding calipers can be pushed back without removing them from the caliper mounts. Be sure it's not seized sliding surfaces that are binding, though.

Brake lights in front of you flash suddenly, and for no discernible reason. You mash the brake pedal to avoid rear-ending a driver who must have suddenly remembered leaving the stove on. Your car stops, but you've swerved well into the adjacent lane, where, fortunately, an alert driver slowed and pulled over to give you some clearance. Too close. Your panic braking attempt should have had two different results: Your car should have slowed more rapidly, and should have tracked in a straight line as it slowed. There's obviously something wrong.

Check It Out

Your car pulled to the left, but the problem is more likely to be at the right front wheel. It didn't brake as well as the left, so the car pulled in the opposite direction.

Remove the wheel and take a look. Don't be surprised if you see that both shoes have a reasonable amount of brake lining left. The problem is more likely to be a bad caliper, that big hydraulic clamp that forces the brake shoes against the disc-brake rotor.

Most cars have sliding calipers with a single hydraulic piston. When you step on the brake pedal and that piston is pushed out, it forces the inner brake shoe against the inner side of the rotor. The caliper simultaneously slides inward, pulling the brake shoe in the outer side of the caliper against the outer surface of the rotor. Some cars have fixed calipers with one or two pistons in each side of the caliper (two or four pistons total). When you step on the pedal, all of the pistons force the brake shoes against the rotor.

Check the brake fluid level in the reservoir and if it's been topped up to maximum level, siphon out a small amount (otherwise it may overflow during the test). Set up the clamp so the top edge of the C is against the midpoint of the inboard side of the caliper (behind the piston) with the tip of the forcing screw directly opposite. Depending on the shape of the caliper, the forcing screw could also bear against the back side of the outboard brake shoe or even against the rotor surface. If the vehicle has antilock brakes, loosen the brake bleeder screw (see page 188).

Turn the forcing screw and the caliper should move smoothly as the piston is pushed back.

If the caliper passes this test, make one more (if you haven't already): Check the bleeder valve to be sure it loosens (**Fig. 1**). If it doesn't, you can drill it out and install a replacement. But this can be a difficult operation, and installing a remanufactured caliper is a safer bet. If the caliper binds or moves only with unusual effort, there are two possibilities: The caliper piston is frozen in place (the most common), or the caliper is not able to slide because of corroded bolts (along which most calipers slide), cocked or damaged bushings through which the bolts go, or rusted sliding "ways" (guides) in the anchor that holds the caliper. You'll have to remove the caliper to see.

Sliding Caliper

Taking off a sliding caliper is a straightforward operation (**Fig. 2**). Usually there are two bolts that hold it to its anchor frame. Or, there's a single bolt at the bottom and a locating stud at the top, a design called a flip up. Clamp off the brake hose close to the caliper with a C-clamp or padded locking pliers to minimize loss of brake fluid and to reduce the necessity of after-the-job bleeding. Next, loosen the bolt that holds the hose to the caliper. Remove the bolt or bolts holding the caliper.

If the caliper is the type that is held in machined guides, there are two common setups. There is a lower guide that can be unbolted and driven off the anchor. Do this, and then lift the caliper up from the bottom of the anchor and off the upper guide. Or, there are bolt-on retainers at the top and bottom to hold the caliper against the guides. Remove both, then pull the caliper away from the anchor.

Replacement

With the caliper off the bracket, unthread the brake-hose banjo bolt and catch any drops of fluid with a rag. Install a replacement caliper, using new sealing washers for the brake hose (Fig. 3). Many caliper bolts are not recommended for reuse, so if the vehicle manufacturer says this is the case, get new ones (Fig. 4).

Making It Slide

If the caliper piston returns smoothly, you'll likely find that the caliper isn't sliding properly because the movement of the caliper bolts is hindered by corrosion or sticking in the caliper ear bushings. Or, the machined guides are rusted. You can use a wire brush to remove rust from the guides (and the caliper edges), then lubricate both with suitable water-repellent grease. Heavy rust may require a light touchup with a file. However, if the bolts are sticking in the bushings, replace them, and lubricate with the same type of grease.

If the O-rings are deteriorated, replace them, or they could allow the bushings to cock. In any case, clean the caliper ears, then relubricate the O-rings.

The Details

Don't leave out any dust or moisture boots from the caliper ears. And if the old ones are deteriorated, replace them (they could be the cause for corrosion seizure of the sliding caliper bolts in the bushings). If a sliding rear caliper is reusable and it incorporates the parking brake mechanism behind the piston, be sure to thread back the piston (turn back the internal screw mechanism).

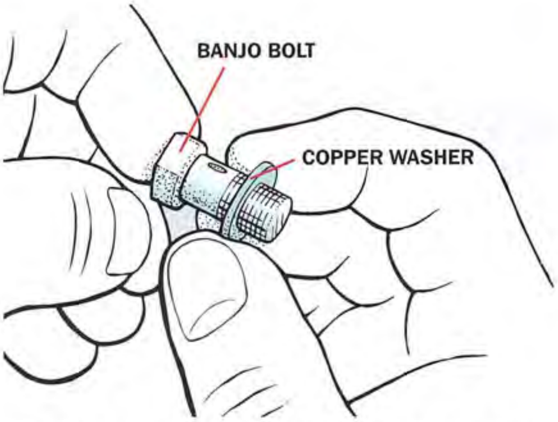

(Fig. 3) Always replace any copper sealing washers on brake line connections with new ones.

Fixed Caliper

With a fixed caliper, you won't know if the pistons are sticking until you take it off. So don't loosen the brake hose (although you should determine if the bleeder valve is frozen). Remove the mounting bolts—usually just two but possibly four. Hang the caliper with wire to relieve any tension on the

HOW IT WORKS

Rear Disc Caliper with Parking Brake

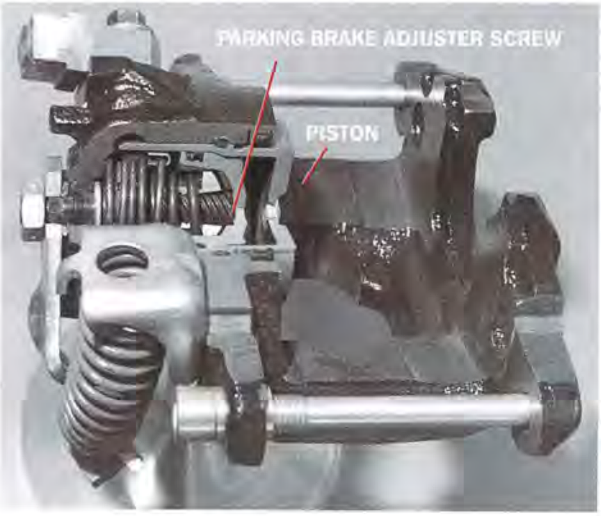

Rear calipers with parking brakes require special techniques to retract the pistons at pad replacement time.

If the parking brake uses the rear disc-brake shoes instead of a separate set of drum-brake shoes, there is a mechanism in the caliper piston cylinder to push the piston and apply the rear brakes when the driver operates the linkage. That mechanism, which is connected to the parking brake linkage, has a screw setup to compensate for the fact that in normal foot-brake operation, the piston automatically moves out to adjust for brake-shoe lining wear. The type shown here is a common design. There is a cone behind the piston and it moves with the piston when the foot brake is applied. When the parking brake is operated, the brake lever turns the screw on which it is mounted (it also is an adjusting screw threaded through an adjusting nut).

The nut can't turn because it's splined into the cone. With the inboard brake shoe in place, the piston (and shoe) can't turn either. So the lever operation causes the piston to be pushed out to apply the disc-brake shoes. When the lining wears, some clearance develops between the piston/cone and the nut. When the foot brake is released, hydraulic pressure in the caliper cylinder also is released. The adjusting spring then causes the nut to thread out on the adjusting screw to take up the clearance.

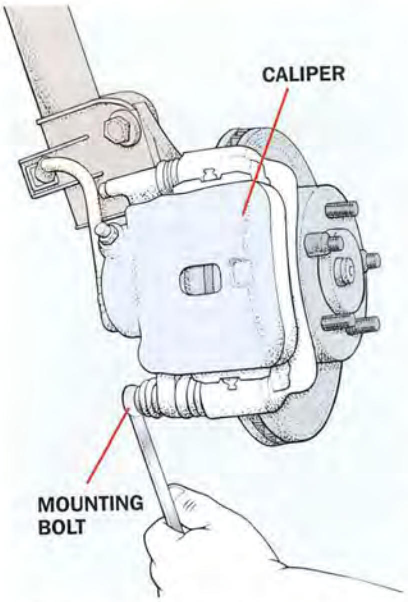

(Fig. 4) Some makers say that the mounting bolts cannot be reused. Check before removing.

brake hose **(Fig. 5)**. Now use the C-clamp to check the pistons. Place a brake shoe against a piston (or pair of pistons) to distribute the force evenly, and put the C-clamp forcing screw against the shoe. Then, turn the forcing screw to determine if the piston moves smoothly. If it's sticking, replace the caliper.

A fixed caliper will be more expensive than a slider, and you may be tempted to salvage one by drilling out a frozen bleeder. Don't take the risk. Let a machine shop do it for you. Also note if the fixed caliper bolts are reusable.

Restricted Brake Hose

It's rare, but you may find that there's no leak and the problem isn't at the caliper. It's possible that the brake isn't applying normally because there's inadequate hydraulic pressure to the caliper, the result of a restricted hose. Have a helper apply the brake and see if you can turn the wheel with a wrench on a lug nut—make a real effort and use a wrench with a long shank. If the wheel can be turned at all, no matter how much effort you're applying, the caliper isn't getting full hydraulic pressure. To check, open the bleeder valve fully (with a hose over it and the other end in a glass jar) and have a helper apply the brakes. Look to see if there's a solid column of fluid coming out of the hose. There isn't? Either a brake line is kinked or the hose is defective.

(Fig. 5) Check for a frozen caliper by attempting to push the piston back. Don't hang the caliper by the hose—use a piece of wire.

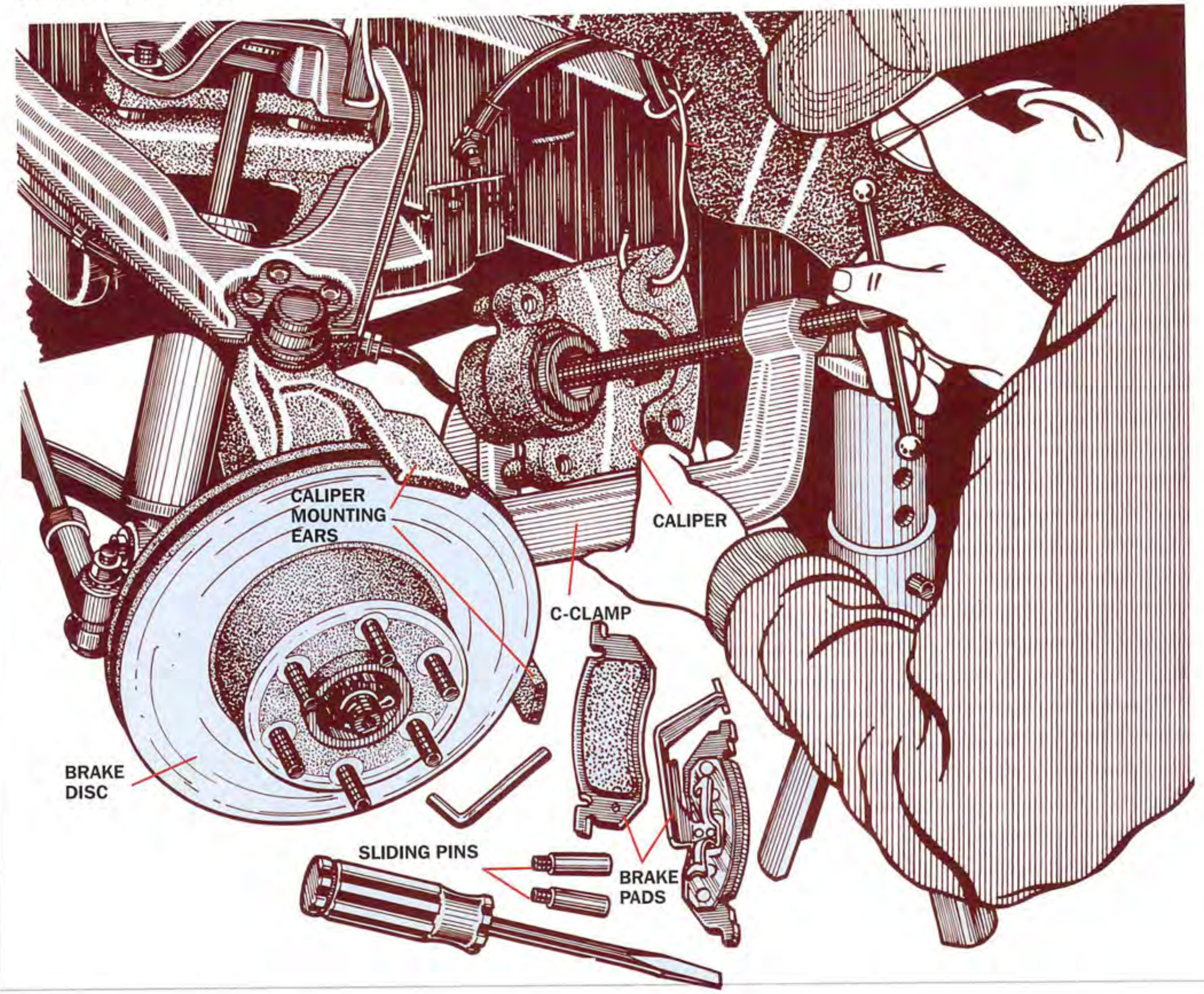

MODERATE

Flushing Your Brake System

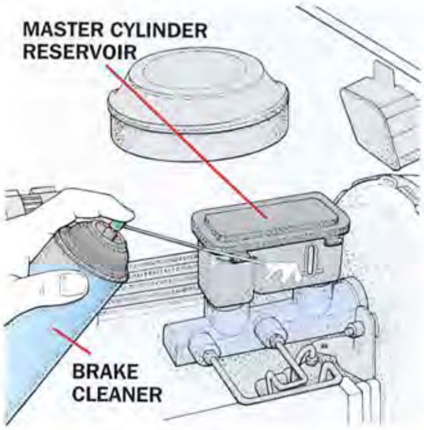

(Fig. 1) Clean the master cylinder reservoir to prevent dirt from falling in.

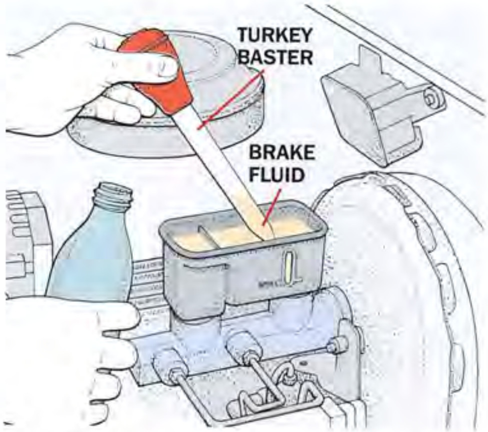

(Fig. 2) Remove almost all of the old brake fluid with a turkey baster before adding fresh fluid.

It's a dark and stormy night, with steady rain and scattered lightning. But traffic moves along smartly—until it doesn't. You wind up hammering the brake pedal to keep from eating the license plate in front of you. Naturally, you expect—you count on—the antilock braking on your new car to ensure a safe, controllable, steady stop. Instead, the car begins to fishtail and it takes a monumental struggle to keep the car in a single lane as it comes to a stop. The extra distance you had from the car in front of you saves a collision, but you know something is wrong with the antilock braking system, a.k.a. ABS.

The shop gives you the bad news: The actuator is defective, and you can figure on $1400 for a replacement (or even worse, it's part of an integrated unit with the brake master cylinder and the price is even higher). No, it didn't log an ABS trouble code, but that doesn't mean anything. It's not a bad connection and it's not a factory defect. What caused it? Maybe even the technician will shrug his shoulders, but there's a good chance it's contaminated brake fluid.

The brake fluid reservoir is vented, so that's an entry point for dirt and moisture. And the rubber brake hoses are permeable—they allow even more moisture to get into the fluid. Dirt and moisture move through the lines. Most goes to the calipers and wheel cylinders, and some may get to the ABS actuator, where it can cause the delicate solenoids or motors inside to behave erratically.

And if you or the shop does a disc caliper brake job, and someone pushes back the caliper pistons, the possibilities are ominous. A lot of the dirty, moisture-laden fluid in the calipers is pushed back, where some of it gets into the ABS actuator. The prospects for an ABS failure go up, up, up.

Even if you don't have ABS, contaminated brake fluid can affect caliper and wheel cylinder bores. The older systems with the reservoir built into the master cylinder are even more prone to contamination, because the reservoir cover gasket may have taken a set and be leaking. Or a lot of dirt may have gotten into the system when the cover was removed as part of a brake job. Even simple moisture in the fluid is a real problem, because the heat of braking will cause it to boil, causing brake fade.

Flushing the system is not a difficult job. There is equipment for a one-man bleeding job, but the good stuff is expensive and we have reservations about trying to use the low-cost alternatives (hose with a one-way valve, and manual vacuum pump are examples). A simpler, effective approach is to have a helper step on the brake pedal. If you can line up someone, the only equipment you need is:

- Ramps or jack and stands to raise the vehicle a few inches in front and back, so you have access to the bleeder valves. However, it's possible to reach them on some vehicles with the wheels on the ground.
- Piece of clear hose to fit on the bleeder valve.
- Tight-fitting wrench for the bleeder (like a tubing wrench). Don't use an ordinary open-end wrench.
- Spray can of automotive cleaning solvent and a spray can of penetrating oil.
- Turkey baster to draw fluid out of the reservoir. A baster costs under $1, so don't try to clean and reuse one from the kitchen.
- A pint container of brake fluid for an econobox, a quart for a larger car. The brake fluid may be labeled DOT 3 (minimum boiling point of 400°F), DOT 4 (minimum boil of 450°) or even DOT 5.0 or 5.1 (500°F). Your system contains DOT 3 or DOT 4. These two are fully compatible, so you can mix them without worry, and one of these is what you should use.

Brake fluid is hygroscopic, which means it's like a sponge. But it's supposed to be, so if any moisture can get in, the fluid embraces it. However, moisture causes the normal boiling point of the fluid to drop. Just 2% moisture in a fluid is considered excessive. There are moisture testers, but we don't know a repair shop that has one. And there's no guarantee that the moisture level in the reservoir is the same as at the caliper—it may be higher at the caliper.

Warning: There are silicone brake fluids (labeled DOT 5) that are immune to moisture (5.1 is nonsilicone, but DOT 3 and 4 are what you should use). Don't even think about using DOT 5 silicone fluid in an automotive braking system, because it may cause faster wear of the seals, and if enough water gets into a system it could cause loss of pedal. DOT 5 and 5.1 have application in some racing vehicles.

Flush Twice

Now for the flushing job itself: Raise the vehicle a few inches on ramps or with a jack and stands.

Begin by cleaning the master cylinder's brake fluid reservoir (or top and sides of the master cylinder with the integral reservoir), using aerosol cleaning solvent and paper towels or a clean rag (Fig. 1). Don't remove the reservoir cover until the area is squeaky clean.

If it's an older style—the master cylinder with the integral reservoir—you'll find a rubber gasket. Inspect the perimeter for deterioration and replace the gasket if necessary. Otherwise, just clean it with some fresh brake fluid. Siphon most of the fluid out of the reservoir (Fig. 2), then add fresh fluid up to the level mark.

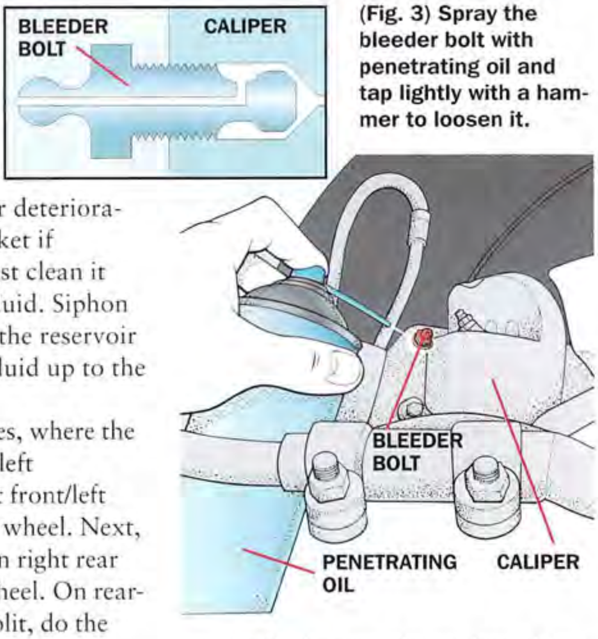

(Fig. 3) Spray the bleeder bolt with penetrating oil and tap lightly with a hammer to loosen it.

On front-drive vehicles, where the hydraulic system is split left front/right rear and right front/left rear, start at the left rear wheel. Next, go to the right front, then right rear and last, the left front wheel. On rear-drives with a diagonal split, do the same. If the rear-drive split is front/rear, however, start at the right rear, then left rear, right front and left front. The object is to begin at the wheel brake farthest away from the master cylinder and gradually work toward the closest.

If the bleeder valve can be loosened with the wrench, you're golden. If not, spray it first with penetrating oil (Fig. 3), let that work in, and try again with the wrench. Won't loosen? Clean the penetrating oil and heat the bleeder with a propane torch. Try again. If your wrench handle isn't that long, or if the fit isn't very tight, you

HOW IT WORKS

Testing Brake Fluid with Dip Strips

If the brake fluid in the reservoir is obviously dirty, flush the system. If it isn't, flush if it contains excess moisture, which you now can determine with "reagent dip strips." Reagents are substances used in a chemical reaction to detect other substances, and Wet Check strips for brake fluid now are marketed by Wagner Brake. Insert the strip into the fluid, remove it in a second, wait 30 seconds and compare the change in color of the two pads with a chart.

WAGNER
Brake Products
WET CHECK

One pad merely identifies the type of fluid—DOT 3 or DOT 4. They're compatible, so this is not critical (and a 50/50 mixture of the two merely produced an ID color close to that of DOT 4, in our tests). The pad closest to the end, however, turns tan if moisture content is excessive.

We tested the strips in separate containers of DOT 3 and DOT 4, with different percentages of water. With pure brake fluid, the color was in the definitely okay range: dark green. At 1.64% water content, there was a very slight indication (some lightening of the green) that the fluids were less than pure.

At 3.25% water, there would be at least a 25% drop in brake fluid boiling point with DOT 3. It could be as much as 50% with DOT 4. The color change to tan was unquestionable in three of four tests. In the fourth, there was some indication of less-than-perfect fluid—fuzziness in the light green—but not tan.

At 5% water the pad turned tan as soon as it was taken out of the fluid—we didn't even have to wait the 30 seconds. Read the strip almost exactly 30 seconds after removal from the fluid. If you read it too early, you might get an okay indication with 3% water. If you put the strip down for a few minutes, it can absorb enough moisture from a humid room to indicate contaminated fluid when none is actually present.

can try locking pliers (Fig. 4). Some tapping on the wrench or pliers end with a hammer may help.

When the bleeder just breaks loose, stop and push the end of the clear hose onto the bleeder nipple. Aim the other end of the hose into a clear glass or bottle (Fig. 5).

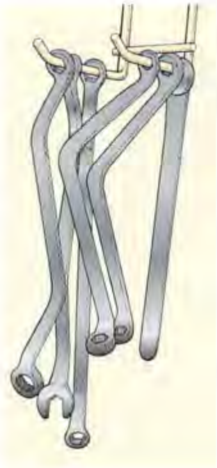

(Fig. 4) You may need to use a special wrench.

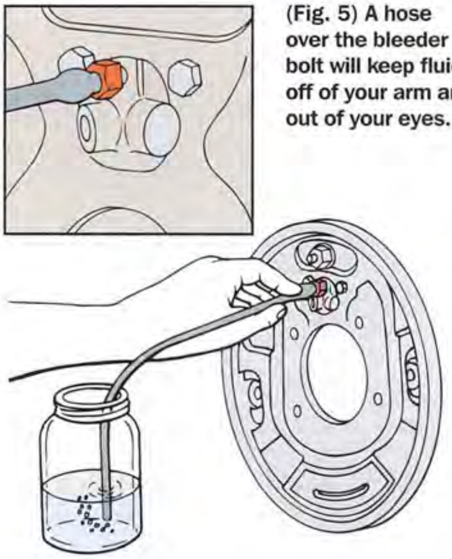

(Fig. 5) A hose over the bleeder bolt will keep fluid off of your arm and out of your eyes.

Put a block of wood under the pedal to prevent the piston from bottoming out in the master cylinder. Then tell your helper to step on the pedal and press down gently but firmly. Open the bleeder valve and watch the fluid as it flows through the clear hose into the clear glass or bottle (Fig. 6). If you don't get any fluid movement at the rear brakes on a car with ABS, close the bleeder valve. Have your helper turn on the ignition and apply the brakes once, then turn off the ignition and just lightly rest his/her foot on the pedal. Open the bleed valve slowly. If a stream of fluid flows from the valve, allow it to come for about 10 seconds, then close the valve and add fluid to the master cylinder. Depending on the system, the pressurized reservoir may push out enough fluid. Or the flow may continue only if the helper presses down on the pedal. On most cars, the flow will start and continue without the preliminaries—only the helper's foot pressing on the pedal.

When the pedal hits the wood block, tell your helper to hold it there. If you had a lot of trouble loosening the bleeder, remove it completely. Apply a thin film of anti-seize compound to the threads of the bleeder valve, then reinstall and close the valve. Add fresh fluid to the master cylinder reservoir again and repeat the procedure. When the fluid color changes from grungy black to something closer to what the new fluid looks like, you can stop. Close the bleeder valve and go to the next wheel brake. Be very sure to add fresh brake fluid up to the level in the reservoir every single time.

Before you do the final top-off, however, have your helper apply the brakes over and over, to exhaust the high-pressure reservoir of the ABS. This will cause the level in the master cylinder reservoir to rise slightly. It may take 25 to 35 pedal applications to do this, but when the level stops rising after a couple of dozen, that should be it. If you don't do this, the reservoir might overflow under some conditions.

When you're done, a hard brake application should extinguish the brake warning light.

And remember when you next do a disc brake job, don't just push back the piston with a C-clamp. First, clamp the brake hose with locking pliers, wrapping the hose with a protective sheath of thick rubber, perhaps from an old radiator hose. Open the bleeder (and with antiseize on the threads, it should cooperate quite nicely actually), attach a hose to the nipple, then push the piston back, collect the fluid in a container and discard.

This job is about more than saving money on an ABS actuator. No matter how new or old your car, it's about saving your skin.

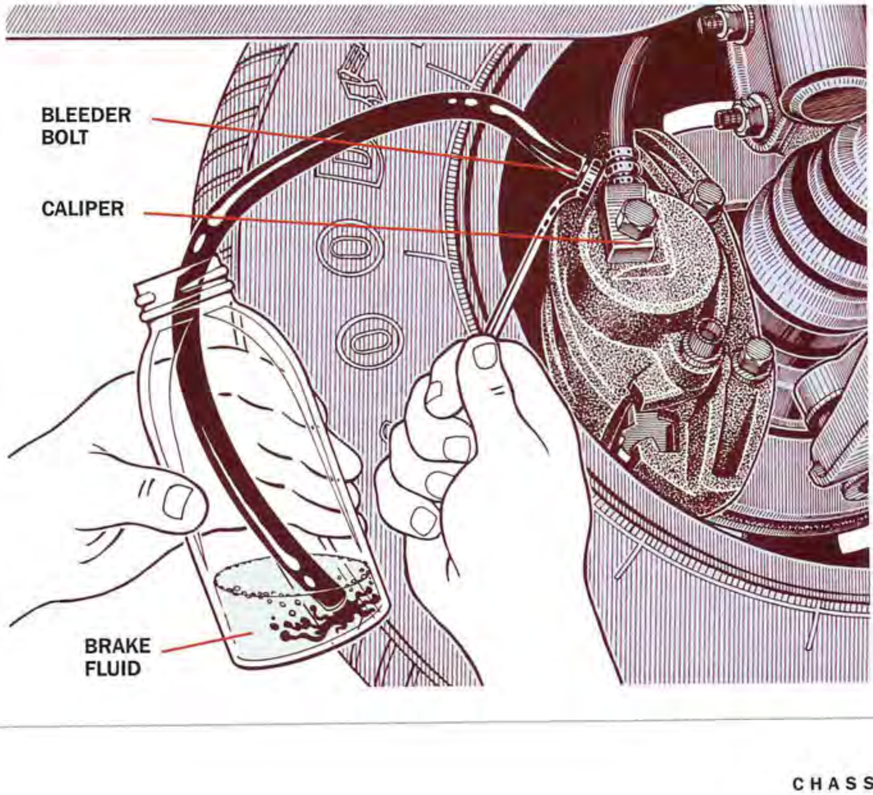

(Fig. 6) Use a box wrench or tubing wrench to loosen the bleeder bolt while a helper depresses the brake pedal. Tighten the bolt after brake fluid stops flowing, but before your helper releases. Repeat until clear fluid comes out.

Servicing Drum Brakes

You're halfway down the hill when you realize how much speed you've picked up. So you hit the brakes. Hard. Too hard, apparently, because the front tires repeatedly lock and aren't able to kill off much velocity. At the first turnoff, you check your front brakes—hot as a cup of McDonald's coffee. But your rear brakes are cool. Obviously, the front disc-brake job you had done over the weekend wasn't enough. It's time to look at those rear drum brakes.

Diagnosis

Block the front wheels. With the vehicle's rear off the ground and on safety stands—and with the parking brake off—turn each wheel. If you hear a uniform light scraping noise, that's okay. Otherwise:

- If the wheel is binding all the way around the rotation, the brakes aren't releasing properly. If that's the case, remove the drum and check for either one of two possibilities: 1. The brake shoes are not retracting fully because the return springs are weak, perhaps combined with the shoes hanging up on the backing plate; or 2. The automatic adjuster for the shoes is overdoing it.
- If you get an intermittent scraping or binding, maybe the drums are out-of-round, the shoes aren't retracting properly, a backing plate is warped or an axle problem exists.
- If there's no hint of a scrape—despite the fact that your rear brakes have been noisy or haven't been contributing to the car's braking—they may need to be adjusted. If there's an access opening in the backing plate or drum, turn the adjuster until the shoes make light contact with the drum. Then rotate the wheel to see if the lining drag is relatively even all around. If you can't turn the adjuster, take off the drum to free it.

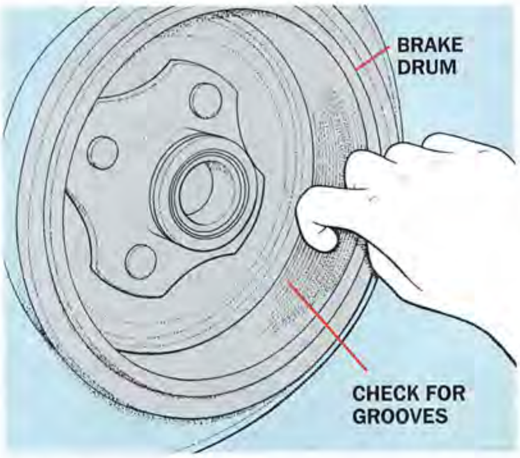

(Fig. 1) The surface of the brake drum should be totally free of any grooves, cracks and shiny spots.

Look As You listen

To gain a better look at the brake drum, remove the tires and wheels. Now, as you turn the brake drum, eyeball its relationship to the backing plate at various points. If the drum seems to move in and out or rock up and down—even slightly—the wheel bearing is bad, or the wheel spindle or axle shaft is bent. If the drum appears to turn evenly but its relationship with the backing plate is uneven, perhaps the plate is bent—also not an unusual problem. (If you're not sure, compare one rear drum brake with the other to see if the appearance is different while turning it.) Either way, you'll have to pull the drums to confirm the cause.

Inspect the drum where the wheel mounts against it. If a wheel is deformed, the drum may be deformed as well. Also, compare the tire sizes side to side and front to rear. Odd-size tires can trigger a vehicle's antilock brakes at the wrong time, affecting the rear-braking performance and, in some systems, setting off an ABS trouble code.

Drums Off

Pulling off stuck brake drums may require a penetrating solvent. You might even have to apply—gently—heat from a propane torch to the stud holes and the drum/hub joint. If a drum is still stuck, you may need to use a screw-type drum puller.

Once the drum is off, check its shoe-contact surface for any grooves in which you can catch a fingernail (Fig. 1). Discs work well enough with deep scores (although they may get noisy), but drums don't. And even if a drum is super-smooth, get it measured at a machine shop to see if it's out-of-round, worn past specs or misshapen in some way. Normally (but not always), a faulty drum will create an uneven lining wear pattern or cause short lining life.

Find a drum that needs cutting? Take both drums to the machine shop, and have an equal amount removed from each. There may not be enough metal remaining, which means new drums. After a drum is machined, clean it out with a detergent and water solution, then wipe it with a lintfree rag and a brake solvent. Wipe until the rag comes clean, and let the drum dry.

Next, inspect the linings. Even if they're thick enough to be left in (⅛ in. or more), they may have a glazed surface and need replacing. Don't try to sand off the glaze. The linings have undergone a chemical change from overheating, possibly combined with the effects of aging as well as climate.

The linings may also be glazed because the brake shoes were dragging. Perhaps the automatic adjust-

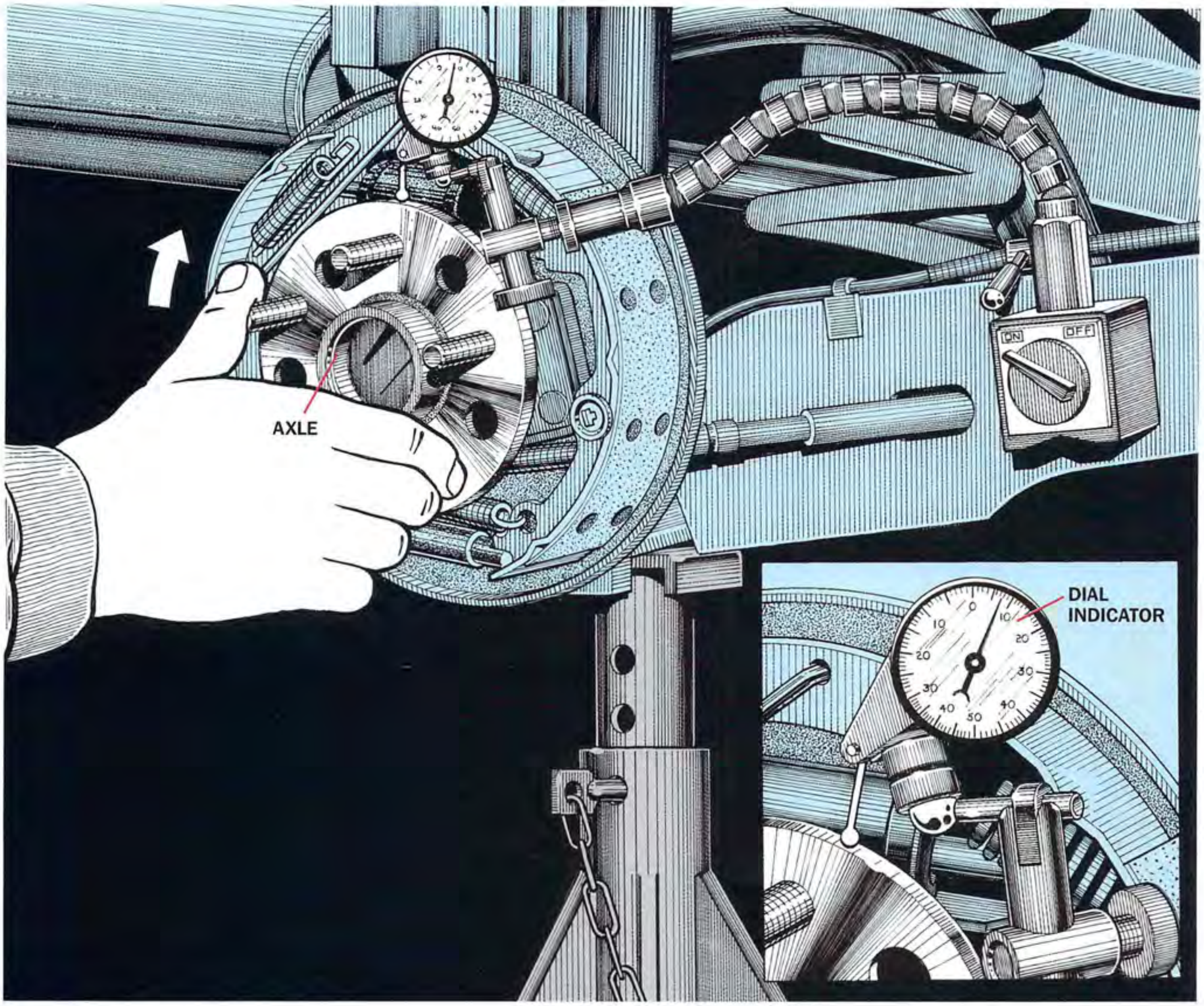

(Fig. 2) After you've removed the brake drum, use a dial indicator to check for both axial play and radial runout of the rear-axle hub. For good braking performance, total runout should be only a few thousandths of an inch. Excess runout probably indicates a bent spindle.

ment mechanism is bad—or someone keeps forgetting to release the parking brake before starting to drive.

Running Out—Literally

Did you encounter that intermittent scrape when you turned the drum in a pre-disassembly check? But despite the noise, did the drum and shoe hardware seem to be good? This is the time to check for a bent spindle or axle shaft, and/or a bad bearing. Use a dial indicator (**Fig. 2**). (If you don't own an indicator, you can rent one from an auto parts store or rental center.)

To check the radial runout, mount the dial indicator on the suspension, body or chassis so that the plunger is against the edge of the spindle or shaft hub. Then turn the axle shaft or spindle. Caution: Turn the axle slowly, so you won't be fooled by a burr or nick on the hub—particularly one on the edge—as you're checking radial runout. To get an accurate reading, you may need to clean up some corrosion on the hub with sandpaper or a wire brush.

To check for lateral runout on axle shafts with a hub, relocate the dial indicator so the plunger is touching the outer face of a hub. Then turn the shaft again.

Runout, in any case, should be within manufacturer's specs—typically just a few thousandths of an inch. If your brakes have objectionable shudder, you'll probably find that the runout is .010 in. or greater.

Putting It All Together

Ready to install new shoes, springs and other hardware that come in the typical drum-brake repair kit? Don't use a screwdriver or your neighbor's 15-year-old set of brake-spring pliers —a spring could fly off and smack you right in the mouth. Instead, get a pair of spring pliers (**Fig. 3**) designed for late-model cars—it has jaws that engage the spring hooks and the shape to actually expand the spring enough to remove it. Ditto for late-model brake holddowns, which come in more than one size and call for a

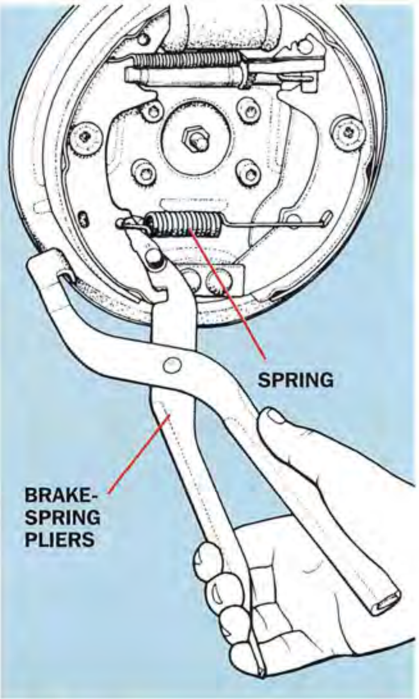

(Fig. 3) Brake-spring pliers make the process of reinstalling springs both safe and easy.

new tool with the two popular-size tips. Clean off the shoe side of the backing plate and make sure that the wheel cylinder isn't leaking.

How should you install semimetallic shoes if you live in an area with cold, wet winters? If they freeze to the drums after an overnight cold snap, the first brake application may be uneven. Chamfer the ends of the shoes (up to the first rivets) and they'll break loose more easily and retract when you take off (Fig. 4). If the problem occurs even at temperatures just above freezing, you might want to check with the car dealer—he may have a substitute lining (with a lower metal content) that's less affected by the cold weather.

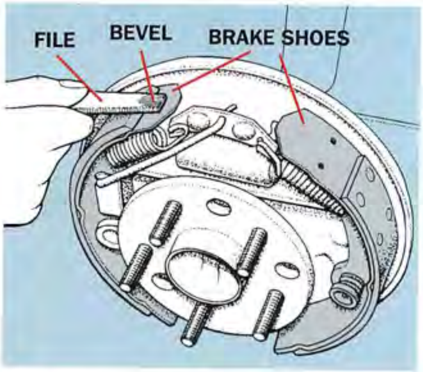

(Fig. 4) Chamfering the leading edges of the lining material can improve your car's braking performance in certain weather conditions.

If pre-disassembly checks indicate a warped backing plate, don't try to bend it back into position—install a new one. On many rear-drive General Motors cars built from the late 1970s through the early 1990s, there may be another problem: wear in the cutout for the clip-held wheel cylinder. The result is a loose cylinder. If you can force it to rotate even slightly—say, by moderate prying on each end of the cylinder with a large screwdriver—that's reason enough to install a new backing plate.

The "by-the-book" method of replacing a brake backing plate on a rear-drive car is to remove the axle shaft. This is a tough job, so if you need a new plate, it's a lot easier to install an aftermarket 2-piece plate (available for many GM cars) (Fig. 5). Just loosen and cut the old one with a hacksaw and you can bolt on the new one without breaking a sweat.

If the platforms on which the shoes slide are burred or rusty, they're probably a contributor (along with weak springs) to the shoes failing to retract quickly—not to mention overheating or premature wear. Smooth off the platforms using fine sandpaper, wipe clean and apply a coat of silicone dielectric grease. Also, you should sand smooth the edges of the new brake shoes where they ride on the platforms.

If you saw a lot of road film on the inside of the backing plate, or evidence of water/snow entry in winter, seal up the back of the plate. If any inspection/adjustment covers are missing, buy and install replacements. And if you can see any light coming through the wheel cylinder cutout in the backing plate, seal it.

Finally, bolt on the wheels and tighten them to specification with a torque wrench, using a crisscross pattern. This is not only important for front-disc rotors, but it will also keep rear drums from warping.

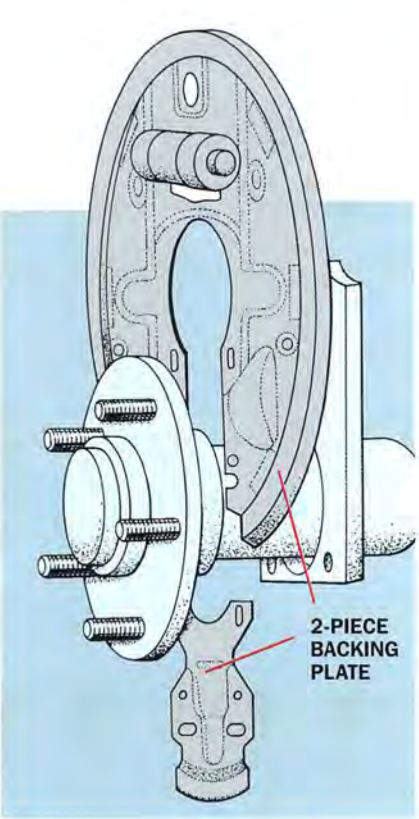

(Fig. 5) Aftermarket 2-piece backing plates can help you avoid the very difficult job of removing the axle.

The bottom line? If you touch all the bases this time around, you'll get 100% braking action the next time you hit that pedal.

ADJUSTERS NEED TO BE free-moving. Check this while lifting the adjuster arm with a screwdriver.

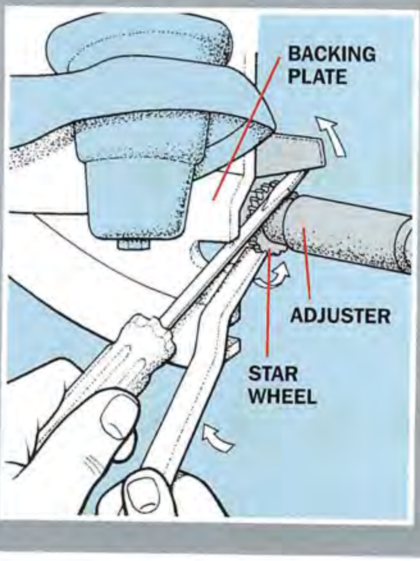

Solving Steering-Pull Problems

Your personal trainer mentions that the diameter of your left forearm is twice that of your right, unusual for a righthanded individual. Popeye would be proud. And your spouse nags you constantly to keep to the right lane of the freeway, because you're always brushing the edge of the shoulder in the fast lane at 5 under. Guess it's time to find out why your car pulls so hard to the left.

The tire pressures are correct. You haven't bounced the front end into a concrete embankment or over a rocky road lately, so there's probably nothing bent in the suspension/steering underbody. Is it just something you'll have to live with? Probably not. The sources of most pulls are pretty easy to pinpoint if you define the basic problem and systematically check the possible causes.

First question: What kind of pull is it? That calls for a road test, made under precise conditions: correct tire pressures when the tires are cold, no wind and a smooth, level road (no crown). Drive at a steady speed (somewhere between 30 and 45 mph), then shift into Neutral. Next, when there's no traffic around, remove your hands from the steering wheel. Note whether the pull occurs in gear and also in Neutral, and if your hands on the steering wheel make any difference.

If the pull is most noticeable under acceleration and goes away when you shift into Neutral, it's called torque steer. And, yes, torque steer also can occur at steady speeds. If it occurs with the transmission in Neutral, it's a typical pull. Still no pull or torque steer? Make a turn and allow the car to return to facing straight ahead. Then, repeat the test to see if the car pulls in the same direction as the turn that was made. This is called memory steer, and you'll see that it's most obvious when you take your hands off the steering wheel.

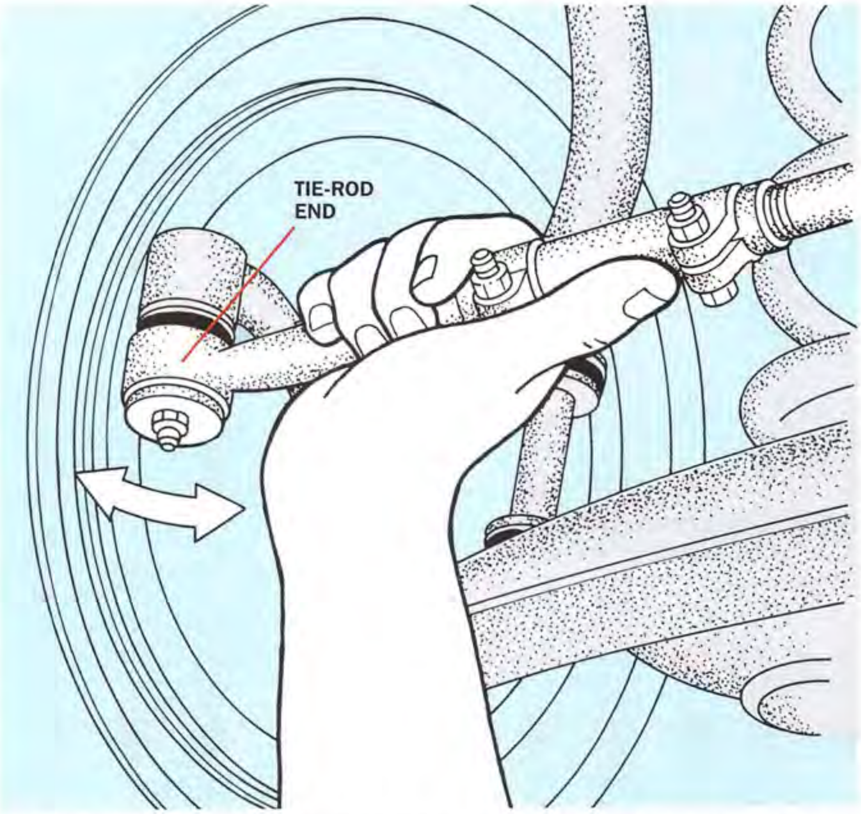

(Fig. 1) Rock the tie-rod ends to check for binding or wear. Some light resistance is fine, but there should be no slop.

Torque Steer

Torque steer results from more possible causes than any other source of pull, so it's tougher to isolate. If you have a powerful engine and you mash the throttle, some torque steer—particularly on acceleration—is normal.

Transverse-engine front-drive cars with high-performance engines are particularly prone to torque steer, although proper suspension design helps minimize it. When significant torque steer is evident, something is wrong and you should be able to trace it.

Checking for brake drag is a good first step. Even slight differences in

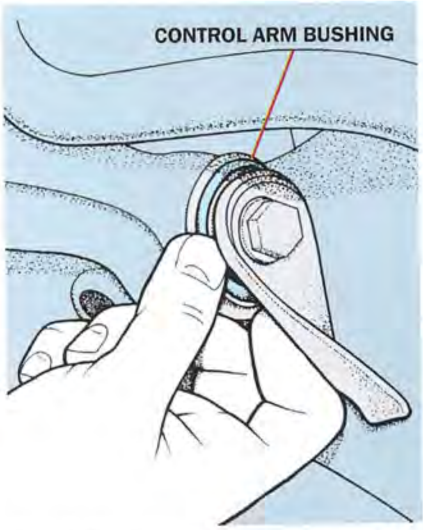

(Fig. 2) Worn bushings can often be detected visually—look for shiny metal where the bushing has been moving.

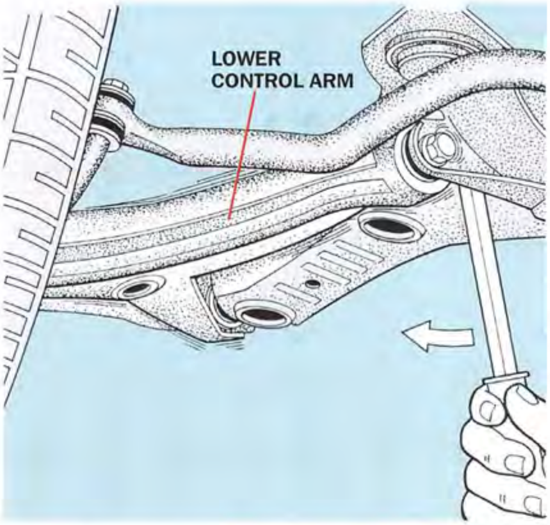

(Fig. 3) Pry the control arm lightly to find worn bushings. The arm should return to its original position when you release the pry bar.

drag or between the front tires can compound any tendency toward torque steer. A second check should be for a difference in looseness of the front suspension or tie-rod ends from one side to the other. You can feel major looseness by grasping and pulling/pushing on the tie rods (Fig. 1). Or you can have a helper hold the steering wheel steady while you rock the road wheel side to side and up and down to look for play in the tie-rod ends.

In addition, eyeball any rubber bushings, particularly at the control arms (Fig. 2). When these bushings are bad, they look it—cracked, distorted. Also, pry the control arms to feel for any looseness (Fig. 3).

Next, measure the front suspension ride height (Fig. 4). Factory service manuals detail the many possible measurement points, but the object is to compare readings between a pair of fixed points—one on the suspension and the other on the chassis. You'll need an absolutely dead-flat piece of pavement for this, and bounce the car up and down on the suspension a couple of times after you park.

One way to check for symmetry—or the lack thereof—in the suspension is to tie a string between the bottoms of the lower control arms (such as from the ball-joint grease fittings). The string should be very tight (and "straight as a string"). Measure from

(Fig. 4) With the car on a dead-level surface, check for correct symmetry in the suspension ride height. Stretch a string between a pair of convenient points on the shocks or ball joints, and measure to the frame rails. It should be level to within less than a half-inch.

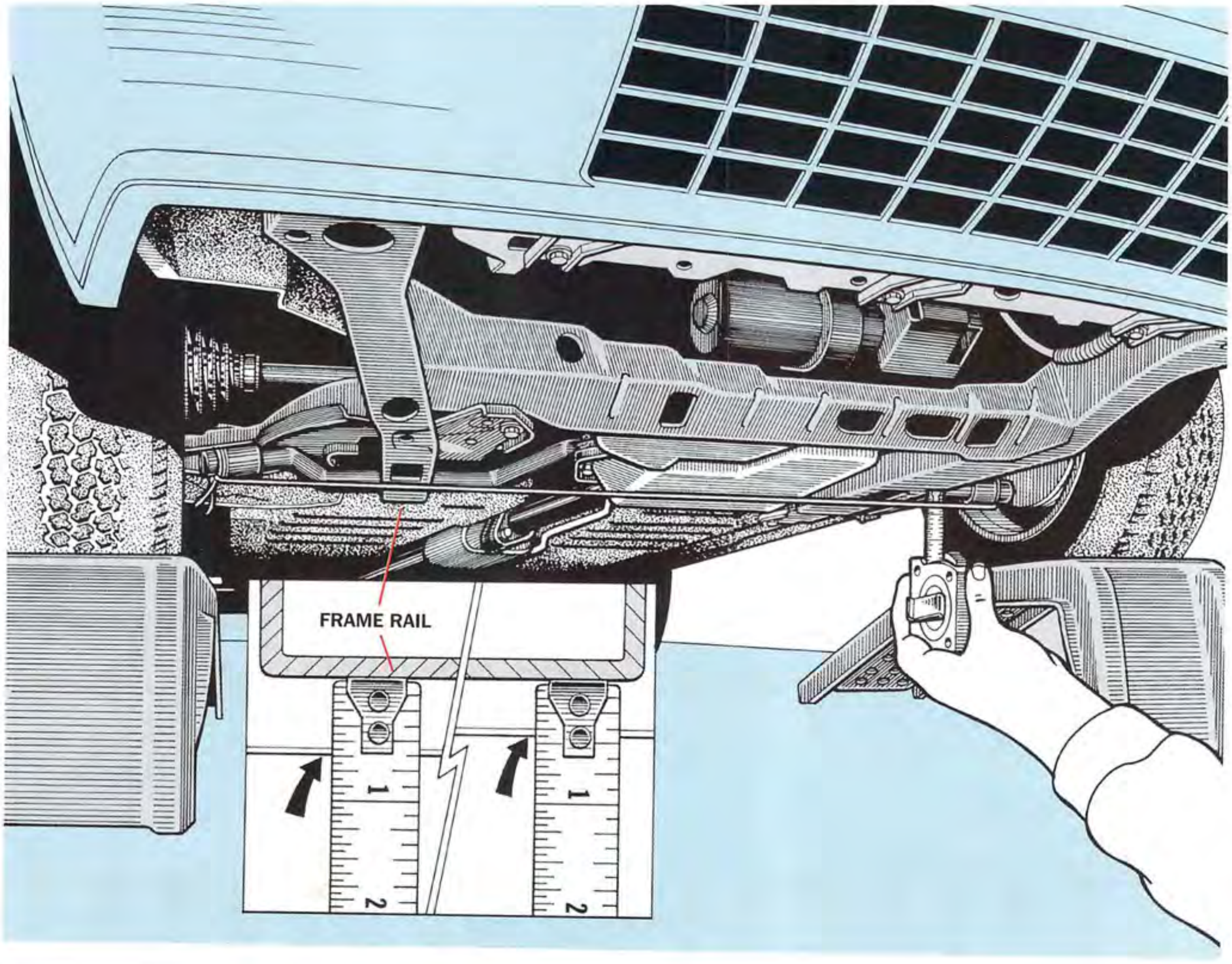

(Fig. 5) You may need to pull the strut apart to properly check the strut bearing for binding.

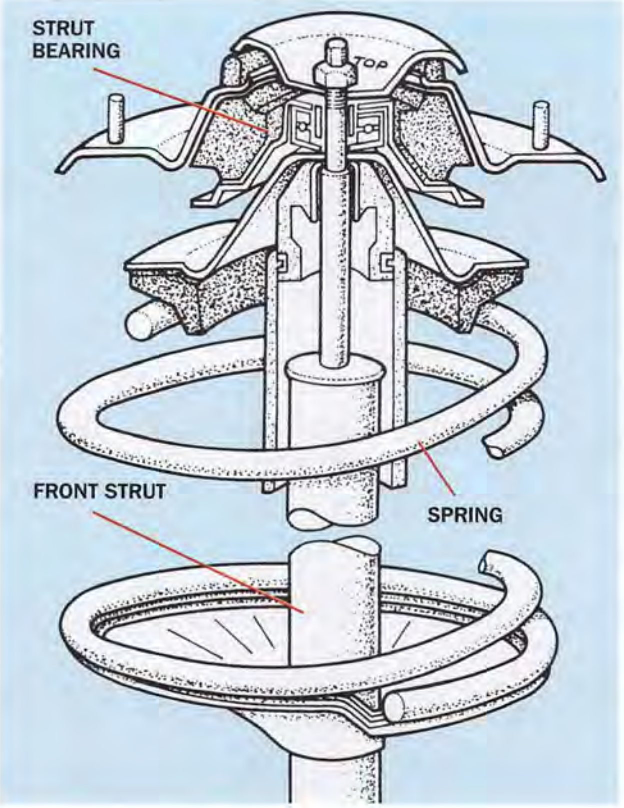

the string to the underside of the frame rail to which the control arm mounts, and the distance should be the same from side to side, within less than a half-inch. If it's greater, there's front spring sag, a common major contributor to torque steer. Replace the springs to fix this problem.

Badly sagged rear springs can cause a vehicle to pull as well, but you'll be able to see the height difference visually before the pull gets bad enough to notice while driving. Just look carefully at your car or truck from the rear when it's parked on that level patch of pavement.

Good springs? Check the engine mounts for deterioration, as weak mounts allow the powertrain to shift unevenly on acceleration, and that can move the car in one direction or the other. If the rear (transaxle) mount is an air-gap type and it's bad, the engine will rock backward, mashing the metal of the mount against the rubber center section. To check for this, have a helper hit the gas pedal while in Neutral.

Also check to see if the transaxle is level, which you can do with a carpenter's level. It doesn't have to be absolutely, perfectly level, but within a quarter-inch or so.

If it's not, try loosening the transaxle mount, running the engine at idle (have the parking and service brakes applied, once more) and, if you have an automatic transmission, shifting the transmission back and forth between Reverse and Drive several times. With a manual box, you can try prying the powertrain fore and aft a few times. This should resettle the powertrain.

Retighten the bolts and recheck the level. If it's still on a tilt, you can try shimming up the low end of the powertrain, using suspension shims.

STRUT BEARINGS CAN bind up and cause the steering to stiffen and pull to one side.

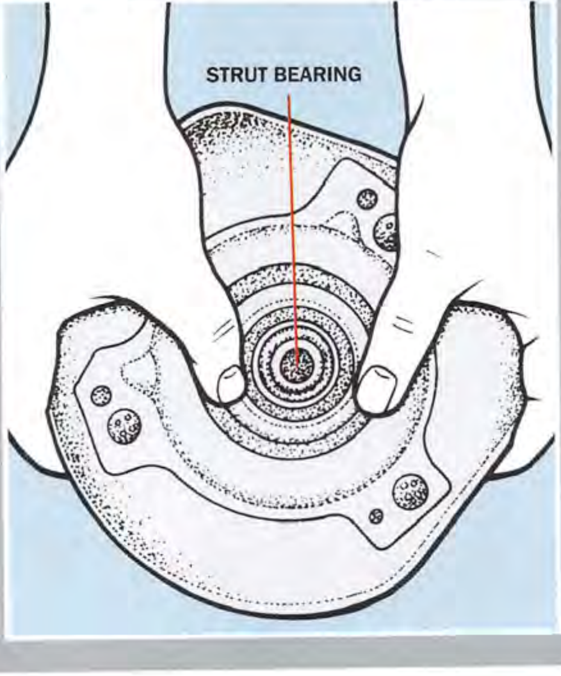

Still looking for answers? It's certainly possible that problems with the front tires, such as misaligned belts—under the tread and out of sight—are a factor. Normally, tire problems are going to cause pull even when you're not accelerating, but it is possible that you are more sensitive to it only during acceleration.

So let's look at the tires, and talk about a straightforward pull.

Tire-Related Pull

Hopefully, you had your alignment checked at a shop that uses first-rate electronic equipment, one that even checks caster and camber on "net-build' (nonadjustable) front suspensions. A good alignment shop knows what is acceptable as a suspension ages and springs sag slightly. And it can slant the camber readings ever so slightly to minimize pull in one direction.

Tires are the first do-it-yourself items to check, even if the pressures are the same and right at specifications. Are they exactly the same size and type, and worn about the same? If not, they're probably at least contributors to any pull. Even if they look exactly the same, the tread belts of one or more may have been damaged.

Get the front end of the car up in the air on proper jackstands for the remainder of these tests. Put the stands under the frame, not the suspension members.

Rotate each front wheel and feel for more than light drag. If you get much more drag at one front wheel than the other, you have a brake-system problem—either a bad caliper or something wrong with the hydraulic system, such as a crimped tube or even a bad master cylinder.

Equal or no drag on both front tires? Switch them left to right. If the pull is in the opposite direction, get new tires. Even if the pull changes somewhat, but not necessarily all the way in the opposite direction, the tires still are the most probable cause of the pull, and a new pair might be worth a try if nothing else pans out. If you're lucky, however, the pull will go away and you can pat yourself on the back for a cheap fix. Stop here if you're checking for a cause of torque steer. Let's try one more thing for a tire-related pull.

If there's no change after the front-tire swap, next switch the wheel on the driver's-side front with the one on the driver's-side rear. If that makes a significant change, get a replacement for the tire now at the driver's-side rear (the one originally at the passenger's-side front). If the tread pattern or amount of wear is significantly different between new tires and the rest that you have, you may have to simply replace all four tires.

Memory Steer

When toe is correct, and the car pulls in the same direction as the last turn, you're dealing with memory steer—almost always caused by binding in the suspension on that side of the car.

On a MacPherson-strut suspension, it's a good bet that the binding is in the upper strut mount bearing. Sometimes, you can feel roughness in the mount by reaching up from underneath and resting your fingers on it, while a helper cranks the wheel back and forth (front end up on safety stands).

More often, you'll have to pull the strut for a better test of the bearing. Check for binding in the bearing after you've pulled the strut (Fig. 5).

No luck? It might be a binding tie-rod end, but that's not likely—they generally get sloppy rather than bind up. Give the tie rods a shake and a twist to check. Also unlikely but possible is a binding ball joint. You may need to disconnect it to confirm this.

It's possible that the binding is in the steering gear. As a final test, you can disconnect the tie rods. With them connected and then disconnected, you should be able to feel the difference by turning the steering wheel in each direction with the engine off.

Unfortunately, not all alignment shops are as methodical as they could be. And some shops will tell you that "all front-wheel-drive cars do that." But unless the pulling problem has been there since the day you received the vehicle, it isn't generic and you should conduct a careful road test and underbody inspection.

MODERATE

Diagnosing and Repairing Wheel Vibration

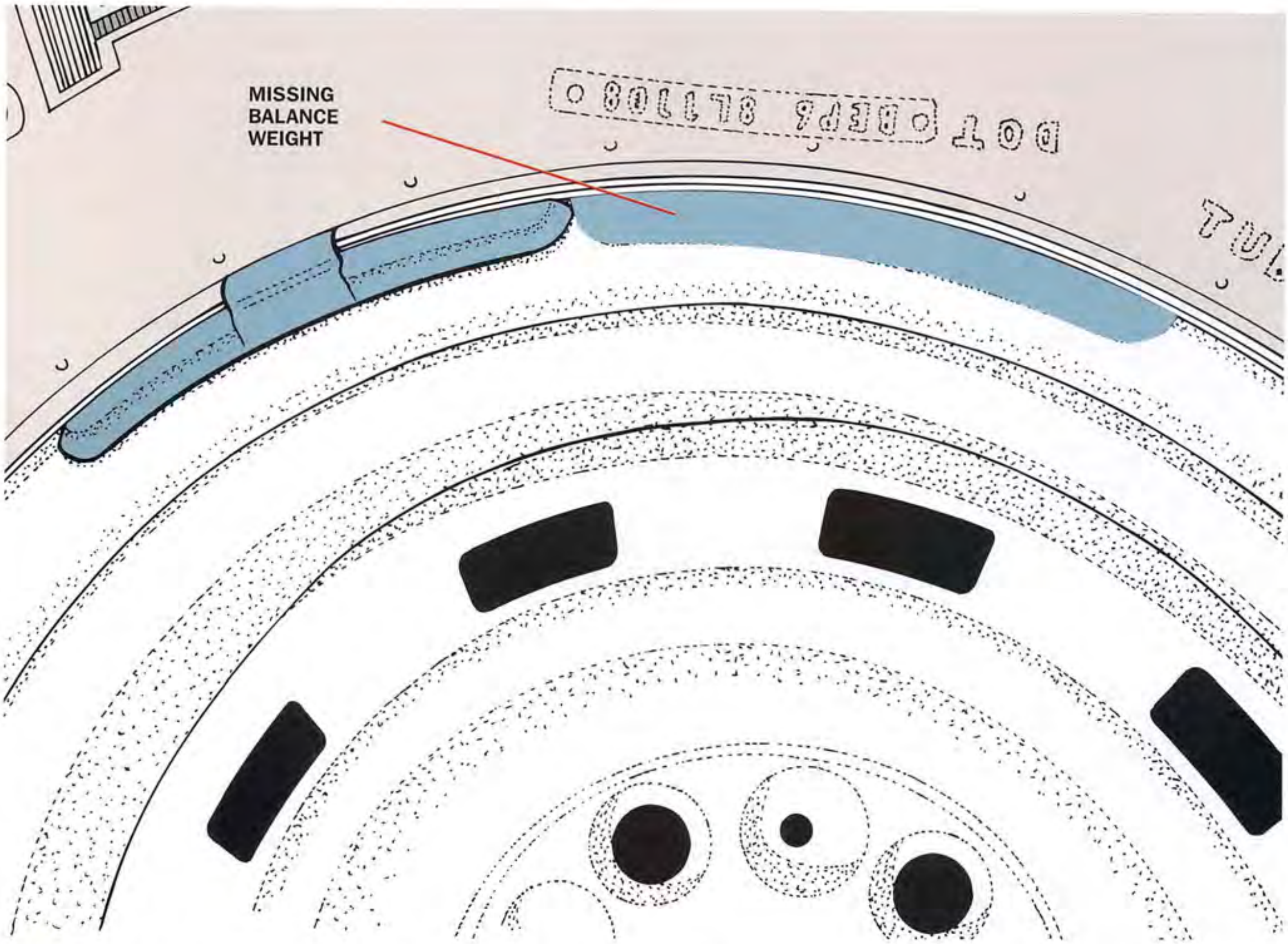

(Fig. 1) Missing wheel weights will leave marks on the rim. Check the inside of the rim, too.

At long last it's the weekend and you're headed for the mountains ... or the desert ... or the shore. Anywhere out of town. And for the first time in weeks you can point your hood ornament at the horizon instead of the license plate in front of you and actually achieve the speed limit on the interstate. Your hands shake with glee.

Actually, that's not glee, or even healthy anticipation. Nor is it some unspeakable neurological syndrome. It's a vibration that isn't even perceptible at lower speeds.

Maintain Your Balance

A simple wheel balance will cure most vibrations. But if that doesn't cure the problem—or if it cropped up suddenly within a reasonable time after a wheel balance—your problems may go deeper.

Begin by cranking the wheels over to the steering stop and looking at the inside of the rim. It's customary to split the amount of the balance weights between the inside and outside of the rim. If a weight (even an old one) has come off, that could be the problem. The weight should leave a clear outline, so you'll know exactly how much is missing (Fig. 1). If the balance weight was added recently, you can take it back to the shop for a replacement.

Also inspect the rim—inside and outside—for any damage. Look for packed mud on the inside of the wheel. Also look at the tires—if you see any bulges or uneven wear of the tires, consider them in the "probable cause" category.

Nothing obvious? Take the car for a test drive. When the vibration occurs, is it while you're accelerating through a bend? That means it's both torque and speed sensitive. When you pull back to your garage, inspect the axle shafts, looking for damage

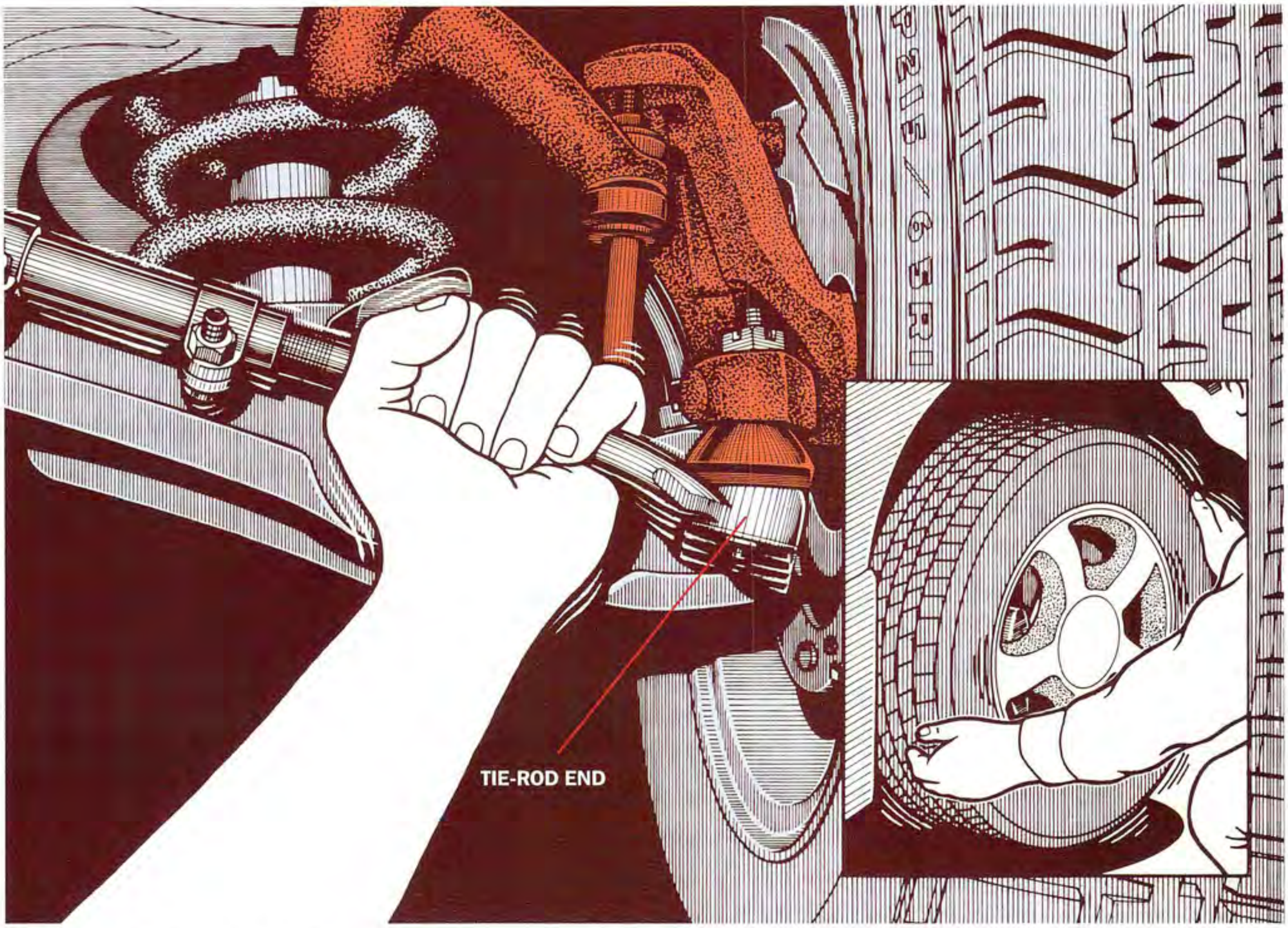

(Fig. 2) With the wheel off the ground, shake it in and out to check for a loose wheel bearing. Then check for loose tie-rod ends.

to the boots. Constant velocity joints can wear out. But if the boots are intact, the clamps are holding them at each end, and there's been no loss of lubricant and no intrusion of road film, then they're probably in good condition.

If the vibration is not related to torque, shift into Neutral and let the vehicle coast at the problem speed. Still have the vibration? It's speed sensitive pure and simple. This could be the source of your troubles, even if the wheels are balanced and the tires are good. It's not a powertrain or driveline issue.

Keeping Your Bearings

Jack up the front wheels by the control arms, so they're off the ground, and support them with safety stands. Grasp each wheel, holding it first at the sides, then at the top and bottom. See if you can rock the wheel in and out and if you can feel any looseness, which indicates a loose wheel or worn wheel hub bearings (Fig. 2). To replace wheel bearings on front-drive cars, you've got to remove the wheel hub. This job requires a slide-hammer puller, a tool typically available from the rental Peg-Board of many auto parts stores, and a torque wrench capable of the high torque usually required for the retaining nut (often well over 200 ft.-lb.). Front-drive wheel bearings (and the front bearings on many rear-drive cars) are well-sealed and often are life-of-the-car without lubrication. However, if you've been on a lot of secondary roads, or glanced off a curb hard enough to bend a rim, they could be worn or damaged.

If you have an older rear-drive car it probably has adjustable front wheel bearings, and finding a lot of

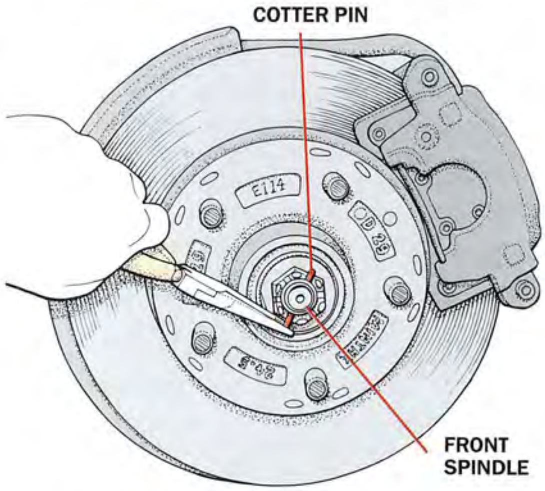

(Fig. 3) Remove the cotter pin to retorque a loose front wheel bearing on a rear-drive car.

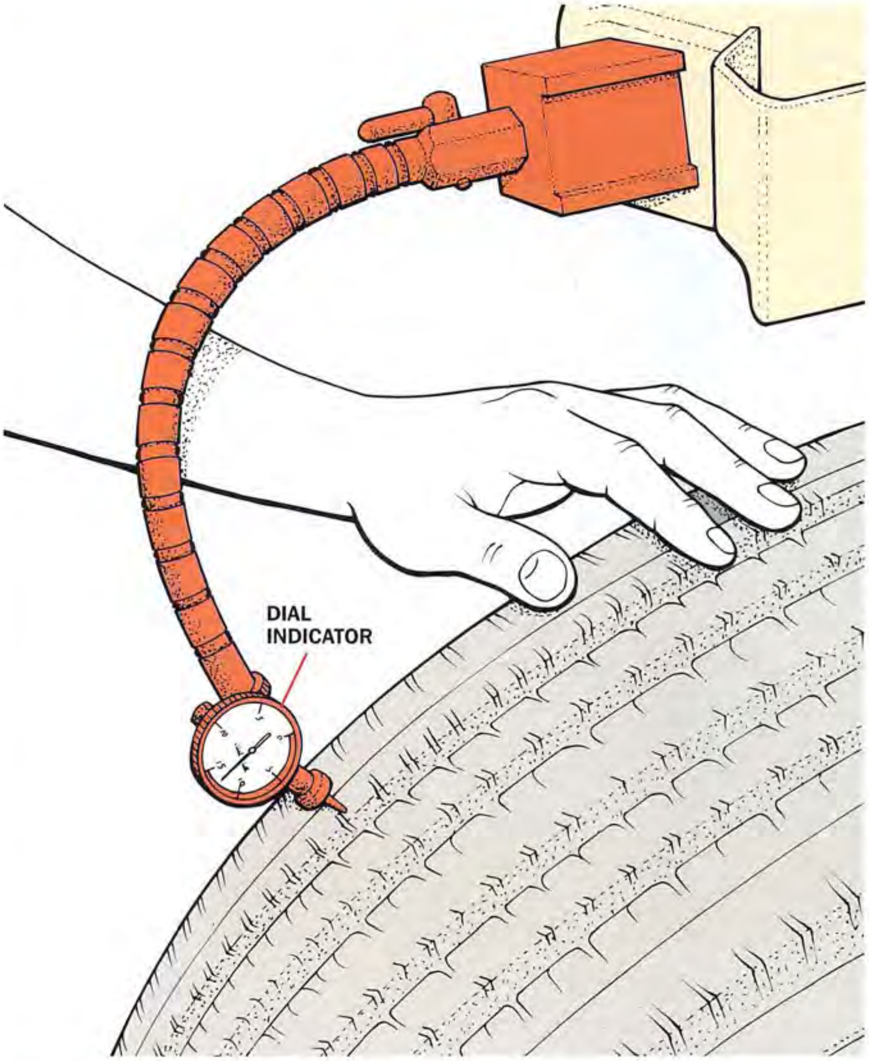

(Fig. 4) Total radial runout at the tire tread should be no more than .050 to .060 in.

free play in these is not surprising. To adjust, remove the cotter pin, tighten the wheel bearing nut to about 20 ft.-lb. to seat the bearings, and back off so they're just free but have so little play that you really can't feel it (Fig. 3). Then line up the slot in the spindle with the nut and insert a new cotter pin.

Steering Your Way

You may not feel free play in a front wheel (front- or rear-drive), but try rocking it in and out with a bit more effort, but not enough to move the steering linkage. That could demonstrate free play from wear in the tie-rod ends or ball joints. If you're not sure where the free play is, pry up on the bottom of the tire and watch the ball joint to see if it has free play—¼ in. is a lot. To check a tie-rod end joint for looseness, try to flex it by hand. A good tie-rod end should feel snug, but not immobile or stiff.

On rack-and-pinion steering, it's a good idea to check the tie rods' inner sockets. They're covered by the steering rack boots, but you can squeeze the boots to hold the inner joint. Jack

HOW IT WORKS

Rack-and-Pinion Steering

There are several types of steering systems, but the rack- and-pinion has become the most popular because of its simplicity and precise response. It's used primarily on passenger cars and also on late-model sport utility vehicles. The rack is a shaft with gear teeth, and it meshes with the pinion, a gear at the end of the steering wheel shaft. The rack is horizontally installed between the front wheels and is connected by a tie rod at each side to a steering knuckle, the pivoting structure to which each front wheel is attached. The tie rod has a flex joint at each end that allows it to flex and pivot in transferring steering wheel motion from the rack to the knuckles. As the steering wheel is turned to either side, the pinion rotates and moves the rack to that side, pivoting the front wheels in the same direction.

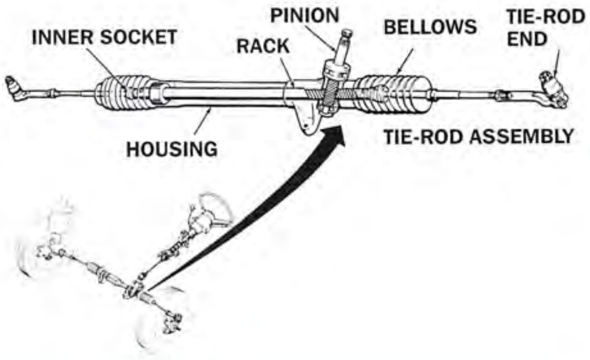

up the front end to take the weight off the front wheels. Have a friend slowly turn the steering wheel a partial turn to each side, while you feel for looseness.

Look Out for Runout

Just because you can't feel a lot of free play or "wobble" in a wheel doesn't mean there isn't enough to cause vibration. It doesn't take a lot to be responsible for objectionable vibration at speeds of 60 to 70 mph and above—any deviation from a truly circular spin is called runout. It can be vertical (up-down) or horizontal (in-out).

The only practical way to check for runout—front or rear—is with a dial indicator, another tool you can rent at many parts stores. There are several different checks to make to pinpoint the source of the runout.

Mount the indicator on something heavy that won't move, such as an anchor plate or wheel hub/knuckle. Position the plunger for the specific runout check. Example: For a radial runout test, rest it against a good tire tread groove. Slowly turn the tire and measure the amount of runout, ignoring jumps in the plunger that result from the shape of the tread or minor imperfections in it. If there are factory specifications for runout, use those.

If you don't have specs, see if the runout is about .050 to .060 in.—this measurement is considered rule of thumb (Fig. 4). The tire almost surely isn't the issue, although there is precision equipment that can check a tire for heavy spots. We know—you don't have it and can't rent it. Most professionals don't have it either, which tells you how common it is.

To isolate the source of the runout, check it at the wheel with the plunger on an underside horizontal surface. Ignore minor imperfections in the wheel finish (paint, weld, tiny dings) that cause the plunger to jump instantaneously. If the runout is over .045 in., the wheel should be replaced.

If radial runout isn't bad, check lateral runout with the plunger against the sidewall, even if the in-out rocking didn't show anything. Obviously, ignore any plunger movement from raised lettering, etc. If the runout is over .045 in., it's too much. Here again, isolate the runout by checking at the wheel with the plunger against a vertical surface. The rule-of-thumb specs are the same as for radial runout.

When the runout at the wheel is excessive, a new wheel normally is the answer, but not always. Remove the wheel and check runout on the wheel hub (Fig. 5). Making a lateral runout check is an obvious procedure because there's a hub face against which you can rest the plunger. For a radial check, it may be more difficult if the top surface of the hub isn't reasonably smooth because you have to use the threaded edges of the studs, and, typically, there are only four or five of those studs. So it does take some careful measuring to see if there's a significant amount. You have to look for the peak reading at each stud to be sure you're measuring at the outermost point. Unless almost all the radial runout is in the bolt circle, and that amount is at least .030 in., go for a new wheel. Replacing the hub and bearing on a front-drive is not a quick and easy job.

It can take a couple of hours to check out the possible causes of high-speed vibration, and you may be tempted to take the car in for wheel alignment to see if that helps before you spend time on all these other things. Sorry. Unless there's some evidence of wheel misalignment (such as irregular tire wear), a wheel alignment is not going to help at all. In fact, until you first isolate and correct the cause of the vibration, alignment would be a waste of time and money.

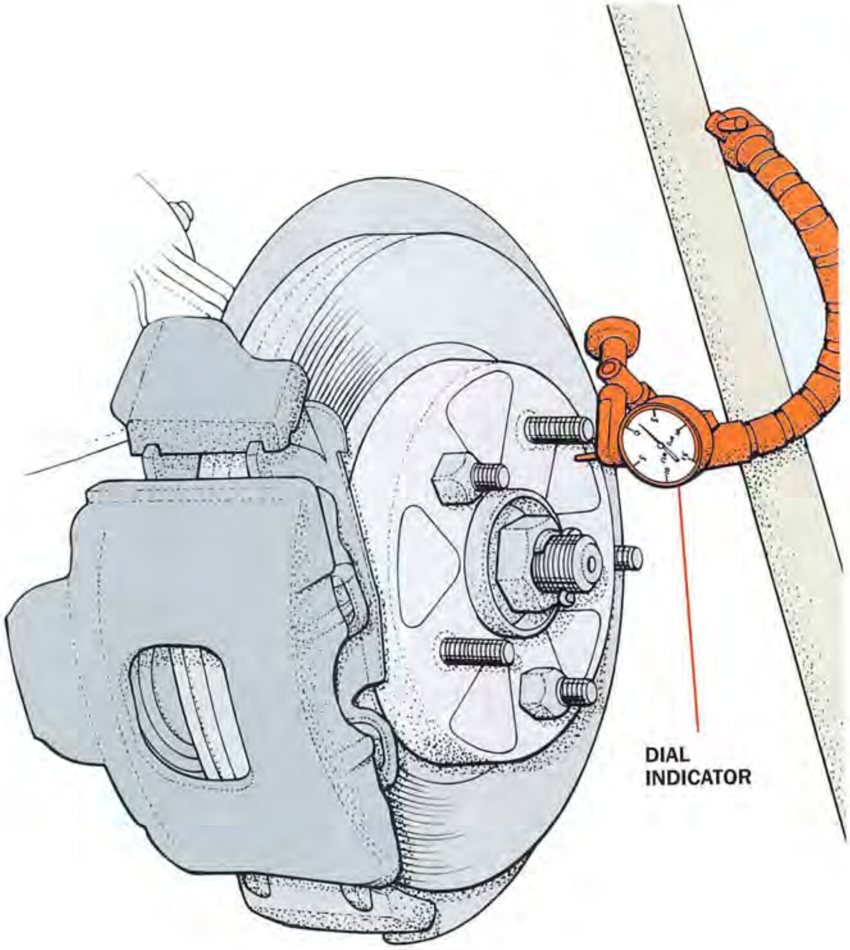

(Fig. 5) If radial or lateral runout is high, check both runouts at the hub to rule out a bent rim.

Replacing A Steering Rack

It's a crisp morning as you carefully back your car down the driveway. A quick turn of the wheel as the front tire clears the curb will swing your car up the street. But the steering wheel is strangely stiff. Instead of swinging the car's nose around, you run straight back, knocking over a trash can across the street. Fortunately, the damage is small: a crunched $14.95 rubbish barrel and a slimy mess to pick up in your second-best suit.

"What happened?" you wonder as you pop the hood. "The belt's still there and the fluid's up—although it does look black. Maybe the belt just slips until it warms up." After tidying up across the street and tidying up yourself, you try backing out again—with both hands on the wheel. The steering feels fine now—no groans and plenty of power steering boost.

If the belt is 4 years old or more, go ahead and change it because it's due anyway. But don't be shocked if the problem returns the next chilly morning. Power steering that awakens slowly on cool days has so-called "morning sickness." It won't go away.

It's caused by wear inside the steering assembly (a.k.a. the "rack"). The fluid's black, metallic look is a result of metal worn from the inside of the housing. The fluid is abrasive, and the wear will worsen with time. The rack is shot—replace it and use the opportunity to flush the junk from the pump and lines.

Can I Do It?

Swapping a rack is simple in some cars, nasty in others. For example, a rear-drive car's rack, if bolted in front of the front crossmember, is usually easy to change. But it's more challenging in a front-driver with the rack bolted to the subframe behind the powertrain or against the firewall.

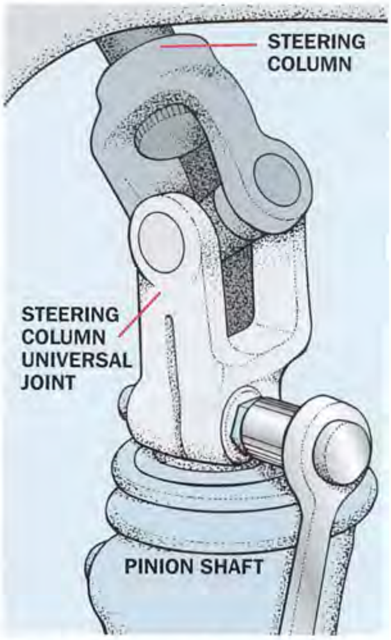

(Fig. 1) Disconnect the rack's pinion shaft from the steering column.

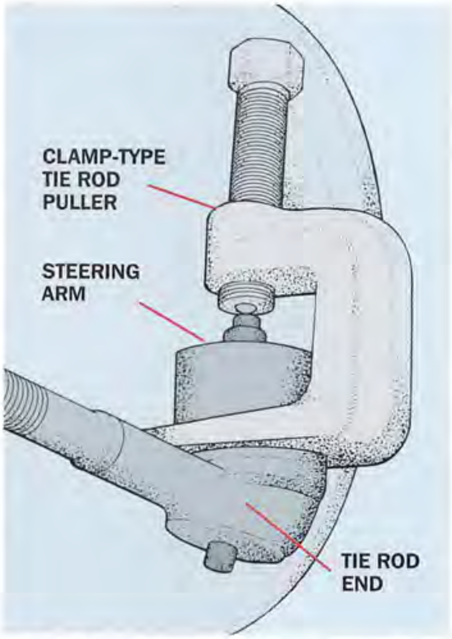

(Fig. 2) Use a special puller to remove the tie rod end from the steering upright.

Should you do it yourself? A shop manual may tell you how tough the job will be, but not always. One General Motors manual says you have to lower the front subframe 3 in. to get enough clearance to slip out the rack. Maybe, maybe not. Legions of technicians just twist and wiggle it through the existing space with no problem.

But you should still check the service manual. For one thing, it'll let you know if there are nuts and bolts that must be replaced with new ones, for safety's sake, when changing the rack. You will also want to know the torque values for all fasteners. Are there any O-rings involved? They'll need replacing. The manual will tell you.

Before chickening out, look at a replacement rack. You'll see the location of bolt holes and the fluid pressure and return line ports—the only fluid lines you'll need to disconnect and reattach at the rack. Safely support the car and check clearances. If you find you can't grasp everything you need to work on from under the car or by reaching around the engine, you may opt to send the job out. Lastly, before deciding, talk with people who know the job, such as the front-end pro who'll align the car after you've swapped the rack, or a dismantler at a local salvage yard. They may know legitimate shortcuts.

Doing It

Ready to go? It will help if you remove the front wheels for better access to the wheel wells. Break all the lug nuts free before using the jack. Lift the car and support it safely on stands. Then, unbolt the clamp that connects the steering column to the pinion shaft (Fig. 1).

Detach the outer tie rod ends. You may need to use a special tie rod end puller to get the tapered bolt to relinquish its grip on the steering upright (Fig. 2). If you're reusing the old outer tie rod ends, don't take them off the knuckles with a fork-type remover—it could damage the grease seals. Use a puller-type remover. Better yet, you probably can leave them in the uprights. Just loosen the jam nuts and turn the tie rods with a pipe wrench. The rods will unscrew from the ends—clockwise on one side, counterclockwise on the other. Unbolt the rack from the chassis. Now you can disconnect the fluid lines and capture the power steering hydraulic fluid as it drains.

Flushing

In spite of having digested considerable amounts of crud, many power steering pumps survive morning sickness intact. Still, you'll need to remove contamination to preserve the new rack. Here's how to flush most of the junk out of the pump and lines.

Disconnect the return line to the fluid tank and let it drain into a small container through the return hose. This is the low-pressure side, usually secured with a hose clamp. Then block the hole. Vinyl vacuum caps work well.

Next, disconnect the pressure line from the rack and aim it into a bucket. Refill with fresh fluid and tap the starter—it may require pumping a quart or more to clean the line. Cap or plug any loose ends.

Now it's time to twist, wiggle and snake the old rack out of the vehicle (Fig. 3). Expel any minors from the garage, as they may find the language necessary to persuade the rack clear of the vehicle unacceptable. You may need to unbolt and slightly move some other components to get the rack out of its tunnel. Lift, twist and wiggle in the new rack and reconnect the fluid lines. It may be easier to get a wrench to swing on a fluid line attachment once you've unbolted the rack and moved it a bit. Also, reattaching the lines may be easier before the new rack is bolted in place.

Use a tape measure to check the overall length of the rack and tie rod assembly. Set the overall length of the new assembly to this same dimension by twisting the tie rod ends on their threads. Keep the rack centered and

(Fig. 3) Snake the old rack out of its mounting tunnel after disconnecting the steering shaft, the tie rod ends and the fluid hoses.

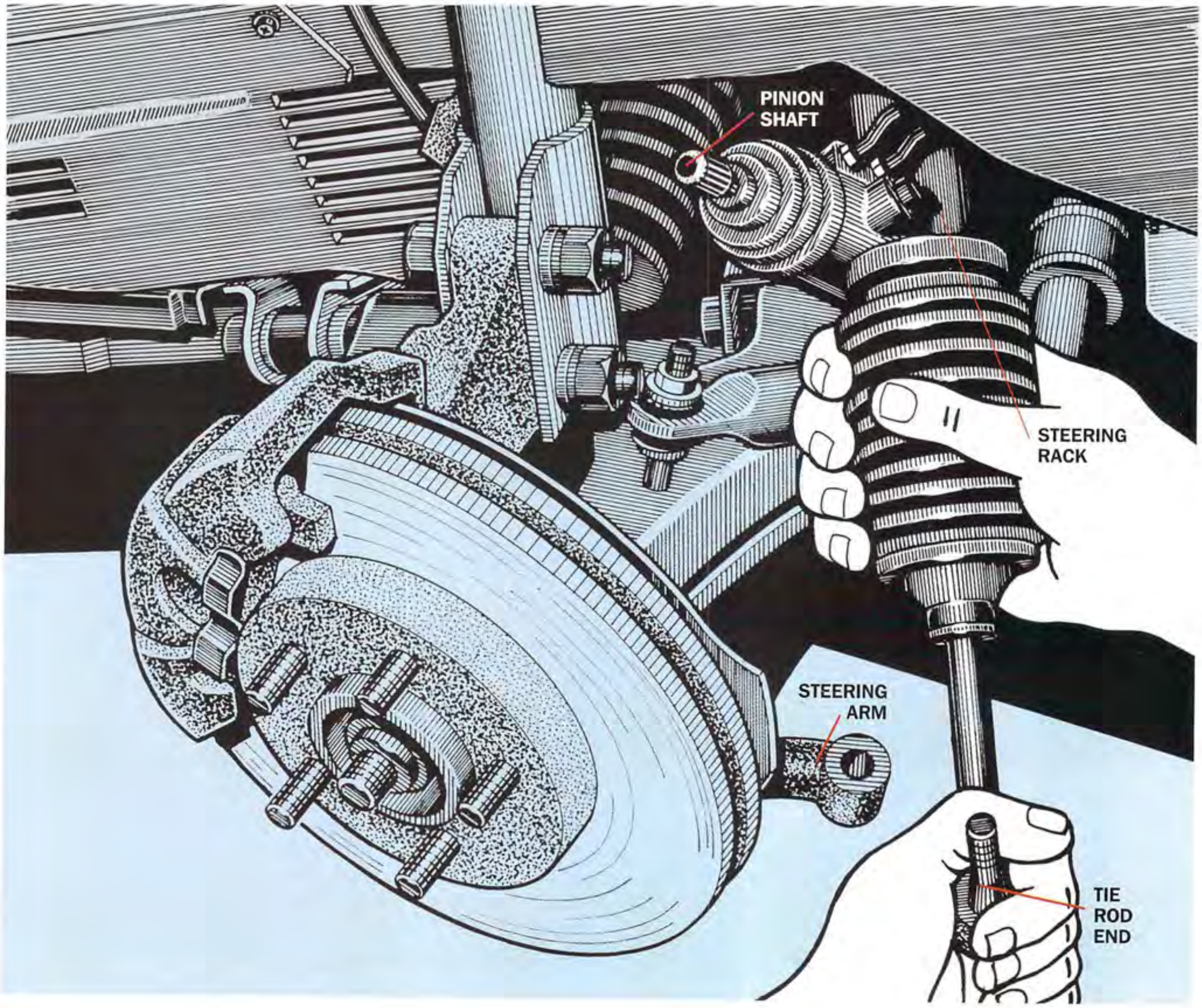

SEALING RINGS can wear grooves into the rack's housing, bypassing hydraulic fluid.

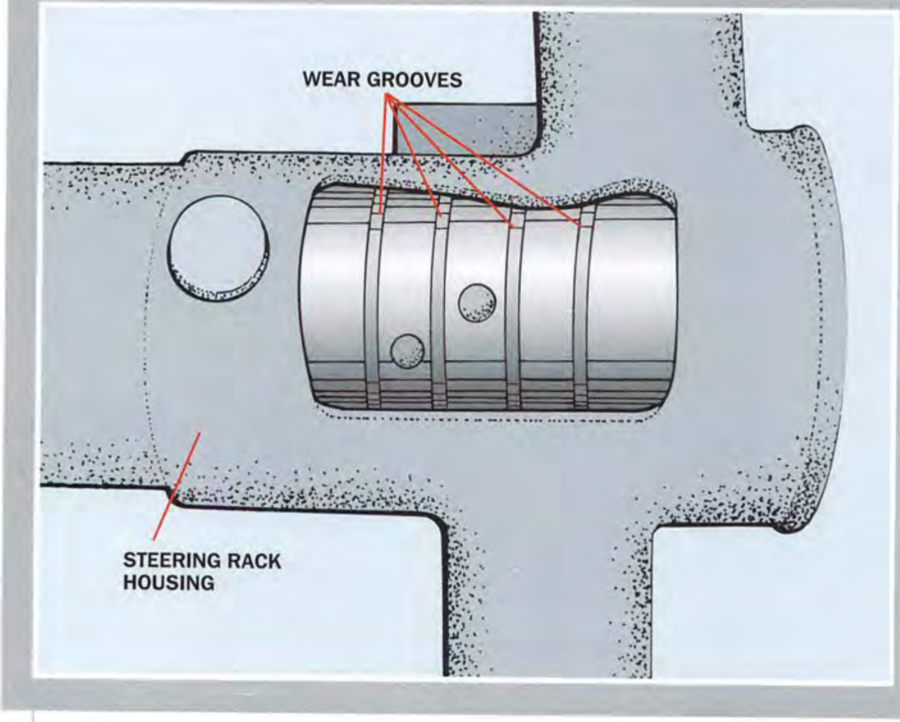

split the overlap difference between the left and right rod ends as you do this, or the steering wheel will be off-center when you're done. Connect all the lines, the steering shaft and the rack-attachment hardware. Use fresh cotter pins in the tie rod ends' castellated nuts. Reattach the front wheels.

Once the new rack's in, reconnect all hoses except the reservoir return line. Point it into a bucket, then refill. Start the car, and run fluid through until it looks clean. Reattach the pressure line. Note: You may be able to install an aftermarket inline filter so any crud you miss won't harm the new rack.

Have an alignment shop reset the toe-in adjustment or the car may handle strangely and wear out its tires rapidly.

The Fluid Situation

The last step is bleeding out all of the trapped air. While the car's still elevated on stands, fill the reservoir, idle the engine and steer side to side, from lock to lock, 10 to 12 times. Be careful to avoid banging into the stops. Tan or foamy fluid contains air. Shut down, let it sit 15 minutes, then top off and start again. Repeat until the fluid looks normal.

Don't forget to properly torque the lug nuts on the front wheels after you've taken the vehicle off the jackstands. And it wouldn't be a bad idea to change the power steering fluid in a few hundred miles to get out any crud you've missed.

HOW IT WORKS

Power Rack-and-Pinion Steering

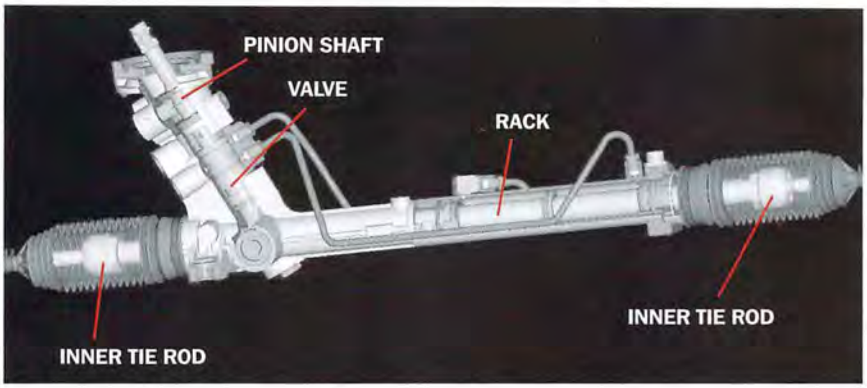

As you turn the steering wheel, a spool valve on the pinion shaft moves relative to ports in the housing. The ports connect to chambers on either side of a piston on the rack shaft. Steer left and the port connected to the chamber on the right side of the piston opens while the other port closes. This lets fluid pressurize the right side of the piston, helping push the rack shaft to the left and pointing the wheels in that direction. (This description assumes the rack assembly is ahead of the axle. If it were behind it, moving the shaft right would point the wheels left.)

Wear allows fluid to bypass the piston, wasting steering assist. "Morning sickness" results from light wear that closes up as heat expands the parts. As wear worsens, expansion will no longer take up the slack.

Intense competition has driven remanufactured rack prices down. A "reman" unit may be your best bet, since many originals wore out because their shafts bore directly on the inside of their aluminum housings. In remanufacturing, the housings are bored, then hardened-steel inserts are installed in areas subject to wear. You may be able to get a recently installed new or reman rack at a salvage yard.

Servicing Brake Discs

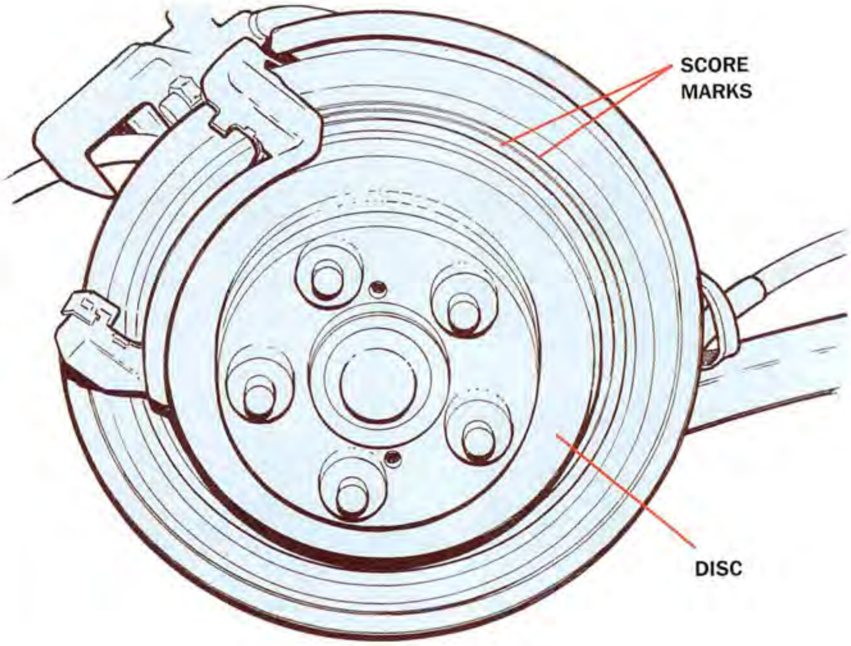

(Fig. 1) Surface grooves deep enough to catch your fingernail will need to be removed by machining the brake on a lathe.

You're up to highway speed and suddenly a car cuts across your lane to make an exit. You manage to slow enough to avoid a nasty accident, but the brakes definitely don't feel right. The brake pedal has had this kind of vibration for a couple of weeks, but only at low speeds. Now it feels like you're stepping on a running chain saw, and the front wheels don't want to go where you want them to.

Forget The Guesswork

Pull the wheels for a closeup look. With the wheels off, you can inspect the rotor-lining contact surfaces. First look for a crack that goes all the way through a vented rotor—or at least through one of its two surfaces. A cracked rotor causes severe pulsation, and the only cure is replacement.

Rust at the rotor edges is meaningless, as are moderate scores in the disc's surface, or a few off-color spots on the shiny areas swept by the pads. Be sure to check both sides of the rotor, however. When you pull the wheels, there should be an accessible area on the back side (even if there's a dust shield) where you can inspect the surface. Severe rusting of the lining contact surfaces is a problem. So are thick, deep scores, uneven thickness and lateral runout (warping).

If the rotors have never been resurfaced, it is theoretically possible to have them machined to remove significant rust and deep scores (about .060 in. deep). However, many late-model rotors are not very thick to start with, and can't take more than a .030-in. resurfacing. Unless the rotor is at least .060 in. thicker than the minimum thickness stamped somewhere on the rotor, forget trying to save it. Buy new rotors. How can you test the depth of a score? If it's thick enough, insert the head of a dime into it. If the dime goes in beyond the top of the president's head, the score is too deep. If a score is too thin for the dime to slip in, it's harmless. Another test is to run your fingernail across the surface radially. If your nail catches, the rotor needs to be resurfaced (Fig. 1). If not, you're probably okay, provided there's enough thickness left on a true-running rotor.

Uniform thickness is critical. Measure each rotor with an outside micrometer at a number of evenly spaced points around the lining contact surface. A variation of .005 in. or more is enough to cause the pulsation you feel.

Lateral runout is caused by a rotor that's uniform in thickness, but is wavy. Uneven thickness and excessive runout often go together. That is, the runout causes the uneven wear on the rotor. So if you catch it early, the rotor may pass a thickness check, but it still could have excessive lateral runout. The high heat (then rapid cooling) from hard braking is a primary cause of runout, but close behind is uneven tightening of wheel lug nuts. Dirt or rust buildup between the rotor and hub also can produce runout (Fig. 2).

To check for excessive runout/warping, you need a dial indicator, which you can rent from many auto parts stores. When you pull the wheels, the rotors might be loose on the studs, so refit the lug nuts and tighten them securely. Install the dial indicator, and mount it on the suspension. Then, aim the indicator plunger at the middle of the lining contact surface, at as close to a 90° angle as possible. Set the plunger so it presses lightly against the rotor-lining contact surface, then zero out the dial. Slowly rotate the rotor and watch the dial. If it exceeds .005 in., the rotor is not running true (Fig. 3).

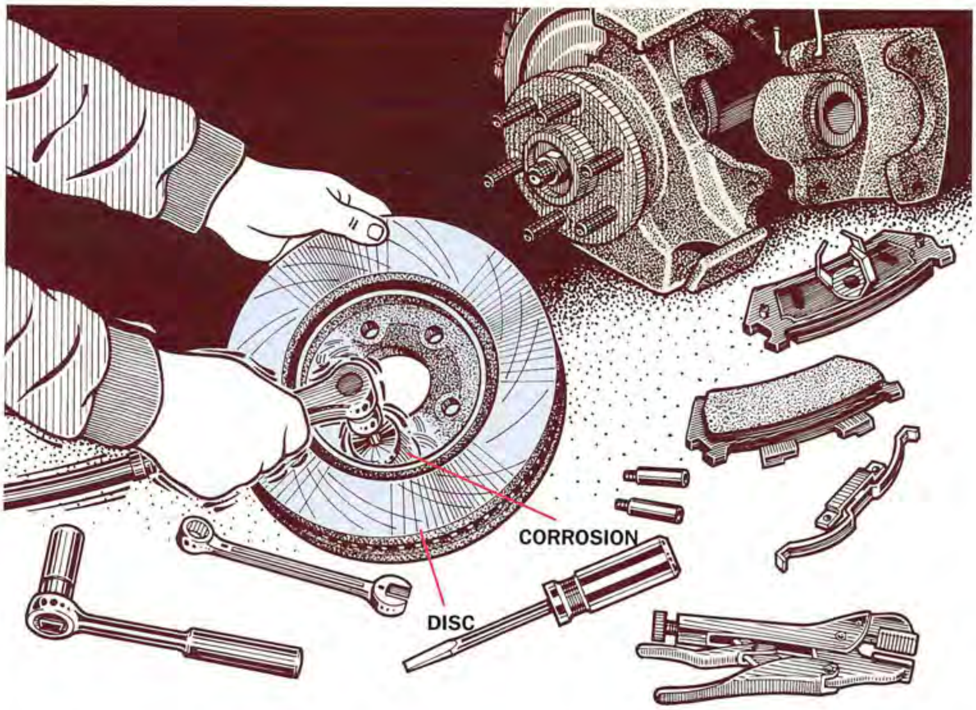

(Fig. 2) Whenever assembling brakes, remove any dirt or rust that might get between the brake and the hub to reduce lateral runout.

You may not have to replace or resurface the rotor. With the rotor off, you can look for and remove dirt or rust buildup. If that's not the problem, you can try re-indexing the rotor. That is, just remove it, rotate it one or two stud holes, install and tighten the lug nuts and recheck. Although the typical rotor just slides off the studs with the wheel removed (some advance spraying with penetrating solvent may be necessary), the caliper has to come off first.

First, remove the bolts that hold the caliper itself to its anchor frame or to the suspension. Then pry the caliper up, lift it off the frame and wire it to the suspension. This will relieve any strain on the brake hose. Many vehicles have "flip up" calipers. Remove one lower retaining bolt, flip up the caliper and slide it off a stud. If the brake shoes stay in the anchor frame, perhaps held by a retaining clip, remove those next.

Finally, unbolt the anchor frame, lift it out and the rotor should slide off. But first, if the rotor is not necessarily being replaced, make alignment marks on the hub section and the stud, so you can reinstall it in the same location. If it binds on the hub, that's normal rust buildup. Just spray the studs and holes and hub/rotor joints with penetrating solvent, rock the rotor with a pulling motion and it should walk off. Clean

HOW IT WORKS

Brake Rotor

The brake rotor, the disc in disc brakes, mounts on the hub with the wheel, and both rotor and hub are held by the same studs and lug nuts. When you apply the brakes, the hydraulic device called the caliper clamps the brake shoes (also called pads) against the rotor, and the friction material on the shoes stops the spinning of the rotor (and therefore the wheel). The friction material on the shoes causes the energy of wheel and rotor motion to be converted to heat, and the rotor temperature can soar to over 1000°F in a hard stop. That heat has to be dissipated very quickly so that the rotor can cool down for the next stop. Because front brakes do about 75 percent of the work in a front-drive car (over 50 percent in a rear-drive), the front rotors have to be able to dissipate a lot more heat than the rear brakes, drum or disc. As a result, the front rotors on virtually all cars today have two contact layers separated by a vent area, which resembles a paddle fan, to scoop up air. This design not only gives each brake shoe a separate disc, but the airflow through the center provides greater cooling. Some high-performance cars also have vented rear discs.

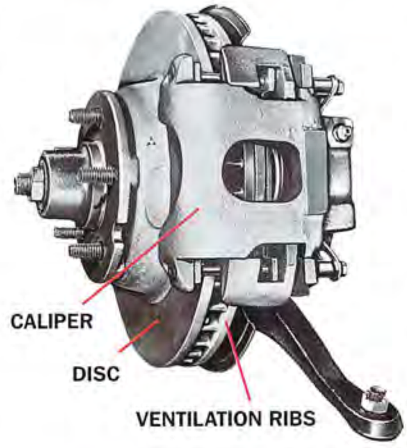

A typical vented disc brake assembly. The hole in the caliper is for inspecting pads.

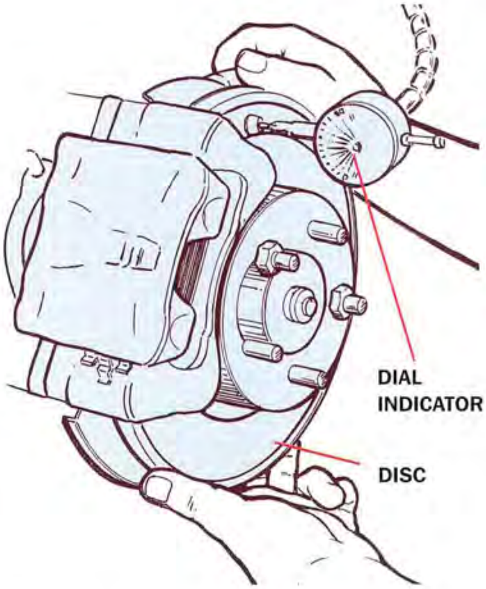

(Fig. 3) A dial indicator will show you the total lateral runout—.005 in. is too much.

rust off the hub with ultrafine sandpaper or a wire brush. Clean the mating face of the rotor as well.

If you're working on rear discs, they're likely to have a mechanical actuator for a parking-brake function or a drum-type parking brake built into the hub. With the drum type, the rotor itself will have a deep "hat" section to fit the hub, and the inside of that hat is the drum for the parking brake. Normally, with the parking brake off, the rotor just slides off, but if the parking brake has a self-adjusting linkage, it's possible that it's overadjusted and the drum shoe linings are locked against the inside of the hat. In that case, you can reach in from the back through an access hole (in the hub and also possibly through a dust shield), and back off the adjuster (Fig. 4). With the type that uses an actuator on the disc brake lining pads, you may be able to just lift the caliper off. At worst, you'll first have to disconnect the brake cable from the actuator on the caliper.

Installation

If you need a new rotor, install it straight out of the box. Some of the lightweight rotors, made of a stamped steel hub with a cast-iron braking surface, could test as warped unless you check them with all lug nuts installed and tightened.

Note: There are some rotors that are a real bear to remove, so if the vehicle manufacturer indicates special tools or if a press is necessary, a rotor can be machined on the vehicle. You're best off leaving this to a properly equipped pro.

Both Sides, Please

Normally, what happens to one wheel happens to the other. If one rotor is gone and the other is marginal for machining, replace both. However, if one is particularly bad, perhaps the vehicle has been pulling, leaving the other side pristine. There may be uneven brake application, perhaps from a hydraulic or mechanical problem at the wheel. Trace the cause in addition to correcting the rotor problem.

Clean lug nuts and stud threads with a mild solvent and wipe dry. Don't lubricate them if the torque specifications are based on clean, dry threads. But a thin film of anti-seize compound normally is beneficial. One of the most important tools you can use is a torque wrench. When you're re-indexing a rotor, be sure to reinstall the lug nuts and torque them to specifications in at least three stages. Use a crisscross pattern for each stage until all are tight.

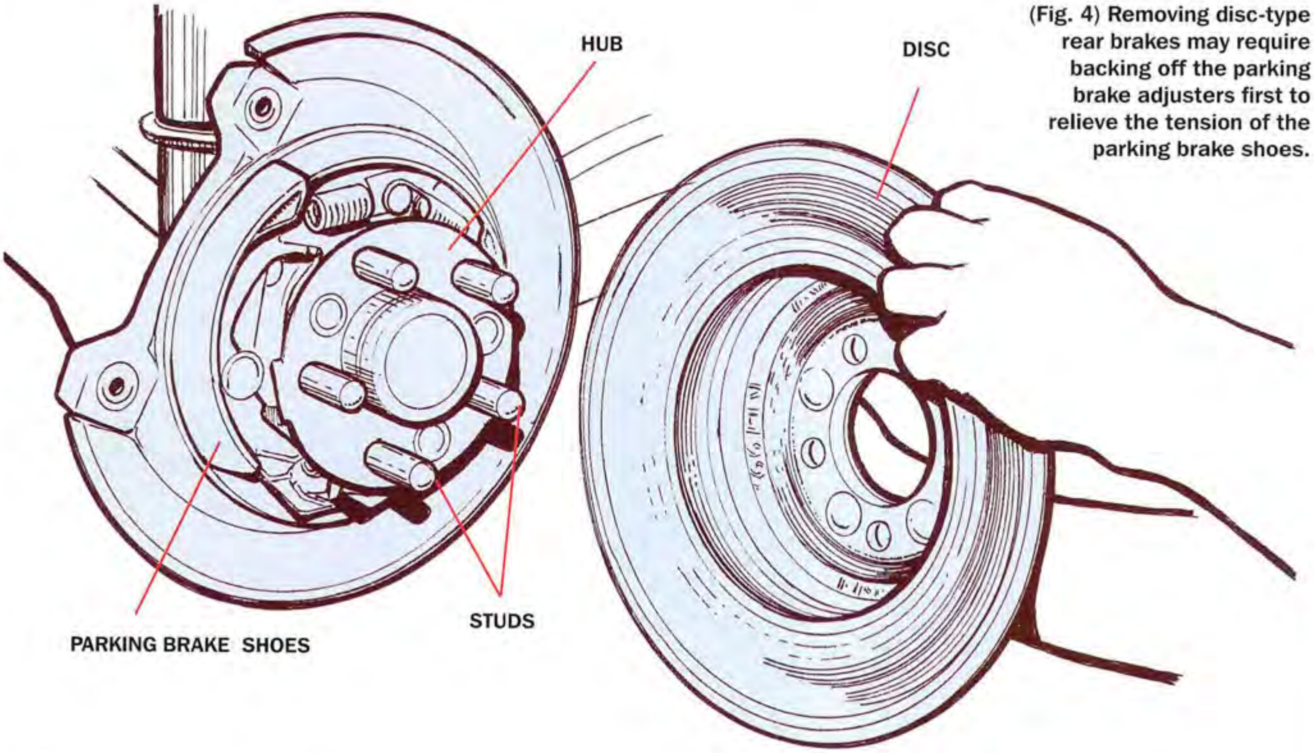

(Fig. 4) Removing disc-type rear brakes may require backing off the parking brake adjusters first to relieve the tension of the parking brake shoes.

DIFFICULT

Fixing Suspension Clunks And Rattles

It's the proverbial 20 miles of bad road. Potholes compete with ruts for the privilege of knocking your Freedom Fries out of the bag and onto the floor. It's worth it, however, for the great weekend of backwoods hiking and knocking around. Heading back to civilization without the heavy load of refreshment and food seems easier on your car and your freshly relaxed psyche. At least until you hit the pavement and the rattling starts. Maybe you couldn't hear it on the unpaved road, but every expansion strip on the interstate makes your car sound like a tin can full of bolts. Something's loose in your suspension.

While there's little chance that your car is going to lose something essential while you're going down the road, chassis and suspension noises definitely have to be checked out for safety's sake. Plus, who wants to drive a vehicle that sounds like it's about to drop to the pavement?

If your car has lots of miles on it—or even if it doesn't, but is "chronologically gifted"—don't be surprised if some portion of the heavy metal that supports it over terra firma starts complaining. Unfortunately, finding the cause of the noise isn't so easy. The dynamics of a rolling vehicle, the complex nature of modern suspensions, and the way sounds can be telegraphed through the chassis and body make it hard to pinpoint the location of a problem.

If you hear a clunk when the suspension works over bumps, the probable cause is excessive clearance in a joint due to wear. It might be as simple as a loose strut gland nut, or something more subtle such as a shrunken, dried-out rubber bushing.

Research, Then Bounce

As a first step, consult whatever literature you have available to see if you can find a Technical Service Bulletin (TSB) that pertains to the noise. Apparently, suspension clatter is a popular problem industrywide because it generates quite a few TSBs. Some of these alert you to redesigned replacement parts, while others say the noise is simply a characteristic of the vehicle and should be accepted as normal.

If no clues are forthcoming, it's

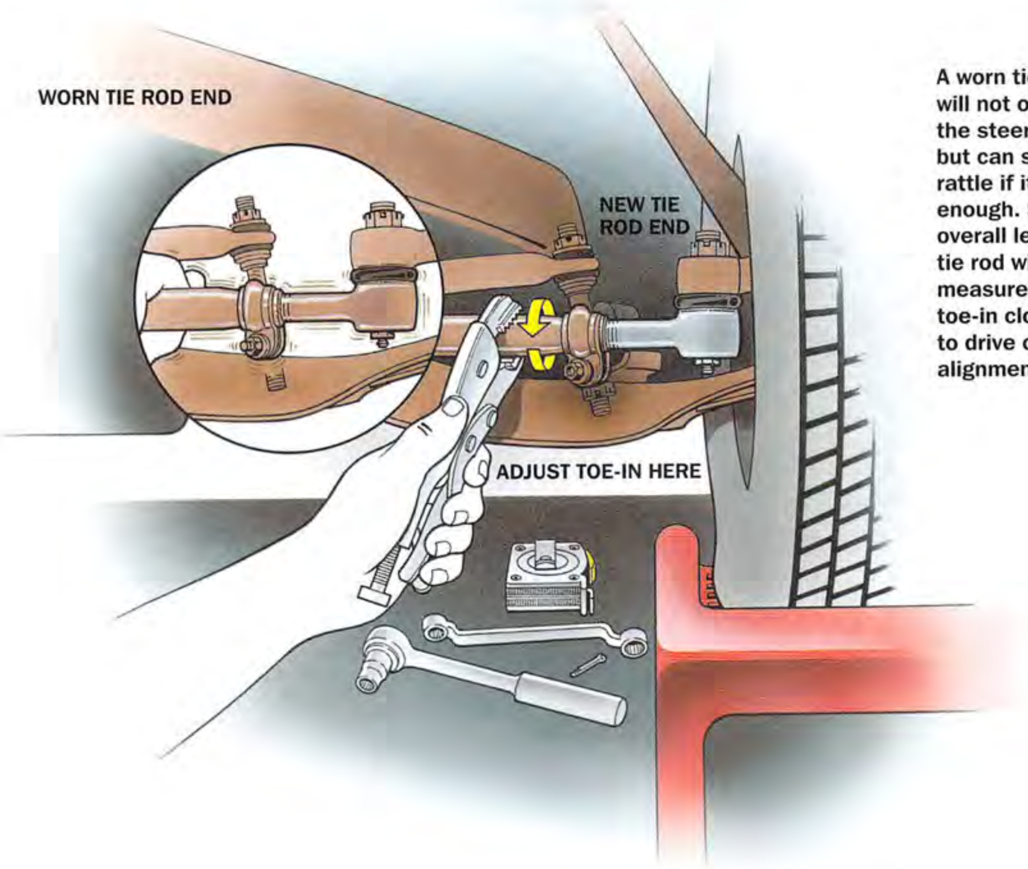

A worn tie rod end will not only make the steering vague, but can set up a rattle if it's bad enough. Set the overall length of the tie rod with a tape measure to get the toe-in close enough to drive over to the alignment shop.

Antiroll bar bushings and brackets can loosen up as they wear. Replace the bushings or tighten the fastening hardware.

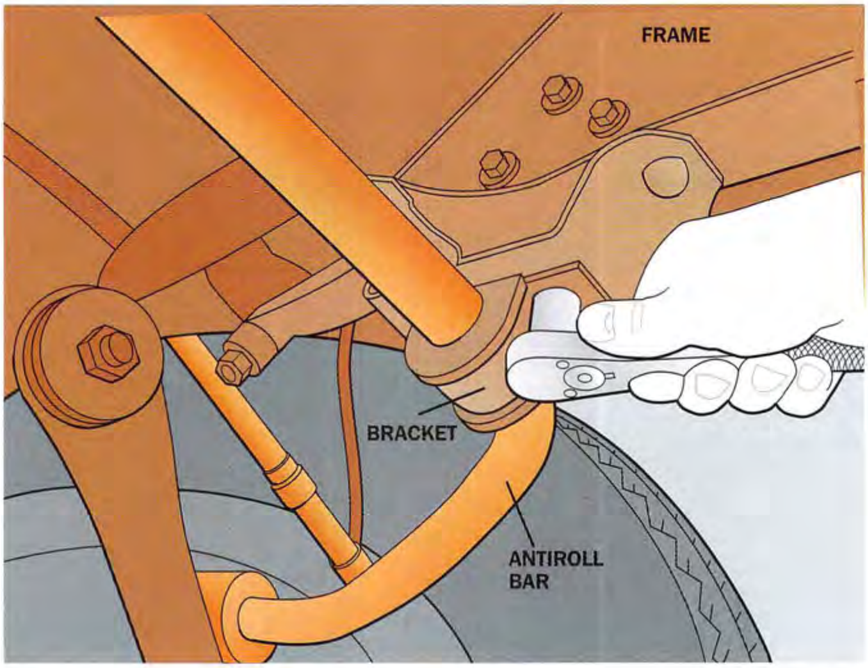

time to go hands-on. It'll be helpful if you can get a friend, preferably a hefty one, to assist. For front-end noises, pop the hood, and have your comrade press down on the bumper or fender, then release and lift repeatedly until the suspension is really working. While he does this, listen carefully and use a good light to examine the upper strut or shock mounts and the control arm joints. If you hear anything untoward, but can't pinpoint the source, place the end of a broomstick against your ear and touch the other end to suspected areas. This works almost as well as a mechanic's stethoscope. Nothing obvious? Then lie down and look underneath with your light, even though your friend's stamina may be taxed by this time.

The "dry park check," which will uncover lash in the steering mechanism, is less physically challenging. Have your helper sit in the driver's seat, turn the key to unlock the column, then rock the steering wheel vigorously from side to side while you watch the steering components. There should be next to no visible lash.

By the way, if you raise the car by the frame, the suspension and steering parts will be hanging at an unnatural angle, which may mask the looseness you're looking for. So, place your jack and jackstands under the control arms or the rear axle to keep the weight on the suspension components.

You can uncover the shortcomings of upper A-frame or control arm bushings by having a helper hold the brakes firmly with the engine idling while shifting from Drive through Neutral to Reverse repeatedly. Look down over the fender as your helper does this.

Some vehicles have substantial horizontal struts that position the lower control arms fore and aft. These are mounted in large rubber bushings, and any clearance here will make itself heard. Also, their mounting points on the frame have been known to rust away, but this causes steering symptoms far more noticeable and worrisome than a mere noise.

Older rear-wheel-drive vehicles with a live rear axle and coil springs may have what's called a "panhard rod" that runs diagonally from the chassis to one side of the axle housing. The rod's bushings are a likely source of a clunking noise.

Worn-out shocks or struts are common culprits here. When the internal hydraulics wear out, the piston will move without the proper resistance, then stop short when you hit a bump. With shocks, another thing to check for is loose or dried-out mounting bushings.

A groaning noise when you turn the wheel typically means there's a dry joint, likely at the idler or pitman arm.

That husky stabilizer bar, the part that helps keep a car level in a curve, is often the source of noise. The links that attach it to the chassis have bushings at both ends, and there is more vulnerable

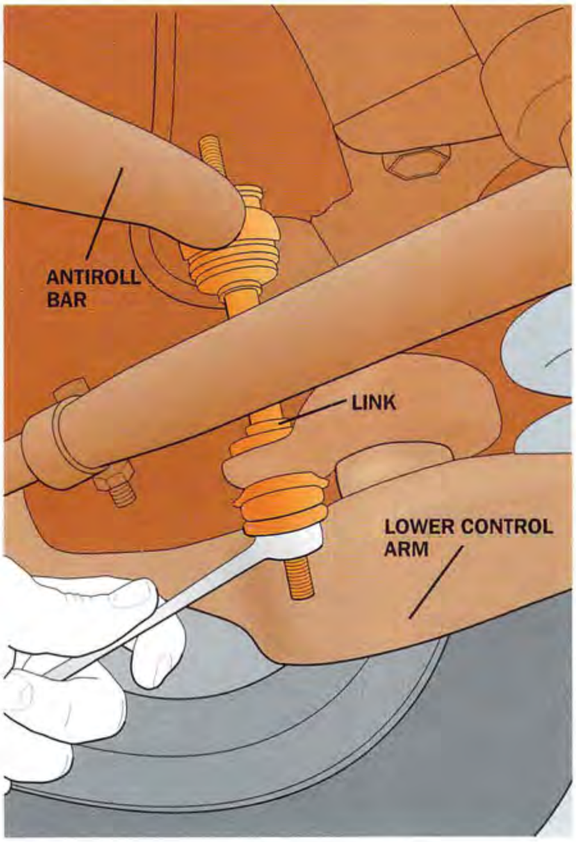

The antiroll bar's attaching links can become loose. Check the bushings.

rubber in its mounts.

A worst-case scenario we often see on unibody cars driven extensively in winter conditions: rusted-out suspension pickup points. The only fix for this is a trip to the frame shop where new metal can be welded on. If this is your vehicle's problem, you have real trouble—because if one point is rusted, its sister on the opposite side is probably rusted also. And probably a bunch of other places as well. Now you'll have to decide how much money you want to spend on a ship that's sinking slowly but surely. It may be time to retire this vehicle.

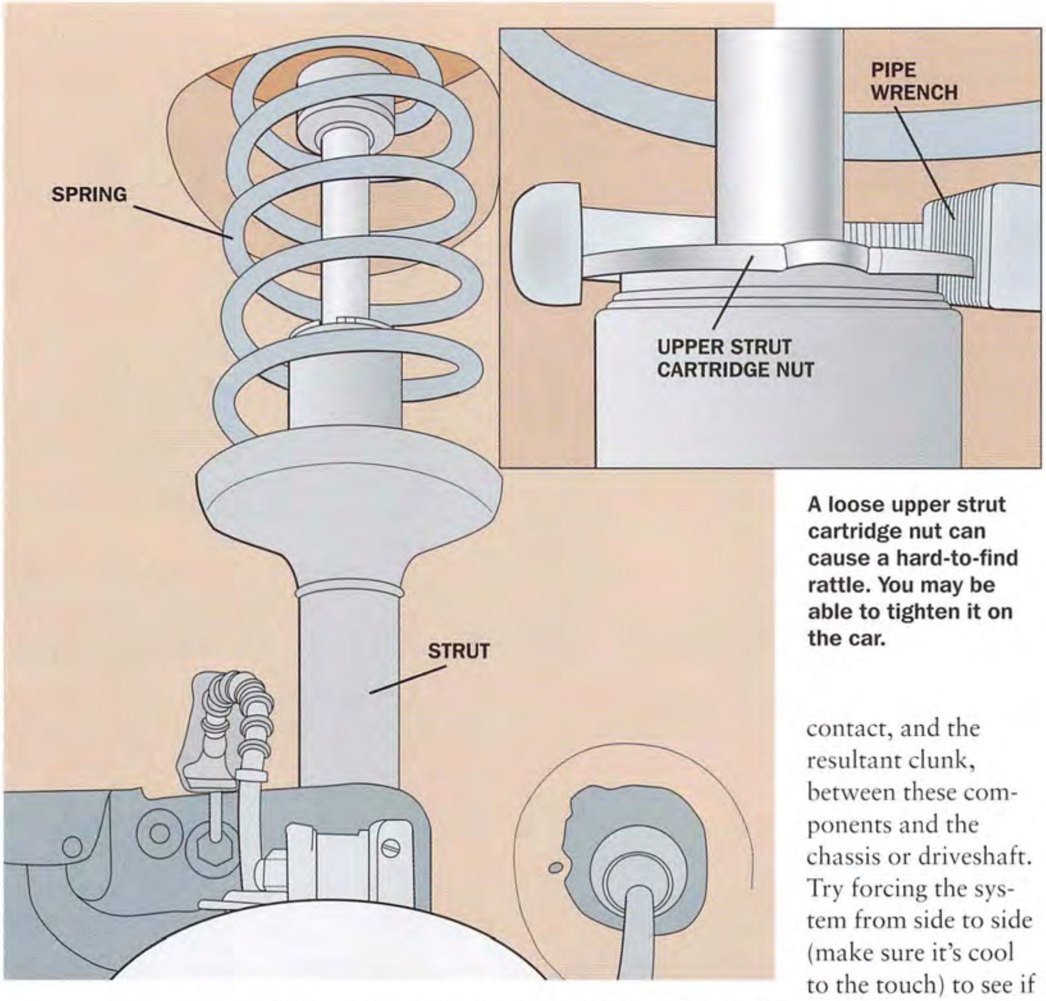

A loose upper strut cartridge nut can cause a hard-to-find rattle. You may be able to tighten it on the car.

Alternate Sources

Of course, some noises that emanate from under a car do not have anything to do with the suspension or steering systems, but it can be difficult to make the distinction.

For instance, take the exhaust system—the muffler, headpipe and tailpipe, and the catalytic converter. If everything isn't positioned just right, or the hangers are loose or broken, it's likely that there'll be occasional contact, and the resultant clunk, between these components and the chassis or driveshaft. Try forcing the system from side to side (make sure it's cool to the touch) to see if you can duplicate the sound.

A broken motor mount can cause a solid thump. Oil soaking may have caused it to delaminate, or perhaps a couple of bolts are loose. This condition will be sensitive to getting on or off the throttle, but won't be detected over bumps.

Rusted-out suspension pickup points are an accident waiting to happen. This calls for welding—or the scrapheap.

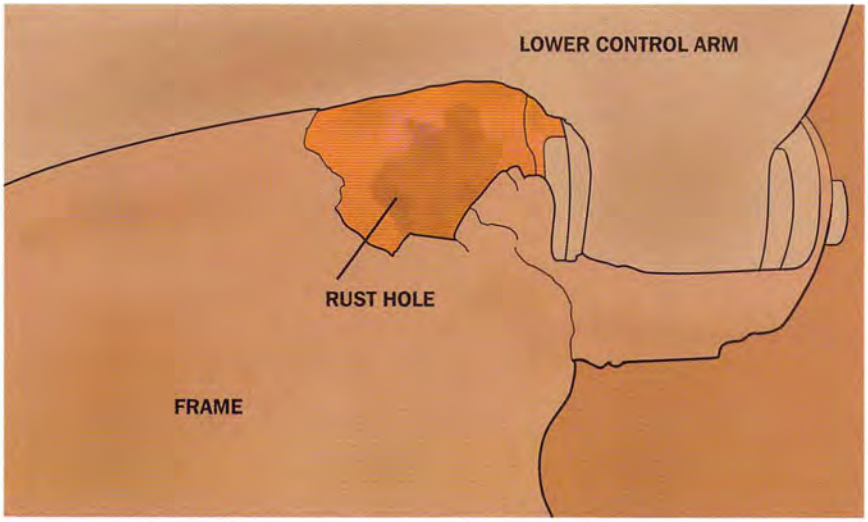

Restoring The Peace

While we've seen people try to take up clearance in a deteriorated bushing with shims made of sheetmetal, screws, etc., the only real fix is replacement. This can be more involved and expensive than you might expect, but it's the price of peace.

On the other hand, some repairs are free, or nearly so. It costs nothing to tighten a strut gland nut (put a drop of anaerobic thread locker on it to avoid a recurrence), and very little to replace shock mounting or stabilizer bar bushings.

DIFFICULT

Curing Steering Wander

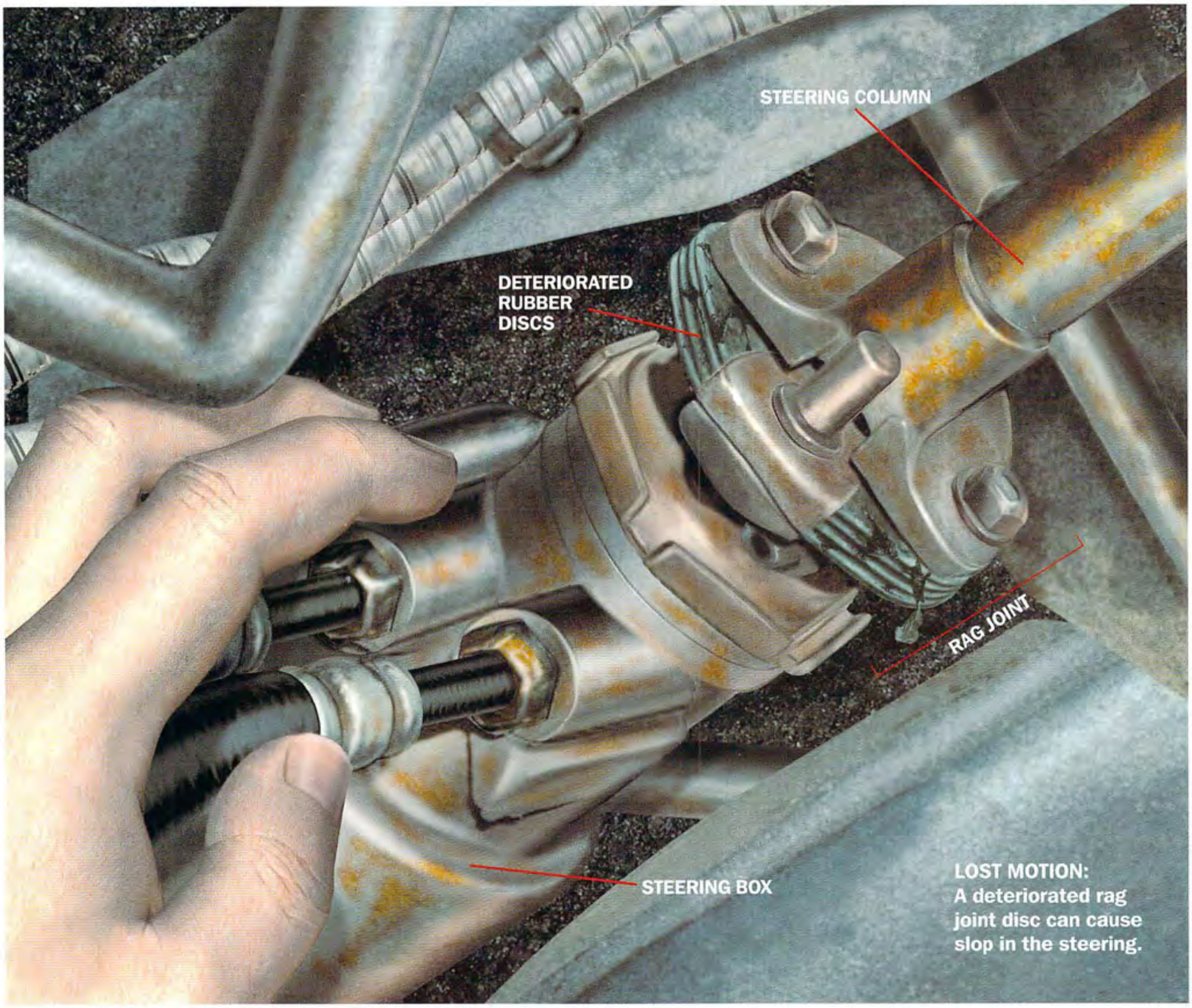

LOST MOTION: A deteriorated rag joint disc can cause slop in the steering.

You're almost enjoying a long drive on a nice, flat interstate. Almost, because it's windy and you need to saw the steering wheel back and forth, tacking like a sailboat, to stay in your lane. That is, until you make a short rest stop and realize that there's no wind whatsoever. The kids in the back seat are screaming for Dramamine, and you're fatigued after only an hour of driving.

There's no doubt about it—you definitely have some steering issues.

Vehicle-to-pavement interface

Before you start thinking about expensive repairs, make sure the basic vehicle-to-road interface is okay, meaning the tires. First, check inflation pressures. (Don't trust the gauge on the quarter-eating pump at the local station—those are often off by as much as 5 pounds. Drop 10 bucks on a good gauge and keep it in your glovebox.) You should check your tire pressures once a month. And that's first thing in the morning, cold. The correct pressure is on a sticker—check your owner's manual for its exact location. A low tire on one side will make a car pull in that direction. This is because its rolling diameter will be smaller than that of its mate on the other side. Also, there'll be more tread-to-pavement drag on the low side, pulling the car in that direction.

If the problem persists, try switching the positions of the right and left tire/wheel assemblies. If the car pulls in the opposite direction after you've done this, you've found tire trouble.

PSSST!: Low tire pressure will cause a wheel to pull.

Check your tires' tread-wear patterns. For example, if a front tire's tread tends to disappear along the outboard edge, it's likely that the camber setting at that corner is too positive, and any pull probably will be toward that side.

Geometrical considerations

Camber refers to the tilt of the tire from the vertical, and this has a profound effect on directional tendencies. Zero camber means the centerline of the tire is perfectly perpendicular to a level surface. If the top of the tire tilts outward from the body, camber is said to be positive. Going too far in this direction will cause a pull to that side because the tire itself forms the shape of a cone.

Don't run to the alignment shop just yet. You can get a pretty good idea of camber by using a carpenter's level, although you'll need to be parked on a perfectly level space. We used two 35mm film canisters held to the edge of the level with rubber bands. The canisters served as feet that we placed at the top and bottom of the wheel's rim. Keep in mind that most vehicles will have a degree or two of negative camber.

Here's where ride height comes in. As springs or torsion bars sag with age, camber changes. Replacing coil springs or adjusting torsion bars can bring alignment back into specs.

Woof!

Even if the front wheels are perfectly aligned and tires properly inflated, you may still have to steer constantly in one direction or the other to keep the vehicle going straight up the road. The problem is that the rear wheels are also trying to steer the vehicle and overtake the fronts. This condition is commonly referred to as dog tracking.

Technically this occurs when your vehicle's "thrust line" and centerline are too far apart. On vehicles with solid rear axles, the thrust line is perpendicular to the rear axle. On vehicles with an independent rear suspension (IRS), the thrust line is determined by splitting the toe-in angle of the rear wheels. For example, if the left rear wheel is toed in at 4° and the right is toed at zero, the thrust line is 2° to the left of the centerline.

In an ideal world the two lines coincide. But given a vehicle's size, manufacturing tolerances, wear and abuse of daily driving, they often do not. If the deviation between the two is great, your vehicle will dog track. Besides having a steering wheel that's not at center when you're going straight, another obvious clue to dog tracking is if you see four distinct tracks in the snow or rain when driving straight.

This is just about impossible to correct without professional 4-wheel-

This side up: A quick camber check can be done with a level and two simple spacers.

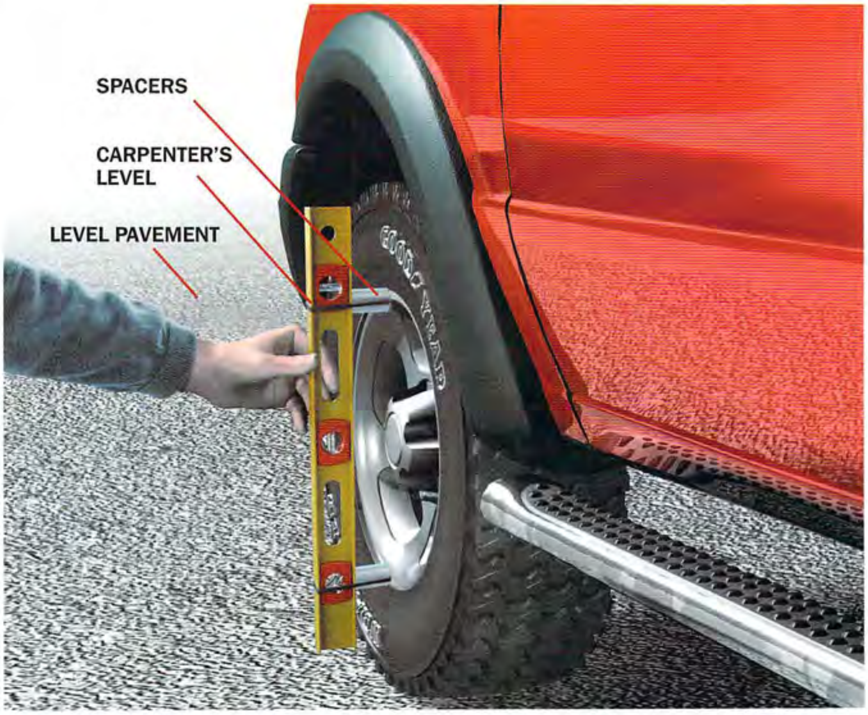

alignment equipment. On IRS cars, tapered shims are typically installed under the rear stub axles to reconcile the thrust line and centerline and to restore harmony. With solid rear axles, the repair will require replacing the rear links or straightening the frame.

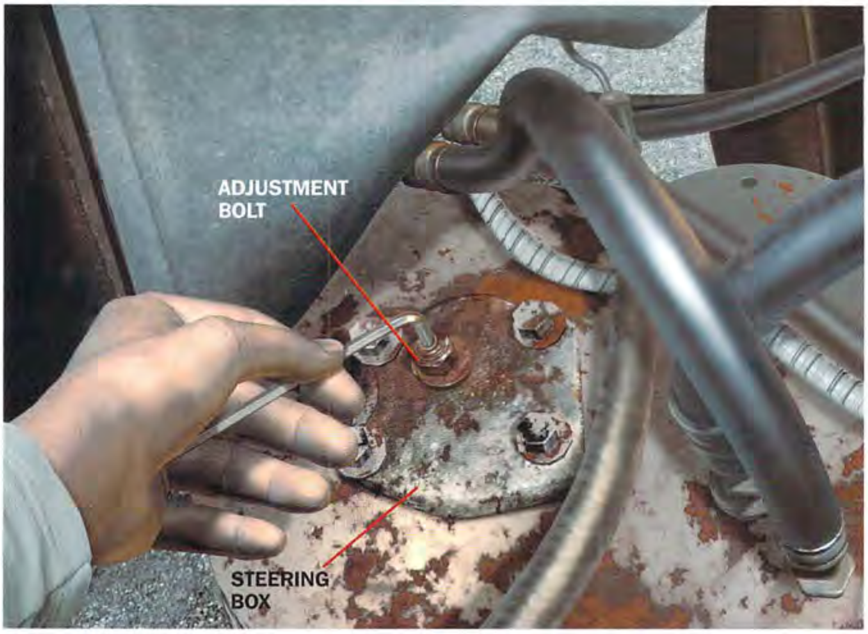

LOOSE: Minor adjustment to the clearance of a high-mileage steering box can tighten up the steering.

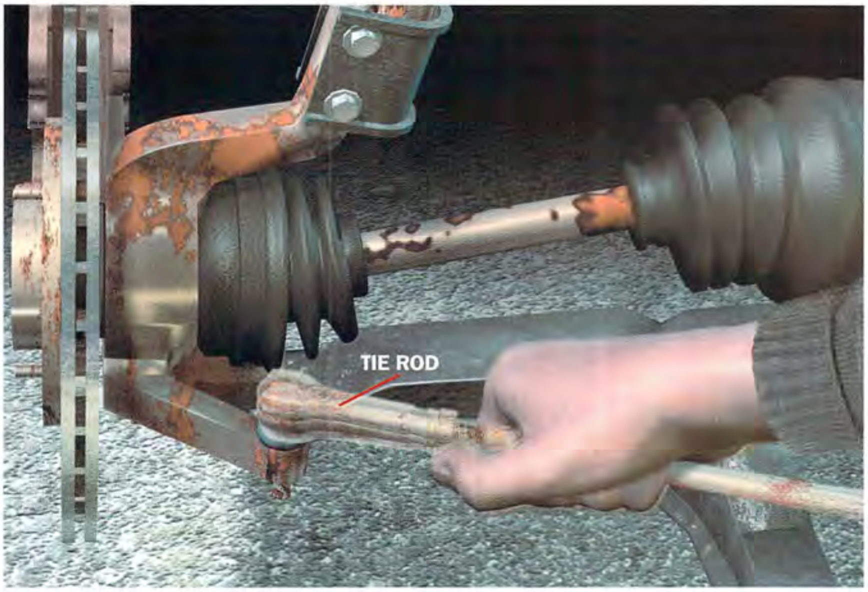

LOOSER: Worn tie-rod ends let the wheels wander, making steering vague.

Wanderlust

If your car doesn't have a definite pull to one side, but instead a sloppy, undisciplined tendency to wander all over the place, the first thing to think about is wear in the steering and suspension.

Direct observation is the best way to find this, but you'll need a helper. With the car sitting with all wheels solidly on the pavement, have him or her unlock the steering column, then rock the wheel vigorously back and forth while you stick your head underneath and inspect with a light. (Needless to say, do this with the engine off.)

With a parallelogram steering system—the kind with a Pitman arm coming out of the steering box—you may see the idler or Pitman arm moving up and down. Or you may notice slop in the tie-rod ends or perhaps more rotational action going into the steering box than coming out. With rack-and-pinion steering, pay special attention to the inner tie-rod ends, and make sure the rack housing itself is firmly mounted.

Another way to inspect is to slightly raise one tire off the ground (place the jack under the lower control arm, then have your assistant rock that tire side to side, then top to bottom, while you take a look).

Steering bag?

With the recirculating-ball steering boxes typically found on big domestic rwd cars, pickups and SUVs, lash will gradually develop between the worm and sector gears. You can eliminate this problem by doing an "over-center adjustment." You'll find an adjusting bolt or screw sticking out of the top of the steering box. With the wheels as close to the straight-ahead position as you can get them, loosen the locknut, then turn the screw clockwise to reduce lash. Do not overdo this or you can reduce interior tolerances too much, causing damage. You'll know you've gone too far if the steering wheel stays where you put it instead of returning to a straight-ahead position under normal road forces, especially when coming out of a turn.

Lash may exist in the joints that allow the steering column to transmit the helmsman's commands to the gearbox. The universal variety typically lasts forever, but the rubberized textile type, fondly known as a "rag joint," often deteriorates to the point at which there's excessive play.

Deteriorated upper control arm bushings can cause serious steering problems, and probably a lot of clunking to boot. Look down on them while your helper holds the brakes and shifts from Drive to Reverse and back. You'll see and hear excessive movement.

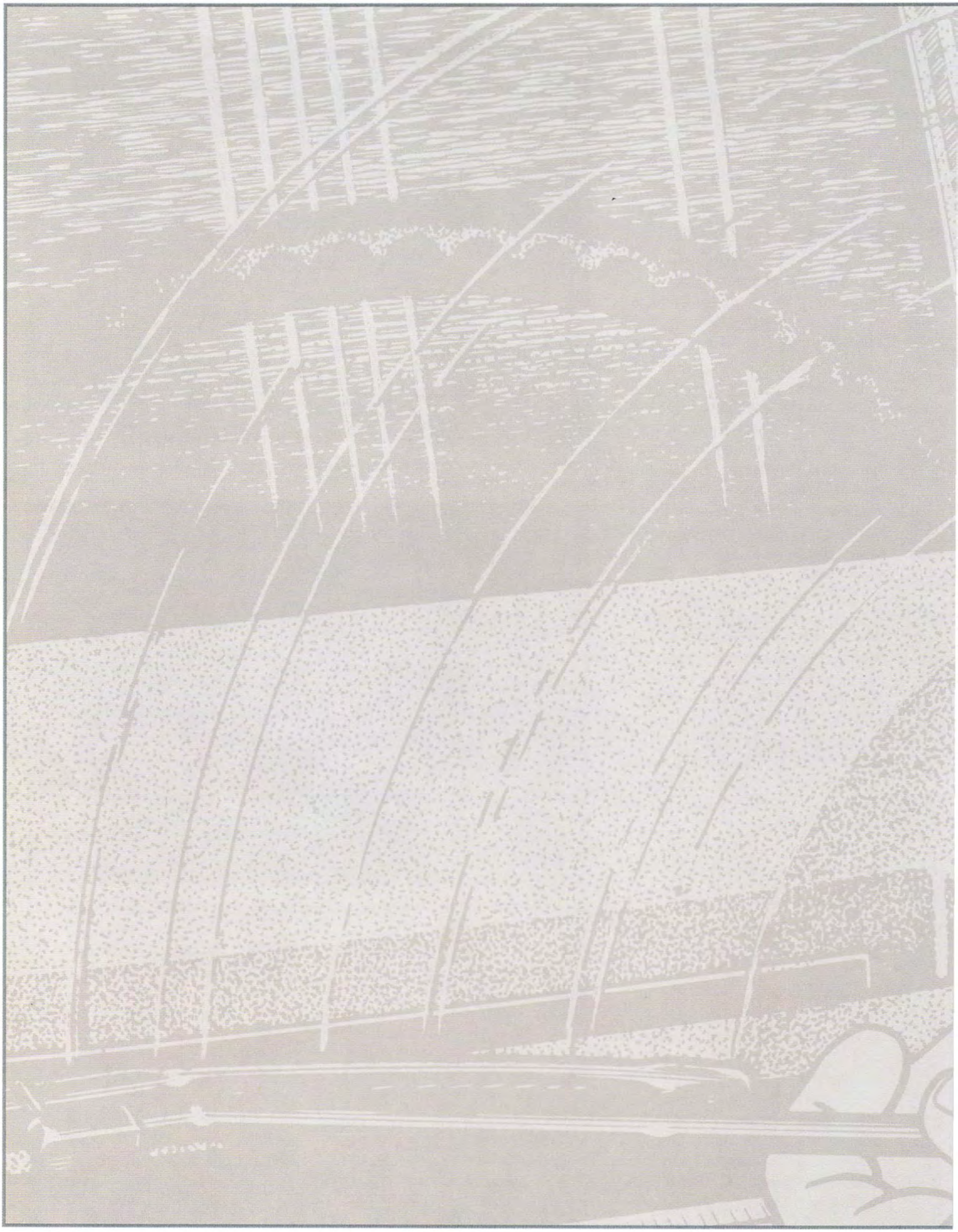

CHAPTER SIX

Interior and Exterior

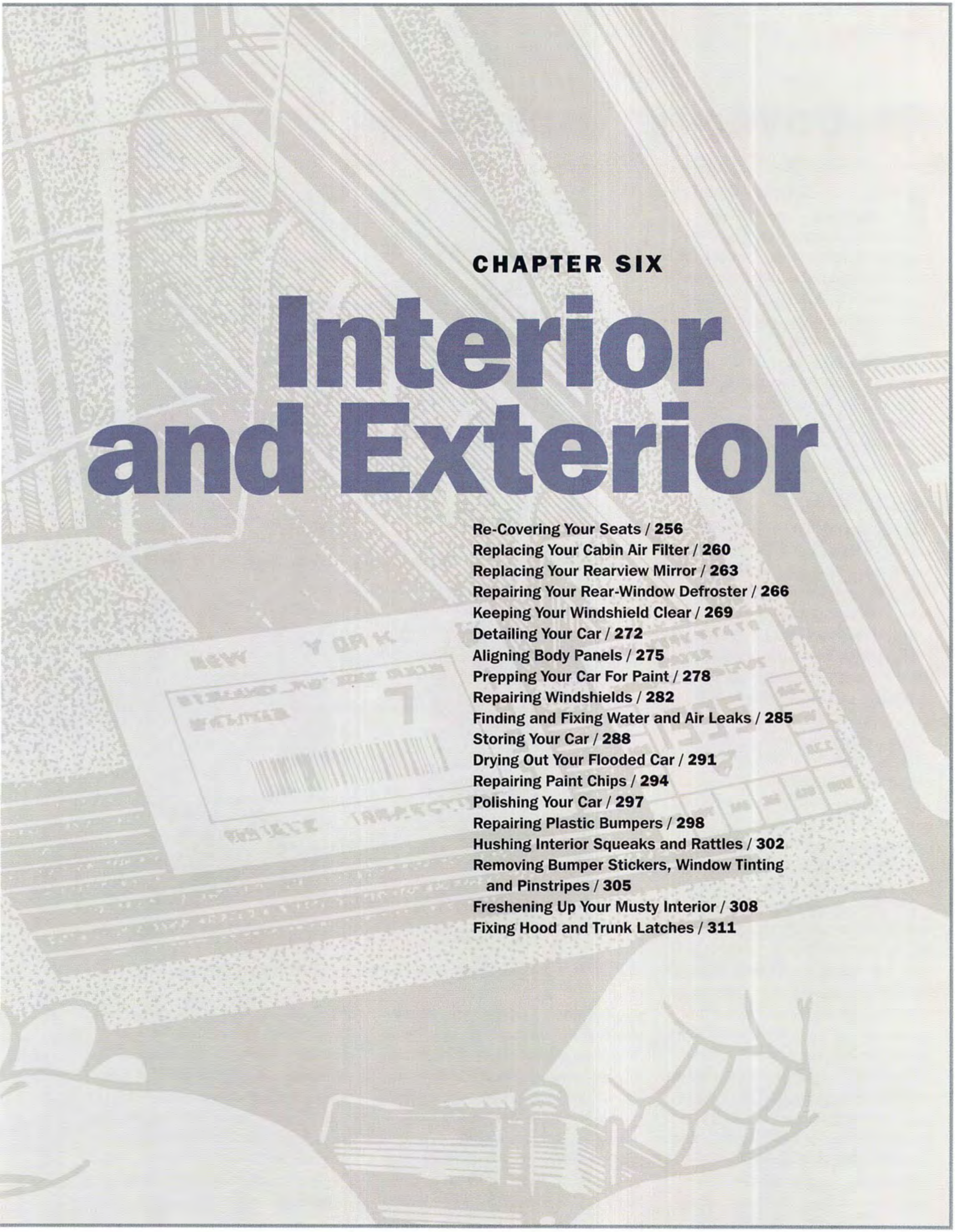

Re-Covering Your Seats / **256**
Replacing Your Cabin Air Filter / **260**
Replacing Your Rearview Mirror / **263**
Repairing Your Rear-Window Defroster / **266**
Keeping Your Windshield Clear / **269**
Detailing Your Car / **272**
Aligning Body Panels / **275**
Prepping Your Car For Paint / **278**
Repairing Windshields / **282**
Finding and Fixing Water and Air Leaks / **285**
Storing Your Car / **288**
Drying Out Your Flooded Car / **291**
Repairing Paint Chips / **294**
Polishing Your Car / **297**
Repairing Plastic Bumpers / **298**
Hushing Interior Squeaks and Rattles / **302**
Removing Bumper Stickers, Window Tinting and Pinstripes / **305**
Freshening Up Your Musty Interior / **308**
Fixing Hood and Trunk Latches / **311**

Re-Covering Your Seats

It's Saturday night, and you're taking the in-laws out to dinner. Unfortunately, the dog has been digging—unbeknownst to you—for bones he thinks are hidden under the back seat. The ensuing divot leaves your spouse's mother sitting at a considerable angle all the way to Denny's.

Morning finds you duct-taping large and small chunks of foam rubber back into place. Late morning finds you calling upholstery shops, followed by a quick balancing of the checkbook to see if you can afford it.

Rolling Your Own

The one-size-fits-all covers you've seen at the discount department stores are certainly inexpensive, but they look like covers more suitable for lawn furniture. And you know that going to the local car upholstery shop is likely to cost you a minor fortune.

"Basically, you have three choices," says Randy Smith of Brougham Seats Inc., "slipcovers, re-covers or getting the damaged seat panels replaced at an upholstery shop." A fourth choice also exists: Get replacement seats of an acceptable cover from a wrecking yard. If all seats have been damaged, consider this possibility.

Slipcovers, which are designed to slide over the car's original seating material, range in price from $25 for the discount store cheapos to $400 for custom-fit covers designed specifically for each vehicle. One company, Performance Products, even sells slipcovers for Toyota trucks and Ford Explorers that are made from scuba-diving wet-suit neoprene.

As the name implies, slipcovers slip over the existing cover. They're typically held in place on back-seat and front-seat bottom cushions with flex cords that attach to both sides of the cover under the cushion. The covers for front-seat back cushions snap on or are held by Velcro at the bottom where the seams can't be seen. Headrest covers attach similarly. Caution: If your vehicle seats have side airbags, covers could affect their deployment.

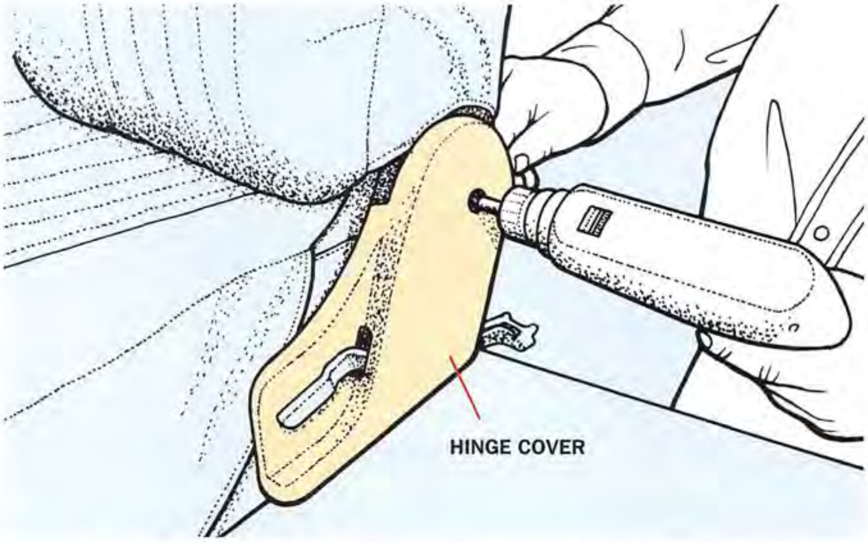

(Fig. 1) Start removing the front-seat top by removing the cover over the hinge.

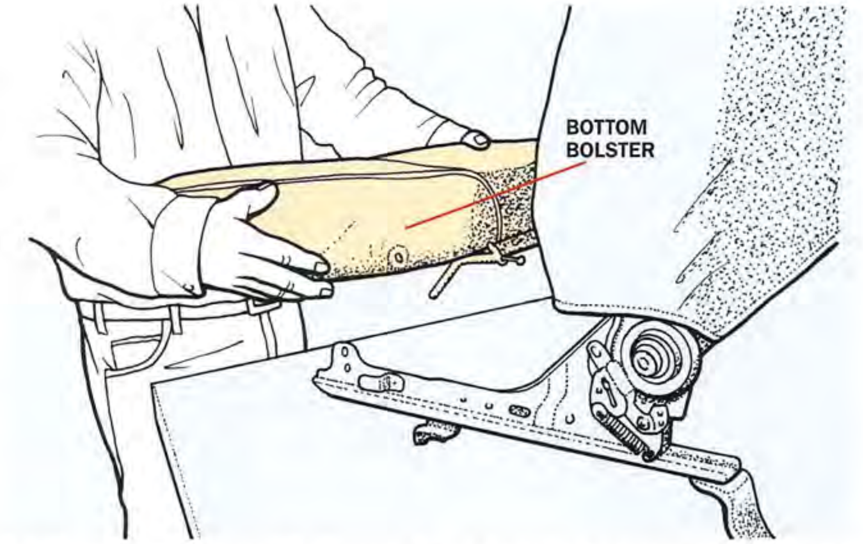

(Fig. 2) Once it's completely disconnected, pull the bottom bolster free of the frame.

Re-covers are designed to completely replace the original factory seat coverings. "It's not a do-it-yourself project," says Smith, whose company makes them. "They can only be purchased by upholstery shops, and they're normally installed in fairly new cars and trucks for customers who want to upgrade their car's interior, say from vinyl to leather." The cost? Typical installation prices range from $1500 to $3000, depending on the vehicle.

The third option—having an upholstery shop repair a vinyl, cloth

(Fig. 3) Rear-seat bottom cushions usually just clip in at the front.

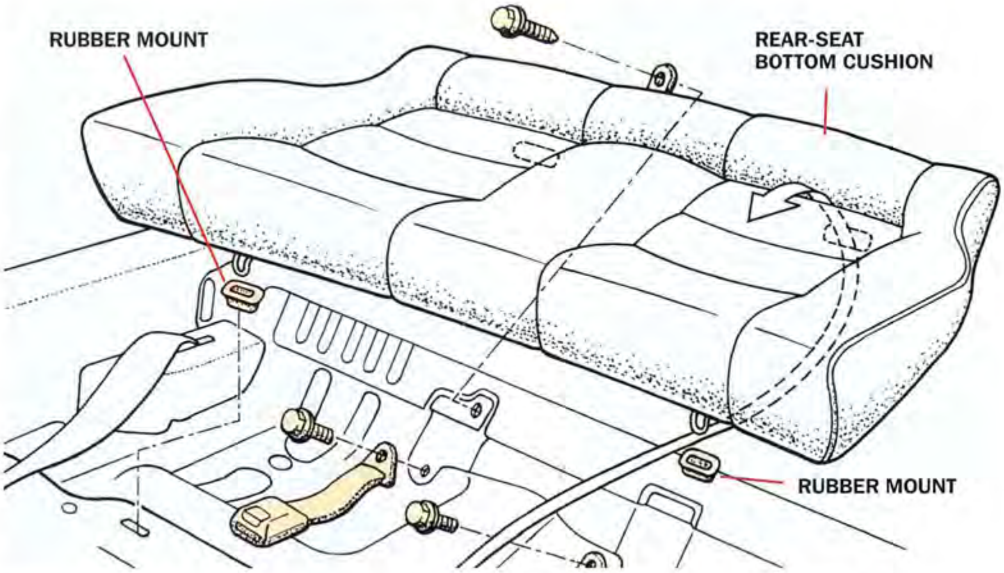

or leather panel on a seat—might actually yield the most attractive result and may even be the least expensive option, if the damage or tearing is limited to a single panel on a bottom seat cushion or seatback. You can save additional money by taking the seat out yourself, removing the hard-plastic skirts, splitting the back from the bottom cushion and delivering the damaged cushions to the shop for repair.

Of course, the drawback is that your car will be out of commission for a while if the work needs to be done to the driver's seat.

If you're planning to go this route, check with the upholstery shop first so it can order matching material ahead of time, then make an appointment to have the work done to reduce your vehicle's downtime.

Seat Removal and Disassembly

Repairing an upholstery panel is not easily tackled at home, but you can remove and disassemble the seats.

And while you don't necessarily have to remove the front seats to install the slipcovers, the finished job will look better and smoother if you take them out and then separate the cushions before putting on the covers.

If you're doing all the seats, remove the front seats first so you'll have room to maneuver the rear-seat cushions from the car. Doing it this way is especially critical on 2-door vehicles.

Most front seats are bolted to the floor pan at all four corners. Begin by removing any plastic covers **(Fig. 1)** and then remove the bolts **(Fig. 2)**. On power seats, disconnect the wires to the seat motor before lifting out the seats. Caution: If your vehicle has side airbags in the seats, refer to a factory service manual for instructions on how to disable them safely, as well as how to reconnect the system.

Typically, rear-seat bottom cushions must be removed before the back cushion can be removed.

Most bottom cushions are "hooked" to the floor pan in front **(Fig. 3)**. Firmly pushing the seat cushion rearward and then lifting upward disengages the cushion from the hooks so the cushion can be lifted out.

The back cushion is usually bolted to the rear bulkhead near the bottom of the cushion, while the top of the back hooks to the top of the bulkhead. Remove the bolts and lift upward to remove the rear cushion.

With the seats missing, you'll want to pick up loose change, petrified food and other materials. Now is also a good time to give the carpet a thorough vacuuming and even a shampooing if it's gotten really disgusting.

To disassemble the front seat, remove the headrest by sliding it from the seatback. You'll probably unbolt the seat top at the hinge pin. First, remove any hard-plastic skirting. If the seat has a reclining lever near the bottom cushion, you'll also have to disconnect the reclining mechanism from the seatback **(Fig. 4)**. After that, unbolt and separate the two cushions.

Installing Slipcovers

If the foam cushions are damaged, this is the time to patch or replace them. If there are chunks missing, you can simply carve matching pieces

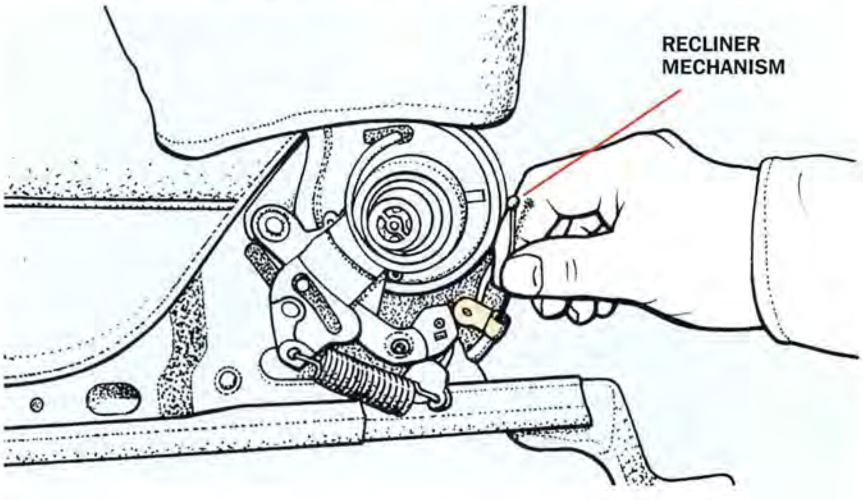

(Fig. 4) Disconnect the seatback recliner mechanism from the seat carefully.

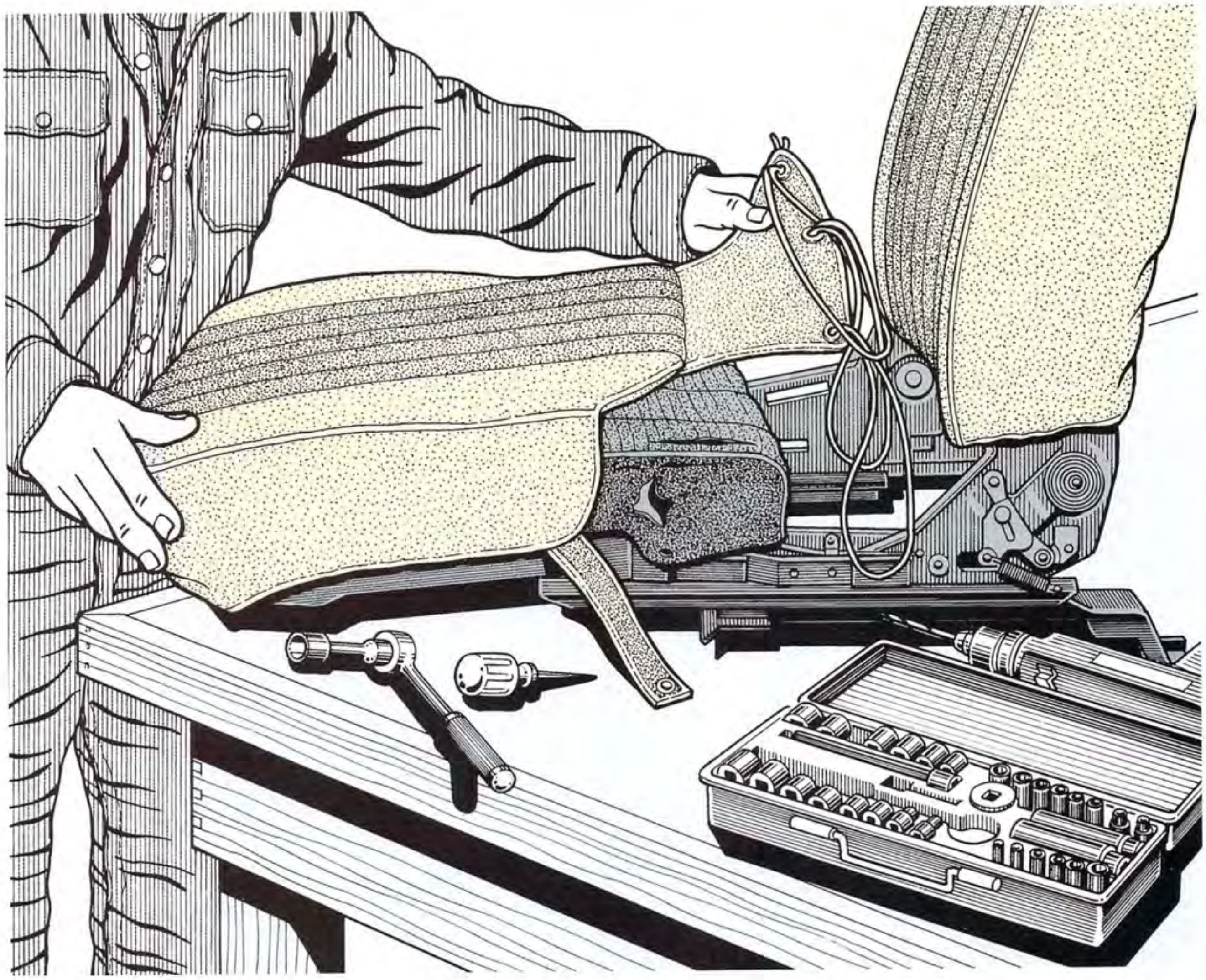

(Fig. 5) After you've removed the front seats from the vehicle, the new seat covers can be tugged over the seat bottoms.

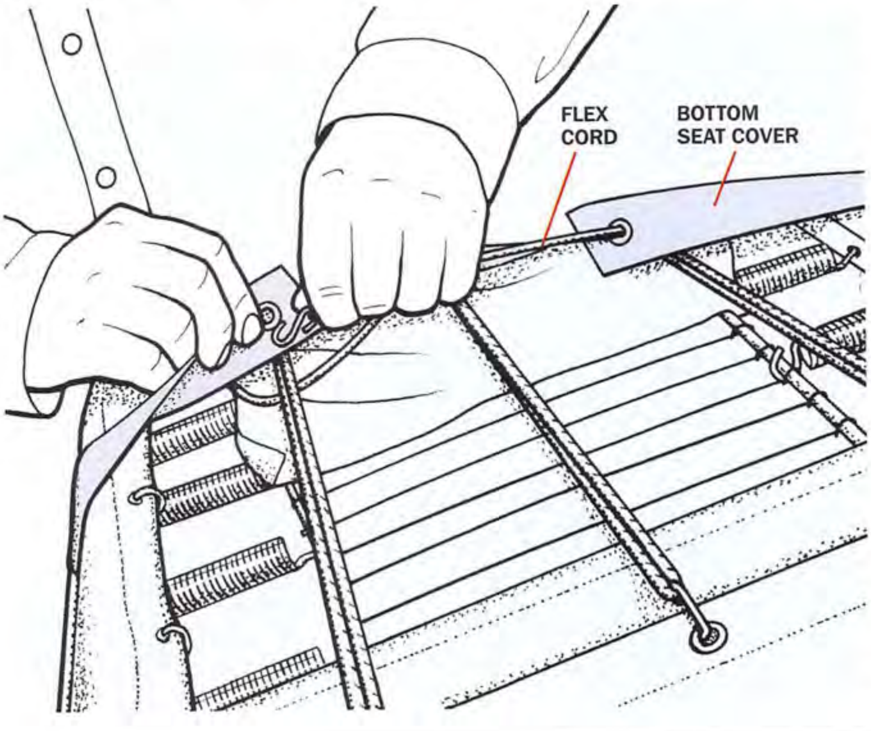

from upholstery foam to fit, gluing them to the existing foam in the cushion with contact cement.

Because durable materials are thicker and a tight fit makes for a better look, the more expensive the slipcovers, the harder they are to install.

Begin by sliding the cover onto the bottom cushion first (Fig. 5). You may need an assistant with strong fingers to help you pull on the cover and get it aligned on the seat. Once the cover is in place, move to the underside of the cushion where you can draw the flex cords to the other side. Attach

(Fig. 6) After the cover is on, add the elastic cords and connect the S-hooks.

them to the eyelets in the cover with the S-hooks that are provided with the cover (Fig. 6). To ensure that the flex cords stay in place, crimp the S-hooks closed with pliers (Fig. 7).

Next, slide the slipcover for the back cushion into place and snap or Velcro the bottom.

Install the headrest covers and any fold-down armrest covers. Before reassembling the seat's back and bottom cushions, make sure the slipcovers are properly fitted with no gaps or loose corners. If there are any, readjust the covers on the cushions until the gaps disappear.

After the seat cushions are reassembled, reattach the recliner mechanism and reinstall the plastic skirt.

The rear-seat slipcovers for both the bottom and back cushions install with flex cords in a manner similar to the front-seat bottom cushion. After the cords are in place, crimp the S-hooks with pliers and reinstall the cushions in the car.

Another suggestion is to take this opportunity to Scotchgard the new fabric while the seats are still out. This water- and dirt-proofing treatment is available at your local discount store in the automotive department, or at the auto parts store. It's a simple aerosol can, and you can avoid cleaning up overspray if you use it outside the car.

Reinstalling Seats

When reinstalling the rear seats, make sure the lap and shoulder belts are properly routed through their seat openings. Also, make sure any hooks and fasteners are firmly engaged (Fig. 8).

When reinstalling the front seat, be sure to torque the mounting bolts to the proper specifications. Properly route the seatbelts and then make sure that both the sliding track and reclining mechanisms work properly.

On electric seats, reconnect the motor wires and then run the seat through its full range of adjustments to ensure the connections are correct.

SUPPLIERS

- The Driver's Seat,
  1400 Glades Road,
  Boca Raton, FL 33431;
  (561) 368-7966.
- JC Whitney,
  1 JC Whitney Way,
  LaSalle, IL 61301;
  (800) 469-3894.
- Performance Products
  (Ford Explorer, Toyota 4x4s, Porsche),
  8000 Haskell Ave.,
  Van Nuys, CA 91406;
  (800) 553-2840.

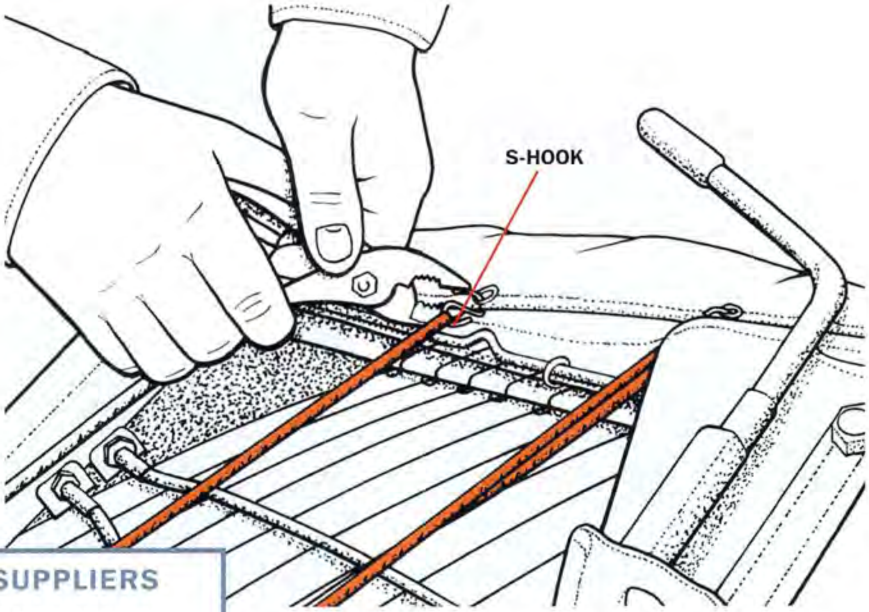

(Fig. 7) Crimp the S-hooks with pliers before reassembly.

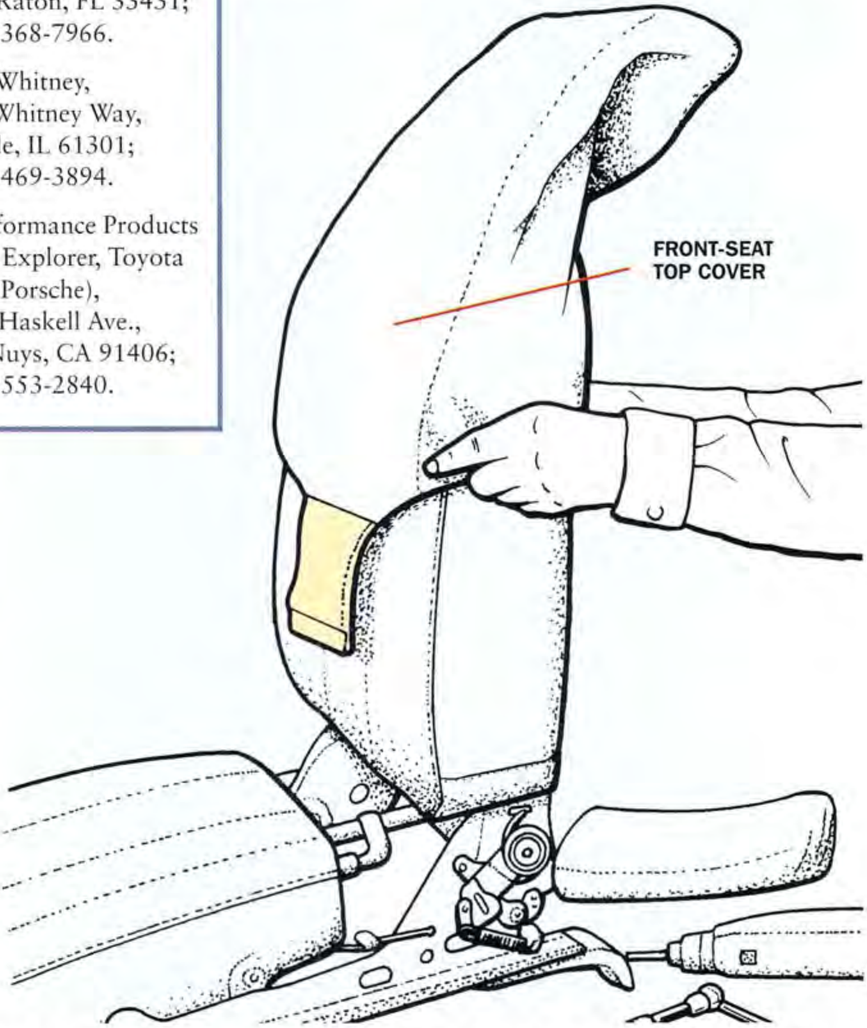

(Fig. 8) Install the front-seat top cover in the same way you would pull on a big, heavy sock.

Replacing Your Cabin Air Filter

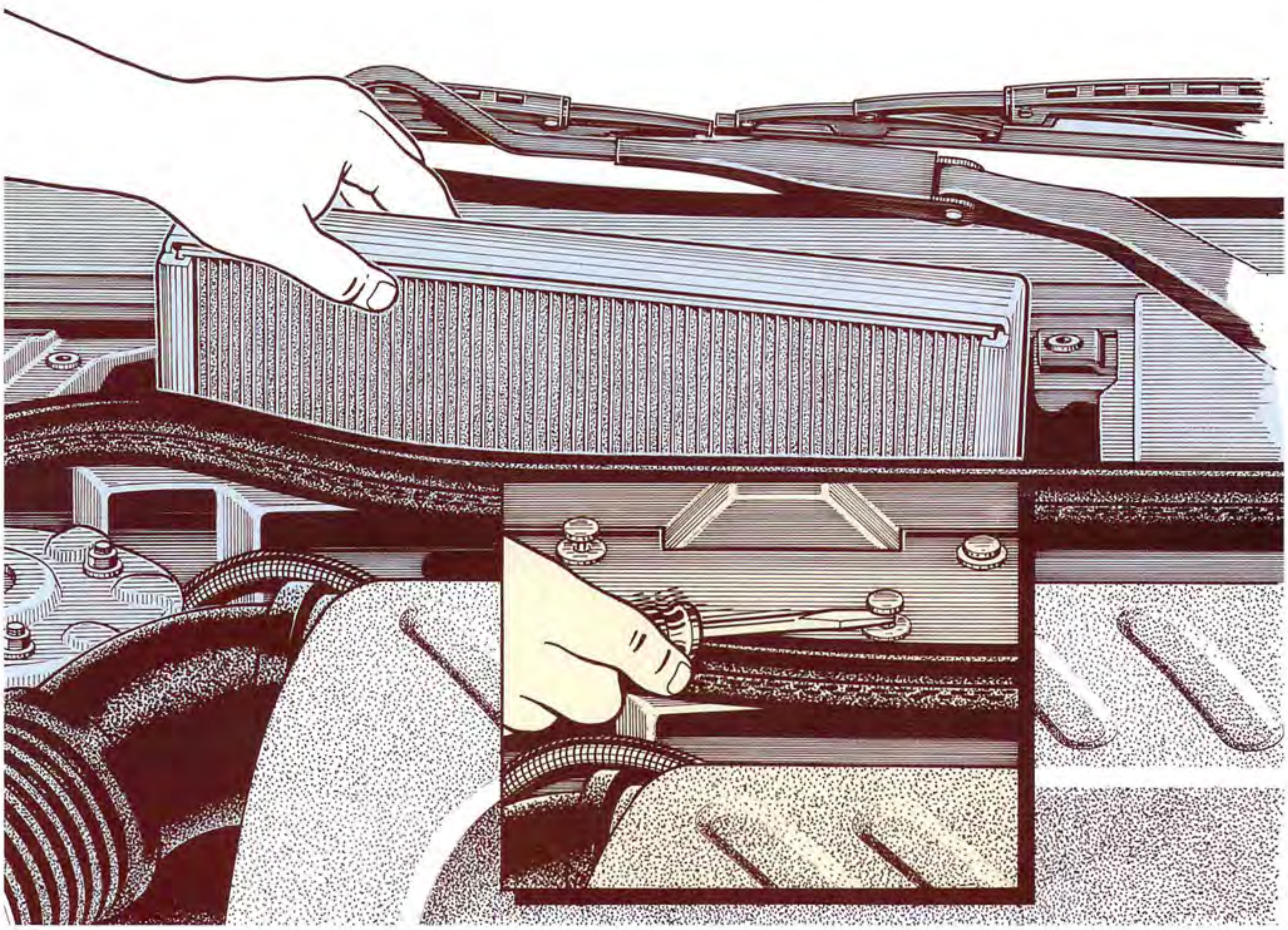

(Fig. 1) Underhood filter covers are often held in place with plastic button fasteners, which need to be pried up carefully.

It's Sunday and the big basketball game is this afternoon, but you can forget about getting tickets. So it's time to gather the crowd and drive to the sports bar. Minutes into the ride, the windshield is so foggy, you've got to hit the Defrost button. But why? The panel is set to outside air for heater operation, and there are no smokers in the car. You put your hand over the defroster outlets and there's hardly any airflow. When you turn the blower switch to high, the noise level goes up but the airflow is still weak.

Maybe you need to change the cabin air filter. If it's plugged, the passenger compartment might as well be a sealed cabin—and with a carload of friends the glass will fog in minutes. Filter life depends on the air quality in your area, but a year, or 12,000 to 15,000 miles, typically is the recommended replacement interval.

Cabin air filter? In the late 1980s and early 1990s, only some European luxury cars had them. It's more common on newer cars. In many premium cars, there may be two filters, even three. Sometimes they're parallel (side by side), sometimes they're in series (one after the other, with some space in between). Sometimes they're nowhere near each other.

Locating the Filters

A cabin air filter may be in the outside air intake, visible with the hood up and perhaps a cover lifted. Or it will be under the dash in one or both of two general locations:

- Above the blower in the back section of the outside air intake, a location well protected from moisture.
- Between the blower and the rest of the HVAC case.

If the owner's manual doesn't tell you the location, and you can't see it with the hood open, check behind the glovebox—perhaps you'll see an obviously removable (with a spring-tab or similar retainer) rectangular plastic cover. Next, look under the dash (on

(Fig. 2) If any debris has contaminated the filter plenum, vacuum it out and then wipe with a damp rag.

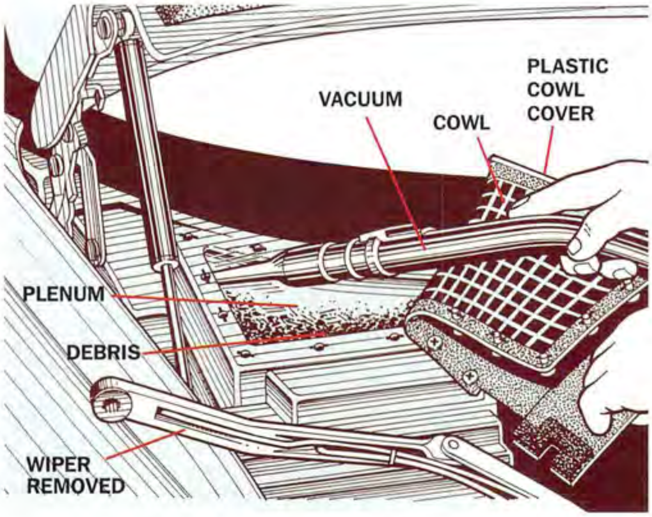

both the driver's and passenger's sides) for a removable plastic cover in the HVAC case. If you can remove its cover, look inside the HVAC case.

Still no answer? Buy the filters, something you have to do anyway. You've got two choices:

- Buy the factory filters at the dealership at full price. This way, you'll know for sure what they look like and get some idea about how they have to be installed. You'll also find out how many your vehicle has. The parts department should have a diagram to show you where the filters are located in the car.
- Buy aftermarket filters that come with step-by-step instructions from an auto parts store. The odds of the original equipment parts coming with instructions are extremely slim. Do all aftermarket filters have them? No, but if you're shopping at a parts store, you should be able to check in the boxes on the shelves. Also, these parts probably will cost less.

Replacing the Filters in the Air Intake

The filter in the outside air intake (even if there are other filters) takes care of most of the dust and pollen, so it's likely to plug first.

The typical procedure begins with removing the air intake screen, which is held by "Christmas tree" plastic fasteners. If the "tree" has a center pin, pry it up with a thin-blade screwdriver and then wiggle the tree out of the hole **(Fig. 1)**. If the tree is a single-piece design, you'll have to pry it out. Most of the time, it will break and you'll have to replace it (blister packs of replacements are sold in most auto parts stores).

If there's a rain/snow deflector over the filter, remove it, then extract the element. The filter element may be held by a plastic tab, or it may slip into a retainer housing that you will reuse with the new element. Some dust may have gotten around or through the filter. Use a vacuum cleaner to clean the outside air intake duct **(Fig. 2)**. Be careful when you reinstall the air intake screen, as well as any cover over the filter element. If the gasket surface around the screen does not seal, or if the cover (actually a shield, in most cases) is not reinstalled, rainwater can flow into the filter, saturating it.

In the Case, above the Blower

If the filter is in the case above the blower, most likely it is behind the glovebox. Removing the glovebox may be easy, but you may not have to remove it at all.

Many filters (such as on most Hondas and Toyotas) are just above the blower behind the glovebox, and you'll have to drop it for access. In the best of cases, you can flex the sides of the box so the pins clear the dash, then pivot the box down for clearance. In others, the entire glovebox assembly has to come out, which means removing a number of retaining screws, not all of which are visible.

Replacing Midcase Filters

The midcase filters may not always be easy, but remember, the setup was designed to allow filter replacement, so don't assume the worst.

There are pull tabs taped in place in GM sedans that have two or three filters in the same HVAC case located

SOME UNDER-DASH FILTERS can be replaced without any tools by simply opening a small access door.

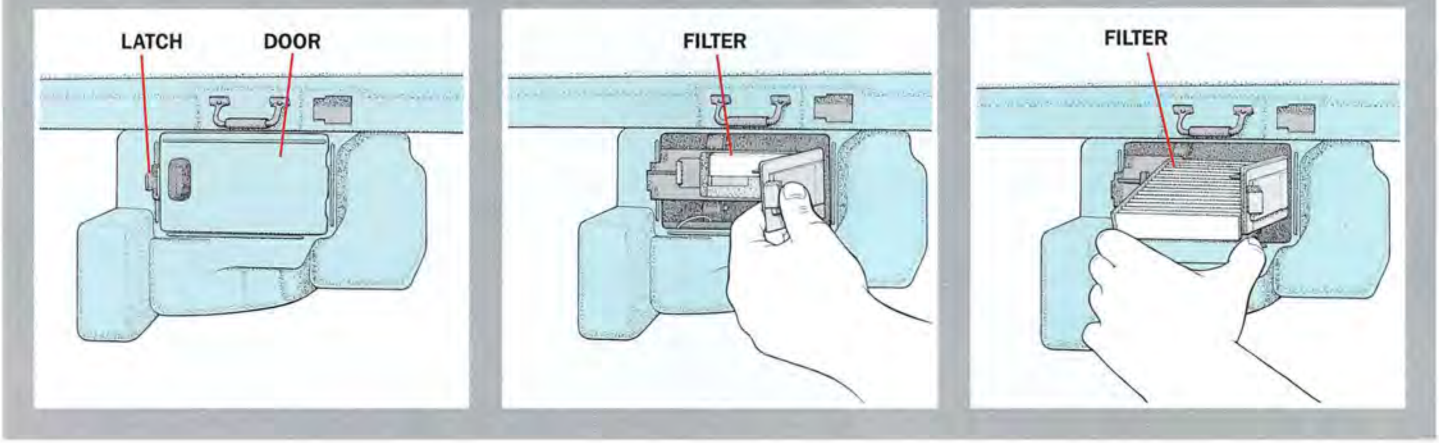

above the gas pedal. Pull a tape to release the tab, then pull the tab to withdraw the filter. Ditto for the others.

All filters need perimeter sealing, so air flows only through the filter element. With side-by-side filters, the sealing may be a simple foam rubber strip. Or there may be interlocking guides. On many cars the filter will come straight down and hit the floor without clearing the HVAC case. Nothing's wrong. Most of these filters are flexible, and you have to bend them to get them out and the new ones in.

INTERIOR-MOUNTED FILTERS may be stacked inside the plenum and housed in a plastic tray.

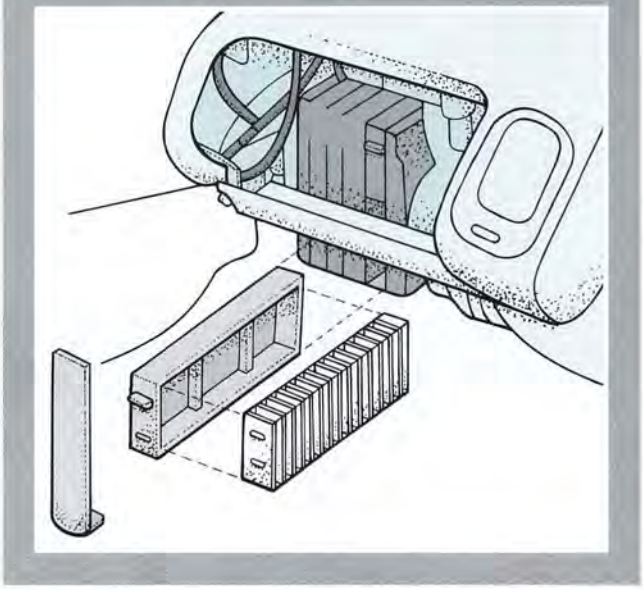

How to Buy

Even if the filter location is accessible, the toughest job may be getting the right replacement filter. One of the most common complaints among professionals is that the aftermarket catalogs haven't kept up with the year-to-year changes, particularly with filters that fit under the dash where the space provided may have changed.

If a vehicle has more than one filter, you may not find both in the same box, especially if they are in different locations. And they may have totally different part numbers. This is particularly true if one filter is in the air intake and others are in the HVAC case. The one in the air intake is more likely to be a simple particulate filter, one that just traps small particles. The one under the dash is more likely to have a charcoal layer for odor control.

No matter what the original equipment element is, the aftermarket one is more likely to be a particulate filter, rather than the more sophisticated type with a charcoal layer. Should you get the charcoal filter, which may cost a lot more? If you're in a densely populated area with slow-moving traffic, or your carpool partner fancies Mexican food, the extra cost is worth it.

If you're not facing an odor problem, you may save money with a low-cost particulate-only replacement for an original-equipment filter with a charcoal layer. An aftermarket particulate element does filter finely enough to remove pollen as well as dust particles, so if pollen and dust are your major concerns, save your money.

HOW IT WORKS

Cabin Filter

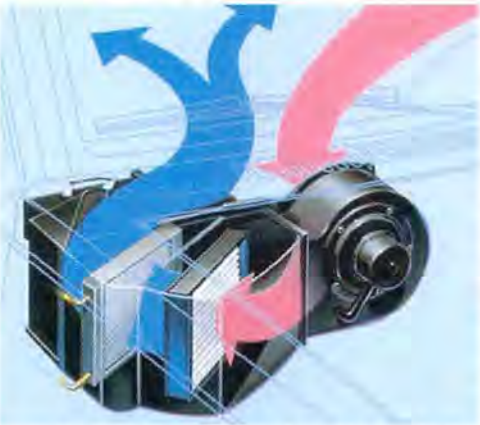

The cabin filter is a type of air filter that in some respects—such as pleating to provide a lot of surface area in a small package—may resemble an engine air filter. However, it usually is made of different materials, the main one being a charcoal layer in many elements. Like some engine air filters, the dust and pollen filtering element is electrostatically charged, which means it's made with fiber particles that have been given either a positive or negative static electrical charge. Airborne particles are negatively or positively charged, so they are attracted to the opposite-charge fibers.

In addition, the filter performs "mechanical filtration." This means that fiber material with pores of a specific size will trap particles larger than the size of the pores. The charcoal filtration layer removes odors.

The charcoal is "activated," which means it is treated with chemicals and heat to give it specific odor-control properties. Each manufacturer has its own recipe, based on tests with individuals who have "sensitive noses," including people who work as perfume testers. Charcoal traps odor-causing gases by adsorption (not absorption), holding them on the surface of the material. Because charcoal is porous, each pore is an exposed surface, so the filter has a very high capacity for trapping noxious gases.

EASY

Replacing Your Rearview Mirror

The traffic seems unusually light this morning. In fact, you haven't seen another car to the rear for 10 miles. Time to make a little time.

But then there's a strange noise following you. Sounds like a siren. Hmmm. No lights behind you, so where is that sound coming from?

It's coming from the police car that's riding your tail. You didn't see it because your glued-on rearview mirror isn't on the windshield. It's lying on the floor under the seat where your 4-year-old has hidden it after discovering he couldn't do chin-ups on it. Tell that one to the judge.

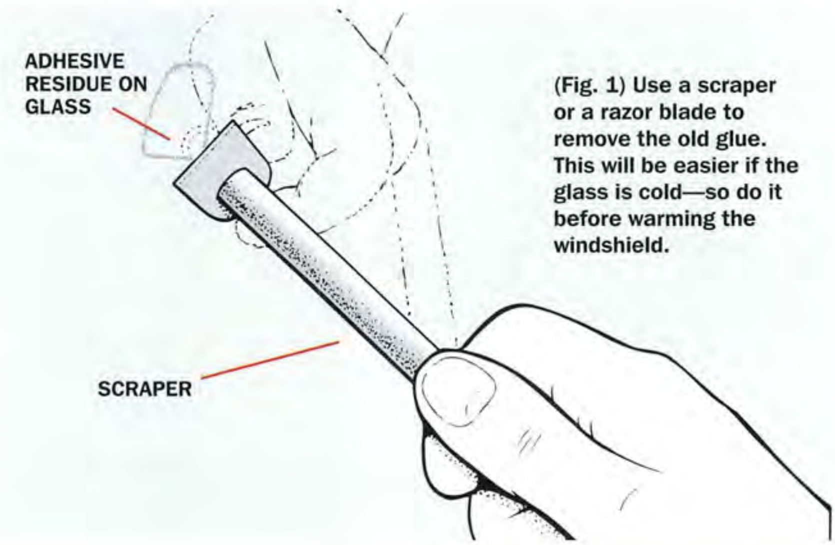

(Fig. 1) Use a scraper or a razor blade to remove the old glue. This will be easier if the glass is cold—so do it before warming the windshield.

HOW IT WORKS

The Antiglare Mirror

DAY

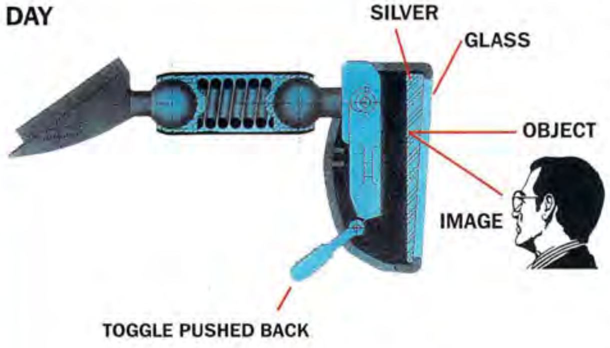

NIGHT

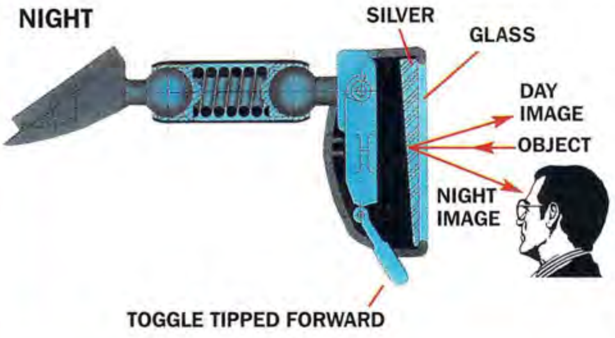

Most cars and trucks come with a manually dimmable day/night prismatic rearview mirror. You toggle a little lever from the day position to the night position. And the mirror glass is not flat, but rather beveled. The top edge is thicker than the bottom edge. Here's how the day/night function works.

In the day position, the image you see is reflected off the silvered surface on the back face of the mirror glass. The image is crisp and bright, because the silver provides an almost total reflection (80%) of the object.

In the night position, the toggle mechanism tilts the mirror glass upward slightly. Now, the bright day image is deflected off the silvered surface and up into the vehicle's headliner. And the front (plain glass) surface of the mirror takes over, providing a 5% reflective image. Everything appears very dark, but you can still monitor traffic behind you.

Many carmakers offer an automatically dimming electrochromic rearview mirror that provides an infinite number of brightness levels. These darken automatically according to ambient light conditions and the amount of headlight glare. Sensors feed input to electronics that determine how much voltage to apply to a special gel sandwiched between thin layers of reflective glass in the rear and clear glass up front. The more glare present, the more voltage is applied, and the darker the gel becomes.

Reattachment

Luckily, gluing your rearview mirror back on is one of the easiest jobs for a Saturday mechanic. What you need most is patience and a clean working environment. Warm weather helps, too, but it isn't mandatory.

Start by running down to your local auto parts store and picking up a rearview-mirror adhesive kit. It'll come with a vial of superstrong glue and a vial of accelerant to help the glue cure very quickly. There are usually instructions on the back of the cardboard packaging, too, should you need some guidance. You'll also need a razor-blade-type scraper, a small Allen wrench that fits the lockscrew on your mirror's baseplate, a grease pencil or some masking tape and, depending on your particular circumstances, a tape measure and a heat gun or hair dryer.

Ready?

Prepare to work outside with the windows open, as the adhesive kit will probably contain acrylic acid, methacrylic ester and trichloroethylene, which you don't want to breathe in large quantities.

Because the adhesive cures faster in warm ambient temperatures, plan on using the heat gun or hair dryer to warm the glass if you're forced to do this job on a cold day. But use common sense. Ice-cold glass will not appreciate being blasted by intense heat. You'll almost definitely crack the windshield this way. To avoid this, hold the heat gun or hair dryer about 12 to 18 in. away from the glass and waft the warm air back and forth to gradually increase the glass temperature. And make sure to scrape off the old glue before warming the glass or you'll create a gooey mess (Fig. 1). An alternative is to park the car facing the sun for a few hours before beginning the job.

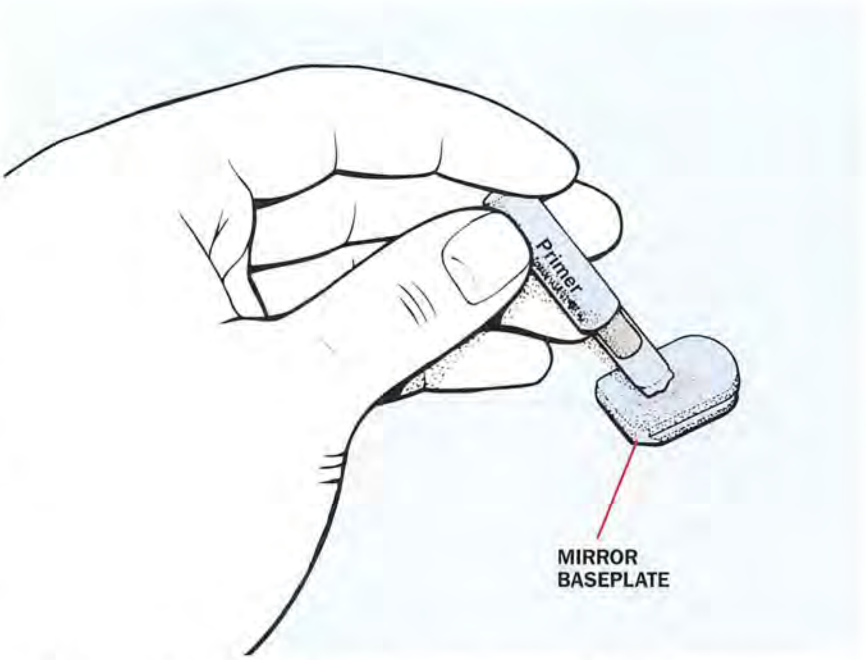

(Fig. 2) Applying accelerant to the mounting surface speeds the adhesive curing. Don't touch the surface with your fingers.

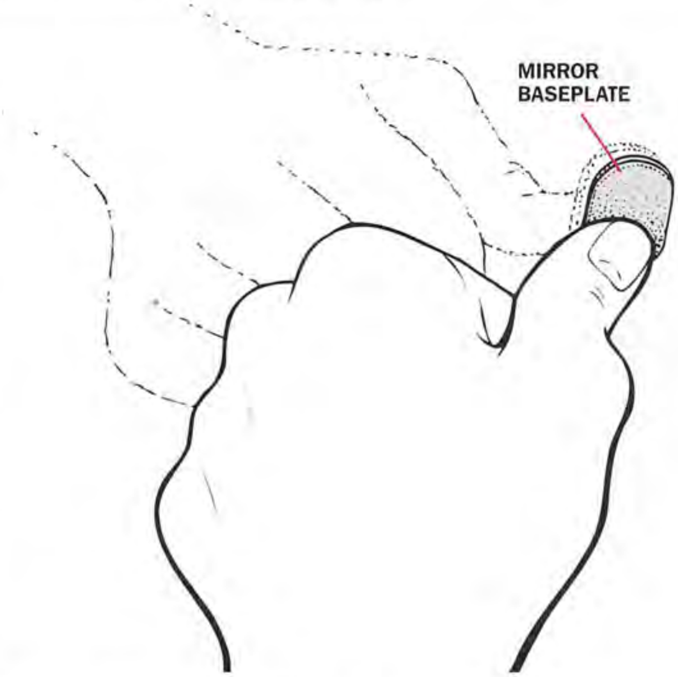

(Fig. 3) Press on the baseplate firmly for at least 2 minutes to allow the fresh adhesive to start to set up on the glass.

If the old mirror left behind an obvious amount of residue on the inside of the glass, you won't need the tape measure to locate the midpoint of the windshield. But if you, say, forgot to mark the spot before you cleaned the glass completely, simply measure halfway across the top of the window and about 3 to 4 in. down from the headliner, and that will be your new mounting point.

You can begin the remounting process by marking the outside of the window with a grease pencil over the area where the old glue is. Or you can cordon off the area on the inside with masking tape. (Some sort of mark is useful should you get interrupted after removing the old glue.) The latter is the neater way to work, but leave adequate room around the work area for scraping and cleaning. Now, go at the old glue with your scraper.

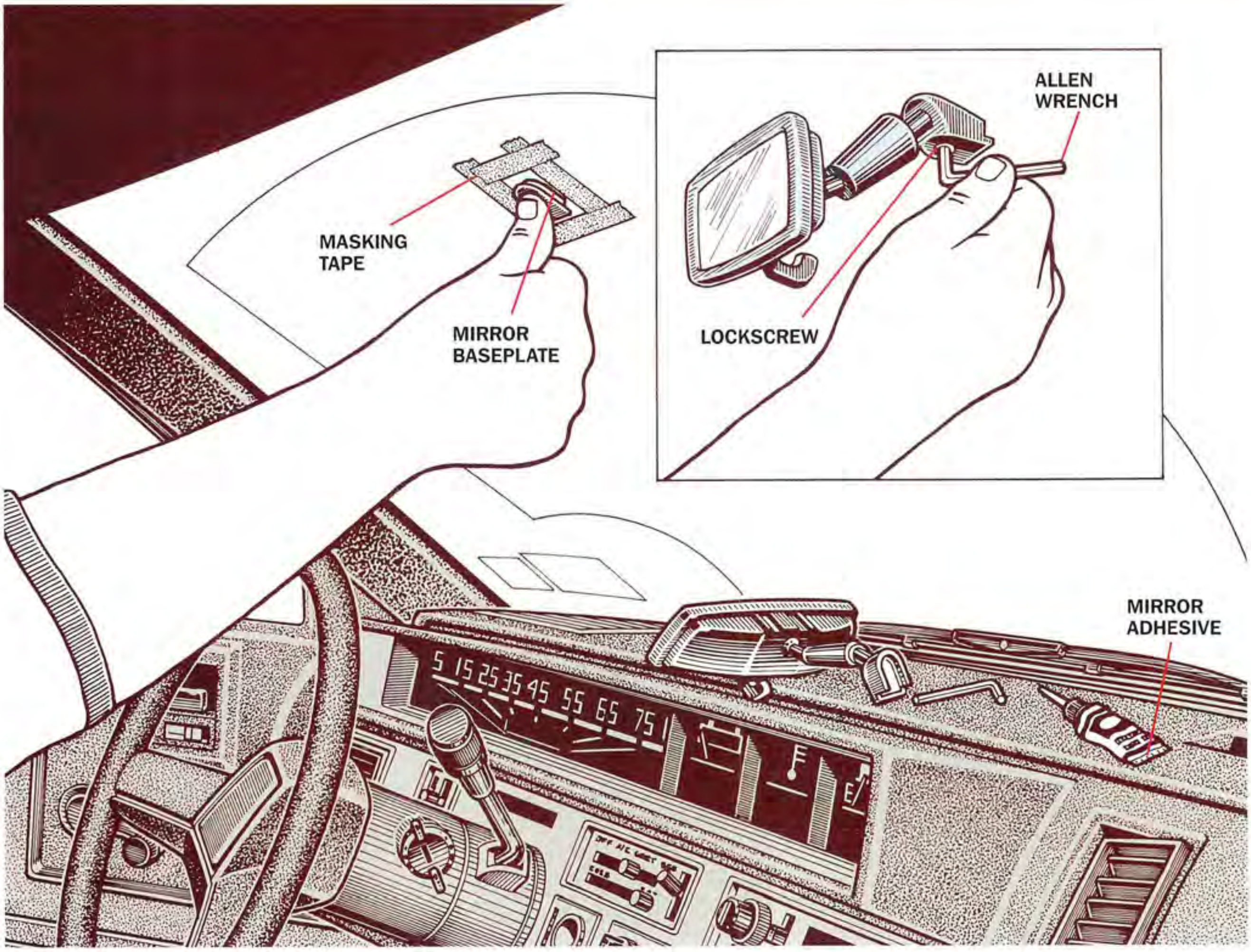

(Fig. 4) After gluing the baseplate onto the glass, mount the mirror and tighten its lockscrew with an Allen wrench.

Once you're down to smooth glass again, wipe the spot down with isopropyl alcohol to get it squeaky clean. Do the same thing on the back of the mirror baseplate so that you have clean metal to work with. Don't touch either surface once it's cleaned or the oil that comes off your fingers will contaminate it.

The Acceleration Phase

The accelerant (primer) vial will have a double wall of plastic. The inside wall is designed to crack open when you bend the tube. The outside wall simply flexes. This allows the accelerant to drench the felt tip of the vial without leaking out all over your hands. Apply the accelerant liberally to the mounting point on the glass. It needs 5 minutes to dry.

In the meantime, use your Allen wrench to undo the lockscrew and release the mirror from its baseplate. It's much easier to glue the baseplate onto the windshield first and install the mirror later. Working with the whole assembly is too awkward.

Once the baseplate is free, goop up its flat mounting surface with the accelerant and let it dry, as well (Fig. 2). Don't touch either of the treated areas as they're now ready to accept the adhesive, which will cure faster because of the accelerant.

Adhesive Time

Now you're ready to apply the adhesive. Neatly goop up the back of the baseplate and carefully press it onto the glass. Squoosh it around just a little bit to get good adhesion—and keep applying pressure. Don't let go for at least 2 minutes while the adhesive begins to set up (Fig. 3). Then let it dry for at least another 15 minutes. If it's really cold out, a little warm air from the heat gun or hair dryer should speed up the curing process. Remember, 15 minutes is a minimum. It's best to let the adhesive cure as long as it is convenient—even overnight—before reattaching the mirror to the baseplate and tightening the lockscrew (Fig. 4).

If you're a neatnick, trim the excess glue around the baseplate with a knife or one of the corners of your scraper blade. Then reattach the mirror and readjust it for a safe rearward view.

Oh. And don't pull on the mirror to see how tightly it's mounted. With Murphy's Law always in effect, you just might rip it off the window again. Just leave the mirror be—and keep your kid out of the front seat. The mirror will probably stay up there for the life of the car.

EASY

Repairing Your Rear-Window Defroster

It was a dark and stormy night. Snow has covered the entire car to a depth of several inches. It takes a few minutes to clear the snow, and by then the engine has warmed up enough to melt the frost clinging to the inside of the windshield, and has actually made inroads to the frost on the side windows. But the rear-window defroster clears only a narrow strip near the top and bottom of the window, leaving a wide strip of glass as translucent as a barrister's door, and absolutely no visibility to the rear.

A worse scenario: The rear-window defroster grid has packed it in completely, requiring you to either drive blind (bad idea), or resort to scraping the rear window clear with your trusty plastic scraper.

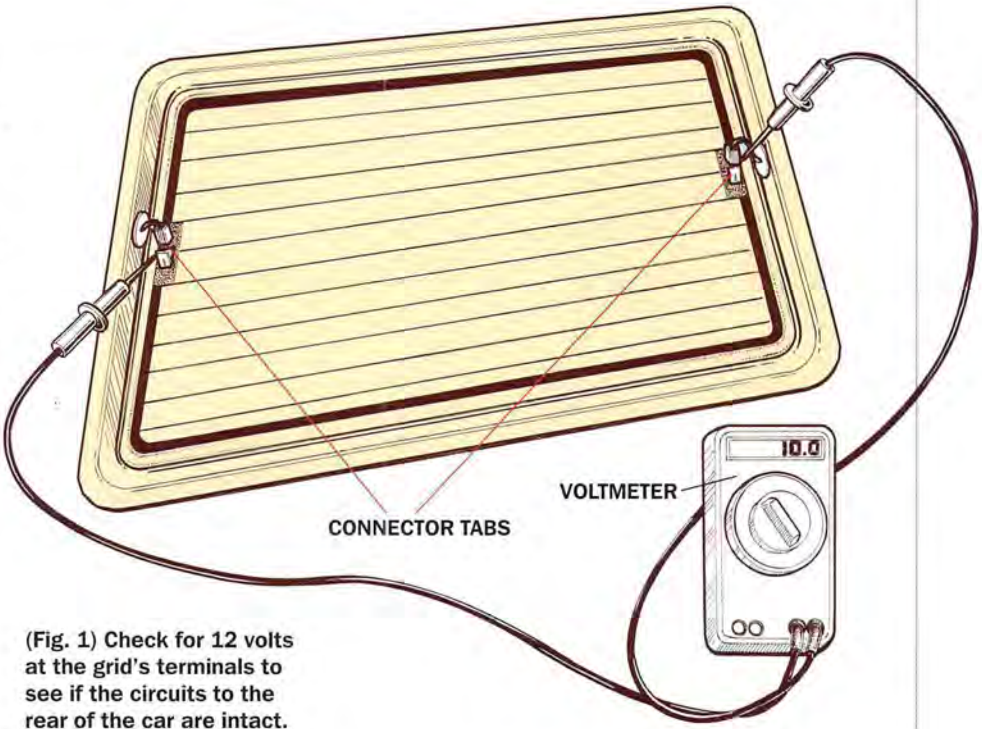

(Fig. 1) Check for 12 volts at the grid's terminals to see if the circuits to the rear of the car are intact.

Danger!

Let us make this perfectly clear: The resistance wires are silk-screened, essentially painted, onto the glass. They are very easy to scratch, and will not work properly if the scratch breaks the continuity along the wire. This means that it's possible for boxes, furniture or any other hard object one might place in a car to scratch the wire. Even a credit card can damage it. Don't cram stuff into the back, and don't let the load shift backward in your minivan so that it touches the glass. If your window has a defroster grid, the only thing that should ever touch the glass is a soft cloth dampened with window cleaner. If you must clean the rear glass, scrub gently, and in the direction of the grid, not across it.

You don't want to have to replace the glass to effect this repair. The compound-curve style of rear window is several hundred dollars, and a large hatchback's backlite on, say, a Camaro, can easily top a grand or more. Add the price of installation at the dealership or a glass shop—it's definitely not covered by insurance.

No Heat at All

Suppose your defroster grid doesn't work at all. First, check the obvious: Is the fuse okay? Defroster grids draw a lot of current (10 to 20 amps), and if the fuse is undersize, it won't last. If the fuse doesn't look blown, check with your voltmeter—with the key on and the defroster on, you should see 12 volts at both grid terminals (Fig. 1).

If the voltage is fine, the problem is somewhere in the wiring or at the grid. Check the connections from the wiring harness to the grid. It's easy for the terminals at the grid to become damaged. Generally, the tab that's attached to the glass breaks off, leaving you with a dangling wire and no way to reattach it. You have two repair options here: soldering and gluing.

If you know how to solder and have a high-capacity soldering iron or gun, solder the tab back on (Fig. 2). It may take a third hand to hold the tab against the grid while you solder it. There's usually a metal strip laid on the glass under the silk-screening. Clean the surfaces with alcohol and use 60-40 rosin-core solder. Work fast, because excess heat may crack the glass.

If you aren't confident about your soldering skills, or aren't ready to take a chance on cracking an expensive piece of glass, there's another way. The dealership and most auto parts stores can sell you a special electrically conductive epoxy to bond the tab back on (Fig. 3). If it's wintertime, you'll need to work in a heated garage, and have the vehicle inside long enough for it to warm up to at least 65° F. Again, clean the area with alcohol. Mask the glass with tape to keep from getting epoxy smeared on it. Mix up a sparing

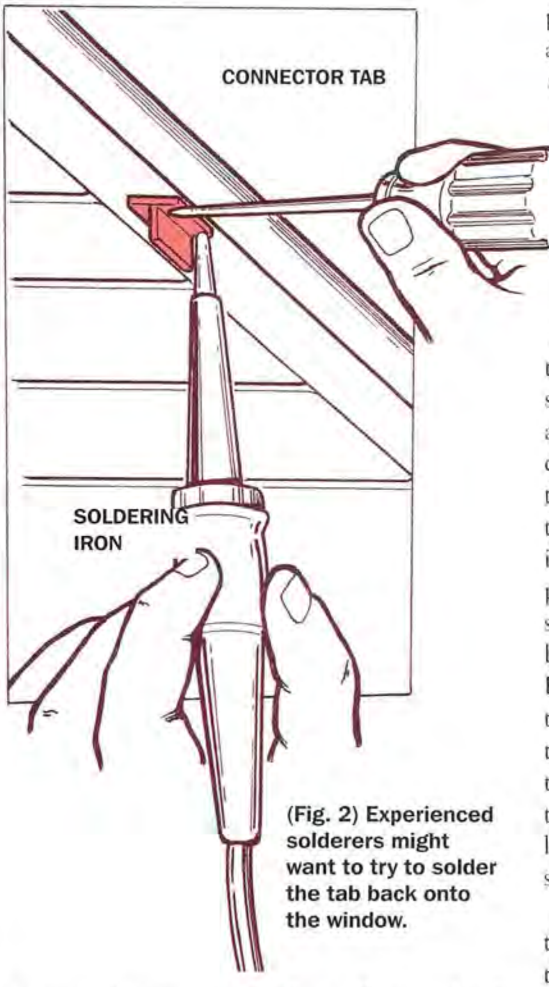

(Fig. 2) Experienced solderers might want to try to solder the tab back onto the window.

amount of epoxy and hardener. Put some epoxy on the tab, and use an ice pick to hold it in place for the 10 minutes or so it will take for the epoxy to harden. You can use a wooden stick or the end of your dampened finger to smear the epoxy within a minute or two of application to improve the cosmetics of the repair. Although the epoxy will set up rapidly, don't attempt to reattach the wiring until it's had 24 hours at 65° F or more to cure and achieve its full strength. The repair will never be as strong as the original wire, so you'll need to be particularly careful not to damage it in the future.

Deeper Problems

Fuse okay? Grid attached to the glass everywhere, but simply no defrosting action at all? Look for a bad switch, relay or timer. For this you'll need a schematic diagram, or considerable experience in troubleshooting wiring problems. Start at the fuse and trace the wiring. If the switch is bad, you'll be able to jumper the switch and get 12 volts beyond there for diagnosis. But you may need to replace the timer—which may be integrated into a larger box of electrical controls buried under the dash. Consult the factory shop manual for your car for a detailed diagnostic procedure. If that's not available to you, you'll have to find the problem the old-fashioned way. Trace the current path from the fuse, to the switch, to the timer, and on back to the window. Remember, the timer will turn off the current within 5 to 10 minutes, so you'll need to keep track of the timer's time window or you'll be looking for current that's not supposed to be there.

Somewhere in the circuit there will be a relay to switch the high current necessary for the grid's operation. This may or may not be integral with the timer. A diagnostic procedure would be to jumper the relay's terminals to see if the relay is bad. You can either jump 12 volts to the relay's coil to make it pull in, or bypass the relay with a large-gauge jumper to see if the grid's wiring is intact between the relay and the window. A separate relay should be inexpensive and available at any auto parts store, but if it's in the same package with the timer you'll pay as much as a hundred bucks.

Don't try to bypass the timer. Rear-window defrosters draw a substantial amount of current—10 to 20 amps depending on the application. Most modern cars use a timer circuit to turn the grid off after a reasonable length of time. There are two reasons for this. The first is to reduce the electrical load on the alternator, which (especially during the wintertime) also supplies electricity for the headlights, heater fan and windshield wipers. Couple that with the extra demands on the battery for starting in cold weather, and there may simply not be enough alternator capacity to keep the battery charged adequately.

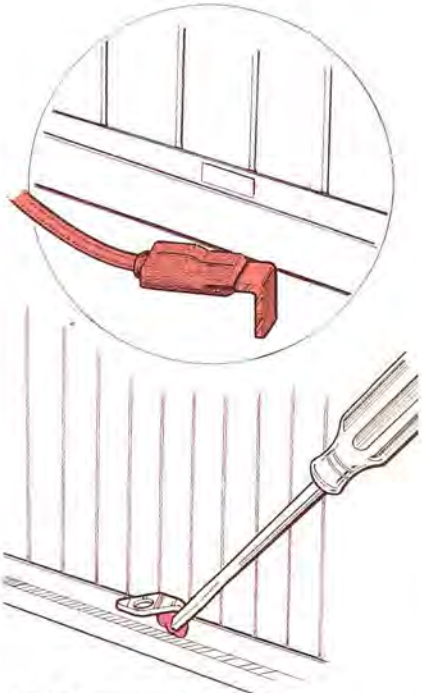

(Fig. 3) Broken-off defroster-grid terminals can be repaired with conductive epoxy adhesive.

The second reason is simpler—the grid may overheat if it's left on too long. Imagine accidentally leaving it on during a long trip on a summer day. The heat from the grid added to the heat of the sun may crack the glass or contribute to deterioration of the window's rubber gaskets.

If you need to replace the timer or switch, you'll probably have to go to the car dealer for the parts.

Easy Fixes

Suppose your grid has several lines that don't heat. You may be able to find the break by simply inspecting the silk-screened grid along its length. This will be easier on hatchbacks, minivans and SUVs because you can open the back up and look at the lines against the sky. Sedans will require you to crawl into the back and poke your head into the area above the rear

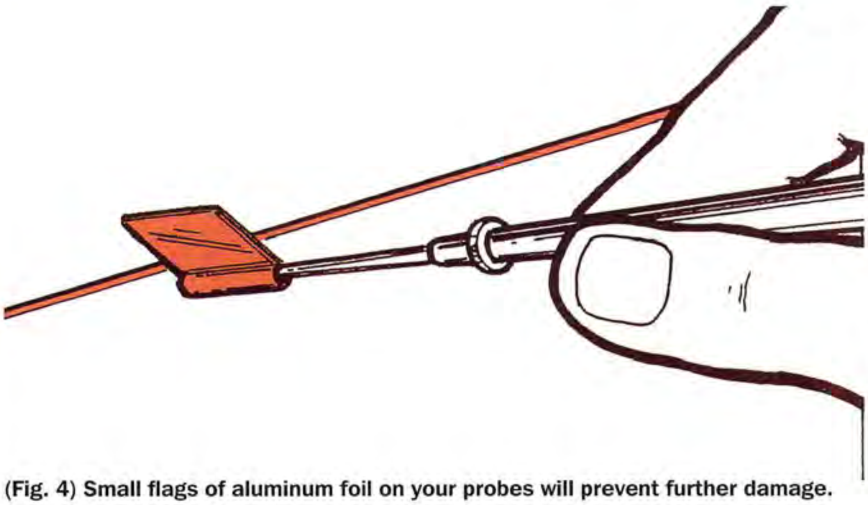

(Fig. 4) Small flags of aluminum foil on your probes will prevent further damage.

deck. Otherwise, you'll need to drag out the trusty DC voltmeter and hunt for it electrically.

Set your voltmeter for the 20-volt scale, and attach a couple of postage-stamp-size pieces of aluminum foil to the leads (Fig. 4). This will prevent the probes from scratching the grid. You can simply lay the aluminum-foil tabs on the glass and press lightly with one finger to make connection with the grid. If the window is large, it may help to have another pair of hands.

Start by measuring the voltage across the entire length of the grid. With the key on and the defroster turned on you should see 12 volts on the meter with one tab at either end of the grid. Now move one tab to the center of the grid and measure again. The voltage should read 12 volts if the break is between the tabs, less if the break is outside of the tabs. Similarly, measure the voltage on a grid line that is working properly, and you should see approximately 6 volts at the center, because you've just turned your defroster grid into a giant rheostat. By moving the tab along the damaged line, you'll see 12 volts on the meter until you reach the break, where the voltage will drop considerably all at once. This should allow you to pinpoint the break.

Repairing the grid is simple. Many auto departments in large stores, and almost any dedicated auto parts store, can sell you a repair kit. Clean the area of the break with alcohol and a fresh, untinted paper towel. Don't use window cleaner, as it may leave a residue of wax or silicone.

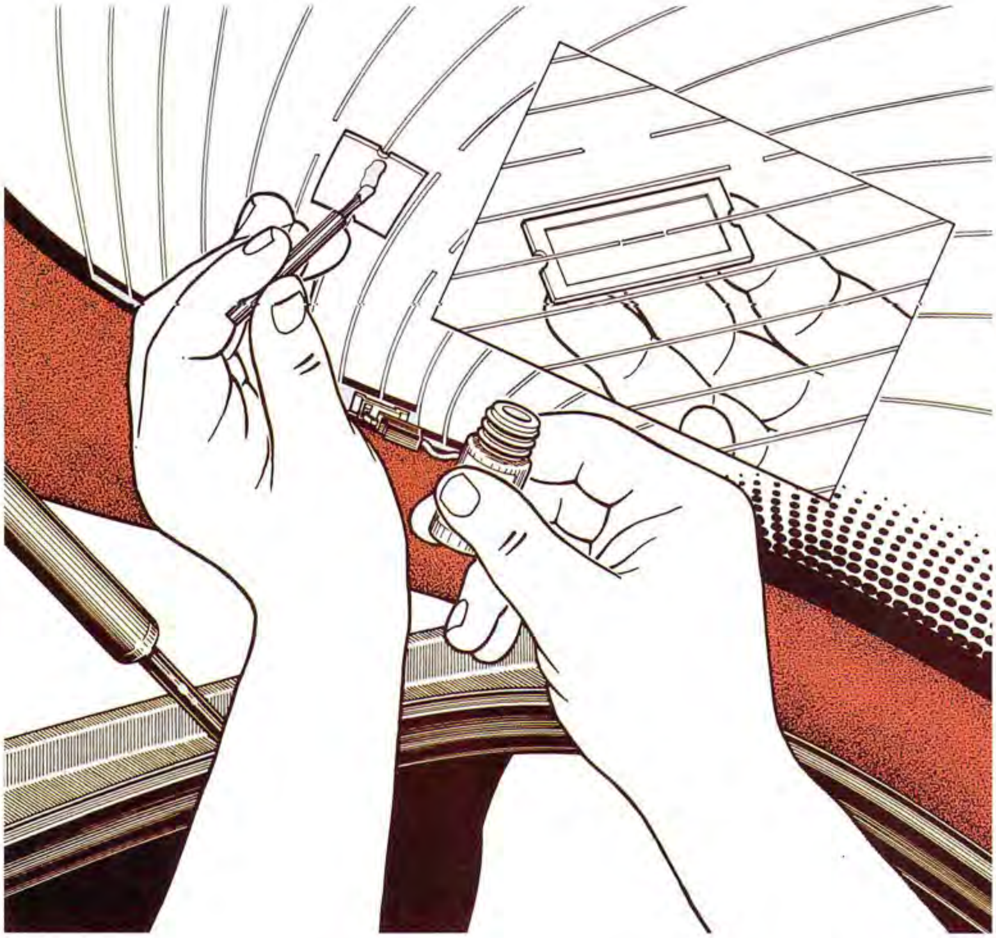

The kit will have an adhesive template to stick over the break, but you can use ordinary masking tape just as easily. The masking tape can be used to make a new line that exactly matches the width of your old grid, if the mask in the kit is too wide or narrow. Paint a stripe of the kit's conductive paint across the break (Fig. 5). Allow it to dry for 10 to 15 minutes and remove the mask.

If you have several grid lines that are damaged, simply repeat the process. If the lines are damaged in more than one place along their length, you'll have to go back to step one and find the next break.

(Fig. 5) Repairing minor breaks in the continuity of the defroster grid lines is as simple as masking them and using a special conductive paint.

Keeping Your Windshield Clear

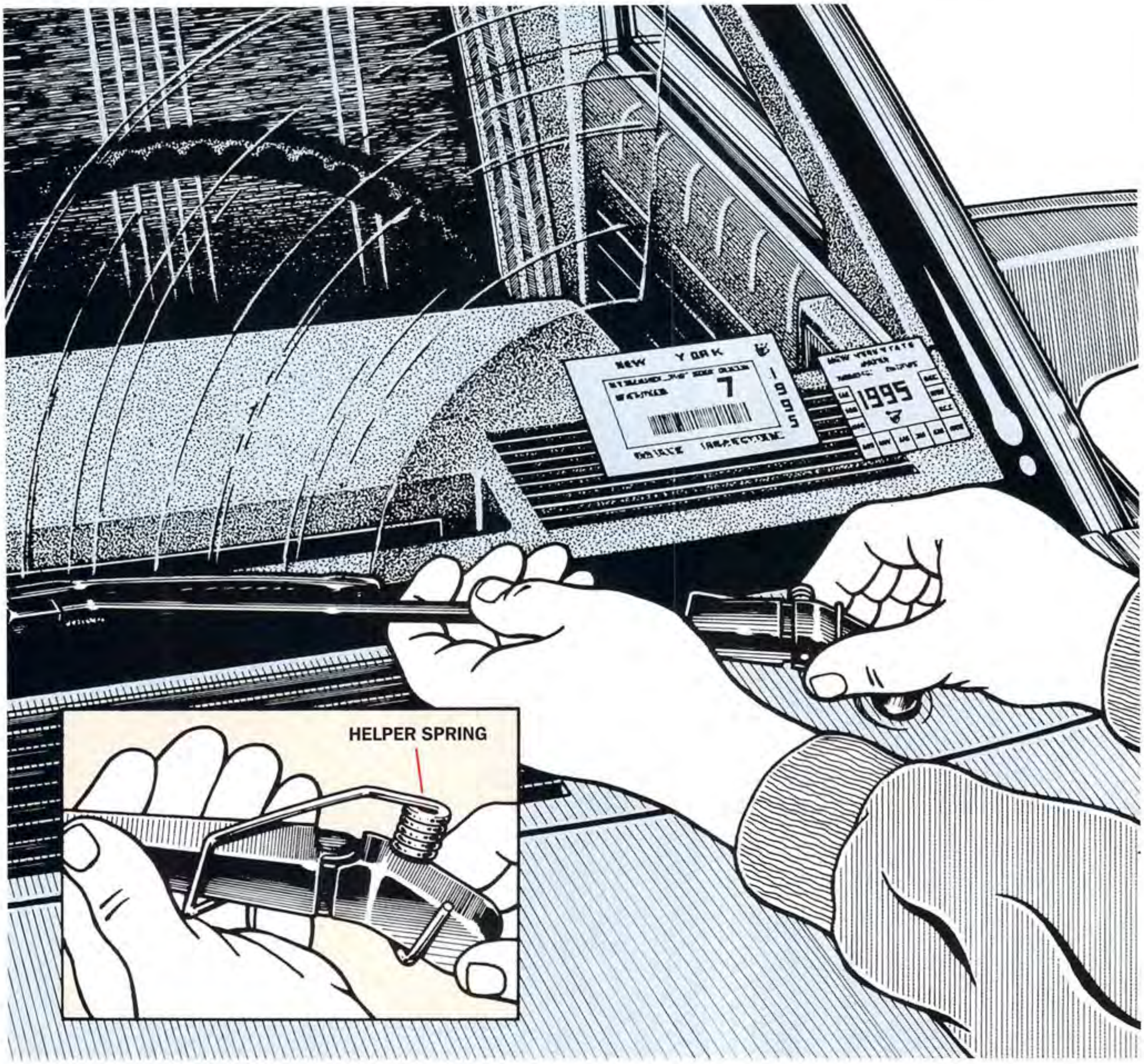

(Fig. 1) **Wimpy wiper-arm springs can be fixed by adding helper springs at the arm joint.**

It's a dark and stormy night, Part II. Every drop of rain that splatters on your windshield seems to hide another vehicle, camouflaged by the smear of light and haze coating the glass. Even with your windshield wipers set on high, you can't see more than a few feet ahead. Every few blocks you have to stop and wipe the condensation from the inside of the windshield with your hand. It's a dangerous situation.

Your car has several systems that are specifically designed to keep a clear piece of glass—and nothing else—between your eyes and the road. Outside the glass are the wiper arms, run by an electric motor. Inside is a defogging blower/heater, integral with the heating/air-conditioning system. Keep them at peak effectiveness.

Slimed

If you get your vehicle washed and waxed regularly, there may be a buildup of silicones on the windshield, layered with the airborne dirt and dust from industrial plants. Even

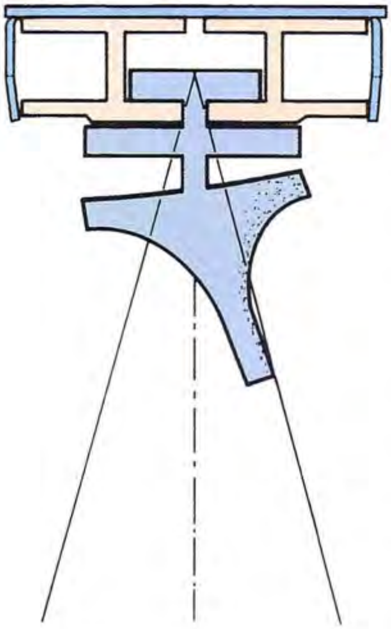

(Fig. 2) Replace the rubber squeegee in the blade if it's taken a set to one side or the other.

good wipers won't remove this kind of mess. It takes a cleaner with a super-fine abrasive that won't scratch the glass, and, although there are household cleaners in that category, auto parts stores now carry ones made specifically for windshields. When a windshield is clean, water from a garden hose should "sheet"—flow across it, not bead up.

For winter-freeze protection, fill the washer reservoir with a cleaning solution with washer antifreeze, not just a mixture of water and alcohol or ordinary antifreeze. And never use an abrasive windshield cleaner in the washer reservoir.

Check the spray pattern of the washer nozzle. Although some are nonadjustable, individual tube-type nozzles can be bent slightly to improve the spray position, and the base of the nontubular type may be shimmed. The common ball-in-socket variety can be re-aimed, using a straight pin carefully placed in the ball's orifice and used as a lever.

If the wipers chatter and the windshield is really clean, inspect the pivots of the blades. If they're corroded or bent, the blade won't conform properly to the glass. Replace it.

Still chattering? Open the hood and examine the windshield-wiper linkage and mounting. If the motor rocks in its mountings, the blades may chatter. The wiper motor is generally mounted in rubber bushings to reduce noise. Look for broken or deteriorated bushings. Next, check the linkage. Worn bushings and pins can also set up a chatter. If there's anything that can't be tightened, you'll have to begin replacing some parts.

Now check the bushings in the fittings where the wiper arms pivot as they pass through the cowl. Some of these bushings are mounted in rubber, and sometimes it's possible to snug them up. Look for a large-diameter nut below the cowl. The pivot pins

(Fig. 3) A small dab of wiper-blade lubricant can help to keep smearing down and also extend squeegee life.

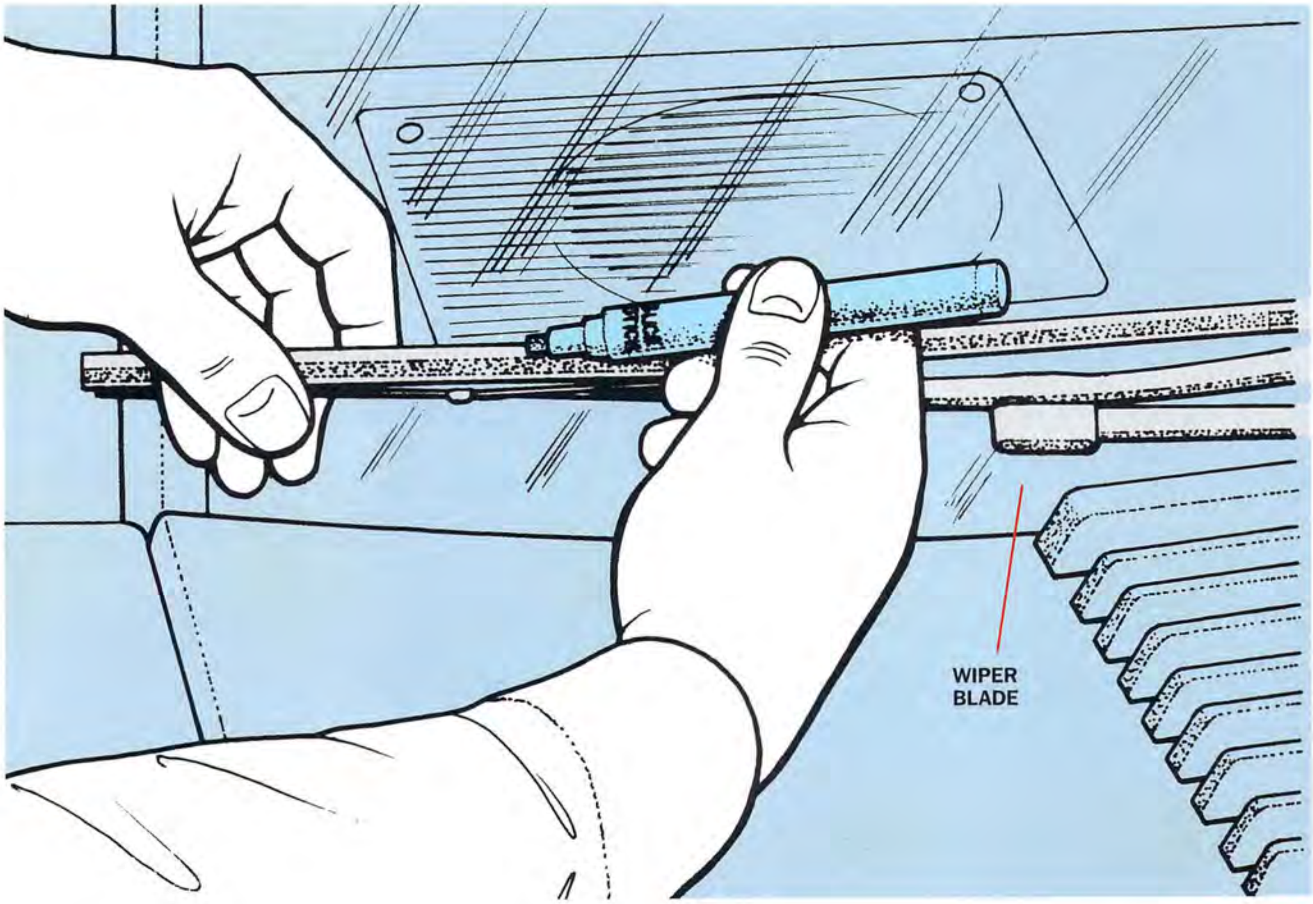

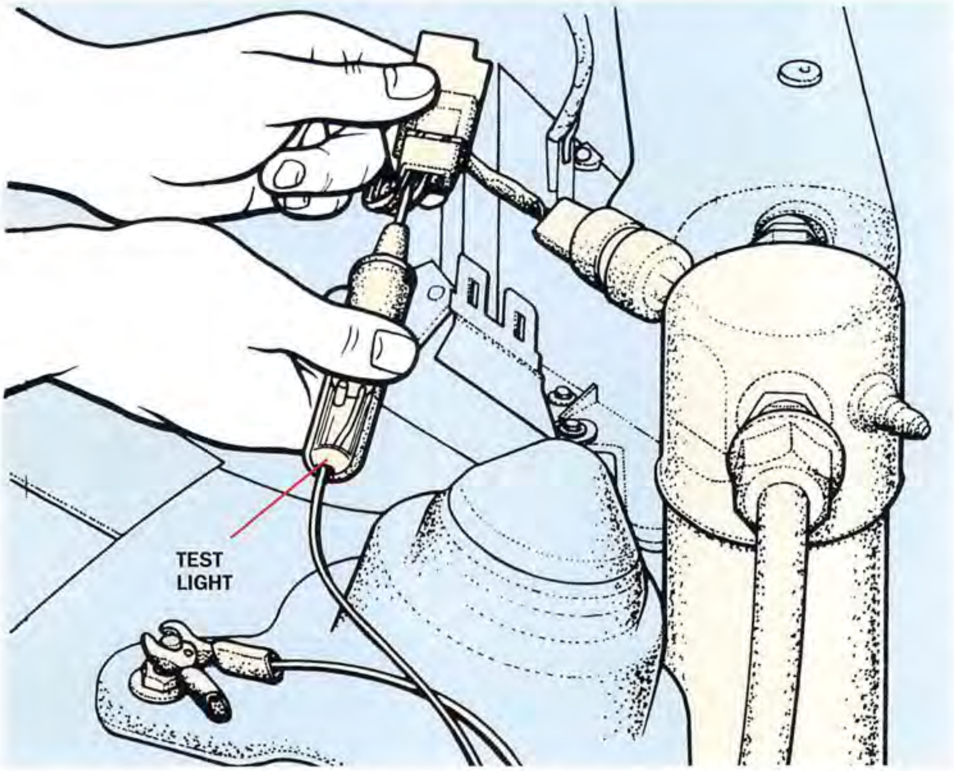

(Fig. 4) **If the blower isn't up to speed, probe the blower-motor relay wiring with a test light.**

themselves can wear, which can also permit chatter to get started. Parts replacement is in order here.

Lift the arm. If it feels weak, compare it with another car, or measure it on a fish scale. The typical arm, lifted from the blade end, should take about 30 ounces. (Note, however, that fish scales aren't much more accurate than a fisherman's estimate of the length of the big one that got away.) If the arm spring is weak, replacements could set you back more than $70. Instead, install helper springs—they're sold in auto parts stores for just a few dollars **(Fig. 1)**.

Inspect the wiper blades' rubber inserts (the squeegees) for cracks or other deterioration. It's best to remove the blade, hold it in front of you and see if the rubber squeegee tip is absolutely straight down the center of the blade. If it's wavy or tilted noticeably to one side, replace it **(Fig. 2)**. If the blade is good, only the squeegee need be replaced. If the squeegee itself is good, apply a lubricant made for natural rubber to both sides. A quick smear of lube once every week or two should not only extend the squeegee's life, but improve its everyday performance **(Fig. 3)**. Again, look in the auto parts store for rubber lubricant. Don't use silicone.

Fogged In

If the inside of the windshield has a permanent "fog," it may be caused by interior air pollution, including tobacco smoke and fumes from plastic trim. Vinegar does a good job of removing it, but be sure to wash the glass to help clear away the odor. Another option is to clean the glass with an antifog formula, which leaves an invisible coating that retards interior fogging. Unfortunately, the formulas we've seen are not recommended for tinted windows, so watch where you apply it, or simply stick with vinegar.

When you use the defroster, many cars' ventilation systems automatically turn on the a/c compressor, even in winter (temperatures above 50° F), to condense the interior moisture. If your system has a separate a/c compressor control, you may have to push the a/c button to get the desired effect. And if your control panel has a separate button or lever for outside air versus recirculated air, choose outside air—it'll help purge the interior moisture that you and your passengers release, which is a major cause of the fog that mars a normally clean windshield. This probably sounds counterintuitive—isn't there plenty of moisture coming in with the fresh air from the cowl vent? Trust us. Fresh air will evaporate more mist.

Just because you hear the blower fan doesn't mean the air is blowing against the windshield, so put your hand over the defroster outlets. If the airflow is weak, either the ventilation-control system is misbehaving, a cabin air filter is plugged, or there's a dashboard duct unplugged, causing the airflow to leak out before it gets to the windshield. If there's strong airflow from another duct when you select defrost, the control panel is probably at fault.

If the blower doesn't seem to come up to speed but is turning, there's a fair chance that the high-blower relay used on many cars is faulty. Inspect the area around the blower motor, and if there's a relay wired to it, check the connection **(Fig. 4)**. And look around for an inline fuse—wired to the relay—to inspect. It could be blown.

Still have trouble seeing what's ahead on the road? Next time, don't forget your eyeglasses.

Detailing Your Car

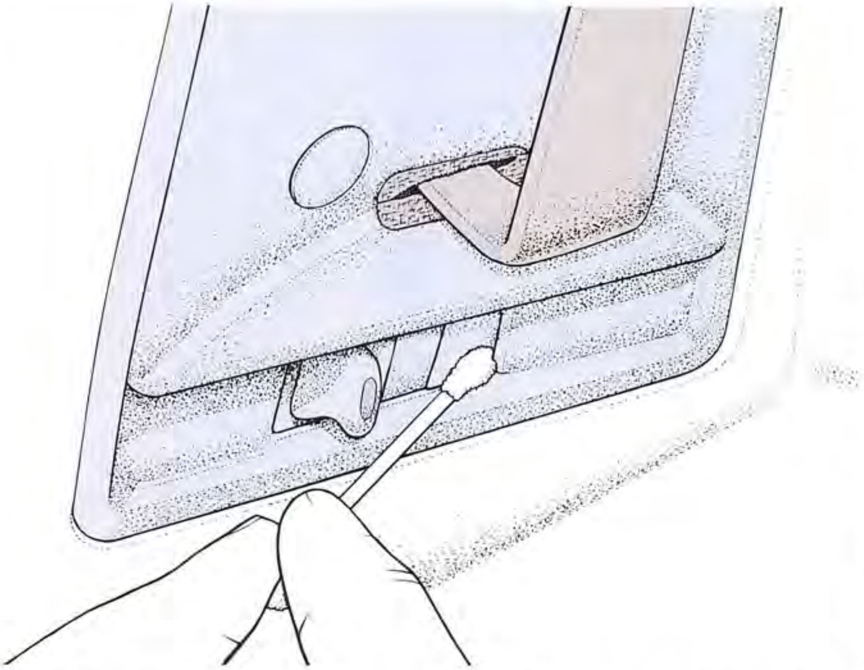

(Fig. 1) Use a Q-tip to scour interior recesses that have collected dust.

It's a Great American Ritual—washing the car. You get your bucket of soapy water and a sponge, fire up the garden hose and go to work. Your neighbor in the adjoining driveway does the same. But he's assembled a whole array of car care implements, including brushes, cotton swabs, a toothbrush and a plethora of cleaning, conditioning and dressing products. By the time you've dried off the body, slapped a coat of wax on the finish and buffed it out, your neighbor is only up to square two. So you move on to the other Great American Ritual—the backyard barbecue, leaving your neighbor elbow-deep in his bucket of soapsuds.

Later, after scarfing down your fifth hot dog and third ear of corn, you saunter over to your neighbor on the pretext of offering him a barbecued chicken thigh. But you really want to see the kind of job he's doing. His vehicle is about the same vintage and in the same condition as yours, but his car gleams like a jewel. Yours just looks clean.

What's your neighbor's secret? He went the detailing route—a process that goes beyond the ordinary wash-and-wax job. Detailing is just that—paying close attention to small details. Sure, it takes a lot more time and effort, but the results can be nothing short of eye popping.

Professional detailers have developed their own tricks of the trade for everything from vehicle washing to cleaning windows to getting ventilation grilles looking supercrisp. To get their results, use products designed for specific areas—wheels, trim, windows, etc. Name-brand products are a safe bet. Be sure to read labels to get the best finish.

Where to Start

The interior is a good starting point, so the dust and dirt you brush out won't settle on a pristine exterior. Remove any floor mats and give the carpeting and upholstery a good vacuuming. Also vacuum the dash and rear parcel shelf. Move the front seats full fore and aft to get to all the accumulated dirt and loose change. If the carpets are clean except for a minor stain or two, use a foaming cleaner to get them out. Saturate the stain with cleaner, working it in with a damp sponge. Let it sit awhile and then blot it out with paper towels or a dry cotton cloth. Repeat if necessary, and then go over the area with a damp sponge before final blotting. Don't oversaturate the carpet and risk getting mildew.

You can repair burns and holes in your carpet by cutting out the offending area with a razor blade or scissors. Then cut a similar-size piece from a hidden spot, such as underneath the seat, and cement it in place using a water-resistant adhesive. Blend in the repair by brushing the nap.

Wash the floor mats, if they're rubber, and apply a dressing that does not leave a slippery finish, for obvious reasons.

Clean interior hard surfaces with a damp cloth and a mild all-purpose cleaner such as Simple Green, diluted about 10:1. If you have vinyl-covered seats, use a conditioner made for that material. Avoid products that give a high-gloss, slippery surface, so passengers won't feel like they're on a roller coaster. If you have leather upholstery, dress the surfaces with a leather conditioner. Never use a vinyl product on leather.

Worn or torn areas of vinyl can be repaired using kits made for this purpose that are available at auto supply stores. Repairs are made with a patch that lets you match the color and grain of your upholstery. Worn areas of leather can be touched up with dyes or a high-grade shoe polish. Just make sure you match the color as closely as possible.

The dash presents a special challenge, with buttons, crevices and bezels that you can't get to with a cleaning rag. You can blast dust and dirt from these areas by using small cans of compressed air made for cleaning camera and computer equipment. Cotton swabs also work well here (Fig. 1). Pay attention to the cleaning products you use on your dash. If your dash has a flat finish, don't use a product on it that will leave you facing a shiny gloss.

Clean air vent grilles with cotton swabs and brighten them up by misting on some spray-on vinyl/rubber dressing or accent spray—just a touch. You can also use these products to cover up light scuff marks on wood trim. Spray the stuff on a soft towel and then apply it to the wood.

Clean glass or plastic gauge lenses with a glass or plastic cleaner, not wax. Pull off any removable knobs to clean the bezels underneath.

MINOR SCRATCHES can be polished off with clearer waxes or mild polish.

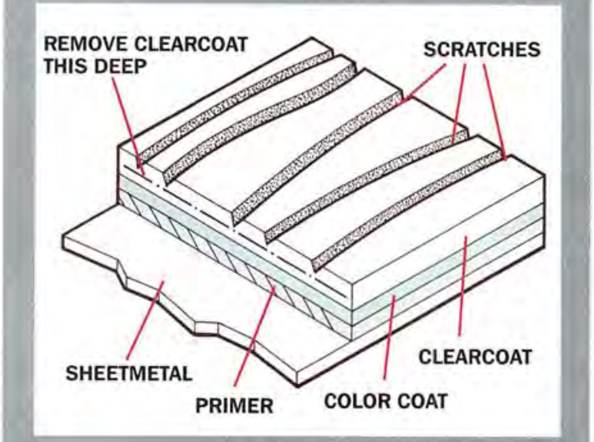

SCRATCHES THAT GO THROUGH the clearcoat into the pigment must be repaired professionally.

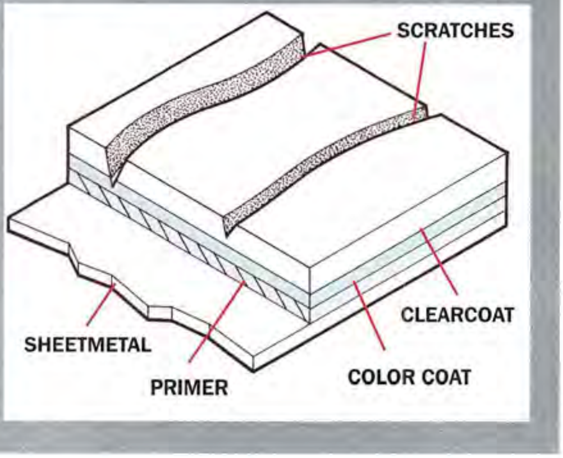

Ever wonder where the haze on the inside of your windshield comes from, since you don't smoke? It consists of plastisols given off as the plastics used in many new cars slowly cure. Not to worry—a good glass cleaner or vinegar should remove it. If your windows are really cruddy, you may have to resort to stronger measures, such as scrubbing with 4-ought steel wool.

A Word of Caution

If you have aftermarket window tint film, it may be degraded by cleaners that contain ammonia or vinegar. Factory tinting is in the glass and is not affected by these cleaners. One trick used by some detailers for the final touch on window glass is to rinse it down with seltzer and do a final wipe with a ball of crumpled newspaper.

When it comes to first impressions, nothing makes a hit like a jewel finish. But this is possible only after any paint problems are corrected. Just about all finishes today are a 2-step (color) basecoat and a protective clearcoat. The top clearcoat is only about 2-3 mils thick, and when it gets scratched or abraded it refracts light and the color coat underneath doesn't shine through clearly. It's like looking through a scratched or foggy lens.

To evaluate your paint, first wash your vehicle. Work in the shade and make sure the surface is cool. Use a carwash soap, not a household detergent, and work in sections, from the top down. The lower panels tend to accumulate more abrasive dirt. To do a final rinse, remove the spray head from the hose and flood the finish. The water will

HOW IT WORKS

Clearcoat Paint

Most modern cars use a clearcoat paint system. The pigmented layer of paint, whether a solid color, metallic or pearlescent, is covered with a layer of clear paint to provide a higher gloss and a "deeper" look. So far so good.

But the surface of any paint, even clearcoat, is vulnerable to scratches and oxidation. In the past, abrasive rubbing compound could be used to polish out these imperfections. Clearcoat is thinner, and using abrasives is trickier. Removing too much paint will leave the pigment layer exposed, necessitating a respray with more clearcoat. The recommended procedure for oxidation and haze is to use a chemical cleaner or cleaner wax. Look for products that specifically say they are appropriate for clearcoat paint on the label. Light scratches can be buffed out with very mild polishing compound. Do not use rubbing compound on a clearcoat finish. Scratches that go through the clearcoat layer cannot be polished out, although some "color in polish" products or pen-type touchups may provide a visually acceptable fix.

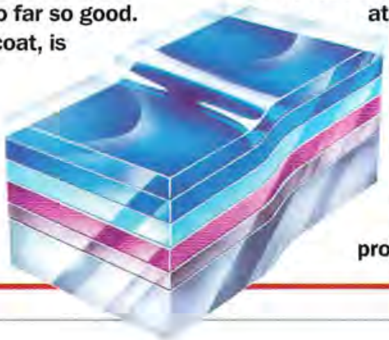

(Fig. 2) Wax your car in the shade, in small sections. A small buffer will help, but hand application will get the job done, too.

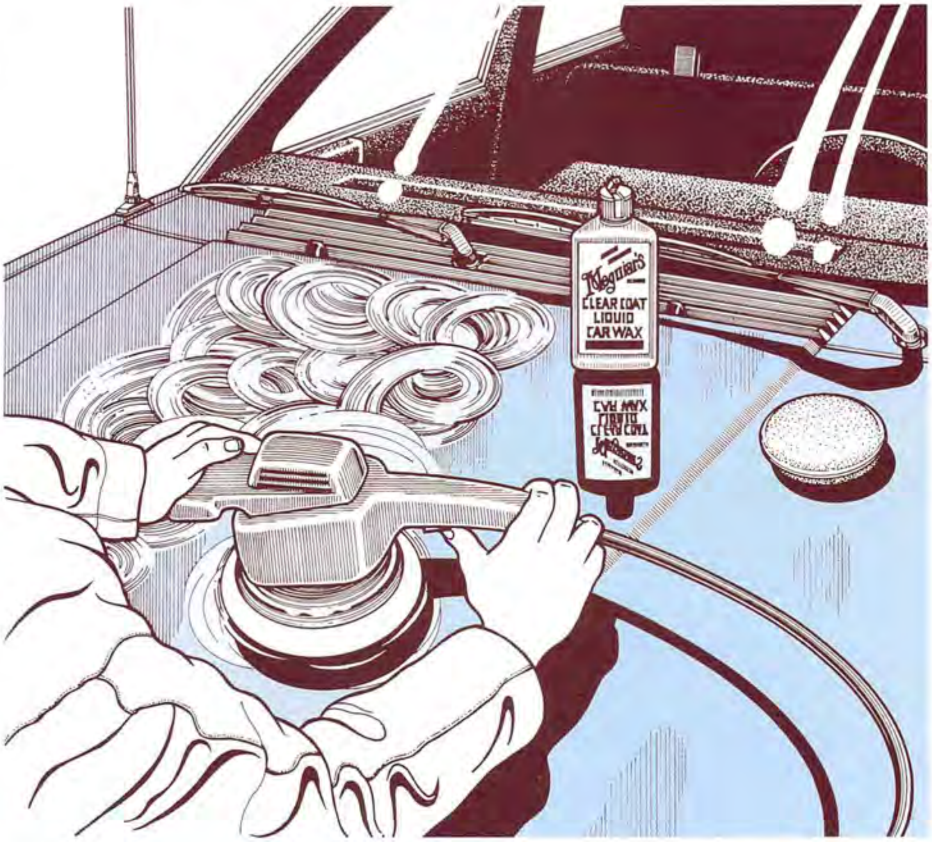

tend to run off in sheets, minimizing spotting. Dry with a good-quality chamois or a soft thick-nap terry cloth towel.

Don't forget the wheel wells. Get the crud out with an all-purpose cleaner and a good high-pressure dousing. After you've finished washing your car, apply a vinyl dressing to add some snap to the wells.

Wash the wheels (make sure they're cool) with a brush made for this purpose, but do not use acid-based cleaners on polished alloy wheels or wheels that are clearcoated. You can use these cleaners on rough-textured alloy wheels. Chrome wheels can be gleamed up with metal polish or glass cleaner.

After washing the car, inspect the paint. Stains and scratches can be attacked with a good clearcoat-safe cleaner. The worse the problem, the more aggressive the cleaner needed. Start off with the least abrasive product and gradually move to coarser cleaners as required. Then machine buff (Fig. 2).

Polishing and/or waxing is next. Be sure to include doorjambs, and the areas beneath door hinges and behind bumpers. Minor blemishes may be neutralized by wrapping a cotton cloth around your index finger and burnishing the polish into the finish. Polish not only gives the finish its gloss, but it feeds the paint with oils to prevent it from drying out. Polymers in the polish fill in minute scratches in the clearcoat layer, restoring its clarity. If you machine-buff the polish/wax to a high luster, go with an orbital rather than a rotary model, which would be more likely to burn the paint. Treat the plastic chrome on today's cars as if it were a painted surface and protect it with a light coat of wax.

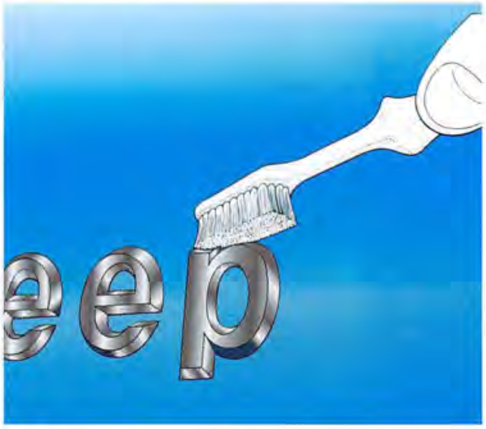

(Fig. 3) A dampened toothbrush can remove wax deposits from exterior trim.

Avoid getting wax or polish on rubber and flat black plastic areas (clean them with a nongloss product), door handles and emblems. If you do get a wax stain on rubber trim, spray it with a mist-and-wipe product and wipe it down with a terry cloth towel. If that doesn't do the trick, this usually works: Microwave some peanut butter and apply it to the stain with a soft toothbrush. Peanut butter's oils dissolve the wax and it's abrasive enough to lift the stain (but it can stick to the roof of your car).

If you get a polish/wax residue around emblems or in crevices, break out the cotton swabs and toothbrushes (Fig. 3). It's important that you first wet the area with a mist-and-wipe product such as Meguiar's Quick Detailer. Never brush on a dry surface.

Moving underhood, protect electronic components by wrapping them in plastic. Then spray on a diluted all-purpose cleaner, hosing it off with light water pressure. Vinyl/rubber protectant will dress up nonmetal areas. Let it soak in if you like the glossy look, or wipe it on and off for a more matte finish.

All that's left now are the tires. Clean them first—whitewall tire cleaner works even on blackwalls—and then apply tire dressing. Here again, to get a gloss finish let the product soak in, or for a matte look wipe it on and off with a cotton cloth. Be sure the tires are dry before driving off, or you'll spatter your nice shiny finish. And maybe even your neighbor.

Aligning Body Panels

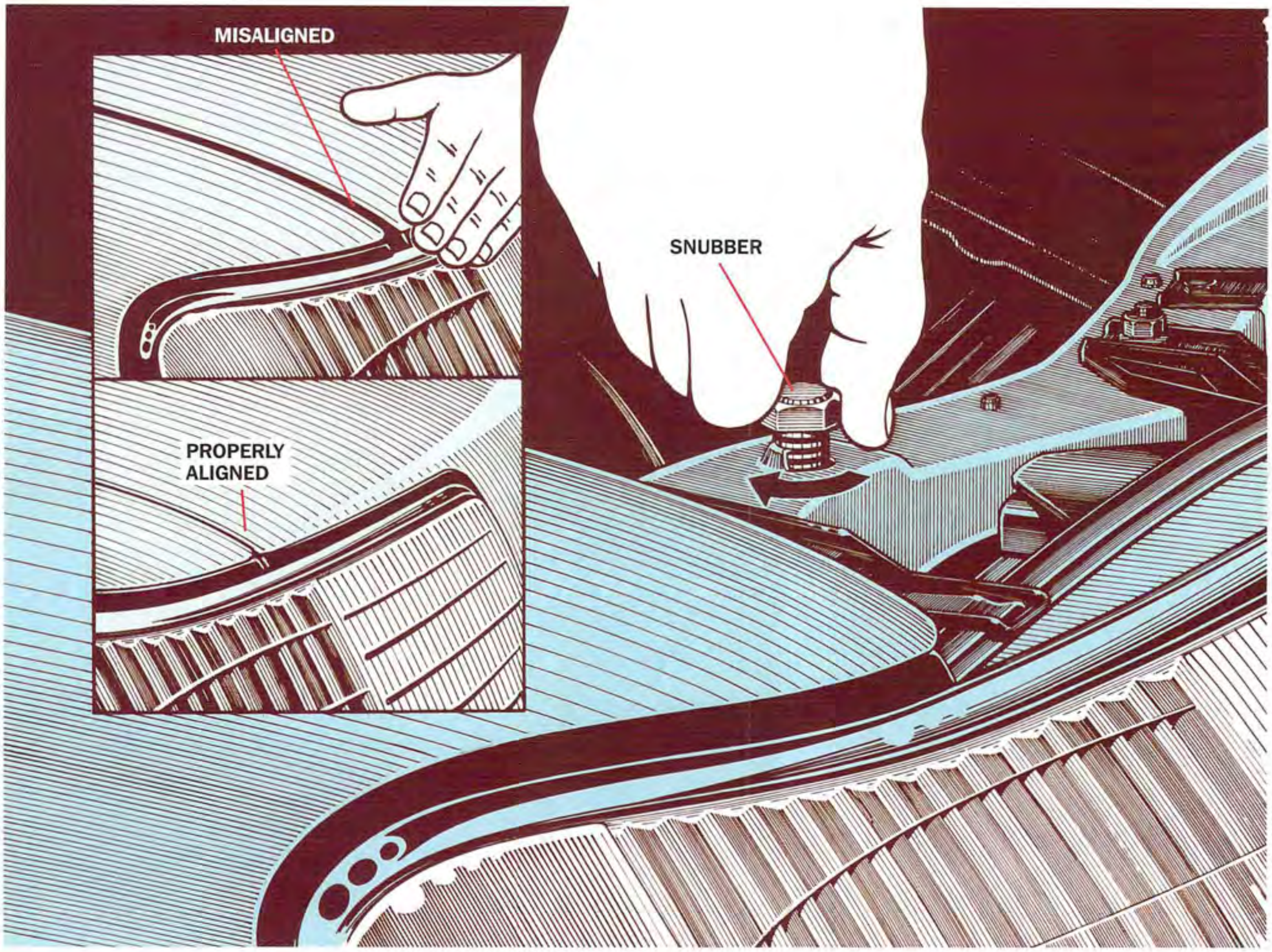

(Fig. 1) Raising or lowering the rubber snubber will raise or lower the corner of a misaligned hood to produce a better fit.

Unpacking the car after a vacation weekend is never fun. Less fun, even, than cramming the family's idea of "just enough clothes for a summer's weekend at the lake" into a modern compact car's diminutive trunk. Were those ski parkas really necessary?

At long last, you head cheerily to the office on Monday morning, glad to be back to the grindstone. At least until you notice, by the cold light of day, the trunklid of your car sitting a good half-inch askew. Must be something still left in the trunk, preventing the lid from closing fully, you say. The trunk, however, is as empty as your wallet. The trunklid is sprung.

Nothing Is Forever

Body panels on cars are made of relatively thin sheetmetal. In the aggregate, they are immensely strong. The panels are particularly strong in protecting the occupants against frontal crashes, telescoping progressively while absorbing energy during impacts. But forcing a trunklid closed against resistance, or slamming the hinge side of a car door on the tail of a ski or even the corner of a duffel bag, can distort the hinges and latches in seconds. Years of driving on pothole-filled streets and countless open-close cycles can make a door that sounded like a bank vault when the car was new now sound like a Cheez Whiz can rolling down a cobblestone hill. Just as the Leaning Tower of Pisa continues its slide toward the courtyard surrounding it, a car's doors, hood and trunk panels can gradually change position, and may need periodic adjustments.

Fortunately, most panel misalignments aren't the result of a serious fender bender. Those usually require the services of a trained body technician and a large hydraulic frame-straightening jig. Simple mis-

(Fig. 2) Hoods and trunks may need to have their latch pins moved up or down to make them close properly.

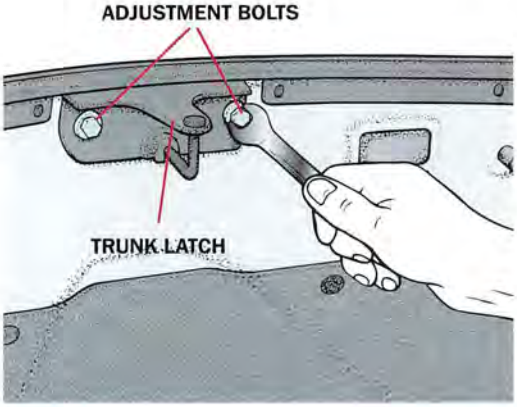

alignments, however, often can be minimized with little more than simple hand tools and a careful eye.

Hoods and Trunks

The simplest misalignment, and the one most often seen while cruising parking lots, is a vertical misalignment of the corner of a hood or trunk. The corner of these opening panels is normally held at the correct height, even with the lip of the fender that abuts it, by a simple rubber snubber. This snubber is usually threaded into a hole in a corner of the panel. Rudimentary threads in the sheetmetal allow the snubber to be screwed in and out by hand. On some vehicles, the snubber is in the hood or trunk panel, sometimes in the bodywork below the panel. Simply screw the snubber in or out until the offending panel is flush (Fig. 1). Some older cars may use a snubber with a threaded rod and jam nut on a steel stud. You'll need to loosen the jam nut, adjust the snubber, and then tighten the jam nut with a wrench to keep it from shifting. Generally, a half-turn or so is sufficient.

More severe misalignment may make the panel's latch either too loose or too tight to open and close properly. Loose latches will rattle. Latches that are too tight will break fingernails, and will eventually break the cable to the interior handle. Fortunately, most hood and trunk latches are adjustable. You may need nothing more than an open-end wrench, or perhaps an Allen wrench (Fig. 2). Sometimes the latch assembly itself is adjustable. Occasionally, the catch on the mating panel can be adjusted instead of the latch assembly. Loosen the appropriate hardware slightly, and tap the assembly just enough to move it almost as far as you moved the snubber. Tighten the hardware, and try closing the hood or trunk. The latch release should pop easily, and the latch shouldn't rattle. Try for about ¼ in. of compression in the snubber on a hood. On a trunk you are likely to find rubber gaskets to compress, and there may be more resistance before the latch pawl snaps home.

If you have an electrically tripped latch, tighten things up slowly. If the latch is too tight, the solenoid may not have the strength to pop the latch open. If you've painted yourself into a corner and can't open the trunk, try having someone else hit the trunk release while you push down on the trunklid. This should release enough pressure on the pawl to let it move.

Fore and Aft

Maybe the hood or trunklid is aligned vertically, but the panel is sitting crooked, or is just too close to the window glass. You'll need to adjust the hinges. Use a felt-tip pen or a pencil to mark the location of the hinges on the inside of the panel. Loosen the attaching hardware just enough to be able to slide the panel to a new position. Do only one side at a time (Fig. 3). Tighten a little, then close the panel and check its alignment. It'll probably take a few iterations, but you should be able to get the panel to center in its intended

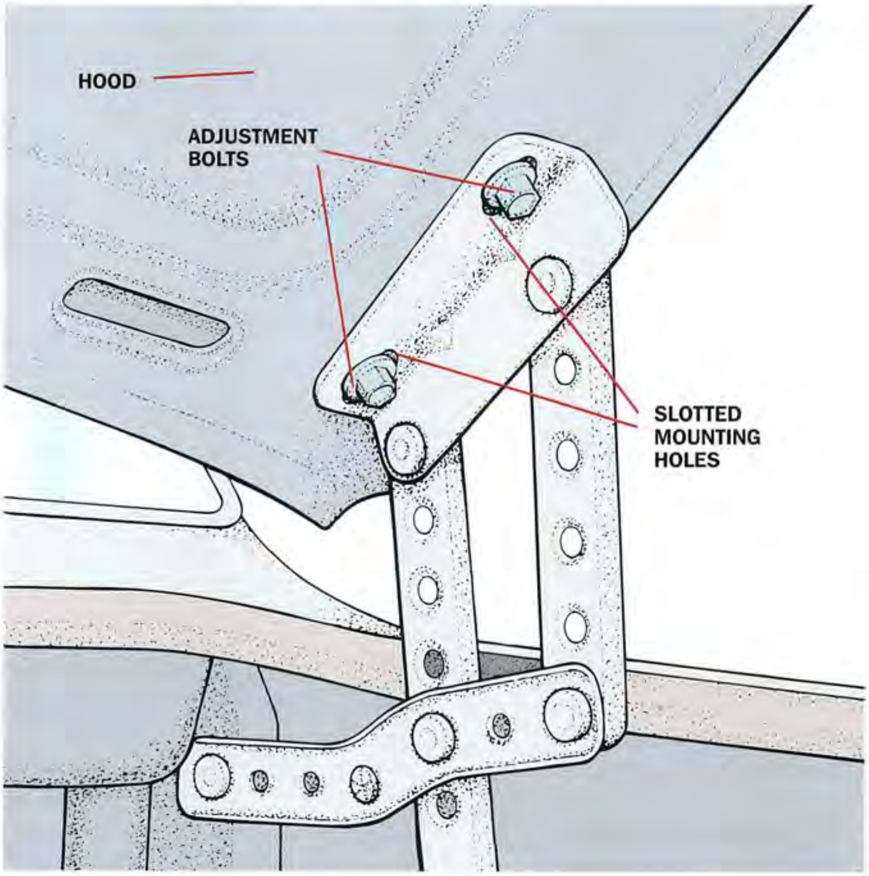

(Fig. 3) Adjust the hood or trunk one side at a time by loosening the attachment bolts and sliding the panel back and forth.

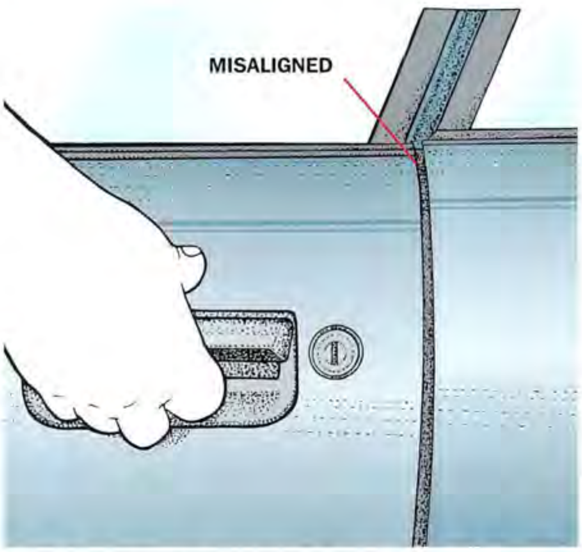

(Fig. 4) If the rear of your car's door sags, the cause may be a door latch that needs to be repositioned, or worn hinge pins.

space. After all is said and done, it may not be possible to get the gaps in the panel to be symmetrical. It's not unusual for the gaps to change their width from the front to the rear of the panel. You may need to strike a compromise between getting the panel to sit square in its intended recess both fore and aft, and making all the gaps symmetrical.

If the gaps are square and the panel is sitting so that it's symmetrical, but the gap is wider on one side than the other, there is a fix. With the adjusting hardware good and tight and the panel adjusted correctly fore and aft, open the hood or trunk fully. With both hands, give it a sharp shove in the correct direction. Close the panel and check. If it still sits in exactly the same position, repeat and shove harder. If it has moved partway back to the correct alignment, shove again. Repeat until done. If you shove too hard, just shove back in the other direction. The first time you try this, close the panel carefully, so you don't scratch the paint if you've moved the panel too far.

That Swinging Door

Fitting a door is more complex. If it's out of alignment when closed and latched, you need to adjust the latch (Fig. 4). Sometimes the latch mechanism on the door is the adjustable piece, sometimes you need to adjust the latch pin on the door pillar that the latch engages. It's a simple matter of slightly loosening the attachment hardware, lightly tapping the latch or pin to an approximation of the correct position, tightening, and trying again (Fig. 5). You can move the door's rear edge in or out and up and down with this procedure.

Older cars can have hardware that's difficult to budge. Hardware that uses hexhead bolts or Torx bolts usually yields to penetrating oil, some moderate tapping with a hammer and the prerequisite profanity. Phillips-head attachment bolts require more finesse. Before you round off the heads be sure you're using a proper-size (usually No. 1 or No. 0) Phillips screwdriver. Some screwdrivers have a flat shank or hex-shaped section on which you can put a wrench. Then you can use one hand to push, hard, in on the handle to seat the screwdriver blade while you gingerly rotate it with your other hand on the wrench. As a last resort, use a handheld hammer-driven impact driver. Just remember you're hammering on sheet-metal parts and excess force can bend something in an unfortunate direction.

(Fig. 5) Adjust the door latch mechanism or latch pin to raise or lower the door, or to move it in or out, for a correct fit to the fender.

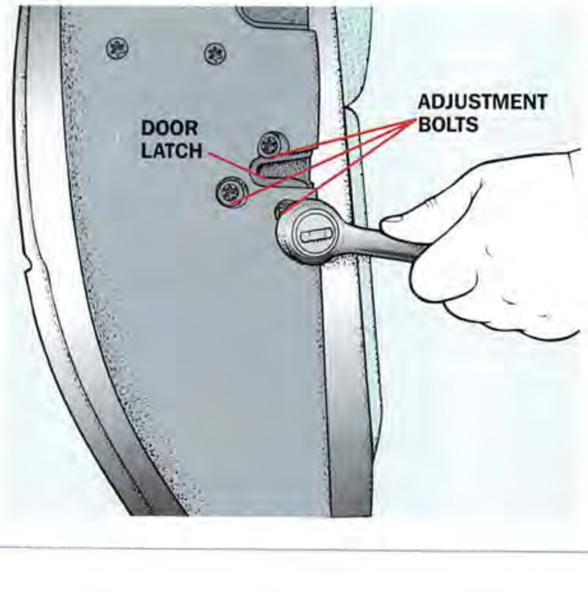

Occasionally, the hinges will fail to keep the front edge of a door in the correct alignment, sometimes because a 9-year-old has been playing at opening and closing the door while sitting in the window frame. Maybe you've managed to close the door on the wooden handle of a leaf rake. Or the hinges could simply be worn.

If the leading edge of the door is too far in or out, you can loosen the hinge bolts and move the hinges in or out one at a time. Don't loosen both hinges together unless the front of the door is too high or low. Generally, this is caused by worn hinges. You can compensate for a small amount of wear by raising the hinges on the pillar. You'll need a helper to hold the door vertically while you tighten the hinge bolts.

Serious wear calls for new hinges. If the hinges are welded to the door, your options are to replace the door or try to find an extremely competent body shop to weld in new hinges.

Another potential problem is a door that sags so badly while it's open that it has to be lifted to engage the latch—which then holds it correctly shut. This is probably caused by a rusted-out A-pillar or very worn hinge pins. A quick short-term fix is to wad up a shop towel and stuff it into the lower hinge. Slowly close the door on the rag until you meet some resistance. This will spread the hinge apart, raising the door's rear edge. Go slowly, and sneak up on the correct adjustment. Seriously worn hinge pins should be replaced.

Prepping Your Car For Paint

Junior has asked to borrow the keys for his big date on Friday. So you reach into your pocket for the keys to your pickup, traditionally his weapon of choice for cruising with the big guys. But this time he demurs—Betty Sue thinks the truck is too nasty-looking, and he'd prefer to use Mom's car.

After the shock of rejection wears off, you stroll out and examine the pickup. After all, when a teenager thinks your truck needs cosmetic improvement, it's time to pull off the blinders and think about a little bodywork.

Painting Is For Pros

Prep work aside, automotive painting couldn't be easier. With a little experience, spraying an entire car should only take about 20 minutes.

All you need is an air compressor and a spray gun. Oh yeah, you'll also need a $30,000 downdraft spray booth (with baking cycle), a supplied-air respirator and a real knack for smoothly flowing on paint without runs, sags or orange peel.

There's the rub. You may have a compressor and spray gun, but even large body shops can barely afford the other stuff. And when you consider that chain refinishers such as Earl Scheib and Maaco can do an adequate job of painting your current ride for $150 to $300, it makes sense to leave paint work to the pros.

Besides, for absolutely optimum results, 90% of a good paint job is not in the painting anyway, it's in the pre-painting preparation. And you can do that yourself.

As you might guess, budget paint jobs don't include a whole lot of prep work. For $149.95, Earl Scheib's preparation includes machine sanding, chemically cleaning the body, spot priming areas that need it and masking over things that aren't supposed to be painted—basically the windows. At some cut-rate shops, chrome and badges are negotiable. Any necessary bodywork costs extra.

Doing your own light bodywork such as fixing small dents and door dings or repairing rusty areas can save you a lot of money and help ensure the quality of the finished job. You can also greatly improve the quality of a budget paint job by removing and sanding under items that the painters would normally mask over, such as badges and assorted trim (Fig. 1). But don't start sanding and dismantling your car first. A body shop will be much more receptive to your plan if you let them know what you're going to do ahead of time.

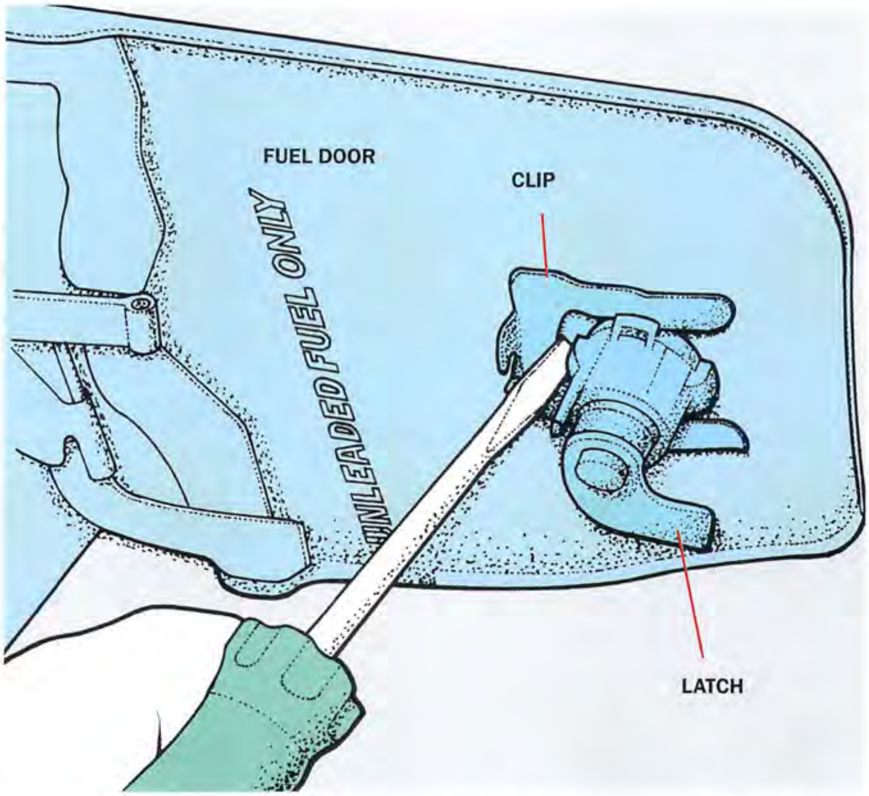

(Fig. 1) Remove as much of the hardware and trim as you can, instead of masking.

Paint Prep Primer

Unless you really hate the color of your vehicle, you'll save yourself work and aggravation if you plan to keep it the same color. Among other things, you can avoid the hours of prep work and extra expense of painting doorjambs and the underside of the trunklid and hood. You also avoid the chance of unpainted spots rearing themselves unexpectedly. A paint match costs extra over the few standard colors of the "basic bargain" price. It's worth the investment.

Start your paint prep by taking your vehicle to a do-it-yourself car wash. Pressure washing the engine, doorjambs, wheel wells and other under-body areas will help to ensure a dust-free paint job later. Cleaning the vehicle's outside will prevent you from sanding in paint-finish-ruining dirt and grime.

Nothing makes paint stick better or

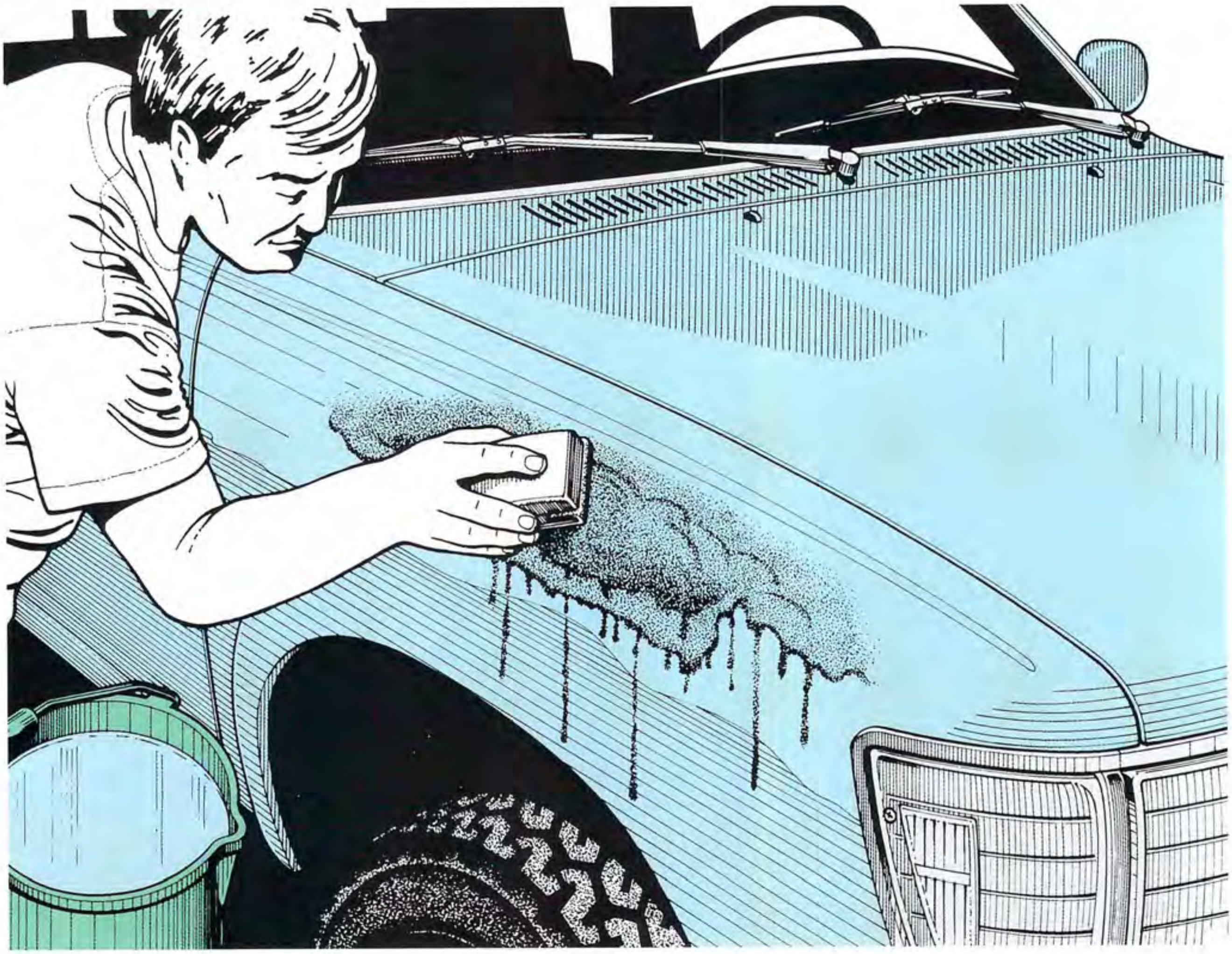

(Fig. 2) The most important thing you can do to make sure the paint shop's paint will adhere is to wet-sand 100% of the existing finish with 400- or 600-grit waterproof sandpaper. Wash the car first, and mask off or remove any chrome or trim.

last longer than a thorough sanding (Fig. 2). Paint needs a microscopically rough, craggy surface to latch on to—professional painters call this effect "tooth"—or the new paint will eventually loosen and fall off the old paint. Sanding large painted surfaces is easy. Problems crop up around areas like badges, antennas, bumpers and door mirrors. That's because no matter how careful you are, it's impossible to remove all traces of shine where the part meets the body. And paint that is applied next to a part instead of underneath causes a paint ledge to form, where dirt, water and ice start prying away. Sooner or later, the paint comes loose. After that, paint-peel is just a car wash away.

The only effective way to keep paint from peeling is to remove each part and sand under it. To have a body shop remove all these pieces can cost hundreds of dollars in labor—which is why the discount shops mask off all those parts (maybe). You can remove them yourself, however, and make a discount paint job look like a top-dollar custom paint job and have the paint finish last years longer. It's not practical to remove the windshield, rear window and door glass. But you can remove items that are common starting points for peeling problems, such as lock cylinders, door handles, luggage racks, radio antennas and side mirrors.

Take It All Off

To remove these parts, it's best to consult your vehicle's factory shop manual, but here are some general procedures. Make sure the window is rolled up, then remove interior door hardware such as the armrest, window crank and inner handle, then pry (or unscrew) the inner panel off the door. Lock cylinders are usually held in place with a spring clip. Door handles and mirrors are usually held by small bolts. The radio antenna, hood emblems and body badges are other places where paint can peel. Usually, it's not necessary to completely remove the antenna. Simply loosen the top retaining nut that holds the antenna in the body, mask the antenna shaft and lower the antenna into the fender or quarter panel. Hood emblems unbolt from under the hood. On older vehicles, badges and letters are mounted through holes in the body and fastened with spring nuts. On newer vehicles, these parts are often fastened with double-sided

(Fig. 3) Use a body hammer and dolly to smooth out minor dents. Take your time.

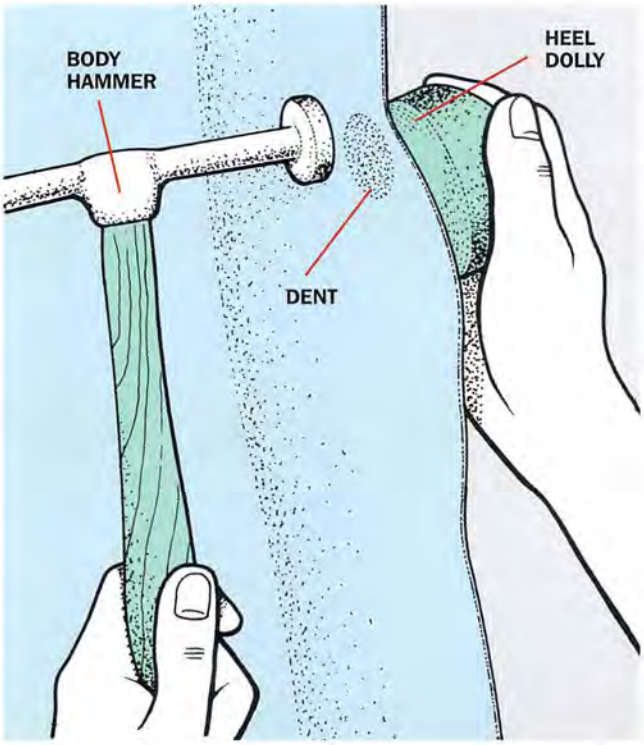

foam tape and are easily pried off with a 1-in. putty knife. If necessary, some time under a heat gun will help convince the adhesive to let go.

Of course, you'll also want to remove all large brightwork, such as chrome bumpers, the grille, headlight doors and taillight bezels. Once everything is removed, dents and dings can be repaired.

But before you start, make an honest assessment of your abilities. Your vehicle is going to end up at a body shop, right? It might make sense to leave those big dents and rust holes for the pros.

Smooth 'n' Fill

For smaller dents that you can reach from behind, use a body hammer and dolly to tap the dent out (Fig. 3). To avoid over-stretching the metal, start at the outer edges first and work to the center. Pound down any high spots in the dent, then use a 36-grit sanding disc in a portable drill to remove all traces of paint and primer, and to also help prepare the surface for plastic filler (Fig. 4).

Mix the filler and hardener according to the instructions on the can, then, working in one direction, apply the filler to low areas using a plastic squeegee. Plastic filler hardens in two stages. First, it hardens to a consistency approximating that of cheddar cheese and remains that way for a few minutes—exactly how long depends on how much hardener you've added and the ambient temperature and humidity. During this critical time, you can use a perforated Surform file—commonly called a "cheese grater" by the pros—to file away large portions of the filler until it's level with the surrounding area.

After filing, the filler will cure to its full hardness and it can be machine sanded smooth with 100-grit paper. Check the contour of the repaired area with your hand. Gently tap down high spots in the filler with the hammer and then refill these areas, file and sand again. Now the area is ready for primer. In addition to providing a surface for paint to adhere to, primer allows you to build up the area so it can be finish-sanded smooth.

Don't waste your time using lacquer primer from a spray can. These primers are very heavily thinned so the paint can easily pass through the can's miniature nozzles. The result is that primer buildup is minimal. Two-part polyester primer-surfacers, such as Marson's Poly-Fill, are by far the easiest to

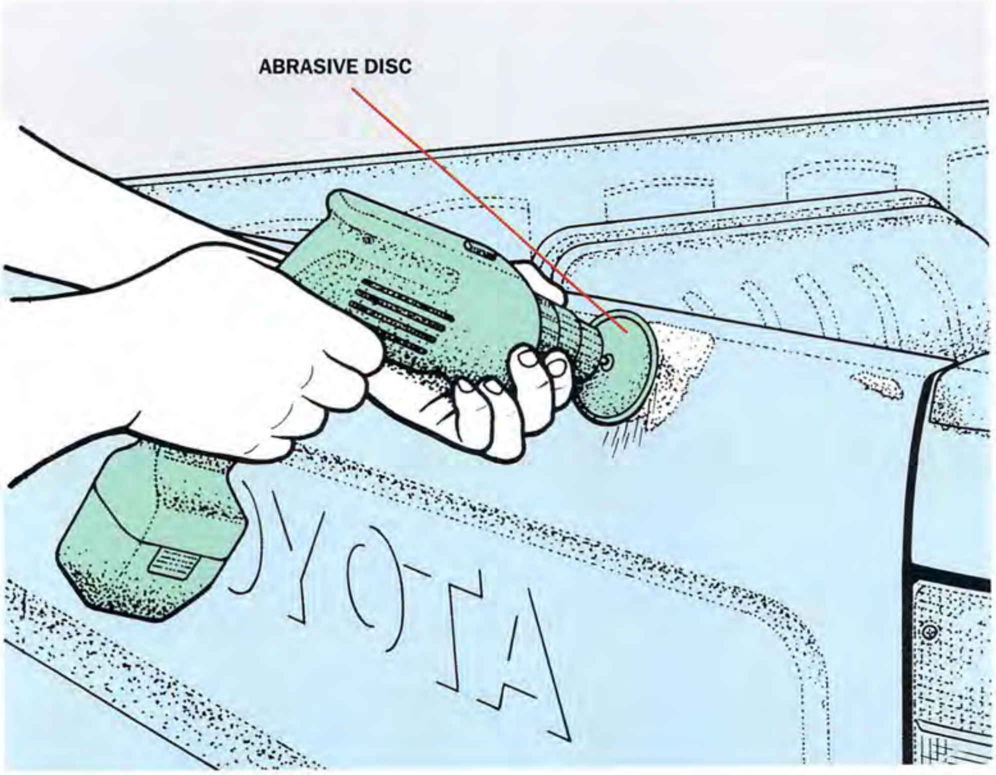

(Fig. 4) Grind all the paint and rust out of dents and creases before applying body filler.

(Fig.5) Prime and resand the areas you've worked on, then sand again before painting.

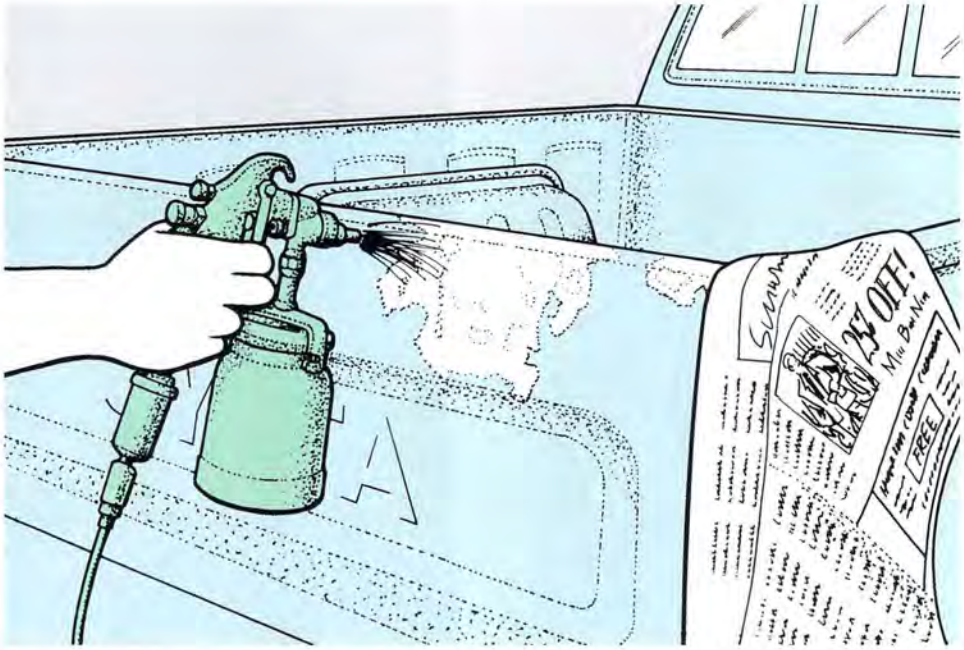

use and give the best surface buildup. Mix the hardener and primer according to the directions, then spray on two or three coats (Fig. 5). Once the primer hardens, machine-sand the primer with 180-grit sandpaper to remove heavy scratches and small waves in the body. Then wet-sand the repair with 400-grit paper.

Sand, Sand, Sand

After all dents and dings are repaired, the whole body should be sanded. The object here is exactly the opposite of what you normally try to do to your car's finish—you want to remove every trace of shine from the body. To do that, wet-sand the entire vehicle using 400-grit waterproof sandpaper. Flood the area with plenty of water while you sand, and don't forget to sand all the areas from where you removed parts. Feather-edge nicks and scratches, paying particular attention to chipped areas around door edges and rocker panels.

Once there's no more shine on the body, wash the vehicle with warm soapy water, rinse it and let it dry. You'll be amazed at how much shine there still is.

Before you hit these areas with the sandpaper again, spot prime any feather-edged areas as well as any areas where you've sanded through to the metal. After the primer has hardened, wet-sand these areas and the leftover shiny areas with 400-grit paper, then repeat the wash and rinse.

Still have some shiny spots? Guess what? Do it again!

Getting to the Shop

While the body shop can mask the large areas that shouldn't be painted, like the windshield and the rear window, it pays to run the first layer of masking tape yourself around areas that are immediately adjacent to the body, like the trim around the windshield and rear window, since you can take the time to be extra careful (Fig. 6).

Since a well-prepped car is devoid of just about everything that makes it legal for driving on the street, such as mirrors, headlights and taillights, you should seriously consider having your car towed to the shop.

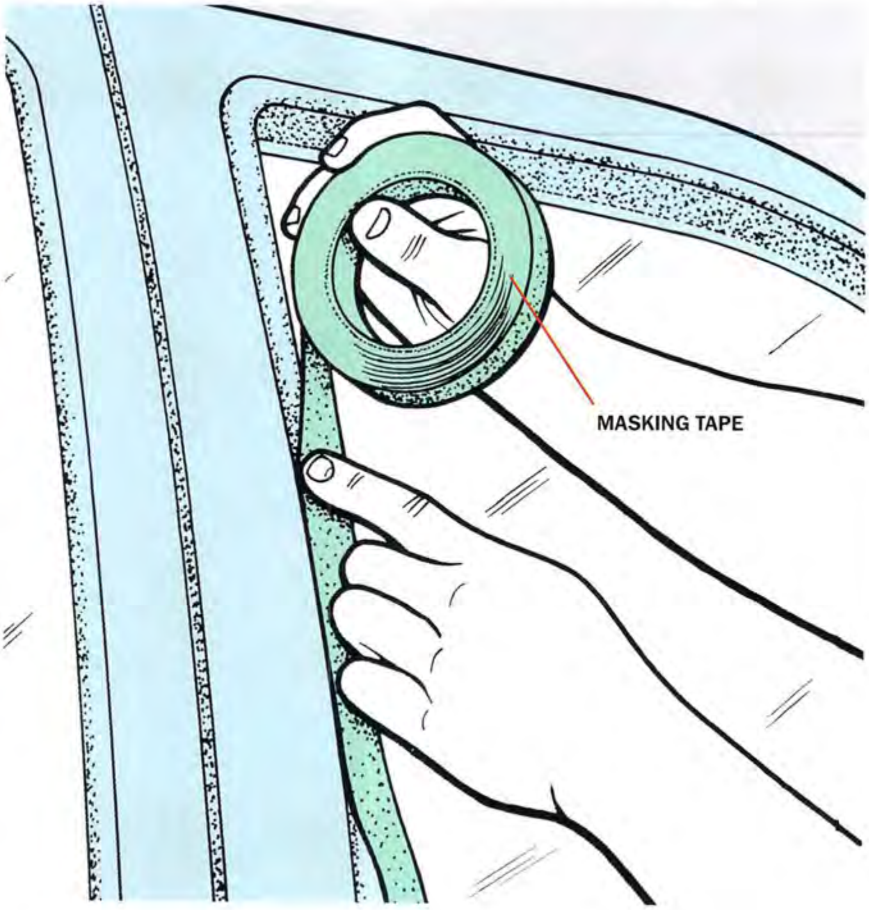

(Fig. 6) Use fresh masking tape to carefully mask all the body trim on the entire car.

EASY

Repairing Windshields

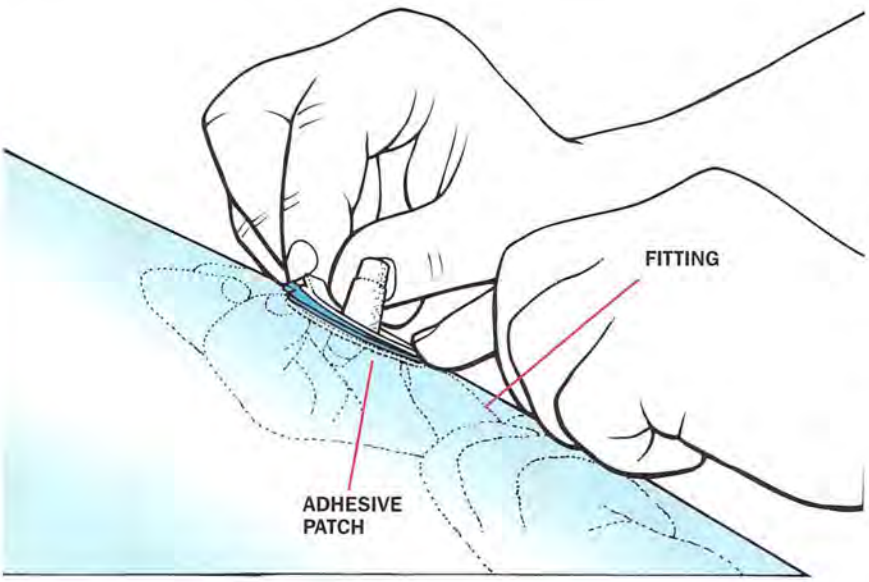

(Fig. 2) **Take care to burnish the adhesive patch to a perfectly clean windshield and adapter, or you'll make a mess later.**

It's a fine day. The sun is shining, traffic is moving along smartly and you haven't a care in the world. Whack! Until now, that is. Now there's a rosebud the size of a quarter in the windshield. What's worse, it's directly in your line of sight. It must have been a hypervelocity railgun pellet fired at you by an Imperial Stormtrooper, because you didn't see it coming or going. And as your heartbeat returns to normal, the awful truth sinks in: You're going to have to have the windshield replaced. This means dealing with the glass shop, being without your car for a day or two, having a potentially leaky windshield and, worst of all, higher insurance premiums.

Actually, it's worse. Some insurance policies won't even cover chipped glass. Maybe if you just raised or lowered your seat an inch so you didn't have to look right through the chip....

Better Living through Chemistry

Before you panic, drive home and get out your magnifying glass. Take a really close look at your new chip. It just might be possible to repair the chip instead of replacing the entire windshield (Fig. 1). The technique is to inject an epoxy or acrylic adhesive or filler into the chip.

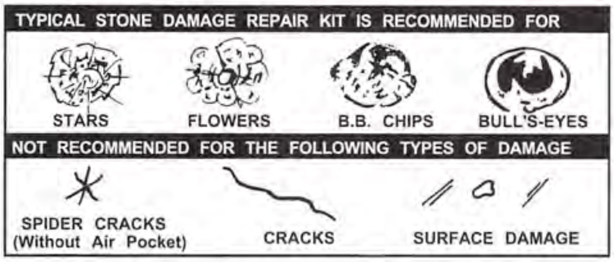
TYPICAL STONE DAMAGE REPAIR KIT IS RECOMMENDED FOR
STARS
FLOWERS
B.B. CHIPS
BULL'S-EYES
NOT RECOMMENDED FOR THE FOLLOWING TYPES OF DAMAGE
SPIDER CRACKS (Without Air Pocket)
CRACKS
SURFACE DAMAGE

(Fig. 1) **Many chips will be repaired almost completely by adhesive injection kits. But some damage requires complete windshield replacement.**

Even if your chip isn't in your direct line of sight, it's a good idea to try and repair it. Water will find its way into the chip, pulled in by surface tension. If the chip goes all the way through the top lamination, any moisture that gets that deep can delaminate the glass from the center membrane. Eventually, the membrane will fog, causing a larger blemish. Water also can freeze in the chip, causing a larger flaw or even a crack. Also, water can carry dirt into the crack—and there's no way to flush it out.

As you can surmise, it's best to do the repair as soon as possible, assuming that it's repairable. Remember that not all chips can be fixed. The best you can hope for is to fill most of the chip. It may still leave a visible flaw. But the improvement on most chips will be dramatic, and at least you've sealed the chip from the atmosphere and probably eliminated the possibility of it growing larger or discoloring in the future.

You can't fix long cracks. So it's critical that you fill chips before they turn into cracks. Basically, any chip that goes into the surface of the glass perpendicular to the surface or at a shallow angle can be repaired. That includes cone-shaped chips, leaf-shaped chips or almost any chip that hasn't flaked a big piece of glass off onto the road.

Let's Go Shopping

Windshield crack repair kits can be found in the auto parts department of many mass merchandisers like Kmart, Sears and Wal-Mart, as well as more traditional auto parts stores like Pep Boys and AutoZone. Failing that, the warehouse-distributor auto parts stores that cater to professional mechanics can supply you.

In the New York area, we found two different types of repair kits and there may be others. Expect to pay around 12 bucks. Procedures differ marginally, but the principle is the same. We fixed a couple of windshields, and the results were excellent.

How Dry I Am

Start with an absolutely dry windshield that's somewhere near room temperature. Tough on a rainy or wintery day, so you may need to park your car indoors for a while to equilibrate. Dry is really essential—you don't want to trap any water in the repair. Use a hair dryer if the window is wet. If the surface is dirty, don't use any detergents or window cleaner. Lighter fluid or acetone can help dry and clean the surface, but don't use so much that it dribbles down the glass and peels the paint, or worse.

Chips Ahoy

We first used a simple kit from Loctite with a one-part adhesive and an uncomplicated syringe to apply it. Start by peeling off the backing film on one side of the precut adhesive strip and applying it to the precleaned glass, centered over the chip. Burnish with the back of your thumbnail or a blunt object (Fig. 2). Now peel the remaining film. Orient the plastic adapter so that the fitting is as close to vertical as possible and stick it to the film. Burnish again.

Pull the cap off the syringe, keeping it pointy-end up so the adhesive doesn't wind up on the fender. Attach the syringe to the adapter. Now here's the tricky part—lay your watch down somewhere so you can see the second hand. Grasp the syringe body with one hand to stabilize it, and pull the handle of the syringe out as far as it will go. Because you've taken your watch off, you can hold the handle in this position for a full minute. This pulls a partial vacuum in the syringe—and in the crack. While you're holding this vacuum, the air in the crack bubbles up through the adhesive in the syringe, while adhesive creeps down toward the glass and chip.

Now let go of the handle. Don't follow the handle, let it go abruptly. The pressure wave from the handle slamming down will force adhesive into the crack. Repeat this suck-and-slam operation a half-dozen times or so, forcing the crack virtually full of adhesive (Fig. 3).

Now remove the syringe, adapter and adhesive sheet. There will be a film of adhesive on the surface of the glass. You can chase that back with an alcohol-dampened paper towel, but leave the pimple of adhesive right above the crack undisturbed until it cures for a few hours.

With the pimple hardened, simply take a single-edge razor blade and shave the protruding adhesive off. Use a sharp blade, and you'll be able to do this in a single pass.

The Other Path

We tried a different crack repair kit as well. This differed in that it used a two-part adhesive. This required us to mix two small vials of adhesive and hardener in the syringe before starting, which was simple. The adhesive disc and syringe adapter were similar (if not identical) to the simpler kit's, and were applied in an identical fashion. The syringe, however, was more complex. It used a wire latch arrange-

HOW IT WORKS

Laminated Safety Glass

Ordinary window glass, like you have around the house, is pretty amazing stuff. It's clear, strong and cheap. But it's also brittle, shattering into long, dangerous, wickedly sharp shards when overstressed. Plastics would be as strong, but not nearly hard enough to resist scratching and remain clear enough for a car window—just look at any plastic-glazed outdoor bus stop or phone booth, with its patina of fine scratches. For the side windows of cars, automakers have come up with a good compromise: tempered glass. It's stronger than standard, but more importantly, when it does shatter it breaks up into small granules. These granules are still sharp, but should do less damage than the long shards of untempered glass. However, for a windshield, constantly bombarded by pebbles, tempered glass would have a short life span.

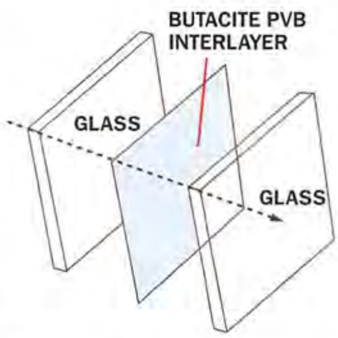

So, many years back, the car manufacturers switched to a laminated glass sandwich for the windshield. It's a simple process. Two thinner sheets of glass are fused to a rubber inner layer. The tempered-glass outer layers are then independent of each other. The rubbery center sheet provides damping to any shock waves from errant stones, reducing the probability of breakage. And if the glass is hit by an object smartly enough, odds are that only the outer sheet will break, as is the case with most stone chips. If a really big piece of debris hits the glass hard enough to break both inner and outer layers, the tough membrane prevents it from winding up in your lap. Even better, the shards of glass from the inner lamination wind up stuck to the membrane, keeping them from spalling away from the windshield at a high velocity, causing great havoc.

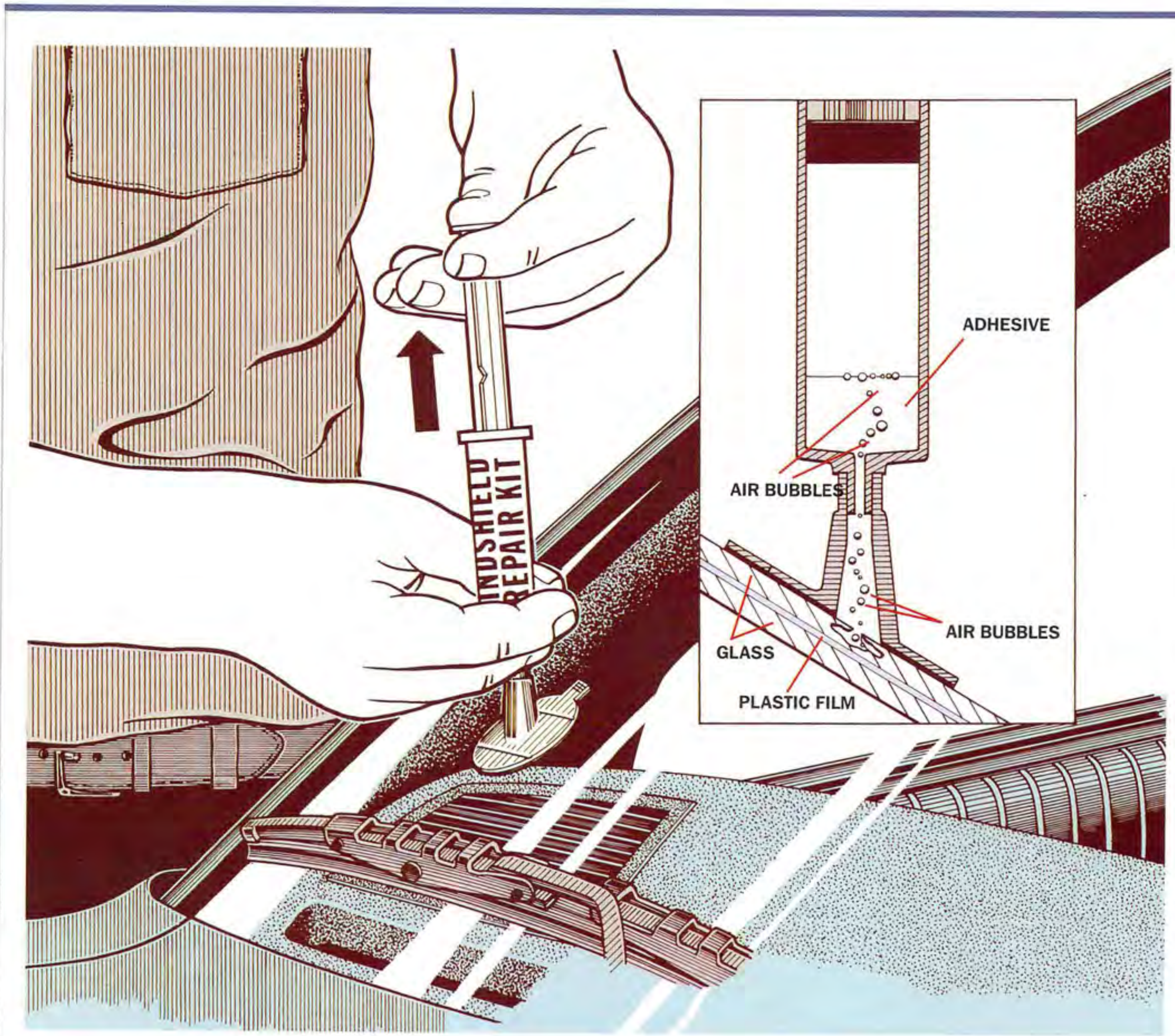

(Fig. 3) Alternating cycles of vacuum and pressure will push adhesive into the chip, and evacuate air from the bottom of the chip.

ment on the body that dropped into two notches on the handle when necessary. A simple pushpin stuck into the body was used as a very crude valve to let air in and out of the body. Once the syringe is attached to the adapter (the adapter already being stuck airtight to the glass), the pushpin is removed. Now bottom the syringe plunger, pushing the air out. Insert the pin, and pull the handle out until the clip clicks into the slot. This will hold the plunger out, and the partial vacuum under it for the designated time.

Now, rather than rapidly releasing the pressure like we did with the earlier kit, remove the pin, admitting air into the syringe barrel. Now replace the pin, sealing the hole. Release the clip, and push the handle in, lightly pressurizing the barrel and forcing adhesive into the chip for a minute or so. The second slot will trap the clip as soon as you've pushed in far enough. Repeat this "vent, suck, vent, squeeze" operation several times to force adhesive into the crack.

While the ingenious clip-and-slot and vent business relieves you of the necessity of constantly grasping the syringe barrel like it's the last beer at the picnic, there are caveats. The vacuum pulled isn't as good, purging less air on every iteration. And the simpler kit seemed to force the adhesive deeper into the crack faster by using the plunger as a piston to rapidly pressurize the system.

Having said that, both kits did a bang-up job. After shaving the excess adhesive off, both cracks are barely visible. The adhesive is clear, and has the same refractive index as the glass so that it renders the crack invisible from almost any angle.

One warning: Don't get cute and try to depress the plunger by hand and squeeze the adhesive deeper into the glass with either kit. The adhesive patch might become unglued and squirt adhesive all over your windshield and fender and shirt. Don't ask how we know this.

EASY

Finding and Fixing Water and Air Leaks

As you finish your morning coffee, you wait for the weather forecast. And it's what you've been dreading: rain. For most people, rain during rush hour means little more than annoyance and perhaps increased driving stress and travel time. But for you, driving in the rain means a miserably wet ride, a ruined suit and possibly ruined shoes as well, unless you take precautions and change into your wet-weather driving clothes. You, like many other folks on the road, have a car that suffers from a water leak.

You're all too familiar with that cold rainwater steadily dripping from under the dash and flowing down your left leg. Your discomfort is exacerbated by the piercing whistle originating at the source of the leak.

You contemplate simply phoning in sick, but then you gird your resolve. Before heading out to the garage, you decide that you're going to find the source of those leaks and squelch them once and for all—even if you have to tear the entire car apart.

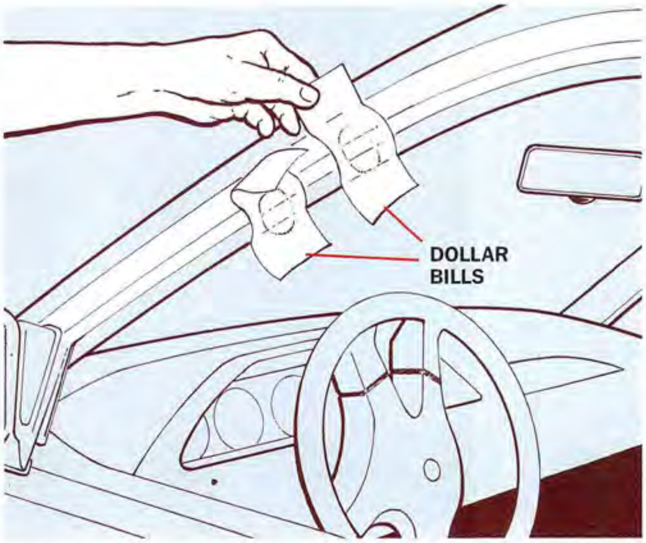

(Fig. 1) A gasket loose enough to let a dollar slide easily might leak air at speed. Tug firmly, but don't tear the bill.

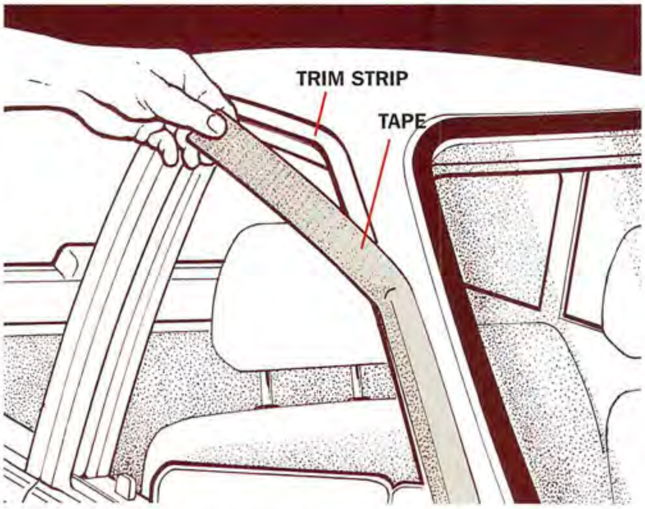

(Fig. 2) Wind can slide under trim pieces and whistle. Use tape to find the culprit, then seal the leak with trim adhesive or silicone seal.

Wind Noise

As you drive, air surrounding the car creates large pressure differentials between the inside and outside of the body. The greatest differential occurs along the sides of the car between the roof and belt line where outside pressure is much lower than the pressure in the passenger compartment.

Weatherstripping and window seals are supposed to keep passenger-compartment air in and outside air out. When they are damaged, however, whistles, hissing and other annoying noises occur.

Therefore, the place to start looking for air leaks is the weatherstripping. Torn, missing or distorted sections are likely sources of leaks that result in wind noise. You should also inspect the door's weatherstrip sealing surface for bumps, lumps or other imperfections that could prevent an otherwise healthy weatherstrip from fully seating.

Replace torn or missing weatherstrips, but do not use a generic weatherstrip intended for storm doors to replace the carefully engineered rubber seal—unless you're desperate. The dealership should be able to order the correct part for you, or you may be able to find what you need on a similar car in a junkyard. It's also possible to repair torn sections of weatherstrip with silicone seal or strong, quick-drying glue. But first clean the pieces with alcohol or lacquer thinner. Don't be afraid to piece short lengths together.

In addition, many firms make replacement gaskets for older cars at very reasonable prices—although you may need to buy an entire kit. J.C. Whitney and other companies also sell weatherstrip by the yard in an array of different profiles.

Distorted sections often can be repaired simply by heating them with a heat gun or a hair dryer and then reforming them by hand.

A quick way to check for air leaks is to rub chalk on the surface of the weatherstrip. When you shut the door, chalk dust will transfer from the

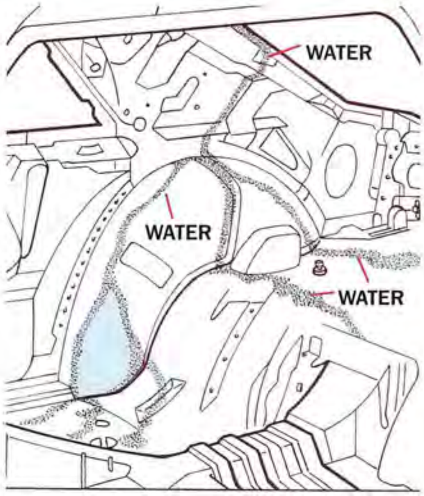

(Fig. 3) Body water leaks are tough to find because the water runs down to a low spot far from the source of the leak.

strip to the door on areas that seal properly. You can also check for leaks in small areas by placing a dollar bill between the weatherstrip and the door (Fig. 1). After closing the door, pull the dollar bill. If it moves too easily, you've found a gap.

Air leaks can also be caused by a misadjusted door or worn door hinges that allow the door to sag. To check door alignment, make sure there are even gaps between the closed door and the body. The outer surface of the door should also be flush with the surrounding sheetmetal. To check for worn hinges, open the door and lift it up and down to see if the hinges allow movement. If they do, they'll have to be replaced.

Pressure Me

You can simulate inside/outside pressure differentials in a parked vehicle using the heater or air-conditioning blower to pressurize the passenger compartment while you look for leaks.

To do this, begin by blocking all interior exhaust vents (they're usually found on the C pillars) with duct tape. Also make sure that any dash-mounted fresh-air vents are closed. Next, start the engine so the vacuum-operated vent controls will operate, then set the HVAC controls to draw in outside air on the highest blower setting. Shut off the engine and turn the key to the accessory position so the blower still operates, and then close the doors.

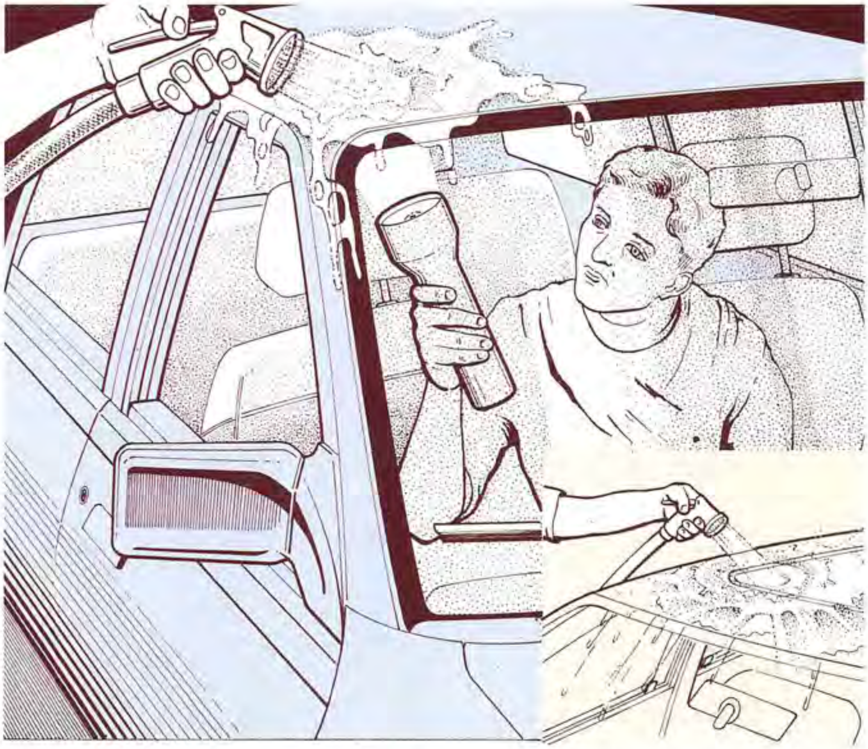

(Fig. 4) Use a helper and a garden hose to follow water leaks back to their source. You may need to remove some trim to find the spot.

In a short while, pressure will build in the passenger compartment. Check for air leaks by moving your hand slowly around window glass and weatherstripping to feel for air leaks. Mark suspect areas with tape so you can correct them later.

Instead of feeling for leaks, you can listen for them by using a stethoscope or by placing one end of a piece of small-diameter tubing at the seal and listening at the other end.

Repair small gaps between the body and the weatherstripping by shimming it with vinyl foam tape. On doors with frames, leaks can also occur around window seals. If the seals are not torn or missing, you can often adjust the window track behind the door panel so the window rides more tightly against the seal.

Wind noise can also be caused by loose driprail covers, molding, the grille and external accessories. If you suspect a noise is coming from one of these areas, cover it or modify its shape with tape to see if the noise disappears (Fig. 2). If it does, you've found the problem.

Water Leaks

You may not realize it, but when it rains, water may flow all over the inside of a car body. Depending on the car's design, water can run down the insides of C pillars, the insides of doors and through the cowl. Fortunately, cars also have systems that channel the water to places that serve as drains. If the drain holes are plugged, water backs up and, before you know it, water is inside the car. To keep this from happening, make sure cowling shields and trunk and hood weatherstripping are in place and in good shape. You should also periodically inspect door- and rocker-panel drainage holes to make sure they are clear of debris that would prevent them from draining freely.

In addition to leaking through weatherstripping and window seals, water can drip through body welds, seams, pinholes, plugs and other areas where gaps or holes exist.

Unlike air, water is affected by gravity, so no matter where the leak originates, the water will eventually wind up in a lower part of the vehicle (Fig. 3). Water can also travel far from its original entry point. For example, it's not unusual for water to enter near the roofline and travel down the inside of the passenger compartment between the trim and body. Whether it becomes visible before it gets to the floor depends on its course.

How do you tell if a water leak starts at a floor-pan seam or the moonroof? Start with common sense: If the floor pan gets wet only when you drive through deep puddles, assume the water is coming from a rust hole, a missing body plug, body-seam leak or other opening down low. The floor pan, fender wells and the lower portion of the engine bulkhead are prime leak areas. If the floor is wet when it rains, but you haven't driven the car, assume that the water is entering at a higher point, such as a moonroof, roof pillar or the window sealing areas. A drip from under the dash might be traced to the cowling below the windshield.

Fight Water with Water

Not surprisingly, you can use water to find water leaks. The best way to spot leaks is to have an assistant inside the car look for leaks while you spray suspect areas with low-pressure spray from a garden hose (Fig. 4).

To accurately pinpoint leaks, you may have to remove interior trim components from the door, roof pillars or even the floor. As you move the water spray from lower to higher points on the vehicle, have your assistant shine a flashlight on areas that correspond to the area that you're spraying.

Leak Repair Products

- **Permatex 65AR: Windshield and glass sealer.**
- **3M 08655: Brushable seam sealer.**
- **3M 08011: Weatherstrip adhesive.**
- **3M 08578: Strip caulk that can be molded by hand to fill gaps, seams and other large areas.**
- **3M 08551: Clear sealer for small leaks around windshield, rear windows, reveal moldings and small seams.**

Leaks in the urethane seal around the windshield or rear window can be especially troublesome to spot. If you suspect an area, wipe the outside down with soapy water, then have your assistant blow compressed air on the area. If the water bubbles on the outside, you've found the leak.

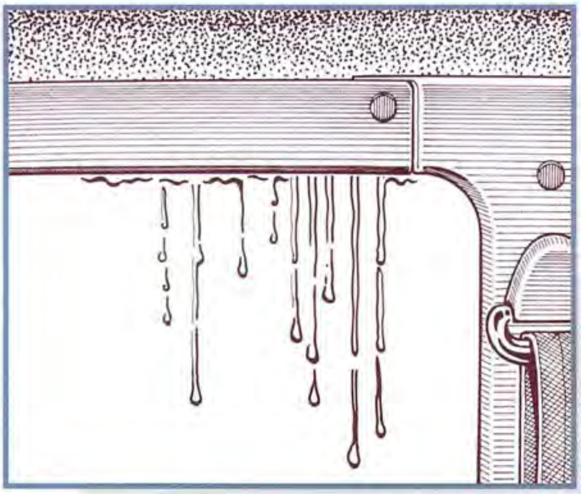

Keep in mind that late-model cars have windshields and rear windows that are installed with urethane sealers. Because they affect the structural integrity of the roof, only glass- or body-repair professionals should attempt to repair these leaks.

Repairs

You'll want to replace or repair all defective seals, weatherstrips and guards (Fig. 5). There are many types of material for filling and sealing body leaks. Consult the box to the left to determine the best material for repairing a leak.

If you want to have a leak repaired by a professional, take your car to a technician who is certified by the inter-industry conference on collision auto repair (I-CAR). I-CAR technicians have been trained in the proper methods of sealing bodies.

(Fig. 5) Sometimes fixing a leak is as simple as reinstalling a misplaced rubber gasket. An auto parts store will have the proper trim adhesive. Clean all the old adhesive off the doorframe and gasket first.

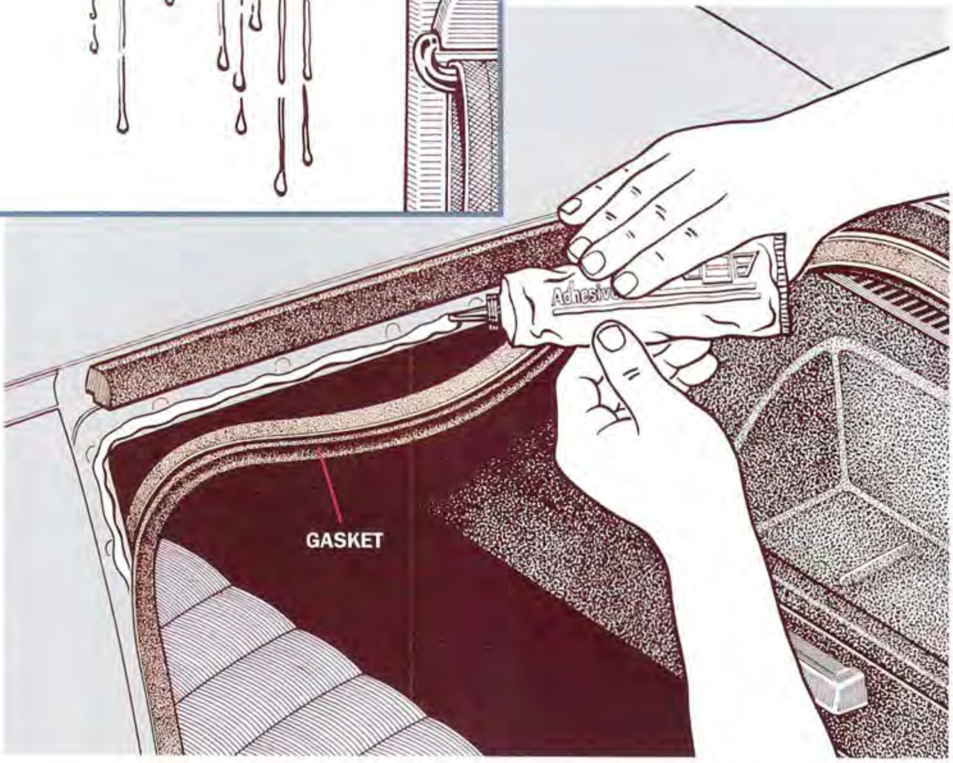

Storing Your Car

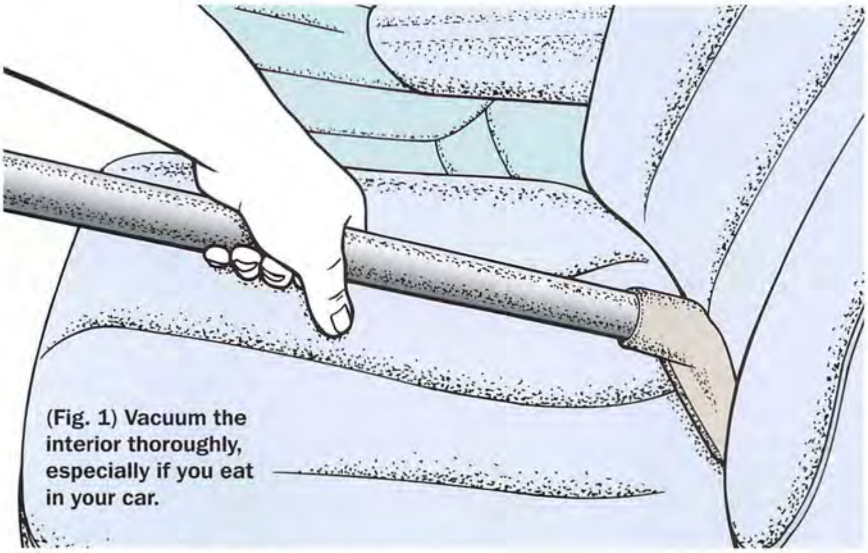

(Fig. 1) Vacuum the interior thoroughly, especially if you eat in your car.

You pull into your driveway, take a loving look at the new convertible and realize: Buying it was the fulfillment of a dream. It's been great fun driving it this summer and fall, but winter is approaching and there's no way you're going to drive it on snow and subject it to corrosive road salt—so you face the problem of storing it until late next spring.

A 2- to 4-month driving season followed by eight to 10 months of storage is something many car enthusiasts

(Fig. 2) To prevent rusting inside the cylinders, spray fogging oil down the plug holes.

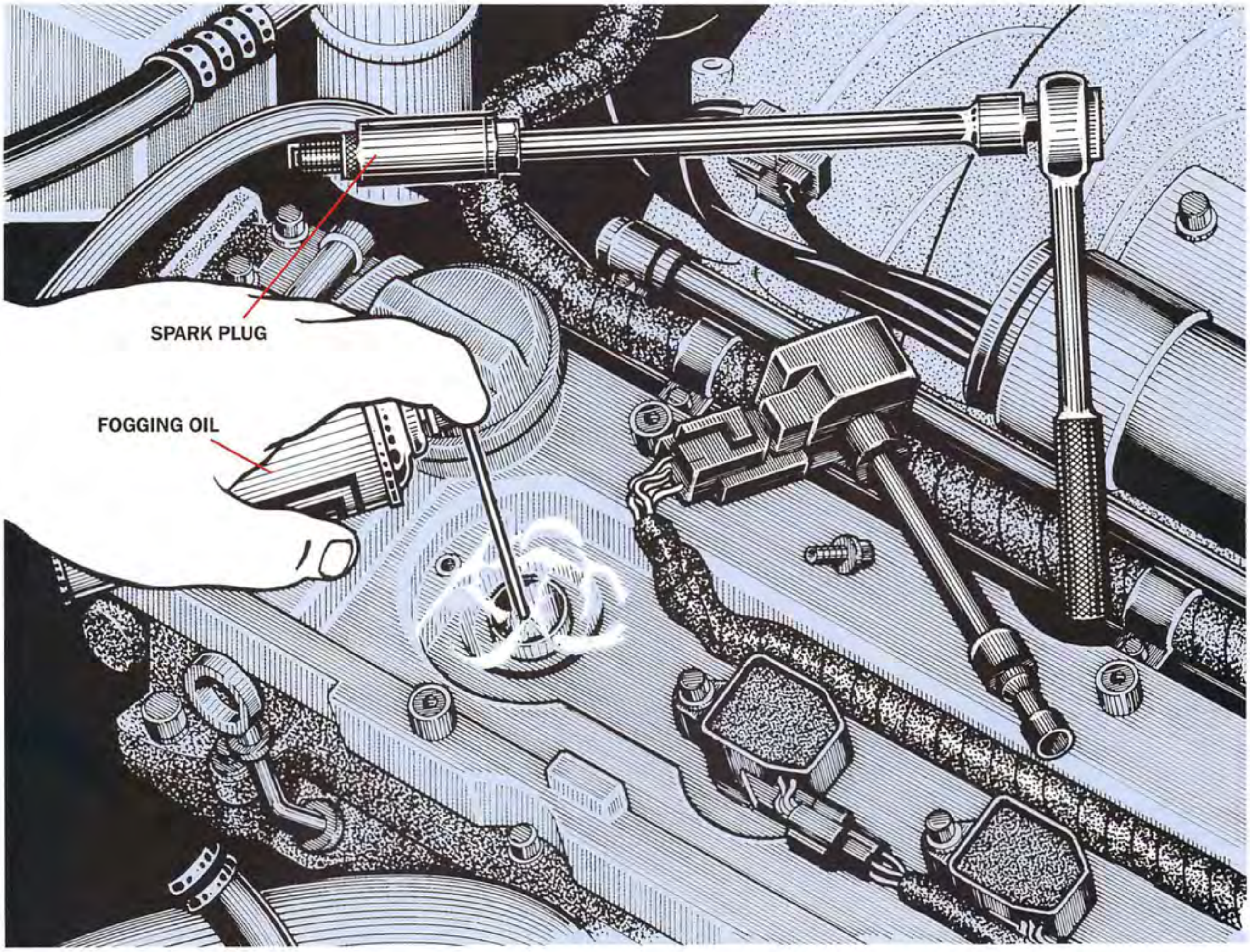

go through every year. Maybe you're a snowbird with a pair of vehicles that go into 6-month storage at both your warm- and cold-weather locations during the away season.

Even more traumatic: You've got a work assignment or a military posting far from home and can't bring the car. Whatever the vehicle, the reason and the season, you want to be able to store the vehicle without big expense, yet with minimum deterioration and an easy return to operation.

The Basics

Indoors is always better, particularly for an older vehicle, even if the storage period is summer in the South. If you're going to be away for up to a couple of years, it definitely has to be kept inside. If you don't have the place, find an indoor storage facility—it will be money well spent. In fact, you also should get someone to take out the car periodically for an "exercising" drive. If you can't afford to do long-term storage right, you might have to pay a lot for restoration when you return.

If outdoor "storage" is your only choice, don't give up. There's still a lot you can do to minimize the damage, particularly for seasonal storage.

Prepping the Vehicle

The cleaner the vehicle the better, and that goes beyond wash, wax and shine. Pick a warm, dry day to do the cleaning. Fill the gas tank, then add an adequate amount of gasoline stabilizer (also made for lawnmowers, snowblowers, etc.) to prevent gum and varnish formation. Drive the car long enough to really warm up the engine and mix the stabilizer with the fuel—at least 30 to 40 miles. Remove dirt from the underbody, particularly from the wheel wells. Dirt holds moisture, and the combination of moisture and air causes iron and steel to rust.

Really, really clean the interior and trunk of the vehicle with a household vacuum cleaner, using those little attachments that reach into nooks and crannies **(Fig. 1)**. The battery-powered car vacuum just doesn't have the suction. The object is to remove all pizza crusts, jellybeans, dog biscuits—anything that could nourish a critter.

Indoor Storage

Allow the car to air-dry. If the garage or shed has a concrete or earth floor, create a floor vapor barrier with plastic sheeting or tarps. With an earth floor, make a drive-along "path" from strips cut from a sheet of plywood and place them over the vapor barrier.

Remove the spark plugs and spray some oil into the cylinders to prevent rust and corrosion. You can use conventional engine oil with a spray-type squirt can or aerosol fogging oil designed for boat storage **(Fig. 2)**. Turn the crankshaft (with a socket and ratchet wrench on the crank pulley bolt) about four to six times to circulate the oil. Reinstall the spark plugs and reconnect the plug wires.

Disconnect the battery cables (ground cable first) and remove the battery. Clean the top and sides of the battery to remove any moisture-retaining, conductive film. Place it on a clean, dry surface such as blocks of wood or a polystyrene spacer **(Fig. 3)**. Connect a float charger, one designed to maintain a battery charge for long periods. Removing the battery also gives you the chance to inspect the battery box for any corrosion, and to clean it out.

(Fig. 3) Remove the battery and keep it clear of damp concrete.

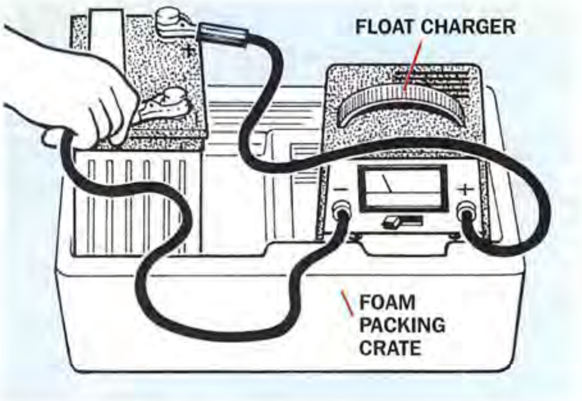

Lubricate the hood release latch, hood and door hinges to protect them from moisture **(Fig. 4)**.

Brake fluid absorbs moisture, which can cause rust and corrosion in the brake system. Flush the old fluid with new. Check the freeze point of the engine coolant with a hydrometer to make sure it's low enough for the ambient temperatures.

Protection from rodents and other critters is important. They not only chew on spark plug wires and other wires, but they can crawl into openings and set up residence. Stuff thick, clean rags into the tailpipe, engine air intake and the fresh air intake in front of the windshield, unless it's covered by mesh (or a cabin air filter). If you know you have mice in the area that may enjoy making nesting material out of your cloth rags, use aluminum foil instead **(Fig. 5)**. If the vehicle is to be left for six months or longer, the issue of flat-spotting the tires is worth considering. Prevention is straightforward if you have a set of four jackstands. Jack up each end of the vehicle so that it's high enough to slip a jackstand, in the lowest position, under each lower arm. Jackstands will eventually sink into dirt floors—use plywood squares under them.

Stuff clean rags between the wiper arms and windshield to hold the blades off the glass (or remove the blades). This will keep them from sticking to the glass, which could both leave marks and ruin the rubber **(Fig. 6)**. Apply a film of rubber lubricant to the squeegees.

Empty a large container of mothballs on the floor all around and under the vehicle to discourage critters.

Outdoor Storage

If you must store outside on an earth surface, the best you can do is park the car on a layer of plywood over a continuous layer of heavy-duty plastic. The combination won't keep all windblown moisture from snow and rain off the underbody. However, to

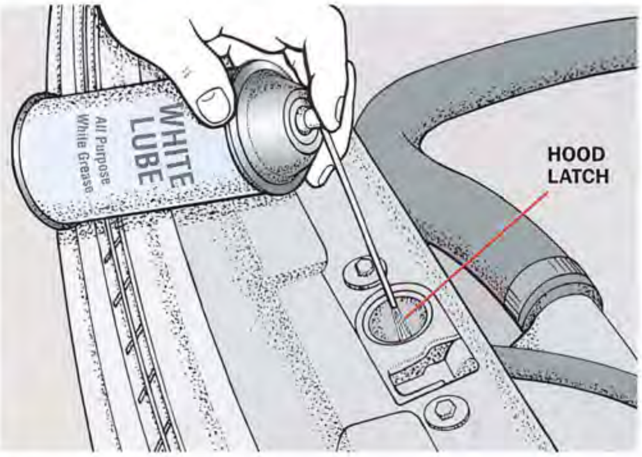

(Fig. 4) Lube all the hinges and latches so you can get them open when you ultimately reanimate your car.

(Fig. 5) Stuff rags or aluminum foil into the tailpipe to keep out rodents and insects.

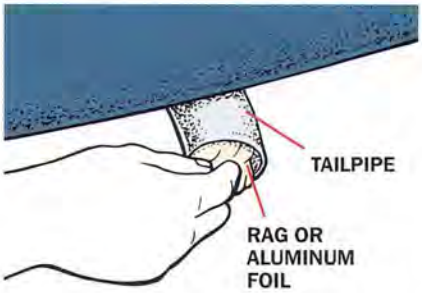

do any better in a windy area you'd have to make the plastic layer oversize. As a final step before placing any cover, lift the ends up and tuck them into the body (trap them in bottom door openings, tape and tie to unpainted trim, etc.) to create a sort of underbody diaper. This is not easy because you don't want plastic sheeting against painted metal body parts, where it could trap moisture and cause rust and paint damage.

Outdoor storage prep, like indoor, starts with a clean, dry vehicle. Put mothballs in the passenger cabin, laying them on sheets of aluminum foil. Leave each window open a half-inch, so the interior can breathe, but cut some strips of fine screening to cover the openings and tape the strips to glass and molding with masking tape. Be careful not to tape to the car body paint surfaces.

Perform all the other indoor prep, including insect- and rodent-proofing and rust and corrosion prevention. Also remove the battery and store it indoors, connected to a trickle charger.

We've saved the toughest topic for last: body covering. The plain fact is that you have to cover the body, and there's no perfect way to do so. That's why we say that indoors is No. 1, and outdoors is No. 2—because it's all that's left.

There are more choices in body covers than we can count, ranging from $20 to more than $300. At the lowest price, you're likely to get just a plastic cover that's made in a few sizes to fit all vehicles in a specific category (car, SUV, truck). It may not fit well and it can trap moisture underneath, damaging the vehicle finish. It's intended as a dust and rain cover for a day or two—at best.

Fit is very important, even if the cover is a "breather" (just porous enough to allow air to pass through, but able to restrict moisture). Wind can whip the inner surface of a loose-fitting cover against the paint and when you peel off the cover, the body may look as if the paint had been sanded. As the prices go up, so does the quality of the fit of the covers—and the materials will be more body-friendly.

(Fig. 6) Prop the wiper arms off the windshield (or backlight) with rags, so they don't stick to the glass.

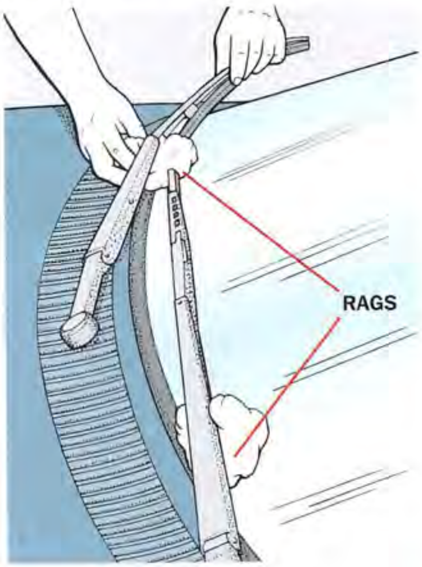

For sunny areas, pick a cover material that keeps out ultraviolet rays to protect the car's finish. But also prep the interior surfaces with suitable protectants such as leather conditioner and plastic treatment.

If you're in a wet 'n' windy area, you can get extra protection by first covering the body paint with soft blankets, tied down with bungee cords. The blankets not only will wick up moisture that gets through, but also provide a protective layer under the vehicle cover.

A premium, breathable custom-fit cover that extends down to cover the wheel wells, and is secured with straps, is your best bet for outside.

Getting Ready to Drive

It's driving season, you've opened the garage door and you're eyeing the car. In addition to unpacking, refitting the battery and so forth, sand most of the rust off the brake rotors with some 100-grit sandpaper on a rubber block, and change the engine oil and filter. You should be ready to roll.

Exercising the Car

Back after an away-from-home work assignment? If you did not opt for the "exercise program," you'll need to take your beloved for a serious drive. Just starting the engine and letting it reach operating temperature is not what we mean. That actually can be harmful, allowing oil dilution by fuel and moisture. A 30-mile drive every 60 days, with brakes, transmission, steering and air conditioning all operating, is what the car really needs. If you don't stint on this, the car will feel every bit as good as you remember.

Drying Out Your Flooded Car

We've all seen it on TV: rainstorms of biblical proportions sweeping entire villages into the sea, and damp, devastated flood victims being interviewed in front of a pile of sodden furnishings and clothing. But then it happens closer to home—only this time it's your car, not the entire village. There's a high-water line halfway up the windshield, and several inches of mud in the interior and the trunk. It smells like a swamp, and it's only getting worse in the hot sun. And the insurance adjuster says he'll be by in a few days.

How Dry I Am

Don't wait for the adjuster to arrive. Mold and corrosion are setting in now. You need to clean out as much liquid and mud as you can and dry out your car as soon as possible. Don't try to start the car. If there's water in the engine, transmission or fuel system, you'll just compound the damage.

Disconnect the battery ground strap first—you must do this, otherwise you'll fry something.

Next, begin assessing just how deep the water got. Frankly, if the waterline is as high as the dashboard, you will probably be better off talking the adjuster into totaling the car and getting another. Double that for salt water. The mechanical systems and the interior can be dried out or cleaned with a lot of labor, but the electrical systems on modern cars are extremely complex. These systems rely on a lot of low-voltage signals from sensors in the engine management system and ABS. These low-voltage signals are extremely sensitive to corrosion on connectors, and problems can crop up for years.

Look for a high-water mark. That can be easy—if the water was muddy or there was a lot of floating grass and leaves. But clean water may leave no residue. Look for water inside the doors and the taillights, and dampness in the carpets and interior trim. This will allow you to eliminate cleaning some areas or systems on the car unnecessarily. Let's go through those systems.

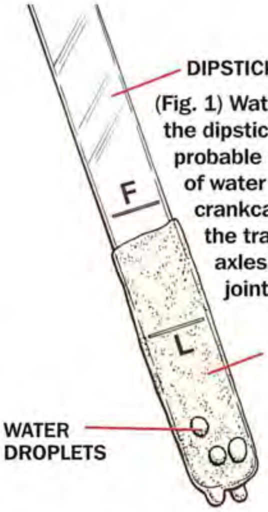

(Fig. 1) Water on the dipstick is a probable indication of water in the crankcase—and the transmission, axles and CV joints, too.

Drivetrain

Check the dipsticks for the engine and transmission. If there are water droplets clinging to the end of either dipstick, you absolutely, positively need to change the oil and filter before even thinking about starting the engine (Fig. 1). If the water was muddy, it's probably wisest to remove the oil pan from the engine and wash the mud out. Change the oil and filter again in a few hundred miles, too.

Late-model cars have sealed fuel systems, and probably won't get any water in them. But that classic '55 T-Bird probably ingested some water if it was deep enough and lingered long enough. Siphon the fuel out into a container and look for water. If you find any, it's probably best to drop the tank and get it cleaned professionally. Blow out the fuel line, and you may need to get water out of the carburetor float bowls as well. If you find evidence of water in a fuel-injected car or truck's tank, replace the fuel filter as well. That paper element may deteriorate if it gets waterlogged. It's not that a few drops of clean water are bad, but floodwater is usually pretty foul with silt and sludge.

Muddy water can infiltrate its way past engine seals within a few hours (Fig. 2). Crankshaft seals, transmission seals and axle and CV joint seals are adequate to keep lubricants in, but they are not designed to keep standing

(Fig. 2) Engine seals are better at keeping oil inside than in keeping water and mud outside.

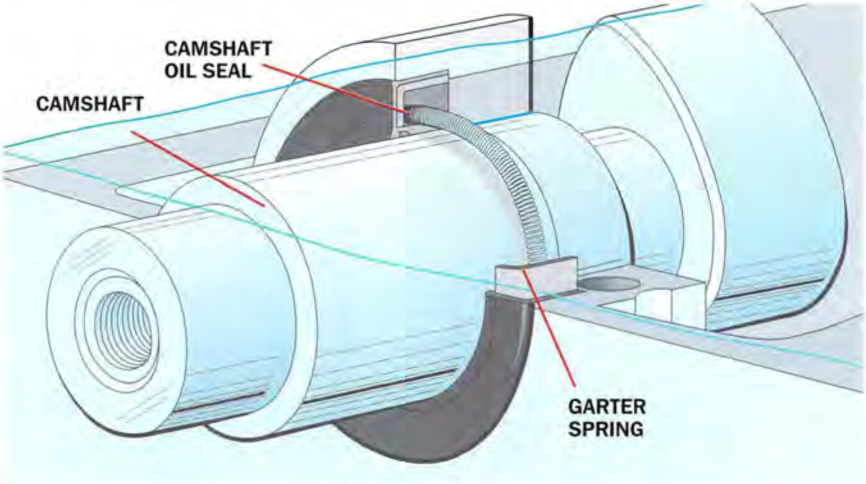

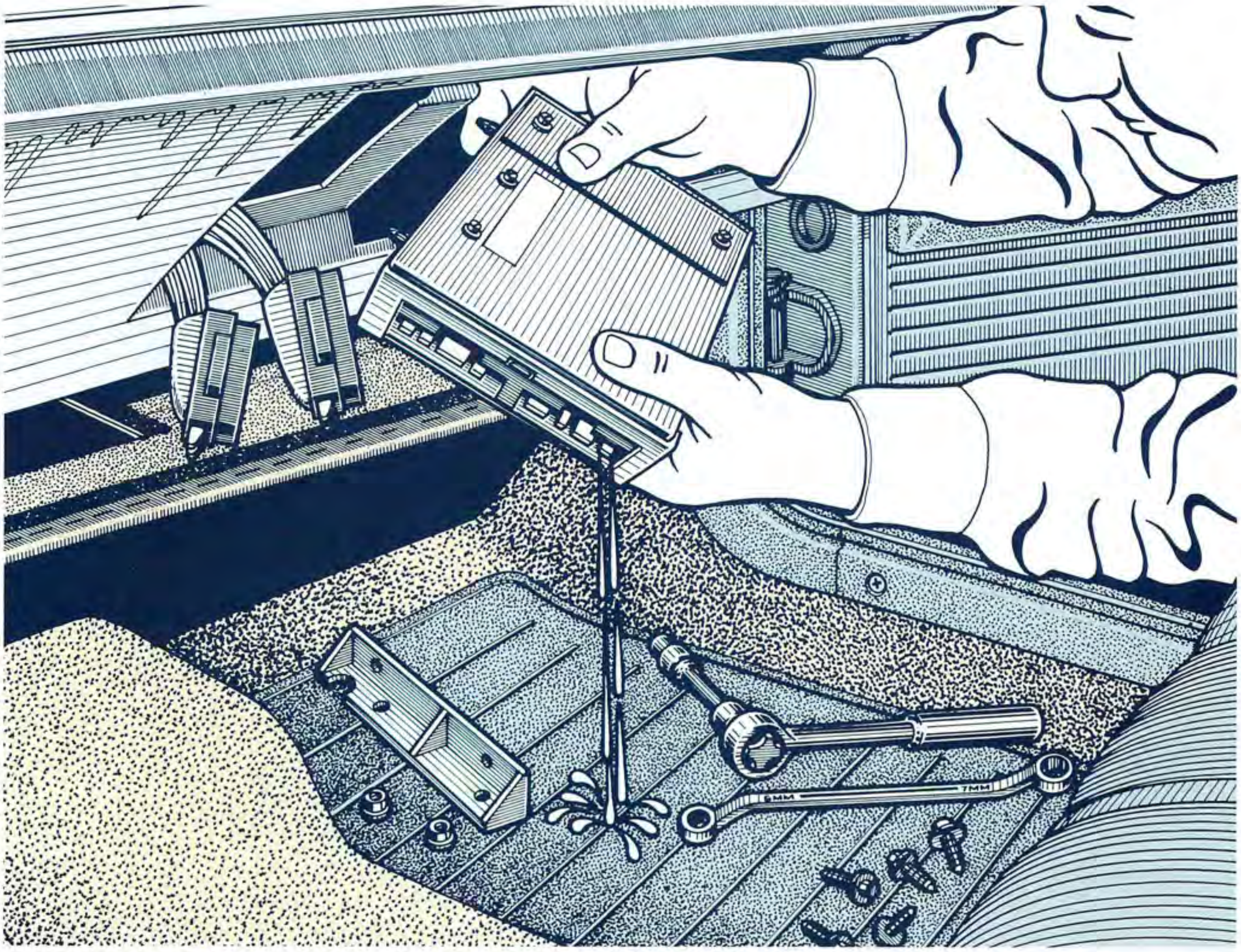

(Fig. 3) Generally mounted in the passenger kickwell, the engine management computer is highly likely to be wet if water rises above the floorboards of your car. Rinse with demineralized water, and dry with a hair dryer. You may be lucky, but more likely, you'll need a replacement.

water from creeping in. Before you start the engine, or tow a car with the wheels on the ground, drain and change the oil, transmission fluid and final-drive lube. Check the dipstick for water droplets. And then change those fluids again in a thousand miles or so if there was evidence of muddy water. And don't forget wheel bearings and constant velocity joints, which will need to be cleaned and repacked. Some front-drive cars have sealed-for-life front axle bearings, and you'll simply have to wait for those to fail, because it's nearly impossible to clean and relube them.

Down Under and In

If the water came up only to the doorsills, you may be in luck. Aside from mud or salt water on the brakes, there's little to damage on the underside of the car. Use fresh water to hose everything down. If the calipers or brake drum cylinders were submerged for more than an hour or two, flush and bleed the brake fluid, in case water seeped backward past the seals.

If the exhaust system was submerged and is full of water, just start the engine after you've determined there's no water in the oil.

Water get high enough to get the interior wet? If the water was clear, fresh rainwater, just vacuum the carpets and let everything air-dry with the windows open in the sun. Odds are that the water was muddy, and that the seats are wet, too. You may not be able to rescue the sponge rubber seat cushions, but it's worth trying. If the seat cushions are wet, rinse with fresh water, and wet-vacuum as much water out as possible. Leave them in the sun to dry. If mildew or mold starts to get a foothold before things dry, use a commercial disinfectant spray to knock it down. Remove as much of the interior as you can, rinse it in fresh water and hang it up to dry. That includes the door panels, and stuff like the interior panels and fiberboard glovebox.

Rinse the inside of the car with clean water and dry it out. This includes the inside of the doors and fenders. Don't forget the trunk.

There's one critical piece of gear you need to deal with immediately if the interior was flooded: the computer. The engine management computer is often mounted behind the passenger kick panel or under the seat, and it's not weatherproofed at all (Fig. 3). Pull the kick panel, remove the computer

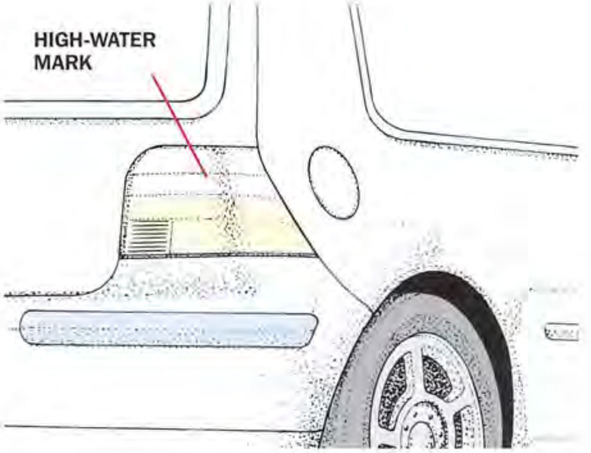

(Fig. 4) Lamp housings can take on water and hold it for weeks.

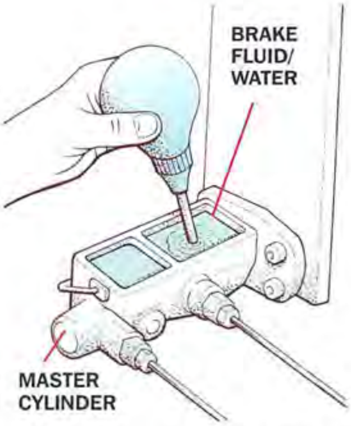

(Fig. 5) Suck most of the brake fluid out and flush the system thoroughly with fresh fluid.

from its harness and get it out. If there's evidence of moisture, rinse in clean water. Then rinse again in demineralized water. Dry with a hair dryer, sunshine or an oven set to 175° F. Clean the electrical contacts on the wiring harness, and lubricate with electrical contact grease. Don't reinstall it until everything is dry.

It's not uncommon to see a high-water mark inside the light housings (Fig. 4), especially if the flood lasted for more than an hour or two. Dampness or water/mud inside the head- and taillight fixtures needs to be dried. Remove the wet assembly, pull the bulb and rinse the light housing thoroughly so the chrome-plated reflector doesn't discolor. Dry in the sun. Remember, if you have quartz-halogen bulbs in your headlights, clean the quartz glass with alcohol and don't touch the bulb while reinstalling it or it'll burn out within a few hours.

Brake fluid is soluble in water, so you won't find any wet evidence of contamination. Water can backtrack past caliper and wheel cylinder seals, so prompt flushing is called for. Suck most of the old fluid/water out of the master cylinder reservoir, and refill with fresh fluid (Fig. 5). Flush one corner of the brake system until you see fresh fluid at the bleeder by pushing down on the pedal while a helper cracks the bleeder, and then hold the pedal down while he closes it. Repeat. Don't allow any air to enter the system at either end. Run the entire quart of brake fluid through to make sure you've gotten it all. Does the car have ABS? Rinse and dry out the ABS computer, although it may be more difficult to find and remove than the engine computer. Check a repair manual to find it, if necessary. Rinse and lubricate the harness connections to the wheel speed sensors. ABS relies on precision metering valves, lots of electrical solenoids and low-voltage sensors, so you can expect problems down the line, especially if your car was immersed in salt water.

Electrical

As mentioned earlier, modern cars have lots of critical, low-voltage, low-current circuits—the kind that are most sensitive to resistance caused by corrosion in the connectors. Fortunately, these connections are pretty well weatherproofed. But it will still pay to dry them out. Systematically disconnect every electrical connector you can find. Do this one at time so you don't wind up reconnecting them to the wrong place, of course. If you find water or contamination, clean with demineralized water to remove salt or mud, and air-dry. Some connections should be lubed with silicone spray or dielectric grease, while others, usually ones with sealed connectors, should be air dried and reassembled dry and clean. Be sure this type is really dry, as any leftover moisture will be trapped inside the connector forever.

On the Road Again

Finally, after everything has been dried out and lubed, put it all back together and give 'er the smoke test—hook up the battery and fire it up. Won't crank? Guess what—the raging floodwaters may have diluted your battery electrolyte. The only fix is a new battery.

HOW IT WORKS

Salvage Titles

As incredible as it may seem, it's possible for a dealer or an individual to acquire, legally, a title for a car that's been flood-damaged and totaled that doesn't reflect the damage. In most states, totaled cars' titles bear a salvage tag on the title. But a dealer can wholesale the car out to a state that issues the fresh title without tagging the car as salvage. Which states? It doesn't matter because once the title has been laundered, it can be retitled in any state, clean as a whistle. And I'd rather not say, so nobody gets any ideas about laundering the title to a damaged car.

So caveat emptor (let the buyer beware). Always examine any potential used car or truck purchases with a jaundiced eye. Look carefully for evidence that the vehicle has been wet: i.e., mud in unusual places in the trunk, water marks inside the instruments, an owner's manual that looks like it's been wet, warped fiberboard door panels or glovebox interiors.

Repairing Paint Chips

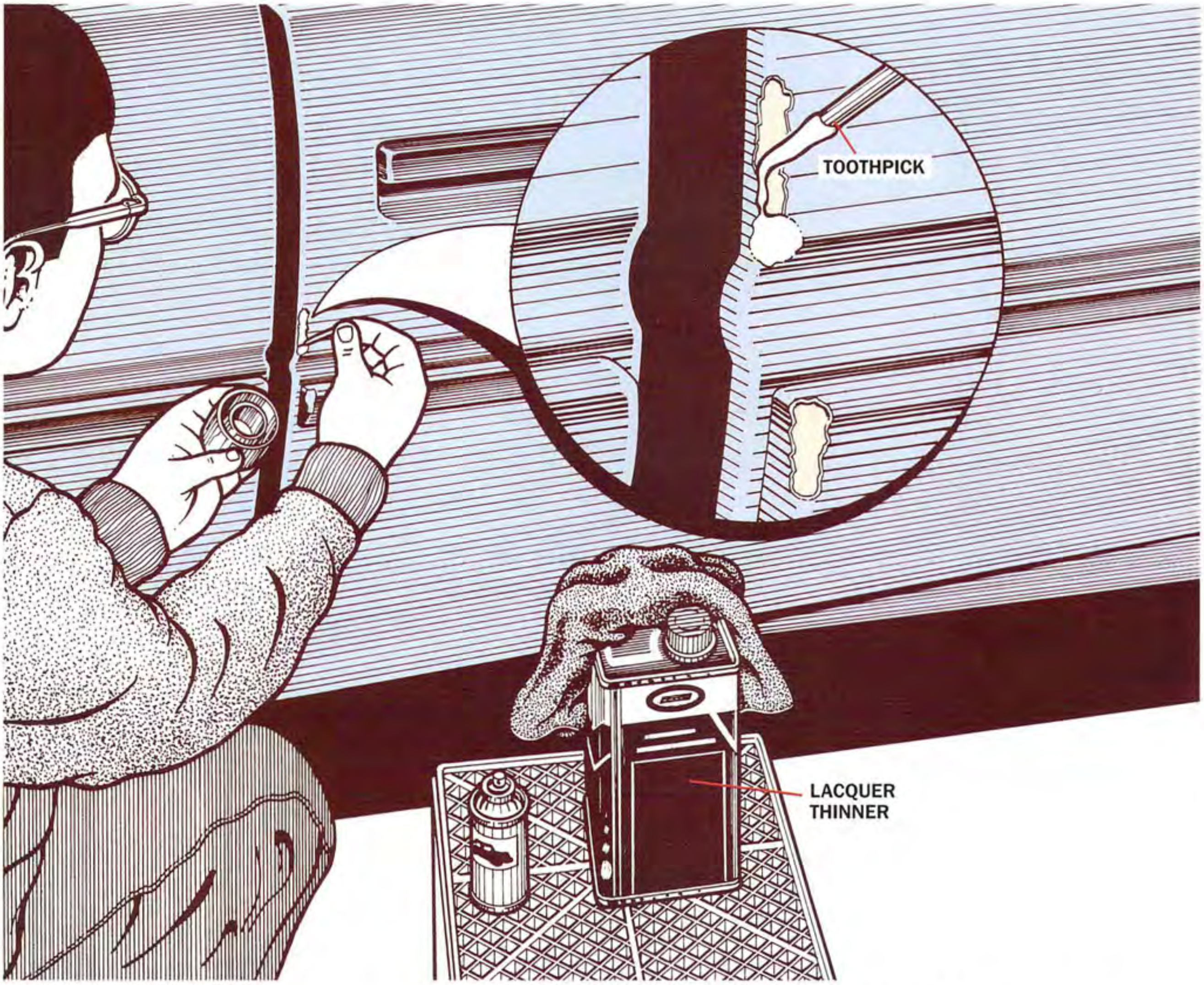

As always, you've parked your new, expensive, shiny car at the far end of the parking lot, several rows away from the nearest vehicle of any sort. You've even been careful to park at the top of the lot's drainage pitch so any errant shopping carts will roll away from, not toward, your car.

But nobody said life was fair. Returning to your car 3 minutes later with a $50 bag of chocolate Napoleons, sushi and okra for your pregnant wife, you find that the unthinkable has happened—there's a rusted, sagging minivan parked only inches from your car. And, yes, the careless driver has managed to chip the paint on your wheel well arch in two places.

Getting It Fixed

The body shop wants $250 to begin repairing it. And the shop foreman says something about clearcoat and not guaranteeing an exact match for your pearlescent mica paint.

Unfortunately, that's about the size of it. It's frustratingly difficult to match many of today's high-tech finishes if you respray an entire panel. If the nick is small, but large enough to go through the clearcoat and into the pigmented paint, your best bet may be a simple touchup, which you can do in the driveway. It won't be perfect, but it may be far less noticeable.

(Fig. 1) Carefully apply touchup paint to fill chips and cracks. Degrease, sand and prime bare metal before applying color.

Paint by Number

This method is best for small scratches or the chips that flake off

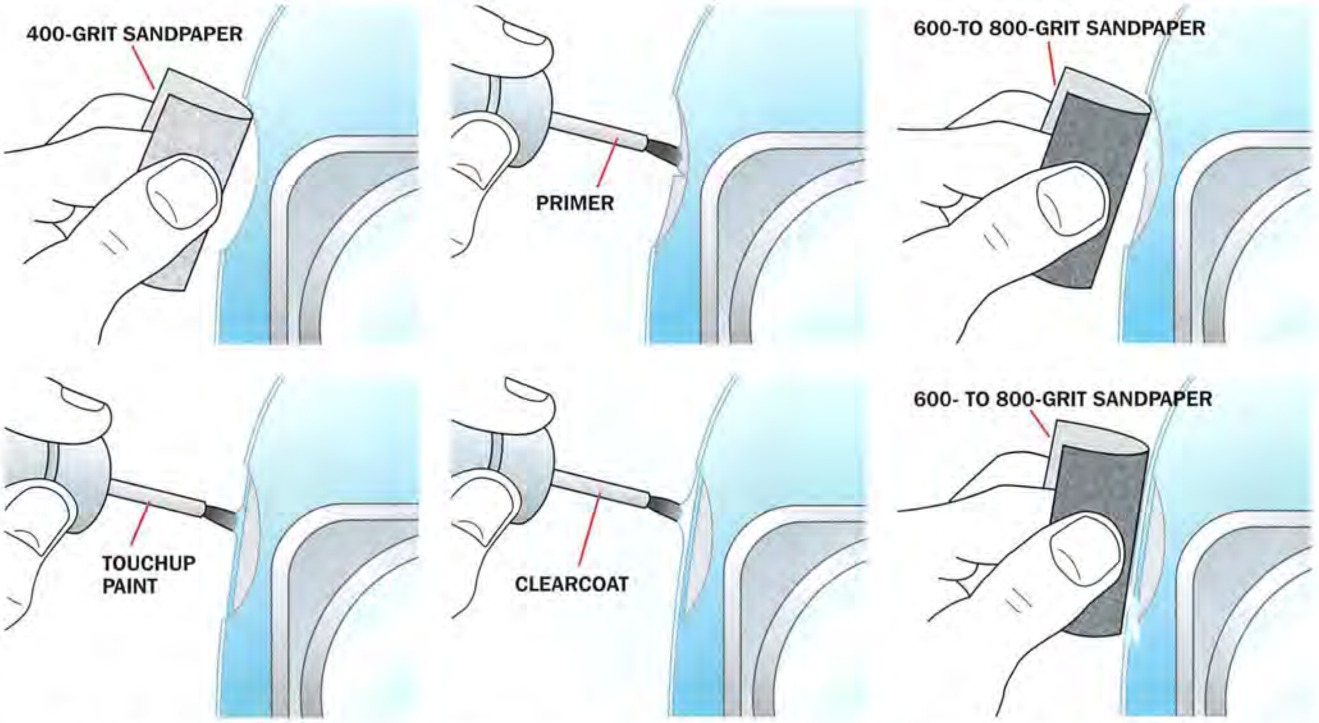

(Fig. 2) Apply primer, color and clearcoat in layers over chipped areas. Sand the area you're working on between layers to keep it smooth.

edges—for example, near trunk and door openings. Work indoors in a heated garage during the cooler months, or outdoors out of the wind and sun. If you live in Truth or Consequences, N.M., or Las Vegas, you may want to work early in the day, before the heat builds up, to keep the paint fluid enough to flow properly.

Go to the car dealership or auto parts store and acquire touchup paint in the appropriate color, clearcoat and, if you've got bare metal showing, primer. Do not use primer intended for lawn furniture or naval vessels—it should be automotive primer, preferably of the same brand as the touchup paint.

Start by using masking tape to isolate the area around the chip, to prevent you from damaging other parts of the finish. Allow about ¼ in. around the damaged spot. Next, clean road grime and wax away from the inside of the chipped area with a cloth moistened with lacquer thinner. (If your car is painted with a lacquer-based paint—which is possible if it dates to the '70s or earlier, or has been repainted—the paint will dissolve in lacquer thinner. Use denatured alcohol instead.) If the paint is chipped down to the metal, use a sharp knife point or some 40-grit sandpaper to rough up the surface, particularly if there's any rust. With the applicator brush in the bottle or with a toothpick, prime any bare metal. Flow primer in a thin coat, but be careful not to lap any primer onto the paint surrounding the chip. Your job at this point is to lay a smooth layer of primer down without any lumps or bumps—just enough to cover the bare metal. Let this dry for at least 24 hours.

Now comes the tricky part. Using the brush or toothpick, fill the chip with paint in a nice, even layer (Fig. 1). Don't try to fill it up, just be sure you get to all the corners. Don't let it sag or bulge. If you get a little overlap, use a cotton swab lightly moistened with lacquer thinner to mop up any excess paint. If the paint is too thin, leave the applicator brush out in the open for a minute or two and the paint will thicken. If it's too gooey, a few drops of lacquer thinner will fix that. Thin sparingly, if you must.

Allow the layer of paint to dry for 24 hours. It will shrink substantially as it dries, and you don't want too much paint. If it gets lumpy, you can try sanding it with 600-grit or 800-grit waterproof sandpaper and water (Fig. 2). Don't sand the paint surrounding the chip—you want that to remain undisturbed.

Build up the color touchup paint until it's nearly flush with the surrounding surface, but definitely recessed. If your car is still wearing its factory coat, that may be only one or two coats of color. Resprayed cars with a thicker layer of chip-prone paint may require four to six coats. Keep chasing the overlap back with cotton swabs and thinner.

After a week's drying time (longer in cool weather) you can coat the repair with clearcoat. If you've had good luck filling the chip, you may simply be able to continue with the process of flowing touchup paint right up to the original paint without overlap, and achieve a nearly invisible repair.

If not, you'll need to blend the repaired area with the surrounding paint. Overlap the chip by ¼in. or so. Add enough coats, a few days apart, to allow for the inevitable shrinkage, to make the repair stand a few thou-

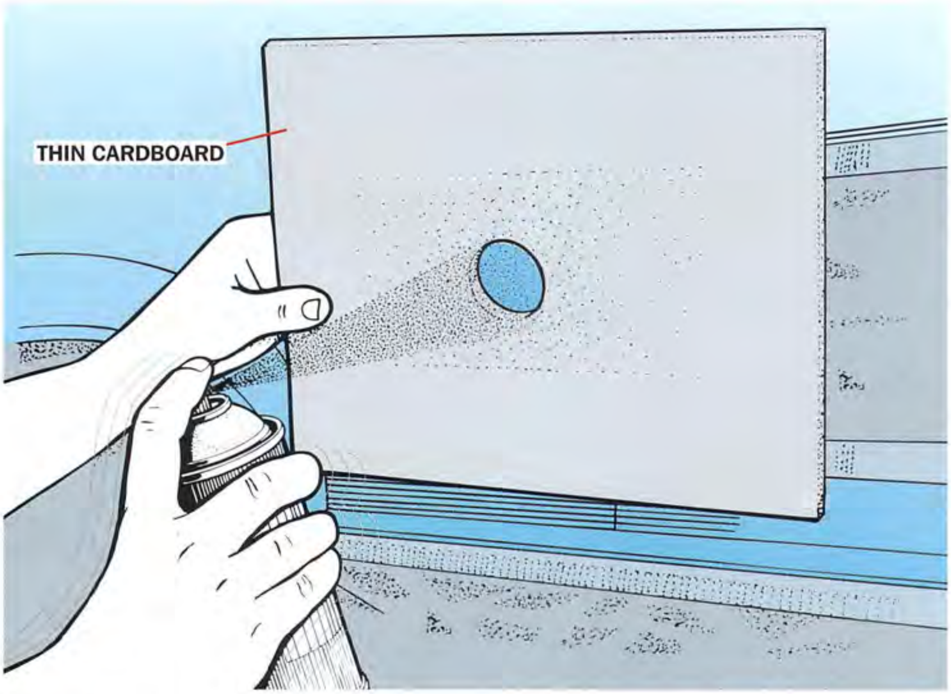

(Fig. 3) Use a cardboard mask to keep overspray to a minimum when spraying.

sandths of an inch—say, the thickness of a sheet of paper—higher than the surrounding paint. Allow the repair to dry and shrink in the sun for a week or so. Now you can gingerly sand with 600- or 800-grit waterproof sandpaper to blend in the color. This will feather the repair into the surrounding clearcoat and smooth the work, but will also leave it with an unattractive matte finish. Use a soft cloth and some rubbing compound to turn the sanded area shiny.

Chipped Off

Got a chip in the middle of a flat panel? If you've tried the above repair and weren't happy with the results, you might get a better, less visible repair with spray paint. Again, you'll need to degrease, derust and sand the area for good adhesion. Cut a 1½-in. hole in a file folder or piece of thin cardboard (Fig. 3). Gently warm a spray can of automotive primer to 100°F with warm water to increase the pressure for a more even spray. Take the spray can in one hand and with your other hold the cardboard with the hole centered over the chip, 2 in. away from the panel. Sweep the spray can back and forth over the hole while spraying three or four passes over the chip. Don't put enough paint on the panel to sag, or even get shiny.

Wait an hour and repeat. Remove the masking, and use 800-grit paper to remove most of the primer from the paint, leaving the primer in the chip to backfill. Now use rubbing compound on a soft cloth to remove the overspray surrounding the area (Fig. 4).

Be aware that primer and glazing putty will shrink in a few days.

Now that the repair is filled nearly level, mask the area again and spray touchup paint in your matching color through the hole in your cardboard mask. Give it about three coats, roughly a half-hour apart. Remove the masking, and let dry at least overnight. Use rubbing compound on the overspray again. Repeat this process with spray-can clearcoat.

You may wish to lightly sand the clearcoat with 800-grit paper, but rubbing compound should bring up the shine and blend the overspray into the panel's original clearcoat. The repair may still be visible, but it should be far less noticeable.

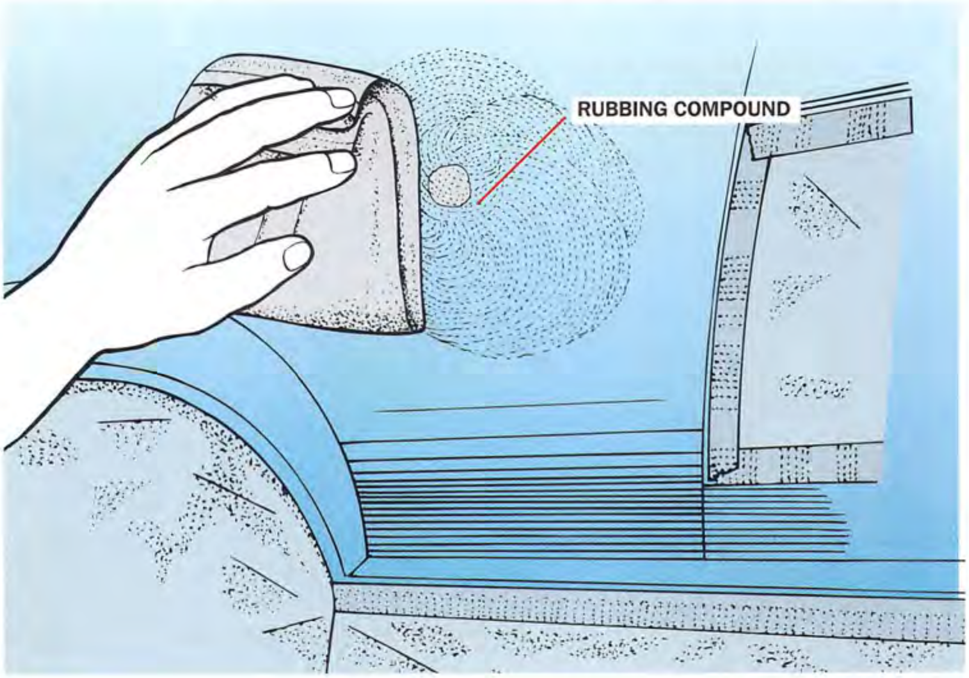

(Fig. 4) Use rubbing compound around the edge of the repair to clean up overspray before spraying the next coat of primer or paint.

Polishing Your Car

In most parts of the country, the winter brings with it some road salt or sand and all types of precipitation. All of the above are less than good for your car's or truck's finish. So spring cleaning not only makes your vehicle look sharper, but it also helps restore and preserve the quality of its finish. Your car will be worth more at trade-in or resale time if it looks good.

You can help ensure that it looks good by regularly shining it up after you've washed it. Before you shine, though, be sure to match the type of polish or wax you use with your car's or truck's paint finish.

Polishing your clean car entails more than simply laying down a thin layer of wax or polish over the paint. The process also involves removing the very fine layer of dull, oxidized and weathered finish on the surface. This dull layer can either be removed chemically or with a fine polishing abrasive. With the oxidized layer removed, the fresh finish beneath it will shine. The wax or silicone in the polish then adds luster and protection to that clean paint.

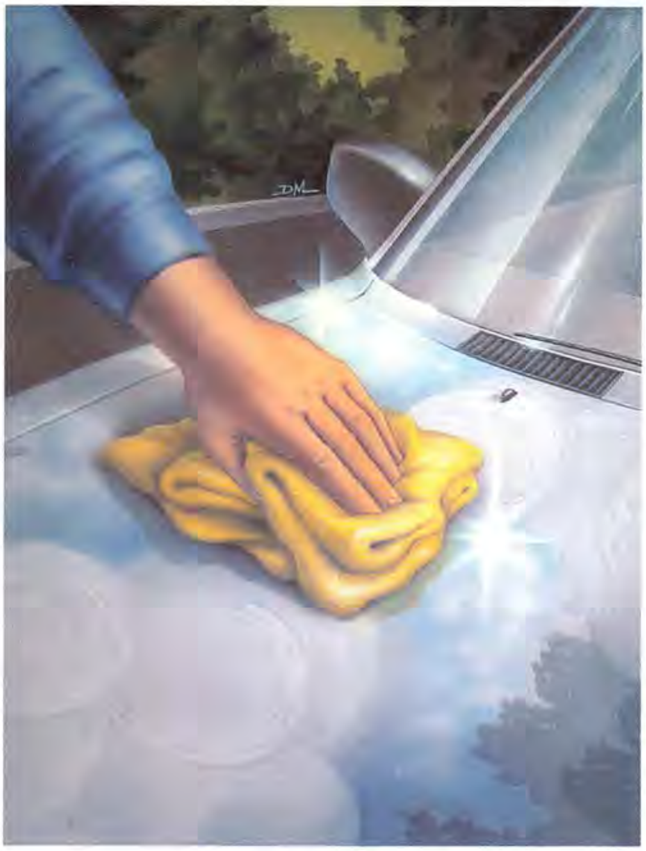

If your car or truck has a clearcoat/basecoat finish, you must use a wax or polish that's specially designed for it. The wrong product may actually dull rather than shine the finish. If you use a nonabrasive product made for a clearcoat finish on an acrylic nonclearcoat finish, the results will be disappointing at best. Be sure to use a product designed for a nonclearcoat paint.

If you opt to shine your car with a wax, use a quality one that lists carnuba wax among its active ingredients. Or use a synthetic that contains silicone resins or amino-active silicones. (Most products have silicone fluids or oils that will make them easier to apply.)

Regardless of the type of finish your vehicle has and the finish-care product you use, the actual task of shining up the car is pretty much the same. Work in the shade, but not beneath the birds' favorite sap-bearing tree. Begin polishing early enough to avoid having to work in the heat of day, and finish the job early enough to let the wax cure before the dew begins to settle.

Start on the roof and work your way down. Apply the wax or polish to a small section no larger than about one-quarter the area of the roof. Use a soft, damp applicator to put down a layer of wax. Small household sponges make excellent, disposable wax applicators. Don't be alarmed if your applicator removes some paint along with the polish on a nonclearcoat finish. This is the dull, oxidized layer coming up. Avoid getting wax or polish on any matte black plastic parts, rubber trim or molding, or on a vinyl roof. Allow the wax to dry to a haze according to the product's instructions.

To get a show-car shine, buff off the wax or polish in a linear rather than a circular motion. This takes more time but it leaves the car's finish with a uniform-looking luster.

If you're using a totally nonabrasive wax or polish and don't mind swirl marks, you can finish the job using a power buffer. Make sure that the power buffer's lamb's-wool bonnet is clean. Change or rinse the bonnet frequently to avoid scratching the finish. Use the bonnet as a mitten to buff in the crevices and other areas that the power buffer can't reach.

To help maintain the shine on your car after you've restored it, try to keep your parked car covered—in the garage or carport or beneath a cover. Also, periodically hose grit and dust off the finish, but avoid the temptation to wipe down the car after a rain—there's all manner of grit in those standing raindrops.

When the finish no longer beads water the way it did after you polished, it's time to wax again—even if it still looks good.

Repairing Plastic Bumpers

After you drove into the mall parking lot and saw how packed it was, you couldn't believe your luck when you spotted a space right next to the building. So you hurried into the space without noticing the extra-high concrete curb—until you heard your flexible plastic bumper strike it. To make matters worse, backing up in a panic put a nice tear in the bumper.

Oh well, there's no use crying over spilled milk. Thankfully, you don't have to cry over a split bumper, either. The 3M Co. makes a repair system designed to fix cuts, tears and abrasions in flexible bumpers. Once repaired, the bumper can then be painted its original color with an automotive paint that has a flexible agent so it won't crack when the bumper flexes.

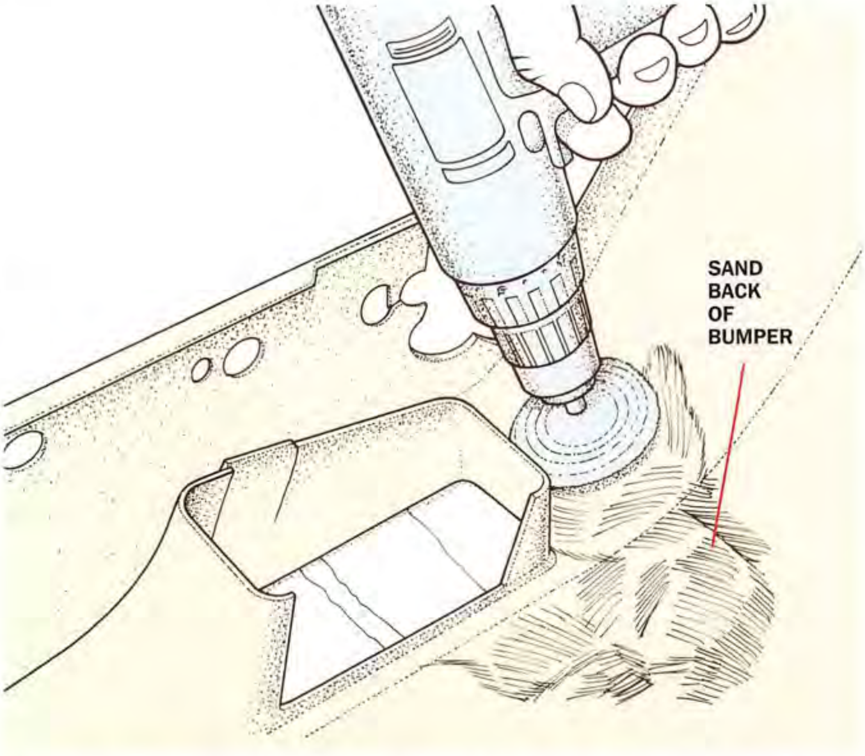

(Fig. 1) Begin the repair by thoroughly cleaning and then sanding the broken part. Then, clamp it in its normal shape.

Velvet Glove

All right, technically it's not really the bumper—it's the bumper's cosmetic cover. The true bumper/impact absorbing system is probably a metal beam that is intended to do the real work. And if *it's* damaged, you'll need to replace the components with new ones to provide your vehicle with the correct crash protection. And don't skimp there. However, the plastic skin isn't structural, despite the fact that it's the most expensive part of both the front and rear bumpers. Odds are that most low-speed mishaps affect only the cover, and those can be repaired if you follow our directions.

The 3M P.R.O. flexible-parts repair system consists of a 2-part flexible filler that is mixed together (like epoxy), a flexible putty and a flexible coating that covers and seals the repair before painting. There's also an adhesion promoter that must be used on plastic parts made with polyolefins. Other items you'll need to make the repair include an electric drill, a 3-in. sanding disc with 36-grit discs, a random-orbit sander with 180-, 240- and 320-grit discs, a rubber sanding block, 240- and 400-grit wet/dry sandpaper, 80- and 120-grit sandpaper, a couple of body-filler spreaders and some method of spraying the coating. While a compressed-air spray gun is ideal, we've illustrated a small aerosol sprayer you can fill with liquid yourself. Pre-Val is one popular brand.

Stocking up

If you've done bodywork in the past, you'll be in good shape to repair these plastic parts. Most of the tools and the style of work is the same, but don't try to use the same materials to repair these flexible parts. The plastic is quite different in its adhesion characteristics from steel or fiberglass. Unless the part is prepped carefully, the repair materials will peel off within a few months, if not sooner. Because the plastic is quite flexible, ordinary body fiberglass resins, plastic body fillers and glazing putty will crack. Permatex and 3M are the major manufacturers who sell appropriate repair materials. The car companies all recommend these products for their warranty repairs.

To find a place that sells them, look in the Yellow Pages for an auto parts store that caters to the body-shop trade. It'll have paint and painting supplies. Sometimes you can find a paint store that also sells automotive paint and supplies. This type of establishment is used to dealing with professional body shop staff, not do-

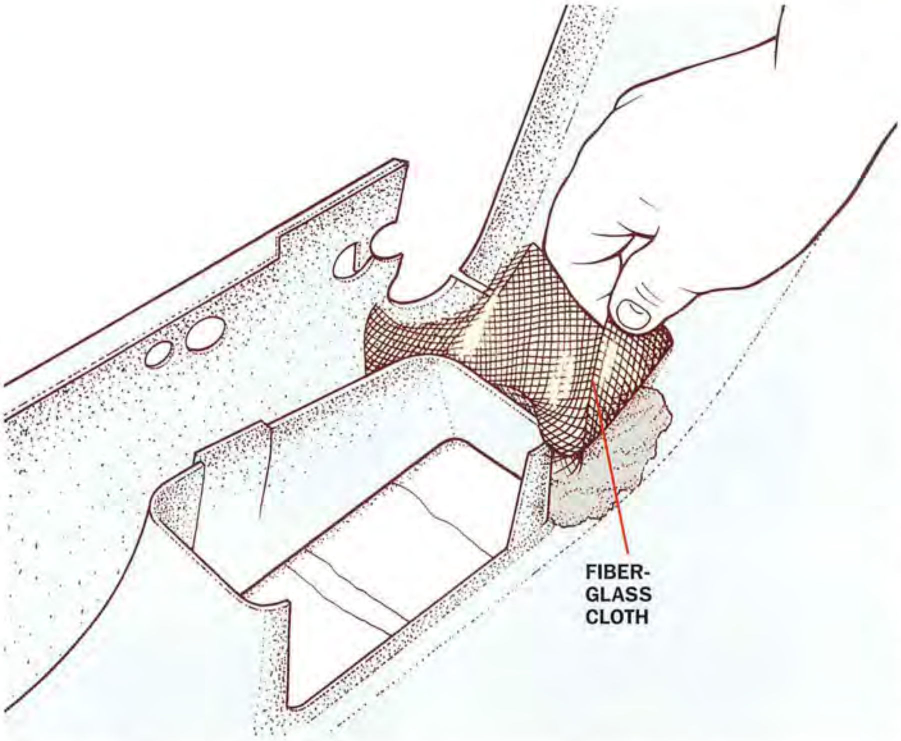

(Fig. 2) Use fiberglass cloth and the special epoxy to reinforce the broken areas. Use your squeegee to work the resin into the cloth.

it-yourselfers, so if you have to ask a lot of questions, be patient and polite.

Most body-working materials are marked "For Professional Use Only." Not to worry, they're not Kryptonite. Use common sense and the same care you would with any epoxy resin or paint with volatile solvents and fine particulates. Avoid skin contact with the uncured resins. Work in a well-ventilated area or use an activated-charcoal painter's respirator to avoid inhaling fumes from the spray gun. Disposable respirators are priced very reasonably. If you plan on doing a lot of bodywork in the future, spring for a better one that uses replaceable activated-charcoal cartridges—it'll be cheaper in the long run. If you're sanding, wear a particle mask. The store that sells you supplies should have all of this equipment.

Plastic Primer

It's possible to repair small cuts, holes, abrasions and other light damage without removing the bumper from the vehicle. However, if the cut is longer than an inch, or it extends to the edge of the part, you'll have to remove the damaged component so you can reinforce the back side of the repair with fiberglass cloth.

As noted earlier, most flexible bumpers are nothing more than bumper covers backed up by other plastic components, a steel structure and collapsible bumper mounts. If you are unsure how to remove the bumper or how to remove the flexible cover from the bumper structure, consult a service manual.

A special adhesion promoter must be used between all the sanding steps on parts made with polyolefin. Therefore, you'll have to determine if the part you're working with is a polyolefin plastic before you begin. In many cases, the letter symbols TPO, PP, E/P or EPDM will be clearly marked on the inside surface of the plastic. Any of these indicate that it is a polyolefin plastic. If you can't find letter symbols on the part, sand the plastic with a 36-grit, 3-in. disc using a drill motor. If the plastic melts or smears, it's a polyolefin plastic. If it sands cleanly, it isn't.

Don't confuse pliable plastic body parts on or around the bumper

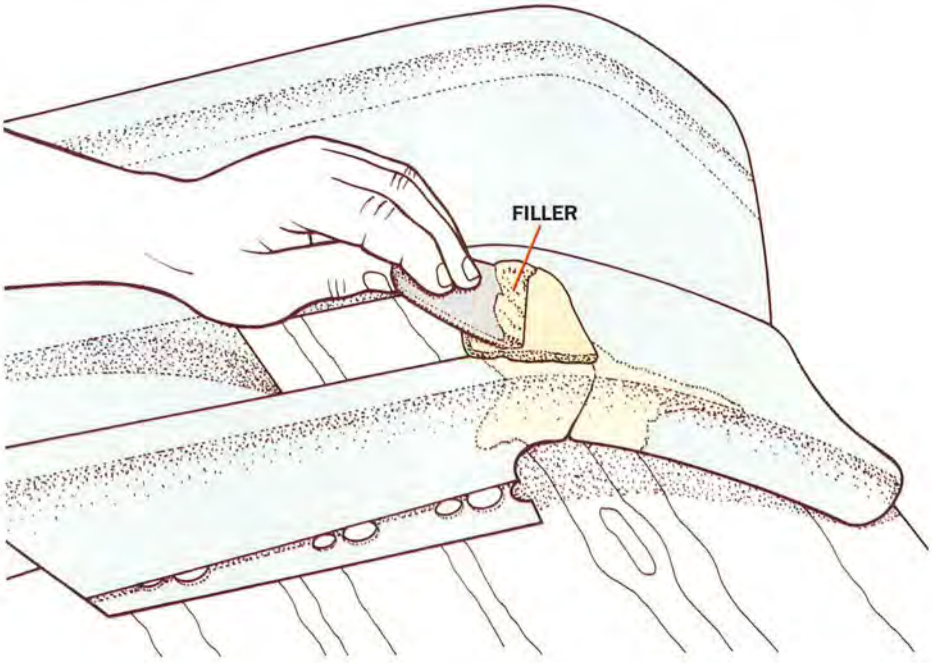

(Fig. 3) Sand a "V" shape into the bottom of the crack. The repair epoxy can then be used to fill in the area. Wait for the resin to cure, then sand.

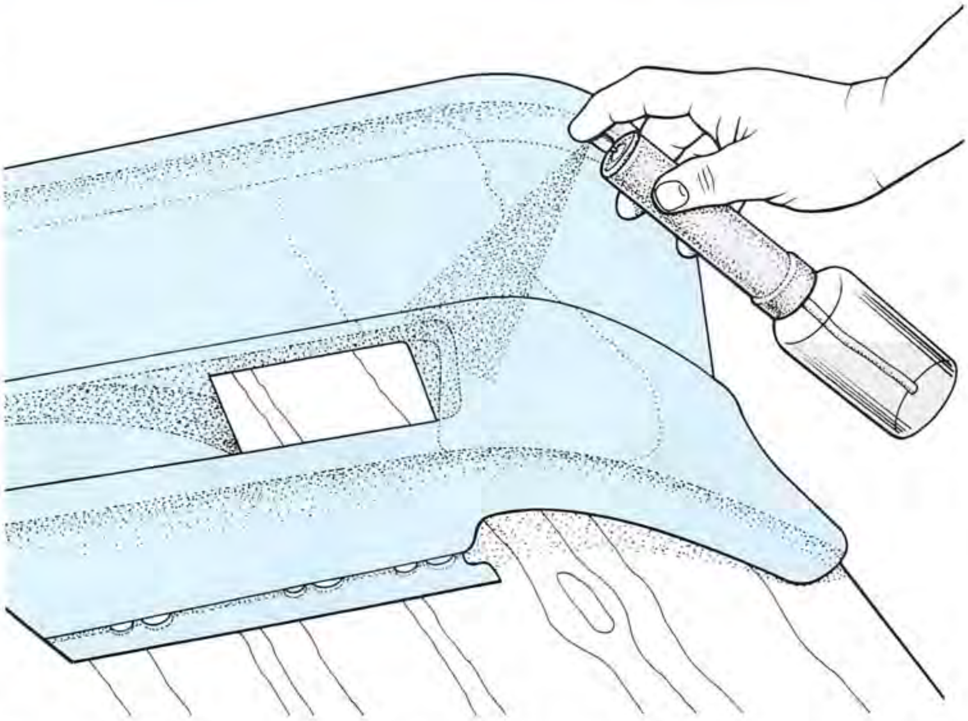

(Fig. 4) An adhesion promoter should be used to make sure the repair materials and paint will adhere. This is necessary with most plastics used to make bumpers.

with the sheet-molding compound (SMC) or fiberglass-reinforced polyester (FRP) parts commonly used on fenders and doors. These rigid, composite materials are identifiable by the fiberglass strands visible at tears or holes, and they require a different repair method.

Getting Started

Once you've determined the type of plastic you're working on, wash all the dirt and grime from the inside and outside of the part with soap and water. After the part dries, clean the area to be repaired with a specially formulated adhesive-cleaning solvent to remove any grease or tar. Be absolutely certain the area is completely clean—don't even touch the area with your fingers after it's cleaned. Use clean white paper towels, without any printed pattern, to apply the cleaner and to dry it off. Greasy shop towels, or any other source of contamination or oil, will prevent adhesion between the repair materials and the base plastic of the bumper, giving you at best a short-lived, sloppy-looking repair.

To reinforce the back side of rips, cuts, holes and tears, grind the back side of the area surrounding the damage with a 36-grit disc, making sure to leave a rough area of at least 1½ in. surrounding the damage. Don't grind all the way through the part. Just rough up the surface (Fig. 1). Next, cut a piece of fiberglass cloth tape to fit over the area. Again, keep the area and your hands squeaky clean. Leave the fiberglass cloth in its original packaging until you need it. Cut the cloth with clean scissors on a clean, grease-free countertop. Then, thoroughly mix together equal amounts of the 2-part repair filler. (If you're working with polyolefin plastic, spray the repair area with adhesion promoter and let it dry.) Next, use a plastic spreader to apply the filler to the area before applying the fiberglass tape. Once the tape is pressed down, apply filler to the top of the tape, using the spreader to force filler into the cloth (Fig. 2).

It takes about 30 minutes for the filler to set thoroughly at 70° F, but temperature and humidity affect setup time considerably. If you need to work in the winter, take the parts and repair materials into a heated area hours ahead of time to let them warm up. If you're forced to work outside, the temperature should be at least 65° F. At lower temperatures, the resins will take longer to harden, and they may never develop their full strength. If the part is badly damaged—say you're reassembling a bumper cover that's been broken into two or more pieces—it may be necessary to repeat the grinding/ squeegeeing/cloth routine several times to achieve adequate strength. You may need to fasten the bumper cover down to something to hold the correct shape and alignment while the resins cure. Once they're set, there's no chance of realigning them—so work carefully. Because you're working on the back of the part, be thorough. You also need to take care while working around the areas where the bumper is bolted to the mounting to be sure it fits properly when you're finished.

Once the reinforced area has set completely, check the stiffness of all the repaired areas and make sure the cover fits over the bumper. It may be necessary to repeat the preceding steps, adding one or more additional layers of fiberglass cloth to build up sufficient strength. Just be sure to roughen the surface with sandpaper to get good grip. Allow adequate time for each layer to cure before adding the next. Flip the part over, and use the 36-grit disc to grind about 1½ in. around the damaged area to remove paint and create a rough, tapered low area for the filler to adhere to (Fig. 3). After grinding, featheredge the paint around the damage using the random-orbit sander and a 180-grit disc or a rubber-block sander. Featheredge for several inches surrounding the repair. Blow dust and debris away with an air hose. Squeeze out equal parts of the filler and mix it thoroughly. Use adhesion promoter (Fig. 4) on the poly-

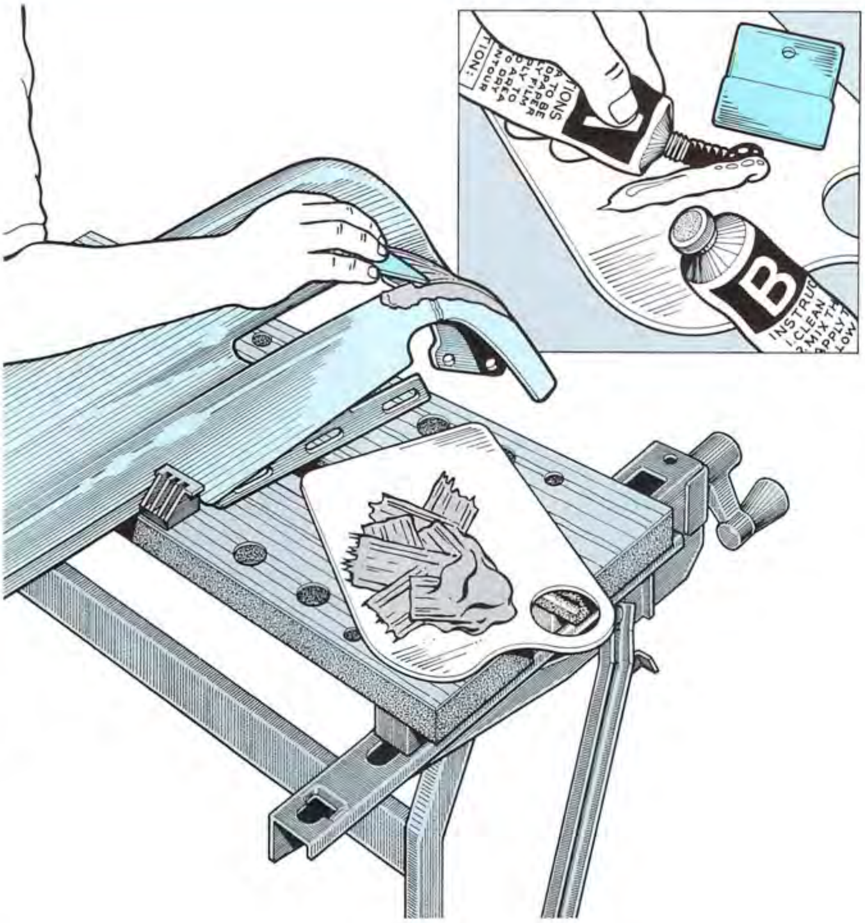

(Fig. 5) Once your bumper cover has been repaired structurally, it's time to begin the cosmetic fix. First, sand the surface. Then, mix equal amounts of the 2-part filler material and use a plastic squeegee to spread it over the repair area.

olefin parts and then spread the filler on the repair, taking care to leave it slightly higher than the surrounding area. If you've ever applied polyester body filler or glazing putty to a repaired steel body panel, you'll feel right at home (Fig. 5).

Going for the finish

After the filler has cured, use 240-grit discs to sand the filler flush with the undamaged surface. Fill any pinholes or other imperfections by applying putty with a plastic spreader (Fig. 6). After waiting at least 15 minutes for the putty to cure, sand the surface with your rubber block and 80-grit paper to achieve a smooth contour. Sand out the 80-grit scratches with 120-grit paper. Then, apply a double wet coat of the coating material, and wait 10 minutes for the coating to "flash" dry. Then, apply a second coat. After 45 minutes of drying time, lightly scuff sand the area by hand with 320-grit paper, being careful not to break through the coating. The bumper is then ready to paint using color coats recommended for plastic body parts.

The painting operation will probably have to be done at a body shop, but you can prep the part by sanding its entire paintable surface with 400-grit wet sandpaper. Minor imperfections can be filled with more repair putty. The body shop will then have to prime it and spray it with a paint that has a special additive to make it slightly flexible.

If you're tempted to just touch up the panel in the affected areas with a spray can of aerosol touchup paint from the auto parts store, fine. Just remember that the panel is flexible, and the touchup paint won't last. In addition, it will be very difficult to feather into the existing flexible paint without leaving an ugly, low-gloss edge.

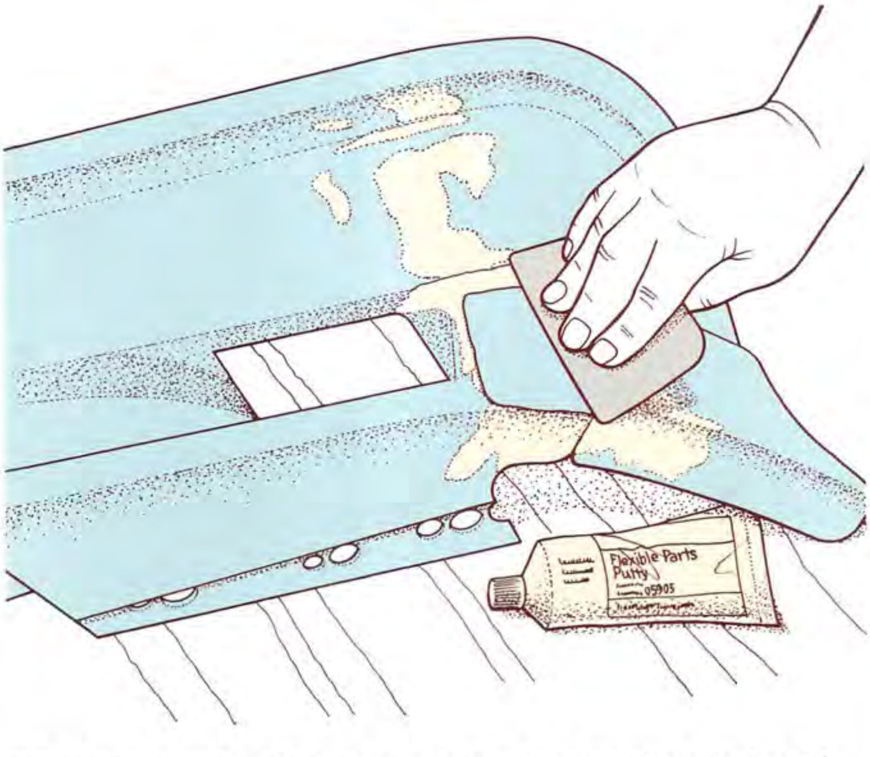

(Fig. 6) The spot repair putty can then be used to fill in scratches, pinholes and small imperfections. It's easier to sand than the repair epoxy.

MODERATE

Hushing Interior Squeaks And Rattles

You're coming to the end of a gas tank's worth of superslab, the pavement as fresh and as smooth as your brother-in-law's balding pate. A quick trip to the Men's and a tankful of gas get you rolling again, this time on a secondary road that's not so smooth. With the new bumps comes a minor rattle that starts under the dash. Within an hour it's as annoying as myelin being flayed from the nerve endings in your inner ears. This is why the term "rattletrap" was coined. You might even begin thinking of the old story about the disgruntled assembly line worker who used to hang empty beer cans inside body panels to sabotage the car company's reputation.

Location, Location, Location

These annoyances can be easily fixed, but finding them is a challenge that requires focused concentration. We once worked with a mechanic who was hearing impaired. He wore an old-fashioned hearing aid that had an earpiece and a wire that ran down to the amplifier, which he kept in his shirt pocket. Whenever the other guys in the shop were presented with a mysterious noise, they'd get him to take a ride with them and he'd pass the amp over the likely spots until he zeroed in on the offending area. Similarly, you can take a piece of heater or garden hose and hold it to your ear. Enlist a helper to drive over the type of pavement that elicits the unpleasant sounds while you climb around,

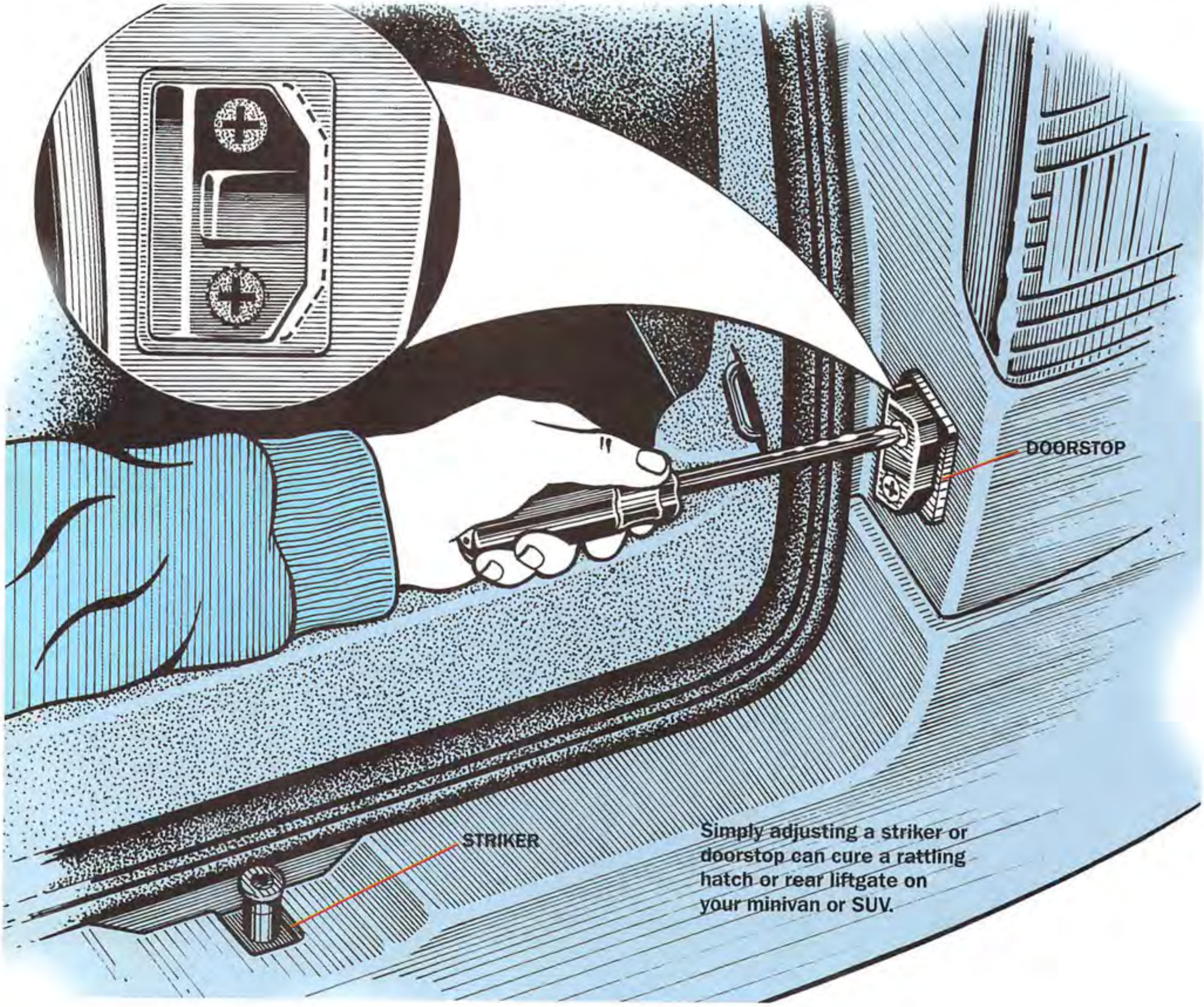

Simply adjusting a striker or doorstop can cure a rattling hatch or rear liftgate on your minivan or SUV.

hose to ear. Or, use a stethoscope (electronic versions are available).

At this point, we should mention the obvious: Make sure the noise isn't being caused by something you've put in the car and forgotten. We've found rattles that turned out to be a socket set in a steel box, other assorted tools, a loose fire extinguisher, etc.

Speaking of loose cargo, are the jack, lug wrench and spare properly secured? This problem will be especially noticeable in vans and SUVs. It can take awhile to figure out what that diagram inside the jack compartment cover is trying to tell you to do, but it's worth looking.

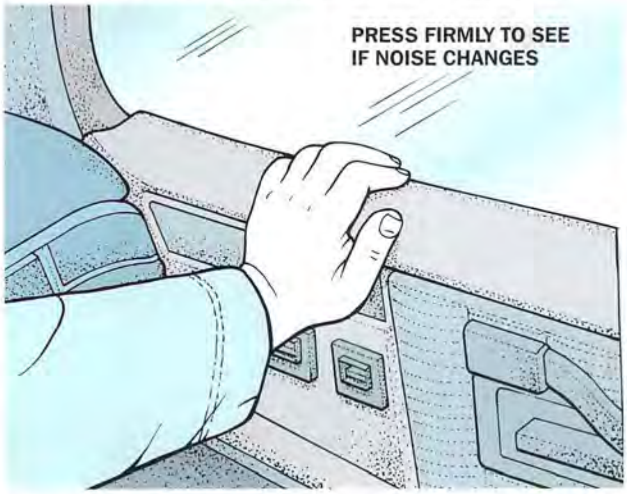
PRESS FIRMLY TO SEE IF NOISE CHANGES

While a patient helper drives, press any potential sources of rattles, such as a door panel.

Pressure Points

Now that you have a general idea of the location, it's time to go hands-on. Press on panels, doors, windows, handles and trim until the sound changes. Sometimes this will require quite a push, especially if you suspect the dashboard or a door panel. Beating on the dash with your fist may also cause a change in the noise, but you've probably done that already as a means of venting your frustration. Besides aural and tactile, this is also a visual inspection. Glove compartment and console doors typically close against little rubber bumpers. Is one missing? Do they look dried out and shrunken? How about the window channels? Are they worn out? Gone altogether? Also look for missing screws.

Once you think you've found the problem, try wedging cardboard or a wooden tongue depressor between the offending components to eliminate vibration.

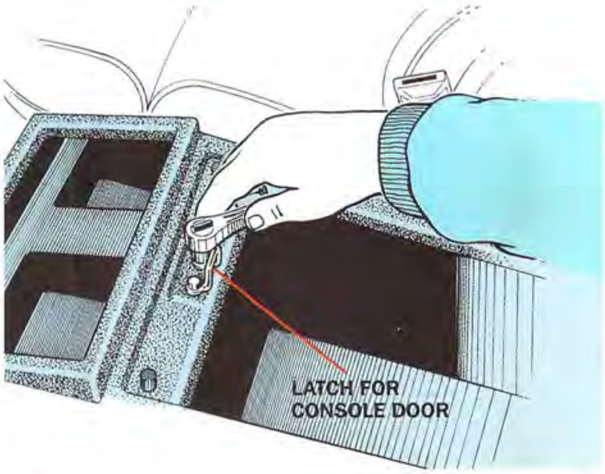

A console's door latch may need minor adjusting.

Close Range

Dashboard noises are particularly annoying. After all, they're right in your face. They may be coming from any of those places where metal and plastic meet, from components such as timers and relays that dangle from the wiring harness, or from heater and air-conditioning control cables. This wouldn't be so bad except that all of this is hidden, requiring you to remove trim panels.

The pushing and pounding routine is the first step. You may find that a high-frequency rattle or buzz is coming from the interface between a piece of trim and the dash itself, or between the dash and the instrument panel. If screws hold these pieces together, try tightening them. If that doesn't work, remove the screws and the component, cut pieces of foam to fit in the seams, and reinstall.

You might be able to find a rattle in the dashboard by reaching underneath and shaking whatever bundles of wires are there. Sometimes you can improve access to the nether regions of the dash by removing the glove compartment liner.

Older-style window mechanisms

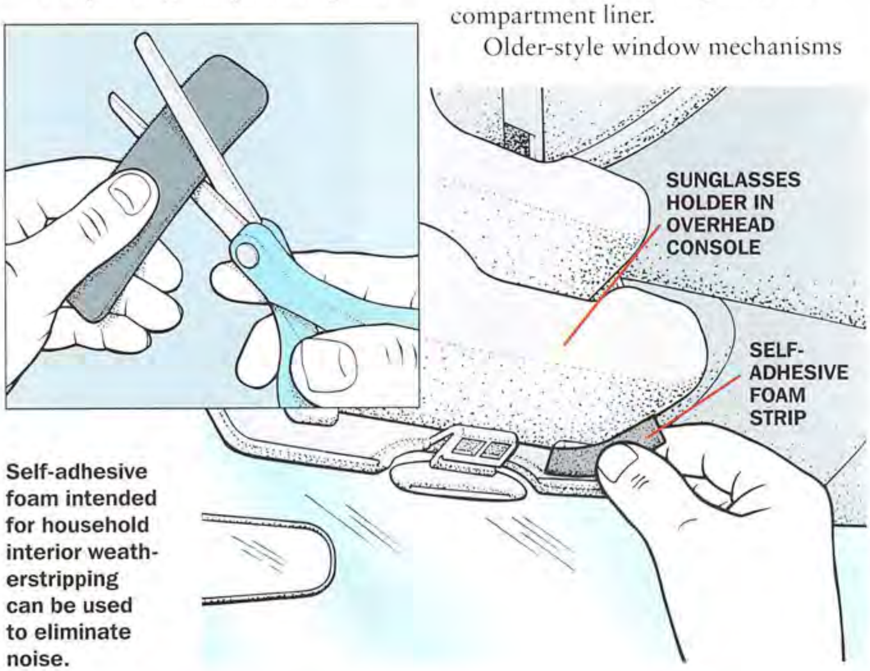
SUNGLASSES HOLDER IN OVERHEAD CONSOLE
SELF-ADHESIVE FOAM STRIP

Self-adhesive foam intended for household interior weatherstripping can be used to eliminate noise.

can become loose and create loud rattles. Power lock linkage is another possibility, as are speakers, especially if they've been added on (stereo shops don't employ NVH engineers). Giving a door a solid blow while it's open, or just shaking it vigorously, may confirm your suspicions as to the location, but you won't really know what's going on without a look-see. You have to remove the door panel, which can be tricky. Check the factory procedures. Special tools are available to help you get under the door panel and pop the clips or push-in plastic plugs with as little damage as possible.

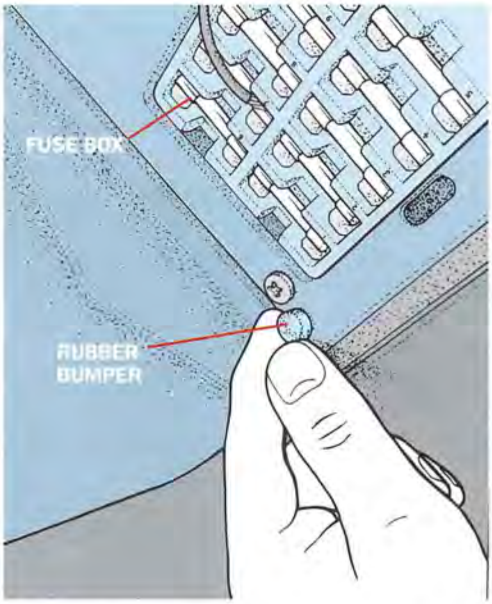

Missing or compressed rubber cushions can let interior panels vibrate.

Likely Suspects

Latches, say, for the glove compartment or console lid, usually can be adjusted. Move the latch in the direction that makes the fit as tight as possible while still being lockable.

Those overhead storage compartments for glasses and garage door openers are sure convenient, but they're made of hard plastic and are likely to generate an excruciating buzz. Unfortunately, the catches on these aren't adjustable. Press various spots while driving to isolate the offending seam, then cut and glue in foam strips of a suitable thickness to make the mating tight.

Sometimes an unoccupied passenger seat will produce a rattle that's downright thunderous. What's the problem? Too much space in the tracks. You should be able to fix this by taking up extra clearance. Loosen the bolts that hold the seat tracks down, then spread them as far as you can and retighten.

Tailgates, whether on a pickup or a truck-based SUV, can sound like a handful of bolts in a cement mixer. Typically, the rubber rings on the hinged arms that keep the tailgate from falling away dry out or break off. As a result, the heavy steel arms come in contact with both the body and the gate. Replace the rubber, or substitute rubber-coated cables for the hinged arms.

Hatches on minivans, SUVs and hatchbacks can drive you crazy. They are big parts of the vehicle structure and naturally make noise as the body twists and bends. Okay, open 'er up and take a look. Are there rubber wedges intended to impede movement? What kind of shape are they in? The sliding doors of vans also can be noisy. Many versions use steel rollers riding on steel tracks—the potential for trouble is obvious. Try to take up excess clearance by loosening the roller-bracket bolts, then moving the bracket down.

Most doors latch onto striker bolts that are surrounded by thin plastic or nylon wear tubes. When these disintegrate, there'll be excessive clearance in the latching mechanism and a good chance of a heavy rattle. For most vehicles, you can buy replacement striker bolts from the aftermarket, but anything exotic will require a trip to the dealer.

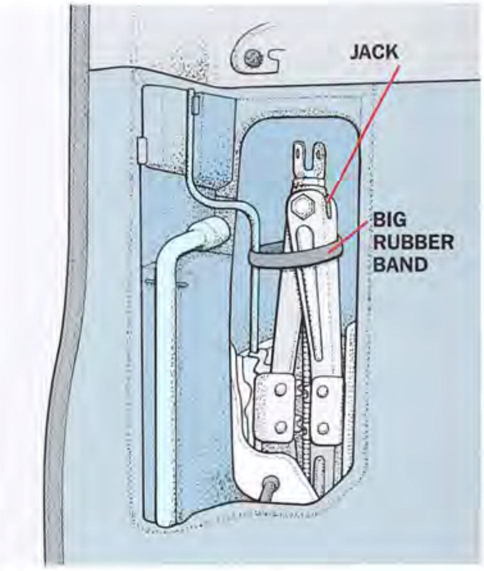

Try using a big rubber band or bungee cord to secure tools and jacks.

Sometimes there's a noise that is not so much a squeak as a creak, which may be the result of torsional flexing in the unibody caused by undulations in the road. You can try jacking each of the corners one at a time and bouncing your own weight around inside until you hear the noise (sometimes you can actually feel it). See if the screws or push-in fasteners that hold the trim panel over the area in question are tight. If that's not the problem, remove the panel and try to duplicate the sound. Can't do it? Then cut foam strips or use foam tape to insulate the panel from the sheetmetal.

Stuffing And Strapping

What materials make sense for insulating one component from another? Whatever will fit easily and stand up to the environment. Try some dense foam rubber. A piece of an old wet suit, perhaps? Or, household door and window weatherstrip foam tape with a peel-off backing. For glue, automotive weatherstrip adhesive is the standard. Finally, get an assortment of nylon wire ties—one of mankind's great inventions—with which you can keep things away from other things, thus restoring blessed silence.

EASY

Removing Bumper Stickers, Window Tinting And Pinstripes

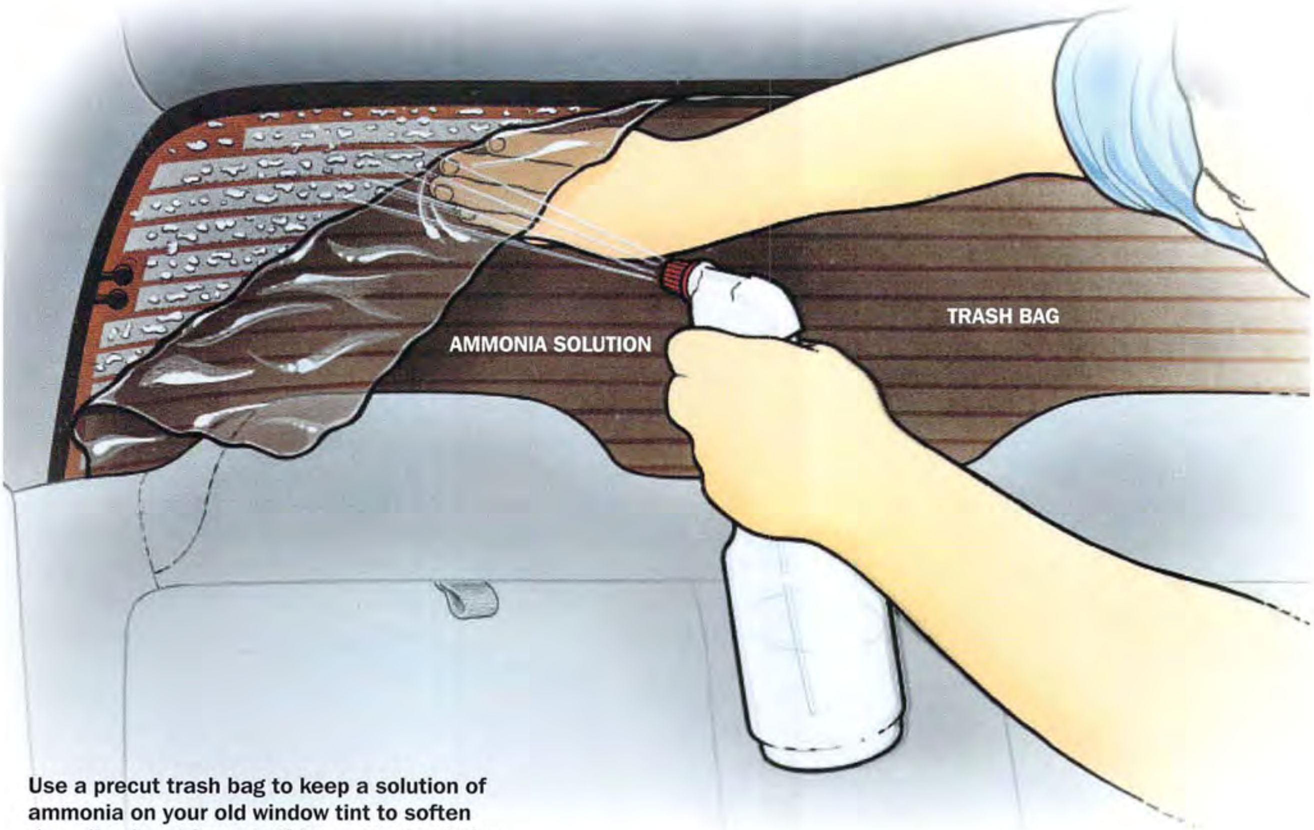

Use a precut trash bag to keep a solution of ammonia on your old window tint to soften the adhesive. After a half-hour you should be able to peel it off.

It's just like the newspaper ad said: one owner, low mileage, good rubber, with all the paperwork from proper dealership maintenance since it was new. There's only one problem—the bumper stickers, all 12 of them, the Day-Glo orange pinstriping, and the peeling, amateurishly applied purple window tinting. (Okay, three problems.) You really don't want the neighbors thinking that you Question Authority or Brake for Hallucinations. But the price is right—if only you weren't ashamed to drive it home. You peel up the corner of one of the bumper stickers with your fingernail, and it breaks off in small slivers and leaves a gooey residue behind. Maybe you should just look for another, less sticky car.

Peel And Unstick

Relax. Removing all that pressure-sensitive film is straightforward. Most of it should just peel off. But time and the ravages of ultraviolet rays from exposure to lots of sunlight will dry out the adhesive and make the film too brittle to peel. Worse, the half-dried adhesive can't be scraped off without the danger of damaging the paint. And that tinting film that's bubbling and wrinkling on the rear window offers up another problem: Scraping it off with something as innocuous as a credit card will certainly damage the silk-screened defroster grid. Here are some tried-and-true, relatively simple but effective techniques.

Most stickers should peel off with the aid of a hair dryer or just plain sunshine.

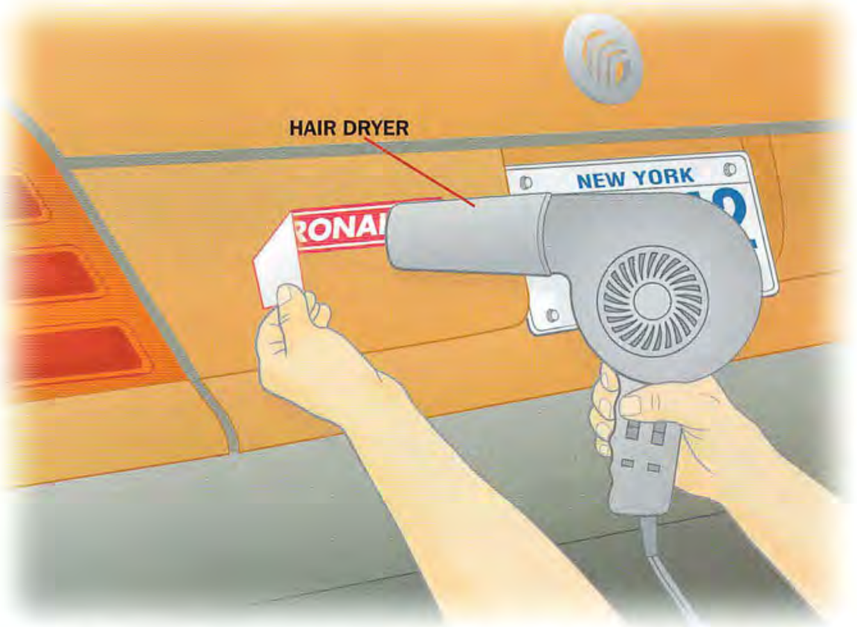

Bumper Stickers

Peel 'em off. This is easiest to do on a sunny day. Just park so the offending sticker is in the sun, wait a half-hour or so for things to get hot, and pry up a corner with a fingernail. Go slowly. Very slowly if the sticker is old and fragile. Sometimes it helps to bend the sticker back nearly flat—by doing it this way, rather than pulling at right angles, you're less likely to rip the film.

It's the dead of winter and you don't want that Grateful Dead dancing bear logo your teenager put on your car last weekend? Try heating the panel with a hair dryer or a heat gun for a couple of minutes to soften the sticker's vinyl film and the adhesive.

Still getting off no more than dime-size shreds of vinyl film? You might try to soften the sun-damaged film and its dried-out adhesive with solvent. We like to use citrus-based cleaners like Goo Gone first—they're less likely to damage paint. Use full-strength and let it soak in for a few minutes. If it's hot, you may need to cover the area with a plastic bag to keep the solvent from evaporating. More stubborn cases might benefit from a lacquer thinner or another serious solvent. This may also remove paint—test it on a hidden area first and don't tell us we didn't warn you. Still stuck? Read on.

Stubborn Cases

What about those giant woodgrain panels on the side of your minivan or vintage station wagon? They're too big to peel by hand unless the peeling is easy—which it probably isn't. 3M, purveyor of all sorts of products that make life easy for professional automotive painters, has an aerosol product called Woodgrain & Stripe Remover. It softens the vinyl so it won't tear, and loosens the glue. Warning: Apply this product carefully, because overspray will loosen the adhesive that holds down exterior trim, rubber gaskets and vinyl roofs. And it probably will delaminate your sneakers if you're not careful. Read the label on the can, and mask off any areas you don't want softened.

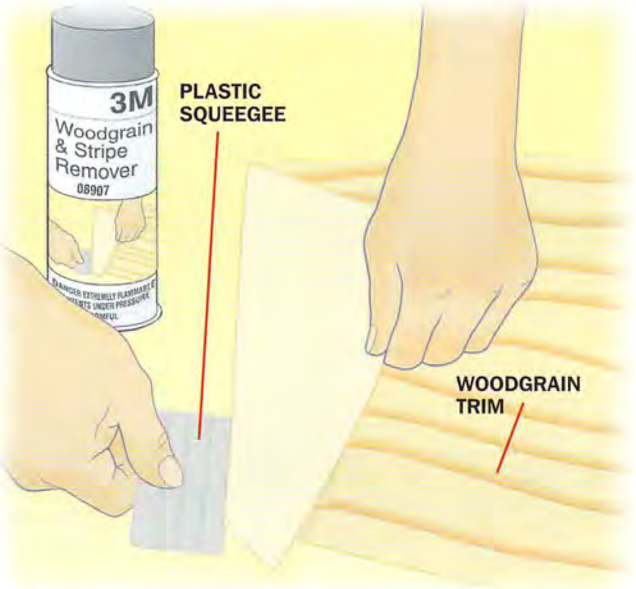

Use 3M aerosol solvent to loosen the film and a plastic squeegee or credit card to help peel off stickers and trim.

Step Back From the Zebra

Pinstripes come in two varieties: tape and painted. Tape stripes will respond to any of the techniques we've already suggested. Another method is to use the edge of a credit card (preferably one that's expired or one of those unsolicited cards you get in the mail) to help peel the stripes up. Leftover adhesive can be removed with solvents.

Got painted stripes? Learn to like them or prepare to have the panel repainted because it's virtually impossible to remove them without damaging the paint.

3M also makes a product called a Stripe Off Wheel, intended for high-speed removal of pinstriping and trim. It consists of expanded polyurethane foam on a wheel that can be chucked into a hand drill. When it is applied lightly to the stripe, the wheel wipes it out like the rubber eraser you used in grade school. The vinyl crumbles, the adhesive rolls up in little balls, and all that's left is a smudge. The product is easy to use—but there are two important caveats. Be sure the panel is very clean or the dirt will scratch the paint. And use a light touch with a low-speed drill—no air tools or Dremels.

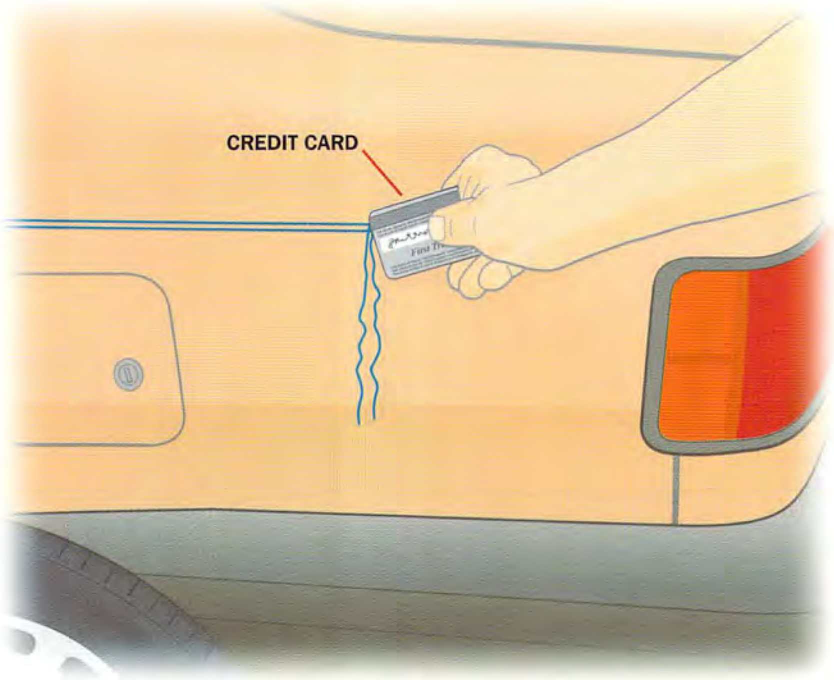

Tape stripes can be peeled back with the edge of a credit card. Painted stripes are tougher to deal with.

Final Prep

Use Goo Gone or some other citrus cleaner to remove the last of the adhesive. Have patience—give the solvent time to work because you don't want to scratch the paint. Then, use a mild polishing compound like swirl mark eliminator (Meguiar's ScratchX is a good one) to remove any minor scratches. To finish, use a cleaner wax to give the panel a consistent, shiny surface.

I'll Be Seein' Ya

Bodywork primo and waxed? Now you need to pull off that faded, bubbled window tint. The side windows are easy—just get out a single-edge razor blade and scrape them clean. Not so the backlite. The silk-screened defroster grid is way too easy to damage. You might try using the hair dryer trick on it, but this polyester (not vinyl) film is probably too brittle to peel. Here's a technique we learned from a local trim shop.

Start by slitting open a large plastic trash bag into a single large panel—any color is fine. Spray a little water on the outside of the back window glass and use it to temporarily stick the bag to the window. Use your single-edge blade to cut the bag approximately 1 in. smaller than the glass. Add a couple of spoonfuls of household ammonia to your water sprayer and spray it on the tinting film. (The ammonia will stink—do this outdoors with the car doors open.) Use this to stick the bag to the film. If needed, add some more spray between the bag and the film to keep things wet and sticking. In about 20 minutes, the adhesive will have softened, and you should be able to peel off the tint more easily. Don't scrape, not even with your credit card. Remove the remaining adhesive with a very soft cloth and more ammonia or Goo Gone. If you must rub the glass to lift off old adhesive, rub along the length of the heater grid conductors, not across them.

A Stripe Off Wheel will remove the most stubborn vinyl films, but large areas will take some time. Use a low-speed drill.

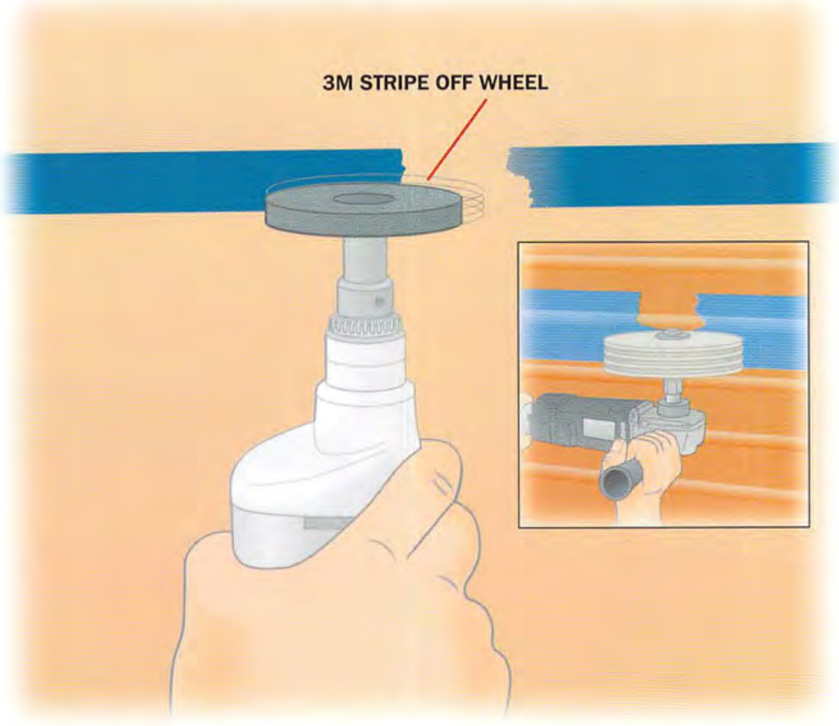

MODERATE

Freshening Up Your Musty Interior

Saturday night, and you're ready to prowl. Actually, it's time to take your potential in-laws out to dinner, an operation possibly more angst-ridden than that first prom date. At least tonight you can drive your own car instead of having Mom squire you around in her minivan. Showered and shaved, you head out to pick up your intended and her kin. You flip on the a/c only to have the passenger compartment fill with the sweet aroma of old sweat socks as the dash vents pump in cold air. Time to pick up a new pine tree air freshener at the carwash? No. Better to get to the bottom of the problem, find the source of that nasty odor and kill it.

Smell Patrol

We know a man who was getting ready to sell his beloved vintage Corvette because the interior reeked. Then he found the cause: A slim package of meat had fallen out of a grocery bag and ended up under the spare tire. While your car may not present such an extreme example, the first step in deodorizing is to remove everything from the interior that's not bolted down or welded in place. Besides the floor mats, that means all the flotsam and jetsam under the seats, the fast-food wrappers from the console and the odd dead chipmunk from the glove compartment. Then, do the same with the trunk.

Now feel around for dampness. If any area is wet, a repair is called for. (We'll get to that.) But that still doesn't relieve you of the task at hand, which is to get the inside of that car as clean as possible.

Vacuum thoroughly, of course. Just a few tiny fragments of food can cause olfactory distress. Then, you may want to get your hands on one of those steam cleaners intended for

DRILL CAREFULLY: Smells in your a/c may require the evaporator core to be coated with an antifungal agent. Apply by opening a hole in the plenum and spraying it in.

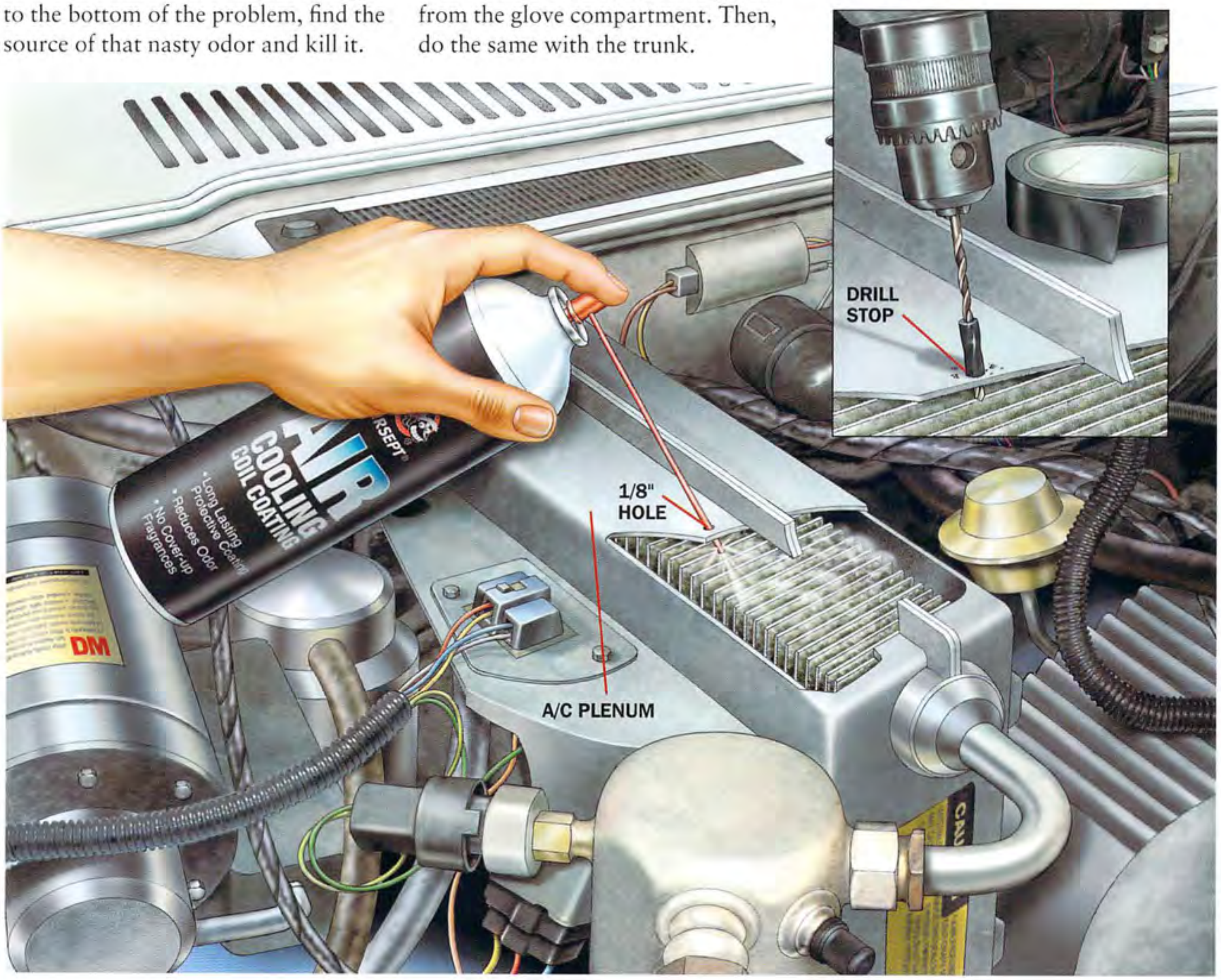

WHAT'S THAT STAIN?: Start by thoroughly cleaning the interior upholstery with detergent to remove any odors and mold spores.

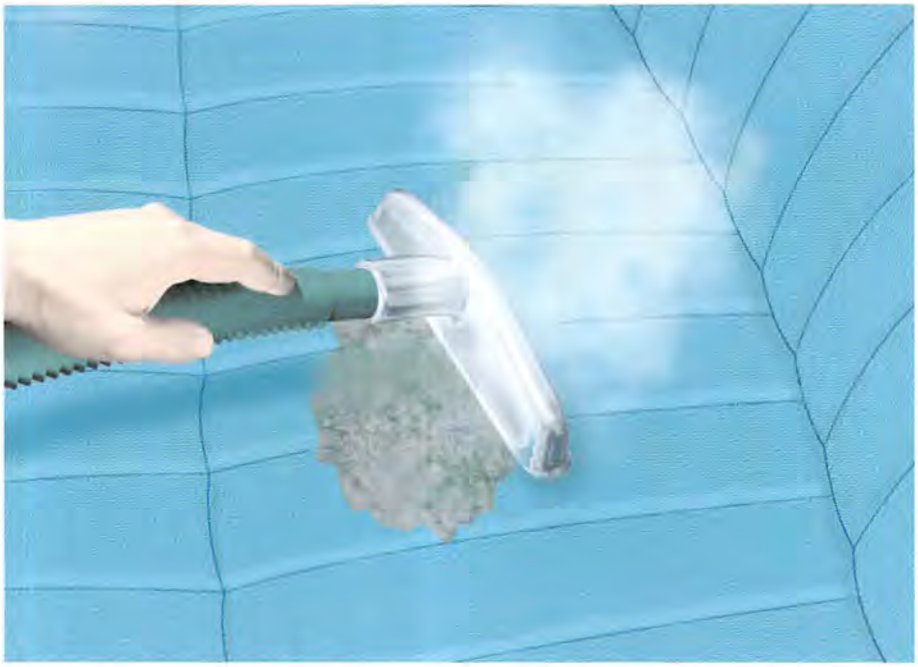

household carpets and upholstery. As they say, "Use as directed." A caveat on headliners: If yours is fabric and doesn't appear to be very well attached to the roof, you probably don't want to be pumping hot vapor and detergent into that tenuous bond, then pulling it downward with vacuum.

Many late models are equipped with cabin air filters, and there's a good chance yours has become odoriferous. Changed yours lately? That job can be quite an exercise in automotive disassembly.

That film on the inside of the glass? It definitely has an aroma, but is practically impervious to ordinary window cleaners. We've found that alcohol really cuts it. Cover the dash and door panels with newspaper (alcohol can cause spots), then douse a clean cloth or pad of paper towels and get to work. An added benefit is that you'll actually be able to see where you're going at night.

When you're finished cleaning everything, leave the car parked out in the sun with the windows open for as long as possible.

Smoke Scream

Smoking in a car with the windows closed is the equivalent of scientifically odorizing the interior. The smell permeates every fiber, even those under the vinyl, and tar coats every surface. A thorough cleaning as already described is about all you can do. Alas, as a used-car dealer we interviewed puts it, "If the vehicle was driven by a heavy smoker, particularly cigars or a pipe, I don't think you can ever get the smell out." He's had a professional automotive deodorizing service treat the cars on his lot, but says he's had only about a 70 percent success rate with smells in general, less with smoke.

Swampy odors

Water provides a habitat that'll grow all kinds of interesting specimens. Think mold, mildew and bacteria.

A common cause of carpet dampness, particularly on the passenger's side, is a clogged air-conditioning condensate drain from the heater/evaporator box. This may be from leaves, twigs and other plant matter that finds its way into the ducts. We've seen it get so bad that you can hear sloshing while you're driving.

Sometimes the drain is simply a hole at the bottom of the box, or it may be a plastic hose or tube that sticks through a hole in the firewall. You've probably noticed a wet spot under the car that appears after you've parked on a hot day—the drain will be directly above that. Use any kind of dull, long probe (a chopstick, an unsharpened pencil, etc.) to clear it out. Something too pointy can damage a rubber drain tube.

HERE'S MUD IN YOUR EYE: Check the a/c evaporator drain line for road debris or clogging.

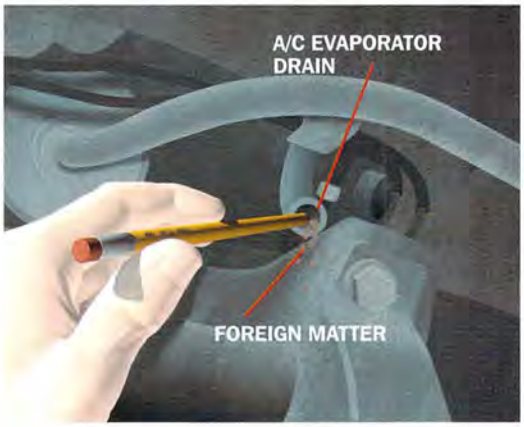

Deteriorated or poorly fitting door, window, trunklid or hatch seals will allow water to seep in. Have somebody hose down the car while you sit inside and look around for drips. Also, look in the trunk or under trim panels for water getting in through the taillight assembly gaskets. Windshield leaks are another possibility, so you may need to visit a professional.

If your vehicle has a spare tire well, it's likely a collection point for water. Remove the spare and check the condition of any rubber plugs in the bottom of the well. You can easily reseal them with RTV silicone caulk.

White out

Household bleach is the most potent weapon against whatever's growing in carpet or upholstery, but you can easily

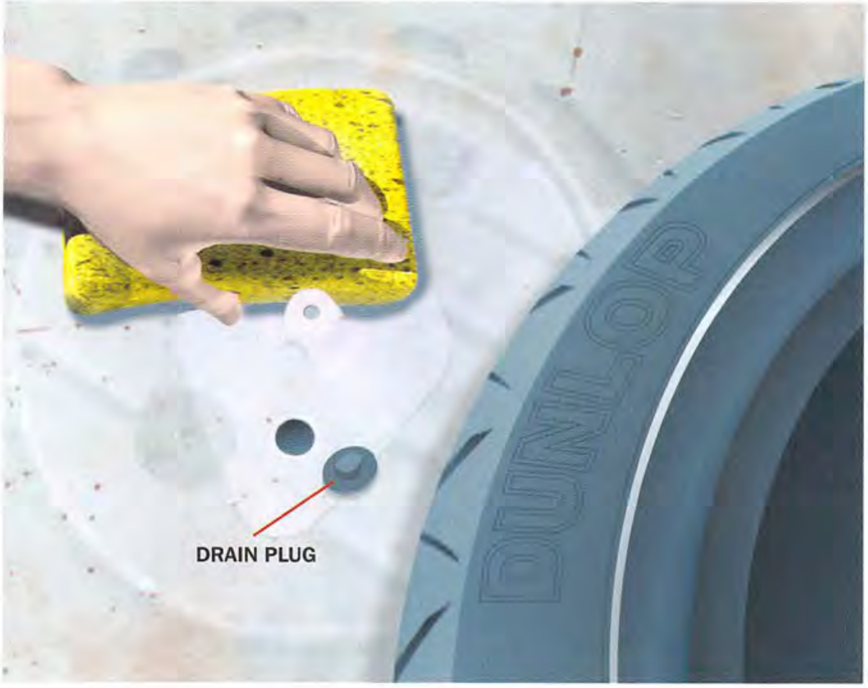

DEAD SEA: Look around for moisture and mold collecting in low spots like the spare tire well or footwell.

take out the dye along with the odor. If the material is a light color, you might want to experiment with a highly diluted solution on an area that's hidden from view and see if you get whitening or spots. If not, sponge it into the suspect area, then use a wet/dry shop vac to extract it.

Locker Room

Sometimes it's easy. You know exactly where the odor is coming from because you get it only through the a/c vents. Condensate moisture that promotes the growth of various microscopic organisms on the air conditioner's evaporator core is the problem. There are two ways to go here. You can buy deodorizers that you spray into the vents, but there's little chance that you'll deliver enough of a dose where it's needed to eradicate flora and microscopic fauna. Professionals inject specially formulated sprays directly into the heater box and really bathe that core. These sprays are available at automotive parts stores and over the Internet, and can be quite expensive. They leave an antifungal coating behind, good for two to three years. Bactine, commonly recommended in automotive folklore, won't cut it.

Another means of keeping the inside of the HVAC system smelling sweet is an afterblower, such as the type sold by AirSept (www.airsept.com). It's an electronic module that you wire into the blower circuit. It keeps the fan running for a predetermined time after you shut off the engine, drying out the fins.

Mouse Condo

Snug, dry and safe from predators—what could be a better place for a mouse to build a nest than a car interior? Once you've smelled that distinctive rodent odor, you'll always recognize it. And this doesn't just happen in vehicles that are left unused for long periods—the busy little devils can make themselves right at home in one night, typically in the glove compartment or HVAC ducts.

Once you've located and vacuumed out all that shredded insulation and paper, you'll find to your disappointment that the odor remains. Try a product intended for pet odors, such as Zep Commercial Pet Spotter.

Gas attack

The most likely source for the odor of gasoline inside your car is the evaporative emissions control system, mandated initially for 1971 models. This collects the fuel that evaporates from the tank, and stores it in a charcoal canister to be metered into the engine intake while the engine is running. The vapor hose from the tank can be cut or become disconnected. The result is fumes escaping into the interior. Find the top of the gas tank, which likely will require that you pull up the trunk mat, then follow the hoses that run from it.

DISNEYLAND?: All your furry little friends will have to find a new home.

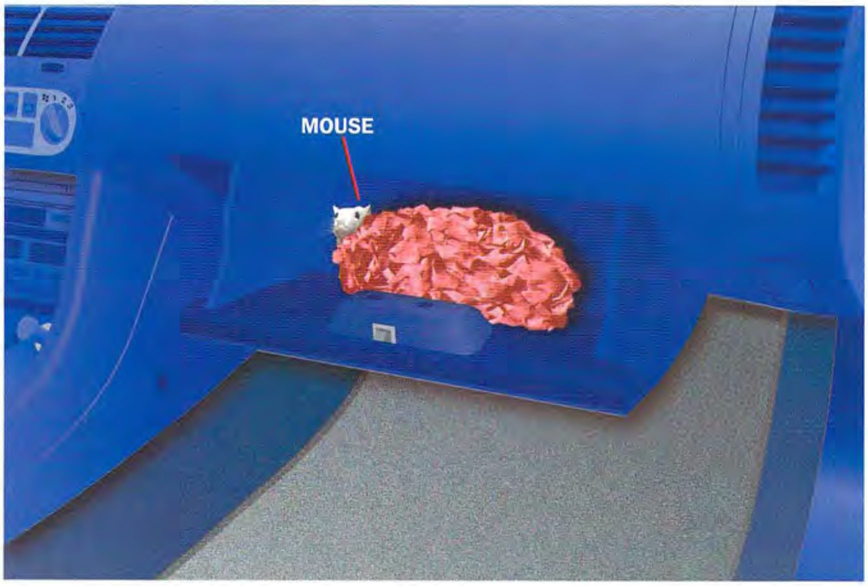

MODERATE

Fixing Hood And Trunk Latches

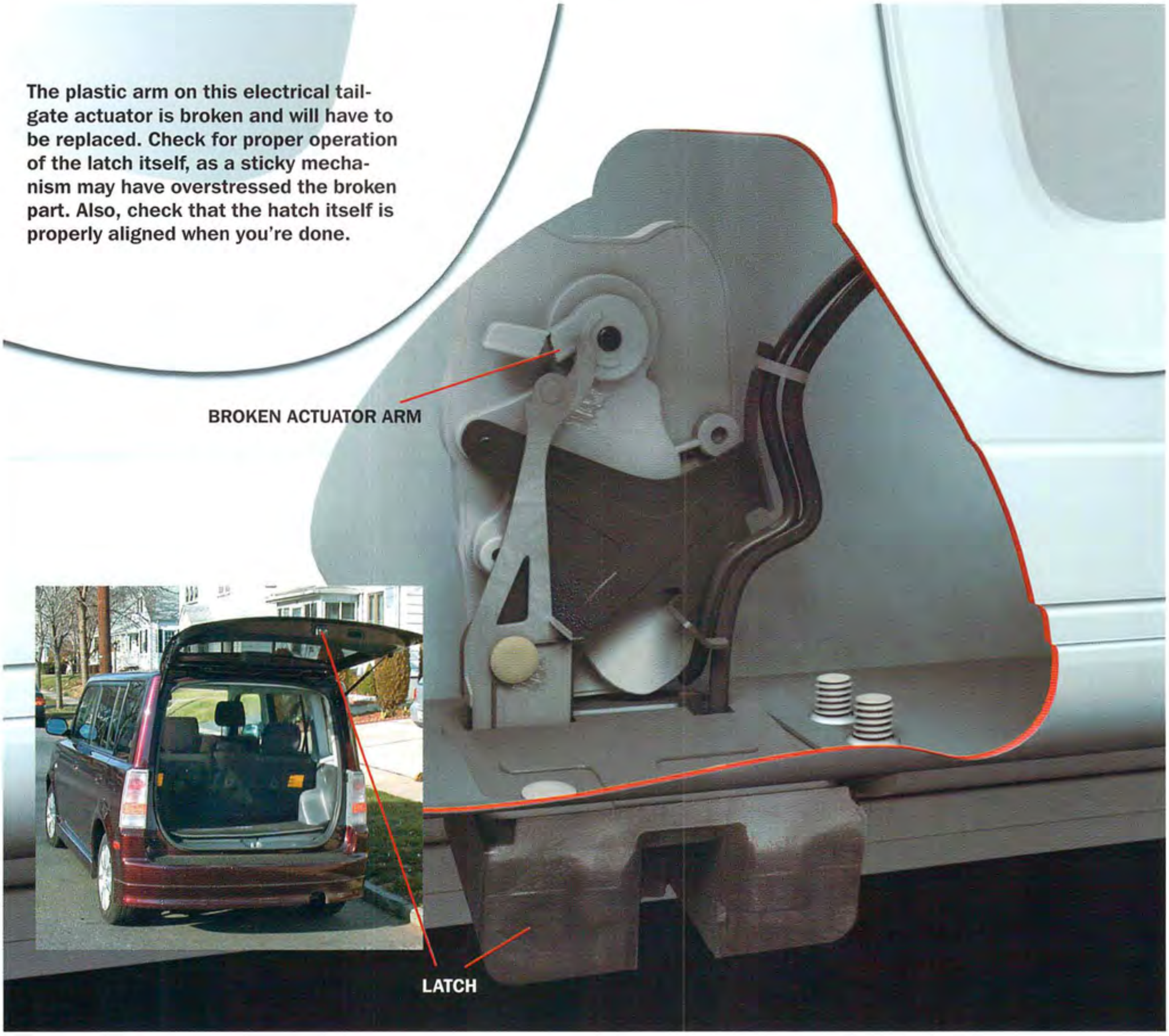

The plastic arm on this electrical tailgate actuator is broken and will have to be replaced. Check for proper operation of the latch itself, as a sticky mechanism may have overstressed the broken part. Also, check that the hatch itself is properly aligned when you're done.

Checking your oil is simple enough—until you attempt to open the hood. You try the remote release, but it seems a little hard to pull. Yank harder, and the familiar clunk of the hood popping up until the safety latch stops it doesn't happen. Your hood release is stuck. The dipstick will have to wait.

Open sesame

Actually, it may not be your hood. Many vehicles in this security-conscious age have remote releases for the trunk, tailgate and even the fuel-filler door. While these may be electrically actuated, the mechanism that latches the tailgate is still mechanical. You may need to use some electrical troubleshooting to repair this end of the system, be it a dead battery in the remote key fob or a logic failure in the body computer's theft-deterrent system. We'll talk about electrical issues in a different Saturday Mechanic.

Slap and tickle

Our most common scenario: Pull the hood release and the hood doesn't budge. But pull the inside release while someone presses the hood down, and the latch trips. There are usually a couple of things, happening concurrently, that cause this.

- The latch mechanism is sticky

from corrosion and/or dirt. Sometimes all it takes to loosen things up is a healthy hosing down with penetrating oil while the mechanism is worked back and forth. Follow up the penetrating oil with spray lithium grease to keep things moving. Don't use silicone spray, which can wind up sucked into your air cleaner, contaminating your O2 sensors.

- The cable that leads back to the interior release may be stretched or frayed. Sometimes the stop on the end of the cable is adjustable—it may have slipped a few millimeters. Readjust it appropriately. Then consider why the cable is stretching. What's binding up and making you pull too hard and overstress things? If the cable feels gritty on operation, try lubricating it. Use an aerosol spray lubricant—not penetrating oil, which is too light—with a straw-style nozzle. Perch the straw so the end butts against the end of the cable sheath and right alongside the inner cable. Fold a clean shop rag over the confluence of the straw and cable, and pinch tightly with your fingers. When you depress the nozzle on the aerosol can, lubricant will be forced between the cable and sheath, making the action of the cable much smoother. If you have compressed air available, you can use a blast of it to force the lube deeper into the sheath. This technique works for almost any Bowden-style cable, including heater-control cables, the throttle on your lawnmower and the brake cable on little Suzy's metallic purple banana-seat bicycle.
- If the cable is starting to corrode and has little steel threads poking out, it has already started to fail. Swap it out for a new one immediately. Splash the new one with some lubricant now and then to keep this from happening down the road.

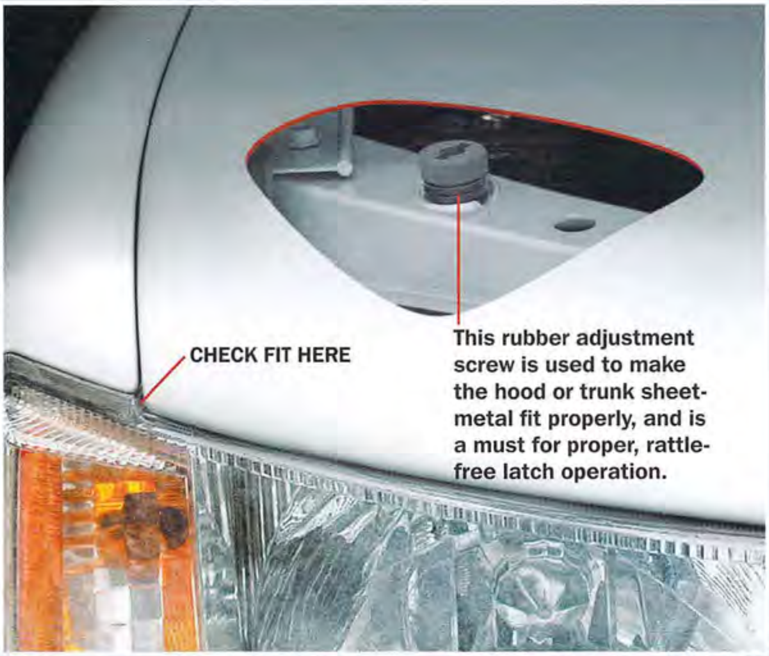

This rubber adjustment screw is used to make the hood or trunk sheetmetal fit properly, and is a must for proper, rattle-free latch operation.

FITNESS TRAINING

If the hood or latch is adjusted incorrectly, it may be difficult to open. This is often seen after a minor fender bender or after a fender or grille has been replaced (for instance, for an engine swap, a radiator swap or major engine work). Start by loosening all the bolts that hold the latch mechanism to the radiator core support, and see if the latch is adjustable. If so, remove the mechanism. Now lower the hood until it sits firmly on its rubber bumpers. Check to see that it fits correctly in the hood opening, with even gaps along the fenders and grille. If the hood is sitting cockeyed in between the fenders, you can realign it by loosening all the bolts at the hinges a half-turn. Jockey it around until it sits square in its intended place and tighten the hinge bolts. Screw the bumpers up or down until the hood sits correctly with respect to the fenders and grille in a vertical plane. Now you can reinstall the latch, snugging up the bolts so the mechanism is just high enough to give a quarter-inch or so of resistance before it closes fully. It shouldn't be necessary to slam the hood, and it should pop up readily when you pull the release. If the vehicle has been in an accident that involved front-end damage, you may have to compromise between an exact fit of the hood and the ability to open and close it easily.

Failure is not an option

Another scenario is that you pull on the release and it moves freely—because the cable is already broken. Now you're screwed. You have no easy way to open the hood, which can make for a long time between tuneups and oil changes. If the cable has come adrift at the handle inside the passenger compartment, you may be able to reattach it. Or, perhaps you can grab the cable with a pair of pliers and yank on it to open the hood. Then you'll be able to access the other end of the cable and replace it. If the failure is at the hood end, check with a flashlight and a small mirror. (We use a small mirror with a swiveling, extendible handle.) Looking through the grille, you may be able to see the lever that the cable pulls on in the latch mechanism. If this is the case, try inserting a long screwdriver or other implement into the gap and tripping the latch. I've even resorted to fabricating a special tool from a piece of flat steel. Find a similar car, open the hood, and figure out how to bend the steel and then how to thread it into the latch and trip it. Practice on your example vehicle to make it easy.

Try to loosen or remove the grille to make access easier. Try not to do this where the local police might take an interest in your activities unless you have your ID and the registration handy, especially late at night.

- If all else fails, it may be necessary to drill a hole in the hood sheetmetal to trip the latch, but this is a last resort. Check the vehicle's shop manual for insight into opening the hood in the event of a cable failure.

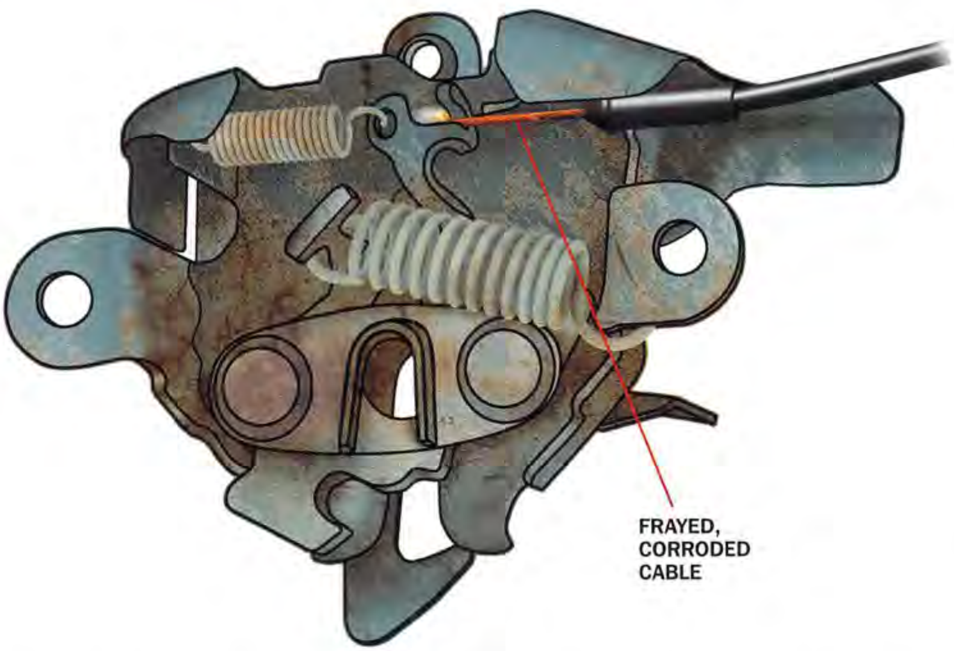

Corrosion and dirt have made this hood latch difficult to operate. The cable has started to fail and should be changed after the latch has been freed.

Trunk line

Like the hood, the trunk should be adjusted on its hinges and rubber stops to close with a proper fit. (See "Fitness Training," page 116.) The latch mechanism can then be tweaked to achieve proper function. If the mechanical pushbutton or key tumbler is hard to actuate, start with adjustments and lubrication.

Many vehicles use electrical actuators for the trunk or hatch. If the electrical actuator is kaput, you'll need to troubleshoot that with a digital voltmeter (rated at 5 megohm, to be safe) and a schematic. But if you can hear the actuator clicking and the latch doesn't open, you've got a hardware problem.

If there's a keyhole, you should be able to open the lid or hatch with a key. There's an urban legend about the owner of the high-priced luxury car who got a flat tire and was unable to open the trunk and access the spare because the remote-control trunk key fob had dead batteries. After waiting 2 hours for the tow truck to arrive, the owner then watched the truck driver easily open the trunk—by turning the key in the lock—and then change the tire. Moral: Don't make this any harder than necessary.

Own a minivan? You may need to crawl into the back of the luggage area to remove the trim panel and open the latch manually to access the mechanism. If your sedan has a balky trunk lid, you may need to fold down or remove the rear seat to get anywhere near the mechanism. Check the owner's manual for hints about emergency access to the trunk—there may be a simple cable to pull in case of major failures. All new cars have an emergency cable release meant to prevent a child from getting trapped inside a locked trunk. If you can reach that release by getting around the rear seatback, you can pop the trunk. Once the hatch or trunk is open, you can clean and lube the parts to free things up.

Trunks and hatches often respond to proper fitting and adjustment, just like hoods. Get the sheetmetal to fit properly, and then adjust the latch and striker for proper function. Don't forget to lubricate the out-of-sight part of the latch mechanism. One important part of a tuneup is to lubricate the door, hood and trunk hinges and latches. If your mechanic is neglecting this, you should take 5 minutes and lubricate these parts yourself.

Most of the time, when there is a failure of the latch mechanism, it's a plastic component that has broken. You'll need to go to the dealer for a replacement part. Sticky or hard-to-operate latches usually need nothing more than lubrication and adjustment.

Fill 'er up

Fuel doors usually have a simple bolt-based latch, with a simple cable to operate it. If the cable is gritty or sticky, try lubing it with the shop towel/spray lube trick mentioned earlier. If the bolt doesn't retract enough, look for an adjustment for length at one end of the cable or the other. Lube the hinges as well, and make sure the spring isn't broken or loose.

FULL HOT
FULL COLD

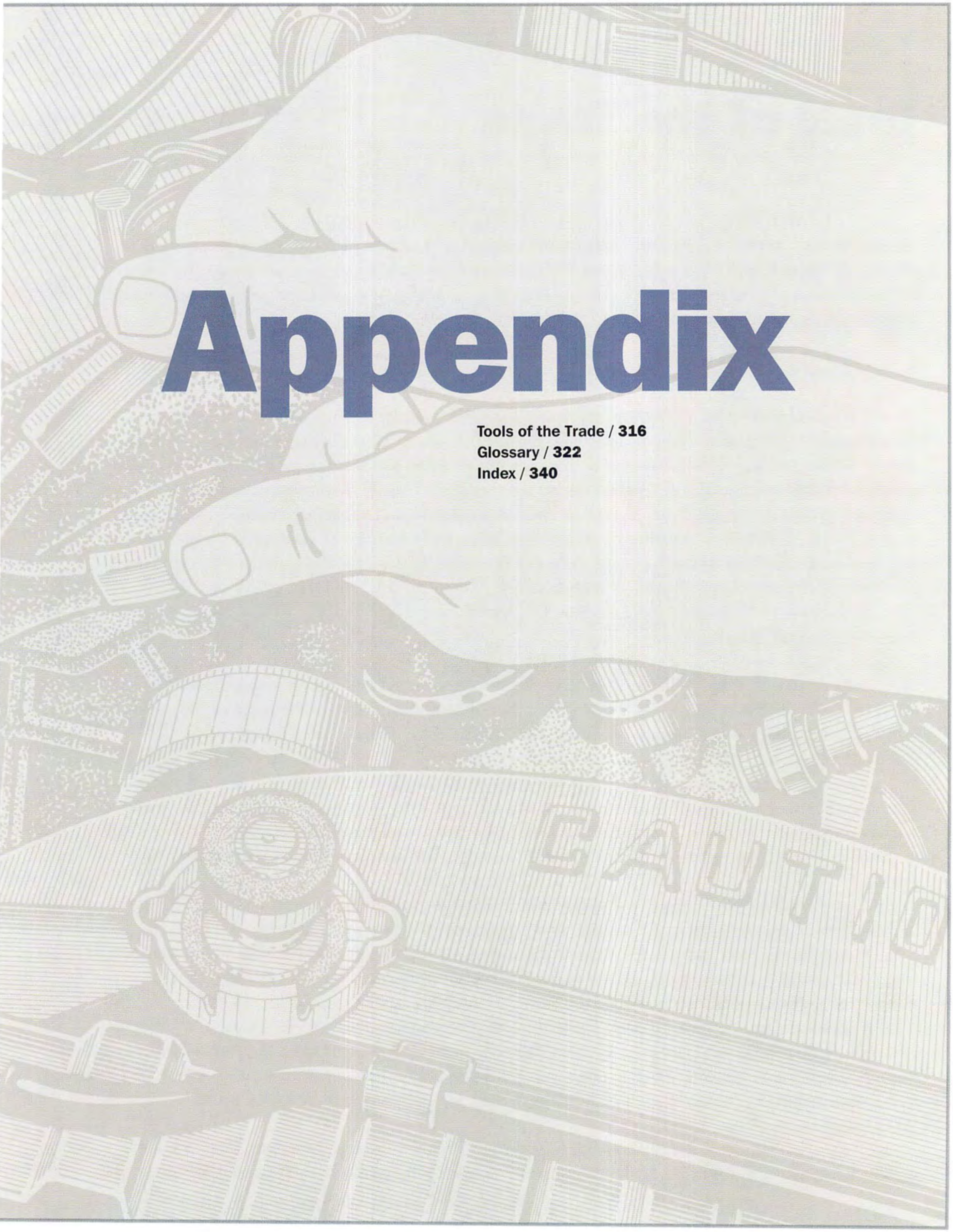

Appendix

Tools of the Trade / **316**
Glossary / **322**
Index / **340**

Tools of the Trade

To a mechanic, "tools" fall into two important categories: "hardware" tools, the implements that actually do the job, and "software" tools, the books and other publications that provide the information necessary to do the job. To work on your own car with any measure of success, you'll have to stock your garage with some of both.

STRATEGIES AND INSTRUCTIONS

There's no good excuse for a driveway mechanic to be without the factory service manual for his automobile. General auto repair manuals that cover a variety of makes are less expensive than several factory manuals if you have more than one car to work on. You can usually buy the factory manual through your car dealer's parts counter. For some cars, more than one instruction book is necessary to fully cover drivetrain as well as body and electronics.The service manuals for all late-models (since 1996) are on websites maintained by the vehicle makers. So instead of a paper manual, you can gain access when you want, typically for about $15 a day. Manuals for older cars are sold by specialty distributors with toll-free numbers.

THE SERVICE MANUAL

Most factory service manuals begin with a general information section, usually with charts that tell you how to interpret your car's VIN number. You'll be able to pinpoint the specs of the engine installed, the plant at which the car was built, and similar data. Other charts interpret body ID plates with paint and interior color codes as well as trim levels. Other general information will include overall specs, fluid capacities of the various systems, and even data on bolt ID and torque. This chapter should also show the correct jack and lift points for raising and supporting your car.

The general information section is often followed by a maintenance and lube chapter. Use this material in conjunction with our car care plans to tailor the maintenance routines to your specific vehicle.

The rest of your service manual is divided into chapters, each of which covers one specific system, such as heating and ventilation, or a group of systems, such as emissions controls.

Most manual chapters include three types of information. First, where applicable, many chapters begin with a short explanation on how a particular system operates. This information can help when it comes to making logical decisions about the possible causes of a given problem.

Next comes diagnostic information. This troubleshooting material can be very general and somewhat brief or it can be incredibly specific and detailed. Typical of the first kind of diagnostic directions are those for engine mechanical diagnosis. Here you'll find lists of possible causes that relate to a symptom. For example, in the 1985 Celebrity manual there are eight potential problem areas listed under the heading "Excessive Oil Loss."

Much more complicated are the diagnostic procedures for driveability problems. Some makers, such as GM, instruct you to begin the diagnosis of a driveability problem by triggering the diagnostic mode of the computer system. You progress from there to troubleshooting individual systems with test lights, ohmmeters, and the like, and sometimes finish up by checking symptom-organized lists of possible causes.

Other manufacturers, notably Chrysler, begin diagnosis of computer-controlled engine driveability problems with visual checks. In either case, you have to read the manual material carefully and do things in the right order. The procedures work on a process-of-elimination basis. If you don't begin at the beginning, you throw the whole diagnostic plan out of whack.

At some point, the troubleshooting of complex systems is sure to include a diagnostic "tree diagram." You begin by performing the test in the box at top, then depending on the results of that first test, you will be led through a series of steps to diagnose your problem. You'll notice that every path eventually concludes with a repair or with a "system OK."

The third type of information, usually at the end of a service manual chapter, is on repairs. It often doesn't include tasks considered simple by professional mechanics.

THE TOOLBOX

How many tools do you need to service your car? Frankly, as many as you can afford. But you can get by with a fairly basic set if you intend to do only maintenance jobs. But if you're going to perform complicated repairs you'll need a lot of special tools, many of them designed to do just one specific job. Check auto parts stores in your area for tool rental peograms.

Plier Tools

From among the vast number of plier tools, you'll need at least three basic types: a conventional slipjoint plier, a pair of side-cutters for cutting and stripping wire and—among many other uses—prying out cotter pins, and a pair of long-nose or needle-nose pliers.

Once you have the basic three you'll eventually want to expand your plier drawer. You might first add a locking plier; the most common type is sold under the Vise-Grip trademark. Among many other possible uses, this tool can be clamped tightly around a stripped bolt for removal or it can be used to hold parts in place for welding. Lineman's pliers with insulated handles are useful for various electrical-system chores. Water-pump pliers offer nearly parallel jaws and lots of leverage. A second pair of long-nose pliers can be sharpened for service on snap rings, or special snap-ring pliers are available as well. Small side-cutters can be of use when working with delicate components. Hose-clamp pliers are grooved to grab the spring-type hose clamps that are standard equipment on many cars. Wire strippers/terminal crimpers are very useful pliers with a row of split circles on the jaws for stripping wires of various gauges and a special crimping area behind the jaws for attaching solderless wire terminals. Tin snips, available for cutting left-hand curves, righthand curves, and straight cuts, are a necessity if you're going to do serious bodywork.

Screwdrivers

For servicing today's cars, you'll need an assortment of flat-blade and Phillips screwdrivers in various sizes. You'll ruin a lot of screwheads if you use the wrong-size driver. In addition to various-sized tips, you'll need some variety in handle lengths. Cheap screwdrivers are not well suited to driving screws, so buy high-quality tools that will not distort the screwheads.

For most cars you'll also need a couple of Torx-style screwdrivers, distinguished by their six-point drives. On some cars, you'll need this type of screwdriver to replace a headlight bulb or seatbelt housing.

Our favorite screwdriver is a high-quality, flexible-shaft ratchet job with interchangeable bits of tool steel. The extra bits fit in the handle, and are available in every size and style you might need. Angle screwdrivers, with a tip mounted perpendicular to the shaft, are useful for screws that can't be reached with anything else.

SOCKET TOOLS

Square-drive tools are probably the most important for someone who intends to service automobiles. To begin, you'll need both 3/8-in.-drive and 1/2-in.-drive ratchets and breaker bars, at least one short and one long extension for each drive, and a complete set of standard sockets for each. If your car is fully metric, you may want to purchase only metric sockets at first. If your car includes both metric and inch sizes, you'll need a set of each.

You'll need a spark-plug socket of either 13/16-in. or 5/8-in. size. Don't just use a deep socket for this purpose. Spark plug sockets have a special rubber insert to hold the insulator, helping prevent breakage while giving you a means to lift the plug away from the engine without dropping it.

A torque wrench is also a necessity, as it should be used for everything from installing wheels to tightening cylinder head or manifold bolts. The most common and least expensive type has a pointer attached to the drive head that rests against a scale near the handle. More expensive types replace the scale and pointer with a dial. A style used by many mechanics has a rotating handle with a scale of torque figures on the tool. You dial in the torque figure you want by turning the handle. Then, when you tighten the bolt, the wrench will click to signal you when you reach the specified figure.

Eventually, you'll want to add universal joints, which can be fitted at the end of an extension; speed handles; T-handles; additional extensions; and a 1/4-in. ratchet, extension, and screwdriver handle for smaller capscrews. Adapters that allow the use of 3/8-in.-drive sockets on 1/2-in.-drive ratchets (or various other combinations) can also be handy at times. You'll want to expand your socket set to include hex or Allen drive sockets, sockets for large slotted screws, deep-well sockets, universal-joint "swivel" sockets, and 1/4-in.-drive sockets, all in both inch and metric sizes.

You'll also want to supplement your standard twelve-point sockets with some eight-, six-, and fourpoint sockets (Fig. 4). Six-point sockets are needed when you have to loosen a bolt with rounded corners where the twelve-point might slip. Our 3/8-in.-drive deep-well sockets are of the six-point type. Thickwall six-point sockets (usually black for identification) are available for use with impact wrenches. Eight- and four-point sockets are for removal of square plugs. If you have the eight-pointers, you really don't need the fours.

If you can't reach the spark plugs on your engine with a conventional plug socket, extension, and ratchet, you may need either a swivel-head plug socket or a 3/8-in.-drive ratchet with a jointed hand that can be rotated to a different angle in respect to its handle.

WRENCHES

A basic set of combination wrenches—box at one end and open end on the other—will get you started in this department. Use a box wrench when you don't have clearance for a socket. Open-end wrenches are really only suited to running down a bolt or backing it out, as they slip quite easily, ruining both the bolthead and your knuckles.

An open-end wrench should be used with the handle angled away from the direction of rotation. Once the bolt is rotated, the wrench is slipped off, slid around in a counterclockwise direction and reinserted. Once you've practiced this technique, an open-end can be used to turn a bolt very quickly. When loosening a fastener with a combination wrench, break it loose with the box end, then flip the wrench around and run out the bolt with the open end.

Once you have a basic wrench set, you'll want to expand with some other types. Those that have a box on both ends generally have a greater offset than the combos, making them useful for certain problem areas. Both box and open-end wrenches are available in a variety of lengths. You'll also find box wrenches shaped like a crescent for inaccessible bolts, box wrenches that ratchet, and sockets attached to a handle with an open-end opposite. We frequently use a very thin open-end "tappet" wrench for any double-nutted bolt or stud or in tight spots.

You'll also want tubing wrenches, which are like a six-point open-end with one side missing. Once slipped over a tube, the five remaining points of the wrench can loosen the soft nut on the end of the tube without rounding it off. Special L-shaped box wrenches that can be driven with a ratchet are available for loosening hard-to-reach distributor clamp bolts.

You might be better off without adjustable crescent wrenches, as these tools are most likely to round off a hexhead. But it's good to have a couple of them around in different sizes for those times when you have to turn a simple-to-reach, not-very-tight bolt, the size of which you are not sure.

HAMMERS

Hammers must be used with great restraint when working on a complicated machine like an automobile. But you will need one from time to time. Begin with a ball-peen for tapping your drift or making a center-punch mark, and supplement it with a rubber mallet for driving things into position. Eventually, you'll want to add a brass hammer for driving lugs into hubs and similar tasks. A plastic-head hammer is useful for dislodging somewhat delicate parts. Finally, you'll want a good hand sledge to drive a large chisel into a large rusty nut.

You'll want to have some chisels, punches, and drifts that you can hit with your hammers. If you work on cars regularly, you'll undoubtedly have to chisel off a rusted nut or bolt before long, probably on a shock absorber. Drifts are used for removing pins from shafts, studs, or hubs. A center punch is necessary to provide a spot to start a drill bit in steel or aluminum.

You'll need only a couple of files at first. A small ignition-point file is good for cleaning up spark plugs or even repairing the threads on a small screw. One medium-size flat file will also come in handy for smoothing rough edges or flattening slightly distorted surfaces. Later, you'll want to add a rattail for smoothing out holes and a threesided file for getting into corners. A hacksaw and a couple of high quality, fine-tooth blades will come in handy if you have to shorten a bolt or fabricate a bracket.

An assortment of tool-steel high quality drill bits and a 1/2-in. or 3/8-in. chuck electric drill is also necessary.

PULLERS AND PLUCKERS

Hub pullers, nut splitters, gear pullers, pickle forks, and other devices of this ilk probably won't be purchased until you need them. Sooner or later you will undoubtedly want them, however.

A good three-pronged gear puller can be used for removing a variety of hubs and gears. This tool is available in several sizes to suit various applications. A puller with slots through which bolts can be inserted is useful for removing any type of hub with threaded holes, such as a steering wheel hub or engine damper pulley. Various other pullers, including slide-hammer devices for removal of rear-drive axles, are also available. Some of them can be rented for one-time use.

A nut splitter is a good substitute for a chisel when you have to remove a rusted nut. A ball-joint spreader and rod-end separator, commonly known as pickle forks, are necessary for suspension-system work. The ball-joint spreader has the wider slot between its prongs.

OIL CHANGE AND LUBE TOOLS

An oil filter wrench is, of course, an absolute necessity for any driveway mechanic. Before you buy, make sure you get one that will work on your car, as there are numerous types available.

Most of the long-handled metal-band jobs work great where there's room for them, but on most cars there isn't. The best of these are adjustable for different type filters.

A handy oil filter wrench is the type that has a sturdy nylon strap attached to a piece of square tubing with a 1/2-in. hole down the center. The strap is placed over the filter and the tubing is rotated counterclockwise until the strap snugs up. Then a 1/2-in. square drive extension and ratchet are used to continue rotating the tubing until the filter is loosened.

PRECISION TOOLS

Every home mechanic should have at least a good pair of vernier calipers for determining the size of pins, bolts, and other similar items as well as a set of feeler gauges and a wire gauge for checking spark plugs. If you're going to rebuild engines, you'll need much more, including micrometers, snap gauges, a dial indicator and magnetic base, a depth micrometer, and—ideally—a dial-bore gauge. A good ring compressor will also be needed for engine work. The best type is a tapered sleeve for each bore size. If you plan on assembling cylinder heads, you'll need a valve-spring compressor that will work on your engine. Those for pushrod engines won't work on overhead cam engines, some of which require specific adapters.

The dial indicator and magnetic base are probably worth having even if you never touch the inside of an engine as they are useful for gauging a lot of different things, such as ball-joint wear and even wheel-bearing adjustment.

OTHER GOOD STUFF

Sooner or later you'll want to invest in a battery charger. A small, 10-amp job can bring a partially discharged battery to full charge in a few hours. You'll need at least a 6-amp model for decent performance. In addition, you'll need a battery post and terminal cleaning tool. You can use a knife for this job, but the special cleaning tool is inexpensive and does a better job. You'll need a timing light to perform tuneups, but be aware many late-models do not require it (no timing marks). If you're willing to suffer, you can use the cheap kind that attaches only to the plug wire. A power timing light, which provides much more illumination and is immeasurably easier to use, costs about three times as much.

And you can't get by without a miniature analyzer, which includes a tachometer, ohmmeter, voltmeter, ammeter, and dwell meter on older cars. Use it for tuneups, charging-system troubleshooting, general electrical-system troubleshooting, and a wide variety of other jobs. On late-models, a scan tool is essential to read trouble codes and engine operating data, including much of the information provided by the tachometer. A volt-ohmmeter, or volt-ohm-ammeter, continues to be useful.

A good set of taps and dies is expensive but necessary when you have to straighten out the threads on a mangled bolt or clean the threads in a rusty cylinder block. Finally, don't forget to be nice to your tools. You don't have to get into public displays of affection, but you should wipe off the grease each time you use one. And you should have an organizational plan for keeping everything in its place, whether you have a twelve-drawer top-and-bottom tool chest or a two-drawer mini toolbox. Knowing where something is when you need it can cut your creeper time in half. It also makes it easier to tell when something is missing.

ABS: see **ANTILOCK BRAKES.**

ACCELERATOR PUMP: a small pump in the carburetor that is activated to spray extra fuel through specific nozzle(s) when the accelerator pedal is suddenly and quickly pressed to the floor on sharp acceleration.

ACTIVE SUSPENSION: a suspension system in which a double-acting hydraulic actuator is present at each wheel. The system is powered by a variable-rate hydraulic pump and controlled by an ECU, which keeps track of body attitude, wheel/hub velocity, forward speed, lateral-versus-longitudinal velocity when cornering, load, and angular displacement. See **SUSPENSION.**

ADJUSTABLE SHOCK: a shock absorber whose road-handling traits can be altered to create a stiffer or softer ride and to adjust for rough roads, heavy loads, etc. While many are adjusted manually, some more recent cars include a switch inside the car that adjusts the shocks electromechanically.

AIR ASPIRATOR VALVE: a device that uses a diaphragm to draw air into the exhaust system to reduce emissions. If a large volume of air is needed, an air pump is used instead.

AIR CLEANER: a housing that holds a filter element that removes dirt from the air flowing into the throttle body, before it is mixed with fuel to form the air-fuel mixture that is burned in the cylinders.

AIR CLEANER INTAKE: the neck, called a snorkel, into which air flows en route to the throttle body.

AIRFLOW SENSOR: see **MASS AIRFLOW SENSOR.**

AIR-FUEL MIXTURE: the mixture of air and fuel necessary to produce power when ignited by a spark plug and burned in a cylinder. The normal mixture is 14.7 parts of air for each part of gasoline.

AIR PUMP: a belt-driven or electrical pump that delivers compressed air to the exhaust system to reduce emissions. Also called **SECONDARY AIR INJECTION** and **THERMACTOR.**

ALL-WHEEL DRIVE (AWD): a drive system in which the driving axle(s) connects the transmission to both the front and the rear wheels. Also called four-wheel-drive, particularly when referring to an SUV or other off-road vehicle.

ALTERNATOR: see **GENERATOR.**

AMMETER: a test meter that reads amps in a circuit.

AMPS (OR AMPERES): the measurement of the energy content of the electricity in a circuit. Also see **VOLTS** and **OHMS.**

ANTIFREEZE AND SUMMER COOLANT: a fluid of ethylene glycol and special additives that when mixed with water lowers its freezing point in winter and raises its boiling point in summer.

ANTILOCK BRAKES: a braking system with electronic and electromechanical components that prevents the wheels from locking up and causing the car to skid when traction at the wheels is uneven.

ANTISEIZE COMPOUND: a paste that is applied to metal parts to keep them from sticking to other metal parts, particularly threaded surfaces such as spark plug threads and wheel studs and bolts.

AUTOMATIC CHOKE: a carburetor choke that operates automatically, according to engine temperature.

AXLE, AXLE HOUSING: a power transfer component with a ball-shaped center section and often two tubes at a 180° angle to each other. In the center section is a gearbox called a differential, with joints for a shaft that goes through each tube. In an automobile, each shaft goes (through a tube, if used) to a wheel to power the car. Or more simply, there is just a suspension assembly that supports the front or rear wheels.

BACKFIRE: an abnormal form of combustion of the air-fuel mixture. The flame may shoot back from the cylinder into the intake manifold and possibly the throttle body. Or it may be delayed until the fuel reaches the exhaust system, in which case it causes a popping sound.

BALL JOINT: a pivot joint, typically used in an automobile's suspension system.

BARREL: the air passage in the carburetor where the throttle is located.

BATTERY: a component that stores electrical energy in the form of chemical energy.

BATTERY TERMINALS: a battery has two terminals, one labeled positive, the other negative, to which thick wires called cables are attached. These cables are the wires for the battery's paths into the electrical system of the car.

BEAM AXLE: an axle that acts as a support for associated wheels, but does not drive them.

BEARING: a smoothly finished round metal sleeve, or a sleeve with smoothly finished metal balls or rollers, to reduce the friction between two parts. In an engine and other components a simple sleeve may be split in two half-circle sections for ease of installation. Also see **BUSHING.**

BEARING CAP: a bolt-down cap used to hold a crankshaft and bearing in position.

BELLHOUSING: the area behind the engine, covered by the front of the transmission, that holds the manual clutch assembly or the torque converter of an automatic transmission.

BLOCK: see **ENGINE BLOCK.**

BEVEL GEARS: a group of gears shaped like slices of a cone, which makes it possible for the axes of the gears to be nonparallel. Bevel gears are used to transmit motion through an angle.

BLEEDING: the process of eliminating air, which can form in bubbles or pockets, from a brake or clutch line by draining the fluid containing the air.

BOTTOM DEAD CENTER (BDC): The position of a piston when it reaches its lowest point in the cylinder.

BORE: the diameter of a cylinder, such as the engine's cylinders or the brake cylinders.

BRAKE BLEEDER: a valve that can be opened as part of a procedure ("bleeding") to expel air from the hydraulic braking system, so the action of the brake pedal is free of sponginess.

BRAKE CALIPER: a C-clamp-like hydraulic device with a piston or pistons to press the brake shoes against a disc to stop the wheel when the driver steps on the brake pedal.

BRAKE DRAG: failure of the brakes to release completely when the driver's foot is removed from the pedal.

BRAKE DYNAMOMETER: a device that measures the power and torque of an engine by restricting its speed (braking) so as to gauge its output.

BRAKE FADE: A deterioration in the brake system, as a result of repeated and protracted use of the brakes, which produces reduced braking efficiency or fade; it is exacerbated by heat.

BRAKE FLUID: a special fluid used in the hydraulic braking system. For automobiles, only a fluid labeled DOT 3 or DOT 4 should be used.

BRAKE LINING WEAR SENSOR: a device on a brake shoe that signals the driver when the lining of friction material on the brake shoe or pad is worn.

BRAKE MASTER CYLINDER: the source of the hydraulic pressure transmitted to the brake system at each wheel. The pressure is developed in this component by the driver stepping on the brake pedal.

BRAKE PAD: The stiff backing plate and friction material which makes contact with the rotating disc on a disc brake when the brake pedal is depressed; it is routinely replaced when excessive wear limits its effectiveness. It also may be called a brake shoe.

BRAKE SHOES: steel plates, either curved or straight, depending on the system, to which friction material called brake lining is bonded or riveted. Also see **DRUM BRAKE, DISC BRAKE,** and **BRAKE PAD.**

BREAKER POINTS: see **POINTS** and **CONDENSOR.**

BULKHEAD: A dividing panel that separates the passenger compartment from the engine (front bulkhead or cowl) and the trunk (rear bulkhead).

BUSHING: a cylindrical sleeve of metal, plastic, or rubber placed between two parts to isolate them from each other. The bushing may absorb shock or help locate the parts.

CAM: a lobe on a camshaft.

CAMBER: a mounting angle of the wheel relative to the road.

CAMSHAFT: a rotating shaft with eccentric lobes that push open the valves for each cylinder. The lobes operate against cylindrical parts called lifters, or other actuating devices called rockers.

CAMSHAFT SENSOR: a sensor that signals to the engine control module the rotational position of the camshaft. This enables the computer to more precisely time the fuel injection and ignition systems for faster starting of the engine.

CAMSHAFT TIMING BELT (OR CHAIN): the rubber belt or metal chain that transfers power from the crankshaft to the camshaft to operate it. The chain or belt must be installed so it maintains the relationship between the camshaft and crankshaft so the valves for each cylinder open and close at the right time for proper engine operation, a factor called camshaft timing.

CARBON CANISTER: a canister filled with activated charcoal. When the engine is off, it absorbs vapors from the fuel system that would otherwise cause air pollution. These vapors are drawn into the engine and burned when the engine is running.

CARBON DEPOSITS: the residue of carbon from burning fuel, which can clog grooves in pistons, combustion chambers, and valves, and cause engine hesitation and other operational maladies.

CARBON TRACKS: fine lines from burned carbon (such as from oil film) that may be found in a distributor cap. Carbon tracks may cause engine misfire.

CARBURETOR: a device that atomizes fuel in air, and meters the resulting air-fuel mixture into the intake manifold. It also includes a throttle body. The carburetor is not used today on other than some racing cars.

CARBURETOR-AIR HORN: the top part of the carburetor, in which the choke is located.

CARBURETOR BASE: the lower part of the carburetor, in which the throttle plate is located.

CARBURETOR CHOKE: see **CHOKE.**

CARBURETOR CLEANER: a petroleum solvent for cleaning the carburetor.

CARBURETOR FLOAT: a device much like the float in a household toilet tank that regulates the amount of fuel in the carburetor fuel bowl.

CARBURETOR FUEL BOWL: a small fuel storage area in the carburetor, at the carburetor fuel inlet. Also called the float bowl because it contains the carburetor float.

CARBURETOR FUEL BOWL VENT: a vent on the bowl. It typically is connected to a carbon canister, which absorbs vapors when the engine is off, and it also may be vented to the atmosphere when the engine is running.

CARBURETOR FUEL INLET: the threaded fitting on the side of the carburetor, to which tubing from the fuel pump is connected. Fuel enters the carburetor at this point.

CARBURETOR NEEDLE AND SEAT VALVE: a valve at the carburetor fuel inlet that is controlled by the up-and-down movement of the carburetor float. When open, it allows fuel to enter the carburetor fuel bowl. When closed, it stops the flow of fuel.

CASTER: an angular relationship between a wheel and its suspension.

CATALYTIC CONVERTER: a component in the exhaust system in which harmful emissions are converted to water and carbon dioxide.

CAVITATION: A condition found in fuel and water pumps as well as fluid couplings in which a partial vacuum forms around the wheels or blades of a pump, lowering the pump's output due to the loss of contact between the blade and the fluid it is pumping.

CHASSIS: the frame or floor pan, suspension, brakes, and steering of a car.

CHECK ENGINE LIGHT (ALSO CALLED MALFUNCTION INDICATOR LAMP): a dashboard light that goes on when the engine control module indicates a failure. It may also be capable of indicating diagnostic trouble codes.

CHLOROFLUOROCARBONS (CFC): The best known of this class of chemicals was R-12 (a popular brand was Freon), which had been used as a refrigerant in a car's air conditioning system but which was phased out due to its negative impact on the earth's ozone layer. It was replaced by R-134a.

CHOKE: a thermostatically controlled plate/flap that pivots to restrict the flow of air through the carburetor when the engine is cold, to improve driveability by temporarily providing a rich fuel mixture. In a radio, a noise suppression device.

CIRCUIT: a complete path for the flow of electricity, such as from a battery, through wiring to a lamp motor, etc., and then back to the battery.

CLEAR FLOOD MODE: a situation in which a carburetor or fuel injection system increases the amount of air or reduces the amount of fuel when necessary to correct a problem called flooding.

CLUTCH: a device used to join moving parts together and disengage them when desired. In a car, a manual clutch, operated by a driver pedal, disengages the engine from the transmission. It consists of a disc coated with friction material, and held in place by a spring-loaded metal plate.

COIL: a winding of wire. See **IGNITION COIL**, for example.

COLD CRANKING AMPS: a measure of the electrical energy a battery can deliver to start ("crank") the engine when it is cold.

COMBUSTION CHAMBER: the area in the cylinder above the piston, where ignition of the air-fuel mixture occurs.

COMPOSITE HEADLIGHT: A non-sealed-beam headlight used in the United States since 1984, which features a separate lens and bulb, each of which can be replaced individually.

COMPRESSION (COMPRESSION STROKE): the squeezing of the air-fuel mixture by the upward movement of the piston into a small space, with the valves closed. It is one of the strokes of the engine. The greater the compression pressure (within limits), the greater is the power produced by the engine when the compressed mixture is burned on its power stroke.

COMPRESSION RATIO: the volume of the combustion chamber with the piston on its compression stroke versus the volume of the cylinder with the piston at the bottom of its stroke.

CONNECTING ROD: a rod that connects the piston to the crankshaft.

COMPRESSOR (A/C): a mechanism in an air conditioning system which draws vaporized refrigerant out of the evaporator, then pressurizes it before sending it to the condenser.

CONDENSER (A/C): a mechanism which converts the vapor in an air conditioning system into the liquid that later vaporizes to cool the air.

CONSTANT VELOCITY (CV) JOINT: a universal joint usually employed to connect the half-shafts to the hubs of front-wheel-drive cars; in a CV joint, the driving and driven portions of the joint rotate at the same (hence "constant") speed.

CONTACT: in an electric switch, the terminals that are bridged or brought together to close the switch.

COOLANT: a mixture of water and antifreeze that absorbs excess heat from the engine and dissipates it into the atmosphere.

COOLING SYSTEM: a system for circulating coolant through the engine, a radiator, and the heater using a water pump to keep the engine from running too hot, and to provide heat for the passenger compartment.

CRANK, CRANKING: the act of starting the engine, in which the starting motor turns a gear on the crankshaft to turn the flywheel fast enough for fuel and ignition systems to quickly go to work and enable the engine to continue running on its own.

CRANKCASE: The portion of the engine which contains and supports the crankshaft; it usually includes the oil pan and the lower part of the cylinder block.

CRANKSHAFT: a rotating shaft that moves the pistons up and down in the cylinders. When it is rotated by the downward force of burning air-fuel mixture oil the pistons, it provides the power that is eventually transferred through the transmission to the wheels to propel the car.

CRANKSHAFT BALANCER (ALSO CALLED HARMONIC BALANCER): a circular device in the front end of the crankshaft, designed to dampen some of the impulses from the combustion events in the cylinders.

CRANKSHAFT JOURNALS: smoothly finished circular surfaces on the crankshaft, some of which are used with bearings to mate the crankshaft to the engine block. These are called main bearing journals. Others, also with bearings, mate to the connecting rods and are called rod journals.

CRANKSHAFT SENSOR: A sensor that determines the rotational position of the crankshaft, and transmits the data to the engine control module.

CYLINDER HEAD: the top part of the engine that contains the cylinder ports, the intake and exhaust valves, on most engines the combustion chamber and on many engines, the camshaft (or two camshafts).

CYLINDER HEAD GASKET: a gasket between the cylinder head and the engine block.

CYLINDER PORTS: passages in the cylinder head, two for each cylinder—one to bring the air-fuel mixture into the cylinder, the other to carry out burned exhaust gases.

DIAGNOSTIC TROUBLE CODE (ALSO TROUBLE CODE): a number stored by the engine control module when it detects a failure in a particular electrical circuit or mechanical system it is capable of monitoring. This number is a useful guide to diagnosis.

DIELECTRIC GREASE (ALSO CALLED SILICONE DIELECTRIC GREASE): a lubricant that repels moisture and has the property of adhering very well to surfaces.

DIESELING (ALSO CALLED RUN-ON): a condition in which a gasoline engine continues to fire after the ignition has been turned off; it is usually the result of excess heat and unusually high manifold pressure.

DIFFERENTIAL: a set of gears that accepts power from the transmission and transfers it to the wheels to drive the car.

DISC BRAKE: a hydraulic braking system in which flat steel plates coated with a lining of friction material (called brake shoes or pads) are pressed by a brake caliper against the sides of a metal disc to which the wheel is attached. The friction material rubs against the sides of the disc to stop the wheel when the brakes are applied.

DISTRIBUTOR: in an ignition system, an electromechanical device that contains a sensor to signal an electronic ignition unit to activate the ignition system, and a rotor to distribute high-voltage electricity to the appropriate spark plug wire, through a cap into which the spark plug wires are fitted. The distributor has been replaced on most engines by electronic devices.

DISTRIBUTOR ADVANCE (WEIGHTS): a pair of spring-loaded centrifugal weights that cause a change in ignition timing at higher engine speeds by pivoting a plate within the distributor; on older cars only.

DISTRIBUTOR CAP: a plastic cover over the distributor. It holds the wires from the spark plugs and the ignition coil.

DISTRIBUTOR DRIVE GEAR: a gear on the distributor that meshes with a gear on the camshaft to cause a shaft in the distributor to turn. The gear is held to the shaft by a pin.

DISTRIBUTORLESS IGNITION: a form of electronic ignition in which the distributor is replaced by an electronic module and a special sensor in the engine.

DRIVEBELT: a rubber belt that transfers power from a pulley on the engine to a pulley on an accessory to operate the accessory. Also see **V-BELT.**

DRIVESHAFT: a shaft from a transmission, either to a wheel or to an axle housing, to transfer power. When used to transfer power to a rear axle housing, it also is called a propeller shaft.

DRIVETRAIN: a collection of components, including the transmission, driveshaft, differential—and any associated shafts, gears, clutches, and joints—which transmits power to the wheels.

DRUM BRAKE: a hydraulic braking system in which half-moon-shaped brake shoes are pushed by pistons in a wheel cylinder against the inside cylindrical surface of a part called the drum, to which the wheel is attached. The brake shoes are coated with a lining of friction material that rubs against the drum surface to stop the wheel when the brakes are applied.

DUAL BRAKING SYSTEM: a braking system which features independent hydraulic circuits for each pair of wheels (divided into front/back or diagonal front/back pairs).

DYNAMIC BALANCING: the process whereby an automotive part, usually a wheel or crankshaft, is balanced so as to correct the tendency of an out-of-balance part to rock or vibrate. The dynamic balancing solution to such rotational problems—as opposed to the method of "static balancing"—is to introduce a counterweight to offset the irregularities.

DYNAMOMETER: A device used to measure horsepower and torque. When applied to an engine it measures the power transmitted to the flywheel; when applied to a chassis, it measures the power transmitted to the wheels.

ECM: see **ENGINE CONTROL MODULE.**

ECU: electronic control unit. It is the "brain" for an electronically-controlled system. One for an engine is called an **ECM**.

EGR: see **EXHAUST GAS RECIRCULATION.**

ELECTROLYTE: A solution used in automobile batteries to conduct electricity; voltage and current are generated when the solution comes into contact with the material on the battery plates. The electrolyte used in batteries is composed of distilled water and sulfuric acid.

ELECTROMAGNETIC PICKUP: see **PICKUP.**

ELECTRONIC IGNITION: the modern type of ignition system, in which electronic components replace some of the electromechanical switches previously used. On newer cars it refers to **DISTRIBUTORLESS IGNITION**.

ELECTRONIC IGNITION MODULE: see **IGNITION CONTROL MODULE.**

EMERGENCY BRAKE (ALSO CALLED PARKING BRAKE): A separate and independent braking system to be used as a backup system in the event of a failure in the primary braking system or to assist in holding the car stationary when parked. A pedal or hand lever is the normal mode of engaging the emergency brake, which is connected mechanically to either the front or rear brakes, although an electric motor and ECU applies the parking brake on some new cars.

EMISSIONS (EXHAUST): those compounds in the exhaust gases that are harmful to air quality. Typically measured are carbon monoxide and and unburned hydrocarbons (gasoline particles), and even oxides of nitrogen, which is a component of groundlevel "smog."

ENGINE BLOCK: the lower part, of the engine, in which the cylinders, pistons, and crankshaft are located.

ENGINE CONTROL MODULE (ECM): the electronic computer that takes readings from various electronic sensors on the engine and possibly the transmission, and performs such functions as controlling idle speed, turning on the air conditioning, regulating fuel mixture and ignition timing and triggering various emission control devices, such as exhaust gas recirculation and the fuel vapor control system.

ENGINE IDLE SPEED: the rotational speed at which the engine runs when the driver's foot is off the gas pedal. The speed is regulated by the engine control module on late-model cars.

ENGINE MOUNTS: The supports that connect the transmission and engine to the car's chassis. Composed of rubber and steel, the engine mounts absorb the motion (twisting, vibrating, etc.) produced by the operation of the engine and transmission; they also assist in reducing the noise and motion transmitted to the passenger compartment. The most efficient recently developed mounts are hydraulic, in some cases electronically-controlled.

ENGINE OIL GALLERY: a series of passages, usually drilled, through which oil circulates to key sections of the engine and to the crankshaft.

ENGINE OVERHEATING: occurs when the coolant in the cooling system is so hot the metals in the engine are at a temperature that may cause damage to them. In addition, the engine runs poorly and usually stalls.

EVAPORATIVE EMISSIONS CANISTER: See **CARBON CANISTER**.

EXHAUST GAS RECIRCULATION (EGR): a system featuring the EGR valve, that controls a portion of the exhaust gas flow. It meters a small amount of burned gases from the exhaust system back into the cylinders to combine with the air-fuel mixture, diluting it. This reduces the fuel mixture combustion temperatures, to reduce formation of oxides of nitrogen.

EXHAUST MANIFOLD: a component with chambers that carry burned exhaust gases from the cylinders into pipes that allow them to flow to the catalytic converter and through that to the muffler and tailpipe.

EXHAUST STROKE: when the piston is moving up in the cylinder, the exhaust valve is open and burned gases are being pushed out into the exhaust system.

EXHAUST SYSTEM: the series of parts that carries burned gases from the cylinder head of the engine out into the atmosphere. It includes one or two exhaust manifolds, on recent-model cars typically one or two catalytic converters, interconnecting pipes, a muffler, and a final pipe called the tailpipe.

EXHAUST VALVE: a valve that when open allows the flow of burned gases from the air-fuel mixture to travel into the exhaust system

FAN BELT: a rubber band used to transmit power from a crankcase-driven pulley to the pulleys driving the fan, alternator, power steering pump, air conditioning pump, etc. It is usually composed of rubber reinforced by steel or cord to minimize the risk of stretching. It may have a V-shaped cross-section or a multi-V shape across the underside (a design called a ribbed belt).

FEELER GAUGE: a blade or piece of wire of predetermined and carefully calibrated thickness, used to measure the gap between two parts. It is most often used as the tool to determine the clearance between a spark plug's inner and outer electrodes.

FIRING ORDER: the sequence in which spark is delivered to the cylinders, which are numbered according to position in the engine. The sequence is determined by what is necessary to provide a smooth flow of power to the crankshaft. A firing order for a six-cylinder engine might be 1-5-3-6-2-4.

FLOODING: a condition in which there is excess fuel flowing from the carburetor or fuel injectors. It may cause the engine to stall or run poorly.

FLUSHING TEE: a device with three hose necks laid out in the shape of the letter T that is spliced into a heater hose, secured with hose clamps. The neck on the stem is covered with a threaded cap, which is removed and a household water hose is attached (with a special adapter) to run water through the cooling system to flush out dirt, rust, etc.

FLYWHEEL: a heavy wheel at the rear of the crankshaft that smooths out its transfer of power to the transmission. It has a gear around its circumference, which meshes with a gear on the starter so the turning of the starter cranks the engine. On automatic transmission cars, it is a thin steel plate, to which the torque converter is attached.

FOUR-STROKE ENGINE: the typical auto engine, with intake, compression, power, and exhaust strokes to form one complete cycle in two complete revolutions of the engine.

FUEL-AIR MIXTURE: see **AIR-FUEL MIXTURE.**

FUEL FILTER: a filter to remove dirt and some water from the fuel. If the filter is installed in the line between the fuel pump and the carburetor or fuel injection system, it is called an in-line filter.

FUEL INJECTION: a system that uses fuel injectors to deliver fuel to the cylinders. Also see **MULTIPORT FUEL INJECTION** and **THROTTLE-BODY FUEL INJECTION**.

FUEL INJECTOR: a device that sprays fuel into a cylinder port or a throttle body. Although most injectors are electromechanical devices, some are purely mechanical.

FUEL PRESSURE: the pressure, measured in pounds per square inch or kilopascals, of the fuel in the line from the fuel pump to the carburetor or fuel injection system. It is measured with a special gauge.

FUEL PUMP: a pump that draws fuel from the tank and delivers it under pressure to the carburetor or fuel injection system. It may be mechanical, operated by the engine, or electrical.

FUEL RAIL: in a multiport fuel injection system, a line of tubing at the engine to which the fuel injectors are attached and from which they receive the fuel they inject into the cylinder ports.

GASKET: a flat material designed to compensate for irregularities and therefore prevent leaks between mating surfaces.

GEARBOX: a housing with gears, such as a transmission.

GENERATOR: a device powered by a drivebelt that converts the mechanical energy of the engine into electrical energy for electrical devices and to recharge the car battery.

GOVERNOR: a mechanical, electrical, or hydraulic device used to restrict the maximum speed of a vehicle in order to reduce wear and tear on the engine and drivetrain and make the operation of the vehicle safer.

GREASE FITTING: a small nipple containing a spring-loaded valve, threaded into a part that needs periodic lubrication with grease. The nipple is shaped to accept the tip of the grease gun.

GREASE GUN: a service tool that injects grease, usually into a grease fitting, but also into hinge joints.

GROUNDING (SUCH AS GROUNDING A TESTER): attaching a wire (such as from a tester) to a ground, a metal part of the engine or car body, or the negative terminal of the car battery.

GROUNDS: the negative terminals of modern automotive circuits. Typically, the engine and metal car body serve this purpose. They provide a complete circuit back to the battery without the need to run many wires to the battery itself.

HALF SHAFT: either of two rotating shafts that connect the transaxle to the wheels in a front-wheel drive car. Half shafts may be either solid or tubular in construction.

HALL (EFFECT) SWITCH: an electronic switch often used as a sensor, such as a type of pickup.

HARNESS: see **WIRING HARNESS.**

HEAD GASKET: see **CYLINDER HEAD GASKET.**

HESITATION: a condition in which the car's engine fails to respond promptly during acceleration.

HYDRAULIC BRAKING SYSTEM: this is the braking system on every modern car. When the driver steps on the pedal, he/she pushes on a piston inside a brake master cylinder filled with fluid. This motion is transferred through the fluid in lines to each wheel, where a disc or drum brake stops the wheel. Use of fluid under pressure to perform work is called hydraulics. On some new cars, an electronic system controls the hydraulic system.

HYDROMETER: a tester for measuring the freeze protection of the engine coolant as well as the specific gravity of battery electrolyte.

IDLE SPEED: see **ENGINE IDLE SPEED.**

IGNITION COIL: a voltage transformer for the ignition system. It converts the 12-volt current from the battery into the high voltage necessary to fire across the spark plugs. The voltage needed typically ranges from under 10,000 to as much as 50,000 volts.

IGNITION CONTROL MODULE: an electronic module that controls the ignition system, perhaps in conjunction with the engine control module.

IGNITION PICKUP: see **PICKUP.**

IGNITION PRIMARY: the low-voltage part of the ignition circuit, such as part of the ignition coil wiring, the pickup, electronic ignition module, and engine control module.

IGNITION SECONDARY: the high-voltage part of the ignition circuit, such as part of the ignition coil wiring, spark plugs, spark plug wires, distributor cap, and rotor.

IGNITION SYSTEM: a group of components in a circuit that transforms the 12-volt battery electricity to approximately 10,000-50,000 volts, and delivers that electricity to each cylinder at the correct instant to ignite the air-fuel mixture.

IGNITION TIMING: See **TIMING.**

INLET SOCK: a coarse fuel filter in an older fuel tank, designed to remove only very large dirt particles; it can last the life of the car.

INTAKE MANIFOLD: a component with passages that lead from the throttle body to the combustion chamber intake port for each cylinder. It carries the air-fuel mixture on engines with carburetors or throttle-body fuel injection, or just the air on engines with a fuel injector at each cylinder (multiport fuel injection).

INTAKE STROKE: when the piston is going down in the cylinder, and the intake valve is open, so the downward movement creates a vacuum that draws in air-fuel mixture.

INTAKE VALVE: a valve that when open allows the flow of air-fuel mixture through the cylinder port into the cylinder.

JUMPED TIME: a situation in which ignition or camshaft timing is incorrect because of a mechanical malfunction.

JUMPER: a wire with a clip at each end. It is used to make a temporary electrical connection.

KNOCK: a knocking sound. It may be produced by worn engine parts, by an ignition system with excessive spark advance, or by low-octane gasoline.

KNOCK SENSOR: a device that senses knock and reports this to the engine control module, which adjusts ignition timing to eliminate it.

LAMINATED WINDSHIELD: a windshield, standard on all American cars, composed of two sheets of glass with a thin layer of plastic between them to keep the windshield from splintering.

LATERAL RUNOUT: a condition in which a wheel wobbles from side to side as it rotates.

LEAF SPRING: a type of spring composed of a flat strip or several long, slightly curved flexible steel or fiberglass leaves (plates). Leaf springs are commonly used in conjunction with the rear axle of trucks.

LEAN FUEL MIXTURE: an air-fuel mixture that contains an excessive amount of air, and thus affects combustion of the mixture in the cylinders.

MALFUNCTION INDICATOR LAMP: See **CHECK ENGINE LIGHT.**

MACPHERSON STRUT: an oversized shock absorber first developed by Earle MacPherson at Ford. It replaces the upper control and ball joint when used on the front suspension; some have removable internal components that can be easily replaced.

MANIFOLD HEAT CONTROL VALVE: a thermostatic valve that controls the flow of exhaust gases so they heat the air-fuel mixture for better driveability when the engine is cold. It was used on engines with carburetors.

MAP SENSOR: a variable resistor used to monitor engine load and accordingly adjust spark timing and fuel mixture.

MASS AIRFLOW SENSOR: a sensor that measures the amount of air flowing to the throttle body in a fuel injection system and reports this information to the engine control module.

MICROMETER: a precision measuring instrument.

MICROPROCESSOR: a small computer.

MID-ENGINE: an engine whose location is wholly or in part within the wheelbase. Due to the more even distribution of weight that it produces, the mid-engine is frequently used in racing and high-performance vehicles.

MISFIRE: erratic operation of an engine caused by failure of one or more cylinders to contribute power. The problem could be in a mechanical system, in the air-fuel delivery, or in the ignition system.

MISS: see **MISFIRE.**

MUFFLER: a part of the exhaust system that contains baffles or special materials to muffle the sound of exhaust gases moving out of the engine.

MULTIPORT FUEL INJECTION (ALSO CALLED MULTIPOINT FUEL INJECTION): a gasoline delivery system in which there is one injector for each cylinder, and that injector is located at the cylinder port.

OHM: the measurement of resistance in an electrical circuit. See **RESISTANCE**; also **AMPS** and **VOLTS.**

OHMMETER: a meter that reads resistance, measured in ohms, in an electrical circuit or wire.

OIL PAN: a removable receptacle that is located at the bottom of the cylinder block and contains the engine's oil.

OIL PUMP: the pump that circulates oil to all the moving parts of an engine. Usually driven from the crankcase by gears or cams, the oil pump is also fitted with an inlet screen to filter the oil before it enters the pump.

OPEN CIRCUIT: a wiring circuit that is interrupted, such as by an open switch, a bad wiring connection, or an internal break in the wiring.

ORIGINAL EQUIPMENT MANUFACTURER (OEM): a generic term that refers to an automobile company or supplier which manufactures the parts used in the original assembly of a car.

O-RING: a type of sealing ring with a cross-section shaped like the letter O, often made of rubber.

OUT OF ROUND: a condition in which a brake drum has lost its original shape due to inconsistent wearing, warping, etc.; a drum that is out of round will produce pulsing, grabby brakes.

OVERHEAD-CAMSHAFT: a type of engine in which the camshaft is mounted in the cylinder head. This design eliminates the use of pushrods.

OVERHEAD VALVE: a type of engine with a camshaft mounted in the engine block, operating valves in the cylinder head through a "train" of valve lifters resting on camshaft lobes, pushrods transmitting the lifting force from the camshaft to rocker arms that pivot to open the overhead valves. Also called "valve in head."

OXIDES OF NITROGEN: an exhaust emission formed by nitrogen in the air at very high fuel mixture combustion temperatures in the combustion chamber.

OXYGEN SENSOR: an electronic device threaded into the exhaust system to measure oxygen content in the exhaust gases. High content usually indicates a lean fuel mixture was burned in the combustion chamber; low content usually indicates a rich fuel mixture.

PASSIVE RESTRAINT: any safety device which restrains the movement of passengers inside a car but which requires no action on the part of the passenger to do so; the most common examples are seat belts and airbags.

PCV: see **POSITIVE CRANKCASE VENTILATION.**

PICKUP: the sensor in the ignition system's distributor. Also a term that may be applied to any sensor or sending unit.

PINTLE: the tip of some types of valves.

PISTON: a cylindrical cap-shaped part that moves from one end to the other in a cylinder, such as in an engine to compress the air-fuel mixture. When the piston moves down under pressure from the burning, expanding mixture, it turns the crankshaft. A piston may also be used to transfer hydraulic force, as in the hydraulic braking system

PISTON SLAP: a slapping noise in the engine caused by piston wear.

PISTON RING: a thin open-ended ring that is installed in a groove on the outer diameter of a piston in order to create a seal between the piston and the associated cylinder and prevent any oil from entering the combustion chamber. Most pistons have three rings, two for compression sealing and one for oil control.

PLASTIGAGE: a plastic material that compresses to the thickness of the clearance between a crankshaft journal and a bearing when the bearing retaining cap is installed, so the clearance can be checked against specifications.

POINTS AND CONDENSER: a simple electromechanical system in which a distributor-controlled switch (the points) and an electrical charge storage device (the condenser) were the control parts for an automobile ignition system. They were replaced by electronic components beginning in the 1960s, although some cars had points and condenser in the early 1980s.

POSITIVE CRANKCASE VENTILATION (PCV): an emissions control system for unburned gasoline droplets that slip past the piston rings and go into the engine crankcase. A PCV valve controls the flow of the unburned fuel back into the cylinders for burning, so that the flow increases with engine speed. If the valve is stuck open, it can upset the engine idle speed.

POWER BRAKE: a device that provides an assist to the driver when the brake pedal is depressed. Although most power brake units are vacuum operated, some use hydraulic pressure.

POWER STROKE: occurs when the piston is pushed down by the expansion of the burning air-fuel mixture in the cylinder.

POWERTRAIN CONTROL MODULE: an engine control module that also controls the automatic transmission.

PREIGNITION: ignition of the air-fuel mixture in the cylinders that occurs prior to the arrival of the spark. It may be caused by a hot spot in the combustion chamber.

PUSHROD: a part of the engine valve system in a "pushrod valve" arrangement. In this system (see **CAMSHAFT**) the camshaft is in the middle of the engine and the pushrod rests in the lifter. The opposite end of the pushrod pushes against a device in the cylinder head called a rocker, which pivots to push open a valve.

RADIATOR: a heat exchanger that dissipates into the atmosphere the heat absorbed by the coolant as it circulates through the engine.

RAM TUBES: tubes of a specific length and shape in the intake manifold that promote performance at certain engine speeds by "ramming" air into the cylinders.

RECHARGE: as in to recharge a battery, which means to restore the electrical energy dissipated in starting the engine and other uses.

REFRIGERANT: the substance used in an air conditioning system that absorbs, carries, and releases heat as it changes from a liquid to a gas and back to a liquid again. The most commonly used refrigerant currently in use is R134a, although older vehicles still may have R-12.

RELAY: an electrical switching device that allows a small amount of current in one circuit to control a much larger flow of current in another circuit.

RESISTANCE: a measurement, in a unit called ohms, of the resistance of electricity to flow in a circuit. It equals the voltage in that circuit divided by the amperes.

RESISTOR: a device that restricts the flow of current in a circuit.

RICH FUEL MIXTURE: an air-fuel mixture that contains more fuel than is necessary for efficient combustion in the cylinders, and so reduces gas mileage. A rich mixture, however, may be necessary for easier cold starting.

ROCKER (ALSO CALLED ROCKER ARM): a pivot installed in the cylinder head. When one end is pushed, the rocker pivots to push open a valve.

ROTOR: another name for the disc in a braking system. See **DISC BRAKE**. Also the term for a rotating part inside the ignition distributor, which serves to distribute high-voltage electricity to the spark plug wires.

RPM: revolutions per minute, the measurement of engine speed.

RUN-ON: the tendency of an engine to run for a brief period after the ignition key is turned off. Also called dieseling and after-run. It may be caused by an engine malfunction or low-octane gasoline.

SAFETY STAND (ALSO CALLED JACK STAND): a device that can be securely locked at a choice of heights, so it can be placed under specific parts of the car underbody to support the weight of the car that has been raised with a jack, and keep the car safely in place.

SCHRADER-TYPE VALVE: a valve with a spring-loaded pin, in which the pin is depressed to open the valve. A Schrader-type valve is used at each tire, on most air-conditioning systems, and on some fuel injection systems. Schrader is a manufacturer of this type of valve.

SENDING UNIT: a device that senses something. On the car it may sense oil pressure in the engine, coolant temperature, air temperature, transmission oil temperature, and fuel level.

SENSOR: a sending unit. Also an electronic device that senses something and reports its reading to an electronic control unit, such as an engine control module. See **OXYGEN SENSOR, THROTTLE POSITION SENSOR, MASS AIRFLOW SENSOR.**

SHOCK ABSORBER: a device that forces fluid through narrow openings to dissipate the energy absorbed by the car's springs when the wheels ride over bumps in the road.

SHOES: see **BRAKE SHOES.**

SHORT CIRCUIT (OR SHORTED CIRCUIT): a wiring circuit that ends short of its completed path because of a wiring fault.

SOLENOID: an electromagnetic switch with an arm or a shaft that can perform a mechanical function when electricity is applied. Most fuel injectors contain a solenoid.

SPARK: in an engine, the high-voltage electricity that jumps an air gap in a spark plug. A check of the delivery of this spark from the end of the spark plug wire is called a **CHECK FOR SPARK** and is a basic automotive test when an engine fails to start.

SPARK ADVANCE: a change in ignition timing, so it occurs earlier than it did, for better performance. Timing may be set back ("retarded") to prevent knock. See **SPARK RETARD.**

SPARK KNOCK: see **KNOCK.**

SPARK PLUG: a part with two electrodes, with a ceramic between them as an insulator, and the electrode tips separated by an air gap, threaded into each cylinder. When high-voltage electricity is applied, it jumps through the air gap between the electrode tips and ignites the air-fuel mixture.

SPARK PLUG BOOT: the nipple end of the plug wire jacket. It covers the terminal on the end of the plug itself.

SPARK PLUG CABLE: another common name for spark plug wire.

SPARK PLUG ELECTRODE: one of two electrical contacts at the tip of a conventional spark plug. The spark is the high-voltage arc that bridges the air gap between them.

SPARK PLUG HEAT RANGE: a spark plug design factor. It refers to the spark plug's ability to dissipate heat.

SPARK PLUG WIRE (ALSO CALLED IGNITION WIRE): a wire with thick rubber insulation that carries the high-voltage spark from the ignition coil to the spark plug.

SPARK RETARD: a change in ignition timing so it occurs later than it previously did, either because of a malfunction or intentionally to stop engine knock caused by the ignition system.

SPEC, SPECIFICATION: a number that is a standard of performance or service adjustment.

STARTER, STARTING MOTOR: the motor that is activated by battery current to start the engine.

STARTER RELAY: a relay used in the starting system.

STARTER SOLENOID: a solenoid used in the starting system, typically built onto the starting motor.

STEERING BOX: a gearbox which converts the circular motion of a steering wheel into the motion of the steering arms, knuckles, and wheels. It is located near the bottom of the steering shaft.

STROKE: the distance a piston travels from bottom dead center to top dead center.

SUSPENSION: a system of bars, springs, and shock absorbers to which the wheels are attached, and which supports the car body and the underbody frame on which it may sit.

TACHOMETER: an instrument for measuring rpm.

TAILPIPE: a pipe that follows the muffler and is the end of the exhaust system.

TEMPERATURE-SENSITIVE BIMETAL COIL: a part made of dissimilar metals that cause the coil to flex with changes in temperature. The flexing bimetal can operate a temperature-sensitive device, such as a carburetor choke, or the hydraulic circuit in most non-electric radiator fans.

TEST LIGHT: a device used to test electrical circuits. It contains a bulb or electronic equivalent, such as a light-emitting diode (LED), that goes on when a circuit is complete and carrying electricity.

THERMODYNAMICS: the science that deals with heat and heat transfer.

THERMOSTAT: a temperature-sensitive valve that regulates the flow of coolant between radiator and engine.

THERMOSTATIC AIR CLEANER: an air cleaner housing with a flap valve that is controlled by a thermostatic device. In one position, it ducts hot air from the area of the exhaust system into the engine for smooth operation when the engine is cold. In the other position, it ducts cool, more dense air from the front of the car into the engine when it is warm for more power. It was used on engines with carburetors.

THERMOSTATIC SWITCH: a switch that opens or closes at a certain temperature.

THROTTLE BODY: the part of the fuel system that holds a flat, generally round plate (the throttle valve) that regulates the flow of air into the cylinders in response to how much the driver steps down on the accelerator. The throttle body may be part of a fuel injection or carburetor fuel system. Regulating the air flow controls the speed of the engine and therefore the amount of power it produces.

THROTTLE-BODY FUEL INJECTION: a simplified fuel injection system with one or two fuel injectors positioned above the throttle valve.

THROTTLE POSITION SENSOR: a sensor that determines how far the gas pedal has been depressed, and delivers that information to the engine control module.

THRUST ANGLE: a locational relationship between the front and rear sets of wheels.

TIE ROD: a rod which is part of the steering system and links the steering knuckles to the steering rack or a center link.

TIME: the state of timing. See **JUMPED TIME.**

TIMING: the regulation of the ignition system such that it produces a spark at the spark plug at the correct instant for combustion; the regulation of the camshaft such that it opens and closes the valves to admit air-fuel mixture and allow exhaust gases to exit the cylinder.

TIMING GUN: see **TIMING LIGHT.**

TIMING LIGHT: a special type of light that when aimed at a mark on a rotating part lights only when a spark plug is triggered, so the mark seems to stand still. The instantaneous alignment of this mark with a fixed mark on the engine permits a check of ignition timing, and adjustment if necessary. See **TIMING MARKS.**

TIMING MARKS: a fixed mark on the engine and a second mark on a part that turns with the crankshaft (such as a belt pulley in front or the engine flywheel in the rear). When these marks align (as checked with a timing light) is an indication of the adjustment of the ignition system. Many late-model vehicles, with computer controls of timing, do not have these marks.

TOE: a measurement of whether the wheels at front or rear point toward each other (called toe-in) or away (called toe-out).

TOP DEAD CENTER (TDC): when the piston is at the very top of its upward movement.

TORQUE CONVERTER: a fluid-filled device with fanlike members that couples an engine to an automatic transmission.

TORQUE WRENCH: a wrench with a dial that reads in a measurement of twisting force, such as pound-feet or ounce-inches, or their metric equivalent, such as Newton-meter, abbreviated N-m.

TORSION BAR: a spring in the shape of a bar. It has been used in the suspensions of many cars.

TORX: a type of screwhead that requires a specific size Torx wrench to be loosened or tightened.

TRANSAXLE: a transmission/axle combination, most often found in front-wheel drive cars, in which the clutch, gearbox, and differential are all housed in a single unit.

TRANSMISSION: a component with gears and shafts that takes the power of the engine and transmits it, through external shafts, to the wheels to move the car. The gears have different ratios to move the car at different speeds.

TRIGGER WHEEL: in the ignition system, a wheel with flat, square "teeth," operated by a shaft in the distributor. It often is used to trigger a Hall-effect pickup to produce a signal to the engine control module.

TROUBLE CODE: see **DIAGNOSTIC TROUBLE CODE.**

TVRS WIRE: a type of resistance wire used for spark plug wires. It minimizes ignition system interference with the radio and other electronic components on the car.

UNIVERSAL JOINT (U-JOINT, ALSO CALLED CARDAN JOINT): a flexible coupling using a double yoke and a four-point cross. It is used most commonly to connect the driveshaft to the transmission and the differential in a front-engine rear-drive car.

VACUUM: pressure that is lower than atmospheric, produced by the engine or a pumping device, measured in inches or millimeters (mm) of mercury. The engine produces about 17-22 inches (approximately 430-560 mm) at engine idle speed. Engine vacuum normally is measured at a hose neck on the intake manifold.

VACUUM GAUGE: a gauge that measures vacuum, such as that produced by an engine.

VACUUM HOSE: a hose that connects a source of vacuum, such as on the engine's intake manifold, to a device that uses it, such as a power brake unit.

VACUUM LEAK: a loss of vacuum from a leaking hose or defective gasket.

VALVE CLEARANCE: an air gap that exists between the tip of a valve and the part that opens it when the valve is fully closed. In some engines this gap is adjustable and must be set to specifications.

VALVE LIFTER: a valve train component. In OHV (overhead valve) engines, the lifter is positioned between the cam lobe and the pushrod. In some overhead camshaft engines, the lifter is positioned between the cam lobe and the valve stem.

VALVE SEAT: a surface in the combustion chamber, against which a valve seats when closed.

VALVE SPRING: a spring that keeps an engine valve closed until pushed open by a rocker or by the valve lifter.

VALVE TRAIN: the system of parts that operates (and includes) the intake and exhaust valves.

V-BELT: a drivebelt with the cross-sectional shape of the letter V that rides in a similarly shaped pulley. Also see **DRIVEBELT.**

VOLTS (OR VOLTAGE): a measurement of electrical pressure. The automobile today uses an electrical system that operates on approximately 12 volts. Most engine control modules and their associated parts contain both 5-volt and 12-volt circuits. A car's ignition system typically develops about 10,000-50,000 volts. Also see **AMPS** and **OHMS.**

VOLTMETER: a test meter that measures volts.

V-TYPE ENGINE: an engine with two banks of cylinders set at an angle, such as a V4, V6, V8, V10, V12, and V16. The number following the V is the total number of cylinders, such as two banks of two cylinders in a V4.

WATER JACKET: passages in an engine through which coolant is circulated.

WATER PUMP: a small pump, driven by a pulley, which circulates coolant through the cooling system; it can be driven by the fan belt, alternator belt, or overhead cam timing belt or chain.

WHEEL ALIGNMENT: adjustment of the suspension and steering of a car so it rides straight down the road and responds predictably in turns. The measurements adjusted or checked include **TOE, CASTER, CAMBER,** and **THRUST ANGLE.**

WHEEL BEARING: a ball or roller bearing assembly that supports whatever part to which the wheel is mounted, so the wheel can spin freely as the car rolls down the road. See **BEARING.**

WHEEL CHOCK: a triangular piece of metal or wood that can be wedged between a tire and the ground to keep the car from rolling.

WHEEL SPINDLE: a flange that holds the wheel bearing assembly. It may be the end of a driveshaft.

WIRING DIAGRAM: a diagram that shows how the wires in the car, or an individual circuit, are connected to components and to one another.

WIRING HARNESS: a group of wires bundled together and covered by a protective jacket.

WIRING HARNESS CONNECTOR: a single connector with many terminals for electrical wiring connections.

INDEX

A

accumulator **142**, 143
actuator 173, **174**, 175, 230
Advanced Metal Diaphragm seal 97
air conditioning system 44, 45, 46, 47, 192. 193, 194, 195
  afterblow module **194**, 195
  air intake 193
  and R-12 refrigerant 44
  clutch relay 45
  compressor **44, 45**
  condenser 46
  coolant temperature sensor 44, 45, 46
  curing odors 192, 193, 194, 195
  disinfectant spray **193**, 194
  evaporator temperature sensor 47
  evaporator tray drain **192**
  servicing of 44, 45, 46, 47
  temperature control systems 192, 193
air filters 7, 8, 9, **90, 260, 261, 262**
  cabin 260, 261, 262
  ducts **8, 9**
  housings 7, 8
  how filters work 9
  installation 8, 9
  replacing filter 7, 8, 9, 90, 93
airflow meter 8, 94
air hose 25
air-powered die grinder **13**, 14
alternator 37, **67**, 95, 183
American Petroleum Institute (API) 38
ammeter 51, **107, 180**
antenna **16, 17, 18, 167**
  and radio noise 167
  power antenna repair 16, 17, 18
  motor **18**
antifreeze 114, 115
automatic shift 140, 141, 142, 143
  all shifts 141
  curing harsh shift 140, 141, 142, 143
  garage shifts 141
  throttle-valve cable **140**
  vacuum modulator 141
automatic transmission 142, 143, 144, 145, 146
  adding oil 146
  draining oil 144
  filter removal 145
  filter replacement 115, 145, 146
  gasket **144**, 145, 146

B

battery **27, 28**, 29, 37, **48, 49**, 50, 51, 52, 53, **92**, 99, 100, **101**, 105, **106**, 182, **183, 184, 185**
  adapter bolts **28**
  charging adapter **182**
  cleaning **37**, 92
  cutoff switch **51**
  eliminating battery drain 182, 183, 184, 185
  ground points 184, 185
  hold-down clamps **50, 184**
  how battery test windows work 52
  how maintenance-free batteries work 49
  how to deal with deeply discharged batteries 28
  hydrometer **49, 50, 51**, 99
  jumper cables **29, 183**, 184
  jumpstarting 27, 28, 29
  maintenance 48, 49, 50
  parasitic-draw adapter **183**
  replacement 51, 52, 53
  side-terminal posts **53**
  trickle chargers **182**
belts 38, 39, **65, 66, 67, 68**, 196, 197
  automatic tensioner 66, 67
  checking tension on 66
  pulleys **66, 67, 68**
  replacement 65, 66, 67, 68
  serpentine belt **66**, 68
  v-belt 65
belt tension gauge 38
bezel attachment screws **186**
body panels 275, 276, 277
  A-pillar 277
  adjustment bolts **276, 277**
  alignment 275, 276, 277
  door 275, 276, **277**
  gaskets 276
  hinges 277
  hood 275, 276
  latches **276, 277**
  snubber adjustment **275**, 276
  trunk 275, 276
boot pliers **21**
brakes 34, 35, 36, 198, 199, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 245, 246, 247
  actuator 247
  antilock 34, 230
  bleeding 223, 224
  caliper **218, 225, 226, 227, 229, 230**
  corrosion 246
  curing premature brake wear 216, 217, 218
  curing pulsing brakes 34, 35, 36
  dealing with a low brake pedal 222, 223, 224
  dial indicator **247**
  discs **34, 35, 216, 218, 245, 246, 247**
  drum brakes 216, 217, **231**, 232, 233
  effect of tire imbalance on 36
  fluid 216, 228, 229, 230
  flushing 228, 229, 230
  four-wheel disc brakes 221
  how brake rotors work 246
  how dual master cylinders work 223
  how pad wear indicators work 217
  inspection 35
  lateral runout 245
  linings 216
  measuring disc runout 35, 36
  "milking" a disc 35
  pads 217
  rear disc **247**
  rear drum 36, 217
  replacing a parking brake cable 219, 220, 221
  restricted brake hose 227
  rotor 245, **246**, 247
  score marks 245
  self-adjuster cable **216**
  servicing brake discs 245, 246, 247
  servicing front brake calipers 225, 226, 227
  shoes 36, 217, 223, 225, **247**
  star-wheel screws **216**, 217
  testing brake fluid 229
bumpers 298, 299, 300, 301
  3M P.R.O. repair system 298
  adhesion promoter **301**
  fiberglass cloth **299**
  fiberglass-reinforced polyester (FRP) 299
  finish 301
  repairing plastic bumpers 298, 299, 300, 301
  particle mask 299
  sanding **298**, 301

  - sheet-molding compound (SMC) 299

C

- cabin air filters **260, 261, 262**
  - case 261
  - cleaning 261
  - covers 260
  - how they work 262
  - how to buy 262
  - HVAC case 260, 261, 262
  - interior-mounted **262**
  - locating 260
  - midcase 261, 262
  - replacement 260, 261, 262
  - under-dash **261**
- calipers 39, **218, 225, 226, 227,** 229, **230,** 236, 246
  - fixed 226, 227
  - how rear disc calipers work with parking brakes 226
  - replacement 226
  - sliding **225,** 226
- camshaft drive belt **108**
- carburetors 37, 93, 106, 124
- catalytic converter 13, 14
- cavitation 96
- chassis 107, 214–253
  - curing premature brake wear 216, 217, 218
  - curing steering wander 251, 252, 253
  - dealing with a low brake pedal 222, 223, 224
  - diagnosing and repairing wheel vibration 238, 239, 240, 241
  - fixing suspension clunks and rattles 248, 249, 250
  - flushing your brake system 228, 229, 230
  - replacing a parking brake cable 219, 220, 221
  - replacing a steering rack 242, 243, 244
  - servicing brake discs 245, 246, 247
  - servicing drum brakes 231, 232, 233
  - servicing front brake calipers 225, 226, 227
  - solving steering-pull problems 234, 235, 236, 237
- clutch **147, 148, 149,** 150, 151
  - cable-operated **147,** 149
  - "clutch pilot tool" 151
  - how it works 148
  - hydraulic **147,** 149
  - inspection 149, 150
  - installation 151
  - mechanical **147**
  - removal 149, 150
  - replacement for front-drive 147, 148, 149, 150, 151
- coil-on-plug iginition (COP) **23**
- constant-tension worm-drive clamp 102, 103, **104**
- coolant 30, 31, 38, 102, 131
- cooling system 102, 103, 104, 112, 113, 114, 115, 132, 133, 134, 196
  - bleeding 113
  - coolant 102
  - draining coolant 112, 113
  - flushing 112, 113, 114, 115
  - hose clamps **102, 103,** 104, 113
  - hose removal 113, 132
  - hoses 104, **113**
  - how coolant hose leaks occur 103
  - how cooling systems work 114
  - overflow tank **112**
  - picking the right antifreeze 114, 115
  - refilling 113
  - repairing leaks 102, 103, 104
- crankshaft 109, 117, 150, 151
- crankshaft sensor 118
- computers **189, 190, 191**
  - computer-safe test light **189**
  - computer system data analyzer 191
  - diagnostic plug 189
  - in air duct assembly **189**
  - installation 191
  - map sensor **189**
  - On-Board Diagnostics II (OBD II) 191
  - replacement 189, 190, 191

D

- dampers *see* shock absorbers
- detailing your car 272, 273, 274
  - brushes 272
  - burn and hole repair 272
  - carpeting 272
  - carwash soap 273
  - chamois 274
  - clearcoat paint 273
  - conditioner 2372
  - cotton swabs **272,** 273
  - dashboards 272
  - floor mats 272
  - gauge lenses 273
  - interior 272
  - *Meguiar's Quick Detailer* 274
  - polish 272, 274
  - seats 272
  - scratches **273**
  - *Simple Green* cleaner 272
  - stain cleaner 272
  - tinted windows 273
  - tires 274
  - toothbrush 272, **274**
  - upholstery 272
  - vacuuming 272
  - ventilation grilles 272, 273
  - vinyl repair kits 272
  - wax 272, 273, **274**
  - wheels 274
  - windshield 273
  - wood trim 273
- Digital Multi-Meter (DMM) 180
- distributor cap **22, 23**
- dome-light switch **184**
- double-wire clamp 102
- drain pan 144
- *The Driver's Seat* 259
- drivetrain 138–157
  - changing your transmission fluid 152, 153, 154
  - curing a harsh-shifting automatic 140, 141, 142, 143
  - replacing an automatic transmission filter 144, 145, 146
  - replacing a front-drive car's clutch 147, 148, 149, 150, 151
  - replacing u-joints 155, 156, 157
- drum brakes 216, 217, **231,** 232, 233
  - chamfering 233
  - diagnosis 231
  - dial indicator **232**
  - installation 232, 233
  - linings 231
  - removal 231, 232
  - servicing 231, 232, 233
  - shoes **233**
- drying out a flooded car 291, 292, 293
  - brake fluid 293
  - doors 291
  - electrical systems 293
  - engine management computer **292**
  - engine seals **291**
  - filter 291
  - headlamps 293

  high-water marks 291, 293
  how salvage titles work 293
  interior damage 291, 292
  oil 291
  taillights 291
  transmission 291

E

Earl Scheib 278
electric cooling fan 125, 126, 127
  checking motors 125, 126
  how it works 127
  problems 125, 126, 127
  relay diagnosis 127
  temperature gauge 125
  testing cooling switch/sensor **127**, 127
electrical and electronic systems 158–213
  changing a fuse 210, 211, 212
  curing air conditioner odors 192, 193, 194, 195
  curing radio noise 166, 167, 168
  eliminating battery drain 182, 183, 184, 185
  fixing a dead horn 160, 161, 162
  fixing faulty cruise control 207, 208, 209
  fixing power windows 176, 177, 178
  low-tech work on high-tech cars 196, 197, 198, 199, 200
  repairing composite headlamps 169, 170, 171
  repairing windshield washers 163, 164, 165
  replacing your car's computer 189, 190, 191
  replacing a fuel pump 179, 180, 181
  replacing headlights 186, 187, 188
  scan tool diagnosis 201, 202, 203
  servicing electronic suspensions 172, 173, 174, 175
  wiring your trailer hitch 204, 205, 206
electrochemical degradation (ECD) 134
electronic ignition system 19
emissions maintenance reminder **73**, 74, 75
  checking emissions system 74
  finding reset location 74, 75
  resetting 73, 74, 75
  service reminder relay 73
  warning light 73
emissions tests 116
engine 88–137
  checking your ignition timing 116, 117, 118
  cleaning throttle bodies 122, 123, 124
  curing a clicking starter 99, 100, 101
  curing slow cranking 105, 106, 107, 108
  electric cooling fan problems 125, 126, 127
  finding engine knock 128, 129, 130, 131
  finding oil leaks 109, 110, 111
  flushing your cooling system 112, 113, 114, 115
  plugging leaks 132, 133, 134
  repairing cooling system leaks 102, 103, 104
  replacing engine mounts 119, 120, 121
  replacing your water pump 96, 97, 98
  smoothing out a lumpy idle 93, 94, 95
  today's tuneup 90, 91, 92
  turning up your heater 135, 136, 137
engine cranking 105, 106, 107, 108
  camshaft drive belt **108**
  curing slow cranking 105, 106, 107, 108
  ignition timing 108
  load test 106
  oil thickness 107
  starter motor 105
  weather related 106
engine idle 93, 94, 95
  checking intake tract **93**
  cleaning throttle bore **94**
  how engine management system works 94
  idle speed controller **95**
  smoothing out a lumpy idle 93, 94, 95
engine knock 128, 129, 130, 131
  compression testing 129, 130, 131
  leakdown testing 129, 130
  locating 128, 129, 130, 131
  testing 129
  use of stethoscope 128, 129
  vacuum testing **129**
  valvetrain 129
engine mounts **119, 120, 121**
  "dog bone" struts 119, 121
  how hydraulic engine mounts work 120
  installation 121
  replacement 119, 120, 121
  rubber bushings 119
  silicone leaks 120
  vacuum pump usage **121**
exhaust gas recirculation (EGR) valve 74, **92**, 94
exhaust manifold 20, 21
exhaust system 13, 14, 62, 63, 64, 197
  hangers 62, 63
  joints 63, 64
  maintenance 62, 63, 64
  muffler 63
  shields 63

F

flat-fix foam 10
flywheel **150**, 151
fuel filter 179, 180
  removal 180
fuel injector 19, 95
fuel pump 179, 180, 181
  filter 179, 180
  fuel pressure gauge **181**
  gerotor **181**
  how it works 181
  relieving pressure 180
  replacement 179, 180, 181
  roller cell **181**
  schrader valve 180
  turbine **181**

G

gaskets 8, 14, 32, 98, 100, 119, 131, 197, 276, 285, 287
  how they work 110
Gate Krik-It gauges 68
Gates Rubber Company 134

H

head gasket 131
headlamps **169**, 170, **171**, **186**, **187**, **188**
  beam height adjustment 170, 171
  bubble level 169, 171
  circuit check 176

  - daytime running lights (DRL) 187, 188
  - high intensity discharge lighting (HID) 186
  - how high intensity discharge (HID) lighting works 170
  - how quartz halogen beams work 187
  - quartz halogen bulbs 169
  - repairing composites 169, 170, 171
  - replaceable bulbs 188
  - replacement 186, 187, 188
  - sealed-beam lamps 188
  - thumbwheel **169,** 171
- high-tech cars 196, 197, 198, 199, 200
  - basic maintenance on 196, 197, 198, 199, 200
  - brakes 198, 199
  - cooling system draining 196
  - drive belt replacement 196, 197
  - electric sytems 199, 200
  - exhaust system 197
  - gaskets 197
  - shocks and struts 199
  - thermostat replacement **197**
  - tuneup 199
- horn **160, 161, 162**
  - button **162**
  - fuse **160**
  - inspection 160, 161, 162
  - mounting 160
  - relay socket **162**
  - repair 160, 161, 162
- hydrometer **49, 50, 106**

I

- ignition timing 116, 117, 118
  - checking 116, 117, 118
  - cleaning timing tab 116
  - finding timing marks 117, 118
  - how a distributorless ignition works 117
  - spout connector **118**
  - timing light 116, **118**
- intake air duct 122
- intake-manifold vacuum 129
- Intake snake 123
- interior and exterior 254–313
  - aligning body parts 275, 276, 277
  - detailing your car 272, 273, 274
  - drying out a flooded car 291, 292, 293
  - finding and fixing water and air leaks 285, 286, 287
  - fixing hood and trunk latches 311, 312, 313
  - freshening up your musty interior 308, 309, 310
  - hushing interior squeaks and rattles 302, 303, 304
  - keeping your windshield clear 269, 270, 271
  - polishing your car 297
  - prepping your car for paint 278, 279, 280, 281
  - re-covering seats 256, 257, 258, 259
  - removing bumper stickers, window tinting and pinstripes 305, 306, 307
  - repairing paint chips 294, 295, 296
  - repairing plastic bumpers 298, 299, 300, 301
  - repairing a rear-window defroster 266, 267, 268
  - repairing windshields 282, 283, 284
  - replacing a cabin air filter 260, 261, 262
  - replacing a rearview mirror 263, 264, 265
  - storing your car 288, 289, 290

J

- jackscrew 67, 68
- *J.C. Whitney* 70, 259, 275
- jounce bumpers 57
- jumper cables 29, 39
- jumpstarting your car 27, 28, 29

L

- leakage tester **130**
- load test 106
- lug nuts 10, **34**
- lug wrench 10, 11

M

- *Maaco* 278
- *MacPherson* struts 58, 237
- maintenance
  - basic 2–87
  - changing a tire 10, 11, 12
  - changing your oil and filter 79, 80, 81
  - charging your battery 85, 86, 87
  - curing pulsing brakes 34, 35, 36
  - jumpstarting your car 27, 28, 29
  - maintaining batteries 48, 49, 50
  - maintaining exhaust systems 62, 63, 64
  - maintaining tires 24, 25, 26
  - maintaining and repairing your suspension 54, 55, 56, 57
  - passing an emissions inspection 76, 77, 78
  - repairing a flat tire 40, 41, 42, 43
  - repairing a leaky radiator 30, 31, 32, 33
  - repairing power antennas 16, 17, 18
  - replacing an air filter 7, 8, 9
  - replacing a battery 51, 52, 53
  - replacing belts 65, 66, 67, 68
  - replacing hood and hatch struts 69, 70, 71, 72
  - replacing shock absorbers 58, 59, 60, 61
  - replacing spark plugs 19, 20
  - replacing spark plug wires 21, 22, 23
  - replacing a tailpipe 13, 14, 15
  - replacing wiper blades 4, 5, 6
  - resetting emissions maintenance reminder (EMR) 73, 74, 75
  - search engines 82, 83, 84
  - servicing your air conditioner 44, 45, 46, 47
  - winterizing your car 37, 38, 39
- MAP (manifold absolute pressure) engine vacuum sensor 91, 93, 95
- *Marson's Poly-Fill* 280
- *Meguiar's Quick Detailer* 274
- *Michael Ostrowski and John Crane International* 97
- Mighty Vac 224
- micrometer 34
- muffler 13, 14, 63
  - how it works 63

N

- National Automotive Radiator Service Association (NARSA) 32

O

- On-Board Diagnostics (OBD) 19, 116, 201, 202, 203
- OBD II computers 92, 116, 191, 201, 202, 203
  - how OBD II transmits data 203

  also see computers
odometer 96, 179
oil leaks 109, 110, 111
  locating 109, 110, 111
  main seal **109**
  oil pan gasket **109**
  use of foot powder 111
  use of trace dye 110, 111
oil viscosity 38
ohmmeter **16**, 17, 44, 46, 47, **126**
oxygen sensor 20

P

paint 273, 278, 279, 280, 281, 294, 295, 296
  body hammer **280**
  bodywork 278, 280
  clearcoat **273, 295**
  Earl Scheib 278
  glazing compound 296
  hardware and trim removal 278, 279, 280
  lacquer thinner **294**
  Maaco 278
  Marson's Poly-Fill 280
  masking 278, 279, **281, 296**
  plastic filler 280
  prepping your car 278, 279, 280, 281
  primer 280, **281, 295**
  professionals 278
  repairing chips 294, 295, 296
  sanding 278, 279, 281, 295
  touchup 294
parking brakes **219, 220, 221**, 222, 223
  and four-wheel disc brakes 221
  cable replacement 219, 220, 221
  how self-adjusting brakes work 221
  installation 220, 221
  lever 220
  pedal **221**
Performance Products 259
plugging leaks 132, 133, 134
  coolant leaks 132
  cooling system pressure tester **133**
  electrochemical degradation (ECD) 134
  finding hidden leaks 133, 134
  fixing obvious leaks 132, 133
  preventing leaks 134
polishing your car **296**
  lamb's-wool bonnet 297
  paint finish 297
  power buffer 297
  sponges 297
  wax 297
portable compressor 40
power steering reservoir 109
powertrain computer 127
powertrain control module (PCM) 94

R

radiators **30, 31**, 32, 37, 97, 119
  downflow 30, 31
  removal 31, 32
  repairing leaks 30, 31, 32, 33
  tank inspection 30
  types of leaks 32, 33
radio 166, 167, 168
  antenna **167**
  curing noise 166, 167, 168
  geiger-counter test 166, 167
  interference sources 166
  radio frequency interference (RFI) 166, 167
  RFI filter **167, 168**
rearview mirror **263, 264, 265**
  accelerant for mounting 264, 265
  adhesive kit 264
  and windshield 263, 264
  baseplate **264, 265**
  how antiglare mirrors work 263
  installation 264, 265
  lockscrew **265**
  reattachment 264
  removing old adhesive 263, 264, 265
  replacement 263, 264, 265
  toggle **263**
rear-window defroster **266, 267, 268**
  abrasive warnings 266
  connector tabs 266, 267
  diagnosing problems 266, 267
  defroster grid 266, 267, **268**
  easy fixes 267, 268
  fuse 266
  gaskets 267
  overheating 267
  relays 267
  repairing 233, 267, 268
  switches 267
  timers 267
  use of conductive epoxy adhesive 267
  use of conductive paint **268**
  use of soldering iron on 266, 267
  voltmeter check 266, 267

S

scan tools **91**, 118, 127, **201, 202, 203**
  Actron Scan Tool 202
  Auto Xray 202, 203
  code readers **203**
  diagnosis 201, 202, 203
  how OBD II transmits data 203
  On-Board Dignostics (OBD) 201, 202, 203
  On-Board Diagnostics II (OBD II) 201, 202, 203
Schrader valve **180**
seats **256, 257, 258, 259**
  backs 256, 257, **258, 259**
  bolsters 256
  choosing slipcovers 256, 257
  cushions 256, 257
  flex cords **258**
  re-covering 256, 257, 258, 259
  disassembly 257, 258
  hinges **256**
  reclining 257, 259
  reinstallation 259
  removal 256, 257, 258
  Scotchgarding 259
  slipcover installation 257, 258
  suppliers 259
serpentine belts **66, 68**, 98
  how they work 68
  routing diagrams 98
shock absorbers **56, 57, 58, 59**, 60, **61, 64, 174**, 199
  how they work 59
  installation 61
  mountings 59
  removal 60
  replacement 58, 59, 60, 61, 174, 175
shrink-band clamps **102**, 103
silicone sealers 98
Simple Green cleaner 272
slave cylinder 148, 149
solenoid valve 120, 121
spark plugs **19, 20**, 91, 95
  choosing the right plug 20
  gapping 95
  removal 131
  replacement of 19, 20, 91
spark plug boot **21, 91**
spark plug pullers 19
spark plug wires **21, 22, 23**

  - how they work 22
  - removal 21, 22
  - replacement of 21, 22, 23
- spring-band clamps **102**, 103
- starters **99**, 100, **101**, **105**
  - curing a clicking starter 99, 100, 101
  - iginition 99
  - relay 99
  - self-tapping repair stud **101**
  - solenoid **99**, 100, 101
  - terminal inspection 100, 101
- steering 234, 235, 236, 237, 240, 241, 242, 243, 244
  - and brake drag 234, 235
  - and suspension 235, 236, 237
  - arm **242, 243**
  - bleeding 244
  - column **242**
  - control arm **235**
  - flushing pump and lines 243
  - how power rack-and-pinion steering works 244
  - how rack-and-pinion steering works 240
  - memory steer 237
  - pull problems 234, 235, 236, 237
  - rack housing **244**
  - steering rack replacement 242, 243, 244
  - tie-rod ends **234**, 235, 237, **239**, 240, **242**, **243**
  - tire pull 236, 237
  - torque steer 234, 235
  - transaxle mount 236
  - wheel 234, 235, 241
- storing your car 288, 289, 290
  - battery removal 289
  - body covers 290
  - exhaust system protection 289, **290**
  - float charger **289**
  - indoor storage 289
  - outdoor storage 289, 290
  - post-storage preparation 290
  - rust prevention 288
  - vacuuming **288**, 289
  - vehicle preparation 289
  - windshield protection 289, **290**
- struts **69, 70, 71, 72**, 199, **236**, 237
  - air strut **172**
  - ball joint mount 69
  - "dog bone" 119, 121
  - flat C-clip 70
  - hood and hatch replacement 69, 70, 71, 72
  - how support struts work 72
  - installation 70, 71
  - wire C-clip 70
- support bar **149**
- suspension 39, 54, 55, 56, 57, 172, 173, 174, 175
  - actuator precautions 175
  - air strut **172**
  - air suspension 173
  - Corvette diagnostics 175
  - electronic problem diagnosis 172, 173
  - how leaf spring suspension works 55
  - leaf springs **54, 55**
  - maintenance and repair 54, 55, 56, 57
  - Selective Ride Control (SRC) system 172, 173
  - servicing electronic suspensions 172, 173, 174, 175
  - shock replacement 174, 175
  - steering stabilizers 56, 57
  - visual inspection 173

T

- tailpipes **13, 14, 15**
  - removal 13, 14
  - replacement 13, 14, 15
- test light (12-volt) **45, 46**
- Teves Mark II ABS systems 224
- throttle bodies 122, 123, 124, 141
  - checking for deposits 122
  - cleaning 122, 123, 124
  - hose disconnection 122, 123
  - how it works 123
  - intake air duct 122
  - Intake snake 123, **124**
- throttle -position sensor (TPS) 47
- tires **10, 11, 12, 24, 25, 26, 40**, 41, 42, 43, **240**, 274
  - checking pressure 10, 24, 25
  - how to read a passenger tire sidewall 26
  - inflation of 43
  - inspection 24, 25
  - maintenance 24, 25, 26
  - radial runout 240, 241
  - repairing a flat 40, 41, 42, 43
  - rotation 26
  - sidewall check 26, 41
  - snow 39
- tire jack **10, 11, 12**
- tire pressure gauge **10, 24**, 25
- tire reamer 42
- Tracer Products Leak-finder Kit 110
- transaxle 149, 150, 151
- transmission 107
- transmission bellhousing 117
- transmission fluid 32
- trickle charger **29, 182**
- tuneups 90, 91, 92
  - battery maintenance **92**
  - exhaust gas recirculation 92
  - filter replacement 90
  - service information 92
  - spark plug replacement 91
  - underhood checks 90, 91

V

- vacuum gauge **129**
- Vadja, Gabe 147
- valvetrain 143
- V-belt 98
- V8 engine 106
- vibration damper 143
- voltage-drop test 180
- voltmeter 99, **100**, 142, **161**, 162, **176, 266**

W

- water and air leaks 285, 286, 287
  - C pillars 286, 287
  - chalk testing 285
  - finding and fixing 285, 286, 287
  - gaskets 285, 287
  - guards 287
  - repair products 287
  - silicone seal 285
  - trimming adhesive 285, **287**
  - weatherstripping 285, 286
  - wind noise 285
  - worn hinges 286
- water pumps 96, **97, 98**, 133
  - Advanced Metal Diaphragm seal 97
  - how shaft seals work 97
  - inspection 96, 97
  - replacement 96, 97, 98
  - seals 96, 97
- wheels 25, 26, 40, **238, 239, 240**, 241, 274
  - alignment 25, 26
  - balancing 25, 238, 239
  - bearings 239, 240
  - constant velocity joints 239

  diagnosing and repairing vibration 238, 239, 240, 241
  how rack-and-pinion steering works 240
  lateral runout 241
  radial runout 240, 241
  rims 241, 274
  tie-rod ends **239**
  tires **240**
  weights 238
  well 40
windows 176, **177**, **178**, **266**, 267, **268**
  backprobing switches 176
  fuses 176, 177
  how automatic windows work 177
  motor **177**
  power 176, 177, 178
  rear 266, 267, 268
  rear defroster repair 266, 267, 268
  reel and cable regulators 178
  repair 176, 177, 178
  weatherstripping 177, **178**
windshield **269**, 270, 271, 273, **282**, 283, **284**
  and dashboard ducts 271
  and wiper system 269, 270, 271
  blower fan **271**
  defogger 269, 271
  fog 271
  interior air pollution 271
  interior moisture 271
  keeping it clear 269, 270, 271
  silicone build-up 269, 270
  smearing and streaking 260, 270
  squeegee replacement 270
  use of vinegar on 271
  ventilation control system 271
  wiper blade lubricant 270
  wiper motor 270
windshield repair 282, 283, 284
  acrylic adhesive 282, 283
  burnishing 283
  chips and cracks **282**
  how laminated safety glass works 283
  preparation 282, 283
  pressure 283, 284
  repair kits 282
  vacuum 283, 284
windshield wiper system 4, 5, 6, 38, 39, 163, 164, 165, 269, 270, 271
  blade replacement **4**, **5**, **6**
  how washer system works 165
  how wipers work 5
  nozzles **164**, 165, 270
  pump **163**
  rear systems 165
  repairing system 163, 164, 165
  reservoir **163**
  squeegee **270**
  spray patterns 164, 165
  types of blades 6
  wiper blade lubricant 270
winterizing your car 37, 38, 39
worm-drive clamp 102

(continued from the front flap)

cars—there's still plenty for you to do and tinker with. Does your water pump need replacing? Do it yourself. Have you been avoiding hills to put off replacing the clutch of your front-wheel-drive car? Don't wait any longer.

Judging the type of job, skills required, and time to tackle it in advance is a simple task. Just check out the easy, moderate, or difficult label for the project you're interested in. No matter what level of expertise is required, you'll accomplish many jobs you may have thought impossible.

Stay within your budget and feel confident about the condition of your car.

Hearst Books
A Division of Sterling Publishing Co., Inc.
New York

www.popularmechanics.com

Printed in China

TRANSPORTATION/AUTOMOTIVE

The Essential Car Care Manual Now Revised and Expanded

Is your rear window defroster not working? Does your fuel pump need replacing? Whatever job you need to do on your car—or discuss intelligently with a professional—is explained in this larger, updated edition of *the* classic automotive guide. The editors of *Popular Mechanics* put their experience to work for you by providing even more advice and additional projects, ranging from basic maintenance to advanced care, in this revised, up-to-date volume.

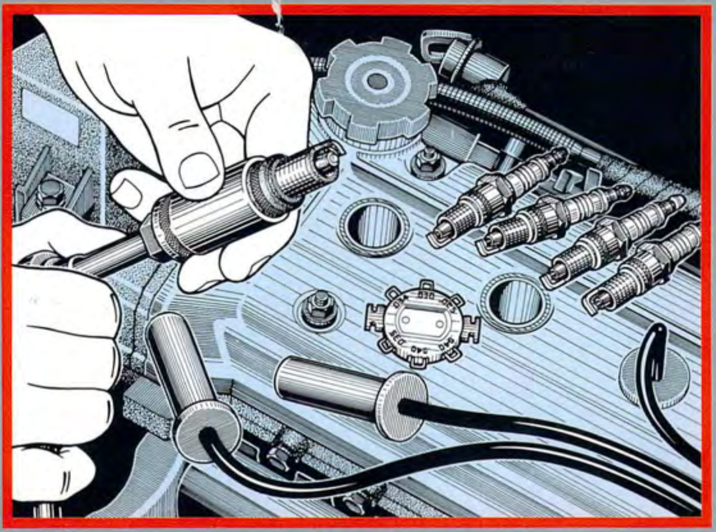

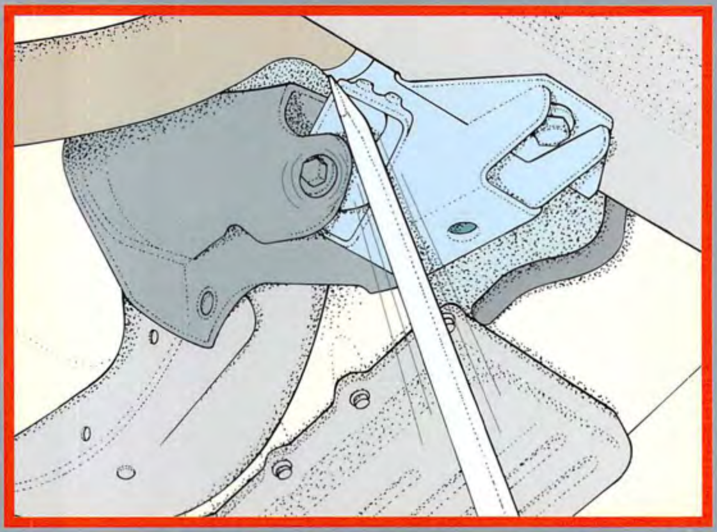

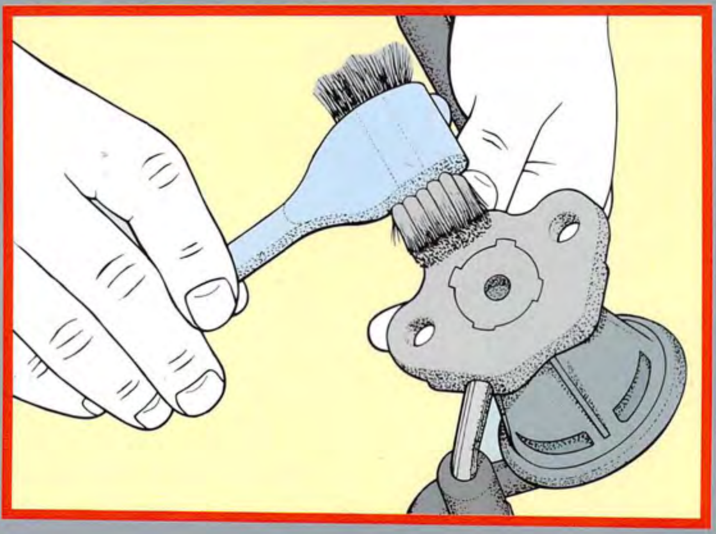

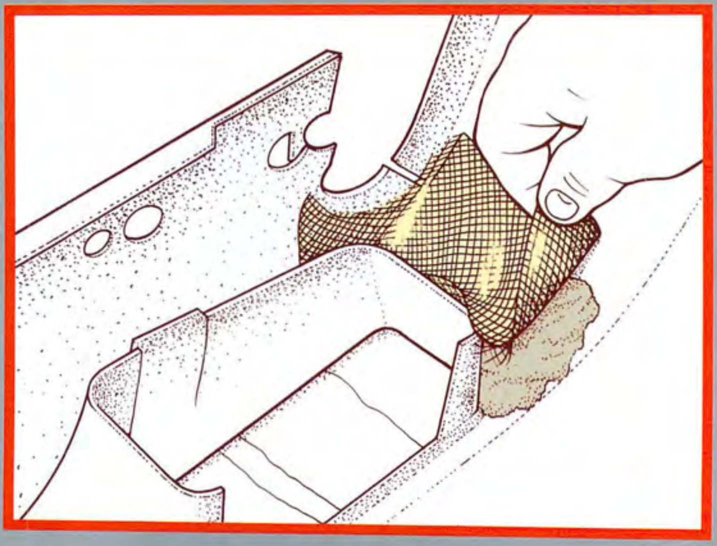

ISBN 1-58816-439-X
EAN
52995>
9 781588 164391

Hearst Books
A Division of Sterling Publishing Co., Inc.
New York

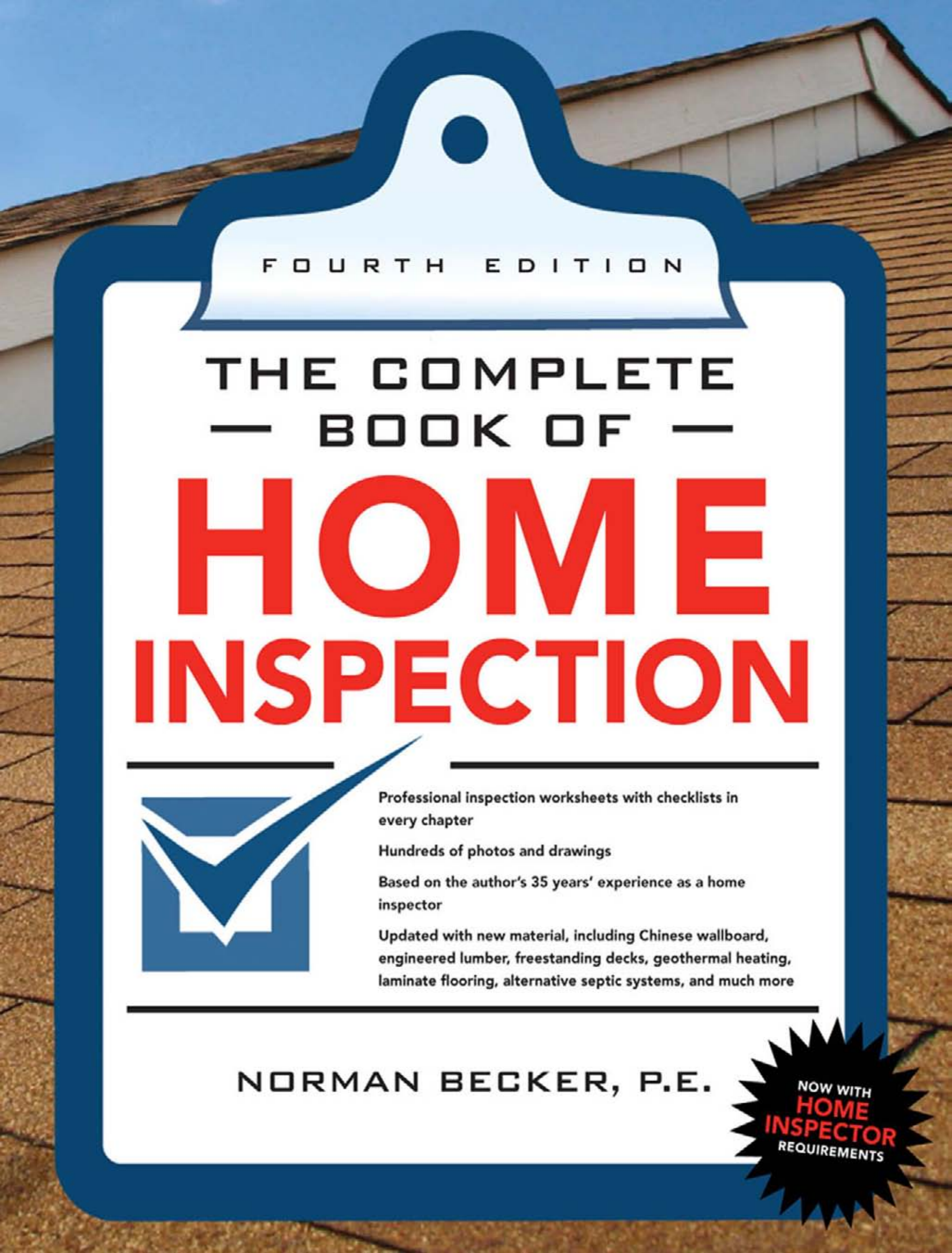

FOURTH EDITION

THE COMPLETE
— BOOK OF —
HOME INSPECTION

Professional inspection worksheets with checklists in every chapter

Hundreds of photos and drawings

Based on the author's 35 years' experience as a home inspector

Updated with new material, including Chinese wallboard, engineered lumber, freestanding decks, geothermal heating, laminate flooring, alternative septic systems, and much more

NORMAN BECKER, P.E.

NOW WITH
HOME
INSPECTOR
REQUIREMENTS

The
Complete Book of
Home Inspection

This page intentionally left blank

The
Complete Book of
Home Inspection

Fourth Edition

Norman Becker, P.E.

Mc
Graw
Hill

New York Chicago San Francisco Lisbon London Madrid
Mexico City Milan New Dehli San Juan Seoul
Singapore Sydney Toronto

The McGraw-Hill Companies

Copyright © 2011, 2002, 1993, 1980 by The McGraw-Hill Companies, Inc. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

ISBN: 978-0-07-170278-2

MHID: 0-07-170278-4

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-170277-5,
MHID: 0-07-170277-6.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. To contact a representative please e-mail us at bulksales@mcgraw-hill.com.

Information contained in this work has been obtained by The McGraw-Hill Companies, Inc. ("McGraw-Hill") from sources believed to be reliable. However, neither McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that McGraw-Hill and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. ("McGrawHill") and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill's prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED "AS IS." McGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

For Renée, Jeffrey, Piper, Ana Maria, Sharon, Mia, and Eli

This page intentionally left blank

About the Author

Norman Becker, P.E. (Hamburg, New Jersey), has more than 35 years' experience in home inspection. He has inspected homes of all ages, from newly constructed to prerevolutionary, from vacation homes to stately mansions. He is one of the founders of the American Society of Home Inspectors (ASHI) and has been qualified in court as an expert witness on the subject of home inspection. He also wrote the widely read "Homeowners Clinic" column for *Popular Mechanics* magazine for 24 years and is the author of the Popular Mechanics book *500 Simple Home Repair Solutions.* Mr. Becker is a licensed professional engineer in New York, New Jersey, and Florida.

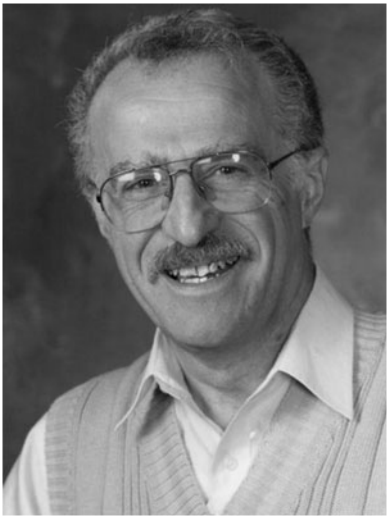

This page intentionally left blank

Contents

Acknowledgments **xvii**

Introduction **xix**

1 Tools and procedure **1**

Tools needed 1
Inspection procedure 2
Exterior inspection 2
Interior inspection 3
Electromechanical systems 4
Final inspection 5

EXTERIOR INSPECTION

2 Roofs **6**

Pitched roofs 6
  Shingles 8
Flat roofs 16
  Built-up roofs (BURs) 17
  Single-ply membrane 19
  Roll roofing 19
  Metal roofing 19
Checkpoint summary 20

3 Roof-mounted structures and projections 22
Chimneys 22
Vent stacks 26
Roof vents 27
Roof hatch 28
Skylights 28
TV antennas 28
Lightning protection 29
Gutters and downspouts 29
  Built-in gutters 30
  Exterior-mounted gutters 30
  Downspouts 32
Checkpoint summary 34

4 Paved areas around the structure 36
Sidewalks 36
Street-level/driveway-level steps 36
Front and side paths 37
Entry steps 37
Patio 38
Driveway 39
Checkpoint summary 41

5 Walls, windows, and doors 42
Exterior walls 43
  Exterior siding 44
  Masonry wall 51
Trim 52
Windows 53
  Inspection 55
Exterior doors 56
Storm windows, screens, and storm doors 58
  Storm windows 58
  Screens 58
  Storm doors 59
  Inspection 59
Caulking 59
Checkpoint summary 60

6 Lot and landscaping 62
Drainage 62
  Groundwater, water table 65
  Inspection 67
Retaining walls 70
  Inspection 71

Landscaping 74
    Lawn 74
    Shrubs 75
    Trees 76
Decks 76
    Free-standing decks 79
    Composite decks 79
Fences 80
Checkpoint summary 81

7 Garage 84
Attached garage 84
    Fire and health hazards 84
    Plumbing check 86
    Flood potential 88
    Doors 88
    General considerations 89
Detached garage 90
    Exterior 90
    Interior 91
    Heat and electricity 92
Checkpoint summary 93

8 Wood-destroying insects and rot 95
Termites 95
    Subterranean termites 98
    Dry-wood termites 105
    Formosan termites 106
Other wood-destroying insects 106
    Carpenter ants 106
    Powder-post beetles 108
Rot 109
    Inspection 110
Checkpoint summary 112

INTERIOR INSPECTION

9 Attic 114
Insulation 115
Violations 116
Leakage 117
Fire hazards 117
Ventilation 118
Attic fan 120
Whole house fans 120
Structural 121
Checkpoint summary 123

10 Interior rooms 124

Walls and ceilings 125
  Mold-resistant drywall 126
  Chinese drywall 128
Floors 128
Heat 131
Windows 131
Electrical outlets 132
Fireplace 134
  Gas-fired fireplace 136
Bedrooms 136
Bathrooms 137
  Water pressure and flow 138
Kitchen 141
Hallway and staircase 142
Checkpoint summary 143

11 Basement and crawl space 147

Foundation 147
  Settlement 148
  Inspection 149
Dampness 154
Water seepage—causes and control 155
  High groundwater level 155
  Hydrostatic pressure—walls 156
  Inspection 158
Furnace room 163
  Ventilation 165
Crawl space 165
  Conditioned crawl space 167
Checkpoint summary 168

ELECTROMECHANICALS

12 Electrical system 170

Inlet electrical service 170
Electrical capacity 171
Fuses and circuit breakers 174
Inlet service panel box 175
  Panel-box interior 176
  Aluminum wiring 176
Grounding 177
  Improper ground connections 177
Interior electrical inspection 178
  Electrical outlets 178

  - Knob-and-tube wiring 180
  - Low-voltage switching systems 180
- Violations 181
- Checkpoint summary 181

13 Plumbing 183

- Water supply and distribution 184
- Fixtures 185
- Drainage system 185
- Waste-disposal system 186
  - Cesspool 187
  - Septic system 187
- Exterior inspection 191
  - Vent stack 191
  - Fresh-air vent 192
  - Lawn sprinkler systems 192
  - Septic system 193
- Interior inspection 193
  - Fixtures 193
  - Cross-connection 193
  - Water pressure, flow 194
  - Plumbing wall hatch 194
  - Pipes 195
- Basement inspection 197
  - Water-supply pipes 197
  - Distribution piping 198
  - Drainage pipes 199
- Well-pumping systems 202
  - Wells 202
  - Well pumps 203
  - Accessory equipment 205
- General considerations 206
- Checkpoint summary 207

14 Heating systems I 209

- Central heating systems 210
- Heating outlets: registers and radiators 211
- Thermostat and master shutoff 211
- Warm-air systems 212
  - Advantages 213
  - Disadvantages 213
  - Gravity warm air 213
  - Forced warm air 215
  - Heat pump 218
  - Geothermal heating/cooling 218

Hot-water systems 219
    Gravity hot water 219
    Forced hot water 221
Steam heating systems 227
    Distribution piping 228
    Controls 229
    Domestic water heater 231
    Advantages and disadvantages 231
Hybrid heating systems 231
    Steam–hot water 231
    Hydro-air 231

15 Heating systems II 233
Oil-fired systems 233
Gas-fired systems 235
Electrical systems 236
Area heaters 236
    Gas-fired units 237
    Electrical units 237
Heating system inspection procedure 237
    Warm-air systems 239
    Hot-water systems 241
    Steam systems 242
    Oil burners 243
    Gas burners 244
Checkpoint summary 244

16 Domestic water heaters 247
Tank water heaters 248
    Reversed connection 248
    Relief valve 249
    Rumbling noise 251
    Exhaust stack 252
    Operational inspection 252
    Water heater replacement 254
    Water heater capacity and recovery 255
Tankless (coil) water heaters 256
    Inspection procedure 258
Tankless (wall-mounted) water heaters 259
Indirect-fired storage water heaters 260
Checkpoint summary 261

17 Air-conditioning 262
Air-conditioning capacity 263
Air-conditioning systems 264
    Integral system 264

  - Split system 264
  - Heat pump 272
  - Evaporative cooler 273
- Checkpoint summary 273

18 Swimming pools 275

- Concrete pools 275
- Vinyl-lined pools 276
- Preformed fiberglass pools 277
- Pool equipment 278
- Pool accessories 280
- Inspection procedure 280
- Checkpoint summary 282

OTHER ISSUES

19 Energy considerations 283

- Energy audit 283
- Insulation 284
- Attic ventilation 287
- Storm windows 288
- Caulking and weatherstripping 289
- Fireplaces and wood-burning stoves 289
- Heating and air-conditioning systems 290

20 Environmental concerns 292

- Radon 292
- Asbestos 295
- Drinking water 297
- Lead 298
- Formaldehyde 300
- Leaky oil tanks 300
- Electromagnetic fields 301
- Mold 302

21 Green home technology 304

- Roof 307
- Roof-mounted structures 307
- Paved areas around the structure 307
- Exterior walls—siding (cladding) 308
- Deck 308
- Windows 308
- Structural framing 309
- Heating, ventilation, and air-conditioning (HVAC) 309
- Water heating 310

22 Conclusion 312

Professional home inspection 312
Home inspection limitations 312
Real estate warranty program 313
Private home inspection warranty programs 314
Manufacturers' warranties 314
Contract 314

Appendix 316
Home inspector requirements 316
Quiz based on *The Complete Book of Home Inspection* 319
Quiz answers 334
ASHI Standards of Practice 335
ASHI Code of Ethics 343
ASHI membership categories and requirements 344

Glossary 345

Inspection worksheets 348

Index 355

Acknowledgments

I would like to thank Victor J. Faggella of Centurion Home Inspections in Mahopac, New York, for his suggestions and overall help and support in bringing this fourth edition to a timely finish. Thanks to David Stewart of Expert Home Inspections in Hamburg, New Jersey, for his helpful comments on New Jersey's licensing law. Also, many thanks to Douglas Hansen in San Rafael, California, for his help with the quiz at the end of this book, which is a test to see if you were asleep when you were reading the book.

This page intentionally left blank

Introduction

Over the years various new products and systems become available for house construction that are of interest to home buyers and owners as well as to home inspectors who have to examine and evaluate them. With each new edition of this book, I have tried to include those products and systems as well as items that may have inadvertently been omitted in the previous edition.

In this fourth edition, most of the chapters have been expanded to include additional items that should be checked during an inspection. Some of these topics are metal shingles, synthetic slate, fiber-cement siding, composite decking, free-standing decks, Superior Walls, conditioned space, Chinese wallboard, mold-resistant drywall, engineered lumber, laminate flooring, engineered hardwood flooring, bamboo flooring, air admittance valves, loop vents for island kitchens, hydro-air, geothermal heating and cooling, electronic low-water cutoffs, alternative septic systems (mound systems and alternating drainfields systems), PEX piping, recirculating hot water systems, tankless water heaters, and arc fault circuit interrupters. In addition, there is a new chapter titled "Green Home Technology" and an appendix that includes home inspector requirements.

By following the procedures outlined in this book, a home buyer will be able to look beyond the cosmetic and have a good idea as to the true condition of the house. The book is also helpful to an owner, who can use it to determine problems or potential problems in the house. There could be problems caused by deterioration as a result of aging or there could be safety or fire hazards. Very often the problems could be quite minor and could be corrected at little or no cost; however, if left unattended they could be quite costly to correct.

The section on home inspector requirements was included because I felt it was important to inform the public about the demanding requirements and the training needed to become a professional home inspector. It will also be of interest to the many engineers, architects, and contractors who would like to work as home inspectors. The book should also be quite helpful to a buyer who wants to hire the services of a home inspector to check the condition of a prospective home.

Norman Becker, P.E.

The
Complete Book of
Home Inspection

This page intentionally left blank

1

Tools and procedure

Tools needed 1
Inspection procedure 2
Exterior inspection 2
Interior inspection 3
Electromechanical systems 4
Final inspection 5

The inspection procedure outlined in this book is similar to the one that I use when inspecting residential structures. Because the various components of homes are basically the same, this procedure is valid regardless of the geographic location of the structure. It has been used on homes of all ages, from newly constructed to pre-Revolutionary, from vacation homes to stately mansions.

Tools needed

To inspect the house properly, you will need the following tools: a flashlight, to see in dark places (and you'll be surprised what you might find); a magnet, to determine whether plumbing pipes are iron; a marble, to note whether the floors are relatively level; an ice pick or screwdriver, to aid in looking for wood rot and termite infestation; a 6-foot stepladder, for those houses that do not provide direct access (built-in or pull-down steps) to the attic; an electrical tester, for checking the electrical ground connection and the electrical polarity, particularly in the kitchen and bathroom outlet receptacles; binoculars, to get a closer look at the roof and roof-mounted structures; a thermometer to check the temperature of the domestic hot water; and a compass, to determine the building's exposure. Knowledge of the exposure is helpful in evaluating the condition of various structural elements and components.

Since the first edition was published, a number of electronic tools have come on the market that are being used by professional home inspectors. The one tool that I would recommend for homeowners and home buyers is a battery-operated moisture meter that checks for water leaks without damaging the surface being tested. One manufacturer of moisture meters is Delmhorst Instrument Co. They can be reached at (800) 222–0638 for information on purchasing a meter. You may have read about infrared cameras that can be used for thermal imaging of a house. The cameras are used for detecting heat loss, insulation deficiencies, and specific areas of air infiltration, and for locating moisture problems. Handling of these cameras, which are quite expensive, requires special training so that the images can be accurately interpreted. Infrared cameras are not normally used during a home inspection. Most professional home inspectors do not have the camera, and those who do charge an additional fee for the service.

When performing the inspection, you should wear old clothes. Areas such as unfinished attics, basements, and crawl spaces are often quite dusty. The last items you need for inspection are a pencil and inspection worksheets. The worksheets are provided in the back of this book and should be completed as you perform the inspection. Later, you can use these worksheets to evaluate the true condition of the house and base your decision on facts rather than emotion.

Inspection procedure

A house, no matter how large or imposing, can be easily inspected if it is divided into its component parts, such as the exterior, interior, and electromechanical systems. The exterior and interior portions can be further subdivided. By approaching the inspection in a systematic order and using the worksheets provided as a guide, all the items of any consequence will be checked.

When driving up to the house, take a moment to notice the overall topography or shape of the land. Often, the topography in the immediate vicinity of the house is level; however, the overall topography might be inclined. Consequently, the possibility exists of subsurface water movement in the direction of the house. When you see inclined topography in the general area of the house, you should be alerted to the possibility of some water seepage into the lower level of the structure.

The exterior of the building should be inspected before the interior. This order is important because it provides you with an overall view of the structure that in turn can reveal the cause for some interior problems. Specifically, water seepage into a lower level can be the result of faulty gutters or downspouts, or improper grading (the ground immediately adjacent to the house slopes toward the house rather than away). A faulty roof can manifest itself in water stains one or two levels below the roof. Cracked and open exterior joints can allow the entry of water, which you will note as cosmetic damage to interior portions of the structure.

Exterior inspection

Before you start the exterior inspection, stand in front of the house and take a compass reading. The exposure for all four sides of the building should be marked on the worksheet (i.e., Front exposure—southerly; Right exposure—easterly, etc.). The exterior inspection is performed while walking around the house twice. The first time, you should look at the roof, gutters, chimney, vent stack, and anything else that is roof-mounted. The details of what to look for and how to inspect the various components of the house are discussed in the chapters that follow. During this first pass around the house, use binoculars so that you can get a closer view of the items on the upper portion of the structure. The binoculars should not have a magnification greater than 8X (eight times actual size). A more powerful pair of binoculars will tend to exaggerate hand movement, making it difficult to see details. The slightest hand movement will cause a blurred image. This problem, however, can be eliminated if you have a (more expensive) pair of 10X power binoculars with image stabilization.

After the first pass, the condition of those items inspected should be noted on the worksheet. If your first pass around the house is in the winter just after a snow, be careful where you walk. There may be a swimming pool behind the house that is covered over with snow. If you are concentrating on the house, the pool may not be that obvious and you might step right into it. (See FIG. 1-1.)

During the second pass, you look for many types of problems. Start at the front of the house and look at all the items that are either on the front of the structure or in the front yard. Examples of these items are paths, entry steps, exterior wall siding, windows, doors, decks, landscaping, fence, and so on. All of the items normally encountered during an inspection are discussed in detail in the following chapters. You should be concerned only with those items that apply to the house you are inspecting.

After inspecting the front of the house, apply the same technique to the left side, the rear, and the right side of the building. If any items of a suspicious nature require further investigation on the interior of the structure, make a note on the worksheet as a reminder. For example, if an elbow is missing from one of the downspouts and no splash plate on the ground deflects the effluent from the downspout away from the building, you should check for water seepage from that area into the lower level of the structure. Noting this fact on your worksheet helps you remember to check the interior wall opposite the downspout for signs of water seepage.

After going around the building the second time, you should be finished with the exterior inspection. Double-check your worksheet to see if you've recorded the condition of all the items inspected. At this point, you should inspect the garage. After the garage inspection, you are ready for the inside of the house.

Interior inspection

Enter the house through the front door. Try the doorbell to make sure it is operational. It's important to remember that you are looking at a house you are interested in buying. If all goes well, this will be your home, so don't be shy or feel embarrassed about doing things that any homeowner would do. As part of the inspection, you should open and close faucets on sinks, tubs, and showers; flush toilets; open and shut doors and windows; turn on the heating system and air-conditioning system by means of the thermostats; feel the airflow from heat/cooling registers; see if radiators

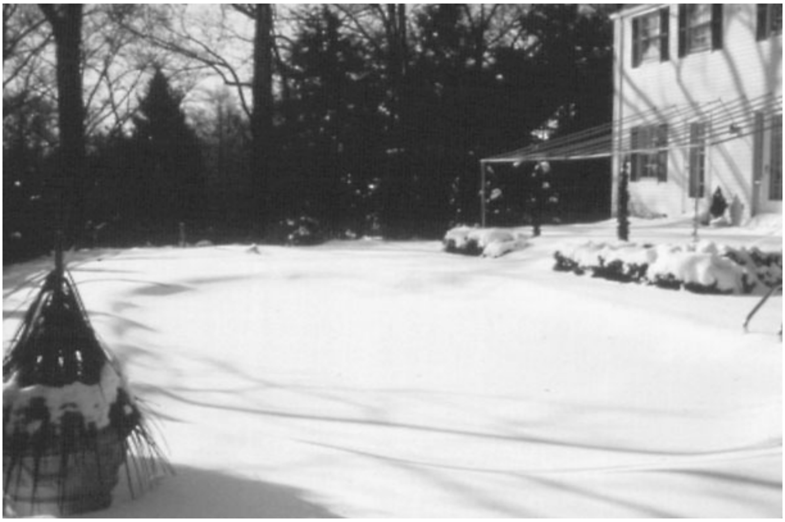

Fig. 1-1. *Snow blanketing a swimming pool cover at the rear of a house.*

get warm; and turn switches controlling lights and fans on and off.

Start the interior inspection at the uppermost portion of the building. If the house has an attic, that's where you start. To inspect the attic, you might need your ladder—check with the owner. Some homes don't have an attic, so begin this portion of the inspection with the rooms directly below the roof.

After the attic inspection, check all of the rooms on the level directly below. In some large homes you can easily miss a room. To avoid overlooking any rooms or items, begin your inspection at the entry to that level. If you start at the entry and walk either clockwise or counterclockwise, looking at each room in order, you will return to your starting point and will have inspected all of the rooms. However, if you jump around from one side to another, you can easily overlook a room or two. Again, there is no substitute for good procedure.

After all the rooms on one level have been checked, proceed to the next lower level, inspecting the connecting staircase along the way. Check all the rooms on this level in the same manner. After all the finished rooms have been checked, inspect any unfinished areas such as the basement and crawl space. This is the end of the interior inspection. At this point, all the rooms, halls, and staircases throughout the house have been checked. To complete the home inspection, you must now check the electromechanical systems.

Electromechanical systems

The systems and associated equipment included in this category are electrical, plumbing, domestic water heater, heating, air-conditioning, and swimming pool. The condition, operation, and adequacy of each system in your house must be checked as described in its respective chapter and recorded on the worksheet.

This final check concludes your home inspection. You have now looked at every item in the house of any consequence and should have recorded on the worksheets all problems and deficiencies. Some of the problems you uncover might require the services of a professional for further investigation. All situations requiring the services of a professional are indicated in the chapters that follow.

Also, after the physical inspection, you might want to have certain tests performed to determine whether the house has environmental problems, such as a high radon concentration, contaminated well water, a leaking buried fuel-oil tank, lead paint, or a mold buildup.

Look at your worksheet and test results and try to evaluate the major problems. Do not expect a perfect house. You will always find minor problems, and the costs for correcting these problems should not concern you. However, if you find many minor problems, the costs for correction can be significant.

Of main concern from a cost point of view are major problems, some of which are defined as follows:

- The need for structural rehabilitation to the foundation
- The need for re-siding the exterior walls
- Water penetration into the basement or lower level
- A malfunctioning or obsolete heating system
- The need for repiping the plumbing system
- The need for a new roof
- A malfunctioning air-conditioning system
- Inadequate electrical service
- Termite infestation
- The need for complete rehabilitation to:
  —paved areas

  - —deck
  - —detached garage
  - —retaining walls
- Environmental problems

If you find a major problem, have a contractor look at it and give you a written estimate on the cost for correction. At this point, you should be able to determine the true cost of buying the house—the purchase price plus the costs for upgrading substandard, deteriorated, or malfunctioning components.

Final inspection

On the day of, but prior to, the contract closing, you should take one final walk through the house and look at the walls, windows, doors, and plumbing fixtures for cracked and broken sections. Sometimes damage occurs when the seller's furniture is being moved out or through vandalism when the house is left vacant for a period of time. Specifically, look for physical changes that occurred between the time of the contract signing and the closing.

During your walk-through, check the operation of the electrical, plumbing, and heating systems. The central air-conditioning system and swimming pool equipment should also be checked if the weather permits. Check the operation of all the appliances that are considered part of the purchase. If any appliances or electromechanical systems are malfunctioning, list them on a sheet of paper, along with any items that have been badly damaged since the contract signing. This list should be taken to the closing and discussed with the seller. Very often, dollar adjustments are made to compensate for the cost of repairs.

2

Roofs

Pitched roofs 6
  Shingles 8
    Asphalt shingles 10
    Wood shingles and shakes 12
    Asbestos-cement shingles 13
    Slate shingles 13
    Clay tiles 14
    Metal shingles 15
Flat roofs 16
  Built-up roofs (BURs) 17
  Single-ply membrane 19
  Roll roofing 19
  Metal roofing 19
Checkpoint summary 20

Every roof has two basic elements—the *deck* and the weather-resistant *covering*. The deck (also called *roof sheathing*) serves as a base for supporting the roof covering that protects the structure from the weather. A proper roof inspection includes an evaluation of both the roof covering and the deck. Even though a covering might be in good condition, the deck underneath might be soft, spongy, and structurally unsound. This condition can be caused by rot or delamination and is not necessarily noticeable in an exterior inspection.

Pitched roofs

The technique used for inspecting the roof differs depending on whether the roof is pitched or flat. Pitched roofs are checked during your initial pass around the house. Because of the hazards involved, I do not recommend that you climb onto a pitched roof. Begin your inspection by stepping far enough away from the house so that you are able to see all exposed sections of the roof as you circle the structure. The use of binoculars is recommended for this inspection to get a close-up view of the roof.

As you walk around the structure looking at the roof, make note of any uneven, sagging, or damaged sections. Unevenness in the roof might be the result of warped sections of deck or a poor installation of a second layer of shingles. This condition is usually not a problem. However, shingles in uneven areas are more vulnerable to damage and water intrusion. Make a note on your worksheet of the approximate locations of the uneven areas. During the attic inspection, you should check these areas for signs of leakage and to verify the cause of the unevenness.

Sagging sections in the roof, on the other hand, might be symptomatic of a structural problem or might reflect a problem that has been corrected. A sagging ridge beam or roof deck could indicate a structural failure, inadequate bracing, or inadequate spacing of wood-frame support members. The condition causing the sagging might have stabilized, so that no further corrective action is necessary. If you see a sagging ridge beam or sagging section of deck during your first pass around the house, have this condition evaluated by a professional.

Damaged sections can occur from falling tree limbs or swaying tree branches that overhang the roof. If you see a damaged section, record its location on your worksheet, since it must be repaired. Usually, patching the damaged area is all that is required.

Ventilation of the area directly below the roof deck is very important, especially in newer buildings where the deck is constructed of plywood panels rather than tongue-and-groove boards. If the area is inadequately ventilated, a moisture buildup can eventually cause the plywood sheathing to delaminate. This moisture problem is particularly acute in homes that have cathedral ceilings constructed in the following manner: The ceiling is plasterboard or an equivalent type of panel nailed directly to the roof rafters. Above the ceiling is insulation, and above this is the roof deck. Often there is a small air space between the insulation and the deck. When the moisture normally generated in the house by cooking, bathing, and so on reaches the area of the deck, there must be vent openings through which it can escape. Otherwise, rot and delamination can occur. A high percentage of the homes built with this type of cathedral ceiling have inadequately ventilated roof structures. Vent openings are needed near the top of each channel formed by the roof rafters and the ceiling and are also needed around the soffit. Often, only the soffit vents are installed. Vent openings for the top portion of the rafter channel can be provided through individual roof vents or a ridge vent. (See FIG. 2-1.) When the cathedral ceiling is the exposed roof-deck planks or panels, there is usually a rigid insulation on the top side between the deck and covering. This type of construction will not result in a deck having a problem with moisture accumulation, and therefore venting is not necessary.

If your house has a cathedral ceiling with no vent openings near the ridge, anticipate problems with the roof deck. You can tell if you have a deck problem by walking on the roof. If sections of the deck yield with each step and feel soft and spongy beneath your

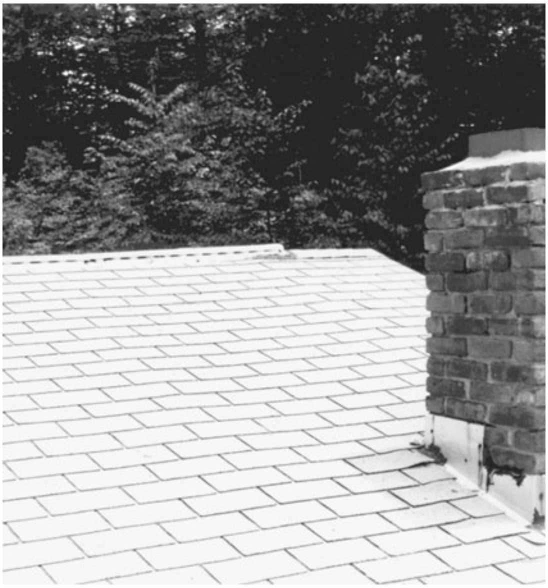

Fig. 2-1. *Ridge vent. Note the ridge vent along the top of the roof. This low-profile ventilator helps circulate air through the area below the roof deck.*

feet, there are problems. Note that you should not attempt to walk on the roof if the pitch is steep or the shingles are a type that can easily be damaged, such as tile or slate. Also, if the roof is not readily accessible from a deck or an intermediate-level area, it is best to leave this part of the inspection to a professional. Even if the roof deck shows no signs of a problem, if the area is not adequately ventilated near the top of the roof, the installation of a ridge vent should be considered to prevent future problems.

Shingles

Pitched roofs are usually covered with shingles applied in an overlapping fashion. The shingles are not intended to be watertight; they protect the structure from rain intrusion by shedding water. The more common types of shingles are made of asphaltic material, wood, asbestos-cement, slate, and clay tiles. When inspecting the roof, pay particular attention to a slope that has a southerly or southwesterly exposure. These slopes receive a maximum sun exposure, and it is the sun's rays that cause the shingles to become brittle and age prematurely. Consequently, the shingles on these exposures will deteriorate more rapidly than the shingles on the other exposures. (See FIG. 2-2.)

Since shingles are intended only to shed water, any water that gets under them will leak into the interior of the structure. Shingles that are lifting, cracked, or broken are vulnerable to this type of water leakage. If you see this problem, it is an indication that some maintenance is needed. In areas where the winter temperature drops below freezing, roof leakage can occur as a result of an ice dam. Because of heat loss through the roof and heat from the sun, snow on a roof can start melting, even in freezing weather. As the water reaches the roof overhang, it often refreezes, forming an ice dam and blocking the melting snow from draining. As the snow continues to melt, the water backs up under the shingles and leaks into the interior. (See FIG. 2-3.)

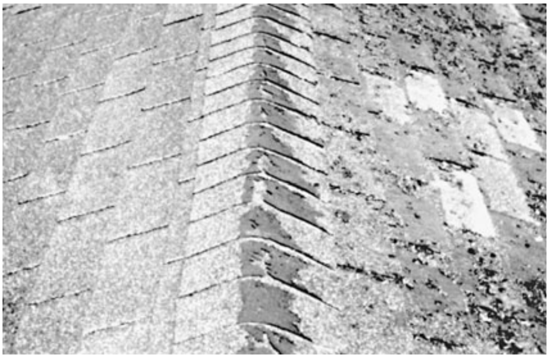

Fig. 2-2. *The orientation of the house can affect the projected life of the roof shingles. The deteriorated shingles on the right slope have a southerly exposure, while the shingles on the left slope have a northerly exposure.*

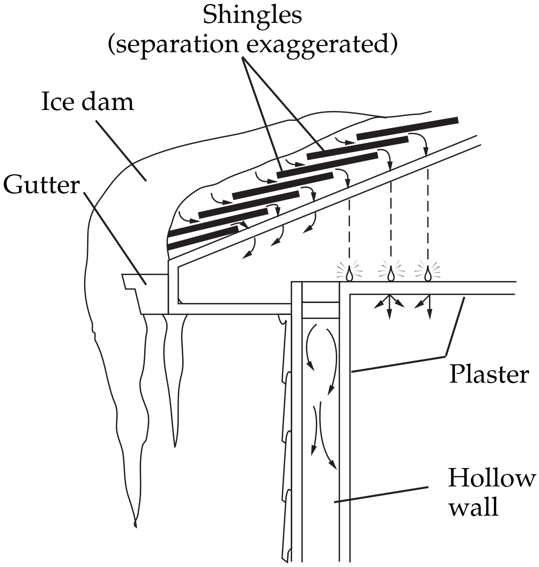

Fig. 2-3. *Ice dam at roof eaves. Because of the ice dam, water from melting snow backs up under the shingles and leaks into the house.*

Water leakage from this type of problem is not an indication of a faulty roof and should not be interpreted as a sign that roof repairs are necessary. It is an indication that adequate precautionary measures were not taken during the installation of the covering to eliminate or minimize the effects of an ice dam. The condition could have been reduced by the installation of eaves flashing. On existing roofs, the condition can be avoided by installing deicing cables along the edge of the roof and in the gutters and downspouts. This may not solve the problem completely, but it is somewhat effective. The deicing cables reduce the ice dam buildup by creating heated channels that allow water to drain into gutters and downspouts. The best method for minimizing an ice dam problem is to maintain what's called a *cold roof*. By overinsulating the attic floor and ventilating the attic profusely, the roof deck temperature will be lowered to the point where the snow won't melt.

When looking at a roof after all the snow has melted, you would never know whether there had been an ice dam and water leakage. Sometimes, however, you can see indications of a past problem—stained or warped sections of soffit trim or water stains on the ceilings of the rooms below, near the exterior walls. I have seen water stains on the ceilings of rooms two levels below the roof that were the result of water leakage because of an ice dam. Ice-dam problems will not necessarily occur every year; they depend on the severity of winter weather conditions.

Portions of the roof particularly vulnerable to leakage are the joints between the roof and roof-mounted structures, such as the chimney; the joint between the roof and a vertical sidewall; and the joint where two sloping sections of the roof intersect. The latter joint is commonly referred to as a *valley*. To protect the joints from water intrusion, they are normally covered with strips of a thin, impervious material called *flashing*. Sheet metal is usually used as a flashing material, with copper flashing as the top of the line; however, roll roofing strips are also used. Valley or sidewall flashing might not be visible, which depends on the type of joint construction. There are three basic types of valley construction: open, closed-cut, and woven valley. (See FIG. 2-4.)

When inspecting the roof, check the condition of the exposed flashing at the various joints. Loose, cracked, and deteriorated sections must be repaired. If there is leakage through any of these joints, it will usually be noted by water stains on the wood framing or roof sheathing in the

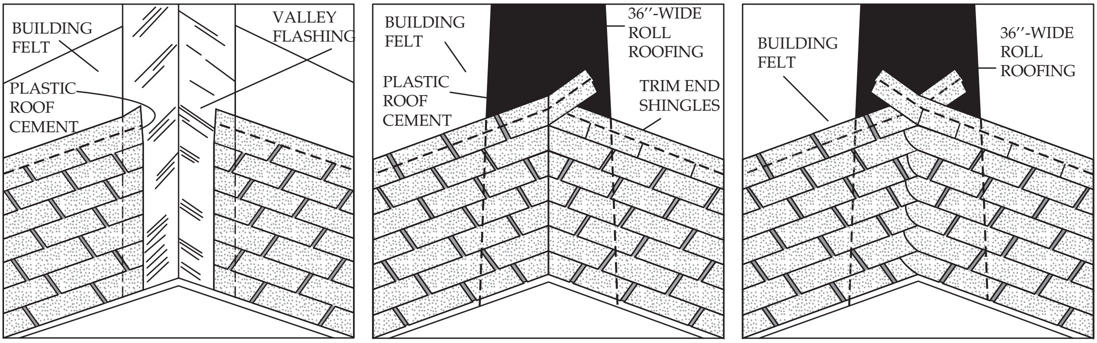

OPEN VALLEY CLOSED-CUT VALLEY WOVEN VALLEY

Fig. 2-4. *The three basic types of roof valley construction.*

attic or by stains in the ceilings of the interior rooms. Faulty joints are often resealed with an asphaltic cement rather than reflashed. The cement, however, is not as durable a seal as sheet-metal flashing, and the joint will often require periodic resealing.

Also check to see if the roof needs cleaning. Most pitched roofs need an occasional cleaning to remove an accumulation of debris—seed pods, twigs, pine needles, and leaves. Accumulated debris must be removed, especially from the spaces between the shingle tabs. If the debris is allowed to remain on the roof, it will retain moisture and promote the growth of moss and fungi, which is detrimental to the roofing. In addition, the litter can impede the runoff of rainwater, resulting in leaks.

Some asphalt shingle roofs develop a discoloration or what appear to be dirty streaks. The condition is often caused by wind-borne microscopic algae or mildew spores, which do not degrade or affect the performance of the shingles. It does, however, detract from the overall aesthetic appearance of the roof. The condition can usually be controlled by installing copper or zinc strips across the length of the roof and every few feet down the roof's slope.

Asphalt shingles The most common type of roof shingle used in this country is asphalt shingle, made by impregnating mats of fiberglass or organic felt materials such as rags, paper, and wood pulp, with asphalt and covering one surface with mineral granules. The *mat* is the vehicle for supporting the asphalt, which is a water-resistant material. The granules protect the shingle from damaging sun rays and provide color. When inspecting asphalt shingles, look for loss of granules, missing and torn sections with erosion of the mat. (See FIG. 2-5.) Some fiberglass-mat shingles have failed prematurely because of cracking, which can take the form of horizontal, vertical, and diagonal cracks across the shingles.

A particularly vulnerable location for leakage is the area between the shingle tabs. The granules in this area tend to come loose before those in other sections, exposing the mat to the weather. Although most roofs have a double and triple layer of shingles, a small section of the area between the shingle tabs has only one-shingle coverage. Thus, an eroded mat in this area is very vulnerable to water leakage. Loss of granules and erosion of the mat between shingle tabs is a deficiency that usually occurs on the roof slope with a

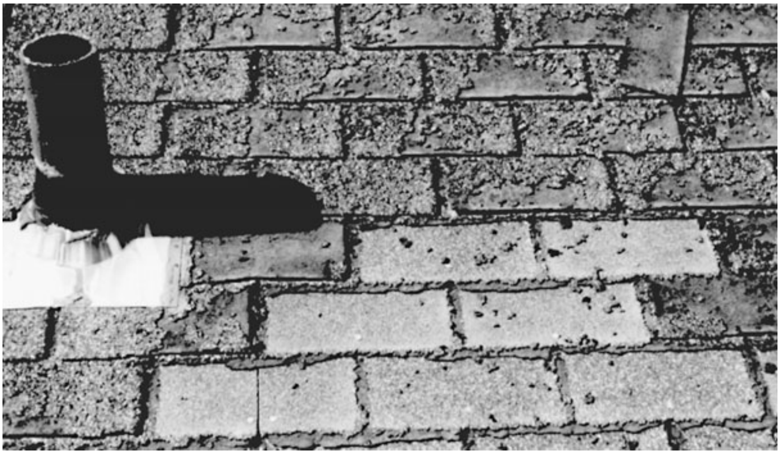

Fig. 2-5. *Deteriorating asphalt roof shingles. Note torn, missing, and brittle shingles with a loss of the granule covering, exposing the roofing mat.*

southerly or southwesterly exposure before other slopes. The condition is usually visible from the ground and can be clearly seen with the aid of your binoculars. When you see such a problem, you should anticipate early replacement of the roof shingles.

Most homes are designed to take three separate layers of shingles, although in some communities only two are allowed. When a new covering is needed on a structure that already has the maximum layers allowed, it is necessary to remove all the layers before installing the new shingles. When reroofing, it is more costly to remove existing layers of shingles than to install a new layer over existing shingles. Therefore, you should try to determine the number of layers. When the roof has an exposed edge, as in the case of a gable roof, look at the thickness of the layers. If you see two to three overlapping shingles, the roof covering is the first layer. In a hip type of roof, since there are no exposed edges, this type of determination cannot be made. In this case, try to find out the age of the house. Asphalt shingles have a projected life of seventeen to twenty-two years. The actual life span of the shingles will depend on the weight of the shingles, the type of mat, and the exposure. Asphalt shingles are classified by weight (pounds per roofing square); a roofing *square* is 100 square feet. Lightweight shingles, the least costly, weigh about 215 pounds per roofing square. Heavyweight shingles weigh about 350 pounds per square and have a longer life expectancy than lightweight ones.

If the covering is over seventeen years old, extended life for the shingles should not be anticipated. Even though the shingles might look all right (lying flat with no noticeable loss of granules or erosion), they are becoming brittle and vulnerable to wind damage. Also, these shingles will be more vulnerable to damage from someone walking on the roof when cleaning the gutters, installing a TV dish antenna, and so on. Often, these shingles will show signs of aging such as curling, cupping, cracking, and pitting. (See FIG. 2-6.) Such shingles are vulnerable to damage and will deteriorate rapidly. An exact estimate of the usable years or months remaining for the shingles is difficult. Some people do not replace an aging roof until there is a leakage problem. Others will replace it before any leakage occurs thus avoiding the cosmetic damage caused by leakage. The life span of an aging roof can be extended by patching and coating exposed cracks and eroded areas. However, even if you see no signs of leakage, the shingles on a roof that can take a second layer should be replaced before they become brittle, curl excessively, crack, and chip, as shown in FIG. 2-7. Because of the physical condition of these shingles, the surface of a second layer of shingles will be uneven, lumpy and aesthetically unattractive. In this case, for a nice even appearance the old shingles need to be stripped off the roof deck before the new shingles are installed.

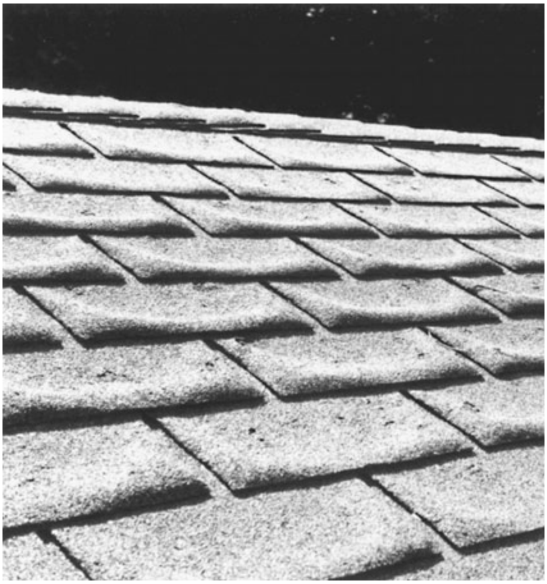

Fig. 2-6. *Aging asphalt roof shingles. Note curling of the edges, with some pitting.*

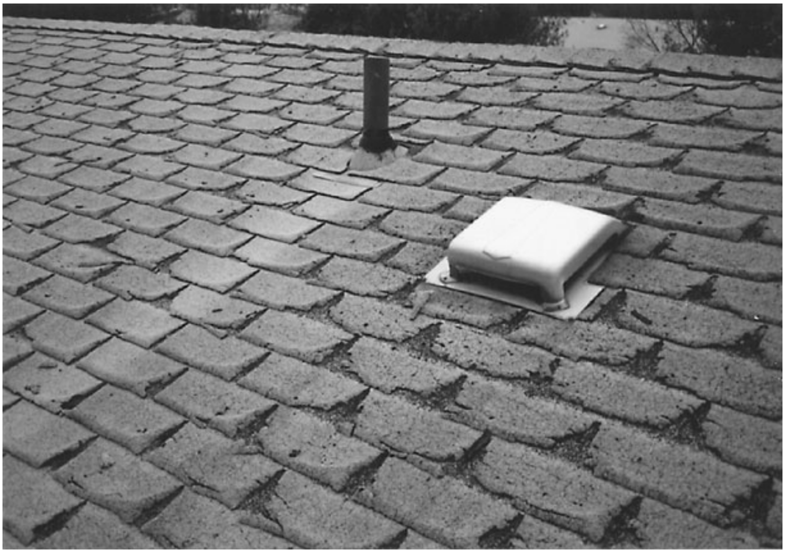

Fig. 2-7. *These old, weathered asphalt roof shingles are dry, brittle, cracked, chipped, and excessively curled.*

If more than approximately one-third of the roof shingles show signs of advanced aging, I recommend reroofing. At this point, attempts to extend the life of the shingles are usually not economically justifiable.

Wood shingles and shakes In many parts of the country, wood shingles and wood shakes are used as a roof covering. The basic differences between the two are appearance and thickness. During the manufacturing process, shingles are sawed; shakes are split. Consequently, wood shingles have a relatively smooth surface, and shakes have a textured surface.

Because of the need for resistance to decay, most wood shingles and shakes (hereafter referred to as shingles) are made from cedar. They are also made from redwood and southern cypress. The shingles, although resistant to decay, are not immune to decay and will rot after prolonged exposure to moisture. (Rot-producing fungi are discussed in chapter 8.) The projected life expectancy for a wood-shingle roof is twenty-five to thirty years. As a wood-shingle roof ages, the shingles dry, crack, curl, and rot. As you walk around the house looking at the roof, be aware of aging shingles. Rotting shingles should be replaced. If you notice loose, damaged, or missing sections, repair is needed, even if you see no signs of water leakage. When approximately one-third of the shingles on a slope show signs of excessive aging (rotting, chipped, cracked, loose, missing, or curling), all the shingles on that slope should be replaced.

On the northern slope or on portions of the roof that are usually shaded, you might see moss growing in clusters between the joints of the shingles. It should be removed. The moss functions like a wick; the root system provides a direct path for water entry. In addition, as the moss cluster builds up, it might lift the shingles slightly, making them more vulnerable to water penetration, particularly during a driving rain.

Wood shingles have traditionally been installed on spaced sheathing, as opposed to solid sheathing, which acts as nailing strips.

This enables air to circulate on the underside of the shingles so that they can dry from both sides. The shingles are spaced between ⅛ and ¼ inch apart to allow for swelling during damp weather. Because of this space and the irregularities of some of the shingles (due to thickness and texture), daylight might be visible through portions of the roof from the attic. If during your inspection of the attic you see daylight through a wood-shingle roof, don't think that roof maintenance is necessary. If daylight is visible through the roof by means of an indirect path, maintenance is not required. On the other hand, if daylight is visible via a direct path, such as a crack, some maintenance is needed. Depending on the pitch of the roof, the shingles are two-, three-, or four-ply and are installed so that the joints between the shingles for the various plies do not line up. When daylight is visible via a direct path, cracks in the shingles line up with the joints. In this case, water can penetrate the roof, and maintenance is needed. On newer construction, you may find the roof deck consisting of solid sheathing. This is not a problem. In such a case, in order to provide the needed air circulation on the underside of the shingles, a product called Cedar Breather is probably installed between the shingles and the roof deck. Cedar Breather is a three-dimensional nylon matrix that is stiff enough to resist crushing, thereby allowing air movement to the underside of the shingles.

Asbestos-cement shingles Asbestos-cement shingles were manufactured by combining asbestos fibers with Portland cement under high pressure. Although the shingles are no longer manufactured, they can be found on many homes. Because of the asbestos ban in the mid-1970s, similar-type shingles are now manufactured using nonasbestos man-made fibers and cement. The shingles possess properties that make them highly suitable for exterior use. They are immune to rot, unaffected by exposure to salt air, and fireproof. One drawback is that they are weak in their resistance to impact and thus are vulnerable to cracking and chipping.

As you walk around the exterior of the building during the roof inspection, look for cracked, loose, chipped, and missing shingles. Note on your worksheet the areas that will require maintenance. Although asbestos-cement shingles individually last many years, an asbestos-cement-shingle roof should not be considered maintenance-free, and periodic repairs should be anticipated. Occasionally, as with a wood-shingle roof, clusters of moss might be found on the northern slope or slopes shaded by trees. If you see this condition, note it on your worksheet. The moss is a potential problem and should be removed.

Slate shingles Of all roof coverings, slate shingles are the most durable. If they are of good quality, they can last indefinitely (at least in excess of one hundred years). The slate roof over the Saxon Chapel at Stratford-on-Avon in England is over eleven hundred years old and, according to the Vermont Structural Slate Company, is in good condition. A slate roof, however, does not remain maintenance-free, even though the slates are of good quality. I have seen very few slate roofs that did not have some cracked, loose, chipped, or missing slates, a condition that requires some repair. These repairs are considered minor roof maintenance and should be anticipated on a periodic basis. Slate roofs are often patched with asphalt cement, which has a tendency to dry and crack and requires periodic application. When inspecting this type of roof from the exterior, look for cracked, loose, chipped, or missing shingles. If you find any, make a note on your worksheet.

Repairs to a slate roof, even minor ones, can be somewhat costly. Several roofers have told me that their fee reflects additional work above and beyond the required repair because they always anticipate accidental cracking

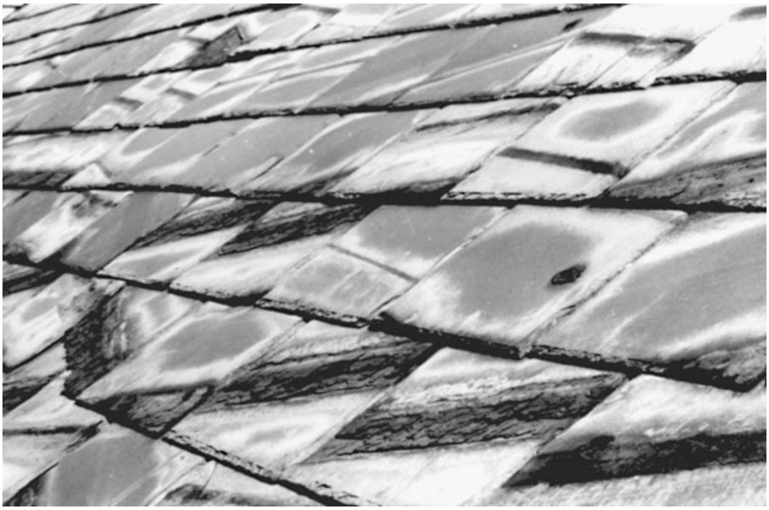

Fig. 2-8. *Slate roof shingles. The ribbon slate is of inferior quality. Cracking often occurs along the ribbon after only 10 years.*

of some of the slates during the repair. One difficulty you should be aware of is that when replacing a slate shingle, the roofer might not be able to match the color of the new slate to the existing weathered shingles.

Sometimes a poor-quality slate, *ribbon slate*, is used as a roof covering. (See FIG. 2-8.) The ribbons within the individual shingles are softer than the normal slate and will cause the shingles to crack along the ribbon. Often, cracking occurs along the ribbon after only ten years. Repairs to these shingles must then be made as needed.

When inspecting the roof, flaking slates may also be noted. Surface flaking is of no concern, since the shingles are at least 3⁄16 inch thick and are basically impervious to water. However, if any of the shingles are deteriorating as a result of excessive splitting and flaking (a condition brought about by winter freeze-thaw cycles), they should be replaced.

There are a number of synthetic slate look-alikes that are now on the market. These products have the durability, texture, and appearance of natural slate and are lighter in weight. The materials used for synthetic slates are quite varied. One product called FlexShake is made from recycled steel-reinforced rubber automobile tire treads. Others are made from recycled postindustrial rubber and plastic waste. There is also a synthetic slate made from ground natural slate, resin, and fiberglass, bonded under high pressure. Although synthetic slates are too new to comment on their life span, manufacturers estimate it at 40 to 60 years or more.

If the house you are inspecting is in the northern part of the country where snow might accumulate, look for snow guards on the lower portion of the roof. (See FIG. 2-9.) In particular, they should be located above doorways, sidewalks, or other areas where people will pass or gather. Snow guards are needed to prevent sliding masses of snow and ice from falling off the roof and damaging the gutters. It might interest you to know that the slate roofs on the buildings of the Harriman estate in New York had 35,000 copper-wire snow guards.

Clay tiles Clay tiles are available in many patterns. The most common are *Spanish*

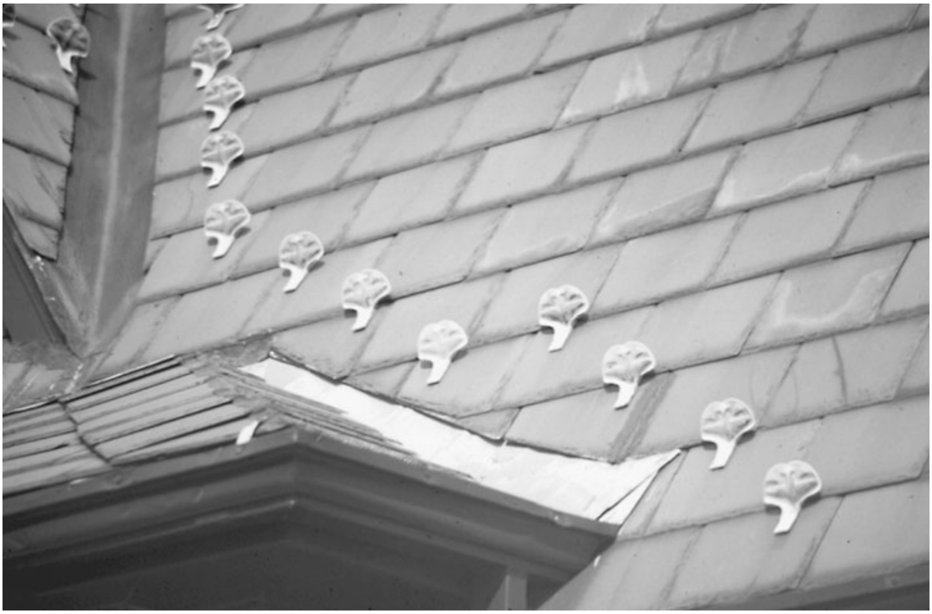

Fig. 2-9. *Snow guards along the edge of the roof will help keep the snow from sliding off.*

and *Mission*. These tiles are made by shaping moist clay in molds and firing the various shapes. They are hard, durable, and fireproof. However, they are also brittle and can be easily damaged by falling tree limbs or climbing on the roof to make repairs. As with slate shingles, repair or replacement of individual tiles is more difficult and costly than that of asphalt shingles. Also, matching new tiles to the weathered tiles is usually a problem. When inspecting this type of roof, look for loose, broken, chipped, cracked, or missing tiles. If any of these conditions is found, it should be noted on the worksheet, as repairs are needed. Tiles can also deteriorate as a result of freeze-thaw cycles. You might find some cracked areas that have been sealed with asphalt cement. This condition is usually an indication of past problems. Since asphalt cement does dry and crack, periodic reapplication should be anticipated.

Check the joint (valley) between two sloping sections of the roof. (See FIG. 2-10.) If it's filled with asphalt cement, it's an indication of a problem condition that has been temporarily corrected. The flashing in that joint should be replaced. As with slate roofs, repairs, even minor ones, can be somewhat costly. This item should be noted on your worksheet.

Metal shingles Metal shingles are becoming more popular for residential roofing. They are available in several different shapes and are primarily made from painted or coated aluminum or steel panels although they are also made from copper. The panels are approximately 4 feet wide by 1 foot high and are formed to resemble wood shakes, slate, or Spanish, Roman, and Mediterranean tiles. The shingles are durable, lightweight, and fire resistant. Some have embedded stone chips for additional texture.

When inspecting the roof look for loose nails or loose shingle panels, both of which can result in leakage. Check to see if any of the joints between panels and valley sections are covered with roofing cement, which indicates

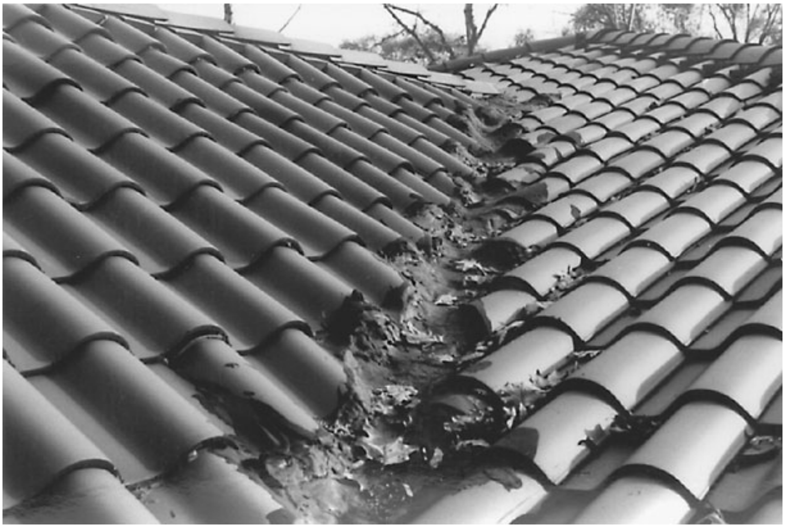

Fig. 2-10. *Valley joint filled with a heavy layer of asphalt cement is an indication of a problem condition.*

past and potential problems. From a cosmetic point of view, is the finish on the shingles fading or chalking? Record your findings on the worksheets.

In addition to shingles, flat sheet metal roofing has also been gaining in popularity for residential pitched roofs. Metal roof panels that come in widths of 12, 16, 18, and 36 inches are installed so that they run down the slope of the roof, essentially from ridge to eave. The joints between the panels are overlapped and interlocked to provide weathertightness. Do not climb up to the roof for an inspection. Metal roofs are slippery when wet, and depending on how the panels are supported they could be dented. The roof should be inspected from the ground with binoculars. If there are any doubts about its condition, have it inspected by a professional roofer.

Flat roofs

A roof that is perfectly level or slightly pitched is referred to as a *flat roof*. Since this type of roof is not visible from the ground, the inspection must be made from the roof itself. As with a pitched roof, a flat roof should be the first item inspected. Safe access to the roof is of prime importance. If the building is higher than one story, the roof should be accessible from the interior. Anything other than an interior means of access is a potential hazard and is considered a deficiency in the structure's design. If the roof you are inspecting is flat and is more than one story high with no interior access, it is best to have it inspected by a professional roofer.

Ventilation of the area directly below the roof deck is needed to minimize the moisture buildup in this area. An excessive moisture condition can result in deterioration of the roof deck, a lowering of the thermal resistance of the insulation, and, eventually, damage to the interior of the structure. During the cooler months, the moisture trapped in the area between the roof and the upper-level ceiling will condense, drip onto the insulation, and cause random water stains on the ceiling. Adequate ventilation of this area is also important in reducing the summer heat load on the rooms located immediately below the roof.

All too often, provisions for ventilation have been omitted by the builder. Therefore, when inspecting a flat roof, be sure to look for ventilation openings. The openings might be in the form of roof vents (vertical pipes protruding through the roof deck) or open areas in the side of the building just below the roof. Roof vents are often shielded from the rain by a cover and should not be confused with the plumbing vent stacks, which also protrude through the roof deck. The plumbing vent stack is connected to the house sewer line and is easily identified by the odor of the discharging gases. If no ventilation openings are noted, that fact should be marked on your work-sheet, and installation of ventilation openings should be considered.

A flat roof must have a watertight covering, rather than one that merely sheds water, to protect the area below the roof from water intrusion. The most common types of flat-roof coverings are built-up, single-ply, roll roofing, and metal. When inspecting a flat roof, look for ponded water. Unless the roof was specifically designed to hold standing water as an energy conservation measure, to reduce the heat load during the warm or hot months, its presence is considered a potential problem. Ponded water can become a breeding place for insects and can promote the growth of vegetation and fungi. The roots of plants growing on the roof can puncture an asphalt covering. The freezing of ponded water that has penetrated into the layers of a built-up roof can delaminate the roof covering. The temperature difference of the wet and dry areas on a randomly ponded roof results in differential expansion that might cause warping and cracking of the roof cover. If you see ponded water or signs of past ponding on the roof during your inspection, note the location on your worksheet as an area that should be drained.

A properly designed roof should have provisions for drainage. Two basic drainage designs are used in a flat roof—the perimeter system and the interior drainage system. In the perimeter system, water that drains from the interior portions of the roof collects in gutters or scuppers (openings in a parapet wall) located along the perimeter and then flows into downspouts or merely drips off the roof. In the interior drainage system, drains are located in the roof itself and are connected to downspouts that run through the interior of the structure. Look for one of these drainage systems as you walk around the roof. The interior roof drains are often clogged with debris and are sometimes set higher than the surrounding area, a condition that results in ponding. Look for cracks around the joints between the roof drain and the roof covering.

When inspecting a flat roof, you should also inspect all roof-mounted structures and projections such as skylights, hatch covers, chimneys, vent stacks, and so on. These items are discussed in chapter 3.

Built-up roofs (BURs)

This type of roof consists of bitumen (asphalt or coal-tar pitch) sandwiched between two to five layers of roofing felts and is usually covered with a mineral aggregate embedded in the top surface. The bitumen is the waterproofing agent, stabilized and reinforced by the roofing felts. The felts restrain the bitumen from flowing in hot weather and help resist cracking in cold weather. The aggregate surfacing generally consists of gravel, slag, or crushed rock. Its purpose is to protect the bitumen from the damaging effects of the sun's infrared and ultraviolet rays. These rays, through a combination of heat and photochemical oxidation, accelerate the aging of bitumen, resulting in premature brittleness and cracking. Depending on the number of layers in the builtup roof and the quality of construction, the projected life can vary from ten to twenty years.

An aggregate-covered built-up roof is difficult to inspect, principally because of the mineral-aggregate topping. Nevertheless, look for telltale signs of past, current, and potential problems. Walk around the roof systematically so that you cover the entire roof area. Look for patched areas, cracking, blistering, surface erosion, alligatoring, and wrinkling. In particular, look for cracks at the joints of roof projections and roof-mounted structures.

Usually, this type of roof has a flanged metallic strip along the perimeter, a *gravel stop*. It provides a finished edge for the built-up roofing and also prevents loose aggregates from washing off the roof. Look for cracks in the joint between the gravel stop and the roof. All cracks in the roof cover should be sealed. Record the location of any cracked or patched areas on your worksheet so that you can check the ceiling of the rooms below these areas for water stains.

Not all builtup roofs have a top surface covered with aggregates. A smooth-surface roof treatment has a weathering surface that is an asphalt-saturated roofing felt. Since the top surface is exposed, defective or deteriorated sections are visible. To maintain the weathering surface in reasonably good condition, the surface should be covered with a protective coating such as an asphalt-based emulsion every three to six years. Excessive application of the emulsion will result in alligator-type surface cracking, which is a potential problem. (See FIG. 2-11.)

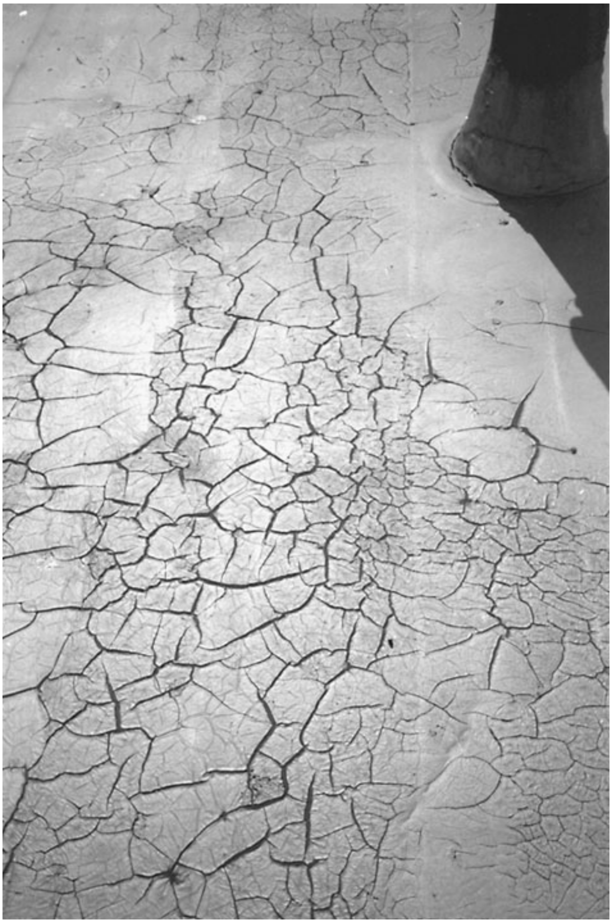

Fig. 2-11. *Roof surface with alligator-type cracks.*

Incidentally, when walking on the roof, be careful not to step on any *blisters*—weak spots in the roof covering, usually the result of air or water vapor being trapped between the layers of roofing felts. Depending on a blister's brittleness, the weight of someone walking on the blister might cause the roofing felt to crack.

Single-ply membrane

In the 1970s, the notion that a single ply of rubber, plastic, or modified bitumen material could provide the same weather protection as a built-up roof gained acceptance. Initially, single-ply membranes were used on commercial and apartment buildings; today, they are also used on residential structures with flat roof decks.

Several methods are used for installing a single-ply membrane. The membrane can be loose-laid over the roof deck and attached only at the perimeters. It is then held in place everywhere else by the weight of gravel or concrete pavers placed on top. Another method is to adhere the membrane to the roof deck fully with a self-adhesive or an adhesive applied at the jobsite. Single-plies can also be installed as a partially adhered system by using mechanical fasteners or applying an adhesive at predetermined intervals.

When inspecting a single-ply roof membrane, look specifically for open joints at seams that are not properly adhered; check for membrane erosion and punctures; and look for cracks, splits, and open sections around the perimeter walls and roof projections. As with a BUR, if the membrane is covered with gravel or paving blocks, a visual inspection of the weathering surface is not practical. In this case, pay particular attention to the condition of the ceilings in the rooms directly below the roof.

Roll roofing

This type of covering is often found on older structures in urban areas and over porch roofs. It is also found on pitched roofs of low-cost homes and outbuildings. It consists basically of an asphalt-saturated roofing felt applied directly over the roof deck and provides only single- or two-ply coverage. Quite often, the outer surface is coated with hot tar to seal joints and small cracks. Roll roofing should be inspected for blistered, cracked, eroded, weathered, torn, and aging sections. (See FIG. 2-12.) Be careful not to walk on blistered sections. Sometimes the lapped joints between the strips curl and lift, making the joints vulnerable to water seepage. Check the joints between the roof and the parapet wall and other roof projections. These joints are vulnerable to cracking and periodically require resealing. (See FIG. 2-13.) If any of the above items are noted, they should be recorded on your worksheet, since they indicate the need for some maintenance.

Metal roofing

In older homes, the roof that covers porches, garages, or an exterior side projection such as a sunroom is usually made of galvanized iron (sheet iron coated with zinc), *terne* metal (steel-coated with a mixture of lead and tin), or copper. A galvanized roof with a heavy zinc coating will last many years without painting. However, a lightly galvanized roof will rust and must be kept well painted. A terne metal roof, if kept well painted, can last for over forty years. A copper roof will also last for many years and does not require painting. Occasionally, a sheet-metal roof covering might be found on the main roof of a Victorian structure. When you are inspecting a sheet-metal roof, look for cracks or open joints at the soldered seams; because of the large amount of expansion that

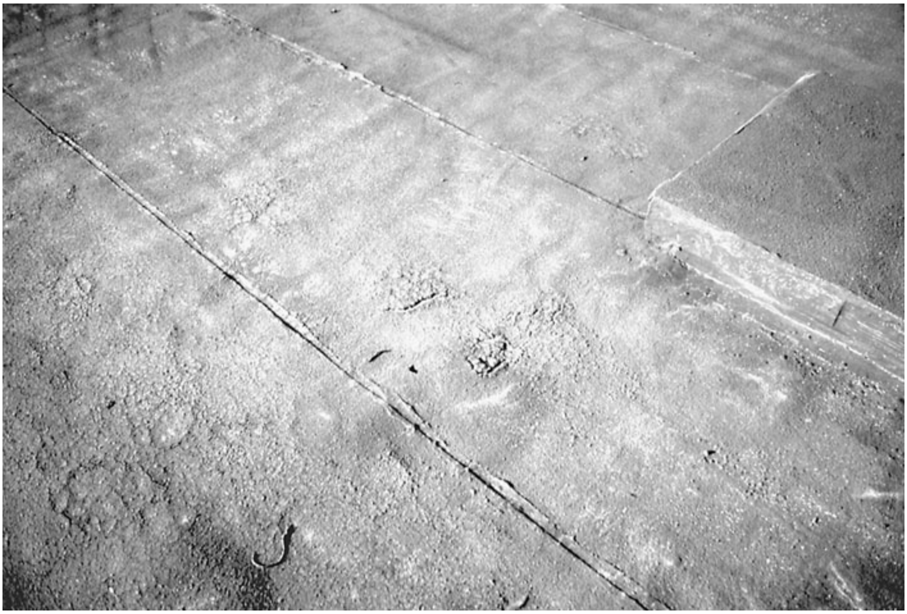

Fig. 2-12. *Weathered, cracked, and aging roof membrane.*

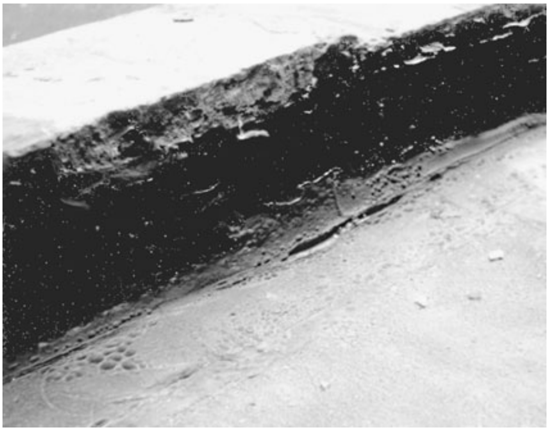

Fig. 2-13. *Cracked and open joint between the roof and the parapet wall.*

takes place, the seam is a vulnerable joint for leakage. Also, look for exposed corroding sections of metal; they must be scraped and coated with paint or a bituminous compound such as asphalt cement. In some cases, problem conditions that have occurred over the years have necessitated coating the entire roof with a bituminous compound as a means of sealing corroded areas. Be aware that this type of roof coating requires periodic maintenance.

Checkpoint summary

Pitched roofs

❍ Visually inspect all portions (slopes) of the roof.
❍ Note any sagging, uneven, damaged, or patched sections.
❍ Look for overhanging tree limbs or branches that can cause damage.
❍ Is the area directly below the roof deck ventilated?
❍ Is the roof in need of a cleaning?

Asphalt shingles

❍ Look for curling, cracked, torn, or missing shingles.

- ❍ Are shingles losing their stone granules?
- ❍ Look for eroded sections in the slot between the shingle tabs.
- ❍ Pay particular attention to slopes with a southerly or southwesterly exposure.
- ❍ How old are the roof shingles?
- ❍ How many layers of shingles?
- ❍ If roof has been recently reshingled, is a guarantee/warranty available?

Wood shingles, shakes

- ❍ Look for rotting, loose, cracked, chipped, or missing sections.
- ❍ Pay particular attention to the slopes with heavy shade.
- ❍ Any signs of moss?

Slate, asbestos-cement, clay tiles

- ❍ Look for missing, cracked, chipped, flaking, or loose sections.
- ❍ Sections patched with asphalt cement?
- ❍ Are there snow guards along the lower edge of the tile roof?
- ❍ Are the valley joints filled with asphalt cement?

Metal shingles

- ❍ Check for loose nails and loose shingle panels.
- ❍ Are any joints or valley sections covered with roofing cement?
- ❍ Are there any chalking, faded, or dented sections?

Flat roofs

- ❍ Safe access available?
- ❍ Cracked, blistered, eroded, split, punctured, or torn sections?
- ❍ Look for open joints and seams.
- ❍ Areas of ponding water or low points where water will accumulate?
- ❍ Drainage system functional?
- ❍ Area below the roof deck adequately ventilated?

Built-up roofs (BURs)

- ❍ Are there areas with missing aggregate?
- ❍ Look for blisters in the membrane.
- ❍ Look for patched areas, surface erosion, and alligatoring.
- ❍ During the interior inspection, check the ceilings of the upper-level rooms for water stains.

Roll roofing (tar paper)

- ❍ Sections drying, eroding, or blistered?
- ❍ Look for open seams.

3

Roof-mounted structures and projections

Chimneys 22
Vent stacks 26
Roof vents 27
Roof hatch 28
Skylights 28
TV antennas 28
Lightning protection 29
Gutters and downspouts 29
  Built-in gutters 30
  Exterior-mounted gutters 30
  Downspouts 32
Checkpoint summary 34

When inspecting a roof, you should also inspect all roof-mounted structures and projections. Specifically, look at chimneys, plumbing vent stacks, roof vents, roof hatch, skylights, TV antennas, and gutters and downspouts. Except for the last two items, an area of concern is the joint between the item and the roof. This joint, although normally sealed with flashing, is vulnerable to water leakage. With years of weathering, the flashing can develop cracks, pinholes, or breaks, resulting in periodic leaks. To correct this type of water leakage, rather than reflashing the joints, the homeowner might often seal the joints with asphalt cement. While this is an effective correction, the asphalt cement will eventually become brittle and crack, and the joints will require periodic resealing.

Chimneys

Chimneys are used to vent smoke and combustion gases from heating units and fireplaces. If, during your exterior inspection, you do not find a chimney, it does not necessarily mean that the house lacks a heating system. In all probability, the house is heated electrically.

Chimneys are normally constructed of masonry (brick, concrete, stone) or are prefabricated from metal or a cement-asbestos material. Masonry chimneys are usually supported by their own footings, which in northern communities extend below the frost line. These chimneys are not dependent on the main structure for support. When inspecting a masonry chimney that extends up the side of a building, if you see open joints between the chimney and the sidewall, it is an indication of some settlement and is often not a problem. (See FIG. 3-1.) However, the open areas should be sealed to prevent water intrusion. On the other hand, if the chimney is no longer vertical, it indicates excessive settlement because of the lack of a footing, or one that's improperly designed, and it should be checked by a professional. Depending on the degree of settlement, the flashing at the joint between the chimney and the roof might need repair. The movement of the chimney could result in open and loose sections of flashing. (See FIG. 3-2.)

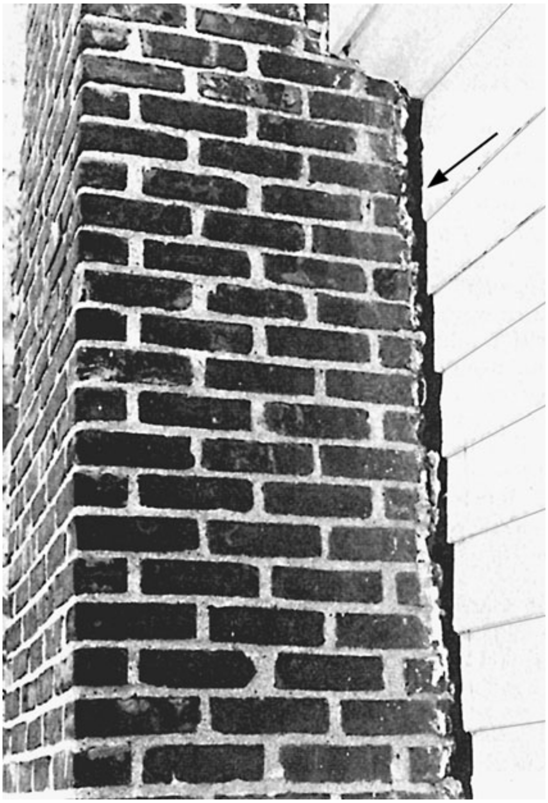

Fig. 3-1. *Uneven settlement of chimney. Note the open joint between the chimney and the sidewall.*

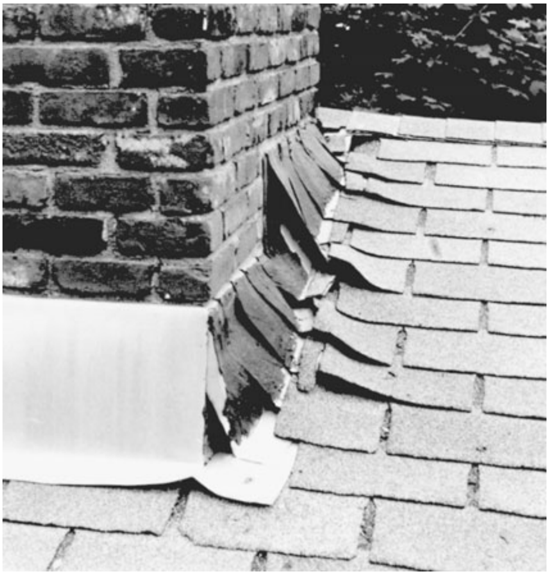

Fig. 3-2. *Open and loose sections of flashing caused by movement of the chimney.*

If your house has a brick chimney, look at the area above the roofline to see that it is vertical. Over the years, the mortar joints can weaken on one side and cause the chimney to lean. (See FIG. 3-3.) A leaning chimney represents a potential safety hazard and should be checked by a professional mason to determine whether or not corrective measures are needed, such as either bracing or rebuilding from the roofline up. If you ask the owner about the chimney, you'll probably learn that it has been in that condition for at least fifteen years. That fact does not mean that corrective action is not needed. A leaning chimney is an indication of weakening mortar joints and should be considered of questionable structural integrity.

Sometimes brick chimneys on relatively new houses might be coated with white mineral deposits called *efflorescence*, a condition often caused by the absorption of water by the bricks. The minerals in the bricks dissolve in the water and then surface when the water

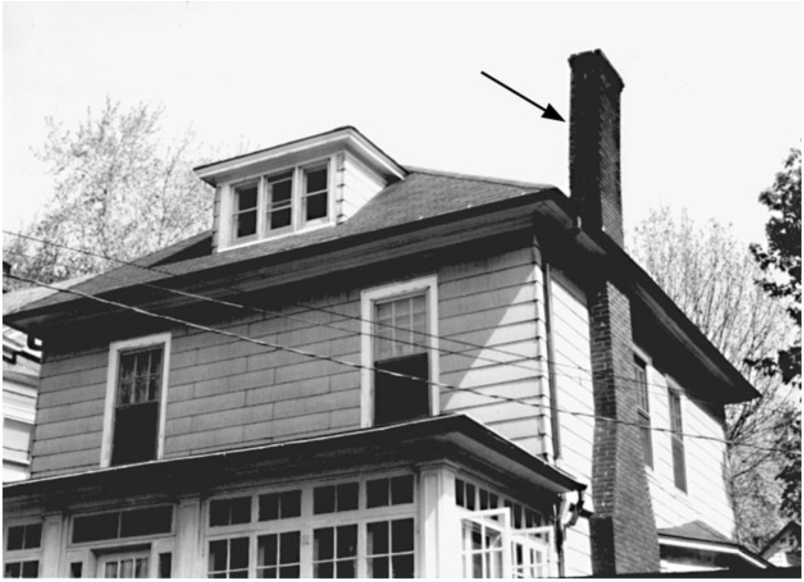

Fig. 3-3. *The portion of the chimney that extends above the roof is leaning, a potentially hazardous condition.*

evaporates. Although efflorescence is quite common in new brickwork, it can easily be scrubbed or washed off with a dilute solution of muriatic acid. Recurrence can usually be controlled by covering the bricks with a vapor permeable water repellant. Efflorescence on a brick chimney that has been up for many years usually means that water is getting inside the chimney, through cracks in the joints on top of the chimney or cracks in the bricks or mortar joints. If you see heavy efflorescence on an older chimney, make a note on your worksheet. The top of a masonry chimney should have a cement finish that slopes from the flue to the edge of the chimney. The purpose of this finish is to deflect rain and protect the joints between the flue and the chimney. This cement finish is vulnerable to cracking, and periodic resealing of this area should be anticipated.

The top of the chimney must extend above the roofline to prevent downdrafts caused by the turbulence of the wind as it sweeps past nearby obstructions or over sloping roofs. The top of the chimney should extend at least 3 feet past the highest point where it passes through the roof, and 2 feet higher than any roof part within 10 feet measured horizontally (FIG. 3-4). When a chimney is 10 horizontal feet beyond a roof ridge, builders often terminate the top of the chimney below the ridge. This is not quality construction, but it satisfies the code. In this case, downdrafts can still occur as a result of air currents formed when the wind hits the side of the building. The downdrafts can affect the efficiency of the heating system or result in backsmoking of the fireplace. This type of problem can usually be controlled by installing a concrete or stone cap about 8 inches above the top of the flue.

Older homes very often have unlined chimney flues. Although these chimneys might operate satisfactorily, they are a potential hazard. Over the years, the corrosive gases can have a deteriorating effect on the mortar joints. (See FIG. 3-5.) If an unlined chimney in your house is connected to a fireplace, and if you intend to use the fireplace, you should have a flue liner installed down the existing flue. This

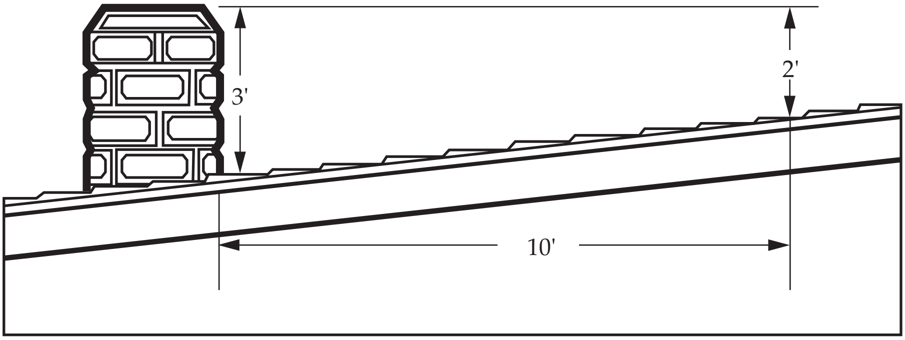

Fig. 3-4. *Chimneys should extend at least 3 feet past the highest point where they pass through the roof, and 2 feet higher than any roof part within 10 feet measured horizontally.*

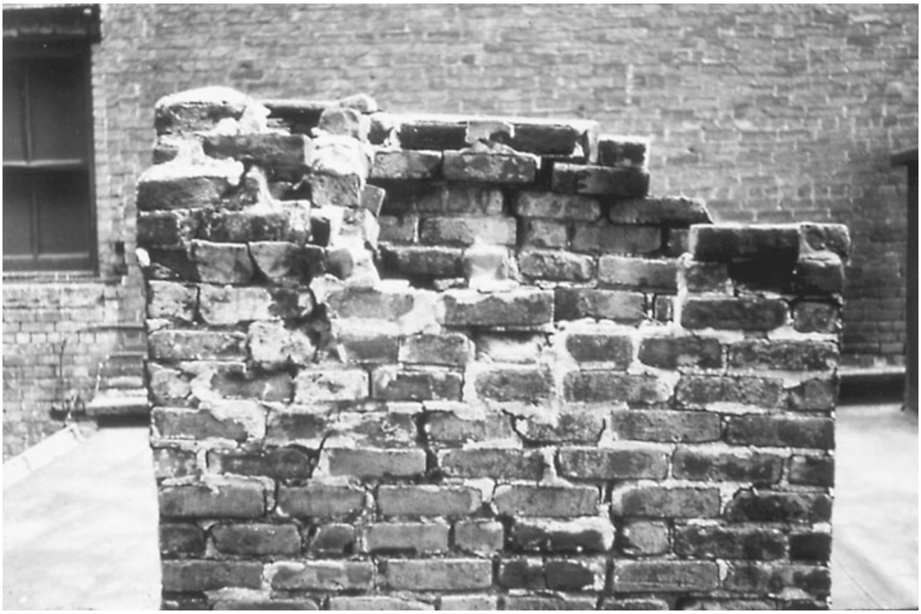

Fig. 3-5. *Deteriorated chimney with no flue liner. Chimney should be rebuilt and flue should be lined.*

is an important fire safety measure and should be recorded on your worksheet.

In many homes when the flue damper in the firebox has deteriorated, it is replaced with a chimney-top damper. (See FIG. 3-6.) This spring-loaded damper is mounted over the flue opening at the top of the chimney. The device has a stainless-steel wire that runs from the damper down the flue into the opening of the fireplace. If you see a chimney-top damper during your inspection, don't forget to check its operation when inspecting the fireplace.

All masonry chimneys should be inspected for cracked, loose, chipped, deteriorating, and missing sections of brick and mortar joints. Some masonry chimneys have a stucco finish. In this case, look for cracked, chipped,

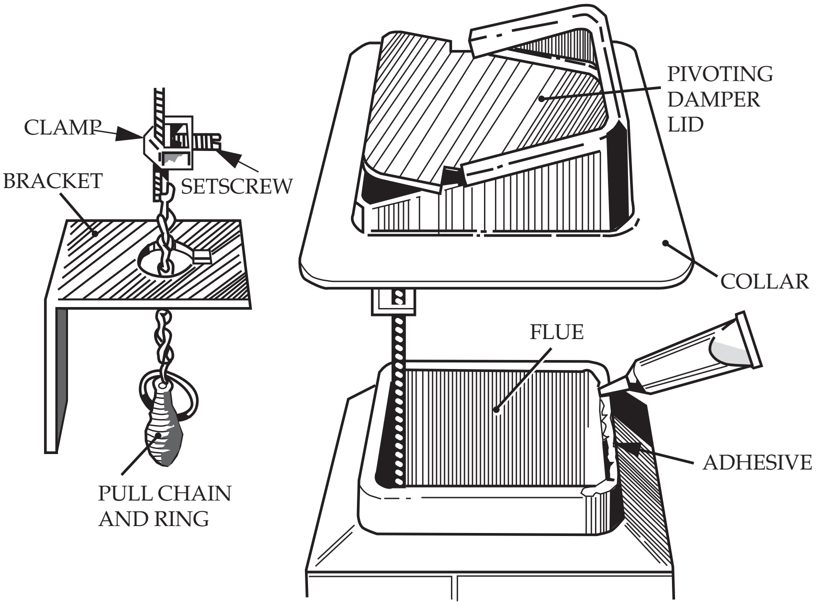

Fig. 3-6. *Spring-loaded chimney-top damper covers flue opening at top of chimney.*

and loose sections of stucco. Brick chimneys on older homes are sometimes covered with an asphalt-type coating, especially above the roofline. This technique is often used to prolong the life of the chimney when there are many cracked mortar joints and deteriorating bricks. It is considered makeshift. If you see any of the above items, they should be recorded on your worksheet.

Is there a *cricket* (also known as a *saddle*) behind the chimney? (See FIG. 3-7.) There should be one if the chimney is located along the slope of a roof and is more than 2 feet wide. The cricket prevents debris or snow and ice from piling up behind the chimney. This can cause rain or melting snow to back up under the shingles and leak into the house. The cricket also deflects water running down the roof around the chimney.

For the most part, very little maintenance is needed for prefabricated chimneys. Metal chimneys have a tendency to rust and should be checked for corrosion holes. If a metal chimney has no rain cover, note it on the worksheet; a cover is recommended.

Vent stacks

The vent stack is part of the plumbing system. Its purpose is to permit adequate circulation of air in all parts of the sanitary drainage system and to allow a means for sewer gas to vent harmlessly to the atmosphere. All homes should have at least one vent stack. (See discussion of air admittance valves in the Violations section of chapter 9.) The absence of a vent stack usually indicates that the plumbing system is not properly vented, almost always a violation of the plumbing code. As you inspect the roof, look for a pipe that projects through the roof and terminates about 8 inches above the roofline. In newer homes, the vent stack is usually fairly obvious. In older homes, however, the vent stack might be missing or at least not visible when looking at the roof. On occasion, I have found vent stacks that terminate in the attic. This condition is a violation of the

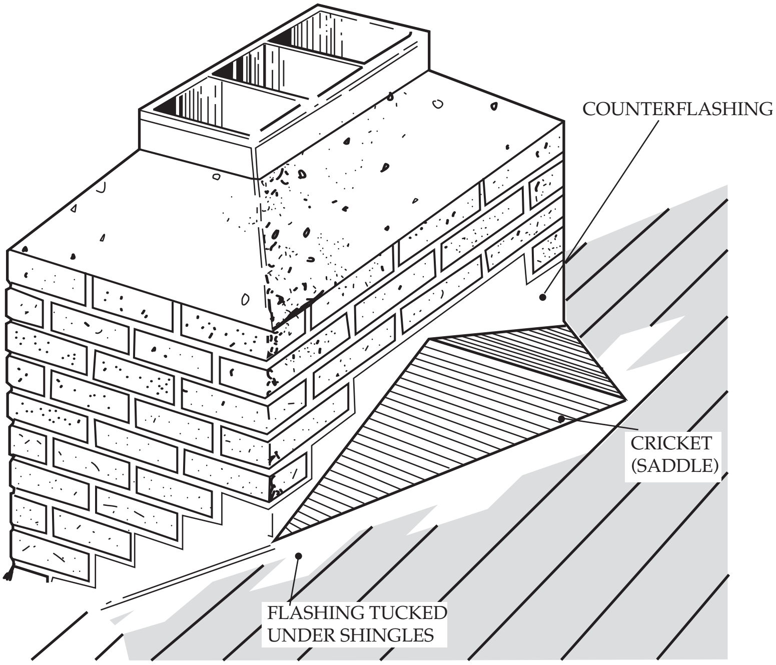

Fig. 3-7. *Cricket (saddle) prevents debris, snow, and ice from accumulating behind the chimney, causing roof leaks.*

plumbing code; the vent stack must extend above the roofline so that the escaping gases can discharge to the outside.

In some older homes, when kitchens or bathrooms are renovated, vent stacks are often run up along the outside of the building to a point above the roofline. (See FIG. 3-8.) If these homes are located in the northern part of the United States, this type of installation is undesirable. Because of the moisture in the escaping gas, during cold weather it could freeze over and eventually block the pipe. Sometimes you find vent stacks that terminate near a window. (See FIG. 3-9.) This installation is also undesirable because the discharging sewer gas might seep into the house when the window is open.

For vent stacks that terminate above the roof, very often you'll see a black ring at the base of the stack adjacent to the flashing. This black ring is asphalt cement that has been used to seal the joint. This joint is vulnerable to leakage and should be periodically checked. On one inspection, I found water dripping from a ceiling tile in a suspended ceiling located two levels below the roof. When I lifted the tile, water came cascading down. After the initial shock and after all the water had spilled away, it became obvious that the water was not caused by a leak in the above laundry room but had collected during the previous day's driving rain when water had leaked into the structure through the small crack between the vent stack and the flashing. This area should be periodically checked and sealed as needed.

Roof vents

These vents can be found on pitched and flat roofs and are available in both round and square hood styles. Normally, there are no problems with these types of vents. However, the joint between the vent and the roof is vulnerable to water leakage. Even though the joint might look okay from the roof side, it is

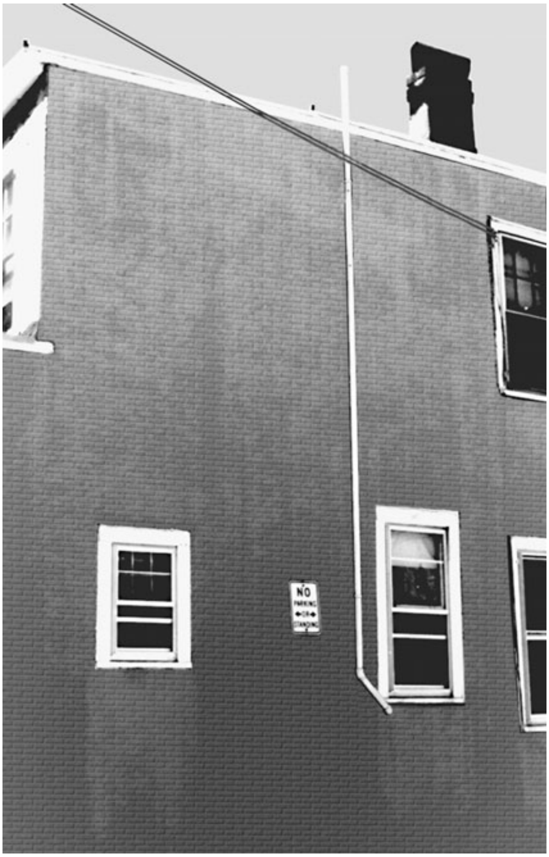

Fig. 3-8. *Exterior-mounted plumbing vent stack.*

best to check the vent openings from the attic periodically. If there is leakage, it will be noted by water stains on the roof deck in those areas. Water leakage around the joints can be easily corrected with asphalt cement or a suitable caulking.

Roof hatch

Access to a flat roof from the interior is provided through a roof hatch. Don't be surprised if you find a roof hatch on a house with a pitched roof. In some older and larger structures, they were installed as a means of easy access. Roof-hatch covers should be checked to determine whether they are operational; they should be. I have found that in about 30 percent of the cases, the hatch covers could not be opened. In all probability, during reroofing or maintenance, asphalt cement accidentally or intentionally sealed the hatch-cover frame. If this condition exists, it should be corrected. Usually, the hatch cover is constructed of wood and is covered with either sheet-metal or asphalt roll-roofing. Of particular concern is the integrity of the waterproofing cover. There should not be any cracked or open joints. The wood framing should not have any cracked or broken sections.

Skylights

Skylights are installed on a structure to provide daylighting and in some cases ventilation. The newer styles are prefabricated, with aluminum frames and plastic or glass panels or domes. Skylights can be found on pitched or flat roofs. These units should be checked for cracked or broken panes and signs of leakage. The leakage can be checked by inspecting the interior area below the skylight. If there are leaks around the skylight, you'll note water stains on the wall or finished ceiling in those areas. The skylights found on older flat-roof homes should be carefully checked for corroding frames and cracked and broken panes. Quite often the frames have corroded through and provide no support for the panes. This condition should be marked on your worksheet, because these skylights require complete rehabilitation.

TV antennas

Although roof-mounted TV antennas were quite common when the first edition of this book was written, for the most part they have been replaced by cable television or dish antennas. Nevertheless, they can still be found on many houses. Some TV antennas are strapped to the chimney for support. This means of

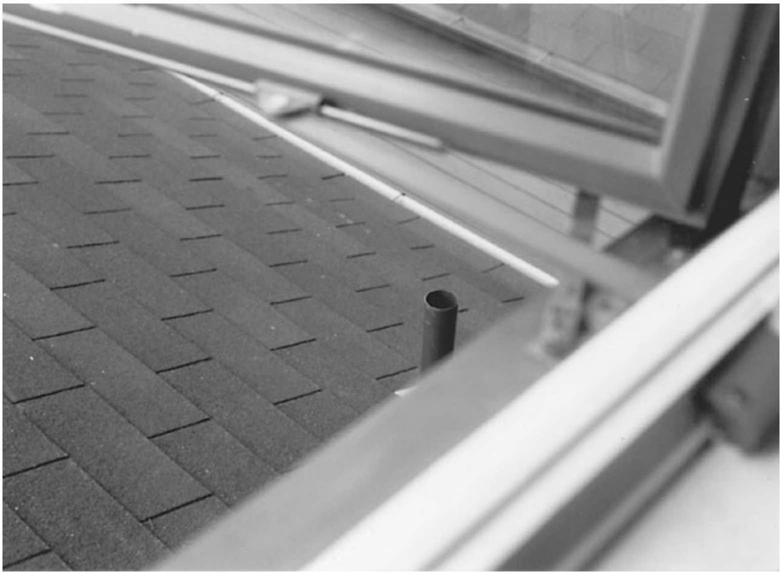

Fig. 3-9. *Plumbing vent stack terminating near window. If the window is open, the discharging sewer gases can be blown into the house.*

bracing the antenna is undesirable because repeated twisting action on the antenna from strong winds can cause stresses on the chimney that can in turn result in cracked mortar joints. When you see a TV antenna strapped to the chimney, look carefully at the joints. Sometimes the antenna is supported on the roof by means of guy wires anchored to the roof deck. When inspecting the roof, check the area around the anchors for deterioration of the roof covering. Sometimes the guy wires are strapped to vent stacks. This practice is undesirable because in a strong wind the movement of the antenna can cause the joint at the base of the vent stack to crack. I have also run into a situation where a homeowner inserted the TV antenna mast down into a vent stack to secure the antenna. If you see this situation, make a note of it so that you can remove the antenna after you take possession of the house. Finally, look for a ground wire that connects the antenna to a metal rod embedded in the ground. According to the National Electrical Code, this wire should not have any intervening splices or connections. The TV ground wire does not protect the house from lightning. Its purpose is to protect the television set in the event of a lightning surge.

Lightning protection

Although there are about 90 million lightning strikes each year in the United States, most homes don't have a lightning protection system. If a house has one, you will notice pointed metal rods projecting up above the high points of the building. These rods are connected to stranded cables, which are then connected to at least two grounded conductors. Of particular concern is whether the connections are properly bonded and if the system is properly grounded. Because of the destructive power of a lightning strike, the integrity of the system should be checked by a company specializing in lightning protection.

Gutters and downspouts

Gutters and downspouts are installed on a structure to control and direct rain runoff from the roof. The absence of gutters might result in water seepage into the basement or crawl space, rotting sections of wood trim, damage to foundation plantings, and the erosion of topsoil. Whether they are masonry-constructed or have long overhanging eaves, most residential structures that are not in the snowbelt would benefit from gutters. In the snowbelt, gutters are considered more an inconvenience than a help because the snow and ice often tear them from the supports and maintenance is constantly necessary. If you do not see gutters on the structure during your inspection, indicate their absence on your worksheet. There are basically two types of gutters, built-in and exterior-mounted.

Built-in gutters

These gutters are essentially extensions of the roof framing with waterways built into the roof surface over the edges. The gutter channel might be lined with asphalt roll-roofing or some other type of impermeable material. These channels require periodic maintenance, such as applications of an asphalt-type cement. Leakage through these channels can often be detected by water stains in the soffit below the leaks or by water stains in the interior. Leaks in this type of gutter often result in rotting sections of trim. If stains or rotting sections are noted, they should be indicated on the worksheet. Built-in gutters are seldom used on modern residential structures.

Exterior-mounted gutters

These gutters can be made from copper, galvanized iron, wood, plastic, or, most commonly, aluminum. Copper gutters are considered the top of the line. They are expensive, virtually corrosion-resistant, and have a projected life in excess of forty years. However, as these gutters age, they corrode and develop tiny holes in the bottom portion of the gutter channel. Depending on the gutters' height, the holes might not be visible to the naked eye from the ground. However, by standing directly below the gutters and looking straight up, the sky is often visible through any corrosion holes.

The joints on copper gutters are always soldered. In some parts of the country, the joints on galvanized gutters are also soldered rather than clipped. When the gutters are painted, you can tell the difference between them by using a magnet. The magnet will not stick to a copper gutter. Sometimes leaks develop around the soldered joints. For the most part, because of their cost, copper gutters and downspouts are no longer used when replacing faulty gutters and downspouts. A number of years ago I inspected a church in Ossining, New York, that had a wet basement. It turned out that most of the copper gutters and downspouts had been stolen from the structure by someone in search of a quick dollar because there was a copper shortage at that time.

Galvanized gutters have been used on many homes because of their low initial cost. However, they rust easily and require periodic maintenance such as patching corrosion holes and repainting. The inside portion of the gutter channel should also be painted. This area is often overlooked by the homeowner.

Aluminum gutters are quite popular because they do not have the corrosion problems of galvanized gutters. Older aluminum gutters on occasion leak around the seam, a condition requiring resealing. Leakage from seams can be noted by discoloration at the joint or some water stains or erosion on the area directly below the joint. Aluminum gutters can now be manufactured to almost any length, producing *seamless gutters*. Leaks in this type of gutter usually occur at corner joints or the joint around the downspout.

Wood gutters are usually made of Douglas fir or red cedar and have a tendency to crack and rot at the various end joints and seams along their length. (See FIG. 3-10.) The joints around the end sections, particularly those where the connection is made to the downspout, deteriorate more rapidly than other portions and should be checked for rot and cracking. Wood gutters should be painted every few years and the inside channel coated with an asphalt-type roof paint.

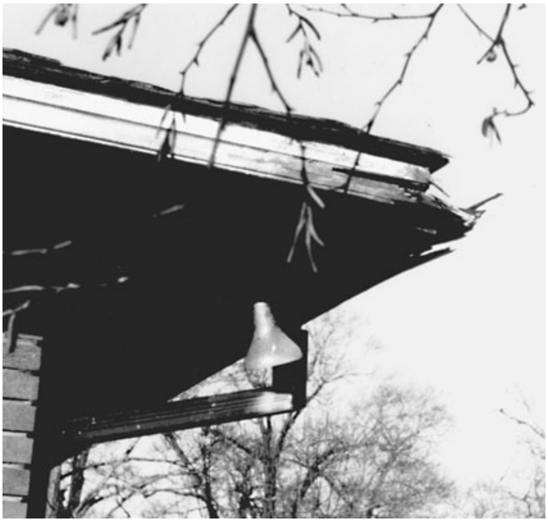

Fig. 3-10. *Wood gutters. Note the cracked and rotting corner joint.*

Plastic gutters, although relatively maintenance-free, have not received wide acceptance. They are found occasionally, but not necessarily, on homes with vinyl siding.

When inspecting the roof with binoculars, you should check the gutters to see if there are any loose support straps or spikes that should be resecured. (See FIG. 3-11.) In addition, an overall view will show whether any gutters are pitched incorrectly or are sagging and should be reset. Sagging is a condition occasionally found on those homes with slate roofs that do not have snow guards. All types of gutters have a tendency to become cluttered with leaves, twigs, seed pods, and mineral granules from roof shingles. They should be cleaned

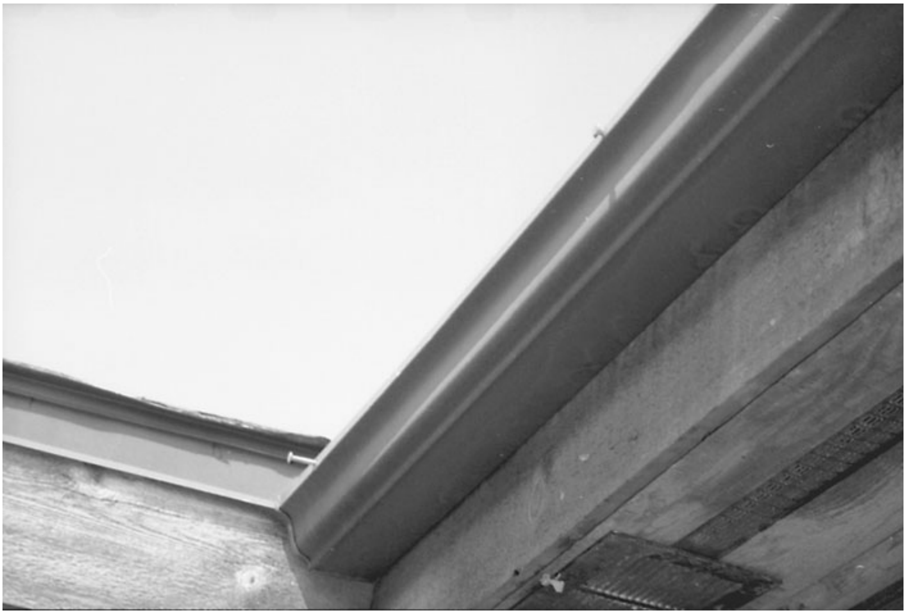

Fig. 3-11. *Loose gutter spike. Should be reset and resecured.*

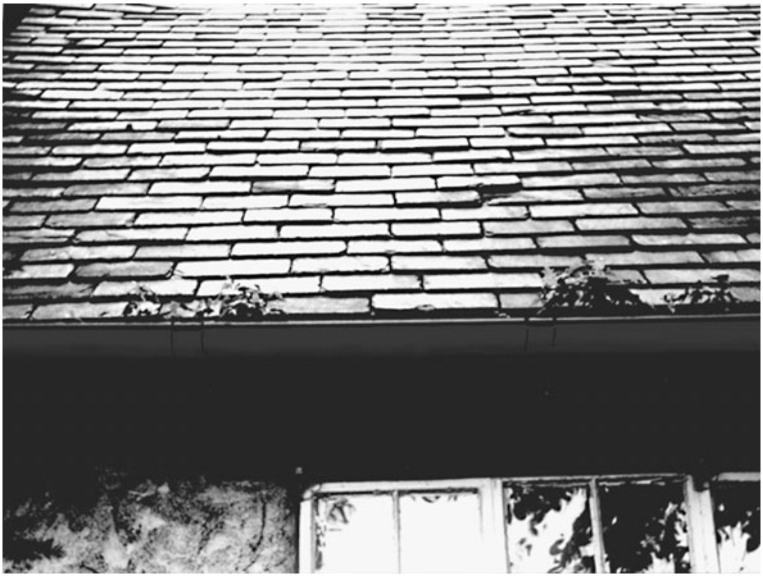

Fig. 3-12. Gutters must be cleaned periodically to be effective. Note the weeds growing out of this gutter.

at least twice a year—once in the spring after the trees have bloomed and once in the fall after the leaves have fallen. Gutter screens are available to help prevent larger items, such as leaves and twigs, from cluttering the gutter channel. However, often a homeowner installs the screens and then forgets about cleaning the gutters. The gutters still require cleaning, although at less frequent intervals. Sometimes you can tell from the ground whether a gutter channel requires cleaning. Figure 3-12 shows weeds growing out of the gutter. Obviously, this gutter has not been cleaned for some time.

Downspouts

Downspouts are normally constructed of aluminum, copper, or galvanized iron. Copper and galvanized downspouts that have aged often have corrosion holes in the elbow sections. In some cases, the copper corrodes so that cracked sections are paper thin and can be stripped away easily. If you see this condition, it indicates that those sections should be replaced. Some downspouts have loose and

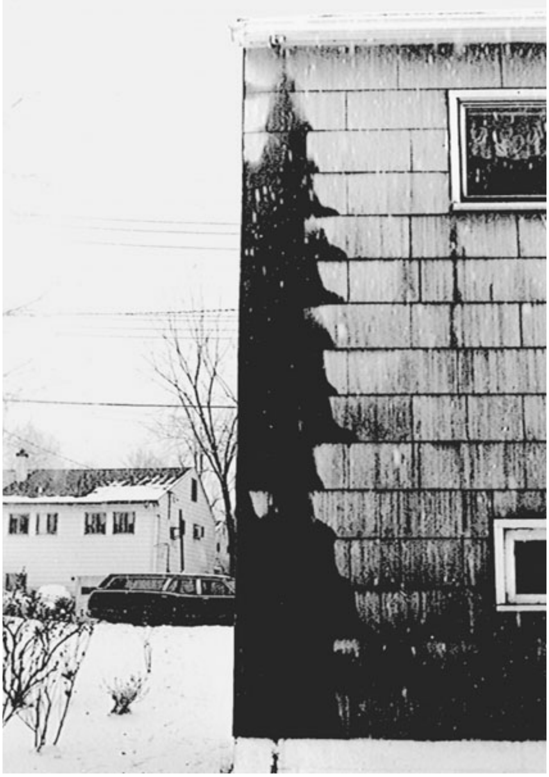

Fig. 3-13. *Water stains on the exterior siding are the result of a missing downspout.*

open seams along their length that interfere with their effectiveness.

If you do not specifically look for the downspouts, you might not realize that some of them are missing. Figure 3-13 shows water stains on the asbestos-cement shingles, the result of a missing downspout. In this case, it is obvious that a downspout is missing; however, if the house is inspected during dry weather, you might not see any stains, and this item can be easily overlooked. Check the assembly of the downspout at the various joints. The lower portion of the downspout should be outside the upper portion. Otherwise, water will leak around the joint. Figure 3-14 shows an incorrect downspout assembly. On occasion some downspouts come loose from their connection to the gutters and should be resecured. Loose support straps around downspouts should be resecured.

The water from downspouts must go somewhere; it can be piped away underground to suitable drainage or discharged onto the ground at the base. Any water discharging from the base of the downspouts must be directed away from the structure. Otherwise, it can accumulate around the foundation and can eventually enter the lower level. To help deflect this water away from the structure, there should be an elbow at the base of the downspout and a splash plate below on the ground. Effluent discharging around the foundation should be indicated on your worksheet.

Sometimes the downspouts terminate directly into the ground. It is usually not possible during an inspection to determine whether the downspouts are connected to free-flowing drain tiles or to dry wells. Free-flowing tiles are more desirable; dry wells can become clogged (see FIG. 3-15) or will become

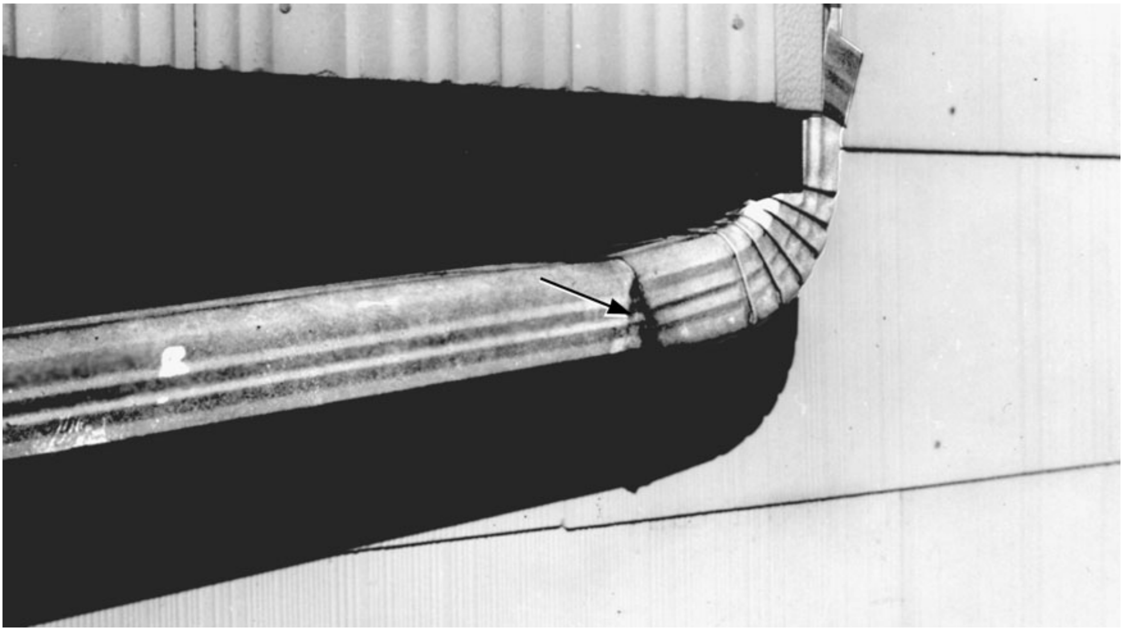

Fig. 3-14. *Incorrectly assembled downspout joint; the lower section is inside the upper section. It should be reversed; otherwise, water will leak out of the joint.*

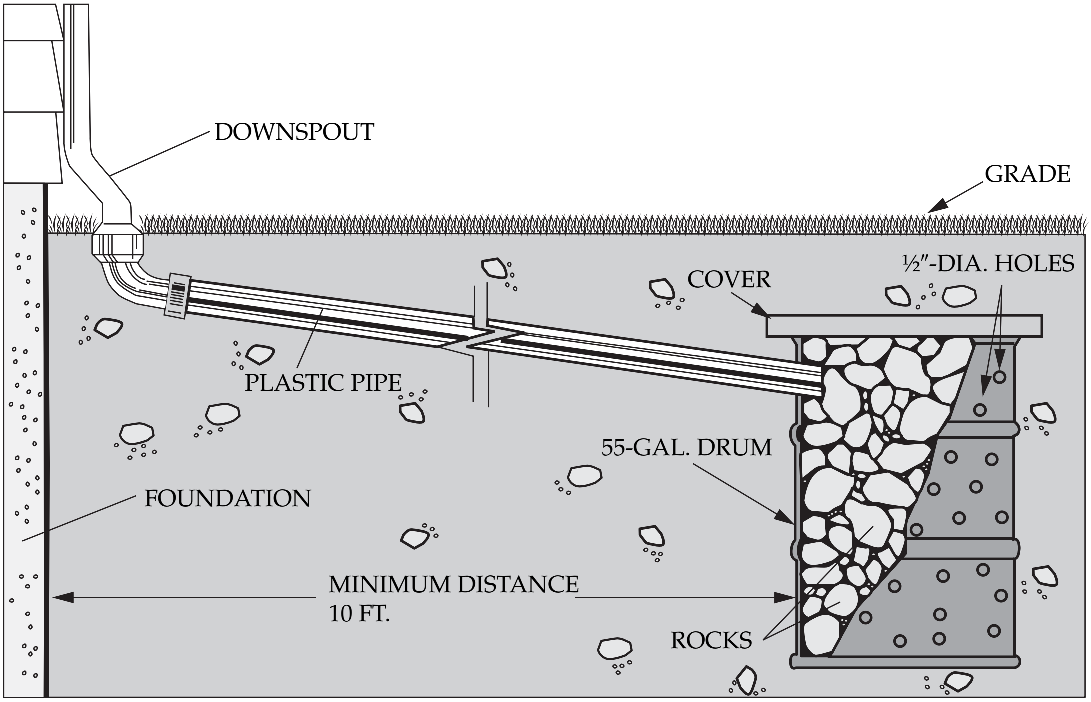

Fig. 3-15. *Over the years dry wells can become clogged, causing rainwater in the downspout to back up.*

ineffective if the level of the subsurface groundwater (water table) is high. Occasionally, some free-flowing tiles are visible at the street curb. If there are outlets at the curb, they must be kept clear.

Checkpoint summary

Chimneys

Masonry type (brick, stone, concrete block)

❍ Inspect for cracked, loose, chipped, eroding, or missing sections of masonry.
❍ Check mortar joints for cracked, loose, and deteriorating sections.
❍ For stucco-finished chimneys, cracked, chipped, missing, or loose sections of stucco?
❍ Chimney vertical?
❍ Open joints between the chimney and the sidewall?
❍ If the roof is flat, does the chimney extend 3 feet above the roofline?
❍ If the roof is pitched, does the chimney extend 2 feet above the roof ridge?
❍ If possible, check for cracked or missing sections of the chimney cap.
❍ If possible, check the chimney flashing for holes, tears, or loose sections. Check these vulnerable areas for leakage again during your attic inspection.
❍ If possible, check to see if the chimney flue is lined.
❍ Is there a chimney top damper?
❍ Is there a cricket (saddle) behind the chimney?

Metal type (prefabricated)

- ❍ Check for corrosion holes, rusting, or missing sections.
- ❍ Rain cover present?
- ❍ Note condition of flashing.

Vent stacks

- ❍ Plumbing vent stacks visible? (Their absence can be verified during attic inspection.)
- ❍ Black rings (asphalt cement) at the base joint of the vent stacks?
- ❍ Note any questionable roof joints and check further during attic inspection.
- ❍ Note vent stacks that
  —terminate near windows
  —run up an exterior side of the house (in northern climates)
  —have TV antennas strapped to them.

Roof vents, hatches, skylights, TV antennas, dish antennas, lightning protection

- ❍ Check all roof joints associated with vents, hatches, and skylights.
- ❍ Note questionable sections and verify tightness of joints during attic inspection.
- ❍ Does structure contain a roof hatch?
- ❍ Can you open it?
- ❍ Cracks or open joints in the cover?
- ❍ Check skylights for corroding frames and cracked or broken panes.
- ❍ Inspect ceiling area below for signs of leakage.
- ❍ Check TV antenna mast and guy wire connections to roof.
- ❍ Is antenna adequately grounded?
- ❍ Are lightning rods secure?

Gutters and downspouts

Exterior-mounted gutters

- ❍ Check for missing sections of gutters.
- ❍ Note type of material: copper, aluminum, galvanized iron, wood.
- ❍ Note gutter sections incorrectly pitched.
- ❍ Check metal gutters for corrosion holes, sagging sections, loose support straps or spikes, and leaking sections.
- ❍ Check wood gutters for cracked sections and areas of rot, particularly at connections and end sections.

Built-in gutters

- ❍ Check for areas with rotting trim.
- ❍ Check for signs of leakage.
- ❍ Note signs of seepage (stains) in soffit trim below gutters.
- ❍ Wherever possible, check the condition of the gutter channel.

Downspouts

- ❍ Note type of material: copper, aluminum, galvanized iron.
- ❍ Check for missing sections and improper joining.
- ❍ Inspect for loose straps, open seams, and corrosion holes at elbows.
- ❍ Do downspouts have elbows at base and extensions and splash plates (where required) to direct roof rain runoff away from the house?
- ❍ If downspouts terminate in the ground, try to find out whether they are connected to dry wells or to free-flowing outlets.

4

Paved areas around the structure

Sidewalks 36
Street-level/driveway-level steps 36
Front and side paths 37
Entry steps 37
Patio 38
Driveway 39
Checkpoint summary 41

As you walk around the house on the second pass, inspect the paths, steps, patio, and driveway. The problems normally encountered with these items usually do not require immediate correction. Nevertheless, a tripping hazard might exist, cosmetic maintenance might be needed, or a condition might make the lower level vulnerable to water penetration. If you see problems in these areas, record them on your worksheet for early correction.

Sidewalks

Sidewalks are usually constructed of concrete, asphalt, or stone embedded in the ground. Not all communities have sidewalks. If a sidewalk is in front of or on the sides of the house, look for cracked or uneven sections because they are a potential tripping hazard. Even though the sidewalk might be on property owned by the town, its maintenance is usually the homeowner's responsibility. If someone should be hurt because of a cracked or uneven section, you would be vulnerable to a lawsuit. Very often uneven sections are caused by the roots of trees growing nearby. Correcting the condition would require chopping that section of root away and resetting the sidewalk. You should be aware of the fact that trees adjacent to the sidewalk are often the property of the local municipality, and prior to altering the roots, permission should be obtained.

Street-level/ driveway-level steps

Steps might lead from the sidewalk or driveway to the front path. These steps should be inspected for cracked, chipped, broken, or uneven sections. Look specifically for dimensional variations in the step risers. The riser (vertical distance) for the top step or the bottom step is often a different size than those for the other steps. This difference is a tripping hazard because it interrupts the natural rhythm of ascending and descending the steps. If there are more than two steps, there should be a handrail as a safety precaution.

Front and side paths

Two general types of paths are used with residential structures. One is a ribbon type, generally constructed of concrete or asphalt; the other is a sectional type, generally inlaid with material such as stones, bricks, sections of tree trunks, or precast concrete blocks. The sectional type normally requires periodic maintenance because of the tendency toward uneven settlement and weed growth between the sections. Sometimes the sections are loose and uneven and present a tripping hazard. Occasionally, the sections are set in mortar. Look for loose, cracked, and chipped mortar joints that require repointing. When inspecting the ribbon-type path, look for cracked, uneven, and broken sections. Very often damage occurs because the base below the path was not properly prepared during construction. Look particularly for settled sections that are sloping toward the house. These areas will direct rain runoff toward the house, so that water accumulates around the foundation. This water can enter the lower level. In this case, the path should be repositioned so that it slopes away from the structure or is completely rehabilitated.

While inspecting the paths, look for small, abrupt changes in the elevation. Occasionally, I have found a single step in the middle of a path. This step is a potential tripping hazard because it often goes unnoticed. A single step in a path, except at an entrance, should be avoided. If a slight elevation change is necessary and there is a step, it should be converted to a ramp, or shrubs should be planted at the step to call attention to the elevation change. An outdoor light should also be placed here. (See FIG. 4-1.) You might find a path partially blocked by overgrown shrubs; consequently it is no longer functional. In this case, if a path is needed, the shrubs should be pruned or the path repositioned.

Entry steps

Entry steps can be made of stone, concrete, brick, metal, or wood. As a safety precaution, when more than two steps are necessary, at least one handrail should be installed especially if the house is located in an area where the temperature drops below freezing.

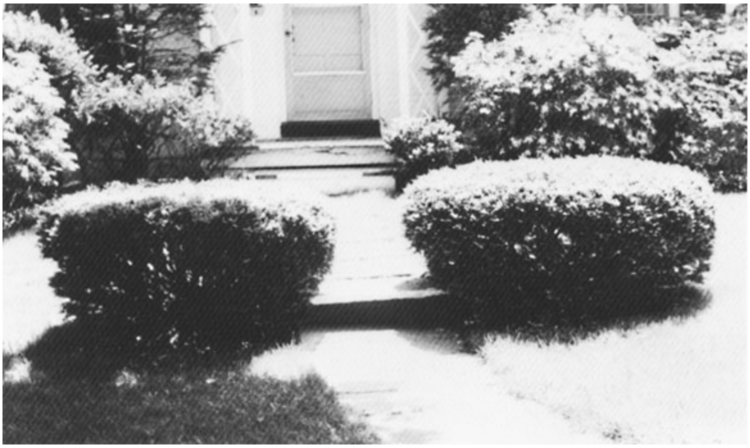

Fig. 4-1. *A single step in the middle of a path is a potential hazard. The shrubs on both sides of this path call attention to the step. However, there should also be an outdoor light in the area.*

In those areas, the steps can be coated with a layer of ice after a freezing rain. (See FIG. 4-2.) Look specifically for differences in the vertical distance between the steps (risers). (See FIG. 4-3.) Any dimensional variations in the risers are potential tripping hazards. Some steps are designed so that the vertical distance between the treads is open (open risers). Although this might be aesthetically pleasing, it is a potential tripping hazard.

If the entry door opens onto the entry steps, a landing platform rather than a step tread is necessary at the doorway. The turnaround area of a single step tread is not considered adequate to operate a door safely. The platform should provide sufficient space to allow adequate standing room while opening the door, which swings over the landing. (See FIG. 4-4.)

When inspecting the steps, look specifically for cracked, broken, rotting, chipped, and loose sections. The treads should be level. Uneven sections are a tripping hazard. If the steps are masonry-constructed, look at the step foundation walls for cracked, broken, and chipped sections. Any wood *stringers* (the side portions of the steps that support the treads) should be resting on a concrete pad rather than on the soil. Check the base of wooden stringers for rot by probing the area with a screwdriver. If the screwdriver penetrates the wood easily, the stringer should be replaced. Metal handrails are often used with exterior steps. Very often, these handrails are corroding and have deep pockets of rust. In this case, the handrails should be scraped, primed, and repainted.

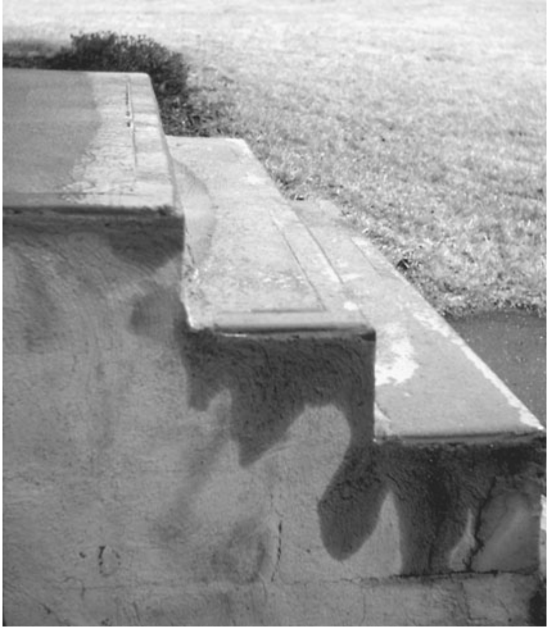

Fig. 4-2. *Handrail is needed for steps, especially when they could be covered with ice.*

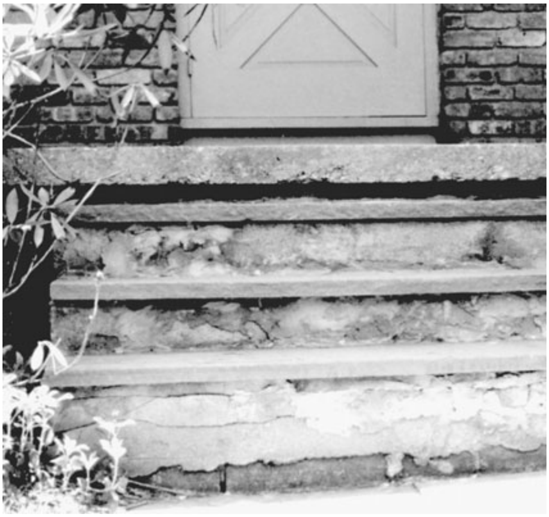

Fig. 4-3. *Note that the vertical distances between the steps vary. These steps are a potential tripping hazard. The bottom riser needs masonry repair.*

Patio

There are probably as many types and styles of patios as there are houses. The more common types are concrete slab, stone set in mortar, and brick or paving blocks set in the ground. In all cases, a particular concern is whether any tripping hazards can result

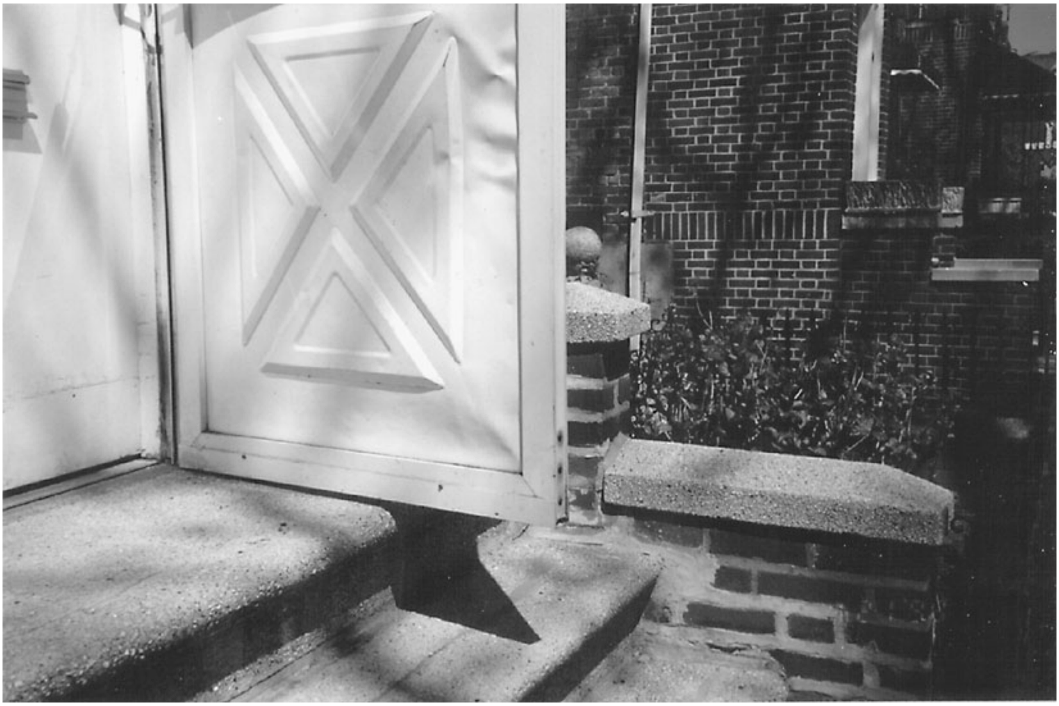

Fig. 4-4. *The turnaround area of the top step is too narrow to operate the door safely.*

from cracked, broken, or uneven sections. Some patios have a grid pattern consisting of wood embedded in the ground around sections of brick or concrete. This wood should be pressure-treated to protect it from rot. However, often the wood is not pressure-treated, so that after a few years it rots and requires replacement. Concrete-slab patios should have expansion joints and control joints for cracking. Very often these items are omitted, and the patio cracks in a random fashion that is aesthetically undesirable. Look at the patio for signs of uneven settlement. If a patio is adjacent to the house and has settled so that it is sloping toward the house, water can accumulate around the joint between the patio and the foundation wall. This water will eventually accumulate around the foundation and if the house has a basement or crawl space, it could enter those areas. This condition should be indicated on your worksheet for later correction and as a reminder for you to check the area of the basement adjacent to the patio.

Driveway

Driveways are normally constructed of asphalt, concrete, gravel, and, more recently, individual paving blocks. They also might just be a clearing with no covering. The latter is not particularly desirable from a cosmetic point of view and because of the ease with which ruts can develop. Gravel driveways on occasion also develop ruts and require periodic replacement of the gravel. Concrete and asphalt driveways should be inspected for cracked, broken, or settled areas that require rehabilitation. Extensive cracking, such as alligator cracks, is usually an indication of poor

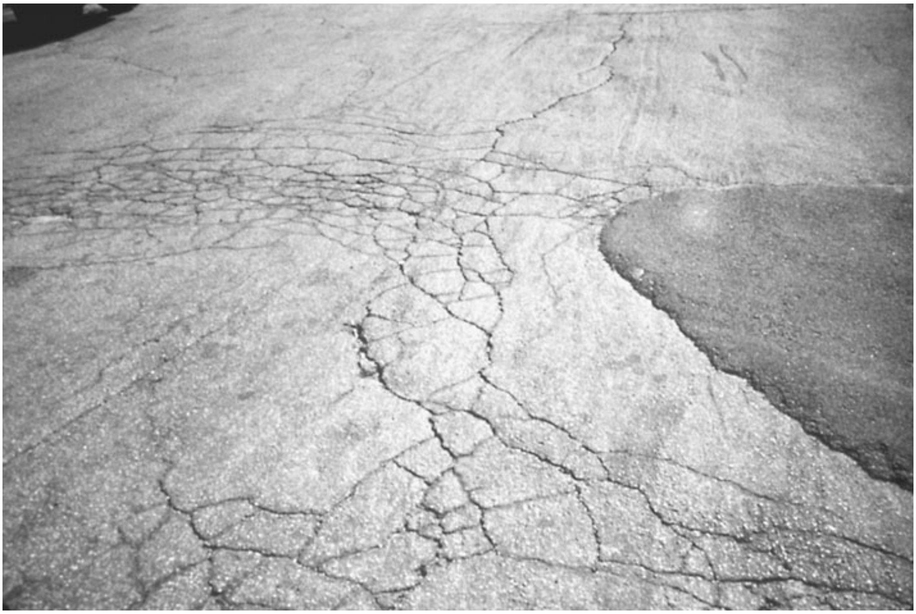

Fig. 4-5. *Extensively cracked and deteriorating driveway.*

drainage of the subbase or poor construction. (See FIG. 4-5.) In this case, the cracks cannot be sealed effectively. The entire area should be patched. If a large portion of the driveway has this type of cracking, resurfacing is required.

If the base below a paving block driveway has not been prepared properly, sections will settle causing water to pond during a rain. (See FIG. 4-6.) Although the water will drain into the joints between the blocks, this item should be recorded on your worksheet so that it can be corrected at a later date.

Sometimes the driveway is on an incline, so that subsurface water flows below the driveway, undermining the subbase. In this case, prior to resurfacing, adequate provision for drainage of the subbase should be made.

When standing at the front of the driveway, look to see if it pitches directly down toward the house. If it does, the garage and any other portion of the lower level are vulnerable to flooding unless a large operational drain at the base of the driveway intercepts the surface water. Usually a channel drain that runs across the width of the driveway is needed to control such water runoff. The drain, however, should have a free-flowing outlet. If the drain is connected to a dry well, it might be ineffective during those months when there is a high water table; one corrective procedure would be

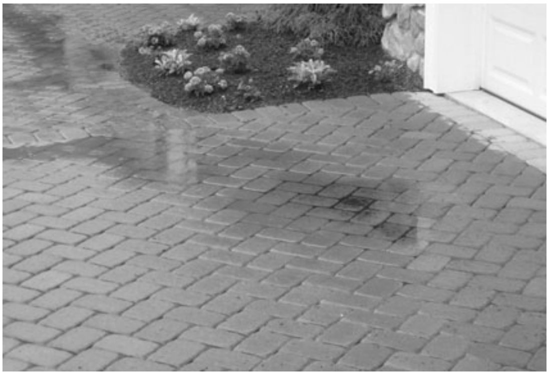

Fig. 4-6. *Water ponding on a paving block driveway.*

to connect the drain to a sump pit and remove the water by means of a sump pump.

When a house is located on a street that is inclined, the curb cut for the driveway should not be feathered into the street. Instead there should be a small ridge on the driveway at the joint between the driveway and the street. This ridge will prevent water that normally accumulates around the curb from flowing onto the driveway. This is normally not a problem in newer structures. However, in older houses the ridge tends to deteriorate, allowing water to overflow. This condition should be corrected.

In some northern areas, raised and inclined driveways have scratch marks that are caused by studded snow tires. This indicates that those driveways will be difficult to negotiate during some winter months; 100 pounds of salt and sand should be stored in the garage or near the driveway.

The minimum width for a driveway is 8 feet, although 9 feet is preferred. If the driveway is used both as an areaway for the car and a walkway in place of a path, it should be at least 10 feet. Anything less will make walking quite difficult when a car is parked in the driveway. Note whether the driveway discharges into a heavily trafficked street. If it does, an area in the driveway should function as a turnaround to allow a driver to head onto the street rather than back out onto the street. Also, look for overgrown trees and shrubs at the end of the driveway that might obstruct the driver's view when entering the street.

Checkpoint summary

Sidewalks, paths

❍ Inspect for cracked, missing, eroding, and uneven sections.

❍ Check for areas that might present a tripping hazard.

❍ Check slope of all paved paths adjacent to the house for improperly pitched sections.

Street-level/driveway-level steps

❍ Inspect for cracked, chipped, broken, or uneven sections.

❍ Check for missing handrails or railings.

Entry steps (masonry, wood)

❍ Inspect for cracked, broken, loose, or deteriorating sections.

❍ Note potential tripping hazards such as variations in riser heights and narrow treads.

❍ Check (probe) for rot in wood stringers, step treads, and handrails.

❍ Are wood stringers supported on concrete pads, or are they resting on the earth?

❍ Check for handrails.

❍ Inspect metal handrails for rusting, loose, and broken sections.

❍ Inspect wooden handrails for cracked, broken, loose, and rotting sections.

❍ Does entry door open onto a step or landing?

Patio

❍ Check for cracked, broken, eroding, and uneven areas.

❍ Inspect for uneven and settled sections adjacent to the house that can allow water to accumulate around the house foundation.

❍ Check for rot and insect damage to embedded wood sections.

Driveway

❍ Inspect for cracked, broken, eroding, or settled areas. Note extensively cracked and deteriorated areas for future rehabilitation.

❍ Check slope of driveway: level, raised, or inclined?

❍ For an inclined driveway, is there an adequate drain at the base?

❍ Does drain discharge to a drywall or to a free-flowing outlet?

❍ Is driveway width adequate (8 feet minimum, 9 feet preferred)?

5

Walls, windows, and doors

- Exterior walls 43
  - Exterior siding 44
    - Wood siding 44
    - Aluminum siding 46
    - Vinyl siding 47
    - Asbestos-cement shingles 47
    - Fiber-cement siding 48
    - Asphalt siding 48
    - Stucco 49
    - Synthetic stucco 49
    - Veneer wall 50
  - Masonry wall 51
- Trim 52
- Windows 53
  - Inspection 55
- Exterior doors 56
- Storm windows, screens, and storm doors 58
  - Storm windows 58
  - Screens 58
  - Storm doors 59
  - Inspection 59
- Caulking 59
- Checkpoint summary 60

As you walk around the house inspecting the paved areas, you should also inspect the exterior walls, windows, trim, and doors. Before actually inspecting these items, however, look at the overall wall area for indications of past or current structural problems. Are the window and door lines square? Are any portions of the walls sagging or bulging? Are the walls and corner sections vertical? A problem condition should be recorded on your worksheet for further investigation. If by the end of the house inspection you cannot determine the cause, you should have the condition checked by a professional.

While inspecting the exterior walls, also note for further investigation any pipe or hood projecting through a wall or basement window. These items are usually not problem conditions, but it is useful to understand their function. The hood is often covering the discharge end of an exhaust fan or clothes-dryer duct, and the pipe may be the discharge line for a sump pump or condensate line from an air-conditioning system. If the pipe is connected to a sump pump, it might indicate a past or current water-seepage problem. Sump pumps and water seepage are discussed in detail in chapter 11.

Exterior walls

The exterior walls in most residential structures will be either wood frame or masonry, sometimes a combination of the two. The latter is commonly called a *veneer wall*. The exterior walls rest directly on the foundation and are *bearing* (load-supporting) walls. They support the roof, floors, and vertical loads imposed by other building components. The outer covering of the exterior walls provides protection from the weather and, if properly installed, minimizes the flow of air, moisture, and heat into or out of the structure.

When the walls are wood frame, the vertical framing members (*studs*) support all the vertical loads, and the outer finish covering (generally called *siding*) provides weather protection. Insulation is normally located in the spaces between the studs. In masonry walls, the masonry (clay tile, brick, stone, concrete block, etc.) provides both the structural support and the weather barrier. A masonry-veneer wall is a wood-frame wall with masonry used in place of the siding. Although the masonry in a veneer wall is not used for supporting the vertical loads, it does support its own weight.

Basically, a wood-frame exterior wall consists of 2-by-4-inch or 2-by-6-inch studs covered on the interior side by materials such as plaster, Sheetrock, plasterboard, wood, or hardboard panels (as described in chapter 10) and on the exterior side by sheathing, sheathing paper, and the finish siding. (See FIG. 5-1.) In some parts of the country where rot and termite activity are problems, metal studs are now being used in place of wood studs.

Sheathing is installed over the studs to provide bracing and minimize air infiltration. Depending on the type of sheathing, it can also be used to form a surface onto which the exterior finish can be nailed. Wood boards, plywood, fiberboard, and plasterboard are often used for wall sheathing. Fiberboard sheathing adds a small amount of insulation to the overall exterior wall; however, it should not be used as a nailing base for the direct attachment of the exterior siding. Rather, the siding should be nailed either to the studding through the sheathing or to wood nailing strips that have been attached to the sheathing. In many communities, when the exterior siding is capable of supplying adequate bracing and weather protection (as with exterior plywood panels), the sheathing is often omitted.

The purpose of the sheathing paper is to resist the direct entry of water during a driving

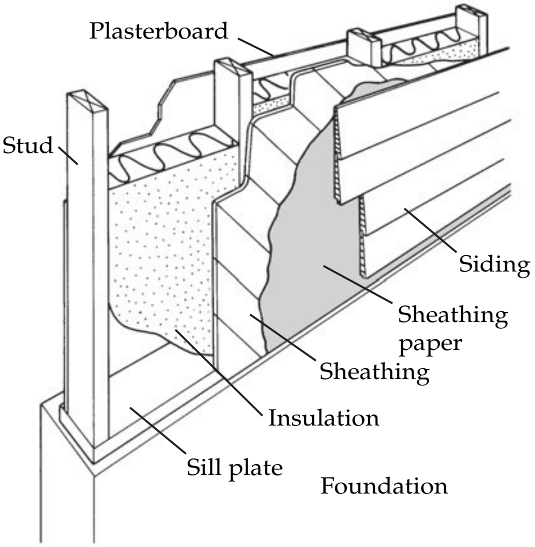

Fig. 5-1. Components of a wood-framed exterior wall.

rain. Sheathing paper (an asphalt-saturated felt) is water-resistant but not vapor-resistant and allows water vapor (which often builds up in the voids of the frame wall) to escape rather than condense and cause problems. Sheathing paper can also be very effective in reducing air infiltration.

Exterior siding

There are many types of exterior siding. If the siding is of good quality and has been maintained, it could last as long as the house. Normally, the type of maintenance required (other than painting) is the repair of cracked, broken, loose, rotting, or missing sections. When maintenance is needed, it is usually on a small section rather than the entire wall. However, some homeowners neglect the siding and allow it to deteriorate to a point where complete re-siding is necessary. Several clients have asked me about the condition of the old siding on a house that had been re-sided. Actually, the degree of deterioration of the old siding does not matter as long as the new siding is properly installed and provides the needed weather protection.

When inspecting the siding, pay particular attention to the sections that are facing south or southwesterly. These areas receive the maximum exposure to the sun and are more vulnerable to weather deterioration. The bottom of exterior siding should not be close to, or in contact with, the ground. Because of the dampness associated with the ground, the bottom of the siding should be at least 8 inches above the finished grade. Otherwise, the wood siding or the wood nailing boards for nonwood siding will be vulnerable to rot and termite infestation.

On occasion I have found vines growing up an exterior wall, in some cases reaching the roof. Although this might be aesthetically pleasing, the vines are undesirable. They can cover a multitude of problems and can cause problems. The vines can conceal termite shelter tubes (see chapter 8) or cracked portions of the siding. They can widen cracks, damage mortar joints, and loosen shingles. In addition, the dampness associated with the vines can promote rot and cause paint to blister and peel. If you find vines growing up an exterior wall, you should consider removing them.

Wood siding Wood siding is a broad classification that includes shingles, shakes, boards (applied vertically and horizontally), plywood panels, and hardboard. When inspecting wood siding, pay particular attention to exterior corner joints and the joints between the siding and window frames and doorframes. In addition, the area where the siding joins a dissimilar material (such as masonry or metal) is vulnerable to water penetration during a driving rain. These areas should be checked for weathertightness and rot.

On wood siding you might find dark, blotchy sections. This condition is generally caused by spores of fungi or mildew and often occurs in shaded areas. It is not a concern because it does not cause the siding to decay. However, it is unsightly. On painted surfaces, it can often be removed by washing with bleach and water. New paint on such areas should contain a mildew-inhibitor additive. On unpainted wood surfaces, this condition can usually be controlled by coating the siding with a penetrating preservative containing a mildewcide. You might also find brown and black discolorations on the siding (FIG. 5-2). This staining is caused by rusting of the nails used to secure the siding. The discolorations could have been avoided if aluminum or galvanized (rust-resistant) nails had been used. Eliminating this condition is somewhat difficult and usually not cost-justified.

Wood shingles/shakes Most shingles and shakes (hereafter referred to as shingles) used for exterior sidewall application are made from cedar or redwood. They are basically the same ones that are used for roofing (as described in chapter 2). However, their application is somewhat different. Because vertical walls present fewer water-penetration problems than roofs, the shingles on walls can be installed with a greater weather exposure than those on roofs. In addition, roofs generally have a three-ply layer of shingles, whereas exterior walls have only a two-ply layer. For full weather protection, the butt joints between the wall shingles for the upper ply should not line up with the vertical joints for the lower ply. Otherwise, water can penetrate the wall during a driving rain.

Since the shingles are decay-resistant, they do not have to be painted for weather protection. However, new shingles that replace deteriorated unpainted weathered shingles will not match the remaining shingles. Many shingles are painted to achieve a color decor. After a number of years, the paint begins to peel and flake. Consequently, once the shingles are painted, they will require periodic repainting for cosmetic purposes.

When inspecting the shingles, look for cracked, loose, chipped, rotting, and missing sections. Check for open joints around window and door trim. (See FIG. 5-3.) Open joints allow water penetration during a rain. Inspect for warped shingles; they generally will be on the sidewall with the southerly or southwesterly exposure. Also look at the quality of the shingles. You might find that the top portions of some are paper thin and can crack or chip very easily. These shingles are of a lower quality and are intended for use as an undercourse or installation with less shingle-length exposure. This type of shingled sidewall has a short projected life, and periodic repairs, with eventual residing, should be anticipated. Try to lift a few shingles gently. They should not lift up. If they do, it is an indication that they were improperly nailed. Shingles that are nailed directly to fiberboard sheathing rather than to wooden nailing strips (attached to the sheathing) will lift up under gentle pressure. Any of the above conditions should be recorded on your worksheet for future correction.

Wood boards Wood-board siding can be applied horizontally or vertically. Horizontal siding tends to make a house appear lower and longer; vertical siding tends to make a house appear taller (and is popular on one-story houses). The wood used for board siding should be free of knots. Otherwise, over a period of time, shrinkage can cause the knotty

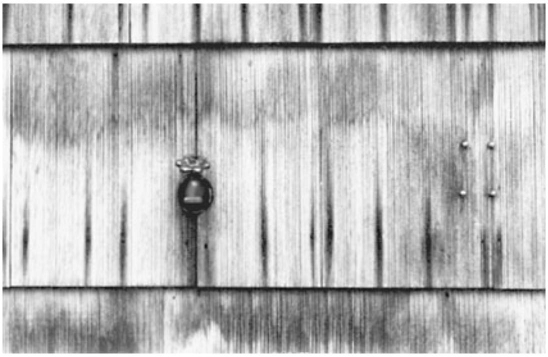

Fig. 5-2. *Discolorations of wood shingles caused by rust stains from iron nails.*

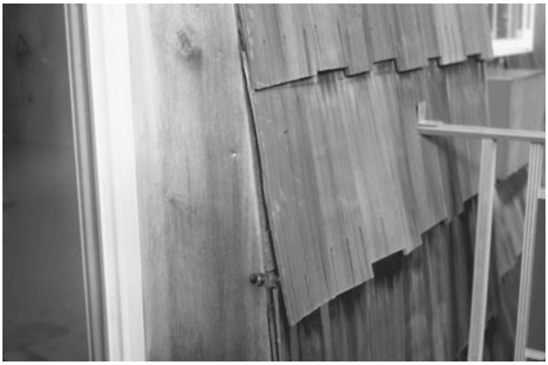

Fig. 5-3. *Open shingle joints allow rainwater penetration.*

cores to drop out, leaving the siding with holes that are vulnerable to water penetration.

With the exception of redwood and cedar boards, most wood siding is painted for protection against weathering and decay. When inspecting the siding, look at the condition of the paint. Are there any bare, peeling, flaking, or blistered sections? If there are, paint touchup or repainting may be needed. Blistered and peeling paint is often caused by moisture in the painted wood, although it can also result from a poor paint job. Determination of the exact cause for blistering and peeling paint cannot be made during a single inspection. Nevertheless, this condition should not affect your thinking about the house. If it is caused by moisture in the wood, it can be corrected after you move in, and usually at minimal expense.

Wood-board siding should be inspected for cracked, loose, and rotting sections. In addition, look for loose or missing knots. All holes should be patched with a wood filler. In vertical siding, check the joints between the vertical sections for weathertightness. In both vertical and horizontal siding, pay particular attention to the outside corner joints. These joints are vulnerable to water penetration during a rain, and any open joints must be sealed.

Plywood panels Plywood panels are also used for siding. They are made from exterior-type plywood in which the veneer layers are bonded together with a waterproof glue. The exterior facing of the panel comes in a variety of surface textures and grooves. The panels are 4 feet wide by 8, 9, or 10 feet long. The thickness of the panel will depend on the depth of the grooves and will generally vary between ⅜ and ⅝ inch. Plywood panels are usually installed in a vertical position with the vertical joints over studs. This minimizes the number of horizontal joints, which is desirable since plywood panels are often applied directly to the studs rather than over sheathing. Because of their vulnerability to water penetration, any horizontal joints should be shiplapped (not visible to the inspector) or protected by metal flashing (visible). When inspecting plywood siding, look for loose, warped, cracked, delaminated, and rotting sections. Also check for open and nonweathertight joints. If the panel siding is painted, check the finish for peeling and flaking paint and blistered sections.

Hardboard siding Hardboard siding is made by bonding (under heat and pressure) wood fibers that have been ground almost to a pulp. The siding is dense and tough, and has a fairly good dimensional stability, although not as good as that of plywood. Hardboard, like plywood, is available in a wide range of textures and surface treatments. It is available in 4-foot-wide panels and 9- and 12-inch-wide planks. When inspecting hardboard siding, look for cracked, chipped, broken, deteriorated, and loose sections. Also check horizontal and vertical joints for weathertightness.

Aluminum siding Aluminum siding is often used in new construction and is often used when re-siding the exterior walls. The siding comes in planks that are either smooth or embossed with a wood-grain texture (to resemble painted wood boards), also as shingles and vertical panels. Aluminum siding is relatively maintenance-free. It is noncorrosive and termite-proof, and will not rot. The siding surface is generally covered with a baked enamel-paint finish that can stand up for many years before it fades, becomes dull, and needs a coat of paint. If the siding is scratched, bare aluminum is exposed. However, since the aluminum does not corrode, the scratch is only of cosmetic concern and can easily be corrected with touch-up paint. One problem with the siding is that it can be dented if struck hard enough—as with a baseball or stone thrown from a power mower. Many communities require that aluminum siding be grounded electrically as a precaution against electrical shock.

Aluminum siding is available with or without insulation backer boards. The insulation is generally a rigid foam (such as polystyrene) or fiberboard. Although the backer boards are only about ⅜ inch thick, they are quite effective as an insulator for a house that has no insulation in the exterior walls. Because of increasing energy costs, even a house with insulation in the exterior walls benefits from the additional insulation. The backer boards reduce heat loss in the winter and heat gain in the summer. They also increase the strength and rigidity of the siding. However, insulation-backed siding (and tight siding jobs) can cause moisture to accumulate within the exterior walls of houses that have no vapor barriers on the inside surface. (Insulation and vapor barriers are discussed in chapter 19.) You can usually tell whether the aluminum siding has an insulation backer board by pressing on it. If the siding is relatively firm, it has a backer board, but if it yields and bends under the pressure, it has no insulation board. Another method is to tap the siding. If it has no insulation, you will get a hollow sound.

When inspecting aluminum siding, look for loose, missing, and dented sections. Check the exterior joints for open sections and weathertightness. In those areas where electrical grounding of the siding is required, you should look for an electrical ground connection, a wire that runs from the siding to the inlet water pipe or a rod or pipe that has been driven into the ground. (See chapter 12.) You can find out whether an electrical ground connection is required by checking with the local municipal building department.

Vinyl siding Vinyl siding is very much like aluminum siding in size, shape, application, and appearance. Quite often close examination of the siding is needed to tell the difference between the two. The coloring in vinyl siding is embedded in the material and is the same throughout its thickness. Since the coloring in aluminum siding is only on the surface, an end cut or scratch in the aluminum reveals the silvery color of the bare metal. To tell whether the siding is vinyl or aluminum, look at an end cut or joint.

Vinyl siding is usually installed with an insulation backer board behind each sheet. In addition to insulation value, the board adds rigidity and strength. Vinyl siding normally does not dent from impact; it merely flexes and springs back to its original shape. However, during very cold weather, the siding becomes brittle, and a hard blow could crack or shatter it. When inspecting vinyl siding, check for cracked and broken sections and loose and sagging sections with open joints.

Vinyl siding expands and contracts as the temperature changes. When the siding is improperly nailed, this movement usually results in waviness and blisters in the vinyl panels. If you see this type of unevenness in the vinyl panels, record the fact on your worksheets for future correction.

Asbestos-cement shingles As with roofing shingles, asbestos-cement siding shingles were manufactured by combining asbestos fibers with Portland cement under pressure. These shingles are currently called *mineral-fiber shingles*. Although the shingles are no longer manufactured, they can be found on many homes in a variety of textures and colors. Since the shingles are unaffected by the weather and are immune to rot and termite activity, they require very little maintenance. However, the shingles are brittle and can be damaged and cracked by impact. The lower courses of the shingles are most vulnerable to damage. Usually damaged shingles are replaced rather than repaired. Depending on the weathered condition of the existing shingles, the new shingles may not match the older ones. (See FIG. 5-4.) When inspecting asbestos shingle siding, look for cracked, chipped, broken, loose, and missing shingles. Shingles that have slipped out of place were usually improperly nailed, or the nails were not rust-resistant and deteriorated. If the condition is caused by the latter, additional shingles will slip out of place in the future, and maintenance should be anticipated.

Asbestos-cement shingles were generally installed with sheathing-paper-backer strips behind the vertical shingle joints. These backer strips provided additional protection against water penetration. You might find sections of backer strips that have slipped out of place or are hanging loose between the shingles. Since the shingles were normally installed over sheathing paper, which is waterproof, replacing the loose backer strips is usually not necessary.

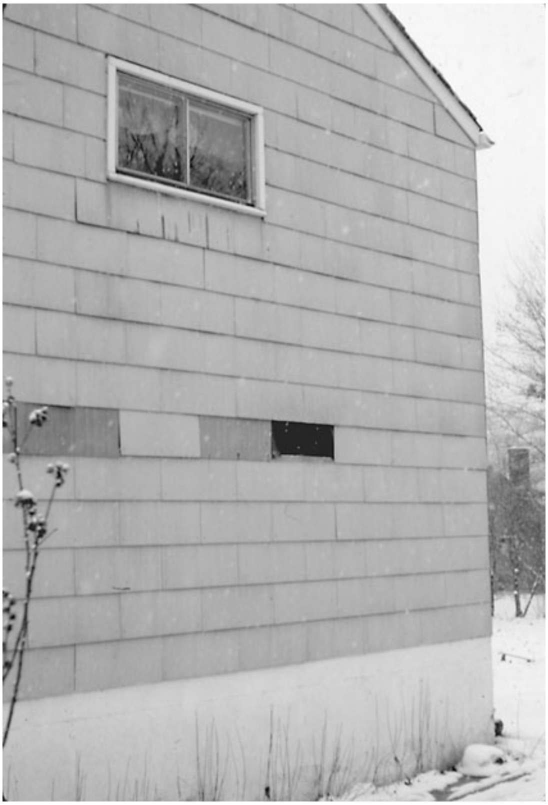

Fig. 5-4. *Missing and patched (mismatched) asbestos-cement shingles.*

Fiber-cement siding Fiber-cement siding is made from a mixture of cement, sand, cellulose fibers, and other components such as water, waxes, and resins. The material then goes through a multistep high-pressure curing process in which the mixture is subjected to highly pressurized steam that presses the material into a variety of siding patterns. There are several companies that manufacture fiber-cement siding; the most popular is James Hardie Building Products. Hardie fiber-cement siding products that you will find on houses are HardiePlank lap siding, shingles, and vertical siding.

Fiber-cement siding is popular because it is fire-resistant and will not rot, warp, or delaminate. It is also resistant to termites, and because it is normally manufactured into a much thicker and more durable product than its vinyl counterparts (according to Hardie, its siding is 5 times thicker than vinyl siding and weighs about 1.5 times as much as wood), it resists cracking and impact damage. The main concern with fiber-cement siding is not the product, but improper installation, which can result in skewed trim, exposed nails, and cracked, chipped, loose, or damaged sections. If you see any problems, record them on your worksheets for future correction.

Asphalt siding Asphalt siding is made by impregnating an organic felt material or glass-fiber mat with asphalt. The siding is available as shingles or as a roll. The exterior surface of the roll material is coated and embossed so that from a distance it looks like bricks. As the siding ages, it becomes dry and brittle, and cracks easily. For the most part, asphalt materials are no longer used for siding or residing residential structures. However, they can be found on existing buildings. When inspecting asphalt siding look for cracked, chipped, and eroded sections. Also check for open and lifting joints and loose, torn, and missing sections. If you find any areas in need of repair or replacement, record them on your worksheet.

Stucco A stucco finish on an exterior wall is basically a concrete sheet that has been built up in layers. It is usually made from a mixture of cement, lime, sand, and water. Stucco is weather-resistant, immune to termite and fungus attack, rigid, and durable—qualities that are desirable for an exterior wall finish. In addition, it can be applied to curved or irregularly shaped surfaces and to wood-frame walls that have been prepared with backing (sheathing) paper and metal lath. The backing paper is needed to resist water penetration through open joints or cracks that might develop in the stucco. The metal lath provides the means for bonding the stucco mix.

Stucco is generally applied in two coats on a masonry wall and three coats on a wood-frame wall. The minimum thickness for a three-coat wall is ⅞ inch; for a two-coat wall, ⅝ inch. The top layer of stucco is the finish coat and can be relatively smooth or have a rough texture. It can be prepared in a wide range of colors or painted.

Because stucco is a rigid material, cracks can develop as a result of a slight movement of the house. (See FIG. 5-5.) Movement occurs from foundation settlement and from wind forces. You generally find more cracks in stucco on a wood-frame house than on a solid masonry house. Shrinkage of the wood-framing members creates stresses in the stucco that often result in cracks. Once a crack develops, water can penetrate into the wall during a driving rain and can cause problems. In time, the portion of the metal lath around the crack rusts and deteriorates, and, depending on the condition of the backing paper, the wood framing and sheathing might rot. In addition, in cold climates, accumulated water behind the stucco can freeze, causing further deterioration as a result of frost action.

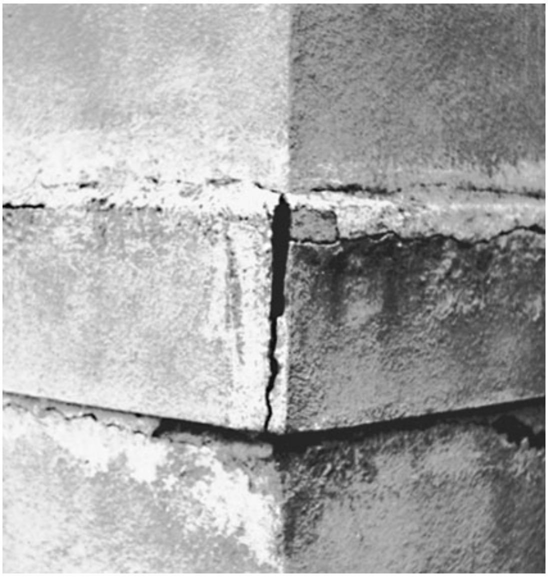

Fig. 5-5. *Cracked stucco wall. If wall was covered with vines, these cracks would be concealed.*

All cracks should be sealed. Hairline cracks and cracks up to 1⁄16 inch generally can be sealed by coating them with a cement-based paint. The only difficulty is in matching the color of the wall. Larger cracks can be sealed by filling them with a mortar mix. Broken and loose sections of stucco must be rehabilitated by a skilled craftsman. When inspecting a stucco wall, look for chipped, cracked, loose, and broken sections. If you find areas in need of repair, record their location on your worksheet. Stucco does not require painting. But if it has been painted, check the condition of the paint. Once a stucco wall has been painted, periodic repainting will be required for cosmetic purposes, although at less frequent intervals than wood.

Synthetic stucco Synthetic stucco, commonly known as an exterior insulation and finish system (EIFS), is an exterior wall siding that consists of four primary components:

- Foam insulation boards attached to the exterior wall sheathing
- A base coat that is applied to the insulation board
- A fiberglass reinforcing mesh embedded in the base coat
- A finish coat applied over the fiber glass mesh

Although the EIFS may resemble stucco in appearance, the two siding systems are quite different, especially with regard to controlling water intrusion. The traditional stucco system anticipates eventual water penetration through open joints and cracks. To prevent rot and deterioration of the wood framing and sheathing in the exterior walls, it uses a housewrap or building paper behind the stucco surface to carry any water that accumulates in that area down and out the bottom of the wall.

The EIFS, on the other hand, was not designed for water intrusion. It was considered a surface barrier system because it resisted water penetration at its outer surface. It was assumed that moisture would not penetrate the surface and reach the wall sheathing and/or framing. The system was not originally intended to drain water that got behind the EIFS cladding. In practice, however, water did penetrate the wall, not through the surface but through jambs and sills of window frames, and at the joints between the exterior walls and door, window, deck, and roof intersections. Water that penetrated the wall could not easily escape. It was trapped between the EIFS cladding and the sheathing.

Over a period of time the water was absorbed by the sheathing and framing, which increased their moisture content to a level above saturation, causing rot and structural damage. Depending on weather conditions and the quality of construction, significant damage due to moisture intrusion could occur. Damage from water intrusion has been found in the exterior walls of houses that are only 3 to 5 years old.

The industry, recognizing the problems that resulted because of the lack of drainage in the wall, modified the system to include the installation of drainage channels and building paper between the foam insulation and the sheathing. In appearance the new system is similar to the old one, and with a visual inspection you cannot tell the difference between the drainable and nondrainable EIFS. Not only is the location of water entry often difficult to see, but moisture damage to the sheathing and framing behind the exterior wall cladding cannot be readily detected by visual inspection.

If the exterior wall cladding is an EIFS, you can do a preliminary inspection by checking for cracked and open joints at the interfaces between the EIFS and dissimilar materials such as windows, doors, and wall penetrations. If any are found that require caulking, record their location on your worksheet. Also, if you have a noninvasive moisture meter, check those areas for water intrusion. However, because of the nature of the potential problems and the high cost for correction, it is recommended that you have the exterior walls inspected and evaluated by a professional.

Veneer wall A veneer wall is a wood-frame wall with an attached masonry facing. Unlike exterior siding, which is held in position by being fastened to the sheathing or studs, the masonry rests on top of the foundation wall and supports its own weight. It is attached to the wood backing by corrosion-resistant metal ties. The ties are considered the weakest point in this type of construction. If the ties have deteriorated or are not properly attached, the masonry facing can pull away from the wood frame.

The masonry—usually clay brick, concrete brick, or split stone—is normally positioned so that there is a 1-inch air space between the veneer wall and the wood backing. (See FIG. 5-6.) Small holes (*weep holes*) are usually installed at the base of a veneer wall. These holes allow water that might accumulate in the air space to drain. When the masonry facing is brick, the weep holes are generally formed by eliminating the mortar in a vertical joint.

Some bricks absorb more water than others. Bricks with high absorption should not be used for the exterior facing of walls, especially in colder climates. Unfortunately, they are used occasionally. Alternate freezing and thawing of bricks that have absorbed water will cause them to deteriorate. Depending on how much water has been absorbed, the interior walls might become damp, a condition that can usually be detected during the interior inspection. This condition can often be controlled by coating the bricks with a silicone sealant.

When inspecting a veneer wall, look for chipped, cracked, loose, deteriorating, and missing bricks or stones. In addition, check for cracked, chipped, and deteriorating mortar joints. Pay particular attention to the mortar joints. Occasionally, because of excessive shrinkage of the wood framing or slight foundation settlement, you might find large open cracks, especially around window frames and doorframes. Also look for loose and bulging sections of veneer wall. If you find any of the above items, record their location on your worksheet for later correction.

Masonry wall

Unlike a wood-frame wall where the structural support and weather barrier are provided by two separate components, studs and siding, the masonry units in a masonry wall (clay tile, brick, stone, or concrete block) provide both the support and the weather protection. Because of the low thermal resistance of masonry, a masonry wall allows greater heat loss than a wood-frame wall.

To reduce the heat loss, insulation can be added by applying a rigid foam insulation board to the interior side. In addition to providing insulation, the board can also be

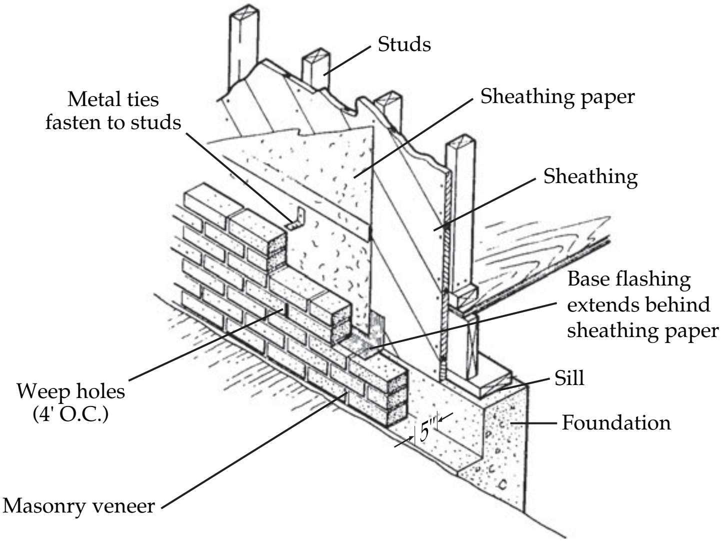

Fig. 5-6. *Components of a brick-veneered exterior wall.*

used as a base for plastering. Another approach is to apply furring strips to the inside wall; the furring strips create an airspace into which insulation can be placed prior to installing the finishing wall panel. In some cases, however, the interior side of the masonry wall is left completely exposed and serves as a decorative element or a base for direct plastering. This is quite wasteful from an energy-conservation point of view.

Because of the rigidity of masonry walls, differential movement within the wall might cause serious cracking. Wall movement might be the result of unequal foundation settlement, or expansion and contraction from temperature and humidity changes. Many cracks are not of structural concern, although they should be sealed to eliminate the possibility of water penetration. If you have any doubt about the severity of a crack, have the condition checked by a professional.

A common problem with masonry walls is *efflorescence* on the exterior surface. Efflorescence is a deposit of soluble salts that were originally within the masonry, usually brought to the surface by water in the wall. When the water evaporates, the salts are deposited on the surface. Efflorescence generally can be removed by scrubbing with a stiff brush or washing with a dilute solution of muriatic acid. However, if the condition is a recurring problem, it is an indication that water is penetrating the wall through cracks or faulty joints or flashing.

When inspecting a masonry wall, pay particular attention to the joints around window and doorframes. All joints should be weathertight. Are there any cracks around the corners of window or door openings? These are areas of high stress concentration and are vulnerable to cracking. Cracked and chipped mortar joints and deteriorated masonry should be indicated on your worksheet for later repair. If you notice bulging sections in the exterior walls or large cracked sections, have the condition checked professionally, since it might indicate structural problems.

Trim

All portions of the exterior finish, other than the wall covering, are generally classified as exterior trim. This includes the moldings and sills around windows and doors, fascia boards, soffits, louvers, shutters, and decorative columns. (See FIG. 5-7.) Trim does not serve a structural function. It is used as finishing around openings and to protect joints, edges, and ends. Most exterior trim is made of wood or wood products, although aluminum and vinyl trim have become quite popular. Many older, traditionally designed homes have decorative sheet-metal cornices, which are considered part of the trim. The problem with sheet-metal trim is that if it is not maintained and kept adequately painted, it will rust and deteriorate.

Wood trim that is exposed to the weather should be decay-resistant so that it does not rot. (See the section on rot in chapter 8.) Some types of preformed trim are factory-treated with a water-repellent preservative to make them water- and decay-resistant. When the trim is cut to size during construction, the ends or miter joints must be treated to make them water-resistant. All too often they are not treated, and the joints, which readily absorb water, begin to rot. When inspecting wood trim, pay particular attention to the joints that are vulnerable to decay. A house with a wide roof overhang at the eaves and gables provides greater weather protection of the sidewalls and trim than one with no roof projection beyond the walls. All nontreated wood continually exposed to moisture is prone to decay. The trim around the edge of the roof is particularly vulnerable. Although the Asphalt Roofing Manufacturers Association recommends the installation of a metal drip edge along the eaves of a roof deck,

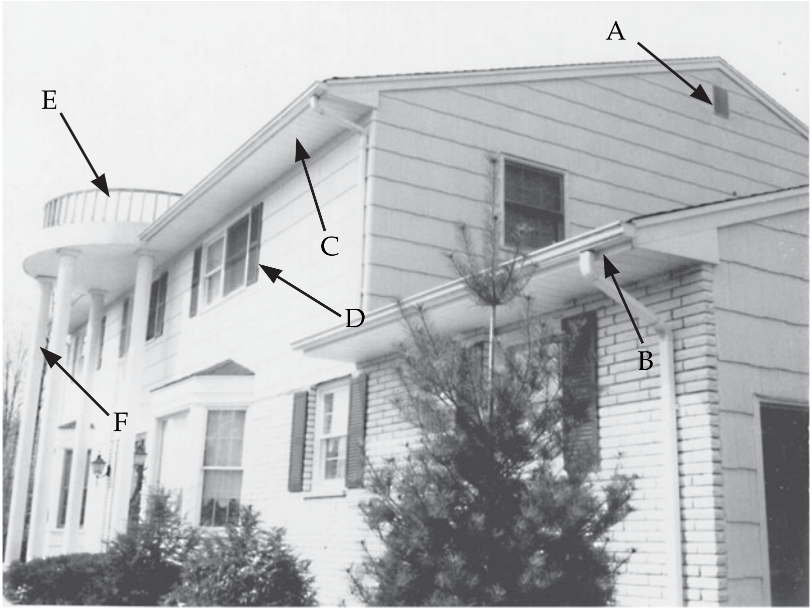

Fig. 5-7. *Exterior trim on a house: A—gable louvers, B—fascia, C— soffit, D—shutters, E— widow's walk balustrade, F—decorative columns.*

in practice it is often omitted. The drip edge is designed to allow water runoff to drip free of the underlying trim. Without it, water tends to curl back under the shingles, wetting the edge of the roof sheathing and trim.

Wood trim should be inspected for cracked, loose, missing, and rotting sections. If the trim is painted, is the paint peeling and flaking in sections? Does the trim need repainting for weather protection? In older Tudor-style houses with timbers embedded in the stucco siding, inspect the timbers at the stucco joints for decay, especially if the joint is horizontal. Over the years, the joints tend to open slightly, allowing water to penetrate. With nonwood trim, check for loose, missing, and deteriorated sections.

Windows

The windows should be checked during the exterior and the interior inspections. The overall condition of the windows should be checked during the exterior inspection; the operation of the windows should be checked during the interior inspection. (See the section on windows in chapter 10.)

Many types of windows are used in residential structures. The most common types, as shown in FIG. 5-8, are double-hung, horizontal sliding, casement, awning, jalousie, and fixed-pane.

The *double-hung window* is the most common window unit in older and newer homes. It consists of upper and lower sashes that slide vertically past each other. The sashes are held in a fixed position within the window frame by a friction fit, counterweights, or spring balances. In some windows, the sashes are removable for ease of maintenance such as cleaning or painting. One variation of the double-hung window is the single-hung window; the upper sash is fixed, and the lower sash is movable.

The sashes in a *horizontal sliding window* slide horizontally on separate tracks. The most

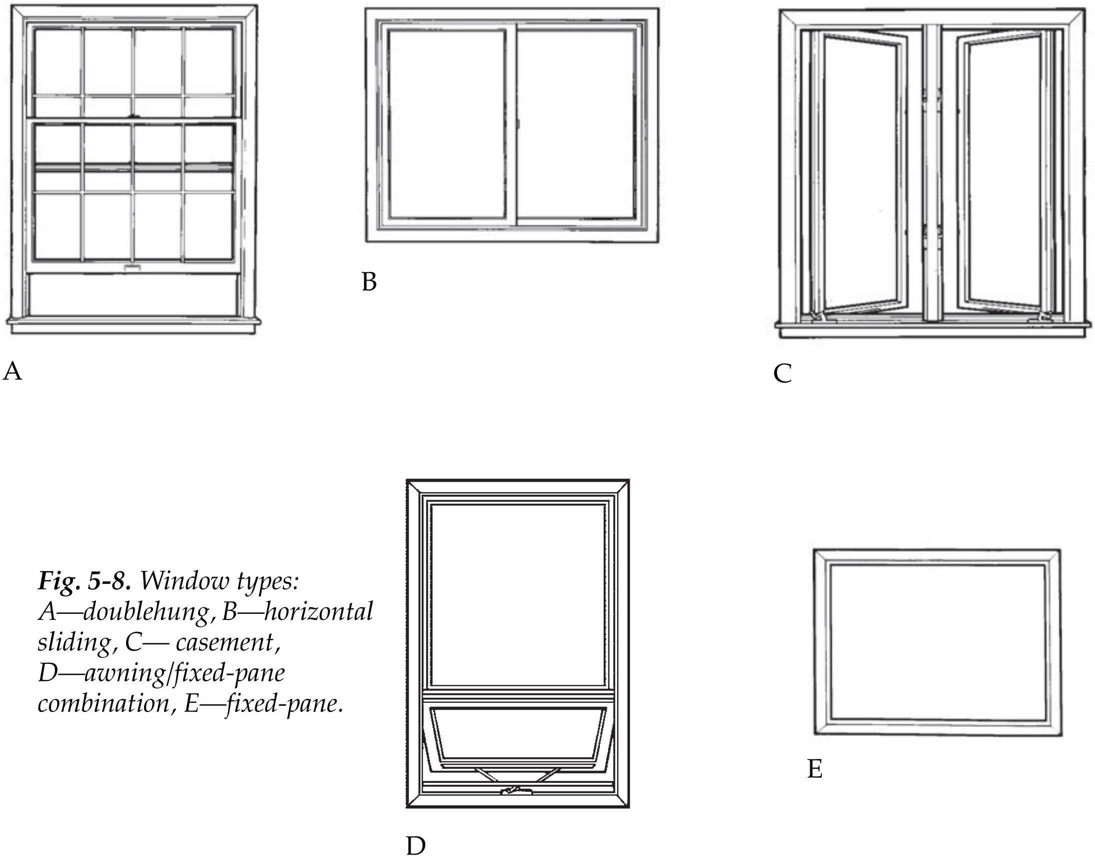

Fig. 5-8. *Window types: A—doublehung, B—horizontal sliding, C— casement, D—awning/fixed-pane combination, E—fixed-pane.*

common design consists of two sashes, both of which are movable. However, sometimes one sash is fixed. In most of these windows, the sash can be removed for cleaning.

Casement windows consist of two or more sashes hinged at the side and mounted so that they swing outward. The sashes are opened and closed by a cranking mechanism, a push bar mounted on the frame, or a handle fastened to the sash. Because the sash opens outward, a storm sash or screen must be attached to the inside of the window.

Awning windows have one or more sashes hinged at the top and mounted so that they swing out at the bottom. They are opened by push bars or cranking mechanisms similar to those on casement windows. As with casement windows, screens and storm sashes are mounted on the inside.

Jalousie windows are basically adjustable louvers. The louvers are glass slats (several inches wide) held by an aluminum frame at each end. The frames are interconnected by levers so that the slats open and close in unison, like venetian blinds. Jalousie windows are crank-operated and provide good ventilation. However, because of the many glass slats, they are troublesome to wash and are not weathertight. Even with a storm sash, cold-air leakage occurs around the windows. In northern climates, jalousies are usually limited to use on porches and breezeways.

Unlike the preceding windows, which are movable and provide ventilation, *fixed-pane windows* are stationary and are used only to provide daylight and outdoor views. They can be used alone or in combination with sliding, double-hung, or swinging windows to achieve a custom design.

With the exception of the jalousie windows, available only in aluminum, all of the windows are available in wood, metal (steel or aluminum), and vinyl-clad frames. If you are not sure whether the metal window frames are steel or aluminum, you can always tell the difference by using a magnet. A magnet sticks to a steel frame but not to an aluminum one.

Metal frames get quite cold during the winter months. Consequently, some of the water vapor in the air inside the house tends to condense on the frames. To reduce this condensation, several manufacturers produce metal-framed windows with a thermal barrier that prevents the outside frame from touching the inside frame. This substantially reduces the condensation but does not eliminate it. Since wood has a greater thermal resistance than metal, wood window frames normally do not get cold enough for condensation to form on their inside surface.

Although wood windows provide a better insulation than metal windows, they do have a tendency to swell or shrink with changes in moisture. A wood sash that absorbs moisture will expand and bind in the frame, so that it does not operate freely. Wood windows should be treated (by the manufacturer) with a water-repellent preservative to resist decay and moisture absorption. In the past, not all windows were so treated; thus rotting sections and binding sashes are occasionally found.

The glass used in windows must be of sufficiently high quality to minimize distortion. Years ago, because of the difficulty of manufacturing large distortion-free panes, window glass was available only in small sheets. Consequently, to fill a large opening such as a window sash, small panes of glass were used and held in position by framing strips called *muntins*. Today, even though large windowpanes are available, muntins are still used to create a special architectural effect. Some window manufacturers provide preassembled wood or plastic dividers that simulate muntins. The dividers merely overlay the large windowpane and snap in and out of the sash. Some companies also install muntins between the panes of double-glazed windows during the manufacturing process.

Windowpanes are also available with single and double glazing. A double-glazed window, a thermal (or insulated) pane, reduces heat loss or gain through an equivalently sized single-pane window by about 50 percent. It also reduces condensation on the inside surface. With some windows, triple glazing is also available. See the section on windows in chapter 10 for a discussion of insulated windows and associated problems.

Inspection

When inspecting windows, look for cracked, broken, and missing panes. Check the joints between the sash and the window frames to see whether they are filled with paint. If they are, the windows might not open, and minor maintenance will be needed. The condition of the joints between the glass pane and the sash should be inspected. Are any panes loose? Most windowpanes are secured to the sash by bedding and sealing with a glazing compound such as putty. If putty is used, are any sections cracked, loose, chipped, or missing? Some panes are secured to the sash by wood strips (trim). Check the strips for cracked, loose, and missing sections. Note any of the above items on your worksheet for later correction.

Look at the overall condition of the window frames and sash. Windows exposed to the elements are vulnerable to weathering deterioration. Are any wood sections cracked and rotted? Steel windows can rust. Depending on the extent of the deterioration, some windows require simple repair; others, replacement. All the windows generally are not visible or accessible during your exterior inspection, especially the windows on upper levels. The exteriors of those windows, and their operation, should be checked during your interior inspection.

Exterior doors

The two basic types of exterior wood doors are flush doors and stile-and-rail doors. Flush doors are made by bonding face panels to solid or hollow cores. Stile-and-rail doors (also referred to as *panel* doors) are solid doors that consist of vertical and horizontal members (called *stiles* and *rails*, respectively) that enclose wood or glass inserts. (See FIG. 5-9.)

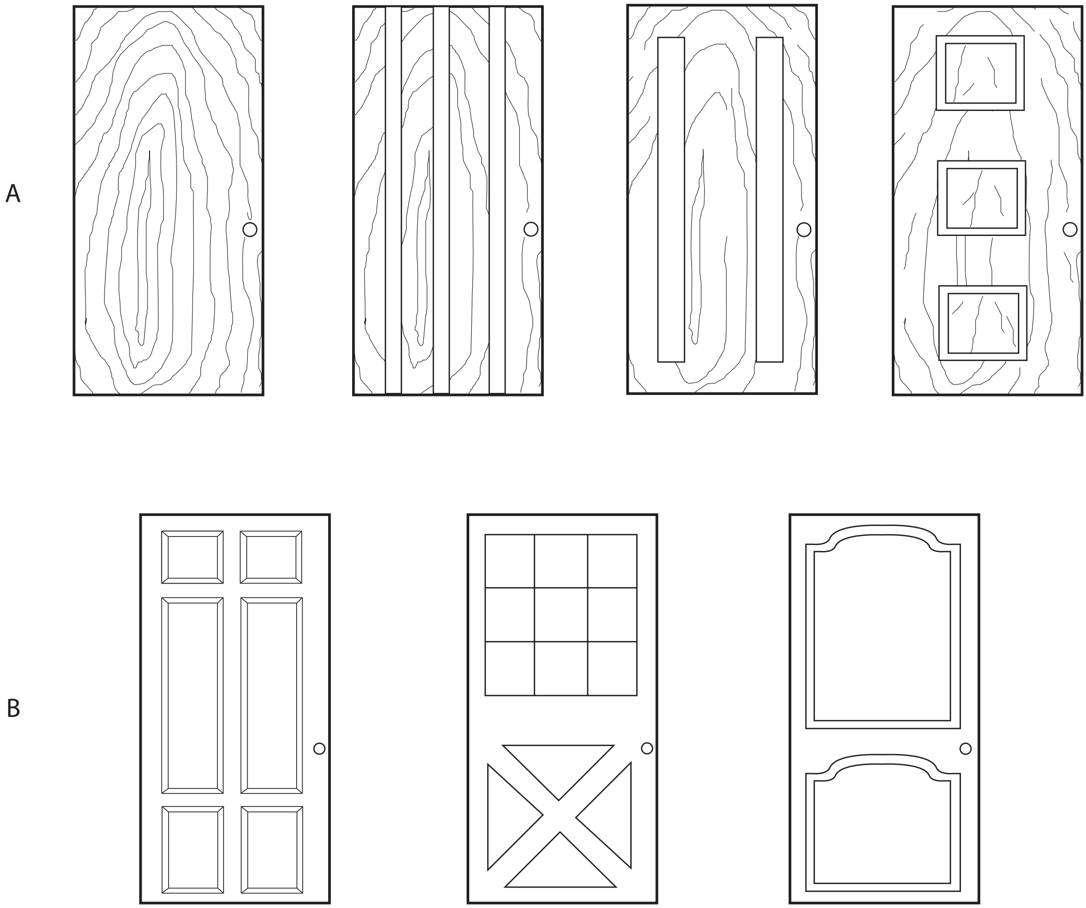

Fig. 5-9. *Exterior doors.*

Most *flush doors* have hardwood veneer face panels, although hardboard and softwood panels are also available. Some doors are made with cutouts for windows or louvers. Flush doors are used as interior and exterior doors. As exterior doors, they should be made with waterproof adhesives rather than water-resistant adhesives. You can often tell whether the proper adhesive was used by looking at the top edge. In many quality doors, you will find a small red plastic plug in the edge. This indicates that the door was bonded with a waterproof adhesive that is suitable for exterior use. When an improper adhesive is used, the exterior face panel eventually begins to delaminate and peel.

A solid-core flush door provides greater heat and sound insulation, fire resistance, and dimensional stability than a hollow-core door. The solid core of a flush door might be made of wood blocks or a composition material that has been formed into a rigid slab. A hollow core is generally made of wood or wood derivatives (cardboard) that have been formed into a honeycomb or parallel strips. When resistance to heat, sound, and fire are not important factors, the hollow-core flush door is sometimes used as an exterior door. This is not considered quality construction. When used on the exterior, the door should be treated with a water-repellent preservative and bonded with a waterproof adhesive. Unfortunately, this is not always the case.

One variation of the flush door is the metal-clad insulated entrance door. This door is available in a selection of surface styles and is becoming more popular. Basically the door consists of metal face panels with an insulating core. Some doors are provided with a thermal break to separate the interior parts of the frame and door panel from the exterior parts, thereby minimizing condensation during the winter months. Because of the insulating characteristics of this type of door, the need for a storm door is essentially eliminated.

Although *stile-and-rail doors* are not used as commonly as flush doors, they are available in a greater variety of designs. In addition to wood, the stile-and-rail doors are available in steel and fiberglass. Because of the number of panels and joints, the stile-and-rail door is not as effective an insulator as a flush door. Also, one or more panels in a wood door might crack as a result of shrinkage. If you stand inside of the house looking at the door, cracked sections are very noticeable. Daylight is visible through the cracks.

Some exterior doors are not considered secure because of the location of the glass panes relative to the door lock. If the door lock can be reached by breaking a glass pane in the door or in side panels, an auxiliary lock is recommended. This lock should be positioned where it is not accessible from the outside. Since you normally do not know whether you were given all the keys to the exterior door locks, it is recommended that after you take possession of the house, you replace all of the door locks or at least have the locks rekeyed.

The condition of the exterior doors should be checked during your exterior inspection, and the operation of the doors should be checked during your interior inspection. Outside, inspect a wood door for cracked, chipped, and delaminating sections. Inside, check the doors for cracks (visible daylight) and ease of operation. Check a steel door for dents and scratches, which will rust if not painted. A fiberglass door can crack under severe impact and should be checked. Does the door open and close easily, or does it bind? Also check for weatherstripping around the exterior joints. Weatherstripping is desirable, since it minimizes air infiltration. If you find any problems with the doors, record them on your worksheet.

Storm windows, screens, and storm doors

Storm windows

Windows are a major source of heat loss. On a per-square-foot basis, more heat is lost through windows than any other area. In fact, the heat lost through a single-pane window is approximately fourteen times greater than that lost through a well-insulated wall of comparable size. However, with storm windows, the heat loss can be reduced by about 50 percent. There are three basic types of storm windows: storm sashes, storm panels, and combination units.

A *storm sash* is a removable sash, usually made of wood, containing a fixed-pane window. The storm sash fits over the window on the outside or inside, depending on how the window opens. Storm sashes are most commonly found on the outside of older double-hung windows. They are not desirable for year-round use because they cannot be opened to admit breezes during the warmer months. Storm windows are generally taken off in early spring and reinstalled in late fall, a task that can be somewhat awkward and time-consuming.

A *storm panel* looks like a storm sash, but it is usually mounted in a narrow metal frame and is attached to the movable window sash rather than being fitted over the entire window opening. Since storm panels are attached directly to the movable sash, they do not interfere with the operation of the window and need not be removed during the warmer months.

Combination units refer to storm and screen sashes combined in a single frame. The unit is mounted over the outside of the window and is therefore effective in reducing air infiltration around the window joints in addition to reducing heat loss. Combination storm and screen windows are available in two- and three-track units. With two-track units, the outside track contains the storm sash in the upper half and a screen sash in the lower half. With the screen in position, the upper storm sash cannot move. The inside track contains the lower storm sash, which can slide up and down. In a triple-track unit, there is a separate track for each of the two storm panes and for the screen. Combination storm and screen units are generally found on double-hung and horizontal sliding windows. Since these units do not interfere with the operation of the movable sash and can also be opened to provide ventilation, they are not normally removed once installed.

Combination units are available in aluminum, steel, or solid vinyl. The aluminum frame with a baked enamel finish looks like a vinyl frame. You can tell the difference by looking at an edge or joint. The aluminum edge will have a silvery color; the vinyl edge will be the same color as the frame. Over the years, many aluminum frames with a mill finish (plain aluminum) show the effects of weathering, such as pitting, corrosion, and a degraded appearance. An anodized or baked enamel finish will offer greater protection against weathering. Combination storm and screen windows made of steel have a tendency to rust and require periodic painting.

Screens

Most window screens for residential structures are either mounted on a wood sash or rimmed with a metal or plastic frame. The wood-framed screen is usually used in conjunction with a storm sash; the metal- or plastic-framed screen is used in the combination storm and screen unit. Metal- or plastic-framed screens are also used for casement and awning-type windows. However, for these windows, the screen must be equipped with panels that provide access to the cranks or push bars.

Another type of screen (not very common) is the roll-up screen. This screen is mounted on the inside of the window and is similar in operation to a roll-up shade. When the screen is not being used, it can be rolled up and hidden from view. The sides of the screen move in metal tracks to prevent insects from flying in around the edges. As the screen ages, the joint between the screen and the track tends to open and become less effective.

Storm doors

In colder climates, unless the exterior doors are insulated, they are often used with storm doors to reduce heat loss and cold-air infiltration around the joints. Storm doors are generally lightly constructed wood or metal stile-and-rail-type–doors with a glass-panel insert. On many of these doors, the glass insert is interchangeable with a screen panel so that they can function as both storm and screen doors. Because storm doors are constantly being opened, they are not as effective as storm windows in reducing heat loss. Nevertheless, they are effective from an overall energy-conservation point of view.

Inspection

When inspecting the house, look for storm windows. If you do not see any or the inspection is being performed during warmer weather when storm sashes are normally not installed, ask the owner whether the house has storm sashes for all the fixed and movable windows. If no storm windows or only a few are available, record the fact on your worksheet. Installing a complete set of storm windows can be quite costly.

Storm windows should be inspected for cracked, broken, and missing panes. When the storm sash is wood, check for cracked, broken, and rotting sections. On combination units, look specifically at the corner joints. These joints should be tight so that no cold air leaks into the unit. Check the overall condition of the frames. Are any sections loose, broken, rusted, or corroded? Are any of the panes loose in the sash? Look for torn sections and holes. In metal- and plastic-rimmed screens, the screening is normally held in position by a spline that has been forced into a groove around the frame. Periodically, I find splines hanging loose within the frame of the combination unit. Loose splines are an indication that the joints in the associated screens must be resecured.

Storm doors should be checked for ease of operation and overall condition. Are sections cracked, broken, loose, rotted, or corroded? Is the glass panel loose, cracked, broken, or missing? If you find any problems with the storm windows, screens, or storm doors, record them on your worksheet as a reminder for later correction.

Caulking

As you walk around the house inspecting the walls, windows, trim, and doors, look for cracked and open joints. All exterior joints should be caulked (sealed) so that they are watertight and airtight. If they are not adequately caulked, wind-driven rain can enter and cause wood members to rot, metal ties to rust, and masonry sections to crack and chip. In addition, cold air can infiltrate the house, resulting in higher heating costs. A vulnerable joint for cracking is one that joins two dissimilar materials; for example, the joint between a brick facing and a non-masonry sidewall. Dissimilar materials usually have different expansion and contraction characteristics, often causing the joints between them to crack and open. These joints should be sealed with a non-shrinking flexible caulking compound.

There are several types of caulking compounds. The four most popular types are oil base, acrylic latex base, butyl-rubber base, and silicone base. *Oil base* caulking compounds are the cheapest and will readily bond to most surfaces—wood, masonry, and metal. However, they are not very durable; they tend to dry and crack after a short time. Joints sealed with this type of caulk require periodic inspection and maintenance. *Acrylic latex* caulking compounds are medium-priced, durable, and flexible, and should last for many years. *Butyl-rubber* caulks are medium-priced, durable, and paintable; however, they exhibit high shrinkage, a characteristic that is acceptable when caulking narrow cracks and inside corners. *Silicone-base* caulks are the most expensive. They are very pliable and are good for sealing joints subject to movement. No one caulking compound is ideally suited for every application. However, based on cost, durability, and ease of application, acrylic-latex caulks are considered by many to be the best choice.

When inspecting the exterior joints, check the condition of the caulking. Look for cracked, chipped, crumbly, and missing caulking compound. The location of joints that need recaulking should be recorded on your worksheet. Recaulking is a relatively simple task and can be done after you move into the house.

Checkpoint summary

General considerations

- ❍ Inspect exterior walls for sagging, bulging sections, and for corners that are not vertical.
- ❍ Check for window frames and door frames that are not square.
- ❍ Structural problems whose cause cannot be determined should be evaluated by a professional.
- ❍ Note wall locations that have pipe or hood projections.
- ❍ Determine their usage (i.e., sump-pump discharge, condensate line, dryer vent, etc.).
- ❍ Check for vines growing up the exterior walls.

Exterior walls

Wood siding (shingles, shakes, boards, plywood panels, hardboard)

- ❍ Check bottom course of siding for sections in contact with, or in close proximity to, the ground (less than 8 inches).
- ❍ Check wood shingles/shakes for open joints, cracked, chipped, loose, or missing sections.
- ❍ Note areas of rot or discolorations.
- ❍ Check for peeling and flaking paint and warped shingles, particularly on sidewalls with a southerly or southwesterly exposure.
- ❍ Inspect for poor-quality shingles and shingles that have been improperly nailed.
- ❍ Check wood boards for open joints, cracked and rotting sections, loose or missing knots, peeling paint, and blistered sections.
- ❍ Inspect plywood panels for open joints, loose, warped, cracked, delaminated, or rotting sections.
- ❍ Check hardboard siding for cracked, deteriorated, or loose sections.

Aluminum/vinyl siding

- ❍ Check aluminum siding for loose, missing, torn, or dented sections.
- ❍ Check joints for open sections and weathertightness.
- ❍ Does siding contain insulation backer boards?
- ❍ Check siding for an electrical ground connection. (This requirement can be verified with the local building department.)
- ❍ Check vinyl siding for open joints, loose, cracked, or sagging sections.
- ❍ Check vinyl panels for waviness and blisters.

Asbestos-cement shingles/asphalt siding

- ❍ Check asbestos-cement shingles for loose or missing sections; cracked, chipped, and broken areas.
- ❍ Inspect asphalt siding for open or lifting joints; missing, loose, torn, cracked, chipped, or eroding sections.

Fiber-cement siding

- ❍ Check for skewed trim, exposed nails, and cracked, chipped, loose, or damaged sections.

Stucco-cement–finished walls

- ❍ Check for bulging, missing, loose, cracked, or chipped sections. Note areas in need of rehabilitation.
- ❍ If stucco is painted, check condition.

Synthetic stucco (EIFS)

- ❍ Check for cracked and open joints at the interface between the EIFS and windows, doors, wall penetrations, and so on.
- ❍ Are there indications of moisture in those areas?

Veneer and masonry walls

- ❍ Inspect for loose or bulging sections and large open cracks, particularly around door and window frames.
- ❍ Check for cracked, chipped, or missing sections of brick or stone.
- ❍ Inspect mortar joints for deterioration, cracked or loose sections.
- ❍ Check exterior surfaces on masonry walls for signs of water seepage (efflorescence).

Trim

- ❍ Check trim for cracked, loose, missing, or rotting sections.
- ❍ Inspect for areas of bare wood, blistered and peeling paint.
- ❍ Check nonwood trim for cracked, torn, missing, or loose sections.

Windows

- ❍ Check for cracked, broken, or missing panes.
- ❍ Are any of the windows painted shut?
- ❍ Are the panes properly secured to the sashes?
- ❍ Check the condition of the window frames and sashes.

Exterior doors

- ❍ Check wood door for cracked, chipped, or delaminating sections.
- ❍ Check metal door for dents or scratches.
- ❍ Check fiberglass door for cracks.
- ❍ Check for weatherstripping around exterior joints.

Storm windows, screens, and storm doors

- ❍ Check for missing units and/or partial installations.
- ❍ Inspect storm windows for loose, cracked, broken, or missing panes.
- ❍ Inspect wood units for cracked, broken, or rotting sections.
- ❍ Inspect combination units for loose, broken, rusting, or corroded sections.
- ❍ Inspect screens for torn sections and holes.
- ❍ Inspect doors for ease of operation; missing glass; cracked, broken, rotting, or corroded sections.

Caulking

- ❍ Check joints for cracked, chipped, crumbly, missing, or loose areas of caulking compound.

6

Lot and landscaping

Drainage 62
  Groundwater, water table 65
  Inspection 67
Retaining walls 70
  Inspection 71
Landscaping 74
  Lawn 74
  Shrubs 75
  Trees 76
Decks 76
  Free-standing decks 79
  Composite decks 79
Fences 80
Checkpoint summary 81

In addition to all of the items mentioned in the previous chapters, you should inspect the drainage around the house and the landscaping. Any retaining walls, decks, or fences also should be inspected.

Drainage

As housing developments and shopping centers sprout up in the countryside, they affect the drainage characteristics of the surrounding areas. Normally, in undeveloped areas most of the water falling to the earth soaks into the ground. The remainder flows over the surface into lakes, rivers, and streams, or accumulates in low-level areas, forming ponds. In built-up areas, thousands of acres that had been soaking up rain have been rendered impervious to water because of buildings and paved areas. The surface-water runoff in these areas might be two to ten times more than it was when the land was undeveloped.

In built-up areas, surface water usually flows into storm drains (catch basins) that in turn discharge into rivers and streams. In some cases established housing developments have been inundated with surface water after a heavy rain because their storm-drainage facilities were not adequate for the increased water flow resulting from adjacent new housing developments. In most cases, the increased runoff results in the rivers and streams swelling, although they are usually contained within their banks. However, after a heavy prolonged rain, a river or stream can overflow its banks and flood the surrounding area. Many people do not realize that even a small creek, which might be a trickle when they see it, can become a raging, destructive torrent following an excessively heavy rain.

The area normally flooded when a river or stream overflows its banks is called a *flood plain*. Between 5 and 10 percent of the land in the United States is in a flood plain. Much of this land is level and from outward appearances seems to be desirable. As land in urban and suburban areas became more and more scarce, builders constructed homes directly on the flood plains of streams and other waterways. These homes are all vulnerable to flooding. (See FIG. 6-1.) In many parts of the country, some owners do not realize that their homes were built in a flood-prone area, and they probably will not realize it until it is too late.

If you have doubts about whether the house is located in a flood plain, you should check with the local town or county engineer. If the engineer is not available, often the local highway superintendent can tell you whether the area periodically floods. In many communities, federal flood insurance is available for those homes located in a flood plain. If you are considering such a home, you should consider purchasing flood insurance.

Surface runoff is of concern to the homeowner because it can result in soil erosion, ponded water, and water in the basement or crawl space. Soil erosion occurs whenever water flows over bare earth. Soil particles are loosened and are carried away by the flow. Water seeks its own level and will therefore flow from a higher elevation to a lower elevation. The paths the water takes

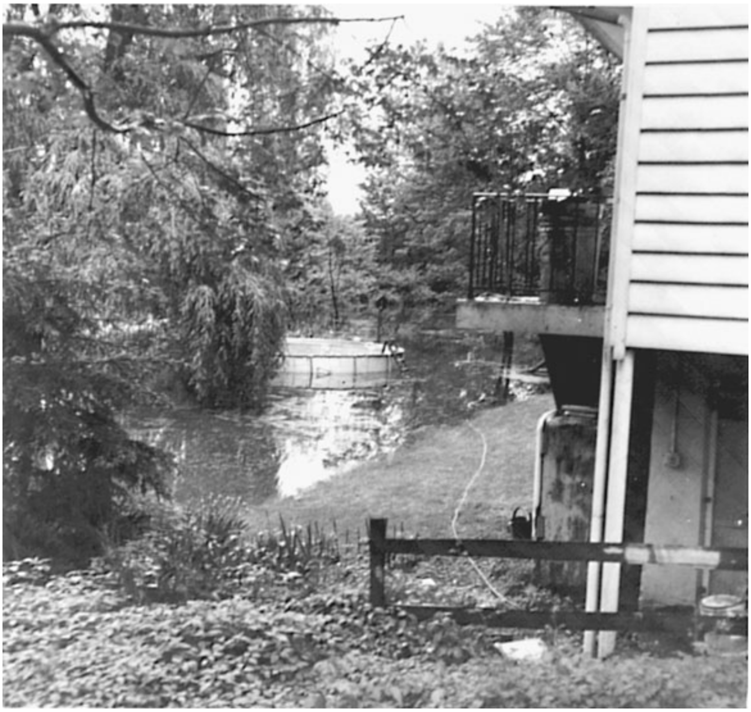

Fig. 6-1. *Backyard flooding of a house located in a flood plain.*

when flowing to a lower level are called *natural drainageways*. Areas particularly vulnerable to erosion are steep banks and drainageways. (See FIG. 6-2.)

The basic principle for preventing or minimizing erosion is to have the ground covered as much as possible with growing vegetation such as grass, trees, bushes, shrubs, and even weeds. If the vegetation does not root and keeps washing out, a substitute cover such as gravel, stones, or mulch can be used. This cover is not as effective, but it does reduce the erosion. Sometimes the banks are too steep for a ground cover of any kind; they must be stabilized by terracing or retaining walls.

In addition to a ground cover, erosion can be reduced by slowing down the water flow. For example, if there is a concentrated surface runoff along a natural drainageway, the water can be diverted to a man-made channel or ridge that follows a level contour. This spreads out the water and slows the flow so that the water does not scour and erode the soil.

The effects of surface runoff can be minimized by reshaping the ground surface. This can be done by terracing and/or regrading the lot into gentle slopes with diversionary ridges and swales. A *swale* is a depression in the ground that like a ridge will intercept surface runoff and redirect it to an area where the water will not cause damage. When a house is located on a sloping lot, a swale or ridge should be in the portion of the lot that slopes toward the house. This type of a diversion will prevent surface water from accumulating around the house. Otherwise, the surface runoff might seep into the basement or crawl space. (See chapter 11.)

Many building lots have low, level areas that will tend to accumulate water after a rain or from surface runoff. When the soil is slow-draining, as with clay and silt, the water will

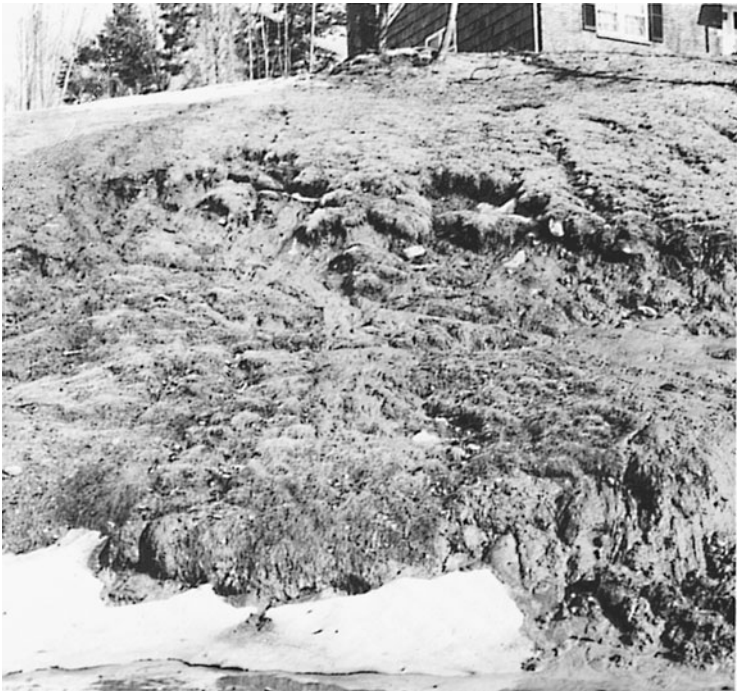

***Fig. 6-2.** Bank erosion on a steep slope. Correcting this condition will require a retaining wall with adequate drainage.*

pond rather than soak into the ground. The ponded areas retain the water until it evaporates or eventually seeps into the ground.

Depending on the location of the pond, the accumulated water might not be a problem. If the pond is over the leaching field of a septic system or is in an area that normally has a lot of foot traffic or is used by children, corrective action is necessary. The problem can often be corrected by bringing in fill and regrading the area. When regrading is not practical, the area can be drained by laying a line of perforated drainpipe through the affected area and directing one end of the pipe to a low spot. The pipe is generally encased in a bed of gravel or broken stones. If conditions warrant it, a concrete-block catch basin with radial spokes of perforated pipe can be installed at the low point. (See FIG. 6-3.) Water collected in the pipes and catch basin can then be directed to another area. If there are no other low areas to which the ponded water can be redirected using a pipe with gravity flow, the water can be directed to a sump pit and pumped to the desired location.

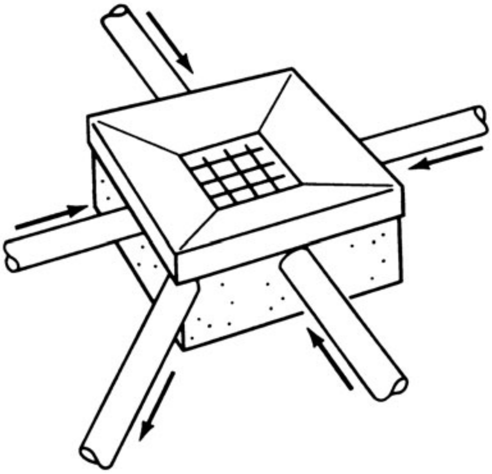

Fig. 6-3. *Concrete block catch basin with several perforated inlet pipes and a solid outlet pipe that drains to a suitable location. Basins of this type are used to drain low-level lawns that are collection ponds for rainwater.*

Groundwater, water table

Water that soaks into the ground eventually percolates downward under the influence of gravity until it reaches an impervious layer it cannot penetrate. After the water reaches the impervious layer, it begins to move in a lateral direction. This underground flow is known as *groundwater*, the top surface of which is commonly called the *water table*. The level of the water table will vary with the amount of rainfall. Consequently, the water table might be several feet higher after a prolonged rainy period than during a prolonged spell of dry weather.

A high water table can result in a flooded basement or failure of the septic system's leaching field. In many parts of the United States, the seasonal high water table is only 2 to 5 feet below the surface. In those areas, houses should be built on a slab or over a crawl space rather than over a full basement. Unfortunately, houses with full basements have been built in areas where the water table (during the wet season) is above the level of the basement floor. This invariably results in water penetration into the basement. Depending on the soil, even when the water table is several feet below the basement floor, some water might seep into the area as a result of the capillary rise of groundwater. (See chapter 11.)

Homes located in areas where the seasonal high water table is only a few feet below the ground surface should not have septic systems for waste disposal. Ideally, they should be connected to a sewer system. For proper operation of a septic system, the water table during the wet season should be at least 4 feet below the bottom of the leaching field or seepage pit. The operation of a septic system is explained in chapter 13.

Basically, the top surface of flowing streams, rivers, lakes, and oceans is the water table. Consequently, the terrain that gradually slopes into the waterway has a high water table. Homes built in these areas are not only vulnerable to problems associated with a high water table but are also vulnerable to flooding. The water table tends to follow the general contour of the land and in some areas might intersect the ground surface, thus forming marshy wetlands. To a homeowner, these wetlands are quite undesirable, because not only are they costly to drain, but they are also a breeding place for insects.

Excessive grading or reshaping of the ground surface (such as cutting out the side of a hill to locate a house, FIG. 6-4) can change the natural drainage patterns and cause groundwater to seep to the surface. I have inspected many such houses and have found (in the early spring or after an excessively heavy rain) water oozing out of the cut side of the hill. This water, if not redirected away from the house, can work its way into the basement or crawl space.

In areas with a seasonal high water table, if the topography is such that the land slopes toward one side of the house, in addition to surface-water runoff, subsurface water will flow toward the house. This water, if allowed to accumulate around the foundation, can seep into the basement or crawl space. This condition can usually be controlled by installing a *curtain drain* in the hillside parallel to the house to divert the water away from the house.

A curtain drain consists of a perforated drainpipe installed in a trench that is filled with gravel and covered with soil. The trench normally extends several feet beyond the house, with one end leading to a suitable disposal area. Incidentally, the perforations in the pipe should be facing downward, not upward as is popularly believed. As the subsurface water level rises, it enters the holes along the length of the pipe. Since water always takes the path of least resistance, once

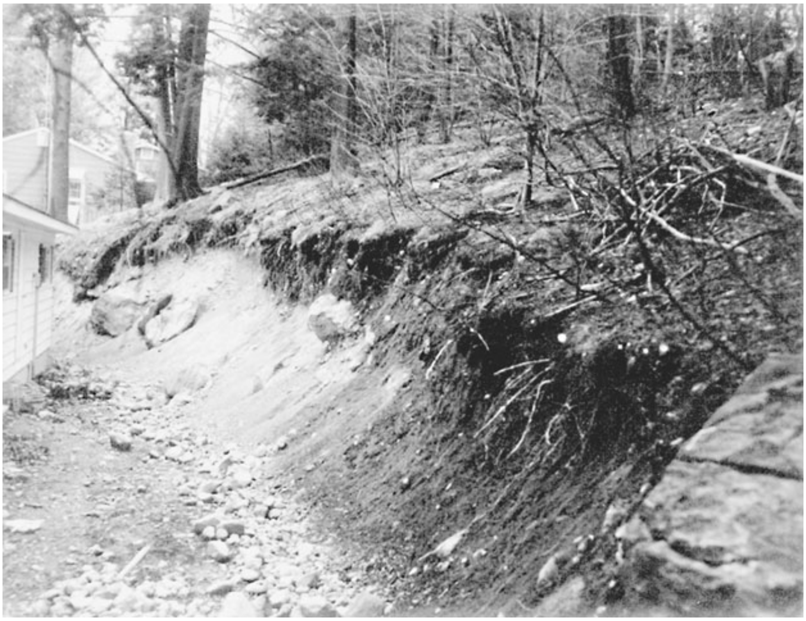

Fig. 6-4. *House located on the cutout side of a hill. The building site requires special provisions to minimize erosion and drainage problems.*

inside the pipe it flows to the outlet, which must be located away from the house and must be unobstructed. The outlet, however, should have an animal screen to prevent a small animal (such as a rabbit) from entering, becoming lodged, and blocking the flow.

In areas with a seasonal high water table or a potential for surface water to accumulate around the foundation, it is advisable to have foundation footing drains (perforated drainpipes) installed parallel and adjacent to the foundation footing. (See FIG. 6-5.) As with curtain drains, footing drains are installed with the holes facing downward. The purpose of the footing drain is to channel the water that accumulates around the foundation to another location. Footing and curtain drains either must have a free-flowing outlet or discharge into a sump pit where the accumulated water can be pumped to the desired location.

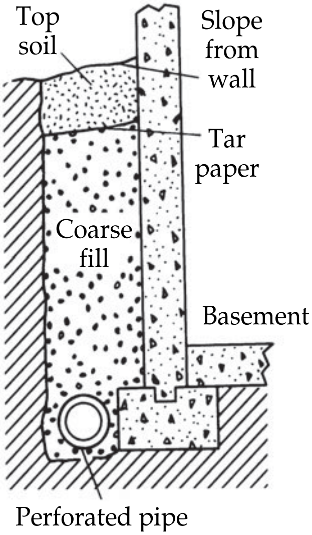

Fig. 6-5. *Foundation footing drain, used for channeling water that accumulates around the foundation to another location.*

Inspection

The drainage inspection should begin as you are driving to the house. When you approach the house, notice the overall topography. Is it level or inclined? If it is inclined, is it a gentle slope or a steep slope? With inclined topography, you should be concerned about the possibility of surface and subsurface water movement toward the house. If the house is located near the bottom of an inclined street, is a storm drain (catch basin) in the street at the low point? There should be, especially if the street is paved. Otherwise, after a rain or a snowmelt, water accumulates at the low area and depending on the amount, can flood the adjacent yards and driveways. Did you notice a waterway (stream, brook, etc.) on the street as you approached the house? If you did, the house might be located in a flood plain.

When you arrive at the house, notice whether the land between the house and the street is above or below the street level. If the land slopes downward from the street to the house, the house is vulnerable to drainage problems. The surface water, if not properly controlled, can accumulate around the foundation or can pond on the lawn or over the entry path. If the house is inspected when it is not raining, you might not see any problems. However, based on the slope and overall grading of the land around the house, you can at least determine the potential for a problem.

As you walk around the house, notice whether the ground immediately adjacent is graded so that it slopes away from the house on all sides. It should be. Otherwise, surface water can run directly to the foundation (see FIG. 6-6), seep down along the foundation walls, and accumulate at the lower section. This usually results in water seepage into the basement or crawl space. A rule of thumb for grading this area is a drop of about 1 inch per foot. The lawn should slope away from the house for at least 10 feet and should be pitched so that there is approximately a 10-inch drop over that distance. (See FIG. 6-7.)

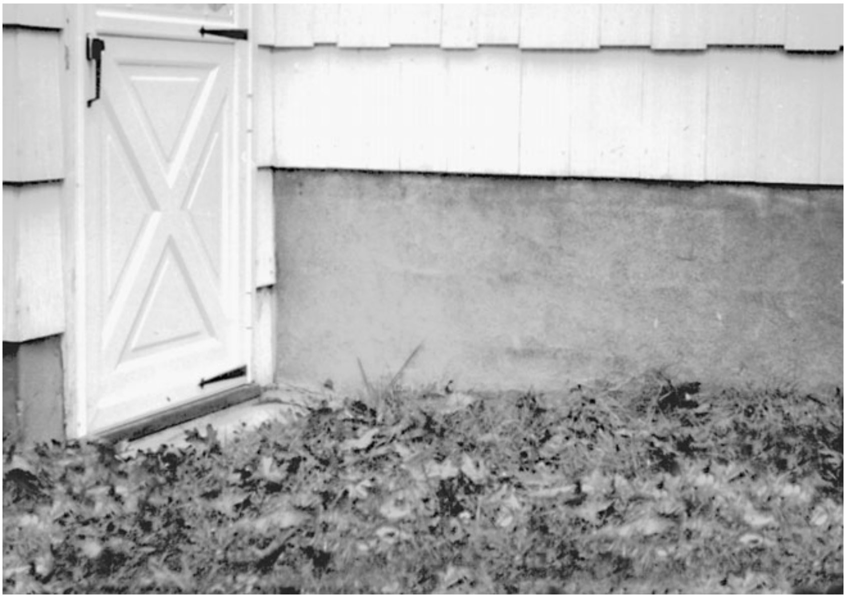

Fig. 6-6. Improper grading of the land adjacent to the house. The lawn pitches toward the door, resulting in surface water ponding in front of the entry area.

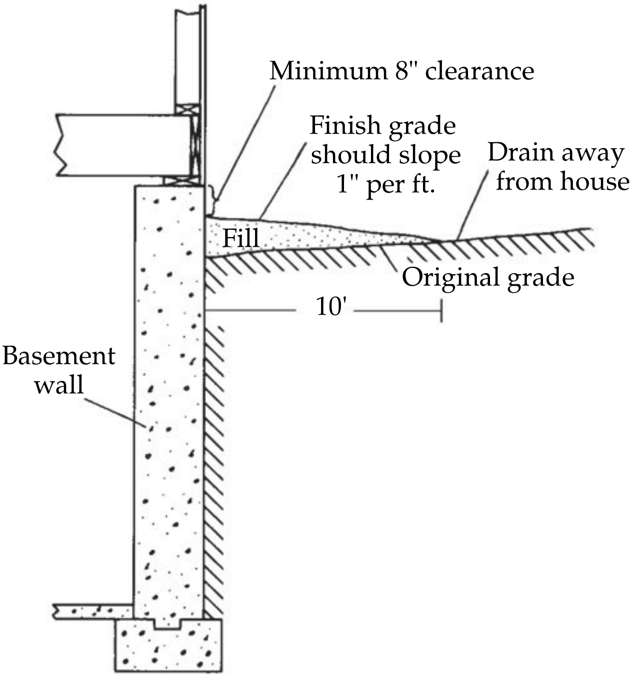
Minimum 8" clearance
Finish grade should slope 1" per ft.
Drain away from house
Fill
Original grade
10'
Basement wall

Fig. 6-7. Finish grade of the ground adjacent to the foundation, sloped for proper drainage.

Because of normal soil settlement and compaction, clogged gutter or faulty downspout, and foot traffic (especially when the ground is wet), the slope of the ground adjacent to the house usually changes with time. Consequently, the area around the foundation should be periodically checked for proper grading.

When checking the grading around the foundation, you might occasionally see a pipe protruding through the foundation wall or from a basement window. Record the fact on your worksheet for further investigation when you inspect the basement. Usually, this pipe is connected to a sump pump and is used for discharging the water that accumulates in the sump pit. The end of the pipe should be extended away from the house so that the water does not accumulate around the foundation. Sometimes the pipe terminates just beyond the foundation wall, which negates the advantage of a sump pump. The discharging water accumulates around and under the foundation and reenters the sump pit, only to be pumped out again. (See FIG. 6-8.)

If you find a stream on the property, you should realize the potential for flooding. As discussed previously, flooding can occur because of increased surface runoff resulting from a prolonged heavy rain. In addition, if the stream channel is blocked by fallen trees, tree limbs, sediment, or trash, flooding can result. Depending on the location of the stream relative to the house, occasional overflowing of the stream banks might or might not be a problem.

As you walk around, look at the overall landscape. If the topography is sloping, does a natural drainageway direct surface runoff away from the house, or does the lot need a swale or ridge? If there is an abrupt change in the grading, is there a need for a retaining wall? Are any areas eroding to the extent that corrective action is necessary? Are low or level areas vulnerable to ponding? Ponded water on the lawn does not necessarily indicate a drainage problem. It might be caused by a malfunctioning septic system, a faulty sewer hookup, or a break in the main water-supply pipe. If you find ponded water during your inspection, try to determine the cause.

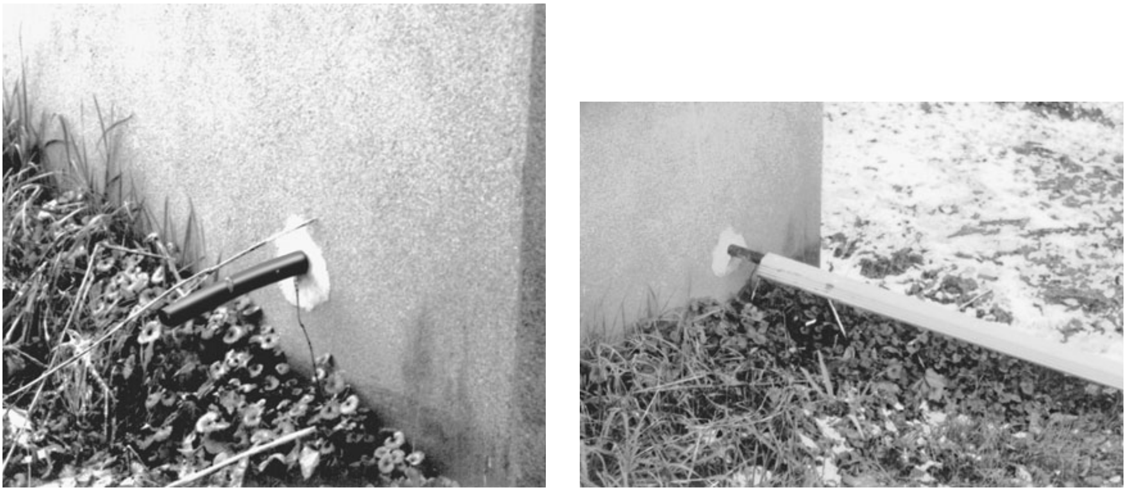

Fig. 6-8. *Sump pump discharge pipes. Left—The pipe terminates just beyond the foundation wall. Right—The discharge pipe has been extended so that effluent discharges away from the foundation.*

If there are footing drains around the foundation or curtain drains in a hillside, they will not be visible during your inspection. For these drains to function properly, they must have a free-flowing outlet. (See FIG. 6-9.) Ask the seller if there are drains, and if so, ask for the location of the outlets. Unfortunately, most homeowners do not know whether there are footing or curtain drains. When a house is sold, this type of information is usually not discussed, although it should be. Consequently, the location of the drain outlet is lost for all future owners. If you are lucky enough to have a seller who knows the location of the drain outlet, you should inspect the opening to make sure that it is not obstructed. If the footing drain discharges into a stream channel, an overflowing stream could be a problem. It all depends on the elevation of the footing drain relative to the level when the stream is running full. If the stream is above the level of the footing drain, then water will back up into the drain around the foundation and could cause a water seepage problem. Look for evidence of seepage when inspecting the basement or crawl space. In new homes, the drain outlet is sometimes inadvertently blocked when the lawn is landscaped. If you are buying a new house, ask the builder to show you the location of the drain outlet.

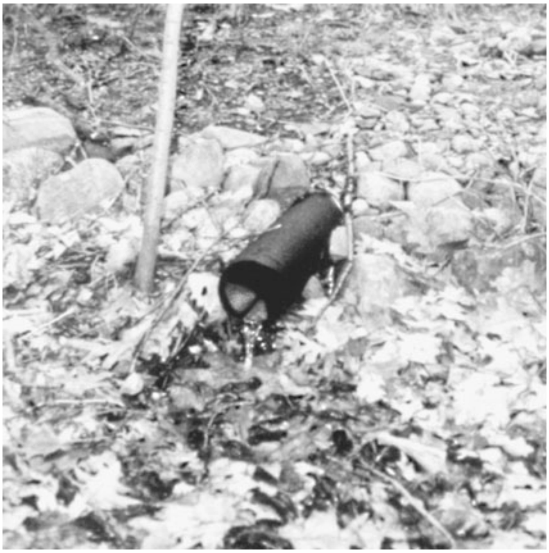

Fig. 6-9. *Free-flowing outlet of curtain drain. Note water discharging from pipe.*

Retaining walls

Retaining walls are mostly used for stabilizing and controlling erosion on steep banks. In some cases, they are used in conjunction with terracing of rear or side yards to provide a level area for recreation. In either case, they must be designed to withstand the lateral pressures being exerted on them by the soil.

Retaining walls are normally built with construction timbers, railroad ties, stone, concrete, or concrete blocks. Some concrete and concrete-block walls have stone or brick veneer facing. On occasion, you might find a *gabion* retaining wall—steel baskets filled with stones. (See FIG. 6-10.) As gabions age, the steel baskets tend to corrode and deteriorate, especially on the side facing the embankment. Over the years, however, soil sediment usually fills the voids between the stones and tends to hold the wall in place.

Stone retaining walls are often referred to as dry or wet, according to whether mortar

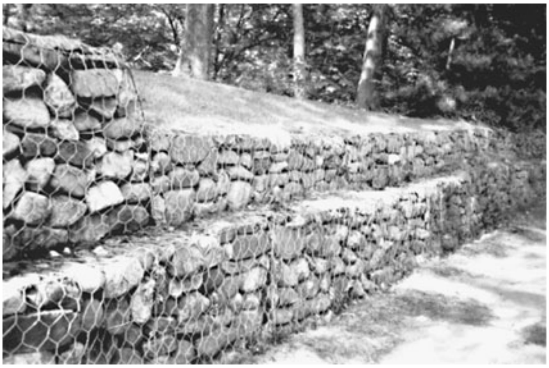

Fig. 6-10. *Gabion retaining wall.*

was used between the stones. A *dry* retaining wall is one that has been constructed without mortar. It depends on the weight and friction of one stone upon another for stability. Frost heaving is not a problem with this type of wall. The stones are not bonded together and will therefore be raised and lowered together by the frost. Consequently, the bottom course of the wall is usually only about 6 inches below grade rather than below the frost line.

A *wet* wall is one with mortar between the stones. The mortar secures one stone to another and thereby achieves a monolithic wall with greater stability. Because this wall is integral, frost heaving will cause cracking. To prevent frost heaving, the bottom of the wall must be below the frost line. A wet wall offers greater solidity, and therefore loose soil will not wash out or run through the voids. Also, a wet wall is less of a hazard because no loose stones can be kicked out of place or fall off the top.

When constructing a retaining wall, provisions must be made for draining the water that normally accumulates behind the wall. Otherwise, a hydrostatic pressure buildup can cause structural failure of the wall. Drainage should be provided by installing a continuous perforated drainpipe at the lower portion of the wall and backfilling the area with broken stones or gravel. The pipe should be directed so that the effluent flows to a suitable location away from the wall. In many monolithic retaining walls, the perforated pipe is replaced by *weep holes*, holes in the wall that run from the front to the gravel backfill. The weep holes allow the water that accumulates in the gravel to drain out through the wall. Weep holes should be placed at a 5- to 10-foot spacing and should be 4 inches in diameter (the same diameter as the drainpipes).

Unfortunately, retaining walls are often built without adequate drainage provisions. The gravel backfill and drainpipe might be omitted, or the weep holes might be too few or too small to be effective. The weep holes must be kept clear so that the water behind the wall can be adequately drained.

A retaining wall built with construction timbers or railroad ties should be anchored into the hillside to provide the resistance to overcome the lateral forces exerted on it. If the wall is not tied back into the earth, it can bow, buckle, or heave and eventually collapse. Anchoring of the wall is achieved by using *tiebacks* and *dead men*. A tieback is a construction timber that has been placed perpendicular to the wall. The front end is flush with the wall and fastened to it with large spikes. The rear end is fastened to a dead man, a small section of timber perpendicular to the tieback and parallel to the wall. (See FIG. 6-11.) When the area around the wall anchor is backfilled with soil, a force is developed on the anchor that resists the lateral force on the wall. Because of the open joints between the railroad ties, weep holes are not needed.

Many railroad-tie or timber retaining walls are not constructed with anchors. You can tell whether anchors were used by looking at the wall. If tiebacks were used, end sections will be visible in the face of the wall. (See FIG. 6-12.) However, from a visual inspection, you cannot tell the length of the tiebacks or whether dead men have been installed. Your inspection, therefore, should concern itself with the condition of the wall rather than its construction, unless it is a new wall. If it is newly constructed, you should inquire about a guarantee.

Inspection

When inspecting a retaining wall, look at its overall condition. With a dry stone wall, look for missing and loose stones and crumbled sections. This type of wall is relatively easy to repair and generally does require periodic maintenance. With a wet stone wall, check

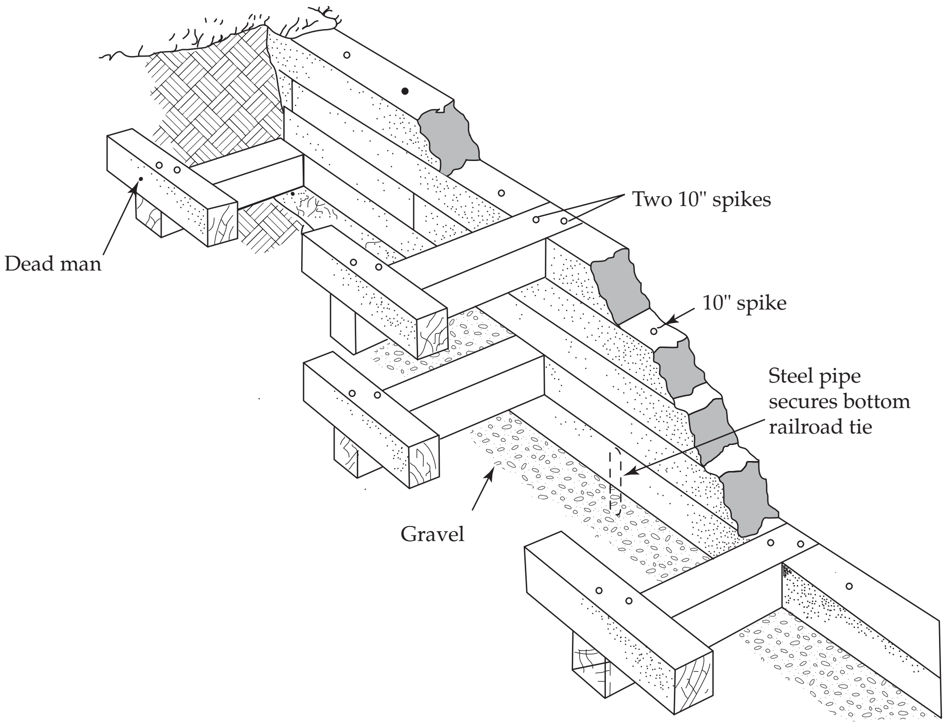
Two 10" spikes
Dead man
10" spike
Steel pipe secures bottom railroad tie
Gravel

Fig. 6-11. *Construction timber/railroad tie retaining wall.*

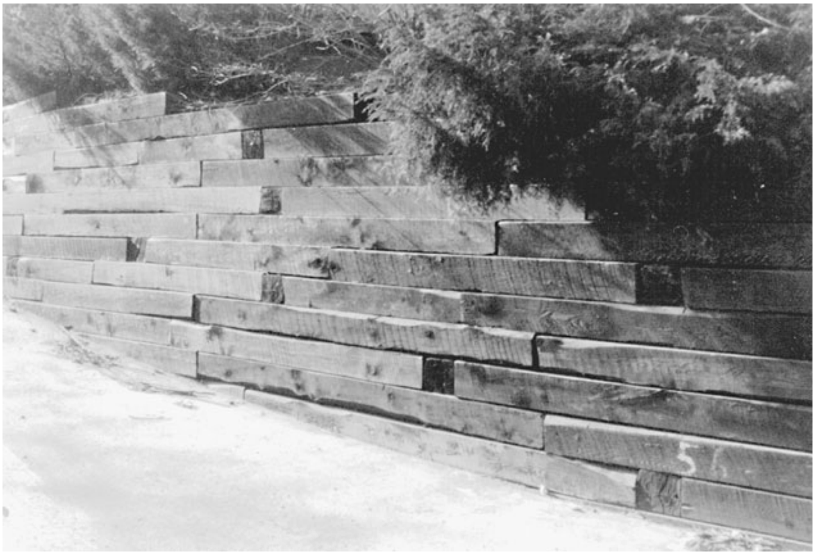

Fig. 6-12. *Timber retaining wall. Note the end sections, indicating tiebacks used for anchoring the wall.*

the mortar joints for cracked, loose, and deteriorated sections. Are there weep holes in the wall? If so, are they adequately sized and unobstructed? Concrete and concrete-block walls should also have weep holes. Are there any cracked and heaved sections? Wood-constructed retaining walls should be checked for cracked, rotting, loose, and heaved sections. Some retaining walls are completely covered with vines. Try to push the vines aside so that you can inspect the wall. Quite often I have found cracked sections in the wall through which the vines were growing. While inspecting the wall, look for loose, heaved, and deteriorated sections.

All retaining walls should be vertical or inclined slightly toward the embankment. They should not be leaning forward. When they are, it is an indication that they could not withstand the lateral forces being exerted on them by the terraced or sloped earth behind. Once a wall cracks and heaves, the pressure that caused the condition is relieved, and the wall might stay in the leaning position for many years. (See FIG. 6-13.) However, additional forces might cause the wall to continue to heave and eventually to collapse. If the heaving is excessive or by the wall's collapsing someone can get hurt, the wall must be rehabilitated or braced. If you have any questions about what is excessive heaving, have the wall inspected by a professional.

In recent years, manufactured concrete blocks have been used for retaining walls for residential construction (see FIG. 6-14). The blocks come in different sizes and shapes depending on the manufacturer. Often these walls are 5 feet high or higher, and many are constructed in tiers. In most states, if the retaining wall is taller than 4 feet, it must be designed or approved by a qualified, licensed professional engineer. In order to prevent failure and to reduce the earth's pressure behind the retaining walls, they are reinforced with horizontal layers of a geosynthetic polymer mesh, which extends back into the

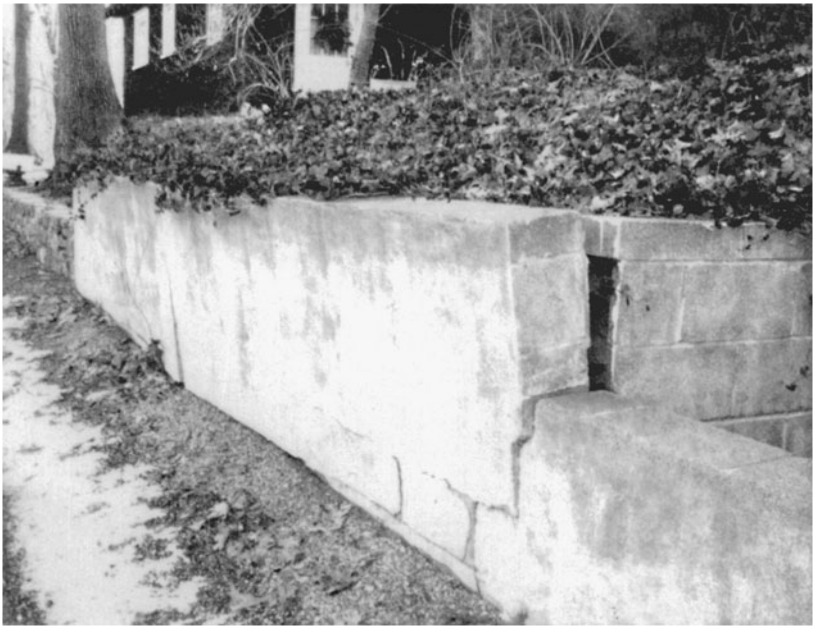

Fig. 6-13. *Cracked and heaved retaining wall.*

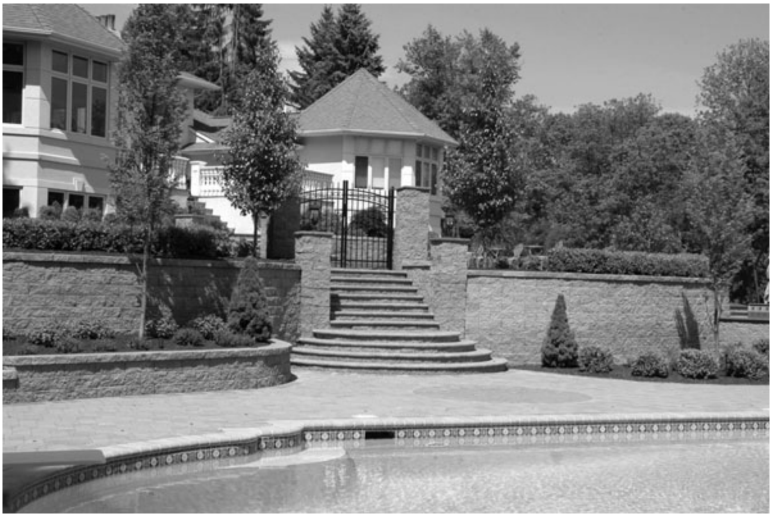

Fig. 6-14. *Concrete block retaining wall.*

embankment. Since the length of the mesh will not be visible, from an inspection point of view all you can do is to check the wall for tilting, cracking, or bowing. If any are noted, it should be recorded on your worksheet for later inspection by a professional engineer.

Landscaping

As you walk around the house, inspect the landscaping in the front, rear, and side yards. Specifically, look at the lawn, shrubs, and trees. A nicely landscaped area greatly enhances the beauty and value of the house. However, do not jump to a hasty conclusion about the house based on the landscaping. I have seen many neglected houses with beautiful landscaping and many well-maintained houses with poor landscaping.

Lawn

A lawn serves two purposes. It adds to the aesthetic beauty of the property, and, more important, it prevents erosion and washout of the topsoil. If you find that a large portion of the lawn consists of crabgrass and other weeds, do not be distressed. With a planned program of weed control, seeding, and fertilizing, you can upgrade the lawn so that it can be the "showcase of the neighborhood" within a few years, and at not too great an expense.

If you find holes or sunken sections in the lawn, they should be filled in, since they represent a potential tripping hazard. Occasionally, sunken sections are caused by the collapse of rotted, decayed, or deteriorated construction debris buried on the site years before. In new homes, all construction debris should be removed from the site rather than buried there.

In some parts of the country, moles are a problem in lawn maintenance. They burrow in the ground near the surface in search of food (grubs, caterpillars, and insects) and in the process create soft ridges (mole hills) that spoil the lawn's appearance. If you see ridges over portions of the lawn and they feel soft when you walk over them, suspect moles. This condition can generally be controlled through soil treatment and should be discussed with the proprietor of a local nursery.

Some lawns have steep sloping areas that from a maintenance point of view are quite difficult to mow, even when mowing across the slope. Because of the danger involved, sit-down riding mowers should not be used when cutting the grass on a steep slope. These mowers have been known to topple over and severely or fatally injure the driver. In some homes, the steep sections of the lawn have been replaced by terraced areas with steps that lead from one level to another. If there are terrace steps on the lawn, you should inspect them for cracked, loose, missing, and deteriorated sections. Also, check for uneven treads and dimensional variations in the risers (a tripping hazard). In addition, if there are more than two steps, a handrail is recommended.

If you are planning to buy a newly constructed house, find out whether it will be your responsibility or the builder's to establish a new lawn. If it is yours, it can be quite expensive, depending on the size of the lawn and whether topsoil needs to be added. Over 100 tons of topsoil is needed to cover an area one-third of an acre to a depth of 2 inches.

Shrubs

When new homes are landscaped, the shrubs are often intentionally planted very close to one another to produce an immediately pleasing effect. Many homeowners do not plant shrubs with the future appearance in mind. Consequently, as the shrubs grow and fill out, they tend to crowd one another, losing their individuality. Eventually, they become unsightly, with portions dying off due to lack of sunlight. In addition, the growing shrubs often block walkways and produce so much shade that the area around the house is always damp, a condition conducive to the growth of decay fungi and mildew. By extensive pruning of the shrubs and transplanting others (if you want to save them), the area can often be completely rejuvenated and restored to its original beauty. For some people, their dream house is one that is covered with ivy. (See FIG. 6-15.) Actually, vines of any type growing up the outside walls of a house are quite undesirable. As vines grow they can cause problems. They can lift wall

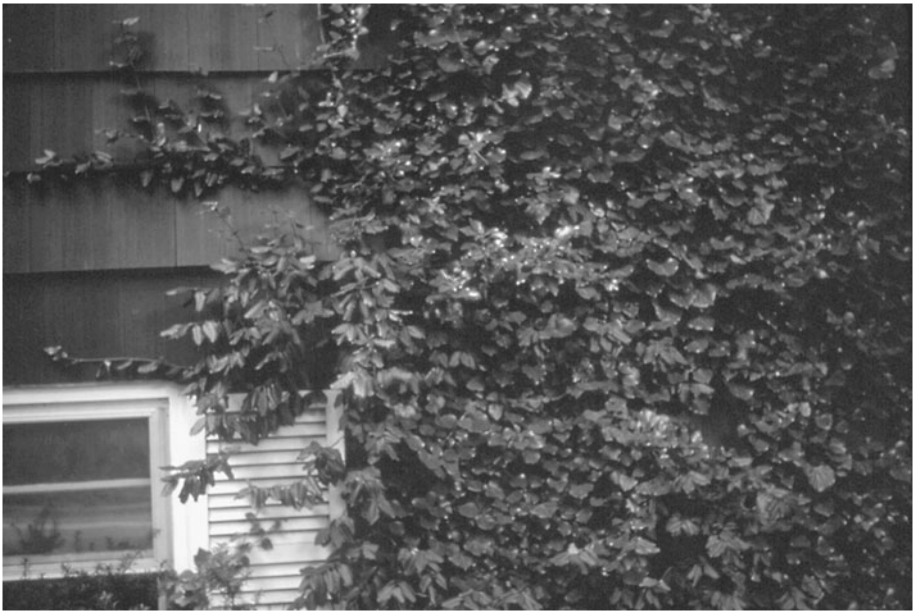

Fig. 6-15. *Vines growing up and partially covering exterior shingle wall.*

shingles and roof shingles. They can also grow into mortar joints and crush downspouts. (See FIG. 6-16.) If you see vines on the house, you should consider their removal.

Trees

If there are trees on the property, they should be checked to see if any are dead or have any large dead branches. All dead trees should be taken down. Because they are vulnerable to insect damage and decay, they are a potential hazard, especially if they are located near the house. Large dead branches are also a hazard. On a windy day, they can break off the trees and fall to the ground or, worse, onto the house.

After a deciduous tree has lost its leaves, it can be somewhat difficult to determine whether it is dead or has any dead branches. However, if you see any limbs with the bark peeled off, you can assume that those branches are dead. (See FIG. 6-17.) If you have any doubt, after you move into the house, you should have the trees checked by a professional, or wait until spring and summer when all the trees are in full bloom.

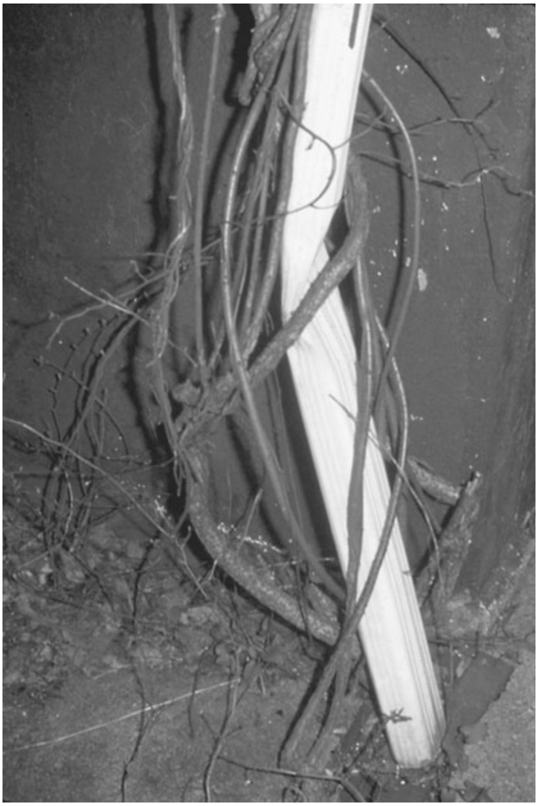
Fig. 6-16. *Downspout crushed by vines.*

If you find any dead trees or dead branches, record their location on your worksheet for later removal. Depending on the size and location of the dead tree, its removal can be somewhat costly. This type of work should be performed only by a professional who is insured in the event that the tree causes damage when it falls to the ground. In addition to dead branches, all limbs that are overhanging or resting on the roof should be pruned back. Otherwise, they might eventually damage the roof.

If you are buying a newly constructed house with trees on the property, you should be aware that the roots of some of the trees might have been damaged during construction. This could occur as a result of heavy equipment (tractors or trucks) being driven too close to the tree. Trees that have had root damage during construction do not necessarily show any immediate effects. However, within a year or two and depending on the degree of damage, the trees may die. If care is taken during construction, this problem can be avoided. Your best bet is to buy a house from a quality builder.

Decks

There are many types and styles of decks. However, from an inspection point of view, your main concern should be safety rather than appearance. When inspecting a deck, unless it is a rooftop or cantilever type, you should begin with the supports on the underside. If the deck is more than a few feet above the ground, it will generally be supported by wood or metal columns (posts). Unless wood

Fig. 6-17. Large, dead branches are a potential safety hazard and should be removed.

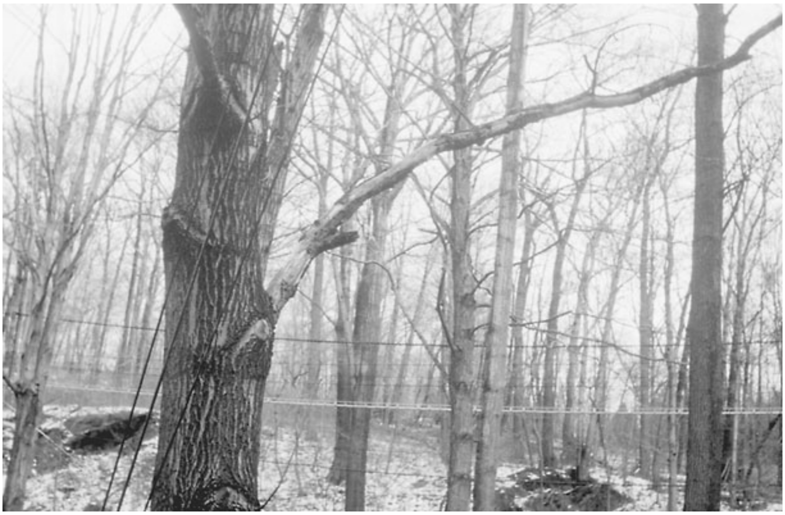

posts have been pressure-treated, they should not be in direct contact with the soil. Untreated wood in contact with the ground is vulnerable to rot and termite activity. Also, the dampness normally associated with the soil can promote rust deterioration of a metal post or connector.

Each post should rest on a concrete pad that has a footing below the frost line. Otherwise, the footing is subject to frost heave. Probe the base sections of the posts with a screwdriver to determine whether there is deterioration. If the screwdriver can penetrate the post beyond the surface, a problem exists that should be corrected. In some cases, the post might require replacement. Push the post to see if it moves. It should not. Occasionally I find columns that are loose and not adequately supporting the deck. (See FIG. 6-18.) The condition usually results from uneven settlement of the support footings and inadequate fastening at the top or bottom of the column. A loose post is a potential hazard and must be resecured as soon as possible.

When the deck is less than a few feet above the ground, it is usually supported by masonry piers. Inspect the piers for cracked, broken, loose, or deteriorated sections. They must be repaired. With some "ground-hugging" decks, usually less than a foot above the ground, this inspection might not be possible. Most of the support piers are not visible.

When one side of the deck is attached to the house, there are usually no support posts below that section. Consequently, if the joint between the house and the deck should weaken, there is a potential for the deck to collapse. Check the joint between the deck and the house to see if it is securely fastened. Is it pulling away from the house? It shouldn't be. This is a potential problem and should be called out on your worksheet for additional bracing. In some cases, the deck is fastened to the house with undersized or too few nails. I know of one community where this type of installation resulted in two decks collapsing. Because of these failures, the town passed an ordinance that requires using lag bolts rather than nails to secure the deck to the house. (See FIG. 6-19.)

Next, check the joist supports at the portion of the deck attached to the house. Since there are no posts, there will not be a girder to support the joists. In this case, the joists should be supported by metal brackets fastened to the

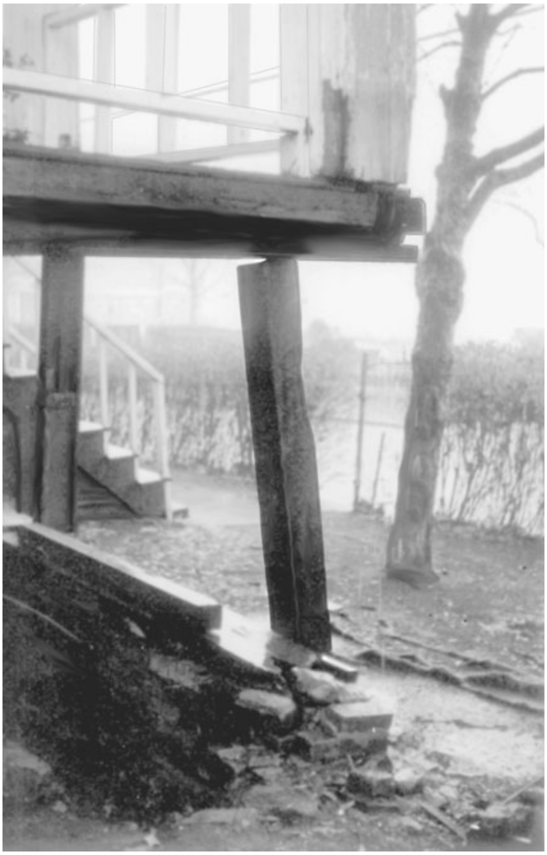

Fig. 6-18. *Inadequately supported deck. Column is loose and can easily be knocked over.*

header or by being toenailed into the header with a ledger below them. The former method is preferred because the ledger used is often skimpy. I have seen many decks where the joists were toenailed into the header but the ledger was never installed. If you find this type of installation, you should install angle brackets to support the joists or at the very least install a ledger as a precautionary measure. Check the condition of any and all metal brackets, connectors, nails, and screws. Over time they can corrode and rust, either because of the weather or because of the chemicals used in the deck's pressure-treated wood.

In addition, depending on the size of the deck, there might be a need for diagonal bracing. This provides additional rigidity to the deck and can be achieved by placing a 2-inch-by-6-inch board between two diagonal corners and nailing it to the underside of each joist.

When deck planks are installed, there should be a space of about ¼ inch between them. This space allows rainwater and melting snow to drain. In some cases, the planks are butted up against one another so that there is virtually no space between them other than the crack of the joint. Water entering this crack does not readily drain and instead promotes decay. If you can reach the underside of the deck planks and the top portion of the joists, probe them for rot. If there are rotting sections and the decay is advanced, you might see the fruiting bodies

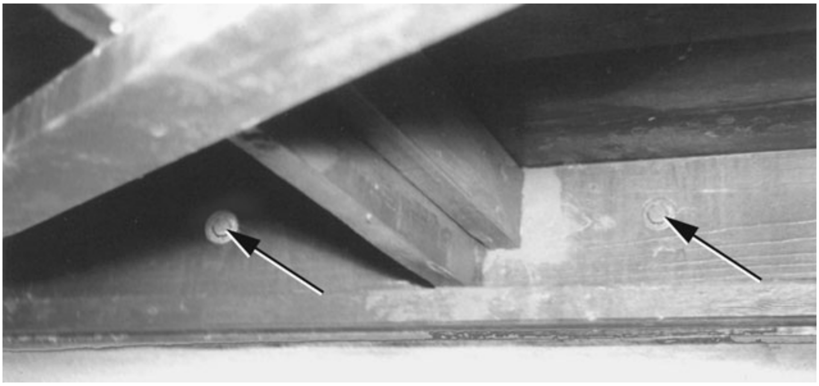

Fig. 6-19. *Lag bolts are used to secure the deck to the house. Note that joist is resting on a ledger board for support.*

of the decay fungi. (This is discussed in detail in the section on rot in chapter 8.) If steps lead to the deck, the treads, stringers, and handrails should be inspected for decay. As with the deck support posts, the stringers should rest on a concrete pad rather than soil.

After inspecting the underside of the deck and the steps, you should inspect the top portion. When the deck is more than 30 inches above the ground, there should be a guardrail around the perimeter as a safety precaution. Naturally, the higher the deck, the greater the need for a rail. Issues such as wobbly railings, loose stairs, and ledgers that appear to be pulling away from the home are all causes for concern. Metal connectors, nails, and screws can corrode over time. Look for rust and other signs of corrosion that can weaken the structure of your deck. Check the rail to see if it is loose and wobbly. If it is, record it on your worksheet to be resecured. If the deck is to be used by small children, additional protection is needed to block the open area between the railing, railing posts, and deck planks. When balusters are used for this purpose, they should be spaced 4 to 5 inches apart.

The guardrails and deck planks should be inspected for cracked, rotting, and loose sections. (See FIG. 6-20.) Depending on the quality of the wood and the upkeep, a deck need not deteriorate to a point where it requires complete rehabilitation. Repairs or replacement of deteriorated sections should be performed as needed. If you find sections of the deck that are in need of repair, you should indicate those areas on your worksheet.

Free-standing decks

Free-standing decks are independent self-supporting structures. If constructed properly they are as safe and secure as attached decks and in some cases more secure. A number of attached decks have collapsed because the joint between the deck and the house has pulled away, because either it was improperly or inadequately secured or it has rotted. Depending on the size of the free-standing deck, it will require an additional beam and two or more posts situated next to the house. The posts should be mounted on footings that extend below the frost line. In order to be able to resist lateral and horizontal movement, diagonal bracing between the posts and the beams must be provided. (See FIG. 6-21.) As with attached decks, these decks should be checked for cracked, broken, loose, or rotting sections.

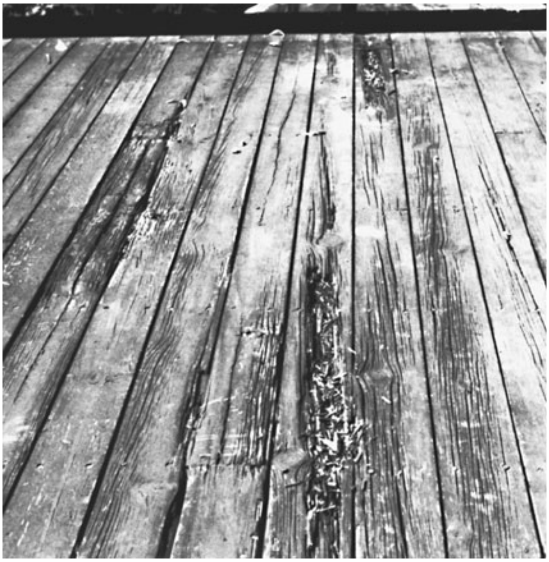

Fig. 6-20. *Cracked and rotting deck planks. Planks should be replaced as needed.*

Composite decks

Most residential decks are constructed entirely of wood. However, the use of composite lumber for the visible portions of the deck, such as the deck planks, handrails, and trim has become very popular. Its popularity stems from the fact that it does not require waterproofing and resists damage from weather and insects. It is nevertheless not maintenance-free. The composites, which are generally made of recycled plastics and waste wood fibers, are

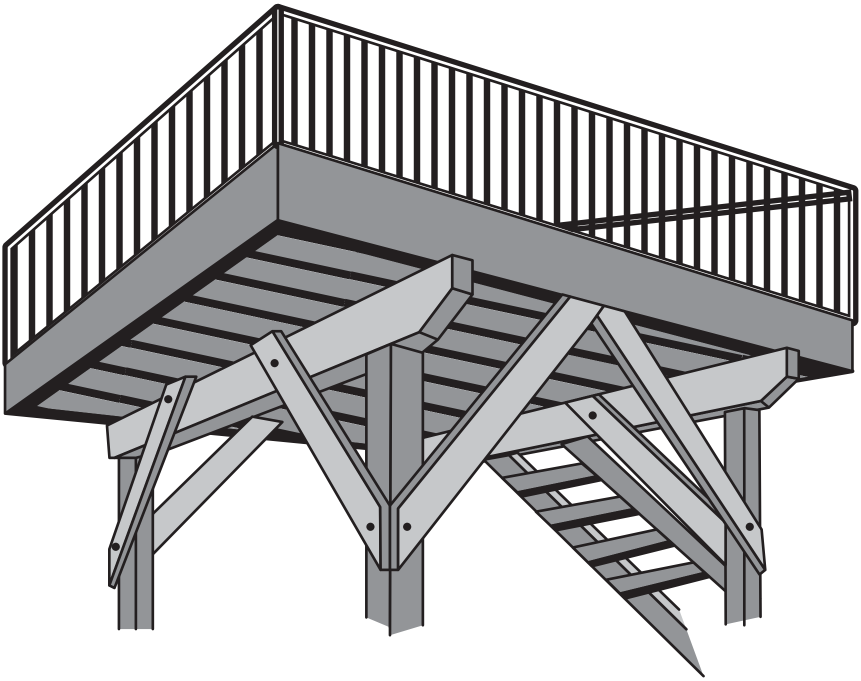

Fig. 6-21. *Diagonal bracing on free-standing deck.*

not intended for structural use. The structural support for those decks that use composite lumber (the joists, beams, and posts) will still be made of wood and should be inspected as previously discussed.

Trex and Timber Tech are the leading brands of composite lumber for decks, although there are a number of companies that also make composite lumber. Not all composite decks are made with the same quality ingredients. Because the material expands and contracts with temperature, proper spacing during installation is critical. Cupping deck planks and joints that don't line up have been noted. In addition, some of the other problems that have been noted, which should be recorded on your worksheet, are cracks, scratches, staining, mold (mildew) blotches, and fading.

One last point: If the deck was not built when the house was constructed or the deck is a complete replacement of a previously deteriorated one, check with the seller to see if a certificate of occupancy (CO) was issued by the local building department at the completion of construction. A CO is required by most municipalities. If one was not issued, record that fact on your worksheet for future discussion with your attorney.

Fences

If there is a fence on the property that you are inspecting, you should check its overall condition. The problems encountered with fences are normally not major and are usually not costly to correct. However, you might find a fence that has deteriorated to a point where it needs complete rehabilitation or replacement.

Wood fences should be inspected for cracked, broken, loose, and missing sections. In addition, they should be checked for deterioration from rot and termite infestation.

(Termites and rot are discussed in chapter 8.) The gates for the fence should also be checked for cracked, loose, or broken sections and ease of operation. Wooden gates often sag as they age and require periodic maintenance.

Metal fences should be inspected for rusting, loose, and deteriorated sections. Rusting sections should be scraped, primed, and painted. Chain link fences are generally constructed of galvanized steel. Galvanizing (zinc coating) protects the steel against rusting and is usually applied by hot dipping or electroplating. The hot-dipped process produces a heavy zinc coating that is very effective, in contrast to the thin coating produced by electroplating. Chain link fences that have been galvanized by electroplating have a tendency to rust and require periodic maintenance. Some chain link fences have a vinyl coating that protects against rust. The vinyl coating is quite effective and lasts for many years.

If there is an in-ground swimming pool on the property, there should be a fence around the pool area. Most communities have an ordinance requiring a fence of a specific height as a protective barrier. If you do not see a fence around a pool area, you should check the requirements with the municipal building department; otherwise, you might find after buying the house that you are legally obliged to install one.

Checkpoint summary

Drainage

❍ When approaching the house, take note of the overall topography.
❍ Is it level or inclined?
❍ Are there gently or steeply sloped areas?
❍ Is the house located near or at the bottom of an inclined street?
❍ Note whether there is a storm drain (catch basin) nearby.
❍ Are there nearby streams or brooks?
❍ Are you able to determine if the house is located in a flood plain or flood-prone area?
❍ Is the ground immediately adjacent to the house graded so that it slopes away on all sides of the structure?
❍ Are there natural drainageways to direct surface water away from the house?
❍ Are there low or level areas that are vulnerable to water ponding?
❍ Are there areas of ponded water on the lot?
❍ Are you able to determine whether the house has footing drains?
❍ Can you locate the outlet for these and any other drainage pipes?

Retaining walls

Timber, railroad tie, dry stone wall, gabion

❍ Inspect for missing, loose, and crumbling sections of stone.
❍ Check timber and railroad-tie walls for cracked, loose, rotting, and heaved sections.
❍ Are the wood-constructed walls properly anchored (tiebacks)?

Concrete, concrete block, wet stone wall

❍ Inspect for cracked and heaved sections.
❍ Check for loose, deteriorated, and missing mortar joints.
❍ Is the wall vertical, or does it lean?
❍ Are portions of the wall heavily covered with vines?
❍ Did you inspect these areas for cracked and heaved sections?
❍ Try to determine whether the area behind the retaining wall is adequately drained.
❍ Are there weep holes at the base of the wall?
❍ Are they blocked?
❍ Are the weep holes adequately sized and spaced?

Landscaping

Lawn

- ❍ Inspect for holes, sunken sections, bald spots, and eroding areas.
- ❍ Estimate areas that will require recultivation.
- ❍ Note soft sections or ridges (possibly due to moles).
- ❍ Inspect terrace steps for cracked, loose, rotting, or missing sections. Check steps for handrails, uneven treads, and variations in riser heights.

Shrubs

- ❍ Inspect shrubbery for overcrowding, dying sections, blocked walkways, steps.
- ❍ Note areas in need of pruning, transplanting, or removal.
- ❍ Note areas of the house that are covered with vines.

Trees

- ❍ Check for dead trees and limbs, especially those close to the house.
- ❍ Note tree limbs that are overhanging or resting on the roof.
- ❍ Note for future professional evaluation any trees that show evidence of rot, split sections, or insect infestation.

Decks

- ❍ Check and inspect the various deck components for safety rather than for appearance.
- ❍ Check concrete or brick piers for cracked, loose, and deteriorated sections.
- ❍ Inspect wood columns for rot and termite activity.
- ❍ Inspect metal columns for rust deterioration.
- ❍ Are columns supported on concrete pads, or are they in contact with the ground?
- ❍ Note any loose columns.
- ❍ Check for open and weakened joints between the deck and the house.
- ❍ Is the deck attached to the house with nails or lag bolts? (Lag bolts are preferred.)
- ❍ If the deck was not built at the time the house was constructed, does it have a certificate of occupancy (CO)?
- ❍ Inspect deck-joist supports at the portion of the deck attached to the house.
- ❍ Are deck joists supported by metal brackets (preferred), or are they toenailed into a header beam with a ledger board below the joist?
- ❍ Where the joists have been toenailed, check for missing ledger boards.
- ❍ Does deck contain diagonal bracings?
- ❍ Inspect the underside of deck (girders, joists, floor planks) for missing, cracked, and rotting members.
- ❍ Inspect wood step treads, stringers, and handrails for cracked, loose, missing, and rotting sections.
- ❍ Are the stringers supported on a concrete pad, or are they in contact with the ground?
- ❍ Check top portion of deck for cracked, loose, missing, and rotting sections of deck planks, railings, and railing posts.
- ❍ On free-standing decks, check for diagonal bracing, as well as for cracked, loose, and rotting members.
- ❍ On composite decks, inspect the wood structural support members. Check the composite decking, handrails, and trim for cracks, scratches, staining, mold (mildew) blotches, fading, and joints that don't line up.

Fences

- ❍ Inspect wood fences for cracked, broken, loose, and missing sections.
- ❍ Check for areas of rot and insect damage.

❍ Inspect metal fencing for loose, missing, and rusting sections.

❍ Check gates (metal and wood) for sag, missing hardware, and cracked, loose, broken, and missing sections.

❍ If property contains an in-ground pool, is the area around the pool adequately fenced off? (It might be a legal requirement.)

7

Garage

Attached garage 84
  Fire and health hazards 84
  Plumbing check 86
  Flood potential 88
  Doors 88
  General considerations 89
Detached garage 90
  Exterior 90
  Interior 91
  Heat and electricity 92
Checkpoint summary 93

The garage should be inspected after the exterior inspection has been completed. There are two basic types of garages: attached and detached. An attached garage is a part of the main building. It might be located below a habitable portion of the structure or connected to the side of the building. A detached garage is a separate structure, not part of the main building. There might be a connecting breezeway or porch between the two structures.

Attached garage

Since the principal use of the garage is car storage, the possibility of dripping oil and gasoline presents a potential fire hazard. Because the attached garage is connected to the main structure, certain precautionary measures should be taken during construction to minimize the hazards. Look around the garage to see if there are any potential problems.

Fire and health hazards

Is there an interior door between the garage and the house? If there is, is there at least one step leading up to the door? There should be. (See FIG. 7-1.) It is surprising how often I find that the garage floor slab is at the same or a higher level than the adjacent living area. (See FIG. 7-2.) The living area should be above the level of the garage floor to prevent toxic exhaust gases and gasoline vapors, which are heavier than air, from entering the house whenever the interior door is opened. As a precautionary measure, the interior door should have a tight seal around the joints to prevent seepage. This door should be fire-resistant, such as metal-clad, solid wood, or hollow core, with a sheet-metal covering on the *garage side*. As a safety feature, the interior

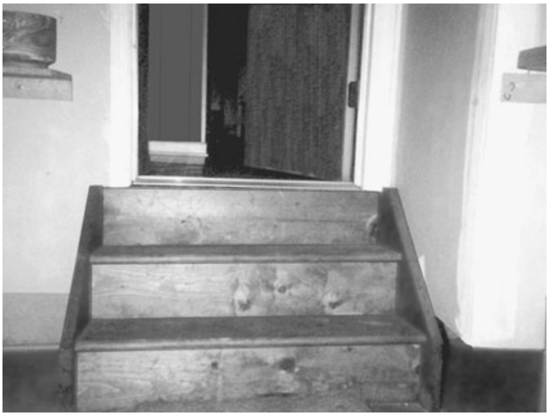

Fig. 7-1. *Steps connecting the garage floor slab to the adjacent living area, which is at a higher level.*

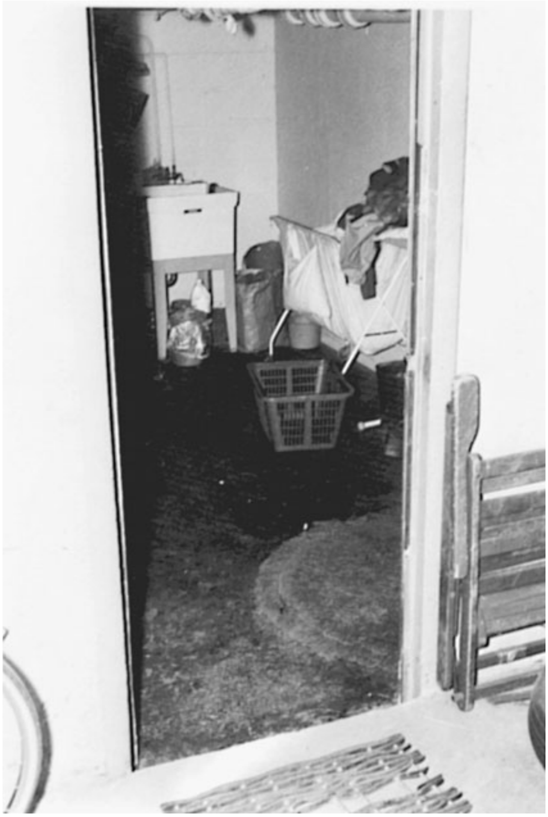

Fig. 7-2. *Garage floor slab at a higher level than the adjacent living area—a potential hazard.*

door should also be self-closing. In most homes it is not. It seems that homeowners have found self-closing doors inconvenient, especially when carrying in packages from the local supermarket. Nevertheless, safety should not be sacrificed for convenience.

Next, look at the walls that separate the garage from the living area. Are there exposed wood-frame members? There should not be. Exposed wood framing in this area is considered a fire hazard and should be covered with a fire-resistant material such as plaster or stucco on lath or ⅝-inch plasterboard. This wall should be insulated to reduce heat loss. If there is a living area above the garage, the ceiling should be insulated and have a fire-retardant covering.

In some garages an access hatch to the attic is located in the ceiling. Occasionally the hatch cover for this opening is missing or open. (See FIG. 7-3.) This is a fire hazard. If a fire should start in the garage, the open area in the ceiling could act as a flue and draw the flames up into the attic where they would quickly engulf the house. The attic hatch cover must be in place at all times.

Some homes have a garage in the basement. The garage is at the basement level with no partition walls separating the garage area from the basement area. This is a fire and health hazard in addition to being inefficient from an energy-conservation point of view. When the garage doors are opened, there will be a loss (from the basement) of warm air in the winter and cool air in the summer.

Occasionally the heating plant (furnace or boiler) is located in the garage. (See FIG. 7-4.) This is perhaps the least desirable location for a heating unit. There is always the possibility that a leak could develop in the gasoline tank or fuel line of an automobile. If the garage is inadequately ventilated, the resultant flammable vapors could be ignited by the flame in the heating system. Of course, this

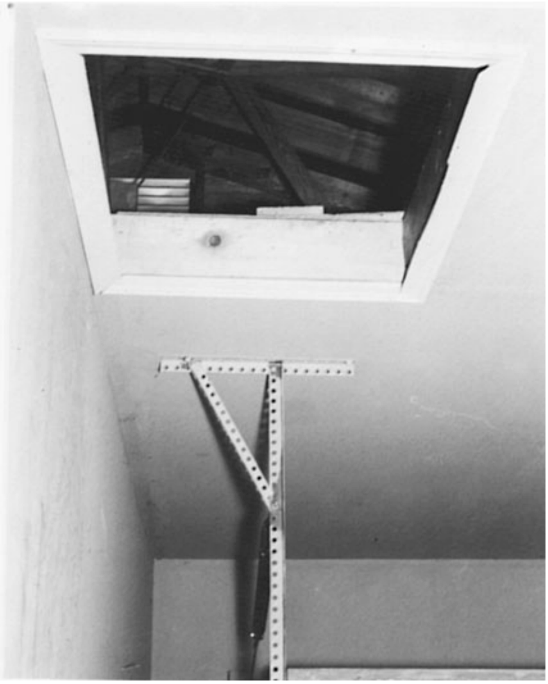

Fig. 7-3. *Open attic-access hatch in the garage ceiling.*

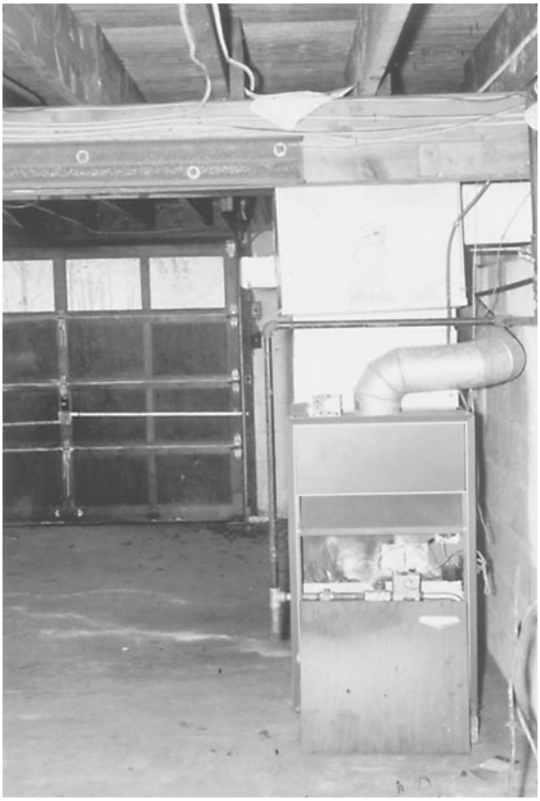

Fig. 7-4. *Furnace located in the garage. Note that the ceiling should have been covered with fire-rated gypsum board.*

risk can be minimized by locating the heating unit on a platform or by building a low wall around the heating system (see FIG. 7-5) (since gasoline vapors are heavier than air and will accumulate at ground level) or by enclosing the heating unit in a room with a tight seal on the door. If the heating unit is enclosed, the area must be vented to provide outside air for combustion. Another problem with a water-heating system located in the garage is that the pipes are more vulnerable to freezing should the system malfunction or run out of fuel oil.

Plumbing check

While looking at the walls and ceiling of the garage, look for signs of plumbing leaks. Check the ceiling for water-leakage stains. If the garage has an overhead door, be sure to close the door and then look at the ceiling. In the open position, the overhead door will block about 25 percent of the ceiling; if there are leakage stains in that section, you might not see them. Ceiling stains are often caused by leakage from a bathroom above the area. (See FIG. 7-6.) When you do the interior inspection of the house, all the plumbing fixtures (sinks, bowls, and tub-shower) should be operated. After that portion of the inspection is completed, if the garage ceiling showed signs of past problems, you should reinspect it for indications of current leakage.

In some cases there are exposed drainpipes or water pipes in the garage. If your home is located in the northern part of the United States, the water pipes should be insulated as a precautionary measure against freezing.

Fig. 7-5. *Heating system (oilfired, forced hot water) boiler located in the garage.*

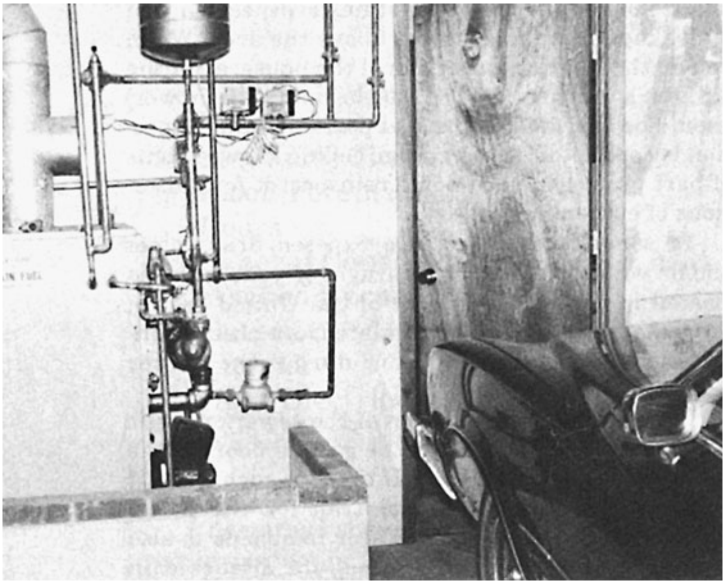

Fig. 7-6. *Plumbing leak caused water-damaged ceiling in garage.*

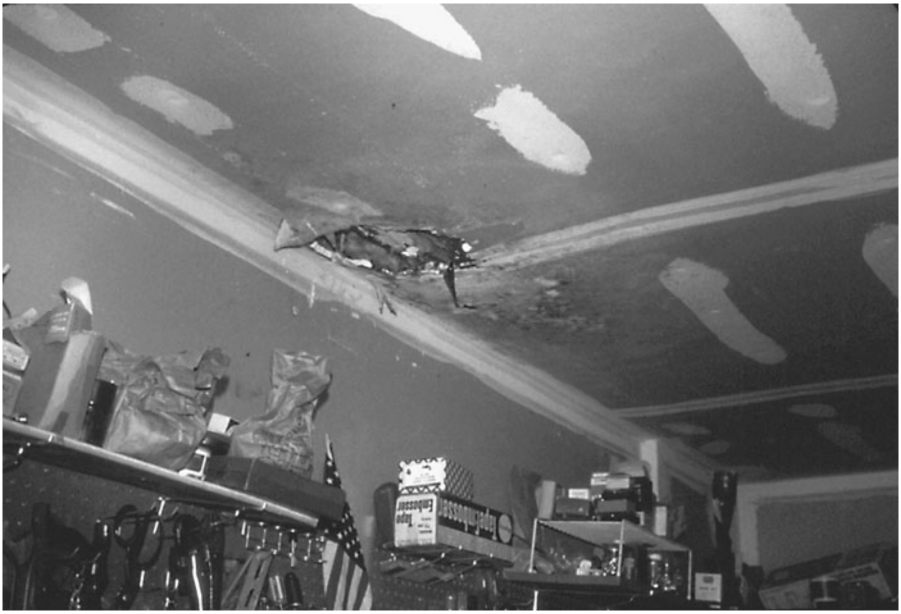

Depending on the location of the sewer or septic tank, there might be a pit in the garage floor covered with a metal plate. It might contain a cleanout and trap for the house waste line. Sometimes the water inlet pipe is also located in this pit. (These items are discussed in detail in chapter 13.) Lift the cover and look inside the pit. Often the builder neglects to remove the wood framing around the sides of the pit (used as a form when constructing the open area). Because of the dampness in the pit, a wood liner will eventually rot and might be termite-infested. (This is an area where termites are often found; see chapter 8 to learn how to determine their presence.) If there is wood in the pit, it should be removed, regardless of its condition.

Flood potential

The bottom of the pit should be relatively dry in all but very wet weather. If the bottom contains water, it is an indication that the level of the subsurface water (water table) in the overall area of the home is high. When this condition exists, there is a possibility that during rainy periods the water level can rise and seep into the garage through the pit or through cracks in the floor slab. (Water seepage into this area is discussed in detail in chapter 11.)

Cracks in the floor slab can be caused by shrinkage or differential settlement and are usually not a concerning factor. They should, however, be sealed because they can allow water to seep into the garage. An extensively cracked or heaved floor slab is of concern because it may indicate a water problem. Heaving and extensive cracking are very often caused by water pressure being exerted on the underside of the floor slab. This condition should be evaluated by a professional.

Note whether there is a drain in the floor slab. The floor should be pitched toward that drain. If there is no drain, the floor should have a slight pitch toward the automobile entry door. This will allow water from melting snow to drain to the exterior rather than puddle on the floor. Also, the floor slab should be slightly above the level of the driveway to reduce the possibility of water entry. It should be noted that a garage at the base of an inclined driveway is *always* vulnerable to water penetration. (This condition is discussed in the driveway section of chapter 4.)

Doors

When you inspect the garage, you should always check the operation of the exterior door or doors. If the door(s) does not have an automatic control, open and close each door and note whether it operates relatively easily. The most common type of door for an attached garage is the sectional overhead type. This door has the advantage of not taking up usable space when open. Look for obvious deficiencies such as broken or missing springs or guide wheels, loose and misaligned tracks, and so on. Check the door's operation. If it is difficult to lift, stuck in a fixed position, out of plumb, or does not stay in the up position, some minor maintenance is needed. When closing the door, give it a start and let it come down by itself. If the door closes rapidly and heavily, it is a hazard, especially for small children. Adjustment is needed to the spring tension, or a new spring should be installed. Check to see if there is a restraining cable that runs through the center of the spring. (See FIG. 7-7.) The cable is a safety feature that prevents the spring from whipping around and injuring a person or damaging a car in the event that the spring breaks. When the door is closed, use the lock mechanism. You might find that the lock bars need to be reset.

Many overhead doors are opened by an automatic control. Operate the control. The doors should open and close smoothly without binding in the tracks. Check to see if the control unit has a reversing feature. This is very important from a safety point of view. When the door is closing, exert a force in the upward direction at the bottom of the door. The door should stop and then reverse its downward travel. If it doesn't, record that fact on your worksheet. Repair or replacement of the control unit is recommended because it is a potential safety hazard. The newer garage door opener assemblies include a photoelectric

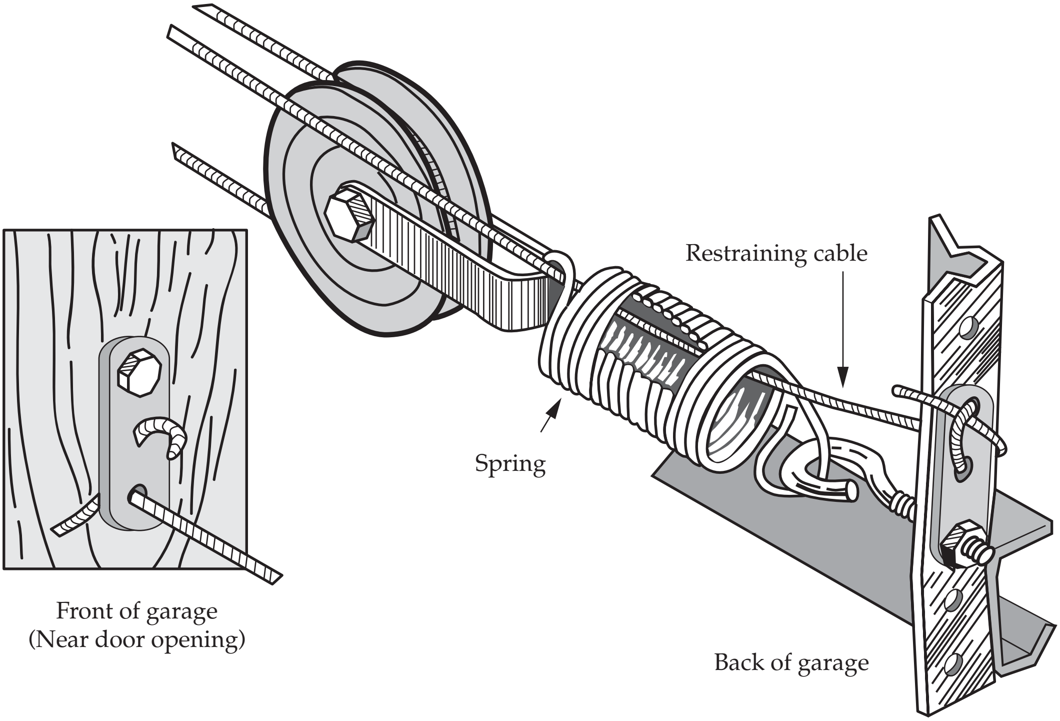

Fig. 7-7. *A restraining cable through the garage door spring prevents the spring from whipping around if it breaks.*

sensor that is mounted on both side tracks, about 12 inches above the floor. The sensor activates the controller, which will reverse the door when it is closing if a person or animal breaks the beam by passing through the opening. If the door does not reverse, it is very often because the sensor is misaligned causing the electronic beam to not contact the receiver's "bulls eye." This can be easily corrected by repositioning the sensor. Make certain that the radio controllers work. Ask the owner to demonstrate that they exist and work.

Some overhead doors are the one-piece, swing-up type rather than the sectional, roll-up type. These doors often require additional efforts to open, particularly during periods of snow and wind. Other doors found on a garage are the sliding and folding types. Sliding doors usually hang from overhead tracks. One disadvantage of such doors is that they take up valuable wall space when open. Also, small pieces of debris on the ground can interfere with their operation. Fold-out doors often sag, have loose hinges, and drag on the ground, making opening and closing quite difficult. In general, they require more frequent maintenance. These conditions can and should be corrected.

General considerations

Look around the garage for an electrical outlet. There should be at least one three-prong convenience outlet. Also, there should be overhead lights controlled by a switch near the interior door and the exterior door. Two desirable features, although not necessary, are windows that provide daylight and ventilation and a service door that can be used as access to the garage without opening the automobile entry doors.

Depending on the location of the house, the garage might be heated. Heat is usually provided by extending the central heating system (hot water, steam, or warm air) into the garage area. The radiators or heat registers providing the heat should be checked as part of the overall heating-system inspection. If there are warm air ducts in the garage, look for return grilles. There shouldn't be any because through them poisonous exhaust fumes could be brought back into the system and circulated throughout the house. When a garage is used solely for storage of automobiles, it is necessary (and not even always) to bring the temperature only above freezing. Heating the garage above that temperature is wasteful of energy.

Last, when inspecting an attached garage, look for termites in any exposed wood-frame members. The area most vulnerable to termite infestation (aside from wood found in a sewer cleanout pit) is the wood framing around the base of the exterior door.

Detached garage

If your house has a detached garage, you need not be as concerned with the fire and health hazards mentioned with the attached garage. True, the area is still considered a potential fire hazard; however, since the structure is physically apart from the main building, a fire would not usually result in the loss of life. The main concern with this type of garage is its structural integrity.

Exterior

The exterior of the detached garage is checked the same way you inspect the main house. Walk around the outside of the building twice. The first time, look at the roof and gutters. Do any of the roof beams appear to be sagging? If so, additional bracing might be needed. Have a professional make this determination. Do not assume that the roof over the garage and the roof over the house are in the same condition. Although the roof covering on the main house might be in good condition, the covering of the garage roof might be badly worn and require replacement. (Inspecting roofs is discussed in chapter 2.) Are there gutters all around the base of the roofs? If not, make sure you check all wood siding and trim for rot. The rain runoff from the roof can promote rot. A wood-frame garage with a pitched roof should have gutters. If there are long overhanging eaves or the garage is masonry-constructed, gutters are not a necessary feature, although they are often desirable. If there are gutters and downspouts, see if they need repair. (These items are discussed in chapter 3.)

After looking at the roof and gutters, walk around the building once more. This time look at the walls, windows, and doors. If the exterior walls are covered with wood siding, does the base of the siding extend to the ground? It should end about 8 inches above the ground. If the siding is in contact with the ground, it should be checked for termites and rot. Pay particular attention to the rear wall. You might see a wall that is bowed. This is usually caused by a car that did not stop in time. The wall stopped the car, and in the process, the supporting studs were broken. If such is the case, the wall is in need of rehabilitation. Also, you might sometimes see a wall that is offset; the bottom section of the wall extends about 3 feet beyond the upper section. This is done to accommodate longer cars than those for which the garage was constructed.

Finally, check the base of the wood framing and trim around the garage doors. This area is particularly vulnerable to rot and termite activity. (See FIG. 7-8.)

Interior

When entering the garage, check the doors first. Look for broken and cracked sections of wood framing and glass panes. Open and close the doors; they should operate smoothly and have all necessary hardware (see page 88).

Depending on the location of the garage, at times the entire roof is not visible from the outside. After you enter the garage, look up at the underside of the roof. If you can see daylight through a hole or crack, there is a problem with the roof. Look for signs of past water-leakage stains on the wood framing. These stains appear as dark-streaked discolorations on the wood. Leakage stains do not necessarily indicate a current leak—the problem might have been corrected. If you see stains on the wood framing, ask the homeowner whether repairs to the garage roof have been performed. Next, look at the walls. If the garage is located on an incline, look at those sections of walls that are below grade. These walls are usually constructed of brick, concrete, or rubble and also function as retaining walls. If proper drainage provisions have not been made, the walls will tend to crack and heave. (See FIG. 7-9.) If you see cracked and heaved walls, you should have a professional make a determination whether rehabilitation is required. Often the walls are covered with a stucco or plaster finish, and the wood-framing members that form the walls are not visible. However, any exposed studs and bottom plates should be checked for cracked and broken sections, rot, and termite activity.

To reduce the vulnerability of the bottom plate of a wood-frame wall to rot and termite infestation, the plate should be resting on a foundation wall that is at least 4 inches above the garage floor. In many older detached garages, this plate is found directly on the

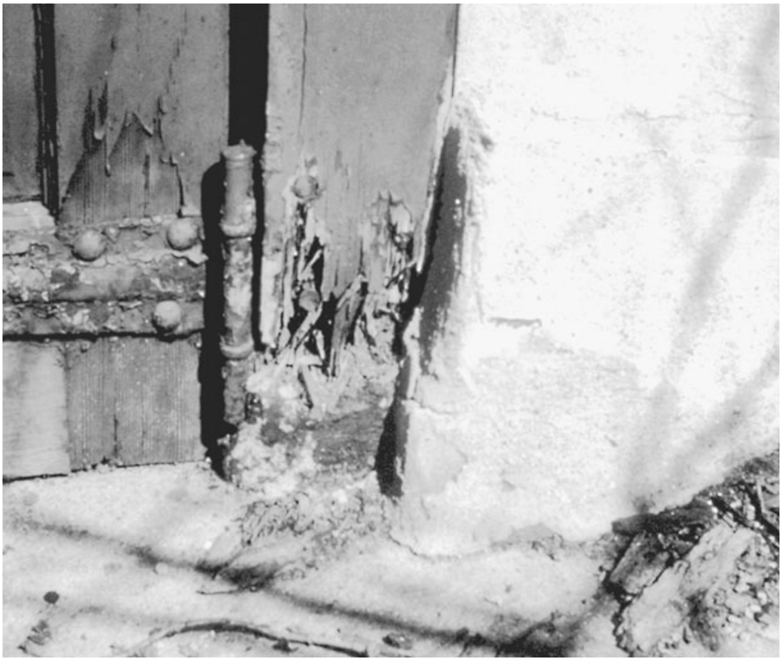

Fig. 7-8. *Termite infestation and rot at base of garage door frame.*

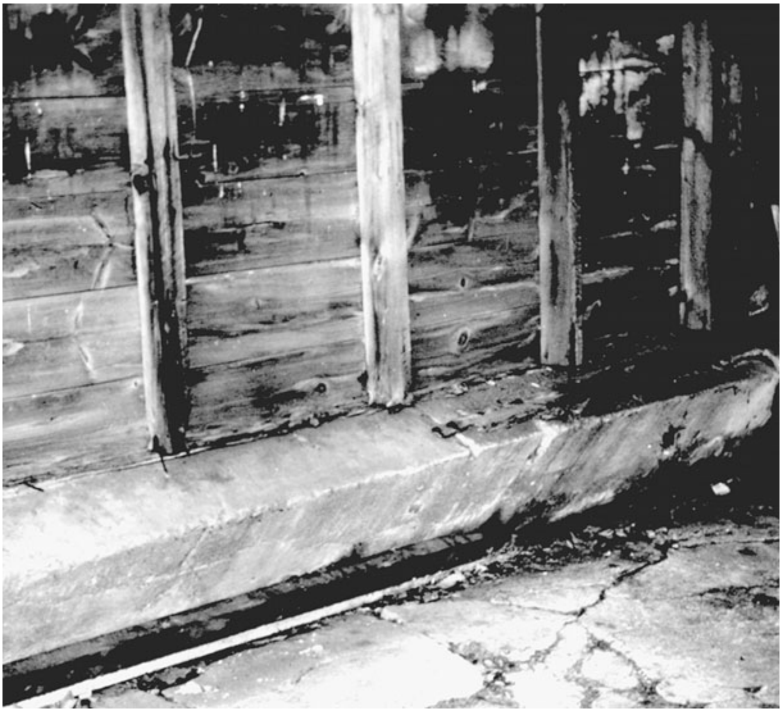

Fig. 7-9. Heaving garage wall. Section of wall was located below grade level.

floor or in contact with the ground. If this is the case, look carefully at the plate and probe it with a screwdriver or ice pick. If it can be penetrated, there is probably rot, termite, or carpenter-ant activity.

Look at the condition of the floor. If there are cracked, broken, and settled sections, often found in older detached garages, rehabilitation is in order. This condition usually does not indicate an undermining of the structural integrity of the garage but a poor installation of the floor slab. In some garages, you will find a dirt floor rather than concrete or asphalt. This is not desirable because the dampness associated with this type of floor promotes rot in the wood-framing members and premature rusting of items stored in the garage.

Heat and electricity

Most detached garages are not heated. However, when they are, heat is usually provided by a space heater rather than by extending the central heating system. The heater should be checked to see if it is operational by turning up the thermostat. The thermostat will be wall-mounted or mounted directly on the unit. Most nonelectric heaters must be vented to the outside and should not have wood framing in contact with the exhaust stack, a fire hazard. If the heater is not vented to the outside, ask the owner to show you proof that the unit has been specifically approved for installation without a flue connection.

In new detached garages, the electrical service and wiring is usually not a problem. There should be an overhead light controlled by a wall switch and at least one three-prong outlet receptacle. If the garage is a distance from the house, a desirable feature would be to have either spotlights or row lights along the path between the two structures. The lights should be controlled by two three-way switches, one at the garage and one at the house. In many older garages, the electrical wiring and service is often makeshift and nonoperational. Look around. If you see loose and hanging wires, exposed junction boxes and wire splices, you are looking at electrical violations. In some cases, the service wire from the main house to the detached garage is interior wire, not exterior. This is a potential hazard and must be corrected. If electrical problems are found during the garage inspection, you should require that the seller provide you at closing with a certificate of approval for the electrical system. The approval should be made by the municipal electrical inspection agency. (See chapter 12.)

Checkpoint summary

Attached garage

Inspecting for fire and health hazards

❍ Are the garage and basement area combined into one open area?
❍ Is the interior garage door located at least one step above the garage floor?
❍ Does this door have a tight seal? Is it self-closing?
❍ Is this door fire-resistant, or does it have a sheet-metal covering on the garage side?
❍ Is the boiler/furnace unit located in the garage?
❍ Has it been placed on a raised slab?
❍ Inspect garage ceiling and walls for exposed wood-frame members.
❍ Check ceiling area for open or missing attic access hatch.
❍ Check for return grilles in warm-air heating systems.

General considerations

❍ Inspect ceiling area for signs of plumbing leaks, stains, and patched sections.
❍ If garage is unheated, are there uninsulated water pipes that are vulnerable to freezing?
❍ Inspect floor for extensively cracked, settled, and heaved sections.
❍ Check these areas for evidence of water seepage and silt deposits.
❍ Does driveway incline make garage vulnerable to flooding?
❍ Is there a drain protecting the garage entry? Is it adequate?
❍ Does garage floor contain a drain?
❍ Inspect exterior doors and trim for cracked, missing, rotting, and insect-damaged sections.
❍ Operate doors. Note broken and missing springs, guide wheels, locks, and misaligned tracks.
❍ Is there a restraining cable running through the spring?
❍ If the overhead door is electrically controlled, does it reverse its downward travel when an upward force is exerted on the door?
❍ Check overhead lights, wall switches, and convenience outlets.

Detached garage

Exterior

❍ Inspect walls/siding for bulging, cracked, loose, missing, and rotting sections.
❍ Note broken windows and patched sections.
❍ Check roof beams for cracked, rotting, and sagging members.
❍ Inspect roof shingles (as outlined in chapter 2 checkpoint summary).
❍ Check type and condition of gutters and downspouts. Note their absence.
❍ Inspect and probe wood framing and trim around doors (particularly doors that are in contact with, or in close proximity to, the ground).

Interior

❍ Check garage doors for broken, cracked, and rotting sections.

❍ Inspect doors for operation, sagging sections, missing hardware, and broken glass panes.
❍ Inspect underside of roof for damaged sheathing and signs of leakage.
❍ Inspect foundation/retaining walls for cracked, bowed, and heaved areas.
❍ Concrete, asphalt, or dirt floor?
❍ Check concrete or asphalt floor for cracked, broken, and heaved sections.
❍ Probe wood sills for insect-infestation damage (particularly if these members are in contact with the ground).
❍ Inspect for loose and hanging electrical wires, exposed junction boxes, wire splices, extension-cord wiring, and makeshift wiring.
❍ Is there a space heater? Check operation.
❍ Is unit properly vented?
❍ Are wood-frame members in contact with the exhaust stack?

8

Wood-destroying insects and rot

Termites 95
  Subterranean termites 98
    Inspection 101
  Dry-wood termites 105
    Control 105
    Inspection 105
  Formosan termites 106
Other wood-destroying insects 106
  Carpenter ants 106
  Powder-post beetles 108
    Inspection 108
Rot 109
  Inspection 110
Checkpoint summary 112

There are many types of wood-destroying insects—subterranean and dry-wood termites; carpenter ants; and powder-post beetles. The one that causes the most damage to residential structures in the United States is the subterranean termite.

Termites

Home buyers generally overreact after discovering a termite condition and on occasion lose interest in the house. Actually, the discovery of termite infestation should not be cause for alarm—concern, maybe, but certainly not alarm.

Termites work very slowly. It takes many years for termites to do serious damage to a house. A mature colony of 60,000 termites eats the equivalent of 2 to 4 feet of 2-by-4-inch board in one year. Some well-established termite colonies have been estimated to contain more than 2 million termites. A termite condition can be controlled through the application of chemical insecticides by constructing a chemical barrier in the soil around and beneath the house. Thus termites attempting to go through the termiticide-treated soil to reach the house are either killed or repelled.

Prior to the mid 1980s, the chemical most often used for termite treatment was *chlordane*. It had an effective life that often exceeded twenty-five years. However, because chlordane is considered a potential risk to human health, it was withdrawn from the U.S. market. Currently several different chemical termiticides are available for use by pest control operators for controlling a termite infestation. All of the chemicals are considered safe and have been found to be effective in the soil for periods of approximately 5 to 10 years.

Around 1995 the termite bait system came on the market for termite control. This system is an option for homeowners that don't want to use a chemical barrier treatment for termite control. Termite baits are considerably less toxic than most liquid termiticides, which are introduced into the soil by the hundreds of gallons in order to effectively control a termite condition in a house. Baits, on the other hand, deliver very small amounts of termiticides over a long period of time.

A bait system for termite control consists of installing plastic tubes or boxes in the ground at various locations around the house. Inside the tubes and boxes is a slow-acting poison combined with a termite food material such as paper or cardboard. Termite control depends on foraging termites finding the bait stations during their random search for a new food source, feeding on it, and carrying it back to the colony where the poisoned food is shared with other termites in the colony.

The length of time for termites to find the bait stations will vary considerably and depends on whether the stations are installed in the southern or northern states. It also seems to depend on whether the bait stations are installed in the spring or late summer. The time to locate the bait has been found to vary from as little as a day to as long as a year or more. Baiting to control a termite problem is a slow long-term solution, and it is not the recommended method to control a heavy infestation problem in a house.

In real estate transactions, if the house has a termite problem, a barrier treatment is the preferred control method rather than baiting. Successful control by baiting is a long-term commitment to frequent inspections and monitoring of the bait stations and rebaiting. In contrast, when necessary, a single application of a barrier treatment can be expected to last from 5 to 10 years after which retreatment may or may not be necessary.

A number of states have regulations requiring a termite inspection by a professional prior to, or as a condition of, the purchase agreement. The cost of this inspection is almost always paid by the seller. If your state has such a requirement, you should ask the seller or real estate agent to have the house inspected by a professional and have a report of the results sent to you. In many states (even in some that do not have a prepurchase termite inspection requirement), if a termite condition is found prior to the sale, the cost for correcting the condition (chemical treatment) is borne by the seller.

When termites are discovered, they should be exterminated professionally. However, because termites work slowly, termite-proofing the house need not be done immediately upon learning of an active infestation. Take your time and get two or three cost estimates from established termite-exterminating firms. After treating a house, most companies provide a one-year guarantee against reinfestation. The guarantee can often be extended annually for a fee, which covers inspection and retreatment if necessary. If the house had been treated previously for termites, find out if the owner has a guarantee and whether it can be transferred to you.

During an inspection, all exposed wood-framing members should be checked for structural deterioration from termite activity. There are very few houses on record that have been damaged by termites to a point where they are considered unsafe. Quite often the damage caused by termites (by the time termite activity is discovered) is minor, and repair or replacement of the infested wood members is not necessary. Even with a heavy infestation, usually only a portion of the house is affected. And even then, only a portion of the wood framing might be damaged to a point where it has lost its structural value. In this case, only the affected members require repair or replacement. If you are in doubt about the structural integrity of any of the affected members, you should consult a professional.

Termites play an important role in the natural ecological cycle. They feed on cellulose, the principal ingredient of wood, and help to break down dead trees in forests and other wooded areas, thus enriching the soil. Termites began attacking houses when the wooded areas were cleared for building construction and there was no other available source of food near their nest. Subterranean termites are found in every state except Alaska. Their overall distribution within the continental United States is shown in FIG. 8-1. As their name implies, subterranean termites live in a colony (nest) that is usually located in the ground below the frost line. Even when a house is infested with termites, they usually do not have a nest in the house. They are there only to gather food. The only condition under which a nest might exist in a house (a rare occurrence) is a constant source of moisture

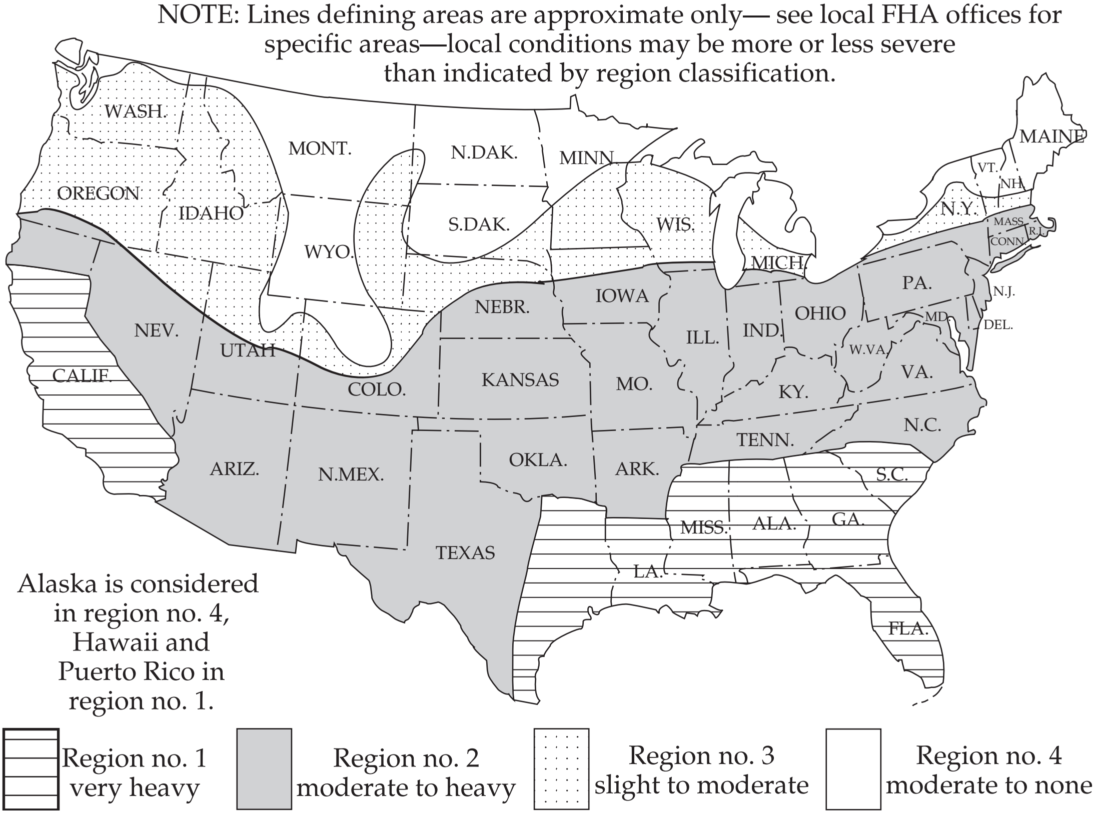
NOTE: Lines defining areas are approximate only— see local FHA offices for specific areas—local conditions may be more or less severe than indicated by region classification.
WASH.
OREGON
IDAHO
MONT.
WYO.
N.DAK.
S.DAK.
MINN.
WIS.
MICH.
MAINE
VT.
NH
N.Y.
MASS.
CONN.
R.I.
PA.
N.J.
DEL.
MD
NEBR.
IOWA
ILL.
IND.
OHIO
W.VA.
VA.
NEV.
UTAH
CALIF.
COLO.
KANSAS
MO.
KY.
N.C.
TENN.
ARIZ.
N.MEX.
OKLA.
ARK.
S.C.
MISS.
ALA.
GA.
TEXAS
LA.
FLA.
Alaska is considered in region no. 4, Hawaii and Puerto Rico in region no. 1.
Region no. 1 very heavy
Region no. 2 moderate to heavy
Region no. 3 slight to moderate
Region no. 4 moderate to none

Fig. 8-1. *Subterranean termite distribution in the United States.*

such as a leaky waterpipe or drainpipe that wets the surrounding area.

Termites are social insects. Within each colony, there is a rigid caste system consisting of a queen and king, workers, soldiers, and reproductives. Each member of the colony instinctively performs its special task. The function of the queen and king is to propagate the colony. The fertilized queen lays the eggs and might live for as long as twenty-five years. The workers care for the eggs, feed the young and the queen, and generally maintain the colony. They also forage from the nest to the wood supply and return with food. The soldiers defend the colony against attack by other insects, mostly ants. The average worker and soldier live only two or three years. The function of the reproductives is to replace the queen and king in the event of their injury or death. They also lay eggs that rapidly increase the termite population.

When a colony matures, reproductives leave the nest (swarm) to set up a new colony. Although thousands of reproductives leave the nest, only a handful survive to establish a new colony. The remainder die because of adverse conditions in the soil or attacks by other insects. Reproductive termites sprout wings for the swarm. With their wings, they are only about ½ inch long. They are considered poor fliers and generally flutter around before falling to the ground. Some, however, might be picked up in the wind and carried great distances. Once the reproductives land, they shed their wings, pair off in couples, and return to the soil in search of a suitable place to build a nest.

In most parts of the country, swarming generally occurs in the spring, sometimes in the fall. However, swarming termites have been found in January in some heated houses. In the warm, humid parts of the country, swarming can occur at any time. Even if there are no other outward signs of termite activity, termite swarming in a house is an indication that there is a healthy established colony nearby from which worker termites are coming in their search for food.

Swarming termites do not attack wood. Their only function is to start a new colony. Even if a swarm is in your house, you might not see it. A swarm might last from fifteen minutes to one hour, and if you are not in the right place at the right time, it can be over by the time you enter the room. However, if there was a swarm, you can tell by the discarded wings. They are often found on windowsills and light fixtures, and beneath doors. Do not confuse swarming termites with swarming ants. To the untrained eye, they appear similar, but there are distinctive differences. (See FIG. 8-2.) The most obvious difference is that termites have a thick waist and ants have a pinched (hourglass) waist.

Subterranean termites

Subterranean termites require a dark, damp environment. In their search for food, worker termites build shelter tubes (tunnels) that help conserve moisture and shield the termites

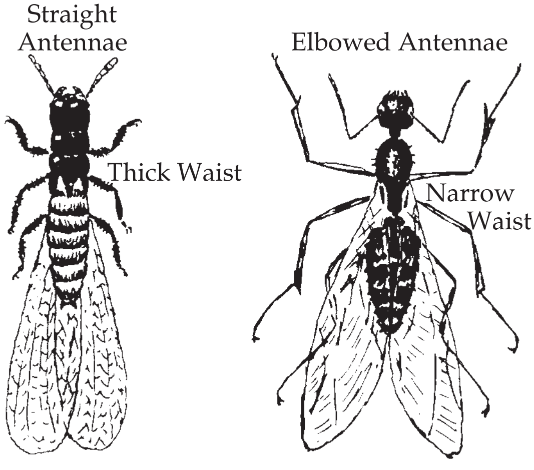

Fig. 8-2. *The difference between swarming ant (right) and swarming termite (actual size ½ inch).*

from the light. (See FIG. 8-3.) The tubes are about ¼- to ½-inch wide and provide a passageway between the ground and the food supply (wood member). They can be built at the rate of several inches per day and are mainly composed of soil, wood particles, and termite excreta. Shelter tubes, which might be noted on foundation walls, on the outside of wood framing, or even freestanding between the ground and an overhead pipe or beam, are visible evidence of termite infestation. If the tube is active, worker termites will be busy using it to go between the nest and the house. By breaking the tube, you can see the workers, who will try to repair the break. They are about ¼-inch long and have a whitish cream coloring.

Some tubes might be abandoned. If you find an inactive shelter tube, it does not mean that termites are no longer in the house. It might, if the house has been termite-proofed. However, if it has not, even an abandoned shelter tube, no matter how small, is sufficient evidence to consider termite treatment. Many shelter tubes emanate from a nest. There might be an active tube inside the voids of a concrete block wall that would not be visible during an inspection.

Whether a new building will be attacked by termites depends on the surrounding area and to a large extent the builder. Certain construction practices tend to increase the probability of termite attack. (See FIG. 8-4.) Some builders have been known to bury tree stumps and wood debris near the foundation or below the basement-floor slab. All stumps and debris should be removed from the building site. All form boards and scrap lumber should be removed before the excavated area around the foundation walls is backfilled. There should be no buried wood around the house. Otherwise, it can provide a source of food for a new termite colony that when it becomes large enough, will attack the house.

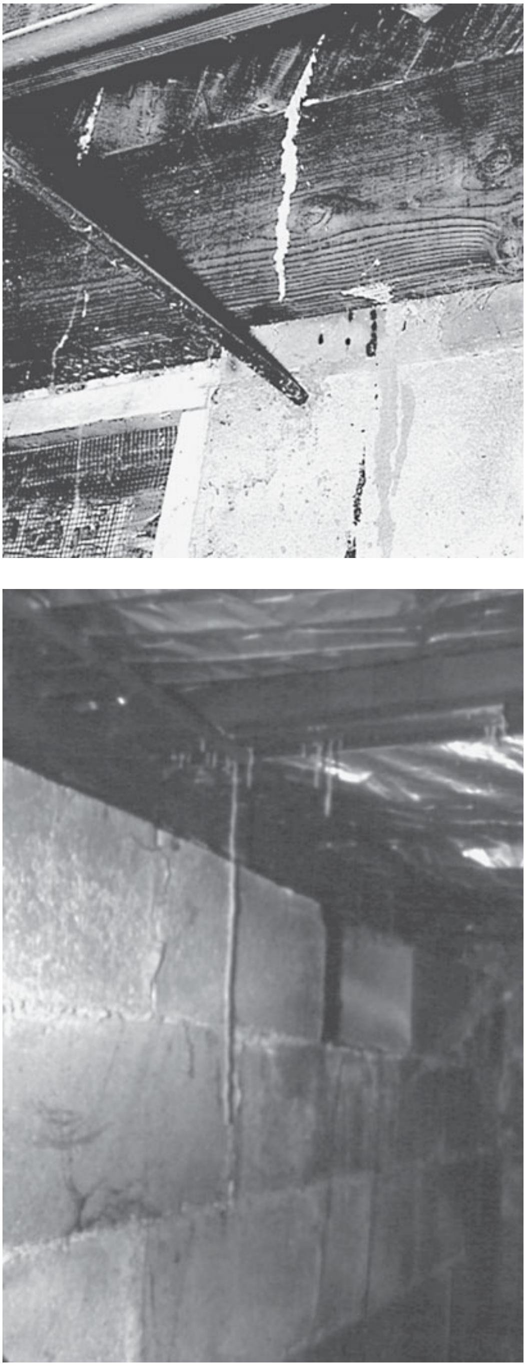

Fig. 8-3. *Termite shelter tubes: on foundation wall, header, and subflooring (top); hanging tubes in crawl space (bottom).*

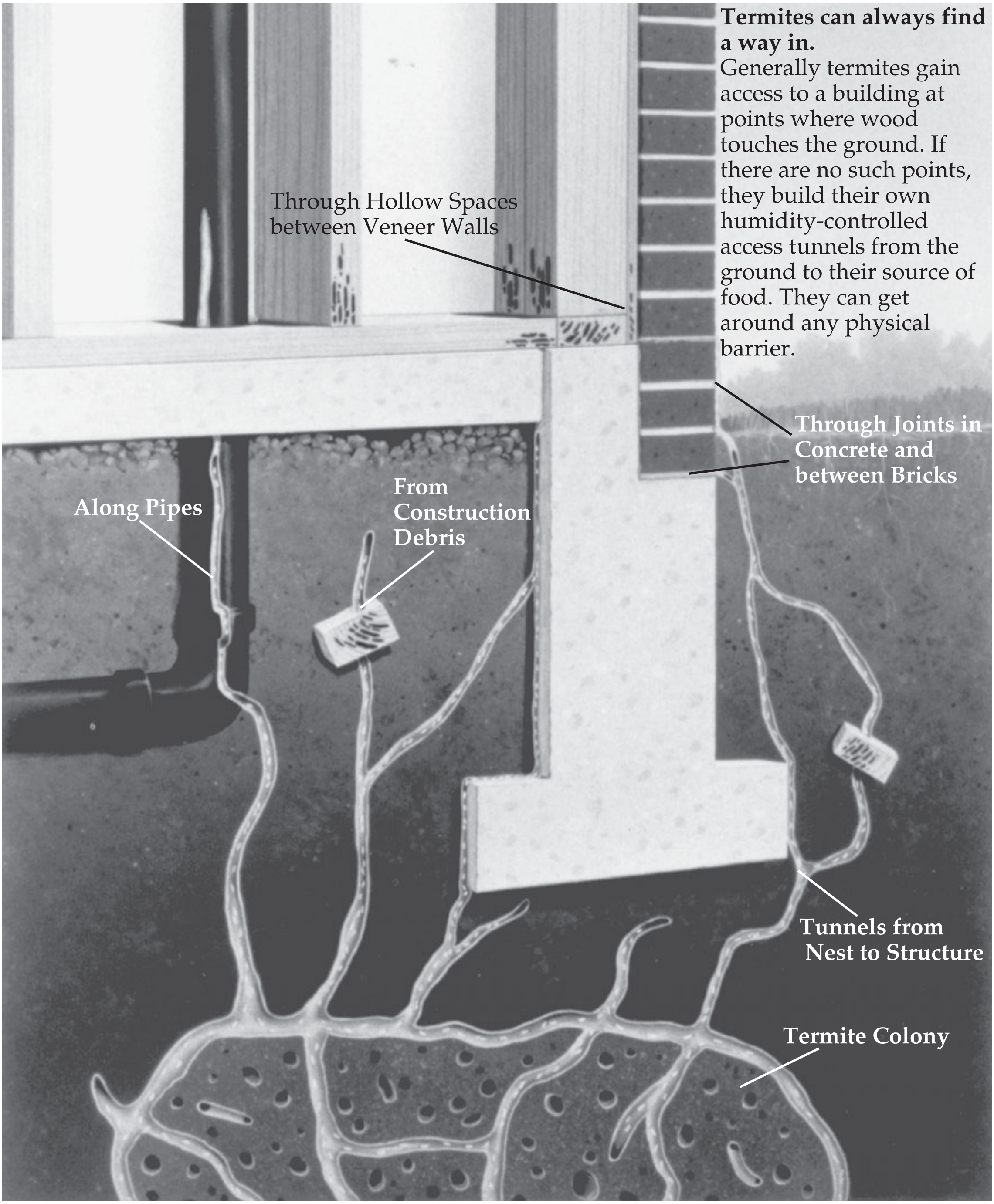

Termites can always find a way in.
Generally termites gain access to a building at points where wood touches the ground. If there are no such points, they build their own humidity-controlled access tunnels from the ground to their source of food. They can get around any physical barrier.

Through Hollow Spaces between Veneer Walls
Through Joints in Concrete and between Bricks
Along Pipes
From Construction Debris
Tunnels from Nest to Structure
Termite Colony

Fig. 8-4. Termite entry points into a house.

Most often termites enter a house by eating their way through untreated wood members that are in direct contact with the ground. Some of the more common points of entry are garage door frames, basement windowsills, wooden steps and supports, wood sills, and headers and studs on foundation walls that are located at or below grade. A particular area of attack is the wood framing adjacent to a concrete-covered, earth-filled porch, patio, or entrance slab. (See FIG. 8-5.) If there is no earth or wood contact, termites can build shelter tubes to provide passageways from the nest to the wood framing in the structure.

Some homes have a strip of metal (termite shield) between the foundation wall and the sill plate, that rests on top of the foundation wall. The purpose of the termite shield is to act as a barrier between the nest and the food supply. In most cases, the termite shield gives the homeowner a false sense of security. The shield does not prevent infestation. It only deters an attack. The problem is that the termite shield is rarely installed properly. An opening at a seam or a hole as small as 1⁄32 inch is large enough for termites to pass through. All seams should be soldered, and any holes around bolts and pipes should be filled with coal-tar pitch. Even if you see a termite shield, you should look for termite infestation.

Inspection A complete subterranean termite inspection consists of an interior and exterior check of that portion of the house that is close to, or in contact with, the

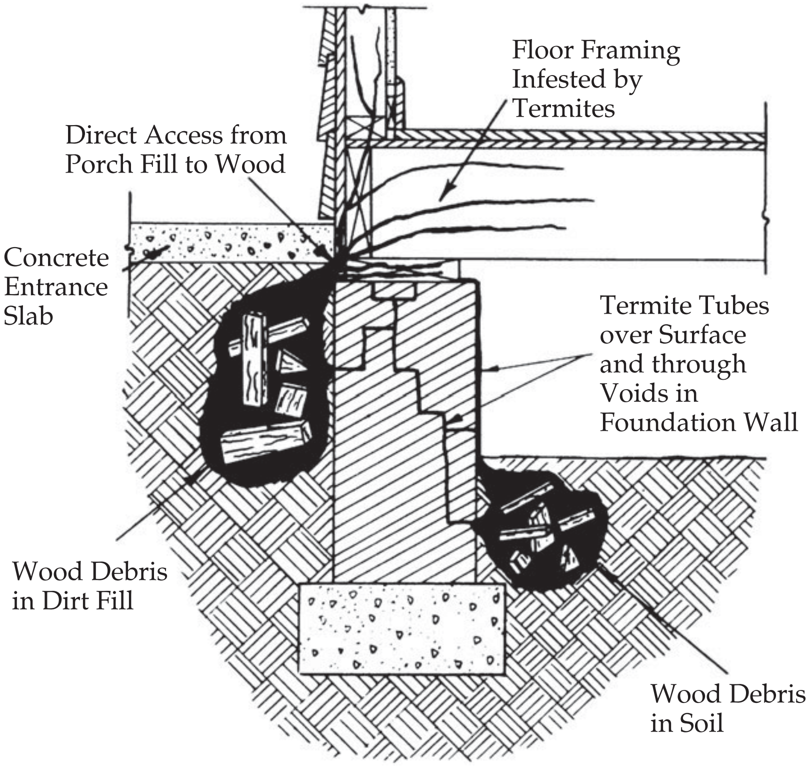
Floor Framing
Infested by
Termites
Direct Access from
Porch Fill to Wood
Concrete
Entrance
Slab
Termite Tubes
over Surface
and through
Voids in
Foundation Wall
Wood Debris
in Dirt Fill
Wood Debris
in Soil

Fig. 8-5. *Termite colonies can develop in buried wood debris and gain entrance into a building, particularly at earth-filled concrete entrance slabs or patios.*

ground. The exterior termite inspection can be performed concurrently with the normal exterior inspection as described in chapter 1. As you walk around the outside of the house, look for termite shelter tubes along the outside foundation walls. (See FIG. 8-6.) In many homes, this area is covered or partially blocked by shrubbery. Part the shrubbery so that you can see the wall. This is especially important for areas just below a garden-hose spigot.

In some cases the base of the exterior wood siding is in contact with the ground, so that the foundation wall is not visible. This is a poor construction practice but unfortunately is fairly common and usually occurs during final grading and landscaping. The base of the wood siding should terminate at least 6 inches above the finished grade. Redwood or cedar exterior siding is often used. Both types of siding are resistant to termite attack and rot. However, it is important to understand that they are not immune to attack and might eventually succumb. If the base of the siding is in contact with the ground, probe it with an ice pick or a screwdriver. If the wood has not been attacked by rot, termites, or other wood-destroying insects, your probe will not penetrate much beyond the surface. If the probe penetrates the wood deeply, the wood has been attacked. To determine the cause for the deterioration, it is necessary to break open a section and look at the condition of the wood. Caution: Do not proceed beyond the probing without consent of the homeowner; in fact, it would be wise to obtain the homeowner's consent for the probing.

A section of termite-damaged wood reveals *galleries* (channels) that run parallel with the grain. (See FIG. 8-7.) Termites attack the softer portion of the wood grown during the spring and not the denser summer wood. The channels will not look polished, as they do with carpenter ants. Portions will be lined with grayish specks that consist of excrement and earth. Deteriorated wood could be the result of a combination of causes such as rot, termites, and carpenter ants. You should become familiar with the telltale signs. If you

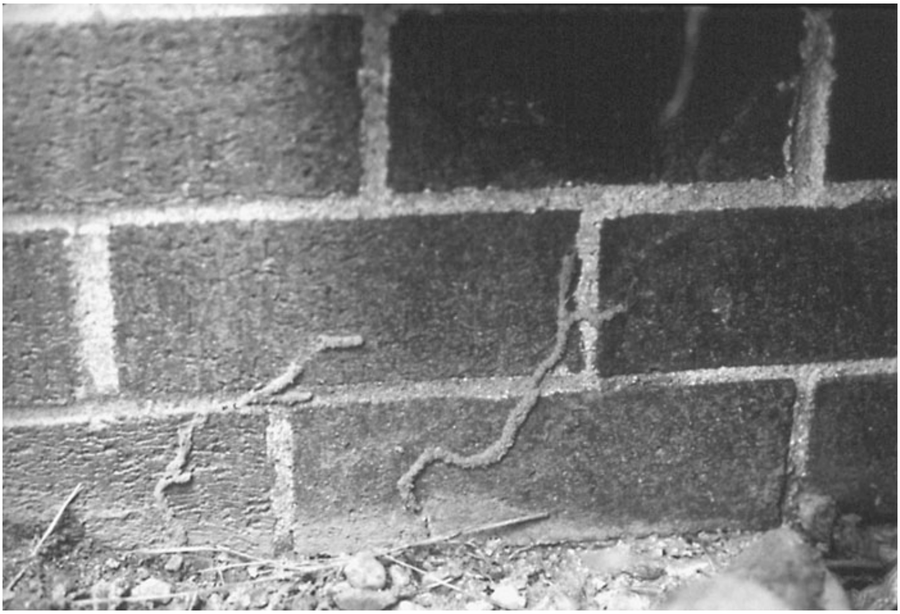

Fig. 8-6. *Termite shelter tubes at base of exterior wall.*

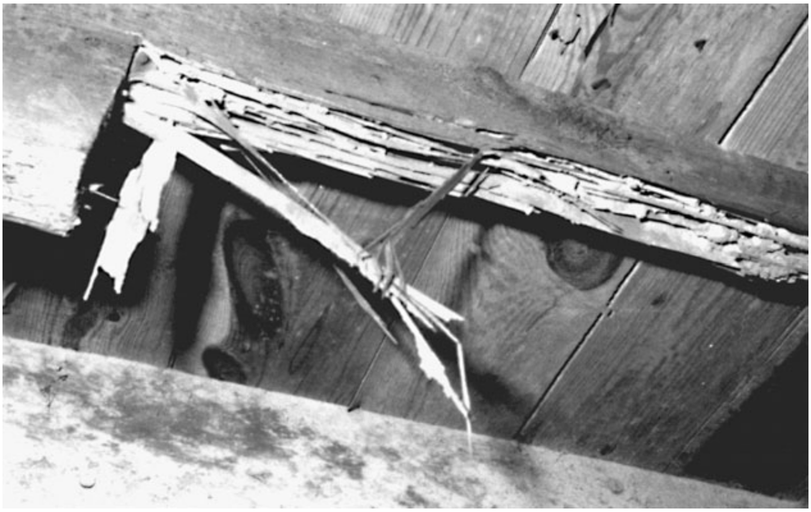

Fig. 8-7. *Termite-damaged wood. Note galleries that run parallel with the grain.*

are not certain of the cause, have the wood evaluated by a professional. The appearance of wood damaged by carpenter ants, powder-post beetles, and rot will be discussed in their respective sections.

As you walk around the outside of the house, probe the attached wood trim, posts, and framing members that are on or close to the ground. Specifically probe garage door frames, basement or lower-level window frames, step stringers, deck posts, and the entry-door riser. Termite activity is of concern only when it is found in the house or in an attached structure such as a garage or deck. If you find termite damage or shelter tubes on a fence post in the yard (FIG. 8-8) or in a piece of wood debris on the ground in the yard, all that means is that those pieces of wood have had termite infestation. It does not mean that the house is infested with termites and should be treated.

After checking the outside of the house, the next place to look is in the crawl space beneath the house or porch. (Not all houses have a crawl space, in which case this step is omitted.) If the access to the crawl space is from the interior, it should be checked as part of the interior termite inspection. In the crawl space, probe the sills and headers for termite damage. Also check the first 15 inches of each

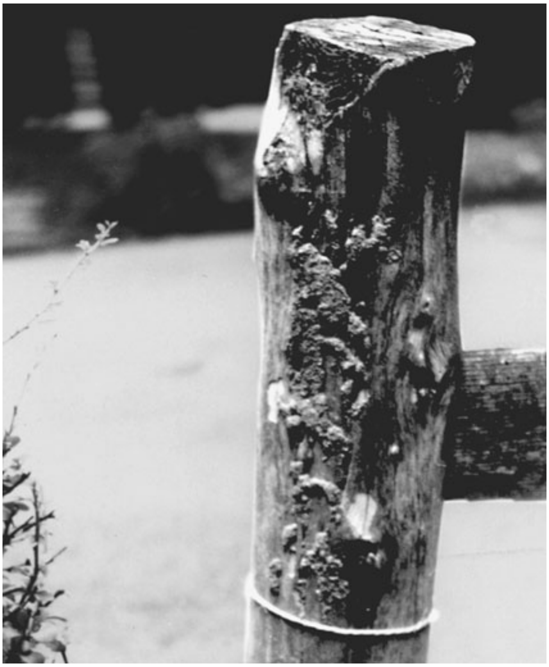

Fig. 8-8. *Termite shelter tubes on a fence post.*

joist that rests on the sill or foundation. As you move around in the crawl space, look for termite shelter tubes on the foundation walls and piers. Sometimes the shelter tubes can be spotted between *double joists* (two joists nailed together, used to provide additional support for a heavy load). While in the crawl space, note any items of wood storage, remembering that they are vulnerable to infestation.

The interior inspection for subterranean termites is generally conducted in the basement and crawl space. If the basement is finished so that there are no exposed sections of foundation wall or wood framing, a thorough inspection cannot be performed, even by a professional. There are, however, some sections more vulnerable to termite attack than others, such as sill plates, headers, and joists below grade or adjacent to a dirt-filled, cement-covered patio. If the ceiling is covered with suspended tiles, the tiles can be lifted or moved to expose the wood framing. If there are no accessible areas, termite activity will have to be determined by a swarm or exterior inspection. In areas where termite infestation is heavy, such as the South and Southwest, a termite inspection should be performed every year. In other areas, an inspection every two year is adequate.

When the basement is unfinished, the wood framing (sill plate) on top of the foundation wall should be inspected by probing as described for the exterior crawl space. (See FIG. 8-9.) Also look for termite shelter tubes. You might find a tube that appears to start in the center of the foundation wall. Actually, the tube is connected to the earth through a small crack in the wall at that point. As previously mentioned, termites can work their way through cracks as small as 1⁄32 inch. While in the basement, pay particular attention to the area around the furnace/boiler and pipes through the foundation or floor slab. Often activity is discovered in these areas.

Homes built on a slab without a basement or crawl space are also vulnerable to attack by termites. Because there are generally no areas with exposed beams or foundation, detecting termites is quite difficult unless the infestation has advanced to the point where shelter tubes are

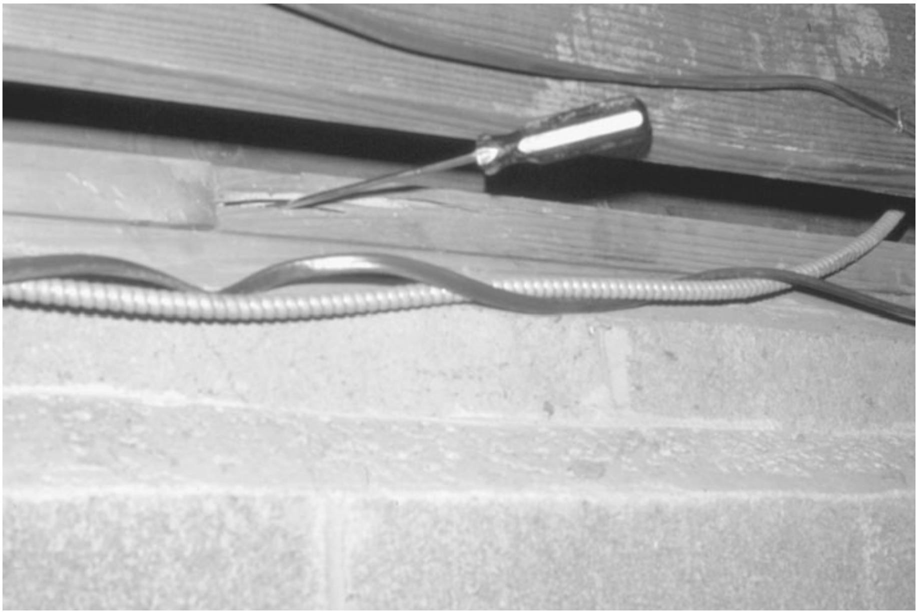

Fig. 8-9. *Termite-damaged sill plate found by probing with a screwdriver.*

visible in the finished rooms. When inspecting a slab house, look for soft spots in the baseboard trim along the exterior wall. Also check for shelter tubes around openings in the floor slab, such as around plumbing or heating pipes.

Dry-wood termites

Homes in the South and Southwest are vulnerable to attack by dry-wood termites as well as subterranean termites. Dry-wood termites cannot live outdoors in northern climates and have not established themselves in those areas. Isolated cases of dry-wood termite infestation have been found in homes located as far north as New York and Ohio, but those are rare.

Dry-wood termites are so called because they build their nest in perfectly good wood that is not decayed and not in contact with the ground. In fact, in many cases they establish a colony in the wood-framing members of the attic. As with subterranean termites, reproductive dry-wood termites swarm from the nest periodically in an attempt to establish a new colony. The swarming termites often do not fly more than a few feet before settling down. However, if aided by air currents, they can fly more than a mile. Once paired, the king and queen seek cracks or checks in nearby wood, whether a roof or lumber pile, and set up housekeeping. In homes, dry-wood termites can be found in rafters, studs, joists, sheathing, floorboards, window frames, door frames, and exterior trim.

Once a colony is established, it feeds on the wood around the nest. Dry-wood termites are general feeders and eat spring- and summer-grown wood. The galleries thus formed will cut *across* the grain. The cavities contain pellets of partially digested wood. These pellets are tiny, seedlike, and usually straw-colored. On occasion, some of the pellets are pushed through openings in the wood surface. If there is not much accumulation, the pellets can easily be overlooked. They are, however, often the first sign of infestation.

The maximum population of a fully established dry-wood termite colony is estimated to be about 3,000. This is considerably less than the typical number of subterranean termites in an established colony, which is estimated to be between 60,000 and 250,000. Consequently, it takes a longer period of time for serious damage to occur with dry-wood termites than with subterranean termites.

Control Since dry-wood termites do not nest in the ground, they must be chemically treated at the source of infestation. This is usually done by injecting insecticide into the galleries, a procedure that should be performed by a professional. To reach the galleries, holes are drilled in the infested wood members. The insecticide will be a liquid or a dust. When the wood is dry, dusting can be effective as far as 15 feet from the point of application. However, when the wood is wet, dusting usually is not effective because the dust tends to cake in the moist galleries. If dry-wood infestation is found in a fence or pole on the property, those wood members should be treated since the termites represent a potential source of infestation for the house.

For severe infestation, treatment is usually by fumigation. The entire house, including the roof, is wrapped with a plastic covering. After all the openings are sealed, a poisonous fumigant is introduced. The house should remain under fumigation for at least forty-eight hours. Since the insecticide is poisonous to humans, fumigation should be undertaken only by experienced fumigators.

Inspection Since dry-wood termites do not leave their nest in search of food, there are no telltale signs of infestation such as shelter tubes. They can be detected, however, by their pellets, which tend to accumulate in a small pile after being pushed from the wood by the termites or falling through a crack in the infested wood. An infestation in the house can also be detected by the presence of a swarm if you happen to be in the room during an occurrence. Since dry-wood termites can attack wood located anywhere in the house, from the attic to the crawl space, all the exposed wood should be checked for signs of infestation. The wood should be gently probed, so as not to break the surface. Infested wood has hollow sections and if heavily probed, can break open, spilling the seedlike pellets. Dry-wood termites often consume wood up to the paint itself, forming what appears to be a paint blister. If the slightest pressure is applied to the blister, it can break. Care should be exercised to maintain the integrity of the wood surface; a broken gallery is difficult to treat with insecticide.

Formosan termites

The Formosan termite is a subterranean termite native to the Far East. It was first discovered in the United States in a Houston, Texas, shipyard in 1965 and has since spread to a number of southern coastal and Gulf states. All indications are that Formosan termites might eventually establish nests in some northern states as long as the temperature and moisture conditions are satisfactory.

An established Formosan termite colony, which can have more than a million termites, is extremely destructive. They can destroy wood six times faster than our native species. Formosan termites have been known to penetrate lead, plastic, rubber, mortar, and plaster to get to their food—wood. They are able to penetrate the above materials by secreting an acid substance from their frontal glands. In at least one recorded case, they have caused short circuits by damaging electric cables.

Identifying the worker Formosan termite is difficult because there are no obvious characteristics distinguishing them from our native species. However, in the latter colony, only 2 percent of the population are soldiers; whereas with the Formosan's, 25 percent are soldiers. As a result, if an active Formosan termite shelter tube is broken or an infested section of framing is examined, many more soldiers would be visible than if it was from our native species.

Native subterranean termites ordinarily have a ground connection and live partly in ground and partly in wood. However, the Formosan termite can live without a ground connection if there is a suitably located constant source of moisture. This characteristic is what makes the control of Formosan termites so difficult. For example, because of a plumbing leak, the termites might have a nest in an exterior wall. Consequently, chemical treatment of the soil around the house, which is the control procedure for native subterranean termites, will not exterminate the termites in the wall. Fumigation is needed.

Since the location of a termite nest is normally not known—it can be in the ground or in a wall—to control an infestation, it would be necessary to treat the soil around the house chemically and fumigate. With a single detached building, this combination of control procedures is possible; however, in urban areas with row houses and attached structures, fumigation is not a viable procedure. Unfortunately for these buildings, the condition cannot be corrected.

Other wood-destroying insects

Carpenter ants

Carpenter ants are also social insects. They live in colonies with a rigid caste system consisting of a queen, workers, and reproductives (sexually mature males and females that periodically swarm and set up new colonies).

Although worker ants can live four to seven years and the queen for as long as fifteen, colonies have been known to last for thirty to forty years. When the queen dies or is accidentally killed, specially fed workers take over the egg-laying function.

Carpenter ants differ from termites in that they do not eat wood. They merely excavate it to build a nest. The nest consists of irregularly shaped galleries that generally follow the grain. The small fragments of shredded wood that are generated during the excavation are removed from the galleries and deposited on the outside. Consequently, the galleries do not have the earthy appearance of the termite galleries. Rather, they have a polished or sandpapered appearance.

A carpenter-ant colony can be located on the ground in a decaying log or tree trunk or in the roof framing of a house. The ants also nest high in trees and can fly from there to set up new colonies in a house. They build their nests in a variety of locations, preferring wood that is moist or softened by decay. However, they will also build their nests in wood that is perfectly dry and sound.

When inspecting for carpenter ants, look specifically at sections of wood that have begun to decay as a result of a past or current moisture condition. Even though the source of the moisture might have been eliminated (as by correcting a leak), an ant colony might have been established already. Typical locations to inspect are portions of the wood framing, siding, or trim that are in contact with the ground; wood that has been dampened by the overflow from defective roof gutters; the area around a damaged section of siding or flashing; the base of hollow porch posts and columns, and areas with large open joints as might occasionally be found around exterior windows and doors. These areas should be probed with a screwdriver or an ice pick. If the wood yields, breaks, or cracks and ants come crawling out, there is a good chance that you have located a nest.

Consider yourself lucky if you do, because a carpenter ant nest is usually quite difficult to locate; it is often established in an inaccessible location in the wall or roof assembly. One indication of the existence of a colony is unexplained piles of sawdust. (See FIG. 8-10.) Some people, however, think that piles of sawdust are an indication of termites. They are not. Subterranean termites completely devour the wood that they are attacking and leave absolutely no trace of wood particles. Drywood termites also eat the wood completely. However, they do drop tiny, well-formed seedlike pellets. If the pellets are observed

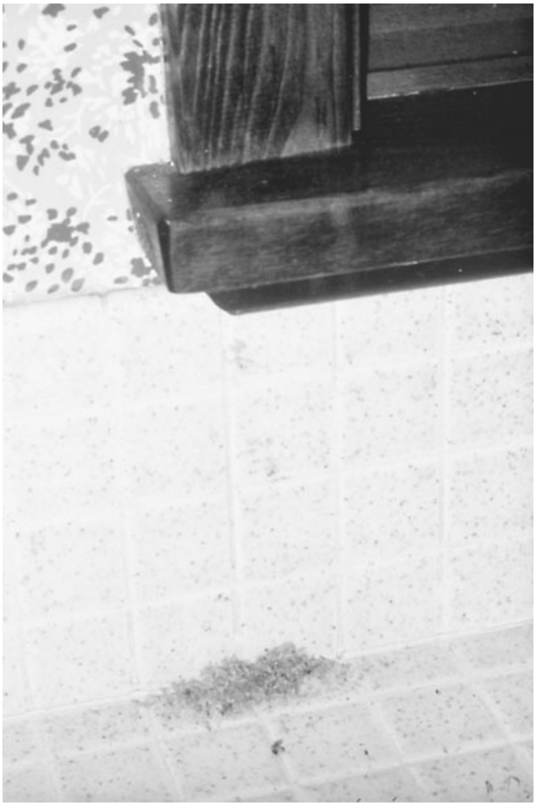

Fig. 8-10. *Sawdust is an indication of carpenter ant activity in window trim.*

closely, they can be differentiated easily from irregularly shaped particles of sawdust.

When a house is infested with carpenter ants, there is little likelihood that the people living there are unaware of the condition. Numerous worker ants will be seen walking around the rooms as if they live there—which indeed they do, with free room and board. These ants feed on sweets, crumbs, and other foodstuffs normally found or spilled on a kitchen counter or floor. Carpenter ants are easy to recognize. They are among the largest ants in the United States, worker ants varying in size between ¼ to ½ inch long. They are black or black with a reddish brown midsection.

While the first sign of infestation is usually the presence of carpenter ants in the house, the fact that they are there does not mean that the nest is inside the house. It might be outdoors, and the ants may have entered the house foraging for food. A carpenter-ant infestation can be controlled only by destroying the nest, either directly or indirectly. Nests can sometimes be located by watching the ant traffic. Ants continually entering and leaving an area are generally an indication of the nest location. If the nest is found, it can be treated directly with insecticide. If not, dusts or sprays can be used where the ants are commonly seen. The latter might not eliminate the infestation, but it should reduce it.

Powder-post beetles

There are many types of wood-boring beetles. The ones whose larvae or grubs feed on seasoned wood and break it down to a powdery residue are commonly called powder-post beetles. These beetles exist all over the United States, although the greatest concentration will be found in those states with a warm, humid climate. The two principal varieties of powder-post beetles are the lyctid and the anobiid beetles. The lyctid beetle attacks only hardwoods; the anobiid beetle attacks both soft and hardwood timbers.

For the most part, powder-post beetles are usually brought into the house via the wood that had been used in its construction. Building materials might become infested while being stockpiled in the lumberyard. The insects might also be brought into the house in finished wood products such as oak flooring, paneling, and furniture.

The beetles lay eggs in the open pores, cracks, and crevices in the surface of unfinished wood. After the eggs hatch, the larvae feed and tunnel their way through the wood, reducing it to a powder. Depending on the temperature and moisture content of the wood, the larval stage can be as short as a few months or as long as a few years. Just prior to emerging, the newly formed adults chew small round exit holes in the wood surface (1⁄32 to ⅛ inch). In the process of emerging, finely powdered wood called *frass* is usually pushed out in front of the body. This is often the first external sign of infestation. Shortly after emerging, the beetles mate and lay eggs. They occasionally deposit the eggs in the mouth of an old exit hole, thereby reinfesting the same piece of wood. Some wood members have shown signs of extensive damage as a result of infestation by several generations of beetles. (See FIG. 8-11.)

On occasion the homeowner might be the one responsible for the powder-post beetle infestation in the house. Under natural conditions, the beetles breed in dead branches and limbs of trees. When gathering wood for the fireplace, it is possible to pick up infested pieces and store them in the basement or under the stairs for later use. If wood is left in storage through the following spring and summer, the emerging beetles might attack unfinished lumber such as girders, joists, studs, sill plates, and subflooring.

Inspection Inspection for powder-post beetles should be performed along with the

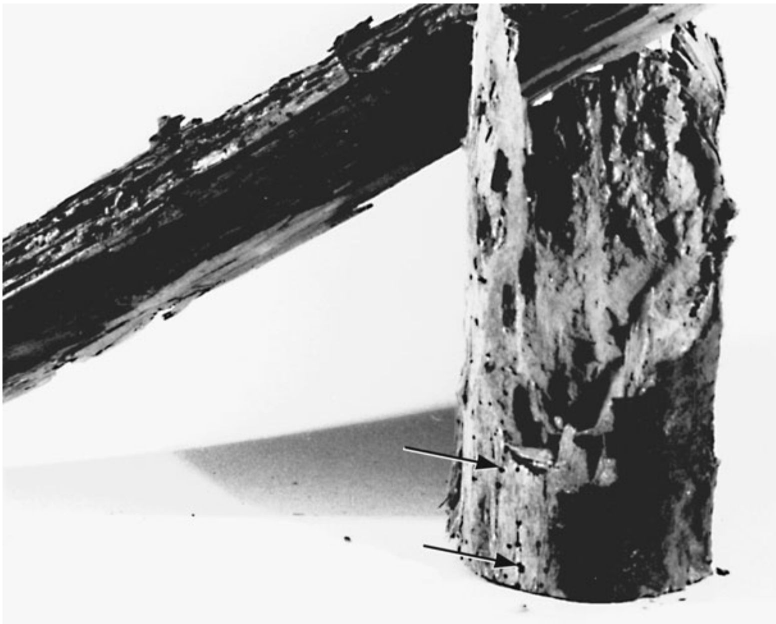

Fig. 8-11. *Wood post with extensive deterioration caused by powder-post beetle infestation. Note beetle exit holes at the lower section of the post.*

inspection for termites. When probing the exposed wood-framing members, look for the small round emergence holes of the beetles. Since it is possible for the beetles to emerge without reinfesting the wood, the fact that there are emergence holes does not mean that the wood member is currently infested. Newly formed flight holes are light and clean in appearance, like a fresh saw cut; older holes are darker in color. If the infestation is well established, there will usually be more than thirty exit holes per square foot of surface. Even though the wood might no longer be infested, small amounts of larvae frass might continue to sift through the holes for many years as a result of normal vibrations of the wood. Look at the frass. If the infestation is no longer active, it will have a yellowish appearance or will be caked.

If there is any doubt about whether the infestation is active, call a professional pest-control operator for an evaluation. An infestation in a single wood-framing member can often be controlled by replacing that piece of wood or coating it with an appropriate insecticide. However, if the infestation is widespread, chemical treatment by a professional is necessary.

Rot

Wood products used in construction are susceptible to decay (rot). However, if properly maintained, they can easily last for hundreds of years. There are three basic types of fungi that attack wood: *stain*, *mold*, and *decay*. Wood rot is caused by an attack of the decay fungi. Stain and mold fungi mainly grow on the wood surface, causing discoloration. By themselves, they do not weaken the wood; but their presence does indicate a moisture problem and should serve as a warning of conditions favorable to the growth of the decay fungi. The decay fungi are microscopic threadlike plants that grow in the wood and attack its thick cell walls. They break down the walls and feed on the contents of the cells. With the destruction of the cells, the wood disintegrates, and decay becomes evident.

In the early stages, it is difficult to recognize that a section of wood has been attacked by the decay fungus. The wood might merely be discolored. However, the advanced stages of decay are easily recognizable because the wood undergoes changes in properties and appearance. The affected wood might be brownish and crumbly or white and spongy. In either case, the decay greatly reduces the strength and structural value of the wood member. The brown, crumbly rotted sections readily break into small cubes and in the final stage of deterioration are often quite dry. Most people when seeing this condition refer to it as "dry rot." This is really a misnomer. The actual decay occurred when the wood was wet since decay fungi cannot survive in dry wood.

Sometimes the physical changes in the wood are not apparent on the surface. They can be detected easily, however, by probing the wood with an ice pick or a screwdriver. If the wood is in good condition, the probe will not penetrate much beyond the surface. However, if the wood has deteriorated, the probe will easily penetrate into the wood. The conditions that promote the growth of the decay fungi also promote subterranean-termite and carpenter-ant activity. Consequently, when probing wood-framing members, you might find deteriorated sections that are caused by a combination of insect damage and decay. Under suitable conditions of temperature and humidity, the decay fungus gives rise to a fruiting body that contains enormous numbers of microscopic spores. The spores are the seeds of a new generation of decay fungi and are readily distributed by air currents. The spores are always present in the air and under normal conditions cannot be kept away from wood. The presence of decay-fungi spores on wood is of no concern unless the moisture content of the wood and the temperature are such that the spores will germinate and grow. Figure 8-12 shows the decay hazard zones in the United States.

Decay fungi will grow and develop only when the moisture content of the wood is in excess of 20 percent and the temperature is in a range from 40° F to 115° F. Temperatures above the upper limit kill the decay fungi; temperatures below the lower limit cause the fungi to become dormant. The latter is a condition that readily occurs in the northern states during the winter months. In the spring, when the temperature rises, the decay fungi in infested lumber resumes growth, assuming the moisture content of the wood has not changed. The moisture content of wood is defined as the weight of water in the wood expressed as a percentage of the weight of the wood when oven-dry. The decay fungi thrive when the moisture content is about 25 percent. However, when the wood is saturated with moisture, the decay fungi are inhibited from growing because of the lack of oxygen.

The moisture content of green lumber can be as high as 200 percent; after the lumber is kiln-dried, its moisture content may be as low as 7 to 10 percent. Wood in a house that has been properly constructed and maintained will seldom have a moisture content over 15 percent. However, once the moisture content exceeds 20 percent, the wood becomes vulnerable to deterioration by the decay fungi. If the building design is such that some of the wood must be subjected to damp or wet conditions, those sections should be treated with toxic chemicals to prevent decay or be made from the heartwood of certain species (cypress, cedar, or redwood) that are resistant to rot.

Inspection

As with the inspection for wood-destroying insects, the inspection for rot should be conducted on the exterior and interior of the

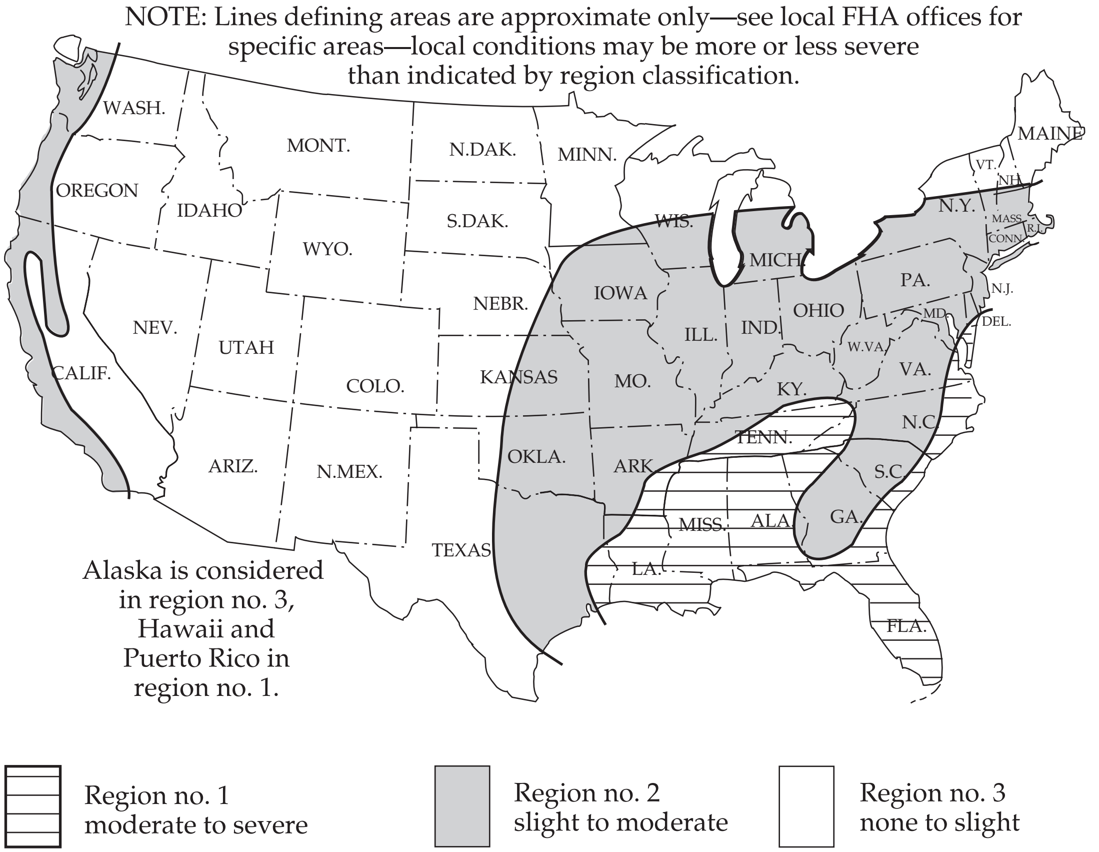
NOTE: Lines defining areas are approximate only—see local FHA offices for specific areas—local conditions may be more or less severe than indicated by region classification.
WASH.
MONT.
N.DAK.
MINN.
MAINE
VT.
NH
OREGON
IDAHO
S.DAK.
WIS.
N.Y.
MASS.
R.I.
CONN.
WYO.
MICH.
PA.
IOWA
NEBR.
OHIO
N.J.
MD
DEL.
NEV.
UTAH
ILL.
IND.
W.VA.
CALIF.
COLO.
KANSAS
MO.
KY.
VA.
N.C.
TENN.
OKLA.
ARK.
S.C.
ARIZ.
N.MEX.
MISS.
ALA.
GA.
TEXAS
LA.
FLA.
Alaska is considered in region no. 3, Hawaii and Puerto Rico in region no. 1.
Region no. 1 moderate to severe
Region no. 2 slight to moderate
Region no. 3 none to slight

Fig. 8-12. *Decay hazard regions in the United States.*

structure. It can be performed concurrently with the inspection for termites. Areas that are vulnerable to attack by subterranean termites or carpenter ants are also conducive to the growth of the decay fungi. During the exterior inspection, in addition to probing the wood that is located near, or is in contact with, the ground, you should check wood members having cracks and open joints that are subjected to periodic wetting from rain. One area particularly vulnerable to decay is the end cut of exterior wood framing, trim, or siding. An exposed end cut (across the grain) absorbs water much more readily than a section that has been cut parallel with the grain.

During the interior inspection, you should check the unfinished attic and basement or crawl space for evidence of rot. In the attic the wood members can rot as a result of water intrusion through the roof because of a faulty roof covering or leakage around the joints of roof projections. Conditions conducive to rot can also result from condensation due to inadequate attic ventilation. In the basement or crawl space, the decay fungi thrives in the wood members if the humidity is constantly high. Check the overhead wood framing and subflooring for signs of rot. Probe the wood members above and around the top of the foundation wall and base of wood posts. Vulnerable locations for decay are the wood members through which plumbing pipes pass—because of the possibility of leakage or condensation. Probe the joists and subflooring below kitchen or bathroom fixtures.

When inspecting for rot, note on your worksheet all wood-framing members that have decayed to a point where they can no longer provide structural support. These sections should be replaced or rehabilitated. If in doubt, consult a professional. If natural decay-resistant wood or preservative-treated wood is not used for the replacement of rotted sections and the source of the moisture has not been eliminated, then it will only be a matter of time before the new sections begin to decay. Rotted wood trim that serves no structural function need not be replaced except for cosmetic reasons. Once the infected wood has been dried out and the source of the wetting has been eliminated, the decay is permanently arrested.

Checkpoint summary

Subterranean termites

Exterior inspection

❍ Check all exterior areas of the structure that have wood in contact with, or close proximity to, the ground.
❍ Note any termite shelter tubes on the foundation walls.
❍ Probe vulnerable areas such as garage door frames, basement windowsills and frames, deck posts, step stringers, and entry-door risers.
❍ Probe wood-frame members adjacent to concrete-covered, earth-filled porches.
❍ Inspect crawl areas under steps, porches, and so on.
❍ Probe sills and headers.
❍ Check wood fencing, dead tree stumps, wood debris, or stored firewood in close proximity to the house for infestation and rot.

Interior inspection

❍ Inspect for shelter tubes on foundation walls and piers and around all plumbing pipes that pass through the foundation.
❍ Pay particular attention to areas around the heating system.
❍ Probe exposed sill plates, headers, joists, and girders.
❍ Inspect wood support posts for infestation and rot.
❍ If the house is built on a slab, note any soft spots in the baseboard trim.

Dry-wood termites

❍ Inspect property fencing for infestation.
❍ Probe exposed wood framing throughout the house, from attic to crawl space.

Carpenter ants

❍ Look for small piles of sawdust below or around wood members.
❍ Did you see any ants walking around in the rooms, particularly the kitchen?
❍ Probe the sections of wood framing, siding, and trim that show evidence of decay or past wetting.

Powder-post beetles

❍ Inspect wood framing for clusters of small round holes.
❍ Newly formed holes are the color of a fresh saw cut and indicate an active infestation.
❍ Probe these wood sections for deterioration.

Rot

❍ Probe vulnerable areas such as wood members that are subject to periodic wetting from rain or garden sprinklers.

❍ Inspect roof sheathing from the attic for decaying sections around chimney, vents, and so on.

❍ Check subflooring and support joists below kitchen and bathroom fixtures and around plumbing pipes.

❍ Probe sill plates, headers, and the ends of joists and girders.

9

Attic

Insulation 115
Violations 116
Leakage 117
Fire hazards 117
Ventilation 118
Attic fan 120
Whole house fan 120
Structural 121
Checkpoint summary 123

Once you are inside the house, if there is an accessible attic, it should be the first area inspected. There are basically two types of attic: full and crawl.

A *full* attic is one in which a person can easily walk around. Usually there is a floor in this type of attic, although the walls and ceilings are unfinished. There might be partition walls forming finished rooms with sloping or horizontal ceilings. Access to a full attic is usually through a finished staircase.

In a *crawl* attic, which is completely unfinished, the roof is sufficiently close to the floor so that to get around it is necessary to crawl or stoop over. The crawl attic usually does not have a structural floor. The ceiling joists from the level below are exposed. When getting around in this type of attic, be careful to walk only on the exposed joists. If you accidentally step between the joists, you will probably frighten the pants off of anyone in the room below because your foot will go right through the ceiling. Access to a crawl attic is usually through a ceiling hatch located in a closet or hallway or through hidden folding or sliding steps.

Inspecting an attic can reveal problems of which most homeowners are not aware, some of which might be potentially dangerous or costly to repair. It is not uncommon for a homeowner to say that he has lived in his house for over twenty years and has never gone into the attic. Incidentally, if the homeowner does say this, you can be sure that at the very least, the attic is inadequately insulated by current energy standards.

Insulation and roof leakage are probably the only items most people consider when thinking about the attic. However, there are other items of importance and concern, such as ventilation and its associated problems, fire hazards, electrical and plumbing violations, improperly discharging vents, and open duct joints.

Insulation

The attic area should be adequately insulated to minimize heat loss. The insulation needed in the attic will of course depend on the geographic location of the structure. (See chapter 19 to determine the proper insulation for your area and for a general description of the various types of insulation.) In both crawl and unfinished full attics, the insulation should be located in the floor (between the floor joists) and not between the roof rafters. (See FIG. 9-1.) Otherwise, heat from the rooms below will escape into the attic.

The insulation should be installed with a vapor barrier facing the heated portion of the structure and not the unfinished attic area. A *vapor barrier* is aluminum foil, a plastic sheet, or an asphalt-impregnated paper that prevents moisture movement from the heated portion of the house into the unfinished attic area. If the vapor barrier is incorrectly positioned (facing up into the unheated, unfinished attic), condensation problems can develop during cool weather. Moisture rising from the heated areas below condenses upon contacting the cool vapor barrier. Depending on the amount of vapor, the resulting condensation buildup can reduce the effectiveness of the insulation and cause peeling and flaking of the painted ceilings and walls in the rooms below.

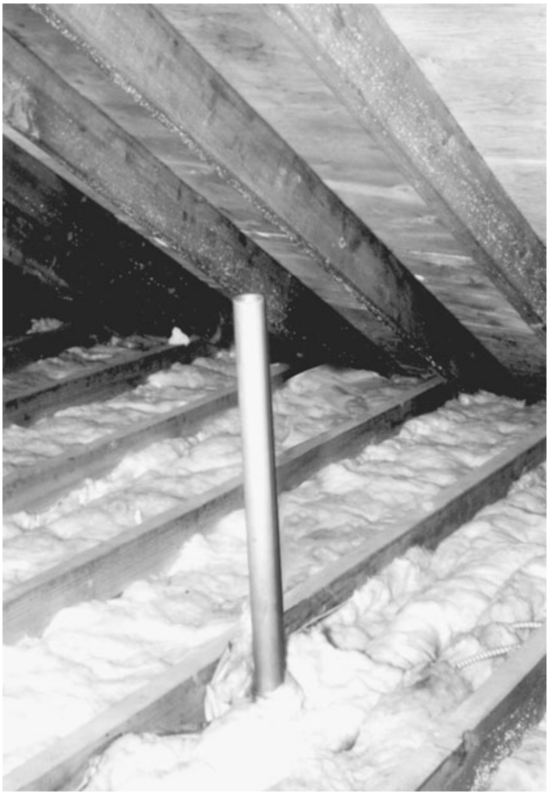

Fig. 9-1. *The proper installation of insulation in the attic, between the floor joists and not the roof rafters. Note plumbing vent stack terminating in the attic—a violation of the plumbing code.*

If there are heating or air-conditioning ducts in the attic, check to see whether they are insulated. Metal ducts often have insulation in the inside, so tap the duct with your flashlight. If you hear a hollow sound, there is no insulation. If there is a dull thud, insulation is present. The more insulation on a duct, the less the heat loss during the winter and, in air-conditioning ducts, the less the heat gain during the summer. Generally, a minimum of 3 inches of insulation wrapped around the outside of the duct will substitute for missing insulation. If the duct is used for air-conditioning, the insulation should be covered with a vapor barrier to prevent condensation from forming.

In the northern sections of the country, when there is a furnace in the attic, insulation between the roof rafters as well as in the attic floor is recommended. To ensure proper ventilation and avoid condensation problems between the rafters, it is important to leave a ventilated air space between the top of the insulation and the underside of the roof deck. The insulation and the associated air space will help keep the roof deck cool. If the insulation is not between the rafters, heat from the furnace will warm the roof deck and melt the bottom layer of snow that has accumulated on the roof. This often results in an ice dam (see chapter 2).

Often, a homeowner adds additional insulation to the attic to bring the total insulation up to current energy standards. The insulation added should not have a vapor barrier if the existing insulation has one. All too often the homeowner adds insulation with a vapor barrier, which can then cause condensation problems. Look at the insulation in the floor. If there are two layers of insulation and both have vapor barriers, the upper barrier should be slit with a razor blade to allow moisture movement.

If there is a full attic with partition walls forming rooms, the insulation should be located on the unfinished sides of the partition walls and on the ceilings of the rooms. Occasionally, insulation is placed between the roof rafters and not between the floor joists and the partition walls. This installation is inefficient because the heat will escape into the unfinished areas.

Violations

When inspecting the attic, look for vent stacks that terminate in the attic area. (See FIG. 9-1.) This is a violation of the plumbing code, a condition that should be corrected. The vent stack should extend through the roof so that the sewer gases can discharge to the outside. In the past, according to code, every vent stack in the attic terminated above the roof. This requirement can now be modified if the local municipality approves it. With the advent of air admittance valves (AAVs), at least one vent stack must terminate above the roof, as long as all the other vent stacks in the attic are capped with an AAV. This valve is a pressure-actuated one-way mechanical vent that is designed to maintain a plumbing fixture's trap seal and also prevent sewer gases from discharging into the attic. The AAV (see FIG. 9-2) opens only under negative pressure, which is created when a toilet is flushed or when water is flowing down the drain. The valve closes by gravity

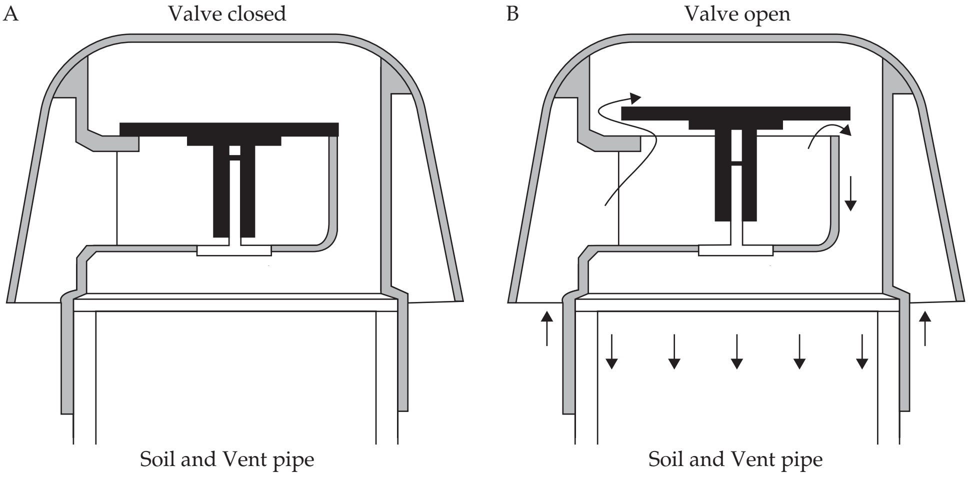

Fig. 9-2. *Air admittance valve (AAV). A—During times of nonactivity in the drainage system, the diaphragm forms a tight seal to close the valve. B—When a toilet is flushed, a negative pressure develops in the pipe, which causes the diaphragm seal to open and air to enter the vent pipe.*

or slight spring pressure to form a tight seal when the water flow stops, or when there is a positive pressure, which would be the result of a blockage in the drain. AAVs come in various designs; however, they all operate under the same principle.

While in the attic, also look for ducts. Whether they are air-conditioning ducts or ducts from an exhaust fan, they should not have any open joints. (See FIG. 9-3.) All open joints should be resecured. Sometimes the exhaust fan from a bathroom discharges its moisture-laden air into the attic. This is undesirable because the moisture can cause condensation problems. The duct from such an exhaust fan should be extended above the roofline so that the exhaust is discharged into the atmosphere. Also look for open electrical junction boxes and makeshift electrical wiring such as extension-cord wiring and "pigtailed" hanging light fixtures. These are electrical violations that should be corrected.

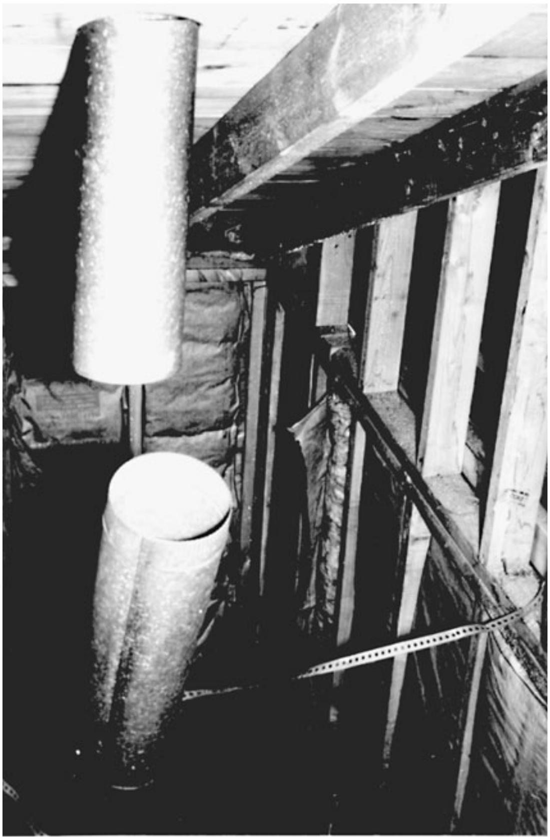

***Fig. 9-3.** Open joint in duct from kitchen exhaust fan.*

Leakage

It is important to check the underside of the roof for signs of past water leakage and, if the structure is inspected during a rain, current water leakage. Water stains can show up on the sheathing or roof rafters and appear as dark-streaked discolorations on the wood. Sometimes there is a separate masonry chimney for the fireplace and a prefabricated chimney for the heating system. Joints vulnerable to water leakage are those between the chimney and the roof and between vent stacks and the roof. These joints should be checked for leakage. Water leakage through joints is a relatively minor problem and can usually be corrected by sealing the joints with an asphalt cement.

Fire hazards

Of particular concern is the joint between a prefabricated chimney and the attic floor. (See FIG. 9-4.) According to building codes, there should be 2 inches minimum clearance between the chimney and adjacent wood framing. The clearance is a fire safety measure because wood, which normally burns at temperatures between 400 and 600°F, can ignite spontaneously at a reduced temperature of about 200°F if it has been exposed over the years to temperatures between 150 and 250°F.

If your house has a prefabricated chimney, check the clearance between it and adjacent wood framing. While the clearance space around the chimney prevents a problem, it also creates one because the open space around the chimney generally runs from the

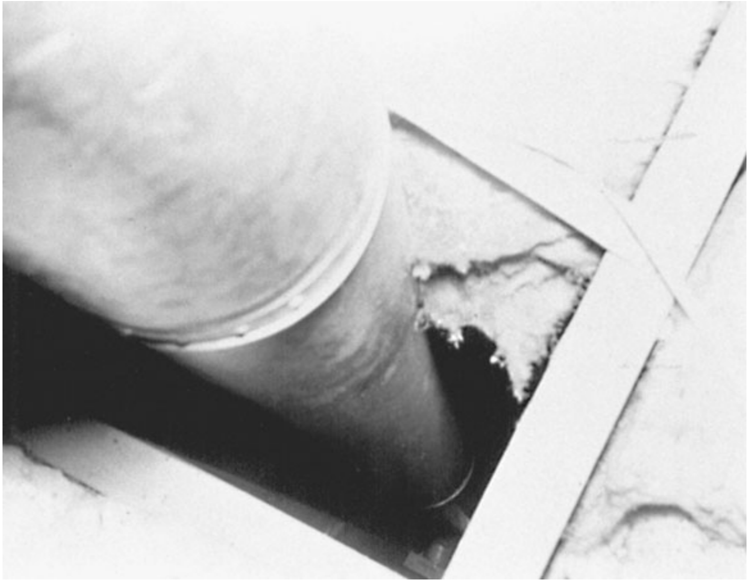

Fig. 9-4. *Open joint between chimney and attic floor—a potential fire hazard.*

boiler/furnace room to the attic. If a fire should develop in the boiler/furnace room, the open area around the chimney will act as a flue and draw the flames up into the attic where they can very rapidly consume the structure. Fortunately, the condition can easily and inexpensively be corrected by blocking (fire-stopping) the opening with a noncombustible material such as sheet metal.

If the house has a fireplace or wood-burning stove, check the section of the chimney that is exposed in the attic. Look for soot or a creosote buildup around the joints. This indicates cracked or open joints through which the exhaust smoke is seeping into the attic. This problem exists more often with chimneys that have unlined flues than those with lined flues. This condition is a potential fire hazard that must be corrected. Record this fact on your worksheet.

Ventilation

Ventilation in an attic is very important. It allows moisture that accumulates in the area to dissipate and also helps to reduce the heat buildup that normally develops during the summer months. Some of the moisture generated in the structure by bathing and cooking works its way up into the attic. If the area is inadequately ventilated, this moisture can eventually cause problems such as delaminating roof sheathing, water streaks on interior walls, peeling and flaking paint, and in severe cases some rotting of the wood framing. When walking around the attic, make sure that all vents are completely unblocked. Sometimes the homeowner blocks a vent opening (see FIG. 9-5) to cut down on the cold air entering the attic and reduce the heat loss through that opening. What he doesn't realize is that he is creating a problem. If the attic is properly insulated, the heat loss through the vent openings is minimal. Of greater importance is the need for vent openings so that moisture can escape. Look around; you'll be able to tell whether the attic is adequately ventilated.

Look at the roofing nails projecting through the sheathing. If you happen to be in the attic on a winter day when the temperature is around 20° F and the area is inadequately

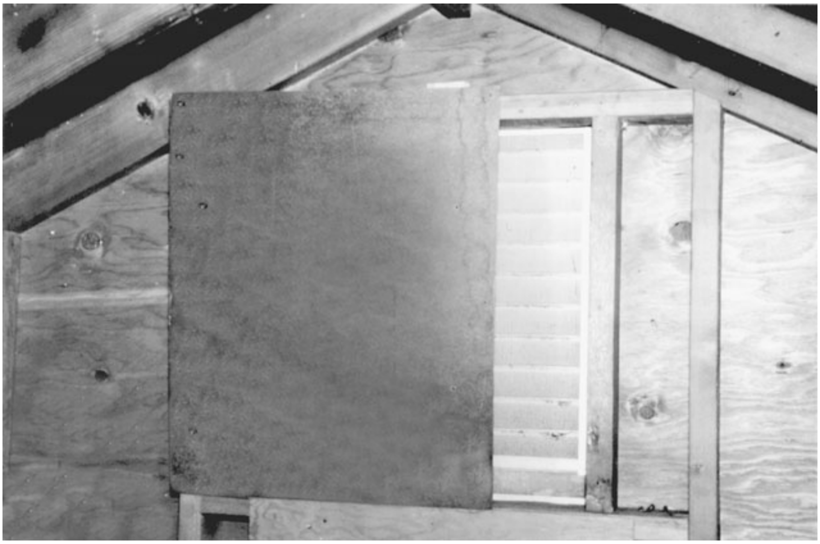

***Fig. 9-5.** Partially blocked gable louvers. Louvers are used to ventilate the attic.*

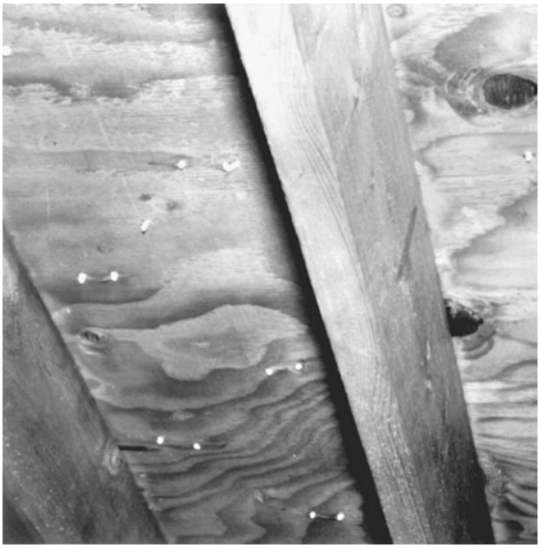

***Fig. 9-6.** Frost on roofing nails due to inadequate attic ventilation.*

ventilated, you will find frost on the roofing nails. (See FIG. 9-6.) The frost that forms on the nails melts and drips onto the floor during the warmer periods of the day. Many droplets of water show up as circular stains on the floorboards and on top of the insulation. (See FIG. 9-7.) If you are in the attic during the warmer months, you'll find rust stains and mildew on the sheathing near the nails. (See FIG. 9-8.)

When inspecting the attic, pay particular attention to the north slope of the roof. If the area is inadequately ventilated and plywood is used for the roof sheathing, delamination of the sheathing might be a problem. The northerly slope begins to delaminate before the southerly slope. Figure 9-9 shows an advanced stage of delamination as a result of moisture buildup in the attic. At this stage the entire roof sheathing and shingles must be replaced. The condition could have been avoided if the attic area had been adequately ventilated. Correcting an inadequately ventilated attic is relatively easy. All that is required is to increase the size of the existing vent openings (if there are any) or to provide additional openings such as roof vents, ridge vents, soffit vents, or a power ventilator. These vent openings can be used individually or in combination. The exact number of vent openings needed should be determined by a professional.

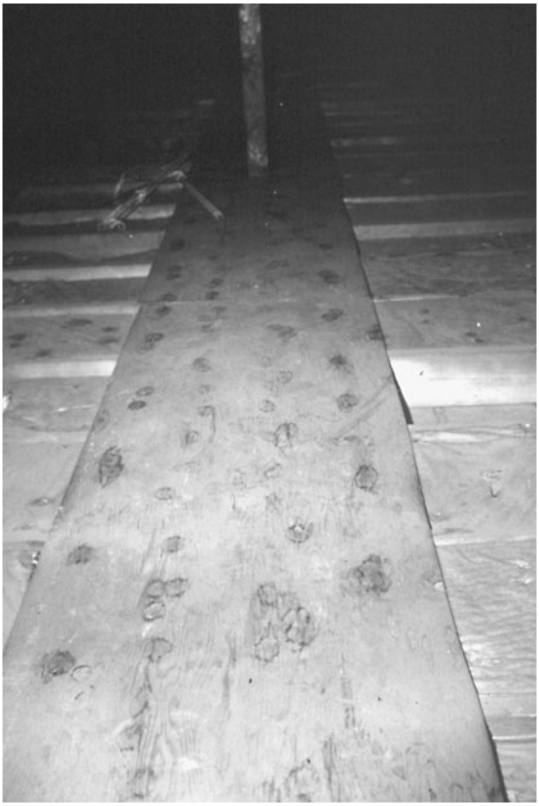

Fig. 9-7. *Water-droplet stains on catwalk in attic due to condensation on protruding roofing nails because of inadequate attic ventilation.*

Attic fan

If there is a fan in the attic, it is considered a desirable feature. (See FIG. 9-10.) Because of trapped air, an attic area can reach 150° F on a hot summer day. An attic fan that is thermostatically controlled greatly reduces the heat load on the building. Usually the thermostat is set to activate the fan when the attic temperature is about 100° F. Along with a thermostatic control, some attic fans also have an additional humidistat control. The humidistat turns the fan on and off regardless of temperature whenever the relative humidity is too high in the attic. The fan provides two additional benefits regarding air conditioning. By reducing the heat load on the structure, less electrical energy is required for air-conditioning units. And for those homes with a central air-conditioning blower coil located in the attic, the unit operates more efficiently at lower surrounding temperatures.

The attic fan should normally be operating during the warm days of summer. If it isn't, either the unit is malfunctioning or the thermostat is improperly set. On many units the thermostat is exposed and can be manually adjusted. But if the thermostat is factory set and not accessible, the unit will require professional maintenance. On cool but not cold days, you can usually check the attic fan, assuming the thermostat is exposed, by lowering the temperature control setting until the fan is activated.

Whole house fans

Some structures have whole house fans, which draw outside air in through screened windows and doors and blow it out through the opening around the fan or movable louvered vents. Check the fan before you go into the attic. If the fan is turned on while you are in the attic, be careful when walking around. In some homes, the area of the vent openings for the fan is too small for the size of the fan; consequently, the air that is moved by the fan is partially blocked from leaving the structure. As a result, pressure is built up in the attic, which decreases the efficiency of the fan. If the fan is turned on before you go into the attic, put your hand over any electrical outlet box located on the level just below the attic. If you feel air rushing onto your hand, the attic vent openings are inadequate and should be increased. Some fans are controlled by manual snap switches. From a convenience point of view, this is not as desirable as a timer switch or a thermostat switch. If this is the case, consider its replacement.

Whole house fans are mounted either in the attic floor or on the exterior sidewall of the attic.

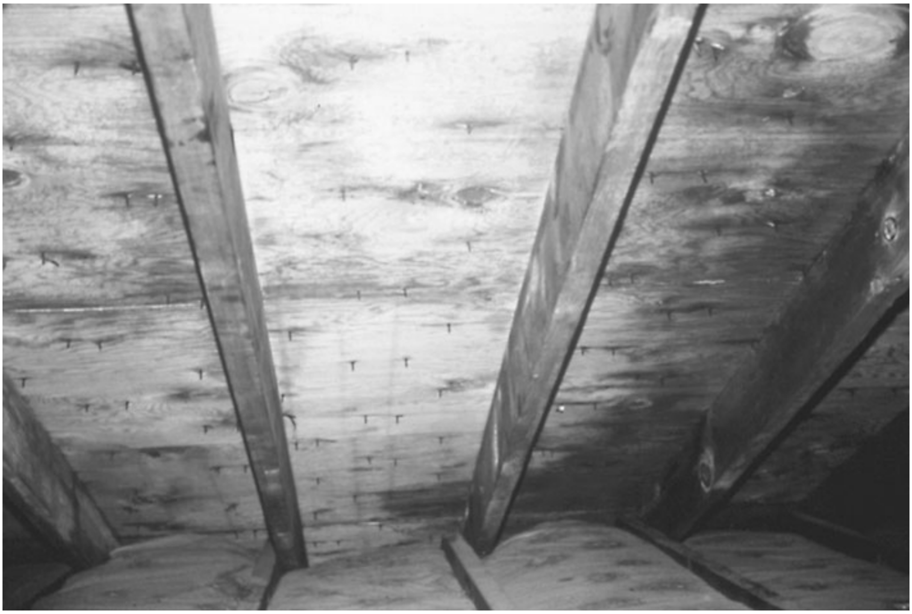

Fig. 9-8. *Mildew on underside of roofing deck.*

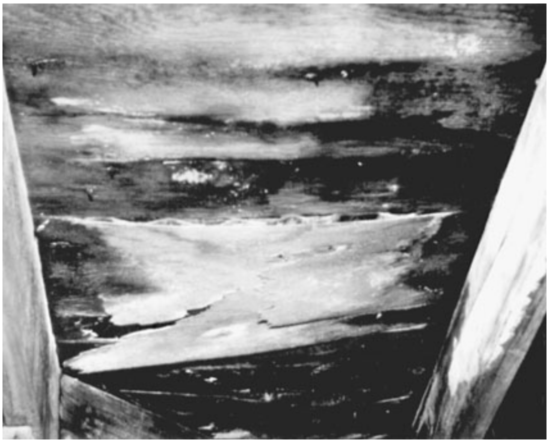

Fig. 9-9. *Delamination of roof sheathing resulting from inadequate ventilation of attic.*

They have movable louvers that open when the fan is activated and remain closed when the fan is off. This operation creates a potential problem. During the winter months and during portions of the summer months when the fan is not operating, the louvers will be in a closed position and will thereby completely close off the vent opening, causing the attic to be inadequately ventilated. This problem can be avoided by adjusting the louvers so that even when the fan is not operating, they remain in a partly open position.

Some houses have no access to the attic area. If such is the case in your home, the probability is very great that the area is inadequately insulated by current energy standards. The amount of ventilation in the area is also questionable. Installing an access hatch to the attic is recommended.

Structural

While in the attic, be sure to inspect the roof framing. Check the rafters and trusses for cracked, broken, and sagging sections. (See FIG. 9-11.) Are the rafters spreading apart near the ridge? Although these problems are uncommon, they indicate a structural defect that should be evaluated by a professional. Look at the floor

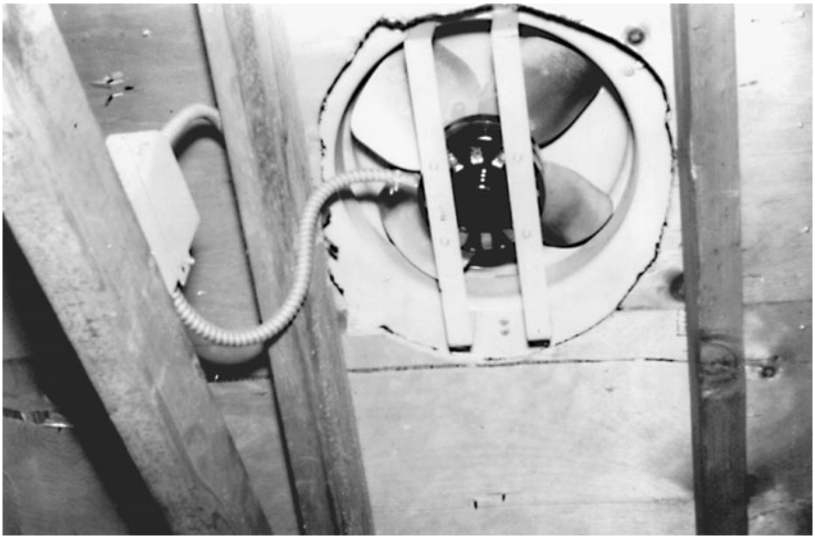

Fig. 9-10. *Thermostatically controlled attic fan mounted between the roof rafters.*

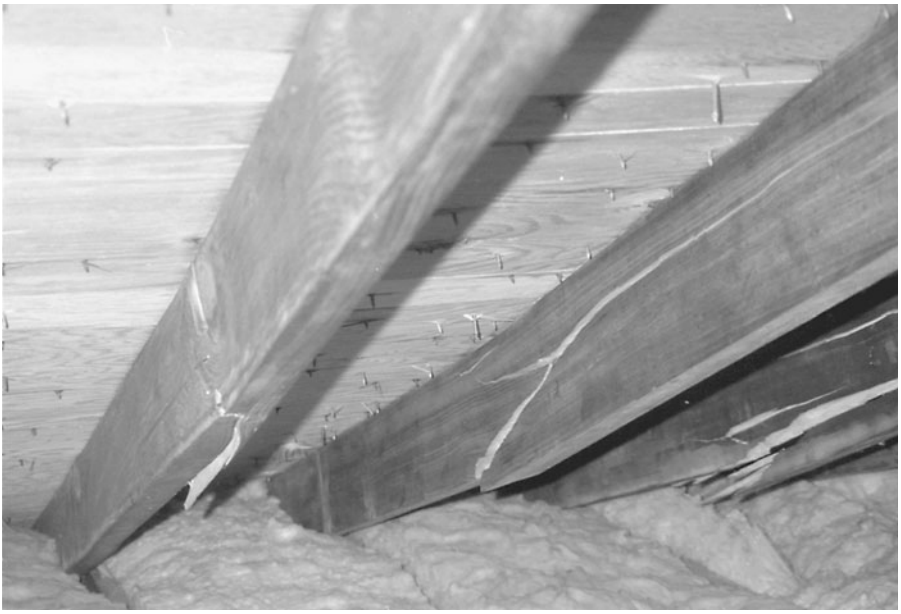

Fig. 9-11. *Cracked and broken roof rafters.*

joists. If they've been cut to accommodate an opening for an attic fan or pull-down stairs, have the cut ends been properly secured to a header? They should be. If not, record that fact on your worksheet for future correction.

When trusses have been used for roof framing, check to see if any of the webs or chords have been removed. Homeowners have been known to cut and remove them to accommodate storage items. This affects the structural integrity of the truss and should also be recorded on your worksheet for needed repairs.

Checkpoint summary

- ❍ Is attic insulated?
- ❍ Is additional insulation needed?
- ❍ Does existing insulation contain a vapor barrier?
- ❍ Is insulation properly installed?
- ❍ Is ventilation provided?
- ❍ Are vent openings blocked?
- ❍ Check operation of attic fan.
- ❍ If attic is poorly ventilated, look particularly at northerly slope for delaminating plywood or warped roofing boards.
- ❍ Look for signs of past or current roof leakage.
- ❍ Pay particular attention to area around the chimney and plumbing vent stacks.
- ❍ Are air-conditioning and heating ducts insulated?
- ❍ Look for open joints in the duct work.
- ❍ Do any plumbing vent stacks terminate in the attic? If so, are they capped with AAVs?
- ❍ Are kitchen or bathroom exhaust fans discharging into the attic?
- ❍ Are there open electrical junction boxes or makeshift electrical wiring?
- ❍ If house has a prefabricated chimney, is there a need for fire-stopping around the joint between the chimney and the attic floor?
- ❍ Are there any cracked, broken or sagging sections of rafters or truss members?

10

Interior rooms

Walls and ceilings 125
  Mold-resistant drywall 126
  Chinese drywall 128
Floors 128
Heat 131
Windows 131
Electrical outlets 132
Fireplace 134
  Gas-fired fireplace 136
Bedrooms 136
Bathrooms 137
  Water pressure and flow 138
Kitchen 141
Hallway and staircase 142
Checkpoint summary 143

After inspecting the attic, you should inspect every room and closet in the house. Don't pass a door without opening it and looking inside. If a door is open, close it to see if it operates properly. All *pocket doors* should be checked to see if they slide open and close easily (pocket doors are doors that when open are housed within the wall cavity). If there is a problem, it generally requires opening a section of wall to provide access to the tracks for repair.

Start your inspection of the rooms at the upper level and work your way down to the basement. As you walk from one floor to another, inspect the hallways and connecting staircase.

The problems normally encountered during an interior room inspection are usually of a cosmetic nature and are not costly to correct. Occasionally you might see cracks or uneven floors or walls that are symptomatic of structural problems, but that is rare. Usually cracks and uneven floors are caused by shrinkage, warpage, or slight movement of the house. Slight movement is considered quite normal. Few people realize that a house is constantly in motion. As the outdoor temperature and humidity vary with the season and the time of day, they cause differential expansion, swelling, and contraction of the various structural and nonstructural elements. This movement, although slight, is very often enough to cause cracks at points of stress concentration, such as over windows or doors.

When you walk into a room, try to look beyond the cosmetics. Don't dwell on the position of the furniture or the pictures on the walls. Look at the walls, the floor, ceiling, and trim, but do not be concerned about minor cosmetic problems. Hairline cracks, small holes, chipped sections, dirty and marked-up areas can all be corrected easily by spackle, wood filler, paint, or stain, depending on the finished surface.

Walls and ceilings

The materials most often used for covering walls and ceilings are plaster and plasterboard, also referred to as gypsum wallboard, drywall, or Sheetrock, which is a trademark of the U.S. Gypsum Company. Plaster—usually made from a mixture of gypsum (a common mineral), sand, and water—has qualities that are very desirable for a wall finish: structural rigidity, durability, resistance to sound transmissions, and high fire resistance. In addition, it can readily be applied to curved or irregularly shaped surfaces. A plaster wall is usually applied in two or three coats to a backing called lath. Depending on the type of lath, the thickness of the plaster usually varies between ½ and ¾ inch.

Plasterboard is a sheet material that basically consists of a gypsum mixture surfaced with a treated paper. Depending on the application, the thickness of the plasterboard can range from ¼ to ⅝ inch. The sheets are usually 4 feet wide and are available in lengths up to 12 feet. When constructing walls using this material, the plasterboard is fastened directly to the studs, and the joints are covered with tape. The edges of the plasterboard are recessed slightly so that when the tape is applied, it is level with the surface. This type of construction requires that the wood framing be perfectly straight and true. Otherwise, the wall surface will be uneven, and there will be visible joint lines. If the moisture content of the wood used in construction is not very nearly that which will be attained in service, there might be nail pops and joint cracks. Popping nails is a condition most often found with plasterboard construction. It can be corrected easily. It often does, however, indicate low-quality installation.

In houses where the roof framing consists of trusses rather than rafters, you might see cracks at the wall-ceiling intersection in the interior rooms below the attic. The cracks are usually the result of truss uplift, which causes the ceiling to move up and down in the course of a yearly weather cycle. The size of the cracks can vary from a hairline to an actual opening of an inch or more. This condition is a cosmetic problem rather than structural. The easiest cosmetic repair is to cover the crack with decorative molding that is nailed to the ceiling and not to the wall. As long as the molding is wide enough, it will cover the crack as the ceiling moves up.

When you inspect a room, look specifically for areas that need major cosmetic rehabilitation, such as broken walls or ceiling, loose and bulging sections of plaster, and disintegrating plaster. Sagging or bulging sections in a plaster ceiling are a safety concern because of the possibility of collapse. This problem, if it exists, is usually found in older homes. Plaster is applied over a wood, metal, or gypsum lath. The sagging condition results from broken plaster keys, which no longer lock the plaster layers to the supporting lath. (See FIG. 10-1.) The condition occurs over a period of time and is usually caused by vibration and wetting from roof and plumbing leaks, which weaken and crack the keys. If there are sagging sections of the ceiling, record this problem on your worksheet for future repair.

Very often, major cosmetic damage is the result of a water condition. As you look at these areas, also look for water stains. Stains

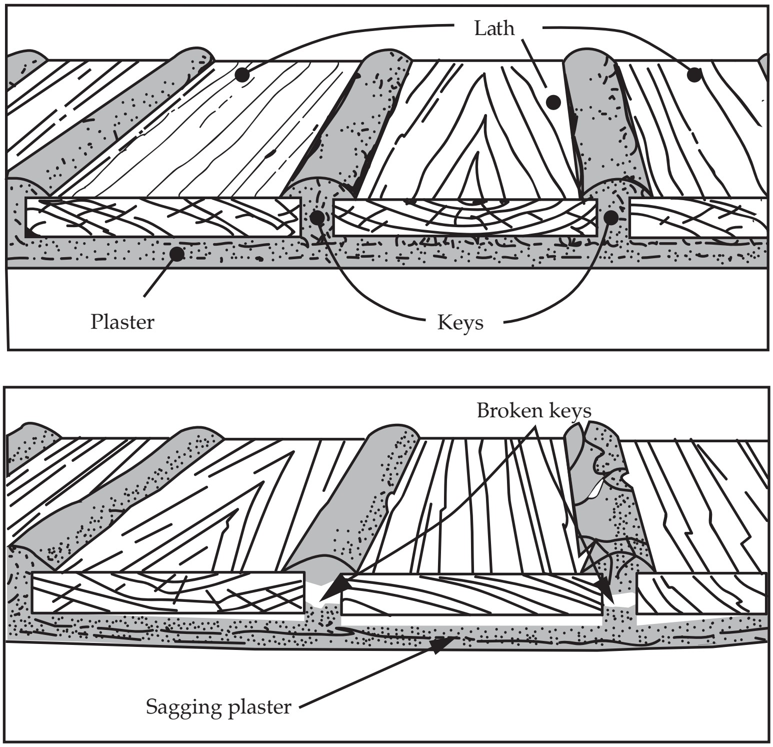

Fig. 10-1. *Ceiling plaster can sag down from lath when keys securing base coat give way.*

often show up as discolorations circled by a light brownish ring. If you see water stains, try to determine the source of the water. All stains do not indicate a current problem condition. You might be looking at the results of a problem that has been corrected. If you have a moisture meter, you should check all water stains to determine whether or not the leakage condition causing the stain is active. If it is, record it on your worksheet.

Mold-resistant drywall

Because of the concern in recent years about mold developing on wet or damp drywall, manufacturers have developed and are now supplying a mold-resistant gypsum wallboard. Mold growth requires moisture and a food source. When traditional drywall gets wet, the paper becomes the food source. With mold-resistant wallboard, either the paper has been chemically treated to resist or minimize the absorption of water or the paper has been replaced with a fiberglass mat facing. Mold-resistant wallboard will reduce the chances of mold developing. However, it does not mean that mold cannot grow. The installation of mold-resistant wallboard is exactly the same as for traditional wallboard.

Water stains on a ceiling can be caused by several types of problems, such as water leaking between the tile joints around a shower or tub located in a bathroom above the ceiling, an ice dam on the roof, an open joint between a plumbing vent stack or chimney and the roof, leaky plumbing, or condensation on cold-water pipes located above the ceiling. Sometimes the cause is obvious, such as a faulty roof. However, you should be aware that leaks that originate in the roof can stain ceilings two or more levels below. The stains are not limited to ceilings just beneath the roof.

Sometimes the source of a stain cannot be explained by any of the above. On one inspection, I noticed that the ceiling of one room had several stains that were still damp. I knew that there were no plumbing pipes above the ceiling and that the roof was in good shape. Frankly, I couldn't figure out why the ceiling was damp. All of a sudden, I noticed that water was dripping from one section of the ceiling. I then went upstairs to the room above and found what was causing the leak. Apparently, the owner kept a puppy that was not housebroken in the room. Whenever the puppy missed the newspapers that were placed there for training purposes, the urine seeped through the floor joints, wetting the ceiling below.

Ceilings can also be covered with composition tiles. These tiles are generally made from asbestos, glass fiber, or fiberboard. They can be applied directly over plaster or plasterboard ceilings or over furring strips and are usually interconnected with tongue-and-groove joints. If you are inspecting an older home with plaster walls and ceilings and find yourself in a room that has a tile ceiling, the tiles were probably installed to cover broken and cracked sections in the original ceiling. It is becoming increasingly difficult to find a good plasterer. Consequently, when renovating an older home, many people cover a ceiling with tiles or plasterboard rather than have it replastered. Look specifically for loose and sagging sections. These sections must be resecured; otherwise, they can cause adjacent tiles to loosen, which can eventually result in a "domino" effect, causing most of the tiles to come down. (See FIG. 10-2.) I inspected a house the day after two-thirds of the ceiling tiles in one room came thundering down. The owner told me that there always had been a few loose tiles.

Water stains on walls facing the exterior might be an indication of water seepage through open joints in the exterior siding. In particular, this condition has been found in older homes with an exterior brick or stone siding and is usually caused by wind-driven rain penetrating cracked and deteriorating mortar joints. When water stains and peeling and flaking paint are found on walls just below windows, it is usually an indication of

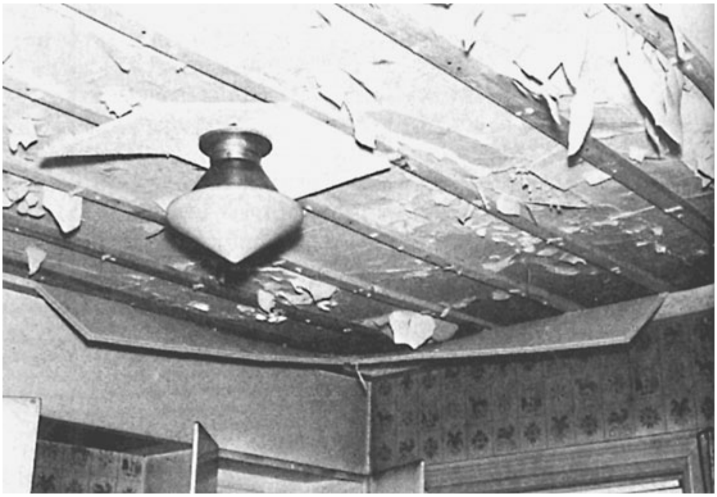

Fig. 10-2. *Ceiling that had been covered with tiles, which came loose and fell down. Note that the tiles had been installed to cover a cracked, broken, and peeling ceiling.*

leakage through open exterior joints around the window.

Cracks in walls are normally merely of cosmetic concern. However, when walls are cracked and the windows and door frames in that room are not level, the condition might be indicative of a structural problem and should be evaluated by a professional.

When the walls are covered with wood or hardboard panels, cracks will not be visible. In quality homes, panels are normally installed over a plasterboard wall. This type of backing provides added rigidity and, when applied on exterior walls, added insulation. If the room you are inspecting has a panel wall, push on the panel at a point between the studs (the studs are usually located 16 inches apart, in some houses 24 inches apart). If the wall panel yields under your push, there is no plasterboard backing. I inspected a new house just after construction and prior to occupancy. The buyer specifically asked me to check the plasterboard because he was paying extra for ⅝-inch rather than ½-inch plasterboard. The builder did install the thicker plasterboard, but in those rooms that were paneled, the builder completely omitted the plasterboard behind the panels because it was not visible to the buyer.

The trim around the various joints in the room should be inspected and checked for missing, loose, cracked, and broken sections. Although a minor element, trim in this condition is an indication of shoddy workmanship in new construction and neglect in a resale.

Chinese drywall

Between 2004 and 2006, because of the housing boom and the need to rebuild the homes destroyed by Hurricane Katrina, many millions of pounds of drywall was imported from China. It is estimated that enough material was imported to build about 100,000 homes. Shortly thereafter, problems began to develop in some of the houses in the southern states where the interior walls were built using drywall made in China. Chinese drywall emits higher levels of sulfuric and organic compounds than drywall made in the United States. The sulfur fumes emitted caused air-conditioner coils, copper plumbing, and electrical wiring to corrode, as well as producing a foul odor such as that of a burning match or rotten eggs. The fumes have also been associated with respiratory and sinus problems in some residents. So far, most of the drywall problems have been limited to the southern states, where a warm, humid climate encourages the emission of sulfur fumes. If the drywall was also used in other parts of the country where it is dryer and cooler, it could be years before homeowners begin seeing the problems associated with the material. Correcting the problem would require stripping and removing all the drywall. There are also indications that it may be necessary to replace some of the material in the house such as carpeting and furniture.

Floors

When walking around the house, you might notice that the floor squeaks on one or more sections. This condition is found in both new and older homes and indicates slightly loose floorboards. It is usually difficult to eliminate. One corrective procedure is to wedge the floor from below. However, the underside is usually not accessible. If a hardwood floor is nailed from above, the nail holes can ruin the finish. With regard to the structural integrity of the house, squeaking floors are not a concern, although if excessive, they can be annoying.

If the floors in a portion of the house are finished hardwood and the remaining floors (other than those in the bathrooms and kitchen) are covered with wall-to-wall carpet, do not assume that there is a hardwood floor below the carpeting. There might be, but quite often there isn't. It is not uncommon for a builder to give a buyer the option of hardwood floors or carpeting. When the buyer selects carpeting, it is laid over a plywood floor. If you are inspecting a room and see a hardwood floor in a closet, though the floor in the main portion of the room is carpeted, do not assume that there is a hardwood floor beneath the carpeting. If you cannot see it, you really cannot be sure. If a hardwood floor is important to you, get a representation in writing from the owner that the floors beneath the carpets are hardwood.

Just because the floor looks like a hardwood floor, it may not be, in the classic sense; that is, solid wood from top to bottom. It may be an engineered hardwood floor or a laminate floor. An engineered hardwood floor is made up of three to 10 thin layers of wood with a core of hardwood, plywood, or high-density fiber. The top layer is a hardwood veneer that is glued on the top surface of the core and is available in almost any hardwood species. It is less expensive than hardwood, and due to its multi-ply structure, engineered wood is much more stable than solid wood and is less susceptible to shrinking and expanding with changes in temperatures and humidity. It can be installed on any grade level, directly over concrete and can also be installed below ground level, whereas solid hardwood should only be installed above grade. Unlike engineered wood, moisture and extreme temperature changes can cause solid wood to shrink and expand, potentially causing gaps between boards during colder or drier seasons. If there is excessive moisture on the underside of the solid hardwood floor, it will cause the top surface to cup, that is, to form a concave surface with the edges raised.

Laminate flooring is the least expensive between solid hardwood and engineered flooring. Although it looks like wood flooring, there is actually no solid wood used in its construction. Laminate floors are made up of several materials bonded together under high pressure. Most laminate flooring consists of a moisture-resistant layer under a layer of HDF (high-density fiberboard). This is topped with a high resolution photographic image of natural hardwood flooring, which accurately reproduces the grain and color of natural hardwood. To protect the laminate, it is then finished with an extremely hard, durable clear coating made from special resin. Laminate flooring can be installed above or below grade and over virtually any other flooring surface. It is usually laid as a floating floor. It is not nailed or glued to the subfloor.

If the wall-to-wall carpet covering the hardwood floors is brand new, ask the seller for a representation in writing that the surfaces of the hardwood floors are in good condition. Several years ago I inspected a split-level house with hardwood floors. All the floors were covered with new wall-to-wall carpet. Lifting a corner of the carpet in each room did not reveal a problem. However, after the buyer moved in, he removed the carpet from all the hardwood floors and was dismayed by what he saw. Every one of the floors was very badly stained by cat or dog urine and had to be refinished.

Another type of hardwood floor is a bamboo floor. The flooring is typically made by slicing bamboo poles into strips. To prevent potential mold, fungus, or insect problems, these strips are boiled and then sent to a drying kiln to reduce moisture levels. The strips are then glued and assembled under extremely high pressure and heat either face up for horizontal flooring or side by side for vertical flooring. Although vertical flooring results in a smooth uniform appearance, horizontal flooring has a more natural look. The pressure and heat applied assure a strong, stable bond. After milling, the bamboo is given an aluminum oxide coating to increase its strength and endurance, and to protect it from stains. Direct sunlight can cause some types of bamboo floor to become discolored. It may happen if the floor is exposed to highly concentrated sun rays for long periods. Bamboo flooring needs to be kept free of dust, dirt, and grit. Even the tiniest particles of dirt can cause minute scratches on its surface. It is like any wood floor—it is damaged by dents, scratches, and shoes with high heels.

If you walk into a room and notice that the floor is not level, do not be alarmed. In all probability, it is a condition caused by past shrinkage, warpage, and settlement of the wood framing and is not a concern. If the floor is sagging in one section, it might indicate that that portion of the floor is or was improperly supported. Go down to the level below to check the ceiling to see if it too is sagging. If it is, have a professional evaluate the condition. Occasionally this condition occurs in a kitchen floor when the refrigerator is placed in a location other than that intended by the builder.

Depending on the moisture content in the wood used for framing, you might see a large open joint between the floor and the partition walls. The joint might be open as much as 1 inch and might run the entire length of the partition wall. When you see this condition for the first time, it is quite unnerving. The trim that normally covers the joint between the floor and the wall is about an inch above the floor, and the joint is wide open. The condition is caused by excessive shrinkage of the wood framing and is only a cosmetic problem. By covering the open joint with trim or lowering the existing trim, the room will look almost as good as new.

The floor on the lower level of bilevels is a concrete slab. Depending on the quality of construction, the concrete floor slab could settle to a point where there is also a large open joint between the floor and the walls. (See FIG. 10-3.) During construction, if the ground below the floor slab is not properly compacted, the slab might eventually settle, resulting in open joints. The foundation walls all have support footings and are normally independent of the floor slab. Consequently, settlement of the floor slab usually does not indicate a problem with the main structure. Nevertheless, the wall above the open joint should be shimmed; otherwise, the floor above might sag. If the settlement of the concrete floor slab is accompanied by cracked and settled sections of the foundation wall, a severe problem does exist and should be evaluated by a professional.

Sometimes the concrete floor slab has a raised wood floor that might be covered with carpeting or resilient floor tiling. When walking on this floor, test for soft or spongy

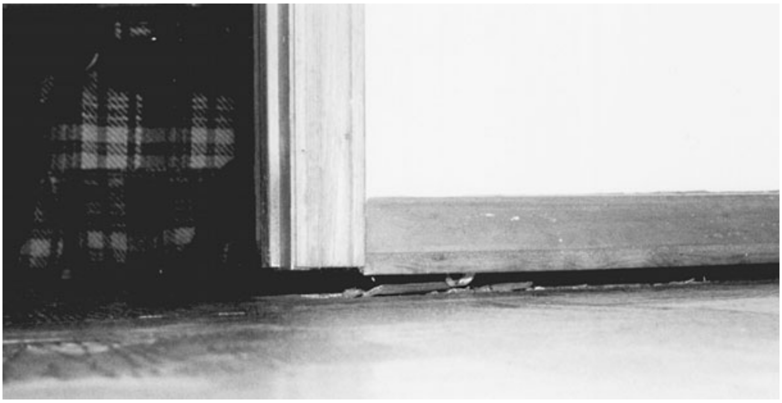

Fig. 10-3. *Settled concrete floor slab. Note open joint between floor and partition walls.*

sections. This condition can be caused by inadequate spacing of the wood framing. However, more often than not, it is caused by rotting wood. If water seeps into the area between the wood floor and the concrete slab, it eventually causes the framing to rot and the subflooring to delaminate. (See FIG. 10-4.) See chapter 11 for a discussion of water seepage into the basement level.

Heat

With the exception of homes in the sunbelt, all the finished rooms in the house should have provisions for heating. Depending on the type of heating (see chapter 14), look for a radiator or heat register. If you don't see any, ask the owner how the room is heated. It might be heated by radiant panels in the floor, walls, or ceiling that would not be visible. If there is no source of heat, record the fact on your worksheet. Adding heat to a room by extending the existing heating system can be costly. If there is a means of heat supply, is it properly located? For maximum efficiency, radiators and heat-supply registers should be located on an exterior wall, preferably below a window. This allows the heat to mix with the cooler outside air that normally infiltrates the interior from around the windows.

Windows

Although the condition of most of the windows is checked during the exterior inspection (see chapter 5), the condition of the windows on the upper levels and the operation of the windows should be checked during an interior inspection. Also, if the windows are the casement or awning type, the storms and screens are often not visible from the exterior and should be looked for during the interior room inspection.

The cost of repairing a broken window frame or replacing a broken pane is usually not high on an individual basis unless the window is one of unusual design or a large thermal pane (double or triple glazing). Keep track of the number of windows that need repair or rehabilitation. It is surprising how fast the dollars add up when you multiply the

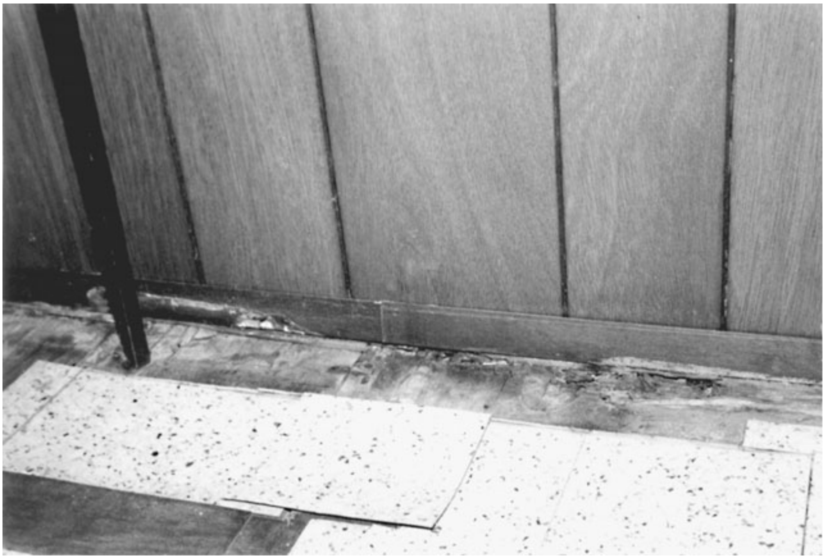

Fig. 10-4. *Rotting and delaminating section of raised wood floor in finished basement. Condition is caused by constant wetting from water seepage into the basement.*

cost for one repair by the number of windows that are faulty.

Specifically, check the windows for cracked and broken panes that require replacement. BB holes or small cracks in the corner of a pane can be overlooked, providing that the window is covered by a storm pane. Is the putty around the pane cracked, dry, chipped, or missing? If so, the window joints must be reputtied. Look at wood window frames and exterior sills for cracked and rotting sections. Some windows have metal frames. Metal frames get quite cold during the winter in the northern part of the United States and have a tendency toward excessive condensation. This can cause peeling and flaking paint or disintegrating sections of plaster near the window frame.

Pay particular attention to steel casement windows. These windows are usually a problem. They rust easily and must be painted every few years to prevent deterioration. In many cases, they do not close properly, a condition often caused by a sprung frame or excessive layers of paint around the joints. Also, almost invariably you will find cracked panes. Usually casement windows are crank-operated. Check the hardware. Sometimes the cranking mechanism is not operational.

Open and close the windows. They should operate relatively easily without sticking or binding. Double-hung windows should not rattle in the channel and should stay in a fixed position when opened fully or partially. Many older double-hung windows use a counterweight to hold the sash in a fixed open position. The weight is usually tied to a sash by a cord or a chain. If a cord is used, check to see if it is broken. Broken cords are not uncommon and should be replaced. If the cord is frayed, you can anticipate its early replacement.

When looking at the windows, check to see if the glazing is a single pane or a thermal pane, that is, two pieces of glass separated by a sealed air space. You can tell by looking at the thickness of the joint between the pane and the frame. A single pane is usually no more than $\frac{3}{16}$ inch thick; a thermal pane is about $\frac{3}{8}$ inch to 1 inch thick. You can also tell by looking very closely at the pane. Usually, you can see dust or dirty spots on the opposite side that reveal the thickness of the pane. Sometimes fixed-pane windows and sliding glass doors have thermal panes, even though the openable windows throughout the house have only a single pane. One of the well-known names for thermal glass is Thermopane. It is also referred to as insulated glazing. Each manufacturer has its own term for it. You can usually find the manufacturer's name in the corner of the window pane.

Thermal-pane windows are fabricated by hermetically sealing dry air or an inert gas between the panes. This is done to eliminate the possibility of future condensation problems between the panes. (See FIG. 10-5.) When the seal breaks, accidentally or otherwise, water vapor can enter the space between the panes. Look at the windows for signs of a faulty seal. If a portion of the window appears cloudy (FIG. 10-6) or there are water droplets between the panes, the seal has been broken. Although the insulating characteristics of a thermal-pane window with a faulty seal are at least as good as a storm window, the pane is not desirable from a visibility and a cosmetic point of view. If you see a thermal-pane window with a faulty seal, record it on your worksheet. The window requires replacement. The better window manufacturers offer a 20-year warranty on the seal of these windows.

Electrical outlets

While inspecting the interior rooms, look for electrical hazards and violations and whether there are an adequate number of electrical outlets on the walls. These items are discussed in chapter 12. According to the electrical code,

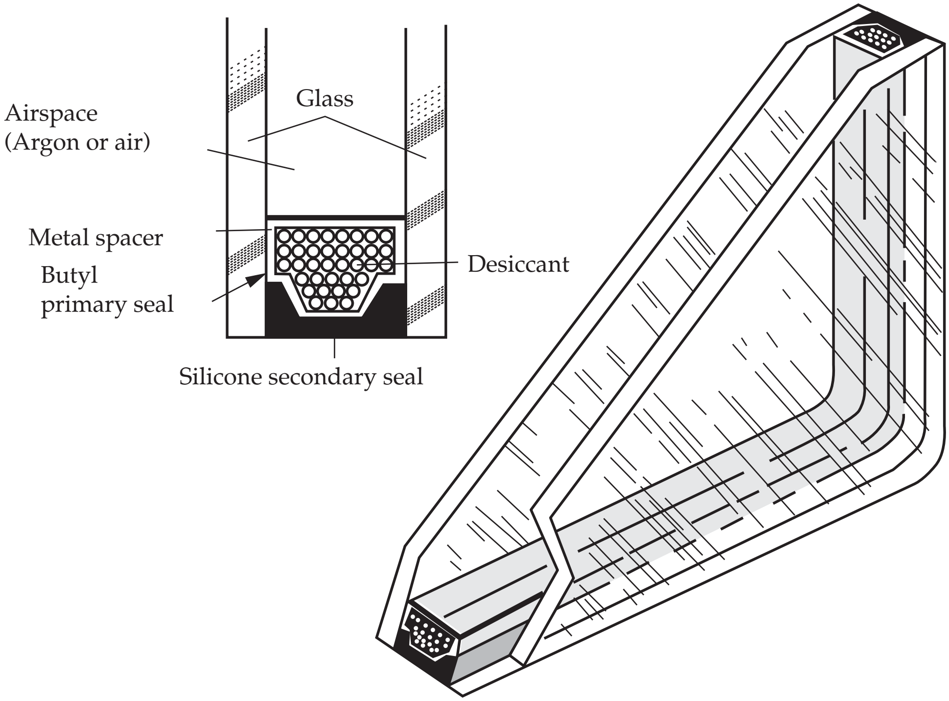

Fig. 10-5. *With an insulated glass pane, if there is a defective seal, moisture will condense between the panes.*

the outlets in a new house must be located such that no point on the wall is more than 6 feet horizontally from an outlet. However, the actual number of outlets needed for a specific room depends on the room's usage and the position of the furniture.

Look at the electrical outlets. In new homes and in recently rewired older homes, the receptacles have three slots rather than two. The third slot is a ground connection that is used in conjunction with the three-prong plugs found on most modern appliances. It is a safety feature and is used for grounding the casing of electrical appliances and equipment. Should an internal short develop between the wiring and the equipment or appliance casing, the ground connection directs the leakage current harmlessly to the ground rather than through the user.

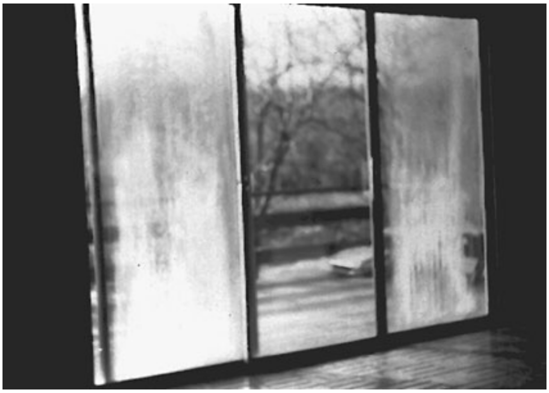

Fig. 10-6. *Faulty seal in thermal-pane windows adjacent to a sliding glass door. Restricted visibility is caused by condensation between the glass panes.*

In homes built after 2008, check to see if the outlet receptacles are tamper resistant. If they are, the letters “TR” will be marked on the face of the receptacle. In 2008, the National Electrical Code (NEC) mandated the use of tamper-resistant receptacles in new or renovated construction. Tamper-resistant receptacles have a shutter mechanism that in most instances will block access to the electrical contacts from anything but an electrical plug. Although a worthwhile safety feature, not all municipalities have adopted the new code. Check with your local Building Department.

Not all electrical appliances and equipment have a grounding wire (three-prong plug). Some are made with double-insulated plastic cases that prevent the user from touching anything electrically charged, even if an internal short should develop. This type of equipment does not require a ground connection and can therefore be used with the older two-slot receptacles. As a safety precaution, you should use only electrical appliances and equipment that have been approved by Underwriters Laboratories, Inc., or some other nationally recognized testing agency.

The two-slot outlet receptacles can be used for the grounding of appliances with three-prong plugs. Adapters are available at hardware and electrical supply stores to enable two-slot receptacles to accept the three-prong plugs. To complete the ground connection when using the adapter, remember that the small wire (pigtail) must be secured to the center screw on the receptacle cover plate.

Check the receptacles (two- and three-slot) to determine if they are electrically hot, whether the receptacles have reverse polarity, and whether they are properly grounded. This can be done using a simple plug-in tester available at all electrical supply stores.

Reverse polarity is a problem because it indicates that the hot and neutral wires are connected to the wrong terminals in the receptacle. This is important because the switch on an appliance should disconnect the hot wire. This keeps everything beyond the switch inside the appliance from being electrically hot when the switch is off. With reverse polarity the switch disconnects the neutral wire so that everything beyond the switch will be hot, which could result in an electrical shock. The diagrams in FIG. 10-7 show how a polarized circuit can protect you from a shock hazard.

It is particularly important to check the receptacles in the bathroom and kitchen to see if they have ground-fault circuit protection. Receptacles that are not properly protected for ground faults are potential hazards. See chapter 12 for discussion of ground-fault circuit interrupters (GFCIs).

Fireplace

If a house has a fireplace, it is most often located in the living room or family room, although

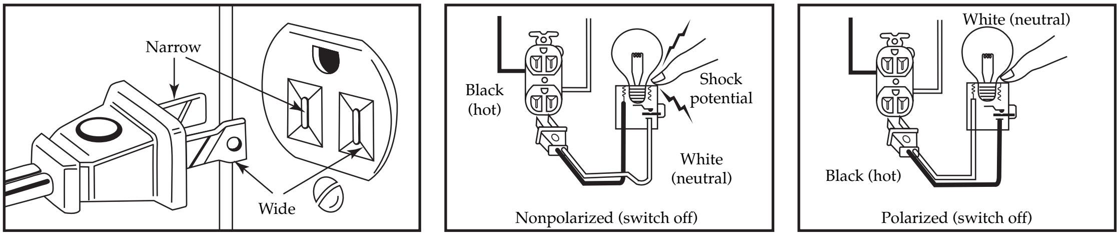

Fig. 10-7. *How a polarized circuit can protect you from a shock hazard.*

you might find a fireplace in a bedroom or kitchen. When inspecting the fireplace, look at the front face just above the firebox. If this area has a blackish tint or color, it is usually an indication of a smoky fireplace and is the result of a buildup over the years of layers of soot and creosote. This problem can usually be corrected.

A smoky fireplace might be the result of too small a flue for the size of the firebox opening. If this is the case, reducing the size of the opening by raising the hearth or installing a canopy on the top portion of the opening very often corrects the problem. However, determining the amount to raise the hearth or the size and shape of canopy to use requires experimentation. Sometimes the soot buildup is caused by backsmoking as a result of downdrafts. This is the result of wind currents bouncing off the side of the building or tall trees and then down the chimney. If the smoking is caused by downdrafts, the problem can usually be corrected by installing a concrete, stone, or metal cap on top of the chimney. Another possible reason for poor draft or backsmoking is that the chimney is undersized. A rule of thumb for the minimum height of the chimney states that the total height from the firebox floor to the top of the chimney should never be less than 15 feet.

Backsmoking can also be caused by a negative pressure condition within the house. This is a fairly common phenomenon in newer houses. To conserve energy, newer homes are better caulked, weatherstripped, and tighter than older homes. The negative pressure results when more air in the house exhausts to the outside from fans, heating-system chimney, and so on than flows in from infiltration. Because of the negative pressure, when a fire is first lit in the fireplace and the damper is opened, there is an onrush of incoming air down the flue. This condition can be eliminated by slightly opening a window or in newer fireplaces opening the fresh-air vent associated with the fireplace.

Look inside the firebox for a damper. The damper is used to close the flue when the fireplace is not in use. It prevents heat loss through the flue in the winter and also prevents small animals such as squirrels and racoons from entering the room through the flue. Check the operation of the damper. On many older fireplaces, dampers were omitted. If the fireplace does not have a damper, or if the damper is defective, record that fact on your worksheet. A damper or equivalent is considered necessary, and one should be installed. If there is a chimney-top damper, check its operation. See chapter 3, page 25.

The bricks or stones lining the firebox should be checked for cracked, chipped, broken, and disintegrating sections. Are the mortar joints intact, or are they in need of repointing? Cracked or open sections inside the firebox are a potential fire hazard and must be repaired.

Look up at the flue from inside the firebox. Normally this area will be coated with a layer of soot and creosote. If the layer is thick, have the chimney cleaned to minimize the possibility of a chimney fire. If the creosote layer is powdery, the flue can be cleaned by a chimney sweep with a brush. However, if the creosote buildup is of a tar consistency or a hard glaze, it cannot be brushed out by conventional means. It must be removed by chemical or mechanical means. Is there an obstruction in the flue, such as a bird's nest? If the flue has a slight offset and you can see daylight, there is no obstruction. When the flue is offset so that you cannot see straight up, determining whether there is an obstruction is difficult without lighting a fire. One trick is to blow up at the flue to dislodge fine particles of soot. If the flue is not obstructed, they will float up the chimney.

In some homes you might find a prefabricated wood-burning fireplace. These units are usually available with chimneys and have a specially insulated firebox shell. They are light in weight, do not require a special foundation, and can be wall-mounted or freestanding. The fireplace can be located in practically any part of a house, depending on local codes. If you find such a fireplace, check to see if it has been approved by Underwriters Laboratories, or some other nationally recognized testing agency.

Gas-fired fireplace

There are two basic types of gas fireplaces. They are either vent-free or have a direct vent. With direct vent fireplaces, the combustion gases are vented to the outside by a two-layer pipe that runs through a hole in the wall behind the unit. The outer pipe draws air in from the outside, which is needed for proper combustion of the gas. The inner pipe channels the gaseous products of combustion to the outside, or up a chimney. If a chimney is used, the local building codes may require the damper in your chimney to be permanently set to an open position because of the gas pilot light.

With a vent-free fireplace there is no need for a damper. These appliances have a thermal sensor located near the top of the fireplace opening. This is a safety device that in a properly operating fireplace would never be triggered. The sensor cannot be reset. It comes with the unit and is set in the factory. Vent-free fireplaces also have an oxygen depletion sensor. If the fire takes too much oxygen out of the room's air and is not replaced, the sensor is designed to turn off the gas before carbon monoxide reaches a dangerous level in the room.

Bedrooms

Every habitable room in the house must have at least one openable window. Sometimes do-it-yourself homeowners finish off an area in the basement or attic for use as an extra bedroom even though there is no window in the room. If you find such a room, be advised that it is a potential fire hazard.

When a room is used as a bedroom, certain items are necessary as fire-safety measures. As a means of escaping in the event of a fire, each bedroom should have at least one outside window whose sill height is not more than 44 inches above the floor. The window should also have a minimum net clear openable area of 5.7 square feet with no dimension less than 20 inches in width and 24 inches in height. If a bedroom is located two stories above the ground, a rope ladder or equivalent device should be provided in the room to enable the inhabitants to escape once they climb through the window. Some municipalities have local ordinances against converting a third-level attic into bedrooms without adequate fire protection. If you find bedrooms in a converted attic, you should check with the local building department to see if a certificate of occupancy had been issued for those rooms.

Last, from a fire-safety point of view, every bedroom must have an entry door. This door should be closed when the occupant goes to sleep. A closed door reduces drafts, thus reducing flame-spread time, and drastically reduces smoke infiltration. Check to see if there is a smoke detector on the wall or ceiling. There should be. In new construction most states require that all of the smoke detectors in the house be hard wired and interconnected. Check the operation of the detector. If it's interconnected with the others, hold your ears. The alarm noise is quite loud.

Don't forget to look for a closet. Is there a ceiling light fixture in the closet? More often than not the light fixture is a porcelain base with an exposed incandescent bulb. Check to see if the bulb is relatively close to the closet shelf. If it is, it's a potential fire hazard. If clothing or other fabrics come in contact with a bulb that has been inadvertently left on for an extended period of time, they could ignite and start a fire. In some homes where bedrooms have been added, closets are omitted in the renovation. There should be a closet in every bedroom. If there isn't one, record that fact on your worksheet.

Bathrooms

Because of the nature of the room, the bathroom must be well ventilated. Ventilation can be provided naturally through an openable window or mechanically by an exhaust fan. If there is an exhaust fan, try to determine where it discharges. The air drawn through the fan should discharge into the atmosphere. This can be accomplished through the use of a duct that terminates on the side of the building or one that extends up through the attic and terminates on the roof. When the bathroom is located on the level just beneath the attic, the exhaust fan very often discharges directly into the attic. This is undesirable because the moisture-laden air can cause condensation problems in the attic. From a convenience point of view, many builders often connect the exhaust fan to the lighting circuit so that both are controlled by one switch. This is not an energy-efficient installation, since the exhaust fan is not always needed. If you see such a setup, you might want to consider rewiring the fan so that it can be controlled by a separate switch.

A recurring problem in bathrooms is water leakage through cracked and open tile joints around the tub or shower. This condition requires periodic maintenance and if neglected can cause considerable cosmetic damage. Check the tiles for cracked and open joints. Lean over the tub or go into the shower and press on the tiles, particularly at the lower portion of the walls. Loose tiles will move slightly or might even come out. If the tiles yield, it is usually an indication that the plaster or plasterboard backing has suffered some deterioration from water seepage. Depending on the severity of the deterioration, it might be necessary to rehabilitate that portion of the wall. Sometimes the cracks are not readily visible but will nevertheless allow water to seep through and wet the wall behind the tiles. You can check for leakage by using your moisture meter on the tile wall.

If the walls around the tub and shower are not kept watertight, water can leak around open areas, wet the ceiling below, and eventually rot the wood framing and cause the ceiling to deteriorate. Missing tiles must be replaced; loose tiles must be resecured; and cracked and open joints must be regrouted or caulked. Caulking is the procedure most often used for making repairs. A tube of caulking compound is available at any hardware store.

If the walls around the tub or shower have a panel finish rather than tile, check the joints for cracked or open sections. Sometimes the tub or shower and its associated walls are an integral unit made of molded plastic. In this case, leaks through open joints are not a problem. Many homes have showers and tubs with doors to prevent water from splashing onto the floor. The doors must be made with safety glass or plastic and should be checked for cracked panes and ease of operation.

When the shower base is covered with ceramic tiles rather than molded plastic or is terrazzo-constructed, there is the potential for a problem because of a faulty shower pan. A large lead or plastic sheet is normally installed below the tiles at the base of the shower to collect water that seeps through cracked and open tile joints. If the shower pan is intact, it will direct the water down the drain without incident. However, as the shower pan ages, the joints very often deteriorate, resulting in water leakage through those joints. If you notice large water stains on or damage to the ceiling of an area below the shower, you should suspect a shower-pan problem. (See FIG. 10-8.)

If you have the seller's permission, you can test for a faulty shower pan by covering the shower drain and filling the shower base with about an inch of water. Let the water stand in the base for about forty-five minutes. If the shower pan is faulty and there are cracks in the base tile joints, water will seep through and wet the ceiling below. When the tile joints at the base of the shower are all sealed, even if the lead pan is faulty, there will be no leakage. Instead of replacing a faulty shower pan, many homeowners simply recaulk the tile joints at the base of the shower. This is considered a makeshift fix. If you find a heavy layer of caulking in the shower base, even if you do not find water stains on the ceiling below (the ceiling could have been repainted after the makeshift fix), you should suspect shower-pan problems.

Water pressure and flow

The water pressure in a house is an item of concern to most home buyers, and rightly so. However, water pressure per se is not usually the problem. If you see water trickling out of a faucet, it does not necessarily mean that the pressure is low. It means that the water flow is low. Low water flow can be caused by low pressure. However, more often than not, low flow is caused by a constriction in the supply pipes. The constriction can be the result of mineral or corrosive deposits on the inside diameter of the pipes, a kink in the pipe, or small-diameter piping. Quite often low flow at a fixture is caused by a partially clogged faucet aerator, which requires only cleaning. Water flow is measured in gallons per minute, and a sink faucet should be able to deliver 4 to 5 gallons per minute.

The water pressure available to a house at the meter varies, but it is usually in a range of 20 to 60 pounds per square inch (psi). I inspected

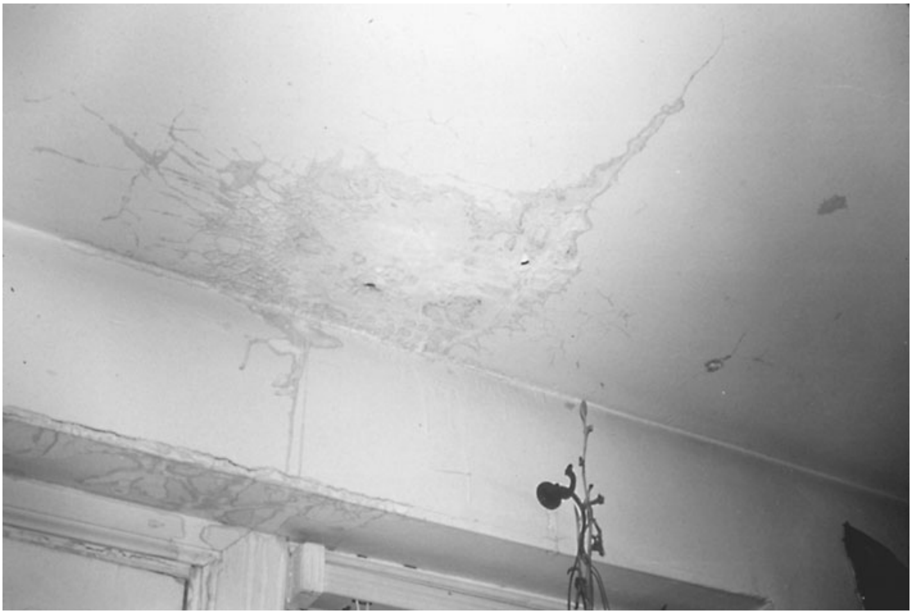

Fig. 10-8. *Water stains on ceilings caused by leaks from radiator in room above.*

a house that had a pressure gauge on the house side of the water meter. (A pressure gauge in the plumbing system is desirable but is not very common.) The gauge indicated that the available water pressure was 110 psi. Yet when three faucets were turned on at the same time, the water flow from each faucet was just a trickle. Regardless of how high the pressure is, if a constriction in the water pipes limits the water flow, the amount of water discharging from the faucet will be noticeably low. In this case, the pipes were made of galvanized iron. Over the years, corrosion and deposits had coated the inside of the distribution piping to a point where the pipes required replacement.

When you inspect the bathroom, check the water flow by simultaneously turning on the faucets in the sink and tub and flushing the bowl. There will usually be a drop in flow when the second faucet is turned on. If it is not very noticeable, do not worry about it—it is normal. You can trust your eye. If the water flow from the faucets looks good, then for all practical purposes, it is good. If the flow is noticeably low, record it on your worksheet. The condition requires correction. If there is a shower in the tub, check the operation of the shower diverter valve. When the shower is on, there should be very little or no water leaking out of the tub spout. If there is, the diverter valve is defective and will require replacement. Record this on your worksheet.

Check the flow for both the hot water and the cold water, but not at the same time. It is possible for the cold-water flow to be good and the hot-water flow fair or poor. Occasionally, when the domestic hot water is generated through the heating system, deposits form in the tankless coil, restricting the flow. If the hot-water flow is poor, you probably need to replace the tankless coil. (See chapter 16.)

Open and close the faucets rapidly. Do you hear a hammering and vibrating noise? You should not, but if you do, you are hearing *water hammer*, which occurs when the water flowing in the pipe comes to an abrupt stop. It introduces hydraulic shock and vibrations that can in extreme cases damage the pipe or fittings. The condition can be corrected easily by installing an air chamber or antiknock coil. These units provide a cushion of air to absorb the shock when the water flow stops.

While you are checking the water flow, look for water leakage around the faucets, and the drainage trap below the sink. If you notice that a portion of the trap is taped over, it is usually a makeshift correction for a leak. The drainage trap is not installed below the sink to catch rings or other valuables that fall down the drain, although very often this happens. The purpose of the trap is to provide a water seal that blocks noxious gases from seeping back into the room. If a sink trap is not properly vented, it can lose its seal from a siphoning action when the water is draining. (See FIG. 10-9.)

Usually the P-type sink trap is vented properly. However, the S-type trap is often not vented properly. You can tell by making the following test, although it is not foolproof. Fill the sink until the water is almost at the rim; then let the water drain. When most of the water has drained, if you hear a sucking and gurgling noise, the trap is not vented, and the water seal has been lost. When all the water has drained, if you let the water run freely for about five seconds, you will reestablish the water seal. If the main plumbing waste line has a house trap, the problem is minimized.

Check the drainage in the sink, tub, and bowl. Wrinkle up some toilet paper into a ball and throw it into the bowl. When the bowl is flushed, the paper should be carried down the drain. If the water in the bowl starts to rise to the top and then settles down without flushing, there is a blockage in the drain line that must be checked further and corrected. If the house has a septic system, this condition might indicate a problem with the septic

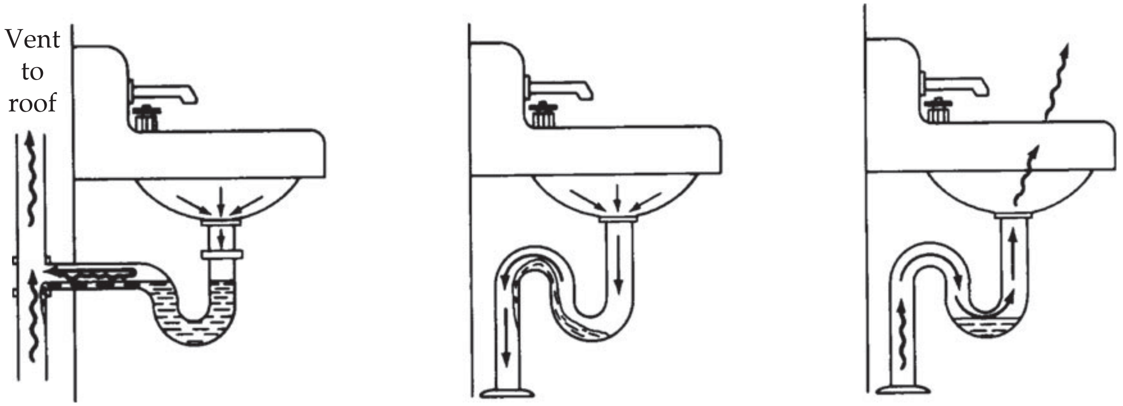

"P" Trap

When the sink trap is properly vented, the trap holds enough water to form a seal against the entry of sewer gases. The gases vent harmlessly through the vent stack to the outside.

A

"S" Trap

When the sink trap is not properly vented, the rush of the waste water can siphon the water seal out of the trap.

With no water seal, sewer gases can escape into the house.

B

Fig. 10-9. *Sink traps: A—Properly vented P-type trap; B—S-type trap, usually not properly vented.*

system. If the water in the sink and tub does not readily drain away or drains sluggishly, there is a blockage in those drains that must also be corrected. In this case, the correction might be as simple as removing hair that has accumulated around the pop-up drain plug.

Look at the sink and tub faucets. When the tub and sink are filled to the flood rim, there should be an air gap between the top of the water and the bottom of the faucet spout. (See FIG. 10-10.) If there isn't, the faucet is vulnerable to back siphonage (the flowing back of dirty water from the sink or tub into the potable water supply because of a negative pressure in the line). This condition is a violation of the plumbing code and is found periodically in old homes. However, it is not limited to old homes. The "do-it-yourself" homeowner could have installed the faucet without being aware that it is a potential problem. Are there shutoff valves below the sink and bowl for the supply pipes? Although not a necessary feature, they are very convenient when making minor repairs.

If the tub is equipped with a whirlpool bath, check its operation. Fill the tub to a level above the jet ports before turning on the bath's pump. Record any problems on your worksheet. For health reasons, before the whirlpool bath is used, it should be professionally cleaned to remove bath residue and scale deposits because no inspection method is available to ensure that proper maintenance procedures were previously followed. Also check to see if there is an access hatch for the pump motor. This is occasionally overlooked during the installation, and it is helpful in the event that maintenance is required.

After you flush the bowl, if the toilet is the tank type, lift the tank lid and look inside. If the lid has a cloth cover, be careful when lifting

Fig. 10-10. Air gap between bottom of faucet spout and flood rim of sink prevents the backflow of nonpotable water into the potable water supply.

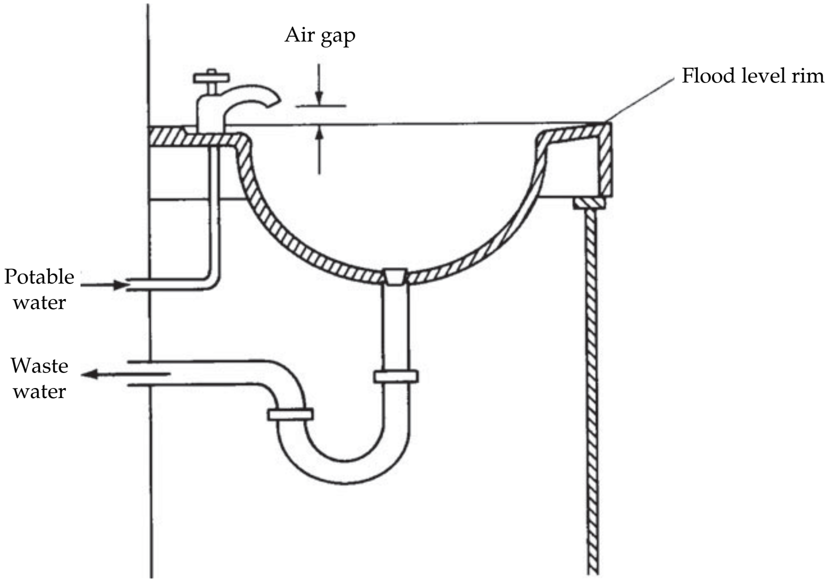

it. It might be cracked or chipped. If water is spurting out of the fill valve, maintenance is needed. You might even find the trip lever connected to the flush valve by a string rather than a metal chain or strip. Repair or replacement of the flushing mechanism is not costly.

Occasionally the seller, not being the original owner, inadvertently misrepresents the age of the house. Depending on the manufacturer of the toilet bowl, you might be able to verify the house's age by looking on the underside of the tank lid. There might be a date stamped on the underside that can be a clue, assuming the toilet is the original and was not installed during a later renovation. The date is the date of manufacture, usually within a few months of installation.

Kitchen

When inspecting the kitchen, in addition to the plumbing items mentioned above, the condition of the cabinets and counters should be checked. The appliances, although important, need not be inspected at this time. Because appliances can break down at any time, it is recommended that on the day of, but prior to, the contract closing, you come back to the house and operationally check every appliance included in the purchase. If an appliance is not operational, you can have your attorney request an adjustment at the closing for the cost of repairs.

Cabinets should be inspected for missing, cracked, and loose-fitting doors and drawers. Missing hardware for doors and drawers should be noted. The shelves should be checked to see if they are adequately supported. The counters should be inspected for cracked, burned, blistered, and loose sections. If there is a cutting board or hotplate on the counter, lift it up or move it aside. You might find that it is concealing a damaged section of the counter.

When inspecting the sink, in addition to checking water flow and drainage, look for a sprayer. If there is one, see if it is operational. I have seen many sprayers with a disconnected hose mounted in the sink fitting. The sprayers

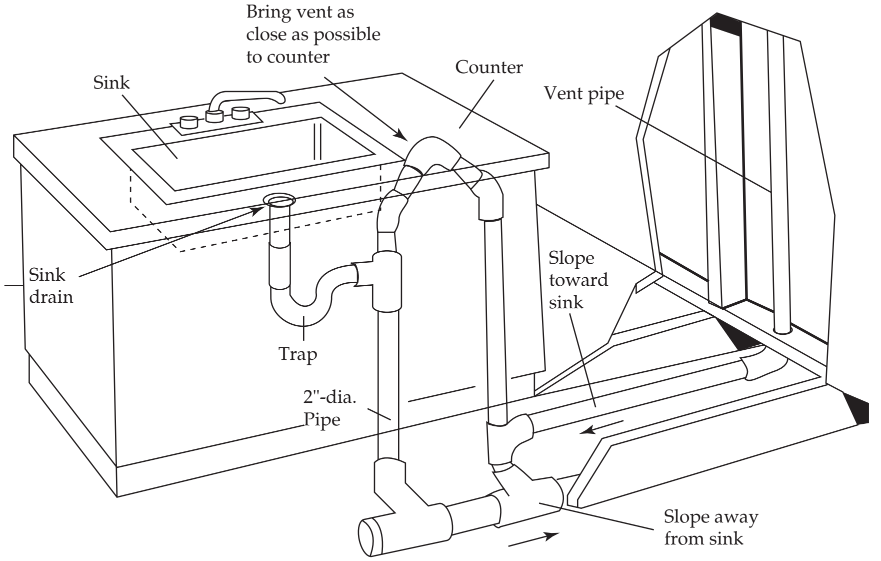

Fig. 10-11. *Loop vent—venting of an island sink.*

were not functional and served only as a decoration to cover the opening in the sink. Is there a sink in a central island cabinet? If there is, check the drain piping. Island sinks are notorious for being improperly vented. If the sink drain has an S-type trap rather than a P-type trap, it is not properly vented. Record this item on your worksheet. If approved by local codes, this condition can be corrected with a minor plumbing change by connecting an air admittance valve (AAV) to the drain. Another way of venting an island sink is with a "loop vent" (see FIG. 10-11), which is usually, but not always, installed during the construction phase of the house.

Is there a garbage-disposal unit connected to the sink drain? If there is and the house has a septic tank, there might be a problem. A garbage-disposal unit introduces solid wastes into the septic system at a greater rate than normal. To avoid overloading the system, some states have a design criterion calling for a larger-capacity septic tank when there is a garbage-disposal unit. Other states recommend that the tank be cleaned at more frequent intervals. (See chapter 13 for a discussion of septic systems.)

If the disposal unit was added after the house was constructed and provisions were not made for a larger septic tank or more frequent cleaning, the disposal unit might have been overloading the septic system. This can result in premature failure of the system. If the house has a garbage-disposal unit and a septic system, check with the local municipal building department to determine if the system was designed to accommodate the wastes from the disposal unit. Also, check with the owner to find out when the septic tank was last cleaned. If it was not cleaned or at least inspected for sludge buildup within the last three years, record the fact on your worksheet. The tank should be cleaned after you move in.

Hallway and staircase

The remaining rooms in the house should be checked as described previously. In addition, the connecting hallway should be inspected as you walk from one part of the house to another. The hallway should be treated as an interior room, and its walls, floor, ceiling, and trim should be inspected. Look for an overhead light in the hall. Is it controlled by three-way switches located at both ends of the hall? It should be. As a fire-safety measure, a smoke detector or equivalent should be mounted on the ceiling of the hallway leading to the bedrooms. If there is a smoke detector, is it properly located? The corners of a hallway where the walls and ceiling meet is considered dead-air space. This means that even though smoke will circulate and accumulate near the ceiling, it will not penetrate into those corners until the hall is completely filled with smoke. According to the National Fire Code: "Spot-type smoke detectors shall be located on the ceiling not less than six inches from a side wall, or if on the side wall, between six to twelve inches from the ceiling." Neither wall- nor ceiling-mounted detectors should be placed near a light fixture or a ventilation grille that could block smoke from reaching the detector. If you see an incorrectly located smoke detector, you should alert the seller and record it on your worksheet.

As you proceed from one level to another, check the connecting staircase. Squeaky treads, as with squeaky floors, indicate loose sections. They are not a concerning factor, but if excessive, they are annoying. Correcting the problem is often difficult because the underside is usually concealed. The steps should be uniformly spaced without any dimensional variation. It is not uncommon to find basement steps with uneven risers for the bottom and top steps. This condition is a potential hazard, since someone can easily trip.

As a safety precaution, all steps should have handrails. Check to see if both ends of the handrail have a return to the wall. Rails with returned ends will prevent clothing from catching on the end, which could result in a fall. Loose handrails must be resecured. Sometimes the handrail is secured in such a manner that allows little finger room. This is a hazardous condition, especially for children. Is there a window at the base of the staircase or on the landing? If there is, its sill should be at least 36 inches above the floor. This will prevent someone from falling through the window in the event of a fall down the stairs. If the sill is less than 36 inches high, a window guard should be installed as a precautionary measure.

Lighting the stairway is essential, especially for areas that are often neglected, such as basement steps. The location of the light fixture is unimportant as long as the light provides sufficient illumination for the entire staircase. The light should be controlled by three-way switches, one at the top and the other at the bottom of the staircase.

Checkpoint summary

Walls and ceilings

- ❍ Do not be concerned with minor cosmetic problems. Specifically check for:
  - — broken walls and ceilings;
  - — loose, missing, and bulging areas of plaster or plasterboard;
  - — missing, loose, and sagging sections of ceiling tile;
  - — sagging sections of plaster;
  - — truss uplift cracks (at wall-ceiling intersection);
  - — water stains on ceilings, particularly below roofs, bathrooms, and kitchens;
  - — disintegrating plaster, peeling and flaking paint, and water stains on walls facing the exterior (usually caused by open or exposed exterior joints).
- ❍ If there is a sulfur-based odor in the house, suspect Chinese drywall, especially if the copper plumbing and electrical wiring show signs of corrosion.

- ❍ Note rooms that have cracked walls in which the windows and door frames are not level (professional evaluation is recommended).
- ❍ Check paneled walls for plasterboard backing (often omitted).
- ❍ Check trim for missing, loose, cracked, or broken sections.

Floors

- ❍ Inspect for floors that are not level, have loose floorboards, or have sagging areas. Sagging floors that are also noted in the ceilings below should be evaluated by a professional.
- ❍ Inspect for large open joints between floor and partition walls (usually caused by excessive shrinkage).
- ❍ Note floors that have wall-to-wall carpeting. Do not assume that there is a hardwood floor beneath the carpeting. Request that owner make this representation in writing.
- ❍ Inspect concrete floor slabs for cracked and settled areas. Note areas that have large open joints between the floor slab and the walls.
- ❍ Check raised wood floors over concrete slabs for soft, spongy, and delaminated sections. Often these conditions are a result of water seepage with associated rot in wood framing.

Heat

- ❍ Check interior rooms for missing radiators or heat registers.
- ❍ If rooms are heated by other means (radiant panels), verify this with the owner.
- ❍ Are radiators and heat-supply registers efficiently located (preferably on an exterior wall and below a window)?

Windows

- ❍ Check windows for ease of operation.
- ❍ Inspect for cracked and broken panes; chipped, cracked, and missing putty.
- ❍ Check exterior sills for cracked and rotting sections.
- ❍ Inspect double-hung windows for broken or missing sash cords, loose or binding sashes, and missing hardware.
- ❍ Inspect steel-casement windows for cracked panes, rusting and sprung frames, loose, missing, or inoperative hardware.
- ❍ Check thermal-pane windows for faulty seals (water droplets or cloudy areas between the glass panes).
- ❍ Check casement and awning windows for interior storm windows and screens.

Electrical outlets

- ❍ Inspect rooms, stairways, and hallways for electrical hazards and violations (see chapter 12 summary).
- ❍ Note all rooms in which there are insufficient outlets and outlets that are loose or have missing cover plates.
- ❍ Check whether outlets are electrically "hot" and properly grounded. Do any have reverse polarity?
- ❍ Are the kitchen and bathroom receptacles GFI protected?

Fireplace

- ❍ Inspect brick or stone firebox lining for cracked, chipped, broken, or deteriorating sections.
- ❍ Check for cracked, loose, or disintegrating mortar joints.
- ❍ Check top of firebox for an operational damper.
- ❍ Is there a chimney top damper?
- ❍ Inspect area for obstructions.
- ❍ Is the chimney flue lined?
- ❍ Check for heavy layers of soot and creosote (heavy layers indicate the need for chimney cleaning).

Bedrooms

- ❍ Check that all bedrooms have at least one openable window with the following criteria:
  - — sill height not *more* than 44 inches above the floor;
  - — minimum openable area of 5.7 square feet with no dimension less than 20 inches.
- ❍ Do all bedrooms have entry doors and closets?
- ❍ Is there a smoke detector on the wall or ceiling?
- ❍ If a portion of the attic or basement has been converted to a bedroom, is there a certificate of occupancy for the room?

Bathrooms

- ❍ Check bathrooms for adequate ventilation.
- ❍ If there is an exhaust fan
  - — is it operational?
  - — does it have a separate on-off switch?
  - — can you determine where the fan exhaust discharges?
- ❍ Inspect tiled areas, particularly around the tub or shower, for open joints, cracked, loose, and missing tiles.
- ❍ Note wall areas in the tub or shower that show evidence of deterioration (spongy or loose sections).
- ❍ Check shower doors for cracked panes (should be safety glass) and ease of operation.
- ❍ Check sinks, bowl, and tub or shower for cracked, chipped, and stained areas. Check that sinks and bowls are properly secured.
- ❍ Inspect sink and tub faucets for proper air gap (potential back siphonage).
- ❍ Check fixtures for individual shutoff valves.
- ❍ Inspect fixture plumbing for leaks, kinked lines, patched and makeshift corrections (taped joints and rubber-hose connections).
- ❍ Inspect sink drain lines for improper venting (S-type traps).
- ❍ Does shower diverter valve in the tub leak?
- ❍ Check operation of whirlpool bath.
- ❍ Is there an access hatch for the pump motor for the whirlpool bath?

Water pressure, flow

- ❍ Check cold-water flow by simultaneously turning on the faucets in the sink and tub or shower and flushing the bowl.
- ❍ Perform a similar check for hot-water flow.
- ❍ Check for water hammer when faucets are opened and closed rapidly.

Kitchen

- ❍ Check sink for low water flow and proper drainage.
- ❍ If there is an island sink, is the drain vented properly?
- ❍ If sink contains a sprayer, check operation (often this unit is disconnected).
- ❍ If sink drain contains a garbage-disposal unit and house has a septic tank, you should determine the following:
  - — Was disposal unit added after the house was constructed? (Septic tank might be undersized.)
  - — When was septic tank last cleaned? (If over three years, the tank should be cleaned.)
- ❍ Inspect cabinets for missing, cracked, or loose-fitting doors and drawers.
- ❍ Check for missing hardware on cabinet doors and drawers.
- ❍ Check shelving for adequate support, cracked, warped, or missing sections.
- ❍ Inspect counter and countertops for cracked, burned, blistered, or loose sections.
- ❍ Check all appliances for operational integrity on the day of, but prior to, contract closing.

Hallway and staircase

❍ Check for properly located smoke detectors in hallway areas leading to the bedrooms.

❍ Check hallways and staircases for adequate lighting. Are three-way switches located at both ends of the hallway?

❍ Check walls, floor, ceiling, trim, and so on as you would for interior rooms.

❍ Inspect stairways for uneven risers, loose treads, missing handrails, and handrails with tight finger room. Do the ends of the handrail have a return to the wall?

❍ If there is a window at the base of the stairway or landing, is the sill less than 36 inches above the floor? (If this condition exists, you should install a window guard.)

11

Basement and crawl space

Foundation 147
  Settlement 148
  Inspection 149
    Cracks 149
    Superior Walls 150
    Structural support framing 151
    Engineered lumber 153
Dampness 154
Water seepage—causes and control 155
  High groundwater level 155
  Hydrostatic pressure—walls 156
  Inspection 158
    Seepage indications in an unfinished basement 159
    Seepage indications in a finished basement 163
    Walls 160
    Floors and floor joints 160
Furnace room 163
  Ventilation 165
Crawl space 165
  Conditioned crawl space 167
Checkpoint summary 168

Problems in a basement or crawl space are often among the most costly to correct. Look specifically for signs of water penetration, structural deterioration of the wood support members, and structural deficiencies of the foundation wall.

Foundation

The purpose of the foundation is to support the main portion of the house and transmit its load to the ground. The foundation of most residential structures consists of walls that rest on an enlarged base called a *footing*. The footing spreads the transmitted load directly to the supporting soil and is usually resting on undisturbed earth. Where there are freezing temperatures, the footings must be located below the frost line. Otherwise, the footings become vulnerable to frost heaving resulting from the freezing of soil moisture. Pilasters, piers, and columns are also used to support the main structure. (See FIG. 11-1.)

Foundation walls are normally designed to support the vertical loads from the house and to resist the horizontal forces resulting

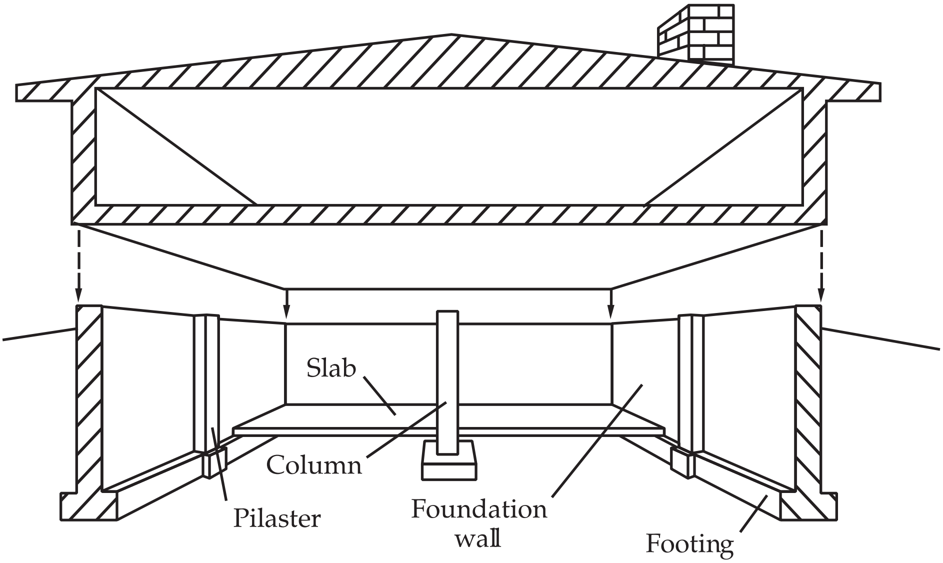

Fig. 11-1. *Elements of a foundation.*

from the earth's pressure. In most parts of the country, the house must be anchored to the foundation to provide resistance to wind and/or earthquake forces. High wind forces or rapid ground movement can cause a structure that is not properly anchored to lift, shift, or rotate slightly. When the house is properly anchored, the resistance to this movement produces stresses that are within the foundation wall. Consequently, the foundation wall must be designed to withstand these forces. In some cases, the combined vertical and horizontal forces acting on the foundation wall is great enough to require additional bracing and stiffening of the walls. In these cases, pilasters are often used to provide the additional support.

Settlement

All soils compress to some extent, and if the footing is not resting on bedrock, the foundation is subject to settlement. When the settlement is slight or uniform, or was anticipated in the design, it is of little concern. However, when the foundation settles unevenly (*differential* settlement), there is generally some cause for concern. Differential settlement introduces stresses that might seriously weaken the building. It often causes cracks in the foundation walls, unsightly cracks in the finished walls and ceilings, sloping floors and windows, and doors that bind. When the differential settlement is excessive, it can result in a structural failure of a portion of the foundation wall, a condition that can be quite costly to correct.

The principal cause of foundation settlement is a reduction of the volume of the voids in the soil supporting the foundation. The voids are the spaces between the soil particles and contain air and water. Sandy soil contains large granular particles with a relatively small volume of voids; clay soil contains fine-grained particles with a large volume of voids. The settlement of foundations built on a sandy soil tends to be quick and slight, whereas the settlement of foundations built on a clay soil tends to be greater and might occur over many years.

Usually you do not have to worry about foundation settlement in older homes because by the time you are considering buying the house, the settlement has already occurred, along with any accompanying problems. And what you see is what you get. However, if what you see is a house out of plumb, with sagging sections, it might have structural problems, and as a precautionary measure, you should have the house evaluated by a professional.

Some soils are considered poor for building-construction purposes. Homes built on a highly organic soil are vulnerable to excessive settlement when the soil dries out. Organic soil acts like a sponge, and as the soil dries, the organic matter shrinks. Depending on the amount of organic matter in the soil, the condition can cause differential foundation settlement.

Some soils on steep hillsides are subject to severe slippage. When the amount of buildable soil above an impervious subsoil layer is shallow, if the soil becomes saturated (because of an extended heavy rain), the entire upper layer of the hillside can slip to the bottom, along with any houses built on the side of the hill.

Some soils with a high clay content can swell or shrink up to 50 percent between wet and dry conditions. These soils are considered unstable, and unless special provisions are made during construction, the condition can cause excessive differential settlement cracks in the foundation.

If you have any questions about the soil the house is built on, you can usually get them answered at the local office of the U.S. Department of Agriculture, Soil Conservation Service, or at the office of the local county soil and water conservation district.

Inspection

Most houses built in the last seventy years have foundation walls of poured concrete, concrete blocks, or cinder blocks, whereas older homes generally have stone or brick foundation walls. When inspecting stone or brick foundations, especially in homes built before or around 1900, pay particular attention to the mortar joints. I have found that in homes of this vintage, many of the mortar joints in the foundation walls have deteriorated to a point where they are no longer functional. Some of the joints might have holes, and some may be filled with soft, crumbling mortar that can easily be raked out. In some cases, the joints between the foundation stones might have been filled with earth or mud rather than mortar.

Deteriorated mortar joints in foundation walls should be repointed, not so much because they represent a weakened structural condition (they do, but this is usually not a severe problem) but because they can allow water to penetrate into the basement (see FIG. 11-2) and enable mice and other pests to enter the area.

In some older homes with brick foundation walls, you might find soft, crumbled, and flaking bricks so badly deteriorated that they should be replaced. This condition is usually the result of using underburned bricks in the construction of the wall. Bricks that are underburned during the manufacturing process are softer and absorb water more readily. If you see this condition, it should be recorded on your worksheet, since masonry rehabilitation is needed.

Cracks Cracks in poured concrete or concrete-block foundation walls can result from shrinkage, differential settlement, lateral pressure on the wall by the soil, or poor-quality workmanship. It is not uncommon to find short cracks in foundation walls. These cracks might be vertical, horizontal, inclined, stepped, smooth, or irregular and are usually of no structural concern. All cracks, however, should be sealed as a precautionary measure

Fig. ***11-2.*** *Deteriorated mortar joints in fieldstone foundation wall. Note water seeping through the wall at base.*

against water penetration into the basement. Otherwise, if there should be a hydrostatic pressure buildup in the soil against the foundation, water will seep through the cracks.

If there are long narrow cracks in the wall and both sides of the cracks line up so that there is no noticeable differential settlement, it is usually not a serious condition and can be controlled by sealing the cracks. However, when both sides of the cracks do not line up or there are long open cracks, a more serious condition of differential settlement exists. Since it is not possible to determine from a single inspection whether the differential settlement is active or dormant, the wall should be checked for incremental movement over a period of time, usually several months. In most cases, after some time, the differential settlement stabilizes with little effect on the house other than functional annoyances such as binding windows or a floor that might not be level. However, because of excessive settlement, an unstable condition can occur either in the foundation wall or the wood framing being supported by the wall. If there is any doubt in your mind about the condition of the foundation wall, you should have it checked by a professional.

Another type of crack of concern is a long, open horizontal crack in the foundation wall, especially if the wall shows signs of bowing. This condition is principally caused by an excessive horizontal pressure being exerted on the foundation wall by the earth backfill and indicates that the wall cannot adequately withstand these external lateral forces. Structurally the wall has failed; this condition should be recorded on your worksheet as a condition that requires further investigation and depending on the extent of bowing, reinforcement, or replacement.

Superior Walls In the 1980s, insulated, precast, ribbed concrete wall panels called *Superior Walls* were introduced to the market to be used for building foundations. (See FIG. 11-3.) The panels can be installed in a fraction of the time that it would take for poured concrete

Fig. 11-3. *Installing precast concrete "Superior Walls."*

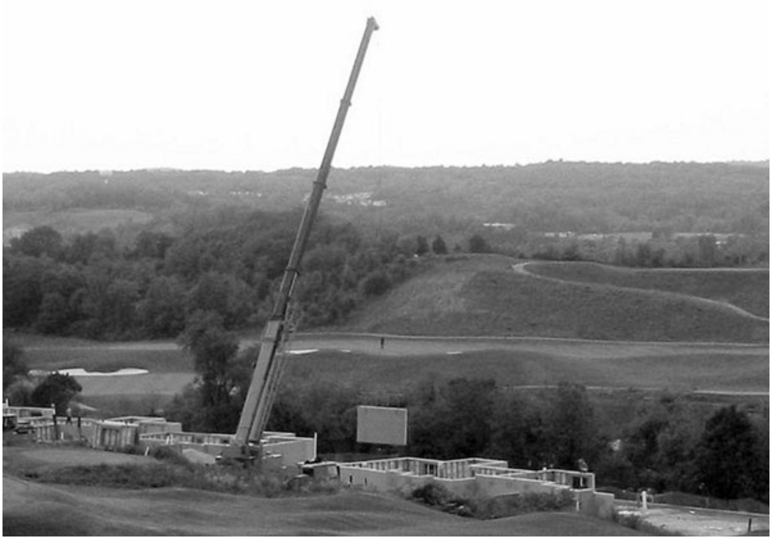

foundations and traditional exterior walls. According to the manufacturer, when panels are delivered to the site, they are often erected in three or four hours. When installing a foundation, the Superior Walls are bolted together at the top and the bottom of each panel, and the joints between the panels are sealed with a polyurethane sealant. When inspecting the foundation, check for open joints and signs of wall tilt, and as with other foundations, check for cracks and signs of water penetration.

Structural support framing Because of the vulnerability to deterioration from rot and wood-destroying insects, all exposed wooden support members (girders, joists, posts, and sill plates) should be checked for structural integrity. Steel beams and columns, on the other hand, need be checked only for degree of rust or corrosion. Usually the rusting is only a surface defect that can be corrected by scraping, priming, and painting.

Wooden support members should be probed with a screwdriver or an ice pick, as described in chapter 8. If the wood is in good condition, the probe will not penetrate much below the surface. The portions of the joists and girders that are most vulnerable to deterioration are those resting on or adjacent to the foundation wall. While probing these areas, you should also check the sill plate (which is anchored to the top of the foundation) for structural integrity. Many older houses have wood support posts rather than steel columns. If there are wood posts, check their bases for decay. The base has a tendency to rot because of periodic dampness or seepage through the floor. A post with a rotted base should be replaced. Joists with deteriorated end sections, however, need not be replaced. They can be rehabilitated. In most cases, all that is needed is to place a similar size wood member alongside the affected joist and secure it to the portion of the joist that has not deteriorated.

Girders (the main support beams) are often supported at both ends by the foundation wall. If the girder is wood-constructed, the notch in the foundation wall in which the girder rests should be large enough to allow the sides and end of the girder to be ventilated. Otherwise, there can be a moisture buildup, which will promote decay. There should normally be at least a ½-inch clearance around the sides and end of the girder. For adequate support, there should be a minimum of a 3-inch end bearing in the foundation notch. (See FIG. 11-4.)

If you find that the end of the girder has deteriorated so that it is no longer providing adequate support, do not be overly concerned. The condition can be corrected at a reasonable cost without replacing the girder. In many cases, all that is needed is to support the girder by a column or pier located near the foundation wall. This repair, however, should be performed by a professional, since the column or pier must have an adequate size footing to spread the load.

When girders or joists have sagging sections, these are usually noticeable on the floors above. During your interior inspection, if you find floors that are not level, you should be alerted to the possibility of sagging support beams. Sagging sections can be braced with adjustable screw-type columns to prevent further sag. The adjustable columns can be used for releveling sagging sections. This procedure, however, should be approached with caution. The leveling process introduces new stresses that in turn can cause cracking.

During the wood-framing inspection, you might find joists or girders that have been notched at the top or bottom to accommodate the passage of pipes. Notching a beam at the

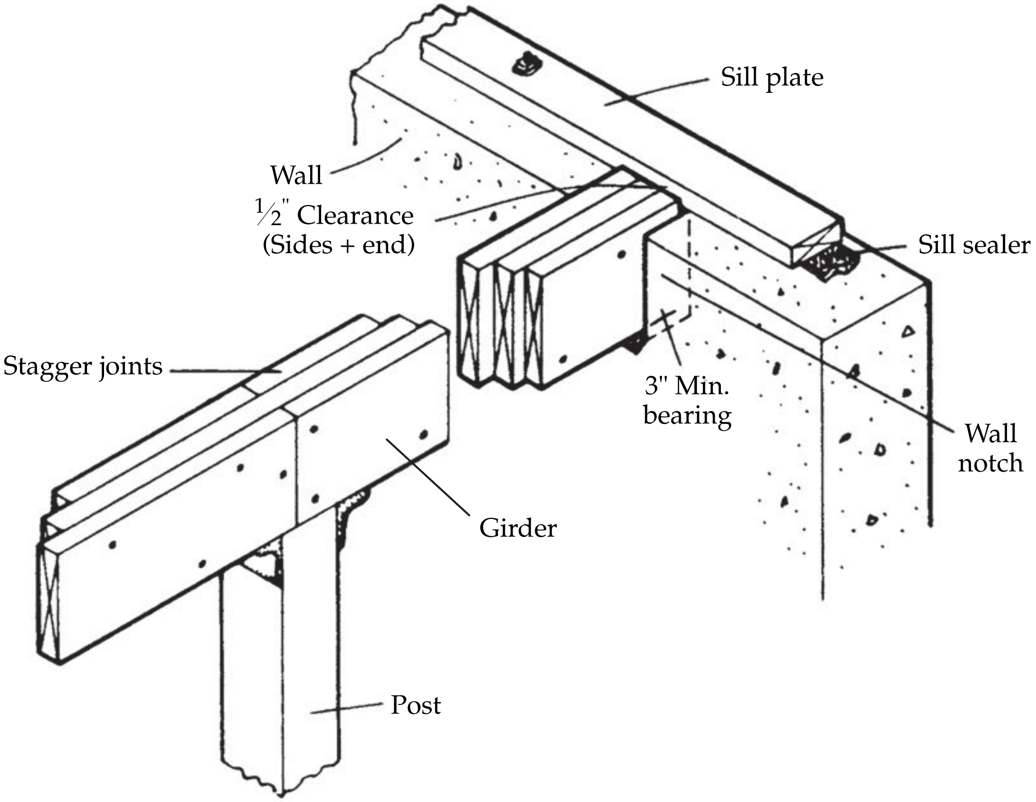

Fig. 11-4. *Notch at top of foundation wall to support the girder.*

top or bottom will reduce its strength. Just as a chain is as strong as its weakest link, a beam is as strong as its narrowest section. For example, if a 2-by-8-inch joist is notched to a depth of 2 inches, its strength will be reduced to that of a 2-by-6-inch joist. When a beam deflects, the top of the beam is in compression and the bottom is in tension. At a point approximately midway between the depth, the stresses change from one to the other so that at this point there is no compression or tension. Consequently, if a hole is cut midway between the top and bottom of a beam (as long as the hole is less than one-fourth the depth), there will be no effective reduction in strength. (See FIG. 11-5.) Using the above example, if a 2-by-8-inch joist has a 2-inch hole in its center, it will still be as strong as it was without the hole. If you find any notched beams, look specifically for signs of excessive deflection. This is an indication that reinforcement or bracing of the beam is necessary.

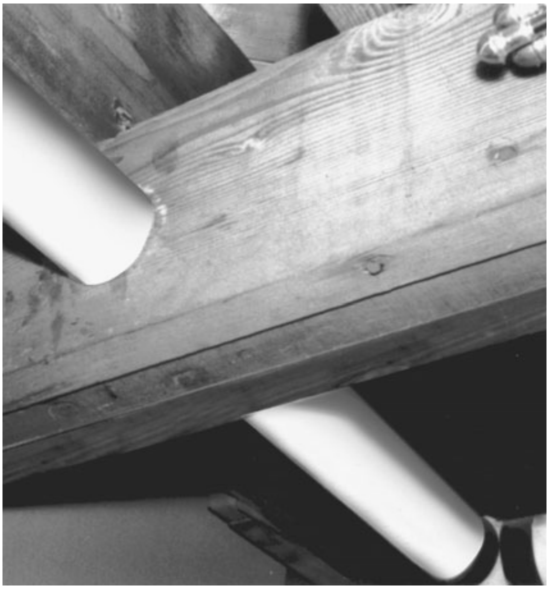

Fig. 11-5. *Hole in center of beam width does not effectively reduce beam strength.*

Engineered lumber Engineered lumber has replaced solid-sawn lumber as the framing and structural support member of choice by builders in many newer homes and in homes that are currently being built. The decision is based on the fact that although the lumber is initially more expensive, the total installed costs are less than that of dimensional lumber. The following products fall in the category of engineered lumber: glued laminated (GLULAM) beams, laminated veneer lumber (LVL), I-joists, and open web truss-joists. The advantage of engineered lumber is that it is stronger, stiffer, and more dimensionally stable than solid-sawn dimensional lumber. GLULAM beams and wood I-joists can carry greater loads over longer spans than is possible with solid-sawn lumber of the same size. Because of their "I" cross-sectional shape, wood I-joists weigh up to 60 percent less than solid lumber joists, making them easier to handle. They generally do not shrink, warp, cup, crown, or twist.

When inspecting the floor framing I-joists, look specifically for problem conditions such as notches in the flanges or holes in the webs that are larger than manufacturer's recommendations. Holes less than 1.5 inches in diameter are generally okay as long as they're not on or into the flange. In certain applications where joists line up under a load-bearing partition, depending on the load, there may be web stiffeners between the flanges of the I-joists. Web stiffeners are used to prevent problem conditions such as the web buckling out of plane or "knifing" through the flange. Also squash blocks are used in situations where there is a load-bearing wall above that is stacked over the wall below. The squash blocks transfer the load from the wall above to the wall below without relying on the joists for this support. (See FIG. 11-6.)

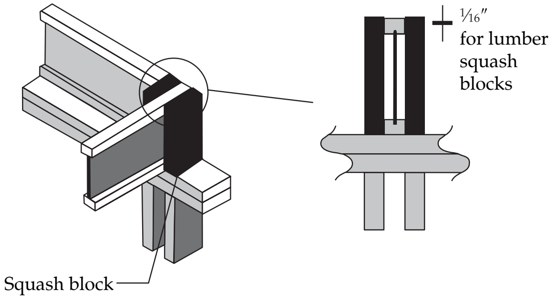

Fig. 11-6. *Squash block installation.*

Check to see if the rim boards are solid lumber or engineered lumber. Only engineered lumber should be used with I-joists. Solid lumber typically has a higher moisture content than engineered lumber, and if used as rim boards, the shrinkage can cause the I-joists to carry the wall load, which could cause web failure.

Most building codes require cross or solid bridging between the floor joists. (See FIG. 11-7.) The purpose of bridging is to hold the joists in a vertical position and to transfer the floor load from one joist to another. Cross bridging consists of pieces of wood or preformed metal set in a diagonal position between the joists to form an X. Solid bridging consists of solid blocks set between the joists. While inspecting the overhead floor joists for rot, insect damage, notching, and sagging sections, you should also check for the presence of bridging. If it is missing, record the fact on your worksheet for later installation.

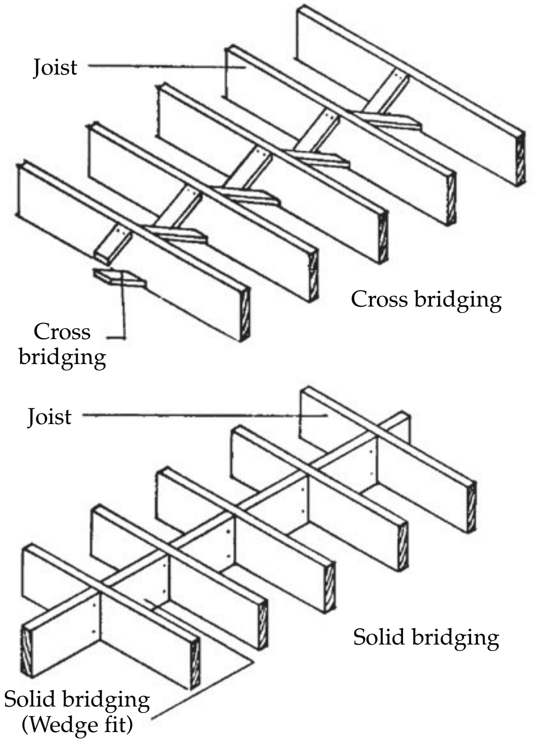

Fig. 11-7. *Bridging. Cross and solid bridging between floor joists. Cross bridging can be wood or metal.*

Dampness

In many parts of the country, the basement or lower level might be damp during portions of the late spring and summer. Dampness in a basement is a normal phenomenon that occurs because cool air cannot hold as much moisture as warm air. It does not necessarily indicate that the basement has a water problem. The temperature of the air in the basement or crawl space during the late spring and summer is always cooler than the outside air. Consequently, when outside air infiltrates into the basement through open windows, doors, cracks, or joints, the temperature of that air drops. This cooler air cannot hold as much moisture and results in a higher relative humidity of the air that entered the basement. Depending on the temperature and the amount of moisture present in the air, some moisture might condense on cool surfaces such as foundation walls and cold-water pipes.

Sometimes water droplets on the foundation wall caused by condensation are erroneously diagnosed as caused by seepage through the wall. If you see a damp-wet foundation wall, you can easily check whether the condition is caused by seepage or condensation. Simply fasten a small piece (4 by 8 inches) of aluminum foil to the foundation wall. Using wide strips of an adhesive tape, seal all the edges of the foil to the wall. After the foil has been on the wall for at least twenty-four hours, examine its surface. If it is moist, the condition is caused by condensation. However, if the foil surface is dry and the area behind the foil is damp, the condition is caused by moisture seeping through the foundation wall from the outside. Of course, it is possible for both the foil and the wall to be damp, indicating both seepage and condensation.

Dampness in a basement or crawl space should be controlled. It can produce conditions conducive to the growth of mold and decay fungi. Dampness can often be detected by musty odors or a clammy, close feeling. In the drier portions of the country, normal dampness (not caused by seepage) in the basement can be controlled by opening the windows and ventilating the area. However, in areas where the climate is hot and humid during the summer, the benefit gained by ventilating the area is lost by the introduction of moist air into the basement. In these areas, the dampness in the lower level can be controlled with one or more electric dehumidifiers. Most of these units have a humidity control that automatically shuts the dehumidifier off when the moisture content in the air reaches a preset level. Depending on the weather and the size of the dehumidifier, the unit might have to run for many hours during each day in order to wring out sufficient moisture from the air so that it is not uncomfortably damp.

Water seepage—causes and control

Depending on the topography, drainage conditions of the soil, and groundwater level (water table), the basement or crawl space might be vulnerable to water seepage. *Water seepage*, as used herein, is a general term that refers to water intrusion into the lower level of the structure. It might manifest itself as a small wet area, a puddle, or layer of water completely covering the floor. If the ground under and around the house is wet, water can seep into the basement through cracks and open joints in the foundation walls or floor slab. Since water seepage can be caused by a number of factors and water can leak into the basement at any number of locations, it is important to determine the cause and source of the seepage so that the proper corrective action can be taken.

For example, if water is entering the basement through the foundation walls, installing a sump pit and pump below the floor slab will not correct the problem. Similarly, if water is seeping into the basement through the floor slab, sealing the walls will not correct the water-infiltration problem. All too often the unsuspecting homeowner is talked into a full waterproofing job, which can cost several thousand dollars, when all that might be needed is to redirect the water discharging from the roof drainage system (gutters and downspouts) so that the water does not accumulate around the foundation.

High groundwater level

Water entering the basement through the floor slab is an indication that water pressure is being exerted on the underside of the floor. When the level of the water below the house is sufficiently high (due to a seasonal high water table or improper drainage) so that it pushes on the underside of the floor slab, it seeps into the basement through cracks, open joints, or porous sections of the slab. If the pressure is great enough, it can cause the floor to crack and heave.

If the seepage is minor, it can often be controlled by sealing cracks and open joints with a hydraulic cement and coating any porous areas of the slab with a cement-base or epoxy sealant. However, a better solution would be to lower the level of the water below the floor slab. This can be done by installing a sump pump below the slab. Subsurface water then flows into the sump pit in the manner of water flowing into a hole dug at the seashore. The water is then removed by the pump and discharged either into a storm drain or at a point sufficiently far from the house so that it will not be absorbed by the ground and flow back under the basement floor slab.

Depending on how the floor slab was constructed, a single sump pump might or might not be adequate to lower the level of the subsurface water. In areas with a seasonal high water table, a concrete floor slab should be installed over a gravel base. Water that accumulates below the slab can then flow through the voids between the gravel and drain away or flow into a sump pit. However, in many houses the floor slab has been installed directly over soil with poor drainage characteristics or over an inadequate gravel bed. In this case, water in the saturated area below the slab will not readily flow into a sump pit, and to control the water buildup, it is necessary to install a series of perforated drainpipes below the floor slab that terminate in the sump pit. Caution should be observed when lowering the level of the groundwater below the basement floor. With some slow-draining soils such as silts and clays, some soil can wash out from around the foundation footing. This can result in unequal settlement, which could crack the walls. Whether a sump pump or drainpipes are needed below the floor slab is an evaluation that should be determined by a professional.

If the house is located in an area with a high incidence of power failures, you should not depend solely on an electrically driven sump pump to control groundwater seepage. It is possible for the power to be knocked out when the water level below the floor slab is rising. As a precautionary measure, there should be an auxiliary backup sump pump in the sump pit. A backup system is particularly helpful in vacation homes where the house will be vacant for extended periods. One type of backup pump is a water-actuated (non-electrical) ejector pump. The Zoeller Pump Company, Louisville, Kentucky, manufactures this type of pump. The pump is connected to the house water supply and is activated by a float control. However, it will be of no help if the water to the house is supplied by an electrically driven well pump. If you install a water-actuated sump pump, it's important that you include a backflow preventer on the water supply because of the potential for contamination as a result of the cross connection. Another type of backup system is a battery-operated sump pump. This system will take over automatically to protect against flood damage when the power fails. The Zoeller Pump Company also manufactures this type of system.

Hydrostatic pressure—walls

Water seeping or leaking through the foundation walls into the basement is due to a hydrostatic pressure being exerted on the walls by saturated soil. This condition is the result of water accumulation around the foundation. The best way to control this type of problem is to minimize the amount of water that accumulates around the foundation. The following are some of the more common causes of water accumulation around the foundation, which can easily be detected and corrected by the homeowner.

- *Missing or defective gutters and downspouts to handle the rain runoff from the roof.* The downspouts must discharge the water away from the structure. All too often, an elbow or splash plate at the base of a downspout is missing, so that the water is discharged directly around the foundation.

- *Improper grading.* The ground immediately adjacent to the structure should be pitched so that it slopes away from the building. Around many homes this area is incorrectly pitched, resulting in surface water (rain or melting snow) collecting around the foundation.
- *Unprotected basement-window wells.* The area around basement windows, if not shielded from rain or serviced with a drain, can easily accumulate water that can leak through window joints or seep down around the foundation.
- *Uneven settlement of walkways or patio.* Occasionally I find that the walkways around the house or the patio have settled and are sloping toward the house. As with improper grading, this condition can cause surface water to collect around the foundation.
- *Leaky garden spigots.* Most homes have exterior-mounted spigots for connection to a garden hose. If the valve is faulty or is not tightened properly, water will drip or leak around the foundation. Water dripping at a rate that fills 1 cup per minute results in 90 gallons of water per day accumulating around the foundation. This water enters the basement through cracks or open joints in the foundation wall.

When the house is located on an inclined lot, surface and subsurface water flows toward the house from the higher portions of the lot. In this case, depending on the incline and the amount of water involved, water-flow control measures will include grading the lot on the high side, so that there is a swale to collect and redirect surface water around the house, and installing a French drain (curtain drain) below the ground to intercept subsurface water and direct it away from the house.

If the amount of water that accumulates around the foundation walls is not excessive, it can be prevented from penetrating into the interior by sealing cracks and open joints on the inside walls with a hydraulic cement and then coating the walls with a cement-base or epoxy sealant. Coating the wall is particularly helpful when the wall is porous, like a cinderblock wall. However, when an excessive amount of water accumulates around the foundation wall, as with a poorly drained soil such as clay, waterproofing the exterior surface of the basement walls might be more effective than treating the interior surface. In addition, a perforated drainpipe is normally installed near and parallel with the foundation footing. (See FIG. 6-5.) The purpose of this footing drain is to carry away water that is accumulating around the foundation and thereby reduce the hydrostatic pressure.

For the footing drain to operate properly, it must have a free-flowing outlet. I know of several cases where builders installed faulty footing drains around houses during construction. The problem was that the drains completely encircled the houses like a doughnut and had no free-flowing outlets. These footing drains were of absolutely no value. Even though initially a footing drain might function properly, over the years it can malfunction because the perforations in the drainpipe or the outlet become clogged. Also, many a footing drain has been damaged during a later modification of or addition to the structure. If the house has a footing drain, you should ask the owner to show you the location of the outlet. The drain outlet should be kept clear and should be checked occasionally during a heavy rain to ensure that it is operating properly.

Even though waterproofing the exterior surface of the foundation wall is more effective than treating the interior surface, quite often an interior treatment is chosen because of the costs involved in excavating around the foundation and temporarily relocating trees and shrubbery. For excessive water accumulation around the foundation, an interior treatment includes sealing the cracks and coating the walls to make them watertight and installing a drainpipe along the foundation footing below the floor slab that discharges into a sump pit.

Just a word about waterproofing the exterior surface of the foundation wall using a pressure-pumping process that requires no digging or relocation of plantings: *Caution*. In this process, a sealant, pumped through tubes that are inserted into the ground, is supposed to coat the wall and render it watertight. The effectiveness of this treatment depends on the condition and porosity of the ground around the foundation. Since contractors doing this work do not always take test borings and analyze the soil, the process is usually not effective, and additional measures are invariably necessary.

Inspection

The inspection for water seepage into the basement or crawl space should begin during your exterior inspection. As you walk around the house, record on your worksheet the location of those conditions that can cause water to accumulate around the foundation: faulty gutters and downspouts, improper grading, settlement of walkways, and so on. When you go into the basement, if there are problem conditions on the exterior, the walls and floor opposite those areas should be checked first for signs of water penetration. Figure 11-8 shows water stains and deposits in the corner of a foundation wall as a result of a faulty downspout. Even if there are no indications of past or current water seepage, the exterior problem conditions should be corrected.

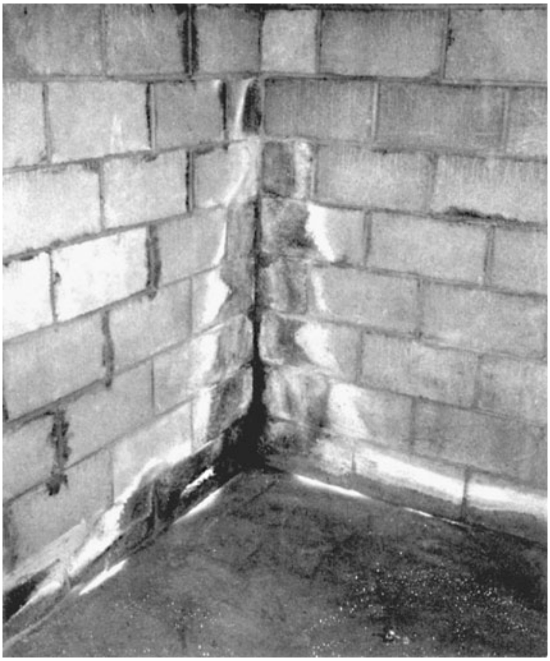

Fig. 11-8. *Water stains and mineral deposits (efflorescence) in the corner of a foundation wall.*

A basement can have a water-seepage problem and be dry when you look at it. Many clients have told me that there were no seepage problems in the basement. After all, they looked at the basement during a heavy rain and found it "bone dry." So how could there be a problem? Well, as previously discussed, there are many causes for water seepage, and depending on the cause, a single rain might not result in seepage. For example, if water penetrates into the basement through the floor slab as a result of a seasonally high water table and the basement is inspected when the water table is 1 or 2 feet below the high level, a heavy rain will not raise the groundwater level sufficiently to cause water to seep through the floor slab. It takes time for rainwater to percolate into the ground and raise the water table.

Water puddles or flooded areas in the basement are obvious signs of a water problem. In most cases, however, you will not see standing water, and you must then make an evaluation of whether there is a condition of water intrusion based on other, more subtle signs. Water-seepage signs indicate only that water has seeped into the basement in the past. They do not indicate the frequency of the seepage or its extent. Consequently, if you see indications of water seepage, you should not engage a contractor to waterproof the house immediately upon taking possession. If you do, it could prove quite costly.

First talk with the homeowner about the condition. It is possible that whatever it was that caused the past seepage has already been corrected. (If the problem was corrected by installing buried drainpipes or coating the outside surface of the foundation wall, the correction would not be visible.) If the homeowner indicates that the problem has been corrected, you should ask to see a copy of the paid bill. Or get the name of the contractor so that you can call to find out exactly what corrective steps were taken. Quite often a contractor provides a guarantee against water seepage. If there is such a guarantee, you should find out whether it is transferable.

The possibility exists that even though there are signs of water seepage, the actual seepage might occur very infrequently—such as only after an excessively heavy rain as might occur every few years. In this case, depending on the extent of the seepage and the projected usage of the basement, costly waterproofing measures might not be justified. The best approach when considering the correction of water seepage is to correct immediately any obvious problem conditions such as faulty gutters and downspouts, improper grading, cracks through which water is actively leaking, and so on. However, before undertaking any major water-seepage control measures, such as excavating and coating the exterior surface of the foundation wall, inserting perforated drainpipes below the floor slab, or trenching and installing buried drainpipes in the yard, you should live in the house for at least one full year. This will enable you to evaluate the degree and extent of the seepage over a full weather cycle. If it turns out that the year is particularly dry so that there is no seepage, well and good. Wait another year. By not taking a "shotgun" approach and waterproofing everything, as recommended by many contractors, you might be able to resolve the problem at a cost that truly reflects the work needed to stop the seepage.

Seepage indications in an unfinished basement When looking for indications of water seepage, you should check the walls, the floor, the joint between the walls and the floor, and the bases of all the items stored or standing on the floor. Specifically, look for white powdery deposits on masonry foundation walls and floor. (See FIG. 11-9.) The

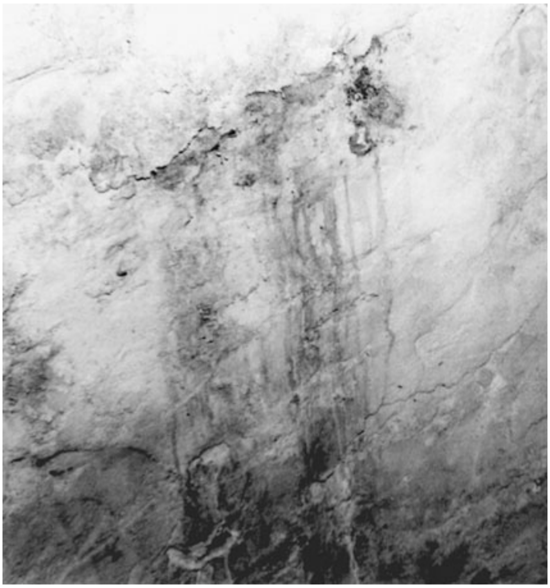

Fig. 11-9. *Efflorescence and water stains on foundation wall.*

deposits, called *efflorescence*, are mineral salts in the masonry that dissolve in the water as it passes through the walls or floor. When the water evaporates from the surface of the walls or floor, it deposits these salts. A thick layer of efflorescence is usually an indication of considerable seepage.

Walls Look for efflorescence, peeling and flaking paint, and scaling sections (surface deterioration) on the foundation wall. Any one of these items can indicate some degree of seepage. Porous walls, such as those made of cinder blocks, may have damp spots. Masonry-block walls are constructed with interior voids. When the hydrostatic pressure on the exterior portion of the wall is high, the voids often fill with water. As a result, the wall might be quite wet to the touch. (Caution: This might also be caused by condensation.) Vulnerable areas for seepage are cracks and the joints around pipes passing through the wall, such as the inlet water pipe and the drainpipe leading to the sewer. Look closely at these areas for water streak stains and efflorescence.

A poured concrete foundation wall is supposed to be more watertight than a concrete-block wall. This, however, assumes that the poured concrete wall is properly constructed. Quite often, it isn’t. If the entire wall is not constructed with a single pouring, the joints between the sections constructed with each pour are vulnerable to water leakage. I inspected a new house just after construction was completed. The inspection was performed during a heavy rain, which was opportune, although not planned. While inspecting the basement, I found water leaking out of the joint at the seam between the individually poured sections of the foundation wall. The builder’s mason apparently had not properly prepared the joint for a new pour, and consequently a cold joint with a poor bond was formed.

Look for seepage in a poured concrete wall around the tie-rod holes—holes in the concrete wall around the small-diameter metal rods that are used to hold (tie) the forms together when the wall is being poured. More often than not, these holes have been patched over. Also, these tie rods can corrode away over a period of time and when below grade are vulnerable areas for water intrusion. Sometimes you see efflorescence and water streaks just under the hole or patched sections. Occasionally I find these holes plugged with corks. This is not considered a permanent patch, and if seepage should develop, they should be plugged with hydraulic cement.

Floors and floor joints A vulnerable place for water seepage is the joint between the foundation wall and the floor. Look closely at this joint as you walk around the entire basement. Water stains and efflorescence are an indication of seepage. (See FIG. 11-10.) You might find silt deposits at the joint. This is also an indication of some degree of seepage. The fine silt is in suspension in the water as the water seeps in from the exterior. When the water evaporates, the silt is deposited. If you find evidence of water seepage through the joint between the floor slab and the foundation wall, you should record it on your worksheet for later correction. The joint should be sealed with either hot tar or a hydraulic cement.

Cracks tend to develop in the floor slab near the base of metal columns. Look for these cracks and any other cracks in the floor slab. Specifically, look for water stains, efflorescence, and silt deposits. Cracks in basement floors are a common phenomenon and are generally caused by slight settlement or shrinkage in the concrete. Usually they are of no concern other than the fact that water can seep through them. Therefore, they should be sealed. However, if the cracks are extremely wide or show evidence of heaving, they are of concern and should be checked further. In all

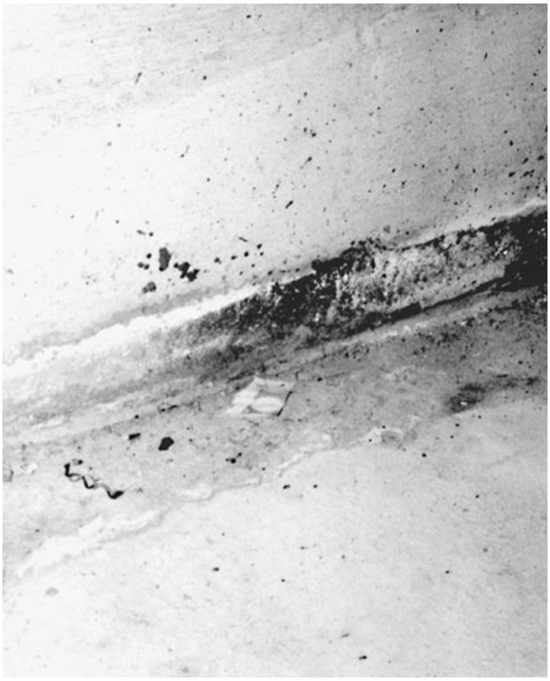

Fig. 11-10. Signs of water seepage at joint between floor slab and foundation wall.

probability, a cracked and heaved floor slab is the result of water pressure being exerted on the underside of the slab by a high water table.

Some homes have a cleanout and trap for the house waste line located in a pit below the basement floor slab. The bottom of the pit should be dry. If it is wet, it is an indication of a high groundwater level or a crack in the drain line. Occasionally I find that the top of the cleanout is open. (See FIG. 11-11.) It should normally be plugged. Some homeowners remove the plug so that the open cleanout can function as a drain if the basement becomes flooded. This is definitely not the way to eliminate the water in a flooded basement. If the basement periodically floods and there is no drain in the floor, a sump pump can be installed in the lowest section of the floor. The water can then be pumped out of the basement.

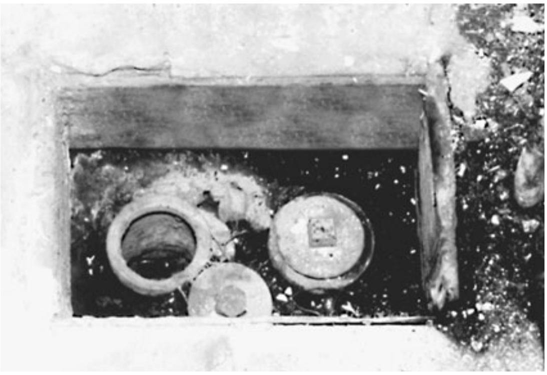

***Fig. 11-11.** If the cleanout plug is removed, sewage can flood the basement.*

With the cleanout plug removed from the top of the house trap, the possibility exists that if the sewer line becomes overloaded (as is the case in some communities), sewage can back up and flood the basement. I know of several homes where this has happened. It was quite unpleasant.

Some people lose interest in a house when they find a sump pump in the basement. They think that the house has water problems. That is not necessarily so. There might have been periodic problems resulting from a seasonal high water table, but the sump pump might have controlled the water level. Or the pump might have been installed when the house was built in order to prevent a problem. To determine whether there still are problems, you must look beyond the sump pump. Look for signs of water seepage.

If the house has a sump pump, look down into the pit. If there is water in the pit and you are inspecting it during the dry season (the water table is usually highest during the spring), the sump pump will probably be operating continuously during the spring. Try to check the sump-pump operation. However, do not actuate a pump unless there is water in the pit. To check the operation of the pump and motor when there is no water in the pit, fill the pit using a garden hose or gently tip in a couple of large buckets of water. If you notice that the water is disappearing into the ground or under the slab before the pump has a chance to activate, wrap a plastic sheet around the wall of the pit. You may have to line the pit bottom with plastic as well. When you actuate the pump, watch the water level to see if it drops. You might find that the motor that drives the pump is operational but the coupling between the pump and the motor is broken. In this case, it will sound as if the pump is working, but it is not, and the water level will not drop. A sump pump is relatively inexpensive and can be repaired or replaced easily.

Although a sump pump is sometimes located in the low section of the floor (when it is being used to collect surface water), it should not be located where it will present a tripping hazard. (See FIG. 11-12.) The pump can be placed in a corner where it will not take up valuable floor space and can be connected to the low spot in the floor by a drainpipe placed below the floor slab that discharges into the sump pit.

Check to see where the sump pump effluent discharges. It should not discharge into the house drain, which in turn discharges into the sanitary sewer or, if there is one, the septic system. As discussed in chapter 6, the sump pump should discharge onto the lawn sufficiently far away from the foundation so as not to seep back into the house.

After inspecting the walls and the floor for signs of water seepage, you should check the bases of items stored or standing on the floor for water stains and rust. You might find that the walls and the floor of the basement appear to be freshly painted. If you see this, you should be suspicious because the new layer of paint will cover almost all the signs of a water-seepage problem. There are some areas, however, that are often omitted when painting the basement. These should be checked for water stains and rust. Specifically, look at the base of the steps leading to the basement and in particular the back of the bottom step. Also look at the base of columns. Wood columns might have stains or may be rotting, and metal columns might be rusting. The base of the inside portion of a

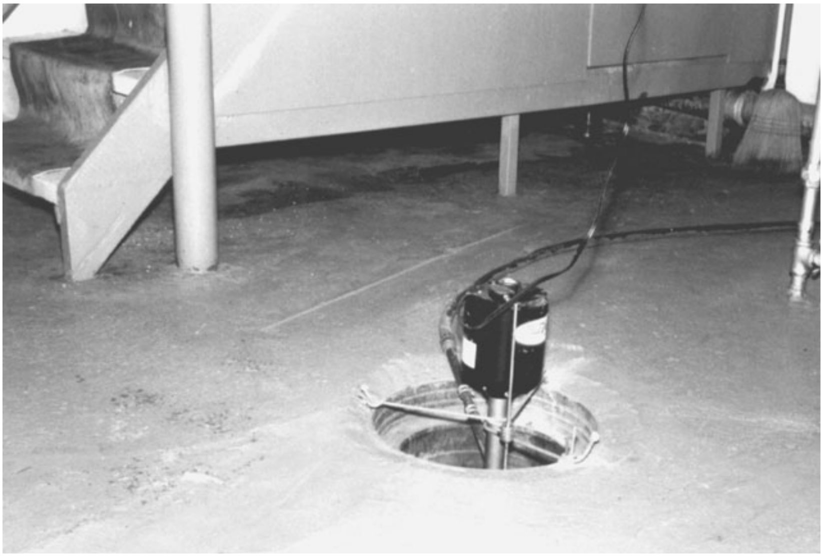

Fig. 11-12. *Sump pit in center of basement floor—a tripping hazard.*

furnace sheet-metal casing is often overlooked when painting. Look inside. Is there extensive rust? If there is, it might have been caused by a past flood. However, it could also have been caused by a faulty humidifier, so do not jump to a quick conclusion.

Seepage indications in a finished basement Usually the finished walls in a basement are a few inches or more away from the foundation walls. Therefore, seepage indications on the foundation wall will not be reflected in the middle and upper portions of the finished wall. Look at the bottom portion of the wall for signs of water intrusion. If the wall is paneled, look for rotted and warped sections and water stains. Sometimes there are grayish mold spots or mold fungi on the walls, a condition caused by excessive dampness. With a plasterboard wall, look for water stains and blackened areas and spots. The latter is mold and mildew. In some cases, the lower portion of the plasterboard wall might have deteriorated.

Next, look at the floor. If the floor is raised above the level of the concrete floor slab, the wood flooring and the wood-framing members used to raise the floor are vulnerable to rot in the event of seepage. The wood framing below the floor should be pressure-treated but often is not. Try to walk all over the floor, especially around the perimeter. If there are any rotting sections, they will feel soft and spongy beneath your feet. If the floor is covered with resilient tiles, the rotting sections might be visible, since the dampness in the wood tends to loosen the tiles. (See FIG. 10-4.)

An area that is particularly vulnerable to water seepage is the portion of the basement that faces a yard where the overall topography is inclined toward the house. Even if the ground adjacent to the house is graded properly, there will still be subsurface water moving toward the house. The walls and the floor of the vulnerable section should always be checked for signs of seepage. Sometimes this is quite difficult; some homeowners might inadvertently (or intentionally) block the area with furniture. If the area is blocked, ask the homeowner if it is all right to move the furniture.

If the floor slab is covered with resilient tiles and there is a problem with heavy seepage through the floor, it is usually noticeable by looking at the tiles. The joints between the tiles become swollen and filled with a white crusting of mineral deposits (efflorescence). Some tiles become loose, and efflorescence is noted below them. Occasionally the tiles are covered with wall-to-wall carpeting. In this case, the tiles cannot be examined for signs of seepage. However, some checking for seepage can be done. Ask the homeowner for permission to lift the edge of the carpet off the tacking strip along a section of the exterior wall. If there is seepage in this area, the tacks will be rusted and the wood strips water stained or rotted.

If the basement is heated with baseboard convectors mounted on an exterior wall, another place to look for signs of seepage is below the convector. If there is some seepage in that area (the joint between the floor slab and the wall), the base of the convector will be rusted.

On occasion a portion of the basement might become flooded as a result of faulty plumbing, an overflowing sink or tub, a malfunctioning water heater, and so on. Flooding from the above is basically a one-time affair and will not cause the type of damage that results from repeated wetting. There will be water stains, but there usually will not be rot, peeling or flaking paint, efflorescence, or heavy rusting.

Furnace room

The heating plant in a house is normally a furnace or a boiler. (See chapter 14.) In this chapter, the heating plant is called a furnace and the room in which it is located is called the furnace room.

In most homes, the furnace is located in the basement or lower level. It can be located in a large open area or confined in a relatively small room. In either case, the area around the furnace is considered a potential fire hazard. Consequently, there should not be any exposed wood framing in the ceiling or partition walls that are in close proximity to the heating plant. (See FIG. 11-13.) Exposed overhead floor joists or wall studs that are near the furnace should be covered with Type X fire-code plasterboard as a safety precaution. Sometimes I find that wood paneling has been installed around the furnace to make the area more attractive. However, it also makes the area more of a fire hazard, especially the portion of the wall around the chimney. (See FIG. 11-14.) If you find wood paneling near the furnace, you might consider covering it with fire-code plasterboard. There should be a minimum 2-inch clearance between the chimney and any wood framing or paneling. Note that the area around the furnace should not be used for storage. On many occasions I have seen combustible items actually stored against the furnace. Fortunately for the families involved, the items never ignited, but they could have!

Many heating systems have prefabricated chimneys that extend from the furnace room up through the interior portion of the structure, terminating above the roof. If the house has such a chimney, look at the joint between the chimney and the ceiling of the furnace room. If there are large openings, they represent a potential fire hazard and should be covered with a noncombustible material such as sheet metal. The open area around the chimney, if not properly blocked, can act as a flue and in the event of a fire in the furnace room, will draw the flames up to the attic.

While in the furnace room, also check for asbestos insulation around the furnace and heating pipes. (Asbestos as a health hazard is discussed in chapter 20.) The door to the furnace may be a louvered door. Depending on

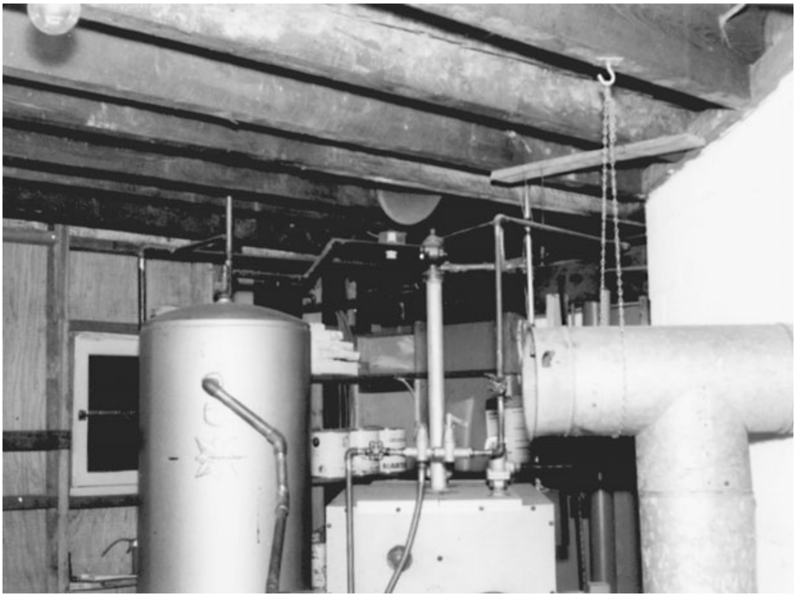

Fig. 11-13. *Exposed wood joists above heating-system boiler—a potential fire hazard—should be covered with fire-code plasterboard.*

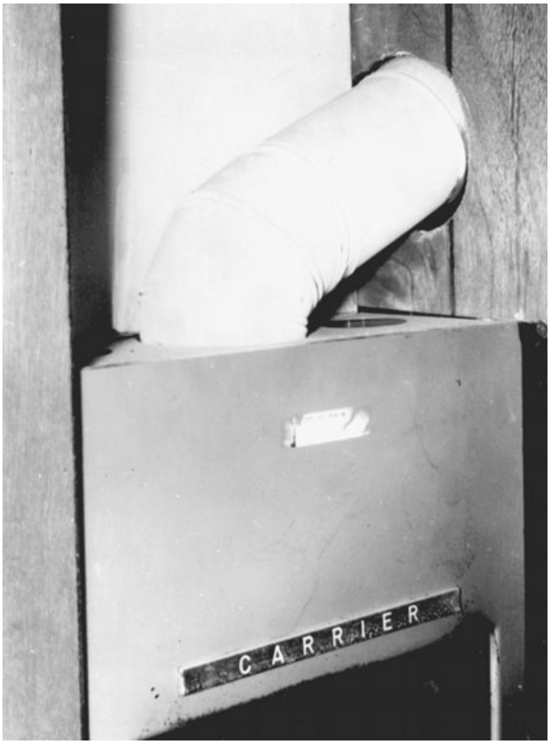

Fig. 11-14. *Wood paneling near furnace—a potential fire hazard.*

the local fire code requirements, louvered doors may not be permitted between the living space and the furnace room. The code may require that the door be fire rated. In this case check the adequacy of the ventilation; see below.

Ventilation

All fuel-burning heating systems must have adequate ventilation for proper operation. If the furnace is located in an unconfined space, the normal air infiltration into the area provides ventilation. However, when the furnace is located in a confined space such as a small room, inlet and outlet ventilation openings must be provided. The vent openings can lead directly to the outside or to a large unconfined area within the structure. The size of the openings depends on the total input (Btu/hour) rating of all the fuel-burning equipment in the enclosure. For most residential structures, an unobstructed inlet and outlet opening of 15 inches by 15 inches is sufficient.

If you would like to calculate the size of the vent openings needed, a safe formula to use is 1 square inch per 1,000 Btu/hour of input rating for both the inlet and outlet vents. The input rating will usually be found on a data plate mounted directly on the equipment. If the vent opening is covered by louvers, remember that metal louvers reduce the effective opening by about 25 percent and wood louvers by about 50 percent. An insect screen covering the louvers reduces the effective opening by another 25 percent.

Some homes have louvered entry doors to the furnace room, which provide the means for ventilation. Occasionally I find that for decorative reasons, the louvers have been covered over, blocking the effective ventilation opening. If the louvers have been covered, look for additional vent openings.

Crawl space

The foundation walls, piers, posts, and wood-support framing in a crawl space should be inspected (as described in the section on the unfinished basement) for deterioration, structural deficiencies, and evidence of water seepage. Pay particular attention to the wood-framing members that are very vulnerable to decay and termite infestation. (See FIG. 11-15.)

Many homes, however, have crawl spaces that are inaccessible and cannot be inspected. These homes were built in accordance with the Federal Housing Administration (FHA) Minimum Property Standards, which permit the ground level to be 18 inches below the bottom of the floor joists and 12 inches below the bottom of the girders. Even if there is an access opening to the area (see FIG. 11-16), the

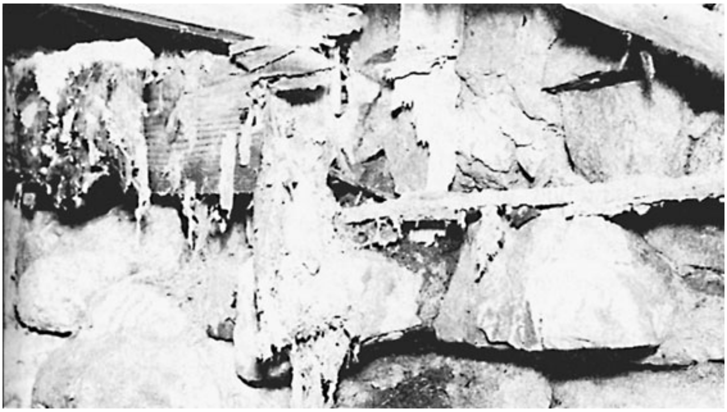

Fig. 11-15. *Cracked and rotting wood-frame members in crawl space.*

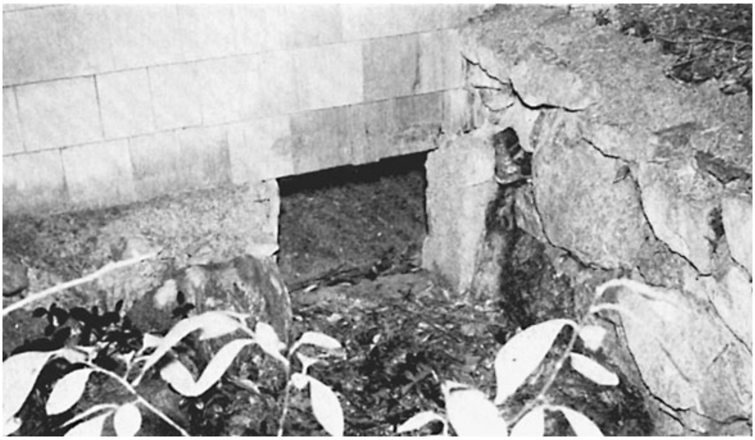

Fig. 11-16. *Access hatch to crawl area.*

clearance in the crawl space is too low for a person to maneuver around easily and perform a detailed inspection. However, you should look into the area from the access opening (using your flashlight) to discover any obvious problems or problem conditions.

In most parts of the country, the crawl space will be quite damp, even though there are no problems with water seepage. The dampness is the result of the capillary rise of ground moisture. Unlike a basement where the ground is usually covered with a concrete floor slab, the floor in a crawl space is often bare earth. Even though the soil might appear dry and dusty, moisture can be present. In some soils, the capillary rise is more than 11 feet above the water table.

Dampness associated with capillary moisture can be effectively reduced by covering the ground with a vapor barrier, such as 4- to 6-mil polyethylene. Roll roofing is also a good vapor barrier, but it tends to deteriorate from fungi. If you do not notice a vapor-barrier ground cover, you should consider its installation.

To help minimize the dampness, the crawl space must be ventilated. There should be at least two vent openings on opposite sides of the foundation with a total free area of 1 square foot for each 1,500 square feet of crawl space area, providing there is a ground cover. When no vapor barrier is used, there should be at least four vent openings (one on each side) with ten times the total free area. Look for vent openings. If you do not see any or they have been permanently blocked, put a note on your worksheet to that effect.

Crawl spaces are usually not heated. Consequently, unless there is some insulation between the floor joists, there will be heat loss between the heated room above and the unheated crawl area. Since moisture from the house can travel down through the floor, the insulation should have a vapor barrier on one side to reduce further moisture entry into the crawl space. The vapor barrier should be located above the insulation, facing the heated room (rather than below, facing the crawl area). If this vapor barrier is located below the insulation, the vapor will condense on its surface during cool weather. Depending on the amount of vapor, the resulting condensation buildup can reduce the effectiveness of the insulation. Look for insulation. You might find missing, loose, or hanging sections, which should be replaced or resecured. (See FIG. 11-17.) If you find uninsulated heating ducts or pipes in the crawl space, record this fact on your worksheet as a reminder to insulate the exposed sections.

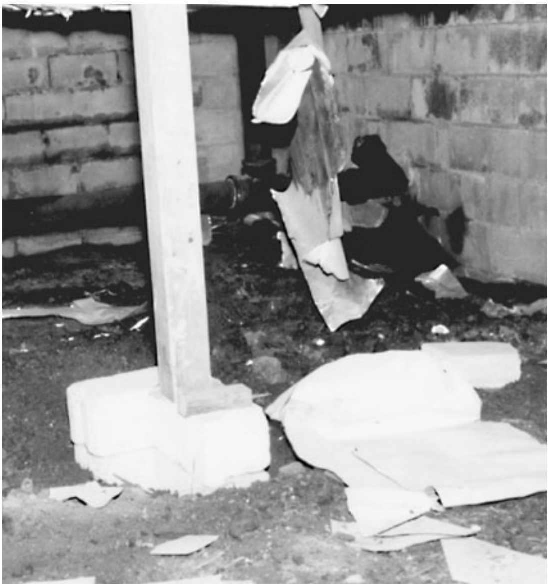

Fig. 11-17. *Loose, hanging insulation in crawl space.*

Some homes are constructed with both a basement and a crawl space. In this case, the crawl space need not be vented to the outside but can be vented to the basement. Look for evidence of water seepage in the crawl space. Even though there might be no signs of seepage in the basement, there might be some in the crawl area. I recently inspected a home that had a combination basement–crawl space. The basement had been painted, and there were no visible signs of a past water condition. The crawl space was separated from the basement by plywood doors that were painted on the basement side and looked good. However, when I inspected the crawl space, I found evidence of a previous water condition. Apparently, the backside of the plywood doors (facing the crawl area) had not been painted over. There were water stains on the lower section. (See FIG. 11-18.) While in the crawl space, check subflooring and support joists below kitchen and bathroom fixtures for evidence of decay and plumbing leaks.

Conditioned crawl space

During your crawl space inspection, you may notice that the foundation walls are insulated, the wall vents are all closed, the dirt floor is covered with large plastic sheets, and the area is heated or air-conditioned. Basically the area is dry and comfortable. In this case the traditional crawl space has been converted to a conditioned crawl space. In the past few years, the concept of a conditioned crawl space has gotten some traction in the building

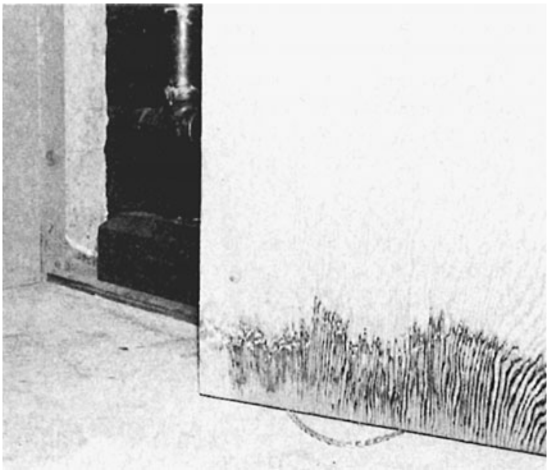

Fig. 11-18. *Water stains on back side of entry door to a crawl area.*

industry. Limited research has shown that a conditioned crawl space is more energy efficient than the traditional vented crawl space and will eliminate some of its problems. With a vented crawl space, warm, moist summer air will enter the cool crawl space and condense. This could lead to a buildup of mold and wood rot and is conducive to termite activity. During the winter months, the cold air could contribute to heat loss and freezing pipes. With a few minor changes, the traditional vented crawl space can be retrofitted and upgraded to a conditioned crawl space. The main concern is that the area must be kept dry. Prior to upgrading, all possible sources of water penetration into the crawl space must be corrected.

Checkpoint summary

Foundation

○ Note foundation-wall construction type: poured concrete, precast concrete panels, concrete block, brick, stone, and so on.
○ Check for cracked areas of concrete; crumbled and flaking bricks; cracked, loose, missing, and eroding mortar joints.
○ Check for long open cracks that do not line up and have shifted sections.
○ Note long, open, horizontal cracks and signs of bowing in the foundation wall.
○ Are sections of the structure sagging and no longer vertical? (Consult a professional.)

Wood-support framing

○ Inspect all vulnerable wood-support members (sill plates, girders, joists) resting on the foundation wall for rot and insect damage.
○ Note floor joists or girders that sag or have notched sections.
○ Is there bridging or blocking between the floor joists?
○ Inspect wood columns, support joists, and subflooring for cracked sections and evidence of rot.
○ Are there any notches in the flanges of I-joists? Any holes in the webs larger than 1.5 inches in diameter?

Water seepage

○ Is the ground adjacent to the house pitched so that it slopes away from the structure?
○ Are there concrete patios and paths that are improperly pitched (toward the house)?
○ Are there basement windows or stairwells that are vulnerable to flooding?
○ Do downspouts
  —have extensions?
  —discharge against the foundation?
  —terminate in the ground?
○ Is there a sump pump?
○ Is there water in the sump pit?
○ Is the sump pump operational? Does the sump pump discharge into the house drain?
○ Is the water being discharged away from the house or to a dry well?
○ If the structure has been waterproofed, is there a guarantee or warranty available? Did you request a copy?

Basement walls

- ❍ Check for areas of scaling, peeling and flaking paint, damp spots, and signs of efflorescence.
- ❍ Check construction joints, tie-rod holes, and pipe openings for signs of seepage.
- ❍ Inspect wall paneling and base trim for stains, warped sections, and rot.
- ❍ Check underside of basement steps for water marks.
- ❍ Note areas of rust at base of metal columns and sheet-metal furnace casing.
- ❍ Check for dampness, noting musty odors and signs of mildew.

Basement floors

- ❍ Check for extensively cracked and heaved floor sections (usually the result of a high water table).
- ❍ Record all areas of active seepage and puddling.
- ❍ Check joint between foundation wall and floor slab for silt deposits.
- ❍ Check for porous areas and signs of efflorescence on floor and around perimeter.
- ❍ If floor is covered with tiles, are there swollen floor-tile joints?
- ❍ Is there efflorescence between joints?
- ❍ Inspect the house trap pit. Is it dry? Is the cleanout plug secure or loose?

Furnace room

- ❍ Check for exposed wood-frame members (wall studs, ceiling joists) that are in close proximity to the boiler or furnace.
- ❍ Check for large openings between the ceiling and the chimney.
- ❍ Is the room adequately ventilated?
- ❍ Check for asbestos insulation around the furnace and heating pipes.

Crawl space

- ❍ Inspect foundation walls, posts, and wood-support framing for deterioration and signs of water seepage.
- ❍ Check overhead subflooring and support joists for insect damage and/or rot.
- ❍ Check for adequate ventilation.
- ❍ Is this area damp?
- ❍ Is there a dirt floor? Is it covered with a vapor barrier?
- ❍ Is area insulated? Is insulation loose or incorrectly placed?
- ❍ Are there water-supply pipes that are vulnerable to freezing?
- ❍ Are there heat-supply ducts or pipes that should be insulated?
- ❍ Is the area conditioned crawl space?

12

Electrical system

Inlet electrical service 170
Electrical capacity 171
Fuses and circuit breakers 174
Inlet service panel box 175
  Panel-box interior 176
    Subpanel 176
  Aluminum wiring 176
Grounding 177
  Improper ground connections 177
Interior electrical inspection 178
  Electrical outlets 178
  Knob-and-tube wiring 180
  Low-voltage switching systems 180
Violations 181
Checkpoint summary 181

The electrical system of a house can be compared to the nervous system of the human body. Just as every part of the body is supplied by nerves connected to the brain, every part of a house is (or at least should be) supplied by electrical wires connected to the inlet service panel box. These wires are called *branch circuits*. A proper electrical system is essential for a healthy house.

Inlet electrical service

The electrical service is provided through overhead wires or underground cables. If there are overhead wires, they can be seen around the outside of the house. These wires usually run from a utility pole to the house and are fastened to the structure at a point at least 10 feet above ground level. Count the number of wires coming into the house. If there are two wires (FIG. 12-1), the electrical service for the house is inadequate. Two-wire service provides only 110 volts, not 110/220 volts. There should be three wires coming into the house from the electrical service entry. (See FIG. 12-2.) Notice in FIG. 12-2 that the wires loop down before entering the weather-head. This is called a drip loop. It causes water that travels along the wires during a rain to drip down off the wire, thereby preventing it from entering the weather-head and flowing into the service entrance cable. Three-wire service provides 110/220 volts. In some cases, there might be four wires. Four-wire service also provides 110/220 volts but is

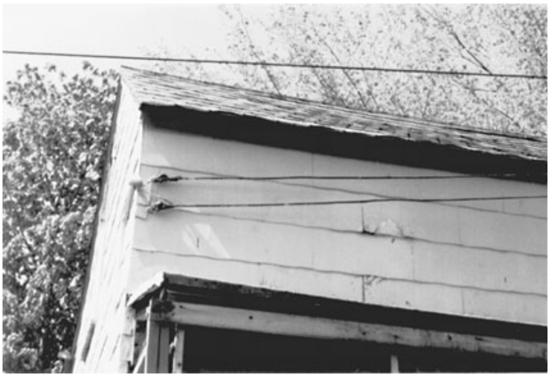

Fig. 12-1. *Two-wire inlet service provides 110 volts.*

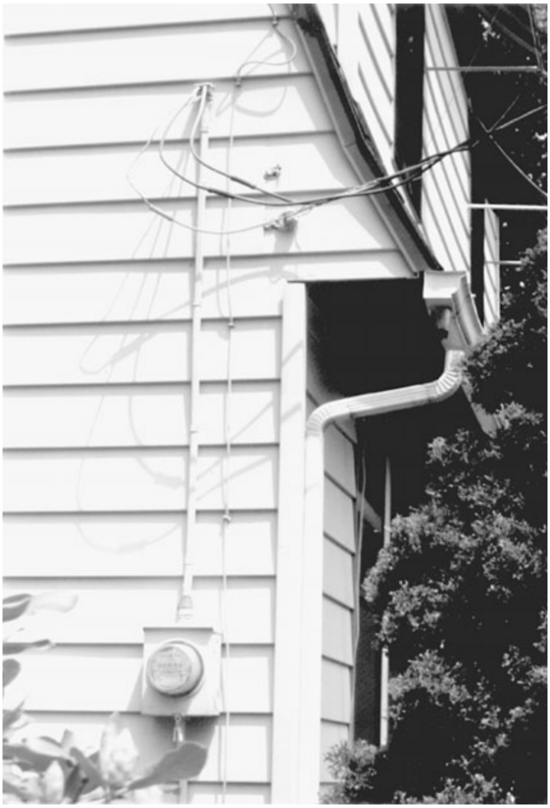

Fig. 12-2. *Three-wire inlet service provides 110/220 volts.*

unusual in a residential structure. That service is three-phase and provides additional capacity, so that you may find it on some very large homes. It is usually found on a structure with heavy electrical demands such as an industrial or commercial building. The inlet service voltage might vary slightly so that in some areas it might be 120/240 rather than 110/220 volts. The difference is of no concern to the homeowner.

Before going into the house, look to see if any tree branches overhang or hit the inlet service wires. They should be pruned. Otherwise, the wire's outer insulation can be worn away, exposing the electrically hot conductors. This could be potentially dangerous if that section of conductors is within reach, as from an upper-level porch, or if it contacts a metal gutter.

While looking at the inlet service wires, check to see if they are securely fastened to the house and whether there are any frayed sections of the outer insulation. In many communities, the inlet service wire (from the attachment to the house to the electrical meter) is the responsibility of the homeowner, not the utility company. Service-entry wires that are badly frayed should be replaced because they can eventually result in a hazardous condition.

The inlet service wires, whether they are overhead or underground, terminate at the electrical meter, which can be mounted on the exterior or interior of the structure. The wires then run from the meter to a panel box, sometimes called the *service switch*. The panel box is basically a distribution center. The branch circuits throughout the house terminate in this box. See FIG. 12-3 for a generalized residential wiring diagram.

Electrical capacity

The unit of electrical power is called a *watt*. Watts are equivalent to volts times amps ($W = V \times A$). The electrical service that I have

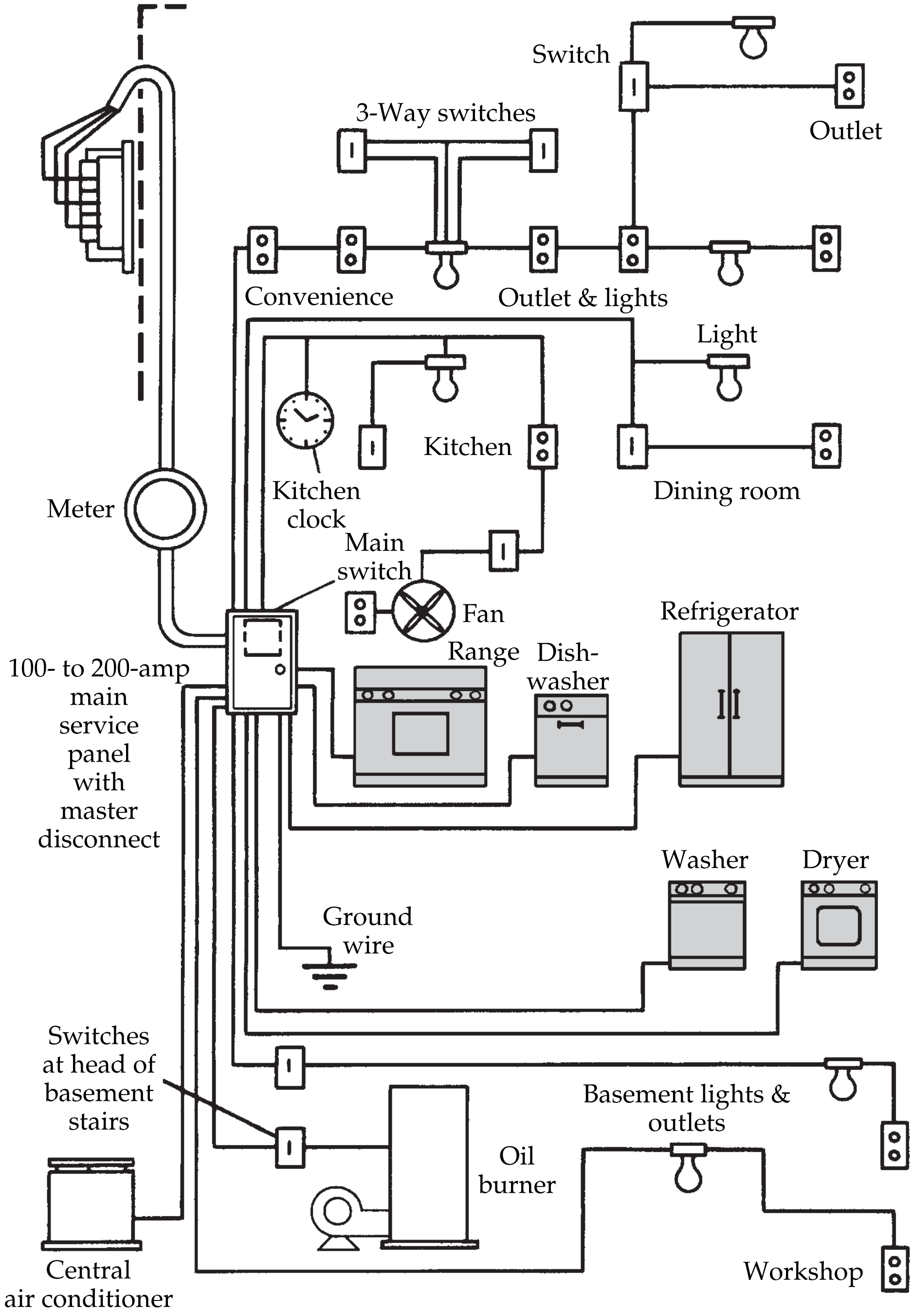
Switch
3-Way switches
Outlet
Convenience
Outlet & lights
Light
Kitchen
Dining room
Kitchen clock
Meter
Main switch
Fan
Refrigerator
Range
Dish-washer
100- to 200-amp main service panel with master disconnect
Washer
Dryer
Ground wire
Switches at head of basement stairs
Basement lights & outlets
Oil burner
Central air conditioner
Workshop

Fig. 12-3. *Generalized wiring diagram for a residential structure.*

found in houses and the corresponding equivalent power capacity is shown in TABLE 12-1.

In older houses, unless the electrical service has been upgraded, you are likely to find 30 and 60 amps at 110 or 110/220 volts. Any service less than 60 amps at 110/220 volts is considered inadequate. Homeowners with such electrical service might indicate that they have lived in the house for over forty years and have found the service totally acceptable. That might be true—for them—because they have adjusted to the lower electrical capacity by not using any major electrical appliances. Nevertheless, if you buy a house with low electrical service, you will probably have to upgrade the service after you move in.

The capacity of the electrical service should be great enough to satisfy the power requirements of the various electrical appliances to be used. Table 12-2 shows some typical

Table 12-1. Power capacity.

Amps	Volts	Watts	Evaluation
30	110	3,300	Inadequate
30	110/220	6,600	Inadequate
60	110	6,600	Inadequate
60	110/220	13,200	Marginal (small house only; no major appliances)
100	110/220	22,000	Minimum
150	110/220	33,000	Good
200	110/220	44,000	Very good

Table 12-2. Power requirements of appliances.

Appliance	Power requirement (watts)
Attic fan	400
Central air conditioning	6,000
Clothes dryer	4,500
Computer	200–400
Copy machine	1,300
Dishwasher	1,500
Fax machine	65
Forced-air furnace (electric heat)	28,000
Freezer	575
Garbage disposal	900
Hand iron	1,000
Instant hot water dispenser	1,000
Lamp (each bulb)	25–150
Microwave oven	500
Range, electric	8,000
Portable room heater	1,600
Printer	23
Room air conditioner	1,100
Sauna	8,000
Steam bath generator	7,500
Television (color)	150–450
Water heater	2,500–4,500

electrical appliances and their associated power requirements.

Considering the many electrical appliances that are available to the homeowner, the minimum electrical service is 100 amps at 110/220 volts. However, if the house is small and the intention is not to use many electrical appliances, 60 amps at 110/220 volts will probably suffice. When a house is equipped with an electric water heater, electric range, electric clothes dryer, and a central air-conditioning system, it should have at least 150-amp service. If electric heat is used in addition to the above appliances, the house should have 200-amp service.

Fuses and circuit breakers

The panel box contains either circuit breakers or fuses to protect individual branch circuits from an overload. Overloaded circuits are one of the chief causes of home electrical fires, and proper protection of branch circuits is essential. FIG. 12-4 shows a circuit-breaker panel box, and FIG. 12-5 shows a fuse panel box. Circuit breakers are more convenient than fuses. Once they have been tripped, they can be reset like a switch, whereas once a fuse has blown, it must be replaced. On the other hand, circuit breakers are somewhat less reliable than fuses. Circuit breakers have been known to "freeze" in the *on* position and should be manually tripped periodically to ensure operational integrity.

Is the panel box a Federal Pacific Electric (FPE) service panel? There have been numerous reports questioning the safety of this panel box and its Stab Lok circuit breakers. Some of the breakers have failed to shut off, or "trip," in response to an overload and arcing has occurred in the panel box. As a safety precaution, replacing the panel box is recommended.

It is important that the capacity of a fuse be matched with the current-carrying capacity of a branch circuit. When a fuse has been blown,

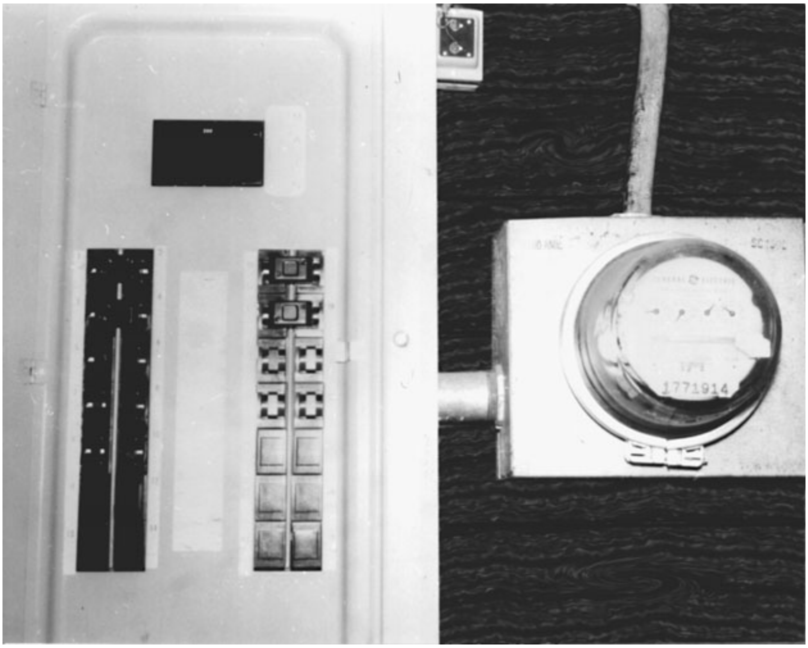

Fig. 12-4. *Circuit-breaker panel box with master disconnect.*

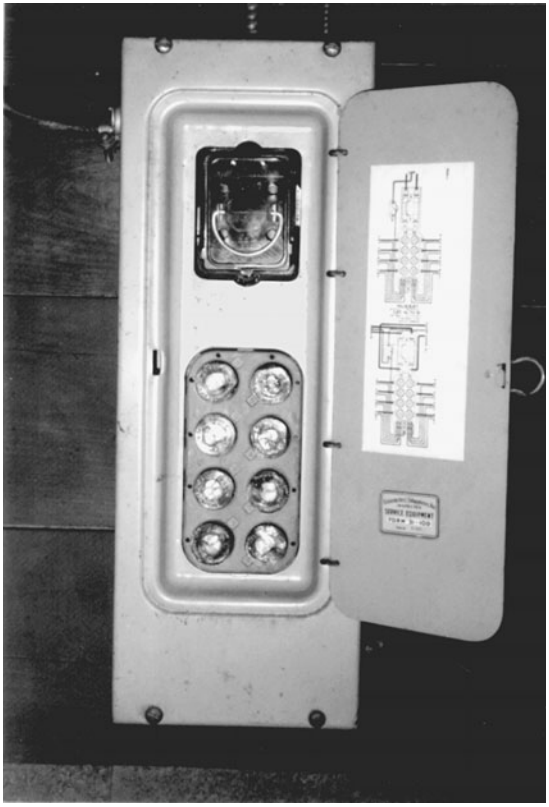

Fig. 12-5. *Fuse panel box with master disconnect.*

it must be replaced by a new fuse with the same current-carrying capacity. Too often, a homeowner replaces a 15-amp fuse with a 20- or 30-amp fuse, not realizing that all three are physically but not electrically interchangeable. In this case, if there is an overload on that branch circuit, the fuse will not blow, but the wires will become excessively hot and possibly cause a fire. This type of problem can be circumvented by replacing the fuse with a Fustat. (See FIG. 12-6.) A Fustat is basically a fuse with an adapter that fits into the fuse holder in the panel box. Once the adapter is inserted, it usually cannot be removed. The adapters are sized so that they accept only fuses of a specific current capacity and are not interchangeable.

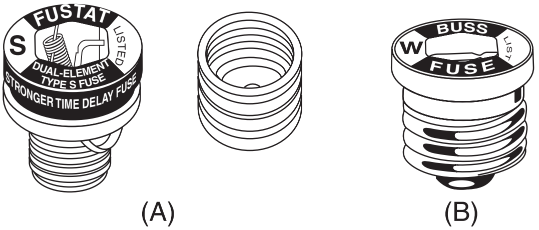

Fig. 12-6. *A—Fustat with adapter; B—plug fuse. Bussman Mfg.*

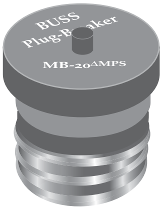

Fig. 12-7. *Minibreaker/circuit breaker fuse.*

There are also fuses that act like circuit breakers. (See FIG. 12-7.) Cooper-Bussmann manufactures 15- and 20-amp mini-circuit-breaker plug fuses that fit into any standard fuse socket. The minibreakers have a little push button that pops out when the circuit is overloaded. To reset the circuit breaker/fuse, simply push the button back in. The minibreakers are also designed for time delay to handle temporary starting loads so they do not trip unnecessarily when motors or appliances start up.

Inlet service panel box

Some panel-box covers have a door. If there is a door on the panel box, lift or swing it open. Warning: Do not remove the panel cover. The cover should be removed only by a professional inspector or electrician because of the danger involved with exposed electrically hot wires. Some homeowners remove the panel cover when doing home wiring and then forget to replace it. Others remove the cover and then cover the panel box with a picture. They think this provides easy access. It does—for children too—and is very dangerous. Panel-box covers should be mounted and securely in place at all times.

When looking at the panel-box cover, note whether any knockout plates or fuses are missing. They represent a potential hazard; it is possible for a child to stick his or her finger in these openings and be electrocuted. If there are openings, they must be permanently blocked off. Check the size and number of the fuses or circuit breakers. There should be at least two 20-amp appliance circuits and one 15-amp lighting circuit for each 500 square feet of floor area. Anything less than this is not considered adequate. Also, if the branch circuits are protected by fuses, note if there are many blown fuses in the area indicating that one or more of the circuits might often be overloaded, a condition that might require installing additional branch circuits.

Panel-box interior

Because of the hazards involved, certain items should not be checked by the home buyer-owner, specifically those items found inside the main panel box. A professional home inspector or licensed electrical contractor checking the interior of the panel box can tell you whether the fuses or circuit breakers are properly sized for their respective branch circuits. Improperly protected branch circuits are found fairly often.

Once the panel cover has been removed, the actual inlet service capacity can be determined. It depends on the size of the inlet service wire only, not on the rated capacity of the panel box or the size of the fuse or circuit breaker used for the master disconnect, if there is one. Some municipalities require that the main electrical disconnect be located on the exterior of the house. Figure 12-4 shows a circuit-breaker panel box with a 200-amp master disconnect. A home buyer seeing this would normally believe that the house is supplied with 200-amp electrical service, and rightly so. In this case, however, the inlet service wire is too small to supply 200 amps. When the cover was removed, it was determined that the inlet service was 100 amps and that the master disconnect circuit breaker did not provide the overload protection.

Subpanel When finishing a basement or constructing an addition on a house, a subpanel is generally used to provide the needed branch circuits, rather than running those circuits from the main panel. During your inspection, you might see a subpanel. As with the main panel, because of the potential hazard, the home buyer-owner should not remove the subpanel cover. A professional home inspector or licensed electrician should check the interior to determine whether or not the grounding wire and the neutral are bonded together as in the main panel. They should definitely **not** be bonded in the subpanel. If you bond the neutral and ground on the same bus or ground bar at the subpanel, you create a parallel neutral current path that could energize all metal parts of the electrical installation. Whereas, by isolating the neutral from ground at the subpanel, any current would then go back to the main panel and go to the service ground.

Aluminum wiring

Inspecting the interior of the panel box also reveals whether aluminum wiring was used for the branch circuits rather than copper wiring. Aluminum wiring is considered a potential fire hazard. Between 1965 and 1973, about 1.5 million homes were wired with aluminum, which at the time was approved by the National Electrical Code. It was later found that dangerous overheating occurred at some of the connections in the 15- and 20-amp branch circuits, which resulted in fires.

If aluminum wiring was used, the electrical connections to the receptacles, switches, and light fixtures throughout the structure should be checked by a competent electrical contractor to determine if they have been properly made or show evidence of problems. If the house does have aluminum branch circuits, you should be aware of the following trouble signs:

- Unusually warm cover plates on switches and outlet receptacles
- A distinctive or strange odor in the vicinity of a receptacle or switch
- Sparks or arcing at switches and outlets
- Periodic flickering of lights (sometimes traceable to faulty appliances or fixtures)

Correcting the problem does not require rewiring the house. Switches and outlet receptacles that are unmarked or marked AL/CU should be replaced with devices that are marked CO/ALR. Or existing switches and outlets can be used, provided short copper pigtails are attached to the ends of the aluminum wires and the devices. It's also necessary to connect light fixtures with copper pigtails. The home buyer-owner should not attempt to correct this condition.

Grounding

The electrical system must be grounded as a safety precaution. This means that a portion of the wiring in the main panel box must be deliberately connected to the ground. This is done by connecting the wiring to a grounding wire that in turn is clamped to a metallic inlet water pipe or to a rod driven into the ground.

Check to see if the electrical ground has been properly connected. There should be a wire coming out of the main panel box that runs to the inlet water pipe. Sometimes the wire is not visible, so go directly to the inlet water pipe. This pipe is located by the water meter (if there is one) and usually protrudes through the foundation wall or lower-level floor slab. The electrical ground wire should be clamped to the water pipe on the street side of the water meter. (See FIG. 12-8.) It might also be connected on the house side, but in such cases there must be a jumper cable running around the meter to the street side.

Improper ground connections

It is surprising how many houses have loose clamps on the water pipes, resulting in an improperly grounded electrical system. Many homeowners do not realize that the wire clamped to the water pipe is a grounding wire for the electrical system, and they loosen the clamp when making a repair or finishing the basement, often forgetting to reclamp the wire because nothing "apparent" happened when it was disconnected.

Look at the clamp. Is it loose? If it is, it must be resecured. This can usually be done with a screwdriver. Sometimes the clamp or screws have corroded and must be replaced. If there is no clamp or ground wire visible, you should question the integrity of the ground system. Sometimes the ground wire is clamped to a water pipe other than the inlet pipe. This procedure provides a false sense of security because a section of the pipe

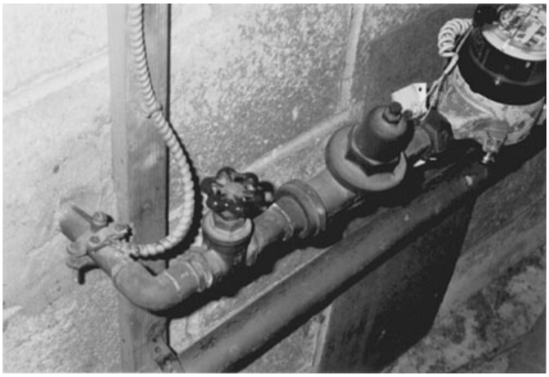

Fig. 12-8. *Ground wire clamped to inlet water pipe.*

can be removed, causing an open circuit in the electrical ground. Some states require a jumper wire between the hot- and cold-water pipes that are connected to the tank-type water heater. This provides electrical bonding between the two pipes and prevents an interrupt in the ground circuit during that period in which the water heater is being replaced. These states may also require a jumper wire around the water meter (if there is one) and also around a plastic water filter in the main water line (if there is one). These requirements can be checked with the local authorities.

The inlet pipe to which the electrical ground is connected must be a functioning water pipe. This ensures that the pipe extends a considerable distance into the ground. According to the electrical code, the pipe should extend at least 10 feet into the ground. If a new inlet water pipe has been installed, the ground wire should be moved to the new pipe. Otherwise, there is no way of knowing how far the old pipe extends into the ground. Often, when replacing an old water pipe, the portion outside the structure is cut off at the foundation wall. This cannot be determined when looking at the pipe from inside the house.

When the inlet water pipe is plastic, as is often the case with a well, the electrical system is grounded by connecting the grounding wire to a rod driven into the ground. If this is the case, check the clamp connection and the rod. If either the clamp or rod is loose, the integrity of the electrical ground should be questioned.

Interior electrical inspection

Now that the basic system has been checked, walk around the interior portion of the house and determine the adequacy of the electrical distribution and possible violations.

Electrical outlets

Check each room for electrical outlets. You might feel that the number of outlets in each room is not very important. This is not so. If there are an insufficient number of outlets, the homeowner will tend to use extension cords. Most extension cords have a lower electrical current capacity than the outlet; and an overload on the extension cord can result in a fire rather than a blown fuse. Figure 12-9 shows one arrangement that is a potential hazard. The convenience outlet in the kitchen is capable of carrying 20 amps, whereas the extension cord can safely carry only 10 amps. Kitchens should ideally have 20-amp appliance outlets that are conveniently located (about every 24 inches) above a working counter.

The number of outlets needed for each room depends on the room's usage and the position of the furniture. Generally speaking, one receptacle outlet per wall for an average-size room (10 feet by 12 feet) is adequate. The outlets in larger rooms should be spaced so that no point along the wall is more than 6 feet (measured horizontally) from an outlet.

In addition to checking whether there are a sufficient number of outlets in each room, you should inspect the outlets to determine if they are functional and whether they are grounded properly. You can check the outlets with a simple plug-in tester available at electrical supply stores. Do the outlet receptacles have two or three slots? Newer outlets have three slots and are intended for use with appliances that have three-prong plugs, although two-prong plugs can also be used. The third slot is a grounding connection for grounding appliances. This is particularly important for appliances that are not double-insulated or have a metal casing. (See the discussion on electrical outlets in chapter 10.) Even though an appliance has a three-prong plug and requires grounding, it can be used with a properly grounded two-slot receptacle

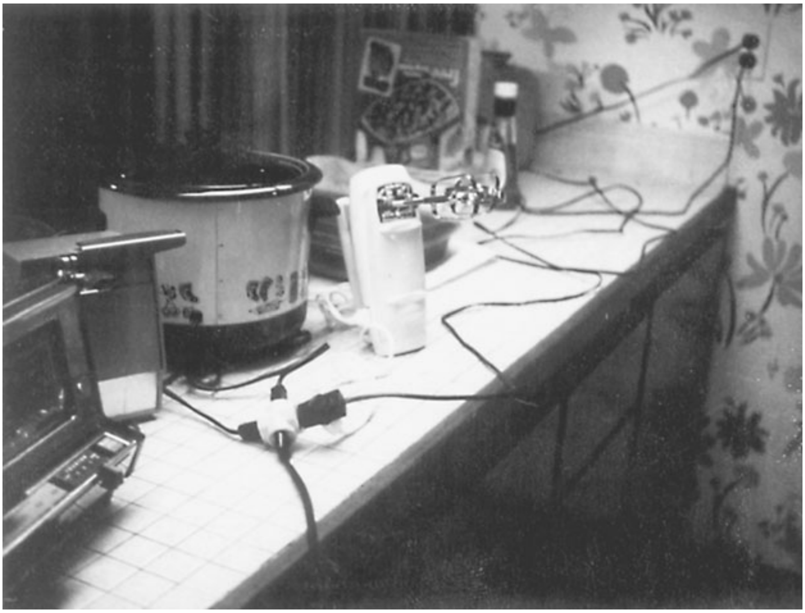

Fig. 12-9. *The home wiring "octopus"—a potential fire hazard.*

with an adapter. There are, however, many electrical appliances that are double-insulated against shock hazards and thus do not require a grounding connection. These appliances can be safely used with outlets that have only two slots.

Bathrooms should have at least one, preferably two, outlets that are readily accessible. Often there is only a single outlet in the wall-mounted light fixture located above the medicine chest, which usually cannot be conveniently reached by anyone less than 6 feet tall. As a safety feature, the outlets and switches in the bathroom must not be reachable from the tub or shower.

According to the National Electrical Code, receptacle outlets in bathrooms, kitchens, garages, crawl spaces, and unfinished basements of new construction must have ground-fault circuit protection. The protection can be achieved by using a special receptacle or a circuit breaker that has been equipped with a ground-fault interrupter (GFI). A GFI is an electronic device that trips (opens) the circuit when it senses a potentially hazardous condition. It is very sensitive and operates very quickly. The GFI interrupts the power in less than $\frac{1}{40}$ second if it senses an imbalance in the electrical current of as little as 0.005 amps. The quick response time in interrupting the power is fast enough to prevent injury to anyone in normal health. Ground-fault protection, however, is not a retroactive requirement and thus will probably not be found in most homes. If there is a GFI circuit breaker or receptacle, it should be tested to see if the fault-sensing function is operational. These units are equipped with manually operated test buttons that trip the circuit when they are operating properly. The GFI should then be reset. Underwriter's Laboratories, Inc., recommends that all GFIs be tested monthly.

In addition to circuit breakers that protect against ground faults, there is a new circuit breaker that protects against electrical arcing conditions, referred to as arc faults. Arc faults are considered one of the major causes of fires in the home. The circuit breaker is an arc fault circuit interrupter (AFCI). It is designed to detect unexpected arcing in electrical wiring and disconnect the power before the arc generates a high temperature that can ignite nearby combustibles such as wood, paper, and carpets. Arc faults can occur in wiring or electrical cords where the insulation has become brittle or cracked, or where it has been punctured by staples or nails. It can also occur at a loose connection where wires are attached to switches and outlets, and where furniture has been pushed up against plugs. As with a GFCI, the unit is equipped with manually operated test button that will trip the circuit when the unit is operating properly. It is recommended that all AFCIs be tested monthly.

When walking through the hallways, rooms, and up and down steps, it should be possible to light the path ahead and to turn off the light without retracing one's steps. This can be done with three-way switches. Also, there should be a convenience outlet in hall areas for nightlights and cleaning equipment.

Knob-and-tube wiring

Knob-and-tube wiring is no longer used in new construction. However, since you might be considering an older house, the building might have knob-and-tube wiring throughout or in portions of the structure. A knob-and-tube wiring system uses porcelain insulating knobs, tubes, and flexible nonmetallic tubing for the protection and support of single-insulated conductors. (See FIG. 12-10.)

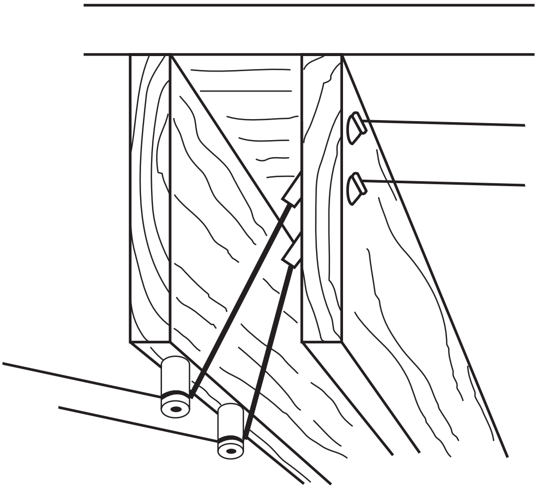

Fig. 12-10. *Knob-and-tube wiring, which is obsolete.*

Obviously you cannot see the type of wiring behind the walls. However, very often there are exposed wires in the unfinished attic or basement. If there are exposed sections of knob-and-tube wiring, the outer insulation covering should be checked for broken and open sections. The insulation is often dry and brittle and chips easily. If there are exposed conductors, the exposed areas must be covered with electrical tape as a safety measure. Knob-and-tube wiring, although obsolete, is considered safe, providing that no modifications are made to that portion of the electrical system. If any changes or extensions are made, rewiring of that branch circuit might be necessary. All modifications to knob-and-tube wiring must be made by a licensed electrician. In addition, the outlet receptacles of a knob-and-tube-wired system are not grounded. Consequently, as a precautionary measure, they should be used only with appliances that do not require grounding.

Low-voltage switching systems

Light fixtures are normally controlled by switches that interrupt the electrical flow in a 110-volt circuit. However, in some houses, the light switches operate at about 24 volts. This type of wiring is generally installed so that the lights can be controlled from three or more locations. With this system, it is possible to control all the lights in the house from a master panel located anywhere in the house. The circuit for a low-voltage system includes a transformer and an electrically operated switch (relay) that is usually mounted near the fixture. If your house has this type of system, make sure that there are replacement relays. Many electrical-supply houses do not stock these relays, and they would have to be specially ordered from the supplier. If you do not have replacement relays and a relay breaks down, you might be without light from that fixture for several weeks before the condition can be repaired.

Violations

Electrical violations should concern the home buyer-owner because they represent potential safety or fire hazards. As you walk around the interior and exterior of the structure, keep in mind those items that are considered common electrical violations. If there is an outside pole lamp, are the wires leading to the lamp buried? They should be. If a portion of the wire is exposed, is it stamped *UF—Sunlight Resistant*? This indicates that it can be used as exterior wire. Many times, homeowners unknowingly use interior-type wire as exterior wire. This is quite dangerous because sun, rain, or soil conditions can cause the insulation covering the wire to deteriorate, exposing the conductors. Incidentally, if there is an outside pole lamp, it should be turned on to see whether it is operational. If it is not, the problem may be as simple to correct as changing the bulb or replacing the switch. However, there might be faulty underground wiring. Do all of the outside electrical outlets have exterior-type covers that provide protection against water penetration? Although not a retroactive requirement, it is recommended that all outside lights and outlets be protected with ground-fault interrupters.

Inside the house, particularly in the basement, look for open junction boxes, loose wires, and exposed wiring and splices (unless the wires are low voltage). Recently, an inspection of a four-story multifamily brownstone walkup in New York City revealed several violations in one circuit located in the basement. The wire coming from the panel box was rated at 15 amps, but it was protected in the panel box by a 30-amp fuse. The wire was hanging in loops from the basement ceiling and had open splices with exposed conductors (splices in a branch circuit should be contained in a closed junction box). One of the hanging loops with exposed conductors was resting on the inlet water pipe leading to the heating-system boiler, and the floor below the pipe was wet. You seldom run into a more dangerous condition. Yet there were ten families living in that house.

If the wiring looks makeshift or appears to have been modified by a nonprofessional, you should request that the seller provide you at contract closing with the Board of Fire Underwriters' certificate of approval for the electrical wiring currently in the house. In your area, the Board of Fire Underwriters might be called by a different name, or a private commercial service might be used.

Finally, when walking through the house, see if all of the receptacle outlets and switches have cover plates. They should. Also, when using an extension cord, the homeowner sometimes runs the cord through the inside of a partition wall. This is a violation of the electrical code. If an outlet is needed, a permanent one should be installed.

Checkpoint summary

Exterior

- ❍ Is electrical service provided by underground cables or overhead wires?
- ❍ Are overhead service wires securely fastened to the house?
- ❍ Count the number of service wires. Note that a two-wire service provides only 110 volts, not 220 volts.

○ Inspect inlet service wires for cracked, missing, and frayed sections of insulation.
○ Note overhanging dead tree limbs or branches in contact with service wires.
○ Inspect for exterior-mounted main panel boxes. (Replacement on the interior is recommended.)
○ Check outside electrical outlets for weather protection.
○ Inspect exterior wiring for proper type. It should be marked *UF—Sunlight Resistant*.
○ Inspect exterior lights and outlets for operation.
○ Note inoperative fixtures or fixtures that are missing, loose, or hanging by wires.

Interior

○ Inspect main panel box. Is it a fuse or circuit-breaker type?
○ Are the GFCIs operational, and are there any AFCIs?
○ Is it a Federal Pacific Electric service panel box?
○ Does system contain a main disconnect?
○ Note the following (do not remove the panel cover):
    — Loose or missing cover.
    — Missing knockout plates or fuses.
    — Are there at least two 20-amp appliance circuits?
    — Is there at least one 15-amp lighting circuit for each 500 square feet of floor area?
    — Are there many spare or burned-out fuses present?
○ If panel-box cover is removed by an electrician or an inspector, he should inform you of
    — the amount of service;
    — circuits that are improperly protected (overfused);
    — circuits that have aluminum wiring;
    — evidence of water leakage or corrosion deposits.

Grounding

○ Inspect electrical system for proper ground protection.
○ If there is a municipal water supply, is the ground wire from the main panel box fastened on the street side of the water meter?
○ Inspect the connection for tightness of fit and corrosion.
○ If the inlet water pipe is plastic (often in a well-pumping system), check on the exterior for a rod or pipe to which the ground wire should be clamped.
○ Note whether the ground wire is missing or has a loose or corroding section.

Interior wiring, outlets/switches, violations

○ Does house contain old or obsolete wiring (i.e., knob-and-tube-type wiring)?
○ Inspect for cracked and open sections of insulation.
○ Inspect wiring in basement and attic areas for
    — loose and hanging sections;
    — extension-cord-type outlets;
    — open junction boxes;
    — exposed splices;
    — makeshift or nonprofessional alterations.
○ Inspect each room for electrical outlets
    — at least one outlet in the bathroom.
    — at least one outlet per wall for an average size room (10 by 12 feet).
○ Are outlets functioning (electrically hot)?
○ Check for loose outlets, switches, and missing cover plates.
○ Are stairways and hallways adequately lit?
○ Are there three-way switches?
○ Are there outlets in the hallways for nightlights and cleaning equipment?
○ Note violations such as
    — open splices;
    — fixtures hanging by wires;
    — extension-cord wiring that passes through partitions or around door openings.

13

Plumbing

Water supply and distribution 184
Fixtures 185
Drainage system 185
Waste-disposal system 186
  Cesspool 187
  Septic system 187
    Alternating drainfields septic system 190
    Mound septic system 191
Exterior inspection 191
  Vent stack 191
  Fresh-air vent 192
  Lawn sprinkler systems 192
  Septic system 193
Interior inspection 193
  Fixtures 193
  Cross-connection 193
  Water pressure, flow 194
  Plumbing wall hatch 194
  Pipes 195
    PEX (plastic) pipe 195
    Polybutylene pipe 196
Basement inspection 197
  Water-supply pipes 197
  Distribution piping 198
  Drainage pipes 199
Well-pumping systems 202
  Wells 202
  Well pumps 203
    Piston pump 203
    Jet pump 203
    Submersible pump 204
  Accessory equipment 205
    Storage tanks 205
    Pressure switch, gauge, and relief valve 206
General considerations 206
Checkpoint summary 207

As with the electrical system, the major portion of the plumbing system is concealed behind the walls and below the floors. Nevertheless, the part of the plumbing system that is accessible for inspection permits you to make a meaningful evaluation of its condition. A basic plumbing system consists of a water-supply source, distribution piping, fixtures, drainage piping, and a waste-disposal system. Figure 13-1 shows the layout of a plumbing system for a typical one-family, two-story house.

Water supply and distribution

Water is supplied to the property line from the street water mains of the local utility company. At the property line, there is a shutoff valve called a *curb valve* that can be used to control the water supply to the

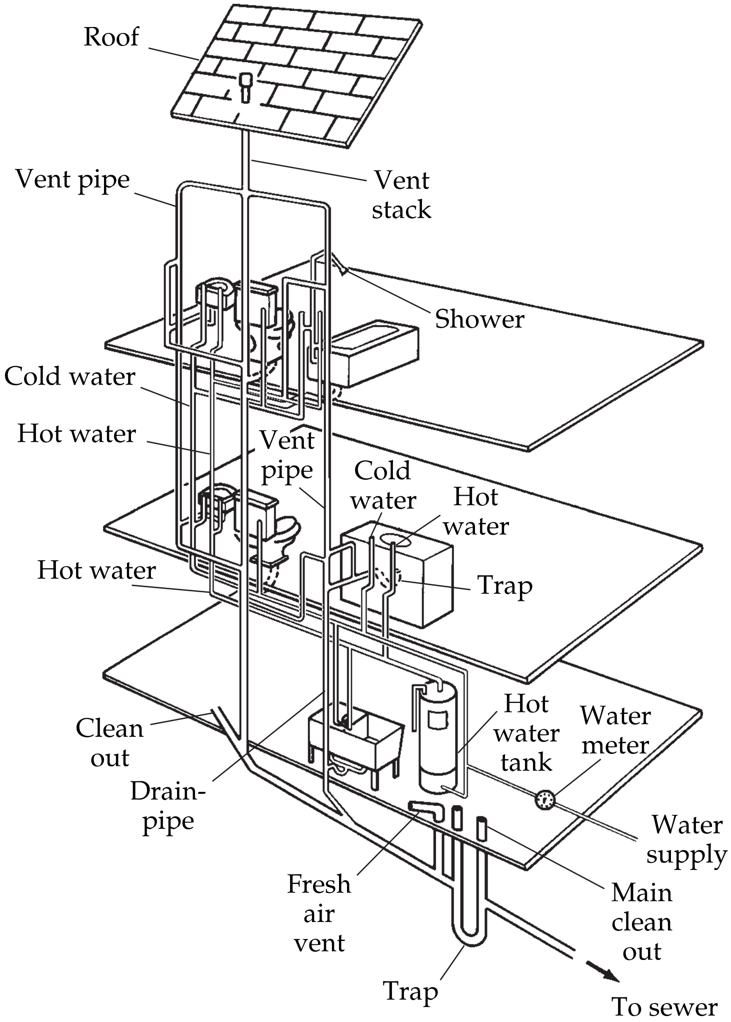
Roof
Vent pipe
Vent stack
Shower
Cold water
Hot water
Vent pipe
Cold water
Hot water
Hot water
Trap
Clean out
Drain-pipe
Hot water tank
Water meter
Water supply
Fresh air vent
Main clean out
Trap
To sewer

Fig. 13-1. *Plumbing system layout, showing water supply, drain and vent pipes for a typical one-family, two-story house.*

house. If you do not pay your water bill on time, the water company can close this valve. The pipe that actually delivers the water to the house, the *house service main*, runs from the curb valve to just inside the house and is the responsibility of the homeowner. Should this pipe need repair or replacement, it is at the homeowner's expense. If water is not supplied by a utility company, it will be supplied by a well-pumping system.

Inside the house, the cold water is distributed to the various fixtures located throughout the house. At the house inlet, there is a shutoff valve that the homeowner can use to close the water supply for the entire house. If there is a water meter in the system, it will usually be located inside the house near the inlet before any branch connections. Sometimes it is located near the curb valve. The cold-water pipe has a branch connection (usually near the inlet) that directs a portion of the water to a heater for generating the domestic hot water. The water heater can be a separate stand-alone unit or part of the heating system and is discussed in chapter 16. Distribution piping for the hot water will run from the heater to the various fixtures. The hot-water pipes are very often near and parallel to the cold-water pipes. However, they should be at least 6 inches apart so that the heat from the hot-water pipes does not affect the temperature of the cold-water system.

Fixtures

Plumbing fixtures are located at the end of the water supply and the beginning of the drainage system. They provide a means by which the water brought into the house can be used. Depending on their purposes, the fixtures have either hot water, cold water, or both. They also have a drain for the removal of the waste water. This waste water is channeled away from the fixtures through drainpipes to a sanitary sewer or a private sewage-disposal system such as a septic tank or cesspool.

Drainage system

The drainage system is more complex than the water distribution system, since it consists of three parts—traps, drainpipes, and vents. The drainage system begins just below the fixture with a water-filled trap. The trap is generally U-shaped and should have water at the bottom portion. The water in the trap forms a seal to prevent sewer gases usually in the drain line from entering the room.

The wastes flow from the fixture trap down the drain line and out to the sewer or private sewage-disposal system. Unlike the water distribution system, where the flow is under pressure, the drainage-system flow is entirely by gravity. Consequently, the drainpipes are larger in diameter than the water pipes, varying from 1½ inches to 4 inches, compared to ⅜ inch to 1 inch for the water pipes. In some communities, the house drain line leading to the sewer must have a house trap. The trap is usually located inside the house near the foundation wall. Its purpose is to provide a seal and prevent the gases that occur in the sanitary sewer from circulating back through the plumbing system. When there is a house trap, there should also be a fresh-air inlet pipe connected to the main drain. This air inlet pipe is located on the house side of the drain approximately 1 foot from the trap. (See FIG. 13-2.) In cold climates, it is located about 5 feet from the house trap to prevent the water seal from freezing during the winter. When the outer end of the fresh-air inlet terminates on the outside of the foundation wall, it should be covered with a perforated metal plate that admits the air and prevents obstruction. When it is freestanding, it should be covered with a cowl or gooseneck. (See FIG. 13-3.) The

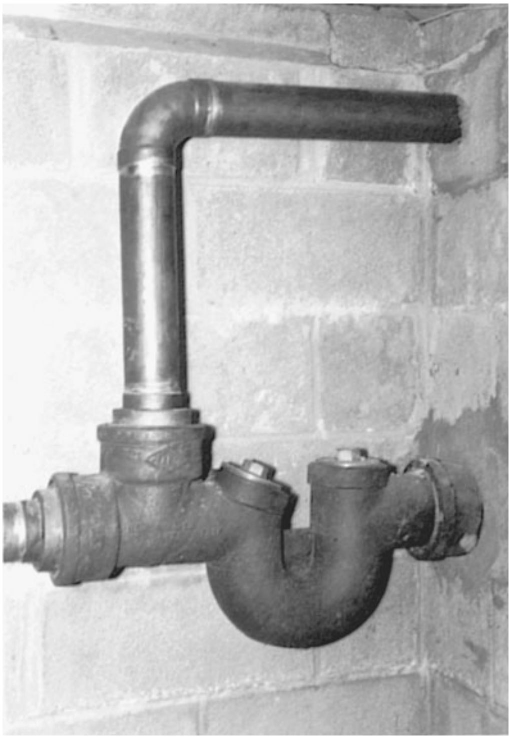

Fig. 13-2. *House trap on main drain line leading to a sewer. Fresh-air inlet pipe on house side of the trap terminates on the outside of the foundation wall.*

function of the fresh-air inlet is to maintain atmospheric pressure at the house trap and to ensure complete air movement within the drainage system. With a private sewage system (septic tank), a house trap on the drain line is not needed. The gases that are generated in septic tanks are usually discharged to the atmosphere through the house drainage-vent system.

Venting is needed in the drainage system, since it provides a means to discharge to the atmosphere gases that develop in the system. It equalizes the air pressure in the drainage system by allowing air to flow into and out of the drainpipes. This free air movement maintains atmospheric pressure at the various fixture traps, which prevents the waste water from siphoning the water seal out of the drain trap. (See discussion on fixture traps in chapter 10, page 140.) Venting in the drainage system is achieved by vent pipes connected to the drain line near each fixture trap and to a pipe that terminates above the roof line. This pipe, called the *vent stack*, is visible from the outside. Vent pipes must be unobstructed. They carry no water or wastes.

Waste-disposal system

Waste disposal from a residential structure will be either through sewers connected to a community waste-treatment plant or through a private disposal system such as a septic tank or cesspool. When all other items are equal, a house with a sewer is more desirable than one with a private disposal system. Sewers are relatively maintenance-free. On occasion, there might be a blockage, which can usually be cleared at low cost by using a drain auger ("plumber's snake"). Maintaining a private disposal system, on the other hand, can be quite costly.

Whether the house is serviced by a sewer or is connected to a septic tank usually cannot be determined during an inspection. The house drain line passing through the foundation wall is the same regardless of whether there is an exterior connection to a sewer or to a septic tank. Since not all municipalities require a house trap on the main drain, the absence of a trap does not mean that there are no sewers. And just because there is a sanitary sewer in the street, it should not be assumed that the house is connected to the sewer line. Tying into the sanitary sewer (if the connection is permitted) is at the homeowner's expense. I know of several instances where homeowners elected to stay with their septic systems rather than go to the expense of tying into the sewer. In some communities, however, once a sewer

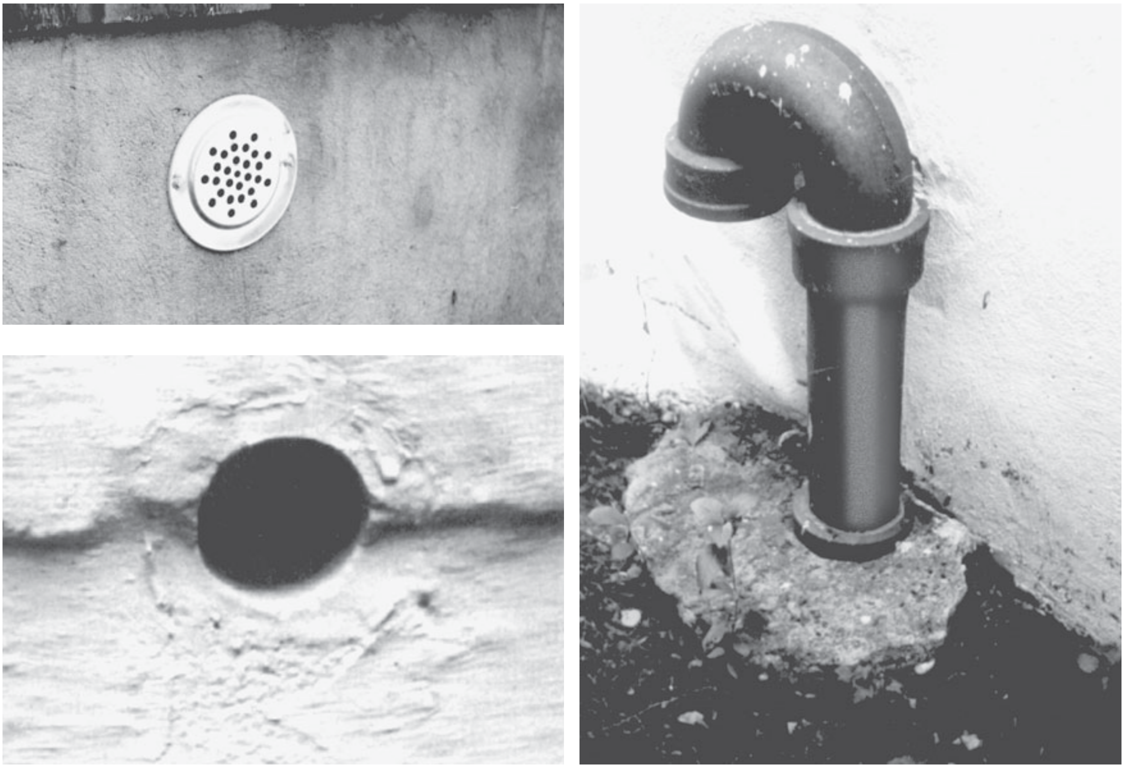

Fig. 13-3. *Terminations of fresh air vent: top left—perforated plate covers opening in foundation wall; bottom left—perforated cover plate for vent opening is missing, a common condition; right—a freestanding "gooseneck" cover.*

line is installed, all the homes on that street are legally obliged to connect. Your best bet is to contact the local municipal building department and ask them if this house is connected to the sanitary sewer system. It might save you a lot of aggravation later on.

Cesspool

A cesspool is basically a hole in the ground that has been lined with stone, brick, or some other material. It is constructed to allow raw contaminated liquid sewage to leach into the soil while retaining the organic matter and solids. Because of environmental and health considerations, most communities no longer allow cesspools in new construction. The older homes that have cesspool disposal systems are not required to upgrade them to septic systems as long as they are functioning properly. However, when problems develop, the homeowner is often legally obligated to replace the cesspool with a septic system rather than repair it. Depending on soil conditions, topography, and available space, the installation of a new septic system may cost over ten thousand dollars.

Septic system

The conventional septic system described below is the most common method of on-site waste water treatment and disposal in the

United States. It consists of a watertight container that functions as a detention tank for sewage sludge and a disposal field for the absorption of the liquid wastes. A septic tank is usually made of concrete but might be made of steel or fiberglass. Raw sewage from the house is discharged into the septic tank through the house drain line. (See FIG. 13-4.) After the sewage settles, the solids are decomposed by bacterial action and are converted into a liquid and a sludge that accumulates at the bottom of the tank. Several types of gases are by-products of the decomposition process, the most common of which is methane, an odorless and highly inflammable gas. The gases generated in the septic tank usually flow back through the house drain and are discharged harmlessly to the atmosphere at the roof-mounted vent stack.

When the level of the liquid (*effluent*) in the septic tank rises to the outlet port, the effluent flows through the outlet pipe to a drainage field. The drainage field, also called the *leaching field*, consists of a series of perforated pipes set into a bed of gravel. As the effluent flows through these pipes, it trickles through the perforations and is absorbed into the ground. The rate at which the ground absorbs the effluent (*percolation rate*) determines the size of the leaching field. When the topography changes abruptly or the area available for a leaching field is too small for adequate absorption, a seepage pit is used. It is basically a covered pit with an open jointed or perforated lining through which the effluent will seep or leach into the surrounding soil.

To determine the size of the leaching area needed for a house, percolation tests are taken in the area of the proposed sewage-disposal system. The rate of water absorption will depend upon, among other things, the type of soil and the level of the water table in that area. If the percolation tests are taken during a drought or when the water table is low (the

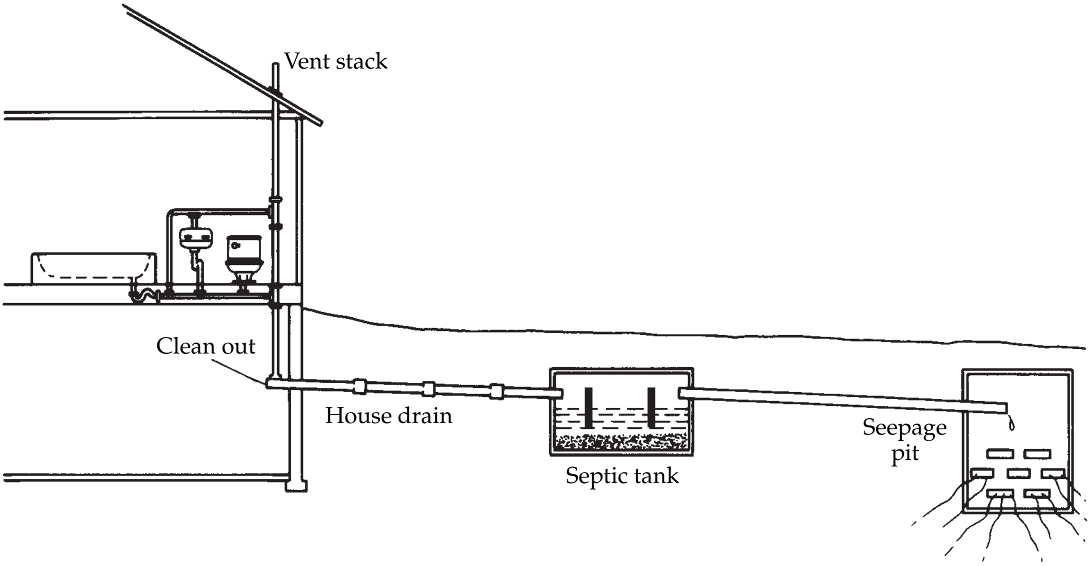

Fig. 13-4. *Typical septic system with seepage pit. Depending on the topography and amount of available land, the pit may be replaced by a leaching field.*

water table rises and drops during the year), the concluding data might result in designing an undersized leaching field that will cause premature failure of the septic system. I know of a village just northwest of New York City where the septic systems in hundreds of homes failed within eight years of installation, some of them within two years. A properly designed and maintained septic system should last between twenty and thirty years; indeed, fifty or more years is not uncommon.

Failure of the septic system will usually show up in the area of the leaching field and not necessarily over the septic tank, although there might be indications in both areas. Look for wet spots or a lush growth of grass. Both can be accompanied by an objectionable odor. When the ground in the seepage field becomes saturated and can no longer absorb the effluent, the liquid will build up and accumulate on the surface. The effluent contains nitrogen and other compounds that are natural fertilizers. When the effluent surfaces, it causes vegetation in the area, particularly grass, to thrive and have a lush green color. However, a healthy-looking patch of green grass over the leaching field is not necessarily an indication of a septic failure. During dry weather, when grass is apt to grow very slowly and turn brownish, the ground over the leaching field (depending on the depth of the field) often contains sufficient moisture to promote the growth of the grass and maintain the green color.

For the leaching field to function properly, there must be sufficient voids in the soil so that the effluent will be absorbed by the ground. Over the years, the voids can be filled with suspended solids, reducing the rate at which the effluent is absorbed into the soil to the point where the effluent surfaces. When this occurs, it is necessary to install a new leaching field. Aside from the cost, a serious problem can arise concerning how to handle the effluent when there is no more room on the property for a new leaching field or seepage pit. If the house has a septic system, find out if there is sufficient room for expansion of the leaching field should it be necessary at a later date.

Premature failure of the septic system can also occur as a result of neglect and abuse by the homeowner. A septic tank should be cleaned or at least inspected for sludge buildup every two to four years. If the tank is not cleaned periodically, the sludge will build up to such a level that the solids are carried out of the tank and into the leaching field. Eventually these solids will clog the voids in the soil or the perforations in the leaching-field pipes, blocking the normal flow of effluent. When this happens, the leaching field requires replacement. This type of problem can be avoided. Unfortunately, many homeowners neglect to inspect and clean the septic tank periodically.

On many occasions I have been told by an owner that the septic tank has never been cleaned. The owner knew from “experience” that if the septic system functioned properly, it was best not to disturb it. After all, his neighbor across the street had had his septic tank cleaned after fifteen years, and two weeks later the system failed and had to be replaced. If he had not touched the system, it would have been all right, wouldn’t it? No! That system apparently had been on the verge of failure before the cleaning. The cleaning had nothing to do with the failure.

Occasionally septic systems can operate satisfactorily for many years without the tank being cleaned. In those cases, the tank might have been grossly oversized for the number of people living in the house, or, perhaps the house was used as a vacation home and occupied only part of the time. One method of extending the life of a septic system is to reduce the volume of water that passes through the tank and leaching field. This is often done by installing a separate drain line for the waste water from fixtures such as the washing machine and connecting it to a dry well or seepage pit that is located away from the leaching field.

Sometimes the homeowner indicates that he has not cleaned the septic tank because he has been using a chemical compound or septic-tank cleaner that he pours down the drain. This cleaner is supposed to improve bacterial activity and eliminate the need for periodic cleaning. The advertising claims for these cleaners are not well founded; there will always be a sludge buildup that must be removed. In addition, some of the cleaners might contain compounds that will actually reduce the bacterial process and can cause the system to fail.

To reduce the possibility of premature failure of the septic system, some homes are equipped with a grease trap in the waste-disposal system. The trap separates grease from the kitchen waste line, thus preventing it from entering the septic system. A buildup of grease in the system can result in clogging or reducing the porosity of the leaching field and can also affect the bacterial action in the septic tank. If there is a grease trap in the waste-disposal system of the house, you should be advised that for maximum effectiveness, it should be frequently cleaned or at least inspected for grease buildup.

Discharging large volumes of water into the septic tank, such as rain runoff from roof gutters or storm drains can over a period of time adversely affect the tank's operation. Large volumes of water can flood the tank, forcing suspended solids into the leaching field where they can eventually block the perforations in the pipes and clog the voids in the soil.

Depending on the soil condition, a water softener can be detrimental to a septic system. If the seepage field consists of a clay-type soil, the waste water from the water-softener regeneration process must not be discharged into the septic tank. The salt brine in the water-softener waste water is not broken down by bacterial action as it passes through the septic tank to the leaching field and can clog the voids in the fine-textured clay soil.

Even though a septic system might be faulty, there might not be any visual indications of a problem when you inspect the house. Whether there is a water (effluent) accumulation over the leaching area or septic tank will depend upon the dryness of the season, the usage of the plumbing system prior to the inspection, and the degree of deterioration of the septic system. One way to check the operation of the system is to "push" it by turning on the water in the tub and letting it run for about one hour. (Before you do this test, ask the owner for permission.) The tub drain must be open so that the water will flow into the septic tank. Assuming an adequate water supply, this pushing should introduce about 275 gallons into the septic system, which very often is enough to cause the effluent in a faulty system to surface and be visible when you reinspect the areas over the tank and leaching field. This test, however, can be misleading if the homeowner had the tank pumped out a few weeks prior to your inspection.

In addition to the conventional septic system described above, there are alternative septic systems available, which are installed when warranted by specific soil conditions, topography, and available space. Alternative septic systems are more complex and costly than conventional systems because of the use of pumps, alarms, and diverter valves. Two such systems are the alternating drainfields septic system and the mound septic system.

Alternating drainfields septic system This system is similar to the conventional septic system's drainfield design. However, as a result of alternating the drainfields, the overall system can be designed so that it takes up less space than a conventional system. Because many municipalities have a requirement that there must be room for a 100-percent drainfield expansion, a smaller drainfield area is a desirable feature, since it will allow construction on small pieces of property, which were considered unusable.

When the drainfield is unsaturated, the treatment of septic effluent is enhanced. As the effluent passes into and through the soil of an unsaturated field, the aerobic bacteria digest the organics in the waste stream and form a filtering layer (a bio-mat) at the soil/gravel edge. On the other hand, saturated conditions and excessive amounts of organics in the effluent reduce the treatment capacity of the system. Excessive amounts of organics escaping from the septic tank can result in a bio-mat growing so thick it will actually prevent adequate water flow into the soil. Consequently, depending on the wastewater flow and the load of suspended solids, some drainfields become clogged and less effective, and in some cases, the fields fail. The alternating drainfields septic system resolves this problem by having two or more drainfields. This system is used primarily to overcome soils with restrictive percolation rates, such as clay. By switching from one field to the other every six months to a year, each drainfield can alternately be used and rested. The resting period causes the bio-mat, which eventually clogs the field, to deteriorate due to lack of nourishment, and allows the drainfield to completely drain, thus renewing the field. Switching from one drainfield to the other is done through a diverter valve, which is housed in a valve box on the effluent line from the septic tank. The valve provides for independent flow to one drainfield or the other, but not both. It must have a watertight lid or cover, which permits unobstructed access for maintenance, inspection, and operation.

Mound septic system Mound systems are used when there are problems with the site for the drainfield, such as a high water table, inadequate percolation rates, and a shallow soil cover over creviced or porous bedrock. This system consists of a septic tank, a dosing chamber or tank with a pump and a high water alarm, and a drainage field constructed in a mound that is raised above the natural grade to maintain a proper distance from the water table or bedrock. (See FIG. 13-5.) The dosing tank is placed between the septic tank and the mound and accumulates septic tank effluent. Float-type control switches inside the tank turn the pump on and off. Once the accumulated effluent in the tank reaches a predetermined volume, the pump delivers a "dose" of effluent to the laterals in the mound by an electrical control system. The mound can be 3 to 5 feet at its highest.

An adequate evaluation of the plumbing system requires an interior and an exterior inspection. Although most of the plumbing inspection is performed in the interior, an exterior inspection can reveal venting problems, violations, and septic-system problems.

Exterior inspection

Vent stack

As you walk around the house looking at the roof, look for a plumbing vent stack. If you do not see one, it indicates either that the plumbing system is not properly vented or that the vent stack terminates in the attic. Both possibilities are violations of the plumbing code. If the construction of the roof is such that there are sections that are not visible from the ground, a vent stack should be looked for during the attic inspection. If there is a vent stack, you will see a pipe coming up from the floor and going out through the roof. Sometimes it is

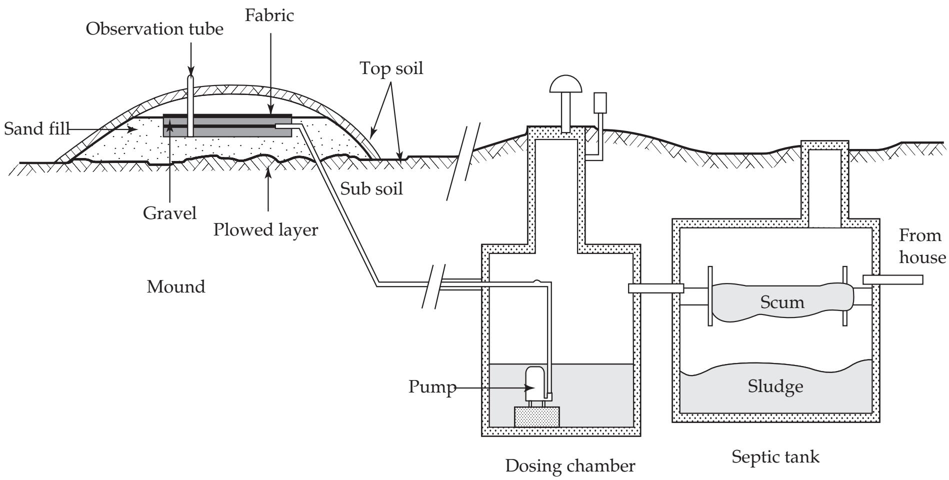

Fig. 13-5. *Schematic of a typical mound septic system.*

difficult to check for the vent stack in the attic because of the restricted space and the lack of adequate lighting.

The vent stack should terminate above the roof. If you see one that terminates near a window or one that runs up along the side of the building (in the northern portion of the country), it is in violation of the National Plumbing Code (as explained in chapter 3).

Fresh-air vent

When there is a fresh-air vent for the house trap, if the outer end terminates on the outside of the foundation wall, the opening should be covered with a screen or perforated metal plate. The covering is needed to prevent children from stuffing toys, balls, or other objects into the opening and blocking air movement. Technically, an unprotected fresh-air vent opening is also a violation of the plumbing code. If you do not see a cover over the fresh-air inlet vent, record the fact on your worksheet as a reminder for later installation. You should also use your flashlight to determine if there is any blockage.

Lawn sprinkler systems

Does the house have an underground lawn sprinkler system? If it does, look specifically for a vacuum breaker (antisiphon device) on the water-supply line for each zone. When the piping arrangement is such that the water is supplied to each zone through a manifold, the vacuum breaker should be on the pipe supplying water to the manifold.

A vacuum breaker is recommended to prevent dirty water or foreign matter that normally accumulates around pop-up spray heads from flowing back into the potable water supply as a result of a negative pressure in the line. It is a relatively simple device and is usually no larger than a few inches in diameter and a few inches high. (See FIG. 13-6.) The vacuum breaker has an atmospheric vent that is sealed when water is flowing to the spray heads. When the water pressure drops to below atmospheric pressure (vacuum) because of a problem in the supply line, a mechanical float drops, opening the atmospheric vent. This allows air to enter the piping system upstream of the spray heads

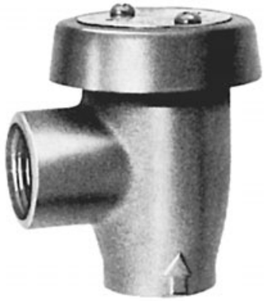

Fig. 13-6. Typical atmospheric vacuum breaker used in lawn irrigation systems. Rain Bird Sprinkle Mfg. Corp.

and thereby prevents dirty water at the spray head from being siphoned back into the supply piping.

Vacuum breakers must be located at least 6 inches above the level of the highest sprinkler head. Look for them near the foundation wall. Sometimes the piping is located in a well adjacent to the foundation, similar to a basement window well. Usually, the vacuum breakers and the electric solenoid valves that control the water flow to the various zones are in the same general area.

Septic system

If the house has a septic system, find out the location of the tank and the leaching area. As you walk around the house performing the general exterior inspection, look specifically for puddles over the leaching area and septic tank. If the puddles have a film on top and there is a foul odor in the area, you can be sure that there is a septic problem. If the puddle is a clear liquid and there is no odor, there might still be a septic problem. The best way to find out whether the liquid is septic effluent or surface water as a result of rain is to collect a sample and have it analyzed. If no puddles are noted during your initial exterior inspection, reinspect the area after "pushing" the septic system as described above.

Interior inspection

Fixtures

Since you will start your general interior inspection at the attic level and work your way down to the basement, the interior plumbing inspection will begin in the first room that has fixtures. (See the "Bathroom" section of chapter 10 for further discussion of plumbing.) The fixtures should be checked for general condition—cracked, chipped, and stained sections—and for operation. Do the faucets function properly, or are there leaks around the handles or spout? Is there an air gap between the spout and the top level of the water when the sink or tub is filled? There should be a gap to prevent back siphonage. Is the sink drain leaking or show signs of past leakage? Occasionally, you might find a pot below the drainage trap to catch dripping water. Sometimes you might find rubber-hose-type connections on the drain or a drain that has been taped up. These are makeshift corrections and require proper attention. Are there individual shutoff valves for the water supply to the various fixtures? Shutoff valves are not necessary but are desirable when making repairs or replacing the faucet. When you open and close the faucet rapidly, do you hear a "water hammer" noise? You shouldn't. But if you do, it can usually be corrected with an antiknock coil or air chamber. Are the toilet bowl flushing and fill valves operating properly? If you hear a whistling noise while the bowl is filling up, the fill valve needs adjustment. After the water closet (tank) has been filled, do you still hear water running? If you do, minor maintenance is needed.

Cross–connection

A cross-connection is where the potable water supply can become contaminated because it is connected to another source of water, which may be nonpotable. Look for situations that could result in a cross-connection. Occasionally, in old homes I find that the faucet spout on the laundry sink extends below the sink's flood rim. This is a contamination problem waiting to happen. If the sink is filled with waste water from a washing machine discharge or some other source, and at the same time a partial vacuum is created in the municipal water supply, the waste water will be drawn back into the potable water, contaminating it. A vacuum can be created because of a number of reasons, such as when a work crew opens a water main to make repairs or when water is drawn from a fire hydrant to fight a fire. If you see this condition, recommend that the faucet be replaced with one where the spout is above the flood rim. The new faucet would create an air gap between the spout and any water in the sink, thereby preventing a cross-connection.

Water pressure, flow

After inspecting the fixtures, faucets, and associated valves, pay particular attention to the water flow and drainage. When there is a considerable flow of water from a faucet, most people say the pressure is good, and when the flow is merely a trickle, they say the pressure is bad. This is a popular misuse of the word *pressure*. It is true that if the pressure is low, the water flow will be low. However, a low-flow condition is not usually caused by low pressure. It is caused by a constriction in the inside of the supply pipes. Depending on the quality of the water, mineral or corrosion deposits can form along the inside of a pipe, reducing the effective pipe opening to that of a straw. In this case, even when the source pressure is good, the flow will be less than minimal.

Check the cold- and the hot-water flow separately by opening two faucets and flushing the bowl at the same time. If the flow appears to be less than adequate, record the fact on your worksheet as a reminder to check further. The low flow might be caused by a kink in the supply pipe, small-diameter distribution piping, or low water pressure at the service entry. However, the probability is greatest that the low flow is caused by a decrease in the inside diameter of one or more sections of pipe supplying the fixtures.

Old galvanized iron piping is particularly bad for water flow. In addition to mineral deposits, there is often a buildup of rust that further constricts the flow. If you find low water flow, especially in an older house, look for galvanized piping. You might find that the house has been partially repiped. Some iron sections might have been replaced with copper or brass. Mixing iron in a copper plumbing system is not desirable, for two reasons. First, just as a chain is as strong as its weakest link, a plumbing system is as good as the weakest section of pipe. Even a small section of iron pipe whose inside diameter has been reduced by a buildup of rust and mineral deposits will lower the flow of water discharging from a faucet. Second, when ferrous (iron) and nonferrous (copper) metals are in contact with each other, a galvanic action takes place between the metals that accelerates the corrosion of the plumbing system at the point of contact.

On occasion, in those homes that have water supplied by a well-pumping system rather than a local utility, you might find that the water discharging from the faucet has a pulsating flow. This condition is caused by a rapidly fluctuating water pressure and is the result of a waterlogged pressure tank. The condition, relatively easy to correct, is discussed on page 205.

Plumbing wall hatch

In many older homes, an opening in the wall of a hallway provides access to the plumbing pipes for the tub. The opening is usually no larger than 2 feet by 3 feet and often has a wooden cover that has been painted like the rest of the wall so that it is not very noticeable. Sometimes this access hatch is in a closet or bedroom. If there is such a hatch, remove the cover. You should be able to see the water-supply pipes, the overflow, and the drainpipes. (See FIG. 13-7.) In newer homes a wall hatch that covers plumbing could also indicate modular construction.

Pipes

Water-distribution pipes can be made of copper, brass, galvanized iron, and more recently plastic (PEX). PEX piping will be discussed separately. Once you become familiar with the types of metal pipe, you will be able to differentiate between them very easily. Here are some pointers. Brass and galvanized pipe have threaded joints; copper joints are soldered. So if the pipe joints are threaded, you will know that they are not copper pipes. Do not try to determine whether the pipe is brass or galvanized by its color. There might be a dirt film on the pipe, or it might be painted. The easiest and most foolproof method is to use your magnet. If the pipe is made of galvanized iron, it will attract the magnet; if it is made of brass, it will not.

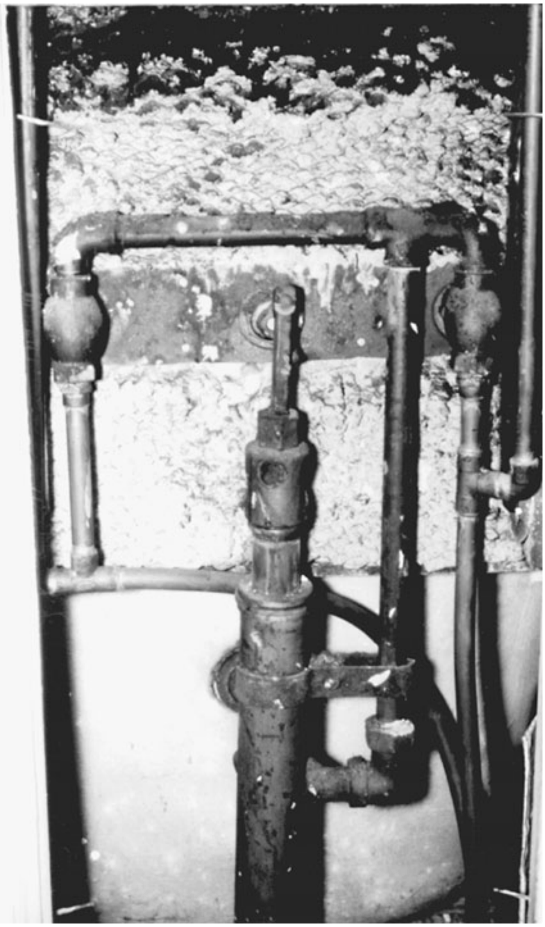

Fig. 13-7. *Wall hatch in hallway provides access to pipes for bathtub.*

Now look to see if the piping is a mixture of copper, brass, and galvanized sections. Even if there is no noticeable drop in water flow, mixed plumbing is not desirable because the resultant galvanic corrosion at the joints will eventually cause leakage. Figure 13-8 shows such a joint. The white encrustations on the pipe are mineral deposits left when the water oozing out of the fitting evaporated. Although deposits on this joint self-sealed past leakage, it is only a temporary correction, for the deposits can come loose at any time. Properly correcting this condition requires replacing the iron section with pipe of copper or brass.

PEX (plastic) pipe PEX piping (tubing) is made from cross-linked high-density polyethylene. Cross-linking is a chemical reaction that permanently links together polymer chains of the polyethylene. PEX was introduced in the United States in the 1980s, but was slow to catch on in the industry. It is now accepted by most municipalities and can be used for all potable water-supply lines, both hot and cold. It has a number of advantages over metal pipe. It is flexible, won't corrode or develop pin holes, has fewer connections and fittings, and is easier and faster to install. Because of its flexibility it is very freeze-break

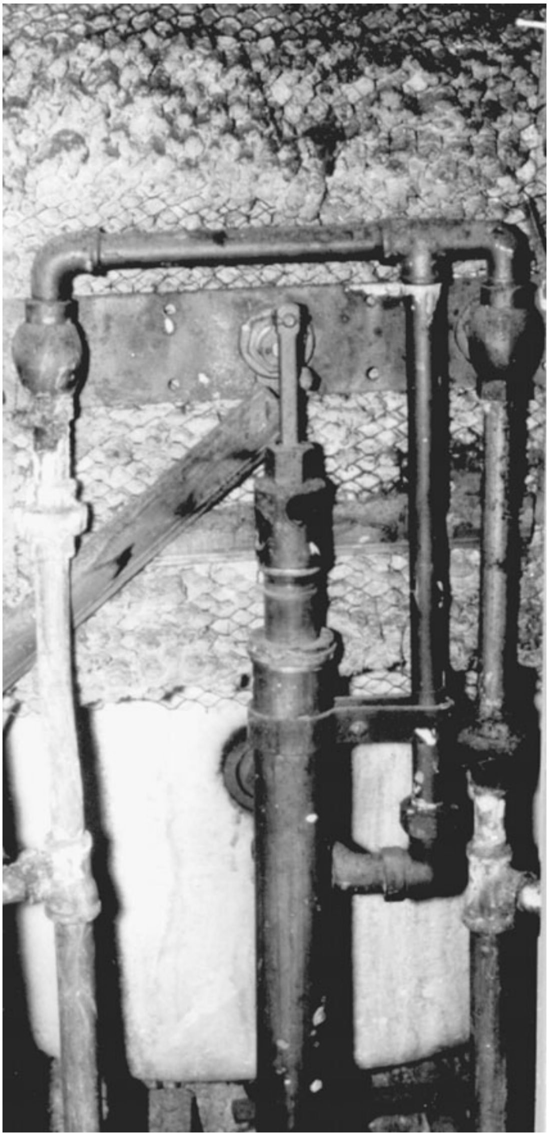

Fig. 13-8. *Mixed plumbing as seen from a hallway wall hatch. Some of the fittings are deteriorating (note the rust and mineral deposits).*

resistant, and a 90-degree corner bend can be made without the need for elbow fittings. Also, PEX is considerably less expensive than copper pipe. PEX comes in three colors: standard white (unpigmented), red, and blue. The red and blue colors are used to help plumbers and homeowners distinguish between hot and cold water-supply lines. Unlike copper pipe, PEX piping does not require soldering at connection joints. There are two popular methods for making connections with PEX: one uses a crimping tool and a crimping ring to secure the tubing over a brass fitting, and the other uses an expansion tool, special expansion fittings, and plastic rings to secure the joints. As you walk around inspecting the pipes, check the exposed joints for signs of past or current leakage. If there is any doubt about the use of PEX tubing for plumbing applications in your house, contact the local authority with jurisdiction over plumbing to verify the acceptance.

Polybutylene pipe There is another plastic pipe that you should be aware of: polybutylene pipe. It's a gray plastic water-supply pipe that was used in the construction of millions of homes throughout the United States between 1978 and 1995. In the 1980s, complaints started to come in that leaks were developing around the fittings and in some cases the pipes were splitting. Some of the leaks were the result of improper installation such as overstressing the pipe at joints and fittings, and overbending the pipe into a tight curve. However, the majority of the leaks were caused by the deterioration of the pipe and/or the plastic fittings. It was determined that polybutylene pipe reacts with oxidants, such as chlorine, normally found in tap water. The reaction, which is internal to the pipe, causes it to scale, flake, and to become brittle, with micro fractures developing. As the pipe ages and reacts with water-soluble oxidants, the condition becomes more severe. Although from a visual inspection, assuming that there are no indications of leakage, the pipe may look to be in good condition. The problem is that since the deterioration is occurring within the pipe, it is very difficult to determine if the pipe is truly in good condition.

The best a home inspector can do if he/she finds polybutylene piping is to alert the buyer about its past history and to anticipate future problems.

Basement inspection

After inspecting all the fixtures, associated pipes, and fittings in the interior rooms, the rest of the plumbing inspection is carried out in the basement. Some homes are built on ground level and do not have a basement. In those cases, this portion of the inspection will be performed as part of the interior room inspection.

Water-supply pipes

Look for the entry of the water-supply pipe. It will be located near the foundation wall and will usually have a meter near the inlet. If you cannot find the water-service entry, ask the homeowner. Sometimes it is concealed behind boxes or storage shelves. The inlet service pipe will generally be made of copper, brass, or galvanized iron. However, in some older homes, you might find that the pipe is made of lead. A lead pipe can be detected by the type of joint between the sections. Lead pipes have wiped lead joints that appear in a horizontal section as a spherical bulge. (See FIG. 13-9.) There is usually a joint near the foundation wall. If you do not see a joint, you can gently scratch the surface of the pipe. If the pipe is made of lead, the surface will be relatively soft, and the scratch will expose an area with a silver-gray color.

Although a lead water pipe might be acceptable in a plumbing sense, it can be a potential health hazard. Depending on the quality of the water, some of the lead might dissolve out. Since the amount of lead that a person can safely absorb is limited and cumulative, by drinking this water over an extended period of time, the maximum tolerance level can be reached. If you find a lead inlet pipe, you should have the water analyzed for lead content. In many communities, the local health department will do the analysis free or for a very nominal charge. If the lead content is high, the pipe should be replaced. Again, remember that replacing the inlet pipe is not the responsibility of the water company; the cost must be borne by the homeowner.

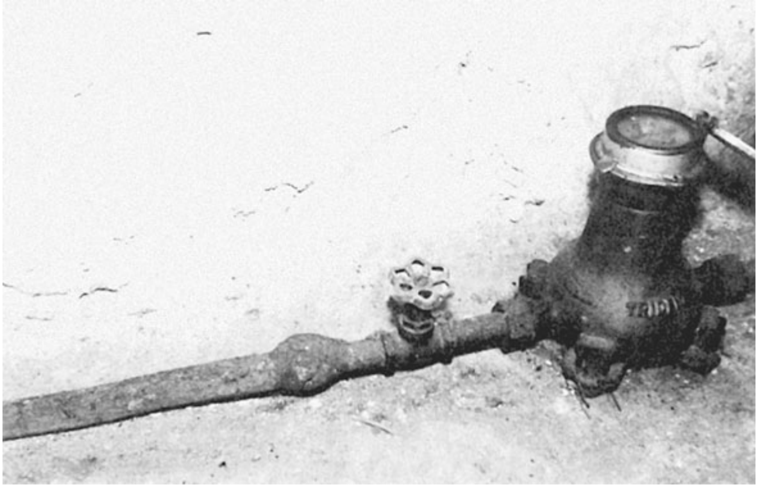

Fig. 13-9. *Lead inlet water pipe. Note wiped joint (bulge) near shutoff valve.*

During my inspections, I periodically find a lead inlet pipe. When I do, I always recommend to the prospective home buyer that as a precautionary measure the water should be analyzed. In one home, located in White Plains, New York, an analysis of the water revealed that it had seven times the allowable concentration of lead. Needless to say, that pipe was replaced.

If the water flow from the plumbing fixtures is low and the piping in the house is good (all copper or brass pipes and fittings with no leaks), the problem will be a constriction in the inlet water supply or low pressure at the street main. On several occasions, I have inspected houses where the old water pipes were completely replaced with new copper pipes and still the water flow was low. Further inspection revealed that even though the house had been repiped, the old galvanized inlet pipe had not been replaced. Under normal soil and water conditions, this pipe should last about forty years. As these pipes age, rust deposits build up on the inside, restricting the flow. Also, a galvanized iron pipe will eventually corrode from the outside and leak. Replacement can cost anywhere from several hundred to several thousand dollars, depending on the length of the line and ground conditions.

When domestic water is supplied by a utility company, the inlet pipe will lead directly to the house distribution piping. In some areas, depending on the quality of the water, there might be a water softener or a filter between the supply and distribution piping. Normally, however, you will not find a storage tank (similar to that needed in a well-pumping system) between the supply and distribution piping. If you do find such a tank, you should suspect a low-pressure condition at the street main. When the water pressure in the street main is low and there are simultaneous demands for water by the houses on the street, the flow to each house may be inadequate. To compensate for this condition, a storage tank is often installed to provide a reservoir to supply water during these periods. Water drawn from the tank is then replaced when the plumbing fixtures in that house are not being used. If you see such a tank between the inlet supply and distribution piping, record the fact on your worksheet.

When domestic water is not supplied by a utility company, it will be supplied by a well-pumping system. Such systems are discussed in the last section of this chapter.

At the house inlet side of the water-supply pipe, there will be a master shutoff valve that can close the water supply to the entire house. Sometimes there are two valves, one on each side of the water meter. See whether the valve is operational. Close and open it. Over the years, because of lack of use, the valve often freezes in the open position.

Distribution piping

Water is supplied to the various fixtures throughout the house by distribution piping. The distribution system begins by the inlet supply pipe just after the water meter and consists of two components: supply mains and fixture risers. The supply mains are usually suspended from the basement ceiling and can be readily inspected in an unfinished basement. The fixture risers run between the supply mains and the fixtures and are usually concealed behind the walls. For the most part, the risers cannot be inspected.

In addition to copper, brass, and galvanized iron pipes, most communities also allow the use of plastic (PEX) pipes for both hot- and cold-water distribution. Trace and inspect all of the exposed water pipes in the basement. You will find that there will be a branch takeoff pipe that leads to the domestic water heater, and if the house is heated by steam or hot water, there will be a cold or hot takeoff leading to the boiler. As a point of interest, over the years, cold-water copper and brass pipes take on a darker color than the hot-water pipes. Sometimes you might find sections of cold piping "sweating" profusely. This is not a problem condition. It is merely condensation and can easily be eliminated by insulating the pipes or reducing the amount of moisture in the air with a dehumidifier.

Faulty plumbing does not necessarily mean that there is a steady stream of water leaking from a pipe or fitting, although if that is the case, immediate correction is necessary. Of particular concern are indications of aging and deterioration. Look specifically for signs of past and current leakage around fittings and valves. Look for mineral deposits, corrosion, and patched sections. The presence of galvanized iron pipes and fittings in a copper and brass plumbing system is a potential problem, as discussed previously. If you see iron pipes, you should make an estimate of the amount and anticipate their replacement. Copper pipes often take on a greenish cast, particularly around the fittings. Although this condition looks as if it could have been caused by water leaking from the joint, it is not. It is usually caused by the soldering flux. A leak, on the other hand, usually shows up as an encrustation of mineral deposits around the joint.

Brass pipes found in residential structures will usually be "red brass" or "yellow brass." You can often tell the difference by the color. Red brass, however, is not really red but a yellowish brown. Yellow-brass water pipes are more vulnerable to corrosion and dezincification (zinc leaching out of the brass into the water) than red-brass pipes. The projected life for yellow-brass pipes is about forty years; seventy-five years for red brass is not uncommon. The weakest part of a brass pipe is the threaded joint. With some older yellow-brass pipes, the threads are paper thin. If force is applied to one of the pipes (with a wrench during a repair or even by leaning on the pipe), the joint could easily rupture. Usually weak joints can be detected by a slight encrustation of mineral deposits. If you find encrusted joints on brass pipes, you should anticipate repair or replacement of those sections.

When the distribution pipes are brass, look along the length of the pipes for signs of leaks. Brass pipes are vulnerable to pinhole leaks along their length. Depending on its chemical quality, the water in the pipes can cause some of the zinc in the brass to dissolve. When this occurs, pinhole openings can be seen along the length of the pipe. Because of the small size of the openings, water drips from the holes very slowly. In many instances, the water will evaporate before it drips, leaving whitish mineral deposits around the opening. (See FIG. 13-10.) Eventually the deposits can self-seal the leak, although the pinhole openings will get larger.

I once inspected a house that had this problem. The owners had moved out in the beginning of the winter before they found a buyer. As a precautionary measure, water was drained from all of the pipes to keep them from freezing over the winter. My inspection took place the following spring when the water was turned on. The sudden surge of pressure in the pipes was enough to loosen all of the deposits, and water started to leak out all along the pipes so that it looked like a sprinkler system. If you see brass pipes with mineral encrustations along their length, even though there might not be any current leakage, those pipes should be replaced.

Drainage pipes

The wastes discharging from toilets and sinks flow from the fixtures down to the sewer or septic tank by means of drainage pipes. As with the distribution piping, only the portion

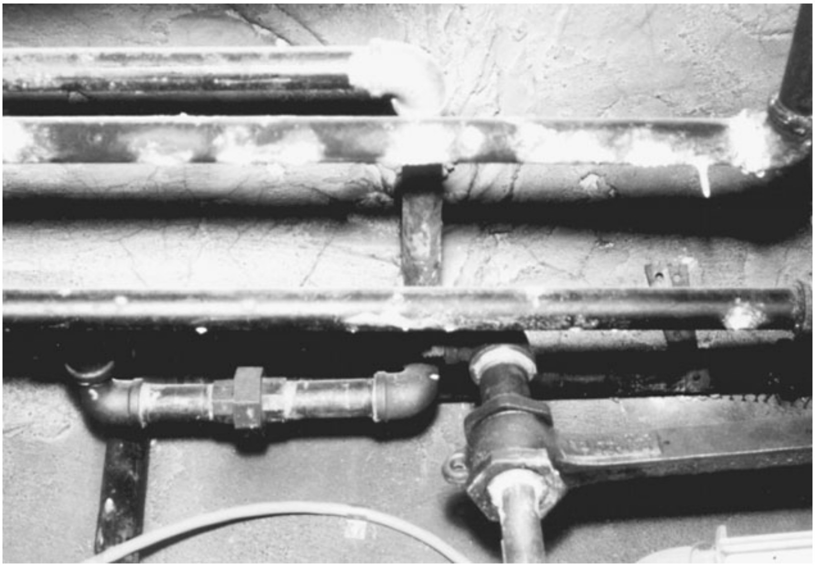

Fig. 13-10. *Brass water pipe with pinhole leaks, a condition caused by leaching zinc. Note mineral deposits on underside of pipe.*

of the drain line that is in the basement will be visible for inspection. In those houses built on ground level or those with finished basements, the drainpipes will probably not be visible and cannot be inspected.

Drainpipes are generally made of cast iron, galvanized iron, copper, lead, or plastic. Very often the drainage system will consist of a combination of the different types of pipes. This is acceptable. The problems encountered in the distribution system when using iron and copper pipes together or lead pipes do not exist in the drainage system.

For most of the day, the drainpipes are empty. The wastes flow down the drain by means of gravity. Consequently, the main house drain in the basement must have a steady downward pitch leading directly to the sewer or septic tank. Look at the drain line. If there is a low point along the length of the pipe, there is a problem. (See FIG. 13-11.) The low point in the pipe will allow grease and sewage

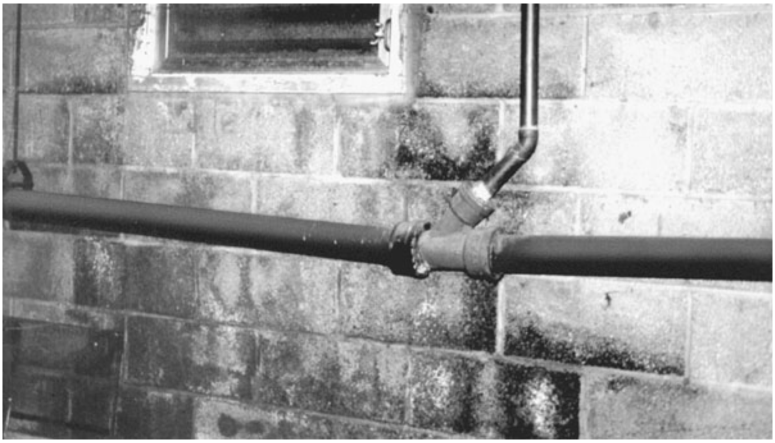

Fig. 13-11. *Low point in drainpipe can cause solid wastes to build up and block the flow.*

solids to settle and eventually block the pipe. This condition is a violation of the plumbing code and must be corrected. When inspecting the drainpipes, look for signs of current and past problems such as cracked and patched sections, improper pitch, and leakage.

One question that I am asked fairly often by prospective buyers is "Can we put a bathroom in the basement?" The only difficulty in installing a bathroom is how to handle the drainage, which must flow by gravity. If the house drain is connected to the sewer at a level above the basement floor, the wastes from the fixtures will have to be pumped up to the house drain so that they can flow out to the sewer. If there is a bathroom in the basement, check to see if the wastewater drainage discharges into a sewer ejector tank. The tank is normally located in an unfinished area of the basement. Since the tank is installed below the floor slab, the only portions visible are the top cover and the associated discharge and vent pipes. (See FIG. 13-12.) The cover should be gasketed and

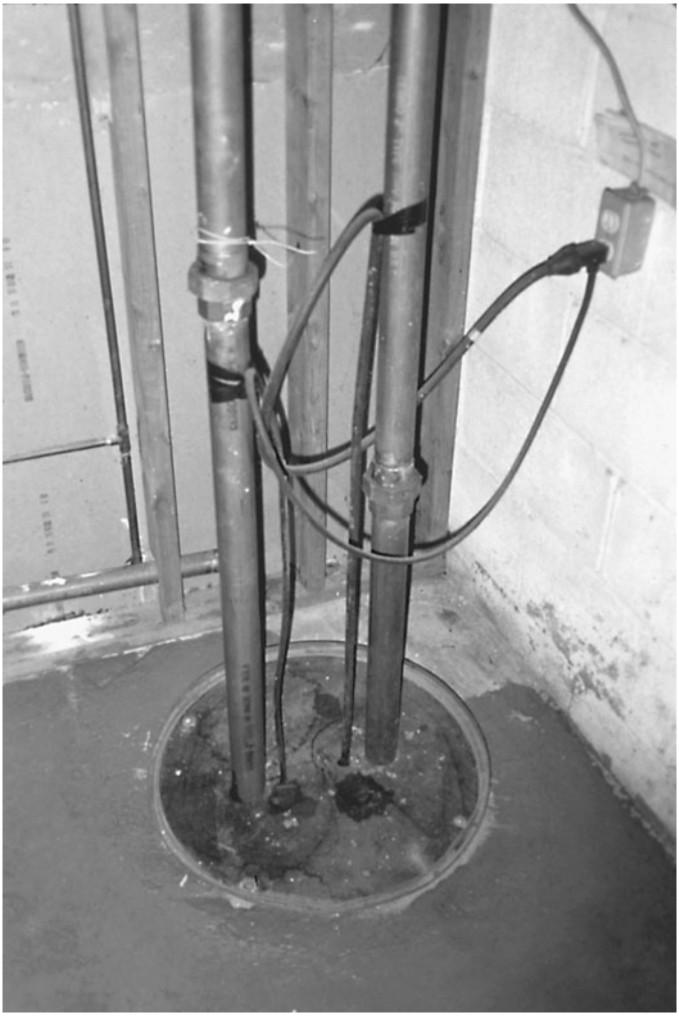

Fig. 13-12. *Sewage ejector pumping system below the floor slab.*

tightly secured to the tank, and there should be adequate seals around the pipe penetrations. There should also be a gate valve and a check valve on the discharge line. A float switch controls the ejector pump. You can check the operation of the pump by flushing the toilet and then letting water run in the sink until the pump is activated. Another approach, although one that is not approved by most building codes, is to use a specially designed toilet that can lift the wastes about 10 feet without the use of a pump. It uses water pressure. This toilet is moderately priced. However, it is quite sensitive and if used for anything other than human wastes and toilet paper, it can become clogged. If only a sink is desired in the basement, a simple, inexpensive lift pump can be used to lift the waste water to the house drain.

When the house drain is located below the basement floor slab, connecting the fixture drains to the house drain will require breaking up sections of the floor. To minimize the installation cost, the proposed bathroom should be located near the existing drain line. In some cases, when the drain line is at or just below the basement floor surface, the toilet is located on a platform to facilitate the connection.

Well-pumping systems

When domestic water is not supplied by a utility company, it will be supplied by a private well-pumping system that includes a well pump, storage tank, and pressure switch.

Wells

As explained in chapter 6, part of the water hitting the surface of the earth as rain, snow, hail, or sleet seeps into the ground and percolates down until it hits an impermeable rock strata through which it cannot penetrate. The water then flows along the strata until it eventually reaches the ocean or a river, which can be more than a thousand miles away. The underground flow is not like a running stream but more like a turtle climbing a rock pile. The water flows through the pores and cracks of rock formations, sometimes surfacing along the way as a river or lake. The water composing this underground flow is *groundwater*, the top surface of which is commonly called the *water table*.

Subsurface rock formations that readily yield the groundwater to wells are *aquifers*. There are two types of wells—shallow and deep wells. A well that draws from an aquifer located less than 25 feet below the earth's surface is a shallow well. When the aquifer is more than 25 feet below the earth's surface, the well is a deep well. Wells over 500 feet deep are not uncommon.

Because of the proximity of the surface, shallow wells are vulnerable to contamination from cesspools, malfunctioning septic systems, barnyard manure, and industrial waste disposal. Deep wells, although less vulnerable to contamination, can also become polluted. Bacterial and chemical pollutants move downward in the soil until they reach the water table and then flow with the groundwater. To a large extent, the soil acts as a natural water purifier for bacterial contamination by filtering small suspended solids and allowing large pollutant particles to settle out. In addition, bacterial pollutants tend to die after a period of time; their life spans are usually short in the unfavorable conditions found in the soil.

Chemical pollution of the water source, however, can persist for years. I recently read a newspaper article about a toxic chemical solvent, trichlorethylene (TCE), that was contaminating the water supply of seventeen private wells. The solvent, which is used for thinning paint or removing grease, can cause neurological problems if inhaled or ingested in high concentrations. The source for the contamination could not be determined. The local health commissioner, however, indicated that he thought it was the result of TCE being dumped in the area many years ago, prior to the homes being built.

Water that has a foul taste or odor and appears dirty may be completely potable, whereas water that is very clear and has a good taste may be polluted. You cannot tell by looking at it or tasting it whether the water is contaminated. As a precautionary measure, well water should be analyzed once a year for both bacterial and chemical pollutants. If the house has a well, have the water analyzed prior to contract closing.

Well pumps

The purpose of the pump is to draw water from the well and push it through the distribution piping with sufficient force that the water overcomes the frictional resistance of the pipes and provides an adequate flow at fixtures.

There are three basic types of well pumps used for residential structures: *submersible*, *jet*, and *piston*. All three can be used for shallow or deep wells. The submersible pump, however, is most frequently used for deep wells. In shallow wells, the jet or piston pumping mechanism is not located in the water. It is located on top of the well. The water is drawn up to the pump by a suction action, not unlike drinking through a straw. A suction will result in a pipe immersed in a body of water when the pressure inside the pipe is reduced below atmospheric pressure (vacuum). Under ideal conditions, the maximum suction lift is 34 feet. However, because of pump inefficiencies and frictional resistance of the pipe walls, the practical limit of suction lift is 25 feet, which is used in defining a shallow well. A deep well, therefore, is one in which water is pumped from a depth that exceeds 25 feet.

Well pumps and their accessory equipment are usually very reliable. Nevertheless, all well-pumping systems require occasional repair or replacement. The projected life expectancy of a pump is seven to ten years, although many pumps run without trouble for twenty to thirty years.

Piston pump These pumps are no longer in general use, although they might be found in older homes. Basically, they are motorized versions of the old hand pump. A motor drives the piston that alternately sucks water into the cylinder and then discharges it on every other stroke. In a shallow well, the pump (motor-piston assembly) is above the ground. In a deep well, the motor is above the ground, and the piston assembly is located in the well. Usually the motor is connected to the piston assembly by a belt and pulley. Inspect the belt for partially torn and frayed sections and adequate tension. Look for signs of leakage around casing joints and the piston rod. There should be none. An overall evaluation of any pumping system must include an inspection of the accessory equipment, which is discussed later in this chapter.

Jet pump The jet pump consists of a jet assembly and a centrifugal pump. The centrifugal pump can be thought of as a small paddle wheel driven by a motor. As the wheel turns, it imparts energy to the water, increasing its velocity and pressure. A portion of the water discharging from the centrifugal pump is diverted to the jet assembly, which has no moving parts. However, it uses this recirculated water to perform two functions. It creates a suction that draws well water into the assembly and pushes this water back up to the centrifugal pump. After passing through the pump, some of the water is again rediverted to the jet assembly and the remainder directed to the plumbing system for distribution.

You can tell whether the pump is a shallow-well or deep-well jet pump by the number of pipes extending into the well. The basic difference between the two pumps is the location of the jet assembly. In a shallow-well jet pump, the jet assembly is built into the centrifugal pump casing and has only one pipe extending into the well. In a deep-well jet pump, the jet assembly must be located within the well (so that the suction lift does not exceed 25 feet). In this case, there are two pipes extending into the well. (See FIG. 13-13.) In areas where the temperature drops below freezing, proper weather protection of jet pumps is important. According to Gould Pumps, Inc., frozen pumps represent one of the most common reasons for pump replacement.

Submersible pump A submersible pump consists of an electrically driven centrifugal pump designed so that both the electric motor and the pump can operate under water. This pump is intended for placement directly in the well and is used primarily for deep wells. However, it can also be used for shallow wells. Water is drawn into the unit through screened openings located between the motor and the pump. A single discharge pipe is connected to the top of the pump and runs to the storage tank, which is usually found in the lower level of the house. When inspecting the pumping system, you will not see the pump, only the accessory equipment. (See FIG. 13-14.)

Because the electric motor is located in the well, in those areas where electrical storms are frequent, it is advisable to have a lightning arrester at the motor power supply. This will conduct high-voltage surges from the line to the ground before they enter and damage the motor. Submersible pumps have the advantage of quiet, dependable operation. They are relatively maintenance-free and are more efficient than jet or piston pumps. However, if a problem develops with the pump or motor, the entire unit must be withdrawn from the well.

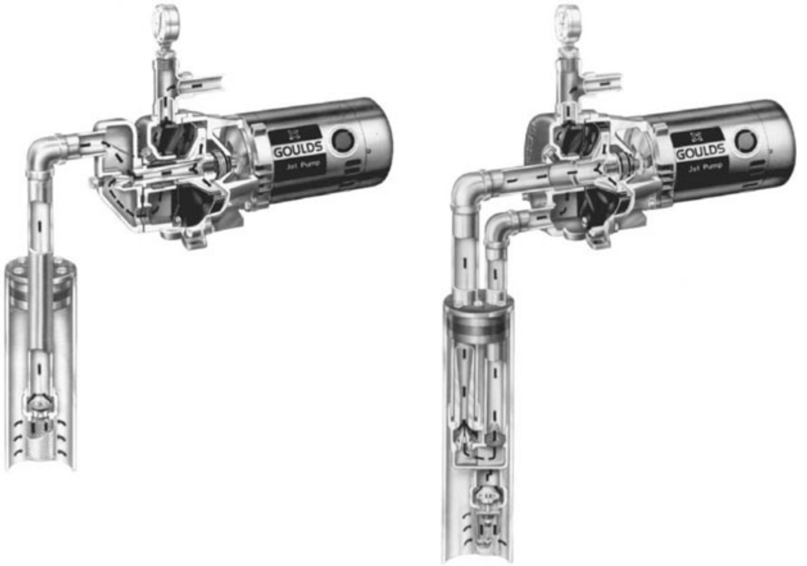

Fig. 13-13. *Jet pumps: left—shallow-well jet pump, one pipe extending into well; right—deep-well jet pump, two pipes extending into well. Gould Pumps, Inc.*

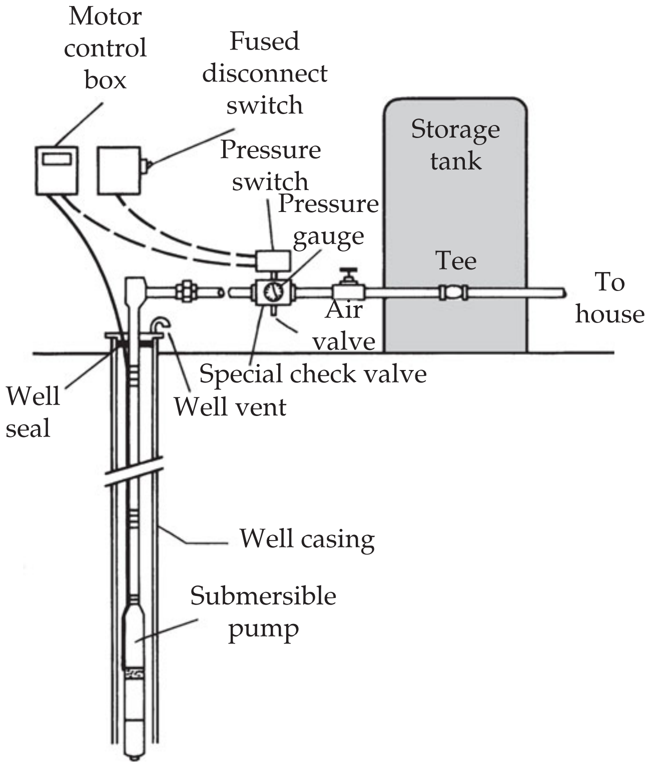
Motor control box
Fused disconnect switch
Storage tank
Pressure switch
Pressure gauge
Tee
To house
Air valve
Special check valve
Well seal
Well vent
Well casing
Submersible pump

Fig. 13-14. *Typical submersible pump installation. Only the controls and storage tank are visible, since the pump is located in the well.*

Accessory equipment

For a private well-pumping system to provide water service comparable to that offered by a utility company, accessory equipment is needed.

Storage tanks The pump discharge line must be connected to a storage tank. The tank, also called a *pressure tank*, is generally located on the lower level of the house but might also be located in an outside pump house. Water from the tank is forced into the house supply pipe whenever there is a demand at one of the plumbing fixtures. A properly functioning tank provides a water reservoir that balances the capacity of the pump against the usage demand. It prevents excessive short-cycling (too rapid starting and stopping), which can cause switch and motor trouble.

The water in the storage tank is under pressure. Since water cannot be compressed, the tank must be partially filled with air to function properly. Over time, the water in the tank absorbs the air, so that the tank eventually becomes completely filled with water. The tank is then "waterlogged," and the pump performs as if no tank were used. Any small request for water, such as filling a glass, will cause the pump to cycle rapidly. This in turn will cause premature wear on the pump, motor, and switch.

The pressure range normally used for well-pumping systems is between 20 to 40 or 30 to 50 pounds per square inch (psi). If a waterlogged condition exists and there is a demand for water, you will hear the pump starting and stopping rapidly if the pump is the jet or piston type. However, when a submersible pump is used, you will not hear the pump. In this case, you can tell that a waterlogged condition exists because the pressure switch will be clicking on and off. Also, the pointer on the pressure gauge will be fluctuating between the high- and low-pressure limits. If you find a waterlogged tank during your inspection, record it on your worksheet. The condition can be easily corrected by draining and then injecting air into the tank.

Some pumping systems have an air-charger apparatus that introduces air into the tank with each cycle to avoid waterlogging. Some prepressurized tanks claim to achieve a permanent separation of air and water by means of a plastic or rubber diaphragm or bag. These tanks should also be checked to determine whether they are waterlogged. I have found that on occasion they are, a condition that indicates a faulty diaphragm or bag.

In most areas of the country, the storage tank should be insulated to prevent condensation during the summer. Some tanks show signs of deterioration because of excessive rusting, a condition brought about over the years by moisture condensation on the tank. Depending on the degree of deterioration, the tank might have to be replaced or scraped, repainted, and insulated.

Pressure switch, gauge, and relief valve All systems must have a pressure switch and pressure gauge. The switch automatically starts and stops the pump at predetermined pressures. The pressure differential between start and stop is usually about 20 psi. The normal pressure range, as mentioned earlier, is 20–40 psi or 30–50 psi, sometimes 40–60 psi. Pressure in excess of 65 psi is abnormal and should be checked out by a pump service company. It might be caused by a faulty pressure switch.

The switch can easily be checked by turning on a faucet. Look at the pressure gauge and watch the pressure drop until the pump is activated by the low limit. Have someone turn off the faucet while you watch the pressure building up on the pressure gauge. When the upper limit is reached, the pump should stop. If it does not or it cuts out at too high a pressure, there is a problem. If the pressure gauge is broken, this test cannot be performed. All too often, I have found inoperative pressure gauges. If you find one, it should be replaced.

Standard tanks are normally rated for a maximum pressure of 75 psi. As a precautionary measure, there should be an automatic pressure relief valve on the storage tank or associated piping. In many jurisdictions it is a code requirement. The relief valve will prevent problems associated with excessive pressure buildup if the pressure switch malfunctions and allows the pump to continue running.

Look at the pressure gauge to see if the system will hold pressure when the pump is not running and no water is being used. If the pointer on the gauge drops, it indicates a leak. If no signs of leaks were noted during the plumbing inspection, the leak is probably between the storage tank and the well. Note this item on your worksheet, for it must be corrected.

General considerations

When the yield of a well is less than 5 gallons per minute (gpm), some municipalities require an auxiliary storage tank from which water can be drawn during periods of peak demand. The rate at which water will be used in a home can vary from 1 gpm (rinsing hands) to a peak rate of 12 gpm or more, depending on personal habits and plumbing fixtures available. The approximate rates at which the various home fixtures use water is shown in TABLE 13-1. This assumes adequate-size distribution piping.

Table 13-1. Water use rates (gpm).

Bathroom sink (lavatory) faucet	3
Water closet (toilet)	4
Bathtub	5
Shower	5
Dishwasher	2
Washing machine (laundry)	5
Garden hose	3
Lawn sprinkler	2

Table 13-2 shows the approximate water-supply requirements of home fixtures.

Obviously, all of the fixtures will not be in operation at the same time. Nevertheless, for a home with two full bathrooms, the pumping system should be designed so that it can supply a peak of about 10 gpm, even if this is greater than the yield of the well. If you find an auxiliary storage tank or the controls for a storage tank (the tank might be buried), you should try to find out the design criteria for the well-pumping system. You might find that the water flow at the design peak demand is less than you require. The design criteria might be known by the seller or might be available through local health department records or the company that installed the system.

Table 13-2. Water supply requirements.

Filling bathroom sink (lavatory)	2 gallons
Filling average bathtub	30 gallons
Each shower	Up to 60 gallons
Older water closet (toilet)	4–5 gallons per flush
Newer toilets	1.6 gallons per flush
Dishwater	3 gallons per load
Washing machine (laundry)	Up to 50 gallons per load

*This figure will, of course, vary with each individual. Also, shower heads are available that will reduce the flow rate.

Checkpoint summary

Exterior inspection

○ Did you note any vent stacks that
  — terminate near windows?
  — run up an exterior side of the house (in northern climates)?
  — have TV antennas and so on strapped to them?
○ Is the drainage system connected to a municipal sewer, a septic tank, or a cesspool?
○ Do you know where the septic system or cesspool is located?
○ If house is connected to a septic tank, has the tank ever been cleaned? When?
○ Did you note any wet spots or any foul odors in the area of the septic system?
○ Are there any areas where liquids are oozing from the ground?
○ Does property contain a lawn sprinkler system?
○ Is sprinkler water-supply line protected by a vacuum breaker?

Interior inspection

Fixtures (operation and condition)

○ Check all plumbing fixtures for operation.
○ Note cracked, chipped, or stained areas.
○ Are sinks or bowls loose?
○ Do faucets leak around handles or spouts?
○ Do sinks, bowls, tubs, and showers drain properly, or are they sluggish?
○ Do sink and tub drains open and close properly?
○ Are there any missing or inoperative "pop-up" units?
○ Does toilet bowl fill and shut off properly?
○ Do any fixture drain lines leak, have make-shift patches or missing traps?
○ Do fixtures have individual shutoff valves on supply lines?

Water pressure, flow

- ❍ Check individual fixtures for low hot- or cold-water flow.
- ❍ Is water flow adequate?
- ❍ Note if there are knocks (water hammer) when faucets are opened and closed rapidly.
- ❍ Note any fixtures with galvanized iron piping or kinked lines (copper).

Piping

Inlet service

- ❍ If water is supplied by a utility company, locate the inlet pipe and the water meter if any.
- ❍ Is the inlet pipe made of iron, brass, copper, or lead?
- ❍ If inlet pipe is lead, take a water sample for analysis.
- ❍ Is there a master shutoff valve? Check its operation.

Distribution piping (supply mains, fixture risers)

- ❍ Are these pipes copper, brass, galvanized iron, plastic, or a combination?
- ❍ Are there signs of leakage, patched or corroding pipe sections or valves?
- ❍ If system is basically brass, note any mineral deposits along the undersides of pipes or around threaded joints.
- ❍ Pipes located in an unheated area such as a crawl space, garage, and so on may be vulnerable to freezing and should be insulated.
- ❍ Are any pipes sweating?
- ❍ Are any pipes improperly supported?
- ❍ Are hot- and cold-water lines adequately spaced apart?

Drainage pipes

- ❍ These pipes are generally made out of cast iron, galvanized iron, copper, lead, or plastic.
- ❍ Look for low points or sagging sections where solid wastes can accumulate.
- ❍ Are any visible drainage lines improperly pitched?
- ❍ Note any signs of leaking, cracked or patched sections.
- ❍ Is there a sewage ejector tank?
- ❍ If there is, is it operational?

Well-pumping systems

- ❍ Have a water sample analyzed for contamination.
- ❍ Is there a deep- or shallow-type well?
- ❍ Is well pump a piston, jet, or submersible type?
- ❍ Do you know the design criteria (gallons/minute) of the system?
- ❍ Are installation records and recorded flow available?

Accessory equipment

- ❍ Is storage tank insulated?
- ❍ Are there signs of rust or corroding areas?
- ❍ Does tank contain a pressure-relief valve?
- ❍ Is pressure gauge operational?
- ❍ When system is active (pumping), note the pressure differential.
- ❍ Does the pressure exceed 65 psi?
- ❍ Does gauge fluctuate rapidly, or does pump cycle on and off?
- ❍ Does system hold its pressure when all faucets are shut and there are no interior plumbing leaks?

14

Heating systems I

Central heating systems 210
Heating outlets: registers and radiators 211
Thermostat and master shutoff 211
Warm-air systems 212
  Advantages 213
  Disadvantages 213
  Gravity warm air 213
  Forced warm air 215
    Controls 215
    Distribution systems 216
    Supply registers and return grille 216
  Heat pump 218
  Geothermal heating/cooling 218
Hot-water systems 219
  Gravity hot water 219
  Forced hot water 221
    Boilers 221
    Advantages and disadvantages 223
    Distribution piping 223
    Radiators 224
    Panel heating 224
    Controls 225
    Low water cutoff 227
Steam heating systems 227
  Distribution piping 228
  Controls 229
  Domestic water heater 231
  Advantages and disadvantages 231
Hybrid heating systems 231
  Steam–hot water 231
  Hydro-air 231

Heating systems fall into two principal categories—central heating and area heaters. In a central heating system, warm air, hot water, or steam is generated in one location of the house and is distributed through ducts or pipes to heat other portions of the house. An area heater is basically a space heater and is used to provide warmth to the room in which it is located, as with a fireplace or potbellied stove. The area heaters of today, however, are much more sophisticated and are equipped with temperature and safety controls.

The principal energy sources used in heating systems are gas, oil, and electricity. To some extent, coal, wood, and solar energy are also used for heating residential structures.

Central heating systems

Most of the homes in the United States have central heating systems. The basic components of these systems are

- A burner for converting gas or oil to heat or a resistance coil for converting electrical energy to heat.
- A heat exchanger for transferring this heat to the air or water. When the heat exchanger is used to produce warm air, it is a *furnace*. When it produces hot water or steam, it is a *boiler*. Because many people have unknowingly been calling a boiler a furnace, the term *furnace* is pretty much the generic name for the heat exchanger.
- A distribution system, consisting of ducts or pipes for conveying the warm air, hot water, or steam to the various parts of the house.
- Heat outlets such as registers (vents) or radiators for transferring heat into the room.
- Automatic safety and temperature controls.

A central heating system provides heat to the rooms throughout the house and in some cases to the nonhabitable areas such as the garage or unfinished basement. In many cases, the system can be extended to provide heat to additions or modifications to the house such as a finished attic or a new dormer. Whether the heating system can be extended depends on its heating capacity and the configuration of the distribution system. If you are thinking about extending the heating system of the house, you should consult a professional to determine its feasibility.

One advantage of central heating is that the distribution system can be designed so that the house is divided into separate heating areas, *zones*. A multizoned house is more economical and more efficient to heat. Zoning is used to maintain the same or different temperatures in various parts of the house. Only those rooms that require it need be heated. Zone control is automatic. Each zone has a thermostat control that opens and closes valves for hot-water and steam systems or dampers for a warm-air system. If the house was originally designed with only one heating zone and you are considering converting it to a multizone system, have the conversion plans checked out by a professional. The configuration of the distribution system might be such that conversion to a multizone system is not economically justifiable. Reaching the break-even point on your investment (cost versus fuel savings) can take many years.

One of the problems of a central heating system is distributing the heat evenly to all parts of the house. The larger the house, the more difficult it is to obtain an even distribution. Often the registers or radiators farthest from the furnace or boiler do not supply as much heat as those that are closest. Minimizing this problem is called "balancing the heating system" and requires the use of dampers for warm-air systems, throttling valves for hot-water systems, and certain types of air valves for steam systems. Balancing the heating system to your family's requirements is best performed after living in the house for a while. It might be that the system is already properly balanced for your needs and requires no further adjustment. Balancing is discussed in this chapter in the sections pertaining to the various systems.

Heating outlets: registers and radiators

The most effective location in a room for forced-warm-air registers and hot-water and steam radiators is along the exterior wall near windows or doors. This enables the heated air to mix with the cold air that very often infiltrates into the interior through the joints around windows and doors. Of course, if the joints are properly caulked and weatherstripped, the air infiltration on a cold, windy day will be minimized. Most homes, however, are not adequately caulked and weatherstripped. (See the section on caulking and weatherstripping in chapter 19.) The mixing of warm air with the cold air around the exterior walls will eliminate cold spots, reduce drafts, and produce a more uniform heat distribution in the room. If the registers or radiators are not located along the exterior walls in any of the rooms, the overall heat distribution in those rooms might be less than what you consider desirable.

All central heating systems have advantages and limitations. If you feel that the type of heating system is crucial to your decision on whether to buy the house, these advantages and limitations are important. Your decision, however, should be based on fact and not hearsay. I have had many clients tell me that they would consider only a house heated with a hot-water system because a warm-air system is "too dry." Well, all heating systems are "too dry" unless the air is intentionally humidified. I have been in homes heated with a hot-water system where the humidity was less than 10 percent, and I have been in homes heated with warm humidified air where the humidity was over 30 percent. Since a hot-water system does not have ducts, the house cannot be humidified from a central location as with a warm-air system. The advantages and limitations of various heating systems are discussed later in this chapter.

Thermostat and master shutoff

There are many types of controls for heating systems. The ones most familiar to homeowners are the thermostat and the master shutoff. The thermostat is used to turn on or shut down the heating system on an as-needed basis automatically. It is a temperature-sensitive switch that normally operates at low voltage (24 volts), although some operate at line voltage (110 volts). As the temperature drops below the thermostat setting, contacts within the thermostat close, activating the heating system. When the temperature in the room containing the thermostat rises above the setting, the contacts open, shutting down the heating system.

In some thermostats, the contacts are exposed to air and dust and should be cleaned periodically. Otherwise, a dust layer can form on the contacts that prevents them from operating properly. Over the years, the contacts in some thermostats become worn because of cleaning and no longer close properly. These thermostats require replacement. In newer thermostats, the contacts are encased in a glass enclosure or have been replaced by a sealed mercury switch.

In a house with a multizoned heating system, each zone will be controlled by a thermostat. The placement of the thermostat is quite important. Since the thermostat senses only the temperature in the surrounding area, the placement of the thermostat must be such that the temperature in that area is representative of the temperature for the entire house or for that zone. The thermostat must never be placed in a draft or in an area where air circulation is blocked.

Research has indicated that by lowering the thermostat setting 5 to 10 degrees prior to going to sleep and resetting it in the morning, you can save 5 to 15 percent of your fuel bill, depending on your geographical location. If you want to take advantage of these savings but feel that you might forget to lower the thermostat setting each night, you can replace the regular thermostat with a clock thermostat. A clock thermostat will automatically lower the temperature setting each night and raise it each morning. From a convenience point of view, a clock thermostat is very worthwhile. If the house does not have a clock thermostat, you should consider its installation. Some units allow for a double setback, which is useful in houses that are empty during the day.

Every heating system should have at least one master shutoff switch. Usually, the switch is located near the furnace or boiler. Sometimes the switch is located at the top of the stairs that lead to the basement. In the event of a problem with the heating system, this switch can be used as an emergency shutoff for the burner. The switch is also used by a repairman when servicing the system. By turning the master switch off, no one can inadvertently turn the heating system on by raising the thermostat.

Warm-air systems

Air heated in a furnace travels via supply ducts to the rooms. The warm air enters the rooms and is discharged through wall or floor registers or ceiling diffusers. The cooler air in the room, being displaced by the heated air, travels through return ducts back to the furnace where it is reheated and recirculated. If the system does not have cold-air return ducts, the cold air travels back to the furnace via gravity. Usually the stairway allows the cold air to travel to the first floor, and grilles in the floor of the first level then allow it to recirculate to the furnace. Of course, doors must be open to permit the air movement. Where doors are closed, the heat rises in a very restricted manner.

The air in a warm-air system can be supplied to the furnace by recirculating heated air through ducts or by drawing in cool air from the basement. The latter method is no longer used in modern warm-air systems because of the potential danger of exposure to poisonous gas and because of the heating inefficiency. If the chimney is clogged, the exhaust gas, which contains poisonous carbon monoxide, backs up into the basement. It is then drawn into the furnace and distributed throughout the house. Also, since the temperature of the basement air is lower than the temperature of the recirculated air, more fuel will be required to heat it to the desired temperature. Although this method of supplying air to the system is for the most part no longer used, it will still be found in many older systems. You can easily spot it. Look for a large opening in the furnace casing.

There is only one condition under which a furnace or its heat exchanger must be replaced: when the walls in the heat exchanger that separate the circulating air from the hot exhaust gases deteriorate because of age, premature corrosion, or cracks and thus allow the exhaust gases to mix with the circulating air. Included in the exhaust gases is carbon monoxide which is poisonous. The mixture of air and exhaust gases circulating around the house is quite dangerous. The "life expectancy" of a furnace refers to the average number of years of usage that can be expected before the walls of the heat exchanger deteriorate. For many modern furnaces, the projected life is between fifteen and twenty years, although some older ones have been safely operational for well over thirty years. A number of manufacturers are now providing replacement heat exchangers so that the overall life expectancy of the furnace can be extended indefinitely.

Advantages

Warm-air systems have an advantage over other types of heating systems in that the air in the house can be cleaned (dust particles removed by filtering) and humidified. Most systems use either inexpensive disposable filters or permanent-type filters that require periodic washing. Some systems utilize an electronic filter, which is very effective in removing dust and pollen from the air.

Not all warm-air furnaces are equipped with a humidifier for adding moisture to the circulating air. If you do not see one, you should consider installing one. The humidifier may be mounted in the main return or supply duct and is usually located near the furnace. Humidifiers such as the evaporative-plate or wick type add some moisture to the circulating air. However, they are not totally effective. A more positive introduction of moisture into the airstream can be achieved with a power spray humidifier that is controlled by a humidity-sensing device.

Additional advantages of a warm-air system:

- Adaptability to a central air-conditioning system. (See chapter 17.) The distribution ducts and the furnace blower can be used to circulate the cool air. This results in considerable cost savings when installing a central cooling system.
- There are no distribution pipes to freeze and burst. Consequently, if the heating system is not operational for several days during the winter, as would be the case in the event of an extended power failure, there would be no need to worry about the distribution system freezing (a condition that can occur with a hot-water heating system). Of course, regardless of the type of heating system, during an extended power failure, the domestic water pipes are vulnerable to freezing.
- The replacement cost for a new warm-air furnace is less than the replacement cost for a new hot-water or steam boiler.

Disadvantages

The major disadvantage of a warm-air system is that in the event of a faulty heat exchanger, the exhaust gases will mix with the circulating air and be distributed around the house.

In a multizoned warm-air heating system, the zones are not totally independent of one another. The zones are controlled by motorized dampers located in the ducts. When the dampers are closed, they block the airflow in their respective ducts, preventing that portion of the house from being heated. However, some air will always flow around the closed damper, decreasing the overall efficiency of operation. In quality-constructed homes, zone control is often obtained by using two separate furnaces with separate distribution systems rather than by using motorized dampers.

Gravity warm air

A gravity warm-air system is a very simple system that is often found in older homes. Generally, the furnace is quite large and often looks like a mechanical “octopus” with many ducts sticking out of the upper portion. (See FIG. 14-1.) There are no moving parts, motors, or electrical connections other than those required for the thermostat and burner control. The basic principle of operation is the fact that warm air rises and as it does

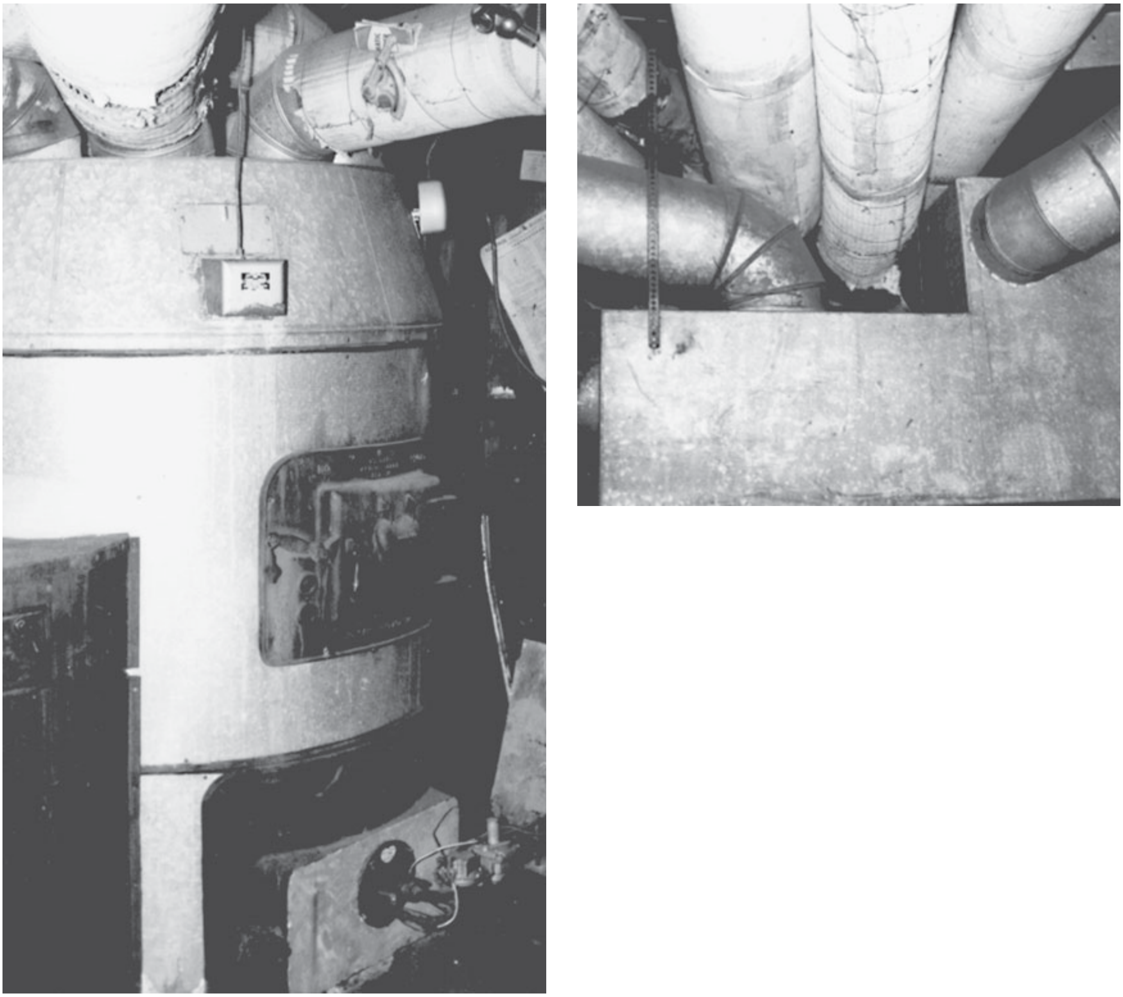

Fig. 14-1. Left—Old "octopus" warm-air furnace; system was converted from gravity to forced warm air by installing a blower unit, which is located in the black casing on left side; right—Heat-supply ducts at top portion of "octopus" furnace.

displaces the cool air, which in turn results in air circulation.

The heat distribution in a house heated by a gravity system is often less than desirable. To obtain a uniform heat distribution, there must be good air circulation. And in order to have good circulation, there must be a considerable temperature difference between the warm air entering the room and the cooler air in the room. As the room begins to warm, the temperature difference decreases, thereby reducing the air circulation.

The cooler air returning to the furnace is drawn into the return duct by a natural draft that is the result of thermal air currents. The moving air does not have much force. Consequently, there will usually not be a filter in this system for cleaning the air. The resistance offered by a filter (especially a dirty one) is often enough to block the airflow.

Although a gravity system is considered obsolete, it can be updated to a forced warm-air system by installing a blower unit (fan and motor assembly). For the most part, the old gravity systems are quite rugged and can last for many years. I have seen many units that are over fifty years old. They are usually made of heavy-gauge cast iron or steel and can probably be used for many more years, assuming that efficiency of operation is not a consideration. As mentioned earlier, the only time a furnace must be replaced is when there is a crack in the heat exchanger that allows the exhaust gases to mix with the circulating air.

Forced warm air

Because it is economical to install and is versatile, the forced warm-air heating system is found in more homes than any other central heating system. The basic difference between this system and the gravity warm-air system is a blower in the heat exchanger that circulates the warm air. Since the warm air is distributed under a draft, comfort heating can be achieved at lower furnace temperatures with lower fuel consumption. In addition, the supply and return ducts need not be as large as those of a gravity system.

Controls There are three basic controls for this system: a thermostat, a fan control, and a high-temperature limit control. The thermostat has been discussed earlier in this chapter. The purpose of the fan control is to prevent the fan from circulating cool air around the house. The fan control is a temperature-sensitive switch that turns the blower on and off at preset air temperatures. It is independent of the thermostat. When the thermostat calls for heat, only the burners should fire. The fan should not begin to operate. If it does, the fan control is either faulty or in need of adjustment. After the heat exchanger warms up to a temperature of about 110° F to 120° F, the fan should begin to operate. When the temperature setting on the thermostat is satisfied, the thermostat will shut off the burner but not the fan. The fan will continue to operate until the temperature in the heat exchanger drops to about 85° F.

If the heat exchanger gets too hot, the high-temperature limit control will shut off the burner. The limit control is usually set at about 175° F. For proper operation, the fan should begin to operate before the burners are shut off by the limit control. Otherwise, the temperature of the air discharging from the registers will be too high for comfort heating.

Condensing furnaces Because of the need to conserve energy and the escalating cost of fuel after the energy crunch of the 1970s, the heating industry developed a new generation of gas-fired forced-warm-air furnaces. These furnaces have been able to achieve an overall operating efficiency of 90 to 97 percent, whereas many of the conventional furnaces found in most homes today have an overall operating efficiency of around 60 to 65 percent. Because the U.S. Department of Energy set a standard in 1992 that requires furnaces to have an overall efficiency in the 80 percent range, conventional furnaces are no longer installed in new homes or as replacement units.

The new furnaces are called *condensing furnaces*. The increase in efficiency is the result of the installation of a secondary heat exchanger that extracts heat from the exhaust gases that normally flow up the chimney with conventional furnaces. During this process, the temperature of the exhaust gases drops to a point where the water vapor in the exhaust gases condenses, thereby releasing additional heat. With the conventional furnace the temperature of the flue gases is about 450–550° F. With the high-efficiency units, it's about 120–130° F.

To achieve the increased efficiency, it was necessary to incorporate additional components into the overall furnace package. Included with the furnace are a power vent fan (also called an *induced draft blower*), plastic piping to vent the flue gases through a side wall or the roof, condensate drainage piping, and an intake air duct for those units that have a sealed combustion system using outside air.

A power vent fan is needed to overcome the additional resistance to the flow of the exhaust gases caused by the secondary heat exchanger. Also, with the low temperature of the exhaust gases, it is not necessary to use the conventional chimney.

Pulse-combustion furnace Another type of high-efficiency furnace is the pulse-combustion furnace. As with the condensing furnaces, the high efficiency of a pulse unit results from the extraction of heat from the exhaust gases until the associated water vapor condenses. However, the combustion process is completely different. Whereas the heat in a condensing furnace results from a continuous burning of fuel, in a pulse combustion furnace, it results from sixty to seventy tiny explosions of a gas-air mixture per second in the combustion chamber.

Distribution systems There are two basic configurations used for the distribution of a forced-warm-air heating system—the *extended-plenum* and the *radial* configuration. (See FIG. 14-2.) Regardless of the configuration, there should always be a physical separation from the furnace at the beginning of the duct. The separate sections are usually connected by a heavy canvas fabric. (See FIG. 14-3.) The separation is intended to isolate the distribution system from the furnace so that blower noise will not be transmitted throughout the house. The isolation fabric in many older homes contains asbestos. If you suspect the fabric contains asbestos, recommend that it be tested and replaced if necessary. (See chapter 20, Environmental concerns.)

In the extended-plenum configuration, a large rectangular supply duct extends in a straight line from the plenum mounted on the furnace. From this main duct, branch ducts supply warm air to the various rooms. The large supply duct results in less resistance to airflow and produces a more effective heat distribution for those rooms farthest from the furnace.

In the radial configuration, there is no main supply duct. Each branch duct takes off directly from the plenum and runs to the individual room registers. This system is usually found in smaller houses. A variation on the radial configuration is the perimeter-loop-duct arrangement, which is intended for houses built on a slab. This configuration uses a duct that encircles the perimeter of the floor slab and is connected to the furnace by feeder ducts. (See FIG. 14-4.)

The branch ducts for both the extended-plenum and the radial configuration will be either rectangular or round. The round ducts are usually relatively small in diameter (4 to 6 inches), have a higher resistance to airflow, and are normally not used for air-conditioning. (Do not confuse these ducts with the insulated flexible ducts used in many central air-conditioning systems.)

The return duct for both configurations is usually made of sheet metal and has a rectangular cross section. However, in some homes with basements, you might find that a portion of the return duct has been formed using a section of the overhead wood framing. This is done by covering the channel that is formed by adjacent joists with sheet metal. (See FIG. 14-5.)

Supply registers and return grille As discussed earlier, the most effective location for the warm-air supply registers is along the outside walls. If the supply register is not located along the outside wall, and often it is not, the

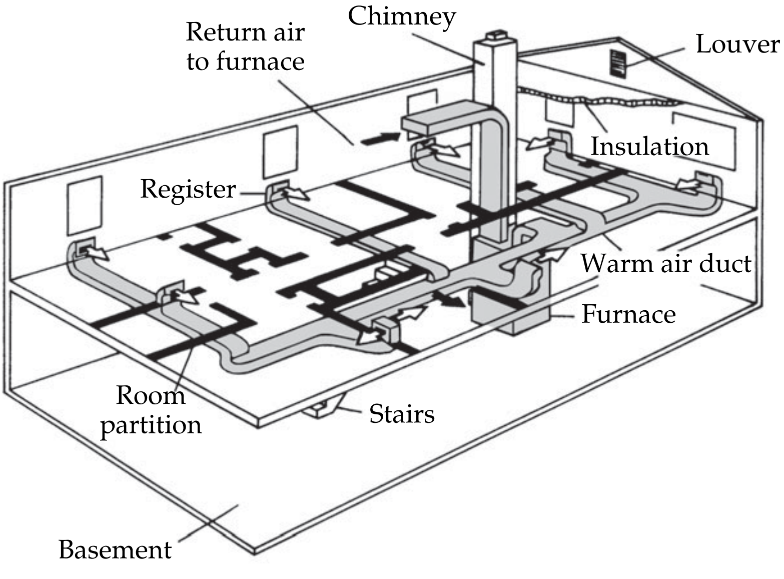
Chimney
Return air
to furnace
Louver
Insulation
Register
Warm air duct
Furnace
Room
partition
Stairs
Basement

Fig. 14-2. Warm-air distribution systems. Top—extended plenum. Bottom—radial.

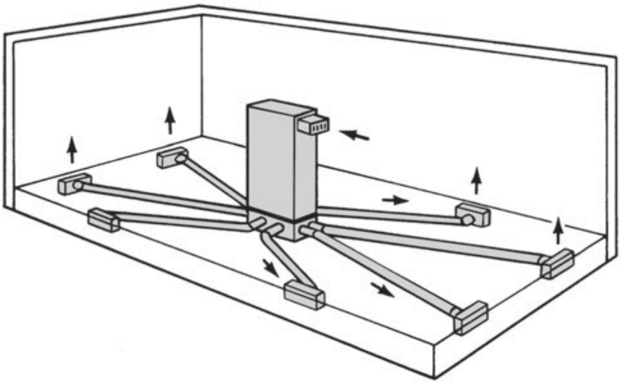

location of the return grille will be important for developing a uniform heat distribution. For optimum distribution, warm air entering the room from the supply register circulates around the room and then leaves through the return grille. If the return grille is located on the same wall as or a wall adjacent to the supply register, there will be a short cycle of the air circulation, which reduces its effectiveness. In this case, warm air discharging from the supply register can be drawn into the return grille before it has a chance to circulate.

You should look for separate return grilles for each room or a centrally located return grille for each floor. When each room has a separate return, the grille should be located on the wall opposite the supply register. If there is a centrally located return, the supply register should be located on the wall farthest from the door. This will provide a better

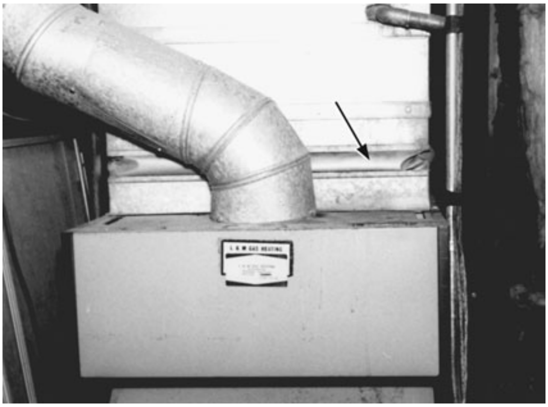

Fig. 14-3. *Canvas fabric cover over the separation between the furnace and the main duct. The separation isolates the distribution system from noises and vibrations that may develop in the furnace.*

circulation, since the air will flow through the door opening. Also, about 1 inch should be cut off the bottom of the door to allow for air movement while the door is closed.

Balancing the warm-air flow for your personal needs can best be accomplished after you move into the house. The various supply ducts will have dampers that can vary the airflow. By reducing the flow to the registers closest to the furnace, more warm air will flow to the registers farthest from the furnace. In addition, each register will have dampers that can be used for further restricting and "fine-tuning" the airflow.

Heat pump

Central heating using a heat pump is basically the same as a forced-warm-air system. However, the means by which the furnace is heated differs. The heating element is not a gas or oil burner but a component of a reversed-cycle air-conditioning system. The heat-pump system is discussed in detail in chapter 17.

Geothermal heating/cooling

A geothermal heating/cooling system is basically a modified heat pump system. It is often referred to as a ground source heat pump. Unlike an air source heat pump where the compressor/condenser is located on the outside of the house, the geothermal heat pump is generally located in the basement or utility room. It is packaged in a single cabinet that includes the compressor, refrigerant heat exchanger, and various controls. The system uses the earth as a heat source and a heat sink. It takes advantage of the earth's relatively constant temperature, at depths below 4 feet, of 40 to 55 degrees year round. The system circulates a water/antifreeze solution through a

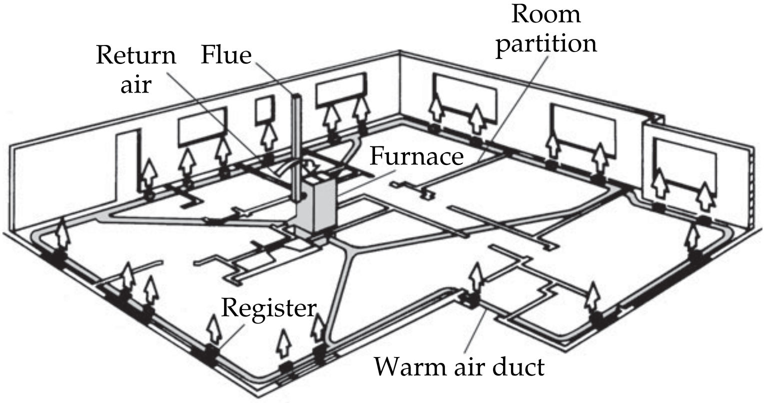

Fig. 14-4. *Perimeter-loop duct arrangement, a variation on the radial configuration.*

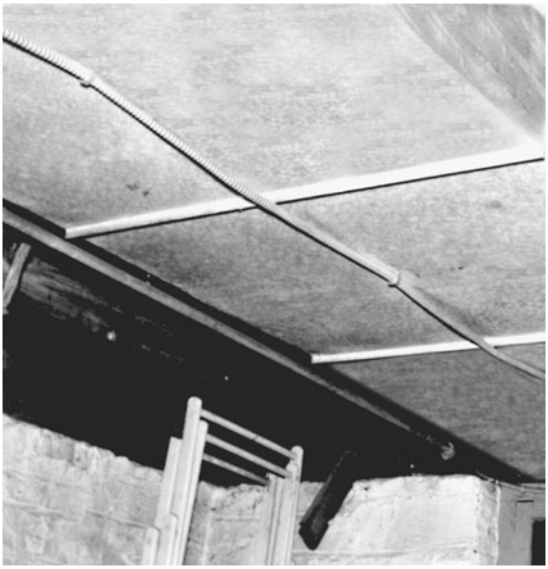

Fig. 14-5. *Return duct for warm-air heating system. The duct was made by enclosing the channel formed by adjacent ceiling joists with a sheetmetal covering. Usually found in older systems.*

closed loop of polyethylene pipe that is buried in the ground or set beneath the water. Depending on which cycle the system is operating under (heating or cooling), the solution adds or extracts heat from the heat exchanger in the heat pump cabinet. Geothermal heating/cooling is discussed in chapter 21.

Hot-water systems

This system operates on the principle of circulation and recirculation. Water heated in a boiler is transmitted through pipes to radiators located throughout the house. At the radiators, the hot water gives up some of its heat. The cooler water then continues to flow back to the boiler where it is reheated and recirculated.

In this system, the boiler, distribution piping, and all the radiators are completely filled with water. When heated, water expands, increasing its overall volume. Consequently, all hot-water heating systems must be equipped with an expansion tank to store the increased water volume temporarily. When the system is shut and the circulating water cools, the volume decreases, drawing the water back from the expansion tank. Without such a tank, excessive pressures could be built up in the system that could rupture the distribution pipes and fittings.

The water circulating within the heating system operates under a pressure that normally ranges from 12 to 22 psi. Although the water is constantly recirculating and there is no need for additional water, an automatic water-feed device is provided with all systems as a precautionary measure. The automatic water-feed device is a pressure-reducing valve. The water supply to the boiler is taken from the house water supply. Since the house supply pressure is normally in a range from 30 to 60 psi, it must be reduced before being introduced into the boiler. The reducing valve is usually preset by the manufacturer to 12 psi.

There are two basic types of hot-water heating systems—gravity and forced. They are classified according to the means by which the water within the system circulates.

Gravity hot water

As with the gravity warm-air heating system, the gravity hot-water system is inefficient, not very responsive to changing demands for heat, and no longer installed in new construction. However, it might be found in many older homes. The principle of operation is similar to that of a gravity warm-air system: As the water is heated, it becomes lighter than the cooler water and tends to rise. Since the system is filled with water, as the hot water rises, it displaces the cooler water, forcing it to return to the boiler for reheating, and thus induces circulation. To keep the resistance to flow at a minimum, the size of the distribution piping is relatively large—about 3 inches in diameter—compared to the distribution piping in a forced system—about 1 inch in diameter.

Gravity systems might also be classified by whether they are *open* or *closed.* In an open system, the expansion tank has an overflow pipe that is open to the atmosphere. The expansion tank must be located above the highest radiator to ensure that the radiator will be filled with water. It is usually found in the attic with the overflow pipe extending through the roof or side of the building. (See FIG. 14-6.) If the attic has been partitioned off into rooms, the expansion tank might be found in a closet or corner. When the tank is located in an unfinished, unheated area, it must be insulated to minimize heat loss and to protect against freezing should the system malfunction.

Hot-water and steam systems are normally equipped with automatic pressure-relief valves that discharge when the pressure exceeds 30 psi and 15 psi, respectively. A relief valve, however, is not needed in an open gravity hot-water system because if the pressure should build up to a point where it exceeds the design pressure, the water will simply discharge through the overflow pipe in the expansion tank.

In the closed system, no portion of the expansion tank is open to the atmosphere. In this case, the expansion tank can be located anywhere in the system, usually near the boiler. Because the system is closed, it can operate at higher pressures, and therefore higher temperatures, without turning to steam. The higher temperatures permit the use of smaller radiators.

There are two types of expansion tanks used in closed systems—air cushion and diaphragm. In the air-cushion type, air is initially trapped in the tank to provide a cushion, which is compressed as the water in the system expands and enters the tank. Over the years, the air in the tank can be absorbed by the water resulting in a "waterlogged" expansion tank. In this case, when the heating system is turned on, since there is no room for expansion, the pressure will increase rapidly until the relief valve discharges. When this happens, the expansion tank must be flushed and the air cushion reinstalled.

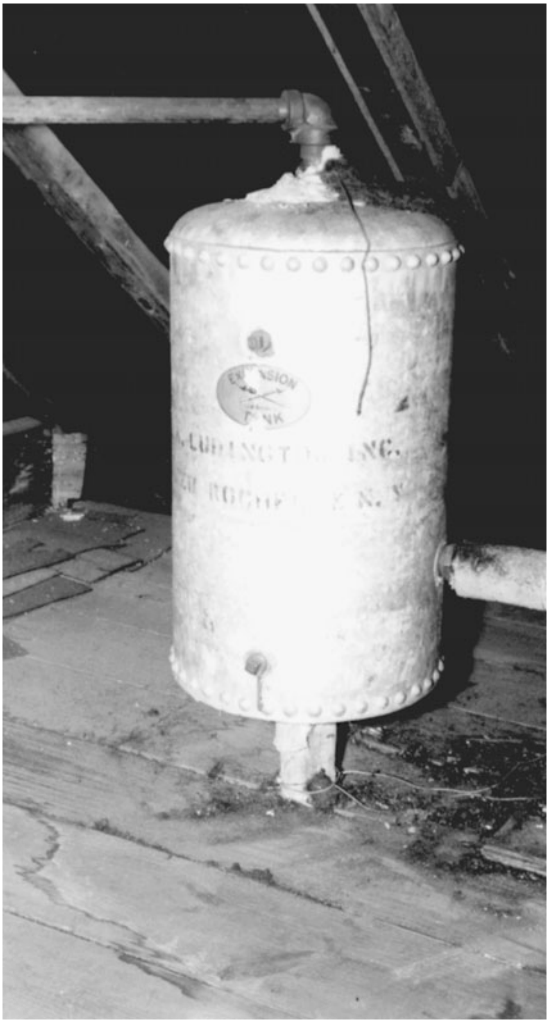

Fig. 14-6. *Expansion tank for "open" gravity hotwater system—located in attic.*

To eliminate waterlogging, the diaphragm expansion tank was developed. This tank uses a rubberized diaphragm to separate the air cushion from the water. For the most part, the diaphragm does eliminate waterlogging. However, where the diaphragm is faulty, waterlogging does occur.

If the heating system is gravity hot water, do not fret. It can be converted into a forced-hot-water system by installing a centrifugal pump with the associated controls to circulate the water through the distribution pipes and radiators. If the system is the open type, the expansion tank will have to be replaced with a closed tank.

Forced hot water

Quite often my clients comment on the small size of the boiler for the forced-hot-water system. They apparently are used to seeing a boiler originally designed for use in a gravity system or one that was converted from coal-fired to oil- or gas-fired, both of which are considerably larger—two to six times as large, depending on the manufacturer. Regardless of the size, you can easily recognize a forced-hot-water system by the presence of a circulating pump in the distribution return pipe just before the connection to the boiler. (See FIG. 14-7.)

Boilers The boilers used in these systems will be made of either cast iron or steel. Cast-iron boilers are more resistant to corrosion than steel boilers and thus have a longer projected life. The projected life span of a modern cast-iron boiler is about twenty-five to thirty years. However, many of the boiler manufacturers will provide only a twenty-year warranty. The older gravity-type boilers were probably made with a heavier-gauge metal. I have inspected many that were over fifty years old and still going strong. (See FIG. 14-8.) Steel boilers, being vulnerable to corrosion, have a shorter projected life, usually about twenty years. I have seen steel boilers that required replacement after fifteen years.

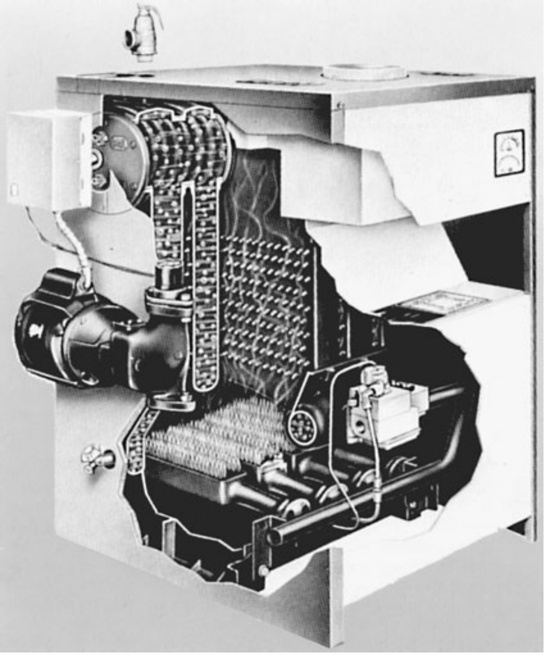

Fig. 14-7. *Boiler for forced-hot-water heating system. Note the circulating pump.*

Other than for reasons of efficiency and economy, the only time a boiler must be replaced is when a leak has developed that cannot be effectively patched. Sometimes, just after firing a boiler that has not been operational for a day or more, you might see a slight amount of water dripping into the firebox (assuming that the firebox is accessible for visual inspection). This is often the result of condensation caused by cold water circulating into the boiler or some slight joint movement. As the boiler heats up, the various sections tend to move slightly. In some cases, this results in a slightly open joint that allows water to drip out. However, as the sections continue to heat up, they expand, compressing the joint and sealing the leak. Although this condition is usually not a problem, if you see water dripping into the firebox, it would be best to have the condition checked by a professional.

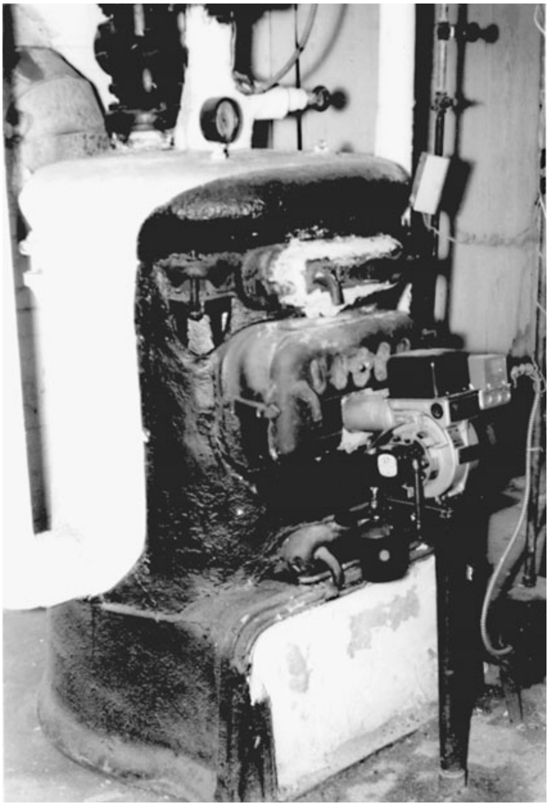

Fig. 14-8. Old cast-iron boiler with new oil burner. Boiler was converted from a coal burner.

On occasion, I have found a regular tank-type domestic water heater (of the type described in chapter 16) being used in the heating system in place of a boiler. From a safety point of view, this is acceptable because there is a high-temperature and pressure control along with a thermocouple control for the gas valve. However, water heaters have a projected life of seven to ten years and are often guaranteed by the manufacturer for only five years. In addition, with the exception of those homes located in the sunbelt, the Btu (heat) output of these units per hour is less than that needed to heat most homes adequately. This type of setup may be effective for heating an addition to a house where the existing heating system cannot be extended. The only drawback is the short projected life of the water heater.

Condensing boilers In addition to developing high-efficiency condensing furnaces, some manufacturers also developed high-efficiency boilers. As with the furnaces, the increase in efficiency results from extracting the heat from the exhaust gases that normally flow up the chimney with conventional boilers. By redesigning the heat exchanger within the high-efficiency boiler, more of the heat extracted from the exhaust (flue) gases is transferred to the water. In the process, the temperature of the flue gases drops to a point where the water vapor in the gases condenses, thereby giving up more heat. This process raises the overall operating efficiency of the boiler to about 87 percent, compared to about 60 percent for the conventional boiler.

Although boilers heat water and furnaces heat air, there are many similarities between high-efficiency boilers and furnaces. Neither requires a chimney for venting hot exhaust gases. They can vent the relatively cool flue gases through a side wall using a plastic pipe. Both require an intake air duct for those units that use outside air for combustion. Both require an induced draft blower (power vent). Both also require provisions for condensate drainage.

Pulse boiler Another type of high-efficiency boiler is the gas-fired Hydro-Pulse boiler. Its overall operating efficiency is about 90 percent. As with the boiler discussed above, the high operating efficiency is the result of extracting so much heat from the exhaust gases that the water vapor contained within condenses. However, the combustion process differs. With the gas-fired high-efficiency boiler, as well as the conventional boiler, heat is generated as a result of a continuous burning of the gas-air mixture, whereas with the Pulse boiler, heat is generated as a result of sixty to seventy mini explosions of the gas-air mixture per second. This results in a noisier operation than with a conventional or high-efficiency boiler. Consequently, vibration isolators are usually installed between the boiler and connecting pipes to minimize noise transmission.

Advantages and disadvantages Because the forced-hot-water system operates under pressure with the water being circulated by a pump, it is very flexible. It can be used to heat an area below the level of the boiler. Additional, completely independent heating zones can be readily installed. Because the hot water remains in the pipes after the boiler is no longer being heated, there is less heat fluctuation and a more even temperature distribution. Also, the operation of this system is relatively quiet.

One of the disadvantages of this system is that with water in the pipes, the distribution system is vulnerable to freezing temperatures. In the event of an extended power failure or if an oil-fired boiler runs out of oil for several days during a cold spell, the pipes can freeze and burst. I know of several cases where antifreeze was introduced into the circulating water to prevent that type of a problem. In addition, this system is not adaptable to central conditioning of the air such as cooling, humidifying, and filtering.

Distribution piping There are three basic types of distribution piping arrangements for forced-hot-water systems: series loop, one pipe, and two pipe. The specific distribution system installed usually depends on the size and cost of the house.

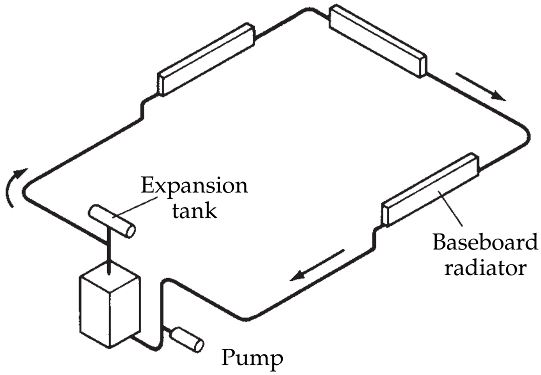

Fig. 14-9. *Series-loop piping configuration for forced-hot-water heating system.*

Series loop This is the simplest and least expensive piping arrangement to install. It is usually used in smaller homes where room-by-room radiator adjustments are not needed for balancing the heat distribution. In a series loop, the radiators, usually baseboard convectors, are an integral part of the supply piping. (See FIG. 14-9.) If a radiator is shut off, the flow throughout the system is stopped. In this arrangement, the temperature of the water entering the last radiator will be considerably less than it was when it entered the first radiator. To minimize the difference in temperature and produce a more uniform heat distribution, the heating system for larger homes is often designed so that the house is divided into two or more heating zones. Each zone has a separate piping configuration and either has a separate circulating pump or shares a common circulating pump with the other zones, but has a separate thermostatically controlled valve in the main supply pipe.

One pipe In a one-pipe distribution configuration, as with the series loop, a single pipe makes a complete circuit from the boiler and back again, serving as both the supply and the return. In this case, however, rather than the radiators being integral with the supply pipe, they are attached to it by two risers, one connected to each end of the radiator. (See FIG. 14-10.) Each radiator will also have a shutoff valve located at the inlet riser. Each radiator can be shut without affecting the water flow in the supply main. Consequently, this system can provide room-by-room heat control. However, as with the series loop, there is a considerable temperature difference between the water entering the first and last radiators. To compensate for the cooler water entering

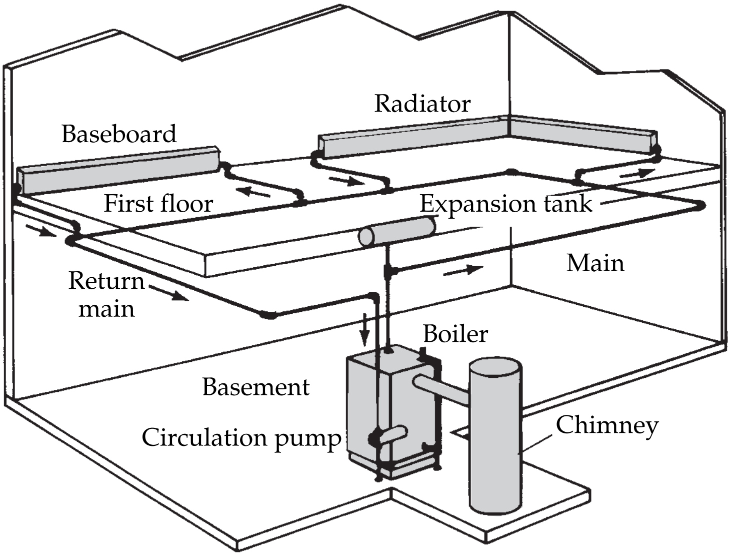

Fig. 14-10. *One-pipe distribution configuration for forced-hotwater heating system.*

the radiators downstream, larger radiators are often installed. They emit heat comparable to the smaller ones closer to the boiler.

Two pipe The two-pipe distribution configuration is the most costly to install, but it overcomes the deficiencies of the other configurations. There are two main pipes, one for the supply and the other for the return. The inlet and outlet ports of the radiators are attached to the mains by risers. (See FIG. 14-11.) The cool water leaving the radiator does not mix with the hot water flowing through the supply main, so the temperature difference between the water entering the first and last radiator is small.

Even though both the one- and two-pipe configurations will provide heat control for individual rooms if the radiator shutoff valve is manually closed or partially closed, the configurations are also often used in homes with zone heating. Zone heating is automatic. All that has to be done is to set the thermostat to the desired temperature for that portion of the house.

Radiators As discussed previously, the optimum location for a radiator is along an outside wall, preferably near or under a window. There are three basic types of radiators found in hot-water heating systems: freestanding cast iron, freestanding convector, and baseboard convector. Freestanding cast-iron radiators are found mostly in older homes. Unless an old home is being restored, a freestanding radiator is usually considered undesirable. Many clients have asked whether the radiators can be replaced with the newer finned-tube baseboard radiators. Yes, they can. Baseboard convector radiators are usually preferred because they distribute the heat better than cast-iron or freestanding convector radiators and are less conspicuous. They achieve an even temperature throughout the room because they distribute the heat near the floor. Natural convection causes the heated air to rise and warm the outside wall.

Panel heating Another method of heating using a forced-hot-water system is to embed the distribution piping into the walls, floors, or ceilings and let those areas (panels) function as radiators. (See FIG. 14-12.) The heat from the distribution piping is conducted to

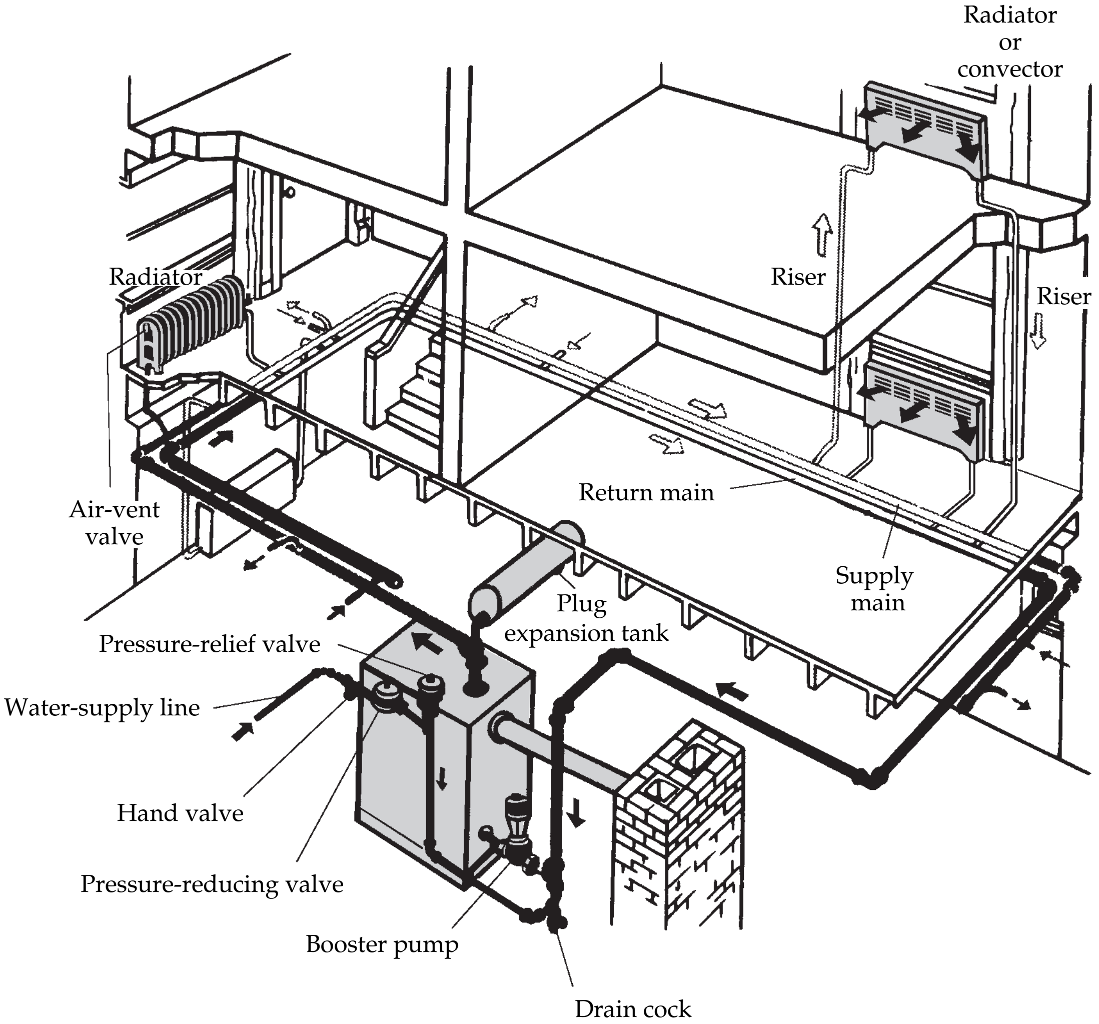
Radiator
or
convector
Radiator
Riser
Riser
Air-vent
valve
Return main
Supply
main
Plug
expansion tank
Pressure-relief valve
Water-supply line
Hand valve
Pressure-reducing valve
Booster pump
Drain cock

Fig. 14-11. *Two-pipe distribution configuration for forced-hot-water heating system.*

the surface of the panels; which in turn heats the room by radiation and convection. This system produces a very uniform temperature distribution and is particularly effective when the heating panel is the floor slab in a basementless house.

Controls In addition to the thermostat, hot-water systems have safety and operational controls. In a simple forced-hot-water system without individual zones, there is a high-temperature limit control that prevents the boiler water from exceeding a preset temperature and a controller for the circulating pump. Depending on the system design, the circulating pump can operate in any one of three modes: constant-running circulator, aquastat-controlled circulator, and relay-controlled circulator.

In the constant-running circulator, the pump is energized by a manual power switch

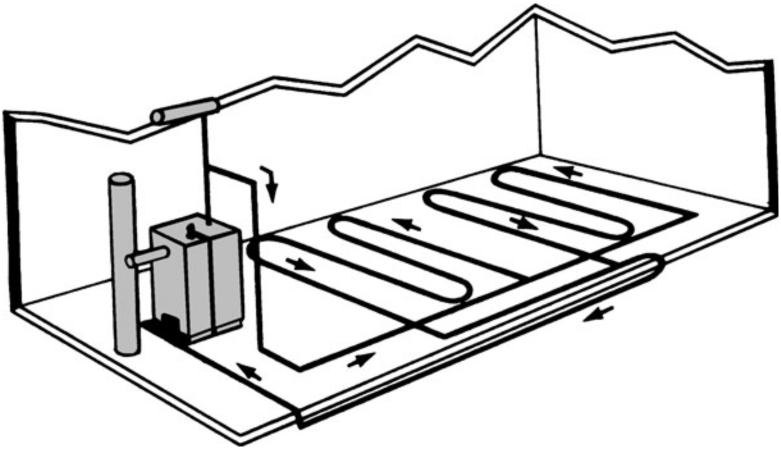

Fig. 14-12. *Forced-hot-water radiant heating panels.*

and operates continuously during the heating season. When the thermostat calls for heat, it fires the burner that heats the boiler. When the heating season is over, the circulating pump must be manually shut off. Sometimes the homeowner forgets to shut the pump, and the pump runs all summer. This type of operation is somewhat wasteful of electrical energy. However, this wastefulness must be weighed against the fact that intermittently operated circulating pumps tend to break down sooner than continuously operated pumps. The constant starting and stopping causes the bearings to wear out sooner. A constant-operating circulating pump is the least expensive to install because there is no relay or temperature controller. It can be converted to either—aquastat-controlled or relay-controlled.

In the aquastat-controlled circulator mode, as with the constant-running circulator, the thermostat will control only the burner. When the boiler water reaches a preset temperature (approximately 120° F), the circulator pump begins to operate. After the thermostat is satisfied and shuts off the burner, the circulator continues to operate until the water temperature drops below the temperature setting of the circulator control. In the relay-controlled mode, the thermostat simultaneously activates the burner and the circulating pump. When the thermostat is satisfied, it will simultaneously shut off the burner and the circulating pump.

Zone control Heating zones in a forced-hot-water system can have a separate circulating pump for each zone or a single circulating pump that services all the zones. When there is a single pump, each zone is controlled by an electrically activated valve that is in turn controlled by a thermostat. When one of the zone thermostats calls for heat, it opens the respective zone valve, fires the burner that heats the boiler water, and starts the circulating pump. After the system is operational, if another zone thermostat calls for heat, it merely opens the zone valve. This allows the hot water to circulate through that zone.

When there are separate circulating pumps for each zone, the first thermostat calling for heat will fire the burner and activate the pump. Thereafter, as long as the system is operational, other thermostats calling for heat will merely activate their respective circulating pumps.

Domestic water heater Most boilers used in hot-water systems can be equipped so that they will also heat the domestic water (water used for washing and bathing). This is discussed in detail in chapter 16. When a heating-system boiler produces domestic hot water, the associated burner must fire all year long and not just for the heating season. In this case, the thermostat does not control the burner. It controls only the circulating pump. The burner is activated by an aquastat that controls the boiler water temperature.

Since the boiler will have hot water during those periods when heat is not required, a flow-control valve must be installed on the supply main. This valve prevents the hot water from rising into the distribution piping and heating the house like a gravity hot-water system. When heat is required, the circulating pump produces sufficient force to lift the flow-control valve and circulate the hot water. In a multizoned system, if zone valves are used, flow-control valves are not needed. When the zone valves are closed, water will not circulate in the distribution piping.

Relief valve Every forced-hot-water heating system must be equipped with an automatic pressure-relief valve as a safety control. To help ensure its effectiveness, the relief valve should be mounted directly on the boiler. Many systems have a relief valve that is mounted in the boiler feed line several feet away from the boiler. This is an improper location because over the years the relief valve can become isolated from the boiler as a result of a lime or scale buildup in the feed line. If this should happen, the valve would be completely ineffective in the event of a pressure buildup. If the heating system does not have a boiler-mounted relief valve, you should have one installed.

Pressure, temperature gauge Although not safety or operational controls, all forced-hot-water systems should have a pressure and a temperature gauge. Newer systems have a combination gauge that measures both pressure and temperature. In older systems, you probably will find a separate pressure gauge and a pencil-type thermometer attached to the boiler. The combination gauge often has two pressure scales, one in pounds per square inch (psi) and the other in altitude (feet of a column of water). (The altitude scale can basically be ignored, although for your information, one psi is equivalent to a column of water measuring 2.31 feet high.) The normal fill pressure of the boiler is 12 psi. This, then, is equivalent to an altitude of 27.7 feet. If the highest radiator in your house is more than 27.7 feet above the boiler, then a higher boiler-water pressure will be needed. Some pressure gauges have two points—one fixed and one movable. The fixed pointer is usually positioned over the normal altitude setting for that house. The movable pointer shows the actual working pressure of the heating system. If this pressure exceeds 30 psi, the relief valve will discharge.

Low water cutoff An automatic water feed valve is normally installed with a hot water boiler. It provides make up water in the event of a low water condition caused by a leak in the system or a discharging relief valve. Over the years these valves can become clogged with sediment and become ineffective. Electronic low water cutoffs are now being installed on new boilers to counter this potential problem. These units have a probe that projects into the water. As long as the water is covering the probe, an electronic circuit is maintained. If the water level drops below the probe, the circuit is broken, which automatically shuts off the burner and prevents damage to the boiler.

Steam heating systems

For the most part, steam heating systems are no longer used in new residential construction, but you will find them in older homes. In fact, in some of these homes, the boiler and its associated controls may be relatively new.

The boiler used in this system is basically the same as that used in a hot-water system. It will be made of either cast iron or steel. (See the discussion on boilers on pages 221–223. You can tell whether a boiler is being used for a steam system or a hot-water system by the type of controls and gauges used. A quick indication is a water-level gauge. (See FIG. 14-15.) If you see such a gauge, the boiler is being used to generate steam.

Unlike the hot-water system that is completely filled with water, the steam boiler is only partially (about three-quarters) filled with water, depending on the size and make. The remaining portion of the boiler, the distribution piping, and the radiators will be filled with air. After the system is fired, the water heats up until it boils, as in a tea kettle. The steam thus formed rises in the pipes of its own accord, without the aid of a fan or pump, pushing the air ahead of it as it moves along. The air from the pipes and radiators is then dissipated into the rooms through air-vent valves that, depending on the piping configuration, are located on the radiator or near the end of the steam main.

When the steam comes into contact with the cool radiator surface, it condenses back into water and in the process gives up its heat. The resulting water then flows back to the boiler for reheating. If because of a blockage in an air vent (often the result of painting the vent) the air cannot be evacuated from a radiator, a pressure will be built up within the radiator that will prevent the steam from entering. In this case, the radiator will be ineffective because it will not heat up. This condition can be easily corrected by replacing the air vent.

Distribution piping

There are two basic distribution piping configurations used with steam heating systems: one pipe and two pipe. You can easily tell which configuration is being used by the number of pipes connected to the radiator. In a one-pipe configuration, there is only one pipe connected to the radiator; in a two-pipe configuration, there are two.

In a one-pipe arrangement, the steam is distributed to the various radiators through the same pipe that carries the condensate back to the boiler. (See FIG. 14-13.) The radiators used in this system must be pitched so that the condensate flows back through the supply valve. Otherwise, the condensate can accumulate and block the steam flow. Each radiator has a manually operated supply valve and an air vent. Some radiators are equipped with an air-vent valve that has an adjustable opening. This opening can be increased or decreased in size, thus allowing the air in the radiators to be evacuated at a faster or slower rate. Adjustable air vents are often found in larger homes and are used as a means of providing a uniform steam supply to all of the radiators. Radiators closest to the boiler will receive steam before those that are farther away. In some cases, depending on the distance apart and the size of the radiator, those closest to the boiler can be fully heated before those farthest away receive any steam. By decreasing the vent opening on those radiators closest to the boiler and increasing the opening on those farthest away, it is possible for all of the radiators to receive steam at about the same time.

In a two-pipe configuration, the steam is supplied to the radiator by one pipe and the condensate returned to the boiler through another. The radiators in this system are not equipped with individual air-vent valves. They have a steam trap on the condensate return pipe. A steam trap allows the air bound in the radiator and the condensate to flow in the return pipe but closes on steam contact and does not allow the passage of steam. The air in the return line is then vented by a main vent. A two-pipe steam system can be converted to a forced-hot-water system. This cannot be done

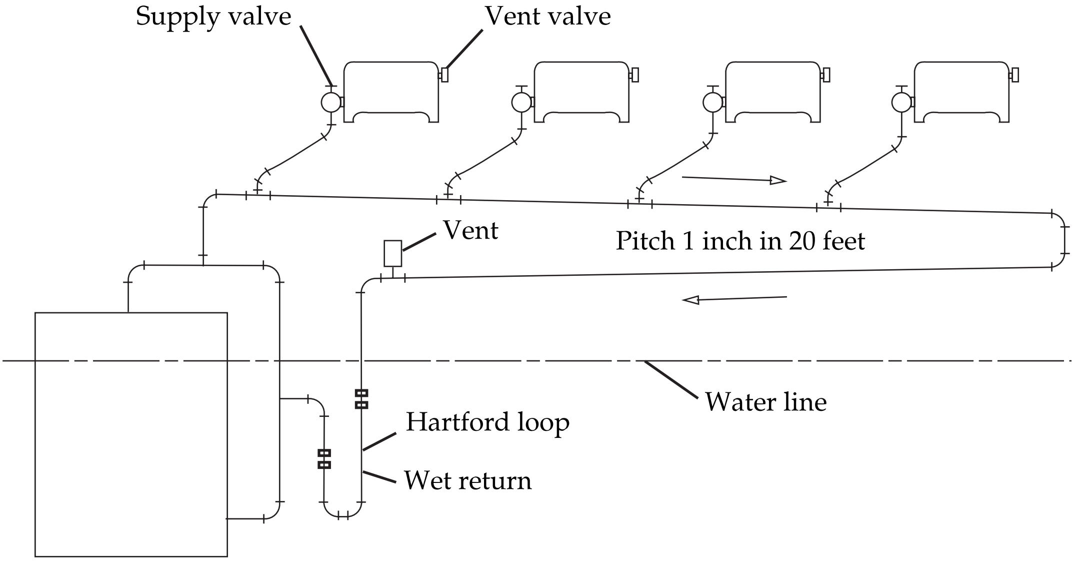
Supply valve
Vent valve
Vent
Pitch 1 inch in 20 feet
Water line
Hartford loop
Wet return

Fig. 14-13. *One-pipe distribution configuration for steam heating system.*

with a one-pipe steam system. In both one- and two-pipe systems, when the condensate is returned to the boiler, if the return line in the boiler room is above the boiler-water level, it is called a *dry return*. If the return line is below the boiler water level, it is called a *wet return*. When the system has a wet return, there should be a special piping arrangement at the boiler, a *Hartford loop*. (See FIG. 14-14.) The purpose of the Hartford loop is to prevent water from draining out of the boiler in the event of a leak in the wet-return piping. If a leak occurs in the return line, boiler water will drain down only until it reaches the top of the Hartford loop. There will still be sufficient water to prevent damage to the boiler if it continues to fire. If the heating system has a wet return, look for a Hartford loop. If you do not see one, you should consider its installation.

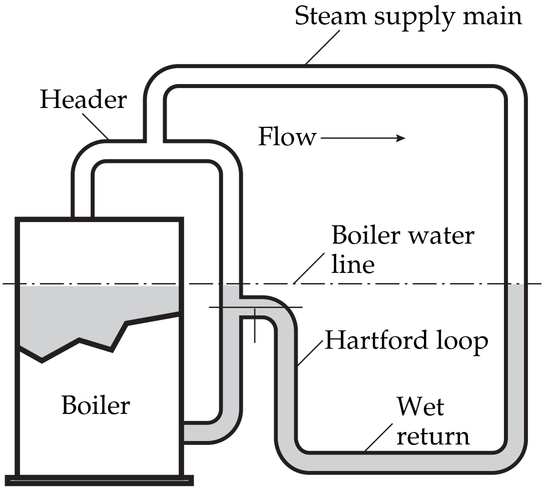
Steam supply main
Header
Flow
Boiler water line
Hartford loop
Boiler
Wet return

Fig. 14-14. *"Hartford loop" in the condensate return line of a steam heating system. The Hartford loop prevents water from draining out of the boiler in the event of a leak in the wet-return piping.*

Controls

In addition to the thermostat, every steam system should have a high-pressure limit switch, a low-water cutoff, and an automatic pressure-relief valve. Also, to determine whether the boiler is operating properly, there should be a water-level gauge and a pressure gauge. (See FIG. 14-15.) The high-pressure limit switch is connected electrically to the burner control. When the steam pressure exceeds a predetermined setting, the limit switch will shut down the burner, thereby preventing the pressure from building up further. The limit control should be physically connected to the boiler with a pipe that has a curl that looks like a pigtail. The pigtail has water in the bottom of the loop, which prevents the corrosive action of the steam from affecting the control.

The low-water cutoff is a control that shuts down the burner when the level of the water in the boiler drops below the design level. There are two types of low-water controls—one mounted inside of the boiler and one externally mounted. The latter is preferred because it provides a convenient means for testing its operation. This unit has a blow-off valve that when opened drops the water level in the control, causing it to operate. In the past, manufacturers recommended that the low water cutoff be opened and blown down once every month to prevent sludge accumulation, which can affect its operation. They are currently recommending that it should be blown down once a week. All too often, homeowners neglect to perform this simple operation. I have checked many a unit that apparently had not been flushed in years; when I opened the blow-off valve, either nothing flowed out or there was a thick dark-brown sludge oozing out. The low water cutoff units that are mounted in the boiler have an electronic probe that projects into the water. As long as the water covers the probe, an electronic circuit is maintained. If the water level drops below the probe, the circuit is broken, which automatically shuts off the burner and prevents damage to the boiler.

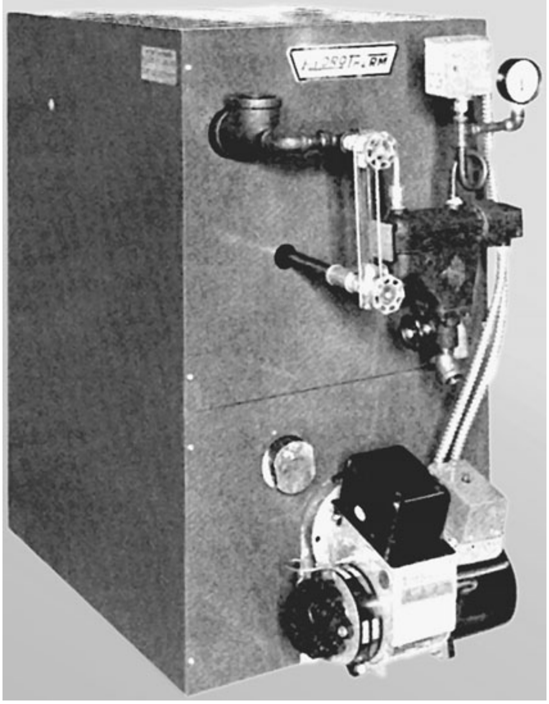

***Fig. 14-15.** Oil-fired steam boiler, showing the level gauge, low-water cutoff, high-pressure limit switch, and pressure gauge.*

The water-level gauge provides a convenient means for determining the level of the water in the boiler. It is usually mounted on the side of the boiler; when there is an exterior-mounted low-water cutoff, it is often part of that assembly. The water level should be at the midpoint or two-thirds up the glass gauge. The exact position is not important. What is important is that you are able to see the water level. If the entire level gauge is filled with water, there is too much water in the system. In fact, the system can be flooded. If too much water is introduced into the boiler, the water level will rise until it fills the entire distribution system and radiators. If any of the radiator valves or fittings are not watertight, and sometimes they are not, the water will leak out all over the room.

When there is no water visible in the gauge, water must be introduced into the boiler. This can be done by manually opening the fill valve in the water-supply line. Some systems have an automatic boiler-water feeder that introduces water to the required level as needed. This is a desirable feature, but occasionally such units malfunction. Often level gauges are coated with sediment so that the water level is not visible. In this case, the glass gauge must be cleaned and all the accumulated sediment removed.

The relief valve is a safety valve that automatically discharges when the operating pressure exceeds the design pressure. For a residential steam system, the relief valve is set to discharge at 15 psi, although the normal operating pressure is considerably less, usually about 1 to 2 psi. Some large systems may even operate at a negative pressure (vacuum).

Domestic water heater

As with hot-water boilers, steam boilers can also be equipped for generating domestic hot water. When it produces domestic hot water, the boiler must be fired all year long. When heat is not required, the boiler-water temperature is controlled by an aquastat. The aquastat activates the burner when the boiler-water temperature drops below a preset figure and shuts off the burner when the boiler-water temperature rises to about 200° F. When steam heat is required, the burner is activated by the thermostat. In this case, the thermostat overrides the aquastat control so that steam is produced.

In a forced-hot-water heating system that also produces domestic hot water, a flow-control valve is needed to prevent the boiler water from circulating as a gravity system during those months when heat is not required. In a comparable steam system, no flow-control valve or equivalent is necessary because the boiler water is not heated sufficiently to produce steam.

Advantages and disadvantages

The advantages of a steam heating system can be appreciated more in a large building than in a residential structure. It is a relatively simple system that does not require a pump or fan for circulation of the steam. Since there is no water in the pipes when the system is not operating, there is no problem of the pipes freezing and bursting. If a repair or replacement is needed to a section of pipe or a fitting, it is not necessary to drain the system. Also, a steam leak would result in very little water accumulation.

On the other hand, a steam system is slow in responding to an initial rapid change in heat demand because the boiler-water temperature must be brought up to 212° F before steam circulation begins. Unless the condensate is returned to the boiler by means of a pump, the boiler must be located below the radiator in the lowest rooms.

Hybrid heating systems

Depending on the extent to which the heating system has been modified or expanded, you may find a hybrid heating system. A hybrid system is composed of two separate heating systems working together.

Steam–hot water

I found the following in an old house. The system functioned as a two-zone system. The main portion of the house was heated by a steam system, and the remaining portion of the house (the lower sections) was heated by a forced-hot-water system. The distribution pipe for the hot-water system was tapped directly off the portion of the steam boiler that was below the water level, and a pump was used to circulate the hot water. The thermostat for this zone controlled the pump and the burner. There was also an aquastat temperature control for the boiler water to keep it from generating steam when the main zone was not calling for heat.

Hydro-air

Hydro-air is basically a hot water–warm air system. It is becoming more popular, especially in larger homes. The system essentially consists of an air handler, which looks like the blower compartment of a furnace, a hot water boiler, and a duct system. Hot water from the boiler is pumped through a series of tubular coils that are mounted above the fan in the blower compartment. As air flows over the coil, it is heated and blown through the duct system to various areas of the house. There are a number of advantages to this system. You can easily add a number of separate zones to the boiler piping; one zone to an indirect fired water heater, another zone to a room that is a new addition and that is heated with baseboard, and a third to another air handler. Another advantage to hydro-air is that the same duct system can be used for the air-conditioning. Also, since no combustion takes place inside the air handler, there is no possibility of combustion gases being distributed throughout the house because of a cracked heat exchanger.

In large homes with more than one air handler, at least one air handler is usually located in the attic. In the event of a prolonged power failure, the water in the coils is subjected to freezing. If the home you are considering has a unit in the attic, as a precaution you should check to determine whether antifreeze has been added to the system either by a tag or a marking on the boiler or by asking the seller.

15

Heating systems II

Oil-fired systems 233
Gas-fired systems 235
Electrical systems 236
Area heaters 236
  Gas-fired units 237
  Electrical units 237
Heating system inspection procedure 237
  Warm-air systems 239
  Hot-water systems 241
  Steam systems 242
  Oil burners 243
  Gas burners 244
Checkpoint summary 244

The boilers and furnaces found in most homes today are heated with oil or gas burners or electrical resistance coils. In many older homes, you will find a boiler or furnace that was originally designed to burn coal but was converted to oil or gas for reasons of economy and convenience. The accessories associated with each heating element are described below.

Oil-fired systems

The most common type of oil burner used in residential heating is the high-pressure or gun-type burner. This unit has a pump that forces oil through a nozzle and produces an oil mist. It also has a fan that mixes the oil mist with a measured amount of air. The resulting flammable mixture is then ignited by an electric spark and burns in a refractory-lined firebox. If the mixture fails to ignite, a safety control will shut off the oil-pump motor. This control is located either in the exhaust stack as a heat-sensitive switch or in the burner as a light-detecting photocell.

The oil burner normally associated with the heating system will generally be either an older conventional unit or a high-efficiency flame-retention burner. Although both burners look somewhat the same to the untrained eye, there is a considerable difference in the flame pattern and overall efficiency. The flammable mixture of oil and air that is produced with a conventional burner is inconsistent. Some parts of the mixture will be fuel-rich and others fuel-lean. Consequently, these burners draw in a large quantity of air so that the fuel-rich portion of the flammable mixture will burn completely and thereby reduce the level of smoke produced.

The large quantity of air and the irregular flame pattern results in inefficient operation of the conventional burner. The flame-retention burner is more efficient. Because of its design, it produces a more uniform mixture of oil and air, requiring a smaller amount of air for complete combustion. This in turn produces a smaller, more compact flame that is hotter. A flame-retention burner gets more Btu out of a gallon of oil than a conventional burner. In fact, a new flame-retention burner is about 16–20 percent more efficient than a new, well-adjusted conventional burner. You can check with the company that services the oil burner to determine whether it is a conventional unit or a flame-retention burner.

The proper draft over the firebox is very important for efficient operation of the oil burner. To ensure the proper draft, most oil burners have a draft regulator mounted in the exhaust stack near the boiler or furnace. The regulator is basically a small swinging damper that can be adjusted to open an inch or two when the burner is firing. (See FIG. 16-7.) I have seen many draft regulators that have been made inoperative by someone cementing shut the opening. This can affect the efficiency of operation and result in greater fuel consumption. Also, to ensure the proper draft, the section of the exhaust stack between the boiler or furnace and the chimney must have an upward pitch.

Another type of oil burner occasionally found in central heating systems is the pot-type vaporizing burner. It contains few moving parts and therefore operates quietly. Basically, the unit consists of a pot containing a pool of oil and a control for regulating the oil flow to the pot. Air needed to produce a flammable mixture is introduced by a small fan. Once the mixture starts burning, the heat produced vaporizes the oil mixed with the air, thus maintaining the burning operation. The initial cost for this burner is less than that for a high-pressure gun burner. However, for efficient operation, it requires a more costly high-grade oil that vaporizes easily.

Oil for the burner is stored in a tank that is either buried in the ground just outside the house or located in the house not too far from the burner. The interior tank should be at least 7 feet from any flame and should have an outside fill connection. These tanks will generally have a 275-gallon capacity; buried tanks generally have a 550-, 1,000-, or 1,500-gallon capacity. Most municipalities allow a maximum of two 275-gallon tanks to be stored in the structure. Oil tanks have a projected life of about thirty years. After that time, pitting and corrosion holes tend to form on the bottom of the tank. This condition is the result of an accumulation of water and corrosive acids that settle on the bottom of the tank. Check to see if the bottom of the tank is at least 6 inches off the ground and whether the tank is at least 2 inches away from the walls. In many municipalities this is a code requirement. Run your hand across the bottom of the tank. If your hand has an oily film, the tank is leaking, even if it is not dripping oil. Recommend that the tank be replaced. Be advised that if there is a buried oil tank that is leaking, it is the homeowner's responsibility to clean it up, including removal of contaminated soil. See "Leaky oil tanks" in chapter 20.

All oil burners require periodic inspection, cleaning, lubrication, and adjustment to ensure an efficient soot- and odor-free operation. During the inspection, which should be performed at least once a year, the exhaust gas should be analyzed to determine whether the burner is operating at maximum efficiency. Preventive maintenance to the oil burner is usually provided by the company that supplies the oil. For an additional fee, oil companies usually issue a maintenance contract in which they agree to maintain and tune the system and provide emergency service. If you enter into a maintenance contract with your oil company, make sure you know exactly what will be maintained. Some companies will inspect and maintain only the oil burner and not any of the peripheral equipment or controls.

I inspected a house one summer that had an oil-fired forced-hot-water heating system that also generated the domestic hot water. The burner, of course, had to fire during the summer to produce the domestic hot water. The circulating pump, however, had not operated for a few months because heat was not needed. When I turned on the heating system (by turning up the thermostat), the circulating pump started to smoke, and I smelled insulation burning. The homeowner, who was with me at the time, said, "How can this be? The oil company just serviced the system two days ago." I called the oil company to find out what kind of maintenance was performed. They informed me that they maintained the oil burner only. They did not even look at the circulating pump. Do not let a maintenance contract lull you into a false sense of security, thinking that your entire heating system is being maintained. It might be. But it must state so in the contract.

Gas-fired systems

If you are given an option and the fixed costs are comparable, a gas burner is preferred to an oil burner. It does not require annual maintenance and is less costly to install. Gas burners have a cleaner, quieter operation, and the supply of gas is not dependent on the weather. In addition, replacement costs are lower.

In this system, gas is supplied at low pressure to the burner through an automatic gas valve. The valve in turn is controlled by the thermostat or aquastat (if the system generates domestic hot water). At the burner, the gas that is mixed with air is ignited by a pilot light. Associated with the pilot light is a safety control, a *thermocouple*. The thermocouple closes the gas valve when the pilot light goes out. Basically, the thermocouple converts heat (from the pilot light) into a small electric current that keeps an electrically operated valve in the main gas line open. When the pilot light goes out, the current ceases, and the valve closes. If the thermocouple is faulty, the pilot will not light. Rather than have a pilot light that is constantly lit, many new boilers use an electronic ignition to fire up the burner.

Some gas-fired systems are equipped with a self-energizing control unit (Powerpile) that enables them to operate in the event of an electrical power outage. This is a particularly desirable feature, especially for those areas prone to power failures. The control unit basically consists of a special pilot-thermocouple assembly, gas valve, and thermostat. It can be used on steam, hot-water, and warm-air systems. With a self-energizing control unit during a power failure, a steam system will operate normally and a forced-hot-water or warm-air system will function as a gravity system.

As with oil burners, gas burners require a small but steady draft. However, rather than using a draft regulator, gas-burning equipment uses a draft diverter-hood. (See FIG. 16-7.) The hood is usually located on the exhaust stack above a boiler and is often built into the sheet-metal casing of a furnace. A draft-diverter hood is completely open on the underside and thus prevents air currents in the chimney (resulting from downdrafts) from blowing out the pilot-light flame.

All gas-burning equipment should have the American Gas Association's seal of approval. This indicates that a similar unit has been tested for safety. Look for the seal as you inspect the equipment. It is usually located on a plate mounted on the front of the boiler or furnace.

What type of pipe is used to supply gas to the furnace or boiler? It should be a rigid black iron pipe rather than a flexible pipe or a copper pipe. Many utility companies buy their gas from the cheapest supplier, and therefore, the gas can contain impurities that will react with copper. Although most municipalities allow flexible gas pipe for movable appliances such as clothes dryers and stoves, they generally require a rigid gas-supply pipe for stationary equipment such as a boiler, furnace, or domestic water heater. If the gas-supply pipe to the heating system is other than a black iron pipe, check with your local utility company for their requirements.

Check the gas meter to see if it is adequately sized to supply gas to the heating system and domestic water heater during peak periods. An inadequately sized gas meter might be found in a house that originally had a gas-fired water heater and an oil-fired heating system. Sometimes when the oil-fired system is converted to or replaced by a gas-fired system, the installer does not notify the utility company to replace the meter, which was sized to supply gas for only the water heater and kitchen stove.

A gas meter is rated in cubic feet per hour, and one cubic foot of natural gas is approximately equal to 1,000 Btu. Therefore, a meter rated at 250 cubic feet per hour can supply gas that is approximately equivalent to 250,000 Btu per hour. You can check to see if the gas meter is adequately sized by adding up the input Btu/hour requirements for the boiler or furnace and the water heater, and comparing the total with the meter. The input Btu will generally be found on the data plate that is mounted on the outer casing of the appliance. (The gas requirement for the kitchen stove is considered negligible for this comparison). The Btu capacity of the meter should be equal to or greater than the sum of the input Btu requirements for the heating system and water heater. If it isn't, make a note on your worksheet to notify the utility company.

Electrical systems

For the most part, electrical resistance heating is used for area heaters such as panel or baseboard heaters. It is also used, although less often, for central heating. The heating mechanism in a boiler or furnace is very simple. Electrical resistance coils are immersed directly into the water (for a boiler) or airstream (for a furnace). As the electrical current passes through the coils, they get hot and directly transmit their heat to the air or water. Unlike oil and gas energy, the conversion of electrical energy to heat is 100 percent. There is no heat lost to the exhaust gas because there are no exhaust gases. There is no fuel combustion with this system. Consequently, there is no need for a chimney.

From an installation point of view, electrical resistance coils are less expensive than oil or gas burners. However, for most parts of the country, they are more expensive to operate. To keep the operating expense from becoming excessive, electrically heated homes must be well insulated. In evaluating the cost of heating a house, do not depend on the owner's assurance ("Oh, it only costs a few hundred dollars to heat the house"). Insist upon seeing the most current year's bills.

Area heaters

These units are self-contained space heaters that are used to provide warmth to a single room. Sometimes they warm up two or three adjacent rooms if the doors between the rooms are kept open. The heat distribution is not very desirable from a comfort point of view. Other than a fireplace or stove, most area heaters of today will be gas-fired or electric.

Gas-fired units

Gas-fired space heaters are basically small warm-air furnaces and can operate either as gravity or forced-air units. Some of the smaller gravity heaters have an open combustion area exposing the flames. This is a potential fire hazard and is particularly undesirable when there are small children in the house. The exhaust gases from all gas-fired space heaters must be vented to the outside. In addition, the units should have the American Gas Association seal of approval. These units are controlled either automatically by a wall-mounted thermostat or manually by a calibrated knob on the gas valve. They are inexpensive to install, relatively maintenance-free, and require only periodic cleaning and adjustment.

Electrical units

Electrical space heaters have an advantage over central heating systems in that each room is usually individually wired, so there are as many independent heating zones as there are rooms. There are three general types of electrical area heaters: panel, baseboard, and wall. *Panel* heating, sometimes called *radiant heating*, is similar in concept to forced-hot-water panel (radiant) heating, as described in chapter 14. However, instead of hot-water pipes being embedded in the walls or ceiling, electrical resistance cables are installed with each room having a separate control. Similarly, *baseboard* heaters are like baseboard convectors (see chapter 14) but they contain electrical resistance coils rather than circulating hot water. They are very popular because of the ease of installation and low initial cost. They also produce a uniform heat distribution. Some baseboard heaters have a thermostat mounted directly on the unit rather than on the wall. This is not a desirable thermostat location because in many cases it will be necessary to move furniture every time a temperature change is desired. *Wall heaters* are compact units mounted on or recessed into a wall. Most heaters operate in conjunction with a fan that blows the warm air into the surrounding area. They are usually used to provide heat in nonhabitable areas such as a garage or workshop and are sometimes used to supplement a central heating system. These heaters do not produce a uniform heat distribution but are very responsive to a call for heat. They are usually manually controlled rather than automatically controlled.

Heating system inspection procedure

A full heating system inspection consists of an evaluation of the operation of the boiler or furnace, the burner, the condition of the distribution system (wherever visible), and heat outlets—radiators or registers. During your interior inspection, each room should be checked to determine whether there is a heat outlet and if there is, whether it is properly located for maximum effectiveness. The area below the radiators should be checked for signs of leakage, and the dampers in registers should be checked for ease of operation. Distribution piping or ducts, which are often visible in the basement, crawl space, attic, and garage, should be checked for leaky joints and the need for insulation. Also, the thermostats should be checked for location (should not be in a draft), condition, and type. An automatic clock-type thermostat is more convenient and if used properly, will result in fuel savings. A broken thermostat must be replaced, and one that is loosely mounted must be resecured.

The boiler or furnace and the associated burner are inspected after completing the interior inspection. By the time you are ready to perform this inspection, you should know whether the house is heated by warm air, hot water, or steam. You can tell by the type of heat outlets—registers for warm air or radiators for hot water and steam. A radiator with one pipe attached is a steam radiator. When there are two pipes, it might be difficult to tell whether the radiator is used for hot water or steam. In this case, you should look at the boiler.

At first glance, two or more thermostats can lead to a false conclusion—that the house has a multizoned heating system. A house with a multizoned heating system will have more than one thermostat. However, a house with more than one thermostat need not have a multizoned heating system. The number of heating zones should be verified when you inspect the boiler or furnace. In a hot-water system, the zones are controlled by electrically operated valves, circulating pumps, or a combination of the two. For a warm-air system, the zones are controlled by electrically operated damper motors. Usually the number of thermostats indicates the number of independent heating zones. However, you might find more thermostats than zone controllers. I have inspected homes that had a one-zone heating system that was activated by two separate thermostats. When either thermostat called for heat, the entire house would be heated. Also, in older homes you might find an old nonfunctioning thermostat on a wall.

When you are ready to inspect the boiler or furnace, stand where you can see the burner, about 3 feet away. Have someone turn up the thermostat and activate the burner. It is important to be near the unit when the burner ignites to determine if there is a problem condition. A puffback with an oil burner, or flames licking back under the cover plate with a gas burner, is an abnormal condition and a potential hazard that must be corrected. Record it on your worksheet. If the unit does not fire, check the master switch. If the switch is on and there is an oil burner, push the *reset* button once. If the burner does not fire, there is a problem condition that must be corrected. If the gas burner does not fire and the pilot is lit, there is also a problem.

Once the unit is firing, check the overall condition. Is the boiler or furnace an aging unit? Do the burner and boiler or furnace appear to have been neglected? Are there signs of excessive corrosion (rust), dust, and flakes? Are there mineral deposits, indicating a past or current water leakage? If there is more than one heating zone, check each zone independently. This can be done by waiting until the burner or circulating pump (if a forced-hot-water system) shuts down and then activating the thermostat that controls the zone being checked. The thermostat should activate the burner or circulating pump, depending on the system (as described in chapter 14). Sometimes there is a time delay of about one minute or so between engaging the thermostat and activating the burner or circulating pump. This occurs because of the time that it takes to open the zone valve or damper physically. Also, some gas valves have a built-in time delay. However, if after five minutes nothing happens, there is a problem with the controls.

Check the condition of the smoke pipe. This is the horizontal section of sheet-metal pipe that connects the boiler or furnace to the chimney. In some cases, the boiler or furnace is connected directly to a prefabricated metal chimney, and there is no horizontal run. When there is, however, the pipe must have a slight upward pitch from the boiler or furnace to the chimney. If the pipe is long, it should be supported to prevent sagging sections. The pipe should not have corrosion holes, and the joints between sections should be tight. This pipe gets very hot and should not be within several inches of combustible material.

In recent years, *direct-vent appliances* (appliances that vent exhaust gas through a sidewall) have become more common. During your exterior inspection, if you see a direct-vent terminal below a window, check the distance between the top of the vent and the bottom of the window. If the terminal is too close and the window is open, flue gases could enter the house. The recommended distance for an appliance with an input of 50,000 Btu per hour is a minimum of 9 inches, and for an appliance with an input of over 50,000 Btu per hour, the vent terminal clearance is at least 12 inches. Also the bottom of the vent terminal should be at least 12 inches above the ground.

The various types of heating systems have specific items that should be checked during an inspection. Each is discussed below. At the conclusion of the heating system inspection, remember to turn the thermostat back to its original setting.

Of particular concern is whether there is an adequate supply of air for ventilation and combustion for the heating equipment. When the heating system is located in an unconfined space such as an unfinished basement, there is generally sufficient infiltration so that the air supply is usually not a problem. However, if the system is located in a confined space such as a small furnace room, provisions must be made for an adequate air supply.

To provide the needed air supply, openings must be made within 12 inches of the top and 12 inches of the bottom in one of the walls of the enclosure. Each opening should have a *net free area* of at least 1 square inch per 1,000 Btu of total input rating of the equipment in the enclosed space. The size of the opening can be reduced if the air supplied to the furnace room comes from the outdoors rather than from inside the building. If the adequacy of the air supply is a concern, you should have it checked by a professional.

Warm-air systems

This system will be either gravity or forced. You can determine which by whether there is a fan. In some of the old "octopus"-type furnaces, the fan might be difficult to locate. It might be on top of the unit 5½ feet off the ground, or there might be fans in return ducts.

After the burner has fired, wait until the fan begins to operate. It should begin before the burner is shut down by the high-temperature limit control. If it begins after the burners are shut off, the fan controller is either faulty or out of adjustment. Also, if there is a canvas connection between the furnace and the main supply duct, check its condition. Torn or open sections should be patched. The canvas connection is used to isolate the supply ducts from the vibrations and sounds that develop within the furnace.

Be advised that with some older systems, the fabric connector may contain asbestos. Since warm air is constantly flowing across the fabric when the heating system is operating, the presence of asbestos would be considered a potential health hazard. See "Asbestos" in chapter 20.

When the fan is operating, is there excessive noise or vibration? There should be none. If there is a power humidifier, it is usually wired so that it will operate only when the fan is running. Turn the humidifier on by turning the humidity control up. If there are portholes in the unit, see if the fan or drum is rotating. If there are no portholes, listen to hear if the motor goes on and shuts off when the humidity control is lowered. Very often the humidifier is in need of a tune-up. This is usually indicated by excessive mineral deposits around the humidifier and the duct to which it is attached.

When the system is operational, check the flow and temperature of the air discharging from the various heat-supply registers. Put your hand in front of the register. A weak flow can be the result of air leakage in the distribution ducts, a dirty air filter, a loose fan belt, or the need for rebalancing. When the temperature of the discharging air in one room is lower than in other rooms, it is probably caused by a heat loss in the branch duct leading to that room. The return register can be checked by placing a tissue on the grille. Because the return duct draws air in, the tissue should stay in place without falling down.

The item of main concern is the heat exchanger. When this portion of the heating system breaks down, the furnace or its heat exchanger must be replaced. When the furnace is aging or excessive corrosion is noted, the condition of the heat exchanger should be suspect. Also, if the flames in a gas-fired furnace are unstable and appear to be dancing, this often indicates a crack in the heat exchanger. It is recommended that prior to the contract closing, the furnace and associated equipment be tuned by a competent service organization. At that time, the serviceman can check the condition of the heat exchanger. Checking the heat exchanger usually requires the use of sprays, propane torches, and mirrors or disassembling a portion of the furnace and should be performed by a professional.

If the furnace is the high-efficiency condensing type, check the unit and associated piping for corrosion. Corrosion is a major concern in these furnaces because of the acidic nature of the condensate. Sulfur in the fuel oil or natural gas can combine with the flue-gas condensate to form sulfuric acid. Also, when indoor air is used for combustion, if the air contains chlorides, the chlorides can combine with the water vapor in the flue gas to form a corrosive condensate containing hydrochloric acid. Chlorides in the air can come from laundry bleach, household cleaners, even chlorinated tap water. To counter these problems, most condensing furnaces use outdoor air for combustion. They also generally use a plastic pipe for the exhaust stack and condensate drain.

Check the condensate drainpipe for blockage and loose sections. The condensate pipe should discharge into a floor drain that is connected to the house sewer. If the floor drain is not lower than the condensate drain or there is no floor drain, a condensate pump should be used. It is not good practice to have the condensate discharge into a hole in the floor slab.

The air supply should also be checked for blockage. If the supply pipe is run through the roof, the exposed end should be shaped like an inverted U with the open end facing down. This keeps the rain out, and since the supply and exhaust pipes are generally near each other, it prevents exhaust gases from being drawn into the supply pipe.

With a condensing furnace, the exhaust vent should not be combined with the exhaust from another appliance such as a water heater. The exhaust gases can be vented through a side wall, the roof, or a masonry chimney. However, if connected to a chimney, the plastic pipe should run up to the top of the chimney where the exhaust gas can vent to the atmosphere. If the exhaust pipe terminates at the base of the chimney, the corrosive condensate that forms will attack the masonry.

If the exhaust pipe has horizontal sections, see if any low spots or sagging sections could trap condensate and thereby block the exhaust gas from venting. Horizontal sections should have a slight pitch back to the furnace. This is especially important if the section is near the termination by a side wall or attic. If the exhaust pipe is sloped away from the furnace, the condensing water vapor in the exhaust gas could freeze and eventually block the opening in the pipe.

Condensing furnaces are noisier than conventional furnaces. They have a higher airflow, which is needed to absorb the additional heat resulting from the added heat exchanger. The higher airflow produces more sound and vibration. There is also a draft-inducing blower that is needed to overcome the increased resistance to the flowing exhaust gases. Although the operation of a condensing furnace is noisier than that of a conventional furnace, it is generally not considered annoying. If you find it bothersome, record that fact on your worksheet. The system is in need of a tune-up and adjustment by a competent heating contractor.

Condensation of water vapor in the exhaust gases also occurs in a pulse-combustion furnace. Consequently, the overall concerns of exhaust-gas venting and condensate removal, as discussed previously for condensing furnaces, apply. Because of the nature of the combustion process in a pulse furnace, the sound of sixty to seventy tiny explosions per second can be heard outdoors at the side-wall terminations of the air-supply and exhaust-vent piping. In those cases, where the sound is considered objectionable, mufflers can be installed on the supply and exhaust pipes to reduce the sound level.

Hot-water systems

When inspecting a hot-water heating system, you should first determine whether it is a gravity or forced system. Sometimes an open expansion tank is found in the attic during the interior inspection. This would normally indicate a gravity system. Often, however, the expansion tank is no longer functional and was replaced by a closed expansion tank when the system was converted to forced circulation. You can easily tell whether a system has forced or gravity circulation by checking the equipment associated with the boiler. A forced system will have a circulating pump in the return line near the boiler. Record the type of system on your worksheet.

Next, look at the pressure gauge and record the operating pressure. The normal operating pressure should be between 12 and 22 psi. Since the water-makeup valve (pressure-reducing valve) is set to introduce water into the system when the pressure drops below 12 psi, any pressure below that value indicates either a malfunctioning valve or the need for adjustment. A low-pressure reading is fairly common.

If the pressure gauge indicates an operating pressure of 30 psi, the relief valve should be discharging. Sometimes the relief valve has already discharged and lowered the pressure to about 28 psi. In this case, the exposed end of the relief valve will have water dripping from it, and there will be a pool of water on the floor. A high-pressure condition that results in the relief valve discharging will usually be caused by a waterlogged or undersized expansion tank. If domestic hot water is generated through the heating system, another possible cause for a high-pressure condition is pinholes in the coils of the water heater. Regardless of the cause of the high pressure, the condition must be corrected.

Locate the relief valve. From a safety point of view, it should be boiler-mounted. If there is no boiler-mounted relief valve (and often there is not), the installation of one should be considered. Do not check the valve to determine whether it is operational. Bits of corroded material, sediment, or mineral deposits may prevent the valve from reseating properly and shutting off. Your best bet is to check the valve after you move into the house. At that time, if it does not reseat properly, it should be replaced. A relief valve is inexpensive. These valves should be checked at least once a year by pulling the lever and allowing a small quantity of water to flow. This action flushes any sediment that tends to build up before it has a chance to clog the valve. Some relief valves do not have levers for testing the unit. If you have such a relief valve on your boiler, you should consider replacing it.

As the various zones are being checked, inspect the zone valves for dripping water and deposits. Even though the valves might be working during your inspection, if water is dripping from the valves, they should be replaced. The dripping water results in a mineral-deposit buildup that can eventually cause the valve to "freeze" in an open or closed position. This might happen, and usually does, on the coldest night of the year. Zone valves are perhaps the weakest link in the heating system, and their occasional replacement should be anticipated.

Is the circulating pump operating properly? Just because you can hear the pump operating or feel its vibration does not mean that it is functioning properly. The coupler between the pump and the motor may be broken, and the noise that you hear or the vibration that you feel may be only the motor turning, not the pump. You can be sure the pump is operating by putting your hand on the return pipe associated with that pump. After a while, the pipe will get hot if the pump is operational. Check the gasket between the pump and the motor for signs of current or past leakage. This is a vulnerable joint for leakage.

Is the circulating pump emitting any unusual or loud sounds? If it is, the noises might indicate faulty bearings, and the condition should be checked by a service organization. If you see smoke coming from the circulating pump or smell electrical insulation burning, the entire system must be shut down immediately and the condition corrected. One last check of the circulating pump is to determine whether it operates continuously or intermittently. If the circulating pump continues to run long after the thermostat is not calling for heat, it is in continuous-mode operation. This mode, discussed in chapter 14, can be modified to an intermittent operation.

If the heating system is used for generating domestic hot water, the thermostat will not control the burner, only the circulating pump. In this case, check the temperature of the boiler water (look at the temperature gauge). If the temperature drops below 150° F and the burner does not fire, it will indicate a faulty or improperly set aquastat. In most cases, the burner will be activated when the temperature drops below 180° F.

In addition to the above, with high-efficiency condensing and pulse boilers, look at the piping associated with the inlet air supply, exhaust gas vent, and condensate drain. Check to see that there are no loose sections, that there is no blockage at the pipe terminations, that the exhaust pipe is not combined with the exhaust from the water heater, and that the horizontal portions of the exhaust pipe slope back to the boiler. Also check that the condensate pipe discharges into a floor drain or condensate pump rather than a hole in the floor slab.

Steam systems

In this system, a water line showing the level of water in the boiler should be visible in the level gauge. Look for the gauge. If it is dirty, it must be cleaned. A gauge that is completely filled with water indicates too much water in the system, and an empty gauge indicates an insufficient amount of water in the boiler. When the water level in the gauge is unsteady (rising and dropping), it indicates a problem condition that might reflect an excessive grease and dirt buildup in the boiler or that the boiler is operating at an excessive output. In either case, it should be corrected.

Check the heating unit to determine whether it has all the required safety controls. A steam system should have a low-water cutoff, a high-pressure limit switch, and a relief valve. If any of these items are missing, record the fact on your worksheet. Look for the low-water cutoff. Is it the built-in type or externally mounted? If it is externally mounted, you should test its operation. This is a normal procedure that the manufacturer recommends the homeowner perform periodically. If the relief valve is old, then as a precautionary measure it should be replaced after you take possession of the house.

Look for the condensate return line to the boiler. Is it a wet or dry return? If it is a wet return, it should be connected to the boiler by means of a special piping arrangement, a *Hartford loop*. Is there one? If not, the installation of a loop should be considered.

Oil burners

When the heating system is oil-fired, the first thing you should note is whether the top portion of the chimney (as seen during your exterior inspection) is covered with soot. This is an indication that the oil burner has not been operating efficiently. When you are inspecting the heating unit, look for a card or tag near or on the unit that contains the maintenance service record. Whenever a serviceperson performs any type of maintenance, even a tune-up, he or she records the date on this card. This will at least give you an idea of whether the burner has been maintained or neglected. The burner should be tuned up and checked for combustion efficiency at least once a year.

Other telltale signs indicate the need for oil-burner maintenance. Is there a smoke odor in the furnace room? When the oil burner first fired, was there a puffback of smoke and fire? At startup, while running, or at shutdown, do you feel pulsations in the boiler room? Are there particles of soot on the furnace or around the burner? Is the oil burner emitting an abnormal noise during operation? If any of the above questions are answered yes, oil-burner maintenance is needed. There are also tests that can be made, such as sampling the exhaust gas for combustion efficiency. This analysis, however, should be performed by a competent serviceperson.

The proper draft over the firebox for an oil-fired system is critical to efficient operation. Therefore, there must not be any cracks or open joints that will allow excess air to infiltrate into the combustion chamber or smoke pipe. Look for cracks and open joints at the front of the boiler near the burner, especially around the air tube, mounting plate, and front boiler doors. If you can see the flame through the cracks or open joints, the open areas should be sealed with a refractory cement. A vulnerable joint for air leakage is the joint between the smoke pipe and the chimney. If this joint is cracked, it must be sealed. Also, to ensure a proper draft, there should be a functioning draft regulator on or near the smoke pipe. Look for one. If it is missing or inoperative, record the fact on your worksheet.

If you see a small hole in the flue stack about ¼ inch in diameter, don't be concerned. It is a test hole that is used by the oil-burner serviceperson for measuring operational parameters such as draft, percentage of carbon dioxide, and stack temperature.

Next, look inside the firebox. The refractory lining of the fire chamber should be intact. There should not be any broken sections or holes in the lining. If there are, maintenance is needed. While looking inside the firebox, if the heating unit is a boiler rather than a furnace, look for water dripping or seeping around the top or sides. As discussed earlier, this condition might be caused by joint movement as a result of thermal expansion. Nevertheless, as a precautionary measure, since it can also be caused by a cracked boiler, it should be checked by a professional.

The oil burner receives its oil through a feed line that runs to the oil storage tank. The feed line is usually a small-diameter copper tube that can easily be damaged and must be protected. Look for the line. It should not run exposed over the floor. (See FIG. 15-1.) Someone could accidentally step on and damage the feed line, restricting the

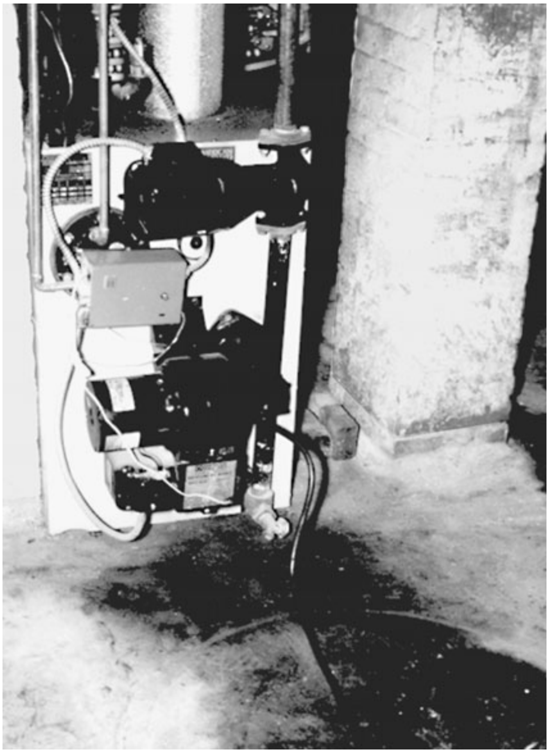

Fig. 15-1. *Oil-fired forced-hot-water boiler with an exposed oil-feed line. The feed line is a small-diameter copper tube that can easily be damaged. It should be protected. Usually the feed line is located below the floor slab and emerges near the oil burner.*

flow to the burner and causing oil to leak on the floor. If the storage tank is the interior type, the underside of the bottom section should be checked for oil leakage. Run your hand along the underside of the tank. If your hand has a slight or heavy film of oil, the unit may be deteriorating, in which case patching repairs would be in order. The average projected life for these tanks is about twenty years.

Gas burners

Gas burners do not require as much preventive maintenance as oil burners. All that is normally needed is periodic cleaning and adjustment. Look at the gas burners. In neglected units, they are covered with excessive corrosion (rust) dust. Quite often some of the gas ports are clogged by the rust particles and are no longer effective. This condition requires a professional cleaning. The quality of the flames on the gas burner is important. The flame should be primarily bluish in color, with little or no yellow. If the flame is lifting off the burner head, it is an indication that too much air is being introduced into the mixture.

A yellow tip in the flame will result in a soot layer building up in the flue passage of a boiler. After some years, the soot buildup can completely block the flue passage. I found this condition in a boiler that was only seven years old. A blocked flue passage is a potentially dangerous condition because it prevents the exhaust gas from rising up to the chimney. Instead, the gas, which contains poisonous carbon monoxide, seeps into the house. Fortunately, this condition can be easily corrected. Clean the boiler flue passages with a long stiff brush to remove the soot buildup and adjust the air intake to the burner.

Another safety check that should be made is to test the draft in the chimney. After the heating unit has been operational for a few minutes, put your hand near the draft diverter (located on the exhaust stack above a boiler and built into the sheet-metal casing of a furnace). There should be no outward flow of hot exhaust gas toward your hand. If you do feel a flow on your hand, a potentially dangerous condition exists that should be checked by a professional.

Checkpoint summary

General considerations

❍ How old is the heating system?

❍ Is the boiler or furnace an obsolete unit (i.e., a converted coal-fired boiler or an "octopus" warm-air furnace)?

❍ Check each room for heat-supply registers or radiators.

❍ Are supply registers or radiators located on or near exterior walls (preferably under the windows or next to exterior door openings)?

❍ Check supply-register dampers for ease of operation.

❍ Are there separate return grilles for each room or centrally located returns for each floor?

❍ Check radiators for broken shutoff valves and blocked air vents (steam systems).

❍ Inspect area below radiators for signs of leakage.

❍ Check visible portions of ductwork and piping (basement, attic, garage, etc.) for open or leaking joints and uninsulated sections.

❍ Are the thermostats properly located (no drafts)?

❍ Note whether the thermostats are manual or automatic clock units.

❍ Is the clock functional?

❍ Check each heating zone and thermostat control independently.

❍ Is there an adequate air supply for the heating system?

❍ If the heating unit has a direct vent, is it properly located?

Boiler or furnace

❍ Check boiler or furnace during startup. Note any puffback with an oil burner or a licking back of flames on a gas burner.

❍ Is the boiler a steel or cast-iron unit? Is it aging? Is it neglected?

❍ Check for signs of excessive corrosion (rust), dust, flaking metal, and mineral deposits indicating past or current leakage.

❍ Inspect firebox, if accessible, for water seepage through joints and dripping water. Have this condition checked by a professional.

❍ Check smoke pipe for corrosion holes, open joints, sagging sections, and adequate clearance.

❍ If furnace or boiler is the condensing type, check for corrosion, signs of condensate leakage, blockage in the supply and vent pipes, and horizontal vent pipe sections sloping back to the furnace or boiler.

❍ Check master shutoff switches for proper operation.

Warm-air systems

❍ Check furnace, burner area, and sheet-metal casing for excessive corrosion, rust, and flaking metal.

❍ Check for unstable or dancing flames (gas-fired furnace). If these conditions are noted, have the heat exchanger checked by a professional.

❍ If there is a canvas connection between furnace and main supply duct, check it for torn and open sections. This material may contain asbestos.

❍ Check fan controller during furnace operation. If fan starts after the burners shut off, controller is either faulty or out of adjustment.

❍ Check fan operation for excessive noise or vibrations.

❍ If system contains a power humidifier, check for operation, signs of leakage, and excessive mineral deposits.

❍ During furnace operation, check airflow and temperature in room supply registers.

❍ Check return grilles for proper air circulation (tissue test).

Hot-water systems

❍ Check whether system is forced or gravity circulation.

❍ During operation, check pressure gauge. System pressure should generally be between 12 and 22 psi.

❍ Inspect for relief valve, which should be boiler-mounted. Do not check valve operation.
❍ Check circulating pumps for proper operation (motor and pump both functioning). Note any unusual loud sounds.
❍ Inspect pumps for leaking fittings or gaskets.
❍ Check each zone valve independently. Inspect valves for dripping water and deposits.
❍ If boiler also generates domestic hot water, check temperature gauge to determine whether aquastat is faulty or improperly set.

Steam systems

❍ Check water-level gauge for proper level, accumulated sediment, and fluctuations (rising and dropping).
❍ Is the water supply to the boiler a manual or an automatic feed?
❍ Check boiler for required safety controls: a low-water cutoff, a high-pressure limit switch, and an automatic relief valve.
❍ If low-water cutoff is an external type, check operation by flushing the unit.
❍ Inspect boiler relief valve. *Do not operate.* If valve is old, consider replacement.
❍ Inspect the condensate return line. Is it a wet or dry return? If a wet return, is there a Hartford loop?

Oil burners

❍ During exterior inspection, check top portion of the chimney for accumulated soot.
❍ Has oil burner been maintained or neglected? Look for a dated maintenance service card.
❍ During operation, note any smoke odors or soot particles in the furnace room.
❍ Are any abnormal noises or pulsations emitted from the oil burner during operation?
❍ Check for cracks and open joints near the burner, particularly around the air tube, mounting plate, and front boiler doors.
❍ Check for air leakage at the joint between the smoke pipe and the chimney.
❍ Check exhaust stack for an operational draft control.
❍ Check refractory lining of the fire chamber for cracked and deteriorating sections.
❍ Check oil-burner feed lines (copper tubing). Are there kinked or exposed sections vulnerable to damage?
❍ Check underside of oil storage tank (interior type) for leakage.

Gas burners

❍ Check gas burners for corrosion dust and clogged gas ports.
❍ Check flame pattern for safe and efficient operation.
❍ Check draft diverter on the exhaust stack for backflow of exhaust gases.
❍ Is the gas meter adequately sized?

16

Domestic water heaters

Tank water heaters 248
  Reversed connection 248
  Relief valve 249
  Rumbling noise 251
  Exhaust stack 252
  Operational inspection 252
  Water heater replacement 254
  Water heater capacity and recovery 255
Tankless (coil) water heaters 256
  Inspection procedure 258
Tankless (wall-mounted) water heaters 259
Indirect-fired storage water heaters 260
Checkpoint summary 261

Hot water for bathing and washing (domestic hot water) is produced through a separate tank-type water heater, through the heating-system boiler (steam or hot water) by a tankless coil, or through an indirect-fired storage water heater. Domestic hot water is considered by manufacturers as cold water whose temperature has been raised 100° F. An important consideration in the design of water heaters is the ability of the unit to raise the temperature of 1 gallon of water 100° F. This criterion is used in evaluating the recovery rate for tank heaters and the water-flow rate for tankless coil heaters. These items will be discussed in the sections that follow.

The temperature of domestic hot water is important for energy conservation and safety. The operating temperature range for most water heaters is between 120° and 160° F, although in some cases it is possible to achieve temperatures of 180° F. Normally, a thermostat setting that produces a water temperature of 140° F will be adequate for household chores, including the use of automatic clotheswashers and dishwashers. However, for washing dishes by hand or bathing, temperatures of 120° F are more than adequate. In fact, 120° F water is above the tolerance level of most people and will have to be mixed with cold water. Setting the thermostat so that the water temperature exceeds 140° F is wasteful of energy and will shorten the life of the water heater. In addition, a water temperature in excess of 160° F is a potential hazard because of the possibility of being scalded while showering (if the mixing valve should be faulty).

Tank water heaters

Most tank water heaters are located near the heating system or in a utility room, sometimes in the kitchen or a closet. These heaters basically consist of a lined steel tank that is covered with an insulated sheet-metal jacket to reduce heat loss. The lining in the tank is usually vitreous enamel (glass) but can be concrete (stone) or copper. To minimize tank corrosion, most heaters are equipped with a "sacrificial" magnesium anode rod that is suspended inside the tank. The electrochemical action causing corrosion takes place between the water and the anode rather than between the water and the tank, thereby extending the life of the tank. Some tanks are constructed so that the magnesium anode can be inspected and replaced if necessary. Tank water heaters are gas- or oil-fired or electric. (See FIG. 16-1.)

Reversed connection

Many water heaters have at the top of the tank hot-water outlet and cold-water inlet fittings that are clearly marked HOT and COLD, respectively. Yet it is surprising how many times I have found heaters with a reversed water-line installation. You can tell whether the installation is reversed by touching a pipe about 4 feet away from the unit and comparing it with the HOT outlet fitting on top of the tank. A reversed connection results in an inefficient operation of the heater. Even though there is a cold-water inlet tap at the top, the cold water does not enter the tank at the top. There is a pipe (dip tube) inside the tank connected to the inlet fitting that carries the cold water to the bottom of the tank. Because the hot water tends to rise above the cold water, the hot water is normally at the top of the tank where it flows through the outlet fitting. When the inlet-outlet

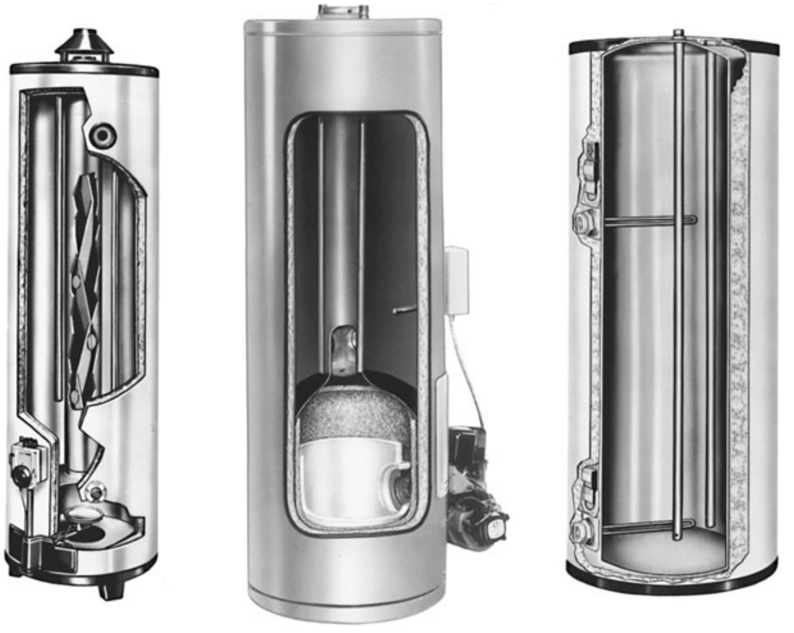

Fig. 16-1. *Tank water heaters. Left—gas-fired. Center—oil-fired. Right—electric. Rheem Mfg. Corp.*

connection is reversed, the cold water enters the tank at the top and mixes with the hot water as it settles to the bottom near the outlet fitting. Consequently, for the same thermostat setting, the temperature of the hot water is lower than it would otherwise have been. To compensate for the lower temperature, homeowners usually turn the thermostat to the high setting, thereby increasing the fuel cost and decreasing the life of the water heater. Some heaters have interchangeable cold- and hot-water fittings. In these units, the dip tube can be shifted from one side to the other.

Relief valve

All water heaters must have a relief valve that is both temperature- and pressure-sensitive. A relief valve that responds to pressure only will not offer adequate protection for the homeowner. An excessively high pressure alone can cause the tank to rupture. Though this will cause water damage, it is not necessarily dangerous. However, if the tank should rupture as a result of a high-temperature failure, the superheated water contained therein will flash into steam, instantaneously releasing high energy in an explosion. A relief valve that is sensitive to both temperature and pressure can avoid this problem.

The relief valve should have an extension that will allow it to discharge into a bucket on the floor. Sometimes, a relief valve is found without this extension. (See FIG. 16-2.) This is a potential hazard because a person standing near the water heater when the relief valve discharges could be scalded. I have also found an extension on the relief valve that leads to a sink or floor drain located near the water heater or in the next room. (See FIG. 16-3.) This type of piping arrangement is usually installed to eliminate water damage if the relief valve discharges. Water discharging from the valve will flow directly down the drain rather than

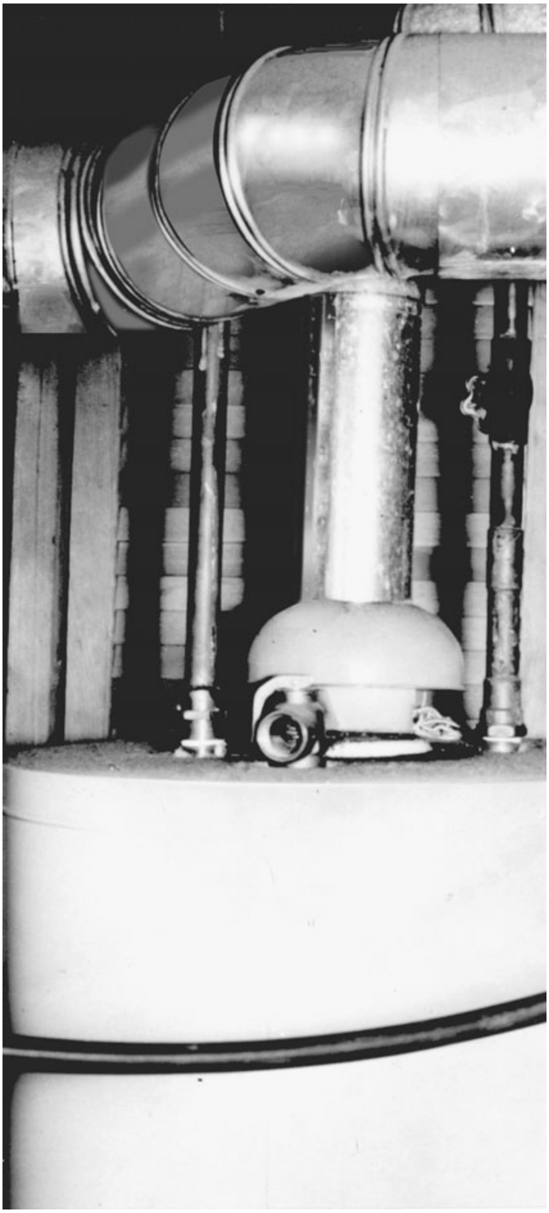

Fig. 16-2. *If this relief valve discharges, someone standing nearby could be scalded. The relief valve should have an extension that discharges into a bucket located on the floor.*

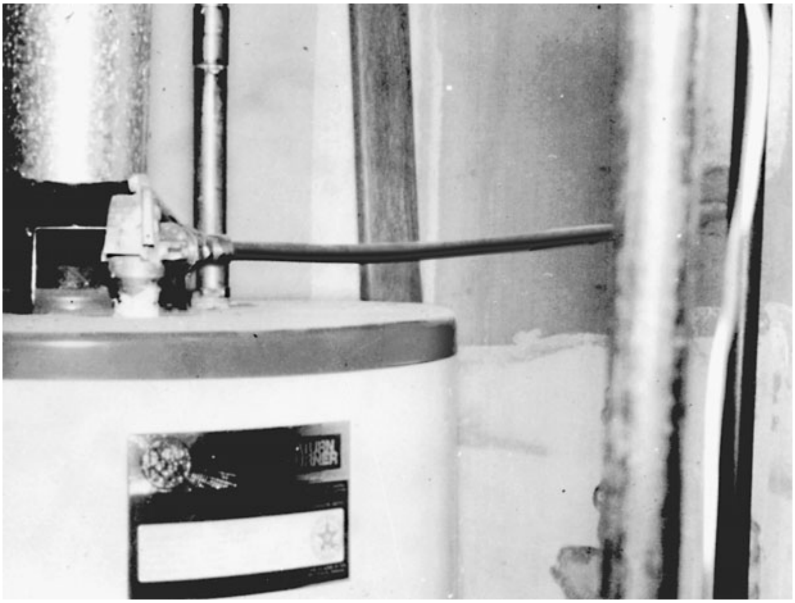

Fig. 16-3. The relief valve has an extension that leads to a sink in the adjacent room. This is undesirable because the effluent will go down the drain, and the homeowner will not be aware of a discharge condition.

accumulate on the floor. The disadvantage of this installation is that the homeowner might never know a discharge condition had occurred. A discharging relief valve indicates an abnormal condition of some type that should be checked out. Obviously, if the homeowner does not know that there is a problem, he or she will not have it corrected. By having the relief valve discharge into a bucket, water damage can be eliminated, and the presence of a problem condition will be indicated.

If you notice that the relief valve is dripping and the temperature of the hot water is not excessive, the problem will be either a defective relief valve or the result of thermal expansion. Many municipalities require a backflow preventer or a nonbypass pressure regulator on the inlet water service. This creates a closed system when a faucet is not drawing water. When the water is heated and expands, the resulting pressure buildup causes the relief valve to drip. Installing an expansion tank on the cold-water inlet to the water heater can control this condition. (See FIG. 16-4.) Record a dripping relief valve on your worksheet for later correction.

It's also important that the Btu capacity of the relief valve exceed the Btu input of the water heater. If the capacity is less, the relief valve will not be able to discharge the overheated water at the same rate that it's produced, resulting in an unsafe condition or possible explosion. Check the capacity of the relief valve by looking at the nameplate mounted on the valve. Specifically look for the AGA (American Gas Association) rating given in Btu/hour (see FIG. 16-5). This rating must exceed the Btu input to the water heater listed on the data plate mounted on the side or top of the tank casing. You can also check the Btu rating of electric water heaters even though they are rated in watts rather than Btu/hour. You can determine the electrical energy input to the heater by converting the total wattage of the heater, which is listed on the data plate, to Btu/hour. 1 watt = 3.413 Btu/hour.

To be effective, the relief valve should be located so that its sensing element is immersed

Fig. 16-4. *Bladder-type expansion tank installed to prevent relief valve from leaking due to thermal expansion of the water in the domestic water heater tank.*

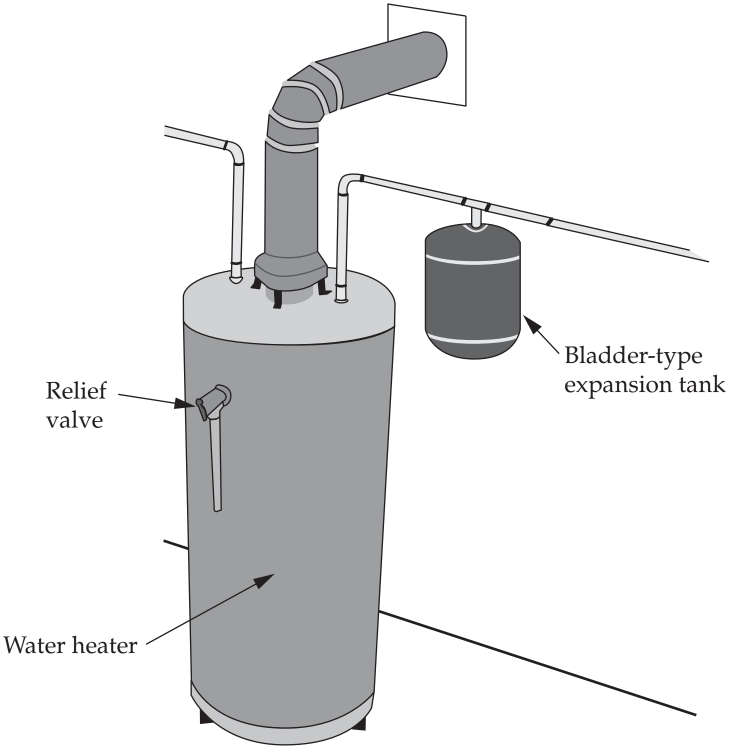

in the top 6 inches of water in the tank rather than in the hot-water outlet pipe. This is important because there is a temperature difference between the tank and the hot-water outlet pipe. For instance, when the relief valve is installed in the hot-water outlet pipe 5 inches away from the tank, the water in the tank could be raised to 250° F before the temperature at the valve reaches 210° F.

During my inspections, I have often found the relief valve located on the cold-water inlet pipe. This location negates the effectiveness of the valve. If you find a relief valve on the cold-water inlet pipe, you should notify the homeowner of the potential hazard and have it relocated after you take possession of the house.

It is important that the diameter of the relief-valve extension pipe be the same as the outlet opening of the relief valve. If the extension pipe has a smaller diameter, it reduces the discharge capability of the valve and reduces its effectiveness as a safety device.

Rumbling noise

Over the years, sediment scale and mineral deposits tend to build up at the base of the tank. Manufacturers suggest that a few quarts of water be periodically drained from the water heater to help remove these deposits. However, this practice is not always effective, and if sufficient deposits accumulate, a rumbling or pounding sound can be heard when the unit is firing. If you hear a rumbling noise while inspecting the water heater, do not be alarmed. It is not a dangerous condition, although the noise can be annoying. In addition to their annoyance value, accumulated deposits at the

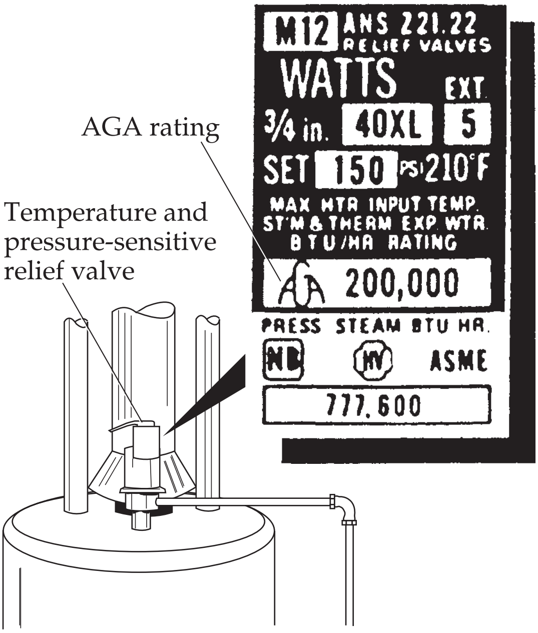
AGA rating
Temperature and pressure-sensitive relief valve
M12
ANS 221.22
RELIEF VALVES
WATTS
EXT
3/4 in.
40XL
5
SET
150
PSI
210°F
MAX HTR INPUT TEMP.
ST'M & THERM EXP. WTR.
BTU/HR RATING
AGA
200,000
PRESS STEAM BTU HR.
NB
NV
ASME
777,600

Fig. 16-5. *Look for the AGA rating on the plate hanging from the relief valve. The rating must exceed the Btu input to the water heater.*

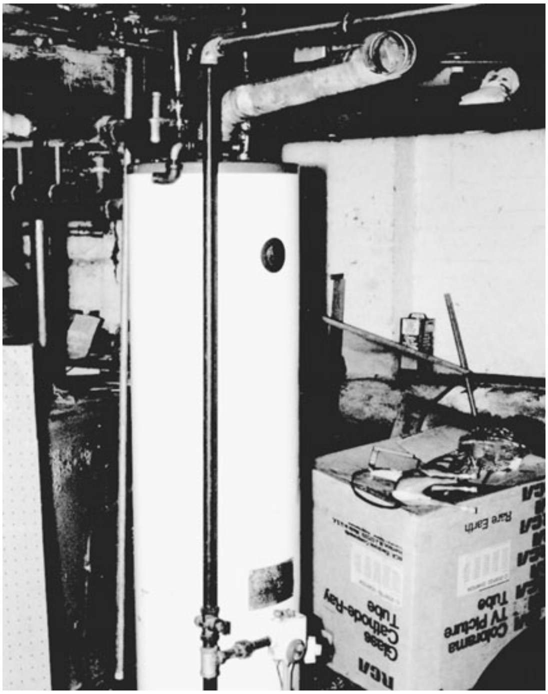

Fig. 16-6. *Disconnected exhaust stack. Exhaust gases from this gas-fired water heater discharge into the basement rather than a chimney—a potentially dangerous condition.*

base of the tank can act as an insulator between the water and the flame, decreasing the overall efficiency of operation.

Exhaust stack

All oil- and gas-fired water heaters must have an exhaust stack to vent the products of combustion to the outside. If you find an exhaust stack that is loose or broken, so that the exhaust gases are discharging directly into the house (see FIG. 16-6), you should notify the homeowner of this potentially dangerous condition. As with a heating system, the exhaust pipe should have an upward pitch from the heater to the chimney connection. The exhaust pipe gets quite hot and therefore should not be in contact with combustible material.

Operational inspection

The water heater should be inspected while the burner is operational. This can be achieved by opening the hot-water faucet at a nearby sink. Within a few minutes, depending on the size of the tank and the water flow at the sink, the burner should fire. (This assumes that the pilot is lit for a gas burner or that the electronic ignition is operational for an oil burner, and that there is oil in the storage tank.) If the burner fails to fire, there is a problem condition that should be checked. Oil and gas burners and their controls are discussed in detail in chapter 15.

To provide an adequate draft for combustion gases, oil-fired units should have a draft regulator, and gas-fired units should have a draft diverter. (See FIG. 16-7.) After the unit has been firing for several minutes, check the draft in a gas-fired unit. Put your hand near the draft diverter. If you feel a flow of hot gases on your hand, there is either a blockage in

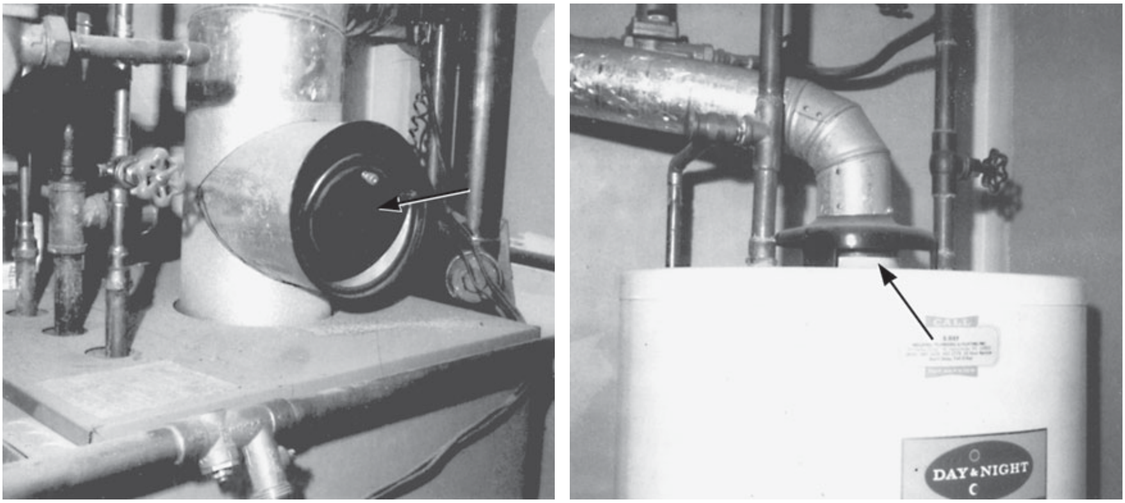

Fig. 16-7. *Draft regulator (left) for an oil-fired water heater; draft diverter (right) for a gas-fired water heater.*

the flue passage or there is a negative pressure in the house. This is a condition that must be corrected because the exhaust gas contains carbon monoxide. This colorless, odorless gas is poisonous and can cause asphyxiation.

Another check for faulty venting of the exhaust gases is to hold a lit match at the draft diverter opening. If the exhaust gases spill out into the room, they will blow out the flame. If the venting system is functioning properly, the flame will be sucked into the draft diverter hood. As a safety feature, some water heaters have a spill switch sensor mounted on the draft diverter. (See FIG. 16-8.) When combustion gases spill out of the draft diverter, the spill switches sense the heat from the exhaust gases. This causes the spillage sensor circuit to open, which shuts the pilot, causing the burner to shut down.

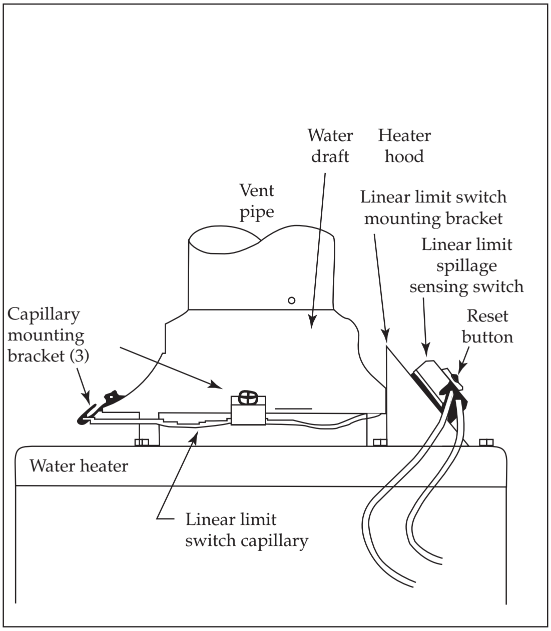
Water draft
Heater hood
Vent pipe
Linear limit switch mounting bracket
Linear limit spillage sensing switch
Capillary mounting bracket (3)
Reset button
Water heater
Linear limit switch capillary

Fig. 16-8. *Exhaust gas spill switch sensor. Courtesy of: Tjernlund Product, Inc.*

Whenever there is a heavy draw of water, if the temperature of the inlet water is low, condensation can form on the tank. If this should occur, you might see some water beneath the heater; in a gas-fired unit, you may hear sizzling or pinging sounds caused by water droplets falling on the burner. Consequently, if you see a little water on the floor or hear the pinging noise of droplets, do not assume that the water heater is leaking; it may be condensation. If it is, the condition will disappear after the water in the heater becomes heated. If the condition does not disappear, a leak should be suspected, and the unit may require replacement.

If you notice a water pipe connected to the drain fitting at the bottom of the tank, it is probably a hot-water return line. (See FIG. 16-9.) This is the typical connection for providing "instant" hot water at the bathroom fixture that is farthest from the water heater. It is more often found in large pricey houses than in more modest homes. You can check out the effectiveness of the instant hot-water system during the interior inspection of the bathrooms.

As with a gas-fired heating system, the pipe supplying gas to a domestic water heater should be rigid black iron and not a copper pipe or a flexible pipe. If the pipe is not black iron, check with the utility company for its requirements.

Some gas-fired water heaters use bottled gas (propane) because natural gas is not available. In that case, a copper gas-supply pipe is usually acceptable because the supplier provides a quality control over the impurities in the gas. All impurities that would react with the copper are removed.

Water heater replacement

A water heater requires replacement only when the tank is leaking. If there are problems with any of the controls, they can be replaced individually as needed. You cannot tell by looking at a unit when it will start leaking. The projected life for an oil- or gas-fired water heater is from seven to ten years, although units have been known to last over fifteen years. Electrical units have a longer projected life than oil- and

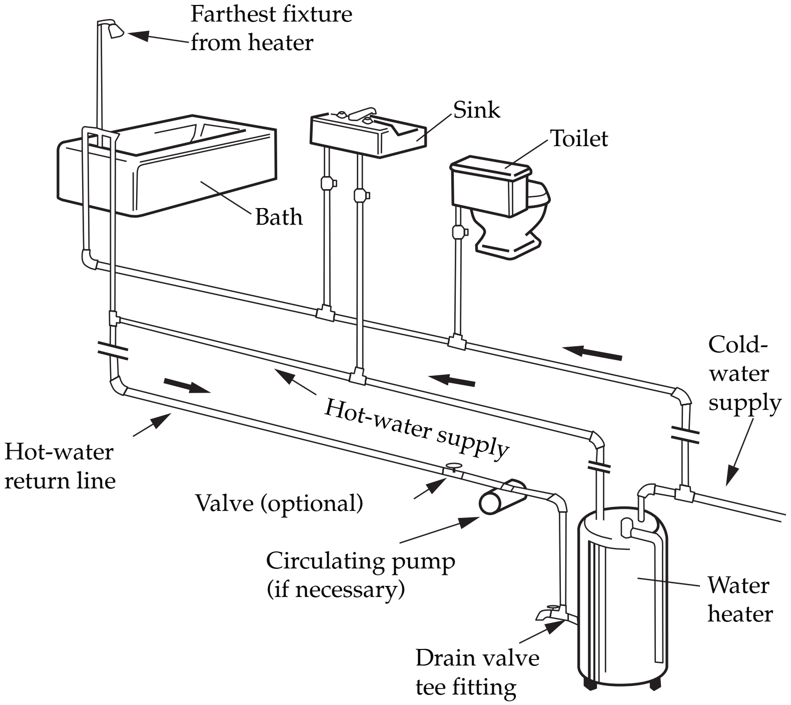

Fig. 16-9. *Hot-water return line supplies "instant" hot water to a fixture. This results from the constant recirculation of the water back to the water heater.*

gas-fired heaters. This is because the electrical heating element is immersed in the water, imparting its heat directly to the water and not to the tank, whereas in an oil- or gas-fired heater the flame imparts its heat directly to the base of the tank. The intense heat results in expansion and contraction stresses that effectively shorten the life of the tank. Look at the temperature control knob (thermostat) on the tank. If it is at the maximum setting, the unit has been operating at a high temperature, and a shorter life span should be anticipated.

The age of a tank water heater can generally be determined from the serial number found on the tank's data plate. The numbering system, however, varies by manufacturer. For example, for Rheem or Rudd water heaters, the first four numbers of the serial number indicate the month and year of production (thus 0286 would indicate that the heater was manufactured in February 1986). Usually a water heater is installed within several months of its manufacture.

With newer A. O. Smith water heaters, the date of manufacture is listed on the data plate. However, with older A. O. Smith water heaters, the date can be found by checking the serial number. For example, in serial number 800-A-89-12345, the letter stands for the month of manufacture (January in this case); and the middle number (89) stands for the year of manufacture, 1989. The letter code for month of manufacture is

A—January
B—February
C—March
D—April
E—May
F—June
G—July
H—August
J—September
K—October
L—November
M—December

If you determine that the water heater is over ten years old, you should not anticipate an extended life for the unit, even though it might look relatively new (shiny and clean on the outside).

I have been asked by clients whether it is necessary to replace a gas-fired water heater that has been in a flooded basement. It is not necessary to replace the unit. The tank is designed to hold water and therefore cannot be damaged by it. However, the burner assembly and controls (pilot, thermocouple, and thermostat) can be damaged and should be checked by a competent serviceman. If any of the controls are malfunctioning, they should be replaced as needed. Also, when the insulation between the tank and the jacket gets wet to the extent that it becomes thoroughly soaked, it can sag and be less effective as an insulator. If this should occur, you will feel "hot spots" on the jacket. This condition can easily be remedied. Insulation jackets are available that fit over water heaters to reduce heat loss.

Water heater capacity and recovery

The amount of hot water that your family requires depends on the number of people, their bathing and washing habits, and the number of tubs and showers available for simultaneous use. Do not judge whether a water heater is adequate by the capacity of the tank alone. Equally as important is its recovery rate—the volume of water that will have its temperature raised by 100° F in one hour. The capacity of residential water heaters typically varies between 30 and 82 gallons. The following example, although somewhat exaggerated, illustrates the point.

Let us take an average home. Assuming normal water pressure and flow, the typical faucet delivers about 5 gallons of water per minute. Let us further assume that there is approximately a sixty-minute demand for hot water flowing at the rate of 2 gallons per minute. This hot water will usually be mixed with cold water for washing and bathing. If the home is equipped with an electric water heater that has a 60-gallon capacity and a 10-gallon recovery rate, after about thirty minutes there would be no more hot water. The low recovery rate of 0.17 gallons per minute is less than 10 percent of that required and would be totally ineffective in satisfying the hot-water demand. On the other hand, if the home had been equipped with an oil-fired water heater that had a capacity of only 30 gallons (half the capacity of the electric heater) and a 120-gallon recovery rate, the hot water would never run out, and the demand would be completely satisfied.

What is the capacity and the recovery rate of the water heater that you are inspecting? The capacity of all water heaters will usually be stamped on the data plate. Gas-fired units might also have the recovery rate on the plate. If yours doesn't, you can get an approximate value by dividing the number of input Btu (on the data plate) by 1,000 and multiplying the results by 0.85. Oil-fired water heaters might not indicate their recovery rate. This should be of little concern because oil-fired units have a high recovery rate, often on the order of 120 gallons per hour. Electric units will not have their recovery rate on the data plate. However, the wattage of the upper heating element, the lower heating element, and the total wattage will usually be stamped on the plate. You can easily compute the approximate recovery rate for an electric water heater by using the following formula: Each 250 watts heats about 1 gallon of water 100° F in one hour. Thus a water heater rated at 4,500 watts has a recovery rate of only 18 gallons per hour. This is a typical recovery rate for an electric water heater and is quite low in comparison to a gas- or oil-fired unit. To compensate for the low recovery rate, electric units should, and often do, have large-capacity tanks.

I have found the following rule of thumb effective in determining the adequacy of a water heater for a house, regardless of whether the unit is gas- or oil-fired or electric: For a house with one full bathroom (sink, bowl, and shower-tub), the sum of the capacity plus the recovery rate should total about 70. For two full bathrooms, the total should be about 90; for three full bathrooms, about 105; and for four full bathrooms, about 115. This also assumes that there is a washing machine in the house that might be operating while all the showers-tubs are being used. Of course, your requirements might be different. If you buy a house with three full bathrooms and there are only two people in your family, you obviously do not need a water heater whose capacity and recovery rate total 105.

From an energy-conservation point of view, a small tank with a high recovery rate would be better than a large tank with a low recovery rate. In addition, it is more costly to maintain a large tank of hot water, especially during periods when there is no hot-water demand.

Tankless (coil) water heaters

Tankless water heaters, sometimes called "instantaneous water heaters," should not be confused with tankless (wall-mounted) water heaters, which will be discussed in the next section. Tankless coil-type water heaters are used in conjunction with steam and hot-water boilers. Most tankless heaters consist of a small-diameter pipe shaped in the form of a coil that is located inside of the boiler (internal generator). (See FIG. 16-10.) The coil might also be located in a casing outside, but connected to, the boiler (external generator). The coil is surrounded by hot boiler water that gives up its heat to the water flowing in the coil. Cold water enters the coil and leaves as hot water with its temperature increased by 100° F. These units are designed for a specific

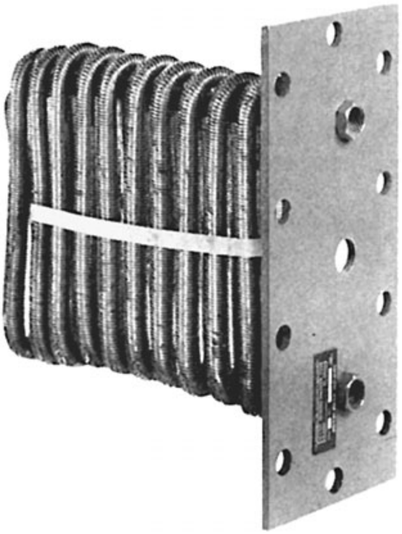

Fig. 16-10. *Tankless water heater. Unit is mounted inside a boiler. Triangle Tube and Specialty Co., Inc.*

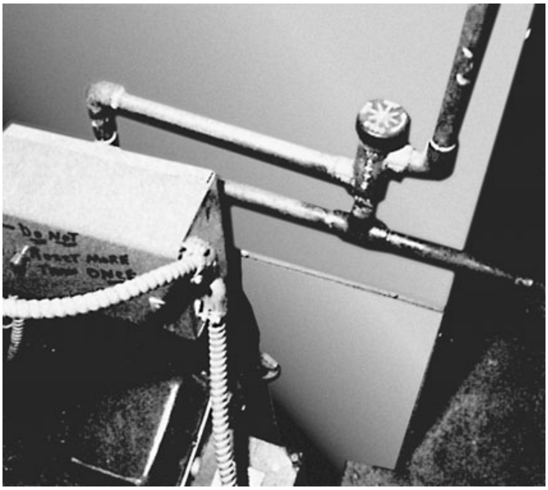

Fig. 16-11. *Mixing valve for tankless water heater.*

water-flow rate (gallons/minute) through the coil to achieve the desired increase in temperature.

If water flows through the coil at a greater rate, less heat will be transferred to each gallon, resulting in a lower outlet temperature. Consequently, many units have a flow-regulating valve installed in the cold-water supply pipe to the tankless heater to limit the water flow to the capacity of the heater. The design flow through these heaters is usually about 3 to 4 gallons per minute. Assuming a normal water supply to the house, the typical cold-water flow will be between 4 to 8 gallons per minute.

To regulate the temperature of the hot water being distributed throughout the house, there should be a mixing valve between the cold-water inlet and the hot-water outlet pipe. (See FIG. 16-11.) A mixing valve is considered a necessary feature to eliminate the possibility of scalding. Over the years, mineral deposits tend to build up inside the heater coil, especially in hard-water areas. This in turn further reduces the water flow through the heater and results in a higher water temperature at the outlet. The high-temperature water, which is a potential hazard, can be cooled to the design temperature by mixing it with cold water in the mixing valve. This water can then be distributed to the plumbing fixtures throughout the house.

As mineral deposits form inside the heater coil, they restrict the water flow. It is possible for the deposits to build up to such a point that the hot water flow becomes a trickle when two or more faucets are turned on at the same time. The hot-water flow should be checked during your interior inspection, as discussed in chapter 10. A low hot-water flow that is caused by a blockage in the tankless coil can often be corrected by chemically flushing the unit and dissolving the deposits. However, when the flow becomes a trickle, the heater coil should be replaced.

A tankless (coil) has virtually no storage capacity. Consequently, if there is a hot-water demand that exceeds the design flow (and it often does), there will not be enough hot water. To provide additional hot water, some installations are equipped with a storage tank. Water heated in the tankless coil will be circulated to the tank by gravity flow or a pump. The pipe that distributes hot water to the plumbing fixtures will be connected to the storage tank rather than the tankless heater. When the hot-water flow between the tankless heater and the storage tank is by gravity, the storage tank must be located above the boiler. When the flow is induced by a circulating pump, the tank can be located alongside the boiler. (See FIG. 16-12.)

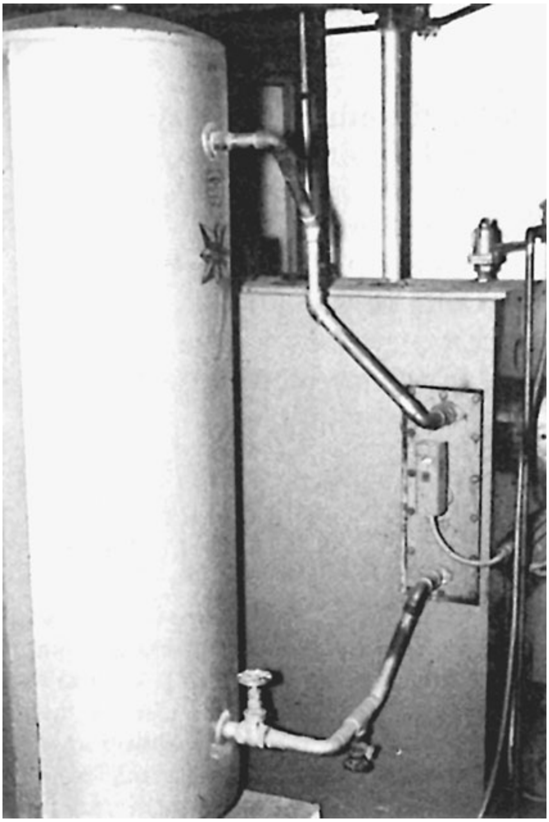

Fig. 16-12. *Storage tank connected to tankless water heater. The hot water circulates between the tank and the coil to maintain high temperature. The tank increases the overall volume of available hot water.*

From an energy-conservation point of view, a tankless (coil-type) water heater is not desirable. It is inefficient and wasteful of energy. To produce domestic hot water, there must be a double heat transfer. The boiler water will first be heated with oil or gas, and it will then transfer its heat to the water in the tankless coil. To produce domestic hot water, the boiler must be heated all year long, winter and summer. As discussed in chapter 14, controls associated with the heating system prevent steam or hot water from being distributed throughout the house when heat is not required. Nevertheless, the heated boiler water during those months does represent wasted energy. It also introduces an additional heat load on the house at a time when heat is not desired. If you should decide to replace a tankless coil with a tank-type water heater, it can be done at a reasonable cost. An oil-fired water heater, however, is more expensive than a gas-fired or an electric unit.

Inspection procedure

The first portion of the tankless water heater inspection is performed in the bathroom. Turn on the hot water at the sink and the tub simultaneously and look at the flow. If the flow appears to be low, mineral deposits might have built up in the tankless coil, and maintenance might be needed. On the other hand, the low flow might reflect a constriction in the distribution piping. Record the fact that there is a low flow on your worksheet as a reminder that the condition must be corrected regardless of the cause.

At the boiler, look at the joints around the tankless heater mounting plate. This is a vulnerable area for leakage. If there are water stains and deposits around the joint, even if you do not see water dripping, maintenance is needed. The deposits tend to self-seal the leak, but this is not a permanent fix. Any movement or vibration of the boiler can cause the deposits to loosen, reactivating the leak. Often, all that is needed is tightening of the mounting bolts around the joint or replacing the gasket around the joint. Look at the inlet and outlet pipes to the heater and the mixing valve (if there is one) for signs of past or current leakage.

The temperature of the boiler water that heats the domestic hot water is controlled by an aquastat. You can check the operation of the aquastat by opening one of the hot-water faucets. After a while, the temperature of the boiler water will drop to a point where the aquastat will cause the oil or gas burner to fire. Tankless coils are designed to operate with a boiler-water temperature of about 200° F. Look at the boiler-mounted thermometer when the burner fires. If the boiler-water temperature is less than 180° F, the aquastat may be in need of adjustment. In some homes, the aquastat is intentionally set lower. This, of course, will result in a lower hot-water temperature.

Look for a temperature-pressure relief valve on the hot-water outlet line of the tankless coil. If there is a storage tank associated with the tankless coil, the relief valve might be located on the tank. A temperature-sensitive relief valve is a necessary safety control and is normally factory-set to discharge when the domestic hot water reaches a scalding temperature. If you do not find a relief valve, record the fact on your worksheet as a reminder to have one installed after you move in.

Tankless (wall-mounted) water heaters

A tankless (wall-mounted) water heater, also sometimes called an "instantaneous water heater," is an appliance that heats water as the water is used. It's about the size of a medium-capacity suitcase. Operating a hot-water faucet activates either a burner (natural gas or propane) or an electrical resistance heater that transfers thermal energy to a heat exchanger. As cold water flows through the heat exchanger, it is heated to a preset temperature, which is normally 120° F. However, for most units there is a temperature controller within the unit, which is accessible by removing the cover, that can adjust the temperature setting from 100° to 140° F. Some units have a remote control option for the homeowner to control the temperature. Tankless water heaters are designed for a maximum flow rate through the heat exchanger; exceeding the flow rate results in a cooler water outlet temperature. However, for high-demand applications, two or more tankless units can be connected together to operate as one. Closing the hot water faucet deactivates the heating mechanism.

One advantage of this water heater is that because of its compact size it can be placed close to the point of use so that very little of the water's heat is lost between where the water is heated and where it is used. However, if a gas-fired unit is installed as a retrofit, as opposed to an electric unit, it must be located in an area where the exhaust gases can be vented to the outside. One item of concern with a retro fit is the size of the gas line. It's important that the gas pipe being used meets the requirements of the new gas-fired tankless water heater. The gas requirement of the tankless unit often exceeds that of the existing tank-style water heater. If the water heater is a retrofit, you should check with the owner to determine whether the installation was approved by the local municipal building department. If the house is located in an area where there are exceptionally hard-water conditions, the heat exchanger inside the unit may require periodic flushing by a plumber to remove built-up scale. To minimize this problem, a water softener or a descaling device should be installed in the water line before the heater.

Indirect-fired storage water heaters

In the mid-1970s, a new type of water heater, an *indirect-fired storage water heater,* was developed. The unit is called indirect-fired because the medium used for heating the domestic water is boiler water, which itself is heated in a separate boiler used for either central heating or space heating. The boiler water circulates through a heat exchanger that is located either within the heater tank or in an external compartment. Figure 16-13 shows one type of indirect-fired storage water heater. This unit has a specially designed heat-exchanger coil mounted in the base of the tank. Hot boiler water flows through the coil, which is surrounded by domestic water. The design of the heat exchanger, which consists of an inner core surrounded by an outer helical coil, is such that it results in a relatively high recovery rate.

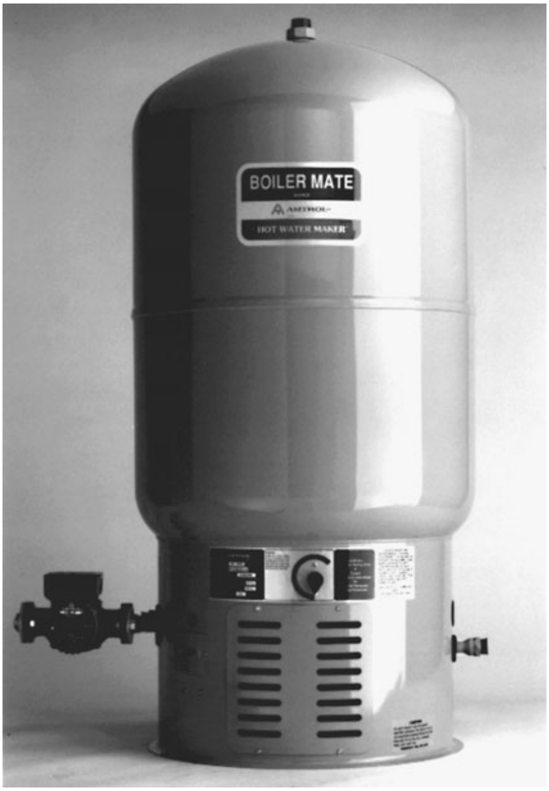

Fig. 16-13. *Indirect-fired storage water heater.*

The typical capacity of this residential indirect-fired water heater is about 40 gallons. With water heaters, whether there is an adequate supply will depend not only on the capacity but also on the recovery rate of the unit. The actual rate will depend on the Btu rating of the boiler that feeds the heat exchanger. For a typical house with a boiler that has an output rating of about 100,000 Btu per hour, the recovery rate is about 90 gallons per hour.

The indirect-fired storage water heater works off the heating system as a separate zone, utilizing its own circulator pump. According to manufacturers, 15–20 gallons of domestic hot water is removed before the zone aquastat calls for hot water from the boiler. Consequently, with the high recovery rate, the boiler firings are kept to a minimum.

In checking the unit, either raise the thermostat setting to HOT or let the hot water run until the boiler fires. You can tell if the pump is operating by touching the water supply to the heater before and after the boiler fires. The pipe should get hot. Look for signs of leakage around the base of the unit. Also, make sure that there is a pressure-temperature relief valve mounted directly on the tank casing. If you adjust the thermostat to activate the boiler and circulator pump, don’t forget to reset it.

In hard-water areas, it is possible that even though the boiler is functioning properly, the domestic water does not get very hot. This is caused by a lime buildup on the heat exchanger, which acts as an insulator and inhibits the heat transfer. In this case, the coil is in need of a cleaning, although it may require replacement.

Checkpoint summary

- ❍ Is hot water supplied by a separate-tank type of water heater, a tankless coil, or by a wall-mounted tankless heater?
- ❍ If you have a tank-type heater, record the capacity (gallons), the recovery rate (gallons per hour), and the age (from serial number).
- ❍ Are exhaust gases spilling out of the draft diverter?
- ❍ Is tank-type heater adequate for house? (See table.)

Full bathrooms	Sum of tank capacity plus recovery rate
1	70
2	90
3	105
4	115

- ❍ Is hot-water flow adequate when two fixtures are turned on?
- ❍ Are hot and cold supply lines incorrectly installed (reversed)?
- ❍ Does water heater contain a properly installed temperature-pressure relief valve?
- ❍ Is the relief valve dripping?
- ❍ Is there an expansion tank on the cold water line?
- ❍ Look for signs of corrosion or past leakage:
  - — Tank type (gas-fired unit): corrosion dust and flaking metal in burner area.
  - — Tankless coil: rust and deposits around fittings and gasket.
- ❍ Is there an "instant" hot-water return line?

17

Air-conditioning

Air-conditioning capacity 263
Air-conditioning systems 264
  Integral system 264
  Split system 264
    Inspection procedure 265
  Heat pump 272
  Evaporative cooler 273
Checkpoint summary 273

Before outlining the procedure for inspecting a central air-conditioning system, let me discuss the components of a typical residential system and their functions. Central air-conditioning systems provide comfort cooling by lowering the air temperature and removing excess moisture. This is achieved by recirculating air from the house across a cooling coil. As the air flows across the coil, its temperature drops, causing some of the moisture in the air to condense out on the coil. The cool, dehumidified air is then distributed throughout the house, and the moisture is disposed of through a drain.

The basic components of an air-conditioning system are the compressor, condenser, expansion device, and evaporator. The evaporator is the cooling coil mentioned above. The other components simply provide the means for the evaporator coil to cool the circulating air. This is done using a cooling medium, a *refrigerant*, that cycles between the components. (See FIG. 17-1.) The refrigerant is normally a gas at atmospheric pressure and temperature. By (1) applying pressure to the refrigerant and (2) removing absorbed heat, the refrigerant will change from a gas to a liquid. At this point, the refrigerant exists as a high-pressure liquid. When the pressure on the liquid refrigerant is released, the refrigerant (3) expands and changes back to a gas. In the process of changing from a liquid to a gas, the refrigerant (4) absorbs heat from its surroundings, thus cooling the air passing over the coil.

An air-conditioning system is a closed system, and theoretically there should never be a need for additional refrigerant. However, in practice the various fittings on the connecting pipes can loosen or develop hairline cracks that can allow some of the refrigerant gas to escape. Usually, when a system is low in refrigerant, the practice is to introduce additional refrigerant and not to look for the openings in the system that allowed the refrigerant to escape. However, if the air-conditioning system cannot hold a refrigerant charge for at least one season, the

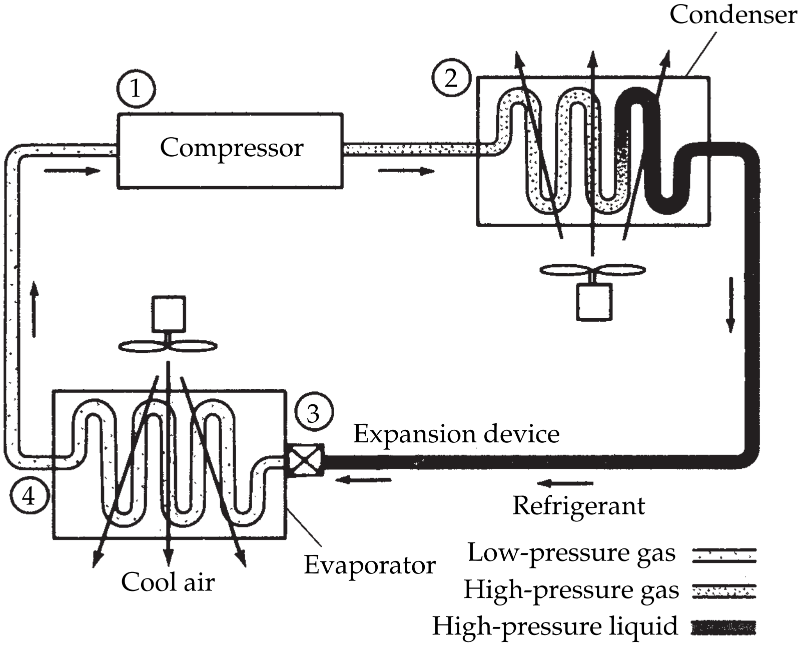

Fig. 17-1. *Basic air-conditioning cycle. Normally the compressor and condenser are contained in a single housing.*

leaks in the pipes or fittings should be located and corrected.

Air-conditioning capacity

The overall cooling capacity of an air-conditioning system is usually measured in Btus (British thermal units) or tonnage. One Btu is the heat required to raise the temperature of 1 pound of water 1° F. A rating in Btus per hour indicates the amount of heat that the unit will remove in that time. A *ton* of air-conditioning, historically, represented the cooling effect achieved by melting 1 ton of ice in twenty-four hours. An air-conditioning system that is rated at 1 ton will provide 12,000 Btus of cooling in an hour.

The cooling needed for a house depends on several variables, such as size of the rooms, whether there are cathedral ceilings, the number and type of windows, the amount of insulation in the walls and ceilings, and whether shading is provided by trees. The following rule of thumb can be used to determine the size of the air-conditioning system needed to cool the house that you are inspecting. One ton of cooling capacity (12,000 Btu) is needed for every 550 square feet in the structure. Therefore, if you are inspecting a house with 1,800 square feet, you will require an air-conditioning system rated at 3 tons. More recently, because newer homes are tighter and are better insulated, a figure of 700 square feet per ton has been used. The exact requirement can be determined by a professional. However, the rule of thumb is quite effective in determining whether the unit servicing the house is grossly oversized or undersized. Calculate your cooling requirements and then ask the homeowner the capacity of the air-conditioning system.

You can check the capacity yourself by looking at the data plate on the compressor-condenser. Even though most manufacturers donot put the system's capacity on the data plate, it can be approximated. The model number of the compressor on most air-conditioning systems contains digits that represent the approximate number of thousands of Btu of cooling capacity. See TABLE 17-1.

Another figure on the data plate that can be used for determining the cooling capacity is the full load amperage (FLA) or rated load amps (RLA). There are approximately 7 amps per ton of cooling. The FLA for the Tappan compressor in TABLE 17-1 is 27.1. When this figure is divided by 7, you get approximately 4 tons, which verifies the figure indicated by the model number.

An air-conditioning system should be properly sized or slightly undersized, but it should not be oversized. A unit that is too large operates intermittently. It quickly chills the air and then shuts down. When the system is shut off, moisture in the air does not condense, and the system does not dehumidify the air. For the system to remove enough moisture to make the air comfortably dry, the evaporator coil must be kept cold. This means that the compressor should run almost continuously for maximum comfort. Even though a system that is too small will run continuously, its cooling capacity is not adequate to remove enough heat to cool the room to a comfortable level.

Air-conditioning systems

There are basically two types of central air-conditioning systems used in residential structures—the *integral* system and the *split* system.

Integral system

This system, sometimes referred to as a *single-package unit*, is self-contained. That is, all the components—compressor, condenser, expansion device, and evaporator, plus the electrical controls and fans—are contained in a single housing. This system, which must be vibration-mounted, is often installed in the attic or a crawl space with ducts projecting through the exterior wall or roof to provide air for cooling the condenser. The integral system is less expensive than the split system. However, the noise level from the compressor makes an interior installation less desirable.

Split system

In this system, the compressor-condenser is physically apart from the evaporator coil. To eliminate the interior noise and provide outside air for cooling the condenser, the compressor and condenser are housed in a unit located outside the structure, usually at the rear or side. The evaporator coil, on the other hand, is located in the house, either in the attic or inside the heating system. The specific location often depends on the type of heating system. When the house is heated by forced warm air, the evaporator coil is usually located in the furnace plenum and utilizes the furnace blower to move the air through the coil. When the house is heated by steam or hot water, the evaporator coil is usually located in the attic with a separate blower (fan) to move the air through the coil and distribute it throughout the house. When there

Table 17-1. Air conditioner capacity.

Manufacturer	Model number	Btu/hour	Tons
General Electric	BTB390A	30,000	2½
Bryant	567CO36RCU	36,000	3
Carrier	38CC042-1	42,000	3½
Tappan Company	CM48-42C	48,000	4

is no attic, the evaporator unit can be located in a closet or the basement.

The compressor-condenser is connected to the evaporator coil by two copper pipes that contain the refrigerant that cycles between the two units. (See FIG. 17-2.) The diameters of the two pipes are different—one pipe is the size of a pencil, the other the size of a broom handle. The small pipe (pencil size) is the liquid line; it carries the high-pressure liquid refrigerant from the condenser to the expansion valve. The larger pipe (broom-handle size) is the suction line; it carries low-pressure refrigerant gas from the evaporator coil to the compressor. The suction line should be insulated. Usually it is covered with black foam-rubber type of insulation.

Inspection procedure Whether the air-conditioning system can be checked operationally depends on the outside air temperature. Most manufacturers do not recommend turning on the system at temperatures below 60° F because of the possibility of damage to the compressor. If the outside temperature is below 60° F during your inspection, do not start up the air conditioner. If the system cannot be checked prior to purchasing the house, the seller should provide you with a guarantee of its operational integrity. If the temperature is above 60° F during your inspection, walk over to the compressor and have someone turn down the thermostat that controls the air conditioner so that the system will begin to operate.

Compressor The compressor is the most important part of any air-conditioning system and the most costly to replace. Its projected life is about eight to ten years, although units have been known to last over fifteen years. In areas of the sunbelt with long air-conditioning seasons,

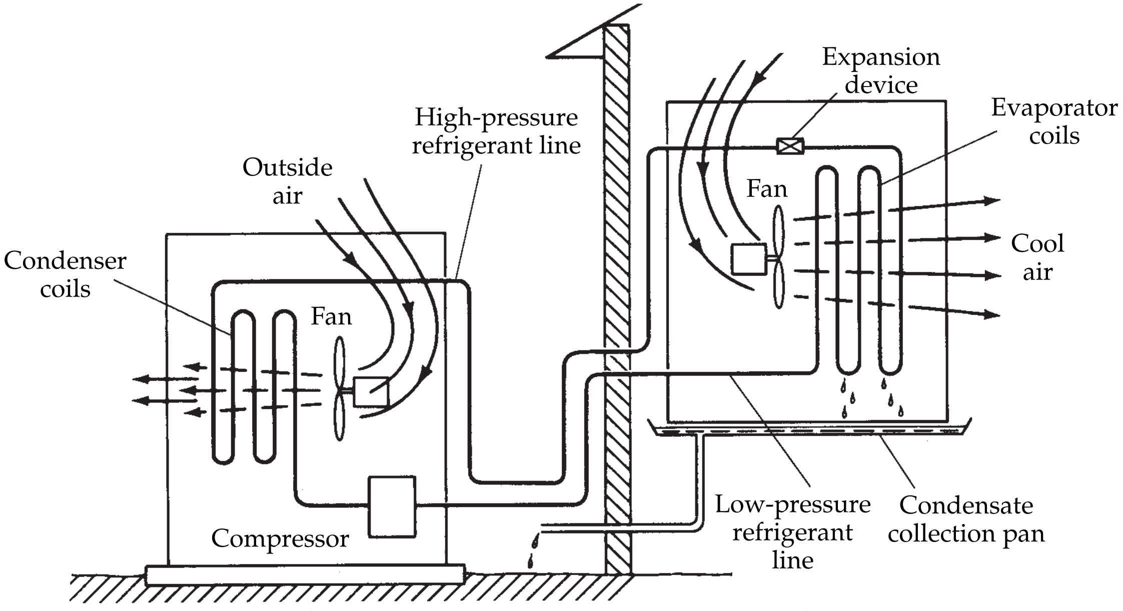
Expansion device
Evaporator coils
High-pressure refrigerant line
Outside air
Fan
Fan
Cool air
Condenser coils
Compressor
Low-pressure refrigerant line
Condensate collection pan
Compressor/condenser
Cooling coil/air handler

Fig. 17-2. *Schematic diagram of a split air-conditioning system. The compressor-condenser is located outside the house, and the cooling coil is inside the house (usually in the furnace plenum or attic).*

the projected life will be somewhat less. After the system has been turned on, listen for any unusual sounds. The compressor should start up smoothly. A straining, grunting, groaning, or squealing noise indicates a problem condition that should be checked and corrected by a competent service organization. Shut the unit off and note the condition on your worksheet. Once the compressor starts up smoothly, it should then operate continuously without any noise (except a low hum) or squeaks. The compressor should not operate in short cycles (on and off repeatedly). If this occurs, there is a problem condition and the unit should be shut down.

When the air-conditioning system is turned on, the fan associated with the condenser should begin to operate at the same time as the compressor. Look inside the unit to see if the fan is turning. If the interior portion is not visible and you cannot see whether the fan is operating, place your hand over the unit. Air rushing over your hand is an indication that it is operating. After the system has been operational for about fifteen minutes, the air being discharged through the condensing coil should be warm. This air is removing the heat that has been generated during the compression of the refrigerant. Air that is not warm is usually an indication that the compressor is not operating properly, a condition that should also be checked by a competent service company.

After the system has been operational for about fifteen minutes, look at the low-pressure refrigerant line (the pipe about the size of a broom handle). Usually the pipe is covered with insulation. If a section of the pipe is exposed, grab it with your hand. If the compressor is working properly and there is an adequate refrigerant charge, the pipe will be quite cool to the touch. On many occasions, the pipe and end fittings will be "sweating" as a result of condensation. This is a normal operating condition. However, a low-pressure line covered with frost usually indicates a deficiency in refrigerant. Even though cooling can be obtained with an air-conditioning system that is deficient in refrigerant, the efficiency of the system is greatly reduced.

Occasionally you will find a sight glass on the small-diameter refrigerant line. The sight glass is a small device installed directly into the high-pressure liquid refrigerant line that allows the homeowner to see whether there is a problem with the flow of the liquid refrigerant. The refrigerant is colorless, and if the system is operating properly, nothing unusual will be noted when looking at the sight glass. However, when the system is low in refrigerant, bubbles will show up in the liquid as it passes under the glass. Record it on your worksheet.

The location of the compressor-condenser is important for efficient operation. (See FIG. 17-3.) The compressor should be located where it will receive a minimum of direct sunlight, since the cooler the air flowing across the condenser, the more efficient the cycle. See if the compressor is positioned so that the condenser air intake is at least 12 inches away from any obstruction

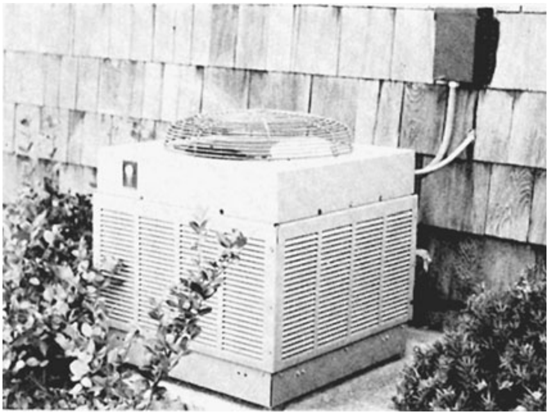

Fig. 17-3. *Typical compressor-condenser for a split air-conditioning system. Airflow through unit must be unobstructed.*

or dense shrubbery. If it isn't, there will not be an adequate airflow for condenser cooling.

The compressor-condenser can be a noisy piece of equipment and should be vibration-mounted on a concrete slab or precast concrete blocks that will not settle. The unit should be level. Excessive uneven settlement can cause fractures in refrigerant-line fittings and thereby allow the refrigerant to escape.

Look for an electrical disconnect switch on the exterior wall near the compressor. The purpose of this switch is to allow the maintenance man to disconnect the unit so that if someone in the house unknowingly turns the thermostat down, the unit will not be activated while he is making repairs. The absence of a disconnect switch should be recorded on your worksheet. The overall compressor-condenser unit should be checked to see whether it is in need of a cleaning. These units require periodic cleaning because leaves, seed pods, twigs, and dust tend to clog the condenser, thus restricting the airflow. The condenser has fins like an automobile radiator's, which can clog easily.

The condenser discussed above is air-cooled. On occasion, you might see a condenser that is water-cooled. A water-cooled unit can be located inside the structure since it does not require outside air. These condensers are generally not used for residential structures because they are quite wasteful and costly to operate. To cool the condenser adequately, cold water flows through a jacket around the coils absorbing heat and is then directed into a sink or floor drain. If the central air-conditioning system operates for twelve hours a day, this type of system can waste several thousand gallons of water a day. If you find this type of condenser in the air-conditioning system, you should consider its replacement.

In large air-conditioning systems such as those found in apartment buildings, water-cooled condensers are not wasteful or costly to operate because the cooling water is recirculated. After the water absorbs heat in the condenser, it is pumped up to a cooling tower, often located on the roof, where it loses its heat and is then recirculated to the condenser.

Evaporator After checking the compressor-condenser, you should inspect the evaporator unit, commonly called the cooling coil. The evaporator will usually be located in the attic or the basement. When the unit is located in the attic, you will often find the refrigerant lines from the compressor running up along the outside of the structure and entering the building at the attic level. When the evaporator coil is located in the basement, either as a separate unit or in the furnace plenum, the refrigerant lines from the compressor are short and run directly into the structure. If possible, the evaporator coil should be observed after the unit has been operational for about thirty minutes. The coil might not always be accessible because it may be covered with a sheet-metal casing that cannot be easily disassembled. If the evaporator is accessible, look at the coil and the associated refrigerant tubing. The evaporator is used to cool and dehumidify the circulating air. If you notice a frosting condition (a buildup of ice) on portions of the coil and refrigerant tubing rather than dripping water, the system is not operating properly. The frosting is usually the result of an insufficient airflow through the evaporator coil or an inadequate amount of refrigerant in the system. This condition should be indicated on your worksheet.

Furnace-mounted evaporator When the house is heated by forced warm air, the most common location for the evaporator is in the *furnace* plenum. A furnace-mounted unit takes advantage of the ducts that have been installed for heating the house and also uses the heating-system blower to circulate the cool air.

You can tell if the evaporator coil is located inside the furnace plenum by whether there are refrigerant lines entering the sheet-metal casing of the plenum. The refrigerant lines inside the casing are connected directly to the evaporator coil. The most common type of coil that is found in a furnace plenum is a two-section design, an *A-coil* (because of its shape). Sometimes an inclined or horizontal coil is used. (See FIG. 17-4.)

Below the evaporator coil is a pan that collects the water condensing out of the circulating air. The water is then removed by means of a plastic drain line that will be visible when looking at the furnace plenum. Look for it. Depending on the location of the furnace, the condensate drain line will run to a nearby sink where the condensate drips down the drain or will run through the foundation wall, where the condensate drips on the outside. When the condensate drain line is extended through the foundation wall, there should be a splash plate below the end of the pipe so that the dripping water can be directed away from the foundation.

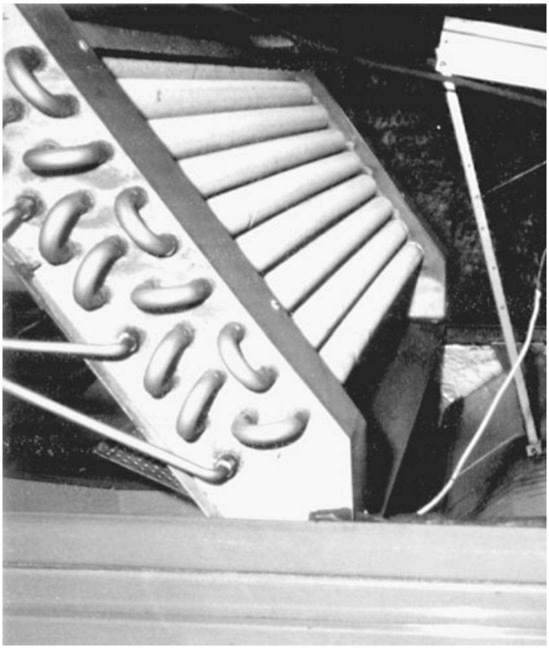

Fig. 17-4. *Evaporator (cooling) coil mounted in furnace plenum.*

Sometimes the condensate drain line runs from the furnace down to a small hole in the floor slab. The condensate trickling out of the drain line accumulates below the slab. This method of removing the condensate is not very desirable in those areas where the water will not readily drain because of a high water table or a high clay content in the soil. Even though the amount of water discharging from the condensate drain is small, the introduction of additional water could aggravate a condition that makes the lower level of the structure vulnerable to water seepage.

The condensate drain line should terminate at the plenum with a U-shaped trap. Since the condensate drain line is an open pipe leading directly into the cooling coil, the trap, which is on the negative side of the blower, prevents the blower from sucking air up through the drain line. This would inhibit the condensate from flowing out and could cause a flooding problem. Look for a trap. If it is missing, record that on your worksheet.

Occasionally the condensate drain discharges into a small rectangular box located near the furnace. This box is the reservoir for a lift pump. The purpose of the pump is to lift the condensate to a level where it can then flow to any desired location. Without a pump, it is often necessary to position the drain line so that it blocks a portion of the room or interferes with foot traffic. Check the pump's operation. These pumps have a float control that is activated when the water reaches a preset level. If there is only a small amount of water in the reservoir, the pump can be checked by pouring water from a glass into the reservoir. A malfunctioning pump should be recorded on your worksheet.

Now look at the overall furnace plenum around the evaporator coil. Rust and mineral deposits indicate a past or present problem in condensate removal. Water overflowing the condensate drain pan can damage the heat exchanger below. If you see this condition, you should have the furnace heat exchanger checked by a heating contractor for signs of deterioration. (See chapter 14.)

When the evaporator coil is located inside the furnace plenum, the blower for the heating system is also used as the blower for circulating the cool air. Because cool air is heavier than warm air, when the blower is used for air-conditioning, it should operate at a higher speed. Most often, however, the furnace is equipped with a one-speed motor. Consequently, the air-conditioning system is often not as effective as it might be. A pair of double- or triple-sheaved hubs can be installed to allow multispeed operation. When the blower is turned off by the master switch, check the tension in the fan belt. There should be no excessive slack. Press the belt midway between the pulleys. If the belt gives more than ¾ inch to an inch, it is too loose, and adjustment is needed. When the blower is operating, listen for any unusual noises or vibrations. They should be recorded on your worksheet. You might also ask the owner when the unit was last serviced—there is no substitute for periodic maintenance.

Blower coil When the evaporator coil is housed in a separate casing that contains a blower for circulating the cool air, the coil is commonly referred to as a *blower coil*. Most often, the blower coil is located in the attic. However, it can be located in a closet or in the basement. The blower coil should be vibration-mounted to prevent the noise of the blower unit from being transmitted into the living area. Vibration mounting can be achieved by placing the unit on rubber, cork, or styrofoam pads. (See FIG. 17-5.) The vibrations might also be isolated in the attic by suspending the unit from the roof rafters.

The base of the blower coil is basically a condensate collection pan. The accumulated condensate is removed by means of a drain line that will extend through the exterior wall, terminating on the outside, or extend through the lower portion of the roof, terminating in the gutter. Sometimes the condensate drain line terminates in the plumbing vent stack located in the attic. (See FIG. 17-6.) In many communities, this type of termination is not permitted because it is not in compliance with the plumbing code. If the drain terminates in the vent stack, record that on your worksheet. The legality of this type of connection should then be verified with the local building department.

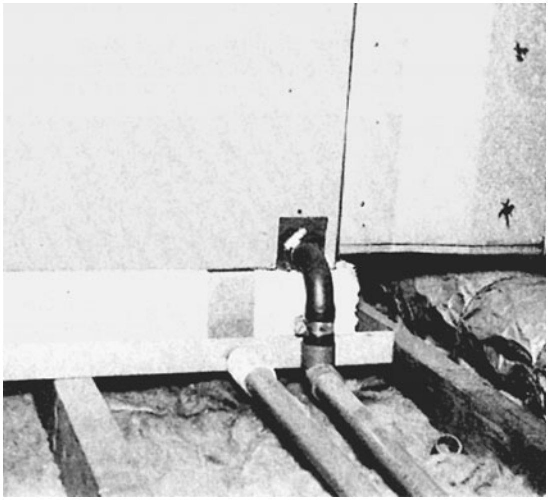

Fig. 17-5. *Attic-mounted blower coil. Unit is resting on a styrofoam pad to minimize vibrations. Below the unit is an auxiliary condensate drain pan and associated drain line. Note that the main condensate drain line does not have a U-shaped trap.*

The purpose of the vent stack is to channel sewer gases in the plumbing system to the outside. If the condensate drain line is connected to the vent stack and there is no trap on the drain line, the sewer gases may back up into the condensate drain line, enter the blower coil, and be circulated throughout the house. An additional concern is bacterial growth developing in the air handler. The condensate drain line should have a U-shaped trap near

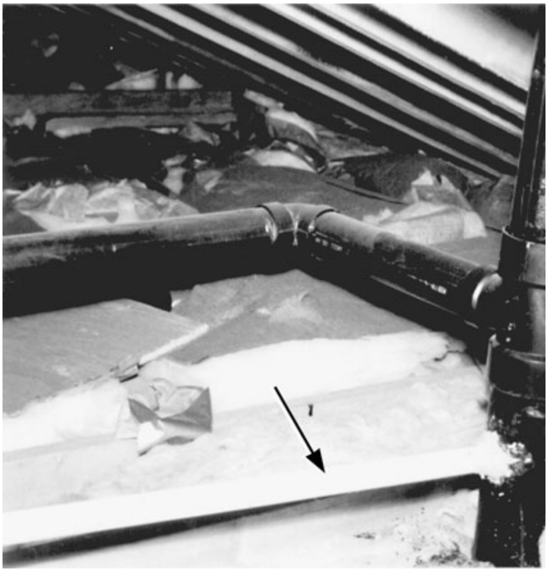

Fig. 17-6. *Air-conditioning condensate drain line terminating in plumbing vent stack. In many communities, this type of termination does not comply with the plumbing code.*

its connection to the blower-coil housing. On many installations, this trap is omitted. Look for it. If it is missing, one should be installed.

When the blower coil is located in the attic, certain steps must be taken to prevent cosmetic damage to the ceiling below in the event of a blockage in the main condensate drain line. Some blower-coil housings have a fitting for an auxiliary drain line that is located just above the main condensate drain fitting. If the main drain becomes clogged, the level of the condensate will rise and be drawn off by the auxiliary drain.

For those blower coils that do not have a fitting in the housing for an auxiliary drain line, there should be an auxiliary drain pan below the unit. The auxiliary pan will collect any condensate that overflows from the main pan when there is a blockage in the main drain line. Look for an auxiliary drain pan. In some parts of the country auxiliary drain pans are installed when the blower coil is located over any furred space, even when the blower housing has an auxiliary drain fitting. Record the absence of one on your worksheet. Unfortunately, many air-conditioning contractors do not install the auxiliary drain or drain pan. Because rising costs make it difficult to remain competitive, they cut costs wherever they can.

The auxiliary drain pan must have a separate drain line that discharges to the outside. It should not be connected to the main drain line. (See FIG. 17-7.) If it is, it reflects poor-quality workmanship; if the main drain line becomes clogged near the discharge end, the auxiliary drain line will also not function.

If the evaporator coil is accessible, it should be inspected for frost buildup. From an efficiency point of view, the attic is the least desirable area for locating the blower coil because of the high temperatures, easily reaching the 140° to 150° F that normally occur during the summer. Even though the blower coil is insulated, there will be a heat gain because of this high temperature. The overall attic temperature, however, can be lowered by increasing the number or size of the attic vent openings. A ridge vent is quite effective, as is a thermostatically controlled power ventilator.

Ducts After the air-conditioning system has been operational for about fifteen minutes, the air discharging from the registers should be felt to determine whether it is relatively cool. The temperature of the air discharging from the supply registers should be about 15 degrees lower than the temperature in the room. If the air does not have a slight chill, it might be because there is a heat gain along the duct leading to that register as a result of inadequate insulation, or the system may be undersized or low in refrigerant.

While checking the temperature of the air leaving the supply registers, also check the airflow. If the air discharging from the registers has a low flow and appears to be

Fig. 17-7. *Auxiliary condensate drain line connected to main condensate line. This negates the use of an auxiliary drain and reflects poor-quality workmanship. If the main drain becomes clogged near the discharge end, the auxiliary drain will not function. Note that the U-shaped trap is missing.*

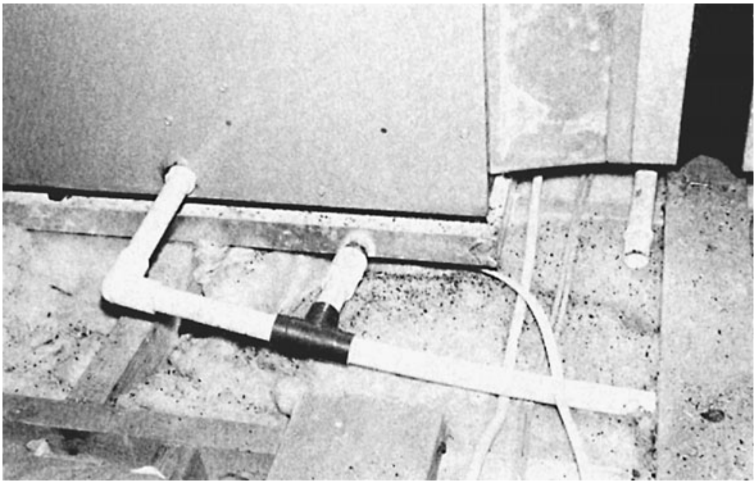

sluggish, it might indicate that there is an obstruction within the system caused by dirty filters or icing on the evaporator coils. Sometimes the condition is caused by an undersized fan or the need for balancing the airflow between the registers. In any case, the condition is abnormal and should be recorded on your worksheet.

As with a heating system, the location of the supply registers is important for effective air-conditioning. Since cool air is heavier than warm air, the cool air will tend to accumulate near the lower portion of the room and the warm air near the top. As a result, there is usually a temperature difference between the ceiling and floor. This stratification of heat layers can be minimized by adequate circulation in the room. Adequate circulation can be achieved by locating the supply registers on the opposite side of the room from the return grille. When the return grille is near the supply register, the air discharging from the supply is drawn in by the return grille and does not have a chance to circulate adequately around the room. In many houses, the rooms do not have individual return grilles. Instead there might be a large central return located in the hall. In these cases, the supply registers should be located on a wall that will allow the supply air to circulate completely prior to being drawn off and returned to the central grille. Also, the doors to the individual rooms must be undercut so that when they are closed, the supply air will be able to flow to the return grille.

Ideally, air-conditioning supply registers should be located in or near the ceiling. To minimize air-conditioning installation costs, rather than install new ducts many new homes use the ducts and registers provided with a forced-air heating system. These registers are usually located near or at the floor level and are quite effective for heating purposes. However, when they are used for air-conditioning, they are less effective and tend to increase the stratification effect. Some houses that have forced-warm-air heating have what are called *high-low registers*. The duct supplying the heat register is extended vertically to a point near the ceiling level where it terminates at another register. When the system is used for heating, the damper controlling the upper register is manually closed, and the lower register is opened. When the system is used for cooling, the damper controlling the lower register is closed, and the upper register is opened. This type of arrangement is very desirable and is often found in high-quality construction.

The presence of a central air-conditioning system does not mean that the entire house is air-conditioned. Look specifically for registers as you walk through the house. In many raised ranches, I have found that the upper level is air-conditioned and the lower level is not. If you have any questions, check each room while the air-conditioning system is operating.

There are two basic types of ducts used in residential structures—sheet metal and glass fiber. While checking the distribution portion of the air-conditioning system, look for exposed ducts. The glass-fiber type of duct is by its very nature insulated. However, the metal duct may or may not be insulated. The fact that there is exposed metal on the outside does not mean that the duct is not insulated. The insulation might be located inside the duct. Whether the metal duct is insulated can be determined by feeling the duct when the system is operating (if no insulation, the duct will be quite cool) or by striking the duct with your fingernail. If there is no insulation, you will hear a ringing sound, and if there is insulation, you will hear a dull thud. All ducts that lead through unfinished areas such as crawl spaces and attics must be insulated so that the cool air flowing through the ducts will not absorb heat from its surroundings. Pay particular attention to the joints for indications of air leakage. Any open joints should be sealed with inexpensive duct tape. Also, whether the evaporator coil is located inside the furnace plenum or in the attic, check the joints around the housing for air leakage. Very often, there are open joints that must be sealed.

Heat pump

A heat pump is a year-round air-conditioning system that provides warm air during the winter and cool air during the summer. It is basically a compressor-cycle air-conditioning system (similar to the one described previously) that can operate in reverse. During the reverse operation, the condenser functions as an evaporator, and the evaporator functions as the condenser. The overall refrigerant cycle, however, remains the same. (See page 262 and FIG. 17-1 for the operational flow details.)

When the system is operating, the condenser (which is located in the house) is cooled by the air that is circulated around the house. As this air passes over the condenser, it absorbs heat that is used for heating the house. For the cycle to operate properly, the evaporator (which is located on the outside) must absorb heat from the outside air. Even when the outside temperature drops to as low as 20° F. the evaporator can absorb heat because the refrigerant within the evaporator is at a lower temperature. However, as the temperature of the outside air drops, the ability of the evaporator to absorb heat decreases, decreasing the effectiveness of the heat pump. Even though a heat pump may be operational at temperatures as low as 20° F, the Btu output is sufficiently reduced so that auxiliary heaters are usually required.

A heat pump should be sized for the air-conditioning load and not the heating load. Otherwise, the air conditioner will be oversized. Except for a small section of the Deep South, the heating load on a house will always be greater than the cooling load. A heat pump will produce approximately 20 percent more Btus per hour for heating than it does for cooling. Consequently, in most parts of the country, when a house is heated by a heat pump, auxiliary heat will also be needed. For example, in the New York area, a typical eight-room house would require a furnace that could produce about 100,000 Btus per hour for winter heating and also a 3½ ton (42,000 Btu/hour) air-conditioning unit for summer cooling. If a properly sized heat pump was used for heating, it would produce only 50,400 Btu/hour. The remaining 49,600 Btu/hour would have to be provided by auxiliary heaters, which are usually electrical resistance heaters.

The auxiliary heaters in heat pumps are automatically activated when the pump cannot supply sufficient heat to keep up with the heat loss of the structure during the winter. In northern communities where there is a considerable amount of moisture and low temperatures during the winter, there is a tendency toward an ice buildup on the metal fins of the outdoor evaporator coil. An excessive ice buildup could cut off air circulation across the coil and result in a loss of heating capacity. With some heat pumps, this icing condition is automatically controlled by a defrost cycle that reverses the flow of the refrigerant for a short time. The hot refrigerant heats the outdoor coil and melts the ice. During the defrost cycle, the auxiliary heaters are usually energized to offset the cycle's cooling effect on the indoor circulating air.

A heat pump can operate in either the heating or air-conditioning mode. Most manufacturers suggest that the unit be operated in the air-conditioning mode when the outdoor air temperature is above 65° F (unlike a regular air conditioner, where the recommended temperature is 60° F) and in the heating mode when the temperature is below 65° F. Operating a heat pump in the wrong mode can result in damage to the compressor.

The components and problems of heat pumps are basically similar to those of air-conditioning systems. Consequently, the overall inspection procedure outlined earlier in this chapter for air conditioners should be used when inspecting heat pumps. However, you should not check both modes of operation. As long as the unit is functioning properly in the mode tested, it is an indication that the major and most costly components (compressor, fans, and coils) are operational. The system should then also function in the opposite mode. If it doesn't, a faulty reversing valve is usually the cause.

Evaporative cooler

One of the benefits of air conditioning is the dehumidification of the circulating air. This benefit is not without cost. Cooling and dehumidifying the air is more costly than cooling alone. In the southwestern part of the United States, the outdoor air is relatively dry, and dehumidification is not necessary. In this area, cooling can be achieved by means of an evaporative cooler. Because of the low humidity, water readily evaporates. In the evaporation process, the water absorbs heat from its surroundings and lowers the temperature.

The typical evaporative cooler consists of a sheet-metal and plastic casing containing a fan, pads, filter, and a water source. The pads, which hold the water, can be wetted by a spray, by a trickling stream, or by passing through a reservoir on a rotating drum. In some units, the wetted pads also function as air filters. As the air passes over or through the pads, it is cooled by the evaporating water and then distributed throughout the house.

When inspecting an evaporative cooler, turn the unit on and listen for any unusual sounds or vibrations in the blower compartment. Also look for signs of water leaks and check the pads for deposits and crusting. For efficient operation, the pads may require cleaning or replacement.

Checkpoint summary

General considerations

❍ How old is the air-conditioning system?
❍ When was the unit last serviced?
❍ Do not turn system on if the outside air temperature is below 60° F.

Compressor-condenser

- ❍ Check compressor during startup. Listen for any unusual sounds, such as straining, groaning, or squealing.
- ❍ Does compressor operate smoothly without short-cycling (repeated startup and shutdown)?
- ❍ Is condenser fan operating properly?
- ❍ Does compressor appear to be functioning properly? (Warm air should be discharging from unit.)
- ❍ Are there indications that the system is low in refrigerant (frosting on low-pressure refrigerant line or air bubbles in sight glass)?
- ❍ Is compressor-condenser located properly for maximum effectiveness (minimum sun exposure and unrestricted airflow)?
- ❍ Is unit in need of a cleaning (clogged with leaves, twigs, dust, etc.)?
- ❍ Check that unit is level and adequately supported by a concrete pad or blocks.
- ❍ For safety and maintenance, check for a main electrical disconnect for the compressor located near the unit.

Evaporator (cooling coil)

- ❍ During operation if possible, check cooling coil for frosting (ice buildup), usually the result of an insufficient airflow or lack of refrigerant.

Furnace-mounted evaporators (installed in the furnace plenum)

- ❍ Check for signs of leakage, mineral deposits, and areas of rust and corrosion.
- ❍ Note method of condensate discharge. Is it
  —to a nearby sink?
  —to the exterior?
  —to a floor drain?
  —to a hole in the floor slab (less desirable)?
- ❍ If condensate discharges into a reservoir lift pump (small rectangular box), check operation of pump.
- ❍ Check blower motor for unusual noises and vibrations. Is blower a single- or two-speed unit?

Blower coil (housed in a separate casing and most often located in the attic)

- ❍ Is unit vibration-mounted?
- ❍ If access is available, inspect evaporator coil for frost buildup.
- ❍ Check for a condensate drain line.
- ❍ Does this drain line discharge the condensate to a nearby roof gutter or directly into a plumbing vent stack? (The latter type of connection is usually not permitted and should be verified with the local building department.)
- ❍ Check blower-coil casing for auxiliary drain-line fitting; if not present, check for auxiliary drain pan below the unit. The auxiliary pan should contain an independent drain line that is *not* connected to the main drain line.
- ❍ If blower coil is located in attic, is the attic adequately ventilated?
- ❍ Is ventilation provided by ridge vent or thermostatically controlled power ventilator?

Ducts or registers

- ❍ Check airflow and temperature after fifteen minutes of operation.
- ❍ Note type of supply registers—ceiling units, combined type (heating and air conditioning), or high-low registers.
- ❍ Check whether all rooms are air-conditioned.
- ❍ Are supply registers and return grilles efficiently located?
- ❍ If rooms do not have individual return grilles, check for a large central return grille (often located in the hall).
- ❍ Are doors to the rooms undercut to permit proper air circulation?
- ❍ Check ductwork for open joints, signs of air leakage, and uninsulated ducts (particularly in the attic and crawl spaces).

18

Swimming pools

Concrete pools 275
Vinyl-lined pools 276
Preformed fiberglass pools 277
Pool equipment 278
Pool accessories 280
Inspection procedure 280
Checkpoint summary 282

Swimming pools, like central air-conditioning systems, are no longer considered a luxury that only a privileged few can afford. New construction methods and materials have considerably lowered the cost of construction. Swimming pools can be installed either in-ground or aboveground. You will find them with various sizes and shapes that can fit into almost every budget. Because the shape of an in-ground pool can be freeform, the shapes available are limitless. On the other hand, the shape of aboveground pools is basically limited to circular, rectangular, or oval. There are three types of construction for the shell of an in-ground pool: concrete, vinyl lined with sidewall supports, and preformed fiberglass. These items as well as the pool's associated equipment and accessories will be discussed below.

Concrete pools

Concrete pools can be constructed using any of four methods: shotcrete, gunite, poured concrete, and concrete block. Both shotcrete and gunite are applied and sprayed from a hose that is directed behind, over, and above previously installed reinforcing rods (re-bars). (See FIG. 18-1.) Shotcrete is premixed wet concrete, and gunite is a dry mix, which combines with water as it discharges from the hose. The spraying of the concrete mix allows complete freedom of size and shape because it can follow the contours of any excavated shape. Poured concrete and concrete block walls are more restrictive with regard to the shape of the pool. Whichever method is used to construct the shell, an interior finish must be applied to provide a waterproof surface.

The most common finish used on a concrete pool is plaster. It gives a smooth finish to the pool and also provides a nonskid walking surface. (See FIG. 18-2.) The projected life of a plaster finish is about 7 to 10 years depending on how well the pool and water quality has been maintained over the years. In many homes where the pool and the chemistry of the water have been diligently maintained, the plaster finish has lasted considerably longer.

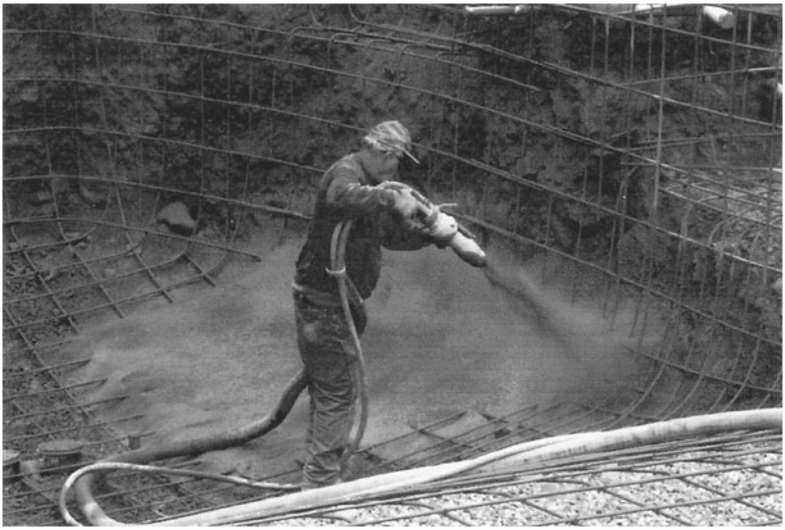

Fig. 18-1. *Gunite being sprayed around reinforcing rods for a concrete swimming pool.*

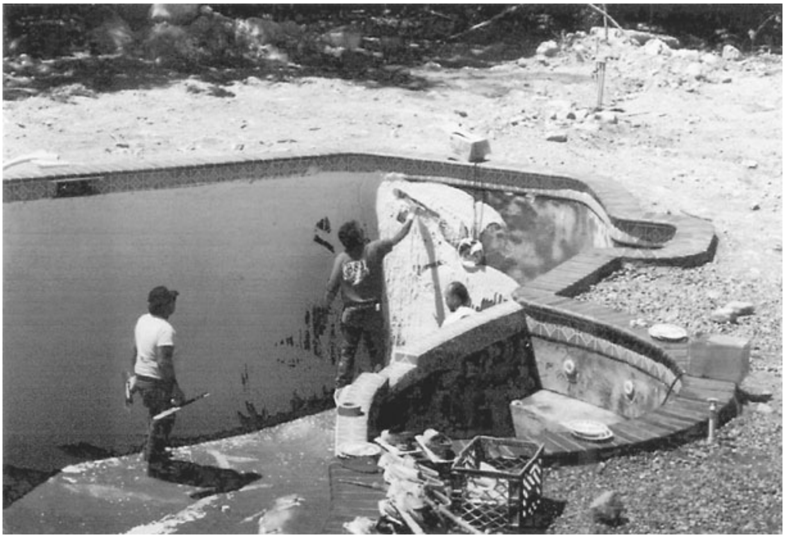

Fig. 18-2. *Plaster finish being applied to a concrete pool.*

Other finishes are paint, fiberglass coating, and ceramic tile. Tile is by far the most expensive; so most homeowners opt for a decorative ribbon of tile above the water line. The top edge of the pool shell is covered with coping, which prevents water from getting behind the shell. It can also serve as a handhold for swimmers and as a shove-off point into the water. The coping can be in the form of precast coping stones, flagstones, or brick, or it can even be the extension of a concrete deck over the edge of the pool shell.

Vinyl-lined pools

The advent of vinyl-lined pools brought the ownership of swimming pools to within the reach of many families. Construction of the pool shell consists of two phases: building of the perimeter frame or sidewalls and installing the liner. The most common sidewall panels are galvanized steel or aluminum. The frame must be structurally adequate to support the various forces exerted on the pool. After the sidewalls are bolted together, the bracing is installed. (See FIG. 18-3.) Then the area around the bracing is backfilled so that it is flush with the surrounding terrain. The sidewalls must have a smooth surface so as not to abrade the vinyl. They also don't have to be waterproof since the vinyl lining is. Prior to installing the liner, the base must be prepared. The bottom of the liner generally rests on a 2- to 3-inch sand base, and the top is secured with a special coping. The life of a vinyl liner is affected by the sun's ultraviolet light and the chemistry of the water. Manufacturers generally guarantee a liner against defective workmanship for 10 years. As vinyl pool liners age, the material becomes less supple and more prone to leaks from cracks or damage. It is possible to patch a small hole in a vinyl pool with a similar vinyl sheet material and a waterproof adhesive. Patching, if done correctly, can be effective, but once a liner has reached the age when several patches are needed, it is worth considering replacing the liner.

Aboveground pools are also vinyl lined. They are considered portable pools in that they can be dismantled and moved to a new location; however, many pools are partially or fully surrounded by a deck and are more permanent. The walls of the pool must be self-supporting and capable of withstanding the pressure being exerted on them by the water. The walls are usually galvanized steel, plastic, or aluminum and are 48 or 52 inches high. Because of the height of the walls, most municipalities do not require a safety fence surrounding the pool.

Preformed fiberglass pools

There are a number of advantages to a preformed fiberglass pool. The major advantage is its durability; the pool can flex and not get damaged. Also time is not lost in constructing the pool shell. After the site is excavated and prepared, the pool shell, which is brought to the site by truck, is lifted by a crane and placed into the excavation. (See FIG. 18-4.) Another advantage is the low maintenance. It is difficult for algae to cling to the sides of the pool because of its very smooth surface. As a result, the walls are very easy to clean. The main drawback of preformed pools is that

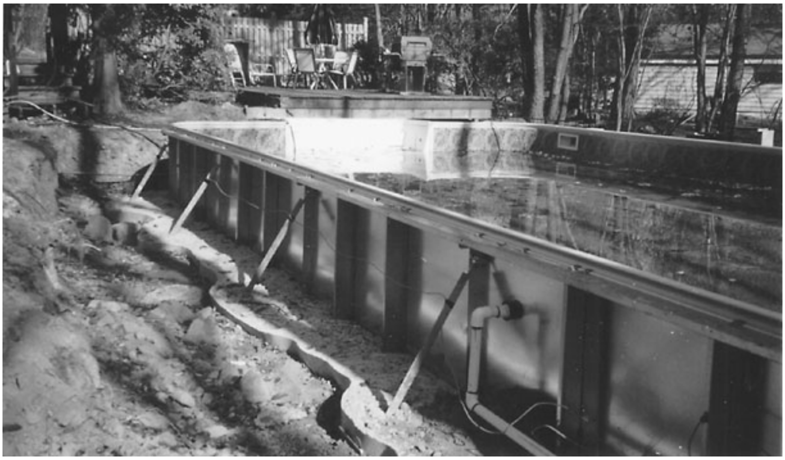

Fig. 18-3. *Steel bracing for the shell of a vinyl-lined pool.*

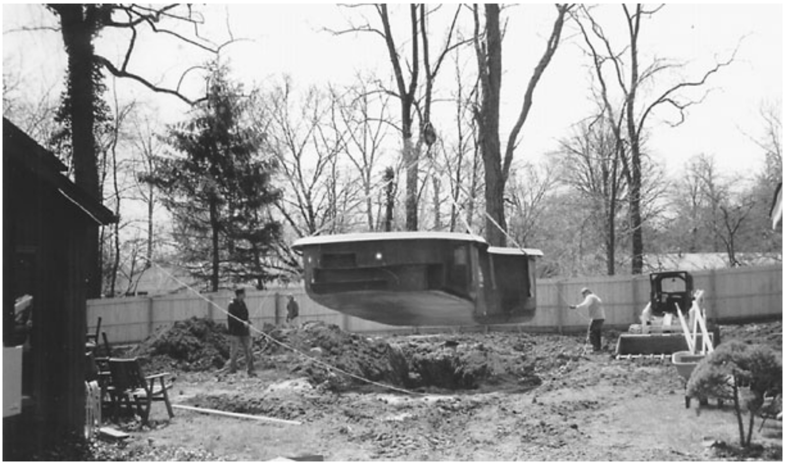

Fig. 18-4. *Preformed fiberglass pool shell being installed in excavated area.*

they are limited in size and shape. However, once the installation of the pool is complete, you cannot readily tell the difference between it and an on-site constructed pool.

Pool equipment

The standard equipment associated with circulating the water in a swimming pool are the pump and motor, filter, heater, and surface skimmer. The water circulation system is a closed system. The pump, which is driven by the motor, is normally located on the intake side of the filter. It draws water from the pool and forces it through the filter where dirt particles are removed. The water then flows to the heater and then back to the pool. To prevent the pump from getting clogged, there is a strainer basket on the water inlet side of the pump, which catches hair, lint, and other debris.

There are three basic types of filters available for swimming pools: sand filters, cartridge filters, and diatomaceous earth (DE) filters. The filters have no moving parts, and any one of the three will provide effective filtration. No matter which filter is used, the filtering medium will require periodic cleaning. The process used to clean DE and sand filters is backwashing, which is sending water backward through the filter, thereby flushing the debris onto the lawn, into the street, or to an approved drain line or sewer. Cartridge filters are not cleaned by backwashing; the cartridge is simply removed from the casing and washed. Most filters have a pressure gauge that is mounted on the top of the casing. The normal operating pressure is usually around 10 to 12 psi, although for some of the newer filters, it may be between 5 and 7 psi. As the filtering medium becomes dirty and clogged, it takes more pressure for the water to flow. It is generally time to clean the filter when the filter pressure increases approximately 10 psi above the normal operating range.

A swimming pool heater is not mandatory, but most houses with pools have one for comfort and because the heater can extend the swimming season. Most swimming pools are heated with a gas-fired heater, although they can also be heated with an electric heater and to a limited extent with solar heating or a heat pump. The ignition system in gas-fired heaters will be either a pilot light or electronic spark ignition. Either natural gas or propane gas (LP) can be used as the fuel for firing the heater. However, a heater that is designed for natural gas should not be used with propane gas and vice versa, because it will not operate properly. Keeping the pool water in the proper chemical balance is very important not only for health reasons and maintaining the quality of the water but also to extend the life of the pool heater. Water that is out of balance will result in a scale buildup within the heater, which if not corrected will cause a blockage of water circulation.

Most swimming pool shells have at least one to two built-in surface skimmers with skimmer baskets that are tied into the water's circulation system. When the pump is operating, dirt, leaves, oils from lotions, algae, and other debris that float on the water's surface are drawn into the skimmer by floating over the entrance weir. (See FIG. 18-5.) The baskets trap and collect the larger debris, and the oils and dirt particles circulate back to the filter where they are removed from the water. Some pools have skimmers that are also connected to a pipe located below the water level. This pipe prevents air from being sucked into the circulation system when the water level is down. Otherwise, when the pool water level is below the bottom of the weir, air will be drawn into the circulation system, which could damage the pump. The pump is not designed to run dry. Running a pump with air in the system or running it dry can

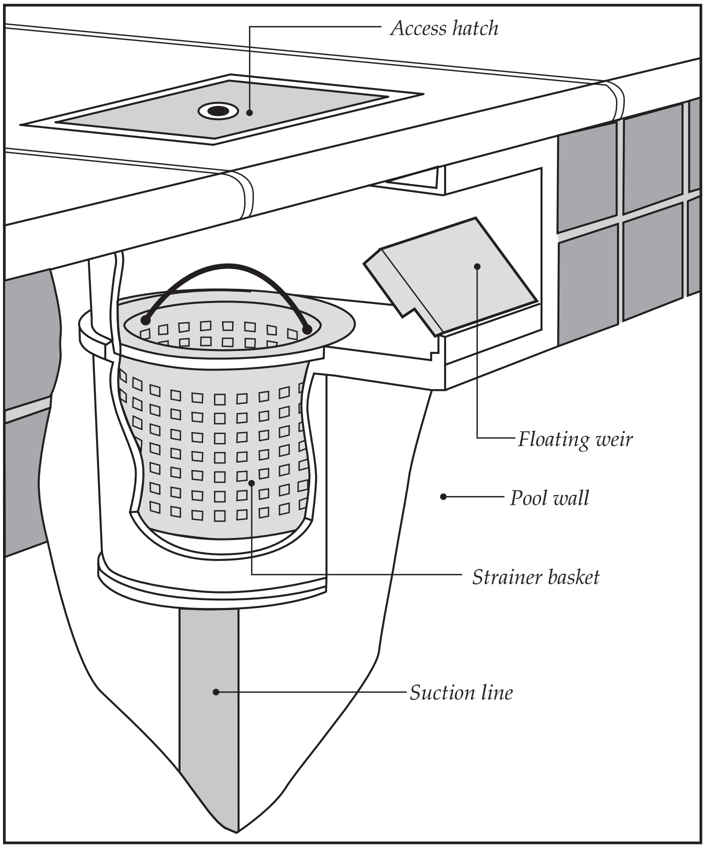

Fig. 18-5. *Cutaway view of a typical surface skimmer for a concrete pool.*

cause overheating and seriously damage both the pump and the motor.

Pool accessories

There are a number of accessories associated with a swimming pool, but the ones of concern from an inspection point of view are the pool cover, diving board, ladder, slide, and grab rails.

The main use for a pool cover is to keep debris out of the pool. It is mostly used when the pool will not be used for an extended period of time, such as during the winter months. (See FIG. 18-6.) However, depending on the material and design, the covers are also used by some homeowners during the swimming season to minimize heat loss and maximize heat gain. There are also safety covers that prevent small children or animals from falling into the pool. These covers are reinforced mesh with spring-loaded straps that hook onto the deck. There is also a "thermal or solar" cover available for a pool. It is essentially a sheet of vinyl bubble wrap that floats on the water. One puts it over the pool during swimming season because it allows the sun to heat the water and because it provides a barrier against evaporation.

Diving boards are available in spring-assisted and simple platform models. They are covered with fiberglass to make them waterproof and topped with a nonslip coating. Because of the danger of injury, some homes have jump boards rather than diving boards. Jump boards are shorter and are considered safer than diving boards. The concern about ladders, slides, and grab rails is whether they are adequately anchored to the deck.

Inspection procedure

In most municipalities a swimming pool is considered to be an auxiliary structure, and, as such, a building permit is required prior to construction. After completion, depending on the municipality, a Final Inspection Certificate, a Certificate of Compliance, or a Certificate of Occupancy is normally issued. If the house that you are inspecting has a swimming pool, check to see if there is a completion certificate for the pool.

There are certain limitations in a home inspection with regard to a swimming pool that you should be aware of. With no splashing, the average water loss in one week from evaporation is less than 1 inch. There

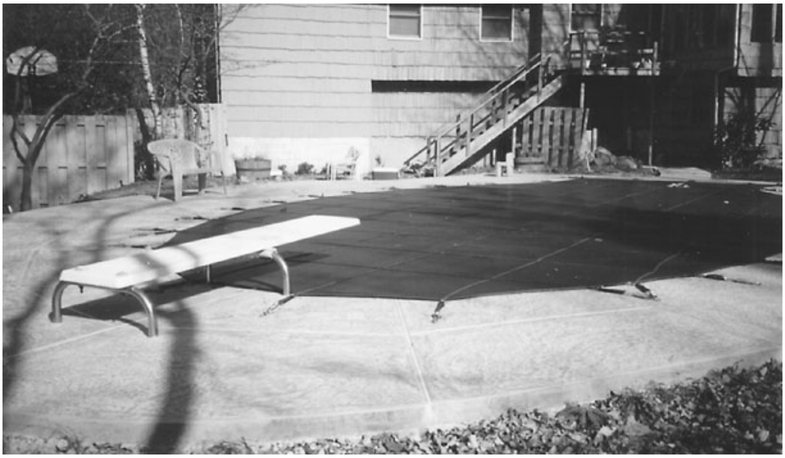

Fig. 18-6. *Pool cover keeps debris out when pool is not in use for extended period of time.*

could be a leak in the underground piping that causes the water level to drop more than what would be expected over a 24-hour period. This is a major problem to correct, but it will never be picked up during a home inspection. Pools lose water continuously through evaporation, splashing, and on the bodies of swimmers as they get out of the water. This is normal. However, a steady drop in the water line even when the pool is not being used may indicate a leak. This item should be discussed with the seller. In addition, the chemical composition of the water is generally not checked as part of a home inspection because it is considered to be normal ongoing maintenance.

As you walk around the house during your exterior inspection, check to see if there is a fence that encloses the swimming pool. This is important from a safety point of view. The fence should provide limited access to the pool. The gate(s) must be self-closing and self-latching. Check the condition of the fence and the operation of the gate. If maintenance is needed, record it on your worksheet for later correction.

The swimming pool and equipment can be inspected either before or after the inspection of the house, whichever is convenient. Check the condition of the deck that surrounds the pool. With an in-ground pool the deck is often concrete, although it could be stone, tile, or brick. Specifically look for cracked, chipped, or settled sections. Uneven joints between the sections are considered tripping hazards. If there are any open joints, particularly between the deck and the coping, they must be sealed. While on the deck if there is a grab rail, ladder, and slide, check to see if they are adequately anchored to the deck. If there is a diving board, look for cracks. If there are any, the board should be replaced. Is the board topped with a nonslip coating? Many older pools with diving boards do not meet new diving board standards. The safety upgrades outlined in the standards are mandatory. Check with a local pool company for the latest standards. With an aboveground pool the deck is generally wood constructed. Check the decking and railing for cracked, broken, or rotting sections. If there is access to the area below the deck, inspect the support framing.

After the deck inspection check the condition of the pool's sidewalls. With a vinyl-lined pool, look for staining, discolorations, or tears in the lining. Faded liners are more likely to tear. Has any portion of the lining pulled out of the edge retainer? If there are stains in the liner at the bottom of the pool that don't continue up the sides, it may indicate a fungal growth in the sand below the liner. With tile-lined concrete walls, look for cracked, loose, chipped, or missing tiles. If the walls are finished with plaster, check for spalling (flaking or chipped) or cracked sections. With painted walls, look for flaking or faded paint and cracks. As you walk around the pool looking at the sidewalls, check the skimmer. Does the weir move freely? Is it broken or missing? Is the strainer basket in place? Is it damaged?

Before you inspect the pump, filter, and heater, look closely at the pool water. If it is turbid, it would indicate that the filter is in need of a cleaning. If there are tiny bubbles discharging at the water-supply outlet, it would indicate that air is getting into the system. This condition must be corrected to prevent possible damage to the pump. Check the cover to the pump's hair and lint strainer. Most covers are transparent. If bubbles are visible, it would normally indicate that the cover is not tightly secured or that the gasket needs replacement. However, if there are no bubbles, you should suspect a leak in the water circulation system. In this case, further investigation is required by a pool service company.

When inspecting the pump, listen for any abnormal sounds. If it is noisy, there may be impeller damage or worn bearings. A pump that is very hot to the touch is also a problem condition. It could indicate that the pump is running dry. If there is water dripping around the pump, it may indicate leakage around the shaft because of worn seals or because the hair and lint strainer cap is not tightly secured. Check the pressure gauge on the filter. Is it inoperative? Is the pressure excessively high? If the heater is not on, turn it on. If it doesn't fire up, record it on your worksheet. If the heater cycles on and off, it may indicate low water flow, a condition that must be corrected. To whatever extent the burner area is visible, look for a scale buildup with rust flakes and dust. Is water dripping in that area? If it is, you should suspect pin-hole leaks. If this condition exists, have the heater inspected by a pool service company. If there is an underwater pool light, check to see if it is operational. Also if there are electrical outlets nearby, check to see if they are GFI protected.

If the pool is not being used and is covered over, check the condition of the cover. Is it worn, are there any torn or damaged sections? If the cover is held in place with straps, are any straps missing? Since the operation of the pool equipment and the condition of the pool shell cannot be checked at this time, you should contact the company that has been servicing the pool and ask them for a copy of their service record. If the record is not available, have the seller include a clause in the contract to guarantee that the pool shell and equipment are in good operating condition.

Checkpoint summary

General considerations

❍ Has a Certificate of Occupancy been issued for the pool?
❍ Is there a fence that encloses the swimming pool?
❍ Are fence gate(s) self-closing and self-latching?

Deck

❍ Are there any cracked, chipped, or settled sections?
❍ Look for uneven joints between sections.
❍ Cracked or open joints should be sealed.
❍ Are grab rails, ladder, and slide adequately anchored to deck?
❍ Is diving board cracked or warped?

Vinyl-lined pool

❍ Is lining stained, discolored, or torn?
❍ Has lining pulled out of edge retainer?
❍ Is bottom of liner stained?

Concrete pool

❍ Are there cracked, chipped, loose, or missing tiles?
❍ Does the plaster finish have flaking, chipped, or discolored sections?
❍ Are any cracks visible in the sidewalls?
❍ Is the painted surface flaking or faded?
❍ Check the skimmer weir and strainer basket.

Pool equipment

❍ Is pool water turbid?
❍ Are tiny bubbles discharging into pool?
❍ Is the pump noisy? Is it very hot to the touch?
❍ Check for water dripping around the pump.
❍ Is the pressure gauge on the filter operative?
❍ Does heater cycle on and off?
❍ Look for scale buildup with rust flakes and dust around heater.
❍ Is water dripping under or around the heater?
❍ Does pool cover have torn or damaged sections?
❍ Check the operation of underwater pool light(s).
❍ Are electrical outlets GFI protected?

19

Energy considerations

Energy audit 283
Insulation 284
Attic ventilation 287
Storm windows 288
Caulking and weatherstripping 289
Fireplaces and wood-burning stoves 289
Heating and air-conditioning systems 290

Many states have enacted an energy conservation construction code to supplement their building codes. The purpose of the energy conservation code is to provide a construction standard that will minimize energy consumption in a building while maintaining the necessary comfort factors. The conservation code applies to new construction, renovations, and additions to buildings. The code is not retroactive and therefore does not apply to buildings constructed prior to its enactment. You can check with the local building department to determine whether there is an energy conservation construction code in your state and if so, whether it was in effect prior to the construction of the house you are planning to buy.

If you are considering the purchase of an existing building, there is a high probability that the house is not as energy efficient as it could be. Although the energy-deficient items are usually found during a prepurchase home inspection, they are often not upgraded until the buyer takes possession of the house. Consequently, after you move into the house, you should perform an energy audit to bring back into focus those items that are needed for conservation improvements.

Having an energy-efficient house is not only good citizenship; it is also good for your pocketbook, resulting in reduced utility bills. The actual dollar savings will of course depend on the gas, oil, and electricity rates in your area; the climate; and the extent to which your house is already energy efficient. You can often determine the projected savings and payback period for the costs involved in making your home energy efficient by contacting your local utility company. For a nominal fee, many utility companies will analyze your energy-conserving improvements, taking into account current and projected energy costs, and will estimate your dollar savings per year.

Energy audit

An energy audit is an inspection of the house to determine the extent of deficiencies that result in energy being wasted. The decision whether to upgrade deficiencies is usually based on economics. Are the dollars spent in making energy-conservation improvements a wise investment? Will the improvements save you enough money on heating and cooling to pay for themselves? For most homes, the answer is yes, especially with ever-increasing costs for fuel.

One cause for wasted energy in a house, especially an older one, is the lack of adequate insulation. Insulation is a basic energy saver and should be used in all houses regardless of location. In colder climates, it reduces heat loss and thereby reduces fuel costs. In warmer climates, the insulation reduces heat gain and consequently reduces cooling costs (electricity is used for running most residential air-conditioning units). During your energy audit, you should determine whether your house is adequately insulated.

Insulation

Insulation is available in a variety of forms and materials. The three most common forms are flexible insulation, loose-fill insulation, and rigid insulation. *Flexible* insulation is manufactured in two types, batts and blankets. Both are made of fibrous materials such as glass fibers, rock wool, wood fibers, or cotton. Organic fibers are treated chemically to make them resistant to fire and decay. Batts are precut in 4- or 8-foot lengths and are available in thicknesses between 2 and 6 inches. Blankets are furnished in continuous rolls and are available in thicknesses between 1½ and 3 inches. Both batts and blankets are manufactured in 15- and 23-inch widths so that they can be readily used in homes that have been constructed with joist and stud spacing of 16 or 24 inches.

Loose-fill insulation is generally made from rock wool, glass fibers, vermiculite, pearlite, cellulose, granulated cork, shredded redwood bark, sawdust, or wood shavings. It is normally supplied in bags or bales and can be poured, blown, or placed by hand. Loose-fill insulation is suited for use in the sidewalls of homes that were not insulated during construction or between the floor joists of unheated attics. However, if there is no floor covering, it is not recommended for use between the floor joists when an attic fan could blow the loose material around.

Rigid insulation is generally made from extruded polystyrene, polystyrene bead board, urethane, fiberglass, or wood fiberboard. It is often used to insulate masonry walls and comes in widths of 24 and 48 inches. Most rigid insulation boards are not fire resistant and should be covered with at least ½-inch gypsum wallboard to ensure fire safety. Rigid insulation boards are also used as backer boards for aluminum and vinyl exterior siding.

Another type of insulation that was popular in the late 1970s was *foamed-in-place* insulation. It was made from urea formaldehyde. In many homes, the foam ingredients were improperly mixed and installed. This resulted in excessive formaldehyde vapor being released into the house, causing adverse health effects. Urea formaldehyde foam insulation (UFFI) is no longer being installed; however, many homes have UFFI in their walls. This is no longer considered a problem. See "Formaldehyde" in chapter 20.

One measure of the effectiveness of insulation is its resistance to heat flow, the *R-number*. The higher the R-number, the greater the resistance to winter heat loss or summer heat gain. Table 19-1 shows typical R-numbers for various types and thicknesses of insulation.

R-numbers are additive. You can add an insulation rated R-11 to one that is rated R-19 to achieve a resistance value of R-30. The thermal resistance of an area covered with loose-fill or flexible insulation can change over the years. The insulation value depends not only on the

Table 19-1. Insulation ratings.

	R-number					
Insulation type	**11**	**13**	**19**	**22**	**30**	**38**
Batts/blankets						
Fiberglass	3½″	4″	6″	7″	9½″	12″
Rock wool	3″	4″	5½″	6″	8½″	11″
Loose-fill						
Fiberglass	5″	5½″	8½″	10″	13½″	17″
Rock wool	4″	4½″	6½″	8″	10½″	13″
Cellulose	3″	3½″	5½″	6″	8½″	11″
Vermiculite	5″	6″	9″	10″	14″	18″
Rigid board						
Polystyrene (extruded)	3″	3½″	5″	5½″	7½″	9½″
Polystyrene (bead board)	3″	3½″	5½″	6″	8½″	10½″
Urethane	2″	2″	3″	3½″	5″	6″
Fiberglass	3″	3½″	5″	5½″	7½″	9½″

material but also on the amount of trapped air contained within the material. If the loose fill is disturbed or the flexible insulation crushed (because of items being stored on top of it), it will no longer be as thick as when it was installed. Consequently, its effective R-number will be reduced. To determine the current R-number of the insulation in your home, you should measure its thickness.

The insulation recommended for your house can be determined from the map in FIG. 19-1. You might be surprised to learn how much insulation is recommended. The R-numbers, however, are based on current and projected fuel costs. If your house is already insulated, once you determine the amount of existing insulation, you can add the difference. Remember, the R-numbers are additive. In some homes, it might not be economically justifiable to increase the insulation to the recommended value.

In determining whether your house is adequately insulated, you should check the exterior walls and the ceilings and floors that face unheated areas, such as the attic and crawl space. (See FIG. 19-2.) In unfinished areas where the insulation is exposed (often in the floor of an attic or ceiling of a crawl space), the thickness can easily be measured. If the attic floor is covered, you can pry up one board and look for insulation. Determining the amount of insulation in a finished exterior wall is more difficult. However, you can make a quick determination whether the wall is inadequately insulated or has no insulation by feeling the inside surface during the heating season. If the wall feels cold to the touch, insulation is needed.

Sometimes you can determine the amount of insulation by removing the cover to a light switch and peering into the wall space, using a flashlight. Caution should be observed, since the switch is electrically hot. Because of the small amount of open space between the wall and switch box, this method is usually not effective. Also, it is possible that the electrician who installed the wiring might have pulled the insulation away from the switch and outlet boxes to facilitate installation. In this case, you might think that there is no insulation in the wall. The only way to determine positively how

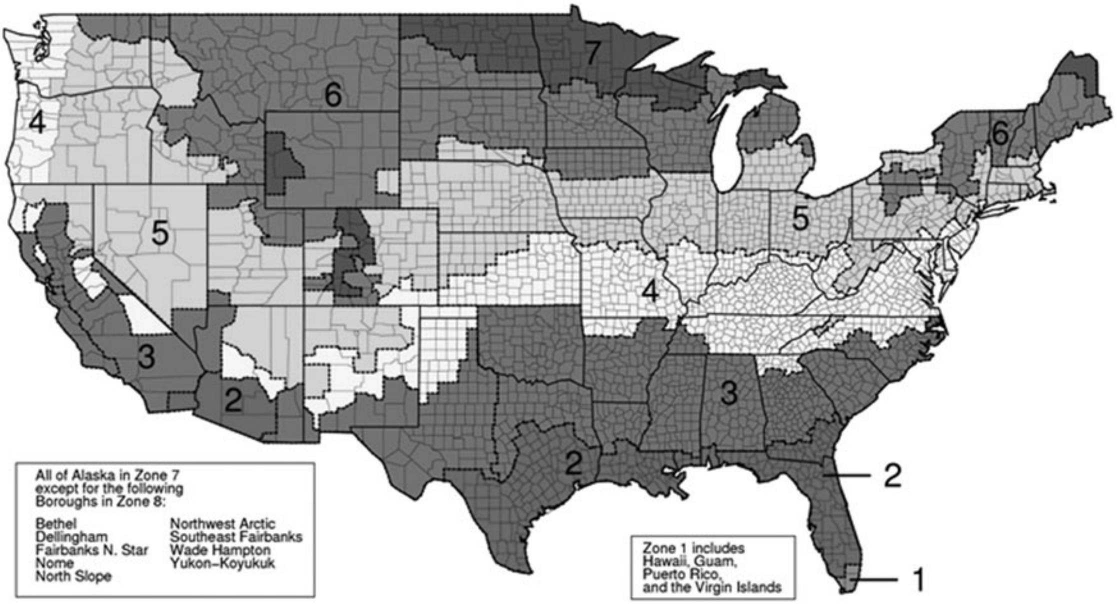
7
6
4
5
3
2
2
4
5
3
6
2
1
All of Alaska in Zone 7 except for the following Boroughs in Zone 8:
Bethel
Dellingham
Fairbanks N. Star
Nome
North Slope
Northwest Arctic
Southeast Fairbanks
Wade Hampton
Yukon–Koyukuk
Zone 1 includes Hawaii, Guam, Puerto Rico, and the Virgin Islands

Zone	Add Insulation to Attic		Floor
	Uninsulated Attic	Existing 3–4 Inches of Insulation	
1	R30 to R49	R25 to R30	R13
2	R30 to R60	R25 to R38	R13 to R19
3	R30 to R60	R25 to R38	R19 to R25
4	R38 to R60	R38	R25 to R30
5 to 8	R49 to R60	R38 to R49	R25 to R30

Fig. 19-1. *Recommended insulation levels for retrofitting existing wood-framed buildings. Courtesy of: Energy Star (U.S. Dept. of Energy and EPA)*

much insulation there is in a finished exterior wall is to make a small hole in the wall (in a nonobvious location such as a closet) and measure it. The hole can then be patched.

While you are determining the amount of insulation, you should also determine whether there is a *vapor barrier* associated with the insulation. A vapor barrier is a thin sheet material such as polyethylene film, aluminum foil, or an asphalt-impregnated kraft paper through which water vapor cannot readily pass. Many insulation materials are produced with a vapor barrier applied on one side. If the insulation does not have a vapor barrier, a separate one can be installed. The purpose of a vapor barrier is to prevent moisture problems in exterior walls and ceilings, and floors that face unheated areas, due to condensation of water vapor (normal in a house) that passes through those surfaces. To be effective, the vapor barrier must be facing the heated room rather than the cool, unheated area. (See FIG. 19-3.)

In addition to the need for insulation of the building shell (exterior walls, ceilings, and floors), all hot-water pipes and heating and cooling ducts that pass through unheated

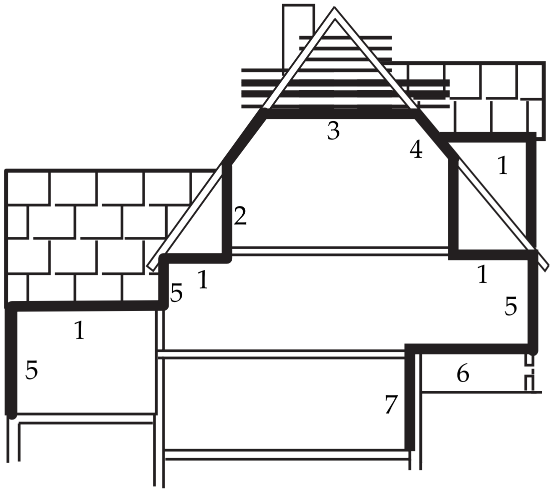

1. Ceilings below an unheated area.
2. "Knee" walls of a finished attic level room.
3. Floor of a crawl attic.
4. The sloping portion of the roof in a finished attic. Leave an airspace between insulation and roof.
5. Exterior walls.
6. Floors above cold crawl spaces. Floors above a porch or an unheated garage.
7. Walls of a heated basement.

Fig. 19-2. *Where to insulate.*

portions of the house (such as a crawl space, garage, or unfinished attic) must be insulated. Most houses usually have no more than 1 or 2 inches of insulation wrapped around ducts in unheated areas. Because of increasing fuel costs, this is considered minimal for most areas, and additional insulation can usually be justified. Check the condition of the insulation. Are there any loose, torn, or missing sections? Also, if there are any exposed duct joints, check them to see if they are sealed tightly. When the ducts are used exclusively for air conditioning or serve a dual function (such as heating and air-conditioning), the outside of the insulation should be covered with a vapor barrier to prevent condensation. A vapor barrier, however, is not needed on ducts used only for heating. If there is a vapor barrier on the ducts, check its condition. Look for torn and missing sections. All vapor-barrier joints must be tightly sealed.

If the domestic hot water is produced in a tank-type water heater located in an unheated area, the tank should be covered with an insulation jacket. These jackets can be purchased in most building-supply or hardware stores. Although tank-type water heaters are normally insulated by the manufacturer, by installing an outer insulation jacket, you will further reduce heat loss and thereby minimize the energy needed to maintain the desired water temperature. The temperature of the hot water should not exceed 140° F. (See chapter 16.) Temperatures in excess of 140° F are not only wasteful of energy but will also shorten the life of the water heater.

Attic ventilation

While in the attic checking the insulation, see if the area is adequately ventilated. As discussed in chapter 9, attic ventilation is necessary not only to prevent condensation problems during the winter months but also from an energy-conservation point of view to reduce the heat load on the structure during the summer. Because of trapped air, attic areas can become excessively hot during the summer, reaching temperatures of about 150° F. If there is an air-conditioning blower coil located in the attic, the high temperature will tend to lessen its efficiency of operation.

Ventilation in an attic should be provided by at least two vent openings located so that air can flow in one opening and out of the other. Vents in the eaves and at gable ends are better than gable vents alone. One of the most effective methods for ventilating the attic is a combination of vents in the eaves and a continuous ridge vent. Remember, the attic must also be adequately

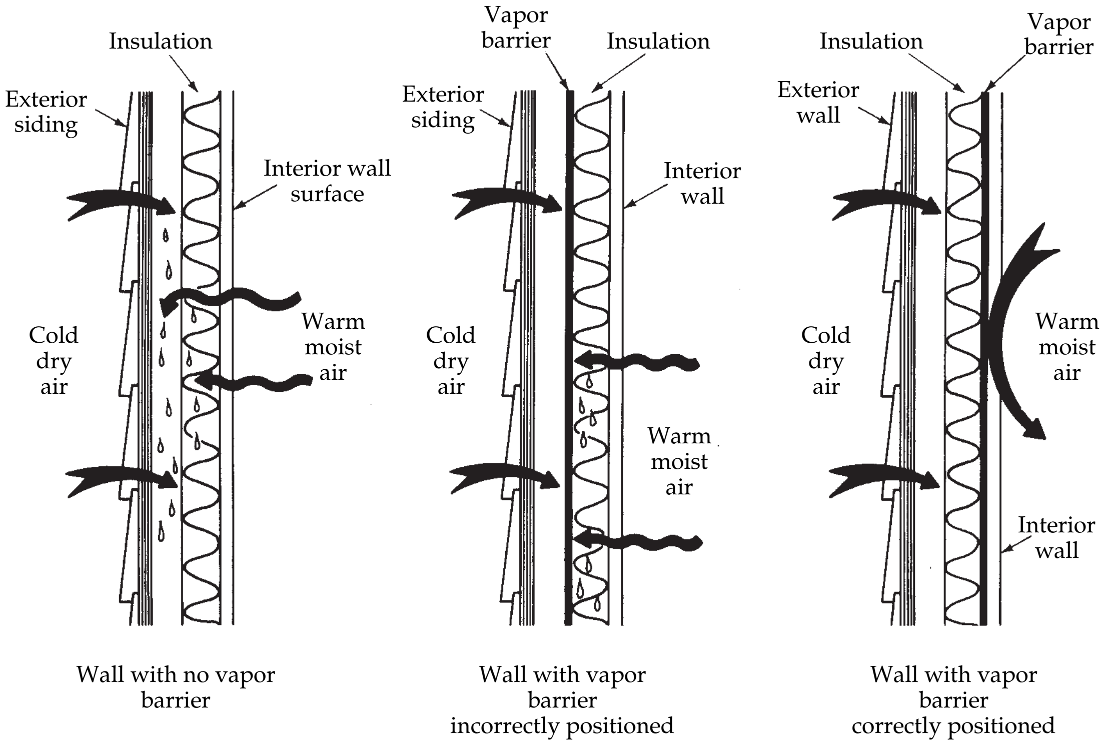
Insulation
Exterior siding
Interior wall surface
Cold dry air
Warm moist air
Wall with no vapor barrier
Vapor barrier
Insulation
Exterior siding
Interior wall
Cold dry air
Warm moist air
Wall with vapor barrier incorrectly positioned
Insulation
Vapor barrier
Exterior wall
Cold dry air
Warm moist air
Interior wall
Wall with vapor barrier correctly positioned

Fig. 19-3. *How a correctly positioned vapor barrier prevents condensation in the exterior wall.*

ventilated during the winter. As long as the attic is adequately insulated, the benefits of ventilation greatly exceed any fuel savings that might result from blocking the vent openings.

Storm windows

Most homes with single-pane windows will benefit by the installation of storm windows. (See chapter 5.) A storm window, whether a storm sash, panel, or combination unit, reduces the heat loss through a single-glazed window by about 50 percent. In cold climates, it also adds to physical comfort by reducing an apparent draft. Body heat radiates toward a cold surface. Also, warm air hitting a cold surface loses its heat, becomes dense, and falls to the floor. This combined effect creates what appears to be a draft to someone sitting or standing near the cold surface (window). If the storm window is a sash or combination unit (both of which cover the window frame), it also helps reduce cold-air infiltration through the movable and fixed joints around the window.

Basically, a storm window is effective because it traps a layer of air between itself and the window. This dead air space acts as an insulator and thus reduces heat loss. If installing storm windows is not economically justified because you are planning to move or your budget does not presently permit the purchase of storm windows, you can still make your windows energy efficient. Simply cover them with plastic sheets and secure the edges with tacks, molding strips, and caulking. Although temporary, these inexpensive homemade storm windows are an effective approach to reducing heat loss. They will more than pay for themselves in fuel savings in the first year.

Caulking and weatherstripping

In a well-insulated house with thermal-pane or storm windows, air leakage is the greatest source of heat loss. To conserve energy further by reducing the cold-air infiltration during the winter and the loss of air-conditioned air in the summer, the movable joints (such as those around windows and doors) should be weatherstripped, and the fixed exterior joints should be caulked. (See chapter 5 for types of caulking compounds.)

Check windows and doors for weatherstripping and tightness of fit. Loose-fitting windows and doors not only lose heat but result in uncomfortable drafts. All window sashes, exterior doors, and interior doors or hatches leading to unheated areas (such as an attic, basement, or crawl space) should be weatherstripped on their sides, tops, and bottoms. Periodically check the condition of the weatherstripping. Over the years, some types of weatherstripping will wear, tear, crack, and generally deteriorate so that replacement is required.

The condition of the caulking on the exterior joints should be checked during your energy audit. Specifically, look at the joints (1) between the exterior siding and the window and door frames, (2) at the inside and outside corners formed by the exterior siding, (3) just under the bottom side of the exterior siding and the foundation wall, (4) between dissimilar siding materials (such as a masonry-and-shingle wall), (5) where the chimney meets the siding, and (6) where storm windows meet the window frame (except for drain holes at the windowsill). If the caulking is old, brittle, broken, or missing, the joints should be recaulked. Good home maintenance includes an annual check of all exterior joints, particularly in view of rising fuel costs.

Fireplaces and wood-burning stoves

Most fireplaces are used for creating a relaxed, cozy atmosphere rather than heating. Usually their use as an auxiliary source of heat cannot be economically justified, since they are very inefficient. A fireplace requires a large volume of air for combustion. Normally this air is drawn from the heated air in the house. Since the combustion air flows up the chimney, all the fuel that had been used in heating the air is wasted. To minimize this problem, a fireplace should have dampered air vents that provide combustion air from the outdoors. Although newer fireplace installations have provisions for outside air, most existing fireplaces do not have this feature and therefore waste more heat than they generate. In fact, in terms of the total heating value of the wood being burned, a fireplace does not generate much heat at all. Only about 10 to 15 percent of the available Btus actually find their way out into the room, while much more warm air disappears up the chimney.

Even when a fireplace is not being used, it is a potential source of heat loss. If the flue is not blocked, heat in the house will be drawn up the chimney. This condition can normally be minimized by closing the damper. Check the fireplace to see whether there is a damper and if so, whether it properly seals the flue opening. Fireplaces in many older houses do not have dampers. If a movable damper is too costly to install, you can always block the flue with a piece of sheet metal supported by guides on the sidewalls. It can be inserted and removed on an as-needed basis.

Some fireplaces have built-in convection ducts around the firebox that function as a heat exchanger. Cool air entering the bottom of the ducts is heated as it circulates around the firebox. The warm air rises and flows out into the room through openings in the upper section of the ducts. This type of fireplace has a greater heating efficiency than the conventional open-faced masonry fireplace. Nevertheless, a much more efficient way to heat the house is by using a wood-burning stove. A good-quality stove will allow only the air needed for combustion into the unit. Consequently, heated room air is not lost up the chimney. If you control the airflow through the stove, a log will produce more heat by burning for a longer period of time. In addition, since the stove is physically located in the room that it is intended to heat, the entire surface of the unit will heat up and warm the room by convection and radiation.

The cost-effectiveness of a wood-burning stove for heating will depend on the cost of the wood being burned. In those areas where wood is inexpensive, a stove is an economically viable method for providing auxiliary heat and reducing your overall heating bill. Even in those areas where wood is more costly, as utility rates increase, the use of a wood-burning stove becomes more attractive. Different woods have different burning characteristics. For example, wood from conifer trees (softwood) such as pine, spruce, and fir burns more quickly and gives less heat than wood from deciduous trees (hardwood) such as maple, oak, and beech. To aid you in making a proper selection for your fireplace or stove, TABLE 19-2, prepared by the Maine Bureau of Forestry, shows the various characteristics of wood.

Heating and air-conditioning systems

During your energy audit, you have been mainly concerned with those items needed to reduce heat loss. Another item of concern is maximizing the Btu output of your heating system. As you walk around the house, check the radiators or heat registers to see that they are unobstructed. If your house is heated with a warm-air furnace, check the filter to see if it needs replacement. If the oil or gas burner for the heating system was not cleaned and tuned up prior to your energy audit, it should be done as soon as possible. A heating system out of adjustment results in

Table 19-2. Characteristics of woods for use in a fireplace or stove.

Species	Ease of starting	Ember generation	Sparks	Fragrance	Heating value
Apple	Poor	Excellent	Few	Excellent	Good
Ash	Fair	Good	Few	Slight	Good
Beech	Poor	Good	Few	Slight	Excellent
Birch (white)	Good	Good	Moderate	Slight	Good
Cherry	Poor	Excellent	Few	Excellent	Good
Cedar	Excellent	Poor	Many	Good	Fair
Elm	Fair	Good	Very few	Fair	Good
Hemlock	Good	Low	Many	Good	Fair
Hickory	Fair	Excellent	Moderate	Slight	Excellent
Locust (black)	Poor	Excellent	Very few	Slight	Excellent
Maple (sugar)	Poor	Excellent	Few	Good	Excellent
Oak (red)	Poor	Excellent	Few	Fair	Excellent
Pine (white)	Excellent	Poor	Moderate	Good	Fair

greater fuel consumption, which can be quite costly. Heating systems require a periodic tune-up for efficient operation.

Various devices on the market to be fitted on or around the chimney flue are supposed to cut fuel consumption and conserve heat. Many of these furnace attachments, however, do not live up to the manufacturers' claims of saving significant energy.

One device of concern from a safety point of view is the *automatic flue damper*. This device is designed to close the furnace chimney after the burner turns off. The intention is to conserve heat by preventing room air from being drawn up the chimney. However, if the automatic flue damper fails to open when the burner fires, poisonous carbon monoxide fumes will vent into the house. If your furnace has such a device, you should find out whether it has been approved by a nationally recognized testing agency and by your local utility company. If it has not been approved, you should consider its removal or replacement as a precautionary measure.

Table 19-3. Energy savings for lowering thermostat.

	Percent of energy savings for an 8-hour temperature setback	
Degree days	**5° F**	**10° F**
5,000	8.1	12.1
6,000	7.2	10.8
7,000	6.1	9.6
8,000	5.2	8.5
8,500	4.6	7.8

By lowering the thermostat setting 5 to 10 degrees each night before going to bed and raising it in the morning, you will reduce fuel consumption and save dollars. The amount of the savings will depend on the duration of the setback, the climate, and the fuel costs in your area. See TABLE 19-3.

The *degree day* is a unit that expresses the severity of the climate in an area. The reference temperature for evaluating degree days is 65° F. Degree days are the number of degrees that the average (of the high and low temperatures for a 24-hour period) is less than 65° F. For example, if for a twenty-four-hour period (during the heating season), the high and low temperatures are 50° F and 30° F, respectively, the average temperature would be 40° F. The degree-day number for that day is then 65–40, or 25. The total number of degree days for an area, therefore, is simply the sum of the degree-day numbers during the heating season. You can check with the local utility company to find the total number of degree days for your area.

From a convenience point of view, if your heating system is controlled by a manual thermostat, you should consider replacing it with an automatic-lock thermostat. With this type of thermostat, you can regulate the amount of the setback and its duration. Depending on your requirements, thermostats with double setbacks are also available.

If there is a central air-conditioning system in your house, it must also be cleaned and tuned up to maximize the efficiency of operation. Prior to the cooling season, have the system checked to see if it needs a refrigerant charge. The system will cool even if it is low in refrigerant; however, it will operate inefficiently. Check the location of the compressor to see if it is in the shade and whether the airflow into and out of the unit is unobstructed. If the compressor is in the midday and afternoon sun, it will not perform efficiently. You should build a sunscreen to shade the unit if necessary. Be careful, however, not to obstruct the airflow.

20

Environmental concerns

Radon 292
Asbestos 295
Drinking water 297
Lead 298
Formaldehyde 300
Leaky oil tanks 300
Electromagnetic fields 301
Mold 302

In the past, whenever people bought a home, their main concern was the physical house; that is, the structural integrity of the building; the condition of the mechanical equipment such as the heating system, plumbing, water heater; the adequacy of the electrical system; the condition of the roof; whether the basement was dry; and whether there was a termite condition. In recent years, another factor has entered into the decision process—environmental problems. Some of these problems, such as a high radon concentration or deteriorating asbestos insulation, are potential health hazards. Others, like a leaky buried oil tank, can contaminate the soil and eventually the aquifer (water table). In all cases, it costs money to correct the problems. This cost should be added to the overall sale price of the house to determine the true cost of purchasing the house.

In this chapter, I will discuss environmental problems that are or should be of concern to the home buyer or homeowner. There is no doubt that in the future, as technology improves and more statistical health information becomes available, additional items will be added to the list of environmental problems.

Radon

Although the health risks associated with exposure to high concentrations of radon have been known for decades because of experience with uranium miners, it wasn't until December 1984 that it was realized that people in homes can also be exposed to high concentrations of radon resulting from uranium deposits in the soil on which the houses are built. A worker in a nuclear generating plant passed through a radiation detection monitor as he entered the plant. It turned out that his home had twenty times more radiation than is allowed in a uranium mine.

Radon is a gas present in varying quantities in the atmosphere and soils around the world. It is colorless, odorless, and tasteless, and is produced by the natural radioactive decay of uranium deposits in the earth. Prolonged exposure to high concentrations of radon can cause cancer. According to the U.S. Environmental Protection Agency (EPA), scientists estimate that between 5,000 and 20,000 lung-cancer deaths a year in the United States can be attributed to radon.

The concentration of radon in the air is measured in units of *picocuries* per liter of air (pCi/l). Based on currently available information, the EPA has set guidelines for radon levels in residential structures. (See TABLE 20-1.) Their "action level" is 4 pCi/l; that is, no action is needed if the radon concentration is below 4 pCi/l. The EPA acknowledges that even the action level is not risk free because exposure to an annual radon level of 4 pCi/l is equivalent to the risk of smoking ten cigarettes a day or having two-hundred chest X-rays a year. The action level was set at 4 pCi/l because in some cases reduction below that level might be difficult or impossible to achieve.

Table 20-1.
U.S. EPA guidelines for residential radon.

Annual average Radon level	**Recommended action**
Over 200 pCi/l	Take action to reduce levels within several weeks.
Between 20 and 200 pCi/l	Take action to reduce levels within several months
Between 4 and 20 pCi/1	Exposures in this range are considered above average You should take action to reduce levels to about 4 pCi/l or below within a few years (sooner if levels are at the upper end of this range).
Below 4 pCi/l	Exposures in this range are considered average or slightly above average. Although exposures in this range do present some increased risk of lung cancer, reductions of levels this low may be difficult, and sometimes impossible, to achieve.
Below 1.0 pCi/l	These are average first-floor residential levels.

Although radon gas is present in varying quantities in soils around the world, not every house has a problem with high radon levels. It is possible for one house to have a very low radon concentration while an adjacent house has a very high level. It depends on the construction of the house, the uranium-radium content in the soil, and geological formation below the house. The only way to know if a house has a radon problem is to test it. The radon concentration in a house varies with time. It is affected by a number of environmental factors such as rain, snow, barometric pressure, and direction of wind relative to open windows, and by induced negative pressures caused by periodic use of exhaust fans, attic fans, fireplaces, and heating systems. Consequently, the most accurate method of determining the average annual radon concentration is a long-term test. However, since time is usually limited in real estate transactions, a long-term test is not practical, and consequently a charcoal canister with a test period of three to seven days is generally used.

Canisters can be purchased in hardware stores and home centers. Also, in many areas professional home-inspection companies offer radon-testing services. If you intend to place the radon-detection canister yourself, make sure that you follow the directions carefully; otherwise, the results might be inaccurate.

Since radon is a gas, it can seep into the lower level of a house through dirt floors, cracks in concrete floors and walls, floor drains, sump pits, open joints and tiny cracks or pores in hollow block walls. (See FIG. 20-1.) If after a house is tested it is determined that it has a high radon concentration, don't panic. The condition can be corrected at a reasonable cost.

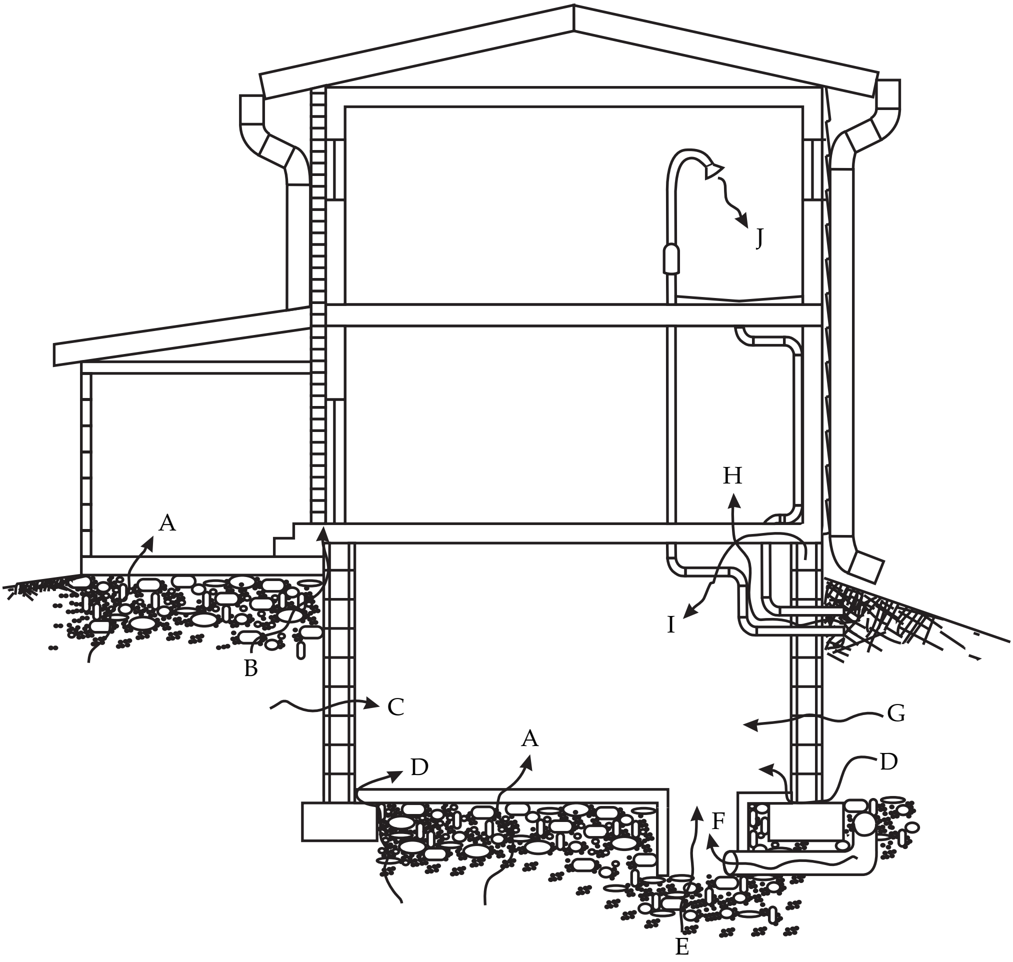

A. Cracks in concrete slabs
B. Spaces behind brick veneer walls that rest on uncapped hollow-block foundation
C. Pores and cracks in concrete blocks
D. Floor-wall joints
E. Exposed soil, as in a sump
F. Weeping (drain) tile, if drained to open sump
G. Mortar joints
H. Loose fitting pipe penetrations
I. Open tops of block walls
J. Shower

Fig. 20-1. *Major radon entry routes.*

A number of methods have been successful in reducing radon concentration levels in buildings to a point below the "action level." A typical mitigation procedure that is very effective is to use a 4-inch plastic pipe with an in-line fan that vents the radon gas from below the floor slab to above the roof. Correcting the problem is not a do-it-yourself task. It should be done by a radon-mitigation contractor that has completed the requirements for listing under the U.S. EPA's Radon Contractor Proficiency Program.

Radon can also enter the house through the domestic water supplied by a private well. It can be transferred into the air during a shower or when water is running in a sink. However, it takes relatively high levels of waterborne radon to produce a significantly elevated level of radon in a house; the large volume of air inside a house dilutes the radon being transferred to the air from the relatively small volume of water. It is estimated that the normal use of a water supply containing 10,000 pCi/l will produce a concentration of 1 pCi/l of radon in the air. It therefore takes a waterborne radon level of 40,000 pCi/l to reach the EPA's action level of 4 pCi/l.

In most parts of the country, radon gas emanating from the soil is the major contributor to indoor airborne radon. As a result, the water from a private well is normally not tested for radon during the initial screening test. However, in some areas of the United States, waterborne radon significantly contributes to the total radon concentration. You can check with your local health department to find out if testing the water is recommended for your area. As of this writing, the EPA has not set a maximum level for radon in drinking water. Nevertheless, if the radon level is greater than 10,000 pCi/l, water-supply mitigation should be considered. There are water-treatment methods available that can be used for removing radon at the point of entry. The specific method should be discussed with the radon-mitigation contractor. Incidentally, do not worry about drinking the water. The health effect of drinking waterborne radon is relatively insignificant compared to that associated with breathing in airborne radon.

Asbestos

A common concern of many home buyers is whether there is any asbestos in the house, and if there is, is it in a condition that would be considered a health hazard? Asbestos, which has been identified as a carcinogen, is a naturally occurring fibrous mineral found in certain types of rock formations throughout the world. Asbestos fibers are strong, won't burn, resist corrosion, and insulate well. These physical properties have made it a staple in architectural and construction applications. When the fibers are mixed during processing with a material which binds them together, they can be used in many construction products such as cement siding and roof shingles, vinyl floor tiles, ceiling tiles, textured paints or coatings, blown-in insulation, flexible fabric connections in ductwork, spackling compounds, boiler insulation, pipe insulation, caulking, putties, door gaskets on wood-burning stoves, and so on. (See FIG. 20-2.) The amount of asbestos in these products varies considerably, from approximately 1 percent to 75 percent.

Asbestos-containing material in the home doesn't necessarily pose a health risk. Asbestos materials become hazardous only when due to damage, disturbance, or deterioration over time, they release fibers into the air. Airborne asbestos fibers can be inhaled through the nose and mouth and lodge in the lungs. According to estimates by the EPA, every year 3,300–12,000 people die from cancer caused by exposure to asbestos. Of particular concern is asbestos-containing material that is sprayed or troweled or that has become friable. (Friable material can be crumbled, pulverized, or reduced to powder by hand pressure.) As long as the asbestos-containing material is intact, it does not pose a health hazard. If the asbestos material is not likely to be disturbed or is in an area where renovations will not occur, the EPA suggests that it is best left alone.

Many houses have old boilers that are insulated with asbestos. The insulation looks like a white plaster coating over the boiler shell. When the boiler is eventually replaced, it is

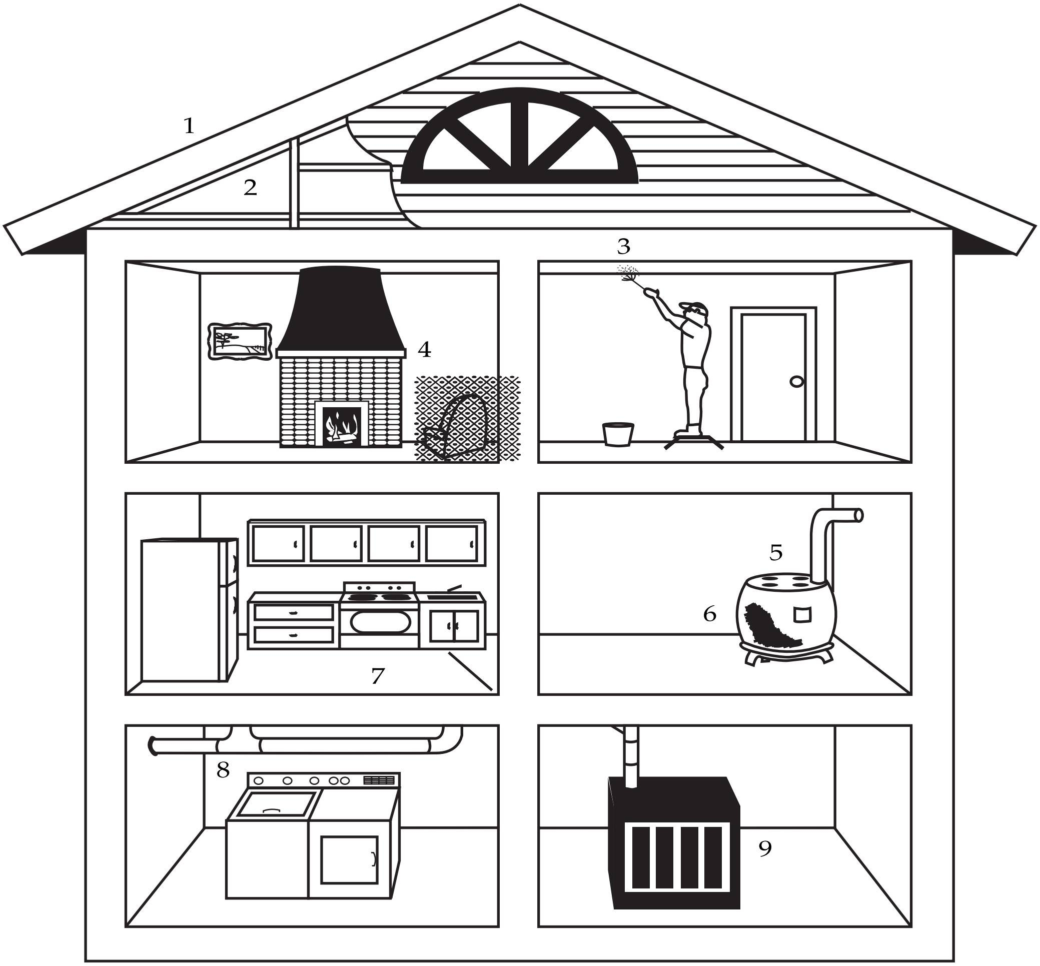

1. Some roofing and siding shingles are made of asbestos cement.
2. Houses built between 1930 and 1950 may have asbestos as insulation.
3. Asbestos may be present in textured paint and in patching compounds used on wall and ceiling joints. Their use was banned in 1977.
4. Artificial ashes and embers sold for use in gas-fired fireplaces may contain asbestos.
5. Older products such as stove-top pads may have some asbestos compounds.
6. Walls and floors around woodburning stoves may be protected with asbestos paper, millboard, or cement sheets.
7. Asbestos is found in some vinyl floor tiles and the backing on vinyl sheet flooring and adhesives.
8. Hot water and steam pipes in older houses may be coated with an asbestos material or covered with an asbestos blanket or tape.
9. Oil and coal furnaces and door gaskets may have asbestos insulation.

Fig. 20-2. *Locations of possible asbestos hazards in the home.*

necessary first to remove the asbestos insulation, even if the insulation is in good condition, before disassembling the boiler. Most heating contractors are not certified in asbestos removal. The job must be done by a qualified asbestos-removal contractor that has been trained and certified. Once removed, the asbestos cannot be thrown out like ordinary household garbage. A licensed industrial-waste hauler must take it, properly identified and contained, to a landfill that accepts asbestos.

Removal of damaged or deteriorating asbestos materials is not always necessary. In fact, it is the least desirable alternative because in the process it creates a considerable amount of airborne asbestos fibers that must be contained and removed. Depending on its condition, the asbestos material can be encapsulated by coating it with a sealant so that the fibers cannot be easily released. It can also be enclosed in airtight walls or ceilings that completely isolate and contain any fibers that become airborne. The decision whether to remove, encapsulate, or enclose deteriorating or damaged asbestos material, as well as the repair of damaged sections, should be made only by a certified trained professional.

Although most of the asbestos materials that had been used in construction are no longer being manufactured, various items can be found in many homes built prior to 1978. Identifying the more common types of asbestos material can generally be done in a visual inspection by home inspectors, asbestos-abatement personnel, or tradespeople who have frequently worked with asbestos material, and even by you. For example, in many older homes, the heating pipes in the basement or crawl spaces are covered with insulation that contains asbestos. The insulation on the straight sections of pipe, when viewed from an end, looks like corrugated cardboard. The angle fittings on the heating pipes are covered with an insulation coating that looks like plaster and is shaped around the fitting. If you see this type of insulation, look for torn, loose, crushed, or otherwise damaged sections. In many cases, it is not readily apparent whether building products and insulation materials contain asbestos. In those cases, positive identification of asbestos can be made by a qualified laboratory after analyzing representative samples of the material.

Drinking water

Domestic water is generally supplied to homes through private wells or public water companies. Water supplied by public water companies is usually safe to drink and does not pose a health risk. The quality of the water supplied by these companies is periodically checked because it must comply with rigid standards set by the U.S. EPA. Nevertheless, the EPA has indicated that some 40 million people have been using drinking water containing potentially hazardous levels of lead. The problem does not originate with the water supply but with distribution piping, solder used at the pipe fittings, and fixtures in the house. In some older homes, the inlet water pipe is made of lead; the solder used on pipe fittings in homes built before 1988 contained lead; and lead is contained in the metal alloy used in the manufacture of many faucets.

The most important factor causing a high concentration of lead in water is the contact time between the water and the lead. Water that is slightly acidic or soft (water that makes soapsuds easily) is corrosive and reacts with lead. When the water stands in pipes or faucets that contain lead for several hours without use, there is a potential for lead to leach, or dissolve, into the water. Also, hot water dissolves lead more quickly than cold water.

Dissolved lead in water has no odor and cannot be seen or tasted. Testing by an approved laboratory is the only way to determine if the drinking water has high levels of dissolved lead. The test is generally conducted in two parts, A "first draw" sample is collected—water that has been sitting in the pipes overnight or at least four hours. Then a "fully purged" sample is collected; the water is turned on and allowed to flow for at least one full minute before a sample is collected. According to preliminary studies at the University of North Carolina, about 30 percent of homes have a high lead concentration in the first draw, but purging corrects the problem more than 90 percent of the time. The current federal standards limit the amount of lead in water to 15 parts per billion (ppb), which is equivalent to 0.015 milligrams per liter (mg/l).

Another domestic water problem is excessive sodium. This condition is the result of using a water softener in the water supply. The softener replaces the calcium in the water with sodium. When the sodium concentration in the water is greater than 28 mg/l, people with high blood pressure and those with low-salt diets should be warned. When using a water softener, the sodium content is usually 100 mg/l or more. The problem can be corrected by having the pipe that supplies water to the kitchen sink bypass the water softener. All drinking and cooking water should then be taken only from the kitchen sink.

In addition to the possibility of high concentrations of lead and sodium, domestic water supplied by a private well can be contaminated by harmful bacteria resulting from faulty septic tanks, chemicals from a toxic spill that occurred years before, leaking underground storage tanks, or pesticides and fertilizers. The only way to tell whether the water is potable is to have it tested. The tests for pesticides and other chemicals are more complex and costly than the routine tests for bacteria or minerals. If you are concerned about pesticide and chemical contamination of your well, first contact your county officials and find out whether contamination problems have been reported in the area. As a precautionary measure, the water from a private well should be analyzed once a year for coliform bacteria to ensure that it is potable.

Lead

Lead poisoning is considered by many public-health officials the number-one environmental threat to children. It dwarfs radon and asbestos. High concentrations of lead in the body can cause permanent brain damage, even death. Low concentrations can result in reading disorders and hyperactivity and can affect a child's ability to perform in school. The government is estimating that one out of every nine children under the age of six has enough lead in his blood to place him at risk. The sad truth is that lead poisoning is entirely preventable.

Lead poisoning occurs in the home. It is not confined to children of low-income families living in the inner-city ghettos but has also been found in children of well-to-do families living in the suburbs. According to the EPA, about two-thirds of the homes built before 1940 and one-third to one-half of the homes built between 1940 and 1960 contain heavily leaded paint. A smaller percentage of homes built between 1960 and 1980 also had surfaces coated with lead-based paint. In 1978, the U.S. Consumer Product Safety Commission (CPSC) lowered the legal maximum lead content in most kinds of paint to trace amounts (0.06%), so houses built after 1978 should be relatively free of lead paint.

It has been known for years that children have been poisoned by eating chips of lead-based paint. However, it was only recently documented that they are also at risk of poisoning from exposure to lead dust in household air. Lead dust particles can become airborne when surfaces covered with lead-based paint are scraped, sanded, or heated with an open flame during paint stripping. Lead dust can also be created by the rubbing and sliding of lead-base-painted window sashes as they open and close.

Lead dust can be inhaled when airborne. When the dust settles on the floor, windowsills, or furniture, it can be ingested by children through their normal hand-to-mouth behavior. Also, settled lead dust particles can become airborne as a result of household cleaning. The particles, which are very fine, can penetrate the filter system of home vacuums and are recirculated in the exhaust airstream. Cleaning of lead dust should be done by a professional who specializes in lead abatement.

The only way to tell whether the paint in a home contains lead is to have samples from different areas tested, such as windowsills, door trim, radiators, banisters, and walls where the paint may be peeling and flaking. Testing by a qualified laboratory is considerably more accurate than using a do-it-yourself kit. Many of these kits aren't very precise, since other metals can cause false-positive results. They are not sensitive to low levels of lead, so that a sample might test negative and still be considered hazardous. Also, the kits cannot tell *how much* lead is in the paint. According to the U.S. Department of Housing and Urban Development (HUD), action should be taken to reduce exposure to lead when the lead content in the paint exceeds 0.5 percent. Measures to reduce exposure to lead are particularly important when the paint is deteriorating or when infants, young children, or pregnant women are present.

If you are planning on buying a house that has lead-painted walls and trim, you have several options not unlike those for asbestos. If it is in good condition and there is little possibility that it will be eaten by children, then leave it undisturbed. If there are damaged or deteriorating sections of paint on the walls or ceilings, you can have those areas covered over with gypsum wallboard or some other building material.

You can also have the lead-based paint removed. This task, however, must be done by professionals trained in removing lead-based paint, since each of the paint-removal methods (sanding, scraping, chemical paint stripping, and heat guns) can produce lead dust or fumes. If not done properly, this option will create a greater health hazard than the original one. In some cases, complete removal and replacement of items such as windows, doors, and wall and door trim might be the best approach because of the cost or difficulty of removing the paint. This task should also be done by professionals, who will control, contain, and remove the lead dust.

If a surface that has been painted with a lead-based paint is intact, painting it with a nonleaded paint is considered a viable method of reducing the hazard associated with lead paint. However, painting the surface is not considered a permanent or long-term solution because the lead-based paint below the top coat might eventually loosen and create lead flakes and dust. Before undertaking any abatement procedures, you should have a qualified lead inspector do a *lead hazard risk analysis* and have him develop an abatement strategy that considers all the options.

Because awareness of the pervasive nature of the lead-dust problem is relatively recent, there is currently a lack of qualified lead inspectors and abatement contractors. However, a number of states are in the process of implementing certification, licensing, or training requirements for abatement contractors. Also, HUD has prepared and published guidelines for identification and abatement of lead-based paint hazards. To find a qualified lead inspector or abatement contractor, check with your state department of health or environmental agency. In those states that have not implemented certification procedures, you might still be able to find qualified inspectors and contractors. Certified lead-abatement training courses are available in Massachusetts and Maryland. A number of companies, including asbestos-abatement contractors from other states, have taken these courses and listed their companies in the telephone yellow pages under the overall heading of "Lead Paint" with subheadings "Inspection" and "Abatement."

Formaldehyde

Although formaldehyde gas had been known for many years to be a powerful irritant that can contaminate the air in a house, it wasn't until the installation of urea formaldehyde foam insulation (UFFI) in houses that state and federal agencies began to focus their attention on it. Exposure to formaldehyde vapor can cause eye, nose, and throat irritation; coughing; skin rashes; headaches; dizziness; and nausea. It has been shown to cause cancer in laboratory animals.

After the oil embargo of 1973, many homeowners, realizing the need for energy conservation, had insulation or additional insulation installed in the exterior walls of their houses. It is estimated that about 500,000 homes were insulated with UFFI, most of which was done in the 1970s. With UFFI, the quality of formulation and installation determines the amount of formaldehyde released. The amount also increases under hot and humid conditions. By the end of 1980, the CPSC received more than 1,500 complaints of adverse health effects associated with the release of formaldehyde gas from UFFI. In 1982, the CPSC banned the use of UFFI in homes; however, in 1983, the ban was overturned by a U.S. court of appeals. By that time, most of the installation contractors were out of business.

At this time, the UFFI is not considered a major source of formaldehyde air contamination. Formaldehyde levels decline rapidly to below 0.1 parts per million (ppm) within the first year of the installation of UFFI. Although people vary in their susceptibility to formaldehyde, most healthy adults do not experience ill effects from exposure below 0.1 ppm. Also, since the UFFI was installed years ago, any vapors from the insulation would probably be negligible.

Notwithstanding the above, there might still be a high formaldehyde vapor concentration in the air. Formaldehyde is found widely in many household and construction products such as plywood, particle board, chipboard, plastic laminates, cosmetics, cleaners, paper products, drapes, carpets, and even tobacco smoke. Symptoms of exposure to formaldehyde generally resemble those associated with a cold or allergies; however, the symptoms usually cease once exposure is discontinued. If you are concerned about excessive formaldehyde concentration in the house, have the air analyzed by a private qualified testing laboratory.

Leaky oil tanks

Out of sight, out of mind. That's exactly how homeowners with oil-fired heating systems think about the buried tank that holds the heating oil. They could be in for a rude awakening. Most homeowners do not know that they are responsible if their oil tank leaks. If a tank leaks and contaminates the surrounding soil, the cost for the cleanup, which includes excavating around the tank, scooping up the oil-saturated soil, and carting it away by a licensed carter, must be paid by the homeowner. This effort can cost thousands of dollars. If the oil leak contaminates an aquifer or a stream, the cost for cleanup can run into tens of thousands.

Most of the buried tanks used to store home heating oil are steel and have a projected life of about twenty years, although the actual life depends on the extent to which water has been accumulating within the tank and on soil conditions. If the oil tank for the house that you want to purchase is that age or older, you have three options. You can ask the seller to provide you with a certification on the integrity of the tank; you can have the tank tested for leaks at your expense; or as a preventive measure you can replace the tank—the most expensive option.

Some homes have buried oil tanks that are no longer in use because their heating system has been converted from oil to gas. If this is the case, ask the seller if the tank has been properly put out of service. If it hasn't, it should be; otherwise, safety and environmental problems can develop. In many communities, properly putting a tank out of service is a legal requirement. An abandoned tank that has not been properly prepared will eventually deteriorate because of corrosion resulting in leakage, collapse, or both. Leakage, of course, contaminates the soil.

To put a tank out of service properly, the residual oil or sludge at the bottom of the tank must be pumped out and disposed of, using a licensed carter. The tank should then be removed entirely or filled with an approved filler such as sand. If the soil below the tank is contaminated, it must be removed. The exact abandonment procedure may vary depending on the municipality. Check with your local department of environmental conservation or health department for the approved method.

Electromagnetic fields

More and more attention by the news media is currently being devoted to electromagnetic fields. Whenever electricity is present, there are both electric fields and magnetic fields. These fields, when taken together, are referred to as electromagnetic fields, or EMF. They exist around electrical transmission lines, transformers on poles, house wiring, appliances, electronic equipment, and so on. An epidemiological study published in 1979 concluded that children living in Denver near certain power lines that were exposed to electromagnetic fields with a magnetic field strength of 2–3 milligauss (mG) were twice as likely to develop leukemia than children living in houses without such exposure. The conclusions of the above study were not universally accepted by the scientific community; some researchers felt that the study was flawed and not scientifically sound. In the early 1980s, the scientific community began looking extensively into whether exposure to electromagnetic fields is dangerous to a person's health. The complexity of the problem is compounded by the fact that the results vary according to the characteristics of the electromagnetic field, such as frequency, intensity, field orientation, period of exposure, and whether the field is continuous or pulsed.

The studies to date have not resulted in definite conclusions. They have shown that there is a statistical association between electromagnetic fields and cancer. However, they have not shown that the fields are involved in causing cancer. Based on all available information, many scientists have concluded that more research is needed before they can understand how EMF interacts with the human body to cause health effects and whether there is indeed a risk from exposure to EMF. Also, since not enough is known about which elements of the exposure are important to health, a safe exposure level has not been established. Both electric fields and magnetic fields are being studied. However, more attention is being focused on magnetic fields because they can penetrate through common objects as thick as concrete walls without losing their strength whereas electric fields can be shielded by materials like wood, aluminum, or insulation around the wires.

How does all of the above affect a person contemplating purchasing a home near power lines or an electrical substation? You can have measurements taken to determine the magnetic-field intensity in and around the house to see if the readings are within typical background levels. A few home inspectors provide this service. Unfortunately, as of this writing, no government regulation or scientific recommendation helps answer the question. Until the research is complete, it is premature to recommend any course of action. Nevertheless, prudent avoidance should be considered, because until the cloud of uncertainty is removed, this problem can affect the resale value of these homes.

Mold

Mold growth in a house is the latest environmental concern. *Mold* is the common name for simple plants or fungi, usually microscopic in size, which are naturally occurring and produce spores that can be found everywhere indoors and outdoors. It has been estimated that there are at least 40,000 known species of fungi. Molds thrive on materials such as cotton, wool, paper, leather, and wood. As a result mold, which will grow anywhere indoors where there is moisture, is quite a common occurrence in buildings. The moisture may be in the host material, on its surface, or in the form of humidity in the air. Although there have been known health problems associated with mold, the one problem that brought it to the attention of the media and therefore the public is a house in Austin, Texas. Because of a series of water leaks in the house, a toxic buildup of mold developed.

It is not unusual for a water leak in a house to cause the mold buildup. However, in this case the mold was the toxic species *Stachybotrys chartarum* (atra), which is black in color and produces airborne toxins. The spores of other molds can appear as black, brown, blue, orange, or white specks although they are not always visible to the naked eye. Because of the airborne toxins, the family living in the house suffered a number of very serious health problems, and the house was deemed uninhabitable.

In order not to overreact to finding a mold buildup in the house, it's important to separate the hype from reality. Fortunately, extensive *Stachybotrys* mold is relatively rare, and the molds usually found in homes are the less toxic species. Exposure to mold spores in a house does not always present a health problem. The testing for mold during a home inspection is not considered necessary even though there are a number of companies selling kits that test for mold. According to the U.S. Center for Disease Control, "It is not necessary to determine what type of mold you may have. All molds should be treated the same with respect to potential health risks and removal."

If there is a mold condition in the house that you are inspecting and you are allergic or otherwise sensitive to mold, you will usually know it because you may experience symptoms such as nasal stuffiness, eye irritation, or wheezing. During your inspection of the interior, check all the rooms, especially a basement, for indications of a possible mold buildup. In this effort your eyes and nose are the best tools. Large mold infestations can usually be seen or smelled. If you see a moldy buildup on a wall or ceiling, record the location on your worksheets for either cleanup or replacement. In most cases you can easily remove the mold by a thorough cleaning with bleach and water. *(Never combine bleach and ammonia because the mixture produces a toxic gas.)* However, if there is an extensive amount of mold, you may want to engage the services of a professional with experience in cleaning mold in homes. Eliminating the mold problem may include discarding moldy items and replacing sections of walls and/or ceilings.

Mold odors can be masked intentionally or unintentionally by a seller. If the house that you are inspecting has an odor because of incense burning, cinnamon water boiling on the stove, or some other pleasant odor, ask the seller if you can return for an additional inspection when there are no extraneous odors.

If you detect a moldy odor in a room, look for water stains, even those stains that are dry to the touch. It is possible that there is mold buildup on the backside of the drywall or paneling. In this case further investigation is required. However, this type of investigation is beyond the scope of a home inspection and should be performed by a company or consultant that specializes in "sick houses." Also, in those carpeted rooms that are vulnerable to occasional or more frequent flooding such as a bathroom or finished basement, if there is a moldy odor, the carpet should be suspect as the source of the mold buildup. If there is mold, the typical solution is to replace the carpet. The hidden dust in fibers and the underside of a wetted carpet are ideal locations for mold growth. In addition to the carpet, it is also possible that there is mold buildup behind the walls.

Since basements are below ground and are therefore vulnerable to water penetration, pay particular attention for any signs of past water problems because mold growth is almost always due to excessive moisture. A moldy odor in the basement during the warmer months is not uncommon and is the result of condensation because of high relative humidity.

The condition can be easily corrected with a dehumidifier. If the dehumidifier doesn't correct the problem, further investigation will be necessary.

21

Green home technology

Roof 307
Roof-mounted structures 307
Paved areas around the structure 307
Exterior walls—siding (cladding) 308
Deck 308
Windows 308
Structural framing 309
Heating, ventilation, and air-conditioning (HVAC) 309
Water heating 310

In the residential real estate market, a "green" home is the latest buzz word. But what does green really mean? There is no single fixed definition. It is interpreted in many different ways. A green home encompasses a myriad of factors such as

- Its impact on the environment
- The site the house is located on
- The actual building design
- The material used for construction
- The use of nontoxic and renewable material for decorating
- The efficient use of energy, which includes heating and air conditioning equipment, appliances, and lighting
- The efficient use of water and its associated fixtures
- The quality of the indoor air
- Whether the project has a global impact

In order for a house to be certified as a green home, it must satisfy certain criteria. There are several shades of "green" depending on the extent to which the criteria have been satisfied. There are a number of organizations that provide guidelines for a green building certification program. However, the two main organizations that certify homes are the U.S. Green Building Council, which issues the Leadership in Energy and Environmental Design (LEED) for Homes certification, and the National Association of Home Builders (NAHB), which issues the National Green Building certification. In 2008, the American National Standards Institute (ANSI) approved the National Green Building Standard, which is used as part of the NAHB National Green Building program.

Each of the above items is subdivided into great detail, and each subitem is assigned a number of points. Based on the total number of points received, houses are certified at one of four levels of green. The green certification levels and the number of points required to achieve that level vary depending on the group that is doing the certifying. The LEED certification levels are Certified (45–59 points), Silver (60–74 points), Gold (75–89 points), and Platinum (90–136 points). The NAHB certification levels are Bronze (237 points), Silver (311 points), and Gold (395 points).

The following checklist is from the U.S. Green Building Council's *The Green Home Guide*. It lists those items in general terms that a person should look for and be aware of when looking for a green home.

- *Location*. New green homes and neighborhoods must not be built on environmentally sensitive sites like prime farmland, wetlands, and endangered species habitats. The greenest development sites are "in-fill" properties like former parking lots, rail yards, shopping malls, and factories. Look for compact development where the average housing density is at least six units per acre. The home should also be within easy walking distance of public transportation—like bus lines, light rail, and subway systems—so that you can leave your car at home. A green home should also be within walking distance of parks, schools, and stores.
- *Size.* No matter how many green building elements go into the home, a 5,000-square-foot green home still consumes many more natural resources than a 2,000-square-foot green home. The larger home will also require more heating, air conditioning, and lighting. If you really want a sustainable home, choose a smaller size.
- *Building design.* The home should be oriented on its site to bring abundant natural daylight into the interior to reduce lighting requirements and to take advantage of any prevailing breezes. Windows, clerestories, skylights, light monitors, light shelves, and other strategies should be used to bring daylight to the interior of the house. The exterior should have shading devices (sunshades, canopies, green screens, and—best of all—trees), particularly on the southern and western facades and over windows and doors, to block the hot summer sun. Dual-glaze windows reduce heat gain in summer and heat loss during cold winter months. The roof should be a light-colored, heat-reflecting Energy Star® roof, or a green (landscaped) roof, to reduce heat absorption.
- *Green building materials.* A green home will have been constructed or renovated with healthy, nontoxic building materials and furnishings, like low and zero volatile organic compound (VOC) paints and sealants and nontoxic materials like strawboard for the subflooring. Wood-based features should come from rapidly renewable sources like bamboo, but if tropical hardwoods are used, they must be certified by the Forest Stewardship Council. A green home uses salvaged materials like kitchen tiles and materials with significant recycled content.
- *Insulation.* A nontoxic insulation, derived from materials like soybean or cotton, with a high R (heat resistance) factor in a home's walls and roof will help prevent cool air leakage in the summer and warm air leakage in the winter.
- *Windows and doors.* Windows and exterior doors should have Energy Star ratings and they should seal their openings tightly to avoid heat gain in summer and heat loss in winter.
- *Energy efficiency.* A green home has energy-efficient lighting, heating,

cooling, and water-heating systems. Appliances should have Energy Star ratings.

- *Renewable energy.* The home should generate some of its own energy with technologies like photovoltaic (PV) systems.
- *Water efficiency.* A green home has a water-conserving irrigation system and water-efficient kitchen and bathroom fixtures. Look for a rainwater collection and storage system, particularly in drier regions where water is increasingly scarce and expensive.
- *Indoor environmental quality.* Natural daylight should reach at least 75 percent of the home's interior. Natural ventilation (via building orientation, operable windows, fans, wind chimneys, and other strategies) should bring plentiful fresh air inside the house. The heating, ventilation, and air conditioning (HVAC) system should filter all incoming air and vent stale air outside. The garage should not have any air handling equipment or return ducts, and it should have an exhaust fan.
- *Landscaping*. Vine-covered green screens, large canopy trees, and other landscaping should shade exterior walls, the driveway, patios, and other paved areas to minimize heat islands. Yards should be landscaped with drought-tolerant plants rather than water-guzzling plants and grass in most regions.

Along with green certification, homes are often rated by Energy Star, the government-backed symbol for superior energy efficiency. Products that can earn the Energy Star include windows, heating and cooling equipment, lighting, and appliances. Energy Star is a joint program of the U.S. Environmental Protection Agency (EPA) and the U.S. Department of Energy. To earn the Energy Star, a home must meet guidelines for energy efficiency set by the U.S. EPA. These homes are at least 15 percent more energy efficient than homes built to the 2004 International Residential Code (IRC), and include additional energy-saving features that typically make them 20 to 30 percent more efficient than standard homes.

Typical features to look for in Energy Star–qualified homes include

- An efficient home envelope, with effective levels of wall, floor, and attic insulation properly installed, comprehensive air barrier details, and high-performance windows
- Efficient air distribution, where ducts are effectively insulated
- Efficient equipment for heating, cooling, and water heating
- Efficient lighting, including fixtures that earn the Energy Star
- Efficient appliances, including Energy Star–qualified dishwashers, refrigerators, and clothes washers

In addition to the above, a green home is one that is environmentally sustainable. That is, the material used for the construction and the utilization of energy during the building's lifecycle will not compromise the health of the occupants or the ability of future generations to meet their own needs. Although a green home may be the ideal, in all probability you are more likely to find a home that is partially green than one that is certified green.

Every year there are millions of homes where one or more of the home's major components is being replaced or upgraded, or a new room addition to the house is being planned or constructed. To a large extent the degree to which these items will conform to green principles will depend on the homeowner's budget and finances.

During the construction phase of an addition, and in upgrading various home components, there are usually a number of different techniques, devices, materials, and equipment that can be used to comply with green principles. Discussing all of them is beyond the scope of this chapter. However, presented here is a sampling of some of the items that are used to green-up an existing house that you are likely to encounter when house hunting.

Roof

Asphalt shingles are the most common type of roof covering. Typically their lifespan has been between 15 to 25 years. When replacing a roof there are a number of "green" options. There are synthetic slate look-alikes that are now on the market. These products have the durability, texture, and appearance of natural slate and are lighter in weight. The materials used for synthetic slates are quite varied. One product is made from recycled steel reinforced rubber automobile tire treads. Others are made from recycled postindustrial rubber and plastic waste. There is also a synthetic slate made from ground natural slate, resin, and fiberglass, bonded under high pressure. Although synthetic slates are too new to comment on their life span, manufacturers estimate it at 40 to 60 years or more.

Metal shingles are becoming more popular for residential roofing. They are available in several different shapes and are primarily made from painted or coated aluminum or steel panels although they are also made from copper. The panels are formed to resemble wood shakes, slate or Spanish, Roman, and Mediterranean tile. The shingles are durable, lightweight, and fire-resistant. Some have embedded stone chips for additional texture. In addition to shingles, flat sheet metal roofing has also been gaining in popularity for residential pitched roofs.

Also, shake-like shingles have been made from 100 percent recycled vinyl and cellulose fibers, and reclaimed wood can be reformed into shakes, or wooden shingles. Photovoltaic (PV) roof shingles are also used. They are most effective when the roof slope faces a southern or southwestern exposure. These shingles convert sunlight into electricity.

Roof-mounted structures

Some homes have thermostatically controlled roof-mounted attic fans, which are wired to the household electrical system (see discussion on attic fans in chapter 9). These units are being replaced with PV solar-powered attic fans, which can be mounted on a roof pitch of 3/12 to 12/12. The PV panel on the unit can be adjusted from laying flat to a 45 degree angle. This unit qualifies for federal tax credits in the Solar Energy Systems category under Photovoltaic Systems as defined in the Energy Improvement and Extension Act of 2008.

Gutters and downspouts can be an important part in water conservation. A simple rainwater collection system can be made by connecting the base of the downspout to a barrel. The water in the barrel can then be used to water plants and shrubs, thereby saving the equivalent amount of potable water.

Paved areas around the structure

The driveways, walkways, and the patios in many homes have a surface area that is impervious to water. As a result, rain washing over those areas picks up contaminants, pollutants, and oils that dripped from the underside of the family auto and carry them to the storm drain where they flow to the local water supply. Some homes have minimized this problem with the use of permeable paving. With permeable paving, water drains naturally through the voids into the ground, where it helps to recharge the groundwater. For the most part, porous paving materials appear nearly indistinguishable from nonporous materials. Permeable paving can be made from asphalt, paving stones, bricks, or concrete. Even though the pavement may appear to be the same as traditional pavement, it is manufactured without the "fine" materials used in traditional pavement, and instead incorporates void spaces that allow for drainage.

Exterior walls—siding (cladding)

For hundreds of years wood, in the form of cedar shingles, shakes, and clapboard, had been the popular choice for siding a house. It was readily available, reasonably priced, and attractive. However, old growth forests have been depleted or are being depleted at a rapid rate, and as a result wood siding is no longer as readily available. Wood siding is still being used on houses, however; for a green house, the wood should be certified by the Forest Stewardship Council (FSC). This indicates that the siding came from a forest that was managed and harvested with sustainable methods. In fact, in addition to siding, all the lumber used for framing a green house should be FSC-certified.

Fiber-cement siding is a popular cladding for green homes. It is made from a mixture of cement, sand, cellulose fibers, and other components. The siding is popular because unlike wood siding it is fire-resistant and will not burn. It will not rot, warp, or delaminate, and is resistant to termites. It is also warranted to last 50 years. Fiber-cement siding can be textured to look like wood lap siding, shingles, and vertical siding.

Vinyl siding, which is commonly used throughout the country because of its low cost and low maintenance, does not have the blessings of many in the green building industry. Because of environmental concerns in the processing, manufacturing, and disposal of vinyl siding, there is controversy concerning its use. Dioxin, an environmental toxin, is a by-product of the manufacturing process. Also, in the event of a fire, when vinyl siding burns, it can release toxic fumes into the atmosphere.

Deck

In the past most residential decks were constructed entirely of wood, and to a large extent many still are. Wood is a renewable resource and can easily be recycled. In addition to using domestic wood for decking, exotic South American hardwoods such as Ipe, Cambara, and Meranti, normally used for furniture, are now also being used for decking. The hardwoods are beautiful, durable, and rot- and insect-resistant. Bamboo is also being used for decking. Of course, all wood decking should be FSC-certified.

In the last 20 years, composite decking has been displacing wood as the decking material of choice. Its popularity stems from the fact that it does not require waterproofing and resists damage from weather and insects, and is splinter-free. They are available in different colors and textures, which give the appearance of wood. The composites are generally made of recycled plastics such as shopping bags and wastewood fibers from reclaimed wood and sawdust. They are not intended for structural use. The structural support for those decks that use composite lumber, the joists, beams, and posts, will still be made of wood.

Windows

A big percentage of the heat loss in a house occurs at the windows. Most of the homes built before 1985 were erected with single glazed window panes. No doubt many windows have already been replaced, with many more due for replacement. A minimum replacement window for a green home is one that is wood framed and double glazed. A better window is one where the void between the panes is filled with an inert gas such as argon or krypton, which are better insulators than air. The best window is one where the glazing has a low-emissivity (low-e) coating. Depending on the climate that the house is located in, the low-e coating can be used to reflect the heat back into the house during the winter and reflect the sun's rays away during the summer.

Vinyl frame windows have also been used in green homes. The downside, however, is that vinyl manufacturing and incineration create dioxin, a highly carcinogenic chemical. Fiberglass frames are also energy-efficient but are more expensive than vinyl. Regardless of which windows are used they should have the Energy Star label indicating that they are high-performance windows.

Structural framing

Most existing homes were built with traditional 2 × 4″, 16″ on-center framing. With green homes the recommended framing is 2 × 6″, 24″ on-center. This will not only save wood, but the additional depth of the wall allows for increased insulation, which in turn reduces heat loss and saves on fuel. Instead of stick-built exterior walls, some green homes have walls built with structural insulated panels (SIPs). These panels consist of a foam core of expanded polystyrene (EPS) sandwiched between two layers of oriented strand board (OSB). Typically, panels are 4 to 24 feet wide, 8 or 9 feet high, and made in thicknesses of 4½ or 6½ inches. SIPs are 35 percent more energy-efficient than stick frame walls and are approximately 50 percent stronger than standard construction.

Framing of the interior portion of the house can be done with conventional lumber; however, engineered lumber has replaced solid-sawn lumber as the framing and structural support members of choice by builders in many newer homes and in homes that are currently being built. The decision is based on the fact that engineered lumber can be ordered to specific dimensions so that there is less wasted wood. Also, the lumber can be made from small-diameter, fast-growing trees that are grown in tree farms. The advantage of engineered lumber is that it is stronger, stiffer, and more dimensionally stable than solid-sawn dimensional lumber. Although the lumber is initially more expensive, the total installed costs are less than that of conventional lumber. The following products fall in the category of engineered lumber: Glued laminated beams (glulam), laminated veneer lumber (LVL), I-joists, and open web truss-joists. Glulam and wood I-joists can carry greater loads over longer spans than is possible with solid-sawn lumber of the same size. Because of their "I" cross-sectional shape, wood I-joists weigh up to 60 percent less than lumber joists, making them easier to handle. They generally do not shrink, warp, cup, crown, or twist. As with all green homes, the wood used for framing should have the Forest Stewardship Council certification.

Heating, ventilation, and air-conditioning (HVAC)

The heating and cooling equipment in a house provides a comfortable environment, but it is also a major consumer of energy. In homes that have been "greened up," the heating unit (furnace or boiler) is often replaced with a unit that has the Energy Star label. That is, the conventional heating unit, which had an operating efficiency of about 60 to 65 percent, has been replaced with a condensing boiler or furnace with an operating efficiency of about 90 percent. The increase in efficiency results from a redesign of the heat exchanger, which extracts the heat from the exhaust gases that normally flow up the chimney. During this process, the temperature of the exhaust gases drops to a point where the water vapor in the exhaust gases condenses, thereby releasing additional heat. With the conventional furnace, the temperature of the flue gases is about 450–550° F. With the high-efficiency units, it's about 120–130° F.

An HVAC system that you are more likely to find in new construction than in existing homes is a geothermal heating and cooling system. This system is basically a modified heat pump system. It is often referred to as a *ground source heat pump*. Unlike an air source heat pump where the compressor/condenser is located on the outside of the house, the geothermal heat pump is generally located in the basement or utility room. It is packaged in a single cabinet that includes the compressor, refrigerant heat exchanger, and various controls. The system uses the earth as a heat source in the winter, and heat sink in the summer. It takes advantage of the earth's relatively constant temperature, at depths below the frost line, of 40° to 55° F year round. The system circulates a water/antifreeze solution through a loop of high-density polyethylene pipe that is buried in the ground. There are two basic types of loops used with this system: closed and open. The simplest, but less common, open loop system circulates a constant source of ground water through the heat pump's heat exchanger, and discharges it back to the same aquifer. However, the supply and return lines must be placed far enough apart to ensure thermal recharge of the source. The open loop can also work with a stream, well, or pond, although in some areas this creates environmental concerns.

In a closed loop system, depending on the ground conditions and available space, the polyethylene pipe can be laid out vertically in holes 75 to 500 feet deep or horizontally in trenches dug below the frost line. Where there is a pond or stream that is deep enough and with enough flow, closed loop coils can be placed on the bottom of the pond or stream. Depending on which cycle the system is operating under (heating or cooling), the water/antifreeze solution circulating through the heat exchanger in the heat pump cabinet adds or extracts heat. It has been estimated that a homeowner may save anywhere from 20 to 60 percent annually on utilities by switching from an ordinary system to a ground source system.

Water heating

Many geothermal systems installed today are equipped with desuperheaters, which provide domestic hot water when the system is providing heat or air conditioning. A desuperheater is a small auxiliary heat-recovery system at the compressor outlet that uses superheated gases from the heat pump's compressor to heat water, which then circulates through a pipe to the home's water heater tank. In the summer, when the air conditioning runs frequently, a desuperheater can provide all the hot water needed by a household virtually for free. However, it provides less hot water during the winter months, and none during the spring and fall months if the system is not operating. The home's conventional water heater is then used to meet household hot water needs in winter if the desuperheater isn't producing enough hot water and in spring and fall when the geothermal heating and cooling system may not be operating at all.

As discussed above, there is no single fixed definition of a green home. It is interpreted in many different ways. It encompasses environmentally sustainable construction, which is the responsible use of natural resources, and how it affects the quality of life for both present and future generations. It's more than energy conservation; it includes the impact on the environment, the site the house is located on, the material used for construction, the use of nontoxic and renewable material for decorating, the efficient use of water and its associated fixtures, the quality of the indoor air, and whether the project has a global impact. At the present time, from a home inspection point of view, you are more likely to find an existing home that has been "greened up" than a home that has been certified as a green home.

22

Conclusion

Professional home inspection 312
Home inspection limitations 312
Real estate warranty program 313
Private home inspection warranty programs 314
Manufacturers' warranties 314
Contract 314

For most families, the purchase of a home is an emotionally charged event filled with anxiety, hopeful dreams, and sometimes disappointments after moving in. The disappointment, however, can be avoided or at least minimized if you know the true condition of the house that you are buying. Prior to purchasing the house, you or a professional should perform a home inspection.

Professional home inspection

The purpose of this book is to provide sufficient background material and an overall procedure, so that you can perform a technical inspection of the house you are considering buying or the one you are currently living in. As previously pointed out, at times the services of a professional home inspector are necessary. A professional inspector, however, will not inspect in as detailed a fashion as suggested in this book. He or she will normally not operate every window; check every light switch; check the temperature of every radiator; indicate every cracked wall, floor, or ceiling (unless the crack represents a structural problem); check the caulking at every joint; and so on.

The main concerns of an inspector are the major problems or deficiencies, ones that will be costly to correct and ones that pose a health, safety, or fire hazard. Because of the inspector's training and experience, he or she is better able than a layperson to pick up problems through a more cursory inspection and relate problem conditions in one area to difficulties in another.

When selecting a professional home inspector, choose one whose sole endeavor is inspection. It is advisable not to use an inspector who is affiliated with an exterminating company, a real estate broker, or a contractor (plumber, electrician, roofer, and so on). The inspector must be completely unbiased and must not have a vested interest in finding problem conditions. Licensed professional engineers who specialize in home inspections have an advantage (because of their training) in that they can evaluate certain problems in depth.

Home inspection limitations

A home inspection is not perfect and will not necessarily reveal every problem, but it is the best thing available. A home inspection is basically a visual inspection as well as an operational check of the various visible elements and components of the house. If performed in the manner suggested herein, you should have a fairly good picture of the condition of the house at the time of the inspection. By taking into account the age and present condition of various items, predictions can usually be made about future problems.

However, a home inspection cannot reveal a deficiency or problem if the conditions causing it are not visible. The following are just a few examples of such possible situations:

- A house with a concrete block foundation may have termite infestation in the voids of the foundation walls. This condition would not be visible during an inspection. Nevertheless, a week later there may be visible termite shelter tubes or a swarm.
- A basement or crawl space can be "bone dry" for the first three years and then flood in the fourth year because of a new building development nearby.
- The operation of a septic system depends on the condition (porosity and degree of saturation) of the subsoil in the leaching field, an item that is not readily apparent during an inspection. Depending on the weather, the level of the water table, and the usage prior to the inspection, it is possible for a malfunctioning septic system to appear to operate properly during the time it takes to perform an inspection.
- If there is an inadequately insulated water pipe in an exterior wall, it can freeze and burst during a severe cold spell. Since the pipe would not be visible during an inspection, the condition would not be detected.

Even with such limitations, a proper home inspection will reveal considerable information about the house that very often even the owner does not know.

Real estate warranty program

Many homes throughout the United States are being sold with a warranty protection plan for the buyer. The warranty usually covers the following components:

- Electrical system
- Heating system
- Water heater
- Plumbing system
- Central air-conditioning system
- Built-in appliances

It is in effect for the first year that the buyer owns the house. The warranties are usually made available to the buyer through the real estate broker. Not all brokers, however, are affiliated with companies that offer home warranties.

The home warranties that are currently available when you buy a house are very limited in scope and should not be considered a replacement for a home inspection. The warranty is basically an insurance plan that will protect the new homeowner against defects or malfunctions in the covered components, and nothing more. When you are planning on buying a house, many questions about the house should be answered so that you can make an intelligent decision. These questions concern items that are not covered by the warranty, and their answers can have a significant impact on the true cost of buying the house. Remember, the true cost of a house is the purchase price plus the costs for upgrading substandard, deteriorated, or malfunctioning components.

For example, the electrical system is covered against defects under the warranty. The fact that the electrical service in the house might be totally inadequate (110 volts at

30 amps) and would have to be upgraded at a considerable expense is not covered by the warranty. Also not covered by the warranty but costly to correct are basements that are subjected to periodic water penetration, structural damage to the foundation and woodframing, termite infestation and deterioration, roof replacement, and improvements needed for energy conservation. In addition, items of concern are general and latent defects such as an inadequately ventilated attic. A home warranty is definitely of value as long as you realize that it is not the answer to all of your questions or potential problems.

Private home inspection warranty programs

For an additional fee, some private home inspection companies offer a warranty protection plan after they inspect the house. The warranty applies for a twelve-month period and is usually effective from the day of the inspection or the transfer of title. The warranty plan is basically an insurance program. In fact, in some states, inspection companies have been forced to drop their warranty programs because they were not licensed by the state insurance commission to sell insurance.

As with all types of insurance, you should know exactly what you are buying before you put your money down. The warranty plan protects the buyer against defects in most of the major housing components, providing that the components are in reasonably good condition to begin with. If during the inspection any covered component is judged to be in less than satisfactory condition, it is excluded from coverage, although this is not reflected in the warranty fee. Also, for an item such as the roof, there is usually a prorated repair schedule that is a function of age. For example, in one plan if the roof is nine years old (which is not considered old or even aging), only 30 percent of the repair or replacement costs are covered. If the deductible (which might be $250 or $500) is subtracted, there might not be any coverage.

The warranty agreement should be read very carefully. In addition to items excluded because of condition or age, many components that are costly to correct are often not included in the plan. As with real estate warranty plans, these warranties can also be of value as long as you know exactly what you are buying and understand the limitations.

Manufacturers' warranties

Many of the components and items in a house have the manufacturer's guarantee or warranty against defects for a specific number of years. Items that are usually covered are the furnace, water heater, roof shingles, well pump, air-conditioning compressor, and electrical appliances. If you are purchasing a relatively new house or are in an older house any of the above components are relatively new (because of replacement), you should obtain the warranties or bills of sale. Many warranties are transferable to a new owner. Also, if the house has been termite-proofed, do not forget to get a copy of the guarantee if it is still in effect.

Contract

Many lawyers indicate that the best time to inspect a house you are considering buying is *before* signing the contract. This avoids the problems that can develop if the contract has already been signed and the subsequent inspection reveals major deficiencies in the house. Often the buyer wants a price adjustment or wants to back out of the deal. By inspecting the house before signing the contract, the correction of problems and deficiencies can be written into the contract.

In a seller's market, however, the buyer might feel pressured into signing the contract prior to the inspection in fear of having the house sold to another buyer before the inspection has been performed.

If you do sign a contract before the house has been inspected, for your protection, the contract should include an "inspection contingency" clause.

It is important that the clause be broad enough to include all the major problems that can occur in a house. Otherwise, there might be legal problems in backing out of the deal because of a major defect that was not covered in the contract. For example, many years ago I received a telephone call from a buyer for whom I had done a home inspection. The buyer indicated that his contract had an inspection clause that provided him with the option of declaring the contract null and void if there were *structural* deficiencies. That was basically the extent of the clause; it was very limited. My inspection had revealed that the electrical service was inadequate, that the heating system had to be replaced, and that there was evidence of water penetration into the basement. After the inspection, the buyer decided that he no longer wanted to buy the house. He called to find out if the problem items were considered structural deficiencies. No, they were not, and unfortunately, because of the wording in the contract, the buyer lost part of his deposit.

One contingency clause that has been used states that the sales agreement is contingent upon having a home inspection within *n* days of the signing. The inspection is defined to include a physical check of the exterior and interior components of the house (including the structural integrity of the building) and the condition of the electrical and mechanical equipment. If the inspection reveals any defects for which the cost exceeds $*n*, the buyer should have the following options: have defects corrected by the seller; negotiate the cost of the corrections with the seller; or declare the sales agreement null and void. The limitation of this clause is in defining the means by which the costs for correcting the defects are determined.

To avoid this type of problem, one attorney I know uses the following clause:

> *This contract is contingent upon the Purchaser's making, or having made on his behalf by another, a physical inspection of the entire premises. If such inspection reveals any condition or state of facts unsatisfactory to the Purchaser, in his sole discretion, the Purchaser shall receive back all monies paid hereunder. Upon receipt of said monies, this agreement shall be null and void, and neither party under any obligation whatsoever to the other.*

This clause might not be valid in your state. To avoid any legal entanglement, your best bet is to consult an attorney.

As mentioned in chapter 1, just before the contract closing you should perform a final "walk-through" inspection of the house. Although most sellers will not deny you the right to make this inspection, it is not ensured. Consequently, the right to a final walk-through inspection should be made part of the contract.

No two houses are alike. Just as a person changes with age, so does a house. Regardless of whether you are a home buyer or a homeowner, it is important to know the true condition of the house. With this book, you are now prepared to find out. Try doing the inspection yourself. It's really a lot of fun. Worksheets are provided on which you can record your observations. Good luck!

Appendix

Home inspector requirements

In a little over 30 years, home inspection as a service to home buyers has morphed from a part-time activity, by a number of individuals working in and around large metropolitan areas, into a mature profession that has been accepted and recommended by government agencies. When I started my home inspection company in 1971 there were no specific requirements covering the items in a house to be inspected, and there were no specific qualifications to be an inspector. At that time anyone could hang out a shingle claiming to be a home inspector. Things started to change when we formed the American Society of Home Inspectors (ASHI). In January 1976 a small group of independent inspectors met and discussed the need to form a nonprofit professional organization. We recognized that in order for inspectors to be accepted by the public as professionals, it would be necessary to establish an organization that has a set of standards regarding those items to be inspected, a code of ethics, and membership requirements.

Today, ASHI is the largest national professional organization of home inspectors with 88 local ASHI chapters servicing thousands of members throughout the United States and Canada. ASHI's Standards of Practice, Code of Ethics, and membership requirements are included at the end of this appendix. As the profession of home inspection matured, additional national organizations were formed to promote excellence and professionalism by also providing a set of standards and a code of ethics and providing educational opportunities such as technical seminars, continuing education courses, and monthly meetings at which there are relevant speakers and where members can exchange ideas and network.

National home inspection organizations and their date of establishment are

- American Society of Home Inspectors (ASHI)—1976
- National Association of Home Inspectors (NAHI)—1987
- National Association of Building Inspection Engineers (NABIE)—1989
- Housing Inspection Foundation (HIF)—1989
- National Association of Certified Home Inspectors (NACHI)—1990
- Canadian Association of Home & Property Inspectors (CAHPI)—1994
- National Society of Home Inspectors (NSHI)—2003

In addition to the national home inspection organizations above, there is also a number of independent state home inspection associations as listed below:

- California Real Estate Inspection Association (CREIA)—1976
- Texas Association of Real Estate Inspectors (TAREI)—1977
- Minnesota Society of Home Inspectors (MSHI)—1979

- Georgia Association of Home Inspectors (GAHI)—1989
- Connecticut Association of Home Inspectors (CAHI)—1992
- Kentucky Real Estate Inspection Association, Inc. (KREIA)—1992
- Arkansas Association of Real Estate Inspectors (AAREI)—1993
- Wisconsin Association of Home Inspectors, Incorporated (WAHI)—1995
- North Carolina Licensed Home Inspector Association (NCLHIA)—1997
- Southern Nevada Association of Professional Property Inspectors (SNAPPI)—1998

As of this writing, 32 states have enacted legislation that requires individuals performing home inspections to become licensed. This further ensures professionalism and provides consumer protection by preventing anyone who is not licensed by the state from holding oneself out as a home inspector, or from performing a home inspection. Although licensing requirements vary from state to state, most require the candidate to take the National Home Inspection Examination or equivalent, as a means of assessing their competence. Some states also require the candidate to successfully complete an exam of ASHI standards and ethics. The national examination is administered by Pearson VUE at more than 215 proctored test center locations throughout the United States, Canada, and Puerto Rico. In addition to the exam, depending on the state, there are requirements that the candidate must successfully complete 60 to 180 hours of classroom instruction, approved by the state, with some states also requiring 40 hours of field-based instruction; moreover, the candidate must have both liability insurance and errors and omission insurance, and must have completed 100 to 200 fee-paid inspections under the supervision of a licensed home inspector.

Included in this appendix is a 128-question quiz along with the answers. The quiz is based on the inspection information that is presented in this book. It is not intended as a study guide for the National Home Inspection Examination. However, it is a good check on your understanding of some of the basics of home inspection. If you are interested in taking the national exam, the national organizations or their local chapters can provide you with information on schools that provide training for the exam.

Once the candidate is approved for licensing, he or she must abide by the rules and regulations set out by the legislation. An example of an item that is included in the legislation of one state is the preinspection agreement. There is a requirement that the home inspector must send a preinspection agreement to the client no later than one business day after the appointment for the home inspection is made. The agreement must contain, among other things, the fee and a description of the systems and components that will be inspected, as well as those items that are not part of the home inspection. The preinspection agreement is also required to contain a description of any and all services not part of the home inspection for which the client is charged an additional fee. The agreement must be executed prior to the start of the home inspection.

In some states, these two areas of concern are not part of a home inspection and require the inspector to have a separate license in order to evaluate them: the presence of wood-destroying organisms and the condition of the septic system. A seperate license is also needed to perform a test or collect samples to check for environmental hazards, such as a high radon concentration in the house and the presence of mold, asbestos, and lead in the water or paint. Those tests or samples must then be analyzed by a certified laboratory.

Depending on the state, the licensing legislation may also contain a standards of practice section, which outlines in considerable detail those systems and components in a house that must be inspected, as well as those items that must be included in a written report to the client. The licensing law can be quite specific. One state, New Jersey, has included in its legislation a mandatory tools and equipment requirement. It states that all home inspectors during the performance of an inspection must be equipped with the following minimum tools and equipment; the inspector, however, is not limited to that list and can use additional tools and equipment.

1. A ladder, minimum 11 feet in length.
2. A flashlight, or another equivalent light source, with a minimum 15,000 candlepower illumination.
3. A flame inspection mirror.
4. An electrical outlet tester with ground fault circuit interrupter (GFCI) test ability.
5. Tools necessary to remove common fasteners on covers or panels that are required to be removed for inspection.
6. A measuring tape.
7. A probe.
8. Thermometers for testing air-conditioning.
9. Binoculars with a magnification between 8 × 42 and 10 × 50.
10. A moisture meter.
11. Combustible gas leak detection equipment.
12. A voltage detector.

Depending on the state requirements, when renewing the home inspection license, the inspector must attest or provide information showing successful completion of 10 to 40 hours of continuing education units (CEUs) per renewal period.

In most of those states that do not license home inspectors, there is usually a statute concerning the field of home inspection. In the statute there is a requirement that the inspector must be a member of a national or a state home inspection organization. This has helped the profession to be self-regulating. When there is a new inspector in town the more-experienced inspectors will check to determine whether or not the new inspector is a member of a national or state inspection organization. They check because the new inspector is more competition for them. If he or she is not a member of a national or state home inspection organization, which have strict membership requirements, the new inspector is encouraged to join one of the organizations.

Quiz based on *The Complete Book of Home Inspection*

1. A home inspection should begin on the
   a) exterior
   b) roof
   c) crawlspace
   d) interior
   e) electrical system

2. Roof sheathing is another term for
   a) wood shingles
   b) the roof deck
   c) thatched roofs
   d) roof runoff
   e) rain gutters

3. The most common form of roof shingle used in this country is
   a) wood shingles
   b) asphalt shingles
   c) concrete tile
   d) synthetic shakes
   e) tar and gravel

4. On a 17-year-old asphalt shingle roof, you would expect to find the shingles on the _________ slope more weathered than the shingles on the other slopes.
   a) easterly
   b) northerly
   c) southerly
   d) westerly
   e) none of the above

5. A Cedar Breather is
   a) a beetle found in cedar shingles
   b) a nylon matrix underlayment for cedar shingles
   c) a finish coating on cedar shingles
   d) used for regenerating the odor in a cedar closet
   e) a device for removing moss from cedar shingles

6. The most durable roof covering is
   a) modified bitumen
   b) treated shakes
   c) tile
   d) 40-year dimensional asphalt shingles
   e) slate

7. Ponded water on a flat roof that was not designed to hold standing water
   a) is desirable because it can reduce the heat load in summer
   b) is especially valuable in freezing weather
   c) must be maintained by hosing the roof in summer
   d) has more disadvantages than advantages and should be avoided
   e) is acceptable only on metal roofs

8. Single-ply membranes
   a) are used as an underlayment below shingle roofs
   b) are used on steeply sloped tile roofs
   c) can only be found on commercial roofs
   d) are vulnerable to leakage at seams
   e) cannot be covered with gravel

9. If the owner says the chimney has been leaning for 15 years
   a) it has withstood the test of time and is presumed safe
   b) corrective action is still needed
   c) it is because of uniform deterioration of the mortar on all sides of the chimney
   d) the chimney is not a potential hazard
   e) the owner probably removed an antenna from the chimney

10. A chimney cricket is
   a) an insect that nests in the chimney
   b) also known as a saddle
   c) found on every chimney

d) a code requirement
e) able to prevent downdrafts

11. Vent stacks
    a) are parts of the plumbing system that project through the roof surface
    b) should not go all the way through the roof
    c) are self-sealing and not vulnerable to roof leakage
    d) need not be checked unless doing a smoke bomb test of the septic tank
    e) should end below the eaves

12. The absence of rain gutters
    a) is a design defect, especially in snow country
    b) indicates that the soils are not expansive
    c) should be noted on your worksheet
    d) is normal in houses with eaves having sufficient overhang
    e) shows that the owner removed them rather than repaint them

13. Rain gutters without seams
    a) are always copper
    b) can be aluminum
    c) are not subject to corrosion
    d) do not require painting
    e) are an impossibility

14. The acceptable difference in height between stair risers
    a) should be less than one inch
    b) can be any amount when the steps are not level
    c) should be zero to avoid potential tripping hazards
    d) can be any amount as long as there is a handrail
    e) can be ½ inch

15. Sheathing on exterior walls
    a) is inappropriate and belongs only on the roof
    b) is installed to provide bracing and minimize air infiltration
    c) cannot also be the exterior siding material
    d) must be made of solid material, not fiberboard or gypsum
    e) is always used as the nailing base for exterior siding

16. When inspecting siding, it is important to pay special attention to
    a) the areas facing the north
    b) the area facing the prevailing winds
    c) the second story
    d) the areas facing east
    e) the sections that face south or southwesterly

17. If you find vines growing up an exterior wall
    a) you should consider removing them
    b) you are not responsible for inspecting the siding in that area
    c) the vines provide extra insulating value
    d) there is no problem as long as this is on a south-facing wall
    e) the paint is being protected from sunlight and will be OK

18. Aluminum siding
    a) is sometimes required to be grounded
    b) never needs repainting
    c) is durable and resistant to mechanical damage from impact
    d) is never used in new construction
    e) cannot incorporate insulation due to condensation problems

19. Stucco siding
    a) is a concrete sheet that is poured first then applied to a wall
    b) is made of recycled marble
    c) can be applied to curved or irregular-shaped surfaces

d) adheres directly to sheathing material
e) is built up in five or six coats

20. EIFS is the acronym for
a) Exterior Interior Flame Safety
b) Easily Installed Flue Shutoff
c) Exterior Insulation Finish System
d) Exterior Interior Final Siding
e) Every Interrupt Fault System

21. Fiber-cement wall shingles are popular because they
a) will not rot
b) will not warp
c) will not delaminate
d) are resistant to termites
e) all of the above

22. In a brick veneer wall, the weakest point is
a) the space between the bricks and the foundation
b) the metal ties to the wall frame
c) the mortar between the bricks
d) the weep holes
e) the area over window or door openings

23. Exterior trim
a) includes all portions of the exterior finish other than the siding
b) serves a structural function
c) is installed in addition to the protection for joints and edges
d) refers only to wood and not to metal
e) cannot be exposed to the weather

24. The most common type of window in older homes is
a) awning window
b) horizontal sliding window
c) casement style
d) Craftsman style
e) double-hung

25. Jalousie refers to a style or type of
a) soffit
b) window
c) porch
d) cornice
e) tango

26. A swale is a
a) measuring stick to determine the water table
b) type of marine mammal
c) leach field for a septic tank system
d) junction box for an underground drainage system
e) depression in the ground to intercept and redirect surface runoff

27. A high water table
a) is especially good when there is a septic tank
b) has nothing to do with local rainfall
c) can be lowered by puncturing the ground
d) will likely lead to flooding in basements
e) is never within 20 feet of the ground surface

28. A dry retaining wall is one constructed
a) without weep holes to let water out
b) only in areas where the soils percolate moisture readily
c) without mortar
d) with the footing at least two feet below the frost line
e) in one piece

29. A retaining wall must be designed or approved by a licensed engineer if it is taller than
a) 3 feet
b) 4 feet
c) 5 feet
d) 6 feet
e) 7 feet

30. Where deck joists attach to a house, they should be secured by
    a) being toe-nailed into a header
    b) toe-nails and a ledger, or by metal brackets
    c) notches in the siding material
    d) diagonal braces
    e) posts set into the siding

31. Wood decks
    a) that show any signs of deterioration should be replaced completely
    b) should have proper drainage slope on the surface
    c) do not require guardrails except on second stories and balconies
    d) should have repairs or replacement of deteriorated planks as needed
    e) must always be constructed with pressure-treated pine

32. Composite decks
    a) are constructed entirely of wood
    b) are maintenance free
    c) do not expand and contract with temperature changes
    d) can be used for structural members
    e) use wood for the structural support members

33. Free-standing decks
    a) require two or more additional posts
    b) are as safe and secure as attached decks
    c) need diagonal bracing between the posts and the beams
    d) all of the above
    e) have posts that do not extend below the frost line

34. A door between the house and the garage
    a) should have a metal surface on the house side
    b) must not be on a higher elevation than the garage floor
    c) should be self-closing
    d) cannot have a tight seal
    e) may not be solid wood

35. A wall between the garage and the house
    a) may not have the attic entrance hatch, even if covered
    b) should have vents to the crawlspace or basement
    c) need not be inspected if covered with storage shelving
    d) may have exposed wood framing if the rest of the garage is also open
    e) must be covered with plaster, stucco, or other fire-resistant material

36. Garage doors with automatic openers
    a) should be checked for automatic reversal of the door
    b) need not be tested if the garage is detached from the main house
    c) should be tested only by disconnecting them from the automatic opener
    d) should only be tested by operating the remote control
    e) cannot be single-piece doors

37. Garage door restraining cables are used to
    a) prevent the door from closing too fast
    b) prevent the spring from whipping around in the event it breaks
    c) prevent the rollers from slipping out of the channels
    d) prevent fingers from getting crushed in door sections
    e) prevent the door from reversing automatically

38. The insect that causes the most damage to houses in the United States is the
    a) powder post beetle
    b) subterranean termite
    c) drywood termite

d) carpenter ant
e) dampwood termite

39. In general, subterranean termites
   a) are very fast working and immediately destructive
   b) are likely to colonize within the home near the heating plant
   c) can be permanently eradicated by pesticides and termite barriers
   d) are social insects with a rigid caste structure
   e) are rare in California and the Southeast

40. Termite baits are
   a) considerably less toxic that most liquid termiticides
   b) slow-acting poisons combined with a termite food material
   c) usually installed in plastic tubes or boxes
   d) used by pest control operators
   e) all of the above

41. Fiberglass batt insulation in an unfinished attic
   a) should be placed between rafters with the foil side showing
   b) should be placed between joists with the foil facing toward the attic
   c) is not necessary in warm climates such as California
   d) should be placed between joists with the foil facing toward the heated portion of the structure
   e) should be slit to allow water vapor to escape

42. Attic ventilation
   a) is not necessary in climates with no snowfall
   b) is very difficult to upgrade
   c) is important in both summer and winter
   d) should not be provided by openings in gable walls
   e) is typically provided by ducts from the kitchen range hood

43. An air admittance valve (AAV)
   a) is a pressure-actuated one-way mechanical vent
   b) is a code requirement
   c) allows sewer gases to vent to the attic
   d) negates the need for vent stacks
   e) is only effective on rainy days

44. Defective Chinese drywall
   a) is a gypsum wallboard made in China
   b) produces a rotten egg odor
   c) causes air-conditioning coils to corrode
   d) causes copper plumbing and electrical wiring to corrode
   e) all of the above

45. Squeaks in flooring
   a) indicate a significant structural defect
   b) are generally difficult to eliminate
   c) can be easily repaired, especially on the second floor
   d) indicate problems with water vapor
   e) can be corrected by carpeting

46. Roof truss uplift causes
   a) cracks in door frames
   b) cracks in joint between floor and walls
   c) cracks in joint between walls and ceiling
   d) cracks in window frames
   e) permanent sag in roof ridge

47. If a thermal pane window appears cloudy
   a) it indicates poor housekeeping on the part of the owner
   b) the hermetic seal should be repaired
   c) water has leaked into the window from the exterior

d) there is no insulating value to the window
e) the window requires replacement

48. Poor water flow is typically caused by
a) too many fixtures on the second floor
b) poor water pressure
c) constriction in the supply pipes
d) inadequate meter sizing
e) clogged aerators

49. S-type traps are
a) used only in second floor bathrooms
b) required in two-pipe steam systems
c) installed on a/c condensate lines
d) not permitted by most plumbing codes
e) installed on water heaters

50. Improper venting at sink traps
a) can be determined by filling the sink with water and listening as it drains
b) is not an important enough problem to be worth fixing
c) is the term for lack of siphoning in a drain trap
d) can be seen by tape on the drain trap
e) will cause water hammer in the drain pipes

51. A rule of thumb for the minimum height of the chimney above the fireplace floor is
a) 8 feet
b) 10 feet
c) 12 feet
d) 15 feet
e) 18 feet

52. The best location for a smoke detector is
a) in the return air duct of the heating system
b) over the stove, since most fires start there
c) on the ceiling of the hallway leading to the bedroom
d) on a hallway wall but no closer than 6 inches from the ceiling
e) in the upper corners of each bedroom

53. Soils with a high clay content can swell or shrink up to
a) 10 percent
b) 20 percent
c) 30 percent
d) 40 percent
e) 50 percent

54. Deteriorated mortar joints in foundation walls
a) do not represent a weakened structural condition
b) should be repointed to improve the appearance of the foundation
c) should be repointed to restore load-bearing capacity
d) should be repointed to prevent basement water entry
e) can be painted to prevent further deterioration

55. Superior walls are
a) walls made from superior materials
b) less likely to crack under pressure
c) insulated, precast, ribbed concrete wall panels
d) made from a mix of concrete and fiber
e) concrete blocks with insulation in the voids

56. Engineered lumber
a) can only be used for roof framing
b) is termite-resistant
c) is stronger and more dimensionally stable than solid-sawn lumber
d) will prevent squeaks in flooring
e) is most effective when installed in cold weather

57. The inspection for water seepage into a basement or crawlspace begins
    a) after you see standing water in the crawlspace
    b) on the roof
    c) on the exterior
    d) by consulting maps available from the U.S. Geological Survey (USGS)
    e) by smelling damp conditions

58. Efflorescence is
    a) a significant structural defect
    b) an enthusiastic spontaneous response from realtors reading a favorable inspection report
    c) a chalky residue of sand and dissolved concrete created by the action of acidic soils
    d) an indication of effective damp proofing
    e) an accumulation of mineral salts created by moisture in masonry or concrete

59. The minimum FHA standards for crawlspace clearance are
    a) 24 inches between any wood and soil
    b) 18 inches soil-to-joist, 12 inches soil-to-girder
    c) 18 inches between any wood and soil
    d) 12 inches between the subfloor and soil
    e) any distance if pressure-treated lumber is used

60. In a conditioned crawl space
    a) the foundation walls are insulated
    b) wall vents are closed
    c) a dirt floor, if there is one, is covered with large plastic sheets
    d) the area is heated or air-conditioned
    e) all of the above

61. 110-volt electrical services
    a) are adequate for all modern appliances
    b) are adequate in most cases
    c) will suffice for all houses except those with air-conditioning
    d) are inadequate by current standards
    e) require that you use a gas refrigerator

62. Circuit breakers
    a) are an inconvenience when compared to fuses
    b) are more reliable than fuses and always provide greater safety
    c) cannot be used to provide overcurrent protection
    d) prevent overloading by allowing just enough voltage to pass through to meet demand
    e) are somewhat less reliable than fuses

63. The capacity of a branch circuit fuse
    a) can be increased by inserting a penny behind it
    b) should be as large as possible for added safety
    c) should match the current-carrying capacity of the branch circuit
    d) is electrically interchangeable with any other Edison fuse
    e) is increased by replacing the fuse with a fusestat

64. Removal of the interior cover of an electrical panel box
    a) is necessary to locate the main circuit breaker
    b) cannot be done without an electrical permit
    c) will enable determination of the true electrical capacity
    d) is not necessary when determining proper sizing of circuit breakers
    e) should be part of routine homeowner maintenance

65. If a house is wired with aluminum
    a) it is normal and acceptable for switches and outlets to be warm to the touch

b) the wiring is unsafe and should be completely replaced
c) there is no problem provided that outlets marked AL/CU were used
d) you should scan all the walls with an infrared heat detector to check for bad wiring
e) the connections should all be checked by a qualified electrical contractor

66. A ground fault circuit interrupter (GFCI) is an electronic device that
    a) can be used instead of a fuse
    b) is retroactively required for all bathroom receptacles
    c) will not work properly if the main grounding is faulty
    d) opens the circuit when it senses a potentially hazardous condition
    e) can be reached when touching another ground source such as a shower head

67. An arc fault circuit interrupter (AFCI) is an electronic device that
    a) is a backup if the GFI fails to operate
    b) does not require periodic testing
    c) disconnects the power if it detects arcing in electrical wiring
    d) is an important part of a fire alarm system
    e) is a retroactive building code requirement

68. Knob-and-tube wiring is
    a) obsolete but usually safe if unmodified
    b) acceptable when it includes grounding conductors
    c) dangerous because the wires are not insulated
    d) exposed and readily accessible in walls and ceilings
    e) found only in Canada, Europe, and the East Coast

69. Systems with light fixture switches operating at 24 volts
    a) indicate a significant loss of current and waste of electricity
    b) can be operated from only one location, and do not work for three-way switches
    c) can be repaired by removing the transformer
    d) should be provided with spare replacement relays
    e) are borrowed from automobile wiring systems

70. The house service main water pipe
    a) is maintained by the water company
    b) must always be at least 1 inch in diameter
    c) would be the homeowners' expense if in need of repair or replacement
    d) must not be made of metal or a dangerous electrical ground would result
    e) should be fully visible and accessible to inspection

71. Hot water piping
    a) should be no further than 4 inches from the parallel cold water piping
    b) cannot be a part of the heating system of the house
    c) is not visually accessible due to insulation
    d) is required for all plumbing fixtures
    e) should be separated from cold water piping by at least 6 inches

72. Drainage systems
    a) are simpler than supply systems
    b) consist of horizontal runs of piping
    c) include traps, drainpipes, and vents
    d) cannot operate properly by gravity alone
    e) must be the same diameter as supply pipes

73. The gases that are generated in a septic tank are usually
  a) explosive and must be kept out of the house drainage-vent system
  b) must be dissipated through a fresh air inlet at the house trap
  c) dissipated into the earth through the leach field
  d) dissipated to the atmosphere through the house drainage-vent system
  e) an indication that the septic tank is sterilized and not functioning

74. A mound septic system
  a) is most effective in areas with hard water
  b) is normally found on top of a mountain ridge
  c) is used when the drainfield has a problem such as inadequate percolation rates
  d) can have a mound 6 to 8 feet high
  e) uses gravity to move the effluent

75. Vent pipes in drainage systems
  a) allow air to flow in and out of drain systems
  b) are not necessary when connected to a public sewer
  c) carry water and wastes
  d) should not connect near fixtures
  e) are not allowed to discharge into the atmosphere

76. The presence of a septic tank
  a) cannot always be determined during an inspection
  b) can be determined by whether or not there is a house trap
  c) can be ruled out if there are sewers for other houses on the street
  d) cannot be determined from the building department or other public records
  e) is a significant defect that must always be reported

77. The appropriate size for a septic leaching field is determined by
  a) the number of seepage pits
  b) the size of the water supply piping
  c) the elevation above sea level
  d) whether or not a garbage disposer is installed
  e) the percolation rate of the ground

78. Septic tank leaching field
  a) must be supplied with fresh leaches each year
  b) works best in clay soils
  c) will receive nitrogen and other natural fertilizers from the septic effluent
  d) definitely failed if the vegetation over them is lush and green
  e) are a less likely area for failure to be visible as compared to the area over the tank

79. An atmospheric vacuum breaker
  a) cannot be used in residential plumbing without the approval of the Department of Air Quality
  b) is required for all lawn sprinkler systems
  c) may not be more than 6 inches above grade
  d) is required at each sprinkler head
  e) prevents dirty water from backing up from the drains

80. A wall hatch providing access to the fixture side of a tub
  a) is a violation of most plumbing codes
  b) indicates that leakage occurred requiring the wall to be opened for repairs
  c) must be distinctly noticeable and not blocked by furniture
  d) is common in many older homes
  e) should never be found above the first floor

81. Inlet service pipes made of lead
   a) can sometimes be identified by a large bulge around a wiped joint
   b) are still installed in modern homes
   c) pose no hazard after prolonged use as interior corrosion prevents direct water contact with lead
   d) are the responsibility of the water supply company not the homeowner
   e) are common, but testing of the effects on water is very expensive and usually unnecessary

82. PEX piping (tubing)
   a) should only be used for cold water
   b) should only be used for hot water
   c) is flexible, and won't corrode or develop pin holes
   d) uses a special adhesive to seal the joints
   e) reacts with water-soluble oxidants

83. Leaching zinc is the cause of
   a) encrustation around fittings of copper pipe
   b) pinhole leaks in brass pipe
   c) rusting in galvanized pipe
   d) galvanic corrosion
   e) a green color at copper pipe fittings

84. A pressure tank
   a) can be located near a well pump house or in a basement
   b) is an indication that the well pump is inadequate
   c) compensates for lack of water in the well
   d) is illegal in most building jurisdictions
   e) is only needed for wells that yield less than 5 gallons per minute

85. Central heating
   a) can be easily extended to room additions
   b) always includes the basement
   c) has fewer moving parts than area heaters
   d) can usually be designed to provide distribution to separate zones
   e) automatically balances itself

86. Every heating system should have
   a) a humidistat
   b) a setback thermostat with timer
   c) multiple zones
   d) a master shutoff switch
   e) a filter

87. A condition under which a furnace must be replaced is
   a) a cracked heat exchanger
   b) introduction of carbon monoxide into the air return
   c) blocking of the chimney
   d) being older than the anticipated service life of the furnace
   e) being over 30 years old

88. Examples of a warm air heating system are
   a) single-pipe steam systems
   b) forced-air heating systems with ducts
   c) two-pipe steam heating system
   d) electric baseboard heating system
   e) all of the above

89. A multizoned forced-air heating system
   a) provides truly independent operation of one zone from another
   b) can only be controlled by a single thermostat
   c) requires motorized duct dampers
   d) bypasses potential dangers of failed heat exchangers
   e) has no moving parts

90. An "octopus" heating system
   a) is another term for a system that smells like dead fish
   b) requires air filters to function properly

c) will not work in thin atmospheres above 1,000 feet
d) requires frequent lubrication of the blower motor
e) is another term for a large gravity warm air central furnace

91. Condensing furnaces are
a) furnaces with reduced clearances that fit into a condensed space
b) the new generation of high efficiency furnaces
c) furnaces that extract the heat from a refrigeration system
d) furnaces without internal air-conditioning coils
e) types with a single heat exchanger cell

92. An extended plenum is
a) a type of warm air distribution system
b) the same as a radial duct system
c) a platform to provide clearance above a garage or basement floor
d) a vibration damper in a duct system
e) always composed of square or rectangular ducts

93. The optimum location for heating registers is
a) high on the wall or on the ceiling opposite from windows
b) in the center of each room's ceiling
c) along the outside walls and below windows
d) as close as possible to the return air inlet
e) where it will produce a "short cycle"

94. All hot water heating systems must be supplied with
a) a pump
b) an expansion tank
c) radiators
d) storage-tank boilers
e) an oil filter

95. Older coal-fired gravity-type boilers
a) cannot be converted to oil or gas
b) should be replaced with domestic water heaters
c) are typically cast iron and very large
d) produce condensation and require drain piping
e) operate at pressures in excess of 100 PSI

96. A gravity hot water heating system does not need an expansion tank
a) True
b) False

97. An advantage of a forced hot-water heating system is
a) immunity from problems with freezing
b) easy adaptability to air-conditioning
c) lack of any moving parts requiring maintenance
d) the absence of a need for any other energy source other than water pressure
e) less heat fluctuation and adaptability to be multiple zones

98. When a boiler is also used to provide heat for domestic hot water, it requires a
a) flow-control valve
b) second pressure relief valve
c) seasonal setting on the thermostat
d) sight glass
e) supplementary circulating pump

99. A pressure-temperature gauge is
a) a necessary safety control on a hot-water heating system
b) an operational control
c) illegal, as there must be separate gauges for each function
d) only necessary on steam heating
e) sometimes provided with a scale in altitude

100. A quick means to determine whether a boiler is for steam or hot water is
   a) whether the system was installed before 1900
   b) if it is hooked up to the municipal steam supply
   c) if it includes a water-level gauge (sight glass)
   d) if the tank has welded pressure boiler plates
   e) the presence of a steam whistle

101. A Hartford loop can be found on
   a) gas-fired water heaters
   b) steam boilers
   c) hot water boilers
   d) oil-fired water heaters
   e) air-conditioning piping

102. The purpose of a Hartford loop is to
   a) provide a more efficient balancing of steam to the radiators
   b) prime the system and prevent freezing
   c) allow expansion and contraction of piping with changes in temperature
   d) prevent water from draining out of the boiler in the event of a leak
   e) qualify the house for insurance by a certain company in Connecticut

103. A blow-off valve is
   a) an automatically self-cleaning valve
   b) a part of the fuel delivery system in an oil burner
   c) part of an externally mounted low water cutoff valve
   d) a phrase to be used when disclaiming any knowledge about steam heating
   e) a type of pressure relief valve

104. A one-pipe steam heating system can be easily converted to an induced hot-water heating system by changing the radiators and adding a circulator pump and expansion tank.
   a) True
   b) False

105. Oil burners
   a) burn cleaner than gas burners
   b) do not require maintenance
   c) must bear an approval stamp from the American Gas Association
   d) are often maintained by the companies that supply the fuel oil
   e) do not require a source of combustion air

106. Condensing furnaces
   a) can be vented into masonry chimneys
   b) are noisier than conventional furnaces
   c) have less air flow than conventional furnaces
   d) must have vertical or near-vertical exhaust piping
   e) can be combined with vents from a gas-fired water heater

107. Gas burners
   a) should have a blue flame with little or no yellow
   b) should have a flame that lifts well above the burner
   c) should have an orange flame
   d) will normally vary in height depending on the barometric pressure
   e) do not produce soot or water vapor when functioning properly

108. A water heater temperature and pressure relief valve
   a) is not required for modern efficient water heaters
   b) can be placed in either the cold or hot water piping near the tank
   c) must always be placed in the lower third of the tank

d) should connect to the inlet with the dip tube
e) should be provided with a discharge pipe ending in a visible location

109. A rumbling noise from a water heater tank
a) is an indication of imminent explosion
b) means that the pressure-relief valve is blocked
c) is a result of loose gas or oil burners
d) results from accumulated sediment
e) means that the water heater will leak soon

110. A dripping water heater relief valve could be an indication of
a) excessive high pressure
b) defective valve
c) thermal expansion
d) all of the above
e) none of the above

111. Recovery rate most often refers to
a) air-conditioning compressor
b) warm air furnace capacity
c) tank-type water heaters
d) hot water heating systems
e) water pressure reducing valves

112. A water heater should be replaced when
a) the burners malfunction
b) the flue is worn out
c) the sacrificial anode contains magnesium
d) it is more than 10 years old
e) it leaks

113. Electric water heaters
a) are usually instantaneous, with no storage tank
b) use much more energy and are not allowed by most cities for energy conservation reasons
c) have lower recovery rates than gas- or oil-fired water heaters
d) have higher recovery rates than gas, so the tanks are usually smaller
e) always require a mixing valve to prevent scalding

114. One ton of cooling capacity
a) means that the air conditioner will produce one ton of ice per hour
b) is needed for approximately every 700 square feet of living space
c) is a measure of pressure in an average size compressor
d) is the weight of the air cooled by a one-ton air conditioner each hour
e) is equal to 2,000 Btu of cooling capacity

115. An air-conditioning condenser
a) draws approximately 7 amps per ton of cooling capacity
b) must be provided with a condensate drain pipe
c) should be located in a basement or furnace room
d) must be protected from rainfall
e) will weigh approximately one ton for each ton of cooling capacity

116. An evaporator coil is
a) part of a hydronic heating system used for cooling interior air
b) a ring on the outside of an air-conditioning compressor
c) an adaptation often found on air-conditioning when distilling spirits
d) located on the building exterior
e) the place where coolant changes state from a liquid to a gas

117. An electrical disconnect switch
a) is needed to deactivate the evaporator coils

- b) must be turned on to energize the condensing coils
- c) is not needed except for the circuit breaker controlling the furnace
- d) is required for a light near the condenser
- e) should be found near the portion of the air-conditioning system containing the compressor

118. A condensate trap in air-conditioning
   - a) is to prevent any condensate from escaping
   - b) means that the system has trapped too much cold air
   - c) prevents the blower from sucking air up through the drain line
   - d) must be vented to dissipate sewer gases
   - e) recycles the coolant to the condenser

119. When a split-system air conditioner is also part of a heating system
   - a) the blower should have at least two speeds
   - b) there should be a separate set of smaller ducts for the air conditioner
   - c) the split should be healed before operating the heating system
   - d) the coolant can be run through the furnace heat exchanger
   - e) the compressor should be next to the furnace plenum

120. The temperature of air discharging from the registers during the cooling cycle
   - a) should be 10 to 20 degrees lower than the outside air
   - b) should be 15 degrees lower than the temperature during the heating cycle
   - c) should be no more than 60 degrees
   - d) should be about 15 degrees lower than the temperature of the air in the room
   - e) should be at least 20 degrees lower than the temperature of the return air

121. A heat pump should
   - a) not ever need supplementary electric-resistance heating
   - b) be sized based upon the cooling load, not the heating load
   - c) be tested in both the cooling and heating modes to ensure that the compressor works
   - d) be located entirely within the building
   - e) not be tested when the outside air temperature is 40 degrees

122. Evaporative coolers are most effective where?
   - a) East Coast
   - b) Florida
   - c) Southwest
   - d) California
   - e) Washington and Oregon

123. If attic insulation does not have a vapor barrier
   - a) one should be laid over the top of it in an attic
   - b) one cannot be added to it or the manufacturer's warranty is void
   - c) one can be laid under the insulation so that it is adjacent to the heated ceiling
   - d) one can be added to the underside of the roof deck
   - e) it is because a vapor barrier is never necessary

124. The average water loss in one week in a swimming pool due to evaporation is
   - a) less than 1 inch
   - b) 1 inch
   - c) 1½ inches
   - d) 2 inches
   - e) 2½ inches

125. Radon gas
   a) is one of the most significant forms of manmade pollution
   b) is odorless, tasteless, invisible, and formed by natural uranium deposits
   c) is a result of high frequency radio waves generated for communications lines
   d) can be prevented from forming when a house is properly insulated
   e) most often enters a home through shower heads

126. Asbestos
   a) kills thousands of people annually from asbestos-related cancers
   b) must always be removed prior to sale of a home
   c) should never be encapsulated, as this only increases its dangers
   d) can be removed only by a licensed heating or insulating contractor
   e) in any level is a health risk within a home

127. Mold growth in a house thrives on moisture and
   a) cotton
   b) wool
   c) paper
   d) wood
   e) all of the above

128. A desuperheater
   a) is used to steam clean clothing
   b) eliminates static electricity in a house
   c) is a device on a geothermal system that is used to heat water
   d) uses gas as an energy source
   e) is a device on newly designed oil burners

Quiz answers

1. a
2. b
3. b
4. c
5. b
6. e
7. d
8. d
9. b
10. b
11. a
12. c
13. b
14. c
15. b
16. e
17. a
18. a
19. c
20. c
21. e
22. b
23. a
24. e
25. b
26. e
27. d
28. c
29. b
30. b
31. d
32. e
33. d
34. c
35. e
36. a
37. b
38. b
39. d
40. e
41. d
42. c
43. a
44. e
45. b
46. c
47. e
48. c
49. d
50. a
51. d
52. c
53. e
54. d
55. c
56. c
57. c
58. e
59. b
60. e
61. d
62. e
63. c
64. c
65. e
66. d
67. c
68. a
69. d
70. c
71. e
72. c
73. d
74. c
75. a
76. a
77. e
78. c
79. b
80. d
81. a
82. c
83. b
84. a
85. d
86. d
87. a
88. b
89. c
90. e
91. b
92. a
93. c
94. b
95. c
96. b
97. e
98. a
99. e
100. c
101. b
102. d
103. c
104. b
105. d
106. b
107. a
108. e
109. d
110. d
111. c
112. e
113. c
114. b
115. a
116. e
117. e
118. c
119. a
120. d
121. b
122. c
123. c
124. a
125. b
126. a
127. e
128. c

ASHI Standards of Practice

1. INTRODUCTION

The American Society of Home Inspectors®, Inc. (ASHI®) is a not-for-profit professional society established in 1976. Membership in ASHI is voluntary and its members are private home *inspectors.* ASHI's objectives include promotion of excellence within the profession and continual improvement of its members' inspection services to the public.

2. PURPOSE AND SCOPE

2.1 The purpose of the Standards of Practice is to establish a minimum and uniform standard for home *inspectors* who subscribe to these Standards of Practice. *Home inspections* performed to these Standards of Practice are intended to provide the client with objective information regarding the condition of the *systems* and *components* of the home as *inspected* at the time of the *home inspection.* Redundancy in the description of the requirements, limitations, and exclusions regarding the scope of the *home inspection* is provided for emphasis only.

2.2 Inspectors shall:

- **A.** adhere to the Code of Ethics of the American Society of Home Inspectors.
- **B.** *inspect readily accessible,* visually observable, *installed systems* and *components* listed in these Standards of Practice.
- **C.** report:
  1. those *systems* and *components inspected* that, in the professional judgment of the *inspector,* are not functioning properly, significantly deficient, *unsafe,* or are near the end of their service lives.
  2. recommendations to correct, or monitor for future correction, the deficiencies *reported* in 2.2.C.1, or items needing *further evaluation.* (*Per* Exclusion 13.2.A.5 *inspectors* are NOT required to determine methods, materials, or costs of corrections.)
  3. reasoning or explanation as to the nature of the deficiencies *reported* in 2.2.C.1, that are not self-evident.
  4. *systems* and *components* designated for inspection in these Standards of Practice that were present at the time of the *home inspection* but were not *inspected* and the reason(s) they were not *inspected.*

2.3 These Standards of Practice are not intended to limit *inspectors* from:

- **A.** including other inspection services or *systems* and *components* in addition to those required in Section 2.2.B.
- **B.** designing or specifying repairs, provided the *inspector* is appropriately qualified and willing to do so.
- **C.** excluding *systems* and *components* from the inspection if requested by the client.

3. STRUCTURAL COMPONENTS

3.1 The *inspector* shall:

- **A.** *inspect*:
  1. *structural components* including the foundation and framing.
  2. by probing a *representative number* of *structural components* where deterioration is suspected or where clear indications of possible deterioration exist. Probing is NOT required when probing would damage any finished surface or where no deterioration is visible or presumed to exist.

B. *describe:*
   1. the methods used to *inspect underfloor crawl spaces* and attics.
   2. the foundation.
   3. the floor structure.
   4. the wall structure.
   5. the ceiling structure.
   6. the roof structure.

3.2 The inspector is NOT required to:

A. provide any *engineering* or architectural services or analysis.

B. offer an opinion as to the adequacy of any *structural system* or *component*

4. EXTERIOR

4.1 The inspector shall:

A. *inspect:*
   1. *siding,* flashing and trim.
   2. all exterior doors.
   3. attached or adjacent decks, balconies, stoops, steps, porches, and their associated railings.
   4. eaves, soffits, and fascias where accessible from the ground level.
   5. vegetation, grading, surface drainage, and retaining walls that are likely to adversely affect the building.
   6. adjacent or entryway walkways, patios, and driveways.

B. *describe:*
   1. *siding.*

4.2 The *inspector* is NOT required to *inspect*

A. screening, shutters, awnings, and similar seasonal accessories.

B. fences.

C. geological and/or soil conditions.

D. *recreational facilities.*

E. outbuildings other than garages and carports.

F. seawalls, break-walls, and docks.

G. erosion control and earth stabilization measures.

5. ROOFING

5.1 The inspector shall:

A. *inspect:*
   1. roofing materials.
   2. *roof drainage systems*.
   3. flashing.
   4. skylights, chimneys, and roof penetrations.

B. *describe:*
   1. roofing materials.
   2. methods used to *inspect* the roofing.

5.2 The inspector is NOT required to inspect:

A. antennae.

B. interiors of flues or chimneys that are not *readily accessible*.

C. other *installed accessories*.

6. PLUMBING

6.1 The inspector shall:

A. *inspect:*
   1. interior water supply and distribution *systems* including all fixtures and faucets.
   2. drain, waste, and vent *systems* including all fixtures.
   3. water heating equipment and hot water supply system.
   4. vent *systems*, flues, and chimneys.
   5. fuel storage and fuel distribution *systems*.
   6. drainage sumps, sump pumps, and related piping.

B. *describe:*
   1. water supply, drain, waste, and vent piping materials.
   2. water heating equipment including energy source(s).
   3. location of main water and fuel shut-off valves.

6.2 The Inspector is NOT required to:

A. *inspect:*
   1. clothes washing machine connections.
   2. interiors of flues or chimneys that are not *readily accessible*.
   3. wells, well pumps, or water storage related equipment.
   4. water conditioning *systems*.
   5. solar water heating *systems*.
   6. fire and lawn sprinkler *systems*.
   7. private waste disposal *systems*.

B. determine:
   1. whether water supply and waste disposal *systems* are public or private.
   2. water supply quantity or quality.

C. operate *automatic safety controls* or manual stop valves.

7. ELECTRICAL

7.1 The inspector shall:

A. *inspect*:
   1. service drop.
   2. service entrance conductors, cables, and raceways.
   3. service equipment and main disconnects.
   4. service grounding.
   5. interior *components* of service panels and sub panels.
   6. conductors.
   7. overcurrent protection devices.
   8. a *representative number* of *installed* lighting fixtures, switches, and receptacles.
   9. ground fault circuit interrupters.

B. *describe:*
   1. amperage and voltage rating of the service.
   2. location of main disconnect(s) and sub panels.
   3. presence of solid conductor aluminum branch circuit wiring.
   4. presence or absence of smoke detectors.
   5. *wiring methods.*

7.2 The inspector is NOT required to:

A. *inspect:*
   1. remote control devices.
   2. *alarm systems* and *components*.
   3. low voltage wiring *systems and components*.
   4. ancillary wiring *systems* and *components* that are not a part of the primary electrical power distribution *system*.

B. measure amperage, voltage, or impedance.

8. HEATING

8.1 The inspector shall:

A. open *readily openable access panels.*

B. *inspect:*
   1. *installed* heating equipment.
   2. vent *systems,* flues, and chimneys.

C. *describe:*
   1. energy source(s).
   2. heating *systems.*

8.2 The inspector is NOT required to:

A. *inspect:*
   1. interiors of flues or chimneys that are not *readily accessible.*
   2. heat exchangers.
   3. humidifiers or dehumidifiers.
   4. electronic air filters
   5. solar space heating *systems.*

B. determine heat supply adequacy or distribution balance.

9. AIR CONDITIONING

9.1 The inspector shall:

A. open *readily openable access panels.*

B. *inspect:*
   1. central and through-wall equipment.
   2. distribution *systems*

C. *describe:*
   1. energy source(s).
   2. cooling *systems.*

9.2 The inspector is NOT required to:

- **A.** *inspect* electronic air filters.
- **B.** determine cooling supply adequacy or distribution balance.
- **C.** *inspect* window air conditioning units.

10. INTERIORS

10.1 The *inspector* shall *inspect:*

- **A.** walls, ceilings, and floors.
- **B.** steps, stairways, and railings.
- **C.** countertops and a *representative number* of *installed* cabinets.
- **D.** a *representative number* of doors and windows.
- **E.** garage doors and garage door operators.

10.2 The *inspector* is NOT required to *inspect:*

- **A.** paint, wallpaper, and other finish treatments.
- **B.** carpeting.
- **C.** window treatments.
- **D.** central vacuum *systems.*
- **E.** *household appliances.*
- **F.** *recreational facilities.*

11. INSULATION & VENTILATION

11.1 The inspector shall:

- **A.** *inspect:*
  1. insulation and vapor retarders in unfinished spaces.
  2. ventilation of attics and foundation areas.
  3. mechanical ventilation *systems.*
- **B.** *describe:*
  1. insulation and vapor retarders in unfinished spaces.
  2. absence of insulation in unfinished spaces at conditioned surfaces.

11.2 The *inspector* is NOT required to disturb insulation.

See 13.2.A.11 and 13.2.A.12.

12. FIREPLACES AND SOLID FUEL BURNING APPLIANCES

12.1 The *inspector* shall:

- **A.** *inspect:*
  1. *system components.*
  2. chimney and vents.
- **B.** *describe:*
  1. fireplaces and *solid fuel burning appli*ances.
  2. chimneys.

12.2 The inspector is NOT required to:

- **A.** *inspect:*
  1. interiors of flues or chimneys.
  2. fire-screens and doors.
  3. seals and gaskets.
  4. automatic fuel feed devices.
  5. mantles and fireplace surrounds.
  6. combustion make-up air devices.
  7. heat distribution assists (gravity fed and fan assisted).
- **B.** ignite or extinguish fires.
- **C.** determine draft characteristics.
- **D.** move fireplace inserts and stoves or firebox contents.

13. GENERAL LIMITATIONS AND EXCLUSIONS

13.1 General limitations:

- **A.** The *inspector is* NOT required to perform any action or make any determination not specifically stated in these Standards of Practice.
- **B.** Inspections performed in accordance with these Standards of Practice:
  1. are not *technically exhaustive.*
  2. are not required to identify concealed. conditions, latent defects, or consequential damage(s).
- **C.** These Standards of Practice are applicable to buildings with four or fewer dwelling units and their garages or carports.

13.2 General exclusions:

A. *Inspectors* are NOT required to determine:

1. conditions of *systems* or *components* that are not *readily accessible.*
2. remaining life expectancy of any *system* or *component.*
3. strength, adequacy, effectiveness, or efficiency of any *system* or *component.*
4. the causes of any condition or deficiency.
5. methods, materials, or costs of corrections.
6. future conditions including but not limited to failure of *systems* and *components.*
7. the suitability of the property for any specialized use.
8. compliance with regulatory requirements (codes, regulations, laws, ordinances, etc.).
9. market value of the property or its marketability.
10. the advisability of purchase of the property.
11. the presence of potentially hazardous plants or animals including, but not limited to, wood destroying organisms or diseases harmful to humans including molds or mold-like substances.
12. the presence of any environmental hazards including, but not limited to, toxins, carcinogens, noise, and contaminants in soil, water, and air.
13. the effectiveness of any *system installed* or method utilized to control or remove suspected hazardous substances.
14. operating costs of *systems* or *components.*
15. acoustical properties of any *system* or *component.*
16. soil conditions relating to geotechnical or hydrologic specialties.

B. *Inspectors* are NOT required to offer:

1. or perform any act or service contrary to law.
2. or perform *engineering* services.
3. or perform any trade or any professional service other than *home inspection.*
4. warranties or guarantees of any kind.

C. *Inspectors* are NOT required to operate:

1. any *system* or *component* that is *shut down* or otherwise inoperable.
2. any *system* or *component* that does not respond to *normal operating controls*.
3. shut-off valves or manual stop valves.

D. *Inspectors* are NOT required to enter:

1. any area that will, in the opinion of the *inspector,* likely be dangerous to the *inspector* or other persons or damage the property or its *systems* or *components.*
2. *under-floor crawl spaces* or attics that are not *readily accessible.*

E. *Inspectors* are NOT required to inspect:

1. underground items including but not limited to underground storage tanks or other underground indications of their presence, whether abandoned or active.
2. items that are not *installed.*
3. *installed decorative* items.
4. items in areas that are not entered in accordance with 13.2.D.
5. detached structures other than garages and carports.

6. common elements or common areas in multi-unit housing, such as condominium properties or cooperative housing.

F. *Inspectors* are NOT required to:

1. perform any procedure or operation that will, in the opinion of the *inspector,* likely be dangerous to the *inspector* or other persons or damage the property or its *systems* or *components.*
2. describe or report on any *system* or *component* that is not included in these Standards and was not *inspected.*
3. move personal property, furniture, equipment, plants, soil, snow, ice, or debris.
4. *dismantle* any *system* or *component*, *except* as explicitly required by these Standards of Practice.

ASHI Standards of Practice glossary of italicized terms

Alarm Systems
Warning devices *installed* or free-standing including but not limited to smoke detectors, carbon monoxide detectors, flue gas, and other spillage detectors, and security equipment

Automatic Safety Controls
Devices designed and *installed* to protect *systems* and *components* from unsafe conditions

Component
A part of a *system*

Decorative
Ornamental; not required for the proper operation of the essential *systems* and *components* of a home

Describe
To identify (in writing) a *system* or *component* by its type or other distinguishing characteristics

Dismantle
To take apart or remove any *component*, device, or piece of equipment that would not be taken apart or removed by a homeowner in the course of normal maintenance

Engineering
The application of scientific knowledge for the design, control, or use of building structures, equipment, or apparatus

Further Evaluation
Examination and analysis by a qualified professional, tradesman, or service technician beyond that provided by the *home inspection*

Home Inspection
The process by which an inspector visually examines the *readily accessible systems* and *components* of a home and which *describes* those *systems* and *components* in accordance with these Standards of Practice

Household Appliances
Kitchen, laundry, and similar appliances, whether *installed* or free-standing

Inspect
To examine any *system* or *component* of a building in accordance with these Standards of Practice, using *normal operating controls* and opening *readily openable access panels*

Inspector
A person hired to examine any *system* or *component* of a building in accordance with these Standards of Practice

Installed
Attached such that removal requires tools

Normal Operating Controls
Devices such as thermostats, switches, or valves intended to be operated by the homeowner

Readily Accessible
Available for visual inspection without requiring moving of personal property, *dismantling*, destructive measures, or any action that will likely involve risk to persons or property

Readily Openable Access Panel
A panel provided for homeowner inspection and maintenance that is *readily accessible*, within normal reach, can be removed by one person, and is not sealed in place

Recreational Facilities
Spas, saunas, steam baths, swimming pools, exercise, entertainment, athletic playground or other similar equipment, and associated accessories

Report
Communicate in writing

Representative Number
One *component* per room for multiple similar interior *components* such as windows, and electric receptacles; one *component* on each side of the building for multiple similar exterior *components*

Root Drainage Systems
Components used to carry water off a roof and away from a building

Shut Down
A state in which a *system* or *component* cannot be operated by *normal operating controls*

Siding
Exterior wall covering and cladding; such as: aluminum, asphalt, brick, cement/asbestos, EIFS, stone, stucco, veneer, vinyl, wood, etc.

Solid Fuel Burning Appliances
A hearth and fire chamber or similar prepared place in which a fire may be built and that is built in conjunction with a chimney; or a listed assembly of a fire chamber, its chimney, and related factory-made parts designed for unit assembly without requiring field construction

Structural Component
A *component* that supports non-variable forces or weights (dead loads) and variable forces or weights (live loads)

System
A combination of interacting or interdependent *components*, assembled to carry out one or more functions.

Technically Exhaustive
An investigation that involves *dismantling*, the extensive use of advanced techniques, measurements, instruments, testing, calculations, or other means

Under-floor Crawl Space
The area within the confines of the foundation and between the ground and the underside of the floor

Unsafe
A condition in a *readily accessible, installed system* or *component* that is judged to be a significant risk of bodily injury during normal, day-to-day use; the risk may be due to damage, deterioration, improper installation, or a change in accepted residential construction standards

Wiring Methods
Identification of electrical conductors or wires by their general type, such as non-metallic sheathed cable, armored cable, or knob and tube, etc.

ASHI Code of Ethics

For the home inspection profession

Integrity, honesty, and objectivity are fundamental principles embodied by this Code, which sets forth obligations of ethical conduct for the home inspection profession. The Membership of ASHI has adopted this Code to provide high ethical standards to safeguard the public and the profession.

Inspectors shall comply with this Code, shall avoid association with any enterprise whose practices violate this Code, and shall strive to uphold, maintain, and improve the integrity, reputation, and practice of the home inspection profession.

1. **Inspectors shall avoid conflicts of interest or activities that compromise, or appear to compromise, professional independence, objectivity, or inspection integrity.**

   A. Inspectors shall not inspect properties for compensation in which they have, or expect to have, a financial interest.
   B. Inspectors shall not inspect properties under contingent arrangements whereby any compensation or future referrals are dependent on reported findings or on the sale of a property.
   C. Inspectors shall not directly or indirectly compensate realty agents, or other parties having a financial interest in closing or settlement of real estate transactions, for the referral of inspections or for inclusion on a list of recommended inspectors, preferred providers, or similar arrangements.
   D. Inspectors shall not receive compensation for an inspection from more than one party unless agreed to by the client(s).
   E. Inspectors shall not accept compensation, directly or indirectly, for recommending contractors, services, or products to inspection clients or other parties having an interest in inspected properties.
   F. Inspectors shall not repair, replace, or upgrade, for compensation, systems or components covered by ASHI Standards of Practice, for one year after the inspection.

2. **Inspectors shall act in good faith toward each client and other interested parties.**

   A. Inspectors shall perform services and express opinions based on genuine conviction and only within their areas of education, training, or experience.
   B. Inspectors shall be objective in their reporting and not knowingly understate or overstate the significance of reported conditions.
   C. Inspectors shall not disclose inspection results or client information without client approval. Inspectors, at their discretion, may disclose observed immediate safety hazards to occupants exposed to such hazards, when feasible.

3. **Inspectors shall avoid activities that may harm the public, discredit themselves, or reduce public confidence in the profession.**

   A. Advertising, marketing, and promotion of inspectors' services or qualifications shall not be fraudulent, false, deceptive, or misleading.
   B. Inspectors shall report substantive and willful violations of this Code to the Society.

ASHI membership categories and requirements

All ASHI members are required to abide by the ASHI Code of Ethics, which specifically forbids home inspectors from active brokerage or sale of real estate, or performing repairs on homes that they inspect. ASHI members also agree to perform and report inspections in accordance with the ASHI Standards of Practice. Not only is the ASHI logo widely regarded as the mark of a professional home inspector, its use by the ASHI membership is also carefully controlled.

There are three designations of active professional home inspector membership in ASHI. These ASHI membership categories are as follows:

ASHI Certified Inspectors—Inspectors in this category have:

- Passed the National Home Inspector Examination and ASHI's Standards and Ethics module.
- Had inspection reports successfully verified for compliance with ASHI's Standards of Practice.
- Submitted valid proof of performance of at least 250 fee-paid home inspections that meet or exceed the ASHI Standards of Practice.

ASHI Associates w/ Logo—Inspectors in this category have:

- Passed the National Home Inspector Examination and ASHI's Standards and Ethics module.
- ASHI has verified performance of 50 fee-paid inspections in substantial compliance with the Standards of Practice.
- Had inspection reports successfully verified for compliance with ASHI's Standards of Practice.

ASHI Associates—Inspectors in this category have:

- Just joined ASHI and may be new to the inspection profession or may be a seasoned inspector who has not yet completed ASHI's requirements to move up in membership and logo use.

© Copyright 2009 American Society of Home Inspectors

Glossary

alligatoring Extensive surface cracking in a pattern that resembles the hide of an alligator.

areaway An open subsurface space around a basement window or doorway; provides light, ventilation, and access.

backfill The gravel or earth replaced in the space around a building wall after the foundation is in place.

bottom plate The bottom horizontal member of a frame wall.

Btu (British thermal unit) The amount of heat required to raise the temperature of 1 pound of water by 1 degree Fahrenheit.

cantilever A structural member that projects beyond its supporting wall or column.

Cedar Breather A three-dimensional nylon matrix that is stiff enough to resist crushing, thereby allowing air movement to the underside of wood shingles and shakes.

Conditioned crawl space A crawl area that has insulated foundation walls, closed vents, is heated and/or air conditioned and has a dirt floor that is covered over with large plastic sheets or cement. Basically an area that is dry and comfortable all year long.

control joint A groove that is formed, sawed, or tooled in a concrete or masonry structure to regulate the location and amount of cracking.

cornice Horizontal projection at the top of a wall or under the overhanging portion of the roof.

cross connection A direct arrangement of a piping line that allows the potable water supply to be connected to a line that contains a contaminant.

cupping An inward-curling distortion at the exposed corners of asphalt shingles.

downdraft A downward current of air in a chimney, often carrying smoke with it.

dry well A covered pit, either with open-jointed lining or filled with coarse aggregate, through which drainage from downspouts or foundation footing drains may seep into the surrounding soil.

eaves The lower edge of a roof that projects beyond the building wall.

efflorescence A white powdery substance appearing on masonry wall surfaces, composed of soluble salts that have been brought to the surface by water or moisture movement.

fascia (or facia) A horizontal board nailed vertically to the ends of roof rafters; sometimes supports a gutter.

feathering The tapering of one surface into another.

flashing Sheet-metal or other thin, impervious material used around roof and wall junctions to protect joints from water penetration.

flue A passageway in a chimney for conveying smoke, gases, or fumes to the outside air.

frost line The depth of frost penetration in soil; varies in different parts of the country.

gable roof A double-sloped roof from the ridge to the eaves; the end section appears as an inverted V.

girder The main structural support beam in a wood-framed floor; a girder supports one end of each joist.

glulam (glued laminated lumber) An engineered structural timber product composed of several layers of dimensioned timber glued together.

By laminating several smaller pieces of timber, a single large, strong, structural member is manufactured.

grade The ground level at the outside walls of a building or elsewhere on a building site.

header A framing member that crosses and supports the ends of joists.

hip roof A roof that slopes upward from all four sides of a building.

hydro-air Basically a hot water/warm air system, which essentially consists of an air handler that looks like the blower compartment of a furnace, a hot water boiler, and a duct system.

joist One of a series of parallel beams used to support floor and ceiling loads, supported in turn by larger beams (*girders*) or bearing walls.

knee wall A wall that acts as a brace by supporting roof rafters at an intermediate position along their length.

Lally column A steel tube, filled with concrete, used to support girders and other floor beams.

lath A building material of wood, metal, gypsum, or insulating board fastened to the frame of a building to act as a plaster base.

lintel A horizontal structural member that supports the load over an opening such as a door or window.

nosing The rounded edge of a stair tread that projects over the riser.

parapet wall The part of a wall that extends above the roofline.

pier A masonry column, usually rectangular in horizontal cross section, used to support other structural members.

pigtail (1) A flexible conductor attached to a light fixture that provides a means of connecting the fixture to a circuit. (2) A short length of copper conductor that is attached to the end of an aluminum branch circuit by a special fastener, then fastened to the terminal of a switch or outlet.

pilaster A pier-type projection of the foundation wall used to support a floor girder or stiffen the wall.

plenum A chamber or large duct above a furnace that serves as a distribution area.

ply A sheet in a layered construction, such as plywood, roofing.

pointing (repointing) Filling open mortar joints; removing deteriorated mortar between joints of masonry units and replacing it with new mortar.

rafter One of a series of inclined structural roof members spanning from an exterior wall to a center ridge beam or ridge board.

refractory A material, usually nonmetallic, used to withstand high temperatures, as in the combustion chamber of an oil-fired heating system.

relay An electromechanical switch; a device in which changes in the current flow in one circuit are used to open or close electrical contacts in a second circuit.

resilient tile A manufactured interior floor covering that is resilient, such as vinyl or vinyl-asbestos tile.

ridge beam The beam or board placed on edge at the ridge (top) of the roof into which the upper ends of the rafters are fastened.

riser The vertical height of a stair step; also the vertical boards that close the space between the treads of a stairway.

sheathing The structural covering, usually wood boards or plywood, over a building's exterior studs or rafters.

sheave A wheel with a grooved rim (pulley).

shelter-tube Mud-type tube (tunnel) built by termites as a passageway between the ground and the source of food (wood).

sill plate The lowest member of the house framing resting on top of the foundation wall. Also called *mud sill*.

soffit The visible underside of a roof overhang or eaves.

stringer (step) One of the enclosed sides of a stair supporting the treads and risers.

stud One of a series of slender wood or metal vertical structural members placed as supporting elements in walls and partitions.

subfloor Boards or plywood laid on joists, over which a finished floor is to be laid.

swale A shallow depression in the ground to form a channel for storm-water drainage.

thermocouple A device consisting of two junctions of dissimilar metals; when the two junctions are at different temperatures, a voltage is generated. Used in controlling gas valves.

toenail Driving a nail at an angle into the corner of one wood-frame member to penetrate into a second member.

tread The horizontal board in a stairway on which feet are placed.

truss A combination of structural members usually arranged in triangular units to form a rigid framework for spanning between load-bearing walls.

weep hole A small opening at the bottom of a retaining wall or the lower section of a masonry veneer facing on a wood-frame exterior wall, that permits water to drain.

Inspection worksheets

Address:______________________________

EXTERIOR

Front Exposure: N E S W

Condition

1. **Sidewalks:** none_____ concrete_____ asphalt_____ stone_____ cracked_____ uneven_____ _______
2. **Exterior walls:** brick_____ stucco_____ stone_____ wood siding_____ fiber-cement siding_____
   wood/asbestos shingle_____ aluminum/vinyl siding_____ other_____ vines growing_____ _______
3. **Trim:** cracked_____ rotting_____ broken_____ loose_____ needs paint touch-up_____ _______
4. **Paths:** concrete_____ stone_____ brick_____ asphalt_____ cracked_____ settled_____
   overgrown_____ _______
5. **Steps:** concrete_____ stone_____ brick_____ wood_____ need repair_____
   Handrail:_____ _______
6. **Porch:** front/rear/left/right: cracked_____ rotting_____ chipped_____ loose_____ _______
7. **Windows:** wood_____ aluminum_____ cracked_____ broken_____ reputty_____
   broken cords_____ needs lubrication/adjustment_____ do not close properly_____
   corroding/rotting frames_____ thermal panes_____ faulty seals_____ _______
8. **Storms/screens:** none_____ partial_____ full_____ cracked_____ broken_____
   torn screens_____ _______
9. **Roof:** asphalt_____ wood_____ slate_____ tile_____ roll roofing_____ other_____
   Shingles: missing_____ loose_____ cracked_____ aging_____ eroded_____ _______
10. **Gutters/downspouts:** none_____ partial_____ full_____ built-in_____ aluminum_____
    copper_____ galvanized_____ wood_____ loose/sagging_____ open joints_____
    need splash plates/elbows_____ need cleaning_____ _______
11. **Chimney:** brick_____ masonry_____ prefabricated_____ needs repair_____ _______

Condition

12. **Garage:** none_____ attached_____ detached_____ car occupancy_____

    Doors: operational_____ need lubrication/adjustment_____ automatic opener_____ _______

13. **Driveway:** cracked_____ uneven_____ broken_____ heaving_____ drain_____ _______

14. **Patio:** yes_____ no_____ concrete_____ stone_____ brick_____ needs repair_____

    cracked_____ settled_____ _______

15. **Deck:** attached_____ free-standing_____ composite decking_____ wood_____

    cracked_____ rotting_____ broken_____ needs additional bracing_____

    loose railings/handrails_____ loose steps_____ loose posts_____ needs CO_____ _______

16. **Landscaping:** grass needs recultivation_____ shrubs need pruning_____

    Trees_____ overhanging roof_____ need pruning_____ dead/dying_____ _______

17. **Retaining walls:** brick_____ stone_____ concrete block_____ wood_____ need repair_____ _______

18. **Fencing:** wood/metal_____ broken_____ rotting/rusting_____ needs painting_____ _______

19. **Drainage/grading:** satisfactory_____ poor_____ low spots_____ needs regrading_____

    stairwell_____ window well_____ _______

20. **Termites:** no visible evidence:_____ visible evidence_____ where_____ _______

21. **Caulking:** needed around exterior joints: yes_____ no_____ _______

    Comments: ____

    ____

ATTIC AREA

○ **Type:** full_____ crawl_____ finished_____ unfinished_____

○ **Accessible:** yes_____ no_____ **Evidence of condensation:** yes_____ no_____

○ **Insulation:** floor_____ roof_____ walls_____ partial_____ adequate_____ inadequate_____
none visible_____ loose sections_____ improperly installed_____
Vapor barrier: yes_____ no_____

○ **Ventilation:** adequate_____ inadequate_____ none_____ need roof/gable vents_____ ridge vent_____
power ventilator_____ Attic fan: _____ operational: yes_____ no_____

○ **Past leakage:** chimney_____ plumbing vent stacks_____ rafters_____ other_____

○ **Firestopping:** needed around chimney and attic floor: yes_____ no_____

CRAWL SPACE (BASEMENT LEVEL)

○ **None:** _____ **Accessible:** yes_____ no_____ Conditioned space: yes_____ no_____

○ **Floor:** dirt_____ cement_____ asphalt_____ other_____
Vapor barrier needed for floor: yes_____ no_____
Past water seepage: yes_____ no_____

○ **Insulation needed:** yes_____ no_____ loose section(s)_____

○ **Ventilation:** adequate_____ inadequate_____ Dehumidifier needed: yes_____ no_____

○ **Termites:** evidence of_____ no visible evidence_____ conditions conducive to activity_____

Comments: ____

____

BASEMENT AND/OR UTILITY ROOM(S)

○ **Evidence of water seepage:** yes_____ no_____
water stains_____ efflorescence: on walls_____ on floor_____
swelled floor tiles/joints_____
peeling and flaking paint on walls_____ corrosion below radiators_____
Seepage condition extensive: yes_____ no_____
Sump pump: yes_____ no_____ Dehumidifier needed: yes_____ no_____

○ **Foundation walls:** concrete_____ block_____ stone_____ brick_____ other_____ **Condition**
cracked_____ reseal_____ cracked or loose mortar joints_____ repoint_____ _______

○ **Girders:** wood_____ steel_____ none_____ not visible_____ _______

○ **Joists:** (partially) covered_____ sagging_____ _______

○ **Columns:** wood_____ steel_____ screw-jack type_____ _______

○ **Mold:** evidence of: yes_____ no_____ _______

Comments: ____

____

ELECTROMECHANICALS

○ **Electricity**_____ amps 110-120 volts circuit breakers_____ fuses_____
Service: adequate_____ marginal_____ inadequate_____
Wiring/outlets: adequate_____ marginal_____ inadequate_____ knob-and-tube wiring_____
aluminum branch circuits_____

○ **Heating system:** forced_____ gravity_____; hot water_____ hot air_____ steam_____
high-efficiency condensing type_____ hydro-air_____ geothermal_____
oil_____ gas_____ electric_____; zones_____
Condition: good_____ fair_____ aging_____ poor_____
Buried oil tank_____ Gas meter adequately sized_____
Furnace room ventilation: adequate_____ inadequate_____
Fire-code plasterboard needed: yes_____ no_____
Asbestos insulation on heating pipes and/or boiler_____

○ **Domestic hot water:** tankless coil thru heating system_____ tankless wall unit_____
separate tank unit_____ capacity_____ recovery_____ age_____
oil_____ gas_____ electric______
Capacity: adequate_____ marginal_____ inadequate_____ recirculating system_____
Relief valve adequately sized_____
Indirect fired storage type_____
Condition: good_____ fair_____ aging_____ poor_____

○ **Plumbing:** Pipes and fittings: copper_____ brass_____ iron_____ PEX_____ other_____
Water flow: adequate_____ low_____
Drain lines: cast iron_____ galvanized_____ copper_____ lead_____ plastic_____
Condition: good_____ fair_____ aging_____ poor_____; leakage_____ corrosion_____
Water inlet: copper_____ iron_____ lead_____ plastic_____ inaccessible_____
Septic system: yes_____ no_____

○ **Well pump:** none_____ jet_____ submersible_____ other_____
Storage tank: yes_____ no_____ Insulation needed: yes_____ no_____
Relief valve needed: yes_____ no_____
Condition: good_____ fair_____ aging_____ poor_____

SWIMMING POOL

In-ground _______ Aboveground ________

○ **Safety fence:** Y___ N___

Needs repair: Y___ N___ Gate self-closing and self-latching: Y___ N___

○ **Masonry deck:** cracked___ chipped___ uneven joints___ open joints___

Grab rail___ adequately secured___ Ladder___ adequately secured___

Slide___ adequately secured___ Diving board___ cracked___ warped___

○ **Wood deck:** cracked___ rotting___ broken___ Handrail: Y___ N___ condition___

○ **Vinyl lined pool:**

○ **Lining:** stained___ tears___ discolored___

Lining pulled out of edge retainer: Y___ N___ Replacement recommended: Y___ N___

○ **Concrete pool side walls:**

Tiles___ cracked___ loose___ chipped___ missing___ Plaster___ flaking___ chipped___ cracked___

Paint___ faded___ flaking___ cracked___

○ **Skimmer:** Weir: moves freely___ broken___ missing___ Strainer basket: damaged___ missing___

○ **Pump:** noisy___ hot to the touch___ water dripping around pump___

Strainer basket: needs cleaning___

○ **Heater:** operational Y___ N___ Cycles on/off: Y___ N___

Rust dust and flakes around base of heater: Y___ N___

○ **Filter:** Pressure gauge operational: Y___ N___ Filter needs cleaning: Y___ N___

○ **Cover:** worn___ torn___ damaged___ Repairs needed___ Replacement recommended___

Comments:____

____

Room_____________

Condition

1. **Ceiling/walls:** plaster_____ sheetrock_____ Chinese wallboard_____
   Cracked: yes_____ no_____ Broken: yes_____ no_____
   Evidence of past leakage: wall_____ ceiling_____ peeling and flaking paint/paper_____
   old layers of paint_____ lead-based paint_____ ________
2. **Floor:** wood_____ bamboo_____ laminate_____ tile_____ concrete_____ carpet_____
   other_____squeaks_____ slopes_____ sagging_____ refinish_____ torn/chipped_____
   loose_____ ________
3. **Trim:** cracked_____ chipped_____ missing_____ lead-based paint_____ ________
4. **Electrical outlets:** not observable_____ number_____ should have_____ ________
5. **Doors:** Exterior:_____ weatherstripped: yes_____ no_____ sliding glass_____
   Interior:_____ cracked_____ chipped_____ broken_____ pocket door_____ ________
6. **Hardware** (locks, hinges, knobs, handles): missing_____ needs repair_____ ________
7. **Closets:** yes_____ no_____ number_____ ________
8. **Heating:** radiators_____ baseboard_____ air registers_____ convectors_____
   radiation panels_____ none_____ ________
9. **Fireplace:** yes_____ no_____ cracked/chipped mortar joints_____ needs cleaning_____
   Damper: missing: yes_____ no_____ needs repair: yes_____ no_____
   indication of backsmoking_____ ________
10. **Cabinets/counter:** none_____ kitchen_____ medicine/vanity_____
    cracked_____ chipped_____ loose_____ ________
11. **Plumbing fixtures:** yes_____ no_____ leaks_____ Operational: yes_____ no_____
    Water flow: adequate_____ low_____ Drainage: adequate_____ blockage_____
    Regrouting/caulking needed at tub/shower joints: yes_____ no_____ ________
12. **Exhaust fan:** yes_____ no_____ operational_____ ________

Comments:____

____

Room_____________

Condition

1. **Ceiling/walls:** plaster_____ sheetrock_____ Chinese wallboard_____
   Cracked: yes_____ no_____ Broken: yes_____ no_____
   Evidence of past leakage: wall_____ ceiling_____ peeling and flaking paint/paper_____
   old layers of paint_____ lead-based paint_____ _______
2. **Floor:** wood_____ bamboo_____ laminate_____ tile_____ concrete_____ carpet_____
   other_____squeaks_____ slopes_____ sagging_____ refinish_____ torn/chipped_____
   loose_____ _______
3. **Trim:** cracked_____ chipped_____ missing_____ lead-based paint_____ _______
4. **Electrical outlets:** not observable_____ number_____ should have_____ _______
5. **Doors:** Exterior:_____ weatherstripped: yes_____ no_____ sliding glass_____
   Interior:_____ cracked_____ chipped_____ broken_____ pocket door_____ _______
6. **Hardware** (locks, hinges, knobs, handles): missing_____ needs repair_____ _______
7. **Closets:** yes_____ no_____ number_____ _______
8. **Heating:** radiators_____ baseboard_____ air registers_____ convectors_____
   radiation panels_____ none_____ _______
9. **Fireplace:** yes_____ no_____ cracked/chipped mortar joints_____ needs cleaning_____
   Damper: missing: yes_____ no_____ needs repair: yes_____ no_____
   indication of backsmoking_____ _______
10. **Cabinets/counter:** none_____ kitchen_____ medicine/vanity_____
    cracked_____ chipped_____ loose_____ _______
11. **Plumbing fixtures:** yes_____ no_____ leaks_____ Operational: yes_____ no_____
    Water flow: adequate_____ low_____ Drainage: adequate_____ blockage_____
    Regrouting/caulking needed at tub/shower joints: yes_____ no_____ _______
12. **Exhaust fan:** yes_____ no_____ operational_____ _______

Comments:____

____

Index

A

air admittance valves (AAVs), 116–117
air-conditioning systems, 262–274
  basic components of, 262
  central, 264–272
  cooling capacity, 263–264
  energy considerations, 290–291
  evaporative cooler, 273
  geothermal heating/cooling, 218–219
  heat pump, 272–273
  inspection checklist, 273–274
antennas, 28–29
  inspection checklist, 35
arc fault circuit interrupter (AFCI), 179–180
American Society of Home Inspectors (ASHI)
  Code of Ethics, 343
  glossary of italicized terms, 341–342
  membership categories and requirements, 344
  Standards of Practice, 335–340
asbestos, 295–297
  shingles, 13, 47–48
attic, 114–123
  crawl, 114
  fans, 120
  fire hazards, 117–118
  full, 114
  heating system located, 115
  inspection checklist, 123
  insulation, 115–116, 284–287
  structure, 121–123
  truss, 122–123, 125
  ventilation, 118–119, 287–288
  violations, 116
  water leakage, 117

B

basement, 147–169
  dampness, 154–155
  floors and floor joints, 160–162
  foundation, 147–151
  furnace room, 163–165
  inspection checklist, 168–169
  plumbing, 197–202
  sump pump, 155–156, 161–162
  walls, 159–160
  water seepage, 155–163
bathrooms, 137–141
  electrical outlets, 179
bathrooms *continued*
  inspection checklist, 143–146
  shower diverter, 139
  water pressure and flow, 138–141
  whirlpool tub, 140
bedrooms, 136–137
  closet, 136–137
  inspection checklist, 145
  smoke detectors, 136
Board of Fire Underwriters, 181
boilers, 210, 221–223
  inspection checklist, 245

C

carpenter ants, 106–108
  inspection checklist, 112
caulking, 59–60
  acrylic latex, 60
  butyl-rubber, 60
  inspection checklist, 61
  oil-base, 60
  silicone-base, 60
  windows and weather-stripping, 289
ceilings, 125–128
  inspection checklist, 143–144
  sagging plaster, 125–126

ceilings *continued*
  tile, 127–128
  water stains, 126
central air, 264–272
  blower coil, 269–270
  compressor, 265–267
  condensate drain line, 268–270
  condenser, 267
  drain pan, 270–271
  ducts, 270–272
  evaporator, 267–269
  inspection procedure, 273–274
  integral system, 264
  split system, 264–265
  types of, 264
  vent stack, 269
cesspool, 187
chimneys, 22–26
  chimney-top damper, 25, 135
  cricket, 26–27
  inspection checklist, 34
  saddle (*see* cricket *above*)
chlordane, 95–96
circuit breakers, 174
contracts, 314–315
crawl space, 165–168
  conditioned crawl space, 167–168
  FHA standards, 165
  inspection checklist, 169

D

decks, 76–80
  bracing, 78
  column support, 77
  composite deck, 79–80
  free-standing deck, 79
  guardrails, 79
  inspection checklist, 82
  joist supports, 77–78
  spacing between planks, 78
domestic water heaters, 247–261
  (*see also* water heaters)
doors
  exterior, 56–57
  flush, 57
  inspection checklist for exterior, 61
  metal-clad insulated, 57
  pocket, 124
  solid-core flush, 57
  stile-and-rail, 57
  storm, 59
drainage, 62–70
  curtain drain, 66–67
  flood plain, 63
  groundwater/water table, 65–67
  inspection checklist, 81
  plumbing system, 185–186
  surface runoff, 63–65
driveways, 39–41
  inspection checklist, 41
dry well, 33–34
drywall (*see* plasterboard)

E

efflorescence, 52, 158–160
EIFS (synthetic stucco), 49–50
electrical heating systems, 236–237
  space heaters, 237
electrical outlets, 132–134
  inspection checklist, 144
electrical system, 170–182
  aluminum wiring, 176–177
  appliance power requirements, 173
  arc fault circuit interrupter (AFCI), 179–180
  capacity, 171–173
  fuses and circuit breakers, 174–175
  ground-fault interrupter (GFI), 179
  grounding, 177–178
  inlet electrical service, 170–171
  inlet service panel box, 175–176
  inspection checklist, 181–182
  knob-and-tube wiring, 180
  low-voltage switching systems, 180–181
  outlets (receptacles), 134, 178–180
  service switch, 171
  tamper-resistant receptacles, 134
  violations, 181
electromagnetic fields (EMF), 301–302
energy considerations, 283–291
engineered lumber, 153–154
environmental concerns, 292–303
  asbestos, 295–297
environmental concerns *continued*
  electromagnetic fields (EMF), 301–302
  formaldehyde, 300
  lead poisoning, 298–300
  leaky oil tanks, 300–301
  radon, 292–295
  water contamination, 297–298

F

Federal Housing Administration (FHA), crawl space standards, 165
fences, 80–81
  inspection checklist, 82
fireplaces, 134–136, 289–290
  chimney-top damper, 25–26, 135
  creosote, 135
  gas-fired, 136
  inspection checklist, 144
  recommended types of wood, 290
  soot (*see* creosote *above*)
  vent-free, 136
flood plain, 63
floors, 128–131
  bamboo flooring, 129–130
  carpet-covered (new), 129
  engineered hardwood floor, 129
  inspection checklist, 144
  laminate flooring, 129
forced hot-water heating systems, 221–227
  advantages/disadvantages, 223
  controls, 225–227
  distribution piping, 223–224
  domestic water heater, 226–227
  panel, 224–225
  pressure/temperature gauge, 227
  radiators, 224
  relief valve, 227
  zone control, 226
forced warm-air heating systems, 215–218
  air distribution systems 216–218
  supply registers and return grille, 216–218

formaldehyde, 300
foundation, 147–151
  bridging, 154
  cracks, 149–150
  girders, 151–152
  inspection checklist, 168
  settlement, 148–149
  structural support framing, 151–154
  Superior Walls, 150–151
furnaces, 210
  condensing, 215–216
  furnace room, 163–165, 169
  inspection checklist, 245
  pulse-combustion, 216
fuses, 174–175

G

garages, 84–94
  attached, 84–90
  detached, 90–93
  door photoelectric eye, 88–89
  door restraining cable, 88–89
  doors, 88–89
  electricity, 89, 92–93
  exterior of detached, 90–91
  fire and health hazards, 84–85
  flood potential, 88
  general considerations, 89–90
  heat, 92
  inspection checklist, 93–94
  interior of detached, 91–92
  plumbing problems, 86–88
gas burner heating systems, 235–236, 244
  inspection checklist, 246
geothermal heating/cooling, 218–219
gravity hot-water heating systems, 219–221
  expansion tanks, 220
gravity warm-air heating systems, 213–215
green home technology, 307–311
groundwater, 65–67, 202
  high level, 155–156
gutters and downspouts, 29–34
  built-in, 30
  exterior-mounted, 30–32
  inspection checklist, 35
  types of material, 30–31
gypsum board (*see* plasterboard)

H

hallway, 142–143
  inspection checklist, 146
heat pump, 218, 272–273
heating systems, 209–246
  central, 210–211
  combustion air, 239
  direct venting, 238–239
  electrical, 236–237
  energy considerations, 290–291
  forced hot-water, 221
  forced warm-air, 215
  gas burners, 244
  geothermal heating/cooling 218–219
  gravity hot-water, 219–221
  gravity warm-air, 213–215
  heat pump, 218, 272–273
  hot-water, 219–227
  hybrid, 231
  hydro-air, 231–232
  inspection checklist, 244–246
  inspection procedures, 237–244
  low-water cutoff, 227, 230
  oil burners, 243–244
  registers and radiators, 211
  space heaters, 236–237
  steam, 227–231
  thermostat and master shut-off, 211–212
  venting, 238–239
  warm-air, 212–219
home inspection
  contracts, 314–315
  electromechanical systems, 4
  exterior, 2–3
  final, 4
  interior, 3–4
  limitations, 312–313
  procedures, 1–4
  professional, 312
  quiz, 319–333
  quiz answers, 334
  warranties (*see* warranties)
  worksheets, 348–354
home inspector requirements, 316–318
hot-water heating systems, 219–227
  inspection checklist, 245–246
  inspection procedures, 241–242
hybrid heating systems, 231
  steam–hot water, 231

I

insects
  carpenter ants, 106–108
  inspection checklist, 112–113
  powder-post beetles, 108–109
  termites, 95–106
  wood-destroying, 95–109
instantaneous water heaters, 254, 259
insulation, 284–287
  attic, 115–116
  determining amount of, 285–286
  flexible, 284
  foamed-in-place, 284
  jacket for water heaters, 287
  loose-fill, 284
  R-value, 285
  recommended amounts of, 286
  types of, 284
  urea formaldehyde foam (UFFI), 284, 300
  vapor barrier, 286–287
interior rooms, 124–143
  bathrooms, 137–141
  bedrooms, 136–137
  cracks (truss uplift), 125
  electrical outlets, 132–134
  floors, 128–131
  hallway, 142–143
  inspection checklist, 143–146
  kitchen, 141–142
  kitchen island sink, 142
  staircase, 143
  walls and ceilings, 125–128

K

kitchen, 141–142
  inspection checklist, 145
  loop vent, 142

L

landscaping, 74–76
  decks, 76–80
  drainage, 62–65
  fences, 80–81
  foundation settlement, 148–149
  inspection checklist, 81–83
  lawn, 74–76

landscaping *continued*
  lawn sprinkler systems, 192–193
  retaining walls, 70–74
  shrubs, 75–76
  trees, 76
lead poisoning, 298–300
lightning protection, 29

M

moisture meter, 1, 126
mold, 302–303

O

oil burner heating systems, 233–235
  inspection procedures, 243–244
oil tanks
  buried, 300–301
  interior location, 234
  leaking, 300–301

P

patios, 38–39
  inspection checklist, 41
plasterboard, 125
plumbing, 184–207
  basement, 197–198
  brass pipes, 199
  curb valve, 184
  distribution piping, 198–199
  drainage pipes, 199–202
  drainage system, 185–186
  fixtures, 185, 193–194
  fresh air vent, 185–187, 192
  house service main, 185
  inspection checklist, 145, 207–208
  island sink vent, 142
  lawn sprinkler systems, 192–193
  PEX, 195–196
  pipes, 195–199
  polybutylene, 196–197
  problems in garages, 86–87
  septic systems, 193, 187–192
  sewage ejector pump, 201–202
  shower diverter, 139
  sink trap, 139–140
  sump pumps, 156, 161–162
  vent stack, 26–27, 116–117, 191–192

plumbing *continued*
  wall hatch, 194–195
  waste disposal system, 186–192
  water pressure and flow, 138–139, 194
  water supply and distribution, 184–185
  water-supply pipes in basement, 197–198
  well-pumping systems, 202–207
powder-post beetles, 108–109
  inspection checklist, 112

R

radon, 292–295
  EPA guidelines for residential, 293
  water supply contamination, 297–298
retaining walls, 70–74
  inspection checklist, 81
  types of, 70
roofs, 6–20
  built-up (BUR), 17–19
  discolored, 10
  flat, 16–20
  hatch, 28
  ice dam, 8
  inspection checklist, 20–21, 34–35
  metal, 15–16, 19–20
  pitched, 6–8
  roll roofing, 19
  shingles (*see* shingles)
  single-ply membrane, 19
  structures and projections, 22–34
  valley construction, 9
  vents, 27–28

S

screens, 58–59
septic tank, 187–192
  alternating drainfields, 190–191
  bacterial cleaner, 190
  cleaning, 190
  mound systems, 191–192
  water softeners, 190
sewage ejector pump, 201–202
sheathing paper, 43–44

shingles, 8
  asbestos-cement, 13, 47–48
  asphalt, 10–12
  clay tiles, 14–15
  life span of, 11
  on pitched roofs, 6–8
  slate, 13–14
  synthetic slate, 14
  wood, 44–45
  wood and shakes, 12–13
sidewalks, 36
  front and side pathways, 37
  inspection checklist, 41
siding, 44–49
  aluminum, 46–47
  asbestos-cement shingles, 47–48
  asphalt, 48–49
  EIFS (synthetic stucco), 49–50
  exterior, 44–49
  fiber-cement siding, 48
  hardboard, 46
  inspection checklist, 60–61
  masonry veneer, 50–52
  mineral-fiber shingles, 47–48
  plywood panels, 46
  stucco, 49
  synthetic stucco (EIFS), 49–50
  vinyl, 47
  wood, 44–46
  wood boards, 45–46
  wood shingles/shakes, 44–45
skylights, 28
smoke detectors, 136, 143
soil erosion, 63–64
space heaters, 236–237
  electric, 237
  gas, 237
staircase, 143
steam heating systems, 227–231
  advantages/disadvantages, 231
  controls, 229–231
  distribution piping, 228–229
  domestic water heater, 231
  inspection checklist, 246
  inspection procedures, 242–243
steps
  entry, 37–39
  inspection checklist, 41
  interior staircase, 143
stucco, 49
sump pumps, 156, 161–162

swimming pools, 275–282
  accessories, 280
  concrete, 275–276
  equipment, 278–279
  inspection checklist, 282
  inspection procedure, 280–282
  preformed fiberglass, 277–278
  vinyl-lined, 276–277
synthetic stucco, 49–50

T

termites, 95–106
  bait, 83
  dry-wood, 105
  Formosan, 106
  inspection checklist, 112
  subterranean, 98–105
  wood damaged by, 102
trim
  exterior, 52–53
  interior, 128

V

vents
  air admittance valves (AAVs), 116–117
  condensate drain line connected to, 269–270
  loop vents, 142
  stacks, 26–27, 115–117, 191–192

W

walls
  basement, 159–160
  Chinese dry wall, 128
  exterior, 43–52
  interior, 125–128
  load-bearing, 43
  masonry, 51–52
  mold-resistant drywall, 126
  siding (*see* siding)
  studs, 43
  Superior Walls, 150–151
  veneer, 43, 50–51
  water stains, 125–128
  wood-framed exterior, 43–44
warm-air heating systems, 212–218
  advantages, 213
  disadvantages, 213
  inspection procedures, 239–241
warranties
  manufacturers', 314
  private home inspection, 314
  real estate, 313–314
waste disposal system, 186–191
  cesspool, 187
  leaching field, 188
  septic tank, 187–188
water
  contamination of, 297–298
  ground, 65–67, 202–203
  high groundwater level, 155–156
  leaking into attic, 117
  pressure and flow, 138–141
  seepage inspection checklist, 168–169
  seeping into basement, 155–156
  seeping through foundation walls, 156–158
  stains on ceilings, 125–126
  supply requirements, 207
  table, 65
  usage rates, 207
water heaters
  capacity and recovery, 255–256
  determining age of, 255
  domestic, 247–260
  exhaust stack, 252
  expansion tank, 250–251
  indirect-fired storage, 260
  inspection checklist, 261
  instantaneous, 259
  insulation jacket for, 287
  operational inspection, 252–254
  pressure/temperature-relief valve, 249–251
  recirculation, 254
  relief valve, 249–251
  replacement, 254–255
water heaters *continued*
  reversed connection, 248–249
  rumbling noise, 251–252
  spill switch, 253
  tank, 248–256
  tankless (coil), 256–258
  tankless (coil) inspection procedures, 258–259
  tankless (wall-mounted) heater, 259
weep holes, 51
well pumps, 202–206
  jet, 203–204
  piston, 203
  submersible, 204
wells, 202
  general considerations, 206–207
  inspection checklist, 207
  pressure switch and gauge, 206
  storage (pressure) tanks, 205–206
windows, 53–56, 131–132
  awning, 54
  casement, 54
  caulking and weather-stripping, 289
  double-hung, 53
  fixed-pane, 55
  frames, 55
  glass used in, 55
  horizontal sliding, 53
  inspection checklist, 61, 144
  jalousie, 54
  repairing broken, 131–132
  screens, 58–59
  storm, 58, 288–289
  thermal-pane, 132–133
wood
  insects that destroy, 95–109
  rotting, 109–112
wood-burning stoves, 289–290
  recommended types of wood, 290
  (*see also* fireplaces)

Y

yard (*see* landscaping)

S. Gernsback's
Radio Encyclopedia

S. GERNSBACK'S

RADIO ENCYCLOPEDIA

SIDNEY GERNSBACK
PUBLISHER
NEW YORK

All rights reserved including that of translating
into foreign languages including the Scandinavian

Copyright 1927
by
SIDNEY GERNSBACK

Printed in the United States of America

A

"A" BATTERY—An appliance for furnishing electrical current to light the *filaments* of a *vacuum tube*, and cause emission of *electrons* (see *Theory of Vacuum Tube Operation*). This battery may be in any of several forms, the two in popular use being the storage type (see *Storage Battery*), and the dry cell (see *Dry Cell*). Fig. 1 shows method of connecting the "A" Battery in a vacuum tube circuit.

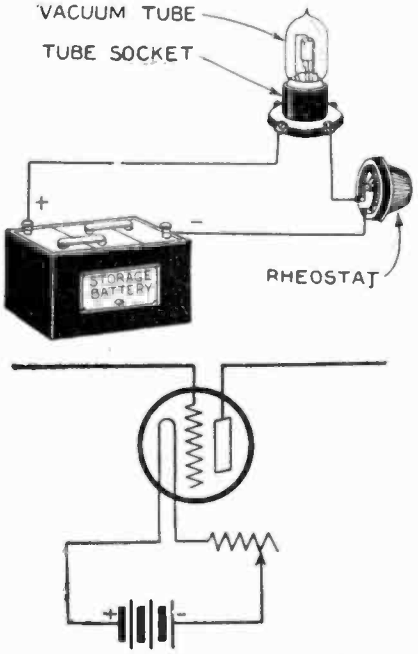
VACUUM TUBE
TUBE SOCKET
+
-
STORAGE BATTERY
RHEOSTAT
+
-

Fig. 1. Method of connecting the "A" battery to a vacuum tube in radio receiving circuits. The lower diagram shows the "A" battery as represented in schematic drawings

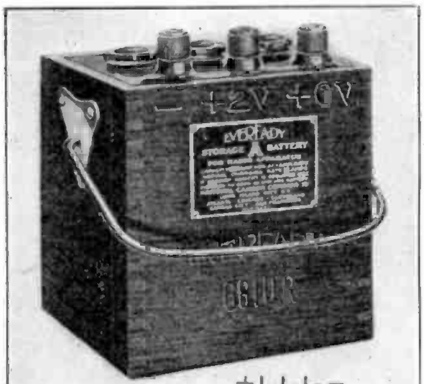
+2V +6V
EVEREADY
STORAGE A BATTERY
SYMBOL
+
-

Photos by courtesy of National Carbon Co., Inc.

Fig. 2. A storage battery such as used for "A" battery current supply

In Fig. 2 a popular type of storage battery is shown. Batteries of this type come in many individual forms, the usual voltage being standardized at 6 volts. However, for use with certain tubes, 2 and 4 volt storage batteries are available. For general radio work, using the low-consumption tubes, the battery, as shown in Fig. 2, can hold a charge of from 60-80 ampere hours. (*See Storage Battery.*)

The other common form of "A" Battery is the dry cell, usually having a difference of potential (voltage) of 1.5 volts. Fig. 3 illustrates a popular type of dry cell. In connecting a number of cells for use with a vacuum tube, the method of connecting will depend on the voltage rating of the particular tube. If the tube is rated at 1.5 volts, a single dry cell may be used as in Fig. 4, or several may be connected in *parallel* (*q.v.*) shown in Fig. 5, the total voltage however, being only the voltage of one cell. Three dry cells connected in *series* (*q.v.*) will give 4½ volts, Fig. 6. A connection for dry cells, known as *series-parallel* or *multiple* (*q.v.*), is shown in Fig. 7. This connection will also produce 4½ volts and at the same time increase the life of the cells. Four dry cells connected in series will give six volts for use with a standard six-volt tube, Fig. 8. Dry cells are only used with six-volt tubes where the tube has a current rating of ¼ ampere, and seldom more than one tube is used in this manner due to the relatively high discharge and consequent short life of the battery.

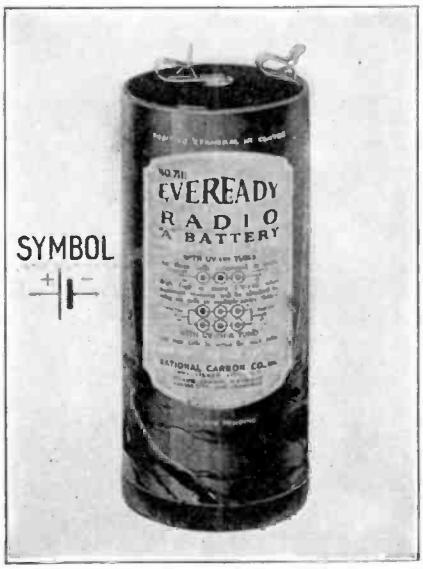
SYMBOL
+
-
No. 711
EVEREADY
RADIO
"A" BATTERY
NATIONAL CARBON CO. INC.

Fig. 3. A dry cell "A" battery

A storage battery can be used for the "A" Battery current with a 4-volt tube if the connection is made as in Fig. 9. Here a standard 6-volt storage battery is tapped at the four volt terminal.

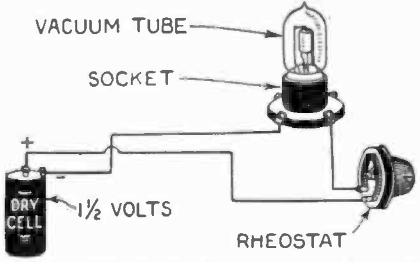
VACUUM TUBE
SOCKET
+
-
DRY CELL
1½ VOLTS
RHEOSTAT

Fig. 4. This circuit shows how one dry cell is used as the "A" battery for types of tubes requiring 1½ volts filament current

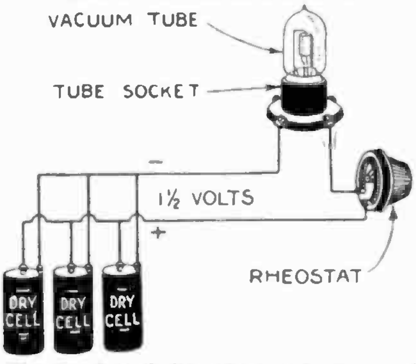
VACUUM TUBE
TUBE SOCKET
-
1½ VOLTS
+
RHEOSTAT
DRY CELL
DRY CELL
DRY CELL

Fig. 5. Several dry cells may be connected in parallel for increased amperage or longer life of the "A" battery

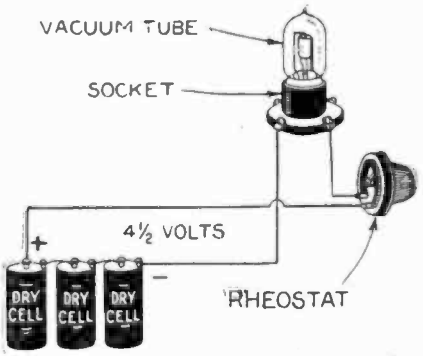
VACUUM TUBE
SOCKET
+
4½ VOLTS
-
DRY CELL
DRY CELL
DRY CELL
RHEOSTAT

Fig. 6. Three dry cells connected in series give 4½ volts, serving as the "A" battery for the U.V.199 or C299 type of tube

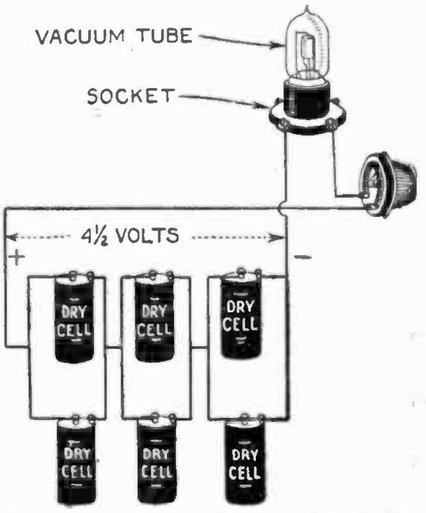
VACUUM TUBE
SOCKET
4½ VOLTS
+
-
DRY CELL
DRY CELL
DRY CELL
DRY CELL
DRY CELL
DRY CELL

Fig. 7. A series—parallel connection of three dry cells for longer life of the "A" battery with 4½ volt tubes

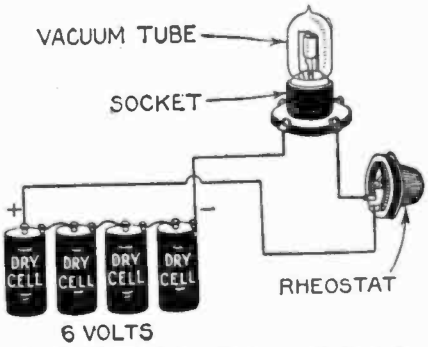
VACUUM TUBE
SOCKET
+
-
DRY CELL
DRY CELL
DRY CELL
DRY CELL
RHEOSTAT
6 VOLTS

Fig. 8. Four dry cells connected in series give 5 volts for low consumption tubes

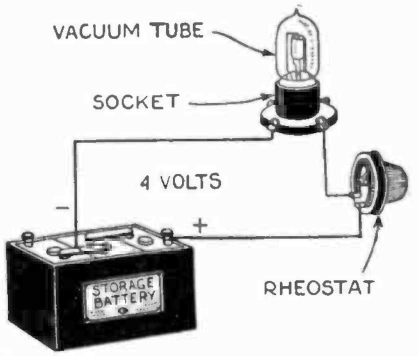
VACUUM TUBE
SOCKET
4 VOLTS
-
+
STORAGE BATTERY
RHEOSTAT

Fig. 9. Showing method of tapping at 4-volt terminal of a 6-volt storage battery for type U.V.199 or C299 tubes or other 4-volt tubes

Another connection for low voltage tubes can be made as illustrated in Fig. 10. Here only one cell of the storage battery is used. This cell can be used until exhausted when the other cells can be used in the same manner. When the three cells are exhausted, the storage battery may be charged by means of a *charger* (*q.v.*).

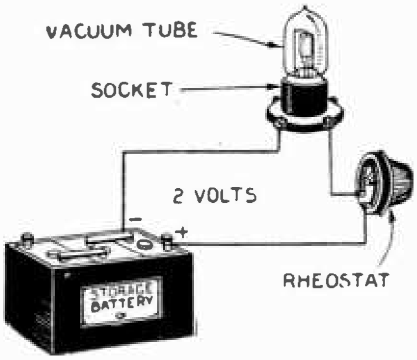

Fig. 10. How a single cell of a 6-volt storage battery may be employed to furnish "A" battery current for 1½ volt tubes

ABAMPERE—A unit of current used in special theoretical work. It has a value of 10 *amperes* (*q.v.*) and is not generally used in electrical work.

ABBREVIATIONS, AMATEUR—The transmitting amateurs of the United States use their own peculiar abbreviations in their inter-communication work. These are used to lessen the interference between transmitting stations as it permits conversation in a much quicker manner. The abbreviations generally used are listed below. Simplified spelling is always resorted to, words being spelled according to their sounds. Some such words are given, but others are so obvious that they are not listed. There are also the standard "Q" signals adopted by the International Radiotelegraphic Convention, to be found in official publications and given under the heading "Abbreviations, International Radiotelegraphic Convention."

ACCW—A transmitter using rectified alternating current as the plate supply.

AFTRN or P.M.—Afternoon.

AMMTR—Ammeter.

ARRL—American Radio Relay League.

BCL—Broadcast listener, one who does not have a transmitting station.

BC STATION—Broadcast Station.

BFRE—or B4—Before.

BLO—Blow.

BOTTLE or V.T.—Vacuum Tube.

C—See or Call.

CGE—Cage.

CM—City Manager of A. R. R. L.

CTPSE—Counterpoise.

CUL—See you later or call you later.

CW—Continuous wave transmitter.

DCCW—A transmitter using direct current as the supply for the plate.

DM—District Manager of A. R. R. L.

DOPE—Information, or in another sense, a varnish for coating inductance coils.

DS—District Superintendent of the A. R. R. L.

DX—Distance, receiving or transmitting.

FB—Fine Business.

FONE—Phone.

GA—Good afternoon.

GLD—Glad.

GM—Good morning.

GN—Good night.

FRQ—Frequency.

HAM—Being a term for the transmitting amateur.

HI—High or I am laughing—it's funny, according to use in sentence.

HI LOS—High loss.

HR—Here.

HV—Have.

KCK BCK—Kick back.

KONGRATS—Congratulations.

KY—Key.

LW LOS—Low loss.

LW WVE—Low waves.

MGHT—Might.

MIKE—Microphone.

MSG—Message.

NITE—Night.

OM—Old man.

OP—Operator.

OPS—Operators.

ORS—Official Radio Station.

OT—Oscillation transformer.

OW—Female operator.

PEEP-PEEP—Continuous wave transmitter.

QCW—Listen for my CW transmitter.

QCW?—Shall I listen to your CW transmitter?

QRMER—An operator who persists in interfering with other stations after being warned to stop.

QSR—I will relay the message.

QSR?—Will you relay the message?

QSS—You are fading.

QSS?—Is my signal fading?

RDO—Radio.

RECVR—Receiver.

RELY—Relay.

ROCKCRUSHER—Spark transmitter.

SA—Say.

SIGS—Signals.

SPIKE—Message file.

SPK—Speak.

SUREFIRE—Something positive in action.

THG—Thing.

THT—That.

TLL—Till.

TM—Traffic manager of A. R. R. L.

TMR—Tomorrow.

TRAFFIC—The handling of messages among the stations.

U—You.

UNLIS—Unlicensed.

UR—Your.

VLTMTR—Voltmeter.

WL—Will.

WRLS—Wireless.

WVL—Wavelength.

"X" LICENSE—Experimental License.

YL—Young lady.

"Z"—Special license.

73—Best regards.

ABBREVIATIONS — RADIO — ELECTRICAL—A letter or group of letters of the alphabet arranged so that there is a definite electrical meaning when used in radio telegraphy or telephony. These abbreviations are the means by which technical writings are simplified, *i.e.*, Alternating Current is simply A.C. when used in general telegraphy. As most of the latest radio text books use these abbreviations, the following are given to aid the reader in understanding the various abbreviations:

A.C.—Alternating Current.

A.F.—Audio Frequency.

B.E.M.F.—Back Electromotive Force.

B. & S.—Brown & Sharpe Wire Gauge.

B.S.W.G.—Birmingham Standard Wire Gauge.

C.—Capacity.

C.G.S. — Centimetre-gramme-second (*q.v.*).

C.W.—Continuous Waves.

D.C.—Direct Current.

D.C.C. — Double Cotton Covered (Wire).

D.S.C.—Double Silk Covered.

E.—Induced Pressure (See Abvolt).

E.M.F.—Electromotive Force.

H.F.—High Frequency.

H.F.C.—High Frequency Current.

I.R.E.—Institute of Radio Engineers.

L.—Inductance (in Formulae).

L.F.—Low Frequency.

L.F.C.—Low Frequency Current.

M.A.—Milliampere.

Mfd.—Microfarad.

P.D.—Potential Difference.

R.—Electrical Resistance.

R.F.—Radio Frequency.

R.P.M.—Revolutions per Minute.

R.P.S.—Revolutions per Second.

S.G.—Specific Gravity.

S.H.M.—Simple Harmonic Motion.

S.S.C.—Single Silk Covered.

S.W.G.—Standard Wire Gauge.

W/L—Wavelength.

X's—Atmospherics.

ABBREVIATIONS, INTERNATIONAL RADIO TELEGRAPHIC CONVENTION—All messages and conversations are not transmitted in regular form, various forms of abbreviations being widely used to increase the speed of transmission and to reduce cost. It also has the tendency to decrease the interference as messages are sent swiftly in half the time. In wireless telegraphy, abbreviations have been adopted as international, and today there are very few ships or shore stations that do not use at least part of them daily. Below we give the International Radiotelegraphic Abbreviations as adopted at The International Radiotelegraphic Convention. An example of the use of the abbreviations is as follows:

A station wishing to learn the name of a ship which is seen off the coast, sends in code the letters QRA?, which means, what is the name of your ship? The station or ship answers with the letters, QRA S.S. Lapland, thus in a few seconds the question had been answered.

Abbreviation	*Question*	*Answer or Notice*
CQ	—.—.—.—.—	Signal of inquiry made by a station desiring to communicate.
TR	— .—.	Signal announcing the sending of particulars concerning a station on shipboard.
	— —. .— —	Signal indicating that a station is about to send a high power.
PRB	Do you wish to communicate by means of the International Signal Code?	I wish to communicate by means of the International Signal Code.
QRA	What ship or coast station is that?	This is ....
QRB	What is your distance?	My distance is ....
QRC	What is your true bearing?	My true bearing is.........degrees.
QRD	Where are you bound for?	I am bound for.............
QRF	Where are you bound from?	I am bound from............
QRG	What line do you belong to?	I belong to the............Line.
QRH	What is your wave length in meters?	My wave length is..........meters.
QRJ	How many words have you to send?	I have .... words to send.
QRK	How do you receive me?	I am receiving well.
QRL	Are you receiving badly? Shall I send 20? ...—. for adjustment?	I am receiving badly. Please send 20. ...—. for adjustment.
QRM	Are you being interfered with?	I am being interfered with.
QRN	Are the atmospherics strong?	Atmospherics are very strong.
QRO	Shall I increase power?	Increase power.
QRP	Shall I decrease power?	Decrease power.
QRQ	Shall I send faster?	Send faster.
QRS	Shall I send slower?	Send slower.
QRT	Shall I stop sending?	Stop sending.
QRU	....	I have nothing to transmit. I have nothing for you.
QRV	Are you ready?	I am ready. All right now.
QRW	Are you busy?	I am busy (or, I am busy with.... Please do not interfere).
QRX	Shall I stand by?	Stand by. I will call you when required.
QRY	When will be my turn?	Your turn will be No. ....
QRZ	Are my signals weak?	Your signals are weak.
QSA	Are my signals strong?	Your signals are strong.
QSB	Is my tone bad?	The tone is bad.
	Is my spark bad?	The spark is bad.
QSC	Is my spacing bad?	Your spacing is bad.

(When an abbreviation is followed by a mark of interrogation, it refers to the question indicated for that abbreviation)

ABSCISSA—A term in geometry adapted to radio use in making *curves* to show various values. The abscissae are the horizontal lines, and the ordinates are the vertical lines. In the diagram, the line AY is the axis of *ordinates* (*q.v.*) and the line AX the axis of abscissae. Fig. 2 shows a simple curve indicating the relation between voltage and current (amperes) in an *arc* (*q.v.*). The volts are given as ordinates and the amperes are shown as abscissae.

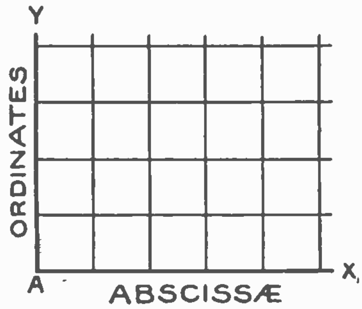

Fig. 1. Indicating axis of abscissae by the line A to X

The distance of any point from the axis of ordinates, measured on a line parallel to the other axis represents the value of the quantity and locates a point on the curve. Where abscissae and ordinates are known, it is a simple matter to plot a series of points and then draw the curve through these points. (See *curve.*) Thus, in Figure 2, the curve shows the operation of the arc to be such that when the voltage is zero, the current is 20 amperes; when the voltage is 60, the current is 6 amperes, etc.

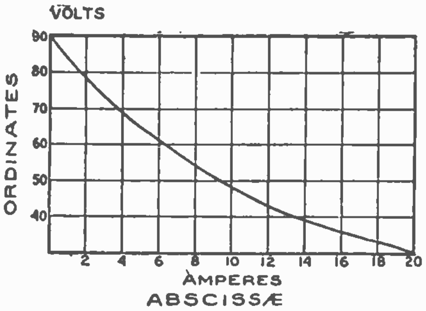

Fig. 2. A simple curve to illustrate the relation between volts and amperes

ABSORPTION, ELECTRIC—(See *Soaking-in.*)

ABSORPTION MODULATION—In the production of modulated (see *modulation*) waves, used to transmit speech in radio, it is necessary first to produce *undamped* (*q.v.*) waves of high or *radio frequency* (*q.v.*). Then other waves of low or *audio frequency* (*q.v.*) are superposed or run together with them. Now, if we insert a *microphone* (*q.v.*) in the *aerial circuit* (*q.v.*) of the transmitting station, providing that the *resistance* (*q.v.*) of that instrument is such as to reproduce accurately the vibrations of the voice or music, it is possible to vary the current output of the system producing the high frequency currents, so as to transmit the voice vibrations. If the microphone has resistance comparably low with the antenna or *antenna resistance* (*q.v.*), then the high frequency power output is handled (apportioned) equally between the aerial and microphone. The result is the production of fairly well modulated waves. There are, however, better methods of modulating the waves for which see *Modulation.*

ABVOLT—The *CGS unit* (*q.v.*) indicating the pressure or induced E.M.F. (*electromotive force*) (*q.v.*) set up between the ends of a wire one centimeter long, moving with a speed of one centimeter per second across the magnetic lines of force (*flux lines*) (*q.v.*) of a uniform magnetic field of unit intensity or flux density. An abvolt is therefore a very small unit and 100,000,000 (10^8) are required to equal one volt. In all formulae for determining the output or E.M.F. that can be expected to be produced by a certain *alternator* (*q.v.*), the result is in abvolts, and the complete formula usually provides for obtaining the actual volts. This is done by dividing the result in abvolts by 100,000,000 or 10^8 as it is generally expressed. (See *Alternator* also *abampere.*)

A. C.—Abbreviation for Alternating Current. (See *Alternating Current.*)

ACCELERATION—The rate at which the speed of a body in motion increases within a certain period of time. If we assume a train running at a certain rate of speed when the brakes are applied, the speed will decrease, and if measured over a period of one second, there will be a definite amount of decrease or a rate of decrease. This is known as *negative* acceleration or *deceleration.* When the train is speeded up there will be a definite change (increase) in velocity. The rate of change of velocity per unit of time is the acceleration. (*See Velocity* also *Frequency.*)

ACCEPTOR—A supplementary combination of *inductance* (*q.v.*) and *capacity* (*q.v.*) tuned to the frequency of the desired signal and connected in series in the receiving antenna. Generally, a *coil* (*q.v.*) and *condenser* (*q.v.*) of a value to correspond with a certain given wavelength—thus literally to "accept" the desired signal. (See also *Rejector.*)

A series combination of *inductance* (*q.v.*) and *capacity* (*q.v.*) connected in series with a high-frequency C.M.F. When tuned to resonance with the impressed E.M.F., the impedance drops to a low value, limited by the resistance of the system

ACCUMULATION OF ELECTRICITY—The process of storing electrical energy as (1)—a *storage battery* (*q.v.*) in which electric charges are placed to be later withdrawn and expended at will for various purposes, or (2)—the collection of energy by a *condenser* (*q.v.*) wherein the *electrons* (*q.v.*)

or minute charges of electricity are stored up momentarily on the plates or electrodes of the *condenser* (*q.v.*) and then released into the circuit.

ACCUMULATOR—(See *Storage Battery.*)

ACID—An active chemical compound formed by combining hydrogen with various acid radicals. Acids are much used in radio, particularly in *storage batteries* (*q.v.*). The most commonly used acid in this connection is sulphuric acid—a compound or union of hydrogen, sulphur and oxygen in certain definite proportions as indicated by the chemical symbol H_2SO_4. (See *Flux, Soldering.*)

ACIDIMETER or ACIDOMETER—An instrument used to determine the *specific gravity* (*q.v.*) of an acid solution. It is similar to a hydrometer and used in the same manner. Much used for testing strength or purity of acids or solutions. Operates on the principle that the strength of an acid will be directly proportional to the quantity of carbonic acid gas which it will liberate from a carbonate of soda or potash. (See *Hydrometer.*)

ACLINIC LINE—An imaginary line on the earth's surface assumed as passing through points where there is no magnetic inclination or dip of the needle of a compass. (See *Agonic Line.*)

ACOUSTICS—The science of sound or the study of the cause and effect of vibrations which effect the hearing. Generally speaking, the production and transmission of sound.

ACOUSTIC WAVE—The term occasionally used to denote a *sound wave.*

ACTINIC RAY—A light ray or beam of invisible *radiant energy* (*q.v.*). Rays of light, generally considered as lying at the blue or extreme left of the *spectrum* (*q.v.*) and being of such short *wavelength* as to be invisible to the eye. These rays have the power to induce or bring about chemical action. The most powerful actinic rays are the violet (X Ray) and the *ultra-violet.*

ACTIVE CONDUCTOR—(See *Conductor.*)

ACTIVE MATERIAL—The spongy part of a plate of a storage battery. The part which changes in nature and appearance due to the flow of electric current and having the ability to redevelop the current by a secondary chemical change.

ACTIVE PRESSURE—The active *electromotive force* (*q.v.*) or the pressure which produces a current. The term is used to distinguish this pressure from one impressed on the circuit.

A component in phase with the current in an *alternating current* (*q.v.*)

ACTIVE SPARK—A spark produced from energy contained in a charged condenser, which produces active *oscillations* (*q.v.*).

Thus, the spark discharge from a coil without any method of storing energy will be "inactive." When a condenser is inserted in such manner as to allow it to store up energy and then discharge, the resultant spark will be "active."

Fig. 1 shows a simple spark coil circuit with the secondary discharging across a gap. This is an Inactive Spark. Fig. 2 represents the same coil with a condenser in the circuit to store up the energy before discharging across the *gap* (*q.v.*). This is an Active Spark.

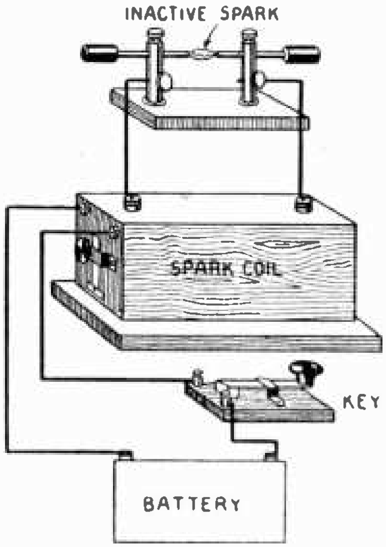

Fig. 1. Showing how an inactive spark is produced

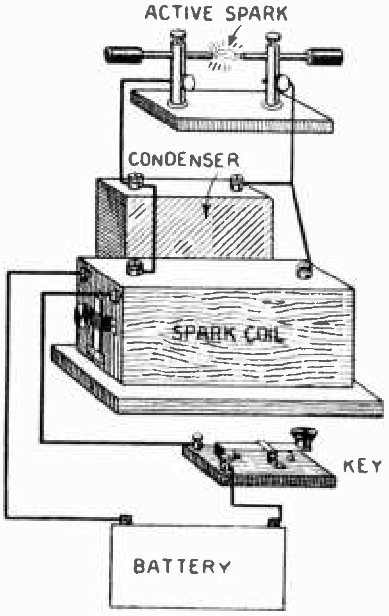

Fig. 2. The active spark is produced by means of a condenser in the circuit as shown

ADAPTER, AERIAL—A device for utilizing an electric light line as an aerial for the reception of radio signals. Essentially, two or more fixed mica condensers of comparatively high capacity, in series with two metal legs attached rigidly to a standard plug for insertion in socket of electric light fixture. Terminals are provided for connection to receiving apparatus. The illustration shows one of the popular types of Aerial Adapters which consists of merely a condenser to prevent the voltage of the line from passing directly to, or through, the instruments.

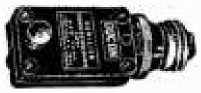

Illustration by courtesy of Dubilier Condenser and Radio Corp.

ADAPTER, TUBE—A device similar in form to a vacuum tube socket. Usually comprises a cylindrical form or shell of either metal or insulating material with legs identical with the tips or legs of a vacuum tube. A receptacle is allowed for the insertion of a tube, and the lead from each tip recess is so connected to a leg or tip at its base, as to permit the use of a non-standard base tube in a standard *socket* (*q.v.*). The illustration shows adapter for U.V. 199 or C-299 type of tube. When it is desired to employ a tube having a non-standard base in a set already equipped with standard sockets, without the necessity of changing the sockets, an adapter may be used. (See *Socket.*)

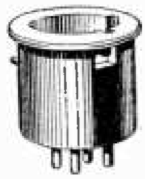

Illustsation by courtesy of Paccnt Elec. Co., Inc.

ADHESION, ELECTRIC—Affinity of one body for another due to dissimilar charges of electricity passing through or being carried by them. The electrical equivalent of magnetic attraction.

ADJUSTABLE CONDENSER—(See *Variable Condenser.*)

ADJUSTABLE GRID LEAK—A form of *grid leak* (*q.v.*) that can be easily adjusted over a wide range of values without removal from the circuit. The illustration shows one type of Adjustable Grid Leak.

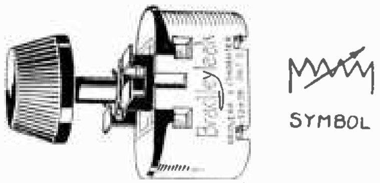

Illustration by courtesy of Allen-Bradley Co.

ADJUSTABLE RESISTANCE—(See *Variable Resistance.*)

ADMITTANCE—The inverse of *impedance* (*q.v.*). In an *alternating current* (*q.v.*) the Admittance acts in the opposite manner to impedance. A circuit having low impedance is said to have relatively high Admittance.

AERIAL—A system of wires suspended in the air or in any form in which they may be insulated or kept free from surrounding objects, used for the purpose of receiving or transmitting impulses as in radio transmission and reception. The term Aerial is actually identical with *Antenna*, but due to the general English definition of Antenna, authorities are gradually coming to distinction between transmission and reception Aerial systems by referring to a receiving device of this nature as an Antenna, and when used to transmit, as an Aerial.

There are different types and forms of Aerials, varying mainly according to purpose and facilities. The standard and perhaps more commonly used receiving Aerial consists of a single wire as in Fig. 1 or several wires as in Fig. 2. The wire used is usually of copper, as this metal combines low resistance with economy, strength, and durability. The theory of the action of the waves either received by or transmitted from an Aerial system will be found under the heading "Electromagnetic Waves."

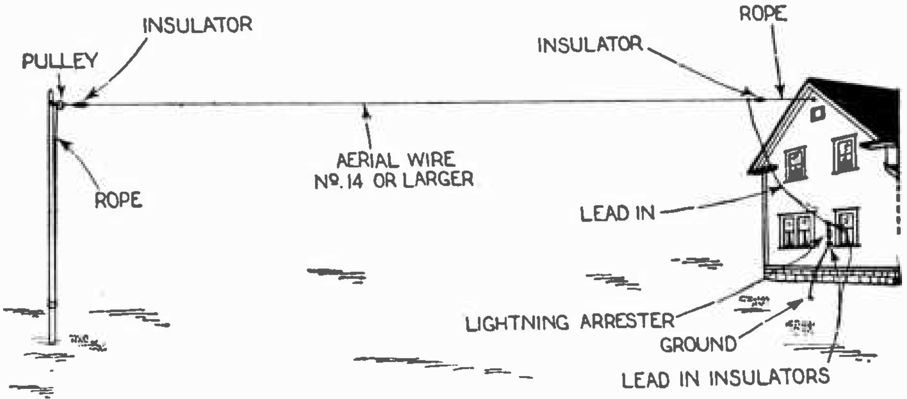

Fig. 1. A single wire aerial in the form of an inverted L. If a long wire is used, the lead-is may be connected in the center, forming a T type

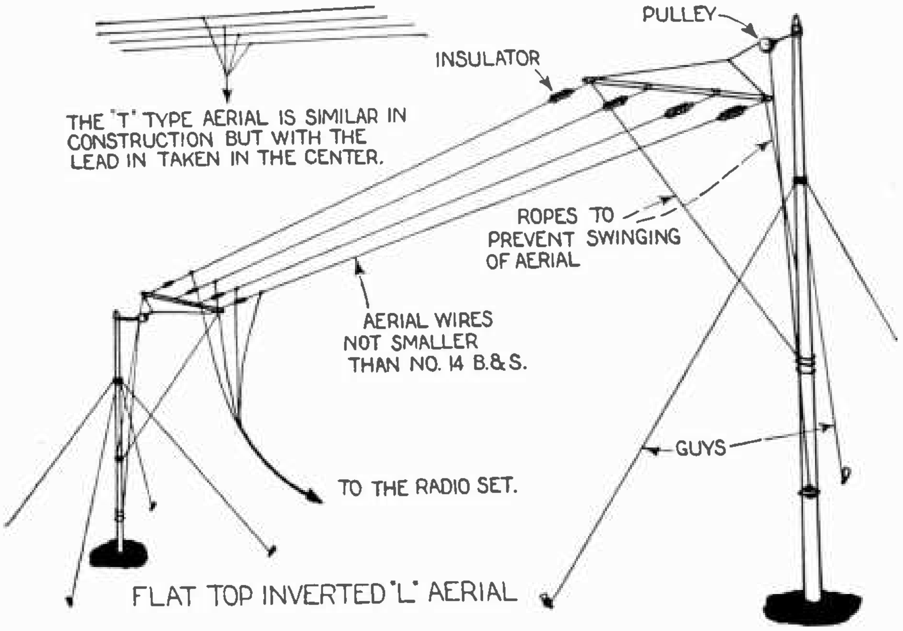
PULLEY
INSULATOR
THE "T" TYPE AERIAL IS SIMILAR IN CONSTRUCTION BUT WITH THE LEAD IN TAKEN IN THE CENTER.
ROPES TO PREVENT SWINGING OF AERIAL
AERIAL WIRES NOT SMALLER THAN NO. 14 B.&S.
GUYS
TO THE RADIO SET.
FLAT TOP INVERTED "L" AERIAL

Fig. 2. Method of erecting a multi-wire aerial. This type of aerial may be made in either the inverted L or T form as shown

The exact type and extent of an Aerial will depend entirely on the available facilities and the particular purpose for which it is intended. These various forms will be taken up under their respective headings. In the receiving Aerials, Figs. 1 and 2, a certain definite *wavelength* (*q.v.*) is obtained by arranging a certain length to the wires of the Aerial system including the *ground* and *lead-in.* (See *Fundamental Wave-Length of Aerials.*) In the single wire Aerial, the lead-in is usually taken from the end nearest the receiving apparatus, but where a particularly long Aerial is to be used, it is customary to attach the lead-in to the centre of the suspended wire. The former is known as an *inverted L* type and the latter as a *T* type Aerial. These designations will apply as well where a number of wires are used as in Fig. 2.

Generally speaking, an Aerial should be as high above the ground as possible, of course within practical limits. From 25 to 100 feet is the most practical height to use. Where the Aerial is suspended at a substantial height the incoming waves will be less likely to be obstructed by surrounding trees or buildings, and in the case of transmission the same factor will apply, inasmuch as the waves may be obstructed in their transmission, if buildings or trees closely surround the Aerial.

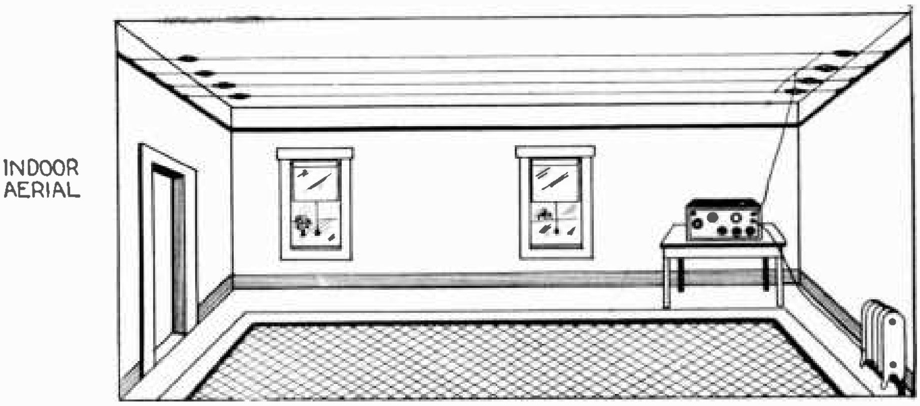

Fig. 3. Showing one method of installing an indoor aerial. While this illustrates a multi-wire type, a single wire run around the moulding of the room may give excellent results

The insulation of an Aerial is of particular importance, and especially so in the case of a transmitting Aerial. The reason for insulating an Aerial is to prevent the currents from escaping to the ground by way of trees, buildings, or the masts to which the Aerial is attached.

Fig. 3 shows a method of arranging an indoor Antenna system where it is not possible to use the out-door type. The construction may be essentially the same as the usual form of outdoor Aerial and can be arranged in a number of different ways. Very often a wire is merely suspended around two sides of a room and insulated from the moulding. One end will be connected as a lead-in to the receiving apparatus. In constructing an indoor Antenna, it is well to remember that the effective length of the Antenna is not much greater than the direct distance from the receiving set to the farthest point of the Antenna.

Fig. 4 shows a different type of Aerial known as the *Cage* type. The illustration shows an inverted L, and the T arrangement. Fig. 5 shows a *Fan Aerial* and Fig. 6 the *umbrella* type. The illustration, Fig. 7, shows one type of *loop Aerial* (*q.v.*) and Fig. 8 the same general type in somewhat different form. Owing to the fact that the theory of operation of a *loop Aerial* is somewhat different than the ordinary types, it will be described more fully under the heading "Loop Aerial." Fig. 9 illustrates a form of *Aerial Adapter* which is arranged to allow the electric light wires to be used as an Aerial. (See *Adapter Aerial.*) Fig. 10 shows an *underground Aerial* system designed by Dr. Rogers and which is not appreciably affected by atmospheric disturbances. (See *Strays.*) This form of Aerial, however, is generally impractical due to the difficulties in construction.

Aerials for transmitting may be of the same general types as receiving Antennas arranged outdoors, with the exception as explained previously, that heavy insulators must be used owing to the high voltage employed. Aerial for transmitting will depend for size on the wave-length desired.

Fig. 11 illustrates several types and arrangements of *counterpoise* (*q.v.*). In the case of the receiving Aerial a single wire may be from 75 to 150 feet in length, or if a number of wires are used, the over-all length may be considerably less. (See *Fundamental Wave-length.*)

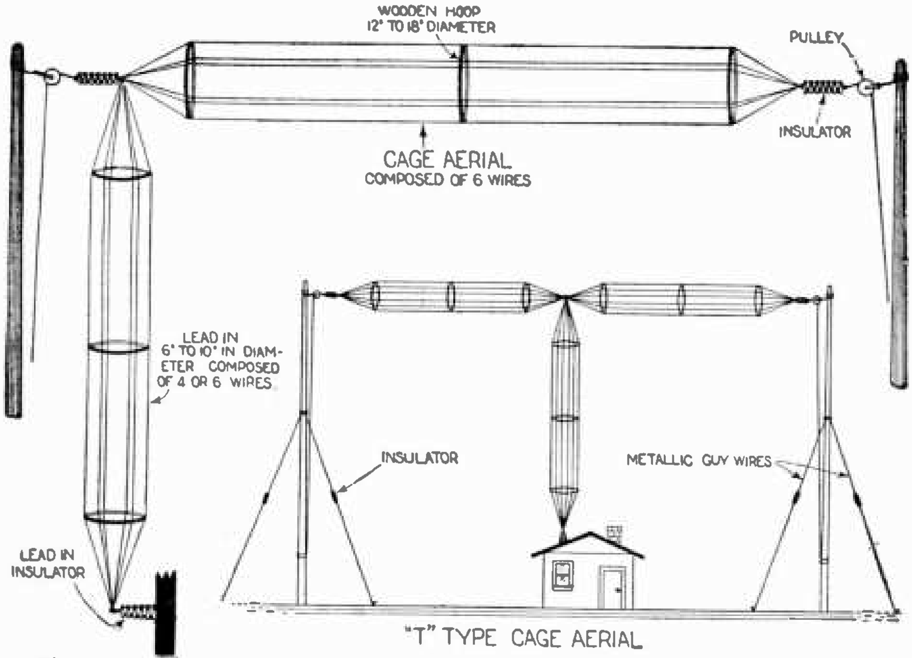
WOODEN HOOP
12" TO 18" DIAMETER
PULLEY
INSULATOR
CAGE AERIAL
COMPOSED OF 6 WIRES
LEAD IN
6" TO 10" IN DIAMETER COMPOSED OF 4 OR 6 WIRES
LEAD IN
INSULATOR
INSULATOR
METALLIC GUY WIRES
"T" TYPE CAGE AERIAL

Fig. 4. Cage aerials are usually employed for transmitting purposes. The inverted L and T types are erected as shown

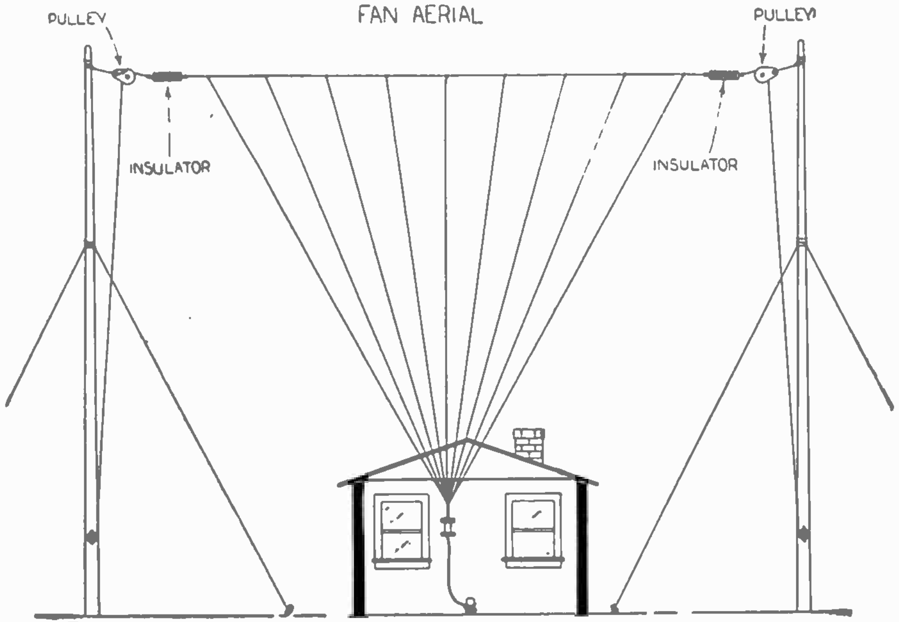
FAN AERIAL
PULLEY
PULLEY
INSULATOR
INSULATOR

Fig. 5. The fan aerial has its wires spread apart at the farthest point, forming a fan

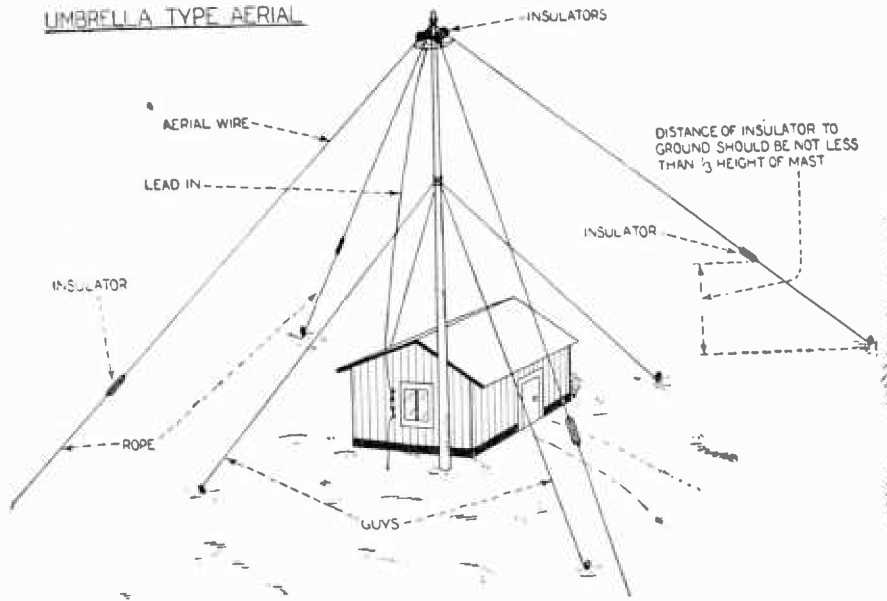
UMBRELLA TYPE AERIAL
INSULATORS
AERIAL WIRE
LEAD IN
DISTANCE OF INSULATOR TO GROUND SHOULD BE NOT LESS THAN 1/3 HEIGHT OF MAST
INSULATOR
INSULATOR
ROPE
GUYS

Fig. 6. The umbrella type of aerial derived its name from the arrangement of its wires in the form of an umbrella

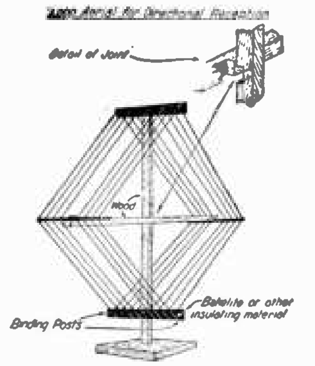
Loop Aerial for Directional Reception
Detail of Joint
Wood
Bakelite or other insulating material
Binding Posts

Fig. 7. The selenoid type loop aerial

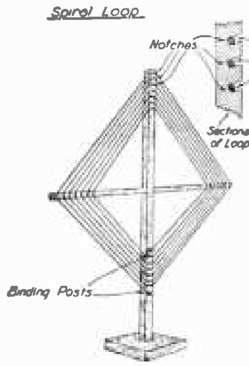
Spiral Loop
Notches
Wires
Sectional View of Loop Arms
Binding Posts

Fig. 8. The spiral type loop aerial

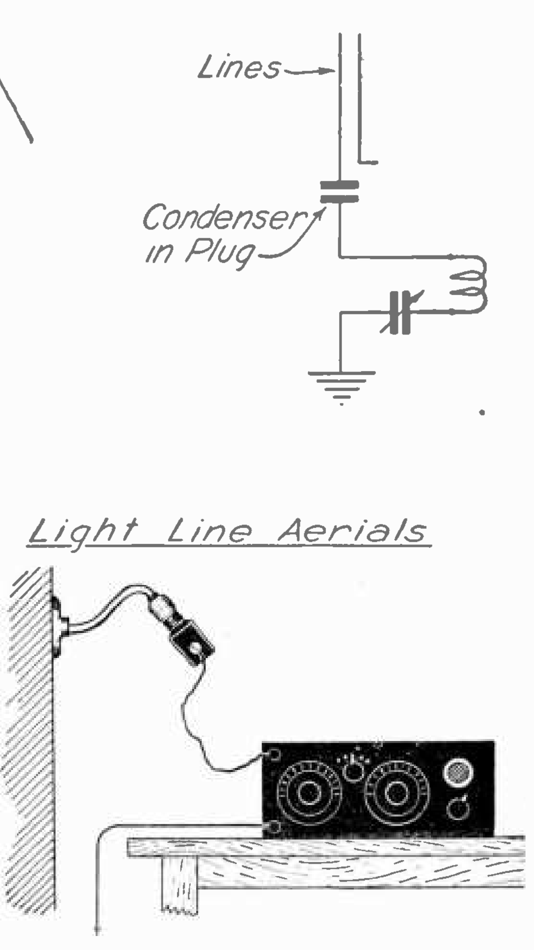
Lines
Condenser in Plug
Light Line Aerials

Fig. 9. Method of employing the electric light line as an aerial with an aerial adapter. Above is shown the diagram of the condenser in connection with the aerial circuit of a radio set

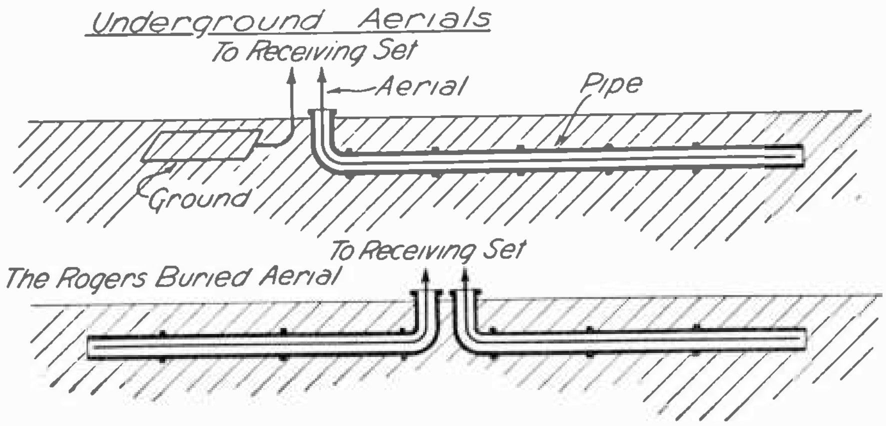

Fig. 10. Showing how wires are buried in the ground for aerials

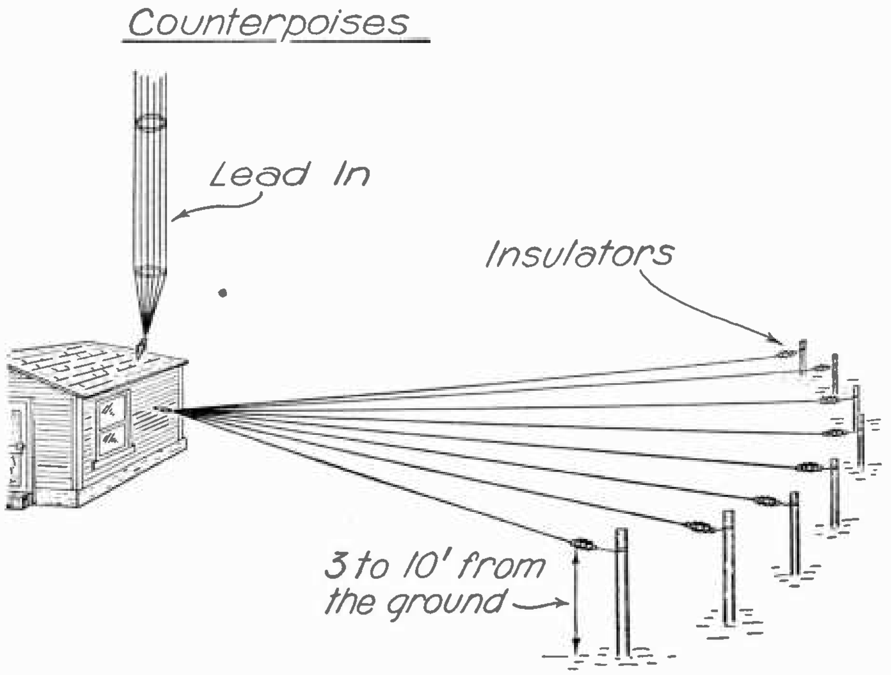

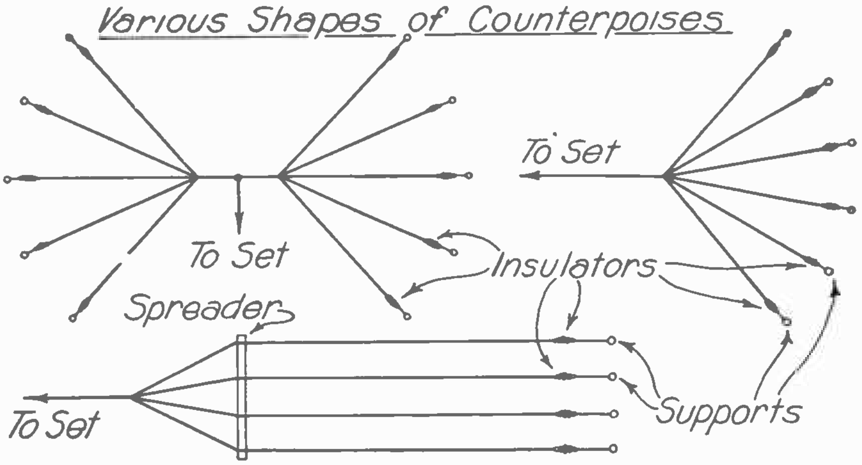

Fig. 11. Different methods of arranging counterpoises for either transmission or reception

AERIAL CAPACITY—(See *Capacity of Antenna.*)

AERIAL CIRCUIT—Consists of *aerial* and *earth* or *ground*, including all *coils* or *inductances*, *condensers*, etc., which may be connected with the aerial and earth and forming a direct path between those points.

AERIAL SWITCH—A device to transfer the *aerial* and *ground* connections from the transmitting circuit to the receiving circuit or vice versa. Its object is to separate the high tension current of the transmitter from the low tension of the receiver. The switch is usually installed in a convenient position so that the operator may change from transmitting to receiving by one motion of the hands. A simple form of aerial switch is illustrated in Fig. 1. The circuit for same is shown in Fig. 2. A more efficient circuit is the one shown in Fig. 3. This circuit requires another blade to control the primary current to the transformer. The latter circuit is most satisfactory as it entirely disconnects the current to the transformer, doing away with the possibility of dangerous shocks to the operator. The switch commonly used in such circuits is shown in Fig. 4.

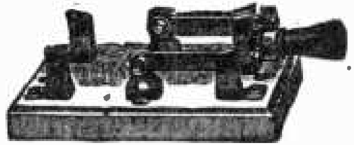

Fig. 1. A simple Aerial Switch

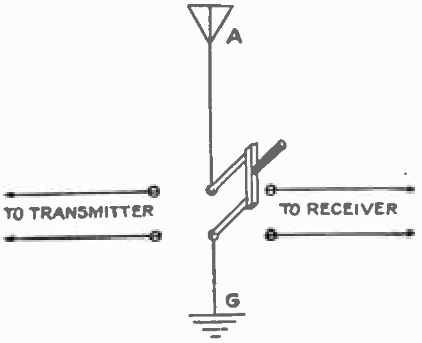

Fig. 2. Circuit for a simple Aerial Switch

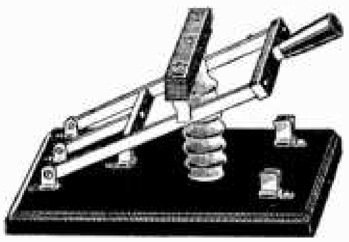

Fig. 4. An Aerial Switch with an extra blade

The switches used in the various commercial wireless stations are of course more complicated. There is a need to control many circuits with the simplest possible method. In order to do this it was found more convenient to only change the aerial connections, and allow the ground connection to both the receiver and transmitter to remain permanently. This of course necessitates a small air gap in the aerial for the transmitter. This is called an *anchor gap* (*q.v.*). This small appliance allows the transmitter to be connected continuously to the aerial as the high voltage from the transformer discharges across the small air gap, as provided for in the *anchor gap* (*q.v.*). Fig. 5 shows the circuit when this type of aerial transfer is to be used. When the switch handle is pushed down it makes connections with the receiver, when it is raised, the transmitter is ready to be used as the primary current is connected also.

AERIAL TUNING CONDENSER—A *condenser* (*q.v.*), usually variable, connected in the *aerial circuit* (*q.v.*) for the purpose of adjusting the natural period of the receiving circuit to the period of the incoming waves (see *Wavelength*) or signals coming from the *aerial*. In the illustration is shown an Aerial Tuning Condenser.

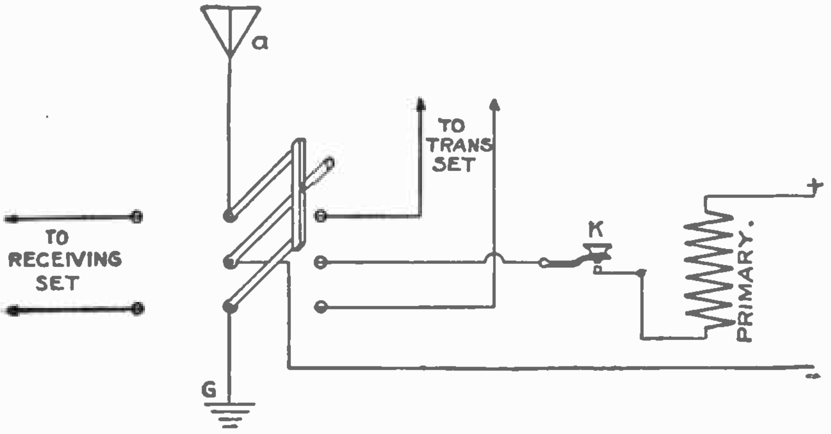

Fig. 3. Circuit for the Aerial Switch shown in Fig. 4 on page 7

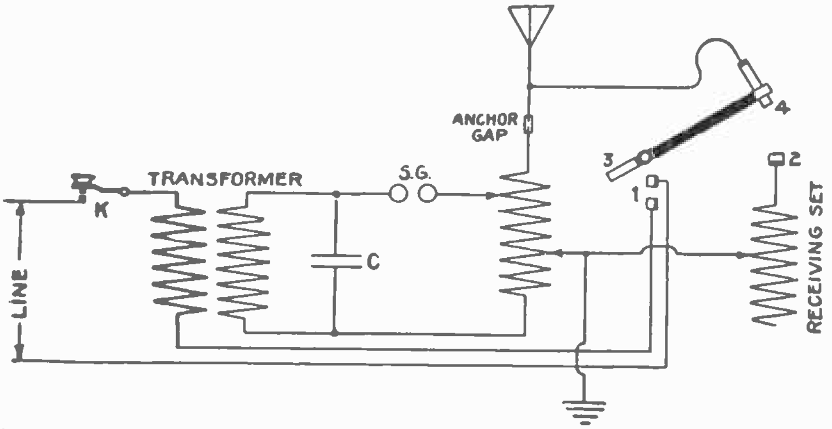

Fig. 5. An Aerial Switch employed in conjunction with an Anchor Gap as explained on page 7

Any *capacitive* (*q.v.*) means of establishing or varying the *oscillation constant* (*q.v.*) of the receiver.

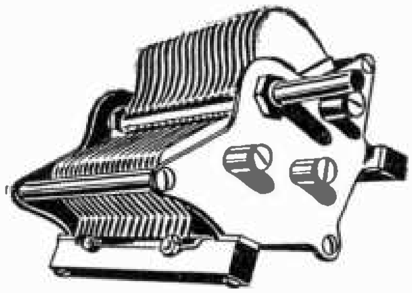

A variable condenser such as commonly used for an Aerial Tuning Condenser (see page 7).

AERIAL WIRE—The wire forming the *aerial* (*q.v.*). Copper is generally used for this purpose as it offers very little *resistance* (*q.v.*) to the feeble currents in receiving.

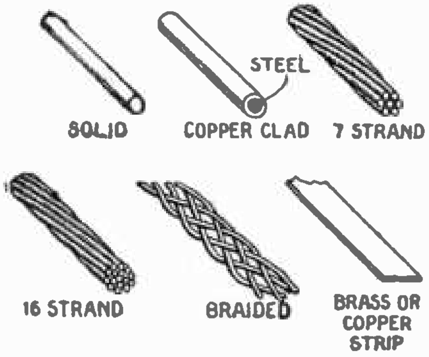

The illustrations show various types of wire as used for aerials. Among them there are the solid copper wire, copper clad steel wire, stranded wire, braided wire and brass or copper strip.

AERO - FERRIC INDUCTANCE—The *inductance* (*q.v.*) of coils wherein the *magnetic circuit* (*q.v.*) is completed through both air and iron.

AEROPLANE ANTENNA—Receiving or transmitting *antenna* (*q.v.*) used on an aeroplane. The commonest form is a suspended wire shown as A in the illustration. The *ground* or *counterpoise* (*q.v.*) in this case may be another wire suspended on and insulated from the fuselage as indicated at B. When the plane is in flight, the hanging wire trails up and backward, depending on the speed of the plane and the size of the weight used. This type of antenna will be slightly *directional* (*q.v.*) in the direction the plane is traveling. If it is impractical to use a trailing wire antenna for any reason, a loop or coil of wire may be used, the results of course being less satisfactory. The metal frame of the plane is often used as a *counterpoise* (*q.v.*)

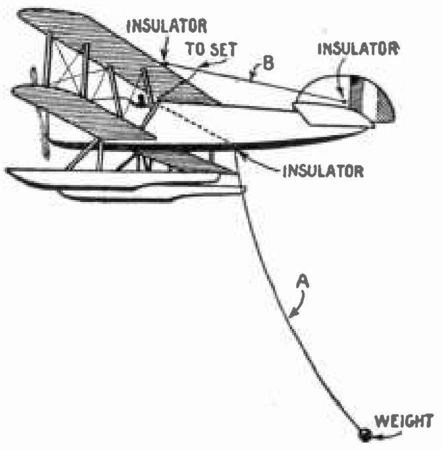

A. F.—Abbreviation for *Audio-Frequency* (*q.v.*)

AFTER-GLOW—Fluorescent phenomena in a *vacuum tube* after current has been withdrawn.

AGONIC LINES—Lines imagined as passing through points on the earth's surface where the magnetic inclination or dip of a magnetic needle (compass) is zero. (See *Aclinic Lines.*)

AGING, VACUUM TUBE—Gradual diminishing of the brilliancy of a tube due to the deterioration of the *filament* (*q.v.*) and the coating or deposit on the bulb. A tube is said to be Aging when it gradually loses its power to emit *electrons* (*q.v.*)

AIR CONDENSER—*Condenser* utilizing air as the *dielectric.* The majority of variable condensers used in radio work employ the air dielectric, that is, the space between plates is merely an air gap. Most *fixed condensers* (*q.v.*) use some substance such as mica, or sheets of waxed paper as a dielectric.

AIR CORE TRANSFORMER—A *transformer* used in radio, having no metal core, *the lines of force* having their path through air. (See *Amplifier, Radio Frequency.*)

ALBIZ, Count—Managing Director of the Compañia National de Telegrafia sin Hiols in 1910, of Scotch origin.

He was born in Madrid in 1858 and received a schooling at the Madrid University and London University College. He was a Tory member of Parliament.

ALEXANDERSON, Ernst Fredrick Werner (1878)—Radio engineer and inventor. Born at Upsala, Sweden, January 25th, 1878, he was educated at the High School and University of Lund, Sweden, at the Royal Institute of Technology, Stockholm, and at Berlin. In 1902 he joined the General Electric Company, and has been for some years their consulting engineer. He holds the post of chief engineer to the Radio Corporation of America, and

Photo by courtesy of General Elec. Co.

E. F. W. Alexanderson

is a member of the American Institute of Radio Engineers. Alexanderson has read many papers on electrical subjects before the chief technical societies of America.

Alexanderson is famous for his work on *high-frequency alternators* used in Radio telegraphy. The Alexanderson alternator is connected directly or *inductively* to the aerial and earth, and constitutes the simplest possible connection for producing *continuous* or *undamped waves.* It is a machine of great speed and many *field poles,* and frequencies as high as 200,000 *cycles* are obtained. For long distance work, employing very *long wavelengths,* this high-frequency alternator is largely used. Alexanderson is also responsible for the *magnetic amplifier,* patented in 1913, and has carried out successful experiments on *duplex wireless telephony.* In the Alexanderson *microphone transmitter* the *modulation* is mainly effected by variations in the tuning of the *aerial circuit.*

ALEXANDERSON ALTERNATOR—A type of *high-frequency alternator* (*q.v.*) which produces *continuous oscillations* (*q.v.*)

ALLOY—A compound of two or more metals. (See *Woods Metal.*)

ALTERNATING CURRENT—An electric current that does not flow steadily in one direction—as from *positive* to *negative*—but completely reverses its direction or *polarity* (*q.v.*) at certain definite intervals. In other words, the current flows first in one direction and then reverses and flows in the opposite direction—these changes being referred to as changes of *phase* (*q.v.*). An electric current is either *direct* (*q.v.*), alternating, or *pulsating.* The latter is an alternating current that has been rectified. (See *Rectified Current.*) Practically the only direct current that is employed or appears in a radio receiver is that supplied by the *"A"* and *"B" Batteries* (*q.v.*). In an alternating current, the change in direction or polarity takes place in a steady, even manner.

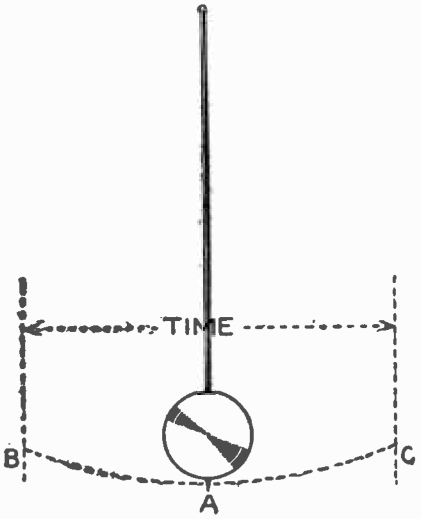

Fig. 1

An understandable analogy for alternating current is its comparison with a pendulum. In Fig. 1 the pendulum is supposed to be the alternating current or impulse. Now when the bob is swung from its natural (neutral) position A, which may be considered as zero value of voltage or current, it will move to a maximum point, either B or C. If we assume B to be the *positive* side, then the bob will move to B which will be the maximum positive *amplitude* (*q.v.*). It will then fall back to zero again, and with the power behind it, the momentum in this case, will rise on the other side to C, which is the maximum *negative* amplitude. The bob will then fall back to A once more and will have undergone a complete change of direction and values; it will have gone through a *cycle* (*q.v.*). That is to say, it will have risen from zero A to positive maximum B, then back to A, rising on the other side to negative maximum C and back to A. Now the length of time required for this complete cycle (see *Alternation*), is known as the *time period* of vibration or *phase change.* If, for example, it required one-sixtieth of a second for this change to take place, then we say that the *frequency* (*q.v.*) is sixty cycles, because obviously, a change requiring one-sixtieth of a second will occur sixty times in one second.

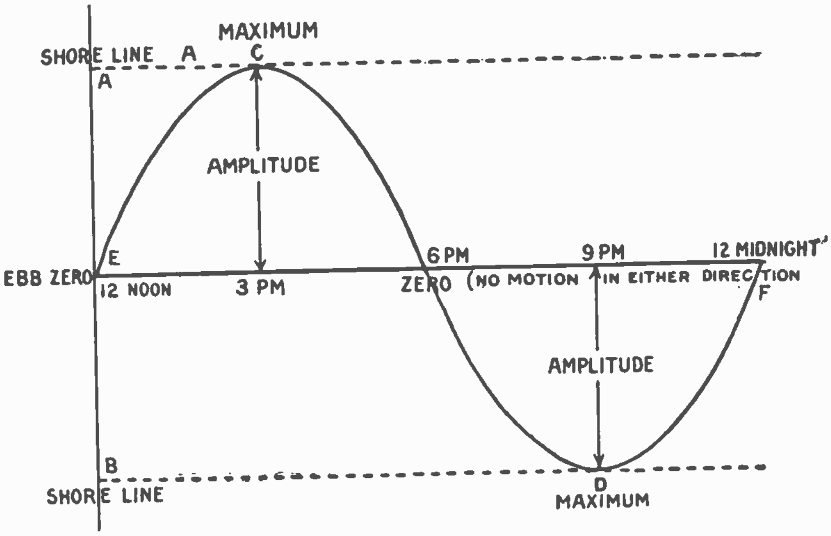

Fig. 2

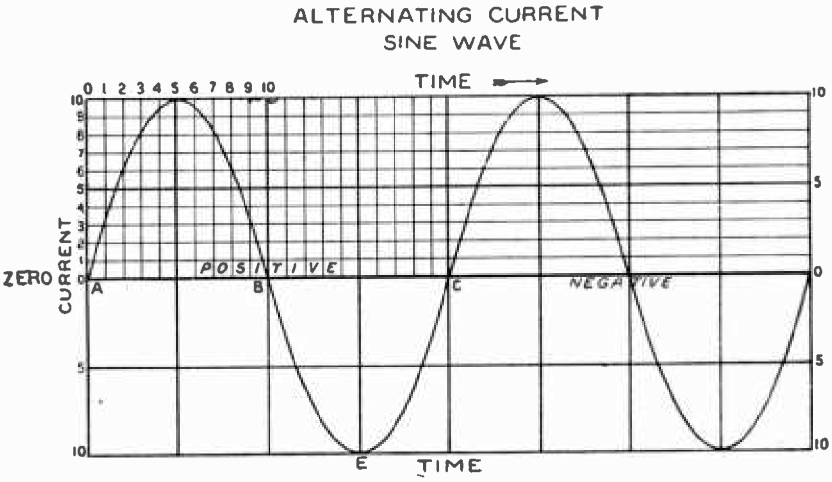

Fig. 3

The action of alternating currents may also be likened to that of ocean tides. In Fig. 2 the curved line represents the water in motion and is to be compared with current in motion in an alternating circuit. For the sake of clarity, we can assume that the tide rises and falls between two shore lines, A and B. At 12 noon the tide is at zero or "ebb" and there is no motion in either direction. Now the flow starts toward shore line A and gradually rises toward that side until at 3 P. M. it has reached its maximum height. In the case of an alternating current, we may consider that at 3 P. M. the current has reached maximum value on the positive side or in a positive direction. It then reverses and falls toward zero until at 6 P. M. it has reached the zero point. In the case of the current, it will have become zero. Motion again starts in the other direction and the tide rises once more, but in the opposite direction, toward shore line B. By 9 P. M. it will have reached maximum height in this direction, or in the case of the current it has reached maximum value or *amplitude* (*q.v.*) at D on the negative side. It then falls back again to zero, and at 12 midnight has come back to its original starting point, having gone through all stages of value or height from zero to maximum at one shore and then back to zero; then from there to maximum at the other shore and back to zero. The action of the alternating current then will have been to flow from zero to positive maximum, back to zero and reversing, from zero to maximum on the negative side, then back to zero. It will have completed one cycle. The time required for the tide to go through these changes E to F is 12 hours and is known as the *time period* or *period* (*q.v.*). Therefore, the time required for a complete change in the case of alternating current will be referred to as the period. The number of times that these changes take place within a certain time limit is known as the *frequency* (*q.v.*)

The current value in Fig. 3 is assumed to be 10 amperes, and the current values at any moment are shown as *ordinates* (*q.v.*), (vertical) while the time is shown in the *Abscissas* (*q.v.*) (horizontals). If the above curve is to be assumed as representing the Sine wave of an Alternating current having a frequency of 60 cycles per second, the total time in the chart would be one-thirtieth of a second. Thus, Time A to C would be one-sixtieth second, the Time of one complete cycle. A B is a positive Alternation and B C a negative Alternation, D and E represent maximum current values in each direction.

If the above Sine Wave is to be used as an example of a high frequency oscillation as in radio, the Time of one cycle can be assumed as about $\frac{1}{1,000,000}$ second. (See *Frequency-Oscillation - Sine - Phase - Characteristic of Alternating Current.*)

The frequency of an alternating current depends on the number of poles (field magnets) and the speed of rotation when the A.C. is generated by an alternator. (See *Alternator.*)

ALTERNATING CURRENT, THEORY OF PRODUCTION OF—Electric current in motion along a conductor always has a magnetic field (*magnetic flux*) surrounding it. If this current is passed near a number of turns of wire, the lines of flux will link with the wire turns in the circuit. The number of flux lines through a coil will depend upon the current. A change of current will change the number of linkages. If two turns of wire are used, the circuit will link twice with the same magnetic flux. Thus, a change in the current or a change in the number of turns of wire will alter the linkages.

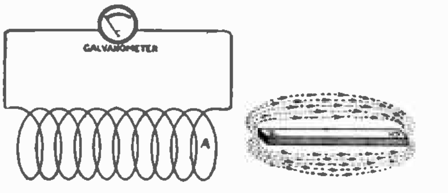

Fig. 1

Now in the case of Fig. 1, the magnet being placed near the coil causes lines of force to pass through the coil. Each turn of wire will link the flux lines. Now any *change* in these linkages between the flux lines and wire turns will produce an E.M.F. (*electro motive force*) in the circuit of the wire turns. This is called an "induced current" because it is induced by the change in linkages. In Fig. 1, we have a coil of wire in a circuit with some sensitive measuring instrument for minute currents, in this case a *galvanometer*. Near it we have a bar magnet. When this magnet is moved toward or within the coil, a current is induced in the circuit by reason of the change in linkages. This current is *always* in a direction to *oppose* the change that causes it. (See *Lenz Law.*) "Whenever an induced current arises by reason of some change in the linkages, the magnetic field about the induced current is in such direction as to oppose the change." When the magnet is moved toward the coil, the change in linkages creates or induces a current that flows in the circuit in such manner as to make A act as a North pole. As the North pole of the magnet is toward this point A, the similar poles will oppose each other according to the simple law of magnetic attraction and repulsion. Now if we attempt to withdraw the magnet, the change will alter the linkages again, this time creating an induced current that makes A act as a South pole, thus *attracting* the magnetic North pole and opposing in the opposite manner, the change in linkages that have created the current.

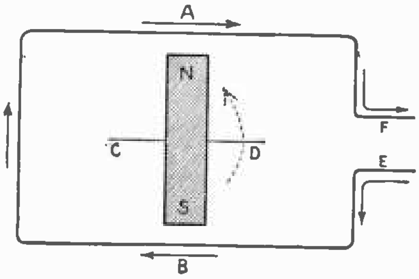

Fig. 2

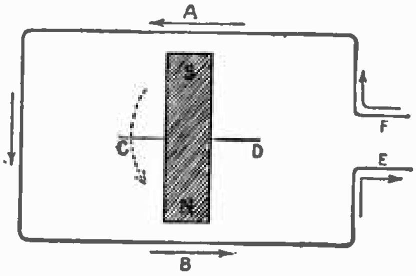

Fig. 3

We already know that if a conductor is moved across a magnetic field, an E.M.F. is induced in the conductor. The action is the same whether the conductor is moved across the magnetic field or vice versa. In Fig. 2, a magnet is shown in position on an axis—C D for rotation within the loop A B. As the magnet is rotated, the North pole passes across A, an E.M.F. is created toward F in the same direction. As the position of the magnet reverses in rotation and S passes A while N passes B, a current is induced in the *opposite* direction, both times in accordance with the law previously stated. Thus, if a meter is connected between the ends of the loop at E F and the magnet rotated, a current will be registered. With each half turn of the magnet, this current will reverse its direction. Therefore, we have produced an alternating current that changes direction with each reversal of the position of the magnets. This is the simple *AC generator* or *alternator*. The illustration, Fig. 4, shows a standard type of low frequency alternator.

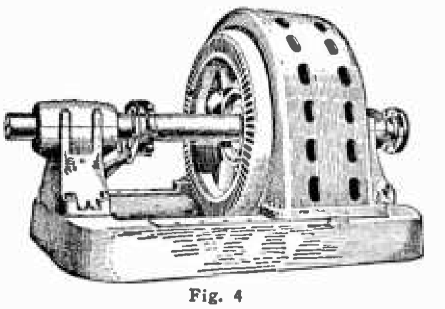

Fig. 4

ALTERNATION—In an alternating current, is the rise from zero to maximum *amplitude* and back to zero on either side, i.e., *positive* or *negative* directions. Two alternations—one positive and one negative—make a complete *cycle*. (See *Alternating Current.*)

ALTERNATOR—A machine for the purpose of changing mechanical energy into electrical energy. Fig. 1 shows an alternator in its elementary or simplest form—merely a conductor, B C moved across a *magnetic flux* (*q.v.*) (Note: For a complete explanation of the production of alternating currents, see *"Alternating Current, Theory of Production of."*) Here the conductor B C is moved to the right *across* the magnetic flux shown as dotted lines between N and S, the poles of the magnets. As the conductor moves across the flux lines of the magnet, an *electro-motive force* (*q.v.*) *or electric* current is set up or induced in the conductor, this force being in such a direction that the current produced will create a flux or magnetic field, in the same direction as the flux on the side of the conductor that comes first into contact with the flux lines N S on moving the conductor. In this case E.M.F. (electromotive force) is in the direction B to C in the conductor. In order to establish a circuit and utilize the current pro-

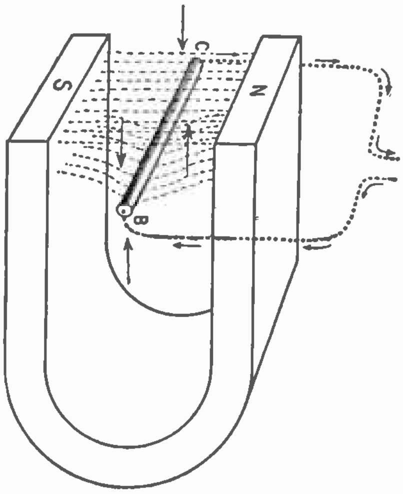

Fig. 1

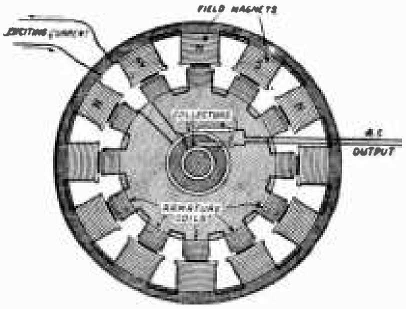

Fig. 2

duced in this manner, it is only necessary to attach wires to both ends of the conductor as shown by the dotted loop. Now at this moment the conductor is being traversed by a current having a definite polarity, that is to say, it is flowing in one direction only. When the direction of the flux lines N S are reversed, as by reversal of the magnet, or the conductor is turned so that the other end comes into contact with the flux lines, then the direction of the current in the conductor will reverse and it will flow in the opposite direction. In other words, with every reversal of the position of the conductor in its relation to the magnetic flux N S, or the reversal of the direction of the flux lines, with respect to the conductor, the E.M.F. that is created will change its direction of flow. (See *Alternating Current.*)

In an alternator it has been shown that the result is approximately the same whether the magnet is rotated or the conductor is rotated through the magnetic flux. In either case the moving member is known as the *rotor* (*q.v.*) and the stationary member is called the *stator* (*q.v.*). The magnets producing the flux lines may be stationary or they may revolve; in either case they are referred to as the *"field"* (*q.v.*).

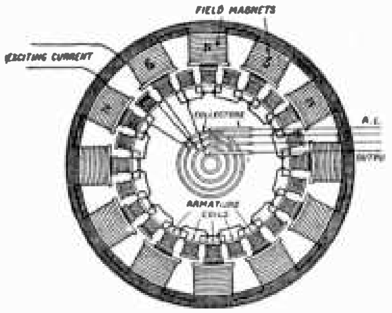

Fig. 3

It has been shown that with every change in the relative position of the field magnets and the *armature* (*q.v.*), irrespective of which member is the rotating one, a change of direction of the induced E.M.F. takes place. Then it is apparent that in order to increase the number of changes, that is, to increase the *frequency* (*q.v.*), it is only necessary to use (1) a greater number of field magnets, or (2) have the armature or conductor pass the magnets more frequently by means of greater speed in the rotating member, or by means of both. Now each of these changes is an alternation (see *Alternating current*) and two such changes will be a *cycle* (*q.v.*). Let us suppose that the moving member of a certain alternator has a rotary speed of 720 revolutions per minute and there are 10 poles or magnets. As frequency is figured by the changes per second, it will be necessary to reduce this number to the revolutions per second. Thus, there will be 12 revolutions per second. This figure multiplied by the number of *poles* (*q.v.*), which we assume as 10, indicates 120 alternations per second. As there are two alternations to each *cycle* (*q.v.*), the frequency in this case will be 60 cycles. The equation is as follows:

$$F = \frac{P}{2} \times \frac{N}{60}$$

where F is the frequency, P, the number of poles and N the speed of rotation per minute. The exact voltage or E.M.F. obtained will depend on the number of poles and the speed of rotation, and also on the distribution of the field windings and total flux from all the poles or magnets. In designing alternators the effective E.M.F. (electromotive force) is expressed by the formula:

$$E = \frac{K P \phi N Z}{10^8}$$

E is the effective electromotive force of the alternator.

K represents the E.M.F. factor of the alternator. This factor depends upon certain charcteristics of the poles and also the distribution of the windings on the armature.

P is the number of poles of the field magnates.

ϕ is the number of lines of magnetic flux that flow from one pole across the gap to the armature—known as the useful magnetic flux per pole; (see *Maxwells*).

N is the speed of rotation of the rotor in revolutions per second.

Z is the total number of conductors on the surface of the armature.

10^8 represents the number 100,000,000—the calculation given would be determined in *abvolts* (*q.v.*), $\frac{1}{100,000,000}$ volt and it is therefore necessary to reduce it to volts by division.

In the "revolving field" alternator, as explained above, the field magnets are rotated and the conductor remains stationary. The windings in which the E.M.F. is produced, are placed in slots in the inner surface of the stator, the stator and the windings composing it being the *armature.*

In the alternators of the revolving armature type, the conductors or armatures are the rotating member and the field magnets remain stationary. Fig. 2 shows the revolving armature type of alternator in diagrammatic form. A small direct current *generator* (*q.v.*) may be used to excite or build up the field. That is to say, a current is sent through the field magnets (windings in this case) for the purpose of increasing or building up the *flux density* (*q.v.*). This is known as an *"exciter"* (*q.v.*). The exciter may be mounted directly on the end of the rotating shaft of the alternator, or mounted separately and connected to the alternator rotor shaft by means of a belt. This permits a heavy field to be built up by the electro-magnets used in place of the ordinary magnet shown in Fig. 1. If the machine is to be *single phase* (*q.v.*) there will be two *collector rings* (*q.v.*), the two phase type will have four rings, and the three phase six rings. The illustration Fig. 2 shows the general design of a single phase alternator. In this case there will be of course only two rings to collect the E.M.F. set up by the machine. Fig. 3 illustrates the general scheme of armature winding for a two phase machine. Here there are actually two separate E.M.F.'s set up by the separate windings and the ends of each of these windings are connected to collector rings as shown. In the three phase arrangement there may be three separate windings and six collector rings.

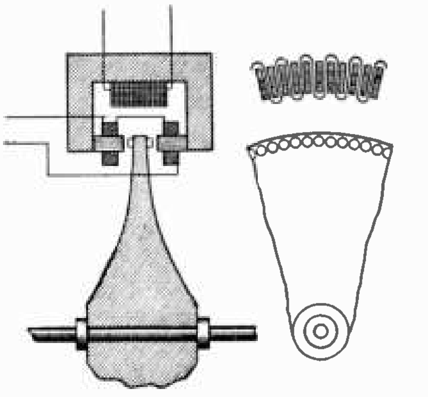

Fig. 4

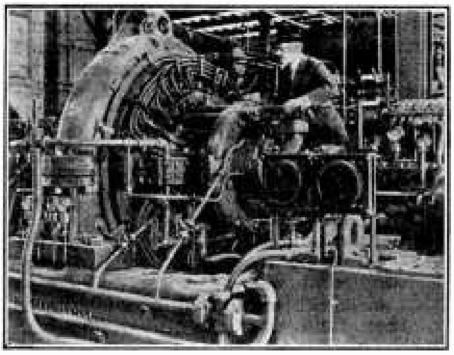

Fig. 5

The use of the revolving armature type for comparatively small voltages is due, as mentioned previously, to the fact that they are easily constructed, and also the advantage that the exciter need not be an external source, but may be a separate winding with a *commutator* (*q.v.*) mounted on the main shaft. The revolving field type of course allows much greater space for the armature winding, inasmuch as the stationary member is not limited to any particular space, whereas the moving member must be arranged as small as possible to do away with mechanical troubles. Thus, the stationary armature can be wound and properly insulated for great voltages. For this reason the revolving field is generally used in high power work. Another point of importance is that where the armature revolves, the windings must be made very carefully as the centrifugal force due to rapid rotation of the members places a great strain on them.

One of the widely used forms of alternator for extremely high frequencies, is the Alexanderson, named after its originator, *Ernst Alexanderson* (*q.v.*), a noted engineer. This machine is of the inductor type. It is a well-known fact that magnetism is more readily established through iron than through air, and therefore when iron is placed in the air gap of a magnetic circuit, the lines of force or flux lines are greatly increased. Then when the iron is withdrawn, the flux lines fall to their original value. We have seen that as changes of flux take place throughout a magnetic circuit an electromotive force is induced in a coil or conductor (armature) surrounding the magnetic circuit. Now, if iron is periodically inserted into and withdrawn from the magnetic field, an alternating current will be set up in the coil surrounding the field. This principle is made use of in the inductor alternator and the field and armature thus remain stationary, contrary to the usual form of an alternator. Now if we arrange to have pieces of iron moved rapidly in the gap between field and armature, changes of flux will take place. Obviously, it is easy in this manner to create many changes of flux within a certain time, and therefore the alternating current set up (E.M.F.) will have a high frequency as seen by the explanation of this phenomena given previously. These moving pieces are so arranged that they practically close the gap between field and armature with only enough clearance to prevent the rotating member striking the stationary members. This is briefly the principle of the Alexanderson alternator. Fig. 4 shows the general plan of an inductor alternator. On the right shows the scheme of winding in slots around the entire form, and as there are necessarily many windings in order to obtain the high frequency, they are generally single turns of wire, each turn having a separate slot. On the left is given a general idea of the arrangement of the rotor, which in this case, as explained previously, is neither field nor armature, but iron teeth arranged on the periphery of the moving disk. The field is increased or excited in the usual manner by means of a source of direct current. In the assembly, the relation of the moving disk to the field and armature is shown. Fig. 5 is a commercial type of Alexanderson Alternator having a frequency (*q.v.*) of 100,000 cycles used for high power radio transmission (*q.v.*).

ALTERNATOR, SINGLE PHASE—(See *Alternator.*)

ALTERNATOR, POLYPHASE—(See *Alternator.*)

AMMETER OR AMPERE METER—An instrument for measuring current in amperes in a circuit. Is connected in series with a circuit. Exists in a variety of forms. The operation of the most commonly used type depends upon the fact that the force a magnet exerts depends upon the number of *ampere turns* (*q.v.*). Therefore, the greater the current sent through its coils, the greater will be its attraction for a balanced armature. Fig. 1 shows a common form of ammeter.

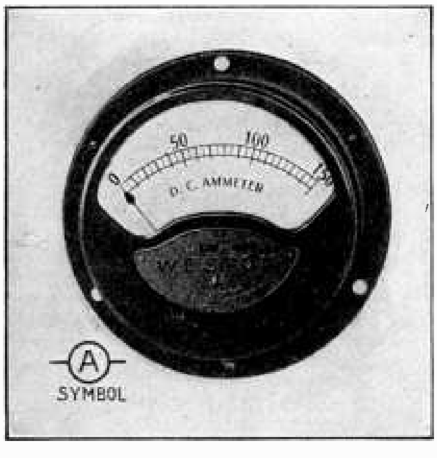

Photo by courtesy of Weston Elec. Inst. Co.

Fig. 1

AMMETER, THERMAL — Commonly known as "*Hotwire Meter.*" Also called "Aerial Ammeter" and "Radio Frequency Ammeter." Operates on the principle of thermal expansion, i.e., the tendency of a wire to expand with heat of current passing through it. Is used to measure current in amperes at extremely high voltages as in radio transmission. Figure 2 shows a Thermal Ammeter which is generally employed in radio work.

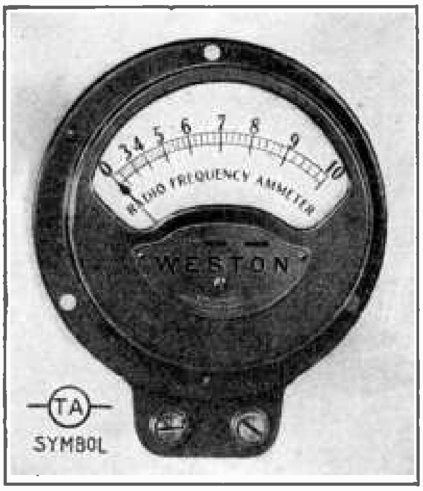

Photo by courtesy of Weston Elec. Inst. Co.

Fig. 2

AMP.—Abbreviation for *Ampere.*

AMPERE—The practical unit of the rate of flow of electricity. It is the unit current, or the rate of flow when one *coulomb* (*q.v.*) flows each second. The value or strength of electric current is almost invariably given in amperes. This unit is used to express the flow of a definite quantity of electricity in a second. If, for example, ten coulombs pass a given point in a circuit in one second, then we say that the current in that circuit is ten amperes. (See *Coulomb, Milli-ampere, Micro-ampere.*)

AMPERE, Andre Marie—A distinguished French physicist, born 1775, died 1836. Noted for his researches in the field of *electro-dynamics.*

AMPERE-HOUR—Commercial unit of quantity. Is that quantity which flows in one hour through a circuit carrying a current of one *ampere.* It is equal to 3,600 coulombs.

AMPERE, INTERNATIONAL—Symbol a—is defined as the unvarying electric current which deposits silver at the rate of 0.00111800 gram per second from a specified solution of nitrate of silver in water.

AMPERE'S RULE FOR DEFLECTION OF NEEDLE—"If one swims with the current and looks at the plus (+) or *north seeking pole* of a magnetic needle it will be deflected to the left, while the negative (—) or *south seeking pole* will be urged to the right."

AMPERE'S THEORY OF MAGNETISM—The theory advanced by *Ampere* that an *electric current continually circulates around each molecule of a magnetic substance and that the process of magnetization is one of arranging these currents in such manner that they all take the same direction.*

AMPERE TURNS—Expressed by the product of number of turns of, and the number of amperes flowing through, the coils of an electro-magnet. Thus, one ampere turn would be one ampere flowing through one turn. The ampere turn is very frequently used as a unit of *magneto-motive force* (*q.v.*) due to the ease with which it can be obtained.

AMPERE VOLT OR VOLT AMPERE—The expression sometimes used to denote *watts* (*q.v.*). For example 1 volt ampere will be one watt, as watts are equal to the product of the volts and amperes in D.C. circuits. (See *Efficiency,* also *Watts.*)

AMPLIFICATION CONSTANT — The factor expressing the maximum voltage *amplification* (*q.v.*) that it is possible to obtain with a given vacuum tube. Thus, if a certain tube will permit amplification of eight times the original voltage, the amplification constant is said to be 8. (See *Vacuum Tube.*)

AMPLIFICATION FACTOR—The ratio of the change of instantaneous voltage, between filament and plate of a *vacuum tube* to a small change of instantaneous voltage between filament and grid for a given constant plate current:

$$y = \frac{\delta V_p}{\delta V_g}$$

The ratio of power (radio signals), voltage or current output of an amplifying device to the power, voltage or current delivered to the input terminals. Generally speaking the degree of increase in amplitude or volume by insertion of an amplifying device in the circuit. (See *Amplification Voltage.*)

AMPLIFICATION, VOLTAGE — In a *vacuum tube* the ratio of the voltage apparent between the terminals of an impedance in series with the *plate,* to the voltage impressed on the *grid.* (See *Amplifier.*)

AMPLIFIER—One of the most important features of the *vacuum tube* (*q.v.*) is its ability under certain conditions to amplify or increase the intensity of electric currents. Such tubes are generally referred to as amplifier tubes and an amplifier is understood to be one or more such tubes with the necessary associated circuits to accomplish amplification.

X—In the accompanying diagram a circuit of a typical one stage *audio frequency* (*q.v.*) amplifier is shown in conjunction with the *detector circuit* (*q.v.*). In this case the *radio frequency signals* (*q.v.*) are impressed at the points A-B from the aerial and ground circuit. The resulting signals at points C-D are audio frequency or voice frequency because they have undergone a change or *rectification* (*q.v.*). Now, if we consider only the solid lines, the circuit will be an ordinary detector circuit and the signals at C-D can be heard by merely connecting the phones at that point in place of the primary of the transformer—P-B. However, by sending these currents through the transformer and impressing them on the second tube it is possible to obtain signals at points E-F of much greater volume and yet having the same general characteristics or sound. The signals or currents will have undergone amplification—the effect of the amplified tube on a local source of power increase the amplitude, but the form or nature of the signals or waves will be the same. (This statement is not entirely true as amplification is generally accomplished only at the expense of some slight changes in the nature of the waves). They will, however, sound about the same in the phones at E-F as they will when the phones are placed at C-D, the volume at E-F being greater due to the amplification. (See *Distortion*). Now it must not be assumed that an amplifier in itself actually has the ability to increase power without some assistance. It is no more possible to obtain something for nothing in electricity than it is in mechanics. What actually takes place is that the amplifier tube and associated circuits control local power—in this case the voltage of "B" battery—in such a manner as to boost or amplify the power of the signals without materially changing their nature. Therefore, an amplifier may be considered as a device which modifies the effect of a local source of power in accordance with the variations of input power, and produces an increased output power or a means of stepping-up or increasing the amplitude or volume of a signal or series of signals. (See *Amplifier Action or Vacuum Tube.*) (See *Amplifier, Audio Frequency, Amplifier Radio Frequency,* also *Transformer.*)

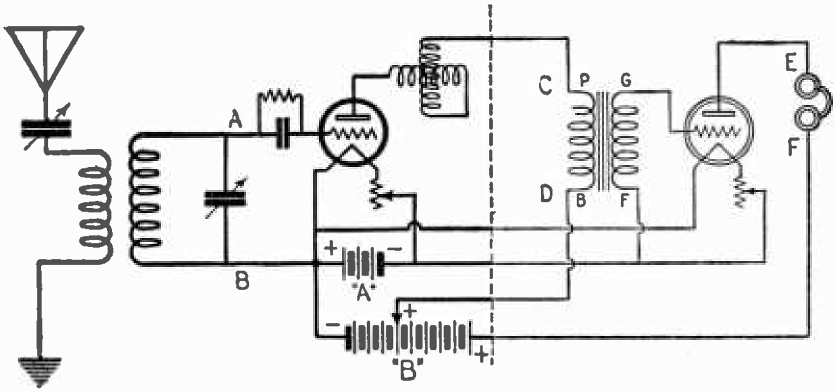

A single stage radio frequency amplifier as used in conjunction with a tuner and detector
Showing the method of amplifying rectified signals at audio frequency

AMPLIFIER ACTION OF VACUUM TUBE—The tendency of a vacuum tube to increase the power or voltage of incoming signals under proper conditions. In the illustration is shown a simple vacuum tube circuit. The incoming signals are impressed on the tube at points A and B. Now it has been noted that small changes of voltage at the *grid* (*q.v.*) will produce much greater changes at the *plate* (*q.v.*). (See *Vacuum Tube, Theory of Operation.*) In other words, the constant changes in value of the incoming signals that are impressed at A-B will produce comparatively larger changes in the current at the plate and therefore larger changes of voltage in the output circuit at C-D. These changes will of course manifest themselves in the form of signals in the phones placed at C-D. Now if these voltages are impressed on a *resistance* (*q.v.*) of suitable value—such as the primary of an amplifying transformer —they can be sent to another tube and again amplified. (See *Amplifier, Audio Frequency.*) In other words the amplifying action of a vacuum tube means the ability of that tube to use input power—the power of the signals delivered to the tube—to control a local source of power such as the "B" battery voltage used on the plate of the tube, and deliver increased power. The signals to be increased or amplified in this manner may be *Radio Frequency* (*q.v.*) or *Audio Frequency* (*q.v.*), each instance requiring a different type of *amplifier* (*q.v.*). The radio frequency signals are amplified by means of a tube and suitable apparatus (transformer or resistance) before they have been changed to audio frequency. The signals that have been *rectified* (*q.v.*) or changed to audio frequency may be amplified by means of a similar arrangement of somewhat different type. The latter would be known as audio frequencv amplification. The increase in power is accomplished without any material change in the nature of the signals or wave form (*q.v.*)

AMPLIFIER, INTERMEDIATE FREQUENCY—An arrangment of a *radio frequency amplifier* to increase the strength of signals at the *intermediate frequency* (*q.v.*) as used in a *Super-Heterodyne receiver.* This is usually a radio frequency amplifier using *transformer coupling* (*q.v.*) the transformers being designed to cover extremely high wave-lengths. (See *Super-Heterodyne,* also *Transformer.*)

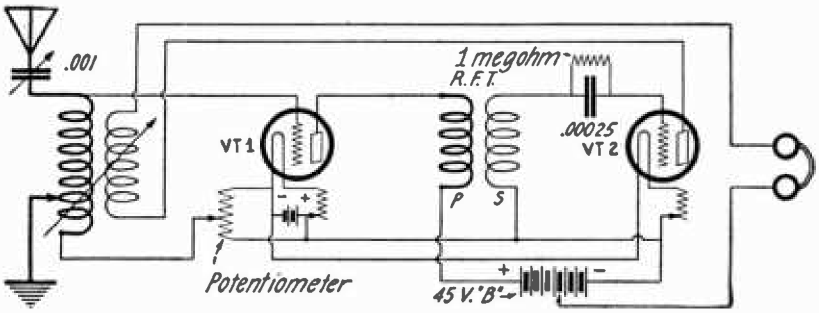

A single stage radio frequency amplifier as used in conjunction with a tuner and detector

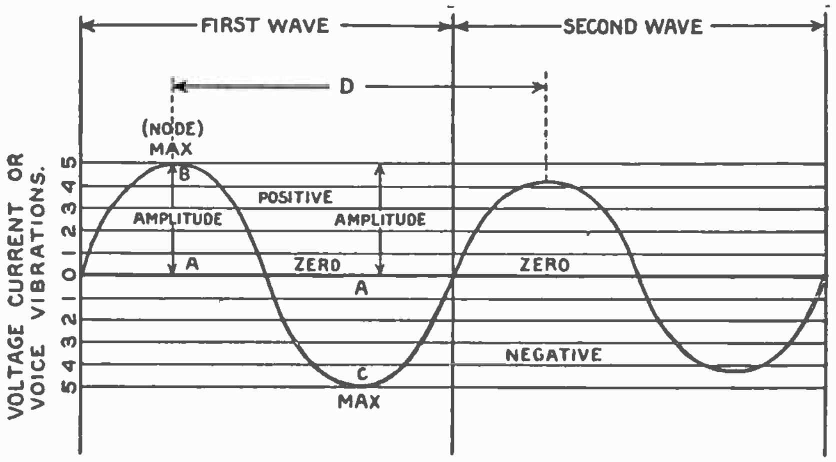

Line wave showing two cycles of an alternating current, indicating the amplitude and damping

AMPLIFIER, RADIO FREQUENCY — An amplifier used to increase the volume of the incoming signals at *radio frequency* (*q.v.*) before they have been *rectified* by the detector. In the illustration is shown a single stage of radio frequency amplification. This is the transformer coupled type. In this case the incoming signals of radio frequency are increased in strength and then passed on to the detector tube VT2. If further amplification is desired, more stages may be used, the number being held within certain limits due to *distortion* (*q.v.*). If more than three stages are used, it is often difficult to control the *feed-back effect* (q.v.) and efficiency is cut down.

AMPLITUDE—The current or voltage on an A.C. or H.F. circuit at any instant.

The greatest vertical distance above or below zero of the crest or trough of a wave is the maximum amplitude of the wave. In the accompanying illustration the numerals at the left (*ordinates q.v.*) may be assumed to be volts, current (amperes), or voice vibrations. A is the zero point; B the maximum *positive,* C the maximum *negative,* and D the length of the wave. Thus, A. B will be the maximum *positive* amplitude, A C the maximum *negative* amplitude, and D the wave length. B is known as the crest or Node—C is the trough or antinode.

X—Now if in a certain train or group of waves, the amplitude of one wave is less than the one immediately preceding it, there is said to be *damping* (*q.v.*), the degree of this damping or diminishing in amplitude being referred to as *decrement* (*q.v.*). Thus, if we assume that in Fig. 1 the current or amplitude of the first wave is 10 amperes, and the next wave has an amplitude of only 5 amperes, then the damping or decrement is 50 per cent, because the amplitude has decreased that much. Similarly, if the first wave amplitude were 10, and the second were 9 amperes, then the decrease

Super Regenerative and *Super Heterodyne.* The principle of regeneration is one that has revolutionized the art of radio reception and broadcasting. The vast majority of present day receivers use regeneration in one form or another. Basically, regeneration is the process of returning a certain portion of the unrectified or radio frequency (*q.v.*) current from the plate circuit of the detector tube to the grid circuit. This has the effect of reducing the resistance of the grid circuit and thus permitting more ready passage of the weak impulses. The result is not only to greatly increase the sensitivity of the detector to weak signals, but actually to step up or amplify the volume of the received signals. In Fig. 1, coils L1 and L2 are the aerial tuning units in conjunction with the variable condenser, C1; C2 and C3 are variable condensers of low capacity; L3 is the plate control for regenerative action; T is the telephone receiver, G, P and F are respectively the grid, plate and filament elements of the tube; R the filament rheostat, and A and B the batteries for filament and plate. In the illustration Fig. 2, C1 is the aerial tuning condenser, L1 the primary coil of a standard vario coupler; V1 the grid variometer and V2 the plate or regeneration variometer. L2 is the secondary of the vario coupler, and the balance of the elements are as in Fig. 1, GL being the customary grid leak and condenser. Regenerative receivers have come into some disfavor owing to their tendency to radiate annoying signals into the antenna, and also the howling within the set when not properly adjusted. While regeneration in some form or other is used in the majority of receivers, it is not widely used now in its original state. (See *Neutrodyne, Tuned Radio Frequency,* etc.)

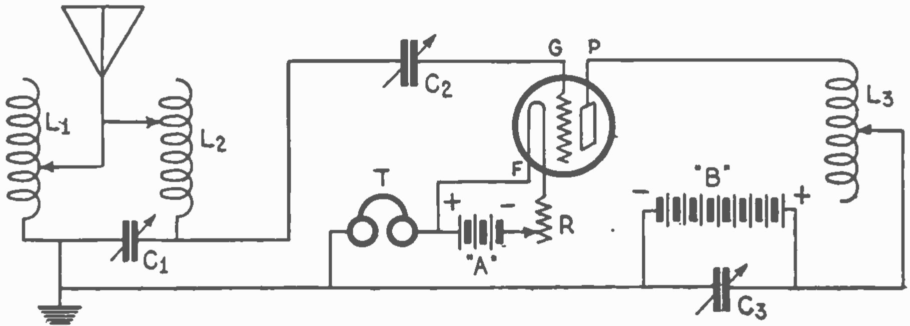

Fig. 1. The Original Armstrong Circuit.

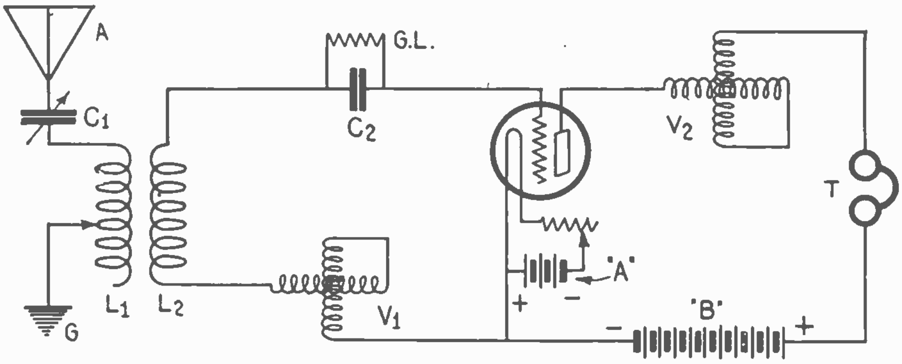

Fig. 2. The Armstrong 3-Circuit Regenerative Receiver.

There are many types of regenerative set, some extremely simple and others more complicated. The original Armstrong regenerative circuit is shown in Fig. 1. Fig. 2 shows a more modern application of the regenerative principle and at the same time a more complicated one. (See *Regeneration, Super Regeneration* and *Super Heterodyne.*)

ARRESTER—A device used for by-passing heavy electrical discharges in the atmosphere in the vicinity of the antenna when they strike the antenna. It ordinarily consists of a small gap between two metallic points connected in series with the antenna and ground in order to protect the apparatus and property against danger. (See *Lightning Arrester.*)

A. R. R. L.—Abbreviation for American Radio Relay League.

ARTIFICIAL ANTENNA — See *Mute Antenna.*

ARTIFICIAL MAGNET—A magnet produced by magnetizing a piece of steel that previously had no magnetic attraction. May be produced by rubbing a *natural magnet* or another artificial magnet across a piece of hard steel in the same direction a number of times. (See *Electro-magnet.*)

ASSUMED DIRECTION OF CURRENT FLOW—The direction which an electric current is assumed to take in its flow. Current is generally considered as leaving the *positive* terminal of its source, flowing through the circuits external to the source, and thence to the *negative* terminal of the source. The direction of flow within the source itself is from the negative to the positive terminal. (See *Theory of Current Flow.*)

ASTATIC COILS—Coils wound in such a manner that, when connected together, they neutralize each other's effect and produce no external *magnetic field.* They are used in the measurement of *inductance.*

ASYMMETRIC CONDUCTOR — One that permits the flow of greater current in one direction than in another. Non-symmetrical as to conductivity.

ASYMMETRICAL EFFECT—The lack of symmetry in the directional effect in a loop or frame aerial due to lack of symmetry in construction. (See *Loop Aerial.*)

ASYNCHRONOUS—A term used in referring to AC (Alternating Current) motors or generators and also to *dischargers* used in radio transmission. As applied to alternating current machines it is signified that the speed of rotation of the machine does not have any definite relation to the *frequency* of the currents produced—thus, out of *synchronism.* (See *Synchronous.*)

An asynchronous radio circuit is one which is not tuned to, or in syntony with, the frequency of the *oscillations* impressed on it. (See *Aperiodic.*)

ASYNCHRONOUS ROTARY DISCHARGE—A spark discharge produced at electrodes of a *rotary discharger* (*q. v.*) for transmitting radio signals, in which the spark rate or number of sparks per second has no relation to the *frequency* of the *alternating current.* (See *Spark Discharge.*)

ASYNCHRONOUS SPARK GAP—See *Spark Gap.*

ATMOSPHERIC ABSORPTION — The portion of the total reduction of radiated power due to atmospheric conductivity, reflection and refraction. The amount of energy loss in radio transmission due to atmospheric factors. This applies only to *electromagnetic waves* (*q.v.*) in *motion in the ether* and should not be confused with the reception losses due to atmospheric disturbances.

ATMOSPHERICS — See *X's, Strays, Static.*

ATTENUATION — of the electric or magnetic intensity or of the average energy density in electric waves refers to the reduction in strength with increase of distance traversed due to absorption or equivalent losses, as distinguished from the reduction in strength due to geometrical divergence. In free space, the geometrical divergence of waves from a small source involves a diminution of the average energy density in accordance with the inverse square law; plane waves (e. g., along wires) suffer no reduction by divergence, but undergo attenuation due to resistance and other line faults. (See *Attenuation, Geometrical Divergence*).

ATTENUATION, GEOMETRICAL DIVERGENCE—The gradual decrease in strength of electromagnetic waves (radio signals) as the distance which they have traversed increases. The effect on the waves as they travel through the air is essentially the same as in the case of light waves. Now if a beam of light is viewed one mile from its source it will have a certain intensity, but if it is viewed again from a distance of two miles the intensity will be much less. In the case of radio waves, the intensity is said to decrease as the square of the increased distance traveled; that is to say, if they have a certain power at a distance of five miles from the transmitter, the same waves at twice the distance or ten miles will be found to have a strength only one-quarter as great. Similarly, if the waves travel fifteen miles or three times the distance, the signal strength will be only one-ninth that of the same waves at a distance of five miles from the transmitter.

AUDIBILITY—The measure of strength of received radio signals as heard in a telephone receiver connected to the radio receiving apparatus. A radio telegraph signal which is at all audible can be read, i.e., the dots and dashes distinguished from one another; radio broadcast programs heard with at least some degree of clarity.

AUDIBILITY (Radio Telegraph) — A measure of the ratio of the telephone current producing a signal in a telephone receiver to that producing a barely audible signal. (A barely audible signal is one which just permits the differentiation of the dot and dash elements of the letters sent in code.) In the simple shunted telephone method of measuring signals the audibility is defined as:

$$\frac{s + t}{s}$$

where s and t are the *impedances* of shunt and telephone respectively.

AUDIBILITY FACTOR—A measure of signal strength obtained by observing the *resistance* to be placed in shunt with the telephones of a receiving station to reduce the signals to unit audibility. Approximately, the audibility factor is proportional to the power absorbed from the signal waves by the receiving apparatus. Unit audibility is the strength of signal at which dots and dashes can just be discriminated.

AUDIBILITY METER — A device for comparison of the strength of received signals, either from different stations or with different circuits by means of a variable *resistance* shunted (placed) across the phones. This resistance is decreased until the signals are just audible in the phones. Now, if another station is tuned in, or the circuit changed in some manner, the signals may either be entirely inaudible or they may be much louder, this giving an idea of the *comparative* strength without actual measurement. A popular type of Audibility Meter is shown in the above illustration.

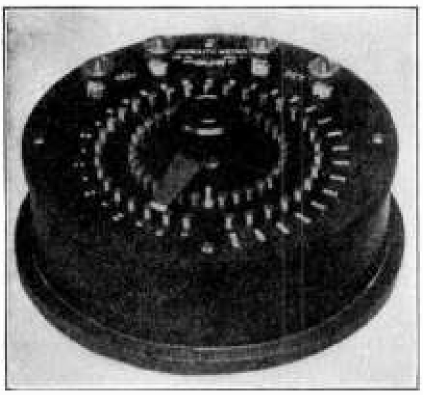

Photo by courtesy of General Radio Co.

An Audibility Meter.

For measurements with an audibility meter, the following shunt formula is used:

$$K = \frac{s + t}{s}$$

Where s is the *impedance* of the *shunt* resistance and t the impedance of the phones, then K is the audibility constant. In general practice a *series* resistance is devised to compensate for the reduction in resistance of the shunt circuit, this keeping the impedance of the plate circuit constant.

AUDIO AMPLIFIER—An *amplifier* used to increase the volume of signals of *audio frequency* (*q. v.*). Audio amplifiers are connected in the circuit of the detector in place of the phones (input) as shown in Fig. 1. This is the standard two stage audio amplifier. The separate stages each consist of a *vacuum tube, a transformer, socket* for the tube and a control *rheostat*. If desired one rheostat may be used for both amplifier tubes as the *filament* current (*q. v.*) can be the same if the tubes are of a similar design. This is known as "transformer coupled amplification." Another much used method is known as "resistance coupled amplification."

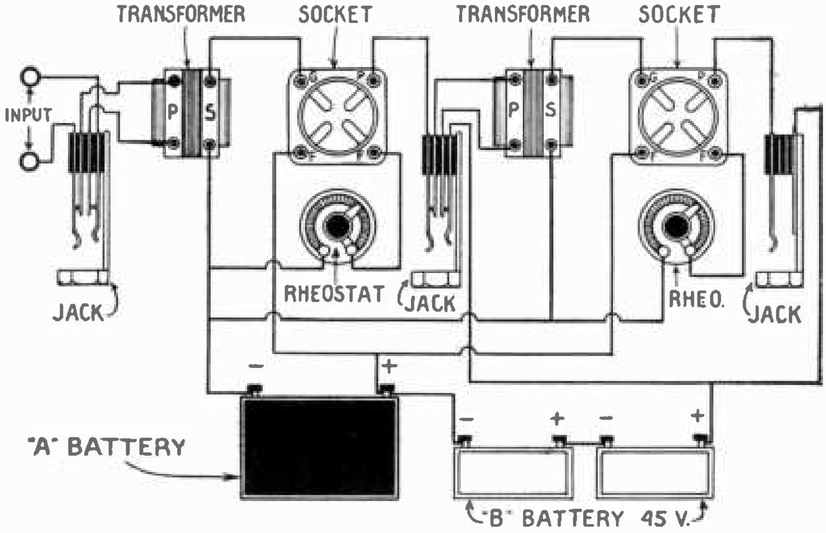

Fig. 1. A standard two stage Audio Amplifier employing phone jacks for detector, first stage, and second stage Audio Amplifiers.

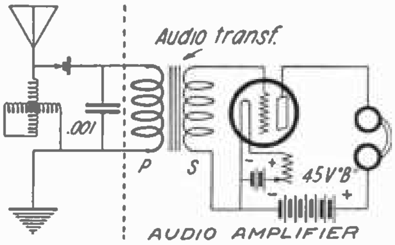

Fig. 2. A one stage Audio Amplifier connected to a crystal receiving set.

Fig. 2 shows one stage of audio frequency amplification connected to a crystal receiving set. There are numerous variations of audio amplification. (See *Push-Pull Amplification*, also *Resistance Coupled Amplifier*, and *Power Amplifier.*)

AUDIO FREQUENCY — A frequency corresponding to normally audible sound waves, i.e., between 40 and 10,000 per second. The limits of audibility are given variously from 16 to 20,000 cycles, but for ordinary usage may be taken as from 40 to 10,000 cycles per second. In audio frequency amplification as used in conjunction with the tuning and detector apparatus of the radio receiving set, the volume of the received sound is amplified or increased over that of the originally rectified sound waves. (See *Radio Frequency;* also *Amplifier.*)

AUDION—The name given by Dr. L. De Forest to a *vacuum tube detector* which possesses a glowing *cathode* and has a steady E. M. F. (*Electromotive force*) permanently applied between *anode* and hot *cathode* and two cold *anodes*. (See *Vacuum Tube.*)

AUSTIN, Louis Winslo, Ph.D.—Born in Ordwell, October 30th, 1867. Graduated at Middlebury College and studied at Clarke University and the Universities of Strasburg and Berlin. Noted for the fact that he is head of the U. S. Naval Telegraphic Laboratory, Washington, D. C., and for his work in measuring high frequency currents.

AUTO-COHERER — A coherer having automatic action to release the particles that cohere when circuit is established. Now obsolete. (See *Coherer.*)

AUTODYNE RECEPTION—See *Self-Heterodyne Reception.*

AUTO INDUCTIVE COUPLING—See *Coupling.*

AUTOMATIC INTERRUPTER—A vibrating electro-mechanical device to make and break an electric circuit by using the energy passing through it. The most common example is that of a *spark coil.* An interrupter may be used for several purposes. In the case of a spark coil it is used to produce sparks for transmission of signals. In another form it may appear in a mechanical battery *charger*, in this instance being used to interrupt the *alternating current* and permit passage of the current in one direction only. A *buzzer* uses a type of interrupter—a vibrating device producing a buzzing sound having a pitch dependent on the rapidity with which the vibrations take place.

AUTOMATIC TRANSMITTER—An apparatus for operating a sending *key* mechanically. (See *Wheatstone Transmitter.*)

AUTOPLEX—A type of radio receiver employing one vacuum tube, operating along the lines of the *Super-regenerative* principle and designed to operate a Loud-speaker. The circuit comprises two standard variometers, a 1500 turn duo-lateral coil, and the usual vacuum

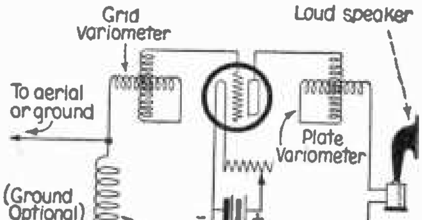

Circuit of the Autoplex radio receiver.

tube, socket, batteries, etc. It is extremely critical in operation, but if carefully assembled will give excellent volume under good conditions. Illustration shows the circuit diagram of the Autoplex receiver.

AUTO-RECEIVER—A device for the automatic reception and recording of radio signals. It is essential in the reception of signals transmitted by an Automatic transmitter (see *Wheatstone Transmitter*) as the speed is too high to permit reception by ear. There are three general classes of automatic receiver: a *resonance intensifier* in conjunction with an inking arrangement; a *phonographic recorder* and a recorder using *photographic means*. In the latter method, a powerful beam of light is projected on to a mirror which is arranged to respond to the signals in approximately the same manner as a mirror *galvanometer*. In the phonographic type, the signals are reproduced on a drum. (See *Resonance Intensifier*, *Photographic Recorder*, also *Phonographic Recorder*.)

AUTO TRANSFORMER—A device for changing voltage in a circuit carrying alternating currents. A form of *transformer* having but one winding, any part of which may be used as the *primary* and any part as the *secondary*. In Fig. 1 the two windings P and S are continuous, although different sizes of wire may be used according to the purpose for which intended. In this case the voltage taken from the secondary D E will be less than the voltage impressed at A B, because the secondary has fewer turns of wire than the primary. Now in Fig. 2, this conditions is reversed, and as the primary has only a few turns of wire and the secondary has a comparatively large number, the voltage obtained from D E will be greater than that impressed at A B. The voltage ratio is approximately proportional to the ratio of wire turns. Thus, if the primary has 100 turns of wire and a pressure of 100 volts is placed at A B, then the voltage at D E would be within certain limits about equal to the number of turns of the secondary. In other words, if the number of turns on the secondary is one-third the number of turns on the primary, then the voltage at D E will be approximately one-third that impressed at A B. The main advantage of the autotransformer is that a saving is accomplished by having the primary and secondary windings combined. The chief disadvantage is that where extremely high voltages are used there is always danger of the two windings becoming crossed or shorted. (See *Short.*) For this reason they are used mainly for comparatively low voltages. One of the most common forms is the step-down type, known as a "bell-ringing transformer." These are used to cut down the voltage of an electric light line to suitable value for use with small toys or bells. If the transformer is used to increase the voltage, it is known as a "step-up" transformer. (See *Transformer—Radio, Audio.*)

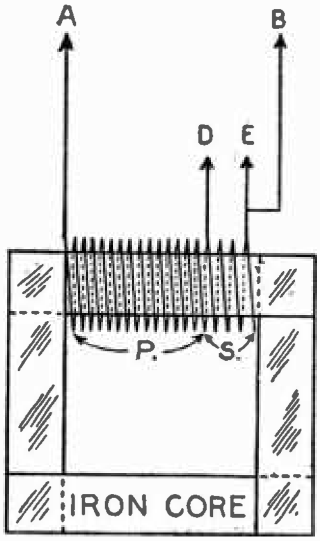

Fig. 1.

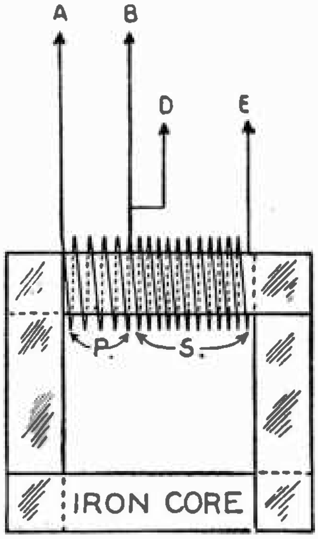

Fig. 2.

AUTO TRANSFORMER COUPLING—See *Coupling.*

B

B—Abbreviation for Baumé, the scale used in a *hydrometer* (*q.v.*)

B. A. AMPERE—The standard ampere fixed by the British Association for the Advancement of Science.

BACK COUPLING—See *Reactance Coils.*

BACK ELECTRO-MOTIVE FORCE—An *Electromotive force* (E.M.F.) which opposes the original electromotive force. The more common term is "*Counter E. M. F.*" In many arrangements of electrical circuits there will be an electromotive force set up in such manner as to oppose the impressed voltage or force. This will be true in the case of a *choke coil* in which any variation in the pressure or electromotive force applied to the coil will result, under proper conditions, in the production of a back or counter-electromotive force that will oppose the change in the original force.

This phenomena is frequently found in electrical circuits; at times introduced for a purpose and at others self-introduced, and therefore demanding consideration in design of apparatus. The rotation of the *armature* of a motor creates a counter E. M. F. opposed to the impressed E. M. F. and upon which the driving power of the motor depends. When an external E. M. F. is impressed on a circuit in which there exists a local E. M. F., the flow of the current through that portion of the circuit containing the local E. M. F. will result in an increase or decrease of energy depending on whether the local E. M. F. is opposed to, or in the same direction as, the external E. M. F. (See *Counter Electromotive force.*)

BACK OSCILLATION—When the *condenser* of a transmitting circuit of the spark type discharges across the *spark gap* some of the *high frequency* current may flow back through the secondary of the transformer instead of taking its normal path across the gap. This back flow is usually referred to as *Back oscillation* or kick back, and is apt to break down (see *Breaking Down of Insulation*) the insulation of the transformer if not prevented or controlled. Back oscillation is retarded ordinarily by means of coils of wire having a certain arbitrary *inductance* value. When the high frequency discharge from the condenser has any tendency to break through, the *reactance* of these coils becomes so great that they act in effect as insulators. As the capacity and inductance of these coils might create an *oscillatory circuit* of the same *natural period* as the primary oscillator, they are shunted with non-inductive carbon resistances. These coils are termed "*choke*" coils.

BAKELITE—An insulating material having widespread use in radio. It possesses great strength and has a high *dielectric* (insulating) value together with the ability to withstand high temperatures. This material is furnished in many forms and under a variety of trade names, being considered one of the best of the insulating compounds.

Chemically known as *Oxybenzylmethylenglycol anhydride*. It is produced by the union under heat and pressure of phenol and formaldehyde together with a small percentage of some alkaline agent. Is heat resisting up to about 500 degrees Fahrenheit—disintegration setting in somewhere above that point.

BALANCED CIRCUIT—An electric circuit arranged in such manner with relation to adjacent or neighboring circuits as to do away with the effect of *mutual induction* (*q.v.*).

BALANCED DETECTORS—A term applied to an arrangement of opposed rectifying *detectors* for reducing the effects of strong *strays* in receiving apparatus. The detectors must be balanced for violent *electro-motive forces*, but not balanced for the moderate E. M. F.'s due to the signals. Actually, a method of eliminating stray impulses in reception.

BALANCED METALLIC CIRCUIT—A circuit through conductors (wire or any metal) in which the total resistance is the same on each side of the circuit. (See *Circuit.*)

BALANCED ROTARY CONDENSER—A variable condenser having some arrangement for keeping the plates accurately balanced in rotation at all points. Fig. 1 shows a form known as the "balanced plate" type. In this case the plates are so arranged that half the movable plates are on one side and half on the other side of the shaft.

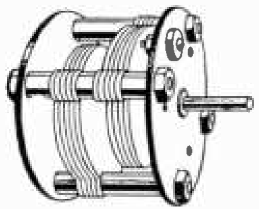
Fig. 1. A "Balanced-Plate" type Rotary Condenser.

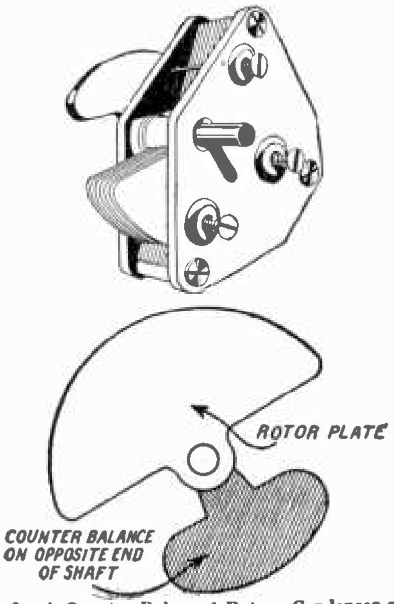

Fig. 2. A Counter-Balanced Rotary Condenser and illustration showing how the counter balance is attached to shaft.

This maintains the balance of the part. Fig. 2 shows another method known as "counter-balanced." The weight of the balance is the same as that of the movable plates and its position on the other side of the shaft serves to maintain balance. The idea is to prevent the plates slipping due to lack of balance and thus keep the setting at the desired point after adjustment.

BALANCING—Any method of eliminating interference by reducing the *amplification* of the undesired signals while permitting full amplification of the desired signals. (See *Interference Eliminator.*)

BALANCING AERIAL—An aerial system designed to permit reception of signals from distant points without interference from a powerful nearby transmitter. The system involves the use of two aerials, one for balancing, the other for receiving. The receiving aerial is tuned to the wavelength of the desired signals, while the balancing aerial is tuned to the wavelength of the nearby station (the interfering signals).

In operation the receiving circuit is coupled to each aerial *electro-magnetically*, and is acted on by both series of *oscillations* (distant and nearby waves). By careful adjustment of *coupling* and phase relation of the *induced currents*, the interfering wave may be balanced out in effect and the desired oscillations received.

BALDWIN RECEIVER—Perhaps the most sensitive type of telephone receiver thus far developed. It employs what is known as a "balanced armature," the signals being produced by the vibrations of a *diaphragm* which is not acted on directly by the magnets. The illustration shows the general plan of a receiver of this type. The *armature* (A) is of soft iron and is pivoted between two U-shaped soft iron pieces (M1, M2) mounted on the ring-shaped horseshoe magnet (M) as indicated. The armature is acted on by the magnets in response to incoming signals and the movement of this element in turn acts on the mica diaphragm (D) by means of a fine brass wire (C). The usual windings (R) are placed between the two pole pieces, the armature being mounted in a central slot. These receivers are much used both for headphones and for loud-speaker units. In the latter case the mica diaphragm is generally made heavier to handle the more powerful vibrations due to the *amplification* of the signals.

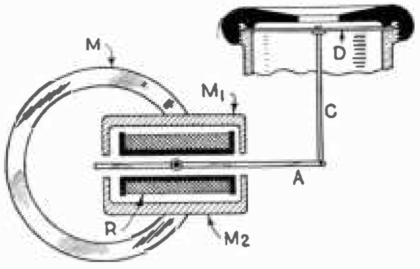

Principle plan of the Baldwin Receiver unit.

When no fluctuating impulses (signals) flow through the windings (R), there is no *magnetic stress* on the armature (A), because this member is suspended centrally between the pieces (M1) and (M2), the magnetic attraction of which are equal. Now when a fluctuating impulse flows through the windings (R) it produces a magnetic flux which combines with the flux of the permanent magnet (M) and the total flux is distributed asymmetrically or unevenly on both sides of the armature. The result is a rocking vibratory movement to this member which is in turn communicated by the juncture wire (C) to the diaphragm (D), producing the audible signals. (For a comparison of this action with the conventional telephone receiver action, see *Telephone Receiver.*)

BALKITE BATTERY CHARGER—An electrolytic battery charger which makes use of a sulphuric acid solution and tantalum as the valve metal. Used for charging *storage batteries* from *alternating current.* See illustration.

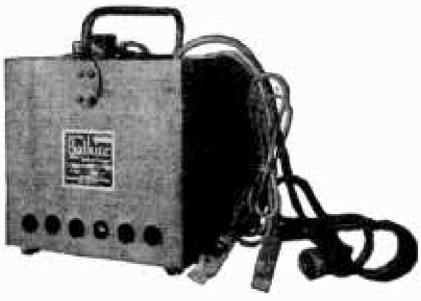
Photo by courtesy of Fansteel Products Co.

A Balkite Battery Charger.

Note: The name Balkite is derived from the name of the man who perfected the process of extraction of tantalum from its ores.

BALLAST TUBE—An automatic means of regulating the *filament current* of vacuum tubes. The device is enclosed in a glass tube and is arranged to permit passage of a certain fixed amount of current. The ballast element is a wire whose *resistance* changes with the current through it, due to the change of resistance with the temperature of the wire. When the *electromotive force* increases beyond normal, the increased current raises the temperature and hence the resistance, so that the current is held to the normal value.

B. and S.—Abbreviation for Brown and Sharpe Gauge (*q.v.*).

BAND OF FREQUENCIES—A continuous range of *frequencies* extending between two frequencies. (See *Band of Wavelengths;* also *Band Pass Filter.*)

BAND PASS FILTER—A *filter* arrangement designed to permit passage of all *frequencies* extending between two definite frequencies but excluding all other frequencies. (See *Band of Frequencies.*)

BAND OF WAVE-LENGTHS—A continuous range of *wave-lengths* extending between two definite wave-lengths. For example, the usual broadcast band of wave-lengths is approximately 250 meters to 550 meters.

BANKED BATTERY—A *battery* that has its *cells* connected in *parallel.*

BANK WINDING—A form of winding for coils in which the turns are staggered so that two or more layers may be used without having large distributed capacity losses. The illustration shows method of bank winding for two layers. Note that in this form the *electrostatic capacity* between any two turns is only that between adjacent turns, representing only a small part of the total winding, whereas on the usual form of winding the adjacent turns would represent a good portion of the total *inductance* and hence the effect of distributed capacity would then be much greater. (See *Distributed Capacity.*)

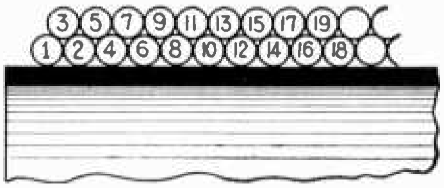

Method of bank winding wire on a coil form.

BARRETER—A receiving instrument consisting essentially of a small mass of conducting material that is heated by the passage of an *oscillatory* current, and arranged so that the consequent alteration of electrical *conductivity* affects an indicating instrument, such as a telephone receiver or *galvanometer.*

BASKET COIL—Variometers—couplers, etc. (See *Basket Wound.*)

BASKET WOUND—A method of winding coils, such as variometers or variocouplers. The winding is a latticework of wires in skeleton form, the object being to save weight and eliminate large losses such as occur in the types using a great deal of solid insulating material. The illustration shows a popular form of basket variometer.

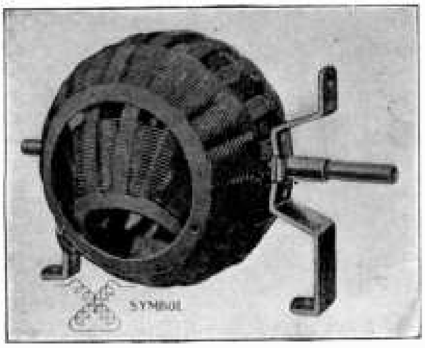

Photo by courtesy of Amer. Radio & Research Corp.

A Basket Wound Variometer.

BATTERY — In the electrical field, a group of *primary or storage cells, dynamos*, or *condensers* grouped together to form a single unit. The term is often used incorrectly to refer to a single cell—as "dry battery," the correct name for which is "dry cell." Thus a group of condensers joined together as one unit is termed a battery, also a number of dynamos arranged as a single unit for the production of electrical energy. (See *Dry Cell*, also *Battery of Alternators, Storage Battery, "A," "B" Battery*, etc.)

"B" BATTERY—An appliance for obtaining electricity of sufficient voltage to supply the plate of a *vacuum tube*. There are two types of "B" batteries, viz., the *dry cell* type and the storage types. In order to obtain the high voltage required for vacuum tube plates, many small cells are connected in series. The usual method is to connect them together and seal in a square package. Often taps are brought out so that various voltages may be obtained for *critical* vacuum tubes. The most popular voltages which dry cell "B" batteries give are 22½ and 45 volts. The storage "B" battery is similar to the ordinary storage battery with the exception that the cells are much smaller, but contain more cells in series. These can be charged in the same manner as the "A" or *storage battery*. In Fig. 1, we have a popular type of "B" battery. This is tapped in order that different voltages may be obtained.

EVEREADY
RADIO

Photo by courtesy of National Carbon Co., Inc.

Fig. 1-A. A popular type of "B" Battery of 22½ volts.

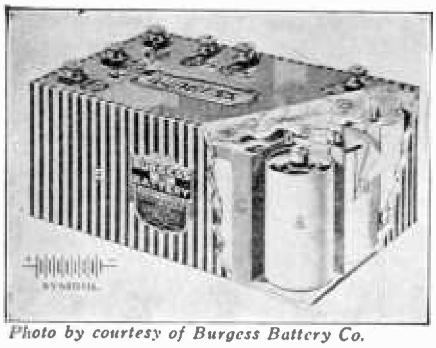

Photo by courtesy of Burgess Battery Co.

Fig. 1-B. A "B" Battery showing the contents of small cells.

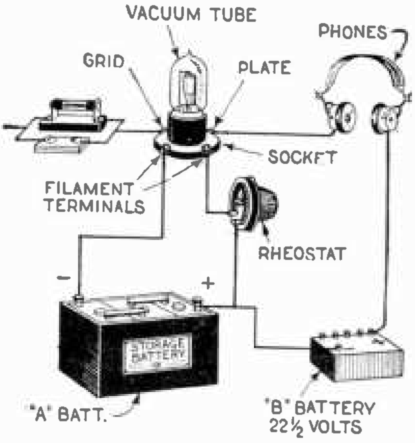

Fig. 2. Manner of connecting a "B" Battery in the circuit of a vacuum tube detector.

The manner of connecting up the "B" batteries in radio receiving is shown in Fig. 2. Here one battery of 22½ volts is used. In Fig. 3, three batteries are connected in series so that 67½ volts are obtained.

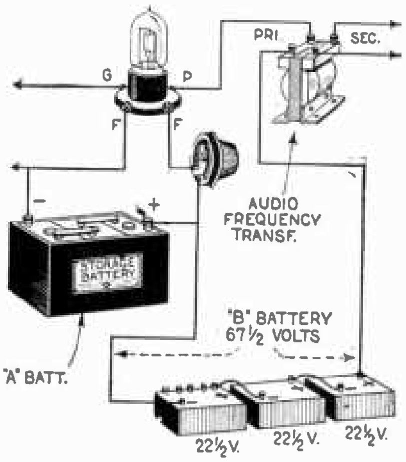

Fig. 3. "B" Battery of 67½ volts connected in an amplifier circuit.

A common "B" battery may be used for the detector tube and the amplifier as shown in Fig. 4.

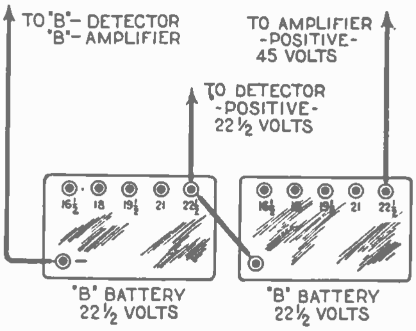

Fig. 4. How a common "B" Battery may be used both for detector and amplifier.

If separate batteries for each tube are desired, the circuit should be connected as in Fig. 5. This lengthens the life of the "B" battery as the drain on the battery is much less, i.e., only one vacuum tube is operated on each battery.

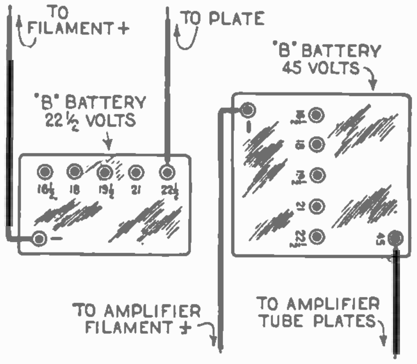

Fig. 5. Method of employing separate "B" Batteries for the detector and amplifier.

BATTERY OF ALTERNATORS — In ordinary electrical work such as house lighting, several *alternators* for the production of alternating current may be used together as one source of power. Also in high-power radio transmission, several *high-frequency alternators* may be used together to produce added power. (See *High-Frequency Alternator*.)

BATTERY TESTING INSTRUMENTS —The various devices used for determining the condition of batteries used as radio apparatus. (See *Ammeter, Voltmeter*, also *Hydrometer*.)

BEAT FREQUENCY—The *frequency* or number of vibrations per second of a series of *oscillations* representing the difference between two other series having different frequencies. It is numerically equal to the difference between the original two series of oscillations. Thus, if we have a series of oscillations having a frequency of 100,000 *cycles* per second, and combine it with another series having a frequency of 90,000 or 110,000 cycles, the *beat frequency* will be 10,000 cycles or the difference between the other two frequencies.

X. The frequency of recurrence per second of either maxima of addition or minima of opposition of two superposed periodic phenomena having the same nature but of different frequencies. The time measure of vibration of a beat or heterodyne. (See *Heterodyne*, also *Beats*.)

BEATS—When two sets of *oscillations* of different *frequencies* occur in the same system, the difference in the frequency or rapidity of vibration causes a new set of oscillations having a frequency equal to the difference between the original two sets. This may be accidental as in the case of certain types of receivers such as the *regenerative* type, in which case the beat oscillations are undesired and cause interference or disturbances in the set. Then again a beat frequency may be purposely introduced as in the case of the *Super-Heterodyne receiver*. Here we have the natural incoming oscillations from the antenna combined with a series of oscillations produced locally in the system. The result is a new set of oscillations where the incoming series and the locally produced series are not of the same frequency. (For more complete explanation see *Super-Heterodyne Receiver*.)

X. Where two *periodic phenomena* are *super-posed* or run together and the frequencies differ, the gradual change in *phase difference* produces a condition wherein the *amplitudes* are in opposition at one instant and in concurrence at a later instant with the various intermediate stages during the interval. (See *Heterodyne*, also *Beat Frequency*.)

BEG OHM—A *resistance* of one billion *ohms*. (See *Megohm*.)

BELIN, EDOUARD—French scientist, the inventor of what is considered the first practical system of radio transmission of photographs. The device can be used for sending photos by cable or telephone lines as well as radio. (See *Transmission of Photographs by Radio*.)

Brown photo.

Edouard Belin.

BELL, Alexander Graham (1847-1922)—Scottish scientist. Born in Edinburgh, March 3rd, 1847, he was educated at the High School and University and graduated as a doctor of medicine. In 1870 he went to Canada, and in 1872 became professor of vocal physiology in the University of Boston. In 1876 he exhibited his apparatus for the transmission of sound, afterwards developed into the telephone. Bell was experimenting with an electric invention by means of which he hoped to make speech visible to the deaf. A delicate metal reed was caused to vibrate by spoken speech and to transmit an electric current to the opposite end of a wire, where the vibrations of the first reed were reproduced in a second reed by a magnet. He found that it was possible to transmit not merely vibrations of the original reed, but to reproduce the sound itself in the vibrations of the second reed. In 1878 he invented the photophone, to enable sound to be transmitted by variations on a beam of light, and later a phonograph. Bell was the author of many scientific papers, was awarded the Albert Medal of the Royal Society of Arts in 1902 and the Hughes Medal of the Royal Society in 1918. He died Aug. 2nd, 1922.

Alexander Graham Bell.

BELLINI, Dr. Ettore—Born in Foligno, Italy, April 13th, 1876. Educated at the University of Naples. Noted as Electrical Engineer to the Royal Italian Navy, and chief of the Naval Electrical Laboratory at Venice. Joint inventor with Captain Tosi of the "Radiogoniometer" (*q. v.*), a device for finding the direction of transmitted radio signals.

BELLINI-TOSI AERIAL—See *Goniometer.*

BELLINI-TOSI DIRECTION FINDER—See *Goniometer.*

BEZEL—In mechanics, generally a groove and flange made to receive a beveled edge. In its adaptation to radio some liberty was taken with its real meaning. Its significance in this sense is a small metal ring having a wire mesh or glass center. It is fitted into a circular hole in the panel of a radio receiver and used as a peep hole to enable the operator to know at all times the conditions of the tubes, whether or not they are properly lighted.

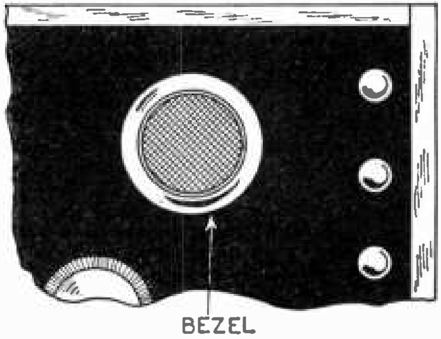

A Bezel on the panel of a radio receiving set.

BILLI CONDENSER—A variable *condenser* of low *capacity*, consisting of two brass tubes, one of which is arranged to slide in and out of the other for the purpose of varying the capacity. Such condensers were much used before the days of broadcasting, in conjunction with any circuit using a crystal and battery for detector. This type of condenser has long been obsolete in the United States.

BINDING POST—A screw and nut arrangement used on electrical units and radio apparatus to make convenient connections from an external source to the desired point within the apparatus. Binding posts are provided, for instance, to make handy connections from batteries to the set and from the aerial and ground leads. The illustration shows a popular type of binding post.

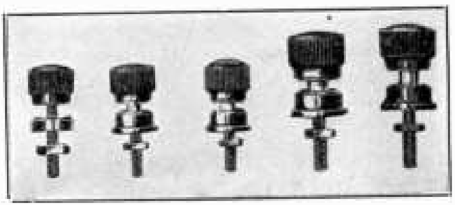

Photo by courtesy of The H. H. Eby Mfg. Co.

A group of Binding Posts as used on electrical and radio apparatus.

BI-POLAR—Having two *poles.* Usually a *dynamo* or motor whose *armature* rotates between a *field magnet* having only two poles. The most modern types are of the Multi-polar variety. (See *Bi-polar Magnetic Field.*)

BI-POLAR ARMATURE WINDING—An *armature* wound in a manner to permit its use with a dynamo having a bi-polar magnetic field.

BI-POLAR MAGNETIC FIELD—The magnetic field (*q.v.*) created between two magnetic poles. The illustration shows a common form of bi-polar magnet for a dynamo. Bi-polar magnetic fields are now used as a rule only for direct current machines of low power, such as those under 5 *Kilowatts.* (See *Dynamo*, also *Generator.*)

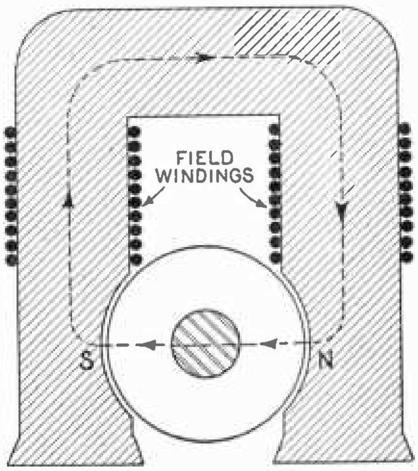

A Bi-Polar Magnet of a dynamo.

BI-TELEPHONE RECEIVER—A headset with two receivers or phones as generally used in radio. In the early days of radio transmission it was the custom to use only one phone.

BLONDEL, Andree—French electrical expert. Born at Chaumont, France, in 1863, he graduated at Paris University, and studied electric waves, on which subject he early contributed a number of papers to various scientific journals. In 1893 he invented the *oscillograph*, an instrument somewhat similar to a mirror *galvanometer*, for showing curves of *oscillating* or *alternating currents.* This invention opened up a fresh field in the study of alternating currents. In the same year Blondel explained for the first time mathematically the effect of inertia in the shunting of *alternators.* He is responsible for a system of acoustically syntonic wireless telegraphy, and for directed waves produced by a double aerial. In 1902 he patented a method for producing electric oscillations for wireless telephony, and has written many papers on microphonic control for transmitters, wireless telephony, the singing arc, etc.

BLONDLOT, Professor Prosper Rene—French wireless expert. Born at Nancy, France, in 1849, studied at

Paris, and became professor at the faculty of sciences, Nancy, and afterwards Honorary Professor and Correspondent of the Institute of France. Professor Blondlot is famous for his studies of electro-magnetic waves, particularly with regard to their speed, and the laws of *propagation* of wireless waves in various media.

BLOWER MOTOR—A motordriven fan used to deliver a high pressure blast at the *spark gap* to prevent *arcing*. (See *Spark Gap*, also *Spark Discharger*.)

BLOWOUT—See *Magnetic Blowout.*

BLUE GLOW—A condition within a *vacuum tube* when the vacuum has become poor. After continued use a tube will often hold a small percentage of gas, which causes a blue glow when current is passed through it. A tube in this condition is a poor *detector* or *amplifier* and should be replaced. The condition should not be confused with the glow often caused by an excessive voltage. In this latter case decreasing the voltage to the normal value will generally permit the tube to be operated efficiently. The blue glow in this case is due to *ionization* of the residual gas by the excessively high *potential* impressed across the elements of the tube.

BOARD OF TRADE UNIT—B.O.T. 1,000 Watt Hours. One and a third Horse Power.

BODY CAPACITY—The effect of the human body when tuning a radio receiving set. The hand when placed on or near the controls very often throws the receiver out of balance with the incoming signals. In the case of very *sharp tuning* this effect is more in evidence and is likely to cause howls due to *self-oscillation*. The howling is due to the production of an audible *beat frequency* in the system, caused by the combination of the local *oscillations* with the incoming signal oscillations. The remedy may be to shield the panel with metal foil or in some cases to merely alter the direction of the leads from the *tuning condenser*. The experienced operator seldom pays much attention to the phenomena, as a little practice enables the listener-in to compensate for the effect. To make this more clear, if, when tuning the set, the withdrawal of the hand from the dial *detunes* the set or throws it out of balance, it is due to the fact that the hand has acted as a certain amount of *capacity* in the circuit, and naturally its withdrawal is equivalent to a change in the condenser setting. A simple system is to turn the control beyond the point of maximum volume, then when the hand is withdrawn the capacity will drop to its proper value. This can be mastered very readily. If the tuning condenser is connected across the secondary, i.e., one side to the grid, and the other to the grid return, the stator plates of the condenser should be connected to the grid of the tube. If metal foil or thin copper sheeting is used for shielding it should be connected to the ground post of the set.

The tendency of the human body to insert an arbitrary capacity in the circuit. The phenomena is especially noted where maximum volume is obtained only by very critical control close to the oscillating point, in which case the circuit may be thrown out of resonance.

BOLITHO CIRCUIT—A *super-regenerative circuit* patented in England in 1919 by Captain J. B. Bolitho. The circuit was originally intended to operate a relay device. The circuit has been adapted to use as an *amplifier* for radio reception.

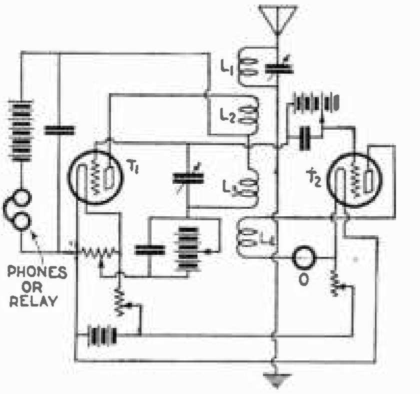

Diagram of the Bolitho Circuit.

In the illustration the phones may be replaced by a loud speaker or relay. In operation, T1 is held at a point just below that where *oscillation* sets in by means of T2 which is excited by an oscillator O in the plate circuit and coupled by means of coil L4 to the tuned grid circuit through L2 and L3, the coupling, as explained, being so arranged that the tube is always just below the oscillating point. A *reaction coil* L4 is placed in the plate circuit of T2 in such manner as to oppose the magnetic linkage between coils L2 and L3. The *frequency* of the generator O is lower than that of the received signals. As indicated, the grids of both tubes T1 and T2 are joined together and the tube T2 maintains the circuit in a receptive condition when tube T1 is tuned just below the oscillating point.

The oscillator O makes the *plate* of T2 alternately *positive* and *negative* with the following effect: When the plate of T2 is negatively charged there is no current in coil L4, this coil therefore not affecting coils L2 and L3, and permitting tube T1 to build up self oscillation. When the plate is made positive by the action of the oscillator, current flows through coil L4, and as it is coupled in opposition to coils L2 and L3 it neutralizes the coupling between L2 and L3 and prevents transfer of energy between the grid and plate circuits of the tube T1. This tends to stop the self oscillation and makes the circuit receptive to the energy (signal) produced by outside signals in the aerial coil L1. The frequency, of course, is determined by the frequency of the oscillator O. The system is especially adapted to the reception of continuous wave signals. (See *Super-regenerative circuit*, also *Feed-back* and *Regenerative circuits*.)

BOLOMETER—Type of *Wheatstone Bridge* having an easily heated *resistance*, such as a very fine wire in one arm. (See *Wheatstone Bridge*.)

BOOSTER—An expression signifying a small *dynamo* used in conjunction with main dynamo to temporarily raise, when necessary, its normal *pressure*. It is generally driven by a motor supplied with energy from the main generator and thus becomes in effect a *continuous current transformer*. Frequently used for charging *accumulators* of a generating plant. The term "Boost" is also used to denote increase or a stepping-up of any electrical quantity. (See *Amplification*.)

BORNITE—A *crystal rectifier* much used in radio reception. It is a natural sulphide of iron and copper, having a metallic blue lustre. This mineral is used in combination with zincite or copper pyrites as a crystal detector. Such combinations are generally known as "crystal to crystal" detectors to distinguish them from the ordinary variety using one mineral and a wire contact.

BORON—A non-metallic chemical element used in radio as one of the electrodes for the *T. Y. K. Arc.* Chemical symbol B, atomic weight 11.0, specific gravity 2.6.

B. O. T.—The customary abbreviation for *Board of Trade Unit.* The unit represents 1,000 *watt hours*. (See *Board of Trade Unit*.)

BOX AERIAL—A term occasionally applied to loop aerials. (See *Loop aerial*, also *Frame Aerial*.)

BRADFIELD INSULATOR—A particular form of *Lead-in Insulator*, consisting of an ebonite tube provided with zinc cone and ebonite spark discs, for breaking up continuous streams of rain running down outside which might cause the aerial to become grounded. The whole is held in position, half way through roof of operating room, by means of a stuffing box. The aerial is led in by means of a conducting rod through center of tube. (See *Petticoat insulator*.)

BRANLY COHERER—Early form of *Marconi Coherer*. (See *Coherer*.)

BRANLY, EDOUARD—French radio expert. Born at Amiens, France, Oct. 23rd, 1844. He was educated at Paris and afterwards became Fellow of the University, doctor of physical science, and doctor of medicine. Branley early made a study of electro-magnetic waves, and in 1890 and 1891 patented methods of operating a local relay circuit from a distance by means of wireless waves. In 1900 he was awarded the Grand Prix by the International Jury of Superior Precept Instruction for his exhibition of radio-conductors.

In 1890 Branly published an account of his very extensive series of observations on the electrical conductivity of loosely packed metal filings, and he made the extremely important observation that an electric spark at a dis-

Edouard Branly.

tance had the power of suddenly changing the electric *conductivity* of loose masses of powdered *conductors*. To Branley is due the *coherer* named after him.

BRAUN, FERDINAND — Professor at the University of Strasburg, and one of the leading world authorities on Radio transmission. As early as 1899 Braun was granted a patent for *closed oscillating* systems with an *inductively coupled* antenna. The system was claimed by Braun to possess a much greater efficiency than the directly coupled systems. The Braun transmitting set, as manufactured by Siemens and Halske, consisted of a large coil worked into an *electrolytic interrupter*, a set of *Leyden jars*, enclosed *spark gap*, *oscillation transformer* wound with insulated wire placed in oil, and for the receiving set the standard type of *coherer relay* and Morse register. In 1899 Braun established communication between Cuxhaven and Heligoland, using aerial wires 90 feet high, and the inductive coupled aerial connection for transmitting. In 1903 Braun joined with Slaby, von Arco, and Siemens to form the Telefunken system of transmission. Professor Braun was awarded the Nobel prize with Marconi in 1900, for his work in wireless. Braun has devised a method of directional wireless which depends upon the interference of *electric waves* travelling in the same direction but different in *phase*. Three simple vertical wire aerials are set up in positions corresponding to the angular points of an equilateral triangle, and *oscillations* are created in these which differ from one another in phase. In 1897 Professon Braun published a description of his cathode ray tube, and afterwards pointed out, in 1902, how such a tube could be used to trace the forms of alternating current waves.

BRITISH STANDARD WIRE GAUGE —The standard wire gauge of Great Britain. The table shows the various diameters in thousandths of an inch (mils.).

WIRE GAUGES IN MILS.

GAUGE	New British Standard
0 000 000	500
000 000	464
00 000	432
0 000	400
000	372
00	378
0	324
1	300
2	276
3	252
4	232
5	212
6	192
7	176
8	160
9	144
10	128
11	116
12	104
13	92
14	80
15	72
16	64
17	56
18	48
19	40
20	36
21	32
22	28
23	24
24	22
25	20
26	18
27	16.4
28	14.8
29	13.6
30	12.4
31	11.6
32	10.8
33	10.0
34	9.2
35	8.4
36	7.6
37	6.8
38	6.0
39	5.2
40	4.8

BREAK or BREAKER — See *Circuit Breaker.*

BREAKDOWN POTENTIAL or BREAKDOWN VOLTAGE—The voltage necessary to *break down* or puncture a *dielectric*. The ability of any *insulating* material to resist breakdown is an important matter. An insulator or dielectric material is usually given a certain rating or point above which it is likely to fracture. Thus, a condenser may have a breakdown rating of 500 volts. That is to say that voltages up to 500 and slightly over may be handled with safety, but any large increase above that figure may result in puncture of the material by the excessive potential or voltage. The relation between dielectric strength and breakdown voltage is shown in the following formula:

$$V = cd\ {}^{2}\!/_{3}$$

Where d is the thickness of dielectric
in millimeters
V the potential difference in volts
c-a constant representing the potential difference required to cause a dielectric breakdown of a sample of the material 1 mm. thick. (See *Dielectric Strength.*)

BREAKING DOWN OF DIELECTRIC—In a *condenser*, the gradual weakening of the *dielectric* or insulating sheets which eventually causes them to permit direct passage of the current. For example, if a force of 1000 volts is applied for a considerable period to a condenser that has been designed for a maximum of 500 volts, the insulating material between the plates would probably be punctured by the current. (See *Break-down Potential.*)

BREAKING DOWN OF INSULATION —The same effect as referred to under *Breaking Down of Dielectric.* The insulation weakens and permits the ready passage of electric currents. (See *Break-down Potential.*)

BRIDGES—See *Capacity Bridge,* also *Wheatstone Bridge.*

BRIDGING—A term occasionally used for *shunt* (q.v.) connections. Thus, a grid condenser is generally shunted or bridged by a *grid leak.*

BROADCAST—Transmission of music, news and other matters of general interest and entertainment by means of *radio telephony*. (See *Broadcasting, General Treatise on Methods.*)

BROADCASTING, GENERAL TREATISE ON METHODS—The application of radio telephony to the transmission through the ether of musical programs, speech or any form of news or entertainment, in such manner that it can be readily received by anyone possessing a radio receiver sufficiently sensitive and capable of being tuned to resonance with the particular waves being sent out. While broadcasting in its present world-wide scope is a comparatively new enterprise, it has actually been used to a more or less degree since the practical development of the wireless telephone. As near as records can be determined, the first actual broadcasting was done in 1912. It was, however, not until after the world war that it began to develop into an industry of prominence. Now it occupies a place of very definite importance in everday life; carrying news and entertainment into the remote corners of the earth, far removed from telephone or telegraph communication.

The fundamental principles of broadcasting are essentially those of radio transmission, with the addition of means of combining the speech or music vibrations with the regular radio waves. First we have to consider the radio waves of high-frequency—alternating current that changes in direction many thousands of times each second. These high frequency waves are known as the carrier waves. The frequency of these waves varies according to the length of the waves (wavelength), in the case of American broadcasting stations, being between 250 and 550 meters in most cases, which means a frequency range of approximately 550,000 to 1,200,000 cycles.

Now by means of a microphone, a device similar in principle to the mouthpiece of an ordinary telephone,

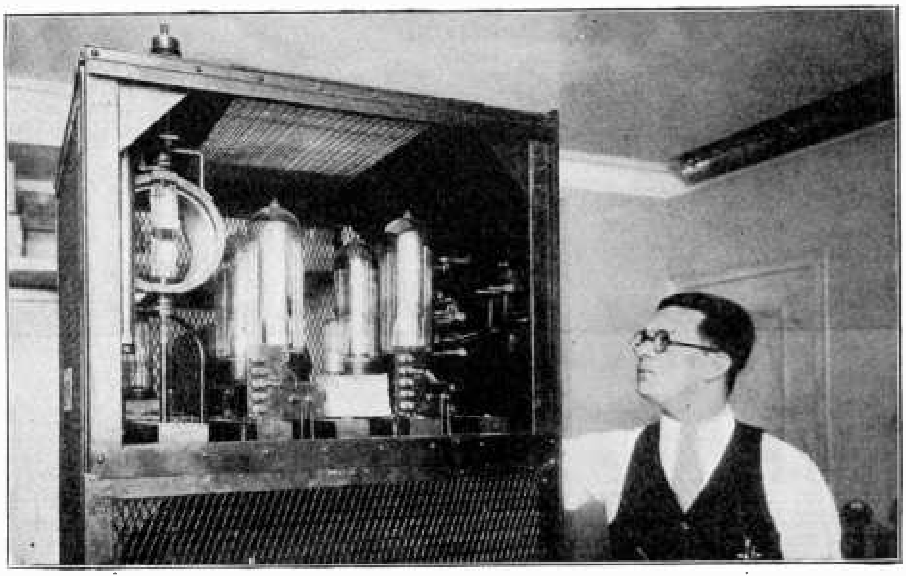

Fig. 1. A bank of 250-watt transmitting tubes in a Radio Broadcast Station. Note the small receiving tube for comparison of size.

but designed to handle large currents, the air pressure waves, i.e., speech or music, are converted into variations of electric current. By use of suitable amplifying apparatus, the strength of the electric waves radiated from the antenna is varied to correspond electrically to the acoustic or sound variations due to the speech or music being impressed on the microphone. The speech or music waves are low frequency waves, generally ranging between 100 and 4000 per second. Thus, while the carrier wave oscillations are of high frequency, the variations in it are low frequency, corresponding to the ordinary speech or music vibrations.

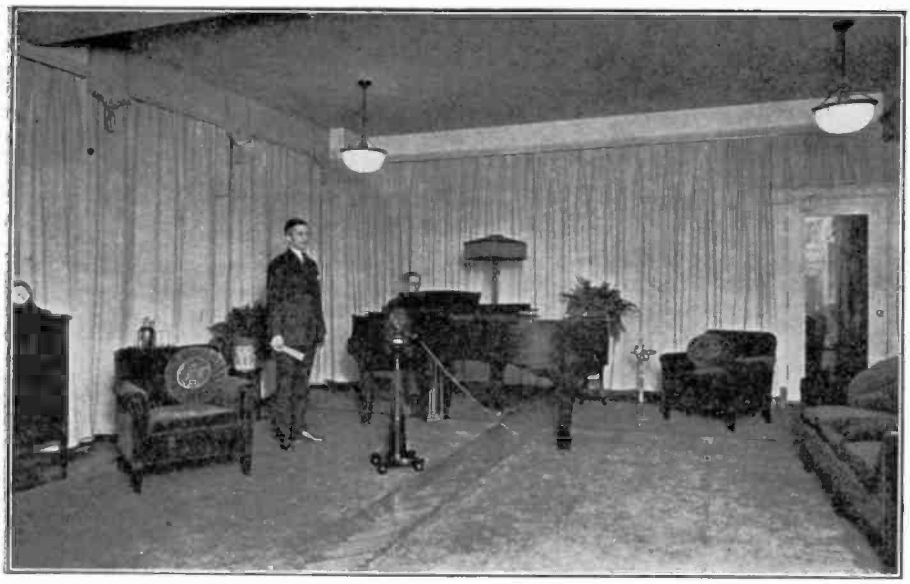

Fig. 2. Studio of a modern Radio Broadcast Station.

The process of varying the radiated wave in accordance with the sound waves is known as *modulation*. This adjustment is of the utmost importance, as defects here result in distorted or improperly modulated sound waves and the reproduction cannot be perfect. The reception of broadcast programs is essentially the same as that of ordinary radio signals, with the exception that care must be used at the receiving end to preserve the form of the original vibrations. When dot and dash signals are received, the tone is of comparatively little importance, whereas with voice or music it is essential that the tone quality be retained.

In broadcasting the high frequency oscillations are produced by means of a tube or group of tubes, practically identical with those used for reception, but designed to handle large currents and voltages. The majority of American stations use vacuum tubes of 250 watts power rating, one or more of them being used to obtain the total power required. These tubes are shown in the illustration, Fig. 1.

When a program is being broadcast, the action is somewhat as follows: The artists are disposed before the microphone with due regard to the effect of certain instruments. That is to say, one type of instrument may be placed near the microphone, while another must be placed farther away, owing to the differences in tone and volume. This is done to prevent any one tone or range of tones from predominating. These speech or music currents are converted into electrical variations by the microphone and then increased many times in power or amplitude by means of a speech amplifier system, comprising vacuum tubes and the necessary associated circuits.

These amplified currents are applied to the grid of a control tube or bank of tubes and result in large variations in the plate current of the tubes. (See *Theory of Vacuum Tube.*) The control tubes are supplied with filament current from storage batteries or special low voltage D. C. generators, and the high voltage applied to the plates of the tubes is furnished by generators. This power will vary in most cases from about 1,000 to 5,000 volts. In some broadcasting stations, the power is obtained through high voltage direct current generators, while in others, the alternating current supply is stepped up or transformed through high voltage transformers and then rectified by means of special vacuum tubes.

The high frequency waves are applied to the antenna system tuned circuits and carefully modulated in accordance with the variations of the speech or music being transmitted. The action of the transmitter and the effect of the modulation must be carefully observed at all times during the transmission of programs. (See *Harmonic Suppressor,* also *Oscillograph.*) Fig. 2 shows the studio of a modern broadcasting station. It is interesting to note that these studios are very carefully arranged and the walls cushioned to prevent troublesome acoustic effects. There must be no reverberation—echoes—or the transmitted program would be distorted by the additional or repeated tones. Fig. 3 shows the main control panels. Fig. 4 illustrates the power generating system of a typical broadcasting station.

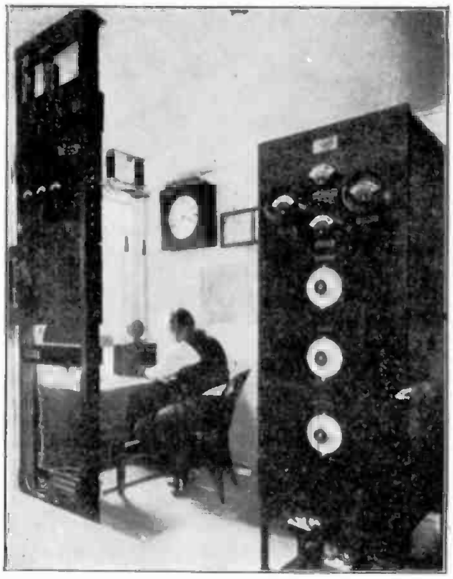

Fig. 3. Main control panels of a Radio Broadcast Station.

The radio waves generated and radiated from the broadcasting station travel away from the antenna at tremendous speed—that of light waves, or about 186,000 miles per second. These waves radiate off into space in

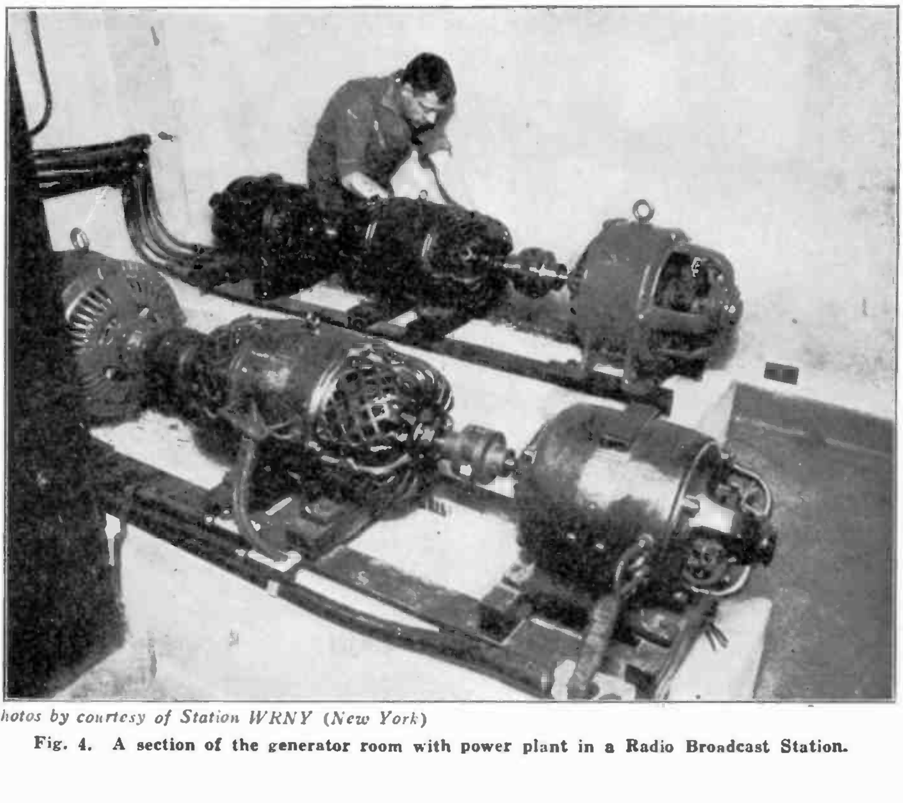

Photos by courtesy of Station WRNY (New York)

Fig. 4. A section of the generator room with power plant in a Radio Broadcast Station.

much the same manner as light waves, but a certain portion apparently follow the curvature of the earth. When these waves impinge on a receiving aerial they follow the path of least resistance, and being then in the nature of electrical impulses, they follow the aerial and lead-in wire to the receiving set, and thence to the ground. The receiver is arranged in such fashion that it corresponds electrically to the antenna circuit of the transmitting station. That is to say, the receiver is placed in resonance with the particular signals desired. The usual system of detecting, either with a simple crystal or by means of vacuum tubes, may be followed by special amplification which will make the signals audible on a loud speaker. In many cases the signals are amplified in their original form (radio frequency amplification), then rectified by the detector and further increased in volume at audio frequency. (See *Modulation*, *Speech Amplifier, Electromagnetic Waves, etc.*)

BROAD TUNING—A term used to designate the lack of *selectivity* in reception or in transmission, the use of several *waves* rather than a sharp, pure wave. If a receiving set is of such a nature that it is difficult or impossible to receive the desired signals without hearing any other, the set is said to be "broadly tuned." (See *Tuning*, also *Selective Tuning.*)

BROKEN CIRCUIT—See *Open Circuit.*

BROWN AMPLIFYING RELAY—An arrangement whereby the comparatively weak signals received by a *crystal detector*, usually of the *carborundum type*, may be *amplified* or increased in intensity, either to allow a recording machine to be used or merely to make the signals more readily audible in the ear phones. The device consists essentially of an arrangement for controlling and stepping up the vibrations in conjunction with a small *battery* or *dry cell.*

One example of such a relay is shown in the illustration. In this case the soft iron *cores* cc are *magnetized* by the *permanent magnets* mm, the poles of which are marked n and s. R1 is a fine wire winding such as is ordinarily used in ear phones and R2 is a larger winding placed over the main magnets mm. Directly above the soft core cc with the windings R1 is placed a steel vibrating tongue V having a contact D which touches lightly contact G. Contact G is of a special alloy of osmium and iridium and the contact D is a carbon button, which in contact with G forms a *microphone*. The winding R2 is joined with the telephone P and the battery B, all in series with the microphonic contact. When current (signal current) enters the leads A—E through the windings R1 the result is a change of *flux* in the core of the magnet which vibrates the tongue V. This in turn varies the current of the battery B and causes correspondingly greater sounds in the ear phones P. Several of these relays may be connected together to permit a much greater *amplification* of the original signals.

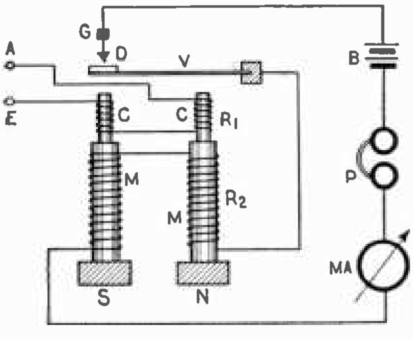

The Brown Amplifying Relay Circuit.

BROWN & SHARPE GAUGE—The American wire gauge adopted as standard for wires for electrical purposes. Commonly known as B. & S. The various sizes run from 40 to 0000, 40 having a diameter of .00314 inch, and 0000 being .46 inches in diameter. We refer to wire having a diameter of .00314 inch as being 40 B. & S. gauge. (*See page 26 for table.*)

BROWN, Sidney George—Born in Chicago, U. S. A., in 1873, of English parents, he was educated at Harrogate and London University. An early study was made by him of the subject of submarine telegraphy. During these investigations he invented the magnifying relay for cables. Another important invention was the cable drum relay. In the year 1898 he invented the magnetic shunt. Since that time many of his activities have been directed to the solution of problems in telephone and wireless telegraphy.

In particular the radio experimenter will know the telephone receivers bearing his name and the important developments of the *microphone relay*. In the field of land telegraphy and telephony, his activities have resulted in the invention of such items as the carbon telephone relay system, which is largely used on land trunk lines for the transmission and reception of telephony.

He is a Fellow of the Royal Society and Vice-President of the Radio Society of Great Britain. His writings on technical subjects are extensive, and numerous valuable patents have been taken out by him, including the vacuum tube *oscillation generator* in 1916 and ionic electric relays in 1918.

BRUSH—A device for collecting current from the *commutator* (*q.v.*) of a *dynamo* or supplying current to the commutator of a motor. Brushes are made in a great variety of forms, the most common being carbon blocks or brass or copper gauze. The illustration shows two different types.

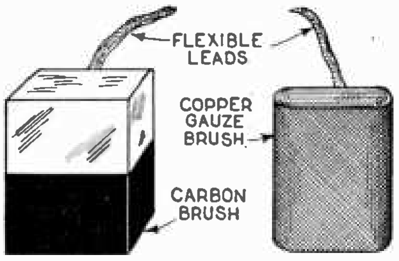

Two different types of Brushes. At the left is shown a carbon brush, and at the right a copper gauze brush.

BRUSH ANGLE—The angle formed by a brush in its contact with the *commutator*. The angle is usually 45°. If the brushes do not bed properly, sparking is apt to occur.

BRUSH DISCHARGE—A faintly luminous discharge which takes place on the surface of *conductors* charged to *high potential*, due to *ionization*. This effect can be noted at times on the *antenna* of a powerful transmitting station. (See *Corona.*)

BRUSH HOLDER—Metal clamp capable of adjustment which holds the *brush* in position on the *commutator* of a dynamo or motor. (See *Brush.*)

BRUSH LOSS—The loss in *watts* or power due to the friction of the brushes against the *commutator*. This loss is at a minimum when the brushes make perfect contact at the proper angle.

BRUSH PRESSURE—A term used to designate the pressure with which the brushes bear on the *commutator* of a dynamo machine and also referring to the voltage or *electrical pressure* delivered at the brushes.

B. S. G.—Abbreviation for British Standard Gauge of Wire, commonly known as B. S. (See *British Standard Wire Gauge.*)

B. T. U.—Abbreviation for "British Thermal Unit," being the commonly used heat unit. The amount of heat required to raise the temperature of a pound of water one degree Fahrenheit at ordinary atmospheric pressure.

BUCHER, Elmer E.—Born Akron, Ohio, Nov. 11, 1885. Educated at high school and private tutors. An American pioneer, Experimental Engineer with De-Forest Wireless Telegraph Company in 1903; constructed several high power stations in Middle West and Gulf Coast and later for United Wireless, which absorbed the Deforest Company. Organized training school for United Wireless and instituted first radio schools for Y. M. C. A. in New York City. Supervised commercial operations and conducted research work for United Wireless; later instructing engineer Marconi Wireless Tel. Co. Technical Editor "Wireless Age" 1913 to 1917. Appointed Commercial Engineer for newly formed Radio Corporation of America in 1920, and since 1922 has been in charge of general sales for that corporation. Author of numerous standard works on radio, including "Practical Wireless Telegraphy," "Wireless Experimenters' Manual," "Vacuum Tubes in Wireless Communication," and many others. Holds numerous U. S. patents on radio systems and devices.

BUCKLED DIAPHRAGM—Warping of the *diaphragm* of a phone as used in radio. Excessive voltages if applied direct to the head-phones will often bend the diaphragms. Any such defect may seriously impair the operation of the part and when it is noted that a diaphragm is bent or warped through rough usage or the application of too great voltages, it is best to replace it immediately.

BUCKLING OF PLATES—During discharge of a *storage battery* the plates gradually expand, owing to the fact that lead sulphate has about twice the volume of the same quantity of lead peroxide. Should this expansion or discharge take place too quickly, the plates will bend or buckle. (See *Storage Battery.*)

BUNSEN BURNER—A form of gas burner frequently used in radio construction for heating the soldering iron where electricity is not available. By special arrangement of the parts, the gas is combined with the proper amount of air before it reaches the burning part. This is done by driving the gas through a tube with small holes drilled to allow the air to be sucked in. This method promotes complete combustion and gives a non-luminous flame that heats the soldering iron without

THE BROWN & SHARPE (B. & S.) COPPER WIRE GAUGE—Below is a comprehensive copper wire table, which is the standard one in use in the United States. The resistance and cross-sectional areas of various sizes of conductors are given, so that considerable calculating can be done with the data at hand. A mil means 1-1000 (one thousandth) of an inch. Hence a wire with a diameter of 9 thousandths of an inch could also be stated as having a diameter of 9 mils. The area of cross-section of the wire in circular mils is found by squaring the diameter in mils; thus the 9 mil conductor would have 9 times 9, or 81 circular mils area. This is the end area of the wire, of course, and is used in all electrical figuring. If the resistance per 1,000 feet of a certain size wire is, say, 50 ohms, then 500 feet of the same size wire would have a proportionately less amount of resistance, or ½ of 50 ohms, that is, 25 ohms. The resistance per foot is found by dividing the resistance in ohms per 1,000 feet by 1,000. This wire table is for bare copper wire. Conductors, according to the B. & S. standard gauge, halve their sectional area in circular mils (Cir. Mils) for every 3 gauge sizes smaller wire; and double their sectional area in Cir. Mils for every 3 gauge sizes increase.

Gauge Number.	Size.		Weight and Length.			Resistance.			Carrying Capacity, 2,000 Amperes p. sq. in. section.
	Diameter in Mils.	Square of Diameter or circular Mils.	Grains per Foot.	Pounds per 1000 Feet.	Feet per Pound.	Ohms per 1000~~0~~ Feet.	Feet per Ohm.	Ohms per Pound.	Amperes.
0000	460.000	211600.0	4477.2	639.60	1.564	.051	19929.7	.0000785	430
000	409.640	167804.9	3550.5	507.22	1.971	.063	15804.9	.000125	262
00	364.800	133079.0	2815.8	402.25	2.486	.080	12534.2	.000198	208
0	324.950	105592.5	2236.2	319.17	3.133	.101	9945.3	.000815	165
1	289.300	83694.49	1770.9	252.98	3.952	.127	7882.8	.000501	130
2	257.630	66373.22	1404.4	200.63	4.994	.160	6251.4	.000799	103
3	229.420	52633.53	1113.6	159.09	6.285	.202	4957.3	.001268	81
4	204.310	41742.57	883.2	126.17	7.925	.254	3931.6	.002016	65
5	181.940	33102.16	700.4	100.05	9.995	.321	3117.8	.003206	52
6	162.020	26250.48	555.4	79.34	12.604	.404	2472.4	.005098	41
7	144.280	20816.72	440.4	62.92	15.893	.509	1960.6	.008106	32
8	128.490	16509.68	349.3	49.90	20.040	.643	1555.0	.01289	26
9	114.430	13094.22	277.1	39.58	25.265	.811	1233.3	.02048	20
10	101.390	10381.57	219.7	31.38	31.867	1.023	977.8	.03259	16
11	90.742	8234.11	174.2	24.89	40.176	1.289	775.5	.05181	13
12	80.808	6529.93	138.2	19.74	50.659	1.626	615.02	.08237	10.2
13	71.961	5178.39	109.6	15.65	63.898	2.048	488.25	.13087	8.1
14	64.084	4106.75	86.87	12.41	80.580	2.585	386.80	.20830	6.4
15	57.068	3256.76	68.88	9.84	101.626	3.177	306.74	.33133	5.1
16	50.820	2582.67	54.67	7.81	128.041	4.582	243.25	.52638	4.0
17	45.257	2048.19	43.33	6.19	161.551	5.183	192.91	.83744	3.2
18	40.303	1624.33	34.37	4.91	203.666	6.536	152.99	1.3312	2.5
19	35.390	1252.45	26.50	3.786	264.136	8.477	117.96	2.2392	1.96
20	31.961	1021.51	21.60	3.086	324.045	10.394	96.21	3.3438	1.60
21	28.462	810.09	17.14	2.448	408.497	13.106	76.30	5.3539	1.28
22	25.347	642.47	13.59	1.942	514.933	16.525	60.51	8.5099	1.08
23	22.571	509.45	10.77	1.539	649.773	20.842	47.98	13.334	.80
24	20.100	404.01	8.55	1.221	819.001	26.284	38.05	21.524	.63
25	17.900	320.41	6.77	.967	1044.126	33.135	30.18	34.298	.50
26	15.940	254.08	5.38	.768	1302.083	41.789	23.93	54.410	.40
27	14.195	201.49	4.26	.608	1644.737	52.687	18.98	86.657	.31
28	12.641	159.79	3.39	.484	2066.116	66.445	15.05	137.283	.25
29	11.257	126.72	2.69	.384	2604.167	83.752	11.94	218.104	.20
30	10.025	100.50	2.11	.302	3311.258	105.641	9.466	349.805	.16
31	8.928	79.71	1.67	.239	4184.100	133.191	7.508	557.286	.13
32	7.950	63.20	1.33	.190	5263.158	168.011	5.952	884.267	.098
33	7.080	50.13	1.06	.151	6622.517	211.820	4.721	1402.78	.078
34	6.304	39.74	.847	.121	8264.463	267.165	3.743	2207.98	.062
35	5.614	31.52	.658	.094	10638.30	336.81	2.969	3583.12	.049
36	5.000	25.00	.525	.075	13333.33	424.65	2.355	5661.71	.039
37	4.453	19.83	.420	.060	16666.66	535.33	1.868	8922.20	.031
38	3.965	15.72	.315	.045	22222.22	675.22	1.481	15000.5	.025
39	3.531	12.47	.266	.038	26315.79	851.789	1.174	22415.5	.020
40	3.144	9.88	.210	.030	33333.33	1074.11	.931	35803.8	.015

depositing soot as in the case of an ordinary luminous flame.

BURSTYN, Dr. W.—Born in Austria, 1877. Educated at Vienna University. Noted as: Developer of the *quenched spark system* in radio transmission, upon which he worked in conjunction with Baron Lepel in 1907 to 1912.

BUS BAR—A single bar, usually copper, which serves as a common connector for a number of pieces of apparatus. Also referred to as Omnibus Bar.

BUS BAR WIRE—Square copper wire, usually tinned, much used in making radio connections. This wire is generally furnished in lengths of two or two and a half feet and is particularly efficient for radio work, as it offers a low *resistance* connection, makes a neat appearance, and is suitable for soldering at joints.

BUSHING — Pieces of insulating material, usually fibre or rubber composition, used to *insulate* parts of radio apparatus or in mounting machine screws or shafts of *condensers* and other devices. The term is also applied to brass or other metal bearings through which the shaft of a piece of apparatus is run.

BUTT JOINT—A method joining two lengths of wire together by placing them end to end or at right angles and soldering the joint. The illustration shows a common butt joint.

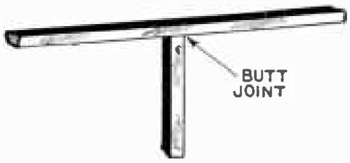

Showing how a Butt Joint is made.

BUZZER—An electric call signal device that makes a buzzing noise caused by the rapid vibrations of a contact breaker in a circuit using a small *battery* or *cell.* (See *Testing Circuit.*)

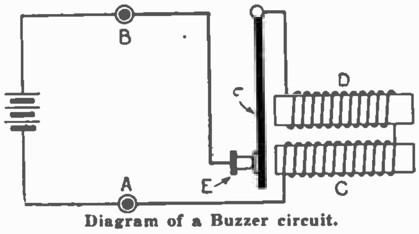

Diagram of a Buzzer circuit.

As the *current* flows from the posts A B the *electro-magnets* C and D are energized and draw down the vibrator spring F. This breaks the contact between E and F and as the current no longer flows the magnets release the spring F, which returns to its normal position and establishes the circuit once more. This making and breaking of the circuit takes place so rapidly that the vibrator makes a buzzing or humming sound, the speed with which it vibrates determining the pitch of the note. (See *Frequency,* also *vibration.*)

BUZZER EXCITER—A method of using a *buzzer* to produce local signals in a crystal receiving circuit to permit sensitive adjustment of the *detector.* The illustration shows a common method of producing signals to test the receiving set in this manner. In this case a small coil having a few turns of wire is placed in the buzzer circuit as indicated and a similar one is placed in the ground lead of the receiving set. When the button is depressed the buzzer is set in vibration and a transfer of energy takes place between L1 and L2. This causes an *oscillating* impulse (*q.v.*) to flow through the receiver circuit. The detector can then be adjusted to the most sensitive point at which it can be left until outside signals are heard. Another but less efficient method of exciting the circuit is to connect a short length of wire from the contact point of the buzzer to the ground connection of the receiving set. The former system is much more practical and furnishes better test signals.

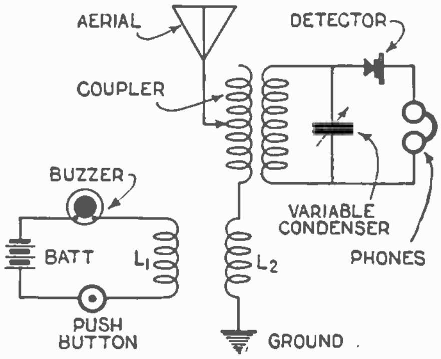

Buzzer Exciter.

Coil L1 being placed in *inductive relation* to coil L2, which is in the antenna circuit, undergoes a change of *flux* when the buzzer is *excited.* This sets up a *difference of potential* across coil L2 which charges the antenna ground system and produces oscillation at a definite *frequency* according to the adjustment. Now if the secondary circuit is tuned to resonance with the primary circuit, the signals will be audible in the phones when the detector is properly adjusted. Also called *Buzzer Wave Generator.*

BUZZER MODULATION—See *Modulation.*

BUZZER PRACTICE SET—A combination of a *buzzer* and signalling key arranged on a baseboard used for the purpose of practicing signalling.

B. W. G.—The abbreviation for Birmingham Wire Gauge—a system of gauging wire used in England. (See *B & S Gauge,* also *British Standard Wire Gauge.*)

BY-PASS CONDENSER — See *Condenser,* also *Filter.*

C

C—The chemical symbol for carbon. When printed in *italics,* it is also the symbol for *coulomb* in electrical practice. It is the international symbol for *capacity.*

CABLE—A number of wires stranded together, used to carry electrical currents, heavily insulated and often covered by a sheath or outer coating of lead, rubber, gutta percha, or braided silk or cotton. The term is used very often in a broad sense to signify any heavy wire or strand of wires that may be insulated more thoroughly than is ordinarily the case. (See *Conductor, Stranded Wire,* etc.)

CADMIUM—A chemical metallic element somewhat resembling zinc. In color, white with a bluish tinge. It is useful in fusible alloys because of its property for uniting readily with other metals, having a low melting point. Its electrical conductivity is about 25% of that of silver. It is very useful in the construction of a crystal detector, the molten metal holding the crystal firmly and as it melts at a low temperature the heat will not affect the sensitivity of the crystal. Cadmium is also used in making cells, notably the *Weston Cell.* (See *Weston Cell.*)

CAGE ANTENNA—An aerial having a number of wires arranged in cylindrical form, suspended either vertically or as a horizontal top. So named because of its resemblance to a cage. (See *Aerial.*)

CALCULATION OF CAPACITY—The mathematical or comparative determination of *capacity.* There are various formulas in use for calculating capacity under various conditions, such as for capacity of various types of condensers, or the capacity of a circuit. Rough approximations of capacity of condensers might be obtained by direct comparison with a known value. Other formulas are used for the calculation of antenna capacity. For determining the capacity of flat condensers, such as fixed condensers used for transmitting or receiving circuits, the following formula is generally used:

$$C = 0.0885\,K\,\frac{S}{r}$$

In this case, C is the capacity in micro-micro farads; to obtain result in the usual units of micro-farads, divide the result by one million. S is the surface area of one plate in square centimeters; r is the thickness of the dielectric in centimeters and K a constant depending on the material between the plates of the condenser. (See *Dielectric Constants.*)

CAL-ELECTRICITY — Electricity produced in the secondary of a transformer due to changes in temperature of the core. This current is in addition to the current normally induced in the secondary. (See *Transformer, Core,* also *Core Losses.*)

CALIBRATION — A process whereby instruments such as *galvanometers* *ammeters, wavemeters,* etc., are made to indicate certain values. Before such

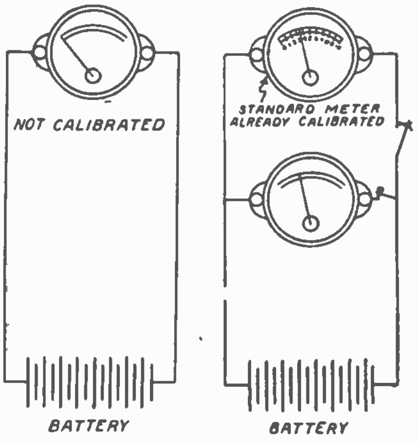

Fig. 1. Fig. 2.
Calibrating a Meter

instruments can be made use of, it is necessary to mark the dials in such a manner that the indicator showing on the dial or scale will have a definite meaning in terms of electrical values. Thus, a voltmeter before calibration is shown in Fig. 1. When it is placed across a battery, the pointer will move to some position on the white space, but without indicating any definite value. Now by testing with another meter which is already marked or calibrated in volts, we find, for example, that when the pointer is at the middle point in its swing, it indicates 10 volts. Similarly various other positions are found to indicate more or less voltage by comparison with the standard. This is shown in Fig. 2. It will be understood that one of the small switches must be open at all times. The reading of the calibrated meter is taken with the other meter out of the circuit (switch open), then the reading is marked on the other meter at the point indicated with the calibrated meter out of the circuit. Such a method would, of course, be inefficient and it is shown only to give a rough idea of the purpose and meaning of calibration. Another example of calibration is in the case of a dial on a receiver. Fig. 3 shows the usual form of dial used in tuning a receiving set. The dial is marked off to scale, from 1 to 100, the numbers having no particular significance. Now if we find that when the tuning dial is set at 20 it is adjusted for a wavelength of 260 meters, we can put that figure in place of the 20. The same procedure can be followed throughout the entire range of the dial, and when thus marked in terms of a definite quantity—wavelength in meters in this case—the dial is said to be calibrated.

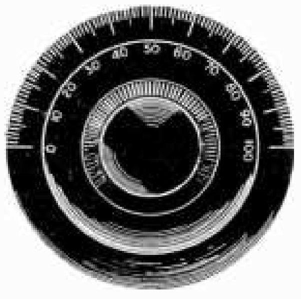

Fig. 3.

A dial as used on radio receiving sets may be Calibrated according to different wavelengths.

The actual method to be followed out in calibrating will depend on the equipment available and the accuracy desired. If a standard of comparison is available, it saves a great deal of work. Thus, a voltmeter can easily be calibrated by means of a standard meter, already calibrated, and a source of steady voltage controlled by a rheostat. An ammeter can be calibrated in the same manner by using a standard ammeter, and a wavemeter can be calibrated from another wavemeter of known values.

Calibration may be done in such a manner that the readings require mathematical solution to determine their value in terms of actual volts, amperes, meters and so on, or it may be *absolute*, in which case the readings are directly in terms of the quantity and require no solution. (See *Voltmeter Calibration, Wavemeter Calibration*, also *Rough Calibration.*)

CALIDO—An alloy containing nickel and chromium with a small percentage of iron. The melting point is very high, about 1550 degrees centigrade. Calido wire is much used as a *resistance*, particularly for heating devices. The chief characteristics are as follows: maximum working temperature 1100 degrees centigrade; microhms per cubic centimeter at 20 degrees centigrade, 100; temperature coefficient per degree centigrade, 0.00034; specific gravity, 8.2.

CALLAND CELL—A primary cell used in French telegraph work. It is a form of gravity cell having a negative electrode of copper and a positive electrode of amalgamated zinc, the electrolyte being zinc sulphate. Crystals of copper sulphate are used as a depolarizer. (See *Gravity Battery.*)

CALORIE—The unit of heat in the C. G. S. (Centimeter-Gram-Second) (q.v.) system. It is equivalent to the amount of heat necessary to raise the temperature of one gram of water from 0° to 1° centigrade. (See *Thermal*, also *C. G. S.*)

CALORIMETER—In electrical practice, an instrument used to measure the heat generated in a conductor carrying an electric current.

CAMBRIC, VARNISHED—Varnished muslin used as an insulating material. Also known as Empire Cloth. (See *Spaghetti.*)

CANAL RAYS—When an electric discharge takes place between the *anode* and a perforated *cathode* in a vacuum tube, fine pencils of light are seen to pass through the perforations in the cathode. These rays are called *canal rays*, and consist of positively charged particles. They produce phosphorescence on the wall of the tube.

Canal rays travel at a lesser velocity and in the opposite direction to the cathode rays. This fact and also the fact that the rays are deflected by powerful electric or magnetic fields in the opposite direction to cathode rays, is taken as proof that they consist of positively charged particles. (See *Cathode Rays.*)

CAPACITANCE—A term very often used as synonymous with *capacity*. Due to the fact that capacity may refer to the current carrying ability of a conductor and also to electrostatic capacity, it has been suggested that capacitance be used to refer only to *electrostatic capacity* of a body or device. Capacity would then refer to *current carrying ability.* (See *Capacity*, also *Electrostatic.*)

CAPACITATIVE REACTANCE—That part of the *reactance* of a circuit carrying alternating current which is due to the *capacity* in the circuit. (See *Reactance* also *Impedance.*)

CAPACITATIVE COUPLING—See *Coupling.*

CAPACITY—Generally speaking, the quantity of electricity in any form which a body is able to store or contain. The term is usually qualified to denote the particular case, such as *electrostatic capacity*, which is a measure of the ability of a condenser to store up energy in the form of electrostatic charges; also *current-carrying capacity* of a conductor, which is considered as the ability of a wire to carry a certain amount of electric current without overheating. The tendency in electrical literature is to use the term capacitance as applying to the electrostatic capacity of a condenser. Thus, a certain condenser may have a capacitance of 2 micro-farads. The unit of electrostatic capacity is the *farad.* This is understood as the capacity of a condenser that will store one *coulomb* of electricty under an electromotive force of one *volt.* When V is expressed in volts, C in farads, and Q in coulombs, $C = Q/V$.

CAPACITY CIRCUIT—An electrical circuit in which the *capacity* is very large compared with the *inductance.* The inductance may be considered negligible.

CAPACITY CONVERSION FACTORS—In the calculation of *capacity*, as the capacity of a condenser, there are several different units. For instance, a *farad* is a large unit of capacity and is equal to one million *microfarads*, or conversely, a microfarad is equal to one millionth of a farad. In the same way a farad is equal to one billionth of an *abfarad.* The four units used in capacity measurement are the *farad, micro-farad, micromicrofarad, statfarad* and *abfarad.* The following table gives the conversion factors for these various units:

$$1 \text{ abfarad} = \begin{cases} 10^{9} \text{ farads} \\ 10^{15} \text{ mfds.} \\ 9 \times 10^{20} \text{ stfds.} \end{cases}$$

$$1 \text{ farad} = \begin{cases} 10^{-9} \text{ abfd.} \\ 10^{6} \text{ mfds.} \\ 9 \times 10^{11} \text{ stfds.} \end{cases}$$

$$1 \text{ mfd.} = \begin{cases} 10^{-15} \text{ abfd.} \\ 10^{-6} \text{ farads} \\ 9 \times 10^{5} \text{ stfds.} \end{cases}$$

$$1 \text{ stfd.} = \begin{cases} 1/9 \times 10^{-20} \text{ abfd.} \\ 1/9 \times 10^{-11} \text{ farad} \\ 1/9 \times 10^{-5} \text{ mfd.} \end{cases}$$

$$1 \text{ mmfd.} = 10^{-6} \text{ mfd.}$$

In the table above, farad is abbreviated Fd.; abfarad is shown as abfd.; microfarad is mfd., micromicrofarad is mmfd. and statfarad is stfd. In dealing with large figures containing many ciphers the amount is shown as 1,000,000 is 10^6 and $\frac{1}{1{,}000{,}000}$ is 10^{-6}, etc. (See *Capacity, Farad*, also *Unit.*)

CAPACITY EARTH or CAPACITY GROUND—A substitute for the usual ground connection where the wires or plates are buried beneath the ground or attached to a water or steam pipe. It is in effect a second aerial, and is placed either under the regular aerial or to one side. It is more used in radio transmission than in reception. (See *Counterpoise.*)

CAPACITY ELECTROSTATIC—See *Capacity* also *Electrostatic Capacity.*

CAPACITY FREQUENCY FACTOR—The relation between the apparent capacity of a condenser and the *electrostatic capacity.* (See *Capacity, Condenser* also *Electrostatic.*)

CAPACITY OF CONDENSERS IN PARALLEL—When several condensers are connected in parallel the resultant capacity is the sum of the individual capacities. It is written:

$$\text{Capacity} = C_1 + C_2 + C_3, \text{etc.}$$

It will be evident that the resultant capacity of condensers connected in parallel is just the reverse of case for resistances, where the total resistance is less than that of the smallest individual resistance in the parallel connection. (See *Resistance Measurement* also *Capacity of Condenser in Series*, also *Condensers in Series* and *Condensers in Parallel.*)

CAPACITY OF CONDENSERS IN SERIES—When several condensers are connected in series with each other the resultant capacity is always less than the capacity of the smallest condenser in the group. The capacity of condensers in series is given as follows:

$$\text{Capacity} = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}, \text{ etc.}}$$

The resultant capacity is thus the reciprocal of the sum of the reciprocals of the several capacities. (See *Condenser, In Series.*)

CAPACITY MEASUREMENT OF ANTENNA—A simple method for the measurement of the capacity of an antenna as given by Austin follows: The illustrations show the general arrangement. The inductance L in Fig. A is a coil that will increase the wavelength (natural) of the antenna system 4 or 5 times; C is a calibrated variable condenser and W a standard wavemeter. In operation the coil L is connected in the antenna circuit, the buzzer is used to excite the antenna circuit and the wavelength is measured by w. The coil L is then taken out of the antenna circuit and shunted by the condenser C. The condenser C is then varied until the wavelength of the combined coil L and condenser C is the same as that obtained before with the coil in the antenna circuit but without the variable condenser. The capacity of the condenser now will be approximately that of the antenna, ignoring the distributed inductance of the antenna system. Fig. A shows the first operation, the measurement of the antenna with the coil in the circuit, and Fig. B shows the measurement of the wavelength of the coil and condenser. (See *Mutual Capacity, Wavemeter, Natural Wavelength,* also *Distributed Capacity.*)

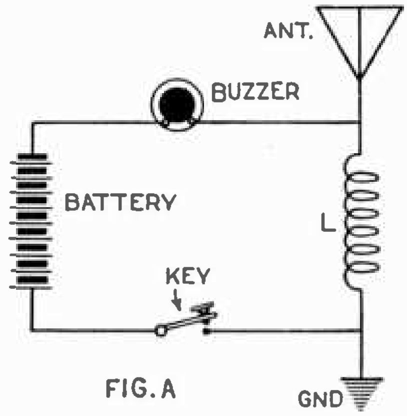

FIG. A

The first operation in the measurement of the antenna. The wavemeter is shown below.

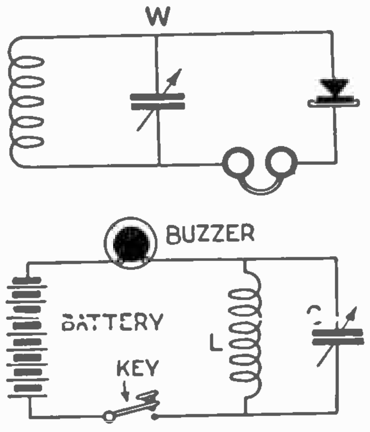

FIG. B

Circuit for the measurement of wavelength of coil and condenser.

CAPACITY MUTUAL—See *Mutual Capacity.*

CAPACITY RESISTANCE—The resistance to alternating currents offered by a body possessing electric capacity. (See *Capacity,* also *Resistance.*)

CAPACITY SWITCH—Any switch used in a circuit to introduce or cut out capacity (condenser).

CAPACITY (UNIT OF)—The unit of capacity is the farad. A condenser has a capacity of 1 farad when 1 coulomb is required to raise its potential from zero to 1 volt. Since the farad is very large, its millionth part, or the microfarad is generally used.

CARBON—A non-metallic chemical element used for many purposes in electricity and radio. It exists in three forms, two crystalline and one non-crystalline. The first two are diamond and graphite, the third charcoal. Carbon is used as an electrode in electric arcs, as one of the poles in certain types of primary cells, as a material in place of wire for various forms of resistance, and as brushes in motors and generators. (See *Diaphragm, Carbon.*)

CARBON RHEOSTAT—A rheostat or variable resistance using carbon in place of the customary wire. In its simplest form a carbon rheostat may be merely a piece of carbon having a sliding contact. This is shown in Fig. 1. Owing to the relatively high resistance of carbon used in this manner, such a controlling resistance will generally be termed a *potentiometer* (q.v.) and used as such.

The most practical form of a carbon rheostat is known as a carbon pile rheostat. This type uses a number of flat pieces of carbon of relatively low resistance placed together and arranged with a knob and screw in such manner that the pressure can be varied. This change in pressure alters the resistance of the mass and allows control of the current for vacuum tubes. The same principle is used for very high resistances as well as for low values. The principle will be seen from the illustration Fig. 2.

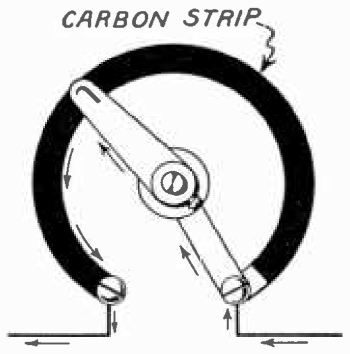

Fig. 1.—A carbon rheostat with a rotary sliding contact.

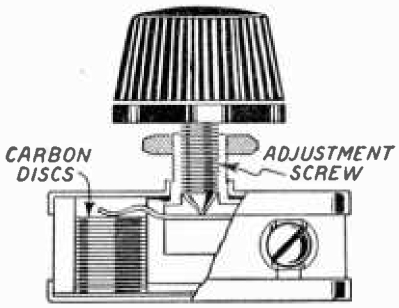

Fig. 2.—A carbon pressure type rheostat.

CARBORUNDUM—An abrasive material, a product of the electric furnace, composed of silica and carbon. It is used in radio as a crystal rectifier, although the strides of radio have practically rendered it obsolete, many other better crystals being in use. The chief difficulty with its use as a rectifier or crystal detector was the fact that for efficient operation it required a cell or battery and potentiometer to regulate the current flowing through it. The simple crystal circuit using a carborundum detector is shown in the illustration. C is a tuning coil, D the carborundum crystal and contact, P a potentiometer and T the head phones. (See *Rectifier,* also *Detector.*)

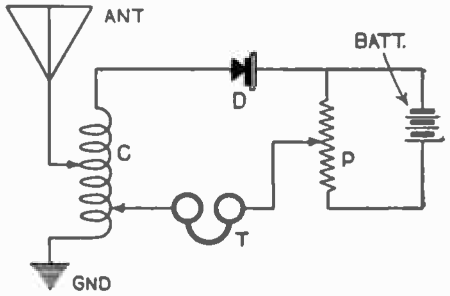

Diagram of a radio circuit employing a carborundum crystal detector.

CARDBOARD TUBING—Tubing made of laminations of paper pasted together. They are used for home made coils of all descriptions, and while not as efficient as the composition tubing, can be used for many purposes without any great loss. As a general rule it is better to use some of the numerous compositions, such as bakelite or rubber compounds.

CARPENTIER, JULES—French radio expert. Born in Paris, in 1854, he joined the Ecole Polytechnique in 1871, and in 1876 was appointed principal stores engineer of the Lyons Railway Company, making a special study of electricity. For his work in electricity he obtained in 1881 the cross of Chevalier of the Legion of Honor. One of the early pioneers of radio in France, he founded the Compagnie Generale Radiotelegraphique, afterwards absorbed in the Compagnie Generale de Telegraphie sans Fil. Carpentier is a member of the Academie des Sciences, Commander of the Legion of Honor, and President of many scientific societies.

CARRIER WAVE—The *continuous wave* (q.v.) of radio frequency generally thought of as carrying the voice or music waves from the radio broadcasting station. Actually it is a radio wave of high frequency, which is altered in amplitude by the music or speech transmitted. In action, this continuous wave has another wave super-imposed on it. This other wave having wave form and amplitudes determined by the voice or other sounds being transmitted. The illustration shows the carrier wave, the waves representing speech or music and the combination of the two. This effect of varying the amplitude or strength of a continuous wave by means of some sort of program is known as *modulation.* (q.v.) The wave representing the speech or music is imposed on the carrier wave resulting in the combined wave which reaches the listener and is translated by suitable apparatus into approximately the same sounds that are transmitted. (See *Modulation, Broadcasting*, also *Speech Vibrations*.)

CARRYING CAPACITY—See *Current Carrying Capacity.*

CARTRIDGE FUSE—A fuse in which the fuse wire is surrounded by some non-inflammable substance, enclosed in a cartridge-like tube and having brass lugs soldered to caps at ends. Used to prevent the hot wire from "flying" when fused. (See *Fuse.*)

CASCADE—A term applied to pieces of apparatus connected together in series, particularly vacuum tubes. In this case, the arrangement would be such that each vacuum tube would amplify the signal output of the preceding tube. The tubes will be so connected that the output of one is introduced as input to the next and so on the total result being greatly increased signal strength from the output of the last tube. This term may apply to either radio frequency or audio frequency amplification. (See *Amplifier, Radio Frequency*, also *Audio Frequency*.) A simple cascade arrangement is shown in the illustration. Fig. 1 is a circuit with two stages of cascade radio frequency amplification, and Fig. 2 a circuit using two stages of audio frequency amplification with the usual detector tube in each case. A combination of both radio frequency and audio frequency stages of amplification might also be used, the term cascade amplification also applying here.

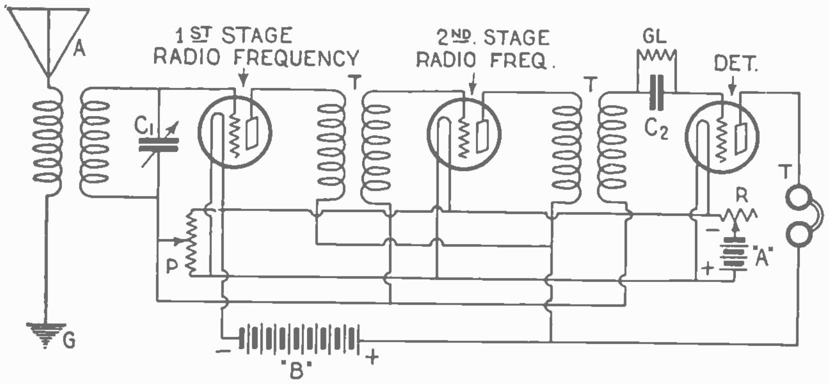

Fig. 1.—Circuit diagram showing two stages of cascade radio frequency amplification and detector.

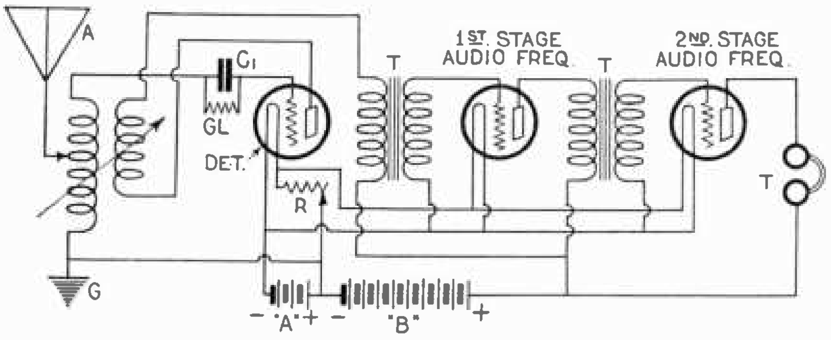

Fig. 2.—Diagram of detector and two stage cascade audio frequency amplifier.

CATHION or CATION—The charged particles which appear at the cathode or move toward it through the electrolyte of an electrolytic cell. (See *Anode.*)

CATHODE or KATHODE—A *negative electrode.* The term is often applied to the filament of a vacuum tube to distinguish it from the anode or plate. (See *Cell, Electrolysis*, also *Anode.*)

CATHODE RAY TUBE—A vacuum tube having a high vacuum, that is from which practically all air has been exhausted, used in the production of cathode rays. These tubes in their modern form are used for many purposes in studying the nature and form of electric waves. The illustration shows one of the latest types of tube, known as the *Cathode Ray Oscillograph.* This tube is used for the following purposes: examination of the wave forms of various rectifiers, study of wave forms of different types of generators, study of vacuum tube characteristics, examination of the characteristics of X-ray and other types of tubes, examination of radio frequency waves modulated as in broadcasting, measurement of hysteresis and dielectric loss in various materials, study of phenomena in arcs and sparks, together with numerous other associated effects. The development of this tube is one of the marvels of science, enabling the research engineer to virtually see the electric wave, and to study exhaustively, electric phenomena otherwise invisible.

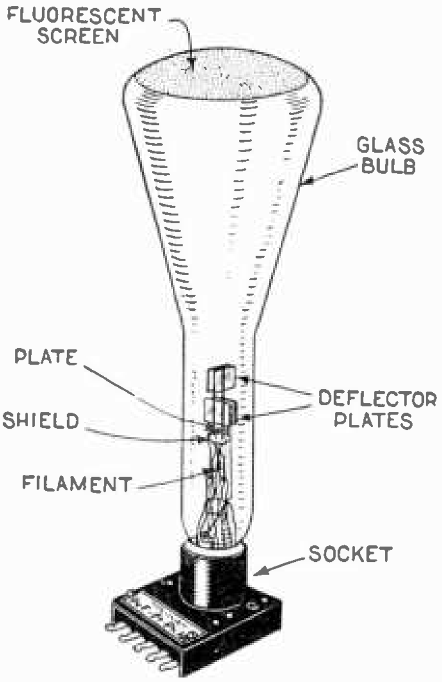

A cathode ray tube with socket.

The Western Electric modification of the Braun Cathode ray tube or oscillograph, is a three electrode glass tube about 30 centimeters (about twelve inches) long, in the form of a cylinder an inch and three-quarters wide, spreading out in conical shape, the tube being about three and a half inches wide at the top. The *cathode* is a filament coated with active oxides and arranged to emit the number of electrons required for the cathode rays, at a dull red heat. The *anode* or plate is a small platinum tube set very near the filament. Between the filament and the anodes is a small circular screen having a hole just a little smaller than the circular filament. A battery of several hundred volts is connected between the filament and plates with the positive terminal connected to the plate. The electrons emitted from the hot filament are controlled by this field in the same manner as in an ordinary vacuum tube. A small portion of these electrons pass completely through the tubular anode and constitute the *cathode* rays. These rays are passed between two pairs of deflecting plates set at right angles to each other, and fly against the large end of the tube. This end is covered with a fluorescent mixture which renders the rays visible. These rays are deflected in various ways. For example, if an alternating current is applied the visible spot on the screen will be drawn out into a line. (See *Electron, Vacuum Tube*, also *Wave Analysis.*)

CATHODE RAYS—The stream of electrons or electrical particles sent out from the *cathode* or filament of a vacuum tube. These rays are negatively charged. (See *Cathode, Electrons, Vacuum Tube* also *Oscillograph.*)

CAT-WHISKER—The fine wire used with certain crystal detectors to make contact with the crystal. Usually a springy metal such as phosphor bronze wire.

"C" BATTERY—One or more small cells having a voltage generally between 2 and 10 volts, used in the grid

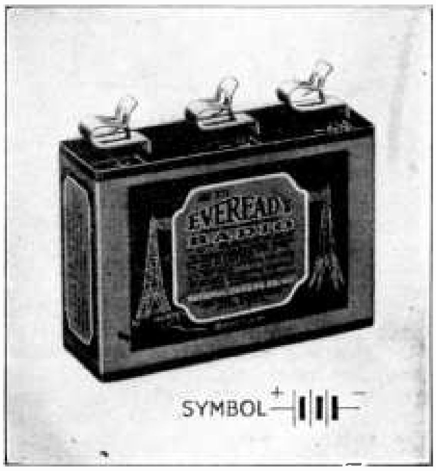

Fig. 1.—A 4½ volt "C" battery. This type of battery has a common positive terminal and is tapped at negative terminals for 3 and 4½ volts.

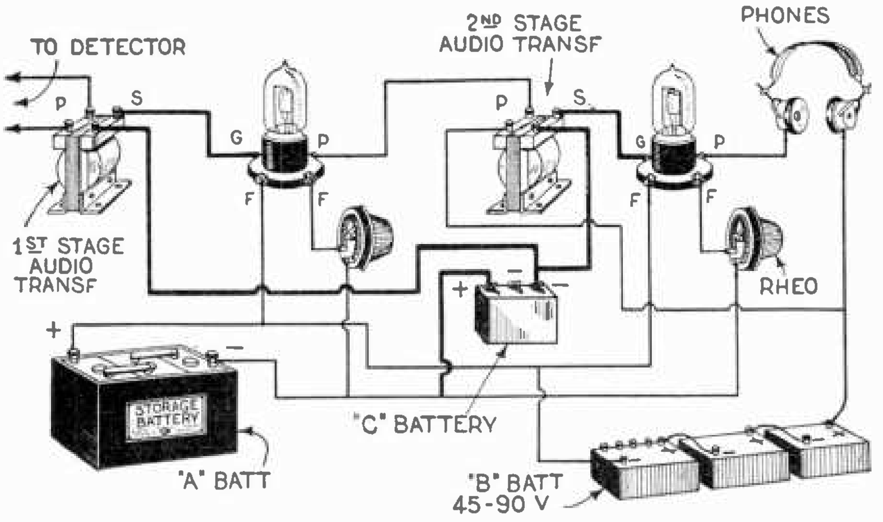

Fig. 2.—A diagram showing how the "C" battery is connected in the audio frequency amplifier.

circuit of an amplifier tube for the purpose of impressing a negative potential on the grid of the tube. The method of connecting it in the circuit is shown in Fig. 2. Fig. 1 shows a typical 4½ volt "C" battery. (See also *Grid Bias* and *"C" Battery Curve.*)

"C" BATTERY CURVE—In order to show the effect of a "C" battery on the consumption of current from the "B" battery in an audio amplifier, a *curve* (q.v.) may be drawn as in the illustration. In this case the test was made with a type UV 201A tube used as a one stage *audio frequency amplifier* in a conventional circuit. It is readily apparent on referring to the lines A and B that when the "C" battery is in use the consumption of current by the plate is much less than without it.

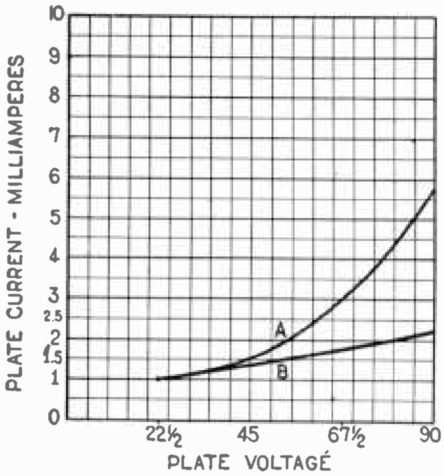

**Test conducted on UV201A tube using single stage audio amplifier with conventional circuit.
Plate Voltage—Abscissae.
Plate Current in milliamperes—ordinates.
A—Consumption curve without "C" Battery.
B—Consumption curve with "C" Battery.
(Note: 1.5 ma. at 45v plate rising to 5.75 ma. at 90v plate without "C". With "C" Battery consumption maximum is 2¼ ma. at 90v.)**

CELL—A cell is one of the four chief sources of electric energy. In a cell the energy is created by an *electrochemical* process, using in its simple form two unlike metals immersed in a dilute acid or alkaline solution. In the illustration Fig. 1 the common form of electric cell is shown. C is a strip of *carbon* and Z a strip of *zinc*. These are placed in the jar in a conducting solution of sal ammoniac. If the exposed terminals from the carbon and zinc strips are connected together by a piece of wire, current will flow from one side of the cell to the other through the wire. The current flows from the carbon or *positive* pole to the *negative* or zinc pole through the wire, completing the circuit through the solution from zinc to carbon. (See *Current, Direction of Flow.*) The cell described is known as a *primary cell.*

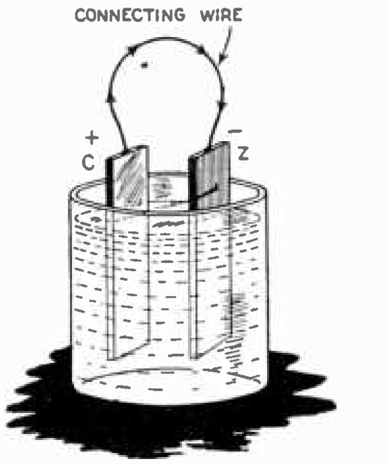

Fig. 1.—Common form of electric cell.

In Fig. 2 a simple *storage cell* or *secondary cell* is shown. In this case there are two lead plates immersed in a dilute solution of sulphuric acid. A source of primary current (primary cell) may be connected to the secondary cell in the manner shown in Fig. 3 and after being connected for a certain length of time it will be found that the energy of the primary cell has been transferred to the secondary cell and stored in it. This process of placing electric charges in a secondary cell is known as *charging*, and the current stored up in it is known as the *charge*. Now, if the wire is connected from one terminal to the other it will be found that electric energy can be withdrawn in a similar manner to that of the primary cell. A secondary cell may contain several plates and several of these cells may be connected together to form a storage battery. (See *Storage Battery, Current, Production of,* also *Anode* and *Cathode.*)

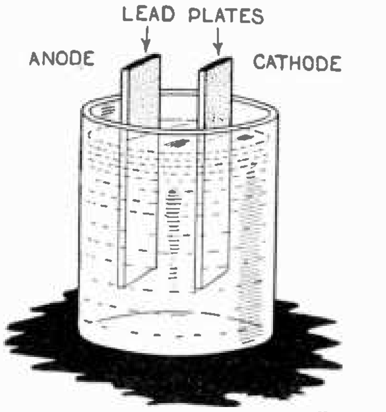

Fig. 2.—Storage or secondary cell.

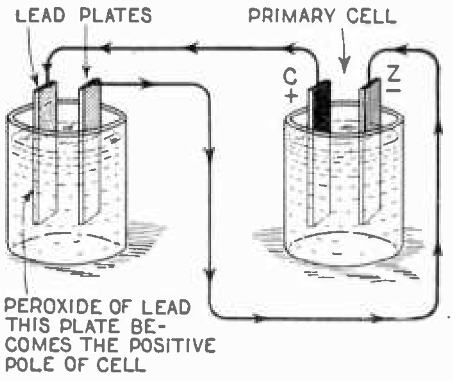
LEAD PLATES
PRIMARY CELL
C
+
Z
-
PEROXIDE OF LEAD THIS PLATE BECOMES THE POSITIVE POLE OF CELL

Fig. 3.—Primary cell connected to secondary or storage cell.

CELL, SECONDARY—When a *primary cell* is connected to a *storage* or *secondary cell*, the current from the primary cell flows into the secondary cell and in effect deposits energy in that cell. Then when the charging source (the primary cell) is withdrawn, the secondary cell can be used in practically the same manner as the primary cell. A wire connected from one terminal to the other will permit passage of current. The action of storing up energy in a storage cell is approximately as follows: When the charging current flows from plate to plate through the solution in the secondary or storage cell, the plate connected to the *positive* (+) *pole* (carbon or copper) of the primary cell, receives a brown coating of *peroxide of lead.* At the same time the other plate becomes spongy or porous. Now as one plate has received a coating while the other remains unchanged as far as its surface is concerned—it acts as a primary cell if the charging source is taken away and the two terminals connected. It now has all the essentials of an ordinary chemical cell. As long as the coating of peroxide of lead remains on the surface of one of the plates, the cell will be capable of delivering current.

When the coating has been worn off or eaten away the cell is said to be in a discharged condition. It will then be necessary to go through the same process of connecting a primary source to the terminals in order to again deposit peroxide of lead on one plate. It will therefore be evident that when the *storage cell* is charged, *current* is not actually stored in it, although in effect it has received potential energy. What has actually occurred is that the current supplied to the cell during the charging process has produced electrochemical changes, making the plates dissimilar and thus producing a difference of potential. (See *Current, Production of*). The *voltage* or *electromotive force* of primary cells varies from .06 to 1.5 volts according to the nature of the elements used and the grade of the *electrolyte* (q.v.). Secondary or *storage cells* of the lead plate type produce from 2.1 to 2.6 volts. (For more complete explanation of the theory and action of a secondary cell, see *Storage Battery*. See also *Electro Motive Force, Current* and *Voltage*.)

CELL, THEORY OF PRIMARY—It has been explained under the heading *Cell*, that a primary cell produces electrical energy by chemical action. More specifically the action of a simple cell is as follows: When the copper and zinc, or carbon and zinc placed in a cell are connected together by a conductor, (wire or other conducting object) the acid in the case of a sulphuric acid solution, attacks the surface of the zinc plate and forms a compound known as zinc sulphate. While this sulphate is being formed some of the hydrogen contained in the sulphuric acid is liberated in the form of bubbles which appear on the copper plate. Some of these bubbles escape into the surrounding air by rising to the surface of the solution, but others cling to the surface of the copper plate and gradually cover it with a film of hydrogen. It may be said here that the decomposition of the zinc plate in the acid solution furnishes the *electro-motive force* required to cause the flow of current between the plates, through the solution and through the conductor. Now, hydrogen is a non-conductor, and as the hydrogen gradually covers the copper plate, the surface of the plate in contact with the solution gradually decreases. The flow of *current* from zinc plate to copper plate in the solution is thus gradually diminished, until eventally we can imagine the copper plate as being entirely covered with hydrogen and no longer in contact with the solution. In addition to this insulating action of the hydrogen, it tends to set up a current within the cell in the opposite direction to the normal flow of current produced by the chemical action. A cell in this condition is said to be "polarized," and various means have been used to cut down or eliminate the *polarization effect*. The purpose of these various means is to prevent the hydrogen bubbles clinging to the surface of the copper plate, thus allowing them to escape to the top of the solution and thence into the surrounding air. (*See Polarization, Current, Voltage, also Storage Cells.*)

CENTIMETER—A measure of length in the metric system. It is one hundredth part of a meter, or approximately .3937 inch. It is used as the unit of length in the Centimeter Gram Second System (q.v.).

CENTIMETER GRAM SECOND or C. G. S. A system of units of measure employed in practically every branch of engineering, particularly in electrical practice. It is a decimal system, the centimeter being the unit of length, the gram the unit of weight and the second the unit of time. All other units are derived from these three. When the magnitude of a quantity is spoken of, it refers to the relative magnitude of the quantity when compared with some other quantity of the same nature. The magnitude used as a standard of comparison is termed a unit. Thus, if the length of a body is measured in feet, as for example, ten feet, it means that the magnitude of the body is ten times unity or ten times one foot. Similarly a *current* of ten *amperes*, is a current of magnitude ten times as great as the unit quantity—one ampere. Now an ampere is the practical unit of rate of flow of electric current. The unit has a certain definite value. (See Ampere.) The magnitude of any other similar quantity measured in amperes will mean a value equal to a certain number of units. In addition to the C. G. S. system of units, there is a practical system. This is commonly used for ordinary calculations. Under this system, the following are the main units: The *volt* is the practical unit of *electro-motive force*;—the force required to maintain a flow of current of one ampere through a resistance of one ohm. The unit of *current strength* or rate of flow of current is the *ampere*;—the strength of current maintained by a force of one volt through a resistance of one *ohm*. The unit of *resistance* is one *ohm*, and is the resistance of a *conductor* or *circuit* that will permit the passage of a current of one ampere under the force of one volt. The unit of *quantity* is the *coulomb*, which is the quantity of current flowing in a circuit when one ampere passes a given point during one second of time. The *watt* is the practical unit of *electrical power*, being the power of a current of one ampere flowing in a circuit under the pressure of one volt.

It will thus appear that the practical system is actually based on the C. G. S. system. Calculations of an involved nature necessarily make use of the C. G. S. system, and when multiples or sub-multiples are used they are expressed in Greek and Latin prefixes. For example, a million ohms is 10^6 ohms and is expressed as a *meg-ohm*. "Meg" being derived from *mega millian*. Similarly a millionth of an ampere is expressed as a *micro-ampere* or 10^{-6} AMPERE. The various units under the C. G. S. and practical systems are taken up in detail under the respective headings, as *JOULE. ERG, DYNE* and so forth. (See also *Inductance (Unit of)*; *Capacity (Unit of)*; *Electrostatic Units and Electromagnetic Units.*)

C. E. M. F.—The abbreviation for *counter electromotive force* (q.v.).

C. G. S. UNIT OF CURRENT—The absolute *electromagnetic unit of current*. It is the current which when passed through a conductor bent into a circle of one centimeter radius will attract or repel a unit magnetic pole placed at its center, with a force of one *dyne* (q.v.).

CERUSITE—A kind of mineral used as a rectifier in some types of *crystal detectors*. It is a carbonate of lead having a whitish-grey color. This mineral is not as good a rectifier as some of the others, but has the advantage of more uniform sensitivity over practically its entire surface. (See *Crystal Detectors.*)

CHALCOPYRITE—A crystal used in a *pericon detector*, which is composed of zincite and chalcopyrite in contact with each other. It is a combination detector suitable for use where it is subject to jarring, since it has nearly uniform characteristics over its entire surface. It works without the aid of an electric battery.

CHARACTERISTICS—A term broadly applied in electricity to refer to any distinguishing feature in electrical apparatus. In radio usage the term is generally applied to *vacuum tubes*. The characteristic of a tube is understood as the relation between the potential (voltage) at the *grid* (q.v.) and the resulting current obtained from the *plate*. (q.v.) It may also occasionally refer to other features of a tube, such as its current requirements or amplifying quality as distinct from the ratio of *grid voltage* to *plate current*. (See *Characteristic Curve, Vacuum Tube Characteristics; also Plate Current.*)

CHARACTERISTIC CURVE—A diagram showing graphically in the form of a curved line the relation of changing values. Thus, in the case of a *vacuum tube*, a change in the *grid* voltage produces certain changes in the *plate* current (q.v.) and these changes may be plotted as curves. The temperature of a wire carrying current increases with increase in the *current* passing through it, and this increase may be shown by a curve. (See *Vacuum Tube Characteristics.*)

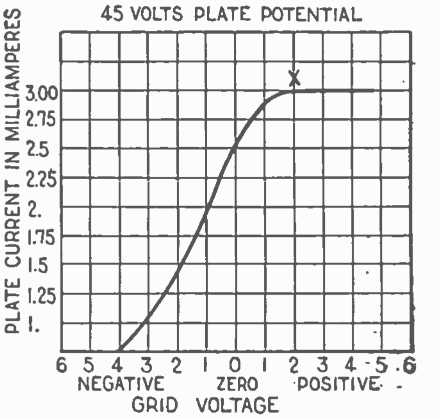

Characteristic chart of vacuum tube grid and plate voltage.

CHARGE—The presence of electricity in a body. Is usually defined as an excess or deficit of *electrons* (q.v.) on an insulated body. In the case of a *condenser* (q.v.) after it has accumulated electrical energy it is said to be in a charged condition, or to have an *electro-static charge*. The charge in a *condenser* is measured in *coulombs*. (q.v.) A *storage battery* is said to contain a charge when it has been connected to a source of direct current for a length of time, and the hydrometer test shows the electrolyte at proper specific gravity.

CHARGE, SPACE—See *Vacuum Tubes*, also *Space Charge.*

CHARGED BODY—Any body that has accumulated electrical energy. (See *Charge.*)

CHARGER, STORAGE BATTERY—Any means of controlling the current delivered from a primary source for the purpose of charging *storage batteries* or other *secondary cells*. Chargers are made in many forms and are designed for either direct or alternating current. Direct current chargers consist of certain resistances, with the necessary meters and switches. A simple form of direct current charger is shown in the illustration. In constructing a direct current charger there are several important details that must be

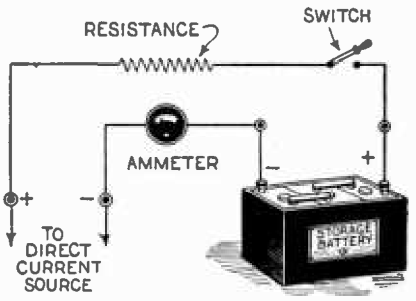

Charging storage battery from direct current source.

taken into consideration. The plates of a storage battery require certain charging rates in amperes, depending on the size and number of the plates. That is to say, the manufacturer specifies that when charging, a certain number of amperes must be passed to the battery from the charging source for a definite number of hours. This figure may be decreased but not increased. As an example, some storage batteries will take a charge of five amperes without injury to the plates, which rate would be ruinous for other batteries. Now the question of obtaining the proper amount of charging current arises. The majority of house lighting systems have a voltage of from 110 to 125 volts. (Reference is now made only to direct current systems, alternating current chargers being taken up later on.) According to *Ohm's Law* (q.v.) the current flowing in a circuit is directly proportional to the *voltage* and inversely proportional to the *resistance*. Voltage is the product of resistance times the current, the resistance in ohms and the current in amperes. Therefore if we desire a current of five amperes from a source of 110 volts, it is necessary to have a resistance of a value in ohms which, when multiplied by the current in amperes,—five in this case—will produce 110. Conversely, we can divide the voltage—110 by the current—5 to obtain the correct resistance. This would indicate a resistance of 22 ohms. There is another factor which enters into the problem however, the back pressure, or Counter Electromotive Force exerted on the charging source by the battery. In the case of a six volt storage battery, this will be six volts. This must be subtracted from the voltage of the charging source to obtain the effective value. Thus, the effective value in this case would be 110 — 6 or 104 volts. Then as explained before, dividing the voltage by the current required, we find that it is 104 ÷ 5 or approximately 21 ohms. Then by placing a resistance of 21 ohms in the circuit, as shown in the illustration, and connecting the battery as indicated, a charge of five amperes can be applied to the battery. The same result can be obtained by placing an ordinary electric lamp in series in place of the resistance. (See *Lamp Bank.*) The *internal resistance* of the battery is very small compared to that of the line resistance, and can be ignored. A point to remember is that the voltage of the charging source must always be higher than that of the storage battery, because, as explained above, the back or *counter electro motive force* of the battery, must be subtracted from the voltage of the charging source to indicate the effective voltage. Thus, if the battery exerts a counter electromotive force of six volts and the charging source is only six volts the counter force would be equal to the charging force and there would be no charging current. When *direct current* is not available and *alternating current* is used to charge the battery, it will be necessary to change it to direct current. For this purpose it is essential that some form of *rectifier* be used. This rectifier may be of a chemical, mechanical or electrical type. These various forms will be taken up under their separate headings. (See *Rectifier, Balkite, Tungar, Rectigon, Electrolytic Rectifier, Tube Rectifier.*)

CHARGING—The act of expending electrical energy to place a charge in or on a battery or body. The term is used generally to indicate the process of impressing electrical energy on the *storage battery* elements for the purpose of making the battery a secondary source of electricty. (See *Charger, Chemical Rectifier* also *Tube Rectifier.*)

CHARGING BOARD—A term used synonymously with *charging panel* to denote an arrangement for controlling the current used for charging *storage batteries*. It generally consists of an insulated panel with various *switches, fuses* and *resistances* thereon to limit the amount of current delivered, and at the same time furnish a ready means of connecting and disconnecting the various circuits connected through it. The fuses act as safety devices and break the circuit when too great a current is delivered. (See *Fuse, Switch, Charger.*)

CHEMICAL RECTIFIER—Device for changing alternating current to direct current. (*See Electrolytic Rectifier.*)

CHLORIDE BATTERY A type of storage battery for heavy duty operation. The positive plate is made of lead with a number of holes pierced in it and each filled with a coil of pure lead. This coil is forced into the holes during manufacture and is said to enable the battery to withstand heavy discharge without damage to the plates. (See *Storage Battery.*)

CHLORIDE CELL—A cell (q.v.) which uses in most cases, a positive plate having lead as the active material, the negative plate being made of metallic zinc. A solution of chloride of zinc is used, from which the name is derived. (See *Chloride Battery.*)

CHOKE OR CHOKING—The action of a device in a circuit wherein it tends to choke, or hold back, certain forms of current or certain *frequencies* while permitting others to pass freely, or to oppose fluctuations in the strength of the current passing through it in the circuit.

CHOKE COIL—A coil of wire possessing considerable *self-inductance* and relatively little *resistance*. These coils are used in various ways and for a variety of purposes. When connected in a direct current circuit, and possessing the proper value, they have a tendency to prevent any fluctuations in the current and thus keep the current smooth. An example of its use is with a *loudspeaker*. A *condenser* will permit passage of *high frequency* (*alternating*) *currents* such as we find in radio reception, but will effectively block direct currents. On the other hand a choke coil will pass direct current without offering any appreciable resistance, but will retard the *modulated currents* (q.v.) It will be apparent then that by a proper combination of the two (Choke Coil and Condenser) as in the illustration, the direct currents can be shunted around the loud speaker and thus prevent injury to it, while the necessary modulated currents are allowed to pass through without difficulty. Another use for the Choke is in a circuit carrying alternating currents. In this case the choke will limit the current in the circuit without loss, whereas an ordinary resistance, while it would limit the current, would dissipate a certain amount of energy. Chokes may be made with iron cores of the open or closed type, for *low frequency currents*, or may only have air for the core when used in *high frequency* circuits. The closed core type is the more common form. The formula most commonly used in determining the inductance of a choke coil in *henries* is as follows:

$$L = \frac{1.257 \times \mu \times A \times N^2 \times 10^{-8}}{l}$$

Where L is the inductance in henries.
μ is the permeability. (This varies from about 1000 to 2000 in most cases).
A is the effective area of iron cross section (square centimeters).
N is the number of turns of wire.
l is the length of iron path (in centimeters).

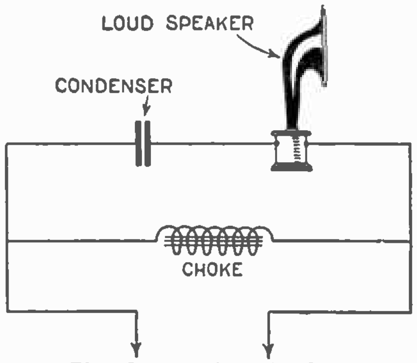

Method of using choke coil in connection with loud speaker.

CHOKE, HIGH FREQUENCY — A choke coil of high *self-inductance* (q.v.) used in transmitting circuits to prevent puncture of the secondary windings of the *transformer* (q.v.). When the condensers in a transmitting circuit are charged or discharged, a back pressure is exerted on the windings of the transformer supplying the high voltage. The illustration shows the position of these choke coils in the transmitting circuit. (See *Back Oscillation.*)

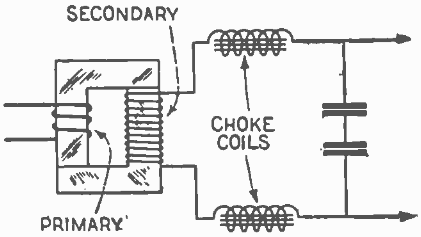

Position of choke coils in transmitting circuit.

CHOPPER—A device used in the aerial circuit of a *continuous wave* (q.v.) transmitter to break up the continuous trains of waves into separate groups and thus permit them to become audible. In the case of reception of ordinary *damped waves* (q.v.) the head phone produces signals that have a frequency of vibration equal to the number of wave trains per second. When the transmitted waves are con-

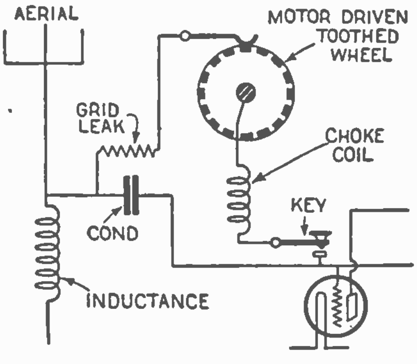

Showing how chopper is used to interrupt signals from continuous wave transmitter.

tinuous there is obviously no interval, that is, there are no groups of waves and they are therefore inaudible in the usual receiving apparatus. The *chopper* is only necessary when *crystal detectors* are used for reception. (See *Ticker, Beat Reception and Continuous Wave Transmission.*)

CIRCUIT—The path followed by an electric current passing from its source through a wire conductor or series of such *conductors*, and back again to its starting point. A *closed circuit* is one that is continuous and will permit the passage of current. An *open circuit* is one that is not continuous and, being broken or open at some point, will not permit passage of the current. The term is used very commonly, and broadly covers any system for the conducting of electric current from its source, through some electrical instrument or appliance and back to the starting point.

CIRCUIT BREAKER—Generally speaking, any device for automatically breaking or opening a circuit and thus preventing flow of current through the circuit. More specifically, it may be a device similar to a *switch*, used to automatically open the circuit when the current falls below a certain point, or rises above a certain fixed value. Thus an *overload circuit breaker* is an arrangement which will automatically open the circuit if the current becomes too great. This is shown in Fig. 1. It is a simple form and consists of an *electro magnet* which requires a certain force to operate it. When the current in the circuit reaches a certain point, the magnet operates, pulling down the bar "A" against the action of the spring "P," and thus breaking the circuit. As long as the current is not more than that fixed by the requirements of the circuit—the magnet being arranged in accordance with the particular requirements — the current flows, the breaker operating to open the circuit only when more than the desired current is introduced into the circuit. In the illustration we can consider "S1" and "S2" as the source of current, "M" as an electro-magnet in series with the motor "D," and "A-B" a switch, in which "A" has attached to it a piece of soft iron. Now if the current required for the motor is ten amperes, the magnet can be so designed that it will not act on the bar "A" until 15 amperes flow through it. Thus, as long as the current is less than this, the circuit is continuous, but if the current is increased to 15 amperes, the magnet operates and attracts the bar "A," breaking its contact with "B" and thus opening the circuit. In Fig. 2 a simple *underload breaker* is shown. In this case the operation is the reverse of that of the *overload breaker*. Here the bar "A" is held down in contact with "B" by the magnet "M" thus keeping the circuit closed through the motor "D". Now assume that the motor requires ten amperes and the magnet requires the same amount to hold the bar "A" in place against "B". If then the current in the circuit falls below ten amperes, the magnet can no longer hold the bar "A" against the action of the spring "P" and it rises, breaking the contact between "A" and "B" and thus opening the circuit. These circuits are of course only assumed for the purpose of illustrating the theory of circuit breakers. Actually the system is generally much more elaborate. Circuit breakers are used for many purposes—chiefly to prevent any sudden increase or decrease of current in a circuit, and to control the flow of current for charging batteries. (See *Charger-Storage Battery.*)

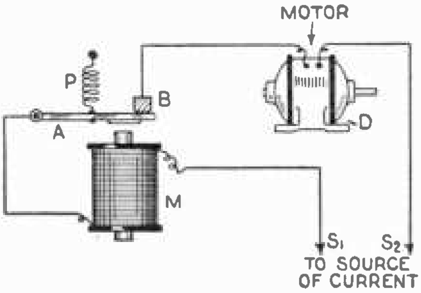

Fig. 1. Overload Circuit Breaker.

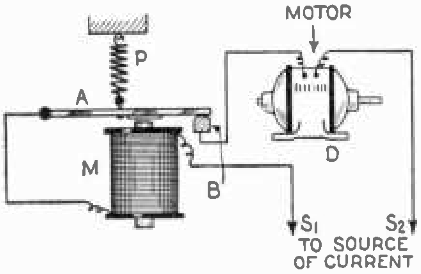

Fig. 2. Underload Circuit Breaker.

CIRCUIT, BROKEN—A circuit that is not continuous, having been opened either purposely, or due to some defect in the circuit. The term is usually applied where the circuit is opened by the breaking of a wire, or through a faulty contact.

CIRCUIT, OPEN — A circuit which is open and through which current cannot flow, or a circuit which is normally open but which can be closed at will by pressing a key or button. A common example of open circuit is in the case of an electric bell circuit. The illustration shows such a circuit. Normally it is open and current does not flow to the bell, but on pressing the button B, the circuit is closed and current passes to the bell. (See *Circuit*, also *Closed Circuit.*)

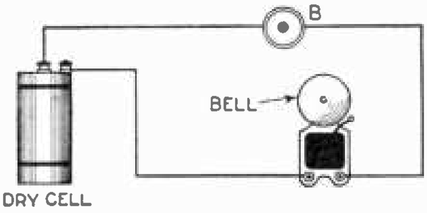

Open circuit which can be closed by pressing button "B".

CIRCUITS, PARALLEL — Circuits starting at a common point and ending at a common point. The illustration shows three *parallel circuits*. Actually each one is a part of the whole circuit. (See *Parallel Resistances.*)

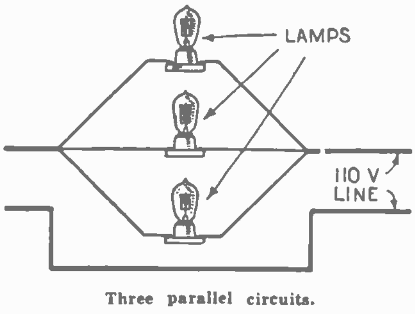

Three parallel circuits.

CIRCULAR MIL—The cross-sectional area of conductors (wires) is usually designated in terms of circular mils. A circular mil is the area of a circle having a diameter of one thousandth of an inch. Thus a wire having a diameter of one quarter of an inch is said to have an area of 250 circular mils. (See *Square Mil.*)

CLEAT—A form of porcelain *insulator* used in house wiring. They are generally arranged in two parts, with corresponding grooves in which the wire is held, and are nailed or otherwise fastened to the wall. Cleats are about the cheapest form of insulator and where the voltages used are low, are quite as effective as the more elaborate types. Cleats are much used in radio work, both for insulating lead in wires coming from the aerial, and for insulating the aerial itself.

CLERK-MAXWELL, JAMES—Born at Edinburgh, Scotland, in 1831, and educated at Edinburgh and Cambridge University, he became Professor of Natural Philosophy at Aberdeen, 1856-69, and Professor of Physics and Astronomy at King's College, London, 1860-65. In 1871 he became the first holder of the new chair of experimental physics at Cambridge, where he died, November 15, 1879. Clerk-Maxwell was recognized in his later years as one of the greatest authorities on physics of his time, and his fame has been steadily increasing since his death. Radio owes Clerk-Maxwell a deep debt. Electricity was the chief study of his lifetime, and his first important paper on the theory of *electro-magnetism* was communicated to the Royal Society in 1867. In 1873 he published his "Electricity and Magnetism," a work on the subject which has never been surpassed. In it he formulated his famous *electro-magnetic theory* of light and his theories on electric waves, which developed into the modern system of wireless telegraphy and telephony through the experiments of Hertz.

James Clerk-Maxwell.

CLIMAX—A high resistance nickel steel alloy, used extensively in *rheostats* (q.v.). It is one of the cheapest and most practical alloys for low-temperature resistance wire. Its chief properties are as follows: Maximum working temperature 540 degrees centigrade, resistance in microhms per cubic centimeter at 20 degrees centigrade, 87.2; temperature coefficient per degree centigrade, .00054, and having a specific gravity of 8.14.

CLIPS—A clamp connector used for various purposes in radio receiving and transmitting circuits. (See *illustration.*) In a receiving set a clip may be used at the end of a flexible wire for the purpose of varying the connection to the various "B" battery taps; clips are very frequently used in making the connections to the storage "A" battery, and for many other purposes about the receiving set. One of their principal uses in the transmitter is in connection with the leads to the oscillation transformer. Flexible leads with clips attached come from the other units of the transmitter and are clamped at will along the various turns of the oscillation transformer until the point giving the desired wave length or highest radiation reading is found.

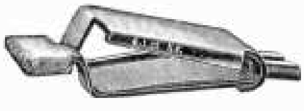

Type of clip in general use.

CLOCKWISE—A term used to indicate that the rotating part of an electrical machine or instrument moves from left to right (when facing it), or as the hands of a clock. In the illustration the needle or pointer of the meter moves in a clock-wise direction. In an electrical measuring instrument this is always true where the zero point is at the left part of the dial. The opposite is known as *counter-clockwise.* When looking at the *armature* of a motor or generator, if the rotating part moves from left to right it is said to have a clock-wise motion.

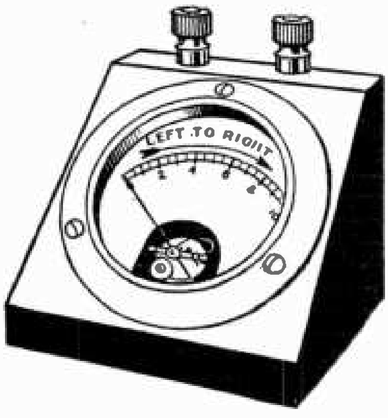

Indicator needle moves from left to right in clockwise motion.

CLOSE COUPLING The arrangement of two coils acting as *primary* and *secondary,* placed close together in such manner that the coupling or electrical relation between them is close. In the illustration the secondary coil is shown entirely within the primary. In other words the transfer of energy will be large in this case as the coupling is close. If the secondary is withdrawn a certain distance from the primary, the effect is referred to as loosening the coupling, and the coils are then said to be *loose coupled,* the transfer of energy from one to the other being considerably less than in the case of close or *tight coupling.* (See *Coupling.*)

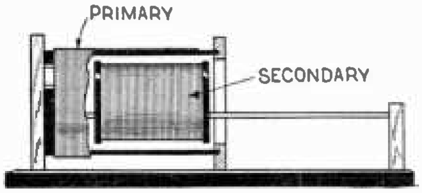

Loose coupler with secondary placed to produce close coupling.

CLOSED CIRCUIT—A circuit which permits the continuous flow of electricity. The illustration shows an ordinary dry cell connected in circuit with a small lamp. In Fig. 1A the circuit is closed and current is permited to flow to the lamp. In Fig. 1B the circuit is open and no current can flow. A closed circuit is therefore an electrical circuit that is continuous. (See *Circuit,* also *Open Circuit.*)

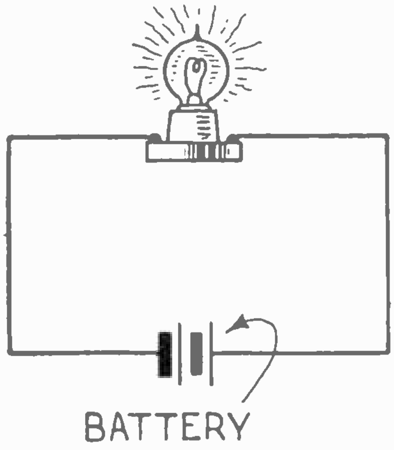

-1A-

Circuit closed.

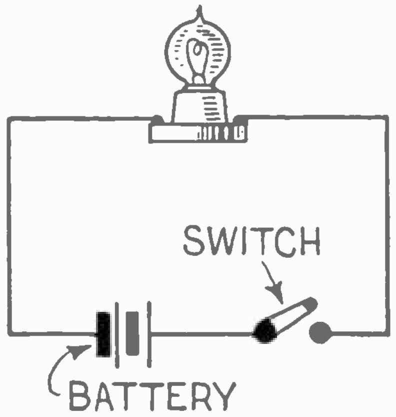

-1B-

Circuit open by means of switch.

CLOSED CIRCUIT CELL—A cell that can be used in a closed circuit for a considerable length of time without the impairing effects of *polarization* (q.v.) An *open circuit* cell is one that can only be used intermittently, as for example, with an electric bell where the circuit is only closed for a brief period. If an open circuit cell is used in a circuit that remains closed for some time, polarization sets in and ruins the cell in a very short time. An open circuit cell might be used with the *test buzzer* (q.v.) because in this case the circuit is closed intermittently as the key is depressed. (See *Cell.*)

CLOSED CORE—A core (q.v.) used in *transformers, chokes,* etc., which is continuous, forming a *closed magnetic path.* The illustration shows the difference between a closed core and an open core. In the open core type, one end is the north pole, the other the south pole, whereas in the closed core, there are obviously no poles. The closed core type is occasionally referred to as a *non-polar* core. (See *Core, Transformer,* also *Choke.*)

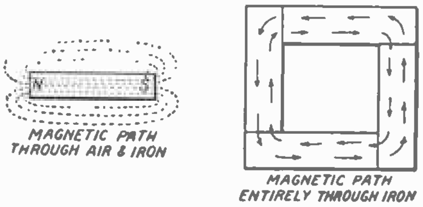
N
S
MAGNETIC PATH THROUGH AIR & IRON
MAGNETIC PATH ENTIRELY THROUGH IRON

Open core above on left. closed on right.

CLOSED CORE TRANSFORMER—A *transformer* (q.v.) having a *closed core.* In the illustration the magnetic path, or core, is continuous, being made of strips of uniform size. This form of core is known as *"laminated."* An efficient transformer is generally made with a closed core. A transformer may be for a variety of purposes, such as *step-up, step-down, audio frequency* or *radio frequency.* (See *Transformer.*)

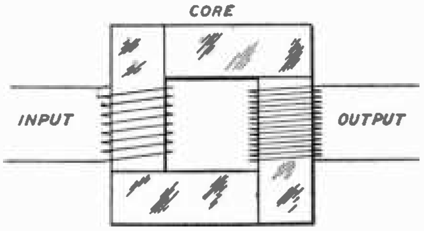

Closed core transformer.

C. M.—Abbreviation for circular mil. (q.v.)

Cm.—The abbreviation for *centimeter,* the unit of length in the C. G. S. system (q.v.).

COCKADAY CIRCUIT — A popular radio receiving circuit devised by the American experimenter, Lawrence M.

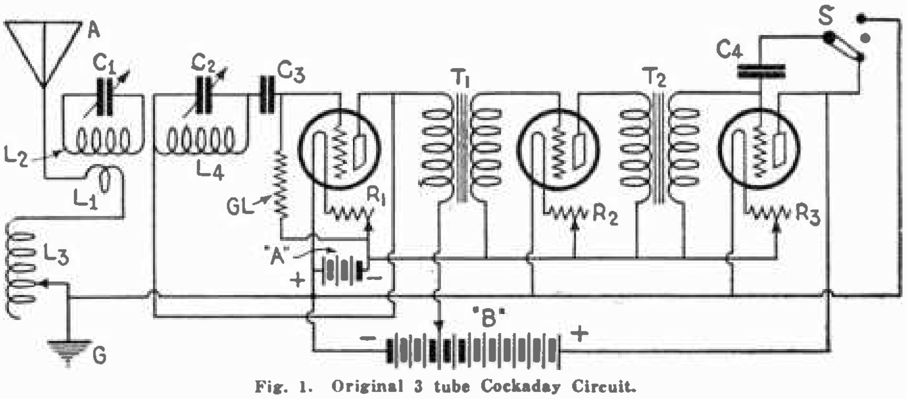

Fig. 1. Original 3 tube Cockaday Circuit.

Cockaday. It includes several radical features not found in the usual systems. Receivers using this circuit are generally made in three, four or five tube types. The chief value of this circuit lies in its extreme selectivity. It is usually referred to as a four circuit receiver, owing to the addition of a sensitizing circuit to the regular three circuit regenerative system. The diagram for the three tube receiver is given in Fig. 1. Here the primary or aerial tuning circuit is seen to have a single turn of wire, L1, wound around another coil L2, which is shunted by a variable condenser C1, thus forming a separate closed, tunable circuit, coupled very loosely with the antenna circuit. The primary, or antenna circuit, has in addition to the single turn, a bank wound coil L3 connected in the ground side of the open circuit and tapped to permit coarse tuning. Thus, the primary is coupled inductively to the separate closed circuit, which in turn is inductively coupled to the grid or secondary circuit containing a coil L4 shunted by the condenser C2, allowing tuning.

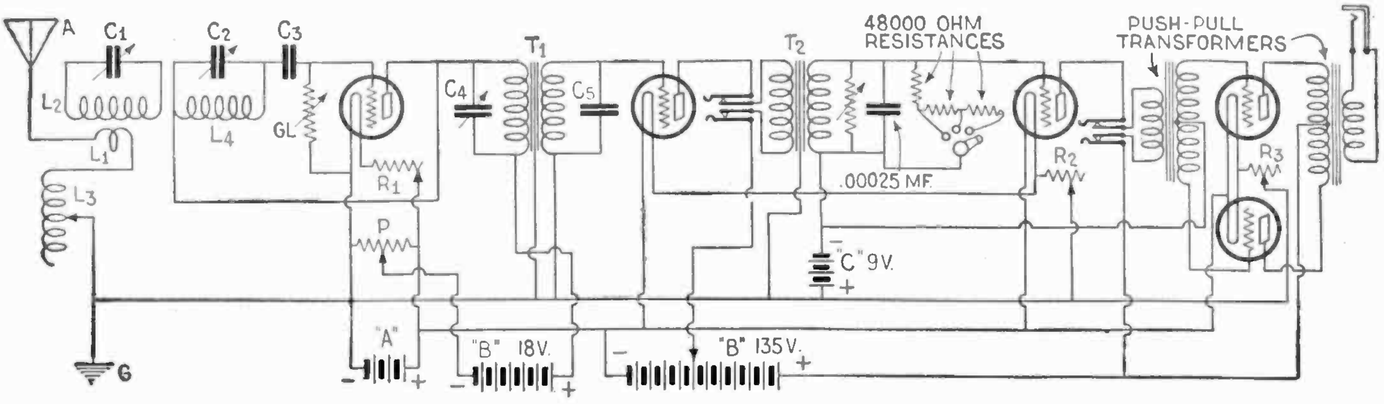

Fig. 2. Cockaday Circuit with push-pull amplifier.

The tone of the music is varied by means of the novel arrangement using a condenser C4 connected to a three point switch. When the switch is placed in the center position the condenser is eliminated from the circuit and the wiring is standard. With the switch on the top tap, the grid of the tube is connected to the ground, thus by-passing a portion of the oscillations and giving a smooth soft tone with a certain loss of volume. With the switch arm on the bottom tap, the grid of the last tube is connected to the plate circuit, thus producing a different note due to the added grid-plate capacity. Coil L2 in the grid circuit is tuned by means of C1, which is a variable condenser of .0005 mfd. capacity.

A peculiarity of the circuit is that the length of the antenna has no bearing on the tuning. This is due to the *semi-aperiodic primary* and the fact that the sensitizing or pick-up coil is only one turn of wire, the inductance of which, compared to the whole inductance of the aerial circuit, is so small that any change in the over-all inductance has little, if any, effect.

This circuit, while originally having three tubes, has been arranged in conjunction with a standard *push-pull audio amplifier*, as shown in Fig. 2, making a five tube set with great power and possessing the characteristic of the original circuit with increased amplification. (See *Push-pull, Cockaday Coil* also *Regenerative Receiver.*)

COCKADAY COIL—The coil used in the *Cockaday circuits.* It comprises four windings, i.e. tapped primary, separate closed circuit coil, grid or secondary winding and a single turn looped around the closed circuit coil. (See *Cockaday Circuit.*)

CODE—Telegraph codes consist of characters formed by combinations of dots, dashes and spaces which represent letters, numerals, and punctuation marks. These characters are sent out by the radio operator with the aid of electrical impulses, and by using suitable receiving apparatus the receiving operator hears the incoming signal and is thus able to interpret the various combinations of dots and dashes and reproduce the original message. The Morse Code of characters came into general use in wire telegraphy shortly after the establishment of that means of communication. In this system, which is also called the "American Morse Code," some of the characters are made up with so-called spaces which are part of the group signal, and are necessary in distinguishing those characters. The International Morse code is a modified form of the American Morse code, and no spaces are used in the characters of the International Code. The International Morse Code is used all over the world for radio telegraphy, and for wire telegraphy in almost every country except the United States, Canada and parts of Australia. The American Morse, owing to the fact that there are fewer dashes in the characters, is about

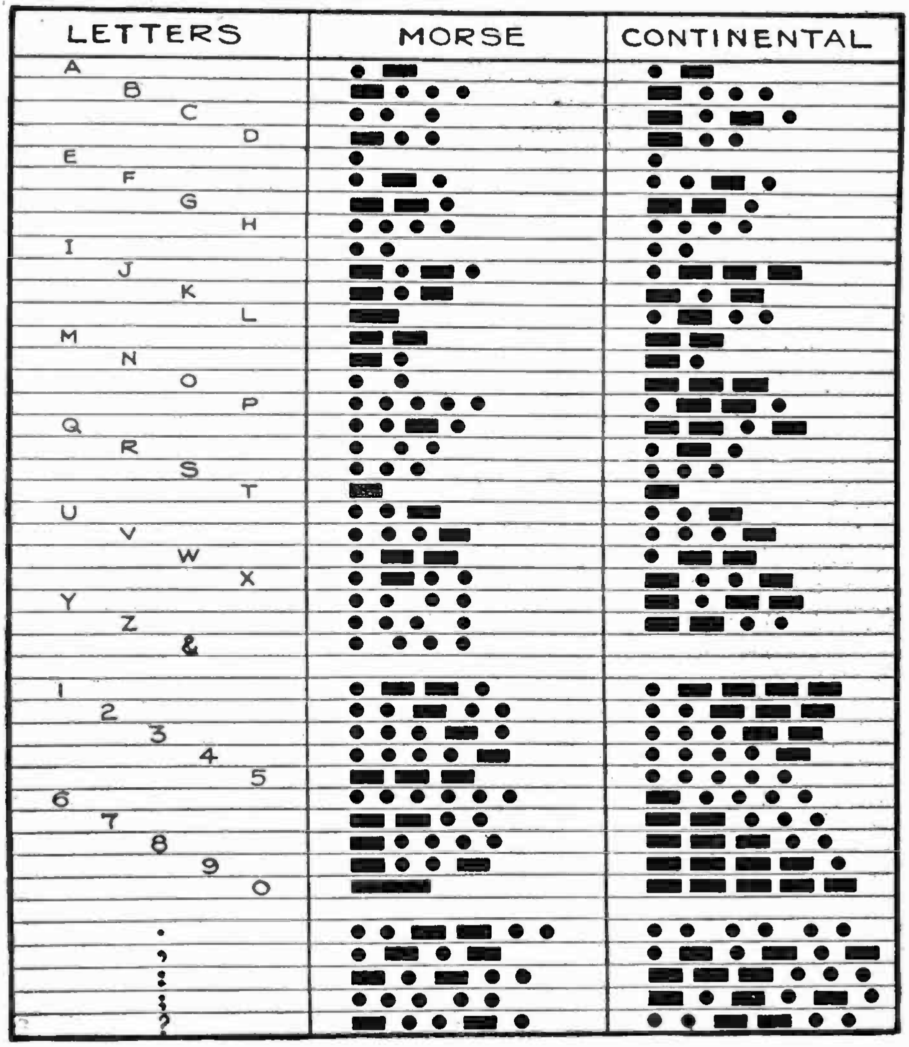

LETTERS	MORSE	CONTINENTAL
A	• —	• —
B	— • • •	— • • •
C	• • •	— • — •
D	— • •	— • •
E	•	•
F	• — •	• • — •
G	— — •	— — •
H	• • • •	• • • •
I	• •	• •
J	— • — •	• — — —
K	— • —	— • —
L	——	• — • •
M	— —	— —
N	— •	— •
O	• •	— — —
P	• • • • •	• — — •
Q	• • — •	— — • —
R	• • •	• — •
S	• • •	• • •
T	—	—
U	• • —	• • —
V	• • • —	• • • —
W	• — —	• — —
X	• — • •	— • • —
Y	• • • •	— • — —
Z	• • • •	— — • •
&	• • • •	
1	• — — •	• — — — —
2	• • — • •	• • — — —
3	• • • — •	• • • — —
4	• • • • —	• • • • —
5	— — —	• • • • •
6	• • • • • •	— • • • •
7	— — • •	— — • • •
8	— • • • •	— — — • •
9	— • • —	— — — — •
0	———	— — — — —
.	• • — — • •	• • • • • •
,	• — • —	• — • — • —
:	— • — • •	— — — • • •
;	• • • • •	— • — • — •
?	— • • — •	• • — — • •

Morse and Continental Codes.

5% more rapid than the International Morse. The American Morse and International (Continental) Codes are given herewith.

COEFFICIENT—In algebra, a number or multiplier of a *symbol*. Thus in the expression 5Y, the coefficient is 5. In electricity, and in its radio application, it is a number expressing the relation or ratio between quantities, one of the quantities generally being unity (1) as a basis or standard for the coefficients. For example, let us assume a copper wire is carrying an electric current, the *resistance* of the wire at 32 degrees Fahrenheit being one *ohm*. If we consider the resistance of the wire at 32 degrees to be unity, or one, then its resistance at any other temperature will be some number or coefficient times one.

To make this clearer, let us take the ratio of an inch to a foot. An inch is one-twelfth, or approximately .0833, of a foot. Any number of inches multiplied by this fraction will give another number in feet. Therefore we say that the coefficient, or multiplier, of an inch as compared to a foot is .0833. A coefficient is thus any number by which a quantity must be multiplied to give another quantity.

In radio the most common cases of coefficients are the following. The *coefficient of coupling* is the ratio between the *mutual inductance* (q.v.) (between the two coils, *primary* and *secondary*) and the square root of the product of the total *self-inductances* (q.v.) in the two circuits containing the coils. Similarly, the coefficient of amplification is the ratio of the effect produced with an *amplifier* (q.v.) to the effect produced without an amplifier. (See *Mutual Induction Coefficient*, also *Self-induction Coefficient*.)

COHEN, LOUIS, Ph. D.—Born in Russia, Dec. 16, 1876. Educated at Armour Institute of Technology, University of Chicago and Columbia University. Noted for his formulae and tables for the calculation of alternating current problems. He has written a book on the subject. Has evolved many formulas for the measurement of inductance and capacity. Has secured a number of patents on radio apparatus.

COHERER—A device formerly employed for detecting *radio waves*. In its original form it consisted of two metal rods separated by a narrow space in a glass tube and the gap between the rods filled with nickel and silver filings. It was found that when the *electromagnetic waves* or signals were passed through the coherer the particles of filings clung together and became *conductors*. Then by using some device such as the tapper hammer of an electric bell to strike the glass tube, again separating the particles after each signal, it was possible to keep the coherer constantly in a state of change between a *conductive* and a *non-conductive* condition, thus making the signals audible. There were many variations of this type of *detector*, which, however, is now obsolete. (See *Marconi Coherer*.)

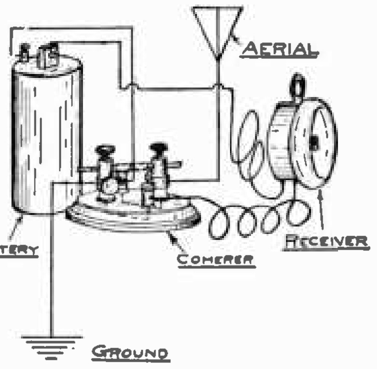

Coherer used in radio receiving circuit.

COIL—A term having very broad application in electricity and radio. It may be used to designate any arrangement of a number of turns of wire, usually copper, and used for almost any purpose in the production of electric and magnetic effects or phenomena. The most common significance of a "coil" in radio is in reference to a number of turns of insulated wire wound on various forms and used for *tuning* (q.v.). The term *inductance coil* is generally used, although actually, any coil of wire will have inductance. Coils come in countless numbers and variety for many purposes. A coil may be used not alone for tuning—that is to adjust a receiver or transmitter to resonance with incoming or outgoing signals; it may be used to connect two circuits, in which case it is known as a *coupling coil;* it may be used with a *core* (q.v.) of iron wire, or laminated pieces, as a *choke coil* (q.v.) or it may be used in connection with *high frequency* apparatus, such as used by physicians. Coils are generally referred to by individual names indicating their construction or use. (See *Inductance, Tuning Coil, Vario-coupler, Choke Coil, Honey Comb Coil, Pancake Coil, Spider Web Coil, Lorenz Coil, Duo Lateral Coils*.)

COIL ANTENNA—Or Loop Aerial. An antenna (aerial) constructed by winding the wire around a square form. The coil (or loop) antenna, because of its comparatively small size, is often mounted on a frame work which may be rotated. If one end of the coil is pointed toward the broadcasting station, the antenna (aerial) will pick up its signals with maximum intensity, but if the coil is rotated from that position the signals from that station grow weaker, until when the plane of the loop is at right angles with the intercepted signals no sound will be obtained. It is a part of the radio compass used at sea. (See *Loop-Aerial*.)

COLLECTOR RINGS—Insulated metal rings attached to the *armature* (q.v.) of an *alternator* (generator of alternating current) in such fashion that the alternating currents can be collected and communicated to the *brushes* without change, these rings are used to collect alternating current from that type of generator and brushes are used to collect direct current from a direct current generator.

COLLODION—A solution used by surgeons to cover slight abrasions or wounds, forming a protective coating similar to skin. It has been adapted to radio use as an insulating coating and is used as a binder on coils of various forms. It has been found more efficient than shellac for this purpose. A solution of pyroxylin or soluble guncotton in ether, with a certain amount of alcohol, depending on the purpose for which intended. When used as a coating for coils it serves as an insulator and also to hold the coils in place, but it should be sparingly applied.

COMBINATION DETECTOR—A detector composed of two different crystals in contact, instead of one crystal with a fine wire to search out the sensitive spots. Some combination detectors are composed of silicon and antimony; zincite and bornite; lenzite and cerusite, and zincite and chalcopyrite. These detectors all function without a battery and are simple to operate.

COMMON CONNECTIONS—A term often applied to a connection that joins several points. For example, in most radio receivers the binding posts for battery connections are arranged so that the "B—" and the "A+" or "A—" are connected to the same point. That is to say the post is a common connection for these leads. The term is also applied to a battery, as a *"B" battery*, which is used for both *detector* and *amplifier*. It is then known as a common "B" battery. The other method would, of course, be to use separate batteries for the detector and amplifier.

COMMUTATOR—In general electrical use, a device for reversing the direction of electric currents in a circuit. A *commutator* as used with a *generator* usually consists of a number of pieces called *segments*, mounted on a circular form but carefully insulated from each other. The illustration shows a common form. In the case of an ordinary motor, the commutator is employed to distribute the current to the windings and in a direct current generator it serves to furnish a direct current to the brushes although the current collected from the windings of the *armature* is alternating.

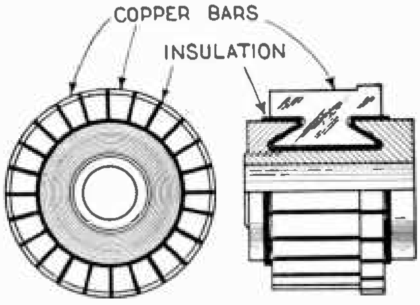

Front and side view of Commutator.

COMPASS, RADIO — See *Direction Finder, Goniometer*, also *Bellin-Tosi Direction Finder*.

COMPONENT—A part of the whole. In mechanics it is one of the parts of a stress or strain (mechanical force). In electricity, *component currents* are the several currents into which a single current may be assumed to be separated, so that if assumed to be acting together they would produce the equivalent effect of the actual current. Thus also, component *E. M. F.'s* (*electro-motive forces*) are understood as the several components into which any electro motive force may be divided. *Impedance* (q.v.) is said to consist of two components, the actual resistance and the apparent resistance present in the opposition offered to the flow of current termed the *reactance* (q.v.) by a circuit. For instance, the total current in the plate circuit of a vacuum tube responding to high *frequency grid voltages* may be considered as made up of two components, viz., a steady or constant current and a high frequency component superimposed on it.

COMPRESSED AIR CONDENSER—A condenser which uses compressed air as the *dielectric* or insulating material between the plates. This type of condenser is used chiefly in transmitting circuits. (See *Condenser*.)

COMPOUND MAGNETS—A combination of several *permanent magnets.* (See *Magnets, Compound.*)

COMPOUND WOUND—A *dynamo* or *motor field magnet* wound with two *field windings,* one of which is connected in series, the other in parallel with the *armature* (q.v.).

CONDENSER—In electricity or radio practice a *condenser* consists of two or more conducting surfaces placed in relation to each other and separated by an insulating medium such as air, impregnated paper, mica, etc. The conducting surfaces are generally referred to as *plates* or *electrodes* (q.v.) and the insulating material is known as the *dielectric* (q.v.). Condensers are used in radio for a variety of purposes, their broad use being to allow tuning. In action, a condenser permits charges of electricity known as *electrostatic charges* (q.v.) to be stored up and released periodically, according to the frequency desired. They may also be used to exclude certain types of energy. (See *Condenser, By Pass,* also *Filter.*) The action of a condenser may be compared to that of a spring when placed under a mechanical strain. In the illustration Fig. 1A a flat spring S is shown in a normal position, NB. In 1B a weight W is placed on the surface of the spring. This weight will force the spring down from its normal position to one of the positions (a), (b), (c) or (d), according to the force exerted and the strength of the spring. If a heavier weight is then placed on the spring it will depress still further until we can imagine a point P in Fig. 1C at which the elasticity of the spring is used up and the opposition is such that it will no longer depress and comes to rest. When the weight is removed the spring will fly back into normal position, Fig. 1D, and in doing so will vibrate back and forth a number of times, thus producing, or rather, returning the energy that was stored in it during the moment of stress.

Considered as analagous to a *condenser,* in the above case the weight can be likened to the electrical charge applied to the condenser, the *resistance* of the spring may be considered somewhat similar to the *capacity* of the condenser, or the power returned on release of the spring may be regarded as the electrical power released by the condenser when the charging force is removed. The capacity of the condenser then will be seen as the ability to store up energy placed in it.

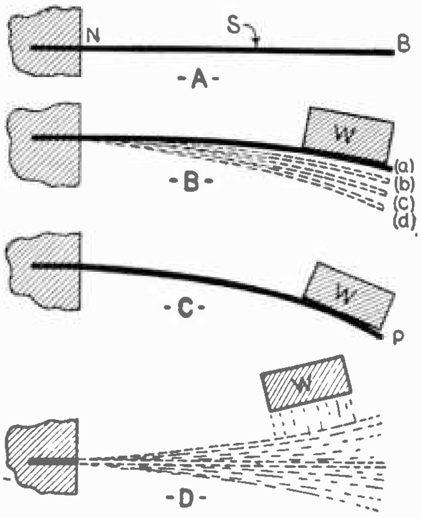

Spring analogy of a condenser.

In the case of the spring, if too great a force is exerted, it may break down or as would actually happen, the elasticity would be lost. The same applies in the case of a condenser. If the charging force is too great, it may exceed the dielectric strength of the condenser and result in puncturing the insulating material between the plates. (See *Breaking Down of Dielectric.*) The energy obtainable from a charged condenser will be equal to the charge placed in it providing there is no leakage or other loss. In a well-designed condenser the leakage will be so negligible that it need not be taken into consideration.

The relation between the charge in a *condenser,* the *voltage* applied and the *capacity of the condenser* is expressed as Q = C × E, where Q is the quantity of the charge in *coulombs,* E the potential in *volts* and C the capacity in *farads.* (See *Capacity, Condenser, Variable; Condenser Fixed; Condenser Transmitting,* also *Condenser Curves.*)

CONDENSER, AIR—See *Air Condenser.*

CONDENSER, ANTENNA—See *Aerial Tuning Condenser.*

CONDENSER, BY-PASS—A condenser, usually of the fixed (q.v.) type arranged in a circuit in such manner that currents of a certain frequency will pass freely around any obstructing object, as, for instance, a high resistance element necessary to the circuit. An example of the use of a condenser in this manner is shown in the case of a "B" battery where the battery has a high internal resistance. A fixed condenser is placed across the terminals of the battery without injury to it and offers a path of low resistance for the high frequency radio currents.

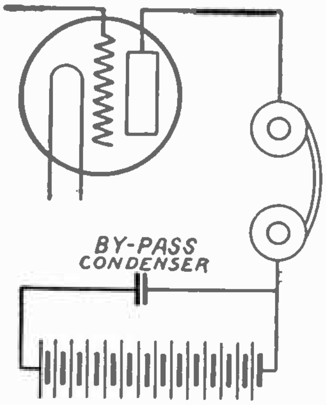

By-pass condenser across "B" battery.

CONDENSER CURVES—Curves showing the relation of variations in wavelength, frequency, and capacity with rotation of the dial of a variable condenser. Three types of condensers are shown here, the *compression,* Fig. 1, the *straight line capacity,* Fig. 2, and the *square law,* Fig. 3. In Fig. 1A is shown the relation of capacity and wavelength changes when the dial of a sample compression condenser was rotated. The condenser was shunted across a fixed coil. It will appear that the increase in capacity is gradual at the lower part of the range but increases very rapidly at the upper ranges. This is due mainly to the relation of the curvature of the upper plate to the lower plate as will be seen from the illustration, Fig. 1. These condensers are very satisfactory on about the lower two-thirds of their range, but through the remainder of their range a very slight adjustment of the knob causes an extremely large change in capacity, making them very difficult to handle near their maximum capacity. Fig. 2 shows the shape of the variable plates in the straight line capacity condenser and Fig. 2A illustrates the relation of wave-length and capacity changes with rotation of the dial. Here the capacity depends upon the overlapping area and is therefore proportional to the angle of rotation of the movable plates. The capacity curve appears as a practically straight line running diagonally across the chart. The wave-length, however, changes rapidly at the lower range and very gradually at the upper range as indicated by the wave-length curve. Fig.

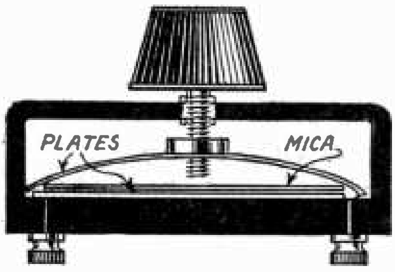

Fig. 1. Compression type condenser.

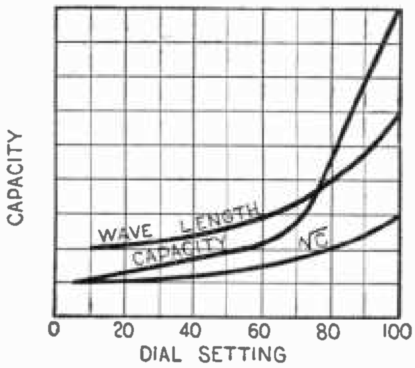

Fig. 1A. Curve for compression type condenser.

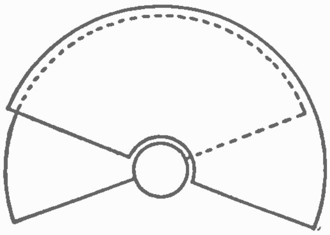

Fig. 2. Straight line condenser.

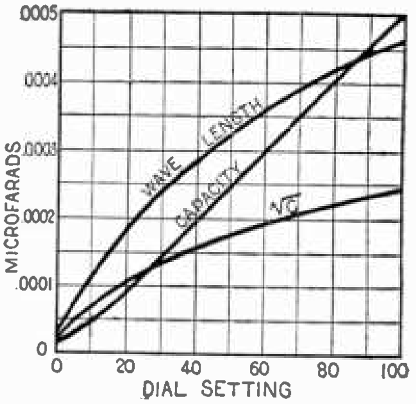

Fig. 2A. Curve for straight line condenser.

3 shows the shape and disposition of rotor and stator plates of a condenser so designed as to read directly in wave-length from the settings of the dial.

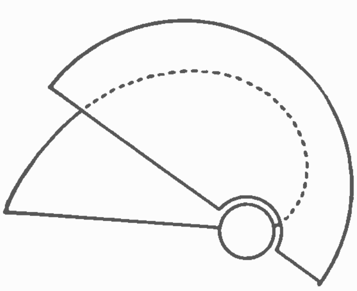

Fig. 3. Square law condenser.

This is known as a square law condenser, and the relational curves are shown in Fig. 3A. Here we find the wave-length curve a straight line and directly proportional to the angle through which the dial is turned. The line indicating the square root of the capacity √C is a straight line, and as the wavelength is the product of the square root of the capacity multiplied by a constant (varying with the individual condenser) the wavelength curve must also be a straight line.

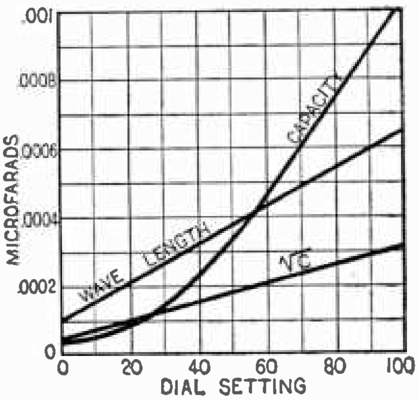

Fig. 3A. Curve for square law condenser.

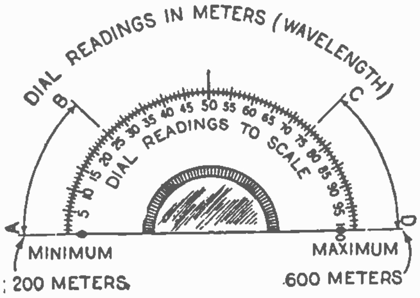

Fig. 3B. Showing effect of straight line condenser on dial reading.

Such condensers are much used in wave meters (q.v.). Condensers are also made so that the frequency curve is a nearly straight line, varying more or less directly as the angle of rotation of the plates. Fig. 3B illustrates the result when an ordinary straight line capacity condenser is arranged with a dial calibrated directly in meters. It appears that the readings are bunched at the lower range (A to B) and spread out at the upper ranges (C to D). The reason for this can readily be seen by glancing at the curve Fig. 2A.

CONDENSER, FIXED—A *condenser* which has a fixed or constant capacity, allowing no variation of the value. (See *Condenser*, also *Condenser, By-Pass.*)

CONDENSER, GRID—A condenser, generally of low capacity, of the fixed or variable type, used in series with the grid of the *detector tube* to prevent excessive charges from affecting the *grid.* These condensers generally have a capacity between .00025 and .0005 *micro farads.* Due to the blocking action of these condensers, a high resistance, known as a *grid leak* (q.v.) is placed across the condenser or from the grid to filament, thus allowing the accumulated charges to leak off the grid of the tube in time for the grid to be free for the succeeding electric charge. The grid leak may be considered as a valve controlling the amount of energy that the detector can efficiently take care of without overload. (See *Grid Leak.*)

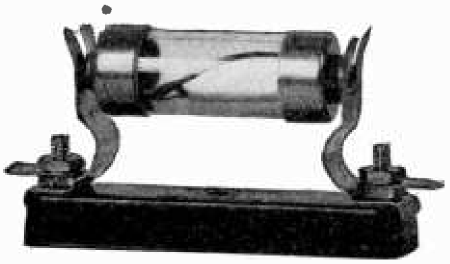

Grid leak and condenser.

CONDENSER, IN PARALLEL—A condenser arranged in a circuit in such manner that the current is divided between it and the other condenser or coil with which it is connected. The illustration shows a *variable condenser* connected in parallel with a coil, both placed in series between aerial and ground and forming a tuned primary or aerial circuit. (See *Parallel Connections.*)

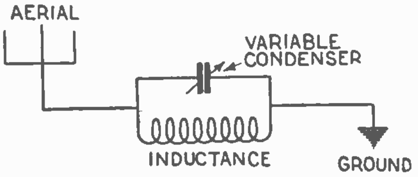

Coil and condenser in parallel.

CONDENSER, IN SERIES—A condenser connected in a circuit with other condensers or coils in such manner as to form a continuous path, allowing the current to pass through each unit without dividing. The illustration shows a condenser connected in series with a *tuning coil,* aerial and ground, thus forming a tuned primary or aerial circuit. (See *Condenser, in Parallel,* also *Series Connections.*)

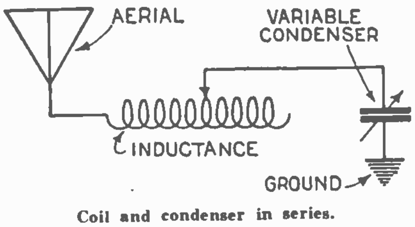

Coil and condenser in series.

CONDENSER, MICA—A condenser, usually of the fixed type, employing sheets of mica as the dielectric or insulating material between the plates. The condensers of small capacity used in receiving circuits are generally made with mica as a dielectric, but owing to its cost and thickness it is seldom used in low-power, high-capacity condensers. Mica is also used in condensers for high-power work such as transmitting, when the capacity is low. (See *Condenser.*)

CONDENSER, PHONE—A fixed condenser of comparatively low capacity used across the *telephone receivers* to allow a by-pass for a certain portion of the currents and thus prevent undue strain on the phones. (See *Condenser, By-Pass.*)

CONDENSER, TRANSMITTING—See *Condenser, Mica.*

CONDENSER, VARIABLE—Any condenser so arranged as to allow variation of its capacity. This may be accomplished in many ways, the usual method being to have the plates divided so that one series is movable with respect to the others, the capacity changing with the rotation of the variable plates. (See *Condenser.*)

CONDUCTANCE — Sometimes called the conductivity. The quality of a given conductor in virtue of which it facilitates the flow of an electric current. A *conductor* (q.v.) is a body having large *conductance* and an *insulator* (q.v.) is one having negligible conductance. *Conductance* is thus the opposite of *resistance.* The unit of conductance is the *mho,* the reciprocal of *ohm.* The symbol is obviously derived from ohm, reversed. It is the conductance of a column of mercury 106.3 cm. (centimeters) long, 1 square mm. (milimeter) in cross-sectional area and of a weight of 14.4521 grams.

CONDUCTIVE COUPLING—The association of one circuit with another by means of inductances (coils) mutual to both circuits.

CONDUCTIVITY, SPECIFIC—The reciprocal of Specific Resistance; being a standard of reference for comparing the conductances of different substances. (See *Specific Resistance, Conductance.*)

CONDUCTOR—Any substance that offers small *resistance* to the passage of electric currents may be termed a *conductor.* This term is the opposite of *insulator,* but there is actually no line of demarcation between the two, since any insulator will permit the passage of current when sufficiently high voltages are used. For all practical purposes, however, a conductor is considered as material that presents little difficulty to the passage of current. The most common conductor is copper, usually in the form of wire. An idea of the relative conductivity (power of carrying electric current possessed by various substances, pure copper being taken as standard) of various substances can be gained from the table following. The materials on the left are comparatively good conductors given in the order of their conductivity, while those on the right are poor conductors and generally used for resistances. (Note: A *resistance* is used to retard the flow of electricity without necessarily stopping it entirely, whereas an insulator is used to stop the passage of current.)

Silver
Copper
Brass
Gold
Aluminum
Zinc
Platinum
Iron
Nickel
Tin

German Silver
Platinum Silver
Manganin
Mercury
Graphite or carbon

There are, of course, many other substances that can be classed as good conductors, or as poor conductors, but the above are more generally known and used. It is a peculiar fact that while good conductors become less conductive as their temperature is raised, insulators lose a certain amount of their insulating properties as their temperature is raised. As too great current in a conductor will cause heat, and as explained above, cause lessening of conducting ability, it will be obvious that for the fullest efficiency a conductor will necessarily be arranged to carry the required current without undue heating. It is understood also that conductors in any part of a radio circuit should be carefully gauged as to size and the effective cross-section of the conductor determined, in order to offset the losses due to heating. It must not be assumed from the foregoing that the larger the wire the greater the efficiency—in fact a large wire is often less satisfactory than a small one in high frequency circuits.

When the conductor is to carry direct current the only consideration is to have the wire large enough to prevent heat losses and at the same time not be too bulky. Where the conductor is to carry *high-frequency* (q.v.) currents there are other factors to consider. High frequency currents are said to travel only on the outside of the wire (see *Skin Effect*) and therefore it is possible to use very fine wire for this purpose when the very low current values obtaining in reception are in question. A phenomenon that must be given attention, however, is that of *distributed capacity* (q.v.). (For more complete explanation of this subject see *Low-loss Coils*. See also *Bus bar*.)

CONNECTING UP or HOOKING UP—These general terms are used in radio to refer to the process of wiring a receiver or any particular part of a receiver or other circuit.

CONSTANT—A quantity or magnitude, which does not vary, derived from actual experiment, which is included as a factor in most formulas for the purpose of bringing theoretical calculations into agreement with experience. Constants are much used in radio calculations. As an illustration, different dielectric or insulating materials used in a condenser have different constants which must be taken into consideration when calculating the capacity of the *condenser*. Thus, as a dielectric, air at ordinary pressure is taken as the standard, the constant being stated as 1. Flint glass has a high dielectric value, its constant being stated as 10.10, other materials varying according to their comparison with air as a standard. (See *Dielectric*.)

CONSTANT CURRENT—An unvarying *current*. The term "constant" is applied to *voltage* when it is steady and unchanging, to *resistance* when it is fixed and unvarying, or to almost any value that does not change.

CONSUMPTION, CURRENT — The term usually employed to denote the current required or consumed by any part of an electrical circuit or by the circuit itself. A resistance in a circuit consumes a certain amount of energy because it dissipates it as heat energy. A filament in the case of a vacuum tube, requires a certain amount of current to heat it to brilliancy. If the filament thus requires one-quarter of an ampere we say its current consumption is one-quarter ampere. (See *Plate Consumption*.)

CONSUMPTION, PLATE—The current required to energize the plate of a vacuum tube. (See *Plate Consumption*.)

CONTACT BREAKER—A contrivance for quickly and automatically making or breaking an electrical circuit normally existing between two contacts. (See *Circuit Breaker*.)

CONTACT DETECTOR—A form of crystal detector in which a small wire, or "catwhisker," as it is called, makes contact with the crystal in the detector and acts as a *rectifier* of the *high frequency waves* received from the broadcasting station. (See *Crystal Detector —Theory of Operation*.)

CONTACT POINTS, or SWITCH POINTS—Generally refers to a small flat-headed machine screw used as a contact point for a *switch*. The illustration shows the types mostly used.

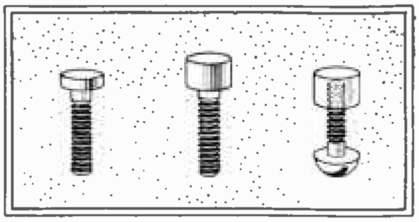

Various contact points.

Leads from various points of a tapped coil may be connected to the several contacts in order to permit variation of the *inductance*. A "contact point" may also mean any point in an electrical circuit at which contact is made to close the circuit.

CONTACT RECTIFIER—See *Contact Detector*.

CONTACT RESISTANCE—When two connecting surfaces in a circuit carrying electrical energy do not make absolute or perfect contact, a certain amount of *resistance* (q.v.) results. In any part of a radio circuit where two wires or *conductors* are joined to form a circuit, it is essential that they make a positive contact. The wires used for connections in a radio receiver are generally soldered or fastened tightly by nuts or bolts, or machine screws in order to form a good contact. If this is not done, energy may be wasted, due to the resistance at the point of contact.

CONTINUOUS OSCILLATIONS—Each oscillation or vibration being of the same amplitude as the preceding one. (See *Sine wave*.)

C. W.—Abbreviation for continuous wave (q.v.).

CONTINUOUS DIRECT CURRENT—A *direct current* which flows steadily without interruption or reversal. The current supplied from a battery for radio use is continuous direct current, while that portion of the *high frequency alternating current* which is rectified by a *vacuum tube* detector is called *pulsating direct current*, as there is an interval of time between the rectified half of each incoming wave, due to the other half not being rectified. (See *Current, Pulsating Direct, Pulsating Current*, also *Rectification*.)

CONTINUOUS WAVES—Also called *Undamped* or *Sustained waves* and usually referred to as *C. W. Electromagnetic waves*, the separate oscillations (each vibration) of which are of constant or unvarying amplitude. In Fig. A a series of *oscillations* (q.v.) are shown. It will appear that one oscillation or *cycle* is of the same amplitude as the preceding one—the amplitude (q.v.) being considered as the height of the wave above the straight line representing time or its depth below that line. Each loop above the line or below it represents an *alteration* (q.v.) (the increase in current to maximum and back to minimum in either direction)—the loops above the line being considered as *positive alternations* and those below the line as *negative alternations*. Now it is evident from the sketch that each succeeding alternation has the same strength (amplitude) as that preceding it, and the same is true of the complete cycle, consisting of one alternation in each direction. Thus the waves are continuous, or constant, as far as amplitude is concerned. Now in Fig. B the amplitude is not constant—it decreases with each succeeding alternation and hence with each succeeding cycle (two alternations). After a comparatively few oscillations the current has decreased to zero, that is, the oscillations have died out completely. In this case the percentage of decrease in amplitude from one cycle to another is known as *decrement* or *damping*. Thus continuous waves may be considered as waves in which, with any train or group of oscillations, each succeeding oscillation has the same amplitude as that preceding it.

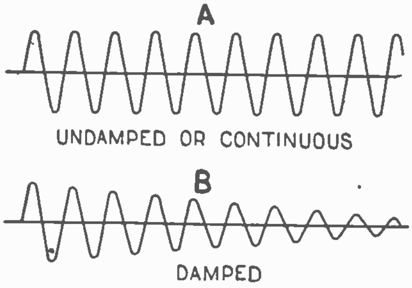

Series of undamped and damped waves.

CONTINUOUS WAVES—KEY MODULATED—Continuous waves (q.v.) which have been broken into dots and dashes of the telegraphic code by means of a *telegraphic key*, these signals being made audible at the receiving end by the use of an oscillating receiver.

CONTINUOUS WAVES—MODULATED AT AUDIO FREQUENCY—The process of variation of the frequency (vibration) of waves radiated by a broadcasting station in accordance with the sound waves is known as modulation. Frequencies from 10,000 down to about 32 are audible to the human ear and are known as audio frequencies. Hence continuous waves which are modulated at *audio frequencies* are waves which have been modulated (varied) by having the voice or music impressed on the *microphone* at the broadcasting station, which in turn causes the strength of the radiated electric wave to vary so as to correspond electrically to the sound vibrations at the microphone.

CONTINUOUS WAVE SIGNALS—See *Continuous Waves*.

CONTINUOUS WAVE TRANSMISSION—System of radio telegraphy using waves as described under "Continuous Waves."

CONVECTION CURRENTS—When a hot and a cold body are separated by a fluid which is free to circulate, heat is transferred from the hot to the cold body by currents of the fluid itself flowing from one to the other; similarly, all parts of a fluid which is being heated quickly come to approximately the same temperature. This transfer of heat by currents of the fluid itself is called "convection." In electricity "Convection Currents" are currents arising from the motion of charged particles thrown off in electrified streams, i.e., the flowing of charged air particles in streams from the pointed end of a highly electrified conductor, sometimes called "electric wind."

CONVENTIONAL—This term is much used in radio to denote a form established by custom. A conventional circuit for some particular purpose is understood as a circuit which has been generally used and which is generally considered as standard. Broadly, the most common form or use of anything. The conventional method of amplifying signals of *audio frequency* is by means of *audio frequency transformers*. Any other means is unconventional, although it may produce similar results.

CONVERSION FACTORS—A numerical factor which gives relation between the magnitude of a quantity expressed in terms of one system of units, and the same quantity expressed in terms of any other system of units. For example, the yard is regarded in the United States as 3600/3937 meter (a meter is approximately 39.37 inches). We say, therefore, that the conversion factor between the yard and the meter is 3600/3937 because any number of yards multiplied by this figure or factor will give the result in meters. (See *Capacity Conversion Factors.*)

CONVERTER—A term applied generally to any mechanical, electrical or electro-chemical device which will convert or change one form of energy to another. More specifically in electricity the term is used in reference to (1) a dynamotor designed with two *armatures*, mounted on one shaft, one being a motor armature driven by the original current, the other armature generates new current; this is also called a "motor-dynamo" and it can transform continuous (direct) current to higher or lower voltages, or, (2) a rotary electrical device used to change direct to alternating current or vice versa. (See *Rotary Converter.*)

COORDINATES—A thing of the same rank with another thing may be termed a coordinate. For example, lines by means of which the position of any point as of a curve may be defined with respect to certain fixed lines. In the illustration Fig. 1 lines known as axes of coordinates are shown. Such a system as this will be the basis for curves (q.v.) showing the relation between varying values. If we assume the horizontal line X as the axis of abscissas, all distances measured parallel to this line will be known as abscissas; distances measured parallel to the vertical line Y being referred to as ordinates. The line A-B is parallel to the axis of abscissas X and is therefore an abscissa. The line B-C is parallel to the axis of ordinates and is referred to as an ordinate. Now if we place a number of lines parallel to each axis (ordinates and abscissas) and assign to each of them a certain value, such as amperes, to each abscissa, and volts to each ordinate, a line can be drawn through a number of points to represent the various values of each quantity as related to the other. Thus, in Fig. 2 we have a purely hypothetical curve representing the assumed value of current in amperes under assumed voltages. In the example, the curve at point A shows approximately three amperes under a pressure of ten volts. This imaginary curve is, of course, shown merely for the purpose of illustrating the meaning of coordinates, ordinates, abscissa, etc. (See *Abscissa, Characteristic curve*, also *Graph.*)

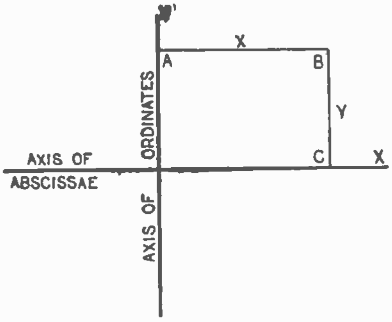

Fig. 1. Axis of coordinates.

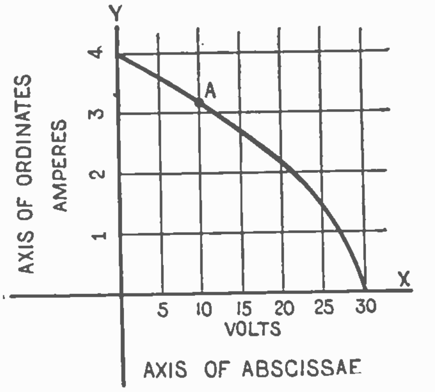

Fig. 2. Curve showing assumed value of current in amperes under assumed voltages.

COPPER PYRITES—Sometimes referred to as *Chalcopyrite*. It is used in radio reception as one of the elements of a *crystal rectifier* (detector) generally in conjunction with zincite or tellurium.

COPPER SULPHATE—A salt of copper. Also known as blue vitriol or copperas. A copper sulphate solution is sometimes used as a preservative for wooden aerial masts. It is used in certain forms of primary cells (devices for transforming chemical action into electric current) and also produces sulphuric acid when subjected to *electrolysis* (q.v.).

CORD, PHONE—The cord or flexible wire used to connect a *headset* (phone) or *loudspeaker* to the terminals of a receiver. Phone cord is usually copper tinsel wrapped around a cotton cord, the whole covered by braided silk or cotton threads. It cannot be used to carry high currents, but is efficient for the comparatively low currents flowing through the phone circuit of a radio receiver. Some phone cords are also formed of a number of fine copper wires covered with a suitable insulating material.

CORE—The central portion of a piece of certain electrical apparatus. Thus the center of a coil will be the *core*, irrespective of its nature, whether air or iron. The core of a *choke coil* (q.v.) is usually made of a bundle of soft iron wires or strips of iron. The core of the closed type is called a *laminated core*, being made up of laminations or layers of the core metal. (See *Choke or Choking Transformer, Closed Core.*)

CORE LOSSES—The losses due to the *stray currents* (q.v.) set up in an iron core while under the stress (strain) of a *magnetic field* (q.v.). Also losses due to *hysteresis* and *eddy-currents*. (See *Magnetic flux, Eddy current*, also *Hysteresis.*)

CORE TRANSFORMER (Iron)—A transformer having the wire wound over an *iron core*, the iron forming a path for the *magnetic lines of force*, as distinguished from a *transformer* in which there is no iron core—called *air core*. (See *Transformer.*)

CORONA—The Latin word meaning crown, originally applied to the crown shaped lights occurring in the *aurora borealis*. In electrical parlance it refers to the luminous discharge which occurs along high tension transmission lines or the aerial wires of a high power radio transmitting station. (See *Brush Discharge;* also *Corona Loss.*)

CORONA LOSS—The energy loss due to corona discharge. (See *Corona;* also *Brush Discharge.*)

CORRECTION FACTORS—A formula for the measurement of certain values is generally made by using certain conditions or types of apparatus as a standard. For example a formula for calculating the inductance of a coil is made by using a particular type and shape of coil as the standard. Therefore the formula will be essentially correct for this type of coil, assuming that its shape and other characteristics are the same. If, however, the nature of the coil to be tested differs from that of the standard coil in some way not covered by the formula, a system of correction must be used to adjust the formula to the particular case. Thus, the formula for inductance of coils takes into consideration the size of the wire or the number of turns, the length of the coil, its diameter and so on. However, it does not take into consideration the ratio between the length of the coil and its diameter. For this purpose a correction factor (necessary consideration) is needed to correct the result of the formula to fit the individual occasion. This correction factor is generally referred to as a *constant* (q.v.), its value being determined generally from a chart prepared to cover various ratios of length to diameter.

CORROSION—Chemical action which causes destruction of the surface of a metal. In the case of iron it is generally rust, corrosion of copper being due to oxidation, or perhaps the action of stray currents causing *electrolysis* (q.v.). This effect is very often noted in the case of an aerial wire where the wire is necessarily exposed to the action of the atmosphere, rain, etc. The effect is particularly important and undesirable where the lead in wire is joined to the aerial.

If the two wires are not soldered securely or at all, corrosion may set in, which will produce in time a high resistance connection and waste the valuable energy of the incoming signals. This is the main point of danger in an aerial system and the best protection is to solder the joint carefully. As a matter of fact the remainder of the wire may corrode to a considerable extent without disastrous effect. In the case of copper, the wire very quickly corrodes when exposed. For this reason insulated wire is sometimes used for the aerial. As the current received by an aerial travels only on the outer surface of the wire (see *Skin Effect*), and a corroded surface is not a particularly efficient conductor, there seems to be some basis for the use of insulated wire for this purpose. Another instance of corrosion is in a soldered joint within a receiving set. All soldered joints should be carefully wiped to remove any trace of *acid flux* (q.v.), which may remain to produce a destroying chemical effect on the wire and cause loss of energy. (See *Electrolysis*, also *High Resistance Joint.*)

COULOMB—The practical unit for electrical quantity. It is defined as the quantity of electricity delivered by a current of one *ampere* flowing for one second. Therefore a *coulomb* of electricity will pass each second in a circuit having one *ohm* of resistance and a pressure of one *volt.* It is also used to express the quantity of electricity which a condenser having one *farad* of capacity will absorb when placed under a pressure of one volt. (See *Capacity, Electro-static, Volt, Ampere Hour, Practical Units, CGS Units.*)

COUNTER ELECTROMOTIVE FORCE—Commonly known as *Counter E. M. F.* It is an opposing electromotive force which tends to resist or act against the original electromotive force. It may be introduced in a circuit to prevent any fluctuation of the original force. (See *Back Electromotive Force*, also *Choke Coil.*)

COUNTERPOISE—A *counterpoise* is generally a second aerial suspended either below the main aerial or to one side of it. It is used more often in transmitting stations in place of the usual ground connection, but is occasionally also used in reception where it is impossible to obtain an efficient ground. (See *Aerial.*)

COUPLE, THERMO ELECTRIC—See *Thermo-Couple.*

COUPLED CIRCUITS—The term generally used to refer to circuits that are joined by *coupling.* (q.v.) In the illustration the *primary circuit* (the aerial ground and tuning devices) is coupled or joined to the *secondary circuit* by *inductive* coupling. That is to say, the signals impressed on the *primary circuit* are transferred to the *secondary circuit* by means of *inductive coupling;* when the lines of magnetic force around the primary coil cut the secondary coil, a voltage is set up in this coil which causes a current to flow in the secondary circuit, and it is then said the coils are "inductively coupled", there being no connecting wires between them. Signals might be transferred to the secondary circuit by any of several means, but in each case the method of joining the two for the purpose of effecting a transfer of energy is referred to as "coupling the circuits." (See *Coupling.*)

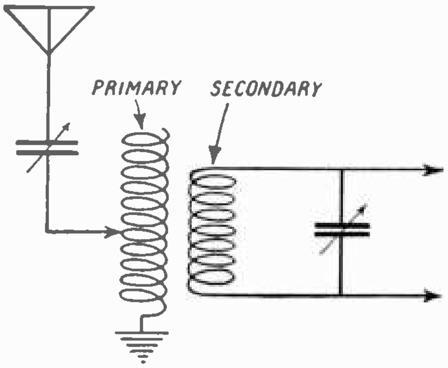

Circuit in which primary and secondary are joined by inductive coupling.

COUPLER—See "Coupler—(Loose)" and "Coupler—(Vario)".

COUPLER—(LOOSE)—A tuner used in a radio receiver, consisting of a primary and secondary coil so constructed that the secondary coil slides on rods in and out of the primary coil. When the secondary is entirely within the primary the coupling is said to be "tight," and as the secondary is withdrawn from the primary the coupling becomes "loose," depending upon the distance which the coils are separated.

COUPLER—(VARIO)—Form of tuner used in a radio receiver, having primary and secondary coils arranged in such a manner that the secondary rotates on a shaft within the primary. The degree of coupling is governed by the variation of the angle between the axes of the two coils.

COUPLING—The relation of one coil of wire carrying alternating currents to another in which energy is set up by the first coil. It is used to define the intimacy between the primary and secondary windings of various types of *oscillation transformers, couplers* and *tuning devices.* This transfer of energy may be produced in several ways, as by *inductive coupling, conductive coupling* or *electro-static coupling.* The more common method is to place the primary and secondary coils in close proximity and produce the energy in the secondary through induction. The illustration Fig. 1 shows a simple two slide tuning coil which is divided into a primary and secondary by *conductive* coupling. Here the transfer of energy from the aerial and ground, or primary circuit, to the secondary circuit (detector and phones) is accomplished with little loss of energy as the circuit is through wire conductors. This is also known as *direct coupling.* Now in Fig. 2 a simple form of *loose-coupler* is shown. This device is made in numerous forms, the principle of operation however, being the same in all cases. Here one coil, the secondary in this case, is made to fit into the larger coil which is the primary. The energy flowing in the primary coil induces or sets up a similar current in the secondary by reason of its proximity. The energy in the secondary, however, will not be as great as in the case of Fig. 1 where it is transferred directly through a wire conductor. The coupling shown in Fig. 2 will on the other hand, permit much closer adjustment to the desired signals.

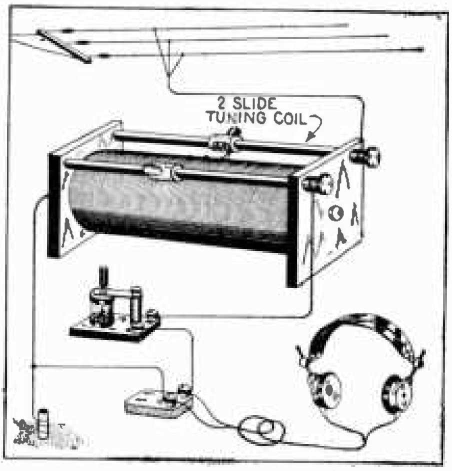

Fig. 1.—Two slide tuner divided into primary and secondary circuits by conductive coupling.

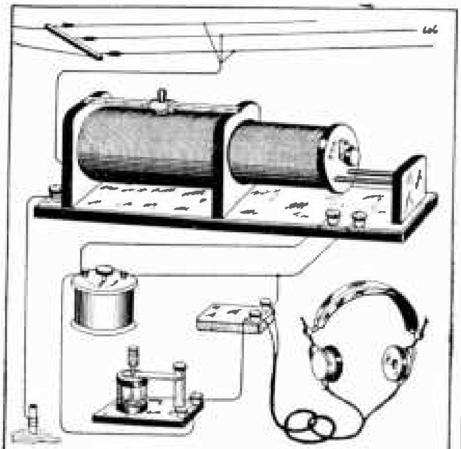

Fig. 2.—Circuit in which primary and secondary are inductively coupled by loose coupler.

The other method is known as *capacitive* or *electrostatic coupling.* This is shown in Fig. 3. Here the secondary coil is not in direct inductive relation to the primary, in fact it might well be almost any distance away were it not for the loss in efficiency due to long connections. The connection between the primary and secondary coils is by means of condensers. The transfer takes place through these and is determined by the setting of the condensers. It will appear that from a standpoint of maximum transfer of energy, the direct or *conductive coupling* system is best. The factor of tuning enters here however, and must be considered. In order to eliminate undesired signals the various systems of coupling are resorted to. It is known that *selectivity*—the ability to tune out undesired signals while receiving the desired signals—is often obtained only at the expense of a certain amount of energy or volume. It is a safe assumption therefore, that if there is any such thing as maximum efficiency in tuning apparatus, it must be a general term covering both selectivity and volume.

Coupling in its various forms is intended to refer not only to receiving circuits but to transmitting circuits as well. Its application in transmitting is covered under the heading *Spark transmission* (q.v.).

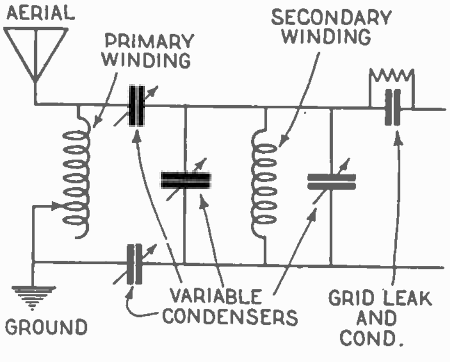

Fig. 3.—Electrostatic coupling between primary and secondary by means of variable condensers.

COUPLING COEFFICIENT — The closeness of *coupling* is specified by a quantity representing the ratio of the *mutual inductance* between two coils to the square root of the product of the two *self inductances.* In its practical sense coupling coefficient refers to the inductive relation of two coils. Coupling coefficient is defined as the ratio

$$\frac{Xm}{X.\ X,}$$

Where Xm is the mutual or common reactance X is the total inductive or capacitive reactance in the primary circuit and X, the similar reactance in the secondary circuit. (See *Coupling, Mutual Inductance.*)

COUPLING, DEGREE OF—The extent of relationship between coupled circuits. In a great number of circuits the coupling is arranged so that it may be adjusted either close or loose, as the requirements of the reception may be. Variable coupling may be obtained either by varying the distance between the two coils of a *loose coupler*, or by changing the angle between the coils of a *variocoupler*. (See *Coupling.*)

COUPLING TRANSFORMER—(OSCILLATION TRANSFORMER) — This is the "tuner" of the transmitting circuit. It consists of primary and secondary windings usually of flat strips of copper, or copper ribbon, instead of wire, in order to reduce losses due to *skin effect* (q.v.). The primary coil receives the energy in the form of alternating currents from the generating device; these currents set up lines of force which interlink with the windings of the secondary coil connected in series with the antenna system, and in this manner the energy of the oscillation (alternating) current is transferred to the radiating system (aerial and ground) and thence into space. The purpose of the oscillation transformer is to produce a sharp wave which is readily tuned at the receiving end and causes a minimum of interference. The greater the distance between the primary and secondary coils the more sharp the wave becomes.

A device serving the same purpose in a receiving set, i.e. to enable the receiver to separate sharply the received signals, is sometimes called a receiving transformer, although commonly known as a tuner, coupler, vario-coupler, etc.

C. Q.—Abbreviation of inquiry made in code by a radio telegraph station desiring to communicate with another station. (See *Abbreviations.*)

C. Q. D.—The original Distress Signal (q.v.) established in 1904 by the Marconi Company. The origin of this signal is said to have developed from the use by telegraphers on land lines sending the C. Q. call when they wished everyone on their circuit to listen to the message being transmitted.

The signal C. Q. D. is now obsolete, having been replaced by S. O. S. at the International Radio Telegraph Convention held at Berlin, Germany, in July, 1908.

CRITICAL—A term often used in reference to the tuning qualities of a radio set, and, less frequently, to indicate a *vacuum tube* that *oscillates* too freely. If a receiving set is not critical in adjustment it is said to tune broadly, that is to say, not sharp in tuning. A critical set has a very narrow margin of control in tuning—the particular signals being receivable within only a very few points on either side of their true wave setting on the tuning dials. If a set is difficult to tune it is termed "too critical". This means that it tunes too sharply and it is difficult to get proper results, due to the possibility of passing by signals unnoticed when turning the dials. Obviously there is a middle point wherein a set may be selective and yet not too difficult to tune. If a vacuum tube goes into *self-oscillation* (q.v.) too readily it is said to be critical. This effect is not desired in a tube except in rare instances. (See *Resonance, Tuning, Broad, etc.*)

CRITICAL CURRENT—The current required to bring about some particular effect in an electrical circuit. In radio use, a common example is the exact amount of current needed by the filament of a vacuum tube to bring it close to the *oscillating point* (q.v.) without actually going into oscillation. (See *Critical.*)

CRITICAL POINT OF CURVE—The peak or point of maximum value in a curve showing the relation of any varying values. The illustration shows the curve indicating relation between voltage and signal strength in two types of vacuum tubes. The points marked "Peak" in each curve represent the critical points of the curves at which maximum audibility is obtained, any further increase in voltage not increasing the audibility. (See *Characteristic Curve*, also *Curve.*)

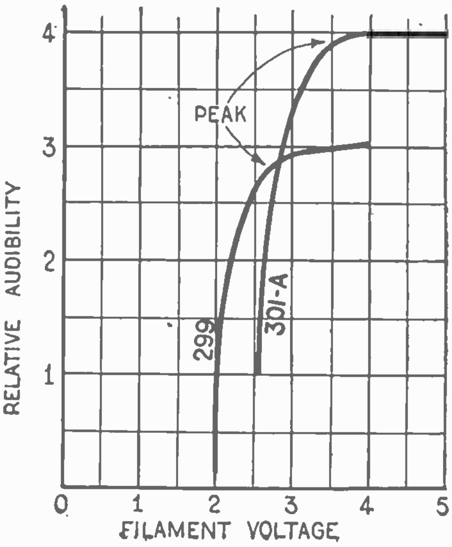

Illustration showing peaks or critical points of curve.

CROOKES, SIR WILLIAM (1832-1919)—English chemist and physicist. Born in London June 17th, 1832, Crookes was educated at the Royal College of Chemistry and afterwards became assistant in the meteorological department of the Radcliffe Observatory, Oxford. In 1855 he obtained a chemical post at Chester.

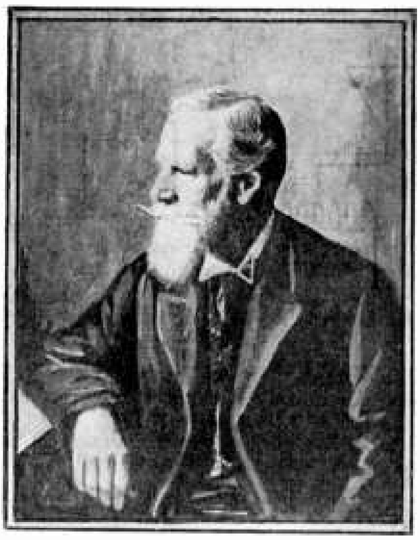

Sir William Crookes.

Crookes early began that series of brilliant researches which has left its mark on the scientific progress of the nineteenth century. In 1861 he isolated the new element *thallium*, during the investigation of the atomic weight of which he made the discovery of his *radiometer*. This led to his researches on the phenomena of electric discharges through highly exhausted tubes, or *Crookes tubes*. The illustration shows such a tube. When a current is passed through it. cathode particles travel along it. His discoveries on these phenomena led directly to the development by Sir J. J. Thomson of the now generally accepted *electron theory*. (q.v.) In 1883 he began his study of the rare earths, and in 1892 he forecast Radio telegraphy on the strength of Lodge's and Hughes' experiments.

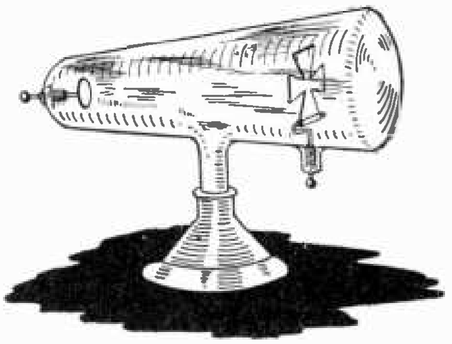

Crookes' tube.

Knighted in 1897, Crookes was awarded the Royal medal in 1875, the Davy medal in 1880, and the Copley medal of the Royal Society in 1904, and the O.M., in 1910. He died in London, April 4th, 1919.

CRYSTAL—A general radio term for a mineral used as a *detector* (q.v.) of radio signals. The Physicist F. Braun noted in 1874 that certain pairs of crystals when arranged so that only a small area of surface was in contact, offered high resistance to the passage of currents in one direction while permitting them to pass readily in the other. This was really the birth of the crystal detector, although radio was not then developed. There are a great many different minerals that will act as detectors for radio signals, some of which operate with a wire, or other metallic contact, others which require a combination of minerals, one in contact with the other. There are also numerous *synthetic crystals* such as *carborundum*. Then again some of the more common natural crystals are subjected to treatment by heat, or other means, to change their nature or to make them more sensitive.

The following crystals are most used in radio reception: *galena, silicon, molybdenum, carborundum, zincite, bornite, chalcopyrite,* and *cerusite*. All of these are taken up under their respective headings. (See *Crystal Detector*, also *Combination Detector.*)

CRYSTAL DETECTOR—A combination of a *crystal rectifier* with contact and supports, used to detect *radio frequency oscillations*. Detectors of this nature are made in a great variety of styles, and use various kinds of crystals. In the more widespread form

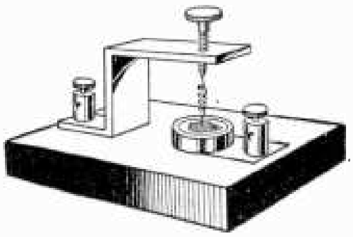

Fig. 1. Crystal detector using contact wire, or "cat-whisker."

the crystal detector consists of a crystal with rectifying properties, held in a cup or spring grip and arranged with a contact wire as shown in the illustration Fig. 1. This may be elaborated upon by supplying a series of flat springs to obtain very fine adjustment of the contact. Fig. 2. shows a combination detector, using two crystals, in contact with each other. This form has several advantages over the wire or *cat-whisker* (as the contact wire is called) type, chief among them the fact that it will withstand severe shocks without losing its adjustment. Fig. 3 illustrates a popular form of fixed crystal detector, in which the contact seldom requires adjustment. This type is of advantage providing it can be made sufficiently sensitive. (See *Carborundum, Vacuum Tube,* also *Crystal Receivers.*)

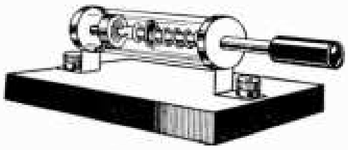

Fig. 2. Combination crystal detector.

Fig. 3. Fixed crystal detector.

CRYSTAL DETECTOR—THEORY OF OPERATION—Theory of operation of a device used to detect radio waves, usually comprising a crystal in combination with a metallic contact, or, in some cases, two crystals arranged to rest against each other. In action this device serves to rectify the incoming signals, that is to say, it changes their nature from an alternating current to pulsating direct current. The incoming waves or signals consist of a series of alternations, the current flowing first in one direction, then reversing and flowing in the opposite direction along the circuit of the receiver. The changes of direction take place many times each second and would not be audible in the head phones in their original state. The crystal, as used in radio reception, has the peculiar property of furnishing a ready path for the flow of current in one direction while offering high resistance to its passage in the other. Now an *alternation* (q.v.) is a rise in current from minimum to maximum in one direction and back to minimum or zero, repeating the process in the other direction. Two of these alternations, one in a positive direction and the other in the negative direction constitute a *cycle.* (q.v.) If the crystal will permit the passage of current in one direction only, it is apparent that only half of each cycle can pass through; the other half of each cycle being lost or blocked out by the action of the crystal.

The illustration "A" gives a graphic representation of the *radio frequency oscillations* as they move over the aerial circuit. The curves above the line show the current in a positive direction while those below the line show it in the opposite or negative direction. If the signals or waves in this state were sent through a telephone receiver or head phone in a radio receiving set no sounds would be reproduced. Then again if the incoming oscillations were continuous, of unvarying strength, the rectified current would merely act on the diaphragm of the phone and hold it in a certain position as long as the signal persisted. This of course would not create any audible or intelligible sounds. If, however, the incoming oscillations were in the forms of groups or trains of waves, in any one of which the successive oscillations were less than the previous ones, the current would be rectified as shown in B and the result would be a series of pulses in the phone as indicated in C. These pulses would occur as often as the trains of oscillations occur (for more complete explanation of the difference between continuous and damped waves, see Damped Waves, also Continuous Waves). It will therefore appear that a *crystal detector* can only be used to receive *damped waves.* When the rectified portion of a damped wave train reaches the phones, the *diaphragm* is depressed and held until the train has died out, at which point the diaphragm is released until another train reaches it. This creates the audible signals and the number of trains occcurring each second will determine the number of depressions of the diaphragm and hence the pitch of the audible signals. (See *Crystal Receiver, Crystal Detectors,* also *Tickler.*)

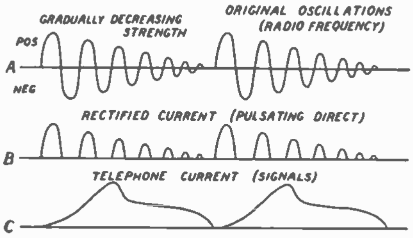
GRADUALLY DECREASING STRENGTH
ORIGINAL OSCILLATIONS (RADIO FREQUENCY)
POS
A
NEG
RECTIFIED CURRENT (PULSATING DIRECT)
B
TELEPHONE CURRENT (SIGNALS)
C

Rectifying process of radio signals.

CRYSTAL RECEIVERS—A crystal receiver is the complete set of apparatus for receiving broadcast programs or *spark transmission* (signals), employing a crystal as the *detector* or *rectifier.* The set comprises a crystal detector, tuning device and phones, together with the usual aerial and ground arrangement. In its simplest form it may be merely a crystal detector, coil of wire and phones as in Fig. 1.

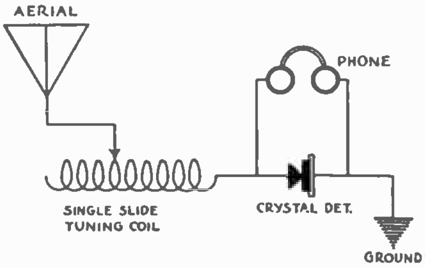

Fig. 1. Simple crystal set with single slide tuner.

Fig. 2 shows a more efficient form, using a two slide tuning coil. This method permits closer adjustment of the receiver to the particular signals desired. A still more efficient form is the *inductively coupled* type shown in Fig. 3. This method allows a very close adjustment and considerably reduces interference. The kind of detector to be used in any of the above types of receiver is immaterial as long as it is sufficiently sensitive. Such an arrangement as shown in Fig. 3 is very inexpensive to construct and when used within ten to twenty miles of the broadcasting station gives excellent results. A very simple arrangement for use with a crystal detector is shown in Fig. 4. This circuit uses a standard *variometer* and eliminates the necessity for a *condenser,* thus having only one control in addition to the adjustment of the crystal.

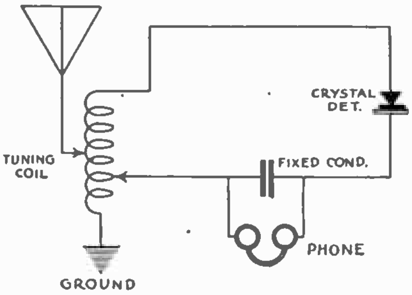

Fig. 2. Crystal receiver using double slide tuner.

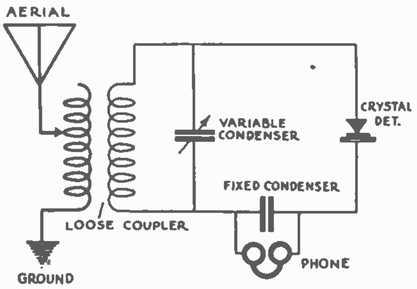

Fig. 3. Selective crystal receiver with loose coupler for tuning.

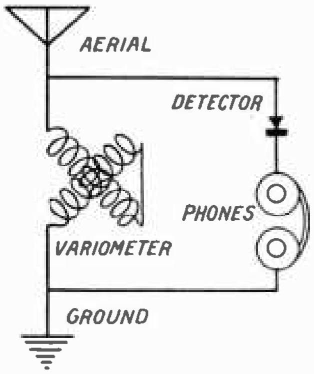

Fig. 4. A variometer is used as the tuning element in this crystal set.

CRYSTAL RECTIFIER—A device for changing an alternating current in a radio receiver into a pulsating direct current. This action is fully explained in "*Crystal Detectors (Theory of Operation).*"

CURRENT—A term used in electrical practise to signify the rate at which electricity flows from point to point in a circuit. The current in a circuit may be likened to the water flowing in a pipe. In this case the water would represent the electricity, the force or pressure represents the voltage, the pipe can be compared to the conductor or wire through which the electricity flows, and the current can be likened to the discharge from the pipe over a given period of time, that is the cubic feet of water per second. For practical purposes the unit of current is considered as the *Ampere.* This is the rate of flow of electricity when one *coulomb* (q.v.) passes a given point in the circuit each second. It will thus be apparent that while the ampere is often considered as the amount of current in a circuit, in reality it is the rate of flow of the electricity, the *coulomb* being the definite unit of quantity. Electric current is of course invisible. The only means of determining when there is current in a circuit is by observing its effect and it is always apparent by one or more of a number of effects. For example, if current is flowing in a wire or other conductor and a compass is placed near the conductor in such a position that the axis on which the needle revolves is parallel to the axis of the wire, the needle will be deflected. If the current reverses and flows in the opposite direction in the wire, the needle will be deflected in the opposite direction. The effect of current is also apparent by reason of the heat which it generates in passing through a wire conductor. If the wire is comparatively small and the current correspondingly large, the heat may be sufficient to be noticeable by placing the hand on the conductor. When the current is small, it can be detected by means of delicate instruments. This heating effect is made use of in various types of *ammeters*. The effect of current may also be noted in an incandescent lamp. In this case the current is made to flow through a fine wire inside a globe from which the air has been extracted, and the heat generated by the passing current causes the filament, or fine wire, to glow.

It is a well-known fact that heat causes expansion and this expansion of a fine wire is the principle on which one type of current measuring device operates.

Current is divided broadly into two divisions, *direct* and *alternating*. These two classes may again be separated into several further groups such as *high* and *low frequency alternating*, *continuous* and *pulsating*. Each phase or characteristic of current is taken up more in detail under its particular heading in this encyclopedia. (See *Current, Direction of Flow,* also *Current, Production of.*)

CURRENT, ALTERNATING—A current which does not flow steadily in one direction but changes its direction in the circuit periodically. (See *Alternating Current.*)

CURRENT CARRYING CAPACITY OF WIRE—The amount of current which a wire or other metallic conductor will carry without over-heating. Whenever current flows through a conductor, heat is generated. The amount of heat will be directly proportional to the resistance of the conductor and the square of the current in it. Now the resistance of a conductor depends on its cross-sectional area, length, and also its nature. It is therefore correct to state that the heating will depend on the amount of current flowing and the cross section and length of the wire. Obviously, a large wire will carry more current without over-heating, than will a small wire. Thus, it is necessary when choosing wire for a conductor to carry a certain amount of current, to select a wire of a cross section sufficient to carry the desired amount of current without undue heating.

The following formula for determining the current carrying capacity of various sizes and kinds of wire uses a certain factor T, which is the permissible rise in temperature above surrounding medium—such as air, earth or water. Where d is the diameter of the conductor in inches, T—the permissible temperature rise in degrees centigrade, N the resistance of conductor in ohms per mil-foot at final temperature, and I the current in amperes, then for solid conductors:

$$I = \frac{K\sqrt{Td^3}}{r}$$ and for stranded conductors: $$I = 0.85\,K\sqrt{\frac{Td^3}{r}}$$

K is a *constant* (q.v.) depending upon the condition of the surface of the wire and upon the amount of heat convection due to air currents. Values of the constant K for air differ according to different authorities from 800 to 1000, the former referring to still air and the latter to open air.

Table of current carrying capacities follows:

Wires and Cables, Insulated: Carrying Capacities, in Amperes, allowed by The Regulations of The National Board of Fire Underwriters for Interior Copper Conductors.

(For Aluminum 84 Per Cent of These Currents is allowed.)

Single Conductor Cables or Each Conductor of Multiple Conductor Cables.

A. W. G.	Area in Circular Mils	Table A. Rubber Insulation	Table B Varnished cloth	Table C. other insulation
18	1,624	3	....	5
16	2,583	6	....	10
14	4,107	15	(18)	20
12	6,530	20	(25)	25
10	10,380	25	(30)	30
8	16,510	35	(40)	50
6	26,250	50	60	70
5	33,100	55	65	80
4	41,740	70	85	90
3	52,630	80	95	100
2	66,370	90	110	125
1	83,690	100	120	150
0	105,500	125	150	200
00	133,100	150	180	225
000	167,800	175	210	275
	200,000	200	240	300
0000	211,600	225	270	325
	250,000	250	300	350
	300,000	275	330	400
	400,000	325	390	500
	500,000	400	480	600
	600,000	450	540	680
	700,000	500	600	760
	800,000	550	660	840
	900,000	600	720	920
	1,000,000	650	780	1000
	1,100,000	690	830	1080
	1,200,000	730	880	1150
	1,300,000	770	920	1220
	1,400,000	810	970	1290
	1,500,000	850	1020	1360
	1,600,000	890	1070	1430
	1,700,000	930	1120	1490
	1,800,000	970	1160	1550
	1,900,000	1010	1210	1610
	2,000,000	1050	1260	1670

Varnished cloth smaller than No. 6 may be used by special permission only.

CURRENT, CONVECTION—See *Convection Current.*

CURRENT, CRITICAL — See *Critical Current.*

CURRENT DENSITY—The number of amperes passing through a wire conductor per square inch of area of the conductor. For example, if a conductor is a wire or bar of copper having a cross section of one square inch and ten amperes are passed through it, the current density will be ten amperes per square inch.

CURRENT, DIRECT—Current that flows steadily, and only in one direction. (See *Direct Current.*)

CURRENT, DIRECTION OF FLOW (ASSUMED) — In electrical practice the *assumed direction of flow* of current is often referred to. This refers to the fact that direct current is assumed to flow from the *positive pole* of a battery, or other source of electrical energy, to the *negative pole* and thence back through the battery from negative to positive. This assumed action is shown in the illustration Fig. 1, where the arrows indicate the assumed direction of flow. Now while the current is stated as flowing from positive to negative outside the source, it is the generally accepted theory that the electron stream moves in the opposite direction, or from negative to positive. In other words, the current as we use the term, flows backward along the path of electrons. (See *Electron Flow.*) This reverse action of current and electrons is shown by the illustration Fig. 2 This apparent anomaly

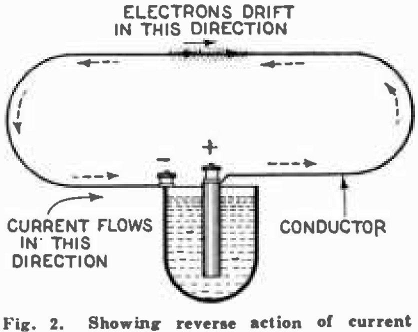
ELECTRONS DRIFT IN THIS DIRECTION
−
+
CURRENT FLOWS IN THIS DIRECTION
CONDUCTOR

Fig. 2. Showing reverse action of current and electrons.

can be likened to a salmon swimming up a rapid. The movement of electrons can be compared to that of the the water in the rapids and while the fish actually swim against the stream, and the electric current flows against the electron flow, the water and the electron stream nevertheless furnish the path of motion in each instance. The reason for the contradiction lies in the unfortunate arbitrary choice of positive and negative poles long before electrons were known. (See *Electron.*)

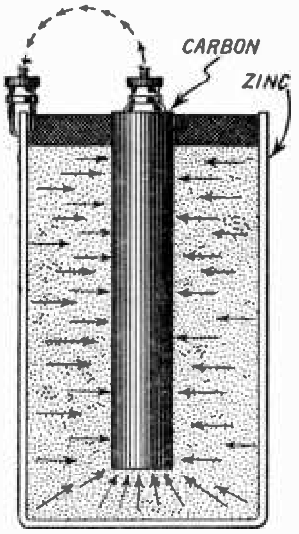

Fig. 1. Illustration showing the assumed direction of current flow in a dry battery.

CURRENT, EDDY—See *Eddy Currents.*

CURRENT, HIGH FREQUENCY—A current that changes direction many times each second. In radio usage, currents having a frequency over ten thousand cycles per second are considered as *high frequency* or *radio frequency currents.* (See *Frequency, Radio Frequency* also *Low Frequency.*)

CURRENT, HIGH TENSION—Alternate term *High Pressure.* When the voltage, tension or pressure, in a circuit is comparatively low, the current is said to be a low pressure or *low tension current.* When the voltage is high, the current is said to be of high tension or pressure. Thus, as applied to radio usage, a dry cell or storage battery giving, respectively, one and a half and six volts for a certain type, are said to produce low tension or low pressure current. A "B" battery (q. v.) giving from 22½ to 45 volts (the usual value) furnishes *high tension* or *high pressure current.* While there is not a very great difference in voltage between a small "B" battery and an "A" type storage battery, the ratio of voltage to current in a "B" battery is far greater than in the ordinary case of the "A" battery and this must be taken into consideration. In general electrical practice the terms are not so arbitrary, and high tension currents might be taken as those above a pressure of several thousand volts.

CURRENT, LOW FREQUENCY—An alternating current that changes its direction of flow a comparatively few times per second. In commercial electrical work any frequency up to about 500 cycles per second is considered as low frequency current. In radio phraseology *audio frequency* (q.v.) currents are termed low frequency. For ordinary radio practice low frequency may be considered as any frequency below about 10,000 cycles. (See *Oscillation* also *Oscillatory current.*)

CURRENT, OSCILLATORY—See *Oscillatory Current.*

CURRENT, PRIMARY—A current which flows directly from its source, such as a cell or generator. The current obtained from any source through a direct conductive circuit is termed a primary current. (See *Secondary Current,* also *Induced Current.*)

CURRENT, NATURE OF ELECTRIC—As explained under the heading *Current,* electric current is manifest or understood by its effects. It may be apparent through the production of heat or light, by producing mechanical action such as required to operate a bell or move an electric motor, or, again, by producing chemical changes (See *Electrolysis*), and still further in its disastrous effects on the human body when the current has sufficient intensity. Without any actual or visible form of manifestation it can best be understood from the *electron* basis. It will be understood that an *electron* (q.v.) small as it is, contains a charge of electricity. When a sufficient number of electrons are in motion in a conductor (Note: current consists of electrons in motion) the current becomes large enough to be measured. As an instance of this fact a current of one ampere intensity (one coulomb per second) requires that about 10^{19} electrons flow past a given point in the circuit each second. When the electrons are not progressing along a conductor, no current flows in the circuit. If no current is being carried by the conductor, the electrons may move about, but they do not progress along the conductor.

CURRENT, PRODUCTION OF ELECTRIC—An electric current is produced in a conductor by applying an *electro-motive force* (voltage) at points of the conductor or by maintaining a difference of *potential* (q.v.) between two points. In the first example we can assume a metal ring or closed coil of wire through which a magnet is pushed or perhaps in the core of which a magnet is being excited, or in the general sense any complete circuit through which the number of magnetic lines of force are being varied. In the second instance the difference of potential may be due to a *primary cell* (q.v.). Thus, the first case can be understood as referring to any mechanical means of producing current, such as *generators.* The other general method is by use of a *voltaic cell,* that is, by chemical action. There are four important methods of producing an *electro-motive force.* The first is by friction, the apparatus used in this connection being termed a static machine; the second is by chemical action, as a primary or secondary cell; third, by mechanical motion, i.e., *dynamos* and *generators,* and fourth, by thermal action—known as a *thermo-junction* which is the wiring of two unlike metals in a conductory circuit so that by heating one of the points of joining an electric current will be set up in the circuit. (See *Static Electricity, Generator, Thermo Couple,* also *Current and Electro-Motive Force.*)

CURRENT, PULSATING DIRECT—When an alternating current is rectified by means of any device which permits current to pass only in one direction, the resultant current will move only in one direction, hence it is known as a *direct current* but unless both negative and positive halves of the *cycle* (q.v.) are rectified or changed to direct current, the resultant current will appear as a series of pulses in one direction. The most common application of this is found in the case of a *crystal rectifier,* or *detector* as used for receiving radio signals or broadcasting.

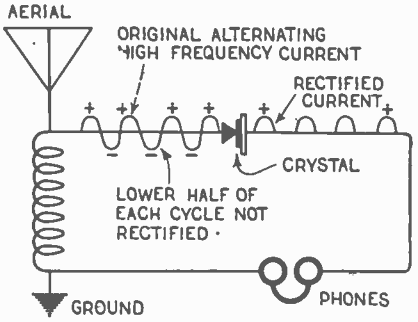
AERIAL
ORIGINAL ALTERNATING HIGH FREQUENCY CURRENT
RECTIFIED CURRENT
CRYSTAL
LOWER HALF OF EACH CYCLE NOT RECTIFIED
GROUND
PHONES

Diagram showing how crystal detector converts alternating high frequency current into pulsating direct current.

When the incoming waves, or high frequency *oscillations* (Q.v.), are impressed on the detector, due to its peculiar properties they can only pass through it in one direction. In other words, granting that the incoming signal, or current, reverses its direction a great many times each second, it will be obvious that at any instant in the *cycle* (q.v.) current is moving in a certain definite direction, as from positive to negative, or vice versa, as stated, due to the peculiar properties of the detector the incoming signal current can only pass through the detector when the current is flowing in the correct direction. This means that only half of the *wave* (q.v.) is passed through and rectified. The result is a series of direct current surges through the balance of the circuit, each surge of current being followed by a space or time interval while the opposite half of the *cycle* or opposite *alternation* (q.v.) is blocked by the uni-directional nature of the crystal. These surges are then audible in the *head phones,* as they will actuate the *diaphragm* (q.v.) whereas a high frequency alternating current such as the original impulse from the broadcasting station would be inaudible. For a more complete explanation of this action see Rectifying action of crystal detector. (See also, *Rectifier, Full Wave Rectification and Rectifying Tube.*)

CURRENT, SECONDARY—A current produced by a *primary current* (q.v.) acting on a *secondary circuit* which does not receive its energy by direct connection with the primary source. The illustration shows both primary and secondary currents and the manner of production. It will be understood that the primary current in this case must be a varying current. (See *Induced Current,* also *Inductive Coupling.*)

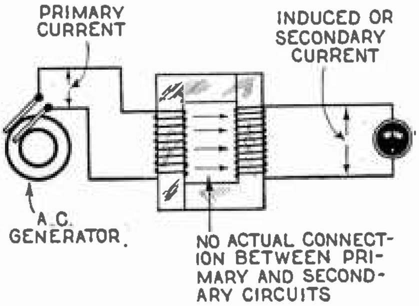
PRIMARY CURRENT
INDUCED OR SECONDARY CURRENT
A.C. GENERATOR
NO ACTUAL CONNECTION BETWEEN PRIMARY AND SECONDARY CIRCUITS

Illustration showing primary and secondary currents and method of production.

CURRENT, THEORY OF ELECTRIC—Under Electromotive force it will be explained that voltage or pressure is the result of a difference of potential (the power possessed by a charge of electricity for doing work) between two points in a conductor. Electric current will flow from the point of higher pressure or *voltage* to the point of lower pressure, the flow continuing as long as the difference is maintained. This action is the same as with water, which will always seek to move from a high point to a lower point. The flow of current is thus an effort to equalize the two *potentials.* (See *Current, Assumed Direction of Flow,* also *Pressure and Voltage.*)

CURVE—A straight or curved line showing the relation between various electrical phenomena or the characteristics (relation of changing values) of an electrical instrument. (See *Characteristic Curve,* also *Sine Wave.*)

CURVE OF SINES—A curved line representing the vibration of a body that oscillates like a pendulum. The illustration shows a *curve of sines* which may be used to represent an alternating current. (See *Alternating Current,* also *Sine Wave* and *Simple Harmonic Vibration.*)

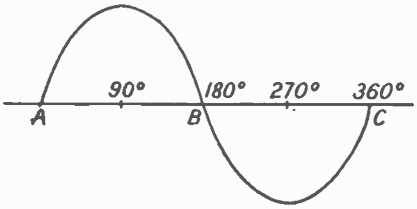

Sine curve.

CUT OUT—An electrical device to interrupt the flow of current through any particular piece of apparatus or instrument, either automatically or by hand; a switch.

CYCLE—In electricity, a period of time during which certain changes take place in an *alternating current,* (q.v.) the same changes occurring again in each successive cycle. We can imagine an alternating current as travelling along a wire in the form of a wave, one half of the wave being in one direction and the other half being in the opposite direction. That is to say, the current flows from positive to negative for one half of the cycle and then in effect, flows from negative to positive the other half of the cycle. (It will be understood that actually current always flows in the same direction, but the action of alternating current is due to a reversal of polarity at the source.) The difference between direct and alternating current is that, in the case of direct, the polarity always remains the same at the source and hence the direction of flow of the current is always in one direction, whereas with alternating current, the polarity changes twice each cycle due to changes at the source, and thus it flows in one direction for a certain fixed period, then reverses and flows in the opposite direction. In the illustration the current is shown as a *sine wave.* (q.v.) We assume that this represents an alternating current of 110 volts pressure such as in the average house lighting system. Now we consider the wave as a circle. The current flows in a positive direction toward its maximum value and at the 90 degrees it has reached the positive maximum. It then falls back to zero at 180 degrees and reverses, rising to maximum on the negative side, which it reaches at 270 degrees. It then falls back to zero at 360 degrees having completed a cycle or two alternations. The operation is repeated over and over a certain number of times each second. Thus, the rise and fall from A to B or 180 degrees is one *alternation,* the rise and fall from B to C is another alternation, the two alternations, one positive and one negative making a complete cycle. If sixty of these cycles occur each second we say that the frequency (q.v.) is sixty cycles. (See *Alternating Current, Phase, Sine Wave.*)

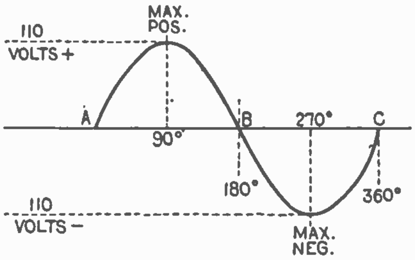

Curve showing action of 110 volts alternating current during one cycle.

CYMOMETER—The name given a type of *wave-meter* designed by Dr. J. A. Fleming. It is used to measure the wave-length of oscillatory (vibrating) circuits. (See *Wave-Meter.*)

CYMOSCOPE—A term used to designate any instrument which enables one to see the effect of electrical waves or to detect their presence. The original form was merely a loop of wire not completely closed as shown in Fig. 1. When held close to a transmitter of short waves a small spark will pass between the ends of the loop providing the *wave-length* is properly adjusted. This was the form used by Hertz (the discoverer of electro-magnetic waves which form the basis of radio telegraphy and telephony) in his early experiments. The more modern form is shown in Fig. 2 where a small lamp is made to light by holding the loop close to the transmitter. A complicated arrangement used to acually see the waves and study their form is called the *Oscillograph.* (See *Oscillograph,* also *Cathode Ray Tube.*)

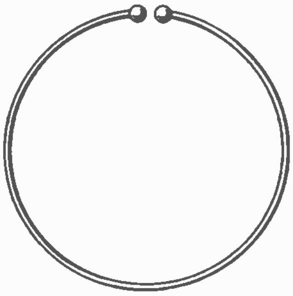

Fig. 1. Early type of cymoscope used to indicate resonance in a circuit.

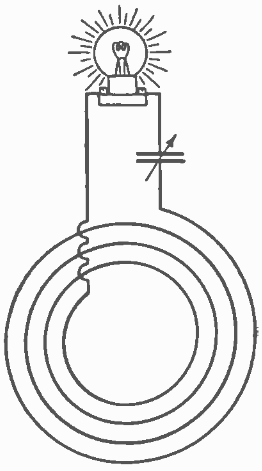

Fig. 2. More modern form of cymoscope.

D

D—The symbol of *electric displacement* (q.v.); also occasionally used as a symbol for diameter in electrical calculations.

DAMPED OSCILLATIONS — Electrical oscillations which die away, each succeeding oscillation in a group having lesser *amplitude* (q.v.) or strength than the preceding one. (See *Damped Waves.*)

DAMPED WAVES—Radio waves in the form of successive trains or groups, in each of which the amplitude or strength decreases with each successive wave. When electrical oscillations are caused by a single impulse, they do not continue indefinitely, but decrease in amplitude or die away. Thus, if oscillations are created in an antenna or other circuit by a discharge from a *condenser* (see spark discharge), each electric spark creates a train of oscillations which die away more or less gradually. Fig. 1 shows a pendulum which is given a momentary impulse. Now as this source of power is only momentary the pendulum will not swing from A to D indefinitely. If the first swing under the impulse carries the pendulum the entire distance A to D, the next swing will cover a lesser arc and so on until the movement has died out altogether.

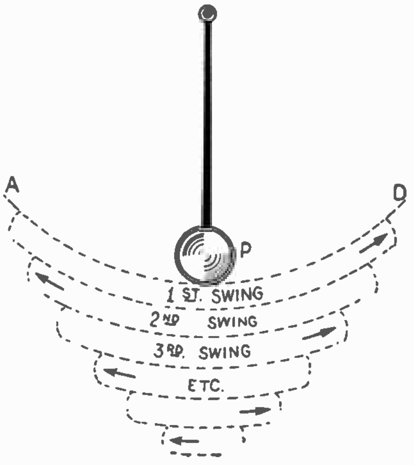

Fig. 1. Pendulum coming to rest after being given a momentary impulse.

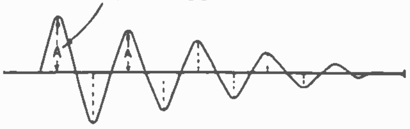

Fig. 2. Train of damped waves.

Now in Fig. 2, the graphic illustration of a train of *damped waves* shows that each crest is a little lower than the one preceding it and after a short period the waves have died out altogether. These curved lines may be considered as an illustration of the path taken by the pendulum in its motion as shown in Fig. 1. The decrease in amplitude of one crest over that preceding is known as the damping, and the naperian logarithm of the ratio of amplitude of any oscillation to the one preceding it is known as the *logarithmic decrement.*

If the impulse of force causing the oscillations is steady—not intermittent—the result will be a sustained series of oscillations that will continue at constant strength or amplitude until the exciting force is withdrawn. Thus, in the case of the pendulum, we can very easily keep exciting it by an evenly exerted force and so insure that the swings will continue without diminishing, as in the case of a momentary impulse. Oscillations or waves that have constant amplitude are known as *Continuous Waves, Undamped Waves* or *Sustained Waves.* (q.v.) (See also *Decrement, Logarithmic;* also *Naperian Base.*)

DAMPING—A progressive decrease in amplitude or strength. (See *Damped Waves, Damping in Instruments.*)

DAMPING COILS—*Short circuited* (q.v.) coils used in electrical apparatus for the purpose of causing damping (i. e., decrease in the intensity of electrical oscillations). The reaction of induced currents in the shorted coil creates the damping effect. (See *Damping.*)

DAMPING DECREMENT—See *Decrement, Logarithmic.*

DAMPING FACTOR—The product of the *logarithmic decrement* (q.v.) and the frequency (rate of alternation) of a damped alternating current. (See *Decrement,* also *Amplitude.*)

DAMPING IN INSTRUMENTS — The term damping is applied in a mechanical sense as in the case of the needle of a measuring instrument such as a *galvanometer.* (q.v.)

In a delicate instrument for measuring current the needle is controlled by the current through the instrument, and when this current is withdrawn the needle is brought back to the neutral position by a fine spring. Now, unless some precaution is taken the needle will have a tendency to swing or vibrate when measurement is being taken. Various devices are used to bring the needle quickly to rest without undue vibration. If the needle comes to rest almost immediately it is said to be a highly damped or *dead beat instrument,* because the oscillations or vibrations die out rapidly the same as in the case of damped waves.

There are two principal methods of damping instruments. When the instrument is of the moving coil type, the coil is generally wound on a light aluminum frame which is mounted between the poles of a permanent magnet. A soft iron core fills the gap between the two poles, allowing space for free movement of the coil. The result is that the metal frame carrying the coil moves through a very strong *magnetic field* and *Eddy currents* (q.v.) are set up in the frame and serve to damp its motion. The other common method is to have a light metallic vane carried by the pointer in its motion and moving in an air chamber to stop vibration. While damping is highly essential in a measuring instrument, too high a degree of damping is not advisable. If there is a very slight tendency on the part of the needle to vibrate, it permits the operator to determine whether it is swinging freely. (See *Voltmeter, Wattmeter.*)

DAMPING MEASUREMENT—See *Decrement, Logarithmic.*

DAMPING WAVES—Term commonly used to denote the gradual decrease in amplitude of a train of oscillations as in radio transmission. If the oscillations persist at constant amplitude or strength, each oscillation being equal in amplitude to the one preceding, there is no damping. When the amplitude of each successive oscillation is less than that of the preceding one there is said to be damping and the oscillations will die out entirely, at a rate dependent on the degree of damping. If the oscillations die out very rapidly they are said to be highly damped; if they die out slowly the damping is said to be feeble.

DANIELL'S CELL—A two fluid *voltaic cell* containing a zinc plate immersed in dilute sulphuric acid, and a copper plate in a saturated solution of copper sulphate; the two solutions being separated by a porous cup. This cell has a constant voltage and shows only slight polarization. Was formerly much used in telegraph work.

D'ARSONVAL GALVANOMETER—A type of galvanometer in which a coil carrying the current to be measured is suspended in a *fixed magnetic field.* (See *Galvanometer.*)

DASH—The long stroke used in *Morse* and *Continental Code.* It is considered as equivalent in length to three dots. (See *Code.*)

D. C. C.—The customary abbreviation for double cotton covered wire.

D. C.—The abbreviation commonly used for *direct current.* (q.v.)

"D" COIL—Occasionally known as Figure Eight coil. A form of inductance, used particularly in tuned radio frequency circuits. The windings are in the shape of a figure eight, either in pancake form (flat wound) or wound on a slotted tube. Owing to the peculiar arrangement of the windings these coils have very little external field and thus can be used in radio frequency amplifiers without the usual oscillating due to conflicting fields. (See *Field, Inductance, Radio Frequency Coils.*)

DEAD BEAT INSTRUMENT—A measuring instrument arranged so that the needle comes quickly to rest with little or no vibration is said to be *dead beat* or highly *damped.* (See *Damping in Instruments.*)

DEAD CELL—A *cell* which has been discharged. (See *Cell.*)

DEAD CRYSTAL—A crystal used in a *crystal detector* (q.v.) that has lost its sensitivity. Very often a crystal becomes dirty or oily from contact with the fingers. In some cases they can be brought back to usefulness by cleaning with alcohol, or if the crystal is dead through causes other than dirt on its surface, it may be broken to present a new surface and thus find a new sensitive spot.

DEAD END EFFECT—The effect on a receiving circuit (oscillatory circuit) resulting from unused turns of wire in a coil. In the customary tuning coil or tapped coil a certain portion of the wire is not in use and this idle section, being in metallic connection with the main portion, but not actually in use, acts as a miniature oscillatory circuit. This tends to reduce the efficiency of the tuning unit as a whole. Dead end effect may be eliminated by using the entire coil at all times, doing the tuning solely by the variable condenser, or it may be reduced by the use of a *dead end switch* (q.v.). (See *Tuning Coil.*)

DEAD END SWITCH—A switch used to cut out unwanted turns of wire in a coil in order to eliminate or reduce the *dead end effect* (q.v.). Such a switch is generally arranged so that instead of taps being taken from various points on the coil, the winding is actually in sections and the switch blade connects into the circuit the desired number of sections.

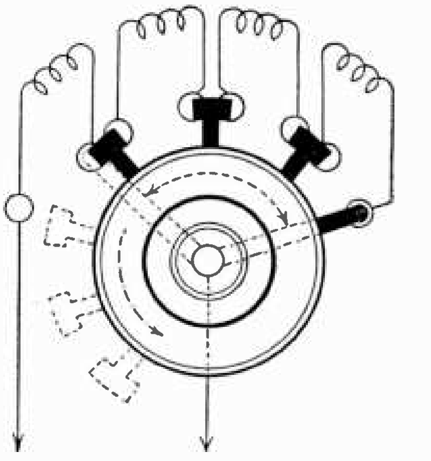

Dead end switch.

DEAD SPACE—In the case of *beat heterodyne* reception if the two frequencies are brought together or the difference is very slight there will be no beat. This space where the two frequencies are identical, or very nearly so, is known as the *dead space.* (See *Beats,* also *Heterodyne.*) The beat resulting from the superposition of two waves differing in frequency is the difference or the sum of the two frequencies. For instance, if two waves are present in the same circuit, of 1,000 and 1,005 kilocycles respectively, the two beat frequencies will be 5 and 2,005 kilocycles. If, however, the two frequencies are made identical, the difference between the two will be zero, and the sum of the two will be twice the original frequency. In the one case, where we take the difference of the frequencies, the beat note is zero, or there is no beat note at all; in the other case, where we add the two frequencies, the resulting frequency is the *harmonic* (q.v.) of the original frequencies, which is the same as an octave in music. In the present instance, this note will not be heard at all as it is above the range of audibility.

DECADENT WAVE—A damped wave. A wave, or oscillation, in a train of *damped waves* (q.v) which is of lesser amplitude or strength than the one preceding it. (See *Damped Waves,* also *Undamped Waves.*)

DECAY OF CURRENT—The gradual dying out of current in an *inductive circuit* (q.v.) after the *impressed potential* has been removed. The presence of inductance in a circuit causes the current to lag in time behind the voltage producing it. As a result, when a highly inductive circuit is broken by opening a switch or by other means, the current still tends to flow for a short interval, and the *induced voltage* in the circuit may become very high. This is the reason why a spark often occurs when the switch in a highly inductive circuit is opened. As a result the current decays gradually after the impressed voltage is removed.

It will be noted that the *decrement* (q.v.) or the rate of decay of the oscillations directly affects the number of oscillations in the wave train that will be useful in transmitting signals. For instance, if a certain oscillation in the figure should represent the smallest voltage that will operate a detector, it is evident that all oscillations in the wave train that have smaller amplitudes will be useless. Consequently, the lower the decrement of the wave, the greater will be the number of oscillations before the amplitude becomes too small for useful purposes. This likewise means that a greater amount of the energy in the wave can be utilized.

DECOHERER—In the type of detector, now obsolete, known as the *coherer*, the incoming signals caused the particles of nickel and silver filings in the glass tube of the device to cohere or cling together. (See *Coherer.*) In order to place the coherer in its original state after each impulse, it was necessary to have some means of separating the particles. These were known as *decoherers* and were in many different forms, some as separate attachments, others as a part of the detector or coherer. (See *Detector.*)

DECOMPOSITION, ELECTROLYTIC—The decomposition of chemicals taking place in a cell due to the action of the electrolyte. (See *Cell*, also *Electrolyte.*)

DECREMENT—A term used to indicate the rate of decay or dying out of an electrical oscillation that is subject to *damping.* (q.v.) In the illustration is shown a series of damped waves. It will be noted that the distance from zero to the crest of wave (C-2) is less than the one preceding it (C-1). This distance represents the amplitude and the ratio of any amplitude to the one preceding it is constant, that is the decay or decrease in amplitude is constant. The *Naperian Logarithm* (q.v.) of the ratio—of one wave to the one preceding it—is called the *logarithmic decrement.* The damping of a train of waves is an important consideration as it affects the tuning qualities to a great extent. This is covered by a U. S. Government Statute Concerning Transmitting Stations, decreeing that the logarithmic decrement per complete *oscillation* must not exceed 0.2, which means that for each single spark discharge from the transmitter there must be not less than 24 complete oscillations in the *antenna system.* (See *Oscillations, Damping, Naperian Logarithm, Sharp Wave.*)

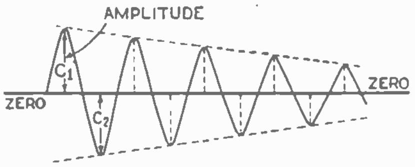

Damped waves gradually dying out.

DECREMETER—An instrument used to measure the *decrement* or degree of damping in an oscillatory current. Such meters are made in several different forms. Some types are direct reading in terms of the logarithmic decrement, while others require mathematical reduction from the readings of a dial. All decremeters operate on the basis of comparison of the resonance current in a tuned circuit with that in a circuit out of tune by a known or unknown percentage. (See *Logarithmic Decrement*, also *Damping.*)

DE FOREST, DR. LEE—Electrical engineer; b. Council Bluffs, Ia., Aug. 26, 1873. Grad. Yale (Sheffield Scientific), Ph. B., 1896; Ph. D., 1899. Inventor De Forest system wireless telegraphy and founder Am. De Forest Wireless Telegraph Co.; system originally adopted largely by U. S. Govt.; demonstrated to the British, Danish, German, Russian and Indian Govts., first used by the U. S. Signal Corps in the war manoeuvres and inst. in Alaska. 1905 invented the Audion or three-electrode vacuum tube. 1906-09 was the pioneer in Radio Telephony. First to broadcast music and opera. 1908 equipped the 24 battleships and destroyers of Admiral Evans for their historic round-the-world cruise. 1913 incorporated De Forest Radio Tel. & Tel. Co., which developed the audion amplifier and oscillator for commercial purposes. These De Forest devices completely revolutionized the radio art and made possible the Bell Tel. Transcontinental System. He was the father of radio broadcasting and all radio broadcasting apparatus is built upon his audion inventions. Since 1919 he has devoted his attention to the phonofilm, photographing sound waves directly on motion picture film. Pres. of the De Forest Phonofilms, Inc.; Dir. of De Forest Phonofilms, Ltd., London. 1921, Awarded Cross of Legion of Honor by the French Govt. in appreciation rendered by audion to the Allies during the war. 1922 awarded Elliott Cresson medal by Franklin Inst. 1922 awarded medal by Inst. of Radio Engineers. Fellow American Inst. of Elec. Engineers. Fellow Inst. of Radio Engineers. Mem. of N. Y. Elec. Society; mem. of Nat. Geog. Soc.

Dr. Lee De Forest.

DE FOREST COILS—A type of coil named after *Lee De Forest* (q.v.). The wires are wound in a diamond shaped pattern in such manner that the turns of one layer cross the turns of the preceding layer at an angle, thus making the *distributed capacity* (q.v.) very small compared with the capacity of other types of coils. (See *Honeycomb Coils.*)

DEGREES, ELECTRICAL—In an alternating current curve each complete oscillation or cycle is shown as two loops, one above the line representing time, and the other below it, the entire cycle being considered as 360 *electrical degrees.* It is possible in this manner to show relations between current and voltage phases, amplitude at any point, etc. (See *Alternating Current*, also *Sine Wave.*)

The idea of electrical degrees comes originally from the *elementary electric generator*, with two-slip-rings taking off the alternating current, and having only two poles. During one-half revolution of the armature turns, the current increases from zero to a maximum and then back to zero again. During the other half revolution the current increases in an opposite direction from a zero value to maximum and then back to zero again. The curve shown in the figure represents the current variation during the complete revolution, a crosssection of the rotating armature wire being shown at A. The angles indicated on the line BC are laid off to any convenient scale, and the intersection of the ordinates from these points with horizontals drawn from the circle as shown, give the points on the curve.

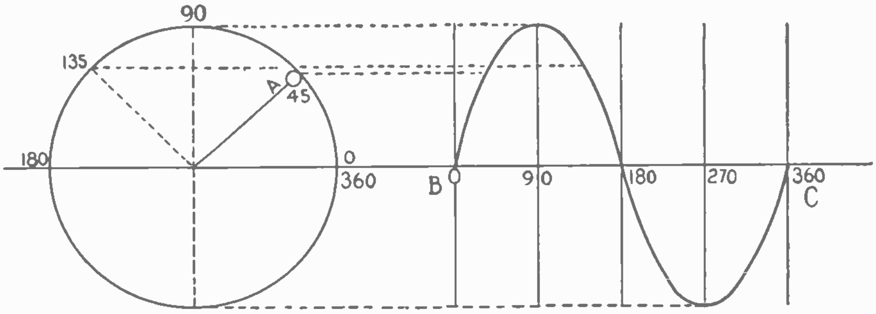

Curve representing current variation during one revolution of the armature of a generator.

The same ideas hold when the generator has more than two poles; in this case a full 360 electrical degrees is taken to represent the armature motion from across two adjacent pairs of poles, so that a complete cycle of voltage or current is generated. In a four pole machine, therefore, a full revolution of the armature could be represented by 2 x 360 or 720 electrical degrees. This corresponds to 2 complete cycles. Or, looking at it the other way, 360 electrical degrees would represent a half revolution of the armature wire.

DEIONIZATION—The action of returning a mass of gas that has become *ionized* (breaking up of the atoms of gas into positive and negative ions, being the elements to which and to whose motion, under the action of electric forces, is supposed to be due their electric conductivity) to its original state. In the case of a *spark gap* (q.v.) the ionization creates a conductive path which remains after the actual spark has passed, thus preventing sharply defined oscillations. There are various methods of deionization, including cooling, absorption, air-blast or magnetic field, all tending to deionize or place the spark gap in its previous condition to prevent an arc from following the discharge. This action is also known as *quenching.* (See *Quenched Gap.*)

DELTA—The fourth letter in the Greek alphabet. The capital letter delta (Δ) is used as a symbol for delta connection of *three phase alternating current generators or transformers.* The small or lower case delta (δ) is used as a symbol for *Logarithmic decrement.* (See *Decrement, also Delta Connection.*)

DELTA CONNECTION—Δ In a three phase *alternating current generator,* where three coils are mounted symmetrically around a shaft rotating in a *magnetic field,* the connection or grouping of the windings is called *delta* after the Greek letter which the grouping somewhat resembles. The illustration shows delta grouping in a three phase system. The sum of the instantaneous Electromotive Forces of the three coils is zero, or in other words the sum of the Electromotive Forces of any two coils is equal and opposite to that of the other coil, the *line voltage* thus being equal to the *phase voltage.* (See *Alternator, Polyphase.*)

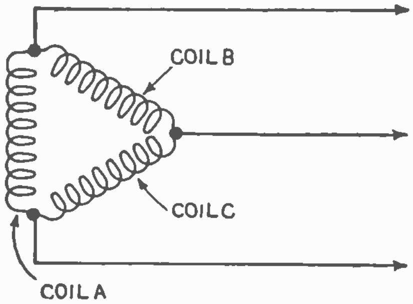

Delta grouping in a three phase system.

DEMAGNETIZATION—The act of returning a body to an unmagnetized state or of reducing the degree of *magnetization.* In the case of phones used in radio reception the permanent magnets are often weakened or *demagnetized* by connecting them in the *electron-tube circuits* in the wrong way. (See *Magnetization,* also *Telephone Receiver.*)

DENSITY, CURRENT—See *Current Density.*

DENSITY, FLUX—See *Flux Density.*

DENSITY OF ELECTROSTATIC CHARGE—The *electrostatic charge* per unit area. (See *Electrostic Charge,* also *Condenser.*)

DEPOLARIZATION—In an electric cell hydrogen bubbles form on the surface of the positive electrode and unless some arrangement is made to counteract this effect, the usefulness of the cell will soon be impaired due to the insulating action caused by this film of gas, known as *polarization.* Usually some means of oxidizing is employed to act upon the hydrogen as fast as it is produced on the positive electrode, thus preventing or reducing the effect of *polarization.* (See *Cell,* also *Polarization,* and *Depolarizer.*)

DEPOLARIZER—The oxidizing agent or other means used in a cell to counteract the effect of polarization. Bichromate of potash or peroxide of manganese is commonly used for this purpose. (See *Dry Battery,* also *Polarization.*)

DETECTOR—A device for converting oscillating currents of high frequency (radio waves) into a form suitable for operating a telephone receiver or sensitive measuring instrument. It is often referred to as a *rectifier,* because it serves to change the incoming currents from *alternating* to *pulsating direct currents. Detectors* vary in type and efficiency, ranging from the now obsolete *coherer,* to the modern sensitive *regenerative vacuum tube.* The different types of detectors are taken up under their various headings. (See *Crystal Detector, Vacuum Tubes, Electrolytic Detector.*)

DETECTOR CIRCUIT—That part of the circuit in a radio receiver which contains the *detector* (q.v.). The *detector circuit* may be closely coupled to the balance of the receiving circuit by means of a tuning coil or it may be inductively coupled by using a variocoupler or loose coupler. (See *Coupling*).

DETECTOR, DAMPED WAVE—Any special type of detector, such as a crystal, etc., used for reception of *damped waves.* (q.v.). (See *Tikker,* also *Heterodyne,* and *Vacuum Tube.*)

DETECTOR, VACUUM TUBE—The *vacuum tube* used as a *detector* of *radio frequency* (*high frequency*) *oscillations.* A vacuum tube may be used to change or rectify the high frequency alternating currents received at the aerial to *pulsating direct currents* (q.v.) capable of operating a telephone receiver or recording device. Formerly vacuum tubes were made for an individual purpose, such as *radio frequency amplifier, audio frequency amplifier, or detector,* the amplifier tubes being highly exhausted and the detector tube having a low vacuum and requiring critical filament adjustment. Of late tubes have been developed to such an extent that no special tube is necessary for the detector circuit, all being adaptable for almost any purpose. Fig. 1 shows the ordinary circuit for a vacuum tube detector. This method is not much used now since the advent of regeneration, except where stages of radio frequency amplification are used ahead of the detector to amplify the incoming signals. In this case incoming signals are impressed on the grid "G" and the slight variations of potential at this point control comparatively large currents at the plate "P," actuating the telephone "T." (See *Vacuum Tube, Theory of Detector Action.*)

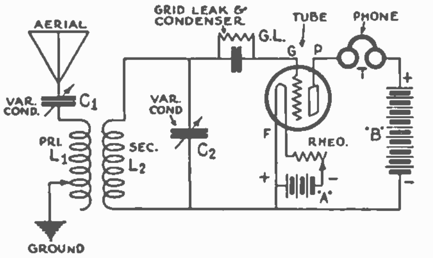

Fig 1. Non-regenerative vacuum tube circuit.

Used in this manner the tube is a considerable improvement over the crystal detector in point of sensitivity. However, by means of regeneration, the sensitivity may be increased many times. Fig. 2 shows a simple *regenerative circuit.* Here the unrectified currents passed through to the *plate* of the vacuum tube are fed back by means of the coil L2 and by being impressed again on the *grid,* further amplification can be attained. (For complete explanation see *Regenerative, Theory of Operation of Vacuum Tube, Detector, Pulsating Direct Current,* also *Crystal Detector.*)

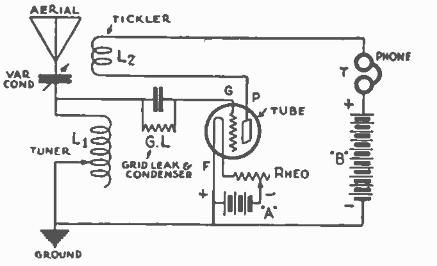

Fig. 2. Simple regenerative circuit.

DETUNING—The opposite of *tuning* (q.v.) The process of varying the effective inductance or capacity or both, to throw the radio receiver out of resonance with the particular signals to which it is tuned. This may be done to decrease the volume of the signals or it may be employed to eliminate or reduce interference from some undesired signals. For example, a station operating with a wavelength of 400 meters might interfere with reception of a station operating on 360 meters, providing it has sufficient power, or the receiver is not particularly selective. If the radio receiver is detuned slightly below 360 meters, the desired signals will still come in, though at lesser volume, but the interfering signals will be lost entirely. Tuning is the act of *producing* resonance and detuning is the act of *destroying* resonance. In the case of a *heterodyne receiver* (q.v.) the heterodyne circuit is slightly detuned from the incoming oscillations, thus producing a difference in frequency or *beat frequency* (q.v.). (See *Resonance,* also *Dead-space.*)

DIAGRAM—A system of lines drawn to represent the circuit or connections for radio receivers or transmitters or associated apparatus, etc. Diagrams may be in any of several forms, the two common methods being known as schematic and perspective. In the schematic form, symbols are generally used to represent the various pieces of apparatus, whereas in the perspective form, drawings of the apparatus are used. Fig. 1 illustrates a typical schematic diagram of a simple crystal receiver. Here the tuning coils, de-

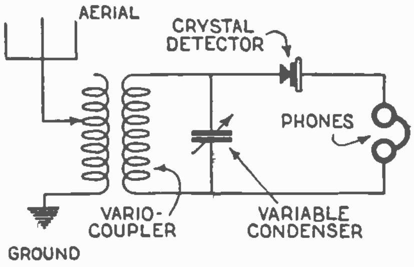

Fig. 1. Schematic diagram of a simple crystal receiver.

tector, phones and so on, are shown in the form of symbols. Fig. 2 shows the same circuit in perspective form, with the tuning coil and other parts actually pictured. Diagrams may be used to show connections for any electrical or radio circuits or for individual parts of instruments. (See *Hookup.*)

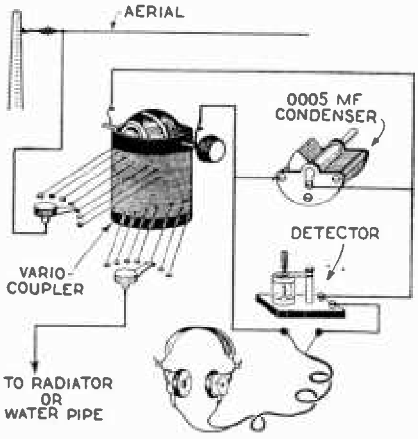

Fig. 2. Showing in perspective form the circuit given in Fig. 1.

DIAL—Devices used to control the movement of condensers or other moving parts of radio apparatus. Most dials are made of a composition such as Bakelite or hard rubber; some are made of metal. The periphery is generally spaced off in numbers from 0 to 100 or 180, for purposes of keeping a record of the positions at which the various stations are received. Dials may be arranged to have the readings refer to definite values by means of *calibration* (q.v.). (See *Dial Vernier.* See *Log.*)

DIAL, VERNIER—A term incorrectly applied to a dial arranged with slow-motion linkage in order to permit very fine adjustment. The illustrations show two types of vernier dials. Vernier dials are generally used on *variable condensers* and occasionally on other instruments, where close adjustment is necessary. They are particularly useful in the case of a *Super-Heterodyne receiver*, the oscillator control being critical, and the wavelength control as well, in certain cases. (See *Dial*, also *Critical.*)

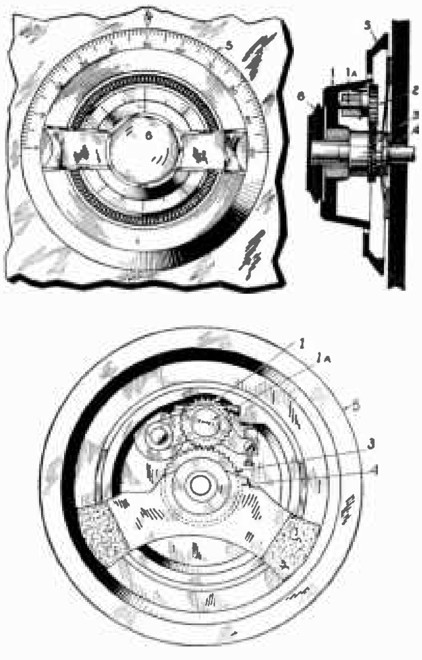

This dial utilizes the slip tooth principle. 5 is the main dial. 6 turns the slip gears (3) and (4), which cause 2 to turn. 2 is mounted on the shaft (1A) fastened in the main dial (1). 5 is the scale, fastened to the condenser shaft.

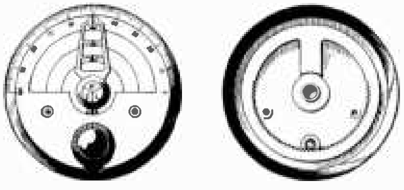

A simple vernier dial using a pinion and internal rack.

DIAMAGNETIC MATERIAL—Substances not readily susceptible to *magnetism.* If a substance does not pass magnetic lines of force, but rather is repelled by a magnetic pole, it is said to be *diamagnetic;* its permeability, or ability to permit magnetism to pass through it, is considered negative or less than unity (air). Diamagnetic substances are apparently repelled from the poles of a magnet. If placed in a magnetic field they have a tendency to diminish the *magnetic induction* (q.v.). (See *Paramagnetic*, also *Permeability.*)

DIAPHRAGM—A thin disk, generally of soft iron used in *telephone receivers* and *microphones* (q.v.) to either produce or detect electrical pulsations by means of vibrations caused by an *electromagnet.* Variations in the current through the electromagnet cause corresponding vibrations in the diaphragm. While the common type of diaphragm is made of soft iron, there are numerous other materials used. Certain alloys have higher permeability, and therefore are more responsive to the attraction of the magnet. In one type of telephone receiver that is particularly sensitive, there are two diaphragms—one that is actuated directly by the electromagnet under the influence of a varying current and another of thin mica, coupled to it by a fine wire. Another type employs a small iron armature suspended on a pivot, and attached to a mica diaphragm by means of a delicate lever, at right angles to the armature. The vibrations of the armature in the magnetic field of the magnets are communicated to the diaphragm by means of this lever. (See *Baldwin Receiver.*) Another type is of metal with fine corrugations and still another is a mica or parchment disk with a piece of soft iron in the center. (See *Microphone, Telephone Receivers.*)

DIAPHRAGM, CARBON—A phone diaphragm in the form of a thin carbon disc. (See *Diaphragm.*)

DIELECTRIC—This term is rather broadly used in electrical work to indicate a non-conductor or *insulator.* A dielectric may be a solid, liquid or gas and is generally employed to separate two conducting surfaces. The most common instance of the application of a *dielectric* is in the case of a condenser. Here the plates are separated by a dielectric material, the nature of this material depending on the particular type and purpose of condenser. A *variable condenser* for receiving purposes generally employs air as the dielectric. That is, the two sets of plates are kept from actual contact while one is being rotated, by an air space. In *fixed condensers* the dielectric is usually waxed paper, mica, glass, oil or compressed air, depending on the purpose for which the condenser is to be used and the dielectric strength necessary. While we refer to a dielectric as a non-conductor or insulator, there is actually no dividing line between conductors and insulators, practically any dielectric permitting passage of a certain amount of current under the proper conditions. If a substance is a very good insulator it is said to have good dielectric properties, while one that is a poor insulator and permits leakage of current is a correspondingly poor dielectric. (See *Insulator, Conductor, Resistance,* also *Dielectric Coefficient and Constant.*)

DIELECTRIC ABSORPTION—The tendency of the dielectric in a condenser to apparently absorb a certain amount of the power applied to it. When a condenser is charged by a direct current source such as a battery, the instantaneous charge is often followed by a small charge or flow of current that steadily decreases as it flows into the condenser. This additional charge is in effect absorbed by the dielectric of the condenser. The reverse effect is obtained when the condenser is discharged, the instantaneous discharge being followed by a steadily decreasing additional discharge. It will be apparent, then, that if the condenser is charged from an *alternating current* source, such as by *radio frequency currents*, the dielectric will show a tendency to withhold a certain part of the charge, and as the charging and discharging goes on rapidly due to the rapid reversals of the current (see *High Frequency Alternating Current*), the dielectric will continue to hold back this portion of the power as long as the charging and discharging process persists.

This loss, due to the absorption in the dielectric, must not be confused with the losses due to faulty dielectric materials, wherein a certain portion of the current is allowed to leak away by conductance through or on the surface of the insulator. *Dielectric absorption* is often referred to as *dielectric viscosity* or *hysteresis* because of its similarity to viscosity in liquids. Viscosity is the property possessed by liquids to resist deformation. The usual method of measuring viscosities is by measuring the time taken by a known volume of the liquid, at a known temperature, in flowing through an aperture of known form and dimensions under a known pressure. Thus tested, water will flow rapidly, while cylinder oil is very sluggish, and hence is said to possess great viscosity. (See *Condenser*, also *Dielectric.*)

DIELECTRIC COEFFICIENT and CONSTANT—The specific *inductive capacity* of a dielectric. Generally speaking, its properties to act as a *dielectric.* In the *Centimeter Gram Second system* (q.v.) the *inductive capacity* and the *dielectric constant* are numerically equal, the constant being the dielectric value of the material as compared with air at ordinary pressure taken as the standard (1). The table of constants for the more important dielectric materials follows.

It will be seen that glass, oils and mica have the highest values, for which reason they are widely used. (See *Specific Inductive Capacity, Dielectric,* also *C. G. S.*)

TABLE OF DIELECTRIC CONSTANTS

Dielectric	*Constant*
Air at ordinary Pressure (Taken as the Standard)	1.000
Manila Paper	1.50
Celluloid	1.555
Parrafine (clear)	1.68 to 2.32
Beeswax	1.86
Parrafine Wax	1.9936 to 2.32
Parrafined Paper	3.65
Hard Rubber (Enonite)	2.05 to 3.15
India Rubber (pure)	2.22 to 2.50
Gutta Percha	2.46 to 4.20
Shellac	2.74 to 3.60
Olive Oil	3.00 to 3.16
Glass (Low Frequency value)	3.25 to 4.00
Glass (High Frequency value)	4.21
Mica (Pure Sheet)	4.00 to 8.00
Porcelain	4.38
Castor Oil	4.80
Flint Glass, very light	6.57
Flint Glass, light	6.85
Flint Glass, very dense	7.40
Flint Glass, double extra dense	10.10

DIELECTRIC CONSTANTS—See *Table of Constants* under *"Dielectric Coefficient* and *Constants."*

DIELECTRIC HYSTERESIS—When the *electric field* in a dielectric material is varied rapidly, as when the condenser is charged by *high frequency currents*, heat may be generated. This is due to *dielectric hysteresis which is* synonymous with *dielectric absorption* (q.v.)

DIELECTRIC STRENGTH—When an electric field is established in a dielectric, that is to say, a charge is applied to a condenser, and the field attains a certain intensity, the dielectric ceases to be an insulator and becomes, in effect, a conductor. This condition generally is accompanied by a puncturing of the dielectric material. When the charge applied to a condenser is too high the spark will burn through the dielectric and the condenser is then said to have broken down. In some cases the condenser may be permanently ruined by this action, and in others, the dielectric may be conductive at a certain voltage, while remaining effective as long as the voltage is held below the safety point. The particular voltage (critical voltage) or field intensity at which the breakdown of the dielectric occurs is called the dielectric strength of the material. (See *Dielectric*, also *Table of Dielectric Strength.)*

DIFFERENCE FREQUENCY—The frequency of oscillations produced by superposing oscillations of one frequency on oscillations of a different frequency. In Super-heterodyne receivers the action is based on the production of a difference or beat frequency. If the incoming oscillations from the antenna are combined with oscillations of a different frequency produced locally in the receiver, a new series of oscillations will be produced, these oscillations having a frequency numerically equal to the difference in frequency between the other two sets of oscillations. (See *Beat Frequency*, also *Super-Heterodyne.)*

DIODE—A thermionic vacuum tube having only two *electrodes*, namely *filament* and *plate*. The original vacuum tubes were known as *Fleming Valves* and contained a hot filament and cold plate but no grid as found in the three element vacuum tube now in general use. Several types of *diode* are still in use as detectors, their chief value lying in the fact that they require no careful adjustments as in the case of a *crystal detector*. (See *Triode*, *Vacuum Tube*, also *Fleming Valve.)*

DIPLEX RECEPTION OR TRANSMISSION—The simultaneous reception or transmission of two series of signals by or from a single operating station. The systems are so arranged that two messages may be sent or received at the same time without interfering with each other. (See *Duplex Signalling.)*

DIRECT CONDUCTIVE CIRCUIT—Any circuit having a direct conductive path and not depending on capacity or electromagnetic coupling. A metallic, conducting circuit.

DIRECT COUPLING—The coupling or relation between two or more coils or circuits, wherein the connection is *metallic*. In the illustration Fig. 1 is shown a two slide tuning coil. In this case A-B is the primary or aerial circuit and C-D the secondary or detector circuit. The two circuits are joined together by metallic connection.

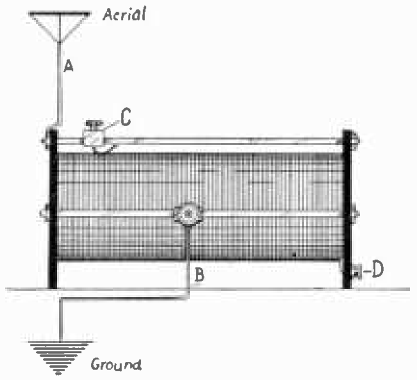

Fig. 1. Double slide tuner which affords direct coupling between the primary and secondary circuits.

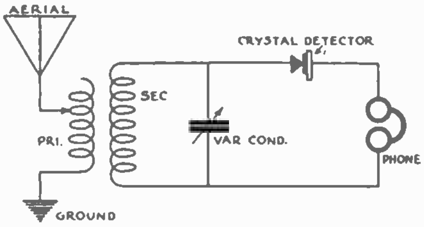

Fig. 2. Circuit in which primary and secondary are inductively coupled.

The illustration Fig. 2 shows essentially the same circuit, except that the primary and secondary are two separate coils. Here the two coils or circuits are joined by *inductive* coupling.

DIRECT COUPLED—Circuit in which primary and secondary circuits are metallically connected. Generally permits little selectivity. For this reason, *inductive coupling* is more generally used to permit close tuning or selectivity (q.v.). (See *Coupling, Inductive Coupling.)*

DIRECT CURRENT—See *Current, Direct.*

DIRECT READING—A term applied to various measuring or recording instruments which are so arranged as to show values directly without requiring mathematical reduction. *Voltmeters, Ammeters* and such instruments, used widely for showing the *voltage, current*, etc., of *batteries* or *cells*, are direct reading, showing the values directly in *volts* or *amperes*. A *decremeter* or *wavemeter* very often shows a reading that requires application of a formula to produce the desired quantity. (See *Calibration.)*

DIRECTION FINDER—See *Radio Compass.*

DIRECTIONAL—The effect of an aerial wherein waves are transmitted better, or entirely, in a certain direction, or received from a certain direction, depending upon the type and direction of aerial. (See *Transmitting Aerial*, also *Receiving Aerial.)*

DIRECTIONAL EFFECT OF ANTENNA—An effect of an antenna (aerial) for reception or transmission wherein signals from a certain direction are more readily received than from another direction, or in the case of transmission, the range being greater in a certain direction.

DISCHARGE—Generally speaking, a comparatively sudden passage of electricity. The term however, as applied to a *storage battery* will mean merely the effect of releasing the *electrical charge* stored up in it, in this case the discharge not being necessarily rapid. A *condenser* has the ability to hold charges of electrical energy and to release them suddenly when the proper contact is made. The term "discharge" as applied to a *transmitter* indicates the passage, usually in the form of a spark or succession of sparks, of the electrical energy across a gap between *electrodes*. The illustration shows spark discharge from spark coil. (See *Spark Discharge, Storage Battery*, also *Discharger.)*

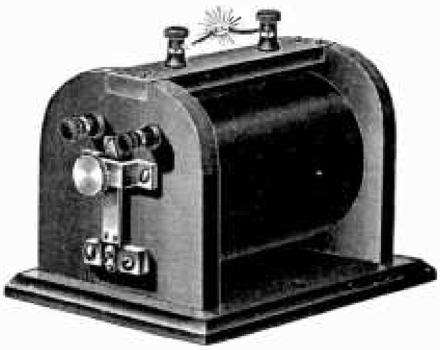

Discharge from a spark coil.

DISCHARGER—Any device allowing a path for an electrical discharge. The term may be considered roughly a synonym for *spark gap*. Usually two or more *electrodes*, either stationary or rotary, spaced a short distance apart to permit the released electrical energy to bridge the space and thus complete an *oscillatory circuit*. (See *Disc Discharger Rotary Gap.)*

DISC DISCHARGER—A form of *discharger* employing one or two rotating discs carrying the sparking surfaces. With the advent of modern undamped transmission methods, these dischargers are gradually passing out. (See *Synchronous Discharger.)*

DISPLACEMENT CURRENT—A current which flows for a short interval in an insulating material or *dielectric* when an *electromotive force* is impressed across it, or when the intensity of the electromotive force impressed across it is increased or decreased. This *displacement current* will flow only when the impressed electromotive force is altered in intensity. After the initial current due to the sudden change in the electromotive force being impressed across the material, the dielectric material or insulating material will remain in a state of strain as long as the *charging force* persists without further change in intensity, and no further displacement current will flow. (See *Current*, also *Dielectric.)*

DISPLACEMENT, PHASE—See *Phase Displacement*, also *Phase Angle.*

DISRUPTIVE VOLTAGE—The voltage sufficient to disrupt or break down a sample of *dielectric* material under given conditions. Known also as *breakdown potential.* For example, if a *condenser* will stand an impressed potential or voltage of 500 volts without injury, but will be punctured by a potential of 1000 volts minimum, the disruptive voltage is said to be 1000. (See *Break Down Potential.*)

DISSOCIATION THEORY—A theory advanced by Arrhenius in 1887, explaining *electrolytic conduction* by the assumption that a substance in solution is dissociated or separated into *positive and negative ions*, these ions carrying their respective charges in opposite directions. (See *Electrolytic Action.*)

DISSONANCE—"Discord; disagreement"—Webster. The antonym of *resonance.* A term broadly applied in radio to indicate lack of resonance, or coordination of signals or impulses. When an alternating current is superimposed on another alternating current of different frequency, the resulting lack of resonance—*dissonance*—is known as a *beat.* (See *Distortion.*)

DISTORTION—Lack of purity or faithfulness in the reproduction of a vibration or series of vibrations. The most common application of the term in broadcast reception is in the case of reproduction by a loud-speaker. It will very often be found that music or speech is not perfectly reproduced, due to any of a variety of causes. Vacuum tubes may themselves cause distortion; too high potential applied to the plate of a tube may result in distortion; many transformers used in audio frequency amplifier circuits may distort notes of certain frequencies, or it may be due to *self-oscillation* or *regeneration* in the receiving set. This effect is also caused by inefficient transmitting (broadcasting) apparatus, or by lack of careful adjustment. The transformation of speech or music into electrical impulses and its subsequent transmission and reception are attended by many difficulties.

The control of wave form in broadcasting is known as *modulation.* (q.v.). If the transformation of speech or music into electrical impulses and its propagation into space is accomplished without materially changing or distorting the wave form from the original voice or music vibrations, the problem of accurate reproduction is entirely dependent on the receiving apparatus. Distortion may often be traced to the diaphragm of the loud-speaker, certain types being more efficient in this respect than others. (See *Loud-speakers.*) If the ordinary type of disc diaphragm is used, it may become bent and thus cause inequalities in the vibrations, producing distortion of the music or speech being reproduced. (See *Wave Form, Modulation,* also *Amplifier.*)

DISTRESS SIGNAL—At the International Radio Telegraph convention held at Berlin, Germany, in July, 1908, the call letters C Q D, established by the Marconi Co. in February, 1904, as the official distress call, were superseded by the letters S O S as the Marine distress signal.

S O S is the International distress call for ships and airships requiring assistance. The letters have no particular significance, being chosen mainly for their distinctive sound. In the International Morse code, S O S is composed of three dots, three dashes and three dots, thus: . . . — — — . . . an unusual combination which permits easy recognition among other messages and calls. When distress signals are heard, the nearest government transmitting station generally sends out immediate notification to all broadcasting stations in the vicinity to suspend operations until the ship has been located and assistance rendered.

DISTRIBUTED CAPACITY—The condenser effect in a coil of wire. Any coil of wire possesses *inherent capacity* to a certain extent, depending on its particular shape, size, etc. In the case of a coil of wire wound on a cylindrical form as indicated in the illustration Fig. 1, the passage of current through the coil sets up a *local magnetic field*, but also an *electrostatic field*, the latter being in the form of lines of force perpendicular to the conductor. Thus in effect, each two adjacent turns of wire act as a miniature condenser (Fig. 2).

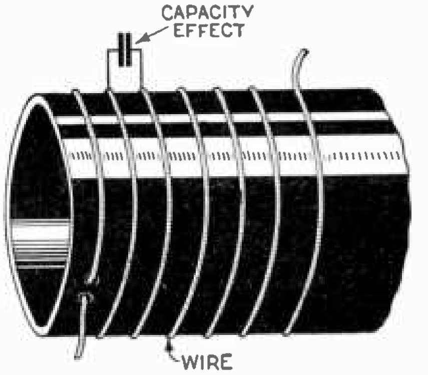

Fig. 1. Coil showing effect of distributed capacity between winding.

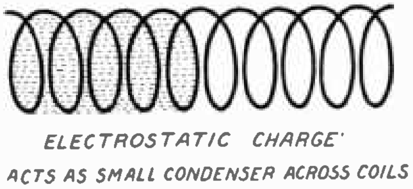

Fig. 2.

There are numerous methods and forms for winding coils in order to do away as much as possible with this effect. (See *Low-Loss Coils.*)

DISTRIBUTED INDUCTANCE—In a long cable or any great length of wire used for electrical purposes, additional inductance (wire) distributed throughout the entire length to compensate for the inherent capacity of the line. (See *Distributed Capacity.*)

DOT—The short signal in the International Morse Code. (See *Code.*)

DOUBLE AMPLIFICATION CIRCUIT—Also known as *dual amplification circuit.* The arrangement whereby a vacuum tube or several tubes may be made to do double duty, acting as both *radio frequency* and *audio frequency amplifiers.* (See *Reflex.*)

DOUBLE COTTON COVERED WIRE—Abbreviation D. C. C.—Cotton covered copper wire widely used in radio work for coils and all forms of inductances. The insulating covering is composed of two distinct layers of cotton, wound on in opposite directions to prevent loosening. Such wire is obtainable in a variety of sizes according to the standard gauges. (See *Wire Gauge.*)

DOUBLE FREQUENCY OSCILLATIONS—Sometimes referred to as a *double humped wave.* An irregular wave resulting from two frequencies, generally due to too close coupling of the *oscillation transformer.* When the coupling between the open and *closed oscillatory circuits* is too tight, the *open circuit* very often oscillates at two frequencies, resulting in the radiation of an irregular or *impure wave.* The illustration Fig. 1 shows in graphic form, such an irregular wave. The two peaks of resonance explain the term double humped wave. Fig. 2 shows another curve illustrating two frequencies.

In this case however, the amplitude of the second wave is considerably less than that of the first or main wave. According to U. S. Government regulations, if the amplitude of the lesser wave is not more than 0.1 of the amplitude of the main wave, it is said to be pure. In other words, where there are two frequencies, but the strength of one less than one tenth that of the other, it is considered negligible as it cannot be expected to cause any appreciable interference. (See *Resonance, Amplitude, Decrement,* etc.)

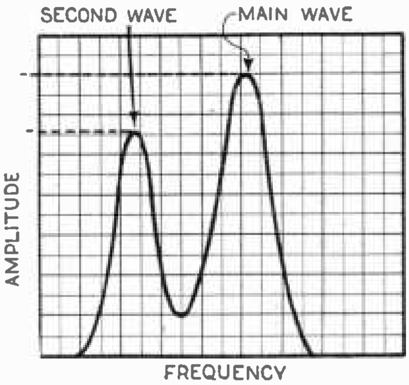

Fig. 1. Resonance curve showing second wave of interfering nature.

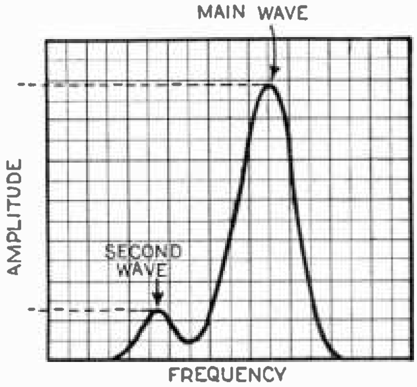

Fig. 2. Resonance curve showing second wave of negligible proportion.

DOUBLE GRID TUBE—A vacuum tube having two distinct *grid* members in addition to the usual *filament* and *plate* elements. Such a tube is used in circuits where the customary "B" battery is not employed, the extra grid acting as a booster for the electron flow, the "A" battery furnishing a small positive potential to the plate. (See *Solodyne,* also *Theory of Vacuum Tubes.*)

DOUBLE MODULATION—Successive modulation of a *radio frequency alternating current* at two lower frequencies. The *intermediate frequency* is usually above *audio frequency* and the lowest frequency is customarily within audio limits or a combination of audio frequencies, as in the case of *radio* *telephony* (broadcast). (See *Modulation.*)

DOUBLE POLE SWITCH—A switch used in electrical practice and in radio installations having two poles or connections, thus permitting both sides of a circuit to be opened or closed simultaneously.

DOUBLE RANGE METERS—Meters used for electrical measurements, arranged to read to two scales. As an example, the voltmeter shown has two scale readings, one in fine degrees for a maximum of 7.5 volts, the other a coarser scale to a maximum of 150 volts. Three binding post connectors are furnished, one being a common positive for both scales. A meter for almost any measurement might be made to have double range, although such instruments are generally confined to measurement of current in amperes and pressure in volts—i. e., ammeters and voltmeters. (See *Meter.*)

Photo by Courtesy of Weston Electrical Instrument Co.

Double range meter, one range to 7½ volts, higher range to 140 volts maximum.

DOUBLE SLIDE TUNER—A tuning coil provided with two sliding contacts, generally used in crystal receivers.

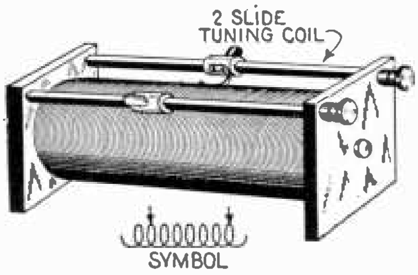

Two-slide tuner.

The illustration shows a common type of two slide tuner. (See *Coupling Crystal Receiver*, also *Tuning.*)

DOUBLE THROW SWITCH—A switch so arranged that a circuit or a certain instrument is connected in either of two different positions by throwing the switch lever. The illustration shows a single pole double throw switch. (See *Switch.*)

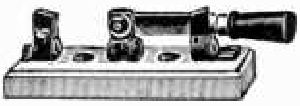

Double throw single pole switch.

DOWN LEAD—More commonly known as *lead-in.* The wire running from the aerial to the receiving or transmitting apparatus, whereby the signal energy is either collected from, or fed to, the aerial. The illustration shows a typical down lead or lead-in for a receiving set. Obviously the chief difference between such a lead for receiving purposes and for connecting a transmitter to the antenna is the matter of insulation. The comparatively feeble impulses received or collected by the aerial when receiving require only ordinary care, whereas the high voltages used in transmitting require careful insulation and much heavier apparatus all around. (*For more complete details regarding insulation, etc., see Aerial.*)

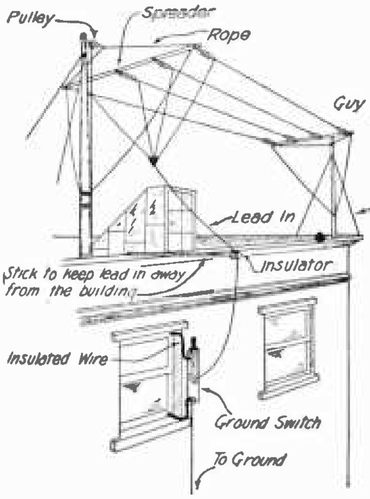

A typical 4-wire aerial showing lead in and method of installing lightning switch.

DRIFT, AVERAGE ELECTRON—The assumed rate of flow or drift of *electrons* under average or specific conditions. (See *Electronic Flow.* Also *Current, Assumed Direction of Flow.*)

DRIVER—A term broadly used to denote any system used to produce *oscillations* (vibrations) of a local nature. More specifically any means of producing oscillatory currents as used to test or make measurements in radio circuits. A buzzer connected inductively to a circuit to produce oscillations in that circuit for the purpose of making measurements of *capacity, wavelength,* etc. Another means might be a circuit involving a vacuum tube and the necessary apparatus to make it oscillate at a given or variable frequency. The term may be used to signify the tube circuit used to produce local oscillations in the case of a *super-heterodyne* (q.v.). Here the driver produces a series of oscillations of a frequency different from the incoming oscillations, the difference between the two being known as the beat frequency. (See *Heterodyne, Local Oscillations, Buzzer, Excitor.*)

DRIVER CIRCUIT—The circuit of the apparatus used to produce oscillations for purposes of test or measurement, or, in the case of a super-heterodyne, to produce a beat effect. (See *Driver.*)

DRUM ARMATURE—A form of armature winding in the approximate shape of a drum. (See *Armature.*)

DRY BATTERY—A battery or group of cells not employing a liquid electrolyte, the cell being filled with a mixture of carbon, manganese dioxide and sawdust (or other absorbent) saturated with a solution of sal ammoniac, these forming a paste, as distinguished from Wet Battery. (See *Battery,* also *Cell* and *B Battery.*)

DRY CELL—See *Dry Battery.*

D. S. C.—The abbreviation for "Double Silk Covered" as applied to copper wire having two distinct layers of silk, wound generally in opposite directions. This wire is furnished in standard sizes according to the various wire gauges. (See *Brown & Sharpe Gauge.*)

DUAL AMPLIFICATION—The process of obtaining both *radio frequency* and *audio frequency amplification* from the same tube instead of using two separate tubes. (See *Double Amplification Circuit,* also *Reflex.*)

DUBILIER, WILLIAM—President and technical director; born, New York, July 25, 1888. Educated New York Schools, Technical Inst. and Cooper Union; Chief Engineer of Continental Wireless Tel. & Tel. Co.; 1910 President of Com. Wireless Tel. Co.; at present Technical Director of Dubilier Condenser and Radio Corp. of New York, The Dubilier Condenser Company Ltd. of London, The Deutsche Dubilier Kondensator Gesellschaft in Berlin, La Protection Electrique Capart-Dubilier in Paris. Has obtained over three hundred patents and applications of electrical devices, which have been purchased or licensed by many companies. Member American Institute Electrical Engineers, Inst. of Radio Engineers, Honorary Member of Societe Academique D'Historie Internationale. Member of Royal Society of Arts.

DUCON—Trade name of a device which can be fitted to any electric light socket and made to serve in place of the usual outdoor aerial. This attachment consists essentially of two condensers, so arranged as to prevent the passage of any direct current or low frequency alternating current from the lighting main to the receiving set, but at the same time to permit passage of the incoming radio signals, the electric light line acting, therefore, as an aerial. (See *Adapter, Aerial.*)

DUDDELL SINGING ARC—Also called *Musical arc.* An *arc* actuated by a source of direct current through a *resistance* and shunted by a *condenser* and *inductance* in series. An *oscillating current* is thus produced in the condenser circuit, the result being a singing note corresponding in pitch to the frequency of the oscillations in the condenser circuit. The schematic arrangement is shown by the illustration. (See *Arc Generator.*)

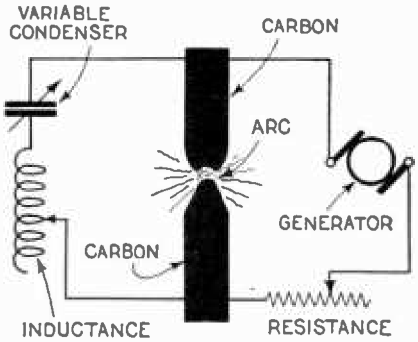

Schematic arrangement for operation of Duddell singing arc.

DULL EMITTER—The English term for vacuum tubes having *thoriated filaments* and operating with low current consumption. While the term is generally used to denote the tubes requiring only one or two dry cells to operate, it may apply as well to storage battery tubes where the current requirements are low. (See *Filament, Thoriated.*)

DUOLATERAL COILS—An alternate term for *honey-comb* coils; inductances wound in diamond-shaped layers to reduce the distributed capacity effect. (See *Honey-Comb Coils.*)

DUPLEX SIGNALLING or DUPLEX TELEGRAPHY OR TELEPHONY—The simultaneous transmission of signals or telephony in both directions between two stations. (See *Diplex Reception or Transmission*, also *Telephony.*)

DUST CORE—A form of core for certain types of transformers, employing iron filings or dust in place of the customary iron wire or laminations. Iron dust is used occasionally where simplicity of construction is desired, the assembly of such a core being obviously much easier than with wire or laminated types. (See *Core.*)

DX—The popular term for long distance; referring to the transmission to or reception from distant points of radio signals or broadcasting.

DYNAMIC CHARACTERISTIC — The curves obtained by impressing an alternating E. M. F. (electromotive force) on the grid of a vacuum tube, as distinguished from the curves obtained by application of a steady direct potential. When alternating potentials such as incoming signals are impressed on the grid of a tube, the curve resulting is apt to take an entirely different form from that of the curve showing *static characteristics* (q.v.). (See *Vacuum Tube characteristics.*)

DYNAMIC CONDENSER—A term occasionally applied to a *synchronous motor* used to improve power factor (in alternating current circuits, the ratio of the electric power in *watts* to the apparent power in *volt amperes* is known as the power factor). Used in this manner the motor has the effect of advancing the current phase in a manner similar to that of a condenser. (See *Angle of Lead or Lag.*)

DYNAMIC ELECTRICITY — A term sometimes used for electric currents to distinguish them from *static electricity* (q.v.).

DYNAMO—A machine for converting mechanical energy into electrical energy. Is also known as *generator*, but usually confined to machines for generating *direct* current. The action of a dynamo is based on the production of an *electromotive force* in a conductor moving in a magnetic field. Fundamentally a dynamo is a machine that generates alternating current, but instead of collecting this current by means of rings, which would cause alternating current to flow in the external circuit, a device known as an *armature* is employed in such a manner that the current collected by brushes and delivered to the external circuit is direct—that is, it flows more or less steadily in one direction. Thus the kind of current delivered for use depends on the method used to collect the currents generated. (For more complete explanation of the production of alternating current see *Alternating Current, Theory of Production.*) In the case of a dynamo, the essential parts are the *field magnets*, the *armature* and the *commutator*. Now in the case of an *alternator*, each end of the loop or armature is connected by a brush bearing against a collector ring and the end of each loop is always connected to the same brush. This *results in an external current that changes its direction with each reversal of direction of the induced electromotive force.* If the external current is to be direct it is necessary to have some means of collecting the current through one brush at the positive instant and the other at the negative instant. This is done by means of an armature, which is in effect a switching arrangement, so designed that it will reverse the connections of the external circuit at the instant of each reversal of current in the armature. The illustration shows a *shunt* wound dynamo. MM are the field magnets, FF the field windings, BB the brushes, A the armature and C the commutator. (See *Alternator, Armature, Generator.*)

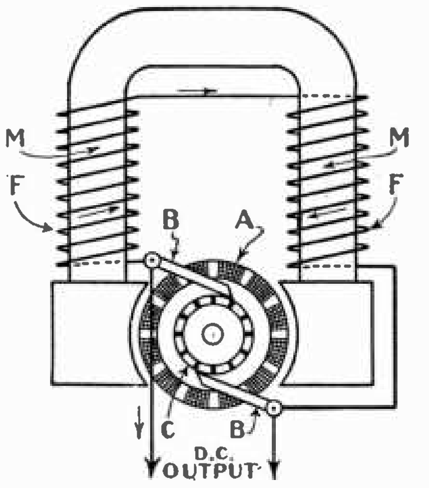

Shunt wound dynamo.

DYNAMOMETER—A delicate and accurate instrument for the measurement of currents and voltages or both. Such instruments depend upon the action of a circuit carrying current upon another carrying the same current. Essentially it comprises two coils, one fixed and the other movable. This system of measurement is used in a *wattmeter*, where it is necessary to measure the instantaneous current and voltage, i.e., power. (See *Wattmeter*, also *Electro Dynamometer.*)

DYNAMOTOR—A direct current machine which acts either as a *motor* or *dynamo*. It has an armature with two separate windings and two separate commutators, one at each end of the armature. Either winding may be used as a motor and the other as the generator winding. Such a machine is used to convert high voltage direct current into low voltage direct current or vice versa, thus performing the same function with direct current as a power transformer performs for alternating current. (See *Converter.*)

DYNATRON—A form of vacuum tube generally used for producing oscillations as in radio telephony, wherein the phenomena of "*secondary electron emission*" is used. In this form of tube the electrons traveling at high speed from the hot filament or cathode are made to collide with a metallic surface. The collision of these electrons on the surface of the interposed element has the effect, under proper conditions, of jarring other electrons out of the metal. The secondary emission thus obtained depends upon the speed of the original electrons which collide with the metal surface. Normally, these secondary electrons will immediately re-enter the surface from which they were emitted, but if another electrode of higher potential is in the vicinity they will travel toward it in the same manner as the electrons are attracted to the plate of an ordinary three element tube. A standard tube can be connected as shown in the illustration to act as a dynatron. Here the filament F emits electrons, some of which pass through the grid G and collide with the plate P. This collision with the plate P may jar loose from the surface additional electrons, which normally would immediately re-enter P. However, as the grid G is held at a higher potential (higher positive voltage) than P, the electrons will be drawn toward it. A tube arranged in this manner may be used for practically any of the purposes of the standard tube, such as regenerative detector, detector of continuous waves or as a generator of high frequency oscillations, but it has not been shown to have any advantages in this respect, being used mainly as an *oscillator* (q.v.). (See *Electron Emission, Vacuum Tube*, etc.)

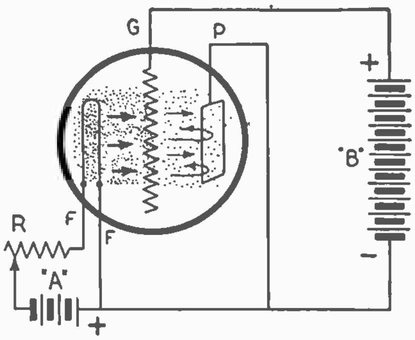

Showing how a standard vacuum tube can be connected to act as a dynatron.

DYNE—The unit of force in the absolute or CGS system of units. It is defined as the force which, acting on a mass of one gram for one second, will impart to the mass a velocity of one centimeter per second. (See *CGS System*, also *Force.*)

E

E—Common symbol for *Electromotive force* (q.v.).

EAR CUSHION—Pads or cushions of soft rubber used in conjunction with head-phones to prevent unpleasant pressure against the ears and also to exclude outside sounds when listening in.

EARTH—An alternate term for *ground*, where the earth or any metallic connection thereto is used as a return in transmission or reception of *electromagnetic waves*. (See *Ground.*)

EARTH CURRENTS—See *Ground Currents.*

EARTH, DEAD—See *Ground.*

EARTH DETECTOR—See *Ground Detector.*

EBONITE—See *Insulating Materials.*

EBURIN—An insulating compound used for *strain insulators* (q.v.). (See *Insulating Materials.*)

EDDY CURRENT LOSS—The portion

of the total loss in electrical apparatus due to *Eddy Currents* (q.v.).

EDDY CURRENTS—Currents induced in the mass of a solid *conductor* due to the action of a varying *magnetic field.* The most common example of eddy currents will be found in the case of a *dynamo* or *generator.* Here the useful current generated in the *armature* is produced by the motion of the armature in the magnetic field. At the same time currents are generated in the iron core due to its motion in the magnetic field. As all currents generated in an electrical machine are produced at the expense of a certain amount of energy, and since the Eddy currents cannot be gathered or put to useful account in this case, they represent a loss. This loss usually is apparent in the form of heat. If the core is made up of solid metal it will be obvious that a good closed path is offered for eddy currents produced. Now if the core is composed of a number of sheets or *laminations,* insulated by thin layers of paper or by an insulating scale (see *Transformer Steel*) the path for the eddy currents is partially broken and the effect is reduced. Eddy currents are produced in the core of a transformer due to the same action—the variation of the magnetic field in which the core is located. In order to reduce the effect, laminated cores are employed as in the case of many generating machines. Eddy current loss is often given as the loss in watts per pound of core material at 10,000 *gausses* (q.v.) and 60 cycles for a sheet 0.0141 inch or 0.0358 centimeter thick. While eddy currents are usually productive of losses they are put to useful account in certain types of instruments. For example, the eddy currents produced in the metal frame of a moving coil meter may be employed to *damp* or retard the free *oscillations* of the needle. (See *Damping of Instruments.*) Eddy currents are also used in a form of speed indicator where the reaction of eddy currents created in a moving disk are made to deflect a pivoted, spring controlled, magnetic needle. (See *Core Loss, Hysteresis,* also *Foucault Currents.*)

EDDY CURRENT COEFFICIENT—The coefficient or numerical multiplier, generally termed K_1, used in calculations of eddy current loss. Its numerical value depends upon the *specific resistance* (q.v.) of the iron used in the core of the machine in question, the character or *wave shape* of the induced voltage, the distribution of *magnetic flux* (q.v.) in the core material, the degree of insulation between sheets, where *laminations* are used, and upon the *flux distribution* (q.v.) due to the shape of the magnetic circuit. Its value will also vary according to the units used in the computation.

EDDY CURRENT LOSS, FORMULA FOR CALCULATING—A common formula for calculation of eddy current loss is the following: $P_\epsilon = \epsilon V(xfB)^2$. Here the loss is expressed in watts as P_ϵ, V is the volume of core metal in cubic centimeters, x the thickness of the sheets in centimeters, f the *frequency* in cycles per second, B the *flux density* (q.v.) in *gausses* and ϵ the eddy current coefficient.

EDGE EFFECT—The effect on the capacity of a condenser due to the curving of the lines of stress at the edge of the plates. This effect is not very pronounced where there is considerable *dielectric* surface (insulating material) extending beyond the edge of the plates. (See *Distributed Capacity.*)

EDISON BATTERY—A number of Edison cells grouped together in one case to supply various currents and voltages as in radio. The illustration shows an Edison "B" battery, used to supply potential to the plates of vacuum tubes. (See *"A" Battery, "B" Battery, Plate, Filament,* also *Storage Battery.*)

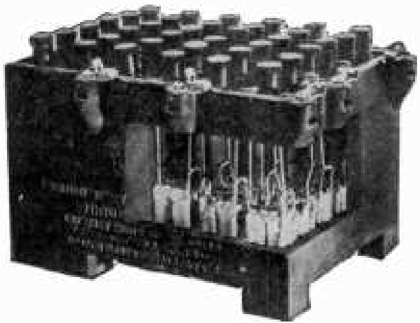

Photo by Courtesy of Edison Storage Battery Co.

Edison "B" battery with portion of case removed to show interior construction.

EDISON CELL—A storage cell employing electrodes of nickeled steel and a solution of potassium hydrate for the electrolyte as distinguished from the usual type of storage cell using lead plates and dilute sulphuric acid as the electrolyte. The chief point of superiority of this type of cell is its comparative ruggedness, due partially to steel construction and also its freedom from the usual ills of cells employing acids and lead plates. The positive electrode in the Edison cell is made by tamping nickel hydrate and alternate layers of pure flake nickel into perforated steel tubes under pressure. The negative electrode is made by pressing iron oxide into flat perforated steel pockets. A number of these pockets are forced into a steel grid to form the complete negative plate. (Note: The *negative* electrode connection on the outside of the cell is referred to as the *positive* pole because it is positive in its relation to the external circuit. Similarly the *positive* electrode connection will be known as the *negative* pole.) The potassium hydrate or hydroxide which takes the place of the acid solution used in the lead plate type of cell acts as a preservative of the steel elements. This naturally means greater life; in fact, it is not subject to the chemical deterioration of other cells. The illustration shows a cell with part of the sides cut away to show the interior construction. (See *Edison Battery.*)

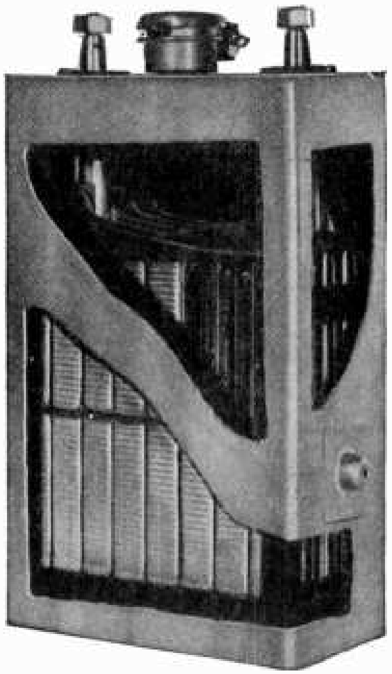

Photo by Courtesy of Edison Storage Battery Co.

Edison cell with part of case cut away to show how constructed.

EDISON, THOMAS ALVA—Born 1847. An inventor famous for his experiments in applied electricity. He began life with newspaper work which he soon abandoned for telegraphy, making many original inventions in duplex systems of operation. After a varied experience in that line he came to New York in 1871, where his talents were recognized and he had opportunity to profitably develop his ideas. The duplex telegraph was made a success the following year, and two years later the quadruplex; and thereupon he began manufacturing on a large scale for the Western Union Telegraph Co. In 1876 he gave up his factory, and established his experimental station at Menlo Park, N. J., where for several years he worked upon the problem of the incandescent electric light, exhibiting a successful bamboo filament lamp in Paris in 1881. He invented the phonograph in 1878. He superintended the construction of the first incandescent lighting station in New York in 1882. Moving his laboratory to Orange, N. J., he established there a large plant for electrical experiment and invention and as a result of his labors there he has taken out 400 patents. Among his inventions may be further named: a type of dynamo, a microphone, the chemical electrical meter, an electric pen, the mimeograph, the magnetic ore separator, dead beat galvanometer, the electric torpedo, a telephone transmitter, and a storage battery. His chief fame rests with his development of the telegraph, his invention of the incandescent lamp and the phonograph.

Thomas A. Edison

EDISON EFFECT—The blackening of the inner surface of an electric light bulb during use. Edison noticed that after the lamp had been burned for some time a coating of black formed on the inner surface and increased in density, becoming finally almost opaque. Edison's experiments, followed by those of *Professor Fleming* (q.v.) resulted in the development of the *Fleming valve* (q.v.)., it having been determined that the black coating was due to the discharge of *electrons* from the hot *filament*. (See *Vacuum Tube, Electron Emission*, etc.)

EFFICIENCY—A very flexible term in electrical practice. Generally speaking, the ratio of useful output of any piece of apparatus to total input. Efficiency is customarily expressed in terms of percentage. Thus, broadly, a piece of apparatus may produce as useful output 80 per cent of the total input. (See *Power Factor*.)

EFFECTIVE ELECTROMOTIVE FORCE—The square root of the mean square of the full alternating current wave. Usually abbreviated R. M. S. (root—mean—square). The effective value of the E. M. F. (electromotive force) is thus taken as the square root of the mean of the squares of the *instantaneous values* over a complete period. (See *Instantaneous Values*.)

EGG INSULATOR—A name applied to a certain type of *strain insulator* due to its egg-like form. The illustration shows such an insulator. The two connections to the insulators are so arranged that the wires will still be looped together in the event that the insulator breaks. (See *Insulator*, also *Insulation of Aerial*.)

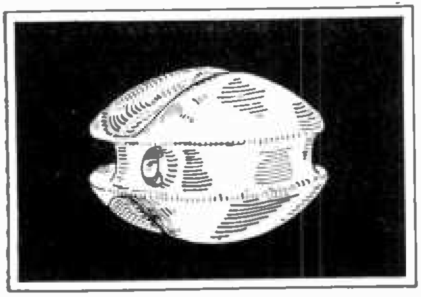

An egg insulator.

ELASTICITY—A property of matter which permits it to resist any change in shape or bulk and permits it after such a change to return to its original state. This property of materials by virtue of which they are enabled to return to their original state after the force causing the distortion has been removed is a very important one in all branches of engineering. Its chief application in radio is in the matter of aerials and aerial guy wires. While the average aerial installation does not represent any serious problems in stress, the large aerial systems of commercial stations must be carefully designed in this respect. For example, a guy wire of one material may have more elasticity than one of similar size in another material. The amount which the wire will stretch under a certain load may be determined by use of a table giving the elasticity of various metals. (See *Electric Elasticity*.)

ELECTRIC ABSORPTION—An effect in *condensers*. That quality of a condenser by means of which it absorbs or "soaks up" a charge of electricity, and conversely, retains part of the charge when the condenser is momentarily discharged. This effect is more apparent in cases where a solid *dielectric* material is used. In fact, *air condensers* (q.v.) are said to have little electric absorption. The reason for this is that mica or other solid dielectrics are strained to a greater extent than air dielectric after the condenser has received a charge. (See *Soaking In*, also *Dielectric Absorption*.)

ELECTRIC DISPLACEMENT—A term proposed by Maxwell to denote a quantity expressing the state or condition of a dielectric in an electric field in accordance with the supposition that such a field is created by the transfer of positive electricity to one end of the field of force and negative electricity to the other end of the so-called *tubes of force* (q.v.). (See *Displacement Current*, also *Electrostatic Induction*, and *Flux Density, Electrostatic*.)

ELECTRIC ELASTICITY—In a dielectric material, the property which permits it to arrest passage of a *displacement current* (q.v.) duc to electric stress. It is equivalent to the electric stress divided by the electric strain. (See *Dielectric*, also *Displacement Current*.)

ELECTRIC FIELD—The area surrounding an electrified body in which the electrical influence of that body can be measured or noticed. Also any area or region in which there is an electric force either steady or varying in intensity. For example, the *dielectric* or insulating surface of a *condenser* (material separating the plates) contains an electric field. Similarly, an electric field always surrounds a wire or other conductor carrying electric current. (See *Field, Magnetic; Field, Electromagnetic; Field, Electrostatic;* also *Flux*.)

ELECTRIC INDUCTION—The transfer of an electric state from a charged or electrified body to a non-electrified body without electrical contact. (See *Induction*.)

ELECTRICITY—The term given to an invisible form of energy. Electricity is not tangible to the human senses, but its effects can readily be detected. The phenomenon of electricity is regarded as due to the separation and independent movement of constituent parts of atoms—known as *electrons*. While electricity is really an intangible thing so far as the human perceptions are concerned, it is known to science not only through its effects or manifestations of energy, but has actually been reduced to quantities. The mass of electricity has been determined, as has also the *rate of flow*, etc. We have units which are used for the basis of all calculations in electricity. Thus, *voltage* is regarded as the propelling force; the quantity of electricity is given in *coulombs*, the *current* or rate of flow of electricity is given in *amperes*, etc. Each term for the various phases of electricity is based on a certain unit which represents definite values under given conditions. The universally accepted theory in chemistry is that all matter is made up of molecules, which in turn are composed of one or more atoms. These atoms now are regarded as comprising minute solar systems made up of positive and negative electrical particles. (For more complete explanation of the *electron theory* see *Electron Theory;* see also *Magnetism, Current, Voltage*.)

ELECTRO-CHEMICAL CONDENSER—See *Electrolytic Condenser*.

ELECTRO-CHEMICAL EQUIVALENT—The amount of a substance liberated by *electrolytic action* (electrolysis) by the passage of one *coulomb* (q.v.) of electric current. It represents the weight in grams of each element of an *electrolyte* (q.v.) which is deposited by one coulomb of electricity. It has been determined that one coulomb will liberate .00001035 grams of hydrogen, which is therefore used as the electro-chemical equivalent of hydrogen. Electro-chemical equivalents are much used in applying the law of electrolysis. This law states that the amount (weight) of an *ion* liberated at an *electrode* each second during *electrolysis* is equal to the strength of the *current* (amperes) multiplied by the electro-chemical equivalent. The corresponding equivalents of other elements compared to hydrogen may be determined by multiplying the figure .00001035 by the atomic weight of the other element in grams and dividing the result by the *valency* (q.v.) of the element in question. (See *Electrolysis*, also *Electrolyte*.)

ELECTRO-CHEMICAL SERIES—A table of metals arranged in accordance with the *potential* they produce in a particular *electrolytic* solution.

ELECTRO-CHEMISTRY—The science dealing with chemical changes due to electricity or effected by means of electricity. (See *Electrolysis*, also *Cell*, and *Storage Batteries*.)

ELECTRODE—The conductor through which electric current enters or leaves an electrical device. That is to say the pole or terminal of the current carrying conductors separated by a medium through which electric current can flow from one to the other. The most common usage of the term is in referring to the *positive* and *negative* poles (q.v.) of a *primary* or *storage cell*. In the case of a *dry cell*, the two electrodes are the carbon rod or block in the center and the zinc container in most cases. In radio it is used to designate the *filament* and *plate* elements of a vacuum tube. The filament is termed the *cathode*, or *negative electrode* and the plate is the *positive electrode* or *anode*. The illustration shows an ordinary *dry cell* or *primary cell* with the two electrodes designated as positive and negative. (See *Anode, Cathode, Battery, Electrolysis, Positive Electrode* and *Negative Electrode*.)

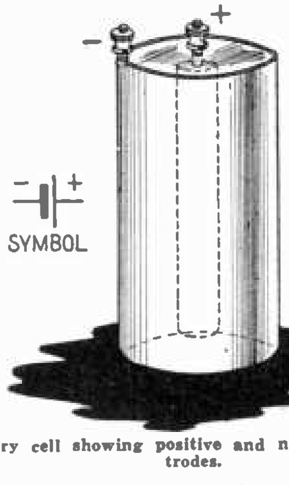

Dry cell showing positive and negative electrodes.

ELECTRODYNAMIC—Term used to denote electric currents or the forces exerted by one current upon another.

ELECTRO-DYNAMICS—A branch of the science of electricity dealing with electricity in motion; the study of the force exerted by electric currents upon each other. This branch of electricity was first accorded serious attention when the French scientist, *A. M. Ampere* (q.v.) announced the result of his investigations in collaboration with D. F. J. Arago. Ampere stated in 1820 that *parallel conductors* through which electric currents were flowing in the same direction had a tendency to be attracted to each other and to be repelled when the currents were flowing in opposite directions. The subject is discussed under its various separate headings. (See *Ampere, Electricity, Magnetism.*)

ELECTRO-DYNAMOMETER—A device for measuring the *current* and *voltage* in a circuit; that is, the actual power in a circuit. It is a form of *wattmeter* (q.v.) designed to indicate the power in *watts.* It is of particular importance in measuring the power in *alternating current* circuits. In the case of *direct current,* the watts in the circuit are the product of the volts multiplied by the amperes, the result being the actual power. In the case of alternating current, however, the product of the volts and amperes is not necessarily, in fact, seldom, the true power or watts. Here we have the problem of true watts and apparent watts, or volt-amperes. If a voltmeter and an ammeter are connected in a circuit and read separately, the readings indicate the maximum voltage and the maximum current or amperage. The product of these two values is the power in watts in a direct current circuit. If the current is an alternating one, the true watts or actual power will be the product of the *instantaneous values* (q.v.) of voltage and current. Under certain conditions in a circuit carrying alternating currents, the current may lag behind or precede the voltage. In other words, the maximum value of current is not reached at the same moment as the maximum value of the voltage. In this case the true watts can only be determined by obtaining the values of volts and amperes at the same instant, i.e., instantaneous values. The electro-dynamometer shown in the illustration gives the instantaneous values, or rather their product, and thus indicates the power or true watts. Here the heavy lines represent the current or moving coil and the light lines the stationary or voltage coil. (See *Dynamometer,* also *Wattmeter.*)

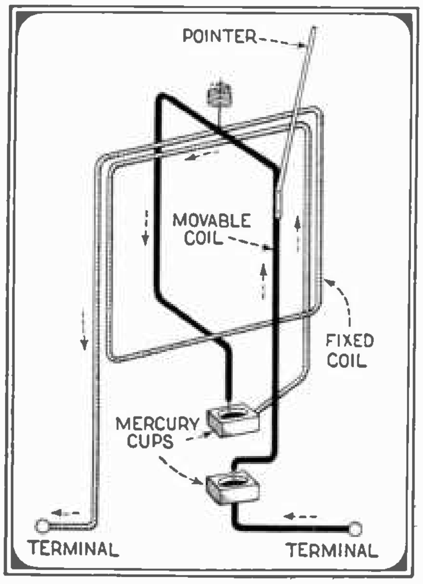

Electro-dynamometer with the current or moving coil shown by heavy lines and the stationary or voltage coil by light lines.

ELECTROLINES—A term advanced by *Professor J. A. Fleming* (q.v.) to designate the lines of electric force radiating from an electron. These lines are pictured as analogous to long straight wires extending in all directions from the center of a small sphere. (See *Electron,* also *Electromagnetic Waves.*)

ELECTROLYSIS—The decomposition of a comopund substance, generally a liquid, into its component parts by the action of an electric current passing through it. While electrolysis is an undesired effect in certain phases of electrical power work, causing composition of grounded structures, its application to radio is mainly in the case of cells for producing an *electromotive force* or to furnish current for lighting tubes. The constituent elements of a chemical compound such as used in a wet cell or any device using an electrolyte for the production of electric current, are known as *ions* (q.v.). These ions, which are separated during electrolysis, are of two kinds. The electro-positive ions are known as *cations,* and the electro-negative ions are known as *anions.* The electro-positive ions appear at the cathode or negative electrode during electrolysis and therefore are of positive origin. The electro-negative ions appear at the positive electrode and are therefore of negative origin. This is easily understood when the ions are considered as moving from the positive electrode or to the negative. Thus the ions appearing at the negative electrode must have come from the positive electrode and are therefore considered electro-positive in nature. The illustration shows the action of the ions in a typical *primary cell* (q.v.). The ions are considered as carrying the current through the electrolyte—each ion carrying a fixed charge of electricity of positive or negative nature. An ion is capable of carrying only a certain fixed charge of electricity and therefore any increase in the current will necessarily be accomplished by an increase in the number of ions. (See *Primary Cell, Electrolyte, Electrolytic Detector, Interrupter* and *Condenser.*)

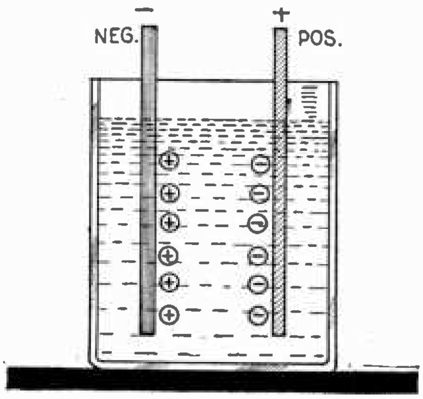

Illustration showing the action of ions in a primary cell.

ELECTROLYTE—The liquid decomposed during electrolysis, as the exciting fluid or solution in a wet cell or primary cell. The liquid used in any primary, secondary or electrolytic cell. (See *Cell,* also *Storage Battery.*)

ELECTROLYTIC ACTION—The decomposition of a chemical compound (electrolyte) into its constituent elements—ions, by the passage of an electric current through it. (See *Electrolysis, Ions,* also *Cell.*)

ELECTROLYTIC BATTERY CHARGER—A device for converting alternating current such as supplied for house lighting purposes, to direct current for the purpose of charging storage batteries. Essentially one or more cells containing an electrolyte and two metal electrodes, generally lead and aluminum, so arranged that current can only pass in one direction, thus making it uni-directional and suitable for charging batteries. Such chargers are arranged to rectify either one-half of the alternating current cycle, or both halves of the wave. (See *Charger, Storage Battery, Electrolytic Rectifier,* also *Electrolytic Cell.*).

ELECTROLYTIC CELL—An arrangement of two *electrodes* in an electrolyte through which electric current can be passed to produce electrolysis. The electrolyte is decomposed or separated into ions by the action of the current, and at each of the electrodes one of the substances of which the electrolyte is composed accumulates. Such a cell is often used to measure electric current, as the amount of a substance deposited on the electrode depends directly on the amount of current passed through the electrolyte. (See *Electrolysis.*)

ELECTROLYTIC CONDENSER—A form of condenser making use of two electrodes placed in an electrolyte and used for alternating current circuits of high voltage and low frequency, low voltage and high frequency or high voltage and high frequency. The action of such an arrangement as a condenser depends on the fact that upon the passage of certain forms of alternating currents through it, a gas is formed around the electrodes, causing *polarization* (q.v.) and opposing the current. This polarization creates a very good insulating medium and a condenser of such nature can be made to have very high capacity. An electrolytic condenser may be used to prevent any sudden surge of pressure due to lightning or other causes. The polarization is instantaneous when any surge of voltage takes place and thus forms an automatic safety device which operates instantly and returns to its normal characteristics as a condenser as soon as the surge has passed. (See *Capacity,* also *Condenser.*)

ELECTROLYTIC CONDUCTION—The action in an electrolyte wherein the ions that have been separated carry positive and negative electric charges in opposite directions to the electrodes. (See *Electrolysis.*)

ELECTROLYTIC DETECTOR A device which converts high frequency currents into direct current pulsations capable of operating a telephone receiver. It is an extremely sensitive detector of electromagnetic waves (radio signals). The illustration shows a simple form as originally used in place of the customary crystal detector. A fine platinum wire dips into a

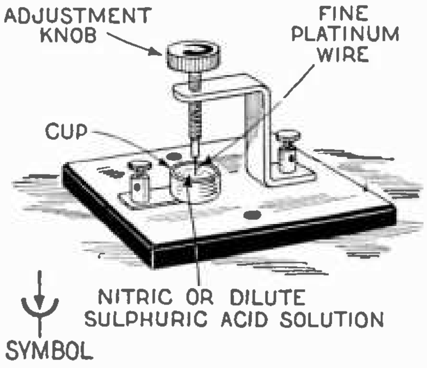

Electrolytic detector originally used in place of a crystal detector for radio reception.

cup of acid solution, generally either nitric or dilute sulphuric. The fine wire is lowered until it just touches the acid solution and must be very carefully adjusted as it has a tendency to curl up, its diameter being usually but one five thousandth of an inch or less. The high frequency currents can pass in only one direction through the detector which gives the rectifying action and makes the signals audible in the head phones. There are several theories regarding the action of an electrolytic detector, the most probable one being that the contact of the wire on the surface of the acid solution gives the effect of a very small *electrolytic condenser.* (q.v.) When the high frequency currents are passed through the detector there is a polarization action which causes high resistance to current in one direction but not in the other. (See *Detector,* also *Polarization.*)

ELECTROLYTIC INTERRUPTER—A device using electrodes immersed in an electrolyte for the purpose of interrupting or breaking up the flow of current. It is used mainly for induction coils for producing high voltages. The common form comprises a lead plate as a cathode or negative element and a platinum wire extending beyond a glass or porcelain tube in which it is tightly imbedded, both immersed in an electrolyte of dilute sulphuric acid. When current passes through, bubbles form on the platinum electrode causing temporary stoppage of the current. The bubbles are automatically dissipated and the current flows again, this operation occurring at rapid intervals and giving as many as 1000 interruptions per second. (See *Interrupter,* also *Induction Coil.*)

ELECTROLYTIC RECTIFIER—A device for changing alternating current into direct current by use of electrodes placed in an electrolytic solution. The action of such electrolytic cells or rectifiers is to permit current to pass only in one direction, thus suppressing one half of each cycle and producing a direct current. The illustration shows one of the customary arrangements for electrolytic rectification, being an electrolytic cell with two electrodes, one being aluminum and the other a lead electrode. One or more lamps are placed in series or shunt in the circuit to permit the proper current flow. The positive pole of a storage battery is connected to the aluminum plate as shown and the lead electrode connects to the other side of the alternating current line and back to the negative side of the battery. At one instant current flows in the normal direction through the battery, but when the current reverses and flows in the opposite direction through the electrolytic cell, oxygen is given off at the aluminum electrode and forms a thin film of aluminum oxide which acts as an insulator and prevents the passage of the current in that direction. Thus with one cell one half of each alternating current cycle is stopped or cut off when the aluminum electrode is made positive and the result is an intermittent direct current suitable for charging batteries. Such rectifiers are also used for supplying current to the filaments of vacuum tubes, or voltage to the plate members of the tubes, particularly for transmitting. (See *Electrolytic Rectifier Full Wave,* also *Balkite Battery Charger.*)

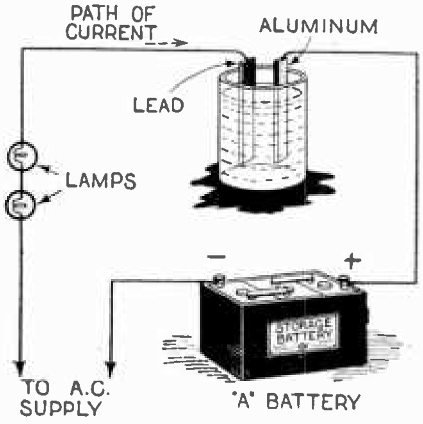

Customary method of electrolytic rectification.

ELECTROLYTIC RECTIFIER FULL WAVE—This system is much used for charging storage batteries from an alternating current house lighting supply. Four cells are used in this instance. The electrolyte may be a solution of hydrolchloric or sulphuric acid, or a saturated solution of sodium phosphate or bicarbonate of soda. Each cell has two electrodes, one iron or aluminum and the other lead. The positive pole of the storage battery is connected to the lead electrodes of two cells and the negative pole is connected to the aluminum electrodes of two cells. By this means both halves of the alternating current cycle are rectified to direct, pulsating current. This is known as *full wave rectification* (q.v.) (See *Charger, Storage Battery, Balkite Battery Charger,* also *Rectifier.*)

ELECTRO-MAGNET—A magnet using a curved or straight piece of soft iron or other magnetic material so arranged as to become magnetized by the passage of electric current through a coil of wire surrounding it. The illustration shows one of the commonest forms of electro-magnet. This type is used in a bell or buzzer to draw down the vibrator bar and break the circuit. (For explanation of action see *Buzzer.*) When the current is withdrawn from an electro-magnet, the magnetic force or attraction ceases. Electro-magnets are used for a variety of purposes in radio, such as for test buzzers, head phones, relays, etc. (See *Magnet,* also *Magnetic Lines of Force.*)

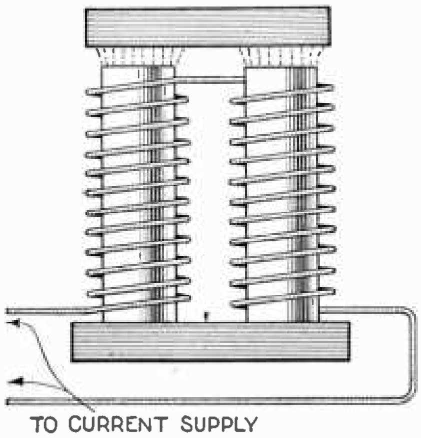

One of the common forms of electro-magnets.

ELECTROMAGNETIC CONTROL—The control of various switches and other apparatus by means of *electromagnets,* usually through an auxiliary contact at a distant point. (See *Remote Control,* also *Electromagnet.*)

ELECTROMAGNETIC FIELD—The space understood to be filled with *electromagnetic lines of force.* (See *Flux,* also *Lines of Force.*)

ELECTROMAGNETIC FLUX—The distribution of the lines of force through an electromagnetic field. (See *Flux,* also *Lines of Force.*)

ELECTROMAGNETIC IMPULSE—An impulse or electromagnetic disturbance conveyed to the ether from a conductor carrying high frequency currents. Another term for the electromagnetic waves involved in radio. (See *Electromagnetic Waves.*)

ELECTROMAGNETIC INDUCTION — The production of electric currents through a change in a *magnetic field* (q.v.). Its production is due to a change in the number of lines of force which are linked with a conducting circuit. The simplest example of this means of producing current is shown in the illustration. Here a conductor is arranged to be moved in the immediate vicinity of a magnet, causing a change in the number of lines of force that link the conductor and creating an induced current. This current will persist only so long as the conductor continues to be moved in the magnetic field. That is to say when there is no longer any change in the magnetic lines of force linking the conductor, there will be no current induced. The discovery of electromagnetic induction and formulation of its laws was primarily due to the scientist *Faraday* (q.v.). Faraday's law governing electromagnetic induction follows: "The induced electromotive force around any circuit is the rate of the decrease of the total flux of magnetic induction through the circuit." (See *Induction, Flux,* and *Lines of Force.*)

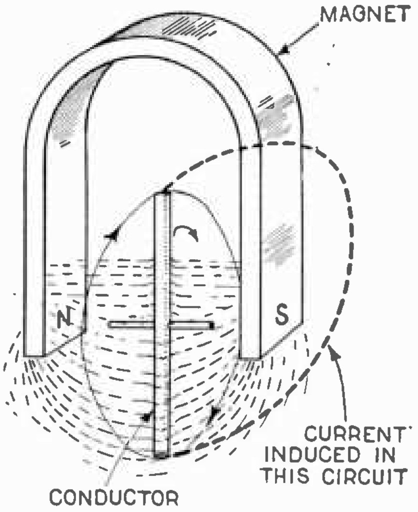

Simple method of producing electric currents by causing changes in the magnetic field.

ELECTROMAGNETIC INSTRUMENTS—Electrical measuring instruments such as voltmeters, ammeters, etc., which depend for their action on the exertion of electromotive forces upon iron armatures. (See *Ammeter, Thermal;* also *Voltmeter.*)

ELECTROMAGNETIC MICROPHONE—A type of microphone in which the sound waves due to speech or music cause vibrations of a suspended coil in a magnetic field, inducing in the coil minute waves of electromotive force corresponding in form to the original sound waves. A form of microphone used in radiophone (broadcasting) transmission when purity of sound is an important consideration. Such a microphone requires that the induced currents be heavily amplified. (See *Microphone*, also *Speech Amplifier.*)

ELECTROMAGNETIC RADIATION—The propagation or radiation of electromagnetic waves through the ether. (See *Electromagnetic Waves,* also *Ether.*)

ELECTROMAGNETIC ATTRACTION—The attraction existing between unlike poles of electromagnets. This attraction is identical with that of two unlike poles of ordinary magnets. (See *Magnetic Attraction and Repulsion.*)

ELECTROMAGNETIC REPULSION—The tendency of two like poles of electromagnets to act against or repel each other. (See *Magnetic Attraction and Repulsion.*)

ELECTROMAGNETIC SWITCH—A switch for breaking electrical circuits by the action of an electromagnet. (See *Electromagnet*, also *Remote Control.*)

ELECTROMAGNETIC THEORY OF LIGHT—The theory and consequent scientific proof that radio waves and light waves are practically identical in nature. Light waves travel at a tremendous speed through the ether, approximately 186,000 miles per second, which is also understood to be the velocity of electromagnetic or radio waves. (See *Electromagnetic Waves.*)

ELECTROMAGNETIC UNITS—The fundamental or absolute units employed in electricity as a basis for calculations and on which the practical units are based. The effects of electricity, and hence electricity itself, can be measured in two ways. The first is the measurement of the quantity of electricity by the force it exerts upon another stationary quantity of electricity. This is known as the *electrostatic system.* The second is by measuring the force which a given quantity of electricity exerts upon a *magnetic pole* when flowing through a nearby conductor. This is referred to as the electromagnetic system. Fundamental or absolute units are derived from the fundamental units of *length, mass* and *time.* These units are the *centimeter of length, the gramme of mass* and the *second of time.* The term centimeter-gramme-second, referring to this system, is abbreviated C.G.S. system. The practical units of electricity are known as *volts, amperes,* etc., and are based on the absolute units. (See *Electrostatic, Units, C.G.S. System,* also *Ampere, Erg, Dyne* and *Coulomb.*)

ELECTROMAGNETIC WAVES—The term applied to the energy radiated from an *antenna* in radio transmission. So called because it is understood to travel in the form of waves. One-half the energy of such a wave is electrical and the other half electromagnetic. *Electromagnetic waves* travel with the speed of light waves or approximately 186,000 miles per second. These waves are of extremely high frequency—that is, their direction changes many thousand times each second. This will be better understood by means of the illustration showing the arrangement of the electric and *magnetic fields* around a radiating antenna. Here the lines represent the electric field and the shaded portion the magnetic field. The lower loops or "feet" of these waves are assumed to pass over the surface of the earth which is a conductor. Thus when electromagnetic waves are being radiated from an antenna, circular distributions of electric charges are in the earth's surface around the antenna, these charges being alternately negative and positive, the bands of charge spreading out or radiating from the antenna in the manner shown in the illustration. These waves are subject to various influences which have a tendency to refract or reflect the waves much in the manner of light waves.

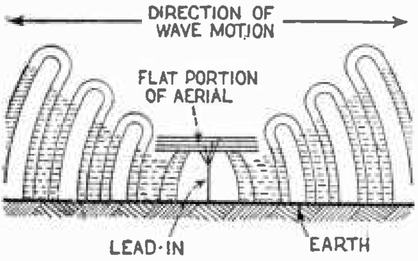

Illustration showing the arrangement of the electric and magnetic field around a radiating antenna.

As these waves move out from the antenna they gradually increase in height, although the distance between the "feet," i.e., from one wave to another, remains the same. This distance is known as the wave-length. (See *Reception of Electromagnetic Waves,* also *Wave-length, Radiation Transmission.*)

ELECTROMETER—An instrument used for measuring potential differences (voltages) and, under certain conditions, power in *watts.* The operation of the instrument depends upon the attraction or repulsion of electrostatic charges and it may be used for either direct or alternating current measurements. (See *Voltmeter, Electrostatic.*)

ELECTROMETRY—The practice or science of electrical measurement. (See *Meter, Voltmeter.*)

ELECTROMOTIVE FORCE—Symbol E or e—abbreviation E.M.F. The force or electrical pressure which starts and maintains a current of electricity through a conductor. Electromotive force is commonly referred to in terms of volts and is abbreviated E.M.F. The electromotive force in electricity is analogous to the difference of level which produces pressure in pipes carrying water and forces the water to flow. Electromotive force is produced by a difference of potential, and this in turn sets up the flow of current. The illustration *Fig. 1A* shows a simple primary cell or electrolytic cell, and *Fig. 1B* is the analogous action of two tanks holding water, connected by a pipe and so arranged that the water level in one tank is higher than in the other, thus having a difference of level and creating pressure which makes the water flow. The water in tank A being at a higher level than in the other, the pressure causes the water to seek its common level and it flowed into tank B through the pipe. This flow continues as long as there is difference of level. In the case of the cell *Fig. 1A* there are two electrodes, one of which is at a higher potential than the other. That is to say, one is in a state of higher electrification than the other, causing a difference of potential and producing an electromotive force. This force is in the direction from the point of greater potential to point of lesser potential as in the case of the water, and causes a flow of current in that direction. Here the flow is from B to A in the electrolytic solution and from A to B externally through the conductor C. Thus as long as there exists a difference of potential, current will flow in an attempt to equalize the potentials. (This statement holds true only under ideal conditions. See *Polarization.*) This flow of current in the illustration *Fig. 1A* is according to *Ohm's law* (q.v.), i.e., a difference of potential of one volt will cause one ampere to flow through the external circuit when the resistance of the circuit is one ohm. (See *Ampere, Ohm, Volt, Resistance.*)

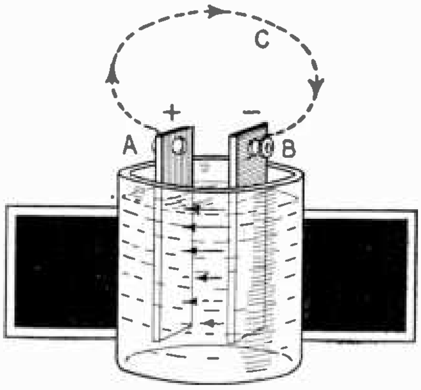

Fig. 1-A. A simple primary or electrolytic cell.

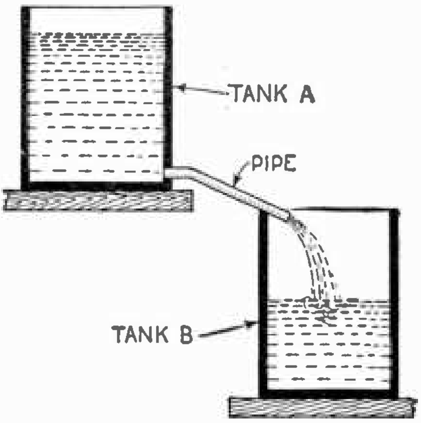

Fig. 1-B. Water in tank "A" seeks its level by discharging into tank "B" in a manner analogous to the flow of electric current where a potential difference exists.

ELECTRON—The smallest negative charge of electricity known. An English investigator, Dr. Johnstone Stoney, is credited with having named the electron, although it was so named by him in anticipation of its being isolated and measured. The actual proof of the earler assumptions was due to Sir J. J. Thomson and others, who actually measured, and in all effect weighed, the minute particles of negative electricity. (See *Electron Theory.*)

ELECTRON FLOW—The theory of movement of free electrons along a conductor through which an electric current is flowing. (See *Electron,* also *Current, Assumed Direction of Flow.*)

ELECTRON EMISSION—The emission or discharge of electrons or minute particles of negative electricity from a heated body, particularly in a vacuum. The emission of electrons in a vacuum tube is the basis of operation of all vacuum tubes as used in radio. (See *Electron Theory,* also *Vacuum Tube, Theory of Operation.*)

ELECTRON THEORY—The assumption that the atoms of all substances contain one or more particles of negative electricity, which is of course paramount to saying that all matter is electricity. These particles are known as electrons, and while the assumption is referred to as a theory, it has been rather definitely proven by numerous experiments, and electrons have actually been measured. Under the electron theory it is assumed that the charge carried by a single electron is the smallest possible charge which can exist in nature and that no charge of electricity exists or can be produced which is not an integral multiple of this charge. (Note: This is merely a mathematical way of stating that any charge of electricity is equivalent to the combined charges of a certain number of electrons.) Probably the most accurate conclusion regarding the charge carried by an electron is due to experiments carried on by Prof. A. E. Millikan of the University of California. His measurements state this charge of a single electron as being 4.77×10^{-10} electrostatic unit, which is 1.59×10^{-19} of a *coulomb* (q.v.). While the mass or weight of an electron is said to vary with its velocity, assuming a velocity of less than one-tenth that of light, the mass will be 8.9×10^{-28} grams. This is approximately one seventeen hundredth of the weight or mass of an atom of hydrogen. For purposes of calculation the electron is assumed to be in the shape of a sphere. The radius has been estimated at approximately 10^{-13} centimeter. When it is considered that a centimeter is somewhat less than half an inch it is readily seen that an electron is an infinitesimal affair.

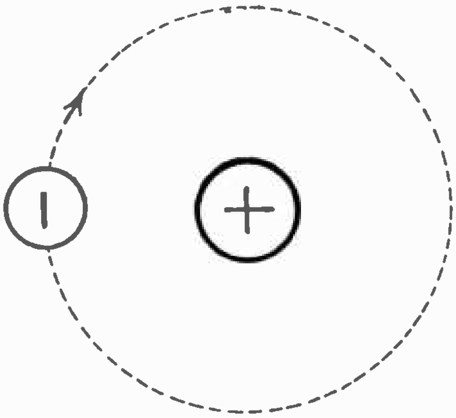

Illustration showing the electron, or negative particle, revolving around the proton or positive nucleus.

The figures as to velocity of electrons vary somewhat with different authorities, but it is probably about 100 kilometers (approximately 62½ miles) per second at a temperature of zero degrees centigrade, the speed increasing as the temperature rises. Electrons may be understood as comprising minute solar systems of their own. In the illustration the *electron* or negative particle is shown revolving around the *proton* (q.v.) or positive nucleus in the manner of a planet revolving around the sun. This represents a normal atom of hydrogen, the simplest of all atoms, which, as shown, is composed of a single positive nucleus and a single electron revolving around it. In the case of an atom of helium there is a doubly charged positive nucleus, or proton, and two negative electrons revolving around it. Other atoms have as many as 92 electrons.

All vacuum tubes used in radio operate on the basis of the emission of electrons from a heated filament within the tube. These electrons are being emitted from the hot filament at all times during operation of the tube. When the incoming signals are such as to charge the grid negatively they are repelled and held back to the filament, thus offering no path for the signals. When the incoming signals have reversed their polarity, as they do thousands of times each second, the grid will be positively charged and the electrons will be aided in their travel to the plate, thus forming a conductive path for the currents. This, of course, results in the elimination of one-half of each cycle of the incoming currents or oscillations. That is to say, they are rectified to direct pulsating currents. (See *Amplifier, Vacuum Tube, Theory of Operation* and *Space Charge.*)

ELECTRON RELAY—A term occasionally applied to a three element tube (vacuum tube). This alternate term is obviously derived from the fact that the vacuum tube makes use of the path of electrons and in effect acts as a relay. (See *Electron.*)

ELECTRON TUBE—Any tube using a high vacuum and depending for operation on the emission of electrons from a heated *cathode.* (See *Electron,* also *Vacuum Tube.*)

ELECTRONEGATIVE—Having the property of being attracted to the positive pole of an electrolytic cell. The electronegative component. (See *Anion,* also *Cathion.*)

ELECTROPOSITIVE—Having the property of being attracted to the negative pole of an electrolytic cell during electrolysis. The electropositive component. (See *Anion,* also *Cathion.*)

ELECTROSCOPE—A device for detecting minute differences of potential (voltages). Such a machine will not measure the voltage but merely indicate its presence. The simplest form of electroscope is shown in the illustration. Here a ball B, composed of the pith of some wood, preferably elder due to its extreme lightness, is suspended by a thin silk thread from an insulated stand. The object under investigation is moved near the pith ball, which will be repelled or attracted by the object if the object is in a state of electrification; but will remain motionless if no difference of potential exists. (See *Voltmeter, Potential, Difference of.*)

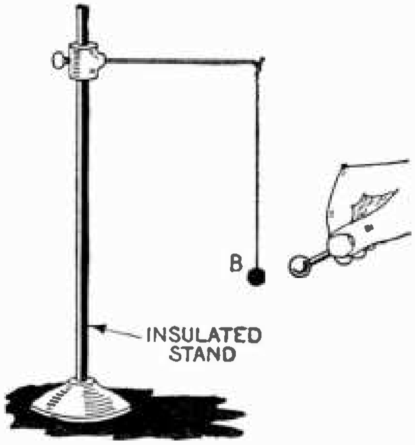

Simplest form of electroscope.

ELECTROSE—An insulating composition much used in electrical power work and for radio purposes. It has very high compressive strength, is hard and tough without being brittle, and is moisture, water and oil proof. Electrose is used extensively for antenna insulation. (See *Insulator*).

ELECTROSTATIC—The term used to designate EMF's (electromotive forces) or potentials, charges, etc., and the resultant forces. That is to say, electricity at rest or stationary in a body in order to distinguish it from the forces due to *electro-dynamics* (q.v.).

ELECTROSTATIC CAPACITY—The amount of electricity which a condenser or similar body is capable of storing. If a condenser is capable of receiving and storing one *coulomb* of electricity under a potential of one *volt* it is said to have an electrostatic capacity of one *farad*, which is the unit of electrostatic capacity. The electrostatic capacity of a condenser depends on several factors, mainly the size of the plates and the nature and thickness of the dielectric material (insulating sheets). (See *Capacity,* also *Condenser.*)

ELECTROSTATIC CHARGE—The presence of electricity in the form of electrostatic lines of force in a body such as a charged conductor. (See *Charge.*)

ELECTROSTATIC COUPLING—The association or coupling of circuits one to another by means of electrostatic action, i.e., condensers. Any capacitive association of one circuit with another. (See *Coupling.*)

ELECTROSTATIC FIELD—A region wherein forces are present and exerted due to the presence of electric charges. It is the area immediately surrounding a charged body. (See *Field, Electromagnetic,* also *Field, Electrostatic*).

ELECTROSTATIC FLUX—Lines of force making up an electrostatic field. (See *Field, Flux,* also *Flux, Electrostatic*).

ELECTROSTATIC FLUX DENSITY—The density or number of electrostatic lines of force in a unit area of an electrostatic field. (See *Electrostatic Field,* also *Flux Density.*)

ELECTROSTATIC FORCE—The force due to interaction of electric charges. (See *Electrostatic Field,* also *Electrostatic Flux Density.*)

ELECTROSTATIC INDUCTION—The production of a charge in a body due to the presence of an opposite charge in a nearby conductor, or synonymous with displacement in the case of a condenser. Electrostatic influence or induction is the basis on which condensers operate. In the illustration *Fig.* 1 is a simple circuit containing a source of potential, which in this case is a battery. As the circuit is closed through the key a momentary flow of current takes place. This flow of current is accompanied by a condition of electric strain in the insulating material of the condenser. The electrically strained dielectric or insulating material exerts a back pressure against the pressure or force of the battery, and when this back pressure is equal to the original battery pressure the flow ceases. The strain in this dielectric is known as *electric displacement* and this displacement can be transmitted or passed on to neighboring parts of the medium and thence into space. Under the proper conditions this displacement and the magnetic field which accompanies it may be detached from the circuit and move independently in space. This is the fundamental process in the radiation of *electromagnetic waves* (q.v.). Now, in *Fig.* 2, A and B represent two conductors, both insulated. If conduc-

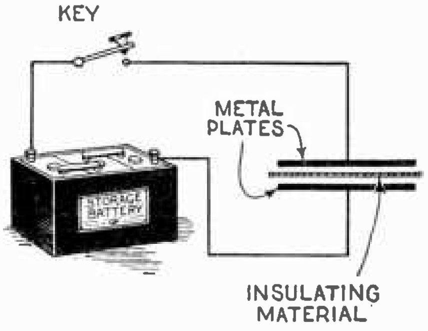

Fig. 1. A simple circuit for illustrating electrostatic induction.

tor A has a positive charge and is brought close to conductor B, although not in actual contact with it, the result is a displacement of electric charge which results in the near side of conductor B becoming negative, while the opposite side assumes the same sign (positive) as conductor A. Actually B has received no charge, but has been placed in a state of strain due to electrostatic induction between it and the conductor A. Due to the negative charge induced on B, the positive potential difference of A has been decreased and in order to maintain the constant potential it must be given a further supply of electricity. In this manner it is actually possible to store the electricity by electrostatic induction between two conductors or plates, by reason of the strain in the insulating medium—air in the case of the two conductors—which in turn passes the displacement on to conductor B. (See *Conductor, Capacity, Electrostatic,* also *Induction.*)

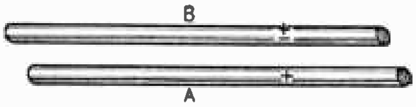

Fig. 2. Illustration of displacement of electric charge due to proximity of charged conductors.

ELECTROSTATIC LEAKAGE—The gradual loss through leakage, of the electrostatic charge in a body, due to the fact that perfect insulation is impossible to obtain. (See *Electrostatic Induction.*)

ELECTROSTATIC LINES OF FORCE—The lines of force traversing an electrostatic field. That is to say the *electrostatic flux* (q.v.).

ELECTROSTATIC STRAIN—The strain or state of stress in which a body is placed when in an electrostatic field. In a condenser, the strain experienced by the *dielectric* (q.v.) when the condenser is in a state of charge. (See *Condenser,* also *Capacity, Electrostatic*).

ELECTROSTATIC SYSTEM—The electrostatic system of C.G.S. units. One of the absolute unit systems in electrical practice. (See *C.G.S.*, also *Electrostatic Units* and *Electromagnetic Units.*)

ELECTROSTATIC UNITS—A system of measurement used in electrical engineering based on the unit of electric quantity under the *C.G.S. system* (q.v.). This unit quantity of electricity is the quantity of electricity which, placed on a small sphere will repel an equal quantity of electricity of the same sign (both negative or both positive) on a similar sphere with a force of one *dyne.* (q.v.) when the centers of the respective spheres are one centimeter apart. With this definition it is assumed that the spheres are in a vacuum, otherwise the distance between centers would vary according to the dielectric or insulating medium. (See *C.G.S., Electromagnetic Units,* also *Conversion Factors.*)

ELECTROSTATIC VOLTMETER — A meter for measuring voltage, depending for its action on the electrostatic action between a fixed and a moving plate connected respectively to the two poles or terminals between which the source of potential is connected. (See *Electrostatic Force,* also *Voltmeter.*)

ELECTROSTATIC WATTMETER — A type of wattmeter depending on the electrostatic action between two sets of fixed and moving plates. (See *Electrostatic Force,* also *Wattmeter.*)

ELECTRO-THERMAL METER — The term sometimes applied to a *hot-wire ammeter* (q.v.) or voltmeter operating on the principle that the heat generated by the passage of an electric current through a constant resistance will be equal to the square of the current. (See *Thermal Ammeter.*)

ELEMENT—A term primarily applied in chemistry to the substances which are not subject to decomposition by electric analysis (see *Electrolysis*). These elements are hydrogen, helium, etc. The term is also used quite generally to denote components such as the electrodes in a primary cell or the plate, grid and filament members of a vacuum tube. (See *Cell,* also *Vacuum Tube.*)

EMANATE—"To flow forth or proceed, as from a source."—Webster. Thus, emanating, issuing from, as the emanation of electrons from a hot filament in a vacuum tube. The throwing off of electrons. (See *Electron Emission.*)

E.M.F. — The abbreviation for electromotive force. The practical unit of E.M.F. is the volt. (See *Electromotive Force,* also *Volt.*)

EMISSION—The act of sending forth. Used chiefly to designate the emission of electrons from the filament of a vacuum tube as used in radio. The flow of these electrons between filament and plate is controlled by the grid. (See *Electron Emission,* also *Vacuum Tube, Theory of Operation.*)

EMISSION CURRENT—In a vacuum tube for transmitting it is the product of the tube current multiplied by the plate voltage and is generally given in watts. Generally speaking, the power for which the tube is designed. (See *Transmitting Tube, Power Rating,* also *Vacuum Tubes.*)

EMISSIVITY OF FILAMENT — The ability of a filament to emit electrons when heated by an electric current. (See *Electron,* also *Vacuum Tube.*)

EMPIRE CLOTH OR PAPER—A closely woven cambric coated with two or more films of an oxidized oil, or a tough paper treated with some oxidized oil. Empire products serve as insulating materials for transformers, condensers and various other devices requiring thin insulating sheets. In a transformer for relatively high voltages, insulated paper is used between the layers of windings and in certain types of fixed condensers it is used as a dielectric material. The cloth has a higher breakdown strength; that is, it will stand higher voltages than paper. (See *Breakdown Potential,* also *Dielectric.*)

ENAMELED WIRE — Wire, generally copper, covered with a coating of enamel as an insulating covering and to protect it from moisture and corrosion. (See *Magnet Wire.*)

ENDODYNE—A term occasionally used to describe reception of radio signals by means of locally generated oscillations such as in the case of the *heterodyne* (q.v.), but wherein these oscillations are created by a circuit or system that is part of the receiving circuit and not a separate arrangement. (See *Self Heterodyne.*)

ENERGY—The capacity for doing work in an electrical sense. Energy is something furnished to a body by means of work done on it. Thus, if a ball is thrown, the operation of throwing is the work, and the ball itself in traveling performs work, the means of doing this work being the energy. There are two kinds of energy, *potential,* as in the case of electricity, and *kinetic,* the energy of a body due to motion. Kinetic energy belongs properly in the realms of mechanics and so will be passed over here with only this brief mention.

When we speak of a difference of potential in a primary cell, it means in effect that the body or cell has potential energy. That is to say, the energy is present although not active. If, now, a wire is connected across the two terminals of the cell, the difference of potential will cause electric current to flow from one pole to the other. Actually, in a cell of this type, we have chemical energy stored up. When a circuit is completed, the stored chemical energy is converted into electrical energy or creates the difference of potential, which in turn causes current to flow and electrical *work* is thus done. (See *Potential,* also *Voltage* and *Generator.*)

EPSTEIN HYSTERESIS TESTER — A device for measuring *hysteresis* and *Eddy current loss* in a sample of sheet iron by using an alternating current wattmeter. The device is used for determining the properties of samples of iron, particularly those to be used for transformers. (See *Eddy Currents,* also *Transformer, Hysteresis* and *Wattmeter.*)

EQUIVALENT RESISTANCE — The value which a resistance would be required to have in order to permit the same quantity of current to flow with the same applied electromotive force (voltage) as in the case of a piece of apparatus wherein there are other factors besides pure resistance to determine the amount of current. (See *Impedance,* also *Reactance.*)

ERG—The fundamental unit of mechanical work under the *C.G.S. system* (q.v.). This is defined as the work done by a force of one *dyne* (q.v.) when the body producing the force and the body on which the force acts are one centimeter apart. (See *Joule,* also *Practical Units.*)

E.S.U. — Abbreviation sometimes used for *electrostatic units* (q.v.). (See *Electromagnetic Units.*)

ETHER—or **AETHER**—The term used to denote the supposed or hypothetical medium by means of which electromagnetic and light waves are propagated. (See *Ether Waves.*)

ETHER WAVES—Term sometimes applied to electromagnetic or radio waves on account of the assumption that they travel through the supposed medium ether. The assumption of an all-pervading medium such as ether has been the basis of the major part of the extensive research work carried on with light and electromagnetic waves for many years. Various experiments have shown beyond reasonable doubt that there is some such medium, capable of conducting or rather permitting passage of light waves and radio waves. So accurate has this work been that it has been proven beyond question that light waves travel through the ether with the tremendous velocity of 186,000 miles per second (approximately). It has also been determined that radio waves travel at this same speed through the ether. It has further been proven that radio waves all travel at the same identical velocity, regardless of their length, form and other factors. (See *Electromagnetic Waves,* also *Theory of Propagation of Electromagnetic Waves.*)

EXCITER—A small direct current dynamo used for the purpose of exciting the field magnets of an alternating current generator. Such a device is generally a part of the main machine and mounted on the main shaft, but it may also be a separate unit. The purpose is to furnish current to the field magnets (electromagnets) to create a magnetic flux. (See *Alternator*, also *Generator.*)

EXPLORING COIL—A coil of insulated wire wound on a rotating form as an integral part of a direction finder or *goniometer*. (See *Direction Finder*, also *Goniometer.*)

F

F—The abbreviation for *dielectric field intensity* (see *Field*, also *Dielectric*), also used as the symbol for the unit of *magneto-motive force*, the *gilbert* (q.v.). The small or lower case f is sometimes used to denote the *farad*, the practical unit of *electrostatic capacity* (q.v.) and also as a symbol for *frequency* (q.v.).

FADING—The tendency of radio signals to decrease in volume or amplitude, or to fade. The strength of a signal is seldom the same at different hours of the day or night, generally being stronger during the dark hours of the night. This has been attributed to a variety of causes, the most plausible being that the air is more free of disturbances during the night and thus permits more ready passage of the waves through the ether. The difference between the strength of signals in the daytime and at night is sometimes very pronounced, but the phenomenon of fading is usually noticed only when distant signals are being received. It has been fairly definitely established that fading of signals is due to the reflection of the transmitted waves. (See *Wave Distortion*, also *Reflection and Refraction.*)

FALL OF POTENTIAL—See *Potential Drop.*

FAN AERIAL—An aerial, usually for transmission, constructed in the shape of a fan. (See *Aerial.*)

FAN CONNECTOR—A small triangular or fan-shaped piece of metal arranged with several holes along one edge, with set screws for fastening in each the several wires of the lead-in of a multi-wire aerial. The lower apex is fitted with some sort of connecting device to permit a single wire to be led in to the receiver. (See *Lead-In.*)

Illustration by Courtesy of Radio Specialty Co., Inc.

Fan Connector—The leads from each wire of a multi-wire aerial go to the wide portion and are held by the set screws: the lead to the set goes to the small end.

FARAD — The unit of *electrostatic capacity*. It is the practical unit of capacity and is defined as the capacity of a condenser that is capable of being charged to a potential difference of one volt by a charge of one ampere for one second; i.e., one *coulomb* (q.v.). "Farad" is a contraction of the name of the distinguished English scientist *Michael Faraday* (q.v.). Inasmuch as the farad is too large a unit for practical purposes, the *micro-farad*, or one millionth of a farad, is commonly used. (See *Electrostatic Capacity*, also *Condenser.*)

FARADAY, MICHAEL. Born 1791, died 1867. An English scientist famous for his discoveries in chemistry, electricity and magnetism. He first produced the rotation of the magnetic needle around the electric current in 1821, based upon Oersted's discovery of electromagnetism in 1820; he discovered electromagnetic induction (1831), a principle upon which is founded the development of dynamo machinery; specific inductive capacity (1838); magnetic polarization of light (1845); diamagnetism (1846). He was a brilliant experimenter, and contributed greatly to the knowledge upon which is based present-day practice of electricity.

Michael Faraday.

FEEBLY DAMPED—In spark transmission, signals that are only slightly damped. Where each succeeding oscillation in a train of oscillations is of slightly lesser *amplitude* (q.v.) than the one preceding it, the signals are said to be slightly damped to distinguish them from signals that are highly damped and thus die out quickly. (See *Damping*, also *Damped Waves.*)

FEED—In electrical parlance, to furnish with a current of electricity. (See *Feed-Back.*)

FEED-BACK—A term applied in radio to the coupling of one circuit to another, whereby a portion of the current present in one circuit is fed back or returned to the other circuit. The common use of the term is in reference to regeneration, where a certain portion of the radio frequency current present in the plate circuit of a detector is returned to the grid circuit of the tube by some form of coupling between the two circuits. It may also apply to the tendency of high-frequency currents to become transferred from one circuit to another due to undesired coupling between two or more circuits. The illustration shows a simple feed-back or regenerative circuit. The incoming signal energy is transferred from the primary of the coupler to the secondary by means of inductive coupling and tuned by the condenser C. A grid condenser and leak GC are placed in the grid circuit in the customary manner. A feed-back or reaction coil T is placed in the plate circuit and inductively coupled to the main coil. A portion of the radio frequency current in the plate circuit or output circuit is transferred back to the grid circuit by the coupling effect between the coil T and the secondary of the coupler. This serves in effect to reduce the resistance of the grid circuit and produces more current in the output or plate circuit, thus acting as a form of amplifier. (See *Regeneration, Tickler.*)

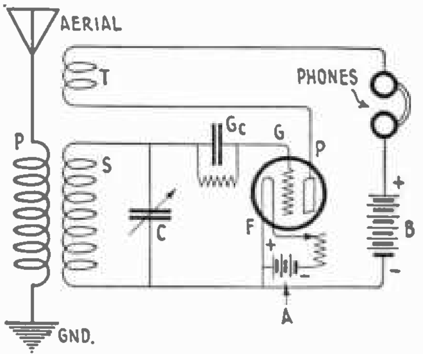

Regenerative circuit in which the coil "T" returns, or "feeds back" to the grid circuit a portion of the radio frequency current in the plate circuit, resulting in increased amplification.

FEED-BACK COIL—An inductance coil placed in the plate circuit of a regenerative receiver and arranged in inductive relation to the grid coil (secondary of tuner) for the purpose of effecting feed-back or regeneration. (See *Feed-Back*, also *Regeneration.*)

FEED-BACK COUPLING—Any process by which the plate circuit of a detector is coupled to the grid circuit to allow a portion of the output current to be transferred back to the grid circuit to produce regeneration. (See *Feed-Back*, also *Regeneration.*)

FEED-BACK EFFECT — The effect of feeding back of currents from one circuit to another as in the case of a *regenerative circuit* (q.v.). Generally the expression "feed-back effect" refers to an undesired feed-back. This may be the action of radio frequency components of the currents present in one part of a circuit being fed back to another part of the circuit due to the coupling effect of various coils or other devices used in the circuit. (See *Feed-Back*, also *Regeneration.*)

FEED-VOLTAGE MODULATION — A process of modulating or varying the amplitude of a radio frequency alternating current to correspond to any wave form of speech vibrations as in radio broadcasting. The system involves the introduction of additional power into the circuit of the radio frequency generator until the desired wave form variations are obtained. This is done particularly in the case of a vacuum tube transmitter, in which case the voltage supplied to the plate of the tube is altered (increased or decreased) to accomplish the purpose. (See *Modulation.*)

FERRO-MAGNETIC MODULATOR—A device for modulating or varying the amplitude of radio frequency current to correspond in wave form to the speech or music vibrations in broadcasting. The system makes use of what is known as the *hysteretic energy absorption* of iron, or in some cases utilizes the variation of inductance of iron-core coils. (See *Modulation*, also *Ferro-magnetic Substances.*)

FERRO-MAGNETIC SUBSTANCES — Elements or compounds such as iron, nickel, cobalt and others which are strongly attracted by magnetic fields of force. Such substances vary widely in their magnetic properties when subjected to temperature changes. Soft iron loses its magnetization almost immediately on removal of the magnetizing source, and for this reason it is used as a pole piece in an electromagnet. (See *Electromagnet*, also *Magnetism* and *Magnetic Properties.*)

FESSENDEN, REGINALD AUBREY. —Canadian-American Radio expert. Born at Milton, Canada, October 6, 1866, and educated at New York and Port Hope, Ontario. Fessenden became inspecting engineer for the Edison Company, New York, and afterwards professor of physics and electrical engineering at Western University, 1892. Professor Fessenden is the author of a well-known system of wireless, and below are briefly described some of the patents bearing his name.

In 1906 and 1907 Fessenden invented a number of microphone transmitters which carried heavy currents for long periods, and also a heavy current telephone relay which allowed the controlling of heavy currents by means of small currents originating in an ordinary microphone circuit or coming from a telephone line. One of these transmitters was called by Fessenden a trough transmitter. It consisted of a soapstone annulus to which were clamped two plates having platinum-iridium electrodes. Through a hole in the center of one plate passed a rod attached at one end to a diaphragm, and at the other to a platinum-iridium spade. The two outside electrodes were water-jacketed. This form of transmitter required no adjusting, all that was necessary being to place about a teaspoonful of carbon granules in the center space. It was able to carry as much as 15 amperes continuously without articulation falling off, and had the advantage that it never packed. By a combination of the trough transmitter and a differential magnetic relay, Fessenden produced a transmitting relay for magnifying very feeble currents. An amplification of fifteen times is possible without any loss of distinctness. Fessenden is also responsible for a duplex system of Radio telephony, and the *heterodyne* method of reception is due to him. Fessenden has written largely on radio subjects, and is one of the leading authorities on both transmission and reception.

Reginald Aubrey Fessenden.

FIBER—or **FIBRE**—A term used for a variety of substances, composed of fine slender thread-like materials, particularly a composition of vulcanized or compressed paper. It is much used in radio as an insulating material for panels and connection blocks. It is not as efficient as insulating medium as bakelite or some forms of rubber composition, but is durable, cheap and comparatively easy to drill. (See *Bakelite.*)

FIELD—The space occupied by electric or magnetic lines of force. (See *Electric Field.*)

FIELD DENSITY—The strength of an electric or magnetic field. It is measured by the number of electric or magnetic lines of force contained in a given cross-sectional area. (See *Density. Flux*, also *Electric Field* and *Field, Magnetic.*)

FIELD, ELECTRIC — The space traversed or occupied by electric lines of force. (See *Electric Field.*)

FIELD, ELECTROMAGNETIC — The total area traversed by electromagnetic lines of force. The region in which electromagnetic lines of force are exerted. (See *Field*, also *Electrostatic Force.*)

FIELD, ELECTROSTATIC—The region in which electrostatic forces are present or exerted. The total area traversed by electrostatic lines of force. (See *Field*, also *Electrostatic Flux.*)

FIELD MAGNET — An *electromagnet*, usually a core of soft iron surrounded by a winding of insulated wire, employed in generating machines to produce a strong magnetic field. As the *armature* of a *generator* revolves the conducting surfaces pass through the *magnetic field* created by the *field magnets.* This field may be merely the ordinary magnetic force due to permanent magnets, but is usually due to the powerful magnetic lines of force resulting from passage of electric current through electromagnets. (See *Alternator*, also *Alternating Current, Theory of Production* and *Electromagnet.*)

FIELD, MAGNETIC—The region through which magnetic flux passes. Usually refers to the space (air) through which the magnetic flux lines pass as distinguished from the activity in the iron path tself. (See *Field*, also *Magnetic Flux.*)

FIELD REGULATOR or RHEOSTAT—Any device for varying the magnetic strength of the field magnets in a dynamo or electric motor, by which variations certain changes may be accomplished in the operation or performance of the machine. A field rheostat is often used in an electric motor to vary the strength of the field and thus either increase or decrease the speed of the machine without the necessity of changing the applied voltage. If the strength of the field (magnetic) is reduced by insertion of resistance, the armature or rotating part speeds up in an attempt to generate the same *counter-electromotive force* (q.v.), which means greater speed. On the other hand, if the resistance is removed—that is, the strength of the field increased—the fact that the armature is running through a stronger field reduces the speed of rotation, the same counter-electromotive force being present. The use of a field regulator or rheostat in connection with a generator is for the purpose of varying the output voltage without the necessity of altering the speed of the machine. The illustration shows the connection of a field rheostat or regulator in the circuit of the *shunt wound generator* (q.v.). (See *Generator, Rheostat, Resistance.*)

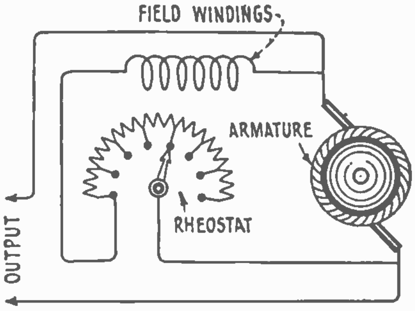

Illustration showing method of connecting a field rheostat or regulator in the circuit of a shunt wound generator.

FILAMENT — A fine wire, one of the elements of a vacuum tube, which is heated to throw off electrons. The term is also used to refer to the fine wire used in an incandescent lamp. The filament of a vacuum tube is generally made of tungsten and is supported by glass pillars within the tube in such manner as to be insulated from the other elements. Filaments are made to be heated with various voltages from 1.1 to 6 volts in most receiving tubes, but much higher in the case of high power transmitting tubes. The filament may consume considerable current as in the old styles of six-volt tubes which consumed one ampere. The modern tubes have thoriated filaments, tungsten treated with thorium, which permits maximum efficiency with a minimum of current consumption. When the filament is heated in the vacuum tube, by having the proper amount of voltage applied, electrons are thrown off and fill the space between filament and grid and grid and plate as shown in the illustration. (See *Electron Theory*, also *Vacuum Tube, Theory of Operation.*)

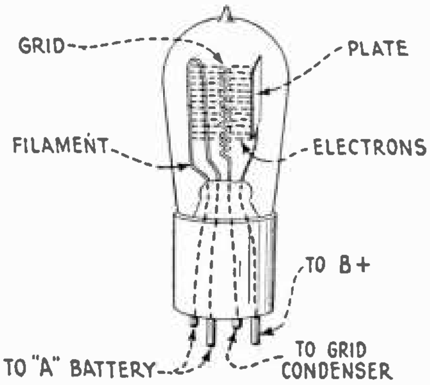

When an "A" battery is connected to the filament terminals, heating the filament, electrons are thrown off and fill the space between filament, grid and plate.

FILAMENT BATTERY—The battery or cell used in lighting the filament of a vacuum tube. These range from a *dry battery* (q.v.) having a voltage of one and a half volts, to storage batteries of six volts. Filament batteries are connected across the filament of a vacuum tube in the manner shown in the illustration. The rheostat is inserted to permit the supply of current to the filament to be varied and thus change the operation of the tube. (See *Filament*, also *Vacuum Tube, Theory of Operation.*)

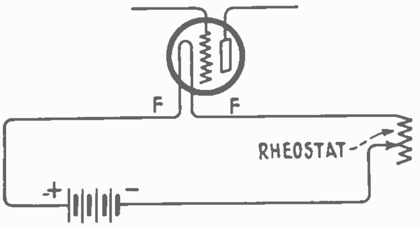

Method of connecting "A" battery in series with rheostat to regulate filament of vacuum tube.

FILAMENT CONSUMPTION — The amount of current consumed by the filament of a vacuum tube. The current requirements of vacuum tubes are always stated by the manufacturer, either in amperes or fractional amperes, or in the case of extremely low consumption, in thousandths of an ampere (milliamperes). Some typical examples are the UV201A and C201A types, which require one-quarter ampere at five volts; the WD11 and 12 and C11 and 12 types, which require one-quarter ampere at approximately 1.1 volts, and the UV200 and C200 tubes, which require one ampere at six volts. Of the so-called standard types, the UV199 and C199 tubes consume the least current, requiring but sixty milliamperes or .06 ampere, with from three to four and a half volts. The statement that a certain tube consumes one-quarter ampere indicates that in operation one-quarter ampere is consumed each hour, this current being supplied by dry cells or storage batteries known as the *"A" battery* (q.v.).

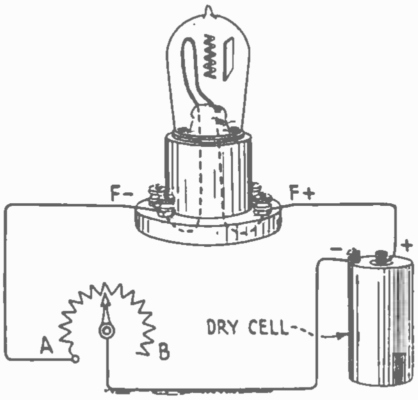

The current produced by the dry cell battery, flowing through the rheostat and filament, is consumed by the filament, at a rate depending upon the type of tube.

It will be apparent from the above that the more current consumed by the filament the greater will be the drain on the dry cells or storage battery. For this reason, and also to lengthen the life of the filaments, it is well to operate the tubes at the lowest efficient brilliancy. The brilliancy of the filament may be controlled by adjustment of the filament rheostat. In the illustration the usual "A" battery, which in this case is a dry cell, is shown connected to the filament of a vacuum tube with a rheostat in series between one filament connection and the negative terminal of the dry cell. Now as the arm of the rheostat is moved from A toward B resistance is introduced into the circuit between the filament and dry cell. As this resistance increases the current in the circuit of the filament will decrease according to *Ohm's Law* (q.v.); i.e., Current = Voltage ÷ Resistance, the current in amperes or fractional amperes and the resistance in ohms. (See *Ampere, Current, Filament Resistance or Rheostat* and *Vacuum Tube.*)

FILAMENT CURRENT — The current used to light the filament of a vacuum tube. Different tubes require different amounts of current; that is, one tube may consume a small fraction of an ampere and another may require a full ampere. Of the tubes most commonly used, the UV199 type uses the least current, this tube requiring only .06 of an ampere each hour. The WD11 and 12 types require one-quarter ampere per hour and also the UV201A, while the now obsolete UV200 and UV201 required a full ampere. The 199 type operates with filament voltage of 3 volts, although it is often used with three dry cells in series which deliver 4½ volts; the WD11 and 12 require 1.1 volts each, which is usually furnished by a dry cell, while the UV201A type is operated with a six-volt storage battery. (See *Vacuum Tubes, Types of*, also *Vacuum Tube, Theory of Operation.*)

FILAMENT, FORMULA FOR ELECTRON FLOW—The extent of electron emission—that is, the number of electrons—emitted from a hot filament in a vacuum tube is subject to certain definite laws. In ordinary practice the actual number of electrons are not referred to, the ability of the filament to emit electrons being measured in fractional amperes of output current. Thus a tungsten filament having a diameter of .0125 centimeters operating with a power dissipation of 3.1 watts per centimeter of length has been found to emit a current of .03 amperes. This result was obtained with a filament having a clean surface and with no absorbed gas on its surface. The formula for electron flow most commonly used is as follows:

$$I_B = AT^{1/2} \epsilon - \frac{b}{T}$$

I_B is the electron current in milliamperes per square centimeter of filament surface, T is the absolute temperature and A and b depend on the metal used for the filament.

FILAMENT RESISTANCE or RHEOSTAT—A device placed in the filament circuit of a vacuum tube, connected between the filament battery and either filament lead on the tube socket for the purpose of reducing or controlling the filament voltage. The illustration Fig. 1 shows one type of filament rheostat which is used to permit variation of the filament current. Fig. 2 shows the manner of connecting such resistance in the filament circuit of a vacuum tube. Fig. 3 illustrates a type of filament control known as an automatic rheostat. In this form of filament resistance a fine wire is placed in a glass tube filled with an inert gas. The wire has the property of increasing its resistance rapidly with any increase in the current passing through it. It is thus possible to arrange such a device to pass only a fixed amount of current for any type of vacuum tube. The device is thus not only an automatic rheostat, but also a safety valve which prevents excessive current reaching the filament and injuring it. These controls are coming into widespread use due to these features and the fact that they reduce the number of controls necessary to operate a receiver. Fig. 4 shows a resistance which is used in series with a rheostat having small resistance in order to enable its use with tubes requiring less current to operate. (See *Filament*, also *Vacuum Tube.*)

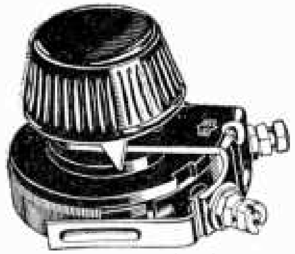

Illustration by Courtesy of The Cutler-Hammer Mfg. Co.

Fig. 1. A type of rheostat used to control the flow of current to the filament of a vacuum tube.

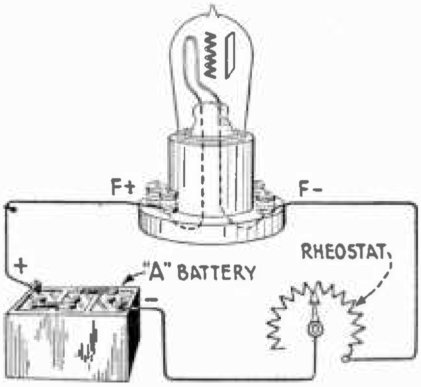

Fig. 2. Showing method of connecting a rheostat in "A" battery circuit of a vacuum tube.

Illustration by Courtesy of the Radiall Co.

Fig. 3. A type of automatic rheostat.

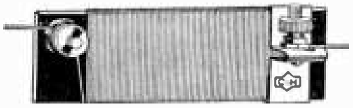

Illustration by Courtesy of The Cutler-Hammer Mfg. Co.

Fig. 4. Resistance used in series with a rheostat having a small resistance in order to enable its use with tubes requiring less current.

FILAMENT, THORIATED—A filament or heating element of a vacuum tube in which a certain amount of *thorium* (q.v.) has been incorporated. The introduction of this rare metal into the filament in small quantities gives it the property of emitting electrons at a relatively low temperature. From a practical standpoint this permits a vacuum tube to be operated with a low *filament consumption* (q.v.). A common example of the use of thorium in filaments is that of the UV201A or C201A vacuum tubes, which operate with a filament voltage of 5 volts and consume but one-quarter ampere of current, giving essentially the same results as the former type UV201 and C201 tubes which consumed one ampere. The value of thorium used in this manner is readily appreciated, the current consumption being materially lessened and the life of the tube conceivably lengthened. (See *Thorium Treatment of Filaments*, also *Vacuum Tubes, Type UV201A and C201A.*)

FILTER—In radio, a device arranged to permit passage through an electrical circuit of currents of certain frequency or *frequencies* (q.v.) while excluding others of different frequencies, or any device for producing smooth direct current from a generating source such as a *direct curent dynamo*, or for smoothing out *rectified direct current.* Filters of one sort or another are much used where rectified direct curent or direct current from a generator is to be used for operation of tubes in radio transmission or reception. In the case of rectified direct current, that is the current derived from any device for changing alternating to direct current, there is always present a tendency on the part of the current to surge or ripple. In other words, it is not smooth and of constant voltage, or there may be present a generator hum which will seriously interfere with proper operation of a tube. The illustration shows the conventional method of filtering rectified direct current for use as plate supply with vacuum tubes in receiving sets. This system may also be used in connection with the output of a direct current generator or dynamo. A filter circuit may be arranged to pass a band of frequencies (see *Band Pass Filter*) or it may be made to pass either high or low frequency current. (*See Filter, High Pass;* also *Filter, Low Pass.*)

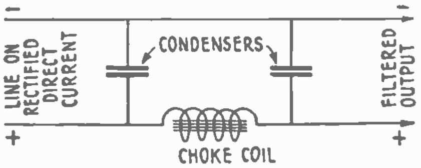

Conventional method of filtering rectified direct current for use as plate supply of a vacuum tube.

FILTER, BAND PASS—See *Band Pass Filter.*

FILTER CONDENSER—The condenser, either fixed or variable, forming a part of a circuit for filtering. (See *Filter*, also *Acceptor, Wave Trap* and *Rejector.*)

FILTER, HIGH PASS—A filter circuit arranged to permit current above a certain frequency to pass freely while blocking out or attenuating all frequencies below it. (See *Filter*, also *Filter, Low Pass.*)

FILTER, LOW PASS—A filter circuit arranged to permit currents below a certain frequency to pass freely while blocking currents of higher frequency. (See *Band of Frequencies*, also *Filter, High Pass.*)

FIXED COILS—Inductance coils as used in radio circuits, having no mechanical means of variation, such as in a *Variocoupler or movable coil.* (q.v.) The majority of tuned radio frequency receivers employ fixed coils, shunted by variable condensers. These are known as *tuned radio frequency transformers.* (q.v.) (See *Coupler, Loose*, also *Coupler, Vario.*)

FIXED CONDENSER — A condenser having a certain definite or fixed value of capacity. Any condenser that is not variable. (See *Condenser*, also *Condenser, Variable.*)

FIXED RESISTANCE—Any resistance device having a fixed or non-variable value. (See *Rheostat*, also *Potentiometer.*)

FIXED DISCHARGER—A fixed *spark gap.* (q.v.) A spark gap used in transmission of radiotelegraphic signals in which the electrodes or sparking surfaces are not moved during operation. (See *Disc Discharger.*)

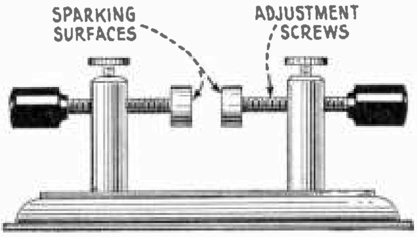

A type of fixed spark gap.

FLAME MICROPHONE—A type of microphone for radio telephone transmission, wherein the height of a coal gas flame is varied by and according to the speech fluctuations and the resulting pulsations made audible by means of a selenium cell and phones. (See *Microphone*, also *Selenium Cell.*)

FLASH SIGNALLING— See *Code*, also *Morse Light.*

FLAT-TOP AERIAL or **ANTENNA**—An aerial, generally composed of a number of wires placed adjacent to each other equally spaced on a spreader, or hoop, and suspended parallel with the earth. A flat-top aerial may be of the kind known as the *T type aerial* (q.v.) or in the form of an inverted L. (See *Aerial.*)

FLAT TUNING—A term sometimes applied where the *resonance curve* (q.v.) has a flat top, that is where the peak of resonance is not sharply defined. The effect is known as *broad tuning.* (q.v.) (See *Resonance*, also *Tuning.*)

FLEMING, JOHN AMBROSE. British radio expert. Born in Lancaster, November 29, 1849, he was educated at University College, London, the Royal School of Mines, and St. John's College, Cambridge.

From 1873-74 Fleming was demonstrator at the Royal College of Chemistry, and in 1877 he began work under Clerk-Maxwell at the Cavendish Laboratory, Cambridge. There he carried out a series of experimental researches on the British Association Standards of Electrical Resistance. In 1881 Dr. Fleming was appointed the first professor of mathematics and physics at University College, Nottingham, but the following year he joined the Edison Electric Light Company, and on the amalgamation of the Edison and Swan Companies, he was appointed advising electrician, a post he held for twenty years. In 1885 Fleming was appointed to the newly founded professorship of electrical engineering at University College, London, and he was entirely responsible for the design and equipment of the new electrical and engineering laboratories which were opened in 1893. In 1892 he was elected Fellow of the Royal Society.

Dr. Fleming has been very closely associated for many years with radio telegraphy and telephony, and has since 1899 acted as the scientific advisor to Marconi's Wireless Telegraph Company. He is well known as the inventor of the thermionic two electrode vacuum tube, or glow-lamp detector, which was one of the greatest steps forward in radio telephony.

Dr. Fleming has written a very large number of papers and books on wireless telegraphy and telephony, and his "Principles of Electric Wave Telegraphy" is the standard treatise on the subject. He has also published an "Elementary Manual of Radio Telegraphy and Radiotelephony"; "The Wonders of Wireless Telegraphy"; "The Thermionic Valve and Its Development"; "A Pocket Book for Wireless Telegraphists," and other books. As a lecturer at the Royal Society of Arts he has obtained a world-wide reputation for his lectures on electric oscillations and electric waves, Hertzian wave telegraphy, high-frequency measurements, etc. He has given many lectures on wireless at the Royal Institution.

John Ambrose Fleming.

Fleming was awarded the Hughes gold medal of the Royal Society in 1910 as an acknowledgement of the value of his work in electrical science and engineering, and he has twice been awarded the Institution Premium of the Institution of Electrical Engineers, the highest award for communicated papers, and he is an Albert Gold Medallist of the Royal Society of Arts awarded to him for the pioneer invention of the thermionic valve.

FLEMING'S RULE—A rule for determining the direction of the *induced current* (q.v.) in a circuit. Hold the right

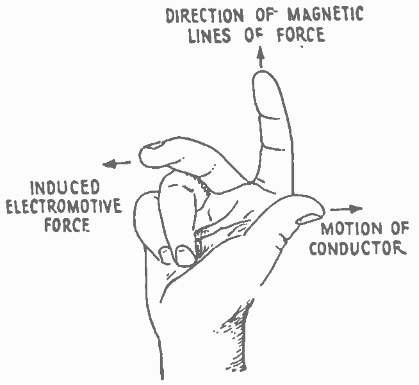
DIRECTION OF MAGNETIC LINES OF FORCE
INDUCED ELECTROMOTIVE FORCE
MOTION OF CONDUCTOR

Method of using the right hand in determining the direction of the induced current.

hand as shown in the illustration with the thumb, forefinger and middle finger as nearly as possible at right angles to each other. If the thumb is then pointing in the direction of motion of the conductor and the forefinger points along the direction of the magnetic lines of force, the middle finger will then point in the direction of the induced electromotive force. (See *Induced Current*, also *Electromotive Force.*)

FLEMING VALVE—The forerunner of the modern vacuum tube. It was invented in 1904 by Professor J. A. Fleming and was for a short time used as a detector in place of the customary crystal or coherer types. The value of the discovery of the Fleming valve lies chiefly in its influence on the course of radio development, as in itself it was not nearly as efficient as a good crystal detector. As a result of the discovery of the Fleming valve, however, many leading scientists bent their efforts toward the discovery of better methods of detecting radio waves, culminating in the three element vacuum tube designed by Dr. Lee De Forest.

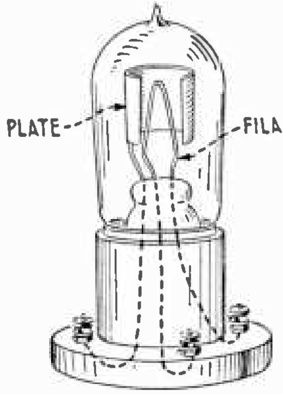

Fig. 1. The Fleming valve, which contained but two elements; i. e., filament and plate. It was the forerunner of the modern vacuum tube.

The Fleming valve consists essentially of an incandescent lamp using a filament of carbon, tungsten or tantalum which can be rendered incandescent by an electric current. The filament is sealed in a glass tube which is highly evacuated, that is from which practically all air has been excluded. Surrounding the filament, but separated from it is a metal cylinder, both ends of the filament and the plate being attached to outside connections by means of leads sealed in the tube. The illustration, Fig. 1, shows the general arrangement of such a tube. The operation of the tube is due to the emission of electrons from the heated filament. It was demonstrated by several investigators that a filament made of certain metals has the property of throwing off minute particles of negative electricity or electrons when heated to a relatively high temperature. It was also found (see *Electron*) that these electrons are thrown off in greater profusion when the filament is placed in high vacuum. It was also found that the escape of negative electrons from the filament would in time leave it positively charged, this charge having an attraction for the negative electrons and preventing their escape from the filament. Hence it was necessary to apply a charge of negative electricity to the filament. The electrons were now found to escape in great number, but had the effect of forming a *space charge* (q.v.) of negative electricity around the filament, which in time served to repel the escape of any more electrons. To counteract this tendency a strong positive charge was applied to the metal plate surrounding the filament. This made the plate positive in respect to the electrons and served to attract them away from the filament. The stream of electrons thus affords a path for current to flow from filament to plate, and the action of the tube as a detector of high frequency waves depends on this principle. If the tube is connected as indicated in the illustration, Fig. 2, it will serve to rectify the incoming signals, that is change them from high frequency alternating currents, to pulsating direct current. The action is as follows: The aerial is connected to the secondary circuit by means of a standard coupler. The coils, marked L2, are the primary and secondary of this coupler; C1 and C2 are variable condensers, F the filament of the Fleming valve, B the battery for lighting the filament of the valve and also to supply voltage to the plate, T a pair of headphones, P a 400 ohm potentiometer and R a standard filament rheostat. The incoming signals in the form of high frequency currents, change direction many times each second. Thus, at one instant the plate will have a positive charge while the next instant it will have a negative charge as a result of the changes in direction of the currents impressed across the tube from the antenna circuit. While the plate is positive, electrons will be attracted away from the filament, which at that instant will be negative with respect to the plate. The current thus flows for an instant, stopping when the impressed current changes its direction and makes the plate negative. In this manner we have a flow of current from filament to plate in one direction in the form of surges appearing at every other alternation or change of current. The diaphragms of the phones will respond to these uni-directional surges and will give out sounds almost identical with the original signals transmitted. The tube thus serves as a rectifier of the incoming signals and is known as a detector of high frequency waves, or oscillations. The Fleming valve principle is used in rectifying tubes for changing alternating current to direct current for charging batteries, its use as a detector for radio reception having died out almost entirely. (See *Charger, Storage Battery, Detector, Tungar, Rectigon,* also *Vacuum Tube.*)

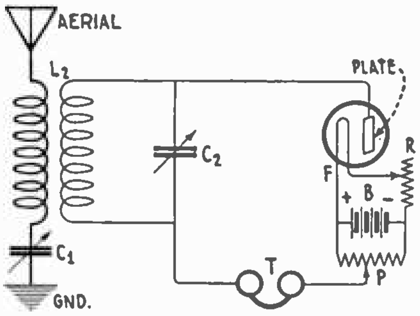

Fig. 2. Method of connecting a Fleming valve in a radio receiving circuit.

FLEWELLING CIRCUIT—A modification of the Armstrong Super-regenerative circuit, designed by the American experimenter, Flewelling. In this system, the tendency toward *self-oscillation* (q.v.) is retarded by the use of fixed condensers which are arranged to have a *damping* (q.v.) effect on the grid of the detector tube, automatically controlling the oscillations of the tube. The circuit controls the self-oscillation of the tube without the necessity for the separate control tube used in the Armstrong circuit. The illustration shows the original Flewelling circuit. It is essentially a standard regenerative circuit with certain necessary modifications.

The coils L1 and L2 give magnetic coupling between the plate circuit and the grid circuit. C1 is a variable condenser for tuning and condensers C2, C3 and C4 are of the fixed type and afford a path of relatively low *impedance* (q.v) for the high frequency currents. C5 is the customary grid blocking condenser across which a grid leak is shunted. The usual telephone receivers T, the plate battery B and the filament battery A are included in the circuit in the customary manner. The coils L1 and L2 may be two honey-comb coils having respectively 50 and 75 turns. The tuning condenser C1 should be .0005 mfd. for the broadcast band of wave-lengths and the loop may be almost any form.

In operation three forms of current pass through the plate circuit including the coil L2, namely, the direct current from the plate battery B, the uni-directional surges of rectified current due to the action of the tube as a rectifier and the high frequency currents transferred through the medium of the magnetic coupling of the two coils L1 and L2. The battery current is present only in the plate circuit, the condensers C2, C3 and C4 blocking the direct current from the grid circuit. These condensers also block the rectified currents but afford two paths for the high frequency currents or oscillations. The condensers C2, C3 and C4 have definite values, respectively .005, .005 and .006 micro-farads, to allow a by-pass through the grid to the filament, thus causing accumulation of electrons on the grid and stopping oscillation of the tube. The charge then leaks off by means of the grid leak connected across C5, the tube then building up oscillations again and continuing the process during operation. The high resistance connected across C3 is variable and permits control of the by-pass of high frequency current of the grid. Careful adjustment is necessary to obtain maximum signal strength without distortion. (See *Armstrong Circuits, Regenerative Circuit,* also *Super-Regenerative Circuit.*)

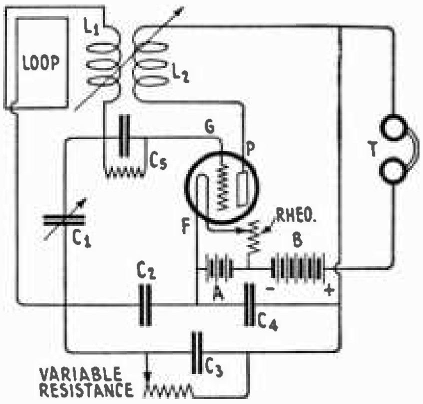

The Flewelling circuit.

FLEX—Term used in Great Britain and occasionally in this country to refer to a number of small wires twisted or stranded together to form a single flexible conductor. (See *Flexible Wire.*)

FLEXIBLE LEAD—Any conductor, particularly the connections from rotary parts of couplers or condensers, in which the conductor is made up of a number of fine wires in such a way that the necessary bending in operation of the unit will not snap the wire as would be the case with a solid conductor. (*See Pig-tail*, also *Flexible Wire.*)

FLEXIBLE WIRE—Any conductor composed of a number of fine wires to permit constant handling and bending without breaking. (See *Flexible Lead.*)

FLOAT, BATTERY — Term sometimes applied to the bulb or float in a hydrometer for testing the condition of storage batteries. (See *Hydrometer.*)

FLOW OF CURRENT—An expression used to denote the passage of electric current along a conductor. (See *Current, Direction of Flow* (*Assumed*), also *Current.*)

FLOW OF MAGNETIC FLUX—The passage of lines of magnetic force through any magnetic circuit. (See *Flux*, also *Field, Magnetic.*)

FLUCTUATING CURRENT—An electric current that does not have steady value, rising and falling in strength or pressure. A current that is not constant. (See *Constant Current*, also *Voltage, Constant.*)

FLUORESCENT SCREEN—A screen or surface coated with fluorescent matter, used to exhibit shadows cast by rays from various types of tubes. Its application in radio is chiefly in the *Cathode Ray Tube.* (q.v.)

FLUX—A very broad term in physics. The three more common examples of flux are: (a) to denote *electrostatic or magnetic flux* (q.v.); (b) a substance used in soldering for the purpose of making the solder flow readily and making it adhere to the parts being united, and (c) in metal smelting, a mineral—customarily chalk or limestone—added to the charge in the furnace to absorb mineral impurities and dispose of them in the form of slag. *Flux* is also used to refer to the amount of heat or light flowing or passing through a given distance or area in a given time. (See *Electrostatic Flux, Flux Density*, also *Flux, Magnetic.*)

FLUX DENSITY—The number of lines of force per square centimeter of sectional area of a magnetic path. The unit of magnetic flux in the C.G.S. (Centimeter Gram Second) or electromagnetic system of units is the Maxwell. It is generally referred to as a "line." As a magnetic flux is merely the effect of magneto-motive force, we can take the rule that the magnetomotive force is 1.257 x ampere turns as the rule for determining the intensity or density of magnetic flux. (See *Flux; Field, Magnetic*, also *Ampere Turns* and *Maxwells.*)

FLUX DISTRIBUTION—The distribution of lines of force in a magnetic or electrostatic field of force. (See *Field*, also *Flux.*)

FLUX, ELECTROSTATIC — Another term for electric displacement as in the case of a condenser. The density of electrostatic flux refers to the intensity of an *electrostatic field.* (q.v.) (See *Flux Density, Field, Electrostatic*, also *Electric Displacement.*)

FLUX LEAKAGE—Any loss or dissipation of magnetic or electrostatic flux. (See *Flux, Electrostatic*, also *Flux, Magnetic.*)

FLUX LINES—The lines of force in a field of force in which electrostatic or electromagnetic force is exerted. An alternate term for *lines of force.* (q.v.)

FLUX, MAGNETIC—The term broadly used to denote the grouping of magnetic lines of force in a *magnetic field.* (q.v.). The most common example of magnetic flux is in the case of an ordinary bar magnet as shown in the illustration Fig. 1. Here the lines of force or magnetic flux create a magnetic field indicated by the dotted lines. The number of lines of force in a field of unit area is used as a basis for determining the density of the flux. (See *Flux Density.*) This effect or magnetic flux field is also present when a current of electricity is allowed to circulate around a coil of wire, the larger the current and the greater the number of turns, the more powerful the flux created. The illustration Fig. 2 shows a coil of wire through which current is assumed to be passing, the dotted lines indicating the flux field. (See *Choke Coil, Transformer, Magnet*, also *Flux, Electrostatic* and *Magneto-motive Force.*)

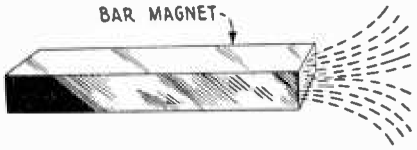

Fig. 1. Bar magnet, showing how the lines of force, or magnetic flux, create a magnetic field as indicated by dotted lines.

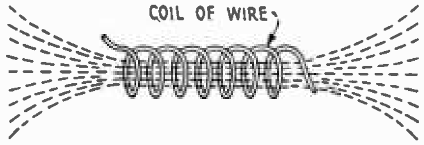

Fig. 2. Coil of wire through which current is assumed to be passing, the dotted lines indicating the flux field.

FLUX MAGNETISM — The magnetic flux, or total number of lines of force passing through a magnetic circuit. It is equal to the product of the magnetic density (see *Flux Density*) and the cross sectional area of the magnetic path. It is also referred to as magnetic flow. (See *Magnetic Flow.*)

FLUXMETER—A device for measuring flux density. The Grassot fluxmeter uses an *exploring coil* (q.v.) in conjunction with a special form of moving coil galvanometer. The coil is suspended on a silk fiber and its movements actuate a pointer across a scale. (See *Flux, Magnetic.*)

FLUX, SOLDERING—See *Soldering Flux, Flux.*

FOCUS or FOCI—The point or points at which converging lines meet. The magnetic foci are the points on the earth's surface in the vicinity of the magnetic poles, where the magnetic force of the earth is the greatest. Essentially a mathematical term.

FORCE—An influence exerted upon a body so as to produce a change or a tendency to change its state of rest or motion. This is mechanical force. In electricity we have its electrical analogy—*electromotive force.* This is the force which causes motion of electricity and sustains the electric currents against the resistance of a circuit. (See *Electromotive Force.*)

FORCE, MAGNETO-MOTIVE—The force which produces magnetic flux as in the case of a magnet moving through a magnetic circuit. The abbreviation is M. M. F. and its unit is the *gilbert.* (q.v.) (See *Magneto-motive Force.*)

FORCED OSCILLATIONS — Electrical vibrations or high frequency oscillations which are imposed upon an *oscillatory circuit* without being in tune or resonance with the natural period of vibration of the circuit. The opposite of a free oscillation in which case the circuit oscillates freely on its own account at the same frequency as the imposed or incoming oscillations or one of the *harmonics* (q.v.) of this frequency. Free and forced oscillations may be present in the same oscillatory circuit, the forced oscillations being set up during the first stages of induction of free oscillations. (See *Free Oscillations*, also *Oscillatory Circuit.*)

FORM FACTOR—The ratio of the *root-mean-square value* (q.v.) to the algebraic mean ordinate taken over one alternation (half cycle), starting with zero value of current. It is the ratio of the r.m.s. value of an alternating electromotive force, current, etc., to its true mean value. (See *Wave Analysis.*)

FORMULA—A principle or rule expressed in mathematical terms or chemical symbols. Formulae are much used in radio to state a problem as, for example, a formula for calculation of wave-length, which is written: Wave-length$=1884\sqrt{LC}$. This states the problem or rule for the determination of wave-length. Here we have certain values represented by symbols. The capacity is represented by C, the inductance by L, the wave-length in meters being the product of the square root of the product of these two multiplied by the factor 1884. (See *Measurement of Wave-length.*)

FOSTER BRIDGE—A form of *Wheatstone Bridge* (q.v) of the slide wire variety designed to permit comparison of two low value resistances that are nearly equal. It was devised by Professor Carey Foster and is used to determine the difference between two resistances as distinguished from the usual bridge for measuring the resistance of a single unit. (See *Wheatstone Bridge.*)

FOUCAULT CURRENT—Another name for *eddy currents*, the *electro-magnetic currents* due to the circulation of induced current in the mass of a conductor. Foucault first proved that these currents were productive of losses of energy as they cannot be utilized except in certain forms of meters. (See *Damping of Instruments.*) Alternating currents tend to set up eddy currents in metal masses near them. Thus in the case of a transformer, eddy currents are produced in the iron core due to the induced currents circulating through its mass. These currents are wasteful and means are therefore used for reducing them. The core of a transformer may be made of numerous flat pieces called laminations, which serve to prevent the free movement of eddy currents in the core. (See *Eddy Currents*, also *Transformer* and *Core Losses.*)

FOUR ELEMENT or FOUR ELECTRODE TUBE—A vacuum tube having four elements instead of the usual three. The standard tube has three elements, namely, grid, filament and

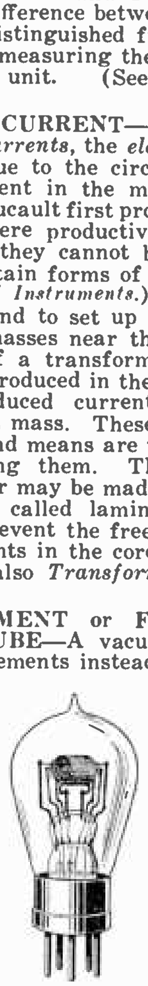

A four element vacuum tube.

plate. The four element tube has an additional grid element. The illustration shows one form of tube having a double grid arrangement. Such a tube can be used for high and low frequency amplification and as a detector simultaneously or it may be used in conjunction with the Solodyne system, which does not use the customary B battery to supply potential to the plate element. (For more complete explanation of the action of a four element tube in the latter manner see *Solodyne*). (See also *Vacuum Tube.*)

FOURTH CIRCUIT—A method of controlling oscillation of a regenerative detector used in the *Cockaday Circuit.* (q.v.)

FRAME AERIAL—An alternate name for *loop aerial,* a type of aerial consisting of a number of turns of wire wound on an insulated frame and used in place of the regular outdoor aerial with sensitive receiving sets. (See *Loop Aerial.*)

FREE ALTERNATING CURRENT—Term applied to alternating current assumed to be produced by the application of a single impulse of electricity to an *oscillatory circuit.* (q.v.) See *Free Oscillations.*)

FREE ELECTRONS—Name given to the outer electrons of an atom, which become detached under various chemical transformations and are assumed to participate in electrical effects or phenomena. Free electrons are so named to distinguish them from the bound electrons of the proton or positive nucleus of an atom. (See *Electron,* also *Proton.*)

FREE MAGNETISM—The portion of magnetism which does not follow the magnetic path through the metal of a magnetized body, but leaves the surface of the magnet. Sometimes referred to as *surface magnetism.* (See *Magnetism.*)

FREE OSCILLATIONS — Oscillations having the natural frequency of the circuit in which they occur. (See *Forced Oscillations.*)

FREQUENCIES—The plural of the term denoting the number of complete cycles per second in an alternating current. In referring to vibrations, particularly audible or so-called *voice frequencies,* the term is used to imply all frequencies that are normally considered audible. (See *Frequency.*)

FREQUENCY—A term adapted from academic English to use in electrical practice. Thus, the number of times a certain action is repeated per unit of time. In the case of an *alternating current* (q.v.) the current reverses its direction of flow at certain definite intervals. If the reversing process takes place 120 times per second—that is 60 reversals from negative to positive and vice versa—the frequency of the current is said to be 60 cycles. In ordinary commercial practice the majority of alternating current generators produce an alternating current or E.M.F. having a frequency from 25 cycles per second to 60 cycles per second. In radio transmission the alternating current or E.M.F. (Electro Motive Force) has an infinitely higher frequency—the reversals taking place many thousands of times each second. (See *Alternating Current, Theory of Production,* also *Alternation, Cycle* and *Periodicity.*)

FREQUENCY, AUDIO—In radio, a frequency within the limits at which audible sounds are produced. It is generally considered as a frequency less than 10,000 cycles per second, although some authorities refer to frequencies up to 16,000 or 20,000 cycles as being within the limits of audibility. For ordinary practice, the human ear is capable of hearing sounds only up to about 2,500 cycles per second. In receiving sets the audio frequency currents appear only in the circuits carrying the rectified currents, although some types of super-heterodyne receiver use an intermediate frequency as low as 30,000 cycles per second, these signals then being rectified and made audible in the telephone receivers. (See *Frequency Circuit, Audio,* also *Frequency, High.*)

FREQUENCY, ALTERNATOR OF—The frequency of an alternating current dynamo or alternator is expressed in cycles per second. With each complete revolution of the *armature* a complete *cycle* or double *alternation* of current is generated. (See *Alternator.*) It is therefore a simple matter to determine the frequency of alternating current generated when the number of revolutions per second and the number of field poles or magnets is known. The frequency will be $f=\frac{N \times S}{2}$, N being the number of field poles and S the number of revolutions per second. The divisor 2 is used to convert the result in *alternations* per second to cycles per second, one cycle representing two alternations. (See *Alternating Current, Theory of Production of,* also *Frequency.*)

FREQUENCY BAND—A term used to denote all the frequencies lying within a certain minimum and maximum limit. Thus, in broadcasting, the frequency band will be the frequencies corresponding to the wave-lengths within the broadcast band. This will be generally between 550,000 and 1,500,000 cycles, or a wave-length band of approximately 200 to 550 meters. (See *Band of Frequencies.*)

FREQUENCY CARRIER—In radio broadcast transmission, the radio frequency (high frequency) wave with which the audio frequency speech variations are superposed. (See *Broadcasting, General Treatise on Methods,* also *Carrier Wave.*)

FREQUENCY CHANGER—A device for converting currents at a certain frequency into currents of a higher or lower frequency, without change of phase or voltage. A frequency converter. Generally an *alternator* driven by a *synchronous motor.* The term is also applied erroneously to certain forms of super-heterodynes. (See *Ultradyne.*)

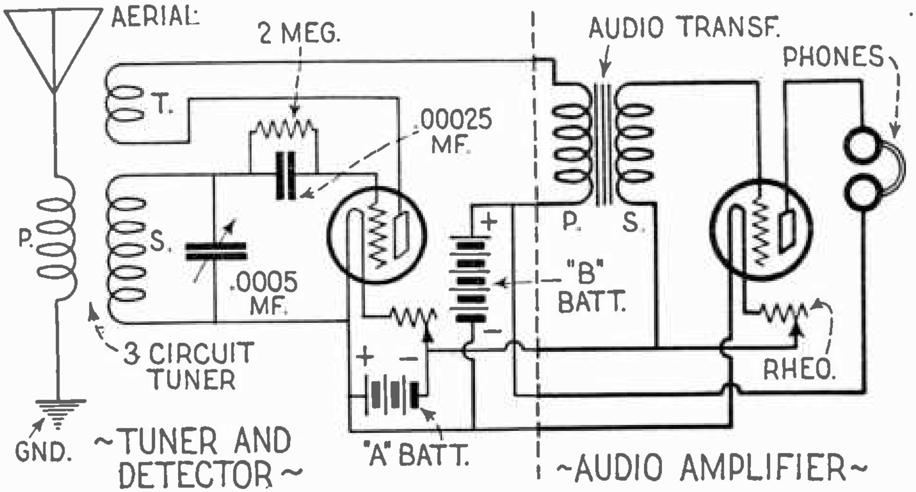

In the diagram above, the heavy lines to the right of the dotted line indicate the audio-frequency circuit of a radio receiver.

FREQUENCY CIRCUIT, AUDIO—Any circuit in a radio receiver or transmitter which carries currents within the arbitrary limits of audibility. In the diagram shown herewith, the heavy lines are the connections of the audio frequency circuit. In this part of the circuit only currents of audio frequencies are present. (See *High Frequency Component of Plate.*)

FREQUENCY CIRCUIT, RADIO—The circuits in a receiver in which only high or radio frequency currents are supposed to be present. (See *Feed-Back, Audio Frequency,* also illustration in connection with *Frequency Circuit, Audio.*)

FREQUENCY, FUNDAMENTAL—The frequency of the fundamental wave. (See *Fundamental Frequency.*)

FREQUENCY, GROUP—The frequency of the separate trains of waves in radio communication. (See *Group Frequency,* also *Wave Train.*)

FREQUENCY, HIGH—Frequencies over certain definite limits, generally from 10,000 to 20,000 cycles per second in radio usage. Commonly used as synonymous with radio frequency. It is the opposite of low or audio frequency. (See *Frequency,* also *Audio Frequency, Limits of Audibility.*)

FREQUENCY, LOW—Frequencies below certain limits, generally 10,000 or 20,000 cycles per second. (See *Frequency; Limits of Audibility.*)

FREQUENCY METER—A device for indicating the frequency of an alternating current in cycles per second. The most simple form uses a system of tuned, vibrating reeds. The illustration shows a frequency meter of the

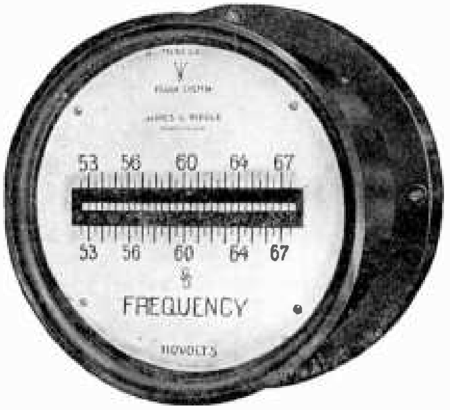

Illustration by Courtesy of Jas. G. Biddle.
A vibrating reed type of frequency meter.

reed type. These reeds may be arranged to respond to various frequencies in a similar manner to the reeds in various musical instruments. These reeds are acted upon by an alternating current magnet. When the magnet is acted upon by an alternating current sent through the windings, the reed nearest to that frequency will vibrate in sympathy with the vibration due to the alternating magnetic effect on the magnet. Such meters and also those of the *induction type* are used for comparatively low frequencies such as occur in commercial lighting systems. For measurement of the high frequencies used in radio communication somewhat different systems are used.

FREQUENCY, RADIO—An alternating current or electromotive force above certain definite limits, generally above 10,000 or 20,000 cycles per second. The line of demarcation between *audio frequency* and *radio frequency* is not very sharply defined, as, in radio usage frequencies that are actually inaudible to the human ear may yet be referred to as audio frequencies. The term is used to differentiate between the rectified or audio currents and the high frequency oscillations such as are impressed on the antenna in receiving or transmitting. (See *Audio Frequency*, also *Frequency, Low.*)

FREQUENCY, UNIT OF—Frequency of alternating currents is expressed in cycles per second and is generally written f. (See *Frequency*, also *Cycle.*)

FREQUENCY, VARIATION OF POWER FACTOR WITH—When a condenser is charged and *dielectric absorption* (q.v.) takes place, it is always accompanied by a loss of power which is apparent in the form of heat produced in the condenser. This loss of power varies with the frequency of the current impressed on the condenser and it is therefore said that the *power factor* (q.v.) varies with the frequency. (See *Condenser, also Dielectric Absorption.*)

FULL LOAD CURRENT—The current delivered from any electrical source, such as a dynamo or generator operating at its maximum load. The maximum current obtainable from an electric source.

FULL WAVE RECTIFICATION — In rectifying alternating current, the process of rectifying both the positive and negative alternations or half cycles. The majority of rectifiers for charging storage batteries rectify only a half wave, or half cycle, and thus do not utilize the full alternating current cycle. (See *Rectifier, Full Wave*, also *Charger, Storage Battery.*)

FULLER CELL—A type of primary cell formerly much used in telephone, telegraph or any intermittent work. (See *Cell, Open Circuit.*) A zinc electrode cast in the shape of a cone is at the end of a soft copper wire and rests at the bottom of a porous cup immersed in a dilute solution of sulphuric acid. The carbon electrode is in an outer jar in a bichromate solution. (See *Cell.*)

FUNDAMENTAL FREQUENCY — The frequency or number of cycles per second of the fundamental wave. (See *Frequency*, also *Fundamental Wave.*)

FUNDAMENTAL UNITS—Units based on the fundamental quantities of length, mass and time. Under the *C. G. S. system* the unit of length is the centimeter; of mass, the gram; and time, the second. The absolute system of units. (See *Centimeter-Gram-Second.*)

FUNDAMENTAL WAVE—In an alternating current, the component of the wave form that is a pure sine wave of the principal frequency. Upon this principal frequency the various higher frequencies (harmonics) may be superposed. A fundamental wave when combined with other frequencies which are a whole multiple of it produce the harmonious note. These terms are derived from acoustics in which the fundamental is the pure sound wave or pure note. (See *Sine Wave of Alternating Current or EMF.*)

FUNDAMENTAL WAVELENGTH — Term used to denote the wave-length at which the inductance and capacity of a coil or antenna would be in resonance. In referring to antennas the fundamental wave-length is the wave-length corresponding to the frequency at which free oscillations will occur. This is often referred to as the natural wave-length and the corresponding frequency is referred to as the natural frequency of the antenna.

For practical purposes the fundamental wave-length of an aerial (antenna) may be computed from the dimensions. Where no loading inductance is used, a single wire flat top antenna (T or inverted L type) will have a fundamental wave-length approximately four times the distance from the ground connection to the extreme end of flat or horizontal portion as measured in meters. Thus, if the antenna is of the L type with a top one hundred feet long and a lead-in and ground connection sixty feet long, the fundamental wave-length will be about $4 \times (100+60) \times 1.1$, the result being given in feet. Now as wave-length is always stated in meters it will be necessary to divide this result by 3.28 which is the rough conversion factor from feet to meters. The multiplier 1.1 adds ten per cent to the figures for necessary correction of length. If the antenna is of the T type, the length of the flat portion would be considered as half of the total horizontal length or 50 feet instead of 100 feet. The illustration shows method of measuring length. (See *Wave-length Calculation for Antennae.*)

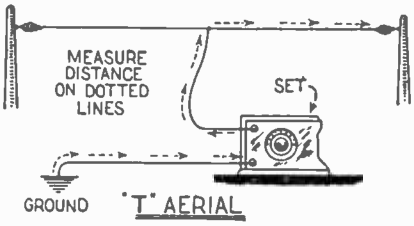

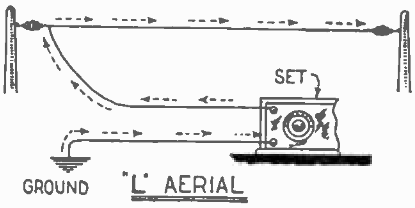

FUSE—A device used in electrical circuits to prevent the passage of excessive currents. Generally a strip of fusible metal, commonly lead with a small percentage of tin, enclosed in a fiber or insulated and fireproof container and inserted in the circuit. When the temperature, due to excessive current, reaches a certain limit, the fuse "blows" causing a break in the circuit. It is thus possible to limit the current flowing in a circuit to a definite maximum in order to protect instruments from excessive charges. The most common use is in commercial lighting and heating systems where the current on any line is generally limited to ten or twenty amperes for the protection of the lines leading in to the building.

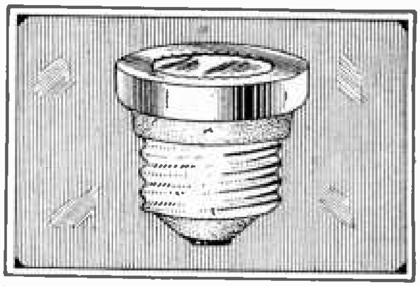

Fig. 1. A common type of fuse used in lighting, heating and power circuits.

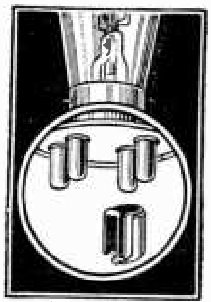

Illustration by Courtesy of Radio Equipment Co.

Fig. 2. A fuse used to protect the filament of a vacuum tube from damage by excessive voltage.

G

GALENA—A natural crystal sulphide of lead. Chemical symbol PbS. It is an ore of lead, grey in color and showing brilliant metallic lustre when fractured in cubical form. Galena crystals are used either alone or in conjunction with other minerals as *crystal rectifiers* (*detectors*) in radio reception. When used in combination it is generally in conjunction with *graphite* or *zincite*. Used alone, the contact may be a fine wire or *cat whisker.* (q.v.) (See *Crystal Detector.*)

GALVANIC BATTERY—The term, now obsolete, formerly applied to a primary battery or group of primary cells.

GALVANI, LUIGI. Born 1737, died 1798.—An Italian physician and physiologist noted as the discoverer of galvanic or current electricity. While dissecting the legs of a frog, they came by accident into contact with dissimilar metals which caused muscular action in them. Galvani thought that he had discovered electricity in animal matter. Volta attributed the action to the metallic contact and thereupon constructed his voltaic pile, the forerunner of the primary cell.

GALVANOMETER—An extremely sensitive instrument used for detecting and measuring electric currents and also for indicating the direction of the current in a circuit. In its simplest form it consists of a small iron needle pivoted in the center of a hollow coil of wire. The needle is controlled by the earth's field and points north and south when at rest. When an electric current is passed through the galvanometer winding the needle is deflected over a graduated scale the extent of the deflection depending on the force of the current. The instrument is arranged in such a manner that the needle will be deflected in one direction when the current flows through the windings in a certain direction, and when the current is reversed the needle will be deflected in the opposite direction. Thus when the connections are known the direction of the current is accurately indicated as well as its strength. The galvanometer is one of the most sensitive current measuring devices, the deflection of the needle being due to the electro-magnetic field created by the passage of electric current through the coil of wire. (See *Ammeter* or *Ampere Meter*, also *Galvanometer, D'Arsonval* and *Galvanometer, Mirror.*)

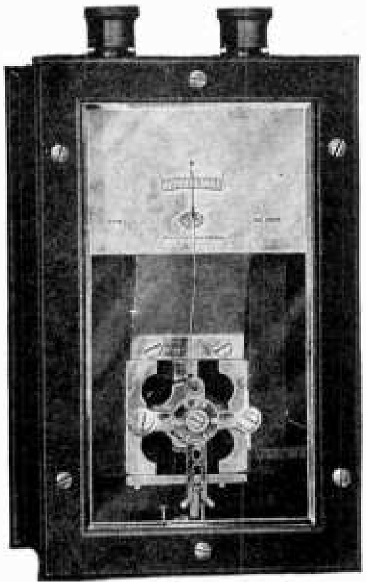

Illustration by
Courtesy of the Roller-Smith Co.

A galvanometer which is used for measuring electric currents, and also for indicating the direction of flow of current in a circuit.

GALVANOMETER CONSTANT—Generally speaking, a number used to change the galvanometer reading to ordinary units. The instrument measures extremely small currents and this number times a certain function of the galvanometer reading gives the value in customary units. Technically, the strength of the field produced at the center of the coil of a galvanometer by the unit of current to be used. (See *Constant.*)

GALVANOMETER, D'ARSONVAL—A form of galvanometer having a dead beat indicator (see *Dead Beat Instrument*) in which the indicating coil is suspended in the field of a powerful horseshoe magnet. It is one of the most sensitive forms of galvanometer. The name was derived from that of the designer, A. D'Arsonval. (See *Galvanometer.*)

GALVANOMETER, MIRROR—A type of galvanometer, the moving coil or indicator of which carries a mirror which reflects a beam of light from a fixed lamp. It is so arranged that the lamp is provided with a lens and the beam is so reflected that a visible spot of light is seen on a scale and deflected when the moving element (carrying the mirror) swings with the passage of current. In the beam of light is a sharp image of fine wire stretched in front of the lamp. This type of galvanometer may also be arranged so that the reflection of a fixed scale may be observed in the mirror by means of a small telescope arrangement. (See *Galvanometer.*)

GAMMA—Γ—The third letter in the Greek alphabet. (See *Gamma Type Antenna.*) The lower case *Gamma* γ is used as a symbol for *conductivity.* (q.v.)

GAMMA TYPE ANTENNA—An alternate name for the inverted L type of antenna, so called because the third letter in the Greek alphabet, Γ, resembles an inverted "L," and a single wire antenna with lead-in from one end is in the shape of an inverted L. (See *T Type Aerial.*)

GAP—See *Spark Gap*, also *Lightning Gap.*

GAP, AIR—Term applied to the gap between sections of an iron core as in a transformer for the purpose of preventing *saturation.* (q.v.) Also applied to speak dischargers where a space is provided between the electrodes across which the spark jumps in operation. The term is used as well in referring to the radial depth of the space between surfaces of the rotor and stator in generating machines or motors. (See *Spark Gap*, also *Rotary Gap.*)

GAP ARRESTER—A form of *lightning arrester* (q.v.) in which small air gaps are provided between non-arcing metal electrodes. (See *Arrester.*)

GAP, MICROMETER—A small gap designed to protect apparatus from excessive potentials. (See *Safety Gap.*)

GASEOUS—Anything of the nature of a gas, having no natural physical boundaries but accommodating itself to the shape of the container, expanding to the dimensions of any space in which it may be confined.

GASEOUS TUBE—A term sometimes applied to a vacuum tube in which a certain amount of gas is present. Such tubes under ordinary conditions are not efficient as amplifiers in vacuum tube circuits, the minute gas atoms presenting a more or less effective obstacle to the free passage of the electrons. (See *Vacuum Tube, Theory of Operation.*)

GAS RECTIFIER—The term applied to a vacuum tube which is only partially exhausted of its residual gas. They are also known as *soft* tubes and are generally used as detectors although the tendency now is toward *hard* or highly evacuated tubes due to their adaptability to all purposes in radio receivers whereas the soft tubes do not function well as *amplifiers.* (See *Vacuum Tube.*)

GASSING—When the charging current is continued after a storage battery has taken a normal charge the free liberation of oxygen and hydrogen takes place. This is manifest in the form of gaseous emanation or liberation of bubbles from the electrolyte. The same effect may be noted when a battery is charged at higher rate than the normal charging rate stated by the manufacturer, and under certain conditions may occur at the later stages of charging, even when the battery is not fully charged. In a storage battery that is in good condition it is a fairly accurate indication of a full charge. (See *Charging*, also *Rate of Charging.*)

GAUGE, WIRE—The various systems of measurement used for denoting the physical dimensions of wire. The standard gauge in the United States is known as the B & S or Brown and Sharpe gauge. (See *Brown & Sharpe Guage.*)

GAUSS—The term adopted by the American Institute of Electrical Engineers as applying to the *magnetic flux density.* (q.v.) It is written B or ℬ and is also used in conjunction with the *Gilbert* to denote magnetizing force, one gauss being one gilbert per Centimeter. It is incorrectly used to apply to *magneto motive force.* (See *Centimeter Gram Second or C. G. S.*)

GENERATOR—Any machine for producing electric energy. Usually a device for converting mechanical power into electrical energy. It may apply to a vacuum tube used to generate oscillations as in transmission or for *heterodyne action.* (See *Heterodyne*, also *Vacuum Tube Generator.*)

GERNSBACK, HUGO—Born 1884. Educated Ecole Industrielle, Luxembourg; Technikum, Bingen, Germany. President, The Experimenter Publishing Company, Publishers of Radio News, Science and Invention, Amazing Stories and Radio Internacional. President of The Consrad Company, Publishers of Radio Review, Radio Listeners' Guide and Call Book, and Money Making. Editor of Radio News, Science and Invention, Amazing Stories. Author of "Radio for All," "Wireless Telephone," "Ralph 124C-41+."

Hugo Gernsback.

GERNSBACK, SIDNEY—Born 1876. Educated College of Luxembourg, Lycee de Nancy, France. Vice-President and Treasurer of The Experimenter Publishing Company, Publishers of Radio News, Science and Invention, Amazing Stories, and Radio Internacional. Vice-President and Treasurer of the Consrad Company, Publishers of Radio Review, Radio Listeners' Guide and Call Book, and Money Making. Editor of Radio Re·

view, Radio Listeners' Guide and Call Book, and Money Making. Author of "A Thousand and One Formulas," "Wireless Course in Twenty Lessons," "Experimental Electricity Course," "Radio Encyclopedia."

Sidney Gernsback.

GILBERT—Symbol *F*, is the unit of magneto-motive force adopted by the American Institute of Electrical Engineers. (See *Magneto-motive Force.*) It is equivalent to $4\pi/10$ or .7958 ampere turns. (See *Ampere Turns.*)

GLASS PLATE CONDENSER—A condenser, generally of the heavy duty fixed type, consisting of alternate layers of metal or foil and thin glass sheets as the *dielectric.* Glass plate condensers are mainly used for transmitting, glass having a very high dielectric constant. Such condensers may be immersed in oil to stand higher potentials without breaking down. (See *Condenser,* also *Dielectric Coefficient and Constant.*)

GLASS SILENCER—A glass tube having wooden ends with holes just permitting discharge electrodes or rods to pass, used in conjunction with spark coils and gaps to enclose the sparking surfaces and thus dampen the noise of the sparking. (See *Spark Gap.*)

GLOW DISCHARGE—See *Corona.*

GLIDING THEORY—A theory regarding the propagation of electromagnetic (radio) waves, which assumes the waves as following the curvature of the earth and not being subject to reflection due to objects in their path or by any reflecting surface above the earth. This theory has been discarded; due to observations of phenomena of transmission it now is generally assumed that there must be a reflecting surface high in the atmosphere. However, it is considered probable that the waves do not extend very high as a general rule, the exceptions being the cases of extreme distance transmission in certain directions. In the upper atmosphere, due to the rarification of the atmosphere *ionization* (q.v.) is very pronounced and the air may become a fairly good conductor. This conducting region or layer may impose severe obstacles to the electromagnetic waves and thus limit them to a certain restricted zone. In the majority of cases this ionized region may thus be regarded as damping or absorbing the waves, and in others it may conceivably intensify or concentrate them in one direction, producing instances of transmission over extreme distances with low power. For more complete discussion along these lines see *Heaviside Layer,* also *Refraction and Reflection of Electromagnetic Waves.*

GOLD LEAF ELECTROSCOPE—A very sensitive device for determining the presence and nature of an electric charge. It comprises, essentially, two strips of gold foil joined at one end by a conductor and suspended in a glass jar. When an electric charge is present the two strips of foil have a tendency to repel one another and thus, by diverging, indicate the presence of the charge. (See *Galvanometer.*)

GOLDSCHMIDT ALTERNATOR—One of the first practical *alternators* for generating *high frequency currents* suitable for *continuous wave* radio transmission. It was designed by *Rudolph Goldschmidt* in 1912 and was used mainly in Europe. This machine operated on a frequency step-up principle. A single phase alternating current was first produced, the frequency being about ten thousand cycles. This was about the highest possible frequency obtainable with an ordinary type of alternator. By means of a *cascade* effect, obtained by an arrangement of tuned circuits connected to stator and rotor of the alternator, a much higher frequency could be impressed on the antenna to radiate continuous waves. The machine had a number of disadvantages which limited its commercial value, chiefly the fact that the air gap between rotor and stator was necessarily very small—a difficult matter with a bulky machine, the rotor of which weighed several tons. (See *Alternator, Continuous Waves,* also *Goldschmidt, Rudolf.*)

GOLDSCHMIDT, RUDOLF—Born March 19, 1876, at Neu-Buckow, Mecklenburg, Germany. After finishing his education at Wiemar Municipal School, he studied engineering at Charlottenburg and Darmstadt Technical High School. In Darmstadt he obtained his degree as electrical engineer in January, 1898, and then became assistant to Professor Kittler. In 1900 he obtained the college and travelling scholarship, which enabled him to visit engineering works in Belgium, England and France. Later in the same year he was appointed engineer in the laboratory of the Allgemeine Elektrizitaets Gesellschaft in Berlin. In 1901-2 he occupied the position of chief laboratory engineer and designer to Kolben & Co., Ltd. in Prague. He came to England in connection with the Willesden Electricity Supply Station and was later appointed chief engineer to Messrs. Crompton & Co., of Chelmsford. In 1905 he joined the Westinghouse Company in Manchester. After private preparation he passed the German abitur-examination and obtained the degree of Dr. Eng. In 1907 he returned to Germany as lecturer at Darmstadt Technical College. Here he practiced as a consulting engineer, and also pursued the development of several inventions, chiefly occupying himself with the invention and design of high-frequency alternators for wireless telegraphy. In 1911 he became manager of the "Hochfrequenz-Machinen Aktiengesellschaft fur drahtlose Telegraphie" in Berlin, a company formed for the utilization of his inventions in wireless telegraphy. In this position he established two large wireless stations at Eilvesen, Province of Hanover, and Tuckerton, New Jersey, for wireless communication between Germany and America.

GOLDSMITH, ALFRED NORTON—Elec. and Radio Engr.; b. New York, N. Y., Sept. 15, 1887; B.S., Coll. of the City of New York, 1907; Ph.D., Columbia Univ., 1911; cons. radio expert, U. S. Dept. of Justice, 1912; cons. radio engr., Atlantic Communication Co., 1914; cons. engr., Gen. Elec. Co., 1915-17; dir. of research, Marconi Wireless Telegraph Co. of America, 1917-19; associated professor of electrical engineering, College of City of New York, since 1924; dir. of research dept., Radio Corp'n of America, 1919-1922; Chief Broadcast Engineer, Radio Corp'n of America, since 1922; editor, Proceedings of the Institute of Radio Engrs., since 1912. Made investigations in simplex and duplex radio telegraphy and telephony, transmission of canal rays, precision measurements in radio eng. Author: "Radio Telephony" (Wireless Press), 1918; "Radio Measurements," "Radio Frequency Changers" (Proc. Inst. Radio Engrs.), 1915; "World Communication" (Jour. A. I. E. E.), 1921. Technical director, U. S. Signal Corps School of Communication, 1917-18; U. S. Naval Radio School, 1917-18. Fellow, A. I. E. E., I. R. E., American Association for Advancement of Science. Hon. mem., Radio Club of America, Radio Society of Great Britain, Am. Physical Soc.

Dr. Alfred N. Goldsmith.

GONIOMETER—A device used in radio reception to determine the direction from which the electromagnetic waves (signals) are coming. The term is also used at times to designate any system of directional transmission or reception. The device was originally designed as an aid to navigation.

The original goniometer was designed by Messrs. Bellini and Tosi (see *Bellini, Dr. Ettore,* also *Tosi, A.*), but numerous changes have since been made, notably by Marconi and his associates, to adapt it to the requirements of ships at sea. The form most used is based on the original Bellini-Tosi machine and is essentially a special form of *oscillation transformer* used in conjunction with a special aerial or loop. As an example of the use of a goniometer in connection with ships at sea, suppose a ship is in distress at a point about two hundred miles off the eastern coast. Perhaps the ship's radio apparatus is working, but for some reason the mariner does not know the ship's position. The radio signals are then received, let us say, by three different land stations at widely separated points. The receiving stations use their goniometers and determine more or less accurately the direction from which the signals are coming. The three stations compare notes on the respective directions and any one of them can readily plot out lines running on a chart from each station in the directions indicated by the goniometers. Then by simple triangulation the location of the ship can be figured, enabling other ships in the vicinity to be instructed accordingly and go to their rescue. There are, of course, many other uses that occur in everyday practice. Goniometers are also much used by radio inspectors to locate stations that are interfering, using an unlicensed transmitter, or are otherwise not complying with the regulations.

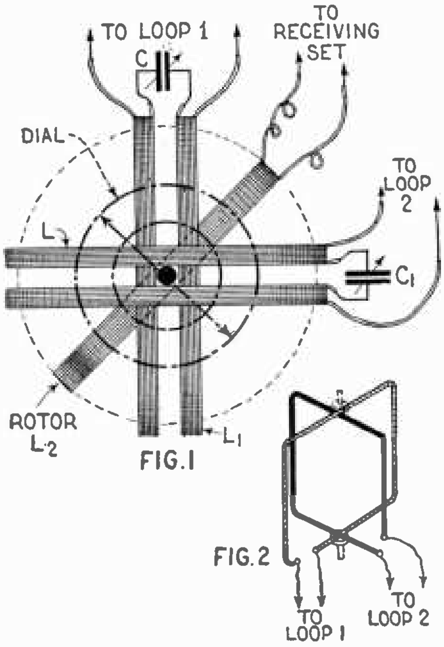

Fig. 1 shows the general arrangement of a goniometer and Fig. 2 shows the loops to which coils L and L-1 are connected.

The illustration Fig. 1 shows the general arrangement of the instrument. Here, coils L and L1 are the main windings of the goniometer. They are wound in two sections with condensers C and C1 inserted between the sections of the two windings for purposes of tuning. Coil L2 is the *exploring coil.* This is connected to the receiving apparatus, which is the customary detector circuit. Coils L and L1 are connected to the two loop aerials shown in Fig. 2. Here it will appear that the loops are placed at right angles, their relation in this respect being variable. In operation the coils L and L1 have minimum *mutual induction*, as they are placed at right angles. The exploring coil L2 can be rotated through a complete circle and the pointer indicates the angle with relation to the fixed coils. When one of the two loops points in the direction of a transmitting station it will pick up the signals, the other being at that moment inoperative, as it is at right angles to the first, as previously explained. If the exploring coil L2 is now turned until maximum signal strength is obtained, the pointer will be in the direction of the plane of the loop. If the advancing wave strikes the two loops simultaneously at any angle, a resulting magnetic field will be set up in the coils L and L1 of the goniometer, the field having a direction at right angles to the direction from which the waves are advancing. Now when the exploring coil L2 is turned to allow maximum signal strength due to this field (permitting maximum inductive transfer), the pointer will be in the direction of the transmitting station. The direction indicated by the pointer is, of course, not the geographical direction, but this can readily be determined by means of a compass. It also appears from the foregoing description that the pointer will indicate two directions, which, however, will be on the same geographic plane. In other words, the pointer may be, for example, exactly north and south.

GRAMME ARMATURE—A form of ring type *armature* (q.v.) used in generators, named for its originator. (See *Generator.*)

GRANULAR CARBON—The carbon grains used in a telephone transmitter to vary the resistance between the electrodes. The highest grade is made by carbonizing pure anthracite coal in an electric furnace, after which it is screened and graded. (See *Microphone, Carbon.*)

GRANULAR COHERER—A form of *coherer* (q.v.) used in the early days of wireless. It employed carbon granules between two electrodes, the whole enclosed in a glass tube. (See *Detector.*)

GRAPH—The presentation in diagrammatic form of facts, formulae, etc., particularly in the electrical and mechanical fields. Graphs are much used in radio to give graphic illustrations of various functions, such as the relation of currents and voltages. For this purpose ruled and squared paper is used. One of the most common and useful forms of graph is known as a characteristic curve of a vacuum tube. These curves may show the relation of changing values, such, for instance, as the relation of grid voltage to plate current, indicating certain of the operating characteristics of the vacuum tube. (See *Characteristic Curve,* also *Sine Curve.*)

GRAPHITE RESISTANCE—A form of resistance composed of graphite, used in any of several ways in radio circuits. The most common usage is in the form of a *grid leak* (q.v.). Here it may take the form of a thin sheet of paper coated with graphite or a glass tube filled with graphite and some binding medium. Such resistances are also used in *resistance coupled amplifiers* (q.v.), in which case they are interposed between the plate and "B" battery and couple the successive stages of amplification. (See *Resistance,* also *Resistance Coupled Amplifier.*)

GRASSOT FLUXMETER—An instrument used to measure the strength of magnetic fields and the regions around magnetic poles. It is of particular value in measuring the relative strengths of various magnets. The instrument comprises a very sensitive galvanometer of the moving coil type with a special exploring coil. This coil is introduced into the magnetic field to be measured and readings of the proportionate strength of the field are obtained directly from the scale of the galvanometer, which is appropriately divided. The field strength is obtained by dividing the deflection value in *Maxwell turns* (q.v.) by the area of the search or exploring coil in square centimeters and the number of turns of the coil. (See *Galvanometer,* also *Flux Density.*)

GRAVITY BATTERY—A combination of gravity cells. A group of cells, each cell having two different electrolytes which remain separated without use of a diaphragm due to their difference in specific gravity. (See *Gravity Cell.*)

GRAVITY CELL—A type of *primary cell* (q.v.) in which the specific gravity of one electrolyte is less than the other, which causes this lighter electrolyte to float on top of the heavier fluid. One of the typical examples of gravity cells is the *Daniell's Cell* (q.v.). (See *Secondary Cell.*)

GRID—Generally, any system of wires placed parallel to each other and in the same plane, as a mesh. Specifically in radio usage, the fine network of wires interposed between the filament and plate elements in a vacuum tube. It may be in the form of a flat network of wires or in the form of an oval cylindrical mesh completely surrounding the filament and in turn surrounded by a similarly shaped plate of metal. The illustration shows the form taken by a typical grid member in a modern type of vacuum tube. Here the plate is not shown in order to permit the grid member to show more clearly.

In receiving circuits the grid acts as a valve to control the flow of electrons from hot filament to cold plate. When signals are impressed on the aerial they are tuned by the customary agencies and passed to the *grid circuit* (q.v.). As the incoming signals are in the form of *high frequency oscillations* (q.v.) changing their polarity (direction) many thousands of times each second, it will be apparent that the grid is altered rapidly from positive to negative and back. As the electrons flowing from the filament are negative, the grid will thus have a tendency to retard or increase the flow. That is to say, at the instant the grid is positive it will attract more electrons, but when it is momentarily negative the like charges will repel each other and less electrons will flow, some of them being held back to the filament. (See *Grid Battery, Grid Bias,* also *Vacuum Tube, Theory of Operation of.*)

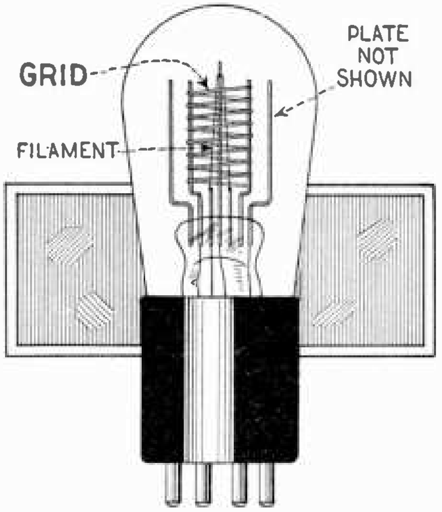

Showing the form taken by a typical grid in a modern vacuum tube.

GRID BATTERY—A battery of small cells connected in the grid circuit of a vacuum tube to place a negative or positive bias or charge on that element. The illustration Fig. 1 shows a small four and one-half volt battery inserted in such a manner that a negative potential is impressed on the grid. In this case the *characteristic curve* (q.v.) of the tube is assumed to indicate that more difference in current occurs between points B and C than between A and B in Fig. 2. For this reason a negative potential is applied as shown in Fig. I. (See *Grid Bias.*)

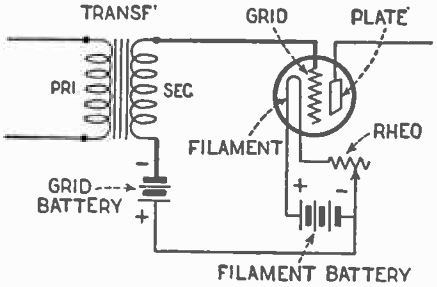

Fig. 1. Method of a 4½ volt battery to impress a negative potential on the grid of a vacuum tube.

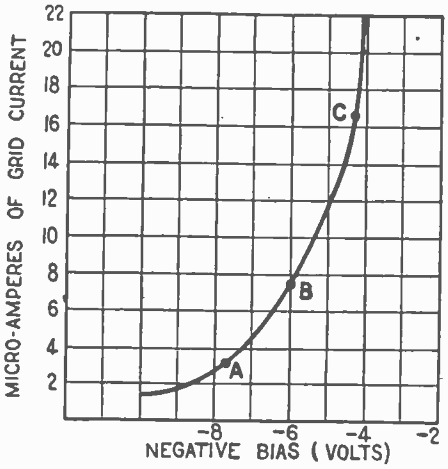

Fig. 2. Characteristic curve of a vacuum tube indicating more difference in current occurs between "B" and "C," than between "A" and "B."

GRID BIAS—A potential of a few volts, generally from four to six, applied to the grid of a vacuum tube to influence its operation by making it more or less negative. The grid bias is usually negative and determines the point of the *characteristic curve* at which the tube will operate. In a sensitive receiver, and particularly where a tube is used as an *amplifier* (q.v.), it is essential to obtain as great a change of grid current as possible. (Note: The greater the change of grid current, the greater the change in plate current and hence the more powerful will be the output.) By applying a negative potential on the grid it is possible to hold it at the point of maximum response. (See *Grid Battery,* also *Vacuum Tube, Theory of Operation of.*)

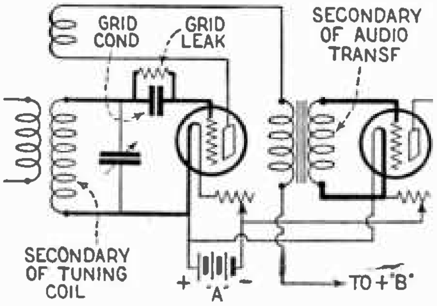

Diagram in which heavy lines at left show the grid circuit of the detector, while those on right indicate audio amplifier grid circuit.

GRID CIRCUIT—In a receiver employing vacuum tubes, the part of the circuit containing the grid of the tube or tubes, but generally referring to the tuning circuit or that part of the system which contains the tuning elements. In the illustration is shown a standard arrangement of a regenerative detector and one stage of audio-frequency amplification. The heavy lines at the left of the diagram represent the tuning circuit or detector *grid circuit,* while the section with heavy lines at the right is the grid circuit of the amplifier tube. The grid circuit of a detector tube may generally be distinguished by means of the grid leak and condenser (so marked in the illustration), which controls the incoming signals and permits the tube to function as a rectifier. (See *Detector, Grid Control,* also *Grid Condenser* and *Grid Leak.*)

GRID CONDENSER—A small fixed condenser, generally from .00025 to .0005 microfarad capacity, inserted in the circuit of a detector tube between the tuning coil and the grid member of the tube. The illustration shows the manner of connecting the grid condenser in a conventional detector circuit. This condenser insulates the grid from the filament by breaking the path F to G and permits the tube to act as a rectifier or detector. A resistance (called *grid leak*) is usually placed across this condenser to allow the accumulated charges on the grid to leak off. (See *Grid Leak,* also *Grid Control* and *Vacuum Tube, Theory of Operation of.*)

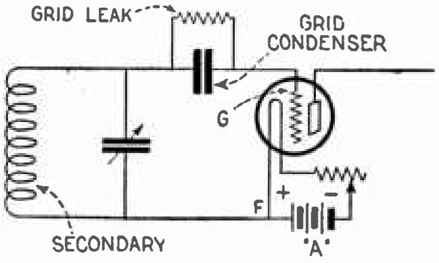

Showing how grid condenser is connected in a detector circuit.

GRID CONTROL—The general term used to designate the various devices and associated connections placed in or allied with the grid circuit of a vacuum tube for the purpose of controlling to a greater or lesser extent the potential of the grid, or to tune or control the incoming oscillations. In the case of a vacuum tube used as a rectifier (detector) the grid condenser and grid leak are placed in series with it. (Note: the grid leak is often connected from the grid directly to the filament instead of across the grid condenser, the operation, however, being approximately the same in either case.) Here the controls permit the tube to function as a detector of the incoming oscillations. Where the tube is used as a radio-frequency (high-frequency) amplifier, the radio-frequency transformer or other amplifying unit controls the grid of the following tube by influencing the fluctuating potential differences impressed on it. (See *Amplifier, Radio-Frequency.*) The most important example of a grid control is by using some means to control the potential applied to the grid, such, for example, as a grid bias or grid battery. (See *Grid Bias, Grid Battery,* also *Vacuum Tube, Theory of Operation of.*)

GRID CURRENT—The current present in the grid circuit of a vacuum tube. This is generally very small, perhaps of the order of fifteen to twenty *microamperes* (q.v.). (See *Plate Current,* also *Grid Bias.*)

GRID LEAK—A high resistance placed in the grid circuit of a vacuum tube to permit the electrons forming the grid current to leak off after each charge, thus preventing their accumulating on the grid in such numbers as to stop the flow from the filament. The electrons being negative, a surplus of them on the grid member of the tube would act as a barrier to the free flow of other electrons from the filament through the grid to the plate, thus diminishing the plate current toward zero. It will be understood that a heavy negative charge on the grid will repel a like or negative charge, in this case the negative electrons being thrown off by the hot filament.

The grid leak may have a value ranging from about 250,000 ohms to several million ohms. The value of a grid leak is generally stated in *megohms*—one megohm being one million ohms. Thus a grid leak of 250,000 ohms resistance would be referred to as a ¼-megohm leak and, similarly, one having a resistance of two million ohms would be referred to as a two-megohm grid leak.

Such resistances are furnished in many forms, the most common being a strip of paper impregnated with graphite or some similar high resistance preparation, placed in a glass tube and sealed to prevent moisture from affecting the value. The illustration, Fig. 1, shows a tubular type of grid leak.

Courtesy of Chas. Freshman Co.

Fig. 1.—A tubular type of grid leak.

For experimental work a fairly satisfactory grid leak can be made by the simple expedient of drawing a line on a piece of cardboard with a soft pencil—the graphite mark acting as the resistance. At each end of the cardboard is placed a binding post in contact with the graphite, each binding post being connected to opposite sides of the grid condensers. Variable grid leaks, the resistance of which can be changed at will, are also furnished in an infinite variety of types, two of which are shown in Fig. 2. (See *Variable Grid Leaks,* also *Grid Control* and *Resistance.*)

CHAS. FRESHMAN CO.
NEW YORK PAT'S PEND.

Courtesy of Chas. Freshman Co.

Fig. 2.—Two types of variable grid leaks with condenser incorporated.

GRID POTENTIAL, MODULATION OF—In the transmission of *undamped waves* (q.v.) generated by a vacuum tube, the process of varying or modulating the potential of the grid element of the tube with respect to the filament. In this method an alternating potential may be applied to the grid circuit, the frequency of this varying potential being within the audio band —possibly about 500 cycles—supplied by a generator of that freqnency. (See *Modulation.*)

GRID POTENTIOMETER—A potentiometer—a variable resistance unit—Fig. 1, used in the grid circuit of a vacuum tube for the purpose of controlling the potential applied to the grid. The conventional method is to place the potentiometer across the "A" battery with the center or variable contact arm connected to the grid of the tube in place of the usual connection to either positive or negative filament lead. This is shown in the illustration, Fig. 2, Here the customary circuit is shown for the "A" battery to light the filament of the tube. Pri. and Sec. are respectively primary and secondary of the transformer (either radio-frequency or audio-frequency) and C the center arm of the potentiometer. The center arm may be moved either to the right or left, permitting either negative or positive potential to be applied to the grid. This system is much used in amplifying tube circuits to control the action of the tube as an amplifier. Another method would be to connect one of the ends of the potentiometer to the negative lead of the "A" battery and the center arm to the grid leak. In this way the potentiometer would be acting as a pure variable resistance, the potential applied to the grid being always negative, but susceptible of variation within the limits of the ressitance and the voltage of the "A" battery. (See *Grid Bias, Grid Battery,* also *Grid Control.*)

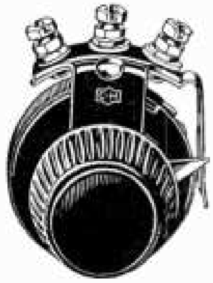

Courtesy of the Cutler-Hammer Co.

Fig. 1.—A type of potentiometer.

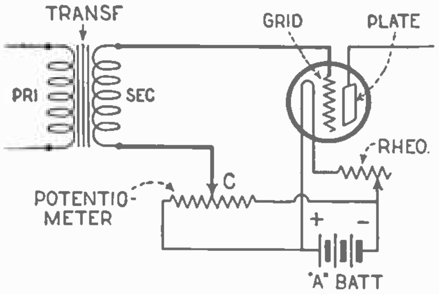

Fig. 2.—Method of connecting potentiometer in grid circuit.

GRINDER—Term occasionally applied to one of the various forms of atmospheric disturbance known under the general heading of *static.* This form of static is more prevalent in warm weather and does not appear to have its origin in local electric storms. Its effect is to produce rattling sounds of no definite pitch. (See *Strays,* also *Static.*)

GROUND—The earth considered as an electric conductor. In radio it is considered as the return circuit for electromagnetic waves. (See *Ground Connection,* also *Theory of Propagation of Electromagnetic Waves.*)

GROUND CLAMP—A strip of metal, generally copper, having some sort of arrangement for fastening rigidly to a water or steam pipe to form the ground connection for a radio receiver. The illustration shows a conventional form of clamp. (See *Ground Connection.*)

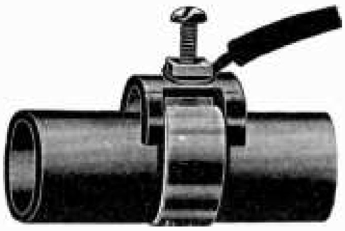

A ground clamp for fastening the ground lead to a water or steam pipe.

GROUND CONNECTION—In radio reception or transmission, the connection between the earth and the apparatus whereby the currents set up by the electromagnetic waves are returned to the earth as one side of the assumed circuit. The ground or earth connection from a radio receiver or transmitter should be as short as possible and preferably a heavy wire as shown in the illustration. Such a connection for receiving purposes may be one or more heavy metal plates sunk in the ground and connected by a heavy wire to the apparatus. The more common method is to connect the set to a convenient water or steam pipe, such pipes being almost invariably well grounded, especially the former. (See *Earth,* also *Ground* and *Transmission.*)

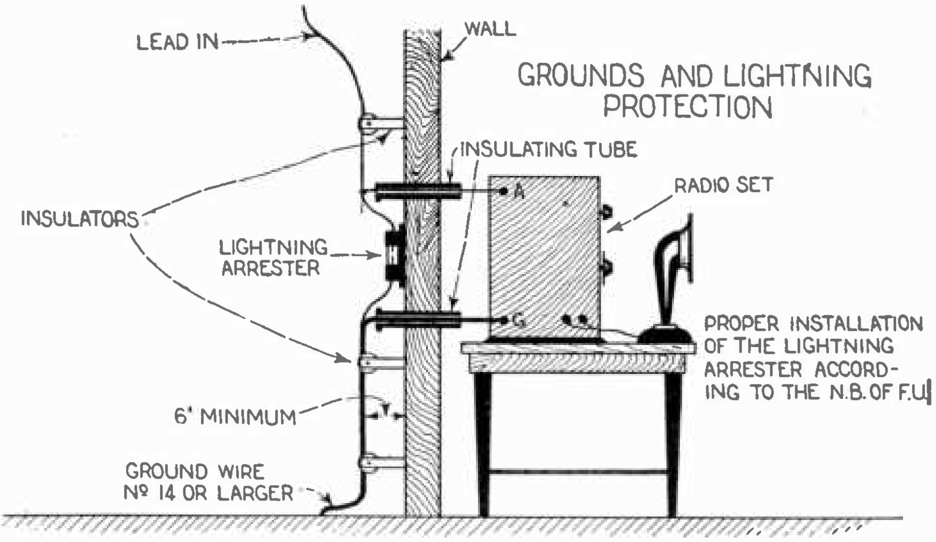
LEAD IN
WALL
GROUNDS AND LIGHTNING PROTECTION
INSULATING TUBE
RADIO SET
A
INSULATORS
LIGHTNING ARRESTER
G
PROPER INSTALLATION OF THE LIGHTNING ARRESTER ACCORDING TO THE N.B.OF F.U
6' MINIMUM
GROUND WIRE Nº 14 OR LARGER

Illustration showing by means of a heavy line, the ground connection of a radio receiver.

GROUND CURRENTS—Electric currents present in the ground due to the grounding of commercial electric machines. These currents very often prove a disturbing factor in radio reception. (See *Ground.*)

GROUND DETECTOR—A device for indicating an accidental ground on a power transmission line or for any conducting systems for electric currents. (See *Detector,* also *Ground.*)

GROUND RESISTANCE—See *Resistance of Ground Connection.*

GROUND SWITCH—A switch, generally of rugged construction, connected to the aerial, ground and apparatus in such manner that at one position the aerial is connected to the set, which in turn is connected to the ground, and in the other position the aerial being connected directly to the ground to protect the set from excessive static when not in use. (See *Lightning Switch,* also *Aerial Switch, Down Lead.*)

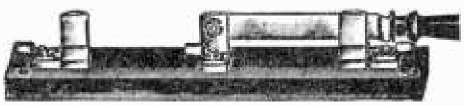

Switch having a break distance of not less than 4 inches.

GROUND WIRE—The wire connecting a receiving or transmitting set to the earth. The underwriters specify not less than No. 14 B & S gauge copper wire. (See *Ground Connection.*)

GROUNDED ROTOR—See *Rotor, Grounded Connection.*

GROUP FREQUENCY—The number of separate trains or groups of waves per second in a damped or undamped system of radio transmission. This frequency is to be distinguished from the frequency of the individual waves. Thus in a spark transmitter, for example, the condenser in the transmitting circuit will be momentarily charged and discharged. The discharges will occur in the form of trains of oscillations, each oscillation having a certain definite frequency, but the trains or groups occurring at definite intervals, the frequency of these intervals being called the group frequency. (See *Frequency,* also *Spark Transmission* and *Undamped Waves.*)

GUARD LAMP—An incandescent lamp having a straight filament, sometimes shunted across the armature and field leads of a rotary converter used in radio transmission. Its purpose is to protect the windings from injury due to the inducing of oscillatory surges from the high-frequency circuits. (See *Rotary Converter.*)

GUTTA PERCHA—An insulating material derived from the milky secretion of certain trees, used as an insulation for submarine cables on account of its ability to resist the action of salt water. It is used to a lesser extent as a general insulating medium for wires, but owing to its scarcity and limited field of usefulness, rubber or rubber compounds are more widely used. (See *Insulating Materials.*)

GUY WIRE—Wires used to brace aerial masts for receiving or transmitting stations. Such guys are generally heavy galvanized wires attached to the mast at various points and extending radially downward and anchored to surrounding rigid objects or by means of some form of anchor in the ground. (See *Aerial.*)

H

H—The chemical symbol for *hydrogen.* In electrical and radio usage it is the symbol for *magnetizing force* or *magnetic field intensity*, adopted as such by the International Electrotechnical Commission and by the American Institute of Electrical Engineers. It has also been adopted by the I. E. C. as the symbol for *Henry* (q.v.). (See *Magnetic Field.*)

HALF WAVE RECTIFICATION—The process of changing alternating currents to pulsating direct currents wherein only one-half of the full alternating current wave or cycle is rectified, the other portion being blocked by the action of the rectifier. In the illustration is shown a typical half wave rectification process. The upper graph shows the sine wave representing an alternating current, such as supplied from an ordinary alternating current house lighting system. Here A-B represents one complete cycle or two alternations, one positive and the other negative. In the lower graph is shown the form taken by the current after passing through a half wave rectifier. It now appears as a pulsating or intermittent direct current, only the upper or positive alternations having passed through the rectifier. The shaded portions below the line represent the negative or unused alternations. (See *Full Wave Rectification, Sine Wave of Alternating Current or E M F*, also *Charger, Storage Battery.*)

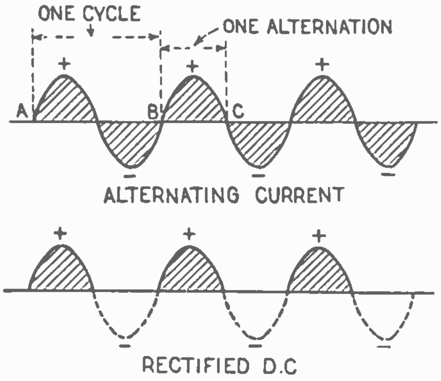

The upper graph represents an alternating current, and the lower graph shows the form taken by the current after passing through a half wave rectifier.

HALF WAVE RECTIFIER—Any system for rectifying or changing alternating currents to pulsating direct currents, wherein only half of the full alternating current wave is used. Essentially such a device is a uni-directional conductor, permitting current to pass in one direction only and submerging or blotting out the surges in the opposite direction. A typical example of a half wave rectifier is the crystal detector used in simple receiving circuits. Here the nature of the crystal is such that the currents are permitted to pass only in one direction. In half wave rectifiers used for charging storage batteries the effect is much the same. Where tubes such as the Tungar or Rectigon are employed, one tube is more commonly used, allowing the current to flow in one direction but submerging half of each cycle. (See *Rectifier, Full Wave*, also *Rectifier and Charger, Storage Battery.*)

HAMMER BREAK—A magnetically operated device for making and breaking contact in an electrical circuit. Hammer breaks are essentially the same as a vibrator or buzzer break but are made to withstand heavy currents without undue arcing, and hence are much slower in action. (See *Buzzer*, also *Spark Coil.*)

HAMMOND, JOHN HAYS, JR.—Inventor. Born San Francisco, Calif., April 13, 1888. B.S., Sheffield Scientific School (Yale), 1910; Sc.D., George Washington University. Inventor of type of torpedo for coast defense, controlled by wireless energy from coast fortifications, which was recommended to Congress for exclusive purchase by U. S. by Board of Ordnance and Fortifications, U. S. A.; invented system of automobile torpedo firing; aluminothermic incendiary projectiles employed by Allied armies in World War; radio system of control of ships, employed on U. S. S. Iowa for target practice; a system of coastal patrol by aeroplane; a system of selective radio telegraphy, adopted by the U. S. Navy, U. S. Signal Corps, and U. S Army; system of aerial coast surveying adopted by the Bartlett Expedition for Polar exploration. President Radio Engineering Company of New York; director Hammond Radio Research Labs.; director Radio Corp. of America; consulting engineer to Radio Corp. of America and General Electric Co. Has licensed the American Tel. & Tel. Co., and the Radio Corp. of America under his electrical patents in U. S. and S. A. for commercial purposes only, U. S. Gov't retaining option on military uses. Has applied for over 224 patents in U. S. and Europe relating to Radio Telegraphy and Telephony and Wirelessly controlled torpedoes and various improvements in pipe organ mechanisms. Formerly member Advisory Board, U. S. Naval Board of Inventors, member Advisory Com. Langley Aerodynamic Lab. of Smithsonian Inst., co-operating with Third Naval Dist.; member Conf. Com. on National Preparedness (Sub.-Com. on Finance); Gov. Aero Club America (mem. Aerodynamic, Tech., Pub. Safety and Map and Landing Places Coms.); v.-p. Am. Soc. Aeronautic Engineers; mem. Inst. Radio Engineers (ex-treas., etc.), Am. Inst. E. E., Royal Soc. Arts (London); asso. mem. Am. Soc. M. E.; hon. mem. Nat. Inst. of Inventors, Harvard Aeronaut. Soc.; Fellow Am. Geog. Soc.; U. S. del. Radio-Telegraphic Com., London, 1912.

John Hays Hammond, Jr.

HAND CAPACITY—See *Body Capacity.*

HARD DRAWN WIRE—In radio usage generally confined to hard drawn copper wire, i. e., copper wire that has been hardened by repeated drawing without being annealed. It is used chiefly for aerials. (See *Aerial.*)

HARD TUBE—A vacuum tube is said to be hard when it has been evacuated of all gases to a high degree. The opposite is a soft tube, one in which the vacuum is not of a very high order and more or less gas particles are present. Hard tubes are used for amplifiers, both audio and radio frequency, while the soft tubes are used mainly as detectors or oscillators. Hard tubes may be used with high plate voltages, whereas soft tubes require comparatively low plate voltages, the tubes having a tendency to turn blue (see *Blue Glow*) under the influence of high plate potentials. (See *Detector*, also *Amplifier.*)

HARMONIC CURRENT—An alternating current which can be represented by a simple harmonic or sine curve. It may be considered as an alternating current, the waves of which are sinusoidal or uniform. (See *Simple Harmonic*, also *Sine Curve.*)

HARMONIC MOTION—A term signifying any periodic oscillatory motion, as, for example, the vibration or beating of a pendulum. The illustration shows the manner of indicating simple harmonic motion. The distance A-B represents 360 degrees of a circle, or, if it is assumed to represent a sine wave of alternating current, it will represent one complete cycle. Curves representing harmonic motion are used extensively in radio engineering to illustrate, in graphic form, the oscillations of an alternating current. (See *Sine Wave*, also *Simple Harmonic Vibration* and *Oscillations.*)

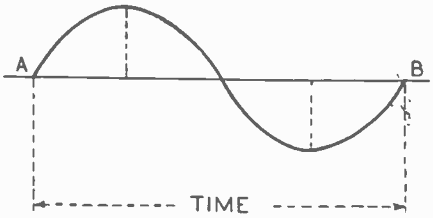

Illustration showing method of indicating simple harmonic motion.

HARMONICS—Oscillations to which a circuit will respond in addition to the basic or fundamental oscillations. The effect is similar to that in sound, where, for instance, the primary or fundamental tone struck by an instrument may be accompanied by notes or tones of a higher pitch. In sound these harmonics are principally the third, fifth, seventh and the octave. In radio, a circuit may respond to oscillations of a frequency either higher or lower than the fundamental frequency, as, for example, one third, one fifth or one seventh, or, again, double or treble the original frequency. As a concrete example, with a fundamental wave-length of 240 meters, the harmonics might be

40 meters or 80 meters. The corresponding frequencies would be approximately 1,250,000 cycles, 7,500,000 cycles and 3,750,000 cycles. (See *Harmonic Suppressor*, *Side Bands*, also *Harmonic Motion*, *Oscillations*, *Wave-length*.)

HARMONIC SUPPRESSOR—A device used in radio transmission and in laboratory work, for the purpose of eliminating undesired frequencies (wavelengths). Essentially such a device is a *filter* (q.v.) so arranged as to suppress or eliminate certain frequencies or *harmonics* (q.v.). In radio broadcasting it is particularly essential that all undesired frequencies or *harmonics* be suppressed. Basically the system involves a special coil shunted by a condenser and tuned to the frequency of the interfering wave, the main tuning circuit at the same time being adjusted to the desired frequency. (See *Broadcasting*, *General Treatise on Methods*, also *Transmission* and *Radio Telephony*.)

HAZELTINE NEUTRODYNE RECEIVER—The radio receiver designed by Professor Hazeltine, employing the Neutrodyne circuit. Essentially these are five-tube receivers employing two stages of tuned and neutralized radio-frequency amplification with the conventional detector and two audio-frequency stages. In addition to the five-tube Neutrodyne receiver there are numerous types involving the Neutrodyne principle in conjunction with *reflex systems*. (See *Neutrodyne*.)

HEAD TELEPHONE OR HEADPHONES—Telephone receivers arranged with a band to fit over the operator's head, thus leaving him free use of both hands. Head-phones are used in commercial radio work, the operator being free to handle a key or the controls of the receiver or transmitter. They are also used in broadcast reception where the strength of signals is not sufficient to permit the use of a loud speaker, and also in some cases to tune in distant signals before attaching a loud speaker. (See *Watch Case Receivers*.)

Illustration by courtesy of C. Brandes, Inc.
A type of headphone used extensively in radio reception.

HEAVISIDE, DR. OLIVER. — Born in London on May 13, 1850, Dr. Heaviside was engaged in telegraph work for a few years, but after 1874 he lived in retirement studying Clerk-Maxwell's theory and applying it to telegraph and wireless problems. Elected a Fellow of the Royal Society in 1891, he was also Faraday Medallist of the Institution of Electrical Engineers and Hon. Ph.D. of Gottingen University. The name of Dr. Heaviside is primarily associated with the theory, propounded by himself, of the existence of a permanently ionized layer in the upper atmosphere capable of deflecting electro-magnetic waves and thus permitting wireless communication round the earth.

HEAVISIDE LAYER—An assumed layer or stratum of the atmosphere supposed to exist from fifty to one hundred miles above the earth's surface according to the Heaviside Theory. Heaviside's theory supposes this stratum to consist of heavily ionized gas which is conductive and acts as a gigantic reflector of electromagnetic waves (radio waves). It is this layer which is supposed to reflect the electric waves and force them to follow the curvature of the earth. This theory partially explains—if true—why radio signals are transmitted over great distances instead of being radiated off into space and lost. It has been accorded the support of many eminent scientists. Elihu Thompson has propounded an alternative theory which assumes the waves as attached to the earth and gliding along over the surface. (See *Gliding Theory*.)

HEAVISIDE LAYER THEORY—The theory of an upper ionized stratum from fifty to one hundred miles above the earth's surface, which is assumed to act as a reflector of radio waves and thus force them to follow the curvature of the earth. None of the theories thus far advanced has been proven, but it has been definitely established that these waves do follow the curvature of the earth. (See *Heaviside Layer*, also *Gliding Theory*.)

HEDGEHOG TRANSFORMER—A type of audio transformer so called because of its peculiar construction. The core is cylindrical in shape and consists of a great number of soft iron wires around which is a bobbin holding the primary and secondary windings. After the windings are completed the wires forming the core are bent back over the form and held in place, completely surrounding the windings. (See *Audio-Frequency*, also *Transformer*.)

HEIGHT OF AERIAL—The elevation of an aerial system as used in radio transmission or reception. The effective height is roughly considered as the height above the earth or surrounding objects. The range of a station having constant power is dependent upon the height of the aerial, other factors of radiation being unchanged. In other words, increase of height increases the range within certain limits. (See *Radiation from Antenna*, also *Transmission*.)

HEISING, RAYMOND A.—Electrical Engineer, born Albert Lea, Minn., August 10, 1888. Educated University of N. D., Elec. Engineer, 1912; University of Wisc., M.S., 1914 Grad. Work in Physics, Research and Development of Radio Tel. since 1914; Trans-Atlantic test in 1915; for Army and Navy use during War, aeroplane, chaser, intership communication; since War continued research and development ship to shore use, trans-Atlantic test, 1923. Author of several technical papers; numerous inventions widely used in present broadcasting stations.

HEISING MODULATION—A system of modulating radio currents for radio transmission, such as broadcasting, developed by *R. A. Heising*. The fundamental principle is shown by the illustration. Here the customary antenna and ground are connected with a high-frequency alternator, A; the field of the alternator, FW, is connected to the grid return and filament of the vacuum tube circuit, and through the "B" battery BB to the plate P of the tube. The microphone circuit comprises a microphone M connected to the primary of a transformer T through the battery MB. The secondary of the transformer is connected to the grid G of the vacuum tube and the filament F. R is the usual rheostat to control the filament of the tube and AB is the filament lighting battery. In operation the speech or sound vibrations are impressed on the microphone M, thus varying the resistance of the current flowing from the battery MB and through the transformer T. This serves to vary the potential of the grid G, and as this potential changes, the current in the plate circuit will also be varied, the same characteristics being maintained. As the field winding FW is in the plate circuit, it is evident that the antenna current will be modulated in accordance with the variations of the speech or sounds introduced at M. (See *Grid Control*, also *Modulation* and *Broadcasting*, *General Treatise on Methods*.)

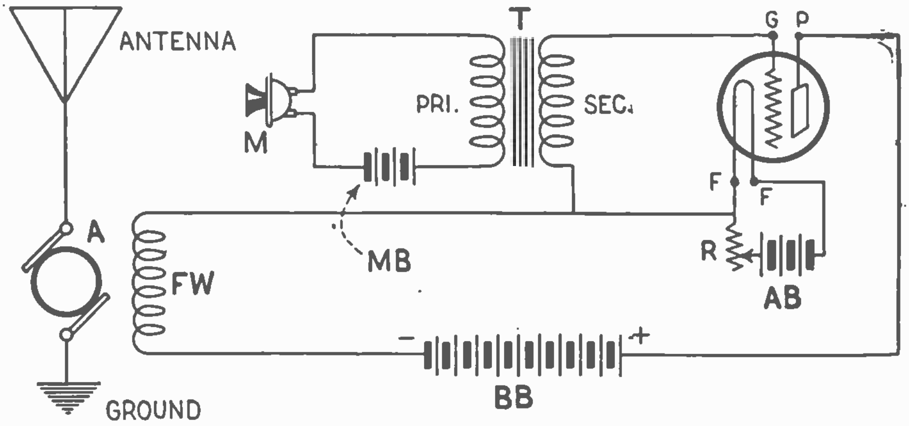

Circuit illustrating the fundamental principle of the Heising system of modulation.

HELIOGRAPH—An instrument involving a movable mirror, by means of which the sun's rays are reflected and made use of as a signalling device. In practice it is usually arranged with a blind or shade to cover the mirror at rapid intervals by pressing on a key device. In this manner the flashes may be made at definite intervals, as in the case of the Morse light. (See *Code*, also *Morse Light*.)

HELIX—The shape assumed by a line or wire wound in a single layer around a cylinder. Generally speaking, a *solenoid* (q.v.). A device in the form of a helix was formally used in transmitting circuits to vary the inductance, or, practically speaking, the wavelength. It usually was a number of turns of heavy copper or aluminum wire wound on an insulating form. Clips were provided to permit a variation in the contact point and hence any necessary change in the value of the inductance. (See *Oscillation Transformer.*)

HENRY, JOSEPH—Born in 1797, died in 1878. An American Physicist, noted for his researches in electromagnetism. He developed the electromagnet which had been invented by Sturgeon in England, so that it became an instrument of far greater power than before. In 1831 he employed a mile of fine copper wire with an electromagnet, causing the current to attract the armature and strike a bell, thereby establishing the principle employed in modern telegraph practice. He was made a professor at Princeton in 1832 and during his experimenting there he devised an arrangement of batteries and electromagnets embodying the principle of the telegraph relay which made possible long distance transmission. He was the first to observe magnetic self-induction, and performed important investigations in oscillating electric discharges (1842), and other electrical phenomena. In 1846 he was chosen secretary of the Smithsonian Inst. at Washington, an office which he held until his death. As chairman of the U. S. Lighthouse Board he made important tests in marine signals and lights.

Joseph Henry.

In meteorology, terrestrial magnetism and acoustics he carried on important researches. Henry enjoyed an international reputation and is acknowledged to be one of America's greatest scientists.

HENRY—The unit of inductance, named after the American physicist, Joseph Henry, who first produced high-frequency oscillations in 1832. A coil of wire is said to have an inductance value of one henry when a current passing through it changes at the rate of one ampere per second and an electromotive force of one volt is produced. As the henry is too large a unit for practical purposes, the thousandth part of a henry, known as the *milli-henry*, or the millionth part, known as the *micro-henry*, are more commonly used. Another much-used term is the centimeter, which is the thousandth part of a micro-henry. The abbreviation for henry is H, and for micro-henry μH. (See *Centimeter Gram Second, Centimeter*, also *Inductance.*)

HERTZ, HEINRICH RUDOLF—Born 1857, died 1894. A German physicist noted as the discoverer of electromagnetic waves (1888) which form the basis of wireless telegraphy. He found that waves produced by the spark of a simple device called the oscillator could be detected by a loop or square of wire known as the resonator, and he was able to show the reflection, refraction, diffraction and poralization of the waves. These remarkable discoveries demonstrated the practical possibilities of radio-telegraphy. He also made valuable experiments with electric phenomena in vacuum tubes.

Heinrich Rudolf Hertz.

HERTZIAN WAVES—Term occasionally applied to radio or electromagnetic waves, they being named after Dr. Heinrich Rudolf Hertz owing to his extensive researches into electromagnetic waves. (See *Hertz*, also *Clerk-Maxwell*, and *Ether Theory.*)

HETERODYNE—The term originally applied by Fessenden to a method of reception of electromagnetic waves, wherein the principle of *Beats* (q.v.) was employed. It is derived from the words "hetero," meaning other or different, and "dyne," meaning power.

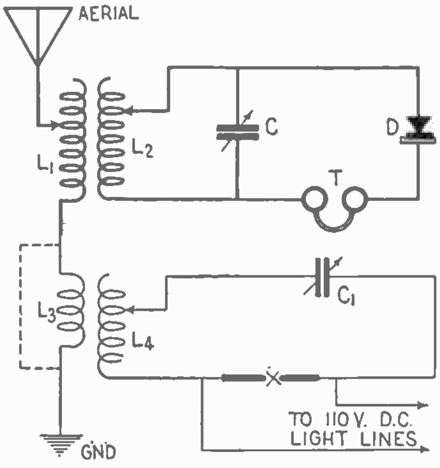

Fig. 1.—The original form of heterodyne receiver.

In its original conception by Fessenden, before the widespread use of vacuum tubes, it was used in conjunction with crystal receivers for the reception of *undamped waves* (q.v.). The heterodyne principle involves production of a local series of oscillations which are tuned to a frequency differing from the frequency of the incoming oscillations from the antenna. The difference between the two is known as the *beat frequency*. (See *Beats*, also *Heterodyne Receiver.*)

HETERODYNE RECEIVER—A receiver for electromagnetitc (radio) waves employing the *heterodyne* principle. The illustration Fig. 1 shows the original form of heterodyne receiver, wherein the local oscillations were produced by means of a direct current arc gap and associated circuits, and combined through a coupling medium, with the oscillations in the antenna circuit. Here the heavy lines represent a conventional crystal receiver circuit. L1 and L2 are respectively primary and secondary of a vario-coupler, C is a variable tuning condenser, D the crystal detector and T the head phones. If the dotted line is assumed as a connection to the ground, this will constitute a complete receiver. For the heterodyne purpose, however, a special coil L3 is employed in the ground circuit. Coupled to this is the secondary L4. C1 is another tuning condenser and A is an arc gap fed by 110 volts direct current. In operation the circuit represented by the light lines acts as a generator of undamped oscillations, these oscillations being supplied to the antenna circuit and the frequency varied by means of the condenser C1 and the coil L4. If the receiving tuner is tuned to a certain wave-length corresponding to incoming signals, and the generator system tuned to a different frequency, the difference between the two frequencies will be a *beat* frequency which will be present in the detector circuit. This new series of oscillations will be rectified by the detector D and will be audible in the head phones, T. The action of this system is shown by the series of curves given in Fig. 2. Here the incoming oscillations (signals) from the antenna are shown in graph 1 at the top of the illustration. These oscillations are of continuous amplitude or *undamped* (q.c.). Graph 2 shows the local oscillations. If we assume that the distance from A to B represents the time, it will be apparent that there are more oscillations per unit of time in the case of the incoming signals—

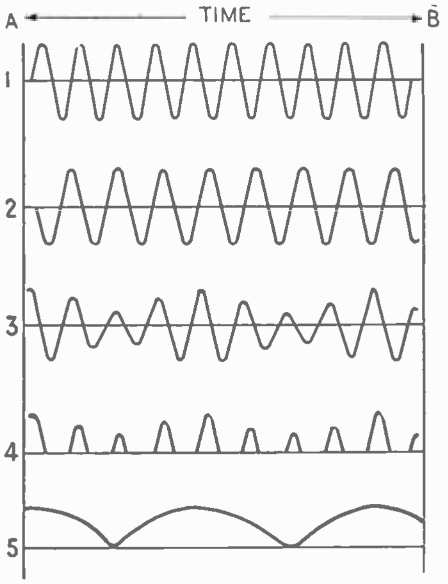

Series of graphs showing the action of a heterodyne receiver.

hence the second graph indicates that the local oscillations have a lower frequency. Graph 3 represents the *beat* resulting from interaction of the two upper series of oscillations and the frequency corresponds to the numerical difference. When the two upper groups of oscillations are opposed, the beat current is zero, and when they are in phase, or assisting each other, the beat current is maximum. When rectified by the crystal detector action the beats take the form shown in graph 4, the lower half of each cycle having been submerged. The telephone current of a periodic nature is indicated by the bottom graph, the original signals being represented in the phones by a series of contractions of the diaphragm. (See *Heterodyne*, also *Beats* and *Super Heterodyne.*)

HETERODYNE, SELF—See *Self-Heterodyne.*

HETERODYNE WAVE-METER—A wave-meter—device for measuring the wave-length of coils, circuits or incoming and outgoing signals—which utilized the heterodyne principle. Briefly the action is as follows: It has already been explained under Heterodyne Receiver that when the local oscillations and the incoming oscillations are opposed, the beat current is zero. It follows, then, that with a pair of phones connected in the circuit, the beat signals will be audible only when the two frequencies are different. Now if a suitable means is present for varying the frequency of the local oscillations, in practice a combination of coil and condenser, the local circuit can be tuned to a point where the two frequencies concur or are similar and no beat is audible. As the frequency is proportional to the capacity and inductance in a circuit, and as wave-length can easily be figured when the frequency of the local generating system is known, the wave-length of the incoming signals must be that of the local circuit when the beat is zero. The condenser dial on the heterodyne wavemeter can thus be calibrated directly in meters and a nearly true reading of the wave-length of incoming signals obtained. (See *Beats*, also *Heterodyne* and *Wave-meter.*)

HIGH FREQUENCY—A very general term applied to alternating electric currents. As a rule, alternating currents having frequencies above one thousand *cycles* (q.v.) per second are referred to as high frequency currents. When the frequency is extremely high, of the order of hundreds of thousands of cycles, the current is usually referred to as high frequency, oscillatory current. There is no definite level or line of distinction between high and low frequency currents. In radio practice, the audio frequency currents are generally considered below twenty thousand cycles per second and known as low frequency currents. The radio frequency or frequencies above twenty thousand are considered as the high frequency currents, although here again there is no definite level or line of demarcation between the two areas. (See *Audio Frequency*, also *Radio Frequency* and *Frequency.*)

HIGH FREQUENCY ALTERNATING CURRENT—See *High Frequency Current.*

HIGH FREQUENCY ALTERNATOR—An alternating current machine or alternator, designed to produce currents having high frequencies, generally above one thousand cycles per second. Such machines are used in high power radio telegraph transmission (not broadcasting) producing undamped or continuous waves. (See *Alternator*, also *Alexanderson Alternator.*)

HIGH FREQUENCY AMPLIFICATION—The process of stepping up or amplifying currents of high frequency. In radio circuits, the high frequency currents are known as radio frequency oscillations. In radio circuits the audio frequency (low frequency) currents and the high or radio frequency currents are usually in different parts of a circuit. When we consider alternating currents having a frequency above, say, twenty thousand cycles per second as radio frequency, and those below that limit as audio frequency, it is easy to consider the radio frequency currents as existing in all stages between the aerial and the detector. Beyond the detector, the alternating currents having become uni-directional surges, we can consider these currents as being audio frequency currents.

Thus we have a definite dividing line in any radio receiver using but one detector, the detector or rectifier marking the transition from radio frequency to audio frequency. Thus, any means of amplifying the signals before they reach the detector will be known as high frequency amplification. (See *Audio Frequency*, also *Amplifier.*)

HIGH FREQUENCY BUZZER—A *buzzer* (q.v.) designed to produce a very high note. The term "high frequency" in this case merely indicates that the pitch or frequency (vibration) is high in comparison with the ordinary buzzer, the sounds of course being actually of audible frequency. The general scheme of such a buzzer is about that of the ordinary type as used in conjunction with electric bells or signal systems, the armature, however, being very thin and thus permitting rapid make and break for the current. Such buzzers are used extensively in connection with crystal receivers for the purpose of testing and also for code practice. (See *Buzzer* also *High Frequency* and *Armature.*)

HIGH FREQUENCY COMPONENT OF PLATE—The high frequency currents present in the plate circuit of a vacuum tube receiver, as distinguished from the audio or low frequency currents. (See *Feed-Back*, also *Plate Component, High Frequency*, and *Regeneration.*)

H.F.C.—Abbreviation used abroad and occasionally in the United States for high frequency current. The more common term is radio frequency current and no abbreviation is used as a rule. (See *Radio Frequency Current*, also *High Frequency.*)

HIGH FREQUENCY CURRENT—Alternating currents having frequencies above certain limits, as, for example, the radio frequency currents employed in radio transmission, which are usually of a frequency above one hundred thousand cycles per second. (See *High Frequency*, also *Low Frequency* and *Radio Frequency.*)

HIGH FREQUENCY OSCILLATION—The term oscillation is generally applied to an alternating current of very high frequency. Thus radio frequency currents, which may run into hundreds of thousands of cycles per second, are called high frequency oscillations. (See *Oscillation.*)

HIGH FREQUENCY RESISTANCE—Resistance to the passage of alternating currents of high frequency. Said of wires, coils, circuits and apparati. The resistance of a certain conductor or part of an electrical circuit, as in radio, may be comparatively higher in the case of high frequency currents than to alternating currents of a low frequency. This is due to the "*skin effect*" (q.v.) it being a phenomenon of radio frequency (high frequency) currents that they have a tendency to concentrate near the surface of the conductor. For this reason, small wires are more apt to have high radio frequency resistance, as obviously the surface on which the currents can travel is restricted. (See *High Frequency*, also *Skin Effect* and *Resistance.*)

HIGH PASS FILTER—An arrangement of electric circuits designed to permit ready passage of high frequency alternating currents while at the same time presenting high resistance or impedance to alternating currents of low frequency, generally within certain definite limits. (See *Filter* also *Filter, High Pass.*)

HIGH PRESSURE—A term occasionally used to denote high potentials or voltages. (See *High Tension*, also *High Potential.*)

HIGH POTENTIAL—One of the several terms used to indicate high voltage. (See *High Potential Battery*, also *High Voltage.*)

HIGH POTENTIAL BATTERY—Another term for "B" battery, the battery used to supply potential to the plate of a vacuum tube. The term "high potential" is used here, only in a comparative sense, the voltage of the battery being high in proportion to that of the filament or "A" battery. High potential batteries for radio usage are generally furnished in blocks of 22½ or 45 volts. (See *High Voltage, High Tension* and *"B" Battery.*)

HIGH RESISTANCE—The property of any circuit whereby it offers considerable opposition to the flow of electric currents. There is no very distinct line of demarcation between high and low resistance. One of the common applications of the term is in the case of a *grid leak* or grid resistance. Here the resistance may be upward of a half million ohms, running into several million ohms in certain cases. An example of low resistance elements is that of a *rheostat* for controlling the filament current. Here the resistance is relatively low, rarely more than thirty ohms. The average resistance of the magnet windings in headphones (as used in radio) is one thousand ohms. The term "high resistance" is also used in a comparative sense in respect to various instruments or parts of radio circuits. Here the resistance in most cases should be as low as possible, and therefore a resistance of even a few ohms may be referred to as *high*. While resistance to direct currents is generally the feature of resistance coils and units, the opposition to currents of an oscillatory nature (high frequency alternating currents) may be of the utmost importance, and it is often the case that a circuit has little direct current resistance but very high resistance or *impedance* (q.v.) to high frequency currents. (See *Resistance, Impedance*, also *Skin Effect.*)

HIGH RESISTANCE JOINT—Any connection between two wires or conductors which offers resistance to the passage of current due to defective contact or imperfect connection through solder or any other medium. When bare wires are joined they should be carefully scraped and soldered. Too much solder is a bad feature, as the resistance of solder is naturally far higher than that of the copper wire. (See *Corrosion*, also *Soldered Joints.*)

HIGH RESISTANCE TELEPHONES—Headphones or telephones used in radio reception, having high resistance windings. In the ordinary telephone receiver used with the commercial telephone systems, the windings of the telephone magnets have a relatively low resistance, generally less than one hundred ohms. Such phones would not be sufficiently sensitive for radio usage, the average resistance in this case being about one thousand ohms. The resistance here refers to the ohmic resistance of the many turns of wire wound around the magnets. This resistance is not a required feature, the factor of importance being the number of turns. Inasmuch as space is a factor to be considered, it is necessary to use very fine wire in order to permit use of the thousands of turns essential. The resistance of this fine wire is of course high. Where very high resistance windings are used—and this indicates small wire—it is sometimes advisable to use a by-pass for the heavy plate potentials to protect the windings. (See *Ampere Turns*, also *Output Transformer* and *High Resistance*.)

HIGH SPEED RECEPTION AND TRANSMISSION—The reception or transmission of radio messages by automatic means in order to obtain much greater speed than would be possible under the manual system or hand opereating and recording. The general practice is to record the messages to be transmitted on a tape by means of a punching machine. They are then sent automatically through a high speed automatic transmitter and received automatically at the same high speed by a machine which punches the messages or records them on some form of tape or sheet. The more common of these methods is the *Wheatstone Transmitter.* (q.v.).

HIGH TENSION—One of the numerous terms used to denote high voltage. "B" batteries as used in radio reception are generally known as high tension batteries to distinguish them from the low voltage storage battery or dry cells used to light the filaments. Abbreviation H. T. mostly used in magazines and books printed in England.

HIGH TENSION BATTERY—Term applied to "B" batteries used in radio reception to supply potential to the plates of the vacuum tubes. These batteries usually are in blocks of twenty-two and a half or forty-five volts. (See *"B" Battery*.)

HIGH TENSION CIRCUIT—Term occasionally applied to the part of a radio circuit in which high potentials are present, as in the plate circuit of a vacuum tube. Here the "B" battery is in the circuit. (See *High Tension*, also *"B" Battery*.)

HIGH TENSION INSULATOR—Insulators of large dimension and rugged design, possessing unusual insulating properties in order to withstand high voltages. As a rule such insulators are made of high grade porcelain and arranged with annular ridges or "petticoats" to give greater surface and reduce the possibility of surface leakage. The illustration shows a form of high tension or high potential insulator. (See *Insulator*, *High Tension*, also *Petticoat Insulators*.)

A high potential insulator.

HIGH VACUUM—A term applied to vacuum tubes when they have been highly exhausted of air. That is, an almost perfect vacuum. When a vacuum tube has a very good vacuum—has been highly evacuated of gases—it is generally called a "hard" tube. A high vacuum is essential in tubes to be used as amplifiers or for transmission. (See *Vacuum Tubes, Theory of Operation*.)

HIGH VOLTAGE—A very indefinite term, used in ordinary electrical practice to denote voltages above 600, but applied in radio practice to much lower potentials, as, for example, a "B" battery, which is generally referred to as a high voltage battery in spite of being generally below 100 volts. The term used in this sense is one of comparison with the low voltage of a storage or other type of "A" battery. (See *Voltagee*, also *Potential*.)

HIGHLY DAMPED WAVES—An expression applied to waves or trains of waves wherein the amplitude or strength of each succeeding oscillation or train of oscillations is greatly decreased. If the progressive decrease in amplitude or strength is very gradual, that is, if the waves die out slowly, they are said to be feebly damped, but if they die out rapidly they are referred to as being highly damped. (See *Damped Waves*, also *Decrement* and *Amplitude*.)

HOGAN, JOHN V. L.—American radio expert. Born at Philadelphia, Pa. He was educated at Sheffield Scientific School, Yale University, where he made a special study of physics and mathematics. In 1906 he became assistant to Dr. Lee De Forest, and in 1909 he joined the National Electric Signalling Co. In 1914 he was appointed chief research engineer to the International Radiotelegraphic Company. Hogan has written many articles on radio and is a past president of the Institute of Radio Engineers, member of the American Institute of Electrical Engineers, of the American Association for the Advancement of Science, of the Radio Club of America and other societies.

John V. L. Hogan.

HOMOPOLAR—Synonymous with unipolar. Having but one pole. Said of a dynamo having its conductor or armature rotating around a single pole of a field magnet. (See *Multi-Polar*.)

HONEYCOMB COIL—A coil used in radio, having a peculiar construction, somewhat similar to a honeycomb. The illustration shows a typical honeycomb coil used as an inductance in receiving circuits. The object of this staggered winding system is to reduce the self-capacity between wires to a minimum without reducing the inductive value. The chief value of such a winding method is that it permits the retaining of high inductance value in a limited space, without undue *high frequency resistancee* (q.v.). (See *Inductance*, *Distributed Capacity* and *High Frequency Resistance*.)

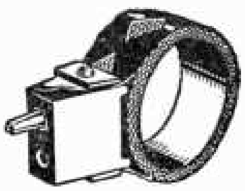

Honeycomb coil used in radio reception.

HOOK UP—The general term in the United States to designate the diagram for any complete radio receiver or transmitter, or for any specific part of such a circuit. Thus, when we refer to a hook-up for Radio receiver, we mean a diagram, either schematic (using symbols for the parts) or perspective (showing pictures of the parts) which gives the various parts necessary and the manner of connecting them to each other to form the complete circuit. Similarly we may use the term "hook-up" to describe a sketch or circuit for connecting various instruments such as *ammeters*, or *voltmeters*, *wavemeters* and the like. (See *Circuit*, also *Diagram*.)

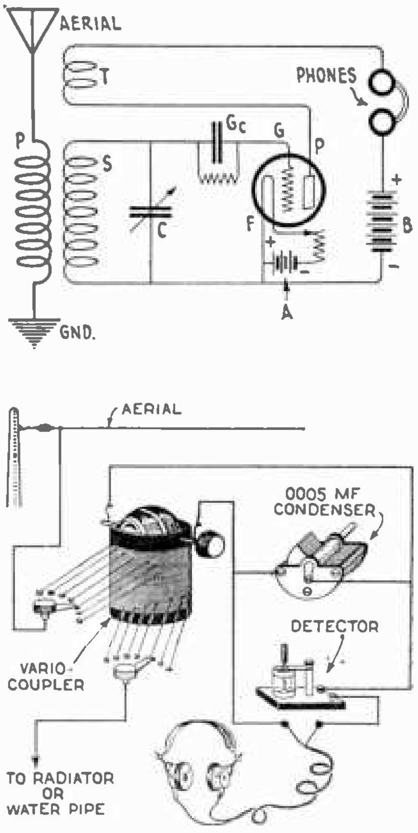

The upper diagram is in schematic form and shows a single tube regenerative receiver, and the lower diagram is in perspective form and shows a crystal receiving set.

HOOPER, STANFORD C.—American Naval wireless expert. Born August 16, 1884, at Colton, California, and educated at San Bernardino, California. He joined the Southern Pacific Company as a telegraph operator. In 1901 he entered the Naval Academy at Annapolis, and entering the Navy was advanced to Commander in 1918. From 1910-11 he was instructor in electrical engineering, physics and chemistry at the U. S. Naval Academy, and in 1912-13 Fleet Radio Officer of the United States Atlantic Fleet. During the great war he was responsible for the supply of wireless instruments, etc., for the American Navy, and he was also responsible for the construction of many of the larger American wireless stations including those at Annapolis, San Diego and Pearl Harbor. Hooper was one of the chief men concerned with the radio-compass system now used in the American Navy.

HORIZONTAL AERIALS—Aerials—overhead wire systems—for the reception or transmission of radio signals, wherein the major portion of the aerial is horizontal—parallel to the earth's surface. Typical examples are the "T" and inverted "L" aerials. In the former type, the flat portion suspended horizontally between two supporting structures and a lead-in wire is attached to the center and thence to the apparatus. In the latter type, the lead-in is attached to either end, giving a shape somewhat similar to an inverted L, from which the name is derived. The more common term is "flat top." This term is now somewhat ambiguous, as single wire aerials are used more or less extensively for receiving purposes, whereas formerly a system of several horizontal wires was employed. (See *Flat Top Aerial or Antenna*, also *Aerial T Type* and *Inverted L Aerial.*)

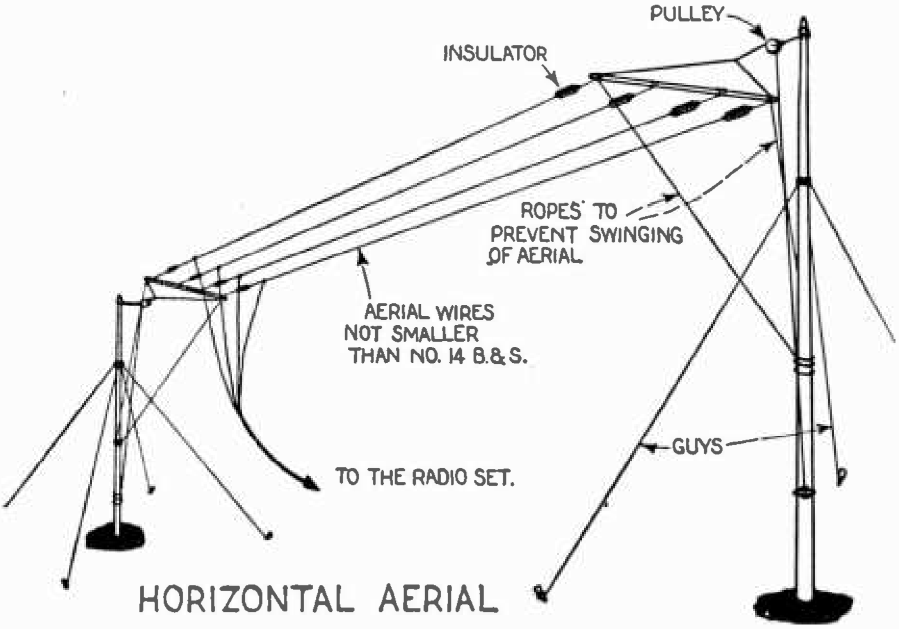
PULLEY
INSULATOR
ROPES TO PREVENT SWINGING OF AERIAL
AERIAL WIRES NOT SMALLER THAN NO. 14 B.&S.
GUYS
TO THE RADIO SET.
HORIZONTAL AERIAL

A horizontal aerial of the inverted "L" type.

HORN, LOUDSPEAKER—A device, generally in the shape of a horn, used in conjunction with a heavy duty telephone receiver or loudspeaker unit, to reproduce loudly the signals or music from a radio receiver. In operation, the horn has attached to it at its small end a unit or telephone receiver. When the diaphragm of the unit is actuated by the signals, speech or music, the column of air in the open tube is displaced, creating audible sounds which can be heard at a distance or throughout a room. Assuming other conditions equal, the volume will be more or less in proportion to the size of the norn, or more specifically to the amount of air displaced by the signals. When a telephone receiver is used without horn, the opening above the diaphragm being unconfined, only a small column of air is displaced and the volume will not be great under ordinary circumstances. When a horn is attached, however, the sound being confined will displace a relatively greater amount of air—depending mainly on the shape and size of the horn—and the volume will be measurably greater. (See *Loudspeaker*, also *Acoustics* and *Loudspeaker Units.*)

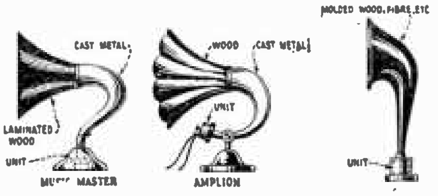

Three types of horn loudspeaker.

HORSE POWER—The unit of power used in the United States and Great Britain. It is defined as the power required to lift 550 pounds to a height of one foot in one second, or similarly, 33,000 pounds to that height in a minute. The term "foot-pounds" is derived from this value. Thus, one horsepower is the power required to perform 550 foot-pounds of work per second of time. (See *Horse Power, Electrical.*)

HORSE POWER, ELECTRICAL—The unit of electrical work similar to the mechanical horse power. It is actually mechanical horse power expressed in *watts* (q.v.). One electrical horse power is equivalent to 746 watts. In the *centimeter gram second* (q.v.) system of units, the unit of work is the *erg.* This unit is too small for ordinary usage and the customary unit is the watt, which is equal to 10^7 ergs per second. The unit used in electrical power work is still larger; it is the *kilowatt* or one thousand watts. (See *Units*, also *Watts.*)

HORSESHOE MAGNET—A *magnet* (q.v.) in the shape of a horseshoe or letter U, both poles being brought nearer together so that they may act on the armature or "keeper." Generally a magnetized steel bar which retains its magnetism for a long period. The illustration, Fig. 1, shows a permanent horseshoe magnet. Such magnets are also used with external power, then being termed *electromagnets.* In this case the magnet is made of soft iron. The illustration, Fig. 2, shows a horseshoe *electromagnet.* Here the magnetic attraction is supplied by application of an external electric force to the windings WW, and the magnet being of soft iron, loses its magnetic attraction almost immediately upon withdrawal of the electric current, magnetizing force. (See *Telephone Receiver*, also *Electromagnet.*)

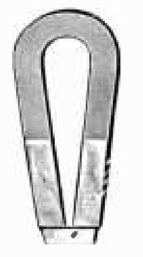

Fig. 1. A permanent horseshoe magnet.

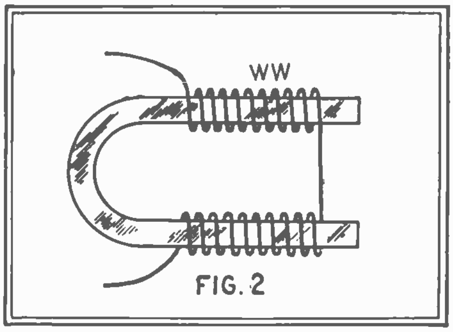

Fig. 2. A horseshoe electromagnet.

HOT-WIRE AMMETER—An ammeter, or instrument for measuring electric current amperes, which depends for its action on the expansion of a fine wire under the influences of the heat produced in it, by the passage of the electric current to be measured. In the illustration the current to be measured is passed through the fine resistance wire A-B. P is a pointer moving over a graduated scale, and attached at its other end to a coiled spring S, which in turn is attached to the wire A-B by the silk thread T. The tendency of the spring is to move the pointer to the right, but owing to the pull of the thread in the normal position of the wire A-B, the pointer rests at zero at the left of the scale. As current is passed through the wire A-B, it heats and expands, thus permitting more or less slack to occur in the thread T and allowing the spring S to move the pointer over the scale toward the right, As soon as the current is withdrawn, the wire cools and contracts to its former size, thus again restoring the tension to the thread and returning the pointer to zero. Hot Wire ammeters are used extensively for measurement of currents at *radio frequency*, that is, extremely high frequency as in radio transmission. The average ammeter using a coil of fine wire would not be adaptable to this use, as the high voltage, high frequency current would burn out the coil. In addition to this. the length of these windings would affect the circuit and probably throw it out of *resonance* (q.v.). (See *Ammeter* or *Ampere Meter, Frequency, Resonance,* also *Instrument Shunts.*)

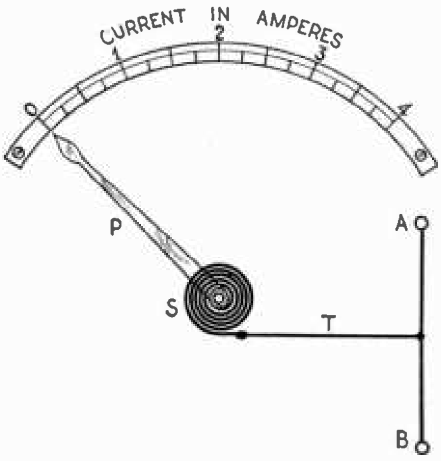

Illustration showing operation of a hot wire ammeter.

HOT WIRE INSTRUMENTS—Current measuring instruments which utilize the tendency of a fine wire to sag or expand under the heat caused by passage of an electric current through it. Such instruments may be used for direct or alternating current. (See *Hot Wire Ammeter.*)

HOT WIRE OSCILLOGRAPH — See *Oscillograph.*

HOT WIRE TELEPHONE—See *Thermal Telephone.*

HOT WIRE WATTMETER—An instrument used to measure combined volts and amperes, i. e., watts in a circuit. It comprises essentially a mirror, which is deflected according to the difference in expansion of two fine wires. (See *Hot Wire Instruments,* also *Wattmeter.*)

HOUR METER, AMPERE—An instrument used extensively in connection with storage batteries, to measure the input or output in amperes per hour. (*Ammeter* or *Ampere Meter.*)

HOUSE MAINS—A term much used in connection with storage battery charger and battery elimination devices. The term implies the main circuit of any standard house lighting system, either direct or alternating current. Thus in the case of a rectifier for charging storage batteries from alternating current, a lead is provided for connecting to the house mains, offering a means for introducing the required current into the device. (See *Current Direct,* also *Alternating Current* and *Chargers, Storage Battery.*)

HOWLING—The general term used to imply any undesired sounds produced in or by a radio receiver. Howling is usually caused by *feed-back effects* (q.v.). The most common example of howling is in the case of a regenerative receiver. Here the detector tube may have a tendency to generate its own oscillations As long as these oscillations are equal in frequency to any incoming signals there will be no howling, but should these local oscillations differ in frequency from the incoming series of signals, the result will be a howl resulting from the *beats* (q.v.) established owing to the difference between the two sets of oscillations. This is essentially the principle of *heterodyne* or "beat" reception, but in the case of a heterodyne, the effect is utilized, whereas in the case of a regenerative receiver it is undesirable. Interaction between various parts or different parts of a circuit may often lead to howling. (See *Feed-Back,* also *Regeneration, Reflex, Beats* and *Re-Radiation.*)

H. P.—Abbreviation sometimes used in radio for *high potential* or *high pressure.* In mechanics it is the standard abbreviation for *horse power.*

HOZIER-BROWN DETECTOR—A form of electrolytic detector, which, however, does not use a liquid electrolyte. It consists of a small pellet of lead oxide held more or less rigidly between two electrodes, one platinum and the other lead. (See *Detector,* also *Electrolytic Detector.*)

HYDROGEN—A gaseous chemical element. The symbol is H and its specific gravity is taken as 1, that of air being 14.4. It is colorless, tasteless and odorless and is the lightest known body. Hydrogen is a constituent of water, which is composed of one atom of oxygen and two atoms of hydrogen according to the chemical formula for water, H_2O. Hydrogen is used in radio transmission in connection with *arc generators.* (q.v.) (See *Poulsen Arc.*)

HYDROMETER—An instrument used to measure the *specific gravity* (q.v.) of a liquid. In radio it has an extensive use in connection with storage batteries. Various types of hydrometers are used to indicate the specific gravity of the electrolyte of a storage battery and hence in a manner, to indicate the state of charge of the battery. The illustration shows the standard form of hydrometer using a weighted float. In the illustration the float is shown inside the barrel of the instrument. In operation the bulb B is compressed and the tube T inserted in the open vent of a storage battery. When pressure on the bulb is released some of the electrolyte is drawn up inside the barrel and the float rises more or less within the barrel in proportion to the specific gravity or density of the electrolyte. With the hydrometer held in the manner shown, that is, upright, the numbers just at the level of the liquid in the barrel indicate the specific gravity. Thus, if the reading is 1250° or higher the battery is considered in fairly good operating condition —full charge being 1280° to 1300°. If the reading is below 1200° it indicates the need of charging, these levels not being arbitrary, a discharged condition being indicated by a reading of 1150° or lower. The use of a hydrometer is generally advisable and the reading accurate under ordinary conditions. If, however, water has just been placed in the battery, the reading may not be accurate for several hours due to the fact that the water and acid may not mix immediately. (See *Storage Battery Tests,* also *Electrolyte.*)

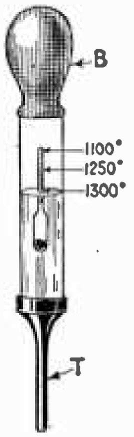

Hydrometer used to test the specific gravity of liquids.

HYGROSCOPIC EFFECT — The effect in any substance which permit it to absorb moisture. In the case of forms or tubes for radio coils and inductances this feature is particularly undesirable, as obviously, moisture lowers the dielectric value of the form and may cause it to warp. (See *Bakelite,* also *Tubing.*)

HYSTERESIS—The tendency of magnetization to lag behind the magnetizing force as, for instance, in an iron core in a *transformer* (q.v.). When the core of such a transformer is undergoing rapid reversals of magnetism, there is often an expenditure of energy which is not useful and which is converted into heat, thus representing a loss in power. This effect is more noticeable with certain qualities of iron. In other words, when rapid reversals of magnetism take place in a poor quality of iron core it may display a certain sluggishness. This is actually the lagging of the magnetic flux behind the magnetizing force producing it. This loss of energy is assumed due to the work required to alter the position of the molecules of the iron composing the core and the less energy expended in this manner the less the hysteretic losses. In designing transformers or armature cores, hysteresis must be taken into consideration and the iron must be of high quality to hold these losses at a minimum. (See *Eddy Currents,* also *Core, Transformer.*)

HYSTERESIS COEFFICIENT — The amount of energy wasted or lost in the process of magnetizing and de-magnetizing a unit volume of magnetic material is referred to as the hysteresis coefficient of that material. (See *Hysteresis,* also *Coefficient.*)

HYSTERESIS LOOP—Graphic illustration of losses due to hysteresis in any sample of core material. Loops are plotted in graphic curves of cycles of magnetization. (See *Curve,* also *Hysteresis.*)

HYSTERESIS LOSSES — The energy wasted or lost through hysteresis. (See *Hysteresis.*)

HYSTERETIC LAG — The lagging of magnetization due to the effect of hysteresis. (See *Hysteresis,* also *Magnetization.*)

I

I—The abbreviation adopted by the *International Electrotechnical Commission* (q.v.) and American Institute of Electrical Engineers, to denote current in amperes. The small or lower case (i) may also be used where the capital letter is not advisable. (See *Current.*)

I ARMATURE—An armature used in generating machines, having a core resembling the letter I in shape. (See *Armature,* also *Core.*)

I.C.W.—The customary abbreviation for "interrupted continuous waves." This is a system of radio transmission wherein the waves are modulated at a constant low frequency. (See *Interrupted Continuous Waves,* also *Modulation.*)

IDLE CURRENT—A name often applied to the component of an alternating current which contributes nothing to the power. This is also known as the *wattless component* (q.v.) or the *wattless current* (q.v.). It represents the component which, being in quadrature with the electromotive force in the circuit, thus has no active value—adds nothing to the total power. (See *Current,* also *Electromotive Force* and *Alternating Current.*)

I.E.C.—The abbreviation for *International Electrotechnical Commission,* the body which suggested the international electrical and magnetic abbreviations and symbols. (See *Symbols,* also *Units.*)

IMPACT OR SHOCK EXCITATION—The method of producing *free oscillations* (q.v.) in a circuit by means of an exciting current of short duration compared to that of the resultant oscillations. The term is a general one and broadly covers the various methods of generating oscillations in which an oscillatory circuit is thrown into electrical vibration at its natural or fundamental frequency by means of an electromotive shock or exciting force of short duration. Also covers the term *Impulse Excitation.* (See *Oscillations,* also *Fundamental Frequency.*)

IMPACT TRANSMITTER—Apparatus for transmitting radio waves wherein the transfer of energy from the primary or exciting circuit to the oscillatory circuit takes place during one pulse or vibration of the exciting current. (See *Impact or Shock Excitation,* also *Transmitter.*)

IMPEDANCE—The total resistance to the flow of alternating current in a circuit. Impedance combines the ohmic resistance and the apparent resistance due to self-induction. When a direct electromotive force is impressed on a circuit, the current flowing in that circuit depends directly on the ohmic resistance of the circuit. The formula is according to Ohm's law $I = \frac{E}{R}$, I representing the current in amperes, R the resistance in ohms and E the electromotive force in volts. Now if this direct electromotive force is replaced by an alternating EMF (electromotive force) and the ohmic resistance remains the same, the total resistance of the circuit to the alternating EMF will nevertheless be greater than in the case of direct current. Now the ohmic resistance is still the same but we have added to it the apparent resistance due to the self-induction of the circuit. In other words the inductance of the circuit has a marked effect on the current flowing in the circuit. What actually occurs is that the self-induction sets up a *counter-electromotive force* (q.v.) which continues to act against the impressed EMF as long as the flow of current persists. The result is a reduction in the current flowing in the circuit, the EMF having been, in effect, reduced, and hence according to the law above, $I = \frac{E}{R}$, the current is less. This added resistance, due to self-induction, is referred to as *reactance* and the flow of current in an alternating current circuit is thus governed by the ohmic resistance and the reactance. The application of the term "apparent resistance" to describe the added resistance due to self-induction, is a concession to the fact that it is not actually resistance in the proper sense of the word. In a direct current circuit, the resistance actually wastes or dissipates a certain amount of energy. In the alternating current circuit, the ohmic resistance has the same effect. The reactance, however, does not dissipate energy. It merely reduces the effective EMF by means of the counter EMF and thus necessitates the application of a higher impressed voltage in order to maintain the same flow of current.

This is one of the major advantages of alternating over direct current. Alternating current can be transformed readily and can also be readily controlled by means of a reactance or choke coil, which by creating a counter EMF limits the current flowing in the circuit without dissipating any great amount of energy. (See *Alternating Current, Current Direct, Choke Coil,* and *Reactance.*)

IMPEDANCE COIL—A coil of wire or inductance coil, generally wound on an iron core, used in alternating current circuits to limit the current flowing in the circuit. In action such a coil serves through the medium of its self-induction, to create a counter electromotive force which acts against the impressed EMF (electromotive force) and thus reduces the current in the circuit according to ohm's law that the current is equivalent to the EMF in volts, divided by the resistance in ohms. An impedance coil is also known as a choke coil although the two may have somewhat different meanings. For example, a choke coil might be used in either direct or alternating current circuits, whereas an impedance coil as its name indicates, offers impedance, an effect only present in alternating current circuits. The simplest formula for impedance of a coil is the following: Impedance $= \sqrt{R^2+P^2L^2}$. Here it is assumed that the current takes the form of a *sine wave* (q.v.), then R is the resistance of the coil, L the coefficient of self-induction in henries and p is 2 π times the number of alternations per second. (See *Choke Coil, Alternating Current, Current Direct, Reactance, Impedance,* also *Impedance Coupled Amplifier.*)

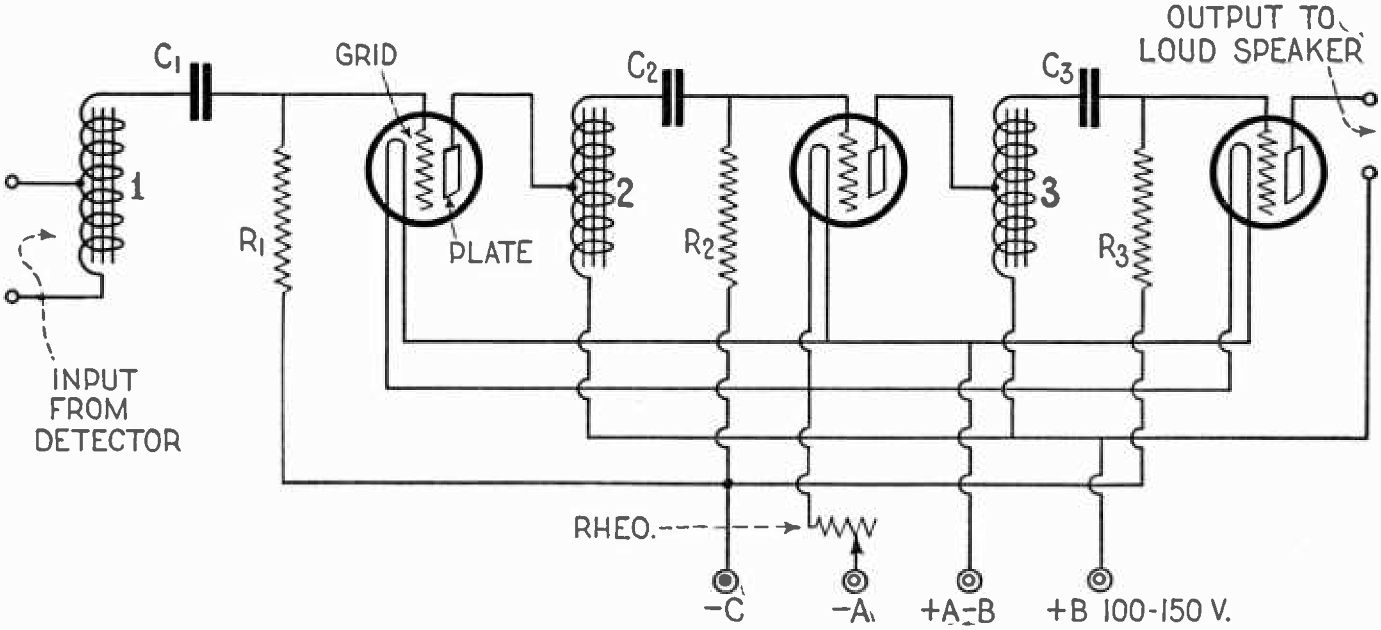

Schematic diagram of an impedance coupled audio frequency amplifier. *(See Impedance Coupled Amplifier.)*

IMPEDANCE COUPLED AMPLIFIER—A system of amplifying radio signals, voice or music, whereby the successive stages are coupled or joined by means of impedance coils. This system is much used in audio amplifiers, the amplification being much more uniform over the necessary range of frequencies than in the case of the ordinary audio frequency transformer. This is also known as choke coil coupling or choke coil amplification. The illustration on the preceding page shows a typical impedance coupled amplifier for audio frequency currents. It will be noted that three stages are used. This is necessary owing to the lower amplification per stage. Using this system, then, three stages are necessary in place of the customary two stages of transformer coupled amplification. In the illustration, which is a schematic diagram of the circuit for such an amplifier, the impedances 1, 2 and 3 are in the form of *auto transformers.* (q.v.) R1, R2 and R3 are grid leaks ranging in value between one quarter megohm and one and one-half megohms. C1, C2 and C3 are stopping condensers of about 1 micro-farad capacity each. The variations in potential are transferred successively from stage to stage, the amplification factor being roughly the sum of the individual factors of the tubes. In other words the ratio of the windings of the impedances is one to one, or unity, and therefore no voltage amplification takes place. The resistances are the usual grid leaks and the condensers C1, C2 and C3 are to block the high plate voltages from the grids of the succeeding tubes. (See *Tuned Impedance,* also *Amplifier* and *Vacuum Tube Amplifiers, Theory of.*)

IMPEDANCE FACTOR—The ratio of the impedance to the ohmic resistance in an alternating current circuit. (See *Impedance,* also *Resistance.*)

IMPEDANCES, MATCHED—Arrangement of an amplifying circuit and the choice of an audio transformer or loudspeaker with due regard for the impedance of the circuit. Thus if the impedance of a certain amplifier tube is 10,000 ohms, and the resistance of a loudspeaker unit being used is 2,500 ohms, then for maximum response when used with an *output transformer* (q.v.) the impedance of the primary of the transformer should be equivalent to the total or 12,500 ohms. (See *Matched Impedances.*)

IMPREGNATED INSULATING MATERIALS—Cloth or paper impregnated with an oily or resinous substance to render it moisture proof and improve the insulating properties. Cardboard is sometimes treated in this manner and used for tubing to wind inductance coils. (See *Empire Cloth or Paper,* also *Insulating Property.*)

IMPRESSED ELECTROMOTIVE FORCE—The term used to distinguish the EMF (electromotive force) impressed on or applied to a certain part of an electrical circuit from the counter or back-electromotive force due to constants of the circuit. The term is sometimes used where a direct current is applied in a circuit, but it should be confined to use where a counter EMF (electromotive force) is present. (See *Counter Electromotive Force,* also *Choke or Choking* and *Reactance.*)

IMPRESSED FIELD—An electric or magnetic field of force impressed on or applied to a body or region (area) for the purpose of producing other fields of force. (See *Field.*)

IMPULSE—Electrically, an electromotive force which produces an impulsive rush or discharge of electricity. The term is used to distinguish between this form and an electromotive force producing a steady flow of current. (See *Current,* also *Electromotive Force.*)

IMPULSE E.M.F.—The more specific term designating an EMF (electromotive force) which produces an impulsive rush or discharge of electricity. It is defiined technically, as having a maximum value which is large compared with its average value, the average value being taken over a time equivalent to the *time-constant* (q.v.) of the circuit on which the EMF is impressed. (See *Current, also Electromotive Force.*)

IMPULSE EXCITATION—Method of exciting or starting oscillations in a circuit, such as the aerial circuit of a transmitter, by a sudden surge of impulse EMF (electromotive force) rather than by application of oscillations of the frequency of the circuit.

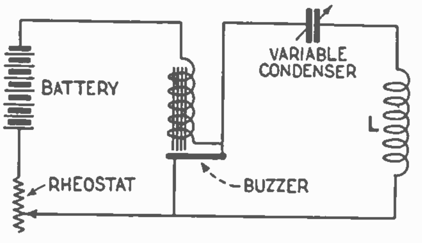

Connections of buzzer for generating damped oscillations suitable for testing by impulse excitation.

This system of impulse or shock excitation may be used for transmission and also has application in connection with testing or comparing of wavemeters. The illustration shows a buzzer connected with the necessary components to form a source of high frequency oscillations. This system is capable of producing through coupling from the coil L, currents by shock excitation. These currents can be used in a variety of ways for testing purposes. In operation the condenser is charged to the voltage of the battery at the instant the buzzer contact is open and when the contact closes, discharges through the coil L thus furnishing a series of oscilations to surge through the coil and by inductive coupling to any other coil placed in proximity to it. (See *Buzzer Exciter.*)

INCANDESCENCE—The glowing of any substance when brought to a white heat. The term is used in radio to signify the condition of the heated cathode or filament in a vacuum tube. In order to produce *electrons* (q.v.) it is necessary to cause the filament to become white hot, the condition being known as incandescence. (See *Filament.*)

INCONDUCTIVITY—A little used term signifying the condition of a substance by virtue of which it will not act as a conductor for electric currents. (See *Conductivity, Specific.*)

INDEPENDENT UNIT—A unit of a quantity, such as length, mass or time, chosen arbitrarily and without having any relation to any other unit. (*Centimeter Gram Second, or C. G. S.,* also *Fundamental Units.*)

INDICATING INSTRUMENTS — Any instrument or device which indicates values, as for example, a *voltmeter* or *ammeter.* (q.v.) (See *Instruments, Measuring.*)

INDIRECT CURRENT—Term for alternating current—now obsolete. (See *Alternating Current.*)

INDUCED CURRENT—The current in a circuit due to an induced EMF (electromotive force). The current produced by an EMF induced in a circuit due to a change in the magnetic flux linking the circuit with another. (See *Induced EMF,* also *Induction.*)

INDUCED E.M.F.—The electromotive force induced in a conductor or body. The phenomena of induced EMF (electromotive force) is the basis of operation of generating machines. Here a conductor is rotated or moved in a field of force, the action causing a change in the flux lines linking the circuits or the conductor. An EMF may also be induced by placing a coil in inductive relation to another coil through which alternating current is passing. Here the transfer is caused by induction. (See *Induction.*)

INDUCTANCE—The broad term used to designate any wire wound coil as used in radio practice for tuning units, inductive couplers and any variety of purpose. Actually the term is derived from self-induction or self-inductance, a property of a coil of wire by virtue of which it sets up a counter EMF (electromotive force) which serves to oppose the original EMF passing through it. When a certain circuit has inductance, it is to say that a number of turns of wire possessing self-induction are present in the circuit. The greater the inductance in a circuit the greater its opposition to the passage of an alternating current. The unit of inductance is the Henry. This is defined as the self-induction of a circuit or coil when the induced EMF is one volt with the inducing current varying at the rate of one ampere per second. (See *Alternating Current, Henry,* also *Inductance Coils.*)

A type of antenna inductance used in radio transmission.

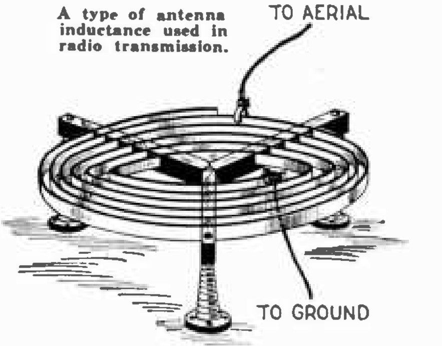

INDUCTANCE, ANTENNA—Inductance (coil) placed in antenna circuit of transmitting set for purpose of increasing wavelength of the open or antenna circuit to any desired level. It is generally referred to as an antenna loading inductance or loading coil. In high power transmitting stations it is very often the case that the natural period or fundamental wavelength of the antenna is not as great as the intended operating wavelength. In this case a loading inductance is used to bring it up to the desired value. In some instances the fundamental wavelength is too high, in which case a condenser is placed in series to reduce the wavelength and then a few turns of inductance placed in series with the condenser and antenna for the purpose of tuning. Such inductances must be of a very rugged nature when high power is used, generally consisting of a comparatively few turns of heavy copper ribbon mounted on high tension insulators. The illustration shows a common form of antenna inductance used in radio transmission. (See *Loading Inductance,* also *Tuned Antenna.*)

INDUCTANCE COILS—Any coil of wire wound in such a manner as to possess the property of self-induction. Such a coil may be in almost any conceivable form, providing the turns are separated or insulated from each other. Inductance coils are a very necessary part of radio circuits, being used in conjunction with condensers to obtain resonance or desired wave-length. Such coils can also be arranged as separate circuits and placed in inductive relation to each other, thus permitting energy present in one to be transferred through induction to the other. The most common example of inductance in a receiving set is that of a tuning coil or vario-coupler. The illustration shows several forms of inductances known as tuning inductances. (See under separate headings as follows: *Inductance, Loading, Tuning, Tapped,* also *Inductance, Units of.*

Honeycomb tvpe inductance coils.

Low-loss air-wound inductance coil made especially for tuned radio frequency work.

INDUCTANCE, DISTRIBUTED—Added inductance distributed throughout the length of a cable or telephone line to compensate for the capacity of the line. (See *Distributed Inductance.*)

INDUCTANCE, LOADING—Literally an inductance to load a circuit. The term is used to designate any inductance coil or unit used in series with the antenna for the purpose of increasing the wave-length of the circuit. (See *Inductance, Antenna.*)

INDUCTANCE, MEASUREMENT OF—Any process or means of determining the inductance value of a coil or circuit. The usual method is by formula. (Fully described under *Measurement of Inductance.*)

INDUCTANCE, MUTUAL—Symbol M. The number of linkages of flux lines or the flux that is common or mutual to two circuits which are adjacent to each other or inductively coupled. It is generally defined as the amount of the *flux* (q.v.) which is common to two linked circuits per unit of current flowing in the inducing circuit. In the illustration circuit A contains an inductance L_1 which is being traversed by an alternating electromotive force. Circuit B contains merely an inductance L_2 and a measuring device, M. Now, as the alternating E.M.F. passes through coil L_1 it sets up a field of force, or rather flux lines as shown. These lines link with the coil L_2 and an alternating E.M.F. is set up. Now, suppose that the inducing current flowing in circuit A is changing at the rate of one ampere per second and an E.M.F. of one volt is induced in the circuit B, or rather in coil L_2. We then say that the mutual inductance of the two circuits will be one henry. This is the same unit used for self-inductance and in fact, mutual inductance and self-inductance are virtually the same thing.

The effect of mutual inductance must be taken into consideration when measuring the fundamental wave-length, capacity or inductance of an antenna. That is to say, where a number of parallel wires are used, the mutual inductance between them must be taken into consideration as well as other values or constants. (See *Self-Inductance, Mutual Inductance, Induction, Coupling* and *Induced E.M.F.*)

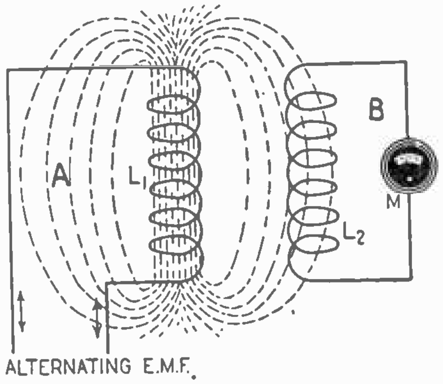

Diagram which illustrates the mutual inductance between coil L_1 in circuit A and coil L_2 in circuit B.

INDUCTANCE, REACTANCE—Symbol x—The *reactance* present in a coil possessing self-inductance. The self-inductance in such a coil opposes the flow of alternating current due to the back *electromotive force* set up as a result of the changing magnetic field of the current. *Inductance reactance* (or *inductive reactance*) is the term used to define this opposition. *Inductance reactance* acts similarly to *resistance* (q.v.) in that it opposes the flow of an electric current. It is measured in *ohms* (q.v.), just as *resistance* is measured. It should be noted, however, that *inductance reactance* only opposes the flow of alternating currents, whereas resistance opposes the flow of both direct and alternating currents. Furthermore, resistance in offering opposition to current flow consumes power, while *inductance reactance* merely takes up energy from the circuit momentarily while current is increasing and thereafter gives it back while the current is decreasing.

The combined opposition offered to the flow of alternating current by the resistance and the inductance reactance of a circuit is called the *impedance* (q.v.) of that circuit. The voltage drop due to resistance is in *phase* (q.v.) with the current. The voltage drop due to *inductance reactance* is 90 degrees in advance of the current.

Expressed as a formula, $x = \sqrt{Z^2 - R^2}$, where x is the inductance reactance, Z is the impedance and R is the resistance, all measured in ohms.

Knowing the *frequency* (q.v.) of the alternating current and the inductance of the circuit through which this current flows, it is possible to calculate the *inductance reactance* by using the formula $x = 2\pi fL$, where π is a constant equal to 3.1416, f is the frequency in *cycles* per second, and L is the *inductance* of the circuit in *henry's* (q.v.). This formula is sometimes abbreviated by substituting the symbol ω (omega) for $2\pi f$. From the formula it can be seen that the *inductance reactance* increases directly both with the inductance of the circuit and with the frequency of the current flowing. (See *Current,* also *Electromotive Force, Resistance, Reactance.*)

INDUCTANCE, RECEIVING—Broadly, any form of inductance used in receiving circuits. The term thus includes all types of coils for tuning purposes, such as couplers, tuned transformers, variometers, etc. The illustration shows a common form of tuning inductance. Here the coil L is an inductance in the aerial circuit and L_1 the coil in the secondary circuit. Here the energy from the aerial traverses coil L and by induction is transferred to coil L_1. (See *Tuning Coil, Vario-Coupler, Receiving Circuit.*)

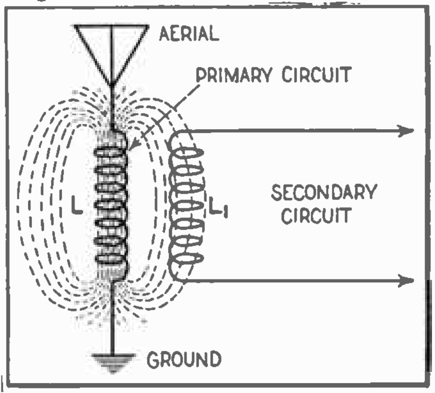

A form of tuning inductance sometimes used in radio receiving sets.

INDUCTANCE, SELF—Symbol L—An alternate term with self-induction for the *inductance* (q.v.) due to the field of force produced by the circuit itself. It is virtually the same as mutual inductance. For example, in a coil of wire being traversed by alternating currents the adjacent turns may act upon each other on the principle of mutual induction between two separate adjacent circuits. The induced currents are the result of self-induction. (See *Self-Induction,* also *Mutual Inductance* and *Induction.*)

INDUCTANCE, TAPPED—Any type of inductance having taps for connecting at different points on the coil in order to vary the inductance value. The illustration shows a form of tapped inductance. Here the taps at different portions of the coil are connected to switch points and the switch lever permits instant change to any of the available values. (See *Inductance Coils,* also *Tuning Coil.*)

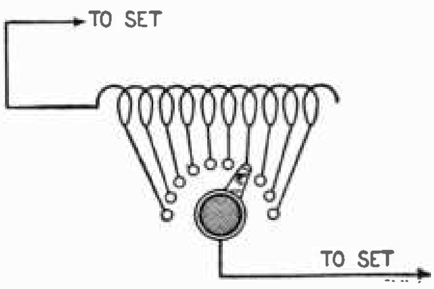

A tapped inductance which permits inductance to be added or taken out of the circuit by rotating the lever switch.

INDUCTANCE, UNITS OF—The measuring units for inductance. The practical unit is the henry (H), the value of inductance which, if current varies at the rate of one ampere each second, results in an *induced electromotive force* (q.v.) of one volt. When we speak of an induced E.M.F. of one volt, the volt is used as a practical unit. Under the electromagnetic system of units, the absolute unit of induced E.M.F. is known as the abvolt. This unit is very small, 100,000,000 of them being required to equal one volt. We thus say that the volt of induced E.M.F. is equal to $1x10^8$ abvolts. The electromagnetic unit of E.M.F., or the abvolt, is defined as the E.M.F. induced in a conductor cutting one line of magnetic force each second. As we have shown that one volt is equal to $1x10^8$ abvolts of induced E.M.F., and since one ampere is equal to $1x10^{-1}$ abamps of current (see *Units, Electromagnetic*) then the unit of inductance or one henry $= \dfrac{1x10^8}{1x10^{-1}}$ or) $1x10^9$ abhenrys. The unit of inductance (the henry) being too large for practical purposes, the *milli-henry* (one-thousandth henry) or the *micro-henry* (one-millionth henry) are commonly used. Occasionally the *centimeter* is used as a measure of inductance, being the thousandth part of a micro-henry or a *milli-micro-henry*. The milli-henry is written as $1x10^{-3}$ henry, the micro-henry as $1x10^{-6}$ henry, and the centimeter as $1x10^{-9}$ henry. (See *Induced Current, Induced E.M.F., Induction, Electromagnetic,* also *Units.*)

INDUCTANCE, VARIABLE—Any coil possessing self-inductance and being arranged in such a manner that the value of the inductance can be changed at will. The simplest form of variable inductance is the sliding contact tuner. This is merely a coil of wire of any predetermined value (maximum) mounted and arranged with a sliding contact moving over the wires, which are bared of insulation at a certain point to permit contact. By using the sliding contact as one connection and either end of the coil as the other lead, the effective value of the inductance in the circuit may be varied within the limits of the coil itself. Other forms of variable inductance include the variometer, vario-coupler and various modifications. The illustration shows the two more common forms of variable inductance. (See *Inductance*, and *Inductance Coils.*)

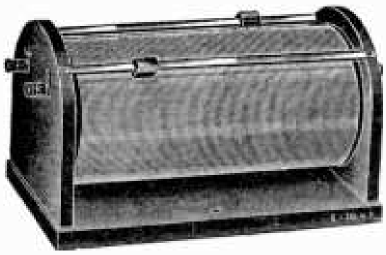
A sliding contact tuner formerly used in radio receiving sets.

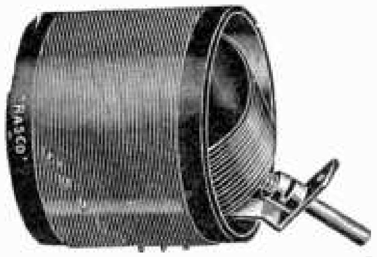

Illustration by courtesy of Radio Specialty Co.
Variable inductance known as a variocoupler.

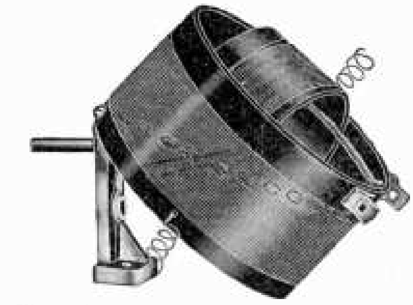
Illustration by courtesy of Radio Specialty Co.
Another type of variocoupler.

INDUCTANCES IN PARALLEL—Coils possessing the property of self-inductance or self-inductance, connected in parallel or shunt to each other. The illustration shows two standard coils connected in a circuit in this manner. The effective inductance is less than that of either coil alone. (See *Inductance Coils.*)

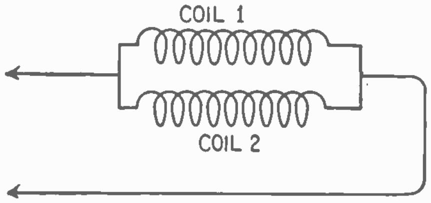

Diagram showing connection of two inductance coils in parallel in a circuit.

INDUCTANCES IN SERIES—Coils possessing the property of self-induction, connected in series in a circuit. The illustration shows two coils connected in series in a circuit. Here the effective inductance value is the sum of the separate inductances. (See *Inductance Coils.*)

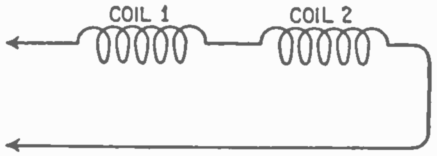

Diagram showing two inductance coils in series in a circuit.

INDUCTION—The influence exerted by one field of force (electromagnetic or electrostatic) on another or on a conductor. Whenever a magnetic field interlinked with an electric circuit is changed, an *electromotive force* (q.v.) is induced in that circuit. The magnetic field is considered to originate from the heart of the conductor, spreading out concentrically from the conductor as the current increases and shrinking back into the center of the conductor as the current decreases. In either case, the conductor is cut by the magnetic field, and in accordance with the *law of electromagnetic induction* (q.v.) an electromotive force is induced in the circuit proportional in amount to the rate of change of the magnetic field and acting in a direction which would, by producing a current, tend to oppose the change.

A bothersome type of electromagnetic induction sometimes occurring in radio takes place when a radio aerial is run parallel to a high-tension line or even a trolley line. The varying field of force about the high-tension line sets up, by induction, a corresponding field in the parallel aerial, thus causing a disturbing hum in the loud speaker. Such a condition can usually be eliminated by running the aerial at right angles to the power line. By doing this, induction between the two conductors is reduced to a minimum.

Electrostatic induction is a rearrangement of an electrostatic field whereby a body in the neighborhood of a charged body will have an opposite charge induced on it. An *induction machine* (q.v.) known as the Wimshurst utilizes the principle of electrostatic induction.

INDUCTION BRIDGE—Also called inductance bridge. A balance arranged similarly to a *Wheatstone Bridge* (q.v.), used in induction measurement. (See *Induction, Measurement of Inductance,* also *Induction Density.*)

INDUCTION COIL—A form of transformer, having a primary winding arranged with a means of making the direct current applied to it, pulsate. Essentially a transformer with open magnetic circuit, carrying a pulsating direct current which induces a high voltage, alternating current in the secondary by means of the step-up effect of the windings. The illustration shows a typical form of induction coil and the associated circuits for producing high voltages. Here, there are two windings, a primary and a secondary, over and insulated from a core, composed of a number of soft iron wires. The primary is arranged with a key, K, and a battery, B, to supply direct current to the circuit. When the key is depressed the current flows through the primary circuit, including the contact C and the vibrator V. The core becomes *saturated* (q.v.) or magnetized and attracts the armature or spring V. As this spring is drawn toward the core it breaks the contact between the spring and the contact post. This, of course, breaks the circuit from the bat-

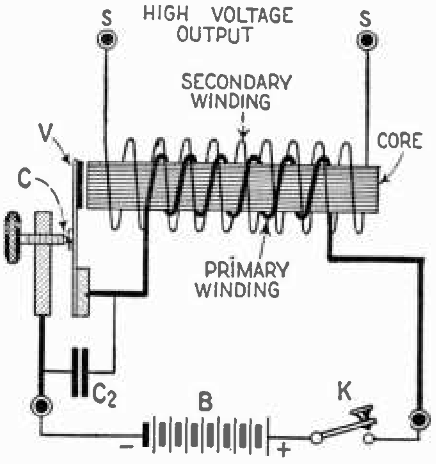

A typical form of induction coil showing diagram of connections, vibrator, core, battery, condenser and key.

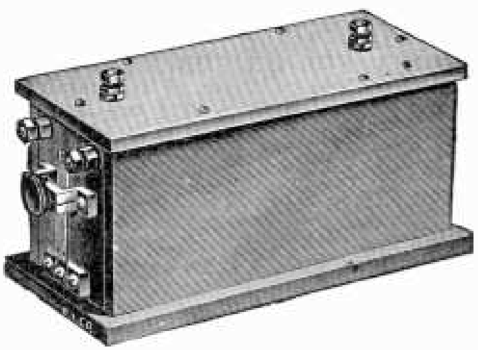
A form of spark coil.

tery and the flow of current stops, resulting in the collapse of the flux through the core. When the magnetic attraction ceases the spring goes back into normal position and the contact is once more complete, allowing current to flow through the windings. This make-and-break process is more or less rapid and is repeated continuously. By induction, an alternating E.M.F. (electromotive force) is created in the secondary windings, and as these windings have many more turns than the primary winding the induced E.M.F. will be proportionately greater. If the breaks occur at a rate of from thirty to one hundred per second, the key can be used to interrupt the primary current and produce at the secondary terminals a high voltage suitable for spark transmission. (See *Induction, Transformer, Spark Coil.*)

INDUCTION COIL, PRIMARY — The simplest form of induction coil. It consists of an iron core wound with a relatively large amount of low resistance copper wire, which is well insulated. The iron core increases the effect of self-induction and the size or length of the core and the number of turns of wire determine the efficiency of the coil. Such coils are used extensively for low tension ignition. (See *Induction Coil.*)

INDUCTION COIL, SECONDARY — A type of induction coil employing a primary and secondary winding. The primary is wound over an iron core and the secondary is generally wound over the primary. (See *Induction Coil.*)

INDUCTION DENSITY—A term occasionally used to refer to the density of flux in a magnetic or electromagnetic field of force. It is synonymous with flux density, under which heading it will be found more fully defined. (See *Induction, Electromagnetic,* also *Gauss,* and *Induction, Magnetic.*)

INDUCTION, ELECTROMAGNETIC — The production or *inducing* of electric currents in a conductor when it is moved in a magnetic field of force in such manner as to cut magnetic lines of force. The simplest example of this phenomena is the Faraday experiment in which he discovered that a wire with its ends joined, when moved rapidly in the field of a magnet, would result in current being induced in the wire or conductor. All dynamo electric machinery is based upon this principle of electromagnetic induction, as are also alternating current transformers and various other electrical devices.

In the illustration is shown the theory of production of current by electromagnetic induction. Fig. 1 shows a coil of wire placed in a uniform magnetic field of force between two magnets. Here the coil is assumed to have been moved from A to B in the same plane. This is known as a *motion of translation,* and the field being uniform at all points the same number of lines pass through the coil at both points. Therefore no current is induced in the coil and a sensitive instrument placed across the leads, C, C, would show no current present. In Fig. 2 the same coil is shown at A in its original position. Here, however, it is rotated, and when in position B it is apparent that less lines of force pass through it. We have thus fulfilled the conditions for inducing current, and if the instrument is placed across the leads, C, C, the presence of electric current will be indicated. If the motion of rotation is continued the number of lines of force passing through the coil will continue to increase and decrease and current will be produced in the coil as long as the motion persists. From the above it will appear that current can be induced either by rotating the coil around any axis in its plane, or by tilting it in its motion across the field. The coil of wire in this case is known as an *inductor* because the current is induced in it. It is also referred to as a *conductor* when describing the production of alternating currents.

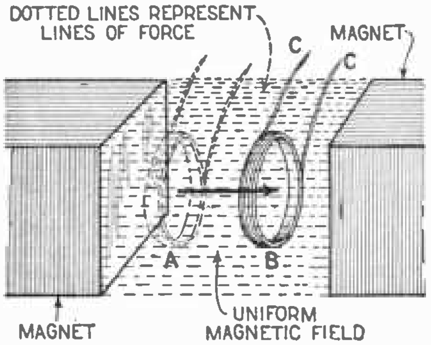

Fig. 1. Coil of wire in a uniform field moving parallel to direction of the field.

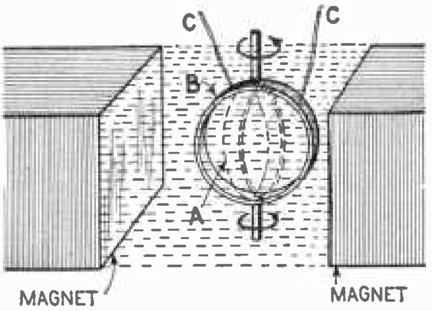

Fig. 2. Same coil, showing cutting of field due to the rotation of the coil.

Now it was stated that current is induced in the inductor whether the lines of force through it are *increased* or *decreased* by its motion through the field. We can reduce the action to two simple rules. First, the relative motion between an inductor and a magnetic field must be such that the number of lines passing through the inductor are altered, i.e., increased or decreased. Second, the electromotive force induced in the inductor circuit will be proportional to the rate of increase or decrease in the number of magnetic lines embraced by the inductor circuit. We know from the *Lenz Law* (see *Alternating Current, Theory of Production of*) that current will flow in the inductor in a direction dependent upon whether the lines of force through it are being increased or decreased, and as the inductor is being rotated continuously the current will change direction with each change in the number of lines. (See *Alternating Current, Alternator, Electromotive Force, Fleming's Rule, Laws of Electromagnetic Induction, Lenz's Law.*)

INDUCTION, ELECTROSTATIC — The production of an electric charge in a body due to the presence of a nearby conductor having an opposite charge. The effect is known as a displacement of electric charge. (See *Electrostatic Induction,* also *Induction.*)

INDUCTION INSTRUMENTS—Measuring instruments used in electrical and radio practice, which make use of the principle of electromagnetic induction. Such instruments are used for measuring alternating currents, the torque or actuating force for the indicator being produced by the effect of a rotating field on a metal disc or drum which is not connected to the circuit carrying the currents. The instruments operate on the same basic principle as the *induction motor* (q.v.).

INDUCTION MACHINE — A machine wherein the primary and secondary windings rotate with respect to each other. Examples of induction machines are induction motors, certain types of frequency converters and induction generators. An entirely different kind of induction machine is one which generates low potentials, utilizing the principle of *electrostatic induction* (q.v.). The most common form of this type is known as the Wimshurst (q.v.).

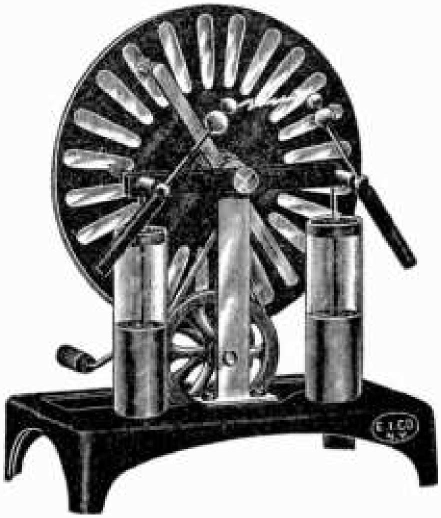

A common type of induction machine known as the Wimshurst.

INDUCTION, MAGNETIC — The communication of magnetism to a metal by the presence of a magnet without any actual contact. The effect of the presence of a magnet in the neighborhood of a piece of magnetic material is to induce magnetic poles in it. (See *Magnetism, Magnetic Poles, Induction,* also *Induction, Electromagnetic.*)

INDUCTION MOTOR — An alternating current motor in which the input current is passed through the field coils only, the armature not being connected to the external circuit. Here the armature is rotated by currents induced by the varying field set up through the field coils. The induction motor is the most widely used of the alternating current types, maintaining a practically constant speed from no load to full load and being simple and rugged in construction. This type of motor has two windings, the primary and secondary, and two members called stator and rotor. The primary winding is usually on the stator and the secondary winding, which is therefore customarily on the rotating member, is not connected electrically to the primary. Current to operate the rotor is induced in the secondary windings by the magnetic action of the current circulating in the primary windings. (See *Alternating Current,* also *Induction, Electromagnetic.*)

INDUCTION, MUTUAL—The interference or mutual effect between two electric or magnetic fields, due entirely to their proximity and without electrical contact. The mutual induction or electromagnetic influence of one circuit upon another is measured by the coefficient of mutual inductance or induction. (See *Mutual Induction,* also *Mutual Induction Coefficient,* and *Mutual Inductance.*)

INDUCTION REACTANCE—The value in ohms of the inductance in an electrical circuit as distinguished from the capacity or capacitive reactance. It is the part of the total impedance in an alternating current circuit that is due to the presence of inductance in the circuit. (See *Inductance Reactance, Ohm, Impedance.*)

INDUCTION REGULATOR—A system of regulating the voltage in alternating current circuits, employing a choke coil with removable core, or some form of variable ratio transformer which will assist or oppose the current by re-adjustment of the ratio. (See *Choke Coil, Reactance,* and *Variable Ratio Transformer.*)

INDUCTION SCREEN—A metal screen or shield placed between two electrified or magnetic bodies to reduce the effect of induction between them. A common example of this in radio usage is the placing of a sheet of metal between two coils, the sheet being generally grounded, or the provision of a metal *shield* (q.v.) entirely around each coil. This reduces the interaction of the fields of the coils, which leads to losses or howling, especially in *tuned radio frequency amplifiers* (q.v.). (See *Shielding.*)

INDUCTION UNIT—The *henry*, defined as the induction in a circuit when the electromotive force induced in that circuit is one *international volt* (see *volt*) while the inducing current varies at the rate of one ampere each second. (See *Henry*, also *Self-Induction.*)

INDUCTION VOLTMETER—See *Induction Wattmeter,* also *Induction Instruments.*

INDUCTION WATTMETER—An instrument for measuring the power delivered to a circuit. The induction type wattmeter operates on a principle used in the induction motor, a magnetic body attached to a pointer being rotated by a changing magnetic field. (See *Induction Motor,* also *Wattmeter.*)

INDUCTIVE CAPACITY—Generally referred to as the specific inductive capacity in rating the dielectric qualities of a substance. It is the quality of any dielectric substance by reason of which it permits electrostatic induction through it. The term is more or less synonymous with *Dielectric Constant*, the latter, however, being generally given in terms of a numerical value representing its comparative specific inductive capacity with some material taken as a standard. (See *Dielectric Coefficient and Constant,* also *Specific Inductive Capacity.*)

INDUCTIVE CIRCUIT—An electrical circuit possessing inductance but not capacity. (See *Inductance,* also *Capacity.*)

INDUCTIVE CONNECTION—A connection between two circuits entirely through the property of induction and without any metallic contact. The two coils, primary and secondary, of a vario-coupler as used in radio receiving circuits, are said to have inductive connection where the windings are separate. Thus, the primary coil will be in one electrical circuit and the secondary coil will receive energy from it by means of induction. (See *Induction, Coupling, Vario-Coupler.*)

INDUCTIVE COUPLER—The inductances or coils placed in inductive relation or inductively connected with each other for the purpose of transferring energy by electromagnetic coupling. Any device consisting of a primary and secondary inductance and used to transfer energy from one circuit to another by means of induction. (See *Coupler, Induction, Vario-Coupler.*)

INDUCTIVE COUPLING—The coupling or connection between two circuits by means of electromagnetic induction.

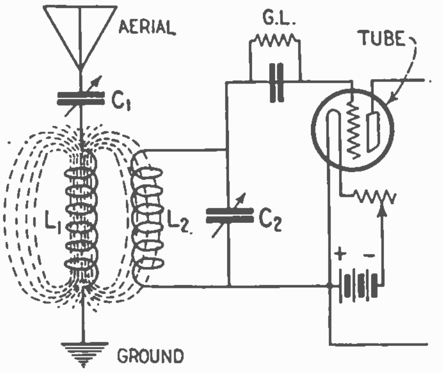

The utilization of a vario-coupler to inductively couple the aerial system with the rest of a receiving circuit.

The most common example of inductive coupling is in radio receiving circuits where coupling coils are used to associate the primary or aerial circuit with the secondary circuit. In the illustration L_1 is the primary of a vario-coupler and L_2 the secondary. The primary is inserted between the ground and aerial and the secondary is connected to the grid and filament of the detector tube. The variable condenser C_1 tunes the primary circuit, the primary coil also being variable by means of taps if desired. The variable condensed C_2 tunes the secondary inductance and the condenser and grid leak GL are the customary elements in the detector grid circuit. As the high frequency currents traverse the primary circuit, passing through coil L_1, lines of force are set up as shown and some of these lines embrace the secondary coil L_2, inducing similar currents in the secondary or vacuum tube circuit. The balance of the circuit is not shown for obvious reasons. This type of coupling between circuits is much more selective than the straight coupling or conductive coupling. (See *Coupling,* also *Conductive Coupling.*)

INDUCTIVE DISTURBANCE—The disturbing effects sometimes experienced in radio broadcast or telegraph reception due to induction from nearby power or telephone lines. This sometimes occurs when an aerial is strung close to and parallel with a line carrying high voltages, and occasionally is experienced with telephone wires near the receiver. (See *Induction.*)

INDUCTIVE DROP IN VOLTAGE—The drop or diminution of voltage in an alternating current circuit due to the presence of inductance. Inductance in a circuit carrying alternating current tends by reason of the self-induction to set up a counter-electromotive force which opposes the original E.M.F. (electromotive force), thus reducing the effective voltage in the line. (See *Inductance,* also *Inductance Reactance.*)

INDUCTIVE EMF—The electromotive force due to induction. That is to say, the electromotive force induced in a circuit. (See *Induction, Induced EMF.*)

INDUCTIVE REACTANCE—The reactance or portion of the total impedance in a circuit, due to the presence of self-inductance. (See *Inductance Reactance,* also *Impedance.*)

INDUCTIVE RESISTANCE—A resistance element possessing inductance, as for example a resistance coil, which also has the property of self-inductance. (See *Resistance,* also *Inductance.*)

INDUCTIVE RISE—The rise in voltage noted in transformers or other alternating current apparatus due to the presence of a *leading current* (q.v.). (See *Inductance,* also *Transformer* and *Angle of Lead or Lag.*)

INDUCTIVITY—Another term used alternately with dielectric constant and conductive capacity, referring to the dielectric properties of materials. (See *Dielectric Coefficient or Constant,* also *Inductive Capacity.*)

INDUCTOR—The coils or conductors in which current is induced through electromagnetic induction, as in an induction motor. Any conductor in which current is induced due to a change in magnetic flux may be called an inductor. (See *Induction, Inductor Alternator,* and *Induction Motor.*)

INDUCTOR ALTERNATOR—An alternator for producing high frequency currents. It employs the *armature* and *field windings* on projections inside the *stator*, the rotor consisting of a drum carrying a magnetic material. By using a *rotor* of solid steel, very high speeds and consequently high frequency currents are obtainable. The *Alexanderson Alternator* (q.v.) is an inductor type alternator. (See *Alternator,* also *Inductor.*)

INDUCTOMETER—A device or instrument used to measure self or mutual inductance. The basic principle depends on the relation of two coils, a primary and secondary, one of which can be moved in its relation to the other, the inductance, either self or mutual, being registered on a scale calibrated in units of inductance. (See *Inductance, Mutual,* also *Inductance, Self.*)

INERTIA—In physics, the property of a body which tends to keep it in a state of rest or to resist any change of motion. (See *Inertia, Electric.*)

INERTIA, ELECTRIC—Term occasionally applied in place of *self-inductance* (q.v.). When a current flows in a circuit possessing inductance and the circuit is broken, current continues to flow for a short time. It thus appears that in this instance the resistance alone will not stop the flow of current on the instant of breaking the circuit, an interval of time being required as in the act of bringing to rest a moving material. (See *Inertia, Inductance, Self,* also *Resistance.*)

INERTIA, ELECTROMAGNETIC—The same as electric inertia, being, in effect, the energy required to start or stop a a current in a circuit possessing self-inductance. (See *Inertia.*)

INFERRED ZERO—A term used in connection with certain instruments having extremely high sensitivity, where the zero is removed from the scale. In a galvanometer, for example, the zero position on the scale may be merely assumed, an electrical or mechanical force being applied to bring the zero off the scale entirely, only a part of the full range of the full scale being utilized. (See *Galvanometer.*)

INFLUENCE—A broad term signifying action at a distance, as by electrostatic induction, without any actual physical contact. It is considered as the effect of a charged body on a neutral body or conductor coming within its field. (See *Field*, also *Induction.*)

INFRA-RED RAYS—The heat waves or rays which lie between light waves and electromagnetic (radio) waves. The infra-red lies at the extreme low frequency end of the invisible spectrum, the wave-length being extremely high and the frequency conversely low. (See *Wave*, also *Ether Waves.*)

INFRA-SATURATION PART OF CHARACTERISTIC—In a characteristic curve of a vacuum tube, this is the portion over which the tube is operable, or the operating portion of the curve. (See *Characteristic Curve*, also *Vacuum Tube, Theory of Operation.*)

INITIAL MAGNETIZATION—The initial stages of magnetization as distinguished from the effect after saturation has been approached or reached. (See *Saturation, Magnetic.*)

INPUT—Generally speaking, the energy absorbed by a machine as distinguished from the output energy. (See *Input Voltage*, etc.)

INPUT CIRCUIT OF VACUUM TUBE—Vacuum tube circuits are generally divided into three separate classes or circuits. The first is the circuit including the filament, i.e., the "A" battery circuit through the filament. The second is the circuit between filament and grid and the third is the circuit between filament and plate. The filament grid circuit is called the input circuit because the incoming oscillations are impressed on this circuit. The filament plate circuit is known as the output because the rectified or amplified currents are taken from across the plate and filament. The illustration shows both input and output circuits, the former in heavy lines and the latter in light lines. The input circuit of a tube is also known as the *grid circuit.* (See *Grid, Filament, Plate*, also *Vacuum Tube, Theory of Operation.*)

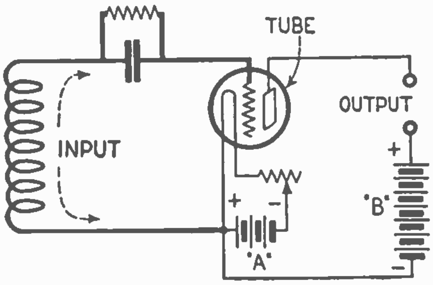

Input and output circuits of a vacuum tube receiver. Heavy lines show the former, light lines the latter.

INPUT VOLTAGE—The voltage impressed on a circuit or machine as distinguished from the output voltage. (See *Input.*)

INSTANTANEOUS CURRENT—The value of current in an alternating current circuit measured at any instant, as distinguished from the average current value. (See *Instantaneous Values.*)

INSTANTANEOUS VALUES—The actual value of an alternating current, E.M.F. (electromotive force), etc., measured at any instant. This value may be anything from zero to maximum or peak It must be remembered that in an alternating current the current and voltage are both variable factors, rising to a maximum in one direction, then falling to a minimum, and then rising to a maximum in the other direction. Thus, a wattmeter measures the actual value of current and voltage at each instant and the instantaneous torque at each instant is proportional to the instantaneous power. The average torque acting on the moving element during each cycle of the current and voltage is then proportional to the average power. If we measure an alternating current with an ammeter, and then measure the voltage with a volt-meter, the product of the two will be the apparent power, (apparent watts). It should be noted that these instruments measure effective and not instantaneous values. In order to ascertain the true *watts* (q.v.), it is necessary to multiply the apparent watts by the power factor. A wattmeter measures true watts. (See *Effective Electromotive Force, Power Factor*, also *Wattmeter.*)

INSTANTANEOUS VOLTAGE OR PRESSURE—The voltage or pressure in an alternating current circuit measured at any instant, as distinguished from the average value of E.M.F. (electromotive force). (See *Instantaneous Values.*)

INSTRUMENTS, MEASURING—In electricity or radio, any device or instrument for recording or indicating values, such as an ammeter for indicating the current in amperes, or a voltmeter to indicate the electromotive force in volts. (See *Ammeter* or *Ampere Meter, Voltmeter, Wattmeter.*)

INSTRUMENT SHUNTS—In electrical measuring instruments, a branch conductor, generally of relatively heavy wire or metal strip, joining the meter circuit at two points and forming a parallel circuit, thus dividing the current and protecting the delicate windings of the instrument. (See *Shunt*, also *Ammeter*, or *Ampere Meter, Hot Wire Ammeter.*)

INSULATE—To provide a protective covering or shield for a part or conductor, preventing flow of electricity from it or the passage of electricity through it. A conductor of electric currents is generally insulated to prevent the flow of the electricity through contact with other parts or conductors except where contact is intended to be made. (See *Insulated Wire.*)

INSULATED WIRE—Wire, generally copper, covered with insulating material to prevent leakage of the electric current. Standard insulated wire of the heavier gauges is usually covered with a layer of rubber and an outer covering of braided cotton. (See *Lead-In*, also *Insulation.*)

INSULATING COMPOUND—Liquid or easily liquifiable mixture of some insulating material, used in radio to coat cardboard tubes and to cover any open point of contact in an otherwise insulated conductor or part. Coils are very often coated with shellac or some other liquid and quick-drying substance. This serves to hold the windings in place and also acts as a protective covering. A very valuable application of an insulating liquid is shown in the illustration. Here an inductance is wound on a form of stiff paper or cardboard tube and after being completed is coated with one or two treatments of collodion or similar liquid. (Shellac introduces losses through the medium of *distributed capacity.*) The form is then cut away and the coil remains self-supporting as shown. This method produces a coil having the least possible dielectric and distributed capacity loss for a cylindrical shape. (See *Dielectric, Inductance Coil, Low Loss Coils, Collodion.*)

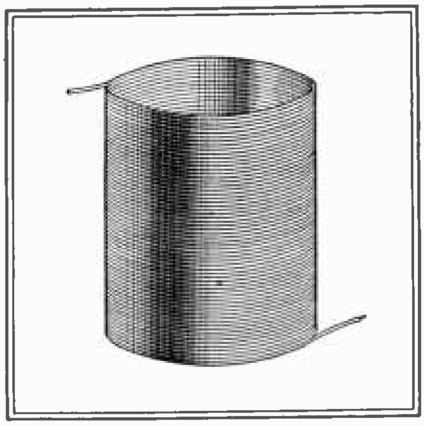

Air core coil made rigid by use of insulating compound.

INSULATING MATERIALS—Non-conducting substances, for example, glass, hard rubber or porcelain, which do not conduct electric currents. Strictly speaking, there are no perfect insulating materials. The most perfect insulators are certain gases, although at low pressure these may act as very good conductors. The more common insulating materials in radio are the following: mica, glass, bakelite and hard rubber, the insulating values being in that order. When we speak of the insulating property of a material we are actually referring to its resistance, a good insulator offering infinite resistance to the passage of electric currents and poor ones being relatively good conductors. The term *specific resistance* is used extensively in connection with the insulating properties of materials. This term is defined as the resistance of a piece of material of unit length and unit cross-sectional area at a given temperature, the value being given in ohms. (See *Dielectric Coefficient and Constant, Table of*, also *Ohms, Resistance* and *Insulate.*)

INSULATING PROPERTY—The ability of any material to *insulate* electric currents. Accepting mica as an insulator, we say that its insulating properties are superior to glass or hard rubber. (See *Insulating Materials*, also *Dielectric Coefficient and Constant* and *Insulation of Aerial.*)

INSULATION—Broadly any material or gas used to insulate electric currents. In radio, insulation plays a most important part, providing protective coverings or separators between conductors of the electric current and between the conductors and the ground. Insulation is really a material or gas having extremely high specific resistance; the better the insulating property, the higher the resistance to passage of electricity. There is no clearly defined line of demarcation between an insulator and a conductor, a poor insulator being to some extent a conductor. (See *Bakelite, Conductor, Dielectric Coefficient and Constant, Insulations of Aerial, Panel.*)

INSULATION OF AERIAL—The means used in connection with an aerial for reception or transmission of radio waves, to prevent leakage or ground-

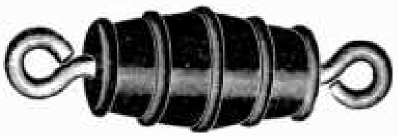

Aerial Insulator.

ing of the currents. The insulation of an aerial for reception purposes should be given very careful consideration. The wires may be suspended by means of small glass or porcelain insulators and the lead-in rubber covered wire, entering the building through a *lead-in insulator* (q.v.). The aerial for transmission must be accorded specially careful attention as far as insulation is concerned. The bare wires must be suspended with heavy duty insulators, capable of withstanding high voltages without *"breaking down."* The lead wire must be conducted into the building through a high-tension insulating tube or other lead-in device to prevent leakage of the current to the ground. (See *Aerial, Insulation of Aerial* and *High Tension Insulators,* also *Lead-In Insulator.*)

INSULATION RESISTANCE—The actual resistance in ohms, generally megohms, of the insulation of any conductor or piece of apparatus. If we assume two conductors, carrying separate currents and crossed so that the only protection would be the actual insulation of the wires, the insulation resistance would be the actual resistance in megohms from one path to the other. Conceivably a small current might leak from one conductor to the other, this current, of course, depending on the resistance of the insulation. The term "insulation resistance" is more easily understood when it is considered that an insulator or insulating material is merely a very poor conductor, or one having extremely high resistance to electric currents, there being actually no perfect insulator. (See *Insulation, Insulated Wire,* also *Conductor* and *Resistance.*)

INSULATION TESTER—An instrument or set of meters for testing the resistance of an insulator in fractional ohms, generally megohms. (See *Insulation Resistance.*)

INSULATOR—Any substance having insulating properties. *Insulators* are used to keep electric currents flowing in predetermined paths. It is necessary to use insulators on all electric circuits in order to prevent *grounds, leakage* and *short circuits.* Insulators are of the utmost importance in radio. At the transmitting end, where high voltages as well as high frequencies are used, the insulators must be specially designed to prevent leakage. A typical installation of insulators in use on the aerial system of a modern broadcasting station utilizes a 12-inch pyrex insulator at each end of each wire between the spreaders. A standard porcelain strain insulator is used every 14 feet on each guy wire. At each end of each guy wire a standard 24-inch porcelain insulator is used.

Insulators used on receiving antennas are made of porcelain, pyrex, rubber and composition. These are usually about 4 inches in length. Sometimes porcelain tubes are used as insulators to bring a *lead-in* (q.v.) into a building, but these have been generally superseded by flat *window lead-ins.*

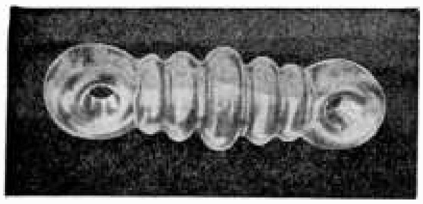

Illustration by courtesy of Corning Glass Works.
A typical small antenna insulator, showing corrugated surface.

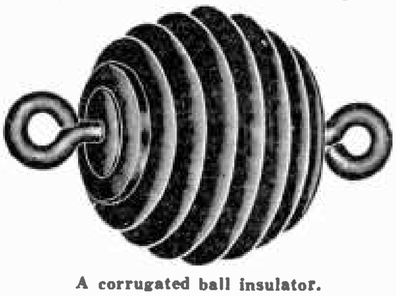

A corrugated ball insulator.

INSULATOR, LEAD IN—The insulating tube through which a lead wire from a radio aerial enters a building, used to prevent grounding or leakage of the current by contact with another conducting surface. (See *Lead-In Insulator.*)

A flat window lead-in which permits entrance of lead-in into building without the necessity of boring holes.

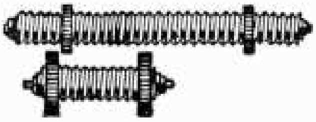

Lead-in insulators of conventional type.

INTEGRATING DETECTOR—The name sometimes given to any detector which yields a response proportional to the total energy received from a spark train rather than to the maximum value of the current or voltage in the train.

INTEGRATING INSTRUMENTS—Meters so designed as to record the total quantity of electricity (ampere hours) or the total energy (watt-hours) passing in a circuit in a given time. (See *Ammeter* or *Ampere Meter,* also *Integrating Wattmeter.*)

INTEGRATING WATTMETER—An instrument which measures and records the total amount of electrical energy being consumed in a circuit. These instruments are also known as *recording watthour meters* (q.v.), *integrating watthour meters,* and sometimes are simply referred to as *watthour meters* (q.v.).

Broadly speaking the *integrating wattmeter* is a type of small motor whose rotations are counted by means of a worm on the armature shaft engaging a set of cogs working a counter. The construction is such that the average *torque* (q.v.) exerted by the motor is proportional to the average power taken by the load. In other words, the motor rotates at a speed directly proportional to the energy being expended. For direct current work, *integrating wattmeters* are either of the commutator type, such as the Thompson, Westinghouse and Duncan meters or of the no-commutator type, such as the Sangamo. The commutator type instruments operate on the *dynamometer* (qv.) principle. Meters of the no-commutator type are essentially motors whose armatures always cut the same direction flux. That is to say, they are essentially *homopolar* (q.v.) motors. The principle of operation of these meters is the law of motor action which states that a conductor, free to move and carrying a current whose direction of flow is at right angles to a fixed magnetic field, will be moved out of this field, and in a direction at right angles both to the direction of current flow and to the field. For alternating current work, *integrating wattmeters* are of the *induction* (q.v.) type. The principle of operation of these meters is identical to that of *induction motors* (q.v) having shunt and series windings stationary and so related and located as to produce a rotating field acting upon a closed rotable secondary. In the induction meter the secondary consists of a light aluminum disc. These meters may be either of *single phase* (q.v.) or *polyphase* (q.v.) construction. (See *Electro-dynamometer.*)

INTEGRATOR—A device which automatically, by means of clockwork, adds up and records on a dial items of calculation or measurement. The system used in a recording or integrating wattmeter. (See *Integrating Wattmeter.*)

INTENSIFICATION—Broadly an intensifying or increasing of the density of electric current, but specifically in radio it refers to the tendency of radio signals to increase in intensity under certain conditions. The phenomena of intensification is naturally closely related to fading, as obviously the intensification process must necessarily follow or be followed by a fading process. The tendency toward intensification is more pronounced toward night and generally on the shorter wavelengths, it being noted that quite often signals from a distant station will be increased in volume or intensity as much as fifty per cent or more between the late afternoon and late evening. While the periodic *fading* of signals and the subsequent increase in volume over regular periods, usually of a few minutes' duration, may be considered as *fading* and *intensification,* it is preferable to think of fading as one operation together with the subsequent increase in volume, and the term *intensification* as referring strictly to the change to greater volume that takes place toward night. (See *Fading,* also *Phenomena of Electric Wave Propagation.*)

INTENSITY OF ELECTRIC CURRENT—A term more or less in use to imply the strength of an electric current in amperes, the symbol I for current being derived from it. (See *Ampere, Current Density,* also *Intensity of Field.*)

INTENSITY OF FIELD—The intensity or strength of the force exerted in a magnetic field. It is measured by its action or effect on a unit pole placed at any point in the field, the intensity of the field, of course, varying at different points. (See *Flux Density, Intensity of Magnetic Flux.*)

INTENSITY OF MAGNETIC FLUX—The strength of the force exerted in a magnetic field. (See *Intensity of Field,* also *Flux Density.*)

INTENSITY OF MAGNETIZATION—The density of force per unit cross-sectional area of a magnetized material. The extent of the magnetizing effect on a material placed in a magnetic field of force. This applies only to material in a magnetic circuit and should not be confused with the lines of force due to a field arising from the presence of neighboring currents or magnets. (See *Magnetization,* also *Intensity of Magnetic Flux.*)

INTERFERENCE—The interruption or interference with desired electromagnetic waves by undesired or extraneous waves. The interference of undesired broadcast or radio-telegraph signals with the desired signals. Also, broadly, the detrimental effect of power lines and other electric circuits on radio communication. Interference is encountered chiefly in receiving sets having poor *selectivity* (q.v.). In some instances interference may be due to the fact that another transmitter is operating on the desired wave-length, and being sufficiently close as to make it practically impossible to obtain clear signals from the desired station. Ordinarily, and where the effect is due to lack of sharp resonance or tuning, certain changes can be made in the receiver to sharpen the tuning, or if the disturbance is due to the transmitting stations and the receiver is normally sharp, various circuits can be employed to filter out the undesired signals. (See *Interference Preventer*, also *Wave Trap.*)

INTERFERENCE ELIMINATOR—A device or system as shown in the illustration for reducing or preventing reception of undesired signals in radio work. Practically all *interference eliminators* make use of a separately tuned inductance across the aerial and ground and in parallel with the receiving set. The inductance is usually tuned by means of a variable condenser which allows it to be tuned to the wave-length of the interfering signal. The interference can thus be by-passed to the ground. *Interference eliminators* of the kind described are generally known as *wave traps* (q.v.). A special form of interference due to atmospheric and electrical disturbances is known as *static* (q.v.). (See *Static Eliminator.*)

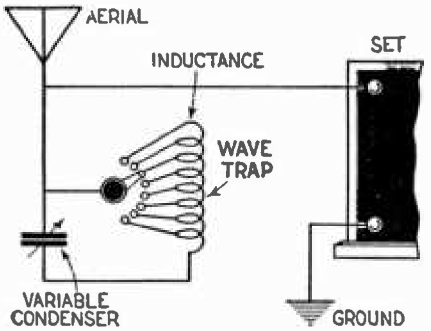

A wave trap or interference eliminator which dispenses with the ground connection.

INTERFERENCE PREVENTERS—Any arrangements either within the receiving set, or external to the set, for limiting or preventing interference. In addition to *interference eliminators* (q.v.) and static *eliminators* (q.v.) various other methods are utilized in radio to prevent interference. Thus the arrangement of a receiving set to give sharp tuning and great selectivity, results in the prevention of much interference. A method of preventing interference which has been used, is the utilization of filter circuits which can be tuned to pass any desired frequency and exclude all others. The use of the *loop aerial* (q.v.) results in great selectivity. The loop has a very pronounced directional effect and this in itself acts to narrow the field of stations offering possible interference. For example, if stations of very nearly the same wave length are in opposite or different directions and are sending at the same time, the loop permits selection of the station wanted, with no interference from the others.

INTERMEDIATE CIRCUIT—A closed circuit consisting of two inductance coils shunted by a variable condenser. One coil acts as primary to the secondary coil of the detector circuit and the other coil acts as secondary to the primary of the aerial circuit. A circuit such as this is sometimes referred to as a *multiple tuner intermediate* circuit and may be used as an interference eliminator.

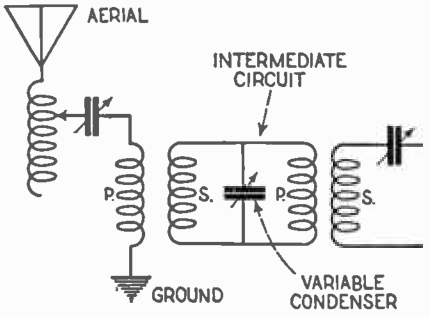

A multiple tuner intermediate circuit.

INTERMEDIATE FREQUENCY—This is a frequency higher in number of oscillations than *audio frequency* (q.v.) but lower than *radio frequency* (q.v.). It is generally around 30,000 cycles or 10,000 meters wave-length. In the *super-heterodyne* (q.v.) the incoming signal is converted from the broadcast frequency to the so-called *intermediate frequency* at which it can be amplified by means of fixed winding transformers to almost any extent desired. The method of converting the incoming signal in a *super-heterodyne* from radio frequency to *intermediate frequency* is by the *beat* method. In this case an oscillating circuit controlled by means of a variable condenser so as to obtain any desired frequency, is brought into interference with the first detector and produces a *beat note* (q.v.) for the *intermediate frequency* which has all the characteristics of the original signal. In this way, no matter what the frequency of the intercepted signal may be, the intermediate frequency amplification is always carried on at a fixed wave-length, thus making for high efficiency. (See *Beat Frequency; Beats.*)

INTERMEMEDIATE FREQUENCY AMPLIFIER—An arrangement of tubes and transformers for stepping up the current, used in the *super-heterodyne* (q.v.) circuit. The intermediate frequency amplifier is located between the first detector and the second detector. (See *Intermediate Frequency.*)

INTERMEDIATE TRANSFORMER—The name given to the transformers used in the *intermediate frequency amplifier* (q.v.) of the *super-heterodyne* circuit. *Intermediate transformers* are often referred to as *intermediate frequency transformers.* They are *radio frequency transformers* (q.v.) but have an iron core and instead of tuning to a maximum of 550 or 600 meters they are designed to cover wave-lengths of 10,000 meters or higher. (See *Intermediate Frequency.*)

INTERMITTENT CURRENT—A current which flows irregularly or which is interrupted at intervals, or without continuity. The current obtained from a *magneto generator* (q.v.) is an interrupted or *intermittent current.* Such a current may be either alternating or direct. A magneto generator constructed to give direct current is equipped with a two segment commutator having a stationary brush in contact with it. (See *Pulsating Current.*)

INTERNAL RESISTANCE—Resistance within an electric source. In the case of a primary cell such as a dry battery, the internal resistance causes an interference with the flow of current. This varies in amount with the construction and materials of the battery.

INTERNATIONAL MORSE CODE—This is also known as the *Continental code.* (See *Code.*)

INTERPOLE—A small magnetic pole placed between the main field poles of an electric generator for the purpose of obtaining better commutation. Interpole windings are connected in series with the armature winding and the load. The action of the interpole is exactly analogous to the shifting of the brushes, but when interpoles are used, brush shifting is dispensed with and the magnetic flux is shifted instead.

INTERRUPTED CONTINUOUS WAVES—abbreviation I.C.W.—These waves are obtained by the modulation at audio frequency, during signalling, of an otherwise *continuous wave* (q.v.). (See *Continuous Waves Key Modulated,* also *Continuous Waves Modulated at Audio Frequency.*)

INTERRUPTED WAVES—abbreviation I.W.—Interrupted waves are waves produced by modulation at audio frequency, of otherwise continuous waves. (See *Interrupted Continuous Waves.*)

INTERRUPTER—A combination of an electro-magnet with a vibrating armature which carries a contact point on a piece of spring steel fastened to the armature. When the electro-magnet attracts the armature the contact points are drawn apart thus interrupting the circuit. This action is only momentary for the spring again brings the circuit back to its original condition. The most important use of the interrupter in radio is as a *rectifying valve* (q.v.) in mechanical chargers. A typical vibrator panel is shown in the illustration. The armature, A, is the moving element. This carries at its outermost end a heavy Tungsten contact, B, which is caused to vibrate in synchronism with the supply current, by the actuating coil, C. When the current from the transformer is of the correct polarity and value the actuating coil attracts the armature, closing the charging circuit to the battery through the fixed contact, D. The contact end of the armature, A, vibrates within the air gap of the pole shoe, G.

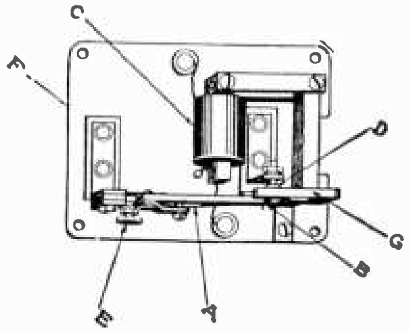

The vibrator panel of an efficient type of mechanical charger.

A common use of the *interrupter* is in the ordinary electric bell. In this apparatus, the electromagnet pulls the armature carrying the spring and contact point, breaking contact and opening the circuit. This immediately de-energizes the magnet since its coils are in series with the circuit. The spring tension pulls the armature back to its original position thus again closing the contact. As soon as the contact is closed, the electromagnet is again energized and it again attracts the armature, opening the circuit. This action keeps on indefinitely as long as a difference of potential is maintained at the terminals of the bell.

The *interrupter* is used in the *induction coil* (q.v.) for breaking up the direct current into a series of impulses, thus producing an intermittent current. (See *Circuit Breaker.*)

INTERRUPTER, ELECTROLYTIC—A jar containing diluted sulphuric acid as the electrolyte, a large sheet of lead as one electrode and a platinum needle point introduced through a glass tube as the other electrode. If these electrodes are connected through an inductance to a source of current supply, the current in the circuit will be rapidly interrupted. *Electrolytic interrupters* are suitable for radio transmitters using small spark coils and are also used in X-ray work. (See *Electrolytic Interrupter.*)

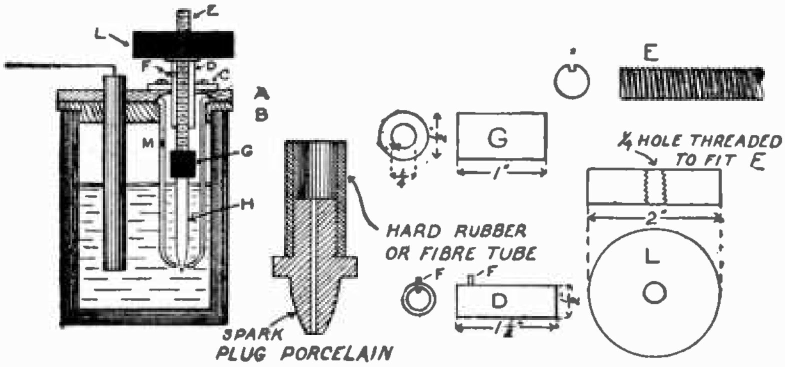

A simple type of Wehnelt interrupter. The positive pole of the direct current supply is connected to the brass part C, while the negative pole is attached to the lead rod. L is a fibre handle; E is a threaded rod, which is constructed to fit the threaded handle, but slotted so as to engage the pin F and prevent it from turning; H is a copper or platinum rod; D is a brass tube; M. is the overflow hole in the fibre tube, into the lower end of which a portion of an old spark plug porcelain part has been driven; finally C is a small brass block as shown. This is the adjustable type of Wehnelt interrupter, since the fibre handle can be operated to allow more or less surface to be exposed to the acidulated water.

INVERTED "L" AERIAL—A flat top aerial (antenna) in which the down lead is tapped off one end of its horizontal span. (See *Length of Aerials, Horizontal Aerial.*)

INVERSE DUPLEX CIRCUIT—A circuit in which the vacuum tubes are utilized both for radio frequency and for audio frequency amplification as in the *reflex* (q.v.) circuit with this important differentiation that the first radio frequency tube is utilized as the second audio frequency tube and the second radio frequency tube is used as the first audio frequency tube. In the ordinary reflex circuit the first radio frequency tube is also the first audio tube and the second radio frequency tube is the second audio tube. The *inverse duplex* circuit has the advantage over the reflex in that the tubes which are reflexed operate more efficiently since the relatively weak radio currents pass through the same tube as the stronger audio currents and vice-versa. The *inverse duplex circuit* is the invention of David Grimes.

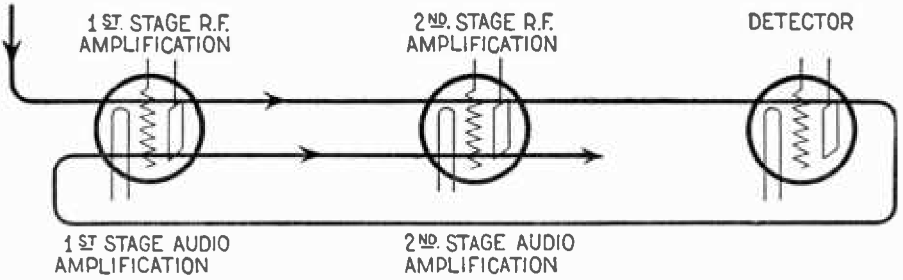
1ST. STAGE R.F. AMPLIFICATION
2ND. STAGE R.F. AMPLIFICATION
DETECTOR
1ST STAGE AUDIO AMPLIFICATION
2ND. STAGE AUDIO AMPLIFICATION

Diagram showing the current path and reflexing in an ordinary reflex circuit.

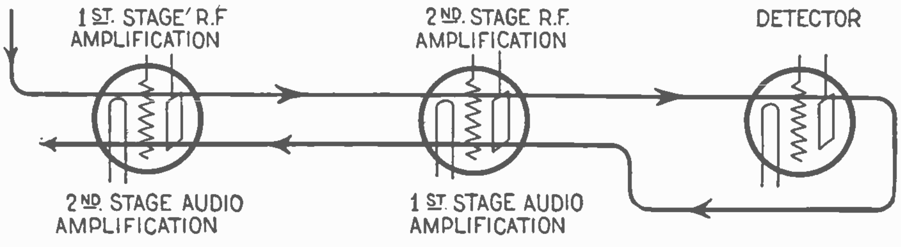
1ST. STAGE R.F AMPLIFICATION
2ND. STAGE R.F. AMPLIFICATION
DETECTOR
2ND. STAGE AUDIO AMPLIFICATION
1ST. STAGE AUDIO AMPLIFICATION

Diagram showing method of reflexing in an inverse duplex circuit.

ION—An atom of matter carrying an *electron* or an *atom* deprived of electrons. *Monad-ion* refers to a unit charge, *dyad-ion, divalent* carries two units, the *triad-ion* carries three unit charges. The positive *ion* is an *atom* minus an *electron*. A negative *ion* is an *atom* plus an *electron*. (See *Anion*, also *Cathion, Electron, Electron Theory.*)

IONIZATION—sometimes spelled **IONISATION**—The splitting up of *molecules* into *ions* (q.v.). *Ionization* may apply either to a chemical compound or to a gas. When a electric current is passed through an *electrolyte* (q.v.) the constituents of the electrolyte which are liberated or deposited at the *electrodes* (q.v.) are called *ions* and this process is called *ionization*. Gases at or near atmospheric pressure are normally good insulators. They are said to undergo *ionization* when they are made conductive, that is when they allow the flow of electricity. A gas may be made conductive by passing *X-rays* (q.v.) through it, by the action of ultra-violet light, by subjecting it to radium rays, by passing it close to a flame or hot piece of metal or by forming in it a brush discharge or *corona* (q.v.). Gases which have been made conductive carry current because of their dissociation into *ions*, which carry positive or negative charges of electricity. (See *Electrolysis, Electrolytic Action*, also *Valve Detectors.*)

I.R. DROP.—The voltage drop due to the resistance in a current-carrying conductor. It is directly proportional to the current flowing in the circuit multiplied by the resistance of that circuit.

I^2R LOSS—The power loss in any circuit due to the resistance offered to current flow. It is proportional to the square of the current flowing in the circuit multiplied by the resistance of the circuit. Whenever current flows there is an I^2R loss. This is measured in *watts* (q.v.) or *kilowatts* (q.v.).

IRON PYRITES—Formula FeS_2—This is a mineral, a disulphide of iron, found in large quantities. It was formerly used in radio as a crystal detector mineral but has been almost entirely superseded by vacuum tube detectors. *Iron pyrites* is flaked with shiny spots which have the appearance of gold, and for this reason it is sometimes called "fools gold." *Iron pyrites* is very hard and brittle. The present commercial use of *iron pyrites* is in the production of sulphuric acid. Ferrous Sulphide, FeS, has also been employed in the past as a radio detector. This substance is chemically formed by fusing equivalent quantities of sulphur and iron together. The resultant substance has a smooth surface. (See *Crystal Detector.*)

ISOCHRONOUS—A term applied to two or more oscillatory circuits, meaning that they have the same *natural frequency* (q.v.). Stated in another way, two or more radio frequency circuits are isochronous when they are in electrical *resonance* (q.v.). Circuits are said to be in resonance when they have the same *oscillation constant* (q.v.). The *oscillation constant* of a circuit is equal to the square root of the product of the inductance multiplied by the capacity of that circuit. Circuits having equal oscillation constants will have the same discharge frequency.

ISODYNAMIC LINES—Lines on a magnetic map connecting all parts of the earth where the magnetic intensity is the same. In other words, these lines pass through points of equal horizontal component of the earths magnetic field.

ISOGONIC LINES—Lines on a magnetic map passing through points of equal *declination*. Lines connecting all points of equal variation of the magnetic meridian from the geographic meridian.

ISOTHERMS—Lines on a meteorological chart joining all points having the same temperature.

ISOTROPIC CONDUCTIVITY—Equal conductivity in every direction. A conductor is said to be *isotropic* when it offers the same resistance to the flow of an electric current in every direction through its mass.

J

JACK—An arrangement of spring terminals as shown in the illustrations which can be connected together or separated thus closing or opening circuits through the insertion of a *plug* which also forms a part of the circuit. The *plug* (q.v.) has a tip and a sleeve at one end (that is at the end inserted into the jack) and at the other end it has the two terminals for connecting the apparatus to be cut into the circuit. As an example, the terminals of the plug may be connected to a *loud speaker* (q.v.). The jack in this case is of the *open circuit* variety. That is to say, the springs of the jack are separated when not in use and the circuit is normally open. When the plug is inserted, the tip of the plug makes contact with one spring and the sleeve makes contact with the other thus closing the circuit through the *loud speaker*. In cases where the plug is used to open a circuit, that is to make the *jack* springs break contact with each other, the *jack* is called a *closed circuit jack* since its spring terminals are normally making contact. In cases where the jack is used not only to put the loud speaker in the circuit, but also to act as a filament switch, it is known as an *automatic filament control jack*. Jacks which are used to control a single circuit only, are referred to as *single circuit jacks*. Those controlling two or more circuits simultaneously are known as *multiple circuit jacks*. *Jacks* are usually mounted on a panel so that the plug can be inserted from the outside of the panel while the spring terminals are concealed behind the panel.

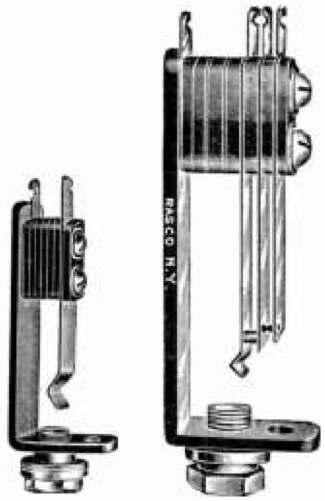

Illustration by courtesy of Radio Specialty Co.

Illustration at left shows one-spring open circuit jack. At the right is a three-spring jack.

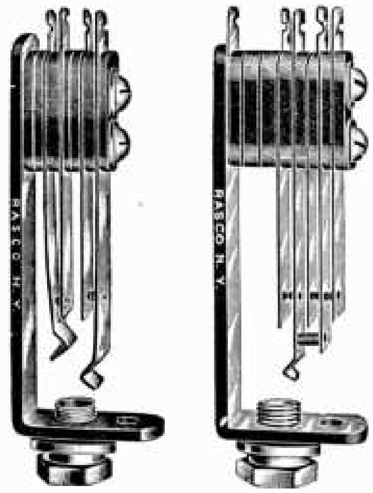

Illustration by courtesy of Radio Specialty Co.

The jack shown at the left is a four-spring double circuit jack. At the right is a five-spring automatic jack.

JAMMING—An expression denoting interference at a receiving station caused by a transmitting station not desired to be heard. More specifically, such interference, prevents the reception of signals, *jamming* the receiving station by drowning out with stronger signals or noises the station desired.

JANET, PAUL—Born in Paris, January 10, 1863. Educated at Lycee Louis-le-Grand and at the high school. He was Professor of Physics at the University of Paris and University of Grenoble. Professor Janet was the author of several important works, and first to make the experiments in electrical resonance successful. The results of these experiments are applied in present day wave meters.

JAR—A unit of electrostatic capacity. It is mainly used in the British Naval service. One *jar* equals 1000 *centimeters* which equals .0011 *microfarads* (q.v.).

JIGGER—A term generally accepted to denote an *oscillation transformer* (q.v.). Used for transforming trains of oscillations from one circuit to another. Dr. J. Erskine Murray refers to high frequency oscillations as *Jigs*. *Jiggers* are also known as *magnetic coupling transformers* (q.v.). They are also sometimes referred to as *auto-jiggers*.

JOULE—symbol J—The practical unit of electrical energy. It is equal to 10^7 *ergs* (q.v.). The *joule* is also the unit of electrical work. It is equal to the work done or the heat generated by a *watt second* (q.v.); or an *ampere* (q.v.) flowing for a second through a resistance of one *ohm* (q.v.). Expressed another way the *joule* is the work done by one *coulomb* (q.v.) flowing under the pressure of one *volt* (q.v.). One *joule* is equal to .73732 foot pounds, or .24 calories (q.v.).

JOULE EFFECT, or JOULEAN HEATTING—This occurs when an electric current traverses a resisting conductor. The rate of generation of heat in the conductor by a steady current equals the square of the current multiplied by the resistance.

JOULES EQUIVALENT—The amount of energy equal to a heat unit.

JOULES LAW—The heat produced in an electric circuit is directly proportional to the square of the current, to the resistance of the conductor and to the time of current flow.

K

KANALSTRAHLEN—A German word meaning Rays of Canal. It has been found that by perforating the cathode of a tube producing *cathode rays* (q.v.), faint luminous streaks will come through these perforations in a direction opposite to that of the *cathode* rays. These are called *kanalstrahlen*. They communicate a positive charge to an insulated conductor. In terms of the *electron theory* (q.v.) they are positively charged *ions* (q.v.) The cathode particle is an electron traveling in one direction, and the kanalstrahlen particle traveling in the opposite direction is what remains of the atom which has lost an electron.

KATHODE—This is a negative *electrode* (q.v.). (See *Cathode*, also *Electrolysis*.)

KATHODE OF CELL—The positive pole of a cell. This is indicated by the plus (+) sign. (See *cell*.)

KATHODE RAYS—The stream of *electrons* (q.v.) thrown off by the *kathode* of a vacuum tube. These produce a glow when they strike the walls forming the tube. They can be deflected by a magnet. (See *Cathode Rays*.)

KATION—The charged particles or *ions* (q.v.) moving in the direction of the *cathode* (q.v.). (See *Cathion*.)

KEEPER—The piece of iron used to close a magnetic circuit to protect it from external disturbances. The term *keeper* is generally used to refer to such a piece of iron when used with a permanent magnet such as a horseshoe magnet. In the case of the piece of iron which closes the magnetic circuit of an electromagnet, this is more often referred to as an *armature* (q.v.). The *keeper* may be used with a straight bar magnet also. In this case, two bar magnets are laid side by side with unlike poles adjacent and two keepers are used, one at each end. The use of a keeper avoids the demagnetizing effect of *leakage lines* (q.v.). (See *Horseshoe Magnet*.)

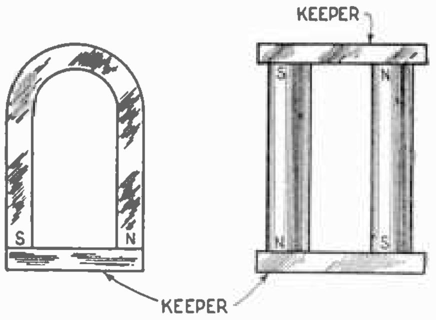

At left is shown keeper used with a horseshoe magnet. At right two keepers are shown used with two bar magnets.

KELVIN, LORD (William Thomson) was born at Belfast in 1824; he entered Glasgow University as a student at 10 years of age; graduated at Cambridge in 1845. When only 22 years of age he was called to occupy the chair of natural philosophy in the University of Glasgow, a chair which he adorned by his genius for fifty years. He was for 40 years or more regarded as the acknowledged leader of British science on its physical and mathematical side. His great inventions in telegraphy, his magnetic compass and sea sounding apparatus brought him fame and fortune. He was knighted in 1866; and in 1892 was raised to the peerage as Baron Kelvin of Largs; he died on December 17, 1907.

Lord Kelvin (William Thomson)

KELVIN'S BRIDGE—A method of measuring low resistance in which the voltage drops, produced by the same current in the resistance under test and in a standard low resistance slide wire are balanced against each other. This *bridge* is a modification of the *Wheatstone Bridge* (q.v.) and was designed by Lord Kelvin to eliminate the errors introduced when measuring resistances much less than one *ohm* (q.v.). In the case of the *Wheatstone Bridge,* such resistances could not be measured accurately on account of the errors produced by the terminal and contact resistances.

KELVIN'S ELECTROSTATIC VOLTMETER—A voltmeter used for measuring high, and in some cases low, alternating current voltages. It is constructed on the principle of an air condenser. One type of electrostatic voltmeter is a high potential instrument having the needle made of a thin aluminum plate suspended vertically on delicate knife edges, with a pointer extending from the upper part to a scale. Two quadrant plates, metallically connected together, are placed on either side of the needle and parallel to its face. These serve as one terminal of the circuit to be measured. The needle acts as the opposite terminal. When there is a difference of potential between the needle and the plates, the needle is deflected out of its neutral position. The value of the scale indications can be changed by hanging calibrated weights on the bottom of the needle.

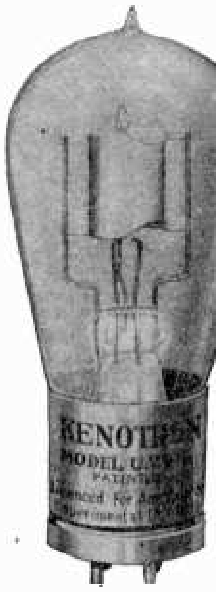

Illustration by courtesy of Radio Corp. of America
A kenotron tube.

KENOTRON—A type of *vacuum tube rectifier* (q.v.) in which the vacuum is extremely high and the discharge is carried almost entirely by electrons, not by gas ions. In these tubes, the plate current is always less than that actually emitted by the hot filament. *Kenotrons* are made sufficiently large to rectify several kilowatts.

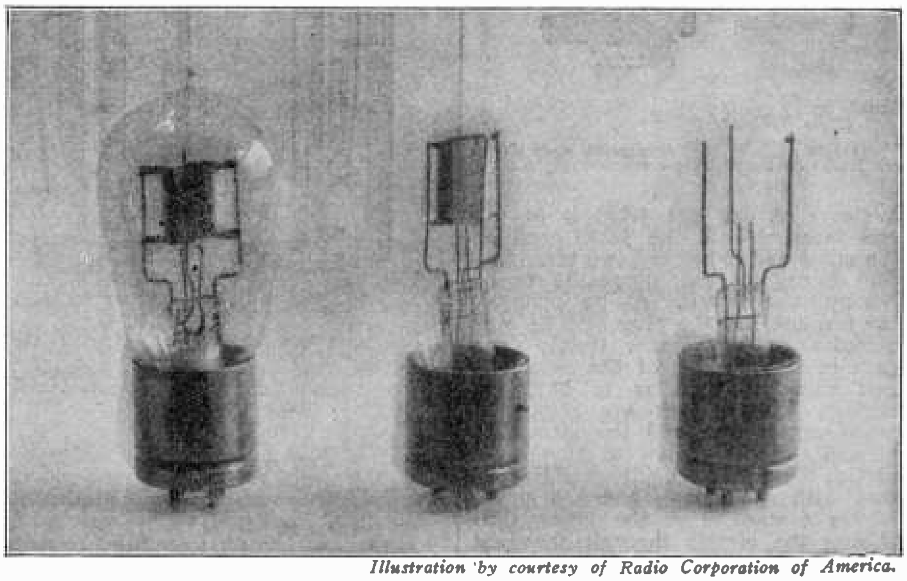

Illustration by courtesy of Radio Corporation of America.
Stages in the construction of a kenotron.

KENNELLY, A. E.—Anglo-American radio expert. Born at Colaba, Bombay, December 17, 1861, he was educated in England, Belgium, France and Italy. In 1875 he became a telegraph operator in the employ of the Eastern Telegraph Company, and in 1886 became the principal electrical assistant to Thomas Alva Edison in the laboratories at Orange, N. J., a post he held until 1892.

He was engineer-in-chief with E. J. Houston, of the Thomson-Houston Company, for the laying of the cables from Vera Cruz to Campeche, 1902. Since 1902 he has been Professor of Electrical Engineering at Harvard University, and since 1914 professor of the same subject at Massachusetts Institute of Technology.

Kennelly has written a large number of books on electricity and radio, and is the author of one of the standard elementary text-books on same. He has written a very large number of papers on radio, and he is an authority on alternating currents.

He is past president of the American Institute of Electrical Engineers, was president in 1916 of the Institute of Radio Engineers, and vice-president of the International Electrical Congresses, held in Paris and Turin. He is a member of many scientific societies, and has received many honorary degrees. In 1921 he was appointed a delegate to the Interallied Radio Technical Committee in Paris.

KEY—A form of switch for conveniently and quickly opening and closing a transmitting circuit, in the act of transmitting signals. An *operating,* or *telegraph,* so-called *Morse key* is a form of *tapping key* designed especially for the rapid sending of signals by the dot and dash code. A tapping key has one contact carried by a springy strip of metal. In signalling, this key is depressed by the fingers, thus closing the circuit by bringing the contact together with a fixed contact. A standard type of *transmitting key* (q.v.) such as used in sending wireless messages is shown in the illustration. These keys must necessarily be made with larger contacts and of more solid construction than those used to interrupt the smaller currents used in wire telegraphy.

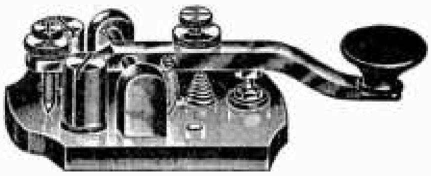

A standard transmitting key.

KEY, ANTI-CAPACITY—This is in reality a form of switch. (See *Switch.*)

KEY, HIGH-SPEED—A transmitting key usually operated from side to side instead of downwards, which makes

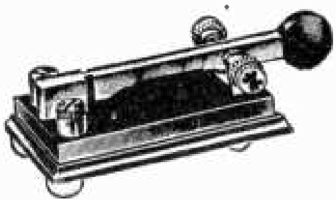

High speed key.

contact both on the right and on the left side. Such a key requires only half the travel of the ordinary transmitting key and thus facilitates high-speed manual transmission of code signals.

KEY SWITCH—A type of switch adapted for operation by a removable portion in the form of a handle or key. In radio work, *key switches* are sometimes used to lock a receiving set so that it cannot be operated unless the filament circuit is closed by the insertion and turning of the key. Key switches are generally operated by a hollow slotted key, although there are other types on the market. The slot in the key fits a projection or pin on the spindle or shaft of the switch which carries the movable contact. When the key is inserted in the socket and turned, it turns the shaft and brings the movable contact against a fixed contact. *Key switches* are also used for turning on or off electric lights and for locking or unlocking the ignition systems of automobiles.

KILO—Symbol K—A prefix placed before the name of a unit to indicate the multiple 1000 or 10^3. In cases where standard units are too small for convenient use, it is customary to utilize this prefix as, for example, in stating the voltage of a high tension transmission line. This would often be given as 10 kilovolts (K.V.) rather than as 10,000 volts. In like manner, in referring to large amounts of power, it would be preferable to give the amounts in kilowatts rather than in watts. Thus the better way to express 10,000,000 watts would be to divide by 1,000, thus converting to 10,000 kilowatts.

KILOAMPERE-TURN—One kiloampere-turn is equivalent to 1,000 ampere-turns. In referring to the magnetic circuits of electrical machinery it is usual to specify the magnetomotive-forces in kiloampere-turns rather than in ampere-turns. (See *ampere-turns*, also *magnetomotive-force*.)

KILOCYCLE—1,000 *cycles*—Two immediately succeeding half waves of an alternating current constitute a *cycle* (q.v.). The number of *cycles* per second is called the *frequency* (q.v.). In house-lighting and similar circuits, where a comparatively low frequency is used, it is usual to refer to the *frequency* in cycles per second. However, in radio circuits, where currents as high as 300,000,000 cycles per second are used, it is preferable to divide by 1,000, thus converting to *kilocycles*. If the wave-length in meters is known, it is a simple matter to determine the frequency in *kilocycles*. This is calculated as follows: The speed of light is 186,000 miles per second and experiments have shown that this is practically the same as the velocity of electromagnetic or radio waves. Expressed in meters this is equal to 300,000,000 (appx.) meters per second. Assuming that a transmitting station is using a wave-length of 500 meters, the frequency will be obtained in cycles per second by dividing the velocity of the waves by the

KILOCYCLES TO METERS, OR METERS TO KILOCYCLES

Meters	Kilocycles	Meters	Kilocycles	Meters	Kilocycles	Meters	Kilocycles	Meters	Kilocycles	Meters	Kilocycles	Meters	Kilocycles
10	29980	720	416.4	1430	209.7	2140	140.1	2850	105.2	4120	72.77	6400	46.85
20	14990	730	410.7	1440	208.2	2150	139.5	2860	104.8	4140	72.42	6450	46.48
30	9994	740	405.2	1450	206.8	2160	138.8	2870	104.5	4160	72.07	6500	46.13
40	7496	750	399.8	1460	205.4	2170	138.1	2880	104.1	4180	71.73	6550	45.77
50	5996	760	394.5	1470	204.0	2180	137.5	2890	103.7	4200	71.39	6600	45.43
60	4997	770	389.4	1480	202.6	2190	136.9	2900	103.4	4220	71.05	6650	45.09
70	4283	780	384.4	1490	201.2	2200	136.3	2910	103.0	4240	70.71	6700	44.75
80	3748	790	379.5	1500	199.9	2210	135.7	2920	102.7	4260	70.38	6750	44.42
90	3331	800	374.8	1510	198.6	2220	135.1	2930	102.3	4280	70.05	6800	55.09
100	2998	810	370.2	1520	197.2	2230	134.4	2940	102.0	4300	69.73	6850	43.77
110	2726	820	365.6	1530	196.0	2240	133.8	2950	101.6	4320	69.40	6900	43.45
120	2499	830	361.2	1540	194.7	2250	133.3	2960	101.3	4340	69.08	6950	43.14
130	2306	840	356.9	1550	193.4	2260	132.7	2970	100.9	4360	68.77	7000	42.83
140	2142	850	362.7	1560	192.2	2270	132.1	2980	100.6	4380	68.45	7050	42.53
150	1999	860	348.6	1570	191.0	2280	131.5	2990	100.3	4400	68.14	7100	42.23
160	1874	870	344.6	1580	189.9	2290	130.9	3000	99.94	4420	67.83	7150	41.93
170	1764	880	340.7	1590	188.6	2300	130.4	3020	99.28	4440	67.53	7200	41.64
180	1666	890	336.9	1600	187.4	2310	129.8	3040	98.62	4460	67.22	7250	41.35
190	1578	900	333.1	1610	186.2	2320	129.2	3060	97.98	4480	66.91	7300	41.07
200	1499	910	329.5	1620	185.1	2330	128.7	3080	97.34	4500	66.63	7350	40.79
210	1428	920	325.9	1630	183.9	2340	128.1	3100	96.72	4520	66.33	7400	40.52
220	1363	930	322.4	1640	182.8	2350	127.6	3120	96.10	4540	66.04	7450	40.24
230	1304	940	319.0	1650	181.7	2360	127.0	3140	95.48	4560	65.75	7500	39.98
240	1249	950	315.6	1660	180.6	2370	126.5	3160	94.88	4580	65.46	7550	39.71
250	1199	960	312.3	1670	179.5	2380	126.0	3180	94.28	4600	65.18	7600	39.45
260	1153	970	309.1	1680	178.5	2390	125.4	3200	93.69	4620	64.90	7650	39.19
270	1110	980	305.9	1690	177.4	2400	124.9	3220	93.11	4640	64.62	7700	38.94
280	1071	990	302.8	1700	176.4	2410	124.4	3240	92.54	4660	64.34	7750	38.69
290	1034	1000	299.8	1710	175.3	2420	123.9	3260	91.97	4680	64.06	7800	38.44
300	999.4	1010	296.9	1720	174.3	2430	123.4	3280	91.41	4700	63.79	7850	38.19
310	967.2	1020	293.9	1730	173.3	2440	122.9	3300	90.86	4720	63.52	7900	37.95
320	936.9	1030	291.1	1740	172.3	2450	122.4	3320	90.31	4740	63.25	7950	37.71
330	908.6	1040	288.3	1750	171.3	2460	121.9	3340	89.77	4760	62.99	8000	37.48
340	881.8	1050	285.5	1760	170.4	2470	121.4	3360	89.23	4780	62.72	8050	37.25
350	856.6	1060	282.8	1770	169.4	2480	120.9	3380	88.70	4800	62.46	8100	37.02
360	832.8	1070	280.2	1780	168.4	2490	120.4	3400	88 18	4820	62.20	8150	36.79
370	810.3	1080	277.6	1790	167.5	2500	119.9	3420	87.67	4840	61.95	8200	36.56
380	789.0	1090	275.1	1800	166.6	2510	119.5	3440	87.16	4860	61.69	8250	36.34
390	768.8	1100	272.6	1810	165.6	2520	119.0	3460	86.65	4880	61.44	8300	36.12
400	749.6	1110	270.1	1820	164.7	2530	118.5	3480	86.16	4900	61.19	8350	35.91
410	731.3	1120	267.7	1830	163.8	2540	118.0	3500	85.66	4920	60.94	8400	35.69
420	713.9	1130	265.3	1840	162.9	2550	117.6	3520	85.18	4940	60.69	8450	35.48
430	697.3	1140	263.0	1850	162.1	2560	117.1	3540	84.70	4960	60.45	8500	35.27
440	681.4	1150	260.7	1860	161.2	2570	116.7	3560	84.22	4980	60.20	8550	35.07
450	666.3	1160	258.5	1870	160.3	2580	116.2	3580	83.75	5000	59.96	8600	34.86
460	651.8	1170	256.3	1880	159.5	2590	115.8	3600	83.28	5050	59.37	8650	34.66
470	637.9	1180	254.1	1890	158.6	2600	115.3	3620	82.82	5100	58.79	8700	34.46
480	624.6	1190	252.0	1900	157.8	2610	114.9	3640	82.37	5150	59.22	8750	34.27
490	611.9	1200	249.9	1910	157.0	2620	114.4	3660	81.92	5200	57.66	8800	34.07
500	599.6	1210	247.8	1920	156.2	2630	114.0	3680	81.47	5250	47.11	8850	33.88
510	587.9	1220	245.8	1930	155.3	2640	113.6	3700	81.03	5300	56.75	8900	33.69
520	576.6	1230	243.8	1940	154.5	2650	113.1	3720	80.60	5350	57.11	8950	33.50
530	565.7	1240	241.8	1950	153.8	2660	112.7	3740	80.17	5400	55.52	9000	33.31
540	555.2	1250	239.9	1960	153.0	2670	112.3	3760	79.74	5450	55.01	9050	33.13
550	545.1	1260	238.0	1970	152.2	2680	111.9	3780	79.32	5500	54.51	9100	32.95
560	535.4	1270	236.1	1980	151.4	2690	111.5	3800	78.90	5550	54.02	9150	32.77
570	526.0	1280	234.2	1990	150.7	2700	111.0	3820	78.49	5600	53.54	9200	32.59
580	516.9	1290	232.4	2000	149.9	2710	110.6	3840	78.08	5650	53.07	9250	32.41
590	508.2	1300	230.6	2010	149.2	2720	110.2	3860	77.67	5700	52.60	9300	32.24
600	499.7	1310	228.9	2020	148.4	2730	109.8	3880	77.27	5750	52.14	9350	32.07
610	491.5	1320	227.1	2030	147.7	2740	109.4	3900	76.88	5800	51.69	9400	31.90
620	483.6	1330	225.4	2040	147.0	2750	109.0	3920	76.49	5850	51.25	9450	31.73
630	475.9	1340	223.7	2050	146.3	2760	108.6	3940	76.10	5900	50.82	9500	31.56
640	468.5	1350	222.1	2060	145.5	2770	108.2	3960	75.71	5950	50.39	9550	31.39
650	461.3	1360	220.4	2070	144.8	2780	107.8	3980	75.33	6000	49.97	9600	31.23
660	454.3	1370	218.8	2080	144.1	2790	107.5	4000	74.96	6050	49.56	9650	31.07
670	447.5	1380	217.3	2090	143.5	2800	107.1	4020	74.58	6100	49.15	9700	30.91
680	440.9	1390	215.7	2100	142.3	2810	106.7	4040	74.21	6150	48.75	9750	30.75
690	434.5	1400	214.2	2110	142.1	2820	106.3	4060	73.85	6200	48.36	9800	30.59
700	428.3	1410	212.6	2120	141.4	2830	105.9	4080	73.49	6250	47.97	9850	30.44
710	422.3	1420	211.1	2130	140.8	2840	105.6	4100	73.13	6300	47.59	9900	30.28
										6350	47.22	9950	30.13

length of a single wave, which in this case would be equal to 300,000,000 (appx.) divided by 500, equalling 600,000 cycles per second or (reduced to kilocycles by dividing by 1,000) 600 kilocycles. On the preceding page is given a chart for converting kilocycles to meters, or meters to kilocycles.

KILOLINE—This is equal to 1,000 *lines of force*, or 1,000 *maxwells* (q.v.). The *maxwell* or the *line of force* is the unit used to measure *magnetic flux*. (See *Line of Force, Maxwell, Magnetic Flux.*)

KILOVOLT — abbreviated K.V. — The kilovolt is equal to 1,000 volts.

KILOVOLT-AMPERE — Abbreviated K.V.A. or kv-a.—The product of the effective volts across the terminals of a circuit by the effective amperes flowing in that circuit, divided by 1,000. Alternators are rated in *kilovolt-amperes* because their load determines the power factor. Thus with a zero *power factor* (q.v.) the *true power* (q.v.) in kilowatts would equal zero even if the alternator were delivering maximum amperes and maximum volts. At 50% power factor the kilowatt capacity would be half that at 100% power factor. On the other hand the kilovolt-ampere capacity remains the same no matter what the power factor.

KILOWATT — abbreviation K.W.—The kilowatt is a unit of electrical power having the value of 1,000 watts. One kilowatt equals 1,000 watts. To reduce the power expressed in kilowatts to watts, it is necessary to multiply the number of kilowatts by 1,000. Most electrical formulas use watts rather than kilowatts. Therefore, in solving such formulas, where the power is given in kilowatts, it is necessary to multiply this by 1,000 before using it in the formula. For example, suppose the power expended in an electrical circuit carrying 100 amperes is 20 kilowatts. The voltage will equal the power in watts divided by the current in amperes. This is $\frac{20 \times 1{,}000}{100} = 200$ volts. Motors are often rated in kilowatts instead of horsepower. For practical calculations, the horsepower rating of a motor is equivalent to four-thirds of its kilowatt rating. (See *watts.*)

KILOWATT-HOUR—The energy expended when work is done for one hour at the rate of one kilowatt. It is possible to calculate the energy expended in an electrical circuit over a certain period of time, knowing the current in amperes and the resistance of the circuit. Using the following formula the energy expended will be obtained in kilowatt-hours:

$$\frac{\text{Work performed or energy expended} = \text{Current}^2 \times \text{Resistance}^* \times \text{Time}^\dagger}{3{,}600{,}000} = \text{kilowatt-hours.}$$

KIMURA SHUNKICHI—Japanese radio authority. Kimura was born in 1866, and educated at the Scientific College of the Tokyo Imperial University, his special study being physics, which he continued to study at Harvard and Yale Universities, United States.

In 1901 Kimura entered the Japanese Navy and first began to study wireless telegraphy, particularly for naval use. In 1906 he was appointed the Japanese delegate to the International Wireless Telegraph Conference, Berlin, and after his retirement from the Navy, 1912, he became the director of the Nippon Radio Telegraph and Telephone Company. Kimura has written largely on radio subjects for various scientific journals.

KINETICS—That branch of dynamics dealing with forces that produce or change motion. Energy is usually classified either as *potential* or *kinetic*. When energy is available for the production of work it is referred to as potential energy. Energy at work is *kinetic energy* (q.v.). The term *kinetic* refers to motion.

KINETIC ENERGY—The work a body is able to do by reason of its motion. The unit of energy is the *erg*. This unit is also used to measure work. (See *Erg.*) It should be noted that *kinetic energy* is stored in the magnetic field. It requires more energy to start a current in a circuit having great inductance than in one having small inductance. This greater amount of energy is stored in the magnetic field and is returned to the circuit when the current decreases to zero.

KINRAIDY SPARK GAP—A form of *quenched spark gap* (q.v.) consisting of two water-cooled flat terminals close together and between which discharge occurs.

KIRCHOFF'S LAWS—These are two in number. First Law: The sum of the currents flowing to a point, in any electrical circuit, is equal to the sum of the currents flowing away from that point. Second Law: In any closed electrical circuit, the sum of the impressed *electromotive forces* (q.v.) will equal the sum of the *voltage drops* (q.v.). This statement requires modification, in so far as "addition" of voltages is concerned. Voltages are added provided that they are in the same direction, but must be subtracted if in opposite directions. These two laws depend on *Ohm's Law* (q.v.) and constitute a further application of that law to more complicated electrical circuits. In addition to simple electrical circuits, conductors may be connected in various complicated networks, all of which come under the heading of *divided circuits*. By means of *Kirchoff's Laws*, the current in any part of a divided circuit may be found, if the resistances of the various parts, and the electromotive forces are given.

KITES AS AERIAL SUPPORTS—During the early experiments with wireless, kites were successfully employed by Marconi as aerial supports.

KLEIN, RENE HENRI—Anglo-French radio expert. Born at Soults-sous-Foret, France, in 1880, he first became interested in wireless telegraphy in 1908. He was the founder and first honorary secretary of the Wireless Society of London, 1913-20, and became a vice-president of the society. He was one of the inventors with H. L. McMichael of the synthetic galena crystal Radiocite, and has written many articles on wireless subjects.

KNIFE SWITCH—A switch in which the movable arm wedges in between two parallel spring clips. Knife switches may be either of the *single pole* (s.p.) or *double pole* (d.p.) types and also may either be *single throw* (s.t.) or *double throw* (d.t.). Knife switches are often referred to as *Lever* switches They are used on direct current circuits up to 250 volts and on alternating current circuits up to 500 volts (See *Ground Switch; Double Throw Switch.*)

KLEMENCIC THERMAL JUNCTION—A device consisting of two thin sheets of brass, designated in the illustration as A, A. These are 4 inches by 11.8 inches and are placed about 1 inch apart. Soldered to one plate is a fine platinum wire and to the other plate is a fine platinum-nickel wire. These two wires are connected together at J to form a *thermal junction* (q.v.) and are then led off at right angles and are soldered at their other ends to the leads R, R, of a sensitive *galvanometer* (q.v.). This resonance system is fixed at the focal line of a suitable cylindrical metallic reflector. When electric waves, with the electric force parallel to A A fall on this receiver, electric oscillations between A and A produce heating at the junction J which is the point of contact between the two dissimilar metals. In consequence, the heat developed gives rise to a thermoelectromotive force at the junction and consequently to a current in the galvanometer. This instrument can be used to detect and measure electric waves.

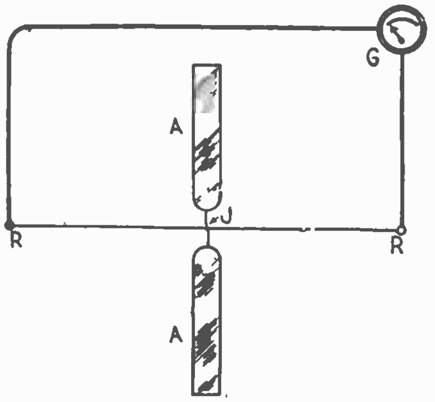

Schematic diagram showing Klemencic thermal junction.

KOEPSEL PERMEAMETER—An instrument used for measuring the *permeability* (q.v.) of an iron or steel bar. In its principle of operation, this device is a movable coil milli-voltmeter in which the permanent magnet is replaced by an electromagnet. The sample to be tested forms a part of this electromagnet. The flux produced in the sample and in the pole pieces of the instrument is measured by a movable coil similar to those used in direct current voltmeters. The coil is supplied with energizing current from a small dry battery. Deflections are indicated by a pointer on a dial, the scale being calibrated directly in magnetic densities.

KOHLRAUSCH'S LAW — This law states that the rate of motion of each atom in a solution undergoing electrolysis for a given liquid is independent of the element with which it may have been in combination.

KOLSTER, FREDERICK A.—American radio authority. Born at Geneva, Switzerland, January 13, 1883, he was educated at Cambridge, Mass., and Harvard University. From 1902-08 he was assistant to J. S. Stone, the radio expert, and from 1909-12 to Dr. Lee de Forest. He was appointed chief of the Radio Section of the Bureau of Standards, a post he held until 1921, when he became research engineer to the Federal Telegraph Company.

Kolster has written a large number of articles on radio subjects, including those on the "Effects of Distributed Capacity in Coils," "Reinforced Harmonics in Radio Transmission," etc. He is the inventor of the Kolster decremeter, of a radio compass, of directional radio systems, etc., and is a member of the American Institute of Radio-Engineers.

---

*Ohms. †Seconds.

KOOMANS, NICHOLAS—Dutch radio authority. Born at Delft, 1879, he was educated there as a mechanical and electrical engineer. In 1908 he received his doctorate for his dissertation "Regarding the Influence of Self-Induction in Telephone Conducting Wires." He was one of the founders of and editor of the "Monthly Review of Telephony and Telegraphy," and was also one of the founders of the Dutch Society for Radio Telegraphy. Koomans is a member of the International Electrotechnical Commission and professor in physics and theoretical electrical engineering at the school of the Dutch Post and Telegraph Administration.

KORDA AIR CONDENSER—A variable *condenser* (q.v.) using air as the *dielectric* (q.v.) consisting of two sets of semi-circular plates, one set being connected together and fixed in position and the other set being also connected together but capable of being rotated about a central axis. Rotating the movable plates brings a greater or less area of the two sets of plates into interlapping positions, thus permitting the capacity of the condenser to be varied. This condenser was patented by Korda in Germany in 1893.

KORN, ARTHUR—German radio expert. Born at Breslau, Germany, 1870, Korn was educated at Leipzig and Paris and became professor of physics at the University of Munich, 1903-8, and afterwards professor at the Polytechnical High School, Charlottenburg, Berlin.

Korn is well known for his experiments on and the invention of a method of transmitting photographs by telegraphy, the first photograph being telegraphed from Munich to Berlin in 1907 by his methods. He is the inventor of a system of telautography and wireless phototelegraphy, and the author of a number of standard books on electricity.

L

L—The symbol of *inductance* (q.v.) or the *coefficient* of *self-induction.* In general, the co-efficient *L* by which the rate of change of current in any circuit must be multiplied to obtain the value of the self-induced *electromotive force* (q.v.) is called the *coefficient* of *self-induction* or simply the *inductance.* (See *Self-induction, coefficient.*)

LAG—One *alternating current* quantity reaching its maximum at a later time than a second alternating current quantity. As an example,—the current is said to *lag* behind the impressed voltage when maximum current occurs after maximum voltage. In other words, the current and voltage in such an alternating current circuit are out of *phase* (q.v.). In the above example, since the current *lags* behind the voltage, the voltage is said to lead the current. *Inductance* (q.v.) in an alternating current circuit will cause the current to *lag* behind the impressed voltage. *Capacity* (q.v.) in such a circuit will cause the current to *lead* (q.v.) the impressed voltage. The word *lag* may be applied to magnetic quantities as well as electrical quantities. Thus, one magnetic flux may *lag* behind another as in the case of an induction motor where the induced *magnetic flux* (q.v.) *lags* behind the impressed magnetic flux. Another type of magnetic *lag* is the retardation of magnetic effects behind their causes, due to *hysteresis* (q.v.) in the magnetized substance. (See *Lead,* also *Leading Current.*)

LAMINATED—Composed of a number of thin plates, one on top of another. Examples of laminated construction are *laminated brushes, laminated cores laminated conductors,* etc.

LAMINATED CORE—An arrangement of soft iron plates or stampings forming the core of a transformer, or of an armature of a motor or generator. Cores are laminated in this way, principally to reduce *eddy current* (q.v.) losses. In the case of an armature core, this must be laminated parallel to the direction of rotation (in other words, perpendicular to the axis of rotation). Laminations are insulated from each other by means of japan, varnish, shellac or simply rust.

LAMP BANK—A bank of incandescent lamps arranged so that they can be connected in series or parallel. *Lamp banks* are used in laboratories in conjunction with electrical machinery under test. In this case, the lamps act as the *load,* the more lamps being connected in *parallel,* the greater the amount of current being drawn. *Lamp banks* are sometimes connected in *series* in a circuit for the purpose of reducing voltage. An example in this connection is the use of a lamp or bank of lamps in charging a 6-volt storage battery from a 110-volt direct current source. Lamp banks are sometimes called *lamp panels* or *lamp batteries.*

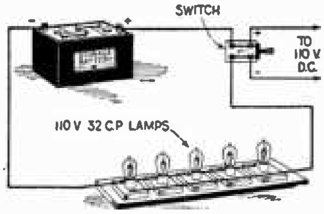

Use of lamp bank in charging storage battery.

LANGMUIR, IRVING—American radio authority. Born at Brooklyn, New York, January 31st, 1881, he was educated at the School of Mines, Columbia University, and was for some time assistant to Professor Nernst at the University of Gottingen. In 1906-9 he was instructor in chemistry at the Stevens Institute of Technology, and in the latter year he became a research assistant to the General Electric Company, at Schenectady. He has carried out a series of brilliant researches on apparatus used in wireless telegraphy and telephony, on electron discharge apparatus, atomic and molecular structure, etc.

Langmuir has paid particular attention to high-vacuum valves, and is the inventor of the Langmuir valve. To him is due the discovery that by treating the tungstic oxide used in the construction of tungsten filaments with certain compounds of thorium the filament becomes thoriated tungsten and the electronic emission is enormously increased. These filaments are the ones used in dull emitter valves, the electrons being given off at comparatively low temperatures Langmuir is the author of many papers to scientific and technical journals, including those on pure electron discharge, thermionic currents in high vacua, etc. He is also well known for his mercury condensation pump for high vacua.

LAP WINDING—An armature winding in which the opposite ends of each coil are connected to adjoining commutator segments. As a result the windings lap back and upon themselves forming loops. *Lap winding* is also called *loop winding, parallel winding* and *multiple winding.* In the elementary form of *lap winding* the method of connection to the commutator bars is as follows: One side of a coil is connected to a commutator bar, and the other side of the coil, which is located 180 electrical degrees (the distance from the center of a north pole to the center of the adjacent south pole) away, has its end connection brought back and soldered to the next adjacent commutator segment. This method of overlapping the ends is continued all the way around the armature until all the slots have been filled and the circuit has been closed on itself. The coils must all be symmetrical. This means that if there is a forward throw of a certain num-

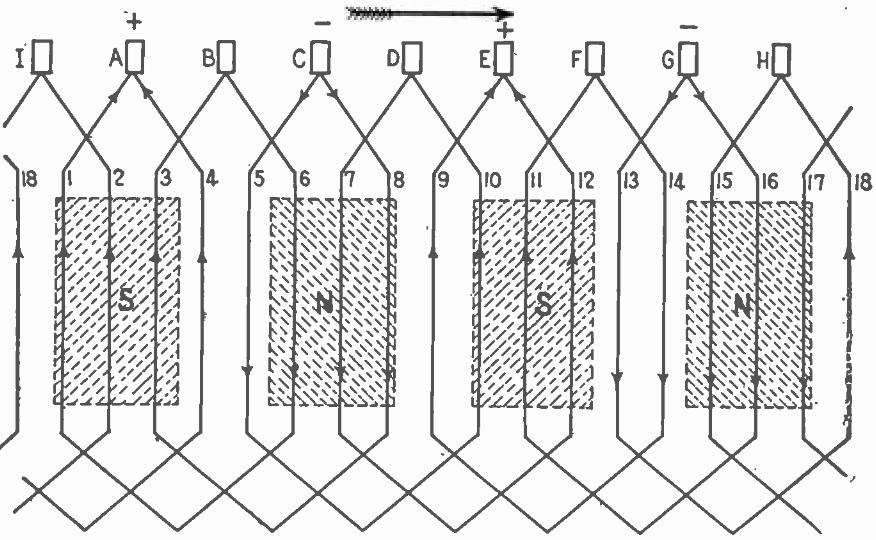

A typical lap winding shown in developed view. In the winding shown each inductor is connected at the rear of the armature to one five slots away. Thus 3 is connected to 8, 5 to 10, etc. After having made one complete "element"—as, for example, A-1-6-B—the winding forms a second "element," B-3-8-C, which laps over the first. This is continued around the armature until the winding closes on itself.

ber of slots on the first coil, all the other coils must have the same forward throw. In like manner all the return throws must be the same. In the *lap wound* generator there are as many paths for the current as there are poles. Hence the name *parallel* or *multiple winding*. This necessitates the use of as many sets of brushes as there are poles. An increased number of paths in parallel means increased current capacity. In the case of a motor, the greater the number of armature paths between brushes, the greater will be the speed, other conditions remaining equal.

LAW OF ELECTROMAGNETIC INDUCTION—Also known as the "*Law of Generator Action*" and as "*Faraday's Law.*" If there is relative motion between an electrical conductor and a magnetic field, such that the conductor cuts *flux*, an *electromotive force* (q.v.) will be induced in the conductor. The amount of the induced electromotive force will depend upon the strength of the magnetic field, upon the speed of cutting, upon the length of the conductor cutting the field and upon the direction of motion between the conductor and the field. Thus the stronger the magnetic field, the greater the induced electromotive force. The induced electromotive force will also increase with increased speed of cutting and with increased conductor length. When the conductor cuts the flux at right angles, the induced electromotive force is maximum. If the flux is cut in any other direction, the induced electromotive force will be less. The induced electromotive force decreases from maximum to zero as the angle of cutting decreases from 90 degrees to zero degrees. If a conductor is of such length and moves in such a direction, and at such a speed, as to cut 100,000,000 lines of force in one second, one *volt* (q.v.) will be induced in it. Expressed as an equation the induced electromotive force (in volts) will equal the flux cut (in *lines* or *maxwells*) divided by the product of the time (in seconds) times 100,000,000. This equation refers to the average value of the electromotive force induced in a single conductor. If a number of conductors are used the same electromotive force will be induced in each. If the conductors are so connected that their electromotive forces add up, the total average electromotive force can be found by multiplying the above equation by the number of conductors.

LATOUR, MARIUS—French radio expert. Born in October, 1875, he was educated at the University of Paris and the Ecole Superieure d'Eléctricité, Paris. He was for many years consulting engineer to the General Electric Company, and he specialized in the construction of high-frequency machines. Latour is the designer of the so-called S. F. R. alternator, in which there are a smaller number of stator slots than usual.

During the World War Latour was engaged in research work under General Ferrie, and he invented a system of elimination of the interference produced in telephone lines by neighboring high-tension power lines which has been installed throughout the greater part of France. Latour is the inventor of the now widely used system of high-frequency multiplex telegraphy and telephony, using the three-electrode valve for generation and reception. He is a member of many scientific societies, including the Institute of Radio Engineers, and of the American Institute of Electrical Engineers. He has written many articles on radio for scientific and technical journals, and is the inventor of a great many important radio developments covering practically the whole range of the art.

LEAD—Pronounced LEED—One alternating current quantity reaching its maximum at an earlier time than a second alternating current quantity. Thus in an alternating current circuit, the current is said to *lead* the voltage if maximum current occurs before maximum voltage. If the current reaches its maximum 90 *electrical degrees* ahead of the voltage, current and voltage are said to be in *quadrature* (q.v.) with current *leading* and voltage *lagging*. (See *Degrees, Electrical*, also *Leading Current* and *Lag*.)

LEAD SULPHATE—Formula $PbSO_4$—A white insoluble salt which is found in native form in the ground and which can be formed artificially by adding sulphuric acid to a soluble lead salt. The active elements of a lead *storage battery* (q.v.) consists of (PbO_2) *lead peroxide* on the positive plate and (Pb) *spongy lead* on the negative plate, with dilute *sulphuric acid* (H_2SO_4) as the *electrolyte* (q.v.). The final result, on discharge of such a battery is the formation of *lead sulphate* on both the positive and negative plates. This is a part of the process of producing current. After a normal discharge, the sulphate is finely crystalline and of such nature that it is easily reduced by the current flowing through the battery on charge. If charging is neglected and the battery is allowed to stand in a discharged state, the sulphate gradually ceases to be crystalline. It fills the pores of the plates and finally makes the active material hard and dense, thus lengthening the time required for a charge. A battery in this condition is said to be *sulphated*. If sulphating proceeds too far, it will result in *buckling* of *plates* (q.v.), loss of active material, greatly decreased capacity, lower efficiency and an increased internal resistance. A moderately sulphated battery can be restored by means of a prolonged charging at a moderate rate. In a badly sulphated battery, a number of successive charges and discharges may be necessary.

LEAD-IN—The wire leading from the aerial to the radio apparatus. *Lead-in* wire, as used for receiving sets, is usually rubber covered wire although bare antenna wire can be used provided it is properly insulated from the walls of the building. A poorly insulated lead-in will greatly reduce the efficiency of the finest antenna. (See *Down Lead.*)

LEAD-IN INSULATOR—Any form of insulator used for passing aerial lead-in through window, wall or roof. These insulators are usually made of porcelain, electrose or insulated copper strips. (See *Insulator, lead-in*, also *Window Lead-in.*)

LEADING CURRENT—An alternating current which reaches its maximum value at an earlier time than the voltage. When an inductance is inserted in an alternating current circuit, the self-induced *electromotive force* (q.v.) will combine with the *impressed* electromotive force and the resultant of the two will be the active electromotive force which causes the current flow. The current will always be exactly in *phase* (q.v.) with and proportional to the resultant electromotive force. The induced electromotive force is 90 degrees behind the current and therefore behind the resultant electromotive force, which is in phase with the current. The impressed electromotive force will lead the current and will be higher than the resultant electromotive force by an amount which at each instant is equal to the counter-electromotive force of the inductance. When a condenser is inserted in an alternating current circuit the current will charge the condenser. As the amount of current stored in the condenser increases, the electromotive force of the condenser increases also until the impressed and the condenser electromotive forces are equal. The condenser electromotive force, being a counter pressure, the current flow will cease when the two electromotive forces balance. The current being zero at this point and the condenser electromotive force a maximum, it follows that the condenser electromotive force is 90 degrees in advance of the current and hence of the resultant electromotive force which is in phase with the current. The impressed electromotive force lags behind the current which in this case is a *leading current* and moreover the impressed electromotive force is greater than the resultant electromotive force. It follows that if either capacity or inductance be placed in an alternating current circuit, the phase of the current with respect to the impressed electromotive force will change and the current flow will be reduced. Capacity sets up an electromotive force 90 degrees in advance of current flow, while inductance sets up an electromotive force of 90 degrees behind current flow. These two effects tend to neutralize each other when the proper amount of inductance and capacity are connected in series in a circuit and when they are just equal, no electromotive force other than the impressed is left to act on the circuit, the resultant and impressed electromotive forces are identical and there is no phase displacement. This is said to be a *resonant* (q.v.) circuit. (See *Lead, Lag*, also *Inductive Rise.*)

LEADS—Wires conveying or "leading" current from one point to another in a circuit. The conductors in any system of electrical distribution are referred to as *leads*. Sometimes the term *lead* is applied to the flexible wire or cable conveying current to or from a piece of apparatus.

LEAK—See *Grid Leak.*

LEAKAGE LINES—Also termed *leakage flux*—That part of the total magnetic flux in a magnet, transformer, generator, etc., which fails to complete the closed magnetic circuit and hence is dissipated or wasted. (See *Keeper.*)

LEAKAGE, SURFACE—The flow of electricity across the surface of insulators. Its amount depends on the material and preparation of the surface and the humidity of the atmosphere and is recorded by measuring the surface *resistivity* (q.v.). Several methods are employed to reduce *surface leakage*. Porcelain *insulators* for outdoor use are often of the *petticoat* type so that a portion of the surface of the insulator remains dry during a rain or snowstorm. Another method very generally used in radio work to reduce surface leakage, is the use of corrugated surfaces on insulators. This method increases the length and hence the resistance of the surface across which leakage is likely to take place.

LENGTH OF AERIAL—The distance from the aerial binding post at the instrument to the farthest point of the antenna system, measured to any of its extremities. Only one component wire of a system is taken into consideration even if two, three or four are used. In a *T aerial* only half of the horizontal span must be added to the length of the down lead. In the *umbrella* type, the length of one radial lead is added to the down lead.

LENZ'S LAW—An induced current always tends to stop the motion which produces it. Stated in another way, whenever current flows in a conductor, due to electromagnetic induction, the action between the magnetic field which is being cut and the field around the conductor is always such as to tend to oppose further motion of the conductor. The following is still another way of stating *Lenz's Law*: The field due to the induced current always acts to prevent a change of the original field.

LEYDEN JAR—A condenser consisting of a glass jar with an outer and an inner coating of tinfoil covering the bottom and the sides nearly to the neck.

A small leyden jar.

A brass rod terminating in an external knob passes through a wooden stopper and is connected to the inner coating by a loose chain. This condenser received its name from Leyden, the city in Holland in which it was invented.

L. F. C.—Abbreviation for *Low Frequency Current*. (See *Frequency, Low*, also *Frequency*.)

LIGHT — Radiant energy or visible *radiation* which affects the eye so that objects become visible. Light consists of *ether* vibrations varying between 400 billions per second producing waves 271 ten-millionths of an inch in length and giving red light, up to 750 billions per second producing waves of 165 ten-millionths in length, giving violet light. Faster or slower vibrations produce waves not visible to the human eye. Shorter waves are called *ultraviolet rays*, while those immediately above the red rays are known as *infra-red rays*. The speed of light waves is identical with that of other electromagnetic waves and electricity, and is 186,000 miles or 300,000,000 meters per second.

LIGHT LINE AERIAL—A device which acts as an aerial for a radio receiving set and which utilizes the electric lighting system by means of a plug to which a condenser is attached. The plug is inserted in any lighting socket and the aerial connection of the receiving set it attached by means of a flexible lead to one terminal of the condenser. (See *Adapter, Aerial*.)

LIGHTNING ARRESTER—(Also called *lightning protector*.) A device for protecting radio or other electrical apparatus from lightning or from atmospheric discharges. Radio *lightning arresters* are provided with two terminals, one being connected with the aerial and the other directly to a ground. Between these terminals is a medium which offers practically no resistance (or technically speaking, which offers low *impedance* (q.v.)) to the flow of very high frequency currents such as lightning discharges. However, this medium acts as an insulator to the normal operating currents. Hence, during the ordinary reception of radio signals, the currents flow past the *arrester* to the receiving instrument but they are unable to pass through the insulating medium of the *ligtning arrester* to the ground. When there is a lightning discharge, the arrester offers the shortest and most direct path to the ground and hence the dangerous current never reaches the apparatus but proceeds directly to the ground through the *lightning arrester*. The more popular designs of lightning arresters are of three types: vacuum, non-air-gap and air-gap. The vacuum type is the more sensitive of the three, due to its having a low functioning voltage. In a usual form, it consists of two terminals in a sealed glass chamber from which air has been partially exhausted. This type of arrester can be discharged repeatedly without impairing its efficiency. The non-air-gap type of arrester involves the use of a carborundum stone, with a special clay binder. This type of arrester has a rectifying action similar to that used in a carborundum detector. The air-gap type of *lightning arrester* depends for its efficiency on a closely spaced air-gap which offers a low break-down voltage. The Underwriters requirements call for the use of an approved lightning protective device on all radio receiving sets where outside aerials are used. Further requirements are that such arresters shall operate on 500 volts or less. (See *Arrester*.)

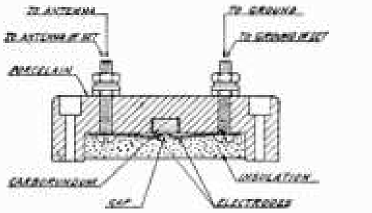

A carborundum type lightning arrester consisting of a block of carborundum held between two pieces of high resistance carbon.

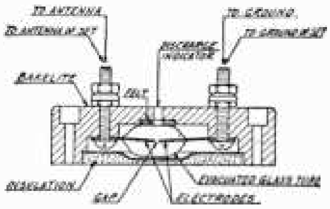

Vacuum type of lightning arrester for indoor work.

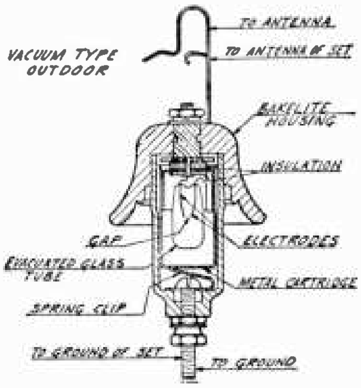

Sectional view of vacuum type lightning arrester used for outdoor work.

Outdoor vacuum type lightning arrester shown above in sectional view.

LIGHTNING GAP—This refers to the *air gap* or *vacuum gap* in a *lightning arrester* (q.v.). This gap is connected across the aerial and ground leads to the receiving set. When the potential in the aerial rises above a certain value (usually 500 volts or less) a spark jumps the gap, discharging the excessive energy to the ground without damage to the radio apparatus. (See *Gap Arrester*.)

LIGHTNING SWITCH—A single pole, double throw knife *switch* (q.v.) used for grounding the antenna during a storm, or in some cases at all times when the apparatus is not in use. The upper terminal of the switch is usually connected to the aerial, the central terminal to which the switch arm is connected, leads to the radio set and the lower terminal is connected to the ground. In most installations, the *lightning switch* is mounted outside the building in which the radio apparatus is located. (See *Ground Switch*.)

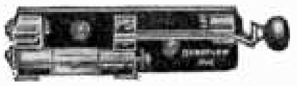

Lightning switch combined with vacuum type lightning arrester.

LIKE CHARGES—Bodies which carry like *electrostatic charge* (q.v.) exert a repulsion against each other.

LIKE POLES—Like magnetic poles repel one another. This is one of the *laws of magnetism*. (See *Magnetism, Laws of*.)

LIMITING VOLTAGE OF BRUSH DISCHARGE—The potential at which a brush discharge passes into an arc or spark. Brush discharges and arcs may be reduced in radio transmitting apparatus by eliminating sharp points or edges on conductors and by coating the edges of metal plates with paraffin. (See *Brush Discharge*, also *Corona*.)

LIMITING VOLTAGE OF GLOW DISCHARGE—The potential at which the glow gives place to the *brush discharge* (q.v.) in a high tension line or in a radio transmitting aerial (q.v.). (See *Corona*.)

LIMITS OF AUDIBILITY—The limiting *frequencies* (q.v.) at which audible sounds are produced. These range from 500 to 1000 cycles per second up to as high as 10,000 cycles per second. *Audio frequencies* (q.v.) are sometimes defined as those conveniently heard in the telephone. (See *Frequency, High*, also *Frequency, Low*.)

LINE—A conducting wire between terminals in a system of electric communication or distribution. The positive wire is sometimes specifically referred to as the *line* and the return wire as *earth*. In many cases the earth or ground is used in the place of a return wire.

LINE OF FORCE—A theoretical conception used to map out a magnetic field and to show the direction along which the magnetism acts. *Lines of force* are conceived as forming closed loops and the path they traverse is called the *magnetic circuit* (q.v.) The positive direction of the *lines of force* is arbitrarily taken as that in which the north-seeking end of a compass moves. *Lines of force* are represented in diagrams by curves so drawn that the direction of the tangent at any point is the direction of the *magnetic force* (q.v.) at that point. The closeness of the lines is used to indicate the intensity of the force. The *line of force* is also used to indicate and map out an *electrostatic* (q.v) field in the same way as it is used for the magnetic field. The *line of force* is also called the *line of flux* or the *line of induction*. This latter term is used especially where the line of force is being considered in respect to its electrical effect. In some cases it is simply referred to as a *line*. Thus a line (*of force*, or *of flux*, or *of induction*) is an electromagnetic unit of magnetic flux and is equal to 10^{-8} *webers* (q.v.). In the practical system of measurements, the *maxwell* (q.v.) equal to one *line*, is the unit of *magnetic flux* (q.v.). When 100,000,000 *maxwells* are cut by a single conductor, in one second, one *volt* (q.v.) will be induced. (See *Law of Electromagnetic Induction*, also *Electromagnetic field*.)

LINEAR OSCILLATOR—See *Oscillator, Linear*.

LINE RADIO—Also known as *Wired Wireless*—A system of communciation devised by Major General George O. Squier which permits the transmission of electromagnetic waves along a wire or conductor. Instead of the waves spreading out in all directions, as in ordinary broadcasting, the waves travel along a predetermined path, viz: the conductor. It is possible to transmit either code or speech by *wired wireless*. Vacuum tubes are used as oscillators at the transmitting end and a specially designed vacuum tube set is used as the receiver. When code transmission is desired, a telegraph key is inserted in the circuit to intermittently stop and start the oscillating waves. This produces dots and dashes and in this way it is possible to send code using undamped radio frequency waves. When it is desired to telephone, the key is omitted and the radio frequency or *carrier waves* (q.v.) are produced continuously except when they are varied or modulated in amplitude corresponding to the voice of the speaker at the transmitting end. The wave which passes from the transmitting end of the wire to the receiving end is of radio frequency and is called the *carrier wave*. In a way, this wave is analogous to the wire which carries the voice current in wire telephony. *Line radio* can be transmitted over ordinary telephone or telegraph lines at the same time as these lines are being used for their ordinary purposes and without any interference whatsoever. High-tension lines are also used for transmission of *wired wireless*.

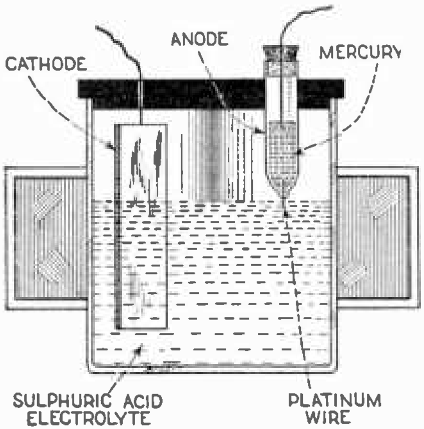

A liquid barreter.

LIQUID BARRETER—An electrolytic detector which utilizes a very fine platinum wire dipped to a small depth in nitric or sulphuric acid. These detectors were formerly used to some extent in receiving wireless code messages. The electrolytic detector of the *barreter* type was invented and named by *Fessenden* (q.v.). The illustration shows a liquid barreter. In the glass jar is a platinum glass-anode, lead cathode and sulphuric acid electrolyte.

LITZENDRAHT—abbreviated Litz—A German word, meaning braided wire. A high frequency conductor built up of a number of very fine conducting strands. The individual strands are enameled and the conductor is usually so constructed that each strand comes to the surface or very near the surface at regular intervals. In the most effective form of conductor the strands are braided so as to form a woven tube. The strands are usually covered by a single braided silk wrapper. The use of *litz* tends to reduce *skin effect* (q.v.). This type of wire has been found to be especially effective when used in the construction of *loop antennas* (q.v.).

LOADSTONE—also spelled *Lode-stone*—Iron ore which shows magnetic properties. The name loadstone is derived from a word meaning leading stone and refers to the property of a compass of pointing to the north magnetic pole. The loadstone is also referred to as a *natural magnet*. (See *Magnetite*, also *Magnet*.)

LOADING—Adding capacity or inductance to a circuit. *Inductance coils* (q.v.) called *loading coils* are used to increase the inductance and condensers are used to increase the capacity. A loading coil may be used to increase the *wave length* (q.v.) of an antenna by placing it in series between the *lead-in* (q.v.) and the receiving set. In the case of a transmitting antenna, the loading coil may be used to tune the aerial for *resonance* (q.v.) at a desired frequency. An example of the use of loading, is that of the addition of a loading coil to a receiving set for the purpose of enabling the set to receive higher wave lengths.

LOADING INDUCTANCE—An inductance coil used to increase the wave length of a circuit. This term usually refers to an inductance coil used in connection with a radio receiving set for increasing its receptive wave length range. Sometimes the term *loading coil* is applied to a unit consisting of an inductance coil, mounted beneath a panel in a small cabinet. On the outside of the panel is a contact arm and a number of studs. These latter lead to taps taken from the inductance coil. By moving the contact arm from one stud to another it is possible to vary the loading of the circuit as desired. A loading coil connected in series with the aerial will increase the natural wave length of the aerial. (See *Loading*, also *Inductance, Antenna*.)

LOCAL ACTION—Electrochemical action within a cell, usually at the positive plate, which does not add to the useful output. In primary cells, zinc is the material usually employed for the positive plate. This zinc practically always contains impurities such as carbon, iron, etc. When the zinc is immersed in the electrolyte, there is a difference of potential between the zinc and the impurities and small local currents flow from the zinc through the electrolyte to the impurities and back to the zinc again. These currents serve no useful purpose and moreover use up the zinc and the electrolyte. *Local action* can be practically eliminated by coating the surface of the zinc with mercury. This is called *amalgamation*. In secondary cells or *accumulators* (q.v.) local action is due to the action which takes place between the lead of the positive and the coating of lead sulphate.

LOCAL OSCILLATIONS—Oscillations set up in the vicinity of a receiving set for test purposes or within the set, as in the *super-heterodyne*, to produce oscillations differing in frequency from the incoming oscillations so as to establish a beat frequency. *Local oscillations* may be produced by means of a buzzer. This is a method frequently used in testing. The use of such oscillations enables a manufacturer of radio receiving sets to test these sets for selectivity, etc., without waiting for actual broadcasting. In the *super-heterodyne* (q.v.) local oscillations are produced by a suitably arranged vacuum tube. (See *Driver*.)

LOCUS—The path of a movement graphically represented by a curve. In mathematical language, the *locus* of an equation is the curve that contains all the points whose co-ordinates satisfy the equation. The circle is defined as the *locus* of all points in a plane equidistant from a fixed point in the plane called the center.

LODGE, SIR OLIVER—British Physicist and radio pioneer. Born at Penkhull, near Stoke-upon-Trent, June 12th, 1851, Oliver Lodge went to the Newport Grammar School at the age of eight, but at fourteen was taken into business to help his father. He worked in the evenings for the intermediate examination in science at the University of London, and eventually took first-class honors in physics. In 1872 he went to University College, London, and in 1877 took his D.Sc. in electricity and became demonstrator, and afterwards assistant professor in physics at University College.

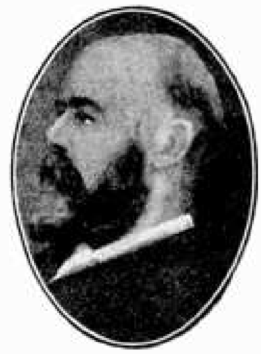

Sir Oliver Lodge.

In 1881 he was appointed first professor of physics at Liverpool, and in 1887 he was made a Fellow at the Royal Society. In 1888 he received the honorary L.L.D. of the University of St. Andrews, the first of a long series of such distinctions. In 1900 he was appointed the first principal of the University of Birmingham, a post from which he retired in 1920. In 1902 he was knighted. Sir Oliver Lodge is one of the foremost physicists of his time, and his brilliant series of researches into electrical phenomena, which he has made his special study have had a profound effect in all branches of that subject. To him, indeed radio owes as much as to any man, for he dealt with fundamentals when the whole subject of electromagnetic radiation was in its infancy. He was led to study the surging and oscillating character of an electrical discharge along wires as a result of his investigations as to the best means of protection against lightning.

One very remarkable experiment of his is now well known under the name of Lodge's resonating jar, now used as a measurer of wave length. In these experiments Lodge was really dealing with the electromagnetic waves discovered by Hertz in 1888. Lodge early recognized the vast importance of the discovery by Hertz, which so brilliantly confirmed the theories of Clerk-Maxwell, and he threw himself into work of investigating these ways with an energy which made him the foremost British physicist on the subject at the time. All this was pioneer work before Marconi and others came along and built upon the solid foundations which Lodge had prepared.

It was during this early period that Lodge discovered his coherer, and with this detector devised the first practical wireless telegraph, sending signals over a distance of several hundred yeards. In May, 1897, Lodge took out his fundamental patent for tuning—a patent that was successfully upheld in the law courts, and extended for seven years by Lord Parker. The patent was acquired by the Marconi Company in 1911. Sir Oliver Lodge has been a pioneer in many other directions. His theories and experiments in connection with the ether, of which he is the leading exponent, are well known, and he has carried out a number of interesting researches on the passage of electricity through liquids. Sir Oliver was the first to devise a simple and direct experiment to show the speed and direction of travel of the constituents of a salt in solution when broken up by a current of electricity. For his researches on electric waves and the passage of light through a moving medium, Sir Oliver Lodge received in 1898 the Rumford Medal of the Royal Society, one of the highest honors the society bestows.

Sir Oliver Lodge has won a well-deserved reputation for the clearness and simplicity with which he is able to explain many of the difficult and fundamental truths of physics, and this clarity of expression is well shown in the many articles he has written.

LODGE DISC DETECTOR—An imperfect contact detector consisting of a steel disc rotating with its sharp knife edge just touching some mercury having a film of oil on it. When an oscillation passes through the circuit, momentary cohesion sets in between the disc and the mercury, but this is immediately broken by the revolving motion of the disc. In this way, oil re-insulates the disc at the completion of each passage of oscillations. Using this type of detector, code signals can be recorded with great precision on a *Morse* tape using a syphon recorder. The *disc detector* is in reality a specialized form of *coherer* (q.v.).

LOG—A list of dial readings of a radio receiving set with corresponding call letters of the broadcasting stations received at these readings. A radio set is said to *log* when the same stations always come in at exactly the same dial readings. Thus on a three dial set, if WEAF comes in at 61, 60 and 60 and always comes in at these identical readings using the same aerial, the set *logs*. All present day sets are constructed so that stations can be accurately logged. Sometimes the term log is used to refer to a graph or chart on which dial settings are plotted against wave lengths of stations recorded. A graph is a diagrammatic representation by means of lines, of the relation between several quantities. Figure 1, gives a simple illustration of this. Suppose it is desired to obtain a general idea of how the radio business varies from year to year or from season to season. On a piece of paper, ruled both vertically and horizontally, as in Figure 1, the months and years are marked at the bottom horizontally.

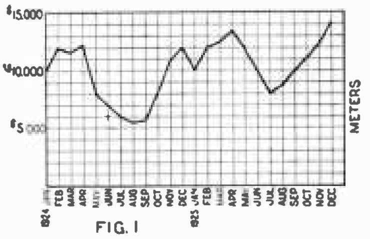

FIG. 1

Log showing variation of a business.

The vertical lines then represent the months and the bottom horizontal line is called the *axis of time*. These vertical lines are known technically as *ordinates*. The vertical line to the extreme left is called the axis of ordinates and after it has been marked as shown in Figure 1, to represent dollars, it may be referred to as the *axis of dollars*.

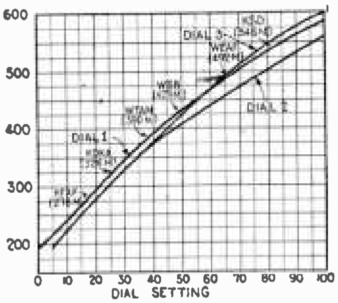

FIG. 3

A three-dial radio log using only one chart.

The horizontal lines are known as *abscissas*, the bottom horizontal line being the *axis of abscissas*. Assume that a certain radio concern did $10,000 worth of business in the month of January, 1924. Looking upward along the vertical line representing this month, until the horizontal line representing $10,000 is reached, a point is put at the intersection of these two lines. Suppose that in the following month, February, 1924, the firm does $12,000 worth of business. The point is found as before by following the line February, 1924, up to the line indicating $12,000. Each horizontal line in the figure represents $1,000, so that this point should be found two lines higher than the one marked for $10,000. This can then be continued indefinitely, plotting the amount of business for each month. Assume that in the summer there is a lull in business. The decline starts, say, in the month of May. The business done amounted to $8,000. By glancing at the chart it is possible to ascertain a number of facts which stand out much clearer than where columns of figures are given. Thus the seasonal character of the business is apparent at a glance. It can also be seen readily that the business was better in 1925 than in 1924, etc. Radio logs are plotted in exactly the same way. Each point must be obtained individually in order for the curve to be absolutely accurate, although in many cases a few well chosen points are sufficiently satisfactory.

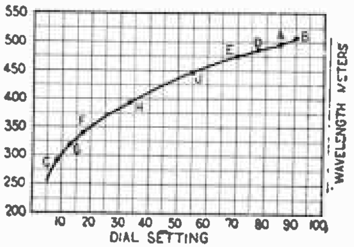

FIG. 2.

A radio log.

In Figure 2, the dial setting is marked along the bottom of the chart. The scale is here chosen at one division on the chart equal to 5 divisions of the dial. If preferred the scale could be chosen so that one division on the chart would equal one division on the dial as shown on chart on the following page, which is a typical log of an actual set. Proceeding, it is first necessary to find out the limits of the chart.

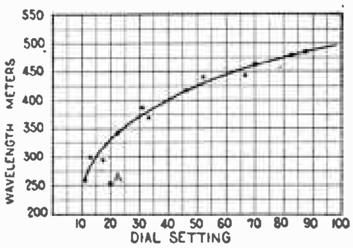

FIG. 4

Radio log showing curve drawn so as to touch as many points as possible.

The limits of the horizontal scale are already known, i.e., zero to 100 on the dial. Looking over the list of broadcasting stations, it is found that they all have wave lengths which fall, roughly between 200 and 600 meters. Therefore it is not necessary to make the vertical scale include wave lengths outside of this range. Consequently the vertical scale is started at 200. Counting upward, two divisions on the chart are taken, for convenience, as equal to 50 meters. Having done this, actual readings are next taken on the radio set. Suppose that the first station tuned in is WEAF and that this is tuned in at 85 on the dial. This station has a wave length of 492 meters. Looking along the bottom of the chart for 85, the vertical line is then followed until the position of 492 meters is reached. This is slightly below the 500 line on the chart, Figure 2, and is indicated by the letter A. Next suppose that WOO, having a wave length of 509 meters is tuned in at 90 on the dial. This, then gives the point marked B. Then another station is tuned in, indicated by the point C,

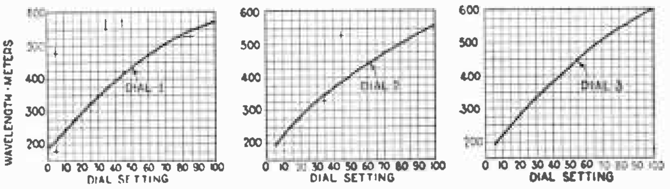

Fig. 5—A three-dial radio log using three separate charts.

Various stations are thus located by points on the chart until enough points are obtained to draw the curve. If the points do not give a smooth curve, it is necessary to approximate a curve, that is to draw it so as to pass through as many of the points as possible. This has been done in Figure 4.

Sometimes points are obtained that do not seem to have any meaning, such as the point A in Fig. 4. This point may have been subject to inaccuracies in reading the dial, or perhaps the operator did not read the wave length correctly from the list of stations. There is always the possibility, also, that he did not plot the point properly. He has plotted the point A as 251 meters at a setting of 20 when it might have been 335 meters at 20 on the dial, in which case it would have fallen directly upon the curve. If none of these difficulties can be found with the point A it can be discarded, however, because it is an absurd point, as the 12 other points, which determine the curve, will testify. Generally it is not good practise to discard points, but the experimenter must learn to use his good judgment.

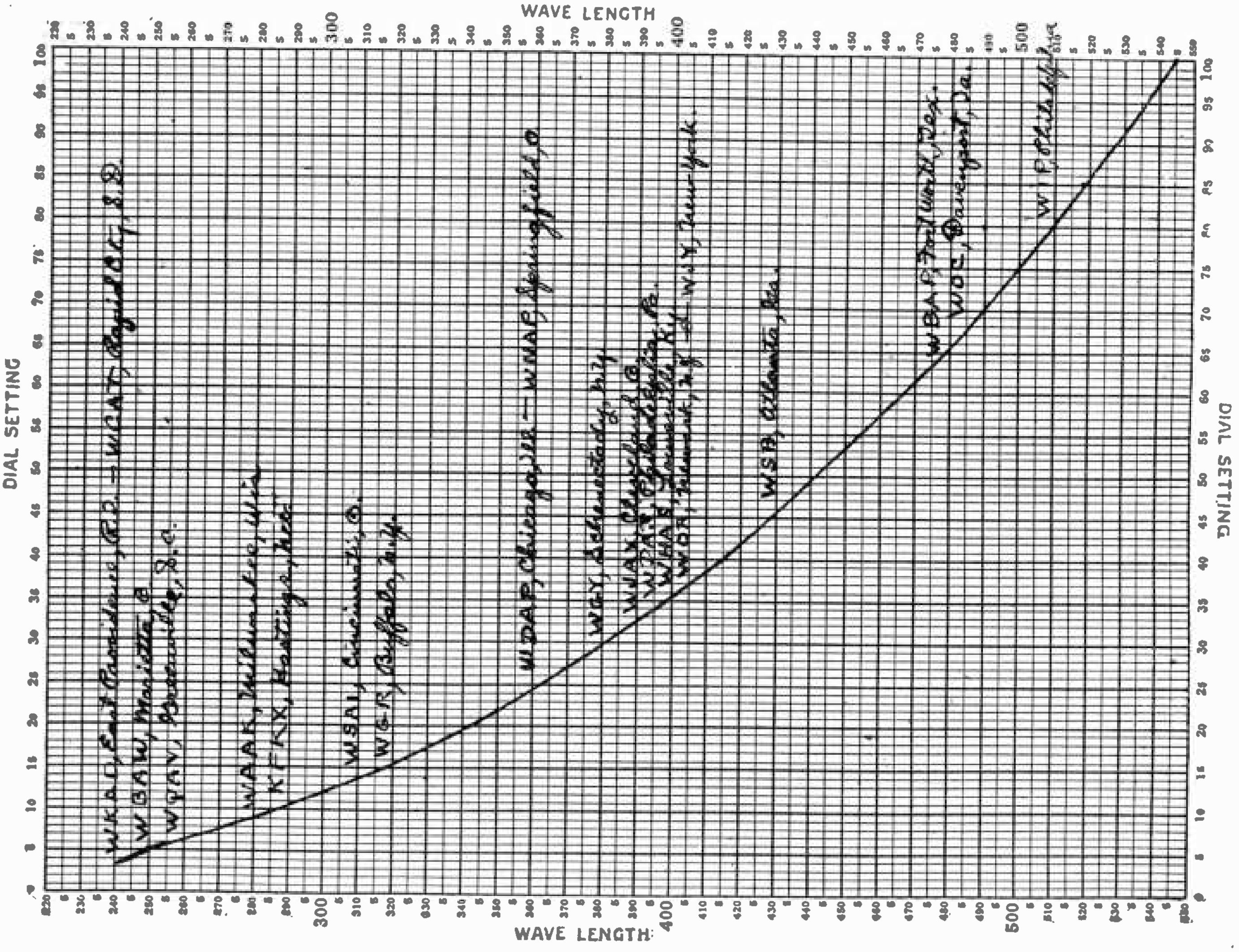

An example of a station log chart.

In single-control sets, where only one dial is employed it is only necessary to have one of these radio log curves. In receivers, such as the Neutrodyne or any two-stage radio frequency amplifier, where three dials are employed, it is generally necessary to log each dial separately. The reason for this is, that although these sets are supposed to tune to stations with identical dial settings, many of them do not do so, a difference as great as 10 divisions on the dial often being found. A three dial log is shown in Fig. 5. Separate charts may be used for each dial, or, for convenience, as has been done in Fig. 3, the three charts may be combined into one. If the difference between the dial settings is very small, say only a division or two, one curve will be sufficient.

Many things can be learned from studying the curves obtained. Take for instance, Fig. 2. This is practically a straight line at the higher wave lengths, changing from about 500 meters to about 400 in 55 divisions of the dial. In the remaining 35 divisions of the dial the wave length changes from 400 to 250. In the first case there was a range of 100 meters for 55 divisions and in the latter case a range of 150 meters in 35 divisions. This is because the curve bends downward. At the low dial settings a small change in the dial reading means a relatively large change in wave length, while at the upper end of the dial the same change in dial reading will mean a relatively small change in wave length. The reason why the stations are so crowded at present, at the small end of the dial, is then very apparent, i.e.:

To illustrate, a station operating on a 200 meter wave corresponds to a frequency of 1500 kilocycles (1 kilocycle equals 1000 cycles), and a station on 200 meters corresponds to a frequency of 1,363.5 kilocycles. In other words the difference in wave length is only 20 meters whereas the difference in kilocycles in 136.5. These two wave bands are at the low end of the present broadcast wave band. Now at the high end of the band we find a wave length of 600 meters corresponds to 500 kilocycles, and a wave of 580 meters corresponds to 517.2 kilocycles, or a difference of but 17.2 kilocycles for the 20 meter difference in wave length. At the lower end of the broadcast band between 200 and 220 meters we found a difference of 20 meters equalled 136.5 kilocycles. This indicates that as the wave length becomes lower the difference in frequency becomes very high.

To bring this most strongly to the mind, go down to the extremely low wave of 5 meters. Here the frequency is found to be 60,000 kilocycles. Going up to 25 meters, this corresponds to 12,000 kilocycles, so that for this difference of 20 meters there is the tremendous difference of 48,000 kilocycles. The accompanying illustration, Fig. 6 will explain the matter very clearly. In changing the dial setting from 100 to, let us say, 90, the wave length is changed roughly from 600 to 580 meters. This is a change from 500 to 517 kilocycles. On the other end of the dial from zero to about 10 we have a change of wave length of roughly 200 to 220 meters, an approx-mate change of from 1500 to 1363 kilocycles. In the first case on the upper part of the dial we have only made a change of 17 kilocycles in the frequency. On the lower end of the dial we have made a change of 137 kilocycles.

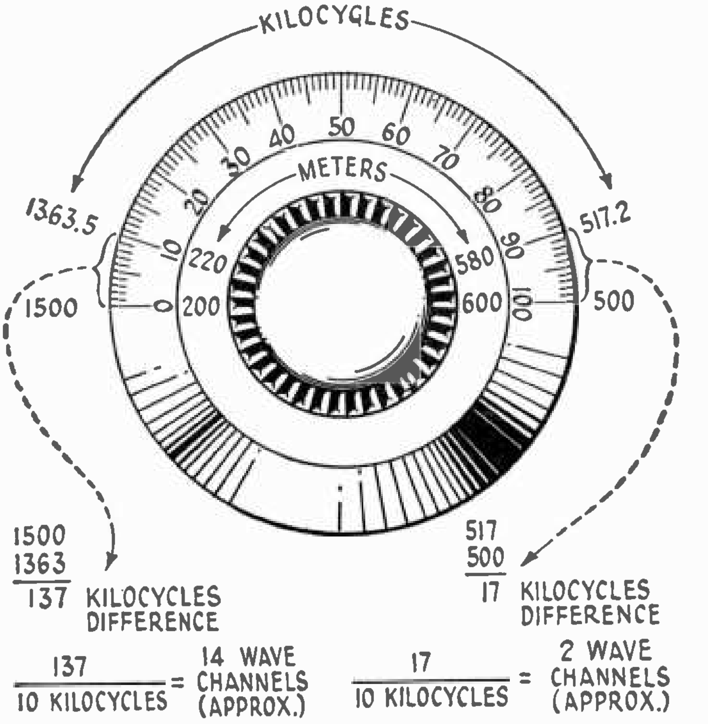
KILOCYCLES
METERS
1363.5
517.2
1500
500
0
10
20
30
40
50
60
70
80
90
100
200
220
580
600
1500
1363
137 KILOCYCLES DIFFERENCE
517
500
17 KILOCYCLES DIFFERENCE
137 / 10 KILOCYCLES = 14 WAVE CHANNELS (APPROX.)
17 / 10 KILOCYCLES = 2 WAVE CHANNELS (APPROX.)

Fig. 6—This illustration shows clearly why stations are crowded on the dial at the lower wave lengths.

According to the frequency allocation as made by the Government, which places broadcast frequencies 10 kilocycles apart, on the upper 10 divisions of the dial there is room for approximately 2 frequency channels whereas on the lower 10 divisions of the dial there is room for roughly 14 frequency channels. This shows very clearly that we would have to be able to separate 14 stations on the lower 10 divisions whereas we have only 2 to separate on the upper 10 divisions of the dial.

So it will be understood why turning the dials but a degree or two at the low setting causes a relatively large change in frequency, tuning in perhaps several stations within one degree of the dial, while at the higher settings, due to the small difference in frequency, as stated, on the higher waves, it may be necessary to turn the dials two or three degrees on either side of the peak of the broadcast wave before it is tuned out. (See *Dial.*)

LOGARITHM—abbreviation LOG—The power to which a number (defined as a base) must be raised in order to equal a given number. As an example, given the number 100, assume that it is desired to determine the *logarithm* of this number to the base 10. (When 10 is used as a base, the *logarithms* obtained are said to belong to the *common* or *Briggs system.*) Then the exponent or power which will raise 10 to the value of 100 is known as the *logarithm* of 100. Obviously, 10^2 equals 100, therefore 2 is the logarithm of 100. *Logarithms* are used in the processes of multiplication, division, raising to powers and taking roots. The whole number, or integer part of a logarithm is called its *characteristic* and the fractional part, is referred to as its *mantissa*. The characteristic of a number greater than unity is positive and is one less than the number of figures to the left of the decimal. The characteristic of a number less than unity is negative and is one greater than the number of ciphers between the decimal and the first significant figure. A *table of logarithms* gives the mantissas only, the characteristic being determined as indicated above.

Logarithms are also used which have as their base the number 2.718 +. These are called *natural* or *Naperian logs*. The base number is usually represented by the symbol ε (epsilon). The natural logarithm of any number may be found from a table of common logarithms by dividing the common log by the logarithm of 2.718 or simply by multiplying the common log by 2.30259. The symbol l_n is often used for the natural logarithm and the abbreviation *log* without a subscript is usually used for the logarithm to the base 10. (See *Naperian Log*, also *Mantissa*, and *Log Table.*)

LOGARITHMIC CURVE—A curve which has no definite minimum in a decreasing curve, or no definite maximum in an increasing curve. A *logarithmic curve* may be further defined as a curve whose ordinates increase arithmetically while its abscissas increase geometrically. A series of damped electrical oscillations diminish according to a logarithmic law and hence a curve drawn tangent to the successive maximum displacements will be a *logarithmic curve.*

LOGARITHMIC DAMPING—See *Damping.*

LOGARITHMIC DECREMENT—The *Naperian logarithm* (q.v.) of the ratio of any maximum to the next following maximum, with the current in the same direction in a decaying wave train. In other words the logarithm of the ratio of two maxima, one cycle apart. The rapidity with which the oscillations decay depends not only on the resistance of the circuit, but on the inductance as well. The greater the resistance and the smaller the inductance, the more rapid is the damping and the rate at which the oscillations increase. If the resistance, inductance and capacity of the circuit have fixed values, each successive maximum of current is the same fraction of the preceding maximum as the latter is of the maximum immediately preceding it. Thus, if the second maximum is 75 percent of the first maximum, the third will be 75 percent of the second, etc. The numerical value of the rate of decrease in this case is .75 and the natural logarithm of .75 is the *logarithmic decrement* or simply the *decrement.* (See *Decrement*, also *Damped Waves.*)

LOG TABLE—A tabulated arrangement of logarithms which gives the *mantissas* (q.v.) only. The numbers whose logarithms are to be looked up are arranged for convenience in consecutive order. (See *Logarithm.*)

LOMBARDI, DR. LUIGI—Italian radio authority. Dr. Lombardi was born at Dronero, August 21, 1876. He graduated from the Royal Engineering School at Turin, Italy. He was Professor of Electricity at the Zurich Polytechnical School, Zurich, Switzerland, and also at the Royal Polytechnic School, Naples, Italy. Dr. Lombardi published many scientific papers including a study of condensers for trans mission and he also invented a special high-tension electrical condenser.

LONGITUDE—The distance east or west of a meridian passing through Greenwich, England, measured in degrees, minutes and seconds.

LOOP ANTENNA—*Loop Aerial*—Also called *Coil Aerial* or simply *Loop*—A receiving or transmitting *aerial* (q.v.) consisting essentially of one or more turns of wire connected to the radio set so as to form a closed circuit. The *loop* acts as a simple inductance coil. The ordinary radio aerial acts as a huge condenser with the aerial system as one plate of the condenser and the *ground* or *counterpoise* (q.v.) as the other. Electromagnetic waves reaching the aerial set up an alternating electromotive force between the wires forming the upper plate of the condenser and the ground or lower plate of the condenser. This action takes place through *electrostatic induction* (q.v.). In the *loop aerial, electromagnetic induction* (q.v.) sets up an induced electromotive force thus causing alternating current to flow to the detecting apparatus. While coil antennas may be used for transmitting, their use in this connection is rather limited. On the other hand, the use of the loop for radio receiving is widespread. The greater the area enclosed

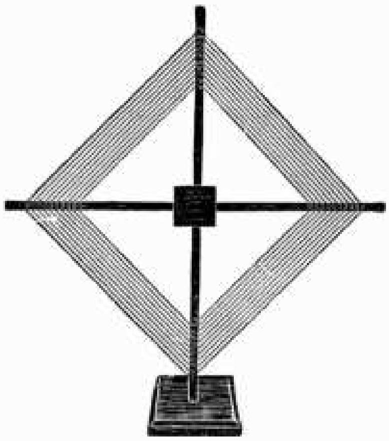

A typical loop antenna.

by a loop the greater the amount of energy it will receive. Loops more than two feet in diameter are usually constructed so as to be collapsible. A common type of loop shown in the illustration is 36 inches high and 28 inches wide. It consists of twelve turns of wire wound in the same vertical plane and held in place by polished hard rubber strips. It has a heavy cast base permitting rotation of the loop as desired. This loop folds up so as to fit into a four inch tube. In another type of loop, sometimes called the *box type*, the wires are wound so as to lie in the same horizontal plane instead of the vertical plane. The use of *Litzendraht* (q.v.) for winding loops has been found to add to their efficiency. Loops are commonly used on *super-heterodyne* (q.v.) sets, *reflex* (q.v.) sets and in some *tuned radio frequency* (q.v.) sets. These latter generally require an extra stage of amplification, although tuned radio frequency sets having only five tubes (detector, two stages of radio frequency and two stages of audio frequency) will operate on a loop and give especially good results when used with the new power tube. As a general rule, it may be stated that any set which will give good results with a loop, will give better results, insofar as distance is concerned, with an outside aerial. For this reason loop operating sets are quite generally constructed with extra binding posts for connecting aerial and ground. The loop is connected directly into the grid circuit, while the aerial and ground are connected to the circuit through an inductance such as an *oscillo-coupler* or antenna coupler. Loops may be used with satisfactory results inside of buildings. Where a compact, portable and selective aerial is desired, the loop is especially useful. The use of the loop, results in a marked directional effect and also has a tendency to reduce *static* (q.v.) and other forms of *interference* (q.v.). Loops are employed in wireless direction finders or *radio compasses* (q.v.) and are also used as aerials in submarine and aeroplane work. Where the loop is used as an aeroplane antenna it may be wound on the wings of the plane or it may be wound on a small frame and placed in the rear of the plane. (See *Coil Antenna*, also *Frame Aerial.*)

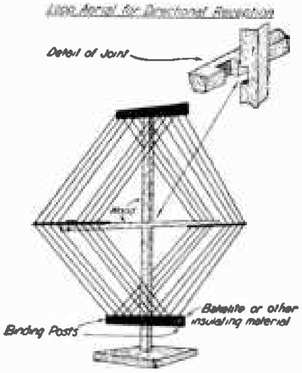
Loop Aerial for Directional Reception
Detail of Joint
Bakelite or other insulating material
Binding Posts

A box-type loop.

LOOPS—The points of greatest amplitude in a wave train. Sometimes called *antinodes*. The term *loop* is also used in describing the electrostatic field spreading out from a charged aerial. Each time an aerial is charged (for example by an induction coil) and discharged across a spark gap a group of gradually decaying oscillations is produced in the aerial called a *train* of oscillations. The interval of time which elapses between a flow of the electricity in the aerial in one direction, and a succeeding similar movement is called the *periodic time* and the reciprocal of this is called the frequency. For example, if the periodic time is one-millionth of a second, the frequency is one million. The diagram shows the manner in which *semi-loops* of electric force represented by the dotted lines move away from the aerial which is represented by the short black line.

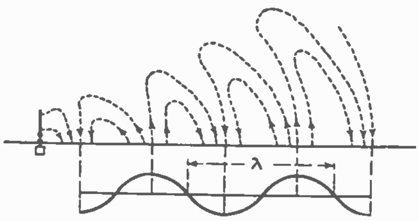

Semi-loops of electric force represented by the dotted lines moving away from the aerial represented by the short black line.

LORENZ COIL—A form of low loss inductance coil, having an air core and of cylindrical shape, being wound in a single layer, with turns slightly kinked to make the coil self-supporting. (See *Coil, Inductance Coils*, also *Low Loss Coils.*)

LOUD SPEAKER—A sound producing device for converting audio frequency currents into sounds loud enough to be heard by an assemblage of people. Nearly all *loud speakers* utilize an electromagnet in order to directly or indirectly cause a diaphragm to vibrate.

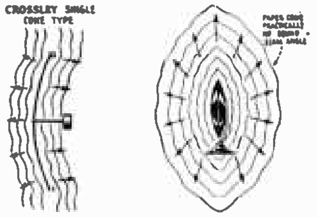

At the left is shown a cone-type speaker having a single diaphragm; at the right one with a double diaphragm.

The audio frequency currents flow through the windings of the electromagnet and the vibrations of the diaphragm correspond to the minute changes and fluctuations of the current thus accurately reproducing sounds. In loud speakers which utilize a metallic diaphragm, the electromagnet directly actuates the diaphragm. In other types, the diaphragm may be of mica, as in the Baldwin unit type, or of parchment or wood as in the cone type speakers. In these speakers, the electromagnet actuates an armature which in turn actuates the diaphragm through a lever action. In certain loud speakers, the electromagnet is of very powerful type and has a very small coil of fine wire attached to a diaphragm but not touching the poles of the magnet. The small coil is connected to the radio set or amplifier and the fluctuating audio frequency currents pass through it. These currents force the coil out of the strong magnetic field in the air gap. In moving, it moves the diaphragm to which it is attached. Speakers of this nature are sometimes called electrodynamic speakers. They are also referred to as power speakers. Such speakers give plenty of volume and in certain makes are reasonably good in tonal quality but they have the disadvantage that the electromagnet must be energized from a source of current such as the storage battery. A loud speaker has been developed in England called the Crystavox for use with crystal sets. This consists essentially of an electromagnetically operated speaker connected with a microphonic relay. The relay is actuated by a storage battery. Its function is to magnify the incoming signals from the crystal set, passing them along with increased strength to the electromagnet of the loud speaker. Still another type of speaker which has been used to a limited extent abroad, dispenses entirely with the electromagnet as the actuating source of diaphragm vibration. In this type, electrostatic principles are involved. An agate cylindrical drum, over which a band of copper or similar metal is wound, is revolved at a uniform rate by a small motor or by clockwork. To one end of the band a diaphragm is fastened mechanically. The other end of the band is secured by a fairly stiff spring. Current from the receiving set is led to the cylinder by means of a brush contact and as the current fluctuates the attraction of the band to the cylinder fluctuates with it and the drag on the diaphragm due to the rotation is increased and decreased. The use of the motor constitutes the most serious disadvantage in this type of loud speaker.

Having discussed the various *types* of loud speakers, from the standpoint of principles of operation, the more important of these will now be considered from the constructional standpoint. The early forms of directly actuated metallic diaphragm, electromagnetic loud speakers consisted essentially of a watch-case telephone receiver, or a pair of receivers attached to a horn. The purpose of the horn was to concentrate and redistribute the sounds caused by the vibration of the diaphragm. One of the first improvements was to modify the shape of the horn so as to make it conform more nearly with correct acoustical principles. Another improvement was to introduce an adjustment by means of which the air gap could be varied. In some cases this was accomplished by moving the magnet closer to the diaphragm and in others by moving the diaphragm itself. This adjustment provided a means of getting greater sensitivity on distant stations (by drawing diaphragm closer to the magnet) and of preventing sticking of diaphragm or rattling, on local stations. The ordinary metallic diaphragm has *one* frequency to which it will respond most strongly. This is called its *resonant* frequency. Special speakers

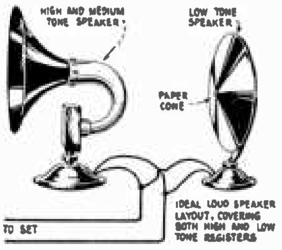
HIGH AND MEDIUM TONE SPEAKER
LOW TONE SPEAKER
PAPER CONE
IDEAL LOUD SPEAKER LAYOUT, COVERING BOTH HIGH AND LOW TONE REGISTERS
TO SET

A high and a low pitched speaker connected in parallel for simultaneous operation.

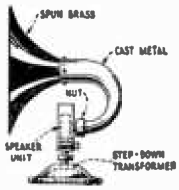

A loud speaker of the horn type having curved cast metal neck and spun brass bell.

called *tuned receivers* have been designed in England which permit the resonant frequency to be varied. In one type, the variations of the magnetic field operate a vibrating reed attached to a non-magnetic diaphragm. The resonant frequency can be varied by adjusting the position of the vibrating reed with a set screw. Another method of getting rid of resonant points in diaphragms has been to use corrugated instead of flat diaphragms. In some of these diaphragms the center is left flat, the remainder of the diaphragm being corrugated with concentric circles either evenly spaced or spaced in section at radii bearing a ratio to each other corresponding to prime numbers. In some cases conical diaphragms with corrugated edges are used. The material as well as the shape of the horn have been found to materially affect the sound reproduction. Horns are constructed of metal, fibre, composition, hard rubber, wood, etc. In general the horn should be of a material which will not give out metallic sounds. The tone arm which carries the speaker unit at one end and the horn at the other is usually constructed of cast metal such as a lead alloy or aluminum. The requirement here is that it must be of an acoustically "dead" material. In some cases the speaker unit and sound chamber are concealed in an artistically designed cabinet thus forming a *cabinet speaker*. Loud speaker units are also made especially for use in conjunction with the sound chamber of a phonograph. These are referred to as *phonograph attachments*. They are ordinarily attached to the phonograph tone arm being put in place of the phonograph reproducer. Some phonographs are now made with provision for attachment of the speaker unit without removal of the reproducer. There is also an attachment on the market for use in connection with a phonograph which has provision for conveniently changing over from the reproducer to the speaker unit or vice versa. Cone speakers may be of the single cone variety or of the double cone style. In these speakers a metal rod extends from the magnet armature and is rigidly fastened to the center of the cone. The most common type of cone speaker has a parchment diaphragm. This seems to give excellent reproduction especially for tones in the lower register. Its points of disadvantage are that it is affected to some extent by moisture and temperature change and furthermore seems to bring out static interference more clearly than some other types. Another type of cone speaker uses a teak wood diaphragm. Some of these speakers have a small fastening screw at the point of attachment between the rod and the diaphragm. This is provided so that tension can be removed from the diaphragm during shipment and also if the speaker is to remain in a room where change in temperature may take place. A cylindrical parchment speaker which utilizes the same principle as the cone speaker, has recently been placed on the market. Its chief point of difference is that the parchment is not under tension as in the case of the cone speaker. Large volume loud speakers are provided with power amplifiers which are often an integral part of the speaker itself. One of the most recent loud speakers is contained in a large cabinet and contains apparatus for utilizing ordinary lighting circuits for operating the set and speaker. The power amplifier unit is also contained within the speaker cabinet and the design of the speaker is such that it can handle any volume without distortion. The rectifier is arranged to supply not only the necessary voltages for the power amplifier but also plate voltage for practically any type of receiving set. (See *Horn, Loudspeaker,* also *Distortion.*)

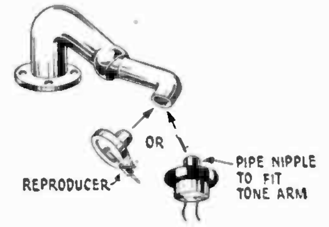

A loud speaker unit used as a phonograph attachment.

LOUDSPEAKER UNIT—The sound reproducing unit attached to the tone arm of the horn type speaker or to the diaphragm of the cone type speaker. Various types of units are described under the heading *Loudspeakers.*

LOW FREQUENCY—See *Low Frequency Current.*

LOW FREQUENCY CURRENT—An alternating current of less than 10,000 cycles. In radio work the alternating currents used are often classified as *low* frequency currents and *high* frequency currents. While there is no fixed value separating high and low frequencies, 10,000 cycles is generally taken as the dividing line. Thus audio frequency currents are considered as low frequencies and radio frequency currents are high frequencies. The term *low frequency* current as applied to power and lighting circuits refers to circuits having frequencies of 25, 60, etc., cycles per second. (See *Current, High Frequency.*)

LOW FREQUENCY IRON CORE INDUCTANCE—A variable inductance having an open-ended iron wire core. This inductance is used with spark transmitters to place the primary transformer circuit in resonance with the alternating current frequency.

LOW LOSS COILS—*Inductance coils* (q.v.) having low resistance and specially wound so as to reduce to a minimum condenser effect and hence dielectric losses. The term low loss has been used rather loosely but for practical purposes coil losses should be taken as referring to self-capacity, often called *distributed capacity* (q.v.) ohmic resistance, *skin effect* (q.v.) in the wire due to the effects of frequency and leakage and absorption in or through the insulating materials upon which the coil is built or with which the wire is covered. While it is impossible to obtain a coil having pure inductance only, it is feasable to design the coil so that the various losses will be reduced to a minimum. Often, it has been found that a design which will greatly reduce one type of loss will result in an increase of another. The illustrations show various types of low loss coils together with curves showing comparative efficiency tests. Figure 1 shows the results of an investigation regarding the distributed

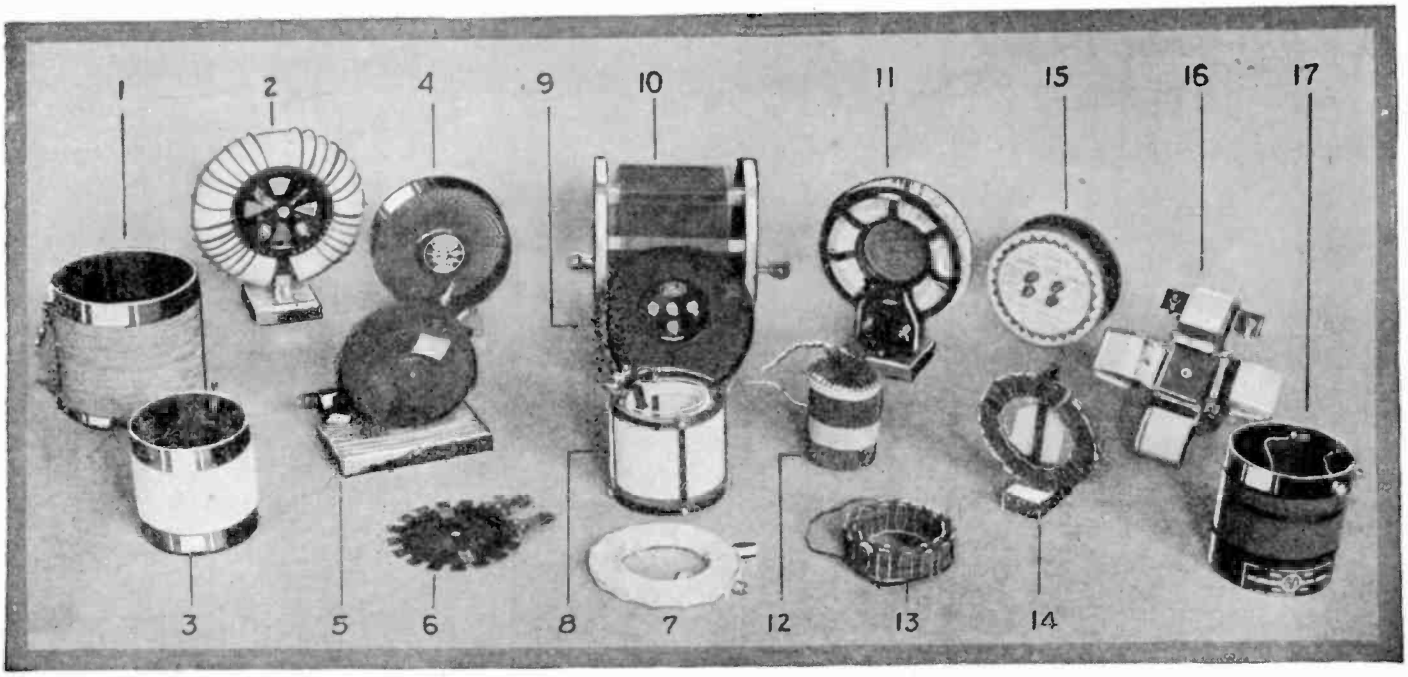

Various types of low loss coils including: 1, Bell wire coil; 2, Naxon; 3, Marco; 4, All-American; 5, Orbit; 6, Turney; 7, Sickles; 8, Aerocoil; 9, Erla; 10, Wavemeter coil; 11, Thorola; 12, Pathe; 13, Freshman; 14, Coast Coil; 15, Summitt; 16, Quadroformer; 17, Bruno.

capacity of various coils. This was obtained by making measurements of

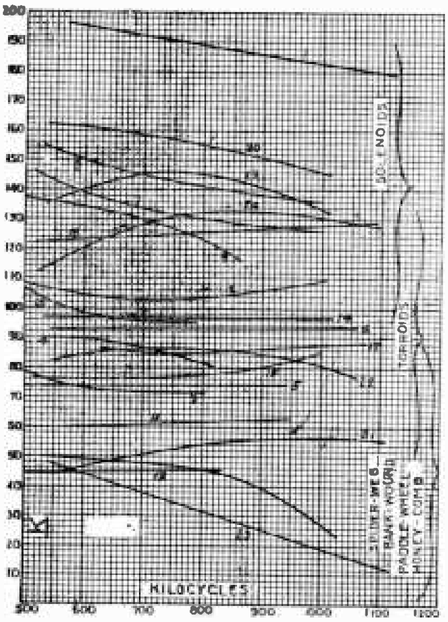

The efficiency of a coil is indicated by the height of its curves on this graph.

the inductance of the coils. If the self-capacity of the coil is very small, the self-inductance will not change perceptibly as the frequency changes. On the other hand, if the self-capacity amounts to anything, the curve showing the relation between inductance

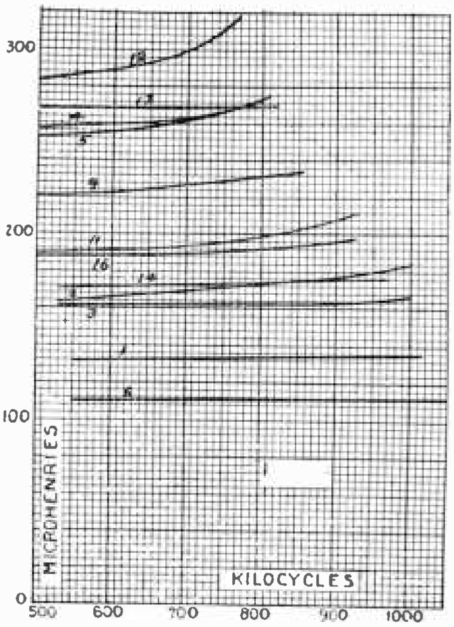

Fig. 1.—The curvature of the graphs indicate the presence of distributed capacity in the coils.

and frequency will bend upward as the frequency increases, and the greater the capacity of the coil, the more sharply will the curve bend. The curves shown in Figure 1 do not give any idea of what value the self-capacity may have, but simply show whether the inductance varies or not, which is one method of determining how harmful is the effect of the distributed capacity. The more sharply the curve bends upward the less desirable is the coil. In general, the smaller the size of the wire (for sizes smaller than about No. 16 B & S) the higher the resistance. The larger the wire and the thinner the insulation, the greater is the skin effect. The greatest inductance is obtained for a given amount of wire when the coil has a true cylindrical shape. The skin-effect in multi-layer coils is much greater than in single-layer coils. The effect of coil capacity and absorption or leakage in insulation is small compared with the skin-effect except in multi-layer coils, where the distributed capacity may become very great. To keep the physical size of toroidal coils within practical limits the diameter of the turns must be relatively small, so that many more turns are required to obtain a given inductance.

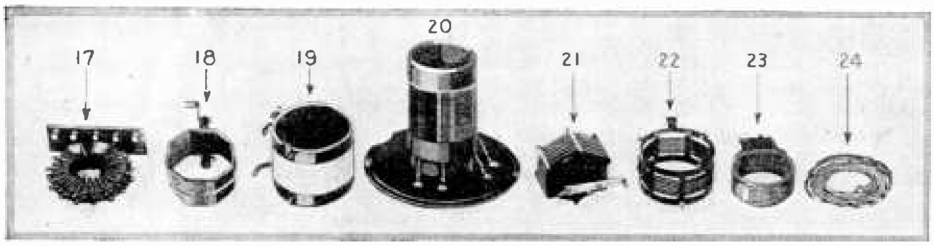

Additional examples of low loss coils: 17, Marwol; 18, Eastern Coil; 19, Workrite; 20, Walbert; 21, Andrews; 22, Bremer Tully; 23, Cotocoil; 24, Kresge Lorenz coil.

LOW TENSION—abbreviation L. T.—A comparatively low voltage. A circuit supplied by a low voltage battery such as the filament circuit of a vacuum tube is called a low tension circuit. In electrical engineering work, any circuit of 600 volts or under is considered to be a low tension circuit.

LUGS—Metal projections used for permanently connecting conductors with electrical apparatus. In one type of lug, the metal is bent over at one end to form a tube into which the conductor can be soldered, while the other end is flat and has a hole drilled to receive any desired terminal.

Typical lugs.

M

MAGNET—A piece of iron or steel having the power to attract other pieces of iron and to attract or repel other magnets. Natural magnets are known as *Loadstones* (q.v.). (See *Electromagnet*, also *Permanent Magnet* and *Temporary Magnets*.)

MAGNET WIRE—Insulated wire for the construction of magnet coils. In radio extremely fine wire is used to wind the coils of the electromagnets used in headsets and loud speaker units. The most common insulation used in radio work for magnet wire is enamel. Other types of magnet wire utilize a single cotton covering, a double cotton covering or a silk covering. (See *Enameled Wire*.)

MAGNETIC ATTRACTION AND REPULSION—The mechanical force tending to draw together two magnetic poles of unlike polarity, or to force apart two magnetic poles of like polarity.

MAGNETIC BLOWOUT—An electromagnet with its pole pieces placed so that the direction of the magnetic field will be at right angles to the movement of the ions in an electric arc. The coil of this magnet is energized in some cases in series with the line and acts to expel or blow out the arc. In the series arrangement, the blow-out action is stronger, the larger the current to be interrupted. For small shunted arcs, permanent magnets are often used. The operating principle of the blowout coil depends upon the fact that in a magnetic field, an electric current tends to move across the lines of force or in other words to cut them. In the illustration, the pole pieces are shown so placed that the general direction of the lines of force is at a right angle to the flow of the ions in the arc. The electromagnetic action causes the arc to move outside the field. This increases the length of the arc, thus rapidly breaking it. Magnetic arcs are used in connection with *keys* (q.v.) used for interrupting a large current, in certain types of *lightning arresters* (q.v.) and also in motor controllers..

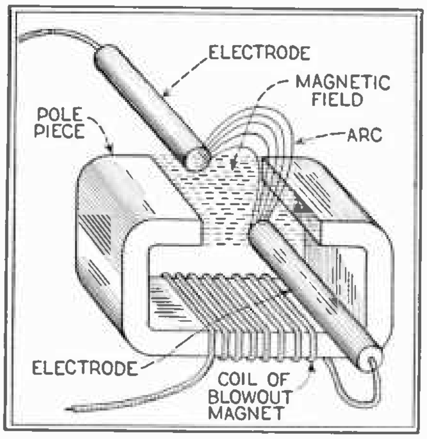

A magnetic blowout.

MAGNETIC CIRCUIT—The path traversed by magnetic lines of force. A magnetic circuit may be entirely through iron, as in the case of a closed core transformer or partly through iron and partly through air as in the case of a motor or generator. (See *Line of Force*.)

MAGNETIC COUPLING—See *Coupling*.

MAGNETIC COUPLING TRANSFORMERS—See *Coupling Transformer*, also *Jigger*.

MAGNETIC FIELD—The whole space over which a magnet exerts influence. The space traversed by the magnetic flux. Wherever magnetism is present the space in the immediate vicinity is called a magnetic field. (See *Electromagnetic Induction.)*

MAGNETIC FLOW—The passage of magnetic lines through a magnetic circuit. This term is sometimes used in the place of magnetic flux. (See *Flux Magnetism.)*

MAGNETIC FLUX—Symbol ϕ—The total magnetism present at any cross-section perpendicular to the lines of force. Magnetic flux is analagous to current flow. That is to say, it is convenient to consider a flow of magnetism in the same way as electricity is considered to flow. The unit of magnetic flux in the c.g.s. (centimeter gram second system) electromagnetic system is the *maxwell* (q.v.) or simply the *line*. The unit of magnetic flux in the practical system is also the *maxwell.* (See *Line of Force, Alternator,* also *Flux, Magnetism.)*

MAGNETIC FLUX DENSITY—The number of lines of force per unit cross-sectional area of a magnetic path. The *gauss* (q.v.) is the unit of magnetic flux density. It is equal to one *line* (q.v.) per square centimeter.

MAGNETIC FORCE—The force with which a magnet attracts or repels any piece of iron or steel. The intensity of a magnetic field at any point as measured by the force which it would exert upon a unit magnetic pole if placed at that point.

MAGNETIC FORCE DUE TO CURRENT—A magnetizing force due to current flow which can be determined mathematically for any point in the surrounding medium. Consider any closed stream line of electric current surrounded by a medium having uniform magnetic properties. It can be demonstrated that each elementary length of the current stream line, dl, contributes to the magnetizing force H, at any point T an amount

$$dH = \frac{(I \sin \Theta)\ dl}{r^2}$$

where I is the current flowing along this stream line, r is the distance from T to dl, and Θ (theta) is the angle between r and dl. The direction of dH is perpendicular to the plane of r and dl. The total magnetizing force at T is the vector sum of dH for all the elementary lengths into which the current stream line is divided.

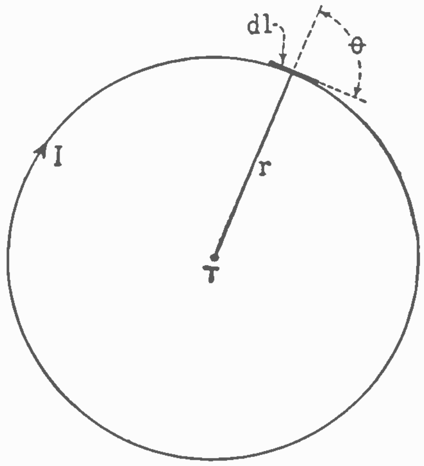

Diagram illustrating method of determining magnetic force due to current.

MAGNETIC HYSTERESIS—The opposition which minute particles of a magnetic material offer to being magnetized or to having their magnetization changed. It is sometimes referred to as molecular friction or as *magnetic lag.* (See *Hysteresis.)*

MAGNETIC INDUCTION—See *Induction,* also *Diamagnetic Material.*

MAGNETIC INDUCTIVE CAPACITY—symbol μ (mu)—Also termed Magnetic Inductivity—Names sometimes used instead of *Permeability* (q.v.). The ratio between the number of lines of force per unit area passing through a magnetizable substance and the magnetizing force which produces them. Stated another way, it is the ratio of flux density to magnetizing force. It can be considered as a measure of the ease with which magnetism passes through any substance. The magnetic inductive capacity of air is taken as unity. The term magnetic inductive capacity is analogous to *specific inductive capacity* (q.v.) which is the ratio of the dielectric constant of a material to that of a vacuum.

MAGNETIC INTERRUPTER—A device energized by an electromagnet for interrupting a circuit. (See *Interrupter.)*

MAGNETIC KEY—An electromagnetic adjunct to a wireless transmitting key for reducing or eliminating sparking due to the self-induction of the heavy currents which are interrupted. A form of magnetic key is shown in the illustration. Underneath the lever arm of the key is placed a spring to which an iron armature C and a platinum contact A_1 are attached. If the key and hence the spring are pressed down contact A_1 touches contact A_2 and the path of the current is completed. The current passes not only through the key, but also through the winding of the electromagnet. If the key is released so as to move upward, the contact A_1 still remains touching A_2, as the magnetic action due to the current continues to hold down the armature C and hence the spring with the contact A_1. Not until the alternating current reaches a zero value is the armature released. This releases A_1 from A_2, at an instant of zero current and hence there is no spark.

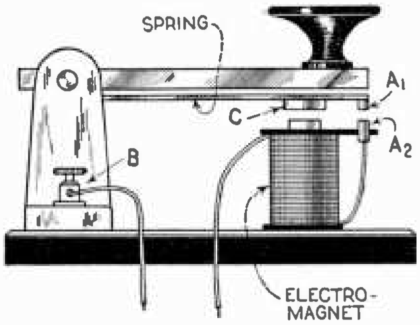

A magnetic key.

MAGNETIC LEAKAGE—Magnetic flux which serves no useful purpose. There are no substances which will prevent the passage of magnetic flux. This fact introduces a certain amount of difficulty into practical magnetic circuit calculations, since it is impossible to confine all the flux into a predetermined magnetic path. Since there is no magnetic insulator there will always be a certain amount of *leakage flux.* However this can be reduced to a minimum by proper design of the magnetic circuit. Magnetic Leakage is also referred to as *Stray Flux.*

MAGNETIC LINES OF FORCE—See *Line of Force,* also *Electromagnet.*

MAGNETIC MOMENT—The product of the strength of pole of a magnet and its virtual length. The torque exerted by a magnetic field upon a magnet depends upon the magnetic moment.

MAGNETIC NEEDLE—A light, thin steel magnet mounted on a pivot or suspended so as to be free to move. It takes a position pointing to the magnetic north.

MAGNETIC POLES—Points on a magnet where the lines of force leaving or entering the iron are concentrated. In a magnetic needle, that end which tends to point north is called a north or positive pole; that end which tends to point south is called a south or negative pole. A unit magnetic pole is defined as one which, when placed at a distance of one centimeter from a like pole of equal strength, will repel it with the force of one dyne. The north magnetic pole of the earth is situated in latitude 70 North and longitude 97 West. The south pole is located at latitude 70 South and longitude 102 East. It should be noted that the magnetic poles do not coincide with the geographical poles. (See *Induction, Magnetic.)*

MAGNETIC PROPERTIES—Materials are classified according to whether or not they offer a good path for the magnetic lines of force. Substances of comparatively low *reluctance* (q.v.) are termed *magnetic.* They are also often referred to as *ferro-magnetic.* They include iron, steel, nickel, cobalt, manganese, chromium, magnetite, certain alloys of copper, manganese and aluminum (called Heusler Alloys), and certain other oxides. The magnetic laws governing these substances can best be set forth by means of magnetization curves. Substances having high reluctance are termed *non-magnetic.* For commercial purposes, air is the most important non-magnetic substance. However, practically all substances, with the exception of the ferro-magnetic ones, follow the same magnetic laws as air. While the reluctance of a non-magnetic substance is constant, the reluctance of magnetic substances varies, as the amount of *flux* (q.v.) in the substance changes. Although air and all other substances, with the exception of the ferro-magnetic ones are classed as non-magnetic substances, and are poor conductors of magnetic flux, when compared to iron, etc., nevertheless they are by no means magnetic insulators. There are no substances which will absolutely prevent the passage of magnetic flux. (See *Ferro-Magnetic Substances.)*

MAGNETIC PYRITES—See *Pyrites.*

MAGNETIC STORM—Sudden and irregular variations of the earths magnetic field. Magnetic storms are sometimes coincident with the appearance of sun spots. These storms often disrupt telegraph service and also cause much interference with radio communication.

MAGNETISM—A phenomena exhibited by certain materials which is characterized by the attraction or repulsion of soft iron or of conductors carrying electricity. Magnetism is one of the manifestations of electricity, since a current flow is always associated or linked with a *magnetic field* (q.v.). (See *Free Magnetism, Magnet,* also *Magnetic Properties.)*

MAGNETISM, LAWS OF—Like magnetic poles repel one another; unlike magnetic poles attract one another. The force exerted between two magnetic poles is proportional to the product of their strengths and is inversely proportional to the square of the distance between them. (See *Like Poles.)*

MAGNETITE — symbol Fe_3O_4 — A mineral composed mainly of magnetic oxide of iron. When found in magnetized condition it is known as a *loadstone* (q.v.) or a natural magnet. It is a black brittle substance having a specific gravity of 5.2.

MAGNETIZATION—(also spelled *Magnetisation*)—The process of communicating magnetism. The act of rendering iron, steel, etc., magnetic. According to Ewing's theory, a substance which is magnetized has its molecules rearranged so that they assume symmetrical positions, each molecule lying in line with or parallel with its neighbor. In this rearranged position each molecule adds its separate magnetic force to every other one and the cumulative effect of this is that the substance exhibits the property of a magnet. (See *Demagnetization.*)

MAGNETO GENERATOR—A small generator consisting essentially of an armature rotating between permanent magnets, usually of the horse-shoe type. Magnetos are mainly used for motor ignition, medical coils, bell ringing and for firing high explosives. (See *Intermittent Current.*)

MAGNETOMOTIVE FORCE OR DIFFERENCE OF MAGNETIC POTENTIAL—abbreviation m.m.f.—The force causing magnetic flow. *Magnetomotive force* is the magnetic cause, while *flux* (q.v.) is the magnetic effect, just as in electrical terms, *electromotive force* (q.v.) is the cause and current is the effect. The c.g.s. (centimeter gram second) unit of magnetomotive force is the gilbert (q.v.). The unit ordinarily employed in practical calculations of the magnetic circuit is the *ampere-turn* (q.v.). (One ampere-turn is equal to 1.2566 gilberts.) An ampere-turn consists of a current of one ampere flowing through one complete turn of a conductor. The expression "turn" means that the conductor is wound in a solenoid. That is to say, it is wrapped around a core which may consist of air or of any other substance. The same magnetic field will result from ten ampere-turns consisting of ten amperes flowing through one turn, as from ten ampere-turns consisting of one ampere flowing through ten turns. (See *Force, Magnetomotive.*)

MAGNIFIER, OR RELAY—An apparatus for intensifying the variations of a current. The term *magnifier* has been incorrectly used for *amplifier* (q.v.). One type of *magnifier* used in radio is the microphonic relay (see loud speaker) by means of which the minute currents in a crystal set can be magnified so as to actuate a loud speaker. In the ordinary telephone repeater, a microphone is acted upon by the receiver of the first line circuit so as to introduce a local circuit which acts inductiveley on the second line wire. Relays are also used in wire telegraphy to boost weak line currents by means of strong local batteries. (See *Relay.*)

MANTISSA — The decimal part of a *logarithm* as differentiated from the integral part or characteristic. For example in the logarithm 1.23578, 1 is the characteristic and the decimal .23578 is the mantissa. (See *Logarithms.*)

MARBLE—The name given to any limestone sufficiently compact to permit being polished. Pure marble is white, the presence of oxides giving the different colors. Marble is used extensively for switchboard work and when so used must be free from metallic veins. Where circuits carrying 1000 volts or more are used, the switchboard should be saturated with an insulating varnish and baked. Inasmuch as the marble shows oil spots, it is sometimes stained black and given a finish called a *marine finish.*

MARCHANT, EDGAR WALFORD. British radio expert. Born in 1876, he was educated at University School, Hastings, and Central Technical College and London University. In 1897 he was appointed superintendent of the laboratory and workshops of Lord Blythswood at Renfrew, where he carried out many experiments in wireless telegraphy. He became chief assistant at Finsbury Technical College, 1900, under Professor Silvanus P. Thompson, and the following year was appointed lecturer in electro-technic at University College, Liverpool. In 1903 he was appointed the first professor of electrical engineering at the university.

Marchant was for some time closely associated with W. Duddell in developing the oscillograph, and the two read a joint paper on the study of the electric arc by the aid of oscillographs to the Institution of Electrical Engineers, of which institution Marchant became vice-president. He is also a vice-president of the Radio Society of Great Britain. He has written many papers on radio, and contributed an article to the proceedings of the Royal Society on the magnetic behaviour of iron under oscillatory discharge of a condenser.

MARCONI FILINGS COHERER—Also called Marconi Coherer—A detector of electric waves consisting of fine nickel and silver filings contained in a suitable glass tube between two electrodes consisting of two silver plugs. These two plugs are connected to platinum wires brought out through the sealed ends of the tube, which is carefully exhausted of its air. The metallic filings in a loose condition have feeble conductivity for the current from a single dry cell. If a very weak electric oscillation is passed through the coherer, the filings cling together or cohere and hence conduct electricity better and they will then pass sufficient current from a single dry cell to operate a telegraphic relay. To bring these filings back to a non-conductive condition, the tube is tapped by an electromagnetic tapper, similar to an electric bell with the gong removed. The Marconi coherer is now obsolete. (See *Coherer.*)

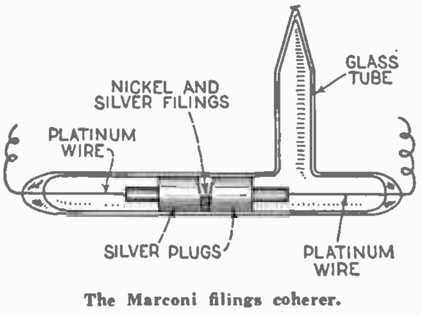

The Marconi filings coherer.

MARCONI, GUGLIELMO—Italian-Irish radio pioneer. Guglielmo Marconi was born at Bologna, Italy, April 25, 1874, and was educated at the Leghorn Technical School, under Professor Rosa. From his earliest years he was keenly interested in communication by means of Hertzian waves, and began his brilliantly successful series of experiments in June, 1895. These first experiments were carried out on his father's estate near Bologna. Marconi soon found that the Hertzian form of resonator gave only feeble signals at a distance, and he substituted a vertical wire, with the result that in 1895 he was able to transmit signals to a distance of one and a half miles. About the same time he improved the Branly coherer, which he was using as a detector, and invented an electric tapping device to decohere the filings.

This early apparatus of Marconi, from which has sprung the far-flung wireless chain of to-day, consisted of a coherer, a relay, a decoherer, and a Morse printing instrument, all working from storage batteries. Between the coherer and the relay Marconi interposed choke coils, and this had a very marked effect on the receptivity of his set. By close attention to the details of his system he was able to carry on his signals at greater ranges than had up to this time been accomplished by other experimenters.

The transmitting apparatus used by Marconi in these early efforts consisted of a large spark gap to which the aerial and earth wires were connected. The high-tension current for the spark was provided through an induction coil from batteries. The spark gap consisted of a ball discharger comprising four brass balls. The two middle balls were separated by a small space filled with vaseline oil, the actual spark jumping from the two end balls to the middle ones and through the vaseline, producing a high-frequency spark.

Marconi came to England in 1896, where he took out the first patent ever granted for a practical system of wireless telegraphy. He made a number of experiments at Westbourne Park, and demonstrated his system before Sir William Preece and other high officials at the Post Office. In the following year he increased the range of his set to nine miles using a 29 in. spark coil and kites to raise the vertical aerials. In July, 1897, in demonstrating before the Italian Government, he covered 12 miles between warships, and he began to instal a number of his sets for lighthouses, and the success of his experiments led to stations being erected for the corporation of Trinity House.

Guglielmo Marconi.

In 1899, the first proof of the advantages of wireless over other forms of communications came with the saving of the lives on board the ship R. F. Matthews, which ran into the East

Goodwin lightship. The latter was equipped with one of Marconi's transmitting sets and was able to communicate with the South Foreland lighthouse and summon assistance.

The genius of Marconi lies not so much in his inventions, as in his far-sightedness in being the first man to realize the immense commercial possibilities of radio, and to make the best use of all the scientific effort of his time to further the object he had in view.

Marconi has received innumerable honors from all countries. He holds the honorary degrees of universities all over the world, and the freedom of a large number of cities. In 1909 with Professor Braun, he was awarded the Nobel Prize for Physics, and in 1914 he was given the Honorary Knighthood of the Grand Cross of the Victorian Order. He has been awarded the Albert Medal of the Royal Society of Arts, the Gold Medal of the Institute of Radio Engineers of New York, the John Fritz Gold Medal for the invention of wireless telegraphy, and the Franklin Gold Medal of the Franklin Institute. He was the Italian delegate to the Peace Conference, and signed on behalf of Italy. He is chairman of the board of directors of the Marconi Company.

MARINE TYPE CHARGING PANEL—A switchboard equipped with resistance coils, carbon filament pilot lights, voltmeter and necessary switches. Manipulation of the switches allows the battery to be placed on charge or on the line. By throwing a switch on this board, it is also possible to work the wireless transmitter either from the ship's generator or from the storage battery.

MASS—The quantity of matter which a body contains. Physical quantities such as force, velocity, etc., are expressed in terms of length, *mass*, and time. The unit of length is the centimeter; of mass, the gram; and of time, the second. Hence this system of measurement is known as the C.G.S. (centimeter gram second) system. The gram is equal to 15.432 grains and represents the mass or quantity of a cubic centimeter of water at 4 degrees centigrade.

MASTER-OSCILLATOR SYSTEM—A radio telephone transmitting circuit in which a small extra *audion* (q.v.) is used to supply the power to excite the grid circuit. This extra oscillator or exciter is referred to as the master-oscillator. The master-oscillator system has several advantages over the self-excited system. Changes in antenna constants, such as might be caused by a swaying aerial or *lead-in* (q.v.) do not affect the transmission where the former system is used. In addition the master-oscillator system is more convenient to work with and the adjustment for maximum output for different wave lengths and antenna resistances is more easily made.

MASTER-OSCILLATOR SYSTEM AMPLIFIER—In the master-oscillator system, the power tube acts as an amplifier of the power supplied by the small extra exciting, or master-oscillator. The exciter must develop sufficient power to supply the losses in its own oscillating circuit and those of the grid circuit of the power tube.

MASTS, ERECTION OF—High powered stations generally use latticed steel masts. In some cases tubular metal masts in telescoping sections are used. Wooden masts have been used to quite an extent. For portable outfits, these are made in sections which can be fitted together like a fishing rod. Guy ropes or wires are necessary in practically every instance. In erecting a mast, it is usual not to make the installation too rigid, but rather to allow some freedom of movement so that the mast may sway slightly in the wind. Where a one-wire aerial is being erected this may be fastened to any support available. In marine construction, the ship's masts are used to support the aerial, the wires being stretched between the two booms or spreaders. from which halyards run to the masts. (See *Aerial.*)

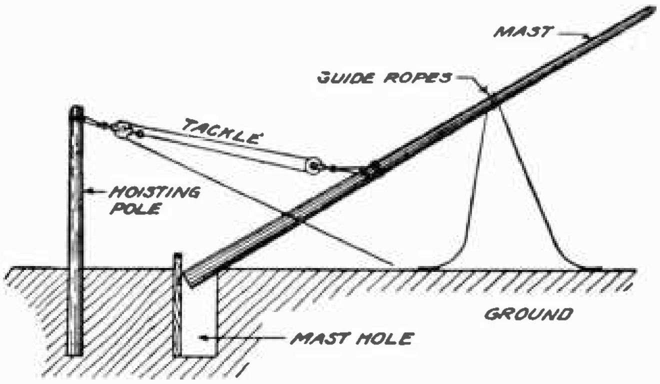

A method of erecting a tall mast using tackle and hoisting pole.

MATCHED IMPEDANCES—Circuits balanced or matched so that the total impedance of one circuit will equal the impedance of another circuit with which it is coupled. (See *Impedances, Matched.*)

MAXWELL—symbol φ—The unit of *magnetic flux* (q.v.). It is also known as the *line* (q.v.). One maxwell is equal to one line of force. (See *Flux Density*, also *Kiloline.*)

MAXWELL'S CORKSCREW RULE—If the direction of travel of a right-handed corkscrew represents the direction of current in a straight conductor, the direction of rotation of the corkscrew will represent the direction of the magnetic lines of force.

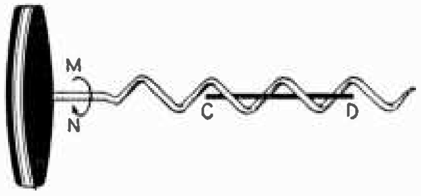

Illustration of Maxwell's corkscrew rule. When current flows from C to D, magnetic lines encircle the wire in the direction of the curved arrow MN.

MAXWELL'S LAWS—The magnetic fields surrounding parallel currents in the same direction will react upon each other in such a way that the conductors will tend to move together. In the case of parallel currents in opposite directions, the fields between the two conductors will be in the same direction and will merge together, thus tending to push the conductors apart. The magnetic fields around two conductors at an angle will react upon each other in such a way as to tend to bring the conductors parallel, and with the currents flowing in the same direction.

MAXWELL'S LAW OF MOTOR ACTION—A conductor carrying a current whose direction of flow is at right angles to a magnetic field, will tend to move out of this field.

MAXWELL'S RULE—Every portion of a circuit is acted upon by a force urging it in such a direction as to make it enclose the greatest possible number of lines of force.

MAXWELL TURNS—The deflection value in a ballistic galvanometer or *fluxmeter* (q.v.). This depends upon the number of turns in the exploring coil and the instantaneous value of the flux in maxwells linked with this coil. (See *Grassot Fluxmeter.*)

McLACHLAN, NORMAN W.—British radio authority. Born at Longtown, Cumberland in 1888, and educated at Carlisle Grammar School and the George Watson and the Heriot Watt Colleges, Edinburgh, and Liverpool University. He was appointed lecturer in Engineering and Mathematics at Newcastle-on-Tyne in 1909. In 1913 McLachlan was appointed supervisor of classes in engineering subjects in the Liverpool Technical Institutes, and after the Great War, during which he was engaged in aeronautical research and the study of anti-submarine devices, he made a special study of magnetos at the National Physical Laboratory, Teddington. Appointed research engineer to the Marconi Company, he is the author of many papers on wireless and electrical subjects in the journal of the Institution of Electrical Engineers and other scientific journals, including those on Characteristic Curves of a Poulsen Arc., the Magnetic Behaviour of Iron at very High Frequencies; and Theory of Iron-cored High Frequency Transformers. He has taken out a number of radio patents.

MEASUREMENT OF INDUCTANCE—Some methods of measuring inductance (also applicable to the measurement of *capacity*) are various bridge methods; by measuring the voltage, current, power, and frequency; and by the wave meter method. Connections in one bridge method for measuring inductance are similar to those in the ordinary *Wheatstone Bridge* (q.v.). A coil having a known inductance is balanced against the inductance to be measured. A device called a *Secohmmeter* is in use to increase the sensibility of bridge measurement of inductance. The Secohmmeter serves the purpose of making an alternating current to use in measurements of self-induction and of commutating such portion of this current as flows in the galvanometer circuit to a direct current. When a source of alternating current of known frequency is available the following method is convenient. Place the inductance across the alternating current mains, measure the power absorbed, the current flowing, the voltage across the unknown inductance and the frequency of the circuit. Then the inductance L is equal to the reactance, x, divided by

2 π times the frequency. Expressed as a formula

$$L = \frac{x}{2\pi f} = \frac{\sqrt{\left(\frac{E}{I}\right)^2 - \left(\frac{\text{Power}}{I}\right)^2}}{2\pi f}$$

The inductance L should be expressed in henries. This method gives results of fair accuracy but is not adapted to measurements of low values of inductance. Small inductances are more usually measured by bridge methods. In the wave meter method, a condenser is discharged through a spark gap in a circuit containing a known capacity and by means of a wave meter measuring the wave length. The expression for the number of oscillations in a closed circuit containing both inductance and capacity is given by the formula

$$n = \frac{1}{2\pi\sqrt{CL}}$$

and the wave length is equal to the velocity of propogation divided by the number of waves and calling λ the wave length we have

$$\lambda = V \text{ times } 2\pi\sqrt{CL}$$

In this last equation everything is known but the inductance which can be calculated. (See *Inductance*, Measurement of.)

MEASUREMENT OF WAVE LENGTH—The usual method of measuring an unknown wave length is by means of a wavemeter. Other apparatus required in this connection is a hot wire ammeter indicator in a coupled circuit, or crystal detector and telephones. Also an antenna excited oscillating circuit supplied by electron tube. A wavemeter placed in inductive coupling with a coil or antenna carrying radio current will show decided increase of current in its own coil and condenser when it is tuned to resonance with the source. The wave length is read directly from the wavemeter setting for resonance or from a calibration curve. A receiving set can be used to measure the wave lengths of received waves if it is first standardized in terms of wave lengths. The operator listens in through the head set attached to the receiving set. As the wavemeter condenser knob is turned the loudest sound is heard when the wavemeter circuit is tuned to the same wave length as that for which the receiving set is adjusted. The wave length is then read from the wavemeter scale or calibration curve. Continuing in this manner, the receiving circuit can be calibrated as a wavemeter by setting it at many different adjustments and reading the wave lengths at resonance each time. The wave length of the waves emitted by an antenna, when no added inductance or capacity is inserted in the antenna circuit, is known as its fundamental wave length. For a single vertical wire grounded antenna the fundamental wave length is slightly greater than four times the length of the wire. The approximate constant often used is 4 2 and this applies also to "L" or "T" type aerials with vertical lead in wire, the total length being measured from the transmitting apparatus up the lead-in wire and over to the end of the flat top. It is of course, more accurate to measure the wave length radiated from an antenna directly by use of a wavemeter. The wavemeter coil needs merely to be brought near the antenna or lead-in wire and the condenser of the wavemeter adjusted to give maximum current in the wavemeter indicator. The wave length corresponding to the wavemeter setting is then the length of the waves radiated by the antenna. The fundamental wave length of the antenna may be determined by gradually decreasing the number of turns in the *loading inductance* (q.v.), measuring the wave length for each setting of the loading coil and plotting a curve showing the wave length corresponding to the various numbers of turns of the loading coil. The fundamental, then, is the wave length corresponding to zero turns and is found at the point where the extension of the curve cuts the wave length axis. (See *Measurement of Inductance*, also *Formula*.)

MECHANICAL MIXTURE— A mixture of two or more elements which leaves them essentially the same. This is sometimes incorrectly called a mechanical compound. Sugar and sand mixed together give an example of a mechanical mixture. A *chemical compound* is formed when two or more substances combine chemically to form another substance having different characteristics. Thus iron filings and sulphur can be combined to form a third substance unlike either of the original two.

MEG OR MEGA—symbol M—The prefix placed before the name of a unit to denote one million times (10^6) that unit, as for example *megohm* (q.v.) *mega-volt, megerg*, etc.

MEGGER—An instrument for measuring insulation resistance or in general any resistance by the direct deflection of a pointer on a scale. The "Evershed" type is made in two forms (1) the *megger* insulation testing set and (2) the bridge-megger testing set which combines the functions of type (1) with those of a *Wheatstone Bridge* (q.v.). Both types consist of an ohmmeter of the moving coil type combined in one box with a hand driven generator for providing the necessary testing voltage and current. When the handle of the generator is not being turned, the pointer is entirely free and will rest anywhere on the scale. The general principles underlying the construction and operation of the megger are shown in the illustration. As can be seen, the megger is a combination of a *magneto generator* (q.v.) shown at the right, with an ohmmeter shown at the left, in a magnetic circuit common to both and consisting of two pairs of field poles braced by strong bar magnets, NS, NS, and forming two bi-polar field magnets in series. In the right hand one, and rotated by a folding handle and spur gearing D, is the armature of the generator, with its brush gear B, and terminal bars marked + and —. In the left hand field is the current coil A, pressure coil P, and compensating coil C of the ohmmeter, connected to resistances Q and R, a *guard plate* G, and the only two external terminals L and E (marked Line and Earth on the actual instrument). The bridge-megger testing set is available for use both as an insulation testing set and as a specialized Wheatstone Bridge. It differs from the ordinary megger in outward appearance only by the addition of two pairs of terminals and of two switches, one of which is the *ratio* switch of the Wheatstone Bridge and the other one is a two-way change over-switch. This latter when set to "Megger" prepares the instrument for measuring insulation resistance or high metallic resistance by coupling the two windings of the constant voltage generator in series and making the two front terminals, marked Line and Earth, the only two available for connection. When the change-over switch is set to "Bridge" the instrument is ready for Wheatstone Bridge work, the two windings of the generator now being connected in parallel in order to increase the current obtainable from it. In this case the ohmmeter part is changed into a galvanometer for the bridge and the arms of the bridge are switched into their appropriate places in the circuit and connected to the terminals marked R and X at the end.

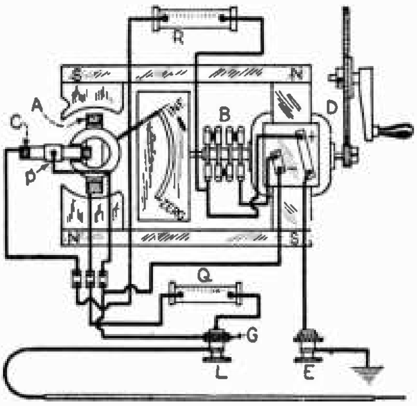

A diagrammatic illustration of a megger.

MEGOHM—One million ohms. This unit is usually employed in the measurement of insulation resistance. It is also used in measuring *grid leaks* (q.v.), etc. See *Ohm*, also *Begohm.*)

MEGOHMITE, also spelled Megohmit—An insulating material similar to micanite, obtainable in several forms, one of which is known as *hard megohmite* and is made of thin sheets of mica built up of shellac. *Flexible megohmite* consists of sheets of mica stuck together by vegetable adhesives. Flexible megohmite covered with Japanese paper is called mica paper while flexible megohmite covered with linen, is called *mica linen.*

MEISSNER, ALEXANDER.— Austrian radio expert. Born at Vienna, 1883, he was educated at the Technical High School and University of Vienna. He joined the Telefunken Company in Berlin, 1907, and is one of the leading authorities in Germany on radio. To him is due a large number of wireless inventions, including the Telefunken compass, musical quenched sparks, interference preventers, the direct current cathode valve relay for the reception of Morse Code, etc. In 1913 he invented the generation of oscillations by means of the three-electrode valve. He holds many patents, including spark gap for impulse excitation, quenched-spark signalling in conjunction with G. Von Arco, etc., and he is the author of many papers for scientific journals and societies on the subject.

MERCURY ARC RECTIFIER (MERCURY VAPOR CONVERTER)—A device for changing alternating current to direct current, which depends for its action upon the unidirectional conductivity of the mercury arc in an exhausted vessel. An idea of the principle of operation of the mercury arc rectifier may be obtained from the illustration. Assume an instant when the terminal H of the supply transformer is positive, the anode A is then positive and the arc is free to flow between A and B, B being the mercury cathode. Following the direction of the arrows still further, the current passes through the lead J, through the reactance coil E and back to the negative terminal G on the transformer. A little later when the impressed electromotive force falls below a value sufficient to maintain the arc against the counter electromotive force of the arc and the lead. the reactance E which heretofore has been charging, now discharges, the discharge current being in the same direction as formerly. This serves to maintain the arc in the rectifier until the electromotive force of the supply has passed through zero, reverses and builds up to such a value as to cause A' to have a sufficiently positive value to start an arc between it and the mercury cathode B. The discharge circuit of the reactance coil E is now through the arc A'B, instead of through its former circuit. Consequently the arc A'B is now supplied with current, partly from the transformer and partly from the reactance coil E. The new circuit from the transformer is indicated by the arrows enclosed in circles. The amount of reactance inserted in the circuit reduces the pulsations of the direct current sufficiently for all ordinary commercial purposes. The mercury rectifier is used in small sizes for charging storage batteries from alternating current circuits and in the larger sizes for operating direct current arc lamps from alternating circuits and also in electric traction work in the place of rotary converters in the substations or even on the train itself.

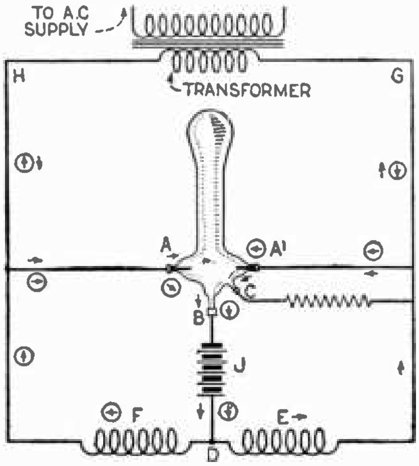

Illustration showing action of mercury arc rectifier.

MERCURY INTERRUPTER (MERCURY JET INTERRUPTER)—A contact breaker in which contact is made between moving metallic contacts and a jet of mercury provided by a small centrifugal pump driven by the same motor as the contacts. When the primary winding of an induction coil is supplied with a continuous current it is necessary to interrupt the current in order that a secondary voltage shall be induced in the secondary winding. The mercury interrupter offers a satisfactory method of accomplishing this. In one type of mercury interrupter a stream of mercury is forced from a jet against a metal plate. The stream of mercury is interrupted by a toothed wheel of insulating material so that the electrical circuit is periodically made and broken between the mercury and the metal plate. In some forms of mercury interrupters, the jet itself revolves, the mercury impinging against a fixed metal plate. The mercury interrupter is used in operating X-ray bulbs from an induction coil where the source is direct current. A form of mercury interrupter is shown in the illustration. A is a tooth-shaped wheel made of insulating material. This wheel is driven by a small motor and the number of revolutions can be varied within wide limits. A small centrifugal pump forces the mercury upon the jet B, while the revolving wheel A interrupts the contact of the mercury with the plate C. The wheel A can be raised or lowered, thereby enabling the duration of contact of the mercury with the plate to be varied without altering the number of interruptions in a given time.

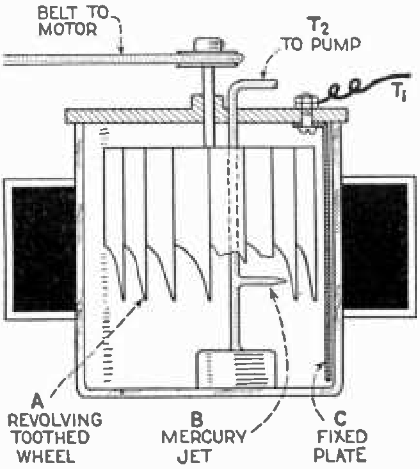

A type of mercury interrupter.

MESH GROUPING—A method of connecting coils in a polyphase circuit whereby they form a closed circuit and have the line wires attached to the points of junction between the coils. The term mesh is used interchangeably with delta. (See *Delta Connection.*)

METER—An instrument used to make measurements. (See *ammeter* or *amperemeter, double range meters, integrating wattmeter, voltmeter, wavemeter, wattmeter.)*

METER (*Measure*)—A unit of length in the *metric system* (q.v.). One meter is equal to 100 centimeters or 39.37 inches. It is equal to the length of a standard platinum bar, kept in Paris, and representing approximately a ten-millionth part of a quadrant of the earth's meridian measured from the equator to the pole through Paris. Wireless waves are measured in meters. The length of a wave is the distance, usually measured in meters between two points in the successive waves where the disturbance is at a maximum or at a minimum or between any two points of equal disturbance. There is a fixed relationship between the length of a single wave, the frequency of oscillation and the velocity of the wave. Thus if 3,000,000 waves pass a given point in one second, each wave must be 100 meters long, since the velocity of radio waves is equal to 300,000,000 meters per second.

METER-AMPERES—The product of the antenna current in amperes at the point of maximum current and the antenna effective height in meters, for any radio transmitting station. It constitutes a factor for indicating the radiating strength of radio transmitting stations.

METRIC SYSTEM—A system of measurement in which the meter is the fundamental unit of length and in which all the units both fundamental and derived are divided decimally and higher units are formed in multiples of ten. (See *Centimeter Gram Second.*)

MFD—Abbreviation for *Microfarad* (q.v.)

MICA—An anhydrous silicate of aluminum and potash or sodium. Mica has very high insulating qualities and can withstand high temperatures. Mica obtained in the natural state is separated into laminations which are sorted and graded. These are then cemented together to form plate or flexible mica of any desired thickness or purity. Mica is used in radio work as the dielectric of small fixed condensers. It is also used as the sound producing diaphragm in certain types of head sets and *loud speakers* (q.v.).

MICABOND CLOTH—India sheet mica faced on one side with muslin and on the other side with Japanese insulating paper bonded together with a special binder. It is used for insulating field coils and transformer coils.

MICABOND PAPER—India sheet mica faced on one side with Japanese insulating paper of one grade and on the other side with paper of a different grade.

MICABOND PLATE—Mica sheets bonded together with orange shellac. Various grades of plate are made some of which are used for magnet spools, others for insulating of commutator segments, while still others are used in electric irons.

MICANITE—A form of reconstructed mica. Canadian amber mica and white India mica are used.

MICRO—A prefix placed before the name of a unit to denote one-millionth part of that unit, as for example *microfarad* (q.v.).

MICRO-AMPERE—One-millionth of an ampere. A unit used in measuring extremely small currents. (See *Ampere.)*

MICROFARAD—abbreviation Mfd.—A unit of electrical capacity equal to one-millionth of a farad. Since the farad is too large for most practical measurements, the microfarad is generally used. (See *Farad,* also *Jar.)*

MICROHM—A unit of resistance equal to one-millionth of an *ohm* (q.v.).

MICROMETER—An instrument for making minute measurements. Usually controlled by an accurate screw of fine pitch.

MICROMETER SPARK GAP—A minute adjustable spark gap of about four-hundredths of an inch placed in the aerial circuit of a multiple tuner to allow heavy charges to pass readily to earth by sparking across its points.

Courtesy of Western Electric Co.

An example of the microphone used in modern broadcasting.

MICROPHONE—A sound magnifier. The ordinary form of microphone consists essentially of a diaphragm set in vibration by the sound waves and causing by its motion variation in the resistance of a mass of loosely packed carbon granules. This variation in resistance causes corresponding variations in the current through the instrument so that at every instant the form of the current corresponds to the variations in sound. This type of microphone is often referred to as a *Carbon Microphone*. The telephone transmitter is a form of microphone, and the microphone (nicknamed *Mike*) used in radio broadcasting is merely a larger and more sensitive variation of the same instrument. (See *Electromagnetic Microphone.*)

MICROPHONE, CARBON—See *Microphone*, also *Granular Carbon*.

MICROPHONE TRANSMITTER—See *Microphone*.

MICROPHONIC JOINT—A loose contact between two solid conductors, so mobile that feeble vibrations vary its resistance.

MICROPHONIC RELAY—An instrument for magnifying the variations of a given current by passing it through an electromagnet that acts on the diaphragm or reed of a microphone. The resistance variations are repeated in the current produced by a constant voltage in a circuit in series with the *microphone* (q.v.). In general a microphonic relay can be described as a microphone combined with a telephone so that a message transmitted over the telephone is repeated by the microphone over another line.

MIL—A unit of length equal to one-thousandth part of an inch.

MIL-FOOT—A wire one foot long having a diameter of one mil. Unit resistance may be measured in *ohms* (q.v.) per mil-foot. This refers to a conductor having a cross-sectional area of a *circular mil* (q.v.) and a length of one foot.

MILLI—The prefix placed before the name of a unit to indicate one-thousandth part of that unit, as for example *milli-ampere* (q.v.).

MILLI-AMPERE—One-thousandth of an *ampere* (q.v.). Sometimes abbreviated to *milli-amp*.

MILLI-MICRO—A prefix denoting one thousand millions. (10^{-9}). The term *Billi* is an alternate prefix.

MILLI-VOLT—One-thousandth of a volt.

MILS, CIRCULAR—See *Circular Mils*.

MIRROR GALVANOMETER—A galvanometer for measuring very small currents which utilizes a reflected beam of light for a pointer. In one type of mirror galvanometer a fixed nearly circular magnet is used and a coil of wire is suspended in its field. When the current flows through the coil it tends to turn the coil so as to set its axis in the direction of the magnetic force. This movement can be observed by means of a beam of light reflected from a mirror fixed to the coil. (See *Galvanometer.*)

MINERALLAC—An insulating compound made in several different grades all of which are impervious to water and show no alkali or acid reactions.

M.M.F.—Abbreviation for *Magnetomotive force* (q.v.).

MODULATED CURRENTS—Currents which have their amplitude periodically varied as in the case of the carrier current used in radio telephony which is varied or modulated in accordance with the vibrations of a microphone. If a sound wave actuates a microphone, its inward and outward displacement, varying the resistance in the aerial circuit, results in a high frequency current in the aerial of variable amplitude, called a modulated high frequency current. The illustration shows an extremely simplified diagram of the apparatus in a broadcasting station. The initial impulse that starts a radio signal towards the receiving set is furnished by the voice of the speaker or singer or by some other sound originating in the studio. This sound strikes the microphone, the operating element of which is a thin metal disk or diaphragm. The sound waves cause this diaphragm to vibrate. In doing so, the disk causes changes to take place in the electrical circuit. Pulsations of current are set up, the pulsations corresponding in strength to the variations of the spoken word, music or other sound as the case may be. These fluctuations are quite weak when they first come from the microphone and therefore they must be strengthened. They go through what is known as a voice amplifier which utilizes vacuum tubes to make the current stronger. The next piece of apparatus to consider is the *oscillator* (q.v.). This consists of one or more vacuum tubes, connected in a circuit of such a type that the tubes, when lighted and properly furnished with a high voltage direct current, will generate another current, alternating in character, which is said to be oscillating at radio frequency. Upon this current so generated, the voice current is impressed or super-imposed. This process is called *modulation* (q.v.). The current from the oscillator is known as the *carrier wave* (q.v.) current and when voice currents are impressed upon it, it is known as modulated radio frequency current. This current flows in the aerial and there sets up radio waves.

MODULATION—Variation of current or wave form to conform to sound waves or to other predetermined forms. Various forms of modulation are dot and dash modulation, chopper modulation, buzzer modulation and *speech modulation* (q.v.). (See *Distortion, Double Modulation*, also *Modulation Frequency Ratio.*)

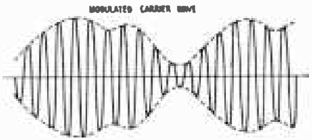

Graph of a modulated carrier wave.

MODULATION, DOUBLE—See *Double Modulation*.

MODULATION FREQUENCY RATIO—The ratio of modulation frequency to wave frequency. An alternating current is said to be *modulated* when the amplitude of its oscillations is varied periodically. The frequency at which the variations occur, that is to say the *modulation frequency* is necessarily less than the frequency of the alternating current which is being modulated. The nature of the variations may assume practically any form. As examples, there are dot and dash modulation, chopper modulation, buzzer modulation and *speech modulation* (q.v.).

MOLECULE—The smallest group of *atoms* of an element or compound which can exist by themselves. A familiar analogy states that if a drop of water could be magnified to the size of the earth, its component molecules would be the size of base-balls.

MOLYBDENITE—symbol MoS_2—A lead gray sulphide of Molybdenum. It is used as a rectifier detector in contact with copper.

MOLYBDENUM—Symbol Mo—A silver white metal found in nature as *molybdenite* (q.v.), a sulphide of molybdenum. When pure this metal is ductile and can be forged. It is used chiefly as an alloy to produce extremely hard steel. Molybdenite is used as a *crystal* (q.v.) in the detection of radio waves.

MORSE LIGHT—A search light used to signal code messages by means of intermittent flashes. (See *Code.*)

MORSE INKER—An instrument for recording Morse code signals in the form of dots and dashes on a moving paper band. In one form of Morse recorder, the attraction of the armature raises an inked wheel against the paper ribbon and thereby prints upon it the dots and dashes of the message.

MOSCICKI CONDENSER—A condenser in which the dielectric thickens out at the edges. It is essentially a form of *Leyden Jar* (q.v.) composed of a glass tube, especially thickened at the neck where the dielectric stress is greatest, coated inside and out with metal foil. The design of this condenser is such as to minimize *brush discharge* (q.v.) or *corona* (q.v.) effect at its edges.

MOTOR—A device for transforming electrical energy into mechanical energy. Motors may be designed for operation on direct current, on alternating current or on either direct or alternating current. These latter are known as *universal* motors. Direct current motors are classified as shunt, series and compound. This classification is arrived at according to the manner in which the field windings are connected. In the shunt motor the field winding is connected in parallel with the brushes (that is with the armature winding). In the series motor the field winding is connected in series with the armature and the line. Compound motors have a shunt winding in parallel with the brushes and a series winding in series with the armature and the line. Compound motors may be further classified as differentially compounded and cumulatively compounded. A differentially compounded motor is one in which the series field winding is so connected as to oppose the shunt field. A cumulatively compounded motor is one in which the series field aids the shunt field. The speed of the shunt motor is nearly constant, falling slightly as the load increases. The speed of the series motor decreases as the load increases. The differential motor may be constructed so that an absolutely constant speed will be maintained at all loads. The speed of the cumulative motor decreases rapidly as the load is increased. Alternating current motors are of the series type, the induction type, commutator type, synchronous type and repulsion type. There are a number of special classifications of alternating current motors such as single phase induction, polyphase induction, repulsion induction, etc. An ordinary direct current series motor used on alternating current will operate but at a low power factor, with a large eddy current loss and with violent sparking at the commutator. In order to adapt the series motor for universal use, it must be specially designed, the field structure being *laminated* (q.v.) in order to avoid *eddy currents* (q.v.) and the field coils having only a few turns to avoid too great *self-inductance* (q.v.) and consequently low *power factor* (q.v.). Sparking at the brushes can be avoided to a certain extent by designing the motor for a small field flux and with but a few turns in series in each armature coil.

MOTOR-ALTERNATOR—A *motor generator* (q.v.) set consisting of an *alternator* (q.v.) driven by a direct or alternating current motor. Such a set may be used for converting from direct to alternating current, or for frequency changing.

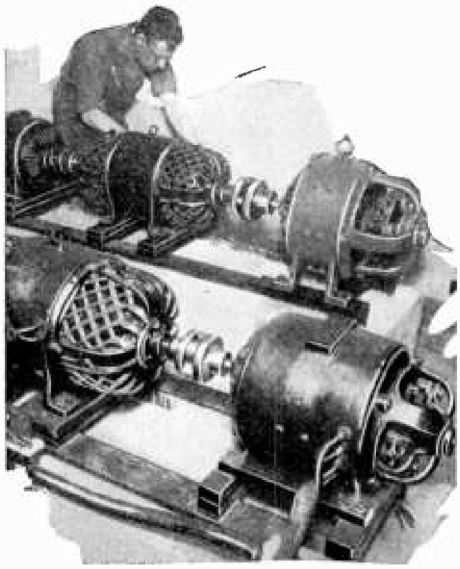

Motor Alternator set installed at Station WRNY.

MOTOR-GENERATOR—A generator or generators driven by an electric motor. Motor generators are used for converting alternating current to direct current or vice versa, for changing from one voltage or from one frequency to another or to obtain a variable from a fixed voltage.

MOVABLE COIL—Inductance coils which can be moved or rotated so as to vary the inductance. Examples of movable coils used in radio are vario couplers, loose couplers, etc. (See *Coupler, Loose; Coupler, Vario,* also *Fixed Coils.*)

MULLARD, S. R.—British radio authority. Born in 1884, he was educated as an electrical engineer and was ultimately appointed head of the research laboratory of Edison & Swan, Ltd., where he developed the Pointolite arc lamp. During the World War, he served in the Royal Air Force and became head of the wireless section laboratory at the Imperial College of Science for the Air Ministry. He carried out a series of researches on vacuum tubes and in 1920 founded the Mullard Radio Valve Company.

MULTIPLE—As applied to an electric circuit, a divided circuit. Two conductors are said to be connected in multiple or *parallel* (q.v.) when their two ends are joined together.

MULTIPLE SERIES—See *Series Parallel.*

MULTIPLIER—A resistance placed in series with a voltmeter for limiting and controlling the amount of current flowing through the voltmeter windings. This resistance may be either internal (enclosed by the voltmeter case) or external.

MULTI-POLAR—Having more than two poles. Usually a generator or motor whose field magnet has more than two poles. Small machines may be made with two poles, *bi-polar* (q.v.), but all others have a number of poles, that is to say they are *multi-polar.* The use of multi-poles results in a lighter machine and incidentally there is a saving of field copper and a reduction in sparking. If there are too many poles, however, there will be excessive *leakage flux* between the poles. (*See Homo-polar.*)

MUSICAL SPARK—A spark giving a regular distinct musical note. This may be produced by a high speed rotary discharger, a *quenched gap* (q.v.) or an arc.

MUSICAL SPARK SIGNALS—Signals in which the sparks occur at regular intervals of time and fast enough to give a musical note. Usually the spark rate is between 100 to 1200 discharges per second.

MUTE ANTENNA—A local circuit or resistor used in testing transmitting apparatus. The mute antenna in this case is substituted in the place of the actual antenna. It is also referred to as a *dummy aerial, phantom antenna, mock antenna,* or *artificial antenna.*

MUTUAL CAPACITY—The capacity effect of one conductor upon another one in the same electrostatic field. The mutual capacity is not the same as the capacity of the two wires regarded as the two plates of a condenser, one positively charged while the other is charged negatively. It really represents a decrease in the capacity of one of the wires with respect to earth caused by the presence of the electrostatic field of the other. The total capacity of the two wires is diminished by the overlapping of the two individual fields. (See *Capacity Measurement of Antenna.*)

MUTUAL INDUCTANCE—Symbol M—The magnetic flux that is common or mutual to two inductively coupled circuits. The mutual inductance of two circuits is the change in the interlinkage of flux that takes place in one circuit for a change of unit current in the other. This is also called the *coefficient of mutual induction* or the *mutual induction coefficient.* The units of mutual inductance are the same as those of *self-inductance.* (See *Inductance, Mutual; Self-Inductance, Induction, Coupling, Induced E.M.F., Induction, Mutual; Coefficient,* and *Mutual Induction Coefficient.*)

MUTUAL INDUCTION—The interaction between two current-carrying coils, not having direct metallic connection. The reactance of a coil such as shown in the illustration at A_1 is reduced by the proximity of another coil, A_2 if the circuit of the latter contains no external source of electromotive force, such as an alternator. The reason for this lessening in the primary reactance is that the flux excited by the first coil induces a current in the secondary coil and this current opposes the action of the primary current. As a result the total *magnetomotive force* (q.v.) the *flux* (q.v.) and consequently the *counter-electromotive force* (q.v) induced in the first coil are reduced. The effect of the secondary coil depends upon its distance from the primary coil and upon the amounts of the resistance R_2 and reactance X_2. If both coils are mounted on the same iron core, the action is greatly increased. In some instances mutual induction is desirable, as in the case of coupled radio circuits, or even in the ordinary transformer. In other cases, mutual inductance may be undesired as in the case where a radio antenna parallels a power line. (See *Induction, Mutual; Mutual Inductance,* also *Mutual Induction Coefficient.*)

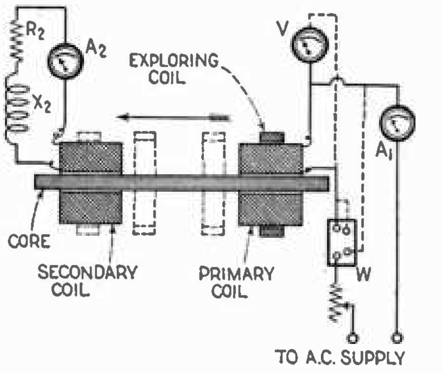

Diagram illustrating mutual inductance. The reactance of the coil A_1 is reduced by the proximity of coil A_2, since the circuit of A_2 contains no source of electromotive force. The effect of the secondary coil depends not only upon its distance from the primary coil but also upon R_2 and X_2. The presence of the iron core greatly increases the action.

MUTUAL INDUCTION COEFFICIENT—Another name for *mutual induction* (q.v.) That coefficient by which the rate of change of the current in a circuit must be multiplied to give the electromotive force induced in an adjacent circuit. From the principle of conservation of energy, it can be shown that the mutual inductance of a circuit A with respect to a second circuit B equals the mutual inductance of B with respect to A. (See *Mutual Inductance, Inductance, Mutual; Self-Inductance, Induction, Coefficient,* also *Coupling.*)

N

NAGYAGITE—A mineral containing lead, gold, antimony, sulphur, and tellurium. It receives its name from the location of its discovery near Nagyag in Transylvania. It is also known as black or leaf tellurium, and has been used as a crystal rectifier in conjunction with zincite.

NALLY, Edward Julian—American radio pioneer. Born April 11th, 1859, in Philadelphia, he joined the Western Union Telegraph Company in St. Louis and worked his way up in the telegraph service until in 1913, he was appointed vice-president and general manager of the Marconi Wireless Telegraph Company of America. Nally was one of the first to see the possibilities of wireless, and under his control the first commercial wireless communication was established between the United States and Japan, in 1914, and in 1920, he founded the first commercial wireless service between the United States and Great Britain, and afterwards to other countries. He was appointed the first president of the

Radio Corporation of America, and is a member of many scientific and other societies.

Edward Julian Nally.

NAME PLATE—A metal plate affixed to a radio set or other apparatus, giving the maker's name, trade name of apparatus, serial number or other details concerning the machine.

NAPERIAN BASE—The base of the so-called natural system of *logarithms* (q.v.). It is usually represented by the symbol ϵ (epsilon) and is numerically equal to 2.718 appx. (See *Damped Waves.*)

NAPERIAN LOG—See *Naperian Logarithms.*

NAPERIAN LOGARITHMS—Also called *Natural Logarithms.* Logarithms to the base ϵ. Logarithms constitute a tabular system of numbers, by which the operation of multiplication can be performed by addition, division by subtraction, involution by a single multiplication, and evolution by a single division. John Napier, Laird of Merchiston, Scotland, is generally regarded as the inventor of logarithms. The logarithms set forth by Napier were those of trigonometric functions. Later on Napier's logarithms were adapted to positive integers, using as a base the number 2.718 appx. and these logarithms are now called natural or Naperian logarithms. (See *Decrement*, also *Logarithms.*)

NATIONAL ELECTRIC CODE—A uniform code of rules, based upon the requirements of fire underwriters, for the electric wiring and electric installations in buildings. Unless these rules are compiled with fire insurance will not be issued or if issued previously will be voided. The National Electric Code contains a section covering special requirements for radio installations. The code is revised annually to meet new conditions and requirements. Copies of the code can be obtained from local insurance agents or from the National Board of Fire Underwriters at New York or Chicago.

NATURAL ELECTRIC WAVES—Wireless waves due to natural causes such as lightning discharges. When received in the radio apparatus, they are called *strays* (q.v.), *static* (q.v.), atmospherics, *X's* (q.v.), and various other names. There are many different causes for these stray waves. Some produce a grinding noise in the loud speaker, others a hissing noise often associated with snow or rain. Nearby lightning produces a sharp snap. Another type of strays cause loud crashes in the speaker.

NATURAL MAGNET—A substance which possesses in a natural state the properties of a magnet. Such a substance is the *lodestone*, a magnetic oxide of iron also called *magnetite* (q.v.). (See *Loadstone.*)

NATURAL FREQUENCY—The frequency at which free oscillations occur in a circuit. If L, C, and R are respectively the inductance, capacity and resistance of a circuit, then free oscillations in the circuit will have the frequency

$$f = \frac{1}{2\pi}\sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}$$

This is the natural or *fundamental frequency* of the circuit. The natural frequency of an aerial is the frequency corresponding to the natural wave length. (See *Natural Wave Length*, also *Measurement of Wave Length.*)

NATURAL RECTIFIER—A mineral which possesses the property of conducting an electric current in one direction only. (See *Crystal*, also *Crystal Detector.*)

NATURAL WAVE-LENGTH—Length of the wave emitted by an aerial when no added inductance or capacity is inserted in the aerial circuit. This is also known as the *fundamental wave length* (q.v.). By putting *inductance coils* (q.v.) (*loading inductance*) (q.v.) in the aerial circuit, longer waves may be radiated, while condensers put in series with the antenna enable it to produce shorter waves than the aerial would ordinarily radiate. The use of a series condenser is avoided wherever possible since it has the effect of decreasing the total capacity of the aerial circuit and thereby decreasing the amount of power which can be given to the antenna. The addition of some inductance has a beneficial effect, since the decrement of the aerial is thereby lessened and a sharper wave results. (See *Measurement of Wave Length; Wave Length.*)

NAUTICAL MILE—abbreviation *Naut*—A marine unit of distance. Equivalent to one minute of longitude at the equator. This unit is also sometimes referred to as a *Telegraph Naut.* It is equal to 1.1528 statute miles (the statute mile being equal to 5,280 feet). The term *nautical mile* is used especially in submarine cable work. and also in matters pertaining to navigation. The term *naut* is differentiated from *knot* in that the former refers to a distance whereas the latter refers to a rate or unit of speed.

NEGATIVE CARRIER[*]—An electron combined with neutral gas molecules. When the velocity of the electrons in a vacuum tube is less than the value necessary to cause ionization by collision, the electrons attract the neutral gas molecules and so form heavy negative carriers. The ease with which this formation of negative carriers takes place depends on the nature of the gas. Such gases as argon and mercury vapor do not readily form negative carriers, while hydrogen and oxygen combine with electrons more easily. The effect of this negative carrier formation is to counteract the reduction of the negative space charge occasioned by the heavy positive ions formed by collision ionization. The positive ions are ions of the gas from which one or more electrons have been removed. The ions therefore have very nearly the same weight as the gas atoms. The negative carriers, on the other hand, may consist of an atom or molecule to which has been attached an electron. It is also possible that the attraction between an electron and the neutral gas molecules can result in the formation of clusters consisting of more than one molecule held together by the electron. These negative carriers, therefore, move as slowly as, and often more slowly than, the positive ions and consequently have a relatively great effect in counteracting the tendency of the positive ions to reduce the negative space charge of the electrons. (See *Electron Theory*, also *Ion.*)

NEGATIVE CORPUSCLE—The natural unit of negative electricity. The *electron* (q.v.). It is more usual to refer to the negative corpuscle simply as the *corpuscle.* This was the name given by Sir J. J. Thomson to the carriers of electricity shot off from the cathodes in vacuum tubes. The corpuscle was found to have a charge equal to the electron and a mass which is $\frac{1}{1845}$ of that of the hydrogen atom.

NEGATIVE ELECTRODE—In a *primary cell* (q.v.) the *cathode* (q.v.) which is the carbon, copper, etc., electrode, is the *negative electrode*, while the *pole* of this electrode is the positive pole, since it is positive in relation to the external circuit. In a *secondary* or *storage cell*, the spongy lead plate, which is the anode during discharge, is called the *negative electrode* and its pole the *negative pole.*

NEGATIVE ION—An atom, which is the smallest particle of an element capable of existing, plus an *electron* (q.v.) (See *Anion, Cathion, Electron Theory*, also *Ion.*)

NEGATIVE POLE—The south-seeking end of a magnet. In a generator, the terminal into which the current returns from the external circuit. In a storage cell, the terminal of the negative plate. In a primary cell, the pole of the positive plate.

NEGATIVE RESISTANCE—A current path within a vacuum tube or an arc in which current decreases as voltage increases. The emission of electrons from cold electrodes under the impact of electrons (known as *secondary electron emission* (q.v.) or *delta rays*) results in a negative resistance or *falling characteristic.* The presence of secondary electrons can be shown by referring to the illustration. The plate P is kept at a constant positive potential with respect to the filament F by means of the battery B. When there is no difference of potential between the filament and the grid G, the cur-

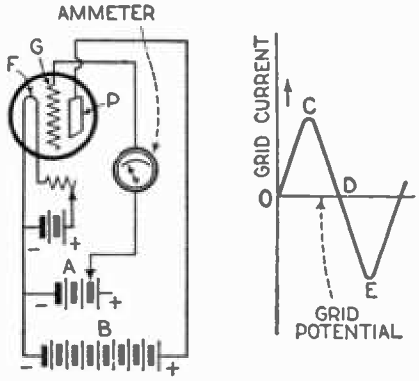

At the left is a hook-up for demonstrating negative resistance. The curve at the right shows the decrease of grid current although grid potential is increased.

---

[*] Van der Bijl—The Thermionic Vacuum Tube.

rent in the circuit FGA is very small, since practically all the electrons emitted from the filament are drawn through the openings of the grid and thrown on to the plate. If the potential of the grid (positive with respect to the filament) is increased, the current to the grid at first increases, as shown by the part OC in the accompanying curve. When the grid potential reaches a certain value, the current as indicated by the ammeter begins to decrease, drops down to zero at D and then reverses its direction of flow. The reason for this is that while the difference of potential between the filament and the grid is small, the electrons that strike the grid enter it, but as the positive grid potential is increased the electrons on striking the grid emit *secondary electrons* from it and these are drawn to the plate which is maintained by the battery B at a positive potential, with respect to the grid. The net current as shown by the ammeter is the sum of the electrons entering the grid and those leaving it. When the velocity with which the electrons strike the grid increases beyond a certain value, one primary electron can knock out more than one secondary electron from the grid and the current in the filament-grid circuit reverses. When the positive grid potential is increased to such an extent that the grid becomes positive with respect to the plate, the secondary electrons are no longer drawn away to the plate, but are driven back to the grid so that the reversed current in the grid circuit again decreases as shown at E on the curve, and finally assumes the original direction. Hence, considering the current as indicated by the ammeter and the voltage between filament and grid, it can be seen that over the portion of the curve CDE, the current decreases as the voltage increases. Therefore CDE represents a *negative resistance* characteristic. It has been found that a device which possesses a negative resistance can function as an amplifier and also as a generator of continuous oscillations.

NEGATRON—An English vacuum tube containing a filament, a grid and two plates. This tube was devised by J. Scott-Taggart, who used it for obtaining *negative resistance* (q.v.) characteristics. The two plates are fixed one on each side of the filament. The accompanying illustration shows a typical negatron. The bulb is tubular in shape and the four electrodes are plainly shown. The grid, in this tube, is a metal rod.

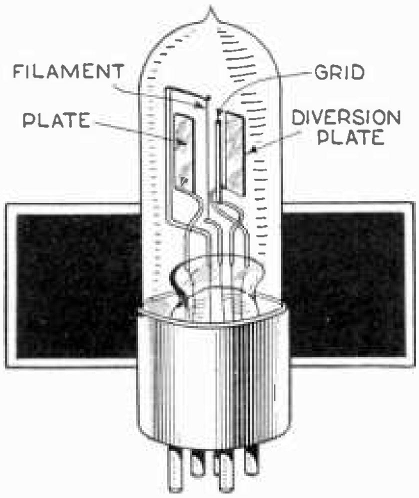

A negatron tube, showing the two plate construction.

NEON LAMP—An incandescent lamp in which a reddish light is produced by the incandescence of neon gas at low pressure. This lamp has very low current consumption and has been applied to pilot lights, signs, etc.

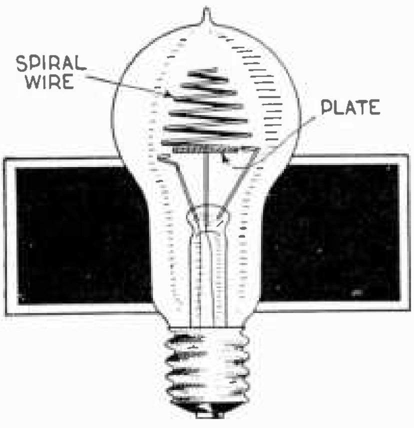

Neon Lamp

NEON TUBE—A vacuum tube containing the gas neon at low pressure used in the Fleming *Cynometer* (q.v.) etc. In this instrument, which is used to measure the frequency of electric waves, wave lengths, etc., a capacity formed by one brass tube sliding over another can be varied simultaneously with an inductance consisting of an air core solenoid and arranged so that the point at which resonance takes place is indicated by the glowing of the *neon tube* and can be read from a calibrated scale.

NERNST LAMP—An incandescent lamp in which the incandescent body consists of a strip of material composed of a mixture of oxides of metal such as zirconium, magnesium and other refractory oxides. The incandescent portion of the lamp is called the *glower* and is conductive only at a high temperature. Consequently it has to be heated by an auxiliary heating coil. The heating coil is automatically cut out of the circuit when the current starts to flow through the glower.

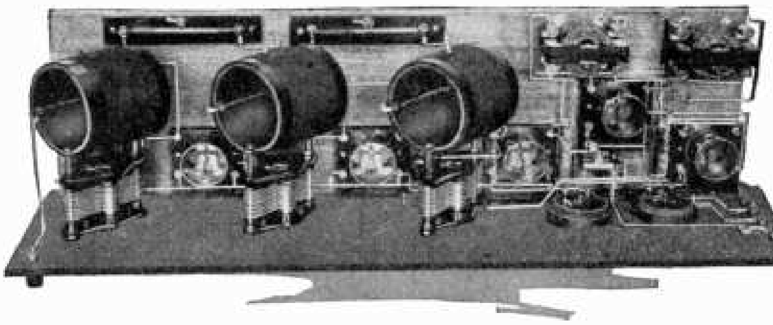

Fig. 2. A typical neutrodyne set.

NEUTRAL BODY—A body void of electrification. A charged body which loses its electrification is said to be *discharged* or *neutralized*. All matter is capable of electrification.

NEUTRAL LINE OF MAGNET—The middle portion between the two magnetic poles of a bar magnet, where there is no manifestation of magnetism.

NEUTRAL WIRE—A middle wire, in a *three wire system* (q.v.) of power distribution, which is kept at a potential mid-way between the positive and the negative mains. In the three-phase system of alternating current distribution, the conductor connected to the neutral point of a "star" or "Y" connected alternator is also known as the neutral wire. The neutral wire is here also known as the *common return*.

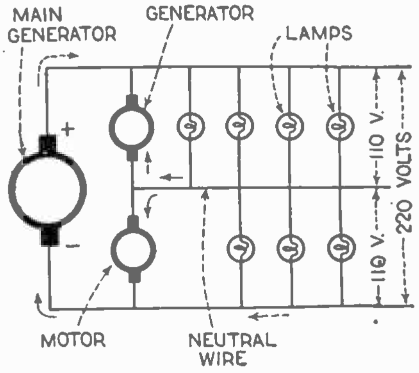

A three wire system with unbalanced load, showing neutral wire.

NEUTRODYNE—See *Neutrodyne Circuit*, also *Hazeltine Neutrodyne Receiver*.

NEUTRODYNE CIRCUIT—A circuit used in radio reception in which radio frequency amplification is used for neutralizing the effects of capacity of the tube and its socket. Figure 1 shows a portion of a receiving circuit. In this diagram the tube capacity is represented by a condenser shown in dotted lines. By studying this circuit, it can be seen that a closed oscillatory circuit exists, made up of the inductances and capacities in the plate and grid circuits. Just as radio frequency currents can pass through condensers in a receiving set, so they can pass through or across the vacuum tube from the plate to the grid, which forms a condenser. This condenser effect is the reason why radio frequency amplifiers oscillate. A difference of potential occurs in the plate side of the coil in the plate circuit. Instead of this being handed on to the next tube for additional amplification, a portion of this potential leaks through the capacity of the tube and its socket and affects the grid in such a way that trouble is experienced with self-oscillation. In the neutrodyne circuit, the

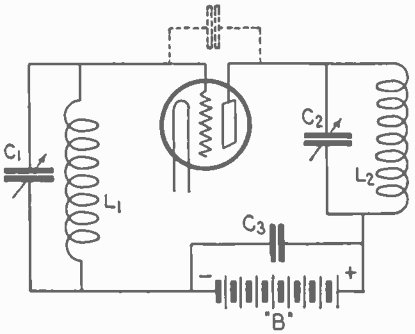

Fig. 1. A portion of a receiving circuit. Dotted line shows tube capacity represented by an equivalent condenser.

inter-element capacity of the amplifier vacuum tubes is neutralized by means of special condensers called *neutrodons*. The capacity of these is very low, being approximately equal to the internal capacity of a vacuum tube. By reason of this equality, any tendency of a large amount of radio frequency current to pass back through the tube through the grid is defeated and instead is neutralized by the combination of the neutralizing capacities, the inter-element capacity of the vacuum tubes and the secondary windings of the tuned radio frequency transformers. This effect is in reality a bucking one, since the current is made to take two paths. Each neutralizing condenser must be adjusted so that its capacity will equal that of the vacuum tube it is connected with. Since it is impossible for a neutrodyne circuit to oscillate, there is no radiation of energy from the set, with consequent disturbance of other receiving sets in the vicinity.

The neutrodyne circuit was invented by Professor L. A. Hazeltine. The typical neutrodyne receiving set uses five tubes, employing two stages of tuned and neutralized radio frequency amplification, with detector and two audio frequency stages. (See *Hazeltine Neutrodyne Receiver.*)

NICHROME WIRE—A nickel-chromium-steel alloy wire, which is able to withstand a bright red heat in the atmosphere without oxidizing. Nichrome is practically non-corrosive. It has a very high melting point, about 1550 degrees Centigrade and is used extensively in electrical heating appliances and resistance elements.

NIGHT EFFECT—Changes in the strength of radio signals noticeable especially at night. This effect is also known as *fading* (q.v.). Fading does not generally occur in the immediate vicinity of the transmitting station. It seems to happen more often on wave lengths below 400 meters. A certain station is being received with normal intensity, when suddenly the sounds become very faint and in a few minutes they may again become normal or may even become stronger than usual. In some cases the variations from strong signals to weak take place every few minutes while in other cases there may be an interval of several hours. It has been observed that fading takes place over land more often than in radio transmission at sea.

NOBLE, Sir William—British engineer. Born in 1861 and educated at Gordon's College, Aberdeen, he thereafter entered the Aberdeen telegraph office. In 1893 he was appointed engineer for the northeast area of Scotland, rising rapidly in the service until he became chief engineer in 1919. He retired from the Post Office in 1922, and became a director of the General Electric Company and of the British Broadcasting Company. Sir William Noble has written many articles on telegraphy and telephony and is considered to be an authority on these subjects.

NODES—In a wave form, such as shown in the illustration which represents an alternating current sine wave, the zero points. Thus in the alternating current sine wave, the points of zero current or potential, at A, C, E, and G, are nodes, and the points B, D, and F are the anti-nodes or *loops* (q.v.).

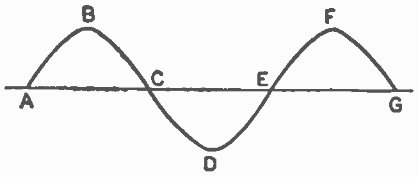

Sine wave illustrating nodes and anti-nodes.

NODEN VALVE—An *electrolytic rectifier* (q.v.) which allows only current in one direction to pass through it. This rectifier utilizes an aluminum rod or cylinder as a cathode and the anode may be of iron, lead or of carbon. Ammonium phosphate is used as the electrolyte. The principal of operation of this rectifier depends upon the fact that the current in the positive direction is stopped at the aluminum plate through the formation of an insulating crust of aluminum phosphate and aluminum oxide. These two substances present an extremely high resistance to the passage of the current in the positive direction, but the compounds dissolve again upon reversal of the current. By suitably combining two or more cells, both the half waves of an alternating current can be rectified. In Fig. 1 is shown the principle upon which the Noden Valve acts, and Fig. 2 shows a hook-up for connecting four cells to obtain full wave rectification. (See *Electrolytic Rectifier Full Wave*, also *Rectifier.*)

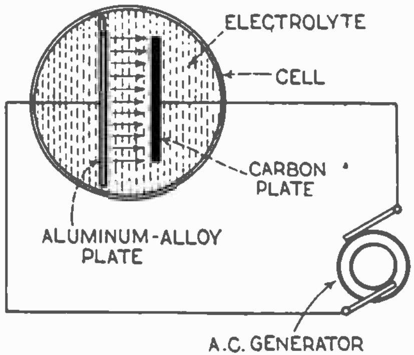

Fig. 1. Diagrammatic illustration of the principle of the Noden valve.

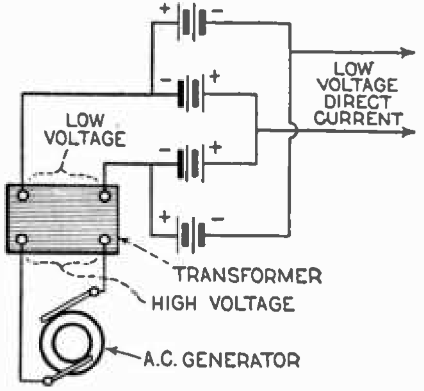

Fig. 2. A hook-up for connecting four Noden cells to obtain full wave rectification.

NON-CONDUCTOR—See *Insulator.*

NON-INDUCTIVE COIL—A coil having negligible inductance. Such a coil may be wound by doubling the wire upon itself and then winding the two parallel halves side by side.

NON-OSCILLATORY—Free from oscillations or vibrations. Without periodicity, that is to say, *aperiodic* (q.v.). A current which is uniform in direction and flow and free from oscillations is said to be a non-oscillatory current. A non-oscillatory circuit is one in which an impressed potential will produce current that gradually diminishes in amplitude without reversing its direction of flow. This is also called an *aperiodic circuit* (q.v.). Such a circuit can have no natural period of oscillation.

NON-PERIODIC—See *Aperiodic*, also *non-oscillatory.*

NON-RADIATIVE ANTENNA—A combination of capacity and inductance equivalent to that of an aerial, used to test transmitting apparatus without radiating waves. (See *Mute Antenna.*)

NORTH MAGNETIC POLE—A point on the earth at a latitude 70 degrees North, longitude 97 West. The north magnetic pole does not coincide with the north (geographic) pole. (See *Magnetic Poles.*)

NORTH POLE—See *Magnetic Poles*, also *North Magnetic Pole.*

NO-VOLT RELEASE—An electro-magnetically controlled device, inserted in the field circuit of a motor so that it holds the handle of the motor starter in place on the last stud, but if the supply current fails, the magnet becomes de-magnetized and allows the spring attached to the starter handle to pull the handle back to the "off" position. This prevents damage to the windings of the motor, in case the current comes on again, since it is then necessary to start up the motor in the usual way, using the starter and gradually increasing the voltage by cutting out the limiting resistances provided in the starter. (See *Overload Release.*)

NULL METHOD—An electrical method of testing in which adjustments are made so that zero deflection is obtained in a galvanometer, as for example when using a *Wheatstone Bridge* (q.v.). This is sometimes referred to as the *zero method.* (Null is also a German word meaning zero.)

O

OCTAHEDRITE—Chemical Symbol TiO_2—Titanium dioxide, an eight sided crystal which has been used as a crystal detector for detecting and rectifying radio currents. (See *Crystal.*)

OERSTED—The unit of *reluctance* (q.v.). The reluctance or in other words, the magnetic resistance offered by a cubic centimeter of vacuum. Unit *magneto-motive-force* (q.v.) will generate a unit of *magnetic flux* (q.v.) through unit reluctance. The oersted was named after the Danish physicist, H. C. Oersted.

OHM—Symbol Ω (Omega)—The unit of electrical resistance. It is the ratio of unit *electromotive force* (q.v.) to unit current. The ohm is the basis of the entire electrical system of units and the volt and the ampere are defined in relation to it. Very elaborate apparatus is required in order to determine the ohm absolutely. However, resistances can be compared easily and once the standard is determined it is a simple matter to make duplicates. In accordance with an act of Congress, the ohm was adopted as a legal unit and defined as follows: The unit of resistance shall be what is known as the international ohm, which is substantially equal to one thousand million units of resistance of the centimeter-second-gram system of electromagnetic units and is represented by the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice, 14.4521 grams in mass, of a constant cross-sectional area and of the length of 106.3 centimeters. Stated in English units, the ohm is specified as the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice (32 degrees Fahrenheit), 0.5050 of an ounce in mass (approximately 3/100ths of a pound), of a constant cross-sectional area and of a length of 3.487 feet.

OHMMETER—An instrument used to measure resistance directly in ohms. It is particularly adapted for measuring high resistances, although certain types of ohmmeters are made for measuring lower resistances. In one form of ohmmeter, the moving system is deflected by forces due to currents in two coils at right angles to one another. These carry currents in one case proportional to the current through the conductor under test, and in the other case proportional to the potential drop across it. This is in effect, a special form of *galvanometer* (q.v.) and when combined with a hand driven generator gives the conventional *megger* (q.v.) The bridge type ohmmeter depends on the *Wheatstone Bridge* (q.v.) principle. In this instrument, however, the rheostat resistance remains constant and the bridge arms are formed by a wire, called a *slide wire*, resistances being varied by means of a moving contact as shown in the illustration. This type of instrument is often referred to as a *slide wire ohmmeter.* While a straight slide wire is used in some types, in others the wire is wound on an insulating cylinder. In the Weston direct-reading ohmmeter, there is a permanent magnet and a moving coil having two windings and an adjustable magnetic shunt. Provisions are made for connecting an unknown resistance and an auxiliary battery in the circuit of the instrument. A checking plug is provided for checking the accuracy of the instrument before use. When the key is closed, the battery current is divided between two windings. The magnetic force exerted on the one winding, which is in series with a fixed resistance, tends to deflect the pointer up the scale, while the force exerted on the other winding tends to return it to zero. With the correct battery voltage the magnetic shunt may be adjusted to bring the pointer to full scale position. When the plug is shifted to the low range position, a resistance equal to that of the low range is removed from the circuit whose magnetic force tends to deflect the pointer towards the zero position. By connecting a resistance equal to that of the low range across the "unknown" binding posts, the previous condition is restored and the pointer will go back to the full scale position. For values of resistance lower than this the opposing force will be greater and hence the coil will be brought back to a lower position on the scale. At zero resistance across the "unknown" binding posts, the magnetic forces of the two windings will balance each other, thus causing the coil to stand at zero on the scale. The Vawter indicating ohmmeter is of the galvanometer type and utilizes two coils carried by a shaft and moving in the field of a permanent magnet. A specially shaped core is used to give the desired scale characteristics. By making connections to the coils through spirals of negligible torque, the position of the pointer does not depend upon the value of the current nor on the strength of the permanent magnet.

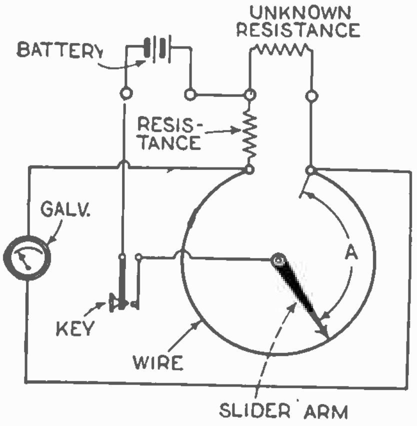

Diagram of connections of a slide wire ohmmeter.

OHMIC COUPLING—See *Resistance Coupling.*

OHMIC DROP—The fall in potential which occurs when an electric current flows through a resistance. (See *IR Drop.*)

OHMIC LOSS—Power or energy loss due to the resistance which an electrical circuit offers to the flow of current. (See *I^2R Loss.*)

OHM'S LAW—Current flowing in a conductor will increase directly with increase of voltage and will decrease directly with increase of resistance. In other words, voltage is the cause, while current is the effect. The effect is directly dependent upon the cause and inversely dependent upon the opposition offered to the cause. Ohm's law is the fundamental law of electrical engineering. Ohm's law can be stated as an equation as follows:

$$\text{Current (in amperes)} = \frac{\text{Voltage (in volts)}}{\text{Resistance (in ohms)}}$$

or

$$\text{Voltage} = \text{Current} \times \text{Resistance}$$

$$\text{Resistance} = \frac{\text{Voltage}}{\text{Current}}$$

The above three equations all mean the same thing and serve to express in various forms, the idea set forth above. Ohm's law can be applied to every conductor, provided that the conductor stays at a constant temperature, that there is no internal electromotive force in the conductor and that the distribution of the stream lines in the conductor remains unchanged. Ohm's law is easy to understand and easy to apply. Thus in a particular circuit, a certain voltage causes a certain current to flow. Double the voltage will cause double the current to flow, if the resistance is the same. If the resistance is doubled, however, the same voltage will result in only half the current flow. Double voltage and double resistance results in unchanged current flow. Ohm's law applies to a portion of an electrical circuit as well as to the entire circuit. Care must be taken when applying the law to only part of the circuit, to consider only the resistance, current and voltage of that part.

OMNIBUS BAR—A conducting bar of copper, mounted in back of a switchboard, or panel and serving as a common connector for two or more pieces of apparatus. This term is more usually shortened to *bus bar* (q.v.)

OMNIGRAPH—An intrument which sends code (dots and dashes) mechanically. It is usually operated by clockwork and is connected with a buzzer. It employs metal disc records each of which has a series of notches which correspond to the dots and dashes of the telegraph code. The omnigraph is used by beginners in learning the code and is also used to test the speed of reception of applicants for radio operators' licenses.

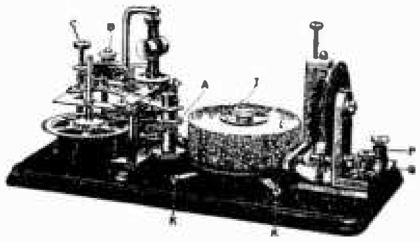

The omnigraph—an instrument which sends code mechanically.

ONDOGRAPH—An instrument for autographically recording the wave forms of varying currents, especially currents which are alternating rapidly. This form of curve tracer was invented by Hospitalier.

ONDOMETER—Another name for a *wave meter* (q.v.)

OPEN ANTENNA—According to the report of the Institute of Radio Engineers standardization committee, this is a condenser antenna, that is to say an antenna consisting of two capacity areas. Antennas may be subdivided into two general classes, those which act principally as condensers, usually simply called antennas, and those which act principally as inductances. These latter are *loop antennas* (q.v.), *coil antennas* (q.v.), etc. When electric waves reach a condenser antenna, they set up an alternating electromotive force between the wires forming the upper plate of the condenser and the ground or the other lower plate of the condenser. As a result of this alternating electromotive force, alternating current will flow in the antenna system. (See *Aerial.*)

OPEN CIRCUIT—An electrical circuit which does not offer a complete path for the flow of electricity. There may be one or more breaks in the path. An open circuit is often referred to as a *broken circuit.* Open circuits often cause trouble in radio apparatus. They may be due to the burning out of a winding, to a connection coming off a terminal, to a soldered connection breaking loose or to similar causes. The method of testing for an open circuit is to connect a battery and a headset in series with the circuit in question. If there is no click in the headset, there is evidently a break in the circuit.

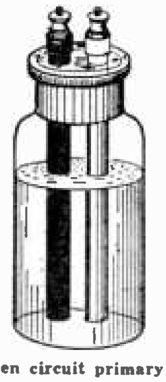

An open circuit primary cell.

OPEN CIRCUIT PRIMARY CELL—A *primary cell* (q.v.) designed for intermittent use and normally kept on open circuit. Such cells usually have a *depolarizer* (q.v.) which acts slowly. In certain cases, no depolarizer whatsoever is employed. Open circuit cells are designed for use during short intervals of time only and must stand for long periods on open circuit, during which the hydrogen is gradually taken away from the negative plate. The operation of the open circuit cell will not be satisfactory if the rest periods are not long enough or are not at frequent intervals.

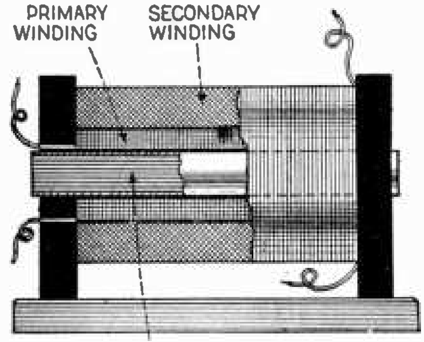

An open-core transformer.

OPEN-CORE TRANSFORMER—A transformer in which the (useful) magnetic circuit is partly through iron and partly through air, instead of entirely through iron as in the *closed-core transformer* (q.v.). It is sometimes referred to as a *polar* transformer. In one type of open-core transformer the arrangement is similar to that of an *induction coil* (q.v.), a hollow iron core, filled with soft iron wires being employed to concentrate the magnetic field.

OPPOSING ELECTROMOTIVE-FORCE — See *Back Electromotive-force.*

OPTOPHONE—An instrument utilizing a telephone receiver controlled by the variations of resistance of a *selenium cell* (q.v.) by means of which the blind are enabled to read ordinary type, through the sense of hearing.

ORDINATE—A mathematical term used in radio in making *logs* (q.v.) or other curves. Defined in mathematical language, the ordinate is that one of the coordinates of a point which is drawn parallel to a line (called the axis of ordinates) to the point from the other axis (called the axis of abscissas) or from the plane of the other axis of coordinates, assumed as a basis of reference. This definition will be clearer if reference is made to the accompanying illustration. The dotted line drawn from the point A to the axis of abscissas is the ordinate of the point A.

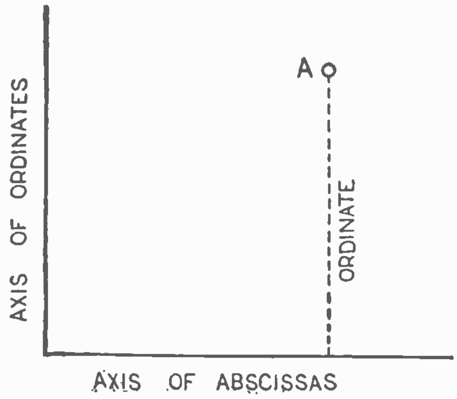

The dotted line is an ordinate.

OSCILLATING CURRENT — An alternating current of high frequency, consisting of a succession of waves of constant length. In some cases, the term oscillating current refers particularly to a current where the amplitude is decreasing in constant proportion due to damping. (See *Oscillation, Oscillation Frequency,* also *Logarithmic Decrement.*)

OSCILLATING IMPULSE — *An oscillating current* set up for the purpose of producing electric waves.

OSCILLATING PERIOD—The time of one complete *oscillation* (q.v.).

OSCILLATION—Periodical surging first in one direction and then in the other. Electrical oscillations are the surgings backward and forward of current periodically, as, for example, in the case of a condenser being discharged in an oscillating circuit. The oscillating action in a discharging condenser is as follows: In the case of a condenser, positively charged, upon being discharged through an inductance, the charge rushes out of the condenser, the current increasing in strength until it reaches a maximum. Then, due to self-induction the current prolongs itself, and it continues until its energy is changed back into an electric charge of an opposite sign, current flowing into the condenser which is then charged negatively. The condenser again discharges giving a maximum current in the opposite direction. the current thereafter flowing back into the condenser with the same direction charge as at first. This action constitutes one complete swing or oscillation and the current continues to oscillate, each time becoming weaker and weaker as the energy is dissipated into heat and also partially into electric waves. It is important to note that the propagation of electric waves could not be accomplished without electrical oscillations. Thus at the transmitting station, the electrical oscillations are converted into electric waves and at the receiving station, the waves are re-converted back into oscillations.

OSCILLATION CONSTANT—The square root of the product of the capacity and the inductance of a circuit. The *natural period* of a circuit is determined by its inductance and capacity in the same manner as the natural period of a steel spring clamped at one end in a vise is determined by its mass and elasticity. Where a simple aerial is used it is an easy matter to calculate the length of a wave radiated from it, provided the oscillation constant has been found. The velocity of electric waves is approximately 3 x 10^{10}cm. per second, and hence the length of the wave radiated is about 60 times the oscillation constant where the wave length is measured in meters. (See *Aerial Tuning Condenser,* also *Isochronous.*)

OSCILLATION FREQUENCY—The number of complete oscillations or cycles flowing in a circuit per second. The oscillation frequency of a circuit varies inversely as the square root of the product of the inductance and capacity. The frequency is inversely proportional to the wave length, hence the addition of either *capacity* (q.v.) or *inductance* (q.v.) to the oscillatory circuit will result in an increase of *wave length* (q.v.). (See *Frequency.*)

OSCILLATION POINT—That current value, known as the *critical current* (q.v.), at which the vacuum tube or tubes in a radio receiving set start to oscillate.

OSCILLATION TRANSFORMER—A high-frequency air-core transformer used for transferring electrical oscillations from one circuit to another. Either one or both the windings may be tuned to the frequency of the oscillations. The transformer used for coupling the oscillator of a transmitting set to the aerial is an example of an oscillation transformer. (See *Coupling, Inductive Coupler,* also *Transformer.*)

OSCILLATION VALVE—See *Oscillator,* also *Vacuum Tube.*

OSCILLATOR—A device or apparatus for producing electrical *oscillations.* The *vacuum tube* (q.v.) under correct conditions may be used to produce sustained oscillations of a definite amplitude and frequency. In order for a vacuum tube to act as an oscillation generator it must be capable of amplifying, part of its output energy must be returned to the input, an oscillation circuit must be combined with the tube, which possesses inductance, capacity and resistances of values such that the tube will oscillate with the desired frequency and finally the tube must have certain characteristics which when combined with the constants of the oscillatory circuit, will determine the amplitude of the oscillations. Oscillations may also be produced by means of a high frequency alternator and by a direct current generator and electric arc. (See *Alternator,* also *Alexanderson.*)

OSCILLATOR, HERTZIAN—Two insulated rods, with axes in line, with their nearer ends forming a spark gap and their outer ends bearing plates or balls to give the system the required capacity. The overall length is usually less than 6½ ft. and the whole constitutes an open oscillator. A Hertzian oscillator is sometimes called a *dipole.* When the rods are brought near each other but not in actual contact and the conductors are connected to the terminals of an *induction coil* (q.v.) a succession of sparks can be caused to jump the gap and these set up electric waves in the *ether* (q.v.).

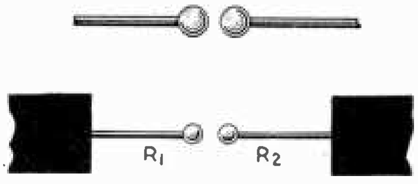

Typical Hertzian Oscillators.

OSCILLATOR, LINEAR—The straight wires connected on either side of the spark gap of an induction coil, traversed by oscillations. Linear oscillators are of the open type. As used by Hertz, they consisted of two rods, having a spark gap between their adjusted ends, while their outer ends terminated in plates or balls in order to increase the capacity of the system.

OSCILLATOR, RIGHI—In this type of oscillator, the spark gap is established between two metal spheres placed between two smaller spheres. (See *Oscillator,* also *Oscillator, Hertzian.*)

OSCILLATOR, THERMIONIC—A vacuum tube used to generate continuous oscillations. (See *Oscillator.*)

OSCILLATORY CIRCUIT—A circuit possessing *inductance* (q.v.) and *capacity* (q.v.), the electrical constants being such that an oscillatory current can be set up. The accompanying figure shows a simple oscillatory circuit, consisting of a condenser, an inductance and a spark gap. When a source of high potential is connected across the spark gap and the latter is properly adjusted, the condenser is charged and there is a spark across the gap. Provided that there is not too much resistance in this circuit, there will be set up an *oscillating current* (q.v.) of high frequency, decreasing in amplitude according to a fixed ratio. Such a current can be used to set up electric waves. Oscillatory circuits are of two general classes, closed as shown in the illustration and open, as in the case of an aerial system. An aerial which has no tuning condenser or inductance in its circuit still has distributed inductance and capacity, whereas an aerial having a tuning inductance and a condenser in its circuit, has a combination of concentrated and distributed inductance and capacity.

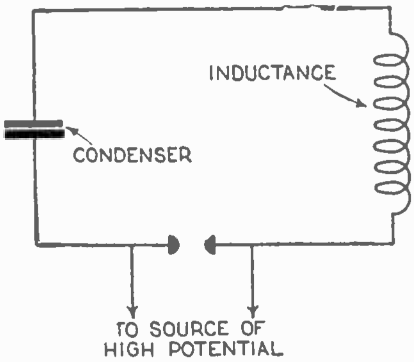

A simple oscillatory circuit.

OSCILLATORY CURRENT—See *Oscillating current.*

OSCILLION—Name which has been applied to the *vacuum tube* (q.v.) especially where used as an oscillation generator.

OSCILLOGRAM — A picture, graph or record of a wave form usually obtained by means of an *oscillograph* (q.v.).

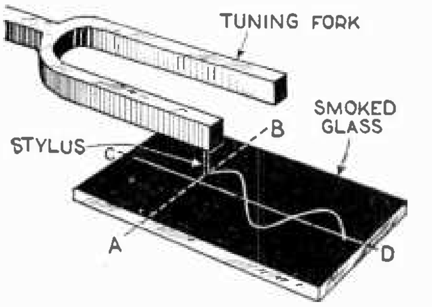

An oscillogram obtained by means of a tuning fork.

The illustration shows an oscillogram obtained on smoked glass by means of a tuning fork. (See *Oscillagraph.)*

OSCILLOGRAPH — An instrument for recording photographically or showing visually the wave form of alternating currents or of other rapidly changing currents or voltages. The essential parts of such an instrument are a moving coil mirror galvanometer, a rotating or vibrating mirror and a moving photographic plate or film. Figure 1 illustrates the principle of the oscillograph and shows the curves it produces. The oscillograph was invented by M. A. Blondel in 1893. By means of the oscillograph the wave form of a current may be shown as a curve and its characteristics may thus be examined and studied. In order to form the curve, it is necessary to have simultaneous motion in two directions. The vibrator of the oscillograph which consists of a delicate coil of wire, with mirror attached, is made extremely light so that the moving system will have as little inertia as possible. The coil is placed in a powerful magnetic field and when an oscillating current is passed through it, it oscillates in synchronism with the current. A spot of light focused on the mirror is made to fall on a screen at the predetermined distance from the vibrator. This spot of light traces out a straight line on the screen. Substituting a photographic film for the screen and arranging the film so that it will move in a vertical path across the moving beam of light (which is assumed to be moving from right to left across the screen) a curve is traced on the film. The commercial form of oscillographs are provided with apparatus whereby the curve may be examined without the necessity of photographing it. In these the beam of light is reflected from a rotating or vibrating mirror in such a way that it receives a horizontal and an up and down motion and is focused onto a curved plate of glass. Tracing paper may be stretched over the glass so that the image of the spot of light traces a visible curve on it. The curve being repeated very rapidly appears to be standing still and hence it is possible to trace over it with a pencil or even to place a piece of photographic paper on the curved glass and make a record of the curve in this way. The motion picture camera has been applied to the oscillograph and is useful where longer records are required. It should be noted that oscillographs which utilize a moving coil, such as the Duddell or Blondin types, will work over a band of frequencies up to a maximum of about 300 cycles per second. Inasmuch as the oscillations common in radio work are of much higher frequency than 300 cycles, it has been found necessary to use an oscillograph whose moving parts are without appreciable weight. The use of electrons as moving parts has solved this problem. A Frenchman, Alexander Dufour has applied the Braun tube to the construction of an oscillograph which has no appreciable inertia and which is capable of operating at a frequency up to a million cycles per second. This device uses only minute amounts of energy in its operation and therefore does not disturb the original circuit. It is able to register both voltage and current simultaneously and is so arranged that a single impulse is sufficient for a photographic impression. The Dufour oscillograph consists essentially of two glass tubes fitted by means of a ground joint into a bronze chamber. The upper glass tube carries a cathode and an anode and the other tube has one pair of deflecting plates for electrostatic deflection of the electron stream. Two sets of coils perpendicular to each other are used for magnetic deflection and these are placed outside the tube and slightly below the deflecting plates. In order to photograph the oscillations, a drum is used which is provided with a film magazine allowing six films to be taken in succession. When visual examinations of the oscillations are to be made, a fluorescent screen is turned up into position covering the opening into the interior of the drum thus preventing the films from being exposed when using the screen. After the films are put into the drum, this is placed within the bronze chamber and locked into position. An air-tight door is used to close the opening and the film changing mechanism is operated by means of external controls. Glass windows on either side of the bronze chamber allow the fluorescent screen to be seen. The accompanying sketch shown at Figure 2, gives a diagrammatic representation of the operation of the Dufour oscillograph.

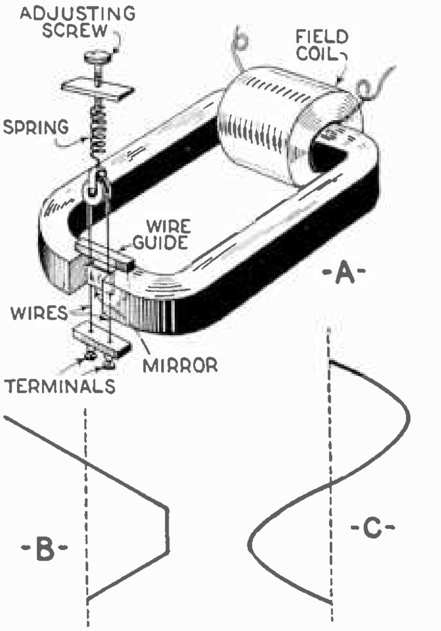

Fig. 1. The principle of operation of the oscillograph is shown at A. At B are shown diagonal and straight paths traced by the beam of light falling on a moving photographic plate. The type of curve which an oscillograph traces for alternating currents is shown at C.

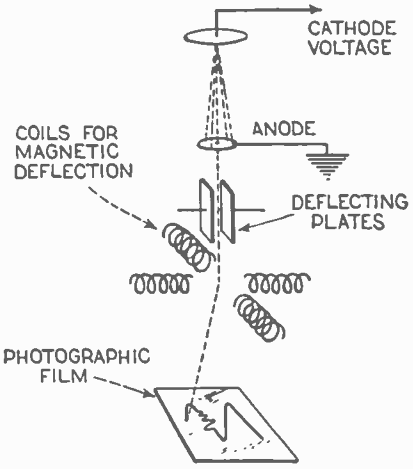

Fig. 2. A diagrammatic representation of the Dufour oscillograph.

OSMOSIS, ELECTRIC—When an electric current is passed through an electrolyte having an anode on one side of a porous diaphragm and a cathode on the other side, there is a tendency for the liquid to pass through the diaphragm towards the cathode raising the level of the liquid on the cathode side. This phenomenon is known as electric osmosis.

OTOPHONE—A device for the hard of hearing, which utilizes an extremely sensitive microphone connected to a two-stage vacuum tube amplifier. The instrument uses dry cell tubes and the batteries are contained in the case with the microphone. A small telephone receiver completes the outfit.

OUDIN RESONATOR — A device used for obtaining high-frequency brush discharges. Discharges of this nature are used in the investigation of resonance effects and also in medical work. The accompanying diagram explains the principle of the Oudin resonator.

B D is an uninsulated coil of copper wire wound on an insulating core with each turn separated from the next adjacent one. At B, the helix is connected to an induction coil and it is tapped by means of a sliding contact at C. The sliding contact is connected through a condenser to the induction coil H, with a spark gap G in parallel in the circuit as shown. The circuit BGFC is a closed oscillatory circuit receiving its excitation from the induction coil H. The circuit C D E is an open circuit in contact with the oscillatory circuit. By moving the sliding contact, C, a point is reached where a brush discharge appears at E, thus indicating that electric oscillations are being set up in the open circuit. (See *Brush Discharge, Corona,* also *Oscillator.*)

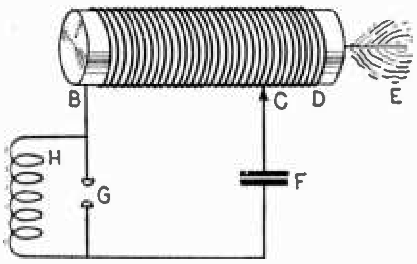

Connections of the Oudin resonator.

OUTPUT—The useful energy given out by a machine, circuit, vacuum tube or other device. (See *Input; Output, Transformer; Output Circuit of Vacuum Tube.*)

OUTPUT CIRCUIT OF VACUUM TUBE—The filament-plate circuit as differentiated from the filament-grid or input circuit. (See *Input Circuit of Vacuum Tube.*)

OUTPUT, TRANSFORMER—The product of the voltage at the secondary terminals of a transformer by the current flowing in the secondary winding. The output of any transformer is always less than the input by an amount equal to the losses. The *efficiency* of a transformer is the ratio of its useful output to the total *input.* (q.v.)

OVERLOAD RELEASE—An electromagnetic device which is used to cause a motor starter handle to return to the off position when the current exceeds a pre-determined overload value.

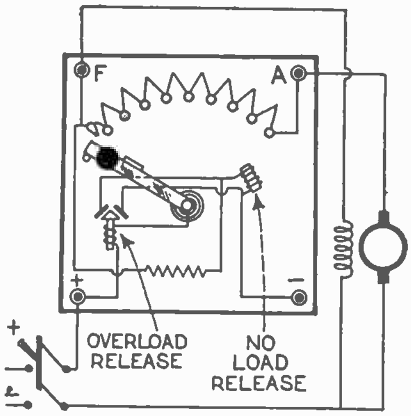

Overload Release.

The overload magnet is usually arranged so that too strong a current will result in short-circuiting the no-volt release thus switching off the driving current. One form of overload release consists of a magnet in series with the line and the armature. This magnet is arranged to attract a pivoted iron keeper whenever excessive current flows. The pivoted arm, when attracted by the magnet, makes connection between two contacts, themselves connected so as to short circuit the holding magnet when the circuit is closed. Therefore, when more than the allowable current is drawn from the line, the overload magnet attracts its pivoted keeper which closes the circuit short-circuiting the holding magnet. The contact arm is then released by the holding magnet and the arm is pulled back to the off position, thus cutting the motor off the line. Another form of overload release utilizes an electromagnet to attract a pivoted arm, arranged so that this will break the circuit between the motor and the line. This results in operating the no-voltage release. (See *No-Volt Release.*)

OVERTONE CURRENTS—Alternating currents which have harmonic frequencies of a higher value than the fundamental or first harmonics.

OVERTONES—Higher harmonics associated with the first harmonic. This term particularly refers to sound waves. In various musical instruments, different overtones are brought out and this results in each instrument having its own tone characteristic.

P

PACKING—(referring to a *microphone*). The tendency of carbon granules in a microphone to settle or pack. In this position they are no longer sensitive and do not readily respond to the vibrations of the diaphragm. Packing may be caused by irregularity in the sizes of the carbon granules. It may be remedied by gently tapping the microphone.

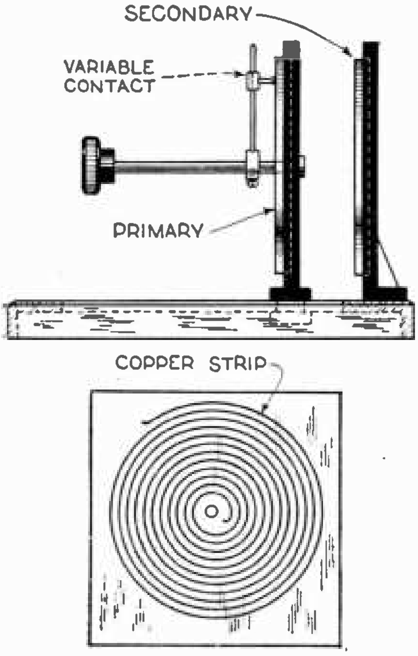

Front and side view of a pancake coil.

PANCAKE COIL—A flat inductance coil. Coils of this type are used in radio receiving sets as radio frequency transformers, also as antenna couplers and oscillator couplers. They may be *basket wound* (q.v.), spider wound, etc. In wireless transmitting apparatus, aerial tuning inductances, loading inductances, and oscillation transformers (illustrated) are frequently of the pancake coil type. (See *Inductance, Antenna.*)

PANEL—A board of insulating material carrying the controlling or measuring devices of an electric circuit. The panel for controlling the charging of storage batteries is called a *charging panel.* A panel carrying generator controls is referred to as a *generator panel.* Where only a single panel is used this is usually called the *switchboard.* The usual switchboard comprises a number of panels. Switchboard panels are commonly made of marble. The usual varieties available are white Italian, pink or grey Tennessee and blue Vermont marbles. Plain slate is sometimes used for panels where the voltage is not too high. When the panels do not require a finish, soapstone is sometimes used. The front of a radio receiving set on which the dials and rheostat controls, etc., are mounted is also called a panel. These panels are usually made of Bakelite, hard rubber, composition or similar insulating substances, although a number of sets are being made with metal panels. The horizontal panel on which the sockets and transformers are usually mounted is known as the *sub-panel* (q.v.).

PANNILL, Charles Jackson—American radio pioneer. Born in Petersburg, Virginia, May 13, 1879. He entered the American Navy, 1898. In 1902 he took a post under Professor R. A. Fessenden and carried out a series of wireless experiments at Hampton Roads, later inaugurating wireless communication between New York and Philadelphia. Pannill was the first man to install wireless on the battleships of the U. S. Navy. Afterwards he carried out a series of communication experiments between various parts of the United States and erected a number of wireless stations. He joined the Marconi Wireless Telegraph Company in 1912 and became a radio adviser to the United States Government, 1914, and assistant director of Naval Communications, 1916. Pannill is a fellow of the Institute of Radio Engineers and a member of the Washington Society of Engineers.

PAPER CONDENSER—A fixed condenser, usually made of tin foil and utilizing a dielectric of paraffin-waxed paper. The 1 microfarad by-pass condenser is a typical example of a paper condenser. (See *Condenser, By-pass.*)

PARALLEL—See *Parallel Connection.*

PARALLEL CONNECTION—Two or more parts of an electrical circuit so connected that the current divides between them. This is also known as a *multiple* (q.v.) connection, *divided circuit,* or *shunt* circuit. Where a number of parts of a circuit are connected in parallel, the same potential is impressed on each part. The current flowing in each branch will depend upon the impressed voltage and the resistance of each branch. (See *Circuit, Parallel.*)

PARALLEL RESISTANCES—Resistances connected so that their terminals have the same difference of potential between them. The greater the number of resistances in parallel, the less will be the total resistance of the circuit. If there are three resistances in parallel of ½, ⅓ and ⅕ ohm, respectively, then the total resistance is equal to the reciprocal (one divided by a number) of the sum of the separate reciprocals. In this case the sum of the reciprocals is 2 plus 3 plus 5 equal to 10 and the reciprocal of this is equal to one-tenth, so that the total resistance of this parallel circuit is one-tenth of an ohm. It can be seen that the total resistance of any circuit having a number of resistances in parallel is less than the resistance of any *one* of the branches. (See *Circuits Parallel,* also *Parallel Connection.*)

PARALLEL RESONANCE—When a concentrated capacity and a concentrated inductance are connected in parallel between terminals to which an alternating electromotive force is applied, and the inductance or capacity or frequency is varied, the condition of parallel resonance exists when the current supplied by the source is a minimum. It should be noted that in series resonance, the total current supplied by the source flows through both the inductance and the capacity whereas in parallel resonance the current supplied by the source is the vectorial sum of the two currents, one flowing through the capacity and one through the inductance.

PARAMAGNETIC—Having a permeability greater than unity. Magnetic as opposed to diamagnetic. The term *paramagnetic* is used in some cases to apply to substances which have only a slightly greater permeability than air such as liquid oxygen, the rare metal erbium, etc. If used in this sense it does not include *ferro-magnetic substances* (q.v.). (See *Diamagnetic material.*)

PARTIAL—An acoustical term denoting any one of the natural vibrations of which a body is capable. An electrical *oscillator* (q.v.) possessing distributed capacity and inductance has, theoretically, an infinite number of possible frequencies, each of which is called a *partial.* The lowest frequency is called *the fundamental.* (See *Harmonics, Harmonic Current, Overtones,* also *Overtone Currents.*)

PARTITION INSULATOR—An ebonite tube having a metallic rod running through its center with wing nuts at each end. It is used for continuing a circuit through a wall or partition.

P.D.—Abbreviation for *potential difference.* (See *Potential, Difference of.*)

PELTIER EFFECT—The change in temperature, either heating or cooling, at the junction of two unlike metals, depending upon whether an electric current originating from an external electromotive force, is sent through the junction in one direction or the other. The Peltier effect is the opposite of the *thermo-electric* effect.

PERCENTAGE COUPLING—The coefficient of coupling between the primary and the secondary of an oscillation transformer in a wireless transmitting system. When the primary winding is placed close to the secondary, the coupling is said to be *close* or *tight.* When the two windings are apart the coupling is said to be *loose.* With the ordinary spark gap in the closed circuit, if the primary and the secondary windings are closely coupled a broad wave will be radiated from the aerial system. If they are loosely coupled, a *sharp* wave will be radiated.

PERIKON DETECTOR—A crystal rectifier utilizing a piece of zincite in firm contact with a piece of chalcopyrite. (See *Combination Detector,* also *Crystal Detector.*)

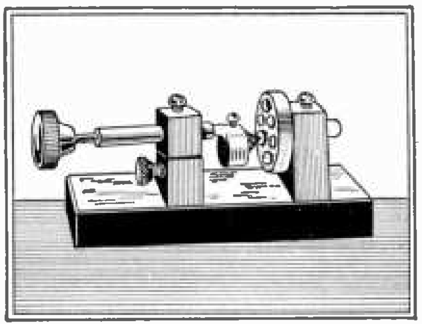

A Perikon detector.

PERIOD—The time required for one complete *cycle.* (q.v.) Referring to an *alternating current* (q.v.) it is the time required for the current or electromotive force to pass through the various values from zero to a positive maximum back to zero again, then to a negative maximum and finally to zero.

PERIODIC—A vibration is said to be periodic if all the phenomena are repeated regularly at equal intervals of time. This time is called *periodic time* or the *period.*

PERIODICITY—A synonym for frequency, or for angular velocity.

PERMANENT MAGNET—Hardened steel which when magnetized retains its magnetism after the removal of the magnetizing force. The usual method of making permanent magnets is to place the steel in a powerful electromagnetic field. They can also be made by stroking with another magnet. Permanent magnets are used in all electromagnetically operated headsets and *loud speakers.* (q.v.) (See *Magnet, Electromagnet,* also *Temporary Magnets.*)

PERMEABILITY—The *permeance* (q.v.) existing between the opposite faces of a cube of the substance each side of which is one centimeter in length. Since the permeability of air is assumed to be unity, the permeability of any substance is the ratio of the flux that passes through it to the flux that would exist in air if the magnetomotive and flux path remained unchanged.

PERMEANCE—That property of a magnetic circuit which allows the flow of magnetic flux. Its reciprocal is *reluctance* (q.v.) which is the property of a magnetic circuit by which it resists the flow of magnetic flux.

PERMITTANCE—A synonym for *capacity* (q.v.) or *capacitance* (q.v.). This property is possessed by every electrical circuit, but does not manifest itself unless there is a change in voltage. The greater the permittance of a circuit, the greater will be the opposition offered to a change in voltage.

PERMITTIVITY—See *Dielectric Coefficient and Constant, Inductive Capacity,* also *Inductivity.*

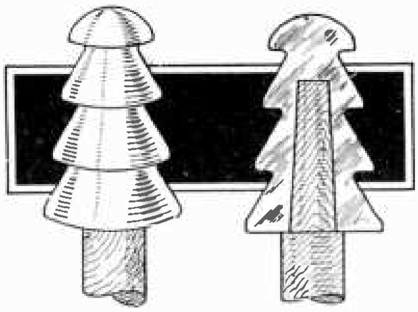

A petticoat insulator. Shown in cross-section at the right.

PETTICOAT INSULATOR—An insulator having two or more flanges or petticoats arranged one on top of the other for the purpose of reducing leakage and preventing accumulation of moisture. Petticoat insulators are usually made of porcelain. Insulators used on low tension lines are also of the petticoat type, in such cases being constructed of glass. (See *High Tension Insulator,* also *Insulator.*)

PHANTOM AERIAL—See *Mute antenna.*

PHASE—A particular state in a regularly recurring cycle of changes. In simple harmonic motion, uniform circular motion, or in periodic changes of any magnitude varying according to a simple harmonic law (as for example an alternating current) the point or stage in the period to which the rotation, oscillation or variation has advanced, considered in its relation to the position or instant of starting or to some other standard position. This relationship is usually expressed in angular measure. Defined in electrical terms, phase is the distance, usually measured as an angle, of the base of any *ordinate* (q.v.) of an alternating wave from any chosen point on the time axis. This distance represents the phase of the ordinate with respect to the point. In any sinusoidal alternating current quantity, the phase at any instant may be shown by the corresponding position of a *vector* (q.v.) revolving about a point with such an angular velocity that its projection at each instant upon a chosen reference line is proportional to the value of the quantity at that instant. As an example, consider a conductor rotating in a magnetic field. The induced electromotive force, in one complete revolution, passes through the following values. It starts at zero and rises to a maximum positive value, then falls to zero and rises to a maximum positive value in the reverse direction, finally falling to zero again. The transitions through these various values constitute a *cycle* (q.v.). The time of this transition is called a *period,* the number of cycles per second being the *frequency* (q.v.). The angle at any instant through which the conductor has rotated since the electromotive force changed its sign from negative to positive is the *phase angle.* This is ordinarily designated by the Greek letter Φ (phi). If ω (omega) is the angular velocity of the rotating conductor and f the frequency, then $\omega = 2\pi f$ (since there are 2π radians in a complete revolution). If the time in seconds which has elapsed since the conductor last passed the point of zero electromotive force be represented by t, then the phase Φ at that particular instant is equal to $\omega t = 2\pi ft$. Periodic variations are conveniently expressed by sine curves and the relative phases are indicated by the relative positions of the *nodes* (q.v.) and *loops* (q.v.).

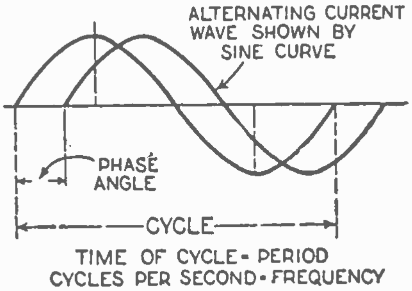
ALTERNATING CURRENT WAVE SHOWN BY SINE CURVE
PHASE ANGLE
CYCLE
TIME OF CYCLE = PERIOD
CYCLES PER SECOND = FREQUENCY

Phase displacement, cycle, etc., illustrated by means of sine curves.

Magnitudes having maximum and minimum values occurring simultaneously are said to be *in phase.* If they do not occur simultaneously, they are *out of phase.* When the corresponding maximum or minimum values of two sinusoidal alternating quantities of the same frequency occur at different instances, the two quantities are said to differ in phase by the angle between their nearest negative maxima or their nearest positive maxima or any other nearest corresponding values. The quantity whose maximum value occurs first is said to *lead* (q.v.) the other. The quantity whose maximum value occurs later is said to *lag* (q.v.). Where one quantity lags behind another by 180 degrees, the values are said to be of *opposite phase.* An alternating current may be *single phase* or *polyphase.* Another term for polyphase is *multiphase.* In a three-phase alternating current system the three separate currents differ in phase from each other by 120 degrees. In a polyphase system, if the various phases each carry or supply equal current, the system is said to be *balanced.*

PHASE ANGLE—See *Phase.*

PHASE DISPLACEMENT—The difference in phase or the phase angle between two alternating quantities, such as voltage and current, of the same frequency. Such quantities do not pass through their respective maximum positive quantities or through their zero values at the same time. The instantaneous values of any sine-shaped wave can be represented as being proportional to the sine of an angle, a complete cycle being equal to 360 degrees. When two quantities are out of phase, the respective angles corresponding to each quantity at any given instant are unequal, the difference between the two angles being called the *phase displacement* or the *phase difference.* In the accompanying illustration, the current and voltage relationship in an alternating current circuit are shown by the two sine curves. As can be seen, the current and voltage are out of phase, the phase displacement being equal to the angle Φ, in this case 90 degrees ($\frac{\pi}{2}$ radians). The current reaches its maximum value at a later time than the voltage. In this case the current *lags* behind the voltage by the angle Φ, the voltage leading the current by this angle. Since the angle of lag of the current is 90 degrees, the current and voltage are said to be in *quadrature.* In an alternating current circuit having pure inductance and no resistance (a theoretical circuit, since all electrical conductors possess resistance), the current would lag by 90 degrees as in the example given. (See *Lag, Lead, Leading Current,* also *Phase.*)

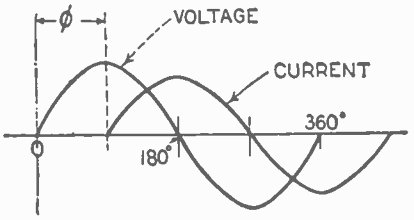

Sine curves of current and voltage, showing difference in phase.

PHENOMENA OF ELECTRIC WAVE PROPAGATION—Electric waves are created by electric *oscillations* (q.v.). They are often referred to as *radio waves.* Electric waves, light waves and radiated heat waves are included in the general term *electromagnetic waves* (q.v.). The velocity of electric waves is the same as that of light waves, 300,000,000 meters per second or 186,000 miles per second (appx.). Electric waves used for radio work have frequencies from about 10,000 to 3,000,000 cycles per second. Since the velocity of electric waves is known, it is possible to calculate the length of a wave if the frequency is known. Thus if the electric wave has a frequency of 10,000 cycles per second, the wavelength is found by dividing the velocity in meters per second by the frequency in cycles per second, which gives in this case a wavelength of 30,000 meters. Electric waves are in reality a combination of electromagnetic and electrostatic disturbances in space. The displacements are at right angles to the motion of the wave train. The electric field and the magnetic field are at right angles and travel together. They spread out or expand from the point of disturbance just as water waves spread out from a point where a stone strikes the surface of the water. Electric waves are propagated from an aerial when certain forms of alternating current flow in the aerial. Since the frequency of the propagated wave corresponds to the frequency of the current in the aerial, high frequency current must flow in the aerial to produce the high frequency electric waves used in radio. Such current may be produced by a high-frequency generator, by a direct current generator and an electric arc, or by means of vacuum tubes. The latter offer the simplest means of producing the rapidly reversing currents. Electric waves may be classified as *continuous* and *discontinuous.* Examples of the former are the waves produced by a high frequency alternator, an oscillating vacuum tube or a Poulsen arc. Examples of the latter are the waves produced by condenser discharges in a spark circuit. In these the amplitude of the waves diminishes in each wave train and the waves are referred to as *damped* waves. The high frequency current used to produce continuous waves used in radio telephony is *modulated* in accordance with the vibrations of a *microphone* (q.v.). As a result modulated waves are sent out from the aerial and these correspond in wave form to the variations of the sound waves. The sensitive apparatus at the distant receiving set picks up or intercepts the modulated electric waves, reconverts them to high frequency electric currents which are rectified into direct currents capable of actuating a 'phone and usually amplified so that they have energy enough to operate a loud speaker. (See *Oscillator, Carrier Wave,* also *Modulated Currents.*)

PHILLIPS, Raymond—British radio expert. Born October 6, 1879, he was educated at Edgebaston, Birmingham Schorne College, Buckinghamshire, and Windsor High School. For three years he was engaged on railway constructions and repairs, with special application to electric railways. In 1902 he invented a system of automatic train control and a number of appliances for electric railways which were widely adopted. In 1905 he took up the study of wireless and specialized in the control of mechanisms at a distance by electric waves. In 1910 he patented a system for controlling airships by radio and by means of a working model demonstrated how an airship could be thus controlled. During the World War he was inspector of ordnance machinery and in 1921 was appointed a member of the Inter-Allied Commission of Control in Germany. He is the author of many articles on the control of machines by electric waves.

PHONOGRAPHIC RECORDING—This refers to the amplifying of code signal currents by means of relays in tandem until the sounds produced in a telephone receiver are loud enough to cut a distinct record on the wax of a phonograph. The signals can be received at high speed and thereafter read at low speed. Instead of a phonograph a telegraphone may be used.

PHOTO-ELECTRIC EFFECT—A change in the electrical conductivity of a gaseous or solid substance when radiations of certain wavelengths come in contact with them. This refers particularly to rays of the visible spectrum and ultra-violet and infra-red rays.

PHOTOGRAPHIC RECORDER—A device for recording high-speed wireless messages by recording the deflections of a sensitive galvanometer on a moving photographic film. (See *Auto-Receiver.*)

PICKARD, Greenleaf Whittier—American radio expert. He was born at Portland, Maine, on February 14, 1877, and was educated at Westbrook Seminary, Lawrence Scientific School, Harvard and Massachusetts Institute of Technology. He made a special study of wireless telegraphy and telephony and has taken out many United States and foreign patents for radio inventions. He is noted for his pioneer work in radio telephony. Pickard began radio work in 1899 at Blue Hill observatory, Milton, Massachusetts, under a grant from the Smithsonian Institute. He was on the engineering staff of the American Telephone and Telegraph Company from 1902 to 1906. Later he was connected with the Wireless Specialty Apparatus Company as consulting engineer. He has an extensive practice as a patent expert in radio patent litigation. Pickard is a fellow of the American Institute of Electrical Engineers, a Member of the American Chemical Society and also is a Member of the Society of Chemical Industry.

Greenleaf Whittier Pickard.

PIG-TAIL—A flexible braided or stranded copper conductor used to carry the current to the rotor of a condenser or a coupler. Pig-tails are also used on the carbon brushes of motors and generators. (See *Flexible Lead.*)

PITCH—Frequency of vibration of a sound. A shrill note is said to be of high pitch, a bass note of low pitch. The pitch of an armature winding refers to the distance from the center of a winding to the center of the next winding.

PLAIN AERIAL—An aerial having the transmitting or receiving circuit connected directly to it without the utilization of inductive coupling or any intermediate tuned circuit. An example of a plain aerial is one used to transmit code, having a spark gap in series with the aerial and the ground.

PLATE—The anode or positive electrode of a vacuum tube. The vacuum tube plate is usually made of nickel. The shape of the plate varies according to the type and construction of the tube. In some tubes the plate is tubular; in others flat. When the cathode or filament is heated, it emits electrons which pass to the grid or through the grid to the plate, depending upon the grid charge. In power tubes (transmitting) it is necessary to use some cooling device to conduct away the heat dissipated at the plate. In some cases a liquid bath or a blast of air is used or the plate may be constructed so that it can be cooled by water circulation. Other methods of preventing overheating of these tubes are increasing the area of the plate as for example increasing the diameter of a tubular plate, or blackening the surface of the plate to increase its heat emissivity.

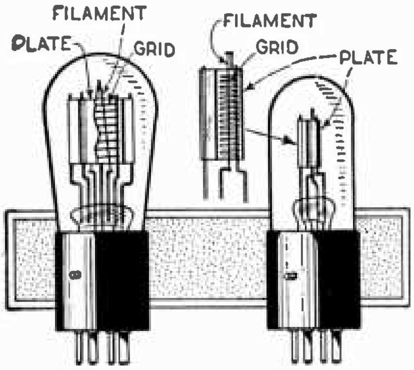

Vacuum tubes, showing construction of plates.

PLATE BATTERY—A battery, usually referred to as the *"B" Battery* having its positive terminal connected to the plate of the three-electrode vacuum tube and its negative terminal leading to the filament. The purpose of this battery is to keep the plate voltage positive with respect to the filament so that the electrons will be attracted sufficiently to the plate. Plate batteries are usually dry cells although there are a number of storage plate or "B" batteries on the market, both of the acid and the alkaline types. It is possible to dispense with the plate battery by using a *"B" eliminator*, which is connected to the ordinary house-lighting circuit. (See *"B" Battery.*)

PLATE CIRCUIT—The circuit connected to the *plate* (q.v.) or anode of a three-electrode vacuum tube.

PLATE COMPONENT, HIGH FREQUENCY—The radio frequency current flowing in the *plate circuit* (q.v.) of a vacuum tube. (See *High Frequency Component of Plate.*)

PLATE CONSUMPTION—This refers to the current consumed in energizing the plate of a vacuum tube. Increasing the grid voltage tends to increase the plate consumption. An increasing plate voltage also increases the plate current, but tends to decrease the relatively small grid current. (See *Consumption, Current;* also *Consumption, Plate.*)

PLATE CONTROL—The vacuum tube starts oscillating at a plate voltage which will depend on the adjustments of the circuit. When the plate voltage is raised, the oscillation current becomes greater, increasing in direct proportion.

PLATE CURRENT—The current in the *plate circuit* (q.v.) of the three-electrode vacuum tube. The current passing between the plate and the heated cathode or filament.

PLATE IMPEDANCE—The total resistance offered to alternating current between the filament and the plate in a vacuum tube. It is often referred to as the *tube impedance.* The impedance is independent of the alternating current input voltage.

PLATE VOLTAGE—The voltage of the plate of a vacuum tube above the filament. The potential difference between the plate and the filament. This term is usually used to refer to the voltage of the *"B" battery* (q.v.) connected to the plate, as for example, if the 22½-volt terminal of the "B" battery is connected to the plate of the detector tube, the plate voltage is said to be 22½ volts. Soft tubes, formerly used as detectors, operated best at this voltage. Most of the amplifier tubes now used as detectors operate best with 45 volts on the plate, although these tubes vary individually, some operating better with slightly increased voltage and others with a voltage less than 45. The plate voltage generally used on the amplifiers is 90 volts. The new power reception tubes use plate voltages up to 135 volts.

PLIODYNATRON—A special vacuum tube, designed by A. W. Hull, having two grids instead of one. This tube utilizes the normal amplifying property of the vacuum tube to obtain a very high voltage amplification, as high as 1,000 fold. When such high amplification is obtained, however, the operation is unstable and extremely careful adjustment is necessary.

PLIOTRON—A name given to a highly evacuated vacuum tube used for transmitting purposes. Large pliotrons are capable of developing as high as 500 watts output since they are so highly exhausted that several thousand volts may be applied to the plate and also since they can carry a plate current of nearly one-half an ampere.

PLUG—A connecting device used extensively in radio and telephone work for making connection or "plugging in" *loud speakers* (q.v.), head sets, power supply, microphone transmitters, etc. The plug has an insulating handle and the metal contact tip is designed to fit into a standard *jack* (q.v.). A flexible wire fits into terminals in the plug, the other end of the wire being connected to the apparatus such as the loud speaker, etc. The plug connection terminals may be of the conventional small binding post type or they may consist of internal spring clips.

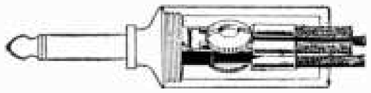

Cross-section of plug.

PLUG-IN COIL—An interchangeable coil, often of the honey-comb type, which can be plugged into a socket, thus permitting it to be quickly and conveniently put into or taken out of a circuit. Plug-in coils are used in some cases in receiving sets designed for operation on widely varying wave bands.

PLUG-IN TRANSFORMER—*A plug-in coil* (q.v.) specifically used as a radio frequency transformer.

POLARITY—In magnets the differentiation between *North Pole* and *South Pole.* In electrical devices or circuits, the difference between *positive* and *negative poles.* It is often necessary to determine which is the positive pole or terminal of a direct current circuit, and which the negative. There are various methods of indicating polarity. A voltmeter or an ammeter will show the direction of current flow. An extremely simple method of determining polarity is to place the two terminals in a glass jar containing salt water or some other electrolyte. A much greater number of bubbles will be observed to collect around the terminal of negative polarity. A piece of used blue print paper can be utilized to indicate polarity. It is simply necessary to moisten the paper and place the terminals on it, at a distance of about ¼ of an inch apart. The paper will be whitened at the negative terminal.

POLARIZATION—In a primary cell, the current liberates hydrogen from the electrolyte in passing through it. The hydrogen is carried to the negative plate with the current and collects there. It thus increases the internal resistance of the cell. This makes it harder for the current to pass through it and as a result the electromotive force available at the terminals is materially lowered. In addition there is a slight difference of potential between the negative plate and the hydrogen and this acts in opposition to the electromotive force of the cell. This further reduces the effective voltage of the cell. The entire effect is known as polarization. Polarization may be carried to such an extent that current can no longer flow from the cell. It may be counteracted or prevented by the use of *depolarizers* (q.v.). There are three different kinds of depolarizers. Chemical depolarizers are placed in the cell near the cathode. Substances such as manganese dioxide, potassium bichromate, nitric acid or cuprous chloride are used. When the hydrogen reaches the depolarizer it combines with it to form a new compound. Electrochemical depolarizers depend for their action upon the use of a substance which will liberate a metal such as copper at the cathode instead of hydrogen. Thus with no hydrogen present there is no polarization. Of course it is necessary to use the proper electrodes and the correct electrolyte to obtain this result. Mechanical depolarizers merely blow or sweep the hydrogen away from the cathode. Being the least effective of the three methods, the latter is seldom used. (See *Local Action.*)

POLARIZATION, ELECTRIC—The state of a dielectric when subjected to electrostatic forces. Electric polarization is synonymous with electric displacement.

POLARIZED RELAY—A relay in which a magnetized swinging arm is placed between poles of two electromagnets. When current passes, one pole must change, so that the arm is attracted by one and repelled by the other. A polarized relay is much more sensitive to weak currents than the ordinary relay.

POLES—See *Magnetic Poles.*

POLYPHASE—Referring to an alternating current system in which the circuits are divided up into two or more branches having their currents displaced in phase from each other. The term *multiphase* has the same meaning as polyphase. (See *Phase.*)

POLYPHASE ALTERNATOR—An alternating current generator wound in such a way that it supplies two or more currents which differ in *phase* (q.v.) from each other by definite phase angles. The usual form of polyphase alternator is the three-phase alternator.

PORCELAIN CLEAT—See *Cleat.*

POSITIVE ELECTRODE—In a primary cell, the zinc plate or *anode* (q.v.) is the *positive electrode,* while the *pole* of this electrode is the *negative pole* (q.v.) because it is negative in relation to the external circuit. In a storage cell, the lead peroxide plate, which is the cathode during discharge, is called the *positive electrode* and its pole the *positive pole.*

POSITIVE TERMINAL POLE—The pole or terminal out of which current is considered to flow. This is more or less of a handy convention, but is contrary to the present day knowledge of electrical flow inasmuch as electrons move from negative to positive.

POTENTIAL—See *Potential, Electric.*

POTENTIAL, DIFFERENCE OF—abbreviation P.D.—Difference of electrical level. Theoretically the potential difference between two points in an electric field is measured by the work done by the electric field on a unit point charge of positive electricity moved from one point to another without disturbing the field. Potential difference is measured in volts. It may refer to any two points in a circuit, even though there is no source of electromotive force between these points. When applied to circuit between two points not possessing a source of electromotive force, the potential difference between the two points under consideration is equal to the product of the current flowing times the resistance of the circuit between the two points. In the case of an open circuit, such as the terminals of a cell, the potential difference is equal to the electromotive force. (See *Potential, Electromotive force,* also *Volt.*)

POTENTIAL DROP—The fall in voltage in a circuit due to the resistance of the conductor. (See *Potential, Electric;* also *Potential, Difference of.*)

POTENTIAL, ELECTRIC—Electrical pressure, or the degree of *electromotive force* (q.v.). Potential is electrical level. Electrical potential can be compared with the "head" of water. Thus it is customary to measure the height or head of water above sea level, which is taken as an arbitrary base. In a somewhat similar manner, the potential of the earth is taken as zero and all potential is measured from this as an arbitrary base. With this understood, it is no longer necessary to refer to the degree of electromotive force, but simply to the electromotive force or the potential. See *Potential, Difference of,* also *Volt.*)

POTENTIAL ENERGY—Energy of position. When energy is available for the production of work, it is referred to as potential energy. Energy at work is *kinetic energy* (q.v.).

POTENTIAL RECTIFIER—A *crystal* (q.v.) or other form of rectifier which requires an initial current to pass through it to become sensitive.

POTENTIOMETER (Potential Divider)—A resistance connected across a battery or other source of electromotive force and equipped with a slider so that the voltage on another circuit may be varied from zero to the full potential of the battery. Where a gas content detector vacuum tube is used, the purpose of the potentiometer is to permit a close adjustment of the *plate battery* (q.v.) potential. The potentiometer is shunted around the filament heating or "A" battery, its resistance being so high that its power consumption is practically negligible. In tuned radio frequency sets, a potentiometer is used to control the tube oscillation. The advantage of being able to make a system oscillate is obvious, as the greatest amplification is obtained when the circuits are operated just at the point before self-oscillation starts. This condition is easily found in operation, since the instant self-oscillation starts, a whistle is heard in the phone or loud speaker when tuning to any station it is desired to receive. The use of a potentiometer to control oscillations in a tuned radio frequency circuit is illustrated in the diagram shown at Fig. 1. Potentiometers used in radio apparatus are usually made either from high resistance wire or from carbon granules. In electrical engineering work, the term potentiometer means, in general, an arrangement of circuits for measuring potential difference.

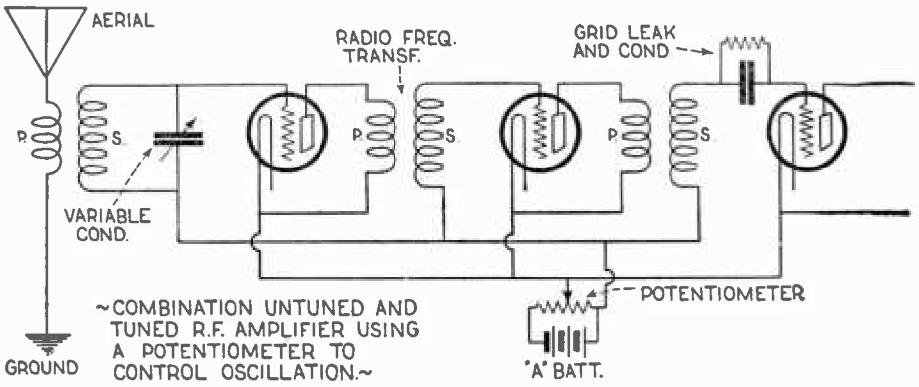
AERIAL
P
S
VARIABLE COND.
RADIO FREQ. TRANSF.
P
S
GRID LEAK AND COND
P
S
~COMBINATION UNTUNED AND TUNED R.F. AMPLIFIER USING A POTENTIOMETER TO CONTROL OSCILLATION.~
GROUND
POTENTIOMETER
'A' BATT.

Fig. 1. Tuned radio frequency circuit showing use of potentiometer to control oscillations.

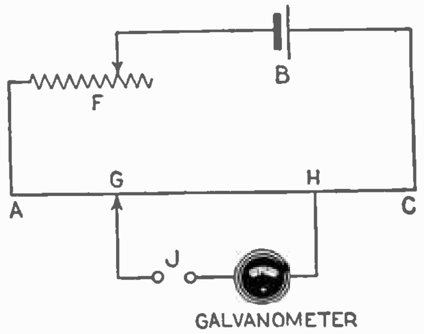

Fig. 2. A simple form of potentiometer used for electrical measurements.

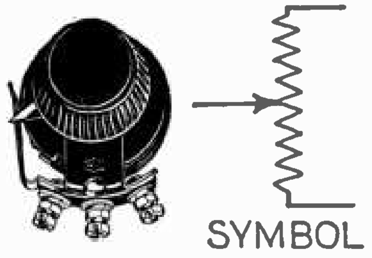

Illustration of a radio potentiometer, showing symbol at the right.

Using certain accessories, laboratory potentiometers can be used to measure voltages over all ranges, and by applying's *Ohm's law* (q.v.) the measurements obtained can be used to determine a wide range of current values.

The simplest form of potentiometer used for electrical measurements is shown in Fig. 2. AC is a resistance in which a constant current from the battery B is flowing. F is a regulating resistance used to compensate for the variations of the electromotive force of the battery. The current in AC is checked for constancy by noting that the drop in potential between the two points selected in it is equal to the electromotive force of a standard cell. The standard cell is connected in the circuit GHJ at J and the regulating resistance is adjusted until the sensitive galvanometer shows no deflection. With the assumption that AC has a uniform resistance throughout its length, and the current in it remaining constant, it is obvious that any other voltage not greater than the drop between A and C can be measured by connecting it in the circuit at J and shifting the points G and H until the galvanometer again comes to a balance.

A direct reading scale may be placed between A and C. For ordinary potentiometer measurements, the drop between A and C is made about 1½ volts as this is approximately the electromotive force of a standard Clark cell.

POULSEN ARC—An arc usually burned in hydrogen between a rotating carbon cathode and a water-cooled copper anode in a magnetic field. This arc is connected to an oscillating circuit, and due to its instability it maintains continuous oscillations which are utilized for the production of continuous waves for radio telegraphy and telephony.

POULSEN, Valdemar—Danish radio pioneer. Born in Copenhagen, November 23, 1869; he was educated at Copenhagen University and in 1893 he joined the Copenhagen Telephone Company. In 1903 he invented his famous system of arc transmission. Poulsen is a fellow of the Danish Society of Sciences. In 1900 he was awarded the Grand Prix at Paris.

POUNDAL—A unit of force. That force which acting on a mass of one pound gives it a velocity of one foot per second. It is equal to the weight of a pound mass divided by the acceleration of gravity; equal to 13,825 dynes or approximately half an ounce.

POWER—The rate of doing work. A certain amount of work may be done in a day or in a week. In the former case, more power is expended than in the latter case. Mechanical power is measured in *horse power* (q.v.). One horse power is equivalent to 33,000 foot-pounds per minute. Since power is the rate of doing work, the expression for power must contain a time element and a work element. Work is a measure of a force exerted through a distance. Hence the foot-pound measures work. Electrical power is the product of voltage times amperage. It is measured in *watts* (q.v.). (See *Kilowatt.*)

POWER AMPLIFIER—See *Power Amplification.*

POWER AMPLIFICATION—Amplification by means of apparatus consisting of a combination of power vacuum tubes, designed to carry larger currents than the ordinary tubes and specially designed transformers. Power amplifiers provide additional audio amplification and are used in connection with radio receiving sets where great volume is required as in the case of apparatus used for public address systems and for use with power operated loud speakers. (See *Amplifier.*)

POWER AMPLIFICATION COEFFICIENT—The ratio of the plate output power to the grid input power of a vacuum tube. This value may be as high as 10,000. The output power of a vacuum tube comes from the plate battery. When the grid is operated at low or negative voltage, the input power needed to produce the changes in grid voltage may be made very small compared with the output power of the tube.

POWER, APPARENT—The product of the voltage and the current in an alternating current circuit expressed in volt-amperes. This expression does not take into account any difference in *phase* (q.v.) between the voltage and the current. Thus in an extreme case (theoretical) where an alternating current flows in a purely inductive circuit, the current will lag behind the voltage by a phase angle of 90 degrees. In such a circuit, the *true power* (q.v.) would be zero, since the cosine of the phase angle between the current and the voltage would be zero, whereas the apparent power might be quite high and a heavy current might be flowing in the circuit. Apparent power is ordinarily expressed in *kilovolt-amperes* (q.v.). (See *Power Factor.*)

POWER FACTOR—The ratio of the *true power* (watts) in an alternating current circuit to the *apparent power* (volt-amperes). Where the wave form of the alternating current is sinusoidal, the power factor is equal to the cosine of the angle between the voltage and the current. Expressed as an equation, power P is equal to the product of the current I and the component of the impressed electromotive force which is in phase with the current, E cos Φ, where Φ is the angle of phase displacement between the current and the voltage ($P = IE \cos \Phi$). Expressed in another way, the power factor is the fractional amount which, if multiplied by the effective values of the current and the voltage, will give the value of the power expended. (See *Effective Electromotive force, Power, Apparent,* also *Power, True; Kilovolt-ampere, Kilowatt.*)

POWER, TRUE—The product of the *apparent power* and the *power factor* (q.v.) in an alternating current circuit. A wattmeter will measure true power, inasmuch as it takes into account the *phase displacement* (q.v.) between the current and the voltage. A voltmeter and an ammeter cannot be used to measure power, however, since the product of the two readings only gives the apparent power.

POWER TUBE—See *Vacuum Tube, Power.*

PRACTICAL UNITS—In the practical system of electrical units, the ohm, the ampere and the volt are the fundamental units. While the practical system is based originally on the Centimeter Gram Second systems, it is in no way dependent upon these systems for practical use. When used with the fundamental units of length and time, all other electrical as well as magnetic units can be derived from the fundamental units of the practical system. At the recommendation of the Chicago International Electrical Congress in 1893, the fundamental electrical units of the practical system were adopted as legal units in the United States by an act of Congress, July 12, 1894.

P R B—International Radiotelegraphic Convention abbreviation. Followed by question mark it means, "Do you wish to communicate by means of the International Signal Code?" Without the question mark it means, "I wish to communicate by means of the International Signal Code." (See *Abbreviations, International Radio Telegraphic Conventional.*)

PRESSURE—See *Potential, Electric.*

PRESSURE, ELECTRICAL—The term is synonymous with *voltage* (q.v.) or *electromotive* force (q.v.). Electrical pressure is the cause of current flow. It is measured in *volts* (q.v.). The flow of electricity along a conductor is analogous to the flow of water through a pipe. In order to make water flow, a pump is often used to force the water or in other words to apply pressure. Similarly, electrical pressure must be applied in order for an electric current to flow.

PRIMARY CELL—A device for transforming chemical energy into electrical energy. It consists of two different conductors placed in a liquid which acts chemically on one conductor, or more on one conductor than on the other, thus keeping a difference of potential between them. While the cell is being discharged either one or both of the conductors is used up and the liquid is also consumed. In many types of primary cells these can be replaced. The liquid is called an *electrolyte* (q.v.). In some cases an acid is used while in others a salt solution is the *electrolyte.* (q.v.) (See *Cell, Cell Secondary; Cell, Theory of Primary; Electrolysis, Local Action, Polarization,* also *Storage Cells.*)

PRIMARY CIRCUIT—A circuit supplying current to another which is called the secondary circuit. The two circuits are usually coupled by means of a transformer, the primary winding being in the primary circuit and the secondary winding being in the secondary circuit.

PRIMARY CURRENT—See *Current, Primary.*

PRIMARY ELECTRONS—The main stream of electrons given off by the hot filament of a vacuum tube, as differentiated from *secondary electrons* (q.v.) which may be detached from the plate by the impact of the primary electrons. (See *Negative Resistance.*)

PRIMARY TUNING INDUCTANCE—A variable inductance in the primary circuit of a radio transmitter. It is also referred to as the *aerial tuning inductance.* The aerial inductance coil may be continuously variable utilizing a sliding contact which bears on an edgewise copper strip or it may be of the drum type, in this case being made of stranded copper cable. (See *Inductance Coils,* also *Pancake Coils.*)

PRINCE, Charles Edmond—British radio expert. Born in 1874 at Cape Town and educated at Clifton and Faraday House, he joined the Marconi Wireless Telegraph Company in 1907, where he was concerned chiefly in research work on wireless telephony. In 1909 he carried out experiments in Italy and Switzerland with the first Marconi Field Station and made a number of valuable improvements in the Bellini-Tosi direction finder. In the World war, he joined the Royal Flying Corps and in 1915 he installed the first radio telephone on aircraft.

PRINTER, MORSE—See *Morse Inker.*

PROTON—A unit of positive electricity. The smallest quantity of positive electricity capable of existing in a free state. A minute particle or quantity of positive electricity carried by the nucleus of a hydrogen atom. See *Electron Theory,* also *Free Electrons.*

PULSATING CURRENT—A periodic current, that is to say a current passing through the successive equal cycles of values, the average value of which is not zero. A pulsating current is the sum of an alternating and a direct current. A current which varies regularly in magnitude is said to be pulsating while if it remains the same in magnitude it is called continuous. (See *Continuous Direct Current,* also *Current, Pulsating Direct.*)

PULSATING DIRECT CURRENT—See *Pulsating Current.*

PUPIN, Michael—American radio authority. Born in Hungary, October 4th, 1858, he came to the United States at the age of 10, and was educated at Columbia University. He also studied at Cambridge University, England. In 1891 he was appointed professor of mathematical physics at Columbia University, where he carried out an important series of researches in radio telegraphy and telephony. He is the holder of a large number of patents relating to telephony and radio, has done valuable research work on reso-

Michael Pupin.

nance and has developed radio circuits of extreme selectivity.

PUSH-PULL AMPLIFICATION—A method of connection in a vacuum tube receiving set, for supplying more power to the loud speaker than that ordinarily obtainable from a one or two stage audio amplifier. In the push-pull amplifier two tubes are used in the final stage of the amplifier, connected in such a way that they are used alternately on the two halves of each audio frequency cycle. A push-pull amplification circuit is shown in Figure 1. The first tube at the left gives the first stage of amplification and the other two in combination give the second stage. An ordinary audio frequency amplifier transformer is connected to the grid and filament of the first tube. Special transformers are usually used for the other two transformers. These have connections brought out from the center point of one of the windings as shown. The transformer between the two stages has a split secondary while the last transformer has a split primary. In cases where a tapped wound transformer is not available, two transformers can be used instead. In this case the two primary windings are connected in series and the two secondary windings are also connected in series. Figure 2 shows the method of connecting two transformers for this purpose, a connection being made to the mid-point between the secondary windings. It is essential that the two windings in series be connected so that they are continuous in the same direction around the transformer cores. In theory, the push-pull amplifier uses one tube during one half of the cycle and the other tube during the other half of the cycle. While the terminal of the secondary transformer winding leading to the grid of the tube is at a positive potential with respect to the middle of the winding, the tube is effective. During the remaining half of the cycle when the terminal is negative with respect to the tap, the tube is not operating. The action however is such that one tube is effective at all times. Although ordinary amplifier tubes can be used in the push-pull amplifier circuit it is desirable to use special power tubes which have a high amplification coefficient.

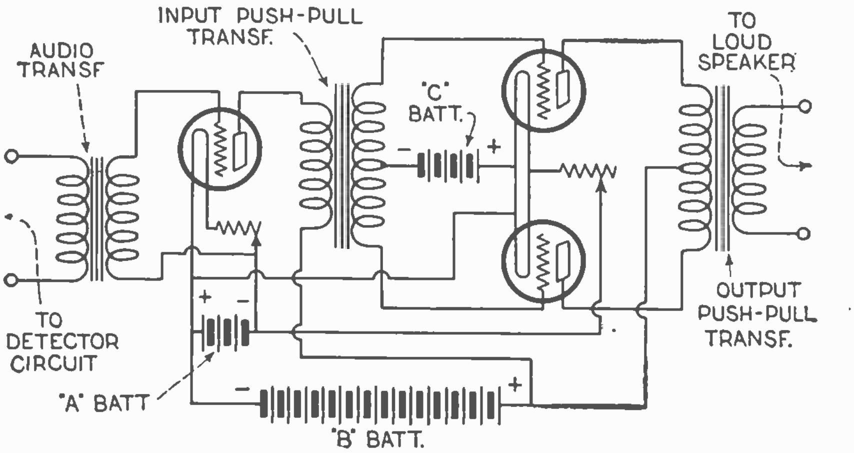

Fig. 1. Diagram of a push-pull amplification circuit.

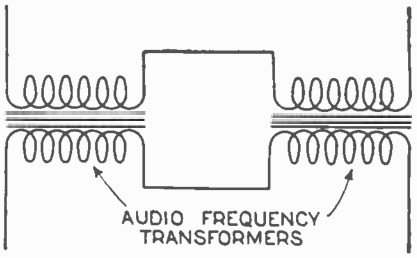

Fig. 2. Method of connecting two transformers where a tapped wound tranformer is not available.

PYRITES—Compounds of iron or copper and sulphur. Three of these are *Ferrous Sulphide* (FeS), *Bisulphide* (FeS_2), and an intermediate known as *Magnetic Pyrites* (Fe_3S_4). Pyrites crystals have been used in radio as detector crystals. (See *Crystal*, also *Crystal Detector.*)

PYRON DETECTOR—A *crystal detector* (q.v.) having an iron pyrites crystal and a copper or other metallic contact point.

Q

Q SIGNALS—International Radiotelegraphic Convention abbreviations. When the signal is followed by a question mark it asks the question. Otherwise the answer is indicated. A list of these abbreviations and their meanings follow:

Abbrev'n	*Question*	*Answer*
QRA	What ship or coast station is that?	This is . . . .
QRB	What is your distance?	My distance is . . . .
QRC	What is your true bearing?	My true bearing is . . . degrees.
QRD	Where are you bound for?	I am bound for . . .
QRF	Where are you bound from?	I am bound from . . . . .
QRG	What line do you belong to?	I belong to the . . . Line.
QRH	What is your wave length in meters?	My wave length is . . . . meters.
QRHH	What tune shall I adjust for?	Adjust to receive on tune . . . .
QRJ	How many words have you to send?	I have . . . . words to send.
QRK	How do you receive me?	I am receiving well.
QRL	Are you receiving badly? Shall I send 20 . . . — . for adjustment?	I am receiving badly. Please send 20 . . . — for adjustment.
QRLL	Request permission to test . . . minutes.	Permission to test granted.
QRM	Are you being interfered with?	I am being interfered with.
QRN	Are the atmospherics strong?	Atmospherics are very strong.
QRO	Shall I increase power?	Increase power.
QRP	Shall I decrease power?	Decrease power.
QRQ	Shall I send faster?	Send faster.
QRS	Shall I send slower?	Send slower.
QRT	Shall I stop sending?	Stop sending.
QRU		I have nothing to transmit. I have nothing for you.
QRV	Are you ready?	I am ready. All right now.
QRW	Are you busy?	I am busy (or, I am busy with . . . . Please do not interfere).
QRX	Shall I stand by?	Stand by. I will call you when required.
QRY	When will be my turn?	Your turn will be No. . . . .
QRZ	Are my signals weak?	Your signals are weak.
QSA	Are my signals strong?	Your signals are strong.
QSB	Is my tone bad? Is my spark bad?	The tone is bad. The spark is bad.
QSC	Is my spacing bad?	Your spacing is bad.
QSD	What is your time?	My time is . . . .
QSF	Is transmission to be in alternate order or in series?	Transmission will be in alternate order.
QSG		Transmission will be in a series of 5 messages.
QSH		Transmission will be in a series of 10 messages.
QSJ	What rate shall I collect for?	Collect . . . .

Abbrev'n	*Question*	*Answer*
QSK	Is the last radiogram cancelled?	The last radiogram is cancelled.
QSL	Did you get my receipt?	Please acknowledge.
QSM	What is your true course?	My true course is . . . . . degrees.
QSN	Are you in communication with land?	I am not in communication with land.
QSO	Are you in communication with any ship or station (or: with . . . .)?	I am in communication with . . . . (through . . . . .
QSP	Shall I inform . . . . that you are calling him?	Inform . . . . that I am calling him.
QSQ	Is . . . . . calling me?	You are being called by . . . .
QSR	Will you forward the radiogram?	I will forward the radiogram.
QSS	Are my signals fading?	Your signals are fading.
QST	Have you received the general call?	General call to all stations.
QSU	Please call me when you have finished (or: at . . o'clock).	Will call when I have finished.
QSV	Is public correspondence being handled?	Public correspondence is being handled. Please do not interfere.
QSW	Shall I increase my spark frequency?	Increase your spark frequency.
QSX	Shall I decrease my spark frequency?	Decrease your spark frequency.
QSY	Shall I send on a wave length of . . . . meters.	Let us change to a wave length of . . . . meters.
QSZ		Send each word twice. I have difficulty in receiving you.
QTA		Repeat the last radiogram.
QTC	Have you anything to transmit?	I have something to transmit. I have one or more radiograms for . . . . . .
QTE	What is my true bearing?	Your true bearing is . . . . degrees from . . . . . .
QTF	What is my true position?	Your true position is . . latitude . . . . . longitude.

QUADRANT—A term formerly in use as the unit of *inductance* (q.v.), now called the *henry* (q.v.). Inductance has the same dimensions as length and this unit is equivalent to 10^9 cm., this being approximately the length of one quadrant of the earth.

QUADRANT ELECTROMETER—A sensitive *electrometer* (q.v.) which in its calibrated form is known as an *electrostatic voltmeter* (q.v.). In the usual type, the "needle" consists of a thin aluminum plate suspended on delicate knife edges, with a pointer extending from the upper part to a scale. Two pairs of quadrant plates, having opposite faces metallically connected together, are placed on either side of the movable plate and parallel to its face. Each pair is insulated from the other pair. When the pairs of quadrants are subjected to a difference of potential, the needle, if highly charged, will be attracted into one pair or repelled out of the other. (See *Kelvin's Electrostatic Voltmeter.*)

QUADRATURE—The differing in *phase* (q.v.) of two alternating current quantities by 90 degrees. For example, when the current lags behind the impressed voltage by 90 degrees, as in the case of pure inductance in a circuit, the current and the voltage are said to be in quadrature. (See *Lag, Lead,* also *Leading Current.*)

QUANTITY OF ELECTRICITY—When referring to static electricity, it is the amount of the charge. Referring to electricity in motion, or an electric current, it is the amount of current flowing in a given time. Quantity of electricity is measured in *coulombs* (q.v.). The coulomb is the quantity of electricity which has passed when a current of one ampere flows for one second. The symbol which is usually used to signify quantity of electricity is Q.

QUANTOMETER—A type of ballistic galvanometer in which the swing of the movable system is proportional to the quantity of electricity passing through the instrument. A *galvanometer* (q.v.) of this type is used for magnetic testing as in the case of the *fluxmeter* (q.v.).

QUARTER PHASE—A two-phase alternating current system. An alternating current system having two currents differing in phase by 90 degrees or in other words by one-quarter of a cycle.

QUARTZ LAMP—A mercury vapor lamp which utilizes a tube of quartz instead of a glass container. Increased current densities can then be employed, since higher temperatures can be used with quartz than with glass.

QUARTZ OSCILLATOR—The piezo-electric quartz plate may be used as a *resonator* (q.v.) or as an *oscillator* (q.v.). When used as an oscillator it may be used either as a *master oscillator* controlling the frequency of a radio transmitting station's output or as a frequency indicator. All three of these uses are applicable to the work of maintaining station frequencies constant. The applicability of the piezo oscillator as a frequency indicator in broadcasting stations has been tested carefully by the United States Bureau of Standards, and found to be useful and satisfactory. This method gives a means of maintaining an extremely accurate check of the transmitting station frequency. (See *Master-Oscillator System,* also *Frequency Meter.*)

QUENCHED GAP—See *Quenched Spark Gap.*

QUENCHED SPARK—See *Quenched Spark Gap.*

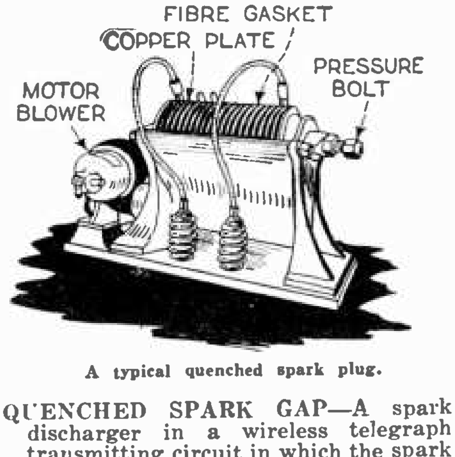

A typical quenched spark plug.

QUENCHED SPARK GAP—A spark discharger in a wireless telegraph transmitting circuit in which the spark is deionized to prevent the gap remaining conductive after the spark has passed. In this way only a few sharply defined oscillations are permitted to occur. The particular advantages of the quenched spark gap is that it enables the aerial to oscillate at its own frequency regardless of the period of the spark system. In one form of construction, a number of heavy copper plates are used, separated by micanite, fiber insulating washers or other insulating material. These are placed in an iron rack and compressed by means of a pressure bolt. The inside edge of the washer rests upon a groove which is cut in each plate. This makes it impossible for the spark to discharge over the edge of the washer. In order to make the discharge surface airtight, the washers are specially treated and as a result this helps in quenching the primary oscillations and makes the discharge noiseless. A small motor driven blower is mounted at the base of the spark discharger for the purpose of forcing air through the base of the rack against the cooling flanges of the plates. (See *Deionization, Musical Spark, Kinraidy Spark Gap,* also *Quenched Spark Transmitter.*

QUENCHED SPARK TRANSMITTER—A radio telegraph transmitter which employs a quenched spark gap. The quenched spark transmitter is not used for very high powers as it is not satisfactory for such use. In this case, a form of disc discharger is usually used. The quenched spark transmitter, however, has no moving parts and is noiseless in operation. In addition it gives large values of antenna current because closer coupling is possible. (See *Quenched Spark Gap.*)

QUICK BREAK SWITCH—A switch provided with auxiliary contacts which are released after the main contacts. The auxiliary contacts are separated rapidly, regardless of the speed with which the switch handle is pulled. This action is obtained by the tension of a spring connected between the main switch blade and the auxiliary blade, as shown in the illustration.

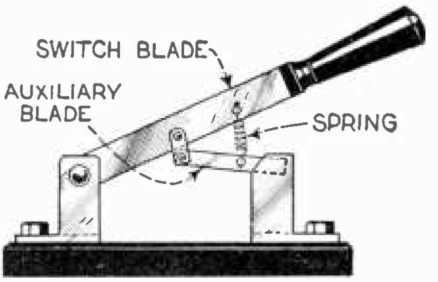

Quick Break Switch. Note the auxiliary spring-controlled blade.

QUIESCENT AERIAL—In a radio telephone transmitter, an aerial operated in such a manner that the carrier wave is suppressed when transmission is not actually taking place.

R

"R"—Symbol for *resistance* (q.v.)

RADIANT ENERGY—Energy transmitted through space by means of vibrations of the ether, such as electromagnetic energy, light or heat. (See *Actinic Ray.*)

RADIANT HEAT—Infra-red *radiation* (q.v.) as for example radiation from substances at temperatures insufficient to cause visible light rays.

RADIATING CIRCUIT—An electrical circuit which emits energy in the form of ether waves. In a transmitting system, the aerial circuit. Any circuit which radiates energy is a radiating circuit.

RADIATION—The transference of energy in waves through space which is not necessarily occupied by matter. Transference of energy through *ether* (q.v.). Electric or electromagnetic radiation refers to the radiation by means of electromagnetic induction. The radiation field is transmitted by wave motion. Various types of electromagnetic waves may be sent out by a radiating system. In one system the waves are known as *damped waves* (q.v.), in another *undamped waves* (q.v.) are used.

RADIATION EFFICIENCY—The radiation efficiency of an antenna at a given wave length is the ratio of the power radiated to the total power delivered to the antenna (resistance).

RADIATION FROM ANTENNA—In the case of a simple vertical antenna, electromagnetic waves are radiated in semi-loops since the lower end of the aerial is grounded. Fig. 1 gives an idea of the way in which the intertwined electric and magnetic fields radiate from the antenna. It should be noted that the electric and the magnetic fields are in planes which are at right angles to each other. The electric lines are shown in elevation while the magnetic lines appear in plan, with the wave form common to both being shown between. At any single point in space, the lines separate and come together like a bellows. After the wave has travelled a certain distance, the lines can be regarded as sections of planes and the electric lines can be represented by the diagram shown in Fig. 2. The magnetic lines can be shown by the same figure, provided this is considered to have been rotated about its horizontal axis through an angle of 90 degrees. Regardless of whether the aerial is of the open or closed coil type, the radiation will take place in the same general way, *loops* (q.v.) being formed and detached as described. (See *Electromagnetic Waves,* also *Height of Aerial.*)

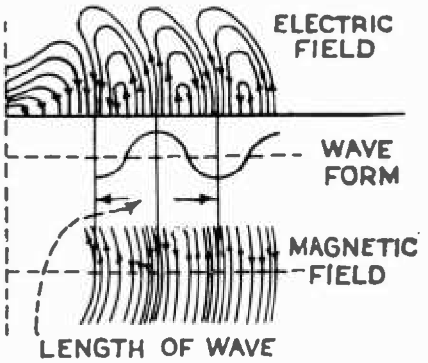

Fig. 1. Electric and magnetic fields about a vertical wire aerial.

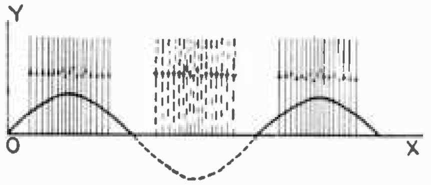

Fig. 2. Electric lines of force shown after the wave has travelled some distance from the aerial.

RADIATION RESISTANCE—The ratio of the total power radiated by an antenna to the square of the effective current at the point of maximum current. The property of an antenna by means of which radiation of energy takes place. Some of the power supplied to a circuit, carrying radio currents, is radiated in the form of electric waves. The radiation is a measure of the useful work obtained from the circuit. The greater the power radiated in comparison with the power dissipated in the circuit itself, the better will be the transmission. At any frequency, the power radiated is proportional to the square of the current flowing, hence the radiative effect may be regarded as causing artificially, an increase in the effective resistance of the circuit. This resistance increase is known as "radiation resistance" and is directly proportional to the square of the frequency or inversely proportional to the square of the wave length. (See *Antenna Resistance.*)

RADIATION TRANSMISSION—See *Radiation from Antenna.*

RADIO—The transmission of intelligence by means of electromagnetic waves through the ether. Radio falls under two general classifications; *radio telephony* (q.v.) and *radio telegraphy* (q.v.). (See *Radio Channel, Radio Communication, Radio Control,* also *Radio Frequencies.*)

RADIO-ACTIVITY—The emanation of energy by certain substances, such as radium, caused by a continuous detachment of electrons from the substance.

RADIOCAST—A term sometimes used instead of *Broadcast* (q.v.). To transmit speech, music or other sounds by *radio telephony* (q.v.).

RADIO CHANNEL—A band of wave lengths or frequencies of a width sufficient to permit of its use for radio communication without the radiation of subsidiary waves of more than a certain intensity at wave lengths or frequencies outside of such a band.

RADIO COMMUNICATION—The transmission of intelligible signals by means of electromagnetic waves. The signals may be in the form of code as in radio telegraphy or in the form of speech or music as in radio telephony. They may be transmitted through the *ether* (q.v.) without the use of intervening conductors as in ordinary radio, or along conductors as in *wired wireless* (q.v.).

RADIO COMPASS—A simple form of radio *direction finder* especially used for marine work. The radio compass depends for its principle of operation upon the directional effect of a coil antenna. Hence a coil antenna mounted on a vertical axis so that it can be rotated freely, is an essential part of the radio direction finder. By rotating the coil while a particular station is transmitting, it is possible to fix the line of direction of the station. The coil may be mounted so that the maximum signal will be received when the transmitting station lies in the plane of the coil. The device may also be set for minimum signal reception in which case the transmitting station lies in a direction perpendicular to the face of the coil. It is possible to obtain a much sharper determination of direction by setting the coil on the minimum position and hence this method is generally used for direction finder work. (See *Goniometer.*)

RADIO CONTROL—The control of motor driven devices such as marine vessels, trains, torpedoes, automobiles, etc., by means of special radio receiving apparatus designed to respond to radio impulses. E. P. Glavin's radio controlled automobile is an example of the radio controlled vehicle. *John Hays Hammond, Jr.* (q.v.) has invented radio controlled torpedoes, radio controlled ships and also radio controlled air craft.

R.F.—Abbreviation for radio frequency.

RADIO FREQUENCIES—Frequencies higher than those corresponding to normally audible sound waves. (See *Frequency, Radio; Frequency, High; Audio Frequency;* also *Frequency, Low.*)

RADIO FREQUENCY COILS—*Inductance coils* (q.v.) used to carry radio frequency currents or to couple radio frequency circuits. Radio frequency *transformers* (q.v.) may be of either iron core or air core types although the air core transformers are more generally used. (See *Low Loss Coils,* also *"D" Coils.*)

RADIO FREQUENCY CURRENT—An alternating current having a frequency of from 20,000 to 2,000,000 cycles per second. There is no hard and fast rule regarding the upper and lower limits, since much higher or much lower frequencies may still be considered as radio frequency currents. (*See High Frequency Current.*)

RADIO FREQUENCY SELECTIVITY—The radio frequency selectivity of a simple element of a receiving system is the ratio of resonant response (in terms of effective voltage or current measured at the indicator) to the nonresonant response when the radio frequency portions of the elements of that system are detuned by one percent of the resonant frequency. In this case, a simple element refers to a combination of an inductance, a capacitance and optionally a resistance.

RADIO FREQUENCY SIGNALS—Currents flowing in a radio receiving circuit up to the point of rectification. The currents flowing in the antenna circuit are radio frequency currents corresponding to the currents flowing in the transmitting aerial. (See *Amplifier, Radio Frequency.*)

RADIO GONIOMETER—See *Goniometer.*

RADIOGRAM—A message sent in code by radio. A radio telegraphic message.

RADIOPHARE—A radiotelegraphic "lighthouse" intended to aid navigation by emitting characteristic signals. By estimating the bearings of two charted radiophares, the mariner may determine the position of his ship.

RADIO TELEGRAPHY—The transmission of messages by means of electromagnetic waves using dot and dash code signals. A transmitter furnishes a source of high frequency alternating current which oscillates rapidly in the antenna system thus sending out electromagnetic waves. These are intercepted at the receiving station by the receiving aerial. At the receiving aerial, the electromagnetic waves are changed back into alternating current which is rectified by a detector in the receiving apparatus, then amplified and made audible in a *head telephone* (q.v.) or *loud speaker* (q.v.). Of course, a means of interrupting the current to produce dots and dashes must be present in the transmitter and suitable tuning devices are used in both the transmitter and the receiver.

RADIO TELEPHONY—The transmission of speech, music or other sounds by means of electromagnetic waves. In the standard systems of radio telephony continuous waves are used of a much higher frequency than those of audible sounds. The amplitude of these is modulated by the connection of a microphone in the transmitting circuit. The receiving apparatus is similar to that employed in radio telegraphy and is arranged so that the head set or loud speaker responds to the modulations of the waves and reproduces the sounds as originally produced at the transmitter.

RADIO TRANSMISSION—The sending of signals, speech, music, or other sounds by means of electromagnetic waves. The transmission of intelligence through the ether. (See *Radio Communication, Radio Control, Radio Telegraphy,* also *Radio Telephony.*)

RAT-TAIL—A single wire, or a group of wires used to connect the main part of an aerial with the *lead-in* (q.v.) wire.

RATIO—Relative values of quantities of the same kind, or number of times one quantity is contained in the other.

RATIO OF TRANSFORMATION—The ratio between the primary and secondary windings of a transformer. This determines the ratio of the primary to the secondary voltages. Thus if a transformer has a ratio of transformation of 3 to 1, then the secondary will have three times as many turns as the primary and the secondary voltage will be, approximately triple that of the primary voltage.

RATIO OF UNITS—The ratio of the unit pole in the electromagnetic system to the unit pole in the electrostatic system, or of the unit charge in the electromagnetic system to the unit charge in the electrostatic system. Its numerical value is approximately 3×10^{10}.

RAYLEIGH, John William Strutt—British physicist. Born at Langford Green, Essex, November 12th, 1842. He was educated at Trinity College, Cambridge. In 1879 he was appointed Cavendish professor of experimental physics in succession to Clerk-Maxwell, a post he held until 1884. In 1887 he became professor of natural philosophy at the Royal Institution and was appointed president of the Royal Society in 1905. In 1908 he was appointed Chancellor of Cambridge University, and died June 30th, 1919. Rayleigh was considered by many to have been one of the most brilliant experimental physicists of the nineteenth century. He not only had a remarkable power of mathematical analysis, but was equally skillful in carrying out the experimental proofs of his theoretical researches. Many of his experiments were made with home-built apparatus. He threw a fresh light on nearly every branch of physics, from the theory of gases to wave theories and electric and magnetic problems. Due to the care with which he carried out his experiments, he was the first to detect the presence of neon in the atmosphere, this being the forerunner of the discovery of a number of inert gases in the atmosphere. The Rayleigh balance for absolute current measurement is as well known as it is important. Lord Rayleigh was one of the original members of the Order of Merit. In 1882 he was awarded the Royal Medal, in 1899 the Copley Medal of the Royal Society and in 1904 the Nobel Prize for physics.

REACTANCE—That property of an electric circuit, aside from its ohmic resistance, which tends to oppose the flow of an alternating current. Mathematically, reactance is defined as the square root of the difference between the square of the *impedance* (q.v.) and the square of the effective resistance of a given portion of an electric circuit. Expressed as an equation reactance $x = \sqrt{z^2 - r^2}$ where z is the impedance and r the resistance of the part of the circuit under consideration. Reactance due to capacity in a circuit is capacity or condensive reactance. The reactance of a condenser of capacity C to an alternating current (of sine wave form) of frequency f is expressed by the equation $X_C = \frac{1}{2\pi fC}$. Reactance due to inductance in a circuit is *inductive reactance* (q.v.). The reactance of an inductance coil of inductance L to an alternating current (of sine wave form) of frequency f is expressed by the equation $X_L = 2\pi fL$. Reactance is measured in *ohms.*

REACTANCE COIL—REACTION COIL—Terms used in England to describe the *tickler coil* (q.v.) used in regenerative sets to couple back part of the energy in the plate or output circuit to the input circuit.

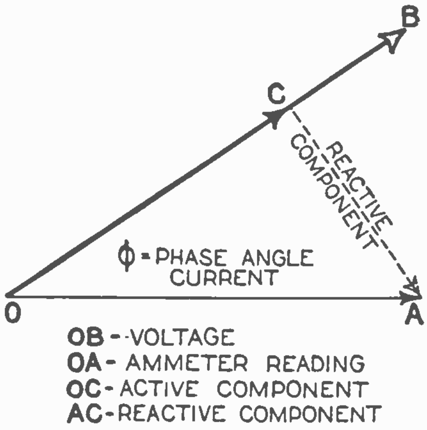

The current represented by vector OA may be considered as divided into two components, OC in phase with the voltage, and CA the reactive component.

REACTIVE COMPONENT—The component of an alternating current used to overcome the reactance of the circuit. This component acts at an angle of 90 degrees from the voltage. It is also known as *wattless current* (q.v.).

REACTIVE CURRENT—See *Reactive Component.*

REACTIVE DROP—The portion of the voltage drop in an alternating current circuit, due to the reactance.

REAL POWER—See *True Power.*

RECEIVER—A term rather loosely applied to cover the entire radio receiving apparatus within a station. Also sometimes used as a synonym for *Detector* (q.v.)

RECEIVING AERIAL—A wire or group of wires suspended at a suitable height and connected to a radio receiving set. The purpose of the receiving aerial is to intercept the electromagnetic waves. It should be placed well above surrounding buildings if possible, and should be well insulated from the ground. (See *Aerial,* also *Directional.*)

RECEIVING CIRCUIT—Any electrical circuit used in connection with the reception of radio. The more important vacuum tube receiving circuits are the *super-heterodyne* (q.v.), the *neutrodyne* (q.v.), the *reflex circuit* (q.v.), the *regenerative circuit* (q.v.) and the *tuned frequency circuit* (q.v.). There are various modifications and combinations of these circuits.

RECEPTACLE—A form of socket into which a plug is inserted for making electrical connections.

RECONSTRUCTED MICA—See *Micanite.*

RECORDER—A telegraphic instrument which automatically records messages sent over a telegraph system.

RECORDING WATTHOUR METER—An instrument for measuring and recording the total amount of electrical energy being consumed in a circuit. (See *Integrating Wattmeter.*)

RECTIFICATION—The changing of alternating current to direct current. In the usual radio receiving circuit rectification is accomplished by means of the vacuum tube, although the ordinary crystal detector rectifies the alternating current also. Before a storage battery can be charged from an alternating current source, the current must first be changed to direct current. Hence the battery charger designed to work on alternating current must be equipped with a rectification device, or rectifier (q.v.).

RECTIFIED CURRENT—An *alternating current* (q.v.) which has been changed to uni-directional or direct current either by preventing alternate waves from flowing or by reversing their direction. (See *Rectified Signals, Rectification,* also *Rectifier.*)

RECTIFIED SIGNALS—Radio signals which have passed through a detector, such as a vacuum tube or a crystal. Rectification of the oscillations results in audible or audio frequency currents which can be heard in a telephone receiver. (See *Rectification, Rectifier, Rectifying Tube,* also *Rectifying Action of Crystal Detector.*)

RECTIFIER—An apparatus for converting alternating current to unidirectional or direct current. Examples of rectifiers are certain crystals used as crystal detectors, the commutators of direct current motors and generators, vacuum tubes, electrolytic devices used for charging batteries, etc. (See *Rectified Signal, Rectification, Rectifying Tube, Rectifying Action of Crystal Detector, Rectified Current, Rectifier, Full Wave, Full Wave Rectification,* also *Half Wave Rectifier, Electrolytic Rectifier, Mercury Arc Rectifier, Charger, Storage Battery; Chemical Rectifier; Tungar; Rectigon.*)

RECTIFIER, FULL WAVE—A device for converting alternating current to direct current, in which both the positive and negative alternations are utilized. This differs from the half-wave rectifier which rectifies by suppressing alternate waves. (See *Full*

Wave Rectification, also *Half Wave Rectifier.)*

RECTIFYING ACTION OF CRYSTAL DETECTOR—The crystal detector has the property of allowing radio frequency alternating currents to flow through it in one direction only. In other words the alternating current is rectified. (See *Crystal Detector-Theory of Operation, Crystal Detector, Crystal Rectifier, Rectified Signals, also Rectified Current.)*

RECTIFYING TUBE—A vacuum tube used to convert alternating current into direct current. Rectification can be accomplished by a two-electrode tube, having filament and plate. In this case the electron flow is from the hot filament to the plate. A two-electrode vacuum tube made especially for rectifying purposes is known as the *kenotron* (q.v.). Another recent type of rectifying tube, known as the *helium tube* (trade name, *Raytheon*) utilizes a different operating principle, known as the "short path" principle, whereby the rarefied gas acts as an insulator between points which are in close proximity. This is in apparent contradiction of the observed phenomenon that the smaller the distance between two points the more readily a spark will jump between them due to the ionization of the gas. However, if the distance is small enough and a suitable gas is used at a low enough pressure, an electron may encounter no gas molecules in its path between the points and there will be no ionization by collision. Consequently, the inert gas helium, may be made to act as a perfect insulator at low pressures. Furthermore, it has been found that when a larger electrode of a gas conduction tube is negative, there is a greater current flow than when the smaller electrode is negative and that the smaller the positive relative to the negative electrode, the smaller is the back current and hence the more complete is the rectification. Figure 1 shows a typical helium tube, while Figure 2 illustrates its construction. The two small positive electrodes *A, A*, are carried through two small glass tubes imbedded in a lava insulating block *L* so as to project very slightly into a lava cup *C* whose wells constitute the negative electrode, being connected to the negative terminal through the base. The diameters of the small wires and the diameters and positions of the holes whereby they enter the cup are so proportioned as to give the necessary short path to the negative electrode. The cup *C* contains helium gas at such low pressure as to prevent gaseous conduction. This gives an insulation of great reliability and long life and makes possible extremely small anode surfaces which reduce the back current to a negligible quantity.

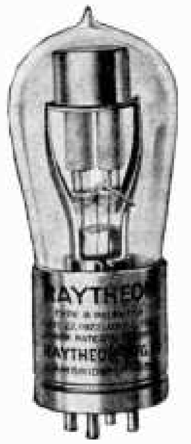

Fig. 1. A Raytheon tube.

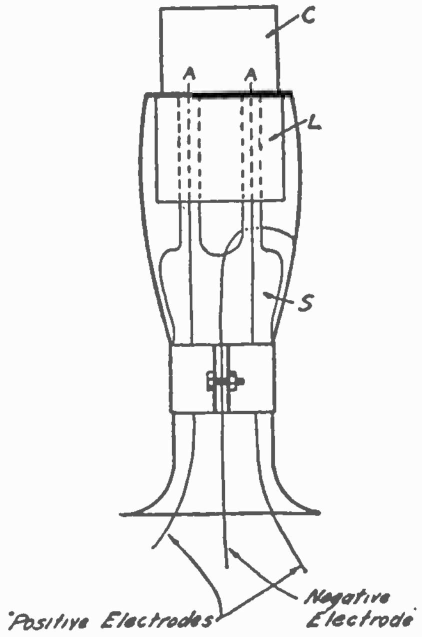

Fig. 2. Construction of the Raytheon tube.

RECTIFYING VALVE—This usually refers to a *rectifying tube* (q.v.) although it may refer to an *electrolytic rectifier* (q.v.)

RECTIGON—The trade name of a storage battery charger which utilizes the vacuum tube rectifier principle. (See *Charger, Storage Battery,* also *Rectifying Tube.)*

RED MAGNETISM—An obsolete term for the lines of force coming from the north pole of a magnet.

REDUCE WAVE-LENGTH OF AERIAL—The wave-length of an antenna may be reduced by the use of series condensers. However, this method should be avoided where possible, since it results in a loss of power. Shortening the aerial is a preferable method of reducing wave-length.

REFLECTING GALVANOMETER—See *Mirror Galvanometer.*

REFLECTION—Radio waves are reflected and also refracted when they pass into a region of different dielectric constant. A perfect conducting sheet would reflect all of the wave energy striking it. A conductor parallel to the direction of motion of a wave serves to guide the direction of the wave. The reflection and other modifications of radio waves serves to explain *fading* (q.v.) and other irregularities of broadcast signals. It has been found that broadcast signals will have most uniform intensity when long waves are used and during the daytime. Moreover, transmission will be better over a uniform conducting surface, such as the ocean. It is assumed that the space through which a radio wave travels is bounded by the surface of the earth and by an upper region which during the day is a fairly good conductor. The transmitted waves are guided by the two conducting mediums but in addition are assisted materially by reflection especially during the time that the air is partially ionized by the radiation of the sun.

REFLECTION COEFFICIENT—(Of a surface of discontinuity between two media.) The ratio of the reflected field intensity near the surface to the incident field intensity near the surface.

REFLEX—See *Reflex Circuit.*

REFLEX CIRCUIT—A radio receiving circuit invented by Marius Latour in which the vacuum tubes are made to perform double duties as both radio and audio frequency amplifiers. The incoming radio frequency current is amplified at radio frequency, rectified by a detector (which may be a crystal detector) and then amplified at audio frequency, using the same tube.

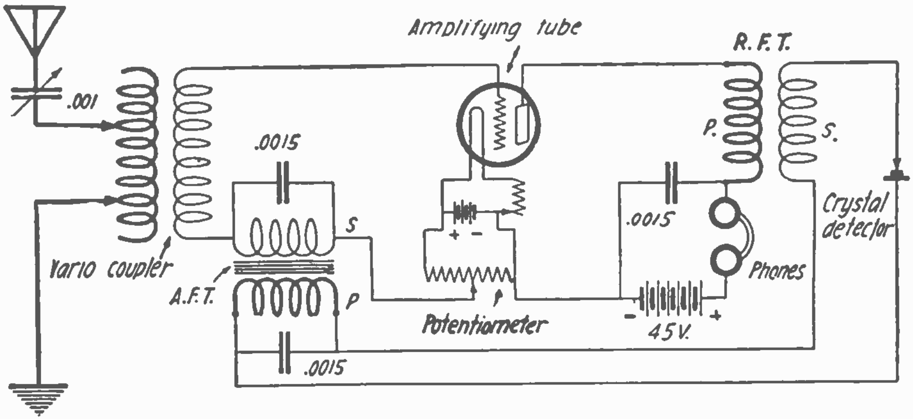

A typical reflex circuit using one tube and a crystal detector.

A typical four tube reflex set uses the second and third tubes as amplifiers simultaneously for both the radio and the audio frequency amplification. As a result the four tubes are made to produce the effect of six. Reflexing may be accomplished in a number of different ways. In some cases all the tubes are made to work twice. In other cases, only a part of the tubes are used for dual amplification. A variation of the reflex circuit, is the *Inverse Duplex Circuit* (q.v.). In this system the incoming signal reaches the first tube where it is amplified at radio frequencies and passes to the second tube, where it is again amplified at radio frequencies. From the second radio frequency tube, it is passed to the detector tube, where it is converted to audio frequency. The current then flows back to the second radio frequency tube which is now used as an audio frequency amplifier. Thence it flows to the first radio frequency tube which acts as the second stage of audio frequency amplification.

REFRACTION—The refraction of electric waves is analogous to the refraction of light. The electric (radio) waves are refracted when they pass into a region of different dielectric constant. That is to say, their direction is slightly altered. (See *Fading, Gliding Theory,* also *Reflection.*)

REFRACTION AND REFLECTION OF ELECTROMAGNETIC WAVES—See *Reflection,* also *Refraction.*

REGENERATION—The amplifying properties of the three-electrode vacuum tube can be used to obtain what is termed *regeneration.* Since it is possible to have greater output energy than input energy, part of the output may be returned to the input side, thus resulting in amplification of energy or in regeneration. By *feeding back* the correct amount of energy and in the right phase relation a constant reamplification may be obtained, and the tube can be made to act as a generator of sustained oscillations. (See *Armstrong Circuits, Feed-Back, Feed-Back Coil, Feed-Back Coupling, Feed-Back Effect,* also *Regenerative Circuit.*)

REGENERATIVE—See *Regenerative Circuit.*

REGENERATIVE CIRCUIT—A method of connecting up vaccum tubes for radio reception in which the amplified variations of the plate circuit are superimposed inductively on the grid or imput circuit thus producing a reinforcing effect which results in increased sensitivity. (See *Armstrong Circuits.*)

REGENERATIVE COUPLING—A radio receiving system designed to increase amplification in a vacuum tube, by coupling the plate or output circuit back to the grid or input circuit thus producing regeneration. (See *Feed-Back, Feed-Back Coil, Feed-Back Coupling, Feed-Back Effect,* also *Regenerative Circuit.*)

REINARTZ CIRCUIT—A selective regenerative circuit devised by John L. Reinartz in which a *spider-web* inductance coil is used, the coil acting as primary, secondary and tickler. Capacity *feed-back* (q.v.) is accomplished by means of a variable condenser. In the improved Reinartz a variocoupler is used, connected as a variometer, in addition to the condenser feed-back.

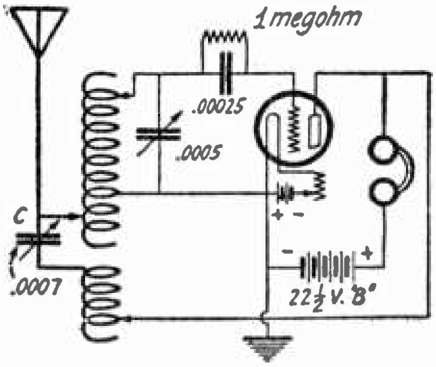

Diagram showing Reinartz circuit

REJECTOR—See *Rejector Circuit.*

REJECTOR CIRCUIT—A circuit consisting of a combination of inductance and capacity for filtering out or preventing the passage of currents of certain frequencies. This circuit is used in connection with a radio receiving set to obtain greater selectivity. The inductance and the capacity are arranged in parallel to form a resonant circuit. The constants of this circuit are chosen so that at the particular frequencies to be filtered, the capacity current will exactly equal the inductive current, thus giving zero resultant current. The primary application of the rejector circuit is as an *interference eliminator* (q.v.) or *wave trap* (q.v.). Figure 1 shows the application of the rejector circuit applied to a radio receiving set. The circuit shown is a combination *acceptor-rejector* circuit. The rejector circuit is shown in heavy lines. The inductance and the capacity are connected in parallel with each other and with the acceptor circuit and this latter is inductively coupled to the receiving set circuit. Both the acceptor and the rejector circuits are tuned to the wave length of the receiving aerial. By using more than one rejector and acceptor circuit increased selectivity can be obtained. A rejector circuit used to filter out undesired spark signals is shown in Figure 2. A small inductance of about ten turns is employed in this circuit and this is inductively coupled to a closed tuned circuit and wired in series with the aerial, and the aerial terminal of the set. If a spider-web coil is used in the closed circuit, a suitable value for broadcast wave lengths is a 50 turn coil, placing the aerial coil in such a position as to get closest coupling. The variable condenser should have a capacity of .0005 mfd. (See *Acceptor.*)

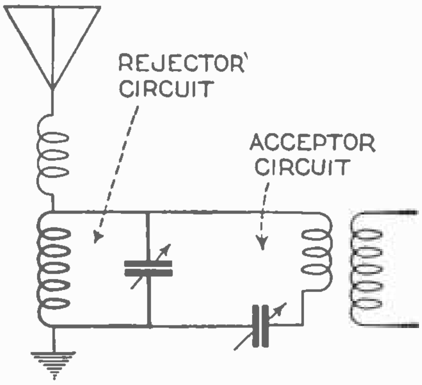

Fig. 1. An acceptor-rejector circuit used as a wave trap.

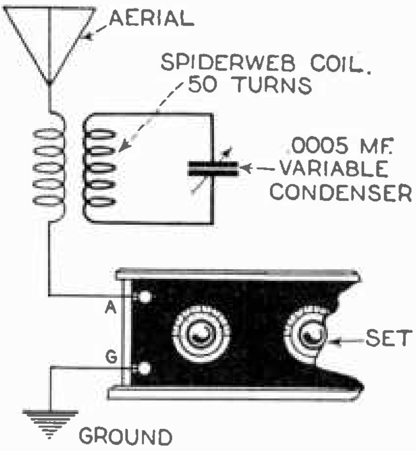

Fig. 2. A rejector circuit designed to filter out undesired spark signals.

RELAY—An electromagnetic device by means of which contacts in one circuit are operated by a change in conditions in the same circuit or in one or more associated circuits. (See *Magnifier.*)

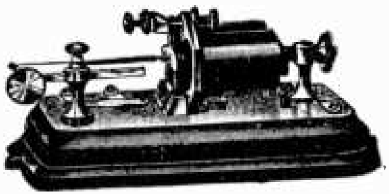

A telegraphic relay.

RELUCTANCE—Magnetic resistance. The reluctance of a magnetic circuit is analogous to electrical resistance. All magnetic circuits offer reluctance to flux. The name *rel* is widely used as the unit of reluctance. The rel is the reluctance of a magnetic circuit which will have one line of force in it, upon application of one *ampere-turn.* The rel can also be specified as the reluctance of a cylinder of air, or of any other non-magnetic material, one square inch in cross-sectional area, and 3.19 inches long. The flux through a magnetic circuit is equal to the *magnetomotive force* (q.v.) divided by the reluctance. (See *Magnetic Properties.*)

RELUCTIVITY—The *reluctance* (q.v.) per unit length and per unit cross-section. The reluctance per inch^3 or per cm^3. Reluctivity may be defined as specific magnetic resistance. It is the reciprocal of *permeability.* (q.v.).

REMANENCE—The term applied to the magnetism remaining in a magnet after the magnetizing force is removed.

REMOTE CONTROL—A control system whereby a radio transmitting or receiving apparatus may be operated at any desired distance from the apparatus. The chief advantage of remote control lies in the fact that the high tension apparatus and the tall aerial masts may be located in outlying districts whereas the studio can be located at any desired, convenient point. In addition it is possible to broadcast concerts, theatrical entertainments, prize fights, etc., direct from the scenes of action by utilizing remote control.

REPEATING COIL—A transformer of unity ratio used in telephone practice.

RE-RADIATION—The relaying of a radio wave through the action of a receiver employing regeneration. This phenomenon is most commonly observed in receivers of the *super-heterodyne* type.

RESIDUAL CHARGE—The residuary electricity retained by a condenser after discharge. This is also known as *electric residue* or *soakage.* After the condenser is discharged in the ordinary way, the residual charge permits a second discharge, smaller than the first. (See *Soaking in.*)

RESIDUAL MAGNETISM—The magnetism retained by a piece of iron, after the magnetizing force is withdrawn. After a mass of iron has become magnetized, some of the magnetism will remain, even though the magnetizing force has been taken away. The amount of residual magnetism depends on the quality of the iron. Pure wrought iron, in general, retains very little residual magnetism, whereas wrought iron which has gone through a hardening process, or which contains a large percentage of impurities and also cast iron possess a much larger amount of residual magnetism.

RESINOUS ELECTRICITY—An obsolete term for negative charges of electricity, adopted because of the fact that resinous bodies become negatively charged by friction. (See *Vitreous.*)

RESISTANCE—Symbol Ω (Omega)—Every conductor of electricity offers opposition to the flow of electricity. Resistance is a property of the conductor itself. It increases with increased length of conductor, decreases with increased cross-sectional area, and is greater or less depending upon the material of the conductor. Resistance is measured in *ohms.* (q.v.). The ohm is equal to 10^9 C.G.S. electromagnetic units.

RESISTANCE BOX—A box containing a number of wire coils of known resistances connected in various combinations, by means of switches or plugs, to produce any desired total resistance.

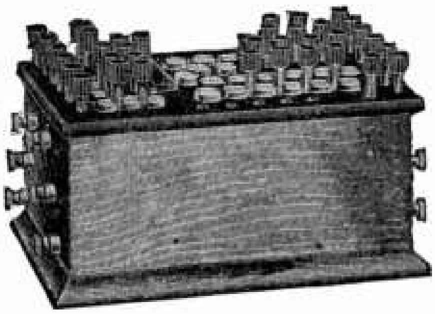

A resistance box.

RESISTANCE COUPLED AMPLIFIER—A type of *amplifier* (q.v.) which utilizes high resistances for interstage coupling. Resistance coupled amplification is generally used only for audio frequency amplification although it may be used for radio frequencies above 1000 meters. The chief advantage of resistance coupling consists in the purity of tone obtainable and the absence of distortion. In addition radio receivers of this type are comparatively inexpensive to build. It should be noted, however, that with resistance coupled amplification, the amplification per tube is much less than with transformer coupled amplification and furthermore the drain on the "B" batteries is also much greater. Of course, where a "B" eliminator is used "B" battery consumption is not a consideration. The resistances used for resistance coupling are non-inductive and may have resistances of from .1 megohm to 2 megohms. Figure 1 shows an elementary circuit utilizing a non-inductive resistance in the plate circuit of the tube. When a signal is impressed upon the grid circuit of this amplifying tube, there is a corresponding voltage set up in the plate circuit. The resistance R in the plate circuit causes a voltage drop to take place in two places in the tube. The total resistance of the plate circuit is equal to the internal resistance of the tube added to the resistance R, and if the "B" battery has 90 volts and if R is equivalent to the tube resistance at that moment the voltage drop across the tube will be 45 volts with the same drop across R. Theoretically it is possible to vary the voltage drop across R from zero to almost 90 degrees. Discounting the small resistances of the "A" and "B" batteries, a changing voltage will occur across the filament and plate when the grid potential varies, and also across the resistance R. Naturally these last variations depend upon the tube resistance and upon the value of the resistance R. These variations will be largest when R has a greater resistance than that of the tube. The resistance or impedance of the tube is constantly in a state of change and can assume very high values. For this reason the value of R should be equal to the higher values of the tube impedance in order to maintain maximum effectiveness. As used with ordinary amplifier tubes, a value of 100,000 ohms has been found to give satisfactory results. The next step is to apply the voltage variations across R to the grid circuit of the next tube. In Figure 2, the resistance R lies in the grid circuit of tube 2 as well as in the plate circuit of tube 1. Hence the magnified changes in tube 1 are transferred from R to the grid and filament of tube 2, thus the latter tube will be affected directly by the output circuit of the first tube and thus there will be amplification in both tubes thereby causing a resultant magnified signal to be heard. However, in such a hook-up, the same "A" and "B" battery cannot be used since this would mean that there would be a positive charge on the grid of tube No. 2, which would then be unable to function. Naturally, it would be impracticable to use a circuit having so many batteries, so a means must be found whereby it is possible to have a power supply consisting of but one filament battery and one plate battery for any given number of tubes in an amplifier. In order to do this the grids of the tubes must be separated electrically from the high voltage of the plates of the preceding tubes, but only as far as the steady voltage is concerned, while the fluctuating voltage of these plates must be impressed on the grids of the following tubes. This is done by introducing condensers between the two resistances. These effectively block the direct currents while they afford easy passage to alternating currents. Figure 3 shows the introduction of the condenser between the plate resistance and the grid resistance. Inspection of this diagram shows that although tube No. 1 remains the same, the plate instead of being connected directly to the grid of tube No. 2 is separated from the latter by means of the condenser C. If voltage fluctuations now occur across the resistance R, they will also be effective across the grid and filament of tube No. 2, because R exerts its influence between plate and filament of tube No. 1. As the condenser, C, conducts the variations very easily, the voltage fluctuations against R are then effective across the filament and grid of tube No. 2. It is necessary that tube No. 2 be provided with a grid leak, R_1 for the accumulated negative charges to leak off the grid to the filament. This grid leak at the same time serves the purpose of a by-pass for voltages, occurring due to possible leaks in the condenser C, which would tend to make the grid positive, thus allowing a grid current to flow and stopping the perfect action of the tube as an amplifier. Since the grid leak R_1 is a direct shunt across the grid and filament of tube No. 2, it must have a relatively high value in order to retain the highest possible amplification. R_1 should have a value lying between 500,000 and 2,000,000 ohms. If this resistance is made variable, it will be possible to control the resulting amplification completely. The value of the condenser C, must be chosen so that its resistance to the flow of alternating current will be as low as possible. Figure 4 shows three stages of resistance coupled amplification with a "C" battery used to give proper grid bias. The value of the resistance R, in the plate circuit of the tubes is 100,000 ohms and the value of the resistance in the grid circuits is 1,000,000 ohms (1 megohm). The condenser C has in each case a capacity of 0.05 mfd. The value of the "C" battery is about 4½ volts, but this should be determined by experiment for best results.

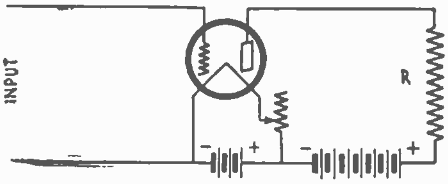

Fig. 1. The introduction of the non-inductive resistance in the plate circuit of the tube is the first step for explaining the principles involved.

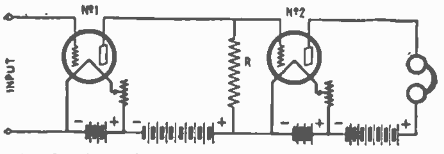

Fig. 2. In this circuit there must be two sets of batteries, but even then the grid of tube No. 2 will have a positive bias, which is undesirable.

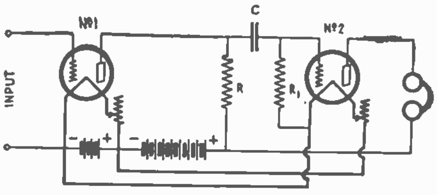

Fig. 3. The resistance coupled amplifier circuit with the introduction of a grid leak and condenser. The reasons for this are explained in the accompanying text.

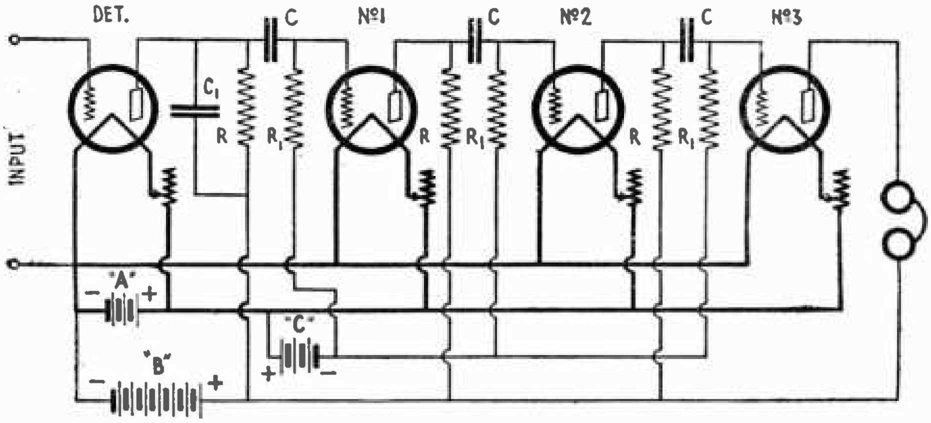

Fig. 4. Diagram showing three stages of resistance coupled amplification.

RESISTANCE COUPLING—RESISTIVE COUPLING—The transference of electrical energy from one circuit to another, by means of resistance common to both. (See *Resistance Coupled Amplifier.*)

RESISTANCE, HIGH FREQUENCY—The increased resistance of a conductor to high frequency currents, due to concentration of current at the surface of the conductor. This phenomenon is known as *skin effect* (q.v.) Skin effect becomes more pronounced, the greater the frequency of the impressed electromotive force. Skin effect also increases with increased cross-section of the conductor, with increased conductivity of the conductor and with increased magnetic *permeability* (q.v.). As a result of skin effect, the actual resistance of a conductor to alternating currents is greater than to direct currents.

RESISTANCE OF GROUND CONNECTION—The connecting wire leading from the radio apparatus to the ground should be of low resistance. If possible, the ground should be made by burying a copper plate in moist earth. The usual way to obtain a ground is to fasten a copper strap or *ground clamp* (q.v.) around a water or a steam pipe.

RESISTANCE (RADIATION)—This is the ratio of the total energy radiated (per second) by the antenna to the square of the root-mean-square current at a potential node (generally the ground connection).

RESISTANCE, RADIO FREQUENCY—The ratio of the heat produced per second, in watts, to the square of the root-mean-square current (radio frequency), in amperes, in a conductor.

RESISTIVITY—See *Resistivity of a Material*, also *Resistivity, Surface.*

RESISTIVITY OF A MATERIAL—SPECIFIC RESISTANCE—Symbol ρ (rho)—The resistance in ohms of unit length and unit cross-section of the material under consideration. Resistivity may be measured in ohms per inch cube or per centimeter cube, but more commonly it is measured in ohms per *mil-foot* (q.v.). The material of a conductor has an important bearing upon its resistance. Thus a unit length and unit cross-section of aluminum has about one and one-half times the resistance of copper having the same dimensions. Platinum has about six times the resistance.

RESISTIVITY, SURFACE—Referring to an insulating substance, surface resistivity is the resistance between two opposite edges of a surface film, 1 centimeter square. Surface resistivity depends to a considerable extent on humidity.

RESONANCE—A condition which exists in an electrical circuit where the frequency of the applied electromotive force is the same as the natural frequency of the circuit. Resonance is the cumulative effect produced by a periodic force in a circuit, so adjusted in frequency that the effect can attain the highest value. The effect measured is usually the current and the circuit when adjusted for maximum current, is said to be in resonance with the force, or with the circuit from which the force comes. The effect of resonance is to bring the current into phase with the electromotive force. Resonance in a circuit has the effect of bringing the reactance due to the inductance (inductive reactance) in exact opposition to the reactance due to capacity (capacity reactance). The total reactance is therefore zero and therefore the impedance is equal to the resistance. When two or more circuits have the same natural frequency they are said to be in resonance. Resonance of a circuit to a given exciting alternating electromotive force is that condition due to variation of the inductance or capacity in which the resultant effective current (or voltage) in that circuit is maximum. Instead of varying the inductance and the capacity, the frequency of the exciting field may be varied. The condition of resonance is determined by the frequency at which the current or voltage is maximum.

RESONANCE CURVE—A curve showing the relationship between the current induced in an oscillatory circuit and the frequency of the inducing current. In the case of two oscillating circuits placed near each other so that oscillations in one circuit will induce oscillations in the other, the induced oscillations will be weak or strong depending whether or not the second circuit is in tune with the first one. The capacity and the inductance of the second circuit can be varied until it is exactly in tune with the first at which time maximum current will flow in it. The resonance curve is obtained by measuring the current in the second circuit for various noted oscillation constants, and plotting the curve with the ordinates representing the current and the abscissae the corresponding natural frequency of the second circuit. The diagram shows a typical resonance curve. The peak of the curve is the point where the two circuits are in tune, or in other words, the point of resonance.

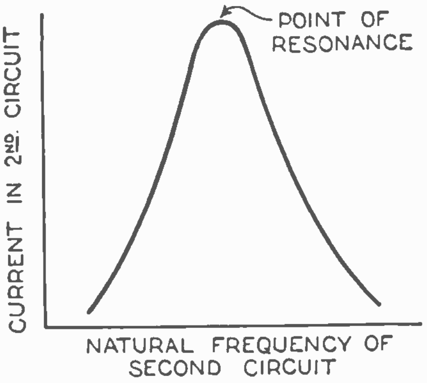

A typical resonance curve.

RESONANCE INTENSIFIER—An automatic receiver used for the reception of high-speed radio telegraph signals.

RESONANCE TRANSFORMER—Any loose-coupled tuning inductance having a primary and secondary each with a variable condenser in the circuit. Tuning the secondary circuit brings it in resonance with the primary, thus enabling signals to be heard with greatest volume.

RESONANT—See *Resonance.*

RESONATOR—A device for detecting by resonance, oscillations produced by an oscillator. This term is also used to refer to the sound box employed with the telegraph sounder.

RESONATOR, ACOUSTIC—A vessel containing a volume of air which can easily be set into vibration by sound waves or by the oscillation of a solid object. Typical forms are cylindrical or spherical and the ear is applied to a small tube opening into the air chamber.

RETARDATION COIL—A reactance coil used in a circuit for the purpose of selectively reacting on currents which vary at different rates.

RHEOSTAT—A variable resistance used to control current flow. Rheostats may be made of resistance wire or of carbon or for special purposes liquids may be used. In radio work, rheostats are connected in series with the filaments of vacuum tubes and the "A" battery in order to control the amount of current flowing in the filaments.

A rheostat constructed of resistance wire.

R. M. S.—Abbreviation for *Root Mean Square* (q.v.)

ROBERTS, Joseph Harrison Thomson—Born in 1890, he was educated at Liverpool and Cambridge Universities. He carried out many physical researches at the Cavendish Laboratory, Cambridge, and did a great deal of work on thermionics and other radio problems. He is a doctor of science and a Fellow of the Institute of Physics. He has several radio patents to his credit and is an expert in electrical, optical and acoustical sciences.

ROBINSON, James—British radio inventor. Born in 1884, he was educated at Durham and Gottingen Universities. He was appointed lecturer in physics at Durham University, 1906-07, lecturer in physics Sheffield University, 1910-12, and lecturer in physics at the East London College, London University, 1912-15. Robinson is known chiefly for his direction finding apparatus. He is a member of the Radio Society of Great Britain.

ROBINSON, Samuel S.—Rear-Admiral, U.S.N. He was born May 10th, 1867, and was educated at the United States Naval Academy, from which he graduated. He wrote the Manual of Wireless Telegraphy for Naval Electricians, and was in charge of the Bureau of Equipment, Navy Department.

ROOT MEAN SQUARE—The square root of the mean value of the square of an alternating current or voltage. An alternating current whose root-mean-square value is one ampere has the same heating effect as a direct current of one ampere. Root mean square is also referred to as *effective* or *virtual value.* The root mean square is derived as follows. In electrical measurements, it is necessary to compare direct and alternating currents. Moreover, the same units of measurement, the *ampere* (q.v.) and the *volt* (q.v.) are used in referring to both direct and alternating currents. However, in a direct current, a certain voltage will produce a steady current, whereas in an alternating current, both the voltage and the current are constantly varying both in magnitude and direction. The common basis of measurement is found in the value of the alternating current which will produce the same amount of heat as a certain direct current. Heating effect, according to *Joules Law* (q.v.) is directly proportional to the square of the current. Hence, in an alternating current varying according to a sine law, it is necessary to square each ordinate, find the mean or average value of these squares, and then extract the square root of this mean. Stated in concise terms, it is necessary to find the root mean square. An alternating current having a *maximum* value of one ampere would have a root mean square value of .707 amp. (See *Form Factor.*)

ROOT-MEAN-SQUARE VALUE—See *Root-Mean-Square.*

ROTARY CONDENSER—A synchronous motor used on a power line to advance the *phase* (q.v.) of the current. As a result there is an improvement in the *power factor* (q.v.). The action of the synchronous motor is similar to that of a condenser, whence the name rotary condenser.

ROTARY CONVERTER—A device for converting alternating current into direct current or vice versa. The rotary converter is also known as a *synchronous converter.* It is built like a direct current generator, except that it has certain commutator segments or conductors in the armature, connected to slip rings. Two, three, four, or six rings may be used. When the armature is rotated there is an alternating voltage between the rings. If the converter is driven by a motor or by an engine, it may be used as a *double current generator,* direct current being obtained from the brushes and alternating current from the slip rings (collector rings). When the collector rings are supplied with alternating current, the converter runs as a synchronous motor and direct current can be obtained from the brushes on the commutator. It should be understood that there is only one set of windings on the rotary converter which act simultaneously as alternating current motor and direct current generator. A converter which is used to convert direct current to alternating current is called an *inverted converter.* In the *single-phase* converter, there are two slip rings connected to the windings by as many equally spaced taps as there are poles. In the three-phase converter, three rings are used, with three equally spaced taps for each pair of poles. The single phase rotary converter must be brought up to synchronous speed by external power as it will not start itself. The polyphase converter will start itself from the alternating current side, but since it draws a very great lagging current, it should be started whenever possible from the direct current side. (See *Converter.*)

ROTARY CURRENT—A term infrequently used to denote *polyphase* (q.v.) current.

ROTARY DISCHARGER—A rotating disc discharger used in wireless transmitters in which the sparks occur between metal studs mounted on a motor driven disc. (See *Disc Discharger.*)

ROTARY FIELD—A magnetic field rotating in space. It may be due to rotating windings or stationary windings properly spaced through which polyphase currents are flowing.

ROTARY GAP—A form of rotating spark gap utilizing two or more rotating discs carrying studs between which the sparks pass. Spark gaps of this type were formerly used in wireless telegraph transmitters. (See *Discharger,* also *Disc Discharger.*)

ROTARY HYSTERESIS—Hysteresis (q.v.) produced by a rotating field as distinguished from that produced by a changing field.

ROTARY SPARK GAP—See *Rotary Discharger.*

ROTARY TICKER—A *ticker* (q.v.) in which the interruptions are produced by the chattering of a springy brush on a rapidly rotating metal wheel.

ROTOR—The rotating part of any electrical apparatus. It may refer to the revolving portion of a *motor* (q.v.) or *generator* (q.v.) or to the rotable part of a variable *condenser* (q.v.). The rotor in a direct current motor or generator is the armature. In an alternator or an alternating current motor, the rotor may be either the armature or the field.

ROTOR, GROUNDED CONNECTION—This refers to the grounding of the rotors of variable condensers in a radio receiving set. This is often done to eliminate hand capacity. Sometimes the rotors are connected to the shielding, which is grounded.

ROUGH CALIBRATION—Meters such as voltmeters, ammeters, etc., are often calibrated by comparison with standard instruments. In case such instruments are not available, it may be necessary to use less accurate methods. As an example of rough calibration, an experimental wave meter could be calibrated roughly by tuning in on various broadcasting stations whose wave lengths are known. (See *Calibration.*)

RUHMKORFF COIL—See *Induction Coil.*

RUTHERFORD, Sir Ernest—British physicist. Born at Nelson, New Zealand, he was educated at Canterbury College, Christchurch, New Zealand University, and Trinity College, Cambridge. From 1898 to 1907 he held the position of MacDonald professor of physics at Manchester University. In the latter year he was appointed Cavendish professor of physics and director of the Cavendish laboratory at Cambridge. Rutherford is an extremely brilliant physicist and has carried out a series of researches on the ultimate composition of matter. Many of these experiments have confirmed the researches of J. J. Thomson and others. Some of Sir Rutherford's experiments have revolutionized the theories of matter. He has received many honors for his brilliant researches including the Nobel prize for chemistry awarded to him in 1908. He has written extensively on radio activity, his books including "Radioactivity," 1904; "Radioactive Transformations," 1906; "Radioactive Substances and their Radiations," 1913; as well as many papers for technical journals. Sir Ernest was knighted in 1914.

S

SAFETY GAP—A spark gap which is connected in parallel with a condenser or other apparatus, and which protects the apparatus from excessive voltage. When the voltage exceeds a predetermined limit, the gap breaks down, thus passing the current around the apparatus. (See *Gap, Micrometer.*)

SAL AMMONIAC—A term commonly used for ammonium chloride. This substance is a white crystal, very soluble in water, and is used as an electrolyte in primary cells and also in many electroplating processes. (See *Cell.*)

SATURATED—The magnetic condition of a substance when increase of magnetizing force produces no further increase in flux density. A point is reached in every magnetic material, where further increase of magnetic intensity will not produce any increase of flux density. At this point, the material is said to be saturated. Cast iron becomes saturated at a lower flux density than steel or wrought iron.

SATURATION—See *Saturated.*

SATURATION, MAGNETIC—See *Saturated.*

SECONDARY CELL—A unit consisting of a positive and a negative electrode immersed in an *electrolyte* (q.v.) which becomes active only after the passage of an electric current. The electrodes are of metal or of a metallic compound. An electric current, flowing into a secondary cell causes certain chemical actions to take place whereby electrical energy is transformed into chemical energy. This process is termed "charging." After the cell is charged, the reverse action takes place and chemical energy is transformed into electrical energy. In this case, the cell is able to deliver an electric current to an outside circuit. This process is termed "discharging." The secondary cell is also known as an *accumulator* or *storage cell.* However, the cell does not store electricity. It merely holds or stores the chemical energy for immediate transformation into electrical energy. It is usual practice for several positive plates to be connected in multiple to form one positive electrode and for several negative plates to be connected in parallel to form a single negative electrode. The *storage battery* consists of two or more storage cells. The operation of the storage cell depends upon the fact that certain metals immersed in an electrolyte may differ in potential. In the secondary cell, no electrical action takes place until after charging. Thereafter, chemical action maintains a difference of potential until the stored energy has been retransformed. While the negative electrode is used up to a certain extent during the process of discharging, it is replaced electrochemically during the charging process, due to the flow of current through the electrolyte in the opposite direction to that of discharge. The two principal types of storage cells are the lead cell and the nickle-iron-alkaline (or Edison Cell). In lead cells, the negative electrode is made of spongy lead and the positive electrode is of lead peroxide. The electrolyte is dilute sulphuric acid. The following chemical action takes place within the lead cell. When the cell is charged, hydrogen is released at the negative plate, and oxygen at the positive plate. The oxygen combines with the lead of the positive plate to form lead peroxide while the hydrogen reduces the oxide in the negative plate. When the cell is discharged, both electrodes are changed to lead sulphate and the liberated hydrogen combines with the oxygen to form water. Care must be taken not to allow the cell to discharge too far as in this case, too great an amount of lead sulphate may form and this may cause buckling of the plates or may make it very difficult to recharge the cell. The indications of a completely charged cell are increased gassing of electrolyte together with a stop in the rising of the voltage and the specific gravity. In the Edison-type secondary cell, the active materials are of nickel and iron. Nickel hydrate forms the active material of the positive plate, while iron oxide forms the active material of the negative plate. The electrolyte is a 21 per cent. solution of potassium hydrate combined with a small amount of lithium hydrate. The chief advantage of the *Edison cell* (q.v.) is that it can be left idle indefinitely in a fully charged, semi-charged, or fully discharged state without any harm to the cell. In addition it is lighter and smaller than the same capacity lead cell. The unit used for measuring the capacity of storage cells is the *ampere-hour* (q.v.). This is based on the rate of discharge normally used in measurement, the eight-hour rate. As an example, an 80 ampere-hour storage battery will furnish a continuous current of 10 amperes for 8 hours. If the battery is discharged at a more rapid rate, it will not have the rated capacity.

SECONDARY CURRENT—The current flowing in the secondary windings of inductive apparatus. Current induced in a secondary winding due to the action of *primary current.* (See *Current, Primary.*)

SECONDARY ELECTRON EMISSION—See *Secondary Electrons,* also *Secondary Emission.*

SECONDARY ELECTRONS—Electrons emitted by the plate of a vacuum tube when struck by the electron stream from the filament. (See *Negative Resistance,* also *Secondary Emission.*)

SECONDARY EMISSION—Electron emission in which the exciting agency is bombardment of the emitting material by electrons. The emission of electrons from cold electrodes under the impact of other electrons. (See *Negative Resistance.*)

SELECTIVE—See *Selectivity.*

SELECTIVITY—The sharpness with which a radio receiving set can be tuned to a particular wavelength. A selective set is one which tunes sharply, whereas a non-selective set tunes broadly. Too long an aerial will often result in making a set tune broadly. A short aerial makes the receiving set more selective but reduces the volume.

SELECTOR—A tuning device connected in a radio receiving circuit for the purpose of eliminating undesired stations. (See *Wave Trap.*)

SELENIUM CELL—A cell constructed of a thin layer of the metal selenium. This metal is extremely sensitive to light. In its crystalline form, it is a conductor of electricity, its conductivity varying with the amount of light allowed to fall on it. As a result it is possible to regulate the flow of current in a circuit by varying the light. The selenium cell thus offers a means of transmitting pictures by wire or radio. Selenium cells are also being used in experimental devices for the transmission of motion pictures and for obtaining *television* (q.v.).

SELF-CAPACITY—The capacity of a circuit or a portion of a circuit. Every electrical circuit possesses inherent capacity, which varies in degree, depending upon the shape, size, etc., of the circuit. Self-capacity is also referred to as *distributed capacity* (q.v.). In general self-capacity is detrimental, especially in radio circuits, since it introduces losses. By using special forms of windings, the self-capacity of coils can be greatly reduced. (See *Low Loss Coils.*)

SELF-EXCITED GENERATOR—A generator which depends upon its own current for its field excitation. Such a generator must build up its magnetic field and hence its voltage, gradually. In order to get a start, it depends upon the *residual magnetism* (q.v.) present in the iron of the poles. There are three types of self-excited generators, the *shunt generator* (q.v.), the *series generator* (q.v.) and the *compound generator.* (See *Compound Wound.*)

SELF-HETERODYNE—A system of reception of continuous wave signals by the production of audio frequency beats, through the use of a device which is both a radio frequency generator and a detector of the audio frequency beat currents produced.

SELF-INDUCTANCES—See *Inductance, Self.*

SELF-INDUCTION—Inductance due to the field produced by the circuit itself. Changes in current are always accompanied by self-induction. When the current is decreasing, the electromotive force of self-induction is in the same direction, thus tending to prevent the decrease. When the current is increasing, the electromotive force of self-induction tends to prevent the increase. (*See Inductance, Self; Mutual Inductance,* also *Induction.*)

SELF-INDUCTION COEFFICIENT—The ratio of the number of lines of induction produced by a current flowing in the circuit divided by the current in the circuit. The coefficient of self-induction of a conductor of finite cross-section may be defined as the ratio of twice the energy of the magnetic field produced by the current flowing to the square of the current. The following factors affect the coefficient of self-induction of a circuit: the cross-section and shape of the conductor, the size and shape of the circuit, the permeabilities of the surrounding medium and of the conductor itself. Where the skin effect is large, the frequency of the current and the specific resistance of the conductor will also affect the coefficient of self-induction.

SELF-OSCILLATION—The tendency for the vacuum tubes in a radio receiving set to act as generators of radio frequency oscillations. (See *Body Capacity,* also *Bolitho Circuit.*)

SENSITIVE SPOTS—Points on the mineral of a crystal detector which show a greater response to incoming oscillations than others. Dust, grease, or dampness reduce the sensitivity of these spots and therefore the crystal should be properly protected.

SEPARATELY EXCITED GENERATOR—A generator in which current for producing the magnetic field is obtained from some source external to the generator itself. This source is either another generator or else a storage battery. Separately excited generators are generally used only for special work such as battery charging, electroplating and testing.

SERIES—See Series Connection.

SERIES CONNECTION—When two or more parts of an electrical circuit are so connected that the same current flows through them they are said to be in *series* and the circuit is called a *series circuit,* the connection being a *series connection.* In order to connect primary or other cells in series, it is necessary to connect the positive terminal of one to the negative terminal of another, etc. In this case, the total voltage of the battery equals the sum of the separate voltages.

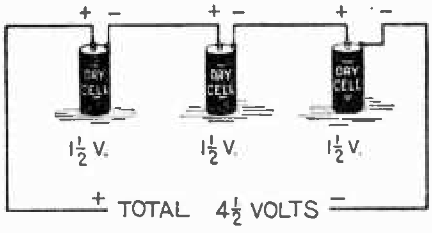

The three dry cells are here shown connected in series.

SERIES GENERATOR—A generator in which the exciting winding is connected in series with the armature and the load. This type of generator is of small commercial importance being used mainly for certain constant current systems.

SERIES MOTOR—A commutator-type motor having the field and armature windings connected in series. Ordinarily, the series motor is a direct current motor, but with minor modifications, such as lamination of the field as well as armature, this motor can be used for alternating current also. With the applied voltage constant, the greater the load, the less will be the speed of a series motor. As the load is lessened, the motor will speed up, and if the load is entirely removed, it will "run away." The *torque* (q.v.) characteristic is the most important feature of a series motor. Instead of varying directly with the current as in the shunt motor, the torque increases with the square of the current.

Hence, if the current is doubled, the torque will be increased four times, while if the current is quadrupled, the torque will be increased sixteen times. For the same current, the torque of the series motor is higher than that of the shunt and in addition, the series motor has a much better starting torque.

SERIES-PARALLEL—This refers to a circuit which is a combination of a *series circuit* and a *parallel circuit.* The term *series-parallel* is generally used to indicate a series circuit having certain minor branches in parallel; the term *parallel-series* denotes a parallel circuit having certain minor branches in series. The illustration shows a series parallel circuit having two dry cells in parallel and three dry cells in series. The push button and the bell are also in series. (See *Parallel Connection, Series Connection,* also *"A" Battery.*)

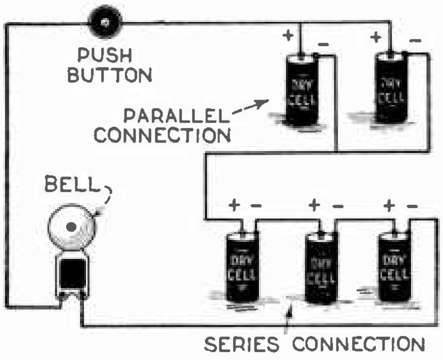

A series-parallel circuit. The three dry cells are connected in series with each other and with the bell and the push button. They are also in series with the other two dry cells, these latter being in parallel with each other.

SERIES RESONANCE—When a single lumped capacity and a single lumped inductance are connected in series between terminals to which an alternating electromotive force is applied, and the inductance or capacity or frequency is varied, the condition of series resonance exists, when the current is a maximum. (See *Resonance,* also *Resonance Curve.*)

SHARPNESS OF RESONANCE—A quantity expressing the fractional change of current in a simple series circuit for a given fractional change in either capacitative or inductive reactance.

SHARPNESS OF TUNING—See *Selectivity.*

SHARP WAVE—A wave in which the energy radiated is confined to a single frequency of oscillation. Tight coupling of the oscillation transformer tends to give a broad wave, whereas loose coupling gives a sharp wave. (See *Decrément.*)

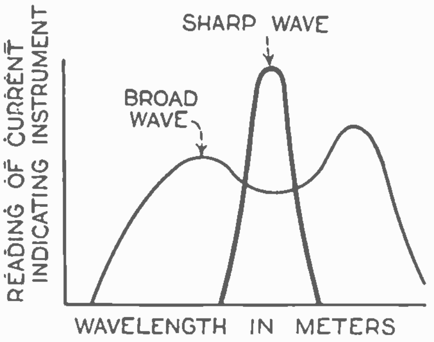

Graph showing difference between sharp and broad waves.

SHIELDED TRANSFORMER—An audio or a radio frequency transformer surrounded by a metal casing. Shielding of radio frequency transformers prevents interference from external fields and also keeps body capacity from affecting the tuning. Audio frequency transformers are shielded mainly to prevent interaction between the flux from the transformer and other parts of the radio set.

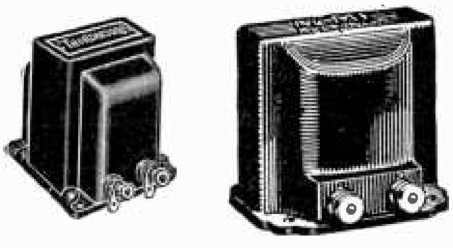

Typical shielded transformers.

SHIELDING—Metallic screening placed around radio apparatus to prevent interference from stray electromagnetic or electrostatic fields. In radio sets, the inside surface of the panel is often shielded by means of a metallic lining, thus eliminating body capacity. It is usual to ground the shielding. In sets having more than two stages of radio frequency amplification complete shielding is essential. (*See Induction Screen.*)

S.H.M.—Abbreviation for *Simple Harmonic Motion* (q.v.).

SHOCK EXCITATION—See *Impact Excitation.*

SHORT—See *Short Circuit.*

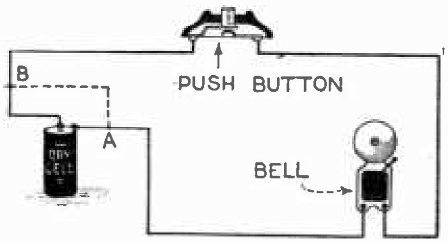

The dotted line between A and B represents a conductor short-circuiting the dry cell.

SHORT CIRCUIT—A short circuit refers to a condition of an electrical circuit whereby an external path of low resistance is connected across the source of electrical energy. A short circuit is accompanied by an increase in current flow, although this current is no longer usefully employed. Apparatus may be protected from short circuits by means of *circuit breakers* (q.v.) or *fuses* (q.v.).

SHORTENING CONDENSER, or ANTENNA TUNING CONDENSER—A condenser connected in series with the antenna and the ground, for the purpose of diminishing the natural wave length.

SHUNT WOUND GENERATOR—A generator in which the field winding is shunted across the armature terminals. The resistance of the shunt winding is made high enough so that only a small percentage of the load current flows through it.

SHUNT WOUND MOTOR—A motor in which the field winding is connected in parallel with the brushes (and the armature winding). The speed of a shunt motor is practically constant, but falls to a slight extent as the load is increased. The drop in speed may be only a few per cent between no load and full load, for large motors, while for small ones it may be as high as 15 per cent. As the load on a shunt motor is increased, the torque which the motor develops will increase and the armature current will also increase in almost direct proportion.

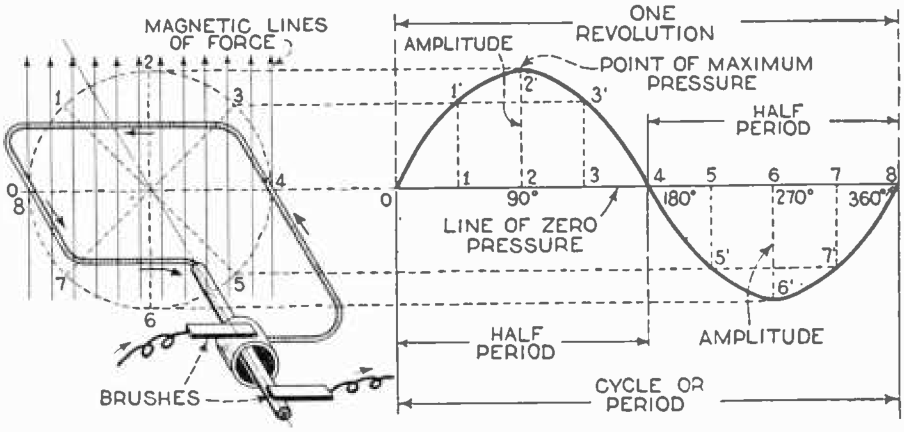

A sine curve is shown at the right. At the left a coil is shown rotating in a magnetic field. As the coil cuts the magnetic lines it generates the varying electromotive force represented by the sine curve.

S.I.C.—Abbreviation for *Specific Inductive Capacity* (q.v.).

SIDE BANDS—A group or band of frequencies formed by the interaction of the *Carrier Wave* (q.v.) and the modulations above and below the main carrier wave frequency. (See *Harmonics.*)

SIGNAL, STRAY RADIO—See *Strays.*

SILICON—A metalloid occurring widely in nature. It is a component of most of the rocks forming the earth's crust. Silicon crystals are used as detectors in radio work. (See *Crystal*, also *Silicon Detector.*)

SILICON DETECTOR—A *crystal detector* (q.v.) using silicon as its crystal. The silicon is used in a crystalline state and a *cat-whisker* (q.v.) is generally used to locate the *sensitive spots* (q.v.) and make contact.

SIMPLE HARMONIC—Referring to a *simple harmonic motion* (q.v.). If a point moves at a uniform rate around a circle and the point be projected in a straight line in the plane of the circle, the back and forth motion of the projected point is referred to as *simple harmonic motion.* True simple harmonic motions are the motions of a pendulum, a vibrating tuning fork, particles of water in a wave, etc. Other motions which are not precisely simple harmonic motions may be treated as resulting from several such motions. Examples of these are motions occurring in alternating currents in light waves, in electric waves and in sound waves.

SIMPLE HARMONIC MOTION—Abbreviation S.H.M.—The motion executed by the foot of a perpendicular let fall on the diameter of a circle from a particle moving with uniform velocity in that circle. The piston rod of a steam engine, turning a crank uniformly approximates a simple harmonic motion.

SIMPLE HARMONIC VIBRATIONS—The vibrations of a body such as a tuning fork which follow the law of *simple harmonic motion* (q.v.). (See *Angular Velocity*, also *Simple Harmonic.*)

SINE CHARACTERISTIC OF ALTERNATING CURRENT—See *Sine Curve*, also *Sine Wave.*

SINE CURVE—A curve whose perpendicular at any point is proportional to the sine of the angle corresponding to that point. The sine curve is used to represent the changes in direction and strength of an alternating current or voltage. The coil illustrated at the left in the accompanying diagram generates a sine wave current as it revolves through the magnetic field shown by the arrows. When the coil is moving parallel to the magnetic flux, no electromotive force is induced and hence current flow is zero. As the coil is rotated in a clockwise direction, it cuts the flux at a greater and greater angle, until the maximum is reached when the flux is cut at a right angle. This gives a gradually increasing electromotive force (and hence current) until maximum is reached, then when the angle at which flux is cut decreases, the electromotive force decreases until the coil again moves parallel to the flux, at which point electromotive force is again zero. Then when the coil starts to cut the flux in the opposite direction, electromotive force is reversed, increasing to a maximum and then decreasing to zero thus completing the *cycle* (q.v.). Current or electromotive force in one direction (positive) is represented by the height of the ordinates of the sine curve above the line, while current in the reverse direction is represented by the height of the ordinates below the line. The horizontal distances (abscissae) represent time generally measured in phase angles (position of coil at any particular instant). (See *Alternating Current*, also *Curve of Sines.*)

SINE WAVE—As referring to an alternating current, a sine wave or a simple alternating current, is a current whose wave shape is sinusoidal. The wave shape is the shape of the curve obtained when the instantaneous values of the current are plotted against time in rectangular co-ordinates. The wave shape is independent of the frequency and of the scale to which the curve is plotted.

SINE WAVE OF ALTERNATING CURRENT OR E.M.F.—See *Sine Wave.*

SINGING ARC—See *Duddell Singing Arc.*

SINGING SPARK—A spark used in a wireless telegraph transmitter which produces a singing note in the receiving head set. A spark which occurs at regular intervals and at an audible frequency such as that obtained in a quenched spark system of wireless telegraphy.

SINGLE CIRCUIT TUNER—A regenerative receiving circuit in which the antenna and grid circuits are conductively coupled. *Feedback* (q.v.) is obtained by placing the plate circuit in inductive relation with the primary.

SINGLE-PHASE—This term is used to refer to an alternating current, an alternating current system of distribution, or to alternating current apparatus. A single phase system or circuit is one energized by a single alternating electromotive force. A single phase circuit is usually supplied through two wires. The currents in these two wires counted outwardly from the source, differ in phase by one-half a cycle (180 degrees). In the single phase generator the voltage per phase is the same as the voltage between the lines and the current per phase is the same as the current per line. Since there is a great saving of copper in the use of three phase transmission of power, three phase generators are more frequently used than single phase machines. (See *Alternator.*)

SINOIDAL VIBRATION—See *Simple Harmonic Motion.*

SIPHON RECORDER—A moving coil galvanometer originally used for recording messages sent over long submarine cables. The received currents are passed to a coil suspended in a strong magnetic field. This coil carries a fine glass syphon which discharges ink onto a moving paper tape. In cable signaling, dots and dashes are usually indicated by opposite deflections which result in a wavy line being produced on the tape.

SKINDERVIKEN TRANSMITTER BUTTON—A small carbon grain microphone (q.v.) button less than an inch in diameter and about half an inch high containing a polished metallic button affixed to a mica diaphragm and a surrounding case of brass, the space between the case and the polished button being partially filled with a good grade of carbon granules. The button is very sensitive to sound waves when it is attached to any form of a diaphragm, thus making it applicable wherever a sensitive telephone transmitter should be used. Fig. 1 shows the external appearance, full size, of the Skinderviken button. This transmitter button has a number of applications in radio work and can be used in connection with speech amplifiers, telephone transmitters, detectaphones, etc. It can be made to transmit in any position, either horizontal or vertical. It is small enough to be readily concealed. Fig. 2 shows how radio music can be re-transmitted to distant points and incidentally amplified at the same time. The microphone button is shown fastened to a cone and mounted in front of the loud speaker. The leads from the button are connected in series with several dry cells and another loud speaker at a distant point. Fig. 3 shows how the transmitter button can be attached to an ordinary receiver to produce the same effect. In this case also, the leads are connected in series with several dry cells and with a low resistance phone at the distant end. In Fig. 4 the transmitter button is shown in use as a grid leak in a regenerative circuit. Where it is desired to get greater sensitivity and volume for either radio telephone transmitting or for ordinary phone transmission the system illustrated in Fig. 5 will be found highly suitable.

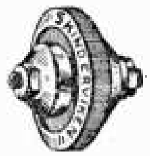

Fig. 1. Skinderviken Transmitter Button.

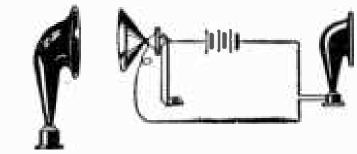

Fig. 2. Retransmitting radio programs.

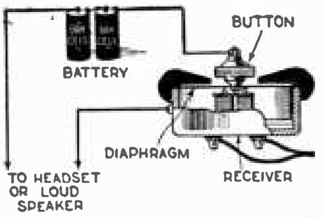

Fig. 3. Retransmission from ear phone.

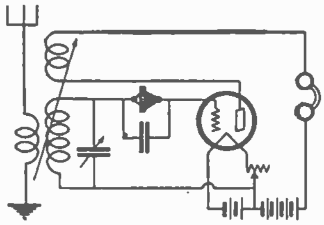

Fig. 4. Button used as grid leak.

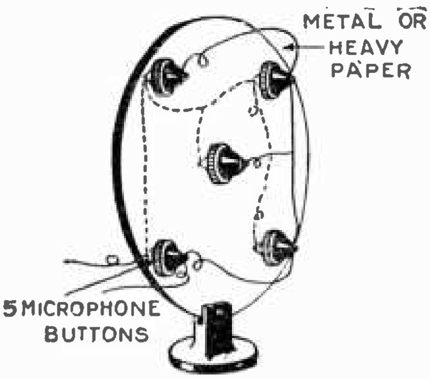

Fig. 5. Combining buttons to form sensitive microphone.

Fig. 6 shows the construction of a hand microphone for transmitting sets or for inter-departmental telephone circuits. A very simple circuit closer is mounted in the handle of the microphone which in this particular case is simply an ordinary piece of wood hollowed out to receive the switch and the wires. This wooden handle is then fastened to the rubber case of a telephone receiver. The coils and magnets in the telephone receiver are removed, a small hole is drilled or punched in the diaphragm, and the transmitter button is fastened as indicated in the diagram.

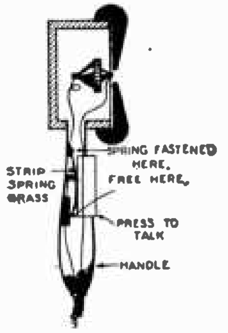

Fig. 6. Construction of hand microphone.

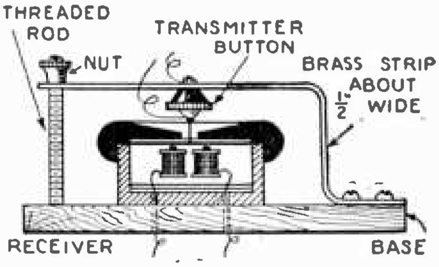

Fig. 7. Method of fastening a button to ear 'phone.

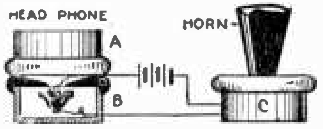

Fig. 8. A simple form of amplifier.

Another method of fastening a transmitter button to a telephone receiver is illustrated in the diagram in Fig. 7. Here the transmitter button is mounted on a brass strip which is bent so that the center of the button will rest upon the diaphragm of the telephone receiver fixed to the base. A threaded rod and nut regulate the pressure of the button against the diaphragm. In the event that it is not desired to make any changes to a pair of telephone receivers, the transmitter button can be fastened in the shell of an old receiver, B in Fig. 8, and the head phone may be rested on the shell. For use with a Baldwin receiver, a microphone transmitter button is fastened to it in the following manner: The diaphragm of the Baldwin receiver is first removed, and a thin rod is soldered to the lever, as indicated in Fig. 9. The transmitter button is then soldered to the rod as shown. Another way of making a loud speaker set from an ordinary one-tube receiving set is illustrated in Fig. 10. Where a phonograph is available, the hook-up illustrated may be employed, placing the reproducer of the phonograph on the receiver, which should be placed in a suitable position on the turntable. Inasmuch as the movement of the diaphragm of the seventy-five ohm receiver is an up and down one, the reproducer of the phonograph should be turned as it would be for "hill and dale" records. A tiny drop of solder should be dropped upon the diaphragm of the seventy-five ohm receiver and a small nick placed in the center of the solder for receiving the needle of the phonograph reproducer.

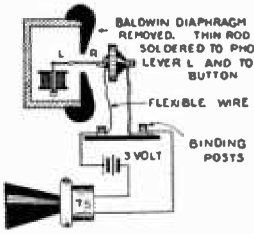
BALDWIN DIAPHRAGM REMOVED. THIN ROD R SOLDERED TO PHONE LEVER L AND TO BUTTON
L
R
FLEXIBLE WIRE
3 VOLT
BINDING POSTS
75

Fig. 9. Button attached to Baldwin 'phone.

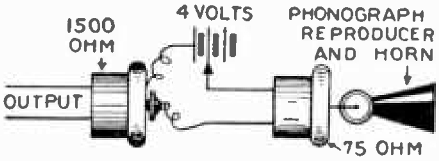
1500 OHM
4 VOLTS
PHONOGRAPH REPRODUCER AND HORN
OUTPUT
75 OHM

Fig. 10. Loud speaker operation from 1-tube set.

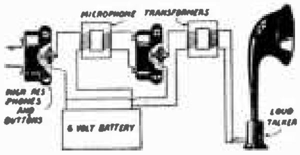
MICROPHONE TRANSFORMERS
HIGH RES PHONES AND BUTTONS
6 VOLT BATTERY
LOUD TALKER

Fig. 11. An effective radio amplifier.

Very often it is desired to further amplify the music or sounds picked up by a transmitter button. For this purpose, microphone transformers, or as they are better known, modulation or telephone induction transformers, are employed. In Fig. 11 this instrument is illustrated, together with the circuit for its use. Here the incoming radio signals are made to operate the microphone fastened to the receiver. The current which the microphone passes then goes through the primary of the microphone transformer, the secondary of which is connected to another receiver similarly fitted with another transmitter button. The output of the second microphone transformer is then connected to a loud speaker. A diagram for using but one stage of amplification through transformers is illustrated in Fig. 12. For the better class of work, however, it is desirable to use a standard vacuum tube amplifier as illustrated in Fig. 13. Here the microphone button is placed in series with the primary of a modulation transformer, and in series with several dry cells or flash light batteries. The secondary of the modulation transformer then connects to the grid and filament, the same as it would in a standard vacuum tube amplifying circuit. The music from the output is, of course, very loud.

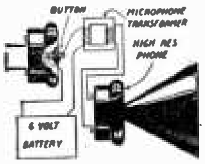
BUTTON
MICROPHONE TRANSFORMER
HIGH RES PHONE
6 VOLT BATTERY

Fig. 12. A single stage amplifier.

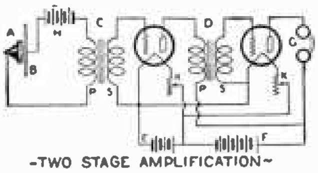
A
B
H
C
D
G
P
S
P
S
K
E
F
-TWO STAGE AMPLIFICATION~

Fig. 13. A 2-stage vacuum tube amplifier.

A very splendid oscillator is illustrated in Fig. 14. Here a tuning fork of standard frequency is mounted on a suitable holder in a cigar box. Two low resistance coils, about five to seventy-five ohms, are then placed under one prong of the tuning fork. These are connected in series with the transmitter button fastened in the cigar box, which serves as a base for mounting the tuning fork. The batteries may be concealed in the cigar box. Thus when the tuning fork is struck, it will continue to vibrate as long as the current is applied to the microphone. To use this as an oscillator, a pair of magnets from a telephone receiver and the permanent magnets are mounted near the other vibrating prong. The current constantly being changed by means of the vibrating bar of steel produces a standard frequency, which when amplified can be used for making radio measurements of different kinds. Fig. 15 shows the transmitter button connected in a radio circuit for use as a detector of radio waves. Fig. 16 shows a method of making a code practice device. An audio frequency howl is produced which will enable an instructor to teach an entire classroom the radio code. Sometimes the key, instead of interrupting the circuit, merely shunts a small portion of the current passing through the transmitter button, and for this purpose the key would have to be placed directly across the transmitter button and in series with a small resistance. Fig. 17 shows how the transmitter button may be used in Heising modulation. Fig. 18 indicates its adaptation to the variation of grid voltage of the oscillator tube and Fig. 19 shows how the transmitter button may be hooked up to modulate the output current. These last three figures indicate the use of transmitter buttons in vacuum tube transmitters.

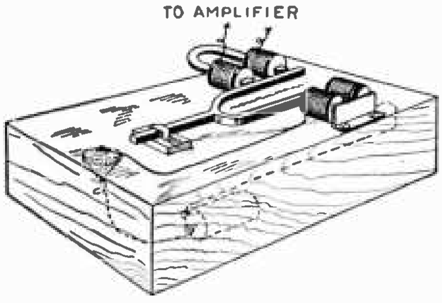
TO AMPLIFIER

Fig. 14. A tuning fork oscillator.

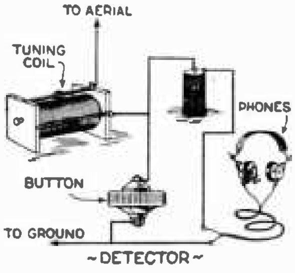
TO AERIAL
TUNING COIL
PHONES
BUTTON
TO GROUND
~DETECTOR~

Fig. 15. Transmitter button used as radio detector.

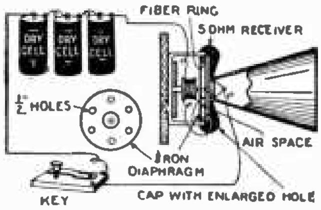
FIBER RING
5 OHM RECEIVER
DRY CELL
DRY CELL
DRY CELL
½" HOLES
AIR SPACE
IRON DIAPHRAGM
KEY
CAP WITH ENLARGED HOLE

Fig. 16. A code practice machine.

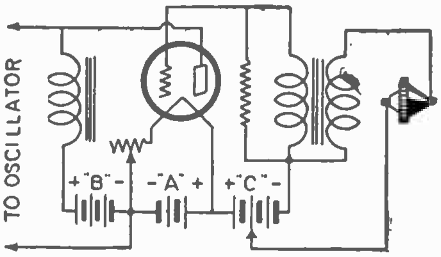
TO OSCILLATOR
+"B"-
-"A"+
+"C"-

Fig. 17. Use of button in Heising modulation system.

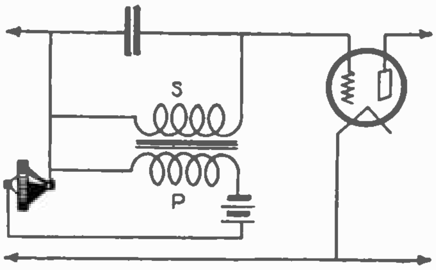

Fig. 18. Method of varying grid voltage of oscillator tube.

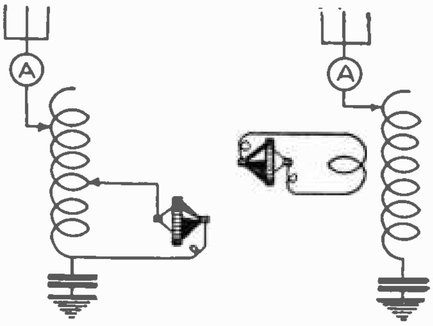

Fig. 19. Method of modulating output current.

SKIN EFFECT — The name given to the crowding of alternating or oscillatory currents into the surface layers of a solid conductor. This phenomenon increases with increased frequency. As a result the resistance of solid conductors is much higher at high than at low frequencies. (See *High Frequency*, also *High Frequency Resistance.*)

SLABY-ARCO-SYSTEM — A system of wireless telegraphy which used a quenched spark gap. This system is now obsolete.

SLABY-ARCO-VACUUM COHERER — A filings coherer enclosed in a sealed glass tube from which the air has been extracted.

A Slaby-Arco vacuum coherer.

SLIP RINGS — Metallic rings, separated by insulators, from which collecting brushes pick up current generated by an alternator. Slip rings take the place of the commutator used on the direct current generator.

SLOTS—Channels cut in the armature core discs and into which the windings fit.

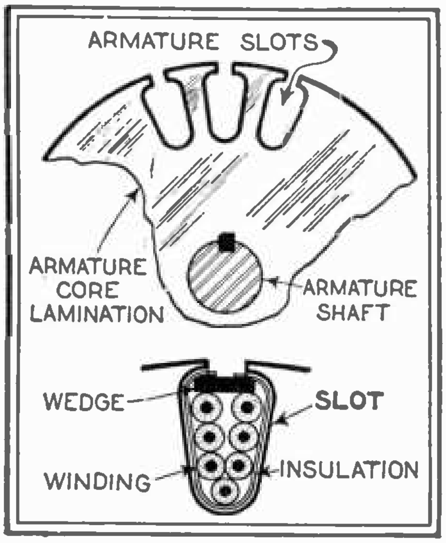

Details of armature slot.

SLUDGING OF TRANSFORMER OIL —The thickening of the oil used for cooling in large transformers. A muddy deposit is formed due to oxidation. The presence of metallic copper in contact with the oil is considered to hasten this action catalytically.

SOAKING-IN—An increase in a condenser charge above its initial value, which takes place gradually if the potential is kept applied. Soaking-in takes place due to a change in the *dielectric* (q.v.).

SOAKING-OUT — A gradual continuation of the discharge of a condenser after the first rapid discharge. In this case, the dielectric is returning to its original state.

SOCKET—A receptacle for holding and making electrical connections with a vacuum tube, an incandescent lamp, or a connection plug. Vacuum tube sockets are made of bakelite, hard rubber, porcelain, pyrex and other suitable *dielectrics* (q.v.). They are usually molded. One type of socket is composed of a shell of insulating material, or in some cases of metal. Four contact blades extend within the base of the socket, for making contact with the prongs of the tube. These blades may be of strong phosphor bronze. Their outer ends are attached to binding posts. It is important that the base supporting the binding posts and contacts be of good insulating material such as bakelite or hard rubber. Modern sockets are of the *spring* or *shock absorbing* type. This construction protects the tube and does away with microphonic sounds. One type of spring socket consists of two molded bakelite parts, the outer one square with a binding post at each corner and the inner one circular with holes for the tube prongs and steel spring contacts fastened underneath the holes.

A metal shell socket.

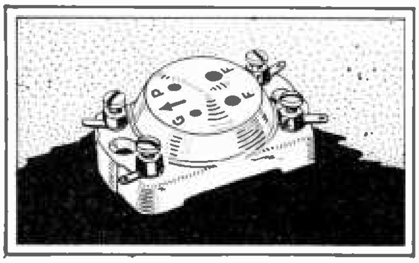

Shock absorbing or spring-type socket.

The circular molding is suspended within the square one by means of two rubber pads. Other types of shock absorbing sockets use spring suspension instead of rubber.

SOFT IRON INSTRUMENTS — Ammeters, voltmeters and other measuring instruments which depend upon the magnetic force between a current carrying coil and a movable core of soft iron to produce the deflection of the needle or pointer. The movement of the iron may be controlled by a spring or by gravity. These instruments may be used for either direct or alternating current measurements. They are not as accurate as instruments of the dynamometer type.

SOFT TUBE — A vacuum tube which contains a gas content instead of being highly evacuated. Soft tubes were formerly recommended for use as detector tubes but at present have been superseded for this use by the ordinary amplifying tube.

SOLDER—An alloy of lead and tin used for making good electrical connections.

SOLDERED JOINT — A connection between two metallic surfaces, cemented together by means of solder. Where the joint is between two electrical conductors, it must be made both mechanically and electrically secure.

SOLDERING — The process of joining together two or more conductors or other metallic surfaces by means of solder, so as to give a good electrical or mechanical connection, or both. After the surfaces to be united are carefully cleaned, a flux is applied to dissolve the oxides which occur on the surface of the parts to be joined with solder.

SOLDERING FLUX—A substance used to dissolve the oxides on the surfaces of the metals to be soldered. When these oxides are dissolved, the flux enables the solder to enter the minute pores of the metal surface, effectually sealing it against the penetration of oxygen. Fluxes range in character from very strong acids to very mild acid bearing substances. For radio use, it is necessary to use a flux which is non-corrosive and which in its use will leave a residual matter that will have no tendency to collect moisture, dust, or other foreign material. Rosin has been found to be very suitable as a flux in radio soldering. (See *Flux*, *Soldering.*)

SOLENOID — A coil of wire of helical form used as an electromagnet, generally to attract a movable iron plunger, which is drawn within the solenoid when the current flows through the windings. The solenoid never has a fixed core. A coil wound on a cylinder so as to give uniform magnetic field within it when traversed by a current is often called a *straight solenoid.*

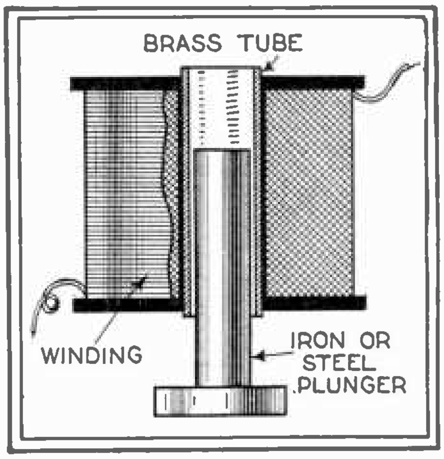

A typical solenoid.

SOLODYNE—A radio circuit which dispenses with high-tension or "B" batteries and which utilizes a double grid vacuum tube. The solodyne principle is also known as the *Unidyne*. The theory of the solodyne is as follows: The small plate current due to the electron emission from the lighted filament passes through the tickler coil which feeds back to the main grid circuit in the usual way, the primary of a step-up transformer, through the telephone receivers, and then back again to the filament of the tube. The electron stream passing from the filament to the plate inside the tube must pass the two grids. The first of these, which may be referred to as the additional grid, is primarily made positive by connecting it directly to the positive terminal of the "A" battery. This in itself tends to assist the electron stream, to reduce the resistance of the vacuum of the tube. But the additional grid is assisted in its work by having impressed upon it the stepped up voltage from the plate circuit, due to the transformer, the secondary of which is in direct connection with the additional grid. Thus a building up process is introduced, every possible electron primarily due to the electron emission of the filament of the tube being made use of and ultimately passing through the telephone receivers to be reproduced in the form of audible signals. The main grid functions in the usual manner, except that this too, can be made to help the additional grid as well, by giving it a strong positive bias. A typical Solodyne circuit is shown in Figure 1. In this circuit L and L^1 may be honeycomb coils or a variocoupler. In Figure 2 is shown the Cowper Solodyne circuit which utilizes standard tubes. In this diagram, the numbers printed near each coil indicate the size of honeycomb or duo-lateral coils to use in each circuit. (See *Four Element Tube.*)

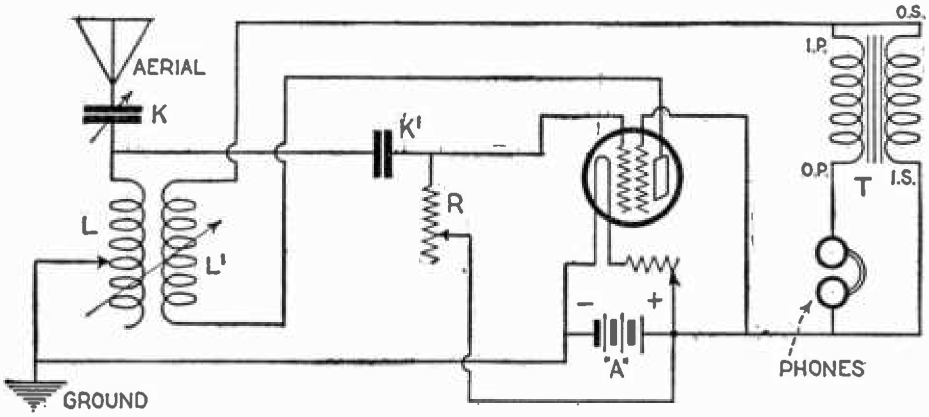

A Solodyne circuit. L and L¹ may be honeycomb coils or a vario-coupler.

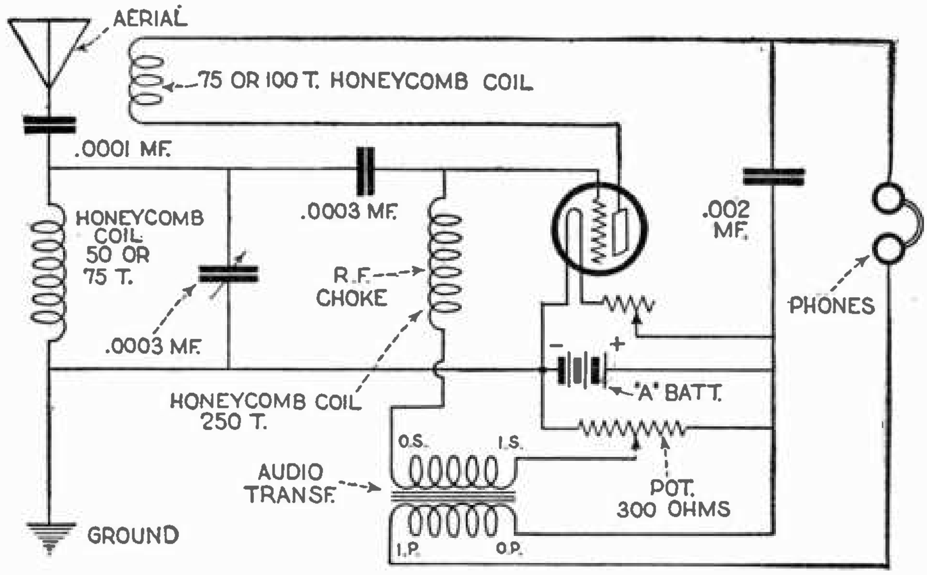

The Cowper circuit uses no "B" battery and is quite stable in operation.

SOUNDER—An instrument used for receiving telegraph signals which utilizes the attraction of an armature by an electromagnet to make sounds as the armature hits against stops at the beginning and end of each current impulse. The sounds thus produced form the characteristic dots and dashes of the telegraph code and can be easily read by the experienced telegraph operator.

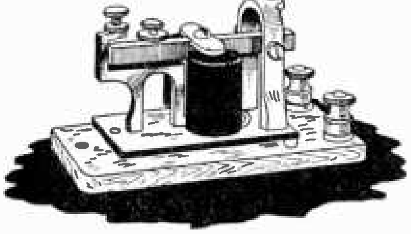

Telegraph sounder.

SOUND WAVE—A wave of alternate condensation and rarefication through an elastic body such as air, water, etc. Sound waves travel at the rate of 1090 feet per second in air.

SOUTH MAGNETIC POLE—Situated in Latitude 70 South and Longitude 102 East. The South Magnetic Pole does not coincide with the geographic South Pole.

SOUTH POLE—When referring to a magnet, the pole which tends to point to the south when the magnet is freely suspended.

SOUTH SEEKING POLE—See *South Pole.*

SPACE CHARGE—The difference between the number of electrons and positive ions in unit volume, multiplied by the charge per ion. This is also called the volume density of electrification. Let us consider the distribution of electrons in a vacuum tube, between the hot filament and the plate. An electron close to the surface of the plate is attracted to the plate by two forces, attraction from the plate and repulsion from the electrons located between it and the filament. An electron close to the surface of the filament, however, is repelled by the electrons between the plate and itself, although it receives some attraction from the plate. It may either move towards the plate or go back to the filament. This will depend upon whether the plate voltage is high enough to result in a force of attraction sufficiently great to overcome the repulsive force of the space charge.

SPACE CURRENT—Current which flows between the cathode and the anode in a vacuum tube. Space current is the result of the motion of electrons through space. It does not follow the same laws which govern the flow of current in a metallic circuit, deviating from Ohm's law to a certain extent, depending upon the relation between the impedances of the tube and the metallic portion of the circuit and also upon the nature of the discharge.

SPAGHETTI—A varnish-impregnated cloth tubing used to insulate bare conductors. (See *Cambric, Varnished.*)

SPARK—See *Spark Discharge.*

SPARK COIL—An *induction coil* (q.v.) used to produce *spark discharges* (q.v.).

SPARK DISCHARGE—A passage of electrical current between two conductors across a previously non-conducting space. The discharge is accompanied by light, heat and sound. It may take place through a liquid, a solid or a gas but whenever it occurs, it means that there has been a breaking down of the dielectric stress. A spark passing through a solid will puncture it. The passage of a spark through a liquid does not affect the dielectric properties of the liquid. On the other hand, a spark passing through a gas, usually ionizes the gas for a period of time after the passage of the discharge and as a result a continuous arc may follow the first spark. An *oscillating current* (q.v.) may produce a rapid train of sparks which persists until the oscillations are damped out. (See *Sparking.*)

SPARK DISCHARGER—A form of *spark gap* (q.v.) such as a *rotary discharger* (q.v.) or a fixed gap, such as a *quenched spark gap* (q.v.). (See *Discharger*, also *Disc Discharger.*)

SPARK GAP—A break in an oscillating circuit which acts as an automatic safety valve to the condenser. Since the air between the gap has a high resistance, the condenser cannot discharge until the potential is sufficient to break down its insulation, thus permitting only heavy discharges to take place. (See *Quenched Spark Gap, Fixed Discharger*, also *Safety Gap.*)

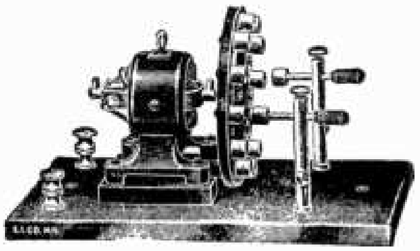

A rotary spark gap.

SPARKING—The production of spark discharges, especially when an inductive circuit is broken. Sparking at the brushes of a motor or generator is caused by armature reaction and self-induction. Sparkless commutation can be obtained by shifting brushes or by the use of *interpoles* (q.v.). Sparking is injurious to contact surfaces and should be avoided if possible. Sparking between contacts of induction coil vibrators is usually reduced by placing a condenser across the gap.

SPARK MICROMETER—A graduated adjustable spark gap for determining sparking distances for various voltages, etc. This device also permits the approximate measurement of high voltages provided the sparking distances for the particular type of terminals, are known.

SPARK RATE, SPARK FREQUENCY, or GROUP FREQUENCY—The number of sparks per second occurring in a spark system of wireless telegraph transmission. It refers to the *Group Frequency* of the wave train rather than to the waves themselves.

SPARK RECORDER—An instrument for recording telegraph signals, in which sparks from an induction coil pass through and mark a paper tape carried on a drum which turns under a metallic pointer. The use of a spark recorder dispenses with the use of ink, thus doing away with the friction of the pen on the paper.

SPARK TRANSMISSION—A wireless telegraph transmitting system which uses a succession of spark discharges in an oscillating circuit, to produce oscillations. These traverse an aerial system and a series of short trains of damped waves are emitted.

SPECIFIC GRAVITY—abbreviation S.G.—The weight of a body compared with that of another, having equal bulk, considered as a standard. The standard for liquids and solids is water, while hydrogen or air is the standard for gases. The specific gravity test of the radio storage battery is an important means of testing the condition of the battery. A *hydrometer* (q.v.) is used for this purpose.

SPECIFIC INDUCTIVE CAPACITY—symbol k—abbreviation S. I. C.—A measure of the degree to which a body permits electrostatic induction through it. The ratio between the capacities of two condensers, one employing the material under consideration as the dielectric and the other using an air or vacuum dielectric. The Specific Inductive Capacity of a material is the *inductivity* (q.v.) of that material relative to that of air. (See *Inductive Capacity* also *Dielectric Coefficient or Constant*.)

SPECIFIC RESISTANCE—The resistance in ohms, of unit length and unit cross-section of a conductor. (See *Resistivity of a Material*, also *Resistivity, Surface*.)

SPECTRUM—An image formed by rays of light or other radiant energy in which the parts are arranged according to their refrangibility or wave length. The image may be visible or invisible. The arrangement is such that all parts of the same wave length fall together, while those of different wave lengths are separated from each other forming a regular series. The spectrum produced by the light of the sun passing through a triangular glass prism and falling on a screen is one of the most common forms. The various colors, since they are unequally refracted, are spread out into a band, showing the seven rainbow colors. The red is at one end (that of the least deviation), while the violet is at the other end.

SPEECH AMPLIFIER—An audio frequency amplifier designed especially for public address systems, where addresses are to be heard over a comparatively wide area. By means of speech amplifiers, it is possible for one speaker to address an unlimited number of listeners.

SPEECH MODULATION—The modulation of radio frequency currents, as utilized in radio telephony. The production of speech modulated waves calls for a source of undamped waves and a method of causing variations in the current output of this source which will accurately correspond to the vibrations of the voice. The source of undamped waves may be a high-frequency alternator or an electron tube generator, the latter being in most common use. The radio frequency antenna current may be varied by inserting a speech-controlled variable resistance, such as a microphone, in the antenna circuit at the transmitter. The microphone may be put in the direct current power supply of the generating system, so that the radio frequency output of the system will be varied as the power input is varied. (See *Modulation*, also *Modulation Frequency Ratio*.)

SPEECH VIBRATIONS—The wave corresponding to a given sound. There are various methods of obtaining graphically, the picture of the wave form of any particular sound. One method is to make a phonograph record of the sound and then as the sound is reproduced from the record, to magnify the movement of the needle, using a lever to trace the form of the waves.

SPIDER WEB COIL—A form of inductance coil in which the wires are wound on a frame consisting of radiating arms similar to the spokes of a wheel. The wire is wound in successive turns in and out around the arms, starting from the center, until it reaches the ends of the arms. An odd number of turns are used so that alternate turns will follow the same wave; that is to say, adjacent turns will be on opposite sides and separated by an arm. Coils of this type have comparatively low distributed capacity and in addition have the advantage of being extremely compact. Figure 1 illustrates the frame on which the wire is wound. This is known as the *spider*. A spider cut from $\frac{1}{16}$th inch bakelite, as shown in Figure 2, can also be used, although, owing to a smaller number of divisions, the latter does not have such a large inductance as the former. In winding, the wire is fastened around one of the pegs and is then taken in and out around the pegs, as shown in Figure 3. When starting the second turn, the wire will be around the pegs the opposite way to the first turn, and all successive odd-number turns will be on the same side as No. 1, while all successive even number turns will be on the same side as No. 2. The finished coil is shown in Figure 4. Spider web coils are also known as *basket wound* (q.v.), *basket woven*, or *stagger wound* coils.

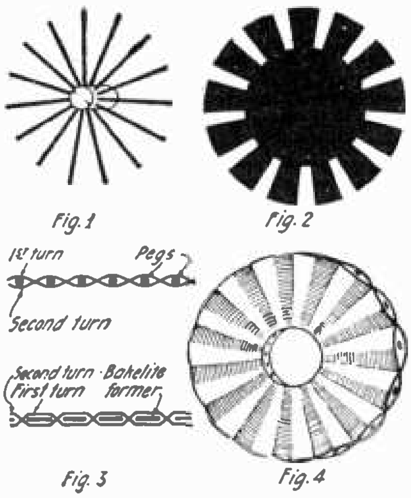

Details of spider web coil construction.

SPLICE—A method of joining two or more conductors by interweaving or entwining the strands, in a similar manner to that of splicing a rope.

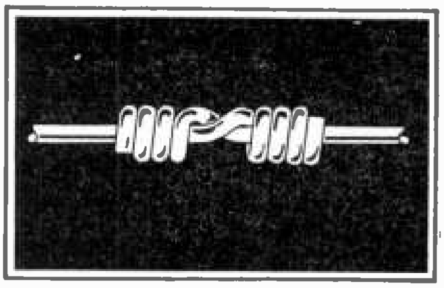

A strong splice.

SPREADER—A spar or pole used on an aerial, where two or more wires are used, to keep these wires properly spread out and parallel to each other.

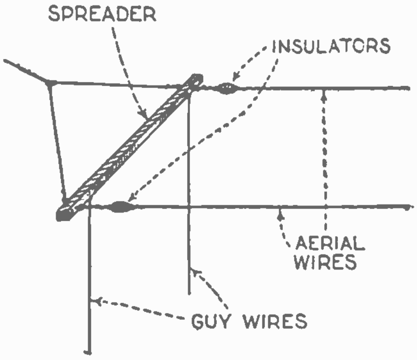

Aerial showing use of spreader.

SQUARE MIL—A square mil is the area of a square, one mil, or 0.001 inch on each side. (See *Circular, Mil*.)

SQUIER, MAJOR-GENERAL SIR GEORGE OWEN—American radio authority. He was educated at the Johns Hopkins University and became a research student under Professor Rowland and Sir William Preece at the British General Post Office. In 1904 he published his famous paper on the absorption of electromagnetic waves by living vegetable organisms and showed how trees could be used for the reception of radio messages. In 1911 he read a paper on multiple telephony before the American Institute of Electrical Engineers. In 1912 he was awarded the Elliot Cresson Gold Medal for his researches in multiplex telephony and in 1919 the Franklin Medal of the Franklin Institute. Major-General Squier was awarded

Major-General George Owen Squier.

the K. C. M. G. for distinguished services during the World War. He is a member of the National Academy of Sciences and the International Electrotechnical Commission. He is the inventor of *Line Radio* (q.v.) which is also known as *Wired Wireless, Wire-Radio Telephony, Guided Wave Telephony,* and *Carrier Current Telephony.*

SQUIRREL CAGE INDUCTION MOTOR—A type of induction motor having a rotor consisting of copper bars connected to rings at each end so as to form a short-circuited system. There are no windings in this type of rotor nor are there any external connections through slip rings. These motors are used for constant speed work, where starting is necessary only at infrequent intervals. The squirrel cage motor draws a large starting current, but has a relatively small starting torque. It is possible, however, by properly designing the rotor so as to have enough resistance, to use small motors of this type for loads requiring frequent starting, rapid acceleration and high starting torque.

SQUIRREL CAGE ROTOR—See *Squirrel Cage Induction Motor.*

STAR GROUPING—A method of connecting up polyphase apparatus or circuits. One end of each phase is connected to a common point, usually called the neutral point. This method of connection is called a "Y" grouping, in the case of a three-phase system.

STATIC—An irregular disturbing noise, heard in the radio loud speaker or head set due to atmospheric discharges, lightning and similar phenomena. A common form of static produces an intermittent crashing sound. Other types cause grinding noises or hissing sounds. Nearby lightning is usually accompanied by a sharp snap in the loud speaker. Various methods have been devised for eliminating or reducing static, although it cannot be said that any of these has been entirely successful. The combination of coil and ground aerials results in the reduction of static; also the use of *beat reception*, using continuous waves as in certain methods of wireless telegraphy. Static is also referred to under various other names, such as *strays, atmospherics, X's*, etc.

STATIC CHARACTERISTIC—The curves obtained by the use of a steady direct-current potential for showing the performance, efficiency, etc., of a vacuum tube. The static characteristics of the tube itself are obtained when the resistance of the external circuit is very small in comparison to the plate resistance of the tube. In cases where the external circuit contains reactance, the true characteristic could not be obtained using direct current. In this case, an alternating electromotive force is impressed on the grid and the curves obtained are known as *dynamic characteristics* (q.v.).

STATIC ELECTRICITY—Electricity which is stored in a circuit, manifesting itself in the form of charges at high potential. Electricity, as produced by frictional or influence machines.

STATIC ELIMINATORS—A device for eliminating or reducing the effect of *static* (q.v.). Various circuits have been proposed for accomplishing static elimination. Some of the most recent and successful of these are the McCaa anti-static circuits described in the July, 1925, issue of RADIO REVIEW, on page 63. There are two types of McCaa circuits, one applicable to radio telegraphy and the other to radio telephony. Both have for their object the reduction of signal static ratio. The wiring diagram shown is that of a standard five-tube neutrodyne receiver plus a McCaa anti-static device of the receiver type. In this diagram Ta is the repeater tube, Tb and Tc are the two radio-frequency amplifier tubes, Td is the detector tube and Te and Tf are the two audio-frequency amplifier tubes. The circuit from tube Tb on is unchanged and is the same as any standard neutrodyne receiver. The input circuit of tube Tb, however, and also the antenna circuit, have been altered to make possible the installation of the anti-static device.

For the neutroformer which usually couples the first tube of a neutrodyne to the antenna circuit, the coils Lo, Ls and Lp have been substituted. These coils may either be homemade coils or honeycomb coils. If honeycomb coils are selected each may have fifty turns. The coupling between Ls and Lp should be fixed so that the coils are about an inch apart and some arrangement should be made for changing the coupling between Lo and Ls. If homemade coils are used each coil should have approximately sixty turns of No. 22 d. c. c. wire on a three-inch diameter tube and they should be coupled the same as the honeycomb coils.

The coils Lr and Lc are also coupled and may be either 50 and 10 turn honeycomb coils, respectively, or 60 and 10 turn homemade coils wound on a three-inch diameter tube with No. 22 d. c. c. wire. The coupling between these coils is not variable. After the proper value of coupling has been found by experiment it may be made permanent.

Coils Ld and Lq are also shown coupled in the diagram. They may be 50 and 75 turn honeycomb coils, respectively, or 60 and 90 turn homemade coils wound on a three-inch diameter tube with No. 22 d. c. c. wire. The coupling between these coils is also fixed after the proper value has been determined by experiment.

The proper values for the various capacities used in this circuit are as follows: Co., .001 mfd. (43 plate); Cs, .000035 mfd. (17 plate); Cr., .00035 mfd. (17 plate); Cd, .0035 mfd. (17 plate).

In building this set it is advisable to place both the anti-static device and the receiver in the same cabinet. This cabinet must be shielded and the shield should be connected to the ground. It is also advisable to shield the anti-static device from the receiver proper. The batteries should also be shielded either by placing them inside the cabinet or in metal boxes which are connected to the ground. The wires from the batteries to the set may be effectively shielded by using BX cable, the outside covering of which should also be connected to the ground. For the phone wires special phone cords which are shielded with a flexible copper braid are desirable.

STATOR—The stationary plates of a rotary condenser. The fixed portion of a motor or a generator, which carries a winding. In induction motors, that portion of the motor which is fixed in position and which carries the winding connected to the external circuit. (See *Rotor*, also *Alternator.*)

STATOR PLATES—The plates of a condenser which remain stationary in position as differentiated from the plates which rotate. (See *Shielding*, also *Stator.*)

STOPPING CONDENSER—A condenser which opposes or prevents the flow of direct current, but which allows alternating current to flow in a circuit.

STORAGE BATTERY—See *Secondary Cell.*

STORAGE BATTERY TESTS—The testing instrument most commonly used in storage battery work is the *hydrometer* (q.v.). The purpose of the hydrometer is to determine the state of the charge of the battery, as indicated by the *relative* specific gravity of the electrolyte. The specific gravity of a fully charged battery is not a fixed

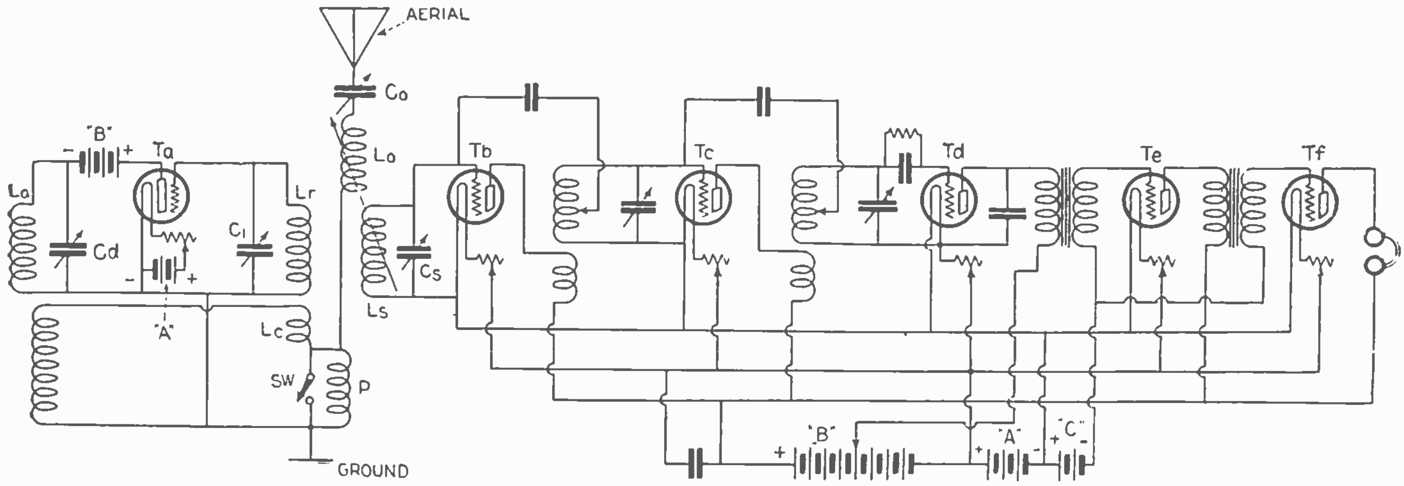

Arrangements of a static reducing device employing repeater circuits.

value but depends entirely upon the standard specific gravity employed by its manufacturer which may be anywhere from 1.250 to 1.300. The user should always ascertain the correct specific gravity for the make of battery before attempting to test it with a hydrometer. A voltmeter may be employed as a means of testing a storage battery provided it is of the proper type and is used correctly. It must be accurate with sufficient length of scale to read tenths of a volt clearly. Voltmeters are obtainable which may be used for both the "A" and "B" batteries whether dry or storage. It is not possible to test the storage "A" battery with cheap types of so-called pocket voltmeters. It must be kept in mind that a voltmeter reading on a storage "A" battery is of value only when the battery is either charging or discharging, except when the battery is practically dead. A voltmeter reading taken when the battery is not connected to the circuit, either charging or discharging, may be misleading. If a reading is taken with the filaments lighted and the voltmeter indicates a full six volts, the battery is charged. Under the same conditions, if the voltmeter lags appreciably below six volts the battery should be charged in order to keep it in the best condition. When it falls to 5.4 volts the battery is discharged and cannot possibly operate the set with any degree of satisfaction. When the battery is connected to the charger and the voltmeter shows from 7.5 to 7.8 volts the battery is fully charged. It is possible to test roughly without instruments. If neither a hydrometer nor a voltmeter is handy it is possible to recognize the necessity for charging by the fact that the amplifying rheostats have to be moved forward of the normal working position, indicating low voltage and that the battery must be charged. When the battery is on the charging line, being charged at the ordinary rate of the commercial charger, and all of the cells of the battery are bubbling, this is an indication that the battery is charged. An interesting meter has recently been placed on the market which should prove of considerable use to owners of storage batteries. It is an ammeter of special design which is to be connected in series with the storage battery on charge. There is a third terminal on back of the instrument by means of which the same instrument can be used to measure the current consumption of the filaments. By pressing a small button in the center of the meter, the needle swings over to the right-hand portion of the scale and indicates whether the battery is fully charged, half charged or low.

STORAGE CELL—See *Secondary Cell.*

STRAIN INSULATOR—An insulator used under tension, as for example the insulators to which the guy wires of an aerial are attached. The simplest form of strain insulator consists of a cylinder or ball having an eye-bolt at each end. Where the aerial guy is fastened to a strain insulator, the purpose of the insulator is to prevent the guy from having a natural wave length nearly the same as the wave length of the antenna. A common form of strain insulator is of nearly spherical form and is so grooved as to carry the two wires firmly, without permitting them to come into contact.

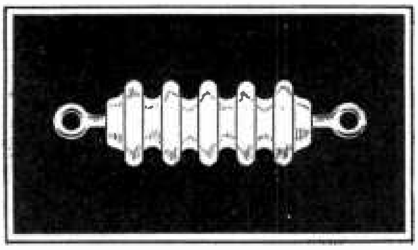

A strain insulator.

STRANDED WIRE—A wire made up of a number of smaller wires, twisted or braided together. An uninsulated, large stranded wire is generally referred to as a bare cable. Stranded wires are of four general classes; (1) wire braid, (2) bunched wire, (3) rope-lay cables, (4) concentric-lay cables. Wire braid is used to afford protection to the insulation of various types of cable. In flat form, it is used as a flexible lead. Bunched wire is a type of cable also referred to as a cord. This form of cable is very flexible, since the individual wires are small. Rope-lay cables are made up of a central core of stranded wire, having one or more layers of stranded wire wound helically about it. Concentric-lay cables are formed in the same way as the rope-lay type, except that the core and layers are of individual solid wires instead of stranded. Stranded wire is very often used in radio work as aerial wire. It has the advantage of being more flexible than solid wire and it also has a lower resistance at high frequencies because of *skin effect* (q.v.). A greater cross-sectional area is available in the stranded conductor than in the solid conductor of the same weight, for carrying current. If used for radio frequency currents, however it is essential that the individual strands be enameled if the lower resistance is to be attained.

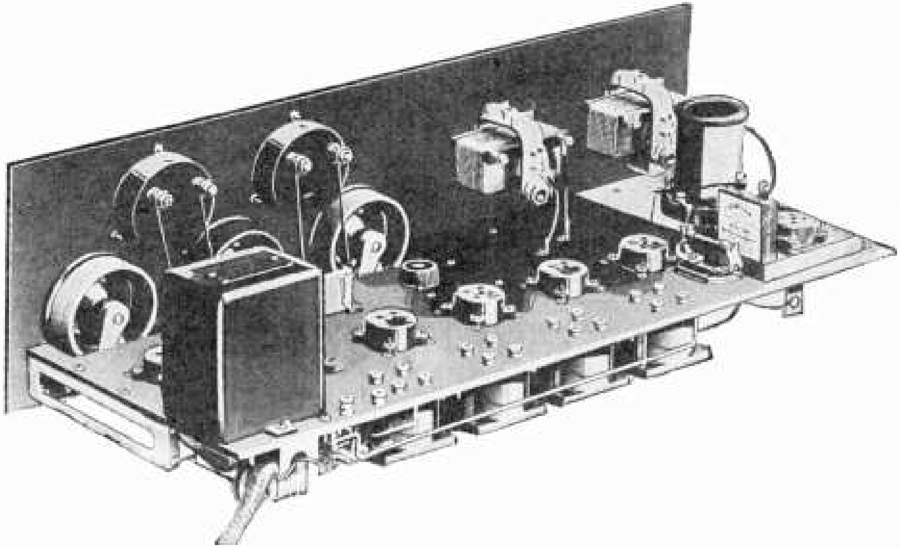

This is the rear view of the Callies Super. Note particularly the long wave transformers at the rear of the set and how they support the sub-panel.

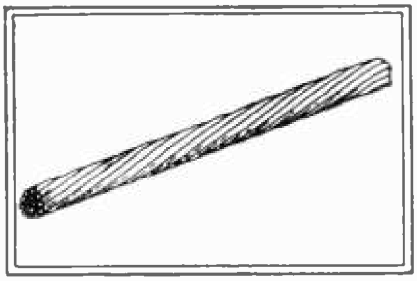

Stranded wire.

STRAY CURRENTS—Currents induced by stray magnetic fields such as eddy currents. Such currents always result in an energy loss. In electric railway systems, the currents returning through the earth, through piping, etc., rather than through the path provided are known as stray currents.

STRAY FLUX—Magnetic flux which is not usefully employed. The *leakage flux* or *leakage lines* (q.v.) which stray from the closed magnetic path provided in a transformer or other electromagnetic or magnetic apparatus. (See *Magnetic Leakage.*)

STRAYS—Atmospheric disturbances which manifest themselves as noises in the radio loud speaker or head set. (See *Static.*)

SUB-PANEL—A secondary panel in a radio set, mounted at an angle with the main panel. The sub-panel usually carries the transformers, sockets, grid condenser and grid leak. Sub-panels are made of bakelite or composition, hard rubber, etc. (See *Panel.*)

SUPER-HETERODYNE—See *Super-Heterodyne Receiver.*

SUPER-HETERODYNE RECEIVER—A circuit used in radio reception in which the wave lengths of the incoming signals are increased to several thousand meters, by the aid of a local source of oscillations. It is a recognized fact that radio signals at "radio frequencies," that is, the original signals as they are impressed on the receiving set from the aerial, must be amplified or built up in some way to operate the detector. However, signals at low wave lengths such as in use for broadcasting, cannot be amplified very efficiently at radio frequencies. By changing the waves from the ordinary broadcast band between about 200 and 550 meters to wave lengths of from 4,500 to 10,000, it is possible to obtain more complete amplification. As a result the receiver is more sensitive and greater distance can be received. In the super-heterodyne receiver, a *difference frequency,* termed a *heterodyne note,* is created. This is done by an arrangement for generating oscillations locally. When the incoming oscillations are combined or superimposed with the local oscillations in such a manner that there is a difference in frequency, this difference will be in the form of a new set of oscillations. If the receiver is arranged in such a form that the new set of oscillations has a comparatively low frequency, the wave length will be correspondingly high. In a standard super-heterodyne receiver the action can be imagined as follows: the signal energy

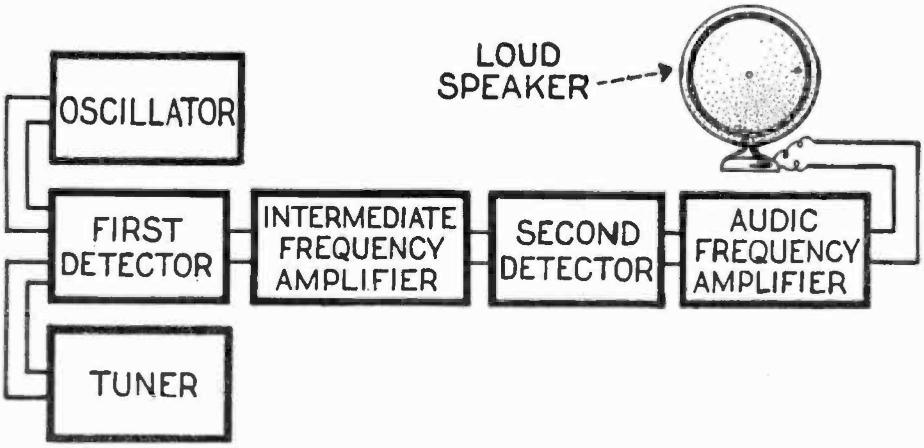
LOUD SPEAKER
OSCILLATOR
FIRST DETECTOR
INTERMEDIATE FREQUENCY AMPLIFIER
SECOND DETECTOR
AUDIC FREQUENCY AMPLIFIER
TUNER

Fig. 1.—This shows the general arrangement of the units of a typical super-heterodyne receiver.

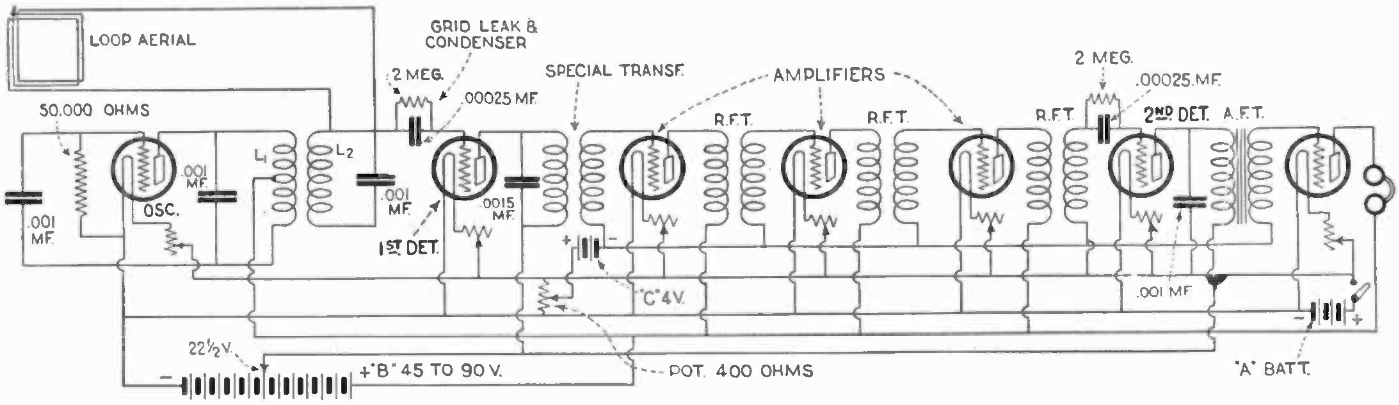
LOOP AERIAL
GRID LEAK & CONDENSER
2 MEG.
.00025 MF
SPECIAL TRANSF.
AMPLIFIERS
2 MEG.
.00025 MF
50,000 OHMS
R.F.T.
R.F.T.
R.F.T.
2ND DET.
A.F.T.
L1
L2
.001 MF
OSC.
.001 MF
.001 MF
.0015 MF
1ST DET.
"C" 4V.
.001 MF
22½V.
+"B" 45 TO 90 V.
POT. 400 OHMS
"A" BATT.

Fig. 2.—Wiring diagram of a standard 7-tube super-heterodyne.

from a certain transmitting station is gathered on an antenna or loop aerial. These signals are tuned in by means of a condenser in the case of a loop aerial, or by a regular tuning device such as a coupler and condenser in the case of an outside antenna. The energy thus obtained is passed through to the grid member of a vacuum tube, which is known as the first detector tube. (In some cases the energy is first amplified by a radio frequency amplifier, consisting essentially of a vacuum tube and a radio frequency transformer). At this point in the circuit, between the tuning device and the first detector, the apparatus for producing the heterodyne action is located. This part of the outfit consists of a vacuum tube with coupled coils in the grid and plate circuits, with a variable condenser (the heterodyne condenser) connected across them, having a suitable value for obtaining any desired local frequency. In this way, the difference between the incoming oscillations and the locally generated oscillations may be adjusted to a prearranged value or frequency. In inductive relation to the grid and plate coils, another coil is arranged, referred to as a *pick-up* coil. The energy obtained from the first detector tube signals having a frequency equal to the difference between that of the incoming signals and the local oscillations, is sent to a *filter transformer* or *tuned filter.* The construction of this filter and the number of windings used is of the utmost importance, as it determines the frequency of the new set of oscillations and will permit the passage of no other frequency than the desired band. This is a very essential feature, as it is necessary to pass only the desired frequency in order to permit maximum transfer of energy. The energy at this predetermined frequency is then passed to the *intermediate frequency amplifier* (q.v.). After being amplified or built up by the intermediate frequency amplifier consisting of tubes and transformers, the energy is then passed to the second detector tube and rectified or changed from radio frequency. The audio frequency signals are now amplified by one or two audio frequency amplifiers consisting of amplifying tubes and transformers and the resulting signals, speech or music

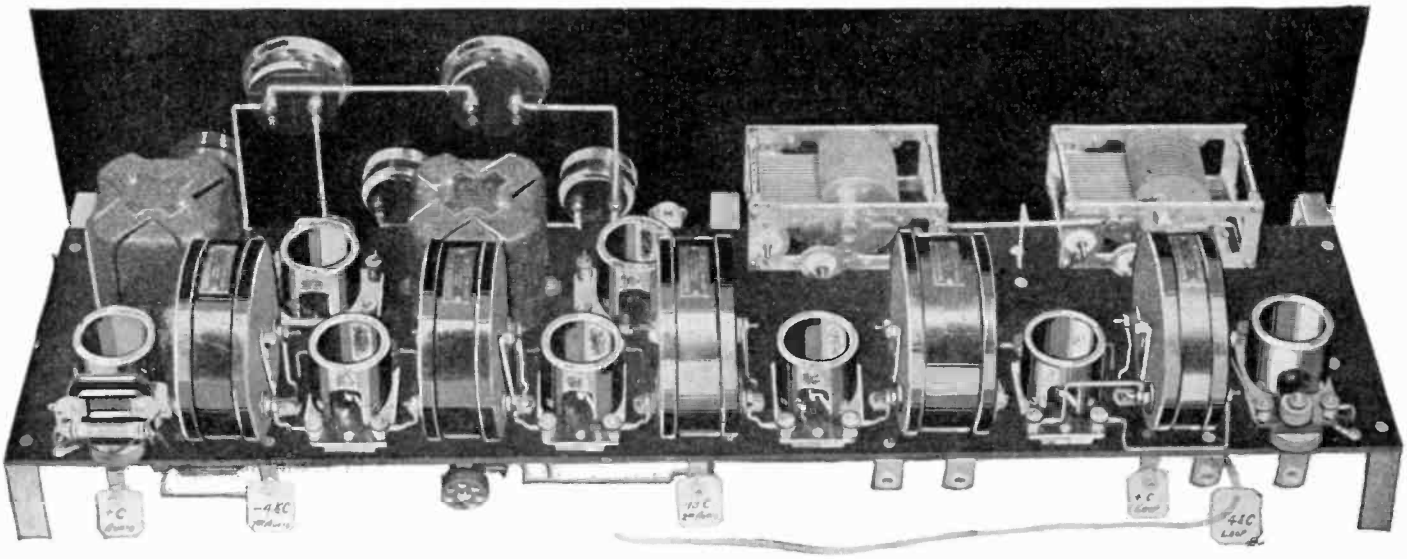

The photograph above shows an eight-tube super-heterodyne set of extra fine quality. Each of the heteroformers are shielded in nickel-plated metal shields. The two variable condensers for tuning the aerial and oscillator circuits should be of the straight line frequency type. The two audio frequency transformers are extra large concert type, the first stage of audio being all that is necessary for average reception on a loud speaker.

are used to operate the loud speaker. Figure 1 shows the general arrangement of the various sections or units of a typical Super-Heterodyne receiver. To go a little more fully into the action of the receiver, let us suppose that a certain station is broadcasting on a wave length of 400 meters. Every wave length is equal to a certain frequency.

That is to say, every wave has a certain number of vibrations per second. The frequency of a wave 400 meters in length will be about 750,000 cycles or 750 kilocycles. By adjusting the wave length condenser, the aerial or loop circuit is placed in resonance with this particular wave length or frequency. This energy is passed through the oscillator pick-up coil and the local oscillations super-imposed on it.

The oscillator condenser, let us say, has been adjusted to permit the oscillator circuit to produce oscillations of a frequency such that the difference between the incoming oscillations and the local ones will be in resonance with the windings of the tuned filter. For instance let us suppose that the tuned filter is arranged for tuning to 10,000 meters, which corresponds to 30,000 cycles.

Now the oscillator circuit must be tuned either 30,000 cycles above or below the incoming wave frequency. In other words, if the incoming wave frequency is as stated—750,000 meters, the oscillator circuit will be tuned to either 780,000 or 720,000 cycles so that the difference between the two frequencies will be 30,000 cycles. No matter what the frequency of the incoming oscillations, the local oscillating circuit will be tuned in a manner to produce the definite difference of 30,000 cycles. These signals will then be passed to the intermediate frequency amplifier and built up before being passed to the second detector tube for rectification.

After this, the signals that are now of audio frequency can be further amplified or built up by the audio frequency amplifier. The two operations that are of the utmost importance in receiving with the super-heterodyne are first, the tuning of the incoming signals and second, the adjustment of the oscillator circuit to produce the difference between two frequencies according to a definite plan. Figure 2 shows the hook-up of a typical Super-Heterodyne receiver.

SUPER-REGENERATION—See *Super-Regenerative Circuit.*

SUPER-REGENERATIVE CIRCUIT—A radio receiving circuit which permits reception under conditions of operation, which in an ordinary regenerative set would result in howling. In the regenerative set, which is tightly coupled, the signal strength increases to a maximum, and thereafter a still further increase in regeneration will result in greatly increased signal strength for a minute fraction of time, followed by violent oscillations. In the super-regenerative circuit, the signals are received at the point of very high amplification by suitably controlling the set. The great advantage of the super-regenerative circuit is the enormous amplification obtainable. Another advantage is the fact that radio code signals are not amplified to the same extent as radio telephony. The super-regenerative circuit was devised by Major Armstrong. The *Flewelling Circuit* (q.v.) is a modification of the Armstrong circuit. (See *Armstrong Circuits.*)

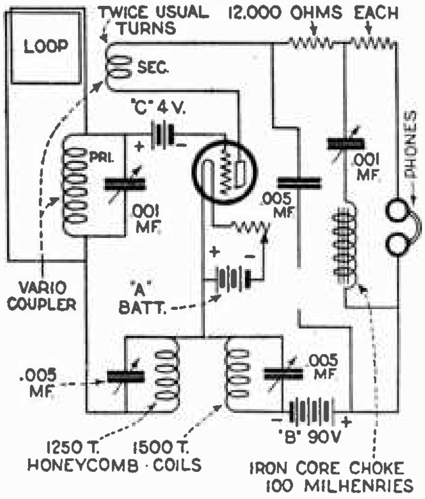

Flivver Armstrong Super-Regenerative circuit.

SURFACE MAGNETISM—A synonym for *Free Magnetism* (q.v.).

SUSTAINED WAVES—Continuous or undamped waves. Waves in which similar current cycles follow each other continuously instead of being broken up into groups like *damped waves* (q.v.). The sustained waves may be interrupted by means of a sending key and thus used for sending radio telegraph messages or they may be modulated and thus utilized in radio telephony. Sustained waves are obtained by the use of high frequency alternators, arc condensers, or thermionic electron tubes (vacuum tubes). The use of sustained waves has made radio telephony practicable. Sustained waves permit sharper tuning, require less energy at the transmitter than in the case of undamped waves, and permit the use of "beat" reception and other sensitive methods of reception. (See *Continuous Waves.*)

S.W.G.—Abbreviation for Standard Wire Gauge. The full name of this wire gage is the *British Standard Wire Gauge*. It is also referred to as the *New British Standard*, the *Imperial Wire Gauge*, and the *English Legal Standard*. It is the legally adopted standard of Great Britain. The following table shows a comparison between the American Wire Gauge (B. & S. Gauge and the Standard Wire Gauge:

Gauge No.	*American Wire Gauge Dia. in Mils*	*Standard Wire Gauge Dia. in Mils*
10	102	128
11	91	116
12	81	104
13	72	92
14	64	80
15	57	72
16	51	64
17	45	56
18	40	48
19	36	40
20	32	36
21	28.5	32
22	25.3	28
23	22.6	24
24	20.1	22
25	17.9	20
26	15.9	18
27	14.2	16.4
28	12.6	14.8
29	11.3	13.6
30	10.0	12.4
31	8.9	11.6
32	8.0	10.8
33	7.1	10
34	6.3	9.2
35	5.6	8.4
36	5.0	7.6
37	4.5	6.8
38	4.0	6.0
39	3.5	5.2
40	3.1	4.8

SWITCH—A device for conveniently making or breaking an electrical circuit. Switches must be designed to carry their rated current without overheating or undue voltage drop, to handle overloads, to prevent arcs on being opened and to properly insulate live parts when switch is open. The most common form of switch is the *knife switch* (q.v.). A switch which opens or closes but one circuit, that is operating in only a single position, is called a *single throw switch.* One operating when thrown in either of two positions is called a *double throw switch* (q.v.). A switch which controls only one side of a circuit is a *single pole switch;* both sides of a circuit, a *double pole switch.* The abbreviation for a single pole, single throw switch is S. P. S. T. The abbreviation for a double pole, double throw switch is D. P. D. T., etc. There are various types of switches ranging from *oil switches* used to handle enormous currents to the small push button or snap switches for turning on or off electric lights. Among the switches used in radio work may be mentioned the *ground switch* (q.v.), *aerial switch* (q.v.), *quick break switch* (q.v.) and *anti-capacity switch.* This latter usually consists of a small handle with a cam attached to its other end which serves to press together or release spring contacts. The electrostatic capacity between the springs is low, due to the construction used. The *plug* (q.v.) and *jack* (q.v.) constitute a form of switch extensively used in radio equipment. The *push-pull switch,* operating as its name implies, is often used to control the filament lighting circuits of radio receiving sets.

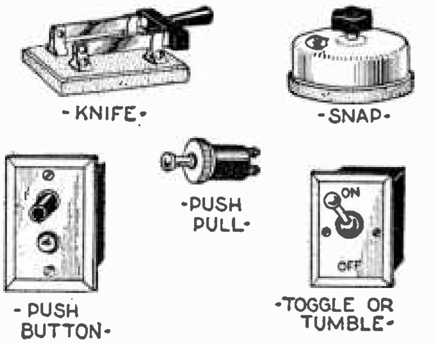

Various types of switches.

SWITCHBOARD—In its broadest sense, this term is applied to any collection of control, operating and measuring apparatus mounted on a panel or panels for the purpose of starting, stopping or otherwise controlling an electrical installation. In a small electrical plant, all control and switch gear may be mounted on a single structure, which is referred to as the switchboard. Switchboards are used at the radio broadcasting station to carry the various meters, control rheostats, switches, etc. (See *Panel.*)

SWITCHGEAR—Apparatus used in connection with the control of an electrical installation, such as switches, rheostats, starters, circuit breakers, etc.

SYMBOLS—Conventional signs used in radio hook-ups to represent various electrical and radio apparatus. The use of symbols has been standardized and results in clearer and more easily drawn diagrams. A table of symbols is given below:

Table of Radio Symbols

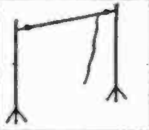	ANTENNA	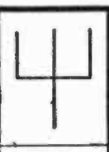	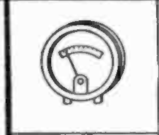	VOLT-METER	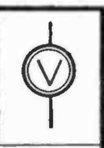	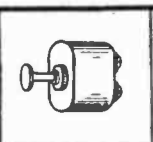	SWITCH FILAMENT	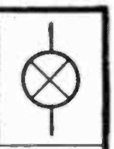
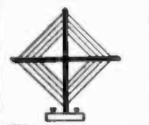	LOOP ANTENNA	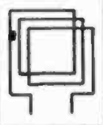	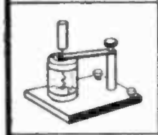	CRYSTAL DETECTOR	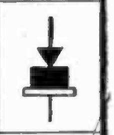	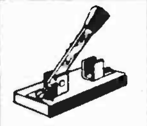	SWITCH S.P. S.T.	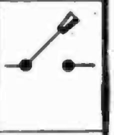
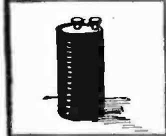	"A" BATTERY DRY CELL	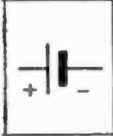	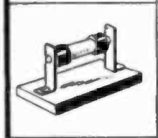	GRID LEAK FIXED	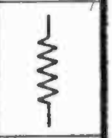	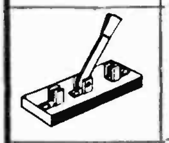	SWITCH S.P. D.T.	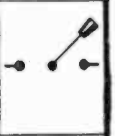
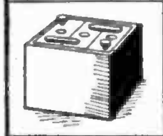	"A" BATTERY STORAGE	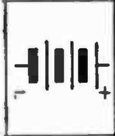	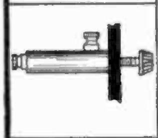	GRID LEAK VARIABLE	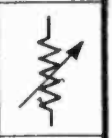	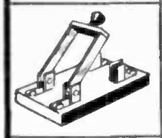	SWITCH D.P. S.T.	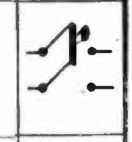
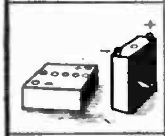	"B" BATTERY	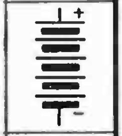	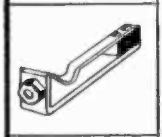	JACK SINGLE CIRCUIT	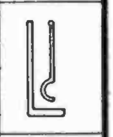	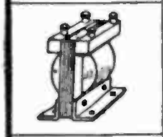	TRANS-FORMER AUDIO FREQ.	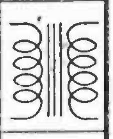
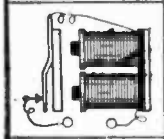	BUZZER	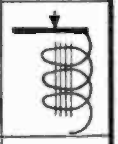	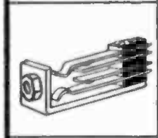	JACK DOUBLE CIRCUIT	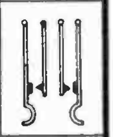	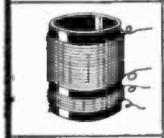	TRANS-FORMER RADIO FREQ.	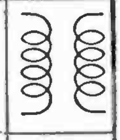
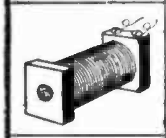	CHOKE COIL AUDIO FREQ.	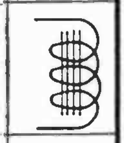	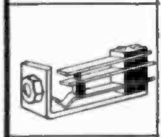	JACK FILAMENT CONTROL	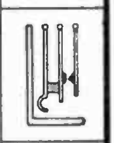	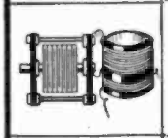	TRANS-FORMER TUNED RADIO FREQ.	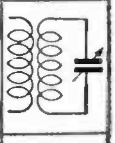
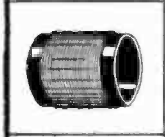	INDUCTANCE COIL FIXED	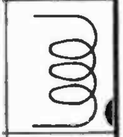	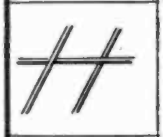	NO CONNECTION	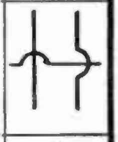	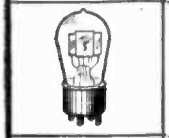	VACUUM TUBE	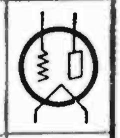
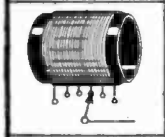	INDUCTANCE COIL TAPPED	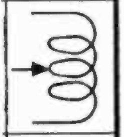	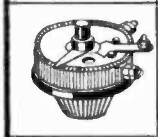	POTENTIO-METER	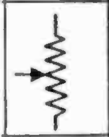	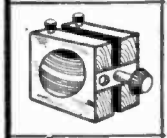	VARIO-METER	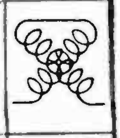
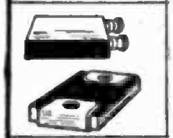	CONDENSER FIXED	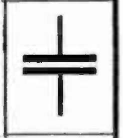	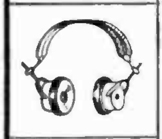	RECEIVERS TELEPHONE	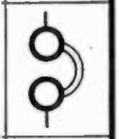	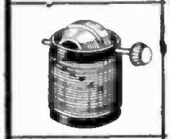	VARIO-COUPLER	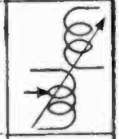
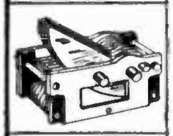	CONDENSER VARIABLE	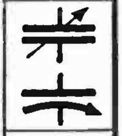	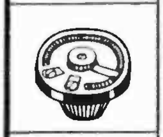	RHEOSTAT	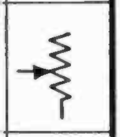	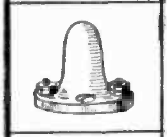	LIGHTNING ARRESTER	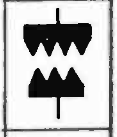
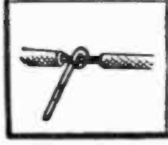	CONNECTION	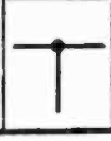	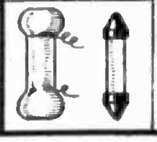	RESISTANCES	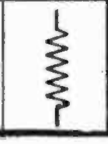	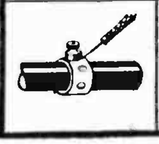	GROUND	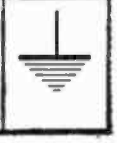

SYNCHRONOUS—As applied to alternating currents, these are said to be synchronous when of the same *frequency* (q.v.) and exactly in *phase* (q.v.).

SYNCHRONOUS DISCHARGER—A rotary gap discharger driven by a synchronous motor from the same line supplying the transformer in a radio telegraph transmitting system or a rotary gap mounted on the shaft of the alternator supplying the 500 cycle or higher frequency current. (See *Disc Discharger.*)

SYNTONY—The relationship between two oscillating currents when one resonates to the waves produced by the oscillations of the other.

T

TABLE OF DIELECTRIC STRENGTH—A table showing the voltages at which certain thicknesses of various materials will break down or puncture. The following table has been obtained from Pender's Handbook for Electrical Engineers:

MATERIAL	DIELECTRIC STRENGTH Specimen thickness mm.	Kv. per mm. (a)
Ambroin	0.84	6.0
Asbestos paper	1.2	4.2
Asphalt (Byerlyte)	3.6	14.0
Bakelite, C-1		up to 27.5
Bakelite, wood molding mixture		17.7 to 21.6
Bakelite, asbestos molding mixture		up to 9.8
Bakelite, Continental	3.2	15.7
Bakelite, Micarta-213		up to 31.4
Bakelite, Micarta-21D		5.9
Bakelite, Micarta-21H		15.7
Bakelite-Dilecto-X	3.2	25.6
Bakelite-Dilecto-XX	3.2	25.6
Celluloid (clear)	0.25	12 to 28
Celluloid (colored)	0.25	10.2 to 18.9
Condensite (molded)	5.7	19.7
Condensite (celoron)	5.7	29.5
Conite	0.13	15.7
Copal	3.0	3.2
Empire cloth, canvas	.41	28.9
Empire cloth, linen	.15	54.0
Empire cloth, muslin	.38	48.0
Empire cloth, silk	.15	48.0
Faturan	3.	10
Fiber, vulcanized, including hard fiber all colors	0.79	8.9 to 16.7
	3.2	4.9 to 10.8
	6.4	3.9 to 8.9
	12.7	3 to 5.9
Galalith (white)		6 to 8.5
Glass (ordinary)		8 to 9
Hermit		2 to 3
Jute (impregnated)	6	1.2
Lava		3 to 10
Litholite	4.5	4.5
Marble		6.5
Mica	6	21 to 28
Micabond, plate	1.6	37.5
Micabond, flexible	1.6	23.1
Micanite, plate	1.6	37.5
Micanite, flexible	1.6	23.1
Minerallac		39.4
Paper	0.13	8.7
Paraffin (parawax)		11.5
Porcelain	20	8
Pressboard (oiled)	0.25	39.3
Pressboard (oiled)	1.58	29.2
Pressboard (oiled)	3.17	21.1
Pressboard (varnished)	0.25	26.3
Pressboard (varnished)	1.58	15.5
Pressboard (varnished)	3.17	9.5
Presspahn		5.2 to 9.3
Redmanol (molded)	5.1	11.8 to 18.5
Redmanol (laminated)	0.8	41 to 51
Rubber (hard)	0.5	70
Slate	10.3	1.3
Vulcabeston	1.9	31.5 to 7.1
Wood (maple), paraffined	15.2	4.6

(a) To obtain volts per mil, multiply kilovolts per millimeter by 25.4.

"T" ANTENNA—A flat top antenna in which the down lead is taken from the center of the horizontal portion. (See *Aerial*, also *Flat Top Aerial.*)

TAILS—Name given to small iron wires forming core of *induction coils* (q.v.).

TANGENT GALVANOMETER—A type of galvanometer in which the strength of the currents which pass through the coil is proportional to the tangent of the angle of deflection of the magnetic needle in the center of the coil. In construction, the tangent galvanometer consists essentially of a small magnetic needle suspended in the center of a large circular coil made of the few turns of insulated wire. Attached to the needle is a light aluminum pointer which allows the deflections to be read on a horizontal scale. The coil is placed in the magnetic meridian with the needle and the coil in the same plane. When the current passes through the coil, the pointer is deflected and the angle of deflection recorded. Knowing the constant of the galvanometer, it is possible to calculate the current in absolute units or in amperes by multiplying the constant by the tangent of the angle recorded.

TANGENT OF AN ANGLE—In a right angle triangle having the given angle as one of its angles, the tangent of the angle under consideration is equal to the ratio of the side opposite the angle to the base of the triangle.

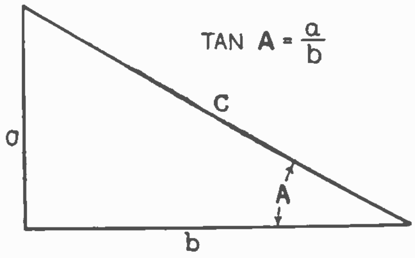

The tangent of angle A is the ratio of side a to side b.

TAP-TAPPING—The connection made to an intermediate point in a winding thus permitting the number of turns in the circuit to be varied at will. Variometers, variocouplers and other coils are made with taps, also transformers, armatures of rotary converters, etc.

TAPPED—See *Tap*, also *Tapped Inductance.*

TAPPER—An electromagnetic device, similar to an electric bell, but with gong removed, arranged so that a hammer gently taps a coherer thus decohering filings after the passage of incoming oscillations. This device was formerly used as a detector of wireless telegraph signals.

TAPPING-BACK—The application of a light blow to a filings coherer to decohere the filings. (See *Tapper.*)

TELEFUNKEN SYSTEM—A wireless telegraph system employing a *quenched spark gap* (q.v.) and giving a singing spark of rather high note. This system has been developed and used by the Germans.

TELEGRAPH—Any system of transmitting intelligence from one point to another over a distance. The word was at one time applied to visual systems such as wig-wagging and semaphore, but at the present is used only to refer to electrical systems. The one exception to this rule is the use of the term to refer to signals sent either mechanically or electrically from the bridge of a ship to the engine room.

TELEGRAPHONE—A magnetic phonograph in which speech currents or signal currents are recorded as permanent magnetic impressions on a moving steel wire. Instruments such as this have been used for recording telephone conversations and also in high speed automatic wireless telegraph transmission. (See *High Speed Reception and Transmission.*)

TELEGRAPH NAUT—A nautical mile of 2,029 yards or 1.1528 statute miles. (See *Nautical Mile.*)

TELEGRAPHY, RADIO—See *Radio Telegraphy.*

TELEPHONE CONDENSER—In radio work, this is a fixed condenser shunted across the head set or loud speaker. Used in this connection, the condenser is also referred to as a by-pass condenser since it offers a path of low resistance to the radio frequency currents. By-pass condensers are usually made of tin foil with a mica or paper dielectric. Condensers used for ordinary telephone work are also referred to as telephone condensers.

TELEPHONE RECEIVER—An electromagnetic instrument by means of which variations of current are caused to reproduce sound waves corresponding to words spoken into a transmitter at a distant point. The essential features of the telephone receiver are an electromagnet, a permanent magnet and a diaphragm held in a suitable case. The diaphragm is supported by the rim of the case, at a short distance from the face of the magnet and it is attracted by the magnet when it is energized by the current flowing in the windings. The steady magnetic flux of the permanent magnet draws the diaphragm to the pole pieces, leaving a small gap between, so that the diaphragm is held under tension. The currents passing through the electromagnet are modulated in accordance with the transmitted speech. These change the tension of the diaphragm due to the change in magnetic pull and as a result the diaphragm vibrates giving forth speech or music. The telephone receivers used in radio work are of the *watch case* type. (See *Head Telephone or Head-Phone.*)

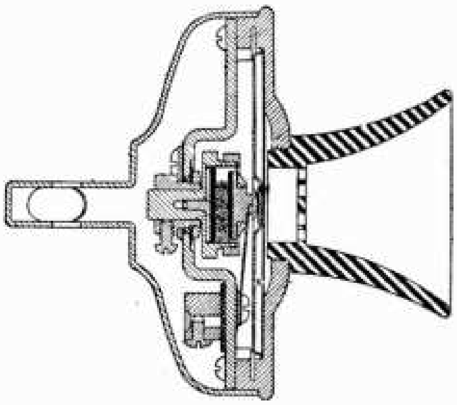

A cross sectional view of a telephone transmitter showing details of construction.

TELEPHONE TRANSMITTER—A sound-wave operated or vibration-operated device designed to produce electromagnetic waves or vibrations which correspond to the sound waves or vibrations actuating it. The standard transmitter used in ordinary telephone work is known as the "solid-back" transmitter. A gong-shaped back supports the transmitter and carries all the parts. A mouth piece of hard rubber is screwed into the front. Just to the rear of the front is placed an aluminum diaphragm which fits into a receptacle cut out for it. An insulated cushion seat of rubber is placed above and below the diaphragm. The diaphragm is held securely by damping springs having soft rubber cushions at their tips, so that it can only assume forced vibrations. An auxiliary diaphragm of mica is used which is fastened to the front electrode of the transmitter button. Both the front and the rear electrodes are made of carbon discs, between which carbon granules are placed in a cylindrical chamber lined with varnished paper. A rigidly connected pin communicates the vibrations of the diaphragm to the front electrode. The movement of the electrode varies the pressure on the carbon granules, thus giving the required variation of resistance which in turn varies the current. The *microphone* (q.v.) used in radio broadcasting is a modification of the telephone transmitter.

TELEPHONE, TUNED—See *Tuned Telephone.*

TELEPHONY, RADIO—See *Radio Telephony.*

TELEVISION—Electric vision at a distance. The transmission by wire or by radio of vision—that is to say the ability to see objects at any distance by changing the light waves to electric currents or to electromagnetic waves, transmitting these to the distant receiving point, and then retransforming them back to the original light waves. Television is still in the experimental stage. Of course, the sending of pictures by radio has been accomplished by Capt. Ranger, in this country, Belin in France, Karolus in Germany, Baird in England, and several others. Both Belin and Baird have been working on the problem of television and are reported to be nearing success. In this country, Mr. C. Francis Jenkins, of Washington, D. C., has constructed a non-commercial (that is in its present stage) machine which will send motion pictures by radio. This is practically the last word in the present development of television. It is possible to broadcast a film picture with this device at the rate of sixteen pictures per second showing the moving outlines as silhouettes, but without the intermediate tones. The method by which a photograph is transmitted by radio in the Jenkins system may be illustrated by the example of a coin covered by a thin sheet of paper on which serried pencil lines are drawn. The design of the coin appears in lines varying in intensity. The variations in the lines may be transformed into pulsing electric current by means of a light sensitive cell passing over the lines in question. Place the lines end to end and the current can modulate a carrier wave. The forming of the continuous line presented a difficulty which was solved in the following way. The inventor proceeded along the line that a prism bends a ray of light and that an image will appear upon a photographic lens placed before a picture. From this it follows that a prism placed near a lens will cause the image to be displaced sideways, the extent of the displacement being dependent on the angle of the prism. If the prism could change its angle, the result would be motion on the part of the image. The prism used is circular, with rapid change of form across the prismatic section. Around the circumference of the disc, the prism changes in thickness according to a graduated scale. Revolving the disc causes the image to move in a straight line. When the end of the prism passes the lens, there is instantaneous snapping back of the picture and the journey is repeated. If a small hole be made in the screen, the light from a given point on the picture will pass through and can be made to operate a light sensitive cell at the back of the screen. When the image is made to move down over the screen by the prismatic ring, the aperture will admit light of varying intensity, according to the light value along a line crossing the picture. This means that the light falling on the photo cell as the image travels will represent a line across the picture from top to bottom. This leads to the creation of the successive lines that make up the picture as a whole. The image is moved along from side to side by means of a second prismatic ring working along at right angles to the first, but at much lower speed. This second prism makes but one revolution during the time that the first makes one hundred, producing the hundred lines assumed to be necessary for making the picture. The cell used may be selenium or thalium oxide. The former requires twenty minutes to cover the picture, the latter six. It is obvious that to reproduce the picture at the other end the process must be reversed, the lines being replaced side by side by means of a pair of prismatic discs similar to those at the transmitting end. As noted above, the main difference between sending photographs and motion pictures is in the necessary speeding up. For this purpose, Mr. Jenkins uses a potassium light-sensitive bulb, the action of which is much more rapid than that of either selenium or thalium oxide. The cell is composed of a bulb with its inner surface coated with metallic potassium. The bulb has a high vacuum. The device has two electrodes, one being a wire through the stem, ending in a loop in the center of the bulb; the other passing through the side of the bulb and connecting with the potassium deposit. Absence of the potassium from a small part of the surface forms a "window" opposite the last named contact. Application of a positive potential to the central electrode, attended by connection with the potassium coating, permits no current to flow as long as no light enters the bulb. The admission of light frees electrons from the potassium, in numbers proportionate to the strength of the light. In this way a current flows through the bulb when light enters the "window." In this cell, there is no appreciable lag due to the amazing speed of the electrons.

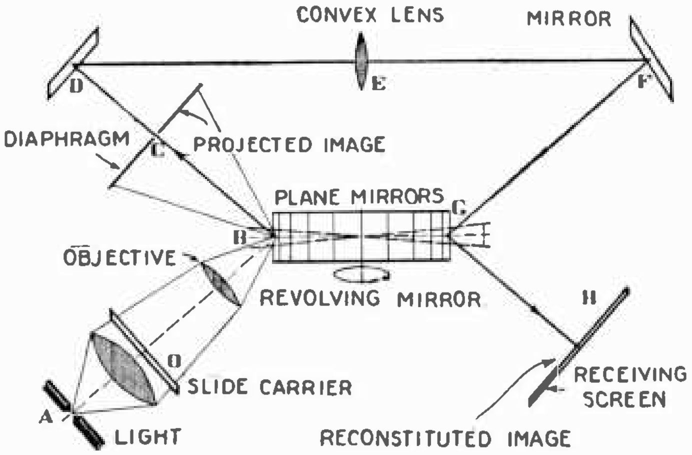

The cylinder is covered with plane mirrors, revolving downward on the side toward the lantern. One ray' at a time from 1/25,000 of the area of the image, passes through the opening in C. The fixed mirrors D and F send it back to the mirror G, opposite B on the cylinder, and it is finally reflected against H in a position corresponding exactly to the portion of the image from which it was first taken. The effect of continuous vision is produced.

The sending station has a motion picture projector, a small machine used to cut up the image, a light-sensitive cell, and a transmitter. The receiving station has a radio receiving set, a machine for reassembling the picture, and a screen for the reproduction of the image.

The lamp required for the receiving station must be able to pass swiftly from darkness to extreme brilliancy, and the ordinary filament lamp could not attain the tremendous speed involved. The lamp used by Mr. Jenkins is the creation of Professor McFarlane Moore, and is capable of handling complete cycles at the rate of 75,000 to the second. This lamp has a glass bulb containing two concentric cylinders, the larger almost filling the bulb, and the smaller almost filling the larger. The inner cylinder has a small axial hole, drilled almost to the bottom and forming a deep cup. The principle of this lamp is much the same as that of the Amrad "S" valve.

It is a truism that the fundamental principle of cinema pictures is that the human eye continues to see an image

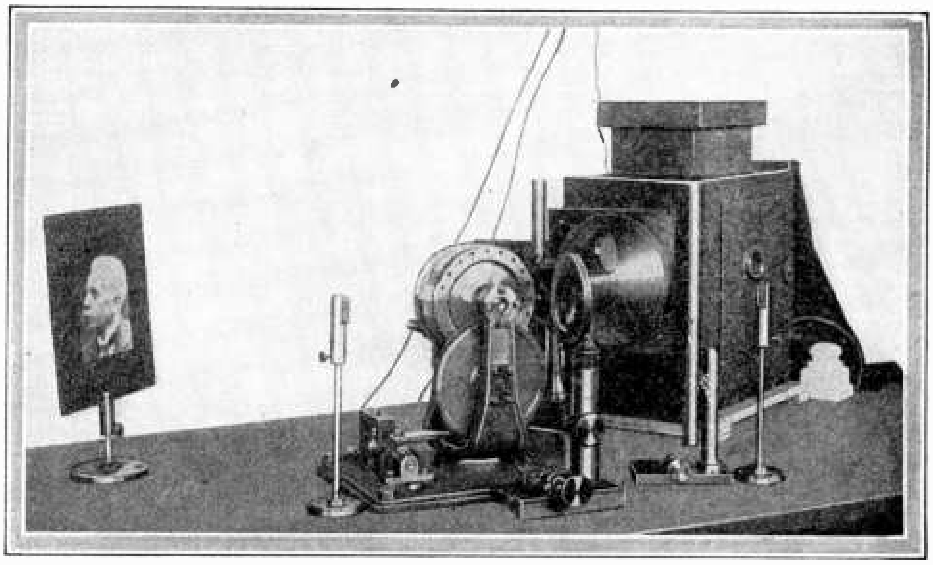

This apparatus is diagrammed above: the projecting lantern at the right; the drum of mirrors, center; and the adjusting stands for diaphragm and fixed mirrors in front. At the left a continuous image appears on the screen, although only 1/25,000 of it is actually projected at any instant.

after that image itself has disappeared. In the case of an electric spark lasting one ten thousandth of a second, the image lasts for at least one-sixteenth of a second as visualized by the eye. Mr. Jenkins assumes that if 10,000 dots of light and shade flash successively on a screen with sufficient speed, the eye will see the picture to which the dots belong even though but one dot is on the screen at a time. He, therefore, aimed at flashing the 10,000 dots on the screen within the space of one-sixteenth of a second.

Theoretically, this feat might be accomplished by halving the slow disc revolving rate to 960 revolutions per minute, keeping the faster disc at a rate exactly one hundred times greater. But the glass disc would be shattered by the centrifugal force before the attainment of a speed of 96,000 revolutions per minute.

Mr. Jenkins gets over this difficulty by the use of forty-eight lenses attached to a large aluminum disc rotating in front of the prismatic disc, thus giving the effect of a slow-moving lens. Each of these forty-eight lenses makes a line across the plate, which thus becomes the screen of the motion picture apparatus. This makes it possible to reduce the speed to one forty-eighth of that which would otherwise be required, that is, to one of 2,000 revolutions per minute.

At the receiving station, the modulated wave is received by a set but slightly different from many in common use. The last amplifier tube is one of 5 watts. By the use of the McFarlane Moore light already described, the modulations in the plate circuit of the last valve are transformed into light, and the rays are distributed over the screen by prismatic and lens discs similar to those used at the sending station. In actual practice, the picture appears on a screen about 6 in. by 8 in. More powerful lamps would allow of larger pictures.

The process described is for sending photographic pictures. But it can be adapted for the sending of direct views without the intervention of photography.

All that is done in this case, is to remove the projector and focus the lens so that it will throw the view as an image.

So far as radio features of the transmission are concerned, there is little to be explained. At the receiving station the circuit resembles the superheterodyne. The picture is at 75 kilocycles frequency and does not interfere with broadcasting.

The carrier wave may be modulated with voice frequency, so that the voice can travel with the picture, which enables vocal explanations to be given simultaneously with the flashing of the film or actual scene being transmitted.

TEMPORARY MAGNETS—Magnets which lose their magnetic properties as soon as the magnetizing force is removed. Soft iron displays the properties of a temporary magnet and hence is ideal for use as the core of electromagnets. (See *Magnet*, *Electromagnet*, also *Permanent Magnet*.)

TERMINAL—A binding post or other fitting attached to the electrodes, ends of windings or other parts of electrical apparatus so as to permit the external circuit to be connected to the apparatus.

TESLA COIL—An oscillation transformer for producing high potential discharges from oscillations of low potential. The *Tesla Coil* or *Tesla Transformer*, as it is sometimes called, is somewhat similar to the ordinary transformer, although it is much more heavily insulated and has the ends of the secondary connected to a condenser which discharges across a spark gap, thus increasing the rapidity of the oscillations, which then pass into a secondary induction coil. This second coil has no iron core. The Tesla Coil consists essentially of a primary winding having a relatively small number of turns of heavy wire and a secondary having a large number of turns of fine wire. An air, glass, or ebonite dielectric separates the primary and the secondary. Due to the sufficient thickness of the dielectric, there is no possibility of a direct discharge between the two windings. Both the primary winding and the secondary winding consist of but one layer each. The primary winding forms part of an ordinary oscillation discharge circuit. A high frequency electromotive force is induced in the secondary winding, due to the high frequency oscillating currents which flow in the primary winding thus setting up an oscillating magnetic field. Due to the high ratio between the windings of the primary and the secondary, very high voltages are set up in the secondary, in some cases as high as a million volts, and hence a very powerful brush discharge can be obtained.

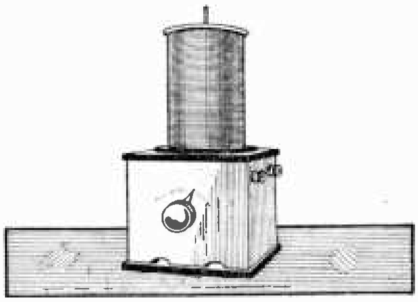

Tesla coil.

TESLA, NIKOLA—American electrical and radio expert. Tesla was born at Smiljan, in Jugo-Slavia, 1857, and was educated at Graz and Prague Universities. He entered the Austrian Telegraph Service and in 1884 he came to the United States where he became an assistant to Edison. He specialized in the study of high frequency and high potential alternating currents. Tesla is noted for his many revolutionary inventions which include polyphase alternating current systems, the rotating field alternating current motor, arc lighting, the *Tesla transformer* (q.v.), and radio apparatus of many kinds including *radio control* (q.v.) devices.

Nikola Tesla.

TESTING CIRCUIT (BUZZER)—A circuit in which a buzzer is used for the purpose of testing or of locating faults in radio apparatus. A buzzer connected in series with a dry cell can be used for locating open circuits and also for testing for short circuits. A buzzer connected in series with a battery and an inductance coil, so as to form a closed circuit and having the inductance coil placed so as to be coupled with the antenna coil of a crystal detector receiving set, can be used to determine the most sensitive point of contact of the crystal. A *tuned buzzer tester* is sometimes used in connection with radio direction finder systems.

THEORY OF CURRENT FLOW—While it is assumed for convenience, that an electric current flows along a conductor from the positive terminal of the source and back to the negative terminal, modern theory is that the electrons actually travel in the opposite direction. Whenever there is a flow of electricity in a conductor, it is believed that extremely small particles of electricity, called *electrons* (q.v.), pass along this conductor. Matter is made up of atoms, which in turn are composed of electrons. The electrons are minute particles of negative electricity revolving around a positive particle. All electric current flow is based upon the motion of electrons within a conductor.

THEORY OF DETECTOR ACTION (OF VACUUM TUBE)—The vacuum tube used as a detector performs a two-fold purpose—it rectifies the current so that it is capable of actuating a telephone receiver and in addition it amplifies or strengthens the current. The rectification takes place due to the fact that the electron flow is only possible from the filament to the plate, but not in the reverse direction. The radio frequency currents applied to the grid regulate the number of electrons passing from the filament to the plate, the variation being exactly in accordance with the variations of the radio frequency current. (See *Detector*, *Vacuum Tube*.)

THEORY OF OPERATION OF VACUUM TUBES—The operation of the three-electrode vacuum tube depends upon the emission of electrons from the hot filament to the plate and the control of the flow of these electrons by means of the potential variations applied to the grid. The vacuum tube consists of an evacuated bulb containing a tungsten filament treated with thorium, an anode or plate, which may be in the form of a cylinder or plate, and a grid consisting of a wire grating inserted between the filament and the plate. When the filament is heated it gives forth electrons which are attracted to the plate under the influence of a difference of potential by which the plate is maintained positive with respect to the filament. The grid is an automatic control, regulating the flow of electrons from filament to plate. Potential variations applied to the grid serve to vary the electron flow between the filament and the plate. The circuit in which the vacuum tube is used may be divided into two parts, the input circuit connecting the filament to the grid through the secondary of a transformer or other method or applying potential changes to the grid, and the output circuit between filament and plate. The vacuum tube may be used as a detector or rectifier, as an oscillator or generator of high frequency oscillations, as an amplifier, as an electrostatic voltmeter, as a voltage and current regulator, a power limiting device, etc.

THEORY OF PROPAGATION OF ELECTROMAGNETIC WAVES—Waves are due to vibrations transmitted through a continuous elastic medium. Electromagnetic waves are due to vibrations caused by an oscillating electric charge in a circuit (the transmitting aerial circuit). The current flowing in the aerial is oscillating (surging back and forth) at an enormously high frequency and in addition to its induction field, the aerial has associated with it a *radiation field* which spreads out in all directions from the aerial. The strength of the radiation field decreases in inverse proportion to the distance from the aerial. At any given point the strength is directly proportional to the frequency. The transmission of electromagnetic waves is considered as due to the motion of the lines of force. There is a displacement of electricity along these lines against the elastic force of the medium, termed for convenience the *ether* (q.v.). Displaced electricity continuously tends to go back to its position of rest under the action of the elastic forces. There is pressure at right angles to the lines of force in addition to the tension along them. The pressures may be considered as due to the repulsion between the displaced charges of the same sign in neighboring lines of force.

THEORY OF RECEPTION OF ELECTROMAGNETIC WAVES—As electromagnetic waves cut across a receiving aerial, the electric field intensity along the aerial alternates in value. There is an alternating potential between the aerial and the ground which gives rise to the flow of an alternating current. Another explanation is based on the principle of induction. The magnetic and electric fields, in cutting the aerial, induce electromotive forces which cause current to flow. This is radio frequency current and it is passed to the grid of the detector tube and rectified as explained under *Theory of Detector Action* (*of Vacuum Tube*). (See *Electromagnetic Waves.*)

THEORY OF VACUUM TUBE OPERATION—See *Theory of Operation of Vacuum Tubes.*

THERMAL—Pertaining to heat. There are various electrical and radio devices which depend upon heating or thermal properties. The *thermal ammeter* (q.v.) is described under the heading *Hot-Wire Ammeter.* The *thermal detector* is taken up under *Barreter* and *Liquid Barreter.*

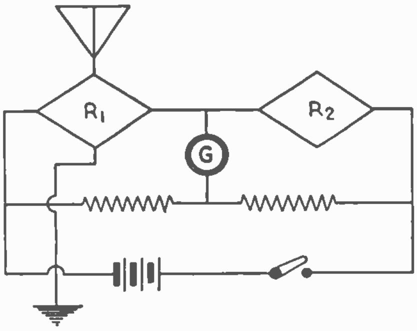

A thermal detector. By means of the change of resistance of a conductor, due to heating, electro-magnetic waves are detected. R_1 and R_2 are two rectangles of very fine iron wire.

THERMAL AMMETER—A current meter, in which the current or a fixed proportion of the current in question, passes along a fine wire. This heats the wire, causing it to expand or sag. This deflects a pointer or mirror across a calibrated scale. The thermal ammeter can be used for measuring direct or alternating currents. (See *Electrothermal Meter,* also *Hot-Wire Ammeter.*)

THERMAL DETECTOR—A radio *detector* (q.v.) which depends for its operation upon the heating of a fine wire by the passage of electrical oscillations. (See *Barreter.*)

THERMAL JUNCTION—THERMO-ELECTRIC JUNCTION—The contact *point* or joining point of the two dissimilar metals of a *thermo-electric couple* (q.v.). (See *Klemencic Thermal Junction.*)

THERMAL TELEPHONE—A telephone receiver in which the movements of the diaphragm are regulated by the variations of expansion of a wire heated by the telephone currents thus reproducing sound waves. In another type of thermal telephone, the diaphragm is dispensed with, sound waves being reproduced directly through the expansion and contraction of air in contact with the heated wire. This device is also known as a *thermal receiver* or a *hot-wire telephone.*

THERMIONIC EMISSION—The emission of a stream of negative electrons from a heated filament (cathode) in a vacuum tube.

THERMO-ELECTRIC COUPLE—THERMO-COUPLE—A pair of dissimilar metal pieces placed in contact and connected by electrical conductors in a closed circuit. When the junction of the two dissimilar metals is heated, an electric current will flow.

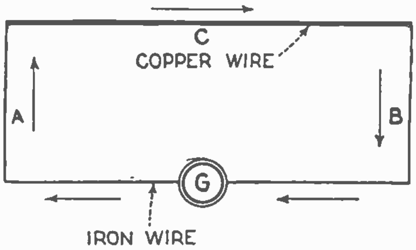

A current is caused to flow by heating the junction of two dissimilar metals. If the end at B is heated the direction of flow will be as shown by the arrows.

THERMO-ELECTRIC CURRENT—A current which is caused to flow due to an electromotive force set up at the junction of two dissimilar metals when these metals are at different temperatures. (See *Peltier Effect.*)

THERMOPILE—A device for magnifying the effect of the thermo-couple by connecting a number of these in series and exposing one set of alternate junctions to the heat, thus causing the electromotive forces to add up.

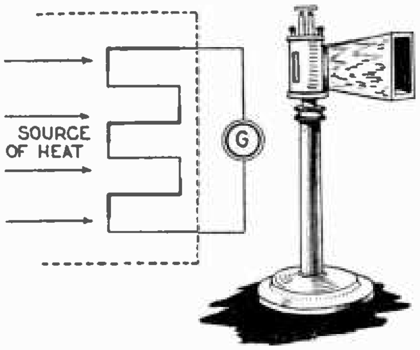

At the left is shown method of joining thermopile junctions to multiply the effect. At the right is a sensitive thermopile.

THOMSON EFFECT—The phenomenon of the appearance or disappearance of heat when a current flows from a cold towards a hot part of a conductor. If copper is unequally heated, heat is liberated at a point when the current and the heat flow in the same direction and is absorbed when they flow in opposite directions. The Thomson effect is measured by a quantity termed "the specific heat of electricity."

THORIUM—A rare metal which in certain compounds in radio-active. This metal has a specific gravity of eleven and an atomic weight of 232.5. Thorium is used for coating Welsbach mantles and is also used in connection with the coating of the tungsten filaments used in vacuum tubes. (See *Filament, Thoriated,* also *Thorium Treatment of Filaments.*)

THORIUM TREATMENT OF FILAMENTS—A compound of thorium, thoria (ThO_2) is used extensively in the manufacture of filaments for vacuum tubes. Sometimes thorium nitrate is used for this purpose. The filament so treated is termed thoriated tungsten. Since the rate of electron emission from a thoriated tungsten filament at a given temperature is several thousand times greater than in the case of the ordinary tungsten filament, it is possible to use the former at a much lower temperature. (See *Filament, Thoriated.*)

THREE CIRCUIT TUNER—A *regenerative circuit* (q.v.) in which the primary, the secondary or grid circuit, and the plate circuit are capable of being tuned in resonance with each other.

THREE-ELECTRODE THERMIONIC TUBE—An evacuated bulb or vessel containing a combination of a heated cathode, a relatively cold anode and a third electrode for controlling the current flowing between the other two electrodes. (See *Triode,* also *Vacuum Tube.*)

THREE-ELECTRODE TUBE—The three element thermionic vacuum tube having a filament or cathode, a plate or anode, and a grid. This tube can be used with minor modifications as a *rectifier* (q.v.) or *detector* (q.v.) of radio signals, an *amplifier* (q.v.), an *oscillation generator* (*oscillator* (q.v.)), an *electrostatic voltmeter* (q.v.), as a *power limiting device,* etc.

THREE WIRE SYSTEM—In direct current work, a system of transmitting electrical energy, which utilizes three wires, one of which is maintained at a potential, midway between the other two. A part of the load is connected between one wire and the neutral and part between the other wire and the neutral. As a result the neutral wire only carries current equal to the difference in the loads. The main advantages of this system are a saving in copper and availability of double voltage. Thus 110 volts may be required for lights, whereas 220 may be wanted for electric motors. In alternating current work reference is sometimes made to a two-phase three wire system. In this system there are two single phase currents which differ in phase by 90 degrees. Each current has a separate outgoing wire, but unites in a common return wire. (See *Neutral Wire.*)

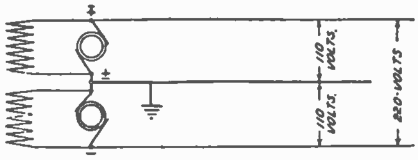

THE EDISON THREE WIRE SYSTEM

TICKER or TIKKER—A form of interrupter or an equivalent apparatus, used in sustained wave telegraphy. It is really a commutator interrupter. In a commercial type illustrated, a disk is mounted on a motor shaft. The disk has a number of teeth filled in between with fibre or other insulating material. The radio frequency currents flow from one brush to the other through the disk which interrupts them at the rate of from 300 to 1000 times per second. A charge is built up in the condenser C_a by resonance with the aerial system and this discharges into the telephone condenser C_b at regular intervals. The telephone condenser then discharges through the head set, thus producing a single sound corresponding to the charge accumulated. Since the tikker discharges the condenser, C_a at different places on the cycle of the incoming oscillations, the note produced is not uniform and is hard to read through atmospheric interference. (See *Chopper.*)

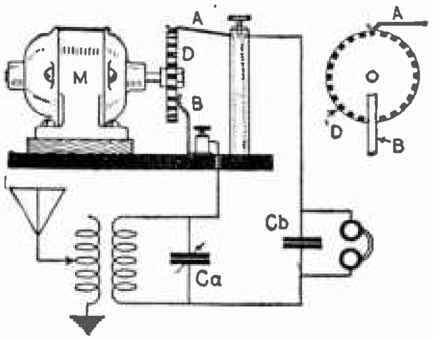

A simple Poulsen ticker. This is used to interrupt circuits of a receiving tuner at a uniform rate per second.

TICKLER—TICKLER COIL—A coil placed in the plate circuit of a regenerative receiver and coupled inductively in the grid coil in order to obtain regeneration. (See *Feed-Back, Feed-Back Coil, Feed-Back Coupling, Regenerative Circuit, Regenerative Coupling,* also *Regeneration.*)

TIME CONSTANT—The ratio, in an electric circuit, of the inductance in henries to the resistance in ohms. It is the time, in seconds, required for a current to attain 63.2 per cent of its ultimate value as given by Ohm's Law. (See *Impulse E.M.F.*)

TONE WHEEL—An apparatus used in the reception of wireless telegraph signals in continuous wave systems, utilizing a wheel in which the conducting teeth are separated by insulating material. A brush is used to make contact with the wheel. The tone wheel is similar in principle to the *tikker* (q.v.) except that it converts the incoming oscillations to audio frequency currents, thus doing away with the non-uniform note of the tikker. In addition to this the tone wheel utilizes more of the incoming energy than does the tikker. The tone wheel is run at such a speed that the frequency of the interruptions differs slightly from that of the incoming oscillations thus giving *beats* (q.v.) which produce audio frequency currents which can be heard in the telephone receiver. Since the tone wheel mechanically converts radio frequency currents into audio frequency currents it is sometimes called a *frequency transformer.*

TORQUE—The moment of a system of forces tending to produce rotation. It is usually measured in foot-pounds. Torque may also be defined as that which produces or tends to produce torsion. Thus if a force acts at a tangent to the periphery of a circle, thus producing a turning moment about the center of the circle, the product of the force and the radius of the circle is called the torque.

TRAIN RADIOPHONE—Communication by means of radio between moving trains and ordinary land-line telephone systems. Such a system is in operation commercially in Germany on trains operating between Berlin and Hamburg. The voice frequency currents from the telephone are impressed on a carrier wave by a method similar to the *Heising Modulation* (q.v.) system. Each train has a special wave length assigned to it. The modulated wave is emitted through a four-wire antenna, strung along the roofs of two coaches. The nearby telegraph wires pick it up and carry it to the receiving sets, located in stations along the line. The ground for the system is obtained through contact of the car wheels with the rails. In reverse order, the voice frequencies of messages from the telephone at the other end are used to modulate a carrier wave (using a different frequency) at the nearest train-telephone switchboard. They are then transmitted over the wires and inductively to the train. An arrangement is provided to prevent interference between the sending and the receiving sets.

TRANSFORMER—A type of induction apparatus, usually stationary, having primary and secondary windings, ordinarily insulated from each other. These are wound about an iron core. The function of the transformer is to change the electrical energy either increasing or decreasing the voltage, current, frequency or phase. Transformers can only be used in connection with alternating currents. Transformers for changing voltage have a different number of turns in the primary than in the secondary. In the step-up transformer, the primary usually consists of a small number of turns of comparatively heavy wire, while the secondary consists of a larger number of turns of fine wire. The ratio of the primary and secondary windings determines the ratio between the input and the output voltages; thus a 10 to 1 step-up transformer could be used to step a primary voltage of ten to a secondary voltage of 100. There are various methods of classifying transformers, namely, according to their operating characteristics, according to their method of cooling, or to their construction. Transformers which are meant to give a constant voltage on the secondary side are known as *constant potential transformers.* Those meant to give constant current are called *constant current transformers.* Transformers may be air cooled, water cooled or oil cooled. In some cases a natural air draft is used as in transformers of the smaller types. Larger transformers may be designed with special ducts through which blowers force a current of air. Oil cooled transformers may have the cores and windings immersed in oil, or the oil may be circulated through external coils. Water cooled transformers are immersed in oil, but have pipes carrying running water submerged in the oil. According to construction, transformers are of the core type and the shell type. The core type transformer is used more for high voltage, low capacity work, while the shell type is used for low voltages and high capacities. Transformers find a wide application in radio apparatus. *Audio transformers* are used to couple audio frequency circuits as for example in the case of an audio frequency amplifier. Audio transformers have an iron core and may be *shielded* or *unshielded. Radio frequency* transformers are used to couple radio frequency circuits. They are usually of air core type, although in special instances iron cores are used. Transformers are also used in alternating current chargers, in "A" and "B" battery eliminators and in power sets which operate directly from alternating current houselight-ing circuits. (See *Charger, Storage Battery, Audio Amplifier, Coupling Transformer, Jigger, Intermediate Transformer, Inductive Coupler, Oscillation Transformer, Air Core Transformer, Core Transfomer, Ratio of Transformation,* also *Shielded Transformer.*)

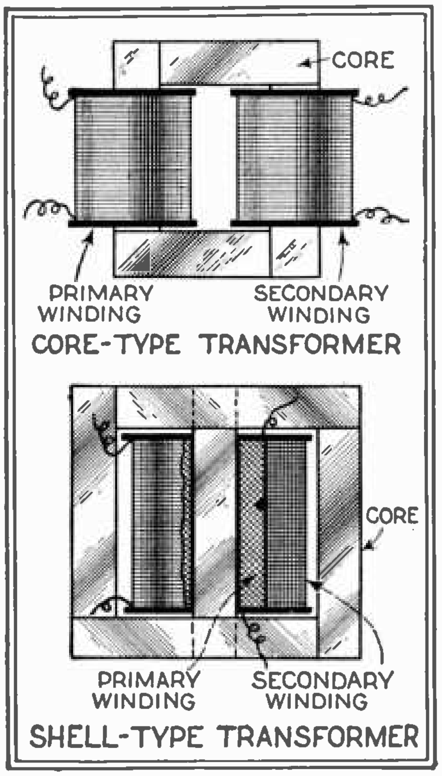

Two standard types of transformers.

TRANSFORMER COUPLING—A method of transferring the electrical energy of an alternating current from one circuit to another inductively by means of a transformer, one of whose windings is connected in one circuit, the other winding being connected in the second circuit. Radio frequency circuits are usually coupled by *radio frequency transformers, audio frequency transformers* being used to couple audio circuits. (See *Coupling,* also *Amplifier, Intermediate Frequency.*)

TRANSFORMER STEEL—A special steel usually containing a high silicon content used for stamping out transformer laminations. (See *Eddy Currents.*)

TRANSFORMER, TUNED—See *Resonance Transformer.*

TRANSMISSION—As applied to radio communication, this is the sending of signals, speech or music from one point to another by means of electromagnetic waves. (See *Radio Telephony, Radio Telegraphy, Theory of Propagation of Electromagnetic Waves.*)

TRANSMISSION OF PHOTOGRAPHS BY RADIO—Various methods have been devised and are now in use for the transmission of photographs by radio. Among these may be mentioned the systems of *Belin* (q.v.), *Baird*, and *Jenkins*. The principles underlying the Jenkins system are explained under the heading of *Television*. Using the system developed by Capt. R. H. Ranger, photographs were transmitted by radio from Honolulu to New York, a distance of 5,136 miles. Recently commercial picture transmission service has been inaugurated between New York and London using the Ranger apparatus. Two distinct methods have been applied for analyzing the picture in the process of transmission. One arrangement consists of producing an image as a non-conducting deposit upon a metal foil which is traversed by a stylus, while the other method makes use of an opaque image deposited upon a transparent film which is traversed by a beam of light, the light interruptions being recorded by a light sensitive cell. The Ranger system makes use of this latter method. The image is photographically recorded upon a celluloid sheet large enough to accommodate easily a picture of half-plate size. In the case of sketches and written messages the image for transmission is made directly by writing upon a piece of transparent film with a dense black ink. The image is then secured to the face of a glass cylinder, and by means of a lamp, focussing lens and reflecting prism, a narrow pencil of light is passed through the film. The cylinder is mounted on a carrier which is caused to be moved backward and forward so that the beam of light is concentrated in turn on all parts of the picture. A rotary motion is applied, as well as the transverse motion, the cylinder being given a slight rotation when it completes each transverse motion. The beam of light is passed through a special photo-electric cell. This consists of a spherical globe, coated on the inside with potassium hydroxide, which is very sensitive to light. The coating is connected to the grid of a vacuum tube, while an "electron collector" near the center of the tube is joined to the plate of the first amplifier. When no light is falling on the deposit on the inner surface of the globe, the grid acquires a negative charge, stopping the flow of electrons between filament and plate, and hence no current flows in the external circuit. The ray of light, however, causes an electron stream to flow between the coating and the collector, and since the coating is connected to the grid, the electron flow constitutes a discharged circuit, so that the grid becomes less negative. The first amplifying tube is a direct current potential amplifier, and is resistance coupled. The grid and plate connections of the amplifier are connected across a condenser which becomes discharged with the fall in the grid to plate resistance of the valve brought about by the grid potential fluctuations. A charging circuit is connected to the condenser and is controlled by a valve, the grid circuit of which operates by variations of the potential across the condenser. The charging current is fed through the plate circuit of this valve, in which a relay is connected, which working through other mechanical relays in cascades, controls the radio transmitter. Wave trains from the transmitting station after detection and amplification, are applied to the picture recorder. The recording mechanism, in order that it may be sensitive to exceedingly small currents, comprises, a small moving coil, in a magnetic field created by three electromagnets. The coil of wire, in moving in the field, as the received fluctuations are applied through its windings, operates a stylus while travelling across the surface of the paper. The stylus traverses the paper in perfect synchrony with the carriage of the transmitter, the paper being lifted each time the stylus completes a forward and backward movement across the paper. A small flashing neon lamp is used to indicate the correct speed adjustment of the driving motor.

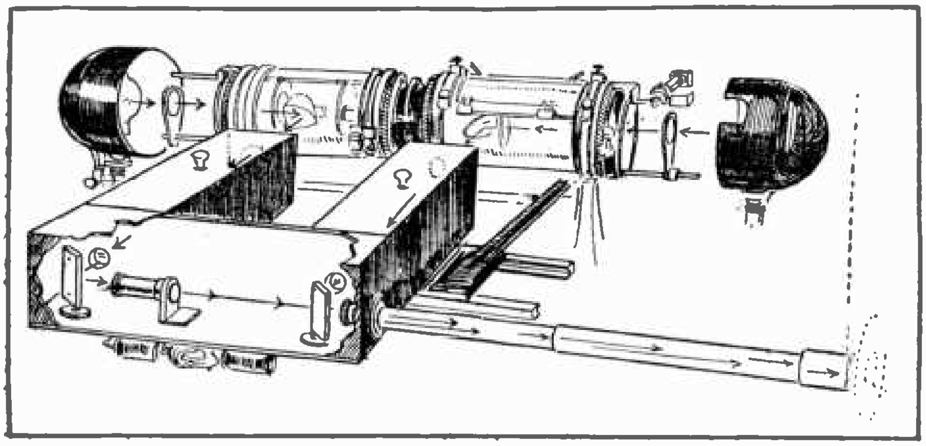

A pencil of light traverses the picture which is attached to the glass drums and is analyzed by a slow rotating action as well as a backwards and forwards movement of the carrier.

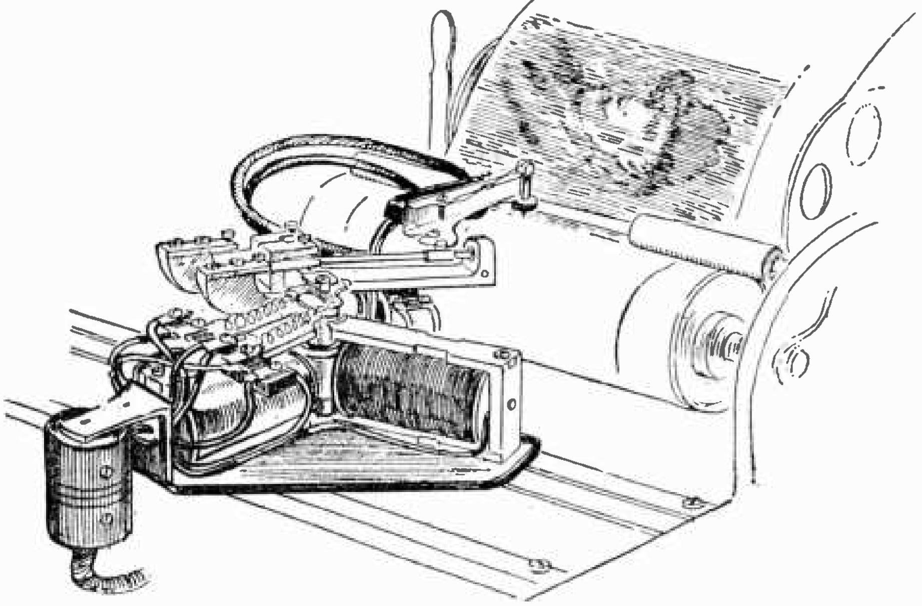

The recording mechanism of the receiver. Three electromagnets produce the magnetic field in which a moving coil controls the stylus.

TRANSMITTER—An apparatus for sending out electrical messages. As applied to radio telegraphy or radio telephony, the transmitter refers to the entire sending apparatus. The term transmitter is often used to refer to a *telephone transmitter* (q.v.). (See *Automatic Transmitter*, also *Wheatstone Transmitter.*)

TRANSMITTING AERIAL—A wire, or more usually a group of wires, suspended at a suitable height and connected to a radio transmitting set. The purpose of the aerial is to facilitate the radiation of the electromagnetic waves generated by the high frequency oscillating current which flows in the aerial. (See *Directional*, also *Receiving Aerial.*)

TRANSMITTING JIGGER—An *oscillation transformer* (q.v.) having a variable secondary, permitting of various degrees of coupling, by adjustment, between the two circuits. (See *Jigger*).

TRANSMITTING KEY—A telegraph key, used in the sending of radio code messages. This key must be of rather heavy construction since it handles larger currents than those used in ordinary wire telegraphy. Pressing the key, closes the circuit, and by holding the key down for a longer or a shorter period, the dots and dashes of the continental code are reproduced. (See *Code, Key, Key High Speed.*)

TRANSMITTING TUBE, POWER RATING—The useful power output from an oscillating tube is the power expended by the oscillating current in the resistance of the output circuit. The power input to the tube, not counting that expended in heating the filament, is the product of the plate supply voltage and the average plate current during an oscillation.

TREMBLER—A springy metallic blade carrying a soft iron armature. The spring makes contact with a fixed contact point, but when current passes through the electromagnet, the armature is attracted and the contact is broken. The spring then resumes its normal position, re-establishing the contact and the same process is repeated again and again. (See *Interrupter.*)

TRIGGER BATTERY—A term (seldom used) denoting a small battery inserted in the grid circuit to give the grid its initial charge when a tube is being used for radio transmission. In this connection, the battery is used to replace the potentiometer. (See *Grid Battery,* also *Grid Bias.*)

TRIODE—A name used to designate the three-electrode type of *thermionic tube.* Other names sometimes applied to the same device are *audion* (q.v.), *audiotron, aerotron, electron relay* (q.v.), *electron tube* (q.v.), *pliotron* (q.v.), *thermionic valve, oscillion* (q.v.), *vacuum tube* (q.v.), etc. (See *Three-Electrode Thermionic Tube.*)

TRUE POWER—The product of the *apparent power* and the *power factor* (q.v.) in an alternating current circuit. (See *Power, True,* also *Power, Apparent.*)

"T" TYPE AERIAL—See *"T" Antenna.*

TUBE OF FORCE—A theoretical conception, similar to *Line of Force* (q.v.) used in mathematical considerations of electrostatic or electromagnetic fields.

TUBE RECTIFIER—A vacuum tube or gas filled tube used to change or rectify alternating current to direct current. The principle of the vacuum tube rectifier is the fact that the electrons will flow from the hot filament to the plate, but not in the reverse direction. Tubes are made for half-wave or full-wave rectification. (See *Rectifying Tube.*)

TUBE, THREE ELEMENT—See *Triode,* also *Three-Electrode Thermionic Tube.*

TUBE, TWO ELEMENT—See *Fleming Valve.*

TUBING—In general, this refers to the bakelite or hard rubber forms used to wind inductance coils on. Another type of radio tubing is defined under *Spaghetti.*

TUDOR ACCUMULATOR—A storage battery having special positive plates formed by the *Plante* process and negative plates of the *Faure* or pasted type. The Tudor positive is of the cast-lead type. This plate has no central web and is made by casting pure soft lead in a mold. Casting is advantageous since it permits the distribution of the metal in the plate without limitations in the manufacturing process. After the plate is removed from the mold, its surface consists of a large number of short vertical ribs which run entirely through the plate. These are bound together by short transverse ribs for the purpose of giving strength to the plate. This gives a large surface and since there is no central web, the electrolyte is enabled to circulate through the plate thus working the active material uniformly.

TUNED ANTENNA—A transmitting aerial, having a loading inductance or a variable condenser or both in series with it, for the purpose of changing the antenna wave length.

TUNED IMPEDANCE—This refers especially to a circuit, or to a method of radio frequency amplification in which a tuned inductance connected in parallel with a variable condenser is used as the method of coupling rather than transformer or resistance coupling. (See *Impedance Coupled Amplifier.*)

TUNED RADIO FREQUENCY—See *Tuned Radio Frequency Circuit.*

TUNED RADIO FREQUENCY AMPLIFIER—A combination of a *radio frequency transformer,* a variable condenser and an amplifying vacuum tube in a radio receiving circuit for the purpose of increasing the intensity or amplifying the radio frequency signals.

TUNED RADIO FREQUENCY CIRCUIT—A radio receiving circuit employing radio frequency amplification in which the radio frequency amplifier circuits may be tuned to the desired wave lengths by varying the inductance or the capacity, or both although the usual method of tuning is by means of a variable condenser in parallel with the secondary of the radio frequency transformer. A typical five-tube tuned radio frequency circuit is shown in the accompanying illustration. This consists of two stages of tuned radio frequency amplification, a detector and two stages of audio frequency amplification.

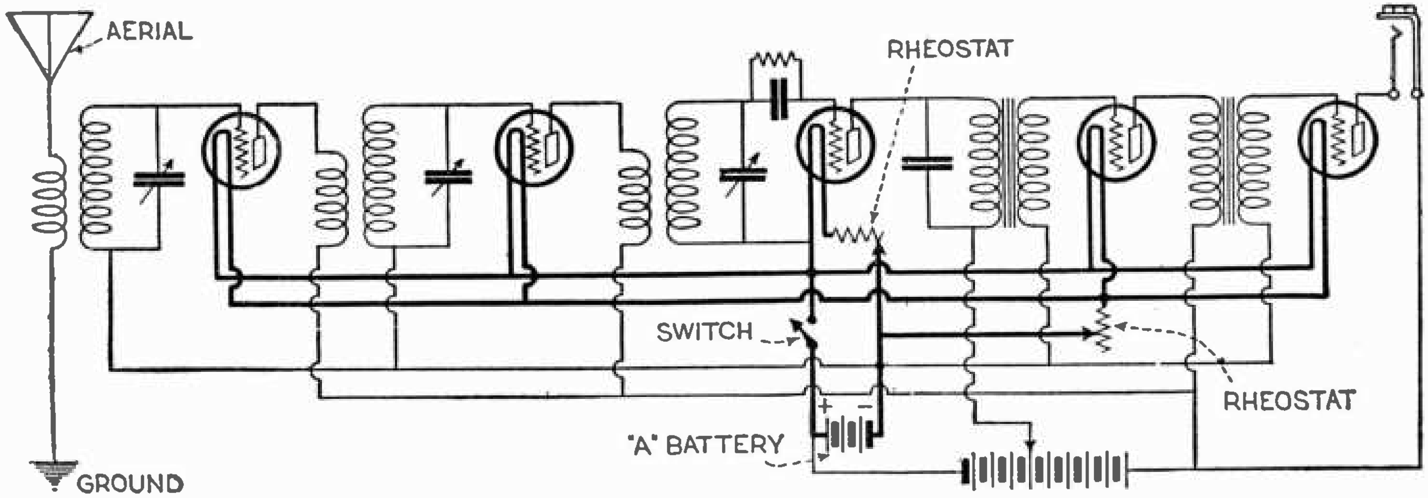

A standard five tube tuned radio frequency circuit.

TUNED RADIO FREQUENCY TRANSFORMERS—An air core *transformer* (q.v.) shunted by a variable condenser used in a *tuned radio frequency circuit* (q.v.). (See *Fixed Coils,* also *Low Loss Coils.*)

TUNED TELEPHONE—One in which the diaphragm is adjusted to vibrate at the same frequency as the current impulses to be indicated. In addition the telephone circuit may have its inductance and capacitance chosen to have the same electrical frequency, and a tuned acoustic resonator may intervene between the diaphragm and the ear.

TUNED TRANSFORMER—See *Resonance Transformer.*

TUNGAR—A trade name for a two-electrode vacuum tube especially designed for rectifying purposes. Tungar tubes are made in two ampere and five ampere capacities. They are used in *chargers* (q.v.) and also in radio sets operating directly from the lighting socket without the use of batteries. (See *Rectifying Tube.*)

A typical rectifier tube of the Tungar type. The screw-base allows it to be inserted in a standard lamp socket from which current is obtained for the filament. Connection is made to the graphite disc by means of a Fahnestock connector which clips on to the wire post projecting from the other end of the tube.

TUNING—The process of regulating the inductance and the capacity of a radio circuit in order to be in unison with a desired wave length.

TUNING, BROAD—Reception of a signal over a wide range, that is on a number of wave lengths, rather than on a single one. A receiving set which tunes broadly cannot be used where the broadcasting stations are transmitting on wave lengths near to one another. Too long an aerial may result in a receiving set tuning broadly. (See *Selectivity.*)

TUNING COIL—A coil of wire so arranged that each turn is electrically insulated from its neighbor and with a device by which contact may be made to bring any desired number of turns into the circuit so as to increase or decrease the inductance of a radio circuit so as to tune it to the desired wave length. A variable *inductance* (q.v.) used for tuning an oscillatory circuit. (See *Lengthening Coil.*)

T.Y.K. ARC—A transmitting arc used for radio telephony invented by W. Torikata, E. Yokoyama and M. Kitamura of Japan. This arc uses magnetite and brass electrodes instead of copper-carbon electrodes. A circuit diagram of the T.Y.K. system is shown in the illustration. The materials used for the electrodes are such that a high resistance film forms on their surfaces. This requires a temporary high voltage to start the discharge. This is accomplished as follows: An armature, A, is attached to one of the electrodes. The two electrodes are in contact, and a steady current is caused to flow in the circuit, through the spark induction coil B. This attracts the armature A, drawing the electrodes apart. The break of the spark-coil current at its interrupter induces a high electromotive force which, acting through the coil and the spark gap in series, breaks down the film on the gap electrodes. The power supplied to the gap is 500 volts and 0.2 ampere. A condenser of approximately .05 mfd. is used in the primary oscillating circuit.

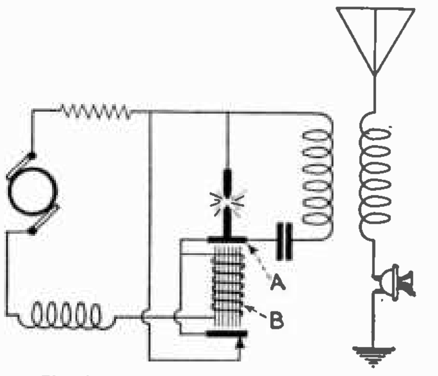

Circuit diagram of the T. Y. K. arc.

T.Y.K. sets have been used only in small units, having a range of 30 or 40 miles.

Several unusual features appear in the circuit shown in the accompanying illustration, which gives the circuit diagram of the T.Y.K. arc.

U

ULTRA-AUDION—A name sometimes applied to the vacuum tube used in connection with *beat* reception for supplying local oscillations. In general, however, the name ultra-audion applies to a circuit for radio reception. (See *Ultra-Audion Circuit.*)

ULTRA-AUDION CIRCUIT—A type of circuit used for long wave radio reception which uses a form of regeneration, without calling for the introduction of auxiliary equipment in the circuit. A typical ultra-audion circuit is illustrated. It will be noticed that while one terminal of the secondary goes to the grid, the other terminal, which usually goes to the filament, in this case goes to the plate. Thus there is a direct plate-filament circuit which does not involve the plate and filament batteries in the usual manner.

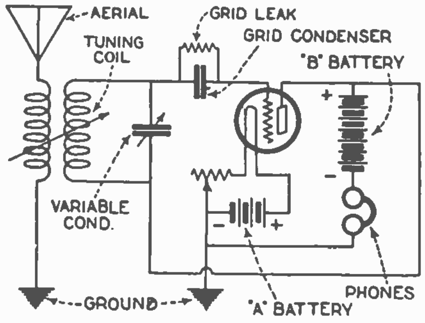

Diagram of an ultra-audion regenerative circuit.

ULTRADYNE—A modification of the *super-heterodyne circuit* (q.v.), which uses the modulation method to produce beats. In this method the incoming signal is caused to modulate the oscillations produced locally in a similar way to that in which the speech or music modulate the carrier wave of a broadcast station. It is claimed that this method is simpler than the ordinary super-heterodyne and also more sensitive to weak signals.

The circuit of the improved model L-2 Ultradyne, incorporating regeneration is shown below. The long wave intermediate transformers are accurately tuned to a wave length of 3000 meters. Referring to the diagram, it will be noted that a special form of oscillator, comprising a grid and plate coil, are used together with an oscillator tuning condenser connected across the grid coil. The first tube, which is usually the detector in the standard super-heterodyne circuit, is known as the modulator tube. (See *Supcr-Heterodyne Receiver.*)

UMBRELLA ANTENNA—An antenna, the conductors of which form elements of a cone with the apex at the top to which the *down lead* (q.v.) is connected.

UNDAMPED—See *Undamped Alternating Current*, also *Undamped Waves.*

UNDAMPED ALTERNATING CURRENT—Periodic current (i.e., current passing through successive equal cycles of values) whose average value is zero.

UNDAMPED OSCILLATIONS—Oscillations which are sustained. *Undamped alternating current* (q.v.). Oscillations such as are generated by a *vacuum tube oscillator* or by an *arc generator.*

UNDAMPED WAVES—Continuous waves. Waves in which similar current cycles follow each other continuously instead of being broken up into groups. (See *Sustained Waves, Decadent Wave,* also *Damped Waves.)*

UNITS—Specified amounts of physical quantities used as a basis of measurement. (See *Practical Units.)*

UNITS, ELECTROMAGNETIC—See *Electromagnetic Units.*

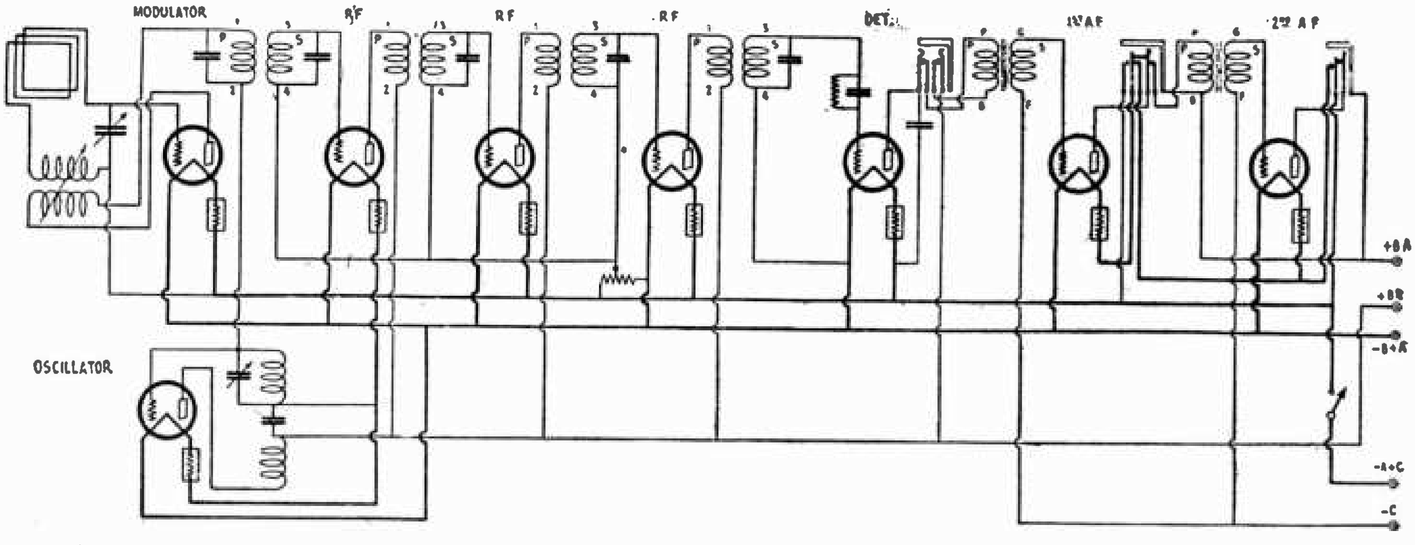

This shows the hook-up of the Ultradyne super-het. This set is not radically different from the average super-heterodyne. The heteroformers are of the tuned air core type and the filter is shown at the input side of the circuit. Aerial and ground as well as loop may be used with this set and instead of the first detector the inventor uses a tube known as the "modulator."

V

VACUUM TUBE—An evacuated bulb, or one containing a rare gas and having three elements, filament, plate and grid. (See *Three-Electrode Thermionic Tube, Triode,* also *Dynatron.*)

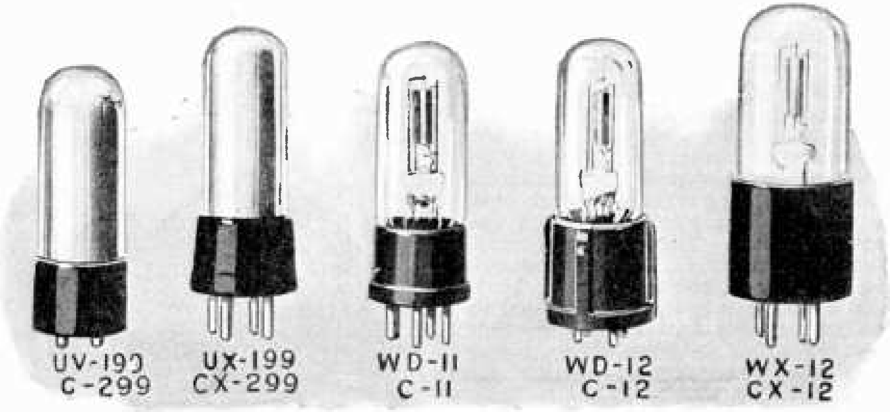

A group of dry-cell vacuum tubes. The first two differ only as to the type of base; both are intended to be operated from a 4½-volt "A" battery, that is, three dry-cells connected in series. The filaments draw very little current, as can be determined from the chart. The last three tubes also have identical characteristics, but are designed to fit different types of sockets. These tubes operate from a single 1½-volt dry cell and draw approximately ¼ ampere of current. The filaments are oxide-coated and very rugged.

VACUUM TUBE AMPLIFIER, THEORY OF—The vacuum tube may be used as an amplifier or magnifier because of the fact that under proper conditions, a small variation of the electromotive force applied between the filament and the grid produces a very large variation in the current flowing in the plate circuit. It is usual to use several amplifying tubes together, suitably coupling the plate of one tube to the grid of the next, thus increasing the amplification to the desired value.

VACUUM TUBE CHARACTERISTICS Performance curves of vacuum tubes. These are obtained experimentally by keeping the filament current constant and applying various known voltages between the plate and the filament and between the grid and the filament and reading the resultant currents that flow to the plate and the grid. The various values obtained are then plotted on cross-section or graph paper. Such curves are convenient for determining the operating characteristics of a particular vacuum tube under consideration. (See *Characteristic Curve.*)

VACUUM TUBE GENERATOR—The vacuum tube can be used to produce sustained oscillations without the necessity of supplying potential variations to the grid from an external source. This can be accomplished as a direct result of the amplification properties of the vacuum tube, since the energy of the output circuit is greater than the energy of the input circuit and part of this output energy may be returned to the input to produce constant reamplification.

VACUUM TUBE, POWER—A tube of special design, made for use in the last audio stage of a radio receiving set, which will handle greater current than the ordinary amplifying tube, thus resulting in more volume and at the same time clearer and more satisfactory tone reproduction. Power tubes now on the market require higher plate voltages than conventional tubes, ranging from 135 volts upwards. The amount of grid bias or "C" battery ranges from approximately 4½ volts to 45 volts, depending upon the type of power tube. Tubes suitable for developing considerable power, in radio transmission are also referred to as power tubes. By increasing the surface area and hence the filament emission or by increasing the degree of evacuation of the bulb, thus allowing a higher plate voltage, the output of the tube may be increased. A form of transmitting power tube is described under the heading *Pliotron.* The illustration below shows one of the new power tubes used for reception.

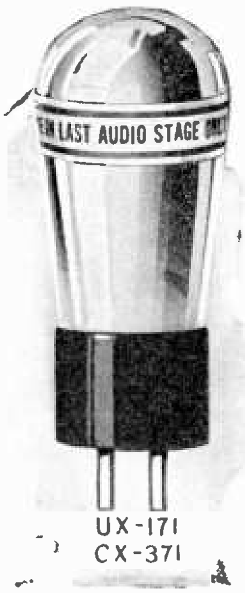

A new storage-battery type power amplifier tube, for use in the last stage audio amplifier only. This tube takes the same filament voltage and filament current as the No. 112 tube but employs a "B" voltage of 180 and 40½-volt "C" battery. Special means should be provided for handling its heavy output.

VACUUM TUBE RECTIFIER—A vacuum tube designed or used to change alternating current to direct current. (See *Rectifying Tube, Kenotron,* also *Tube Rectifier.*)

VACUUM TUBE, THEORY OF OPERATION—See *Theory of Operation of Vacuum Tubes.*

CHART OF VACUUM TUBE TYPES AND CHARACTERISTICS

TYPE	USE	"A" battery volts (Supply)	Filament Terminal Volts	"A" Battery Current Amperes	"B" Battery Volts		Negative "C" Battery Volts	Voltage Amplification Factor	Output Resistance Ohms
					Det.	Amp.			
UV. UX-199 C. CX-299	Det. or Amp.	4.5	3	.06	45	90	4.5	6.25	15,000
UV. UX-200 C. CX-300	Det. only	6	5	1.0	16 to 22½				
UX-200-A CX300-A	Det. only	6	5	.25	Max. 45				
UV. UX.201-A C. CX. 201-A	Det. or Amp.	6	5	.25	45	90 to 135	4.5 9.0	8 8	12,000 11,000
UX-120 CX-220	Pow. Amp. (Lst. stg. only)	4.5	3	.125		135	22.5	3.3	6,600
UX-112 CX-112	Det. or Amp.	6	5	.5	22½ to 45	135	9.0	Var.	Var.
UX-171 CX-371	Pow. Amp. (Lst. stg. only)	6	5	.5		180	40.5		
UX-210 CX-310	Pow. Amp. Oscillator	6	6	1.1		90 to 425	4.5 to 35	Var.	Var.
WD-11 C-11	Det. or Amp.	1.5	1.1	.25	22½	90	4.5	5.6	14,000
WD. WX-12 WX. CX-12	Det. or Amp.	1.5	1.1	.25	22½	90	4.5	5.6	14,000
3VB-199 3VBX-199	Det. or Amp.	4.5	3	.06	20	80	4.5	6.0	
3V-A 3VAX	Det. or Amp.	4.5	3	.12	20	90	4.5	6.5	
5V-A 5VAX	Det. or Amp.	6	5	.25	20	100	4.5 to 9.0	9.4	9,400
5VC 5VX	Pow. Amp. or Det.	6	5	.5	22½	90 to 157½	6 to 10.5	8.6	5,900
99 99X	Det. or Amp.	4.5	3	.06	22½	90 to 150	3 to 12		
O1A O1X	Det. or Amp.	6	5	.25	22½	90 to 150	3 to 12		
MU-20	Audio Amp.	6	6	.25		90 to 150	4.5 to 10.5	20	40,000
MU-6	Pow. Amp. (Lst. stg. only)	6	6	.5		90 to 150	4.5 to 10.5	6	5,000
B-6	Det. only	6	5	.25	16 to 22½				
A	Det. or Amp.	6	5	.25	20	120	4.5 to 9		
B C	Det. or Amp.	4.5	3	.06	20	80	4.5		
E	Pow. Amp. (Lst. stg. only)	4.5	3	.125		135	22.5		
F	Pow. Amp. (Lst. stg. only)	6	5	.5		90 to 180	4.5 to 9		
G	Audio Amp.	6	5	.25		90 to 180	4.5 to 9		
DC—	Det. or Amp.	4.5	3	.06	45	90	4.5	6.3	16,500
DC—	Det. or Amp.	6	5	.25	45	90	4.5	8.5	10,000
DC—	Pow. Amp. (Lst. stg. only)	4.5	3	.125		112 to 135	13 to 22.5	3.3	6,300
DC—	Pow. Amp. (Lst. stg. only)	6	5	.5		90 to 157.5	6 to 10.5	8.0	8,500

VACUUM TUBES, TYPES OF—Vacuum tubes are of various types and designs depending upon the work which they must do. They may be divided into two general classes, receiving tubes and transmitting tubes.

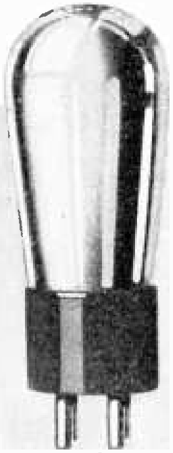

A new type of detector tube which has greater sensitivity than former types, due to the use of an alkaline vapor. It can be used as a detector in any type of receiving circuit.

Tubes for receiving purposes are designed either for dry cell operation or else are of the storage battery type.

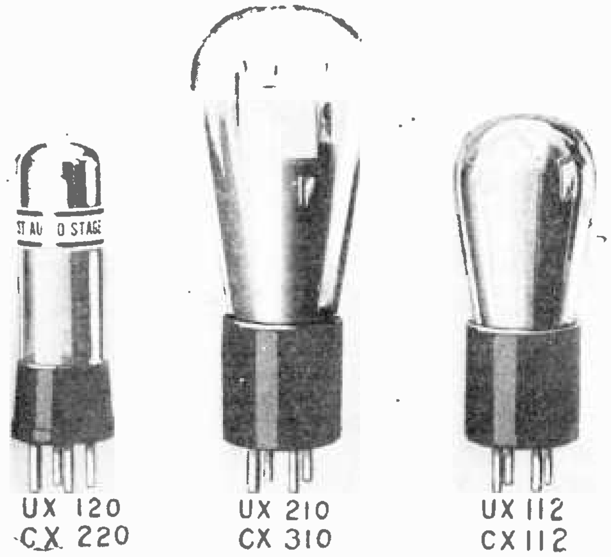

These power amplifier tubes are designed for use in the last stage of an audio frequency amplifier. On the left is a dry cell tube which requires a 4½-volt "A" battery. The center tube employs exceptionally high "B" battery voltage, and is usually employed with an amplifier that operates from the house current. This tube can also be used for transmitting purposes. The right-hand tube is of the storage battery type and has a very rugged filament. It makes an excellent detector tube as well as a power amplifier.

They may be further classified as detector tubes, amplifying tubes or power tubes. There are many special types of tubes, such as *kenotrons* (q.v.), *pliotrons* (q.v.), etc.

VALENCY—The property possessed by elements in combining with or replacing other elements in a certain definite proportion. Also referred to as *valence*. (See *Electro-Chemical Equivalent.*)

VALVE—This generally refers to a vacuum tube used as a detector or rectifier. The name *valve* is used in many foreign countries to refer to the thermionic vacuum tube, as used for any purpose.

VALVE DETECTOR—See *Valve.*

VARIABLE CONDENSER—A condenser having rotable or movable metal plates and an air dielectric. Both the stationary and rotating plates may be made of brass, copper or aluminum. Bakelite, fibre or composition are used to insulate the stationary from the movable plates. Turning the movable plates either increases or decreases the capacity of the condenser. Variable condensers are primarily used in radio for *tuning* (q.v.). They are constructed so as to have straight line frequency characteristics either by undercutting a portion of the movable plates or else by making these of a variable thickness.

VARIABLE GRID LEAK—A variable high resistance unit designed to be placed in the grid circuit of a vacuum tube used as a detector. Certain types of variable grid leaks may have their resistance varied from ¼ megohm to 10 or more megohms by merely tightening a thumb-screw. (See *Grid Leak*, also *Adjustable Grid Leak.*)

VARIABLE RATIO TRANSFORMER—A transformer in which the ratio between the primary and the secondary windings can be varied, usually by means of suitably located taps.

VARIABLE RESISTANCE—Both the *rheostat* and the potentiometer used in radio work, are variable resistances. Variable resistances are usually of high resistance wire type, with a sliding contact so that the resistance can be easily altered.

VARIO COUPLER—A tuner formerly used in radio receiving sets, having a primary and a secondary coil inductively coupled, with the secondary arranged so as to be rotable within the primary. The amount of coupling depends upon the variation of the angle between the axes of the two coils.

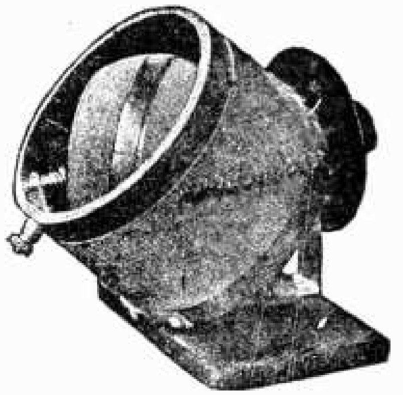

A standard form of vario coupler.

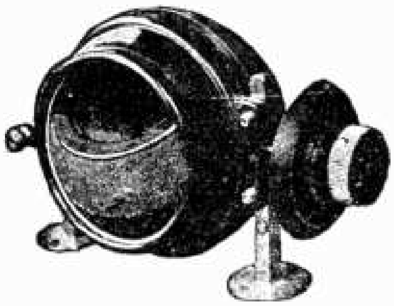

A variometer.

VARIOMETER—A tuner, similar to the vario-coupler, except that the primary and secondary coils are electrically connected.

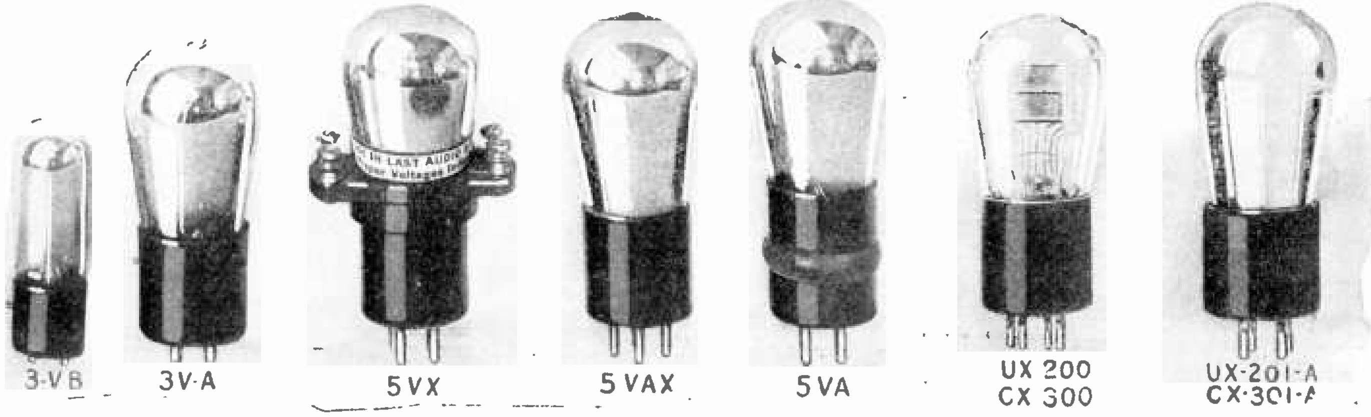

The characteristics of these vacuum tubes are given in the vacuum tube chart on page 187. The first tube is of the dry-cell type, requiring a 4½-volt "A" battery for lighting the filament. 3V-A is a standard 5-volt tube. 5VX is a power amplifier which can be used in the last stage of any audio-frequency amplifier. There is no necessity of changing any of the wiring in the set for the addition of more "B" battery, as binding posts are included on the tube itself. 5VA has a sponge rubber ring included as a part of the base which tends to absorb all vibrations which might otherwise cause the tube to become noisy. The next type of tube is for use as a detector only. The last is the well-known 201-A type, which can be employed as a detector or an amplifier.

VECTOR—A graphical illustration used in mathematical calculations, consisting of a line with an arrow-head at one end, used to show by means of its length, direction and the angle between itself and another vector or vectors, the magnitude, direction and phase angle of alternating current quantities.

VELOCITY—The distance passed through in a certain time. Velocity is measured in feet per second, miles per hour, etc.

VIBRATION—A to and fro motion. An oscillating or swinging motion.

VITREOUS—Consisting of or pertaining to glass. (See *Resinous Electricity.*)

VOLT—The practical unit of *electromotive force* (q.v.). The volt is a measure of the electromotive force which will cause a current of one ampere to flow through a resistance of one ohm. It is equal to 10^8 absolute electromagnetic units (Abvolts).

VOLTAGE—The electrical pressure or the electromotive force between two points in an electrical circuit, measured in volts.

VOLTAGE, CONSTANT—A steady unvarying electrical pressure or electromotive force, as differentiated from a pulsating or fluctuating voltage.

VOLTAGE DROP—The fall in potential caused by the resistance of the conductor through which the current is flowing. The longer the conductor, the greater will be the voltage drop. The greater the cross-section of the conductor, the less will be the voltage drop.

VOLT-AMPERES—The product of volts as measured with a voltmeter, by the current as shown on an ammeter, in an alternating current circuit, gives the apparent power, or in other words the apparent watts. (See *Instantaneous Values.*)

VOLTMETER—An instrument for measuring voltage drop or potential difference. Most of the voltmeters used in radio work operate on the moving coil principle. (See *Galvanometer.*)

A plug-in type of voltmeter.

Panel-type of voltmeter.

VOLTMETER CALIBRATION—A method of correctly marking the voltage readings on the scale of a voltmeter or of checking up the voltmeter readings to make sure that they are accurate. The usual method of calibration is to check the voltmeter reading by comparison with a standard voltmeter. (See *Calibration.*)

V.T.—Abbreviation for *vacuum tube.*

VULCANITE — (EBONITE) — A hard black substance produced by vulcanizing rubber with about 25 per cent sulphur. It is readily polished and has excellent insulating properties.

VULCANIZED RUBBER—Pure rubber, mixed with five per cent sulphur and baked at a temperature of about 150 degrees Centigrade. Vulcanized rubber must not be allowed to come into contact with copper, which is chemically attacked by the sulphur.

W

WALL TUBE—See *Partition Insulator.*

WATCHCASE RECEIVER—A telephone receiver having a shape somewhat similar to a watch. Watch case receivers are compact in construction and are the usual type used in the radio telephone *head set.*

WATT—The unit of electric power. One *Joule* (q.v.) per second. To find power in a direct current circuit, multiply voltage by amperage. The *Kilowatt* (q.v.) equals one thousand watts. Seven hundred and forty-six watts are equivalent to one electrical horsepower.

WATT HOUR—The commercial unit of electrical work. The work done in one hour by a current of one ampere flowing between two points of a conductor having a difference of potential of one volt.

WATT-HOUR METER—An integrating meter which measures energy in watt-hours or in kilowatt-hours. (See *Integrating Wattmeter.*)

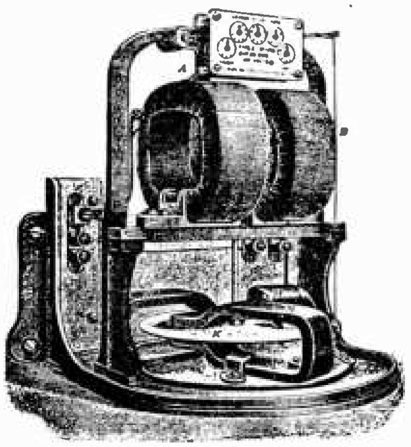

A direct current watt-hour meter.

WATT, INTERNATIONAL—symbol W—The energy expended per second by an unvarying electric current of one International ampere under an electrical pressure of one International volt. $IW = 1$ Joule per second $= 10^7$ ergs per second.

WATTLESS COMPONENT—WATTLESS CURRENT—The component of an alternating current which is in quadrature with the voltage. (See *Reactive Component.*)

WATTMETER—An instrument designed to measure the power being expended in an electrical circuit. The most common form of direct current wattmeter is of the dynamometer type. This has two coils, one fixed and one movable, one coil being connected so as to exert a force proportional to the current flowing in the circuit, and the other coil exerting a force proportional to the electromotive force. The *induction* type of wattmeter is in general use for alternating current circuits. (See *Electro-Dynamometer, Dynamometer,* also *Induction Wattmeter.*)

WATT SECOND — A unit of electrical energy, representing the energy expended by one watt flowing for a second. The watt second is the same as the *Joule* (q.v.) and is usually applied to the measurement of heat developed by an electric current.

WAVE — A periodic alternation of an alternating current. As applied to an electric or electromagnetic disturbance, an electric wave is an undulatory movement of the ether, radiated from conductors carrying electrical oscillations.

WAVE ANALYSIS — A study of the wave form of an alternating current or other type of wave. The wave form of commercial alternating current used for lighting purposes approximates a *sine curve* (q.v.). A convenient instrument for the analysis and study of alternating current wave form is known as the *oscillograph* (q.v.).

WAVE ANTENNA — A horizontal antenna, the physical length of which is approximately equal to the length of signaling waves to be received, and which is so used as to be strongly directional.

WAVE DISTORTION—Alteration in the wave form of a wave after it has traversed a considerable distance. (See *Distortion.*)

WAVE FORM—Referring to an alternator, the "shape" of the curve of the current generated.

WAVE FREQUENCY—See *Frequency.*

WAVE LENGTH—symbol λ (lambda)—The distance between two successive *antinodes* (q.v.) in the same direction. In referring to the wave length of electrical oscillations in a circuit, this means the length of the waves in free space that would have a frequency corresponding to the given oscillations. Electromagnetic waves used in radio work have frequencies of from 10,000 to 3,000,000 cycles per second. Since these waves have a velocity of approximately 300,000,000 meters per second, it is possible to calculate the length of a wave by dividing the velocity by the frequency. In other words meters per second divided by cycles per second gives length of the wave in meters. Thus a wave having a frequency of 100 kilocycles (100,000 cycles) will have a wave length of 300,000,000 divided by 100,000 or 3,000 meters.

WAVE LENGTH ALLOCATIONS — In order to prevent radio transmitting stations from interfering with each other, or creating interference at a receiving station, various wave length bands have been allotted to each type of service. The following table shows the present short wave assignments and the service for which they are used:

Wave-length in Meters	*Service*
109–105	Relay broadcasting only.
105–85.7	Public toll service, Government mobile, and point-to-point communication by electric power supply utilities, and point-to-point and multiple address message service by press organizations only.
85.7 –75.0	Amateur, Army mobile, naval aircraft, and naval vessels working aircraft only.
75.0 –66.3	Public toll service, mobile Government point-to-point and point-to-point public utilities.
66.3 –60.0	Relay broadcasting only.
60.0 –54.5	Public toll service only.
54.5 –52.6	Relay broadcasting only.
52.6 –42.8	Point-to-point only.
42.8 –37.5	Amateur and Army mobile only.
37.5 –33.1	Public toll service, mobile, Government point-to-point, and point-to-point public service utilities.
33.1 –30.0	Relay broadcasting only.
30.0 –27.3	Public toll service only.
27.3 –26.3	Relay broadcasting only.
26.3 –21.4	Public service, mobile, and Government point-to-point.
21.4 –18.7	Amateur only.
18.7 –16.6	Public toll service, mobile and Government point-to-point.
16.6 – 5.35	Experimental.
5.35 – 4.69	Amateur.
4.69 – 0.7496	Experimental.
0.7496– 0.7477	Amateur.

Radio broadcast stations work within the wave length band of 200 to 545 meters at intervals of 10 kilocycles. 600 meters is reserved for distress signals. The higher wave lengths are allocated to various marine, government, university, aircraft and other uses.

WAVE LENGTH CALCULATION FOR ANTENNA—For the ordinary vertical wire grounded antenna, the fundamental wave length is slightly greater than four times the length of the wire. A constant suggested is 4.2 and applies approximately also to flat top antennae having vertical lead ins. For other calculations see *Fundamental Wave Length.*

WAVE LENGTH, FUNDAMENTAL—See *Fundamental Wave Length.*

WAVE LENGTH, NATURAL—In a loaded antenna (that is with series inductance or capacity) the natural wave length corresponds to the lowest free oscillation. (See *Natural Wave Length.*)

WAVE LENGTH OF STATIONARY S. H. M. WAVES—On a straight wire, the smallest distance between two points where the disturbance is of the same amplitude and phase; or, since consecutive loops (q.v.) are in opposite phases, the wave length is double the distance between consecutive loops or consecutive *nodes* (q.v.). (See *S.H.M.*)

WAVEMETER—A radio instrument for measuring frequency. A calibrated resonator of variable frequency, capable of easy excitation by the oscillation under test, combined with a means of indicating the attainment of resonance. The calibration may be given directly in wave lengths or frequencies or both.

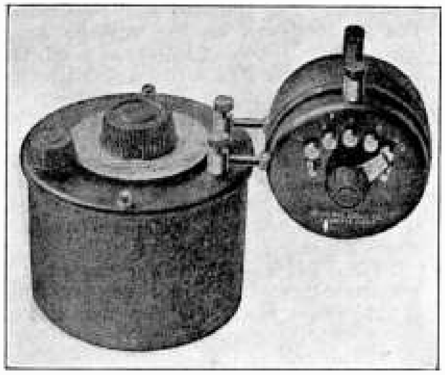

Photo by courtesy of General Radio Co.

The selectivity of the radio set may be greatly increased by the use of this wavemeter as a radio filter.

WAVEMETER CALIBRATION—Wavemeters may be calibrated by comparison with a standard wavemeter. An accurate method of calibrating a wavemeter, known as the *Three-Way Method* is shown in the accompanying diagram. In this illustration A is the standard wavemeter, B the wavemeter to be calibrated, and C is a condenser and coil of size permitting resonance with A. To calibrate, the buzzer should be put in operation, varying the capacity of the condenser C^1 until L^1C^1 is in resonance with L^2C^2 as noted in the head set. The wavemeter B should then be put in inductive relation to C and the condenser C^2 varied until the wavemeter A resonates with C. The wave length of B is now the same as A. This procedure should be carried out over the whole range of the condenser C^3.

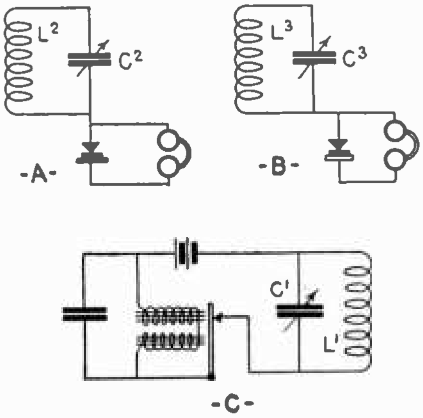

Diagram showing three-way method of wavemeter calibration.

WAVE MOTION—A disturbance of the equilibrium of a medium or body, extended or propagated from point to point with a continuous motion, each particle vibrating only on each side of its position of equilibrium, while each phase of vibration moves onward. Examples of wave motion are the propagation of electromagnetic waves, or the waves on the surface of a body of water.

WAVES, CONTINUOUS—See *Continuous Waves.*

WAVES, CONTINUOUS, KEY MODULATED—See *Continuous Waves, Key Modulated.*

WAVES, CONTINUOUS, MODULATED AT AUDIO FREQUENCY—See *Continuous Waves at Audio Frequency.*

WAVES, ELECTROMAGNETIC—Periodic electromagnetic disturbances progressive through space. (See *Wave.*)

WAVES, SUSTAINED—See *Sustained Waves*, also *Continuous Waves.*

WAVE TRAIN—A series of waves. An example of a wave train is the series of disturbances set up by each spark in a spark system of radio telegraphy.

WAVE TRAP—A device consisting of a variable condenser and coil, usually connected in the antenna circuit of a receiving set, to eliminate interference from undesired stations. (See *Filter, Filter Condenser, Acceptor*, also *Rejector.*)

WAVE-WOUND—A class of drum armature in which connections produce a "stepping forward" in a zig-zag wavy line all the time. Also called *Series* and *Two Circuit Windings.*

WEBER—The *Practical Unit* (q.v.) of magnetic flux. The engineers' unit is the "line" which is also the electromagnetic unit. 1 Weber is equal to 108 electromagnetic units (or "lines"). The Weber is produced by a current of one ampere flowing through a circuit having one henry of inductance.

WEHNELT BREAK—An *electrolytic interrupter* (q.v.).

WEIGHT—The force with which the earth attracts a body. A body varies in weight according to its distance from the centre of the earth, although its mass remains constant.

WESTON CELL—The present standard cell. It consists of mercury with a paste of mercurous and cadmium sulphates which form the cathode. The Weston cell has an anode of 12.5 per cent cadmium amalgan in an electrolyte of saturated solution of cadmium sulphate. It has a constant electromotive force of 1.0125 volts at 20 degrees Centigrade. (See *Electromotive Force*, also *Cadmium.*)

WHEATSTONE BRIDGE—An instrument for measuring resistance in which the current from a battery divides into two parallel circuits, each

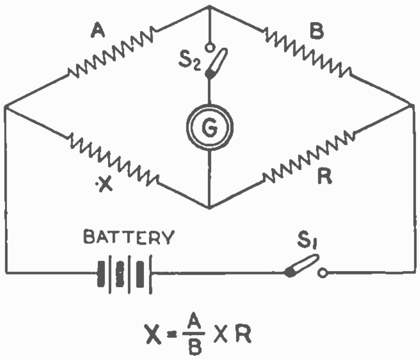

Diagram illustrating principle of the Wheatstone Bridge.

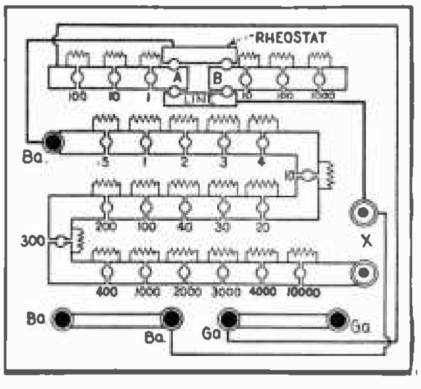

Post-Office type of Wheatstone Bridge.

divided into two "arms" capable of being adjusted so that there is no difference of potential between the dividing points on both sides. The point of zero difference of potential is indicated by the absence of deflection of a galvanometer placed across the dividing points. The ratio of the resistance of the two parts of the arms is then the same, so that if the resistance of three of the arms is known, that of the fourth arm (i.e. the resistance to be measured) can be determined.

WHEATSTONE TRANSMITTER—An apparatus for delivering telegraphic currents to a line at high speed. Control is effected by aid of a moving perforated paper tape, prepared according to a code.

WHIPPING—The binding of string or small wire round the end of a rope or multiple wire to prevent the ends from fraying out.

WIMSHURST—A type of Induction or Influence Machine consisting, in a simple form, of two discs of insulating material revolving in opposite directions. These carry a number of equal sectors which form combined inductors and carriers. There are usually two collectors and two pair of brushes. (See *Induction Machine.*)

WINDOW LEAD-IN—A form of insulator for passing an aerial *lead-in* through a window. In some cases porcelain tubes are used as window lead-in's. A common form of window lead-in consists of a flat copper conductor covered with woven cotton insulation.

WIRED WIRELESS—A method of transmitting radio messages, in which the waves are guided by wires instead of radiating freely through the ether. (See *Line Radio.*)

WIRE GAUGE—A device for measuring the diameter of a round wire in accordance with a predetermined scale. The gauge commonly used in this country is the Brown and Sharpe (B. & S.) or American gauge.

WIRELESS TELEGRAPHY—Also known as **RADIO TELEGRAPHY** (q.v.).—A system of telegraphy, utilizing electromagnetic waves set up by oscillating currents as the means of transmission. The waves generated at the transmitting station spread out from the *aerial* (q.v.) in all directions. They are intercepted by the receiving aerial, detected by a rectifying device such as a crystal or vacuum tube, amplified and delivered to a head set or loud speaker. A key is used at the transmitter to control the duration of the waves so that they will correspond to the dots and dashes of the *continental code* and these dots and dashes are reproduced in the phones at the receiving station.

WOLLASTON WIRE—Platinum wire, drawn extremely fine by the process of coating fine platinum wire with ductile material, drawing the whole down together, and (if required) removing the coating by dissolving it off.

WORK—The action of a force upon a body to overcome resistance. Work is measured by the product of the force exerted and the distance moved. Work may be measured by *ergs* (q.v.), *foot-pounds, horsepower*, etc.

WOOD'S METAL—A soft metallic alloy having a low melting point. This alloy consists of two parts of lead, one part of tin, four parts of bismuth, and one part of cadmium, all by weight. This metal has been used extensively for setting detector crystals in their cups, so as to make a good electrical contact.

X, Y, Z

X—Symbol for *Reactance* (q.v.).

"X" RAYS—Also called *Roentgen Rays.* Electric waves of much higher frequency than light, produced by the striking of *cathode rays* upon a solid substance. These rays are not deflected by electric or magnetic fields and possess the property of penetrating solid substances which could not be penetrated by ordinary light rays. The use of the "X" Rays permit photographs to be taken of bones or other parts of the human body which could not be seen in any other way.

X'S—Disturbances of an erratic nature heard from time to time in the 'phones or loud speaker of a radio receiving set, due to storms, electric disturbances, etc. (See *Atmospherics, Natural Electric Waves, Static,* also *Strays.*)

YAGI SPARK GAP—A form of spark gap originated by H. Yagi. It is a quenched spark gap, the electrodes of which are aluminum and brass. The gap functions in an atmosphere of coal gas.

"Y" GROUPING—See *Star Grouping.*

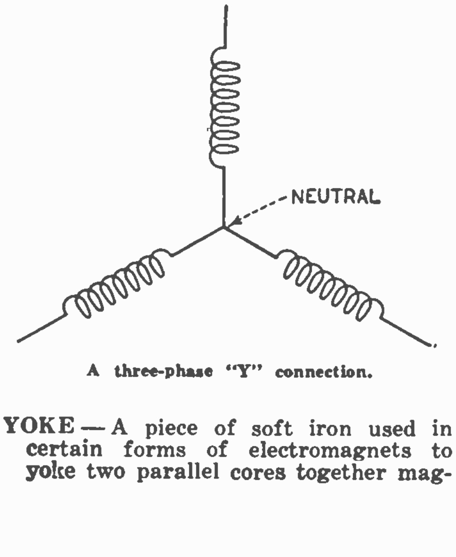

A three-phase "Y" connection.

YOKE—A piece of soft iron used in certain forms of electromagnets to yoke two parallel cores together magnetically. The cores of the ordinary electric bell are fastened together at one end by means of a yoke.

YOKOJAMA, EITARO—Japanese radio expert. He was born in 1883 and was educated at the Engineering College of the Tokyo Imperial University. While at college, he specialized in radio. He was appointed to the Electro-technical Laboratory of the Japanese Ministry of Communications, to carry out research work in wireless telegraphy and telephony. He was one of the inventors of the T.Y.K. oscillation gaps for radio telephony, for which he received many distinctions. In 1910 he was appointed head of the Radio Section. Yokojama, who is one of the most brilliant Japanese radio experts, is a member of the Institute of Radio Engineers and of many other scientific societies.

ZEEMAN EFFECT—Doubling of the spectrum lines of light sources when placed in a strong magnetic field.

ZENNECK, J.—German wireless expert. He was born April 15th, 1871, at Wurtemberg and was educated at Tuebingen. In 1895 he was appointed assistant in the Physical Institute in Strassburg, a post he held until 1899, when he carried out a series of tests in radio telegraphy in the North Sea. In 1905 he was appointed assistant professor of physics at the Institute of Technology, Brunswick and was appointed professor at Munich in 1913. Professor Zenneck has written a number of authoritative books on wireless and also a large number of articles on electro-magnetic oscillations.

ZERO BEAT RECEPTION—The detection of continuous modulated waves using a local source of high frequency current having a frequency equal to that of the incoming wave. When this continuous voltage is impressed on the incoming modulated voltage, (both of the same frequency) the output of the detector contains a current of audio frequency similar in character to the modulated current at the transmitter.

ZERO METHOD—A method of measurement in which various adjustments are made until the current flowing through a galvanometer is reduced to zero, as in a *Wheatstone Bridge* (q.v.). (See *Null Method.*)

ZINC—A metallic element. Its chemical symbol is Zn. In color, zinc is bluish-gray. It is practically non-corrosive in the atmosphere, is capable of taking a high polish, is unaffected by water, but is soluble in nitric acid and in soda and potash solutions. Pure zinc is attacked very slowly by sulphuric acid, but this feature is one of the greatest in the application of zinc in radio work and in electrical work generally.

Zinc is one of the most important components of most dry cells. It forms the negative terminal in most "B" batteries. Zinc in rod form is used in most forms of wet cells.

ZINCITE—An oxide of zinc. Zincite crystals can be distinguished by their red color, often broken up by orange-yellow streaks. This crystal, in combination with several other crystals makes a very excellent detector for radio purposes. The *perikon detector* (q.v.) is a combination of zincite and chalcopyrites. Zincite is also used in combination with bornite, galena, copper, iron pyrites, tellurium and silicon.

ZIRCONIUM—One of the metallic elements. Its chemical symbol is Zr, and its atomic weight is 90.6. Zirconium is an iron-gray powder in one form, or it may be made to crystallize. The crystals look like antimony, are very brittle and extremely hard, being capable of scratching glass and rubies. Zirconium resembles thorium in many of its chemical properties. For the control of the vacuum in high vacuum tubes, a small quantity of thorium or zinconium is included in the tube. These metals combine with hydrogen, oxygen, nitrogen, etc., to form compounds of very low vapor pressure.

SUPPLEMENT

CLASSIFIED INDEX TO S. GERNSBACK'S RADIO ENCYCLOPEDIA

In the following index various co-related subjects have been gathered together for the convenience of the reader or student, under appropriate headings. Thus, under the heading of "Circuits," every circuit contained in the Encyclopedia is listed. Other important subjects are similarly treated. Where the reader is interested in but a single item such as, for example, "Bolitho Circuit," this can be found more conveniently by referring to the Encyclopedia itself, which, of course, is arranged in alphabetical order. All separate and unrelated headings or subjects should be looked up in this manner, that is by referring directly to the Encyclopedia, as these are not listed in the index.

A

AERIAL, (6) — Page
- Balancing ... 21
- Circuit ... 9
- Fan ... 65
- Frame ... 71
- Height of ... 79
- Horizontal ... 83
- Insulation of ... 91
- Insulator ... 92
- Inverted "L" ... 94
- Length of ... 100
- Light Line ... 101
- Plain ... 125
- Quiescent ... 129
- Receiving ... 131
- Switch ... 9
- Transmitting ... 154
- Tuning Condenser ... 9
- Wire ... 10

ALTERNATING CURRENT, (11)
- Theory of Production ... 12
- Undamped ... 156

ALTERNATOR, (12)
- Alexanderson ... 11
- Definition ... 12
- Goldschmidt ... 74
- High Frequency ... 81
- Inductor ... 90
- Laminated Core ... 99
- Polyphase ... 126
- Single Phase ... 139
- Synchronous Motor ... 148

AMPLIFICATION
- Audio Frequency ... 19
- Cascade ... 32
- Constant ... 14
- Factor ... 14
- Impedance Coupled ... 86
- Radio Frequency ... 15
- Resistance Coupled ... 134
- Voltage ... 14

AMPLIFIER, (14)
- Action of Vacuum Tube ... 15
- Audio ... 19
- Intermediate Frequency ... 15, 93
- Radio Frequency ... 15
- Resistance Coupled ... 134

ANTENNA, (6)
- Aeroplane ... 10
- Cage ... 29
- Capacity Measurement ... 31
- Gamma type ... 73
- Inductance ... 80
- Non Radiative ... 118
- Resistance ... 16
- Tuned ... 155
- Tuning Condenser ... 138
- Umbrella ... 156
- Wave ... 159
- Wave Length Calculation for ... 160

ARC
- Characteristic ... 16
- Duddell Singing ... 56
- Generator ... 16
- Generator System ... 16
- Magnetic Blowout for ... 108
- Poulsen ... 126

B

BATTERIES, (22)
- "A" ... 3
- Accumulator ... 136
- "B" ... 22, 125
- "C" ... 32
- "C" Curve ... 32
- Cell ... 33
- Charger, Balkite ... 21
- Chloride ... 35
- Daniell's Cell ... 50
- Electrolytic Cell ... 60
- Filament ... 67
- Float ... 70
- Galvanic ... 72
- Gravity ... 75
- Grid ... 75
- High Potential ... 81
- High Tension ... 82
- Of Alternators ... 22
- Parallel Connection ... 3
- Plate ... 125
- Primary Cell ... 127
- Secondary Cell ... 33, 136
- Series Connection ... 3
- Series-Parallel Connection ... 3
- Storage ... 3, 136
- Testing Instruments ... 22
- Theory of Primary Cell ... 34
- Trigger ... 155

BIOGRAPHIES
- Alexanderson, Ernst F. W. ... 10
- Appleton, Edward Victor ... 16
- Arco, Graf George Von. ... 17
- Armstrong, Edw. H. ... 17
- Austin, Louis Winslo ... 19
- Belin, Edward ... 23
- Bell, Alexander Graham ... 23
- Bellini, Dr. Ettore ... 23
- Blondel, Andree ... 23
- Blondlot, Professor Prosper Reno ... 23
- Branly, Edward ... 24
- Braun, Ferdinand ... 25
- Brown, Sidney George ... 27
- Carpentier, Jules ... 31
- Clerk-Maxwell, James ... 36
- Cohen, Louis ... 39
- Crookes, Sir William ... 45
- De Forest ... 51
- Dubilier, William ... 56
- Edison, Thomas Alva ... 58
- Faraday, Michael ... 65
- Fessenden, Reginald Aubrey ... 66
- Fleming, John Ambrose ... 68
- Gernsback, Hugo ... 73
- Gernsback, Sidney ... 73
- Goldschmidt, Rudolph ... 74
- Goldsmith, Dr. Alfred N. ... 74
- Hammond, John Hays, Jr. ... 78
- Heaviside, Dr. Oliver ... 79
- Heising, Raymond A. ... 79
- Henry, Joseph ... 80
- Hertz, Heinrich Rudolph ... 80
- Hogan, John V. L. ... 82
- Hooper, Stanford C. ... 83
- Janet, Paul ... 95
- Kelvin, Lord William Thompson ... 96
- Kimura, Shunkichi ... 98
- Kolster, Frederick A. ... 98
- Koomans, Nicholas ... 99
- Korn, Arthur ... 99
- Langmuir, Irving ... 99
- Latour, Marius ... 100
- Lodge, Sir Oliver ... 102
- Lombardi, Dr. Luigi ... 105
- Marchant, Edgar Walford ... 110
- Marconi, Guglielmo ... 110
- McLachlan, Norman W. ... 111
- Meissner, Alexander ... 112
- Mullard, S. R. ... 115
- Nally, Edward Julian ... 115
- Noble, Sir William ... 118
- Pannill, Charles Jack ... 122
- Phillips, Raymond ... 124
- Pickard, Greenleaf W ttier ... 124
- Poulsen, Valdemar ... 126
- Prince, Charles Ed d ... 127
- Pupin, Michael ... 127
- Rayleigh, John W am Strutt ... 131
- Roberts, Joseph rrison Thompson ... 135
- Robinson, Jame ... 135
- Robinson, Sam S. ... 135
- Rutherford, S Ernest ... 136
- Squier, Majo eneral George Owen ... 143
- Tesla, Nikol ... 151
- Yokojama, aro ... 161
- Zenneck, J ... 161

C

CAPACITY, (30) Page
- Calculation 29
- Amt. of 34
- Circuit 30
- Condenser 40
- Condenser Curves 40
- Condensers in Parallel 30, 41
- Condensers in Series 31, 41
- Conversion Factor 30
- Definition 30
- Distributed 55
- Earth 30
- Electrostatic 63
- Farad 65
- Frequency Factor 30
- Ground 30
- Measurement of Antenna 31
- Mutual 115
- Resistance 31
- Switch 31

CIRCUITS
- "A" Battery 3
- Armstrong Circuit (Original) 18
- Armstrong 3-Circuit Regenerative Receiver 18
- Armstrong Super-Regenerative 147
- Audio Amplifier 19
- Audio Frequency 71
- Autoplex 19
- Balanced 20
- Balanced Metallic 20
- "B" Battery 22
- Bolitho 24
- Breaker 36
- Broken 36
- Buzzer 29
- Buzzer Exciter 29
- Cascade N. F. and detector 32
- Closed Circuit 37
- Cockaday 37
- Cockaday, with push-pull amplifier 37
- Coupled circuits 44
- Crystal 31, 46
- Definition 36
- Detector 52
- Direct Coupled 54
- Fourth 71
- Heising 79
- Heterodyne 79, 80
- Impedance Coupled 85
- Interference Eliminator 93
- Inverse Duplex 94
- Neutrodyne 117
- Open 36
- Parallel 36
- Receiving 131
- Reflex 132
- Regenerative 65
- Reinartz 133
- Resistance Coupled 134
- Series 3
- Series Parallel 3
- Short 138
- Single Stage Radio Frequency amplifier 15
- Super-Heterodyne 145, 146
- Super-Regenerative, Armstrong 147
- Three-Wire 117
- Tuned Radio Frequency 155
- Two-stage Audio Amplifier 19
- T. Y. K. Arc 156
- Ultradyne 156

Page
- Vacuum Tube, Non-Regenerative 52
- Vacuum Tube, Regenerative 52

COHERER, (39)
- Autocoherer 19
- Branley 24
- Definition 39
- Granular 75
- Lodge Disc Detector 108
- Marconi Filings 110
- Slaby-Arco-Vacuum 141

CONDENSER, (40)—See also Capacity
- Aerial Tuning 9, 138
- Air 10
- Balanced Rotary 21
- Billi 23
- By-Pass 40
- Compressed Air 39
- Curves 40
- Definition 40
- Electrolytic 60
- Filter 68
- Fixed 41
- Grid 41, 76
- In Parallel 41
- In Series 41
- Korda Air 99
- Leyden Jar 101
- Mica 41
- Moscicki 114
- Paper 122
- Phone 41
- Rotary 136
- Shortening 138
- Stopping 144
- Telephone 149
- Variable 41

COILS (39)
- Antenna 39
- Astatic 18
- Basket 21
- Choke 35
- Cockaday 38
- "D" 50
- Damping 50
- Definition 39
- Duolateral 56
- Exploring 65
- Feed-Back 65
- Fixed 68
- Honey-Comb 82
- Impedance 85
- Inductance 87
- Induction 88
- Jigger 95
- Loop Antenna 105
- Lorenz 106
- Low Loss 107
- Movable 115
- Pancake 122
- Plug-In 125
- Primary Induction 89
- Radio Frequency 130
- Reactance 131
- Reaction 131
- Retardation 135
- Repeating 133
- Spark 142
- Spider Web 143
- Tesla 151
- Tickler 153
- Tuning 155

Page
- Vario Coupler 44, 158
- Variometer 158

COUPLING, (44)
- Coefficient 44
- Close 37
- Conductive 41
- Coupled Circuits 44
- Degree of 45
- Definition 44
- Direct 54
- Electrostatic 63
- Feed Back 65
- Inductive 90
- Regenerative 133
- Transformer 45, 153

CRYSTALS
- Bornite 24
- Carborundum 31
- Cat-Whisker for 32
- Cerusite 34
- Chalcopyrite 34
- Definition 45
- Detector 45
- Galena 72
- Iron Pyrites 94
- Molybdenite 114
- Nagyagite 115
- Quartz Oscillator 129
- Receiver 46
- Rectifier 46
- Silicon 138
- Theory of Operation 46
- Zincite 161

CURRENT, (46)
- Alternating 11, 47
- Anode 16
- Assumed Direction of Flow 18
- Audio frequency 19
- Carrying Capacity of Wire 47
- C. G. S. Unit of 34
- Constant 42
- Consumption 42
- Continuous Direct 42
- Convection 42
- Critical 45
- Decay of 50
- Definition 46
- Density 47
- Direct 47
- Direction of Flow (Assumed) 47
- Displacement 54
- Eddy 58
- Electron Theory 62
- Emission 64
- Flow of 70
- Fluctuating 70
- Foucault 70
- Free Alternating 71
- Full Load 72
- Grid 76
- Ground 77
- Harmonic 78
- High Frequency 48, 81
- High Tension 48
- Idle 85
- Intermittent 93
- Leading 100
- Low Frequency 48, 107
- Modulated 114
- Nature of Electric 49

Page

  - Oscillating 120
  - Oscillatory 120
  - Overtone 122
  - Primary 48
  - Production of Electric 48
  - Pulsating 127
  - Pulsating Direct 48
  - Radio Frequency 130
  - Reactive 131
  - Rectified 131
  - Rotary 136
  - Secondary 48, 137
  - Space 142
  - Stray 145
  - Theory of Alternating 11
  - Theory of Electric 49
  - Theory of Flow 151
  - Undamped Alternating 56

D

DETECTOR, (52)
  - Balanced 20
  - Barreter 21
  - Carborundum 31
  - Circuit 52
  - Coherer 39
  - Combination 39
  - Crystal 45
  - Crystal, Theory of Operation 46
  - Cymoscope 49
  - Damped Wave 52
  - Disc, Lodge 103
  - Electrolytic 60
  - Hozier-Brown 84
  - Liquid Barreter 102
  - Lodge Disc 103
  - Marconi Filing Coherer 110
  - Perikon 123
  - Silicon 138
  - Theory of Detector Action 151
  - Thermal 152
  - Ticker 153
  - Tikker 153
  - Tone Wheel 153
  - Vacuum Tube 52, 157, 158
  - Valve 158

DIELECTRIC, (53)
  - Absorption 53
  - Coefficient and Constant 53
  - Frequency 54
  - Constants, Table of 54
  - Hysteresis 54
  - Strength 54

DISCHARGE
  - Brush 27
  - Corona 43
  - Oscillatory 120
  - Oudin Resonator for 121
  - Spark 142

E

ELECTRIC
  - Absorption 59
  - Adhesion 6

Page

  - Displacement 59
  - Elasticity 59
  - Field 59, 66
  - Induction 59
  - Electricity 59
  - Osmosis 121
  - Potential 126

ELECTROLYTIC
  - Action 60
  - Battery Charger 60
  - Cell 60
  - Condenser 60
  - Conduction 60
  - Decomposition 51
  - Detector 60
  - Interrupter 61, 94
  - Rectifier 61
  - Rectifier, Full Wave 61

ELECTROMAGNETIC
  - Attraction 62
  - Control 61
  - Field 61, 66
  - Flux 61
  - Impulse 61
  - Induction 61, 89
  - Inertia 90
  - Instruments 61
  - Microphone 61
  - Radiation 62
  - Repulsion 62
  - Switch 62
  - Theory of Light 62
  - Units 62
  - Waves 62, 160

ELECTROMOTIVE FORCE, (62)
  - Active Pressure 6
  - Back 20
  - Breakdown Potential 25
  - C. E. M. F. 34
  - Counter 44
  - Current, Resistance Relationships (Ohms Law) 119
  - Definition 62
  - Induced 86
  - Kirchoff's Laws 98
  - Potential, Electric 126
  - Pressure, Electrical 127
  - Volt 159
  - Voltage 159
  - Voltage, Constant 159
  - Voltage Drop 159

ELECTRON (62)
  - Flow 62
  - Emission 62
  - Relay 63
  - Theory 62
  - Tube 63

ELECTROSTATIC, (63)
  - Capacity 63
  - Charge 63
  - Coupling 63
  - Field 63, 66
  - Flux 63, 70
  - Flux Density 63
  - Force 63
  - Induction 63, 89
  - Leakage 64
  - Line of Force 64
  - Strain 64

Page

  - Systems 64
  - Units 64
  - Voltmeter 64
  - Wattmeter 64

F

FEED BACK, (65)
  - Armstrong's Circuits 17
  - Coil 65
  - Coupling 65
  - Definition 65
  - Effect 65
  - Regeneration 133
  - Regenerative Circuit 133
  - Reinartz Circuit 133
  - Tickler 153

FIELD, (66)
  - Density 66
  - Electric 66
  - Electromagnetic 66
  - Electrostatic 66
  - Magnet 66
  - Magnetic 66
  - Regulator 66

FILAMENT, (66)
  - Battery 67
  - Consumption 67
  - Current 67
  - Formula for Electron Flow 67
  - Rheostat 67
  - Resistance 67
  - Thoriated 67

FLUX, (70)
  - Density 70
  - Distribution 70
  - Electrostatic 70
  - Leakage 70
  - Lines 70
  - Magnetic 70
  - Magnetism 70

FREQUENCY, (71)
  - Alternator of 71
  - Audio 19, 71
  - Band 71
  - Carrier 71
  - Changer 71
  - Circuit, Audio 71
  - Circuit, Radio 71
  - Cycle 49
  - Definition 71
  - Fundamental 71
  - Group 71, 77
  - High 71
  - Low 71
  - Meter 71
  - Periodicy 123
  - Radio 72
  - Unit of 72
  - Variation of Power Factor with 72

G

GALVANOMETER, (73)
  - Constant 73
  - D'Arsonval 73
  - Mirror 73, 114

GRID, (75) Page
Battery 75
Bias 76
Circuit 76
Condenser 41, 76
Control 76
Current 76
Leak 6, 76
Potential, Modulation of 76
Potentiometer 77

GROUND, (77)
Clamp 77
Connection 77
Currents 77
Detector 77
Switch 77
Wire 77

H

HETERODYNE, (80)
Beats 22
Dissonance 55
Driver 56
Local Oscillations 102
Receiver 80
Self- 137
Super-Heterodyne Receiver 145
Wave Meter 81

HIGH FREQUENCY, (71, 81)
Alternator 81
Amplification 81
Buzzer 81
Choke 35
Current 48, 81
Component of Plate 81
Oscillation 81
Resistance 81, 135

I

IMPEDANCE, (85) Page
Coil 85
Coupled Amplifier 86
Factor 86
Matched 86
Tuned 155

INDUCTANCE, (86)
Antenna 86
Coils 87
Distributed 87
Loading 87
Measurement of 87
Mutual 87
Parallel 88
Reactance 87
Receiving 87
Self 87
Series 88
Tapped 87
Units of 88
Variable 88

INDUCTANCE COIL
Basket Wound 21
Choke 35

Page
Definition 87
Duolateral 56
Honey-Comb 82
Lorenz 106
Spider Web 143
Tuning 155
Vario Coupler 44, 158
Variometer 158

INDUCTION, (88)
Bridge 88
Coil 88
Coil, Primary 89
Coil, Secondary 89
Density 89
Electromagnetic 89
Electrostatic 89
Instruments 89
Machine 89
Magnetic 89
Motor 89
Motor, Squirrel Cage 144
Mutual 89
Reactance 90
Regulator 90
Screen 90
Unit 90
Wattmeter 90

INDUCTIVE
Capacity 90
Circuit 90
Connection 90
Coupler 90
Coupling 90
Drop in Voltage 90
E. M. F. 90
Reactance 90
Rise 90
Resistance 90

INSULATING MATERIALS, (91)
Bakelite 20
Compound 91
Definition 91
Dielectric Constants 5
Ebonite 57, 159
Eburin 57
Electrose 63
Impregnated 86
List of 149
Mica, 113
Micabond Cloth 113
Micabond Paper 113
Micabond Plate 113
Micanite 113
Property 91
Table of Dielectric Strength 149
Vulcanite 159
Vulcanized Rubber 159

K

KEY, (96) Page
Anti-Capacity 96
Definition 96
High-Speed 96
Magnetic 109
Modulated Continuous Waves 42
Switch 97
Transmitting 154

M

MAGNET, (108) Page
Artificial 18
Compound 40
Definition 108
Electro- 61
Horseshoe 83
Natural 116
Negative Pole 116
North Pole 118
Permanent 123
South Pole 142
Temporary 151
Wire 108

MAGNETIC
Attraction and Repulsion 108
Blowout 108
Circuit 108
Field 66, 108
Flow 109
Flux 70, 109
Flux Density 109
Force 109
Force Due to Current 109
Hysteresis 109
Induction 109
Inductive Capacity 109
Interrupter 109
Key 109
Leakage 109
Needle 109
Poles 109
Properties 109
Storm 109

O

OSCILLATIONS
Back 20
Constant 120
Damped 49
Definition 120
Forced 70
Free 71
Frequency 120
High Frequency 81
Impact Excitation 85
Local 102
Logarithmic Decrement 105
Master Oscillator System 111
Oscillion 121
Oscillogram 121
Oscillograph 121
Point 120
Poulsen Arc 126
Resonator 135
Tesla Coil 151
Transformer 45, 120
Undamped 156
Valve 120

OSCILLATOR, (120)
Hertzian 120
Linear 120
Magnetic Blowout for 108
Master 111
Negatron 117

Page

- Righi 120
- Quartz 129
- Thermionic 120

P

PLATE, (125) Page

- Battery 125
- Circuit 125
- Component, High Frequency 125
- Consumption 42, 125
- Control 125
- Current 125
- Impedance 125
- Voltage 125

R

RADIATION, (130)

- Definition 130
- Efficiency 130
- Electromagnetic 62
- From Antenna 130
- Gliding Through 74
- Heaviside Layer 79
- Resistance 130, 135
- Theory of Propagation of Electromagnetic Waves 152
- Theory of Reception of Electromagnetic Waves 152

RADIO, (130)

- Activity 130
- Channel 130
- Communication 130
- Compass 130
- Control 130
- Frequency 72
- Telegraphy 130, 161
- Telephony 130
- Transmission 130

RADIO FREQUENCY, (72)

- Ammeter 14
- Amplifier 15
- Circuit 71
- Coils 130
- Current 130
- Resistance 135
- Selectivity 130
- Signals 130

REACTANCE, (131)

- Capacitative 30
- Coil 131
- Inductance 87

RECTIFIER, (131)

- Electrolytic 61
- Full Wave 131
- Half Wave 78
- Potential 126

RESISTANCE, (134)

- Antenna 16
- Box 134
- Contact 42
- Coupled Amplifier 134

Page

- Coupling 134
- Definition 134
- Field Regulator 66
- Graphite 75
- High Frequency 81, 135
- Litzendraht 102
- Of Ground Connection 135
- Ohm 118
- Ohmic Drop 119
- Ohmic Loss 119
- Ohm's Law 119
- Radiation 135
- Radio Frequency 135
- Variable 158

RESONANCE, (135)

- Curve 135
- Intensifies 135
- Parallel 123
- Series 137
- Transformer 135

S

SWITCH, (147) Page

- Aerial 9
- Board 147
- Capacity 31
- Dead End 50
- Double Throw 56
- Electromagnetic 62
- Gear 148
- Ground 77
- Jack 95
- Key 96
- Key, Anti-Capacity 96
- Knife 98
- Lightning 101
- Plug 125
- Points 42
- Quick Break 129

T

THEORY OF CURRENT FLOW, (151)

- Detector Action (of Vacuum Tube) 151
- Electric Current 49
- Magnetism 14
- Operation of Crystal Detector 46
- Operation of Vacuum Tubes 151
- Primary Cell 34
- Production of Alternating Current 12
- Propagation of Electromagnetic Waves 152
- Reception of Electromagnetic Waves 152
- Vacuum Tube Amplifier 157

TRANSFORMER, (153)

- Air Core 10
- Auto 20
- Closed Core 37
- Coupling 45, 153
- Hedgehog 79
- Intermediate 93
- Open Core 120
- Oscillation 45, 120
- Output 122
- Plug-In 125

Page

- Resonance 135
- Shielded 138
- Steel 153
- Tuned Radio Frequency 155

TRANSMISSION

- Aerial, Cage 8
- Automatic 19
- Broadcasting, General Treatise on Methods. 25
- Carrier Wave 31
- Chopper 31
- Fading 65
- Fan Aerial 65
- Feed-Voltage Modulation 65
- Ferro-Magnetic Modulator 66
- Flame Microphone 68
- Frequency Carrier 71
- Heising Modulation 74
- Key 96, 154
- Key, High Speed 96
- Jamming 95
- Jigger 95, 154
- Microphone 113
- Modulation 114
- Of Photographs by Radio 154
- Phenomena of Electric Wave Propagation.. 124
- Quartz Oscillator 129
- Queenland Gap Transmitter 129
- Radio 131
- Radio Telephony 131
- Radio Telegraphy 130
- Remote Control 133
- Spark Coil 142
- Spark Rate 142
- Speech Modulation 143
- Synchronous Discharger 148
- Telefunken System 149
- Television 150
- Tesla Coil 151
- Theory of Propagation of Electromagnetic Waves 152
- T. Y. K. Arc 156

TUNING, (155), See also Coil; Condenser

- Broad 155
- Coil 155
- Condenser, Aerial 9
- Inductance Coils 87
- Inductance, Variable 88
- Interference Eliminator 93
- Interference Preventers 93
- Rejector Circuit 133
- Selectivity 137
- Sharp 137
- Static Eliminators 144
- Syntony 148
- Wave Trap 160

U

UNITS, (156)

- Ampere 14
- Ampere-Hour 14
- Ampere, International 14
- Ampere Turns 14
- Ampere Volt 14
- B.A. Ampere 20
- Beg Ohm 23
- B. O. T. 24
- B. T. U. 27

针灸

Claudia Focks

Atlas of Acupuncture

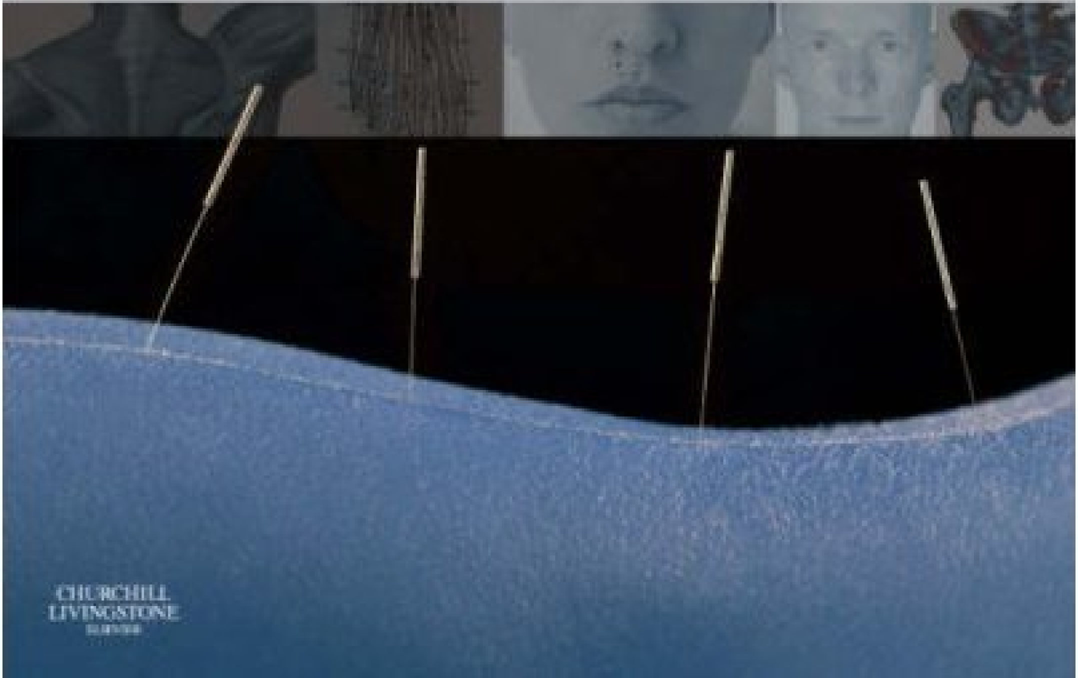

CHURCHILL
LIVINGSTONE
ELSEVIER

Dedication

To my parents Paula and Alfons Focks who have always supported my siblings and me in finding and pursuing our own paths, with love and gratitude

Commissioning Editor: Karen Morley, Claire Wilson
Development Editor: Kerry McGechie, Martin Mellor
Project Manager: Morven Dean
Designer: Stewart Larking
Photographs: Anja Messerschmidt
Graphics: Henriette Rintelen

CHURCHILL
LIVINGSTONE
ELSEVIER

CHURCHILL LIVINGSTONE
An imprint of Elsevier Limited

First and second edition published in German under the title *Atlas Akupunktur*
1. First edition 1998
2. Second edition 2006
© Elsevier GmbH, München

First edition published in English
© 2008, Elsevier Limited. All rights reserved.

The right of Claudia Focks to be identified as editor of this work has been asserted by her in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the Publishers. Permissions may be sought directly from Elsevier's Health Sciences Rights Department, 1600 John F. Kennedy Boulevard, Suite 1800, Philadelphia, PA 19103-2899, USA: phone: (+1) 215 239 3804; fax: (+1) 215 239 3805; or, e-mail: healthpermissions@elsevier.com. You may also complete your request on-line via the Elsevier homepage (http://www.elsevier.com), by selecting 'Support and contact' and then 'Copyright and Permission'.

First edition	**1998**
Second edition	**2006**
English edition	**2008**

ISBN-13: 9780443100284
ISBN-10: 0443100284

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress

Notice

Neither the Publisher nor the Editor assumes any responsibility for any loss or injury and/or damage to persons or property arising out of or related to any use of the material contained in this book. It is the responsibility of the treating practitioner, relying on independent expertise and knowledge of the patient, to determine the best treatment and method of application for the patient.

The Publisher

ELSEVIER your source for books, journals and multimedia in the health sciences
www.elsevierhealth.com

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Printed in China

The
Publisher's
policy is to use
paper manufactured
from sustainable forests

Preface

'An acupuncture point is located where it is located' – like veins, arteries or nerves, the location of acupuncture points may vary. Consequently, locating and stimulating acupuncture points is an individual process in each patient, similar to locating veins or arteries for puncture. The goal of this atlas is therefore to help acupuncture students and therapists find the correct location of acupuncture points. Ultimately, the exact location of an 'acupuncture hole' will be where it can be palpated, addressed energetically in a safe way, and – most importantly – where it will be therapeutically effective.

Acupuncture therapy requires not only the knowledge to locate individual points but also an understanding of the context of these points. This atlas therefore includes both a description of the individual points as well as the central idea – the channel system.

The core chapters 4, 5 and 6 contain single page portraits of all the channel points as well as extra points used in body acupuncture. The clearly structured page layout, complemented by pictograms, allows for quick access to the clinically relevant information for the location and indication of individual points. Clear step-by-step instructions guide the reader through the surrounding surface anatomy to the correct location of the point. Figures with the relevant anatomical structures, as well as drawings showing channel pathways pertaining to the relevant region of the body, provide a further practical aid for correct point location. The text and the photo details also contain information regarding other points located in the vicinity or points in comparable locations in other parts of the body. This not only draws attention away from the individual points to the larger anatomical orientation, but also fosters an understanding of the context between the location and the action of a given point.

Chapter 7 presents the channel points according to the anatomical region – including illustrations of anatomical overview as well as text/picture details. This format will deepen the understanding of the anatomical relationship between the channel points and complete this academic concept. Chapter 2 'Location Methods and Cun Measurements' and Chapter 3 'Anatomical Orientation' provide further support in locating the individual points. The modified illustrations in these chapters are based on the well-known Sobotta Atlas.

Chinese medicine practitioners work with the Qi and its flow. In this context the individual 'acupuncture hole' has both an anatomical and an energetic component and is an intrinsic part of the channel and vessel network, which provides an exchange between the body's Exterior and Interior. In this way Chapter 1 describes the network of channels and vessels, and Chapter 8 the point categories and point combinations, explaining them in terms of their effects with relation to channel energetics, and presenting them in a clearly defined and visually organised manner. Chapter 9 contains up-to-date information and data about the scientifically proven effects or otherwise of individual points.

I hope the new edition of this atlas will provide a valuable help to students and therapists alike in studying acupuncture and applying the information in clinical practice. I look forward to hearing your constructive criticism and ideas.

Claudia Focks, March 2008

Acknowledgements

This Acupuncture Picture Atlas is a joint project of many very helpful contributors who have helped to complete it.

First of all my thanks go to the models and especially to the photographer Anja Messerschmidt. Her understanding and photographic skills provided the visual foundation for the Atlas to also be an artistic ‘feast for the eyes’.

I would like to thank the graphic designer Henriette Rintelen for the successful graphic design and converting my ideas into visual images.

Thanks go to my colleague Ulrich März for the succint arrangement and editing of the chapter on Anatomical Orientation as well as his contributions to the core chapters. I would like to thank my colleague Ingolf Hosbach for his contributions to the chapter on Research as well as his idea about the visually concise pictograms and for partially converting them by computer.

I want to thank all the staff at Elsevier involved in this project for their, once again, very productive co-operation. My thanks go particularly to Rolf Lenzen for agreeing to do the atlas, Christine Kosel for dealing with the gruelling organisation during production, as well as Petra Münzel-Kaiser and Christl Kiener for editing. My thanks extend particularly to Christl Kiener for her numerous ideas and fruitful exchange of concepts.

My heartfelt gratitude also includes my teachers who have sown the nourishing spark of enthusiasm for Chinese and energetic medicine. I thank my patients for their trust and for allowing me to learn from them.

I would also like to thank the strong women in my family and my circle of friends for their inspiration and being a source of vibrant female power. Thank you to my sister Annette Focks for her friendship and wonderful music, which helped me overcome some of the hurdles of this project. I extend particular thanks to Anneliese Solenski and Christa Skopp for their helpful support with my private affairs.

Above all I want to thank my inspirational husband Christoph Ranzinger and our wonderful children Tabea and Lasse for their love and generous tolerance, without which the Atlas would not have been possible.

For the English edition I would like to extend my thanks to the staff at Elsevier in London, particularly Karen Morley for accepting the atlas and to Kerry McGechie and Martin Mellor for the extensive editing work. Thank you also to Johanna Schuster who has translated the Atlas into English with great care and knowledge.

Claudia Focks, Rottweil, March 2008

About the editor and contributors

Dr med. Claudia Focks, Rottweil: anaesthesiologist, additional qualification in naturopathy and homeopathy; training in Chinese Medicine in China, Austria, and Germany, postgraduate training at the ETHNOMED Institut für Ethnomedizin e.V. (Institute for Ethnomedicine, registered charity) in Munich. Clinical focus on Chinese Medicine and energy therapies. Co-editor and co-author of *Leitfaden Chinesische Medizin* (*Chinese Medicine: An Introduction*), co-author of *Leitfaden Akupunktur* (*Acupuncture: An Introduction*), both published by Elsevier Munich. Lectures on Chinese Medicine, both in Germany and abroad.

Dr med. Ulrich März, Ulm: General practitioner with his own practice, focus on acupuncture and Traditional Chinese Medicine. Adjunct lecturer at the University of Ulm for acupuncture and TCM. Project manager for the postgraduate programme for acupuncture and TCM at the Akademie für Wissenschaft, Wirtschaft und Technik (Academy for Science, Economy and Technology) at the University of Ulm. Co-author of *Leitfaden Akupunktur* (*Acupuncture: An Introduction*), published by Elsevier Munich.

Dr med. Ingolf Hosbach, Bochum: Anaesthesiologist, additional qualification in palliative therapy. Employed at the University Clinic Bergmannsheil Bochum. Adjunct lecturer at the Universities of Witten/Herdecke and Bochum. Scientific publications and lectures on palliative therapy, acupuncture and TCM research. Webmaster of the homepage of the Department for Chinese Medicine at the University Witten/Herdecke and of the DWGTCM e.V. (www.dwgtcm.com).

Frontcover-F10028 2/23/08 11:37 AM Page 1

Pictogrammes

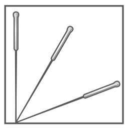

Angle of insertion: The angle of insertion recommended for a particular point (see figure below) is shown in colour. (Note that occasionally more than one angle is possible – this is signified by colouring as appropriate).

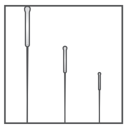

Depth of insertion: Shallow: up to 0.5 cun, **Medium:** 0.5–1 cun, **Deep:** >1 cun. (In some cases more than one depth is possible depending on the angle of insertion – this is signified by colouring as appropriate).

Caution! Organs and structures such as the eyes, nerves, blood vessels, peritoneum (symbolised by intestinal tissue) etc. that might be injured by needling a particular point are shown by a symbol together with a blue exclamation mark. A lightning bolt signifies a point that is painful when needled.

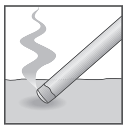

Moxibustion: In principle all points can be treated with moxibustion. The following options are pointed out: **Blue colour below moxa cigar:** moxibustion particularly recommended, **Question mark next to moxa cigar:** moxibustion questionable (contraindicated according to some classical texts, but contradictory information), **Moxa cigar crossed out:** moxibustion contraindicated.

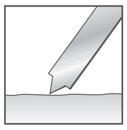

Bloodletting: In principle bloodletting may be applied to all points. **Neutral pictogramme:** bloodletting possible, **Blue droplet of blood:** bloodletting particularly recommended, **Blue question mark:** bloodletting questionable

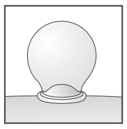

Blue cup: Cupping particularly recommended.

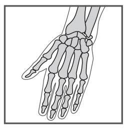

Anatomical orientation: A blue hand signifies cases where anatomical structures are of particular importance for point location (for more detail → chapter 3).

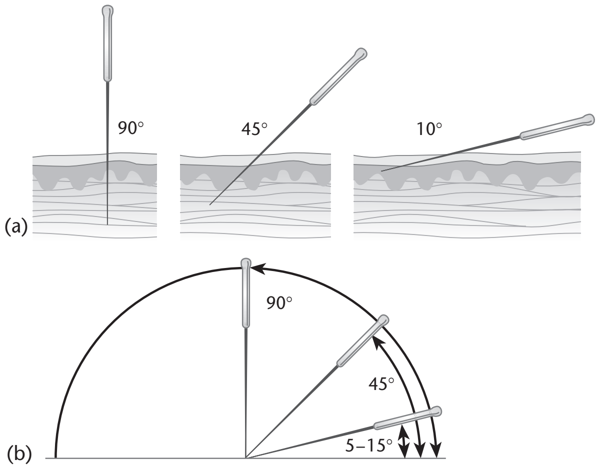

Angle of insertion

a) Angle of insertion and layers of tissue: All needles are inserted to the same ‘depth’ (=the same proportion of the needle is inserted into the tissue), but reaching different levels of tissue.

b) Angle of insertion: transverse (subcutaneously): 5–15°; oblique: 15–45°; perpendicular: 90°.

Point hierarchy (Chapters 4–6)

In the introductions to the channels (chapters 4–6) the importance of points are marked with the following symbols:

🟦🟦: very important, universal point

🟦■: important point

This hierarchy is of course subjective and influenced by the authors’ clinical experience but has proved very helpful for beginners.

1 *Jing luo* (Channel and Network Vessel System)

Claudia Focks

1.1 Introduction

1.1.1 Two Models of Qi Flow in the Channels – Historical Overview

There exist essentially two views about the direction of the Qi flow in the channels, which are based on different understandings of the Qi movement within the body (→ Pirog 1995, Manaka 1995/2004). For a better understanding of either model, imagine a person extending his/her arms towards the sky (→ Figs. 1.1 and 1.2).

Centripetal circulation model

The centre of classical Chinese cosmology, which describes humans as standing between heaven (Yang) and earth (Yin), includes the mutual relationship, influence, and dependence between cosmos and man. Accordingly, early records of the pathways (→ see appendix) depict the (11/12) channels as pathways that originate at the extremities, flow towards the centre of the body and terminate either on the head or the torso. This is the distinguishing feature of the **centripetal circulation model** (→ Fig. 1.1).

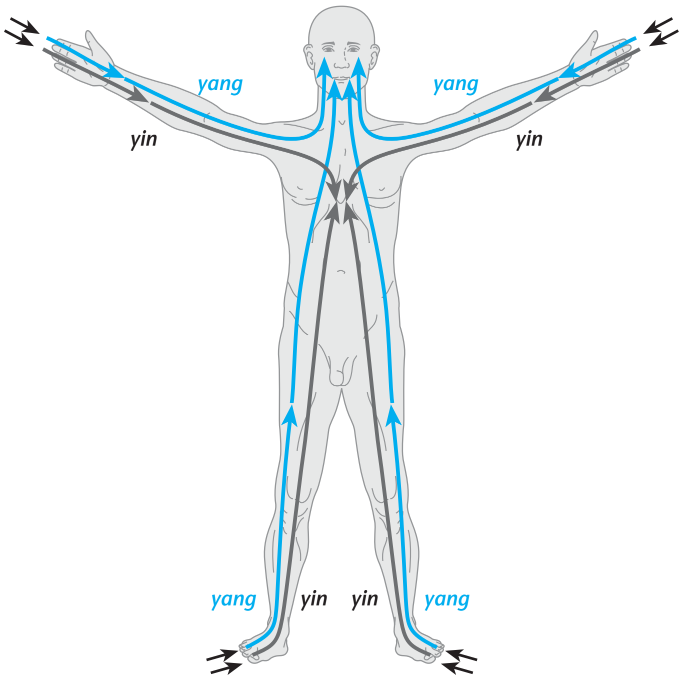

Fig. 1.1 Centripetal circulation model (modified according to Pirog 1996)

This model is also present in the theory of the five *shu*-Transporting points (→ 8.1.6). The macrocosmic Qi enters the body at the tips of the extremities and can be compared to the course of a river. It begins very dynamically as a well, spring and stream, expands to form a river and flows into the wide, deep sea at the elbows and the knees and then further to the internal organs via the channels.

According to this model, the functions of the channels can be compared to antennae, which receive the cosmic influence, transmitting it into the body. The Qi flow in the channels is always from distal (coming from the Exterior, entering at the tips of the extremities) to proximal (towards the centre, flowing towards the internal organs). Each channel connects man with a different part of the cosmos, which can be identified by a numerological structure. Thus a relationship was assumed between the eight extraordinary vessels and the eight trigrams of the *Yijing* (→ 1.7). The primary channels reflect the 12 earthly branches and the 10 heavenly branches. Since the early records only mention 11 channels (→ Appendix 2), they were counted as 10 channels on the arms and 12 channels on the legs. Later, the 12 primary channels were mainly associated with the 12 earthly branches – the 10 heavenly branches rather representing the Five Phases model (*note*: the earthly branches can be included in the latter as well).

Self-contained circulation model

In the course of the development of modern Chinese society, the idea of a close relationship between man and cosmos (as a heaven-earth-man model) became weaker. The channel system was now increasingly compared to phenomena created by man himself such as **canals** and **drainage ditches**. The importance of the original connection with the cosmos declined while that of relationships within society was on the rise. Chinese society became more complex, forming a self-contained unit. Similarly, the concept of the channels as a connection to the macrocosm was partially abandoned and the Qi flow is described as independent and self-contained (→ Fig. 1.2). According to this **self-contained circulation model**, Qi can flow forward as well as backward – from the outer extremities to the Interior of the body and from the Interior to the Exterior.

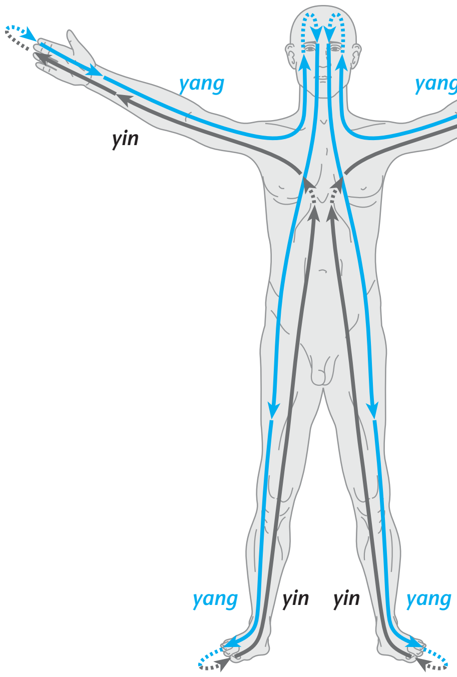

Fig. 1.2 Self-contained circulation model (modified according to Pirog 1996)

Thus the direction of the Qi flow in the primary channels can be from proximal to distal and vice versa, depending on the Yin/Yang polarity of the respective channel, as well as the respective extremity.

This second, probably more recent concept describes the Qi **as circulating continuously** through the body (→ Appendix 2): from the thorax to the hand, to the head, to the foot and back to the thoracic region. These ideas reflect the development of the Chinese civilisation, its agriculture and in particular its **water control** and water **storage** in reservoirs, drainage canals, ditches, etc., which largely form the foundation of the theory of the **channels as water conduits** circulating Qi and Blood. The connections among the channels are considered to be anastomoses (→ 1.2.2), which facilitate the ceaseless, circular flow of Qi from one channel to the next, allowing the exchange of the Qi flow with the Interior.

The self-contained circulation model can thus be described as follows:

- Hand Yin channels run from the thorax to the hand: **LU, HE, P**
- Hand Yang channels run from the hand to the head: **L.I., S.I., T.B.**
- Foot Yang channels run from the head to the foot: **ST, BL, G.B.**
- Foot Yin channels run from the foot to the thorax: **SP, KID, LIV**

More importance is attached to this model of Qi flow in the Western acupuncture tradition, which can in part explain the use of numbers in naming the acupuncture points, rather than names as in China.

However, the structure of this new circulatory model was probably also too rigid to sufficiently explain some of the effects of acpuncture. According to Pirog (1996), this might be the reason why the secondary channels, for example the sinew channels (→ 1.4) and the divergent channels (→ 1.3), with their rather primitive and natural pathways, were integrated into the channel and network vessel system (*jing luo*) in accordance with the **centripetal** circulation model.

Comparison of the two circulation models (modified according to Pirog 1996)

	Centripetal model	**Self-contained model**
Direction of Qi-flow	Always from distal to proximal	Either from proximal to distal or vice versa, depending on the polarity of the channel (Yin/Yang)
Origin of Qi	From outside the body. The distal end of the channel is open in order to receive cosmic Qi	Originates in the inside of the body. The distal end of the channels is connected to the channel that follows next.
Function of the channels	Transporting the Qi from the outer cosmos to the Interior of the body. Supporting the relationship between man and nature/ cosmos	Circulation of the Qi inside the body. Supporting man's relationship with himself

1.1.2 Overview of the Channel and Network Vessel System (*jing luo* System)

In *Ling Shu*, chapter 11, it says: 'Man lives, diseases occur … both the beginner and the experienced master always have to start with the channels and network vessels (*jing luo*).'

In Chinese Medicine, the *jing luo* are considered to be a network of channels and vessels in which the Qi and Blood (*xue*) flow. They are connected to the Organ systems (*zangfu*) and 'water' the whole organism, supplying the body with Qi and Blood (*xue*) on the surface (Exterior) and deep inside the body (Interior), above as well as below.

From a functional point of view, the channels and network vessels (*jing luo*) govern the distribution of Qi and Blood (*xue*), they regulate Yin and Yang and they protect the body. However, they also enable the spreading of diseases. Reactions to any disorders may therefore manifest along these channel pathways. These could be either disorders of the channels themselves, or external reflections of *zangfu* disorders. In clinical practice, the channels and network vessels (*jing luo*) can be utilised in order to send Qi to the diseased parts of the body (for an overview of the classification and nomenclature of the *jing luo* system → Fig. 1.3).

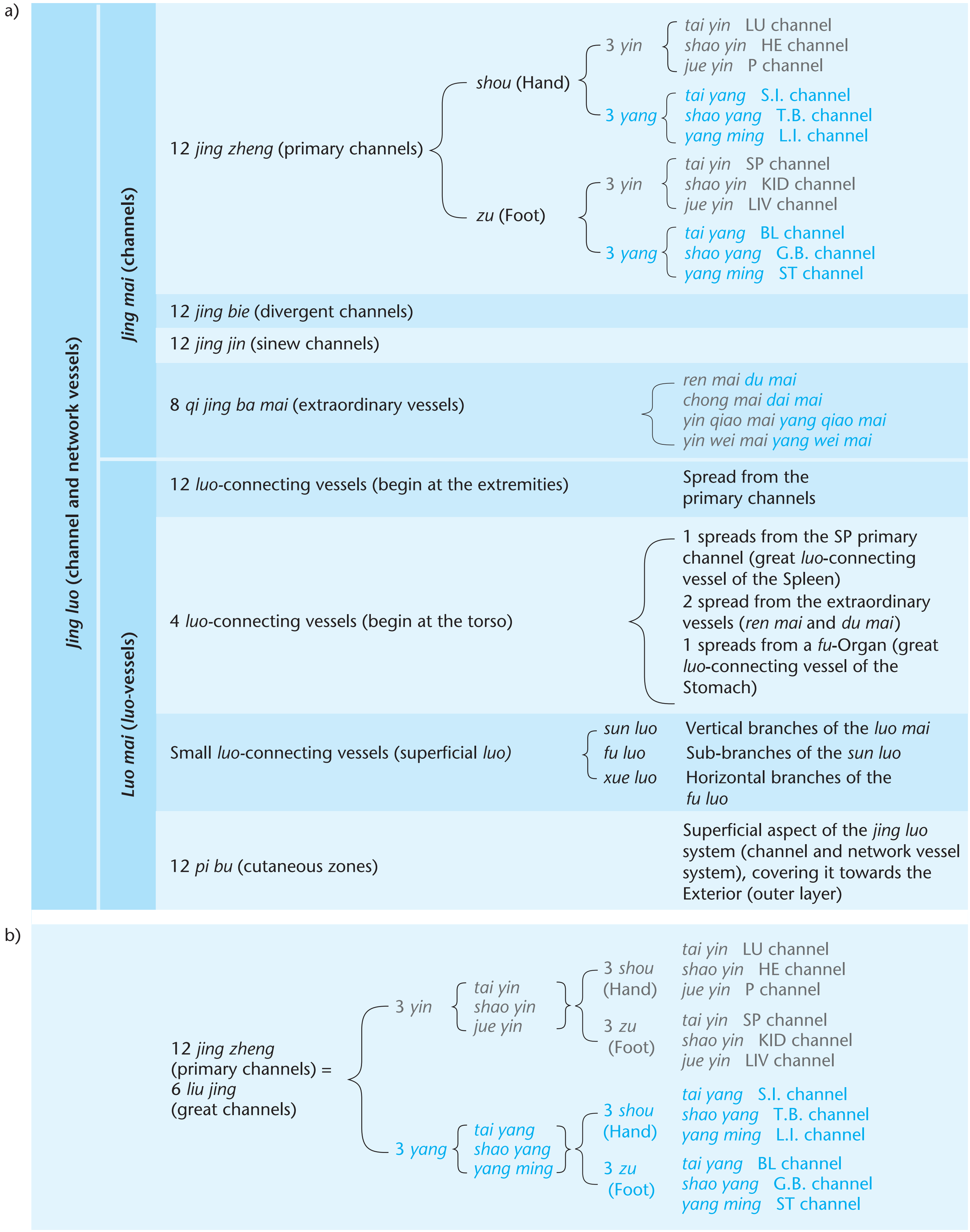
a)
Jing luo (channel and network vessels)
Jing mai (channels)
12 jing zheng (primary channels)
shou (Hand)
3 yin
tai yin LU channel
shao yin HE channel
jue yin P channel
3 yang
tai yang S.I. channel
shao yang T.B. channel
yang ming L.I. channel
zu (Foot)
3 yin
tai yin SP channel
shao yin KID channel
jue yin LIV channel
3 yang
tai yang BL channel
shao yang G.B. channel
yang ming ST channel
12 jing bie (divergent channels)
12 jing jin (sinew channels)
8 qi jing ba mai (extraordinary vessels)
ren mai du mai
chong mai dai mai
yin qiao mai yang qiao mai
yin wei mai yang wei mai
Luo mai (luo-vessels)
12 luo-connecting vessels (begin at the extremities)
Spread from the primary channels
4 luo-connecting vessels (begin at the torso)
1 spreads from the SP primary channel (great luo-connecting vessel of the Spleen)
2 spread from the extraordinary vessels (ren mai and du mai)
1 spreads from a fu-Organ (great luo-connecting vessel of the Stomach)
Small luo-connecting vessels (superficial luo)
sun luo
fu luo
xue luo
Vertical branches of the luo mai
Sub-branches of the sun luo
Horizontal branches of the fu luo
12 pi bu (cutaneous zones)
Superficial aspect of the jing luo system (channel and network vessel system), covering it towards the Exterior (outer layer)
b)
12 jing zheng (primary channels) = 6 liu jing (great channels)
3 yin
tai yin
shao yin
jue yin
3 shou (Hand)
3 zu (Foot)
tai yin LU channel
shao yin HE channel
jue yin P channel
tai yin SP channel
shao yin KID channel
jue yin LIV channel
3 yang
tai yang
shao yang
yang ming
3 shou (Hand)
3 zu (Foot)
tai yang S.I. channel
shao yang T.B. channel
yang ming L.I. channel
tai yang BL channel
shao yang G.B. channel
yang ming ST channel

Fig. 1.3 (a) Overview and classification of the *jing luo* system (channel and network vessel system); (b) the 12 primary channels and the six great channels *(liu jing)*

1.1.3 Distribution and Organisation of the Channel and Network Vessel (*jing luo*) System

According to the law of the Exterior–Interior (*biao-li*) the 'Exterior' communicates with the 'Interior'.

Exterior (*biao*) corresponds to the skin, the muscles and the pathways of the channel and network vessel superficial (*jing luo*) system. The deeper pathways of the channels and the Organ systems (*zangfu*) are attributed to the Interior (*li*). A particular organisational structure within the *jing luo* system is necessary to safeguard the circulation of Qi and the communication between the Exterior and the Interior. In this respect, the extraordinary vessels play a special role. While they play a major role in coordinating and regulating the primary channels and the *jing luo* system in general, they do not directly connect the Interior and the Exterior. There is also no direct connection between the extraordinary vessels and the *zangfu*-Organs (see ➞ 1.7 and Chapter 5).

Depth organisation of the *jing luo* system

Differing ideas exist regarding the depth at which the various channels and vessels are located within the body. The table below and Fig. 1.4 show the ideas according to Solinas et al. (1998) and Deadman et al. (1998) in a modified way. While these authors regard the deep, Interior pathways of the primary and divergent channels as the deepest channel structures, other authors attribute this role to the eight extraordinary vessels (for the eight extraordinary vessels, see ➞ 1.7).

Overview over the possible location with regard to depth of the *jing luo* system

Levels	Channel system
Superficial levels (skin and muscles)	• 12 cutaneous regions • Superficial *luo*-connecting vessels (*xue luo, fu luo, sun luo*) • 12 sinew channels (*jing jin*)
Mid-levels	• 15/16 *luo*-connecting vessels (*luo mai*) • External pathways of the 12 primary channels • External pathways of the 12 divergent channels (*jing bie*) • Eight extraordinary vessels
Deep levels (*zangfu*-Organs)	• Deep (internal) pathways of the 12 primary channels • Deep (internal) pathways of the 12 divergent channels

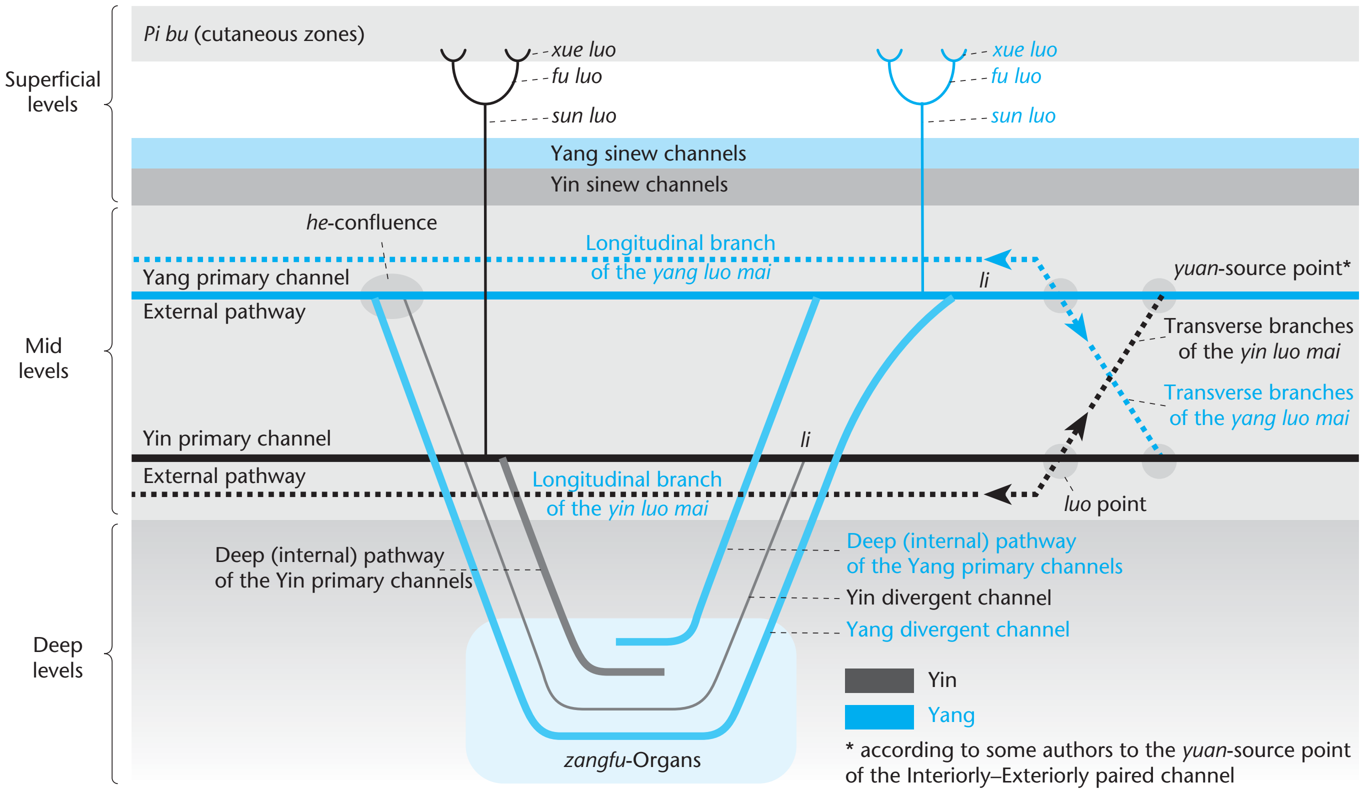
Superficial levels
Mid levels
Deep levels
Pi bu (cutaneous zones)
xue luo
fu luo
sun luo
xue luo
fu luo
sun luo
Yang sinew channels
Yin sinew channels
he-confluence
Longitudinal branch of the *yang luo mai*
Yang primary channel
External pathway
li
yuan-source point*
Transverse branches of the *yin luo mai*
Transverse branches of the *yang luo mai*
Yin primary channel
li
External pathway
Longitudinal branch of the *yin luo mai*
luo point
Deep (internal) pathway of the Yin primary channels
Deep (internal) pathway of the Yang primary channels
Yin divergent channel
Yang divergent channel
zangfu-Organs
Yin
Yang
* according to some authors to the *yuan*-source point of the Interiorly–Exteriorly paired channel

Fig. 1.4 Illustration of the possible organisation of the *jing luo* system (channel and network vessel system)

Channel systems

Each primary channel, together with its pertaining 'secondary channels', forms a complex, multi-layered organisational unit, a 'system'. These systems comprise the various relationships and connections among each other as well as with the eight extraordinary vessels (→ 1.7, → Chapter 5). They support the harmonious regulation within the body. There are 12 channel systems (hand *taiyin*, hand *yangming*, etc.), each of which includes the following:

- a **primary channel** (*jing zheng*) (→ 1.2) with an external and internal pathway, which enters its pertaining *zangfu*-Organ
- a **divergent channel** (*jing bie*) (→ 1.3)
- a **sinew channel** (*jing jin*) (→ 1.4)
- a great ***luo*-connecting vessel** (*luo mai*) (→ 1.5).

In turn, the **cutaneous zones** (*pi bu*) (→ 1.6) wrap around the whole channel system.

1.1.4 Qi Circulation and the Channel System

What moves with and within the channels?

The channel system represents the 'pathway of the Qi'. In acupuncture many schools of thought work with the **True Qi (*zhen qi*)**, which, according to Maciocia (1989), represents the final stage of a process of Qi refinement and transformation: the Gathering Qi (*zong qi*), acted upon by the Original Qi (*yuan qi*) as a catalyst, becomes True Qi (*zhen qi*), which manifests in its two aspects, the Defensive Qi (*wei qi*) and the Nutritive Qi (*ying qi*). According to Larre and Rochat de la Vallée (1986), the True Qi is regarded as the **sum of all Qi mechanisms and/or forms of Qi** in the body. In other words: if there is a balanced and harmonious flow in the channels, this can be called True Qi (*zhen qi*). It means that which circulates in the 'here and now'. The sum total of all forms of Qi in their correct alignment is called **Upright Qi (*zheng qi*)** and forms the counterpart to Evil Qi (*xie qi*), for example to pathogenic factors or counterflow Qi.

Forms of Qi

Essence (Qi) (*jing qi*)

The Essence (Qi) acts within the organism and, according to some authors, to some extent also in the extraordinary vessels (→ 1.7). It represents the combination of Early Heaven Essence inherited from the parents, which, according to many schools of thought, is stored in the Kidneys and has a relationship to the Life Gate (*mingmen*), and the Later Heaven Essence of the Middle Burner. They support and complement each other.

Original Qi (*yuan qi*)

The Original Qi (*yuan qi*) is often described as the **active form of Essence (*jing*)**, circulating in the channels and spreading to the whole organism with the help of the Triple Burner. Like a catalyst, it is the dynamic driving force which wakens and sustains the functional activity of all organs and structures. It has a pre- and a postnatal component. Therefore its proper functioning is dependent on the supply of acquired Qi derived from water and food supplied by the Middle Burner. The Original Qi can be directly accessed and influenced at the *yuan*-source points (→ 8.1.1) or at the points **Ren-17** (*danzhong/shanzhong*), **Ren-12** (*zhongwan*) and **Ren-6** (*qihai*).

Gathering Qi (*zong qi*)

According to Larre and Rochat de la Vallée (1986), the Gathering Qi (*zong qi*) is formed in the centre of the thorax when the essences derived from respiration (Great Qi, *da qi*) meet with the food (*shui gu*, from Grain Qi (*gu qi*) and water) and are set in motion through ancestral mechanisms by the first postnatal breath. Like an 'engine', it dictates the rhythm and circulation, it controls respiration and it regulates the heartbeat. It does not circulate itself, but is like a 'Sea of Qi' that 'collects' in the centre of the thorax behind **Ren-17** (*danzhong/shanzhong*). This sea contains the water from all the rivers (e.g. the vessels), resembling an inexhaustible reservoir that does not overflow but redistributes all its water supplies.

Defensive Qi (*wei qi*)

The *wei qi* is the body's **Defensive Qi**. Thus it defends the body at the level of the skin, the fascia and the muscles. In relation to the Nutritive Qi (*Ying qi*), it is a more slippery and less pure form of Qi that moves dynamically and quickly through the body like a **guard**. Its root lies in the **Lower Burner**, where it is produced by the ***mingmen***-Fire (Life Gate fire), so that it has a considerable genetic or constitutional component. Essence (*jing*) and Original Qi (*yuan qi*), which are stored in the Lower Burner (according to many schools in the Kidneys), are involved in the formation of Defensive Qi (*wei qi*), therefore also playing a role in the defence against Exterior pathogens. Additionally, the *wei qi* is constantly being replenished with pure essence, formed from food by the Spleen and Stomach in the **Middle Burner**. It is dispersed throughout the body by the **Upper Burner**. Proper functioning of the *wei qi* therefore depends on all three Burners.

Nutritive Qi (*ying qi*)

The Nutritive Qi (*ying qi*) is the result of a purification and/or distillation process of pure, clear origin. Compared to Defensive Qi (*wei qi*), it is more Yin in nature. It **nourishes** the whole body and is its '**building substance**', thus often translated as 'constructive Qi' or 'building energy'.

Circulation of Defensive Qi (*wei qi*) and Nutritive Qi (*ying qi*)

Circulation of Defensive Qi (*wei qi*)

In chapter 43 of the *Su Wen* the *wei qi* is described as flowing 'outside of the *mai*' (the channels and vessels). However, it circulates in part along the channel pathways, moving in the space between the skin and the muscles – the ***cou li***.

The *cou li* – often unsatisfactorily translated as ‘pores’ – are the striae or compartments between the skin and the muscles. They have the function of a gateway for the entry and exit of Qi and fluids and serve as protection against the invasion of Exterior pathogenic factors. According to Larre and Rochat de la Vallée (1986), the *cou li*, as the outermost of the wrappings of the Triple Burner, cover the whole body. They connect the surface of the body with its inner organs. This explains the effectiveness of acupuncture and other manual therapies for the treatment of internal disorders as well as showing up the presence of diseases of the inner organs on the Exterior of the body.

In the **superficial layers of the body**, the Defensive Qi (*wei qi*) circulates through the skin and the superficial musculature, warming, nourishing and strengthening them. Through these actions it supports the general defence against Exterior pathogenic factors, acting mainly in the realm of the sinew channels (*jing jin*) (→ 1.4). In the **deep layers of the body**, it plays an important role in the functioning of the ‘diaphragm’. According to Larre and Rochat de la Vallée (1986), this represents not only a membranous barrier between the thorax and the abdomen, but can be regarded as a ‘sac of membranes’, including and connecting the peritoneum, the pleura and the pericardium. Nielsen (1995) interprets this ‘network of caves’ as the **inner aspect of the Triple Burner**, which connects with its **outer aspect, the *cou li***. Following this interpretation, the Defensive Qi (*wei qi*) would also be involved with the **mesenteric defence** and the protection of the inner organs.

According to chapter 75 of the *Ling Shu*, the **circulation of the Defensive Qi (*wei qi*)** is cyclical, changing from day to night and vice versa. At dawn, when the Yin Qi is exhausted, the Yang Qi pours itself from the eyes and the eyes are opened. As a result, the *wei qi* rises from the heel via the (Yin) *qiao mai* upwards to the eye to **BL-1** (*jingming*) and flows into the whole body like a waterfall by following the six great Yang channels – ‘it moves 25 times in the Yang’. At dusk, when the Yang Qi is exhausted, the *wei qi* enters the Interior of the body and moves ‘25 times in the Yin’, following the controlling cycle (*ke* cycle → 8.2.5): from the Kidneys to the Heart, from the Heart to the Lungs, from the Lungs to the Liver, from the Liver to the Spleen, and back to the Kidneys. Therefore the *wei qi* has a relationship with the sleep–wake rhythm: during sleep it withdraws deeper into the body – during daytime it circulates in the outer layers of the body. The two extraordinary vessels, the *yin qiao mai* and the *yang qiao mai* (→ 1.7, → Chapter 5), play an important role in this respect. If their cycle is blocked, the circulation of the *wei qi* will be interrupted and disorders will arise. The *yin qiao mai* ascends, while the *yang qiao mai* descends, both meeting at the eyes at **BL-1** (*jingming*) and forming a cycle, similar to the small heavenly cycle of the *ren mai* and *du mai*. If there is an excess of Yang Qi, this will be drained from the channels into the *yang qiao mai*, ‘the eyes cannot close’ and insomnia will occur. If there is an excess of Yin Qi, this will be drained from the channels into the *yin qiao mai*, ‘the eyes cannot open’ and somnolence will occur. In both cases, a relative disharmony develops between the two opponents.

Circulation of the Nutritive Qi (*ying qi*)

The Nutritive Qi (*ying qi*) circulates in the *mai*. The *mai* include both the primary channels (*jing mai*) and the secondary channels such as the Connecting Vessels (*luo mai*, *sun luo*), the divergent channels (*jing bie*), the extraordinary vessels and the Blood Vessels. Wherever there are channels and vessels, the *ying qi* will circulate. According to chapter 16 and 18 of the *Ling Shu*, the *ying qi* circulates constantly and continuously.

First circulation of the *ying qi* (→ Fig. 1.5)

The first circulation of the *ying qi* encompasses the 12 primary channels. The cycle **starts** at the **primary LU channel** and **ends** at the **primary LIV channel**, which in turn connects with the primary LU channel, thus closing the cycle (→ Fig. 1.5). The **Middle Burner** provides and **distributes** the *ying qi* throughout the whole body. According to some of the classics, the Grain Qi (*gu qi*) derived from food is only transformed into Nutritive Qi (*ying qi*) in the Lungs, not already in the Middle Burner. Since the internal pathway of the primary Lung channel begins in the Middle Burner, these two statements do not contradict each other in principle.

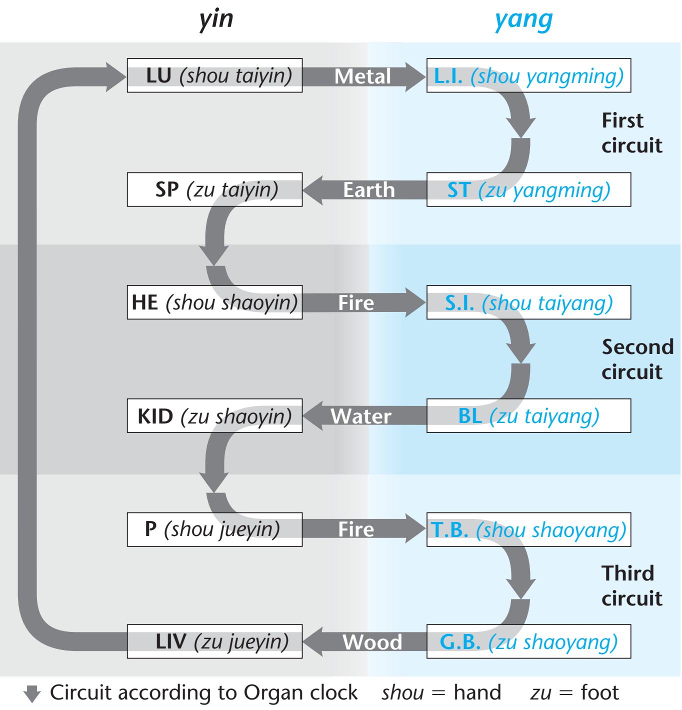

Fig. 1.5 First circulation of the Nutritive Qi (*ying qi*)

The **changes in Yin/Yang polarisation** – the changes from a Yin channel to a Yang channel and vice versa – always take place in the hand or the foot (also → 1.2.2). For example, the Yang energy potential rises from the Yin channel to the Yang channel, and then descends again from the Yang channel to the Yin channel (→ Fig. 1.6).

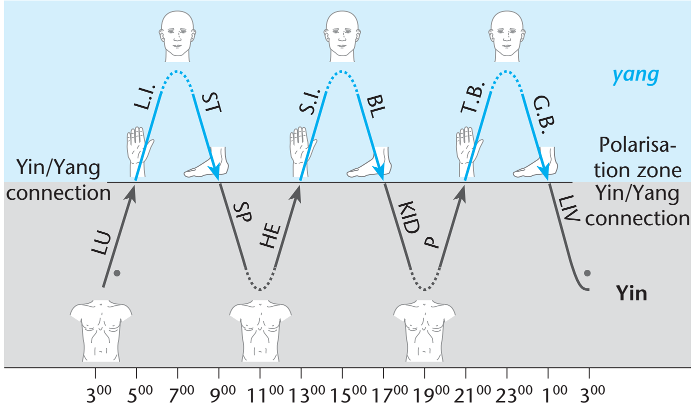
yang
L.I.
ST
S.I.
BL
T.B.
G.B.
Yin/Yang connection
Polarisation zone
Yin/Yang connection
LU
SP
HE
KID
P
LIV
Yin
3^{00} 5^{00} 7^{00} 9^{00} 11^{00} 13^{00} 15^{00} 17^{00} 19^{00} 21^{00} 23^{00} 1^{00} 3^{00}

Fig. 1.6 Changes in Yin/Yang polarisation, shown as a sinus curve

First circulation of the Nutritive Qi (*ying qi*) and the Organ clock

The Nutritive Qi (*ying qi*) circulates in the 12 primary channels in a **circadian 24-hour rhythm** (→ Fig. 1.7) with each **Chinese hour**, named after one of the 'earthly branches', corresponding to two Western hours (→ expanded Organ clock, Fig. 1.8).

Each two-hour slot favours a particular channel, in other words: for two (Western) hours per day, **energy flow peaks** in a particular channel. During this period, the respective channel Qi increases, falling again during the following two hours. However, it will never fall below a particular level, so that there is always a certain portion of energy flowing continuously through the channels.

Example: The channel Qi of the primary Stomach channel will begin to appear more 'powerful', when the channel Qi of the Large Intestine is very strong. Therefore the 'high tide' of the Stomach channel takes place between 5–7 a.m., its peak is between 7–9 a.m. and its 'low tide' between 9–11 a.m. During the peak period of the Stomach channel, the channel on the opposite side of the Organ clock is at its lowest point (→ 8.3.7); while the Stomach channel is peaking, the Pericardium channel is at its lowest point.

In figure 1.8 the 'expanded Organ clock' with its 24-hour circulation through the 12 primary channels is shown in relation to the earthly branches and the hexagrams of the *Yijing*.

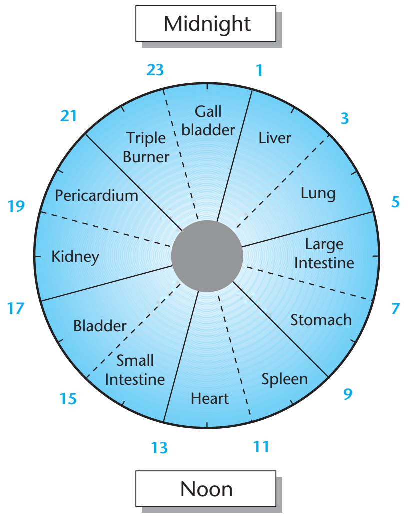
Midnight
23
1
3
5
7
9
11
13
15
17
19
21
Gall bladder
Liver
Lung
Large Intestine
Stomach
Spleen
Heart
Small Intestine
Bladder
Kidney
Pericardium
Triple Burner
Noon

Fig 1.7 Organ Clock

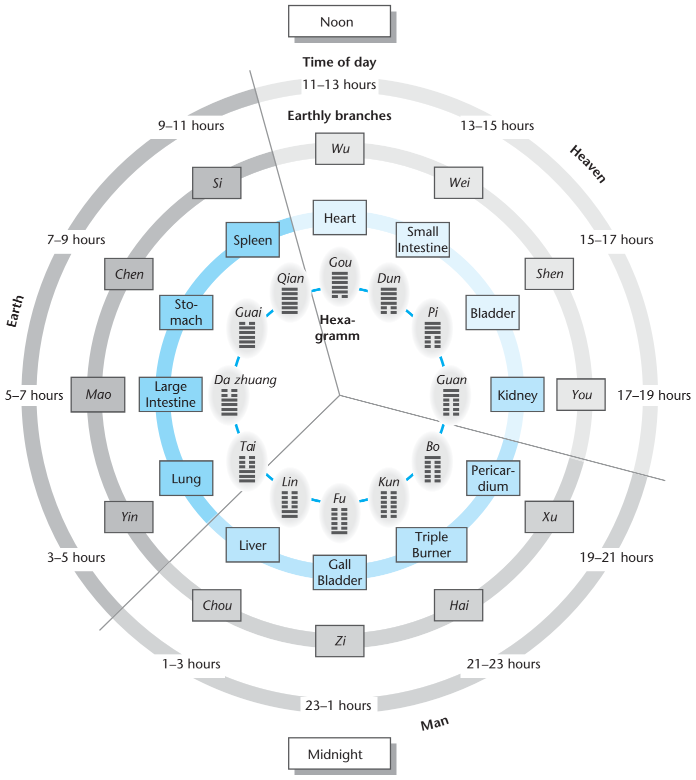
Noon
Time of day
11–13 hours
Earthly branches
9–11 hours
13–15 hours
Heaven
Wu
Si
Wei
Heart
Spleen
Small Intestine
7–9 hours
15–17 hours
Chen
Shen
Gou
Qian
Dun
Sto-mach
Guai
Pi
Bladder
Earth
Hexa-gramm
5–7 hours
Mao
Large Intestine
Da zhuang
Guan
Kidney
You
17–19 hours
Tai
Bo
Lung
Lin
Fu
Kun
Pericar-dium
Yin
Xu
Liver
Triple Burner
3–5 hours
Gall Bladder
19–21 hours
Chou
Hai
Zi
1–3 hours
21–23 hours
23–1 hours
Man
Midnight

Fig. 1.8 Expanded Organ clock

Second circulation of the Nutritive Qi (*ying qi*)

According to chapter 16 of the *Ling Shu*, the Nutritive Qi (*ying qi*), after having circulated through the 12 primary channels, takes a deep pathway of the primary Liver channel, which originates at **LIV-14** (*qimen*), passes the Lungs, the posterior aspect of the neck, the inner nasal passages and reaches **Du-20** (*baihui*), then follows along the *du mai* on the back and the *ren mai* along the midline of the abdomen. At the approximate level of **Ren-22** (*tiantu*) the *ying qi* again reaches the Lung channel, which traverses the supraclavicular fossa (near **ST-12**). Here, a new circulation of the *ying qi* through the 12 primary channels starts at the beginning of the primary LU channel.

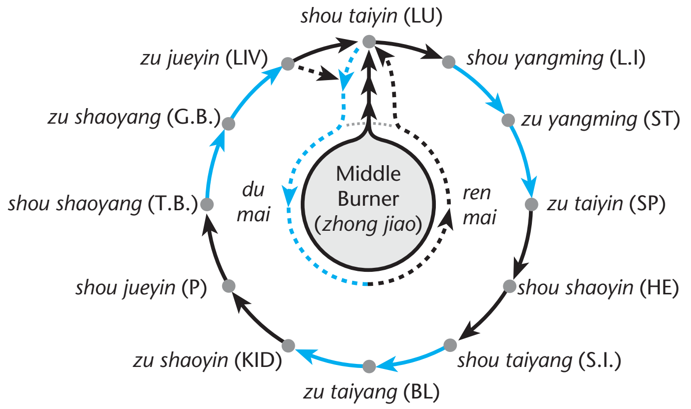

Fig. 1.9 First and second circulation of the Nutritive Qi (*ying qi*) with the Middle Burner

1.2 The 12 Primary Channels (*jing zheng*)

1.2.1 Background Information

Synonyms: Meridians, regular channels. '*zheng*' can be translated as 'mainly' (primary channels), but also relates to terms such as 'straight' and 'direct'.

Chinese terms for the channels

Meanings (also → 1.2.4, the six great channels (*liu jing*)

Tai: means greater, highest. The polarised energy (Yang or Yin) develops and reaches the maximum extent of its activity in the *taiyang* and *taiyin* channels. Here, the maximum extent of their respective polarity is reached and begins to fall again.

Shao: means less or younger. The polarised energy (Yang or Yin) is less in the *shaoyang* and *shaoyin* channels than in the *taiyang* or *taiyin* channels.

Ming: means clear, bright, radiant, shining (for more detail see → 1.2.4, the six great channels).

Jue: means 'absolute', 'at the end', 'exhausted', and according to Wiseman also 'inverted'. The polarised energy (Yin) in the *jueyin* channel represents the terminal phase of the Yin; here polarity changes to Yang.

Composition of the channel names

In the Chinese language, the name for each channel comprises

- the nature or **polarity of the energy** (Yin or Yang) that flows in the channel
- the **quality** or **intensity** (*tai*, *shao*, *jue*, *ming*) of the Yin or Yang energy and
- the **extremity** at which the channel originates or terminates.

Thus the Yin or Yang channels that begin or terminate at the hand carry the name of the primary hand channels (***shou*** *jing zheng*). The Yin and Yang channels that terminate or begin at the feet carry the name of the foot primary channels (***zu*** *jing zheng*) (→ Fig. 1.2).

1.2.2 Communication and Connections

Principle of the primary channel system

The 12 primary channels cover the body **bilaterally**. Each channel has its **individual regular course** with a deep, **internal** and a more superficial **external** pathway.

One can distinguish between Yin and Yang channels, which are Interiorly/Exteriorly paired. While each channel is connected with its pertaining *zang* or *fu* Organ, it is also connected with the Organ of its Interiorly/Exteriorly coupled channel.

All hand Yin channels begin in the region of the thorax and flow to the hand. All hand Yang channels start at the hands and travel to the head, where they meet the foot Yang channels. These descend down to the toes, where they meet the foot Yin channels, which ascend to the thorax and there meet the hand Yin channels.

For a better understanding of this model, imagine a person standing with his/her hands raised to the sky. In this position, all **Yin channels are ascending** (a **Yang phenomenon**), while all **Yang channels are descending** (a **Yin phenomenon**) (→ Fig. 1.10).

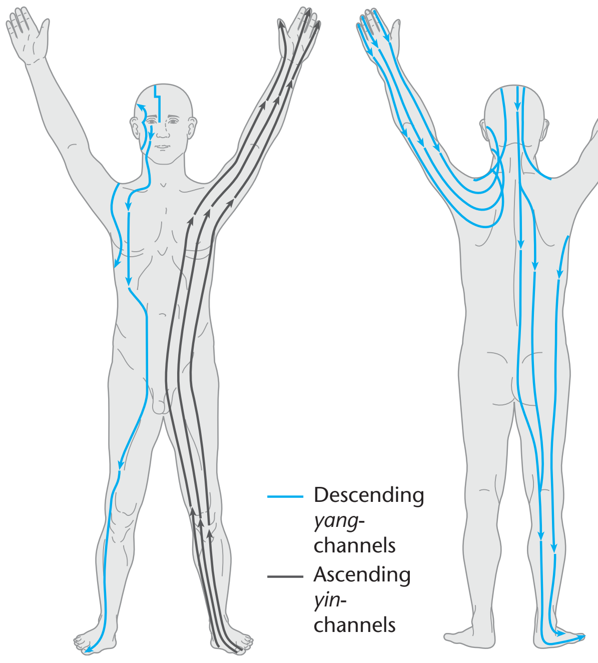

Fig. 1.10 Ascending and descending primary channels (according to Pirog 1996)

Each channel is associated with its own specific pathological symptoms, which are an important diagnostic tool in clinical practice (for channel-specific pathologies see → chapter 4).

Communication between channels and *zangfu*-Organs

The **connections (anastomoses)** between the channels have the function of allowing communication within the channel system as well as between the channels and the Organ systems. This external/internal (*biao-li*) relationship comprises the following aspects:

Communication from the Interior to the Exterior

- The **internal pathways of the hand Yin channels** connect with their pertaining Organ as well as with their Yin/Yang-paired *fu*-Organ. They flow from the thoracic region to the Exterior and connect with their paired Yang channel at the finger tips or on the hand (→ Fig. 1.11 a).
- The **internal pathways of the foot Yang channels** connect with their pertaining *fu*-Organ as well as with their Yin/Yang-paired *zang*-Organ. They then flow from the centre of the body to its Exterior and connect with their paired Yin–channel at the foot (→ Fig. 1.11 b).

Communication from the Exterior to the Interior

- **Each hand Yang channel** begins on the hand, runs to the thoracic region, penetrates the thorax and connects in the Interior of the body with its pertaining *fu*-Organ and its Yin/Yang-paired *zang*-Organ (→ Fig. 1.11 c).
- **Each foot Yin channel** starts at the foot, penetrates the abdomen, reaches the Interior of the body and its pertaining *zang*-Organ as well as its Yin/Yang-paired *fu*-Organ (→ Fig. 1.11 d).

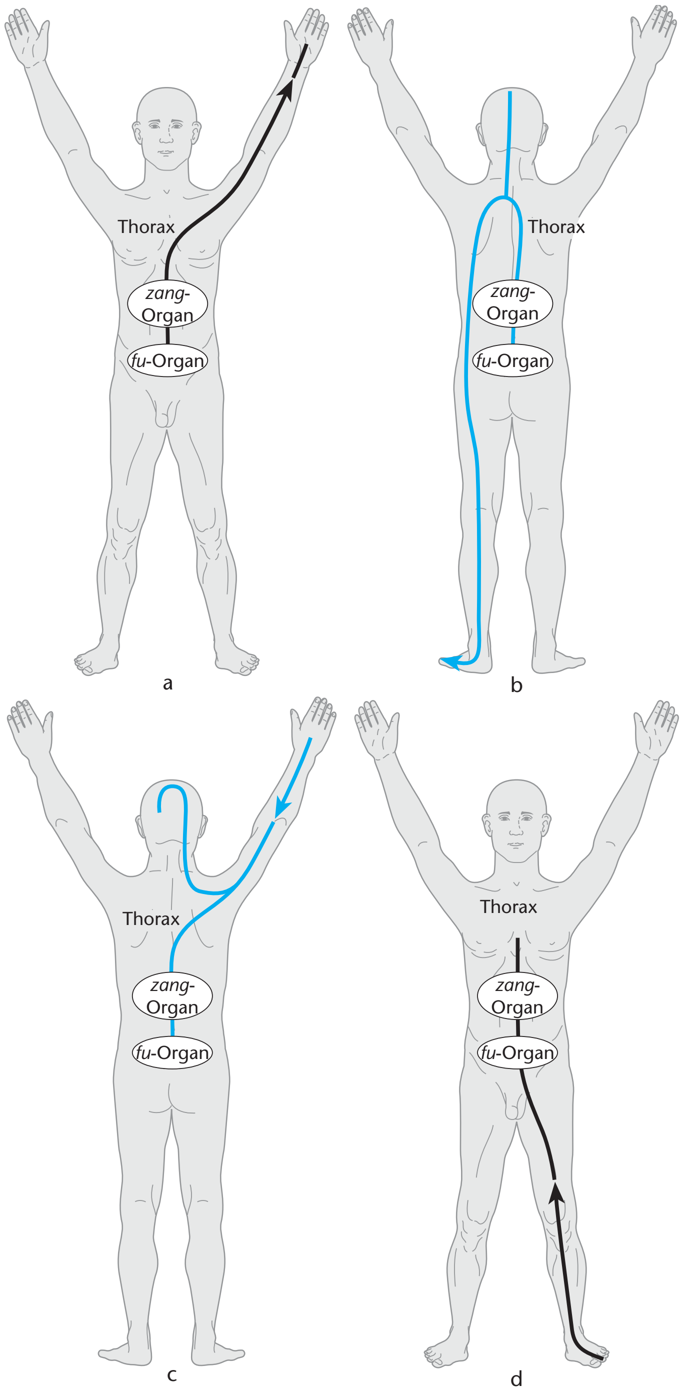

Fig. 1.11 Communication Interior → Exterior, hand-Yin channels (a). Communication Interior → Exterior, foot-Yang channels (b). Communication Exterior → Interior, hand-Yang channels (c). Communication Exterior → Interior, foot-Yin channels (d).

Channel circuits

Within the primary channel system we can distinguish three circuits, each circuit comprising four primary channels. The trajectories of two of the channels in one circuit are located on the Yin side of the body, the more anterior side (Interior), and two are found on the Yang side, the more posterior side of the body (Exterior). The Stomach channel (foot *yangming*, for more detail ➞ 1.2.3) presents the exception to this rule. Despite its major part running along the anterior side of the body, it is considered a Yang channel (➞ Fig. 1.12 and 1.3).

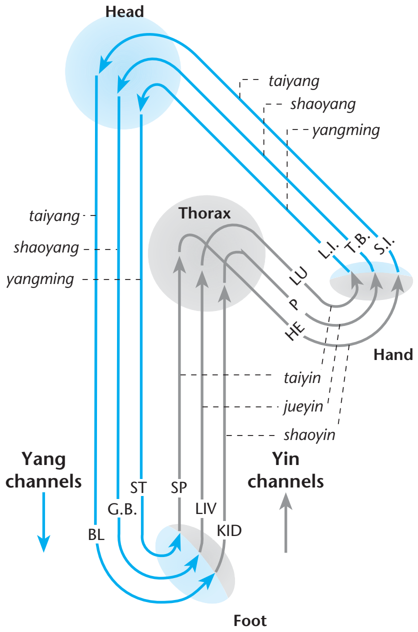
Head
taiyang
shaoyang
yangming
taiyang
shaoyang
yangming
Thorax
L.I.
T.B.
S.I.
LU
P
HE
Hand
taiyin
jueyin
shaoyin
Yang channels
Yin channels
ST
SP
G.B.
LIV
BL
KID
Foot

Fig. 1.12 Principle of the channel circuits

		First circuit			
Channel	**Time**	**Starts**	**Terminates**	**Hand/foot**	
LU	3–5 a.m.	Thorax	Finger tips	Hand (*shou*)	*taiyin*
L.I.	5–7 a.m.	Finger tips	Face	Hand (*shou*)	*yangming*
ST	7–9 a.m.	Face	Tips of the toes	Foot (*zu*)	*yangming*
SP	9–11 a.m.	Tips of the toes	Thorax (Heart)	Foot (*zu*)	*taiyin*
		Second circuit			
H	11 a.m.–1 p.m.	Thorax	Finger tips	Hand (*shou*)	*shaoyin*
S.I.	1–3 p.m.	Finger tips	Face	Hand (*shou*)	*taiyang*
BL	3–5 p.m.	Face	Tips of the toes	Foot (*zu*)	*taiyang*
KID	5–7 p.m	Tips of the toes	Thorax (Pericardium)	Foot (*zu*)	*shaoyin*
		Third circuit			
P	7–9 p.m.	Thorax	Finger tips	Hand (*shou*)	*jueyin*
T.B.	9–11 p.m.	Finger tips	Face	Hand (*shou*)	*shaoyang*
G.B.	11 p.m.–1 a.m.	Face	Tips of the toes	Foot (*zu*)	*shaoyang*
LIV	1–3 a.m.	Tips of the toes	Thorax (Lung)	Foot (*zu*)	*jueyin*

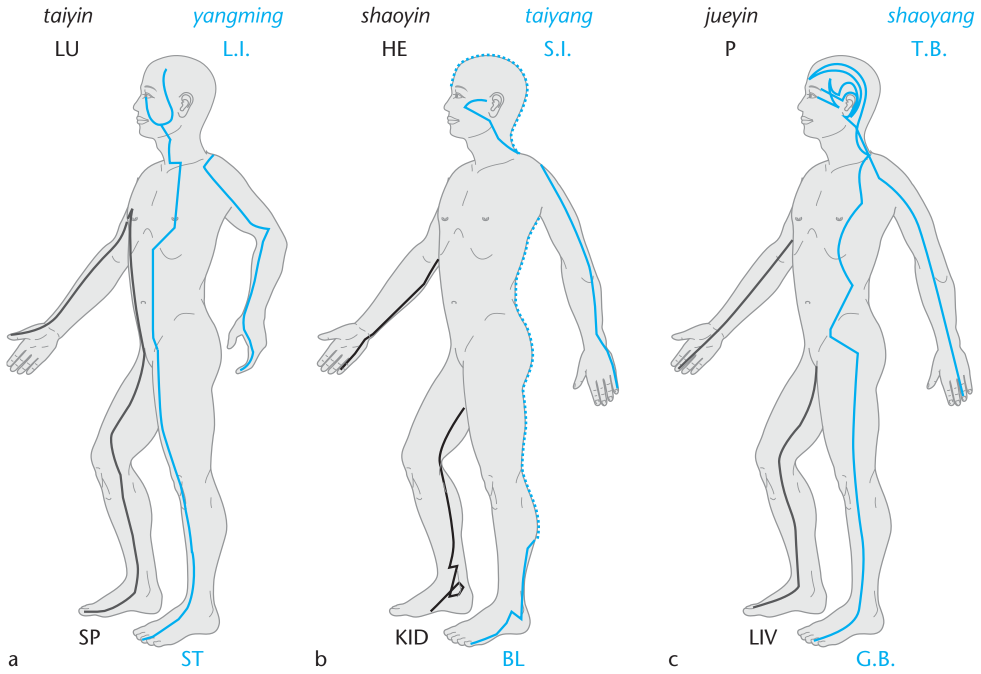

Fig. 1.13 Overview of the channel circuits

Connections among the primary channels

In order to ensure the continuous circulation (also see → 1.1.4) within the primary channel system, there need to be connections among the individual primary channels. Fig. 1.5 illustrates these in a schematic graphic overview.

According to some schools, the connections among the primary channels are facilitated by so-called entry/exit points. Various authors (for example Worsley, Jarrett, Pirog, Hicks et al., for more detail → 8.1.16) describe these as the shunting points between successive channels according to the Organ clock. The exit point marks the point on a channel from which the (internal) flow moves to a point (the entry point) on the channel that succeeds it on the Organ clock. There are some discrepancies regarding the location of the connections as described by the above authors and those described by Solinas et al. (1998), which are marked below as appropriate.

Yin–Yang connections

The connections – anastomoses – between the Yin and Yang channels are located on the hands and feet. The channels are most commonly connected at their respective final points, but sometimes also by branches separating from the main channel. These Yin–Yang connections are places where the Qi can change its polarity: Yin Qi becomes Yang Qi, and Yang Qi becomes Yin Qi. This creates a dynamic balance and allows for a better circulation of the Qi. The Yin–Yang connections follow the order of the basic structure of the circadian rhythm (24-hour cycle).

Hand Yin–Yang connections (→ Fig. 1.14)

The **hand Yin channels** are in most cases connected to the **hand Yang channels** at the tips of the fingers; sometimes they are connected on the hand by a branch that separates from the primary Yin channel and runs to the hand Yang channel:

FIRST CIRCUIT

LU channel (*shou taiyin*) ➡ L.I. channel (*shou yangming*)

LU-7 (*lieque*) ➡ **L.I.-1** (*shangyang*)

A branch separates from the primary LU channel at → **LU-7** (*lieque*) and travels to the primary L.I. channel at → **L.I.-1** (*shangyang*).

Note: According to Hicks et al. (2004, → 8.1.16) **L.I.-4** (*hegu*) is considered to be an entry point.

SECOND CIRCUIT

HE channel (*shou shaoyin*) ➡ S.I. channel (*shou taiyang*)

HE-9 (*shaoyang*) ➡ **S.I.-1** (*shaoze*)

THIRD CIRCUIT

P channel (*shou jueyin*) ➡ T.B. channel (*shou shaoyang*)

P-8 (*laogong*) ➡ **T.B.-1** (*guanchong*)

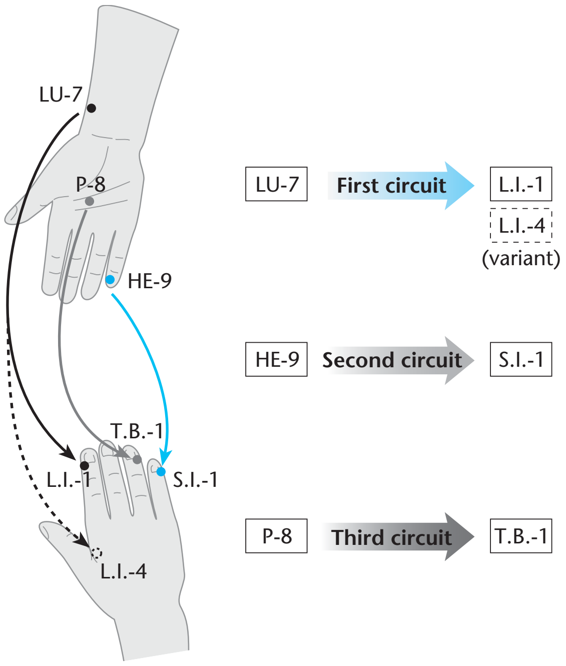
LU-7
P-8
HE-9
T.B.-1
L.I.-1
S.I.-1
L.I.-4
LU-7
First circuit
L.I.-1
L.I.-4
(variant)
HE-9
Second circuit
S.I.-1
P-8
Third circuit
T.B.-1

Fig. 1.14 Hand Yin–Yang connections (primary channels)

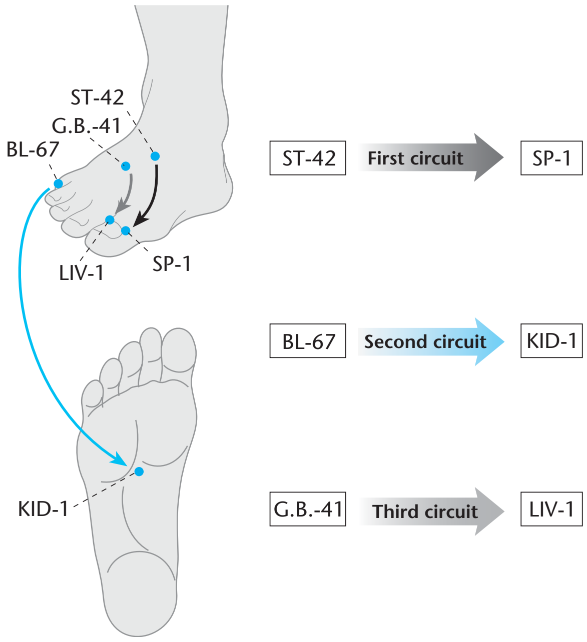
ST-42
G.B.-41
BL-67
LIV-1
SP-1
KID-1
ST-42
First circuit
SP-1
BL-67
Second circuit
KID-1
G.B.-41
Third circuit
LIV-1

Fig. 1.15 Foot Yang–Yin connections (primary channels)

Foot Yang–Yin connections (→ Fig. 1.15)

The **foot Yang channels** are connected to the **foot Yin channels** at the toes by branches that separate from the primary Yang channels and run to the foot Yin channels (→ Fig. 1.15):

First circuit

ST channel (*zu yangming*) ➡ SP channel (*zu taiyin*)

ST-42 (*chongyang*) ➡ **SP-1** (*yinbai*)

At → **ST-42** a branch separates from the primary ST channel and travels to → **SP-1**.

Second circuit

BL channel (*zu taiyang*) ➡ KID channel (*zu shaoyin*)

BL-67 (*zhiyin*) ➡ **KID-1** (*yongquan*)

Third circuit

G.B. channel (*zu shaoyang*) ➡ LIV channel (*zu jueyin*)

G.B.-41 (*zulinqi*) ➡ **LIV-1** (*dadun*)

On the instep of the foot, at → **G.B.-41**, a branch separates from the G.B. channel and travels to → **LIV-1**

Yang–Yang connections (→ Fig. 1.16)

The Yang–Yang connections between the primary channels are located on the **head**. They are rather superficial and follow the order of the channels within the **circadian cycle** (according to the Organ clock → Fig. 1.7). These connections support the communication between the hand Yang and foot Yang primary channels in the cranial region, which convey Qi of the same nature and quality (*yangming*, *taiyang*, *shaoyang*). They also mark the connections between the Yang axes, also referred to as hand–foot pairing (→ 1.2.3), and serve as a link between **above** and **below**.

First circuit

L.I. channel (*shou yangming*) ➡ ST channel (*zu yangming*)

L.I.-20 (*yingxiang*) ➡ **ST-1** (*chengqi*)

Second circuit

S.I. channel (*shou taiyang*) ➡ BL channel (*zu taiyang*)

S.I.-18 (*quanliao*) ➡ **BL-1** (*yingming*)

Third circuit

T.B. channel (*shou shaoyang*) ➡ G.B. channel (*zu shaoyang*)

T.B.-23 (*sizhukong*) ➡ **G.B.-1** (*tongziliao*). Comment: according to Hicks et al. (2004, → entry/exit points, → 8.1.16) **T.B.-22** (*erheliao*) is considered an exit point.

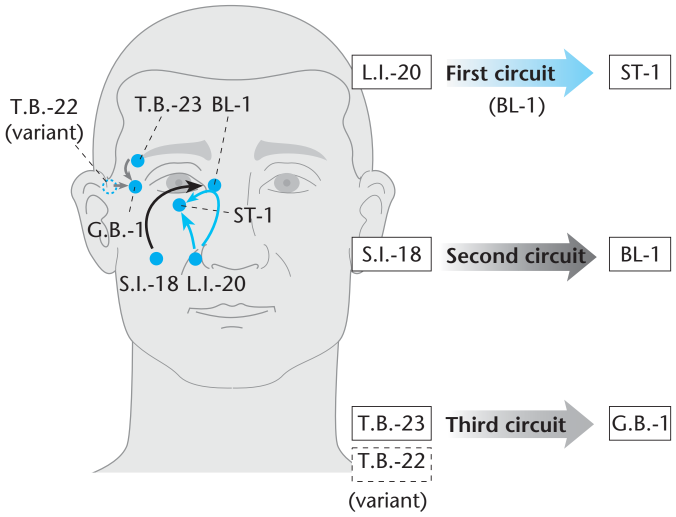
T.B.-22
(variant)
T.B.-23 BL-1
G.B.-1
ST-1
S.I.-18 L.I.-20
L.I.-20
First circuit
(BL-1)
ST-1
S.I.-18
Second circuit
BL-1
T.B.-23
T.B.-22
(variant)
Third circuit
G.B.-1

Fig. 1.16 Yang–Yang connections of the primary channels

Yin–Yin connections

The Yin–Yin connections between the primary channels are located in the **thoracic region** (Fig. 1.17).

In contrast to the more superficial Yang–Yang connections, they are located deep inside the body (internal pathways of the channels).

There are two types of Yin–Yin connection:

- the Yin axes connections or hand–foot pairings (*taiyin*, *jueyin*, *shaoyin*), which are **non-circadian** (do not flow according to the Organ clock) and
- the (deep) Yin–Yin connections, whose Qi flow follows the circadian rhythm of the Organ clock.

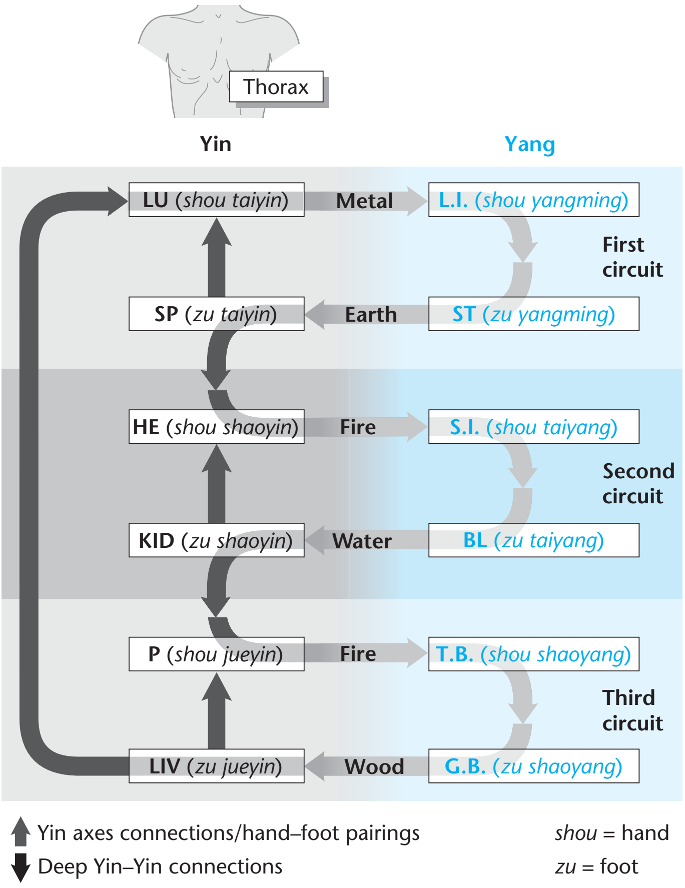
Thorax
Yin
Yang
LU (shou taiyin)
Metal
L.I. (shou yangming)
First circuit
SP (zu taiyin)
Earth
ST (zu yangming)
HE (shou shaoyin)
Fire
S.I. (shou taiyang)
Second circuit
KID (zu shaoyin)
Water
BL (zu taiyang)
P (shou jueyin)
Fire
T.B. (shou shaoyang)
Third circuit
LIV (zu jueyin)
Wood
G.B. (zu shaoyang)
Yin axes connections/hand–foot pairings
Deep Yin–Yin connections
shou = hand
zu = foot

Fig. 1.17 Overview of the Yin–Yin connections (primary channels)

Connections of the Yin axes (hand–foot pairings)

These Yin–Yin connections support the communication between the hand Yin and the foot Yin channels, which convey Qi of the same quality (*taiyin*, *jueyin*, *shaoyin*) and serve as a link between **above** and **below**. Their Qi flow does **not follow the circadian rhythm** of the Organ clock (→ Fig. 1.17).

FIRST CIRCUIT

SP channel (*zu taiyin*) ➡ LU channel (*shou taiyin*)

SP-20 (*zhourong*) ➡ **LU-1** (*zhongfu*)

According to Solinas et al. (1998), a superficial branch separates from the primary SP channel at → **SP-20** and travels to the primary LU channel at → **LU-1**.

SECOND CIRCUIT

LIV channel (*zu jueyin*) ➡ P channel (*shou jueyin*)

LIV ➡ P

An internal branch ascends from the Liver, passes through the diaphragm, connects with the primary P channel below → **P-1** (*tianchi*).

THIRD CIRCUIT

KID channel (*zu shaoyin*) ➡ HE channel (*shou shaoyin*)

KID ➡ HE

An internal branch of the primary KID channel runs to the Liver, penetrates the diaphragm and spreads into the Lungs. From the

Lung a branch travels to the Heart, where it connects with the primary HE channel.
Many authors give **HE-1** (*jiquan*) as the site of connection.

Deep Yin–Yin connections

These Yin–Yin connections facilitate the **circadian cycle** according to the Organ clock and mark the transitions of the channel circuits. They are located deep in the thorax area on the level of the *zang*-Organs.

CONNECTION OF THE FIRST AND SECOND CIRCUIT

SP channel (*zu taiyin*) ➡ HE channel (*shou shaoyin*)

SP ➡ HE

An internal branch of the primary SP channel spreads in the Heart and connects with the HE channel. Note: according to Hicks et al. (2004, → entry/exit points, → 8.1.16) **SP-21** is considered to be the exit point of the SP channel and **HE-1** the entry point of the HE channel.

CONNECTION OF THE SECOND AND THIRD CIRCUIT

KID channel (*zu shaoyin*) ➡ P channel (*shou jueyin*)

KID ➡ P

An internal branch of the primary KID channel travels to the Kidneys, then to the Liver, penetrates the diaphragm and spreads into the Lung. From the Lung, an internal branch runs to the Heart, where it meets the primary P channel and also reaches → **Ren-17** (*shanzhong*/*danzhong*). Note: according to Hicks et al. (2004, → entry/exit points → 8.1.16) **KID-22** is considered to be the exit point of the KID channel and **P-1** the entry point of the P channel.

CONNECTION OF THE THIRD AND FIRST CIRCUIT

LIV channel (*zu jueyin*) ➡ LU channel (*shou taiyin*)

LIV ➡ P

An internal branch of the LIV channel originates in the Liver, passes the diaphragm, spreads in the Lungs and connects with the LU channel through a network of branches. Note: according to Hicks et al. (2004, → entry/exit points → 8.1.16) **LIV-14** is considered to be the exit point of the LIV channel and **LU-1** the entry point of the LU channel.

1.2.3 The Six Great Channels (*liu jing*)

Based on the Yang–Yang and Yin–Yin connections (→ 1.2.2, → Fig. 1.16, → Fig. 1.17), primary channels of the same nature and quality can be divided into **six great channels** (*liu jing*) or **hand–foot pairings (or axes).** They represent the three great Yang channels and the three great Yin channels.

Channels	**Trajectory**		**Amount of Qi and Blood**
Exterior			
Yang channels (hand–foot pairings)			
taiyang **(S.I., BL)**	Latero-posterior aspect of the four extremities, posterior aspect of the head and body	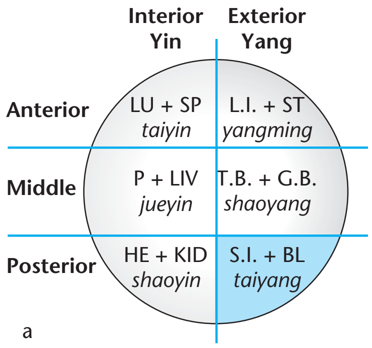 Interior Yin Exterior Yang Anterior LU + SP *taiyin* L.I. + ST *yangming* Middle P + LIV *jueyin* T.B. + G.B. *shaoyang* Posterior HE + KID *shaoyin* S.I. + BL *taiyang* a	Less Qi, more Blood **Qi < Blood**
shaoyang **(T.B., G.B.)**	Lateral aspect of the four extremities and torso	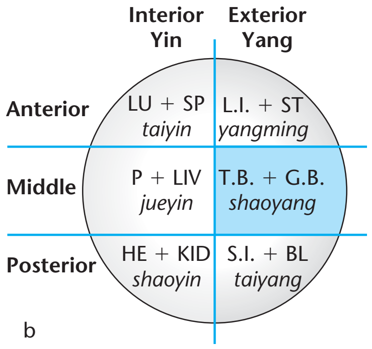 Interior Yin Exterior Yang Anterior LU + SP *taiyin* L.I. + ST *yangming* Middle P + LIV *jueyin* T.B. + G.B. *shaoyang* Posterior HE + KID *shaoyin* S.I. + BL *taiyang* b	More Qi, less Blood **Qi > Blood**

Fig. 1.18 Overview of the six great channels (*liu jing*)

Yang channels (hand–foot pairings)			
yangming **(ST, L.I.)**	Latero-anterior aspect of the four extremities, frontal aspect of the head (face, forehead), anterior aspect of the trunk	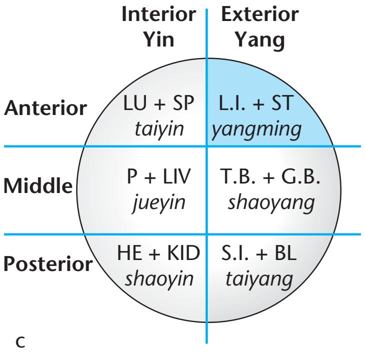 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP *taiyin* L.I. + ST *yangming* P + LIV *jueyin* T.B. + G.B. *shaoyang* HE + KID *shaoyin* S.I. + BL *taiyang* c	More Qi, more Blood **Qi = Blood**
Yin channels (hand–foot pairings)			
taiyin **(LU, SP)**	Medio-anterior aspect of the four extremities as well as the thorax and abdomen	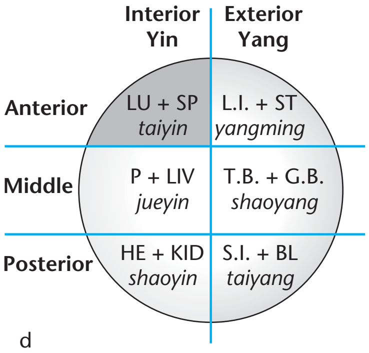 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP *taiyin* L.I. + ST *yangming* P + LIV *jueyin* T.B. + G.B. *shaoyang* HE + KID *shaoyin* S.I. + BL *taiyang* d	More Qi, less Blood **Qi > Blood**
jueyin **(P, LIV)**	Middle of the medial aspect of the four extremities as well as the thorax and abdomen	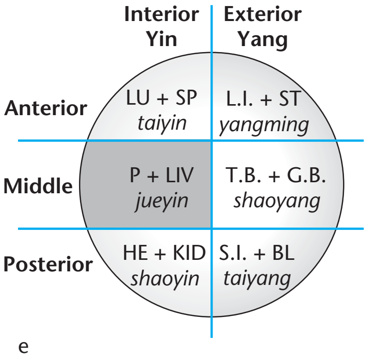 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP *taiyin* L.I. + ST *yangming* P + LIV *jueyin* T.B. + G.B. *shaoyang* HE + KID *shaoyin* S.I. + BL *taiyang* e	Less Qi, more Blood **Qi < Blood**
shaoyin **(HE, KID)**	Medio-anterior aspect of the four extremities as well as the thorax and abdomen	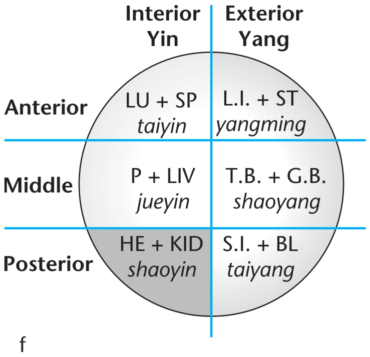 Interior Yin Exterior Yang Anterior Middle Posterior LU + SP *taiyin* L.I. + ST *yangming* P + LIV *jueyin* T.B. + G.B. *shaoyang* HE + KID *shaoyin* S.I. + BL *taiyang* f	More Qi, less Blood **Qi > Blood**
Interior			

Fig. 1.18 (cont'd) Overview of the six great channels (*liu jing*)

Ch01-F10028.qxd 2/22/08 2:30 PM Page 16

Names and functions of the six great channels (*liu jing*)

Tai means 'greater', 'highest'

The polarised energy (Yang or Yin) develops and reaches its maximum activity in the *taiyang* or *taiyin* channels. Here, **the maximum level of the respective polarity** is reached and begins to fall again. '*Taiyang* opens to the outside' means that it spreads to the Exterior (→ see comments below under *yangming*, also → Fig. 1.19).

Shao means 'less' or 'younger'

The polarised energy (Yang or Yin) is less in the *shaoyang* or *shaoyin* channels than in the *taiyang* or *taiyin* channels. This also manifests in their location: the *shaoyang* channels are located midway between the **inside** (or anterior aspect) and the **outside** (or posterior aspect) of the body, acting like a hinge. According to Larre and Rochat de la Vallée (1986), *shaoyang* should be translated as 'young Yang'. In their view it is positioned between the outside (the other two Yang channels) and the inside (the other three Yin channels).

Depending on the interpretation, the *shaoyin* channels are considered either to be located between the *taiyin* and the *jueyin* channels or to be the deepest of the six channels. Thus they are either seen as a hinge or as a fixed central point (for more detail see → Depth organisation of the channels).

Ming means 'clear', 'bright', 'radiant', 'shining'

'*Yangming* closes towards the inside – *yangming* contracts towards the inside.'

The *yangming* channel has some unique features. Based on its course alone, it is **not 'Yang'** like the other two Yang channels, but it develops as a Yang phenomenon within the Yin aspect of the body (partially anterior pathway of the ST channel).

The character for *ming* is composed of the radicals for the sun and the moon. The character for Yang also contains the sun radical. According to the *Su Wen*, the *yangming* is where the two Yang shine together. Thus, the two suns can be symbolically interpreted as a doubling of Yang energy. *Taiyang* and *shaoyang* are 'Yang' simply due to their physical location. *Yangming*, however, is 'Yang' due to the content of its stored energy. This raises the question as to why the *yangming* channel is so abundant, why it carries so much Qi and Blood. Pirog (1996) explains this fact based on embryonic development. The 'closing' of the *yangming*' can be equated with the image of a curled-up **embryo** (→ Fig. 1.19). Because of its contracting, 'closing' position, the energy of the *yangming* channel can be compared to the compressed energy in a pressure cooker. Thus the 'Yin' process of contracting towards the Interior results in producing Yang or, in other words, compression of the energy in the Interior. This naturally compressed energy is utilised in clinical practice. The *yangming* channels, but especially the ST channel, can be accessed, either by massage or by acupuncture, to supply the whole body with the energy stored and accumulated here. By the same token, the *yangming* channels can be needled in order to drain excess Heat – a procedure comparable to a controlled release of steam from a pressure cooker.

Jue means 'absolute', 'at the end', 'exhausted' (also, according to Wiseman, 'to revert')

The polarised energy (Yin) in the *jueyin* channel represents the terminal phase of the Yin; here polarity changes to Yang (→ see below for more detail).

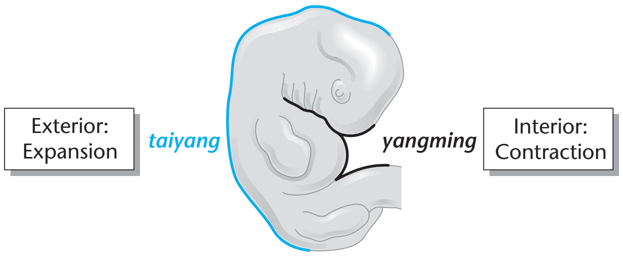

Fig. 1.19 Embryonic development of the *yangming* and *taiyang* channels (according to Pirog 1996)

The six great channels (*liu jing*): relationships and depth organisation

The systemic division of the channels into six levels or stages is mainly associated with the *Shang Han Lun*, a herbal classic. But according to Pirog (1996), this system can also be used as a paradigm for acupuncture, since it provides fundamental ideas regarding the functions of the points and channels as well as the relationships among the channels. It focuses particularly on changes of characteristics of the channel Qi, depending on which level it is acting within the body. The channels can be described as being stacked one upon each other, similar to layered rocks in a quarry. The depth of a channel or a point will be indicative of its function and significance. The postulate is that the more deeply a channel is located in the body, the more deep-seated its functions are and the more deep-seated illnesses it will be able to treat. They are not static, however. Depending on their location in the human body, they are indeed living structures with specific activities such as openers, closers, hinges or pivots, reflected in the meaning of the Chinese names. Fig 1.20 illustrates the arrangement of the six great channels from the outside to the inside, resembling the layers of an onion (→ Fig. 1.20).

Nguyen Van Nghi (1996) compares the three Yin and three Yang channels to doors. The door (*taiyang* or *taiyin*) is closed to prevent the entry of unwelcome guests such as external pathogenic factors. But the door can also be opened to let in friends. Whether the door opens and closes well depends largely on the hinges (*shaoyang*, *shaoyin*). Only if the hinges function smoothly can the energy circulate adequately. The doorframe (or closing mechanism) (*yangming*, *jueyin*) closes towards the Interior; in other words, it protects and supports what is contained on the inside. (→ Fig. 1.21).

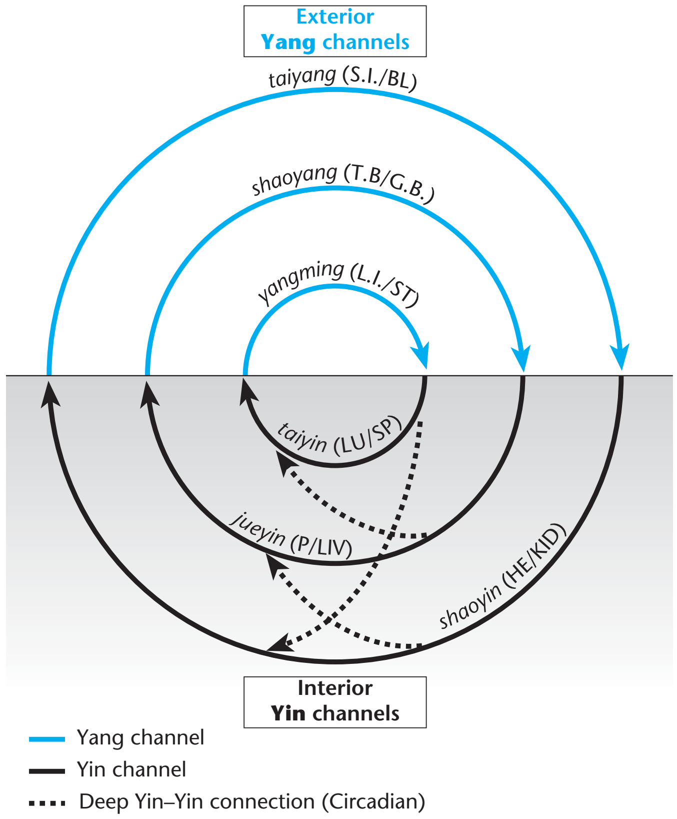
Exterior
Yang channels
taiyang (S.I./BL)
shaoyang (T.B/G.B.)
yangming (L.I./ST)
taiyin (LU/SP)
jueyin (P/LIV)
shaoyin (HE/KID)
Interior
Yin channels
Yang channel
Yin channel
Deep Yin–Yin connection (Circadian)

Fig. 1.20 Relationships of the three Yin and three Yang channels (according to Nguyen Van Nghi, 1996)

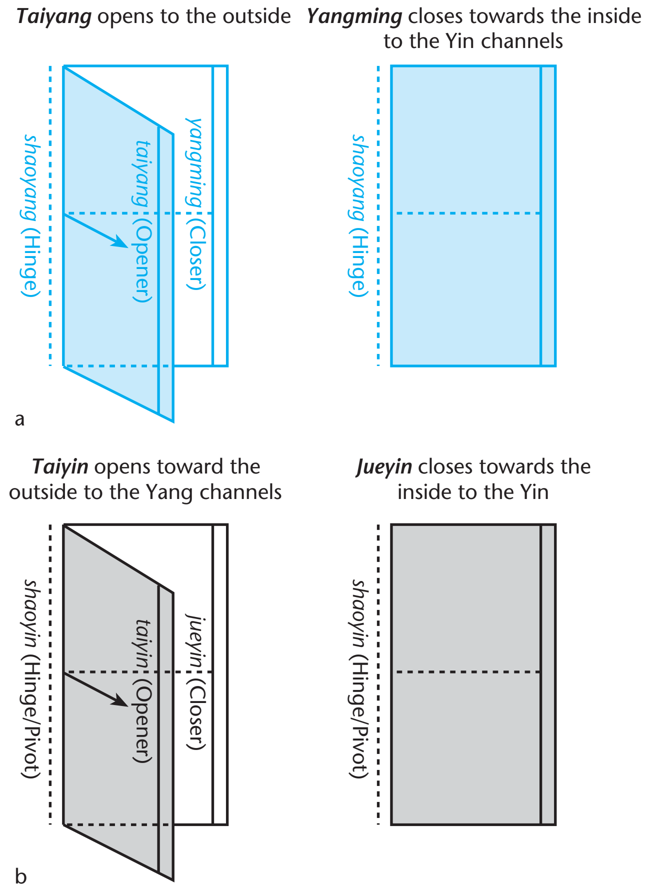
Taiyang opens to the outside
Yangming closes towards the inside to the Yin channels
shaoyang (Hinge)
taiyang (Opener)
yangming (Closer)
shaoyang (Hinge)
a
Taiyin opens toward the outside to the Yang channels
Jueyin closes towards the inside to the Yin
shaoyin (Hinge/Pivot)
taiyin (Opener)
jueyin (Closer)
shaoyin (Hinge/Pivot)
b

Fig. 1.21 The six great channels as opener, hinge and frame of a door

The literature is somewhat contradictory about which are the deeper of the six great channels, the *shaoyin* or the *jueyin* channels. Pirog (1996) has contributed an interesting theory in suggesting that the *shaoyin* channels occupy the deepest position in the body. They are located in such a medial location (near the middle of the body) that, when a person is standing in a normal position, the *shaoyin* channels are covered and not directly visible. According to Pirog, the *shaoyin* does not function like a hinge, but rather like a fixed pivot, similar to the axis of a mill wheel, which itself is immobile, but is able to cause movement. When the Qi has reached the deepest point of the *shaoyin*, and the 'Yin of the *jueyin* is exhausted', it has to revert. In other words: it has to return to a more superficial level on the outside or it has to transform into Yang. In that sense, it is the P and LIV channels that again start the flow of the Qi towards the Yang, from the inside to the outside and from below to above (→ Fig. 1.22).

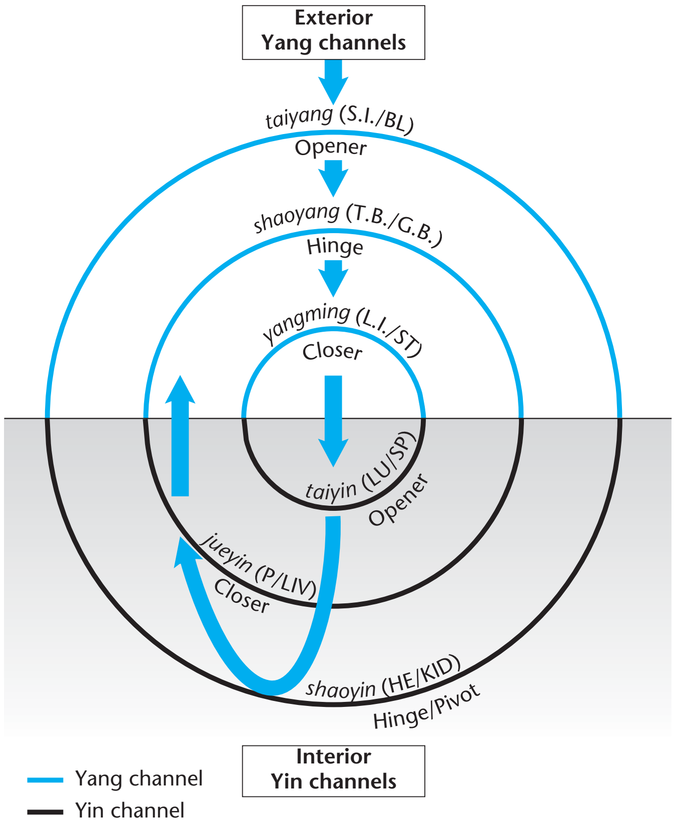
Exterior
Yang channels
taiyang (S.I./BL)
Opener
shaoyang (T.B./G.B.)
Hinge
yangming (L.I./ST)
Closer
taiyin (LU/SP)
Opener
jueyin (P/LIV)
Closer
shaoyin (HE/KID)
Hinge/Pivot
Interior
Yin channels
Yang channel
Yin channel

Fig. 1.22 Depth organisation of the six great channels (according to Pirog, 1996)

Confluent or connecting points of the six great channels (*liu jing*)

The **primary Yin channels** (*liu jing*) connect in the **thoracic region** (see also → Yin axes or hand–foot pairings → 1.2.2, Fig. 1.17). In contrast, the **primary Yang channels** connect in the **cranial region** (see also → Yang axes or hand–foot pairings → 1.2.2, Fig. 1.16). Many schools of thought suggest the connections of the channels of the same nature and quality at the following **six confluent points**:

Yin channels	**LU-1** (*zhongfu*), **P-1** (*tianchi*), **H-1** (*jiquan*)
Yang channels	**BL-1** (*jingming*), **G.B.-1** (*tongziliao*), **ST-1** (*chengqi*)

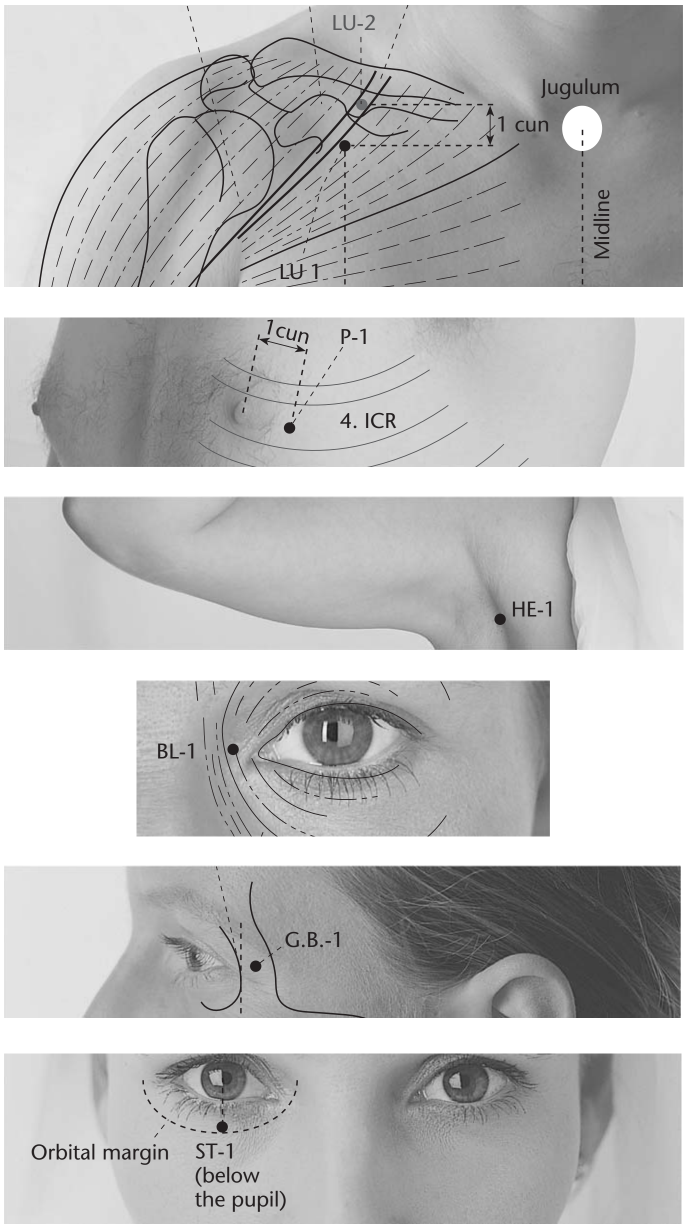
LU-2
Jugulum
1 cun
Midline
LU 1
1 cun
P-1
4. ICR
HE-1
BL-1
G.B.-1
Orbital margin
ST-1
(below
the pupil)

Fig. 1.23 Confluent points of the six great channels

Starting and end points of the six great channels (*liu jing*)

Fig. 1.24 shows the starting and end points of the six great channels on the feet or hands respectively.

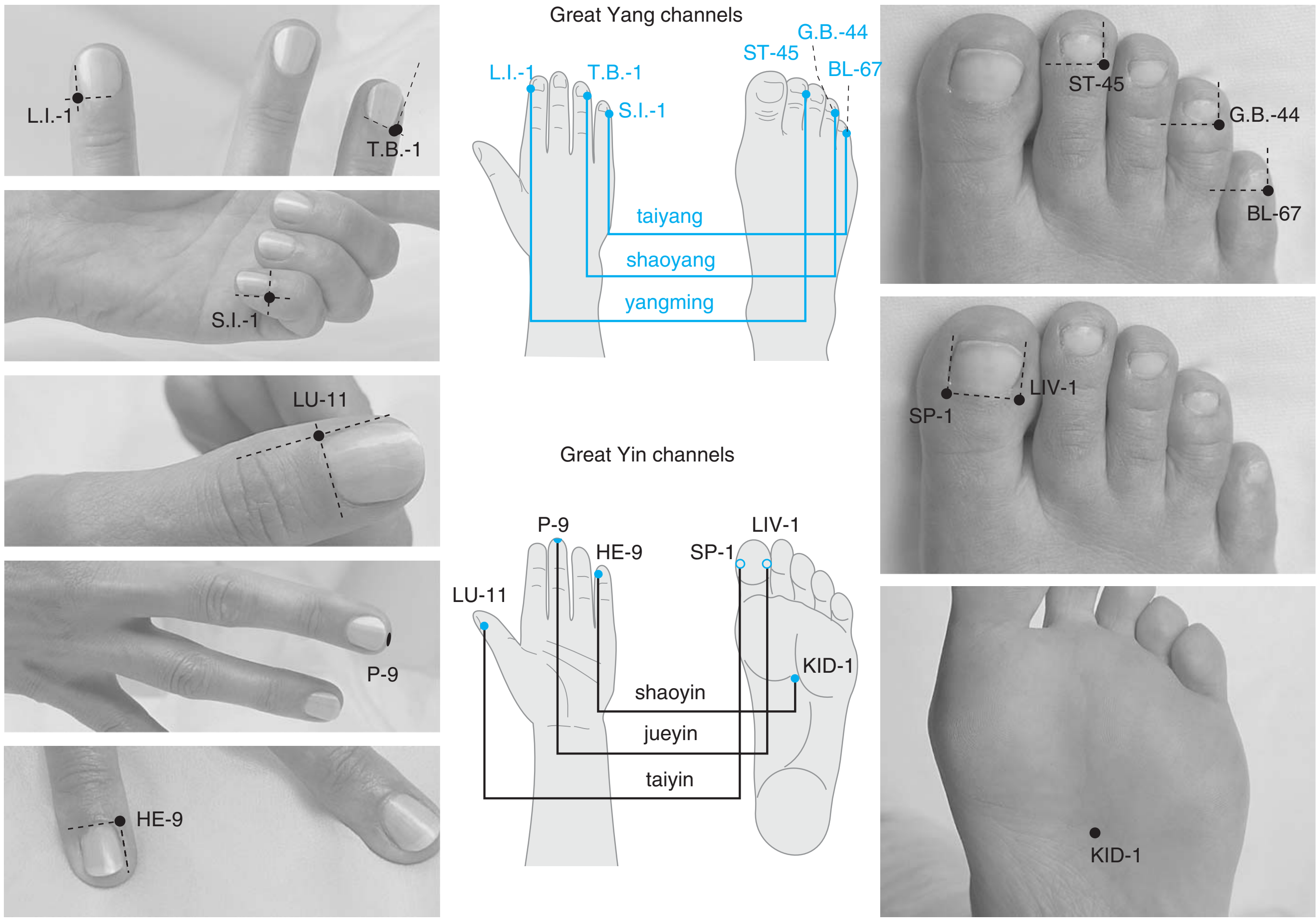
L.I.-1
T.B.-1
S.I.-1
LU-11
P-9
HE-9
Great Yang channels
L.I.-1
T.B.-1
S.I.-1
ST-45
G.B.-44
BL-67
taiyang
shaoyang
yangming
Great Yin channels
P-9
HE-9
LU-11
SP-1
LIV-1
KID-1
shaoyin
jueyin
taiyin
ST-45
G.B.-44
BL-67
SP-1
LIV-1
KID-1

Fig. 1.24 Starting and end points of the *jiu jing*

Root points and binding points of the six great channels (*liu jing*)

According to Chapter 5 of the *Ling Shu*, each of the six great channels (*liu jing*) has a root point (*gen*) and a binding point (*jie*).

Root points (*gen*)

At the root point, the polarised energy (Yin/Yang) of a great channel is at its minimum – this point marks the end of the Yang or the beginning of the Yin. The root points are all located on the foot (→ Fig. 1.24).

Root points of the three Yang channels (→ Fig. 1.24)

These always represent the **terminal point** of each great **Yang** channel.

taiyang (S.I. ➡ BL)	**BL-67** (*zhiyin*)
shaoyang (T.B. ➡ G.B.)	**G.B.-44** (*zuqiaoyin*)
yangming (L.I. ➡ ST)	**ST-45** (*lidui*)

Root points of the three Yin channels (→ Fig. 1.24)

These always represent the **first point** of each great **Yin** channel.

taiying (SP ➡ LU)	**SP-1** (*yinbai*)
jueyin (LIV ➡ P)	**LIV-1** (*dadun*)
shaoyin (KID ➡ H)	**KID-1** (*yongquan*)

Binding points (*jie*)

The binding point is the point at which the polarised energy (Yin/Yang) of a great channel reaches its maximum level. It is at this point that the hand channel and the foot channel are connected.

Binding points of the three Yang channels

These are the **switching points** of the great Yang channels (hand to foot) and are located in the **cranial region**.

taiyang (BL ➡ S.I.)	**BL-1** (*jingming*)
shaoyang (G.B. ➡ T.B.)	**T.B.-21** (*ermen*)*
yangming (ST ➡ L.I.)	**ST-1** (*chengqi*)**

Binding points of the three Yin channels

These are the **switching points** of the great Yin channels (foot to hand) and are located in the **thoracic or abdominal region**.

taiying (SP ➡ LU)	**Ren-12** (*zhongwan*)
jueyin (LIV ➡ P)	**Ren-18** (*yutang*)
shaoyin (KID ➡ H)	**Ren-23** (*lianqian*)

1.3 Divergent Channels (*jing bie*)

Synonyms: Channel divergence (Wiseman), separate pathways (Larre and Rochat de la Vallée).
Character: *bie* means diverging, separating, branching off. In this sense, the 12 divergent channels can be seen as branches of the primary channels.

General pathways of the divergent channels (➞ Fig. 1.25)

The 12 bilateral divergent channels, which have no specific points of their own, are named after and are dependent on their corresponding primary channel.
They can be divided into **six great Yin and six great Yang pairs** and have their own system of six confluences (reunions): BL/KID, G.B./LIV, ST/SP, S.I./H, T.B./P, L.I./LU. According to Ramakers (course material, 2003) the divergent channels follow their own (non-cosmic) daily biorhythm:
BL ➡ G.B. ➡ ST ➡ S.I. ➡ T.B. ➡ L.I. ➡ KID ➡ LIV ➡ SP ➡ H ➡ P ➡ LU.
All divergent channels separate (*li*) from their corresponding primary channel on the extremities, except for the divergent T.B. channel, which branches off on the head. But many authors suspect that their pathways begin before the actual branching off. Based on this assumption, the divergent channels begin at the distal ends of the primary channels, running parallel to them, but as separate entities. The circulation in the divergent channels is of a centripetal nature (➞ 1.1.1); in other words, the flow is always from a distal in a proximal direction, towards the trunk and the head. After separating from the primary channels, they enter deeper (*ru*) into the body and, after completing their internal pathway, they re-emerge close to the surface of the body (*chu*). The divergent channel pairs (Yin/Yang) pass together through the Interior of the body to join (*he*) their pertaining primary Yang channel in the occipital or cranial region. It is this joining of the courses that is referred to as the six confluences (or reunions) (*liu he*).

Pathways of the divergent Yin channels

After separating from their primary channels, they connect with their pertaining *zang*-Organ, then with the Interiorly-Exteriorly paired *fu*-Organ. They continue their flow in the upper half of the body together with their paired divergent Yang channel to join the paired primary Yang channel. In contrast to the divergent Yang channels, the divergent Yin channels will not return to their pertaining primary Yin channel after having completed their internal pathway (➞ Fig. 1.25).

Pathways of the divergent Yang channels

After separating from their primary channels, the divergent Yang channels travel to their pertaining *fu*-Organ, then to the Interiorly-Exteriorly paired *zang*-Organ. They connect again with their pertaining primary channel and also with their Interiorly-Exteriorly paired divergent Yin channel in the upper half of the body (➞ Fig. 1.25).

The divergent channels and the Heart

All divergent channels pass the thoracic region and the Heart, except for the divergent channels of the LU, L.I. and KID. According to Shima and Chase (2001), this means that all divergent channels passing the thorax will communicate with the True Qi (*zhen qi*), representing the synthesis of the Gathering Qi (*zong qi*) and Original Qi (*yuan qi*) (➞ 1.1.4). They suggest that this hypothesis is in accordance with all the information available about the divergent channels, and explains furthermore why the divergent channels are so useful in the treatment of deep-seated Organ disorders as well as for disharmonies of the Defensive Qi (*wei qi*). It supports the image of a divergent channel as a fundamental axis between the Interior and the Exterior of the body. In the opinion of Shima and Chase (2001), the divergent channels are an invaluable connection to facilitate the return of the Defensive Qi (*wei qi*) to the Interior of the body and to transfer the fundamental inner Original Qi (*yuan qi*) to the periphery. In that respect the Original Yang (*yuan yang*) not only touches upon, but also strengthens the Defensive Yang (*wei-yang*).

---

* According to Nguyen Van Nghi (1996) and Solinas et al. (1998), this is G.B.-2 (*tinghui*)

** According to Nguyen Van Nghi (1996) and Solinas et al. (1998), this is ST-8 (*touwei*)

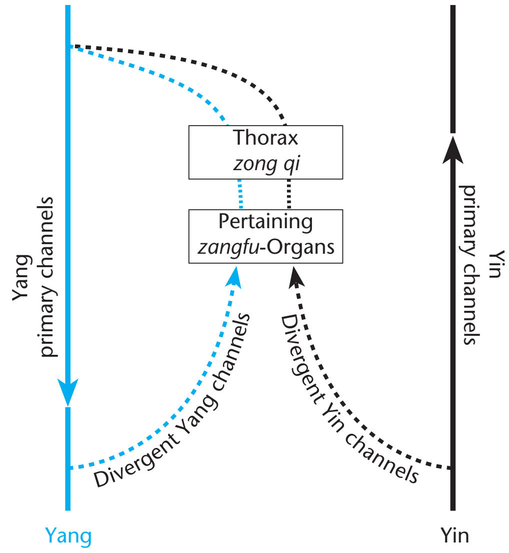

Fig. 1.25 Schematic illustration of the divergent channel pathways (according to Shima and Chase 2001)

Functions

The divergent channels reinforce the circulation of and communication between the primary channels in the Interior of the body. On the one hand, this is achieved due to their connection with the primary Yang channels (→ pathways). But their organisational structure also supplements the primary channels in a functional way – for example they also control areas of the body not covered by the primary channels. Thus points on the primary Yin channels can affect the head and face, in spite of their 'external' pathways ending in the thoracic region. This effect can be explained by the fact that the Yin divergent channels, after branching off from their pertaining primary Yin channel, pour into the Yang divergent channels flowing to the head. A deeper understanding of the trajectory and functions of the divergent channels will in turn lead to a better understanding of the clinical application of some commonly used acupuncture points. Since their pathway passes the Heart and thorax, the divergent channels further strengthen the connection to the Heart (→ explanations above). According to some schools, they can therefore also be used for the treatment of psychosomatic disorders. Many interpretations and variants exist regarding the pathways and clinical application of the divergent channels (for a good overview see → Shima and Chase 2001).

1.4 The 12 Sinew Channels (*jing jin*)

Synonyms and translations: muscle channels, muscular meridians/the muscles in connection with the meridians (Larre and Rochat de la Vallée), muscle regions, channel sinews.

Principles of the sinew channel system

The sinew channels mainly represent the muscles, sinews and ligaments located along the pathways of the primary channels and their pertaining connecting vessels. They do not have their own specific points and also no direct connection to the *zangfu*-Organs. The 12 bilateral sinew channels are connected to and named after their pertaining primary channel. Each sinew channel has its own pathway, which generally follows the trajectory of its pertaining primary channel. However, in contrast to the primary channels, the sinew channels always begin at the extremities, mostly separating from the primary channels at the *jing*-well points (→ 8.1.6) on the tips of the fingers and the toes. Their pathways always run from the extremities to the trunk or further to the head and face (→ centripetal circulation → 1.1.1). They cover a larger area than the primary channels, in the shape of a band and binding (*jie*) like a spindle at the joints and other areas of the body. This allows them to cover areas of the body which are not reached by the primary channels or the divergent channels. It further explains the actions of some of the points located on their pertaining primary channel. Under pathological conditions, the sinew channels will manifest as so-called *ashi*-points (*a shi xue*) or as trigger points.

Terms

Binding (*jie*) refers to spindle-shaped concentrations of the sinew channels, which tend to be located on the bigger muscles and joints (for more details on specific sinew channels see → Chapter 4).
The term *ju* (point of confluence, gathering) is used to signify an area where two or more sinew channels are bundled together. In the larger muscles, the Qi disperses or spreads (*san*), so that they can be divided into smaller muscle groups. The sinew channels connect (*luo*) the muscles, sinews and ligaments with the joints in order to facilitate mobility.

Rhythm of Qi flow

According to Ramakers (course material, 2003) the sinew channels are subject to their own daily rhythm of Qi flow:
BL ➡ G.B. ➡ ST ➡ S.I. ➡ T.B. ➡ L.I. ➡ SP ➡ LU ➡ LIV ➡ P ➡ KID ➡ HE.

General pathways of the sinew channels (→ Fig. 1.26)

- The three hand Yang sinew channels (L.I./S.I./T.B.) flow from the finger tips towards the scapula, ascend to the ear and intersect with **G.B.-13** (*benshen*)
- The three foot Yang sinew channels (ST/BL/G.B.) run from the tips of the toes towards the head, join each other on the cheek and meet with **S.I.-18** (*quanliao*)
- The three hand Yin sinew channels (HE/P/LU) flow from the finger tips towards the thorax, joining within the thoracic musculature and meeting at **G.B.-22** (*yuanye*)
- The three foot Yin sinew channels (LIV/KID/SP) run from the tips of the feet towards the thorax, join each other in the genital region and, depending on the author, intersect in the area of **Ren-2** (*qugu*) or **Ren-3** (*zhongji*)

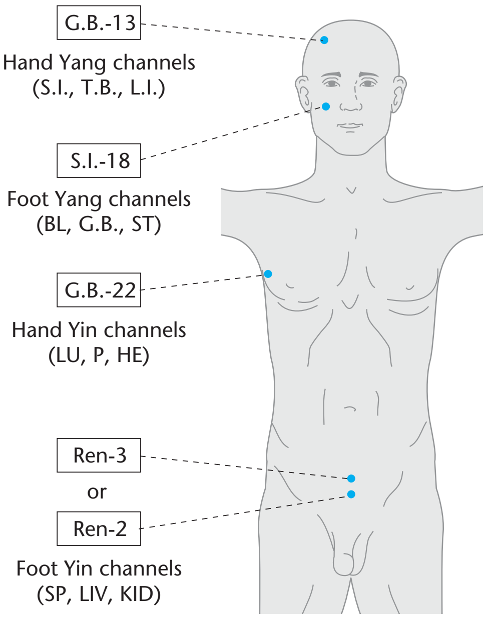

Fig. 1.26 Meeting points of the sinew channels

Functions

The main function of the sinew channels is to distribute Qi and Blood (*xue*) over the surface of the body as well as to integrate the muscles, sinews and ligaments with the joints. Generally, the sinew channels join groups of synergistic muscles, supporting their mobility and natural movement while also serving the function of integrating the surface of the body. At the same time they protect the bones and skeletal structure, linking the structures of the body.

Proper functioning of the muscles is dependent on the Spleen (*pi*) and the Liver (*gan*) and particularly on their Blood (*xue*). While the Yin (or material) aspect of the musculature, the 'flesh', is governed by the nourishing function of the Spleen, the functioning of the muscles and sinews is governed by the Liver. But the Liver is also connected to the mesenchymal structures of the body, especially those surrounding the organs. The sinew channels are not only composed of the muscular structures, but also of the mesenchymal formations, so that according to Larre and Rochat de la Vallée they are also involved in the structural aspect of the diaphragm as well as the thoracoabdominal serous membranes, in particular the pleura, peritoneum and pericardium. In the deeper parts of the body they act through the deep muscular fascia. One of the main functions of the sinew channels is that of a 'protective coat' that shields the body against attacks from the Exterior. In this respect the Defensive Qi (*wei qi*) is of particular importance ((→ 1.1.4): the muscle spindles are crossed by loose and diffuse flows of Qi, which also contain Defensive Qi (*wei qi*). Nutritive Qi (*ying qi*) is also present, particularly at the outer ends of the spindles (→ 1.1.4).

The sinew channels also have the task of barring access to the deeper channels. In other words, in a pathological situation, a deeper channel can only be reached if the resistance of the muscular layers can be broken. An important aim of Tuina therapy is to remove muscular tension. According to Larre and Rochat de la Vallée (1996), the sinew channels also play a role in psychosomatic disorders since they very effectively constitute the muscular carapace. The authors refer here to a bioenergetic interpretation of muscular tension.

Clinical and therapeutic importance

Disorders of the sinew channels will show the following symptoms along their pathway: muscular tension, pain, cramping, spasms, stiffness, but also paralyses, weakness and swellings of the muscles, tendons and ligaments. Any impairment of range of motion of the joints also indicates that the sinew channels are involved. In addition, each sinew channel also has its specific symptoms (→ chapter 4, individual channel descriptions).

The following factors can contribute to disorders of the sinew channels:

- *Bi* syndromes – external invasion of pathogenic Wind, Cold, Damp or Heat
- Traumatic injuries
- Muscle strain through over-use
- Muscle strain and contraction due to longstanding emotional and mental stress

While the sinew channels do not have their own specific points, they can easily be accessed by needling *ashi*-points, as well as through adjunctive therapies such as cupping, Tuina massage, plum blossom needling, moxibustion and *gua sha* (scraping technique). Distal points can also be used in treating disorders of the sinew channels.

For traumatic injuries with limited range of motion, needling of distal points with simultaneous movement of the affected joint has proven to be very effective. The distal points (→ 8.2.1) can be chosen by first determining the affected area, then needling points in the corresponding area on the contralateral side.

For example, for a knee injury, the exact location of the pain is determined, then the corresponding point or area on the contralateral elbow is needled or massaged. Fig. 1.27 illustrates the areas that have been found to be effective. These areas are also referred to as reciprocal areas (for more detail also see → 8.2, 8.3).

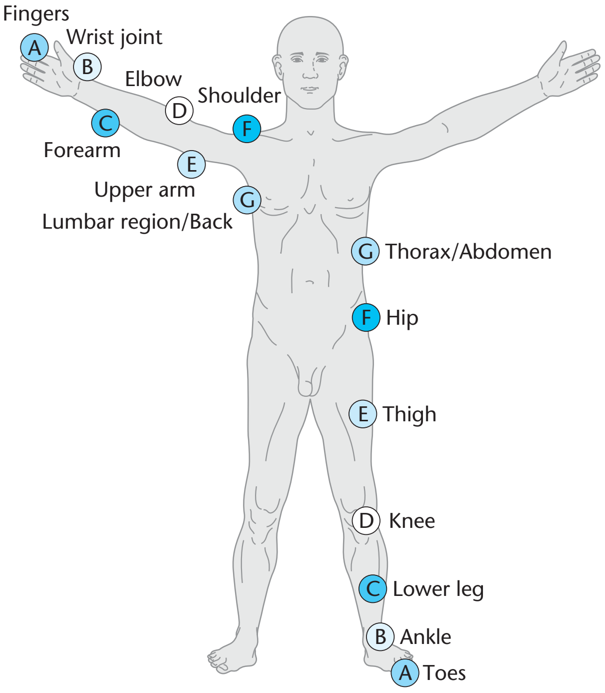

Fig. 1.27 Corresponding areas of the body

1.5 *Luo*-Connecting Vessels

Synonyms: secondary pathways (Larre and Rochat de la Vallée), network vessels, collaterals

Organisation

The *luo*-connecting vessels can be divided into 15/16 great *luo*-connecting vessels as well as into many smaller and more superficial *luo*-connecting vessels. Looking at the *luo*-connecting vessels altogether, they form a network that spans the entire body. And all *luo*-connecting vessels in turn connect with the great *luo*-connecting vessel of the Spleen (at → **SP-21**), which plays a major role within the network of the *luo*-connecting vessels (*Ling Shu*, Chapter 10). Together with the primary channel system, they form a network of vessels which distribute Qi and Blood not only within the body, but also to its surface (→ Fig. 1.3, 1.4).

The 16 great *luo*-connecting vessels

Traditionally there are 15 *luo*-connecting vessels, but if one includes the great *luo*-connecting vessel of the Stomach (*wei zhi da luo* or *xu li*), it raises the number to 16:

- 12 *luo*-connecting vessels (*luo mai*), related to the 12 primary channels and beginning at the extremities
- 4 *luo*-connecting vessels beginning at the thorax:
  - 2 *luo*-connecting vessels of the *ren mai* and *du mai*
  - 1 great *luo*-connecting vessel of the Spleen
  - 1 great *luo*-connecting vessel of the Stomach.

Except for the Great *luo*-connecting vessel of the Stomach (*xu li*) which originates in a *fu*-Organ, all other 15 *luo*-connecting vessels emerge at a specific point on their pertaining primary channel, the *luo*-connecting point (→ 8.1.2). 13 *luo*-connecting points are located on primary channels, two on extraordinary vessels (*ren mai* and *du mai*). The *luo*-connecting points are located on areas of the body, where the *luo*-connecting vessels that connect the Interiorly-Exteriorly pertaining channels with each other divide into smaller vessels. The 15/16 *luo*-connecting vessels thus fulfil a controlling function over all the other, smaller *luo*-connecting vessels. Each *luo*-connecting vessel has its own pathology according to its pertaining primary channel (→ introductions to the primary channels → Chapters 4 and 5).

Superficial *luo*-connecting vessels

The smaller superficial *luo*-connecting vessels are sub-branches of the 'great' *luo*-connecting vessels. They, in turn, are divided into the *sun luo*, vertical branches which are sometimes also referred to as tertiary branches or grandchild-vessels. The *sun luo* not only spread over the surface of the body, but also cover the internal Organs. The *sun luo* further divide into the more superficial *fu luo*. The *fu luo* again divide horizontally into small branches called the *xue luo* (blood vessel network).

Pathogenic factors tend to enter the body through the superficial *luo*-connecting vessels. By the same token, disorders such as Qi or Blood stagnation in a primary channel or an Organ may manifest in the *xue luo* as skin discolorations, spider veins, etc.

Clinical importance

The *luo*-connecting vessels can be accessed through superficial needling techniques as well as skin-sensitising techniques such as plum blossom needling, bloodletting, cupping, *gua sha* and moxibustion.

New diseases are located in the primary channels; old (chronic) diseases lodge in the *luo*-connecting channels.

1.6 Cutaneous Regions (*pi bu*)

Characteristics

The cutaneous regions form the most superficial aspect or outer cover of the primary channel system (*pi* = skin). According to Ramakers (course material, 2003), they can be compared to a 'big bag that holds everything together' and – besides the extraordinary vessels – they represent the oldest tissue in the body, which developed from the ectoderm. The cutaneous zones are not channels *per se*, but rather areas of skin connected to and covering the network of superficial channels and vessels. They are named after the six great channels (*liu jing* ➞ 1.2.3) (Fig. 1.28).

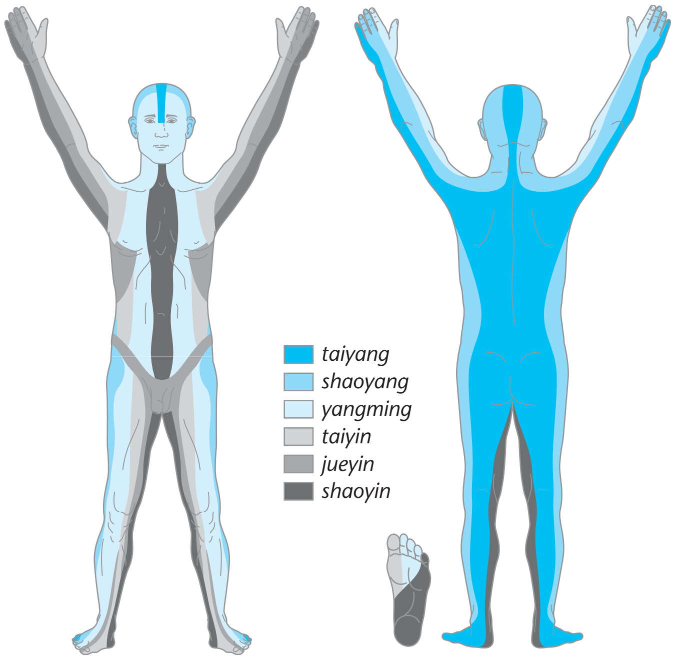

Fig. 1.28 Cutaneous zones (anterior and posterior aspect)

Functions

The cutaneous zones circulate Blood and Qi, especially Defensive Qi (*wei qi*), to the surface of the body and regulate the skin and pores, thus strengthening the body's defences. Proper functioning depends on the Lung Qi, which governs the skin, and the Defensive Qi (*wei qi*) (➞ 1.1.4). Pathogenic factors can enter and exit the body through the cutaneous zones.

Clinical importance

Disorders of the primary channels and their pertaining *zangfu*-Organs may be reflected on the outside of the body by the cutaneous zones, for example pathological skin sensations, lesions or discolorations may occur. A bluish-greenish discoloration may indicate pain or obstructions, a reddish colour Heat and a pale-whitish colour deficiency or Cold. Therefore, the cutaneous zones can serve as a diagnostic tool for skin disorders or pain, specifically for chronic pain. Treatment techniques include superficial needling, plum blossom needling, cupping, massage, and *gua sha* (Chinese scraping technique). The author had good results in the treatment of chronic pain with the somatopic wrist/ankle acupuncture developed by Dr Zhang Xin Shu (➞ JCM, No. 37,11/1991, ➞ Focks and Hillenbrand 2003). The strictly subcutaneous needling used in this technique influences, of course, the cutaneous zones.

1.7 The Eight Extraordinary Vessels (*qi jing ba mai*)

Synonyms: Extraordinary meridians, Wondrous meridians
Meaning of *qi jing ba mai*:
Qi: extraordinary, strange, not paired, miraculous, wonderful
Jing: channels, meridians
Ba: eight (first differentiation)
Mai: vessels

The eight extraordinary vessels are not channels such as the *jing mai*, but they are – as the name implies – extraordinary. While they are often mentioned together with the primary channels (*jing mai* ➞ Fig. 1.3), they still play a special role.

1.7.1 Theories about the Origin of the Extraordinary Vessels

The eight extraordinary vessels are often considered to be the deepest and most original structures of the body and channel system, developing very early on during the embryonic phase. They are considered to 'cover all of human activity' (Larre et al 1986). They also initiate all secondary functions and always remain operative. Therefore, their pathologies manifest with complex symptoms and include pre- and postnatal Essence problems. The *Nanjing* (Unschuld's translation, 1986) is the first to describe the extraordinary vessels as a complete system. Earlier records only mention them individually in the context of their respective functions.

Origins of the extraordinary vessels

In order for the manifestations of life to develop, the body requires structures that keep the rhythm of the Qi and its related processes in harmony and balance. The *Nanjing* introduces the concept of an energetic centre in the human body, the 'moving Qi between the Kidneys', which is often seen as connected with the *mingmen* and *qihai dantian* and is subject to the influences of the cosmic as well as of the prenatal and postnatal Qi.

According to Matsumoto and Birch (1986), the 'moving Qi between the Kidneys' represents the *Taiji* of the body, the Great One, the origin of Yin and Yang. It is considered to be the root and origin of all human systems and structures and, according to some authors, it has a connection to the Original Qi (*yuan qi*) (→ 1.1.4). The *chong mai*, as the oldest vessel, is sometimes equated with the 'moving Qi'. The *ren mai* and *du mai* represent the basic division of the Yin and Yang in the body; in other words, they originate from the 'moving Qi' or *chong mai* as Yin and Yang branches. During the 'division', one movement ascends the front of the body (*ren mai*), the other movement the back (*du mai*).

Heaven and Earth have noon and midnight, the body has the *ren mai* and *du mai* as its polar axes. The *du mai*, *ren mai* and *chong mai* have different names, but in the end they are the same and have the same significance. The *chong mai* ensures the inseparability or unity of the *du mai* and *ren mai* – of Yin and Yang. If we tried to separate Yin and Yang, we would have to realise that they are an 'inseparable whole, a unit'.

Therefore, the extraordinary vessels function as a kind of matrix, which is the source of the energetic functioning and the cellular divisions in the body. This matrix is present from the moment of conception and organises the structuring of the embryonic development from the earliest stages.

Explanation of Fig. 1.29 (after Matsumoto and Birch)

In the beginning, there is the Great One, the *Taiji* (first level). At this level of 'no-form', the Great One begins to separate (second level) and becomes Yin and Yang (third level). At the fourth level, Yin and Yang separate twice (two branches):

- One branch **separates in a threefold way** to become the **Triple** Burner (*san jiao*) and its associated forms of Qi. Further separations lead to the formation of the **six** great channels (*liu jing* → 1.2), the **Five** Phases and **their source**, as well as the 12 primary channels.
- The other branch separates at the fourth level to form the **four** body areas (left, right, above, below) and then the **eight** extraordinary vessels.

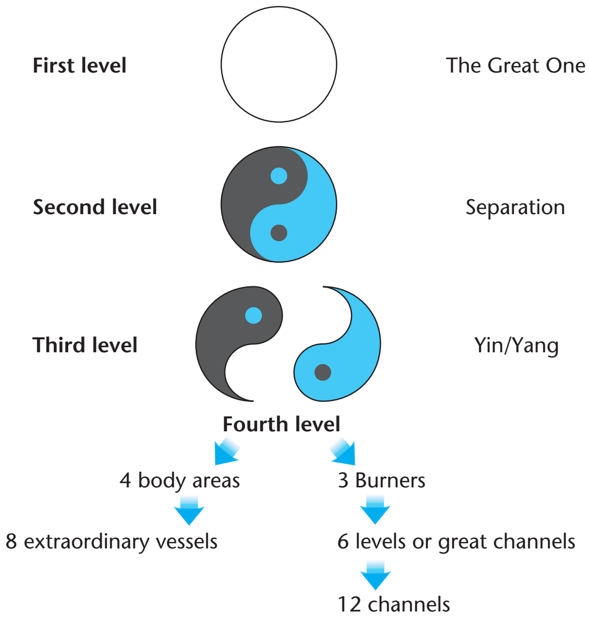

Fig. 1.29 Origin of the channels

Scars of the first cell division

Matsumoto and Birch (1986) comment on the work of Manaka and others, who interpret the extraordinary vessels with regard to embryonic development, regarding them as scars or 'lines of separation of the first cell divisions'.

The *du mai* and *ren mai* are the result of the first division of the fertilised egg, while the *dai mai* develops during the second division. At this stage, the embryo is essentially a torso, while the extremities and the more 'peripheral' extraordinary vessels – the *yang qiao mai* and the *yin qiao mai* as well as the *yang wei mai* and the *yin wei mai* – only form during subsequent divisions.

The eight extraordinary vessels and body structure

- According to Manaka et al (1995), a close connection exists between the eight extraordinary vessels and **body** structure/symmetry.
- The ***du mai*** and the ***ren mai*** separate the body **vertically** into **left** and **right**: the *ren mai* marks the anterior midline, the *du mai* ascends the posterior midline.
- The ***dai mai*** circles the waist **horizontally,** forming a **horizontal axis**. It separates **above** from **below** (→ Fig. 1.30 a).

In a further development of this idea, Manaka describes the division of the body into **eight areas,** using the structure of a topological octahedron for illustration. (→ Fig. 1.30 b). In this model, the eight extraordinary vessels directly relate to the lines 'separating' the faces of the octahedron:

- Due to their location on the anterior/posterior midline, the ***ren mai*** and the ***du mai*** separate the body into its right and left section.
- The ***dai mai*** separates above and below.
- The **G.B. and T.B. channels** compare to the anterior and posterior dividing lines at the lateral edges of the Yang aspect of the body.

- The **P** and **SP channels** compare to the anterior and posterior dividing lines at the lateral edges of the Yin portion of the body.

The opening points of the extraordinary vessels, with the exception of **SP-4** (a point on the LIV channel would be more appropriate), represent this division:

- **P-6** and **T.B.-5** are the opening points of the *yin wei mai* and *yang wei mai* respectively
- **SP-4** and **G.B.-41** are the opening points of the *chong mai* and *dai mai* respectively.

These four channels divide the **anterior** and the **lateral** aspects of the body and connect the **upper** and **lower** parts, the four associated extraordinary vessels functioning accordingly. The *yin qiao mai* and the *yang qiao mai*, due to their connection to the KID and BL channels (opening points **KID-6** and **BL-62**) have a close relationship to the *ren mai* and *du mai*. Furthermore, the pathway of the KID channel is located close to the *ren mai* while the BL channel is close to the *du mai*.

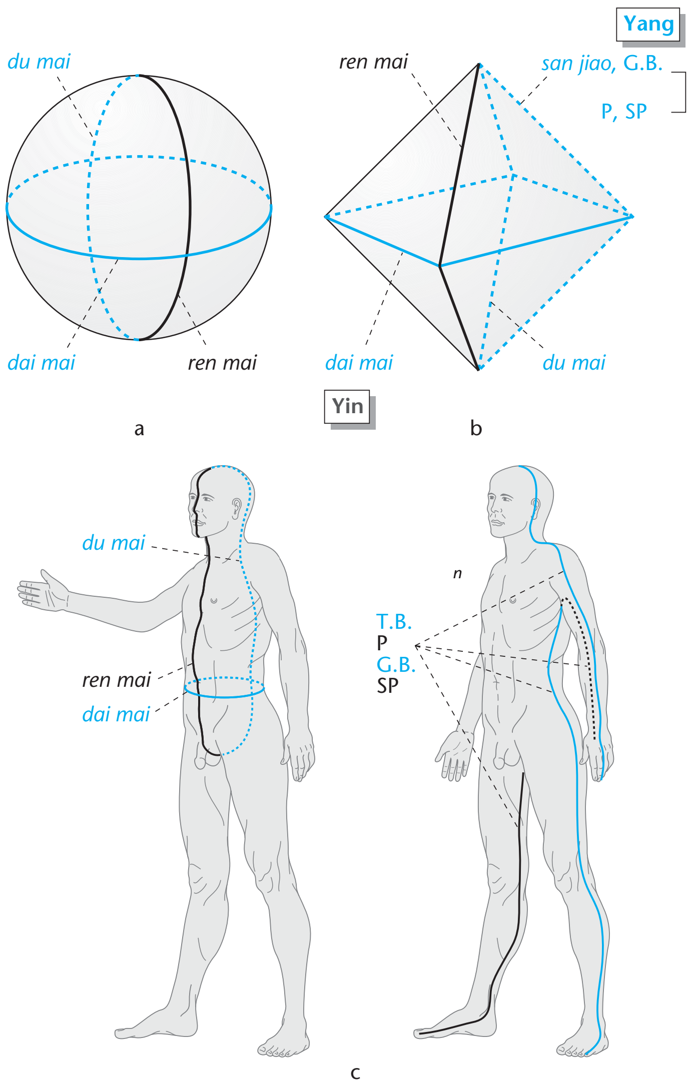

Fig. 1.30 Development of the octahedral structure based on body symmetry

1.7.2 Principles and Functions

'The primary channels are the rivers, the extraordinary vessels are the lakes.'

In some of the classics, the extraordinary vessels are compared to reservoirs, which store excess Yang, Yin, Qi or Blood, releasing it into the primary channel system in times of need. With the exception of the *ren mai* and *du mai* they do not have their own points, but they can be accessed through and are connected by points on the primary channels (see pathways of the primary channels → Chapter 5). All extraordinary vessels (with the exception of the *dai mai* with its horizontal trajectory) begin at the lower extremities or in the lower part of the torso and travel to the head. They can be activated by needling their opening and coupled points (differential therapy → 8.1.8).

Reservoir and compensation function

'*…when the network vessels (the primary channels) are filled to overflowing, none of the… conduits could seize any of their contents, and it is only then that the surplus contents of these vessels flow into the single conduit vessels (the extraordinary vessels)*' (*Nanjing*, translation by Unschuld 1986).

The *Nanjing* compares the extraordinary vessels to reservoirs, which are able to absorb a surplus of Qi and Blood in the primary channels, releasing them again during periods of deficiency. For example, if pathogenic factors penetrate the primary channel system, this excess can be absorbed by the extraordinary vessels. This will ensure that the *zangfu*-Organs remain unaffected. In that sense, the extraordinary vessels have a regulating compensation function.

Reproduction and fertility

The *chong mai* and the *ren mai* in particular, but also the *du mai*, play an important role in reproduction and fertility.

Connecting the primary channels

- The *chong mai*, also known as the '**sea of the 12 primary channels**', connects the primary ST and KID channels. It also strengthens the connection between the *ren mai* and the *du mai*.
- The *ren mai* as the **sea of Yin** connects all Yin channels
- The *du mai* as the **sea of Yang** connects all Yang channels at **Du-14** (*dazhui*).
- The *dai mai* circles the body at the waist and binds the vertical trajectories of all 12 primary channels in general, but especially the *chong mai*, the *ren mai* and the KID, LIV and SP channels.
- The *yin wei mai* **dominates the Interior** of the body and connects the KID, SP and LIV channels and the *ren mai*.
- The *yang wei mai* **controls the Exterior** of the body and connects the BL, G.B., T.B., S.I. and ST channels and the *du mai*.
- The *yin qiao mai* connects the KID and BL channels and controls balance.
- The *yang qiao mai* connects the BL, G.B., S.I., L.I. and ST channels and controls activity.

In addition, the eight extraordinary vessels are regarded as **connecting channels for the extraordinary *fu*-**Organs.

Protection of the body

The *chong mai*, *ren mai* and *du mai* circulate the Defensive Qi (*wei qi*) in the thorax, abdomen and the back, and thus help to protect the body against external pathogenic factors. Due to their origin, they are also directly linked to the Kidneys, the pre-Heaven Qi and to the constitutional forces.

1.7.3 Comparison of the Eight Extraordinary Vessels with the Primary Channels

	Primary channels	**Eight extraordinary vessels**
Points, trajectory	Specific, bilateral points	Only the *ren mai* and *du mai* have their own points (and internal pathways); all other extraordinary vessels 'borrow' points from the primary channels. The four central vessels are unilateral (except for the *chong mai*, which also has a pathway on the extremities); the four peripheral vessels are bilateral.
Circulation	Continuously circulate *Nutritive Qi* and Blood	Circulate Essence Qi, Nutritive Qi and Defensive Qi, but according to some authors they do not have a continuous flow of Qi
Depth	Superficial, external pathway and a deeper, internal pathway; directly connected to their pertaining *zangfu*-Organs	Compared to the internal pathways of the primary channels they are more superficial, compared to the external pathways of the primary channels they are deeper. They have **no direct connection** to the *zangfu*-Organs, but their origin – especially of the four central vessels – lies deep within the body.
Connections	They connect the Exterior and Interior (through internal/ external pathways and through Yin/Yang channels) as well as above and below (hand–foot pairing, axes)	They do not connect the Interior with the Exterior, nor do they connect Yin and Yang.
Functions	Guide excess to the eight extraordinary vessels or to other secondary channels	Reservoir function: excess of Qi and Blood in the primary channels flows to the eight vessels and are stored there. In times of deficiency they are transported back into the primary channels.
Diagnosis	Diagnosis based on channel theory (meridian therapy), in TCM according to the *ba gang* (*zangfu*, pulse and tongue)	Diagnosis is characterised by specific arrays of symptoms and specific pulses.

1.7.4 Pairing and Special Functions

The eight extraordinary vessels can be paired in several ways (→ Fig. 1.31):

- The central vessels and the peripheral vessels can be coupled to form Yin/Yang pairs
- Or a central vessel can be paired with a peripheral vessel. These pairs support particular regions and are very important in acupuncture therapy (→ Opening points 8.1.8).

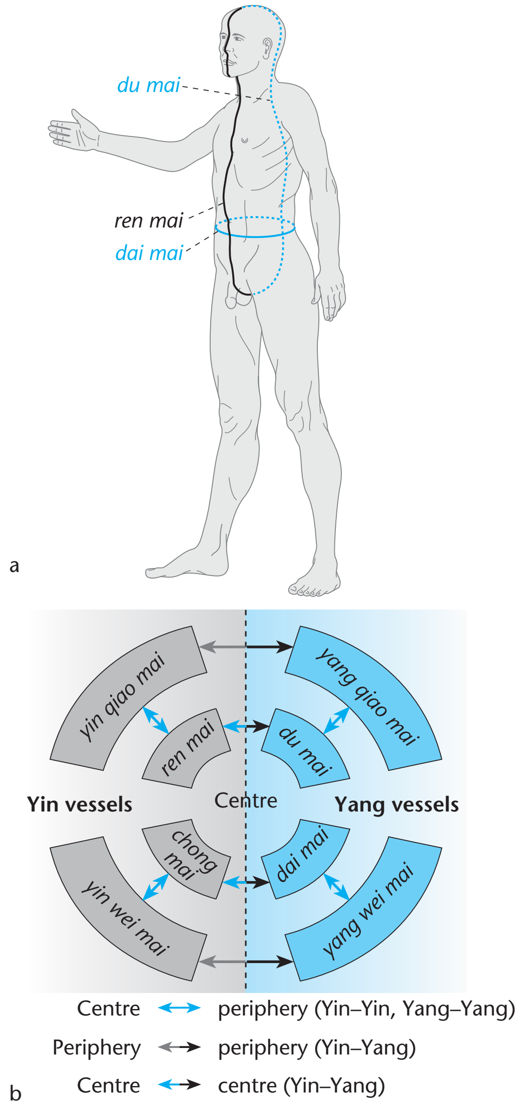
du mai
ren mai
dai mai
a
yin qiao mai
ren mai
du mai
yang qiao mai
Yin vessels
Centre
Yang vessels
chong mai
dai mai
yin wei mai
yang wei mai
Centre ↔ periphery (Yin–Yin, Yang–Yang)
Periphery ↔ periphery (Yin–Yang)
Centre ↔ centre (Yin–Yang)
b

Fig. 1.31 Pairing of the extraordinary vessels (b, graphic overview modified according to Pirog 1996)

Pairing according to Yin/Yang polarity

Pairing of the extraordinary vessels according to Yin/Yang polarity

Yin vessels	Yang vessels	Functions of the Yin/Yang pair	General functions – Centre/periphery
Central vessels (pathways on the head and torso)			
ren mai	*du mai*	The *du mai* can be considered to be the first manifestation of the Qi and Yang, the *ren mai* the first manifestation of the Blood and Yin. This pair **stores the Yin Qi and Yang Qi**.	**Functions:** Storage and distribution (of Yin and Yang). Foundation of the eight extraordinary vessels, stores Essence Qi, is responsible for the support of the internal organs and extraordinary *fu* organs, as well as for the fundamental body structures
chong mai	*dai mai*	The *chong mai* is like a harmonious connection between Yin and Yang, it expresses vitality and sets impulses. The *dai mai* encircles; it is like a container and prepares the circulation of Yin and Yang. The pair **circulates the Essence (*jing*) and Qi between the Yin and Yang regions of the body**.	
Peripheral vessels (pathways from the lower extremities to the torso)			
yin qiao mai	*yang qiao mai*	This pair **creates balance between Yin and Yang** (for example balance between Essence (*jing*) and Qi, Blood and Qi, left and right, Nutritive Qi and Defensive Qi, also on the outside – between day and night), controls and regulates the ascending (*yin qiao mai*) and descending (*yang qiao mai*) movement in the body, is responsible for balance and rhythm between Yin and Yang.	**Functions:** Balance, circulation and connection (of the Yin and Yang aspects). Provides contact between the distal regions of the body and the torso and head.
yin wei mai	*yang wei mai*	The *yin wei mai* governs the Interior (*li*), it organises and creates connections within the *Yin* aspects of the body. The *yang wei mai* governs the Exterior (*biao*), it organises within the Yang aspects of the body. This pair **connects the Yin and Yang respectively.** 'That which moves and circulates between the *Yin* channels is called the *yin wei* – that which moves between the Yang channels is called the *yang wei*.'	

Central–peripheral pairing

Here, a central vessel is combined with a peripheral vessel. This pairing probably refers to regions of the body which are traversed and supplied by the pair.

Central–peripheral pairing of the extraordinary vessels

Pair	Vessel	Relationship	Opening point	Coupling point	Supported regions
Pair 1	*chong mai*	father	SP-4	P-6	Heart, thorax, Stomach
	yin wei mai	mother	P-6	SP-4	
Pair 2	*du mai*	husband	S.I.-3	BL-62	Inner canthus of the eye, occiput, shoulders and back, S.I. and BL channels
	ren mai	wife	BL-62	S.I.-3	
Pair 3	*dai mai*	man	G.B.-41	T.B.-5	Lateral eye and temple region, ears, cheeks, occiput and shoulders
	yang wei mai	woman	T.B.-5	G.B.-41	
Pair 4	*ren mai*	host	LU-7	KID-6	Face, throat, thorax, Lungs, diaphragm, abdomen
	yin qiao mai	guest	KID-6	LU-7	

2 Location Methods and Cun Measurements

Claudia Focks

One of the basic prerequisites for any successful acupuncture treatment, besides a differential Chinese Medicine diagnosis and the relevant point selection, is the correct point location.

2.1 Location Methods for Body Acupuncture

There are various methods for locating acupuncture points:

Anatomical approach: many acupuncture points are situated at clearly defined anatomical locations, for example in depressions, at muscle and tendon insertions, in grooves, at joint clefts, at bony prominences, etc. With some experience, the palpating finger will often automatically 'fall' into the right depressions and holes.

Various anatomical structures and zones form the basis for point location. These are described in more detail in Chapter 3. We can distinguish between the following:

- **Fixed markers** are not affected by the body position or by movement. These include bony landmarks such as depressions or prominences of the skeletal system, but also the finger- and toenails, the nipples, the umbilicus, etc. Most of the proportional cun measurements (→ 2.2) rely on fixed markers.
- **Movable markers** often become more clearly defined with a particular body posture or movement (→ 2.3.2). For example, by flexing the elbow, the cubital crease becomes much more clearly visible to locate **L.I.-11** precisely. Equally, the most distal palmar crease becomes much more visible when making a loose fist, in order to locate **S.I.-3**. Other parameters that might determine the location of a point include changed skin consistency, sensitivity to pressure, puffiness or a resistant sensation when palpating gently with a finger.

Proportional measurements: When trying to locate points not situated at any prominent structures, Chinese Medicine applies the proportional cun measurement (→ 2.2).

Electric tools: These measure the electric resistance of the skin in order to find the correct location of the points. Generally, electric resistance is lower in the immediate area around the point. This method is especially used in auricular acupuncture, but it has not proved to be a practical method for body acupuncture.

Very point technique (after Gleditsch 2005): The needle is held loosely and guided in a 'dancing' fashion over the skin. When the correct point – the 'very point' – is found, the needle will 'catch' on to it and penetrate the skin as if out of its own volition.

2.2 Cun Measurements of the Body

In Chinese medicine, distances on the body are traditionally measured in cun. In contrast to the official Chinese cun (1 cun = 2.5 cm), the cun used in a medical context is a **proportional unit of measurement** that takes the individual proportions of each body into account.

This relative unit of measurement is defined either by using the width of a particular finger or fingers (finger cun) or by the distances between clearly defined landmarks of the body (body cun).

In clinical practice, many points can be located by using **finger cun** measurements (→ Fig. 2.1). It is important, however, to use the **patient's fingers**, not the practitioner's fingers, for reference. If the fingers of both the practitioner and the patient have approximately the same width, the width of the practitioner's fingers can be applied for point location. Otherwise, measurements have to be adapted to correct minor deviations, for example by adjusting the spaces (narrow or wide) between the practitioner's fingers when using finger cun measurements.

The **body or proportional cun** is based on the proportions of particular sections of the body, which are divided into a specific number of units (→ Fig. 2.3). In clinical practice, the measurement of these divisions can be quite cumbersome. Therefore, the spreading hands technique is commonly applied (Koenig and Wancura 1979/1983). It allows the practitioner to quickly determine the midpoint of a given distance (→ 2.3.3) and, combined with finger cun measurements and palpation, is in most cases sufficient for correctly locating a point. For point location on the abdomen, a prepared elastic tape can be useful for measuring proportional cun (→ 2.3.1).

Finger cun

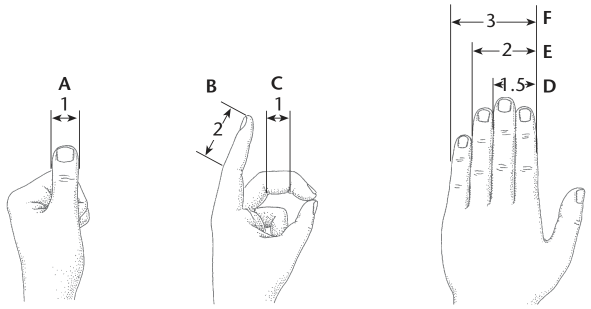
A
1
B
2
C
1
3
F
2
E
1.5
D

A:
1 cun:
thumb

B:
2 cun:
index finger

C:
1 cun:
middle finger

D:
1.5 cun:
finger-width

E:
2 cun:
finger-width

F:
3 cun:
finger-width

Fig 2.1

A	1 cun as thumb measure	The width of the interphalangeal joint of the thumb at its widest point
B	2 cun as index finger measure	The distance of the two distal phalanges of the index finger
C	1 cun as middle finger measure	With the tip of the middle finger placed on the thumb, the distance between the ends of the creases of the interphalangeal joints of the middle finger
D	1.5 cun as two-finger measure	The width of the index and middle finger when placed closely together, at the level of the distal interphalangeal joints
E	2 cun as index, middle and ring finger measure	The width of the index, middle and ring finger when placed closely together, at the level of the distal interphalangeal joints
F	3 cun as four finger measure	The width of the four fingers when placed closely together, at the level of the distal interphalangeal joints (=1 hand-breadth)

Body cun or proportional cun – the most commonly used proportional distances

Tip for locating the end of the axillary fold: The anterior or posterior axillary fold defines the end of the axilla, forming a fold in the skin (both anteriorly and posteriorly), where the arm joins the thorax. The axillary fold is the reference point for any cun measurements on the upper arm (→ Figs 2.2, 2.3, LU-3 and LU-4), which can be found by asking the patient to adduct his/her arm and to insert his/her flat hand into the axilla. The end of the axillary fold points in a superior direction, terminating at the lower border of the pectoralis major muscle, which in this area extends to its insertion in the humerus, below the deltoid muscle. The end of the axillary fold may vary considerably, depending on the individual proportions, positioning of the arm and flexing of the muscles. It should therefore be located with the patient's arms hanging down in a relaxed way.

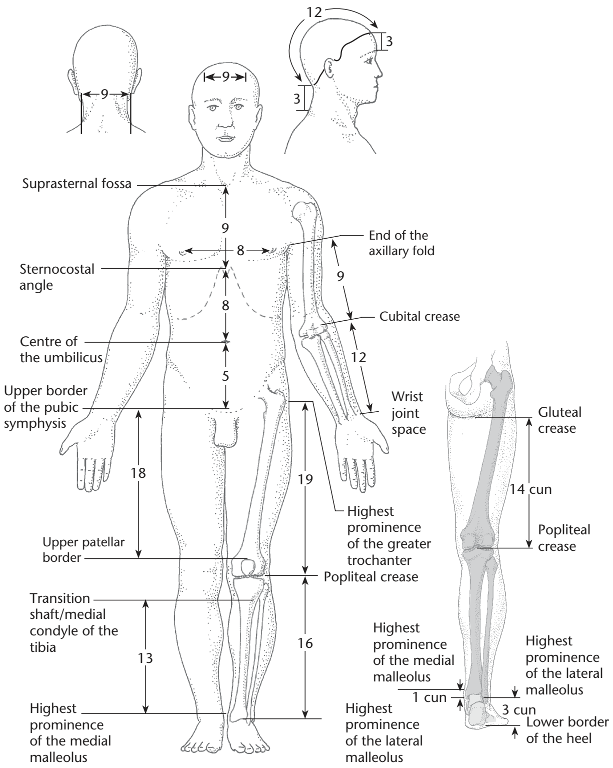
9
9
12
3
3
Suprasternal fossa
9
8
End of the axillary fold
9
Sternocostal angle
8
Cubital crease
Centre of the umbilicus
12
5
Upper border of the pubic symphysis
Wrist joint space
Gluteal crease
18
19
14 cun
Highest prominence of the greater trochanter
Popliteal crease
Upper patellar border
Popliteal crease
Transition shaft/medial condyle of the tibia
Highest prominence of the medial malleolus
Highest prominence of the lateral malleolus
13
16
1 cun
3 cun
Highest prominence of the medial malleolus
Highest prominence of the lateral malleolus
Lower border of the heel

Fig. 2.2

Head		
Longitudinal	12 cun	From the anterior (→ 3.1.1) to the posterior hairline (→ 3.1.5)
	3 cun	From the midpoint of the eyebrows to the anterior hairline (→ 3.1.1)
	3 cun	From the lower border of the spinous process of C7 (→ 3.4.1) to the posterior hairline (→ 3.1.5)
	18 cun	From **Ex-HN-3** (***yintang***) to **Du-14**
Horizontal	9 cun	The distance between the mastoid processes (→ 3.1.4)
	9 cun	From corner to corner on the forehead (**ST-8** to **ST-8**)
Thorax and abdomen		
Longitudinal	9 cun	From the suprasternal depression (**Ren-22**) to the sternocostal angle (→ 3.5, xiphosternal synchondrosis)
	8 cun	From the sternocostal angle (→ 3.5, xiphosternal synchondrosis) to the centre of the umbilicus
	5 cun	From the centre of the umbilicus to the upper border of the pubic symphysis (→ 3.5)
Horizontal	8 cun	The distance between the two nipples
Laterally longitudinal	12 cun	From the end of the axillary fold to the tip of the free end of the 11th rib (**LIV-13**)
Back and lumbar region		
Both the spinous processes and the intercostal spaces are used for orientation. Main reference points are: the spinous process of C7, the inferior angle of the scapula (level with T7), the iliac crest (level with the spinous process of L4), the posterior superior iliac spine (PSIS) (level with S2 or the 2nd sacral foramen). Please note: the position of the patient (standing, sitting, lying) as well as individual anatomy and spinal structure (kyphosis, lordosis etc.) all influence the level of the vertebrae, which consequently may vary considerably in relation to other anatomical structures. For more detail → Chapter 3.		
Longitudinal	30 cun	From the spinous process of the T1 to the tip of the coccyx
Horizontal	3 cun	From the medial border of the scapula to the posterior midline (spinous processes), with the shoulder relaxed and hanging arms
Upper extremity		
Lateral	9 cun	From the end of the anterior axillary fold to the transverse cubital crease
	12 cun	From the transverse cubital crease to the wrist joint space (→ 3.3.3, transverse wrist crease)
Lower extremity		
Lateral	19 cun	From the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease
	16 cun	From the popliteal crease to the highest prominence of the lateral malleolus (→ 3.6.2)
	14 cun	From **G.B.-34** (anterior and inferior to the head of the fibula) to the highest prominence of the lateral malleolus (→ 3.6.2)
	3 cun	From the highest prominence of the lateral malleolus (→ 3.6.2) to the lower border of the heel
Medial	18 cun	From the upper border of the pubic symphysis (3.5) to the upper border of the patella (→ 3.6.1)
	15 cun	From the medial end of the popliteal crease to the highest prominence of the medial malleolus (→ 3.6.2)
	13 cun	From the medial condyle of the tibia (→ 3.6.1, junction of the head and the shaft of the tibia) to the highest prominence of the medial malleolus (→ 3.6.2)
Dorsal	14 cun	From the gluteal crease to the popliteal crease
Difference between the medial and lateral malleolus		
	1 cun	Difference between the level of the highest prominence of the lateral and medial malleolus respectively (→ 3.6.2)

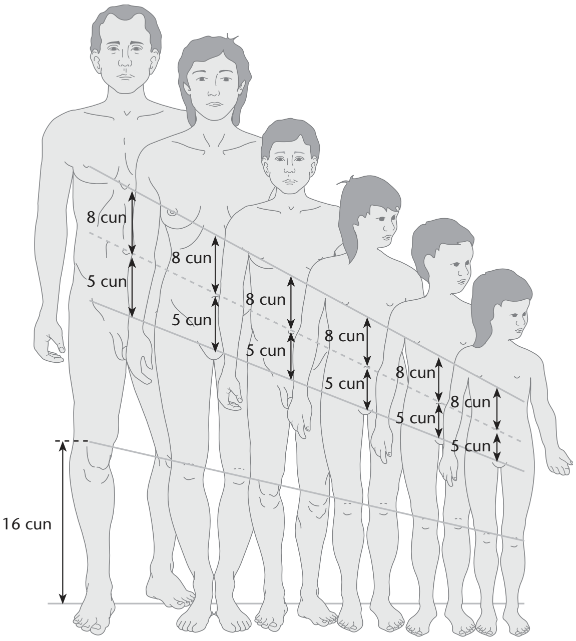
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
8 cun
5 cun
16 cun

Fig. 2.3

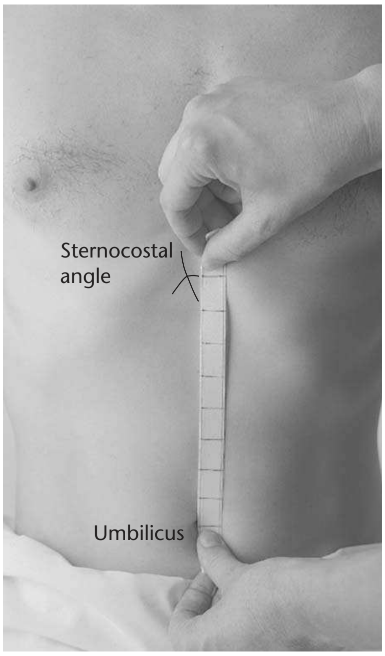
Sternocostal angle
Umbilicus

Fig. 2.4

2.3 Tips for Point Location

2.3.1 Elastic Tape

A marked elastic tape has proven a very useful tool for simplifying cun measurements and facilitating correct point location. It is especially recommended for measurements on the abdomen, the shape of which can vary considerably from patient to patient (Kitzinger 1995). An elastic tape, 1–2 cm wide and approximately 40 cm long, is marked at regular intervals of 2 cm, which can optionally be numbered from 1–20. The tape is placed onto the distance to be measured, stretching it according to the correct number of cun for that particular distance. This method allows the use of proportional cun for each individual patient.
Examples:

- For locating points on the upper abdomen, stretch the tape so that eight units cover the distance between the sternocostal angle and the centre of the umbilicus (➞ Fig. 2.4).
- For locating points on the lower abdomen, stretch the tape so that five units cover the distance between the upper border of the pubic symphysis and the centre of the umbilicus (➞ Fig. 2.5).

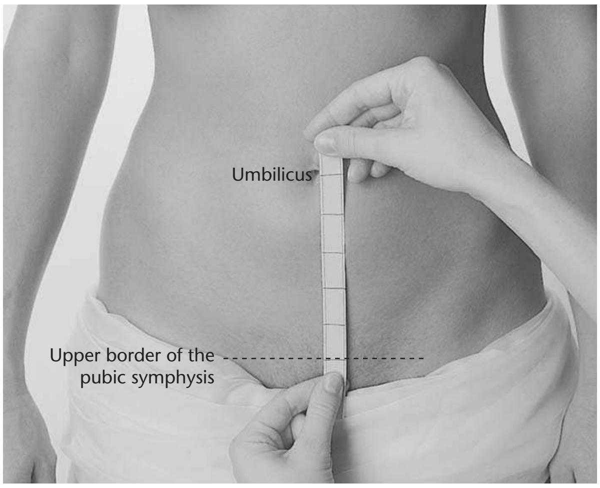
Umbilicus
Upper border of the pubic symphysis

Fig. 2.5

2.3.2 Patient Positioning and Body Movements

To locate certain points, it can be very helpful to ask the patient to assume a particular position or to perform a specific movement.

Examples:

- For locating points on the **P** channel on the forearm, ask the patient to tightly press the thumb and little finger together. This will make the tendons more visible (→ Fig. 2.6)
- 'Tiger mouth grip' as location help for **LU-7** and **L.I.-6** (→ Fig. 2.7)
- For locating **L.I.-4**, ask the patient to press his/her thumb and index finger together. A muscle bulge will form; L.I.-4 is located on its highest point (→ Fig. 2.8).
- When turning the wrist joint from a pronated to a half-supinated position, the palpating finger will glide into a bony cleft just proximal to the styloid process of the ulna (**S.I.-6**, → fig 2.9).
- As the palpating finger glides along the skin, the skin bulge forming in front of it will 'arrest' the finger at the point to be located. This method is very useful for quick location of certain points such as **T.B.-5** or **P-6** (→ Fig. 2.10).

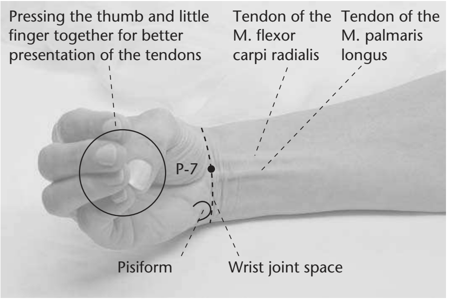
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the M. flexor carpi radialis
Tendon of the M. palmaris longus
P-7
Pisiform
Wrist joint space

Fig. 2.6

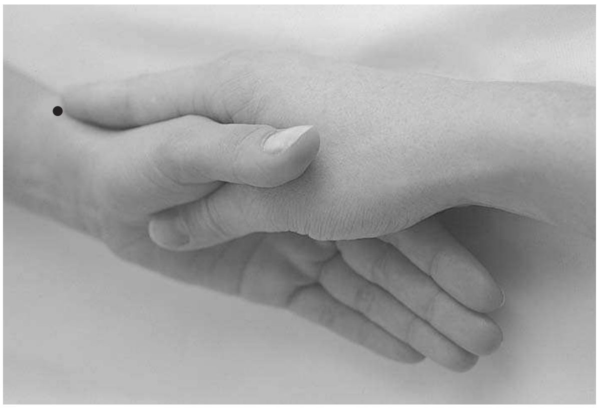

Fig. 2.7

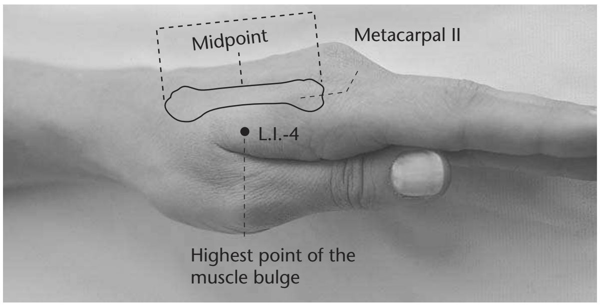
Midpoint
Metacarpal II
L.I.-4
Highest point of the muscle bulge

Fig. 2.8

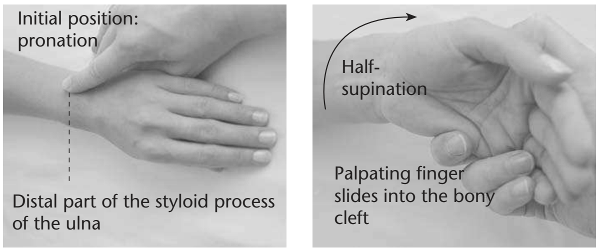
Initial position: pronation
Distal part of the styloid process of the ulna
Half-supination
Palpating finger slides into the bony cleft

Fig. 2.9

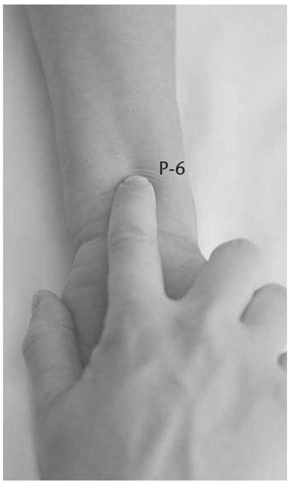

Fig. 2.10

2.3.3 Spreading Hands Technique According to König and Wancura

The spreading hands technique is particularly useful for quickly determining the midpoint of a given distance. For this purpose, the little fingers of each hand are placed on the endpoints of the distance to be measured. Both hands are then evenly spread over the distance. The thumbs are joined in the middle, indicating the midpoint of the distance.

Examples:

- To determine the midpoint of the 16 cun-distance between the popliteal crease and the prominence of the lateral malleolus, the little fingers are placed on both endpoints of this distance (→ Fig. 2.11)
- To locate **Du-20**, the hands are placed on the lateral aspects of the patient's head, with the little fingers touching the apices of the ears. The thumbs are then joined on the vertex of the head, and will be pointing to **Du-20** (→ Fig. 2.12).

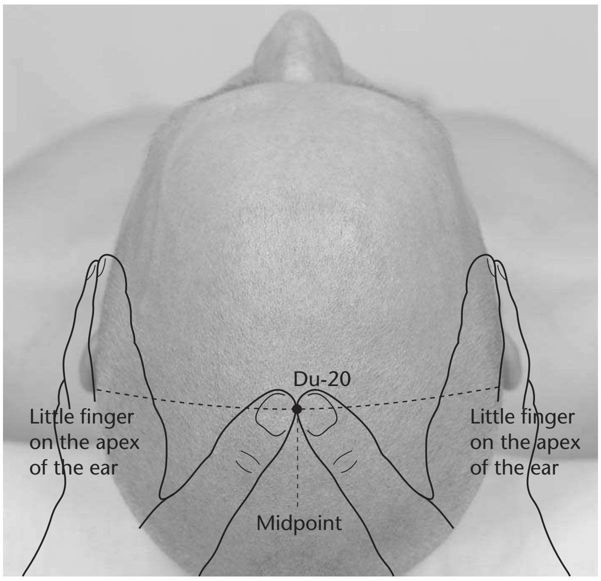
Du-20
Little finger on the apex of the ear
Little finger on the apex of the ear
Midpoint

Fig. 2.12

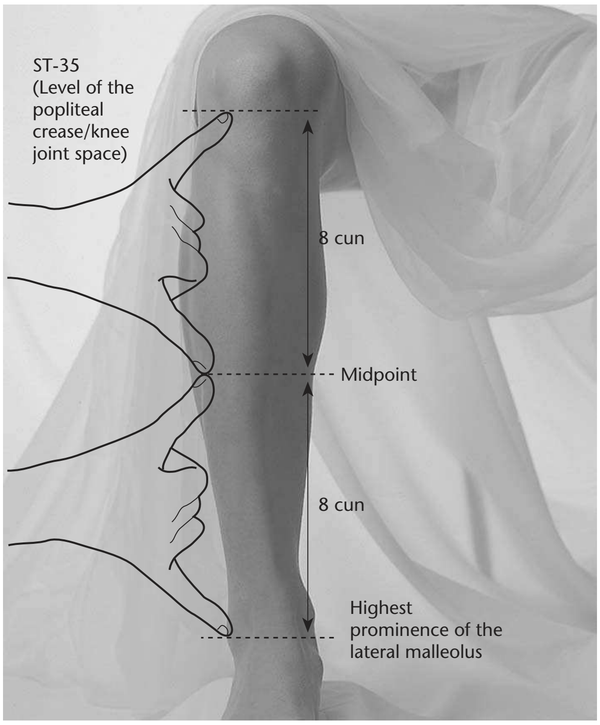
ST-35 (Level of the popliteal crease/knee joint space)
8 cun
Midpoint
8 cun
Highest prominence of the lateral malleolus

Fig. 2.11

3 Anatomical Orientation

Ulrich März

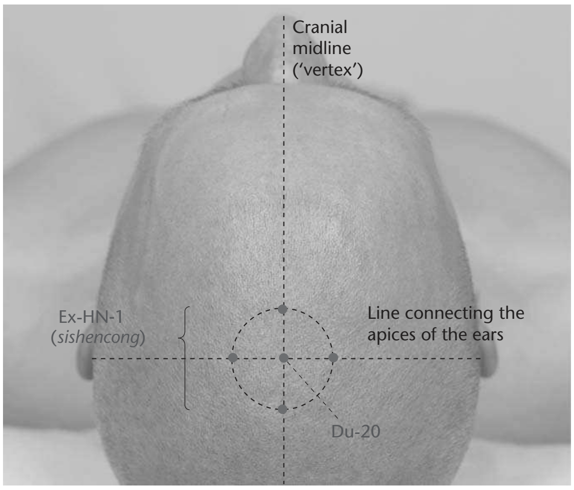
Cranial
midline
('vertex')
Ex-HN-1
(*sishencong*)
Line connecting the
apices of the ears
Du-20

Fig. 3.1

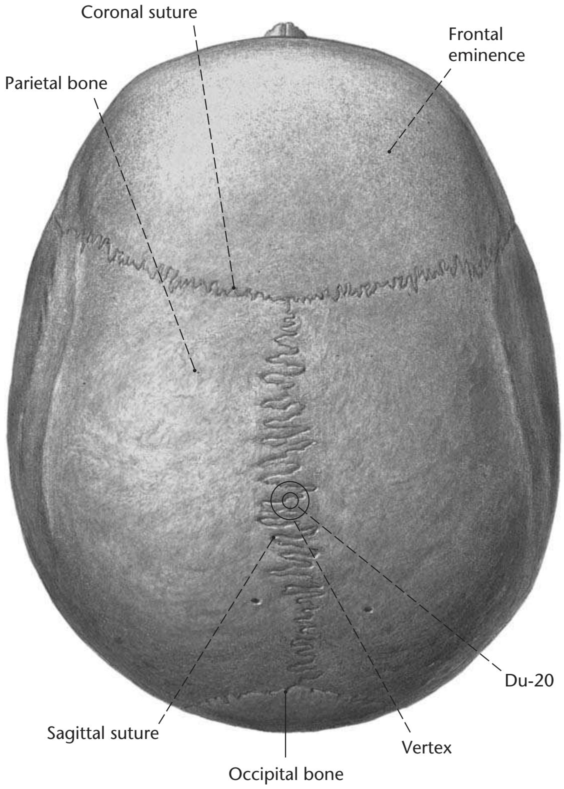
Coronal suture
Frontal
eminence
Parietal bone
Du-20
Sagittal suture
Vertex
Occipital bone

Fig. 3.2

3.1 Head

3.1.1 Vertex, Forehead and Supraorbital Region

Vertex (→ Figs 3.1, 3.2)

→ **Du-20** (*baihui*) is located at the junction of the cranial midline and an imaginary line connecting the apices of the ears. This point is at a distance of 5 cun from the anterior hairline and 7 cun from the posterior hairline, marking the highest point of the cranium ('vertex').

Please note: When locating **Du-20**, the patient's head should be in an upright position. Spreading hands technique (after König and Wancura, → Chapter 2): Place the hands on the lateral aspects of the patient's head, with the little fingers touching the apices of the ears. The thumbs are then joined at the cranial midline and will be pointing to → **Du-20** (*baihui*).

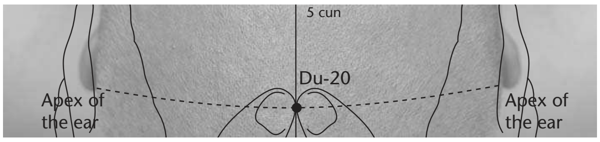
5 cun
Du-20
Apex of
the ear
Apex of
the ear

Fig. 3.3

Anterior hairline (→ Fig. 3.4)

The anterior hairline forms the superior border of the forehead and can vary considerably from person to person. The original line of hair growth is generally considered as the anterior hairline.
In men especially, the hairline may begin to recede early in life due to progressive baldness, so that the existing hairline may not be identical with the original line, which will now be located in an area not covered by hair. In these cases ask the patient to frown: the original hairline will be located superior to the wrinkled forehead, at the junction of the still slightly creased skin with the completely smooth skin. Points in this area are → **Du-20** to **Du-24**, **BL-3** to **BL-7, G.B.-4** to **G.B.-7, G.B.-13** to **G.B.-15, ST-8.**

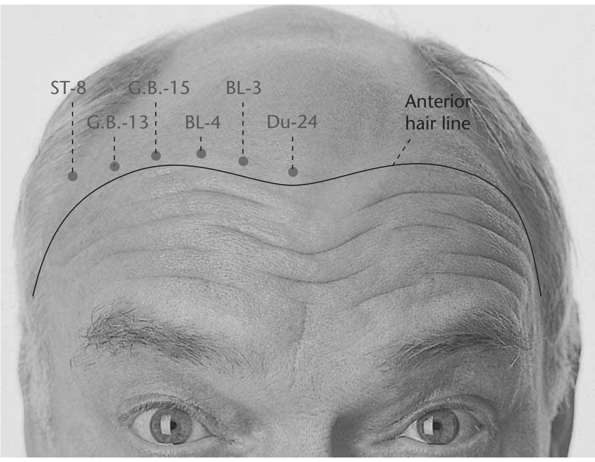
ST-8
G.B.-15
BL-3
G.B.-13
BL-4
Du-24
Anterior hair line

Fig. 3.4

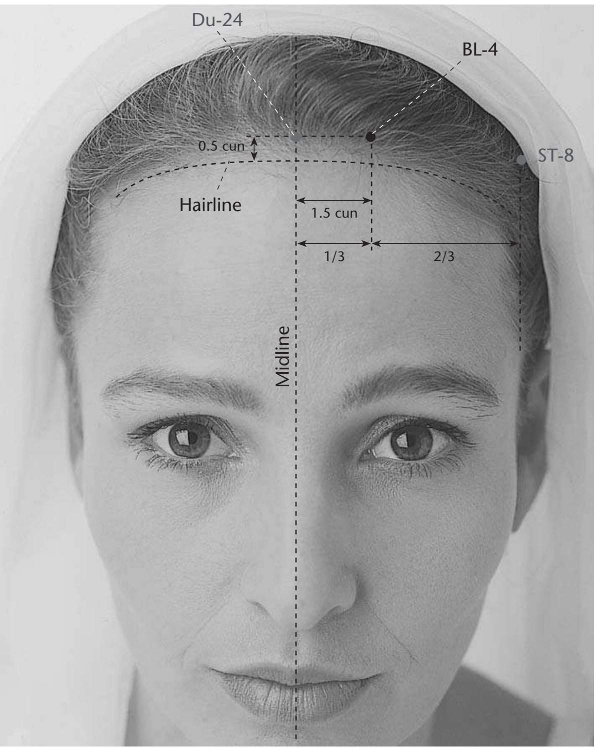
Du-24
BL-4
0.5 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

Fig. 3.5

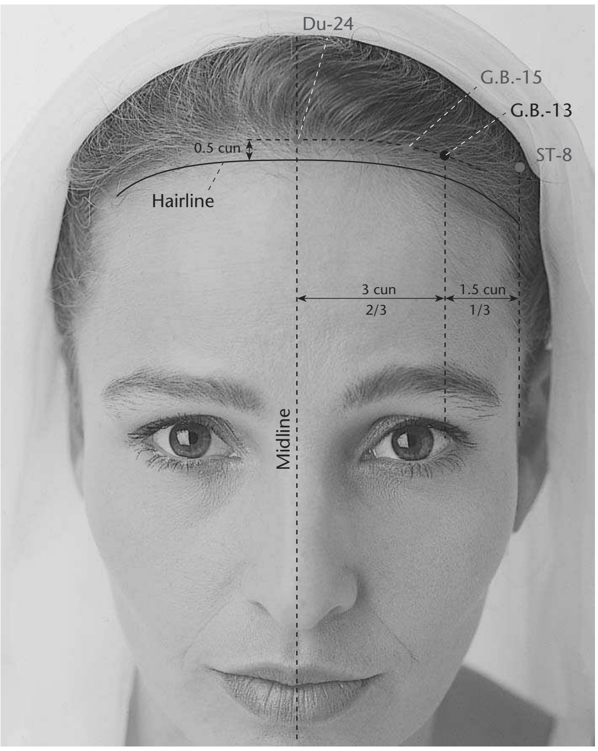
Du-24
G.B.-15
G.B.-13
0.5 cun
ST-8
Hairline
3 cun
2/3
1.5 cun
1/3
Midline

Fig. 3.6

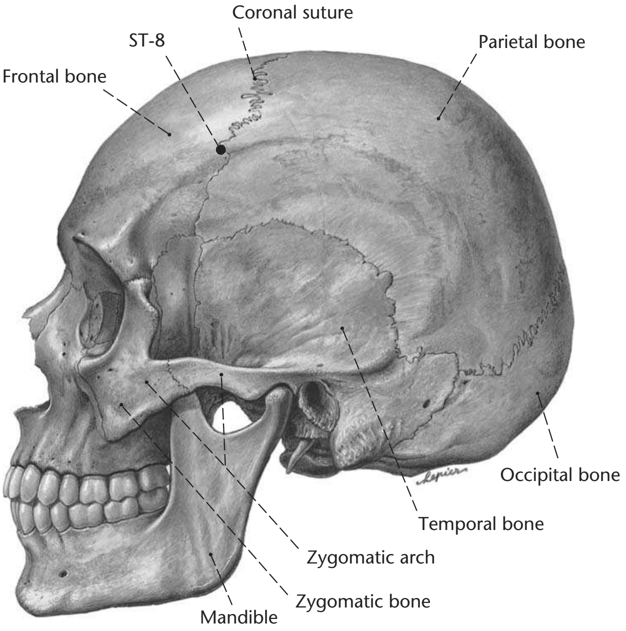
Coronal suture
ST-8
Parietal bone
Frontal bone
Occipital bone
Temporal bone
Zygomatic arch
Zygomatic bone
Mandible

Fig. 3.7

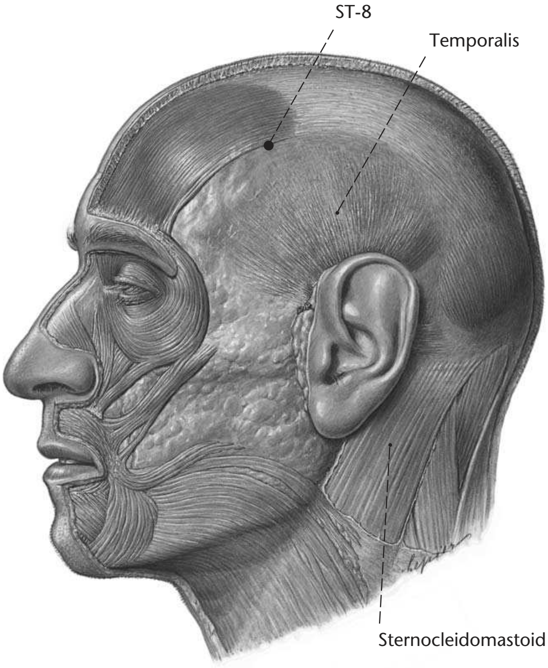
ST-8
Temporalis
Sternocleidomastoid

Fig. 3.8

Temporalis muscle and temporal corner of the forehead (→ Figs 3.7–3.9)

By asking the patient to firmly clench and release his/her teeth, the temporalis muscle can be palpated on the lateral aspect of the forehead. The corner between the forehead and the temples is marked by the junction of the anterior hairline with the temporal hairline, which tends to become more pronounced with a receding hairline.

→ **ST-8** is located in the centre of this area, at the insertion point of the temporalis muscle and 0.5 cun within the natural hairline. Often, the coronal suture can be palpated here.

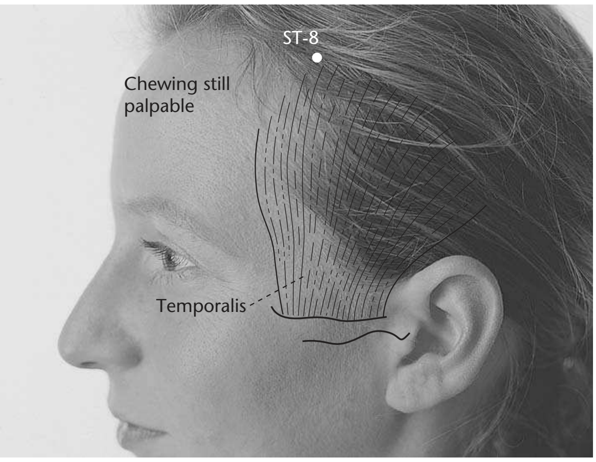
ST-8
Chewing still palpable
Temporalis

Fig. 3.9

Forehead (→ Figs 3.10, 3.11)

The more or less prominent bulge of the frontal eminences can be palpated on both sides of the forehead. → **G.B.-14** is located at their lower border, in a shallow depression at the junction of the eminences and the superciliary arches. With the patient looking straight ahead, G.B.-14 is located directly above the pupil, approximately 1 cun superior to the supraorbital margin.

Glabella, superciliary arches

The glabella is the central bony reference point on the forehead, forming a slightly convex, level area superior to the root of the nose and between the superciliary arches.

The extra point *yintang* (Ex-HN-3) is located at the centre of the glabella. Several points are located along the superciliary arches (from medial to lateral): → **BL-2, *yuyao*** (Ex-HN-4) and **T.B.-23.**

Please note: → **BL-2** is located superior to the inner canthus of the eye, in the area of the frontal (not the supraorbital) incisure, when this is present.

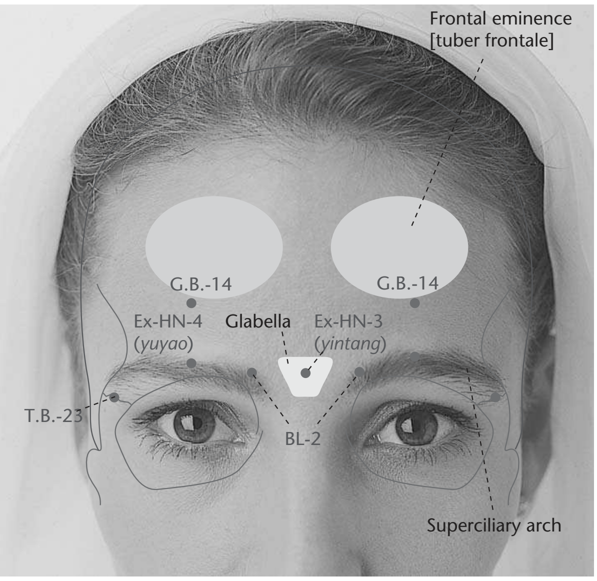
Frontal eminence
[tuber frontale]
G.B.-14
G.B.-14
Ex-HN-4
(*yuyao*)
Glabella
Ex-HN-3
(*yintang*)
T.B.-23
BL-2
Superciliary arch

Fig. 3.10

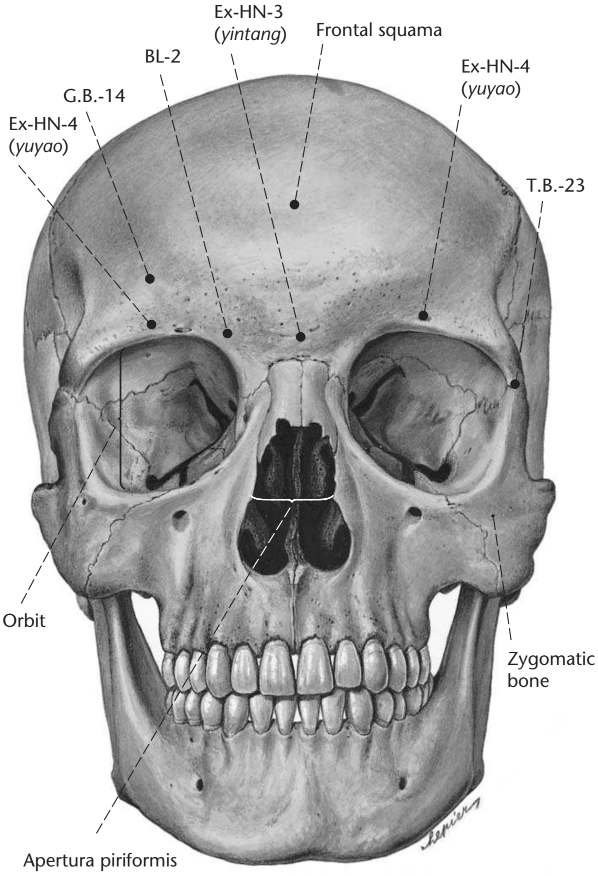
Ex-HN-3
(*yintang*)
Frontal squama
BL-2
Ex-HN-4
(*yuyao*)
G.B.-14
Ex-HN-4
(*yuyao*)
T.B.-23
Orbit
Zygomatic
bone
Apertura piriformis

Fig. 3.11

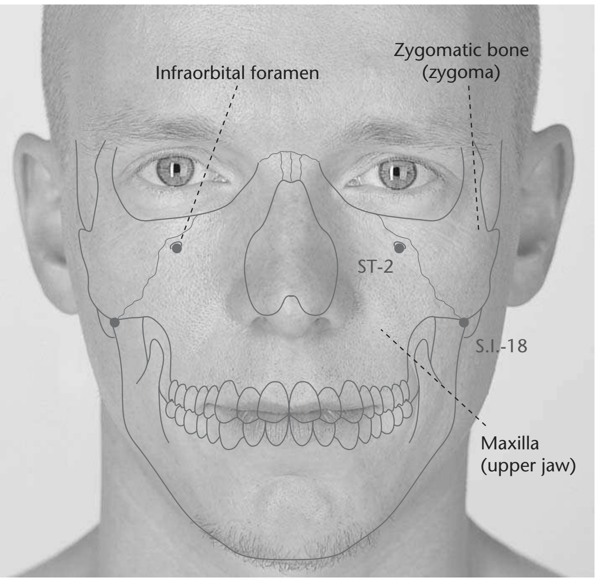
Infraorbital foramen
Zygomatic bone
(zygoma)
ST-2
S.I.-18
Maxilla
(upper jaw)

Fig. 3.12

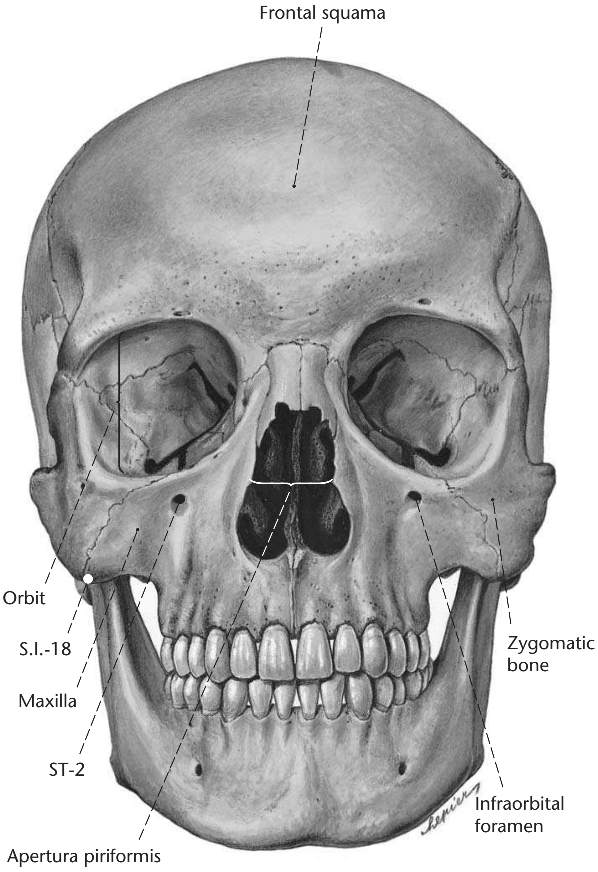
Frontal squama
Orbit
S.I.-18
Maxilla
ST-2
Apertura piriformis
Zygomatic
bone
Infraorbital
foramen

Fig. 3.13

3.1.2 Centre of the Face and Nasal Region

Cheek bones (→ Figs 3.12, 3.13)

Both the maxilla and the zygoma form the cheek bones below the eyes. Important anatomical landmarks are the infraorbital foramen, the lower border of the zygoma and the zygomatic arch of the temporal bone.

Infraorbital foramen (→ Figs 3.12, 3.13)

With the patient looking straight ahead, the infraorbital foramen with → **ST-2** is located on the maxilla, directly below the pupil. The foramen lies 1 finger-width below the infraorbital ridge. By palpating inferiorly from there, it can be felt below the maximal curvature of the ridge, in a small depression that is often sensitive to pressure.

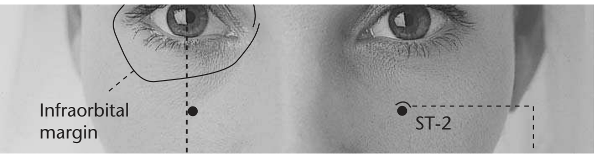
Infraorbital
margin
ST-2

Fig. 3.14

Zygoma (→ Figs 3.15, 3.16)

The bony structures below the lateral orbit are formed by the zygoma. → **S.I.-18** is located directly below the outer canthus of the eye at the lower border of the zygoma, at the junction of the zygoma and the maxilla, on the anterior border of the masseter muscle. (→ Lower face and chin).

Zygomatic arch (→ Figs 3.15, 3.16)

On the lateral aspect of the face, the temporal process of the zygoma articulates with the zygomatic process of the temporal bone to form the zygomatic arch. This is an approximately horizontal bony structure which can be palpated laterally to the ear. → **G.B.-3** and → **T.B.-22** are both located on the upper border of the zygomatic arch, while → **ST-7** and → **T.B.-21** are situated at its lower border.

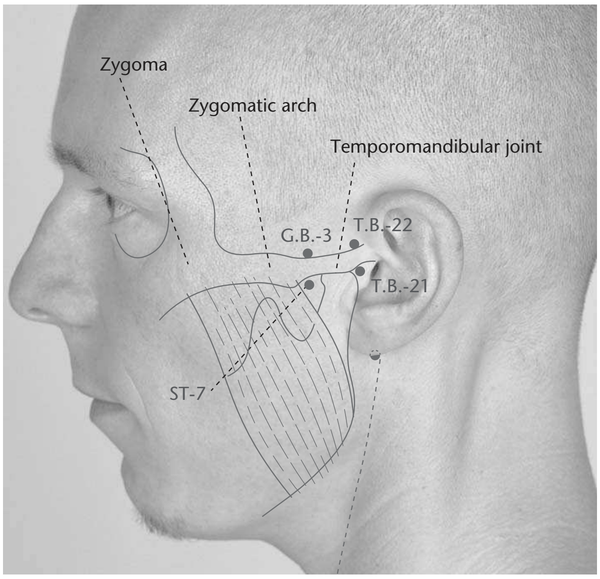
Zygoma
Zygomatic arch
Temporomandibular joint
G.B.-3
T.B.-22
T.B.-21
ST-7

Fig. 3.15

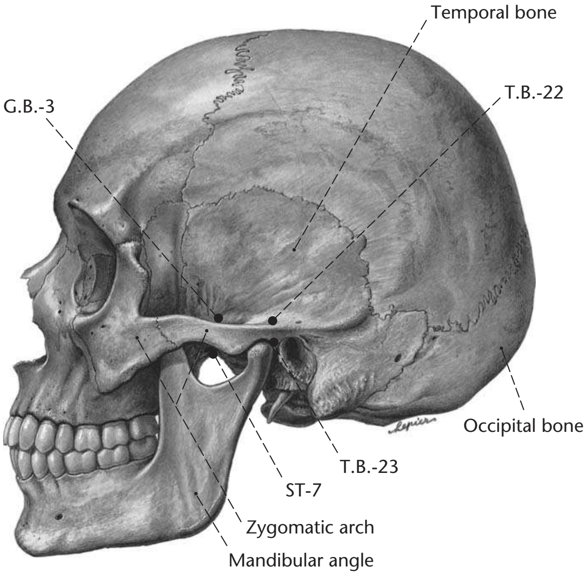
Temporal bone
T.B.-22
G.B.-3
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

Fig. 3.16

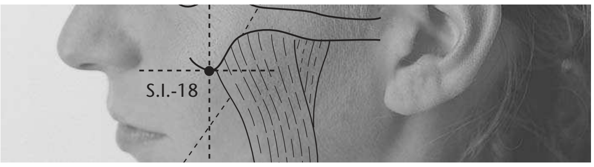

Fig. 3.17

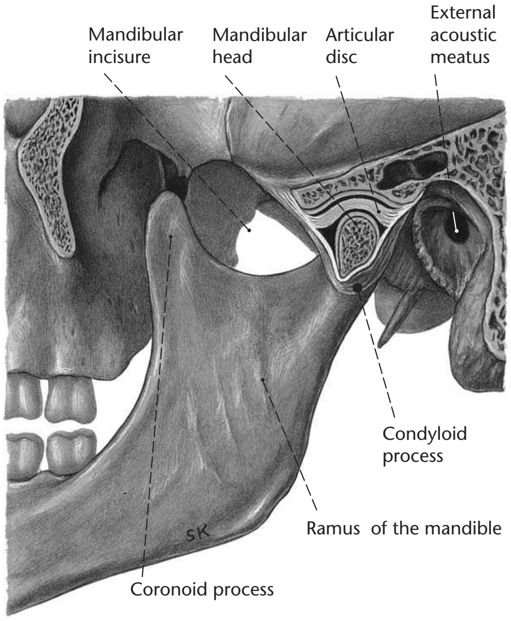
Mandibular incisure
Mandibular head
Articular disc
External acoustic meatus
Condyloid process
Ramus of the mandible
Coronoid process
SK

Fig. 3.18

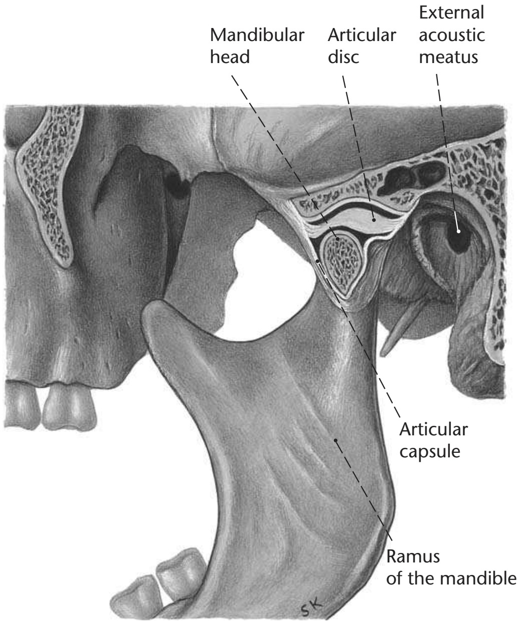
Mandibular head
Articular disc
External acoustic meatus
Articular capsule
Ramus of the mandible
SK

Fig. 3.19

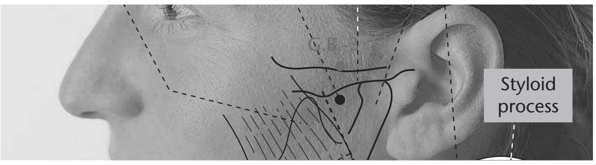
Styloid process

Fig. 3.20

Temporomandibular joint (➞ Figs 3.15, 3.18, 3.19)

The temporomandibular joint is located inferior to the zygomatic arch and anterior to the tragus of the ear. Its movement can easily be palpated by gently opening and closing the mouth. ➞ **ST-7** is located in a depression anterior to the joint and directly inferior to the zygomatic arch. This depression is situated in the mandibular incisure, which is formed by the coronoid and condylar processes of the mandible. By opening the mouth, the palpating finger resting on ➞ **ST-7** is 'lifted' out of the depression by the condyloid process sliding anteriorly.

Nasolabial groove (→ Fig. 3.21)

The nasolabial groove begins lateral to the nose and superior to the ala nasi, at the junction of the cartilage and the bone. From there it runs to the corners of the mouth. → **Ex-HN-8** (*shangyingxiang, bitong*) is located at its upper end, at the junction of the upper jaw and the nasal cavities. → **L.I.-20** is located slightly below *bitong*, between the lateral border of the side of the nose and the nasolabial groove.

3.1.3 Lower Face and Chin

Mandibular angle (→ Fig. 3.22)

The mandibular angle corresponds to the area where the lower jaw line changes its direction from horizontal to vertical. It is located markedly inferior and anterior to the ear lobe. This angle marks the location of → **ST-6** (1 finger-width anterior and superior to the mandibular angle) as well as of → **S.I.-17** and → **T.B.-16** (level with the mandibular angle).

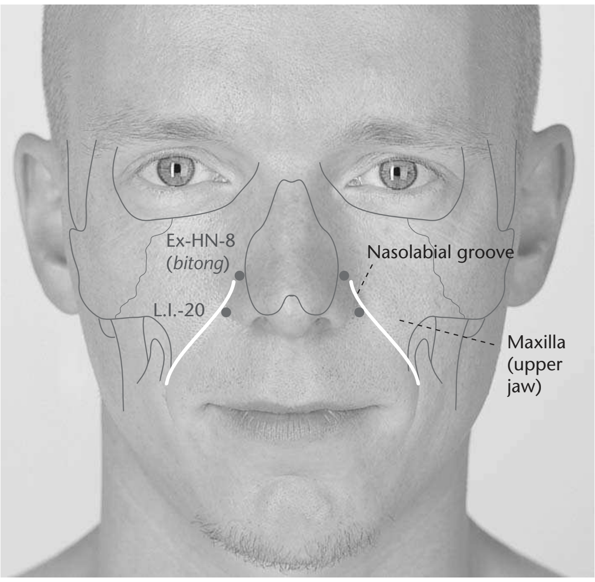
Ex-HN-8
(*bitong*)
L.I.-20
Nasolabial groove
Maxilla
(upper
jaw)

Fig. 3.21

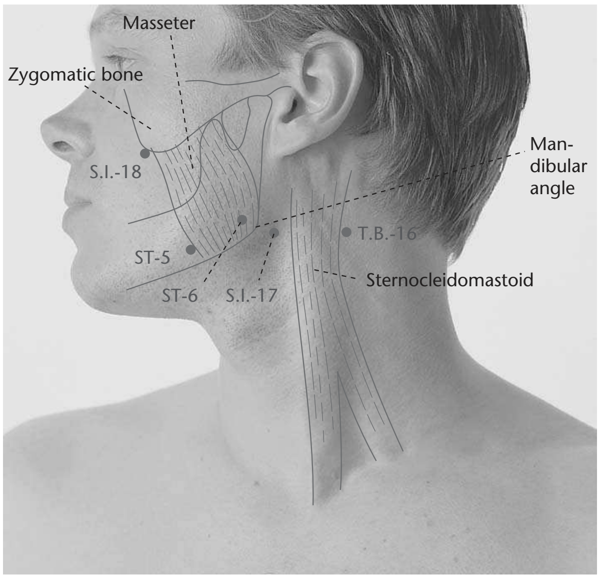
Masseter
Zygomatic bone
S.I.-18
ST-5
ST-6
S.I.-17
T.B.-16
Sternocleidomastoid
Man-
dibular
angle

Fig. 3.22

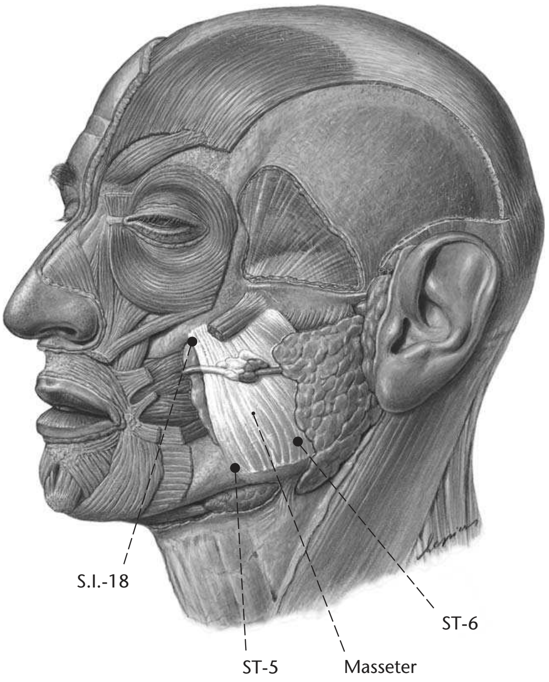
S.I.-18
ST-5
Masseter
ST-6

Fig. 3.23

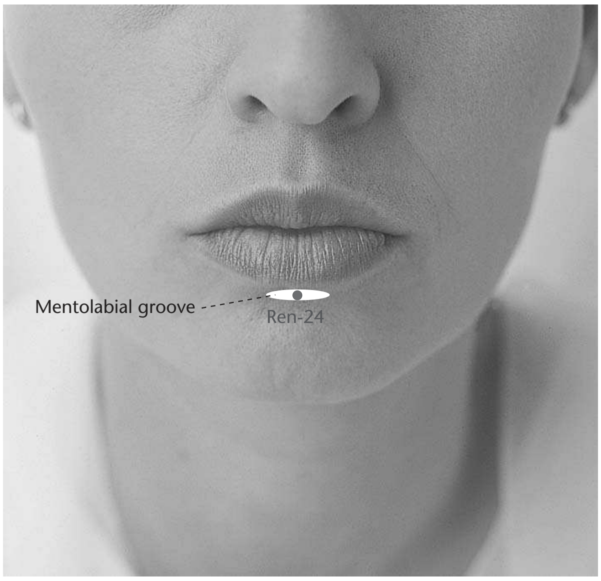
Mentolabial groove
Ren-24

Fig. 3.24

Masseter muscle (→ Fig. 3.23)

By firmly clenching the jaw, the contours of the masseter muscle become well defined, stretching from the lateral aspect of the lower jaw to the zygomatic bone. It originates at the zygomatic arch (→ Fig. 3.15) and inserts at the mandibular angle (→ Fig. 3.22).

Mentolabial groove (→ Fig. 3.24)

This forms a transverse groove at the junction of the chin and the lower lip. → **Ren-24** is located in its centre.

3.1.4 Ear Region

Note: points used in auricular acupuncture are not discussed here.

Helix root (→ Figs 3.25, 3.26)

The helix root borders the upper part of the ear. This is the location of → **T.B.-22**.

Supratragic notch (→ Figs 3.25, 3.26)

The supratragic notch forms an incisure within the cartilage of the ear, separating the helix root from the tragus. → **T.B.-21** is located anterior to it.

Tragus (→ Figs 3.25, 3.26)

→ **S.I.-19** is located anterior to the midpoint of the tragus.

Intertragic notch (→ Figs 3.25, 3.26)

The intertragic notch forms an incisure in the auricular cartilage, separating the tragus from the lobe and antitragus. → **G.B.-2** is located anterior to this notch.

Ear lobe (→ Figs 3.25, 3.26)

Inferior to the earlobe and the ear canal is a depression, which is bordered posteriorly by the mastoid process and anteriorly by the lower jaw. The lower border is formed by the transverse process of the 1st cervical vertebra (atlas). → **T.B.-17** is located in this depression.

Border between the auricle and the face (→ Fig. 3.26)

The border between the auricle and the temple/cheek generally forms a more or less vertical line, which will become more clearly defined by bending the auricle and tragus towards the anterior. Several points are located along this line (from top to bottom): → **T.B.-22, T.B.-21, S.I.-19, G.B.-2.**

Mastoid process (→ Figs 3.26, 3.27)

The mastoid process is a cone-shaped bony structure which can be palpated posterior to the ear. → **G.B.-12** is located at its tip, while → ***anmian*** (Ex-HN) can be found at the border of the mastoid process and the occiput. → ***yiming*** (Ex-HN-14) is located somewhat more inferiorly.

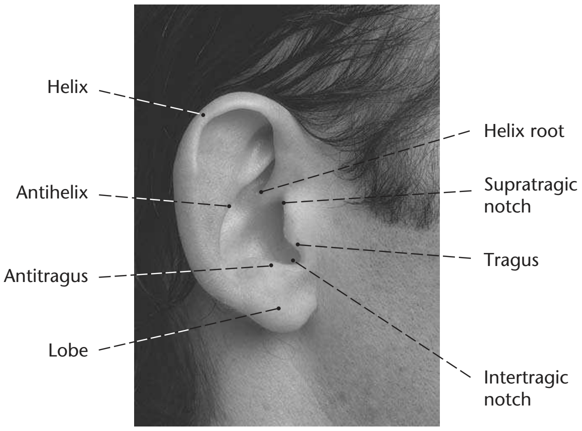
Helix
Helix root
Antihelix
Supratragic notch
Tragus
Antitragus
Lobe
Intertragic notch

Fig. 3.25

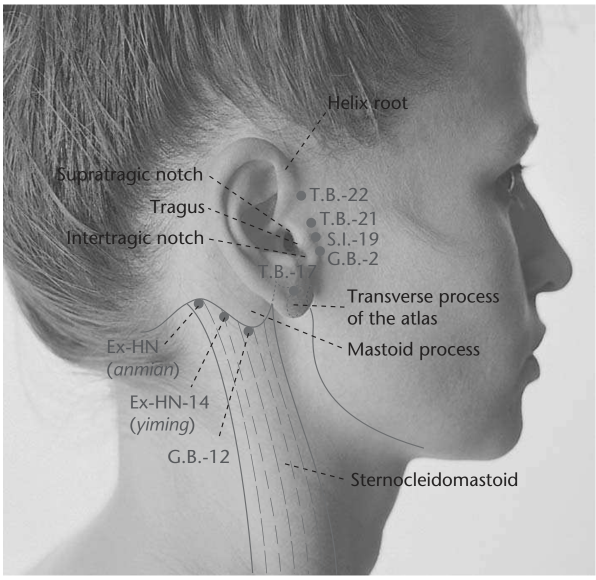
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN-14 (yiming)
G.B.-12
Sternocleidomastoid

Fig. 3.26

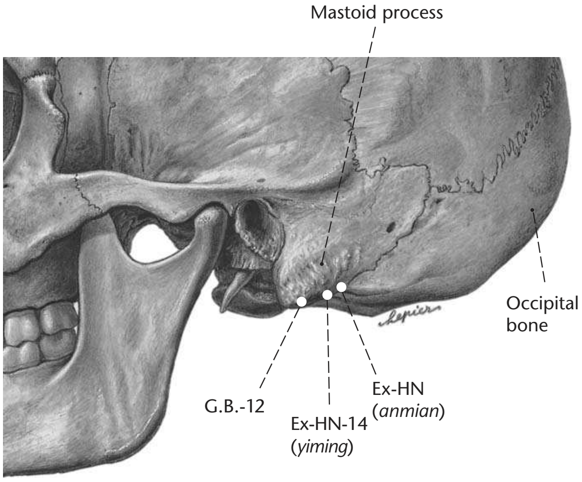
Mastoid process
Occipital bone
G.B.-12
Ex-HN-14 (yiming)
Ex-HN (anmian)

Fig. 3.27

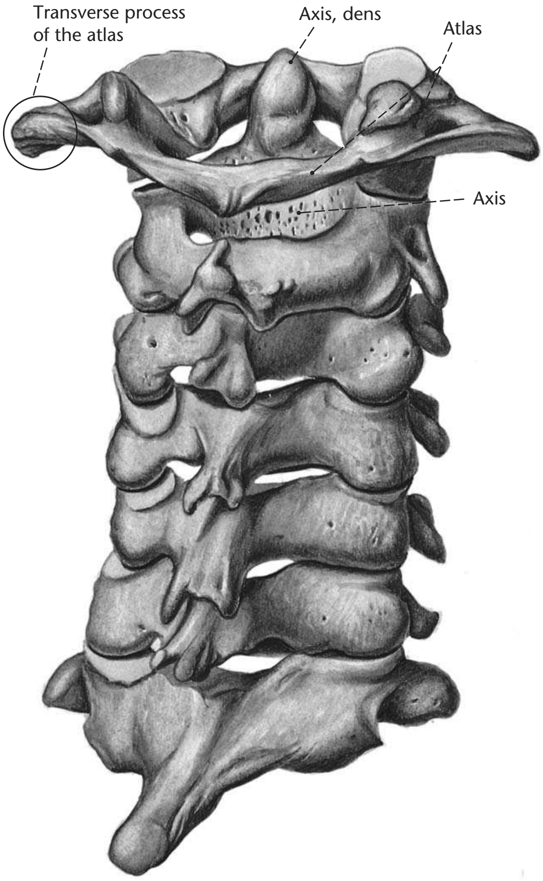
Transverse process of the atlas
Axis, dens
Atlas
Axis

Fig. 3.28

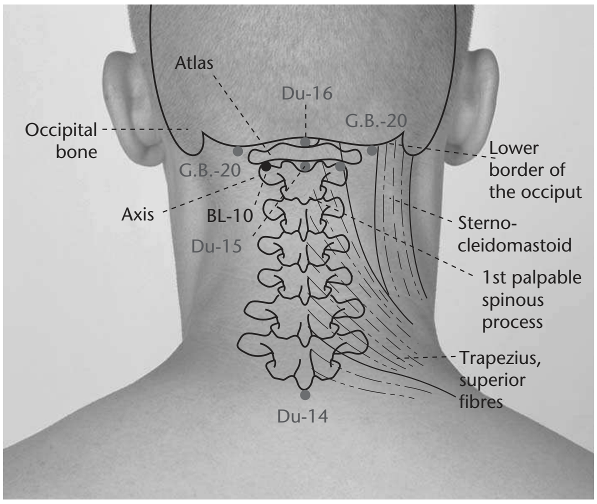
Atlas
Du-16
G.B.-20
Occipital bone
G.B.-20
Lower border of the occiput
Axis
BL-10
Sterno-cleidomastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

Fig. 3.29

Transverse process of the 1st cervical vertebra (atlas) (→ Fig. 3.28)

The transverse process of the atlas can be palpated inferior to the earlobe as a deep bony structure, which in most cases is very sensitive to pressure.

3.1.5 Occiput

Craniocervical transition (→ Figs 3.28–3.32)

The transition from the head to the neck is formed by the mastoid process, the adjacent dorsal musculature and the occiput. The following points are located in this area (from lateral to medial) → **G.B.-12,** ***yiming*** **(Ex-HN-14),** ***anmian*** **(Ex-HN), G.B.-20, BL-10, Du-15** and **Du-16.**

External occipital protuberance (→ Figs 3.32, 3.33)

The external occipital protuberance forms a flat projection on the posterior midline of the occiput, which can be palpated slightly superior to the craniocervical transition. It defines the location of → **Du-16, Du-17** and **BL-9.** Sometimes, and more frequently in women, the protuberance may be defined only poorly or not at all.

Posterior hairline (no figure)

The posterior hairline is used as a reference point for locating points on the occiput, but due to its variable position it is not a very reliable landmark.

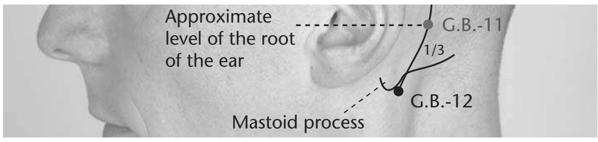
Approximate level of the root of the ear
G.B.-11
1/3
G.B.-12
Mastoid process

Fig. 3.30

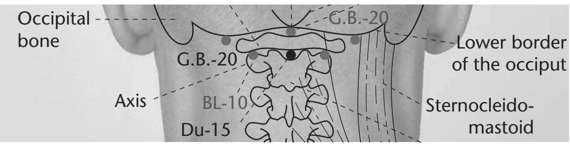
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Du-15
Sternocleido-mastoid

Fig. 3.31

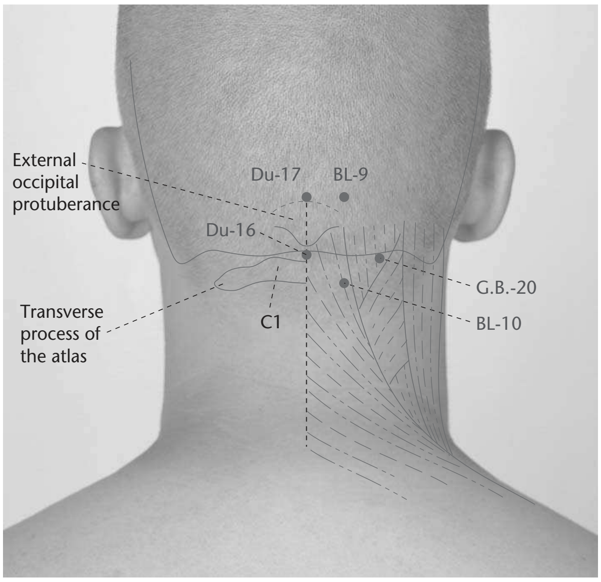
External occipital protuberance
Du-17
BL-9
Du-16
G.B.-20
BL-10
Transverse process of the atlas
C1

Fig. 3.32

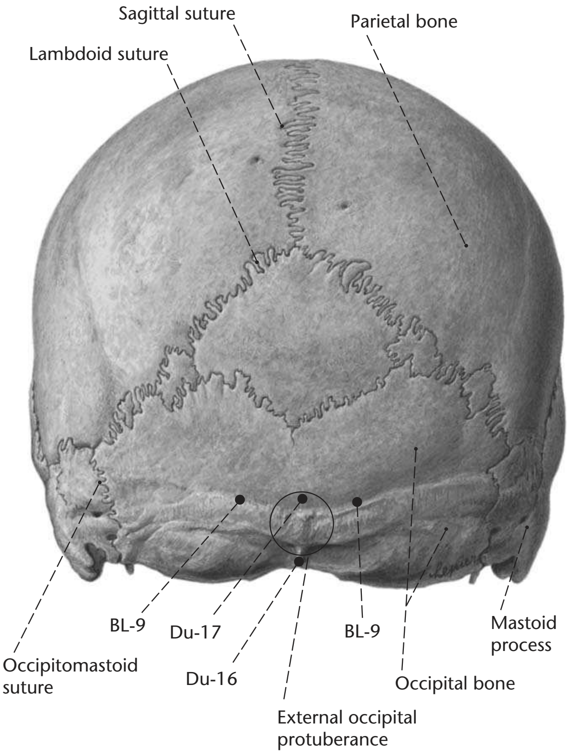
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

Fig. 3.33

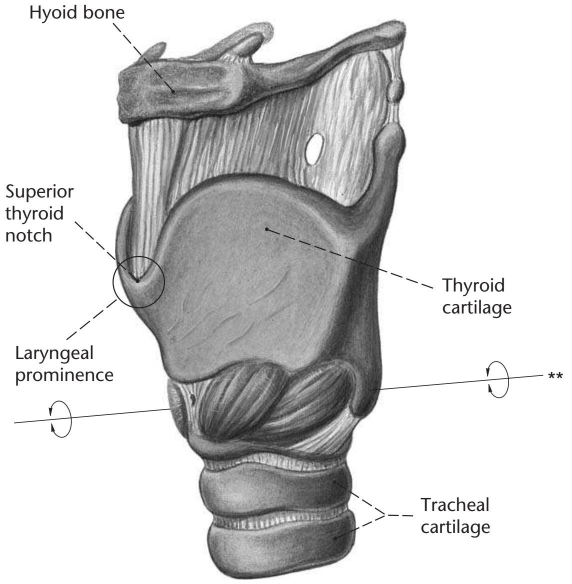
Hyoid bone
Superior thyroid notch
Laryngeal prominence
Thyroid cartilage
**
Tracheal cartilage

Fig. 3.34

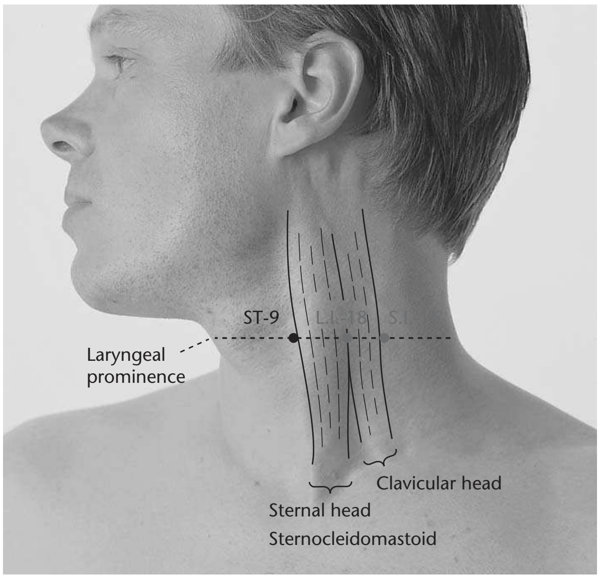
ST-9
L.I.-18
S.I.
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

Fig. 3.35

3.2 Neck

Laryngeal prominence (→ Figs 3.34, 3.35)

In men, the laryngeal prominence as the highest point of the larynx (‘Adam’s apple’) tends to form a clearly palpable and visible structure. In women, visual identification is more difficult, so that palpation is necessary in order to locate the V-shaped incisure on the upper border of the thyroid cartilage on the laryngeal midline. Located at the level of the laryngeal prominence are → **ST-9, L.I.-18** and **S.I.-16**.

Sternocleidomastoid (SCM) muscle (→ Figs 3.36, 3.37)

This muscle forms a distinct, easily identifiable structure on the anterior and lateral aspect of the neck, which becomes visible and palpable by rotating the neck in the opposite direction. This effect can be increased by rotating the head against resistance (by pushing against the chin for example). The more slender clavicular head originates from the anterior aspect of the manubrium, while the more flat clavicular head originates from the medial third of the clavicle, revealing a small triangle between the two heads near the sternoclavicular joint. This is the location of → **ST-11**. By palpating along the SCM muscle in a superior direction, the two heads can be distinguished for the major part of the muscle, before they merge to insert at the posterior part of the mastoid process and the lateral superior nuchal line of the occiput. Besides → **ST-11,** the following points are located alongside the muscle belly of the SCM: → **L.I.-17, L.I.-18, S.I.-16, ST-9, G.B-12** and ***anmian*** **(Ex-HN)**.

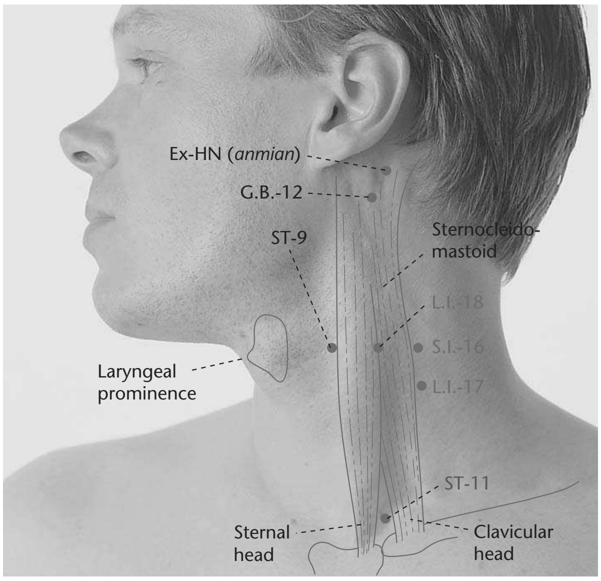
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleido-mastoid
L.I.-18
S.I.-16
Laryngeal prominence
L.I.-17
ST-11
Sternal head
Clavicular head

Fig. 3.36

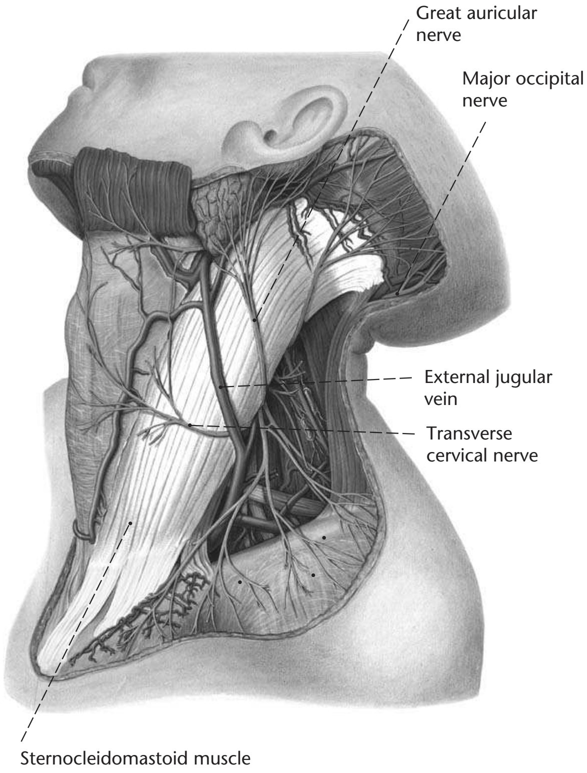
Great auricular nerve
Major occipital nerve
External jugular vein
Transverse cervical nerve
Sternocleidomastoid muscle

Fig. 3.37

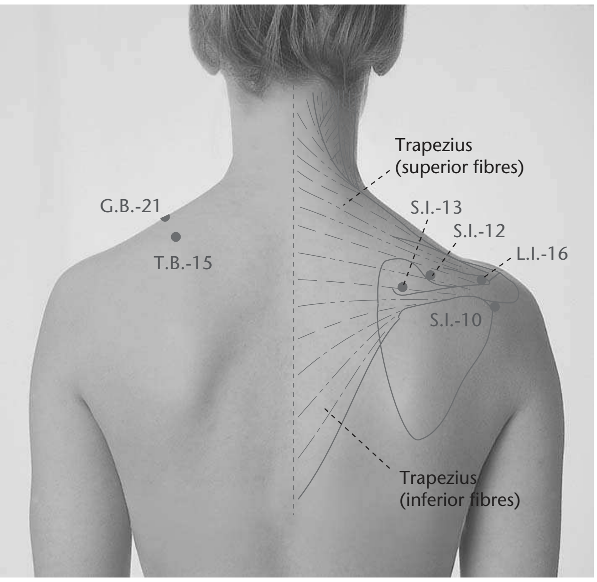
Trapezius
(superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10
Trapezius
(inferior fibres)

Fig. 3.38

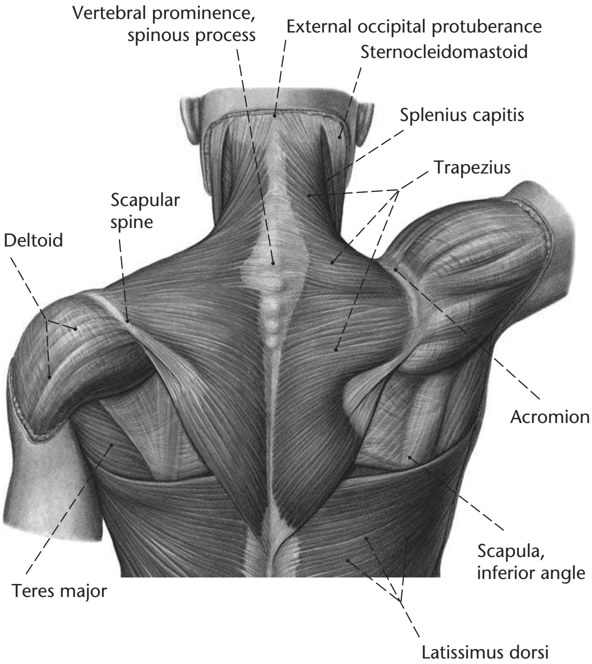
Vertebral prominence,
spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular
spine
Deltoid
Acromion
Scapula,
inferior angle
Teres major
Latissimus dorsi

Fig. 3.39

3.3. Shoulder and Arm

3.3.1 Shoulder and Upper Arm

Trapezius muscle (→ Figs 3.38, 3.39)

This muscle covers the superior aspect of the shoulder, its superior portions extending from the cervical spine to the acromion at the lateral end of the scapula. Located at the midpoint of the superior border of the muscle are → **G.B.-21** and, slightly inferiorly, → **T.B.-15.**

Acromion (→ Figs 3.39, 3.42, 3.43)

When palpating the shoulder joint from a lateral direction, the acromion can be felt as a bony extension with a flat surface, slightly superior to the laterally protruding head of the humerus.

Scapular spine (→ Figs 3.39–3.44)

The scapular spine forms a bony ridge that originates at the acromion and runs in an oblique angle across the scapula, ending at its medial border in an upward opening curve. This is the location of → **S.I.-13**, → **L.I.-16** is located on the most lateral portion of the trapezius and supraspinatus muscles, in the angle between the acromial origin of the scapular spine and the acromioclavicular joint.

Head of the humerus, shoulder hollows (→ Figs 3.42, 3.44)

The head of the humerus is located inferior to the acromion and protrudes slightly on the lateral aspect of the shoulder. When abducting the arm horizontally, two hollows will form at the transition from the arm to the shoulder. The anterior hollow is the location of → **L.I.-15**, while → **T.B.-14** is located in the posterior one. The hollows mark the border between the clavicular and acromial portion and the border between the acromial and spinal portion of the deltoid muscle.

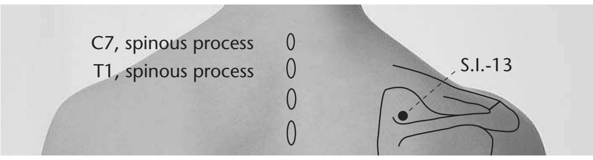
C7, spinous process
T1, spinous process
S.I.-13

Fig. 3.40

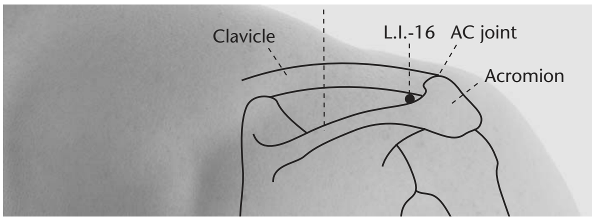
Clavicle
L.I.-16
AC joint
Acromion

Fig. 3.41

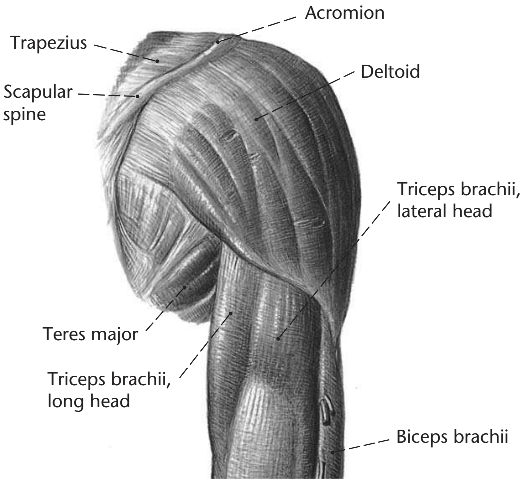
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

Fig. 3.42

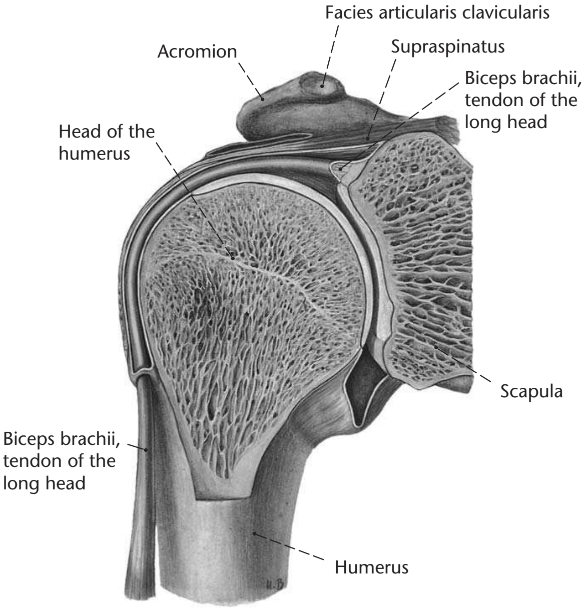
Facies articularis clavicularis
Acromion
Supraspinatus
Biceps brachii, tendon of the long head
Head of the humerus
Scapula
Biceps brachii, tendon of the long head
Humerus

Fig. 3.43

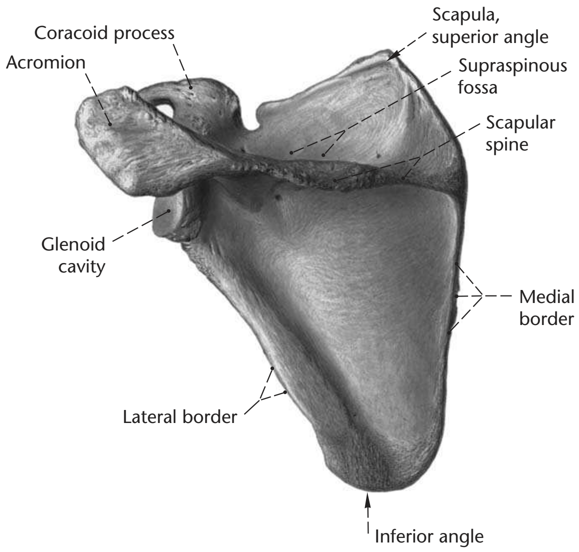
Scapula,
superior angle
Coracoid process
Acromion
Supraspinous
fossa
Scapular
spine
Glenoid
cavity
Medial
border
Lateral border
Inferior angle

Fig. 3.44

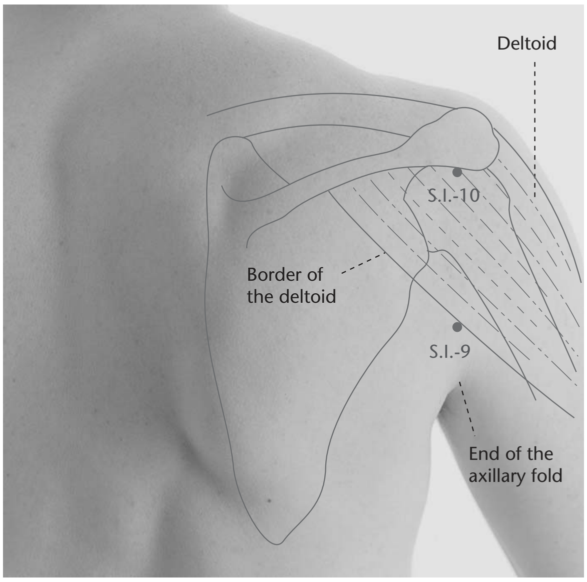
Deltoid
S.I.-10
Border of
the deltoid
S.I.-9
End of the
axillary fold

Fig. 3.45

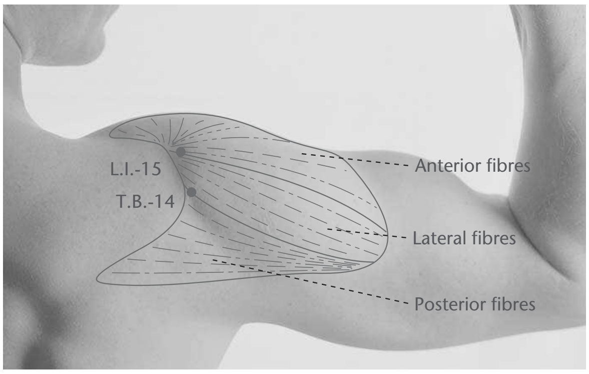
L.I.-15
T.B.-14
Anterior fibres
Lateral fibres
Posterior fibres

Fig. 3.46

Deltoid muscle (→ Figs 3.42, 3.45–3.51)

This muscle covers the anterior, posterior and lateral aspect of the head of the humerus. The following points are located either on the muscle or on its margins: → **S.I.-9, S.I.-10, T.B.-14, L.I.-15,** ***jianquan*** **(Ex-UE), LU-1, LU-2, LU-3, L.I.-14**.

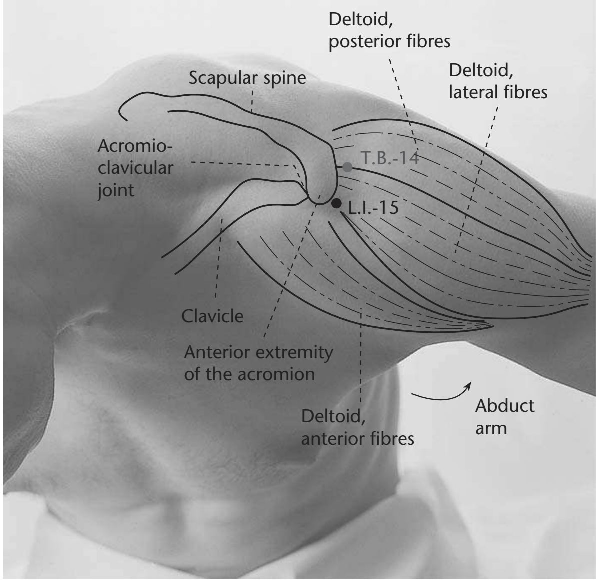
Deltoid,
posterior fibres
Scapular spine
Deltoid,
lateral fibres
Acromio-
clavicular
joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity
of the acromion
Deltoid,
anterior fibres
Abduct
arm

Fig. 3.47

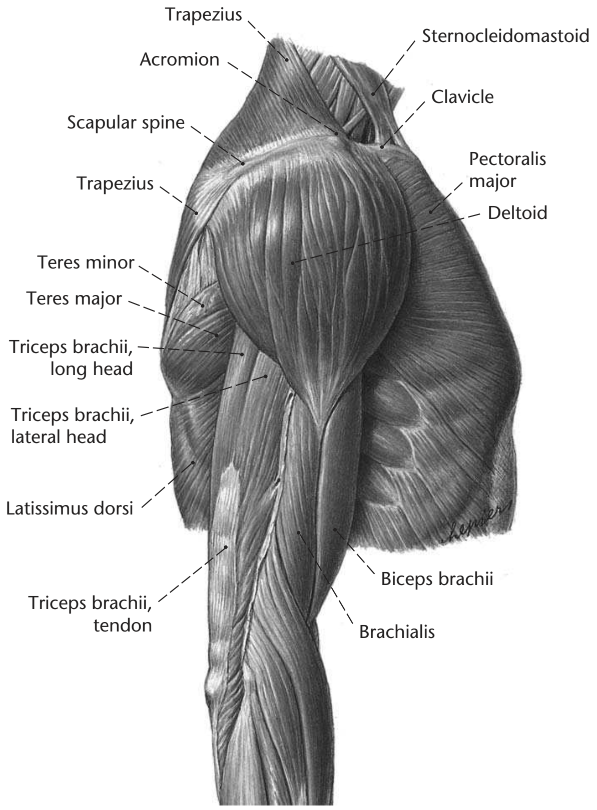
Trapezius
Acromion
Scapular spine
Trapezius
Teres minor
Teres major
Triceps brachii, long head
Triceps brachii, lateral head
Latissimus dorsi
Triceps brachii, tendon
Sternocleidomastoid
Clavicle
Pectoralis major
Deltoid
Biceps brachii
Brachialis

Fig. 3.48

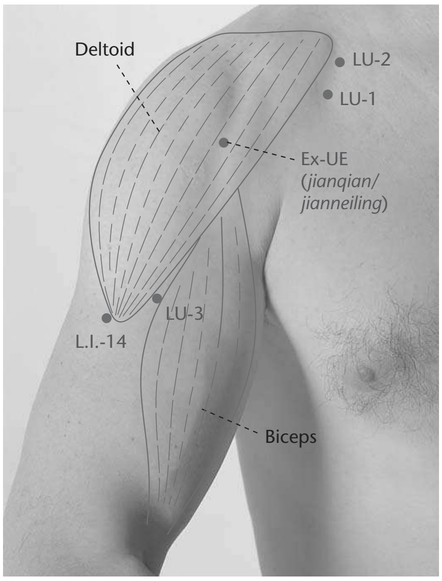
Deltoid
LU-2
LU-1
Ex-UE
(*jianqian/ jianneiling*)
LU-3
L.I.-14
Biceps

Fig. 3.50

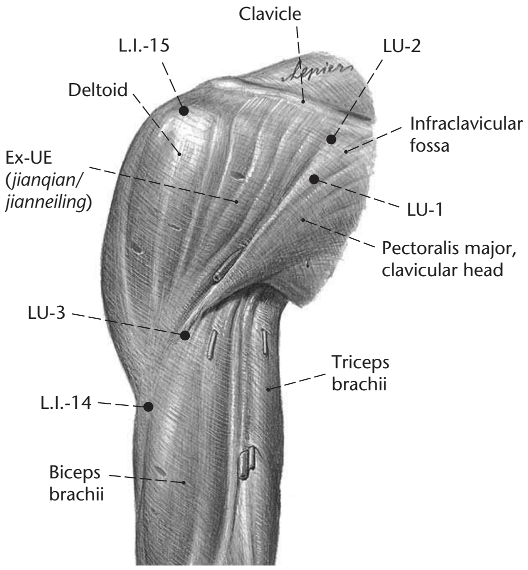
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Ex-UE
(*jianqian/ jianneiling*)
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14
Biceps brachii

Fig. 3.51

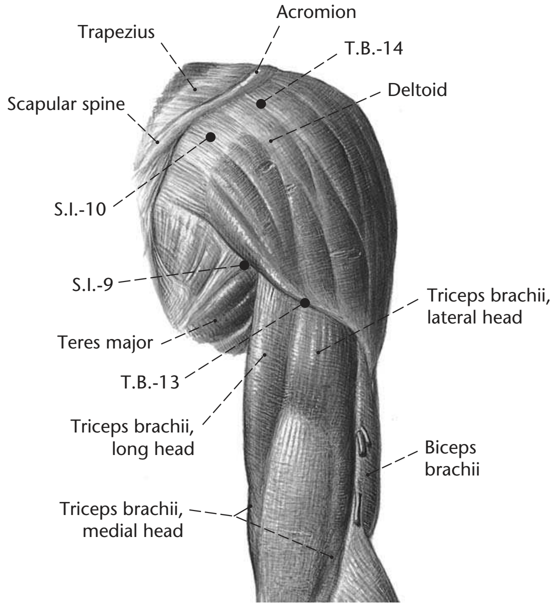
Trapezius
Acromion
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

Fig. 3.49

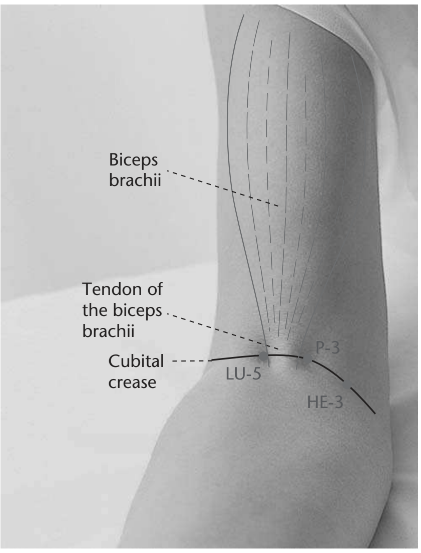
Biceps brachii
Tendon of the biceps brachii
Cubital crease
LU-5
P-3
HE-3

Fig. 3.52

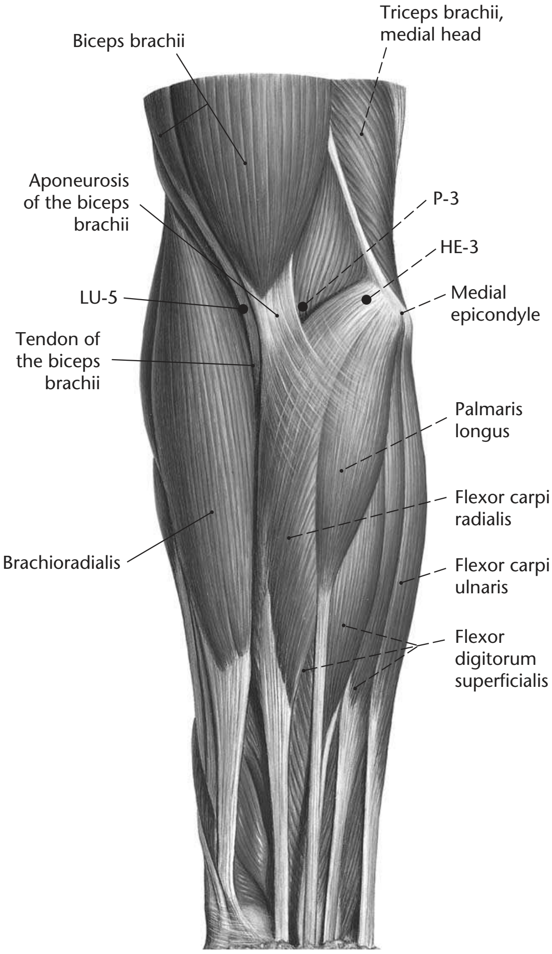
Biceps brachii
Triceps brachii, medial head
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus
Flexor carpi radialis
Brachioradialis
Flexor carpi ulnaris
Flexor digitorum superficialis

Fig. 3.53

3.3.2 Elbow

Elbow, tendon of the biceps muscle (→ Figs 3.52, 3.53)

The anterior aspect of the elbow is marked by the cubital crease, extending from the lateral to the medial epicondyle of the humerus. It is divided by the tendon of the biceps muscle. → **L.I.-11** and → **LU-5** are located lateral to the tendon while → **P-3** and → **HE-3** are located medial to it.

Lateral and medial epicondyles of the humerus (→ Figs 3.53, 3.55, 3.56)

The lateral and medial epicondyles of the humerus form the lateral aspects of the elbow. The olecranon is located between them, on the posterior aspect of the elbow. The medial epicondyle is the origin of the extensor muscles of the wrist joint, while the flexor muscles originate at the medial epicondyle.

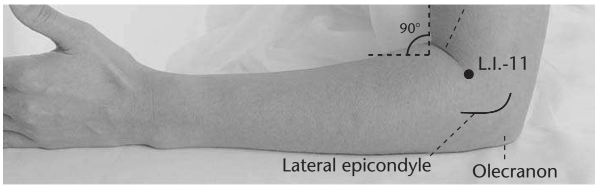
90°
L.I.-11
Lateral epicondyle
Olecranon

Fig. 3.54

Olecranon (→ Figs 3.55, 3.56)

The olecranon forms the tip of the elbow. Located in the ulnar sulcus, between the medial epicondyle and the tip of the olecranon, is → **S.I.-8**, while → **T.B.-10** is located in a depression proximal to the olecranon.

3.3.3 Forearm and Hand

For locating points on the forearm, it is important to note whether the arm is in a pronated or supinated position, since this can have a considerable effect on the relative location of some points (see also location details of specific points). For example, in supination the line connecting → **L.I.-5** and → **L.I.-11** runs along the lateral border of the forearm, while in pronation it traverses the forearm.

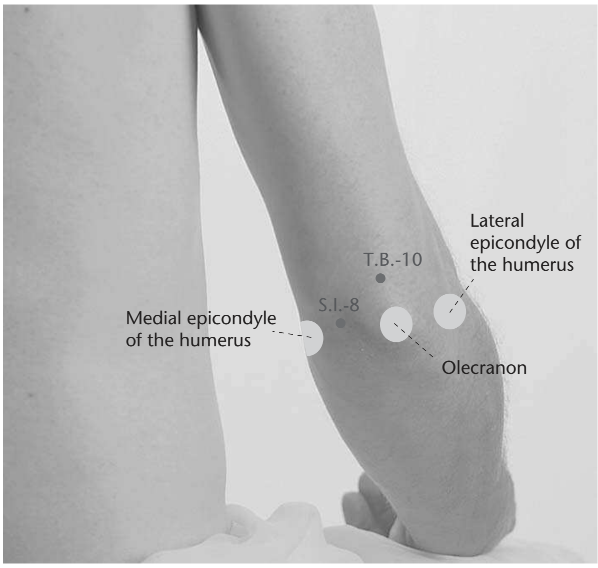
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

Fig. 3.55

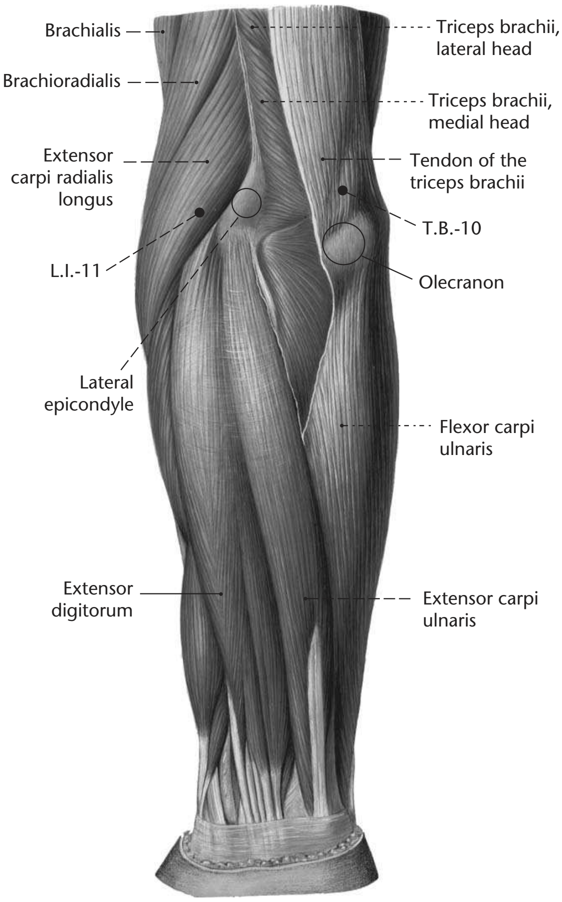
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Extensor digitorum
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris
Extensor carpi ulnaris

Fig. 3.56

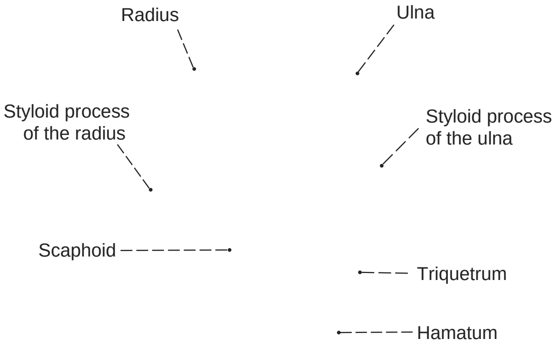
Radius
Ulna
Styloid process of the radius
Styloid process of the ulna
Scaphoid
Triquetrum
Hamatum

Fig. 3.57 Wrist joint, posterior view

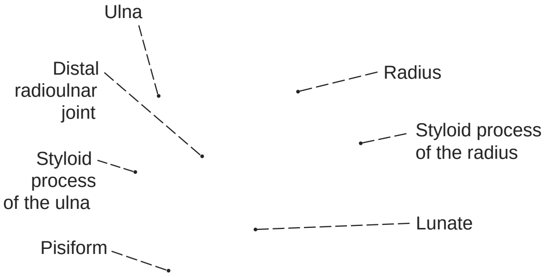
Ulna
Distal radioulnar joint
Radius
Styloid process of the radius
Styloid process of the ulna
Lunate
Pisiform

Fig. 3.58 Wrist joint, anterior view

Styloid process of the radius
(Figs 3.57, 3.58)

The styloid process forms a flat mound on the anterior and lateral aspect of the distal end of the radius. An oblique groove can be palpated at the transition from the process to the shaft of the radius. This is the location of LU-7. LU-8 is located slightly more distally and anteriorly, at the border of the process.

Styloid process of the ulna
(Figs 3.57, 3.58)

The styloid process of the ulna is located at the distal end of the ulna, forming a distinct prominence. With the elbow flexed and the forearm in a supinated position, a small groove becomes palpable (sliding groove of the tendon of the extensor carpi ulnaris muscle). This is the location of S.I.-6.

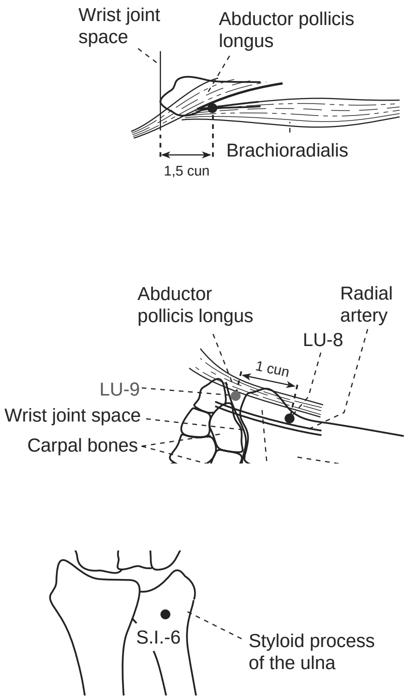
Wrist joint space
Abductor pollicis longus
Brachioradialis
1,5 cun
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Wrist joint space
Carpal bones
S.I.-6
Styloid process of the ulna

Fig. 3.59

Lumbosacral junction

Following orientation in the lumbar region with the help of the iliac crests (→ Iliac crest), the lumbosacral joint can be located by counting the spinous processes in an inferior direction. It can generally be felt as a marked 'kink' or depression, followed by the smaller spinous processes of the slightly convex median sacral crest directly below.

Caution: Note that (according to the literature in up to 15% of cases) there can be lumbarization of the 1st sacral vertebra or sacralization of L5. In these cases there will be either four or six lumbar vertebrae.

Posterior superior iliac spine (PSIS) (→ Figs 3.73, 3.74)

Another important landmark on the lower back is the posterior superior iliac spine (PSIS), which forms the posterior end of the iliac crest lateral to the upper sacral region. The PSIS can often be identified by two superficially visible dimples. It is best located by palpating from an inferior in a superior direction. If the dimples are not visible, they can be located by palpating 3 cun from the rima ani in a superior and lateral direction at an angle of 45°, until a marked bony ridge can be felt.

Generally, the PSIS is located at the level of the space between the 1st and 2nd sacral foramina, so that → **BL-27** is located directly medial or slightly superior and medial to the PSIS, while → **BL-28** is located just inferior and medial to the PSIS.

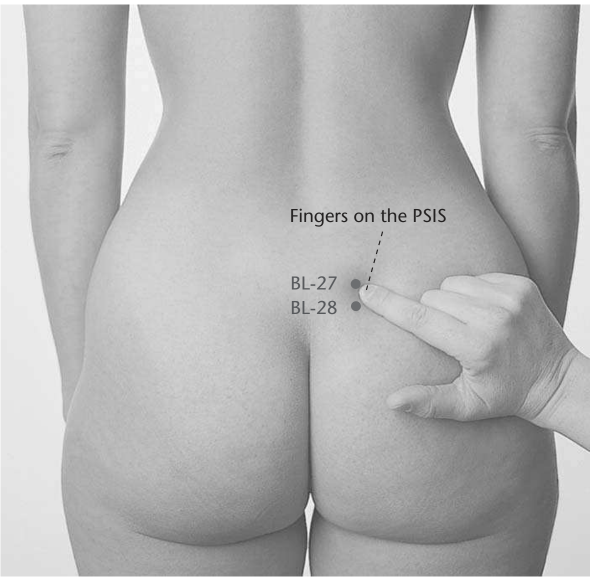

Fig. 3.73

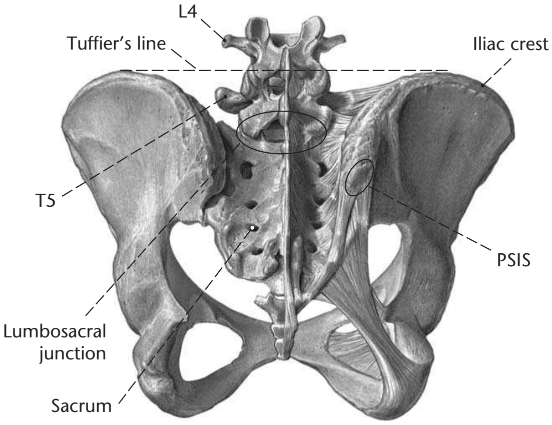

Fig. 3.74

3.4.4 Sacrum and Sacroiliac Joint

Sacrum (→ Figs 3.75, 3.76)

On the sacrum, landmarks for palpation are the median sacral crest and the sacral hiatus as well as the sacral foramina on its lateral aspects.

Sacral crest (→ Figs 3.74, 3.75)

The sacral crest is formed by the narrow, irregular spinous processes, which can be palpated on the sacral midline.

Sacral hiatus (→ Figs 3.75, 3.76)

The sacral hiatus can be palpated as a U-shaped depression at the inferior end of the sacral crest. → **Du-2** is located on the midline, just inferior to it. Both the sacral hiatus and the greater trochanter (→ Lower extremity) are reference points for locating → **G.B.-30.**

Sacral foramina (→ Figs 3.75, 3.76)

The sacral foramina are located between the lumbosacral joint and the sacral hiatus and can usually be palpated. They are in most cases evenly spaced and lie approximately 1 finger-width lateral to the midline, with the distance decreasing in a distal direction. The foramina are the location of → **BL-31–BL-34.**

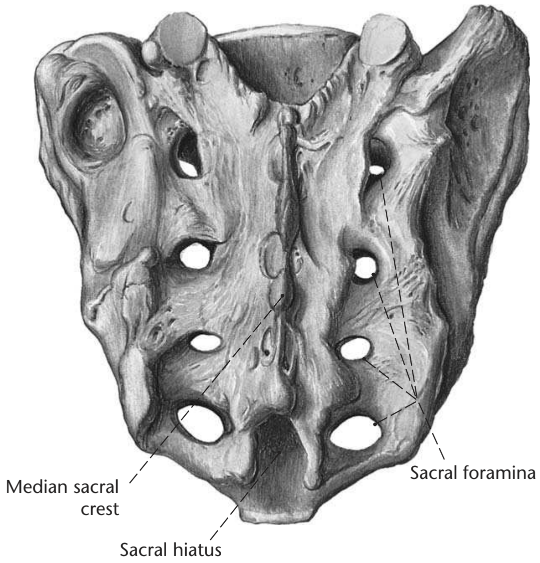
Median sacral crest
Sacral foramina
Sacral hiatus

Fig. 3.75

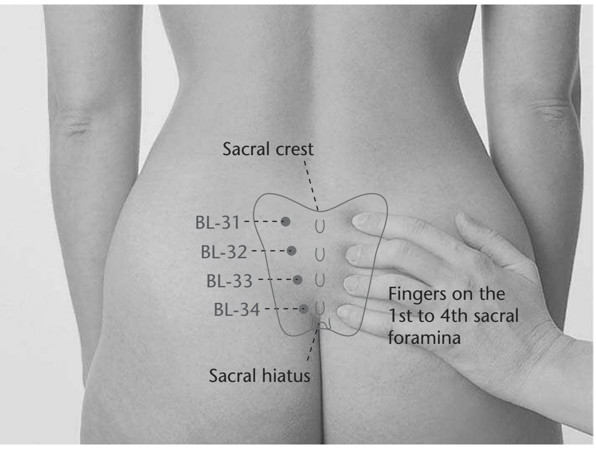
Sacral crest
BL-31
BL-32
BL-33
BL-34
Fingers on the 1st to 4th sacral foramina
Sacral hiatus

Fig. 3.76

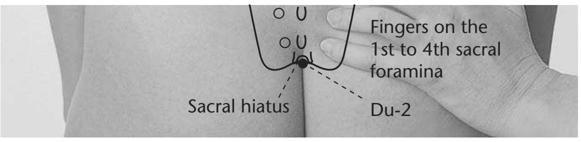
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

Fig. 3.77

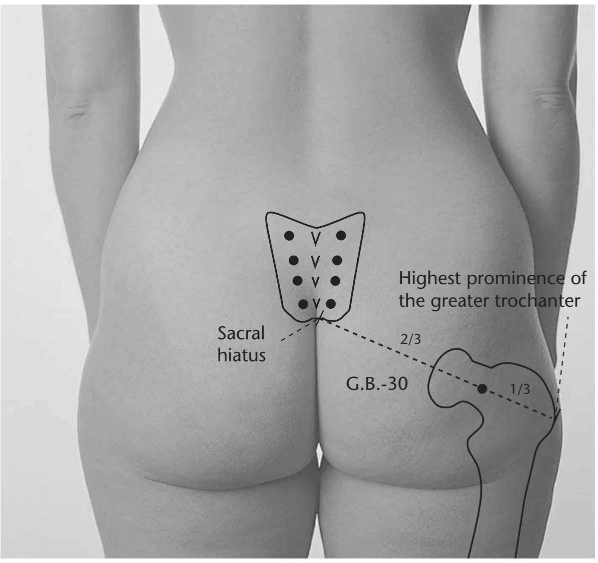
Highest prominence of the greater trochanter
2/3
Sacral hiatus
G.B.-30
1/3

Fig. 3.78

Iliosacral joint (→ Figs 3.80, 3.81)

The iliosacral joint can be palpated bilaterally as a shallow groove between the sacrum, the PSIS (see above) and more inferiorly located ilial structures. → **BL-26–BL-29** are located in this area.

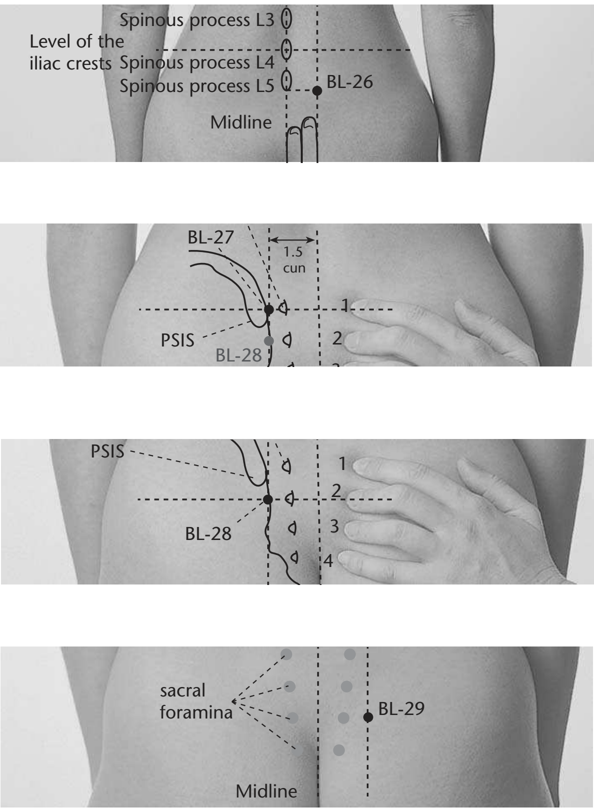
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline
BL-27
1.5 cun
1
2
PSIS
BL-28
PSIS
1
2
BL-28
3
4
sacral foramina
BL-29
Midline

Fig. 3.79

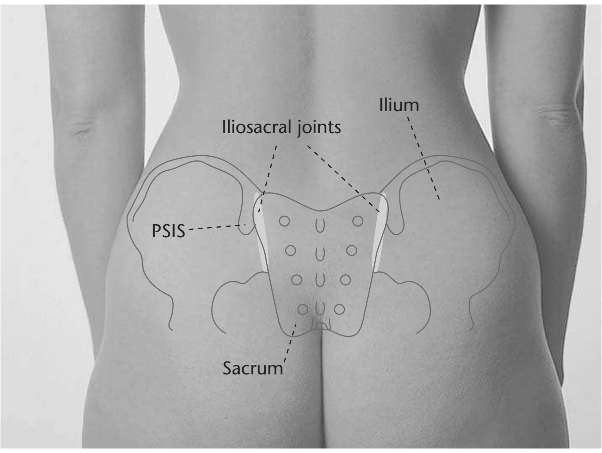
Ilium
Iliosacral joints
PSIS
Sacrum

Fig. 3.80

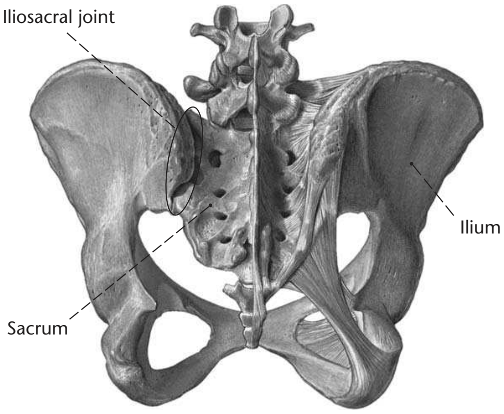
Iliosacral joint
Ilium
Sacrum

Fig. 3.81

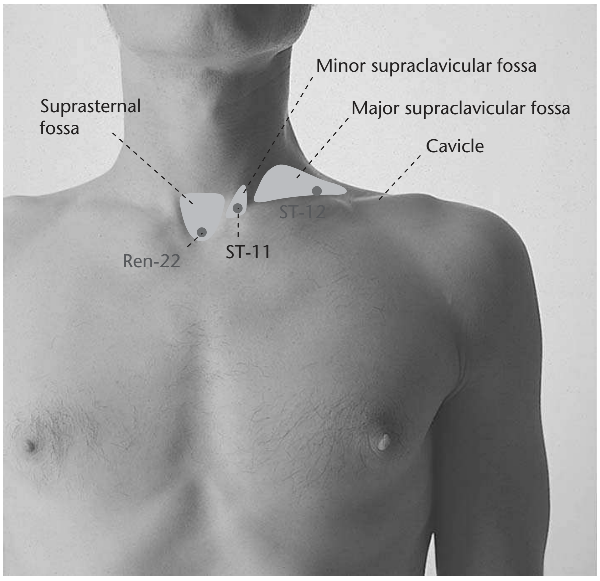
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Cavicle
ST-12
Ren-22
ST-11

Fig. 3.82

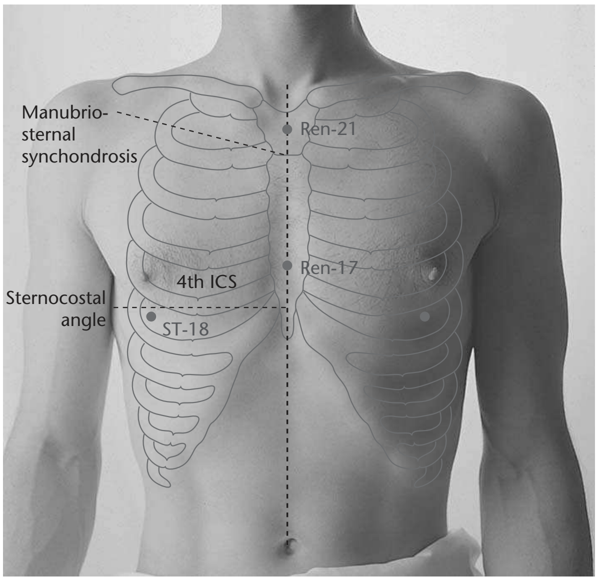
Manubrio-sternal synchondrosis
Ren-21
4th ICS
Ren-17
Sternocostal angle
ST-18

Fig. 3.83

3.5 Anterior and Lateral Aspect of the Thorax and Abdomen

Clavicle (→ Figs 3.82, 3.86)

The clavicle forms the connection between the shoulder girdle and the thorax. It separates the major and minor supraclavicular fossae with → **ST-11, ST-12** and ***jingbi*** (Ex-HN) from the ribcage. Located on the lower border of the clavicle are → **ST-13** (at its midpoint) and → **KID-27** (2 cun lateral to the midline).

Suprasternal fossa (→ Fig. 3.82)

The suprasternal fossa forms a depression superior to the sternum between the sternoclavicular joints. It is the location of → **Ren-22**.

Ribs (→ Figs 3.83, 3.86)

For anatomical orientation of the ribs, the midsection of the clavicle is a good starting point. The first rib lies directly below the clavicle, forming a narrow bow. It articulates with the manubrium just inferior to the sternoclavicular joint.

Please note the following when palpating the ribs and intercostal spaces (ICS) and locating the respective acupuncture points:

- From their articulation with the sternum, the ribs run initially horizontally or in a gentle curve in an inferior direction, while at the lateral aspect of the thorax, they assume a superior direction towards the posterior aspect of the body.
- Inferior to the 4th/5th intercostal spaces, the medial sections of the ribs merge with the costal cartilage, so that the intercostal spaces do not extend very far medially. Therefore it is necessary to palpate somewhat more laterally.

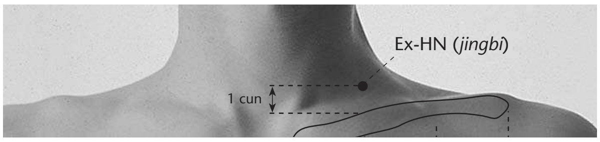
Ex-HN (*jingbi*)
1 cun

Fig. 3.84A

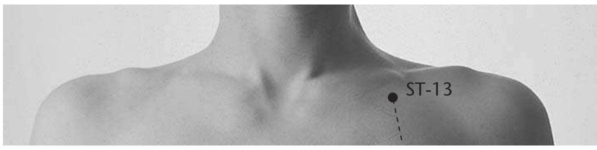

Fig. 3.84B

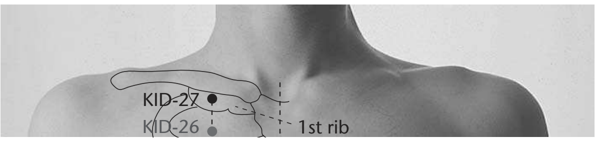
KID-27
KID-26
1st rib

Fig. 3.84C

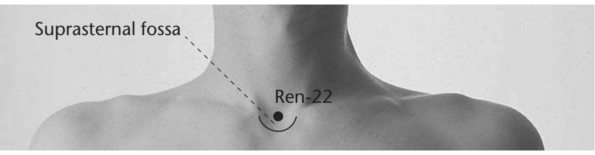
Suprasternal fossa
Ren-22

Fig. 3.84D

1st rib, 1st intercostal space (→ Fig. 3.86)

By palpating parasternally in an inferior direction, the first rib can generally be felt directly below the clavicle. Here, → **KID-27** is located 2 cun lateral to the anterior midline. Below follows the first intercostal space, then the second rib and so on. → **Ren-20** is located on the anterior midline, at the level of the first intercostal space, → **Ren-21** is located slightly superior to it.

2nd rib, manubriosternal synchondrosis (manubriosternal junction) (→ Figs 3.83, 3.86)

The synchondrosis is usually a clearly defined transverse bony structure on the superior section of the sternum, level with the articulation of the second rib. Directly below the synchondrosis is the 2nd intercostal space from where subsequent ribs and intercostal spaces can be counted by palpation.

Laterosuperior aspect of the thorax and coracoid process (→ Fig. 3.87)

The laterosuperior aspect of the thorax is bordered by the clavicle and the deltoid muscle with the coracoid process deep to the latter. The deltopectoral triangle can be located by palpating from the end of the anterior axillary fold along the border of the deltoid muscle towards the shoulder and the lower border of the clavicle. Here, in the centre of the triangle, lies → **LU-2**. Laterally, the triangle is bordered by a clearly palpable bony structure, the coracoid process. In order to distinguish the process from the lesser tubercle of the humerus, the arm is rotated externally with the elbow flexed at the same time. While the process will remain fixed, the tubercle will follow the movement. → **LU-1** is located slightly superior and lateral to → **LU-2** and medial to the lower border of the coracoid process. The contours of the deltoid muscle and the deltopectoral triangle can be emphasised through muscle flexion, for example by pressing the hands against each other in front of the chest.

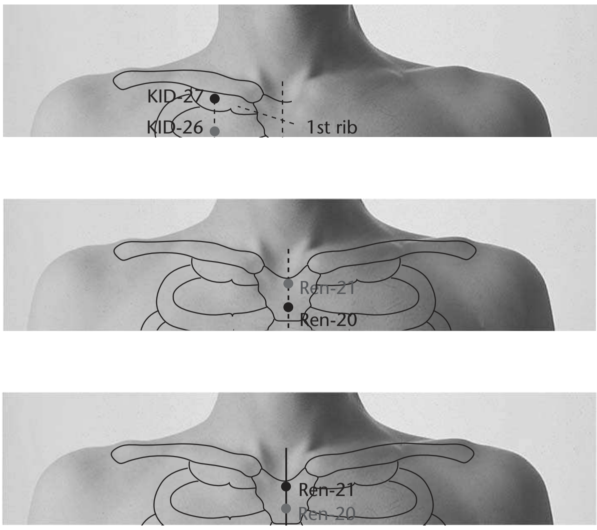
KID-27
KID-26
1st rib
Ren-21
Ren-20
Ren-21
Ren-20

Fig. 3.85

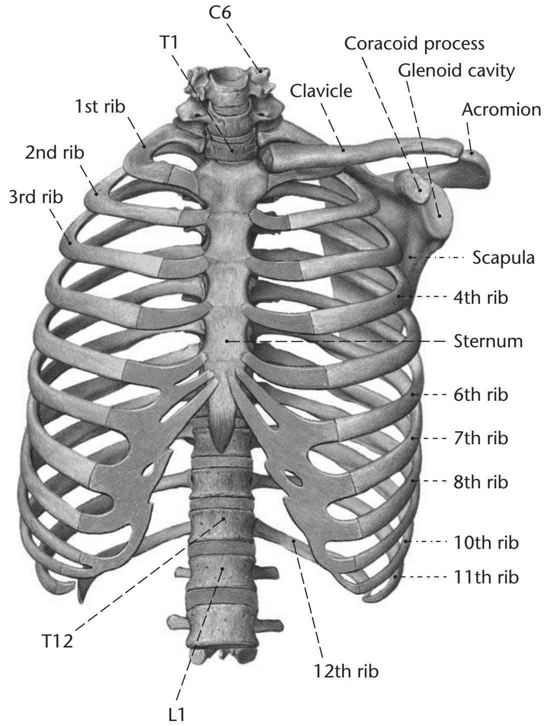
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Fig. 3.86

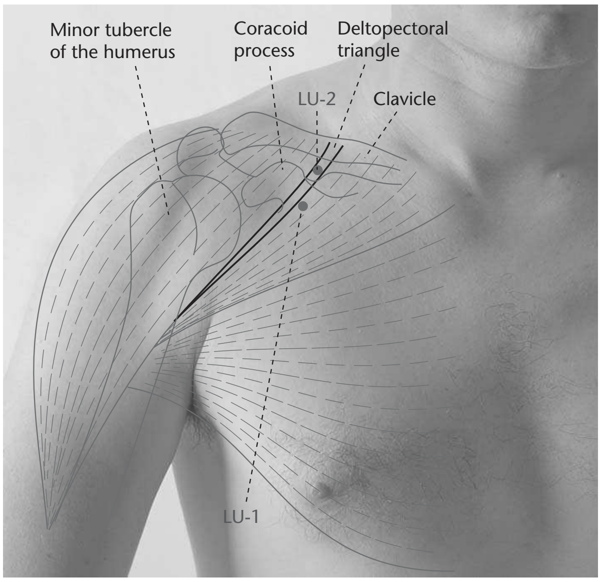
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Clavicle
LU-1

Fig. 3.87

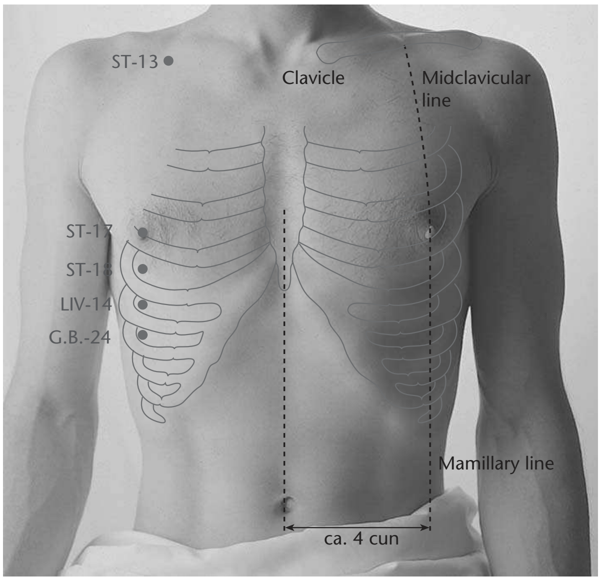

Fig. 3.88

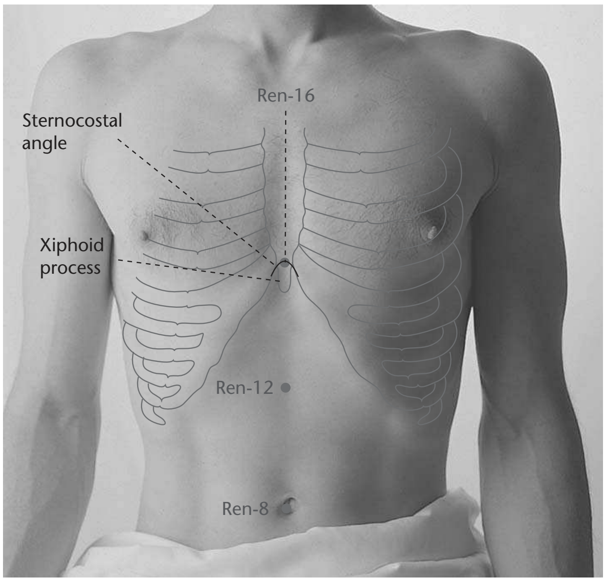

Fig. 3.89

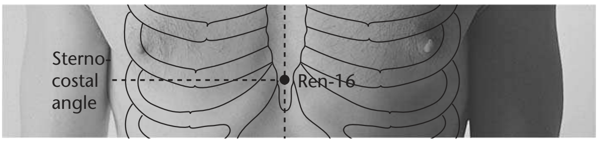

Fig. 3.90

Midclavicular line (→ Fig. 3.88)

The midclavicular line provides a further tool for reference on the upper thorax. It is an imaginary line that runs from the midpoint of the clavicle to the nipple. Since in most cases the nipple lies somewhat more laterally than the clavicular midpoint, the line tends to have a slightly oblique pathway. → **ST-13** to **ST-16** are all located on the line.

Nipple (→ Fig. 3.88)

In men, the nipple is located in the 4th intercostal space, lateral to the midline, while in women, the location can vary depending on the shape of the breast. → **ST-17** is located on the nipple.

Mamillary line (→ Fig. 3.88)

The mamillary line is an imaginary line for orientation on the mediolateral thorax and abdomen. It runs vertically from the nipple in an inferior direction. The points → **ST-18, LIV-14** and **G.B.-24** are all located on this line.

Sternocostal angle, xiphoid process and umbilicus (→ Figs 89, 3.91)

The lower borders of the ribcage unite at the lower sternal border, forming the sternocostal angle. This angle and the umbilicus are the two major anatomical landmarks for point location on the upper abdomen.

Caution: It is important not to mistake the sternocostal angle for the tip of the xiphoid process. The latter is attached to the sternum at the sternocostal angle, projecting inferiorly towards the abdomen. For definite identification, palpate along the costal cartilage in a superior direction to where it meets the sternocostal angle. This is the location of → **Ren-16**.

Axillary line (→ Fig. 3.92)

The axillary line is an imaginary line for orientation on the lateral aspect of the thorax and abdomen. It runs vertically from the middle of the axilla in an inferior direction. Both → **G.B.-22** and **SP-21** are located on this line.

Free end of the 11th rib (→ Figs 3.91, 3.92)

The free end of the 11th rib can be palpated below the costal cartilage, at the transitional zone from the thorax to the abdomen. Location tip: the patient places his/her flexed elbow (90° angle) alongside his/her thorax. The tip of the olecranon will then usually be close to the free end of the 11th rib. → **LIV-13** is located below its anterior lower border.

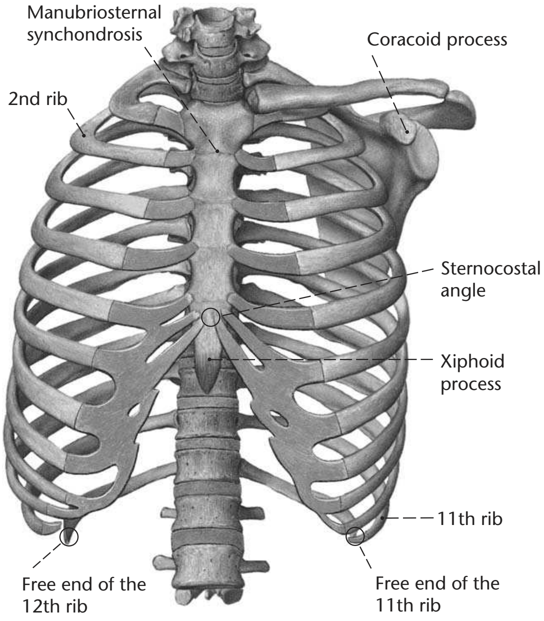
Manubriosternal synchondrosis
Coracoid process
2nd rib
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

Fig. 3.91

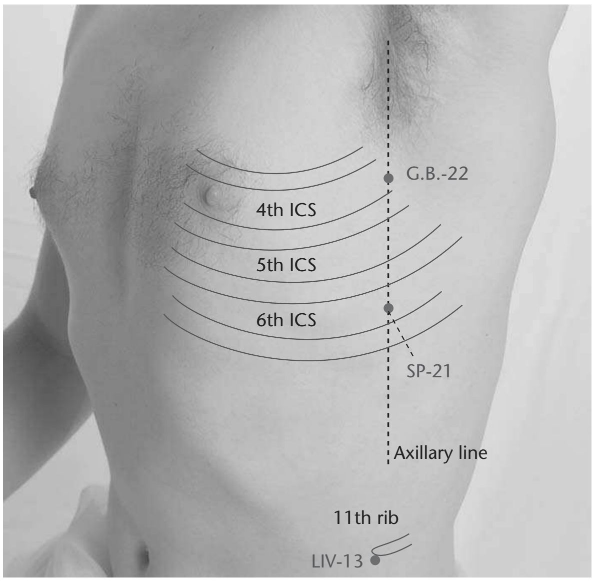
G.B.-22
4th ICS
5th ICS
6th ICS
SP-21
Axillary line
11th rib
LIV-13

Fig. 3.92

Anterior superior iliac spine (ASIS) (→ Figs 3.93–3.95)

By palpating along the iliac crest in an anterior and inferior direction, the anterior superior iliac spine (ASIS) is reached at the lateral end of the inguinal groove. The ASIS is felt as a distinct bony ridge. → **G.B.-27** is located anterior to it, while → **G.B.-28** is located slightly more inferiorly and medially.

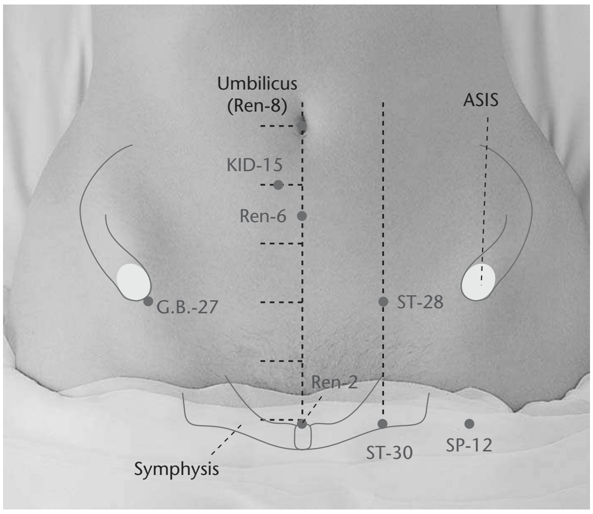
Umbilicus
(Ren-8)
ASIS
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Fig. 3.93

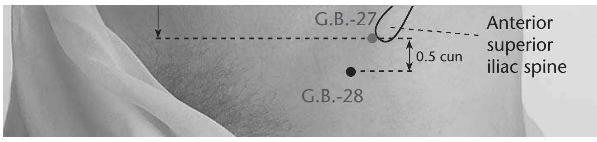
G.B.-27
Anterior
superior
iliac spine
0.5 cun
G.B.-28

Fig. 3.96

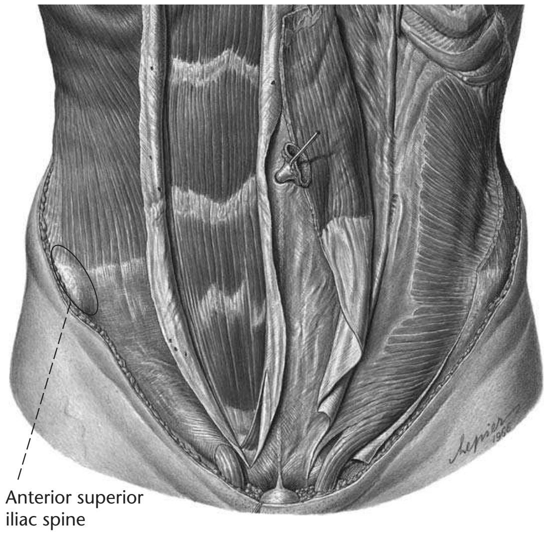
Anterior superior
iliac spine

Fig. 3.94

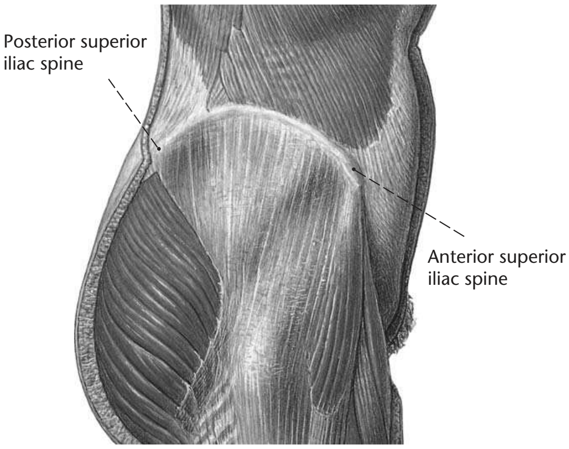
Posterior superior
iliac spine
Anterior superior
iliac spine

Fig. 3.95

Pubic symphysis, umbilicus (→ Fig. 3.98)

The upper border of the pubic symphysis is an important orientation point for point location on the lower abdomen. The distance between the centre of the umbilicus and the upper border of the symphysis can be divided into 5 proportional cun, which can differ significantly from the patient's finger cun (→ Chapter 2). → **Ren-2** is located on the anterior midline and on the upper border of the pubic symphysis.

3.6 Lower Extremity

Greater trochanter (→ Fig. 3.99)

The greater trochanter forms a distinct, projecting bony structure on the hip area. It is one of the endpoints on an imaginary line connecting it with the sacral hiatus. Located on this line is → **G.B.-30**, a third of the distance from the greater trochanter. Another line connecting the greater trochanter and the anterior superior iliac spine (ASIS) is used for locating → **G.B.-29**.

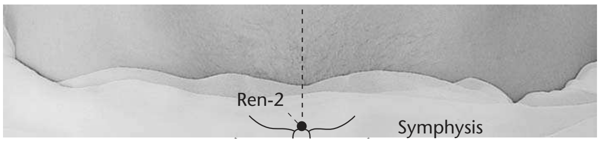
Ren-2
Symphysis

Fig. 3.97

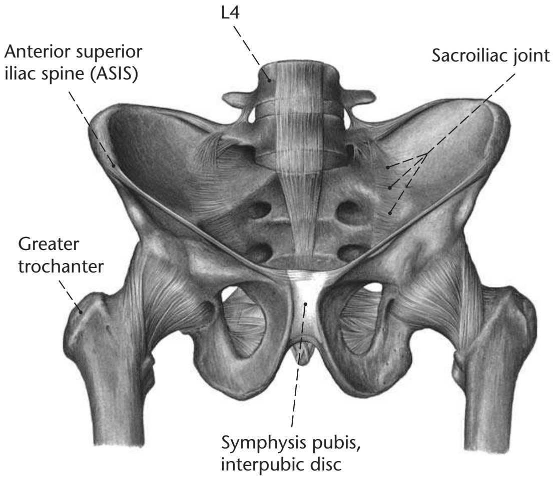
L4
Anterior superior iliac spine (ASIS)
Sacroiliac joint
Greater trochanter
Symphysis pubis, interpubic disc

Fig. 3.98

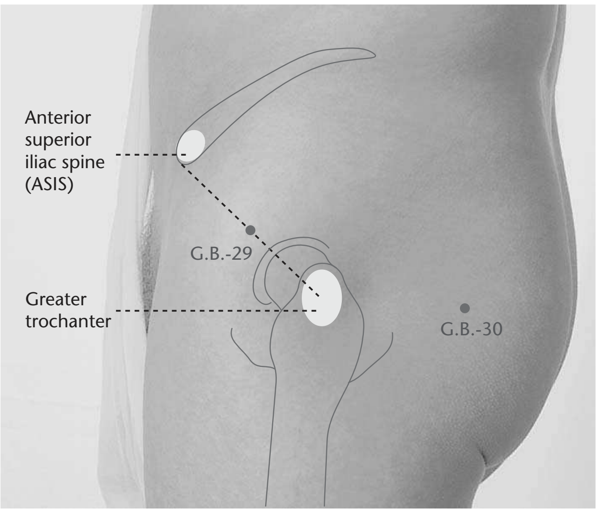
Anterior superior iliac spine (ASIS)
G.B.-29
Greater trochanter
G.B.-30

Fig. 3.99

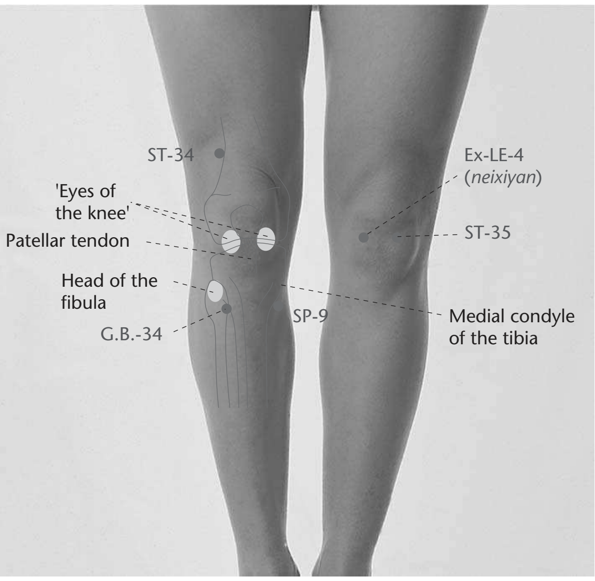
ST-34
'Eyes of the knee'
Patellar tendon
Head of the fibula
G.B.-34
SP-9
Ex-LE-4 (*neixiyan*)
ST-35
Medial condyle of the tibia

Fig. 3.100

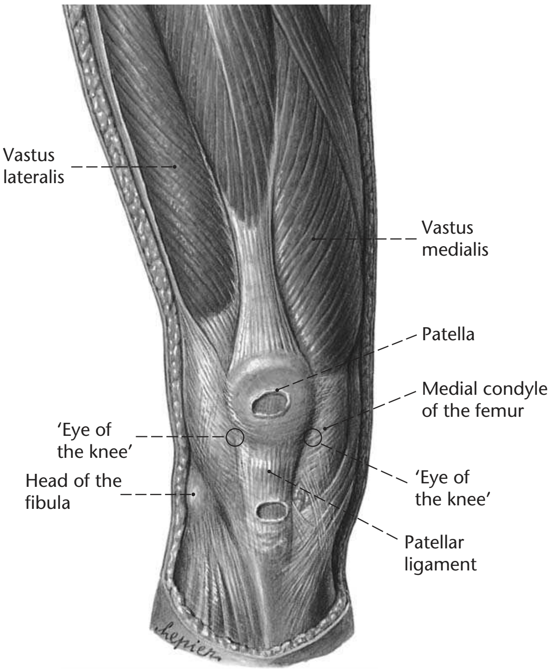
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

Fig. 3.101

3.6.1 Knee Region

For locating and needling points in the knee region, the patient should be in a relaxed supine position, with the knees slightly flexed and supported by pillows. This opens up the knee joint and allows for better palpation of the eyes of the knees. A prone position with a support below the feet and ankles is recommended for treating points in the popliteal crease.

Patella (→ Figs 3.100, 3.101)

The upper patellar border marks the location of → **SP-10** as well as of → **ST-32** to **ST-34**.

Eyes of the knee (→ Figs 3.100, 3.101)

The eyes of the knee are formed by the depressions on either side of the patellar tendon at the level of the lower patellar border. The lateral eye corresponds to the location of → **ST-35**, the medial eye is the location of the extra point → **Ex-LE-4** (*neixiyan*).

Head of the fibula (→ Figs 3.100, 3.101)

The head of the fibula forms a further important reference point inferior to the lateral aspect of the knee joint. For locating → **G.B.-34**, find the head of the fibula (approximately where the seam of a pair of trousers would be) and hold it with two fingers. By sliding downward, the anterior finger will 'drop' into the depression anterior and inferior to the head of the fibula, which marks the location of → **G.B.-34**.

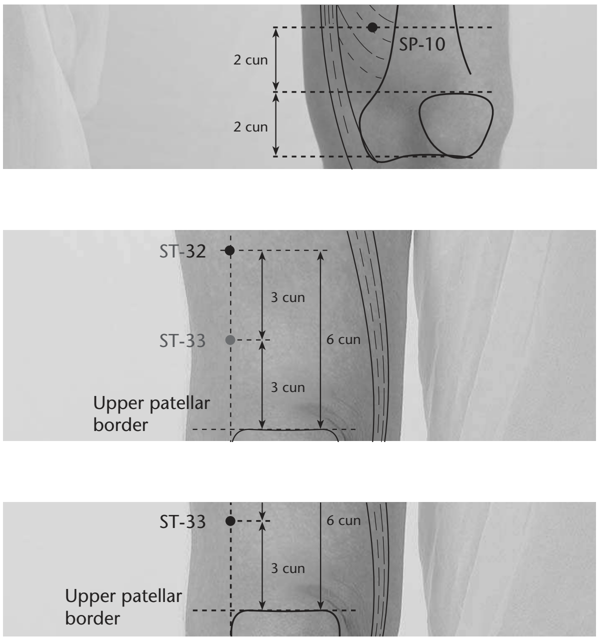
2 cun
SP-10
2 cun
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
ST-33
6 cun
3 cun
Upper patellar border

Fig. 3.102

Tibial shaft and medial condyle of the tibia (→ Fig. 3.103)

Below the medial aspect of the knee joint, the transition from the medial condyle of the tibia to the tibial shaft marks the location of → **SP-9**. This transition is best located by palpating in a superior direction along the posterior border of the medial aspect of the tibia. When using a knee roll, be aware that the soft tissue around this area may be pushed anteriorly, hindering palpation.

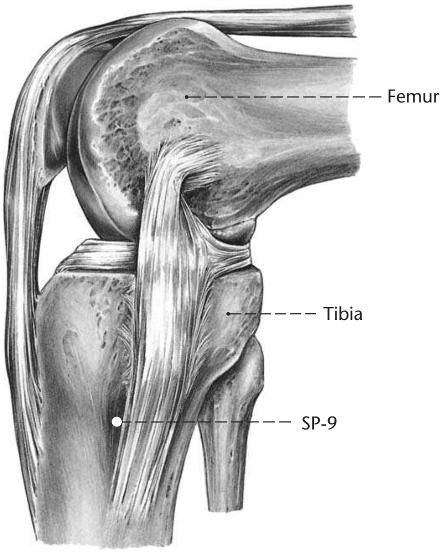
Femur
Tibia
SP-9

Fig. 3.103

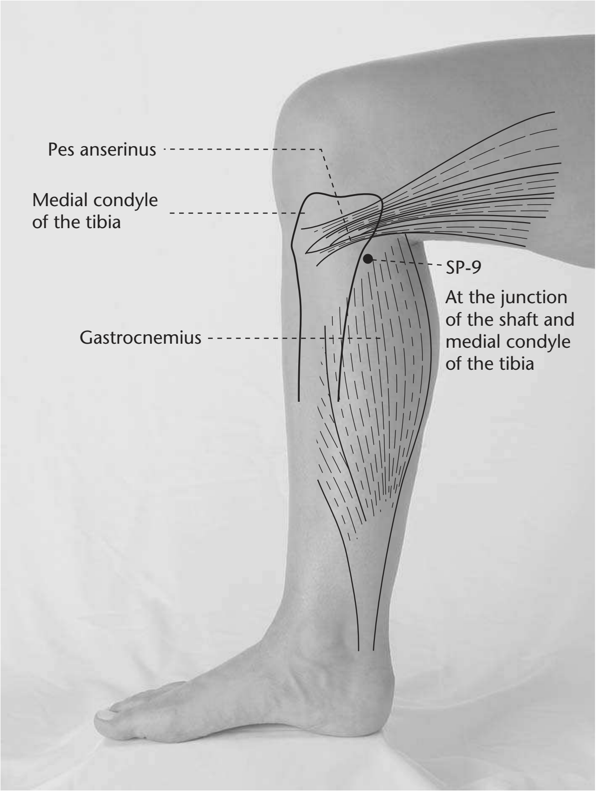
Pes anserinus
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

Fig. 3.104

3.6.2 Ankle and Foot

Highest prominence of the medial/lateral malleolus (→ Figs 3.105, 3.106)

Please note: The highest prominence of the malleolus does not refer to its lower border, but literally to its most protruding tip. The prominences are important reference points for the lower leg and ankle. → **Ex-LE-8** (*neihuajian*) is located on the highest prominence of the lateral malleolus, while → **Ex-LE-9** (*waihuajian*) is its medial counterpart.

Please also note: Contrary to some textbooks, the lateral and medial prominences are anatomically not located at the same level. There is approximately a 1 cun (→ Chapter 1, Fig. 1.2) difference in level. Thus, → **BL-60** and → **KID-3** are not located directly opposite each other, but rather occupy a comparable position.

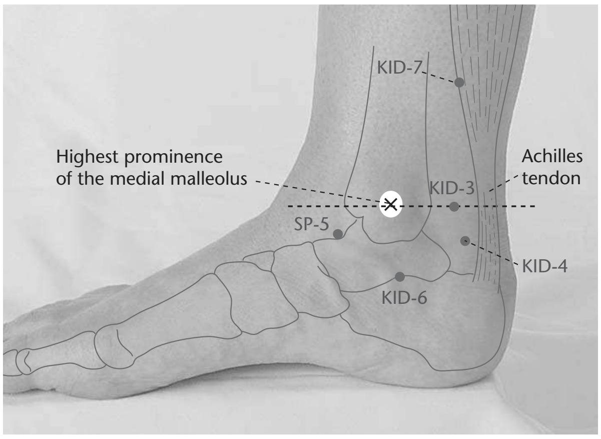
KID-7
Highest prominence of the medial malleolus
Achilles tendon
KID-3
SP-5
KID-4
KID-6

Fig. 3.105

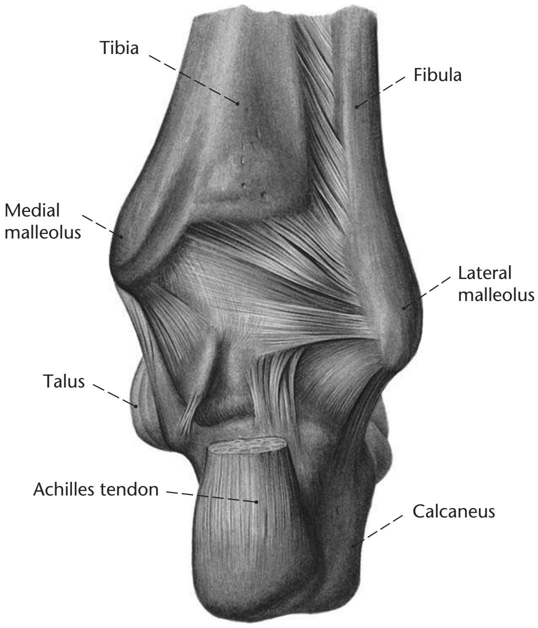
Tibia
Fibula
Medial malleolus
Lateral malleolus
Talus
Achilles tendon
Calcaneus

Fig. 3.106

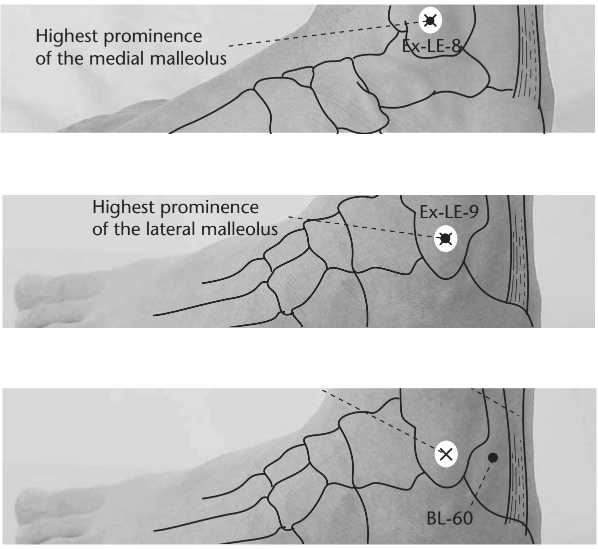
Highest prominence of the medial malleolus
Ex-LE-8
Highest prominence of the lateral malleolus
Ex-LE-9
BL-60

Fig. 3.107

Joint space on the anterior ankle (➞ Figs 3.108, 3.109)

The joint space on the anterior aspect of the ankle (between the malleoli) is best palpated by passively flexing and extending the foot. ➞ **LIV-4** and ➞ **ST-41** are both located here.

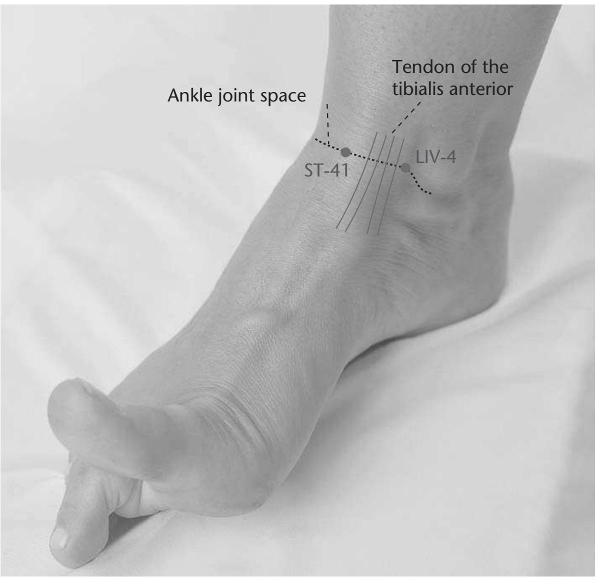
Ankle joint space
Tendon of the tibialis anterior
ST-41
LIV-4

Fig. 3.108

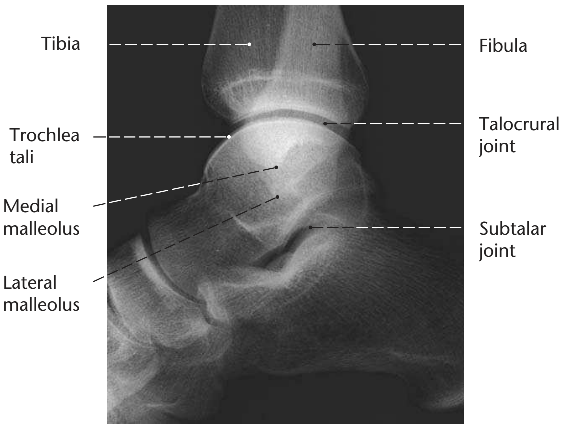
Tibia
Trochlea tali
Medial malleolus
Lateral malleolus
Fibula
Talocrural joint
Subtalar joint

Fig. 3.109

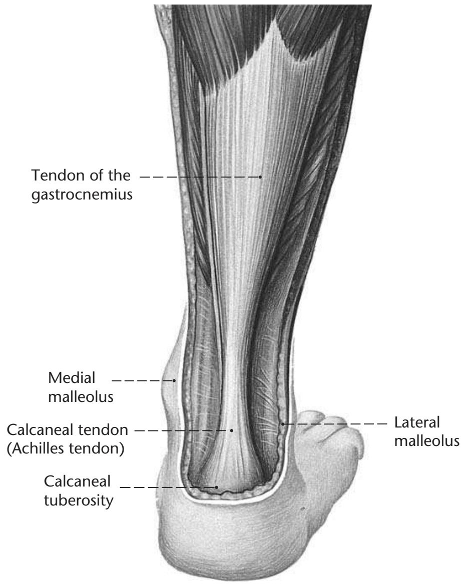
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

Fig. 3.110

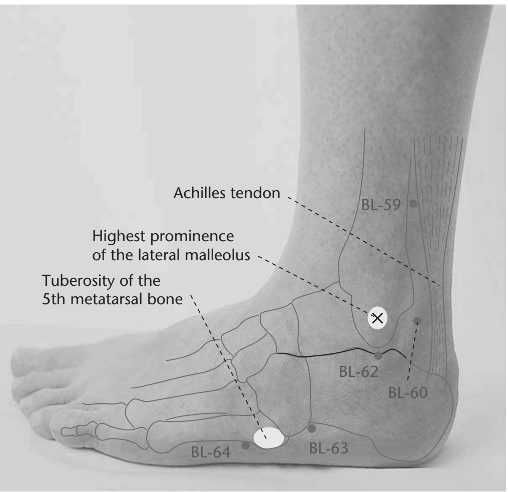
Achilles tendon
Highest prominence of the lateral malleolus
Tuberosity of the 5th metatarsal bone
BL-59
BL-62
BL-60
BL-64
BL-63

Fig. 3.111

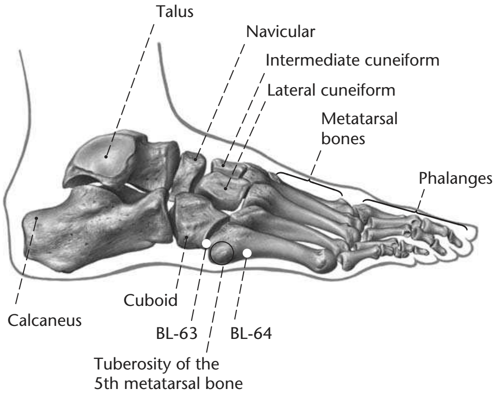
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Calcaneus
Cuboid
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Fig. 3.112

Achilles tendon (→ Fig. 3.110)

On the distal aspect of the lower leg and on the heel, the Achilles tendon defines the location of → **KID-3, KID-4**, and **KID-7** on its medial aspect and of → **BL-59** and **BL-60** on its lateral aspect.

Tuberosity of the 5th metatarsal bone (→ Figs 3.111, 3.112)

The tuberosity at the proximal end of the 5th metatarsal bone is the most palpable orientation point on the lateral aspect of the midfoot. → **BL-63** is located in the depression proximal to the tuberosity, while → **BL-64** is located in the depression distal to it, at the junction of the head and the shaft.

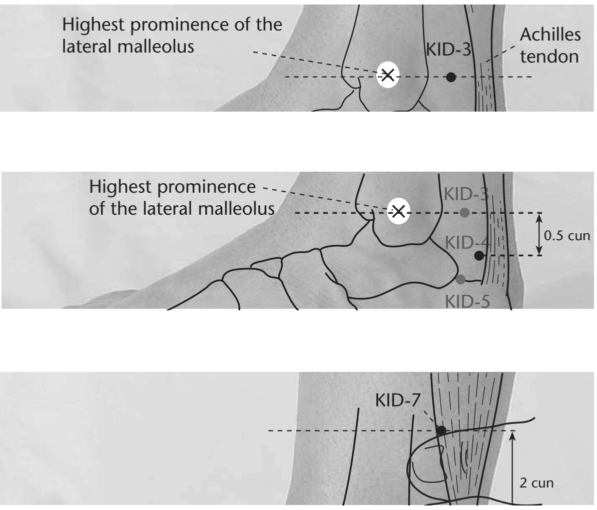
Highest prominence of the lateral malleolus
KID-3
Achilles tendon
Highest prominence of the lateral malleolus
KID-3
KID-4
KID-5
0.5 cun
KID-7
2 cun

Fig. 3.113

Base of the 1st metatarsal bone (→ Figs 3.114, 3.115)

When palpating along the medial aspect of the midfoot in a proximal direction, the proximal end (base) of the 1st metatarsal bone forms the first distinct prominence. → **SP-4** is located distal to this bony orientation point, at the junction of the shaft and the base.

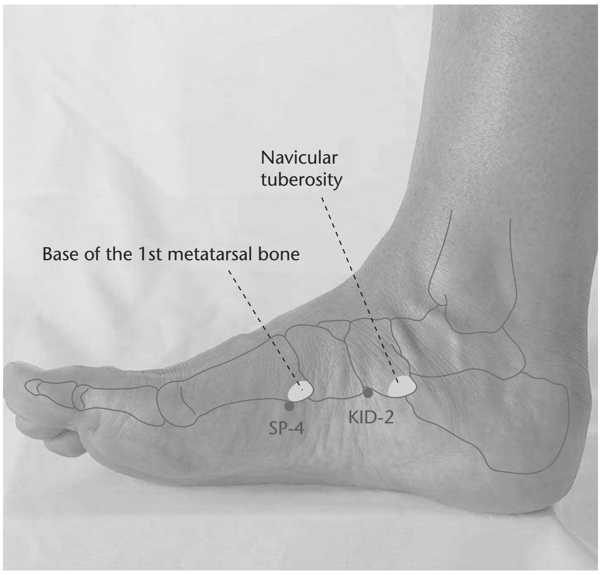
Navicular tuberosity
Base of the 1st metatarsal bone
SP-4
KID-2

Fig. 3.114

Navicular tuberosity (→ Fig. 3.115)

When palpating along the medial side of the foot, past → **SP-4** and the base of the 1st metatarsal bone, a further bony structure can be clearly felt. This is the navicular tuberosity, where → **KID-2** is located.

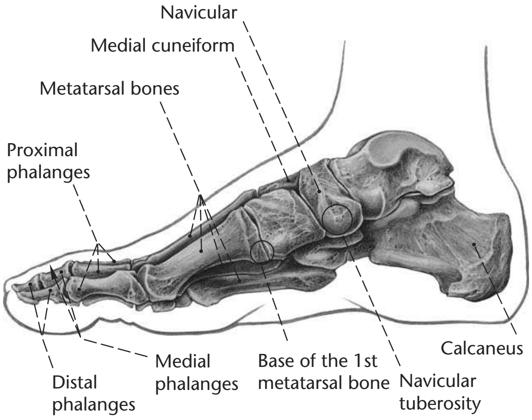
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Calcaneus
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity
Distal phalanges

Fig. 3.115

4 Acupuncture Points of the Twelve Primary Channels

Claudia Focks, Ulrich März

4.1 The Lung Channel System – Hand-*taiyin* (*shou tai yin jing luo*)

4.1.1 The Lung Primary Channel (*shou tai yin jing*)

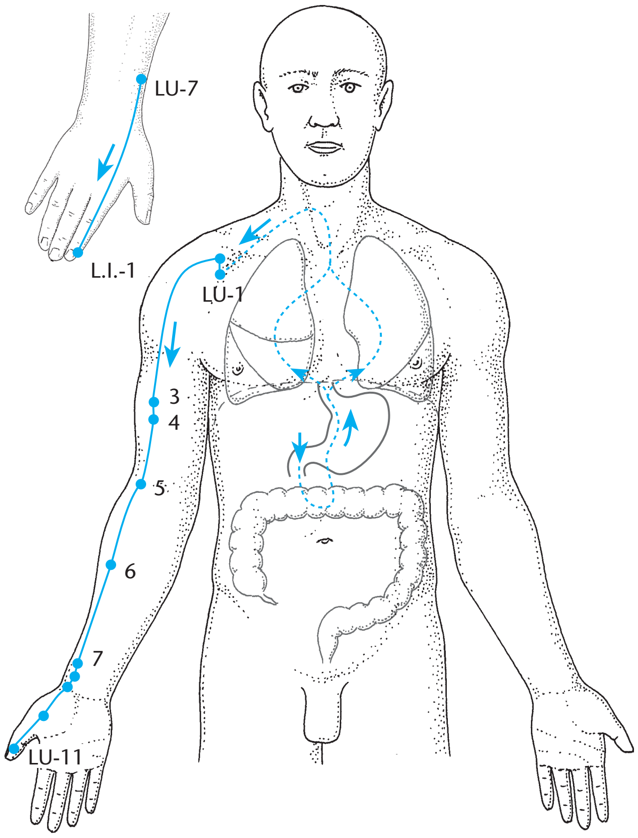
LU-7
L.I.-1
LU-1
3
4
5
6
7
LU-11

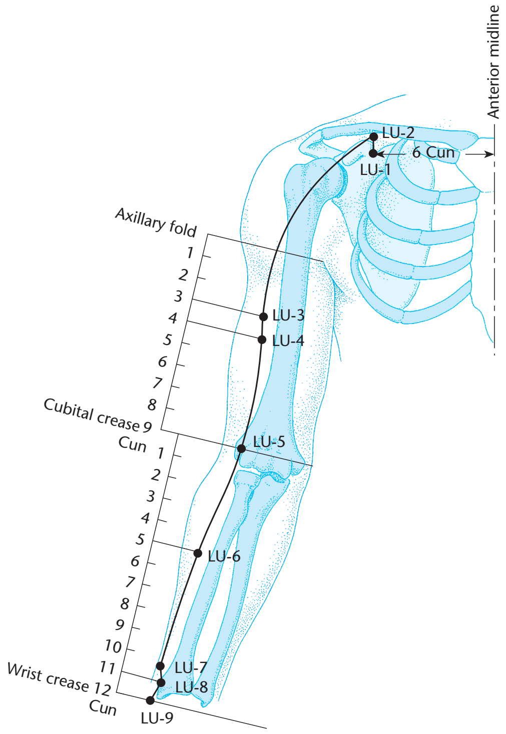
Anterior midline
LU-2
6 Cun
LU-1
Axillary fold
1
2
3
4
5
6
7
8
Cubital crease 9
Cun
LU-3
LU-4
LU-5
1
2
3
4
5
6
7
8
9
10
11
Wrist crease 12
Cun
LU-6
LU-7
LU-8
LU-9

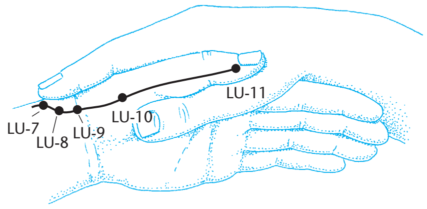
LU-11
LU-10
LU-7
LU-8
LU-9

Pathway

The internal section of the Lung primary channel originates in the Middle Burner in the region of the Stomach (*wei*)

- ➡ descends to connect with the Large Intestine (*dachang*), its paired *fu*-Organ
- ➡ curves upwards again to the Stomach (*wei*)
- ➡ penetrates the diaphragm
- ➡ enters the Lung (*fei*), its pertaining *zang*-Organ
- ➡ ascends to the throat region
- ➡ and curves towards the lateral thoracic region.

The Lung primary channel emerges at LU-1 (*zhongfu*) on the lateral thoracic wall at the level of the first intercostal space

- ➡ travels along the anterolateral aspect of the upper arm and forearm
- ➡ ends at LU-11 (*shaoshang*) on the thumb, at the radial aspect of the corner of the nail.

A branch separates from the primary channel at LU-7 (at the styloid process of the radius) and travels along the radial side of the index finger to its tip, where it connects with → **L.I.-1** (*shangyang*) (hand Yin–Yang connection of the first great circuit).

Ch04.1-F10028.qxd 2/23/08 7:02 PM Page 80

Clinical importance (→ 1.2)

Exterior (***biao***) signs and symptoms: Fever, aversion to cold, nasal blockage, headache, pain in the region of the thorax, shoulder and back, pain with cold sensations along the channel
Interior (***li***) or *zangfu*-Organ signs and symptoms: Cough, asthma, wheezing, dyspnoea, shortness of breath, fullness and tightness of the chest, expectoration of phlegm, dry throat, changed colour of the urine, irritability, heat sensation in the palms, abdominal fullness and distension, diarrhoea
Excess (***shi***): Shoulder and back pain, Wind-Cold attack with spontaneous sweating, frequent urination, yawning
Deficiency (***xu***): Shoulder and back pain with aversion to cold, shortness of breath, changed colour of urine

Connections of the Lung primary channel (→ 1.2)

Connections with other channels

Large Intestine primary channel (*shou yang ming jing*)

Connection: Hand Yin–Yang connection of the first great circuit
Location: LU-7 → L.I.-1 (on the hand). A branch of the Lung primary channel separates at **LU-7** (*lieque*) and connects with the Large Intestine primary channel at **L.I.-1** (*shangyang*), according to some authors also with → **L.I.-4** (*hegu*).
Circulation: Circadian (according to the Organ clock)
Importance: Interior–Exterior relationship

Spleen primary channel (*zu tai yin jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yin* (Yin axes of the first great circuit)
Location: SP-20 → LU-1 (on the thorax). A branch of the Spleen primary channel separates at → **SP-20** (*zhourong*) and connects with the Lung primary channel at → **LU-1** (*zhongfu*).
Circulation: Non-circadian (not according to the Organ clock)
Importance: Above–below relationship

Liver primary channel (*zu jue yin jing*)

Connection: Deep Yin–Yin connection
Location: LIV → LU (on the thorax). An internal branch of the Liver primary channel emerges at the Liver, passes the diaphragm and spreads inside the Lung (*fei*), forming a network with the Lung primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: Closes the circle of the first great circuit of the Nutritive Qi (*ying qi*, → 1.1.4)

Ren mai

Connection: *Ren mai* → Lung primary channel
Importance: Closes the circle of the second circuit of the Nutritive Qi (*ying qi*, → 1.1.4). The Lung primary channel as well as the great circuit (circulation through the 12 primary channels) receives Qi from the small circuit (*ren mai – du mai*).

Connections to the *zangfu*-Organ systems

Lung (*fei*), **Large Intestine (*dachang*)**, Stomach (*wei*)

4.1.2 The Lung Divergent Channel (*shou tai yin jing bie*)

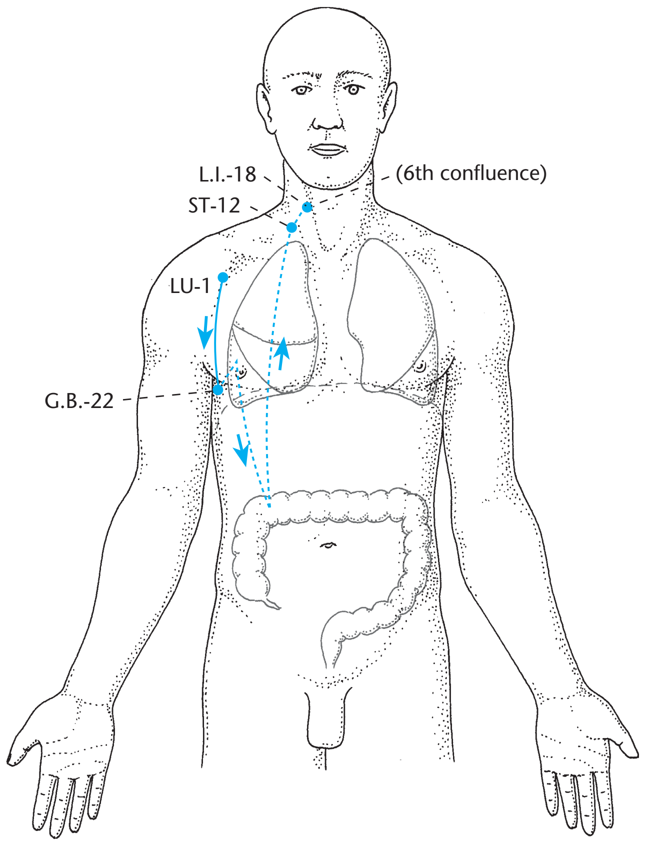

Pathway

The Lung divergent channel separates from the Lung primary channel in the axillary region

➡ travels anteriorly to → **G.B.-22** (*yuanye*) on the midaxillary line, 3 cun inferior to the axilla
➡ enters the thoracic region
➡ disperses in the Lung (*fei*), its pertaining *zang*-Organ
➡ descends to the Large Intestine (*dachang*), its paired *fu*-Organ
➡ ascends again and emerges in the region of the supraclavicular fossa at → **ST-12** (*quepen*)
➡ traverses the neck in a cranial direction and connects with the Large Intestine primary channel at → **L.I.-18** (*futu*) at one of the six *he*-confluences* (L.I./LU as 6th confluence. → 1.3).

There exist variations regarding the channel pathway due to differing interpretations of the *Ling Shu* (→ Solinas et al. 1998).

---

* According to some authors the 6th confluence is localised at → **ST-12** (*quepen*).

Clinical importance

- Strengthens the connection between the Lung and the Large Intestine (*zangfu*-Organ systems). Points on the Large Intestine primary channel can therefore be used for disorders of the Lung, and vice versa points on the Lung primary channel can treat disorders of the Large Intestine.
- Creates a network between the Lung primary channel and the neck region at the confluent point → **L.I.-18**: thus LU points such as → **LU-10** or → **LU-11** may be used to treat the neck region.

4.1.3 The Lung Sinew Channel (*shou tai yin jing jin*)

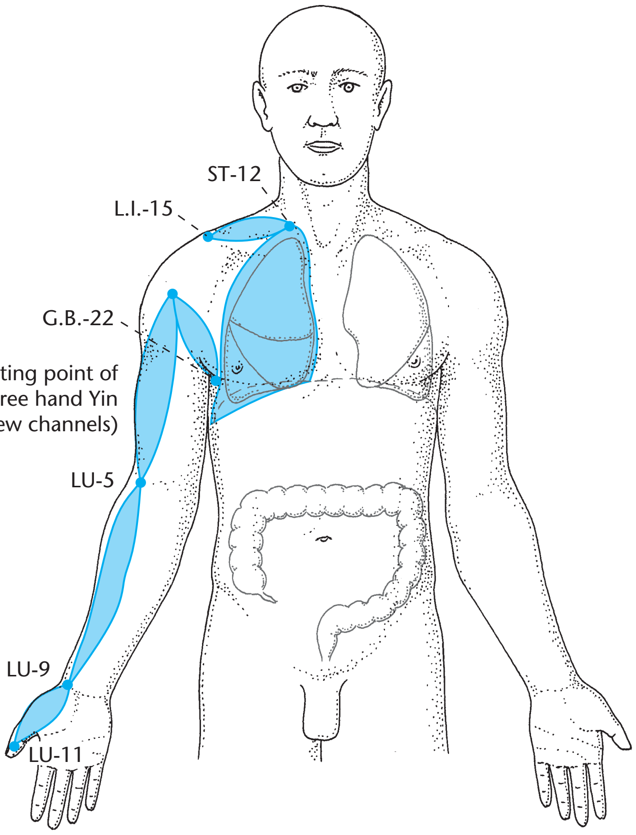

Pathway

The Lung sinew channel

- begins on the thumb at **LU-11** (*shaoshang*), at the radial aspect of the corner of the nail
- spreads along the radial aspect of the thumb and the 1st metacarpal bone as well as the thenar eminence
- meets **LU-9** (*taiyuan*) and binds (*jie*) at the radial wrist joint space
- travels proximally along the anterolateral aspect of the forearm
- reaches the elbow at **LU-5** (*chize*) and binds (*jie*) at the biceps tendon
- follows the biceps brachii muscle and the lateral portion of the deltoid muscle
- binds (*jie*) at the anterior shoulder region near the acromioclavicular joint
- continues to the axilla and meets the other hand Yin sinew channels (P, HE) at → **G.B.-22** (*yuanye*) and binds (*jie*) there
- continues deep to the pectoralis major muscle and enters the supraclavicular fossa at → **ST-12** (*quepen*)
- courses to → **L.I.-15** (*jianyu*) and returns to → **ST-12** (*quepen*)
- penetrates the supraclavicular fossa and disperses in the thoracic and hypochondriac region and the diaphragm (→ Fig.)

Clinical importance

Pathology: Stiffness, cramps and pain along the sinew channel. *Xi fen*: spasms and pain in the thorax and hypochondriac region. In severe cases formation of nodules below the right lateral costal region and distension along the lateral costal region.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the LU channel.

The area covered by the Lung sinew channel is larger than that covered by the Lung primary channel. This explains why the indications of points on the Lung primary channel include disorders and diseases of the thoracic and hypochondriac region.

4.1.4 The Lung *luo*-Connecting Vessel System (*shou tai yin luo mai*)

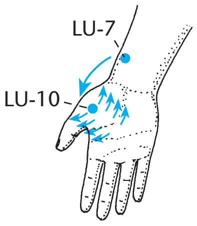

Pathway

The Lung *luo*-connecting channel system separates from the Lung primary channel at **LU-7** (*lieque*) (→ 8.1.2) and forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions course to the Interiorly–Exteriorly paired Large Intestine primary channel; according to some schools (for example Van Nghi, → Appendix) they travel as a transverse Lung *luo* vessel to the *yuan*-source point → **L.I.-4** (*hegu*).
- A longitudinally orientated division spreads on the palm and thenar eminence.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*)**: Hot and burning sensations in the palms and the wrist joints
- **Deficiency (*xu*)**: Yawning, shortness of breath, frequent urination, enuresis

4.1.5 Cutaneous Region (*tai yin pi bu*)

See description and figures → 1.6.

4.1.6 Points of the Lung Primary Channel (Overview)

Specific points according to function

- ***Yuan*-source point (→ 8.1.1): LU-9** (*taiyuan*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): LU-7** (*lieque*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): LU-6** (*kongzui*)
- **Pertaining Back-*shu* point (→ 8.1.4): BL-13** (*feishu*) 🟦⬛
- **Pertaining Front-*mu* point (→ 8.1.5): LU-1** (*zhongfu*) 🟦⬛
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **LU-11** (*shaoshang*) 🟦⬛
  ying-spring point (Fire): **LU-10** (*yuji*)
  shu-stream point (Earth), tonification point: **LU-9** (*taiyuan*) 🟦🟦
  jing-river point (Metal), *ben* point: **LU-8** (*jingqu*) 🟦⬛
  he-sea point (Water), sedation point: **LU-5** (*chize*) 🟦⬛
- **Influential *hui*-meeting point (→ 8.1.7) of the vessels: LU-9** (*taiyuan*) 🟦🟦
- **Opening point (→ 8.1.8) of the *ren mai*: LU-7** (*lieque*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the SP primary channel: **LU-1** (*zhongfu*) 🟦⬛
  – of other channels with the LU channel: –
- **Gao Wu Command point (→ 8.1.11) for the occiput and nape: LU-7** (*lieque*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): LU-3** (*tianfu*)
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): LU-7** (*lieque*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): LU-11** (*shaoshang*) 🟦⬛

Points according to region

- **Local points (→ 8.2.1):** Lung – **LU-1** (*zhongfu*) 🟦⬛; elbow – **LU-5** (*chize*) 🟦⬛; wrist – **LU-7** (*lieque*) 🟦🟦
- **Adjacent points (→ 8.2.1):** hands – **LU-7** (*lieque*) 🟦🟦
- **Distal points (→ 8.2.1):** occiput – **LU-7** (*lieque* 🟦🟦); throat – **LU-10** (*yuji*) – **LU-11** (*shaoshang*); Lung – **LU-5** (*chize*) 🟦⬛, **LU-7** (*lieque*) 🟦🟦; shoulder – **LU-7** (*lieque*) 🟦🟦

Specific points according to the course of the channel (in numeric order)

- **LU-1** (*zhongfu* 🟦⬛): associated Front-*mu* point (→ 8.1.5), *jiaohui*-meeting point with the SP channel (→ 8.1.10)
- **LU-3** (*tianfu*): Window of Heaven point (→ 8.1.2)
- **LU-5** (*chize*) 🟦⬛: *he*-sea point (Water) (→ 8.1.6); sedation point; distal point of the Lung (→ 8.2.1)
- **LU-6** (*kongzui*): *xi*-cleft point (→ 8.13)
- **LU-7** (*lieque*) 🟦🟦: *luo*-connecting point (→ 8.1.2); opening point (→ 8.1.8) of the *ren mai*; Gao Wu Command point (→ 8.1.11) of the nape and occiput; Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the occiput, shoulder and Lungs (→ 8.2.1); adjacent point for the hands (→ 8.2.1)
- **LU-8** (*jingqu*): *jing*-river point (Metal) (→ 8.1.6); *ben* point (Five Phase point)
- **LU-9** (*taiyuan*) 🟦🟦: *yuan*-source point (→ 8.1.1), hui-meeting point (→ 8.1.7) of the vessels; *shu*-stream point (Earth) (→ 8.1.6); tonification point
- **LU-10** (*yuji*): *ying*-spring point (Fire) (→ 8.1.6); distal point for the neck (→ 8.2.1)
- **LU-11** (*shaoshang*) 🟦⬛: Sun Si Miao Ghost point (→ 8.1.15); *jing*-well point (Wood) (→ 8.1.6); distal point for the neck

Central Residence ZHONGFU **LU-1**

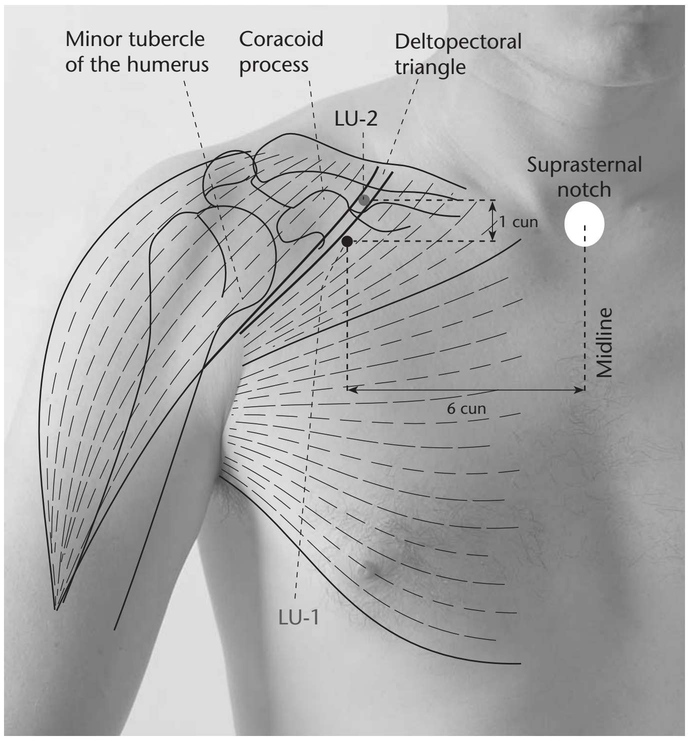
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Suprasternal notch
1 cun
Midline
6 cun
LU-1

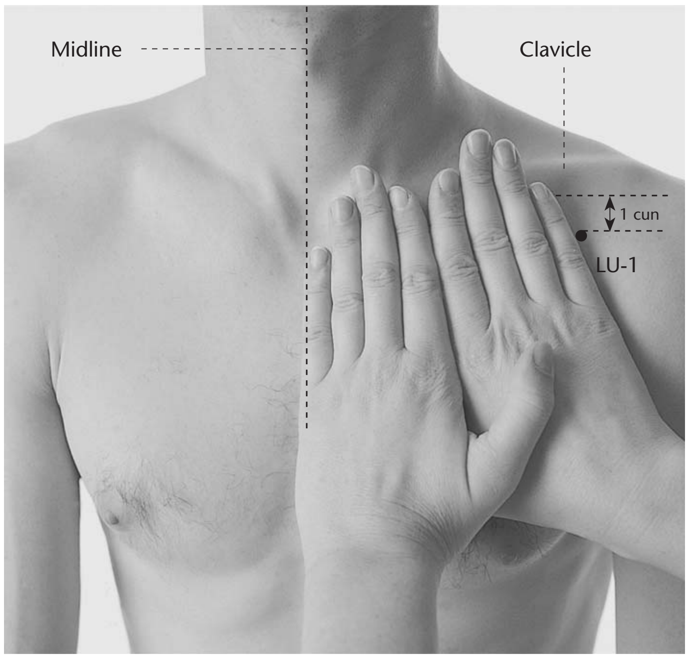
Midline
Clavicle
1 cun
LU-1

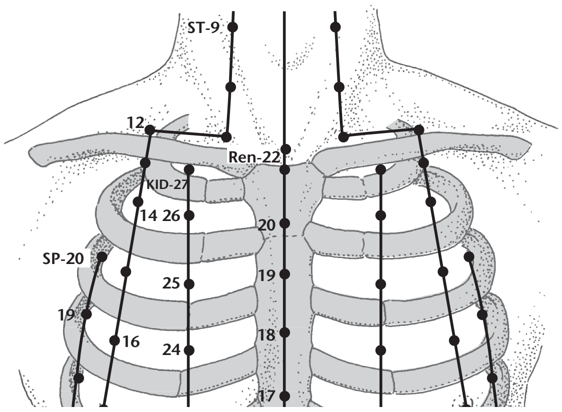
ST-9
12
Ren-22
KID-27
14 26
20
SP-20
25
19
19
16
24
18
17

Location

6 cun lateral of the anterior midline and approximately 1 cun below → **LU-2,** slightly medial to the lower border of the coracoid process.

How to find

First, locate → **LU-2** in the deltopectoral triangle, which is bordered superiorly by the clavicle and laterally by the coracoid process (within the deltoid muscle). From there, palpate along the border of the deltoid for 1 cun and there locate **LU-1** slightly medial to the lower border of the deltoid on the thoracic wall, approximately 6 cun lateral to the midline. For orientation: when externally rotating the arm with the elbow flexed at the same time, the coracoid process will remain static, while the minor tuberculum of the humerus will follow the movement. → **SP-20** is also located 6 cun lateral to the midline on the level of the 2nd intercostal space. Located approximately on the same level are → **Ren-20,** → **KID-26** and → **ST-14**.

Needling

0.5–1 cun obliquely in a cranial (approximately 45°) and lateral direction towards the coracoid process.

Actions/Indications

- Regulates and descends the Lung Qi
- Clears Heat in the Upper Burner, transforms Phlegm
- Regulates the water passages
- Moves Qi locally and in the sinew channel

Special features

Front-*mu* point of the Lung, meeting point with the Spleen channel, entry point. Important point for disorders of the Lung.

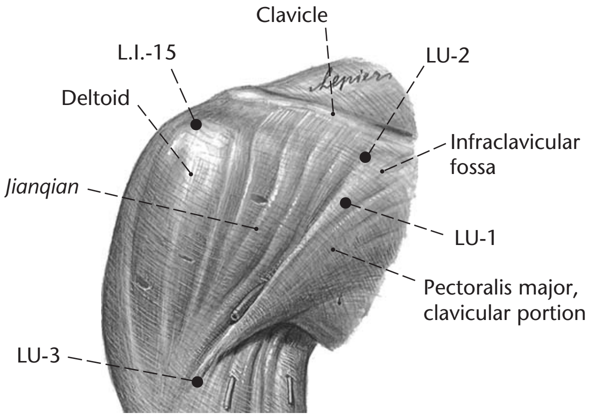
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Jianqian
LU-1
Pectoralis major, clavicular portion
LU-3

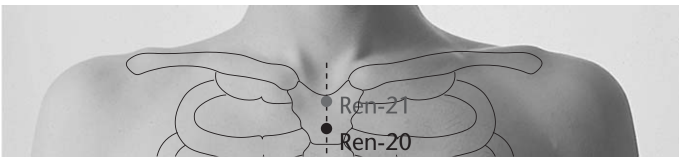
Clavicle
Ren-21
Ren-20

LU-2 *Cloud Gate* YUNMEN

Location

6 cun lateral to the anterior midline, below the clavicle, in the centre of the deltopectoral triangle.

How to find

Find the angle formed by the lateral lower border of the clavicle and the coracoid process (within the deltoid muscle). Locate **LU-2** at the junction of the clavicular insertions of the deltoid and pectoralis muscles (deltopectoral triangle), in the centre of a clearly palpable depression in the thoracic wall, 6 cun lateral to the anterior midline.

→ **ST-13** is also located at the lower border of the clavicle, 4 cun lateral to the anterior midline or at the midpoint of the clavicle. Also located on approximately the same level are → **KID-27** (2 cun lateral to the midline) and → **Ren-21** (on the midline). → **LU-1** is located 1 cun below and in most cases slightly lateral to **LU-2**.

Needling

Obliquely to a maximal depth of 0.8 cun. No needling in a medial direction! Risk of pneumothorax!

Actions/Indications

- Clears Lung Heat, decends the Lung Qi
- Opens the channel and sinew channel
- Clears Heat from the extremities

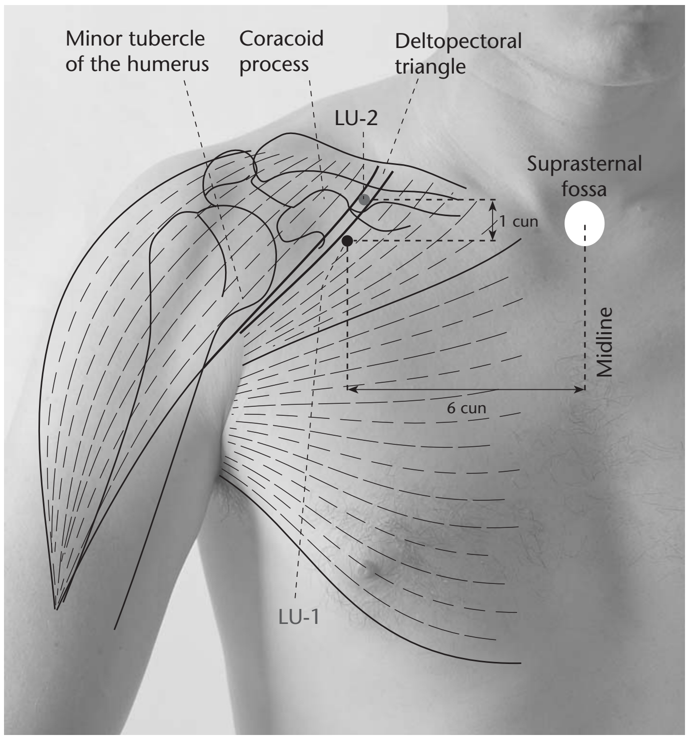
Minor tubercle of the humerus
Coracoid process
Deltopectoral triangle
LU-2
Suprasternal fossa
1 cun
Midline
6 cun
LU-1

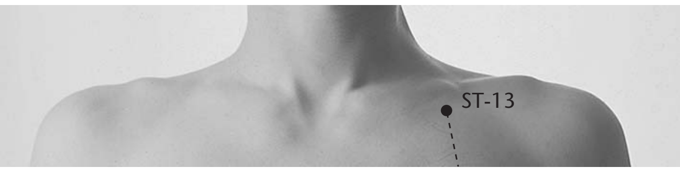

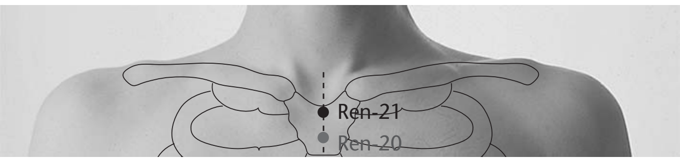
Ren-21
Ren-20

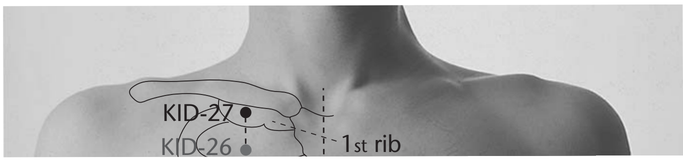
KID-27
KID-26
1st rib

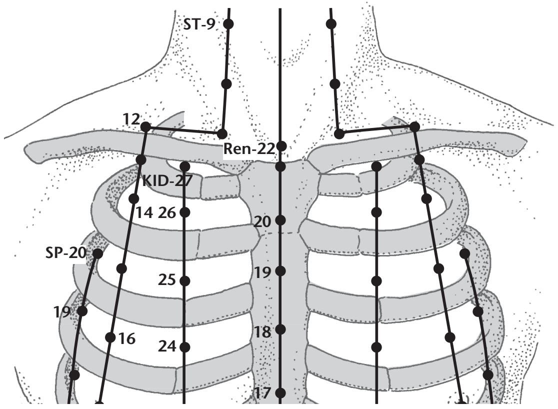
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
19
25
19
16
24
18
17

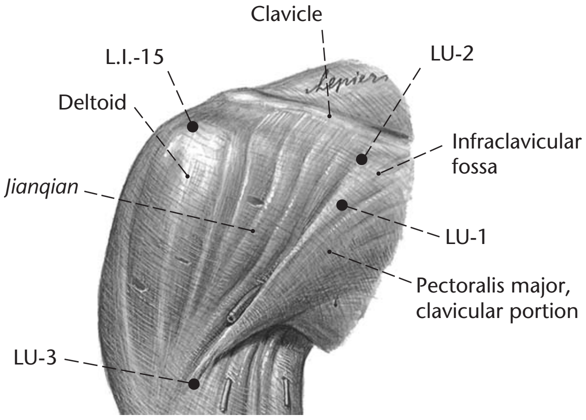
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Jianqian
LU-1
Pectoralis major, clavicular portion
LU-3

Heavenly Residence TIANFU LU-3

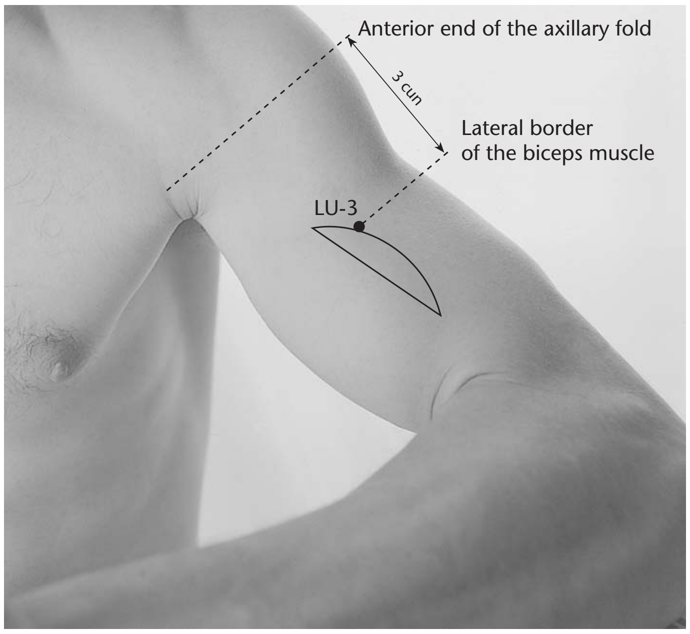
Anterior end of the axillary fold
3 cun
Lateral border of the biceps muscle
LU-3

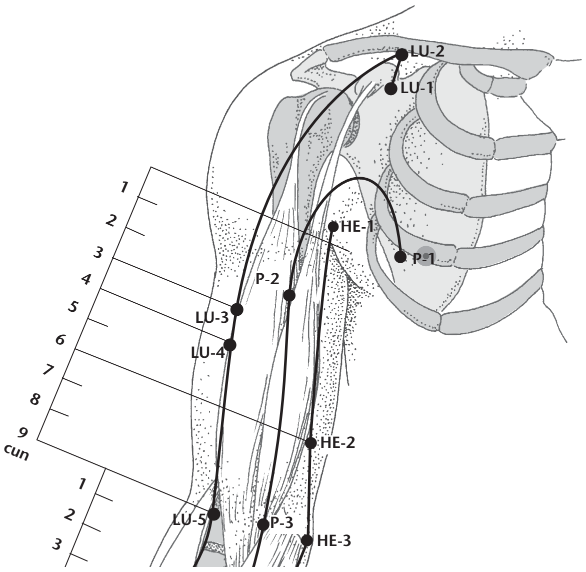
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3

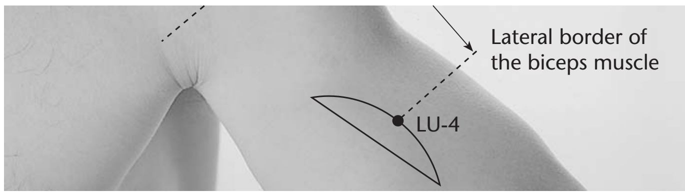
Lateral border of the biceps muscle
LU-4

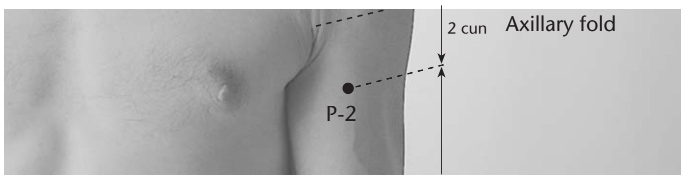
2 cun
Axillary fold
P-2

Location

On the medial aspect of the upper arm, 3 cun distal to the end of the anterior axillary fold, in the groove on the lateral border of biceps brachii.

How to find

Ask the patient to flex the biceps muscle against resistance. Locate **LU-3** 3 cun below the anterior end of the axillary fold (→ 2.2) in a groove on the lateral border of biceps brachii. Sometimes, you can feel the pulse of the brachial artery here.
Or: Divide the distance of 9 cun (→ 2.2) between the end of the axillary fold and the cubital crease (→ **LU-5**) into thirds and locate **LU-3** a third of the distance from the axillary fold.
→ **LU-4** is located in the sulcus, 1 cun distal to **LU-3**. → **P-2** is located 2 cun below the axillary fold, between the two heads of the biceps brachii muscle.

Needling

Vertically 0.5–1 cun. Moxibustion is contraindicated according to some classics.

Actions/Indications

- Regulates and descends the Lung Qi
- Clears Lung Heat
- Cools the Blood and stops bleeding
- Calms the *po* (corporeal soul)
- Opens the channel and alleviates pain

Special features

Window of Heaven point

LU-4 *Clasping the White* XIABAI

Location

On the medial aspect of the upper arm, 4 cun distal to the end of the anterior axillary fold, in the depression on the lateral border of sulcus bicipitalis lateralis.

How to find

Ask the patient to flex their biceps muscle against resistance. Locate **LU-4** 4 cun below the anterior end of the axillary fold (➞ 2.2) in the depression on the lateral border of the sulcus bicipitalis muscle. Sometimes you can feel the pulse of the brachial artery here. **Or quick method:** Spreading hands technique (➞ 2.3.3): place the little fingers on the axillary fold and the cubital crease (at ➞ **LU-5**) respectively (this distance corresponds to 9 proportional cun ➞ 2.2) and determine the midpoint of this distance by joining the thumbs. Locate **LU-4** 0.5 proximal to the midpoint in the sulcus on the lateral border of the biceps muscle. ➞ **LU-3** is also located in the sulcus, 1 cun proximal to **LU-4**.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates and moves Qi and Blood in the chest
- Descends the Lung Qi
- Opens the channel and alleviates pain

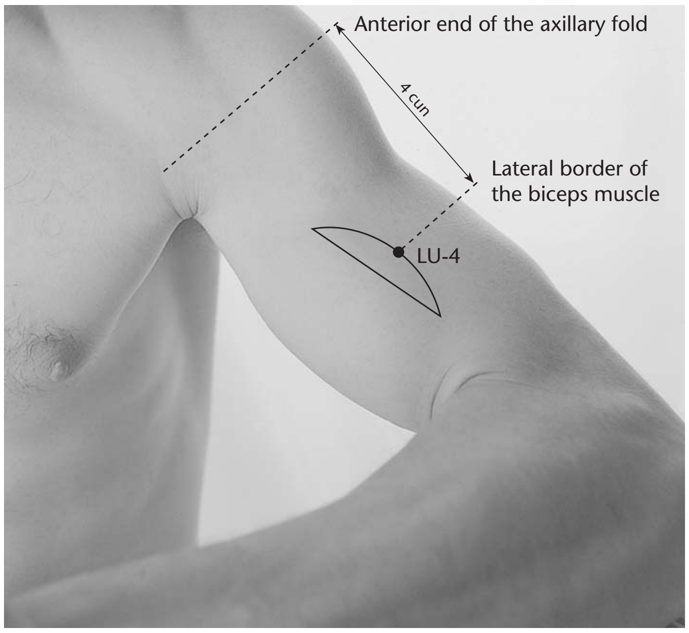
Anterior end of the axillary fold
4 cun
Lateral border of the biceps muscle
LU-4

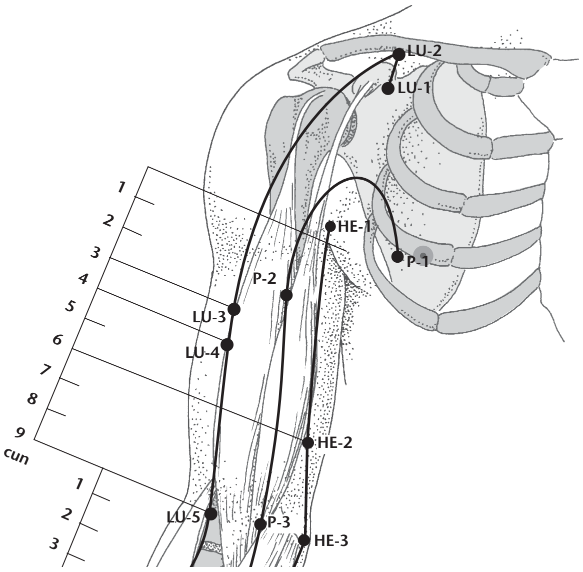
LU-2
LU-1
1
2
3
4
5
6
7
8
9
cun
HE-1
P-1
P-2
LU-3
LU-4
HE-2
1
2
3
LU-5
P-3
HE-3

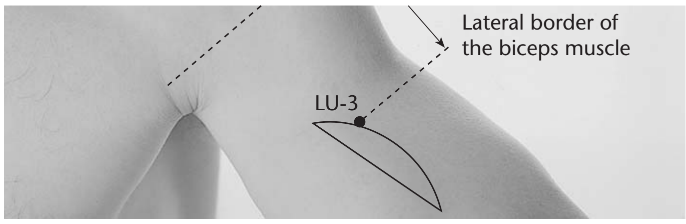
Lateral border of the biceps muscle
LU-3

Cubit Marsh CHIZE LU-5

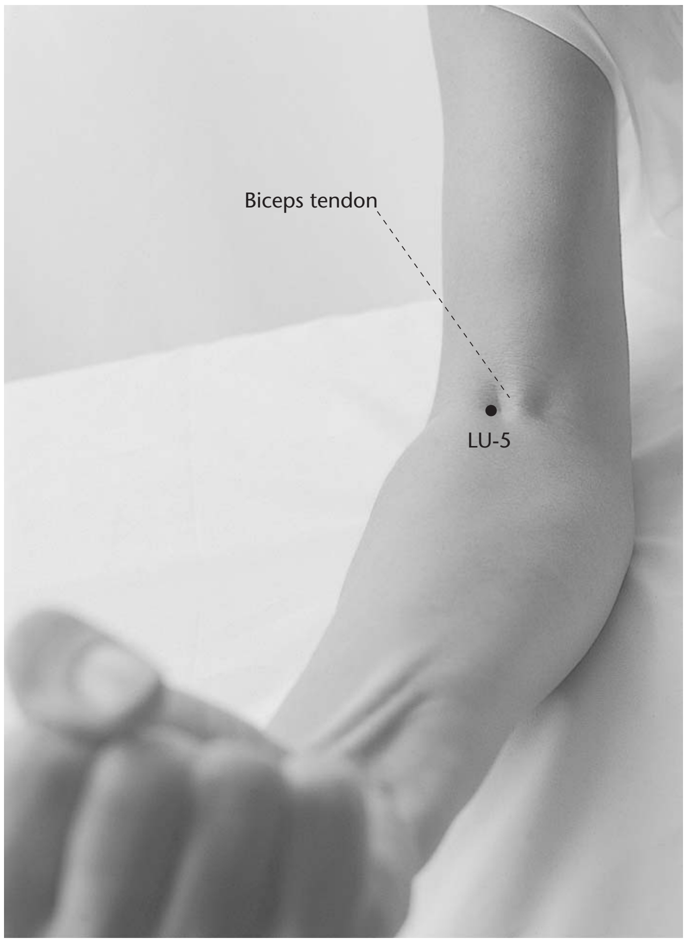
Biceps tendon
LU-5

Right arm

Location

In the cubital crease, on the radial aspect of the tendon of the biceps.

How to find

By slightly flexing the forearm, the biceps tendon becomes more visible. **LU-5** is located on its radial aspect in the cubital crease. Also located in the area of the cubital crease are → **P-3** (ulnar to the tendon), → **HE-3** (at the ulnar end of the cubital crease) and → **L.I.-11** (between the radial end of the cubital crease and the lateral epicondyle).

Needling

Vertically 0.5–1 cun. Needle with the arm slightly flexed. Use bloodletting for Heat disorders. Caution: Avoid cubital vein.

Actions/Indications

- Clears Heat in the Upper Burner
- Descends the Lung Qi
- Regulates the water passages
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

He-sea point, Water point, sedation point, distal point for the Lung, local point for the elbow. Use bloodletting for disorders of the skin and Lung due to excess and/or Heat.

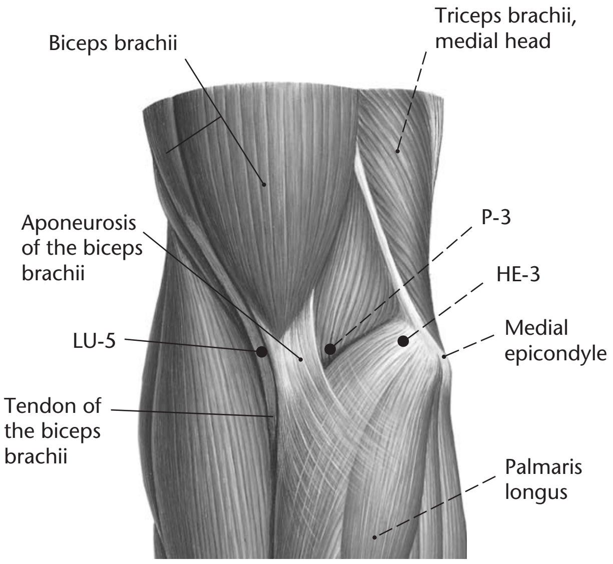
Biceps brachii
Triceps brachii, medial head
Aponeurosis of the biceps brachii
P-3
HE-3
Medial epicondyle
LU-5
Tendon of the biceps brachii
Palmaris longus

Left arm, anterior aspect

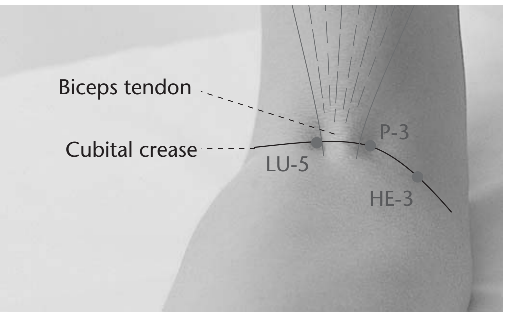
Biceps tendon
Cubital crease
P-3
LU-5
HE-3

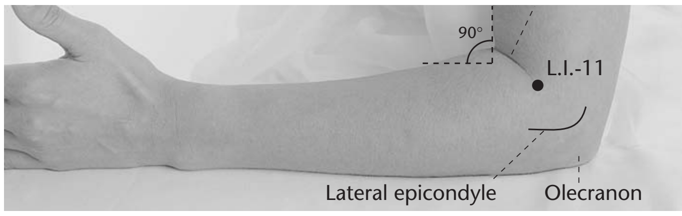
90°
L.I.-11
Lateral epicondyle
Olecranon

LU-6 *Biggest Hole* KONGZUI

Location

On the line connecting → **LU-5** (in the cubital crease) and → **LU-9** (on the wrist crease/joint space), 5 cun distal to → **LU-5** or 7 cun proximal to → **LU-9**.

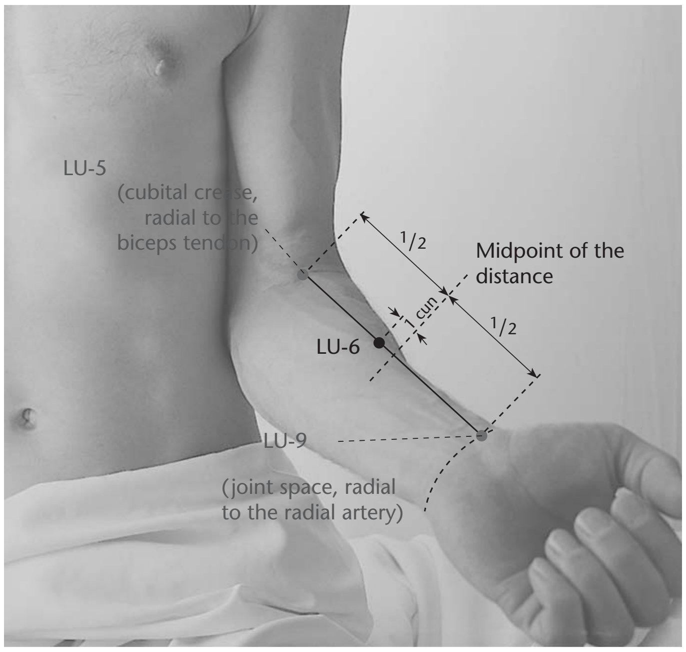
LU-5
(cubital crease,
radial to the
biceps tendon)
1/2
Midpoint of the
distance
1 cun
1/2
LU-6
LU-9
(joint space, radial
to the radial artery)

How to find

The biceps tendon becomes more visible with the arm flexed. First, locate → **LU-5** radial to the biceps tendon in the cubital crease and → **LU-9** radial to the radial artery on the wrist joint space (→ 3.3.3) between the proximal row of carpals and the radius/ulna. By moving the hand, the joint space becomes more easily palpable. Next, find the midpoint of the distance between → **LU-5** and → **LU-9** using the spreading hands technique (→ 2.3.3). **LU-6** is located 1 cun proximal to the midpoint of this distance.

→ **P-4** is located more medially (between the tendons) and distally (1 cun distal to the midpoint of the distance).

Needling

Vertically 0.5–1 cun

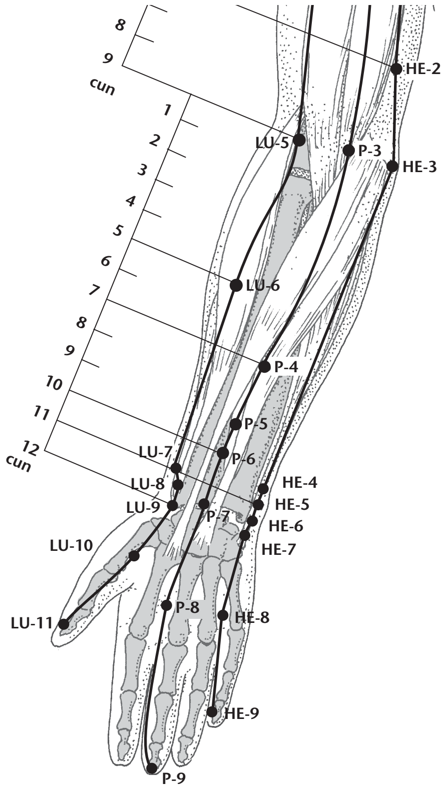
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun
HE-2
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
HE-4
LU-9
P-7
HE-5
HE-6
LU-10
HE-7
P-8
HE-8
LU-11
HE-9
P-9

Actions/Indications

- Descends the Lung Qi
- Clears Lung Heat
- Cools Blood Heat, stops bleeding
- Expels Exterior pathogenic factors
- Opens the channel

Special features

Xi-cleft point

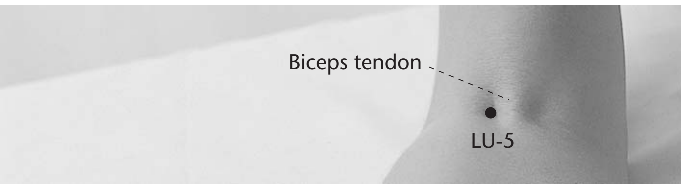
Biceps tendon
LU-5

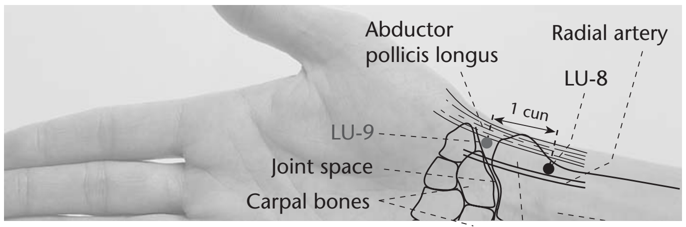
Abductor
pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

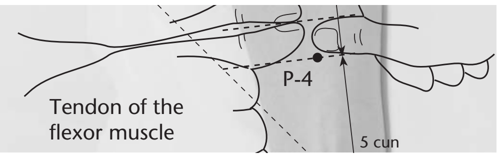
P-4
Tendon of the
flexor muscle
5 cun

Broken Sequence LIEQUE LU-7

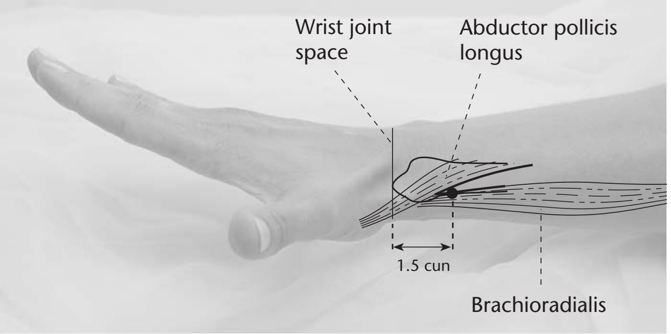
Wrist joint space
Abductor pollicis longus
1.5 cun
Brachioradialis

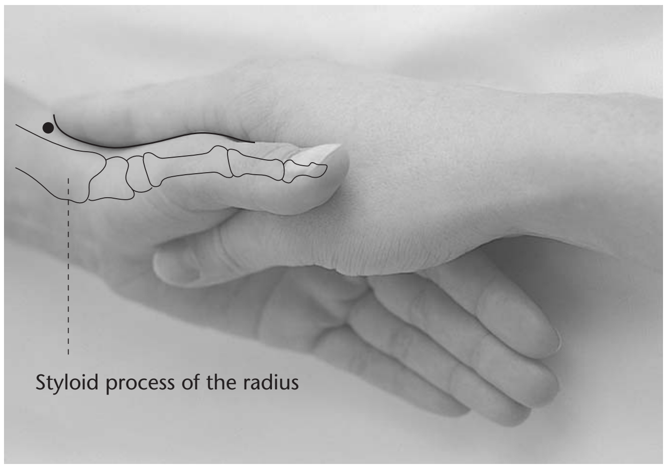
Styloid process of the radius

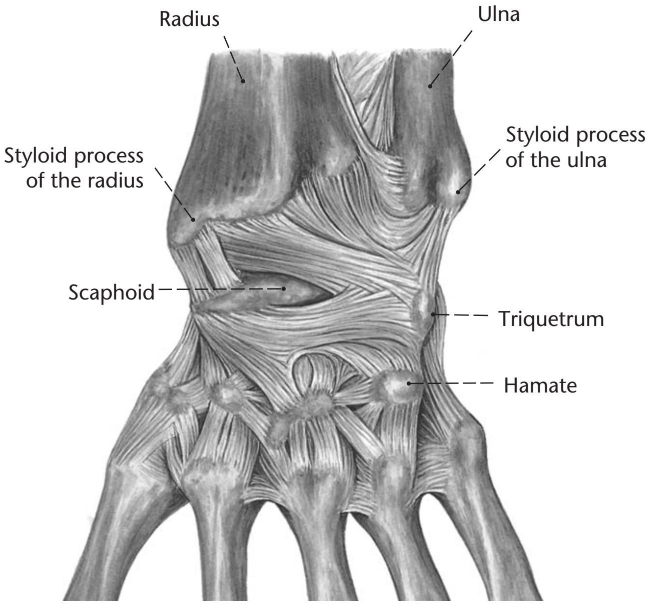
Radius
Ulna
Styloid process of the radius
Styloid process of the ulna
Scaphoid
Triquetrum
Hamate

Left hand, dorsal aspect

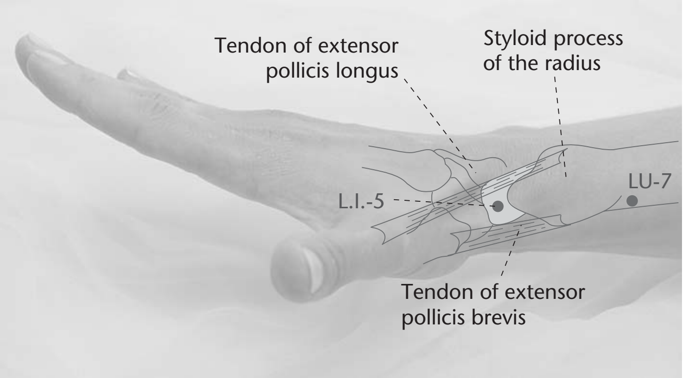
Tendon of extensor pollicis longus
Styloid process of the radius
L.I.-5
LU-7
Tendon of extensor pollicis brevis

Location

On the radial aspect of the forearm, directly above the styloid process of the radius, approximately 1.5 cun proximal to the wrist joint space (wrist crease) in a V-shaped groove.

How to find

Place the palpating forefinger on the anatomical snuffbox (→ **L.I.-5**). From there, glide proximally over the styloid process of the radius and locate **LU-7** where the finger can palpate a cleft between two tendons (brachioradialis/abductor pollicis longus). **Or:** 'Tiger mouth grip' (→ Fig. 2.6): spread the thumb and forefinger of both hands, crossing them so that the forefinger of one hand comes to rest on the styloid process of the other, but avoiding bending the wrist. **LU-7** is located directly below the tip of the forefinger in a V-shaped groove

Needling

Lift the skin above the styloid process by pinching it. Insert the needle 0.5–1 cun obliquely and proximally (reducing) or distally (tonifying). Caution: Cephalic vein.

Actions/Indications

- Supports the occiput and head
- Releases the Exterior, expels Wind, descends the Lung Qi
- Opens and regulates the *ren mai*
- Regulates the water passages
- Opens the channel and *luo*-connecting channel, alleviates pain

Special features

Luo-connecting point, opening point of the *ren mai*, Ma Dan Yang Heavenly Star point, Gao Wu command point of the occiput, exit point.

LU-8 *Channel Gutter* JINGQU

Location

Lateral to the radial artery, 1 cun proximal to the ventral wrist joint space (most distal wrist crease).

How to find

As the location of the wrist crease varies, use the joint space (→ 3.3.3) between the proximal row of carpals and the radius/ulna for orientation. By loosely moving the hand, the joint space is easily palpable. This is the level for palpating the radial pulse (using the finger tips, not the finger pads). → **LU-9** is located on the radial side of the artery, while **LU-8** is located 1 cun proximal to → **LU-9**.
→ **L.I.-5** is located close to **LU-8**, but on the dorsal aspect of the wrist joint in the anatomical snuffbox.

Needling

0.3–0.5 cun proximally or vertically. Avoid the radial artery by gently pushing it to the side when palpating for the point. Moxibustion is contraindicated according to some classics.

Actions/Indications

- Descends the Lung Qi, alleviates cough and wheezing
- Opens the pores and expels Wind in particular
- Opens the channel, especially locally

Special features

Jing-river point, Metal point, *ben* point (Five Phases).

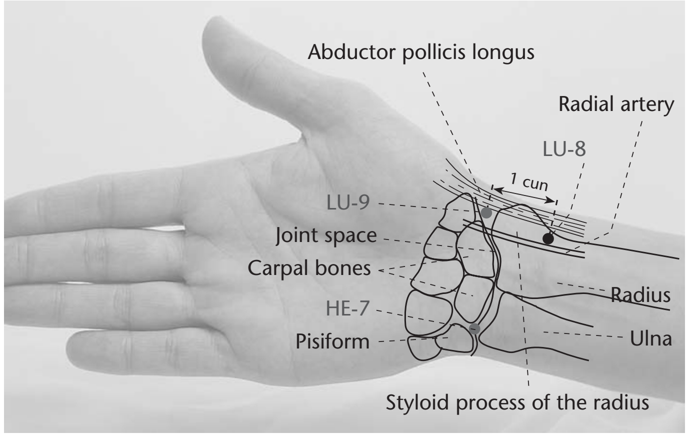
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones
Radius
HE-7
Ulna
Pisiform
Styloid process of the radius

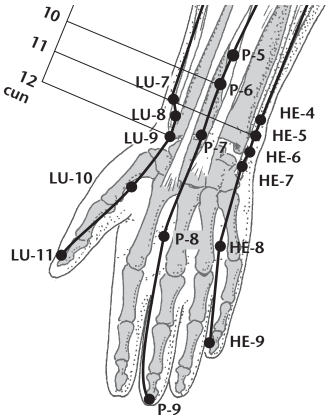
10
11
12
cun
P-5
LU-7
P-6
LU-8
HE-4
LU-9
HE-5
P-7
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

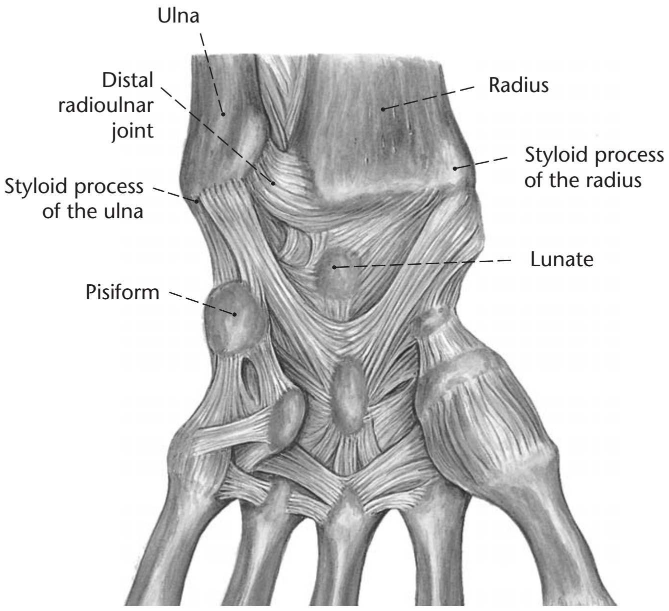
Ulna
Distal radioulnar joint
Radius
Styloid process of the radius
Styloid process of the ulna
Lunate
Pisiform

Left hand, palmar aspect

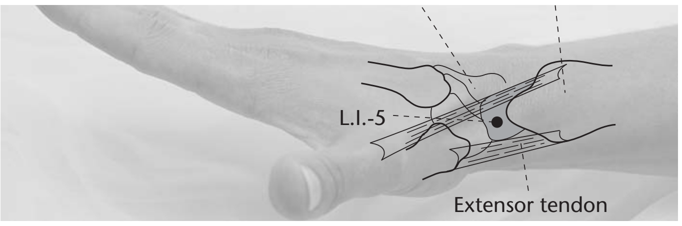
L.I.-5
Extensor tendon

Supreme Abyss TAIYUAN LU-9

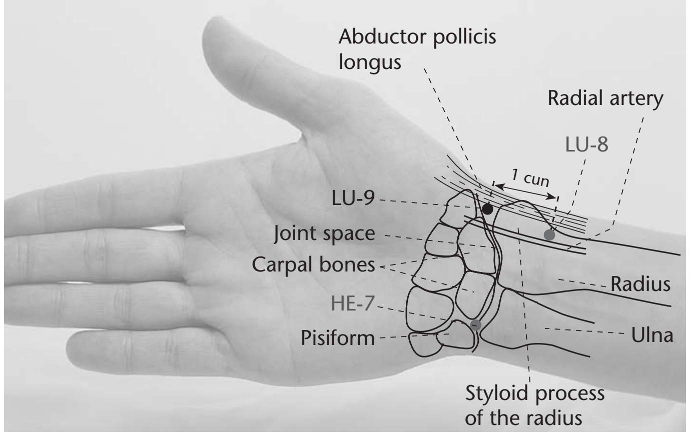
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones
HE-7
Pisiform
Radius
Ulna
Styloid process of the radius

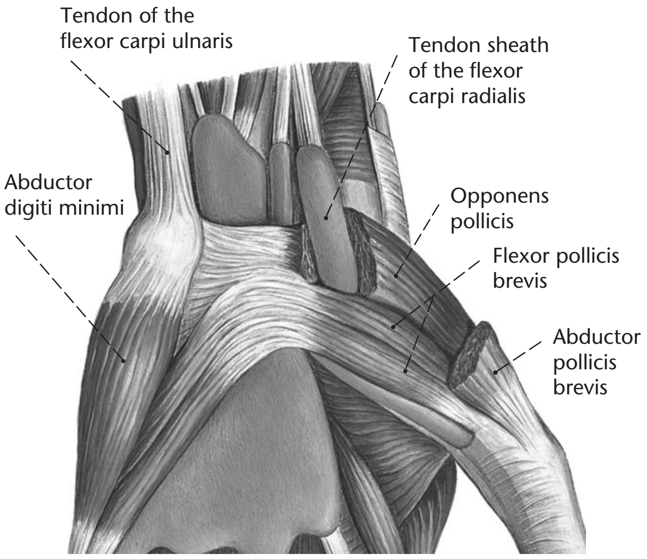
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Abductor digiti minimi
Opponens pollicis
Flexor pollicis brevis
Abductor pollicis brevis

Left hand, palmar aspect

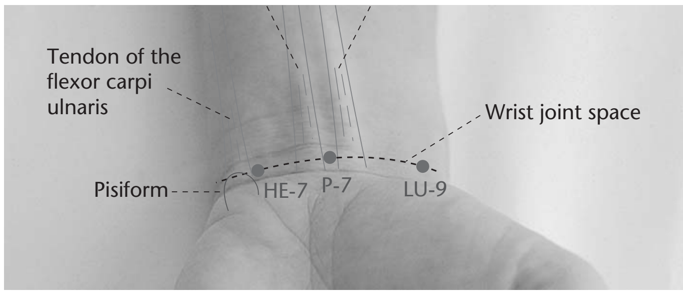
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

Location

On the ventral aspect of the wrist, at the level of the wrist joint space (most distal wrist crease), on the radial aspect of the radial artery and ulnar to the tendon of the abductor pollicis longus muscle.

How to find

As the location of the wrist crease varies, the ventral joint space (➞ 3.3.3) between the proximal row of carpals and the radius/ulna is a more reliable landmark for orientation. By loosely moving the hand, the joint space is easily palpable. Palpate the radial artery (pulsation) on the radial side and locate **LU-9** on its radial aspect. Located on the same level are ➞ **P-7** (between two tendons) and ➞ **HE-7** (radial to the insertion of the tendon of the flexor carpi ulnaris on the pisiform bone).

Needling

Vertically 0.2–0.5 cun. Caution: Avoid radial artery. If the needle is pulsating (desired effect), no additional stimulation!

Actions/Indications

- Strengthens the Lung, transforms Phlegm, descends the Lung Qi
- Regulates and harmonises the vessels (relationship between *zong qi* and Blood circulation)
- Opens the channel, alleviates pain

Special features

Yuan-source point, *shu*-stream point, Earth point, tonification point, *hui*-meeting point of the vessels. Important point for tonifying the Lung Qi and Yin, especially with chronic disorders.

LU-10 *Fish Border* YUJI

Location

At the midpoint of the palmar border of the 1st metacarpal bone.

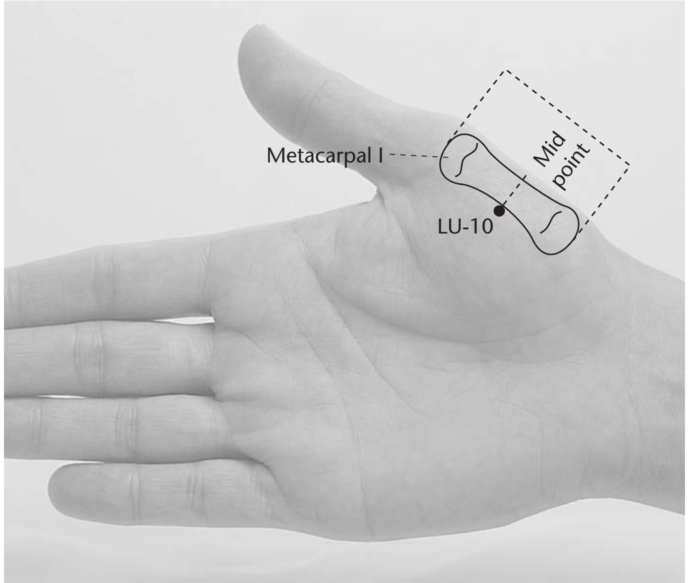
Metacarpal I
Mid point
LU-10

How to find

With the thumb relaxed, palpate the 'belly' of the thenar eminence from palmar to lateral (radial), then palpate the first metacarpal bone. **LU-10** is located at the midpoint of its palmar 'border'.

→ **L.I.-4** is located on the dorsal aspect of the hand, close to the 2nd metacarpal bone.

Needling

Vertically 0.5–1 cun. Caution: Painful!

Actions/Indications

- Regulates the Lung Qi
- Cools Blood Heat
- Expels pathogenic factors from the throat
- Descends rebellious Qi
- Harmonises the Stomach and Heart

Special features

Ying-spring point, Fire point, important distal point for sore throat caused by Wind-Heat.

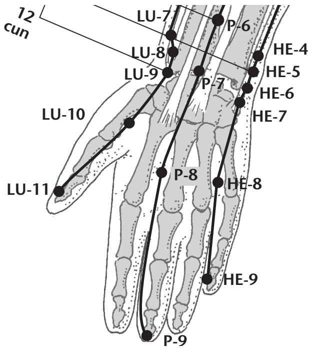
12 cun
LU-7
P-6
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

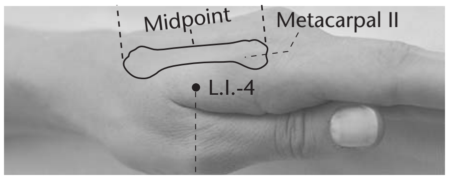
Midpoint
Metacarpal II
L.I.-4

Lesser Shang SHAOSHANG LU-11

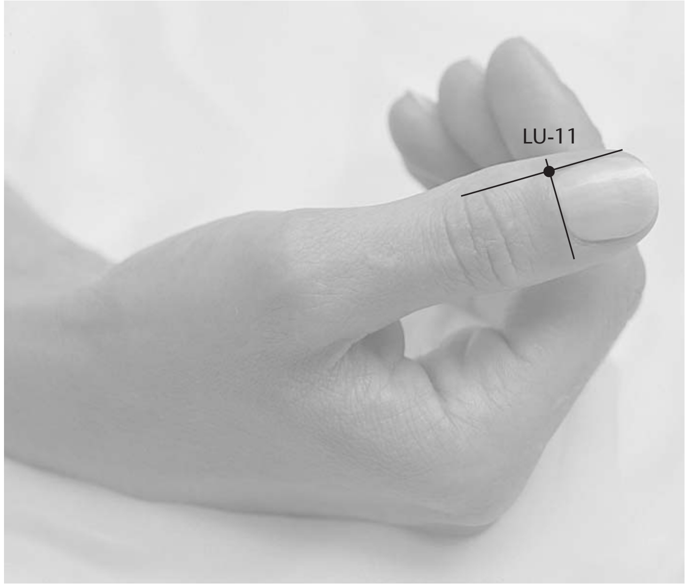

Location

On the thumb, 0.1 cun from the radial corner of the nail.

How to find

The point is located at the junction of two tangents running along the proximal and radial border of the thumb nail, approximately 0.1 cun from the actual border of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed. Avoid needling the perionychium.

Actions/Indications

- Frees the senses
- Regulates the Lung Qi
- Clears Heat (especially from the throat)
- Opens the channel

Special features

Jing-well point, Wood point, Sun Si Miao Ghost point, important distal point for sore throat caused by Wind-Heat.

4.2 The Large Intestine Channel System – Hand *yangming* (*shou yang ming jing luo*)

4.2.1 The Large Intestine Primary Channel (*shou yang ming jing*)

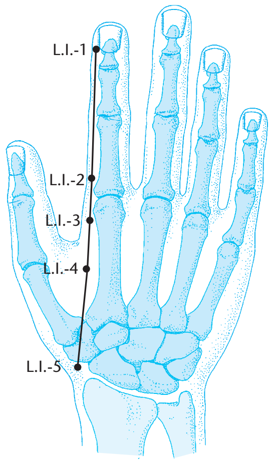
L.I.-1
L.I.-2
L.I.-3
L.I.-4
L.I.-5

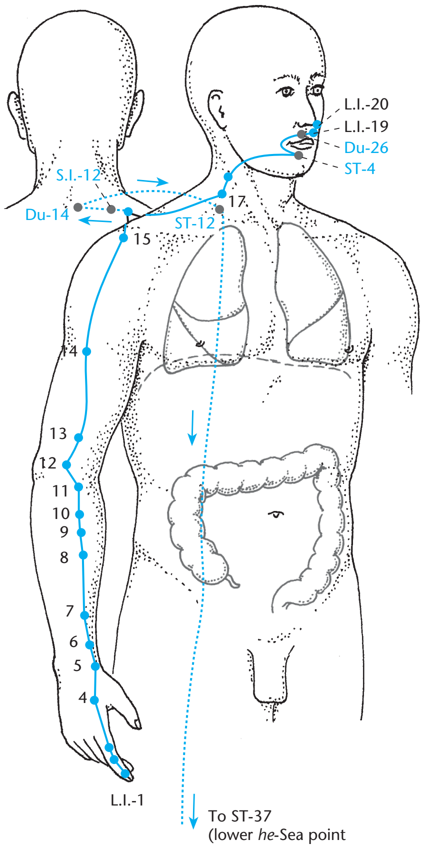
L.I.-20
L.I.-19
Du-26
ST-4
S.I.-12
Du-14
ST-12
17
15
14
13
12
11
10
9
8
7
6
5
4
L.I.-1

To ST-37
(lower *he*-Sea point
of the Large Intestine)

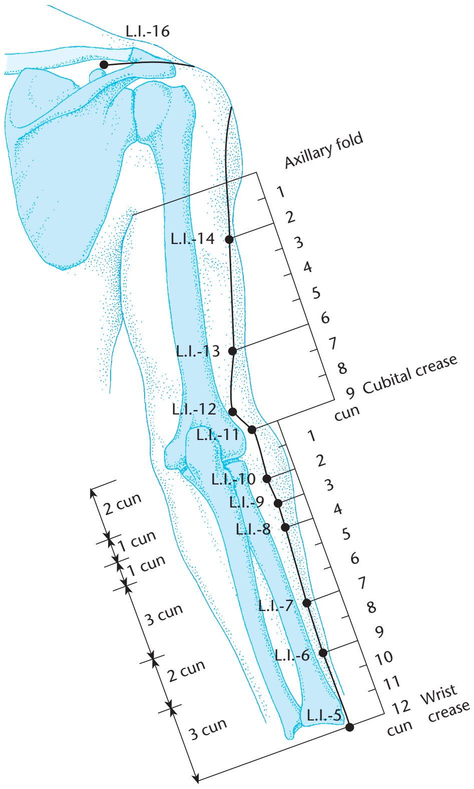
L.I.-16
Axillary fold
1
2
3
4
5
6
7
8
9
cun
Cubital crease
L.I.-14
L.I.-13
L.I.-12
L.I.-11
L.I.-10
L.I.-9
L.I.-8
L.I.-7
L.I.-6
L.I.-5
2 cun
1 cun
1 cun
3 cun
2 cun
3 cun
1
2
3
4
5
6
7
8
9
10
11
12
cun
Wrist crease

Pathway

The **external** part of the Large Intestine primary channel begins on the index finger at → **L.I.-1** (*shangyang*, at the radial aspect of the corner of the nail). This point is reached by a branch which separates from the Lung primary channel at → **LU-7** (*lieque*; hand Yin–Yang connection of the first great circuit). The Large Intestine primary channel runs between the first and second metacarpal bones,

- then runs proximally along the radial aspect of the forearm towards the lateral aspect of the elbow
- travels in a proximal direction along the lateral aspect of the upper arm towards the shoulder
- crosses the Small Intestine primary channel at → **S.I.-12** (*bingfeng*) in the centre of the suprascapular fossa
- continues to → **Du-14** (*dazhui*) inferior to the spinous process of the 7th cervical vertebra, where it meets with the other five Yang primary channels
- crosses the neck and enters the supraclavicular fossa at → **ST-12** (*quepen*).

Here the **internal** pathway branches off and connects with the Lung (*fei*), its paired *zang*-Organ, penetrates the diaphragm and enters the Large Intestine (*dachang*), its pertaining *fu*-Organ. From the Large Intestine (according to some authors: the Large Intestine primary channel) an **internal** branch runs in an inferior direction to → **ST-37** (*shangjuxu*) on the lateral aspect of the lower leg, the lower *he*-sea point of the Large Intestine.

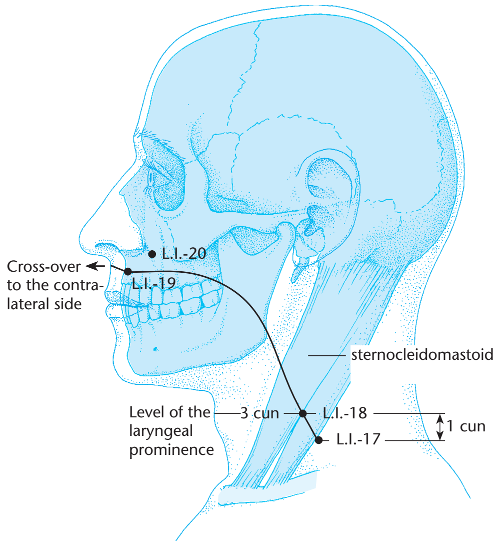
Cross-over to the contra-lateral side
L.I.-20
L.I.-19
sternocleidomastoid
Level of the laryngeal prominence
3 cun
L.I.-18
L.I.-17
1 cun

The **external** branch ascends from the supraclavicular fossa, travels along the lateral aspect of the neck and traverses the cheek.

➡ From here another **internal** branch enters the lower gums, where it extends further.

➡ The **external** course of the primary channel passes → **ST-4** (*dicang*) and curves around the mouth to the upper lip, crossing to the contralateral side of the body at → **Du-26** (*renzhong*), located at the philtrum. On the contralateral side of the body, it reaches → **L.I.-19** (*kouheliao*) and terminates **contralaterally** at → **L.I.-20** (*yingxiang*) on the side of the nose.

Note: Deadman et al. (1998) mention → **Ren-24** (*chengjiang*) as additional meeting point (variant → figure). They also mention → **G.B.-5** (*xuanlu*), → **G.B.-6** (*xuanli*) and → **G.B.-14** (*yangbai*) as classic meeting points with the Large Intestine channel but these are not normally shown in illustrations of the channel.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Fever, dryness of the mouth, thirst, a sore and swollen throat, nosebleed, toothache, red, painful eyes, stiff fingers, painful cold or painful hot swellings of the upper arm and shoulder

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain in the lower abdomen, borborygmus, diarrhoea or yellow stools with mucus, shortness of breath, dyspnoea

Excess (*shi*): Heat sensations along the channel

Deficiency (*xu*): Cold sensations and chills along the channel

Connections of the Large Intestine primary channel (→ 1.2)

With other channels

Lung primary channel (*shou tai yin jing*)

Connection: Hand Yin–Yang connection of the first great circuit
Location: LU-7 → L.I.-1 (on the hand). A branch of the Lung primary channel connects with the Large Intestine primary channel at **L.I.-1** (*shangyang*)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Stomach primary channel (*zu yang ming jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *yangming* (Yang axes of the first great circuit)
Location: L.I.-20 → (BL-1) → ST-1 (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

With the *zangfu*-Organ systems

Lung (*fei*), **Large Intestine (*dachang*)**, Stomach (*wei*)

4.2.2 The Large Intestine Divergent Channel (*shou yang ming jing bie*)

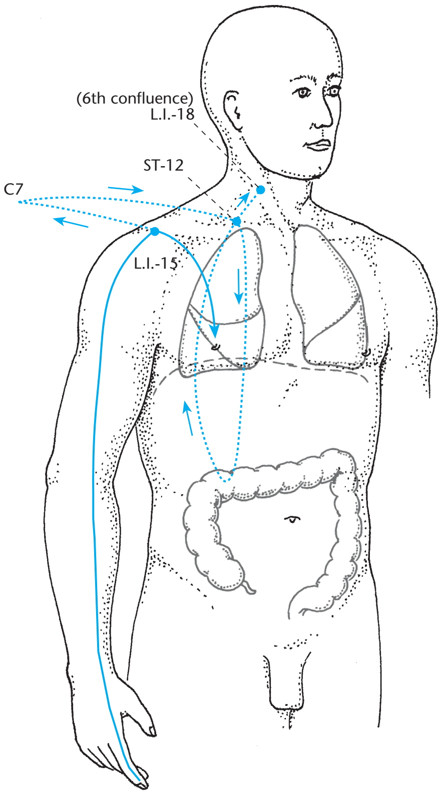

Pathway

The Large Intestine divergent channel separates from the Large Intestine primary channel on the hand,

➡ travels along the arm to the shoulder to → **L.I.-15** (*jianyu*)

➡ from **L.I.-15** a **branch** runs to the thorax and branches out into the breasts

➡ **another branch** runs to the cervical spine at the 7th cervical vertebra, returns from there to the supraclavicular fossa, descends to the Large Intestine (*dachang*), ascends to the Lung (*fei*) and **emerges** to the Exterior at → **ST-12** (*quepen*).*

It then travels along the anterolateral aspect of the neck and connects with the Large Intestine primary channel and the Lung divergent channel at → **L.I.-18** (*futu*) at one of the six *he*-confluences† (here: L.I./LU as 6th confluence → 1.3).

Clinical importance

- Strengthens the connection between the Lung and the Large Intestine (*zangfu*-Organs). Points on the Large Intestine primary channel can therefore be used for disorders of the Lung, and vice versa points on the Lung primary channel can treat disorders of the Large Intestine.
- A branch of the Large Intestine divergent channel spreads (from → **L.I.-15**) across the thorax and the breasts, thus extending the actions of the points on the Large Intestine primary channel to these regions, for example for the treatment of mastitis, myalgias and thoracic pain.

---

* Some authors describe the divergent channel as travelling directly from the spine to the Large Intestine, without meeting → **ST-12** (variant → Fig.).

† According to some authors the 6th *he*-Confluence is located at → **ST-12** (*quepen*).

4.2.3 The Large Intestine Sinew Channel (*shou yang ming jing jin*)

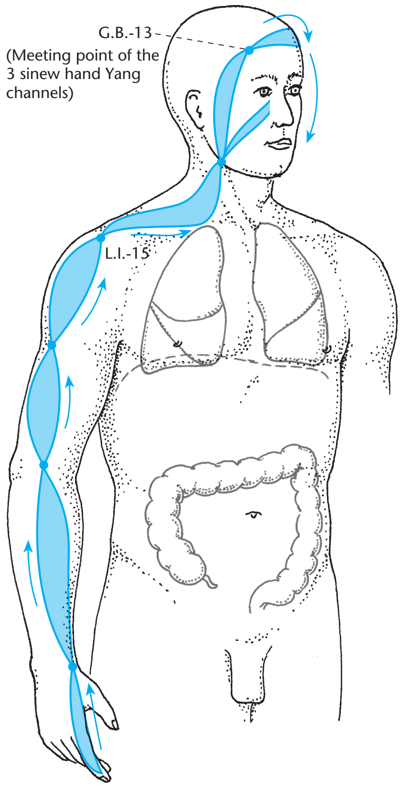

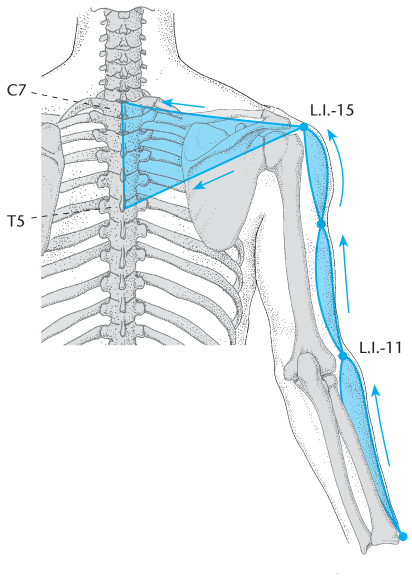

Pathway

The Large Intestine sinew channel

- begins on the index finger at → **L.I.-1** (*shangyang*)
- runs along the index finger and the 2nd metacarpal bone towards the wrist, where it binds (*jie*) at → **L.I.-5** (*yangxi*)
- travels along the posterolateral aspect of the forearm
- reaches the elbow and binds (*jie*) at → **L.I.-11** (*quchi*)
- continues along the lateral aspect of the upper arm and the deltoid muscle to the shoulder, where it binds (*jie*) again at → **L.I.-15** (*jianyu*)
- from **L.I.-15** a branch spreads across the scapula and then attaches to the spine between the 7th cervical vertebra and the 5th thoracic vertebra (or between → **Du-14** (*dazhui*) and → **Du-11** (*shendao*).
- The **main branch** ascends from **L.I.-15**, crosses the supraclavicular fossa and continues from there along the lateral aspect of the neck towards the angle of the jaw where it binds (*jie*).

Here it forks into two branches:

- **One branch** runs across the zygomatic bone towards the lateral side of the nose
- **The other branch** ascends the lateral aspect of the face and anterior to the Small Intestine sinew channel, connects at → **G.B.-13** (*benshen*) with the other sinew Yang channels, crosses over the head and ends at the lower jaw on the opposite side.

Note: According to some schools of thought, the sinew channel only ends at → **L.I.-17** (*tianding*; Solinas et al. 1998).

Clinical importance

Pathology: cramps, pain, tension and stiffness along the Large Intestine sinew channel. Limited range of motion of the upper limbs and shoulders. Occipital pain, stiffness and limited range of motion.

Indication: Mainly for *bi*-syndrome (painful obstruction syndrome) along the Large Intestine channel. The area covered by the Large Intestine sinew channel is larger than that covered by the Large Intestine primary channel. This explains why the indications of points on the Large Intestine primary channel include disorders and diseases of the upper neck and thoracic region as well as of the whole face (for example frontal and temporal headaches).

4.2.4 The Large Intestine *luo*-Connecting Vessel System (*shou yang ming luo mai*)

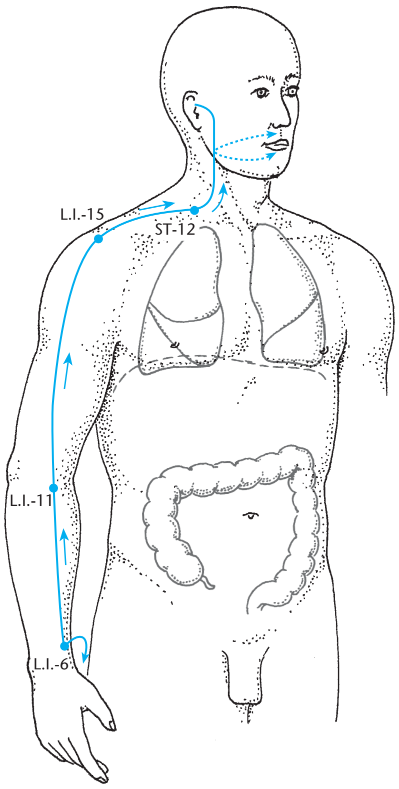

Pathway

The **Large Intestine** *luo*-connecting channel system begins at the *luo*-connecting point → **L.I.-6** (*pianli*) (→ 8.1.2), where it branches off from the primary channel and forms a three-dimensional reticular network, dividing into numerous branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions course to the Interiorly–Exteriorly paired Lung primary channel; according to some schools (for example, Van Nghi → Appendix) they travel as a **transverse** LU *luo*-connecting vessel to the *yuan*-source point → **LU-9** (*taiyuan*).

➡ A **longitudinally** orientated division follows the Large Intestine primary channel to **L.I.-15** (*jianyu*), continues via the supraclavicular fossa to → **ST-12** (*quepen*) and along the neck to the mandibular angle, where it spreads. One branch runs to the teeth; another branch reaches the ear, where it communicates with all channels reaching the ear.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Disorders of the teeth and gums, disorders of the ears
- **Deficiency (*xu*):** Sensitive teeth and gums (such as sensitivity to cold), sensation of tightness in the chest and diaphragm

4.2.5 Cutaneous Region (*yang ming pi bu*)

See description and figures → 1.6

4.2.6 Points of the Large Intestine Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): L.I.-4** (*hegu*) 🟦🟦
- ***Luo*-Connecting point (→ 8.1.2.): L.I.-6** (*pianli*) 🟦■
- ***Xi*-cleft point (→ 8.1.3): L.I.-7** (*wenliu*)
- **Associated Back-*shu* point (→ 8.1.4): BL-25** (*dachangshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): ST-25** (*tianshu*)
- **Five *shu*-transport point (→ 8.1.6):**
  jing-well point (Metal), *ben* point:. **L.-I.1** (*shangyang*)
  ying-spring point (Water), sedation point: **L.I.-2** (*erjian*)
  shu-stream point (Wood): **L.I.-3** (*sanjian*)
  jing-river point (Fire): **L.I.-5** (*yangxi*)
  he-sea point (Earth), tonification point: **L.I.-11** (*quchi*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the Small Intestine primary and BL channel, *yang wei mai*: **L.I.-14** (*binao*)*
  - with the *yang qiao mai*: **L.I.-15** (*jianyu*) 🟦■, **L.I.-16** (*jugu*)
  - with the primary ST channel: **L.I.-20** (*yingjiang*) 🟦■
  - of other channels with the L.I. channel: **ST-4, ST-12, S.I.-12, Du-14, Du-26, Ren-24***; according to some classics: **G.B.-5, G.B.-6, G.B.-14**
- **Gao Wu command point (→ 8.1.11) for the face and mouth: L.I.-4** (*hegu*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): L.I.-18** (*futu*)
- **Point of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): L.I.-4** (*hegu*) 🟦🟦, **L.I.-11** (*quchi*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): L.I.-11** (*quchi*) 🟦🟦

---

* Mentioned by only some authors

Points according to region

- **Local points (→ 8.2.1):** nose – **L.I.-20** (*yingxiang*) 🟦⬛; shoulder – **L.I.-15** (*jianyu*) 🟦⬛; elbow – **L.I.-11** (*quchi*) 🟦🟦; hand – **L.I.-5** (*yangxi*); hand and fingers – **L.I.-4** (*hegu*) 🟦🟦; fingers (numbness and pain) – **L.I.-3** (*sanjian*)
- **Regional points (→ 8.2.1):** elbow – **L.I.-10** (*shousanli*), **L.I.-13** (*shouwuli*); shoulder and elbow – **L.I.-14** (*binao*)
- **Distal points (→ 8.2.1):** forehead, eyes, mouth/teeth – **L.I.-4** (*hegu*) 🟦🟦; nose – **L.I.-3** (*sanjian*), **L.I.-4** (*hegu*) 🟦🟦; neck – **L.I.-4** (*hegu*) 🟦🟦, **L.I.-11** (*quchi*) 🟦🟦; mouth/teeth – **L.I.-2** (*erjian*), **L.I.-4** (*hegu*) 🟦🟦; shoulder, elbow, hand, fingers – **L.I.-4** (*hegu*) 🟦🟦, **L.I.-1** (*shangyang*); knee and foot – **L.I.-5** (*yangxi*); foot – **L.I.-2** (*erjian*), **L.I.-3** (*sanjian*); toes – **L.I.-4** (*hegu*) 🟦🟦

Specific points according to the course of the channel (in numerical order)

- **L.I.-1** (*shangyang*): *jing*-well point (Metal) (→ 8.1.6); *ben* point (Five Phase point); distal point for the elbow and the hand (→ 8.2.1)
- **L.I.-2** (*erjian*): *ying*-spring point (Water) (→ 8.1.6); sedation point; distal point for the mouth/teeth (→ 8.2.1); distal point for the foot (→ 8.2.1)
- **L.I.-3** (*sanjian*): *shu*-stream point (Wood) (→ 8.1.6); local point for the fingers (numbness and pain) (→ 8.2.1); distal point for the nose (→ 8.2.1); distal point for the feet (→ 8.2.1)
- **L.I.-4** (*hegu*) 🟦🟦: *yuan*-source point (→ 8.1.1); Ma Dan Yang Heavenly Star point (→ 8.1.14); Gao Wu command point (→ 8.1.11) for the face and mouth; important point for acupuncture anaesthesia; distal point for the forehead, eyes, nose, mouth/teeth, jaw, shoulder, elbow, hand, fingers (→ 8.2.1); local point for the hand and fingers (stiffness) (→ 8.2.1)
- **L.I.-5** (*yangxi*): *shu*-stream point (Fire) (→ 8.1.6); local point for the hand (→ 8.2.1); distal point for the knees and feet (→ 8.2.1)
- **L.I.-6** (*pianli*): *luo*-connecting point (→ 8.1.2)
- **L.I.-7** (*wenliu*): *xi*-cleft point (→ 8.1.3)
- **L.I.-10** (*shousanli*): adjacent point for the elbow (→ 8.2.1)
- **L.I.-11** (*qichi*) 🟦🟦: Sun Si Miao Ghost point (→ 8.1.15); *he*-sea point (Earth) (→ 8.1.6); tonification point; Ma Dan Yang Heavenly Star point; distal point for the neck (→ 8.2.1); local point for the elbow (→ 8.2.1)
- **L.I.-13** (*shouwuli*): regional point for the elbow (→ 8.2.1)
- **L.I.-14** (*binao*)*: *jiaohui*-meeting point with the S.I. and BL channel and the *yang wei mai** (→ 8.1.10); regional point for the shoulder and elbow (→ 8.2.1)
- **L.I.-15** (*jianyu*) 🟦⬛: *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10); local point for the shoulder (→ 8.2.1)
- **L.I.-16** (*jugu*): *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)
- **L.I.-18** (*futu*): Window of Heaven point (→ 8.1.12)
- **L.I.-20** (*yingxiang*) 🟦⬛: local point for disorders of the nose (→ 8.2.1); *jiaohui*-meeting point with the ST channel (→ 8.1.10)

General location help

- The course of the line connecting **L.I.-5** and **L.I.-11** (12 cun) changes depending on whether the forearm is in a supinated or a pronated position. For location purposes the forearm should be in a midposition with the elbow slightly flexed.
- Since the L.I. channel crosses to the opposite side of the body at **Du-26** (nasolabial groove), locate **L.I.-19** and **L.I.-20** contralaterally.

---

* Mentioned by only a few authors

Shang Yang SHANGYANG L.I.-1

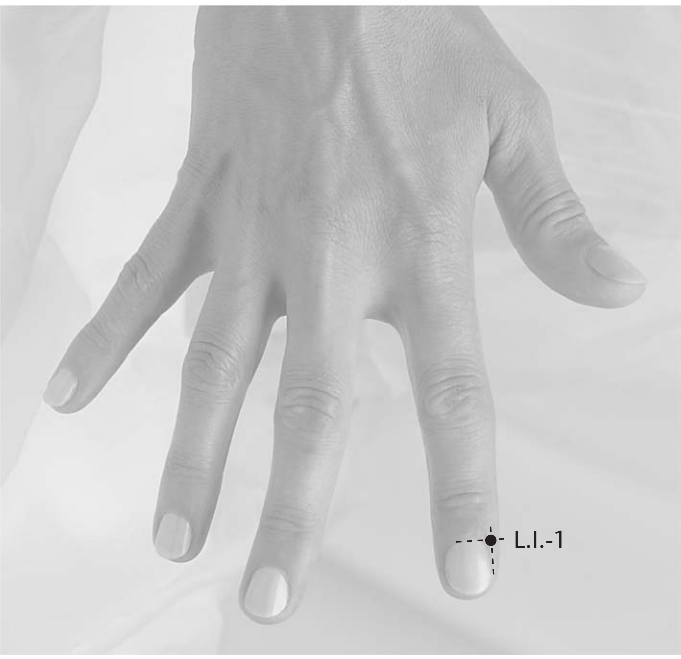

Location

On the index finger, 0.1 cun from the radial corner of the nail.

How to find

This point is located where two tangents bordering the nail of the index finger medially and proximally cross, approximately 0.1 cun from the corner of the nail.

Needling

0.2–0.3 cun vertically or obliquely in a proximal direction. Avoid needling the perionychium. For acute disorders (pain, inflammation), prick to bleed with needle, lancet or three-edged needle.

Actions/Indications

- Expels pathogenic factors and opens the channel
- Frees the senses

Special features

Jing-well point, Metal point, *ben* point (Five Phases), entry point according to some authors.

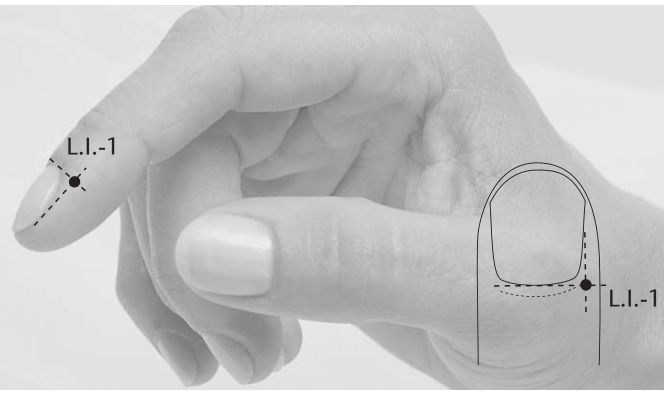

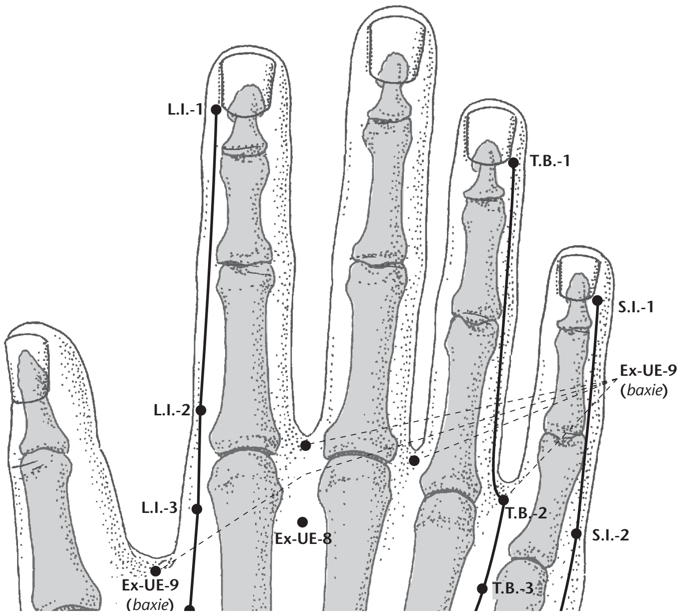

L.I.-2 *Second Space* ERJIAN

Location

On the radial aspect of the index finger, distal to the metacarpophalangeal joint, at the junction of the shaft and the basis of the proximal phalanx.

How to find

With the hand relaxed, palpate the border between the 'red to white' skin along the shaft of the second proximal phalangeal bone towards the metacarpo-phalangeal joint until the angle between the base and the shaft of the bone can be clearly felt. **L.I.-2** is located at the junction of the shaft and the base and slightly below (palmar to) the most external curvature of the bone.

→ **S.I.-2** is located in a similar position at the proximal phalanx of the little finger.

Needling

0.2–0.5 cun obliquely in a proximal direction or distal in a slightly palmar direction.

Actions/Indications

- Clears Heat and expels Wind

Special features

Ying-spring point, Water point, sedation point, distal point for the mouth and teeth

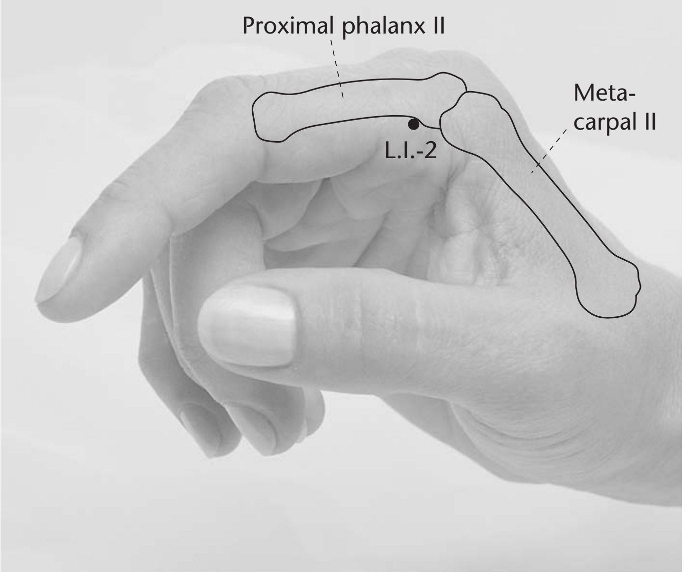

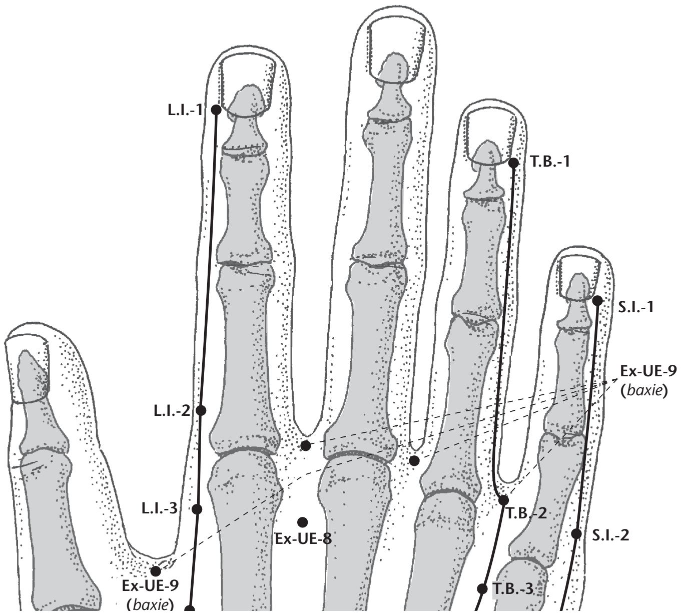

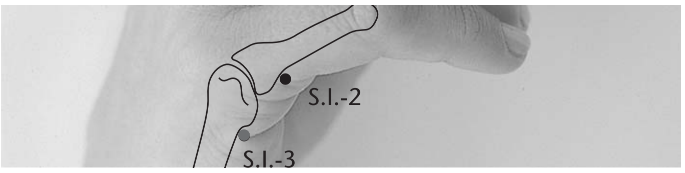

Third Space SANJIAN **L.I.-3**

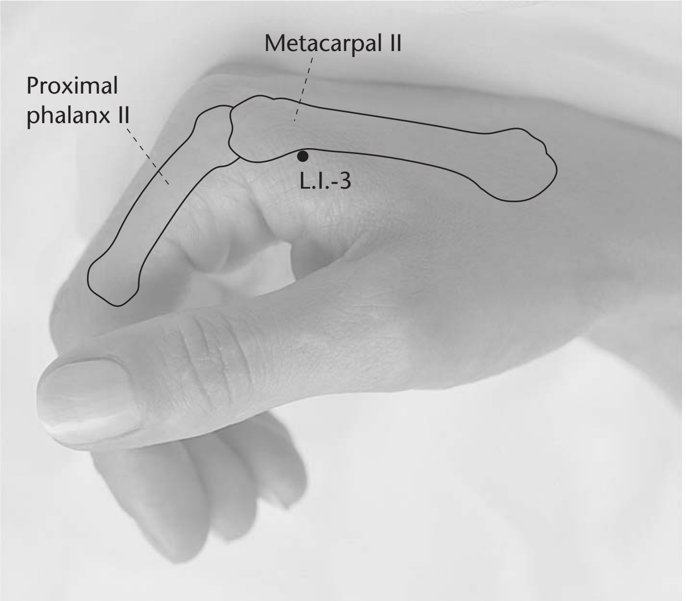
Metacarpal II
Proximal phalanx II
L.I.-3

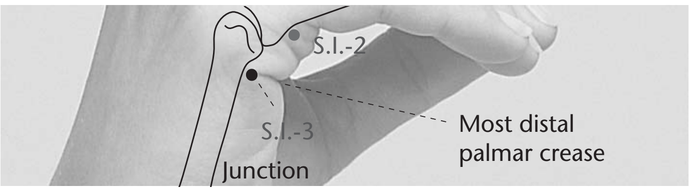
S.I.-2
S.I.-3
Junction
Most distal palmar crease

Location

On the radial aspect of the index finger, proximal to the metacarpo-phalangeal joint, at the junction of the shaft and the head of the 2nd metacarpal bone.

How to find

With the hand relaxed, palpate along the radial aspect of the 2nd metacarpal bone in a distal direction until the angle formed between the head and the shaft can be clearly felt. **L.I-3** is located at the junction of the shaft and the head and slightly below (palmar to) its most external curvature. → **S.I.-3** is located in a comparable position on the ulnar aspect of the 5th metacarpal bone (on the ulnar border of the hand). Located in a comparable location on the feet are → **SP-3** and → **BL-65**.

Needling

With the hand relaxed, 0.3–1 cun vertically directly below the lower border of the metacarpal bone towards → **S.I.-3**.

Actions/Indications

- Clears Heat and expels Wind
- Benefits the throat and teeth
- Alleviates pain

Special features

Shu-stream point, Wood point. Important point for pain in the fingers and the metacarpus.

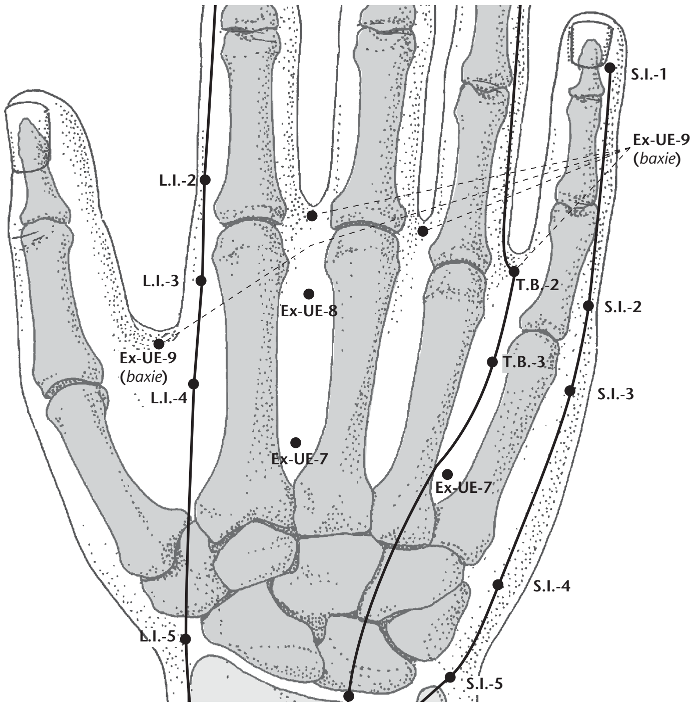
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5

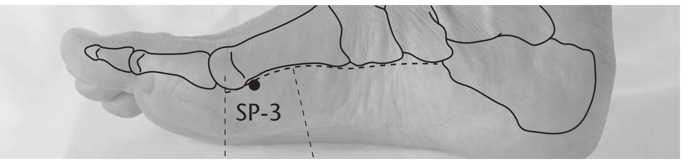

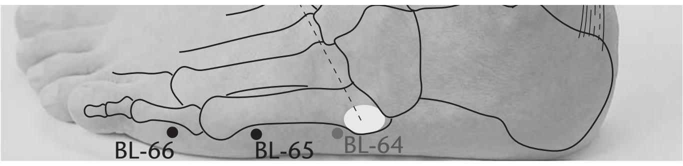
BL-66
BL-65
BL-64

Ch04.2-F10028.qxd 2/23/08 7:05 PM Page 104

L.I.-4 *Joining Valley* HEGU

Location

On the radial aspect of the hand, between the 1st and 2nd metacarpal bones, closer to the 2nd metacarpal bone and approximately at its midpoint.

How to find

When pressing the thumb and index finger together, the adductor pollicis and the interosseus dorsalis muscles will form a bulge. Needle **L.I.-4** at the highest point of the bulge and push it further towards the inferior aspect of the second metacarpal bone. **Or**: Spread the thumb and forefinger and place the distal phalanx of the thumb of the other hand on the webbed border. When flexing the thumb, its tip will point to **L.I.-4. Or**: Insertion on the angle bisector between the 1st and 2nd metacarpal bones, needling in the direction of the midpoint of the second metacarpal bone.

Needling

0.5–1 cun vertically or slightly obliquely towards the palmar aspect of the hand. Caution: Reducing needle techniques are contraindicated during pregnancy; exception: to induce labour.

Actions/Indications

- Releases the Exterior (main point!)
- Regulates the face and head
- Regulates the Defensive Qi (*wei qi*) and sweating
- Opens the channel and *luo*-connecting vessels
- Alleviates pain
- Promotes labour

Special features

Yuan-source point, Gao Wu command point for the head and mouth, Ma Dan Yang Heavenly Star point, entry point according to some authors. Most important analgesic point for the whole body. Most important single distal point for disorders of the face and sensory organs. It is the most commonly used point in clinical practice. The combined bilateral needling of **L.I.-4** and → **LIV-3** (between the 2nd and 3rd metatarsal bones) is known as *si guan* (Four Gates): they strongly regulate the Qi and Blood, stop pain and relieve spasms.

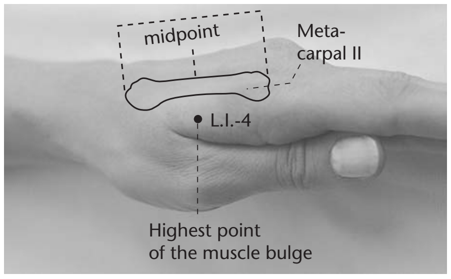
midpoint
Meta-
carpal II
L.I.-4
Highest point
of the muscle bulge

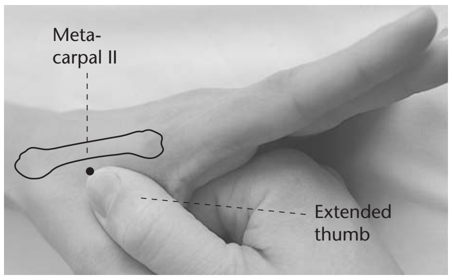
Meta-
carpal II
Extended
thumb

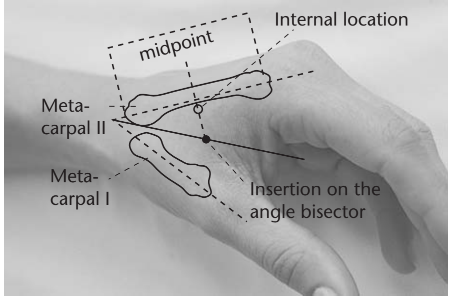
midpoint
Internal location
Meta-
carpal II
Meta-
carpal I
Insertion on the
angle bisector

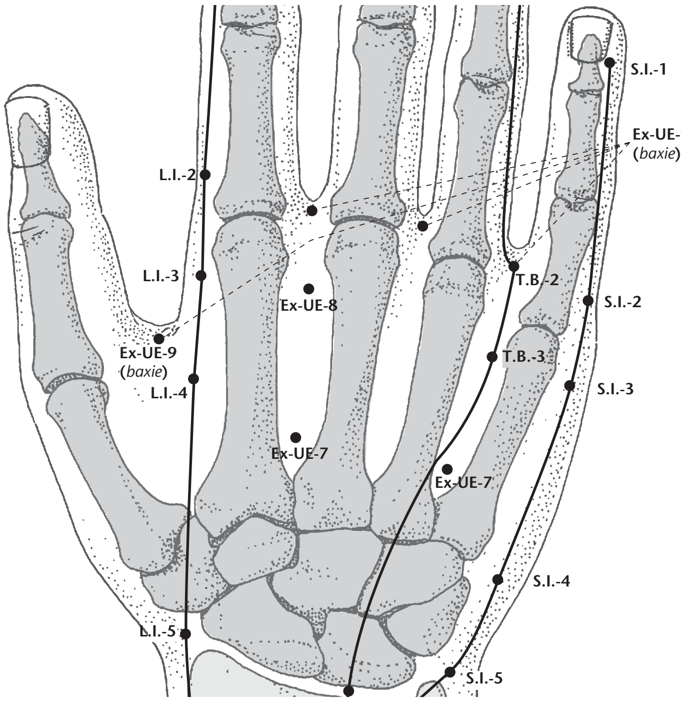
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5

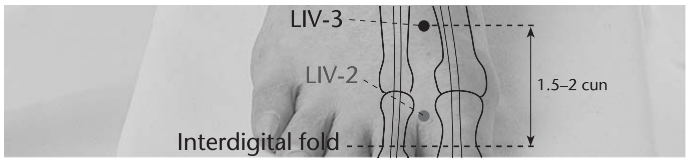
LIV-3
LIV-2
1.5–2 cun
Interdigital fold

Yang Stream YANGXI L.I.-5

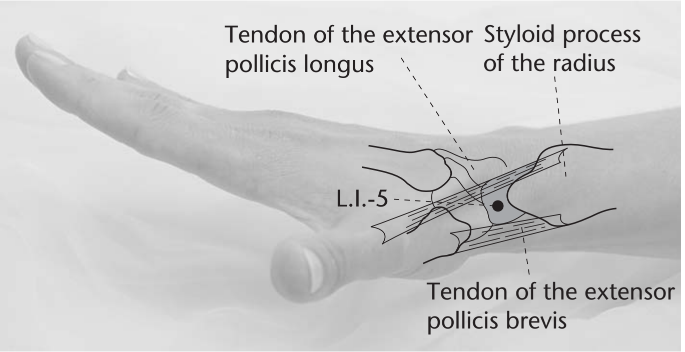
Tendon of the extensor pollicis longus
Styloid process of the radius
L.I.-5
Tendon of the extensor pollicis brevis

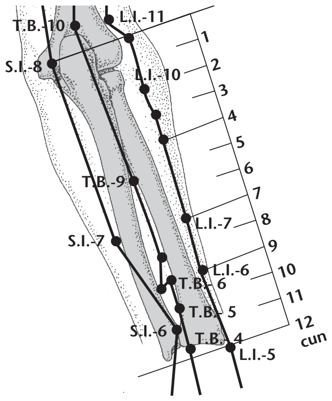
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
8
9
10
11
12
cun
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5

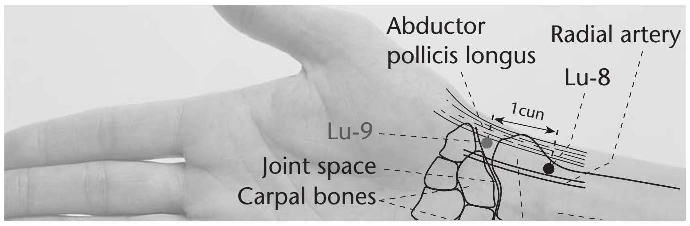
Abductor pollicis longus
Radial artery
Lu-8
1cun
Lu-9
Joint space
Carpal bones

Location

With the thumb abducted, in a depression between the tendons of the extensor pollicis longus and brevis muscles (‘anatomical snuffbox’), on the radial aspect of the wrist.

How to find

The anatomical snuffbox is a depression on the radial aspect of the wrist formed when abducting the thumb. Holding the hand in a horizontal position, it opens towards the body. **L.I.-5** is located on the wrist joint space, which becomes palpable by moving the wrist. → **LU-9** is also located on the wrist joint space and can be found by moving anteriorly from **L.I.-5** across the short tendon of the extensor pollicis muscle.

Needling

Vertically 0.5–1 cun. Caution: Avoid the superficial cephalic vein.

Actions/Indications

- Supports the wrist joint
- Clears Fire and expels Wind
- Clears Fire from the *yangming* (L.I., ST) and calms the *shen*

Special features

Jing-river point, Fire point. Important local point for disorders of the wrist, together with **T.B.-4, S.I.-5, HE-7, P-7** and **LU-9**.

L.I.-6 *Veering Passage* PIANLI

Location

3 cun proximal to → **L.I.-5** (in the centre of the anatomical snuffbox), on the line connecting → **L.I.-5** and → **L.I.-11**, between the abductor pollicis longus and the extensor pollicis brevis muscles, at the level of the junction between the tendon and the muscle.

How to find

Please note: The course of the line connecting → **L.I.-5** and → **L.I.-11** depends on the position of the forearm. For location purposes, a midposition of the forearm with flexed elbow is recommended. To locate **L.I.-6**, measure 3 cun (1 handsbreath) from → **L.I.-5** (centre of the anatomical snuffbox) on the connecting line. **Or**: 'Tiger mouth grip' (→ Fig. 2.6): Spread thumb and index finger of both hands, crossing them so that the forefinger of one hand comes to rest on the styloid process of the other, but avoiding bending the wrist. Place the middle finger against the forefinger: **L.I.-6** is located directly below the tip of the middle finger.
Located at the same level (3 cun proximal to the wrist joint space) on the dorsal side of the forearm are → **T.B.-6** (in the depression between the radius and the muscle) and → **T.B.-7** (in the depression between the ulna and the muscle) as well as → **P-5** (on the ventral side of the forearm between the tendons).

Needling

Obliquely or transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Expels Wind, clears Heat
- Regulates and opens the water passages
- Opens the channel and *luo* vessels

Special features

Luo-connecting point. Major point of the L.I. channel for regulating the water passages.

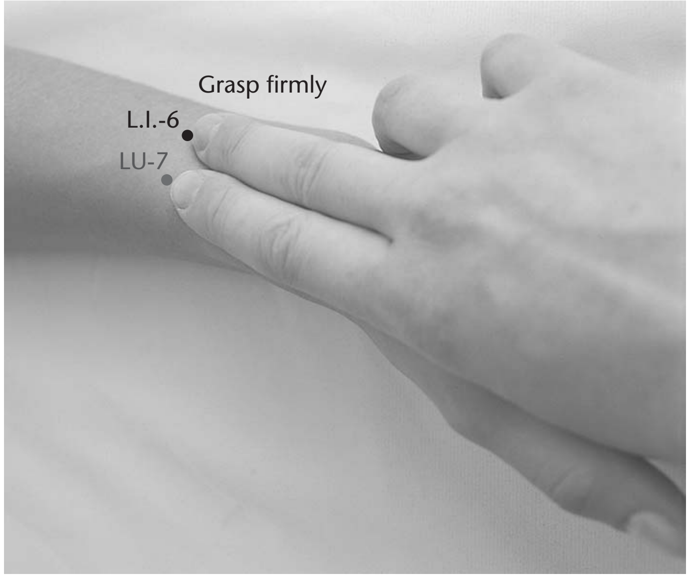

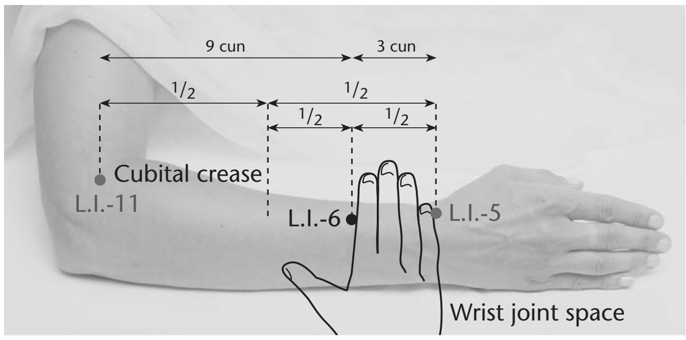

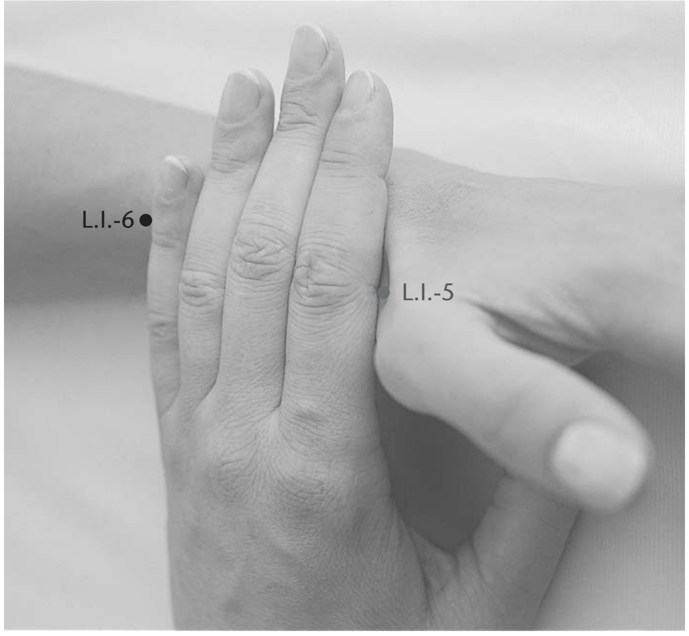

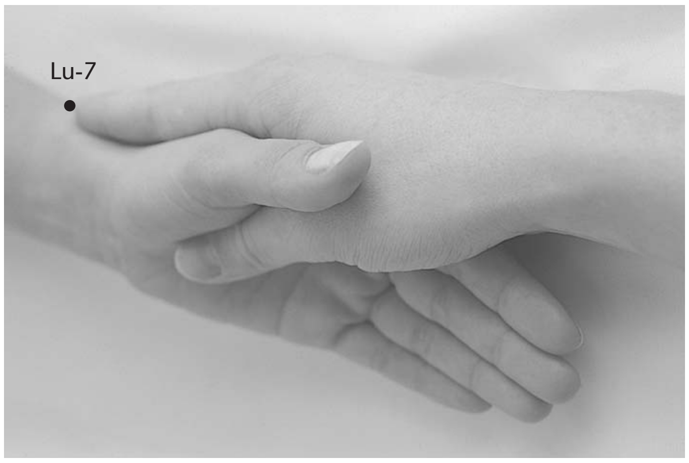

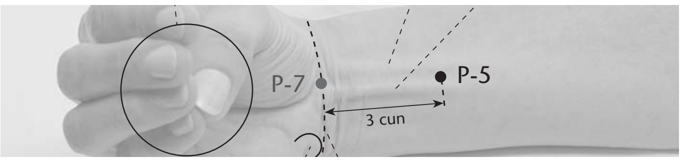

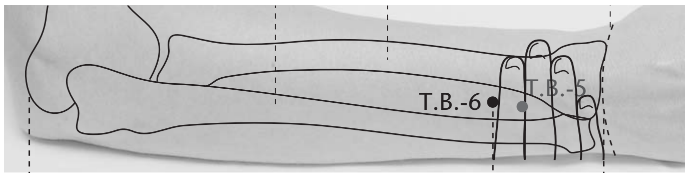

Warm Flow WENLIU L.I.-7

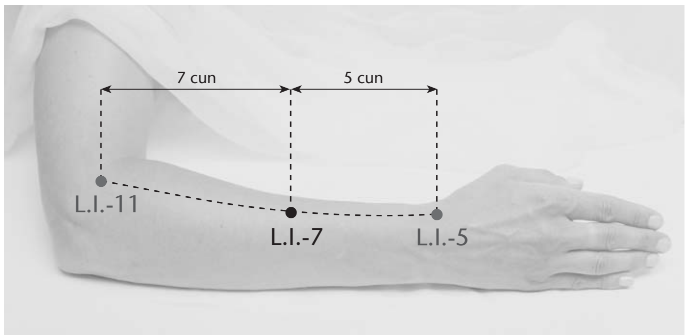
7 cun
5 cun
L.I.-11
L.I.-7
L.I.-5

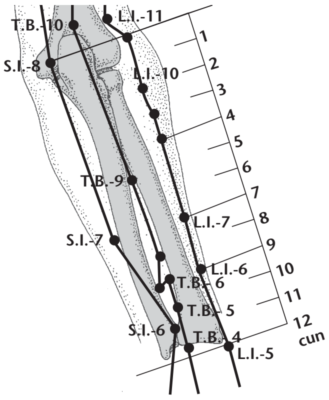
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
T.B.-9
L.I.-7
8
S.I.-7
9
L.I.-6
10
T.B.-6
11
T.B.-5
12
cun
S.I.-6
T.B.-4
L.I.-5

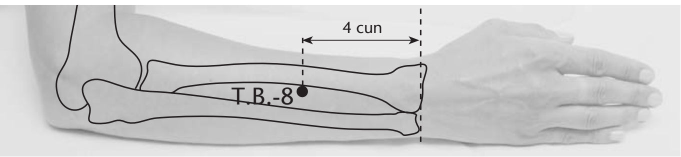

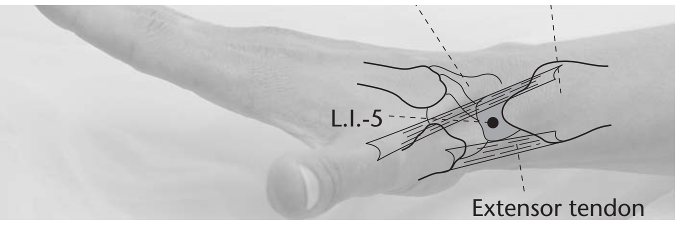
L.I.-5
Extensor tendon

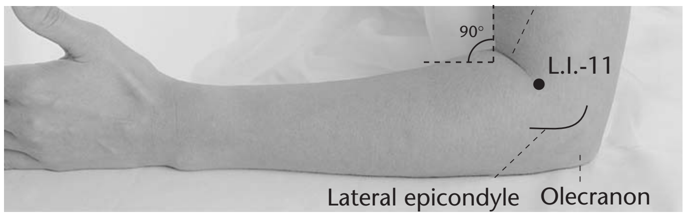
90°
L.I.-11
Lateral epicondyle
Olecranon

Location

5 cun proximal to the anatomical snuffbox in the direction of the lateral end of the elbow crease or 1 cun distal to the midpoint of the line connecting ➞ **L.I.-5** and ➞ **L.I.-11**.

How to find

Please note: With the hand supinated, the line connecting ➞ **L.I.-5** and ➞ **L.I.-11** runs alongside the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-7** lies on the dorso-lateral aspect of the radius and is best located with the forearm in midposition and the elbow flexed. Spreading hands technique: Place the little fingers on ➞ **L.I.-5** and ➞ **L.I.-11** and span the hands evenly so that the thumbs join at the midpoint of the line. From there, measure 1 cun distally. ➞ **T.B.-8** is located 4 cun proximal to the wrist crease between the radius and the ulna in the centre of the dorsal forearm or 1 cun distal to **L.I.-7**.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Treats acute disorders
- Clears Heat and toxic Heat
- Regulates the Stomach and Large Intestine

Special features

Xi-cleft point

L.I.-8 *Lower Angle* XIALIAN

Location

4 cun distal to the lateral end of the elbow crease in the direction of the anatomical snuffbox and on a line connecting → **L.I.-5** and → **L.I.-11**.

How to find

Please note: In supination, the line connecting → **L.I.-5** and → **L.I.-11** runs along the radial margin of the forearm; in pronation, it transverses the forearm. **L.I.-8** lies on the dorsal aspect of the radius and is best located with the forearm in mid-position and the elbow flexed. **L.I.-8** can be found either 4 cun distal to → **L.I.-11** or 2 cun proximal to the midpoint of the line connecting → **L.I.-5** and → **L.I.-11** (spreading hands technique, → 2.3.3).

→ **T.B.-9** is 5 cun distal to the olecranon between the radius and the ulna, about 1 cun distal to **L.I.-8** in the middle of the dorsal forearm.

Needling

Vertically or obliquely 0.5–1 cun

Actions/Indications:

- Opens the channel
- Cools and expels Wind
- Harmonises the Small Intestine

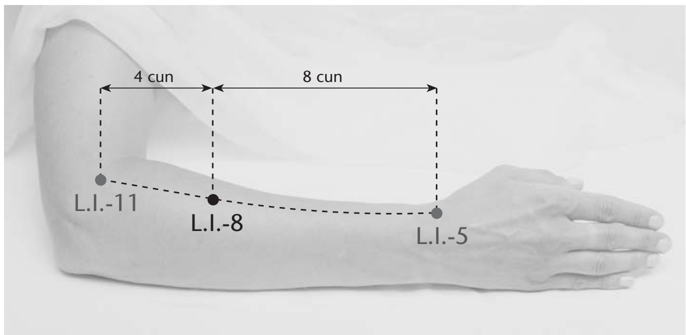
4 cun
8 cun
L.I.-11
L.I.-8
L.I.-5

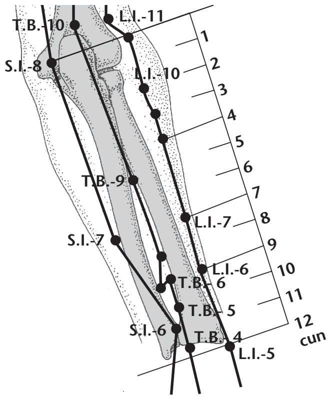
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
1
2
3
4
5
6
7
8
9
10
11
12
cun

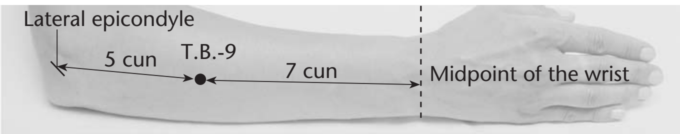
Lateral epicondyle
5 cun
T.B.-9
7 cun
Midpoint of the wrist

Upper Angle SHANGLIAN L.I.-9

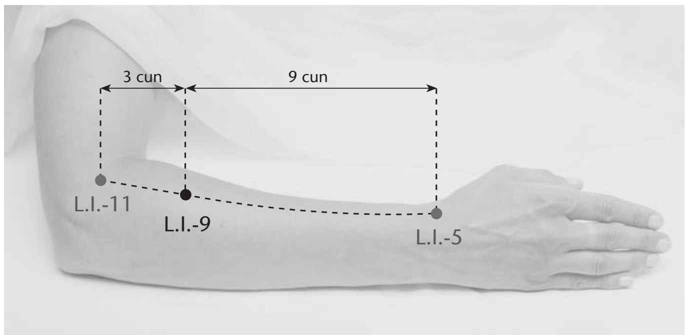

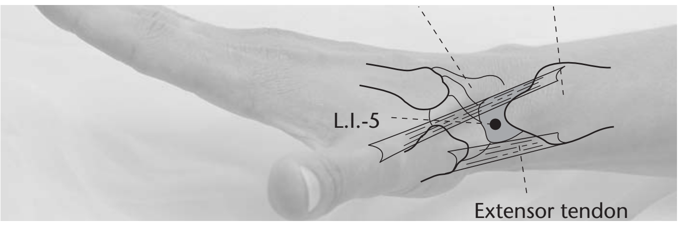

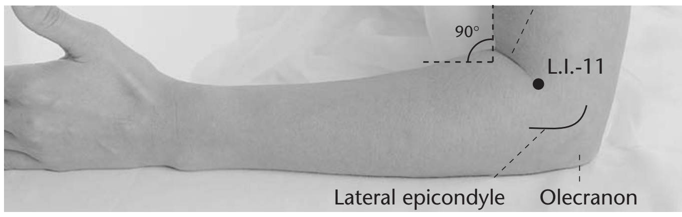

Location

3 cun distal to the lateral end of the elbow crease in the direction of the anatomical snuffbox, on the line connecting → **L.I.-5** and → **L.I.-11**.

How to find

Please note: With the hand supinated, the line connecting → **L.I.-5** and → **L.I.-11** runs along the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-9** lies on the dorsal aspect of the radius and is best located with the forearm in midposition and the elbow flexed. From → **L.I.-11**, measure 3 cun distally. The point is located in an easily palpable depression between two muscle bellies and is often tender with pressure.

Needling

Vertically or obliquely 0.5–1 cun

Actions/Indications

- Opens the channel, cools and expels Wind
- Harmonises the Large Intestine

L.I.-10 *Arm Three Miles* SHOUSANLI

Location

2 cun distal to → **L.I.-11**, on the line connecting → **L.I.-5** and → **L.I.-11**, on the extensor carpi radialis longus muscle; a deeper insertion will reach the supinator muscle.

How to find

Please note: With the hand supinated, the line connecting → **L.I.-5** and → **L.I.-11** runs along the radial margin of the forearm. However, with the arm pronated, it transverses the forearm. **L.I.-10** lies on the dorsal aspect of the radius and is best located with the forearm in midposition and with the elbow flexed. First, palpate → **L.I.-11** in the depression lateral to the radial end of the cubital crease. **L.I.-10**, usually sensitive to pressure, is found by measuring 2 cun distally from → **L.I.-11** on the connecting line.

Needling

Vertically 1–2 cun

Actions/Indications

- Regulates the Qi and Blood (in the upper extremity), opens the channel and *luo* vessels, alleviates pain
- Regulates the Stomach and Intestines

Special features

Important local point, often used in chain-and-lock method with other channel points

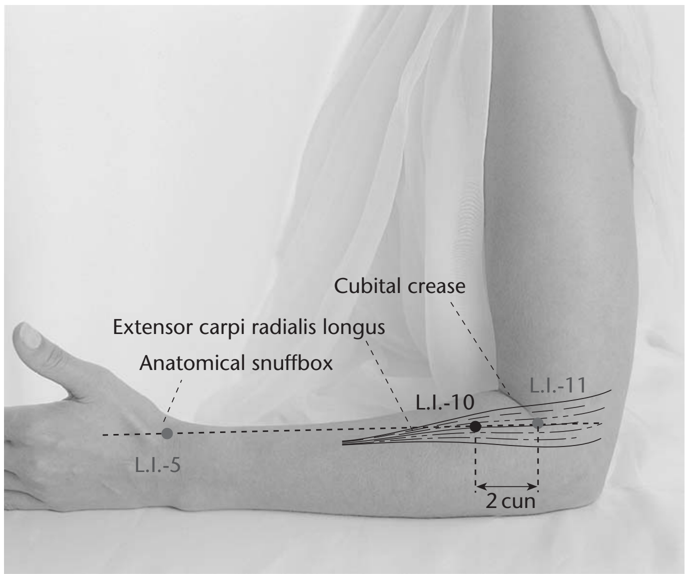
Cubital crease
Extensor carpi radialis longus
Anatomical snuffbox
L.I.-11
L.I.-10
L.I.-5
2 cun

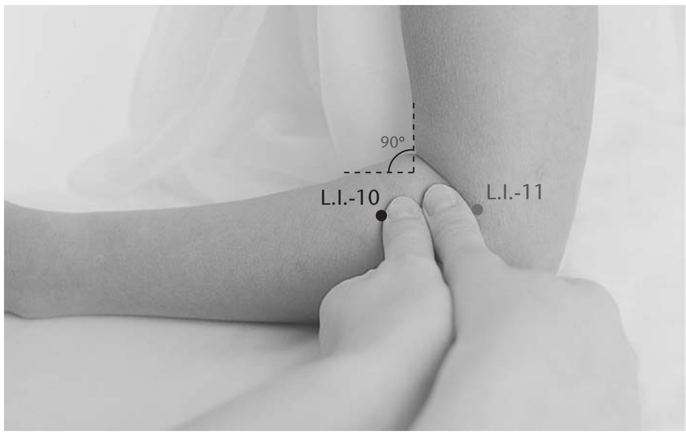
90°
L.I.-10
L.I.-11

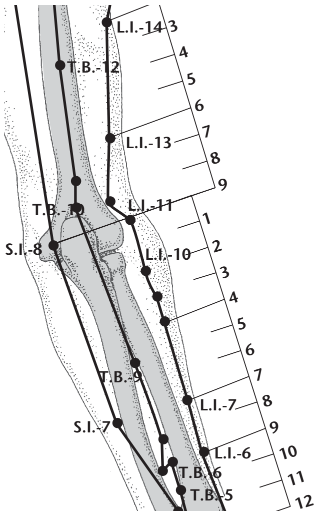
L.I.-14
T.B.-12
L.I.-13
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10
11
12

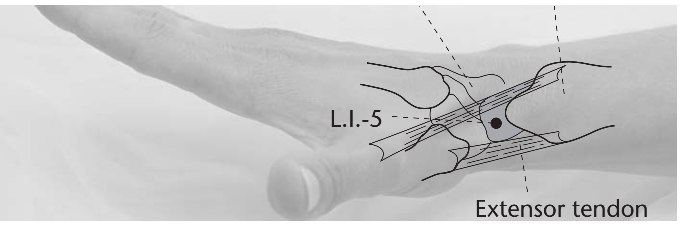
L.I.-5
Extensor tendon

Pool at the Crook QUCHI L.I.-11

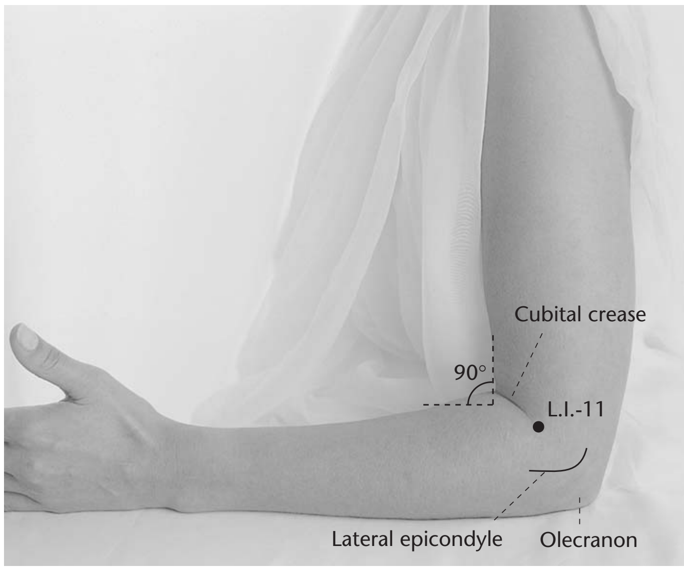
Cubital crease
90°
L.I.-11
Lateral epicondyle
Olecranon

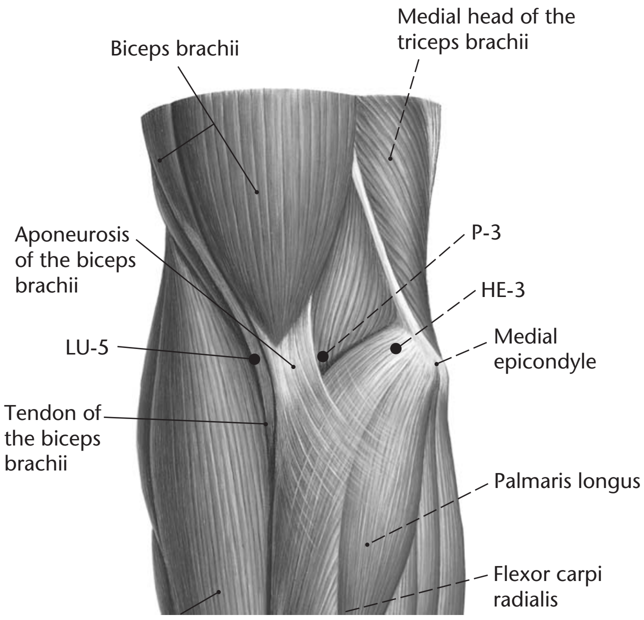
Biceps brachii
Medial head of the triceps brachii
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus
Flexor carpi radialis

Left arm, anterior aspect

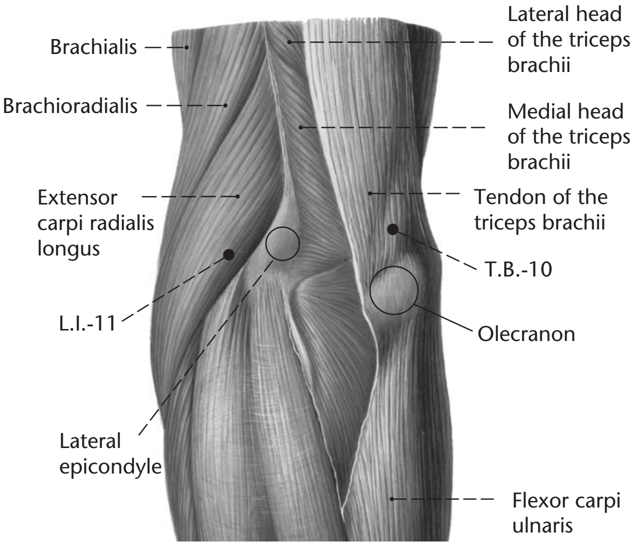
Brachialis
Lateral head of the triceps brachii
Brachioradialis
Medial head of the triceps brachii
Extensor carpi radialis longus
Tendon of the triceps brachii
T.B.-10
L.I.-11
Olecranon
Lateral epicondyle
Flexor carpi ulnaris

Left arm, posterior aspect

Location

With the elbow flexed, on the lateral end of the elbow crease, in a depression between the end of the crease and the lateral epicondyle of the humerus, on the extensor carpi radialis longus muscle.

How to find

First, with the elbow completely flexed, locate the lateral end of the elbow crease. Next, with the elbow flexed at approximately 90°, palpate for a pressure-sensitive depression in this area on the extensor carpi radialis longus muscle. **L.I.-11** is located close to the border of the proximal aspect of the ulna.

Also located at the level of the elbow crease are → **LU-5,** → **P-3** (radial/ulnar of the tendon of the biceps) and → **HE-3** (with the elbow completely flexed at the medial end of the cubital crease).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Heat, expels Wind
- Clears *yangming* – Fire
- Cools the Blood, drains Dampness, alleviates itching
- Opens the channel

Special features

He-sea point, Earth point, tonification point, Sun Si Miao Ghost point, Ma Dan Yang Heavenly Star point. Important point for Heat conditions and disorders of the upper extremity.

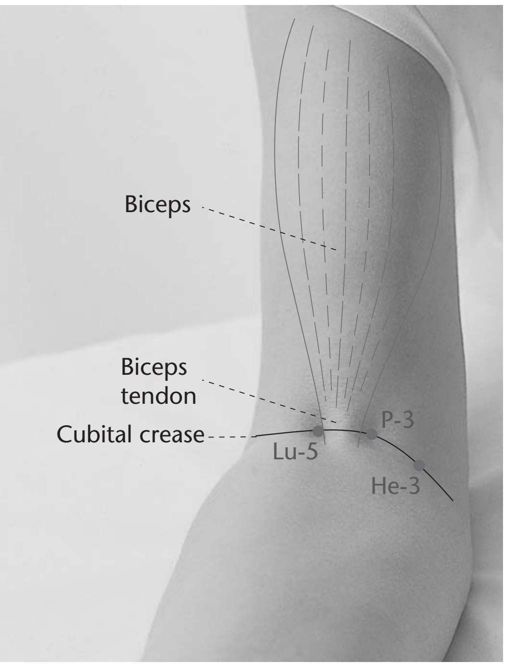
Biceps
Biceps tendon
Cubital crease
P-3
Lu-5
He-3

Right arm, anterior aspect

L.I.-12 *Elbow Crevice* ZHOULIAO

Location

1 cun proximal to the lateral end of the elbow crease (→ **L.I.-11**), on the anterior border of the humerus.

How to find

L.I.-12 is best located with the elbow flexed at 90°. From → **L.I.-11** (on the lateral end of the elbow crease), palpate 1 cun towards proximal. There, palpate for the anterior border of the humerus. **L.I.-12** is located at the junction of the shaft with the lateral epicondyle of the humerus and in front of the anterior border of the bone.

Needling

Vertically 0.5–1 cun, needle insertion between the border of the humerus and the flexor muscles in a medial direction.

Actions/Indications

- Opens the channel

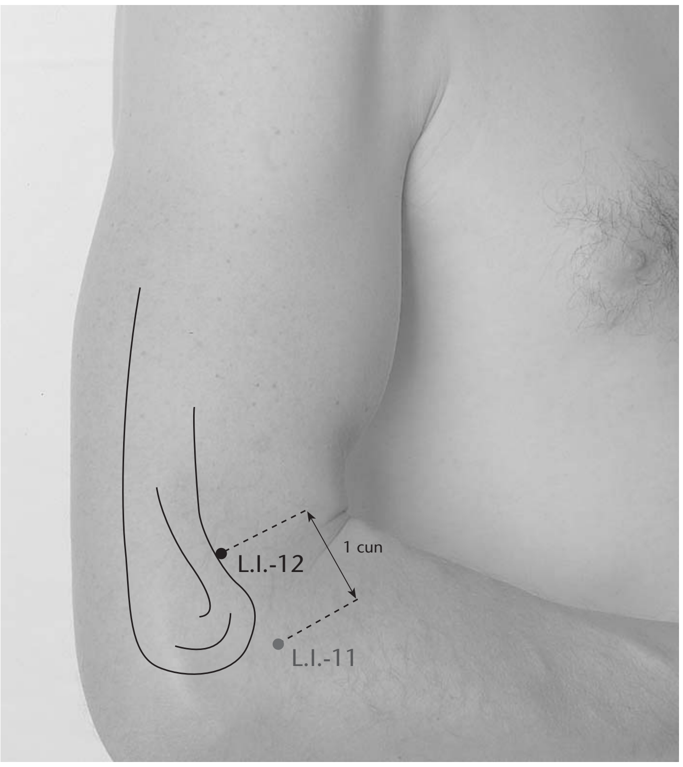
1 cun
L.I.-12
L.I.-11

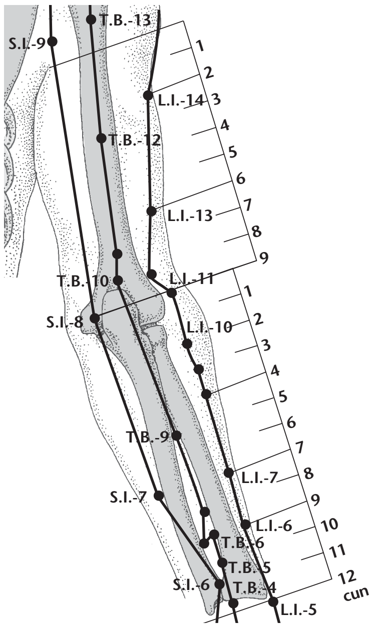
T.B.-13
S.I.-9
1
2
L.I.-14
3
T.B.-12
4
5
6
L.I.-13
7
8
9
T.B.-10
L.I.-11
1
S.I.-8
L.I.-10
2
3
4
5
6
T.B.-9
7
L.I.-7
8
S.I.-7
9
L.I.-6
10
T.B.-6
11
T.B.-5
12
cun
S.I.-6
T.B.-4
L.I.-5

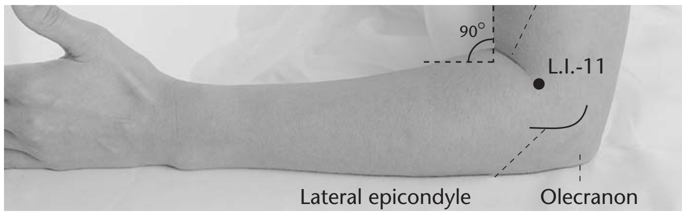
90°
L.I.-11
Lateral epicondyle
Olecranon

Arm Five Miles SHOUWULI L.I.-13

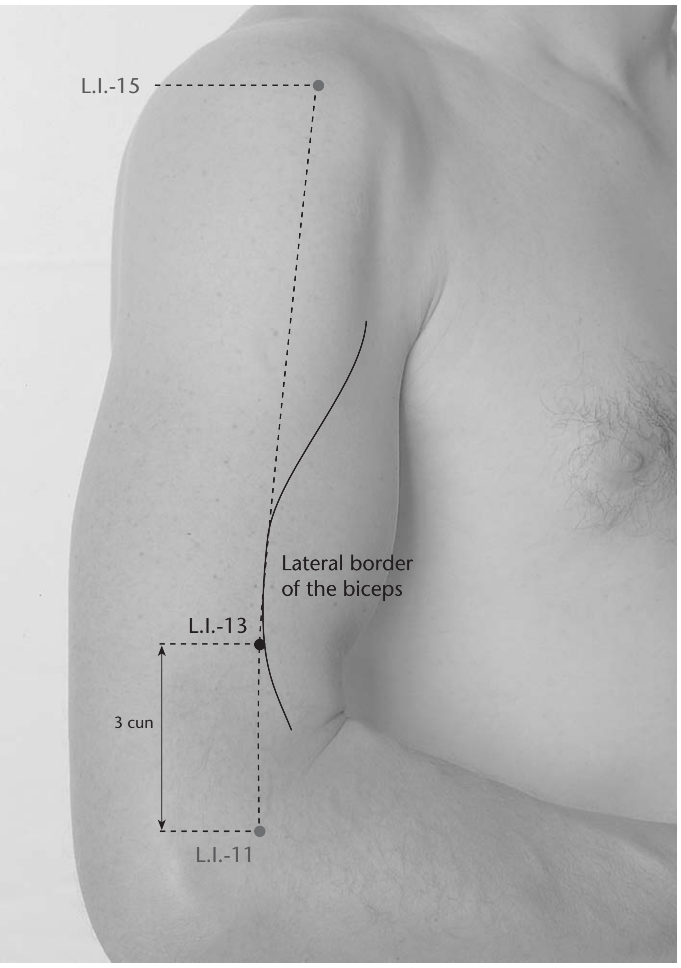
L.I.-15
Lateral border
of the biceps
L.I.-13
3 cun
L.I.-11

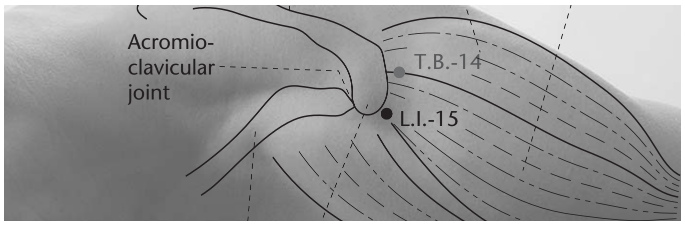
Acromio-
clavicular
joint
T.B.-14
L.I.-15

Location

On the lateral aspect of the upper arm, 3 cun proximal to the lateral end of the elbow crease (➞ **L.I.-11**), in the direction of the head of the humerus.

How to find

This point is located on a line connecting ➞ **L.I.-11** (lateral end of the elbow crease) and ➞ **L.I.-15** (in a depression anterior and inferior to the acromion), on the lateral border of the biceps, in a groove between the biceps and the brachialis muscles. With the elbow flexed at 90°, measure 3 cun from ➞ **L.I.-11** towards proximal. Ask the patient to flex their biceps.

Needling

Vertically 0.5–1 cun, alongside the anterior border of the humerus.

Actions/Indications

- Opens the channel

L.I.-14 *Upper Arm* BINAO

Location

On the lateral aspect of the upper arm, on a line connecting → **L.I.-11** and → **L.I.-15**, 7 cun proximal to → **L.I.-11** and slightly superior to the pointed insertion of the deltoid muscle.

How to find

This point is best located with the elbow flexed and the deltoid muscle flexed against resistance, making the lateral border of the muscle belly more visible. **L.I.-14** is located in a depression on the lower pointed insertion area of the deltoid. 'By moving the upper arm towards the tip of the nose, the patient will touch the approximate location of this point.' **For orientation:** This point lies approximately 2 cun inferior to the anterior end of the axillary fold (→ 2.2).

Needling

Vertically 0.5–1 cun; also obliquely towards the shoulder for disorders of the eyes.

Actions/Indications

- Opens the channel and *luo* vessels, alleviates pain
- Benefits the eyes

Special features

Important local point. According to some authors, meeting point with the *yang wei mai* and the S.I. and BL channels.

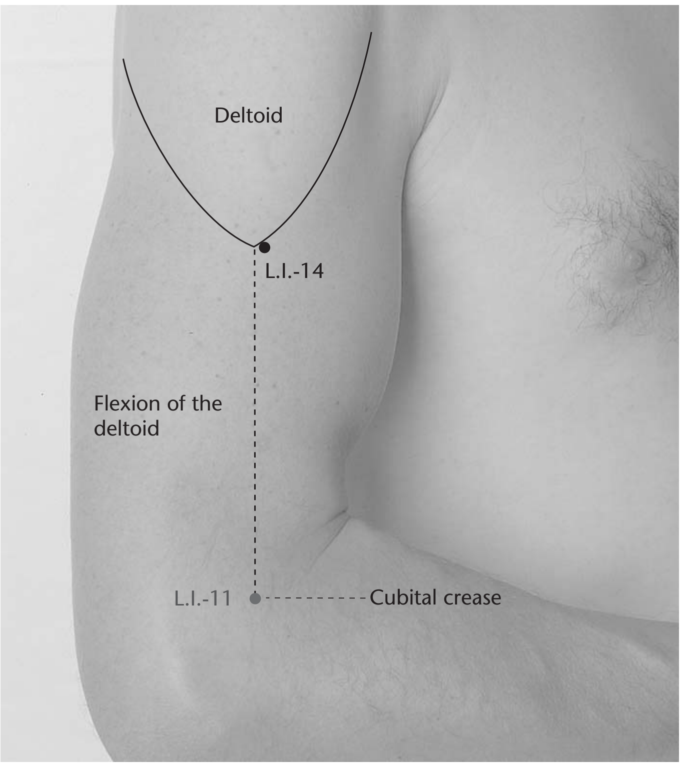
Deltoid
L.I.-14
Flexion of the deltoid
L.I.-11
Cubital crease

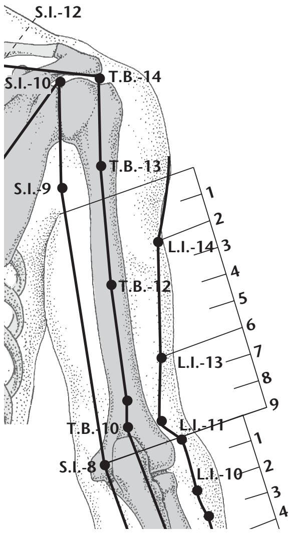
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
7
L.I.-13
8
9
T.B.-10
L.I.-11
S.I.-8
1
L.I.-10
2
3
4

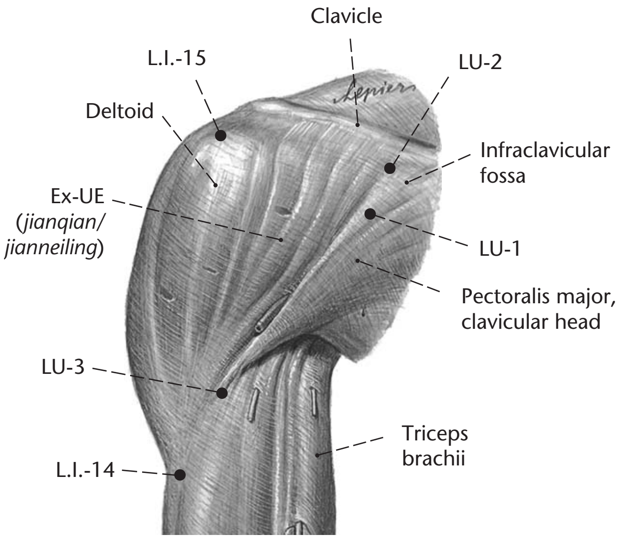
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
Ex-UE (jianqian/ jianneiling)
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14

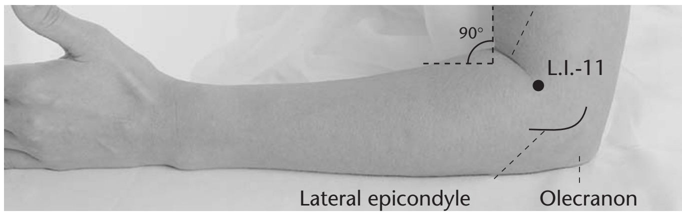
90°
L.I.-11
Lateral epicondyle
Olecranon

Shoulder Bone JIANYU L.I.-15

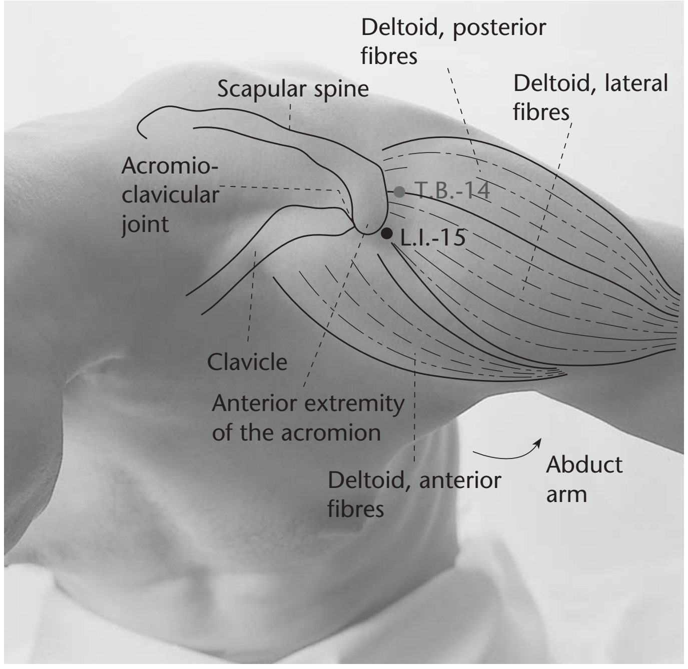
Deltoid, posterior fibres
Scapular spine
Deltoid, lateral fibres
Acromio-clavicular joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity of the acromion
Deltoid, anterior fibres
Abduct arm

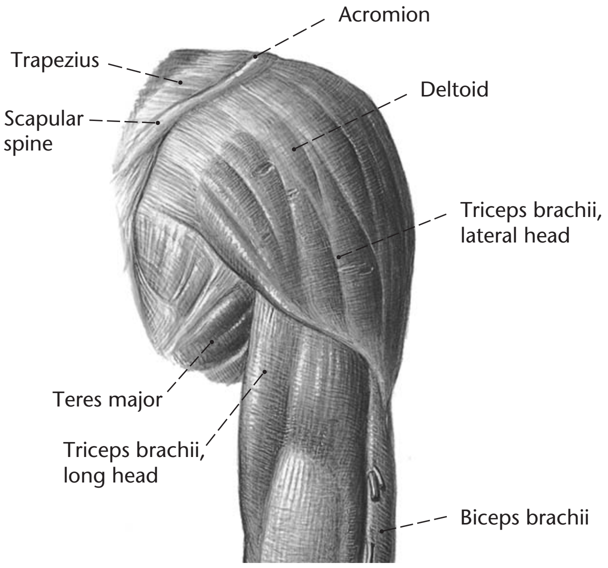
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

Location

In the depression distal and anterior to the acromion, between the clavicular and acromial portions of the deltoid muscle.

How to find

Ask the patient to abduct their arm. With the arm in a horizontal position, two depressions will form in the insertion area of the deltoid, distally to the acromion. By placing the thumb and index finger in those depressions in a pincer-like grip (with thumb and finger 1 thumb's breadth apart), the anterior finger will be in the anterior depression, where **L.I.-15** is located.
→ **T.B.-14** is located under the posterior finger in the dorsal depression.

Needling

With the arm slightly adducted, 1–1.5 cun obliquely in the bony cleft and in a posterior direction; with the arm abducted, 0.5–1 cun vertically. Caution: Shoulder joint.

Actions/Indications

- Expels Wind-Damp, opens the channel, alleviates pain, supports the shoulder joint
- Expels Wind
- Regulates the Qi, resolves Phlegm

Special features

Meeting point with the *yang qiao mai*. Important local point for disorders of the shoulder, often used with chain-and-lock method (→ 8.3.6) with other channel points of the upper extremity.

L.I.-16 *Great Bone* JUGU

Location

In a depression between the acromial extremity of the clavicle and the junction of the scapular spine and the acromion.

How to find

This point is located on the shoulder, at the angle between the acromio-clavicular joint and the junction of the scapular spine and the acromion (→ 3.3.1), above the most lateral portion of the trapezius and supraspinatus muscles. At this point, the tendon of the supraspinatus delves below the acromion, where it often causes problems owing to the cramped anatomical structure (for example, impingement syndrome).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Opens the channel, supports the shoulder joint

Special features

Meeting point with the *yang qiao mai*

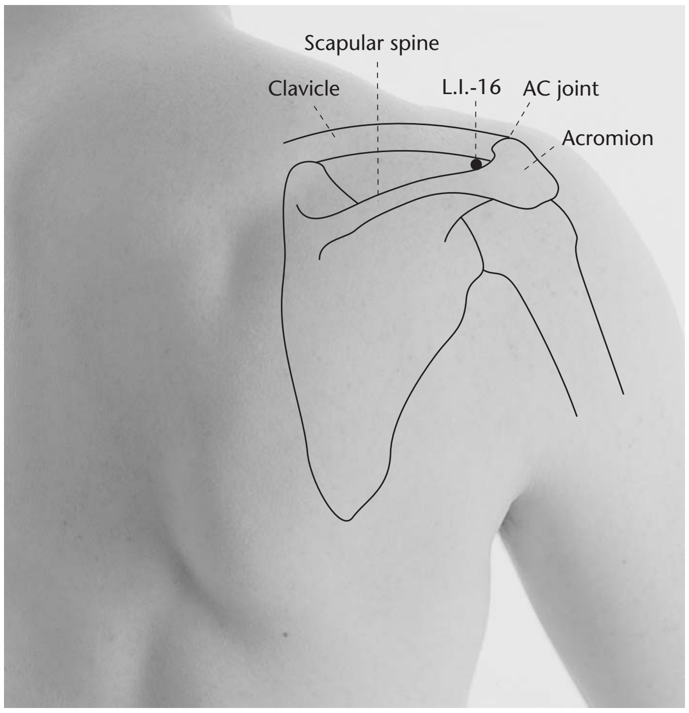
Scapular spine
Clavicle
L.I.-16
AC joint
Acromion

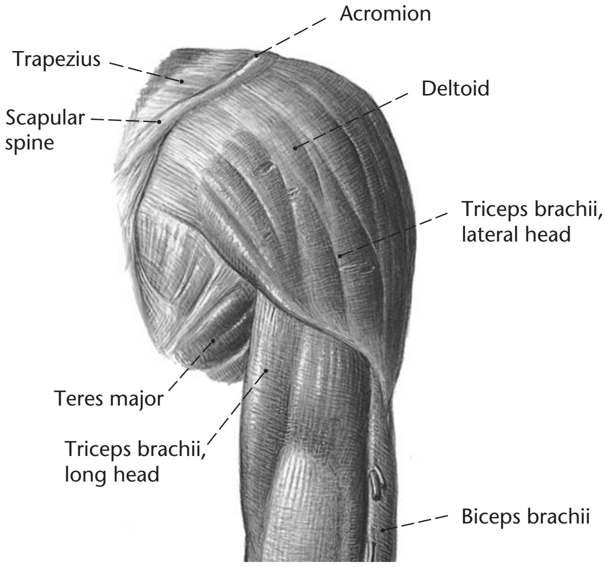
Acromion
Trapezius
Deltoid
Scapular spine
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

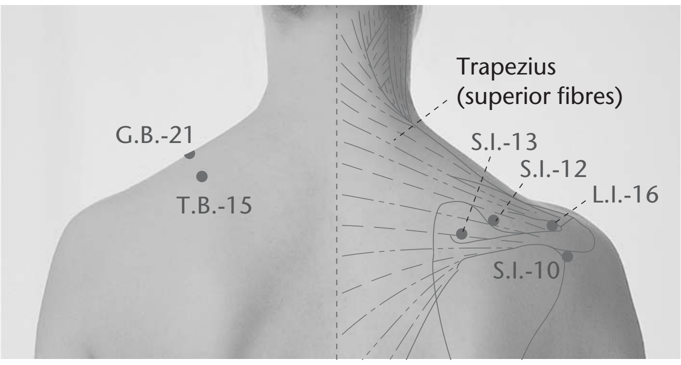
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
T.B.-15
L.I.-16
S.I.-10

Heaven's Tripod TIANDING L.I.-17

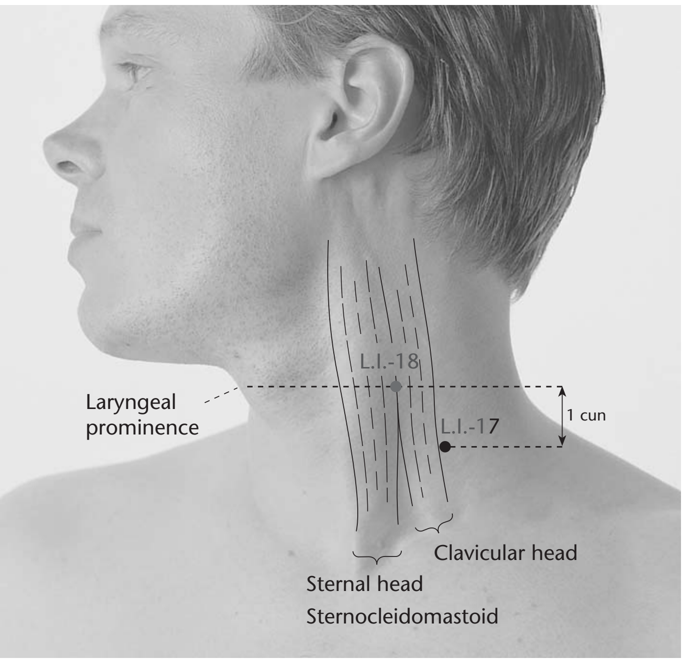
L.I.-18
Laryngeal prominence
1 cun
L.I.-17
Clavicular head
Sternal head
Sternocleidomastoid

Location

On the posterior border of the sternocleidomastoid muscle, 1 cun below the laryngeal prominence (→ 3.2).

How to find

This point is located on the lateral musculature of the neck, directly posterior to the sternocleidomastoid muscle. By rotating the patient's head to the opposite side, this muscle will become more visible and palpable. → **L.I.-18** is located 1 cun superior to **L.I.-17**, between the two heads of the sternocleidomastoid.

Needling

Rotate the head back to its normal position before needle insertion. Vertically 0.3–0.5 cun or obliquely up to 0.8 cun. Caution: Carotid artery, jugular vein.

Actions/Indications

- Benefits the throat and larynx

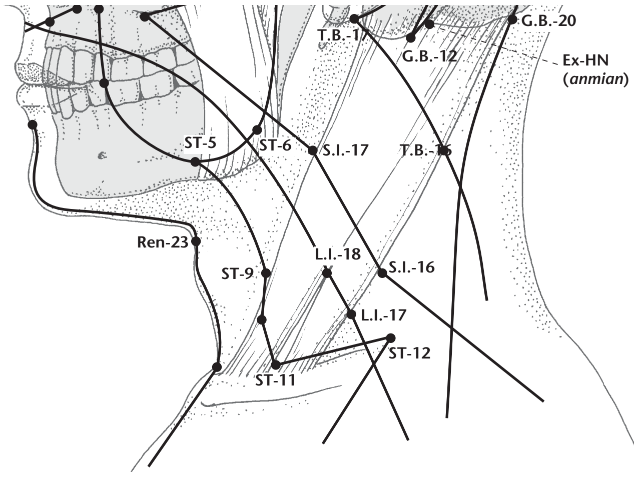
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

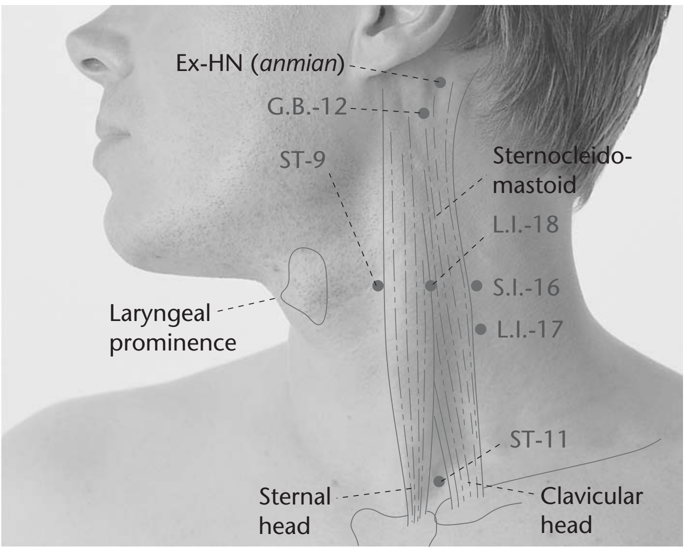
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleido-mastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal prominence
ST-11
Sternal head
Clavicular head

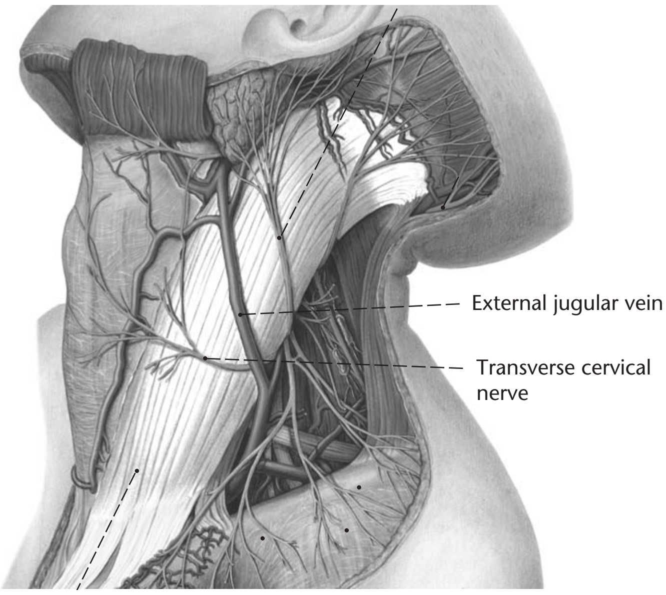
External jugular vein
Transverse cervical nerve

L.I.-18 *Support the Prominence* FUTU

Location

On the lateral aspect of the neck, at the level of the laryngeal prominence, between the two heads of the sternocleidomastoid muscle (→ 3.2).

How to find

From the laryngeal prominence (tip of the Adam's apple), draw a line across the sternocleidomastoid muscle to its posterior border. **L.I.-18** is located between the two heads of this muscle, which becomes more visible and palpable by rotating the head to the opposite side.

Located on the same level are → **ST-9** on the anterior border and → **S.I.-16** on the posterior border of the sternocleidomastoid muscle.

Needling

Rotate the head back to its normal position before inserting the needle. Vertically 0.3–0.5 cun or obliquely up to 0.8 cun. Caution: Carotid artery, jugular vein.

Actions/Indications

- Benefits the throat and larynx

Special features

Window of Heaven point. Often used when swallowing is difficult, for example after a stroke.

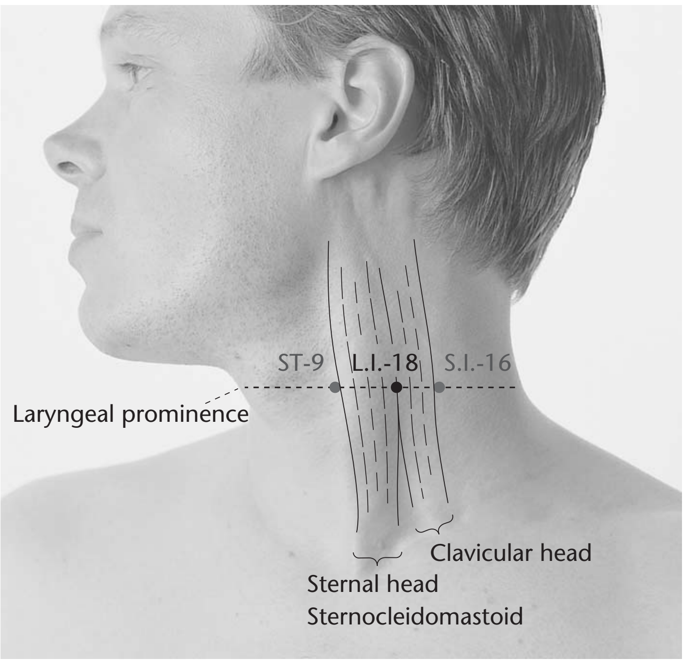
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

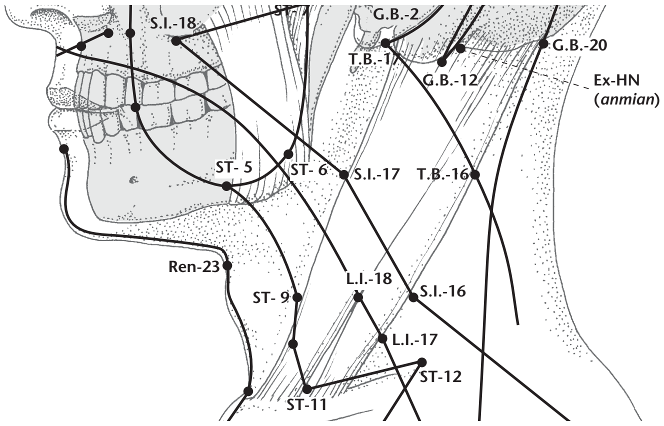
S.I.-18
ST-7
G.B.-2
T.B.-17
G.B.-20
G.B.-12
Ex-HN (*anmian*)
ST- 5
ST- 6
S.I.-17
T.B.-16
Ren-23
ST- 9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

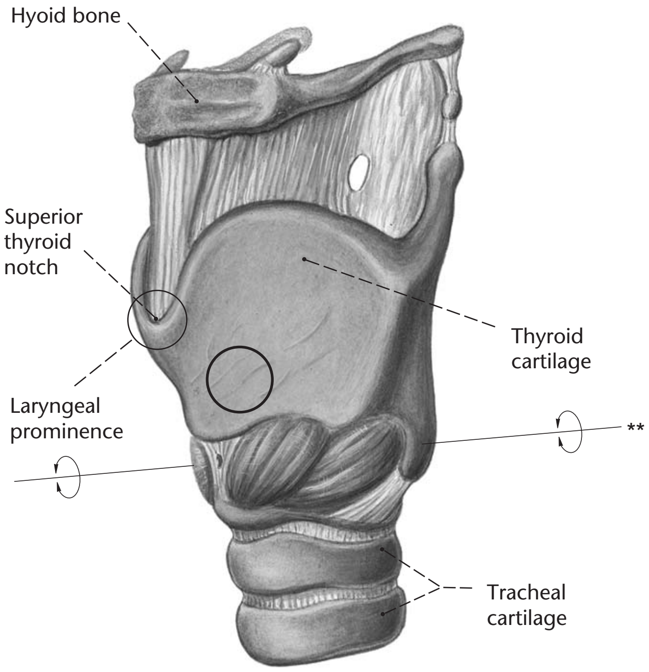
Hyoid bone
Superior thyroid notch
Thyroid cartilage
Laryngeal prominence
**
Tracheal cartilage

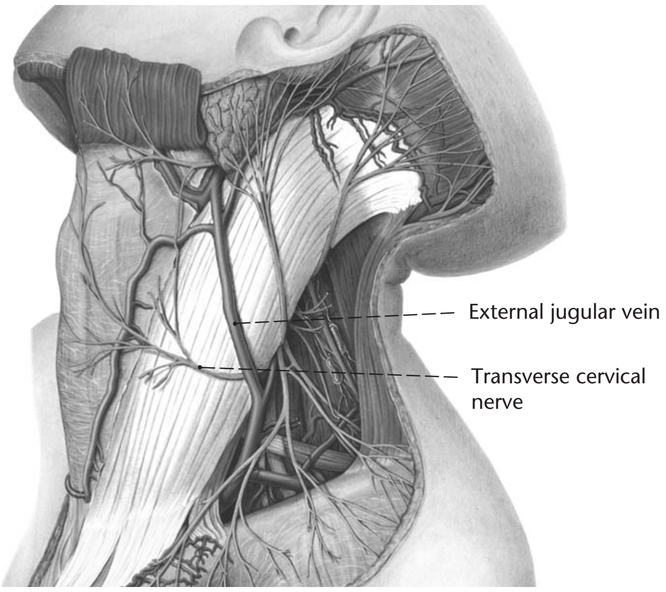
External jugular vein
Transverse cervical nerve

Mouth Grain Crevice KOUHELIAO L.I.-19

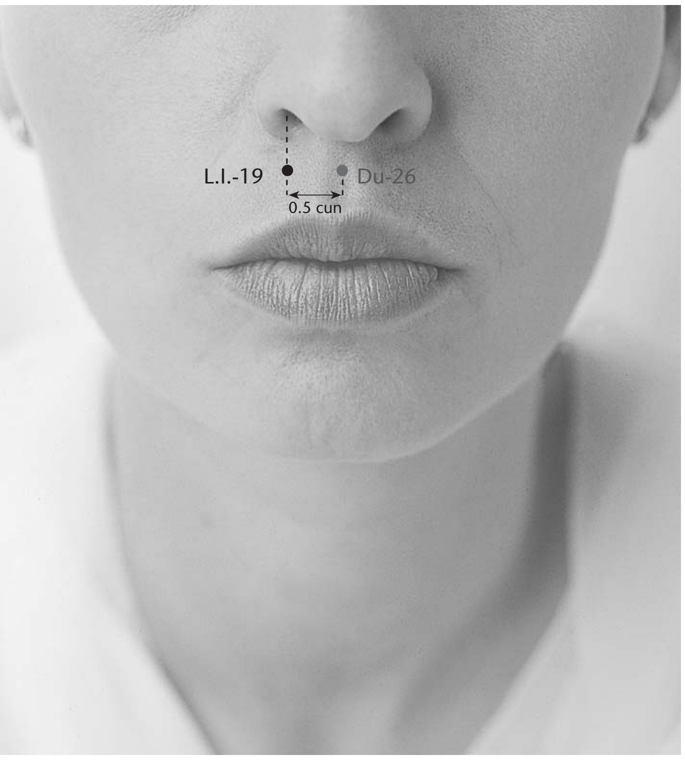
L.I.-19
Du-26
0.5 cun

Location

On the maxilla, slightly below the lateral margin of the nostril.

How to find

By dividing the distance between the margin of the nostril and the margin of the upper lip into three sections, **L.I.-19** is located at the junction of the upper and the middle third.

→ **Du-26** is located on the same level, on the anterior midline and at a distance of approximately 0.5 cun.

Needling

Obliquely 0.3–0.5 cun. Painful point! According to some classics, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and opens the nasal passages

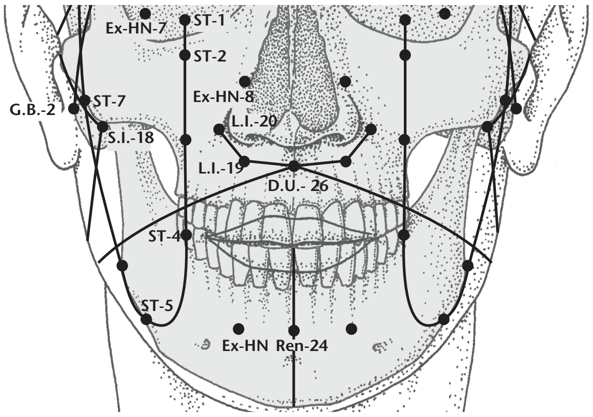
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
D.U.- 26
ST-4
ST-5
Ex-HN
Ren-24

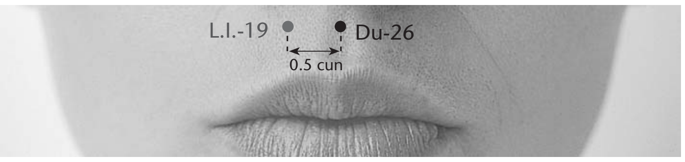
L.I.-19
Du-26
0.5 cun

L.I.-20 *Welcome Fragrance* YINGXIANG

Location

In the nasolabial groove, on the level of the midpoint of the lateral border of the ala nasi.

How to find

L.I.-20 is located by finding the midpoint of the lateral border of the ala nasi (→ 3.1.2) and following it to the nasolabial groove. Tip: The nasolabial groove becomes more pronounced if you ask the patient to smile.

Needling

0.3–0.5 cun vertically, obliquely or transversely (subcutaneously) in a medial and superior direction towards the extra point → **EX-HN-8** (*shangyingxiang* or *bitong*: at the upper end of the nasolabial groove, at the transition between the nasal bone and the cartilage). According to some authors, moxibustion is contraindicated.

Actions/Indications

- Opens the nasal passages, expels Wind, clears Heat

Special features

Meeting point with the ST channel, exit point. Most important local point for disorders of the nose.

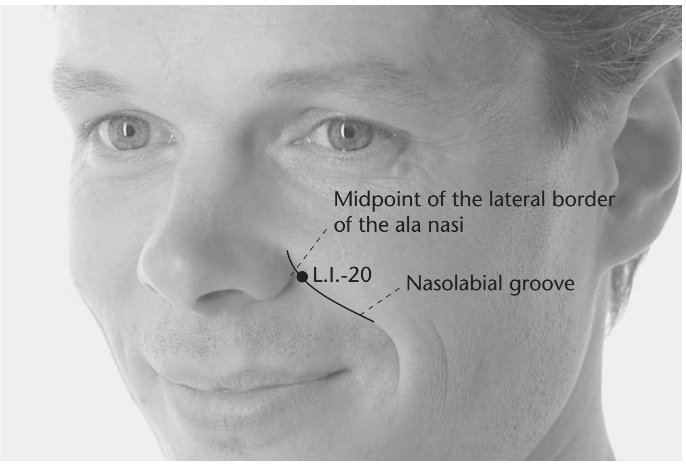
Midpoint of the lateral border of the ala nasi
L.I.-20
Nasolabial groove

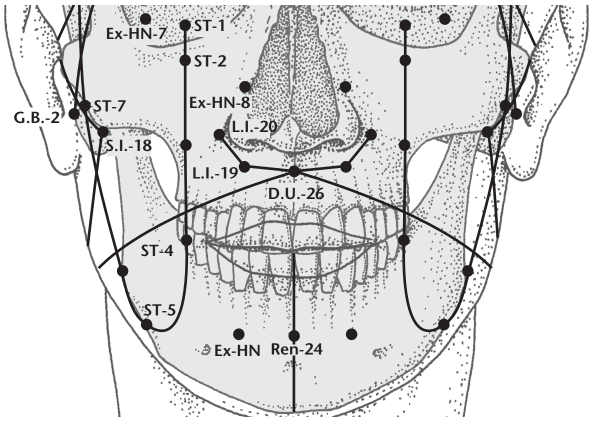
Ex-HN-7
ST-1
ST-2
G.B.-2
ST-7
Ex-HN-8
L.I.-20
S.I.-18
L.I.-19
D.U.-26
ST-4
ST-5
Ex-HN
Ren-24

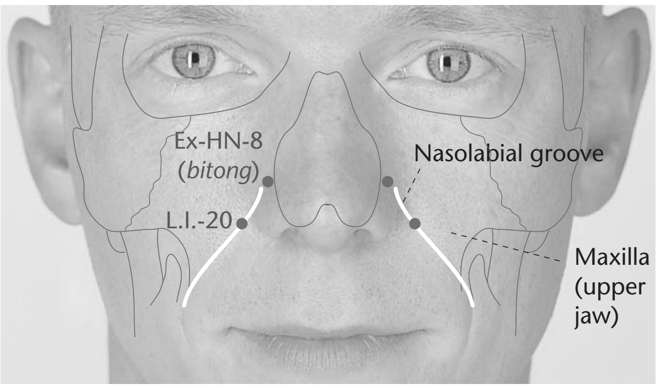
Ex-HN-8
(*bitong*)
L.I.-20
Nasolabial groove
Maxilla
(upper
jaw)

4.3 The Stomach Channel System – Foot *yangming* (*zu yang ming jing luo*)

4.3.1 The Stomach Primary Channel (*zu yang ming jing*)

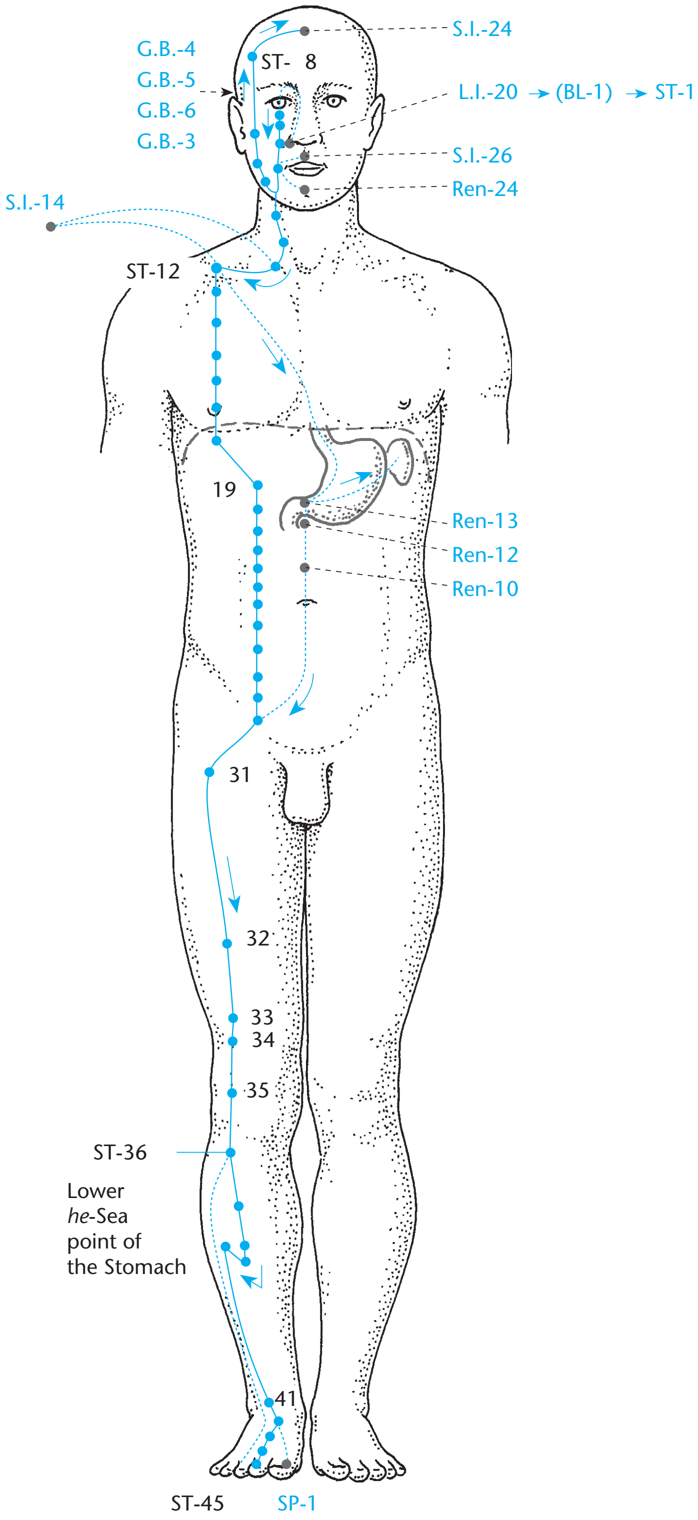
G.B.-4
G.B.-5
G.B.-6
G.B.-3
ST- 8
S.I.-24
L.I.-20 → (BL-1) → ST-1
S.I.-26
Ren-24
S.I.-14
ST-12
19
Ren-13
Ren-12
Ren-10
31
32
33
34
35
ST-36
Lower
he-Sea
point of
the Stomach
41
ST-45
SP-1

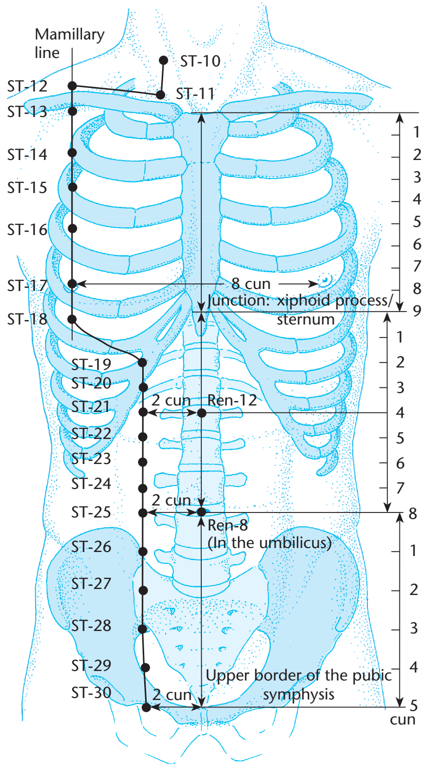
Mamillary
line
ST-10
ST-11
ST-12
ST-13
ST-14
ST-15
ST-16
ST-17
ST-18
8 cun
Junction: xiphoid process/
sternum
1
2
3
4
5
6
7
8
9
ST-19
ST-20
ST-21
ST-22
ST-23
ST-24
ST-25
2 cun
Ren-12
2 cun
Ren-8
(In the umbilicus)
ST-26
ST-27
ST-28
ST-29
ST-30
2 cun
Upper border of the pubic
symphysis
1
2
3
4
5
cun

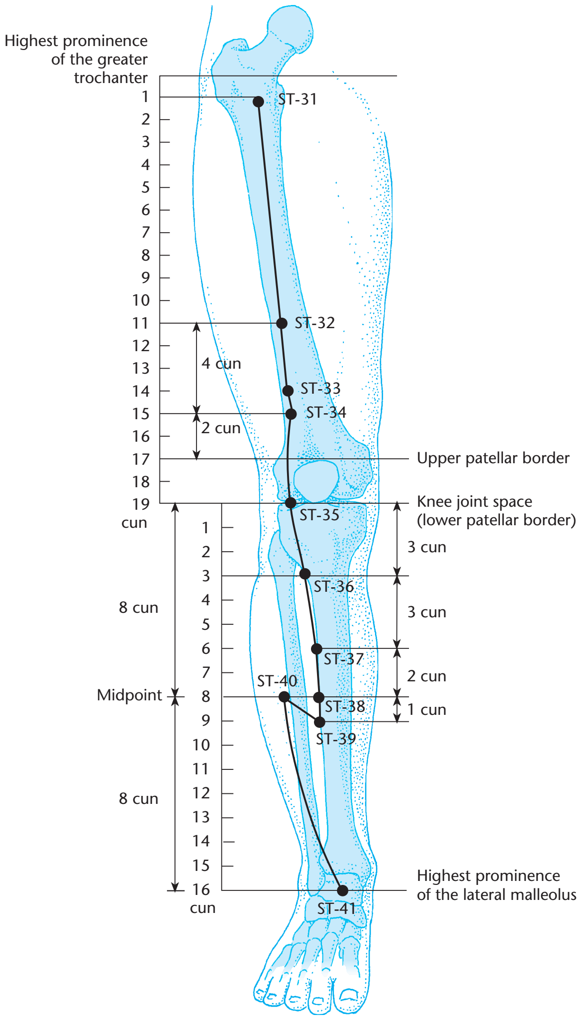
Highest prominence
of the greater
trochanter
ST-31
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
ST-32
4 cun
ST-33
ST-34
2 cun
Upper patellar border
Knee joint space
(lower patellar border)
ST-35
3 cun
ST-36
3 cun
8 cun
ST-37
2 cun
ST-40
ST-38
1 cun
Midpoint
ST-39
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
8 cun
Highest prominence
of the lateral malleolus
ST-41

Ch04.3-F10028.qxd 2/22/08 8:41 PM Page 122

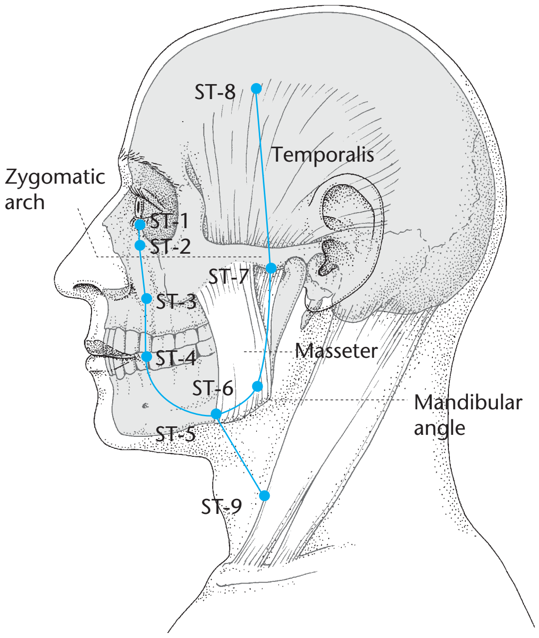

Pathway

The Stomach primary channel begins with an **internal** branch at the Yang axes connection of the first great circuit (*yangming*, hand–foot pairing) at → **LI-20** (*yingxiang*) lateral to the ala nasi

- travels **internally** to → **BL-1** (*jingming*) at the inner canthus of the eye
- emerges at the infraorbital ridge at **ST-1** (*chengqi*), the beginning of the external part of the primary channel
- descends laterally along the nose and **enters** the gums of the upper jaw
- meets the contralateral Stomach primary channel and the *du mai* at → **Du-26** (*renzhong*) and emerges again
- curves around the lips to meet the *ren mai* and again the contralateral Stomach channel at → **Ren-24** (*chengjiang*)
- traverses the lower cheek and winds around the lower angle of the mandible.

Here, at → **ST-5** (*daying*), the external channel **divides into two branches**:

- **One branch** ascends anterior to the ear, travels via → **G.B.-3** (*shangguan*), → **G.B.-6** (*xuanli*), → **G.B.-5** (*xuanlu*) and → **G.B.-4** (*hanyan*) to the temporal region to → **ST-8** (*touwei*) and continues to the forehead to meet → **Du-24** (*shenting*), where it terminates.
- The **other external branch** descends along the lateral aspect of the neck to the supraclavicular fossa and travels posteriorly over the back of the neck to → **Du-14** (*dazhui*) below the spinous process of the 7th cervical vertebra, where it meets the other five Yang primary channels.

An internal branch descends from the supraclavicular fossa, passes the diaphragm, enters first its pertaining *fu*-Organ, the Stomach (*wei*) and then connects with its paired *zang*-Organ, the Spleen (*pi*). Minor branches of the **internal** channel intersect with → **Ren-13** (*shangwan*), → **Ren-12** (*zhongwan*) and → **Ren-10** (*xiawan*). The internal branch then descends towards the inguinal region where it emerges slightly superior to **ST-30** (*qichong*) and reconnects with the **external** branch.

- The **external** channel descends from the supraclavicular fossa, crosses the nipple and laterally passes the umbilicus, travels in a curve to the leg and knee, continues along the anterior crest of the tibia and the dorsum of the foot and terminates on the second toe at → **ST-45** (*lidui*) at the lateral corner of the nail.
- An **internal** branch originates below the knee area at → **ST-36** (*zusanli*) and travels to the lateral aspect of the middle toe.
- A further branch separates from the primary channel at → **ST-42** (*chongyang*) on the dorsum of the foot and connects at → **SP-1** (*yinbai*) with its paired SP channel on the medial aspect of the big toe (foot Yin–Yang connection of the first great circuit).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: High fever, malaria, flushed face, sweating, confused mental state, aversion to cold, eye pain, dry nasal passages, nosebleeds, dry lips and mouth, lesions of the lips and mouth, sore and swollen throat, thoracic pain, red and swollen legs

Interior (*li*) or *zangfu*-Organ signs and symptoms: Abdominal distension, sensation of fullness, oedema, irritability during work and rest, mania and epilepsy, hyperpepsinia, constant hunger, yellow urine

Excess (*shi*): Heat sensations on the anterior aspect of the body, constant hunger, yellow urine

Deficiency (*xu*): Cold sensations on the anterior aspect of the body, chills, Stomach Cold with distension and fullness

Connections and meeting points of the Stomach primary channel (→ 1.2)

Connections with other channels

Spleen primary channel (*zu tai yin jing*)

Connection: Foot Yin–Yang connection of the first great circuit
Location: ST-42 → SP-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Large Intestine primary channel (*shou yang ming jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *yang ming* (Yang axes of the first great circuit)
Location: L.I.-20 → (BL-1) → ST-1 (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Stomach (*wei*), Spleen (*pi*)

4.3.2 The Stomach Divergent Channel (*zu yang ming jing bie*)

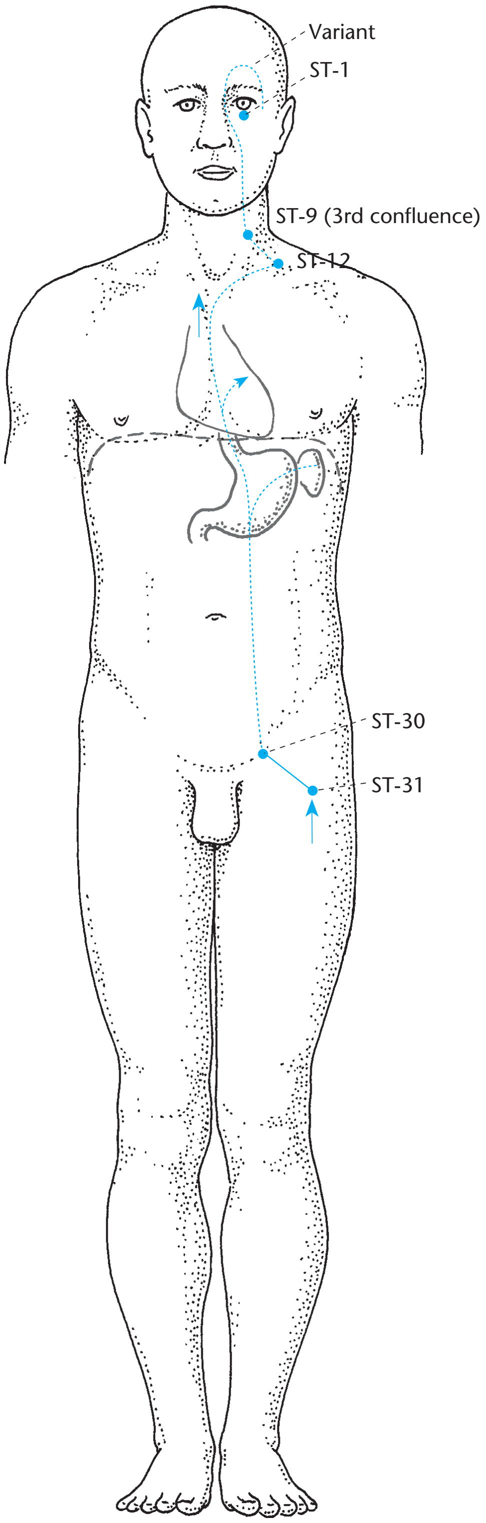

Pathway

The Stomach divergent Channel separates from the Stomach primary channel at ➞ **ST-31** (*biguan*) on the antero-lateral aspect of the thigh

- ➡ penetrates the body at the inguinal region at **ST-30** (*qichong*)
- ➡ continues to the Stomach (*wei*) and disperses in the Spleen (*pi*)
- ➡ ascends and penetrates the Heart (*xin*)
- ➡ ascends along the oesophagus to the suprasternal fossa and travels laterally to the supraclavicular fossa at **ST-12** (*quepen*).

From there it rises along the anterior border of the sternocleidomastoid muscle and meets its primary channel as well as the divergent Spleen channel at **ST-9** (*renying*) at one of the six *he*-confluences (here: ST/SP as third confluence ➞ 1.3).

It then **emerges** at the mouth, ascends along the nose, reaches the inner canthus of the eye and connects with the eye, circles the eye region and meets the Stomach primary channel at **ST-1** (*chengqi*) where it terminates (➞ fig.).

According to some authors, the Stomach divergent channel connects with the Stomach primary channel at **BL-1** (➞ variant, fig.). After intersecting with **BL-1**, the divergent channel could enter the head, disperse in the brain and re-emerge at **ST-1.**

Clinical importance

- Strengthens the relationship between the Stomach and the Spleen (*zangfu*-Organs). Points on the Stomach primary channel can therefore be used for disorders of the Spleen, and vice versa points on the Spleen primary channel can treat disorders of the Stomach.
- Spreads Qi to the face and sensory organs: many points on the ST channel treat disorders of the head and face.
- Strengthens the relationship between the Stomach and the eyes: Heat and excess in this region can be directed downward by using points on the ST channel.

4.3.3 The Stomach Sinew Channel (*zu yang ming jing jin*)

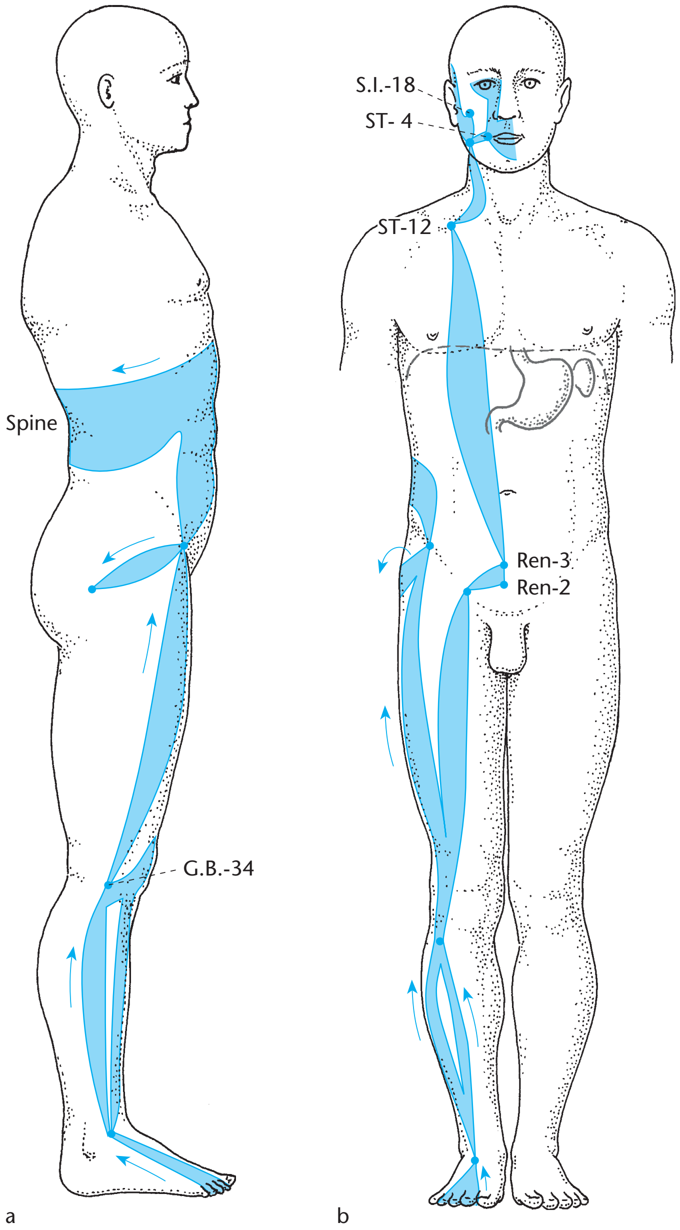

Pathway

The Stomach sinew channel begins as a wider band, encompassing the second, third and fourth toe, traverses the dorsum of the foot, binds (*jie*) at the tibial depression and divides into two branches:

- **One branch** obliquely ascends the **anterolateral aspect** of the leg and binds (*jie*) at the lateral aspect of the knee. It then ascends the anterolateral aspect of the thigh and binds (*jie*) at the anterior hip region. From there it sends **a branch** to → **G.B.-30** (*huantiao*) in the direction of the greater trochanter. The sinew channel travels from the anterior hip region across the lateral abdomen and spreads across the lower ribcage posteriorly to the spine.
- **Another branch** follows the tibia on the **anterior aspect** of the lower leg and binds (*jie*) at the level of the patella. From there a branch runs to the head of the fibula where it meets the G.B. sinew channel at → **G.B.-34** (*yanglingquan*).
- From the patella the channel ascends along the anterior aspect of the thigh, binds (*jie*) in the inguinal region, passes → **Ren-2** (*qugu*) and binds (*jie*) at → **Ren-3** (*zhongji*). It then ascends across the abdominal and thoracic region and reaches the supraclavicular fossa to bind (*jie*) at **ST-12** (*quepen*). It then continues along the anterolateral aspect of the neck and binds (*jie*) at the angle of the mandible.

At the angle of the mandible the channel divides into three branches:

- **One branch** terminates anterior to the ear.
- **A further branch** ascends to the zygomatic bone where it meets the other Yang sinew channels at → **S.I.-18** (*quanliao*).
- **The third branch** circles the mouth, ascends to the upper nose region, there connects with the Bladder sinew channel and then spreads along the lower eyelid. The Bladder sinew channel spreads along the upper eyelid, so that the two channels together form a network around the eyes.

Clinical importance

Pathology: Stiffness and aching of the toes, leg cramps (gastrocnemius and quadriceps muscles), stiffness and pain on the dorsum of the foot (at **ST-41**), swelling and tension in the inguinal region, *shan*-disorders, cramping in the abdomen as well as in the supraclavicular fossa and the face, facial paralysis, weakness and paralysis of the superior rectus muscle.

Indication: Mainly used for *bi*-syndromes (painful obstruction syndromes) along the Stomach channel. The area covered by the Stomach sinew channel is larger than that covered by the Stomach primary channel. This explains why the indications of points on the Stomach primary channel include disorders and diseases of the external genitalia (for example urinary tract disorders such as cystitis, hernia and orchitis, which are often treated in combination with LIV points) as well as disorders of the eyelids and of all sensory organs.

4.3.4 The Stomach *luo*-Connecting Vessel System (*zu yang ming luo mai*)

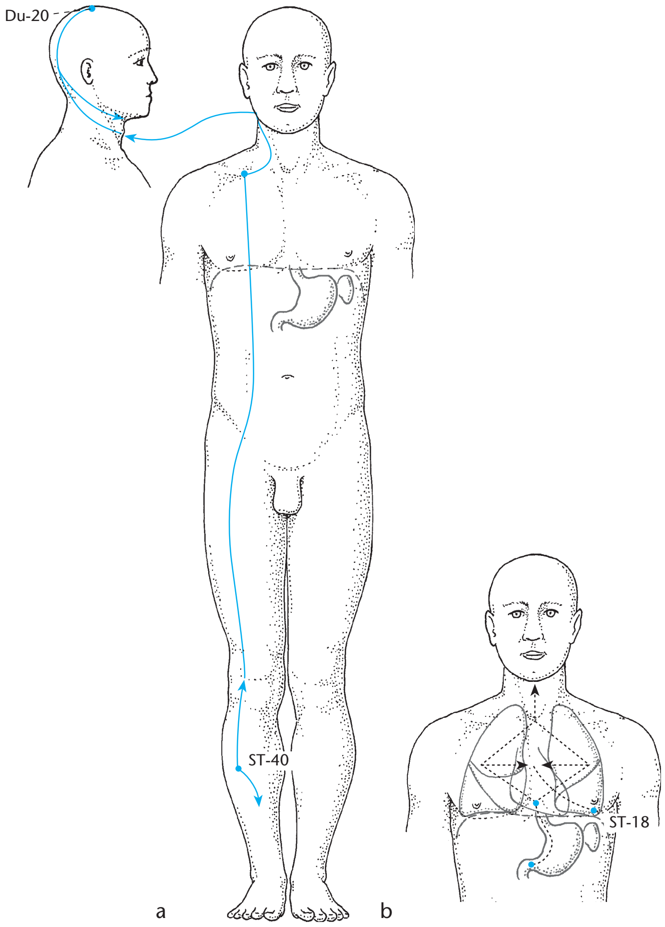

Pathway

The ST *luo*-connecting vessel separates from the Stomach primary channel at its *luo*-connecting point **ST-40** (*fenglong*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches *(sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions run to the Interiorly–Exteriorly paired primary SP channel; according to some schools of thought (for example Van Nghi, → Appendix) they travel as a **transverse** Stomach *luo*-connecting vessel to the *yuan* point → **SP-3** (*taibai*).

➡ A **longitudinal** division ascends along the anterolateral aspect of the leg to the torso and reaches the lateral aspect of the neck at → **ST-12**, where it divides into two branches: **one branch** traverses the neck, the **other branch** runs to the occiput and ascends the head to reach → **Du-20** (*baihui*).

Clinical importance (→ 8.1.2)

Pathology

- **Counterflow Qi:** Swelling and pain of the throat, sudden loss of voice, sensation of tightness in the throat
- **Excess (*shi*):** Psychological disorders, also epilepsy and mania
- **Deficiency (*xu*):** *Wei*-syndrome of the legs (atrophy syndrome with muscle weakness and paralysis)

The great *luo*-connecting vessel of the Stomach (*wei zhi dao luo* or *xu li*)

Pathway

The great Stomach *luo*-connecting vessel begins at the Stomach (*wei*), passes through the diaphragm, intersects with → **Ren-17** (*shanzhong*) and spreads in the Lungs (*fei*), trachea and larynx. From the Lung it travels to the Heart (*xin*) and emerges on the left aspect of the thorax near → **ST-18** (*rugen*) which is located where the heartbeat is visible (→ 8.1.2).

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Dyspnoea
- **Deficiency (*xu*):** Chest tightness, for example with asthma, coughing, angina pectoris, etc.

4.3.5 Cutaneous Region (*yang ming pi bu*)

See description and figures → 1.6

4.3.6 Points of the Stomach Primary Channel (Overview)

Specific points according to their function (→ 8.1)

- ***Yuan*-source point (→ 8.1.1): ST-42** (*chongyang*) 🟦■
- ***Luo*-connecting point (→ 8.1.2.): ST-40** (*fenglong*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): ST-34** (*liangqiu*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-21** (*weishu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): Ren-12** (*zhongwan*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal), sedation point: **ST-45** (*lidui*) 🟦■
  ying-spring point (Water): **ST-44** (*neiting*) 🟦■
  shu-stream point (Wood): **ST-43** (*xiangu*) 🟦■
  jing-river point (Fire), tonification point: **ST-41** (*jiexi*) 🟦■
  he-sea point (Earth), *ben* point: **ST-36** (*zusanli*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea points (→ 8.1.9):**
  of the Stomach: **ST-36** (*zusanli*) 🟦🟦
  of the Large Intestine: **ST-37** (*shangjuxu*) 🟦■
  of the Small Intestine: **ST-39** (*xiajushu*) 🟦■
- ***jiaohui*-meeting points (→ 8.1.10):**

  - with the *ren mai, yang qiao mai, du mai**: **ST-1** (*chengqi*) 🟦■
  - with the *yang qiao mai**: **ST-2** (*sibai*) 🟦■
  - with the *yang qiao mai*: **ST-3** (*juliao*)
  - with the primary L.I. channel, *yang qiao mai, du mai*, ren mai*:* **ST-4** (*dicang*)
  - with the G.B. channel*: **ST-5** (*daying*), **ST-6** (*jiache*)
  - with the G.B. channel: **ST-7** (*xiaguan*) 🟦■, **ST-9** (*renying*)
  - with the G.B. channel, *yang wei mai*: **ST-8** (*touwei*) 🟦■
  - with the primary L.I., S.I., T.B., G.B. channels: **ST-12** (*quepen*)
  - with the *chong mai*, G.B. channel*: **ST-30** (*qichong*) 🟦■
  - other channels with the ST channel: **L.I.-20, BL-1, G.B.-3, G.B.-4, G.B.-5, G.B.-6; Du-14, Du-24; Du-26, Du-28*; Ren-10*, Ren-12, Ren-13, Ren-24; G.B.-14*, G.B.-21***
- **Gao Wu command point** (→ **8.1.11**) **for the abdomen: ST-36** (*zusanli*) ■■
- **Window of Heaven point** (→ **8.1.12**)**: ST-9** (*renying*)
- **Points of the Four Seas** (→ **8.1.13**)**:**
  Sea of Qi point: **ST-9** (*renying*)
  Sea of Grain: **ST-30** (*qichong*) 🟦■, **ST-36** (*zusanli*) 🟦🟦
  Sea of Blood: **ST-37** (*shangjuxu*) 🟦■, **ST-39** (*xiajushu*) 🟦■
- **Ma Dan Yang Heavenly Star points** (→ **8.1.14**)**: ST-36** (*zusanli*) 🟦🟦, **ST-44** (*neiting*) 🟦🟦
- **Sun Si Miao Ghost point** (→ **8.1.15**)**: ST-6** (*jiache*) 🟦■
- **Other functional points:**
  - Front-*mu* point of the Large Intestine: **ST-25** (*tianshu*) 🟦🟦
  - most important acupuncture 'Phlegm'-point: **ST-40** (*fenglong*) 🟦🟦
- Points according to region
- **Local points** (→ **8.2.1**)**:** eyes – **ST-1** (*chengqi*) 🟦■; mouth and teeth – **ST-4** (*dicang*), **ST-6** (*jiache*) 🟦■; jaw – **ST-7** (*xiaguan*) 🟦■; Large Intestine – **ST-25** (*tianshu*) 🟦🟦; Small Intestine and Bladder – **ST-28** (*shuidao*); urogenital region – **ST-30** (*qichong*) 🟦■; knee – **ST-34** (*liangqiu*) 🟦■, **ST-35** (*dubi*) 🟦■, **ST-36** (*zusanli*) 🟦🟦; foot – **ST-41** (*jiexi*) 🟦■
- **Adjacent points** (→ **8.2.1**)**:** forehead – **ST-8** (*touwei*) 🟦■; eyes – **ST-2** (*sibai*) 🟦■; mouth/teeth – **ST-7** (*xiaguan*) 🟦■; throat – **ST-10** (*shuitu*); Spleen, Stomach, Gall Bladder – **ST-21** (*liangmen*); Kidney – **ST-29** (*guilai*); abdominal region – **ST-25** (*tianshu*) 🟦🟦, **ST-26** (*wailing*), **ST-30** (*qichong*) 🟦■; foot – **ST-34** (*liangqiu*) 🟦■, **ST-36** (*zusanli*) 🟦🟦; toes – **ST-41** (*jiexi*) 🟦■
- **Distal points** (→ **8.2.1**)**:** frontal headaches – **ST-44** (*neiting*) 🟦■; nose – **ST-44** (*neiting*) 🟦■; mouth/teeth – **ST-44** (*neiting*) 🟦■; throat – **ST-44** (*neiting*) 🟦■; shoulder region – **ST-38** (*tiakou*) 🟦■; Lung – **ST-40** (*fenglong*) 🟦🟦; Spleen/Stomach – **ST-36** (*zusanli*) 🟦🟦; Large Intestine – **ST-37** (*shangjuxu*) 🟦■; Small Intestine – **ST-39** (*xiajushu*) 🟦■; epigastrium – **ST-36** (*zusanli*) 🟦🟦; shoulder – **ST-36** (*zusanli*) 🟦🟦, **ST-38** (*tiakou*) 🟦■

Specific points according to the channel pathway (in numerical order)

- **ST-1** (*chengqi*) 🟦■: meeting point with the *ren mai, yang qiao mai*, du mai* (→ 8.1.10); local point for the eyes (→ 8.2.1)
- **ST-2** (*sibai*) 🟦■: meeting point with the *yang qiao mai** (→ 8.1.10); adjacent point for the eyes (→ 8.2.1)
- **ST-3** (*juliao*): meeting point with the *yang qiao mai* (→ 8.1.10)
- **ST-4** (*dicang*): meeting point with the Large Intestine channel, *yang qiao mai, du mai*, ren mai** (→ 8.1.10); local point for the mouth/teeth (→ 8.2.1)
- **ST-5** (*daying*): meeting point with the Gall Bladder channel
- **ST-6** (*jiache*) 🟦■*:* Sun Si Miao Ghost point (→ 8.1.15); meeting point with the Gall Bladder channel* (→ 8.1.10); local point for the mouth/teeth (→ 8.2.1)
- **ST-7** (*xiaguan*) 🟦■: meeting point with the Gall Bladder channel (→ 8.1.10); adjacent point for the teeth and jaw (8.2.1); local point for the jaw (→ 8.2.1)
- **ST-8** (*touwei*) 🟦■: meeting point with the Gall Bladder channel, *yang wei mai* (→ 8.1.10); adjacent point for frontal headaches (→ 8.2.1)
- **ST-9** (*renying*): meeting point with the Gall Bladder channel (→ 8.1.10); Sea of Qi point (→ 8.1.13); Window of Heaven point (→ 8.1.12)
- **ST-10** (*shuitu*): adjacent point for the neck (→ 8.2.1)
- **ST-12** (*quepen*): meeting point with the L.I., S.I., T.B. and G.B. channels (→ 8.1.10)
- **ST-21** (*liangmen*): adjacent point for the Spleen, Stomach and Gall Bladder (→ 8.2.1)
- **ST-25** (*tianshu*) 🟦🟦: Front-*mu* point of the Large Intestine (→ 8.1.5); adjacent point for the abdominal region (→ 8.2.1)
- **ST-26** (*wailing*): adjacent point for the abdominal region (→ 8.2.1)
- **ST-28** (*shuidao*): local point for the Small Intestine and Bladder (→ 8.2.1)
- **ST-29** (*guilai*): local point for the urogenital region (→ 8.2.1)
- **ST-30** (*qichong*) 🟦■: meeting point with the *chong mai*, G.B. channel* (→ 8.1.10); Sea of Water and Grain point (→ 8.1.13); local point for the urogenital region (8.2.1); adjacent point for the abdominal region (→ 8.2.1)
- **ST-34** (*liangqiu*) 🟦■: *xi*-cleft point, important adjacent point for disorders of the knee (→ 8.2.1)
- **ST-35** (*dubi*) 🟦■: local point for the knee region (→ 8.2.1)
- **ST-36** (*zusanli*) 🟦🟦: *he*-sea point (Earth) (→ 8.1.6), *ben*-point (Five Phase point); Lower *he*-sea point of the Stomach (→ 8.1.9); *Gao Wu* command point (→ 8.1.11) for the abdomen; Sea of Water and Grain point (→ 8.1.13); *Ma Dan Yang* Heavenly Star point (→ 8.1.14); distal point for the Spleen/Stomach and epigastrium (→ 8.2.1); local point for the knee (→ 8.2.1); distal point for the shoulder (→ 8.2.1); adjacent point for the foot (→ 8.2.1)

---

* Mentioned by only some authors.

- **ST-37** (*shangjuxu*) 🟦⬛: Lower *he*-sea point of the Large Intestine (→ 8.1.9); Sea of Blood point (→ 8.1.13); distal point for the Large Intestine (→ 8.2.1)
- **ST-38** (*tiakou*) 🟦⬛: distal point for the shoulder region (→ 8.2.1)
- **ST-39** (*xiajushu*) 🟦⬛: Lower *he*-sea point of the Small Intestine (→ 8.1.9); Sea of Blood point (→ 8.1.13); distal point for the Small Intestine (→ 8.2.1)
- **ST-40** (*fenglong*) 🟦⬛: *luo*-connecting point (→ 8.1.2); main 'Phlegm'-point; distal point for the Lung (→ 8.2.1)
- **ST-41** (*jiexi*) 🟦⬛: *jing*-river point (Fire) (→ 8.1.6); tonification point; local point for the foot (8.2.1); adjacent point for the toes (→ 8.2.1)
- **ST-42** (*chongyang*) 🟦⬛: *yuan*-source point (→ 8.1.1)
- **ST-43** (*xiangu*) 🟦⬛: *shu*-stream point (Wood) (→ 8.1.6)
- **ST-44** (*neiting*) 🟦🟦: *ying*-spring point (Water) (→ 8.1.6); distal point for frontal headaches, as well as for the nose, mouth/teeth and throat (→ 8.2.1); Ma Dan Yang Heavenly Star point (→ 8.1.14)
- **ST-45** (*lidui*) 🟦⬛: *jing*-well point (Metal) (→ 8.1.6); sedation point.

General location help

- **ST-1** to **ST-4**: with the patient looking straight ahead, these points are located on the vertical pupil line
- **ST12** to **ST-18** are all located on the midclavicular/mamillary line, 4 cun lateral to the anterior midline
  - **ST-12**: in the supraclavicular fossa
  - **ST-13**: on the lower border of the clavicle
  - **ST-14** to **ST-18**: in the first to fifth intercostal spaces
- **ST-19** to **ST-30** are all located on a vertical line midway between the mamillary line and the anterior midline, 2 cun lateral to the anterior midline:
  - **ST-19** to **ST-25** are located on the 8 cun section (→ 2.2) between the sternocostal angle and the umbilicus (**ST-19** to **ST-25** are evenly spread on this line, from 6 cun above the umbilicus to the same level as the umbilicus)
  - **ST-26** to **ST-30** are located on the 5 cun section (→ 2.2) between the umbilicus and the upper border of the pubic symphysis (1 cun steps from 4 cun superior to the upper border of the pubic symphysis to its upper border).

Container of Tears CHENGQI ST-1

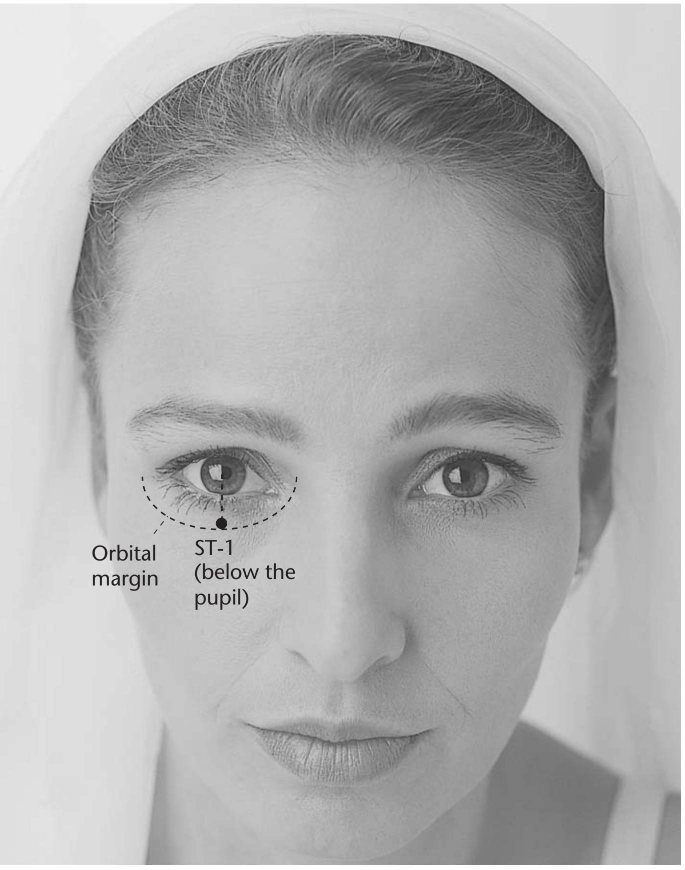
Orbital margin
ST-1 (below the pupil)

Location

With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil. Within the area of the lower eyelid, the infraorbital ridge presents itself as a distinct bony structure. **ST-1** is located on the pupil line, directly superior to the infraorbital ridge.

Needling

Ask the patient to look upward, gently push the eyeball upwards and insert the needle vertically along the orbital ridge in a dorsal direction. Caution: Venous plexus and arteries, avoid injuring the eyeball and periost! Needling recommended only by experienced practitioners! Pay attention to needling pain! No needle manipulation! After removing the needle, press firmly on the insertion site for about one minute to prevent haematoma. Inform the patient that it may still be possible for a haematoma to develop. Moxibustion is contraindicated! Points for disorders of the eyes with less risk of complications include **BL-2**, **T.B.-21**, **G.B.-1**, **ST-2**, **Ex-HN-5** (*taiyang*), **EX-HN-4** (*yuyao*).

Actions/Indications

- Expels Wind and Heat, clears and benefits the eyes

Special features

Meeting point with the *yang qiao mai* and *ren mai*, also with the *du mai* according to some authors, important point for the eyes, entry point.

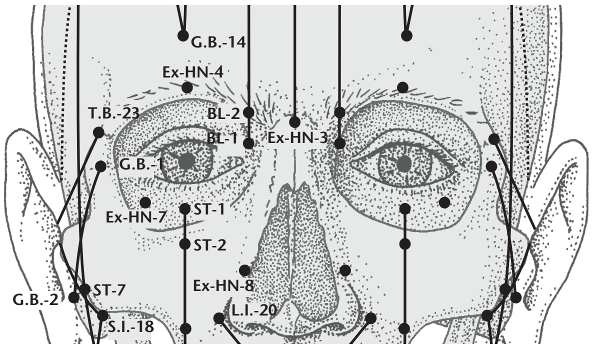
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18

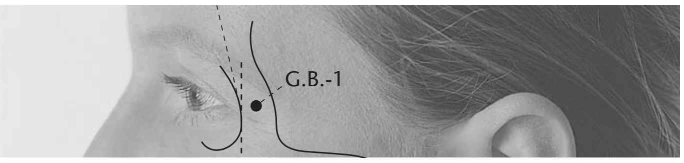

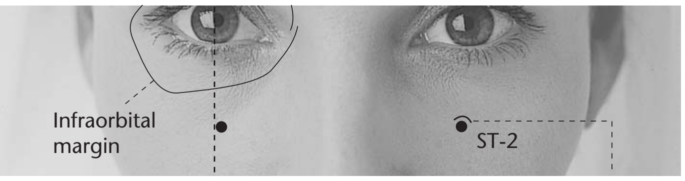
Infraorbital margin
ST-2

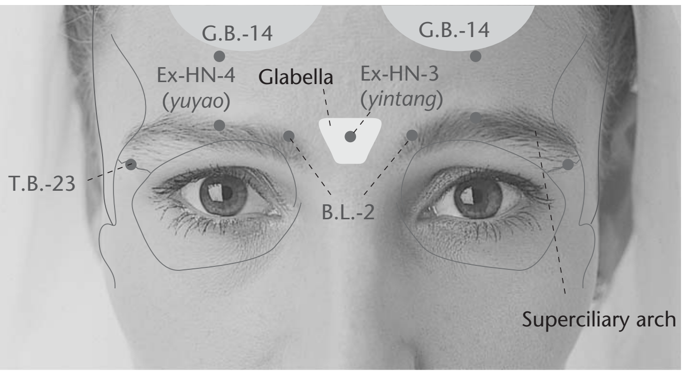
G.B.-14
G.B.-14
Ex-HN-4 (yuyao)
Glabella
Ex-HN-3 (yintang)
T.B.-23
B.L.-2
Superciliary arch

ST-2 *Four Whites* SIBAI

Location

With the eyes looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil. Within the area of the lower eyelid, the infraorbital ridge presents itself as a distinct bony structure. Starting at the infraorbital ridge (➞ 3.1.2), palpate inferiorly until you can feel the depression of the infraorbital foramen (this is often located slightly medial to the pupil line), where **ST-2** is located.

Needling

0.3–0.5 cun vertically or transversely (subcutaneously) towards **➞ S.I.-18** or **➞ L.I.-20** (for example, in cases of facial paralysis). Deep needling (in an oblique cranial direction) is contraindicated. Caution: Infraorbital nerve, eye injury. According to some classic texts, moxibustion is contraindicated.

Actions/Indications

- Benefits the eyes, clears Heat, eliminates Wind

Special features

Important local point for disorders of the eyes, pain and paralysis of the face; meeting point with the *yang qiao mai* according to some authors. A less risky substitute for **➞ ST-1**.

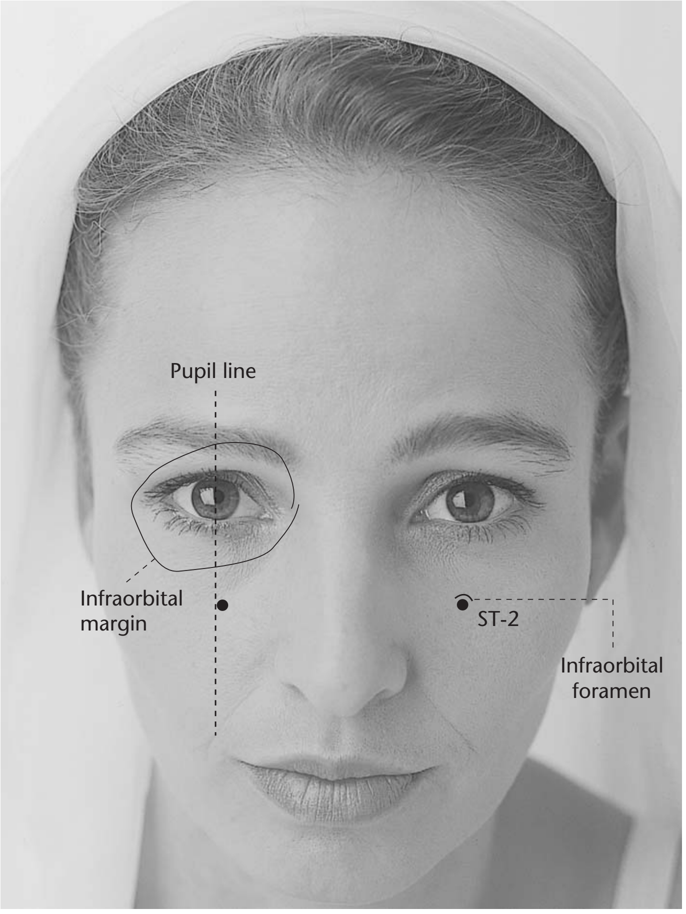

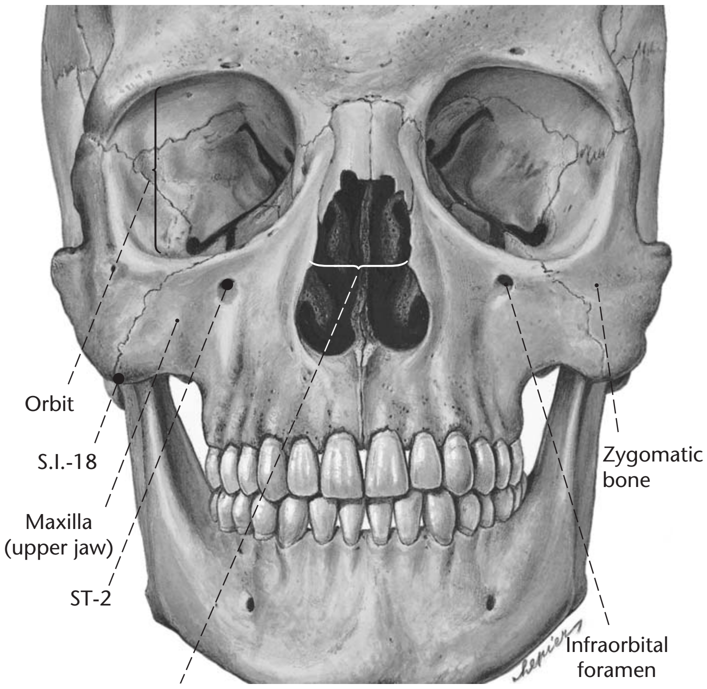

Great Crevice JULIAO **ST-3**

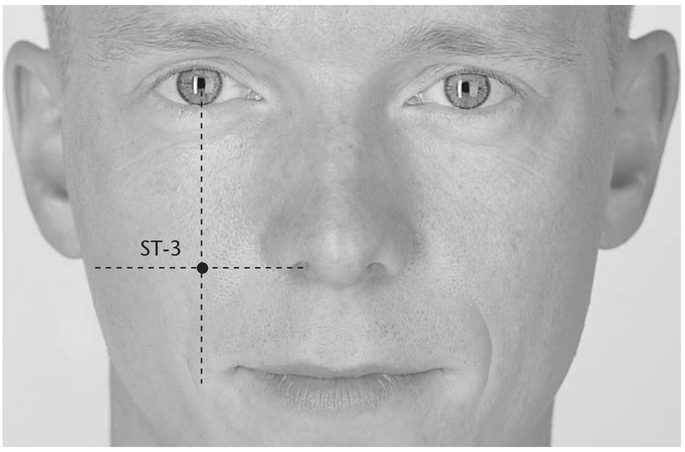

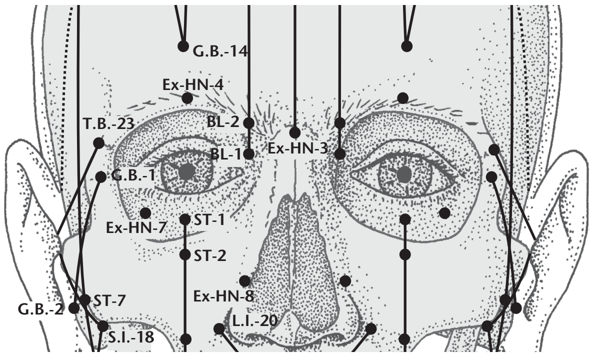
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20
S.I.-18

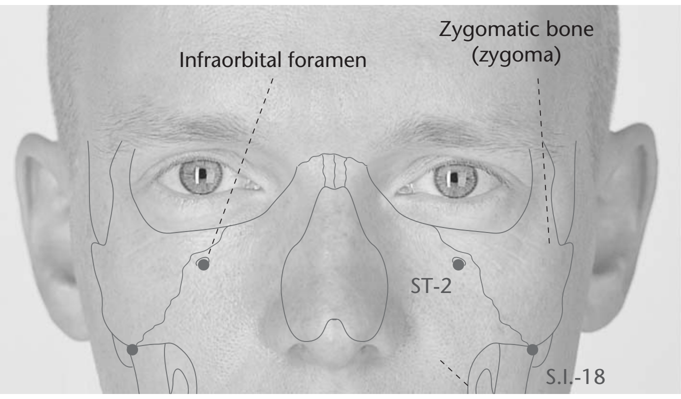
Infraorbital foramen
Zygomatic bone (zygoma)
ST-2
S.I.-18

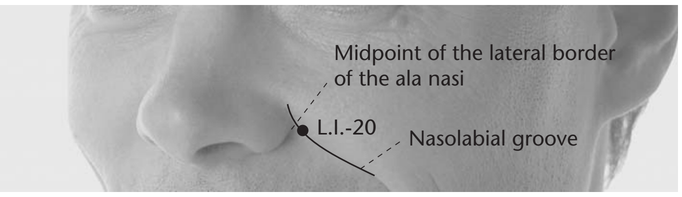
Midpoint of the lateral border of the ala nasi
L.I.-20
Nasolabial groove

Location

With the patient looking straight ahead, directly below the centre of the pupil, on the level of the lower border of the ala nasi.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil (pupil line). **ST-3** is located at the crossing point of the pupil line and a horizontal line on the level of the lower border of the ala nasi.

Needling

Vertically or obliquely 0.3–0.5 cun or penetration method: transversely (subcutaneously) towards → **ST-4,** → **S.I.-18**, etc, for example, in cases of peripheral facial paralysis.

Actions/Indications

- Expels Wind, opens the channel, disperses accumulations

Special features

Meeting point with the *yang qiao mai*

ST-4 *Earth Granary* DICANG

Location

With the patient looking straight ahead, directly below the centre of the pupil and approximately 0.4 cun lateral to the corner of the mouth.

How to find

With the patient looking straight ahead, the first four points of the ST channel are located on a vertical line drawn through the centre of the pupil (pupil line). **ST-4** is located at the crossing point of the pupil line and a horizontal line originating at the corner of the mouth. The point is located in the nasolabial groove (➞ 3.1.2), approximately 0.4 cun lateral to the corner of the mouth. Ask the patient to smile in order to make the line more clearly visible.

Needling

Obliquely in a lateral direction or vertically 0.3–0.5 cun or penetration method: transversely (subcutaneously) approximately 1.5–2 cun in cases of facial paralysis towards ➞ **ST-6**, for trigeminal neuralgia towards ➞ **L.I.-20**. Caution: Facial artery/vein.

Actions/Indications

- Expels Wind (from the face), opens the channel, alleviates pain, relaxes the facial musculature
- Sometimes as distal point for disorders of the leg

Special features

Meeting point with the L.I. channel, the *ren mai* and *yang qiao mai* (also with the *du mai* according to some authors). Important local point for neuralgias and pareses affecting the mouth and cheeks.

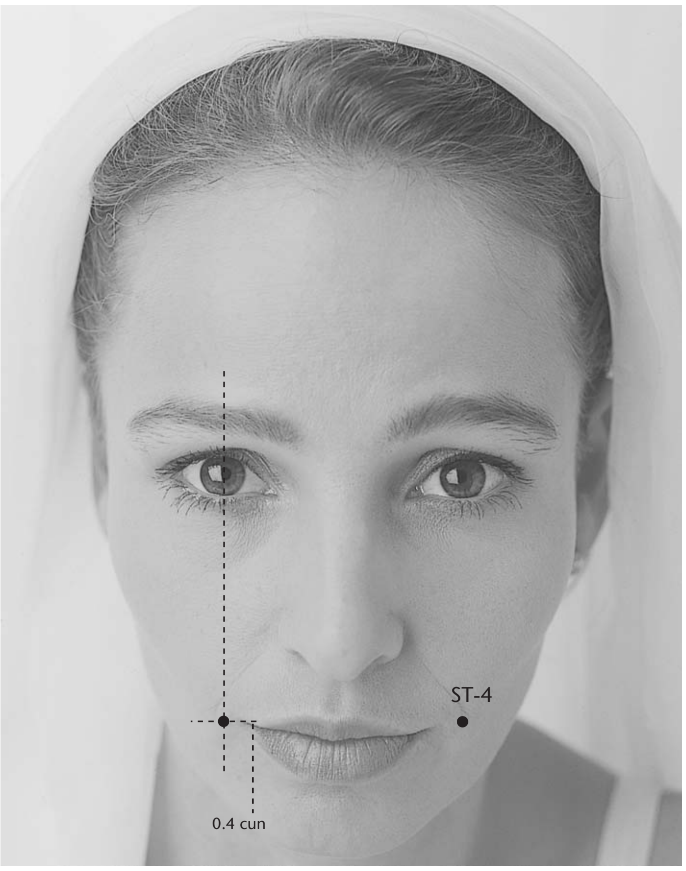

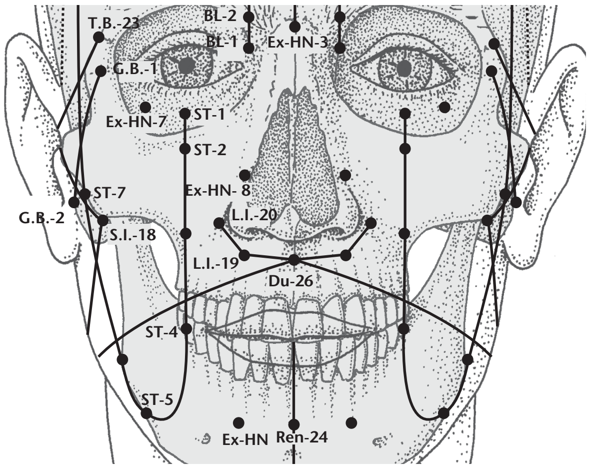

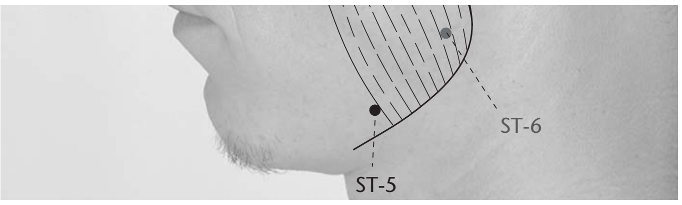

Great Welcome DAYING ST-5

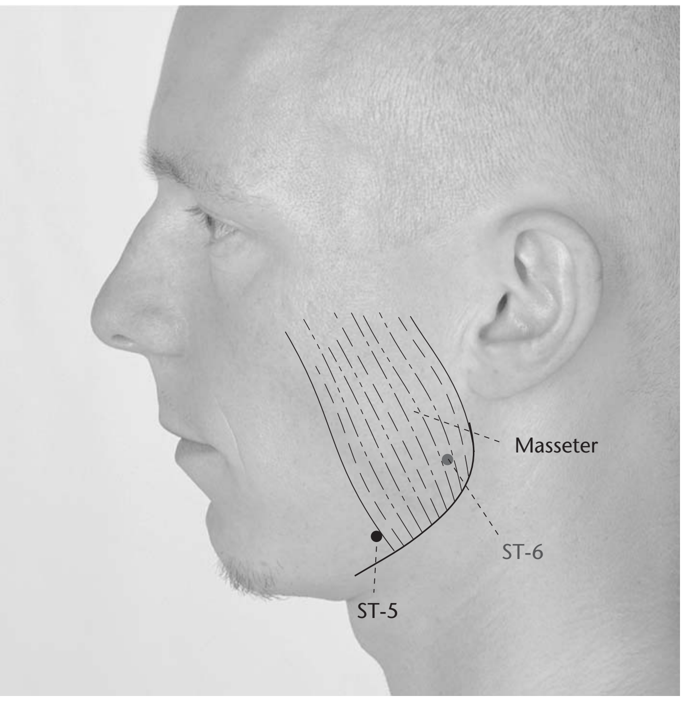
Masseter
ST-6
ST-5

Location

On the lateral mandible, anterior to the masseter muscle.

How to find

By asking the patient to clench their teeth, the anterior border of the masseter muscle becomes clearly palpable. **ST-5** is located directly anterior to the border of the muscle, in a shallow depression on the mandible and slightly above the angle of the jaw. Here, the facial artery can be palpated.

Needling

0.3–0.5 cun obliquely or transversely (subcutaneously) in the direction of → **ST-6**. Caution: Facial artery/vein.

Actions/Indications

- Expels Wind, opens the channels

Special features

Meeting point with the G.B. channel according to some authors

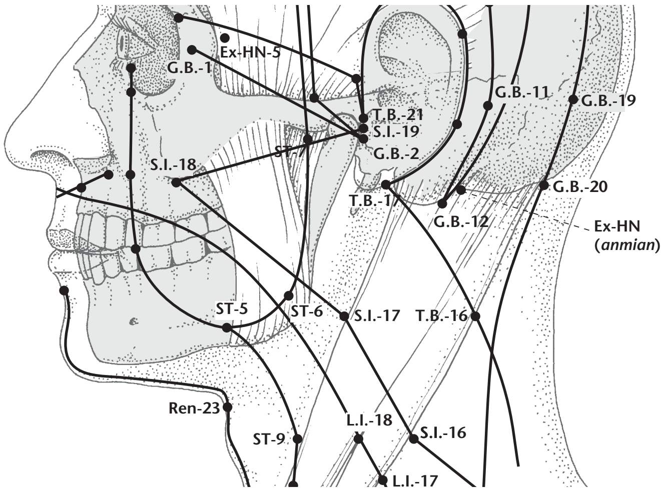
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
ST-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17

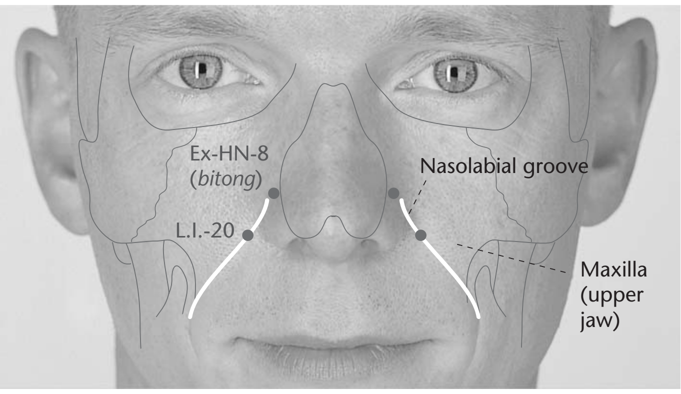
Ex-HN-8
(*bitong*)
Nasolabial groove
L.I.-20
Maxilla
(upper
jaw)

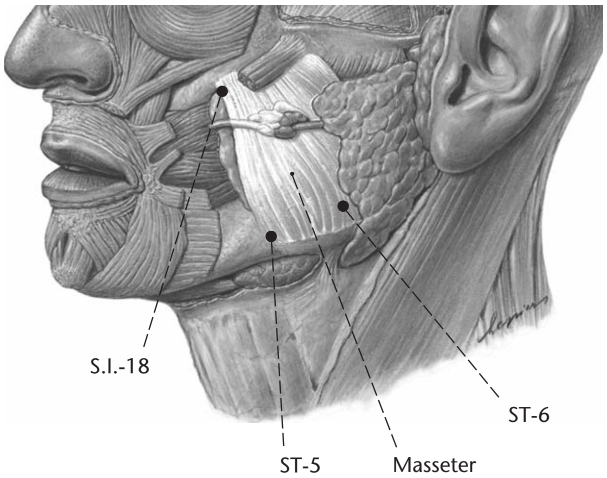
S.I.-18
ST-6
ST-5
Masseter

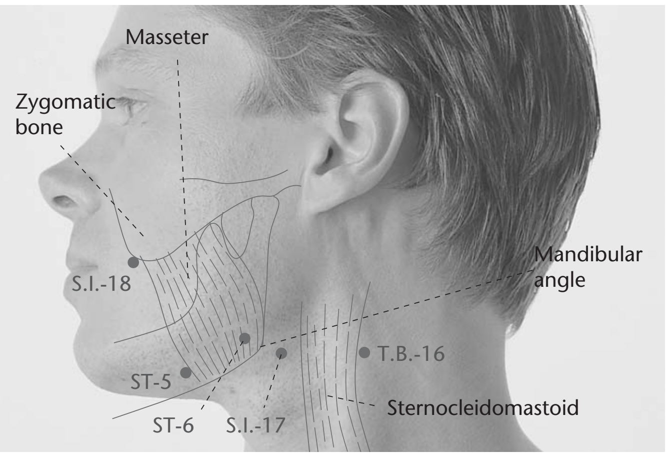
Masseter
Zygomatic
bone
S.I.-18
Mandibular
angle
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

ST-6 *Jaw Bone* JIACHE

Location

With the teeth clenched, on the highest prominence of the masseter muscle, approximately one finger-width (middle finger) anterior and superior to the angle of the jaw.

How to find

Ask the patient to clench their teeth, which will allow the muscle belly of the masseter muscle to bulge. **ST-6** is located on its prominence, a little anterior and superior to the angle of the jaw (mandibular angle → 3.1.3). With chewing movements of the jaw ('biting and relaxing'), the palpating finger will naturally fall into the right depression when the muscle is relaxed.
→ **ST-5** is located more medially in the depression at the anterior border of the masseter muscle.

Needling

Insertion with the patient's mouth slightly opened (relaxed masseter muscle). Vertically 0.3–0.5 cun or obliquely or transversely (subcutaneously) 1–1.5 cun in the direction of → **ST-4**, **ST-5**, **ST-7**, for example, in cases of facial paralysis.

Actions/Indications

- Eliminates Wind, benefits the jaw and teeth, opens the channel and *luo* vessels, alleviates pain

Special features

Sun Si Miao Ghost point. Important local point for the jaw region. Its location corresponds to a commonly used trigger point on the masseter muscle. According to some authors, **ST-6** is a meeting point with the G.B. channel.

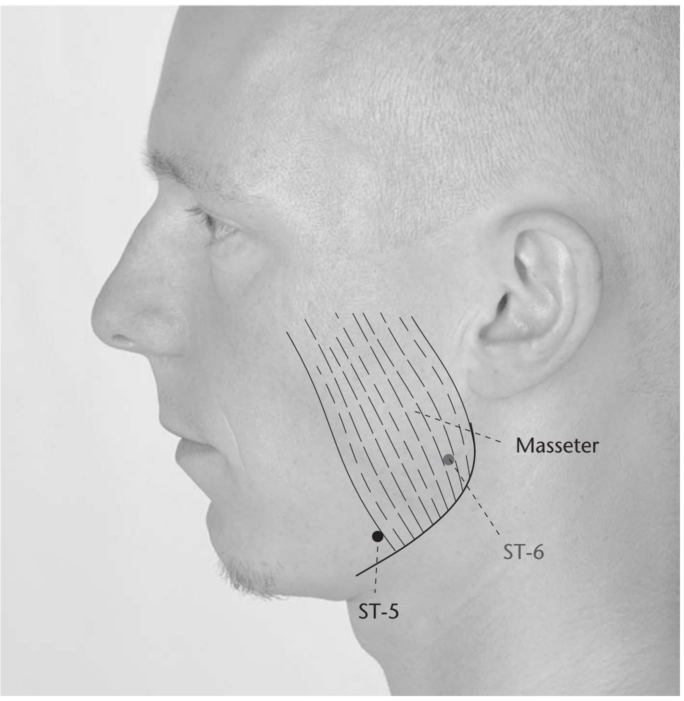
Masseter
ST-6
ST-5

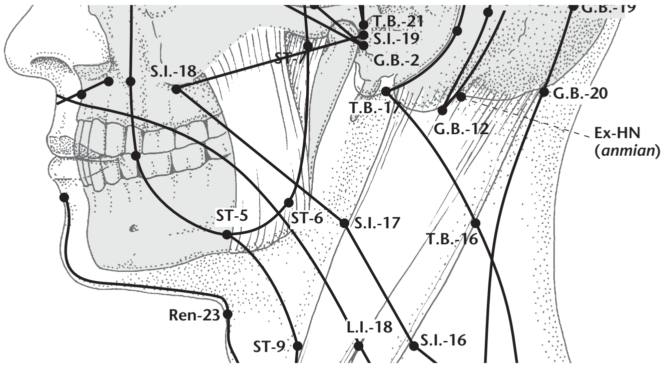
G.B.-19
T.B.-21
S.I.-19
G.B.-2
ST-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
ST-9
S.I.-16

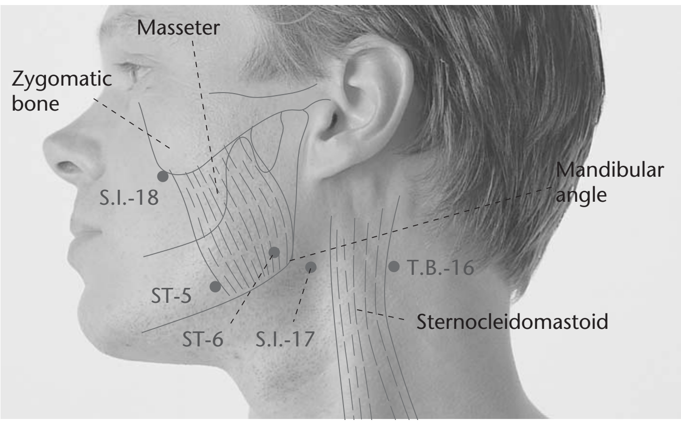
Masseter
Zygomatic
bone
S.I.-18
Mandibular
angle
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

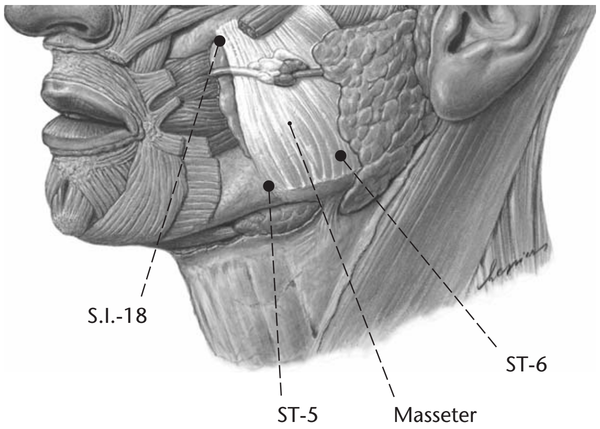
S.I.-18
ST-6
ST-5
Masseter

Below the Joint XIAGUAN ST-7

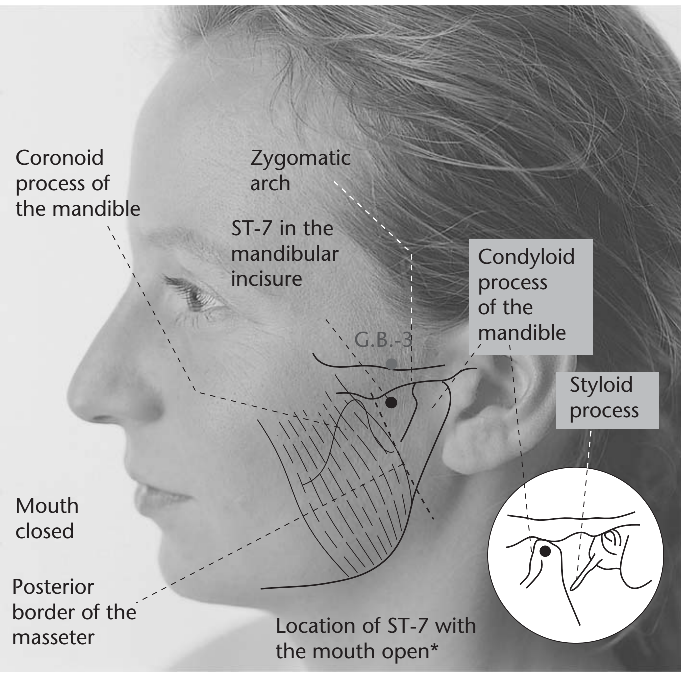
Coronoid process of the mandible
Zygomatic arch
ST-7 in the mandibular incisure
G.B.-3
Condyloid process of the mandible
Styloid process
Mouth closed
Posterior border of the masseter
Location of ST-7 with the mouth open*

Location

With the mouth closed, at the lower border of the zygomatic arch (→ 3.1.2) in the centre of the depression of the mandibular fossa, between the coronoid process and the condyloid process of the mandible.

How to find

Palpate along the lower border of the zygomatic arch (→ 3.1.2) towards the ear. **ST-7** is located in a clearly palpable depression just before the temporomandibular joint and at the posterior border of the masseter muscle (clench teeth). To confirm: When the patient's mouth is fully opened, the condyloid process of the lower jaw slides anteriorly and the depression disappears (figure modified after Deadman et al).

→ **G.B.-3** is located directly superior to **ST-7**, on the upper border of the zygomatic arch.

Needling

Insert the needle with the patient's mouth closed. 0.3–0.5 cun vertically or 1–1.5 cun transversely (subcutaneously) in the direction of → **ST-6**, → **S.I.-19**, → **S.I.-18**.

Actions/Indications

- Opens the channel, benefits the jaw and teeth, alleviates pain
- Benefits the ears

Special features

Meeting point with the G.B. channel. Important local point and frequently used trigger point.

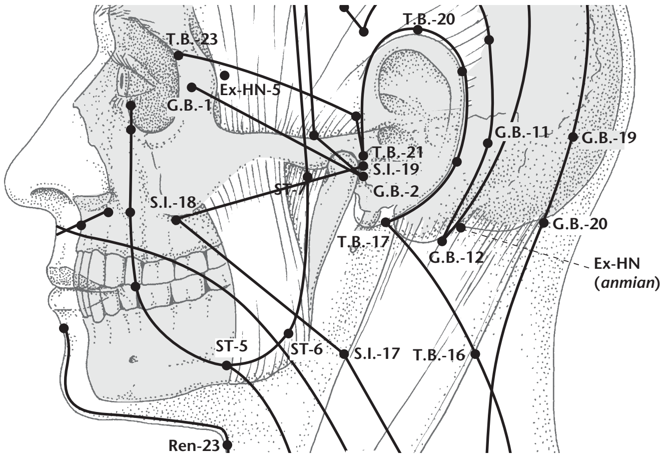
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-20
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23

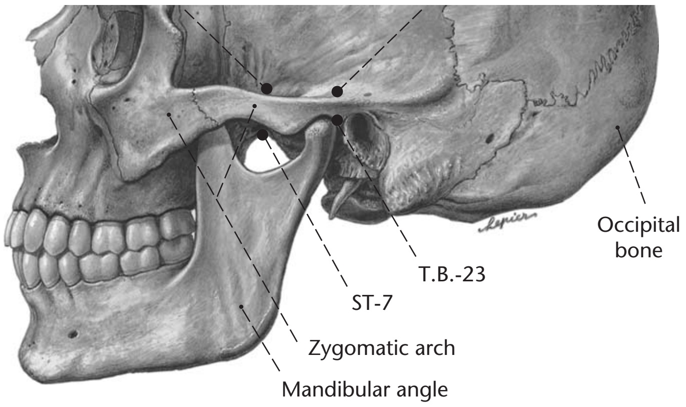
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

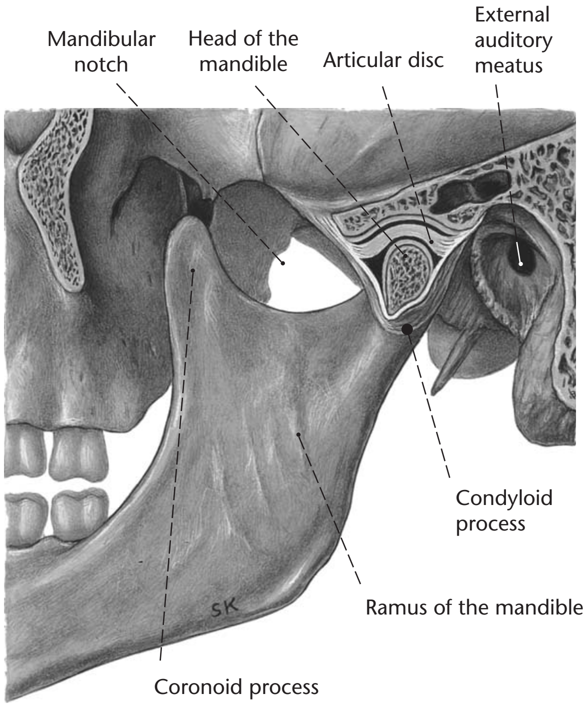
Mandibular notch
Head of the mandible
Articular disc
External auditory meatus
Condyloid process
Ramus of the mandible
Coronoid process
SK

---

* According to Deadman et al 1998.

ST-8 Head's Binding TOUWEI

Location

At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hairline or 4.5 cun lateral to the anterior midline (→ **Du-24**).

How to find

First, find the anterior hairline (→ 3.1.1). Then locate **ST-8** at the temporal corner of the forehead 0.5 cun within the hairline. Chewing movements of the temporalis muscle will still be palpable here.

Also located 0.5 cun superior to the anterior hairline are → **Du-24/BL-3/BL-4/G.B.-15/G.B.-13** (anterior midline/superior to the inner canthus of the eye/1.5 cun lateral to the anterior midline/3 cun lateral to the anterior midline). These distances refer to the distance between → **Du-24** and **ST-8** (=4.5 cun, → 2.2). **ST-8** and → **G.B.-7** (a depression on the level of the apex of the ear, within the temporal hairline) mark the endpoints of a curved line, on which **ST-8 and** → **G.B.-4–7** are located at regular intervals (by dividing the line into five parts).

Needling

0.5–1 cun transversely (subcutaneously) or obliquely in the direction of the disorder. Moxibustion possible (contraindicated according to some classical texts).

Actions/Indications

- Eliminates Wind from the head and eyes, benefits the eyes, alleviates pain

Special features

Meeting point with the *yang wei mai* and the G.B. channel. Important local point for headaches and disorders of the eyes.

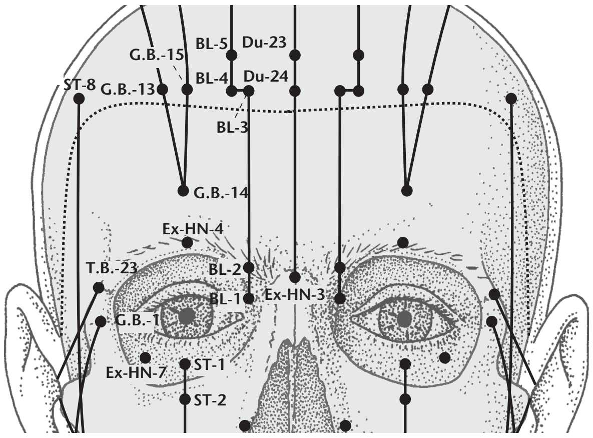
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2

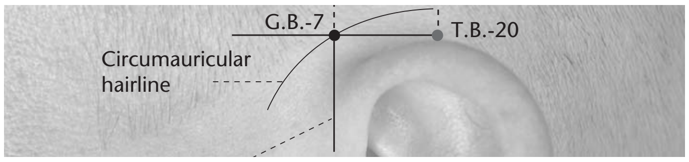
G.B.-7
T.B.-20
Circumauricular
hairline

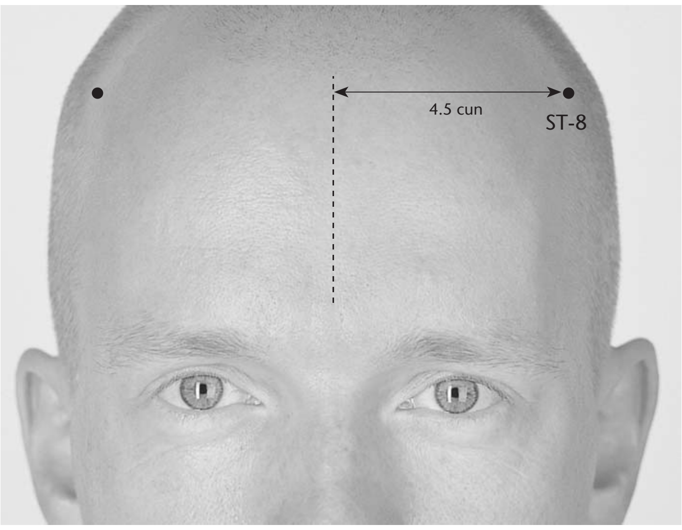

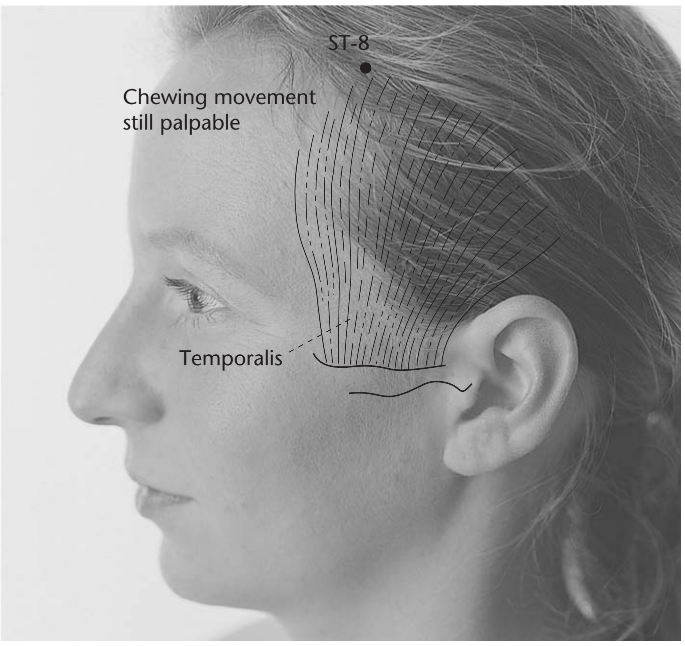
ST-8
Chewing movement
still palpable
Temporalis

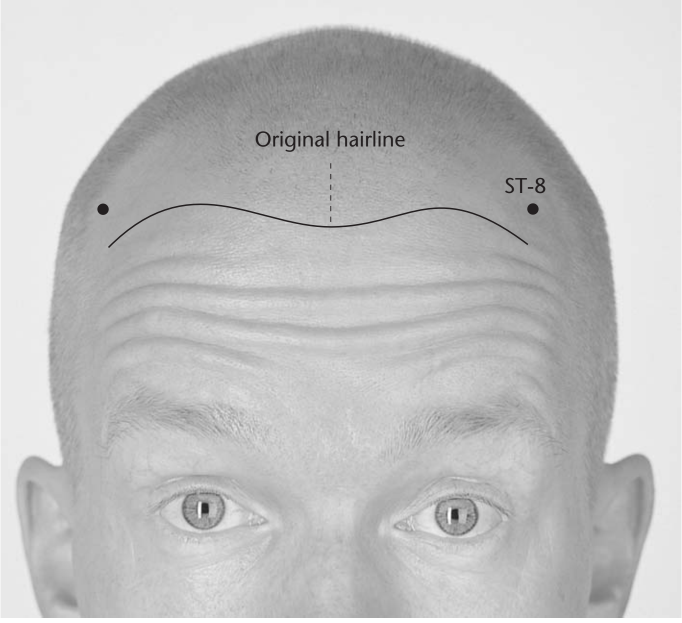
Original hairline
ST-8

Man's Welcome RENYING **ST-9**

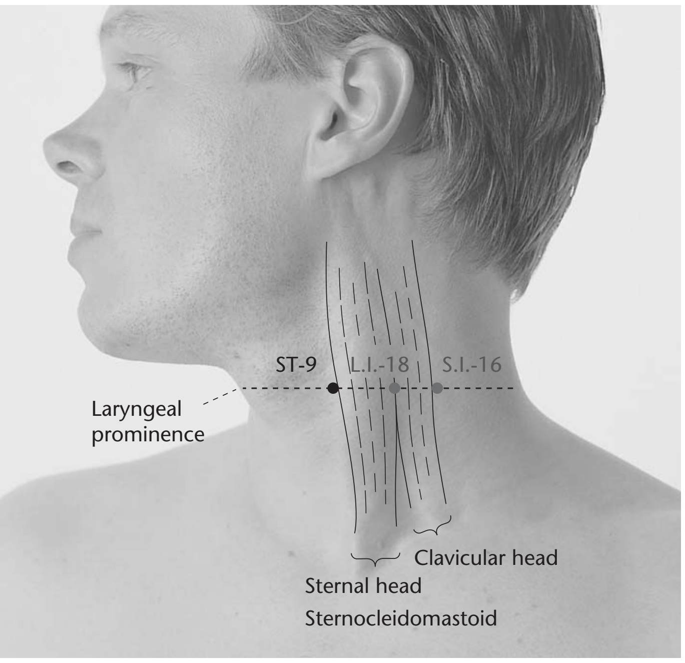
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

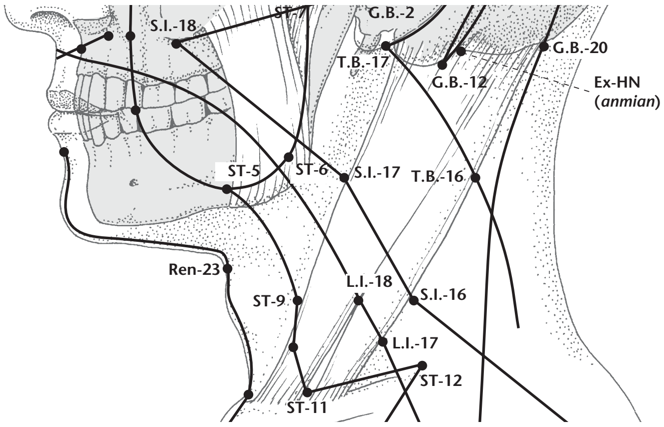
S.I.-18
ST-7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

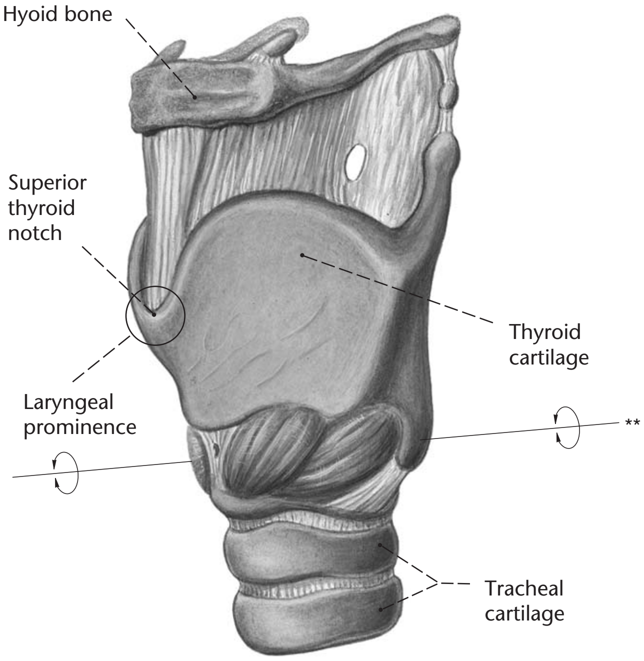
Hyoid bone
Superior thyroid notch
Laryngeal prominence
Thyroid cartilage
**
Tracheal cartilage

Location

1.5 cun lateral to the anterior midline, on the level of the laryngeal prominence and at the anterior border of the sternocleidomastoid muscle.

How to find

From the laryngeal prominence, palpate posteriorly to the anterior border of the sternocleidomastoid muscle (→ 3.2). **ST-9** is located at the anterior border of the muscle, which becomes more clearly visible and palpable if the patient's head is turned in the opposite direction.
Located on the same level are → **S.I.-16** on the posterior border of the sternocleidomastoid muscle and **L.I.-18** between the two heads of the sternocleidomastoid muscle.

Needling

If necessary, move the head to its normal position before needling. Caution: Carotid artery, carotid sinus, jugular veins. The carotid artery is located directly at the anterior border of the sternocleidomastoid muscle. The needle should be inserted anterior to where you can feel the pulsating of the carotid artery. Vertically 0.5–1 cun between the artery and the thyroid cartilage. For reasons of safety, the artery should be held in a lateral position with one finger of the palpating hand above and one finger below the insertion site. Irritation of the carotid sinus can lead to a lowering of the blood pressure and fainting!

Actions/Indications

- Regulates the Qi
- Opens the channel
- Harmonises excess or deficiency in the thorax

Special features

Window of Heaven point, Sea of Qi point, meeting point with the G.B. channel.

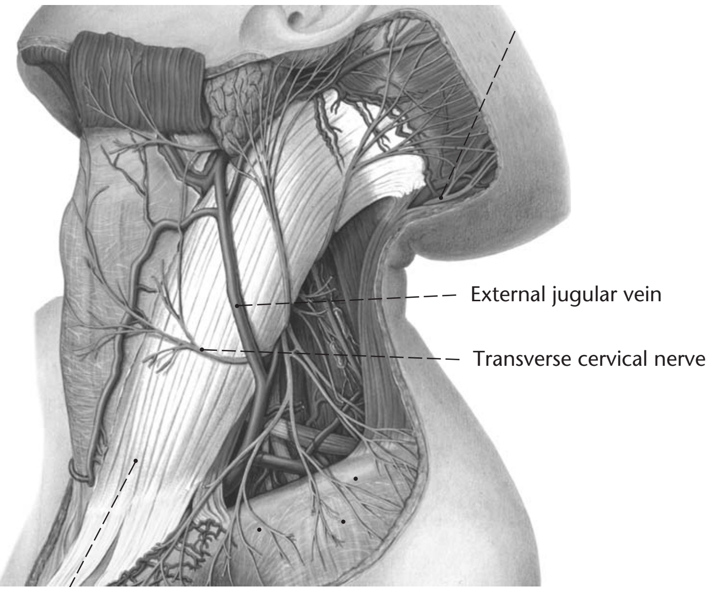
External jugular vein
Transverse cervical nerve

ST-10 *Water Prominence* SHUITU

Location

On the anterior border of the sternocleidomastoid muscle, at the midpoint of a line connecting ➞ **ST-9** and **ST-11**.

How to find

➞ **ST-9** and **ST-10** are both located on the anterior border of the sternocleidomastoid muscle (➞ 3.2). ➞ **ST-11** is located posterior to the sternal head. Place the middle finger of the left hand onto ➞ **ST-9** on the level of the laryngeal prominence and the thumb on ➞ **ST-11** posterior to the sternal insertion. **ST-10** is located at the midpoint between the two fingers. With the patient's head rotated in the opposite direction, the muscle will be more visible and can be more easily palpated.

Needling

0.5–1 cun obliquely in a medial direction, away from the artery. Caution: Carotid artery, jugular veins. If necessary, rotate the head back to a middle position before needling.

Actions/Indications

- Regulates the Lung Qi
- Benefits the throat

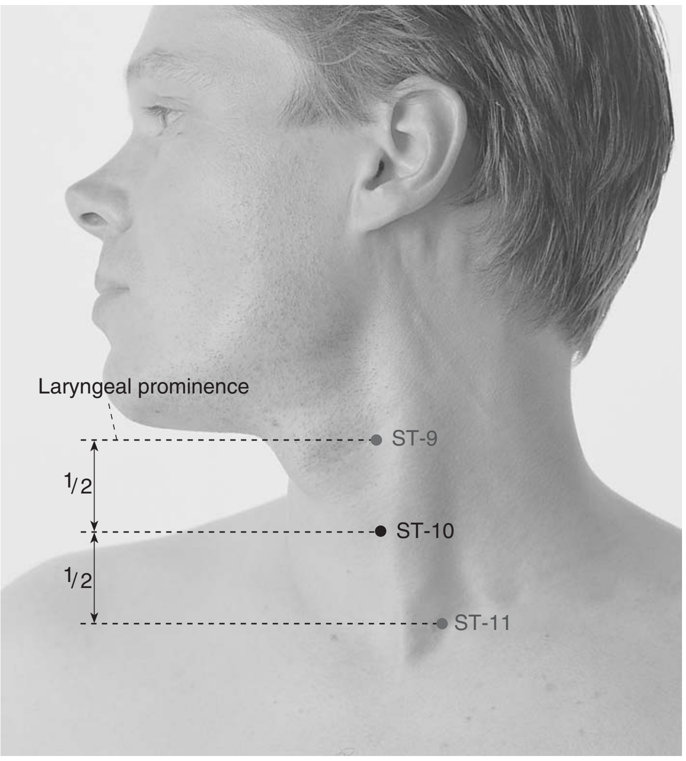
Laryngeal prominence
ST-9
1/2
ST-10
1/2
ST-11

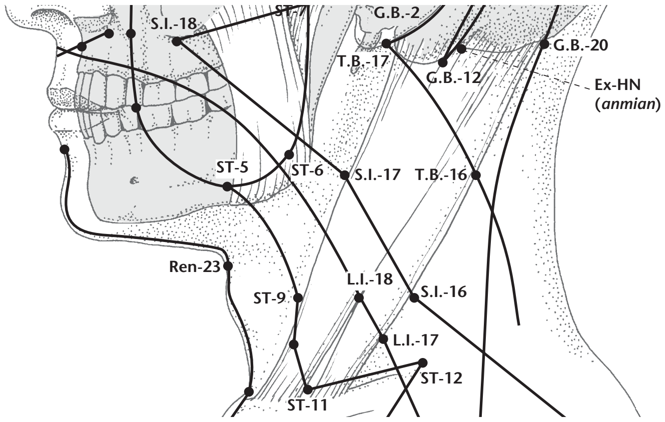
S.I.-18
ST-7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

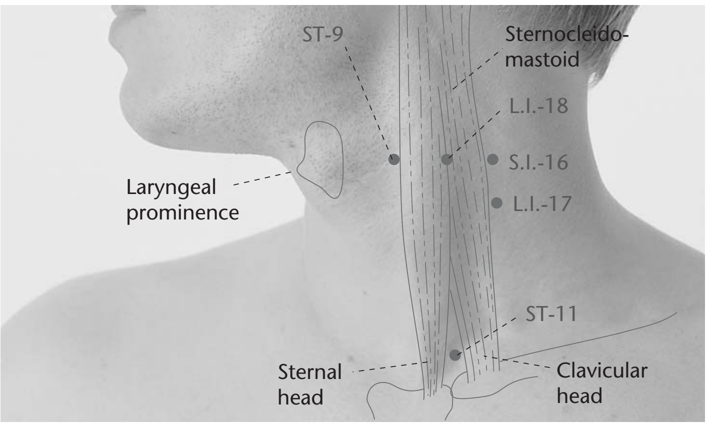
ST-9
Sternocleido-
mastoid
L.I.-18
S.I.-16
Laryngeal
prominence
L.I.-17
ST-11
Sternal
head
Clavicular
head

Qi Abode QISHE ST-11

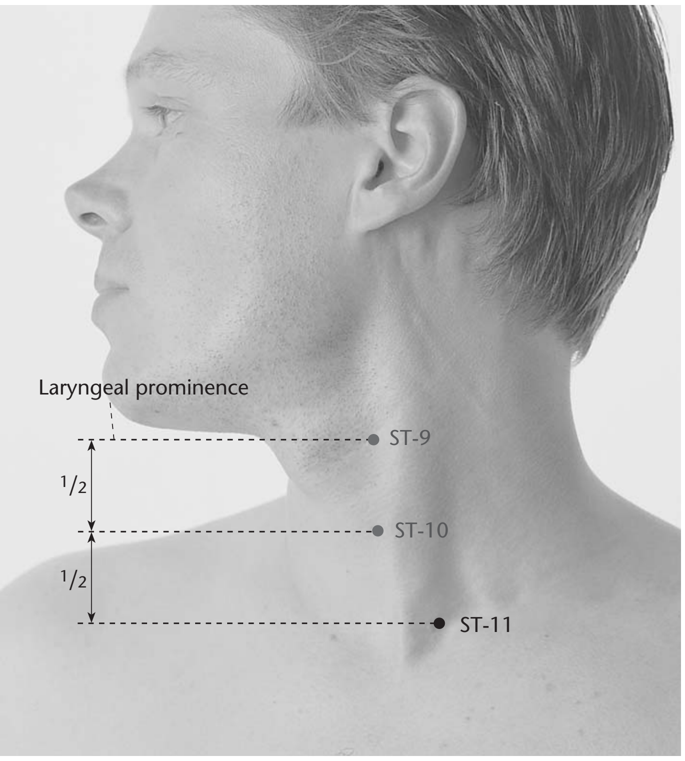
Laryngeal prominence
ST-9
1/2
ST-10
1/2
ST-11

Location

On the upper border of the clavicle, between the tendons of the sternal and clavicular heads of the sternocleidomastoid muscle.

How to find

By rotating the patient's head in the opposite direction (especially against resistance), the ropey sternal insertion of the sternocleidomastoid muscle (→ 3.2) becomes clearly visible. **ST-11** is located in a shallow depression, the minor supraclavicular fossa, between the sternal insertion and the more extended insertion of the clavicular head.

Needling

Vertically 0.3–0.5 cun. Caution: Pneumothorax, jugular veins.

Actions/Indications

- Benefits the throat and neck
- Regulates counterflow Qi

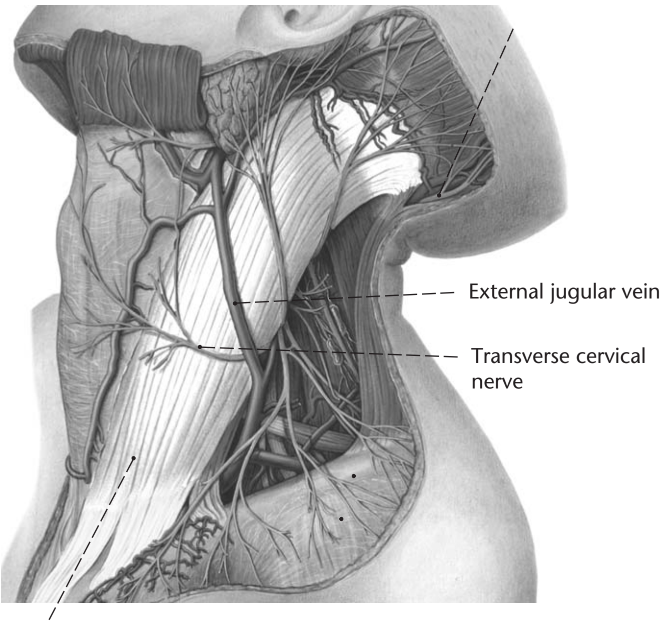
External jugular vein
Transverse cervical nerve

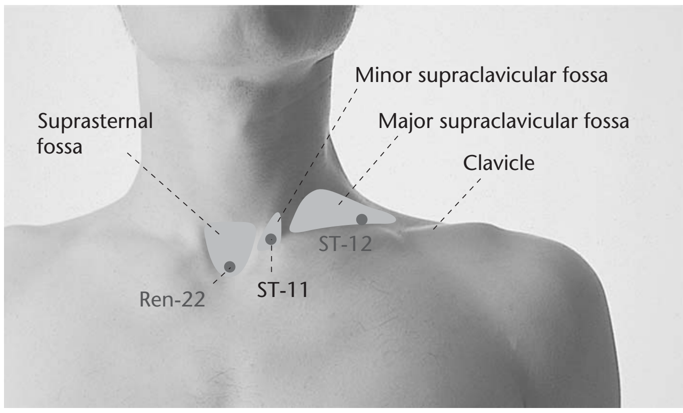
Minor supraclavicular fossa
Suprasternal fossa
Major supraclavicular fossa
Clavicle
ST-12
ST-11
Ren-22

ST-12 *Empty Basin* QUEPEN

Location

In the supraclavicular fossa, superior to the midpoint of the clavicle, approximately 4 cun lateral to the anterior midline.

How to find

From the anterior midline, palpate to the midpoint of the clavicle (=4 cun in a lateral direction). There, an imaginary vertical line marks the midclavicular line. **ST-12** is located on the midclavicular line, at the upper border of the clavicle, in the supraclavicular fossa, lateral to the clavicular head of the sternocleidomastoid muscle.

Needling

Vertically 0.3–0.5 cun. The needle should be guided along the posterior border of the clavicle. Caution: Pneumothorax, especially with emphysema; superficial and deep cervical arteries. According to some authors, this point is contraindicated during pregnancy.

Actions/Indications

- Opens the channel
- Disperses and descends the Lung Qi
- Clears Heat in the throat

Special features

Meeting point with the L.I., S.I., G.B. and T.B. channel

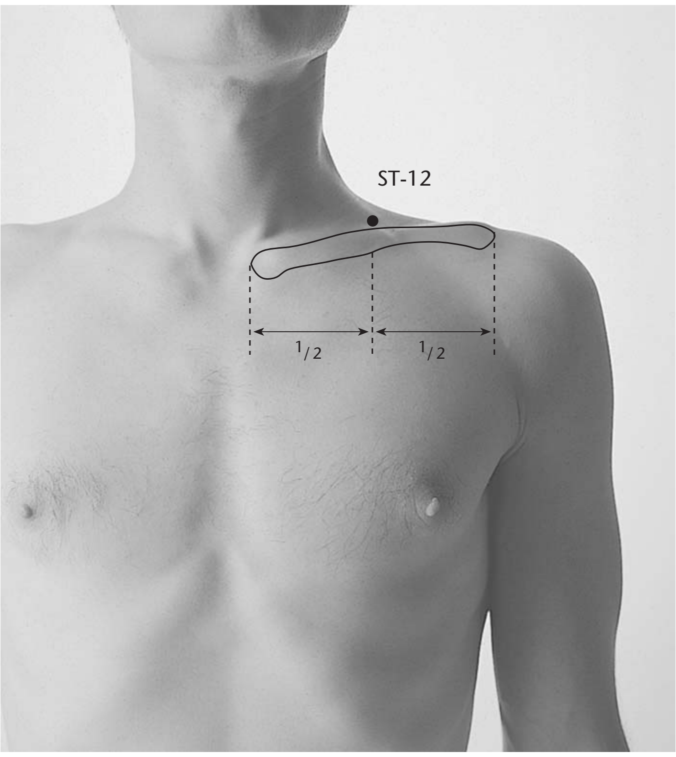

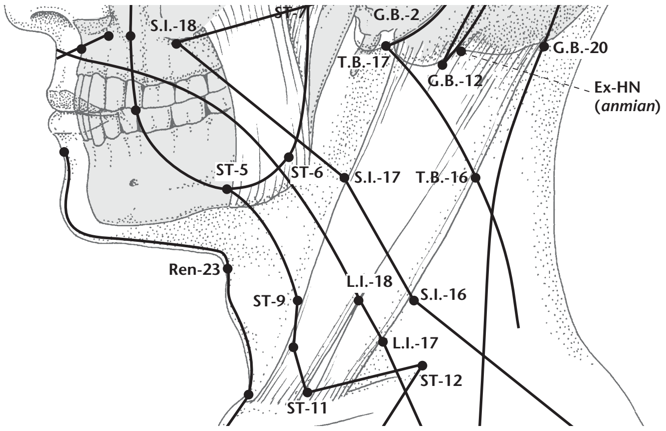
S.I.-18
ST-7
G.B.-2
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

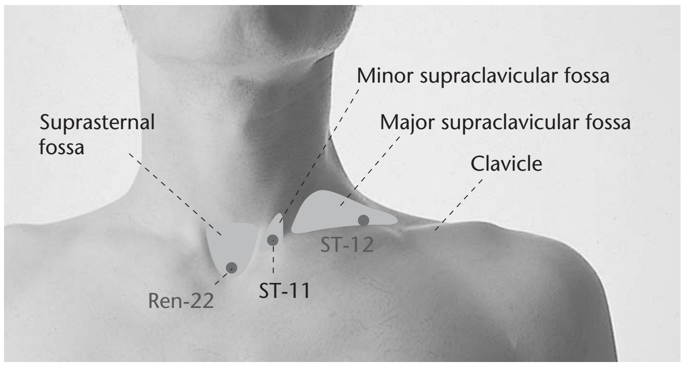
Minor supraclavicular fossa
Suprasternal
fossa
Major supraclavicular fossa
Clavicle
ST-12
ST-11
Ren-22

Ch04.3-F10028.qxd 2/22/08 8:42 PM Page 141

Qi Door QIHU ST-13

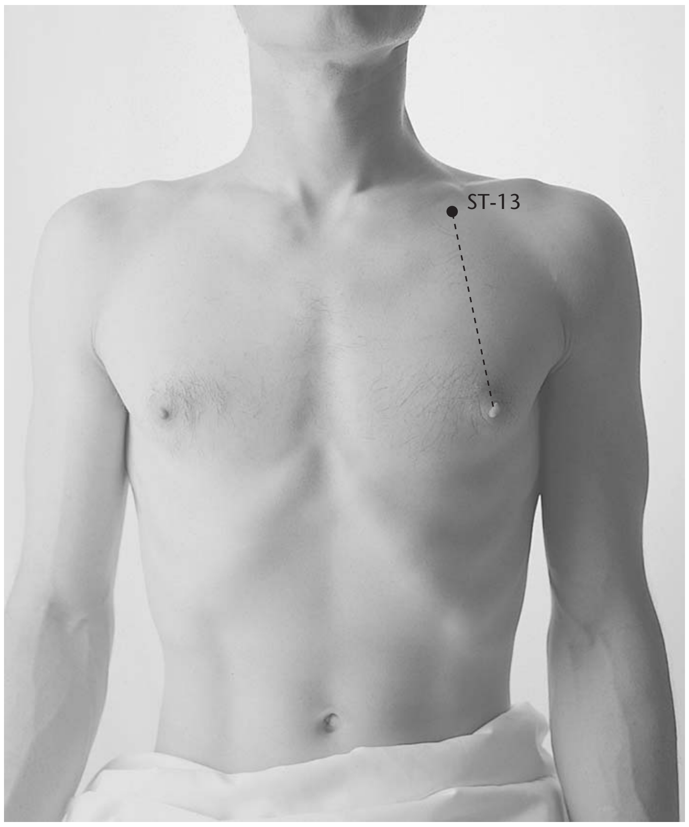

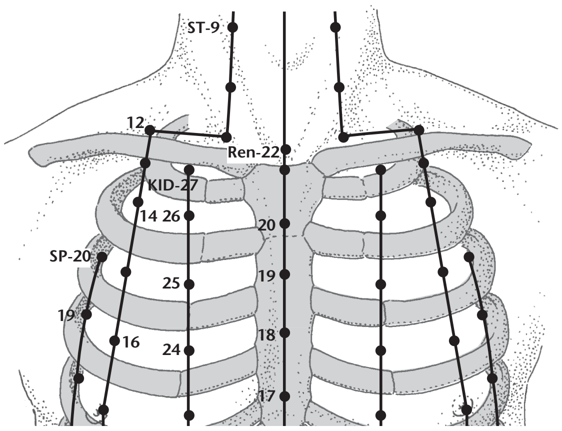

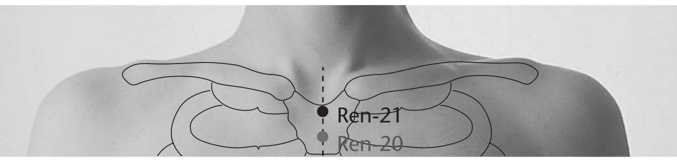

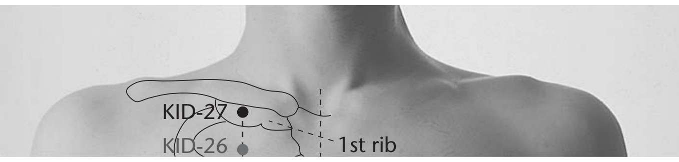

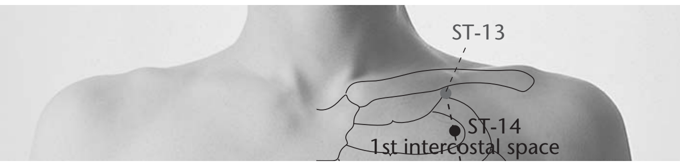

Location

At the midpoint of the clavicle and on its inferior border, 4 cun lateral to the anterior midline.

How to find

On the upper thorax region, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Find the midpoint of the clavicle (or measure 4 cun lateral to the midline) and locate **ST-13** at the lower border of the clavicle. Since the 1st rib curves posteriorly in a short bow, **ST-13** is generally situated lateral to the 1st rib.

→ **KID-27** is located medial to **ST-13**, but inferior to the lower border of the clavicle and 2 cun lateral to the anterior midline. → **Ren-21** is located on the midline. → **ST-14** is located only slightly inferior to **ST-13** in the 1st intercostal space.

Needling

0.3–0.5 cun obliquely in a lateral or medial direction or transversely (subcutaneously) along the channel pathway. Caution: Pneumothorax, subclavicular artery/vein.

Actions/Indications

- Descends counterflow Qi
- Clears Heat
- Unbinds the chest

ST-14 *Store House* KUFANG

Location

On the midclavicular line, in the first intercostal space, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. By palpating across the clavicle from superior to inferior, the first rib will generally be felt directly below the clavicle, although, in some cases, the 1st rib might be completely hidden under the clavicle. The 1st intercostal space follows directly inferior to the 1st rib, where **ST-14** is located.

→ **ST-13** is located only slightly superior to **ST-14** at the lower border of the clavicle. Also located on the level of the first intercostal space are → **Ren-20** (anterior midline), → **KID-26** (2 cun lateral to the anterior midline) and → **LU-1** (6 cun lateral to the midline). For more detail about orientation regarding intercostal spaces → 3.5.

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi
- Unbinds the chest
- Benefits the breasts

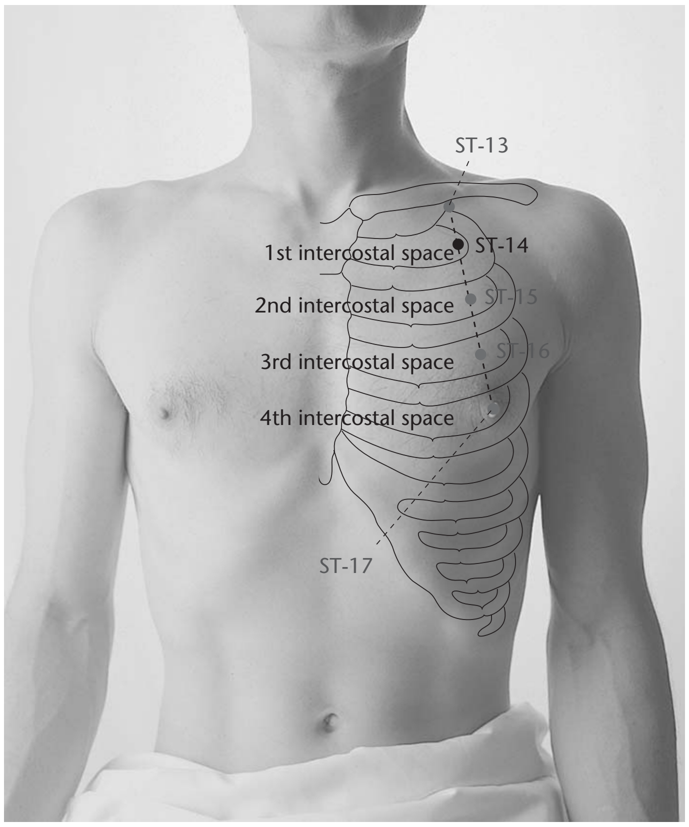
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

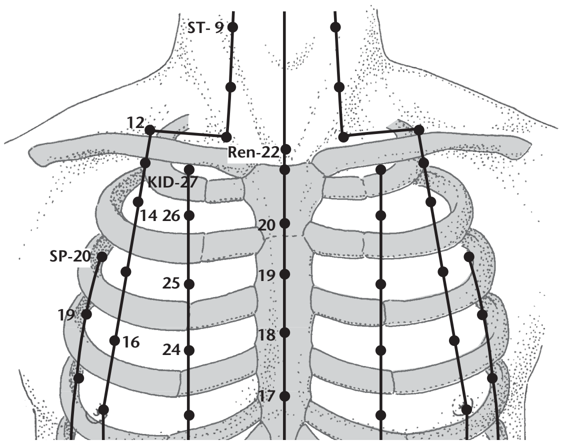
ST- 9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17

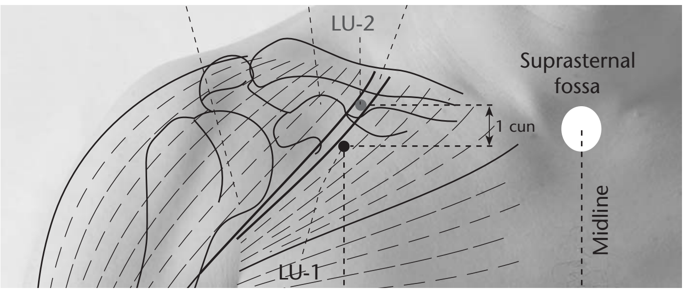
LU-2
Suprasternal fossa
1 cun
Midline
LU-1

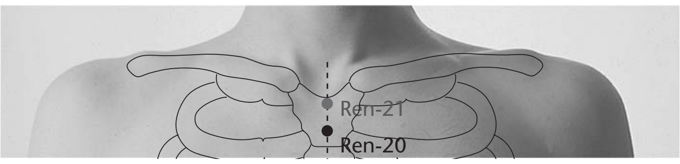
Ren-21
Ren-20

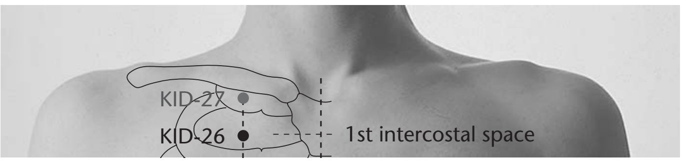
KID-27
KID-26
1st intercostal space

Room Screen WUYI ST-15

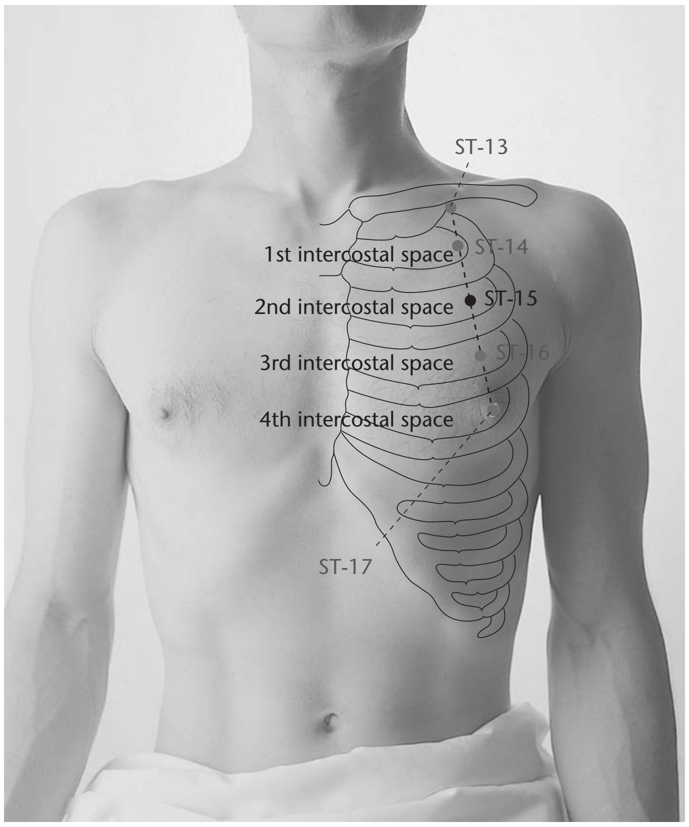
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

Location

In the 2nd intercostal space, on the midclavicular line, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Locate the 2nd intercostal space by palpating downward from the clavicle or by palpating from the manubriosternal synchondrosis (second rib) (→ 3.5). Then palpate along its course to the midclavicular line, where **ST-15** is located. Note the ascending course of the intercostal space! Located on the same level are → **Ren-19** (anterior midline), → **KID-25** (2 cun lateral to the anterior midline) and → **SP-20** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: pneumothorax.

Actions/Indications

- Regulates the Lung Qi
- Unbinds the chest
- Benefits the breasts and clears Heat

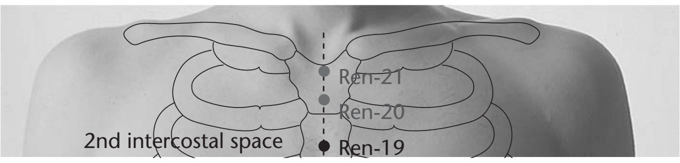
Ren-21
Ren-20
2nd intercostal space
Ren-19

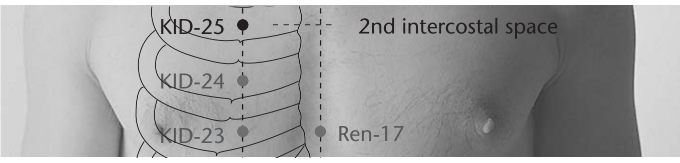
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

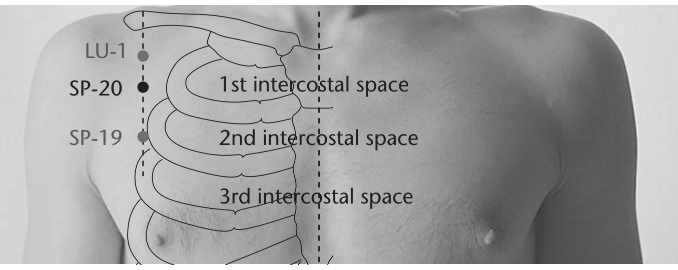
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

ST-16 *Breast Window* YINGCHUANG

Location

In the 3rd intercostal space, on the midclavicular line, 4 cun lateral to the anterior midline.

How to find

On the upper thorax, the midclavicular line (4 cun lateral to the anterior midline) runs slightly obliquely from the clavicular midpoint to the generally more lateral nipple. Locate the 3rd intercostal space by palpating downward from the clavicle or by palpating from the manubriosternal synchondrosis (second rib) (→ 3.5). Then palpate along its course to the midclavicular line, where **ST-16** is located. Note the ascending course of the intercostal space! In male patients, this point can also be located by counting one intercostal space upward from the level of the nipple.

Located on the same level are → **Ren-18** (anterior midline), → **KID-24** (2 cun lateral to the anterior midline) and → **SP-19** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), along the course of the intercostal space or with or against the flow of the channel. Caution: pneumothorax.

Actions/Indications

- Regulates the Qi
- Unbinds the chest
- Benefits the breasts

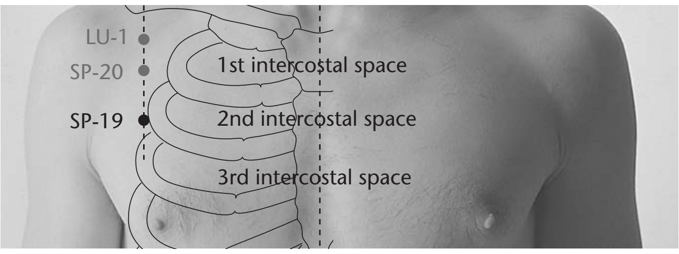
LU-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

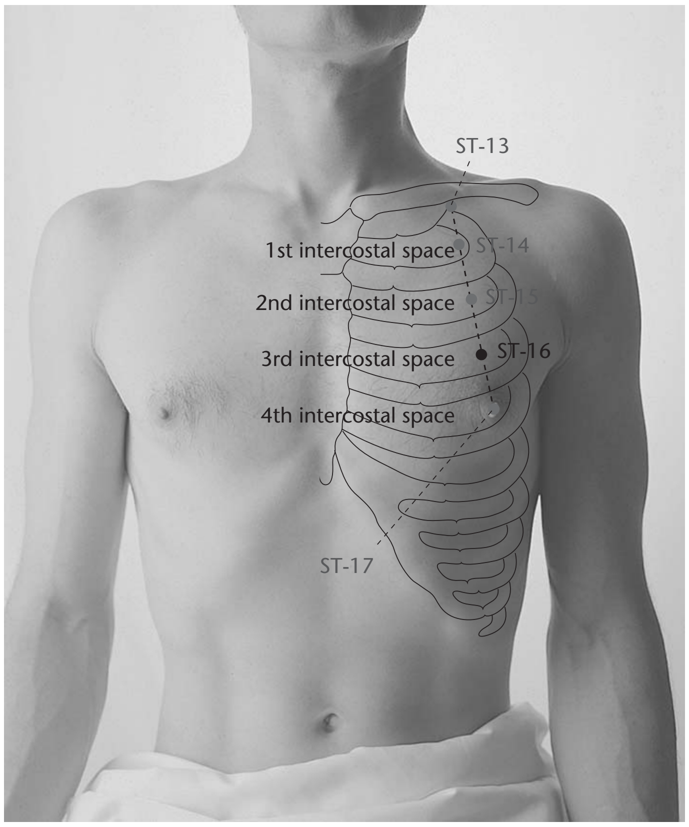
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

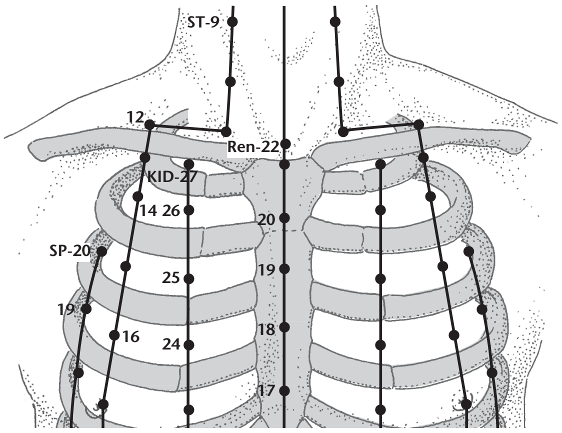
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17

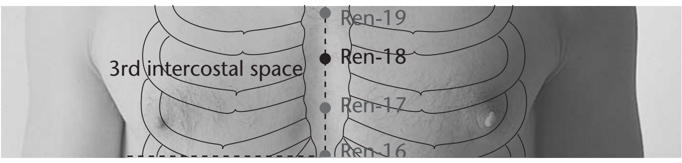
Ren-19
3rd intercostal space
Ren-18
Ren-17
Ren-16

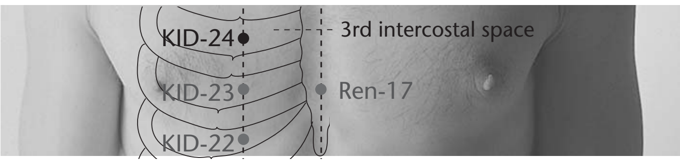
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

Breast Centre RUZHONG ST-17

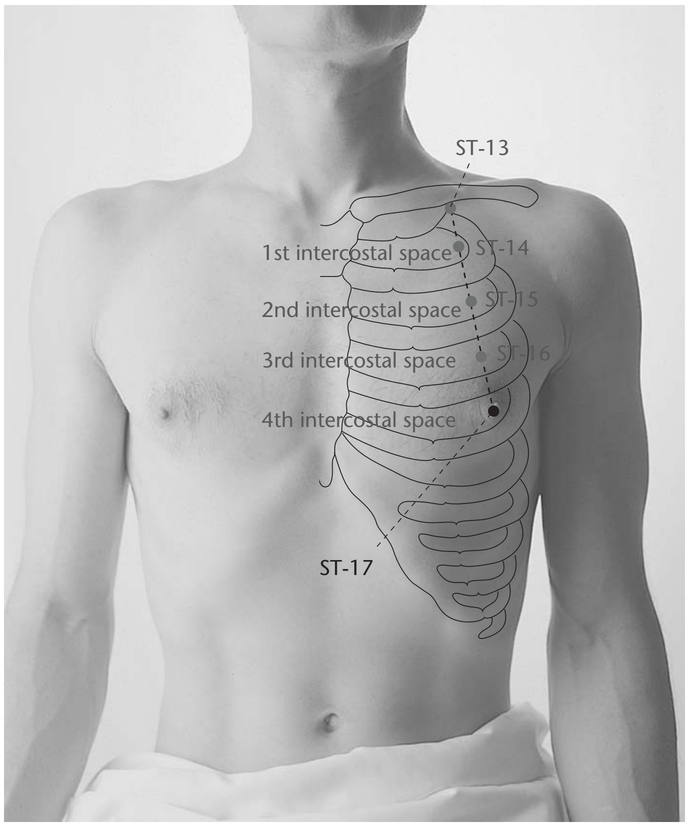
ST-13
1st intercostal space
ST-14
2nd intercostal space
ST-15
3rd intercostal space
ST-16
4th intercostal space
ST-17

Location

In the centre of the nipple.

How to find

In the centre of the nipple. In men, this is usually located 4 cun lateral to the anterior midline, at the level of the 4th intercostal space and is also used for reference.

In men, also located on the same level are → **Ren-17** (anterior midline), → **KID-23** (2 cun lateral to the anterior midline), → **SP-18** (6 cun lateral to the midline), → **P-1** (1 cun lateral to the nipple), → **G.B.-22** and → **G.B.-23**. In women, the location of the nipples varies considerably and should therefore not be used for reference.

Needling

No therapy; this point is for reference only.

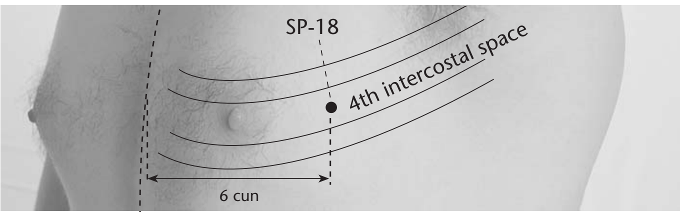
SP-18
4th intercostal space
6 cun

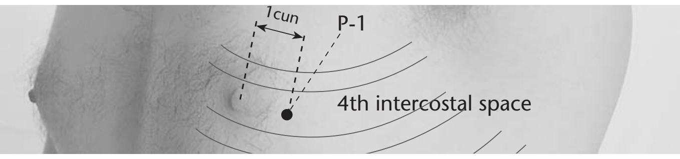
1cun
P-1
4th intercostal space

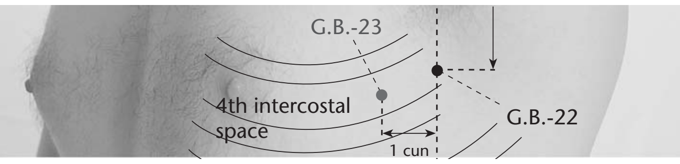
G.B.-23
4th intercostal
space
1 cun
G.B.-22

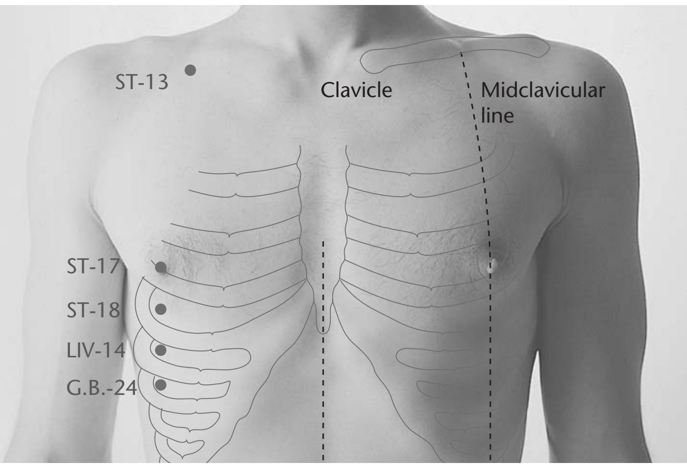
ST-13
Clavicle
Midclavicular
line
ST-17
ST-18
LIV-14
G.B.-24

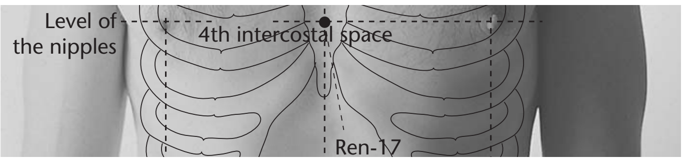
Level of
the nipples
4th intercostal space
Ren-17

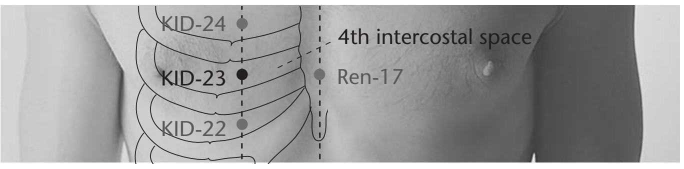
KID-24
4th intercostal space
KID-23
Ren-17
KID-22

ST-18 *Breast Root* RUGEN

Location

In the 5th intercostal space, on the mamillary line, 4 cun lateral to the anterior midline.

How to find

The mamillary line, a vertical line used for reference, is located 4 cun lateral to the anterior midline (→ 3.5). In men, the nipple is generally located on the level of the 4th intercostal space, and **ST-18** is located in the intercostal space directly below. The point is often located at the lower border of the pectoralis major muscle; in women, it is often found at the root of the breast. Location help in women: Find the 5th intercostal space by palpating downward from the clavicle or from the manubriosternal synchondrosis (second rib) (→ 3.5). Then follow the course of the intercostal space laterally to the mamillary line and there locate **ST-18**. Note the ascending course of the intercostal spaces!

Located on the same level are → **Ren-16** (anterior midline), → **KID-22** (2 cun lateral to the anterior midline) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the course of the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Benefits the breasts, moves the Blood
- Unbinds the chest
- Regulates the Lung Qi

Special features

Important local point

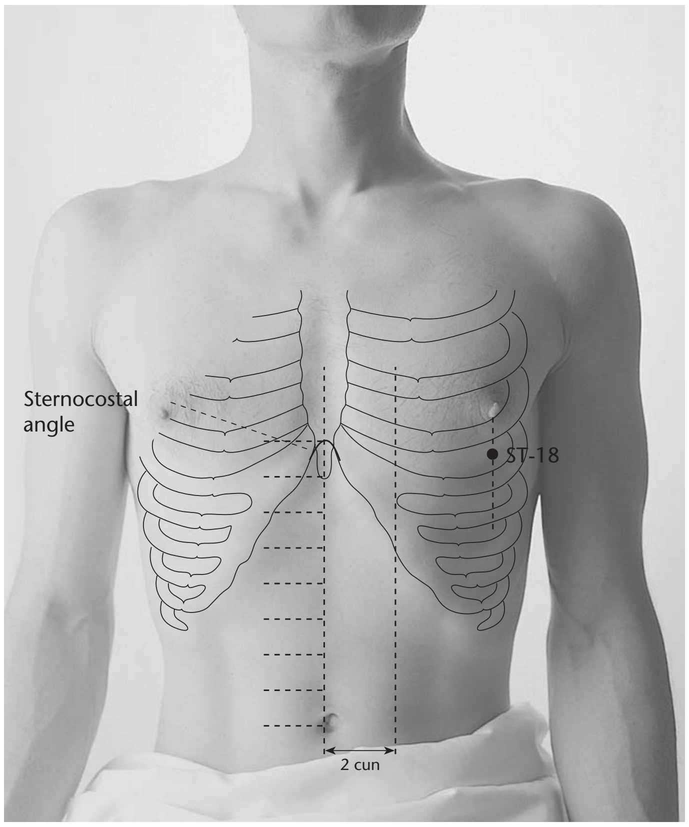
Sternocostal angle
ST-18
2 cun

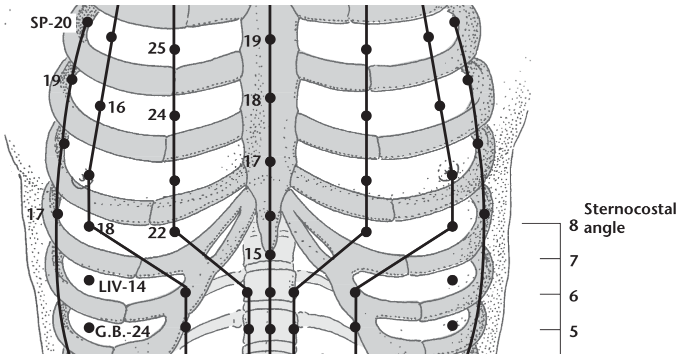
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
Sternocostal angle
8
7
6
5

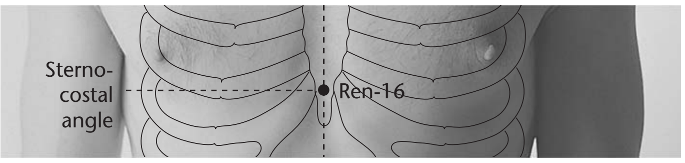
Sterno-costal angle
Ren-16

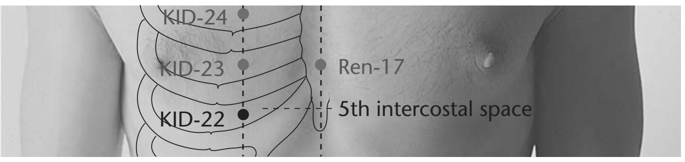
KID-24
KID-23
KID-22
Ren-17
5th intercostal space

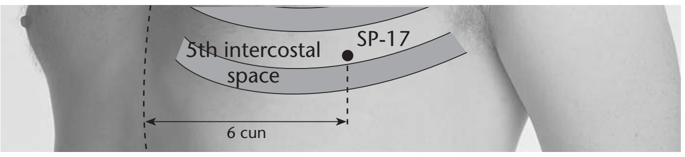
5th intercostal space
SP-17
6 cun

Not Contained BURONG ST-19

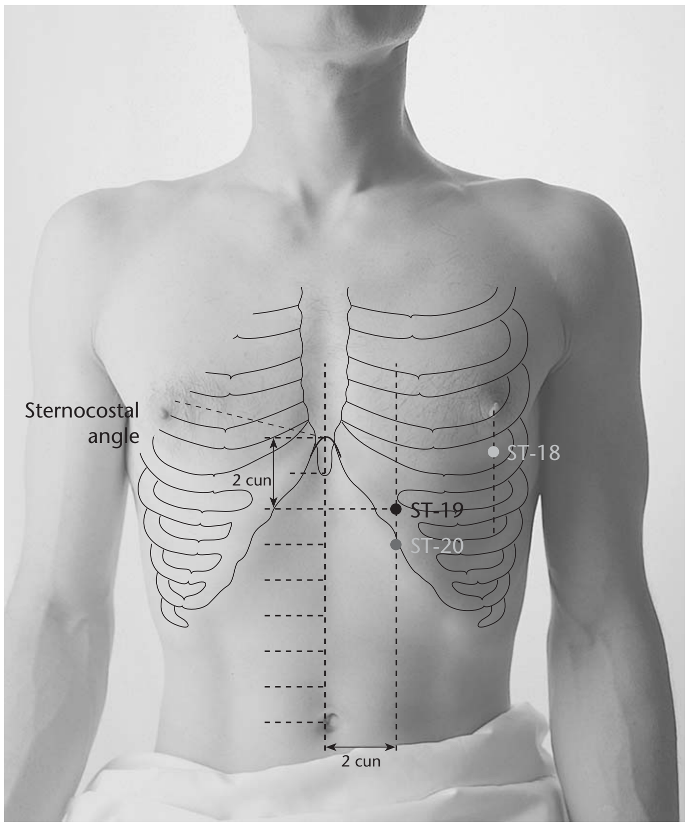

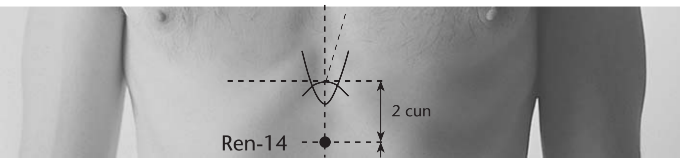

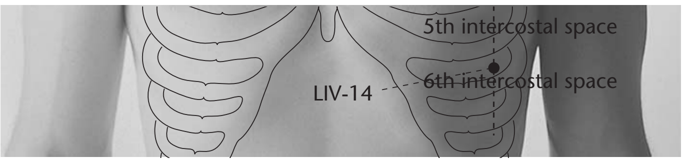

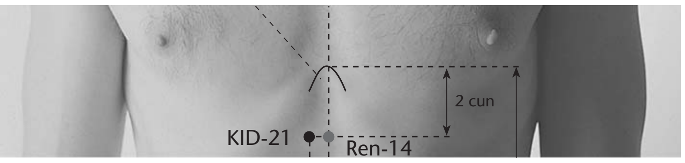

Location

2 cun below the sternocostal angle and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 2 cun down from the sternocostal angle. **Or:** Measure the midpoint of the upper half of the distance between the umbilicus and the subcostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). From there, measure 2 cun in a lateral direction to locate **ST-19**. As its location depends on the shape of the thorax, it might be located either on the rib or on the abdomen. Located on the same level are → **Ren-14** (anterior midline), → **KID-21** (0.5 cun lateral to the anterior midline) and → **LIV-14** (on the mamillary line and in the 6th intercostal space). Please note: From **ST-19** onwards, the ST channel is located only 2 cun lateral to the anterior midline.

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point) Caution: Peritoneum, pericardium.

Actions/Indications

- Regulates the Stomach, Intestines and Middle Burner
- Regulates the Lung Qi

ST-20 *Supporting Fullness* CHENGMAN

Location

3 cun below the sternocostal angle (or 5 cun above the umbilicus) and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 3 cun down from the sternocostal angle. **Or:** Measure 1 cun up from the midpoint of the distance between the umbilicus and the sternocostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). From there, measure 2 cun in a lateral direction to locate **ST-20**. Depending on the shape of the thorax, **ST-20** might be located on the lower border of the ribcage in some cases (narrow subcostal angle).
Located on the same level are → **Ren-13** (anterior midline) and → **KID-20** (0.5 cun lateral to the anterior midline) and → **G.B.-24** (on the mamillary line, in the 7th intercostal space).

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point). Caution: Peritoneum.

Actions/Indications

- Regulates the Middle Burner and the Stomach Qi
- Regulates the Lung Qi

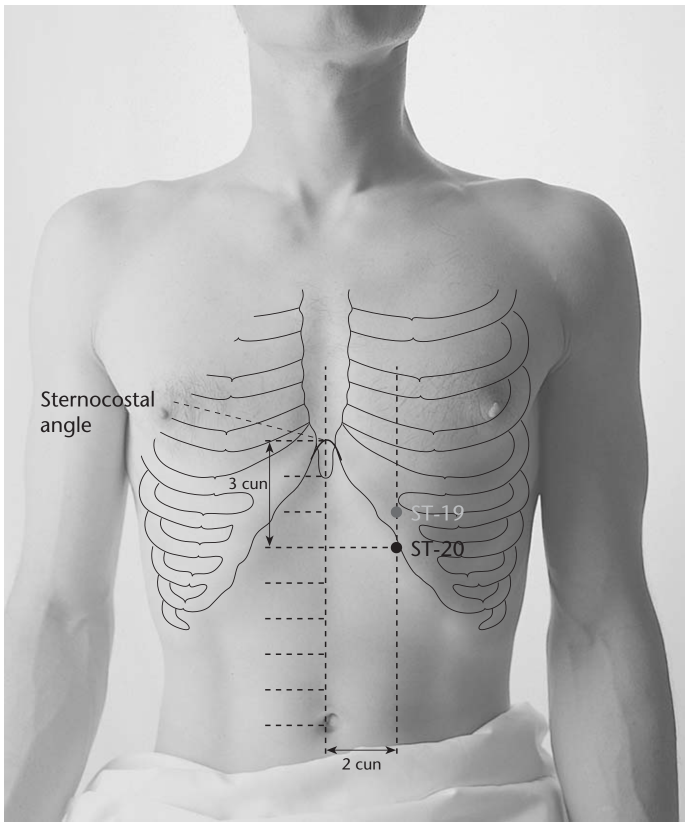
Sternocostal angle
3 cun
ST-19
ST-20
2 cun

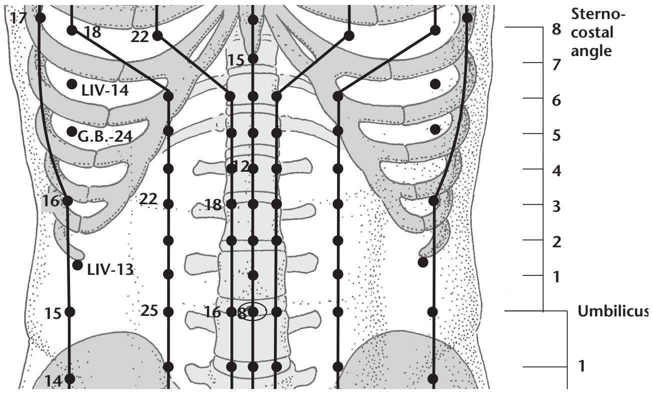
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

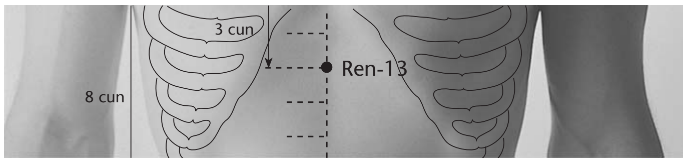
3 cun
Ren-13
8 cun

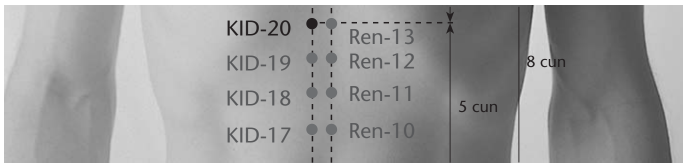
KID-20
KID-19
KID-18
KID-17
Ren-13
Ren-12
Ren-11
Ren-10
5 cun
8 cun

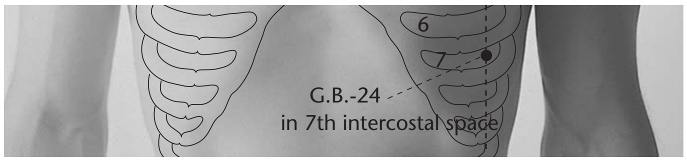
6
7
G.B.-24
in 7th intercostal space

Beam Gate LIANGMEN ST-21

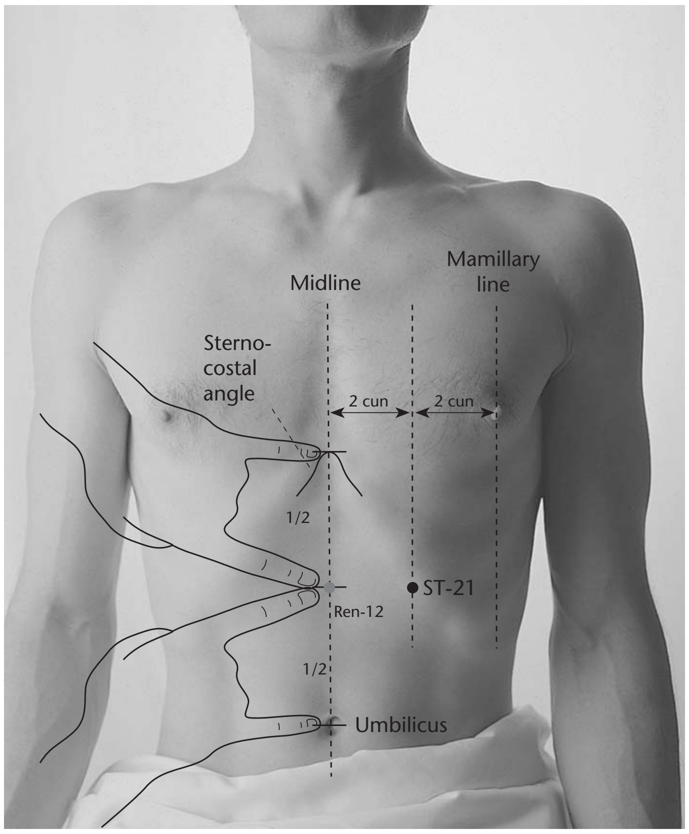
Midline
Mamillary line
Sterno-costal angle
2 cun
2 cun
1/2
Ren-12
ST-21
1/2
Umbilicus

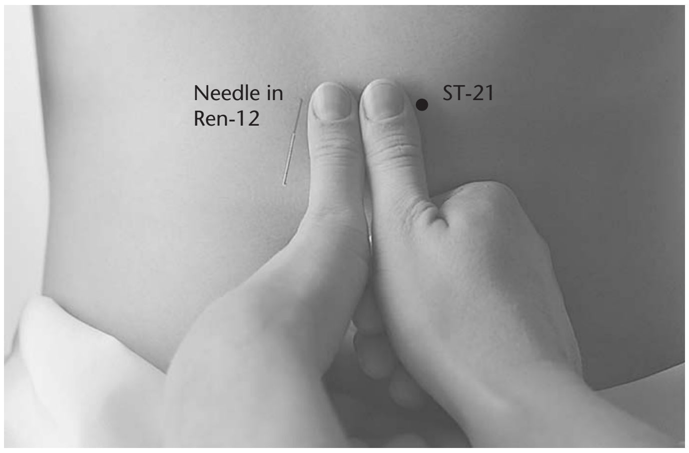
Needle in Ren-12
ST-21

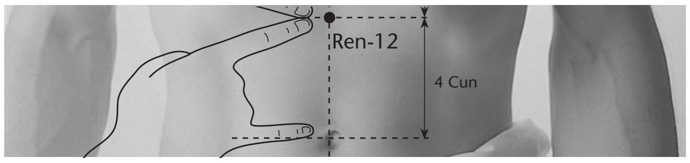

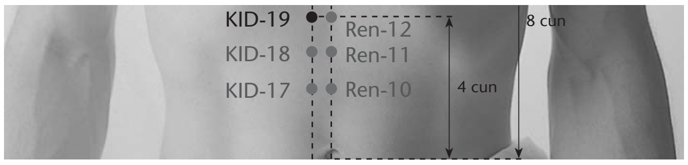
KID-19
KID-18
KID-17
Ren-12
Ren-11
Ren-10
4 cun
8 cun

Location

2 cun lateral to the anterior midline, 4 cun above the umbilicus (or 4 cun below the subcostal angle).

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). **Quick method:** Find the midpoint between the umbilicus and the sternocostal angle (→ **Ren-12**) by using the spreading hands technique (→ 2.3.3), then measure 2 cun in a lateral direction to locate **ST-21**. Depending on the shape of the thorax, this can be located on the lower rib cage.

Located on the same level are → **Ren-12** (anterior midline) and → **KID-19** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. If the correct location projects this point on the lower ribcage, needle transversely on the rib or locate the point more medially (or choose an alternative point). Caution: Peritoneum, enlarged liver (on the right), pregnancy.

Actions/Indications

- Regulates the Qi, harmonises the Middle Burner, eliminates stagnation, raises the Qi and stops diarrhoea

Special features

Important local point for distension and pain in the epigastrium and the lateral abdomen

ST-22 *Pass Gate* GUANMEN

Location

3 cun above the umbilicus (or 5 cun inferior to the sternocostal angle) and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 3 cun up from the umbilicus. **Or:** Measure 1 cun downward from the midpoint of the distance between the umbilicus and the sternocostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). **ST-22** is located 2 cun lateral from here.

Located on the same level are → **Ren-11** (anterior midline), → **KID-18** (0.5 cun lateral to the anterior midline) and → **SP-16** (4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and Intestines
- Regulates the water passages

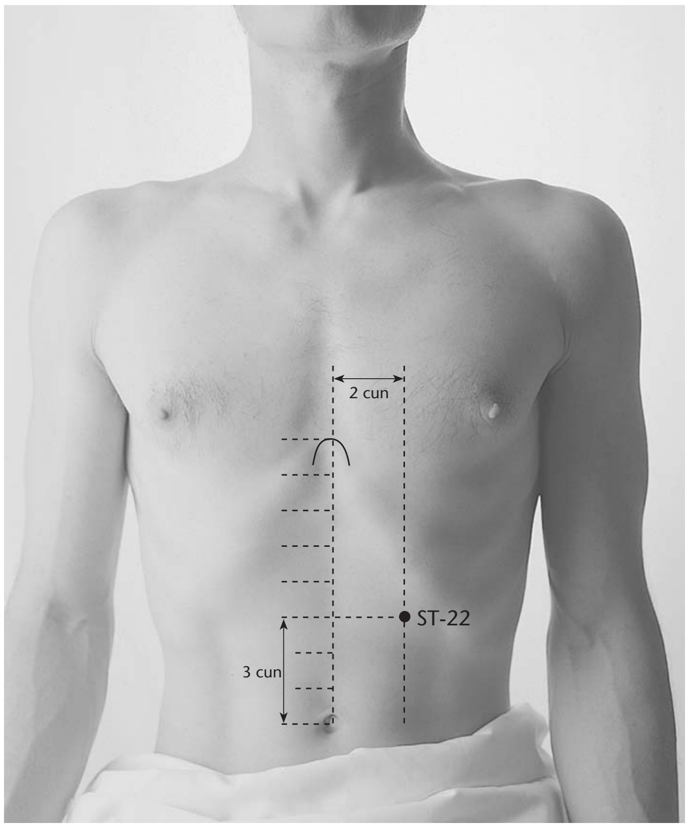
2 cun
ST-22
3 cun

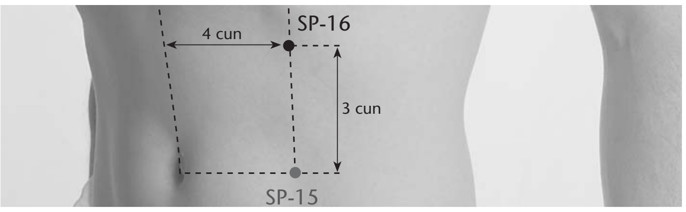
4 cun
SP-16
3 cun
SP-15

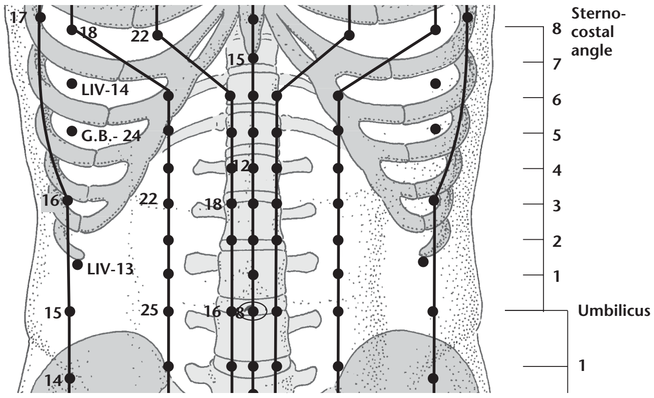
17
18
22
15
LIV-14
G.B.- 24
12
16
22
18
LIV-13
15
25
16
8
14
8
Sterno-
costal
angle
7
6
5
4
3
2
1
Umbilicus
1

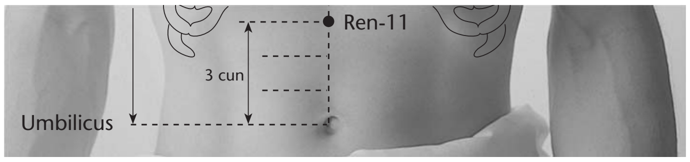
Ren-11
3 cun
Umbilicus

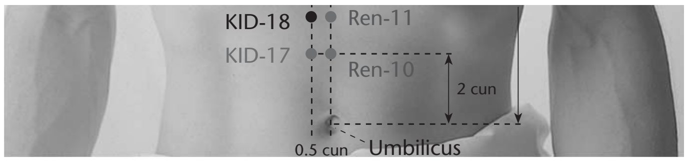
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

Supreme Unity TAIYI **ST-23**

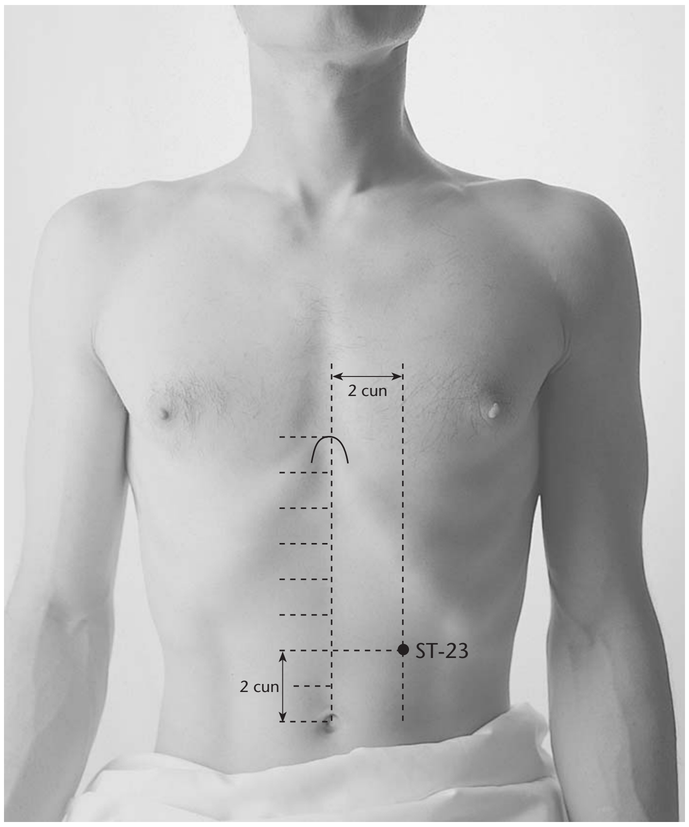

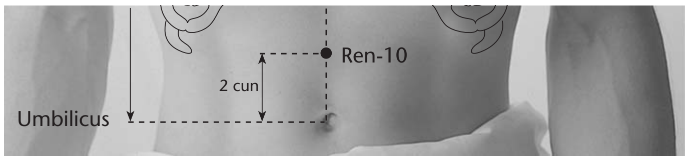

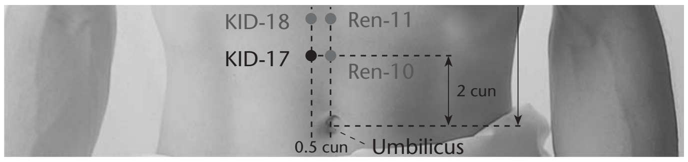

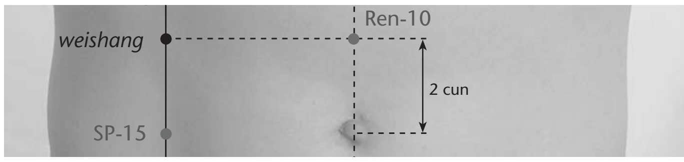

Location

2 cun above the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). First, measure 2 cun up from the umbilicus. **Or:** Measure the midpoint of the lower half of the distance between the umbilicus and the subcostal angle (→ **Ren-12**: spreading hands technique → 2.3.3). **ST-23** is located 2 cun lateral from here.

Located on the same level are → **Ren-10** (anterior midline), → **KID-17** (0.5 cun lateral to the anterior midline) and → **Ex-CA** (*weishang*; 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Harmonises the Middle Burner, strengthens the Spleen
- Benefits the Heart and calms the *shen*

ST-24 *Slippery Flesh Gate* HUAROUMEN

Location

1 cun above the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). **ST-24** is located 1 cun above the umbilicus and 2 cun lateral to the anterior midline.
Located on the same level is → **Ren-19** (anterior midline); often, also → **LIV-13** (at the free end of the 11th rib) will be located at this level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach
- Transforms Phlegm and calms the *shen*

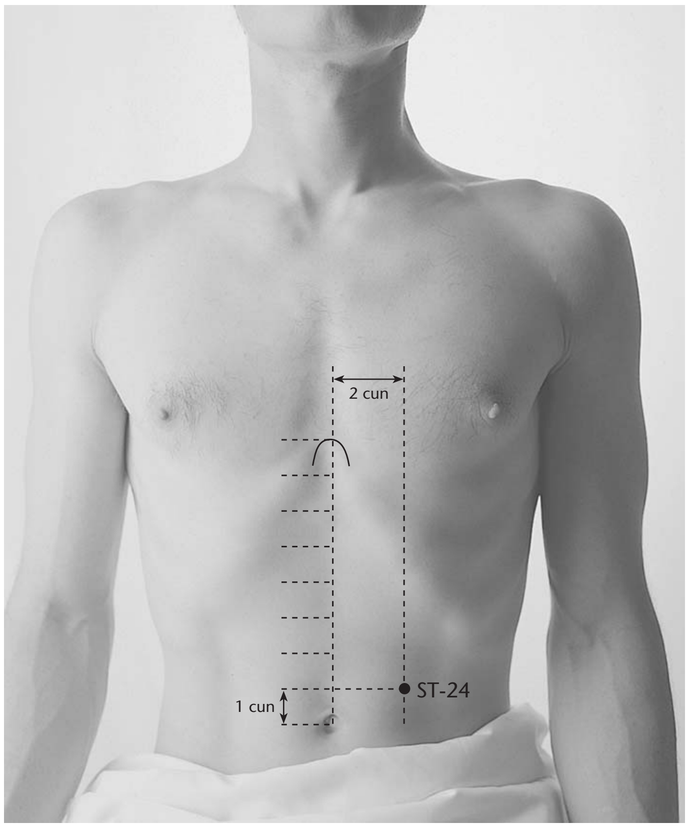
2 cun
ST-24
1 cun

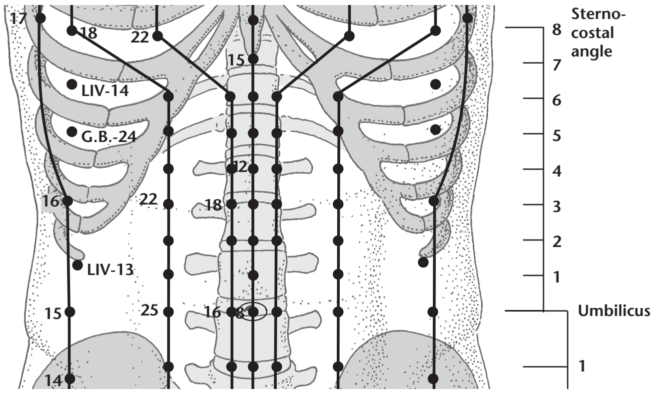
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

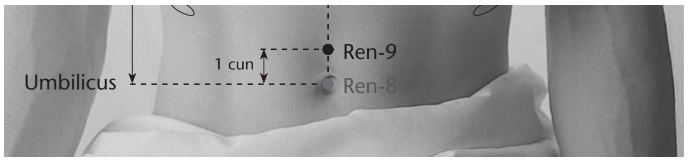
Umbilicus
1 cun
Ren-9
Ren-8

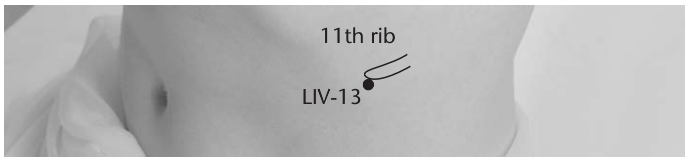
11th rib
LIV-13

Heaven's Pivot TIANSHU ST-25

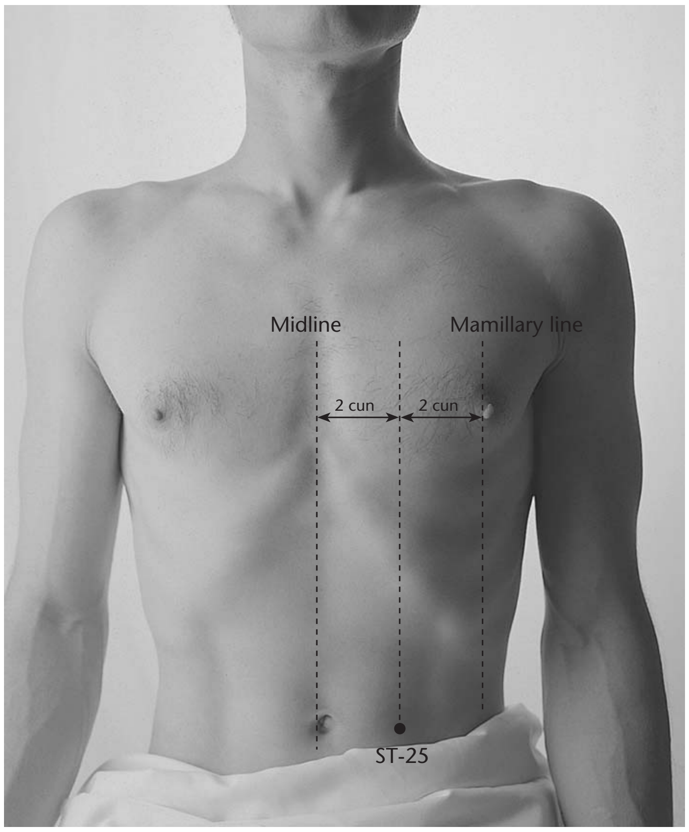

Location

2 cun lateral to the umbilicus.

How to find

Measure 2 cun lateral from the centre of the umbilicus. **For reference:** On the abdomen, 2 cun lateral corresponds to the midpoint between the anterior midline and the mamillary line (→ 3.5). Located on the same level are → **Ren-8/KID-16/SP-15/G.B.-26** (umbilicus/0.5 cun lateral to the umbilicus/mamillary line or 4 cun lateral to the umbilicus/at the level of the umbilicus and directly inferior to the free end of the 11th rib).

Needling

Vertically 0.5–1.5 cun. For gynaecological disorders, obliquely in an inferior direction towards the uterus. Caution: Peritoneum, pregnancy. The needle may reach the aponeuroses of the oblique abdominal musculature at the transition to the sheath of the rectus abdominis muscle.

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, eliminates Dampness and clears Damp-Heat
- Eliminates stagnation

Special features

Front-*mu* point of the Large Intestine. Important point for disorders of the gastrointestinal tract.

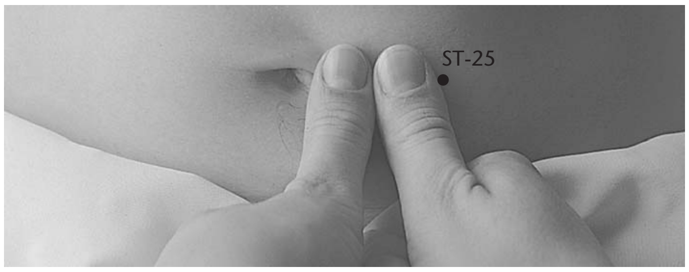

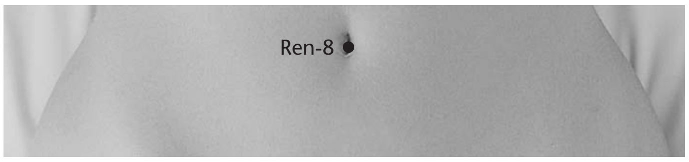

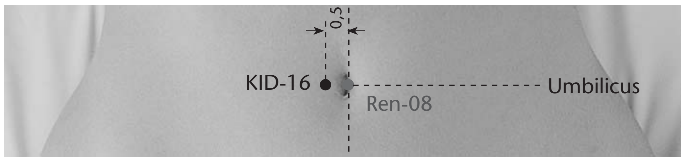

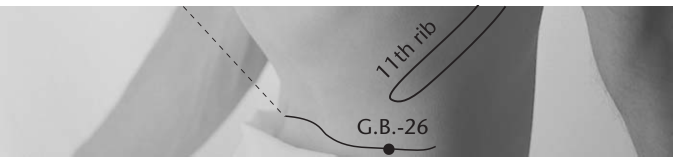

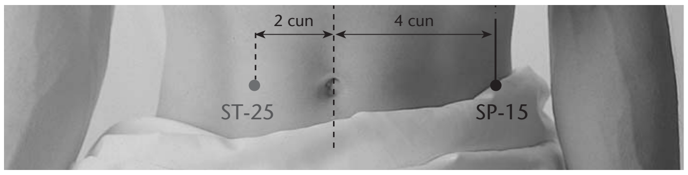

ST-26 *Outer Mound* WAILING

Location

1 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape ➞ 2.3.1). From the umbilicus, measure 1 cun down: **ST-26** is located 2 cun laterally from there.
Located on the same level are ➞ **Ren-7** (anterior midline) and ➞ **KID-15** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Qi, alleviates pain and Cold

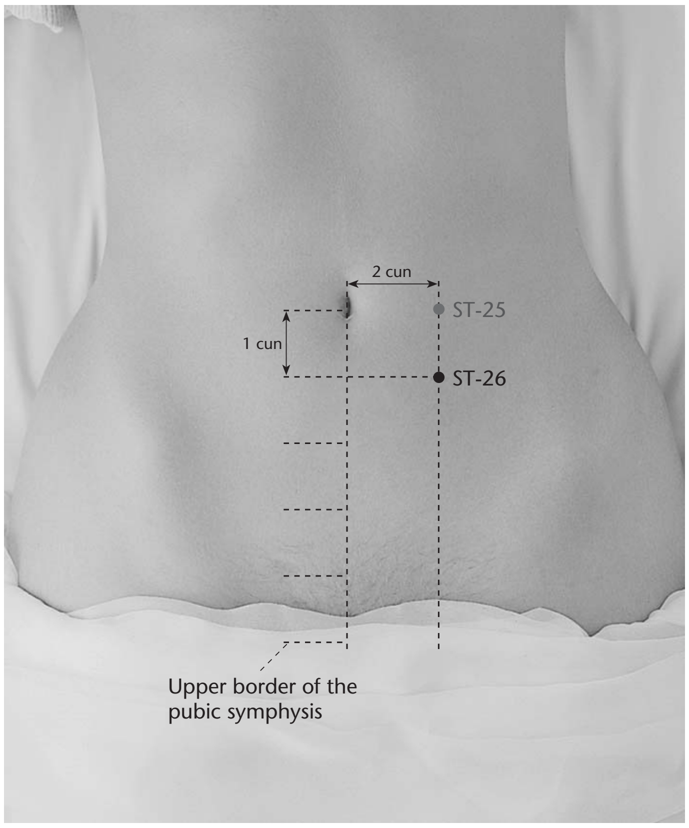
2 cun
ST-25
1 cun
ST-26
Upper border of the pubic symphysis

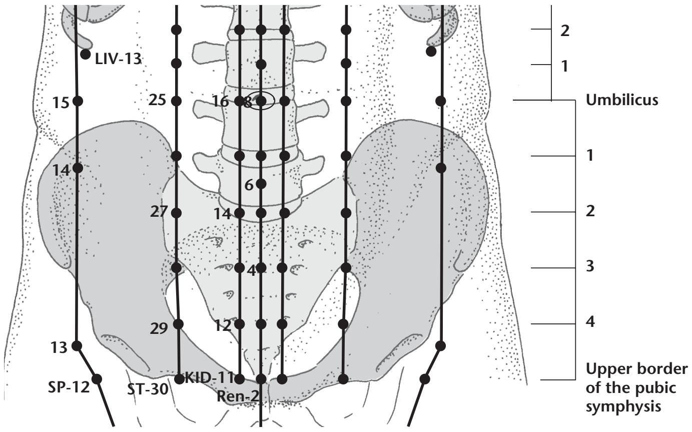
LIV-13
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

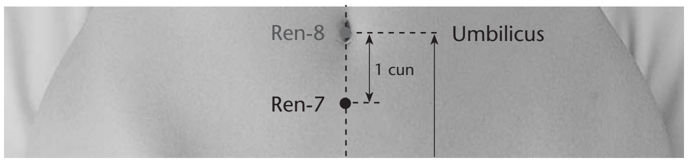
Ren-8
Umbilicus
1 cun
Ren-7

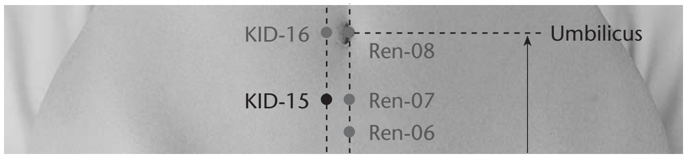
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

The Great DAJU ST-27

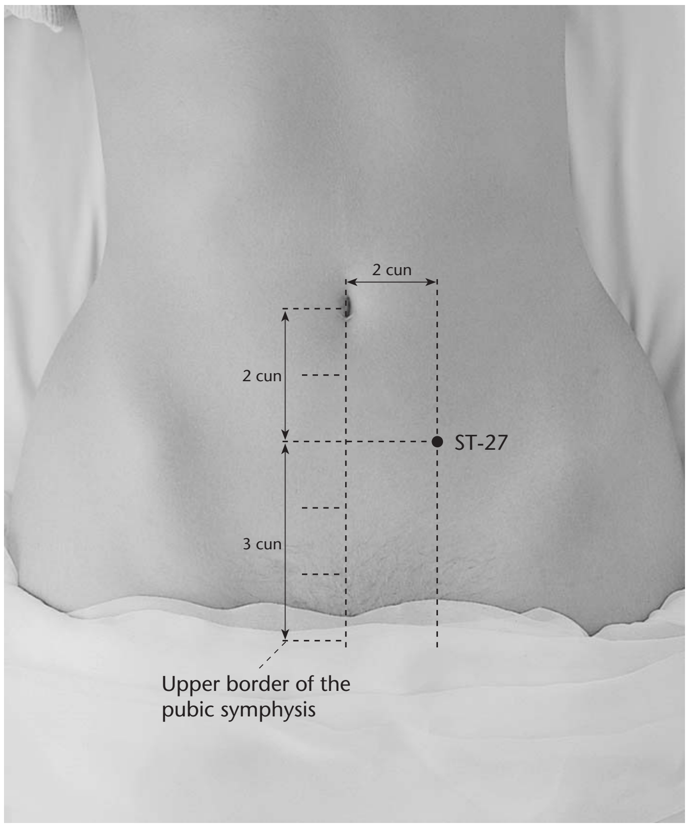

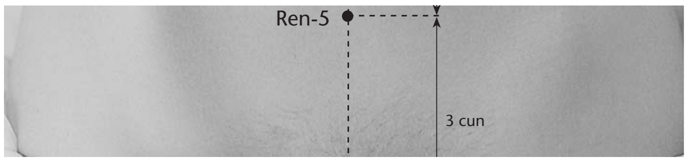

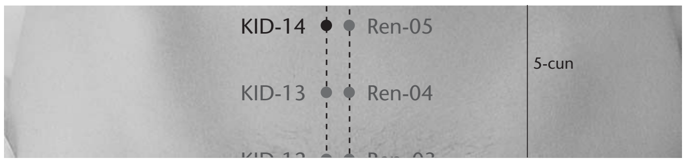

Location

2 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the umbilicus, measure 2 cun down and 2 cun laterally, where **ST-27** is located.

Located on the same level are → **Ren-5** (anterior midline) and → **KID-14** (0.5 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates and strengthens the Qi
- Benefits the Kidneys and promotes diuresis, strengthens the Yang and Essence (*jing*)

ST-28 *Water Passage* SHUIDAO

Location

3 cun below the umbilicus or 2 cun above the upper border of the pubic symphysis and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun up or measure 3 cun down from the umbilicus. **ST-28** is located 2 cun laterally from there.

Located on the same level are → **Ren-4** (anterior midline), → **KID-13** (0.5 cun lateral to the anterior midline) and → **Ex-CA** (*tituo/quimen/yijing*; 4/3/1 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, urinary bladder, pregnancy.

Actions/Indications

- Clears Damp-Heat, benefits the Lower Burner and eliminates stagnation

Special features

Important local point for the urogenital tract

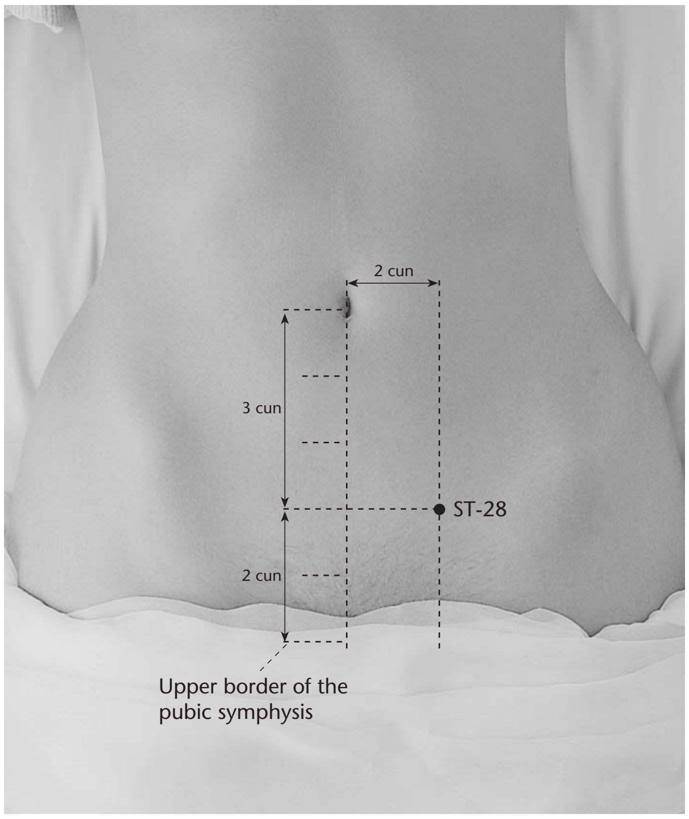
2 cun
3 cun
ST-28
2 cun
Upper border of the pubic symphysis

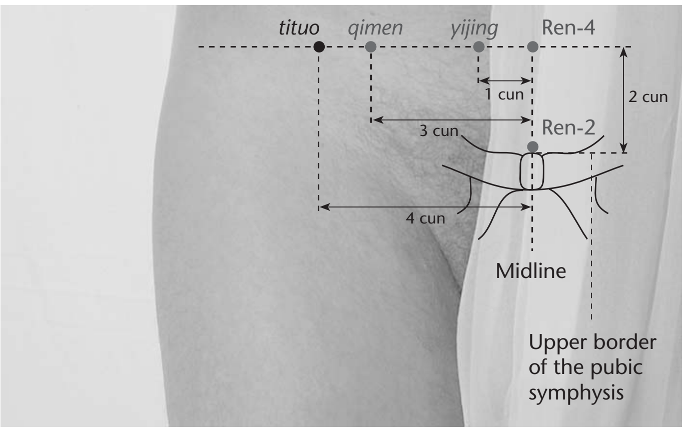
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

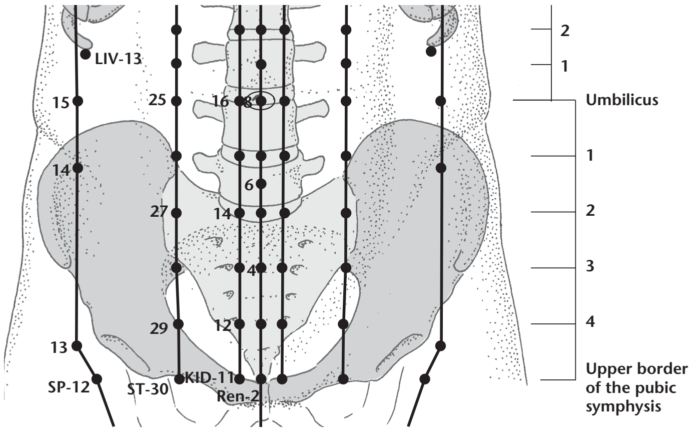
LIV-13
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

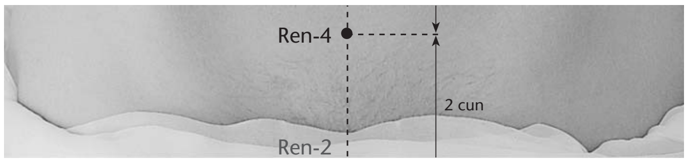
Ren-4
2 cun
Ren-2

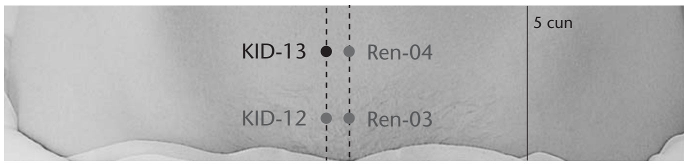
5 cun
KID-13
Ren-04
KID-12
Ren-03

Return GUILAI ST-29

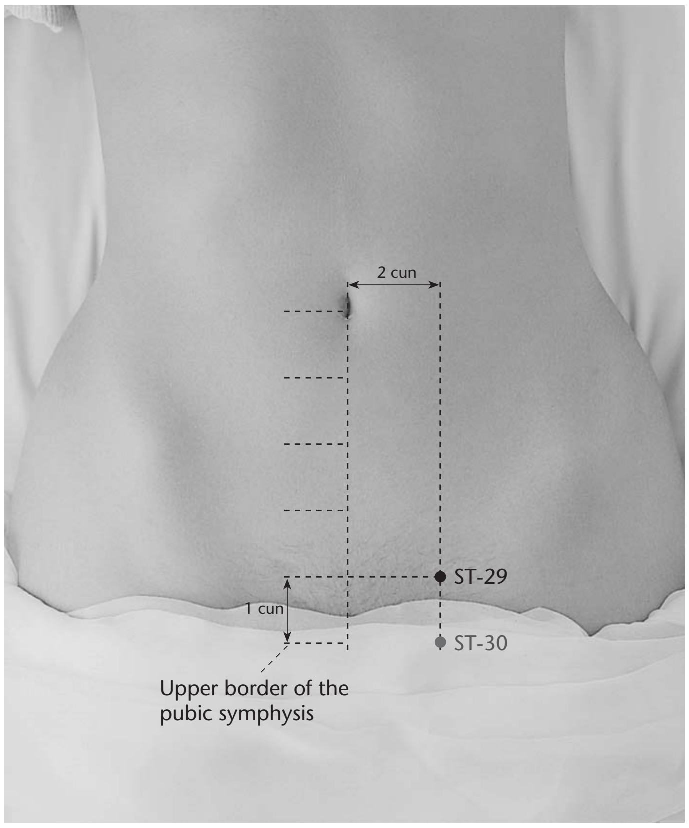
2 cun
ST-29
1 cun
ST-30
Upper border of the pubic symphysis

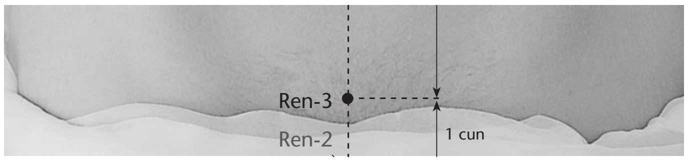
Ren-3
Ren-2
1 cun

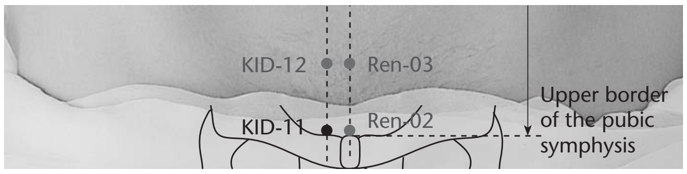
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

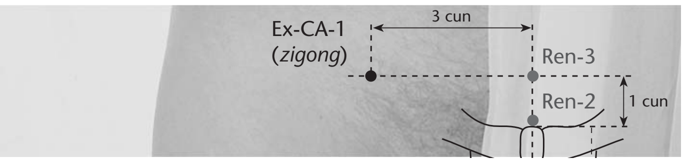
Ex-CA-1 (*zigong*)
3 cun
Ren-3
Ren-2
1 cun

Location

1 cun above the upper border of the pubic symphysis or 4 cun below the umbilicus and 2 cun lateral to the anterior midline.

How to find

The distance between the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun; therefore, only proportional cun measurements should be used (helpful tool: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun up and 2 cun lateral to **ST-29**.

Located on the same level are → **Ren-3** (anterior midline), → **KID-12** (0.5 cun lateral to the anterior midline) and → **Ex-CA** (*zigong*; 3 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, urinary bladder, pregnancy.

Actions/Indications

- Warms and regulates the Lower Burner and menstruation

Special features

Commonly treated with moxibustion, the combination of acupuncture and moxibustion being particularly beneficial. Important local point for the urogenital tract.

ST-30 *Rushing Qi* QICHONG

Location

2 cun lateral to the upper border of the pubic symphysis and medial to the femoral artery and vein, approximately 1 cun superior to the inguinal groove.

How to find

First, locate the upper border of the pubic symphysis (location of → **Ren-2**). Then locate **ST-30** 2 cun lateral from there (on the vertical line halfway between the anterior midline and the mamillary line). This often corresponds with a location 1 cun superior to the inguinal groove and medial to the femoral artery/vein.
Located on the same level are → **Ren-2/KID-11/SP-12** (anterior midline/0.5 cun lateral to the anterior midline/3.5 cun lateral to the anterior midline). → **LIV-12** is located in the inguinal groove, 1 cun inferior and 0.5 cun lateral to **ST-30**.

Needling

0.5–1 cun vertically or 1.5 cun slightly obliquely towards the external genitalia for disorders of the urogenital tract. Caution: Femoral artery, urinary bladder (quite superficial in a superior direction); in men, the spermatic ducts (quite superficial in an inferior direction), pregnancy. Ask the patient to empty their bladder before needling this point.

Actions/Indications

- Regulates the Qi in the Lower Burner
- Strengthens the Sea of Grain (later Heaven Qi)

Special features

Meeting point with the *chong mai* (according to some authors, also with the G.B. channel), Sea of Grain point. Wide usage for disorders of the Lower Burner.

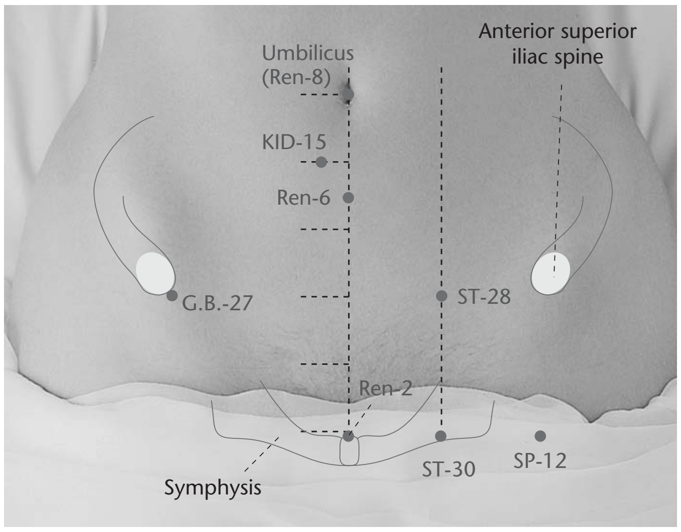
Umbilicus
(Ren-8)
Anterior superior
iliac spine
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

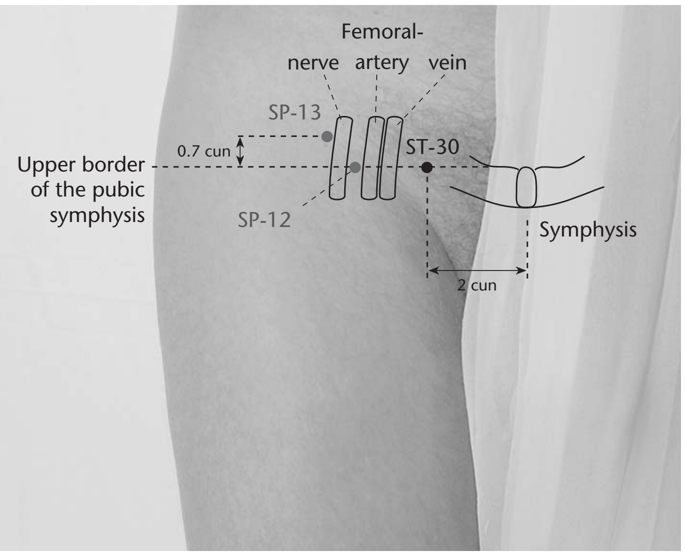
Femoral-
nerve
artery
vein
SP-13
0.7 cun
ST-30
Upper border
of the pubic
symphysis
SP-12
Symphysis
2 cun

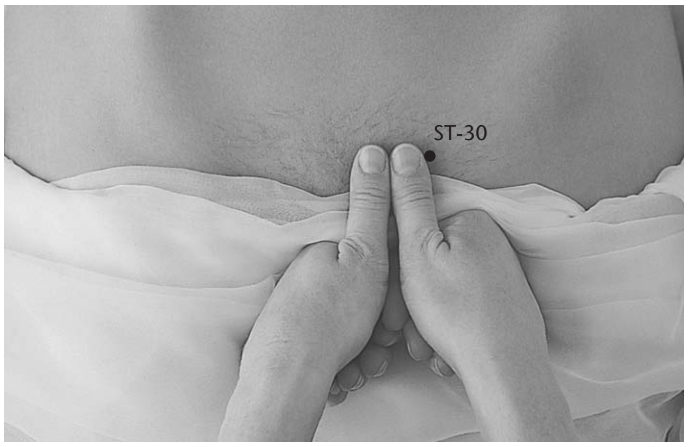

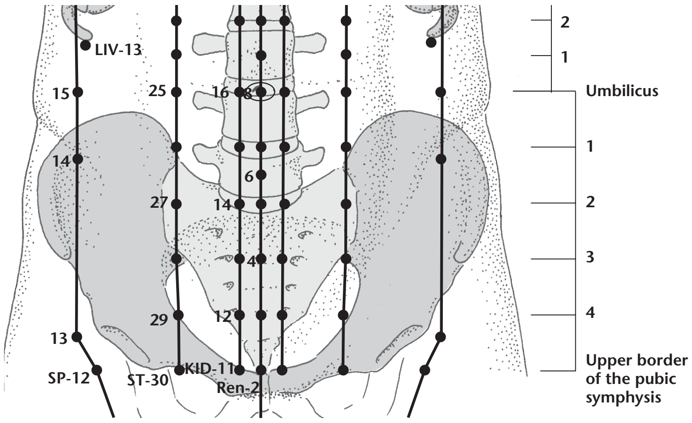
LIV-13
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
2
1
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

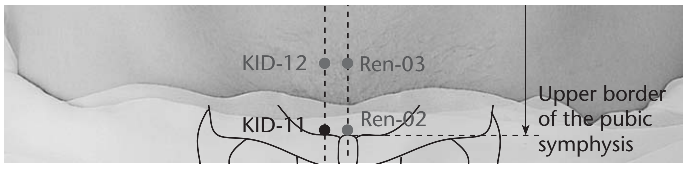
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

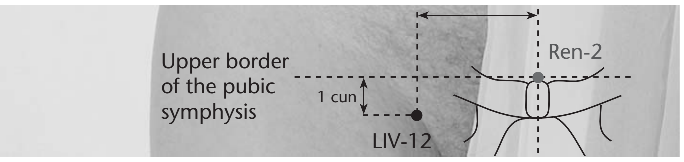
Upper border
of the pubic
symphysis
1 cun
Ren-2
LIV-12

Thigh Gate BIGUAN ST-31

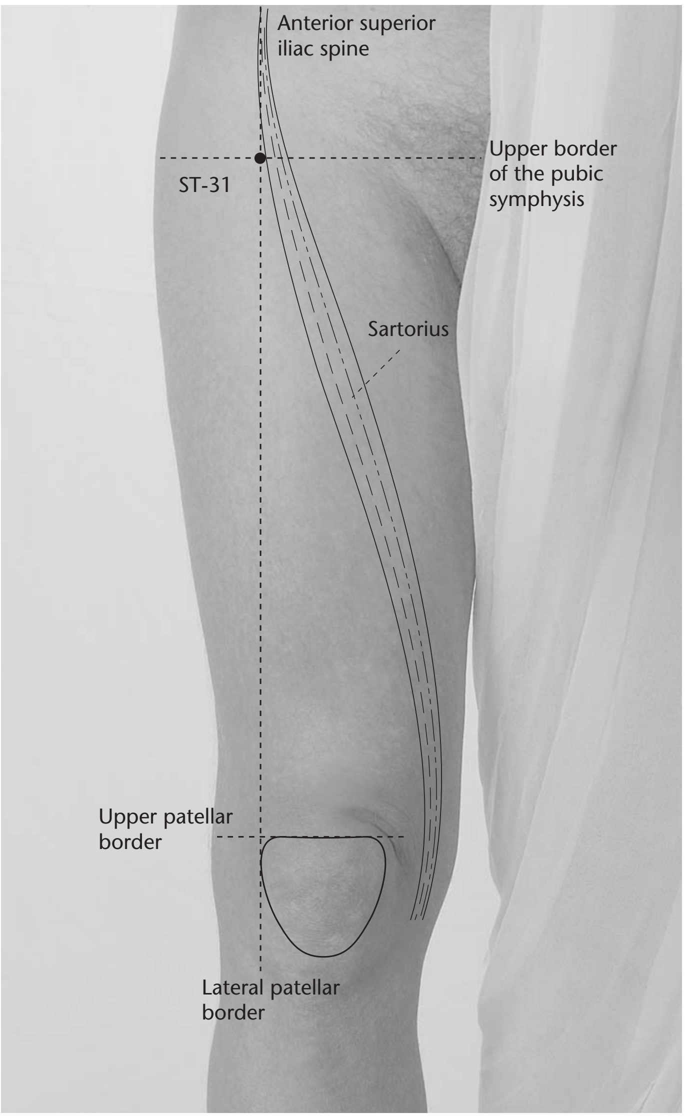
Anterior superior
iliac spine
ST-31
Upper border
of the pubic
symphysis
Sartorius
Upper patellar
border
Lateral patellar
border

Location

Inferior to the anterior superior iliac spine and lateral to the sartorius muscle, at the level of the lower border of the pubic symphysis.

How to find

The sartorius muscle originates at the anterior superior iliac spine (ASIS) and crosses the thigh towards the medial aspect of the knee. With the knee and hip slightly flexed and with active lateral rotation, it can be isolated easily. **ST-31** is located on the lateral border of the muscle, at the crossing point of a vertical line connecting the ASIS and the lateral upper border of the patella and a horizontal line at the level of the lower border of the pubic symphysis.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Opens the channel, warms and expels pathogenic factors

Special features

Particularly useful for Cold *bi*-syndrome, disorders of the knees and lower back. The chain-and-lock method in combination (→ 8) with → **ST-36** and → **ST-41** is often used for paraesthesia and pain in the hip and leg radiating along the channel.

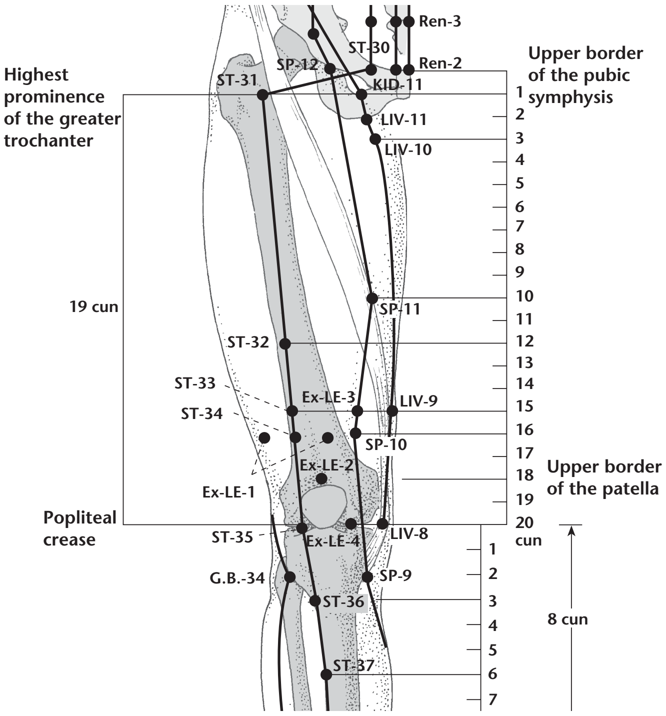
Ren-3
ST-30
SP-12
Ren-2
Upper border
of the pubic
symphysis
Highest
prominence
of the greater
trochanter
ST-31
KID-11
LIV-11
LIV-10
19 cun
SP-11
ST-32
ST-33
Ex-LE-3
LIV-9
ST-34
SP-10
Ex-LE-2
Upper border
of the patella
Ex-LE-1
Popliteal
crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
8 cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
cun
1
2
3
4
5
6
7

ST-32 *Crouching Rabbit* FUTU

Location

On the thigh, on a line connecting the lateral patellar border and the anterior superior iliac spine, 6 cun superior to the upper, lateral border of the patella.

How to find

Divide the distance between the tip of the greater trochanter and the lateral upper border of the patella into three sections. **ST-32** is located at the junction of the lower and the middle third of the thigh in a depression on the quadriceps muscle.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Opens the channel and expels Wind, Dampness and Cold

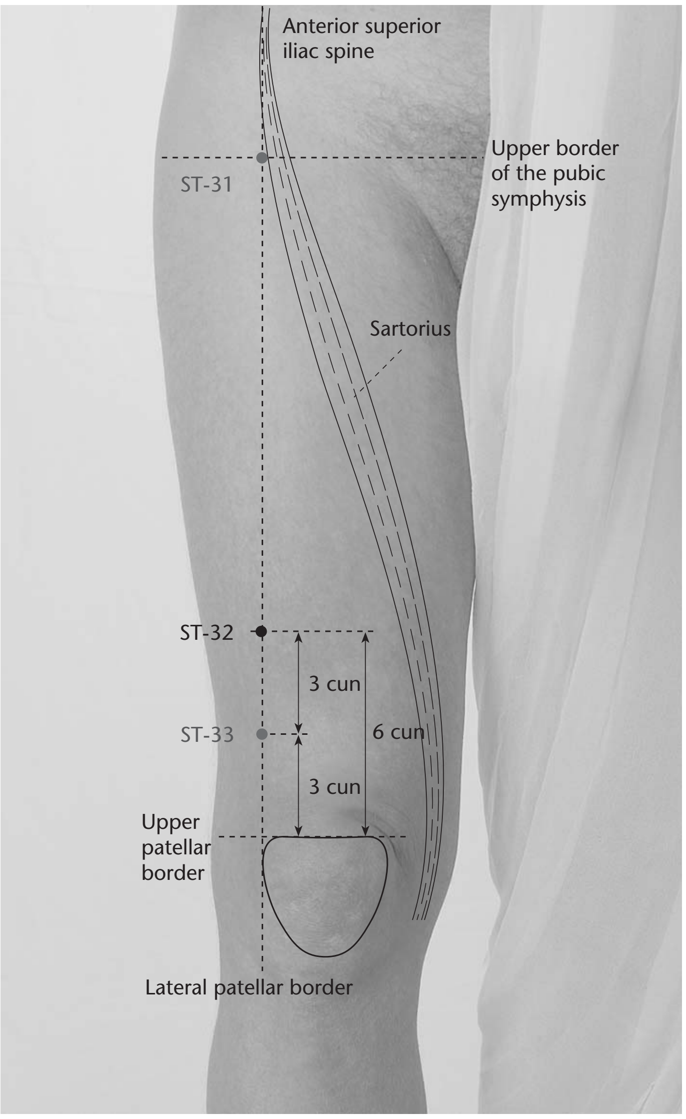
Anterior superior iliac spine
ST-31
Upper border of the pubic symphysis
Sartorius
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
Lateral patellar border

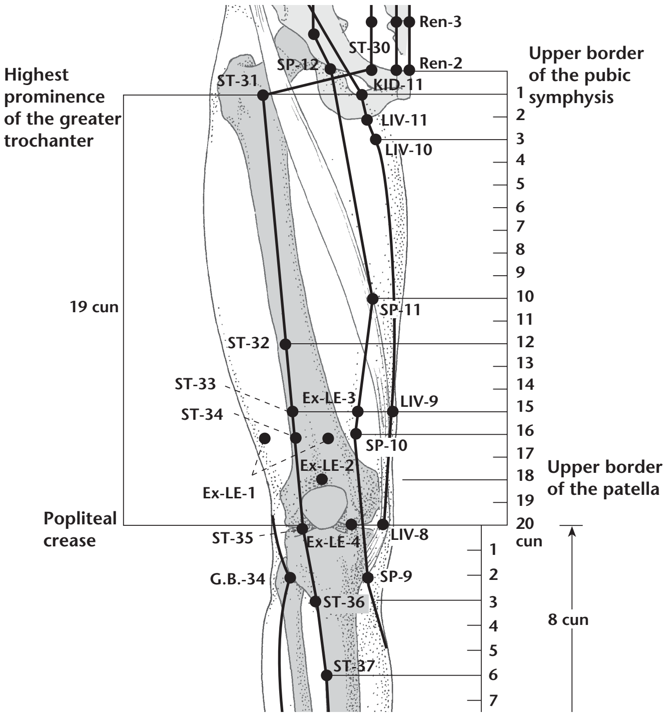
Ren-3
ST-30
Ren-2
SP-12
ST-31
KID-11
Highest prominence of the greater trochanter
Upper border of the pubic symphysis
LIV-11
LIV-10
19 cun
SP-11
ST-32
ST-33
Ex-LE-3
LIV-9
ST-34
SP-10
Ex-LE-2
Ex-LE-1
Upper border of the patella
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
8 cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20 cun
1
2
3
4
5
6
7

Yin Market YINSHI ST-33

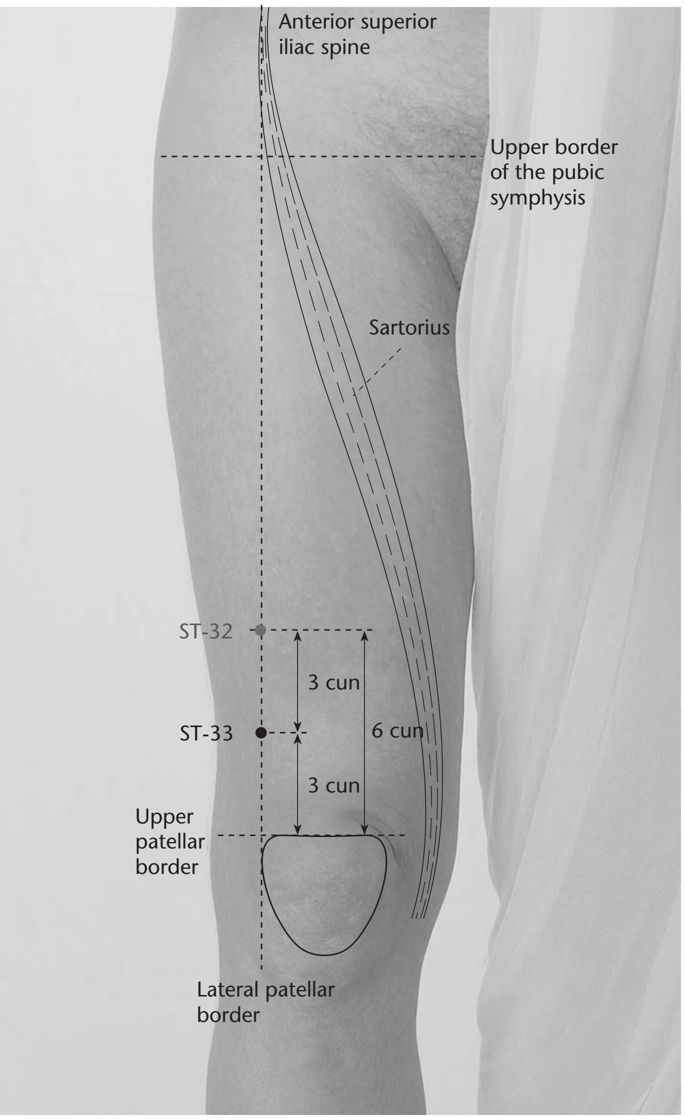
Anterior superior iliac spine
Upper border of the pubic symphysis
Sartorius
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar border
Lateral patellar border

Location

On a line joining the lateral patellar border and the anterior superior iliac spine, 3 cun superior to the upper lateral border of the patella.

How to find

Measure 3 cun up from the upper lateral border of the patella, on the connecting line to the anterior superior iliac spine. **ST-33** is located in a groove at the junction of the rectus femoris and vastus lateralis muscles.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel and expels pathogenic factors

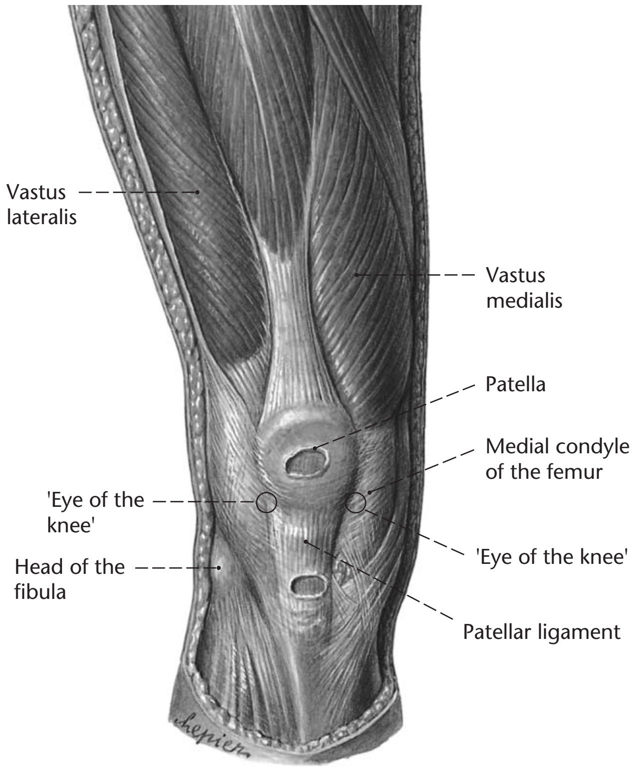
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

ST-34 *Ridge Mound* LIANGQIU

Location

On a line joining the lateral patellar border and the anterior superior iliac spine, 2 cun proximal to the upper lateral border of the patella, in a groove of the vastus lateralis muscle.

How to find

Locate the lateral upper border of the patella and from there measure 2 cun proximally. **ST-34** is located in a palpable groove on the vastus lateralis muscle of the quadriceps femoris group. This point projects on the line joining the lateral upper border of the patella and the anterior superior iliac spine (ASIS → 3.5).
Located on the same level are the points → **Ex-LE-1** (each 1.5 cun lateral and medial to **ST-34**). → **SP-10** is located in a comparable position on the medial aspect of the thigh (2 cun proximal to the medial upper border of the patella and in a groove on the vastus medialis muscle).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Regulates the Stomach Qi, alleviates acute conditions
- Opens the channel and alleviates pain

Special features

Xi-cleft point. Important local point for disorders of the knee. Modern application for acute epigastric pain – strongly reducing needling techniques have an immediate and calming effect on gastrointestinal peristalsis.

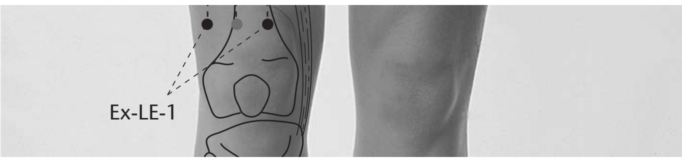

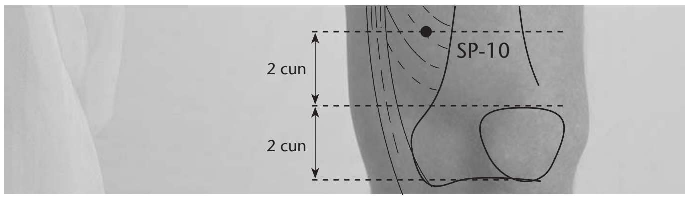
2 cun
2 cun
SP-10

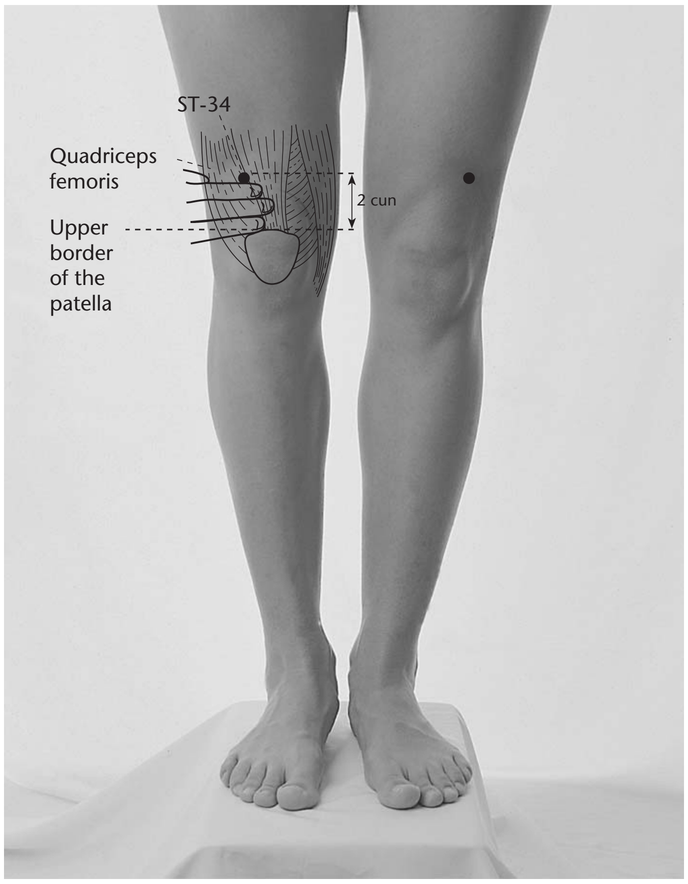
ST-34
Quadriceps femoris
2 cun
Upper border of the patella

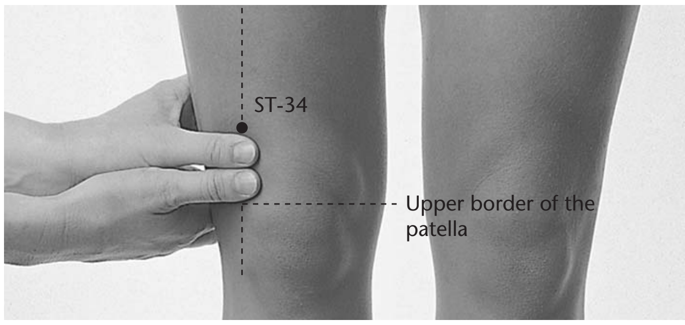
ST-34
Upper border of the patella

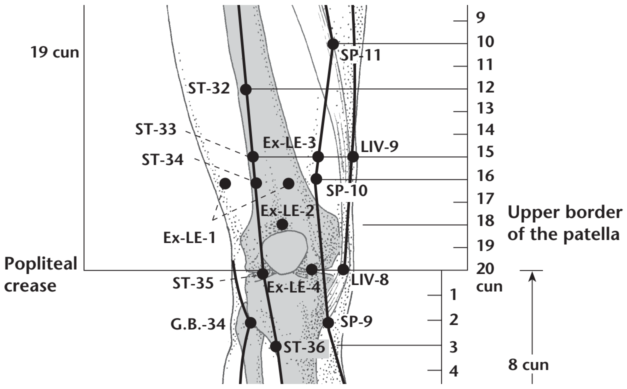
19 cun
ST-32
ST-33
ST-34
Ex-LE-1
Popliteal crease
ST-35
G.B.-34
SP-11
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-4
LIV-8
SP-9
ST-36
9
10
11
12
13
14
15
16
17
18
19
20 cun
1
2
3
4
Upper border of the patella
8 cun

Ch04.3-F10028.qxd 2/22/08 8:43 PM Page 163

Calf's Nose DUBI ST-35

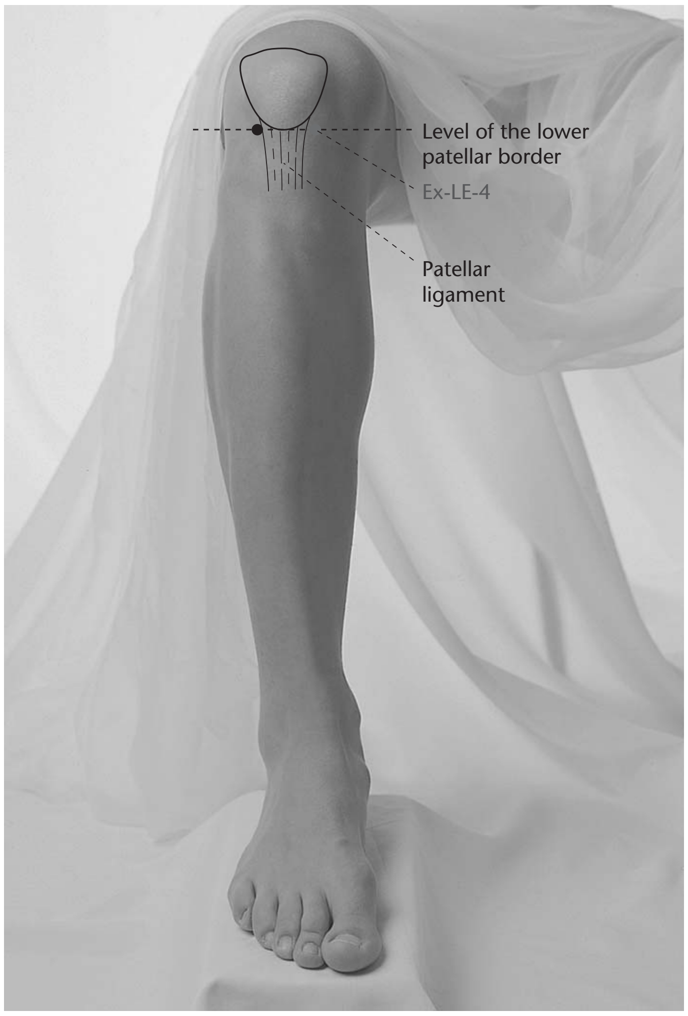
Level of the lower patellar border
Ex-LE-4
Patellar ligament

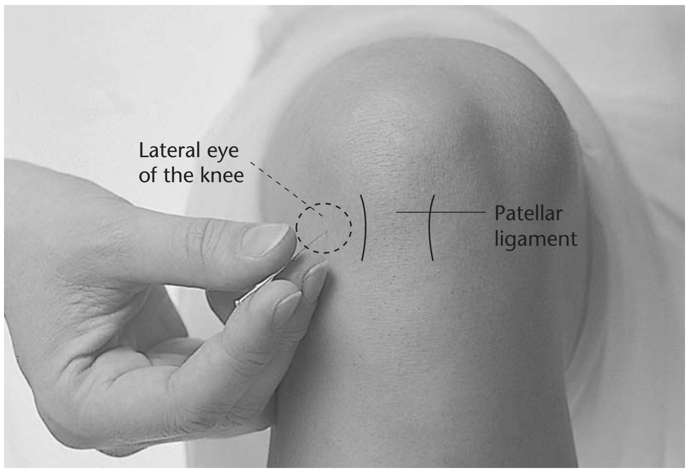
Lateral eye of the knee
Patellar ligament

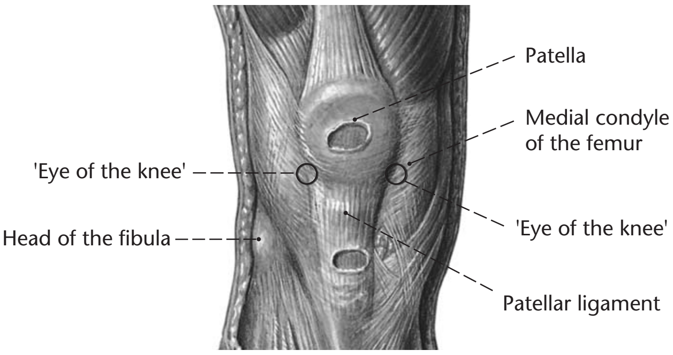
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

Location

With the knee flexed, in the depression inferior to the patella and lateral to the patellar ligament.

How to find

This point is best located and needled with the knee slightly flexed (use knee support). **ST-35** is located in a depression on the lower lateral border of the patella, in a groove lateral to the patellar ligament. It projects onto the level of the knee joint space and more or less corresponds with the arthroscopic access to the joint. **ST-35** is also known as 'lateral eye of the knee' and is part of the extra point → **Ex-LE-5** (*xiyan*).

Needling

0.5–1 cun slightly obliquely in a medial direction, towards the medial eye of the knee → **Ex-LE-4** (*neixiyan*). Caution: Knee joint (the location of this point corresponds approximately with the arthroscopic access to this joint!).

Actions/Indications

- Expels Wind and Dampness, reduces swellings, opens the channel, alleviates pain

Special features

Important local point for all disorders of the knee

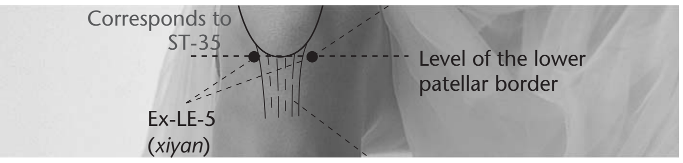
Corresponds to ST-35
Level of the lower patellar border
Ex-LE-5 (xiyan)

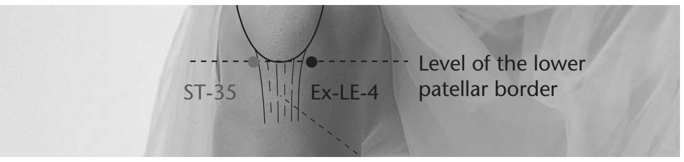
ST-35
Ex-LE-4
Level of the lower patellar border

ST-36 *Leg Three Miles* ZUSANLI

Location

3 cun distal to → **ST-35** ('lateral eye of the knee') and 1 fingerbreadth lateral to the anterior crest of the tibia, on the tibialis anterior muscle.

How to find

First, locate → **ST-35** ('lateral eye of the knee', at the level of the knee joint space). From → **ST-35**, palpate 3 cun (1 handbreadth) down and 1 fingerbreadth towards lateral by placing the middle finger adjacent to the tibial crest. There, **ST-36** is located in a depression, which can be palpated 'dynamically' (choose the point according to pressure sensitivity). **Or:** Palpate the lower border of the tibial tuberosity and 'dynamically' locate **ST-36** 1 fingerbreadth lateral to it.

Needling

Vertically 1–1.5 cun. Moxibustion is often used on this point.

Actions/Indications

- Regulates the Stomach, strengthens the Spleen, transforms Dampness
- Tonifies Qi and Yang, nourishes the Blood and Yin
- Calms the *shen*
- Opens the channel

Special features

He-sea point, Earth point, *ben* point (Five Phases point), lower *he*-sea point of the Stomach, *Gao Wu* command point for disorders of the abdomen, Ma Dan Yang point, Sea of Water and Grain point. Main point to tonify Qi and Blood for general weakness. Very important point.

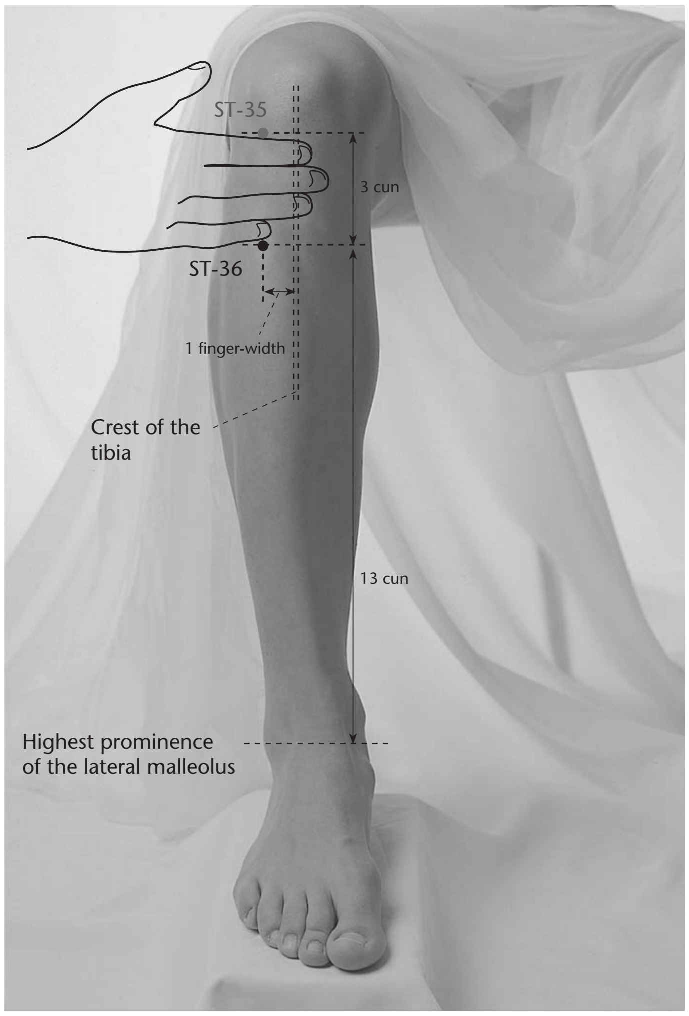
ST-35
3 cun
ST-36
1 finger-width
Crest of the tibia
13 cun
Highest prominence of the lateral malleolus

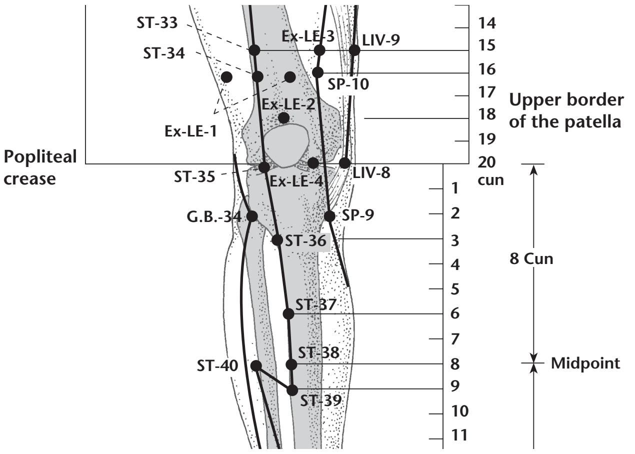
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper border of the patella
1
2
3
4
5
6
7
8
9
10
11
8 Cun
Midpoint

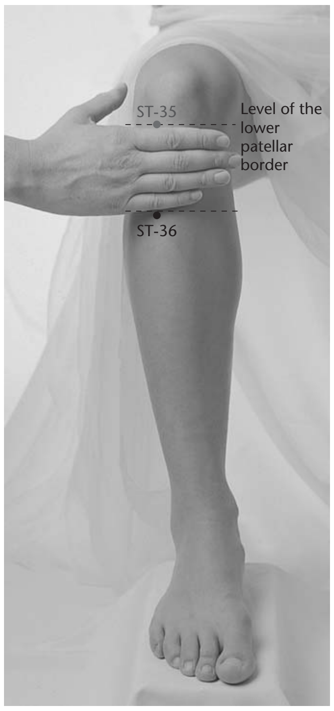
ST-35
Level of the lower patellar border
ST-36

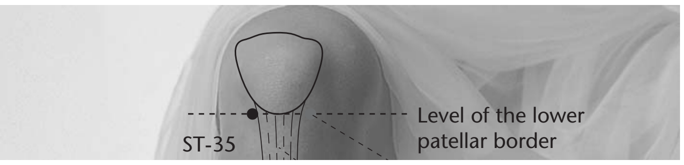
ST-35
Level of the lower patellar border

Upper Great Emptiness SHANGJUXU ST-37

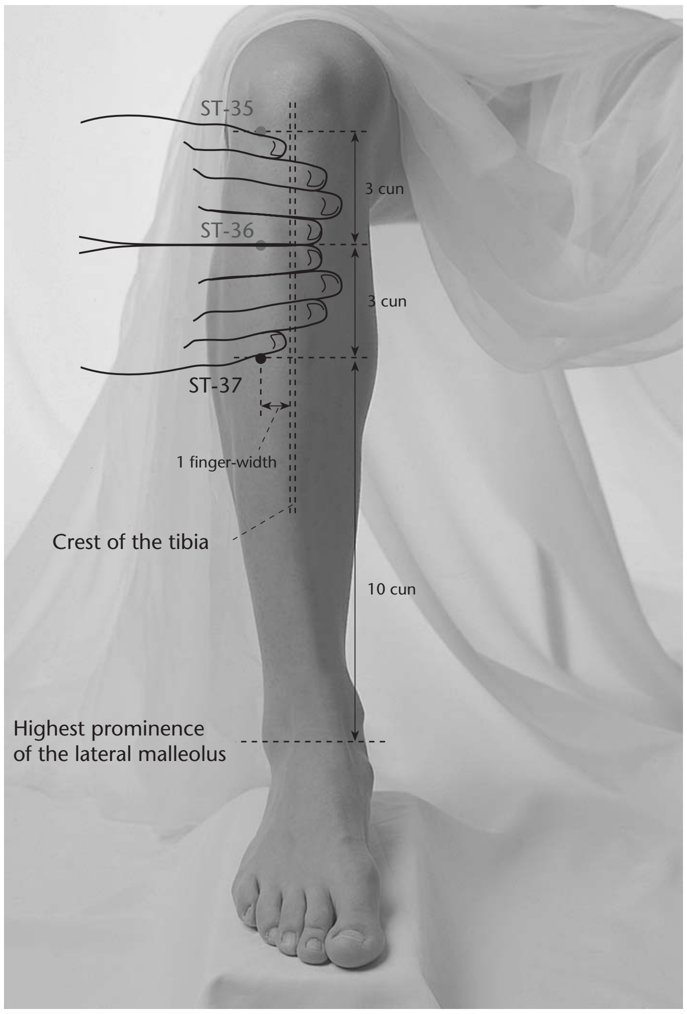
ST-35
3 cun
ST-36
3 cun
ST-37
1 finger-width
Crest of the tibia
10 cun
Highest prominence of the lateral malleolus

Location

6 cun distal to → **ST-35** (at the knee joint space) or 3 cun distal to → **ST-36** and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

First, find → **ST-35** ('lateral eye of the knee', level with the knee joint space). From → **ST-35**, measure 6 cun (2 handbreadths) in a distal direction and, at that level, 1 fingerbreadth towards lateral (place the middle finger onto the anterior crest of the tibia). There, **ST-37** is located in a depression, which can be 'dynamically' palpated (choose the point according to pressure sensitivity). **Or:** Find the midpoint of the line joining → **ST-35** and → **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) by using the spreading hands technique (→ 2.3.3). **ST-37** is located 2 cun proximal and 1 fingerbreadth lateral to this point.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, eliminates stagnation, clears Damp-Heat
- Opens the channel, alleviates pain

Special features

Lower *he*-sea point of the Large Intestine, Sea of Blood point.

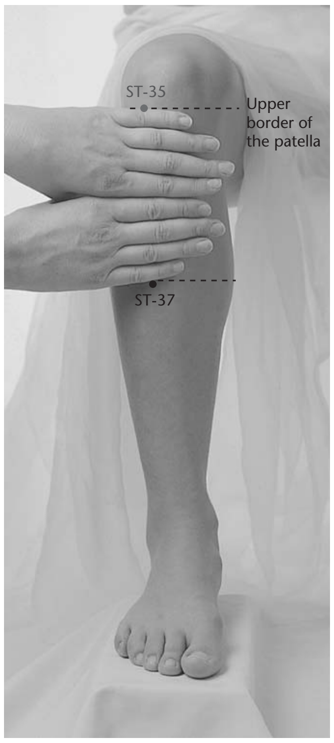
ST-35
Upper border of the patella
ST-37

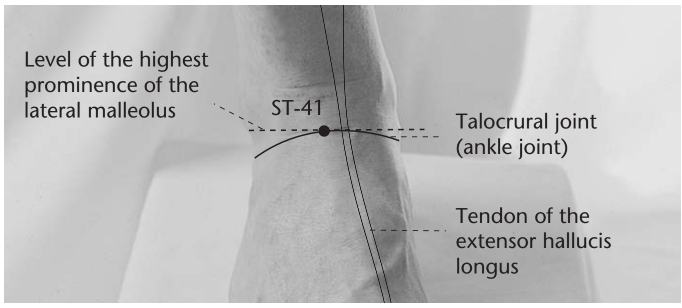
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

ST-38 *Lines Opening* TIAOKOU

Location

At the midpoint of the line joining → **ST-35** and → **ST-41** and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

Quick method: Spreading-hands technique (→ 2.3.3): place the small fingers on → **ST-35** ('lateral eye of the knee', level with the knee joint space) and → **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line, using the thumbs. **ST-38** is located in a depression 1 fingerbreadth (middle fingerbreadth, measured from the anterior crest of the tibia) lateral to the midpoint. **ST-38** may be palpated 'dynamically' (according to pressure sensitivity). **Or:** From → **ST-35**, measure 8 cun (→ 2 handbreadths and 3 fingerbreadths) distally and 1 fingerbreadth laterally.

Located on the same level are → **ST-40** (1 fingerbreadth more laterally) and → **BL-57** (dorsal aspect of the lower leg, on the midpoint of the line joining → **BL-40** and → **BL-60**).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Expels Wind-Damp, opens the channel, alleviates pain, benefits the shoulder

Special features

Very effective distal point for disorders of the shoulder. For limited range of motion of the shoulder joint: apply strong reducing needling techniques while the patient moves their shoulder joint.

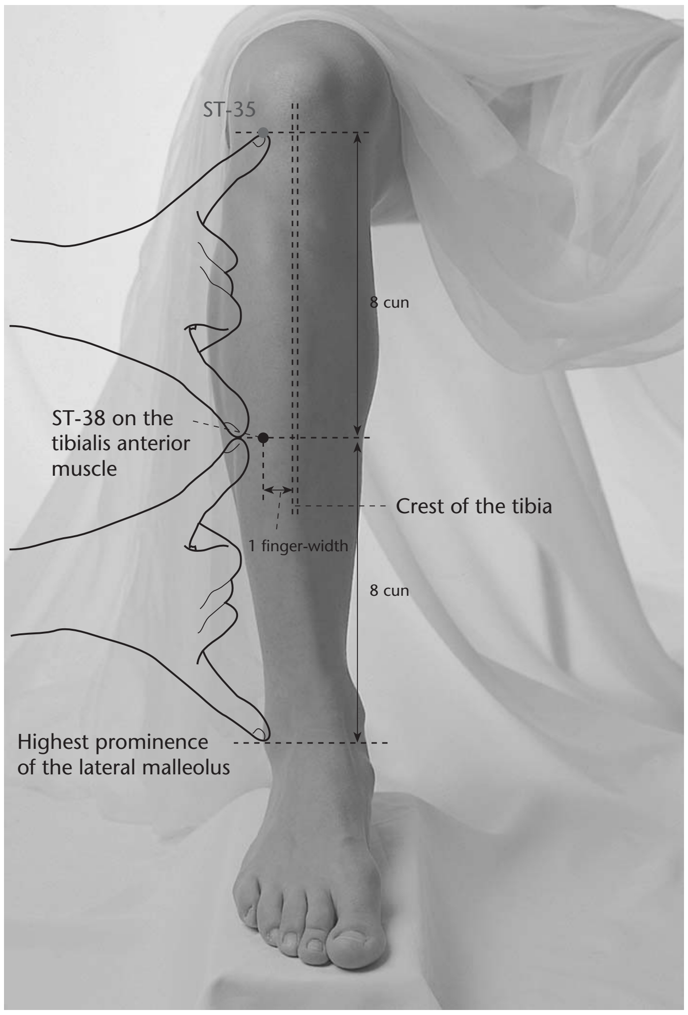
ST-35
8 cun
ST-38 on the tibialis anterior muscle
Crest of the tibia
1 finger-width
8 cun
Highest prominence of the lateral malleolus

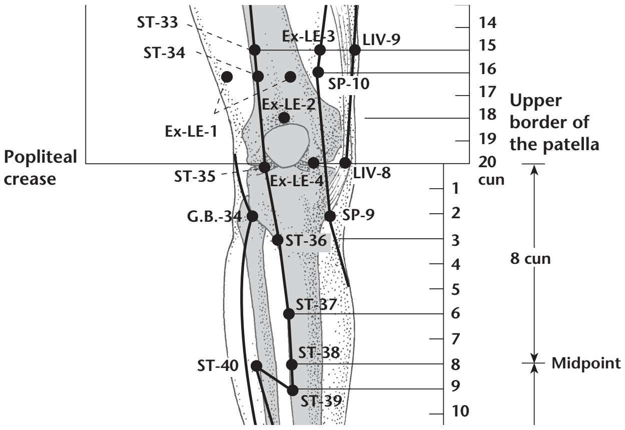
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper border of the patella
1
2
3
4
5
6
7
8
9
10
8 cun
Midpoint

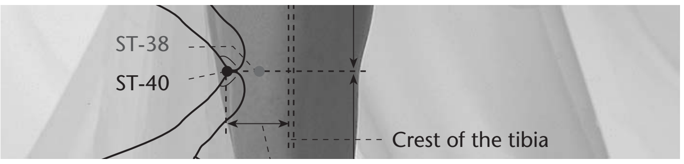
ST-38
ST-40
Crest of the tibia

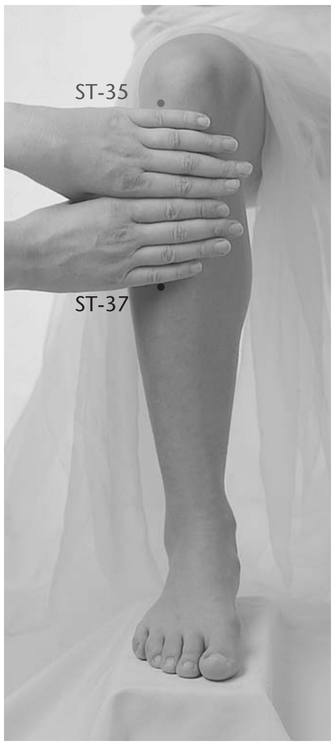

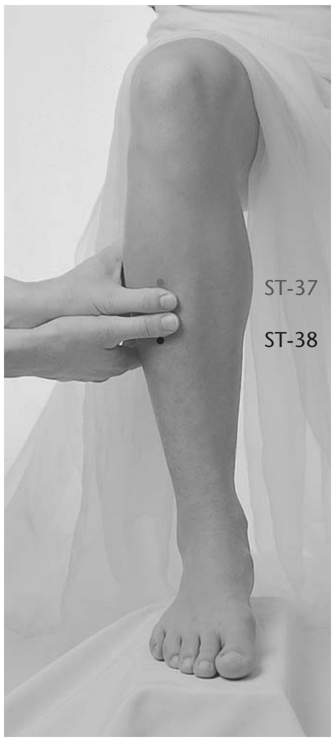

Lower Great Void XIAJUXU ST-39

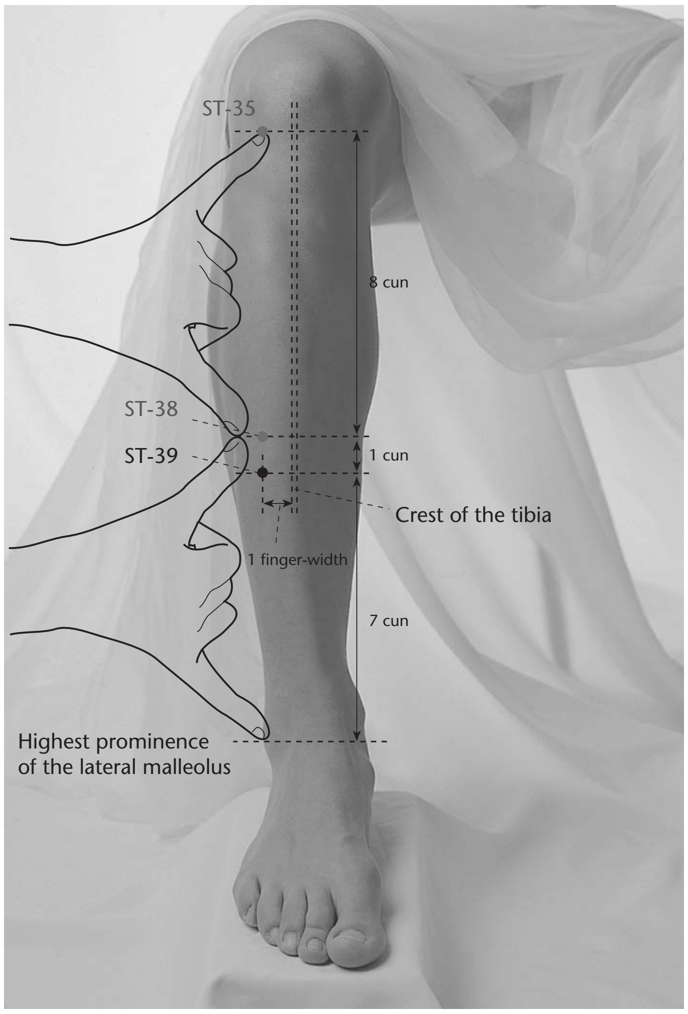
ST-35
8 cun
ST-38
ST-39
1 cun
Crest of the tibia
1 finger-width
7 cun
Highest prominence of the lateral malleolus

Location

1 cun distal to ➞ **ST-38** (midpoint of the line joining ➞ **ST-35** and ➞ **ST-41**) and 1 fingerbreadth lateral to the anterior crest of the tibia.

How to find

Quick method: Spreading-hands technique (➞ 2.3.3): Place the little fingers on ➞ **ST-35** (‘lateral eye of the knee’, level with the knee joint space) and ➞ **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line (level with ➞ **ST-38**), using the thumbs. **ST-39** is located in a depression 1 cun distal and 1 fingerbreadth (middle fingerbreadth, measured from the anterior crest of the tibia) lateral to this point. **ST-39** may be palpated ‘dynamically’ (according to pressure sensitivity).

Located on the same level (7 cun proximal to the lateral malleolus) are ➞ **G.B.-35** (on the posterior border of the fibula), ➞ **G.B.-36** (on the anterior border of the fibula) and ➞ **BL-58** (on the lateral border of the gastrocnemius muscle).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Moves the Small Intestine Qi, eliminates stagnation, clears Damp-Heat
- Opens the channel, alleviates pain

Special features

Lower *he*-sea point of the Small Intestine, Sea of Blood point.

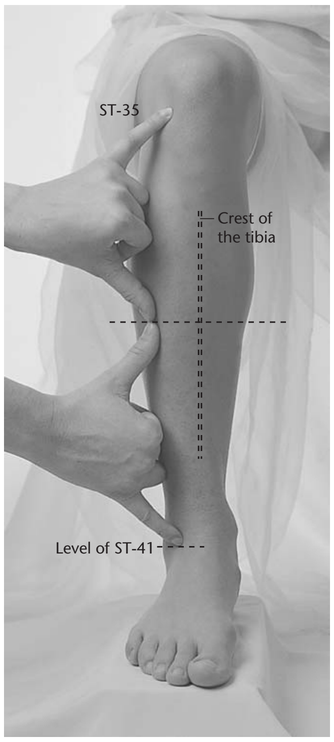
ST-35
Crest of the tibia
Level of ST-41

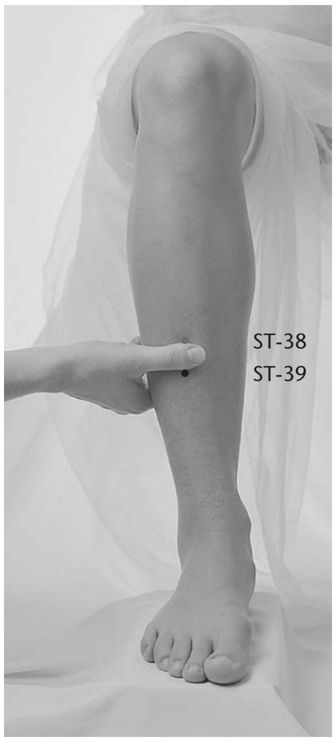

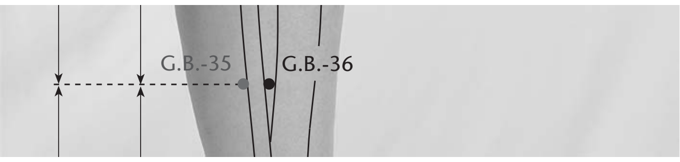

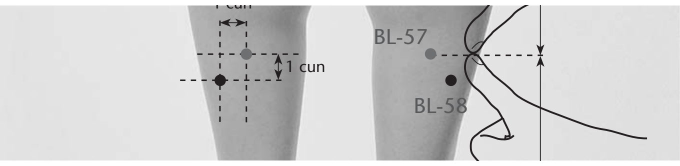
1 cun
BL-57
BL-58

ST-40 *Abundant Bulge* FENGLONG

Location

At the midpoint of the line joining ➞ **ST-35** and ➞ **ST-41** and 2 fingerbreadths lateral to the anterior crest of the tibia or 1 fingerbreadth lateral to ➞ **ST-38**, between the extensor digitorum longus and peroneus brevis muscles.

How to find

Quick method: Spreading-hands technique (➞ 2.3.3): Place the little fingers on ➞ **ST-35** ('lateral eye of the knee', level with the knee joint space) and ➞ **ST-41** (in the depression between the tendons of the extensor digitorum longus and the hallucis longus, level with the highest prominence of the lateral malleolus) and determine the midpoint of this line, using the thumb. **ST-40** is located in a depression 2 fingerbreadths (2 middle fingerbreadths, measured from the anterior crest of the tibia) lateral to this point. **ST-40** may be palpated 'dynamically' (according to pressure sensitivity). Located on the same level (8 cun proximal to the highest prominence of the lateral malleolus) are ➞ **ST-38** (1 fingerbreadth lateral to the tibia) and ➞ **BL-57** (on the dorsal aspect of the lower leg, halfway between ➞ **BL-40** and ➞ **BL-60**).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Transforms Dampness and Phlegm, clears Phlegm in the Lung and Heart, calms the *shen*

Special features

Luo-connecting point. Main acupuncture point for eliminating Phlegm.

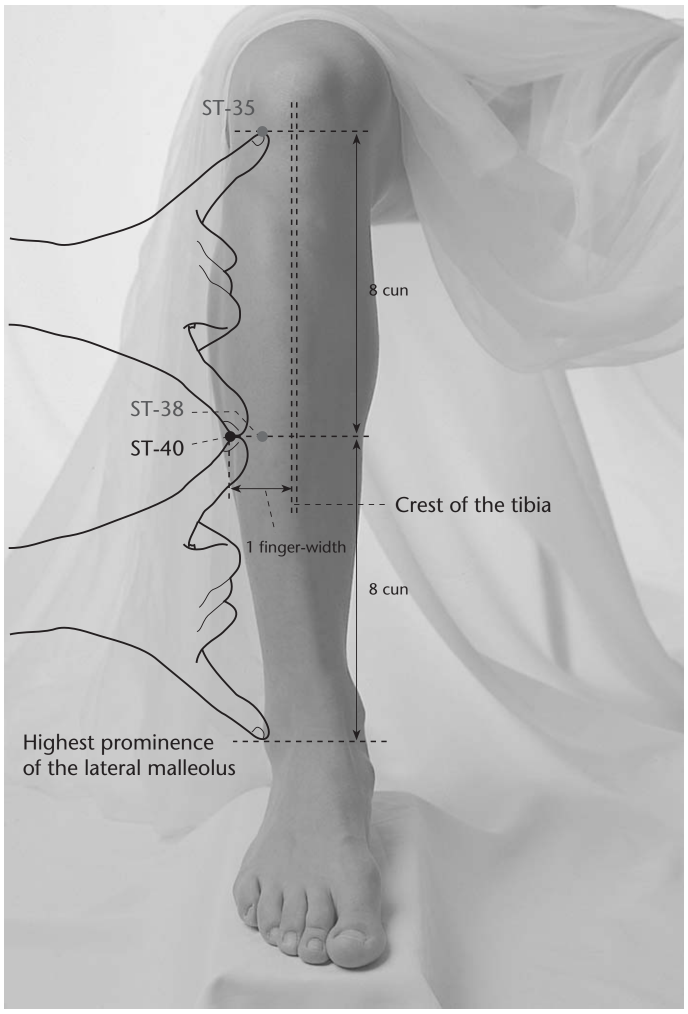
ST-35
8 cun
ST-38
ST-40
Crest of the tibia
1 finger-width
8 cun
Highest prominence of the lateral malleolus

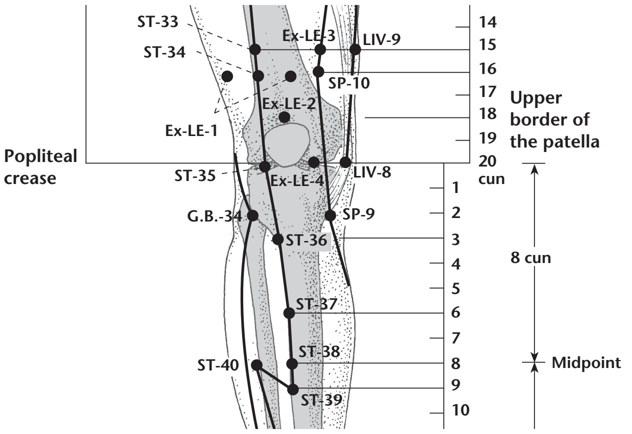
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper border of the patella
1
2
3
4
5
6
7
8
9
10
8 cun
Midpoint

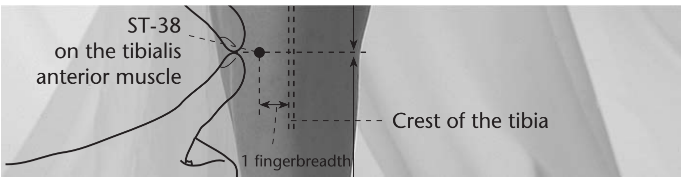
ST-38 on the tibialis anterior muscle
Crest of the tibia
1 fingerbreadth

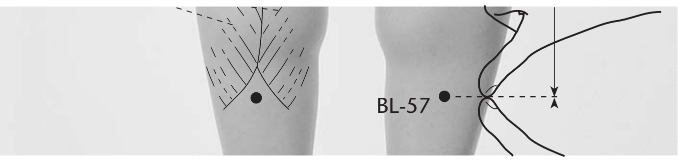

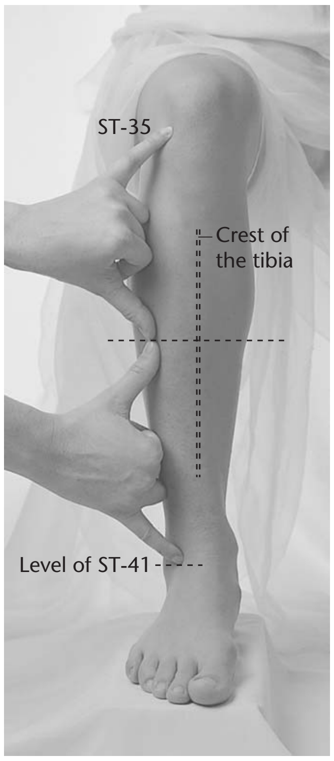
ST-35
Crest of the tibia
Level of ST-41

Stream Divide JIEXI ST-41

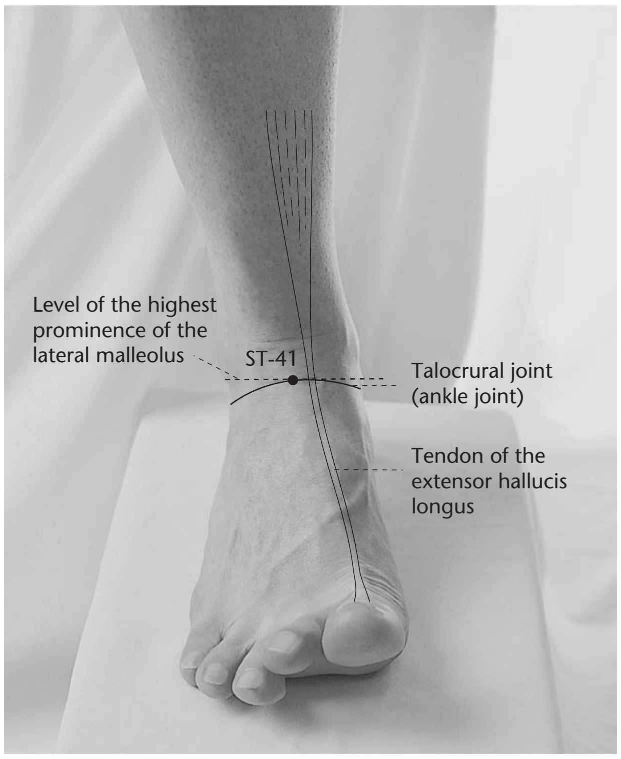
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

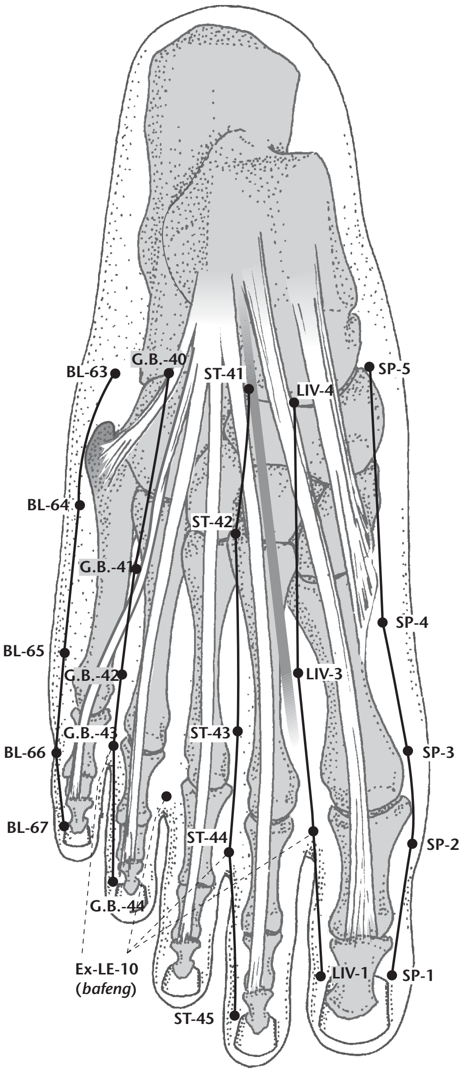
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10 (*bafeng*)
LIV-1
SP-1
ST-45

Location

On the ankle, on the level of the highest prominence of the lateral malleolus, in the depression between the tendons of the extensor digitorum and the extensor hallucis longus.

How to find

With the patient extending their big toe against resistance, the tendon of the extensor hallucis longus muscle will become more distinct. **ST-41** is located lateral to this tendon, in an obvious depression on the level of the highest prominence of the lateral malleolus (→ 3.6.2) and ankle joint space.
Located on the same level is → **LIV-4** (medial to the tendon, halfway between → **SP-5** and **ST-41**).

Needling

0.5–1 cun vertically or obliquely deep to the tendons in a medial direction towards → **SP-5** or in a lateral direction towards → **G.B.-40**. Caution: Tibial artery/vein/nerve lie deep to this point.

Actions/Indications

- Clears Heat from the Stomach and Stomach channel
- Calms the *shen*
- Opens the channel and *luo* vessels, alleviates pain

Special features

Jing-river point, Fire point, tonification point. Important local point for disorders of the ankle joint. Important distal point for frontal headaches caused by Stomach Heat or Stomach Fire.

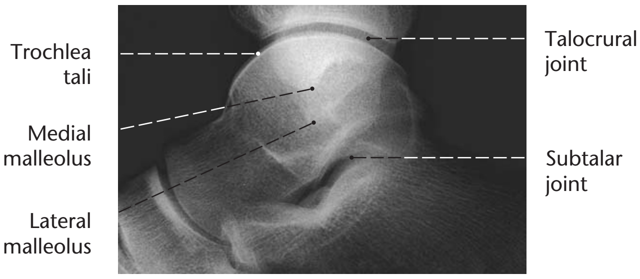
Trochlea tali
Medial malleolus
Lateral malleolus
Talocrural joint
Subtalar joint

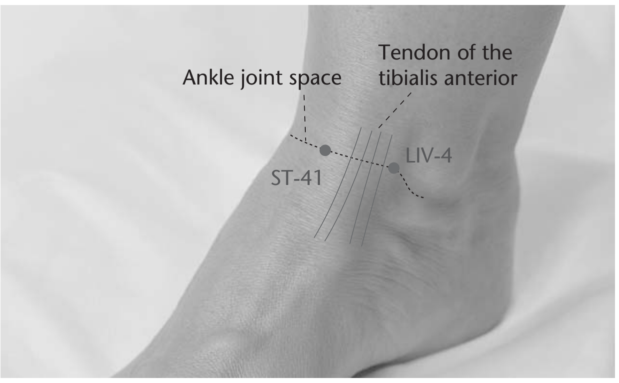
Ankle joint space
Tendon of the tibialis anterior
LIV-4
ST-41

ST-42 *Rushing Yang* CHONGYANG

Location

On the highest point of the dorsum of the foot, between the tendons of the extensor hallucis longus and the extensor digitorum longus, directly lateral to the point where the dorsalis pedis artery may be palpated. The point is bordered proximally by the 2nd and 3rd metatarsal bones and distally by the 2nd and 3rd cuneiform bones. **Alternative location:** Sometimes, this point may be located lateral to the medial portion of the extensor digitorum longus tendon (joining the 2nd toe).

How to find

With the palpating finger, glide from → **ST-43** (in the depression between the base of the 2nd and 3rd metatarsal bones) proximally towards → **ST-41** (in the depression between the tendons of extensor digitorum longus and hallucis longus). Locate **ST-42** at the highest point of the dorsum of the foot, in a depression lateral to the pulse. In most cases, **ST-42** is located lateral to the tendon of the extensor hallucis longus, which becomes more distinct by extending the big toe.

Needling

Vertically 0.2–0.5 cun. Caution: Dorsalis pedis artery.

Actions/Indications

- Clears Heat from the Stomach channel, regulates the Stomach
- Calms the *shen*
- Opens the channel and *luo* vessels

Special features

Yuan-source point, exit point.

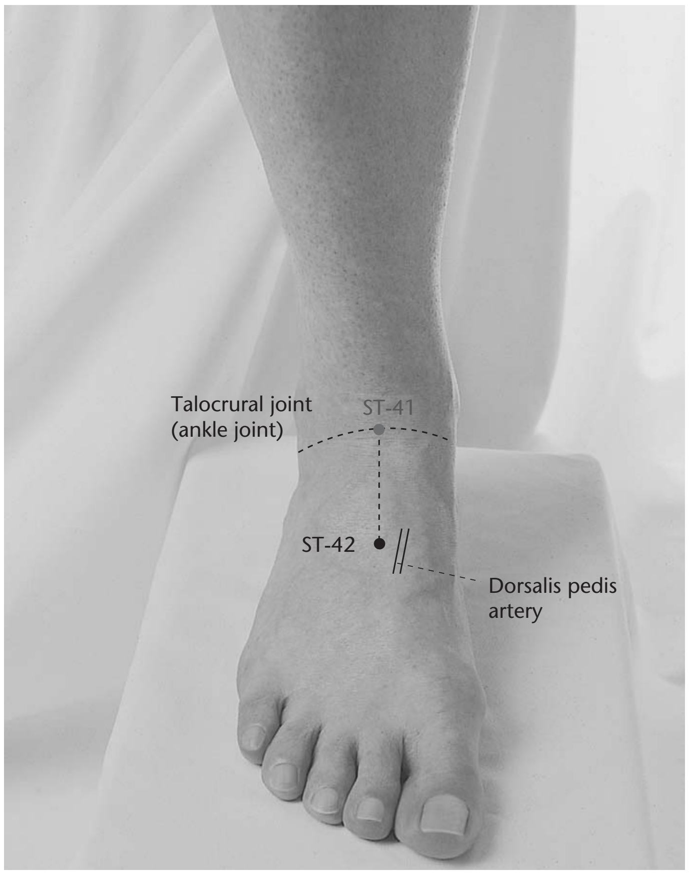
Talocrural joint
(ankle joint)
ST-41
ST-42
Dorsalis pedis
artery

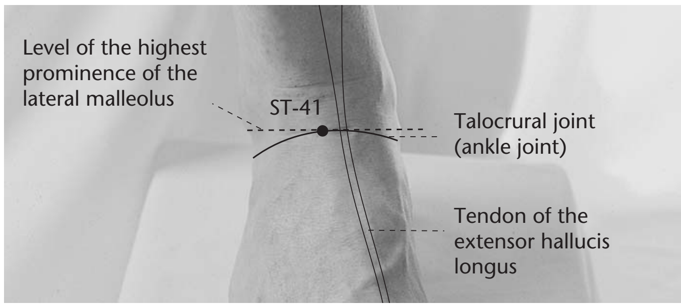
Level of the highest
prominence of the
lateral malleolus
ST-41
Talocrural joint
(ankle joint)
Tendon of the
extensor hallucis
longus

Sunken Valley XIANGU **ST-43**

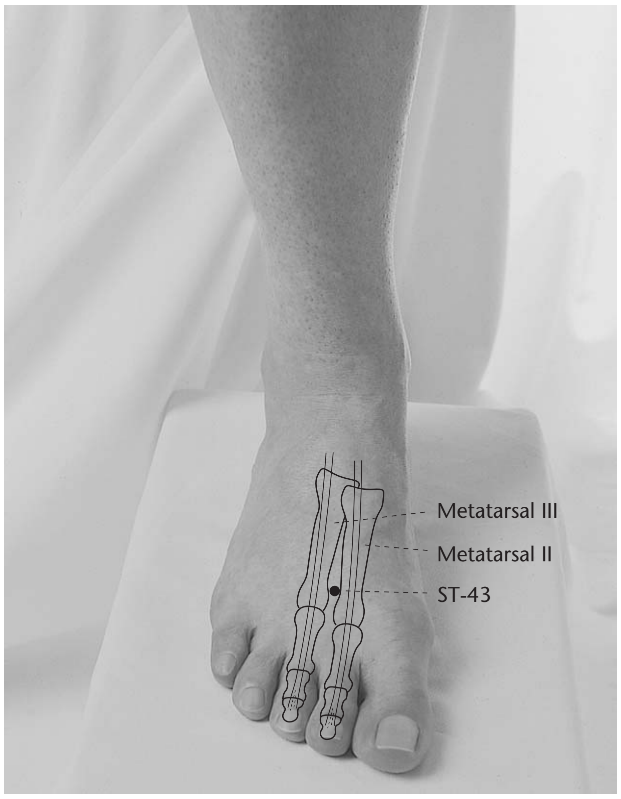
Metatarsal III
Metatarsal II
ST-43

Location

On the dorsum of the foot, in the depression between the 2nd and 3rd metatarsal bones, at the level of the junction of the shafts and the heads of these metatarsal bones.

How to find

Palpate from distal (toes) towards proximal (ankle) between the 2nd and 3rd metatarsal bones, until the finger glides into a depression distal to the metatarsophalangeal joint, where **ST-43** is located. It is approximately on the same level as the junction of the shafts and the heads of the two metatarsal bones.

Needling

Vertically or obliquely 0.5–1 cun

Actions/Indications

- Regulates the Spleen, Stomach and Intestines, reduces oedema

Special features

Shu-stream point, Wood point, generally used for ‘inflammatory’ joint disorders (*bi*-syndromes).

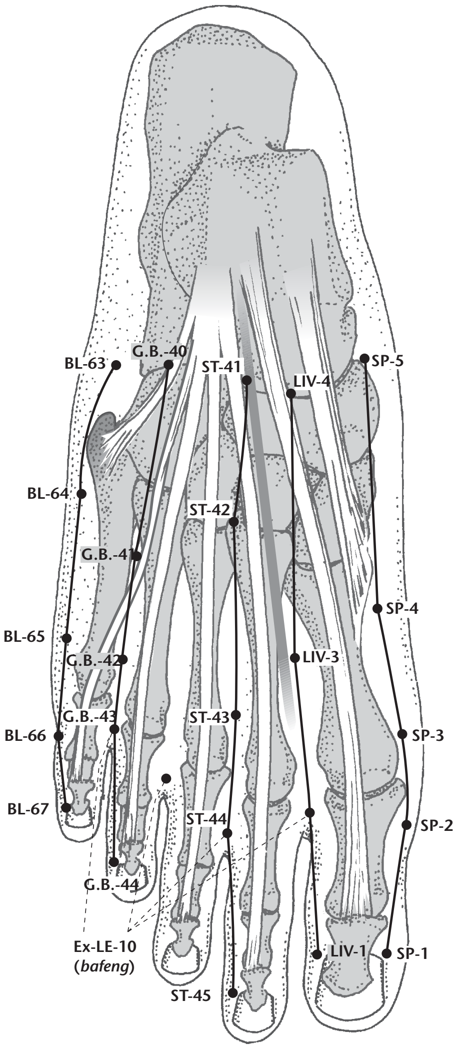
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
SP-3
BL-66
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

ST-44 *Inner Court* NEITING

Location

Between the 2nd and the 3rd toes, proximal to the interdigital fold.

How to find

Locate **ST-44** just proximal to the interdigital fold between the 2nd and the 3rd toes.

Needling

0.5–1 cun vertically or obliquely in a proximal direction

Actions/Indications

- Opens the channel, expels Wind and clears Heat, alleviates pain
- Regulates the Intestines, clears Damp-Heat
- Calms the *shen*

Special features

Ying-spring point, Water point, Ma Dan Yang Heavenly Star point. Major point for clearing Heat from the head and face.

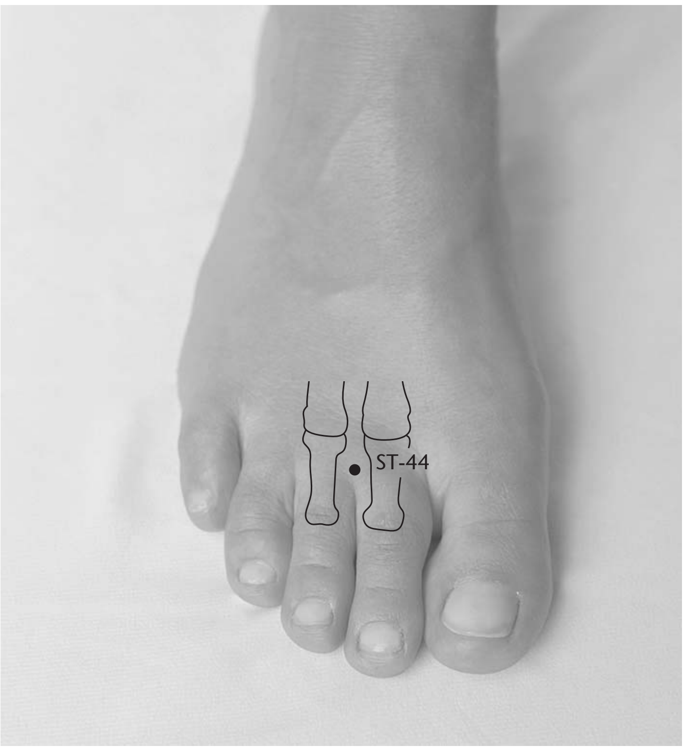

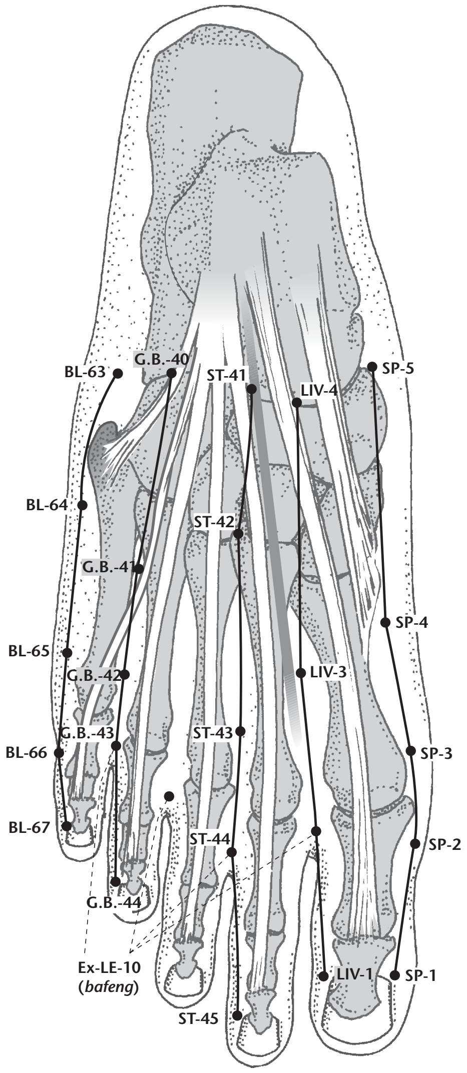
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Strict Exchange LIDUI ST-45

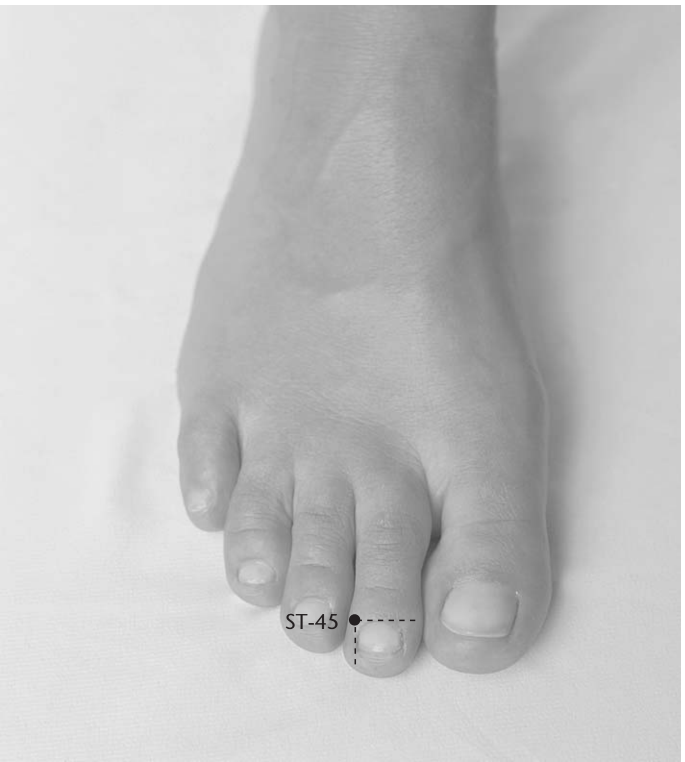

Location

On the 2nd toe, 0.1 cun from the lateral corner of the nail.

How to find

ST-45 is located at the junction of two tangents drawn along the lateral and proximal border of the nail of the 2nd toe, approximately 0.1 cun from the margin of the nail.

Needling

0.1 cun vertically or obliquely in a proximal direction or prick to bleed.

Actions/Indications

- Opens the channel and clears Heat from the *yangming* channels (L.I./ST)
- Calms and regulates the *shen*

Special features

Jing-well point, Metal point, sedation point, important distal point for all Heat disorders affecting the head. For severe insomnia due to Fire or Phlegm-Heat, apply moxibustion (‘small Fire attracting big Fire’ technique).

4.4 The Spleen Channel System – Foot *taiyin* (*zu tai yin jing luo*)

4.4.1 The Spleen Primary Channel (*zu tai yin jing*)

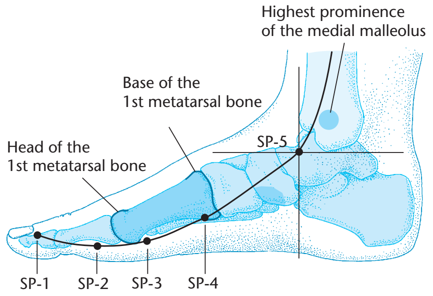
Highest prominence
of the medial malleolus
Base of the
1st metatarsal bone
Head of the
1st metatarsal bone
SP-5
SP-1
SP-2
SP-3
SP-4

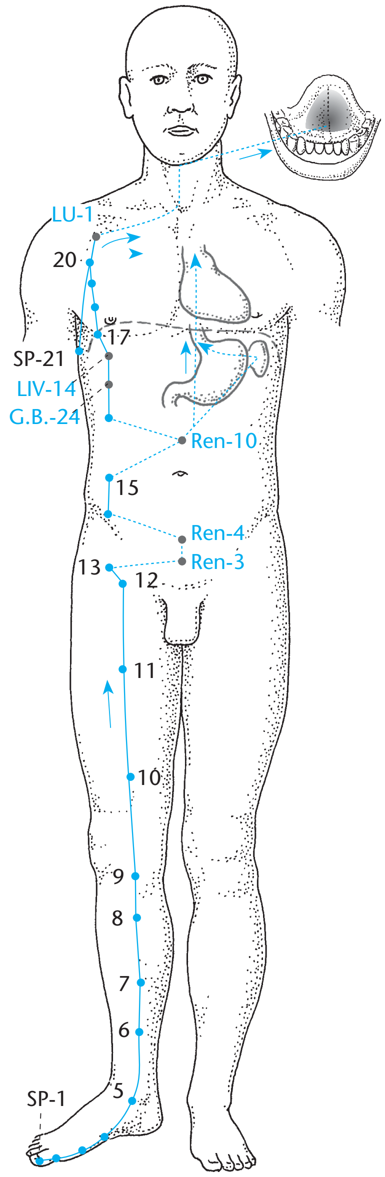
LU-1
20
17
SP-21
LIV-14
G.B.-24
Ren-10
15
Ren-4
Ren-3
13
12
11
10
9
8
7
6
5
SP-1

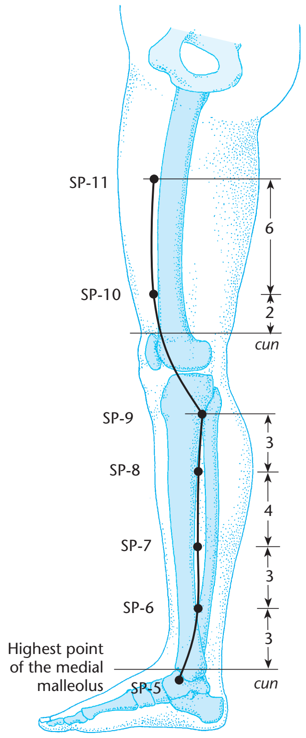
SP-11
6
SP-10
2
cun
SP-9
3
SP-8
4
SP-7
3
SP-6
3
Highest point
of the medial
malleolus
SP-5
cun

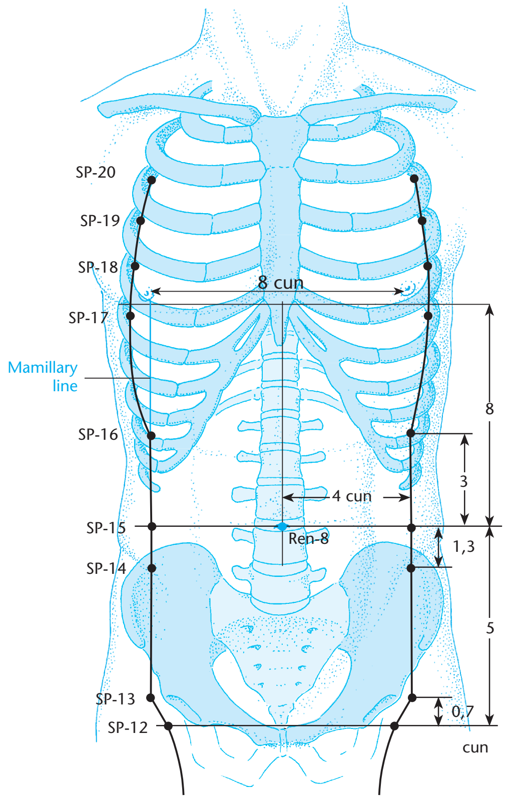
SP-20
SP-19
SP-18
8 cun
SP-17
Mamillary
line
8
SP-16
3
4 cun
SP-15
Ren-8
1,3
SP-14
5
SP-13
0,7
SP-12
cun

Pathway

The **external pathway** of the Spleen primary channel begins at **SP-1** (*yinbai*) on the big toe at the medial corner of the nail. This point is reached by a branch that separates from the Stomach primary channel at → **ST-42** (*chongyang*) (foot Yin–Yang connection of the first great circuit).

➡ From **SP-1** the channel courses along the medial aspect of the foot

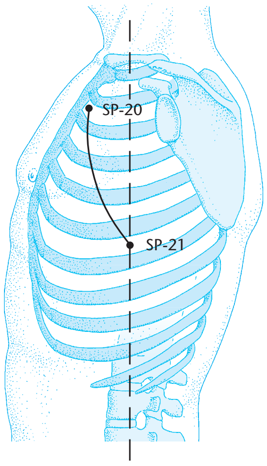

Mid-axillary line

- ascends anterior to the medial malleolus and continues along the posterior border of the medial crest of the tibia
- passes the anteromedial region of the knee and thigh
- ascends the abdomen lateral to the midline to the second intercostal space, passing on its way → **Ren-3** (*zhongji*), → **Ren-4** (*guanyuan*), → **Ren-10** (*xiawan*), → **G.B.-24** (*riyue*) and → **LIV-14** (*qimen*)
- curves downward again and terminates 6 cun inferior to the axilla on the midaxillary line at **SP-21** (*dabao*).

According to Solinas et al (1998), its **inner pathway** branches off at → **Ren-10** (*xiawan*) and enters the Spleen (*pi*), its pertaining *zang*-Organ, and connects with the Stomach (*wei*), its paired *fu*-Organ. From the Stomach (*wei*), an **Interior branch** passes the diaphragm, ascends part of the oesophagus and enters the Heart (*xin*), where it connects with the Heart primary channel (*shou shao yin jing*) (deep Yin–Yin connection → 1.2). From **SP-20** (*zhourong*), a **further branch** ascends to → **LU-1** *(zhongfu)* at the level of the first intercostal space to form the hand–foot pairing (Yin axes of the first great circuit, *taiyin*). The branch ascends further to the tongue, reaches the root of the tongue and disperses on its lower surface.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Heavy sensations of the head and body, sensation of weakness in the legs, general fever, pain in the lower cheek, muscular atrophy and weakness of the legs and medial hip region, cold knees, feet and legs, oedema

Interior (*li*) or *zangfu*-Organ signs and symptoms: Abdominal pain, diarrhoea or loose stools with undigested food, borborygmus, belching and dizziness, abdominal masses, loss of appetite, a sallow complexion, retention of urine

Excess (*shi*) → Spasms, foot pain

Deficiency (*xu*) → Abdominal fullness, borborygmus, diarrhoea with undigested food

Connections and meeting points of the Spleen primary channel (→ 1.2)

Connections with other channels

Stomach primary channel (*zu yang ming jing*)

Connection: Foot Yin–Yang connection of the first great circuit
Location: ST-42 SP-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Interior–Exterior relationship

Lung primary channel (*shou tai yin jing*)

Connection: Paired according to the six channel theory (hand–foot pairing) → *tai yin* (Yin axes of the first great circuit)
Location: SP-20 → LU-1 (on the thorax). A branch of the Spleen primary channel ascends from **SP-20** (*zhourong*) to **LU-1** (*zhongfu*)
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Heart primary channel (*shou shao yin jing*)

Connection: Deep Yin–*Yin* connection
Location: SP → HE (on the thorax). An internal branch of the Spleen primary channel spreads inside the Heart (*xin*) and connects with the Heart primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Heart primary channel receives Nutritive Qi (*ying qi*) from the Spleen primary channel (first circulation of the *ying qi* → 1.1.4)

Connections with other *zangfu*-Organs

Spleen (*pi*), Stomach (*wei*), Heart (*xin*)

4.4.2 The Spleen Divergent Channel (*zu tai yin jing bie*)

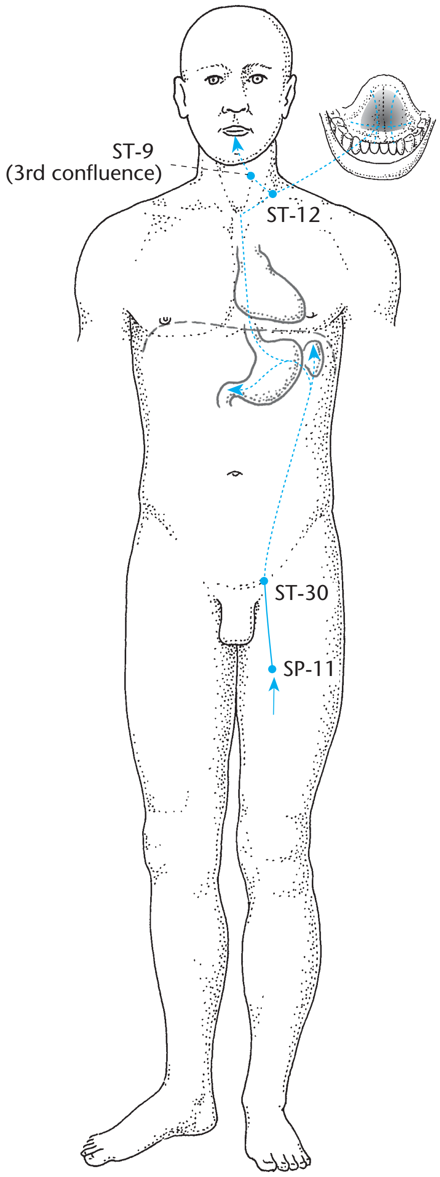

Pathway

The Spleen divergent channel separates from the Spleen primary channel near **SP-11** (*jimen*) on the anteromedial aspect of the thigh, at the medial border of the sartorius muscle. It

➡ penetrates the abdomen in the inguinal region at → **ST-30** (*qichong*)

➡ connects with the Stomach (*wei*), Spleen (*pi*) and Heart (*xin*)

➡ ascends along the oesophagus to the suprasternal fossa and travels laterally to → **ST-12** (*quepen*) in the supraclavicular fossa

➡ crosses the Stomach primary channel and the Stomach divergent channel at → **ST-9** (*renying*) where they meet to connect at one of the six *he*-confluences (here: ST/SP as third confluence → 1.3).

Clinical importance

- Strengthens the relationship between the Spleen and the Stomach (*zangfu*-Organs). Points on the Spleen primary channel can therefore be used for disorders of the Stomach, and vice versa points on the Stomach primary channel can treat disorders of the Spleen. In clinical practice, Spleen points are used for the whole gastrointestinal tract.
- Meets the Stomach primary channel at → **ST-30**; this is also a Sea of Water and Grain point, which can be used for many digestive disorders.
- Links the Spleen primary channel with the Heart; in combination with other points, SP points can be used for Qi and Blood deficiency as well as for psycho-emotional disorders.
- Connects the Spleen primary channel and the neck region and can therefore be used for chronic disorders of the neck and thyroid caused by retention of Phlegm (for example **SP-3** with **ST-40**, **P-5**).
- Connects the Spleen primary channel with the mouth and tongue and can therefore be used for disorders/ulcers of the tongue and mouth caused by retention of Phlegm (for example **SP-2, SP-3, SP-6**).

4.4.3 The Spleen Sinew Channel (*zu tai yin jing jin*)

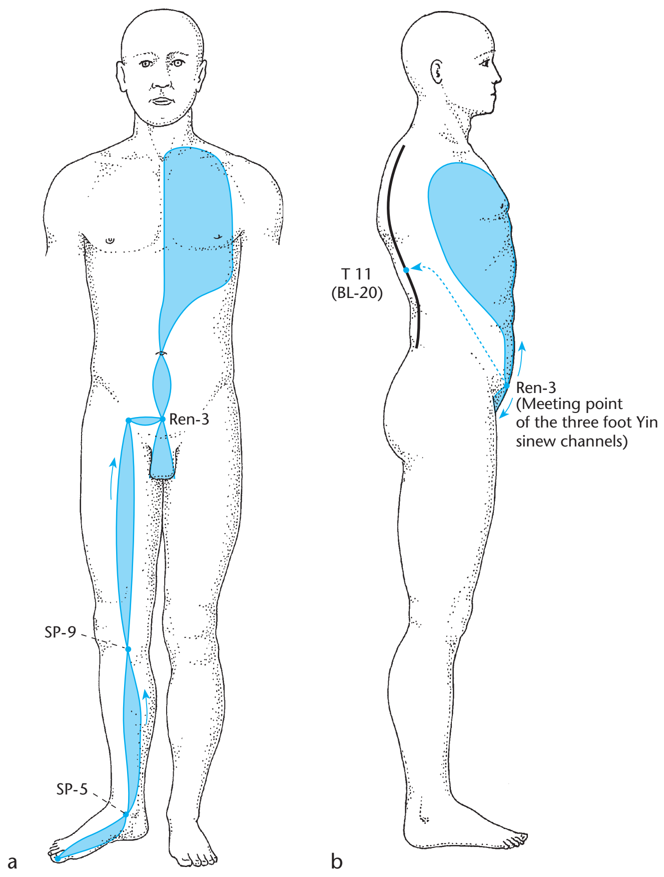

Pathway

The Spleen sinew channel begins on the medial aspect of the big toe at **SP-1** (*yinbai*)

➡ runs along the medial border of the foot

➡ binds (*jie*) anterior to the medial malleolus at **SP-5** (*shangqiu*)

➡ ascends along the medial aspect of the tibia to **SP-9** (*yinlingquan*) and binds (*jie*) anterior to the medial head of the tibia

➡ runs along the medial aspect of the thigh and binds (*jie*) on the medial aspect of the inguinal region

➡ continues to → **Ren-3** (*zhongji*) where it meets the other sinew channels of the foot Yin channels

At → **Ren-3** the sinew channel separates into three branches:

- **One branch** runs to the genital region
- **A further branch** ascends along the anterior midline, covering the *ren mai* and binding (*jie*) at the umbilical region. It penetrates the abdomen and spreads in the costal and thoracic region.
- **A third branch** penetrates deeper into the body and ends at the spinal region at → **BL-20** (*pishu*).

Clinical importance

- **Pathology:** Distending pain of the big toe, pain at the medial malleolus, cramps and pain in the calves, pain of the medial aspect of the knees and thighs, pain in the inguinal and genital region (radiating to the umbilicus, thorax and back).
- **Indication:** Mainly for *bi*-syndromes (Painful Obstruction Syndrome) and *wei*-syndrome (atrophy syndrome) along the Spleen channel, especially along the medial aspect of the leg. The area covered by the Spleen sinew channel is larger than that covered by the Spleen primary channel. This explains why Spleen primary channel points can also be used for disorders and diseases of the external genitalia and the umbilical region.

4.4.4 The Spleen *luo*-Connecting Vessel System (*zu tai yin luo mai*)

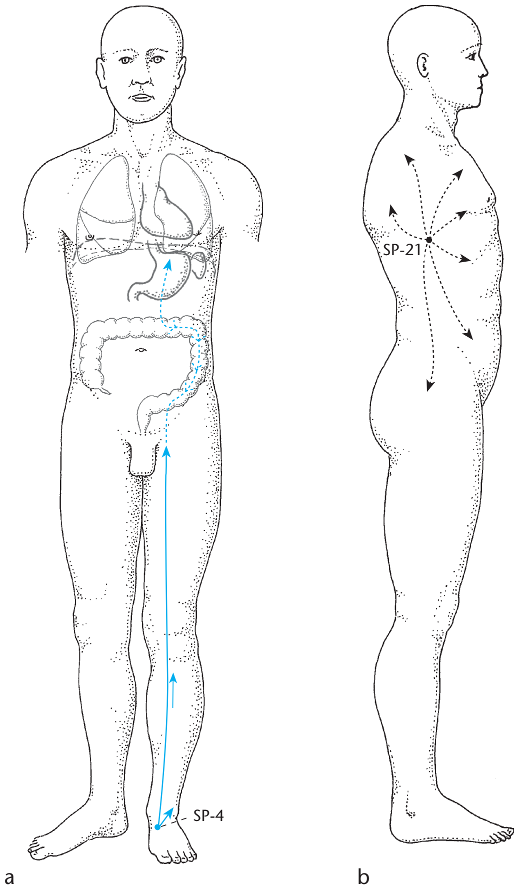

Pathway

The Spleen *luo*-connecting channel separates from the Spleen primary channel at its *luo*-connecting point **SP-4** (*gongsun*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

➡ Horizontal divisions run to the Interiorly–Exteriorly paired Stomach primary channel, according to some schools of thought (for example Ngyen Van Nghi ➞ Appendix) they travel as a **transverse** Spleen *luo*-connecting vessel to the *yuan*-source point ➞ **ST-42** (*chongyang*).

- A **longitudinal** division follows the pathway of the Spleen primary channel, enters the abdomen in the inguinal region and connects with the Large Intestine (*dachang*) and the Stomach (*wei*).

Clinical importance (➞ 8.1.2)

Pathology

- **Excess (*shi*)**: Colicky pain in the Stomach and Large Intestine
- **Deficiency (*xu*)** ➞ *Gu zhang* (drum-like abdominal distension) such as digestive disorders in children, abdominal fullness and distension, ascites
- **Counterflow Qi:** Acute vomiting, diarrhoea, severe abdominal pain

The great *luo*-connecting vessel of the Spleen (*zu tai yin luo mai*)

Pathway

The great *luo*-connecting vessel of the Spleen begins on the lateral aspect of the thorax at ➞ **SP-21** (*dabao*). It contains the Blood of the *luo*-connecting vessel and disperses to the vessels of the whole body, covering them in a net-like fashion (➞ 1.5, ➞ 8.1.2).

Clinical importance (➞ 8.1.2)

- **Excess (*shi*)**: Pain in the whole body, polyarthritis, *bi*-Syndrome (Painful Obstruction Syndrome)
- **Deficiency (*xu*)**: Muscular atrophy and weakness, weakness of the joints, weakness in the whole body.

4.4.5 Cutaneous Region (*tai yin pi bu*)

See description and figures ➞ 1.6

4.4.6 Points of the Spleen Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (➞ 8.1.1): SP-3** (*taibai*) 🟦■
- ***Luo*-connecting point (➞ 8.1.2.): SP-4** (*gongsun*) 🟦■/ SP-21 (*dabao*) 🟦🟦
- ***Xi*-cleft point (➞ 8.1.3): SP-8** (*diji*) 🟦■
- **Associated Back-*shu* point (➞ 8.1.4): BL-20** (*pishu*) 🟦🟦
- **Associated Front-*mu* point (➞ 8.1.5): LIV-13** (*zhangmen*) 🟦🟦
- **Five *shu*-transporting points (➞ 8.1.6):**
  jing-well point (Wood): **SP-1** (*yinbai*) 🟦■
  ying-spring point (Fire), tonification point: **SP-2** (*dadu*)
  shu-stream point (Earth), *ben* point: **SP-3** (*taibai*) 🟦■
  jing-river point (Metal), sedation point: **SP-5** (*shangqiu*) 🟦■
  he-sea point (Water): **SP-9** (*yinlingquan*) 🟦🟦
- ***Hui*-meeting point (➞ 8.1.7): –**
- **Opening point (➞ 8.1.8) of the *chong mai*: SP-4** (*gongsun*) 🟦■
- **Lower *he*-sea points (➞ 8.1.9): –**
- ***Jiaohui*-meeting points (➞ 8.1.10):**
  - with the Kidney and Liver primary channels (meeting point of the three foot *Yin* channels and the *yin qiao mai**: **SP-6** (*sanyinjiao*) 🟦🟦
  - with the LIV and T.B.* channels and the *yin wei mai:* **SP-12** (*chongmen*)
  - with the LIV channel and the *yin wei mai*: **SP-13** (*fushe*)
  - with the *yin wei mai*: **SP-15** (*daheng*), **SP-16** (*fuai*)
  - of other channels with the SP channel: **LU-1**, **G.B.-24**, **LIV-14**, **Ren-3**, **Ren-4**, **Ren-10**, **Ren-17***
- **Gao Wu command point (➞ 8.1.11):** –
- **Window of Heaven point (➞ 8.1.12):** –
- **Points of the Four Seas (➞ 8.1.13):** –
- **Ma Dan Yang Heavenly Star point(➞ 8.1.14):** –
- **Sun Si Miao Ghost point (➞ 8.1.15): SP-1** (*yinbai*) 🟦■
- **Other functional points:**
  – important point for diarrhoea: **SP-4** (*gongsun*) 🟦■
  – important point for Damp *bi*-syndrome: **SP-5** (*shangqiu*) 🟦■
  – important point for eliminating Dampness: **SP-9** (*yinlingquan*) 🟦🟦
  – important point for regulating the Blood: **SP-10** (*xuehai*) 🟦■
  – important point for general pain: **SP-21** (*dabao*) 🟦🟦

Points according to region

- **Local points (➞ 8.2.1):** knee region – **SP-9** (*yinlingquan*) 🟦🟦; Large Intestine – **SP-15** (daheng) 🟦■; knee – **SP-9** (*yinlingquan*) 🟦🟦; foot – **SP-5** (*shangqiu*) 🟦■; toes – **SP-3** (*taibai*) 🟦■
- **Adjacent points (➞ 8.2.1):** Spleen/Stomach – **SP-15** (*daheng*) 🟦■; hypochondrium – **SP-21** (*dabao*) 🟦🟦; knee – **SP-10** (*xuehai*) 🟦■; toes – **SP-4** (*gongsun*) 🟦■, **SP-5** (*shangqiu*) 🟦■, **SP-9** (*yinlingquan*) 🟦🟦
- **Distal points (➞ 8.2.1):** for the Spleen/Stomach, upper abdominal region – **SP-4** (*gongsun*) 🟦■; for the female genital organs, urogenital region – **SP-6** (*sanyinjiao*) 🟦🟦; lower abdomen – **SP-1** (*yinbai*) 🟦■; for the hand and knee – **SP-5** (*shangqiu*) 🟦■

Specific points according to the channel pathway (in numerical order):

- **SP-1** (*yinbai*) 🟦■: Sun Si Miao Ghost point (➞ 8.1.15); distal point for the lower abdomen (➞ 8.2.1); *jing*-well point (Wood) (➞ 8.1.6)
- **SP-2** (*dadu*): *ying*-spring point (Fire) (➞ 8.1.6), tonification point
- **SP-3** (*taibai*) 🟦■: *yuan*-source point (➞ 8.1.1), *shu*-stream point (➞ 8.1.6), *ben* point (Five Phase point) (➞ 8.1.6)

---

* Mentioned by only some authors.

- **SP-4** (*gongsun*) 🟦■: *luo*-connecting point of the Spleen (➞ 8.1.2); opening point of the *chong mai* (➞ 8.1.8); distal point for the Spleen/Stomach, upper abdomen (➞ 8.2.1); important point for diarrhoea; adjacent point for the toes (➞ 8.2.1)
- **SP-5** (*shangqiu*) 🟦■: *jing*-river point (Metal) (➞ 8.1.6); Sedation point; important point for Damp *bi*-syndrome; local point for the foot (➞ 8.2.1); distal point for the hand and knee (➞ 8.2.1)
- **SP-6** (*sanyinjiao*) 🟦🟦: *jiaohui*-meeting point with the KID and LIV channels and the *yin qiao mai* (➞ 8.1.10) (meeting point of the three foot Yin channels); distal point for the female genital system, the urogenital region (➞ 8.2.1)
- **SP-8** (*diji*) 🟦■: *xi*-cleft point (➞ 8.1.3)
- **SP-9** (*yinlingquan*) 🟦🟦: *he*-sea point (Water) (➞ 8.1.6); important point for eliminating Dampness; local point for the knee region (➞ 8.2.1); local point for the knees (➞ 8.2.1)
- **SP-10** (*xuehai*) 🟦■: important point for regulating the Blood; adjacent point for the knees (➞ 8.2.1)
- **SP-12** (*chongmen*): *jiaohui*-meeting point with the LIV and T.B.* channels and the *yin wei mai* (➞ 8.1.10)
- **SP-13** (*fushe*): *jiaohui*-meeting point with the LIV channel and the *yin wei mai* (➞ 8.1.10)
- **SP-15** (*daheng*) 🟦■: local point for the Large Intestine (➞ 8.2.1); adjacent point for the Spleen/Stomach (➞ 8.2.1); *jiaohui*-meeting point with the *yin wei mai* (➞ 8.1.10)
- **SP-16** (*fuai*): *jiaohui*-meeting point with the *yin wei mai* (➞ 8.1.10)
- **SP-21** (*dabao*) 🟦🟦: adjacent point for the hypochondrium (➞ 8.2.1); Main-*luo*-connecting point of the Spleen (➞ 8.1.2); important point for generalised pain

General location help

- **SP-13 – SP-16**: these points are all located 4 cun lateral to the anterior midline (= mamillary line) ➞ **SP-13**: 0.7 cun superior to **SP-12; SP-14**: 1.3 cun inferior to the umbilicus; **SP-15**: at the level of the umbilicus; **SP-16**: 3 cun superior to the umbilicus.
- **SP-17 – SP-20**: these points are all located 6 cun lateral to the anterior midline in the fifth to second intercostal spaces.

---

*Mentioned by only some authors.

Hidden White YINBAI SP-1

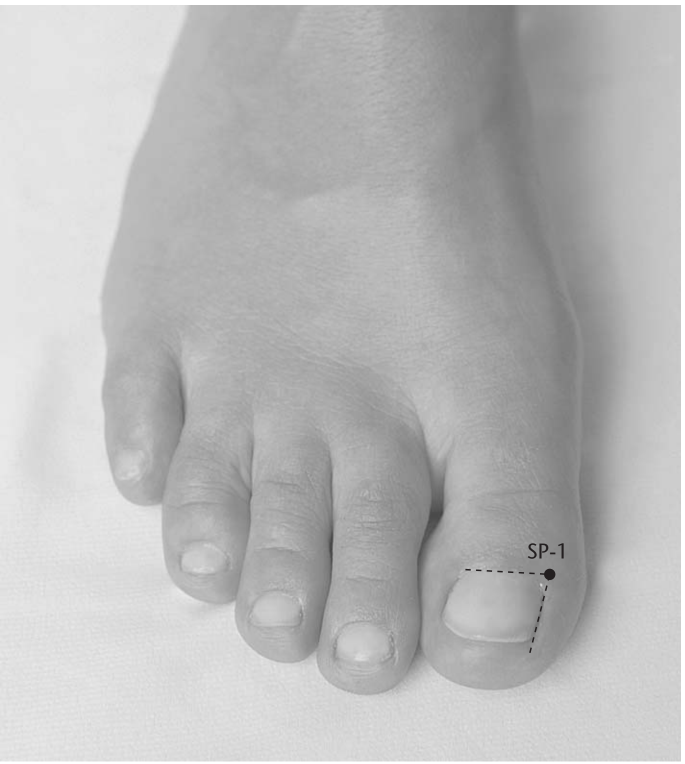

Location

On the big toe, 0.1 cun from the medial corner of the nail.

How to find

SP-1 is located at the junction of two tangents drawn along the medial and proximal border of the nail of the big toe, approximately 0.1 cun from the margin of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed with a needle, lancet or three-edged needle.

Actions/Indications

- Regulates the Blood
- Regulates the Spleen
- Unbinds the chest
- Calms the *shen*
- Revives consciousness

Special features

Jing-well point, Wood point, Sun Si Miao Ghost point, entry point, important distal point for the lower abdomen.

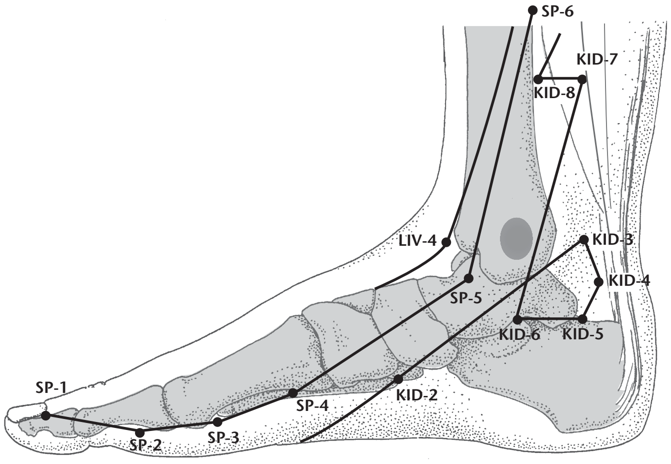
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

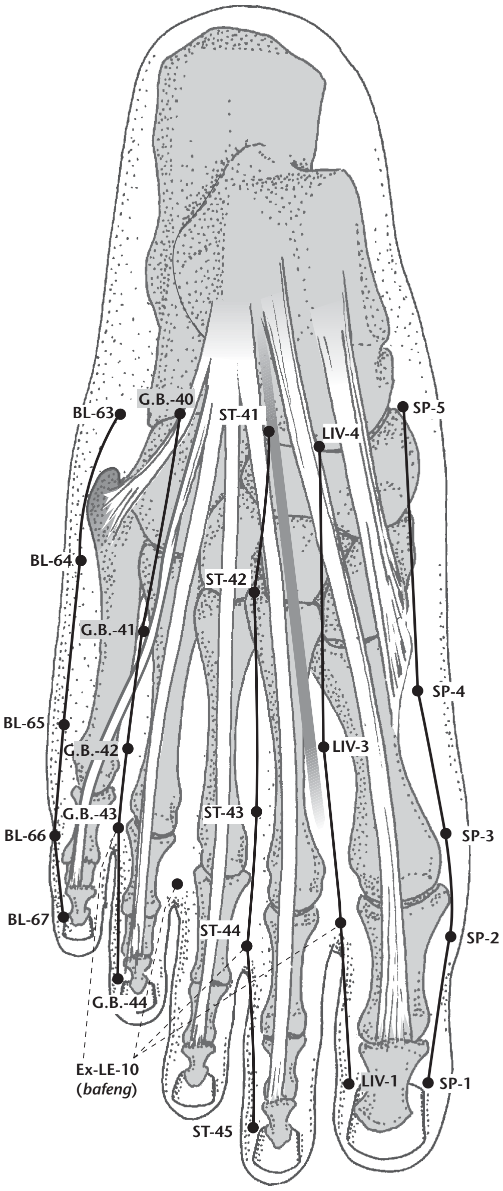
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

SP-2 *Great Metropolis* DADU

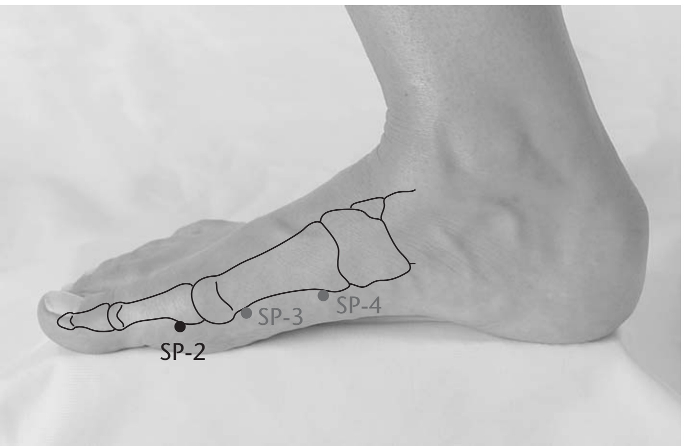

Location

On the medial aspect of the big toe, at the junction of the shaft and the base of the proximal phalanx, distal to the 1st metatarsophalangeal joint.

How to find

On the medial aspect of the big toe, palpate from distal to proximal towards the metatarsophalangeal joint, until you can feel the junction of the shaft with the base of the proximal phalanx. **SP-2** is located distally to the base of the bone, which forms a distinct step, and slightly inferiorly to its outer curvature, at the border of the red and white skin (sole/dorsum of the foot). Located in comparable positions are → **BL-66** on the lateral border of the foot, → **L.I.-2** on radial aspect and → **S.I.-2** on the ulnar aspect of the hand.

Needling

0.2–0.5 cun vertically or obliquely, just below the lower border of the bone. Caution: Painful point.

Actions/Indications

- Regulates and supports the Spleen, harmonises the Lower Burner and clears Heat

Special features

Ying-spring point, Fire point, tonification point.

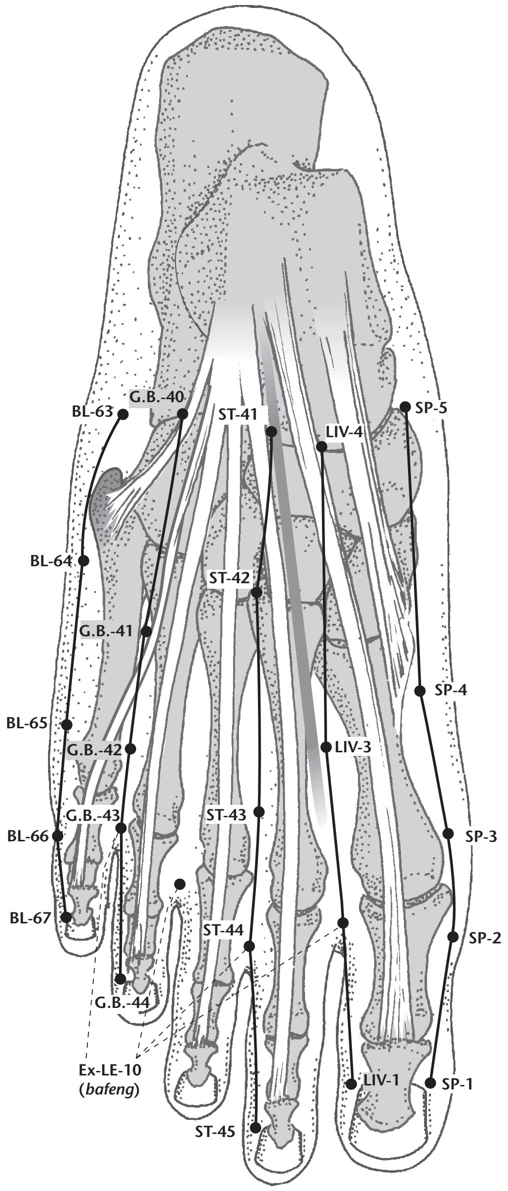
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

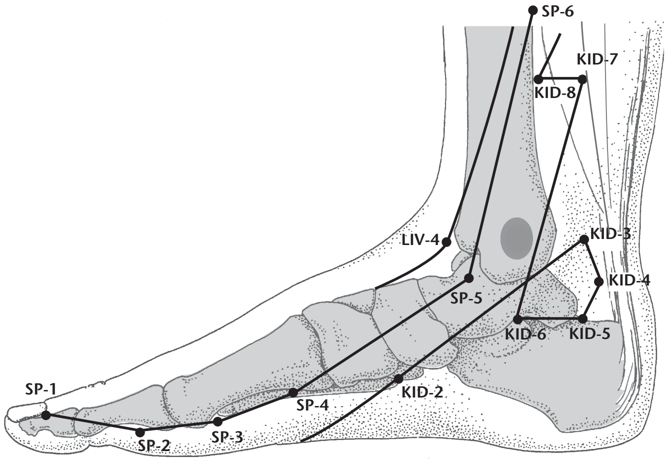
SP-6
KID-7
KID-8
LIV-4
KID-3
SP-5
KID-4
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

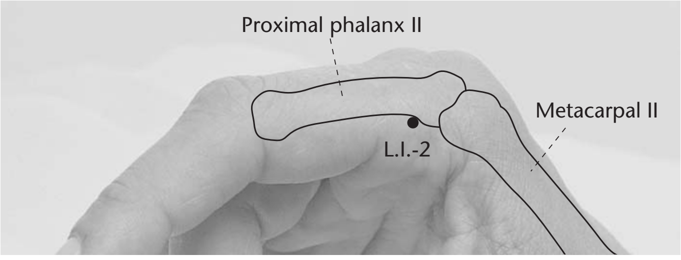
Proximal phalanx II
Metacarpal II
L.I.-2

Supreme White TAIBAI SP-3

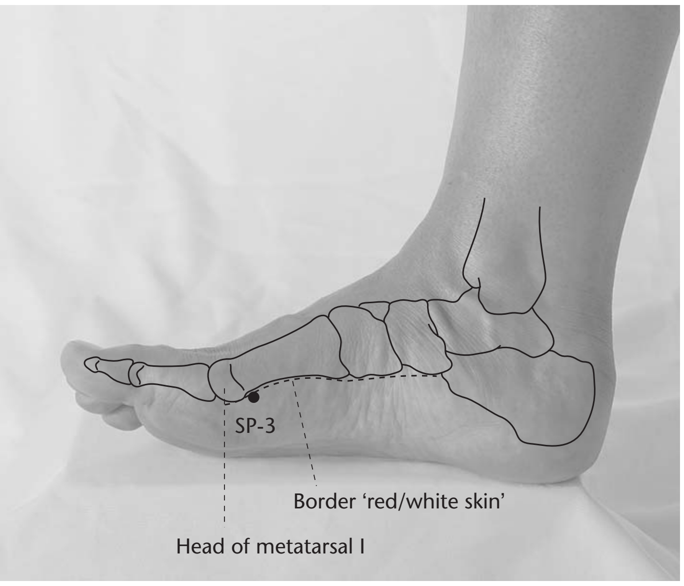
SP-3
Border 'red/white skin'
Head of metatarsal I

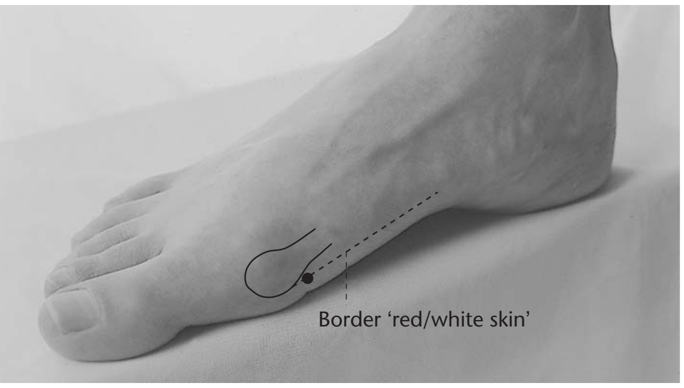
Border 'red/white skin'

Location

On the medial aspect of the foot, in the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin.

How to find

When palpating along the border of the red and white skin on the medial aspect of the foot from distal (toes) to proximal (ankle), the metatarsophalangeal joint forms a prominent landmark. **SP-3** is located just proximal to this joint in a clearly palpable depression, which marks the junction of the head and the shaft of the 1st metatarsal bone.

Located in a comparable position is → **BL-65** (proximal to the head of the 5th metatarsal bone) on the lateral border of the foot. In comparable locations on the borders of the hand are → **L.I.-3** on the radial aspect and → **S.I.-3** on the ulnar aspect of the hand.

Needling

0.5–1 cun vertically or obliquely from medial towards lateral. Caution: Needling this point tends to be painful.

Actions/Indications

- Tonifies the Spleen, harmonises the Lower Burner
- Eliminates Dampness and Damp-Heat
- Regulates Qi locally

Special features

Yuan-source point, *shu*-stream point, Earth point, *ben* point (Five Phases). Important point for tonifying the Spleen.

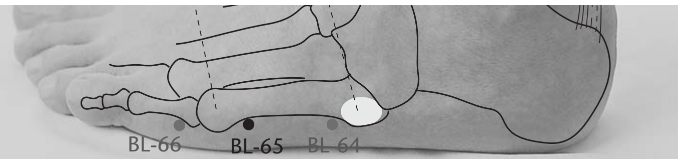
BL-66
BL-65
BL-64

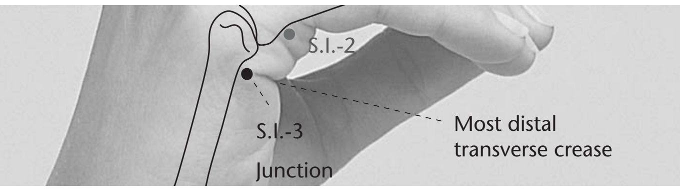
S.I.-2
S.I.-3
Most distal transverse crease
Junction

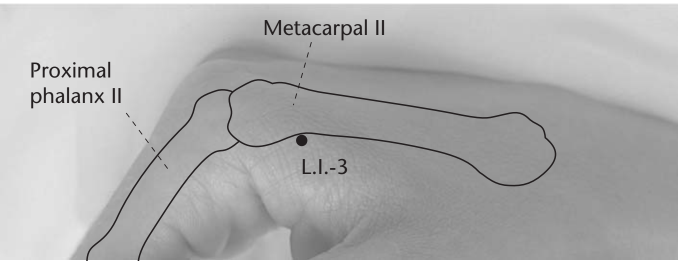
Metacarpal II
Proximal phalanx II
L.I.-3

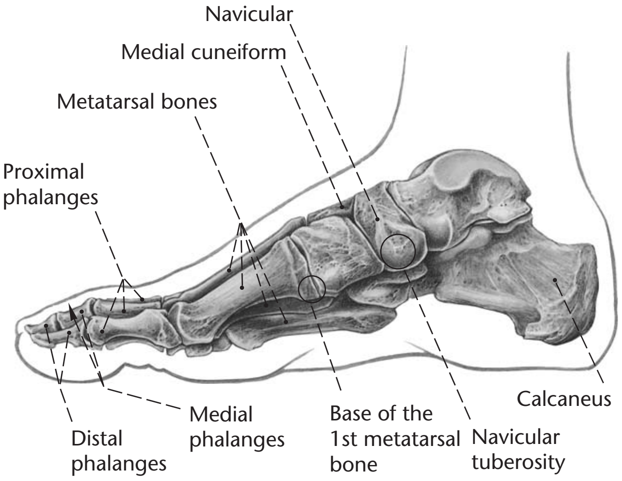
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity
Calcaneus

SP-4 *Grandfather Grandson* GONGSUN

Location

In the depression distal to the base of the 1st metatarsal bone, on the border of the red and white skin.

How to find

When palpating along the border of the red and white skin (sole/dorsum of the foot) on the medial aspect of the midfoot from proximal towards medial, the base of the 1st metatarsal bone (→ 2.6.2) forms a distinct bony prominence. **SP-4** is located in the depression distal to this prominence, at the junction of the shaft and the base of the bone.

→ **SP-3** is located slightly more distally at the junction of the head and the shaft of the 1st metatarsal bone. → **BL-64** is located in a comparable position on the lateral aspect of the midsection of the foot (at the junction of the shaft and the base of the 5th metatarsal bone).

Needling

0.5–1 cun vertically or obliquely from medial towards lateral. Caution: Needling may be painful.

Actions/Indications

- Strengthens and regulates the Spleen Qi, eliminates Dampness
- Calms the *shen*
- Benefits the Heart and chest
- Regulates the *chong mai*
- Local point, for example for metatarsalgia

Special features

Luo-connecting point, opening (master) point of the *chong mai*. Important point for distension and pain in the upper and lower abdomen caused by Qi stagnation, Blood stasis or obstruction due to Dampness.

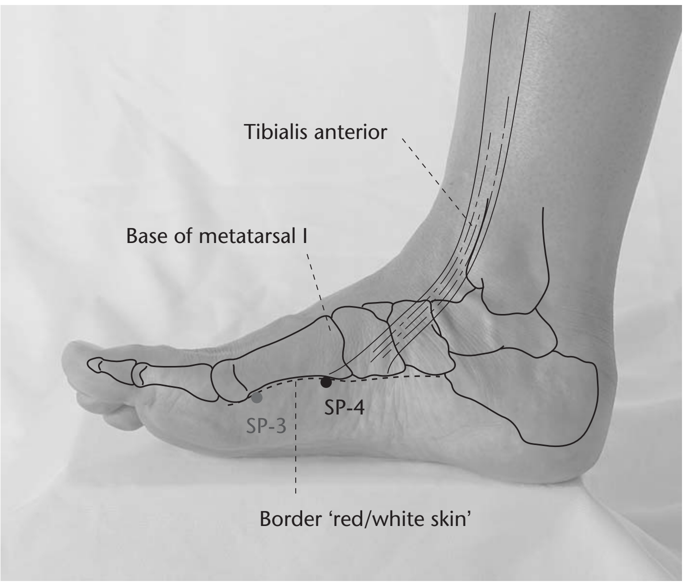
Tibialis anterior
Base of metatarsal I
SP-3
SP-4
Border 'red/white skin'

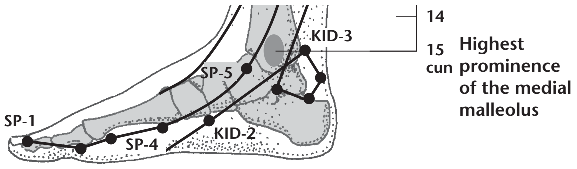
14
15
cun
Highest prominence of the medial malleolus
KID-3
SP-5
SP-1
KID-2
SP-4

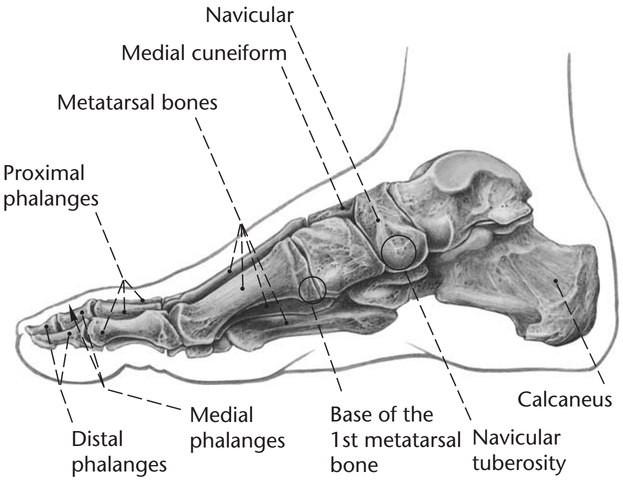
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Navicular tuberosity
Calcaneus

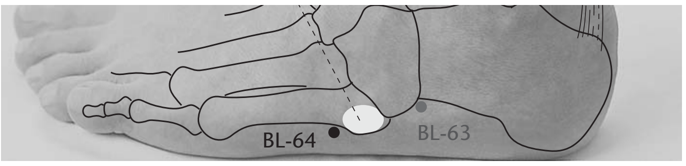

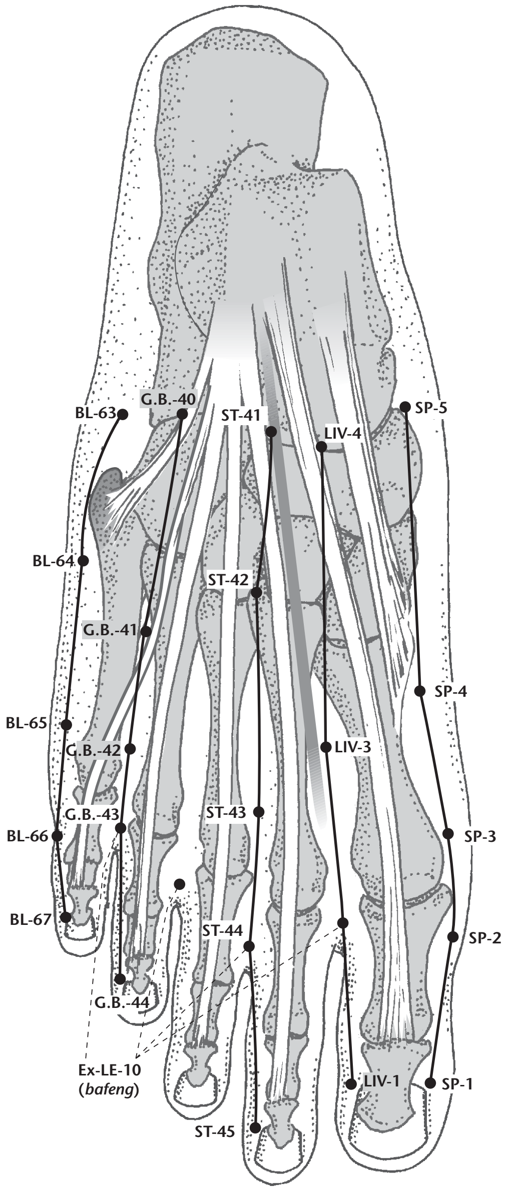
BL-63
G.B.-40
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(bafeng)
LIV-1
SP-1
ST-45

Shang Mound SHANGQIU **SP-5**

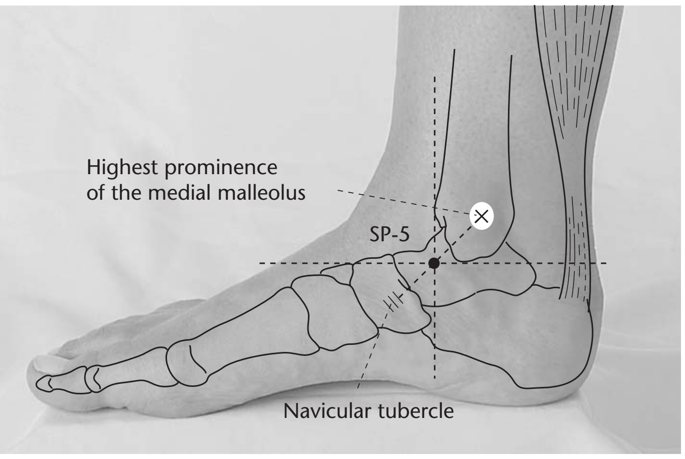
Highest prominence of the medial malleolus
SP-5
Navicular tubercle

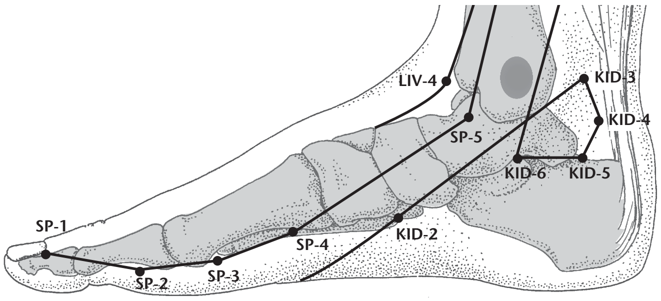
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

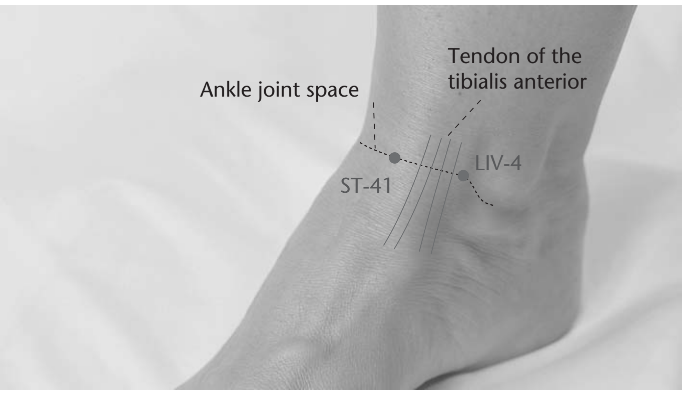
Ankle joint space
Tendon of the tibialis anterior
LIV-4
ST-41

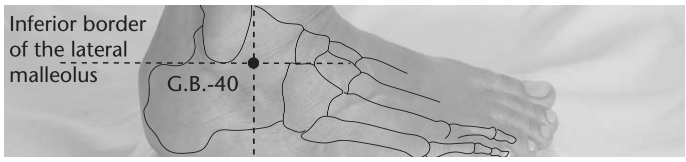
Inferior border of the lateral malleolus
G.B.-40

Location

In the depression at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the lower border of the medial malleolus. **Or**: In the depression halfway between the highest prominence of the medial malleolus and the tubercle of the navicular bone.

How to find

SP-5 is located in a depression anterior and inferior to the medial malleolus, at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the inferior border of the medial malleolus. **SP-5** is located distal to the tendon of the tibialis anterior muscle.

Located in a comparable position on the lateral aspect of the foot is → **G.B.-40** (in a depression at the junction of a vertical line drawn along the anterior border and a horizontal line drawn along the inferior border of the lateral malleolus).

Needling

0.3–0.6 cun vertically, obliquely or transversely (subcutaneously) deep to the tendons towards → **ST-41**.

Actions/Indications

- Tonifies the Spleen, eliminates Dampness, benefits the tendons and bones
- Moves Qi locally
- Calms the *shen*

Special features

Jing-river point, Metal point, sedation point. Important point for treating Damp *bi*-syndrome.

SP-6 Three Yin Intersection SANYINJIAO

Location

3 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia.

How to find

From the highest prominence of the medial malleolus (→ 3.6.2), measure 3 cun (1 handbreadth) in a superior direction, where **SP-6** is located in a depression at the posterior border of the medial crest of the tibia. This point is often sensitive to pressure. Sometimes, this depression is located more anteriorly, and locating **SP-6** should be based on pressure-sensitivity.

Located in a comparable position is → **G.B.-39** (3 cun proximal to the prominence of the lateral malleolus, on the anterior border of the fibula, intersecting point of the three foot Yang channels).

Needling

Vertically or obliquely 1–2 cun. Caution: Contraindicated during pregnancy, particularly with reducing needle techniques; exception: to promote labour.

Actions/Indications

- Tonifies the Spleen and the Stomach, eliminates Dampness
- Nourishes the Blood and Yin, regulates menstruation, promotes labour
- Regulates urination, benefits the genitals, harmonises the Lower Burner
- Calms the *shen*

Special features

Meeting point with the Liver and Kidney channels (group *luo*-connecting point) and the *yin qiao mai*. Very important point for disorders of the Lower Burner (especially for gynaecological disorders, birth and urinary disorders), as well as for many digestive disorders and psychological disorders (caused by Blood and Yin deficiency).

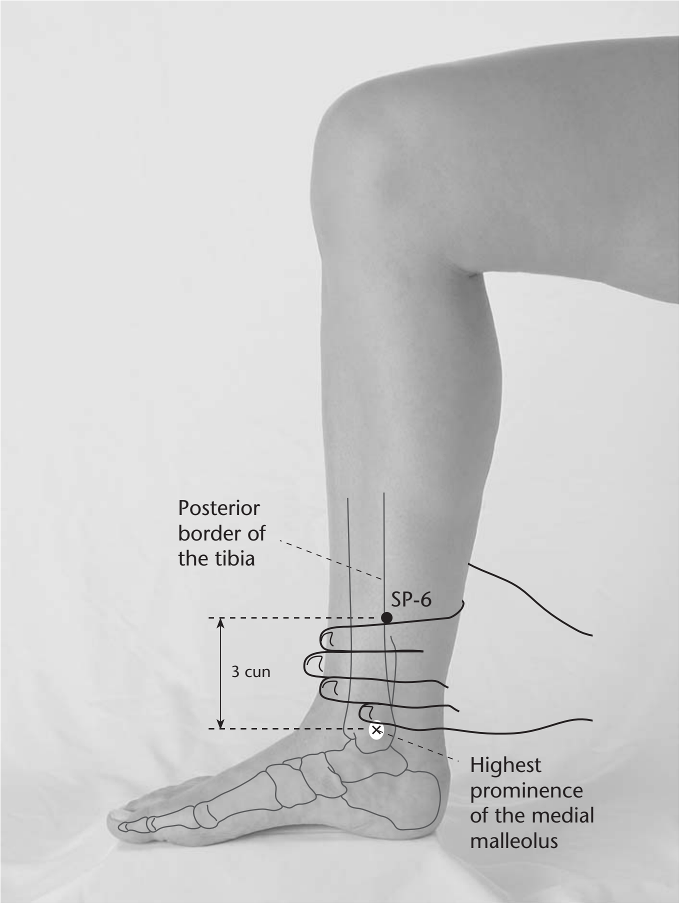
Posterior border of the tibia
SP-6
3 cun
Highest prominence of the medial malleolus

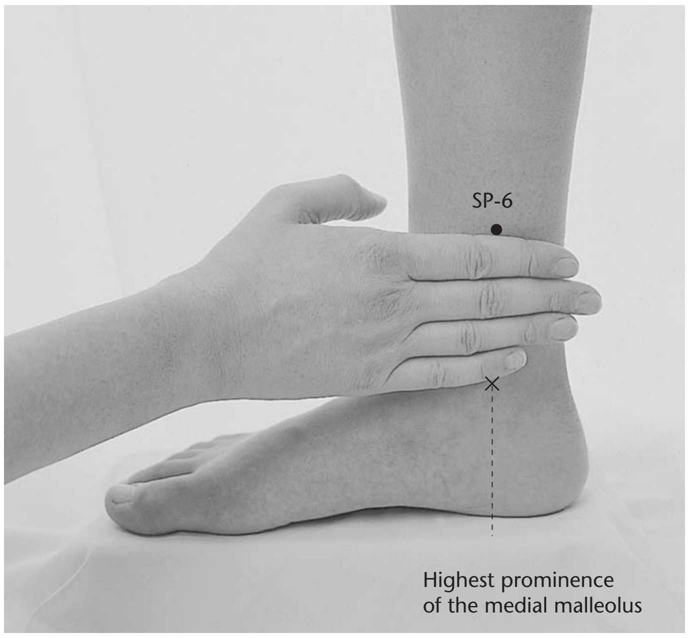
SP-6
Highest prominence of the medial malleolus

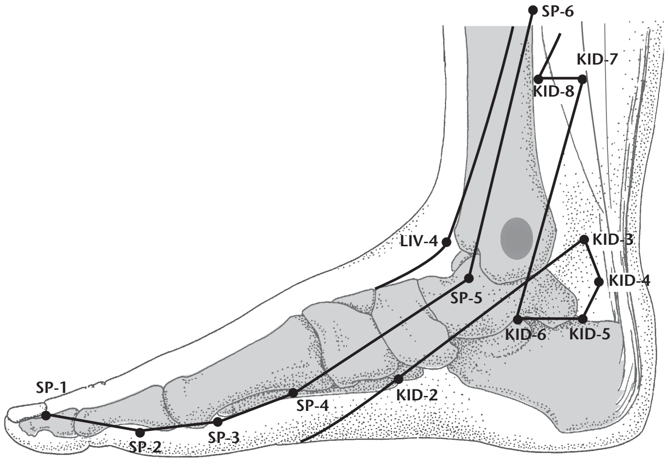
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

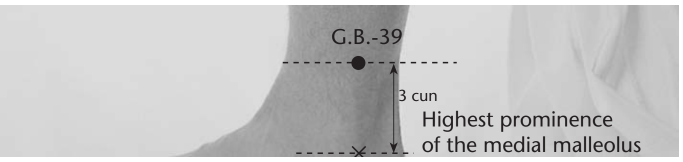
G.B.-39
3 cun
Highest prominence of the medial malleolus

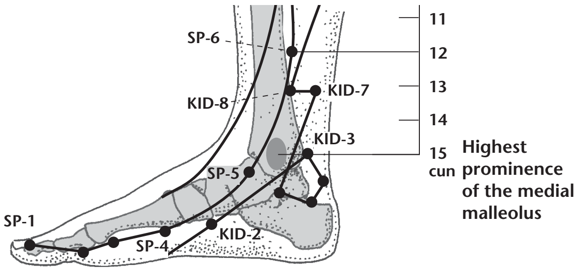
SP-6
KID-8
KID-7
KID-3
SP-5
SP-1
KID-2
SP-4
11
12
13
14
15
cun
Highest prominence of the medial malleolus

Dripping Valley LOUGU SP-7

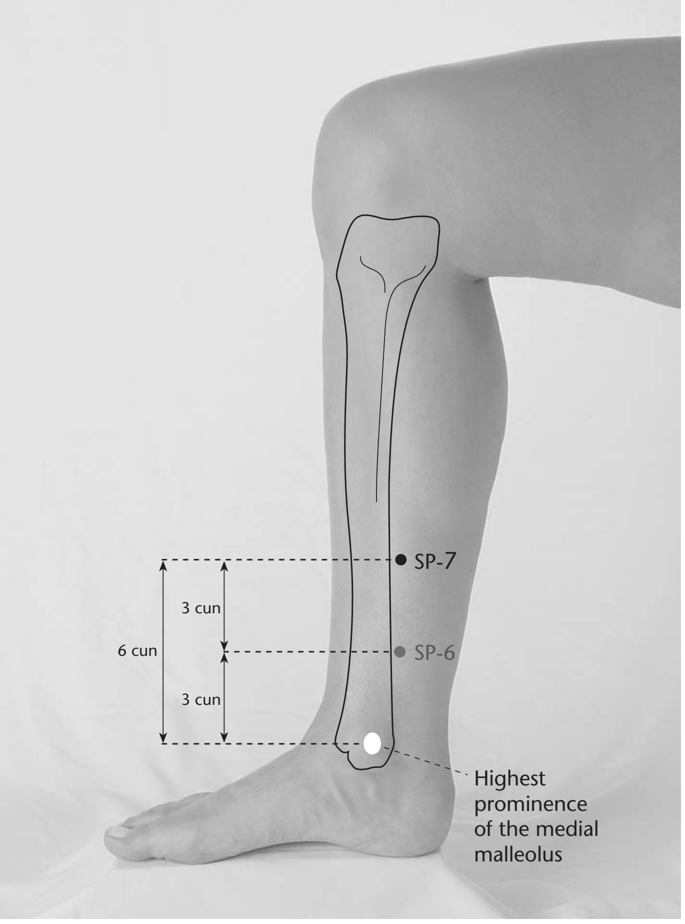
SP-7
3 cun
6 cun
SP-6
3 cun
Highest prominence of the medial malleolus

Location

6 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia.

How to find

From the highest prominence of the medial malleolus, measure 6 cun in a superior direction, where **SP-7** is located at the posterior border of the medial crest of the tibia. **Or**: Spreading hands technique (→ 2.3.3): **SP-7** is located 0.5 cun distal to the midpoint of a line connecting → **SP-9** (distal to the tibial condyle) and the highest prominence of the medial malleolus. → **SP-6** is located midway between the tip of the malleolus and **SP-7** (3 cun superior/inferior respectively).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Strengthens the Spleen, drains Dampness, opens the channel
- Promotes urination

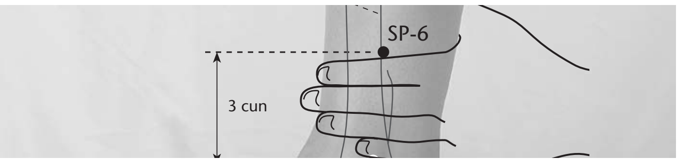

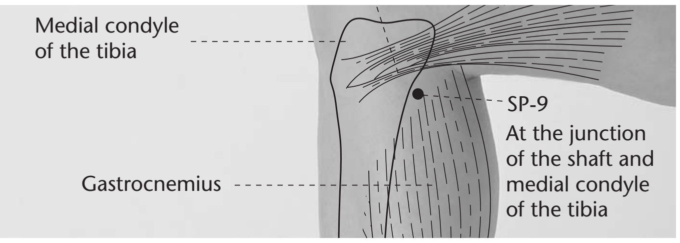
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

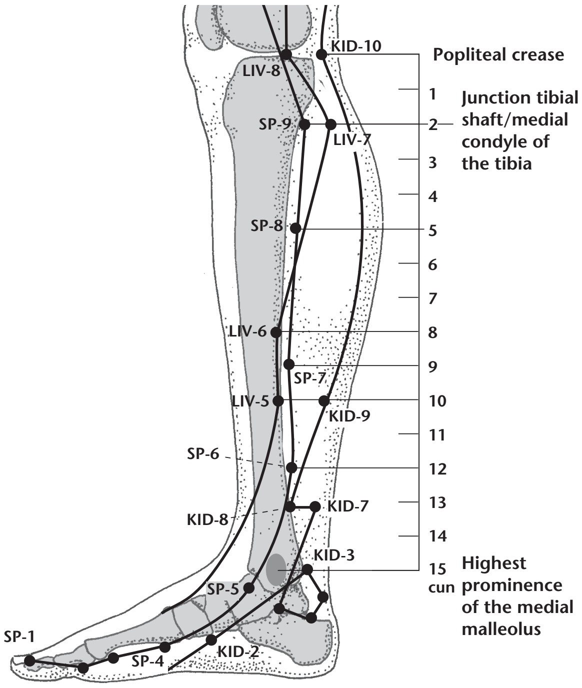
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
10
LIV-5
KID-9
11
SP-6
12
KID-7
13
KID-8
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

SP-8 *Earth Pivot* DIJI

Location

3 cun distal to the junction of the shaft and the medial condyle of the tibia (→ **SP-9**), at the posterior border of the medial crest of the tibia.

How to find

Measure 1 handbreadth directly inferior to the medial condyle of the tibia (→ **SP-9**). **SP-8** is located 3 cun distal to → **SP-9** on the posterior border of the medial crest of the tibia, on a line connecting → **SP-9** and the highest prominence of the medial malleolus.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Regulates the uterus and the Blood
- Regulates the Spleen and eliminates Dampness

Special features

Xi-cleft point. Important point for acute dysmenorrhoea, often used in combination with → **L.I.-4**.

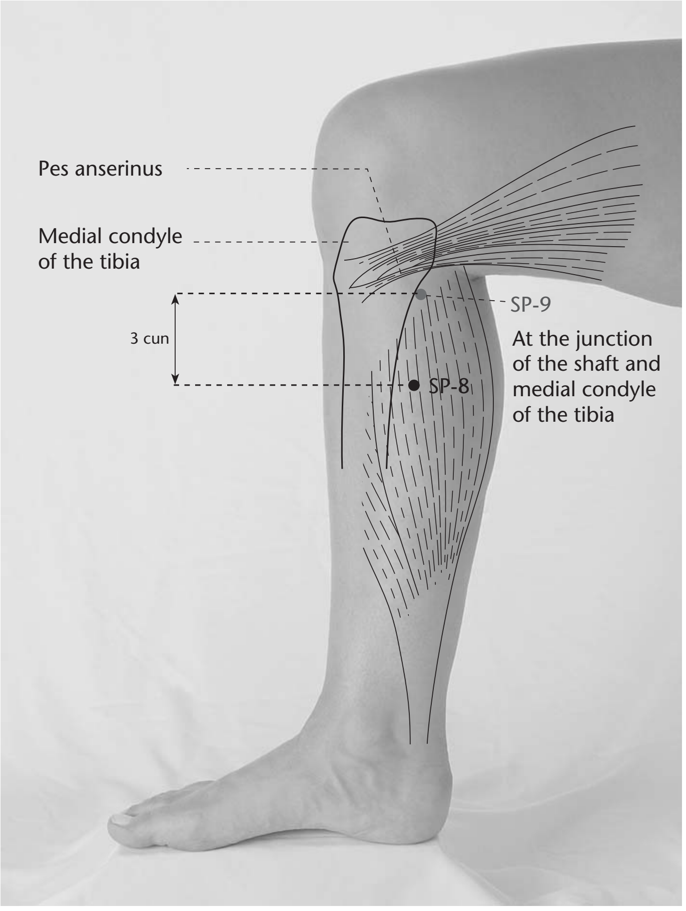
Pes anserinus
Medial condyle of the tibia
3 cun
SP-9
At the junction of the shaft and medial condyle of the tibia
SP-8

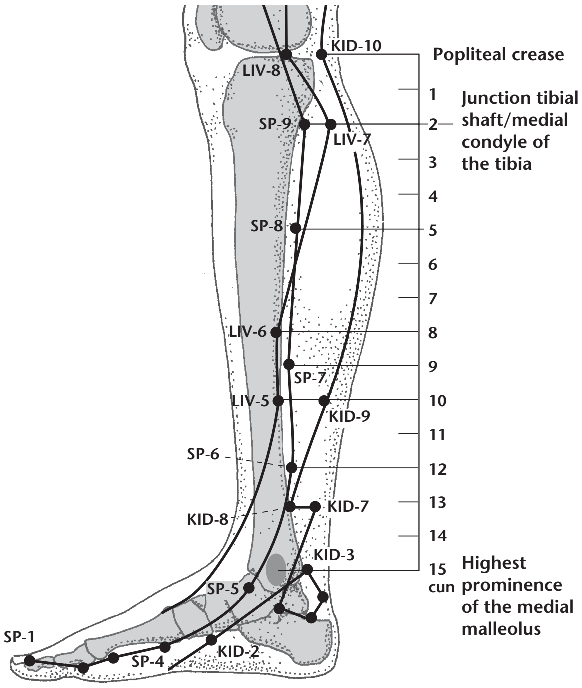
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-7
13
KID-8
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

Yin Mound Spring YINLINGQUAN SP-9

Location

With the knee flexed, this point is located in a depression distal to the medial condyle of the tibia, at the junction of the shaft and the medial condyle.

How to find

Patient's position: Preferably locate and needle **SP-9** with the knee bent and the hip slightly externally rotated, so that the knee slightly lifts from the surface, or rest the knee in an externally rotated position on a knee support. **Location**: Palpate along the posterior border of the medial crest of the tibia towards the head of the tibia to locate **SP-9** in a pressure-sensitive depression at the junction of the shaft and the head of the tibia. This depression is located between the pes anserinus superficialis and the belly of the gastrocnemius muscle (→ 3.6.1).

Located on the same level but 1 cun posterior to **SP-9** is → **LIV-7**.

Located on the lateral aspect of the lower leg is → **G.B.-34** (anterior and inferior to the head of the fibula).

Needling

1–1.5 cun vertically, along the posterior border of the tibia.

Actions/Indications

- Regulates the Spleen, transforms Dampness, regulates the water passages, benefits the Lower Burner
- As local point for disorders of the knee, especially if there are swellings

Special features

He-sea point, Water point. One of the major points for eliminating Dampness anywhere in the body. Important local point for disorders of the knee. **SP-9** is located approximately on the same level as → **G.B.-34**. If indicated, this point tends to be very tender upon pressure.

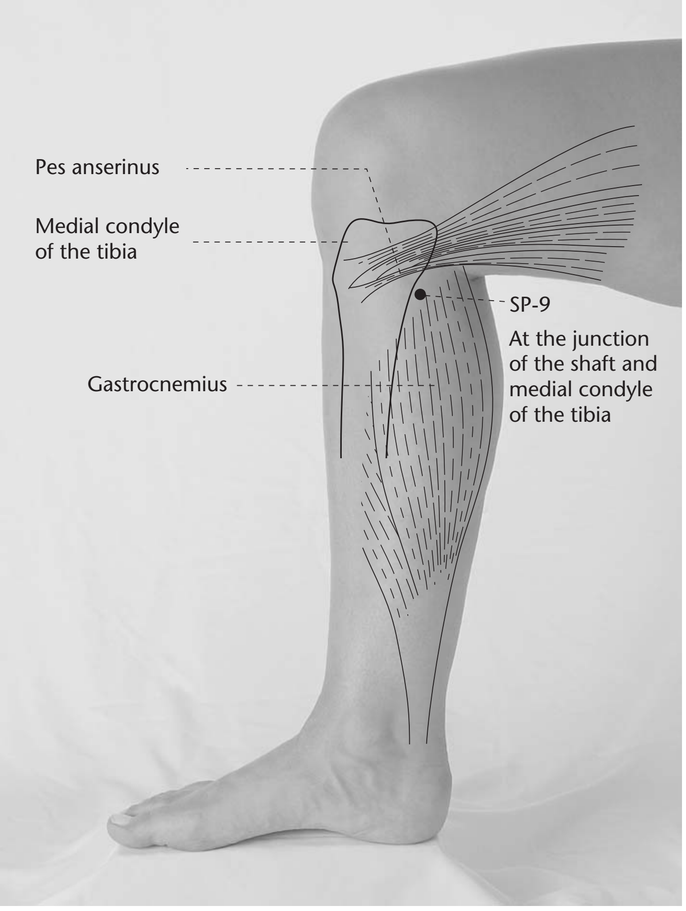
Pes anserinus
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

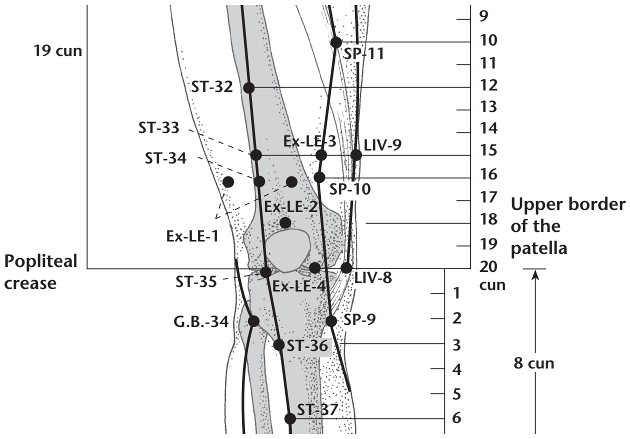
19 cun
ST-32
ST-33
ST-34
Ex-LE-1
Popliteal crease
ST-35
G.B.-34
SP-11
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-4
LIV-8
SP-9
ST-36
ST-37
9
10
11
12
13
14
15
16
17
18
19
20
cun
1
2
3
4
5
6
Upper border of the patella
8 cun

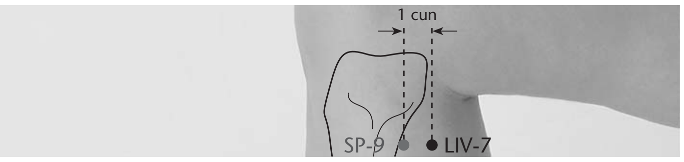
1 cun
SP-9
LIV-7

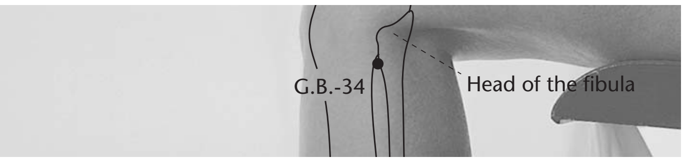
G.B.-34
Head of the fibula

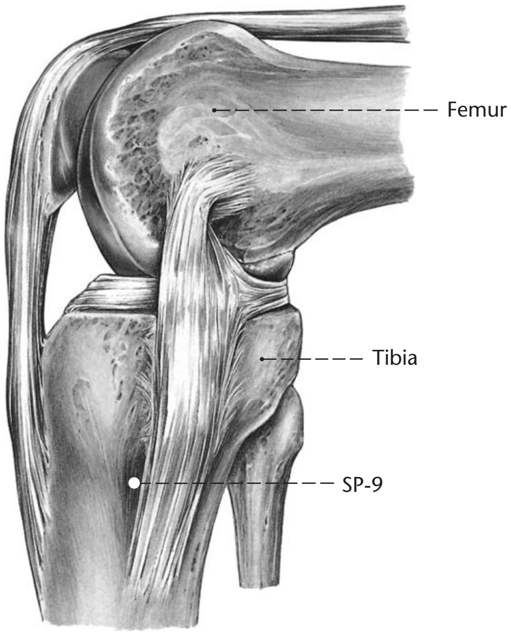
Femur
Tibia
SP-9

SP-10 *Sea of Blood* XUEHAI

Location

With the knee flexed, **SP-10** is located 2 cun proximal and slightly medial to the medial superior border of the patella, in a depression on the vastus medialis muscle.

How to find

Measure 2 cun proximal to the medial superior border of the patella and locate **SP-10** by palpating slightly medially for a soft depression. **Or**: The practitioner places their left hand on the patient's right patella, with the heel of the palm level with the inferior patellar border and the extended fingers pointing in a superior direction. With the thumb and index finger forming an angle of 45°, the tip of the thumb will point to **SP-10**.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Strengthens the Blood, dispels Blood stasis, cools the Blood, stops bleeding, regulates menstruation, benefits the skin
- Local point for disorders of the knee

Special features

Important point for regulating the Blood

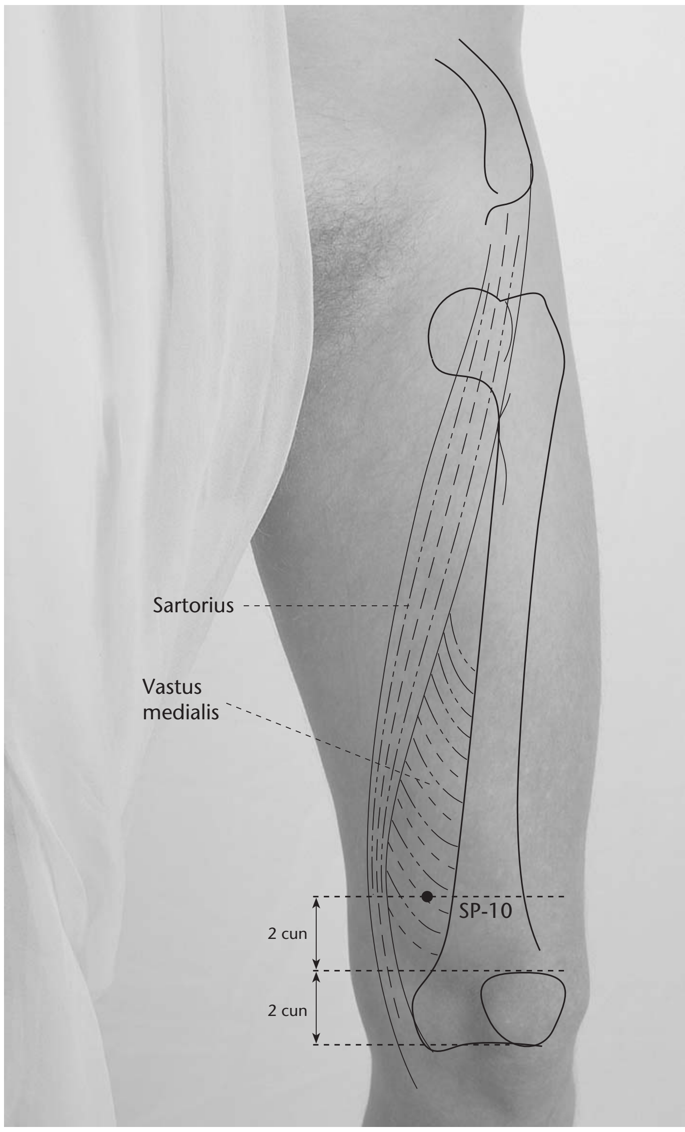

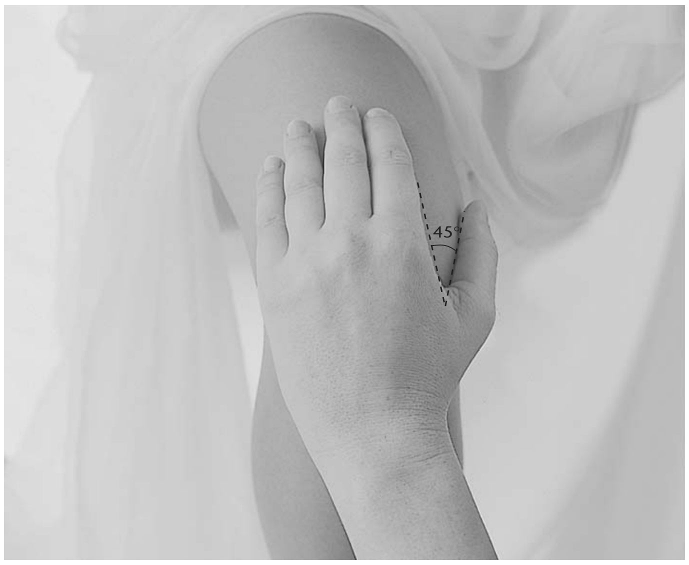

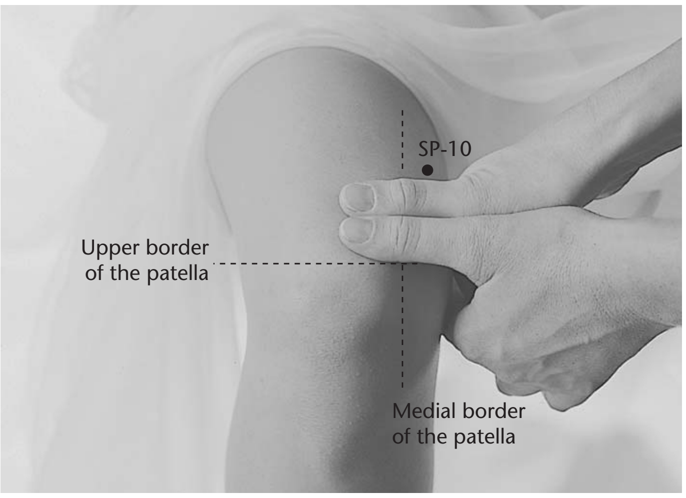

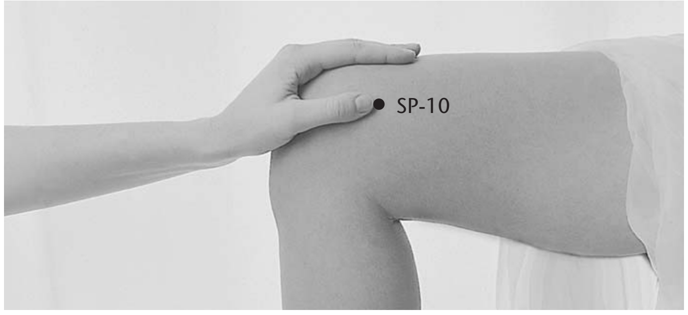

Winnowing Gate JIMEN **SP-11**

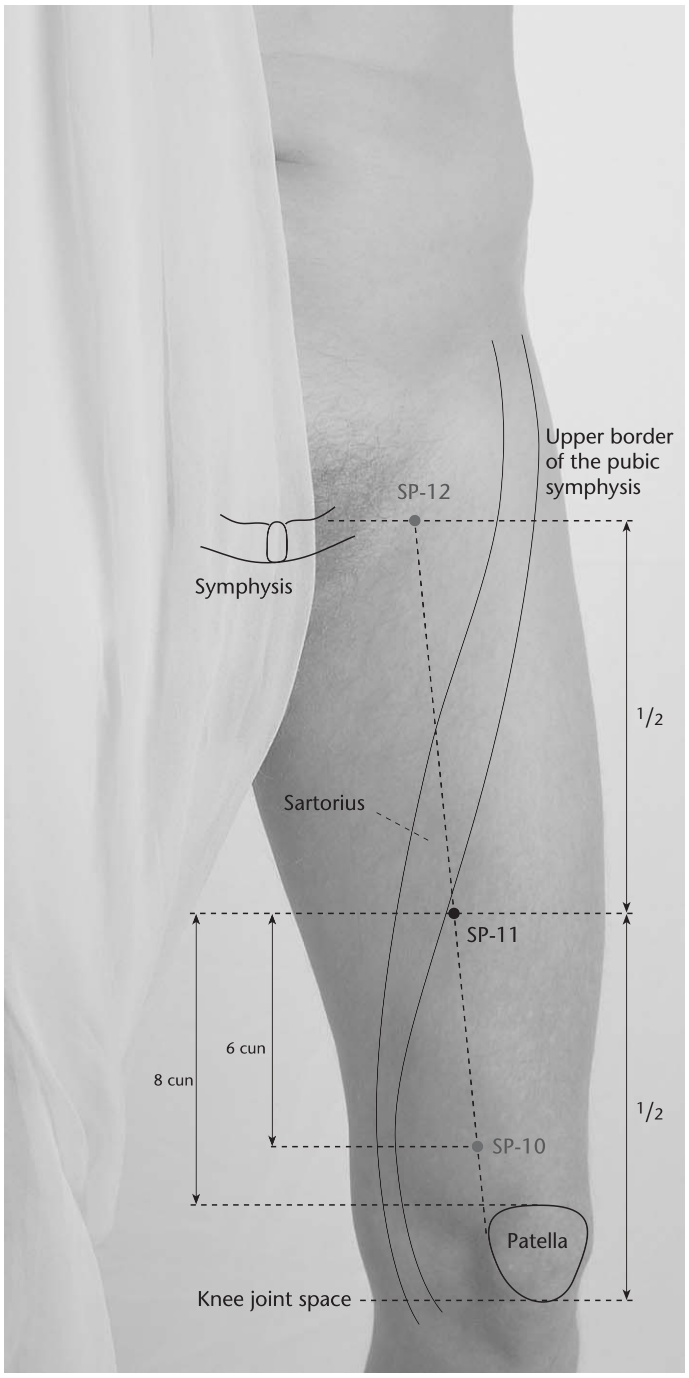
Upper border
of the pubic
symphysis
SP-12
Symphysis
1/2
Sartorius
SP-11
6 cun
8 cun
1/2
SP-10
Patella
Knee joint space

Location

6 cun proximal to → **SP-10** or 8 cun proximal to the medial upper border of the patella, at the midpoint of the femur between the sartorius and vastus lateralis muscles.

How to find

First, locate → **SP-10** by measuring 2 cun proximal and 1 cun medial from the medial upper border of the patella. From there, measure 6 cun vertically in a proximal direction to locate **SP-11** in a depression on the lateral border of the sartorius muscle. **Or**: Spreading hands technique (→ 2.3.3): **SP-11** is located midway between the knee joint space and → **SP-12** (2 cun lateral to the anterior midline, on a level with the upper border of the pubic symphysis).

Needling

Vertically 0.5–1 cun. Caution: Femoral artery/vein.

Actions/Indications

- Regulates the Lower Burner and urination, drains Dampness, clears Heat

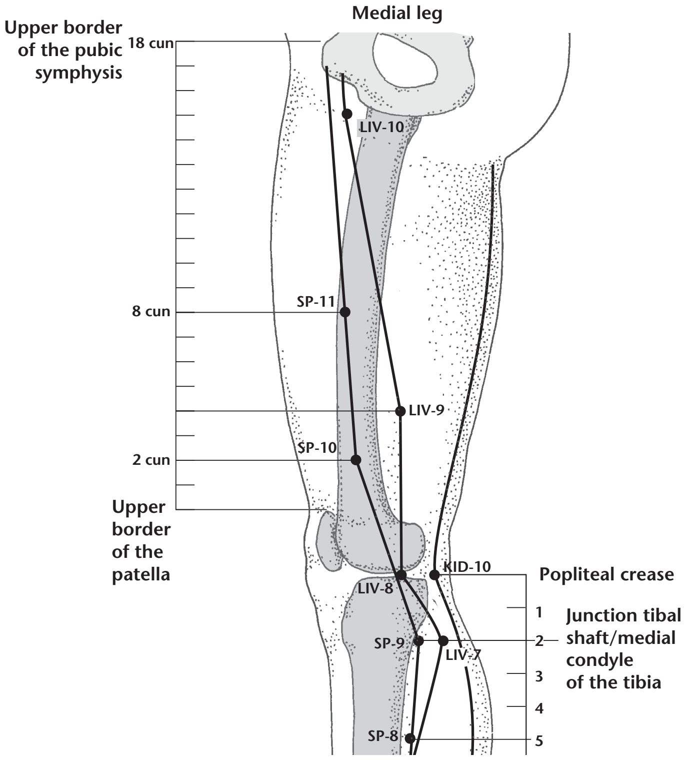
Medial leg
Upper border
of the pubic
symphysis
18 cun
LIV-10
SP-11
8 cun
LIV-9
SP-10
2 cun
Upper
border
of the
patella
KID-10
Popliteal crease
LIV-8
1
Junction tibal
shaft/medial
condyle
of the tibia
2
SP-9
LIV-7
3
4
SP-8
5

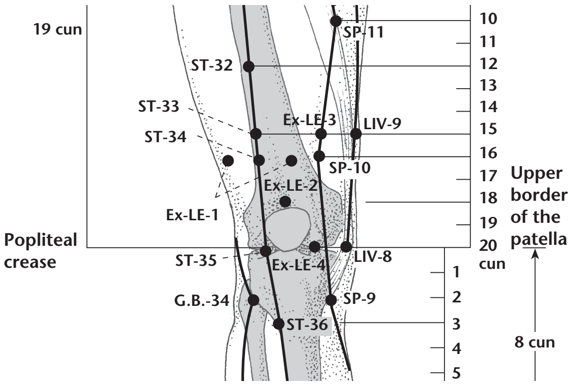
19 cun
SP-11
10
11
ST-32
12
13
ST-33
14
Ex-LE-3
LIV-9
ST-34
15
16
SP-10
17
Upper
border
of the
patella
Ex-LE-2
18
Ex-LE-1
19
Popliteal
crease
20
cun
ST-35
LIV-8
Ex-LE-4
1
G.B.-34
SP-9
2
ST-36
3
4
8 cun
5

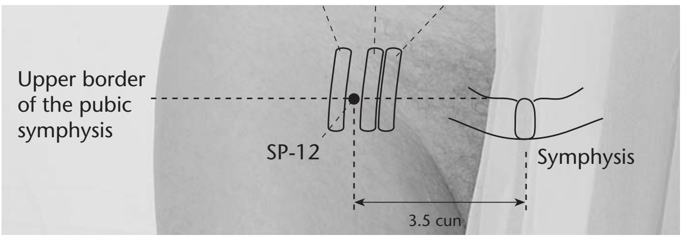
Upper border
of the pubic
symphysis
SP-12
Symphysis
3.5 cun

SP-12 *Rushing Gate* CHONGMEN

Location

3.5 cun lateral to the midline, at the level of the upper border of the pubic symphysis, lateral to the femoral artery.

How to find

From the upper border of the pubic symphysis (→ **Ren-2**), measure 3.5 cun in a lateral direction. There, you can feel the pulsation of the femoral artery. **SP-12** is located in a depression immediately lateral to the artery.

Located on the same level are **Ren-2** (on the midline), → **KID-11** (0.5 cun lateral to the midline) and → **ST-30** (2 cun lateral to the midline). → **LIV-12** is located 2.5 cun lateral to the midline, but 1 cun inferior to the upper border of the pubic symphysis.

Needling

Vertically 0.5–1 cun. Caution: Needling in a medial direction may puncture the artery; needling in a lateral direction may injure the femoral nerve.

Actions/Indications

- Regulates Qi and Blood in the Lower Burner
- Descends rising fetal Qi
- Clears Heat and drains Dampness

Special features

Meeting point with the LIV and T.B.* channels and the *yin wei mai*

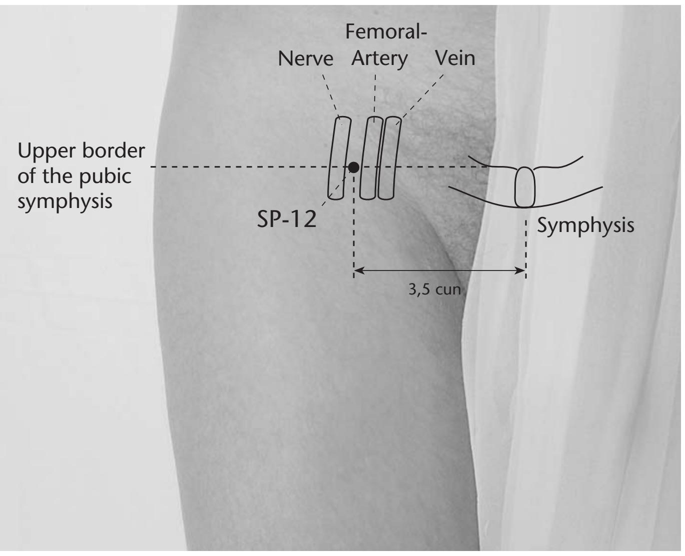
Femoral-
Nerve
Artery
Vein
Upper border of the pubic symphysis
SP-12
Symphysis
3,5 cun

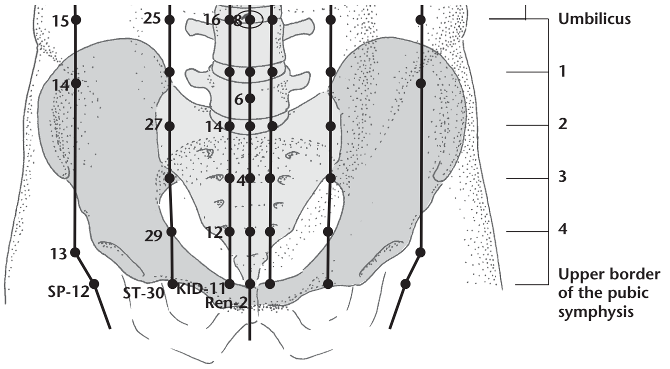
15
25
16
8
6
14
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

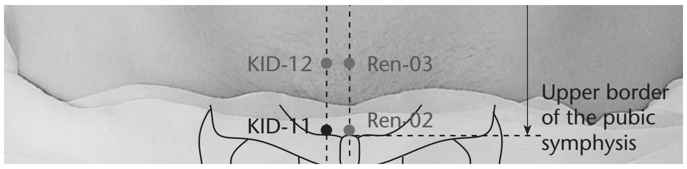
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

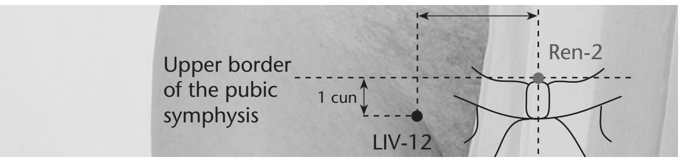
Ren-2
Upper border of the pubic symphysis
1 cun
LIV-12

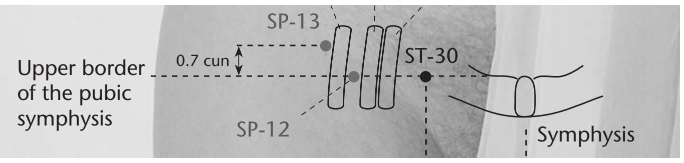
SP-13
0.7 cun
ST-30
Upper border of the pubic symphysis
SP-12
Symphysis

---

* According to some authors

Abode of the **Fu** FUSHE **SP-13**

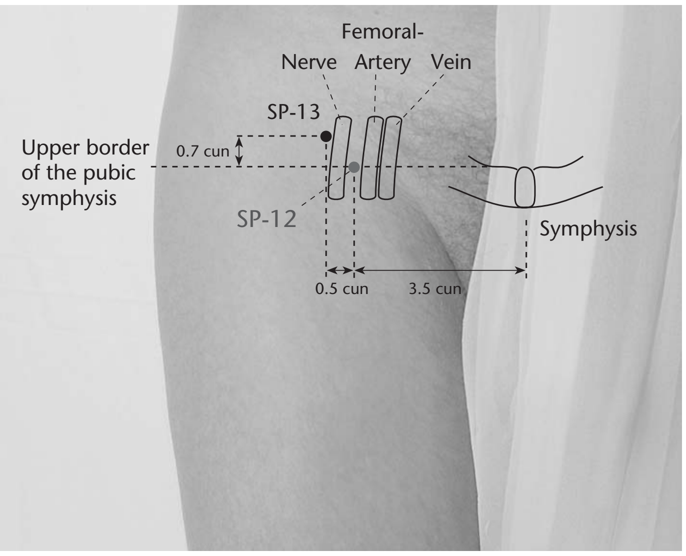
Femoral-
Nerve
Artery
Vein
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

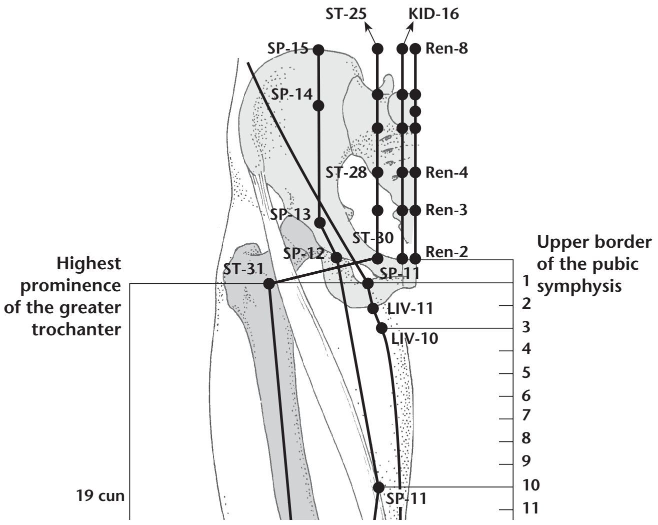
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
SP-13
ST-30
SP-12
Ren-2
Upper border of the pubic symphysis
Highest prominence of the greater trochanter
ST-31
SP-11
1
2
LIV-11
3
LIV-10
4
5
6
7
8
9
10
19 cun
SP-11
11

Location

4 cun lateral to the anterior midline and 0.7 cun superior to the upper border of the pubic symphysis.

How to find

First, measure 4 cun in a lateral direction from the upper border of the pubic symphysis and from there 0.7 cun in a superior direction to locate **SP-13** in a depression in the inguinal groove. **SP-13** is located 0.5 cun lateral and 0.7 cun superior to → **SP-12** (→ 3.5 cun lateral to the upper border of the pubic symphysis).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum.

Actions/Indications

- Disperses accumulations
- Regulates the Qi

Special features

Meeting point with the *yin wei mai* and the Liver channel

SP-14 *Abdomen Knot* FUJIE

Location

4 cun lateral to the anterior midline, on the mamillary line, 3 cun superior to → **SP-13** or 1.3 cun inferior to → **SP-15**.

How to find

First, measure 4 cun from the umbilicus in a lateral direction in order to locate the mamillary line. This is the point → **SP-15**. From there, measure 1.3 cun in an inferior direction to locate **SP-14**.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Moves Qi and Blood, warms the middle, dispels Cold
- Regulates the Lower Burner and descends counterflow Qi

Special features

Sensitive to pressure with appendicitis

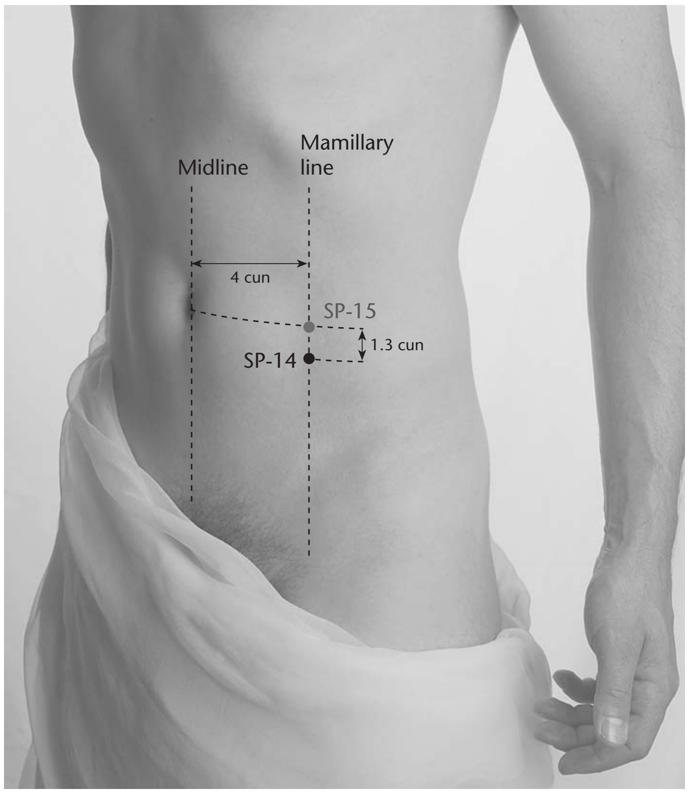
Midline
Mamillary line
4 cun
SP-15
1.3 cun
SP-14

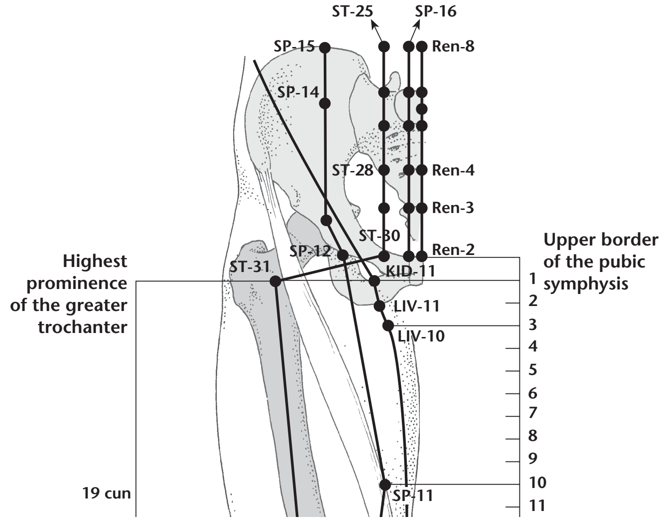
ST-25
SP-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
Ren-2
SP-12
ST-31
KID-11
Highest prominence of the greater trochanter
Upper border of the pubic symphysis
1
2
3
4
5
6
7
8
9
10
11
LIV-11
LIV-10
19 cun
SP-11

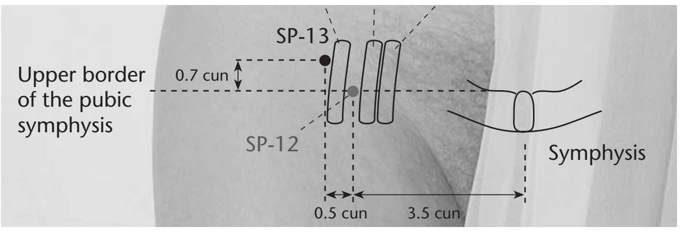
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

Great Horizontal DAHENG SP-15

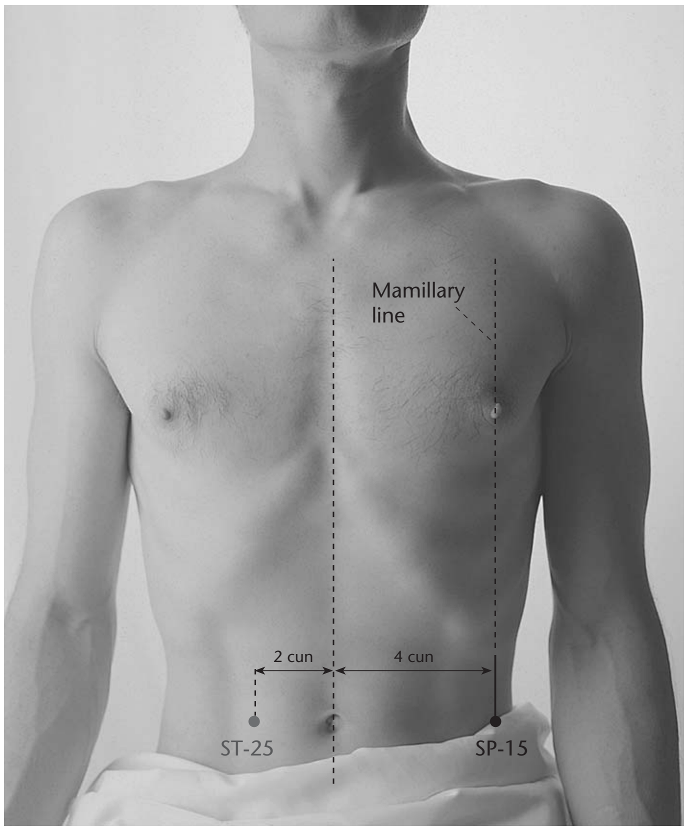
Mamillary line
2 cun
4 cun
ST-25
SP-15

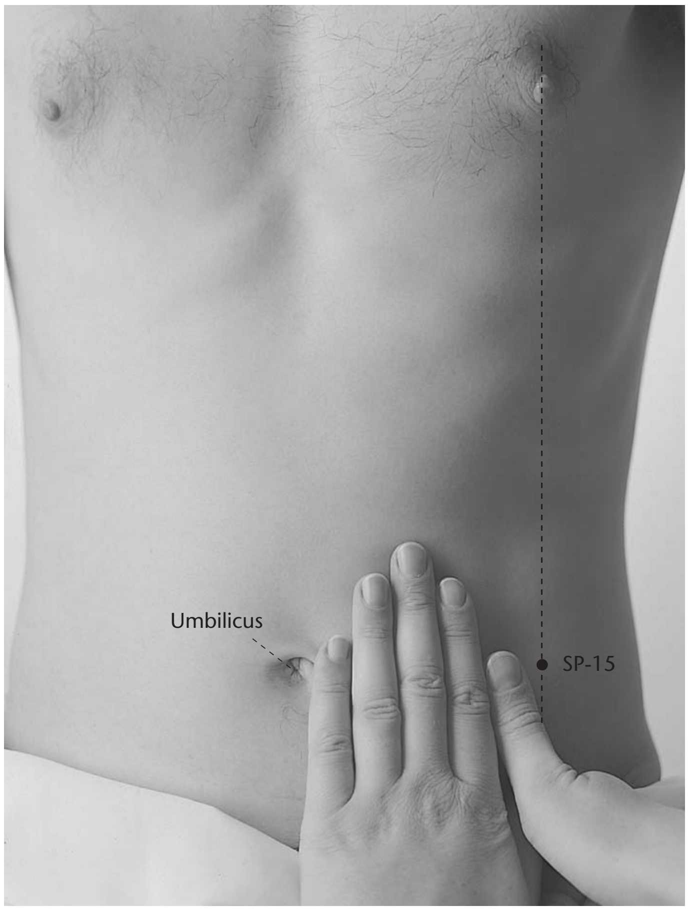
Umbilicus
SP-15

Location

4 cun lateral to the centre of the umbilicus, on the mamillary line.

How to find

Locate **SP-15** 4 cun lateral to the umbilicus, using the mamillary line for reference, especially in male patients.
Located on the same level are → **Ren-8/KID-16/ST-25/G.B.-26** (umbilicus/0.5 cun lateral to the umbilicus/2 cun lateral to the umbilicus/below the free end of the 11th rib at the level of the umbilicus).

Needling

Vertically 1–1.5 cun. Caution: Peritoneum, pregnancy. Deep needling may reach the aponeuroses of the transverse abdominal musculature.

Actions/Indications

- Regulates and moves Qi in the Intestines (promotes bowel movement by stimulating peristalsis)

Special features

Meeting point with the *yin wei mai*. Important point for regulating the Large Intestine Qi.

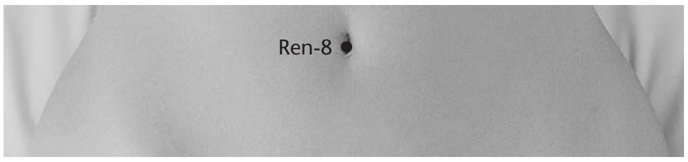

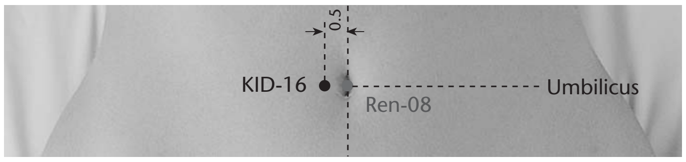
0.5
KID-16
Umbilicus
Ren-08

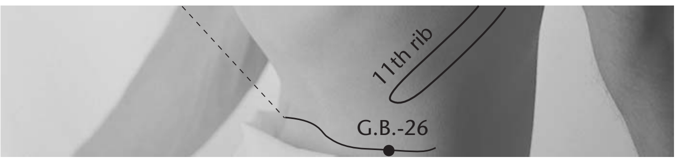
11th rib
G.B.-26

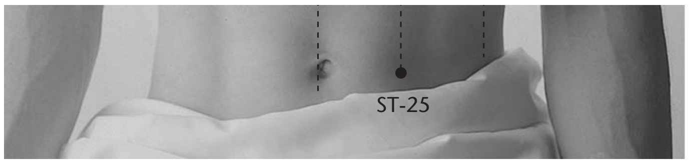

Ch04.4-F10028.qxd 2/23/08 7:09 PM Page 196

SP-16 *Abdomen Sorrow* FUAI

Location

3 cun above the centre of the umbilicus and 4 cun lateral to the anterior midline, on the mamillary line.

How to find

The distance between the sternocostal angle and the centre of the umbilicus can be divided into 8 proportional cun, which can differ considerably from the patient's finger cun; therefore, only proportional cun measurements should be applied (elastic tape → 2.3.1). First, measure 3 cun from the umbilicus in a superior direction, then 4 cun in a lateral direction, where **SP-16** is located on the mamillary line. Depending on the shape of the thorax, this point will either be located on the abdomen or it will project onto the lower border of the ribcage.

Located on the same level are → **Ren-11** (anterior midline), → **KID-18** (0.5 cun lateral to the anterior midline) and → **ST-22** (2 cun lateral to the anterior midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy. Depending on the shape of the thorax, the correct location of this point might be on the costal cartilage. In this case, either transverse (subcutaneous) needling or locate the point more medially (or choose an alternative point).

Actions/Indications

- Regulates Qi in the Intestines
- Drains Dampness

Special features

Meeting point with the *yin wei mai*

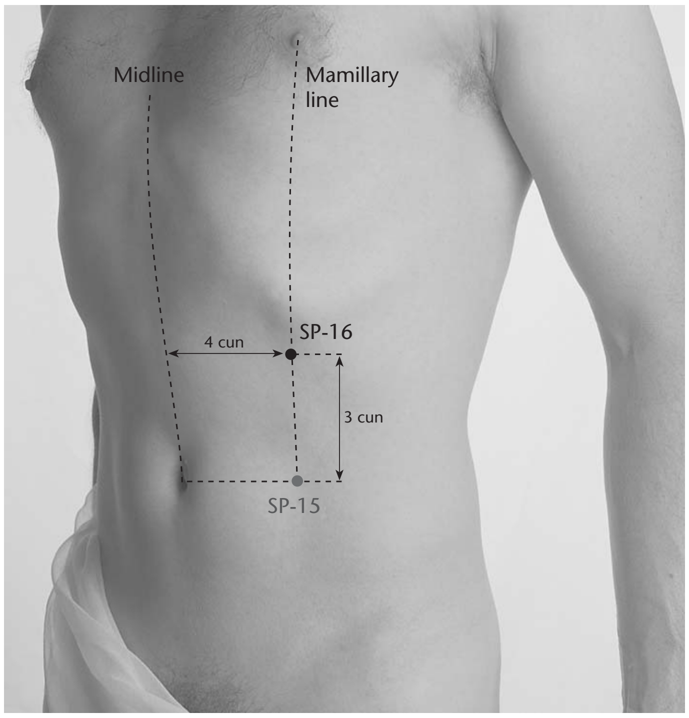
Midline
Mamillary line
4 cun
SP-16
3 cun
SP-15

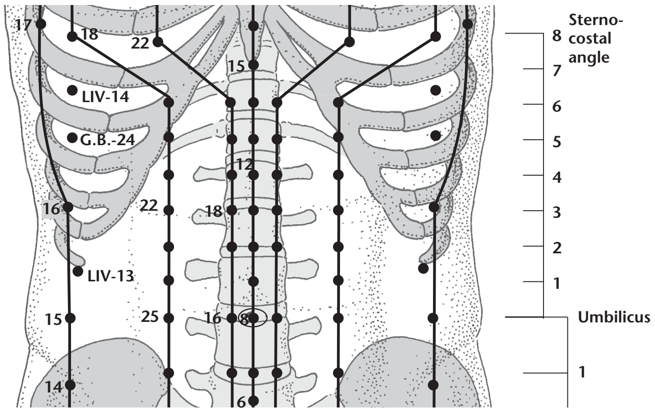
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
6
8
Sterno-costal angle
7
6
5
4
3
2
1
Umbilicus
1

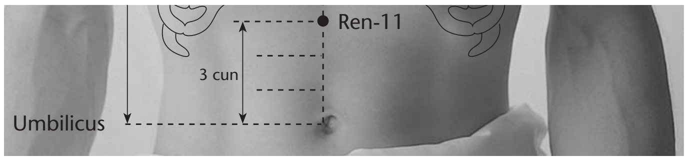
Ren-11
3 cun
Umbilicus

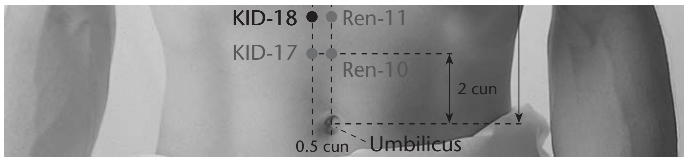
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

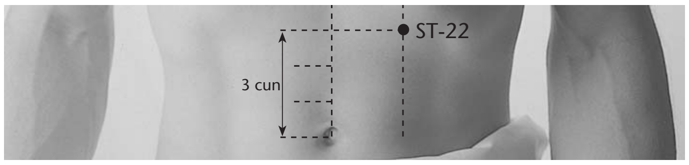

Ch04.4-F10028.qxd 2/23/08 7:09 PM Page 197

Food Cavity SHIDOU SP-17

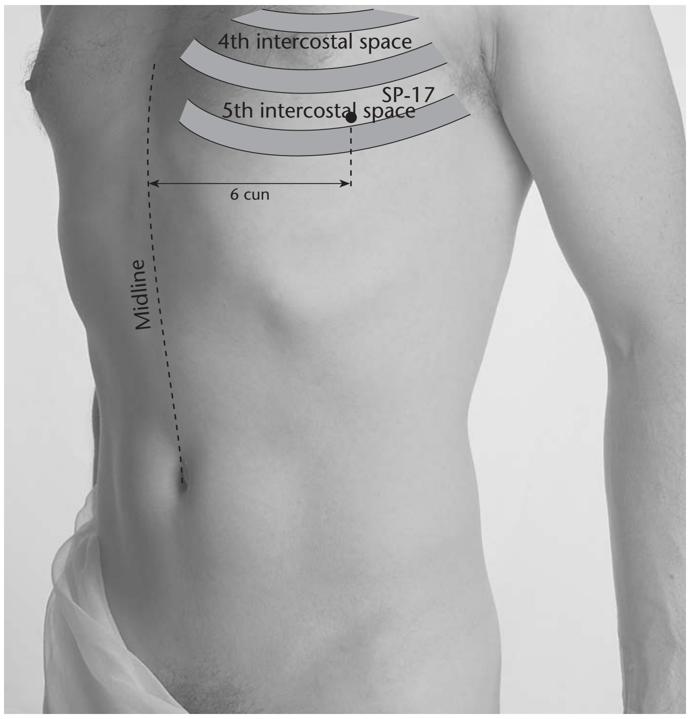
4th intercostal space
SP-17
5th intercostal space
6 cun
Midline

Location

In the 5th intercostal space, 6 cun lateral to the anterior midline (in male patients, 2 cun lateral to the nipple).

How to find

In men, the nipple is generally located on the level of the 4th intercostal space, while in women, when in a supine position, it is not always at this level. Locate **SP-17** in the 5th intercostal space, 2 cun lateral to the mamillary line (note that the intercostal space laterally curves upward). Both the clavicle and the manubriosternal synchondrosis are good reference points for locating the 5th intercostal space. **SP-17** can be located by following the course of the intercostal space in a lateral direction and measuring 2 cun lateral to the mamillary line.

Also located on the level of the 5th intercostal space are → **Ren-16** (anterior midline), → **KID-22** (2 cun lateral to the anterior midline) and → **ST-18** (4 cun lateral to the anterior midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi
- Harmonises the Stomach and eliminates stagnation
- Drains Dampness

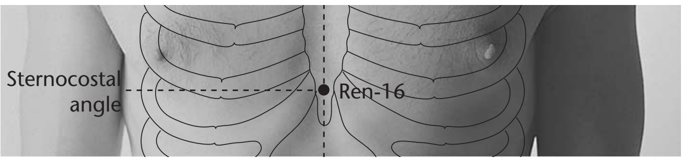
Sternocostal angle
Ren-16

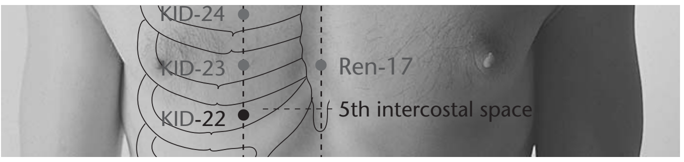
KID-24
KID-23
KID-22
Ren-17
5th intercostal space

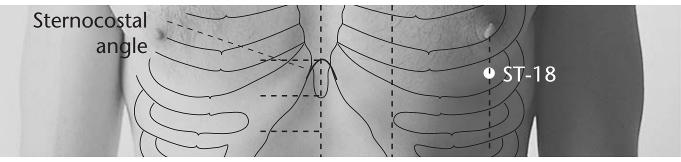
Sternocostal angle
ST-18

SP-18 *Heavenly Stream* TIANXI

Location

In the 4th intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

In men, the nipple is generally located on the level of the 4th intercostal space, while in women, when in a supine position, it is not always at this level. Follow the course of the 4th intercostal space in a lateral direction and locate **SP-18** 2 cun lateral to the nipple (note that the intercostal space laterally curves upward). In women, either the clavicle or the manubriosternal synchondrosis are good reference points for locating the 4th intercostal space (→ 3.5). **SP-18** can be located by following the course of the intercostal space in a lateral direction while measuring 6 cun laterally from the anterior midline or 2 cun laterally from the nipple.

Also located on the level of the 4th intercostal space are → **Ren-17** (anterior midline), → **KID-23** (2 cun lateral to the anterior midline), → **ST-17** (nipple), → **P-1** (1 cun lateral to the nipple) and → **G.B.-23** (1 cun anterior to **G.B.-22**).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications:

- Unbinds the chest
- Descends the Qi
- Benefits the breasts

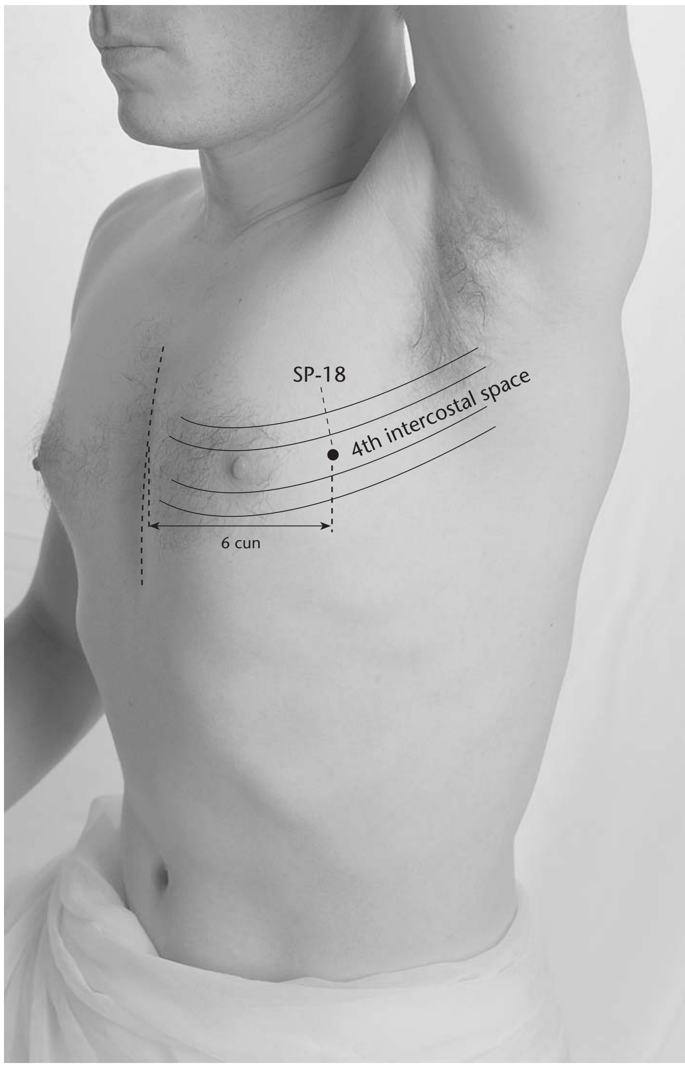
SP-18
4th intercostal space
6 cun

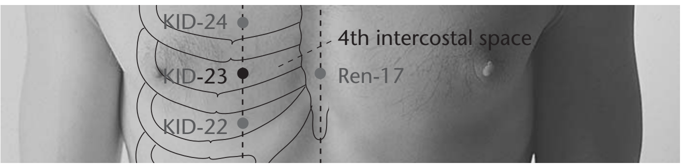
KID-24
KID-23
KID-22
4th intercostal space
Ren-17

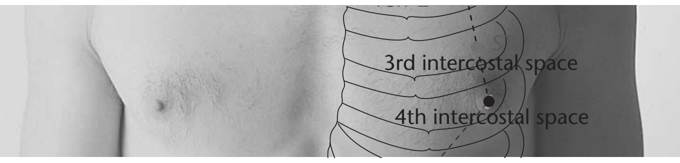
3rd intercostal space
4th intercostal space

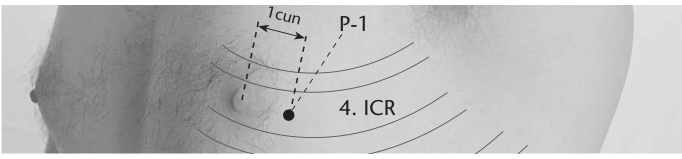
1 cun
P-1
4. ICR

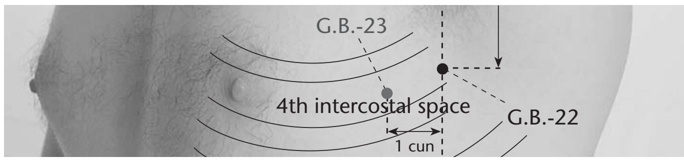
G.B.-23
4th intercostal space
G.B.-22
1 cun

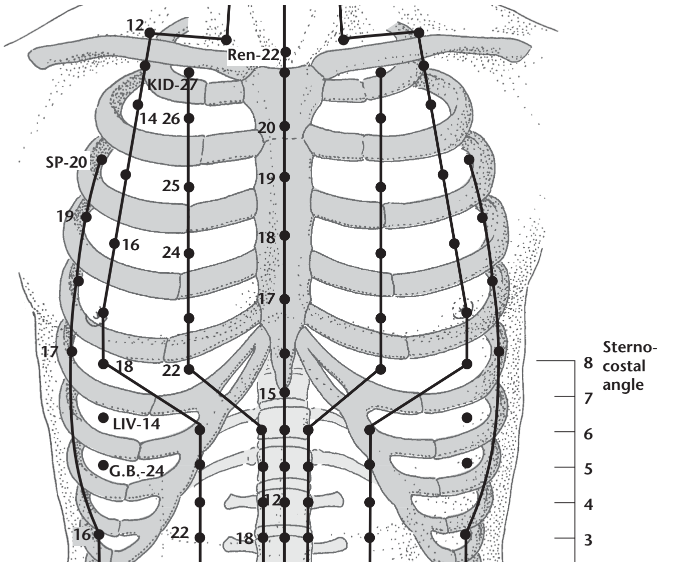
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
Sterno-costal angle
8
7
6
5
4
3

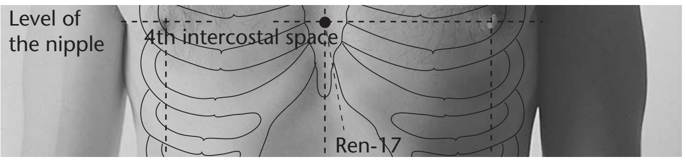
Level of the nipple
4th intercostal space
Ren-17

Chest Village **XIONGXIANG** **SP-19**

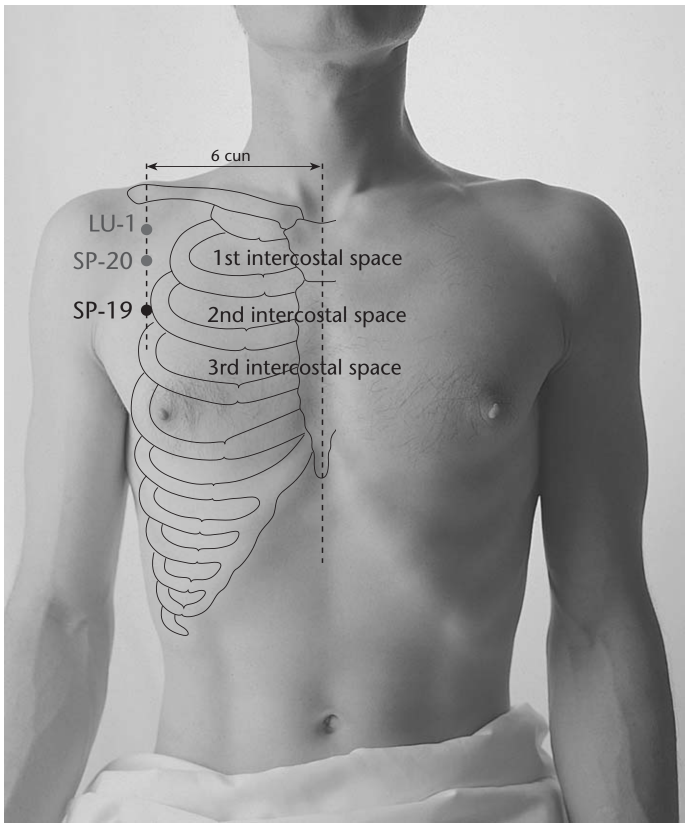
6 cun
LU-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

Location

In the 3rd intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

Locate the 3rd intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-19** 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line (note: the intercostal space laterally curves upward).

Located on the same level (note: course of the intercostal space) are → **Ren-18** (anterior midline), → **KID-24** (2 cun lateral to the midline) and → **ST-16** (4 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space, or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Disperses and descends the Lung Qi
- Unbinds the chest

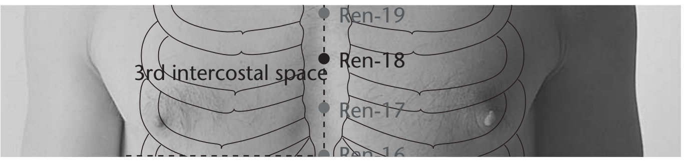
Ren-19
3rd intercostal space
Ren-18
Ren-17
Ren-16

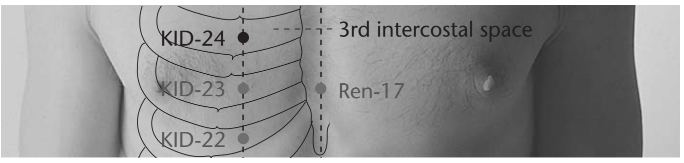
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

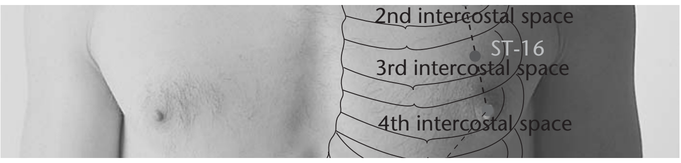
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

SP-20 *Encircling Glory* ZHOURONG

Location

In the 2nd intercostal space, 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line.

How to find

Locate the 2nd intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-20** 6 cun lateral to the anterior midline or 2 cun lateral to the midclavicular line (note: the intercostal space laterally curves upward). **SP-20** is located 1 cun inferior to → **LU-1**. Located on the same level (note: course of the intercostal space) are → **Ren-19** (anterior midline), → **KID-25** (2 cun lateral to the midline) and → **ST-15** (4 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Unbinds the chest
- Regulates the Lung Qi

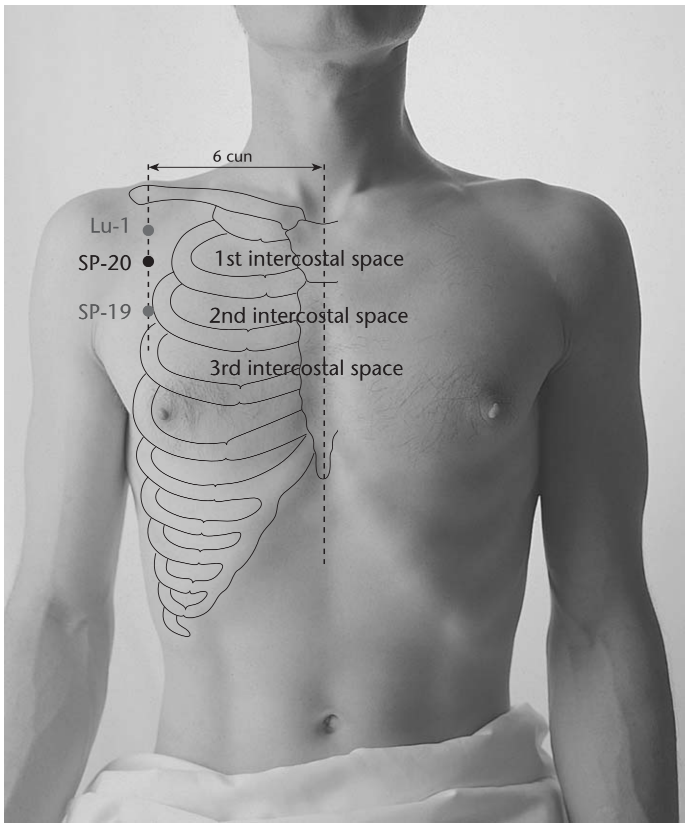
6 cun
Lu-1
SP-20
SP-19
1st intercostal space
2nd intercostal space
3rd intercostal space

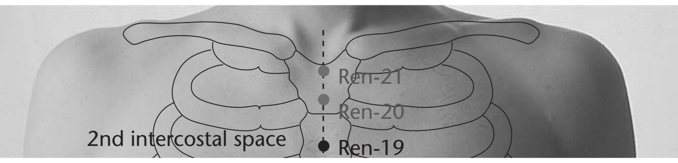
Ren-21
Ren-20
2nd intercostal space
Ren-19

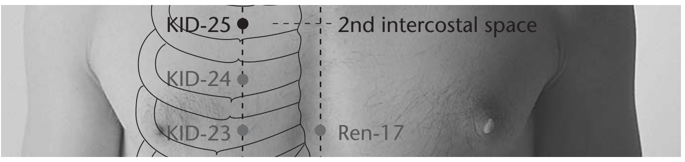
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

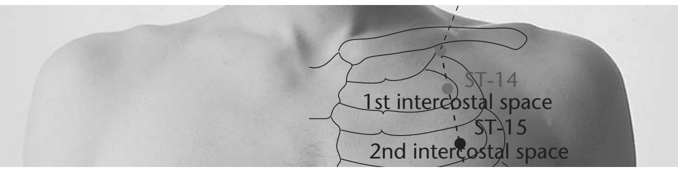
ST-14
1st intercostal space
ST-15
2nd intercostal space

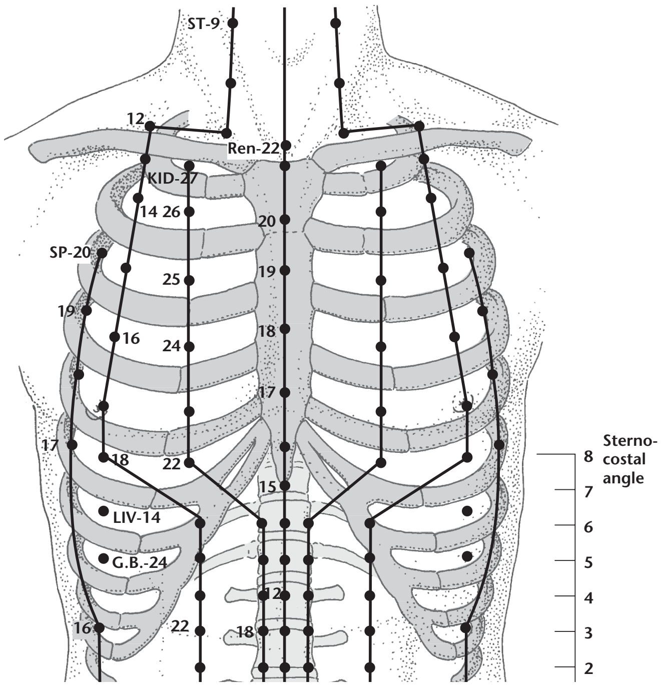
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
Sterno-costal angle
8
7
6
5
4
3
2

Great Wrapping DABAO SP-21

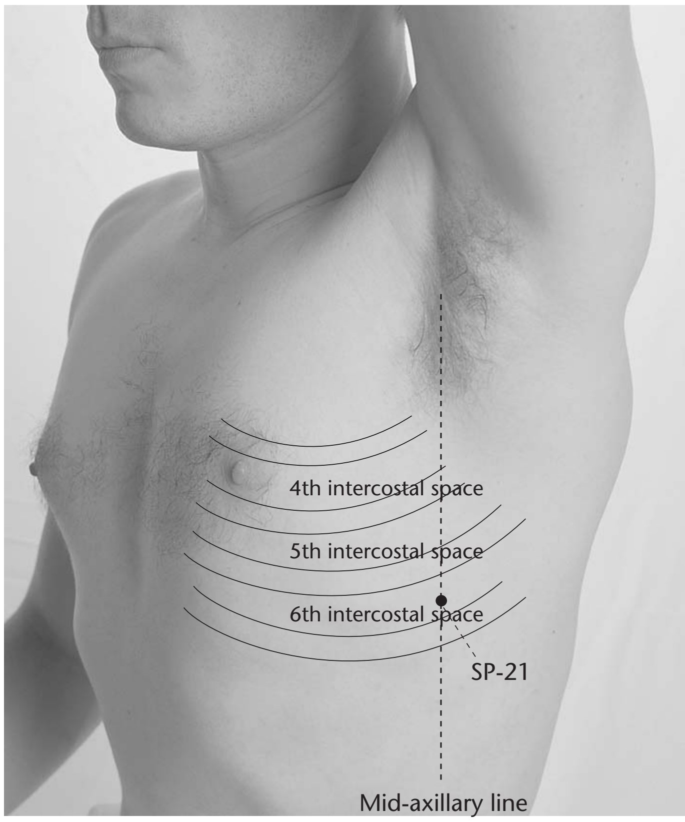
4th intercostal space
5th intercostal space
6th intercostal space
SP-21
Mid-axillary line

Location

On the midaxillary line, in the 6th intercostal space. Note: According to some texts, this point is located in the 7th intercostal space.

How to find

Locate the 6th intercostal space by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 3.5). Follow the course of the intercostal space in a lateral direction and locate **SP-21** on the axillary midline (note: the intercostal space laterally curves upward). **SP-21** is often sensitive to pressure. Or: Locate **SP-21** on the midaxillary line, midway between the axilla and the end of the 11th rib.

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously), parallel to the intercostal space or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates Qi and Blood, governs and wraps all the *luo*-connecting vessels of the body
- Unbinds the chest

Special features

Great *luo*-connecting point of the spleen (→ 8.1.2), exit point.

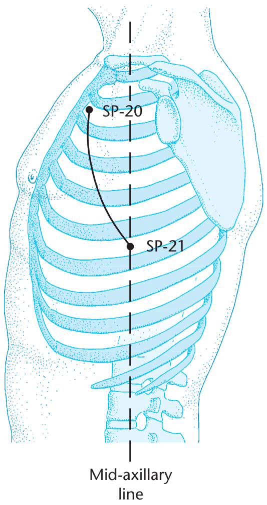
SP-20
SP-21
Mid-axillary line

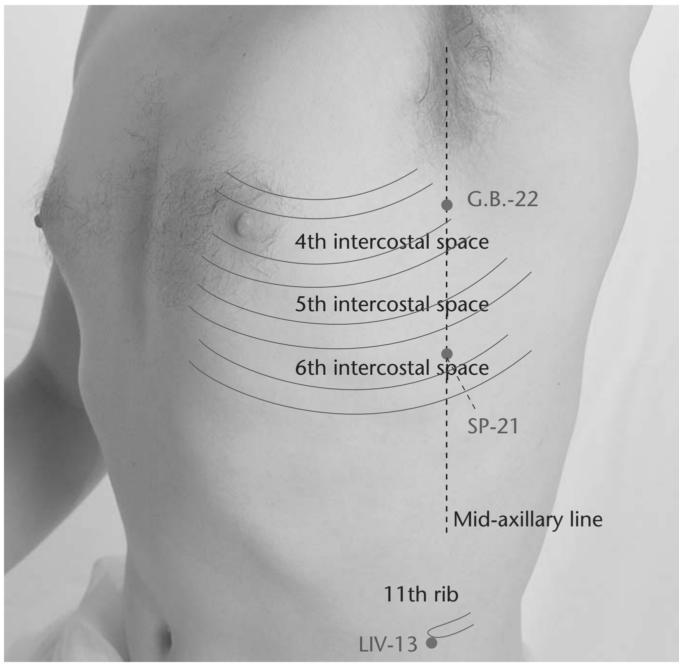
G.B.-22
4th intercostal space
5th intercostal space
6th intercostal space
SP-21
Mid-axillary line
11th rib
LIV-13

4.5 The Heart Channel System – Hand *shaoyin* (*shou shao yin jing luo*)

4.5.1 The Heart Primary Channel (*shou shao yin jing*)

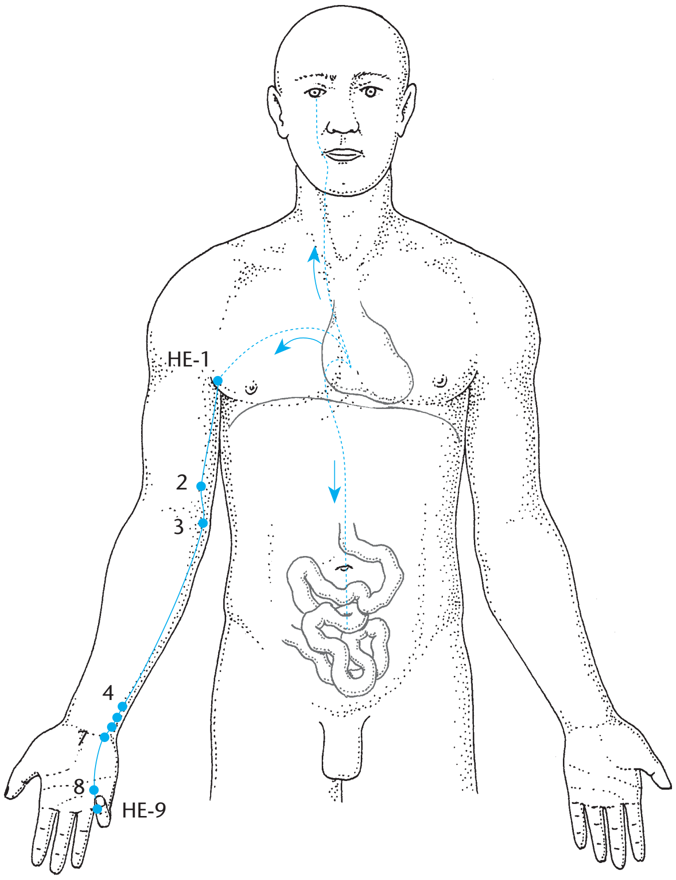
HE-1
2
3
4
7
8
HE-9

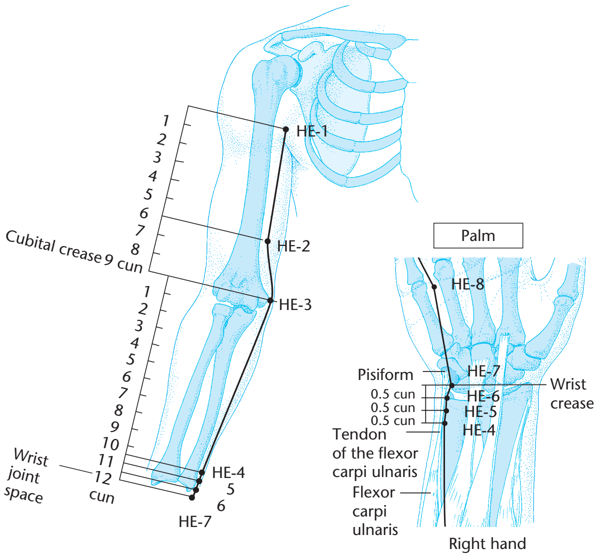
1
2
3
4
5
6
7
8
Cubital crease 9 cun
HE-1
HE-2
HE-3
1
2
3
4
5
6
7
8
9
10
11
12
Wrist joint space
cun
HE-4
5
6
HE-7
Palm
HE-8
Pisiform
HE-7
0.5 cun
0.5 cun
0.5 cun
HE-6
HE-5
HE-4
Wrist crease
Tendon of the flexor carpi ulnaris
Flexor carpi ulnaris
Right hand

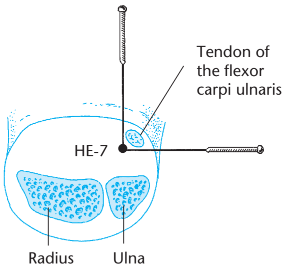
Tendon of the flexor carpi ulnaris
HE-7
Radius
Ulna

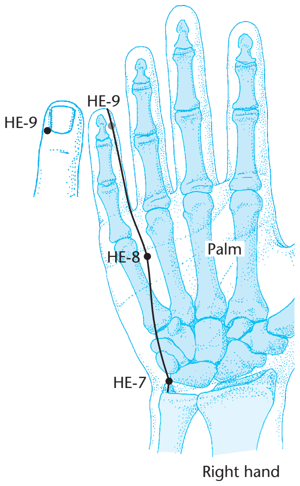
HE-9
HE-9
HE-8
Palm
HE-7
Right hand

Pathway

The **internal** pathway of the Heart primary channel originates in its related *zang*-Organ, the Heart (*xin*). An **internal branch** of the Spleen primary channel reaches the Heart and connects with the Heart primary channel (deep Yin–Yin connection ➞ 1.2).

The **internal pathway** of the Heart primary channel splits into three branches:

- ➡ **the first branch** penetrates the diaphragm and connects with the internally–externally related *fu*-Organ, the Small Intestine (*xiaochang*)
- ➡ **the second branch** ascends along the oesophagus and the neck region, disperses in the tissues around the eyes and, according to some authors, also spreads inside the brain.
- ➡ **the main branch** runs from the Heart to the Lung (*fei*) and descends again to emerge in the centre of the axilla at ➞ **HE-1** (*jiquan*) (beginning of the external pathway).

From here the **external pathway** runs along the anteromedial aspect of the upper arm and forearm, traverses the palm between the fourth and fifth metacarpal bones, continues along the radial aspect of the little finger and ends at the radial corner of its nail at **HE-9** (*shaoyang*).

From **HE-9** (*shaoyang*) a branch crosses to the ulnar border of the little finger's nail to ➞ **S.I.-1** (*shaoze*) (hand Yin–Yang connection of the second great circuit).

Clinical importance (➞ 1.2)

Exterior (*biao*) signs and symptoms: General fever, headaches, pain in the thorax and back, thirst with a desire to drink, Heat sensation in the palms, Cold invasion in the extremities, pain in the shoulder and the medial aspect of the upper arm

Interior (*li*) or *zangfu*-Organ signs and symptoms: Heart pain, pain and fullness in the thorax and lateral costal region, hypochondriac pain, irritability, shortness of breath, restlessness, dizziness, mental confusion

Connections of the Heart primary channel (→ 1.2)

Connections with other channels

Small Intestine primary channel (*shou tai yang jing*)

Connection: Hand Yin–Yang connection of the second great circuit
Location: H-9 → S.I.-1 (on the hand).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Kidney primary channel (zu shao yin jing)

Connection: Paired according to the six channel theory (hand–foot pairing): *shao yin* (Yin axes of the second great circuit)
Location: KID → HE (on the thorax). An internal branch of the Kidney primary channel travels to the Liver (*gan*), penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung a branch runs to Heart (*xin*) and there connects with the Heart primary channel.
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Spleen primary channel (*zu tai yin jing*)

Connection: Deep Yin–Yin connection
Location: SP → HE (on the thorax). An **internal** branch of the Spleen primary channel spreads in the Heart (*xin*) and connects with the Heart primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Heart primary channel receives Nutritive Qi (*ying qi*) from the Spleen primary channel (first circulation of the *ying qi* → 1.1.4)

Connections with other *zangfu*-Organ systems

Small Intestine (*xiaochang*), Heart (*xin*), Lung (*fei*)

4.5.2 The Heart Divergent Channel (*zu shao yin jing bie*)

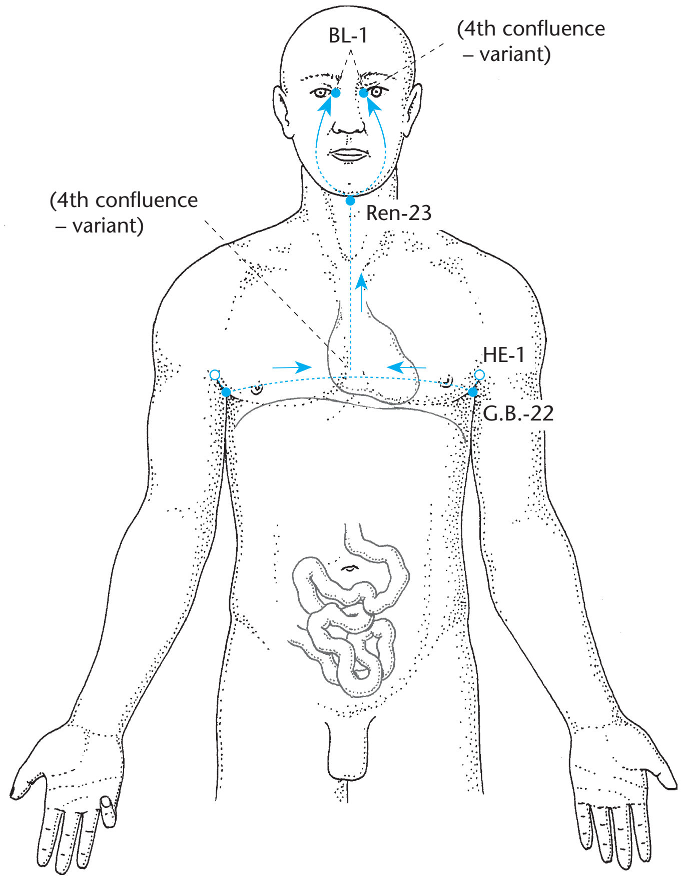

Pathway

The Heart divergent channel separates from the Heart primary channel at **HE-1** (*jiquan*) in the axilla

- ➡ penetrates the thorax at → **G.B.-22** (*yuanyue*) on the mid-axillary line and 3 cun distal to the axilla
- ➡ and reaches the Heart (*xin*) (see commentary below).
- ➡ It then ascends to the neck and crosses → **Ren-23** (*lianquan*)
- ➡ disperses in the cheek and reaches the medial canthus of the eye at → **BL-1** (*jingming*), where, according to most authors, it connects with a small branch coming from → **S.I.-18** (*quanliao*) to form one of the six he-confluences (here: S.I./BL as fourth confluence → 1.3).

Commentary: according to Solinas et al (1998) the Heart divergent channel meets the Small Intestine divergent channel and the Small Intestine primary channel at the Heart (*xin*) in order to bind with one of the six he-confluences.

Clinical importance

- Strengthens the relationship between the Heart and the thoracic region. Points on the Heart primary channel mainly treat disorders of the Heart and the thorax. There is no connection with the Small Intestine (*xiaochang*).

- Connects the Heart primary channel with the neck. Points on the Heart primary channel can be used to treat disorders of the throat and vocal chords such as laryngitis, pharyngitis, etc. Use points such as **HE-5, HE-8**.
- Strengthens the connection between the Heart and the inner canthus of the eye. Heart points can thus be used for disorders of the eyes.

4.5.3 The Heart Sinew Channel (*shou shao yin jing jin*)

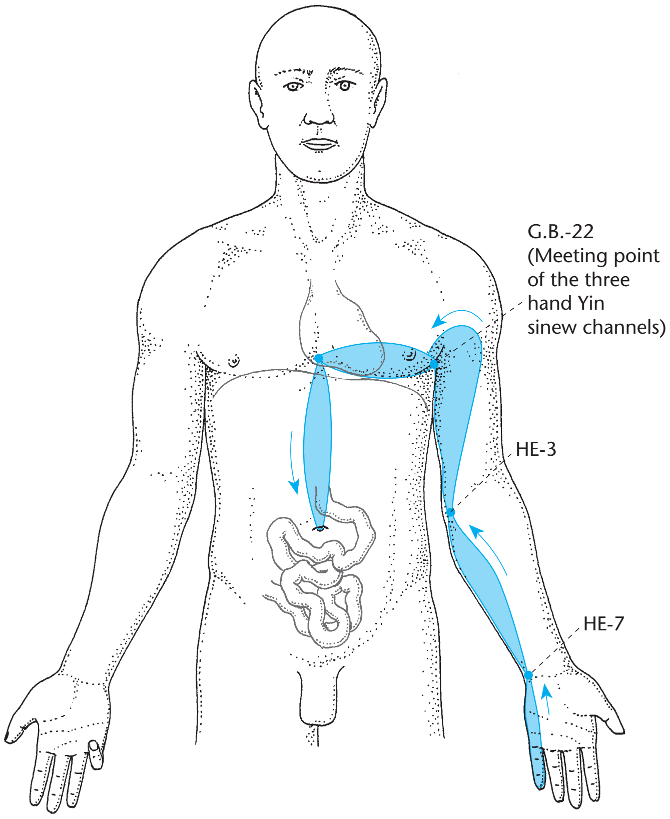

Pathway

The Heart sinew channel begins on the radial aspect of the little finger

➡ continues along the anterior and radial aspect of the little finger and the hand

➡ binds (*jie*) at the pisiform bone at **HE-7** (*shenmen*)

➡ ascends the anterior and ulnar aspect of the forearm, binds (*jie*) at the medial epicondyle at **HE-3** (*shaohai*) and continues to the axillary region, where it meets the Lung sinew channel, binds (*jie*) in the axilla and meets the other hand Yin sinew channels near → **G.B.-22** (*yuanye*)

➡ from here it continues at a deeper level, crosses the thoracic region, binds (*jie*) at → **Ren-17** (*shanzhong*), descends to the upper cardiac orifice, penetrates the diaphragm and terminates in the umbilical region.

Clinical importance

Pathology: Stiffness, cramps and pain along the pathway of the Heart sinew channel. Internal cramping in the cardiac region, for example disorders with angina-like pain due to stress, hiatal hernia and disorders of the stomach.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the pathway of the Heart channel. The area covered by the Heart sinew channel is larger than that covered by the Heart primary channel. This explains why Heart primary channel points can also be used for disorders and diseases of the thorax and the chest, for example pain and tightness in the chest, mastitis and mastopathy. Since the Heart sinew channel also penetrates the diaphragm, it is also indicated for disorders such as hiccups, acid reflux and oesophagitis. Due to its connection with the umbilical region, the Heart sinew channel has a direct link to the Original Qi (*yuan qi*): if the Heart (*shen*) has suffered a shock, for example due to fainting or collapse, direct moxibustion at **Ren-8** can be very helpful.

4.5.4 The Heart *luo*-Connecting Vessel System (*shou shao yin luo mai*)

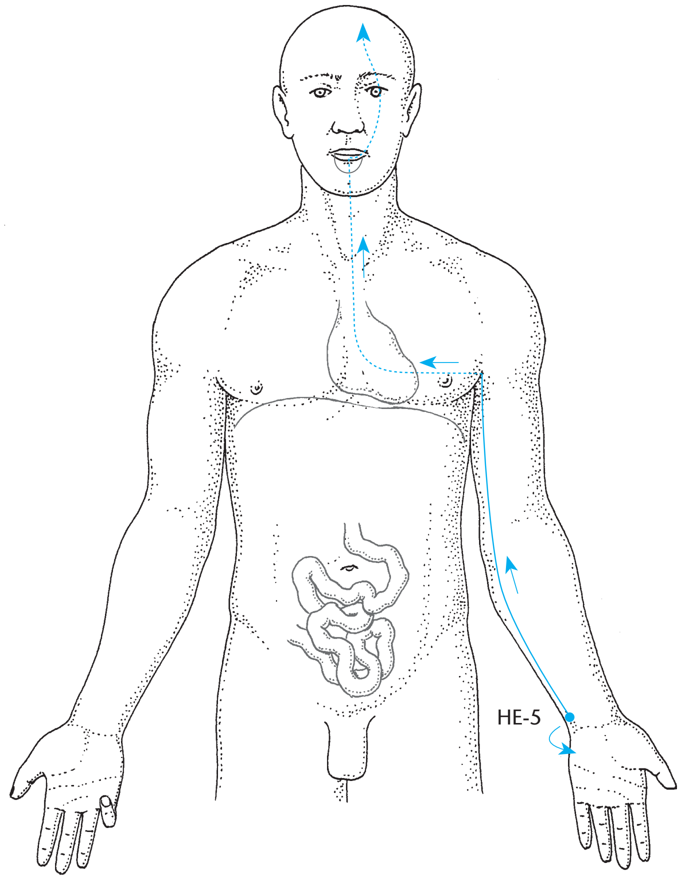

Pathway

The HE *luo*-connecting vessel separates from the Heart primary channel at its *luo*-connecting point **HE-5** (*tongli*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Small Intestine primary channel; according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** HE *luo*-connecting vessel to the *yuan*-source point → **S.I.-4** (*wangu*).
- A **longitudinal** division follows the pathway of the Heart primary channel along the anteromedial aspect of the arm, ascends to the axilla, reaches the Heart, ascends to the root of the tongue and terminates in the eye while further branches penetrate the brain.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*)**: Distension and fullness of the thorax and diaphragm
- **Deficiency (*xu*)**: Aphasia, disorders of the vocal cords

4.5.5 Cutaneous Region (*shao yin pi bu*)

See description and figures → 1.6.

4.5.6 Points of the Heart Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): HE-7** (*shenmen*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): HE-5** (*tongli*) 🟦⬛
- ***Xi*-cleft point (→ 8.1.3): HE-6 (*yinxi*)** 🟦⬛
- **Associated Back-*shu* point (→ 8.1.4): BL-15** (*xinshu*) 🟦⬛
- **Associated Front-*mu* point (→ 8.1.5): Ren-14** (*juque*) 🟦⬛
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **HE-9** (*shaochong*) 🟦⬛
  ying-spring point (Fire), *ben* point (Five Phases): **HE-8** (*shaofu*) 🟦⬛
  shu-stream point (Earth), sedation point: **HE-7** (*shenmen*) 🟦🟦
  jing-river point (Metal): **HE-4** (*lingdao*)
  he-sea point (Water): **HE-3** (*shaohai*) 🟦⬛
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10): –**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): HE-5** (*tongli*) 🟦⬛
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points: –**

Points according to region

- **Local point (→ 8.2.1):** elbow – **HE-3** (*shaohai*) 🟦⬛
- **Distal points (→ 8.2.1):** tongue – **HE-5** (*tongli*) 🟦⬛; Heart – **HE-7** (*shenmen*) 🟦🟦, **HE-5** (*tongli*) 🟦⬛, **HE-3** (*shaohai*) 🟦⬛

Specific points according to the channel pathway (in numeric order)

- **HE-3** (*shaohai*) 🟦⬛: *he*-sea point (Water); local point for the elbow (→ 8.2.1); distal point for the Heart (→ 8.2.1)
- **HE-4** (*lingdao*): *jing*-river point (Metal)
- **HE-5** (*tongli*) 🟦⬛: *luo*-connecting point (→ 8.1.2); Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the tongue and Heart (→ 8.2.1)
- **HE-6** (*yinxi*) 🟦⬛: *xi*-cleft point (→ 8.1.3)
- **HE-7** (*shenmen*) 🟦🟦: *yuan*-source point (→ 8.1.1); distal point for the Heart (→ 8.2.1); *shu*-stream point (Earth), sedation point
- **HE-8** (*shaofu*) 🟦⬛: *ying*-spring point (Fire); *ben* point (Five Phases)
- **HE-9** (*shaochong*) 🟦⬛: *jing*-well point (Wood), tonification point

Summit Spring JIQUAN HE-1

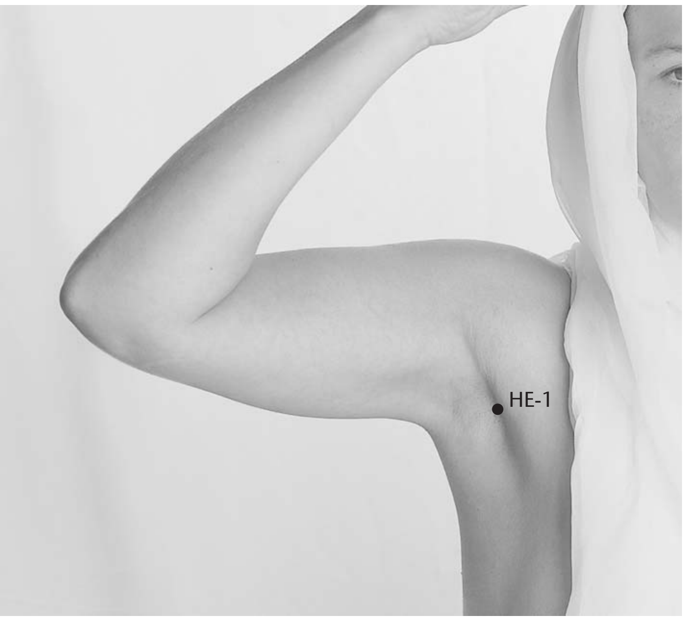

Location

With the arm abducted, in the centre of the axilla, medial to the axillary artery.

How to find

With the arm abducted, palpate along the thoracic wall in a superior direction towards the centre of the axilla until you reach the highest point.

Needling

Palpate for the artery before needling! 0.5–1 cun vertically in the direction of → **G.B.-21** (border of the trapezius, middle of the shoulder). Caution: Axillary plexus and artery. Needling in a medial direction may puncture the lung!

Actions/Indications

- Unbinds the chest
- Opens the channel
- Nourishes the Heart Yin

Special features

Entry point. Important point in Qigong practice: During practice (and in everyday life), the axillary region should always be slightly open to allow for the free circulation of Qi and Blood to and from the upper extremity.

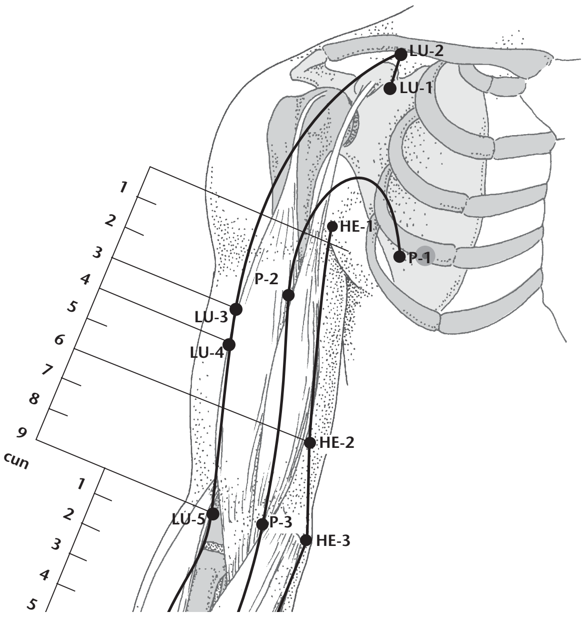

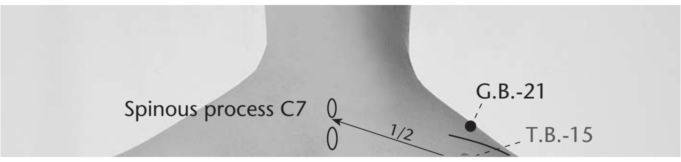

HE-2 *Green Spirit* QINGLING

Location

3 cun proximal to the cubital crease, on the medial border of the biceps brachii muscle.

How to find

With the elbow flexed, measure 3 cun from the ulnar end of the cubital crease (→ **HE-3**) towards the axilla. **HE-2** is located in a groove on the medial border of the biceps brachii muscle. Ask the patient to flex this muscle for easier location.

Needling

Vertically or obliquely 0.5–1 cun. Caution: Brachial artery! According to classic texts, needling is contraindicated, only moxibustion is permitted.

Actions/Indications

- Opens the channel

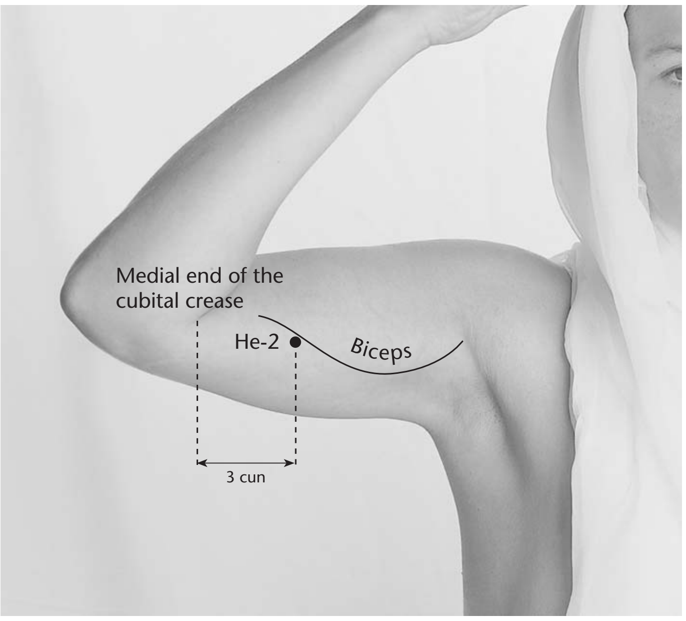
Medial end of the cubital crease
He-2
Biceps
3 cun

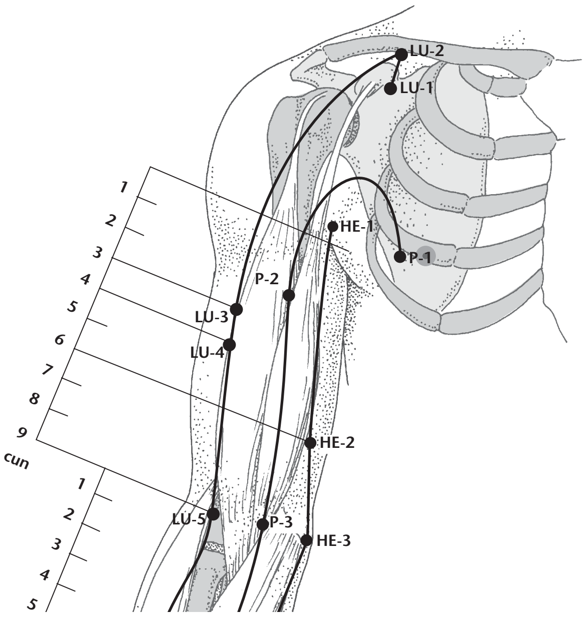
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3
4
5

Lesser Sea SHAOHAI **HE-3**

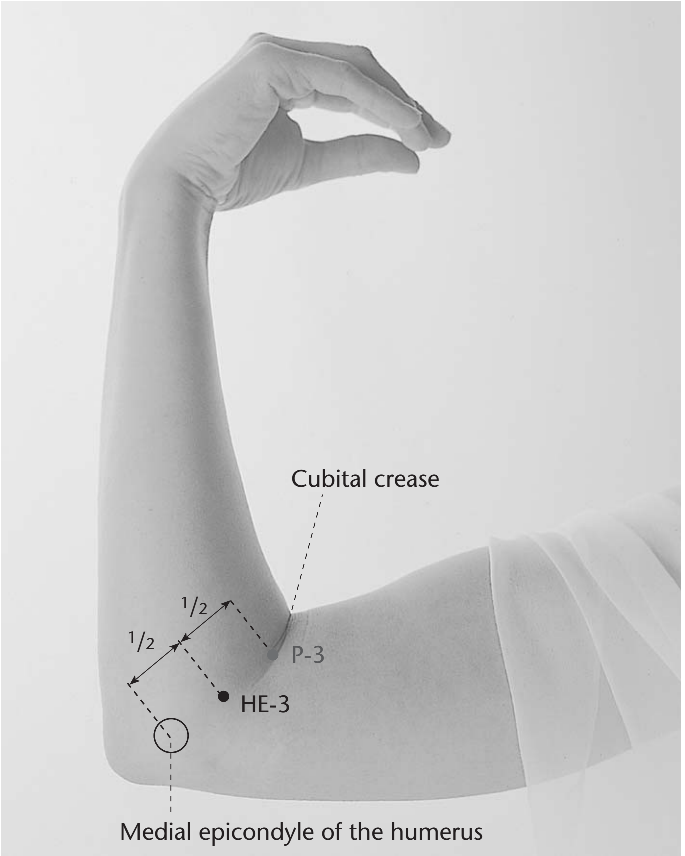
Cubital crease
1/2
1/2
P-3
HE-3
Medial epicondyle of the humerus

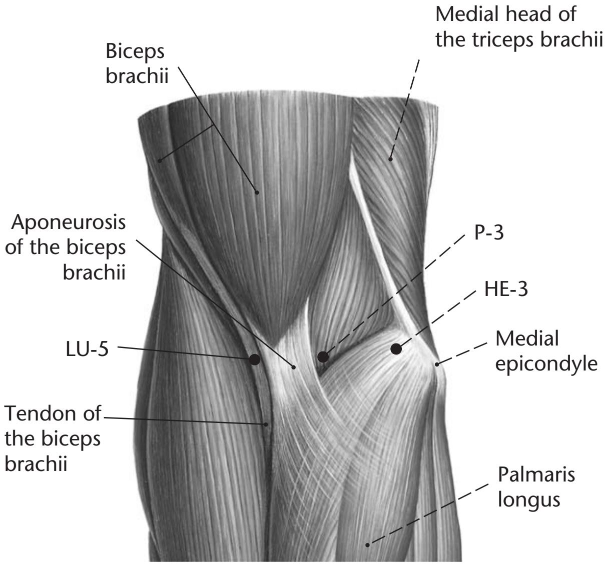
Biceps brachii
Medial head of the triceps brachii
Aponeurosis of the biceps brachii
P-3
HE-3
LU-5
Medial epicondyle
Tendon of the biceps brachii
Palmaris longus

Right arm anterior aspect

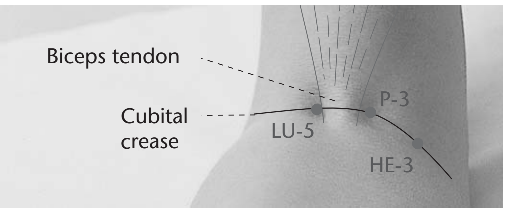
Biceps tendon
Cubital crease
P-3
LU-5
HE-3

Right arm

Location

With the elbow flexed, this point is in a depression between the ulnar end of the transverse cubital crease and the medial epicondyle of the humerus.

How to find

This point should be located with the elbow slightly flexed and the forearm in a supinated position. Find the ulnar end of the transverse cubital crease and palpate towards the medial epicondyle of the humerus (➞ 3.2.2). **HE-3** is located in a depression approximately midway between ➞ **P-3** and the medial epicondyle. **Or**: With the elbow fully flexed, **HE-3** is located directly at the medial end of the transverse crease.

Located on the same level are ➞ **P-3** (ulnar to the tendon of the biceps muscle), ➞ **LU-5** (radial to the tendon of the biceps muscle) and ➞ **L.I.-11** (at the radial end of the crease).

Needling

0.5–1.5 cun obliquely in a distal or proximal direction or vertically towards ➞ **L.I.-11**.

Actions/Indications

- Transforms Phlegm, clears Heat (from the heart), calms the *shen*
- Opens the channel and *luo*-vessels

Special features

He-sea point, Water point. Important local point, important calming point.

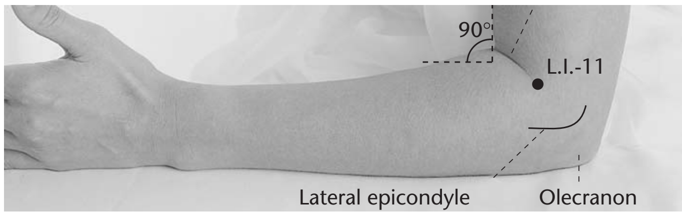
90°
L.I.-11
Lateral epicondyle
Olecranon

HE-4 *Spirit Path* LINGDAO

Location

1.5 cun proximal to the transverse wrist crease, on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

The flexor carpi radialis muscle has a distinct tendon which can be easily palpated ulnar and proximal to the wrist crease. It attaches to the pisiform bone (→ 3.3.3). **HE-4** is located directly on the radial side of the tendon.

→ **HE-5**, → **HE-6** and → **HE-7** are located at 0.5 cun intervals distal to **HE-4**. → **LU-7** is located approximately 1.5 cun proximal to the wrist crease on the radial aspect of the forearm.

Needling

0.3–1 cun vertically or obliquely in a proximal or distal direction.

Actions/Indications

- Opens the channel
- Nourishes the Heart, calms the *shen*
- Strengthens the voice

Special features

Jing-river point, Metal point.

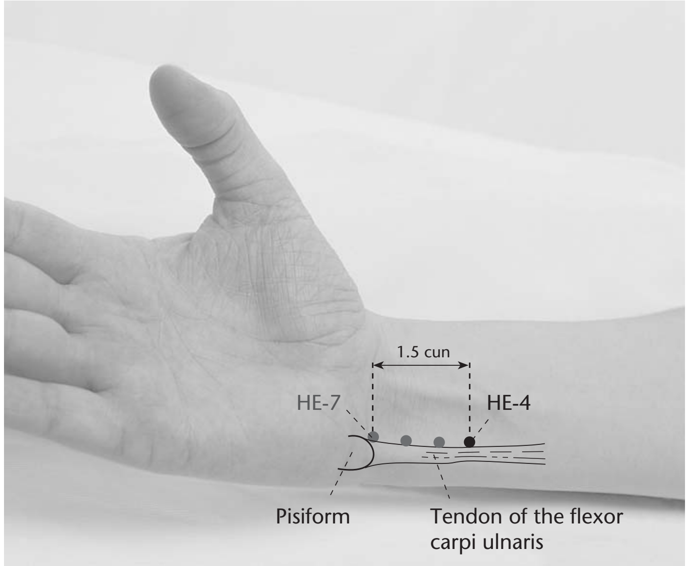
1.5 cun
HE-7
HE-4
Pisiform
Tendon of the flexor carpi ulnaris

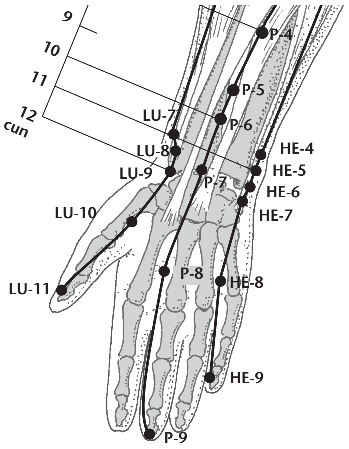
9
10
11
12 cun
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

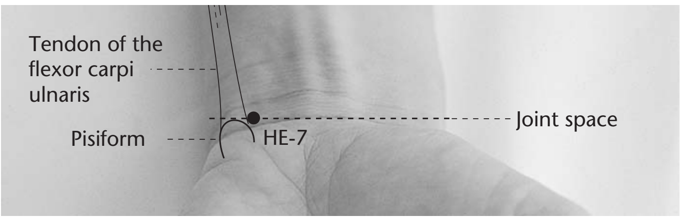
Tendon of the flexor carpi ulnaris
Pisiform
HE-7
Joint space

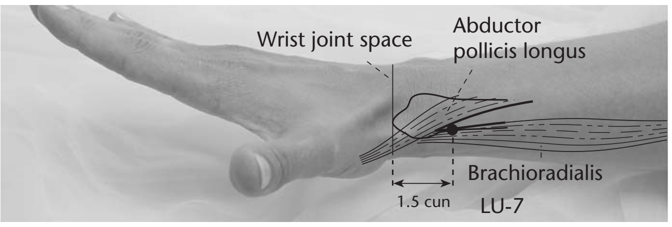
Wrist joint space
Abductor pollicis longus
Brachioradialis
1.5 cun
LU-7

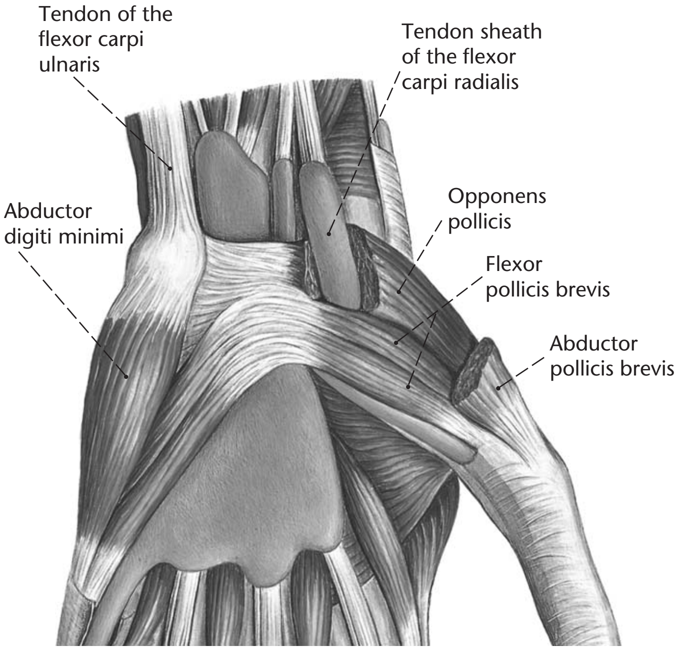
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Opponens pollicis
Abductor digiti minimi
Flexor pollicis brevis
Abductor pollicis brevis

Penetrating the Interior TONGLI HE-5

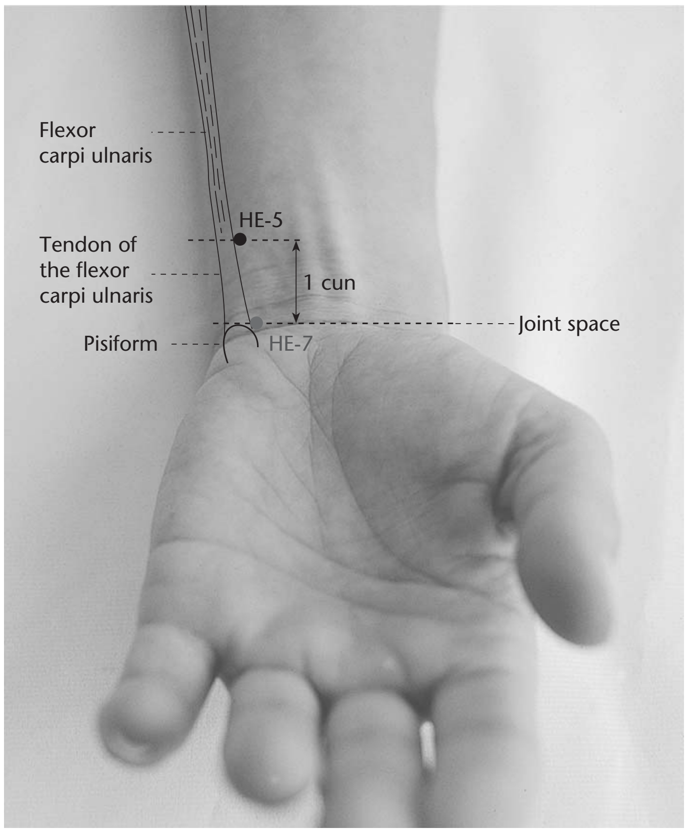
Flexor carpi ulnaris
HE-5
Tendon of the flexor carpi ulnaris
1 cun
Joint space
Pisiform
HE-7

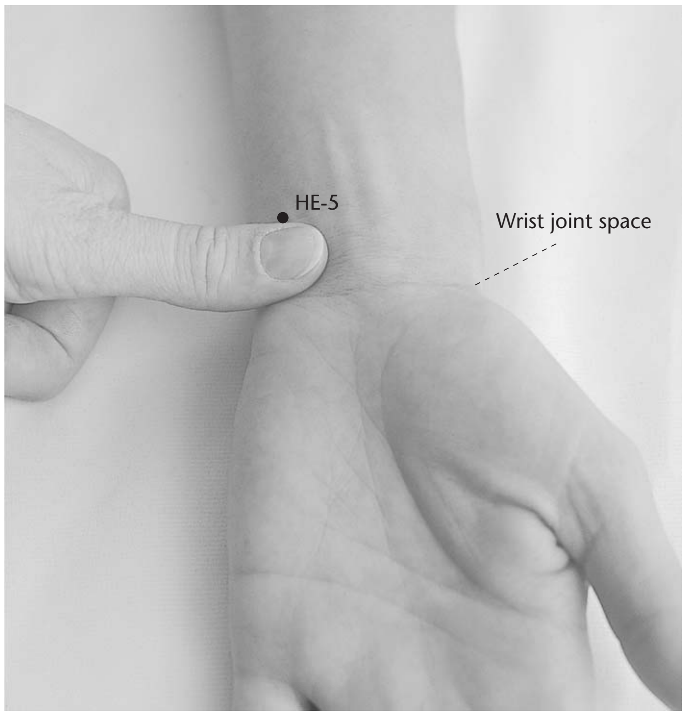
HE-5
Wrist joint space

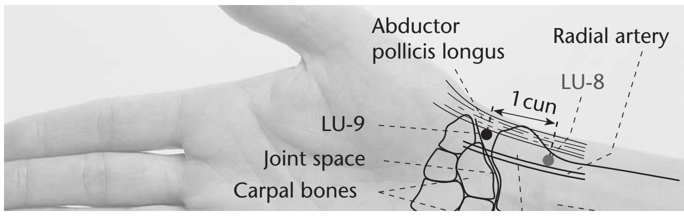
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

Location

1 cun proximal to the ventral wrist joint space (‘most distal transverse wrist crease’), on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

Position the forearm in a relaxed fashion. The location of the ventral wrist crease is variable; therefore, the proximal border of the pisiform bone (→ 3.3.2) provides a more reliable landmark. → **HE-7** is located on the wrist joint space, radial to the distinct tendon of the flexor carpi ulnaris muscle. Locate **HE-5** 1 cun proximal to → **HE-7**, also on the radial side of the tendon.
Located along a line on the radial aspect of the tendon are → **HE-6** (0.5 cun proximal to → **HE-7**) and → **HE-4** (1.5 cun proximal to **HE-7**). → **LU-8** is also located 1 cun proximal to the ventral wrist joint space, but on the radial aspect of the forearm.

Needling

Vertically 0.3–0.5 cun or obliquely 0.5–1 cun in a proximal or distal direction. Caution: Needling may be painful.

Actions/Indications

- Regulates and strengthens the Heart Qi (main point!) and the Heart rhythm
- Calms the *shen*
- Benefits the tongue
- Benefits the Bladder (via the *taiyang* channel)
- Opens the channel and *luo*-connecting vessels

Special features

Luo-connecting point, Ma Dan Yang Heavenly Star point. Main point for regulating and strengthening the Heart Qi.

HE-6 *Yin Cleft* YINXI

Location

0.5 cun proximal to the transverse wrist crease, on the radial side of the tendon of the flexor carpi ulnaris muscle.

How to find

The flexor carpi ulnaris muscle has a distinct tendon on the ventral aspect of the forearm that can be easily palpated ulnar and proximal to the wrist. **HE-6** is located on the radial aspect of the tendon, 0.5 cun proximal to → **HE-7** (located on the joint space).

→ **HE-5** and → **HE-4** are located proximally on the same line at 0.5 cun intervals.

Needling

0.3–1 cun vertically or obliquely in a proximal or distal direction. Caution: Ulnar artery/nerve. Painful point!

Actions/Indications

- Tonifies and regulates the Heart Yin and Blood, clears empty Heat in the Heart, calms the *shen*

Special features

Xi-cleft point

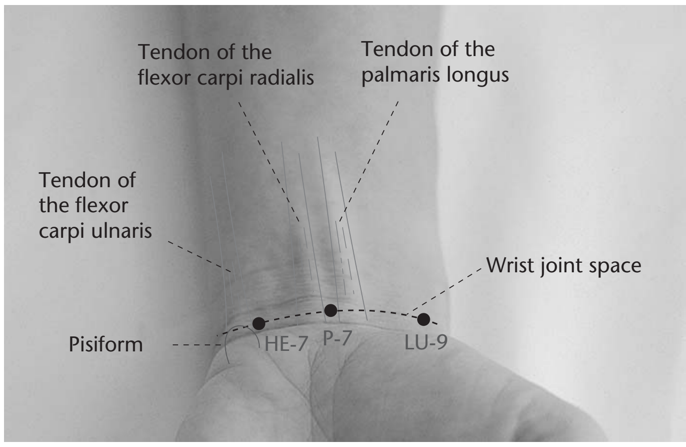
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

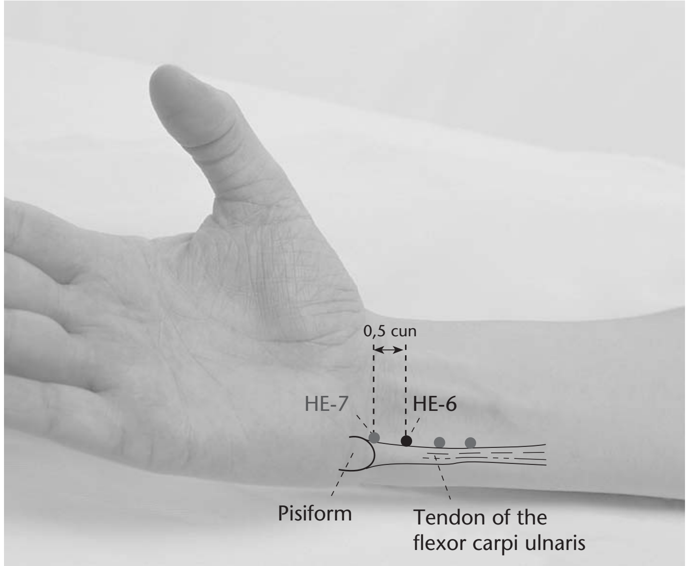
0,5 cun
HE-7
HE-6
Pisiform
Tendon of the flexor carpi ulnaris

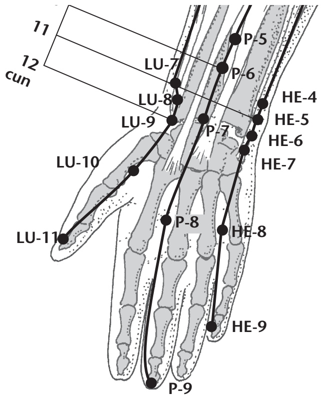
11
12
cun
P-5
LU-7
P-6
LU-8
HE-4
LU-9
HE-5
P-7
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

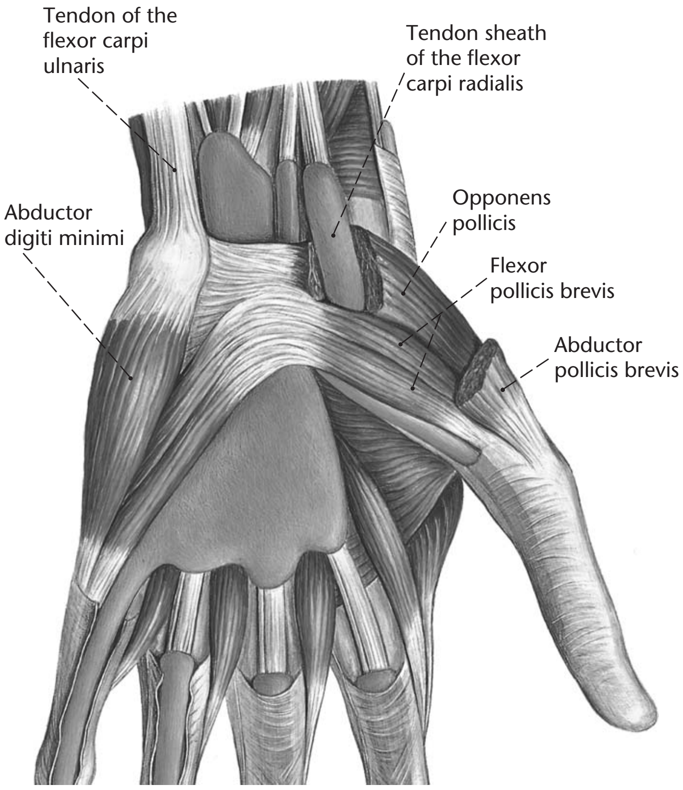
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Opponens pollicis
Abductor digiti minimi
Flexor pollicis brevis
Abductor pollicis brevis

Spirit Gate SHENMEN HE-7

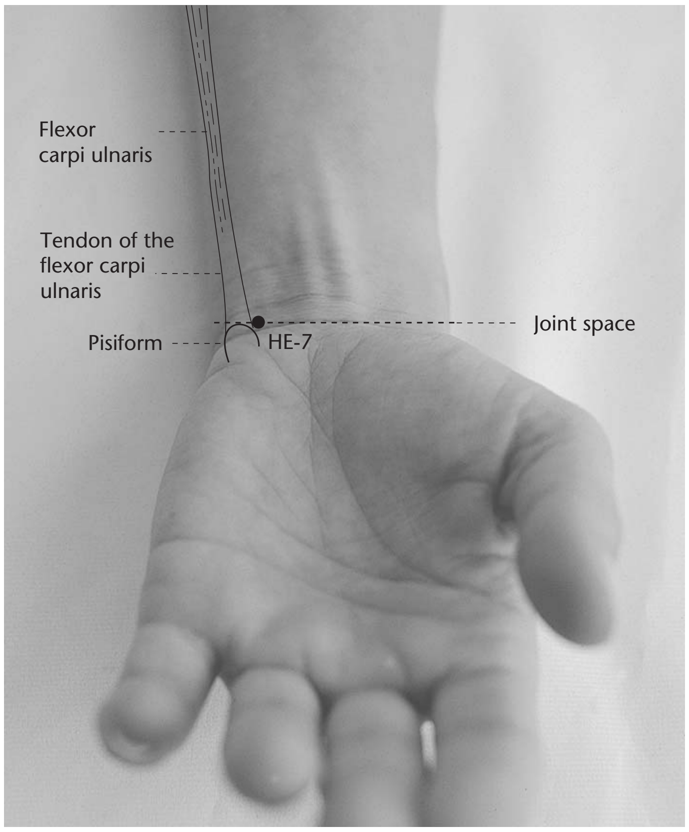
Flexor carpi ulnaris
Tendon of the flexor carpi ulnaris
Pisiform
HE-7
Joint space

Location

On the ventral aspect of the wrist joint space ('most distal wrist crease'), radial to the insertion of the tendon of the flexor carpi ulnaris muscle.

How to find

Position the forearm in a relaxed supinated position. As the location of the wrist crease can vary, the proximal border of the pisiform bone (→ 3.3.3) should be used for reference instead. **HE-7** is located radial to the insertion of the tendon of the flexor carpi ulnaris muscle, in a depression proximal to the pisiform bone.

→ **HE-6**, → **HE-5** and → **HE-4** are all located on a line on the radial aspect of the tendon at 0.5 cun intervals. Also located on the wrist crease are → **P-7** (between the tendons of the palmaris longus and the flexor carpi radialis muscles) and → **LU-9** (on the radial aspect, lateral to the radial artery).

Needling

0.3–0.5 cun vertically or transversely deep to the tendon (parallel to the wrist crease 0.5–0.8 cun). Caution: Ulnar artery/nerve. Painful point!

Actions/Indications

- Regulates and strengthens the Heart (Blood and Yin), calms the *shen*
- Clears Heat from the Heart channel
- Moves Qi locally for disorders of the wrist

Special features

Yuan-source point, *shu*-stream point, Earth point, sedation point. Major point for calming the *shen*.

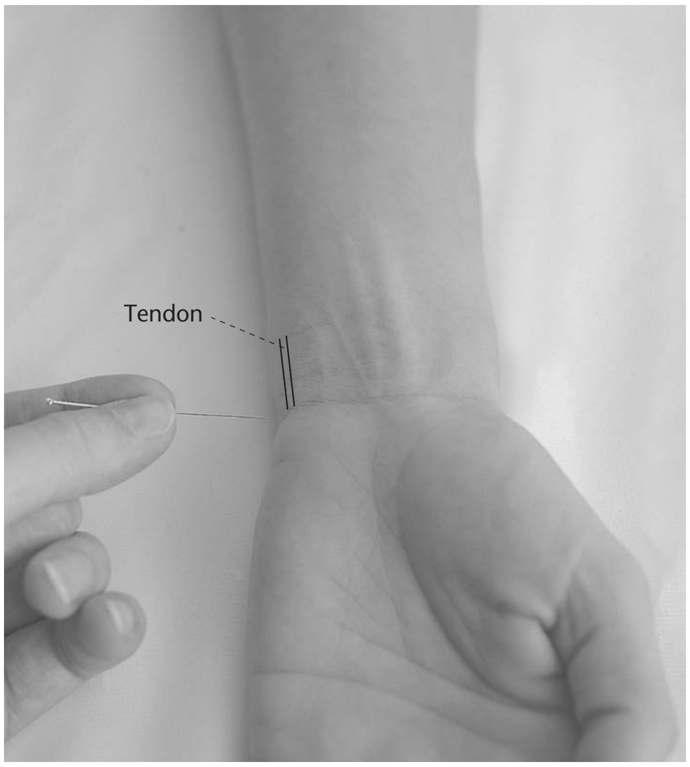

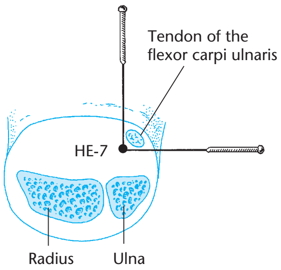
Tendon of the flexor carpi ulnaris
HE-7
Radius
Ulna

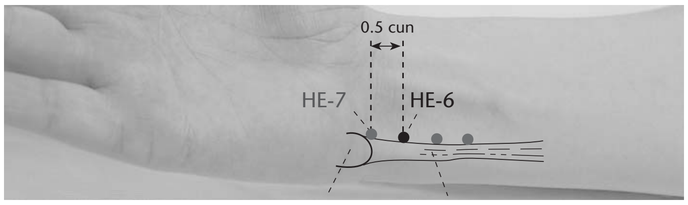
0.5 cun
HE-7
HE-6

HE-8 *Lesser Palace* SHAOFU

Location

On the palm, between the 4th and 5th metacarpal bones.

How to find

When making a fist, **HE-8** is generally located below the tip of the little finger, in the depression between the 4th and 5th metacarpal bones. **HE-8** often lies between the two transverse palmar creases. It is often sensitive to pressure.

→ **P-8** is also located on the palm, generally below the tip of the middle finger when making a fist (between the 2nd and 3rd metacarpal bones).

Needling

Vertically 0.5–1 cun. Caution: Needling this point may be painful. Provide a competitive stimulus by pressing the palm with the fingernail close to the point before inserting the needle. Ask the patient to breathe out during needle insertion.

Actions/Indications

- Clears Heat in the Heart and Small Intestine
- Regulates the Heart Qi for Liver Qi stagnation
- Calms the *shen* and strengthens the Heart Qi
- Opens the channel, especially for contractures of the fingers

Special features

Ying-spring point, Fire point, *ben* point (Five Phases).

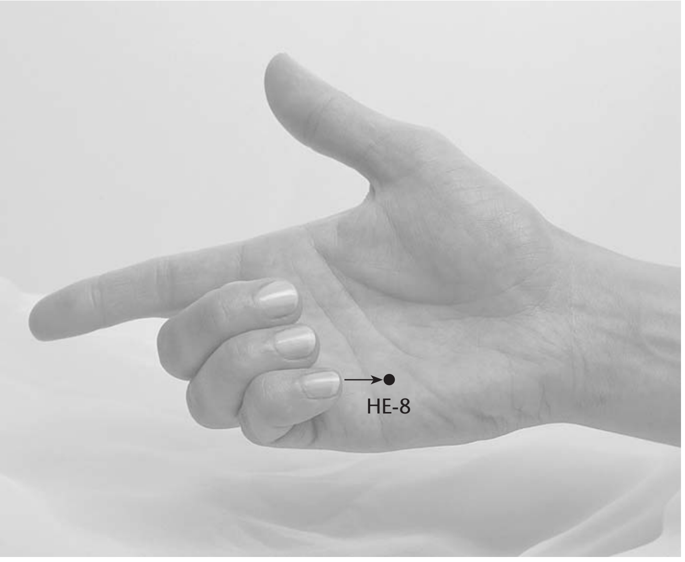

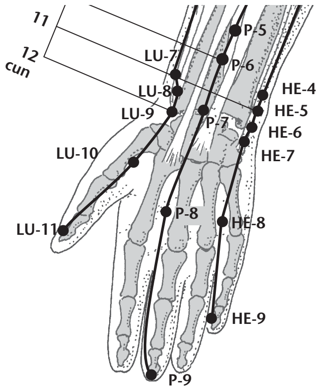

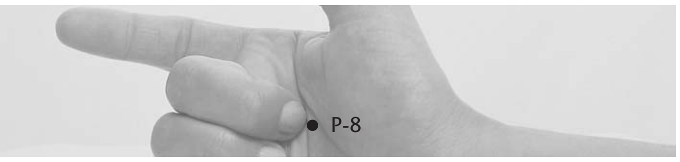

Lesser Rushing SHAOCHONG HE-9

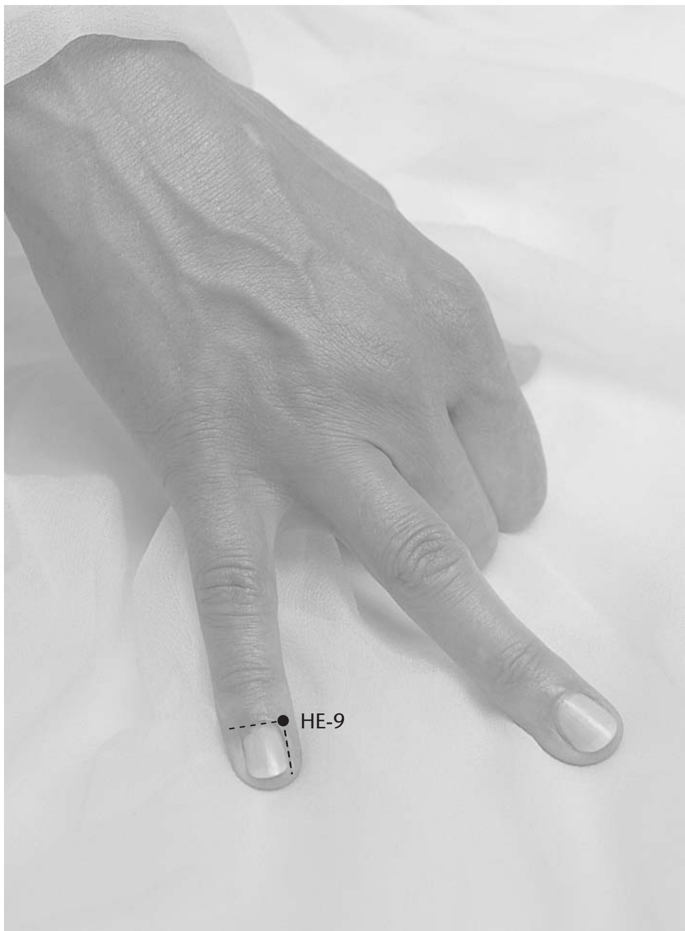

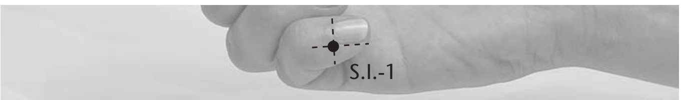

Location

On the little finger, 0.1 cun lateral to the radial corner of the nail.

How to find

This point is located at the junction of two tangents that border the radial and proximal aspect of the nail, approximately 0.1 cun from the actual corner of the nail. → **S.I.-1** is located on the opposite (ulnar) aspect of the nail.

Needling

0.2–0.3 cun vertically or obliquely in a proximal direction. For acute disorders (pain, inflammation), prick to bleed with a needle, lancet or three-edged needle. Caution: Painful point!

Actions/Indications

- Revives consciousness
- Clears Heat from the opposite end of the channel
- Regulates Qi in the chest and clears Heat in the Heart
- Opens the channel

Special features

Jing-well point, Wood point, tonification point, exit point.

4.6 The Small Intestine Channel System – Hand *taiyang* (*shou tai yang jing luo*)

4.6.1 The Small Intestine Primary Channel (*shou tai yang jing*)

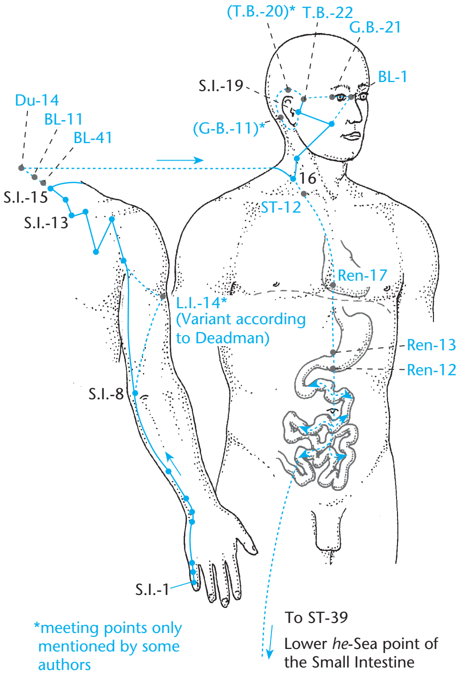
(T.B.-20)*
T.B.-22
G.B.-21
BL-1
S.I.-19
(G-B.-11)*
Du-14
BL-11
BL-41
S.I.-15
S.I.-13
16
ST-12
Ren-17
L.I.-14*
(Variant according to Deadman)
Ren-13
Ren-12
S.I.-8
S.I.-1

*meeting points only mentioned by some authors

To ST-39
Lower *he*-Sea point of the Small Intestine

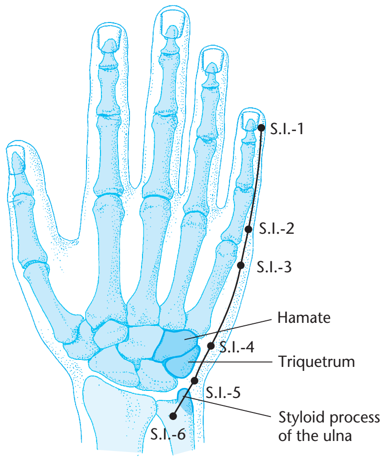
S.I.-1
S.I.-2
S.I.-3
Hamate
S.I.-4
Triquetrum
S.I.-5
Styloid process of the ulna
S.I.-6

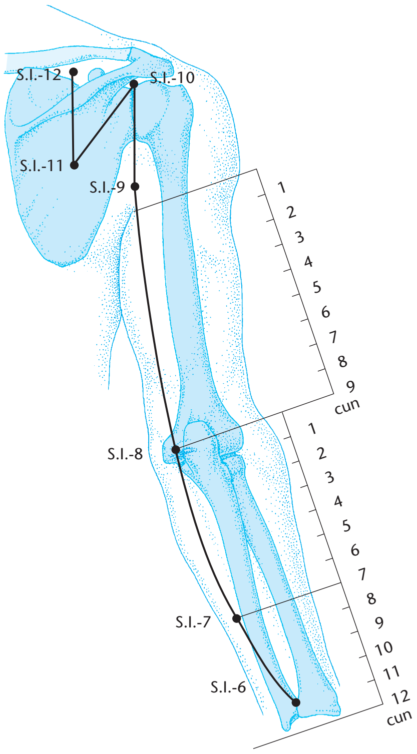
S.I.-12
S.I.-10
S.I.-11
S.I.-9
1
2
3
4
5
6
7
8
9
cun
S.I.-8
1
2
3
4
5
6
7
8
9
10
11
12
cun
S.I.-7
S.I.-6

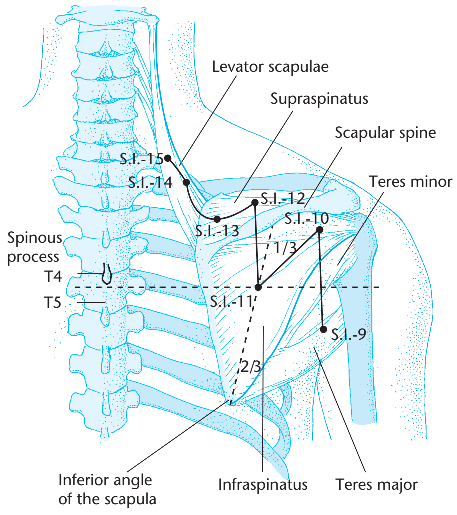
Levator scapulae
Supraspinatus
Scapular spine
Teres minor
S.I.-15
S.I.-14
S.I.-12
S.I.-13
S.I.-10
1/3
Spinous process
T4
T5
S.I.-11
S.I.-9
2/3
Inferior angle of the scapula
Infraspinatus
Teres major

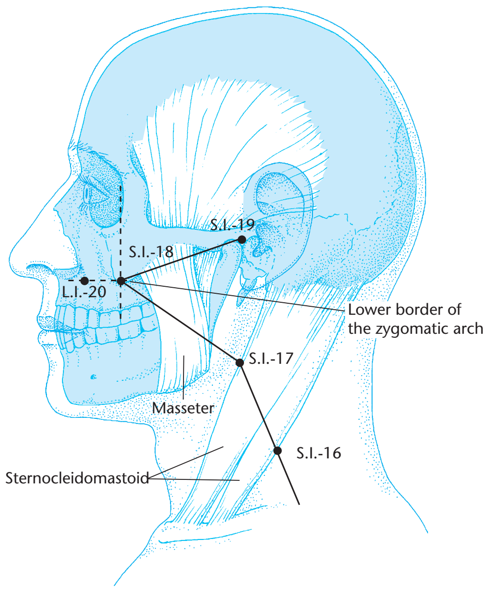
S.I.-19
S.I.-18
L.I.-20
Lower border of the zygomatic arch
S.I.-17
Masseter
S.I.-16
Sternocleidomastoid

Pathway

The **external** pathway of the Small Intestine primary channel begins on the little finger, on the ulnar corner of the nail, at **S.I.-1** (*shaoze*). This point is reached by a branch that separates from the Heart primary channel at → **HE-9** (*shaochong*) (hand Yin–Yang connection of the second great circuit).

The channel runs along the ulnar border of the hand, passes the wrist and continues along the ulna to the medial aspects of the elbow.

- passes between the olecranon and the medial epicondyle, ascends the posterior aspect of the humerus to the shoulder joint and, according to some authors, meets → **L.I.-14**
- zig-zags across the scapula, passes → **BL-41** (*fufen*) and → **BL-11** (*dazhu*) and meets the extraordinary vessel *du mai* and the other Yang primary channels at → **Du-14** (*dazhui*), below the spinous process of C7
- continues anteriorly to the supraclavicular fossa to → **ST-12** (*quepen*).

An **internal** branch originates at → **ST-12** and runs inferiorly to the Heart (*xin*), the paired *zang*-Organ, continues to → **Ren-17** and descends along the oesophagus, penetrates the diaphragm and continues to the Stomach (*wei*). Here it crosses → **Ren-13** (*shangwan*) and → **Ren-12** (*zhongwan*), further descends and spreads in its pertaining *fu*-Organ, the Small Intestine (*xiaochang*).

According to the *Neijing Ling Shu*, an **internal** branch courses from the Small Intestine (according to some authors from the Small Intestine primary channel) to → **ST-39** (*xiajuxu*), the lower *he*-sea point of the Small Intestine.

The **external** pathway of the Small Intestine primary channel continues from the supraclavicular fossa along the neck and ascends to the cheek to **S.I.-18** (*quanliao*).

From **S.I.-18**

- **a branch** courses to → **BL-1** (*jingming*) at the inner canthus of the eye, where it connects with the Bladder primary channel (six channel theory, Yang axes: *taiyang*)
- **a further branch** travels to the outer canthus of the eye, intersects with → **G.B.-1** (*tongziliao*), according to some authors also with → **T.B.-20** and → **G.B.-11**, continues to → **T.B-22** (*erheliao*) anterior to the ear and terminates anterior to the ear deep to **S.I.-19** (*tinggong*)

Clinical importance (➡ 1.2)

Exterior (*biao*) signs and symptoms: Mouth and tongue ulcers, cheek pain, neck pain, occipital stiffness, pain on the lateral aspect of the shoulder and the upper arm

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain and distension of the upper abdomen, radiating to the genital region, diarrhoea, epigastric pain with constipation

Connections of the Small Intestine primary channel

Connections with other channels

Heart primary channel (*shou shao yin jing*)

Connection: Hand Yin–Yang connection of the second great circuit
Location: **H-9** → **S.I.-1** (on the hand)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Bladder primary channel (*zu tai yang jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yang* (Yang axes of the second great circuit)
Location: **S.I.-18** → **BL-1** (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Heart (*xin*), Stomach (*wei*), **Small Intestine (*xiaochang*)**

4.6.2 The Small Intestine Divergent Channel (*shou tai yang jing bie*)

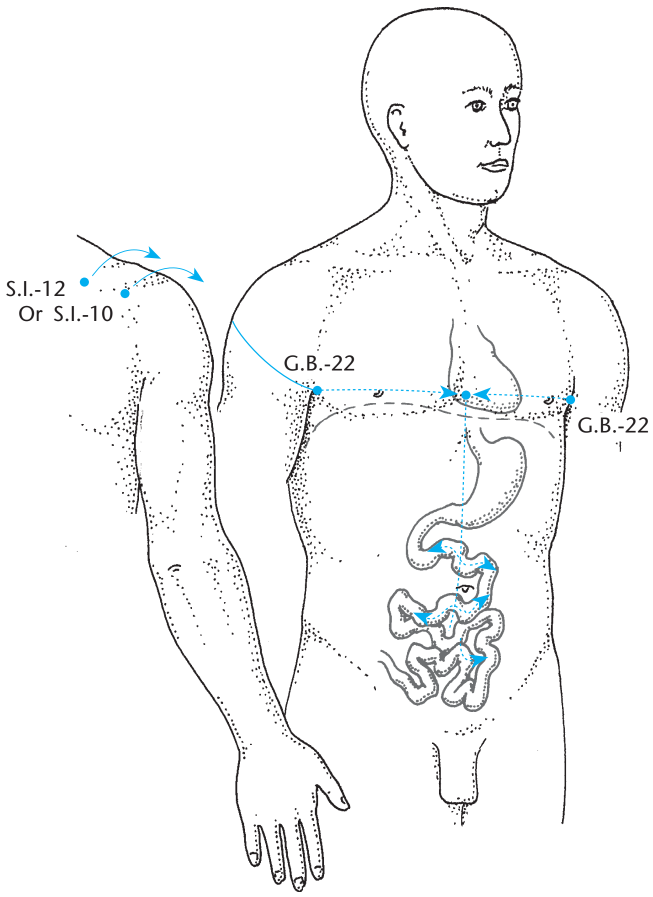

Pathway

The Small Intestine divergent channel separates from the Small Intestine primary channel at the shoulder, depending on the school of thought in the area of → **S.I.-12** (*bingfeng*) or in the area of → **S.I.-10** (*naoshu*)

- descends towards the axilla and continues to → **G.B.-22** (*yuanye*)
- penetrates the thorax and spreads in the Heart (*xin*) where, according to most authors, it forms the fourth confluence
- penetrates the diaphragm to connect with its pertaining *fu*-Organ, the Small Intestine

Ch04.6-F10028.qxd 2/22/08 9:47 PM Page 219

Clinical importance

- Strengthens the relationship between the Heart and the Small Intestine (*zangfu*-Organ system). It provides an explanation for the pattern of Heart Fire transferring to the Small Intestine and the Bladder.

4.6.3 The Small Intestine Sinew Channel (*shou tai yang yin jing jin*)

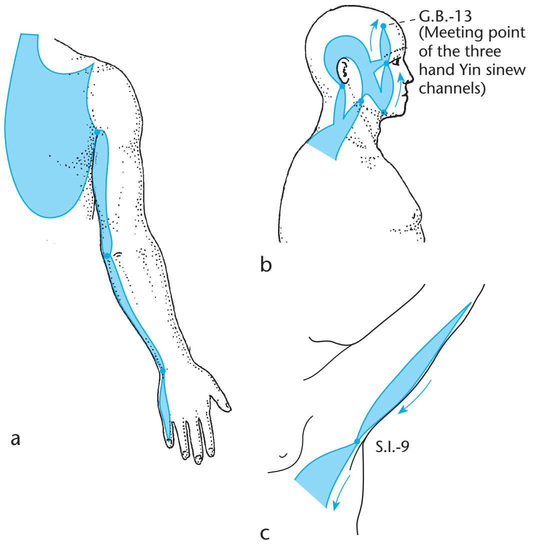

Pathway

The Small Intestine sinew channel begins on the ulnar aspect of the little finger at ➞ **S.I.-1** (*shaoze*)

- ➡ continues along the ulnar aspect of the little finger and the ulnar border of the hand
- ➡ runs to **S.I.-5** (*yanggu*) distal to the styloid process of the ulna and binds (*jie*) at the wrist
- ➡ travels along the ulnar aspect of the forearm to the elbow, continues to the axilla, binding at **S.I.-8** (*xiaohai*) and the posterior aspect of the axilla.

From here a branch courses into the axilla where it binds (*jie*). Another branch spreads across the scapula and ascends anterior to the Bladder and Gall Bladder sinew channels and posterior to the Stomach sinew channel to the lateral aspect of the neck, where it meets the Triple Burner sinew channel.

At the neck it divides into **two branches:**

- ➡ the **anterior branch** runs to the mandibular angle where it binds (*jie*), continues along the mandible to the ear and the zygomatic bone to bind (*jie*) at the lateral aspect of the orbit
- ➡ the **posterior branch** binds (*jie*) at the mastoid process (from here a small branch runs to the ear). The branch curves around the ear, descends to the cheek and binds (*jie*) at the insertion of the masseter muscle on the mandible, continues to the lateral aspect of the orbit and reaches ➞ **G.B.-13** (*benshen*) on the frontoparietal aspect of the head, where it meets the other hand Yang sinew channels.

Clinical importance

Pathology: Distending sensations, stiffness and pain of the little finger, impaired range of motion of the upper extremity and shoulder. Pain, stiffness and impaired range of motion of the cervical spine and occiput. Tinnitus and ear pain radiating to the mandible.

Indication: Mainly for *bi*-syndromes (painful obstruction syndrome) along the pathway of the S.I. channel. The area covered by the Small Intestine sinew channel is larger than that covered by the Small Intestine primary channel. This explains why Small Intestine primary channel points can also be used for disorders and diseases of the cervical and thoracic spine and the whole face (for example frontal and lateral headaches).

4.6.4 The Small Intestine *luo*-Connecting Vessel System (*shou tai yang luo mai*)

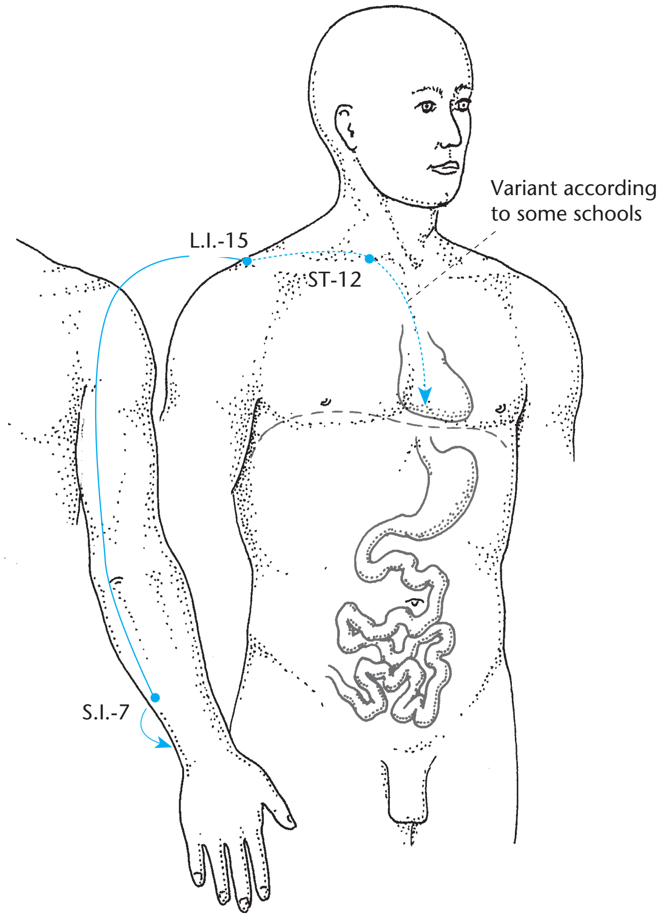

Pathway

The S.I. *luo*-connecting vessel separates from the Small Intestine primary channel at its *luo*-connecting point **S.I.-7** (*zhizheng*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Heart primary channel, according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** S.I. *luo*-connecting vessel to the *yuan*-source point → **HE-7** (*shenmen*).
- A **longitudinal** division runs to the medial condyle of the humerus, ascends to the shoulder to → **L.I.-15** (*jianyu*). According to some schools of thought the vessel spreads and terminates here, while according to other schools it traverses the acromion, continues to → **ST-12** (*quepen*) in the supraclavicular fossa, penetrating the thorax in order to connect with the Heart.

Clinical importance (➡ 8.1.2)

Pathology

- **Excess** (***shi***): Weak and unstable joints, weakness and paralysis of the elbow and the arm
- **Deficiency** (***xu***): Warts

4.6.5 Cutaneous Region (*tai yang pi bu*)

See description and figures → 1.6.

4.6.6 Points of the Small Intestine Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): S.I.-4** (*wangu*)
- ***Luo*-connecting point (→ 8.1.2.): S.I.-7** (*zhizheng*)
- ***Xi*-cleft point (→ 8.1.3): S.I.-6** (*yanglao*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-27** (*xiaochangshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): Ren-4** (*guanyuan*) 🟦🟦
- **Five *shu*-transporting point (→ 8.1.6):**
  jing-well point (Metal): **S.I.-1** (*shaoze*)
  ying-spring point (Water): **S.I.-2** (*qiangu*)
  shu-stream point (Wood), tonification point: **S.I.-3** (*houxi*) 🟦■
  jing-river point (Fire), *ben* point: **S.I.-5** (*yanggu*)
  he-sea point (Earth), tonification point: **S.I.-8** (*xiaohai*)
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *du mai*: S.I.-3** (*houxi*) 🟦🟦
- **Lower *he*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the BL channel, the *yang wei mai, yang qiao mai*: **S.I.-10** (*naoshu*)
  - with the L.I., T.B., G.B. channels: **S.I.-12** (*bingfeng*)
  - with the T.B. channel: **S.I.-18** (*quanlao*) 🟦■
  - with the T.B., G.B.: **S.I.-19** (*tinggong*) 🟦■
  - of other channels with the S.I. channel: **L.I.-14*, Du-14, BL-41, BL-11, ST-12, Ren-17*, Ren-13, Ren-12, G.B.-1, G.B.-11*, T.B.-20*, T.B.-22, BL-1**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): S.I.-16** (*tianchuang*), **S.I.-17** (*tianrong*) 🟦■
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**

Points according to region

- **Local points (→ 8.2.1):** face and cheek – **S.I.-18** (*quanlao*) 🟦■; jaw – **S.I.-19** (*tinggong*) 🟦■; neck – **S.I.-17** (*tianrong*) 🟦■; ears – **S.I.-19** (*tinggong*) 🟦■; shoulder – **S.I.-9** (*jianzhen*), **S.I.-10** (*naoshu*), **S.I.-11** (*tianzong*), **S.I.-12** (*bingfeng*) ■■, **S.I.-13** (*quyuan*), **S.I.-14** (*jianwaishu*) 🟦■; elbow – **S.I.-8** (*xiaohai*); hand – **S.I.-4** (*wangu*), **S.I.-5** (*yanggu*); finger (numbness and pain) – **S.I.-3** (*houxi*) 🟦■; finger (stiffness) – **S.I.-3** (*houxi*) 🟦■, **S.I.-5** (*yanggu*)
- **Adjacent points (→ 8.2.1):** shoulder – **S.I.-9–S.I.-15**; finger – **S.I.-5** (*yanggu*)
- **Distal points (→ 8.2.1):** for the occiput – **S.I.-3** (*houxi*) 🟦■; for the eyes – **S.I.-6** (*yanglao*) 🟦■; for the cervical and thoracic spine – **S.I.-3** (*houxi*) 🟦■; for the thoracic spine – **S.I.-6** (*yanglao*) 🟦■; for the lumbar region – **S.I.-3** (*houxi*) 🟦■, **S.I.-6** (*yanglao*) 🟦■

Specific points according to the channel pathway (in numerical order)

- **S.I.-1** (*shaoze*): *jing*-well point (Metal)
- **S.I.-2** (*qiangu*): *ying*-spring point (Water)
- **S.I.-3** (*houxi*) 🟦■: *shu*-stream point (Wood), tonification point, opening point (→ 8.1.8) of the *du mai*, distal point for the occiput and the lumbar region (→ 8.2.1); local point for the fingers (→ 8.2.1); distal point for the cervical and thoracic spine (→ 8.2.1)
- **S.I.-4** (*wangu*): *yuan*-source point (→ 8.1.1); local point for the hand (→ 8.2.1)
- **S.I.-5** (*yanggu*): *jing*-river point (Fire), *ben* point (Five Phases); local point for the hand and fingers (stiffness) (→ 8.2.1); regional point for the fingers (→ 8.2.1)
- **S.I.-6** (*yanglao*) 🟦■: *xi*-cleft point (→ 8.1.3); distal point for the eyes, the thoracic and lumbar spine (→ 8.2.1)
- **S.I.-7** (*zhizheng*): *luo*-connecting point (→ 8.1.2)
- **S.I.-8** (*xiaohai*): *he*-sea point (Earth), sedation point; local point for the elbow (→ 8.2.1)
- **S.I.-9** *(jianzhen):* local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-9 – S.I.-15**: adjacent points for the shoulder (often sensitive to pressure) (→ 8.2.1)

---

* According to Deadman et al 1998.

Ch04.6-F10028.qxd 2/22/08 9:47 PM Page 221

- **S.I.-10** (*naoshu*): *jiaohui*-meeting point with the BL channel, the *yang wei mai, yang qiao mai* (→ 8.1.10); local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-11** (*tianzong*) 🟦■: local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-12** (*bingfeng*): *jiaohui*-meeting point with the L.I., T.B. and G.B. channels (→ 8.1.10); local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-13** (*quyuan*): local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-14** (*jianwaishu*) 🟦■: local point for the shoulder (often sensitive to pressure) (→ 8.2.1)
- **S.I.-16** (*tianchuang*): Window of Heaven point (→ 8.1.12)
- **S.I.-17** (*tianrong*) 🟦■: Window of Heaven point (→ 8.1.12); local point for the neck (→ 8.2.1)
- **S.I.-18** (*quanlao*) 🟦■: local point for the face and cheek (→ 8.2.1); *jiaohui*-meeting point with the T.B. channel (→ 8.1.10)
- **S.I.-19** (*tinggong*) 🟦■: *jiaohui*-meeting point with the T.B., G.B. channels (→ 8.1.10); local point for disorders of the ear and jaw (→ 8.2.1)

Lesser Marsh SHAOZE S.I.-1

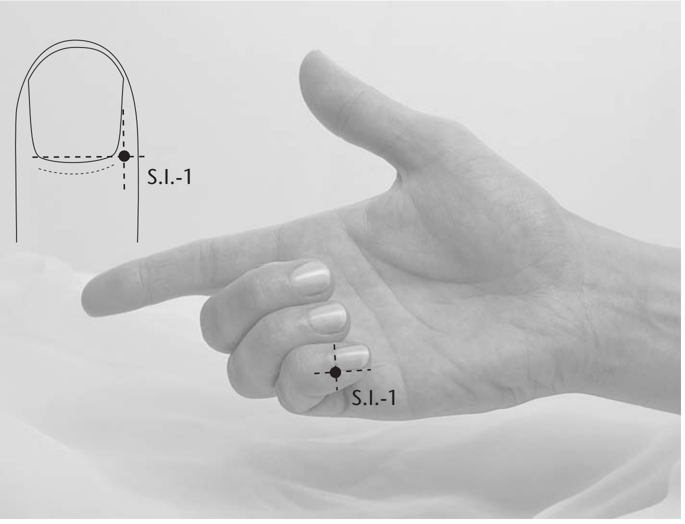

Location

On the little finger, 0.1 cun from the ulnar corner of the nail.

How to find

This point is located at the junction of two tangents drawn along the lateral and proximal borders of the nail, approximately 0.1 cun from the actual corner of the nail. → **HE-9** is located at the radial corner of the nail.

Needling

0.1–0.2 cun vertically or obliquely in a proximal direction or prick to bleed with a needle, lancet or three-edged needle. Caution: Painful point.

Actions/Indications

- Expels Wind and Heat
- Opens the orifices
- Revives consciousness
- Opens the channel
- Benefits the breasts

Special features

Jing-well point, Metal point, entry point.

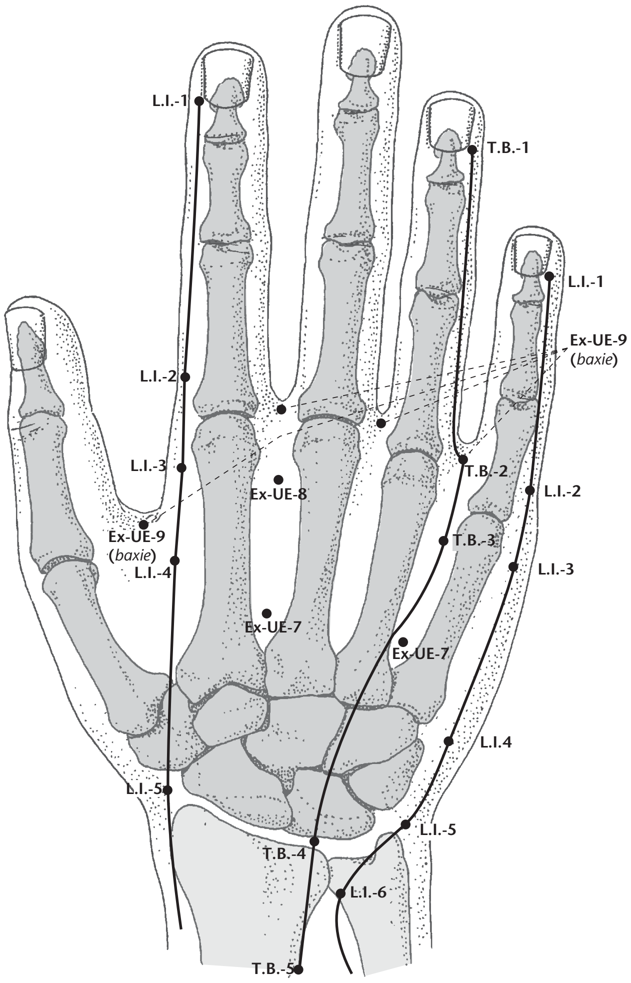
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
L.I.-2
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
L.I.-3
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

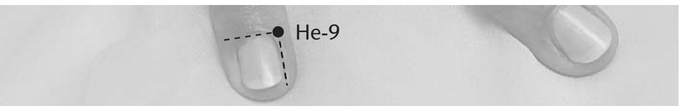

S.I.-2 *Front Valley* QIANGU

Location

On the ulnar aspect of the little finger, distal to the metacarpophalangeal joint, at the junction of the shaft and the base of the proximal phalanx.

How to find

On the ulnar aspect of the little finger, palpate along the border of the 'red and white' skin in the direction of the metacarpophalangeal joint, until you can clearly feel the junction between the shaft and the base. **S.I.-2** is located at the junction of the shaft and the base and slightly inferior (palmar) to the exterior curvature of the bone. → **L.I.-2** is located in a comparable position on the proximal phalanx of the index finger. In a comparable location on the foot are → **BL-66** (on the lateral aspect) and → **SP-2** (on the medial aspect).

Needling

0.2–0.5 cun obliquely in a proximal or distal direction and slightly towards the palm. Caution: Painful point.

Actions/Indications

- Drains Heat, expels Wind
- Disperses accumulations

Special features

Ying-spring point, Water point.

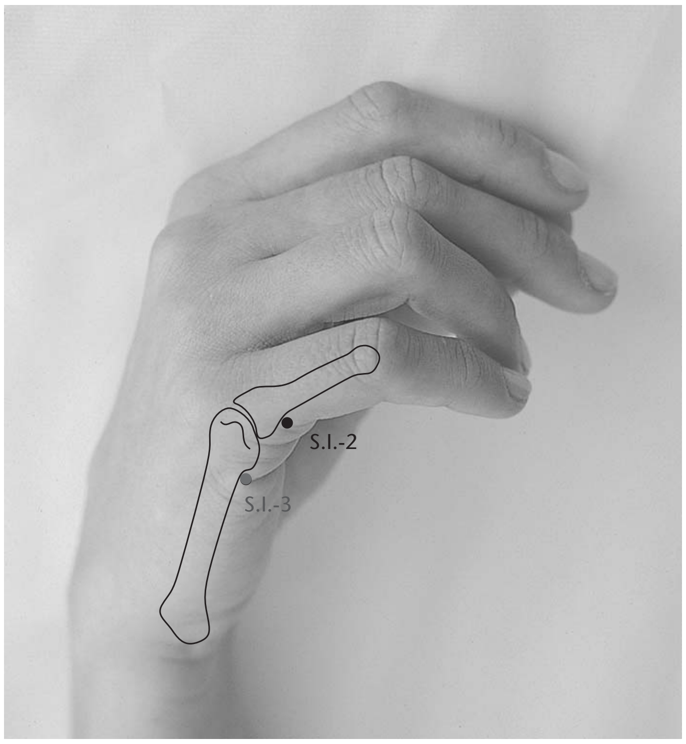

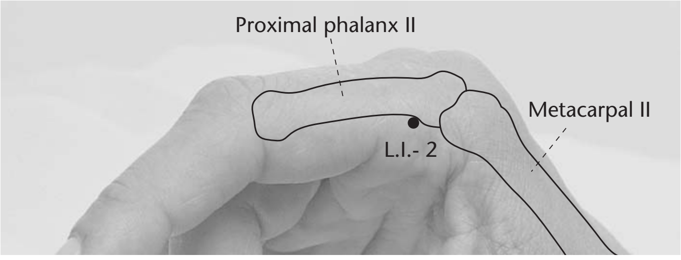
Proximal phalanx II
Metacarpal II
L.I.- 2

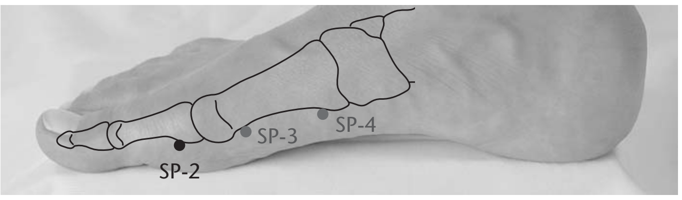

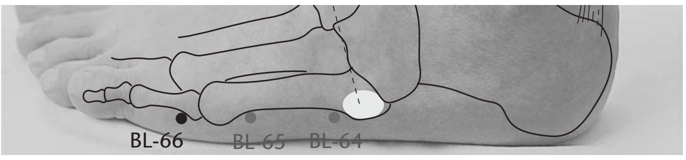
BL-66
BL-65
BL-64

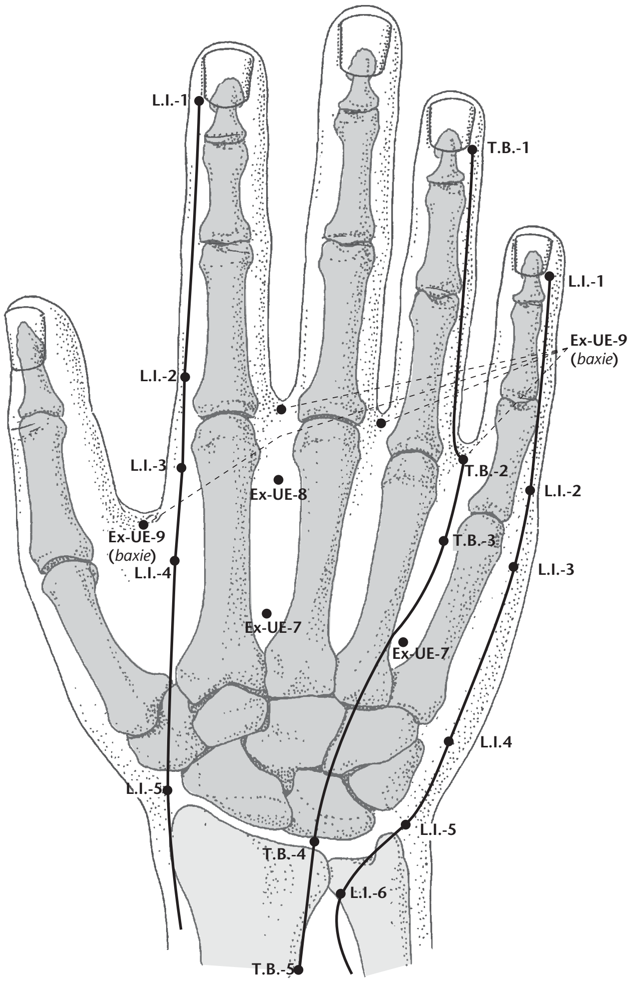
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
L.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
L.I.-3
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

Back Stream HOUXI S.I.-3

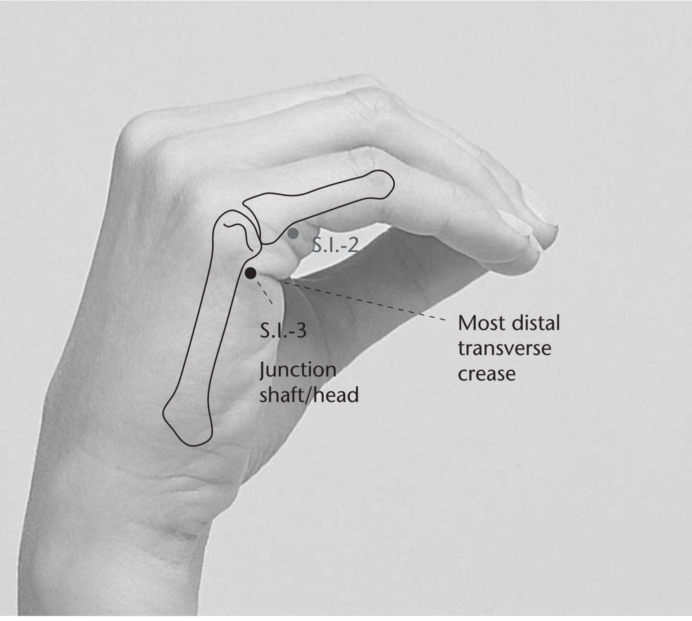

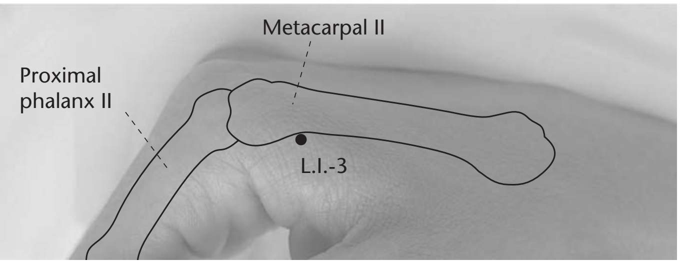

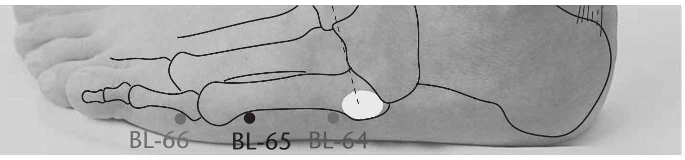

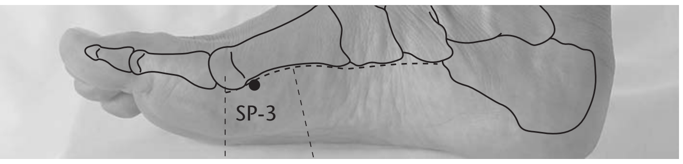

Location

On the ulnar border of the hand, in the depression proximal to the metacarpophalangeal joint, at the border of the 'red and white' skin.

How to find

Quick method: Slide from proximal (wrist) to distal (little finger) along the ulnar border of the hand, until the palpating finger comes to rest at the junction between the base and the head of the 5th metacarpal bone proximal to the metacarpophalangeal joint. Here, locate **S.I.-3. Or:** Locate, with the patient making a loose fist. Locate **S.I.-3** at the end of the most distal transverse crease (this often originates between the index and middle fingers and runs to the little finger). When making a fist, a small bulge forms at the end of the crease. There, locate **S.I.-3** in a palpable depression slightly proximal and dorsal to the bulge.
→ **L.I.-3** is located in a comparable position on the radial border of the hand. Located in a comparable position on the foot are → **SP-3** (on the medial aspect) and → **BL-65** (on the lateral aspect).

Needling

0.5–1 cun horizontally below the bone – that is, vertically to the skin towards the palm – or 0.2–0.3 cun obliquely in a distal direction. Caution: Painful point.

Actions/Indications

- Eliminates Wind and Heat from the *taiyang* (S.I., BL)
- Opens the channel and the *luo* vessels
- Clears Heat and benefits the orifices
- Regulates the *du mai*, calms the *shen*

Special features

Shu-stream point, Wood point, tonification point, opening (master) point of the *du mai*. Important distal point for pain, stiffness and contractures along the course of the channel and for disorders of the cervical spine.

Ch04.6-F10028.qxd 2/22/08 9:47 PM Page 226

S.I.-4 *Wrist Bone* WANGU

Location

On the ulnar border of the hand, between the 5th metacarpal bone and the carpal bones, on the border of the red and white skin.

How to find

On the ulnar aspect of the hand, palpate along the shaft of the 5th metacarpal bone in a proximal direction, until the finger glides over the curvature of the base and into the cleft between the metacarpal and carpal bones.

→ **S.I.-5** is located on the same line, proximal to the pisiform bone and at the level of the wrist crease.

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Clears pathogenic factors from the *taiyang* channel
- Clears Damp-Heat in the Small Intestine

Special features

Yuan-source point

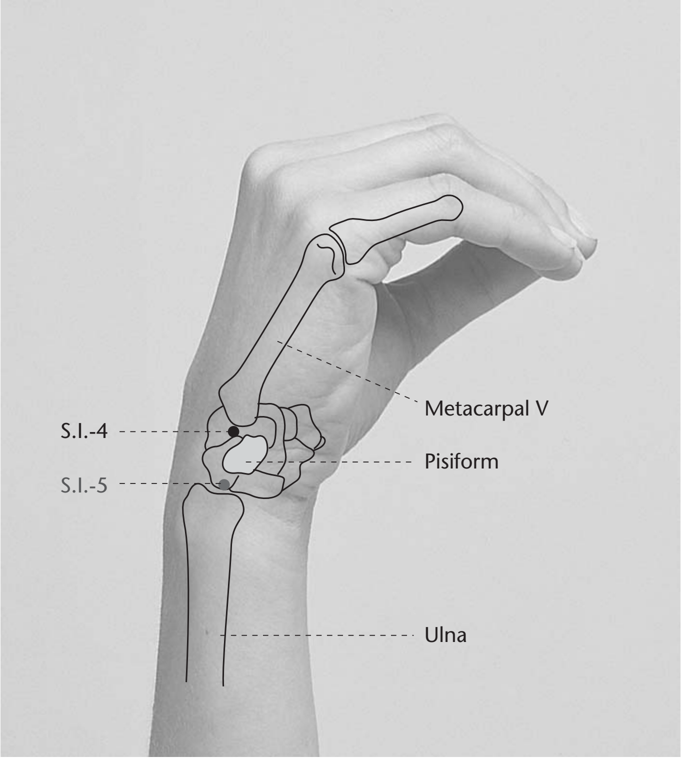
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Ulna

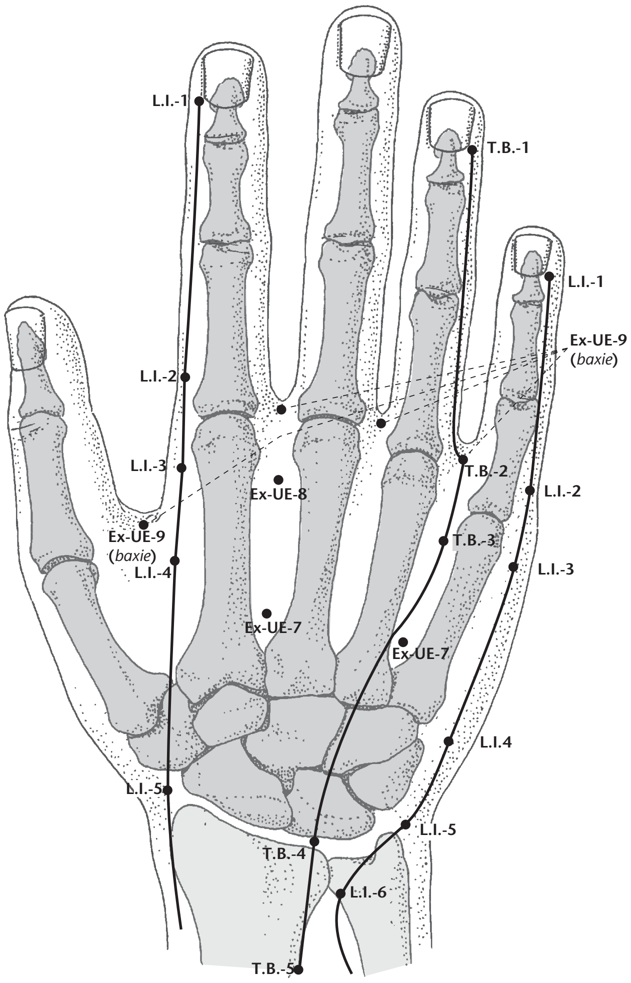
L.I.-1
T.B.-1
L.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
T.B.-2
L.I.-2
Ex-UE-8
T.B.-3
Ex-UE-9
(*baxie*)
L.I.-3
L.I.-4
Ex-UE-7
Ex-UE-7
L.I.4
L.I.-5
L.I.-5
T.B.-4
L.I.-6
T.B.-5

Yang Valley YANGGU S.I.-5

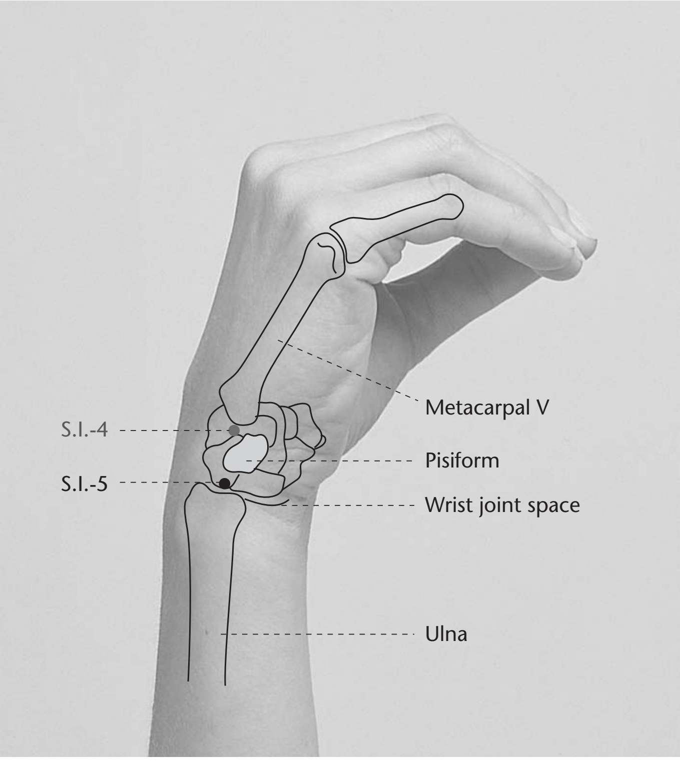
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Wrist joint space
Ulna

Location

On the ulnar aspect of the wrist, at the level of the lateral joint space.

How to find

By moving the wrist in a relaxed way, the wrist joint space can be palpated on the ulnar side. **S.I.-5** is located directly distal to the styloid process of the ulna (➞ 3.3.3), on a line running along the ulnar border of the hand and continuing to the wrist. Located on the same line and distal to the prominence of the pisiform bone is ➞ **S.I.-4**.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Clears Heat and reduces swellings

Special features

Jing-river point, Fire point, *ben* point (Five Phases point).

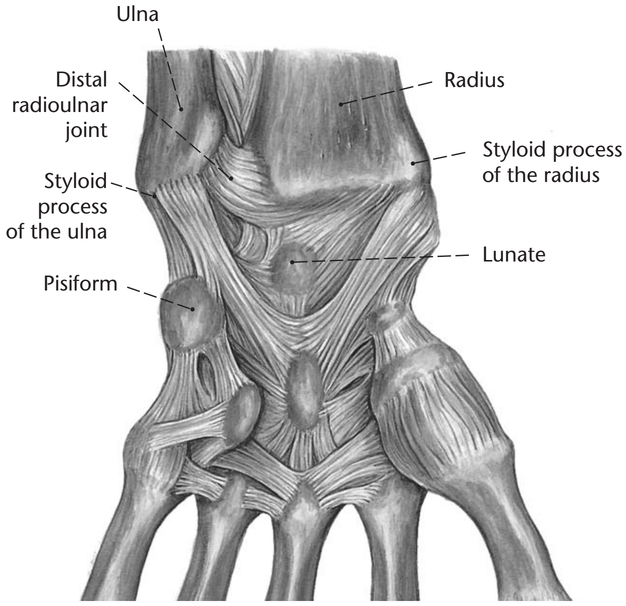
Ulna
Distal radioulnar joint
Styloid process of the ulna
Pisiform
Radius
Styloid process of the radius
Lunate

S.I.-6 *Support the Aged* YANGLAO

Location

In the depression radial and proximal to the styloid process of the ulna, which forms when the hand is turned from a pronated to a supinated position.

How to find

This point is best located with the elbow slightly flexed. Place the palpating finger on the distal portion of the styloid process of the ulna (→ 3.3.3). When turning the hand from a pronated to a supinated position (mnemonic: cut **br**ead and eat **soup**), the palpating finger will glide into a bony cleft at the proximal aspect of the styloid process (the tendon of the extensor carpi ulnari muscle runs along this groove). This cleft also becomes palpable when the patient rests the hand on the chest (half-supinated position). Locate **S.I.-6** in this bony groove.

Needling

0.3–0.8 cun vertically or slightly obliquely in a proximal direction towards → **P-6** or along the tendon towards the elbow joint.

Actions/Indications

- Opens the channel, alleviates pain, benefits the shoulder and arm, alleviates acute conditions
- Benefits the eyes

Special features

Xi-cleft point. Important distal point for pain and limited range of motion of the occiput and shoulder (for example, acute torticollis) and the lumbar region (for example, acute sciatica). For acute and painful impairment of motion, use strong reducing needling technique while the patient gently moves the affected region.

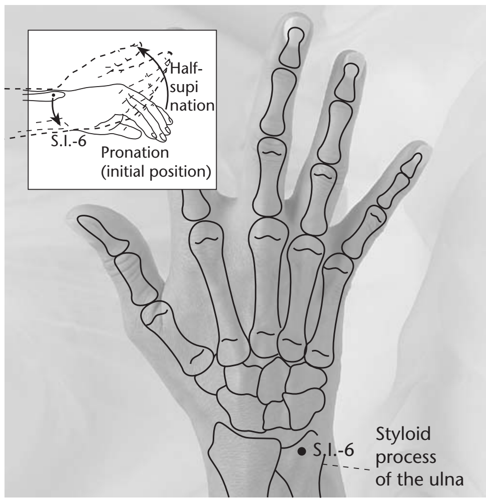
Half-
supi
nation
S.I.-6
Pronation
(initial position)
S.I.-6
Styloid
process
of the ulna

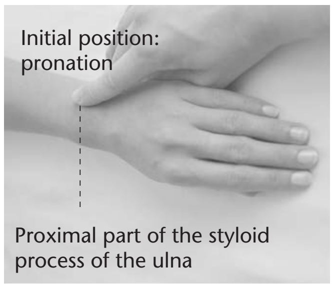
Initial position:
pronation
Proximal part of the styloid
process of the ulna

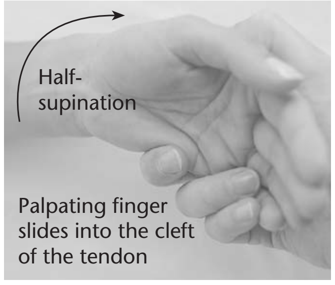
Half-
supination
Palpating finger
slides into the cleft
of the tendon

Branch of the Upright ZHIZHENG S.I.-7

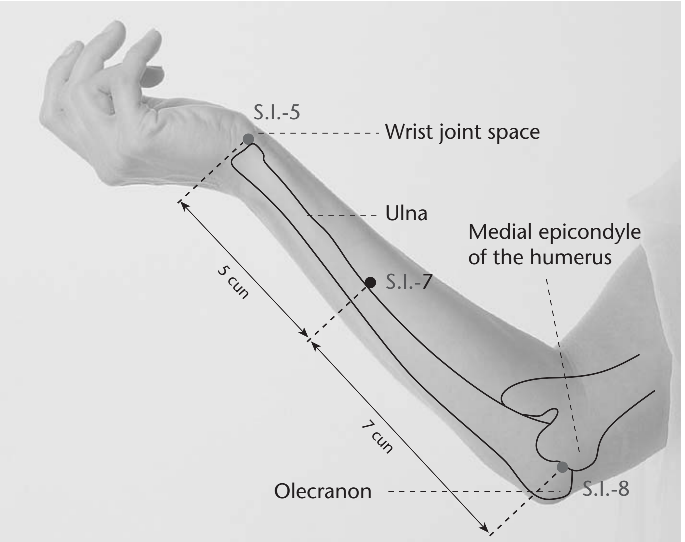
S.I.-5
Wrist joint space
Ulna
Medial epicondyle of the humerus
5 cun
S.I.-7
7 cun
Olecranon
S.I.-8

Location

5 cun proximal to the wrist joint, on the line connecting the ulnar wrist joint space with the ulnar sulcus of the elbow (➞ **S.I.-5** to ➞ **S.I.-8**) or 1 cun distal to the midpoint of this line.

How to find

Find the midpoint of the distance between ➞ **S.I.-5** and ➞ **S.I.-8** by using the spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **S.I.-5** and ➞ **S.I.-8** respectively and span the hands evenly, with the thumbs joining at the midpoint. From the midpoint, measure 1 cun in a distal direction. The point lies between the palpable border of the ulna and the more anteriorly located flexor carpi ulnaris muscle.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Opens the channel
- Releases the Exterior
- Calms the *shen*

Special features

Luo-connecting point

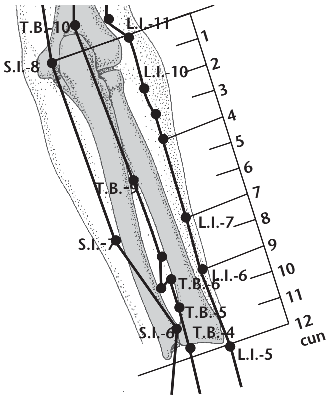
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
8
9
10
11
12
cun
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5

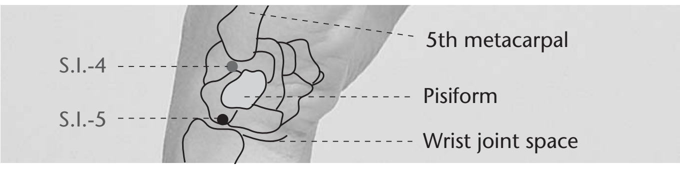
S.I.-4
S.I.-5
5th metacarpal
Pisiform
Wrist joint space

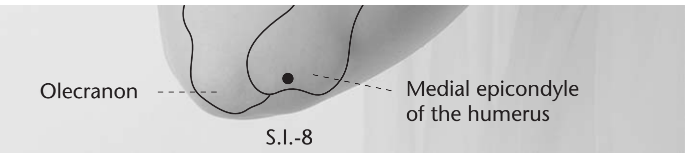
Olecranon
Medial epicondyle of the humerus
S.I.-8

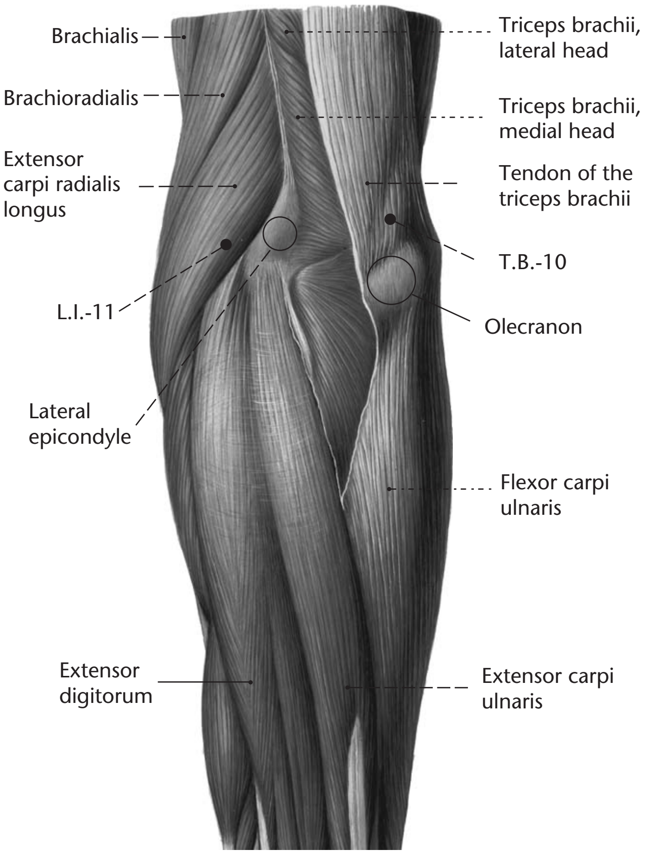
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Extensor digitorum
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris
Extensor carpi ulnaris

S.I.-8 *Small Sea* XIAOHAI

Location

With the elbow flexed, in the depression between the olecranon process of the ulna and the medial epicondyle of the humerus.

How to find

S.I.-8 is located on a line connecting the tip of the olecranon and the tip of the medial epicondyle, at the deepest point of the groove between those two bony protrusions.

Needling

0.3–0.5 cun vertically or obliquely with or against the flow of the channel. Caution: The ulnar nerve is located in this groove.

Actions/Indications

- Opens the channel

Special features

He-sea point, Earth point, sedation point.

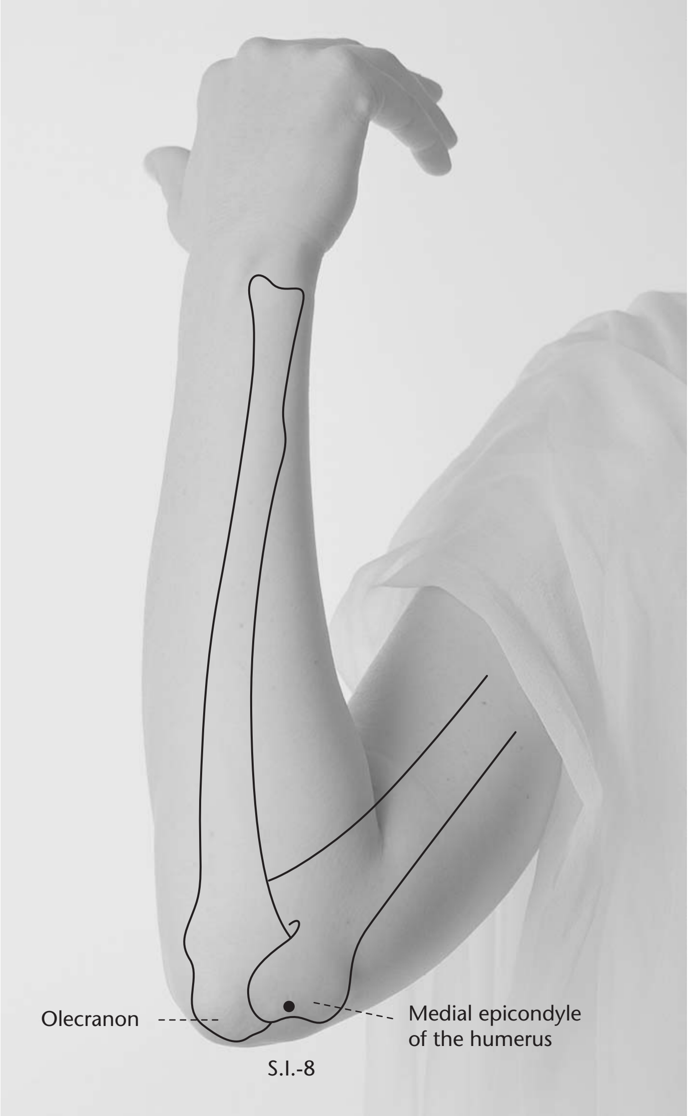
Olecranon
Medial epicondyle of the humerus
S.I.-8

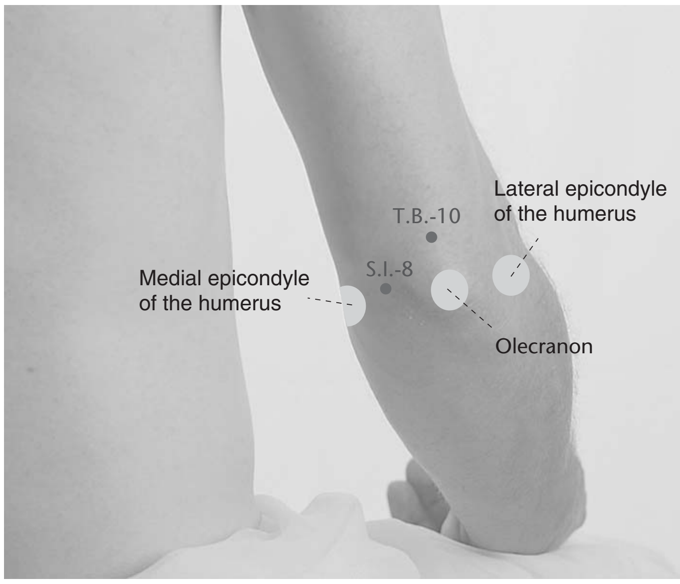
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

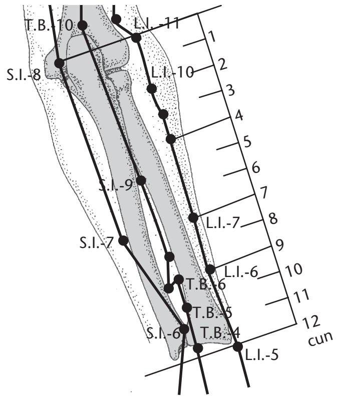
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
8
9
10
11
12
cun
S.I.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5

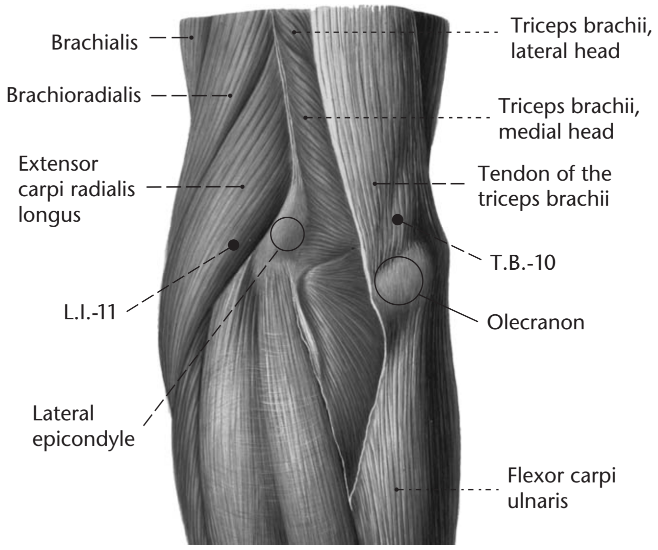
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris

True Shoulder JIANZHEN S.I.-9

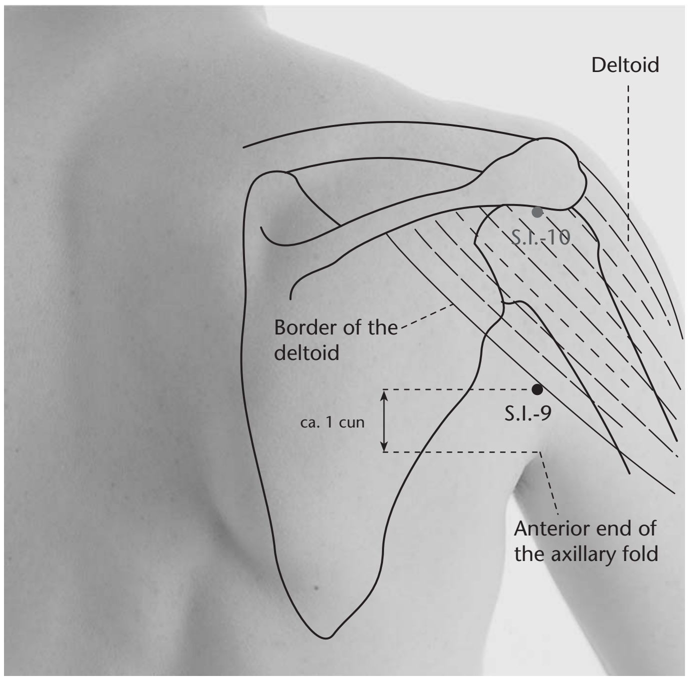
Deltoid
S.I.-10
Border of the deltoid
ca. 1 cun
S.I.-9
Anterior end of the axillary fold

Location

With the arm adducted, 1 cun superior to the posterior axillary fold, on the lower border of the deltoid muscle.

How to find

With the patient sitting upright, palpate upward from the end of the posterior axillary fold, until you can feel the lower border of the deltoid muscle. If in doubt, ask the patient to flex their muscle. **S.I.-9** is located on its lower border.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel, expels Wind and benefits the shoulder

Special features

S.I.-9 corresponds to a trigger point that is often sensitive to pressure.

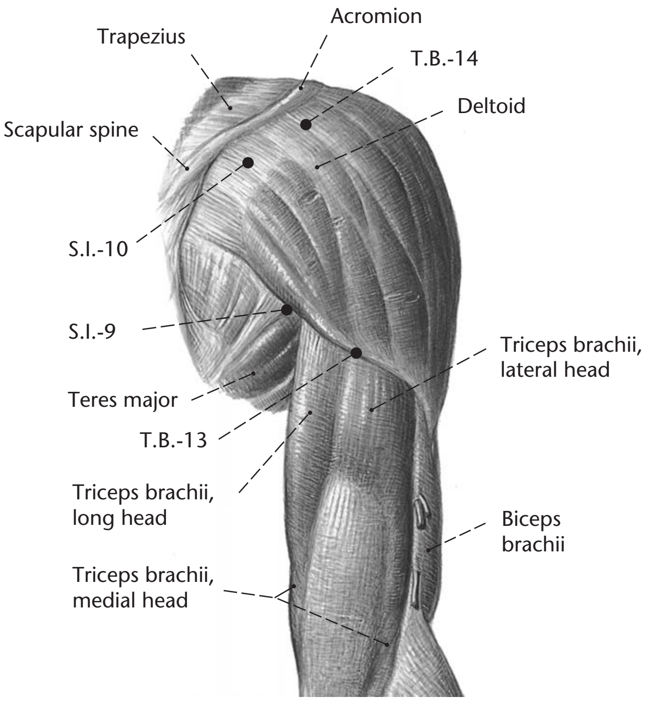
Trapezius
Acromion
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

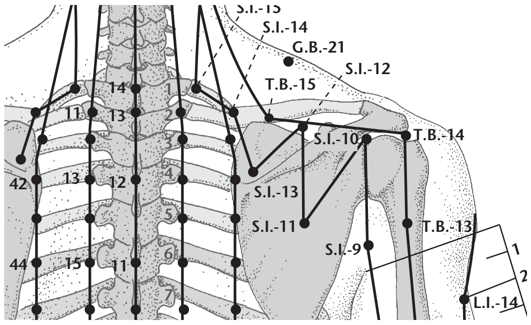
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
3
S.I.-10
T.B.-14
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
1
44
15
11
6
2
7
L.I.-14

Ch04.6-F10028.qxd 2/22/08 9:48 PM Page 232

S.I.-10 *Upper Arm* Shu NAOSHU

Location

With the arm adducted, on an imaginary line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine.

How to find

With the patient sitting upright, palpate from the posterior axillary fold upward, past the lower border of the deltoid muscle (→ **S.I.-9**) to the bony resistance of the scapular spine and acromion, which here form a curve that opens towards the inferior. **S.I.-10** is located below its bony border.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel and relaxes the sinews

Special features

Meeting point of the *yang wei mai* and the *yang qiao mai*. S.I.10 corresponds to a trigger point that is often sensitive to pressure. It also corresponds to the posterior access for puncturing the shoulder joint.

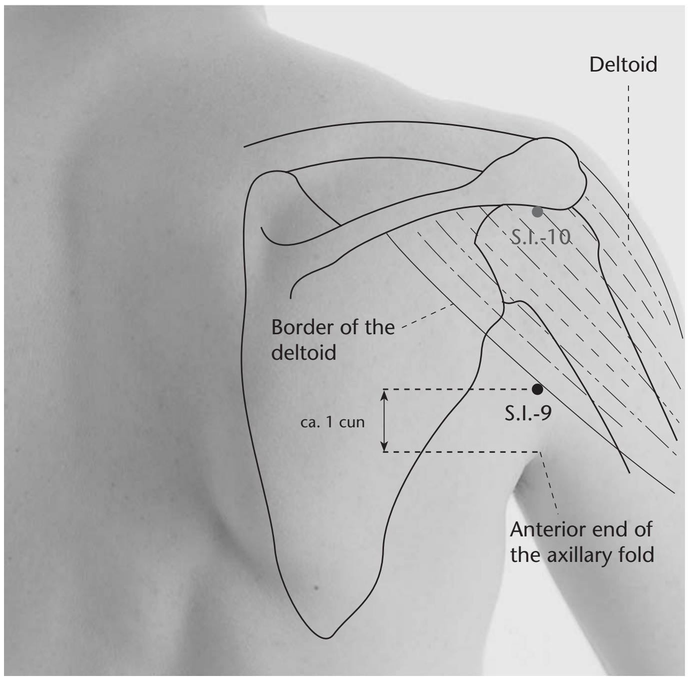
Deltoid
S.I.-10
Border of the deltoid
ca. 1 cun
S.I.-9
Anterior end of the axillary fold

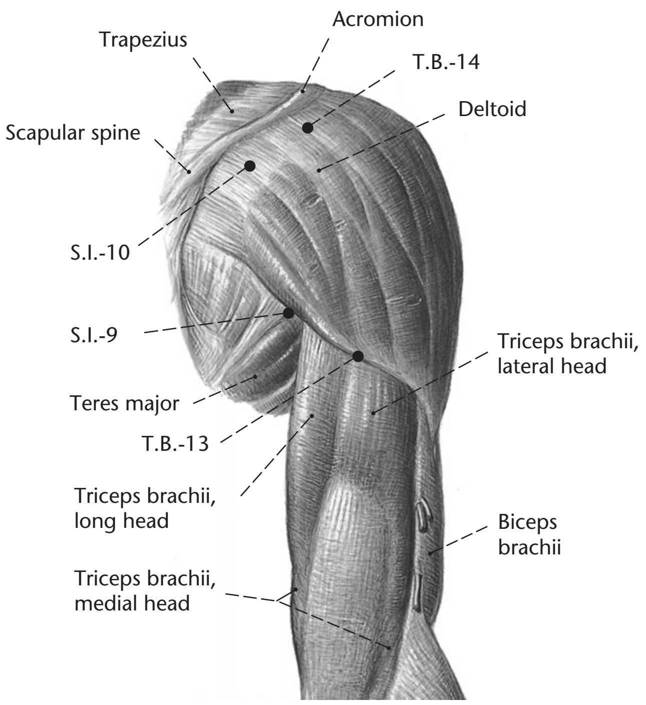
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

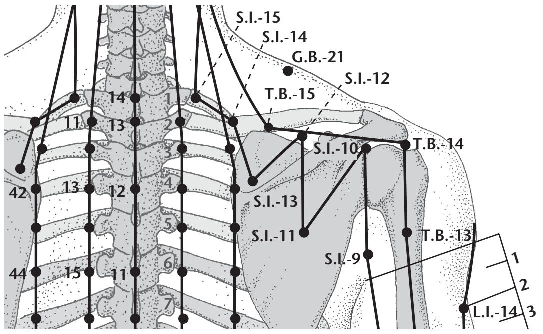
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
S.I.-10
T.B.-14
3
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
1
44
15
11
6
2
7
L.I.-14
3

Heavenly Gathering TIANZONG S.I.-11

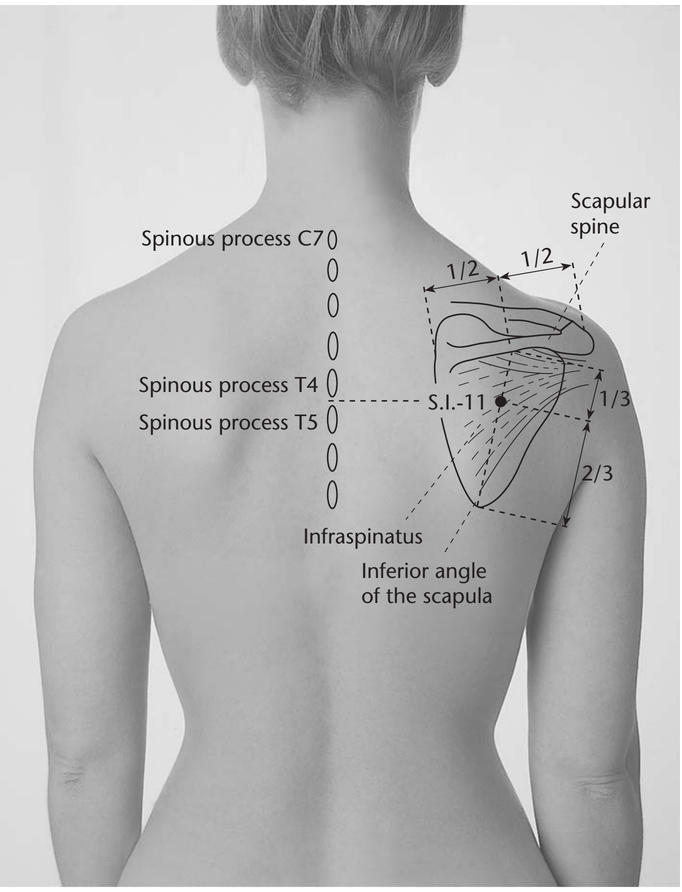
Scapular spine
Spinous process C7
1/2
1/2
Spinous process T4
Spinous process T5
S.I.-11
1/3
2/3
Infraspinatus
Inferior angle of the scapula

Location

On the scapula, in a depression on the infraspinatus muscle, one third of the distance from the midpoint of the scapular spine and the inferior angle of the scapula.

How to find

Patient's position: Prone position or better seated with relaxed shoulders. Draw an imaginary line from the midpoint of the scapular spine (→ 3.3.1) to the inferior angle of the scapula. **S.I.-11** is located at the junction of the upper and middle third of this line, in a depression on the infraspinatus muscle, which is often sensitive to pressure. In a seated or standing position with the arms hanging down, it will generally be level with the spinous process of the 4th thoracic vertebra or the body of the 5th thoracic vertebra.

→ **S.I.-12** is located directly above **S.I.-11**, in the centre of the supraspinous fossa.

Needling

0.5–1.5 cun vertically or obliquely into the muscle tissue

Actions/Indications

- Opens the channel, alleviates pain, moves Qi locally, unbinds the chest and lateral costal region
- Benefits the breasts

Special features

S.I.-11 is often a trigger point on the infraspinatus muscle.

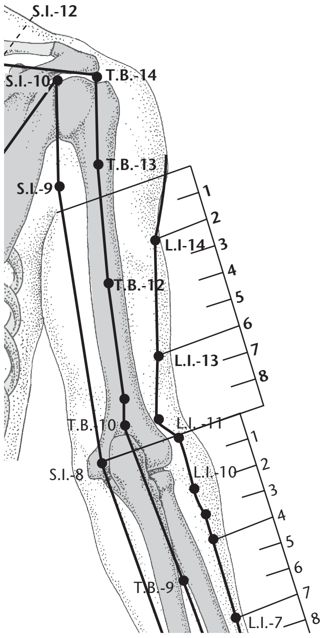
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I-14
3
4
5
T.B.-12
6
7
L.I.-13
8
T.B.-10
L.I.-11
1
S.I.-8
L.I.-10
2
3
4
5
6
7
T.B.-9
L.I.-7
8

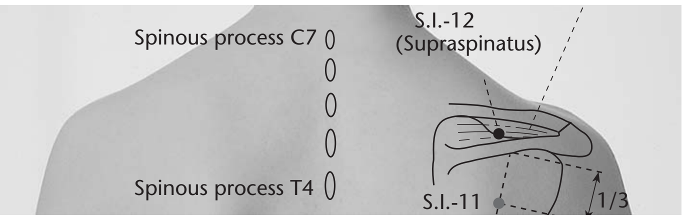
Spinous process C7
S.I.-12 (Supraspinatus)
Spinous process T4
S.I.-11
1/3

S.I.-12 *Grasping the Wind* BINGFENG

Location

Directly above → **S.I.-11**, in the centre of the supraspinous fossa.

How to find

Patient's position: Prone position or better seated with relaxed shoulders. First, locate → **S.I.-11** (one third of the distance between the midpoint of the scapular spine and the inferior angle of the scapula). **S.I.-12** is located directly above → **S.I.-11** in the centre of the supraspinous fossa.

Needling

0.5–1 cun vertically or obliquely into the supraspinatus muscle in the direction of the shoulder. Caution: Pneumothorax.

Actions/Indications

- Eliminates Wind, benefits the shoulder and scapula, opens the channel and the *luo* vessels

Special features

Meeting point with the Gall Bladder, Triple Burner and Large Intestine channels. **S.I.-12** is often a trigger point in the supraspinatus muscle.

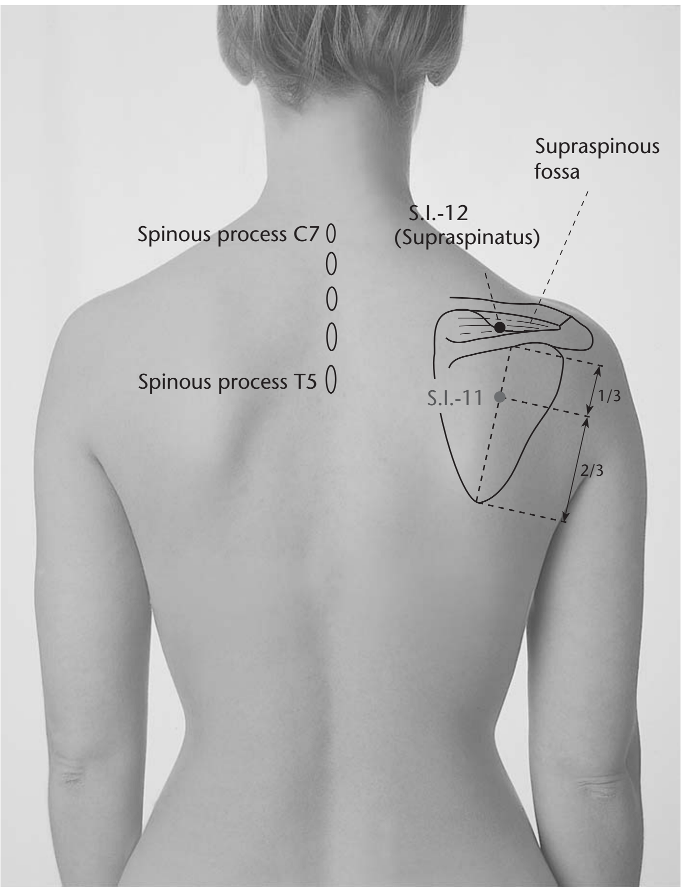
Supraspinous fossa
S.I.-12 (Supraspinatus)
Spinous process C7
Spinous process T5
S.I.-11
1/3
2/3

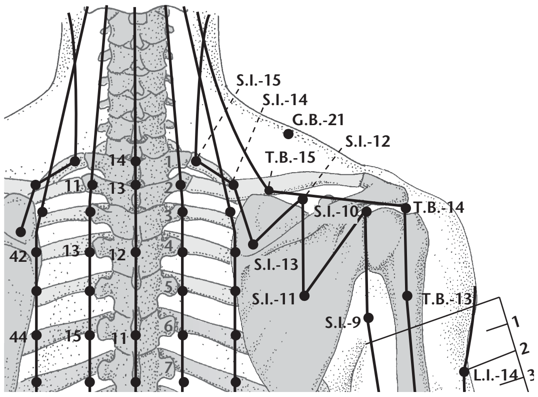
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
3
S.I.-10
T.B.-14
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
44
15
11
6
7
1
2
L.I.-14
3

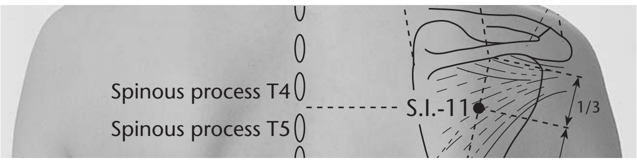
Spinous process T4
Spinous process T5
S.I.-11
1/3

Crooked Wall QUYUAN S.I.-13

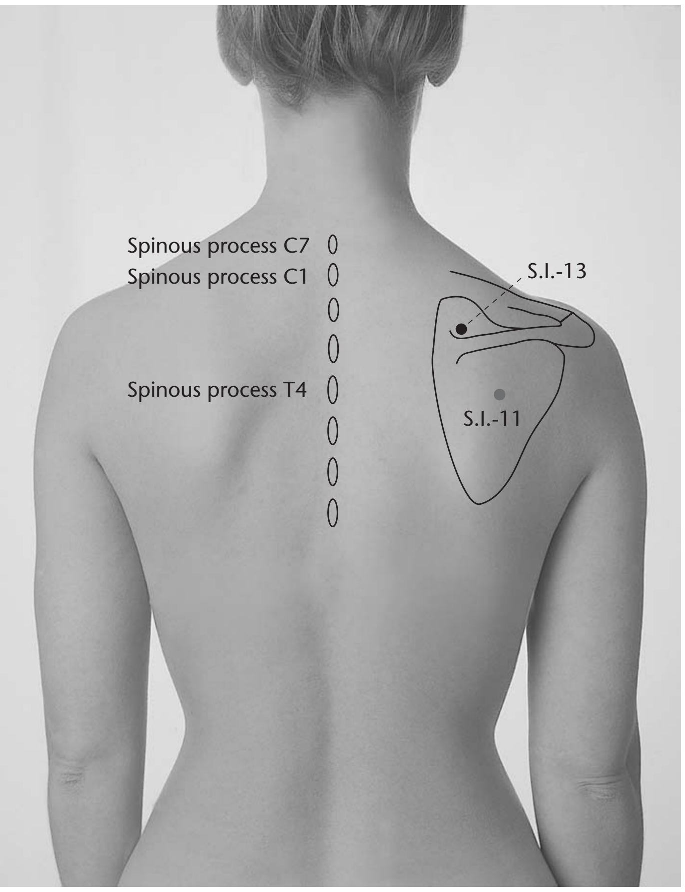
Spinous process C7
Spinous process C1
Spinous process T4
S.I.-13
S.I.-11

Location

At the medial end of the supraspinous fossa.

How to find

At its medial end, the scapular spine fans out in a deltoid shape, curving round in a superior direction. When palpating, this feels like a ‘crooked wall’, hence its name. **S.I.-13** is located on the scapula, in the slightly concave bend, at the medial origin of the supraspinatus muscle.

Needling

0.5–1 cun vertically or obliquely in a lateral direction. No danger of puncturing the pleura if located correctly.

Actions/Indications

- Opens the channel

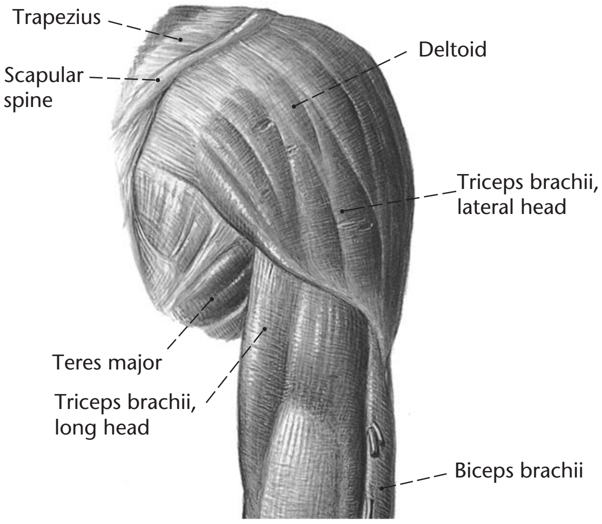
Trapezius
Scapular spine
Deltoid
Triceps brachii, lateral head
Teres major
Triceps brachii, long head
Biceps brachii

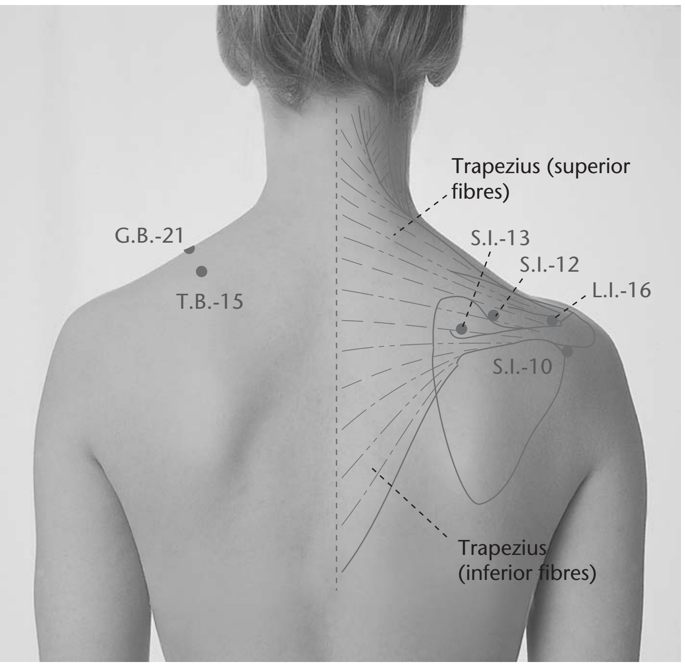
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
T.B.-15
L.I.-16
S.I.-10
Trapezius (inferior fibres)

S.I.-14 *Outer Shoulder Shu* JIANWAISHU

Location

3 cun lateral to the lower border of the spinous process of the 1st thoracic vertebra, at the insertion of the levator scapulae muscle.

How to find

The levator scapulae muscle originates on the upper aspect of the medial border and the superior angle of the scapula. In this area, it is often myogelotic and therefore well defined and often sensitive to pressure (trigger point). **S.I.-14** is located just superior to the actual muscle insertion; in contrast to → **S.I.-13**, it is not located on the scapula.

Located on approximately the same level are → **Du-13** (below the spinous process of the 1st thoracic vertebra) and → **BL-11** (1.5 cun distal to the posterior midline).

Needling

0.5–1 cun vertically or obliquely in a medial direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel
- Expels Wind-Cold

Special features

Trigger point on the levator scapulae muscle

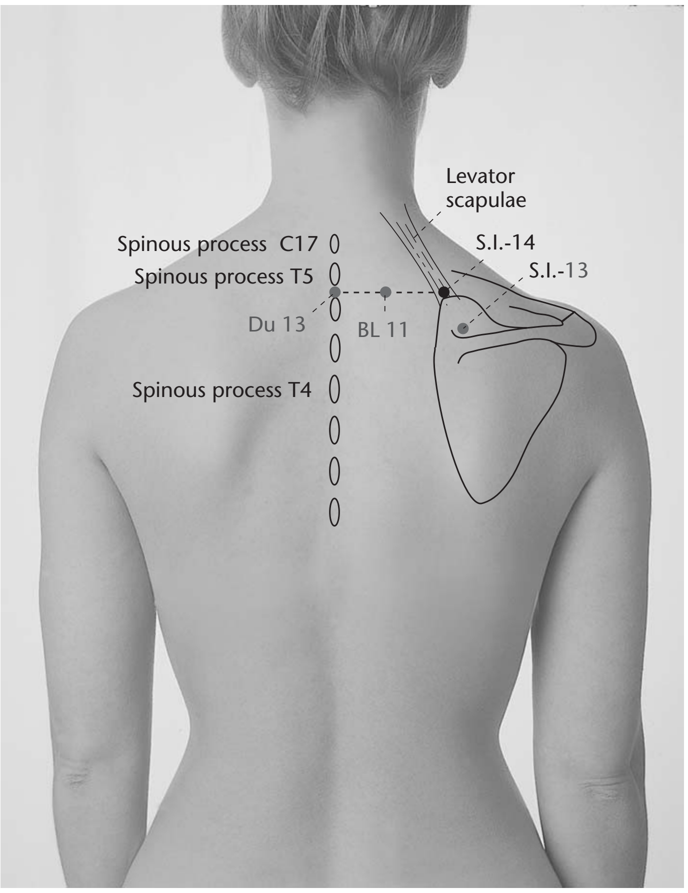
Levator scapulae
Spinous process C17
Spinous process T5
S.I.-14
S.I.-13
Du 13
BL 11
Spinous process T4

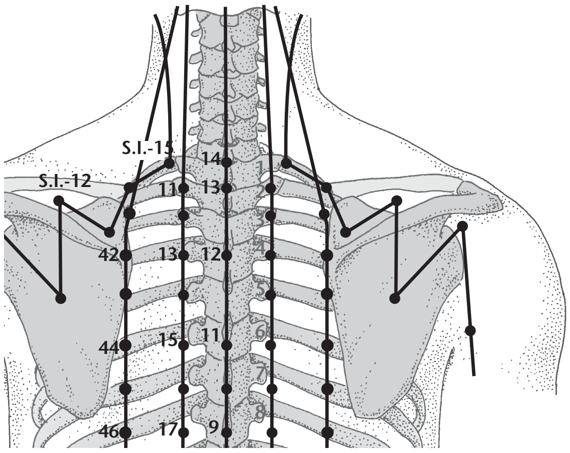
S.I.-15
S.I.-12
14
11
13
42
13
12
44
15
11
46
17
9
1
2
3
4
5
6
7
8

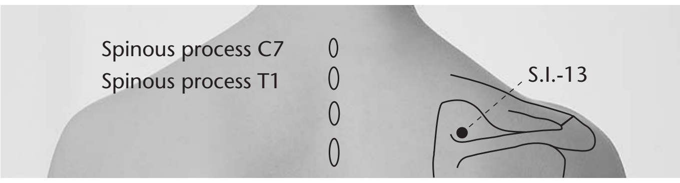
Spinous process C7
Spinous process T1
S.I.-13

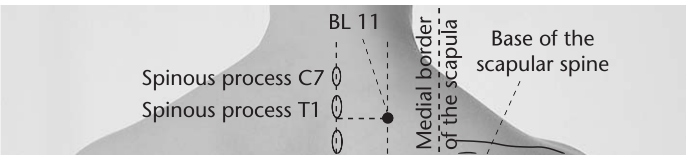
BL 11
Spinous process C7
Spinous process T1
Medial border of the scapula
Base of the scapular spine

Middle Shoulder Shu JIANZHONGSHU S.I.-15

Location

2 cun lateral to the lower border of the spinous process of the 7th thoracic vertebra.

How to find

For locating the 7th cervical vertebra, see → 3.4.
Located on the same level is → **Du-14** below the spinous process of the 7th cervical vertebra. Locate **S.I.-15** by measuring 2 cun in a lateral direction.

Needling

0.5–1 cun vertically or obliquely in a medial direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel

Special features

Trigger point on the levator scapulae muscle, effective point for disorders of the cervicodorsal area.

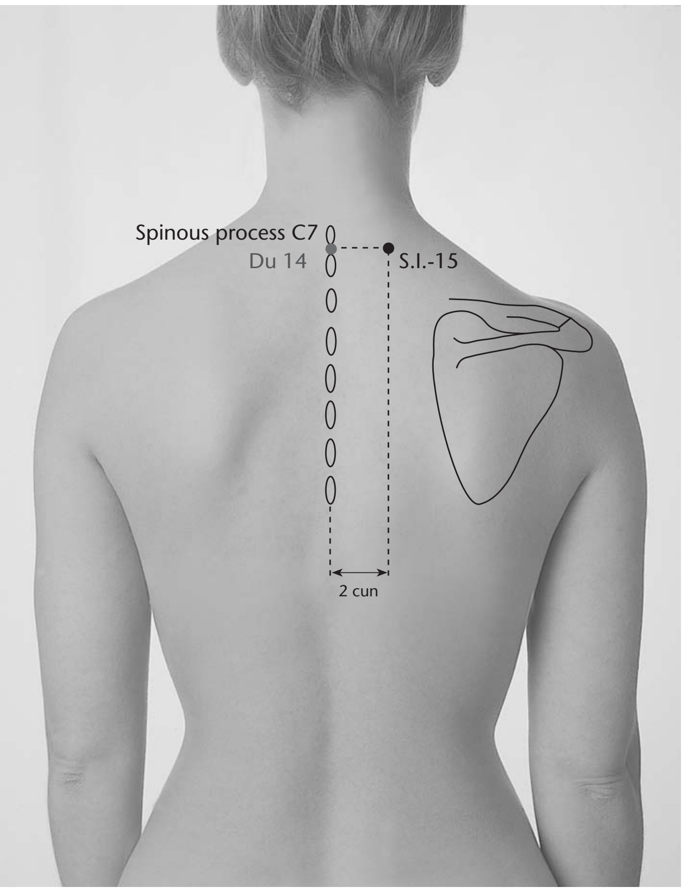
Spinous process C7
Du 14
S.I.-15
2 cun

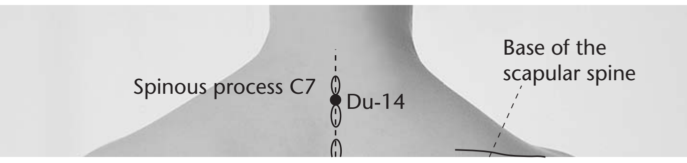
Spinous process C7
Du-14
Base of the scapular spine

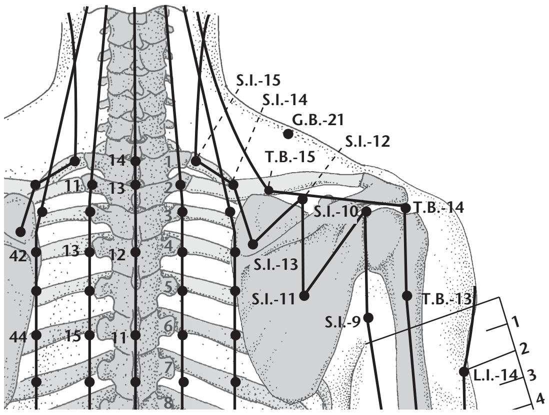
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
S.I.-10
T.B.-14
3
42
13
12
4
S.I.-13
5
S.I.-11
T.B.-13
S.I.-9
1
44
15
11
6
2
L.I.-14
3
7
8
4

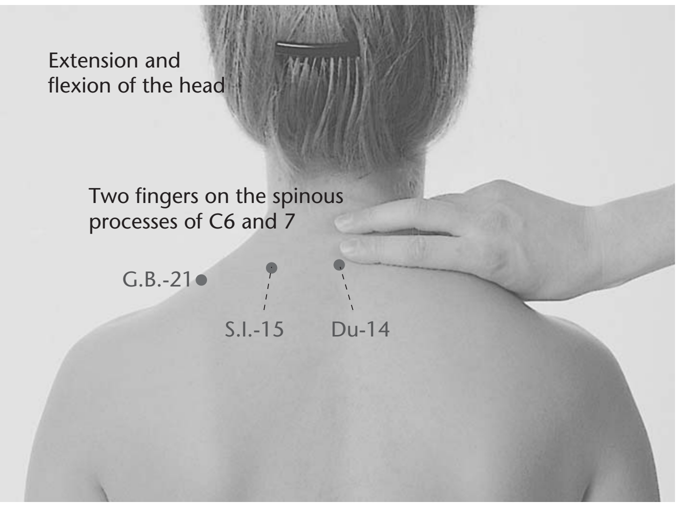
Extension and flexion of the head
Two fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

S.I.-16 *Heavenly Window* TIANCHUANG

Location

Approximately 3.5 cun lateral to the anterior midline, at the level of the laryngeal prominence, on the posterior border of the sternocleidomastoid muscle.

How to find

From the laryngeal prominence (→ 3.2), draw a line to the posterior border of the sternocleidomastoid muscle. **S.I.-16** is located directly at its posterior border, which can be made more clearly visible and palpable by rotating the patient's head in the opposite direction.

Located on the same level are → **ST-9** (on the anterior border of the sternocleidomastoid muscle) and → **L.I.-18** (between the two heads of the SCM muscle).

Needling

Vertically 0.5–1 cun. If necessary, rotate the patient's head back to its normal position before needle insertion.

Actions/Indications

- Expels Wind and Heat, benefits the throat, ears and voice
- Opens the channel

Special features

Window of Heaven point

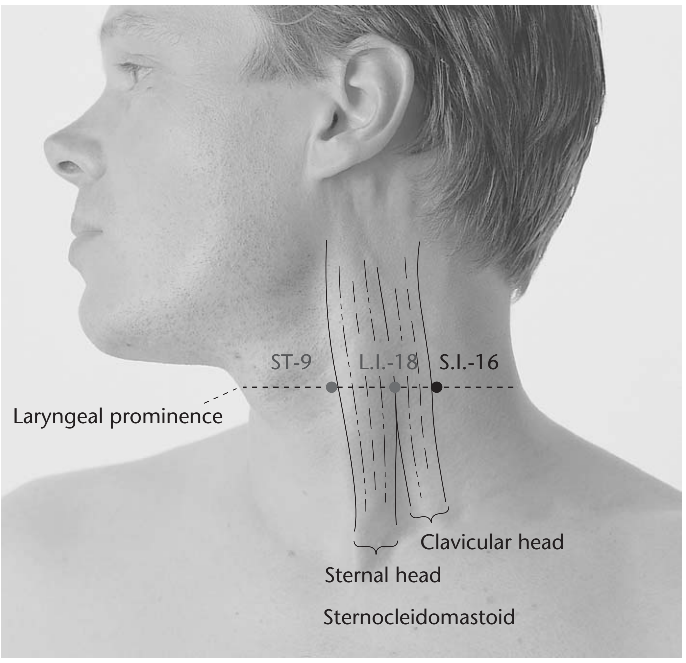
ST-9
L.I.-18
S.I.-16
Laryngeal prominence
Clavicular head
Sternal head
Sternocleidomastoid

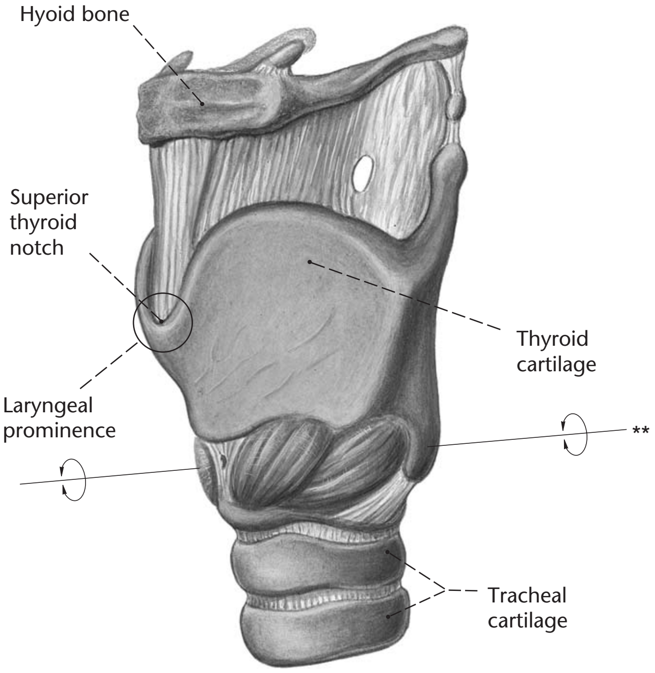
Hyoid bone
Superior thyroid notch
Thyroid cartilage
Laryngeal prominence
**
Tracheal cartilage

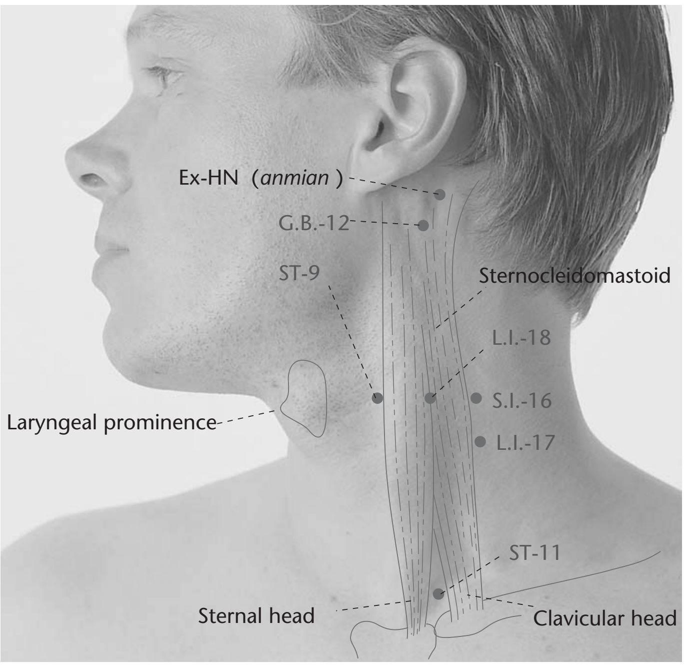
Ex-HN (*anmian*)
G.B.-12
ST-9
Sternocleidomastoid
L.I.-18
S.I.-16
L.I.-17
Laryngeal prominence
ST-11
Sternal head
Clavicular head

Heavenly Appearance TIANRONG S.I.-17

Location

Posterior to the mandibular angle, on the anterior border of the sternocleidomastoid muscle.

How to find

By rotating the patient's head in the opposite direction, the border of the sternocleidomastoid muscle becomes more pronounced. **S.I.-17** is located halfway between the angle of the mandible and the anterior border of the sternocleidomastoid muscle.

Needling

Vertically 0.5–1 cun. Caution: Internal carotid artery, internal and external jugular veins.

Actions/Indications

- Expels Wind, opens the channel, disperses accumulations, benefits the ears and throat

Special features

Window of Heaven point

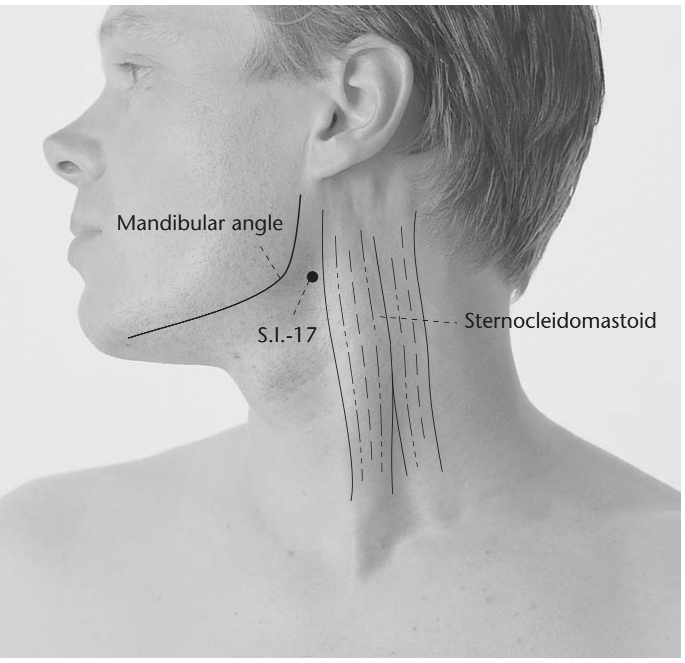
Mandibular angle
S.I.-17
Sternocleidomastoid

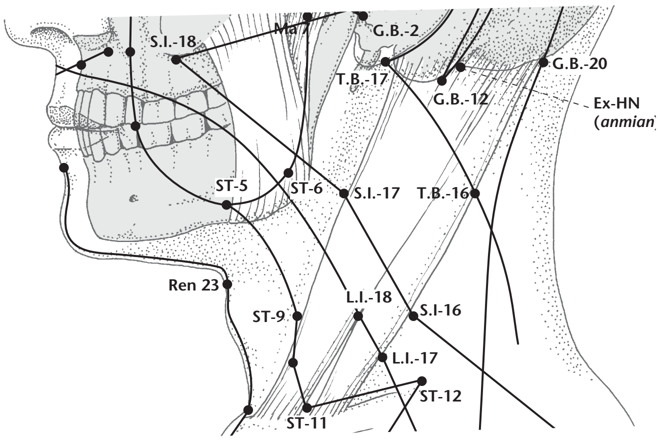
S.I.-18
Ma 7
G.B.-2
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren 23
ST-9
L.I.-18
S.I-16
L.I.-17
ST-12
ST-11

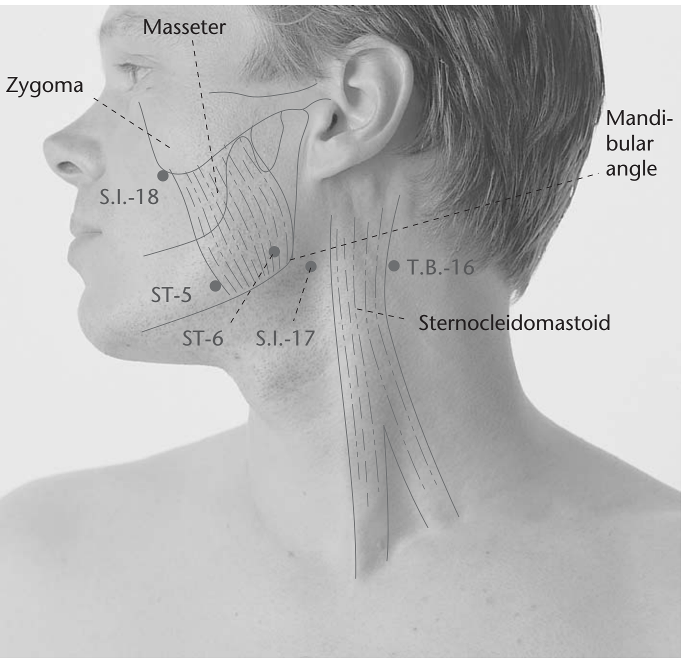
Masseter
Zygoma
Mandi-
bular
angle
S.I.-18
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

S.I.-18 *Cheekbone Crevice* QUANLIAO

Location

At the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone, on the anterior border of the masseter muscle.

How to find

Draw an imaginary vertical line from the outer canthus of the eye to the lower border of the zygomatic bone. **S.I.-18** is located in a depression on the anterior border of the masseter muscle, approximately on the level of the lateral border of the ala nasi (→ **L.I.-20**). By making chewing movements, the anterior border of the masseter muscle becomes easily palpable.

Needling

0.3–0.5 cun vertically or 1–1.5 cun transversely (subcutaneously) in the direction of → **ST-4,** → **ST-7** and → **L.I.-20**. According to some texts, moxibustion is contraindicated.

Actions/Indications

- Eliminates Wind, alleviates pain, clears Heat, reduces swellings

Special features

Meeting point with the Triple Burner channel. Important local point for disorders of the face (caused by internal or external Wind).

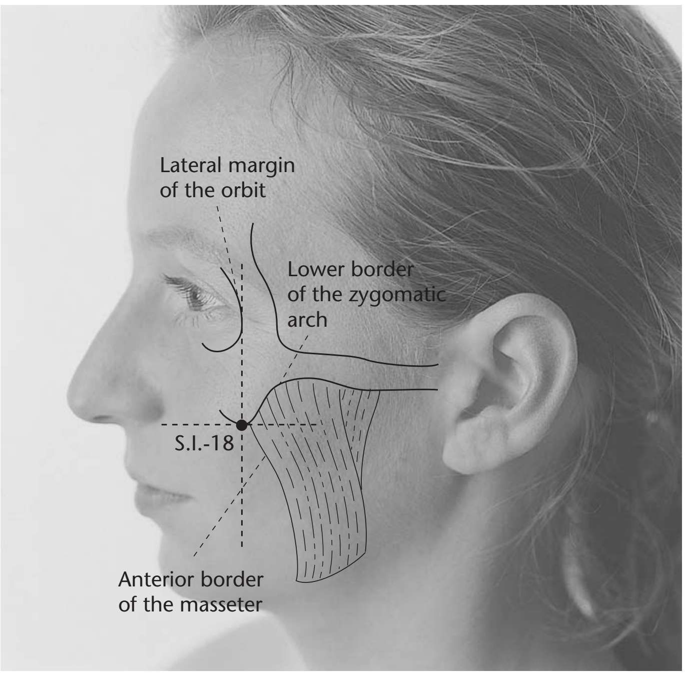
Lateral margin of the orbit
Lower border of the zygomatic arch
S.I.-18
Anterior border of the masseter

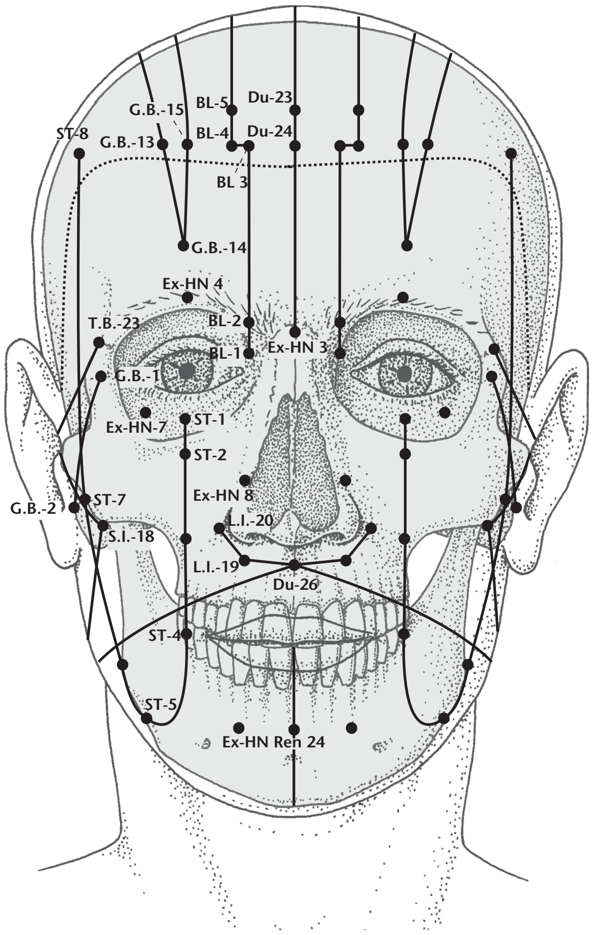
ST-8
G.B.-15
BL-5
Du-23
G.B.-13
BL-4
Du-24
BL 3
G.B.-14
Ex-HN 4
T.B.-23
BL-2
BL-1
Ex-HN 3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN 8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren 24

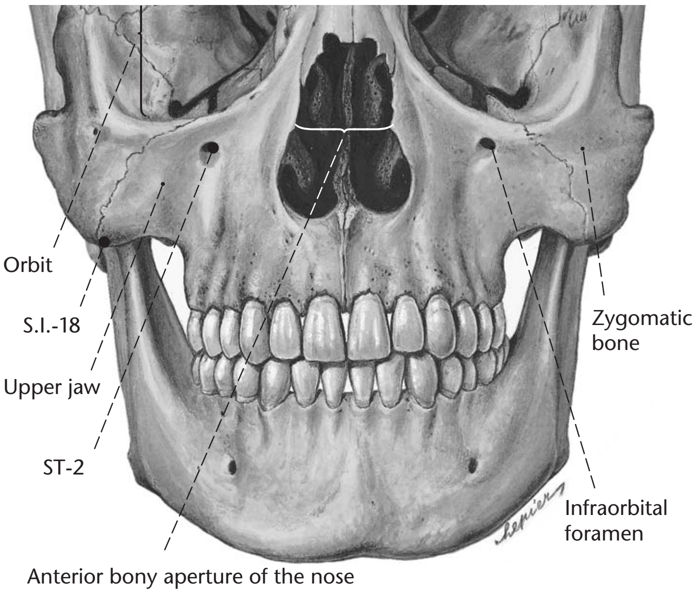
Orbit
S.I.-18
Upper jaw
ST-2
Zygomatic bone
Infraorbital foramen
Anterior bony aperture of the nose

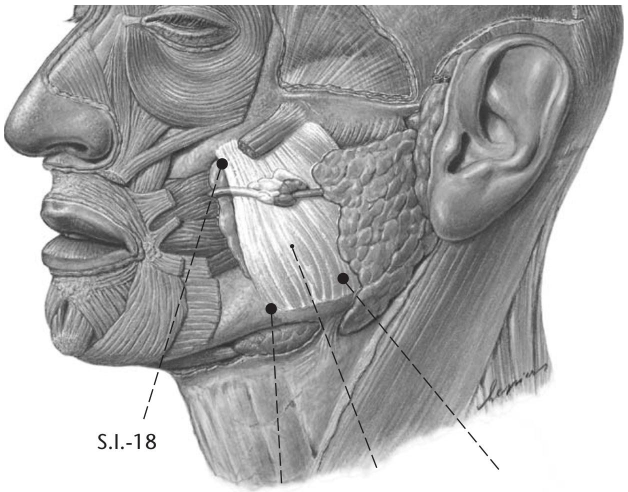

Palace of Hearing TINGGONG S.I.-19

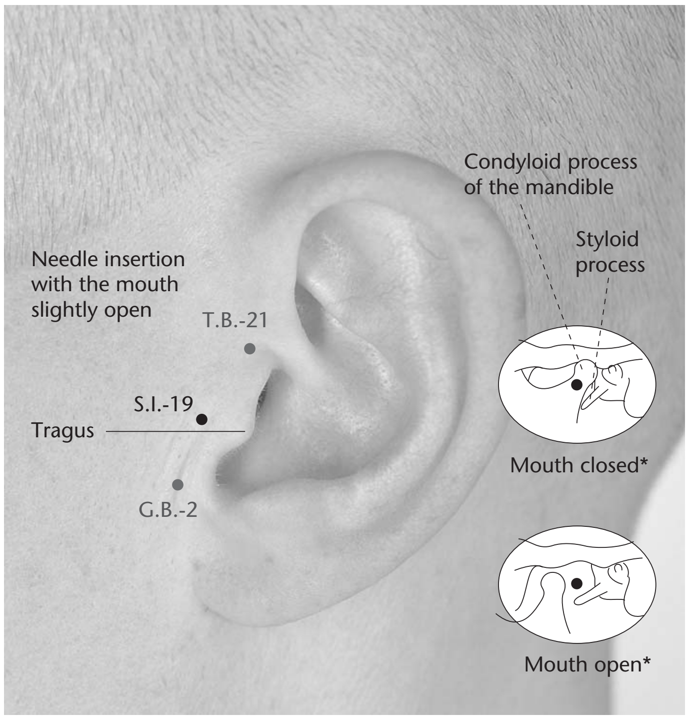
Condyloid process of the mandible
Styloid process
Needle insertion with the mouth slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed*
Mouth open*

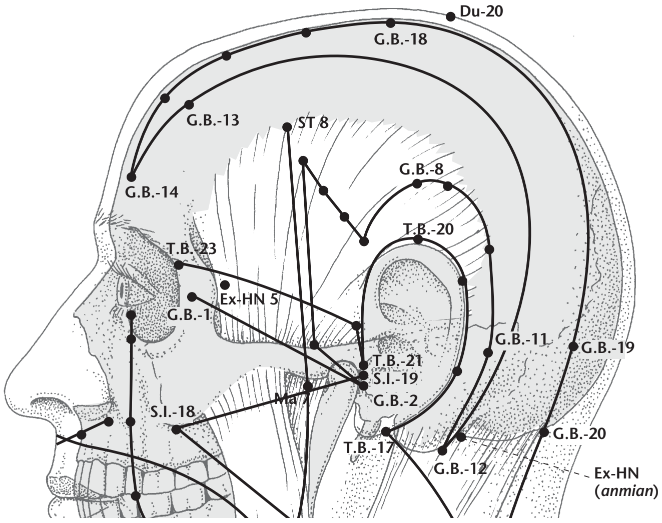
Du-20
G.B.-18
G.B.-13
ST 8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN 5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma
T.B.-17
G.B.-12
G.B.-20
Ex-HN
(anmian)

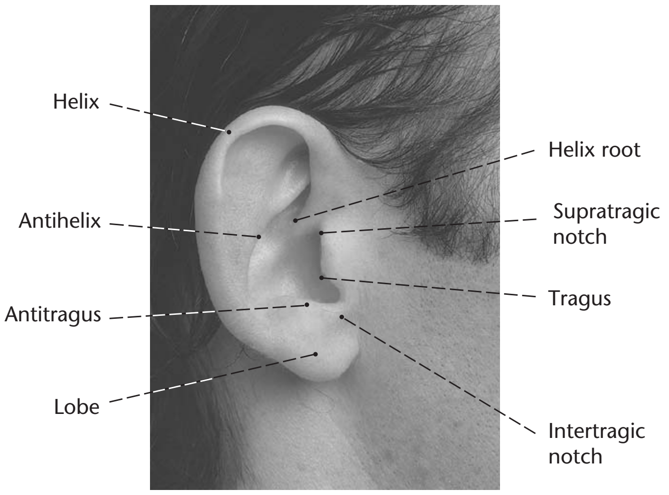
Helix
Helix root
Antihelix
Supratragic notch
Antitragus
Tragus
Lobe
Intertragic notch

Location

Anterior to the ear, with the mouth open at the level of the middle of the tragus, in a depression between the tragus and condyloid process of the mandible.

How to find

Locate and needle this point with the mouth open. This allows the condyloid process of the mandible to slide anteriorly and reveals the depression. Locate the vertical sulcus at the junction between the ear and the cheek (this is not always very pronounced). Then locate **S.I.-19** on the level of the tragus on the sulcus. If you cannot clearly define the sulcus (it becomes more pronounced with increasing age), an (ear) point locator can be used to locate the point (the patient should have their mouth open in order to reveal the depression).

S.I.-19 is the centre point of three points located anteriorly to the ear (→ **T.B.-21** is located proximal and → **G.B.-2** is located distal to **S.I.-19**).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the mouth open (to avoid intra-articular needling). Caution: This point is located close to the superficial temporal artery and the auriculotemporal nerve (as are → **T.B.-21** and → **G.B.-2**).

Actions/Indications

- Benefits the ears
- Calms the *shen*

Special features

Meeting point with the Gall Bladder and Triple Burner channels, exit point. Important local point for disorders of the ears. For temporomandibular joint dysfunction and pain, rather use → **G.B.-2** as a local point.

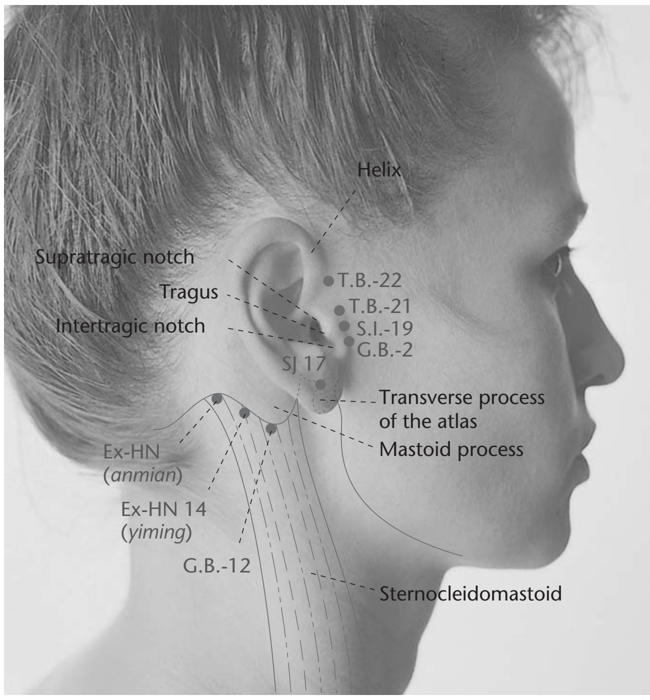
Helix
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
SJ 17
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN 14 (yiming)
G.B.-12
Sternocleidomastoid

---

* According to Deadman et al 1998.

4.7 The Bladder Channel System – Foot *taiyang* (*zu tai yang jing luo*)

4.7.1 The Bladder Primary Channel (*zu tai yang jing*)

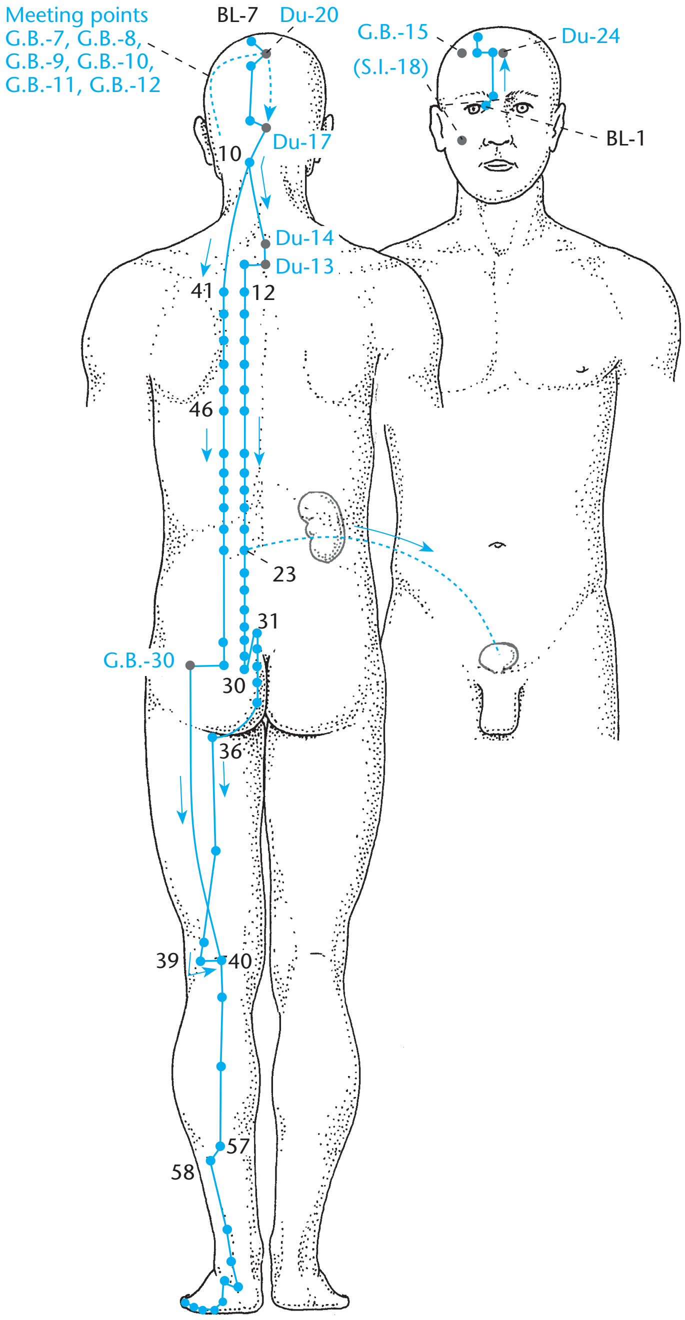
Meeting points
G.B.-7, G.B.-8,
G.B.-9, G.B.-10,
G.B.-11, G.B.-12
BL-7
Du-20
G.B.-15
(S.I.-18)
Du-24
BL-1
Du-17
10
Du-14
Du-13
41
12
46
23
31
G.B.-30
30
36
39
40
57
58

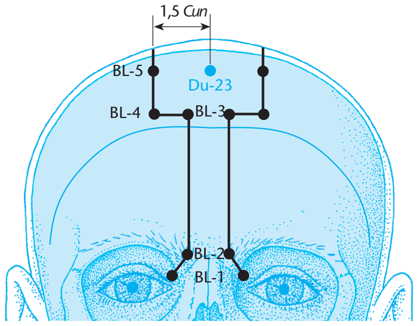
1,5 Cun
BL-5
Du-23
BL-4
BL-3
BL-2
BL-1

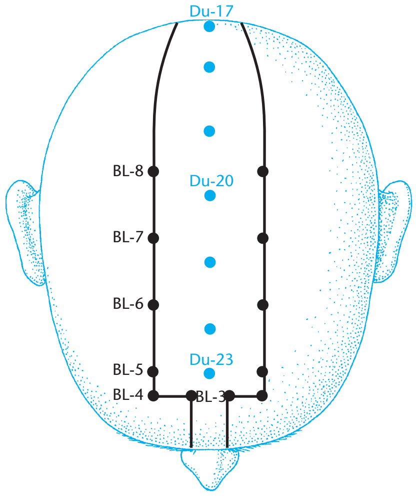
Du-17
BL-8
Du-20
BL-7
BL-6
BL-5
Du-23
BL-4
BL-3

Pathway

The **external** pathway of the Bladder primary channel begins at the inner canthus of the eye at ➞ **BL-1** (*jingming*). This point is reached by a branch that separates from the Small Intestine primary channel at ➞ **S.I.-18** (*quanliao*) (hand–foot pairing of the second great circuit, Yang axes: *taiyang*).

The **external** pathway ascends the forehead, there intersects with ➞ **Du-24** (*shenting*) and ➞ **G.B.-15** (*toulinqi*), continues to the vertex and there connects with the *du mai* at ➞ **Du-20** (*baihui*).

At ➞ **Du-20 two branches** separate from the primary channel:

- ➡ **one branch** travels to the temporal region and intersects the Gall Bladder primary channel at **G.B.-8** (*shuaigu*), **G.B.-7** (*qubin*), **G.B.-9** (*tianchong*), **G.B.-10** (*fubai*), **G.B.-11** (*touqiaoyin*) and **G.B.-12** (*wangu*)
- ➡ the **other branch** penetrates deeper into the brain and re-emerges at ➞ **Du-17** (*naohu*) or, according to some authors, at **BL-8** (*luoque*).

From ➞ **Du-20**, the **external pathway** runs to **BL-8** (*luoque*) and **BL-9** (*yuzhen*), passes ➞ **Du-17** (*naohu*) and continues to **BL-10** (*tianzhu*) in the occipital region. Here, the **external pathway** of the channel separates into **two branches**:

- ➡ the **medial** branch runs to ➞ **Du-14** (*dazhui*) below the spinal process of C7 where it meets the other Yang primary channels, then continues to ➞ **Du-13** (*taodao*). From T1 it runs parallel to the posterior midline, 1.5 cun lateral to it.

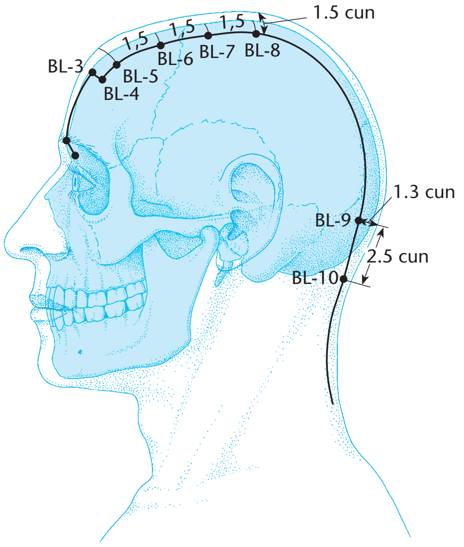
1.5 cun
1.5
1.5
1.5
BL-3
BL-6
BL-7
BL-8
BL-5
BL-4
1.3 cun
BL-9
2.5 cun
BL-10

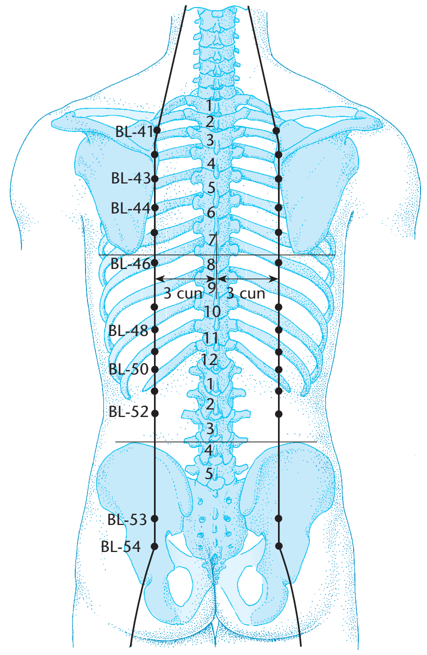
BL-41
1
2
3
BL-43
4
5
BL-44
6
7
BL-46
8
9
3 cun
3 cun
10
BL-48
11
12
BL-50
1
BL-52
2
3
4
5
BL-53
BL-54

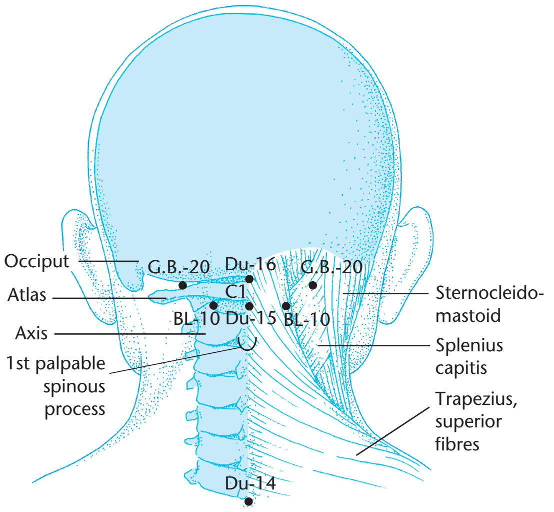
Occiput
G.B.-20
Du-16
G.B.-20
Atlas
C1
Sternocleido-mastoid
BL-10
Du-15
BL-10
Axis
Splenius capitis
1st palpable spinous process
Trapezius, superior fibres
Du-14

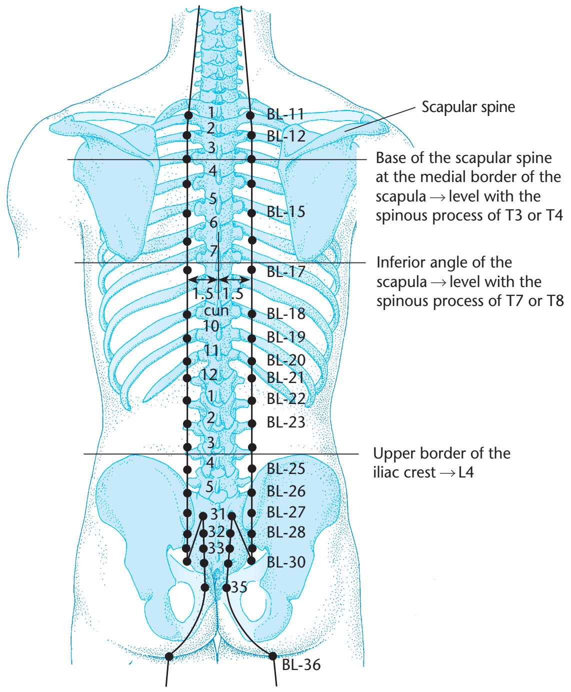
1
2
3
4
5
6
7
BL-11
BL-12
BL-15
BL-17
Scapular spine
Base of the scapular spine at the medial border of the scapula → level with the spinous process of T3 or T4
Inferior angle of the scapula → level with the spinous process of T7 or T8
1.5
1.5
cun
10
11
12
1
2
3
4
5
BL-18
BL-19
BL-20
BL-21
BL-22
BL-23
Upper border of the iliac crest → L4
BL-25
BL-26
BL-27
BL-28
31
32
33
35
BL-30
BL-36

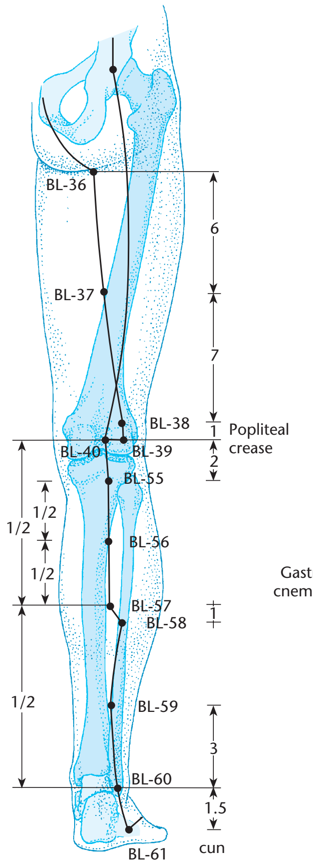
BL-36
6
BL-37
7
BL-38
1
Popliteal crease
BL-40
BL-39
2
BL-55
1/2
1/2
1/2
BL-56
BL-57
BL-58
1
1/2
BL-59
3
BL-60
1.5
cun
BL-61

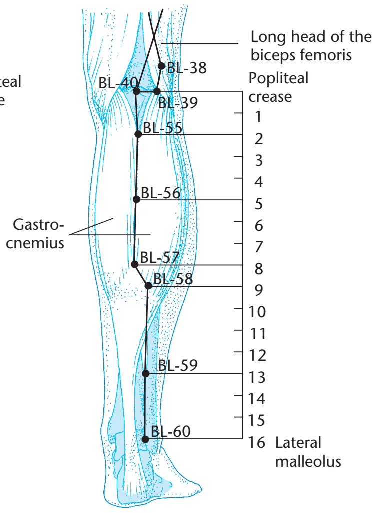
Long head of the biceps femoris
BL-38
BL-40
Popliteal crease
BL-39
BL-55
1
2
3
4
BL-56
5
6
Gastro-cnemius
7
BL-57
8
BL-58
9
10
11
12
BL-59
13
14
15
BL-60
16
Lateral malleolus

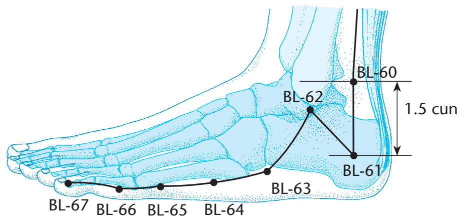
BL-60
BL-62
1.5 cun
BL-61
BL-63
BL-67
BL-66
BL-65
BL-64

At the level of L2, it **enters** the abdomen and reaches its paired *zang*-Organ, the Kidney (*shen*) and its pertaining *fu*-Organ, the Bladder (*pangguang*). The **external** branch runs through the gluteal region and the thigh to the popliteal fossa (**BL-40**), where it re-unites with the lateral external branch.

- the **lateral branch** runs from **BL-10** (*tianzhu*) to **BL-41** (*fufen*) at the level of T2, from where it descends down the back, 3 cun lateral to the midline to the level of the fourth sacral foramen. From there it traverses the gluteal region to → **G.B.-30** (*huantiao*), continues along the posterolateral aspect of the thigh and meets the medial external branch at the popliteal fossa.

From the popliteal fossa (**BL-40**) the channel descends along the posterolateral aspect of the foreleg to the foot

- passes the lateral malleolus
- travels along the lateral border of the foot to the lateral corner of the little toenail to **BL-67** (*zhiyin*). From here, a branch travels to → **KID-1** (*yongquan*) located on the Interiorly–Exteriorly paired Kidney primary channel (foot Yin–Yang pairing of the second great circuit).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Chills and fever, headaches, occipital stiffness, pain in the lumbar region, obstruction of the nasal passages, eye pain with tearing, pain along the posterior aspect of the hip, knee and foreleg, foot pain

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain and tension in the lower abdomen, retention of urine, enuresis, psycho-emotional disorders, opisthotonus

Connections and meeting points of the Bladder primary channel (→ 1.2)

Connections with other channels

Kidney primary channel (*zu shao yin jing*)

Connection: Foot Yin–Yang pairing of the second great circuit
Location: BL-67 → KID-1 (on the foot).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Small Intestine primary channel (*shou tai yang jing*)

Connection: Paired according to the six channel theory (hand–foot pairing): *tai yang* (Yang axes of the second great circuit)
Location: S.I.-18 → BL-1 (on the head).
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with other *zangfu*-Organ systems

Kidney (*shen*), Bladder (*pangguang*)

4.7.2 The Bladder Divergent Channel (*zu tai yang jing bie*)

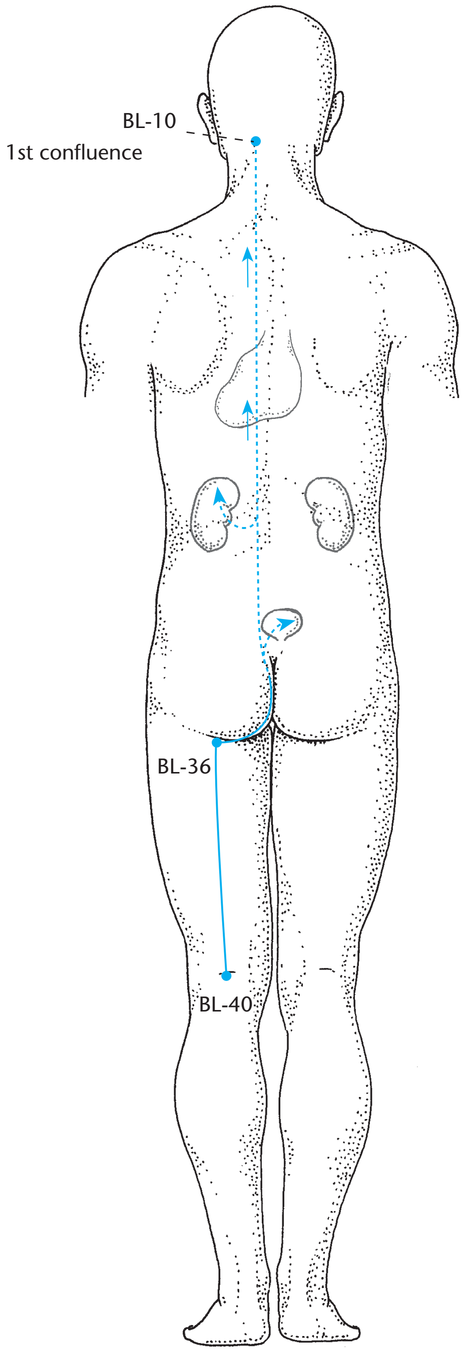

Pathway

The Bladder divergent channel separates from the Bladder primary channel in the popliteal fossa at **BL-40** (*weizhong*)

- ascends to the centre of the gluteal crease at **BL-36** (*chengfu*)
- ascends further and penetrates the anus, continues to the Bladder (*pangguang*), its pertaining *fu*-Organ, and the Kidney (*shen*), its paired *zang*-Organ
- rises along the spine and spreads in the Heart (*xin*)

➡ from here it ascends to the occiput, where it unites at **BL-10** (*tianzhu*) with the Kidney divergent channel and the Bladder primary channel to form one of the six *he*-confluences (here: KID/BL as first confluence → 1.3)

Clinical importance

- Strengthens the relationship between the Bladder and the Kidney (*zangfu*-Organs). Points on the Bladder primary channel can therefore be used for disorders of the Kidney, and vice versa points on the Kidney primary channel can treat disorders of the Bladder.
- Supports the connection between the Bladder primary channel and the Heart
- Supports the connection between the Bladder primary channel and the anus
- Creates a connection between the popliteal region (**BL-40**) and the sacral region (**BL-32**, **BL-31**) as well as with the occipital region (**BL-10**).

4.7.3 The Bladder Sinew Channel (*zu tai yang jing jin*)

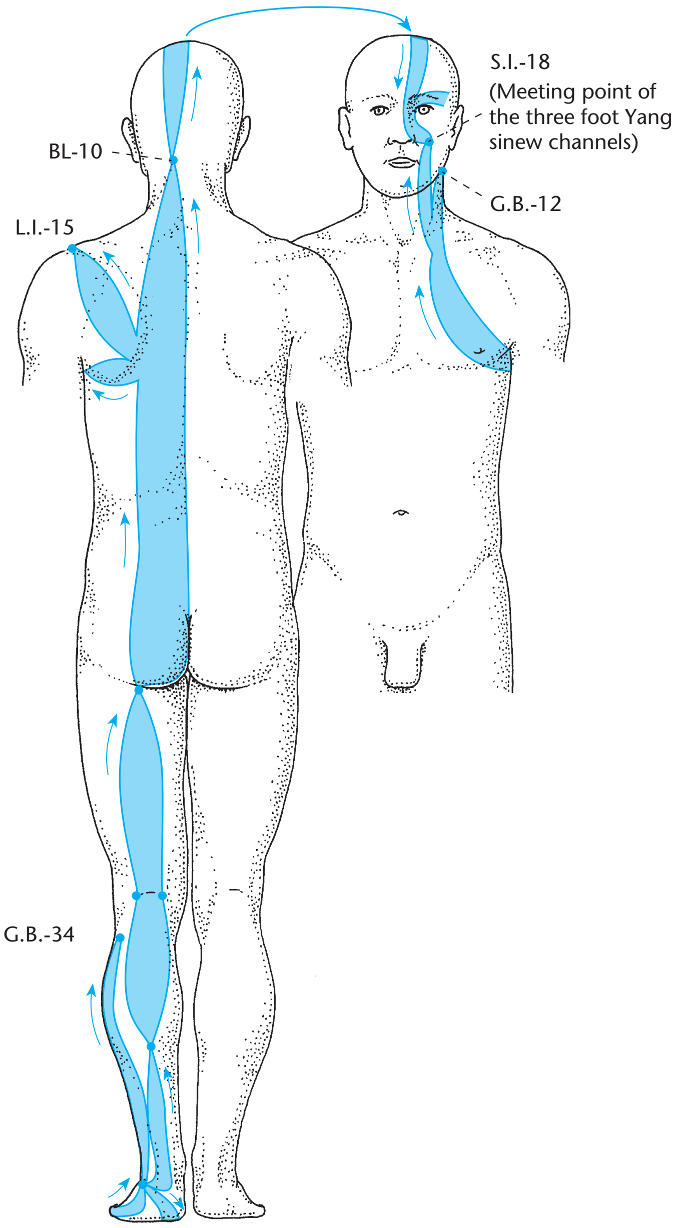

Pathway

The Bladder sinew channel begins on the lateral aspect of the little toe and travels along the lateral border of the foot to the lateral malleolus, where it binds (*jie*).

At the lateral malleolus **three branches** separate from the sinew channel:

➡ **one branch** runs to the calcaneus where it binds (*jie*) and disperses over the heel

➡ **a further branch** ascends the lateral aspect of the lower leg and binds (*jie*) at the head of the fibula, where it also meets the Gall Bladder and Stmach sinew channels at → **G.B.-34** (*yanglingquan*)

➡ **the main branch** travels to the Achilles tendon, binds (*jie*) in the area of the two heads of the gastrocnemius muscle, ascends the posterior aspect of the leg along the gastrocnemius and soleus muscles and binds (*jie*) on either side of the popliteal crease, ascends the posterior aspect of the thigh and binds (*jie*) at the centre of the gluteal crease, ascends as a broad band along the spine and binds (*jie*) at **BL-10** (*tianzhu*).

Two branches separate from the main branch in the thoracic region at the level of the axilla:

- ➡ **one branch** ascends to the shoulder and reaches ➞ **L.I.-15** (*jianyu*)
- ➡ **the other branch** penetrates the axilla, ascends the thorax and runs to the supraclavicular fossa, where again **two sub-branches** separate:
  - **one sub-branch** ascends to the mastoid process to ➞ **G.B.-12** (*wangu*)
  - **the other sub-branch** crosses the cheek to ➞ **S.I.-18** (*quanliao*) where it meets with the other foot Yang sinew channels

The **main branch** divides at **BL-10** (*tianzhu*) into **two branches:**

- ➡ **one branch** penetrates deeper to the root of the tongue
- ➡ **the other branch** crosses the head lateral to the midline, reaches the inner canthus of the eye and separates into **two branches:**
  - **one branch** binds (*jie*) lateral to the nose and descends the cheek
  - **the other branch** crosses the upper eyelid and connects with the Stomach sinew channel, together forming a network around the eye. It then penetrates deeper into the upper orbital region and ends at the roof of the mouth.

Clinical importance

Pathology: Muscular cramping in the region of the little toe. Tension and pain in the heel. Stiffness and limited range of motion of the joints. Tension and stiffness in the back and along the spine, and difficulty when bending forward. Inability to lift the arms above the shoulder region. Stiffness or distending pain in the axillary region. Pain and tense musculature in the supraclavicular fossa. Limited range of motion of the cervical spine.

Indication: Mainly with *bi*-syndromes (painful obstruction syndrome) along the Bladder channel. The area covered by the Bladder sinew channel is larger than that covered by the Bladder primary channel. This explains why Bladder primary channel points can also be used for disorders of the scapula and the shoulder region as well as for disorders of the lower eyelid and the nasal region.

4.7.4 The Bladder *luo*-Connecting Vessel System (*zu tai yang luo mai*)

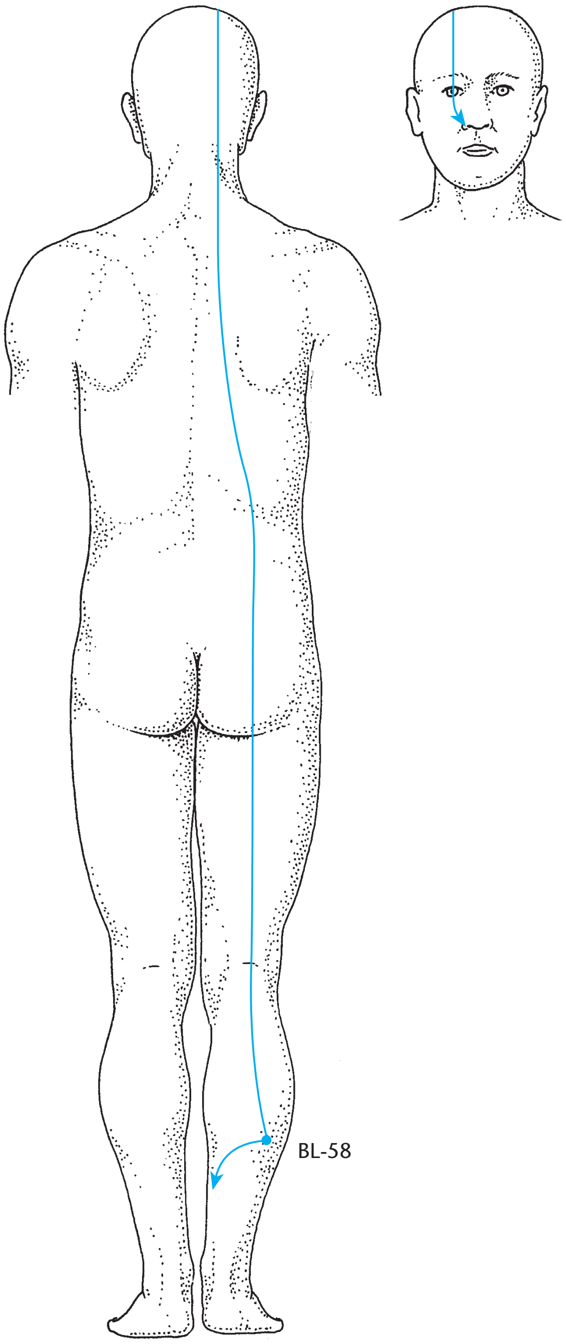

Pathway

The Bladder *luo*-connecting channel separates from the Bladder primary channel at **BL-58** (*feiyang*) (➞ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* ➞ 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Kidney primary channel; according to some schools of thought (for example Nguyen Van Nghi ➞ Appendix) they travel as a **transverse** BL *luo*-connecting vessel to the *yuan*-source point ➞ **KID-3** (*taixi*).

- Solinas et al. (1998, variant → fig.) describe a **longitudinal** division following the pathway of the Bladder primary channel, reaching the head and penetrating the nose. While this division is not mentioned in the classics, according to Solinas et al it becomes evident based on the classic descriptions of the pathologies of the *luo*-connecting vessels.

Clinical importance (→ 8.1.2)

Pathology

- **Excess (*shi*):** Obstruction of the nasal passages, clear watery nasal discharge, headaches, back pain
- **Deficiency (*xu*):** Nosebleeds, chronic clear nasal discharge, sinusitis

4.7.5 Cutaneous Region (*tai yang pi bu*)

See description and figures → 1.6.

4.7.6 Points of the Bladder Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): BL-64** (*jinggu*)
- ***Luo*-connecting point (→ 8.1.2): BL-58** (*feiyang*)
- ***Xi*-cleft point (→ 8.1.3): BL-63** (*jinmen*)
- ***Xi*-cleft point of the *yang qiao mai*: BL-59** (*fuyang*)
- **Associated Back-*shu* point (→ 8.1.4): BL-28** (*pangguangshu*) 🟦🟦
  - Back-*shu* point of the Lung: **BL-13** (*feishu*) 🟦🟦
  - Back-*shu* point of the Pericardium: **BL-14** (*jueyinshu*)
  - Back-*shu* point of the Heart: **BL-15** (*xinshu*) 🟦🟦
  - Back-*shu* point of the *du mai*: **BL-16** (*dushu*)
  - Back-*shu* point of the diaphragm: **BL-17** (*geshu*) 🟦🟦
  - Back-*shu* point of the Liver: **BL-18** (*ganshu*) 🟦🟦
  - Back-*shu* point of the Gall Bladder: **BL-19** (*danshu*) 🟦■
  - Back-*shu* point of the Spleen: **BL-20** (*pishu*) 🟦🟦
  - Back-*shu* point of the Stomach: **BL-21** (*weishu*) 🟦🟦
  - Back-*shu* point of the Triple Burner: **BL-22** (*sanjiaoshu*) 🟦■
  - Back-*shu* point of the Kidneys: **BL-23** (*shenshu*) 🟦🟦
  - Back-*shu* point of the Large Intestine: **BL-25** (*dachangshu*) 🟦■
  - Back-*shu* point of the Small Intestine: **BL-27** (*xiaochangshu*)
  - Back-*shu* point of the Bladder: **BL-28** (*pangguangshu*) 🟦🟦
- **Associated Front-*mu* point (→ 8.1.5): Ren-3** (*zhongji*) 🟦🟦
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal), tonification point: **BL-67** (*zhiyin*) 🟦■
  ying-spring point (Water), *ben* point (Five Phases): **BL-66** (*zutonggu*)
  shu-stream point (Wood), sedation point: **BL-65** (*shugu*)
  jing-river point (Fire): **BL-60** (*kunlun*) 🟦🟦
  he-sea point (Earth): **BL-40** (*weizhong*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7):**
  – of the Blood: **BL-17** (*geshu*) 🟦🟦
  – of the Bones: **BL-11** (*dazhu*) 🟦■
- **Opening point (→ 8.1.8) of the *yang qiao mai*: BL-62** (*shenmai*) 🟦🟦
- **Lower *he*-sea points (→ 8.1.9):**
  of the Triple Burner: **BL-39** (*weiyang*) 🟦■
  of the Bladder: **BL-40** (*weizhong*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the ST, T.B.* channels, the *yin qiao mai*, *yang qiao mai*, S.I., G.B.* channels, *du mai*: **BL-1** (*jingming*) 🟦■
  – with the S.I. channel, (T.B., G.B. channels, *du mai**): **BL-11** (*dazhu*) 🟦■
  – with the *du mai*: **BL-12** (*fengmen*) 🟦■
  – with the *du mai**: **BL-23** (*shenshu*) 🟦🟦
  – with the G.B. channel: **BL-31–BL-34** 🟦■
  – with the S.I. channel: **BL-41** (*fufen*)
  – with the *yang qiao mai*: (**BL-59** (*fuyang*)*), **BL-61** (*pucan*), **BL-62*** (*shenmai*) 🟦🟦
  – with the *yang wei mai*: **BL-63** (*jinmen*)
  – of other channels with the Bladder channel: **Du-24, G.B.-15, Du-20, (G.B.-6*), G.B.-7, G.B.-8, G.B.-9, G.B.-10, G.B.-11, G.B.-12, Du-17, Du-16*, Du-14, Du-13, G.B.-30**
- **Gao Wu command point (→ 8.1.11) for the Back and lumbar region: BL-40** (*weizhong*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): BL-10** (*tianzhu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13):**
  – Sea of Blood point: **BL-11** (*dazhu*) 🟦■
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): BL-40** (*weizhong*) 🟦🟦, **BL-57** (*chengshan*) 🟦■, **BL-60** (*kunlun*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): BL-62** (*shenmai*) 🟦🟦
- **Other functional points:**
  – *xi*-cleft point of the *yang qiao mai*: **BL-59** (*fuyang*)

Points according to region

- **Local points (→ 8.2.1):** occiput – **BL-10** (*tianzhu*) 🟦🟦; eyes – **BL-1** (*jingming*) 🟦■; Stomach/Spleen – **BL-20** (*pishu*) 🟦🟦, **BL-21** (*weishu*) 🟦🟦; Kidney – **BL-23** (*shenshu*) 🟦🟦, **BL-62** (*shenmai*) 🟦🟦; rectum – **BL-35** (*huiyang*); cervical spine – **BL-10** (*tianzhu*) 🟦🟦; thoracic spine – **BL-17** – **BL-23**; lumbar spine – **BL-23** – **BL-26**, **BL-52** (*zhishi*); iliosacral region – **BL-27** (*xiaochangshu*), **BL-28** (*pangguangshu*); hip – **BL-49** (*yishe*), **BL-50** (*weicang*), **BL-54** (*zhibian*); knee – **BL-40** (*weizhong*) 🟦🟦; foot – **BL-60** (*kunlun*) 🟦🟦, **BL-62** (*shenmai*) 🟦🟦
- **Adjacent points (→ 8.2.1):** eyes – **BL-2** (*zanzhu*) 🟦■; nose – **BL-7** (*tongtian*); Lung – **BL-13** (*feishu*) 🟦🟦, **BL-43** (*gaohuang*); Heart – **BL-14** (*jueyinshu*), **BL-15** (*xinshu*) 🟦🟦; Liver – **BL-18** (*ganshu*) 🟦🟦, **BL-20** (*pishu*) 🟦🟦; Gall Bladder – **BL-19** (*danshu*) 🟦■; Large Intestine – **BL-25** (*dachangshu*) 🟦🟦; Small Intestine – **BL-27** (*xiaochangshu*) 🟦■; rectum – **BL-30** (*baihuanshu*), **BL-34** (*xialiao*) 🟦■; Bladder – **BL-28** (*pangguangshu*) 🟦🟦, **BL-23** (*shenshu*) 🟦🟦; urogenital region – **BL-23** (*shenshu*) 🟦🟦, **BL-32** (*ciliao*); cervical spine – **BL-11** (*dazhu*) 🟦■;

---

* Mentioned by only some authors

iliosacral region – **BL-23** (*shenshu*) 🟦🟦; thoracic spine – **BL-38** – **BL-47**; toes – **BL-60** (*kunlun*) 🟦🟦, **BL-65** (*shugu*)

- **Distal points (→ 8.2.1):** occipital region – **BL-65** (*shugu*), **BL-60** (*kunlun*); rectum – **BL-57** (*chengshan*) 🟦🟦; thoracic spine – **BL-57** (*chengshan*) 🟦⬛; cervical, thoracic, lumbar spine – **BL-60** (*kunlun*) 🟦🟦; shoulder and iliosacral region – **BL-58** (*feiyang*); for the thoracic spine, lumbar spine and iliosacral region – **BL-40** (*weizhong*) 🟦🟦; lumbar spine – **BL-40** (*weizhong*) 🟦🟦, **BL-59** (*fuyang*), **BL-60** (*kunlun*) 🟦🟦; lumbar spine and hip – **BL-62** (*shenmai*) 🟦🟦

Specific points according to the channel pathway (in numerical order):

- **BL-1** (*jingming*) 🟦⬛: *jiaohui*-meeting point with the ST, T.B.* channels, the *yin qiao mai, yang qiao mai*, S.I., G.B.* channels, *du mai* (→ 8.1.10): local point for the eyes (→ 8.2.1)
- **BL-2** (*zanzhu*) 🟦⬛: adjacent point for the eyes (→ 8.2.1)
- **BL-7** (*tongtian*): adjacent point for the nose (→ 8.2.1)
- **BL-10** (*tianzhu*) 🟦🟦: Window of Heaven point (→ 8.1.12); important point for eliminating Wind; local point for the occiput and the cervical spine (→ 8.2.1)
- **BL-11** (*dazhu*) 🟦⬛: *hui*-meeting point (→ 8.1.7) of the bones; Sea of Blood point (→ 8.1.13); *jiaohui*-meeting point with the S.I. channel (T.B., G.B. channels, *du mai**) (→ 8.1.10); adjacent point for the cervical spine (→ 8.1.10)
- **BL-12** (*fengmen*) 🟦⬛: *jiaohui*-meeting point with the *du mai* (→ 8.1.10)
- **BL-13** (*feishu*) 🟦🟦: Back-*shu* point of the Lung (→ 8.1.4); adjacent point for the Lung (→ 8.2.1)
- **BL-14** (*jueyinshu*): Back-*shu* point of the Pericardium (→ 8.1.4); adjacent point for the Heart (→ 8.2.1)
- **BL-15** (*xinshu*) 🟦🟦: Back-*shu* point of the Heart (→ 8.1.4); adjacent point for the Heart (→ 8.2.1)
- **BL-16** (*dushu*): Back-*shu* point of the *du mai* (→ 8.1.4);
- **BL-17** (*geshu*) 🟦🟦: Back-*shu* point of the diaphragm (→ 8.1.4); *hui*-meeting point of the Blood (→ 8.1.7)
- **BL-17 to BL-23**: local points for the thoracic spine (→ 8.2.1)
- **BL-18** (*ganshu*) 🟦🟦: Back-*shu* point of the Liver (→ 8.1.4); adjacent point for the Liver (→ 8.2.1)
- **BL-19** (*danshu*) 🟦⬛: Back-*shu* point of the Gall Bladder (→ 8.1.4); adjacent point for the Gall Bladder (→ 8.2.1)
- **BL-20** (*pishu*) 🟦🟦: Back-*shu* point of the Spleen (→ 8.1.4); adjacent point for the Liver (→ 8.2.1)
- **BL-21** (*weishu*) 🟦🟦: local point for the Stomach (→ 8.2.1); Back-*shu* point of the Stomach (→ 8.1.4)
- **BL-22** (*sanjiaoshu*) 🟦⬛: Back-*shu* point of the Triple Burner (→ 8.1.4)
- **BL-23** (*shenshu*) 🟦🟦: *jiaohui*-meeting point with the *du mai** (→ 8.1.10); local point for the Kidneys and lumbar region (→ 8.2.1); adjacent point for the Bladder, urogenital region and iliosacral region (→ 8.2.1); Back-*shu* point of the Kidney (→ 8.1.4)
- **BL-24** (*qihaishu*): local point for the lumbar spine (→ 8.2.1)
- **BL-25** (*dachangshu*) 🟦🟦: Back-*shu* point of the Large Intestine (→ 8.1.4); adjacent point for the lumbar spine (→ 8.2.1)
- **BL-26** (*guanyuanshu*): local point for the lumbar spine (→ 8.2.1)
- **BL-27** (*xiaochangshu*) 🟦⬛: Back-*shu* point of the Small Intestine (→ 8.1.4); adjacent point for the Small Intestine (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-28** (*pangguangshu*) 🟦🟦: Back-*shu* point of the Bladder (→ 8.1.4); adjacent point for the Bladder (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-30** (*baihuanshu*): regional point for the rectum (→ 8.2.1)
- **BL-31** – **BL-34**: *jiaohui*-meeting points with the G.B. channel (→ 8.1.10)
- **BL-32** (*ciliao*): regional point for the urogenital region (→ 8.2.1); local point for the iliosacral region (→ 8.2.1)
- **BL-34** (*xiaoliao*): regional point for the rectum (→ 8.2.1)
- **BL-35** (*huiyang*): local point for the rectum (→ 8.2.1)
- **BL-38** – **BL-47**: adjacent points for the thoracic spine (→ 8.2.1)
- **BL-39** (*weiyang*) 🟦⬛: lower *he*-sea point of the Triple Burner (→ 8.1.9)
- **BL-40** (*weizhong*) 🟦🟦: *he*-sea point (Earth); lower *he*-sea point of the Bladder (→ 8.1.9); Gao Wu command point (→ 8.1.11) of the back and lumbar region; Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the thoracic and lumbar spine as well as for the iliosacral region (→ 8.2.1); local point for the knee (→ 8.2.1)
- **BL-41** (*fufen*): *jiaohui*-meeting point with the S.I. channel (→ 8.1.10)
- **BL-43** (*gaohuang*): adjacent point for the Lung (→ 8.2.1)
- **BL-49** (*yishe*): local point for the hip (→ 8.2.1)
- **BL-50** (*weicang*): local point for the hip (→ 8.2.1)
- **BL-52** (*zhishi*): local point for the lumbar region (→ 8.2.1)
- **BL-54** (*zhibian*): local point for the hip (→ 8.2.1)
- **BL-57** (*chengshan*) 🟦⬛: Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the rectum (→ 8.2.1); distal point for the thoracic spine (→ 8.2.1)
- **BL-58** (*feiyang*): *luo*-connecting point (→ 8.1.2); distal point for the shoulder and iliosacral region (→ 8.2.1)
- **BL-59** (*fuyang*): *xi*-cleft point of the *yang qiao mai; jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)*; distal point for the lumbar spine (→ 8.2.1)
- **BL-60** (*kunlun*) 🟦🟦: *jing*-river point (Fire); Ma Dan Yang Heavenly Star point (→ 8.1.14); distal point for the cervical, thoracic and lumbar spine (→ 8.2.1); distal point for the occipital and lumbar region (→ 8.2.1); local point for the foot (→ 8.2.1); regional point for the toes (→ 8.2.1)
- **BL-61** (*pucan*): *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10)
- **BL-62** (*shenmai*) 🟦🟦:Opening point of the *yang qiao mai* (→ 8.1.8); Sun Si Miao Ghost point (→ 8.1.15); *jiaohui*-meeting point with the *yang qiao mai* (→ 8.1.10); local point

---

* Mentioned by only some authors

for the Kidney (→ 8.2.1); local point for the foot (→ 8.2.1); distal point for the lumbar spine and hip (→ 8.2.1)
- **BL-63** (*jinmen*): *xi*-cleft point (→ 8.1.3); *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **BL-64** (*jinggu*): *yuan*-source point (→ 8.1.1)
- **BL-65** (*shugu*): *shu*-stream point (Wood); sedation point; distal point for the occiput (→ 8.2.1); adjacent point for the toes (→ 8.2.1)
- **BL-66** (*zutongu*): *ying*-spring point (Water); *ben* point (Five Phases point)
- **BL-67** (*zhiyin*) ■■: *jing*-well point; tonification point

General location help

- **The medial branch of the Bladder channel** runs from **BL-11** (below the spinal process of the T1) to **BL-30** (level with the fourth sacral foramen), 1.5 cun lateral to the midline

Memorisation help for the medial branch of the BL channel:
- **BL11 to BL-17**: the second digit of the point number corresponds to the level of the respective thoracic vertebra. Example: **BL-13** is located below the spinal process of T3.
- **BL-18 to BL-21**: second digit of the point number + 1. *Example*: **BL-19** (9 + 1) is located below the spinal process of the T10.

The outer branch of the Bladder channel runs from **BL-41** (below the spinal process of the T2) to **BL-54** (below the spinal process of S4/level with the 4th sacral foramen) 3 cun lateral to the midline.

Memorisation help for the lateral branch of the BL channel:
- **BL41 to BL-46**: second digit of the point number + 1. *Example*: **BL-42** (2 + 1) corresponds to T3
- **BL-47 to BL-49**: second digit of the point number + 2. *Example*: **BL-47** (7 + 2) corresponds to the level of T9.
- **BL-50**: second digit of the point number + 12.

Needling: The Back-*shu* points are also commonly needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape).

Bright Eyes JINGMING BL-1

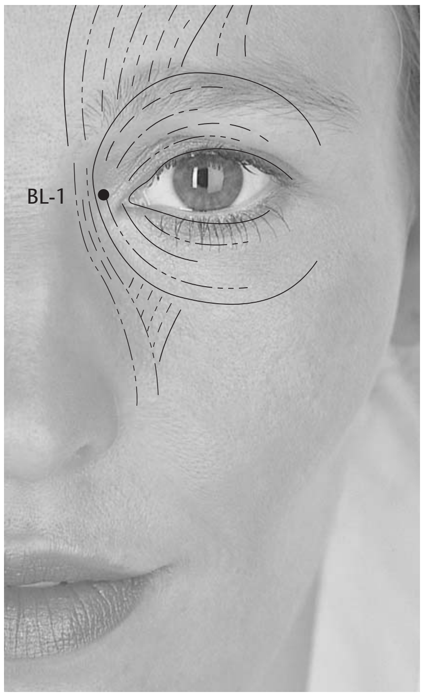

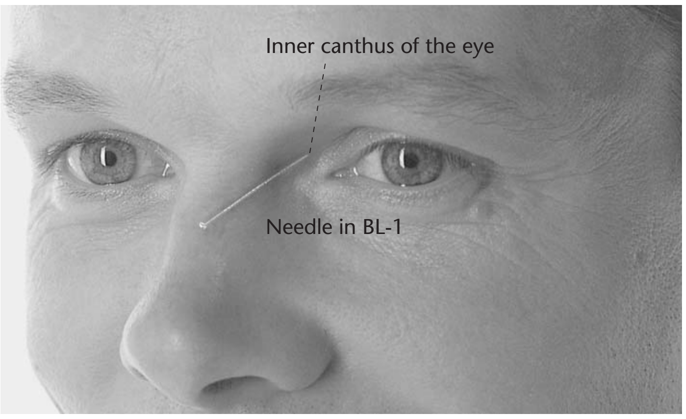
Inner canthus of the eye
Needle in BL-1

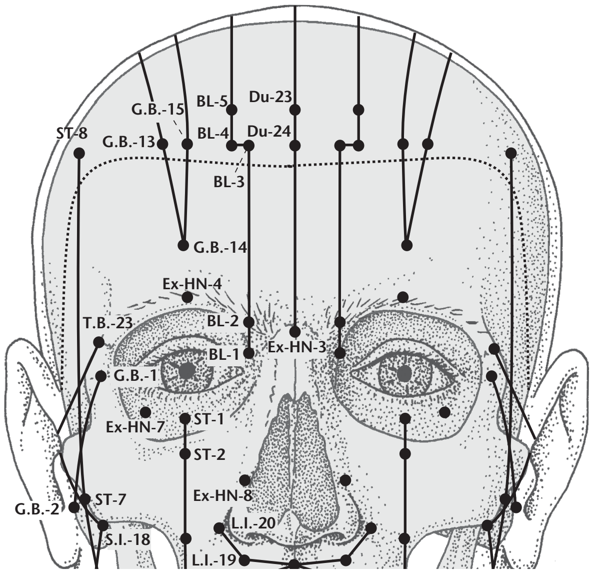
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19

Location

0.1 cun superior and medial to the inner canthus of the eye, in a depression.

How to find

BL-1 is located in a small depression near the inner canthus of the eye, superior to the insertion of the upper eyelid. Medially, it is bordered by the bony root of the nose. ➞ **BL-2** is located directly above **BL-1** in a depression at the medial end of the eyebrow. ➞ **ST-1** is located on the pupil line between the eyeball and the border of the infraorbital ridge, ➞ **G.B.-1** and ➞ **T.B.-23** are located on the outer canthus of the eye.

Needling

Ask the patient to look downward and outward, with their eyes closed. Gently push the eyeball downward and outward from above. Needle perpendicularly along the supraorbital ridge in a posterior direction. Caution: Angular artery and vein, eyeball, periosteum. This point should only be needled by experienced practitioners. After removing the needle, press on the insertion site for 10 minutes. Despite this precaution, a bruise may still form (inform the patient). **No** needle stimulation! Moxibustion is contraindicated. Alternative point: ➞ **BL-2** or transverse (subcutaneous) insertion at **BL-1** in the direction of ➞ **BL-2**.

Actions/Indications

- Expels Wind and clears Heat, benefits the eyes

Special features

Meeting point with the *du mai*, *yin qiao mai*, *yang qiao mai* as well as the ST, S.I., G.B. channels and, according to some authors, also the T.B. channel. Entry point.

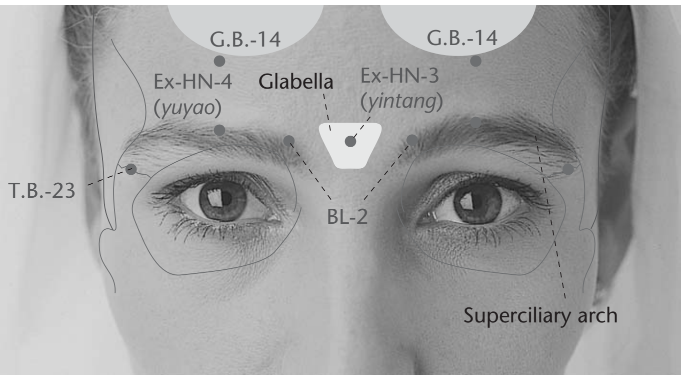
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2
Superciliary arch

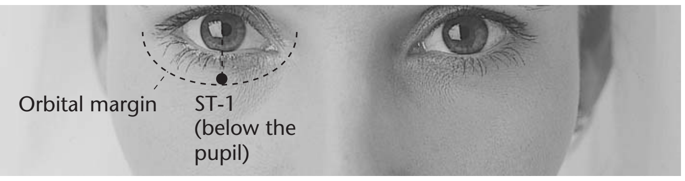
Orbital margin
ST-1
(below the
pupil)

BL-2 *Gathered Bamboo* ZANZHU

Location

In a depression at the medial end of the eyebrow, directly above the inner canthus of the eye.

How to find

As the location of the eyebrow can vary, the medial canthus of the eye should be used for reference (➞ **BL-1**). Palpate for a generally pressure-sensitive depression in this area of the orbital ridge. Caution: According to Sobotta, except in rare cases, **BL-2** is not located on the supraorbital foramen (mainly located more laterally), but in the area where the supratrochlear artery and the supraorbital nerve emerge (frontal notch).

➞ **BL-1** is located inferior to **BL-2**.

Needling

0.3–0.5 cun obliquely from lateral to medial or transversely (subcutaneously) in the direction of the disorder. Caution: Branches of the frontal nerve, facial nerve, blood vessels! According to the majority of authors, moxibustion is contraindicted. Considered a less risky alternative to ➞ **BL-1**.

Actions/Indications

- Benefits the eyes
- Expels external pathogenic factors such as Wind, clears Heat
- Clears the head
- Moves Qi in the Bladder divergent channel (for example, for painful haemorrhoids)

Special features

Important and frequently used local point for the eye region. Good point for excess and Heat syndromes of the eye region: prick to bleed (puncture the point and let it bleed), but avoid facial haematoma.

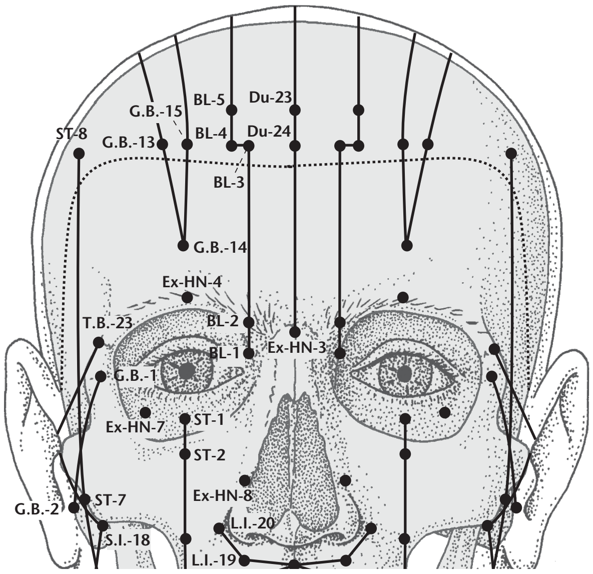

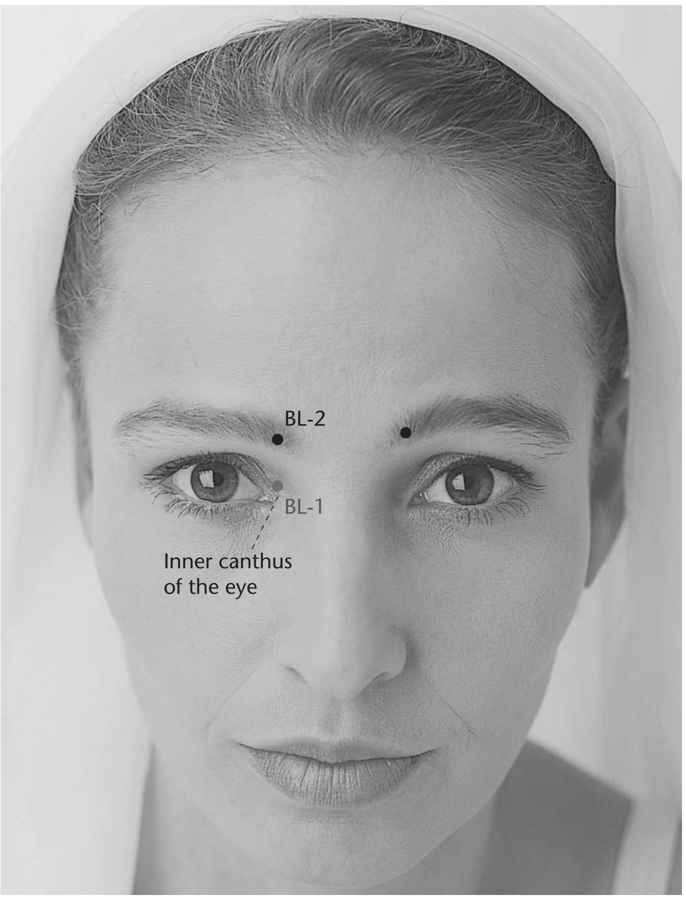

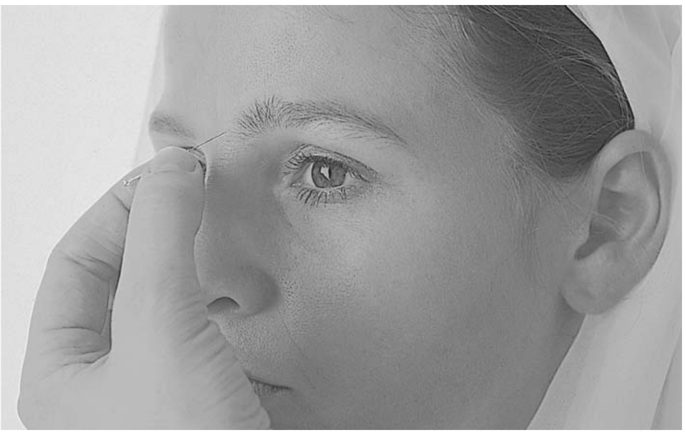

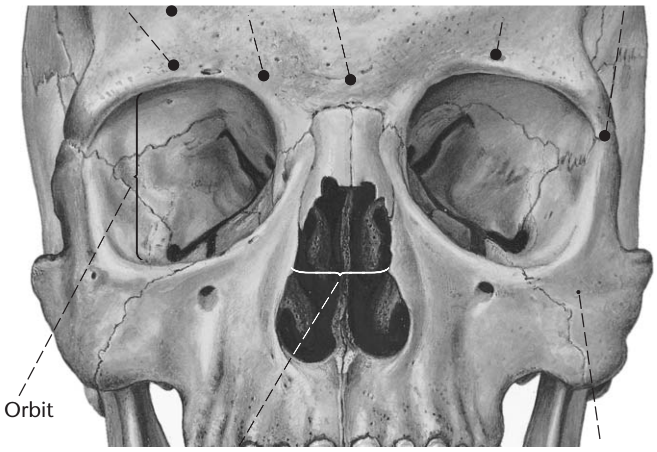

Eyebrows' Pouring MEICHONG BL-3

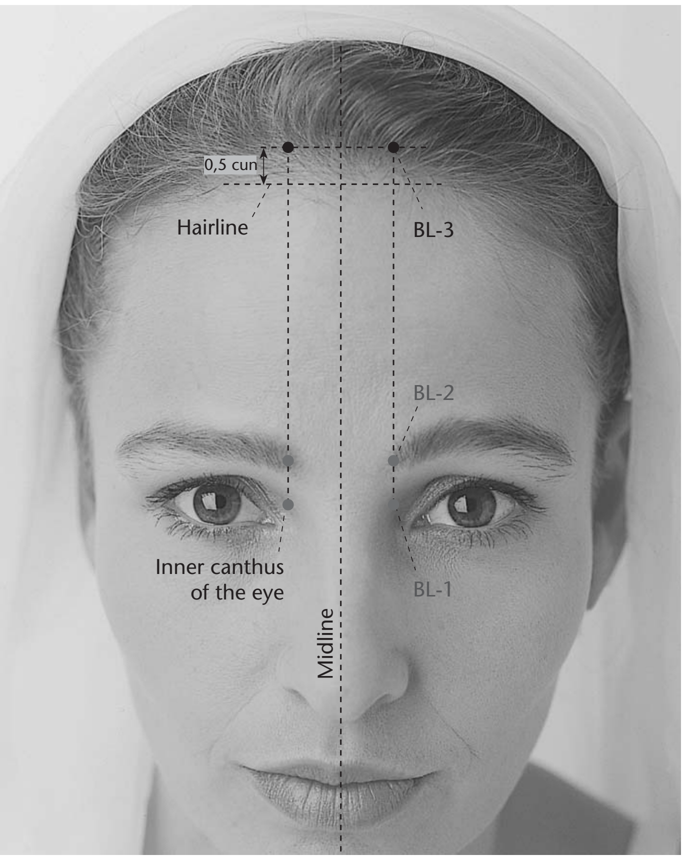
0,5 cun
Hairline
BL-3
BL-2
Inner canthus of the eye
BL-1
Midline

Location

0.5 cun within the anterior hairline, vertically above the medial canthus of the eye.

How to find

BL-3 is located directly above the medial canthus of the eye (➞ **BL-1**) and 0.5 cun within the anterior hairline (➞ 3.1.1). Located on the same level (0.5 cun within the anterior hairline) are ➞ **Du-24** (on the midline), ➞ **G.B.-15** (on the pupil line), ➞ **G.B.-13** (3 cun lateral to the midline) and ➞ **ST-8** (on the corner of the forehead).

Needling

0.5–1 cun obliquely in a dorsal direction or transversely (subcutaneously). According to some classic texts, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and clears Heat from the head and forehead

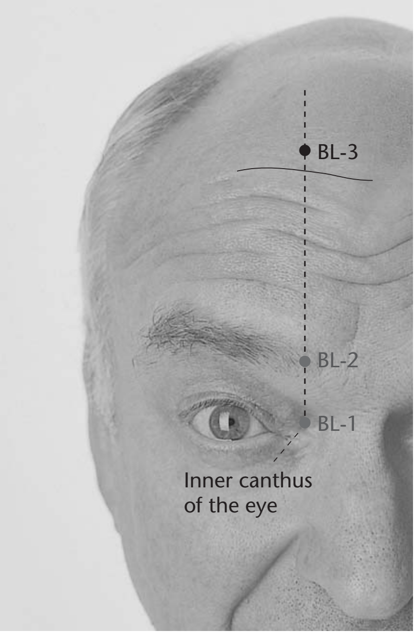
BL-3
BL-2
BL-1
Inner canthus of the eye

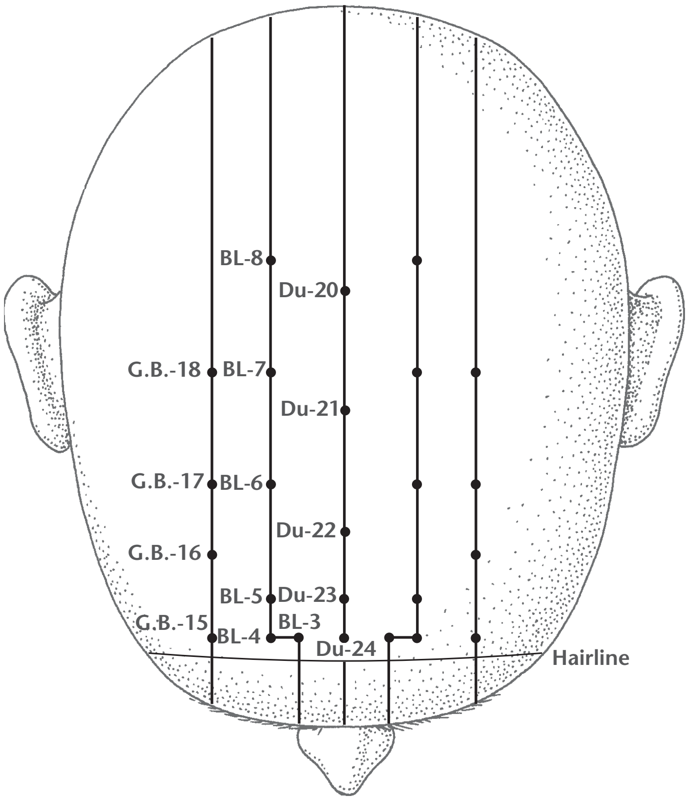
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
G.B.-15
BL-4
BL-3
Du-24
Hairline

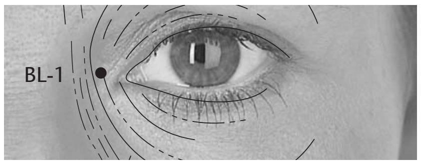

BL-4 *Crooked Curve* QUCHA

Location

0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline (or one third of the distance between → **Du-24** and → **ST-8**).

How to find

While → **BL-3** is still located on a vertical line through the inner canthus of the eye, **BL-4** is located slightly more laterally (hence its name). The distance of 1.5 cun from the midline refers to the proportional cun measurements along the anterior hairline (→ 3.1.1): the distance between → **Du-24** (on the midline) and → **ST-8** (on the corner of the forehead) is 4.5 proportional cun; therefore, the distance between **BL-4** and the midline is one third of this distance.

Located on the same level (0.5 cun within the anterior hairline) are → **Du-24** (on the midline), → **BL-3** (superior to the inner canthus of the eye), → **G.B.-15** (on the pupil line), → **G.B.-13** (3 cun lateral to the midline) and → **ST-8** (on the corner of the forehead).

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Expels Wind and Wind-Heat, especially from the eyes and the nose
- Clears the head, opens the orifices, alleviates pain

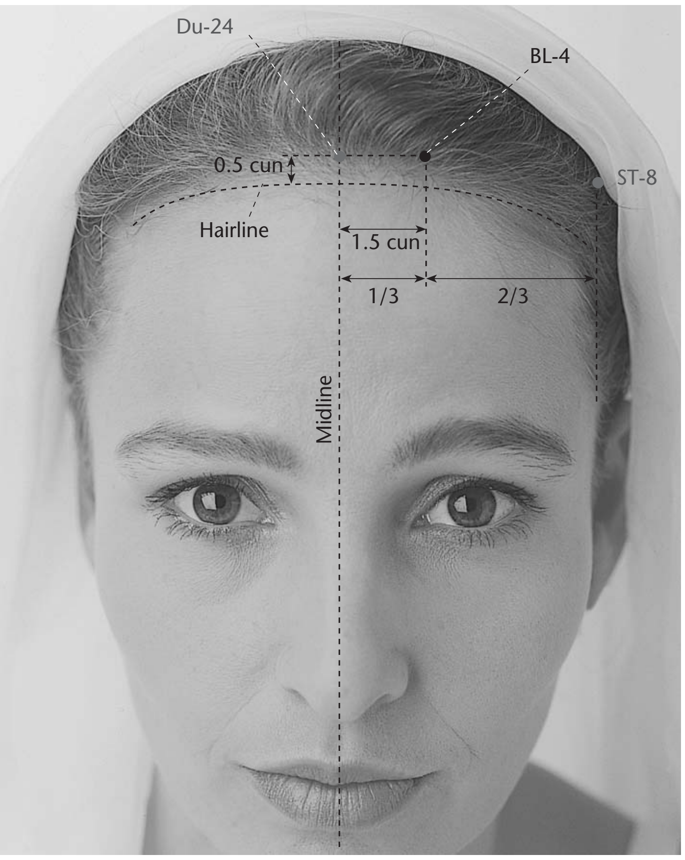
Du-24
BL-4
0.5 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

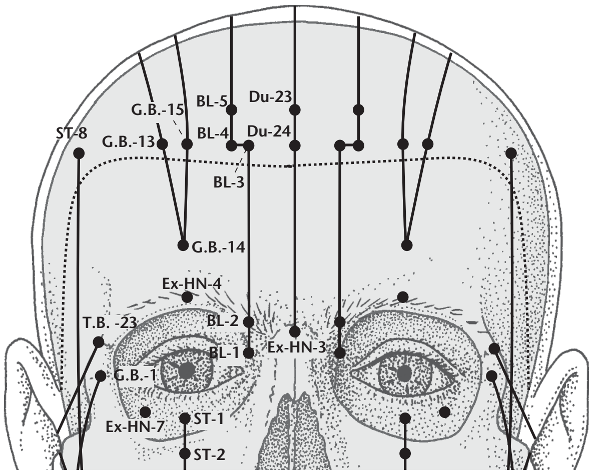
BL-5
Du-23
G.B.-15
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B. -23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2

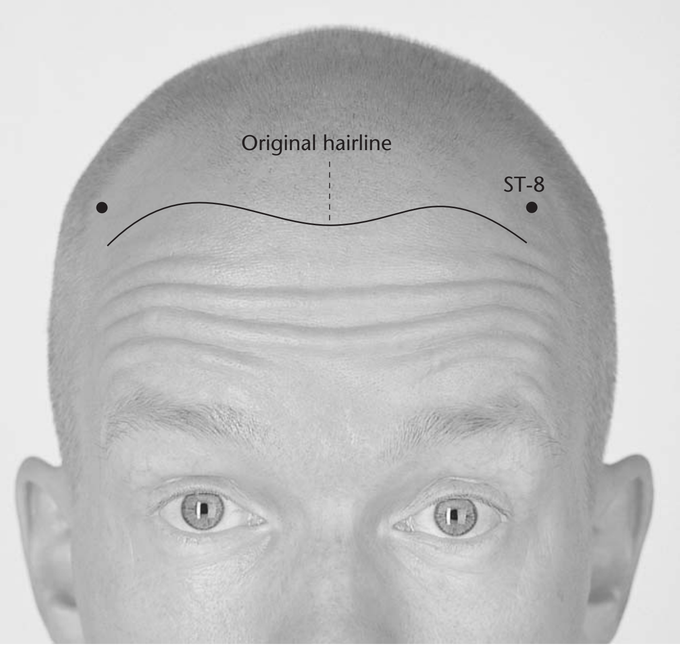
Original hairline
ST-8

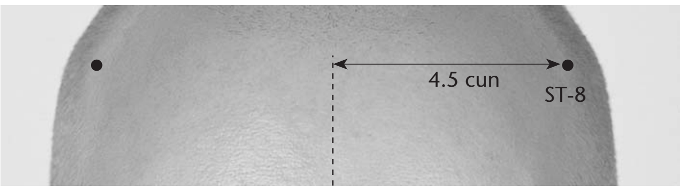

Fifth Place WUCHU BL-5

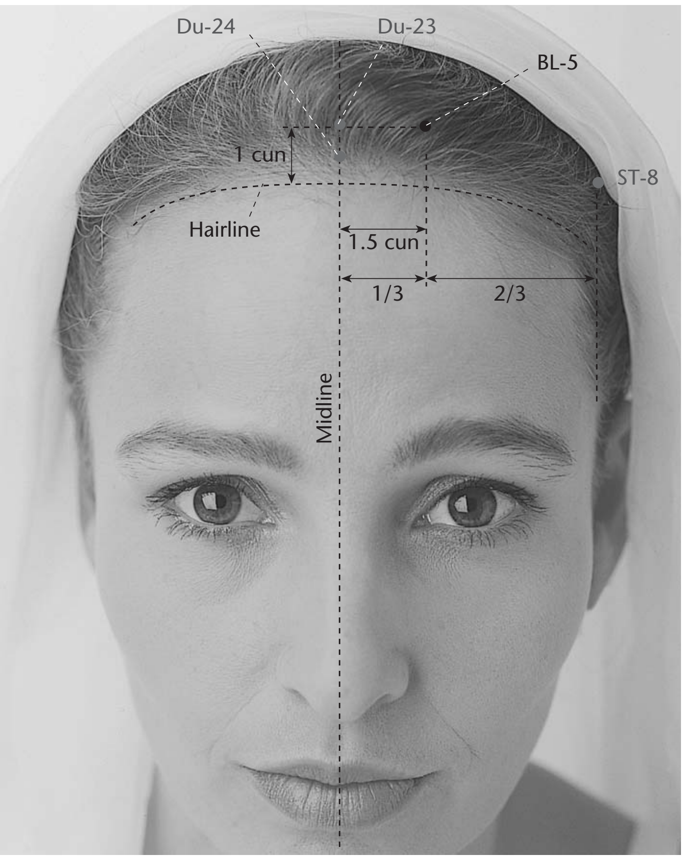
Du-24
Du-23
BL-5
1 cun
ST-8
Hairline
1.5 cun
1/3
2/3
Midline

Location

1 cun superior to the anterior hairline and 1.5 cun lateral to the midline or one third of the distance between → **Du-24** and → **ST-8**.

How to find

The distance of 1.5 cun refers to the proportional cun measurement for the distance of 4.5 cun between → **Du-24** (midline, 0.5 cun superior to the anterior hairline) and → **ST-8** (corner of the forehead), (→ 2.2). By dividing this distance into thirds, **BL-5** can be located one third of the distance from → **Du-24** and 1 cun superior to the anterior hairline.
Located on the same level is → **Du-23** (on the midline).

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion is contraindicated according to some texts.

Actions/Indications

- Releases the Exterior, expels Wind and Heat from the head and eyes
- Opens the channel and the *luo* vessels
- Opens the orifices (especially the eyes) and revives consciousness

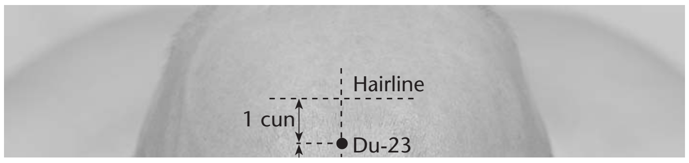
Hairline
1 cun
Du-23

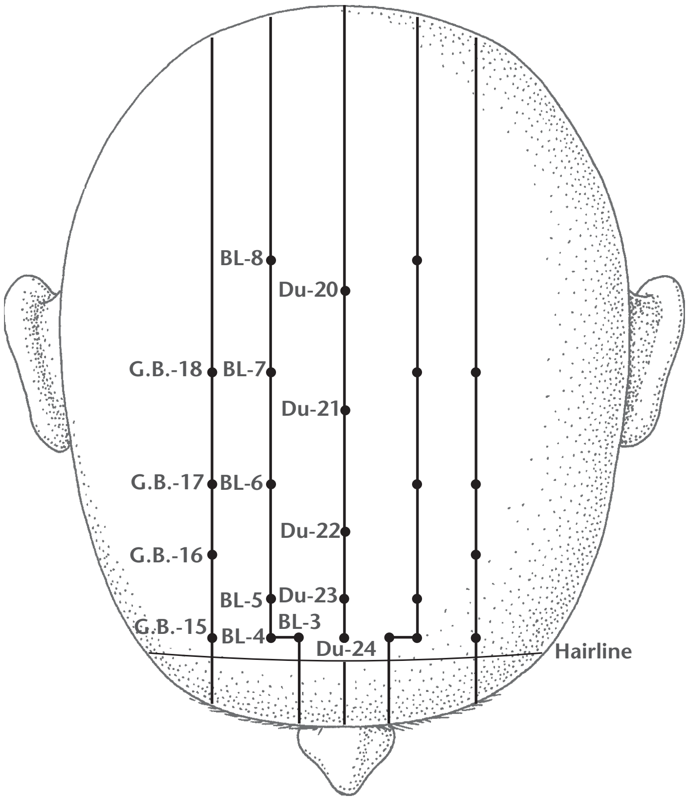
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
G.B.-15
BL-4
BL-3
Du-24
Hairline

BL-6 *Receiving Light* CHENGGUAN

Location

2.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline or one third of the distance between → **Du-24** and → **ST-8**.

How to find

Spreading hands technique (→ 2.3.3): First, locate the anterior midline at the level of the anterior hairline (→ 3.1.1) as well as → **Du-20** (5 cun superior to the anterior hairline, on the vertex, at the junction of the midline with a line connecting the apices of the ears). Next, determine the midpoint between those two reference points (→ 2.5 cun superior to the anterior hairline). From this point, measure 1.5 cun laterally (or one third of the distance between → **Du-24** and → **ST-8**) to locate **BL-6**.
Located on the same level (→ 2.5 cun superior to the anterior hairline), but more laterally (pupil line), is → **G.B.-17**.

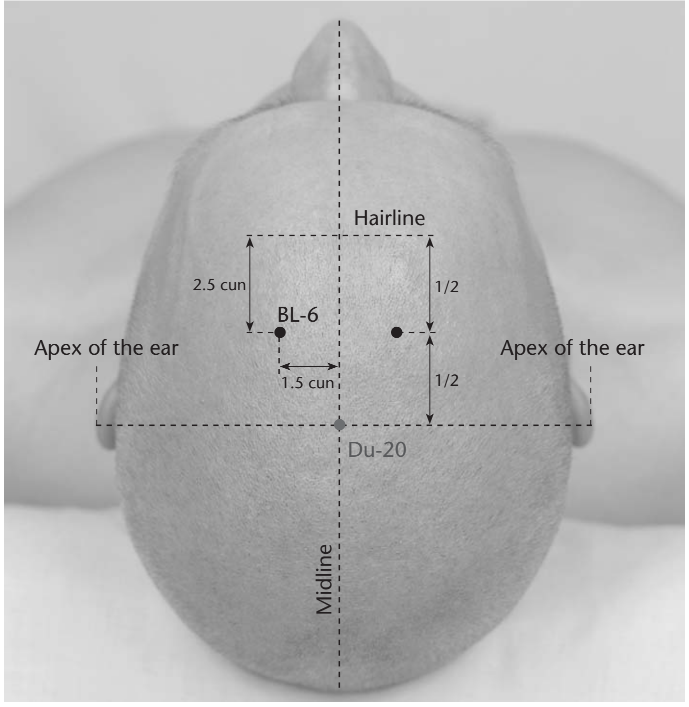
Hairline
2.5 cun
BL-6
1/2
Apex of the ear
Apex of the ear
1.5 cun
1/2
Du-20
Midline

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion only following careful diagnosis.

Actions/Indications

- Expels Wind, especially from the head and eyes
- Clears Heat from the head, especially from the eyes and nose

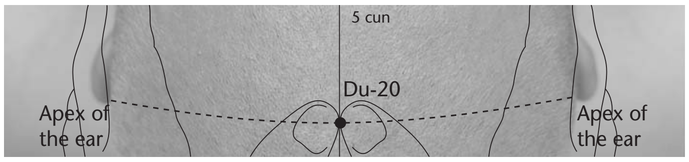
5 cun
Du-20
Apex of the ear
Apex of the ear

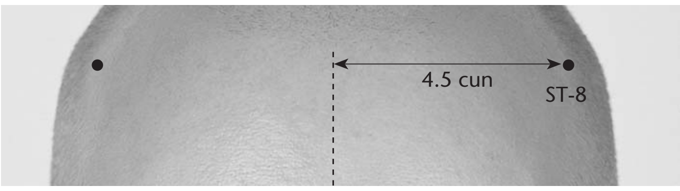

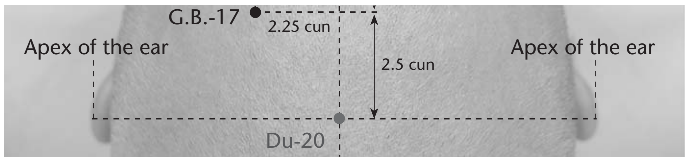
G.B.-17
2.25 cun
Apex of the ear
Apex of the ear
2.5 cun
Du-20

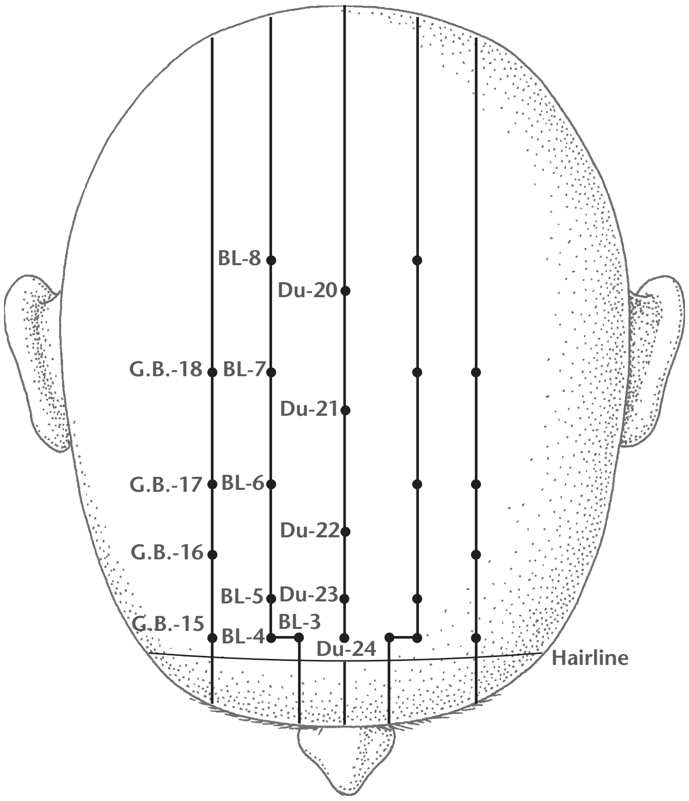
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Heavenly Connection TONGTIAN **BL-7**

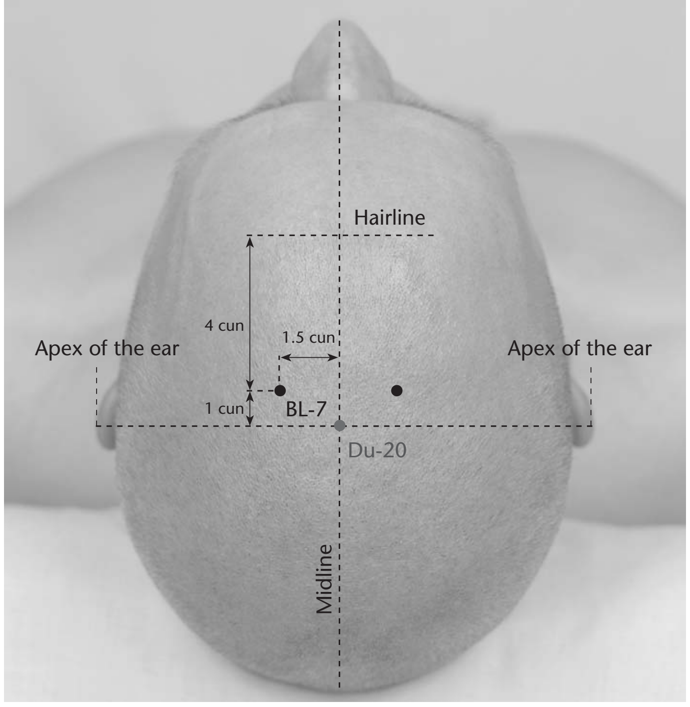
Hairline
4 cun
1.5 cun
Apex of the ear
Apex of the ear
1 cun
BL-7
Du-20
Midline

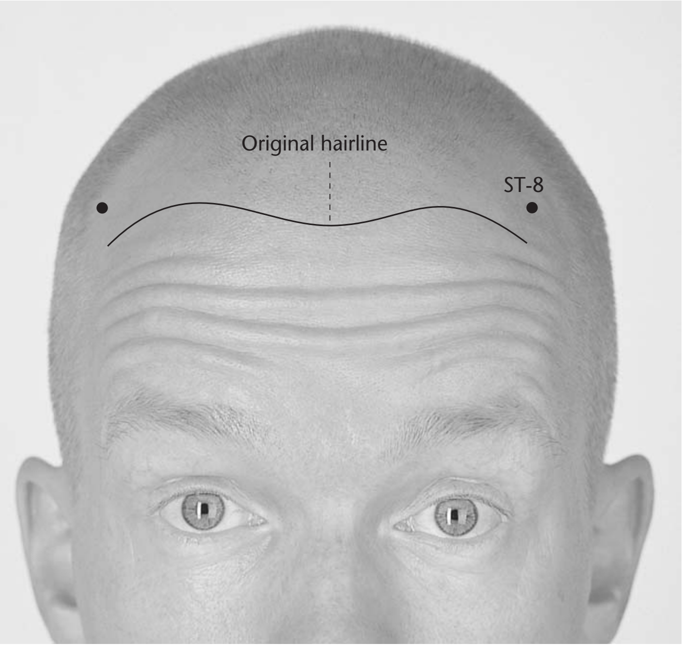
Original hairline
ST-8

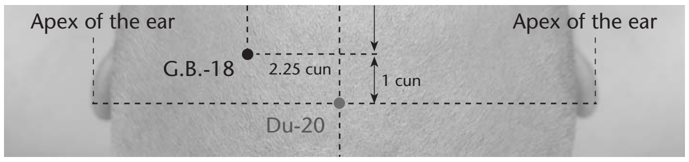
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

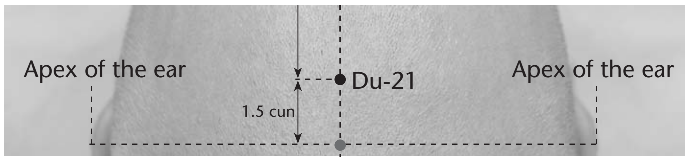
Apex of the ear
Du-21
Apex of the ear
1.5 cun

Location

1.5 cun lateral to the midline and 4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20**.

How to find

BL-7 is best located by using → **Du-20** (5 cun superior to the anterior hairline, on the vertex, at the junction of the midline with a line connecting the apices of the ears) as a reference point. Measure 1 cun anterior from → **Du-20** and 1.5 cun lateral to the midline (or one third of the distance between → **Du-24** and **ST-8,** → 2.2) to locate **BL-7**. Located on the same level (4 cun superior to the anterior hairline), but more laterally (pupil line), is → **G.B.-18**. → **Du-21** is located on the midline, 1.5 cun anterior to → **Du-20**.

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Clears the head, benefits especially the nose

Special features

Important point for treating disorders of the nose

BL-8 *Declining Connection* LUOQUE

Location

1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline or 0.5 cun posterior to → **Du-20**.

How to find

BL-8 is best located by using → **Du-20** as a reference point. → **Du-20** is located at the junction of the vertical midline with a line connecting the apices of the ears, 5 cun superior to the anterior hairline (→ 3.1.1). To locate **BL-8**, measure 0.5 cun from → **Du-20** in a posterior direction and 1.5 cun in a lateral direction.

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion possible.

Actions/Indications

- Opens the orifices
- Subdues Wind
- Transforms Phlegm
- Calms the *shen*

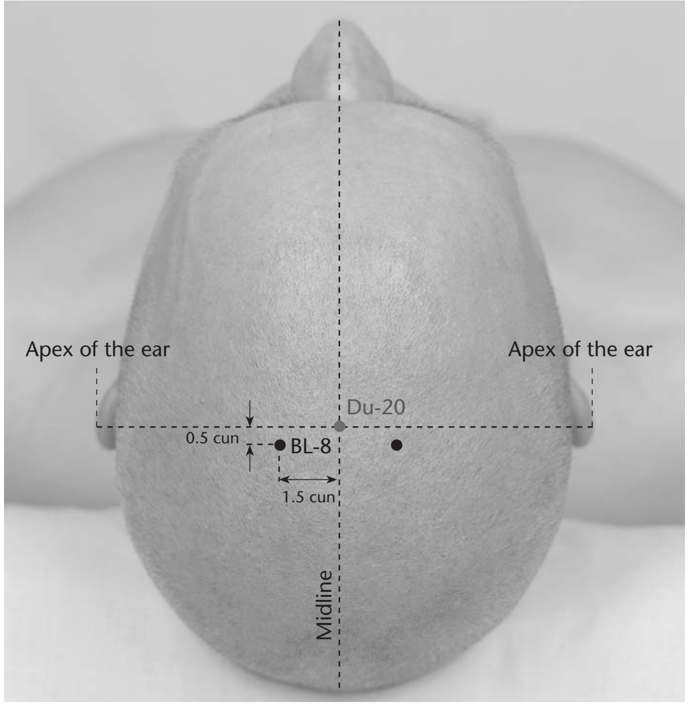
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8
1.5 cun
Midline

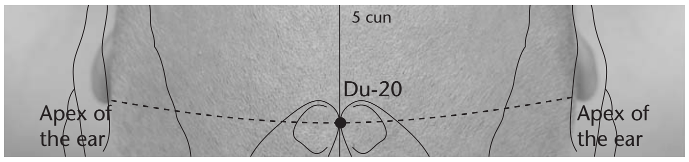
5 cun
Du-20
Apex of the ear
Apex of the ear

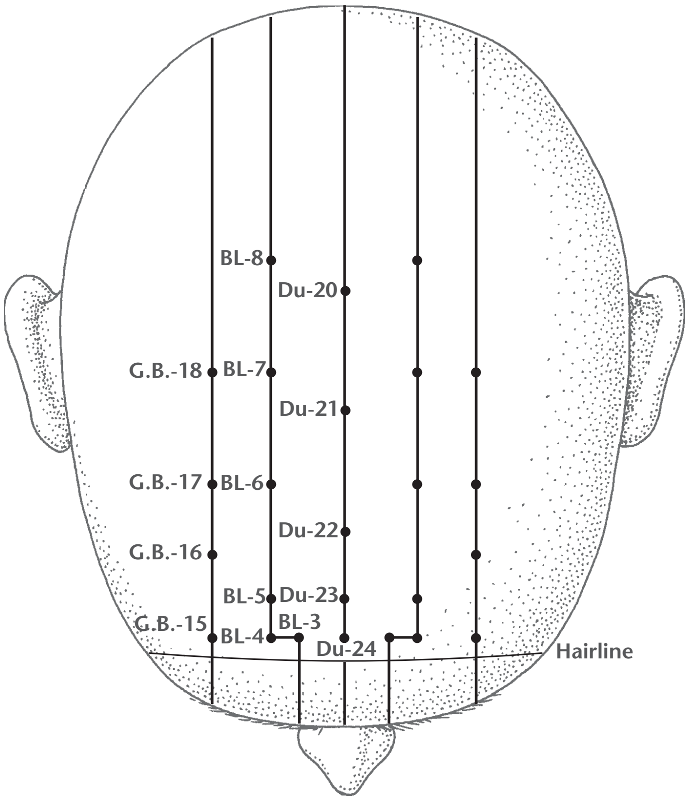
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Ch04.7a-F10028.qxd 2/23/08 3:53 PM Page 259

Jade Pillow YUZHEN BL-9

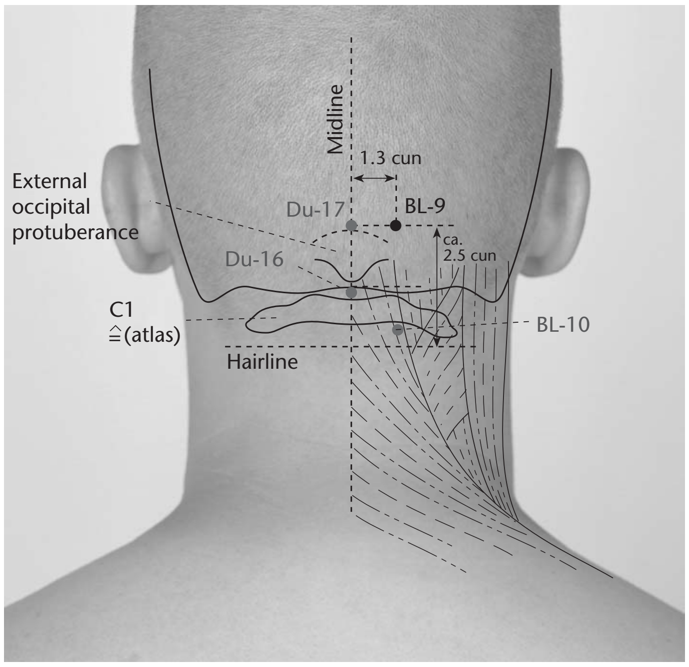
Midline
1.3 cun
External occipital protuberance
Du-17
BL-9
ca. 2.5 cun
Du-16
C1 ≙(atlas)
BL-10
Hairline

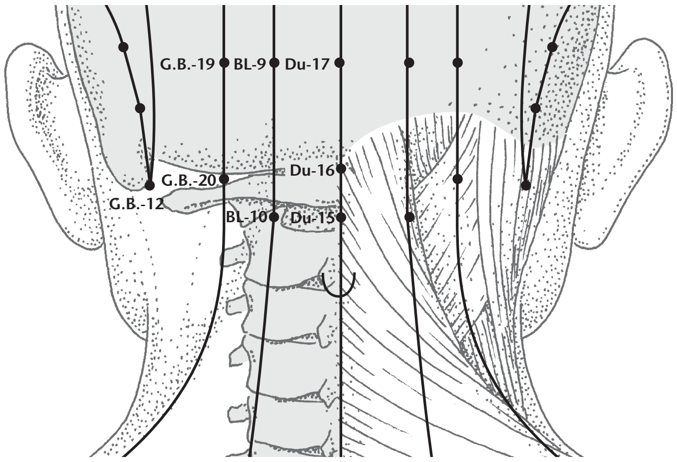
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

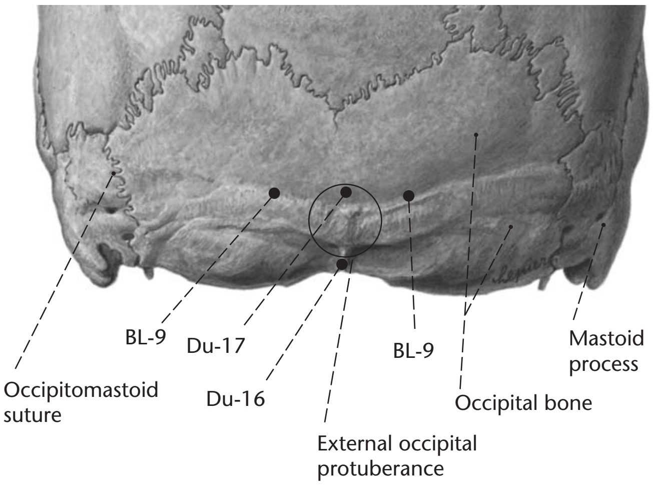
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

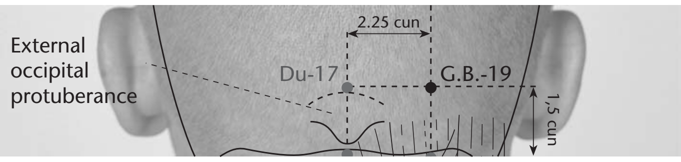
External occipital protuberance
2.25 cun
Du-17
G.B.-19
1.5 cun

Location

2.5 cun superior to the posterior hairline and 1.3 cun lateral to the midline or → **Du-17** (directly superior to the external occipital protuberance).

How to find

To locate **BL-9,** → **Du-17** is a good reference point, located on the midline, in the depression directly superior to the upper border of the external occipital protuberance (→ 3.1.5). From there, measure 1.3 cun in a lateral direction to locate **BL-9**, which is generally located on an imaginary vertical line through → **BL-10.**
Or: Locate the posterior hairline (often not very accurate → 3.1.5) and, on the midline, measure 2.5 cun in a superior direction, then 1.3 cun towards lateral.
Located on the same level are → **Du-17** (midline, in the depression at the upper border of the external occipital protuberance) and → **G.B.-19** (more laterally, on a vertical line through → **G.B.-20**).

Needling

Transversely (subcutaneously) 0.5–1 cun. Moxibustion possible.

Actions/Indications

- Expels Wind and Cold
- Opens the channel and *luo* vessels, alleviates pain
- Benefits the eyes and nose

Special features

In Qigong practice, **BL-9** plays an important role as one of the three gates (*san guan*).

BL-10 *Heavenly Pillar* TIANZHU

Location

Approximately 1.3 cun lateral to ➞ **Du-15**, where the trapezius muscle inserts on the lower border of the occiput, close to where the major occipital nerve emerges.

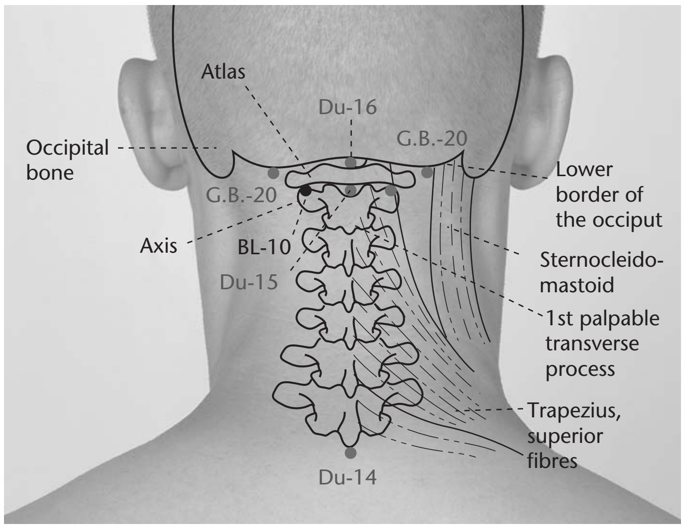
Atlas
Du-16
G.B.-20
Occipital bone
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable transverse process
Trapezius, superior fibres
Du-14

How to find

With the patient's head in an upright and relaxed position, first locate ➞ **G.B.-20** below the lower border of the occiput and between the insertions of the sternocleidomastoid and trapezius muscles. From there, palpate in a medial and inferior direction at a 45° angle, until the finger comes to a halt against the bulge of the trapezius muscle. This is the location of **BL-10**. For reference: If the practitioner places their left middle finger on ➞ **G.B.-20**, the (shorter) index finger will be pointing to **BL-10**. Located on the same level is ➞ **Du-15** (approximately 1.3 cun lateral to the midline and superior to the axis (first palpable spinous process)).

Needling

0.5–1 cun vertically or slightly obliquely in a dorsal direction. Caution: Do not needle in a superior direction.

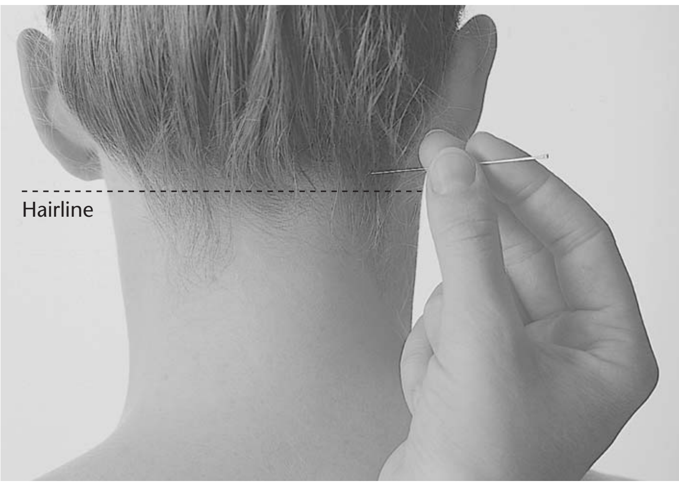

Actions/Indications

- Regulates the Qi, calms Wind and the *shen*, benefits the head and sensory orifices
- Opens the channel
- Strengthens the lower back
- Expels Wind

Special features

Window of Heaven point. Important point for expelling (internal and external) Wind. In contrast to ➞ **G.B.-20**, **BL-10** is used in clinical practice mainly for pacifiying internal Wind.

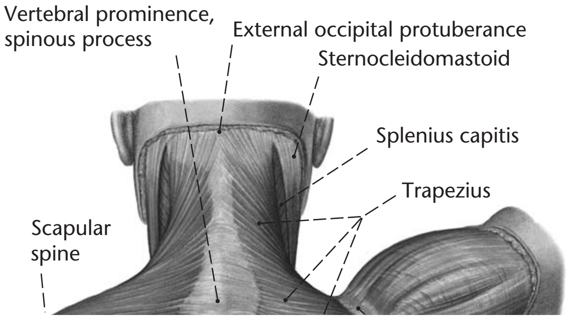
Vertebral prominence, spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular spine

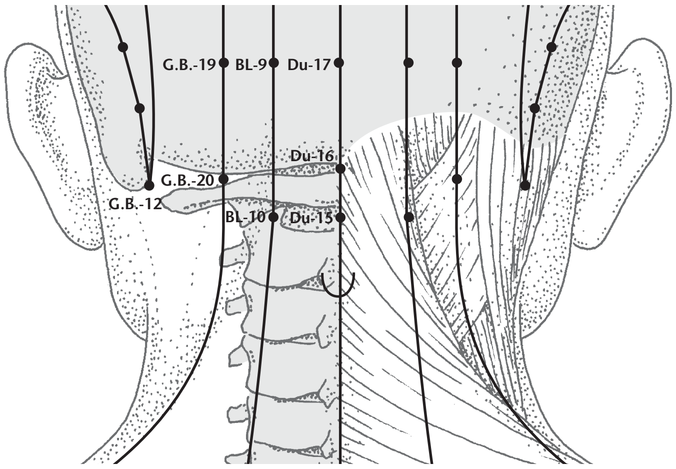
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

Great Shuttle DAZHU BL-11

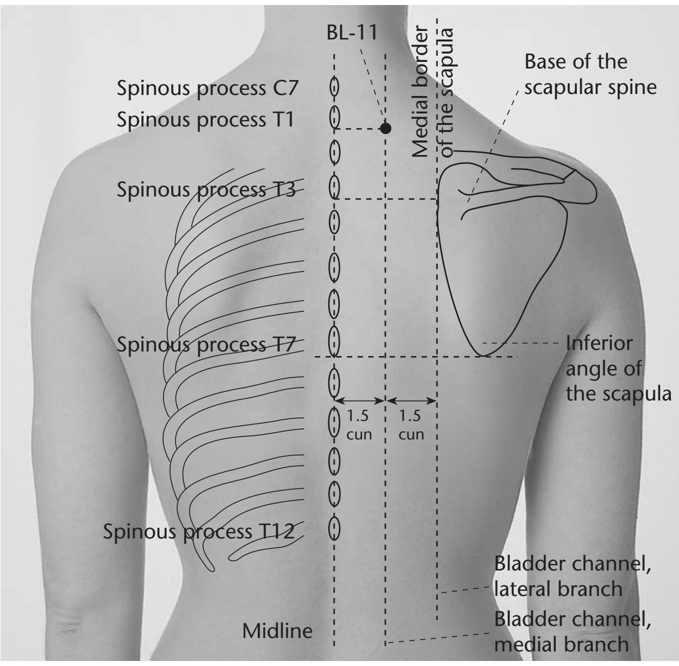
BL-11
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T1
Spinous process T3
Spinous process T7
Inferior angle of the scapula
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

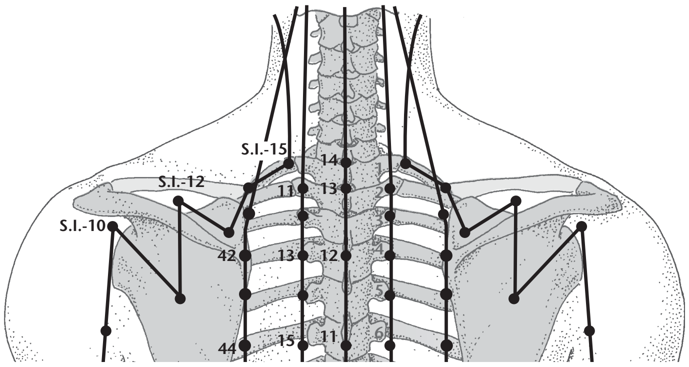
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11

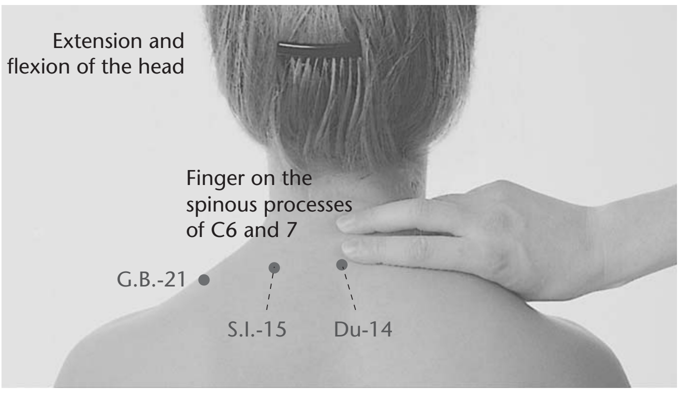
Extension and flexion of the head
Finger on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

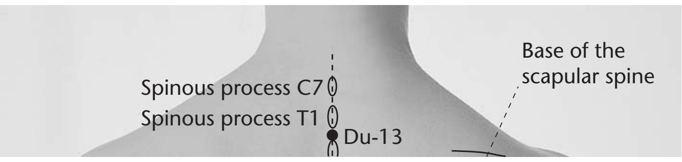
Base of the scapular spine
Spinous process C7
Spinous process T1
Du-13

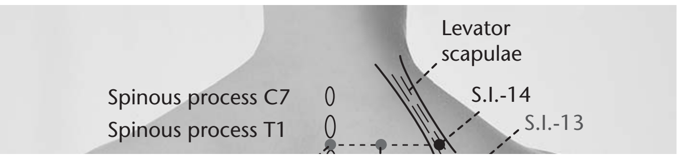
Levator scapulae
Spinous process C7
Spinous process T1
S.I.-14
S.I.-13

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st thoracic vertebra (T1, with hanging shoulders on the level of the acromion).

How to find

First, locate the spinous process of C7 (→ 3.4.1). The next spinous process inferior to this belongs to the T1. Locate **BL-11** 1.5 cun lateral to the lower border of T1, at the highest point of the paraspinal musculature.

Located on the same level are → **Du-13** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **S.I.-14** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors (Wind-Heat)
- Regulates the Lung Qi
- Opens the channel, benefits the bones and joints

Special features

Hui-meeting point with the bones; meeting point with the S.I. channel, according to some authors also with the T.B., G.B. channels and the *du mai*; Sea of Blood point.

BL-12 *Wind Gate* FENGMEN

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2).

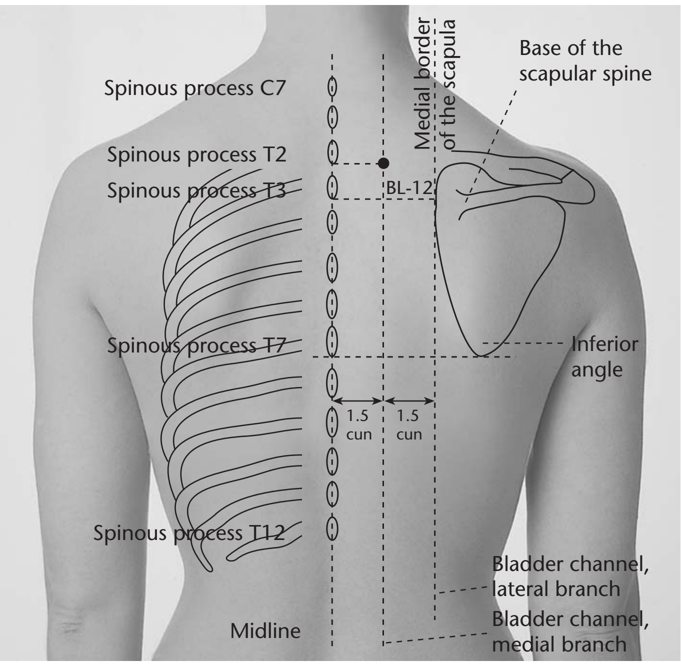
Spinous process C7
Spinous process T2
Spinous process T3
Spinous process T7
Spinous process T12
Midline
Medial border of the scapula
Base of the scapular spine
BL-12
Inferior angle
1.5 cun
1.5 cun
Bladder channel, lateral branch
Bladder channel, medial branch

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down two spinous processes to the lower border of the spinous process of T2. Measure 1.5 cun laterally to locate **BL-12** on the highest point of the paraspinal musculature.

Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **BL-41** (3 cun lateral to the midline) and → **S.I.-13** (on the scapula, on the medial end of the scapular spine).

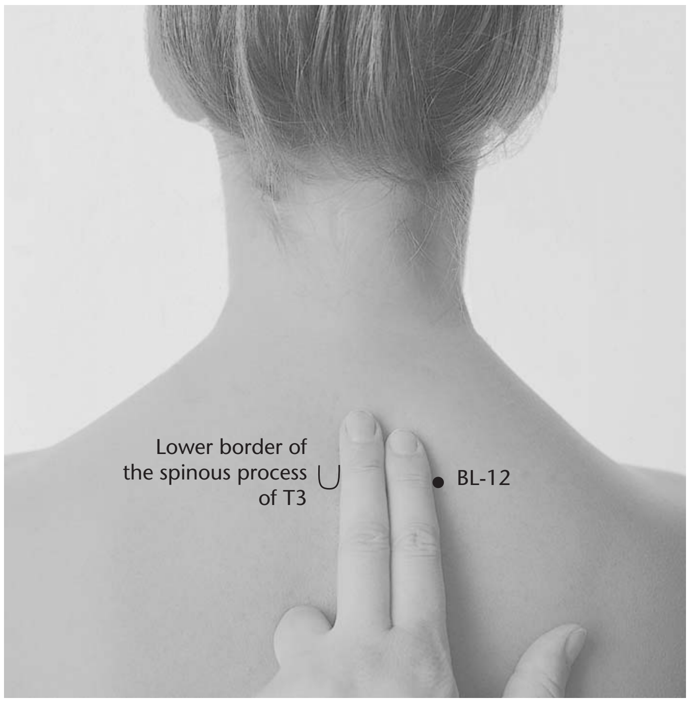
Lower border of the spinous process of T3
BL-12

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax. For acute febrile colds, use reducing needling techniques or cupping. For susceptibility to catching colds, use tonifying needling techniques and/or moxibustion.

Actions/Indications

- Expels (external) Wind and releases the Exterior
- Spreads and descends the Lung Qi
- Tonifies the defensive Qi (*wei qi*), stabilises the Exterior
- Benefits the nose
- Opens the channel

Special features

Meeting point with the *du mai*. Important point for expelling pathogenic factors, especially Wind.

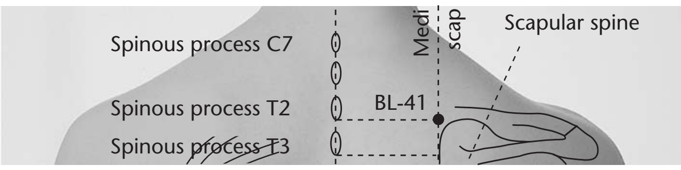
Spinous process C7
Spinous process T2
Spinous process T3
Medi
scap
Scapular spine
BL-41

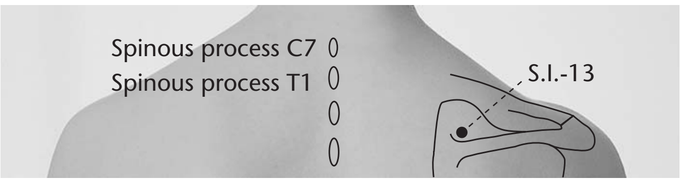
Spinous process C7
Spinous process T1
S.I.-13

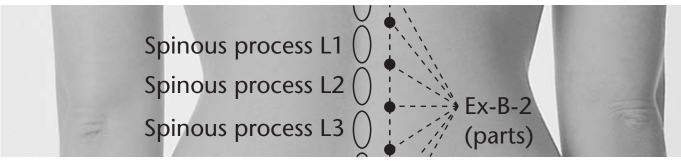
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

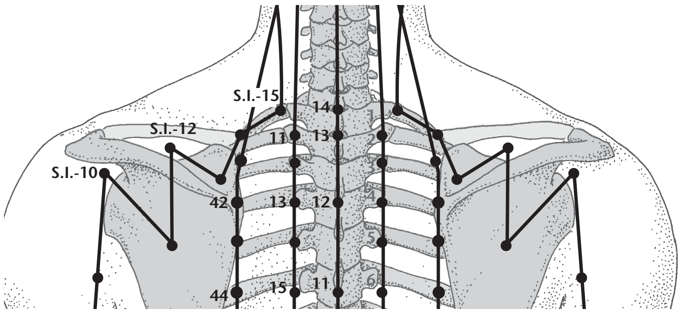
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
1
2
3
4
5
6

Lung Shu FEISHU BL-13

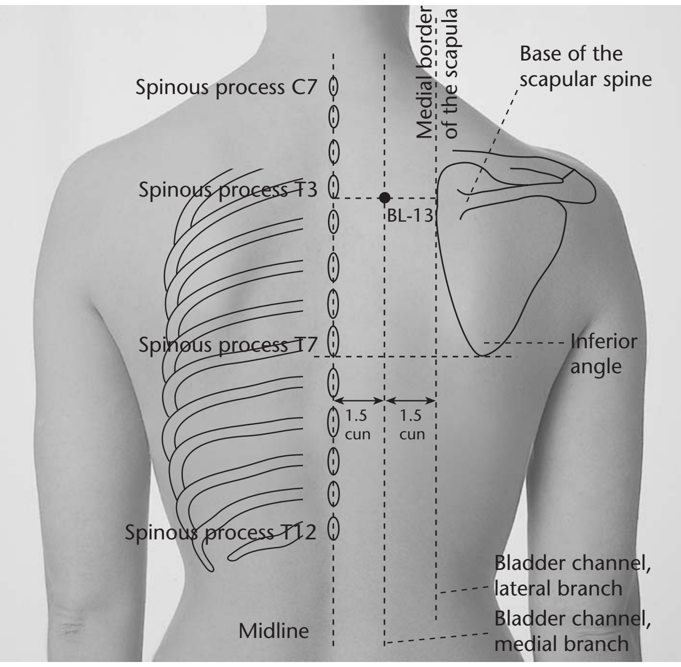
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-13
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

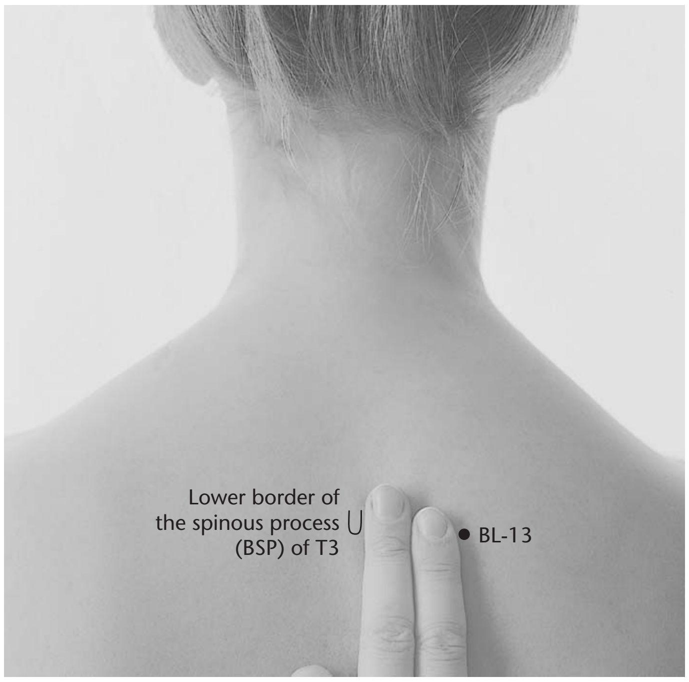
Lower border of the spinous process (BSP) of T3
BL-13

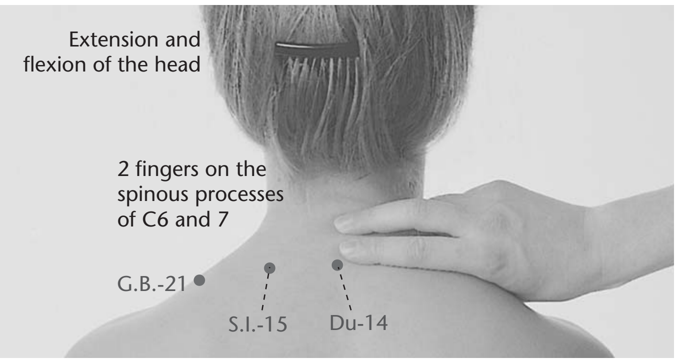
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd thoracic vertebra (T3).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down three spinous processes to the lower border of the spinous process of the T3. Measure 1.5 cun laterally to locate **BL-13** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of the T3 is generally on the level of the medial end of the scapular spine (→ 3.4.2). Locate **BL-13** 1.5 cun lateral to the midline.

Located on the same level are → **Du-12** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and **BL-42** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax. For chronic Lung Qi deficiency without Heat signs, use moxibustion.

Actions/Indications

- Tonifies, spreads and descends the Lung Qi, nourishes the Lung Yin
- Clears Heat from the Lung
- Releases the Exterior
- Opens the channel, alleviates pain

Special features

Back-*shu* point of the lung. Major point for all disorders of the Lung (Lung deficiency and excess patterns).

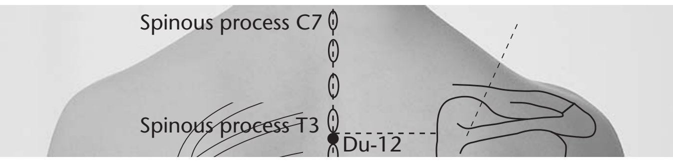
Spinous process C7
Spinous process T3
Du-12

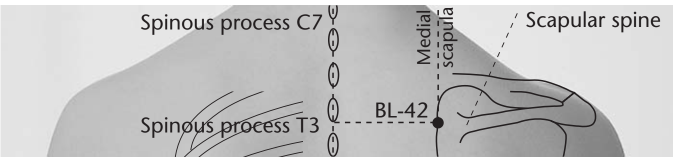
Spinous process C7
Medial scapula
Scapular spine
Spinous process T3
BL-42

BL-14 Jueyin Shu JUEYINSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th thoracic vertebra (T4).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down four spinous processes to the lower border of the spinous process of T4. Measure 1.5 cun laterally to locate **BL-14** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of T3 is generally on the level of the medial end of the scapular spine (→ 3.4.2). The next spinous process in an inferior direction belongs to the T4. Locate **BL-14** 1.5 cun lateral to the midline. Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-43** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Regulates and tonifies the Heart
- Unbinds the chest, regulates the Qi
- Calms the *shen*

Special features

Back-*shu* point of the Pericardium

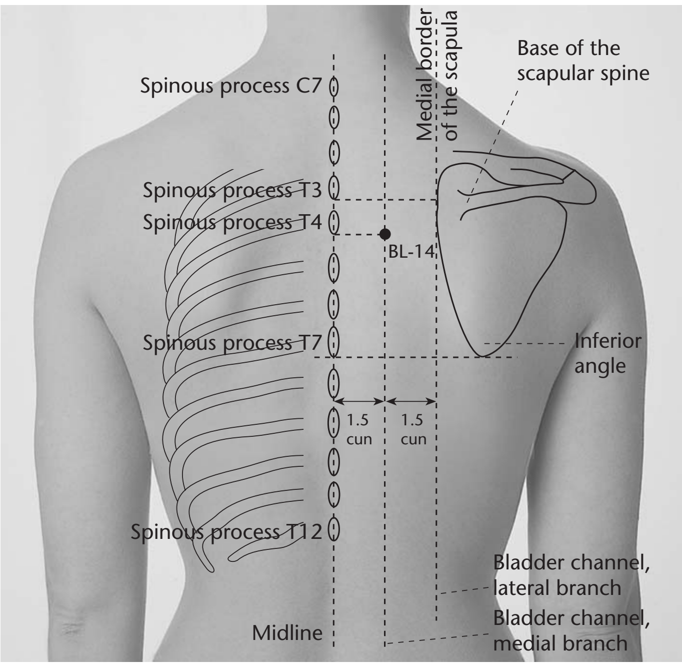
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T4
BL-14
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

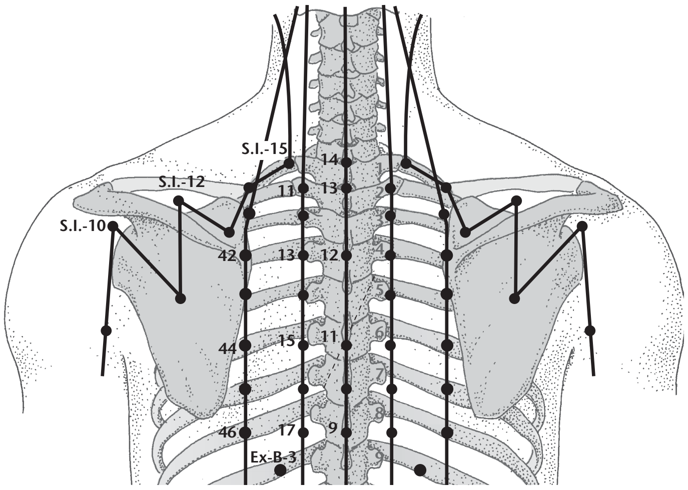
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3

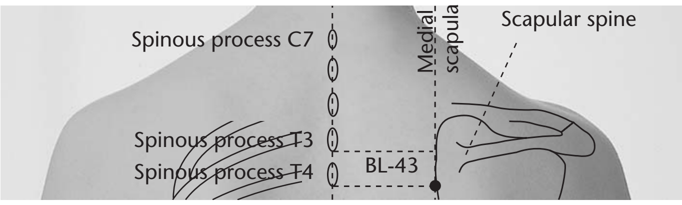
Spinous process C7
Medial scapula
Scapular spine
Spinous process T3
Spinous process T4
BL-43

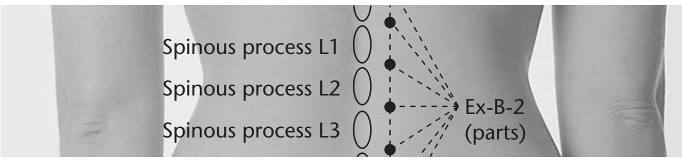
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ch04.7a-F10028.qxd 2/23/08 3:54 PM Page 265

Heart Shu XINSHU BL-15

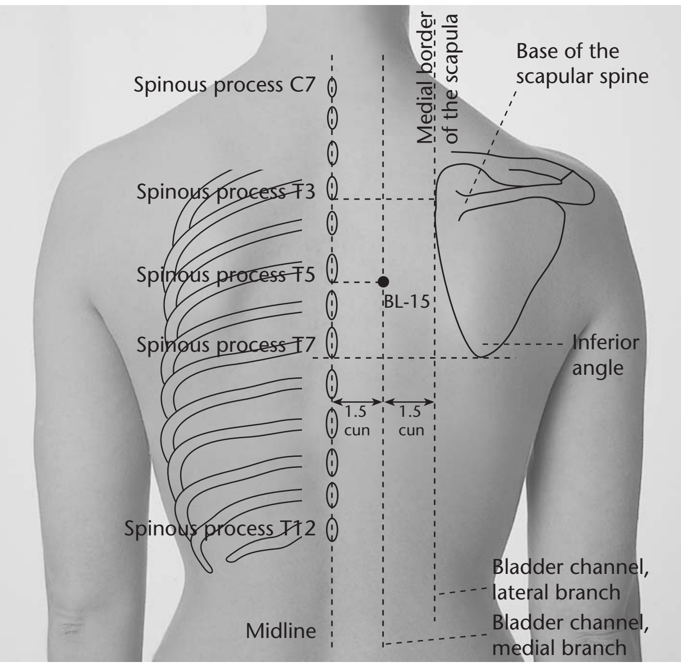
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T5
BL-15
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

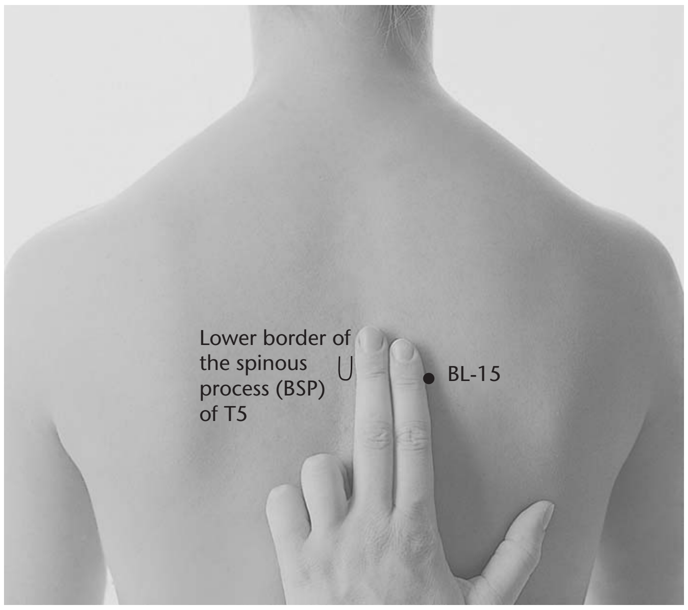
Lower border of the spinous process (BSP) of T5
BL-15

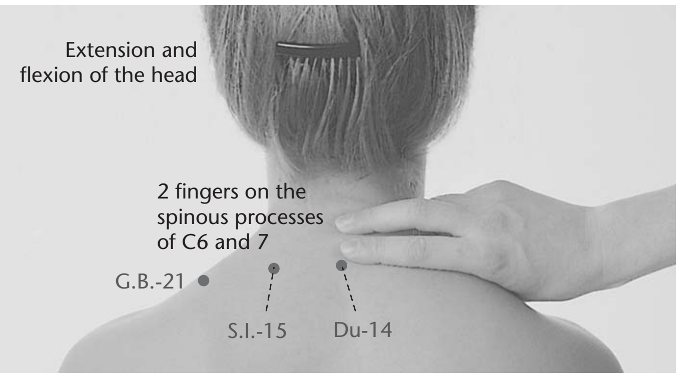
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th thoracic vertebra (T5).

How to find

First, locate the spinous process of C7 (→ 3.4.1). From there, count down five spinous processes to the lower border of the spinous process of T5. Measure 1.5 cun laterally to locate **BL-15** on the highest point of the paraspinal musculature. **Or:** With the patient seated and their arms hanging down, the spinous process of the T3 is generally on the level of the easily palpable medial end of the scapular spine (→ 3.4.2). From there, count down two spinous processes in an inferior direction to the T5. Locate **BL-15** 1.5 cun lateral to the midline.

Located on the same level are → **Du-11** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-44** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the Heart, regulates the Heart Qi, unbinds the chest and removes Blood stasis, calms the *shen*, clears Fire, moves Qi locally and along the channel pathway: for example, for disorders of the thoracic spine and intercostal region

Special features

Back-*shu* point of the Heart

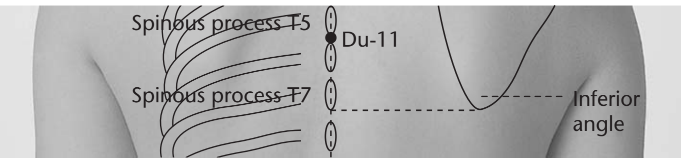
Spinous process T5
Du-11
Spinous process T7
Inferior angle

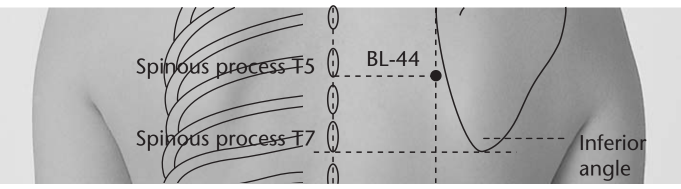
Spinous process T5
BL-44
Spinous process T7
Inferior angle

BL-16 *Governing* Shu DUSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 6th thoracic vertebra (T6).

How to find

First, locate the spinous process of the C7 (→ 3.4.1). From there, count down six spinous processes to the lower border of the spinous process of T6. Measure 1.5 cun laterally to locate **BL-16** on the highest point of the paraspinal musculature. **Or:** With the patient standing and their arms hanging down, the spinous process of T7 is level with the inferior angle of the scapula (→ 3.4.2). The next spinous process in a superior direction belongs to T6. Locate **BL-16** 1.5 cun lateral to the midline.

Located on the same level are → **Du-10** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-45** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine or transversely (subcutaneously). Caution: Pneumothorax.

Actions/Indications

- Unbinds the chest and regulates the Qi movement in the thorax and the abdomen

Special features

Despite its designation as Back-*shu* point of the *du mai*, this point is hardly mentioned in classical or modern texts.

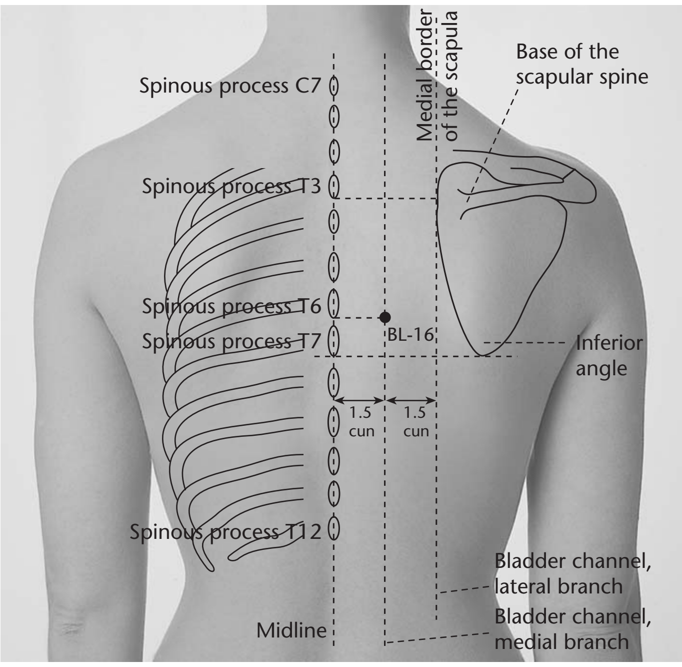
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T6
Spinous process T7
BL-16
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

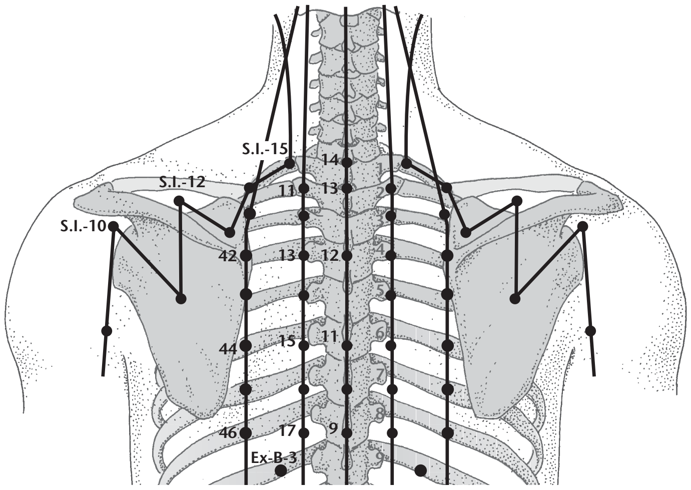
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3
1
2
3
4
5
6
7
8
9

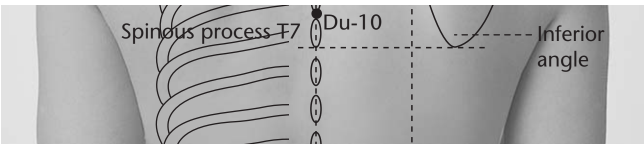
Spinous process T7
Du-10
Inferior angle

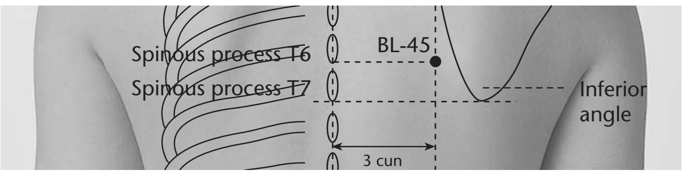
Spinous process T6
Spinous process T7
BL-45
Inferior angle
3 cun

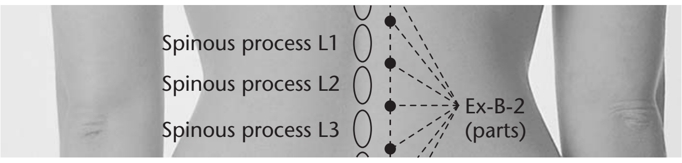
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Diaphragm Shu GESHU BL-17

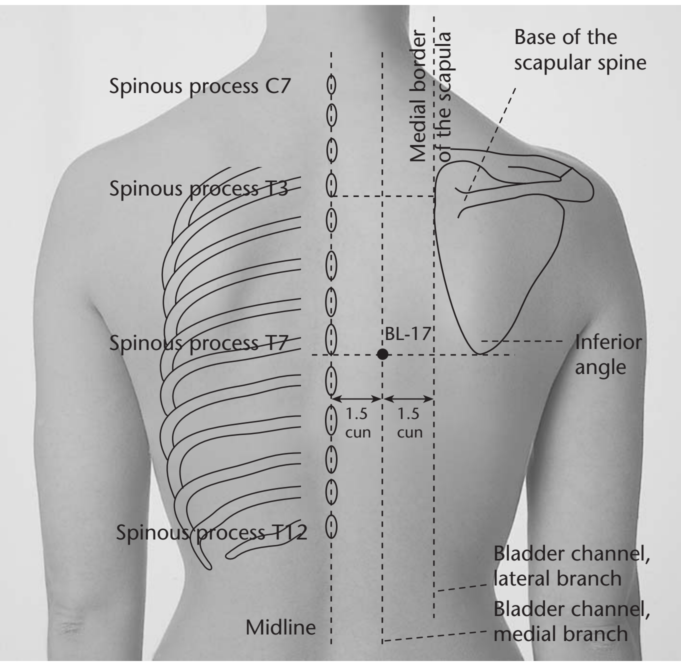
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-17
Spinous process T7
Inferior angle
1.5 cun
1.5 cun
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

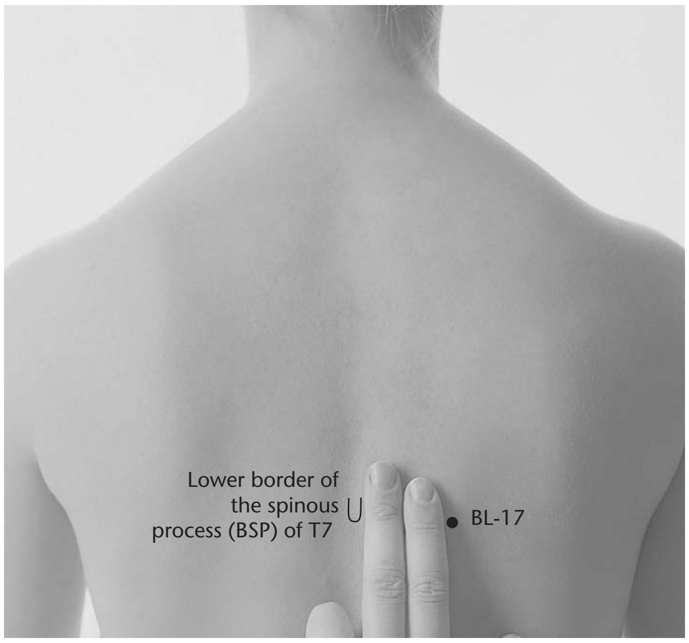
Lower border of the spinous process (BSP) of T7
BL-17

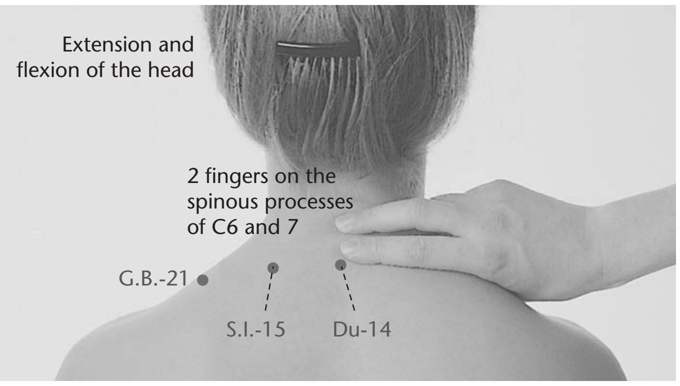
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 7th thoracic vertebra (T7).

How to find

The spinous process of the T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). **BL-17** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use the spinous process of the T3 for reference (generally on the level of the medial end of the scapular spine) (→ 3.4.2). **Or:** Use C7 as a reference point (→ 3.4.1).
Located on the same level are → **Du-9** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-46** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Cools Blood-Heat, stops bleeding, removes Blood stasis, nourishes and harmonises Blood (and Yin)
- Regulates the diaphragm, descends counterflow Qi
- Moves Qi locally and along the channel pathway

Special features

Back-*shu* point of the Diaphragm, *hui*-meeting point of the Blood.

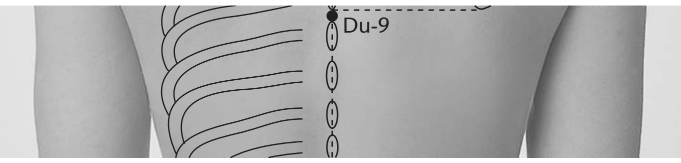

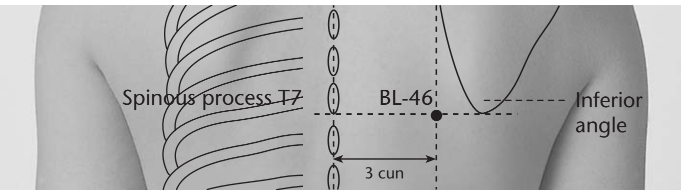
Spinous process T7
BL-46
Inferior angle
3 cun

BL-18 *Liver* Shu GANSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 9th thoracic vertebra (T9).

How to find

In a standing position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down two spinous processes to the spinous process of T9. **BL-18** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use the C7 as a reference point (→ 3.4.1). From there, count down nine spinous processes to T9 and locate **BL-18** 1.5 cun lateral to the lower border of its spinous process.

Located on the same level are → **Du-8** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-47** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Spreads Liver Qi, cools Fire, clears Damp-Heat, regulates and nourishes Liver Blood, subdues (internal) Wind
- Benefits the eyes
- Benefits the sinews

Special features

Back-*shu* point of the Liver

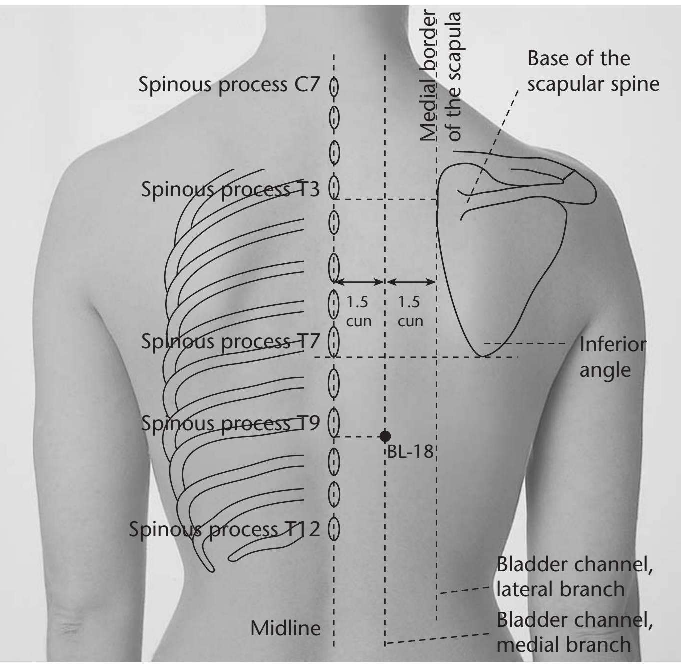
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T9
BL-18
Spinous process T12
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

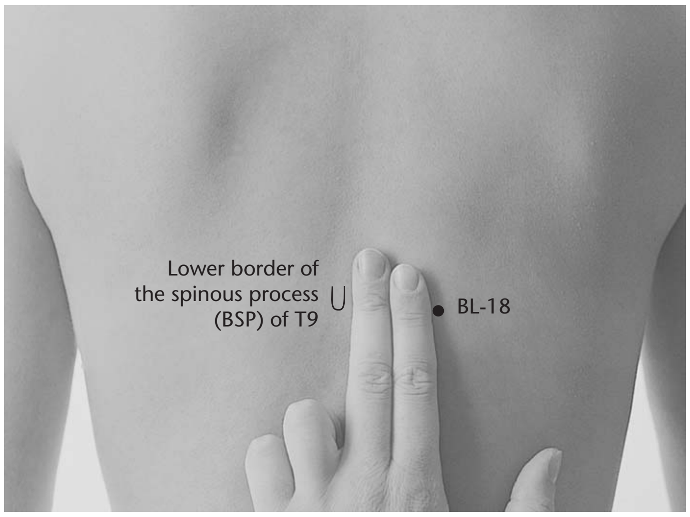
Lower border of the spinous process (BSP) of T9
BL-18

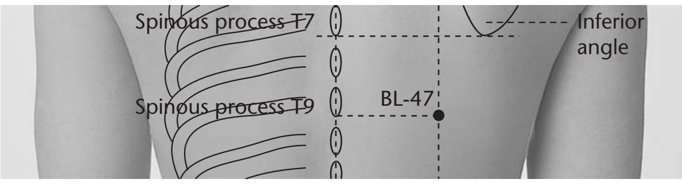
Spinous process T7
Inferior angle
Spinous process T9
BL-47

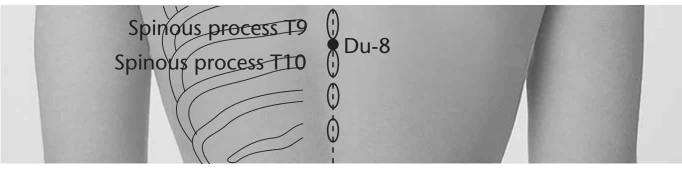
Spinous process T9
Spinous process T10
Du-8

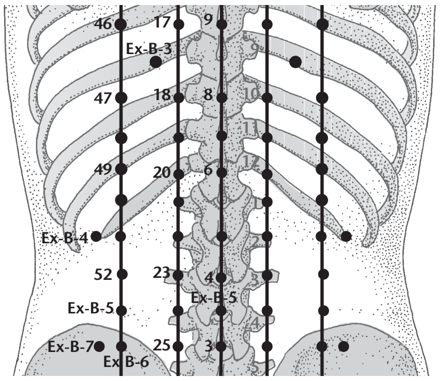
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

Gall Bladder Shu DANSHU BL-19

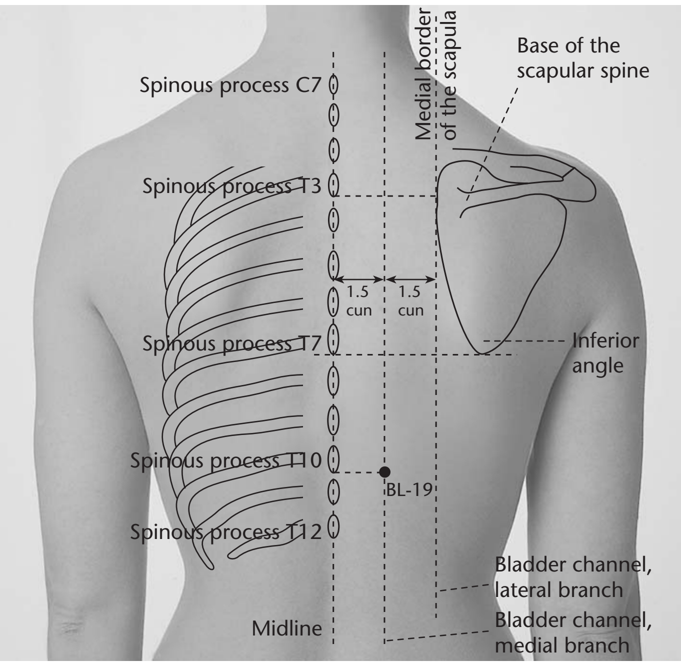
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19
Spinous process T12
Bladder channel, lateral branch
Bladder channel, medial branch
Midline

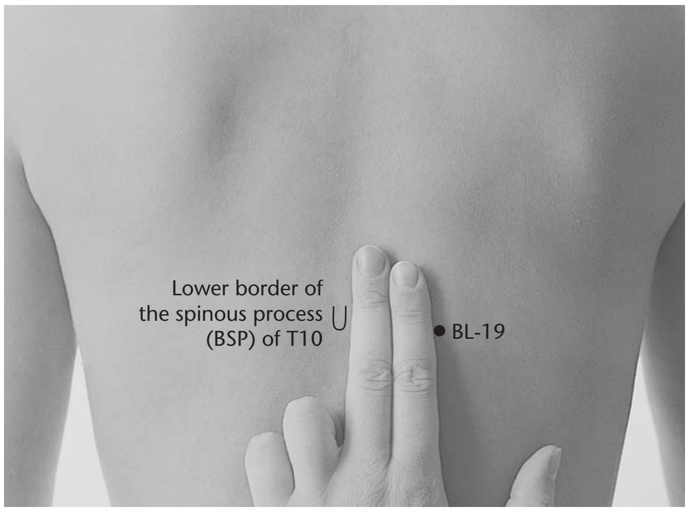
Lower border of the spinous process (BSP) of T10
BL-19

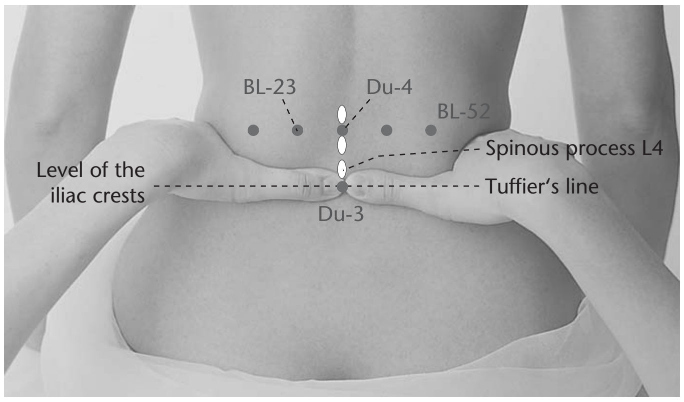
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

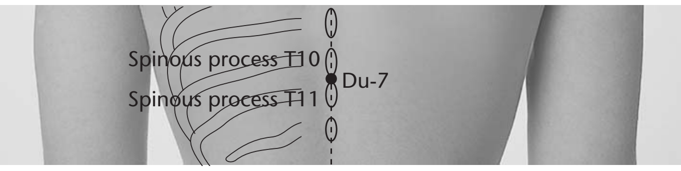
Spinous process T10
Du-7
Spinous process T11

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 10th thoracic vertebra (T10).

How to find

In a standing position with the patient's arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down three spinous processes to the spinous process of T10. **BL-19** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Use C7 as a reference point (→ 3.4.1). From there, count down 10 spinous processes to T10 and locate **BL-19** 1.5 cun lateral to the lower border of its spinous process. **Or:** Use the lumbar spine as a reference point (→ 3.4.3).

Located on the same level are → **Du-7** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-48** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Clears Damp-Heat from the Liver and Gall Bladder
- Expels pathogenic factors from the *shaoyang*
- Tonifies and regulates the Gall Bladder Qi
- Unbinds the chest and relaxes the pathway of the channel

Special features

Back-*shu* point of the Gall Bladder. Important point for clearing Damp-Heat from the Liver and Gall Bladder.

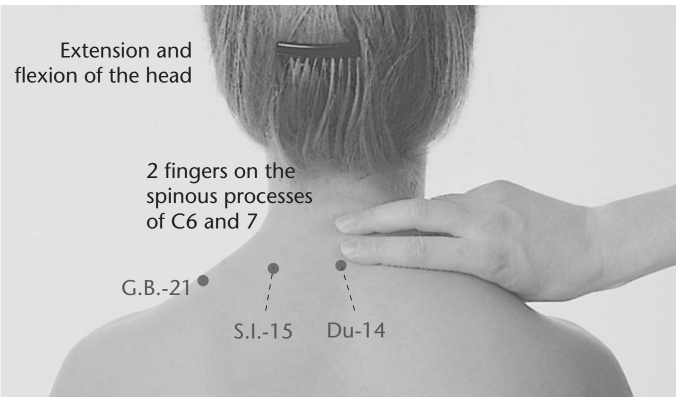
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

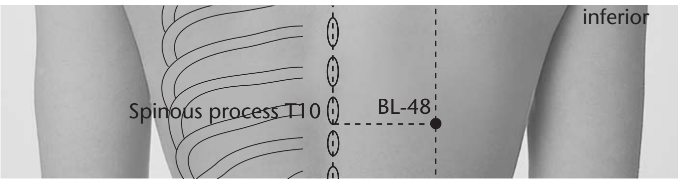
inferior
Spinous process T10
BL-48

BL-20 *Spleen* Shu PISHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 11th thoracic vertebra (T11).

How to find

In a seated position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down four spinous processes to the spinous process of T11. **BL-20** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Locate the articulation of the most inferior rib (T12) and palpate in a superior direction to T11. **Or:** Use the lumbar spine for reference (→ 3.4.3).

Located on the same level are → **Du-6** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and **BL-49** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Tonifies the Spleen Qi and Yang, regulates Qi in the Middle, raises Qi
- Transforms Dampness
- Holds and nourishes Blood

Special features

Back-*shu* point of the spleen. Major point for tonifying the Middle, especially for Spleen deficiency syndromes. Use moxibustion or tonifying needle techniques for nourishing the Blood.

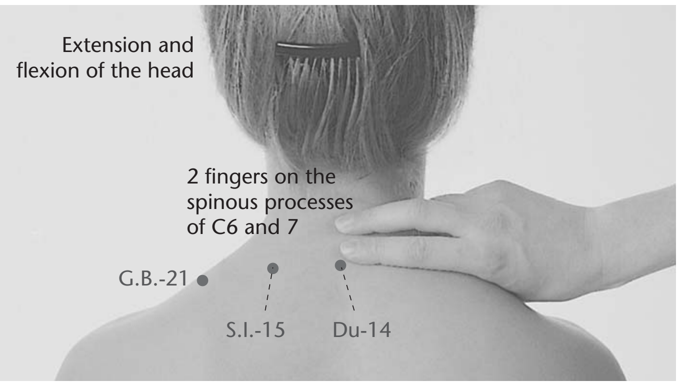
Extension and
flexion of the head
2 fingers on the
spinous processes
of C6 and 7
G.B.-21
S.I.-15
Du-14

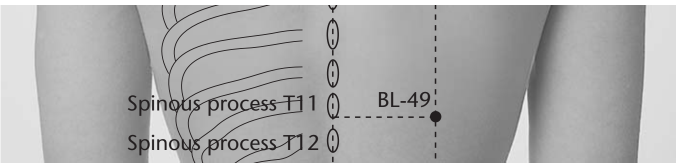
Spinous process T11
Spinous process T12
BL-49

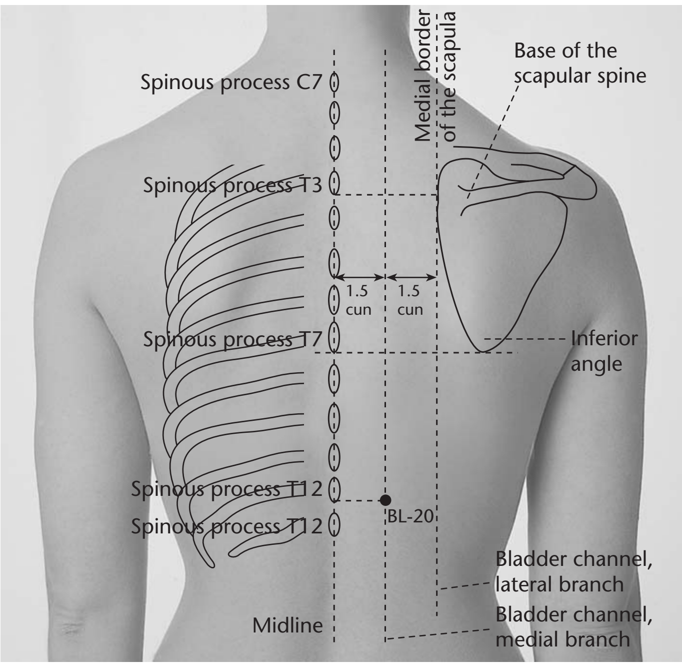
Spinous process C7
Medial border
of the scapula
Base of the
scapular spine
Spinous process T3
1.5
cun
1.5
cun
Spinous process T7
Inferior
angle
Spinous process T12
Spinous process T12
BL-20
Bladder channel,
lateral branch
Midline
Bladder channel,
medial branch

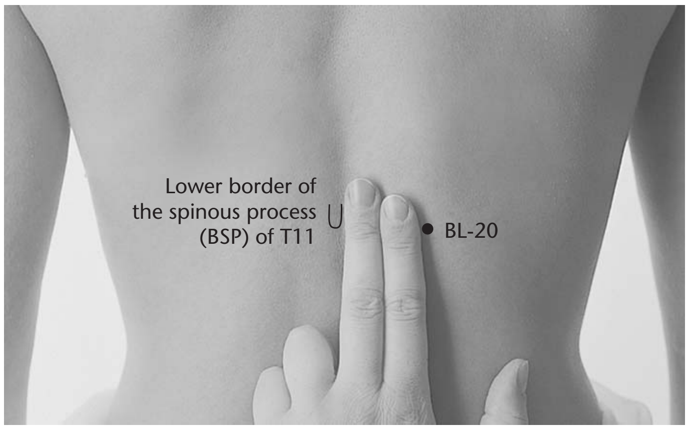
Lower border of
the spinous process
(BSP) of T11
BL-20

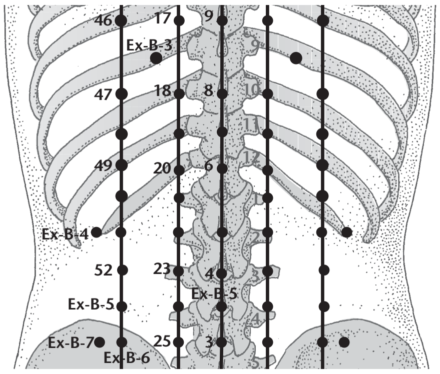
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
Ex-B-4
1
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

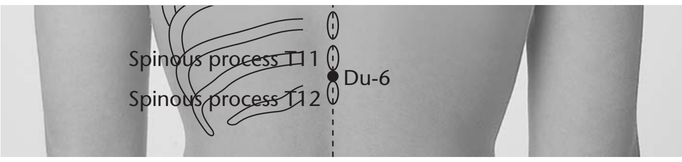
Spinous process T11
Du-6
Spinous process T12

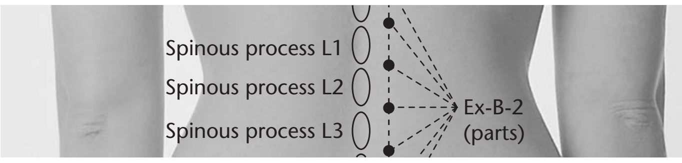
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

Stomach Shu **WEISHU** **BL-21**

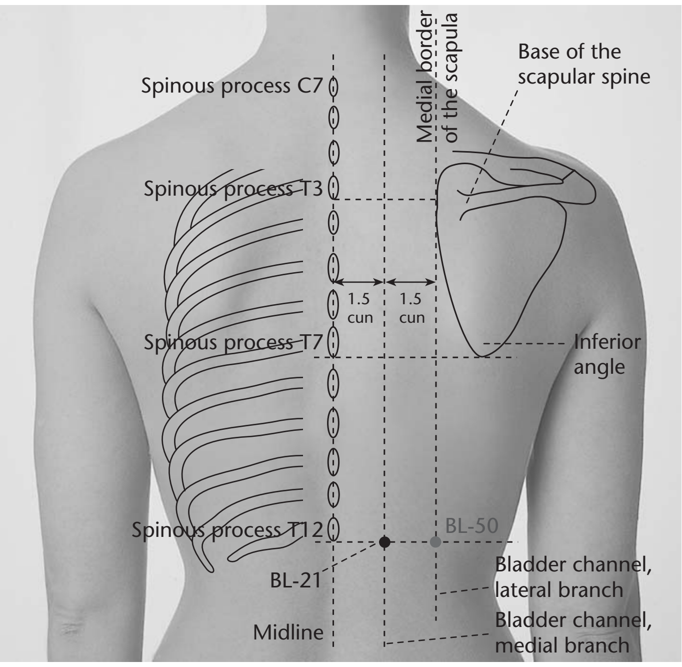
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T12
BL-50
BL-21
Bladder channel, lateral branch
Midline
Bladder channel, medial branch

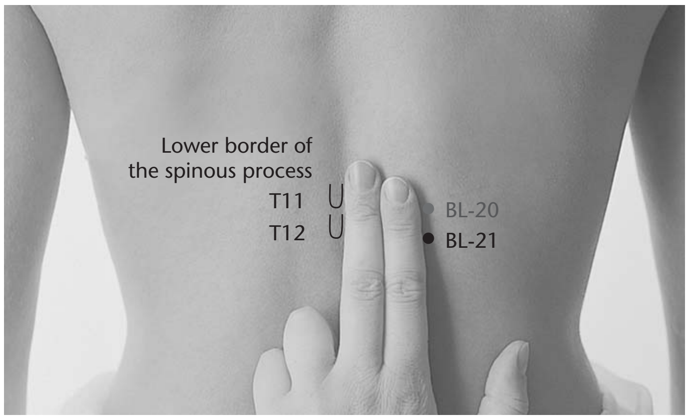
Lower border of the spinous process
T11
T12
BL-20
BL-21

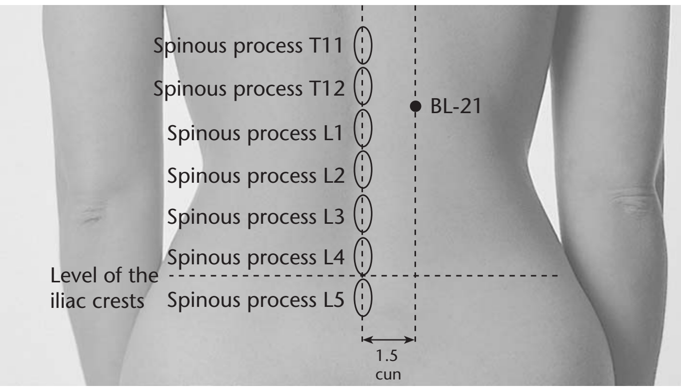
Spinous process T11
Spinous process T12
BL-21
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Level of the iliac crests
Spinous process L5
1.5 cun

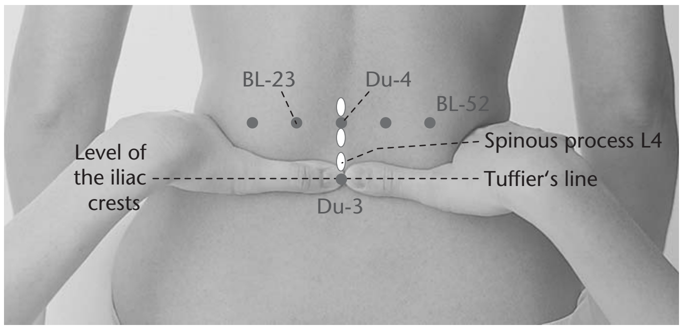
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 12th thoracic vertebra (T12).

How to find

In a seated position with the arms hanging down, the spinous process of T7 is approximately on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down five spinous processes to the spinous process of T12. **BL-21** is located on the level of the lower border of its spinous process and 1.5 cun lateral to the midline at the highest point of the paraspinal musculature. **Or:** Palpate for the articulation of the most inferior rib with T12 and locate **BL-21** 1.5 cun lateral to the lower border of its spinous process. **Or:** Use the lumbar spine for reference (→ 3.4.3). Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-50** (3 cun lateral to the midline).

Needling

0.5–1 cun obliquely towards the spine. In clinical practice, this point is often needled transversely (subcutaneously) towards the spine. This allows the patient to lie in a supine position during needle retention (the needles can be secured with tape). Caution: Pneumothorax.

Actions/Indications

- Regulates the Stomach, descends counterflow Qi, harmonises the Middle Burner, eliminates Dampness and food stagnation
- Opens the channel

Special features

Back-*shu* point of the Stomach, main point for regulating any dysfunction of the Stomach.

BL-22 San Jiao Shu SANJIAOSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line (a line connecting the highest points of the iliac crests), which in most cases intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L1 and, on this level, locate **BL-22** 1.5 cun in a lateral direction, on the highest point of the paraspinal musculature. Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2/BL-51/Ex-B-4** (0.5 cun/3 cun/3.5cun lateral of the midline).

Needling

Obliquely 0.5–1.5 cun. Caution: Kidneys.

Actions/Indications

- Regulates the Triple Burner
- Opens the water passages and promotes urination
- Has an effect on the *shaoyang*

Special features

Back-*shu* point of the Triple Burner. Especially regulates the relationship between the Middle and Lower Burner.

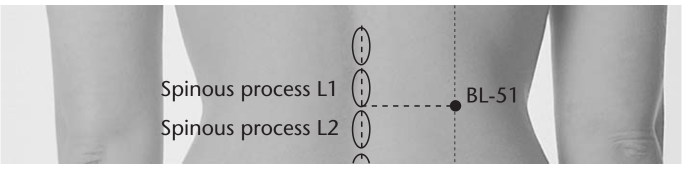
Spinous process L1
Spinous process L2
BL-51

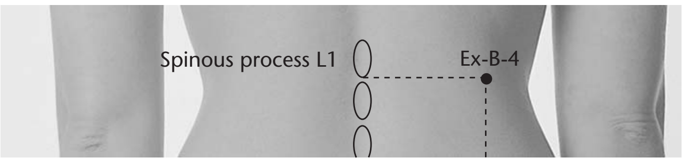
Spinous process L1
Ex-B-4

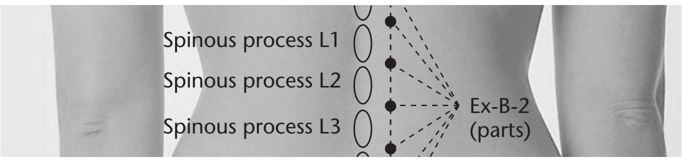
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

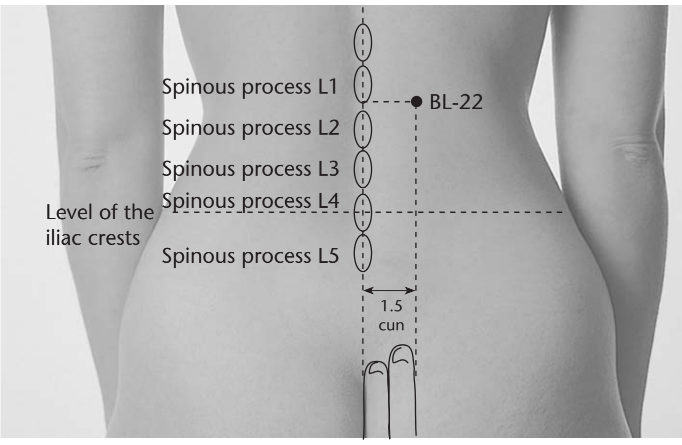
Spinous process L1
BL-22
Spinous process L2
Spinous process L3
Spinous process L4
Level of the
iliac crests
Spinous process L5
1.5
cun

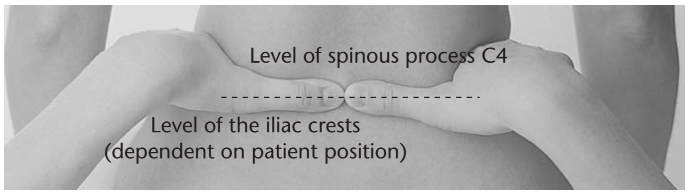
Level of spinous process C4
Level of the iliac crests
(dependent on patient position)

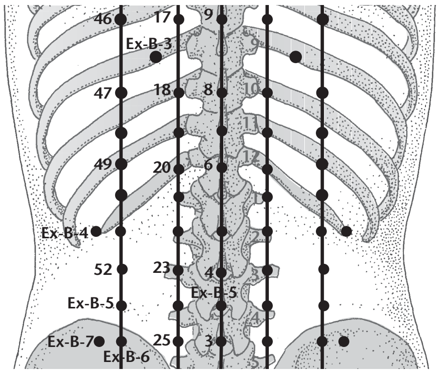
46
17
9
Ex-B-3
9
47
18
8
10
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5

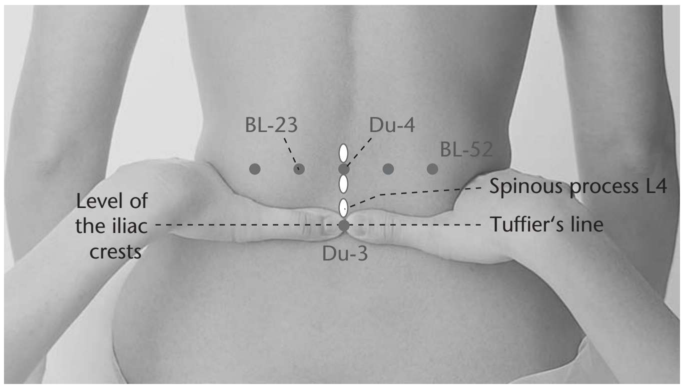
BL-23
Du-4
BL-52
Spinous process L4
Level of
the iliac
crests
Tuffier's line
Du-3

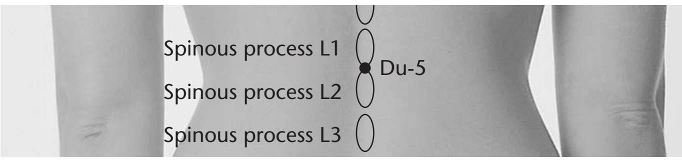
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

Ch04.7a-F10028.qxd 2/23/08 3:54 PM Page 273

Kidney Shu SHENSHU BL-23

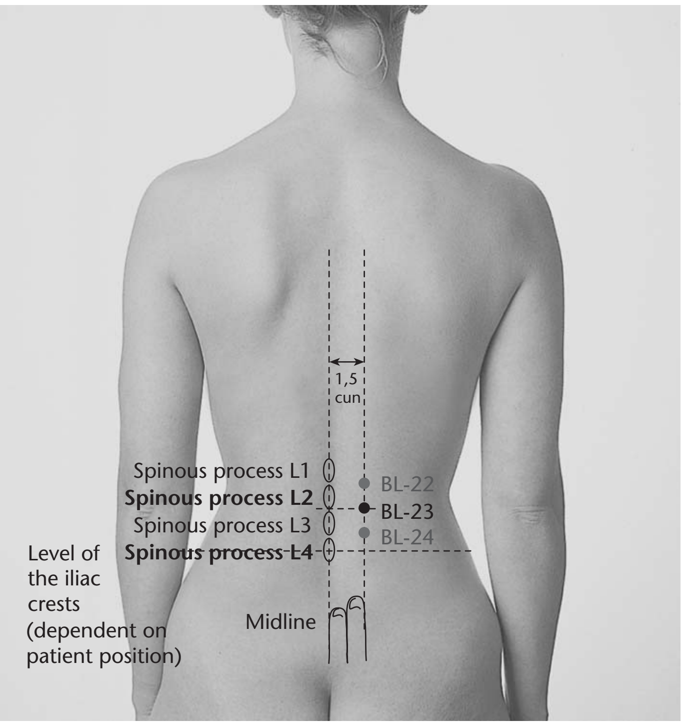
1,5
cun
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
BL-22
BL-23
BL-24
Level of the iliac crests (dependent on patient position)
Midline

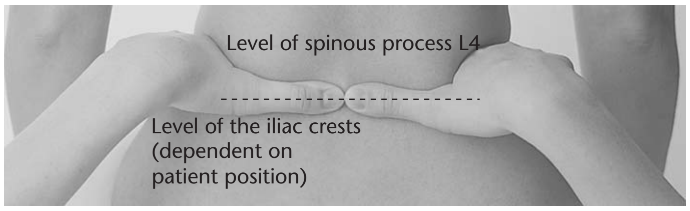
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

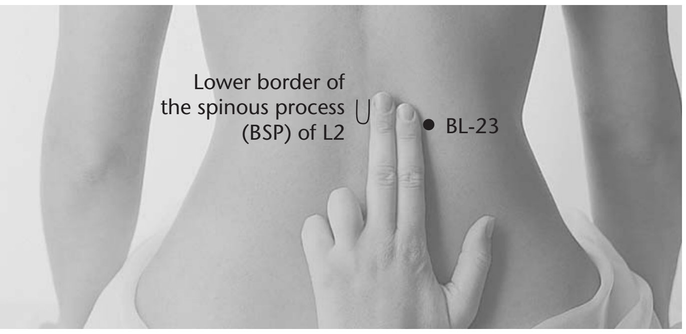
Lower border of the spinous process (BSP) of L2
BL-23

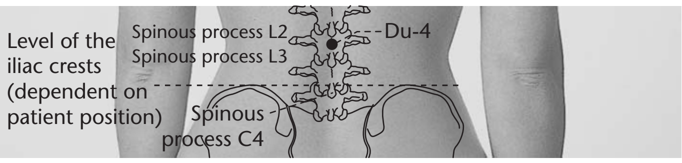
Level of the iliac crests (dependent on patient position)
Spinous process L2
Spinous process L3
Du-4
Spinous process C4

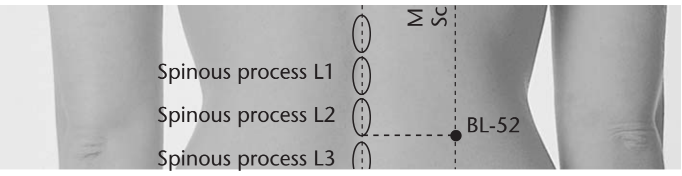
Spinous process L1
Spinous process L2
Spinous process L3
BL-52

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line (a line connecting the highest points of the iliac crests), which in most cases intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L2 and, on this level, locate **BL-23** 1.5 cun in a lateral direction, on the highest point of the paraspinal musculature. Located on the same level are → **Du-4** (on the midline), a point of → **Ex-B-2/BL-52** (0.5 cun/3 cun lateral of the midline).

Needling

Vertically or obliquely 0.5–1.5 cun. Caution: Kidneys.

Actions/Indications

- Strengthens the Kidneys, tonifies the Kidney Qi and Yang, benefits the Essence (*jing*), nourishes the Kidney Yin
- Regulates the Lower Burner, benefits the uterus
- Benefits the bones and the marrow
- Benefits the eyes and ears
- Strengthens the lower back

Special features

Back-*shu* point of the Kidneys, according to some authors, meeting point with the *du mai*. Major point for tonifying the Kidneys. Use tonifying needle techniques for all Kidney deficiency syndromes. Moxibustion tonifies especially the Kidney Yang. **Caution**: Moxibustion is contraindicated for Kidney Yin deficiency with empty Heat (Heat in the Five Hearts (Heat in the thorax, the palms and the soles of the feet), night sweats, restlessness, often a red tongue).

BL-24 *Sea of Qi* QIHAISHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd lumbar vertebra (L3).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). From there, count upward to the lower border of the spinous process of L3 and, on this level, locate **BL-24** 1.5 cun in a lateral direction.

Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **Ex-B-5** (*xiazhishi*; 3 cun lateral to the midline according to the WHO location notes).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Benefits the back and strengthens the knees
- Regulates and tonifies Qi and Blood of the Lower Burner

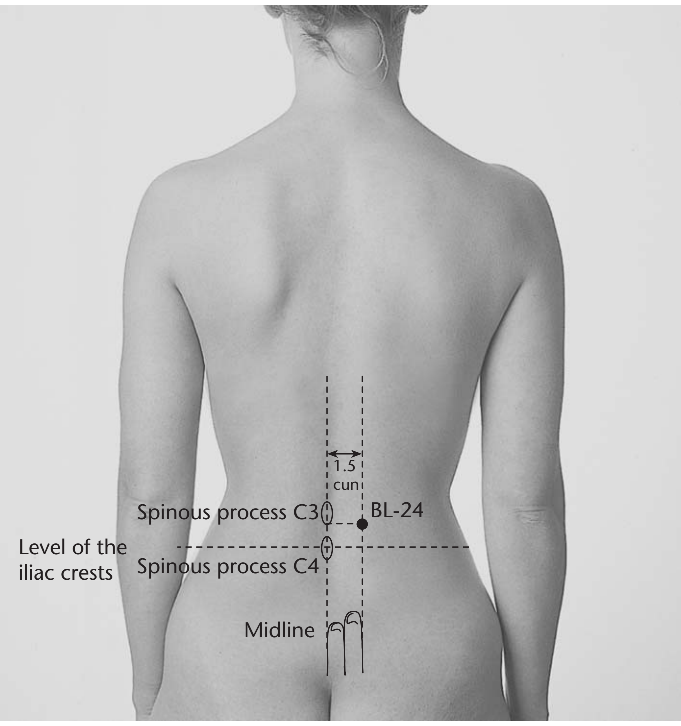
1.5 cun
Spinous process C3
BL-24
Level of the iliac crests
Spinous process C4
Midline

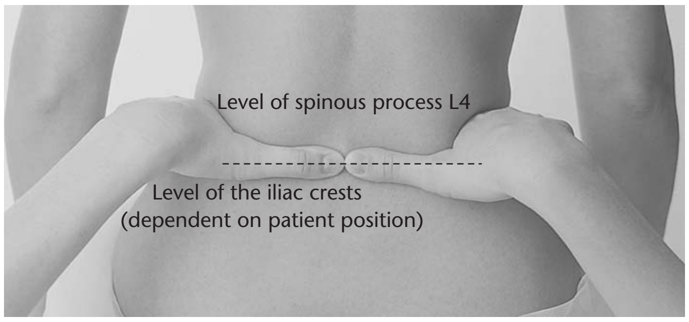
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

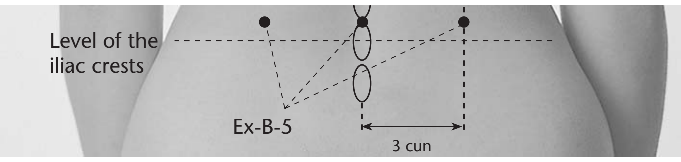
Level of the iliac crests
Ex-B-5
3 cun

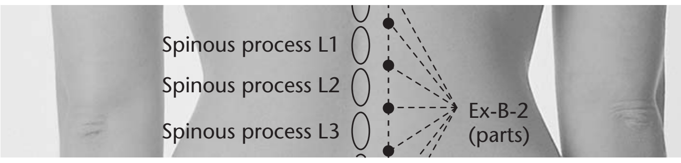
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

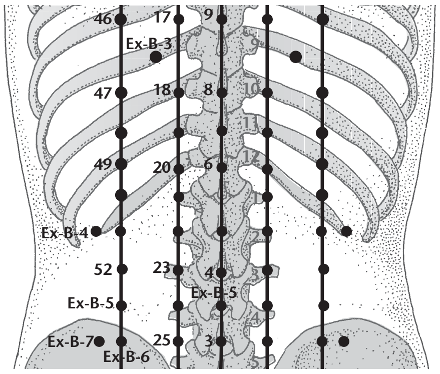
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6

Large Intestine Shu DACHANGSHU BL-25

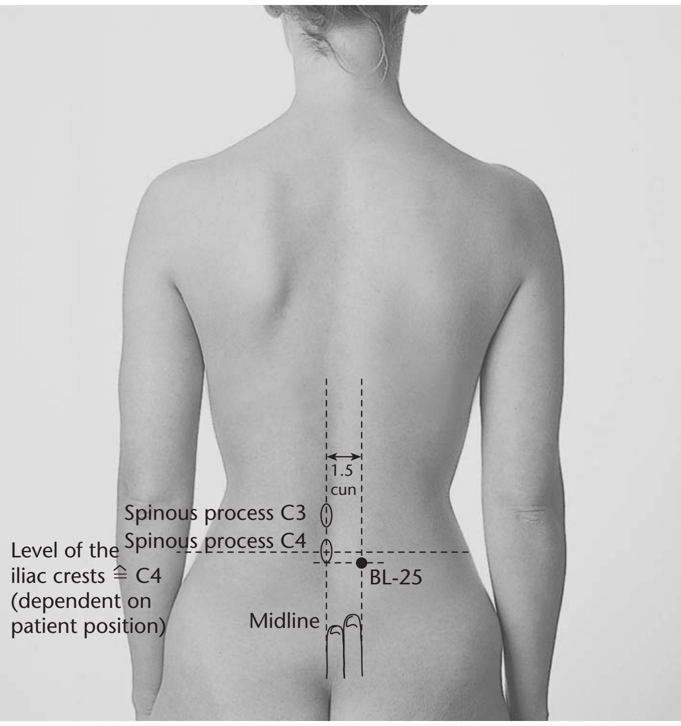
1.5 cun
Spinous process C3
Spinous process C4
Level of the iliac crests ≙ C4 (dependent on patient position)
BL-25
Midline

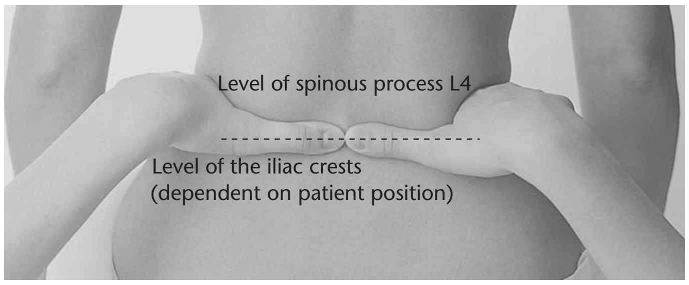
Level of spinous process L4
Level of the iliac crests (dependent on patient position)

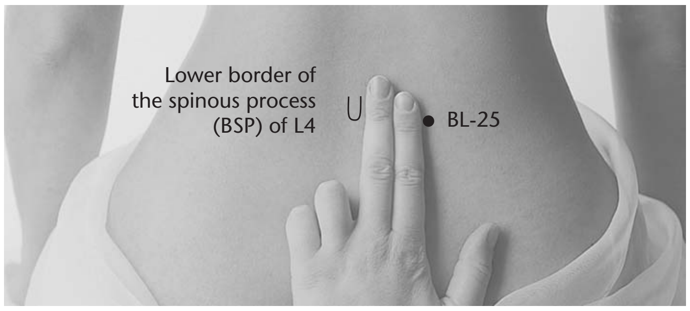
Lower border of the spinous process (BSP) of L4
BL-25

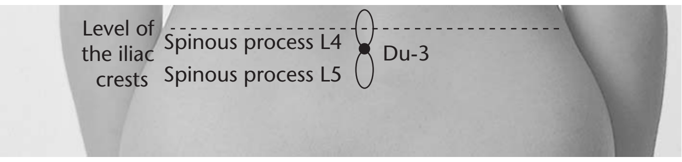
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

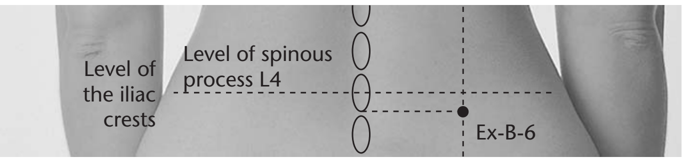
Level of the iliac crests
Level of spinous process L4
Ex-B-6

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.3). On this level, locate **BL-25** 1.5 cun in a lateral direction.

Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **Ex-B-6** (3 cun lateral to the midline) and → **Ex-B-7** (3.5 cun lateral to the midline).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Regulates the intestines and promotes the Qi flow
- Strengthens the lower back

Special features

Back-*shu* point of the Large Intestine. Important point for regulating the Large Intestine Qi. Important local point for acute and chronic lumbar disorders.

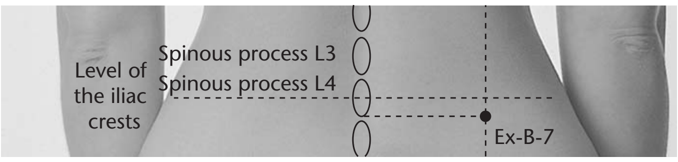
Level of the iliac crests
Spinous process L3
Spinous process L4
Ex-B-7

Ch04.7a-F10028.qxd 2/23/08 3:54 PM Page 276

BL-26 *Gate of Origin* Shu GUANYUANSHU

Location

1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th lumbar vertebra (L5).

How to find

For orientation in the lumbar region (→ 3.4.3), first, determine the Tuffier's line. Place both hands directly superior or lateral to the highest points of the iliac crests, joining the thumbs on the midline at the lumbar spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position; for more detail on orientation in the lumbar/sacral region → 3.4.3, 3.4.4). Next, locate the lower border of the spinous process of L5 and, on this level, locate **BL-26** 1.5 cun in a lateral direction.

Located on the same level are → **Ex-B-8** (on the midline) and a point of → **Ex-B-2** (0.5 cun lateral to the midline).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the back, especially for Kidney deficiency syndromes
- Regulates the Lower Burner

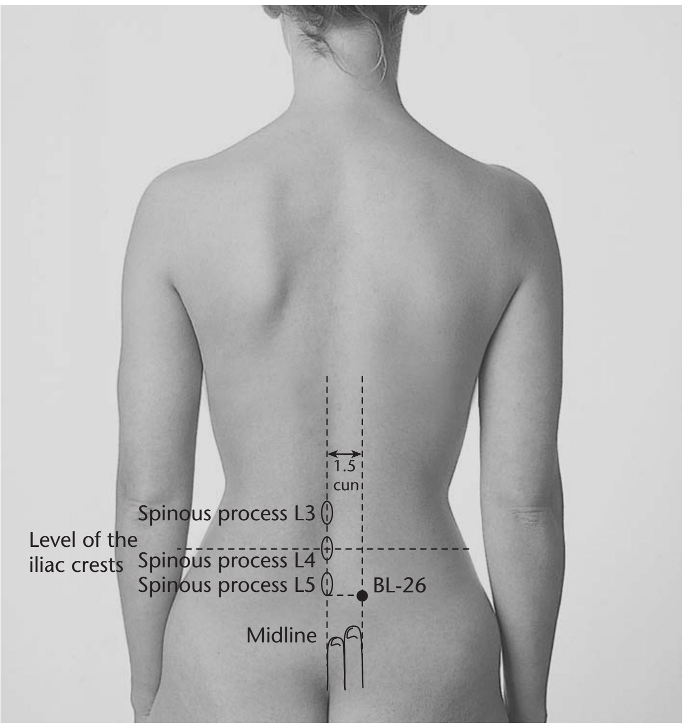
1.5 cun
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline

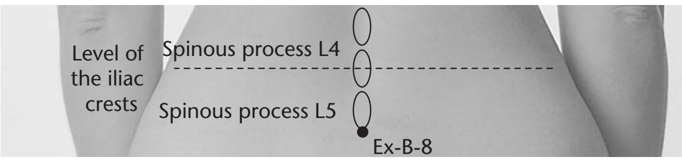
Level of the iliac crests
Spinous process L4
Spinous process L5
Ex-B-8

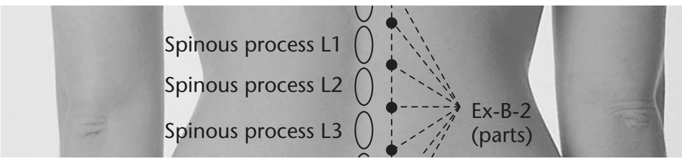
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

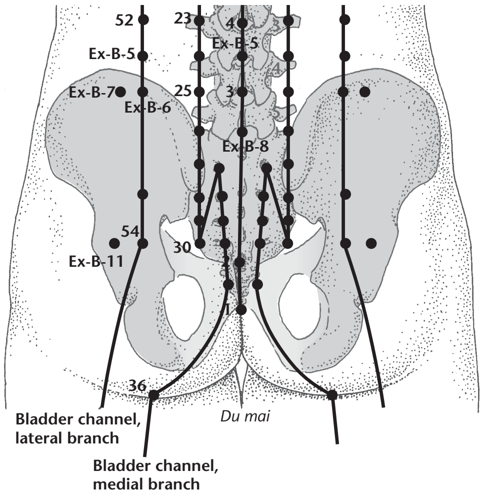
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

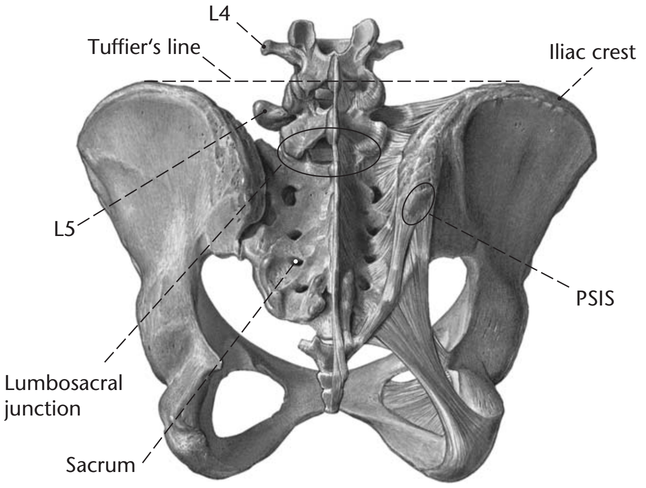
L4
Tuffier's line
Iliac crest
L5
PSIS
Lumbosacral junction
Sacrum

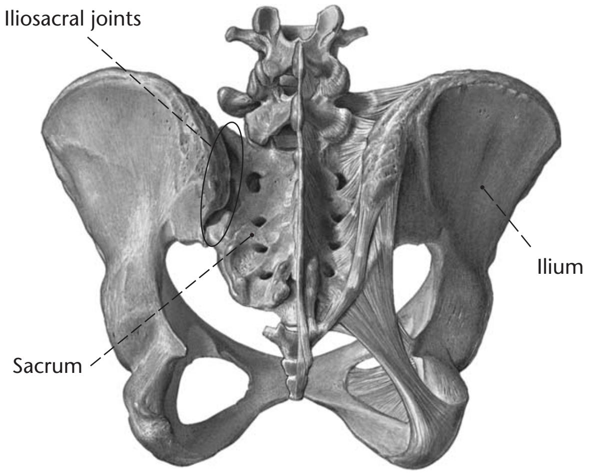
Iliosacral joints
Ilium
Sacrum

Small Intestine Shu XIAOCHANGSHU BL-27

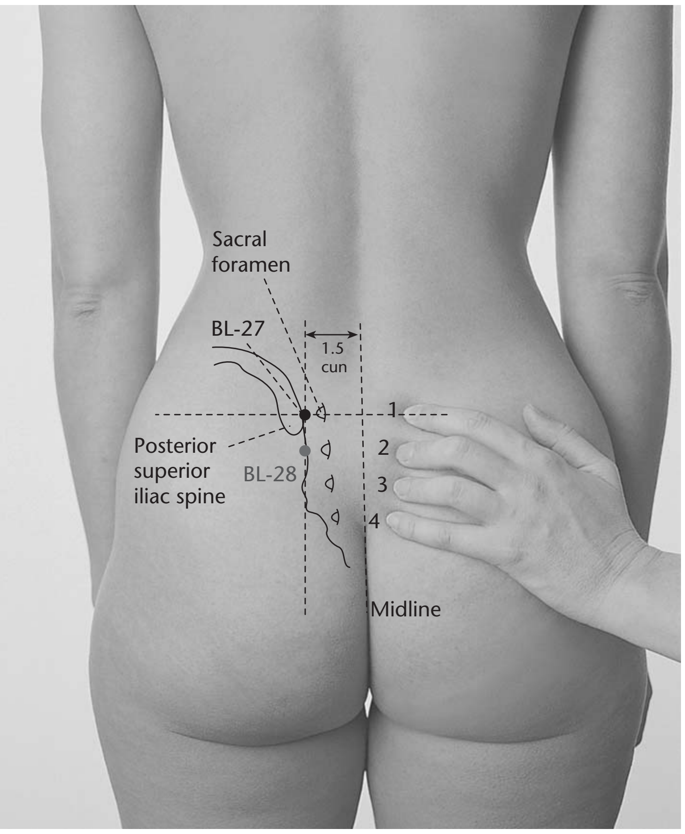
Sacral
foramen
BL-27
1.5
cun
1
2
3
4
Posterior
superior
iliac spine
BL-28
Midline

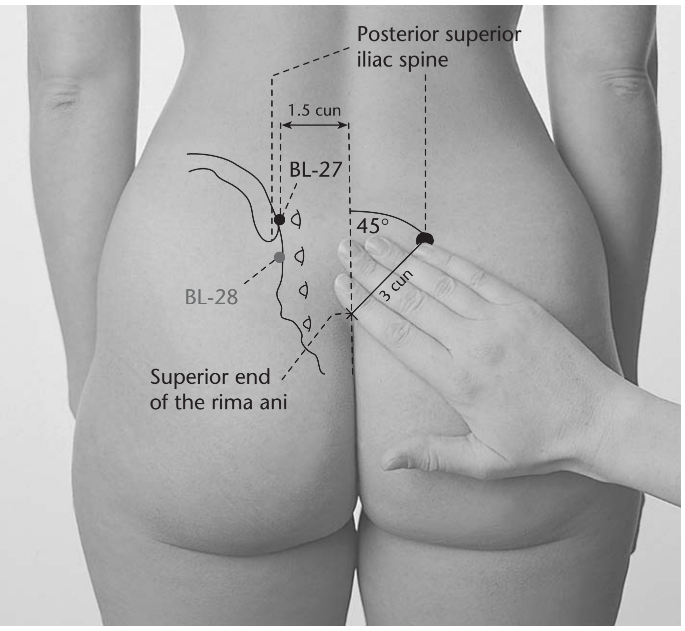
Posterior superior
iliac spine
1.5 cun
BL-27
45°
3 cun
BL-28
Superior end
of the rima ani

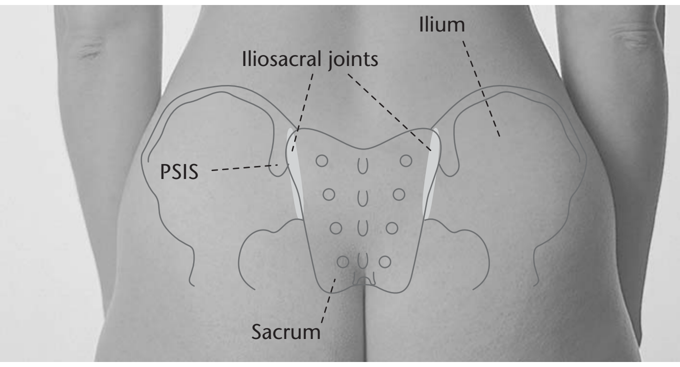
Ilium
Iliosacral joints
PSIS
Sacrum

Location

1.5 cun lateral to the posterior midline, on the level of the first sacral foramen.

How to find

Quick method: Locate the posterior superior iliac spine (PSIS ➞ 3.4.3): from the rima ani, palpate 3 cun (1 handbreadth) in a superior and lateral direction at a 45° angle, until you can feel a distinct bony ridge (often superficially visible by a dimple). **BL-27** is located slightly superiorly and medially to the PSIS, 1.5 cun lateral to the midline, on the level of the first sacral foramen. For more detail on orientation in the lumbar and sacral region ➞ 3.4.3, 3.4.4.

Located on the same level are ➞ **BL-31** (over the 1st sacral foramen) and ➞ **BL-28** (slightly below and medial to the PSIS at the level of the 2nd sacral foramen).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Tonifies the Small Intestine Qi
- Drains Dampness and Damp-Heat
- Regulates the water passages

Special features

Back-*shu* point of the Small Intestine

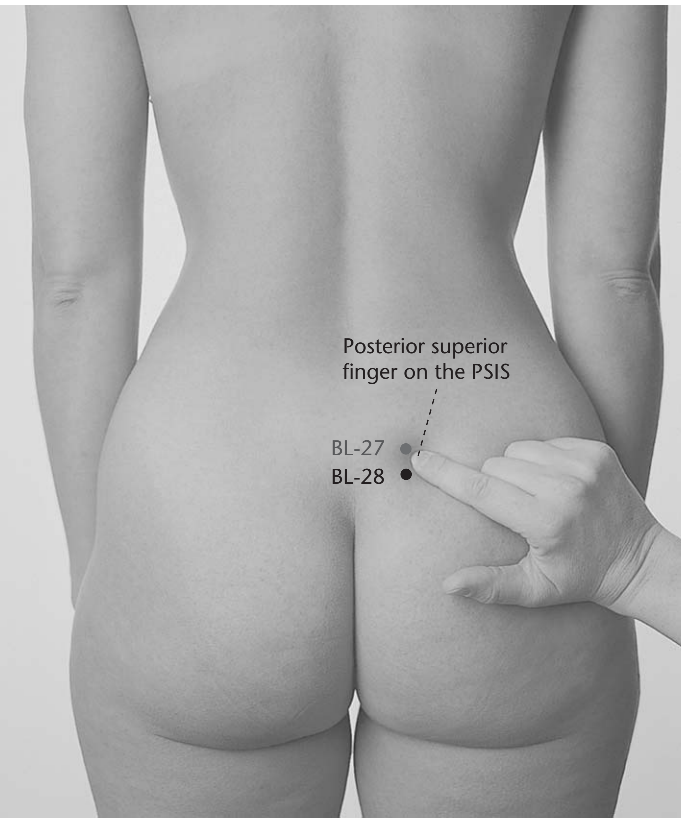
Posterior superior
finger on the PSIS
BL-27
BL-28

BL-28 *Bladder* Shu PANGGUANGSHU

Location

1.5 cun lateral to the posterior midline, on the level of the second sacral foramen.

How to find

Quick method: Locate the posterior superior iliac spine (PSIS → 3.4.3): from the rima ani, palpate 3 cun (1 handbreadth) in a superior and lateral direction at a 45° angle, until you can feel a distinct bony ridge (often superficially visible by a dimple). **BL-28** is located slightly inferiorly and medially to the PSIS, 1.5 cun lateral to the midline on the level of the 2nd sacral foramen. For more detail on orientation in the lumbar and sacral region → 3.4.3, 3.4.4.

Located on the same level are → **BL-32** (over the 2nd sacral foramen) and → **BL-53** (3 cun lateral to the midline). → **BL-27** is located slightly superiorly and medially to the PSIS on the level of the 1st sacral foramen.

Needling

Vertically 0.5–1.5 cun. For local disorders, slightly oblique insertion towards the iliosacral joint.

Actions/Indications

- Regulates the Bladder and the Lower Burner, clears Damp-Heat from the Lower Burner, removes stagnation and transforms masses
- Benefits the lumbar region and the legs

Special features

Back-*shu* point of the Bladder

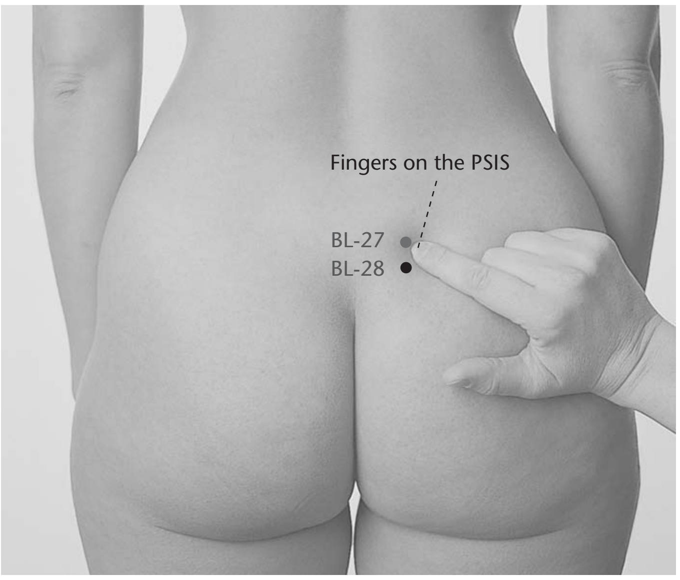
Fingers on the PSIS
BL-27
BL-28

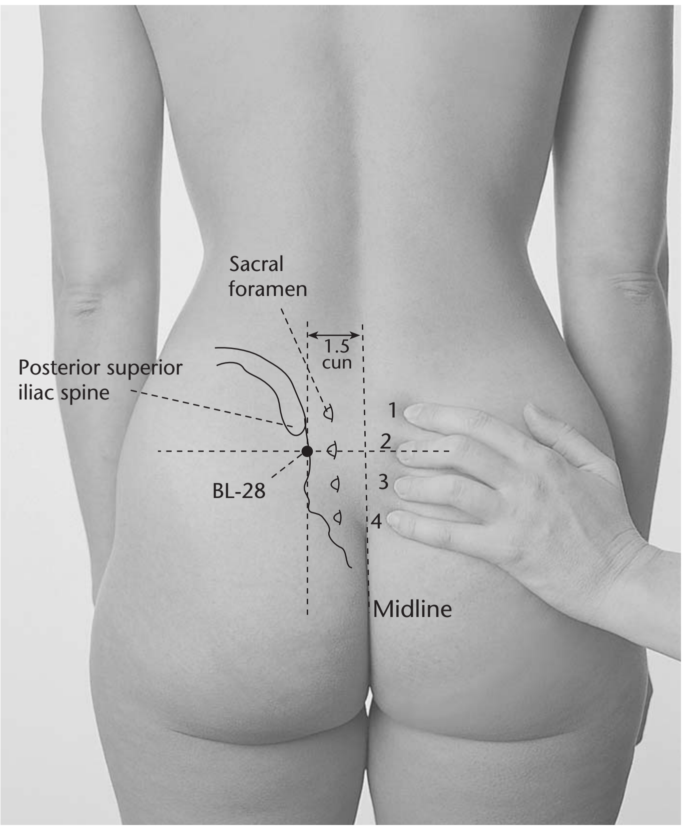
Sacral foramen
1.5 cun
Posterior superior iliac spine
1
2
3
4
BL-28
Midline

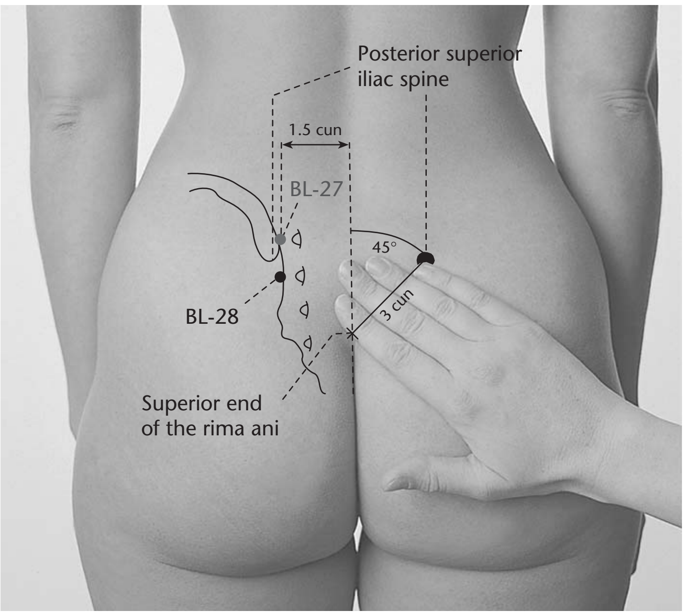
Posterior superior iliac spine
1.5 cun
BL-27
45°
3 cun
BL-28
Superior end of the rima ani

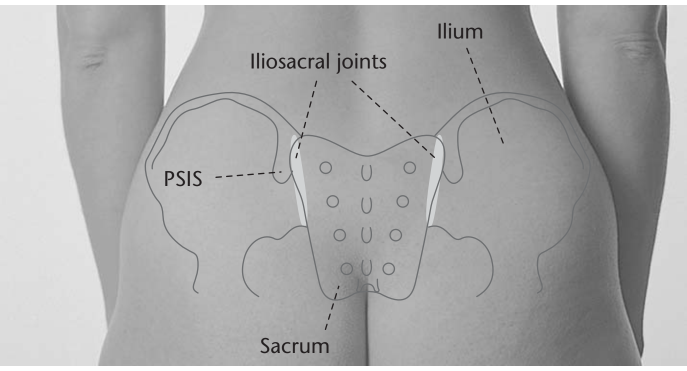
Ilium
Iliosacral joints
PSIS
Sacrum

Ch04.7a-F10028.qxd 2/23/08 3:54 PM Page 279

Mid Spine Shu ZHONGLÜSHU BL-29

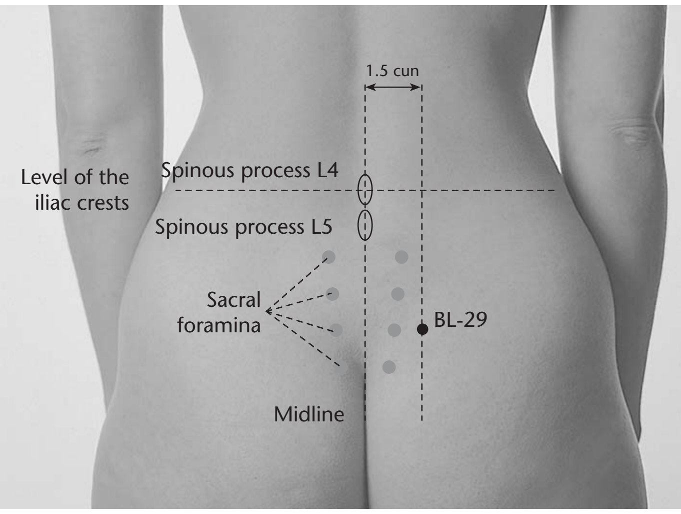
1.5 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-29
Midline

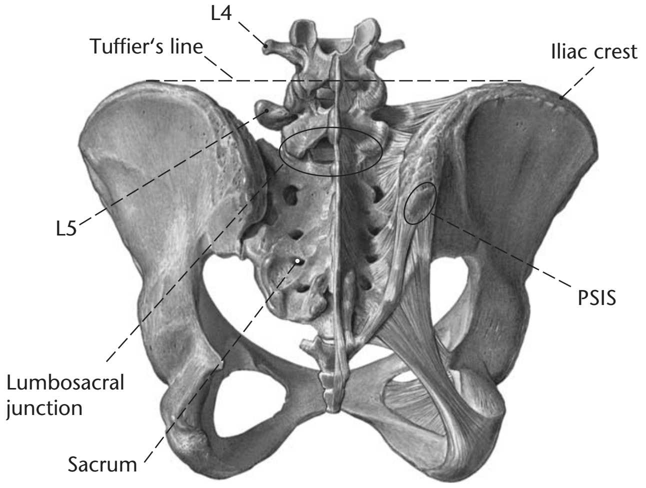
L4
Tuffier's line
Iliac crest
L5
PSIS
Lumbosacral junction
Sacrum

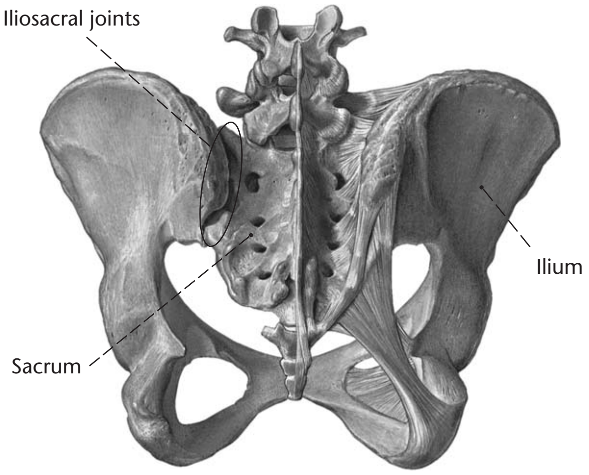
Iliosacral joints
Ilium
Sacrum

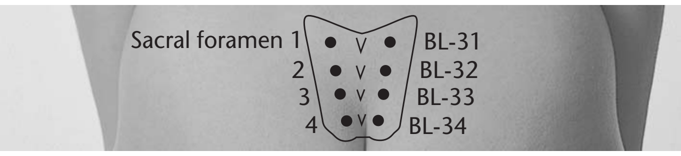
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

Location

1.5 cun lateral to the posterior midline, on the level of the 3rd sacral foramen.

How to find

For more detail on orientation in the lumbar and sacral region, see ➞ 3.4.3, 3.4.4. First of all find L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in an inferior direction for two further sacral processes and locate the 3rd sacral foramen in a depression less than 1 cun lateral to the 3rd process. **BL-29** is located on the same level (3rd sacral foramen) and 1.5 cun lateral to the midline.

Also located on this level is ➞ **BL-33** (over the 3rd sacral foramen).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the lower back and the Kidneys
- Expels Cold
- Regulates the Lower Burner and stops diarrhoea

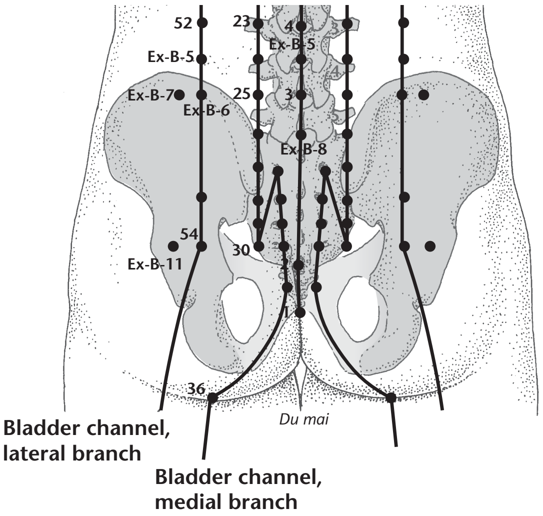
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
Ex-B-6
25
3
Ex-B-8
54
Ex-B-11
30
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

BL-30 *White Ring* Shu BAI HUAN SHU

Location

1.5 cun lateral to the posterior midline, on the level of the 4th sacral foramen.

How to find

For more detail on orientation in the lumbar and sacral region, see ➞ 3.4.3, 3.4.4. First, find L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in an inferior direction for three further sacral processes and locate the 4th sacral foramen in a depression less than 1 cun lateral to the 4th process. **BL-30** is located on the level of the 4th foramen, 1.5 cun lateral to the midline.

Located on the same level are ➞ **BL-34** (over the 4th sacral foramen) and ➞ **BL-54** (3 cun lateral to the midline).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Strengthens the lower back and the knees
- Eliminates Damp-Heat from the Lower Burner
- Regulates menstruation

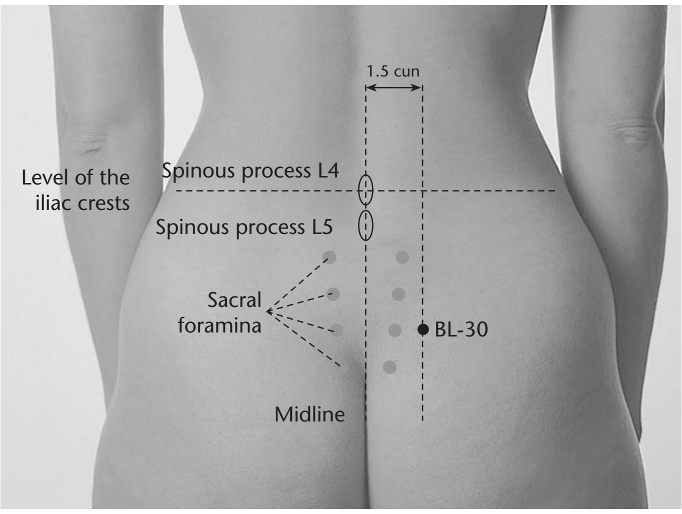
1.5 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-30
Midline

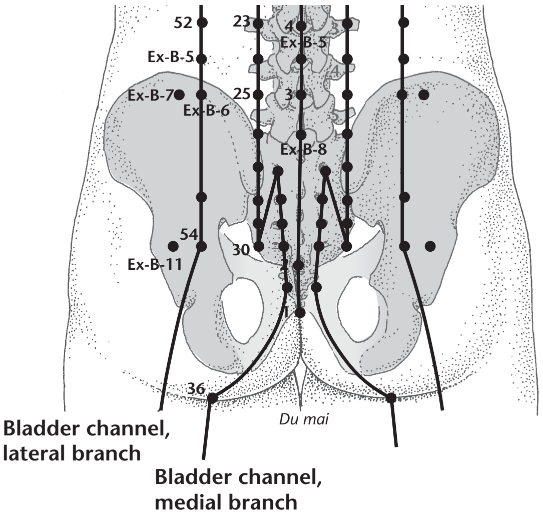
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

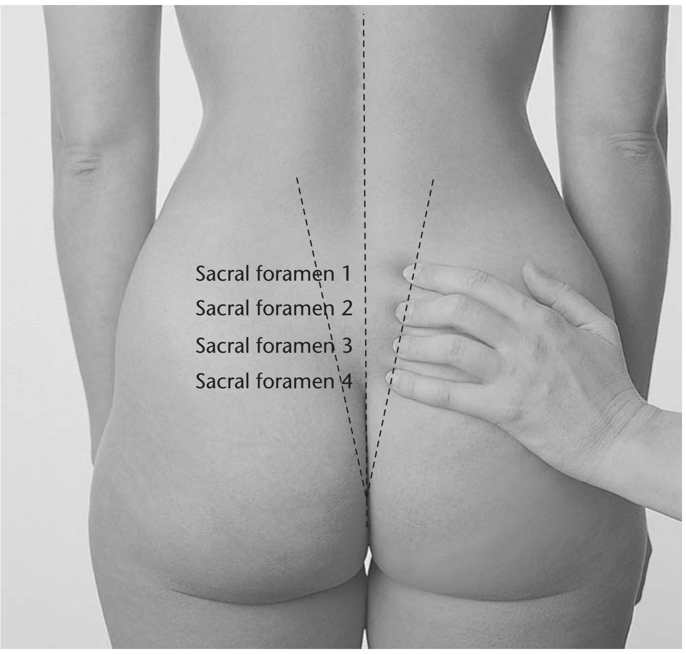
Sacral foramen 1
Sacral foramen 2
Sacral foramen 3
Sacral foramen 4

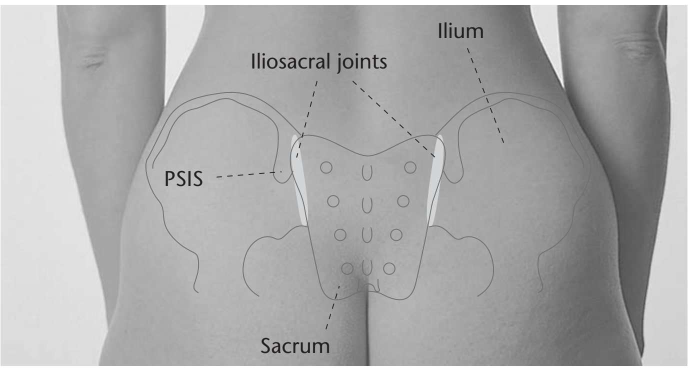
Ilium
Iliosacral joints
PSIS
Sacrum

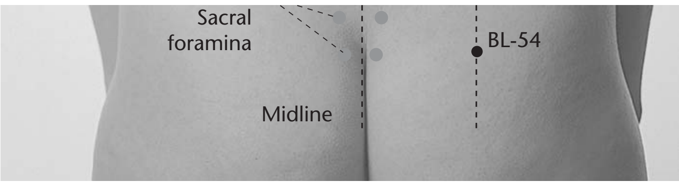
Sacral foramina
BL-54
Midline

Eight **Liao** BALIAO BL 31–BL 34

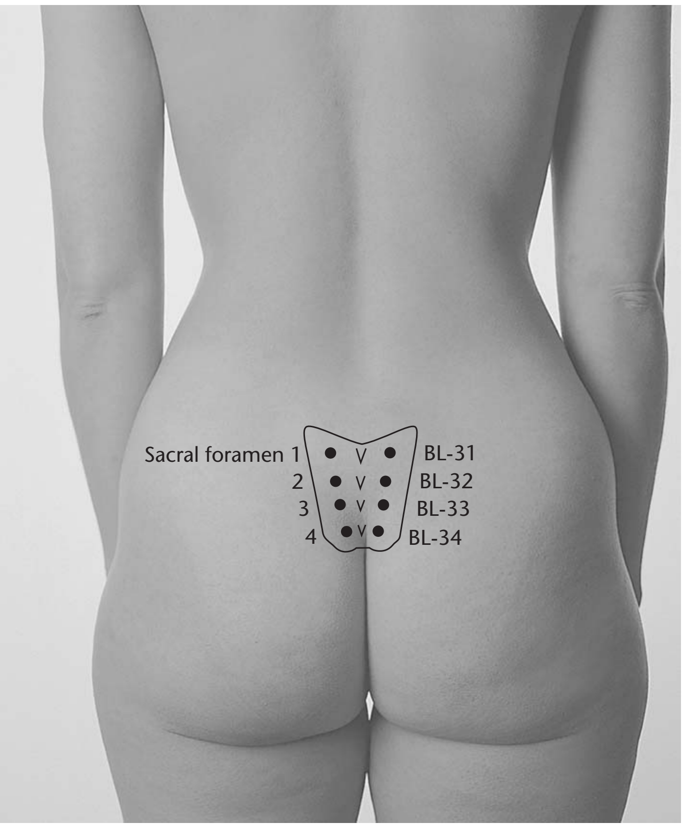

These points are described together because of their similar actions and indications.

BL-31 (*shangliao*) 'Upper Crevice', 1st sacral foramen
BL-32 (*ciliao*) 'Second Crevice', 2nd sacral foramen
BL-33 (*zhongliao*) 'Middle Crevice', 3rd sacral foramen
BL-34 (*xialiao*) 'Lower Crevice', 4th sacral foramen

Location

BL-31: Over the 1st sacral foramen
BL-32: Over the 2nd sacral foramen
BL-33: Over the 3rd sacral foramen
BL-34: Over the 4th sacral foramen

How to find

For more detail on orientation in the lumbar and sacral region, → 3.4.3, 3.4.4. First, locate L5. The next palpable structure in an inferior direction is the most superior sacral spinous process. Palpate in a lateral direction on the level of the sacral processes for the depressions of the sacral foramina. **Or:** Locate the 4th sacral foramen on the level of the superior end of the rima ani (sometimes a bit more superiorly or inferiorly). Place the little finger on the 4th foramen, then the ring, middle and index fingers on a slightly V-shaped line in a superior direction, palpating for the depressions of the sacral foramina. With the fingers in this position, the index finger will be resting on the 1st foramen (**BL-31**), the middle finger on the 2nd foramen (**BL-32**), the ring finger on the 3rd foramen (**BL-33**) and the little finger on the 4th foramen (**BL-34**). Located on the same level as **BL-31–BL-34** are the points → **BL-27–BL-30** on the inner branch of the BL channel (1.5 cun lateral to the midline) as well as → **BL-53** (on the level of **BL-32**) and → **BL-54** (on the level of **BL-34**), both 3 cun lateral to the midline. Also on the level of **BL-34** is → **Ex-B-11** (3.5 cun lateral to the midline).

Needling

Vertically 0.7–1.5 cun. Caution: Pregnancy (especially reducing needling techniques are contraindicated); exception: to promote labour.

Actions/Indications

- Regulates the Lower Burner, promotes urination (all eight points) (**BL-32** and **BL-33** have the strongest action in cases of urinary disorders, while **BL-34** has the strongest action in cases of genital disorders)
- Strengthens the Kidneys and Essence (*jing*)
- Benefits the Intestines (**BL-34** has the strongest action in cases of intestinal disorders)
- Promotes labour (for example, use **BL-32** with electro-acupuncture)
- Benefits the lumbar region (**BL-32** and **BL-33** have the strongest action)

Special features

According to some authors, meeting point with the G.B. channel. Clinically, **BL-32** is the most commonly used point owing to its widest range of indications. Electro-acupuncture at **BL-32** is useful for alleviating pain during labour: after needling this point, bend the handle of the needle in a superior direction at a 90° angle and tape it to the skin. Now attach the electro-stimulation. This method allows the woman to lie down if desired.
Moxibustion of the *baliao* is indicated for difficult urination or defaecation due to Kidney deficiency in elderly people.

BL-35 *Meeting of Yang* HUIYANG

Location

0.5 cun lateral to the posterior midline, on the level of the tip of the coccyx.

How to find

Locate the coccyx superior to the anus. In contrast to the sacrum, the coccyx is movable. **BL-35** is located 0.5 cun lateral to the midline on the level of the tip of the coccyx.

→ **Du-2** is located on the midline, on the level of the sacral hiatus (→ 3.4.4) and superior to **BL-35** and the tip of the coccyx.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Damp-Heat from the Lower Burner
- Treats haemorrhoids and rectal prolapse
- Benefits the coccyx

Special features

Especially in patients with a weakened immune system, this point should be needled only if strictly necessary and after using a disinfecting wipe three times.

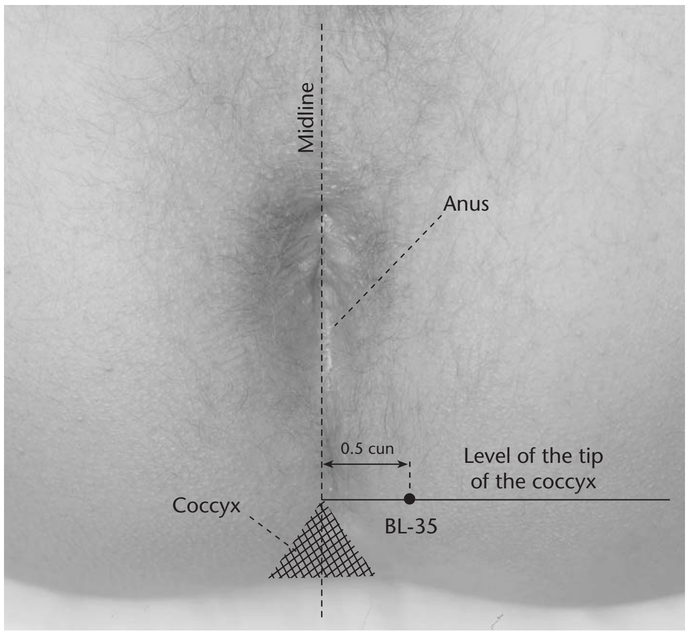
Midline
Anus
0.5 cun
Level of the tip of the coccyx
Coccyx
BL-35

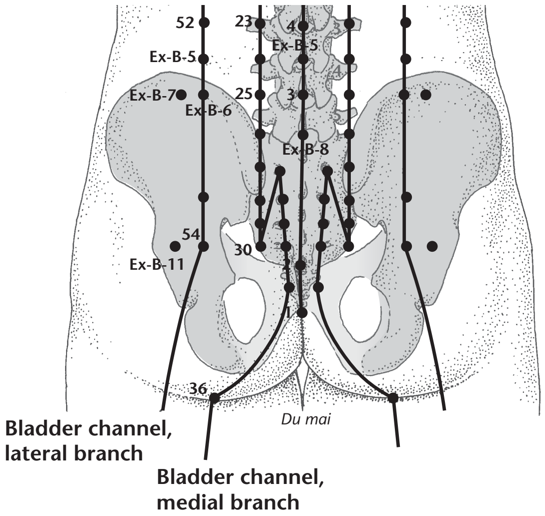
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

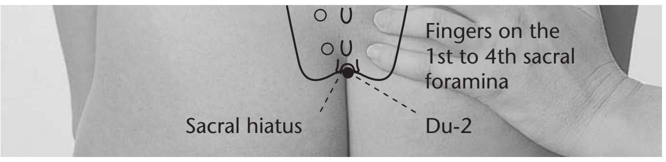
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

Hold and Support CHENGFU BL-36

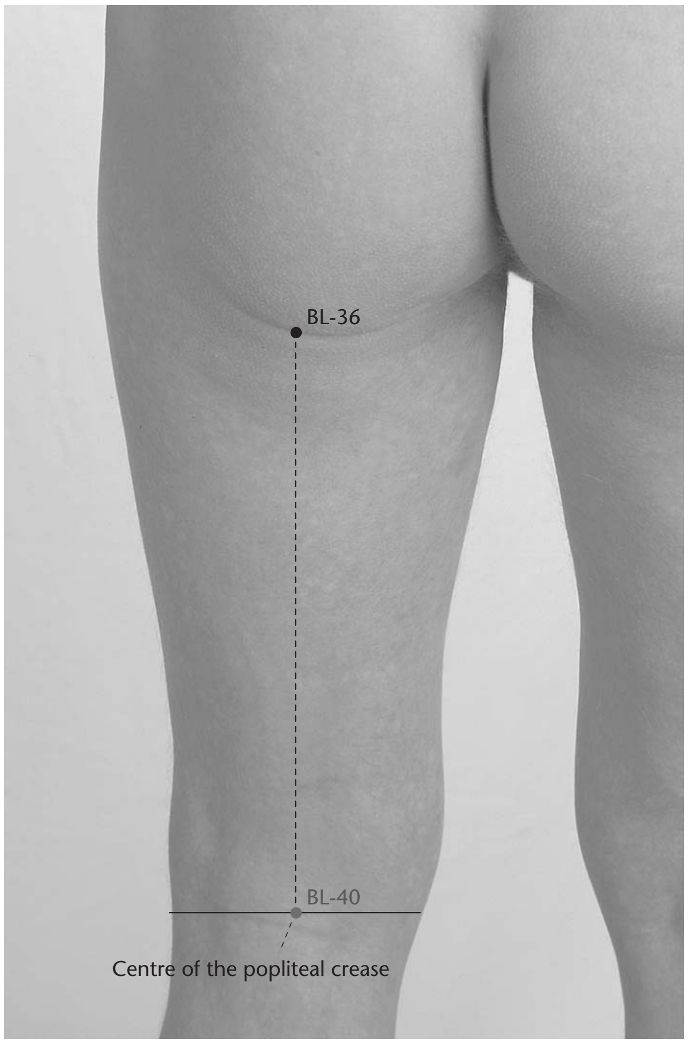
BL-36
BL-40
Centre of the popliteal crease

Location

In the gluteal crease, superior to the midpoint of the popliteal crease (→ **BL-40**).

How to find

Locate the midpoint of the popliteal crease (in obese patients, use bony/muscular structures to find the midpoint). **BL-36** is located directly vertically above the midpoint of the popliteal crease, in the gluteal crease at the transition between the buttocks and the posterior aspect of the thigh (palpate for a depression).

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel, alleviates pain
- Relaxes the tendons
- Regulates the Lower Burner and treats haemorrhoids

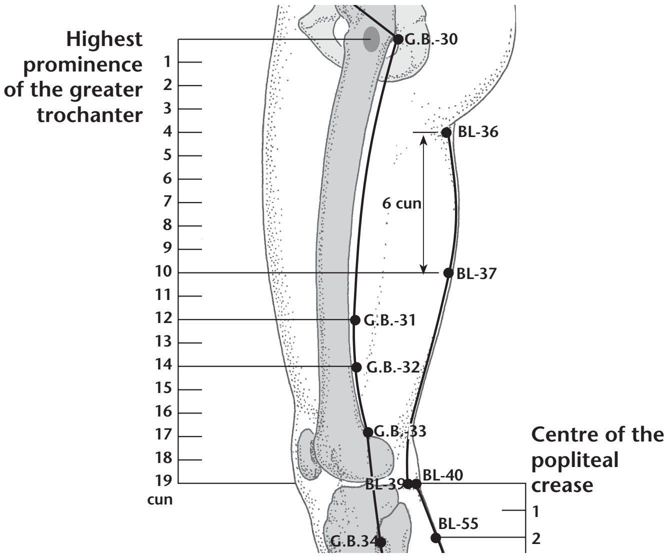
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
BL-55
G.B.34
Centre of the popliteal crease
1
2

BL-37 *Gate of Abundance* YINMEN

Location

6 cun distal to → **BL-36** (gluteal crease), on a line connecting → **BL-36** and → **BL-40** (in the popliteal crease), in a gap in the musculature.

How to find

First, locate → **BL-36** in the centre of the gluteal crease and, from there, palpate 6 cun along the posterior aspect of the thigh in an inferior direction towards the centre of the popliteal crease (→ **BL-40**). This is the location of **BL-37**, which in slim patients can be located where the long head of the biceps femoris muscle and the semitendinosus muscle converge (usually 8 cun superior to the centre of the popliteal crease. **Or:** Spreading hands technique (→ 2.3.3 or elastic tape): Place the little fingers on the centre of the popliteal crease (→ **BL-40**) and the centre of the gluteal crease (→ **BL-36**). Locate **BL-37** 1–2 cun proximal to the midpoint of the distance, in a depression between the two muscle bellies.

Needling

Vertically 0.5–2 cun

Actions/Indications

- Relaxes the tendons
- Opens the channel and *luo* vessels
- Benefits the lower back

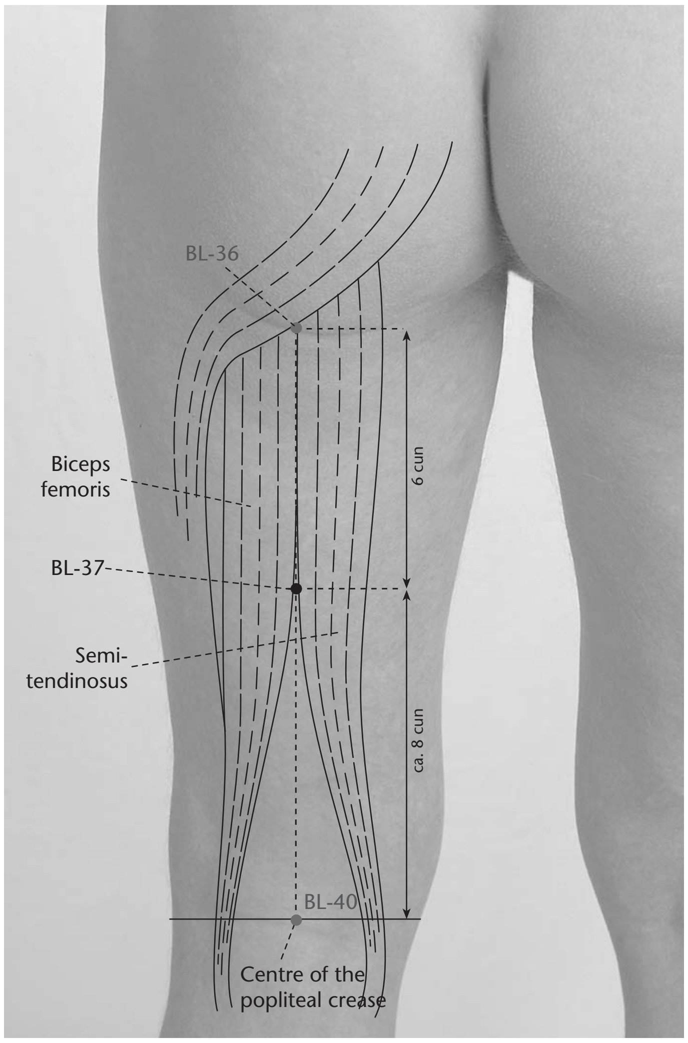
BL-36
Biceps femoris
BL-37
Semi-tendinosus
6 cun
ca. 8 cun
BL-40
Centre of the popliteal crease

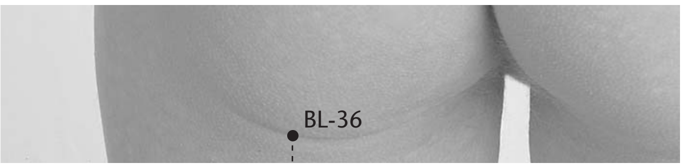

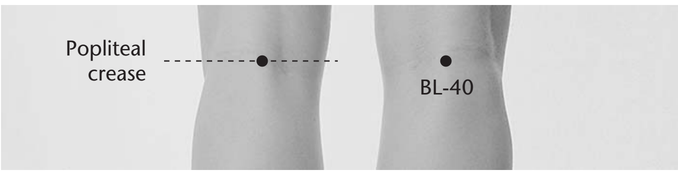
Popliteal crease
BL-40

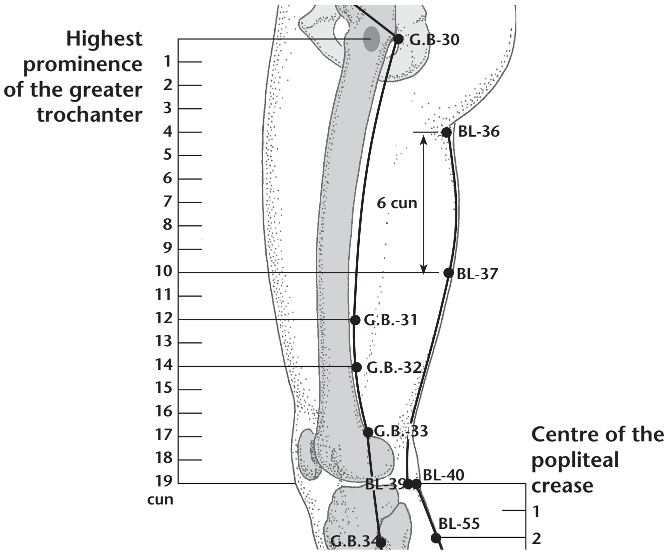
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-39
BL-40
BL-55
G.B.34
Centre of the popliteal crease
1
2

Floating Cleft **FUXI** **BL-38**

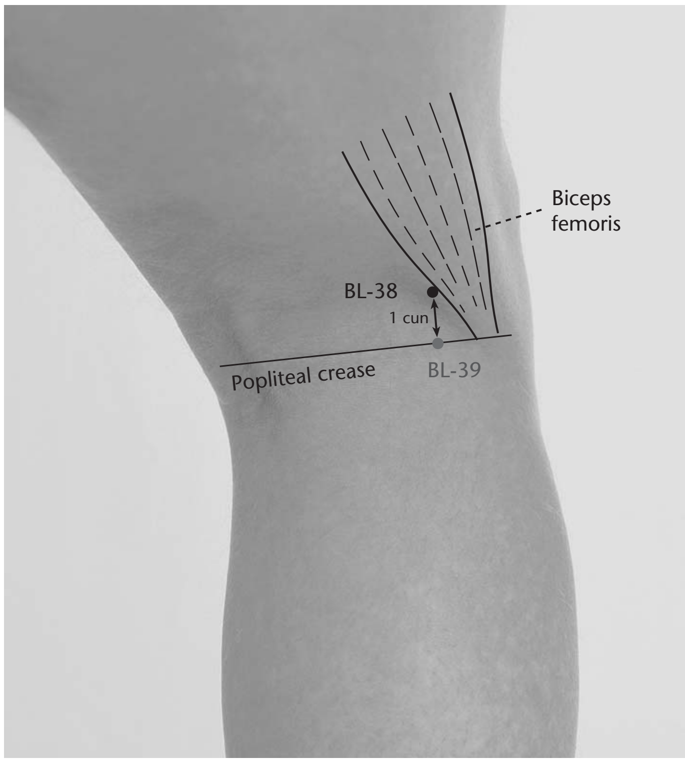

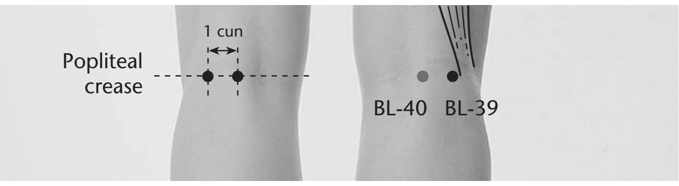

Location

On the laterodorsal aspect of the knee, 1 cun superior and lateral to the centre of the popliteal crease (→ **BL-40**), medial to the biceps femoris muscle, or 1 cun proximal to → **BL-39**.

How to find

To locate the popliteal crease, ask the patient to flex their knee slightly and palpate for the knee joint space. By flexing the knee against resistance, the tendon of the biceps femoris muscle will become more pronounced and visible. Locate **BL-38** 1 cun superior to the knee joint on the medial border of the tendon.

→ **BL-39** is located 1 cun distal to **BL-38** on the level of the popliteal crease.

Needling

Vertically 1–2 cun

Actions/Indications

- Relaxes the muscles and tendons
- Opens the channel and *luo* vessels
- Clears Heat (from the Small Intestine)

BL-39 *Outside of the Crook* WEIYANG

Location

At the lateral end of the popliteal crease, on the medial side of the tendon of the long head of the biceps femoris muscle, 1 cun lateral to ➞ **BL-40** (in the centre of the popliteal crease).

How to find

This point is best located with the patient's knee slightly flexed. Find the midpoint of the popliteal crease (➞ **BL-40**) and locate **BL-39** approximately 1 cun lateral in a depression medial to the tendon of the long head of the biceps femoris muscle. Located on the same level is ➞ **BL-40** at the centre of the popliteal crease.

Needling

Vertically 0.5–1.5 cun. Caution: N. fibularis communis.

Actions/Indications

- Harmonises the Triple Burner and regulates the water passages
- Opens the channel and the *luo* vessels, alleviates pain

Special features

Lower *he*-sea point of the Triple Burner

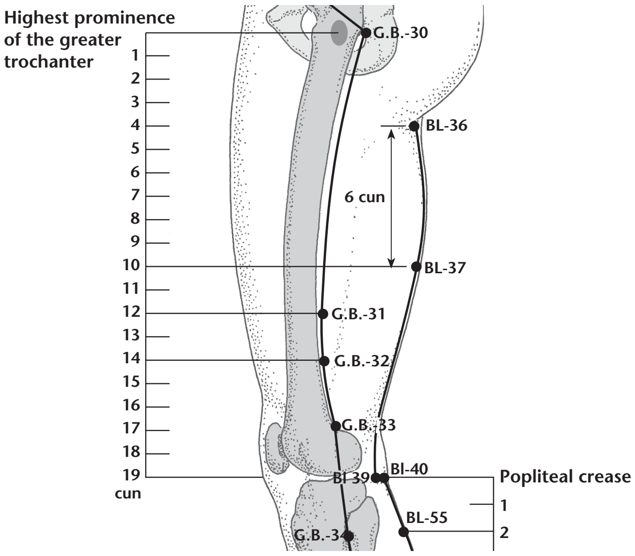
Highest prominence of the greater trochanter
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
Bl-40
Bl 39
Popliteal crease
1
BL-55
2
G.B.-34

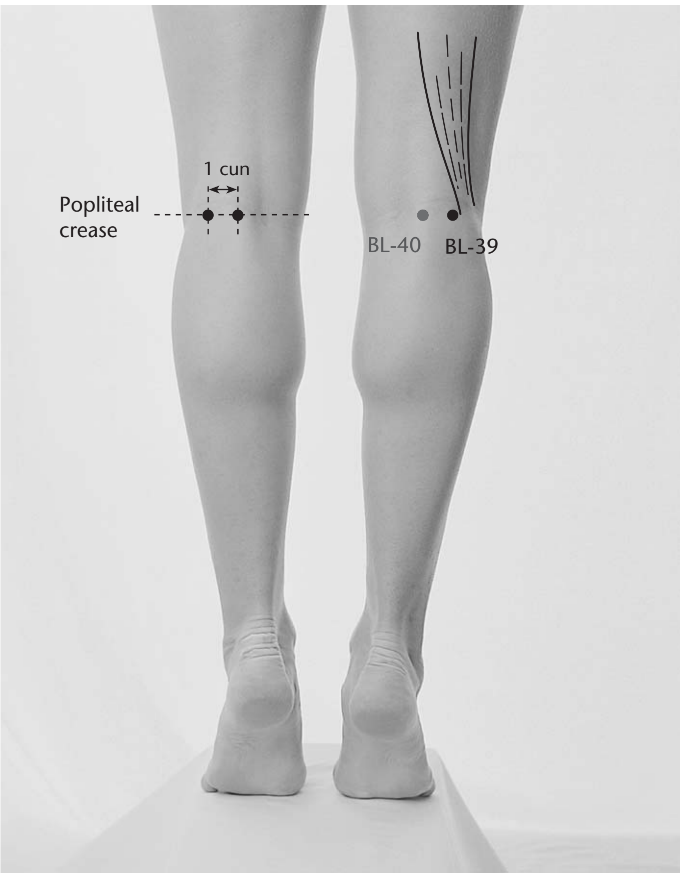
1 cun
Popliteal crease
BL-40
BL-39

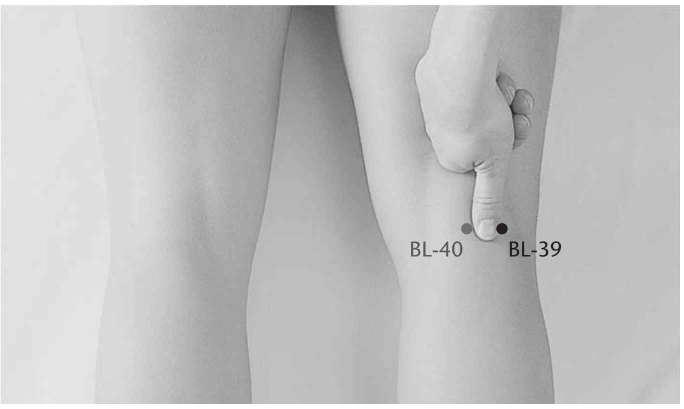

Middle of the Crook WEIZHONG BL-40

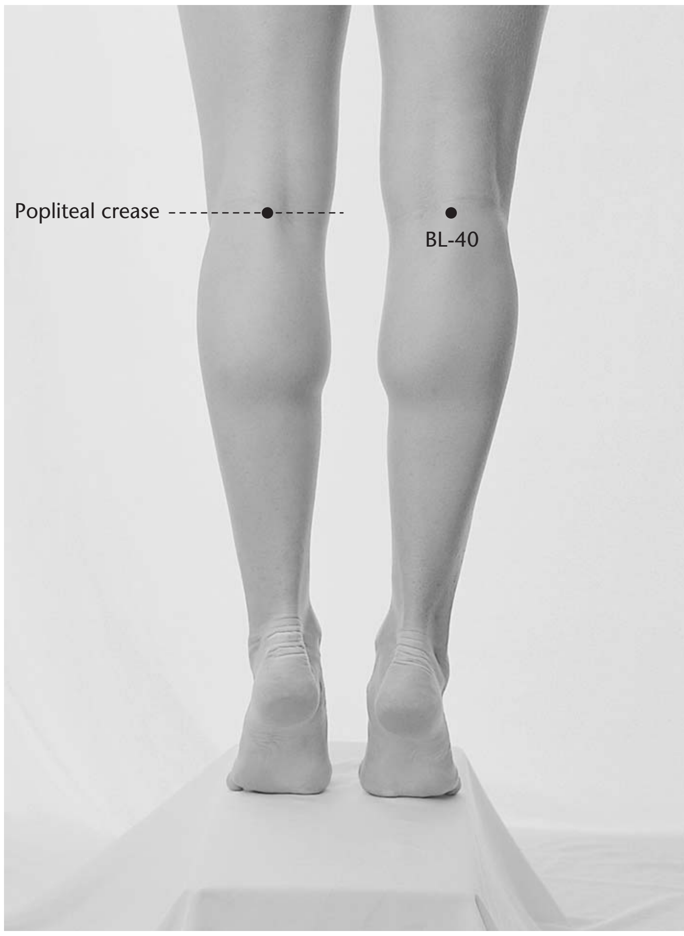

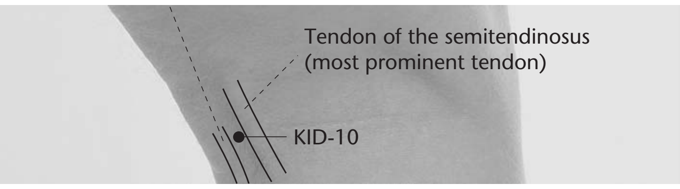

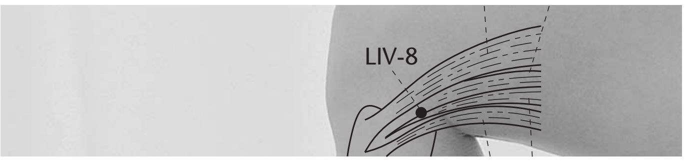

Location

In the centre of the popliteal crease, between the tendons of the biceps femoris and semitendinosus muscles.

How to find

This point is best located with the patient's knee slightly flexed. Locate the centre of the popliteal crease (a pulse may be palpable) and there locate **BL-40**.

Located on the same level are → **BL-39** (1 cun laterally), → **KID-10** (more medially, between the tendons of the semimembranosus and semitendinosus muscles) and → **LIV-8** (more medially, anterior to the tendons of the semimembranosus and semitendinosus muscles).

Needling

Vertically 0.5–1.5 cun. Caution: Popliteal nerve, artery and vein lie deep to this point. Use moxibustion with caution (contraindicated according to some authors). To clear Heat and eliminate stasis: bleed the superficial veins, possibly followed by brief cupping. Caution: Only use for excess conditions and in constitutionally robust patients. **BL-40** is also indicated as a distal point for the lower back: needle with the patient standing (for safety and balance, use a treatment couch or chair as a support for the hands) and ask them to move their lower back gently while manipulating the needle (only short needle retention). When needling this point with the patient in a prone position, place a supporting roll or pillow under the ankles to relax the popliteal crease.

Actions/Indications

- Clears (summer) Heat, stops vomiting and diarrhoea
- Cools the Blood
- Benefits the lower back and knees, opens the channel and *luo* vessels, alleviates pain

Special features

He-sea point, Earth point, lower *he*-sea point of the Bladder, Ma Dan Yang Heavenly Star point, Gao Wu command point of the lumbar region. Important distal point for the lower back, point with a wide range of indications.

BL-41 *Attached Branch* FUFEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2).

How to find

After locating the spinous process of C7 (➞ 3.4.1), count down 2 spinous processes to the lower border of the spinous process of T2. Locate **BL-41** 3 cun lateral from the midline (with the shoulders relaxed and the arms hanging loosely down, 3 cun corresponds on this level to the distance from the midline to the medial border of the scapula). Note: From **BL-41** to ➞ **BL-54**, all points on the BL channel are located on its outer branch.

Located on the same level are a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline), ➞ **BL-12** (1.5 cun lateral to the midline) and ➞ **S.I.-13** (more laterally, on the scapula).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors such as Wind and Cold
- Opens the channel and the *luo* vessels, alleviates pain

Special features

Meeting point with the S.I. channel

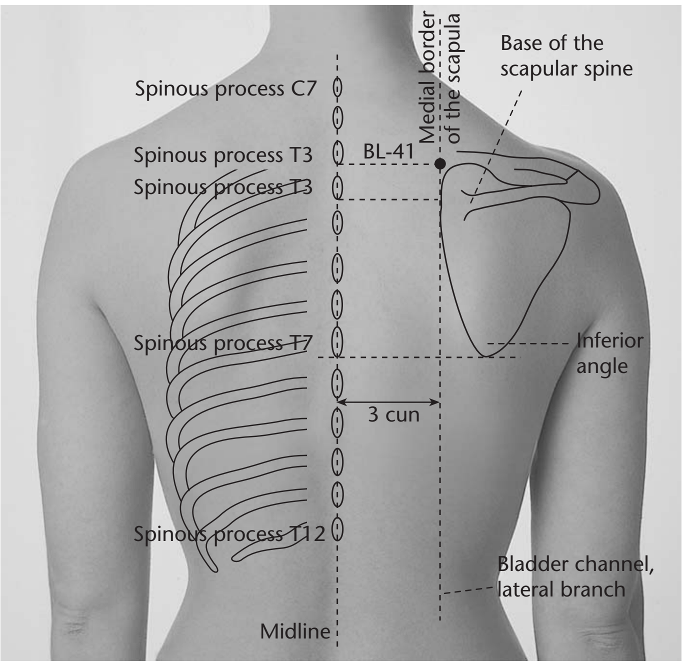
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-41
Spinous process T3
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

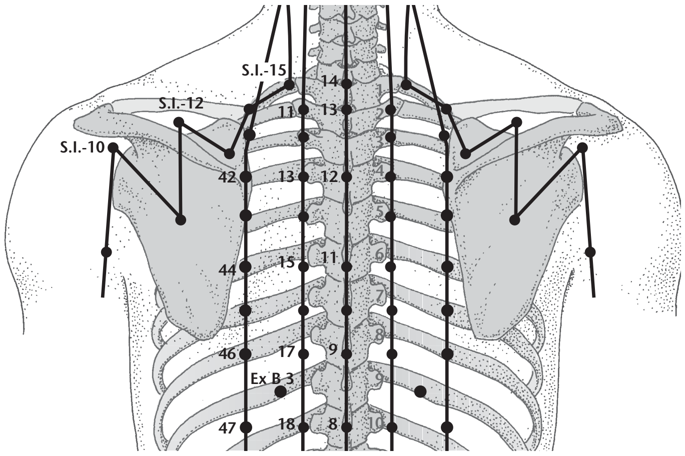
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex B 3
47
18
8

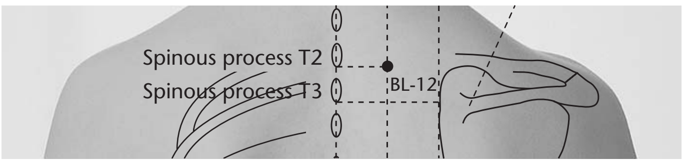
Spinous process T2
Spinous process T3
BL-12

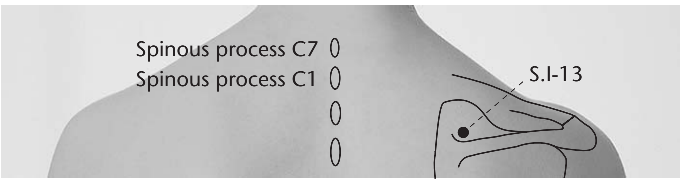
Spinous process C7
Spinous process C1
S.I-13

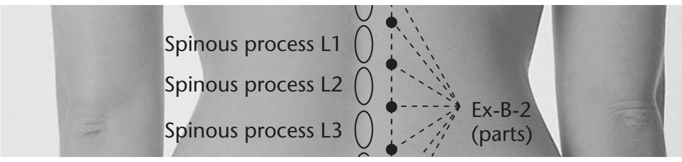
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Door of the Corporeal Soul POHU BL-42

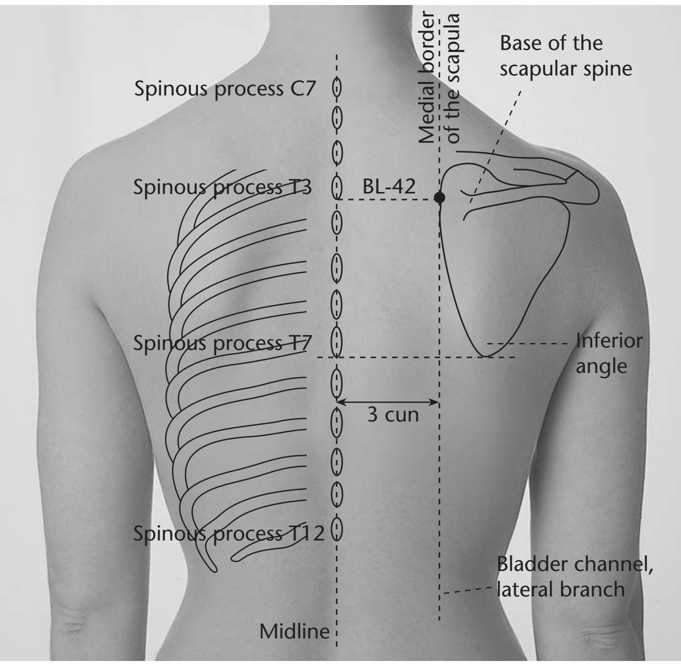
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
BL-42
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

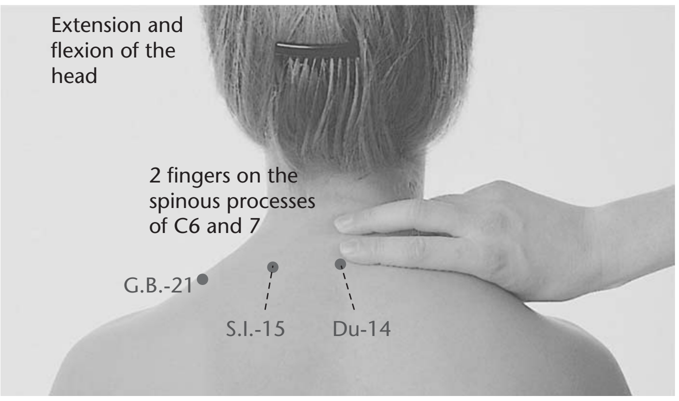
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

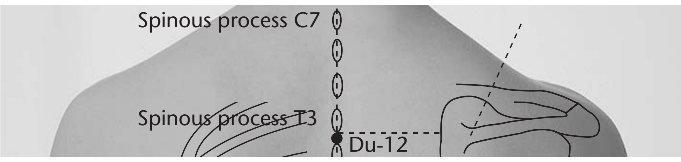
Spinous process C7
Spinous process T3
Du-12

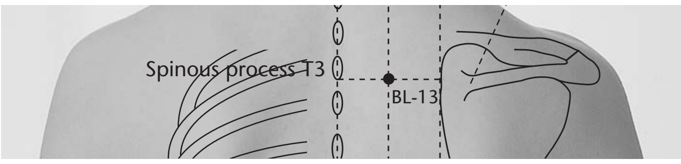
Spinous process T3
BL-13

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 3rd thoracic vertebra (T3).

How to find

After locating the spinous process of C7 (→ 3.4.1), count down 3 spinous processes to the lower border of the spinous process of T3. Located on this level and 3 cun lateral to the midline is **BL-42.**

Or: With the patient seated and their shoulders in a relaxed position, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable structure (→ 3.4.2). Locate **BL-42** on the level of the lower border of the spinous process of T3, 3 cun lateral to the midline. With the patient's shoulders relaxed, 3 cun corresponds on this level to the distance from the midline to the medial border of the scapula.

Located on the same level are → **Du-12** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-13** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the Lung
- Alleviates wheezing and coughing, calms the corporeal soul (*po*)
- Opens the channel and alleviates pain
- Clears Heat from the Lung

BL-43 *Vital Region* Shu GAOHUANG

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 4th thoracic vertebra (T4).

How to find

After locating the spinous process of C7 (➞ 3.4.1), count down 4 spinous processes to the lower border of the spinous process of T4. Located on this level and 3 cun lateral to the midline is **BL-43. Or:** Locate **BL-43** by using the spinous process of T3 (generally, on the level of the easily palpable medial end of the scapular spine, 3.4.2) as a reference point.

Located on the same level are a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-14** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Moxibustion is often used on this point. Caution: Pneumothorax.

Actions/Indications

- Tonifies and nourishes the five *zang*-Organs: Lung, Heart, Kidneys, Stomach and Spleen
- Nourishes Yin, clears Heat
- Calms the *shen*
- Strengthens the Original Qi (*yuan qi*)
- Eliminates Phlegm

Special features

Important point for deficiency syndromes (classic indication)

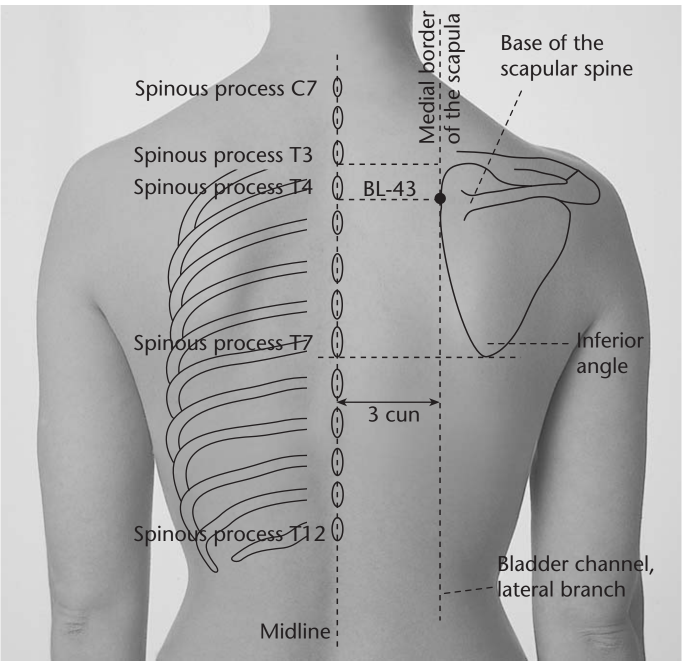
Medial border of the scapula
Base of the scapular spine
Spinous process C7
Spinous process T3
Spinous process T4
BL-43
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

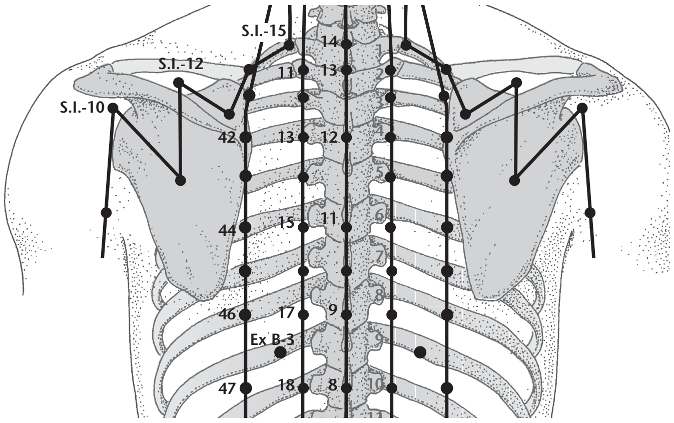
S.I.-15
14
S.I.-12
11
13
1
2
3
S.I.-10
42
13
12
4
5
44
15
11
6
7
8
46
17
9
Ex B-3
9
47
18
8
10
11

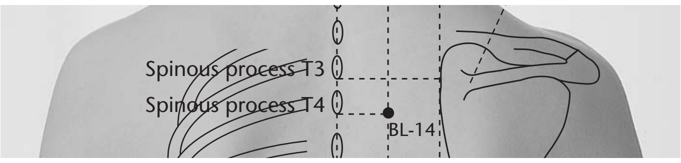
Spinous process T3
Spinous process T4
BL-14

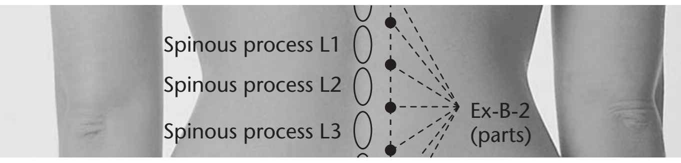
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Spirit Hall SHENTANG BL-44

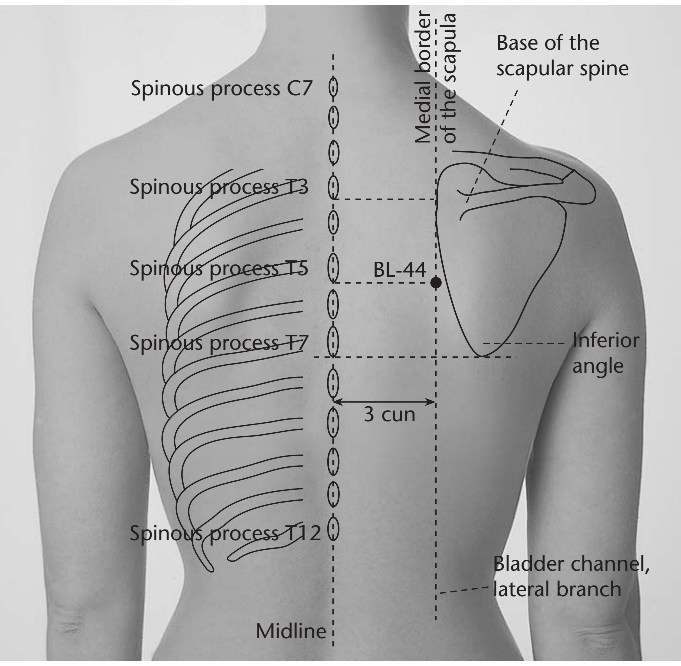
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T5
BL-44
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

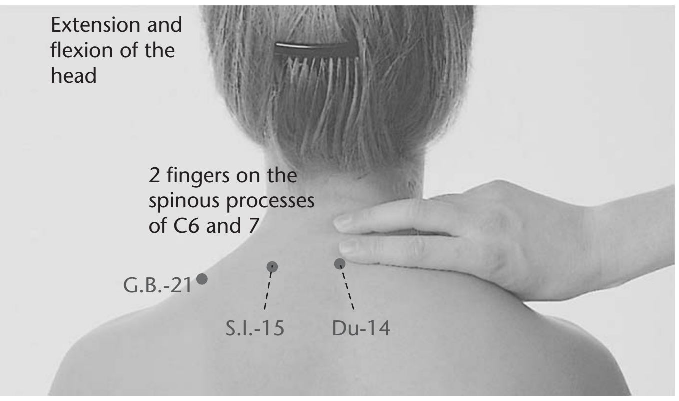
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

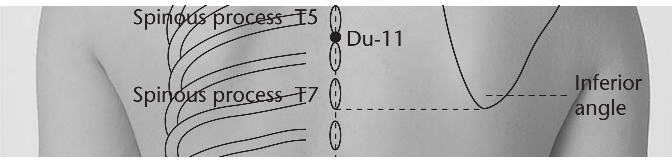
Spinous process T5
Du-11
Spinous process T7
Inferior angle

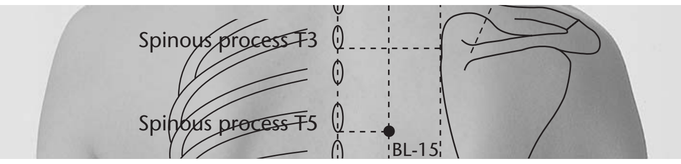
Spinous process T3
Spinous process T5
BL-15

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 5th thoracic vertebra (T5).

How to find

After locating the spinous process of C7 (➞ 3.4.1), count down 5 spinous processes to the lower border of the spinous process of T5. Located on this level and 3 cun lateral to the midline is **BL-44. Or:** With the patient seated and their shoulders in a relaxed position, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable structure (➞ 3.4.2). Locate **BL-44** by counting down to the level of the lower border of the spinous process of T5 and measure 3 cun lateral to the midline.

Located on the same level are ➞ **Du-11** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-15** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the Upper Burner, unbinds the chest
- Opens the channel and alleviates pain

Ch04.7b-F10028.qxd 2/22/08 5:00 PM Page 292

BL-45 *Cry of Pain* YIXI

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 6th thoracic vertebra (T6).

How to find

After locating the spinous process of C7 (➞ 3.4.1), count down 6 spinous processes to the lower border of the spinous process of T6. Located on this level and 3 cun lateral to the midline is **BL-45. Or:** With the patient seated and their shoulders in a relaxed position, the spinous process of T7 is generally on the level of the inferior angle of the scapula (➞ 3.4.2). Locate **BL-45** on the level of the spinous process of T6, 3 cun lateral to the midline. Located on the same level are ➞ **Du-10** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-16** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Expels pathogenic factors, especially Wind and Heat
- Descends the Lung Qi and unbinds the chest
- Strengthens the Qi and Blood and alleviates pain

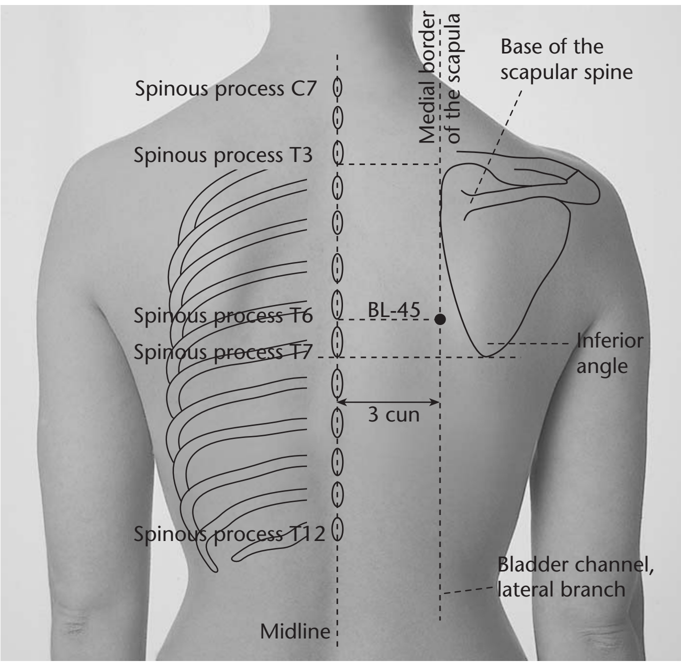
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T6
BL-45
Spinous process T7
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

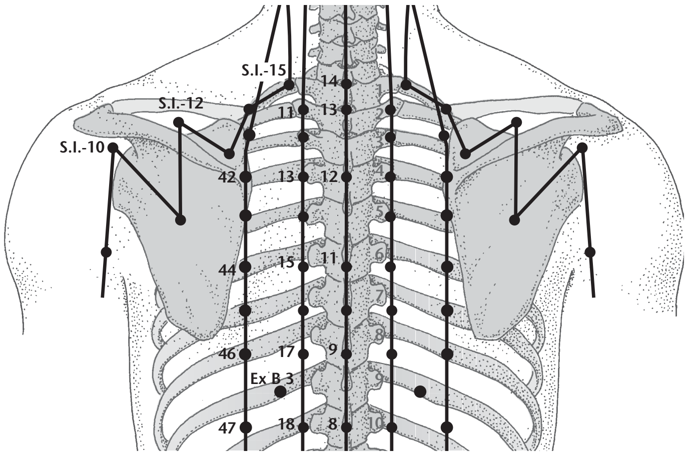
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex B 3
47
18
8

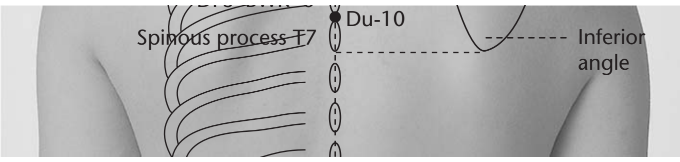
Du-10
Spinous process T7
Inferior angle

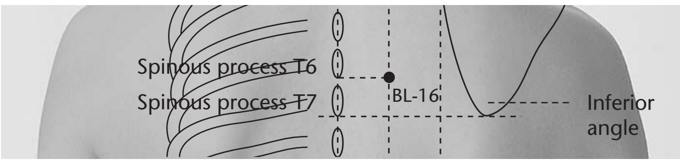
Spinous process T6
Spinous process T7
BL-16
Inferior angle

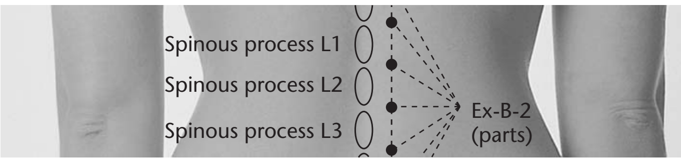
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Diaphragm's Gate GEGUAN BL-46

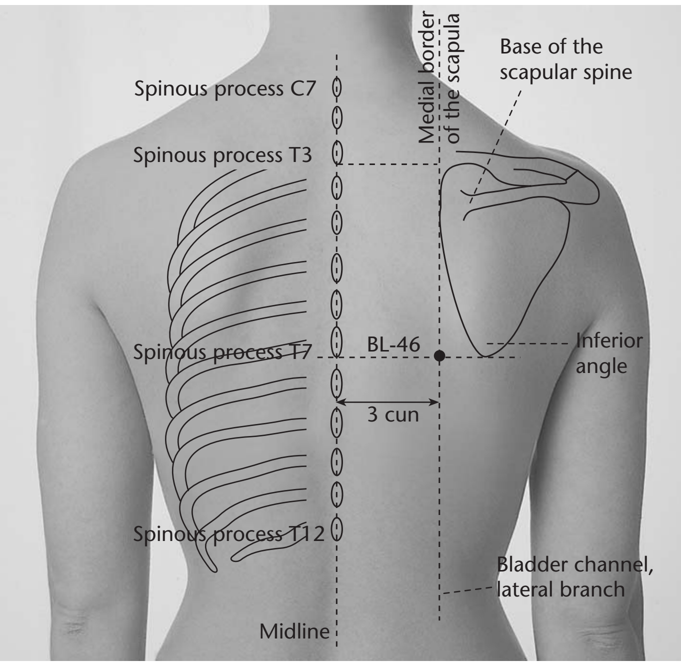
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
Spinous process T7
BL-46
Inferior angle
3 cun
Spinous process T12
Bladder channel, lateral branch
Midline

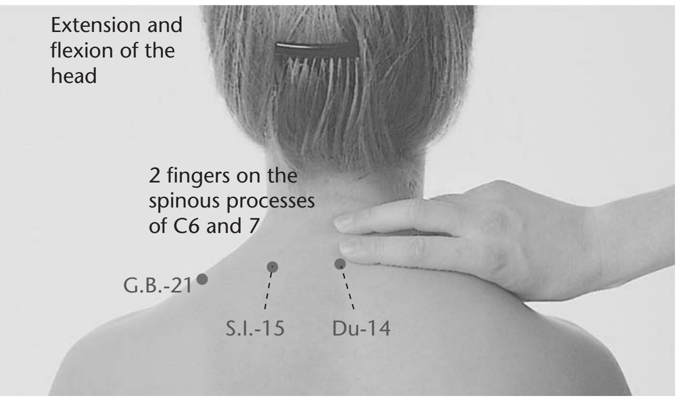
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

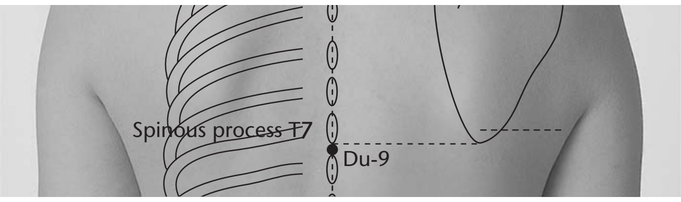
Spinous process T7
Du-9

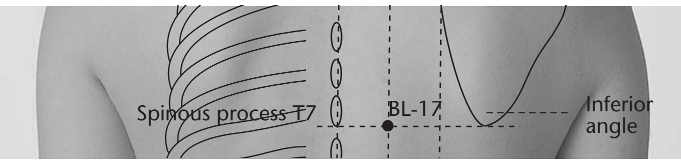
Spinous process T7
BL-17
Inferior angle

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 7th thoracic vertebra (T7).

How to find

With the patient standing and their shoulders in a relaxed position, the spinous process of T7 is generally on the level of the inferior angle of the scapula (➞ 3.4.2). Locate **BL-46** on the level of the spinous process of T7, 3 cun lateral to the midline. **Or:** With the patient seated, the spinous process of T3 is generally on the level of the medial end of the scapular spine, an easily palpable bony protrusion at the medial border of the scapula (➞ 3.4.2). From there, count down 4 spinous processes to the spinous process of T7. Locate **BL-46** on this level, 3 cun lateral to the midline. **Or:** Find the spinous process of C7 (➞ 3.4.1). From there, count down 7 spinous processes in an inferior direction to the spinous process of T7. Locate **BL-46** on this level, 3 cun lateral to the midline.

Located on the same level are ➞ **Du-9** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-17** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the diaphragm and descends counterflow Qi
- Harmonises the Middle Burner
- Opens the channel and alleviates pain

BL-47 *Gate of the Ethereal Soul* HUNMEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 9th thoracic vertebra (T9).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down 2 processes to the lower border of the spinous process of T9. **BL-47** is located 3 cun lateral to this point. **Or:** Locate the spinous process of C7 (→ 3.4.1). From there, count down 9 processes to the lower border of the spinous process of T9 and locate **BL-47** 3 cun lateral to it. Located on the same level are → **Du-8** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-18** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.5 cun. Caution: Pneumothorax.

Actions/Indications

- Spreads the Liver Qi
- Harmonises the Spleen and the Stomach
- Relaxes the tendons

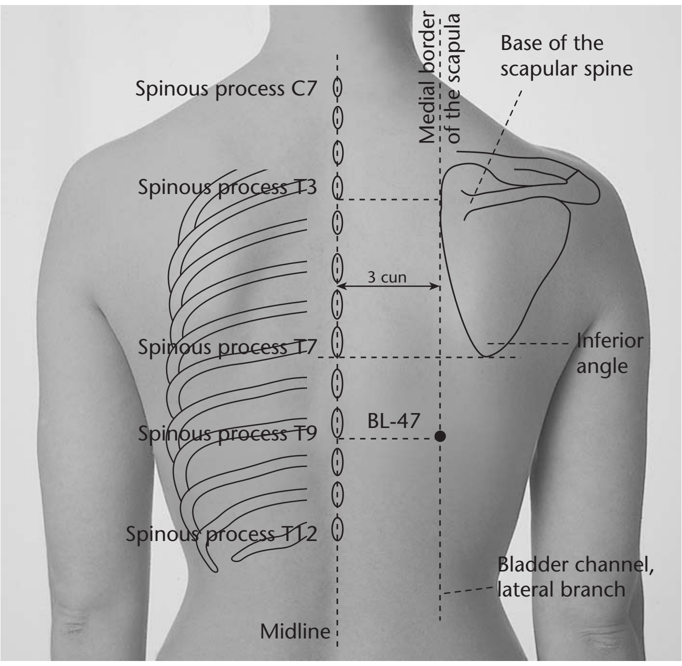
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
BL-47
Spinous process T9
Spinous process T12
Bladder channel, lateral branch
Midline

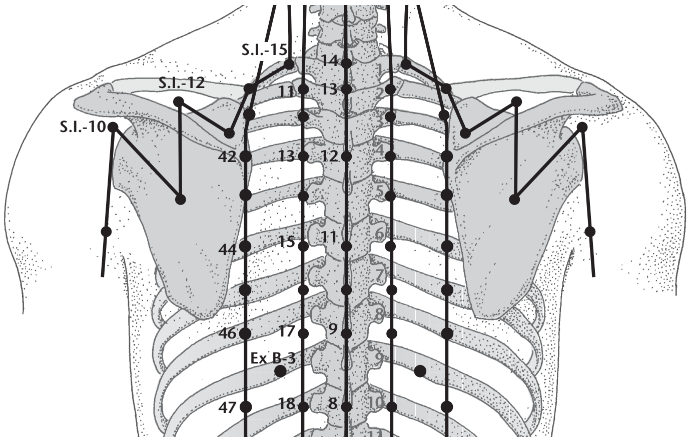
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex B-3
47
18
8

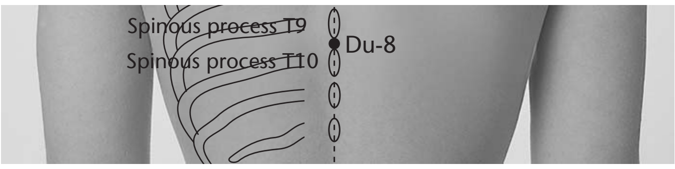
Spinous process T9
Spinous process T10
Du-8

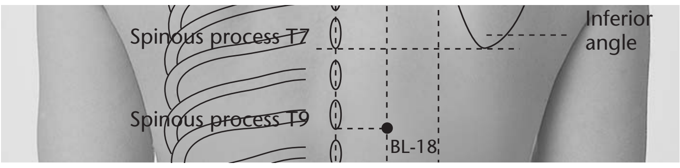
Spinous process T7
Inferior angle
Spinous process T9
BL-18

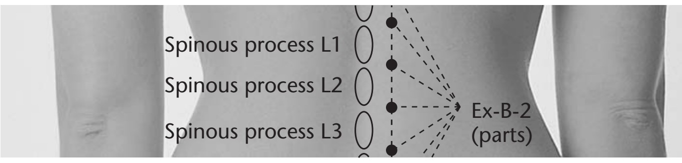
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

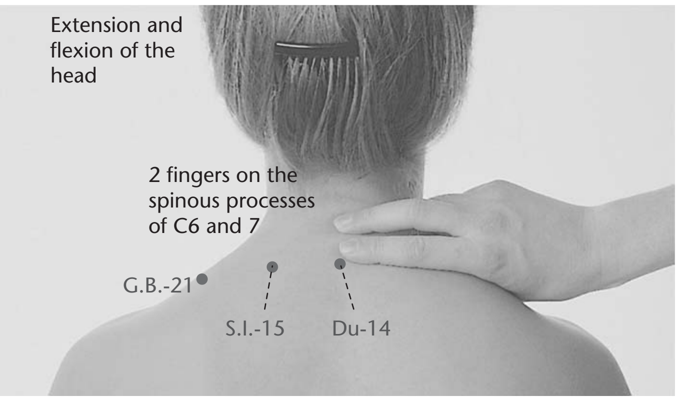
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Yang's Key Link YANGGANG BL-48

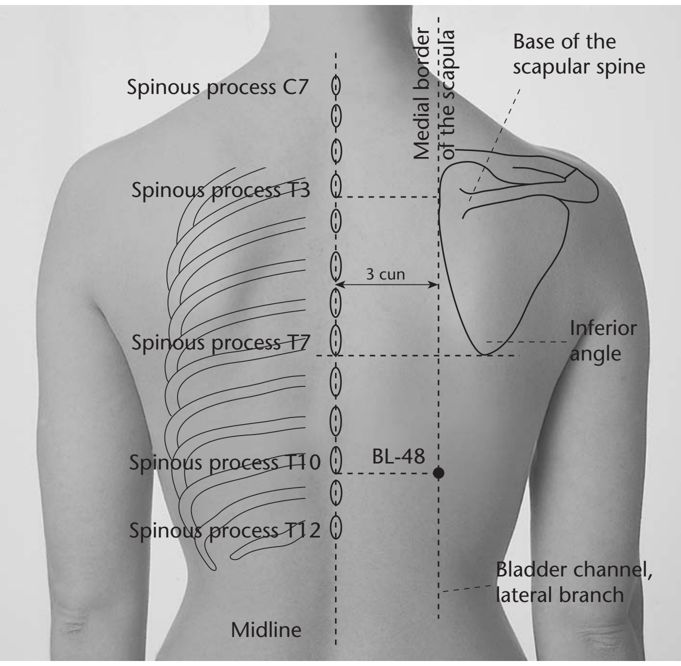
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-48
Spinous process T12
Bladder channel, lateral branch
Midline

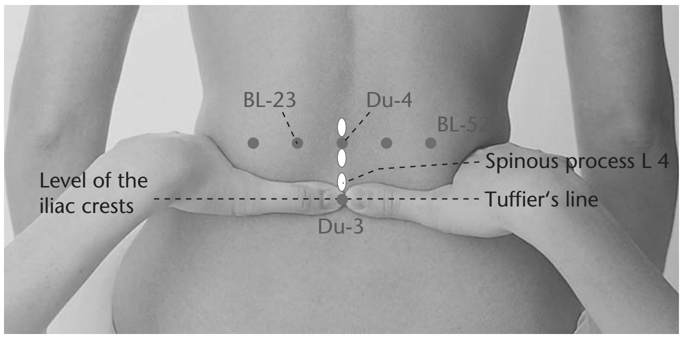
BL-23
Du-4
BL-52
Spinous process L 4
Level of the iliac crests
Tuffier's line
Du-3

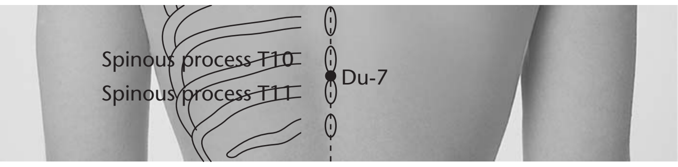
Spinous process T10
Spinous process T11
Du-7

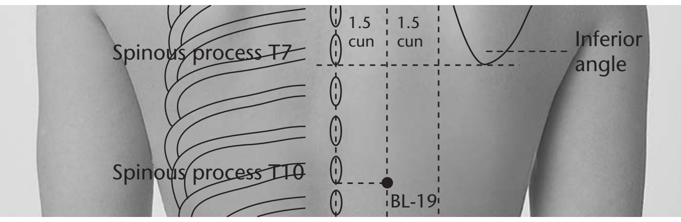
1.5 cun
1.5 cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 10th thoracic vertebra (T10).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down 3 processes to the lower border of the spinous process of T10. **BL-48** is located 3 cun lateral to this point. **Or:** Locate the spinous process of C7 (→ 3.4.1). From there, count down 10 processes to the lower border of the spinous process of T10 and locate **BL-48** 3 cun lateral to it.

Located on the same level are → **Du-7** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-19** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Regulates the Gall Bladder
- Clears Damp-Heat
- Harmonises the Middle Burner

BL-49 *Abode of Thought* YISHE

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 11th thoracic vertebra (T11).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down 4 spinous processes to the lower border of the spinous process of T11 and locate **BL-49** 3 cun lateral to this point. **Or:** From the articulation of the most inferior rib, palpate in a superior direction to T11. **Or:** Use the lumbar spine for reference (→ 3.4.3).

Located on the same level are → **Du-6** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-20** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Clears Damp-Heat
- Harmonises the Spleen and the Stomach

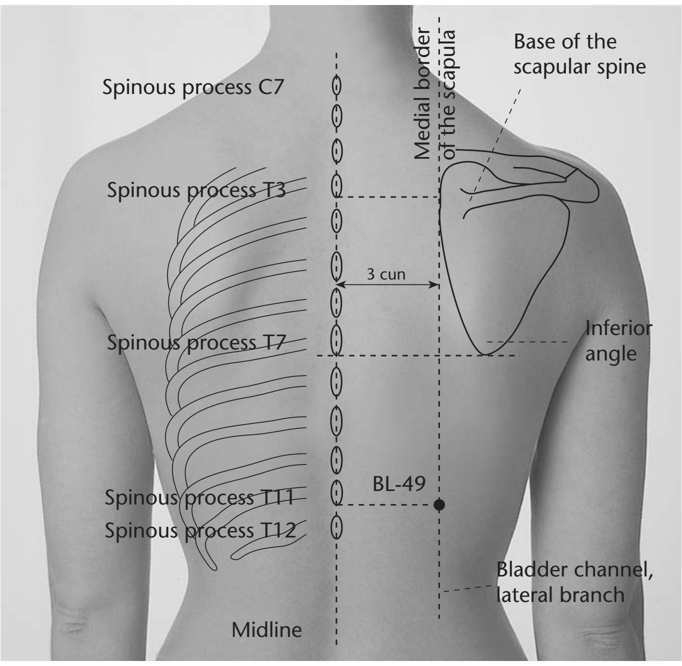
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
BL-49
Spinous process T11
Spinous process T12
Bladder channel, lateral branch
Midline

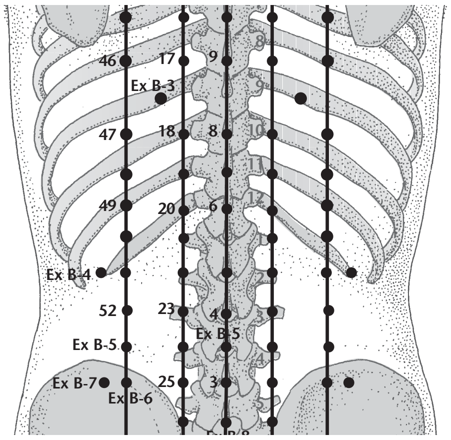
46
17
9
Ex B-3
47
18
8
10
11
49
20
6
12
Ex B-4
52
23
4
Ex B-5
Ex B-5
Ex B-7
25
3
Ex B-6

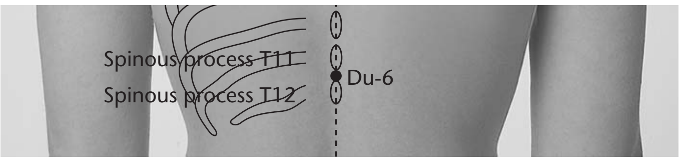
Spinous process T11
Spinous process T12
Du-6

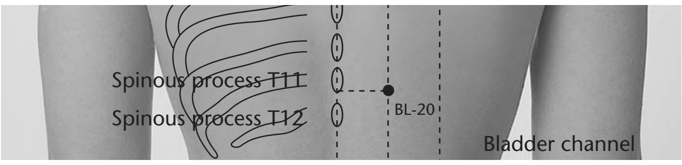
Spinous process T11
Spinous process T12
BL-20
Bladder channel

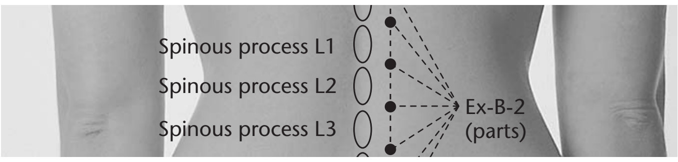
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ch04.7b-F10028.qxd 2/22/08 5:00 PM Page 297

Stomach Granary WEICANG BL-50

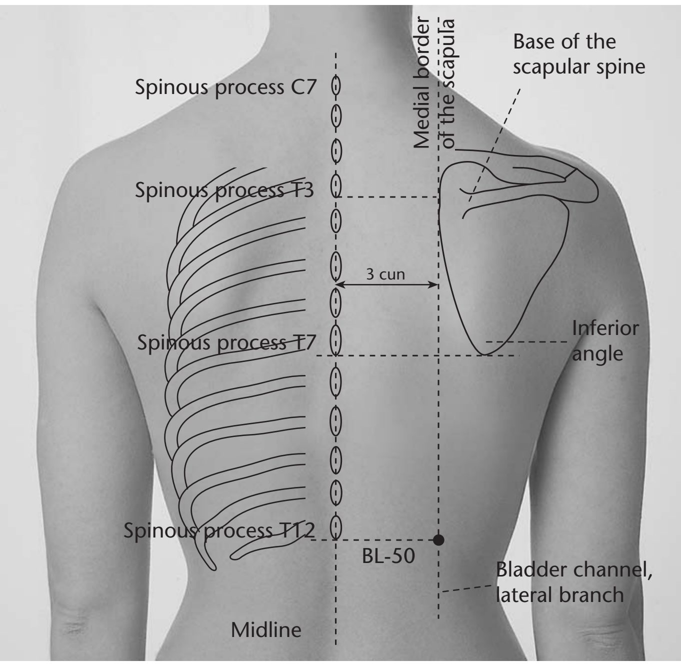
Spinous process C7
Medial border of the scapula
Base of the scapular spine
Spinous process T3
3 cun
Spinous process T7
Inferior angle
Spinous process T12
BL-50
Bladder channel, lateral branch
Midline

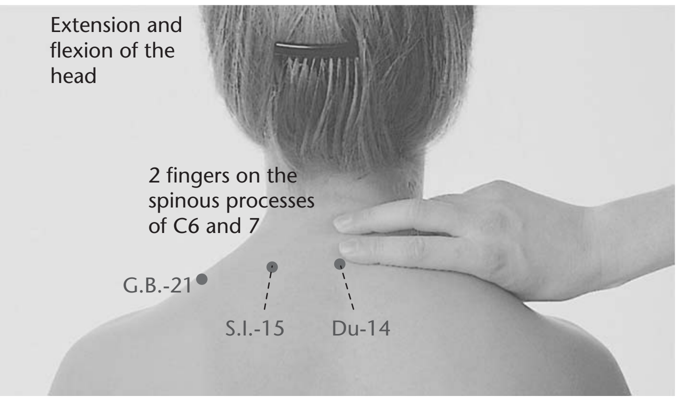
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

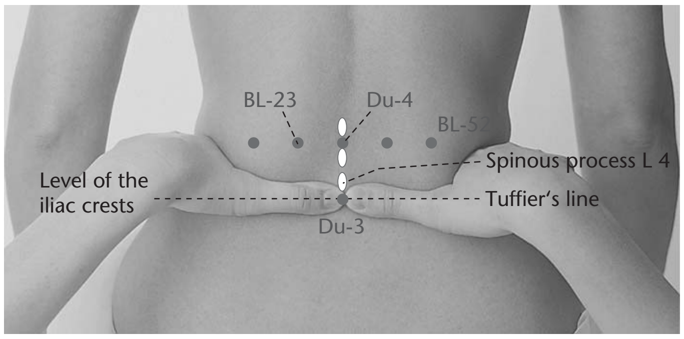
BL-23
Du-4
BL-52
Spinous process L 4
Level of the iliac crests
Tuffier's line
Du-3

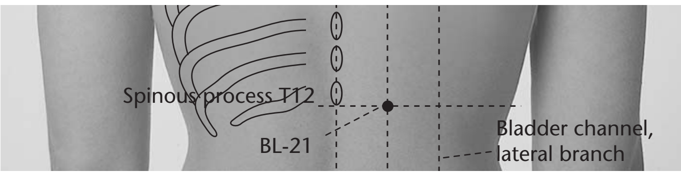
Spinous process T12
BL-21
Bladder channel, lateral branch

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 12th thoracic vertebra (T12).

How to find

In a sitting position with the arms hanging down, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down 5 spinous processes to the lower border of the spinous process of the T12 and locate **BL-50** 3 cun lateral to this point. **Or:** Locate the articulation of the most inferior rib with T12 and palpate 3 cun in a lateral direction for **BL-50. Or:** Orientation from the lumbar spine (→ 3.4.3). Located on the same level are a point of → **Ex-B-2** (0.5 cun lateral to the midline) and → **BL-21** (1.5 cun lateral to the midline).

Needling

Obliquely 0.3–0.8 cun. Caution: Pneumothorax, especially with asthenic patients or patients with emphysema.

Actions/Indications

- Harmonises the Middle Burner
- Opens the channel

BL-51 *Vitals Gate* HUANGMEN

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, the Tuffier's line: place both hands superior or lateral to the highest points of the iliac crest, joining the thumbs in the middle above the spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position → 3.4.). From there, count upward to the lower border of the spinous process of L1 and, on this level, locate **BL-51** 3 cun in a lateral direction. **Or:** With the patient seated and their arms hanging down in a relaxed way, the spinous process of T7 is generally on the level of the inferior angle of the scapula (→ 3.4.2). From there, count down six spinous processes and locate **BL-51** 3 cun lateral to the spinous process of L1.

Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2** (0.5 cun lateral to the midline), → **BL-22** (1.5 cun lateral to the midline) and → **Ex-B-4** (→ 3.5 cun lateral to the midline).

Needling

Vertically or obliquely 0.5–1 cun. Caution: Kidneys.

Actions/Indications

- Moves the Qi and resolves stagnation
- Benefits the breasts

Special features

Distal point for disorders of the breasts

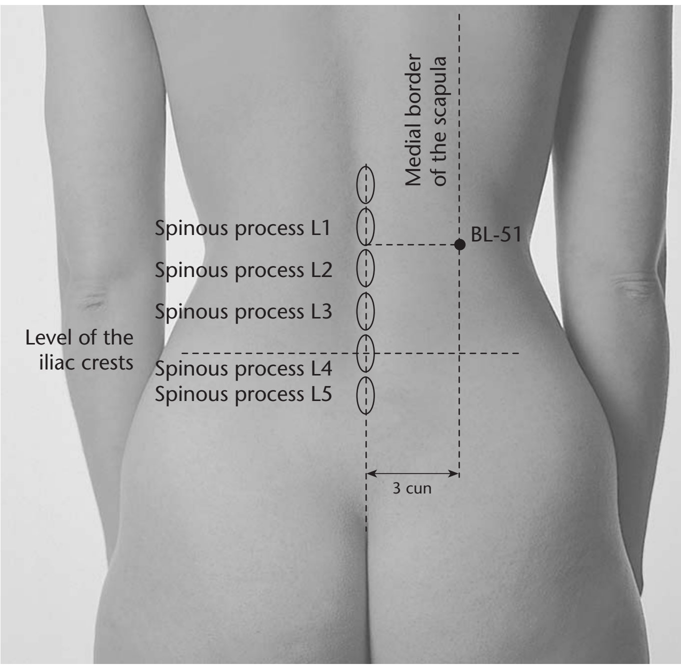
Medial border of the scapula
BL-51
Spinous process L1
Spinous process L2
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
3 cun

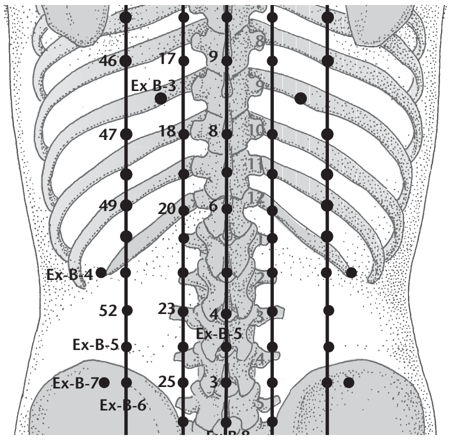
46
17
9
Ex B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6

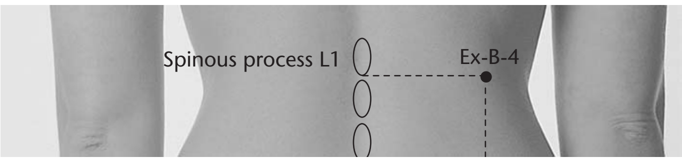
Spinous process L1
Ex-B-4

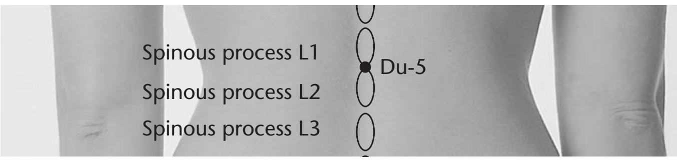
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

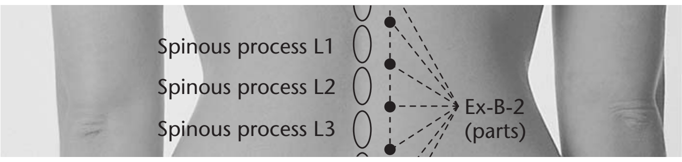
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

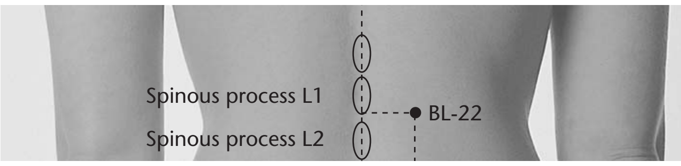
Spinous process L1
BL-22
Spinous process L2

Residence of the Will ZHISHI BL-52

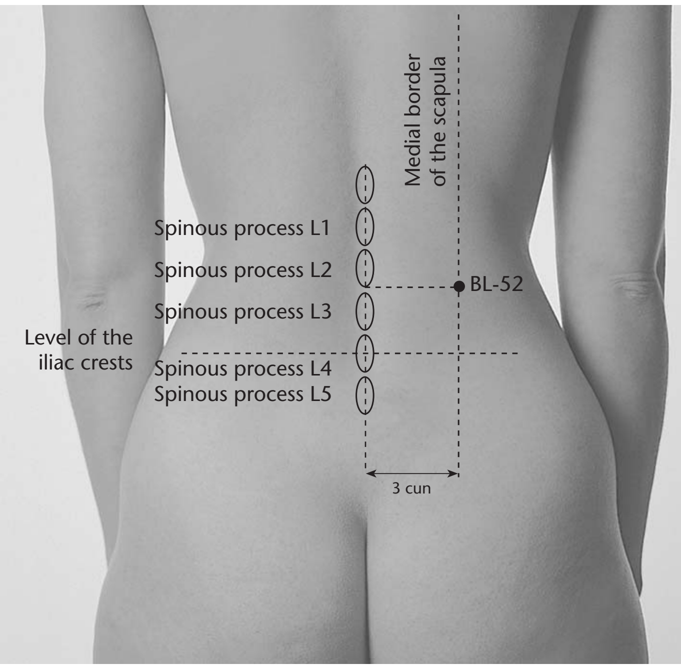
Medial border of the scapula
Spinous process L1
Spinous process L2
BL-52
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
3 cun

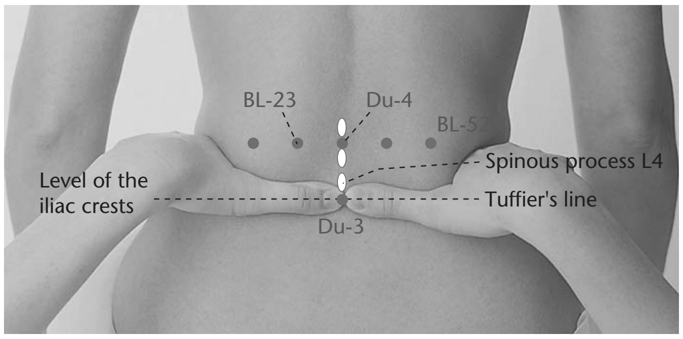
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

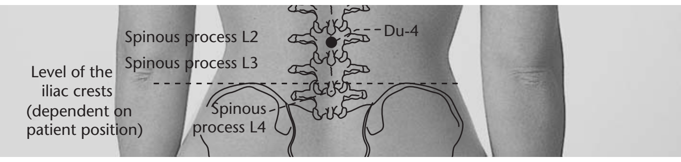
Spinous process L2
Du-4
Spinous process L3
Level of the iliac crests (dependent on patient position)
Spinous process L4

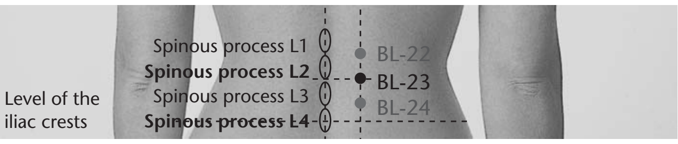
Spinous process L1
BL-22
Spinous process L2
BL-23
Spinous process L3
BL-24
Level of the iliac crests
Spinous process L4

Location

3 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (➞ 3.4.3), first, locate the Tuffier's line: place both hands superior or lateral to the highest points of the iliac crest, joining the thumbs in the middle above the spine. In most cases, this line intersects with the spinous process of L4 (note: this varies depending on the patient's position ➞ 3.4.). From there, count upward to the lower border of the spinous process of L2 and, on this level, locate **BL-52** 3 cun in a lateral direction.

Located on the same level are ➞ **Du-4** (on the midline), a point of ➞ **Ex-B-2** (0.5 cun lateral to the midline) and ➞ **BL-23** (1.5 cun lateral to the midline).

Needling

Vertically or obliquely 0.5–1 cun. Caution: Kidneys.

Actions/Indications

- Tonifies the Kidneys and Essence (*jing*), regulates urination
- Benefits the lumbar region

BL-53 *Bladder Vitals* BAOHUANG

Location

3 cun lateral to the posterior midline, on the level of the 2nd sacral foramen.

How to find

For details on orientation in the lumbar and sacral region, → 3.4.3, 3.4.4. First, locate L5. The next palpable bony structure inferior to L5 is the highest process of the medial sacral crest. Palpate inferiorly for the 2nd process and locate the 2nd sacral foramen slightly less than 1 cun lateral to the process. Next, locate **BL-53** 3 cun lateral to the midline on the level of the 2nd foramen. **Or:** Use the sacral foramina as a reference point (→ 3.4.4). **BL-53** is located on the level of the 2nd foramen, 3 cun lateral to the midline.

Located on the same level are → **BL-32** (over the 2nd sacral foramen) and → **BL-28** (1.5 cun lateral to the midline).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Benefits the lumbar region, opens the channel and alleviates pain
- Regulates the Lower Burner

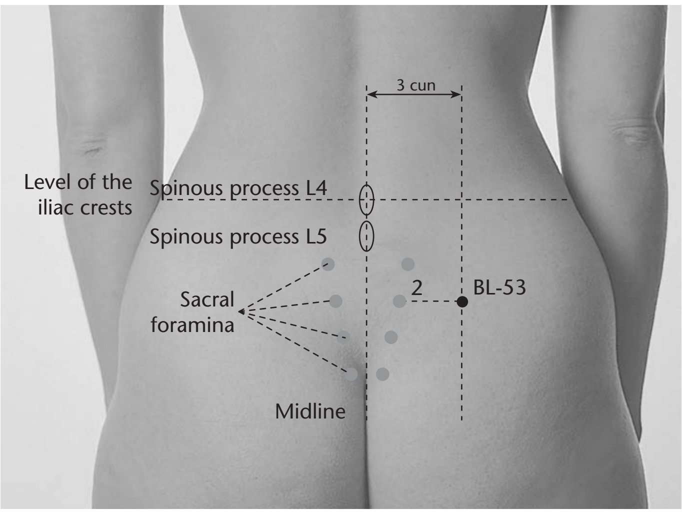
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
2
BL-53
Midline

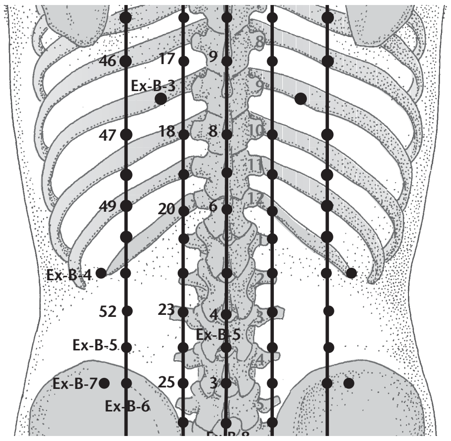
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6

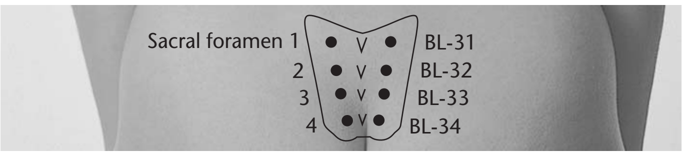
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

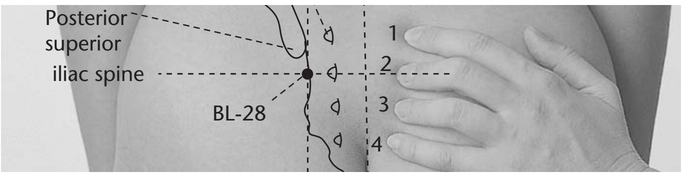
Posterior superior iliac spine
BL-28
1
2
3
4

Order's Limit ZHIBIAN BL-54

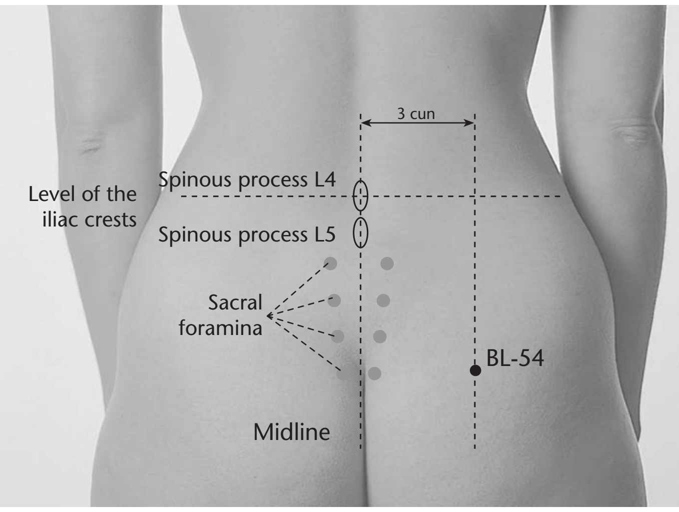
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
BL-54
Midline

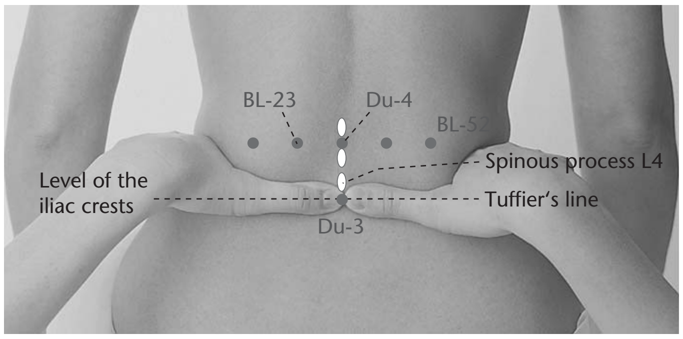
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

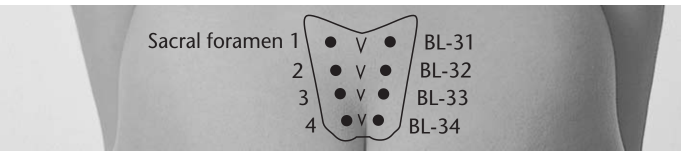
Sacral foramen 1
2
3
4
BL-31
BL-32
BL-33
BL-34

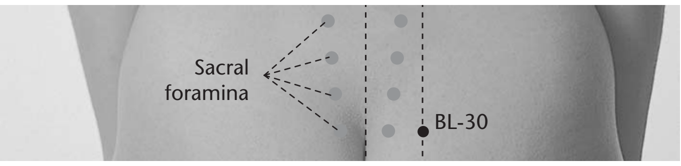
Sacral foramina
BL-30

Location

3 cun lateral to the posterior midline, on the level of the 4th sacral foramen, approximately at the centre of the buttock.

How to find

For details on orientation in the lumbar and sacral region, see → 3.4.3, 3.4.4. First, locate L5. Inferior to it, the next palpable bony structure is the highest point on the medial sacral crest (= sacral spinous processes). Palpate downward for 3 more bony elevations and locate the depression of the 4th sacral foramen less than 1 cun lateral to it. This is on the level of **BL-54**, located 3 cun lateral to the midline. For further reference: The 4th sacral foramen is mostly level with the superior end of the rima ani. Located on the same level are → **BL-34** (over the 4th sacral foramen) and → **BL-30** (1.5 cun lateral to the midline).

Needling

Vertically 1.5–2 cun or 2–3 cun in the direction of the anus or the genital region, depending on the indication.

Actions/Indications

- Opens the channel and *luo* vessels, eliminates stagnation
- Regulates urination
- Eliminates haemorrhoids

Special features

Important local point

BL-55 *Confluence of Yang* HEYANG

Location

2 cun inferior to the midpoint of the popliteal crease, in a depression between the two bellies of the gastrocnemius muscle.

How to find

Locate the popliteal crease by flexing the knee. In slim patients, palpate for the joint space. From the centre of the popliteal crease, palpate 2 cun in an inferior direction and locate **BL-55** in a depression between the two bellies of the gastrocnemius muscle.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel, alleviates pain (also in the urogenital region)
- Stops uterine bleeding

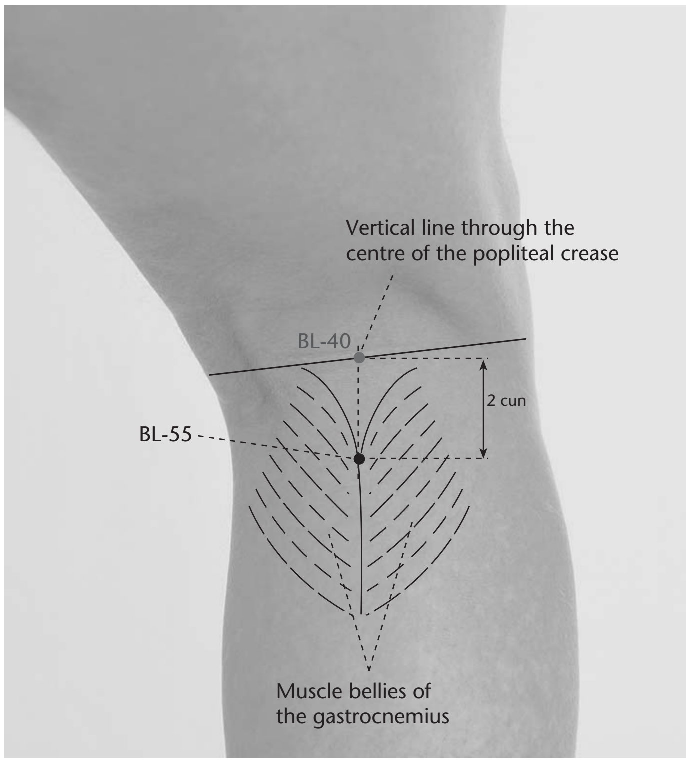
Vertical line through the centre of the popliteal crease
BL-40
2 cun
BL-55
Muscle bellies of the gastrocnemius

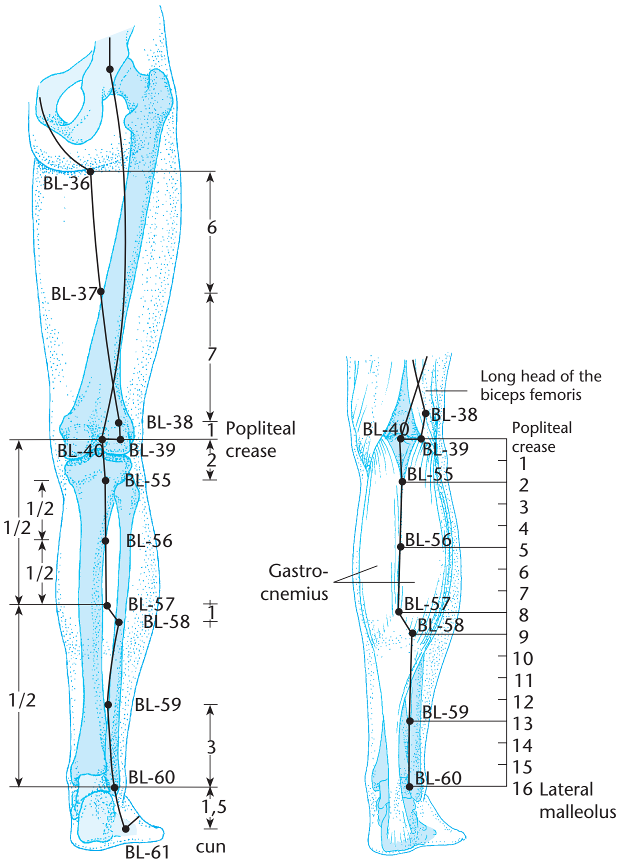
BL-36
6
BL-37
7
BL-38
1
Popliteal crease
BL-40
BL-39
2
BL-55
1/2
1/2
BL-56
1/2
BL-57
BL-58
1
1/2
BL-59
3
BL-60
1,5
BL-61
cun
Long head of the biceps femoris
BL-38
BL-40
BL-39
Popliteal crease
1
BL-55
2
3
4
BL-56
5
6
Gastro-cnemius
7
BL-57
8
BL-58
9
10
11
12
BL-59
13
14
15
BL-60
16 Lateral malleolus

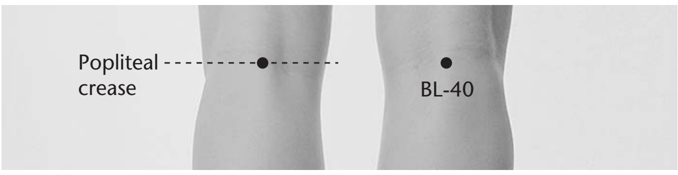
Popliteal crease
BL-40

Support the Sinews CHENGJIN **BL-56**

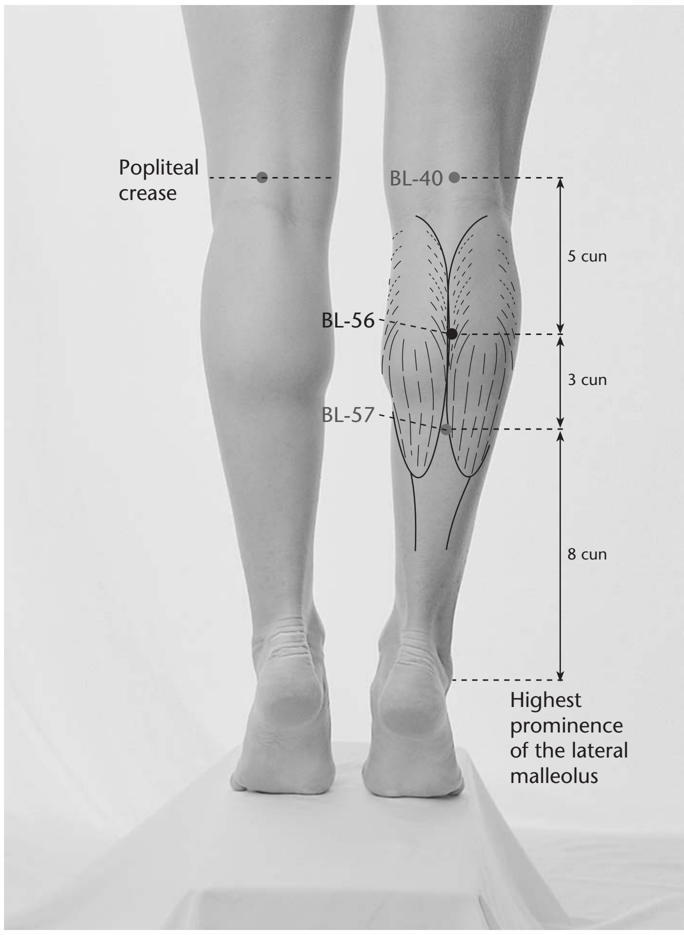
Popliteal crease
BL-40
5 cun
BL-56
3 cun
BL-57
8 cun
Highest prominence of the lateral malleolus

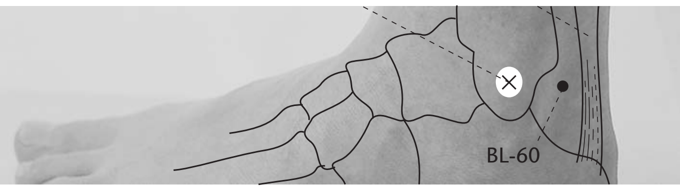

Location

5 cun inferior to the midpoint of the popliteal crease, between the two bellies of the gastrocnemius muscle.

How to find

Locate the popliteal crease by flexing the knee. In slim patients, palpate for the joint space. **BL-56** is located 5 cun distal to the popliteal crease, in a depression between the two bellies of the gastrocnemius muscle.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Relaxes the muscles and tendons
- Opens the channel
- Eliminates haemorrhoids

BL-57 *Supporting Mountain* CHENGSHAN

Location

In the middle of the calf, between the two heads of the gastrocnemius muscle, on a line connecting → **BL-40** and → **BL-60**, approximately 8 cun distal to → **BL-40**.

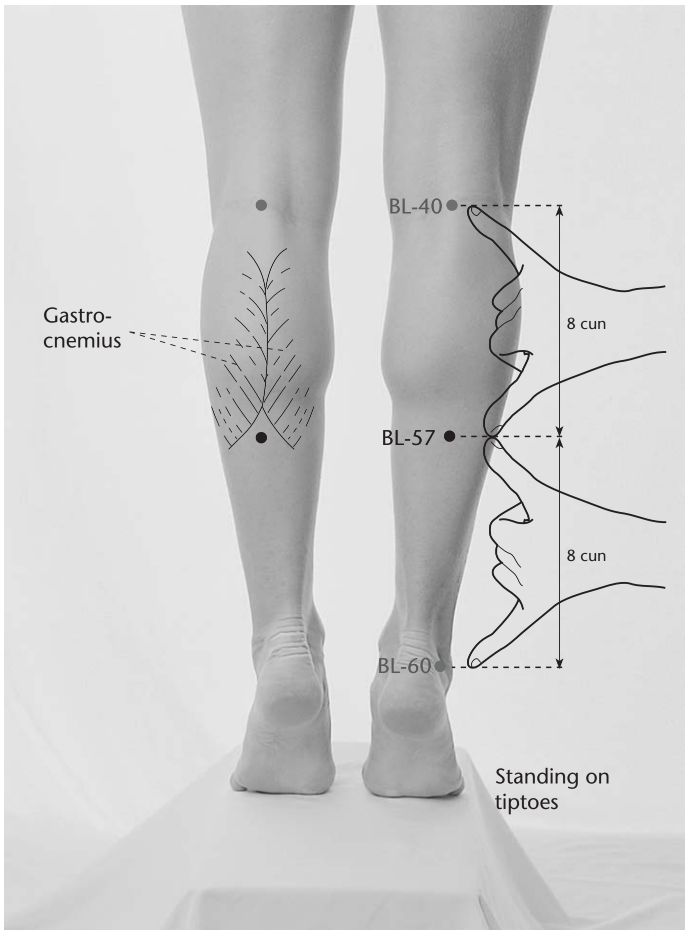
BL-40
Gastro-
cnemius
8 cun
BL-57
8 cun
BL-60
Standing on
tiptoes

How to find

Slide along the dorsal aspect of the Achilles tendon in a superior direction until you can feel the depression between the two bellies of the gastrocnemius muscle. This will become more defined when the patient's muscles are flexed. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on → **BL-40** (centre of the popliteal crease) and → **BL-60** (depression between the Achilles tendon and the highest prominence of the lateral malleolus). The midpoint of this distance, where the thumbs meet, marks the location of **BL-57**.

Located on the same level on the anterolateral aspect of the lower leg are → **ST-38** (1 finger-width lateral to the tibia) and → **ST-40** (2 fingerbreadths lateral to the tibia).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Relaxes the muscles and tendons, opens the channel
- Eliminates haemorrhoids (pathway of the Bladder divergent channel)

Special features

Ma Dan Yang Heavenly Star point. Important local point with distal effect on the lumbar and anal region.

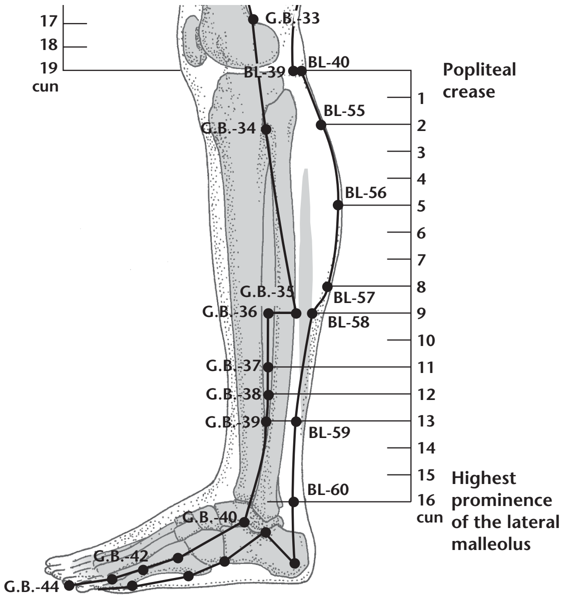
17
18
19
cun
G.B.-33
BL-40
BL-39
Popliteal
crease
1
BL-55
G.B.-34
2
3
4
BL-56
5
6
7
8
G.B.-35
BL-57
G.B.-36
9
BL-58
10
G.B.-37
11
G.B.-38
12
G.B.-39
13
BL-59
14
15
Highest
prominence
of the lateral
malleolus
BL-60
16
cun
G.B.-40
G.B.-42
G.B.-44

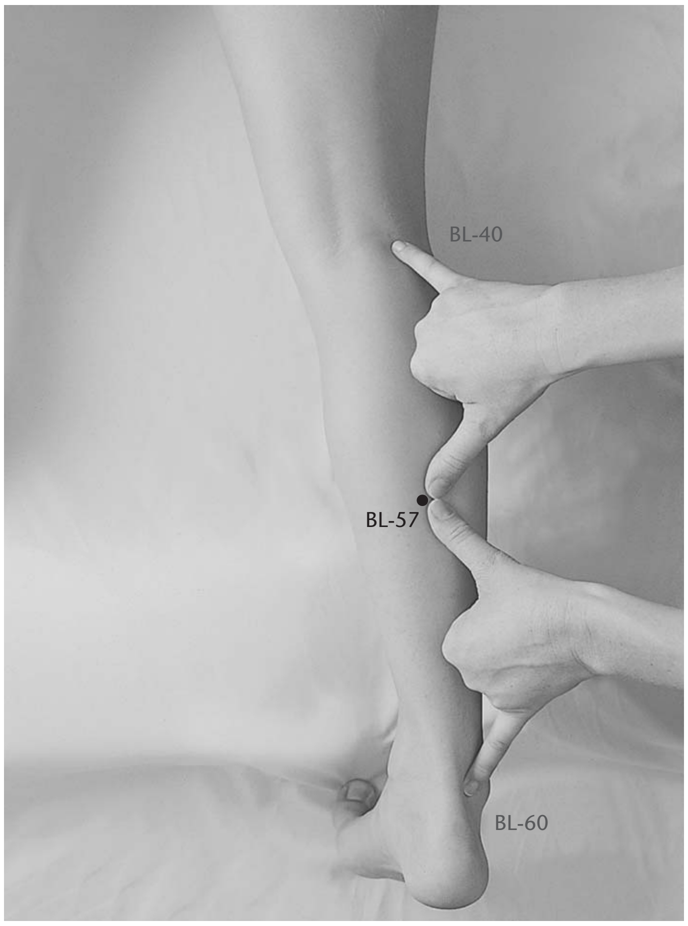
BL-40
BL-57
BL-60

Soaring Upward FEIYANG BL-58

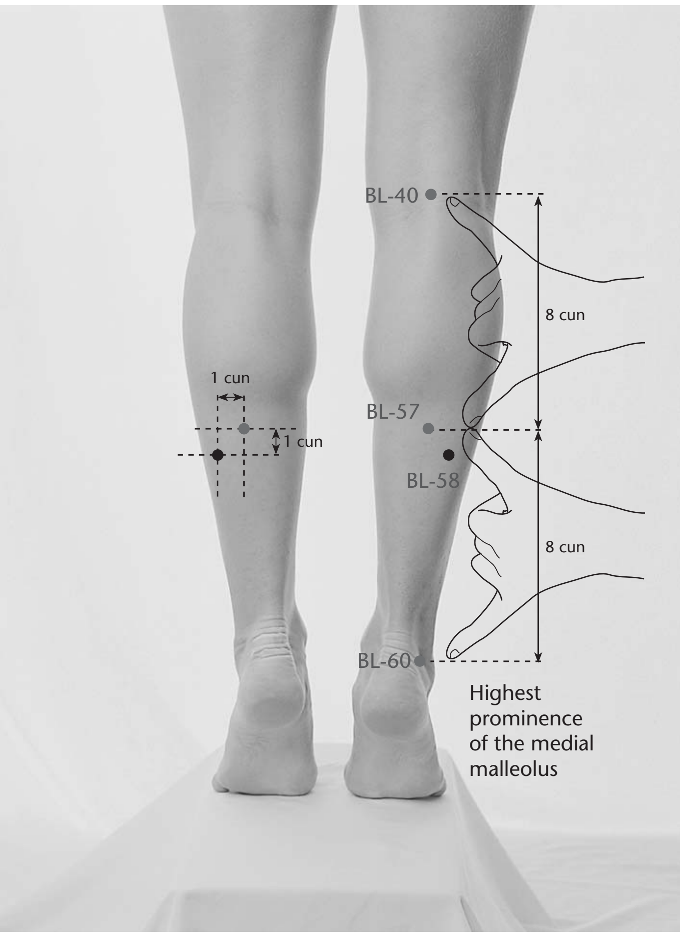
BL-40
8 cun
1 cun
1 cun
BL-57
BL-58
8 cun
BL-60
Highest prominence of the medial malleolus

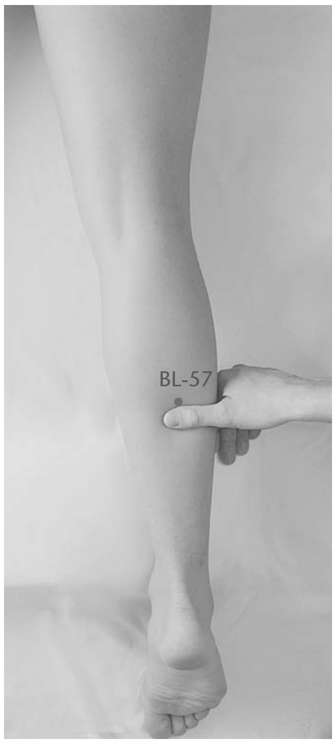

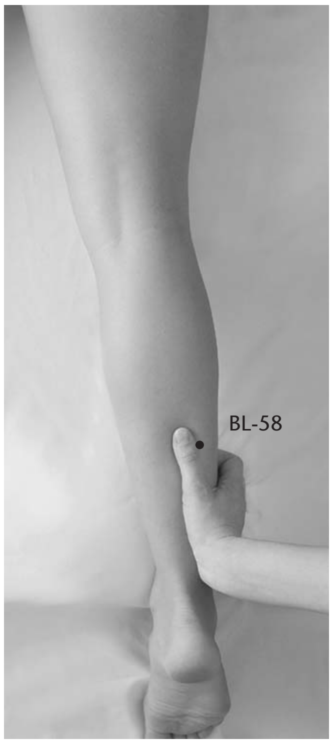

Location

1 cun distal and 1 cun lateral to → **BL-57** or 7 cun proximal to → **BL-60**, at the posterior border of the fibula and at the lower border of the gastrocnemius muscle.

How to find

First, locate → **BL-57** (in the centre of the calf, between the two bellies of the gastrocnemius muscle). From → **BL-57**, palpate 1 cun in a distal direction and 1 cun in a lateral direction. There, locate **BL-58** on the lower border of the gastrocnemius muscle. Located on the same level (7 cun superior to the highest prominence of the lateral malleolus) are → **G.B.-35** (at the posterior border of the fibula), → **G.B.-36** (at the anterior border of the fibula) and → **ST-39** (1 cun distal to the midpoint of the line connecting → **ST-35** and → **ST-41** and 1 fingerbreadth lateral to the anterior crest of the tibia).

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels, alleviates pain
- Expels pathogenic factors from the *taiyang* channels (S.I., BL)
- Harmonises above and below
- Treats haemorrhoids

Special features

Luo-connecting point

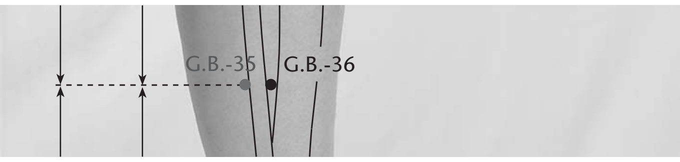

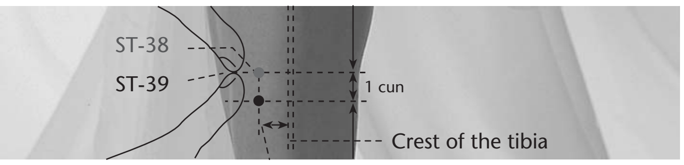
ST-38
ST-39
1 cun
Crest of the tibia

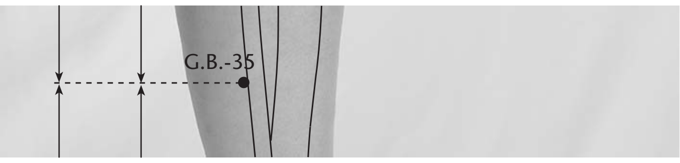

Ch04.7b-F10028.qxd 2/22/08 5:01 PM Page 306

BL-59 *Instep Yang* FUYANG

Location

On the lateral aspect of the lower leg, 3 cun superior to → **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon).

How to find

First, locate → **BL-60** in the depression between the highest prominence of the lateral malleolus and the Achilles tendon. With oedema of the lower leg, this point is not visible but palpable. From → **BL-60**, measure 3 cun (1 handbreadth) in a superior direction. There, locate **BL-59** in a depression between the Achilles tendon and the tendons of the peroneous longus and brevis muscles.

Located on the same level is → **G.B.-39**, 3 cun directly above the prominence of the lateral malleolus.

Needling

Vertically or obliquely 1–1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels
- Benefits the lower back
- Benefits the head, activates the *yang qiao mai*

Special features

Xi-cleft point of the *yang qiao mai*, meeting point with the *yang qiao mai* according to some authors.

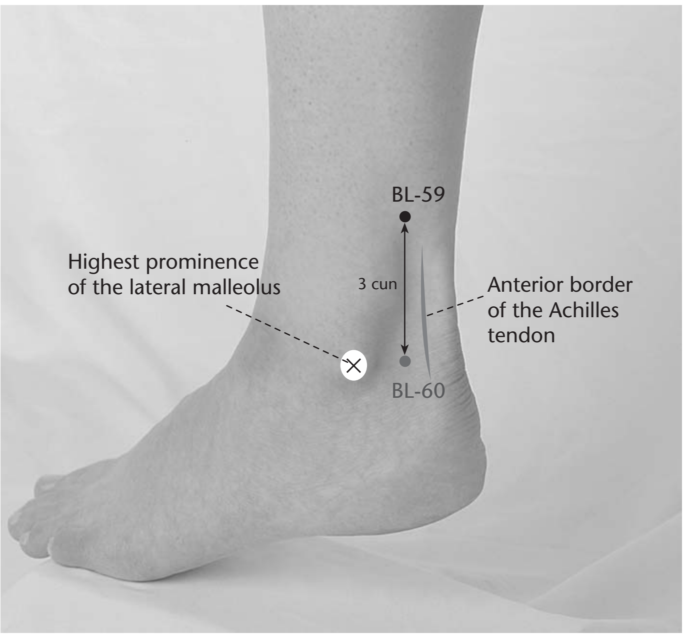
BL-59
Highest prominence of the lateral malleolus
3 cun
Anterior border of the Achilles tendon
BL-60

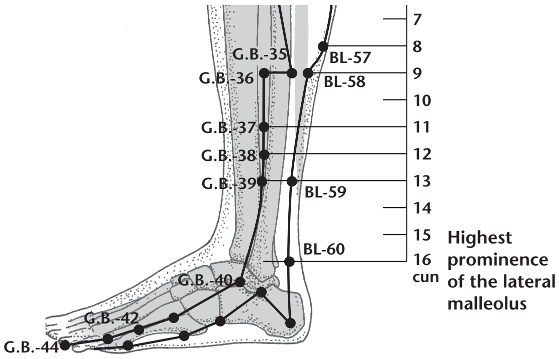
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

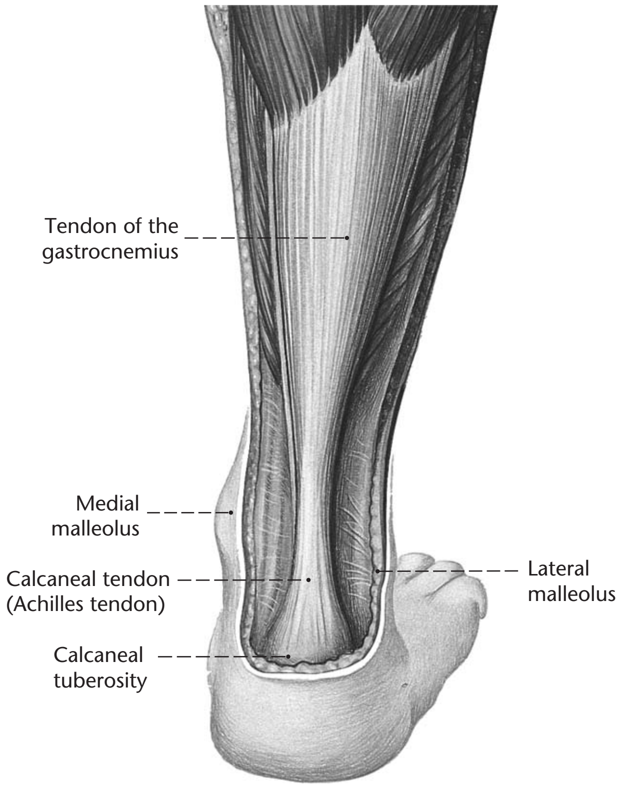
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Lateral malleolus
Calcaneal tuberosity

Kunlun Mountains KUNLUN **BL-60**

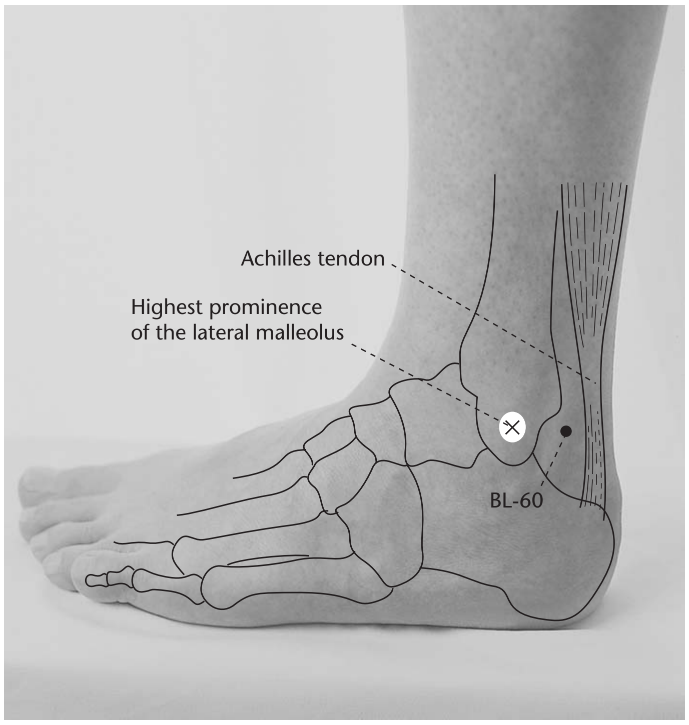
Achilles tendon
Highest prominence of the lateral malleolus
BL-60

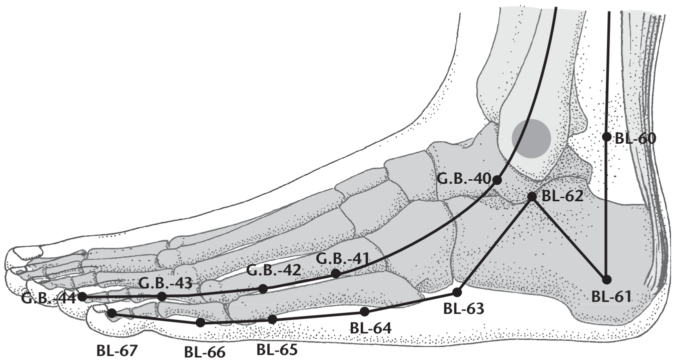
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

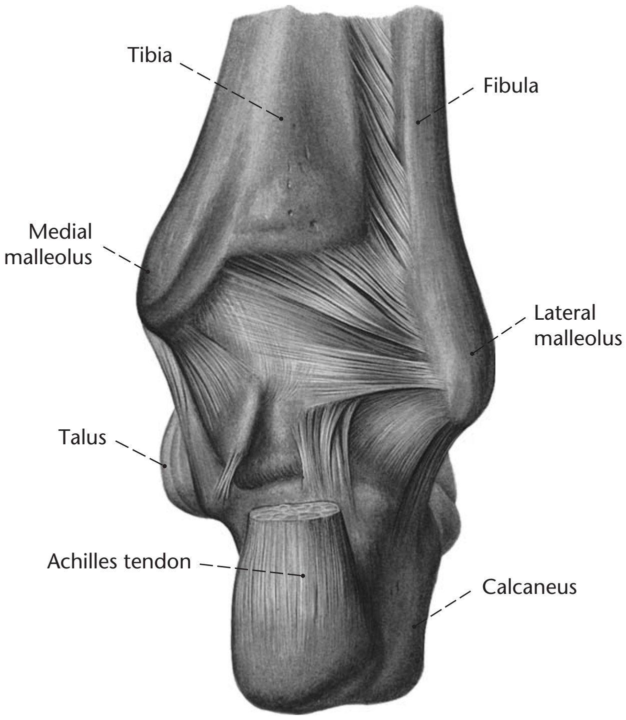
Tibia
Fibula
Medial malleolus
Lateral malleolus
Talus
Achilles tendon
Calcaneus

Location

In the depression on the line connecting the Achilles tendon and the highest prominence of the lateral malleolus.

How to find

Locate the highest prominence of the lateral malleolus (→ 3.6.2). From there, palpate horizontally towards the Achilles tendon and locate **BL-60** in a depression anterior to the tendon.

Needling

Vertically 0.5–1 cun. Caution: Reducing needle techniques contraindicated during pregnancy. Exception: A supportive point during labour. If the patient has cold feet, the therapeutic effect can be increased by using a heatlamp or a hot cherry-pit bag under the soles of the feet.

Actions/Indications

- Clears Heat, subdues Yang and Wind and eliminates excess, especially in the head
- Opens the channel, alleviates pain, relaxes the tendons, strengthens the lumbar region
- Promotes labour

Special features

Jing-river point, Fire point, Ma Dan Yang Heavenly Star point. Important distal point for the cervical, thoracic and lumbar spine, especially for chronic cases.

BL-61 *Servant's Respect* PUCAN

Location

On the lateral aspect of the heel, 1.5 cun inferior to → **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon), in a depression on the calcaneus.

How to find

First, locate → **BL-60** in the depression between the highest prominence of the lateral malleolus and the Achilles tendon. With oedema of the lower leg, the depression is not visible but palpable. From → **BL-60**, measure 1.5 cun in a distal direction. There, locate **BL-61** in a depression on the calcaneus, approximately on the midpoint of a line connecting the sole of the foot and → **BL-60**.

Needling

Vertically or obliquely 0.3–0.5 cun

Actions/Indications

- Relaxes the tendons
- Opens the channel, alleviates pain

Special features

Meeting point with the *yang qiao mai*

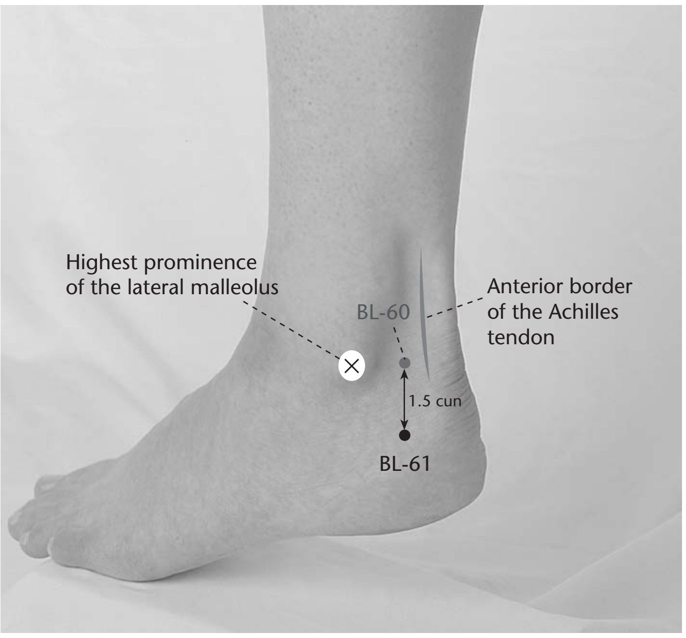
Highest prominence of the lateral malleolus
BL-60
Anterior border of the Achilles tendon
1.5 cun
BL-61

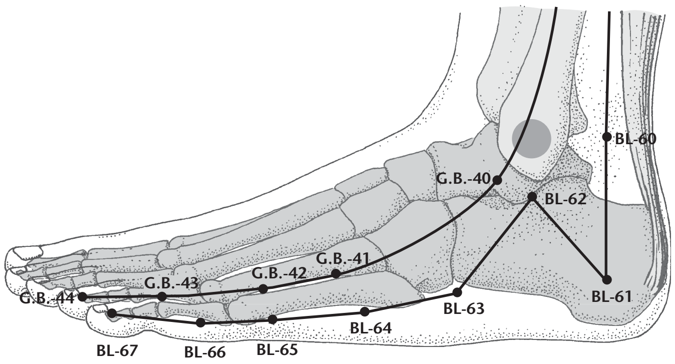
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

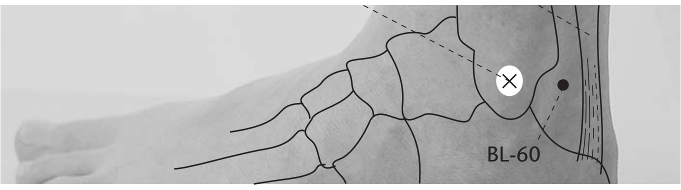

Extending Vessel SHENMAI BL-62

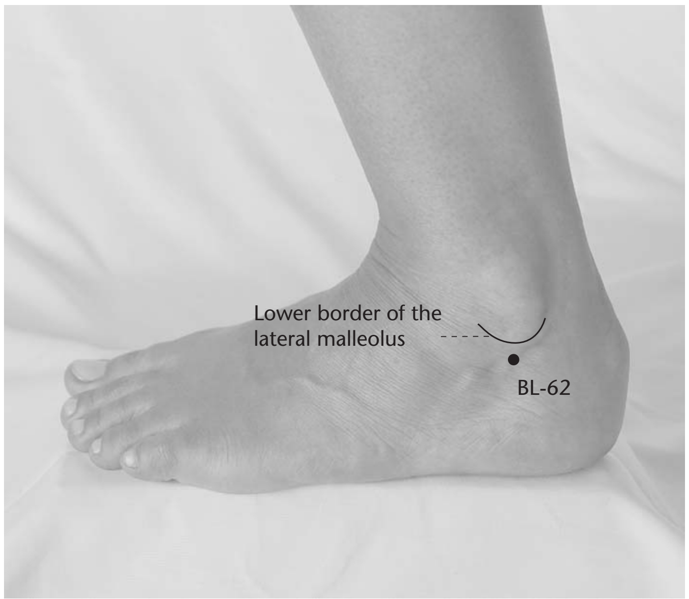
Lower border of the lateral malleolus
BL-62

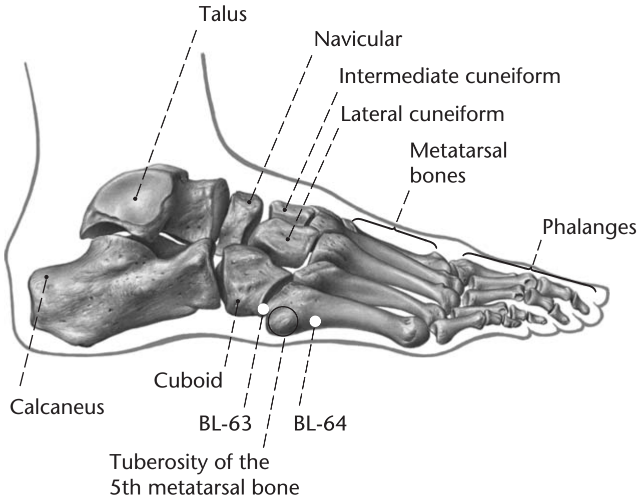
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Cuboid
Calcaneus
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Location

In a depression directly inferior to the highest prominence of the lateral malleolus, above the joint space between the talus and calcaneus.

How to find

First, locate the highest prominence of the lateral malleolus (→ 3.6.2). **BL-62** is located directly below it, in a depression inferior to the lower border of the malleolus and superior to the tendons of the peroneus muscle, in the joint space between the talus and calcaneus.

Needling

Vertically or obliquely 0.3–0.5 cun. The needle may reach the calcaneofibular ligaments superior to the tendons of the peroneus longus and brevis muscles, possibly also the joint space.

Actions/Indications

- Subdues (internal) Wind, clears Heat from the head, calms the *shen*, benefits the head and eyes
- Expels external Wind
- Opens and regulates the *yang qiao mai*
- Opens the channel, alleviates pain
- Moves the Qi in the Bladder sinew channel

Special features

Opening (master) point of the *yang qiao mai*, Sun Si Miao Ghost point.

BL-63 Golden Gate JINMEN

Location

On the lateral aspect of the foot, proximal to the tuberosity of the 5th metatarsal bone, in a depression anterior and inferior to → **BL-62** between the calcaneus and the cuboid bone. Note: Some authors locate **BL-63** between the cuboid bone and the tuberosity of the 5th metatarsal bone (see figure on **p. 311**); the more tender point should be selected.

How to find

When palpating along the lateral aspect of the foot on the level of the border of the 'red and white' skin, a bony structure can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal point (→ 3.6.2). Slightly proximal to it (towards the heel), you can palpate a depression between the calcaneus and the cuboid bone. There, locate **BL-63**.

→ **BL-64** is located distal (towards the toes) to the tuberosity of the 5th metatarsal bone.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Opens the channel, alleviates pain and eliminates stagnation (*xi*-cleft point)
- Subdues internal Wind and calms the *shen*

Special features

Xi-cleft point, meeting point with the *yang wei mai*.

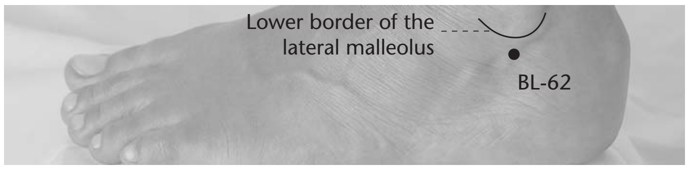
Lower border of the lateral malleolus
BL-62

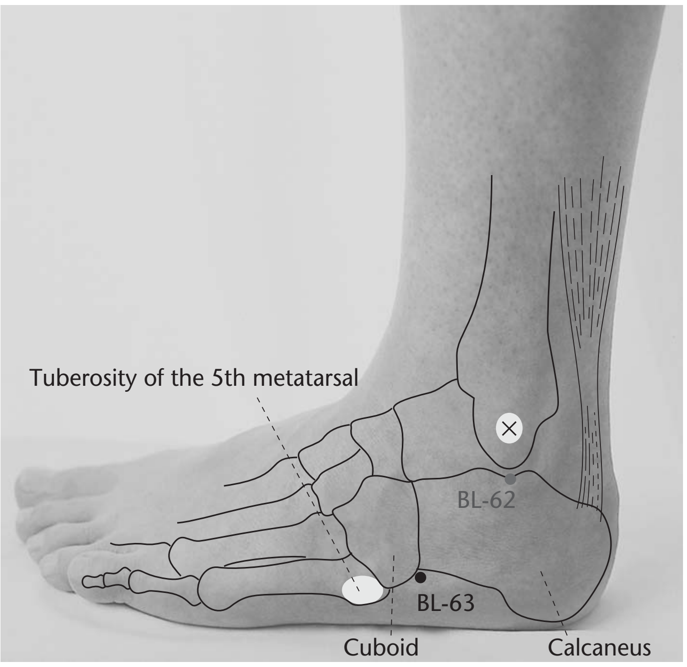
Tuberosity of the 5th metatarsal
BL-62
BL-63
Cuboid
Calcaneus

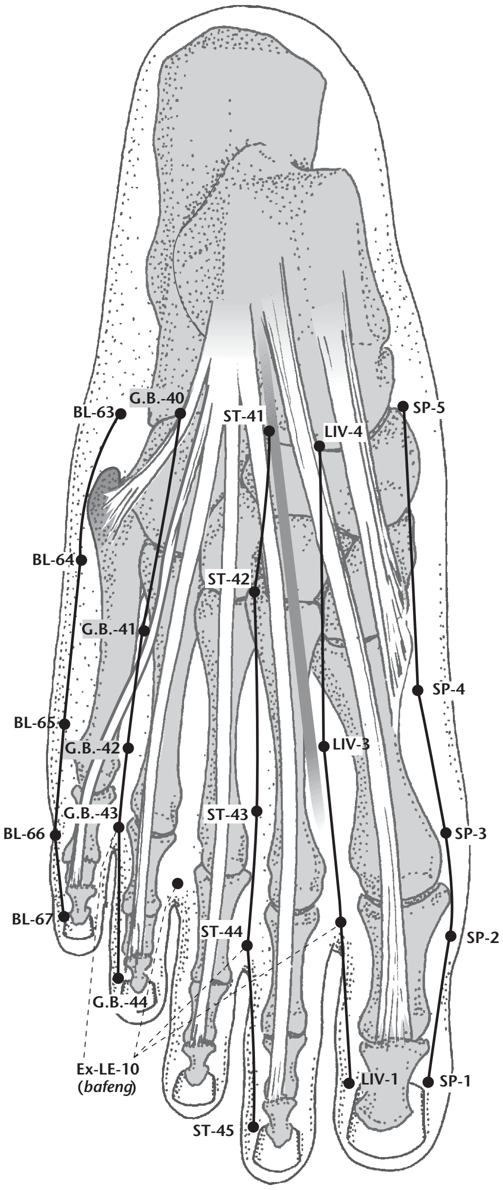
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

Capital Bone JINGGU BL-64

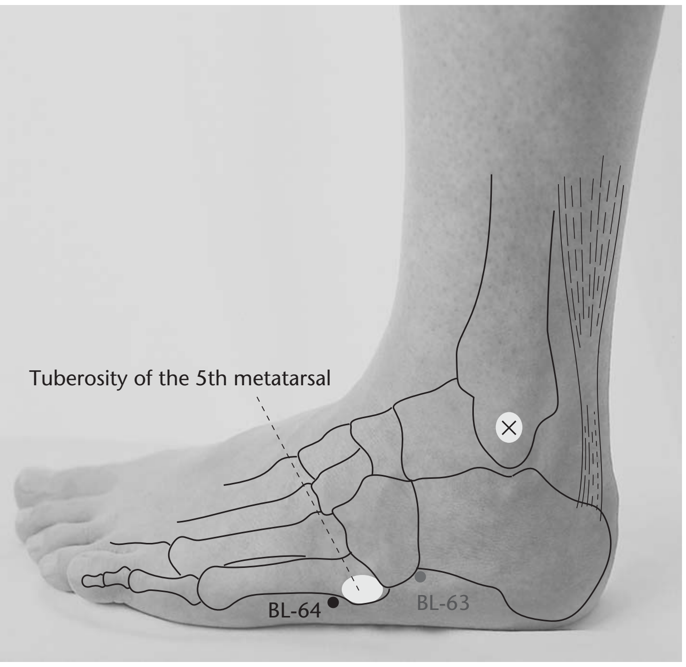
Tuberosity of the 5th metatarsal
BL-64
BL-63

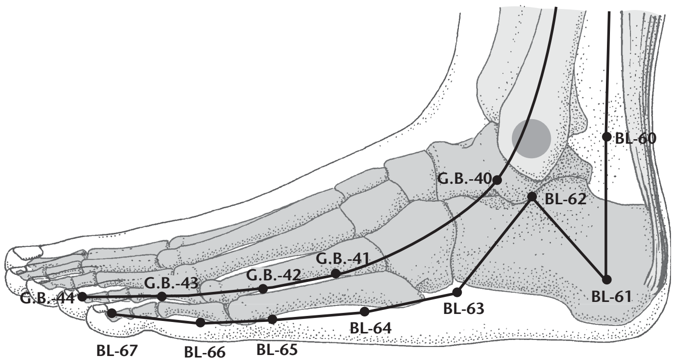
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

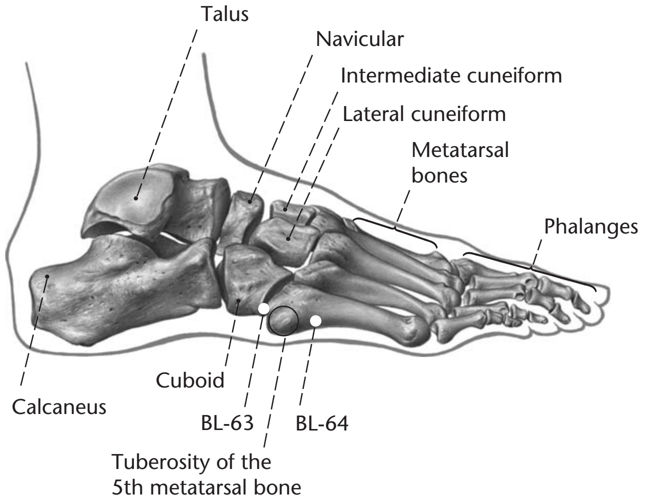
Talus
Navicular
Intermediate cuneiform
Lateral cuneiform
Metatarsal bones
Phalanges
Cuboid
Calcaneus
BL-63
BL-64
Tuberosity of the 5th metatarsal bone

Location

On the lateral aspect of the foot, distal to the tuberosity of the 5th metatarsal bone.

How to find

When palpating along the lateral aspect of the foot on the level of the border of the 'red and white' skin, a bony structure can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal bone (→ 3.6.2). **BL-64** is located directly distal to it (towards the toes), at the junction between the base and the shaft of the 5th metatarsal bone.

→ **BL-63** is located proximal to the tuberosity of the 5th metatarsal bone, in a depression between the calcaneus and cuboid bone (or between the tuberosity and the cuboid bone). → **SP-4** is located in a comparable position on the medial aspect of the foot, in a depression distal to the base of the 1st metatarsal bone.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Clears Wind and Heat from the head and eyes
- Calms the *shen*
- Opens the channel

Special features

Yuan-source point

BL-65 *Restraining Bone* SHUGU

Location

On the lateral aspect of the foot, in the depression proximal to the head of the 5th metatarsal bone.

How to find

When palpating along the bony structures on the lateral aspect of the foot, a prominent bony landmark can be felt approximately at the midpoint of the foot. This is the tuberosity of the 5th metatarsal bone. Distal to it (towards the toes), you can feel a further bony structure, the head of the 5th metatarsal bone. Directly proximal to it, at the border of the 'red and white' skin (border dorsum/sole of the foot), **BL-65** can be palpated in a depression. → **BL-66** is located in a depression distal to the head of the 5th metatarsal bone (distal to the metatarsophalangeal joint), at the junction of the base and the shaft of the proximal phalanx of the little toe.
→ **SP-3** is located in a comparable position on the medial aspect of the foot (proximal to the head of the 1st metatarsal bone). → **S.I.-3** and → **L.I.-3** are located in comparable positions at the ulnar/lateral border of the hand.

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Expels pathogenic factors (especially from the head)
- Clears Heat
- Opens the channel and *luo* vessels, alleviates pain
- Calms the *shen*

Special features

Shu-stream point, Wood point, sedation point.

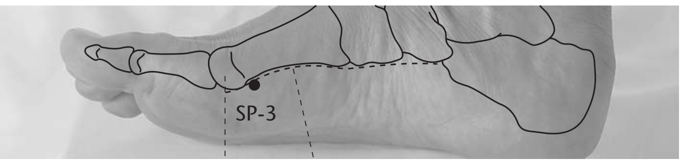

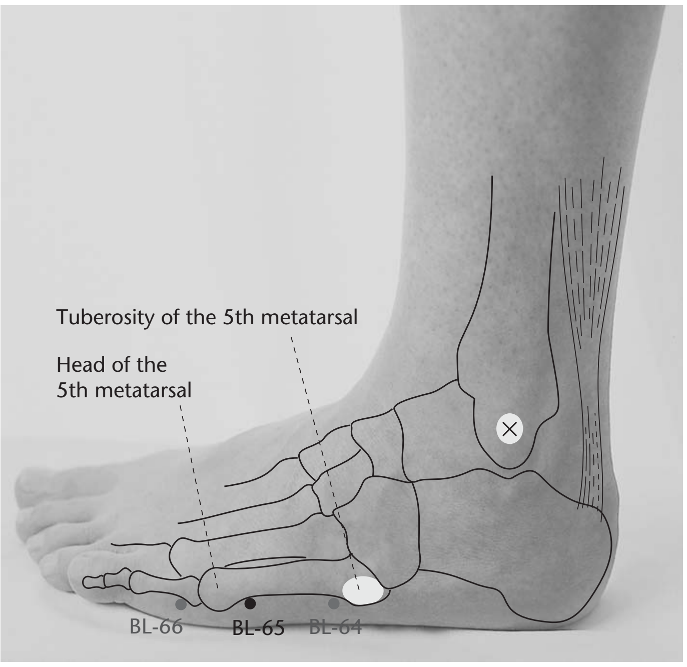
Tuberosity of the 5th metatarsal
Head of the 5th metatarsal
BL-66
BL-65
BL-64

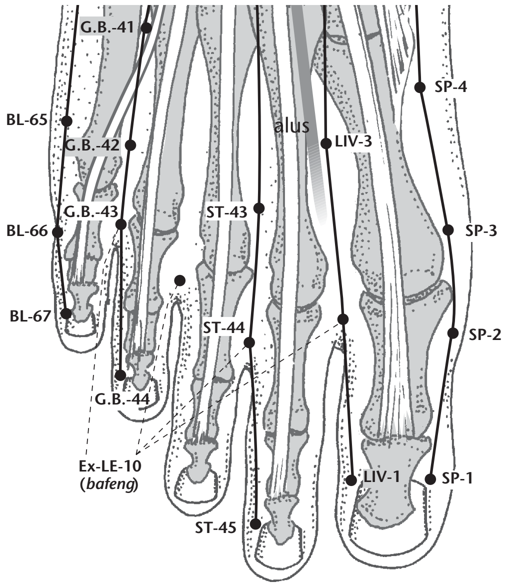
G.B.-41
SP-4
BL-65
G.B.-42
alus
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(bafeng)
LIV-1
SP-1
ST-45

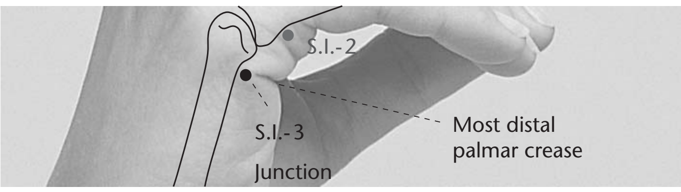
S.I.-2
S.I.-3
Junction
Most distal palmar crease

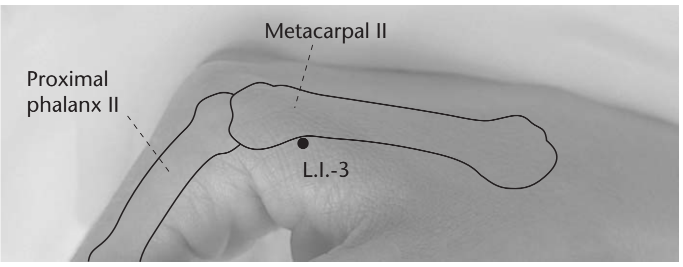
Metacarpal II
Proximal phalanx II
L.I.-3

Foot Connecting Valley ZUTONGGU BL-66

Location

At the lateral border of the foot, in the depression distal to the metatarsophalangeal joint of the little toe.

How to find

When palpating along the border of the 'red and white' skin on the lateral aspect of the foot, at approximately the midpoint of the length of the foot, a clearly defined bony structure can be felt. This is the tuberosity of the 5th metatarsal bone. Distal to it (towards the toes), you can feel a further prominent bony structure, the head of the 5th metatarsal bone (or the 5th metatarsophalangeal joint). **BL-66** is located directly distal to it, in a depression at the junction of the base and the shaft of the proximal phalanx of the little toe.

→ **BL-65** is located proximal to the head of the 5th metatarsal bone. → **SP-2** is located in a comparable position on the medial aspect of the foot. Located in comparable positions on the hand are → **S.I.-2** (on the ulnar aspect) and → **L.I.-2** (on the radial aspect).

Needling

Vertically 0.3–0.5 cun. Caution: Painful point.

Actions/Indications

- Clears the head
- Descends the Lung and Stomach Qi
- Calms the *shen*

Special features

Ying-spring point, Water point, *ben* point (Five Phases point).

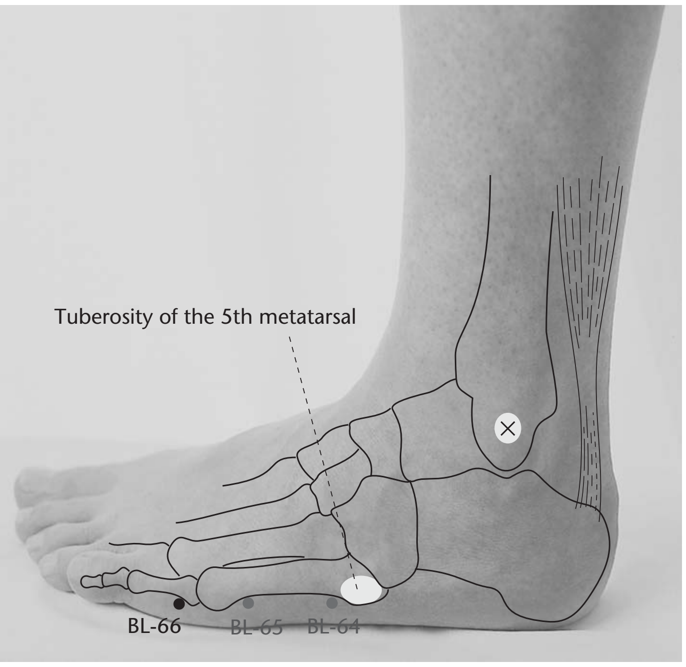
Tuberosity of the 5th metatarsal
BL-66
BL-65
BL-64

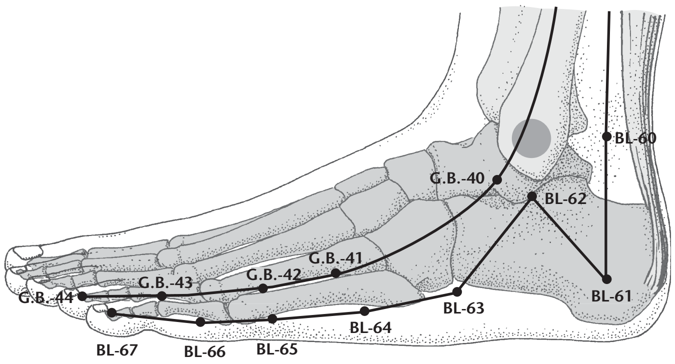
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

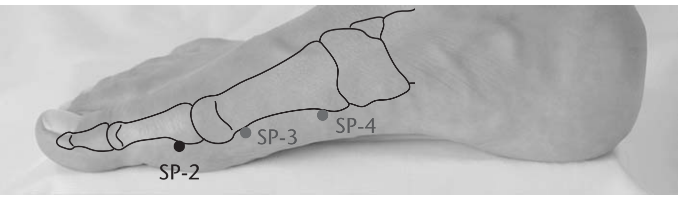

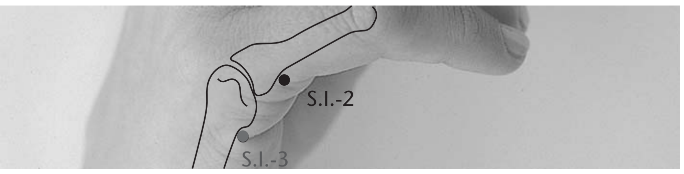

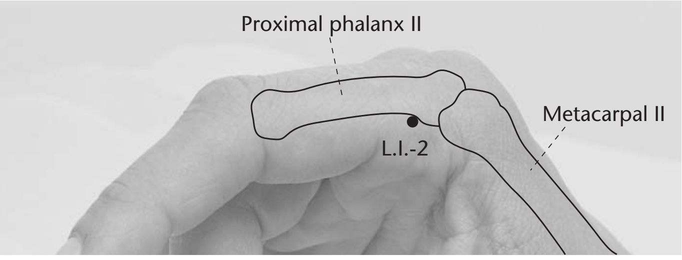
Proximal phalanx II
Metacarpal II
L.I.-2

BL-67 *Reaching Yin* ZHIYIN

Location

On the dorsal aspect of the little toe, 0.1 cun from the lateral corner of the nail.

How to find

This point is located at the junction of two tangents that border the nail of the little toe laterally and proximally, approximately 0.1 cun from the actual corner of the nail.

Needling

Vertically 0.1 cun or obliquely 0.2 cun in the direction of the foot. Prick to bleed. Caution: Painful point.

Actions/Indications

- Scientifically established: conversion of breech presentations prior to birth: moxibustion and electro-acupuncture at this point seem to be most effective (approximately 70–80%)
- Facilitates labour
- Expels pathogenic factors from the opposite end of the channel
- Regulates Yin and Yang of the Water element

Special features

Jing-Well point, Metal point, tonification point, exit point. An isolated case has been reported regarding a pregnant woman who was treated with moxibustion at **BL-67**. This resulted in a fetomaternal transfusion of 300 ml of blood, endangering the fetus. For this reason, treatments should be controlled by cardiotocography.

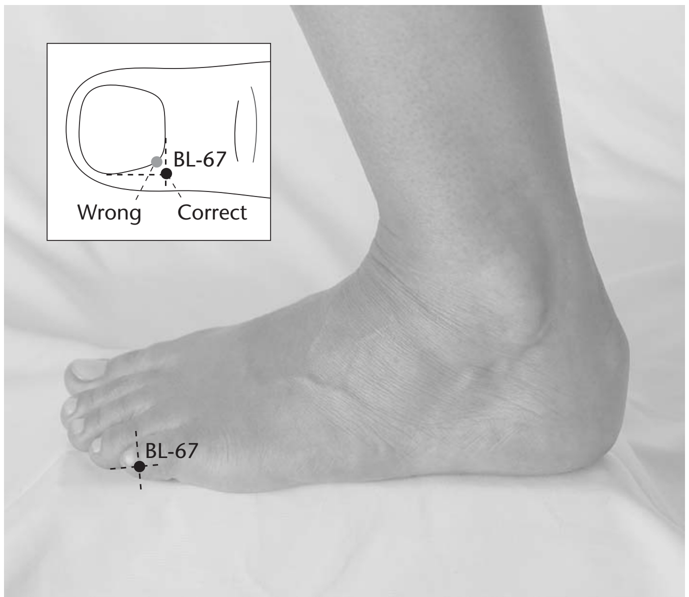
BL-67
Wrong
Correct
BL-67

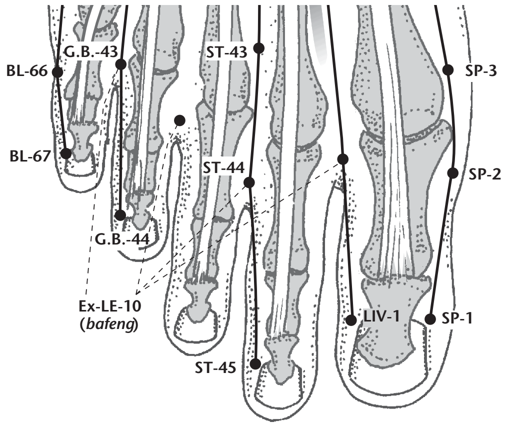
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

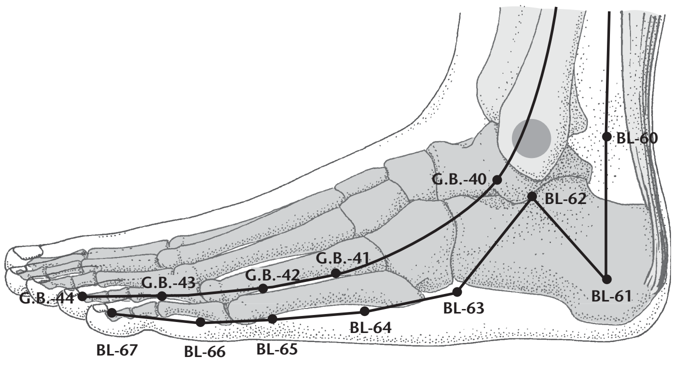
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

4.8 The Kidney Channel System – Foot *shaoyin* (*zu shao yin jing luo*)

4.8.1 The Kidney Primary Channel (*zu shao yin jing*)

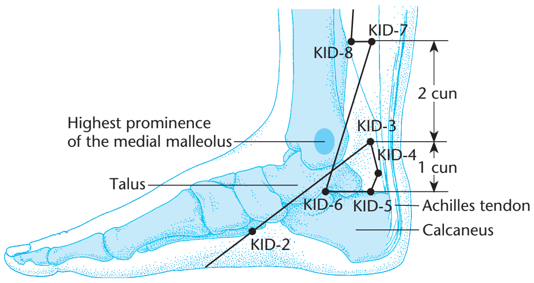
KID-7
KID-8
2 cun
Highest prominence of the medial malleolus
KID-3
KID-4
1 cun
Talus
KID-6
KID-5
Achilles tendon
Calcaneus
KID-2

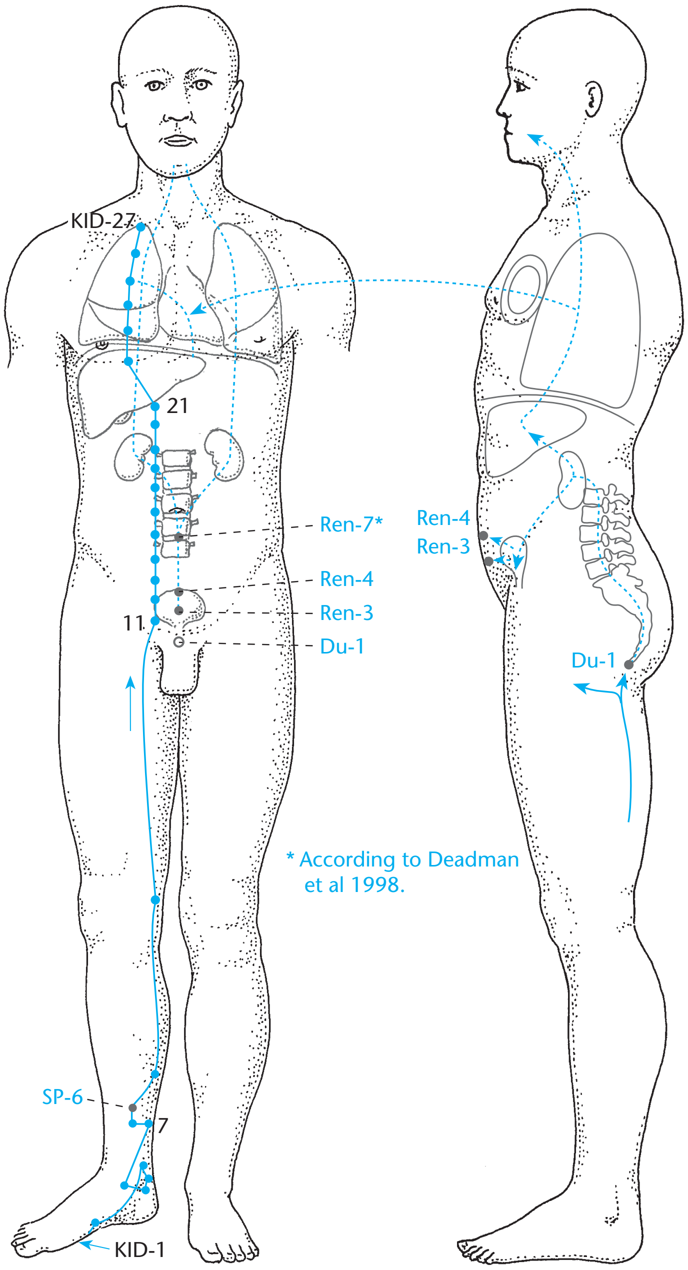
KID-27
21
Ren-7*
Ren-4
Ren-3
Du-1
11
Ren-4
Ren-3
Du-1
SP-6
7
KID-1

* According to Deadman et al 1998.

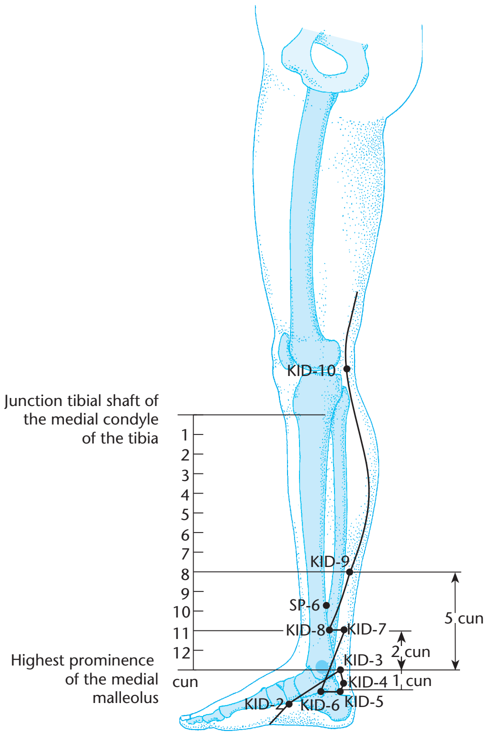
KID-10
Junction tibial shaft of the medial condyle of the tibia
1
2
3
4
5
6
7
8
9
10
11
12
cun
KID-9
SP-6
KID-8
KID-7
5 cun
2 cun
KID-3
KID-4
1 cun
Highest prominence of the medial malleolus
KID-2
KID-6
KID-5

Pathway

The Kidney primary channel begins beneath the little toe, which is reached by **a branch** separating from the Bladder primary channel at its endpoint at → **BL-67** (*zhiyin*) (foot Yin–Yang connection of the second great circuit).

The Kidney primary channel diagonally crosses the sole of the foot to **KID-1** (*yongquan*) and to **KID-2** (*rangu*) below the navicular tuberosity, continues to the posterior aspect of the medial malleolus and descends to enter the heel. From here it ascends the medial aspect of the lower leg, intersecting with → **SP-6** (*yinlingquan*) and continuing to ascend the postero-medial aspect of the thigh towards the perineum.

In the perineal region the primary channel divides into **two branches**:

- the **deep, internal** branch travels to → **Du-1** (*changqiang*) and ascends alongside the spine to connect with its pertaining *zang*-Organ, the Kidneys (*shen*), and its paired *fu*-Organ, the Bladder (*pangguang*). Here it divides into smaller branches which connect with → **Ren-4** (*guanyuan*) and → **Ren-3** (*zhongji*), and, according to some authors, also with → **Ren-7** (*yinjiao*). From the Kidney (*shen*) a branch ascends to the Liver (*gan*), runs to and spreads in the Lung (*fei*), continues to the trachea and terminates at the root of the tongue. From the Lung (*fei*), an **inner branch** travels to the Heart (*xin*), where it meets the Pericardium primary channel (deep Yin–Yin connection). It then disperses in the chest and reaches → **Ren-17** (*shanzhong*).
- the **external** branch separates in the perineal region and runs to the lower abdomen at **KID-11** (*henggu*). From the pubic

Ch04.8-F10028.qxd 2/22/08 6:13 PM Page 316

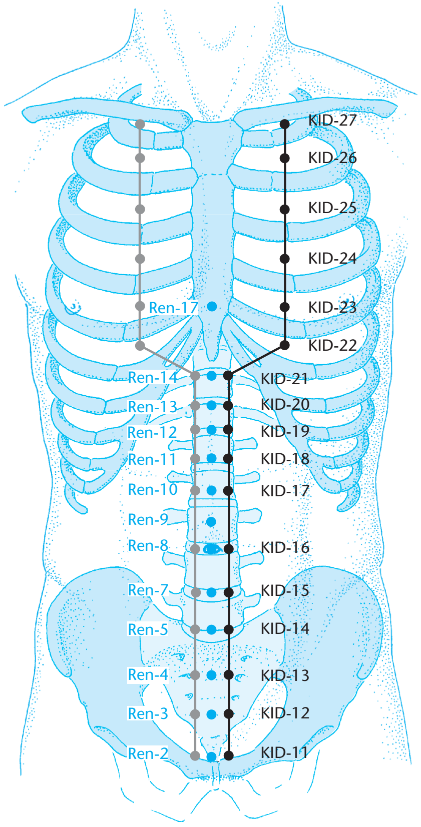

bone it ascends initially 0.5 cun lateral to the anterior midline, from the 5th intercostal space (**KID-22**) 2 cun lateral to the midline to the infraclavicular fossa.

Clinical importance (→ 1.2)

Exterior (*biao*) **signs and symptoms:** Pain in the lumbar region, counterflow Cold or weakness of the legs, dry mouth, sore throat, pain in the lateral gluteal region and on the posterior thigh, pain on the soles of the feet.
Interior (*li*) **or** ***zangfu*****-Organ signs and symptoms:** Pain, facial oedema, dark rings under the eyes, shortness of breath, somnolence, restlessness, diarrhoea, thin or dry stools, abdominal distension, nausea and vomiting, impotence.

Connections of the Kidney primary channel

Connections with other channels

Bladder primary channel (*zu tai yang jing*)

Connection: Foot Yin–Yang connection of the second great circuit
Location: BL-67 → KID-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Heart primary channel (*shou shao yin jing*)

Connection: Paired according to the Six-Channel theory (hand–foot pairing): *shao yin* (Yin axes of the second great circuit)
Location: KID → HE (on the thorax). A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung a branch runs to the Heart (*xin*), there connecting with the Heart primary channel.
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Pericardium primary channel (*shou jue yin jing*)

Connection: Deep Yin–Yin connection
Location: KID → P. A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung an internal branch courses to the Heart (*xin*), where it joins the Pericardium primary channel (deep Yin–Yin connection) and also reaches → **Ren-17** (*shanzhong*).
Circulation: Circadian (according to the Organ clock)
Importance: The Pericardium primary channel receives Nutritive Qi (*ying qi*) from the Kidney primary channel (first circulation of the *ying qi* → 1.1.4).

Connections with other *zangfu*-Organ systems

Kidney (***shen***), **Bladder** (***pangguang***), Liver (*gan*), Lung (*fei*), Pericardium (*xin bao*), Heart (*xin*)

4.8.2 The Kidney Divergent Channel (*zu shao yin jing bie*)

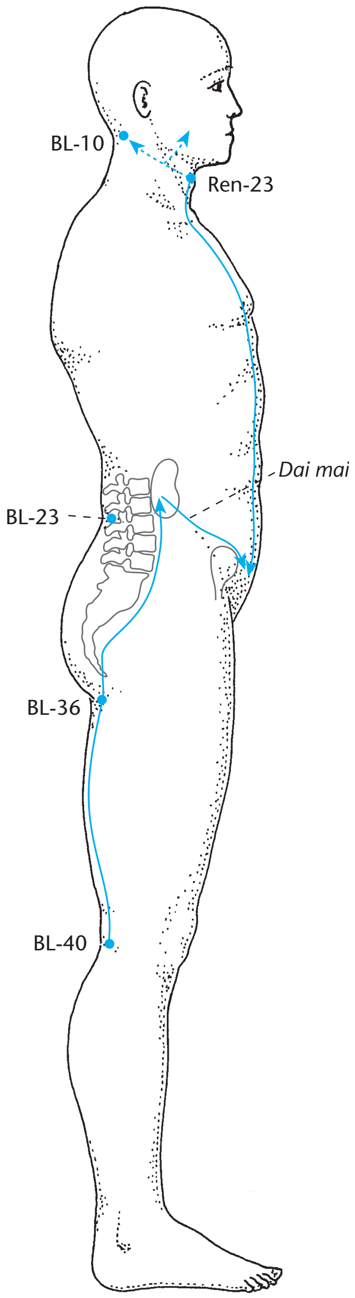

Pathway

The Kidney divergent channel separates from the Kidney primary channel in the popliteal fossa at → **KID-10** (*yinggu*). It

- ➡ meets → **BL-40** (*weizhong*) in the popliteal fossa
- ➡ ascends to → **BL-36** (*chengfu*) in the centre of the gluteal crease
- ➡ penetrates the anus and reaches the Kidney (*shen*) and the Bladder (*pangguang*)
- ➡ ascends to → **BL-23** (*shenshu*) at the level of the lower border of the spinous process of L2, where it also meets the *dai mai*, following it round the waist to the abdomen
- ➡ ascends lateral to the midline to the infraclavicular fossa
- ➡ reaches the neck at → **Ren-23** (*liangquan*)
- ➡ emerges at the nape of the neck at → **BL-10** (*tianzhu*) and converges with the Bladder primary channel and the Bladder divergent channel to form one of the 6 *he*-confluences (here: BL/KID as first confluence → 1.3).

Clinical importance

- Strengthens the relationship between the Kidneys and the Bladder (*zangfu*-Organ systems). Points on the Kidney primary channel can therefore be used for disorders of the Bladder, and vice versa points on the Bladder primary channel can treat disorders of the Kidneys.
- Through the connection with the *dai m*ai, KID points can be used for disorders of the *dai mai*.
- Kidney Qi and Essence (*jing*) are sent to the Brain through the Kidney divergent channel via the spinal canal, thus strengthening the connection between the Kidneys, the Marrow and the Brain.
- Supports the connection to the tongue.

4.8.3 The Kidney Sinew Channel (*zu shao yin jing jin*)

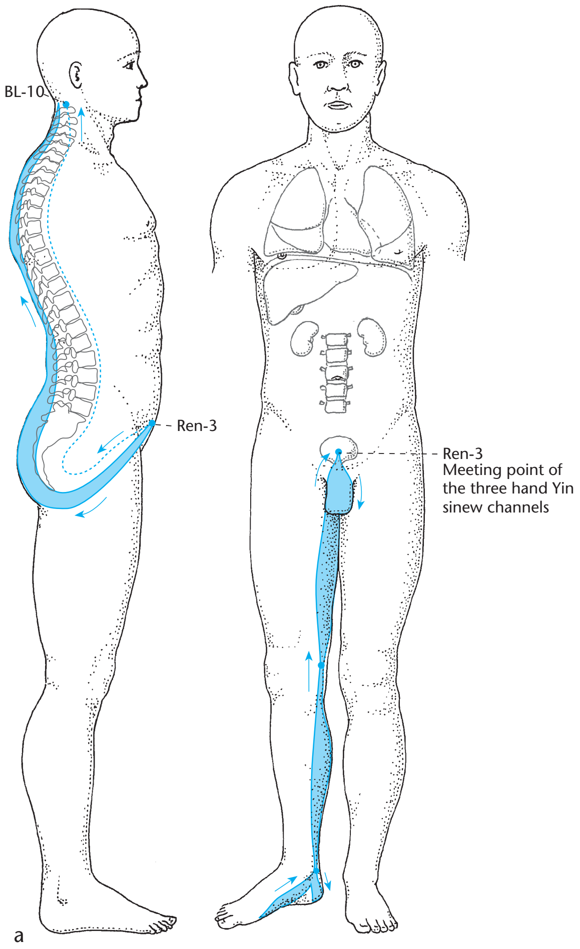

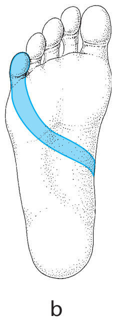

Pathway

The Kidney sinew channel originates beneath the little toe

- diagonally crosses the sole of the foot, meets the Spleen sinew channel and continues to the medial malleolus where it binds (*jie*). From here a **minor branch** travels to the medial portion of the calcaneus. The **main branch** ascends the posteromedial aspect of the leg, binds (*jie*) at the medial condyle of the femur, crosses the Bladder sinew channel, ascends the posteromedial aspect of the thigh to the genital region to → **Ren-3** (*zhongji*) and → **Ren-2** (→ 1.3), where it meets the other Yin sinew channels. It then penetrates the abdomen, runs along the spine to the nape of the neck and there meets the Bladder sinew channel.

From the genital region **an internal branch** traverses the gluteal region and follows the deep musculature along the spine to the occiput, where it meets the Bladder sinew channel.

Clinical importance

Pathology: Spasms on the sole of the foot, pain or spasms along the posteromedial aspect of the lower extremity, the back and the occiput. Chronic *bi*-syndromes. Epileptic seizures, limited range of motion with spinal flexion and extension. Degenerative disorders of the bones and joints, spondylosis, osteoarthritis.

Indication: Mainly for chronic pain, spasms and Cold sensations of the lower back and spine, accompanied by limited range of motion.

4.8.4 The Kidney *luo*-Connecting Vessel System (*zu shao yin luo mai*)

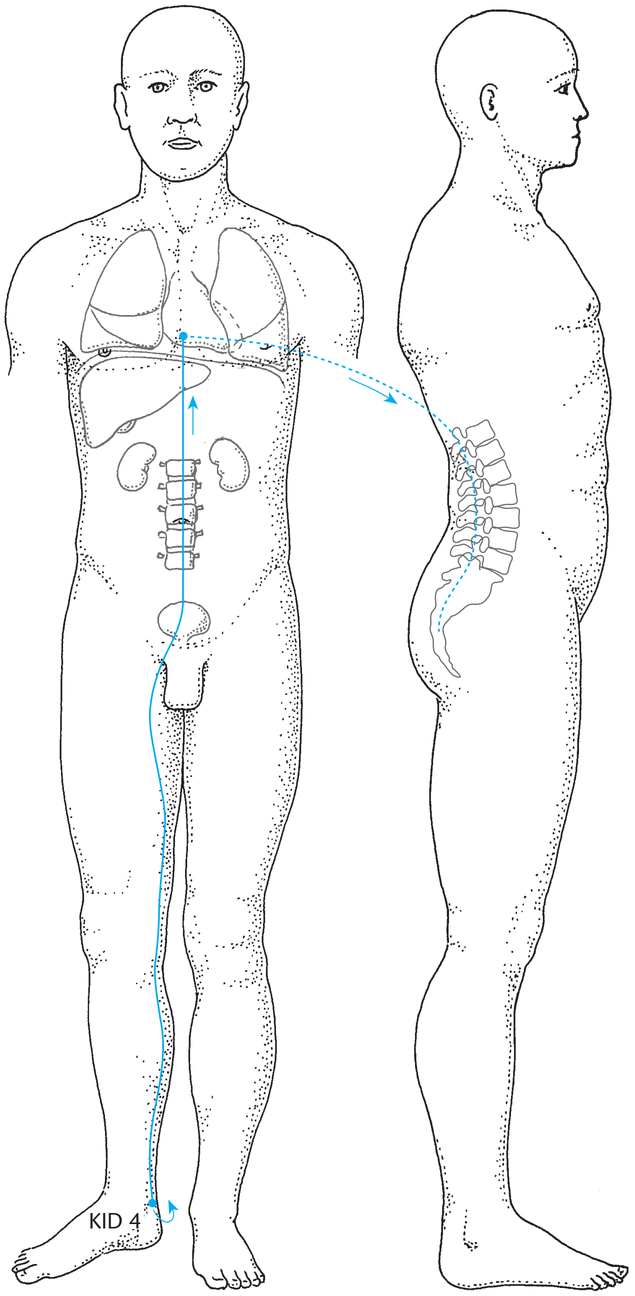

Pathway

The Kidney *luo*-connecting channel separates from the Kidney primary channel at **KID-4** (*dazhong*). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo*, *fu luo*, *xue luo* → 1.5) within the surrounding tissue.

- ➡ Horizontal divisions run to the Interiorly–Exteriorly paired Bladder primary channel; according to some schools of thought (for example Ngyen Van Nghi → Appendix) they travel as a **transverse** KID *luo*-connecting vessel to the *yuan*-source point **BL-64** (*jinggu*).
- ➡ A **longitudinal** division follows the Kidney primary channel to → **Ren-17** (*shanzhong*), penetrates the thorax and reaches the spine in the lumbar region.

Clinical importance (→ 8.1.2)

Pathology
Excess (*shi*): Retention of urine
Deficiency (*xu*): Pain in the lumbar region
Counterflow Qi: Restlessness, anxiety, fear, depression, fullness in the chest and epigastrium.

4.8.5 Cutaneous Region (*shao yin pi bu*)

See description and figures → 1.6.

4.8.6 Points of the Kidney Primary Channel (Overview)

Specific points according to their functions

- ***Yuan*-source point (→ 8.1.1): KID-3** (*taixi*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): KID-4** (*dazhong*)
- ***Xi*-cleft point (→ 8.1.3): KID-5** (*shuiquan*)
- **Associated Back-*shu* point (→ 8.1.4): BL-23** (*shenshu*) 🟦🟦
- **Associated Front-*mu* point (→ 8.1.5): G.B.-25** (*jingmen*) ■■
- **Five *shu*-transporting point (→ 8.1.6):**
  jing-well point (Wood), sedation point: **KID-1** (*yongquan*)
  ying-spring point (Fire): **KID-2** (*rangu*)
  shu-stream point (Earth): **KID-3** (*taixi*) 🟦■
  jing-river point (Metal), tonification point: **KID-7** (*fuliu*) 🟦■
  he-sea point (Water), *ben* point (Five Phases): **KID-10** (*yingu*) 🟦■
- *Hui*-meeting point (→ 8.1.7): –
- Opening point (→ 8.1.8) of the *yin qiao mai*: **KID-6** (*zhaohai*) 🟦■
- Lower *he*-sea points (→ 8.1.9): –
- *Jiaohui*-meeting points (→ 8.1.10):
  - with the *yin qiao mai*: **KID-2** (*rangu*)*, (**KID-6** (*zhaohai*)*, **KID-8** (*jiaoxin*)*)
  - with the *yin wei mai*: **KID-9** (*zhubin*)
  - with the *chong mai*: **KID-11–KID-21**
  - of other channels with the Kidney channel: **SP-6**, **Du-1**, **Ren-4**, **Ren-3**, **Ren-7***, **Ren-17**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**

---

* Mentioned by only some authors

- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - *xi*-cleft point of the *yin qiao mai*: **KID-8** (*jiaoxin*)
  - *xi*-cleft point of the *yin wei mai*: **KID-9** (*zhubin*)

Points according to region

- **Local points** (→ 8.2.1): foot – **KID-6** (*zhaohai*) 🟦🟦; knee – **KID-10** (*yingu*) 🟦⬛
- **Adjacent points** (→ 8.2.1): foot – **KID-7** (*fuliu*) 🟦⬛; toes – **KID-6** (*zhaohai*) 🟦🟦
- **Distal points** (→ 8.2.1): for the throat – **KID-6** (*zhaohai*) 🟦🟦; for the tongue – **KID-6** (*zhaohai*) 🟦🟦; for the Kidneys and the Bladder – **KID-3** (*taixi*) 🟦⬛, **KID-7** (*fuliu*) 🟦🟦; for the genito-urinary region – **KID-3** (*taixi*) 🟦🟦

Specific points according to the channel pathway (in numerical order)

- **KID-1** (*yongquan*) 🟦⬛: *jing*-well point (Wood); sedation point
- **KID-2** (*rangu*): *ying*-spring point (Fire); *jiaohui*-meeting point with the *yin qiao mai* (→ 8.1.10)
- **KID-3** (*taixi*) 🟦🟦: *shu*-stream point (Earth); distal point for the Kidneys, Bladder, urogenital region (→ 8.2.1); *yuan*-source point (→ 8.1.1)
- **KID-4** (*dazhong*): *luo*-connecting point (→ 8.1.2)
- **KID-5** (*shuiquan*): *xi*-cleft point (→ 8.1.3)
- **KID-6** (*zhaohai*) 🟦🟦: opening point (→ 8.1.8) of the *yin qiao mai*; distal point for the tongue and throat (→ 8.2.1); *jiaohui*-meeting point with the *yin qiao mai** (→ 8.1.10); local point for the foot (→ 8.2.1)
- **KID-7** (*fuliu*) 🟦⬛: distal point for the Kidneys and the Bladder (→ 8.2.1); *jing*-river point (Metal); tonification point; regional point for the foot (→ 8.2.1)
- **KID-8** (*jiaoxin*): *xi*-cleft point of the *yin qiao mai*; *jiaohui*-meeting point with the *yin qiao mai** (→ 8.1.10)
- **KID-9** (*zhubin*): *xi*-cleft point of the *yin wei mai*; *jiaohui*-meeting point with the *yin wei mai** (→ 8.1.10)
- **KID-10** (*yingu*) 🟦⬛: *he*-sea point (Water), *ben* point (Five Phases); local point for the knee (→ 8.2.1)
- **KID-11** to **KID-21**: *jiaohui*-meeting points with the *chong mai* (→ 8.1.10)

General location help

- **KID-11** to **KID-21** are all located 0.5 cun lateral to the midline
  - **KID-11** to **KID-15** are spread on the 5 cun distance (→ 1.2) between the upper border of the pubic symphysis and the umbilicus (they are spaced at 1 cun intervals from the upper border of the pubic symphysis to 4 cun above it)
  - **KID-16** to **KID-21** are spread on the 8 cun distance (→ 1.2) between the umbilicus and the sternocostal angle (they are spaced at 1 cun intervals from the level of the umbilicus to 6 cun superior to it).
- **KID-22** to **KID-27** are all located 2 cun lateral to the midline
  - **KID-22** to **KID-26** are located in the 5th to 1st intercostal spaces
  - **KID-27** is located on the lower border of the sternoclavicular joint

---

* Mentioned by only some authors

Gushing Spring YONGQUAN **KID-1**

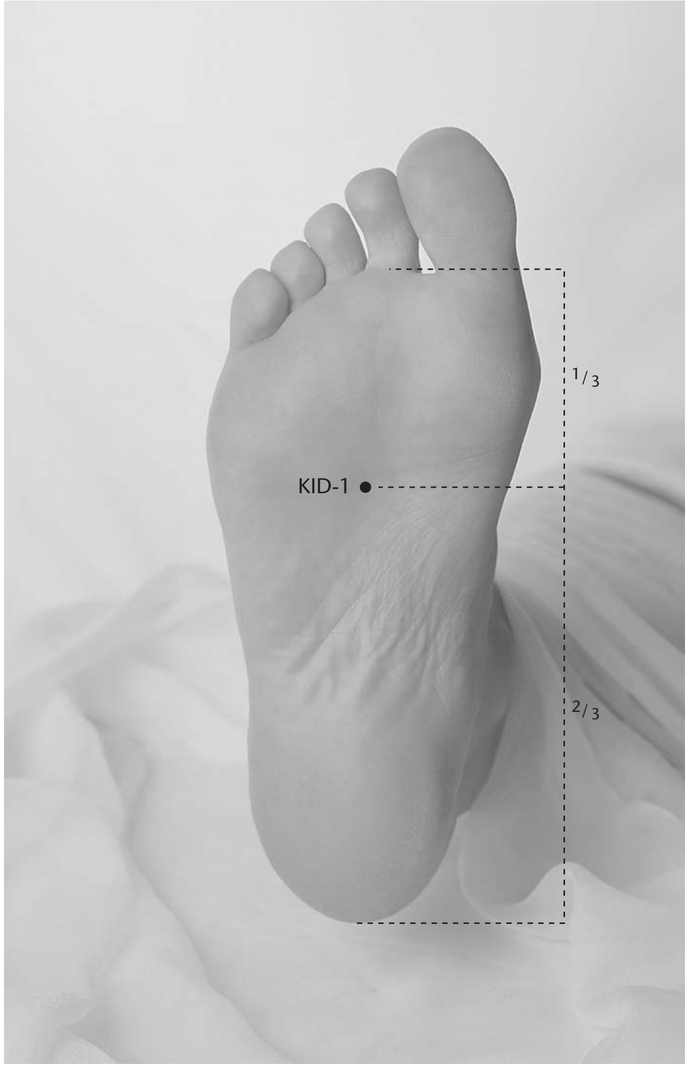

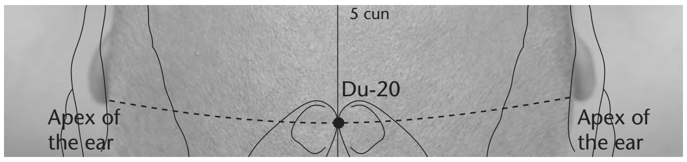

Location

On the sole of the foot, in a depression between the 2nd and 3rd metatarsal bones, at the junction of the anterior third and the posterior two-thirds of the sole.

How to find

Measure one third from the anterior border of the sole of the foot. Palpate for a pressure-sensitive point dorsal to the balls of the foot, at the midpoint of the width of the foot.

Needling

Needling this point is very painful and is therefore only recommended for severe disorders or in an emergency. Acupressure is commonly used instead; moxibustion is possible. For the treatment of collapse, unconsciousness, shock or a severe excess condition, this point should be strongly stimulated; in pronounced deficiency conditions, stimulation should be applied more cautiously.

Actions/Indications

- Revives collapsed Yang
- Clears excess, Heat and Wind from the head, lowers Yang
- Calms the *shen*

Special features

Jing-well point, Wood point, sedation point, entry point. **KID-1** is the lowest acupuncture point on the body and the only point on the sole of the foot. For this reason, it plays an important role in Qigong, forming a grounding point in the centre of the foot, where the energies of the Earth and Man unite. During practice, the centre of gravity should be above this point.

KID-2 *Blazing Valley* RANGU

Location

At the medial border of the foot, in a depression at the anterior border of the navicular bone, at the border of the 'red and white' skin.

How to find

Palpate from distal to proximal along the tarsal section of the medial aspect of the foot, past the shaft and the head of the 1st metatarsal bone (→ **SP-4**) and the medial cuneiform bone, finally reaching the prominent navicular bone. Locate **KID-2** anterior to the navicular bone, at the inferior angle of the joint between the medial cuneiform bone and the navicular bone.

Needling

From the medial aspect of the foot, 0.5–1 cun vertically below the border of the bone. Use reducing needling techniques for empty Heat. Use tonifying techniques for Kidney Yang/Qi deficiency.

Actions/Indications

- Clears empty Heat
- Regulates the Lower Burner and the Kidneys
- Local point

Special features

Ying-spring point, Fire point, meeting point with the *yin qiao mai*.

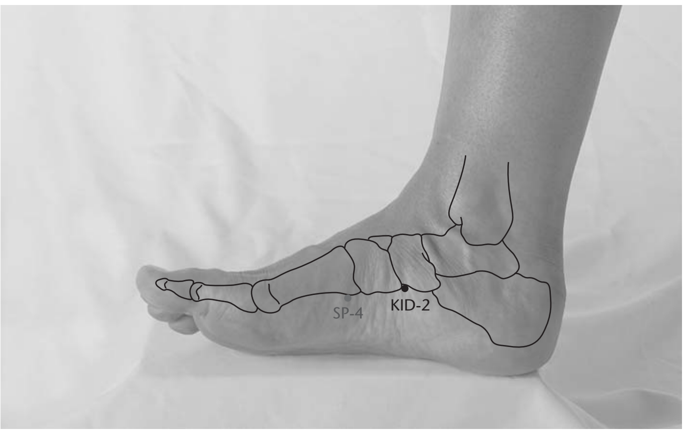

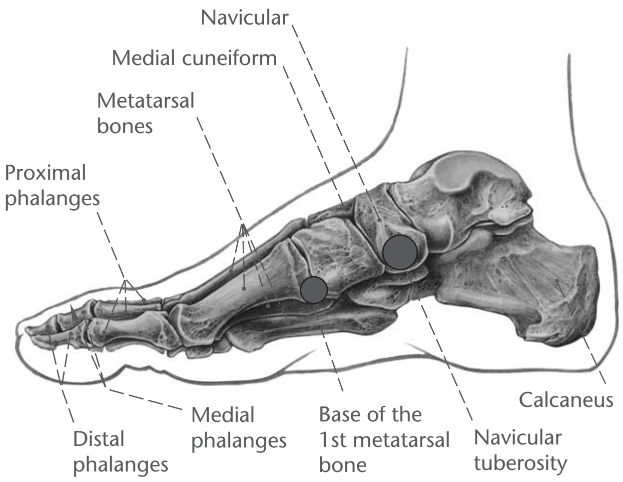
Navicular
Medial cuneiform
Metatarsal bones
Proximal phalanges
Distal phalanges
Medial phalanges
Base of the 1st metatarsal bone
Calcaneus
Navicular tuberosity

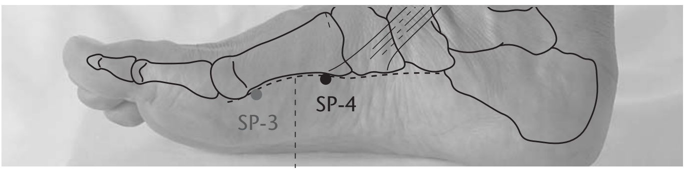

Ch04.8-F10028.qxd 2/22/08 6:13 PM Page 323

Supreme Stream TAIXI KID-3

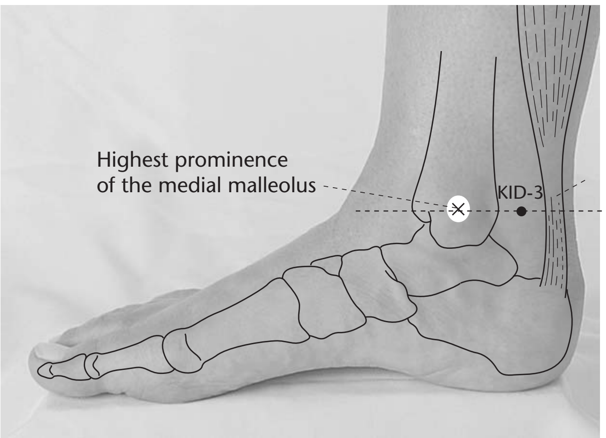
Highest prominence
of the medial malleolus
KID-3

Location

In the depression between the highest prominence of the medial malleolus and the Achilles tendon.

How to find

Locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate horizontally towards the Achilles tendon. **KID-3** is located in a depression that can be palpated anterior to the tendon.

Needling

Vertically 0.3–1 cun

Actions/Indications

- Nourishes Kidney Yin and clears deficiency Heat, tonifies Kidney Yang, stabilises the Kidney Qi and the Lungs (helping the Kidneys absorb the Qi), regulates menstruation
- Strengthens the lower back, local point

Special features

Yuan-source point, *shu*-stream point, Earth point. Important point for tonifying the Kidneys (especially Kidney Yin).

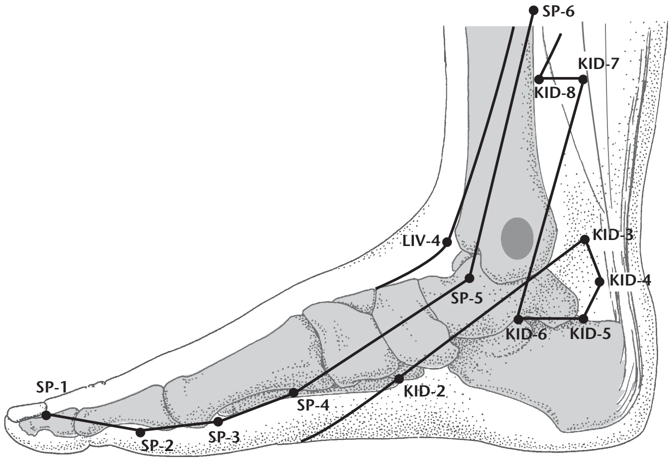
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

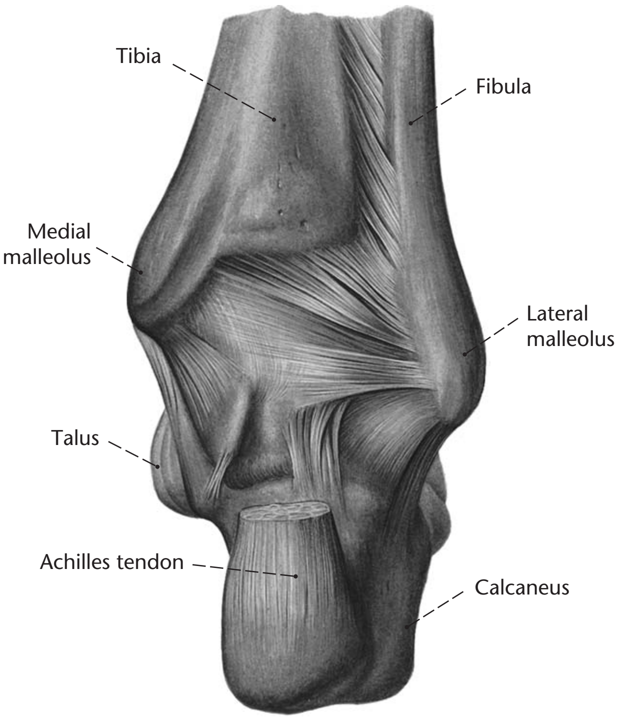
Tibia
Fibula
Medial malleolus
Lateral malleolus
Talus
Achilles tendon
Calcaneus

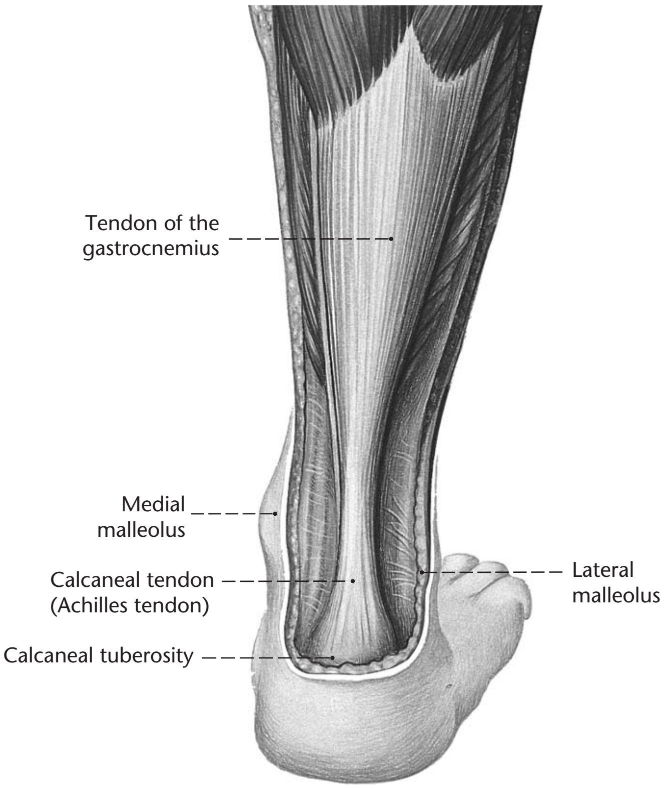
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

KID-4 *Great Bell* DAZHONG

Location

Anterior to the medial border of the Achilles tendon, superior to its insertion at the calcaneus.

How to find

From the highest prominence of the medial malleolus, draw a horizontal line to the medial border of the Achilles tendon. From there, measure 0.5 cun in a distal direction. **KID-4** is located in a depression anterior to the Achilles tendon, slightly superior to its insertion at the calcaneus. **Or: KID-4** is located posterior to the midpoint of a line connecting → **KID-3** and → **KID-5** anterior to the Achilles tendon.

Needling

Vertically 0.3–0.5 cun. Avoid puncturing the tendon.

Actions/Indications

- Firms the Kidney Qi (helping the Kidneys absorb the Qi) and supports the Lungs
- Strengthens the Kidneys and cools empty Heat
- Strengthens the will, dispels fear
- Local point

Special features

Luo-connecting point

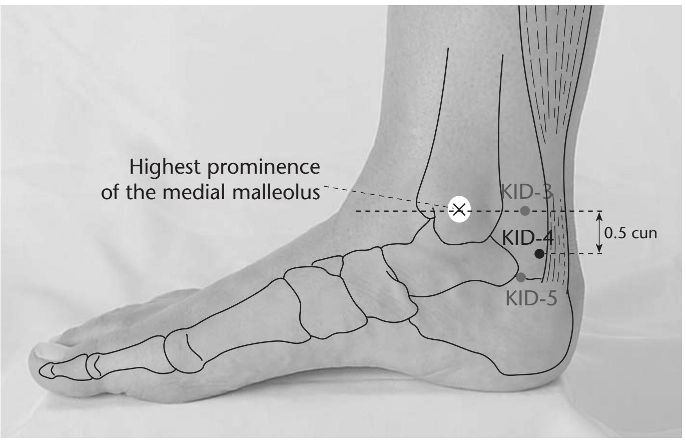
Highest prominence of the medial malleolus
KID-3
KID-4
0.5 cun
KID-5

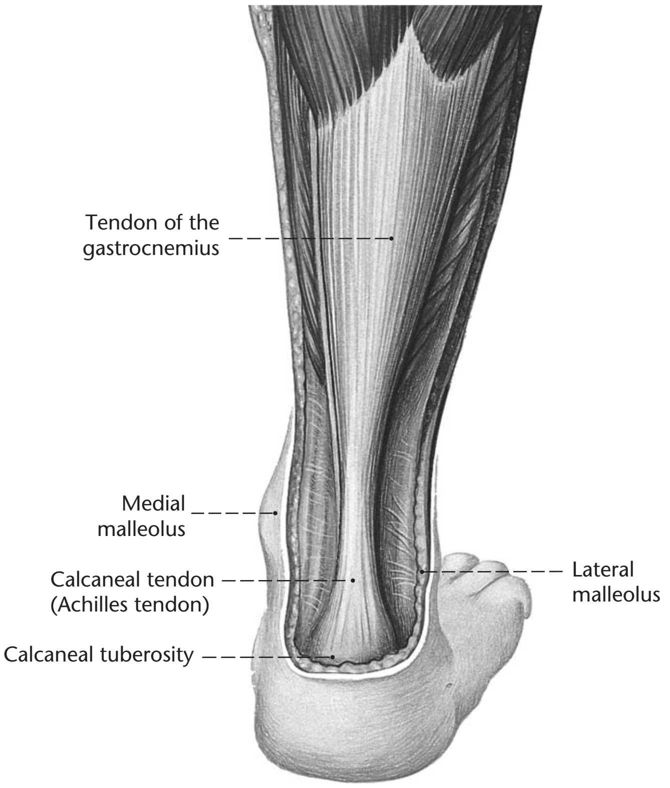
Tendon of the gastrocnemius
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

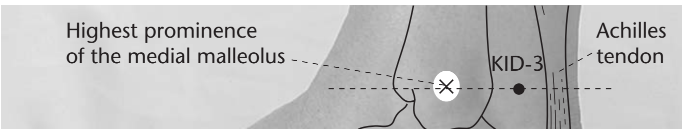
Highest prominence of the medial malleolus
KID-3
Achilles tendon

Water Spring SHUIQUAN KID-5

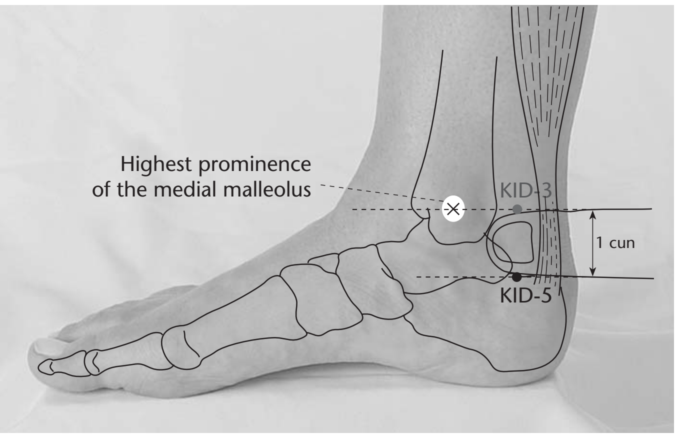
Highest prominence of the medial malleolus
KID-3
1 cun
KID-5

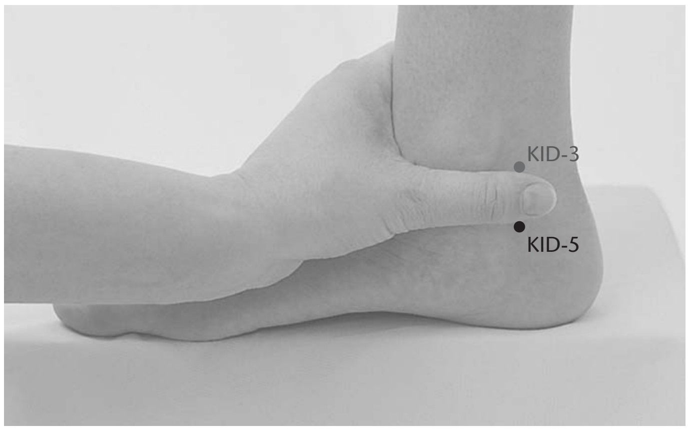

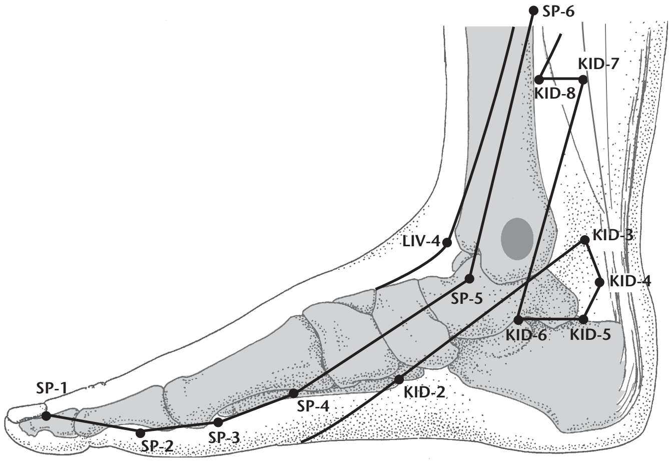
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

Location

1 cun distal to → **KID-3**, in a depression over the joint space between the talus and calcaneus.

How to find

First, locate → **KID-3** on the level of the highest prominence of the medial malleolus (→ 3.6.2), in a depression between the malleolus and the Achilles tendon. From **KID-3**, palpate 1 cun in an inferior direction to a palpable depression over the joint space between the talus and calcaneus. This is the location of **KID-5**.

Needling

0.3–0.5 cun obliquely to vertically superior to the margin of the bone.

Actions/Indications

- Regulates the *chong mai* and *ren mai*, benefits menstruation and urination

Special features

Xi-cleft point. Especially indicated for acute disorders and painful conditions.

KID-6 *Shining Sea* ZHAOHAI

Location

Approximately 1 cun inferior to the highest prominence of the medial malleolus, over the joint space between the talus and the calcaneus.

How to find

From the highest prominence of the medial malleolus (→ 3.6.2), palpate distally, until the palpating finger can feel the depression of the joint space between the talus and the calcaneus. (**KID-6** is often located between the tendons of the tibialis and flexor digitorum longus muscles.) Directly distal to this point, you can palpate a small bony prominence, the sustentaculum tali. With the patient's foot in a supinated position, a visible skin crease forms on the level of the point. Note: There are several variations regarding the exact location of this point. However, the determining aspect for its location is not the cun measurement but the joint space below the prominence.

Needling

0.3–0.5 cun vertically or obliquely in a proximal direction. The needle may reach the deltoid ligament.

Actions/Indications

- Nourishes Kidney Yin, clears empty Heat, benefits the throat, regulates the Lower Burner
- Calms the *shen*
- Moves Qi locally

Special features

Opening (master) point of the *yin qiao mai*. Important point, especially for strengthening the Kidney Yin.

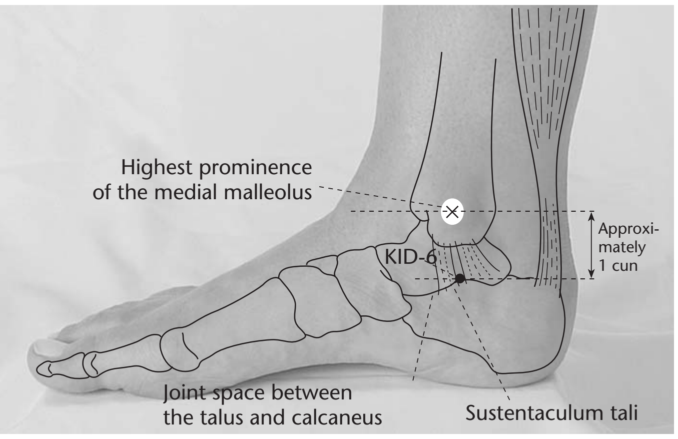
Highest prominence
of the medial malleolus
KID-6
Approxi-
mately
1 cun
Joint space between
the talus and calcaneus
Sustentaculum tali

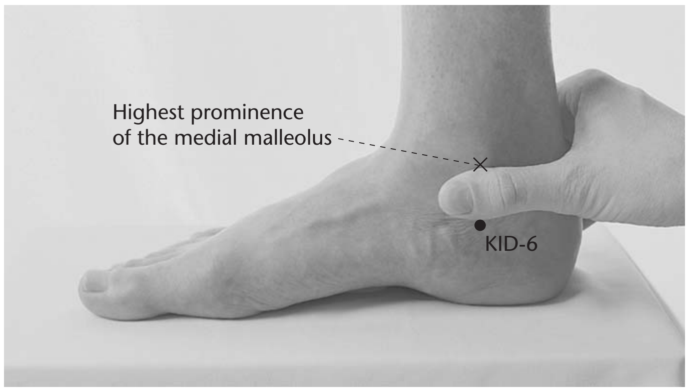
Highest prominence
of the medial malleolus
KID-6

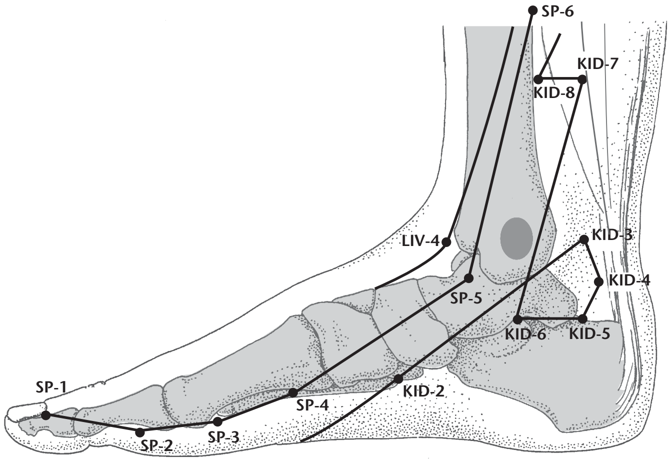
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

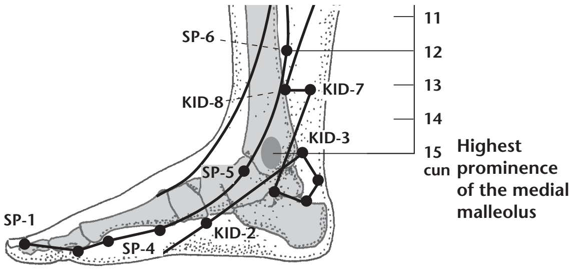
11
12
13
14
15
cun
SP-6
KID-8
KID-7
KID-3
SP-5
SP-1
KID-2
SP-4
Highest
prominence
of the medial
malleolus

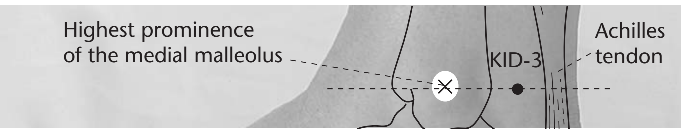
Highest prominence
of the medial malleolus
KID-3
Achilles
tendon

Returning Current FULIU KID-7

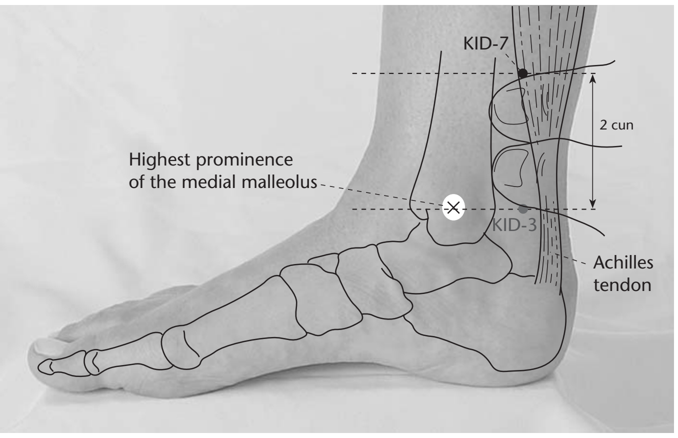
KID-7
2 cun
Highest prominence of the medial malleolus
KID-3
Achilles tendon

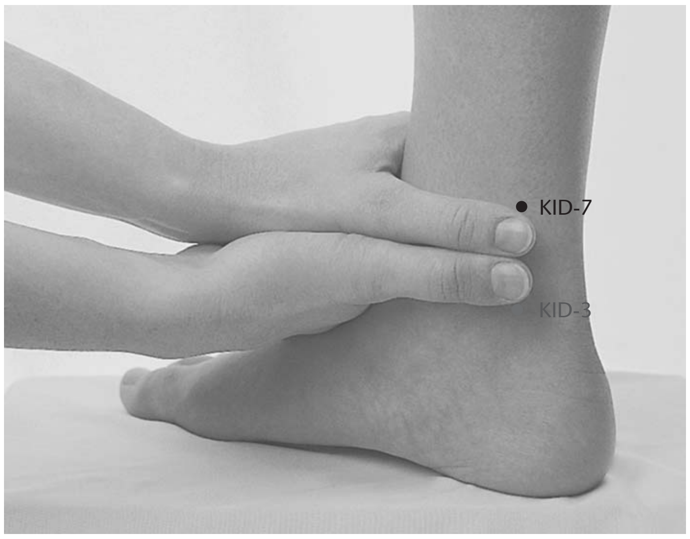

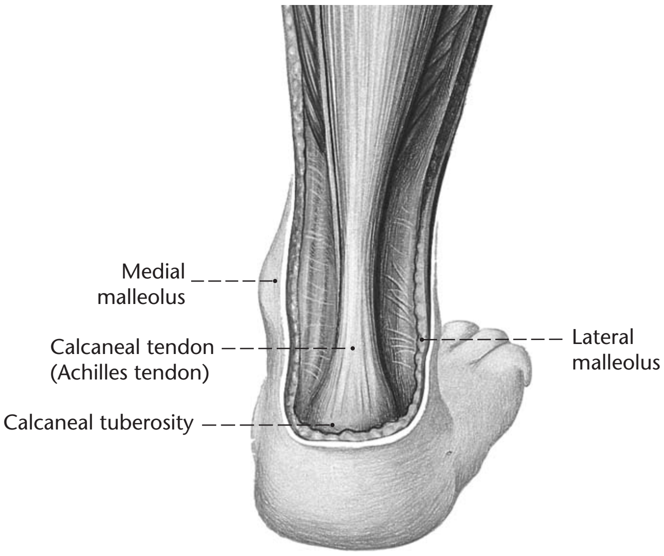
Medial malleolus
Calcaneal tendon (Achilles tendon)
Calcaneal tuberosity
Lateral malleolus

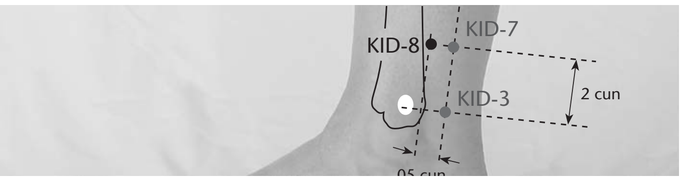
KID-8
KID-7
KID-3
2 cun

Location

2 cun proximal to → **KID-3**, at the anterior border of the Achilles tendon.

How to find

First, locate → **KID-3** on the level of the higest prominence of the medial malleolus (→ 3.6.2), in the depression between the malleolus and the Achilles tendon. From → **KID-3**, measure 2 cun in a proximal direction (towards the knee joint) and locate **KID-7** in a depression at the anterior border of the Achilles tendon.
Located on the same level, but more medially, is → **KID-8** (2 cun directly proximal to the medial malleolus).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates the water passages and treats oedema, strengthens the Kidneys (especially Kidney Yang), drains Dampness and clears Damp-Heat
- For disorders of the Intestines due to Damp-Heat
- Strengthens the lumbar region

Special features

Jing-river point, Metal point, tonification point

KID-8 *Exchange Belief* JIAOXIN

Location

2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia.

How to find

KID-8 is located on the same level as → **KID-7**, but slightly closer to the border of the tibia (approximately 0.5 cun).

Needling

Vertically 0.5–1 cun

Actions/Indications

- Regulates menstruation, regulates the *ren mai* and *chong mai*
- Clears Heat and eliminates Dampness from the Lower Burner
- Treats blockages of the *yin qiao mai*

Special features

Xi-cleft point of the *yin qiao mai*

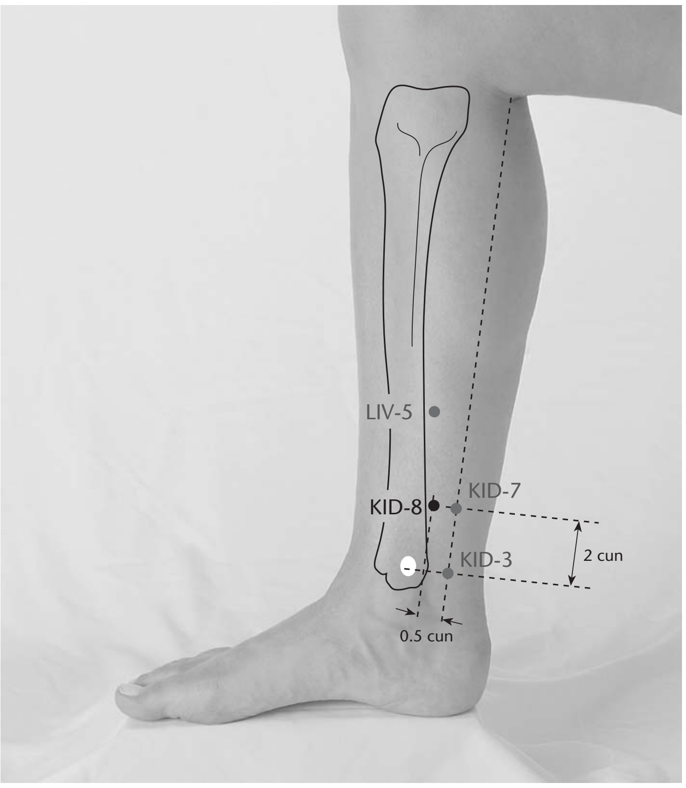
LIV-5
KID-7
KID-8
2 cun
KID-3
0.5 cun

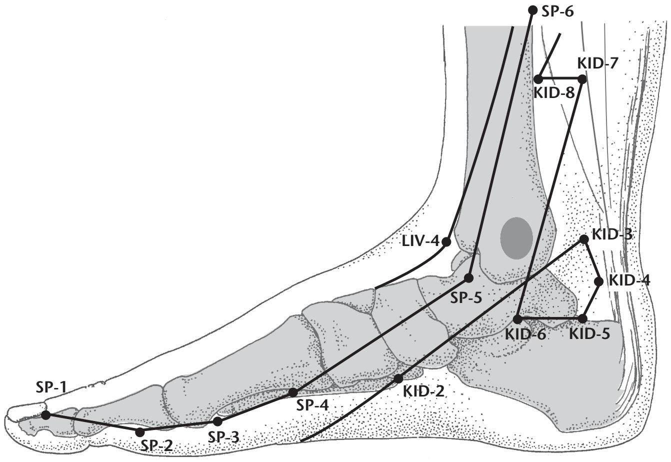
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-2
SP-3

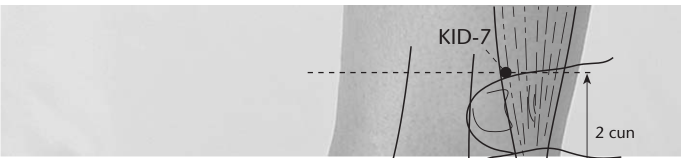

Guesthouse ZHUBIN KID-9

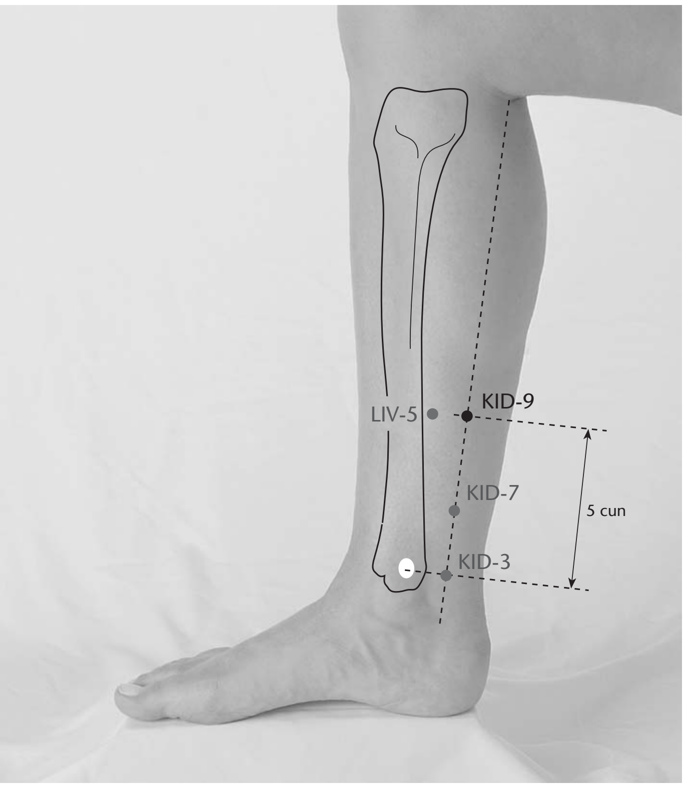
LIV-5
KID-9
KID-7
KID-3
5 cun

Location

5 cun proximal to the highest prominence of the medial malleolus, 2 cun posterior to the medial border of the tibia.

How to find

On a line connecting → **KID-3** (posterior to the medial malleolus) and → **KID-10** (medial aspect of the popliteal crease), measure 5 cun from the level of the prominence of the medial malleolus. Here, **KID-9** is located 2 cun posterior to the medial border of the tibia.

Located on the same level, but directly posterior to the medial border of the tibia, is → **LIV-5**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears the Heart and transforms Phlegm
- Regulates the Qi and alleviates pain

Special features

Xi-cleft point of the *yin wei mai*

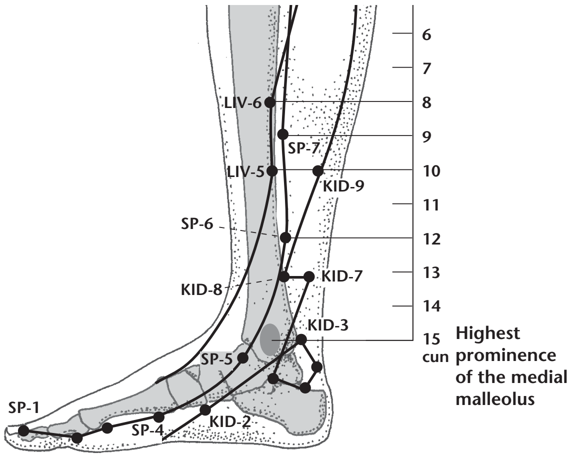
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
13
KID-8
KID-7
14
KID-3
15
cun
Highest prominence of the medial malleolus
SP-5
SP-1
KID-2
SP-4

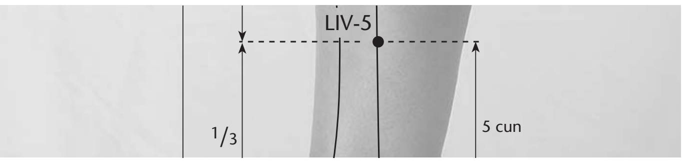
LIV-5
1/3
5 cun

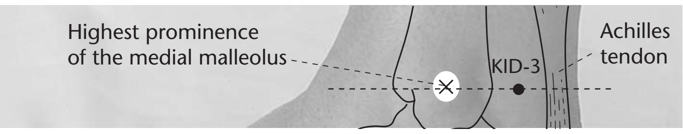
Highest prominence of the medial malleolus
KID-3
Achilles tendon

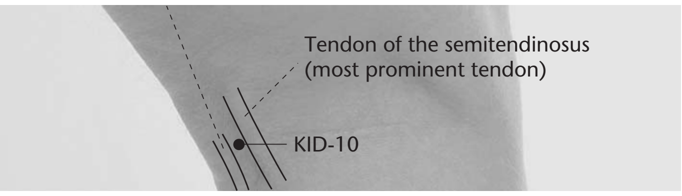
Tendon of the semitendinosus (most prominent tendon)
KID-10

KID-10 *Yin Valley* YINGU

Location

At the medial end of the popliteal crease, between the tendons of the semimembranosus and semitendinosus muscles, on the level of the knee joint space.

How to find

With the knee flexed less than 90°, ask the patient to press their heel against the treatment couch so that the two tendons will become more pronounced in the popliteal crease. From a medial direction, **KID-10** is located in a small gap anterior to the more prominent tendon of the semitendinosus muscle and posterior to the less well-defined tendon of the semimembranosus muscle.
→ **BL-40** is located on the same level, in the centre of the popliteal crease.

Needling

Needle 1–1.5 cun vertically from a dorsomedial aspect towards the tibial tuberosity. Avoid puncturing the tendon.

Actions/Indications

- Drains Damp-Heat from the Lower Burner and benefits the Kidneys
- Opens the channel

Special features

He-sea point, *ben* point (Five Phases point); for recurring urinary disorders with Kidney Yin deficiency.

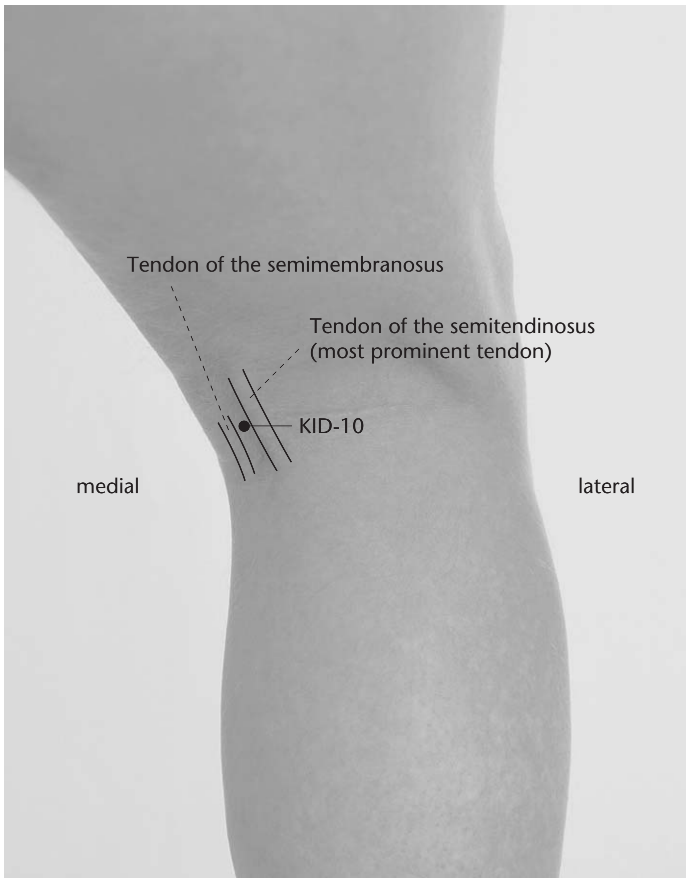
Tendon of the semimembranosus
Tendon of the semitendinosus (most prominent tendon)
KID-10
medial
lateral

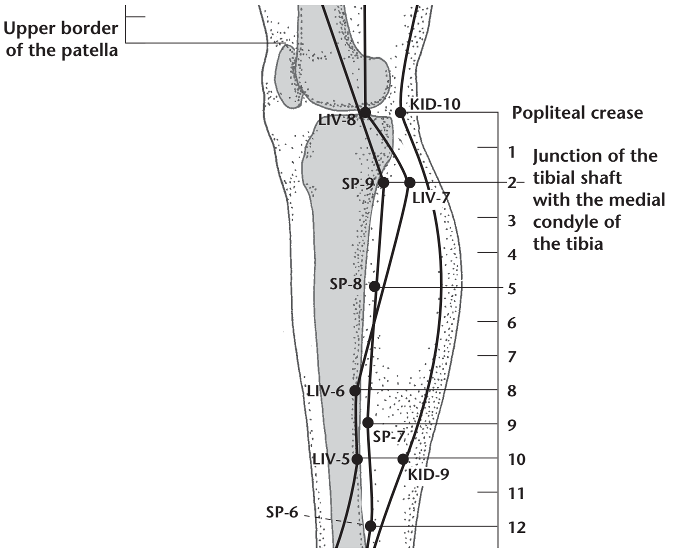
Upper border of the patella
KID-10
Popliteal crease
LIV-8
SP-9
LIV-7
1
2
3
4
Junction of the tibial shaft with the medial condyle of the tibia
SP-8
5
6
7
8
LIV-6
9
SP-7
10
LIV-5
KID-9
11
SP-6
12

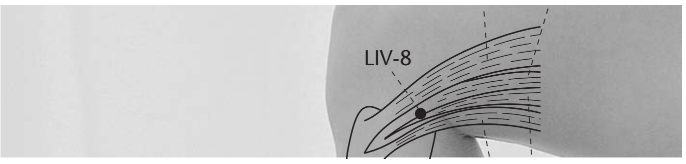

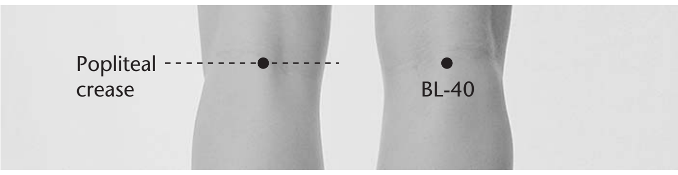
Popliteal crease
BL-40

Pubic Bone HENGGU KID-11

Location

At the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

Close to the midline, the upper border of the pubic symphysis is generally easily palpable. **KID-11** is located directly superior to its bony margin and 0.5 cun lateral to the midline.

Located on the same level are → **Ren-2** (on the midline), → **ST-30** (2 cun lateral to the midline) and → **SP-12** (3.5 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner and the water passages, strengthens the Kidneys

Special features

Meeting point with the *chong mai*

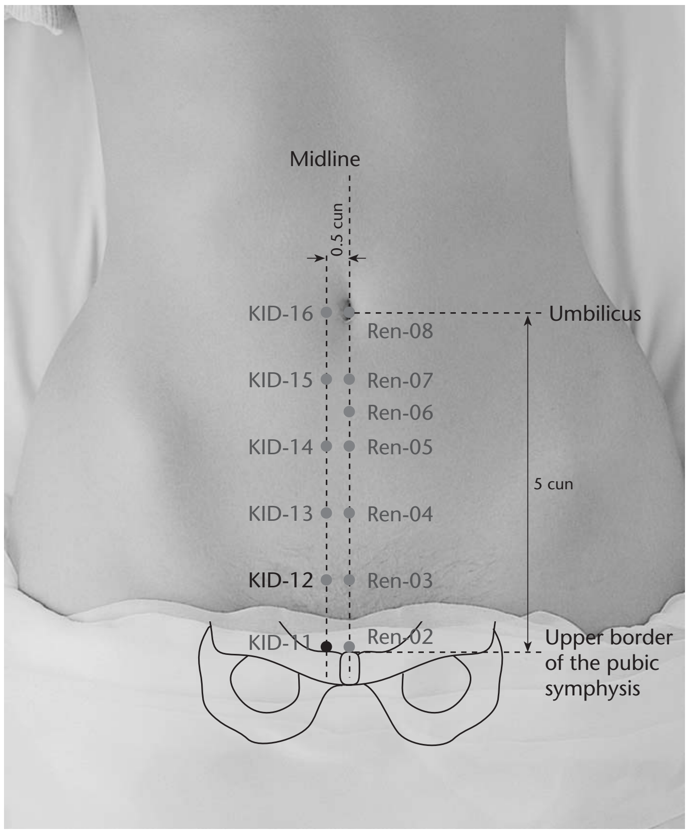
Midline
0.5 cun
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

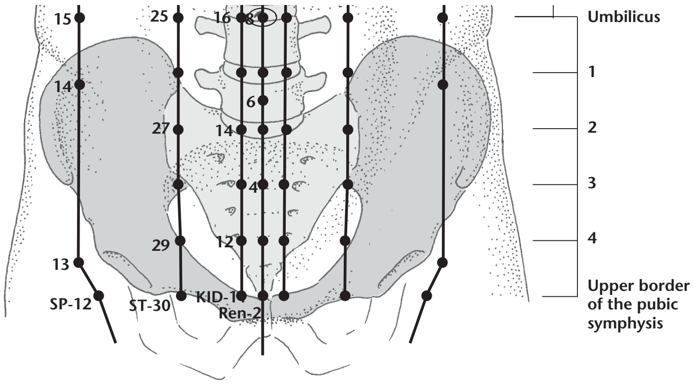
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

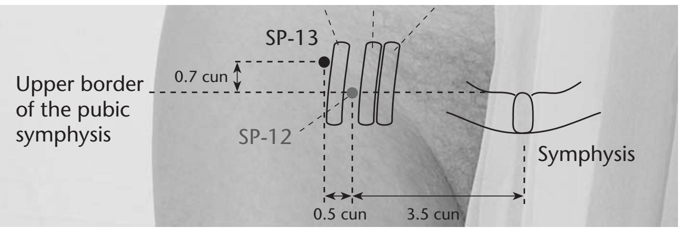
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

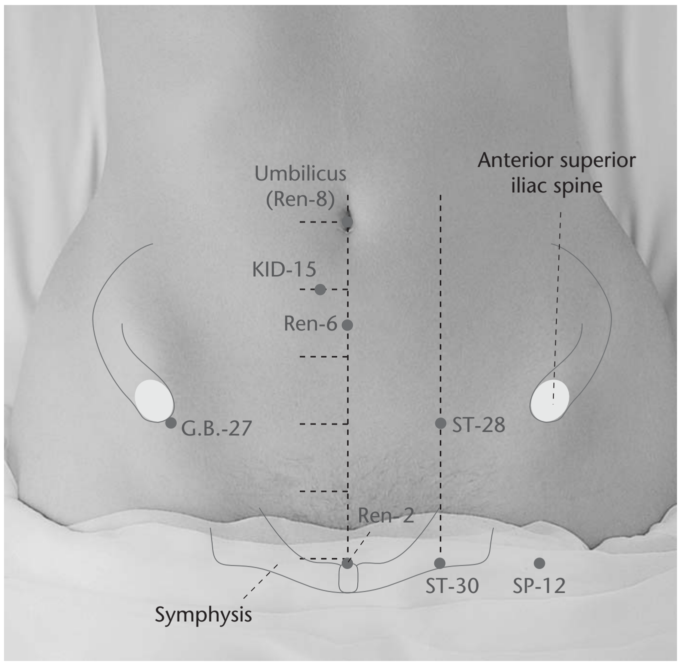
Umbilicus (Ren-8)
Anterior superior iliac spine
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
ST-30
SP-12
Symphysis

KID-12 *Great Luminance* DAHE

Location

1 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun in a superior direction and on this level locate **KID-12** 0.5 cun lateral to the midline. Located on the same level are → **Ren-3** (on the midline), → **ST-29** (2 cun lateral to the midline) and → **Ex-CA-1** (*zigong*; 3 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner, strengthens the Kidneys and the Essence (*jing*)

Special features

Meeting point with the *chong mai*

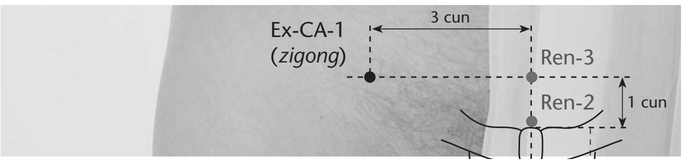
Ex-CA-1
(zigong)
3 cun
Ren-3
Ren-2
1 cun

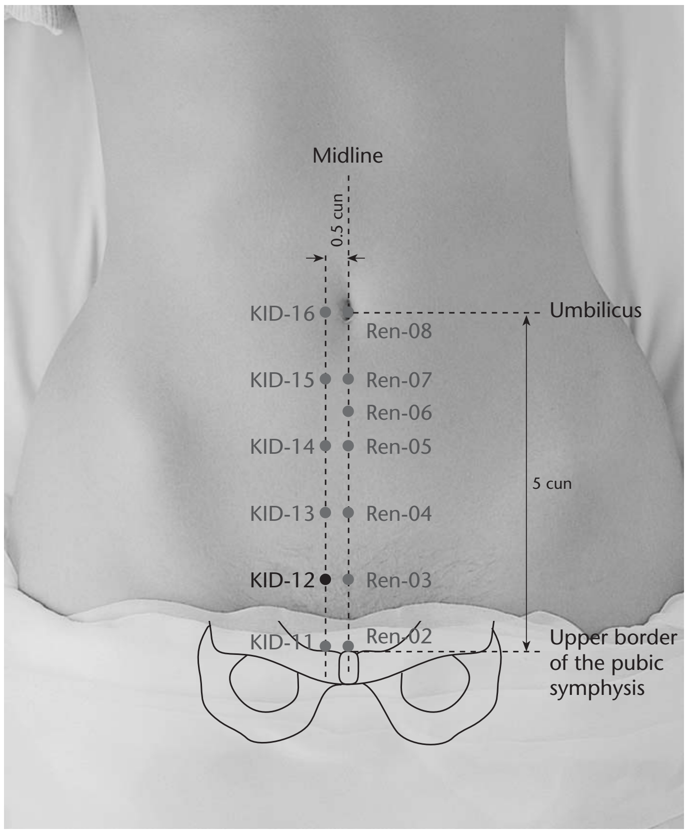
Midline
0.5 cun
KID-16
Ren-08
Umbilicus
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

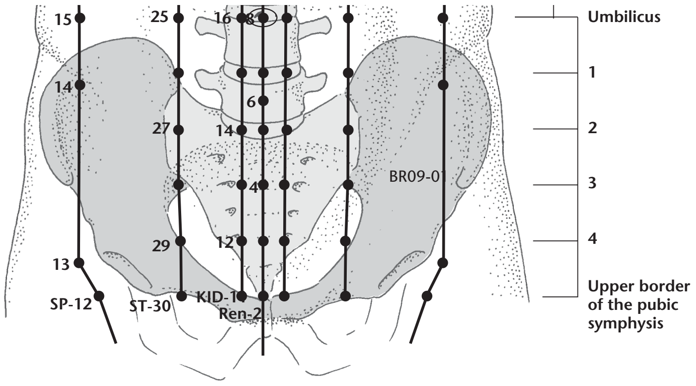
15
25
16
8
14
6
27
14
BR09-01
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

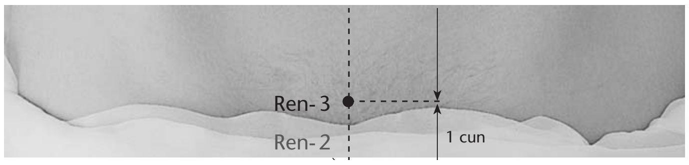
Ren-3
Ren-2
1 cun

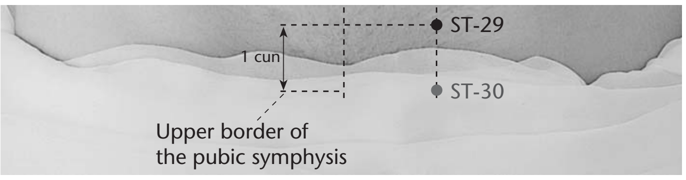
ST-29
1 cun
ST-30
Upper border of
the pubic symphysis

Qi Cave QIXUE KID-13

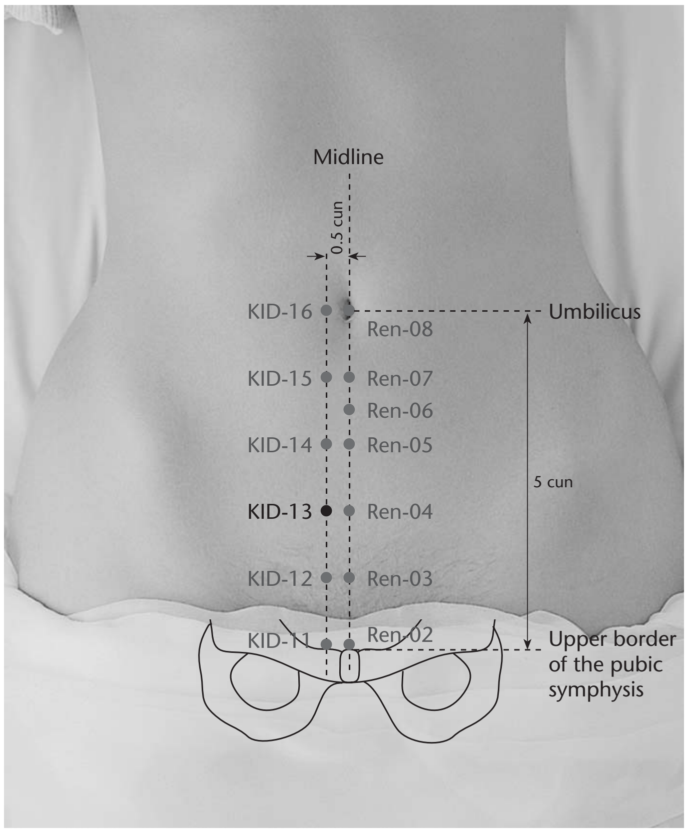
Midline
0.5 cun
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

Location

2 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a superior direction and on this level locate **KID-13** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-4** (on the midline), → **ST-28** (2 cun lateral to the midline), → **Ex-CA** (*tituo*/*qimen*/*yijing*: 4/3/1 cun lateral to the midline) and → **G.B.-27** (anterior and medial to the ASIS).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Regulates the Lower Burner, regulates the *chong mai* and *ren mai*, strengthens the Kidneys and the essence (*jing*)

Special features

Meeting point with the *chong mai*

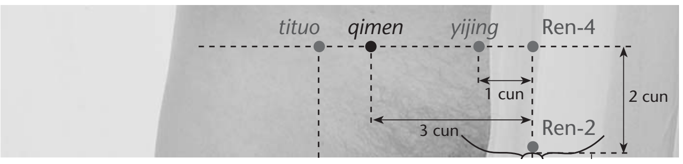
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

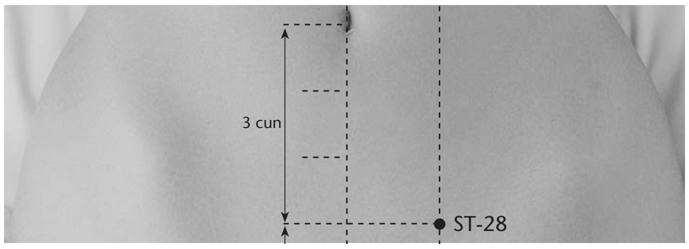

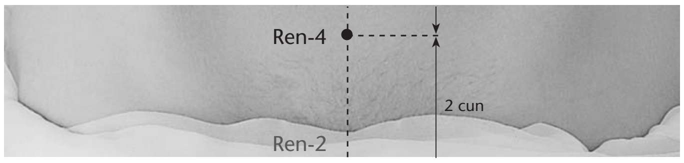
Ren-4
2 cun
Ren-2

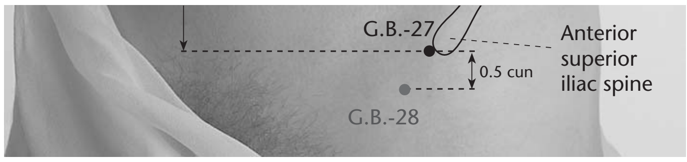
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

KID-14 *Four Fullnesses* SIMAN

Location

2 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape ➞ 2.3.1). From the centre of the umbilicus, measure 2 cun in an inferior direction and on this level locate **KID-14** 0.5 cun lateral to the midline.

Located on the same level are ➞ **Ren-5** (on the midline) and ➞ **ST-27** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Lower Burner and benefits the Kidneys
- Regulates the water passages
- Regulates the Qi and moves Blood stasis

Special features

Meeting point with the *chong mai*

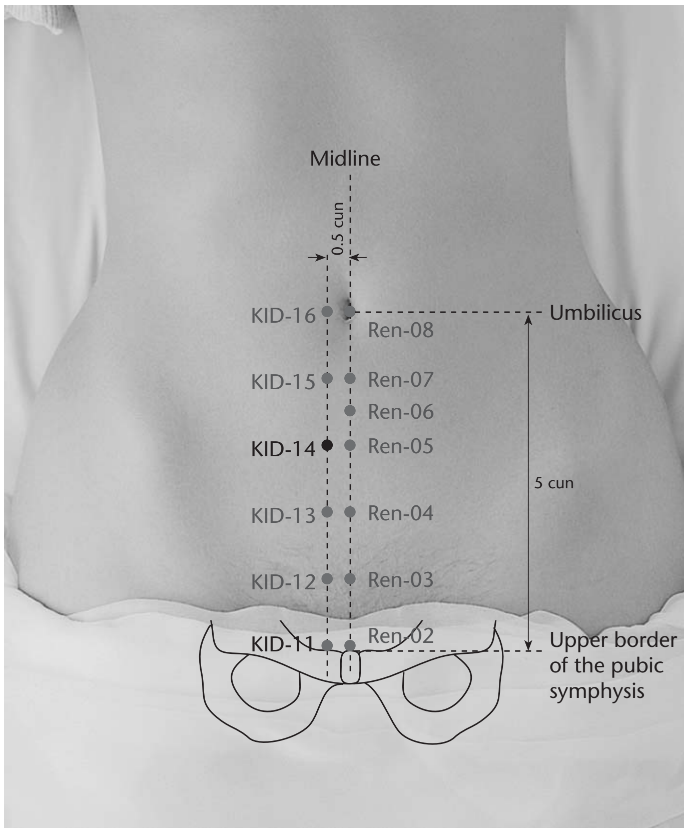
Midline
0.5 cun
KID-16
Ren-08
Umbilicus
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

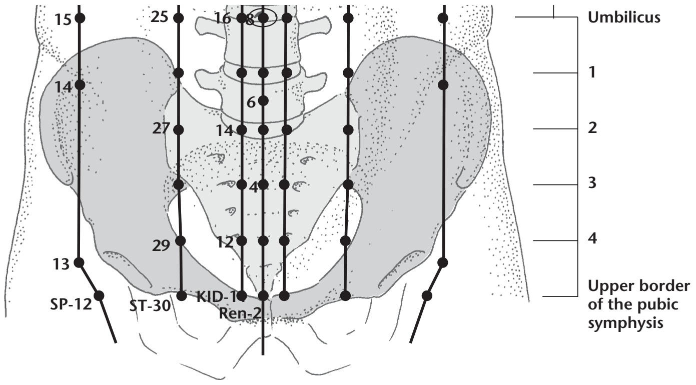
15
25
16
8
Umbilicus
1
14
6
27
14
2
3
4
29
12
4
13
SP-12
ST-30
KID-11
Ren-2
Upper border
of the pubic
symphysis

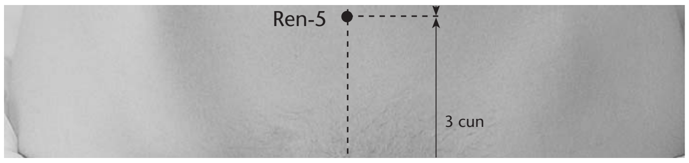

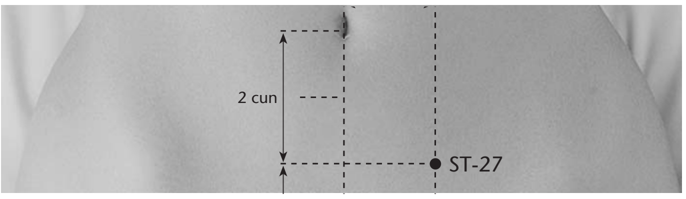

Middle Flow ZHONGZHU KID-15

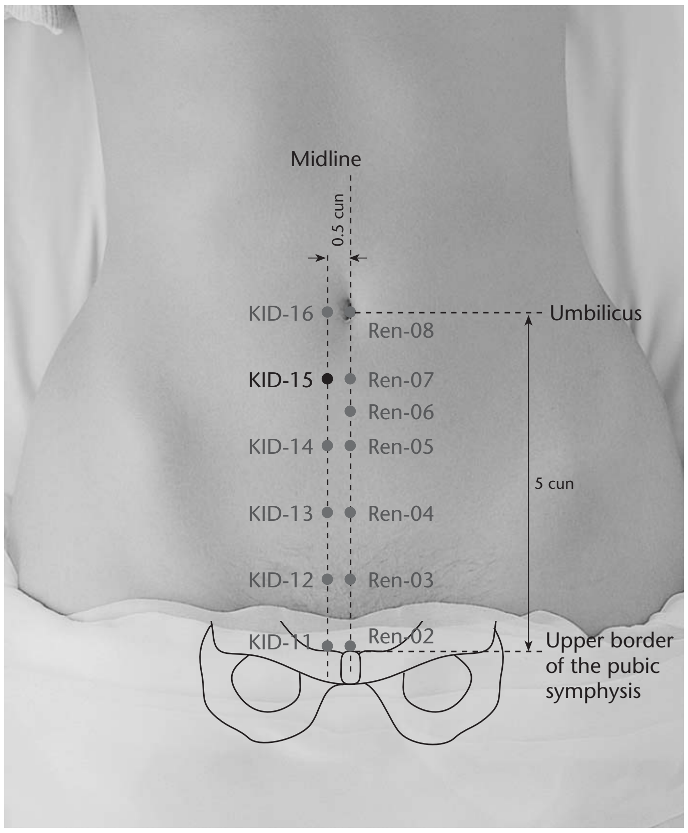
Midline
0.5 cun
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06
KID-14
Ren-05
5 cun
KID-13
Ren-04
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

Location

1 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun. As these can vary considerably from finger cun measurements, only proportional cun measurements should be applied (elastic tape → 2.3.1). From the centre of the umbilicus, measure 1 cun in an inferior direction and on this level locate **KID-15** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-7** (on the midline) and → **ST-26** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Intestines
- Regulates the Lower Burner

Special features

Meeting point with the *chong mai*

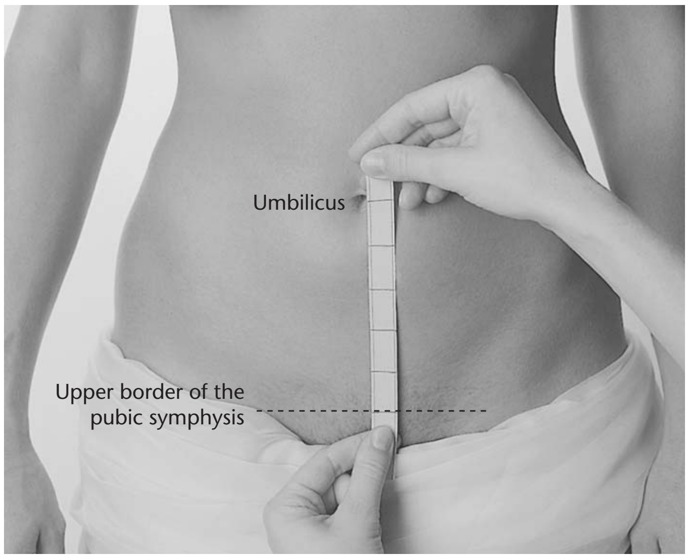
Umbilicus
Upper border of the pubic symphysis

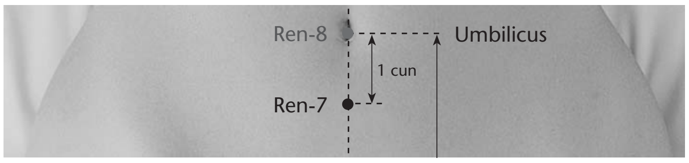
Ren-8
Umbilicus
1 cun
Ren-7

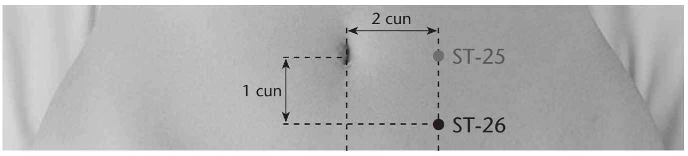
2 cun
ST-25
1 cun
ST-26

KID-16 *Vitals* Shu HUANGSHU

Location

0.5 cun lateral to the centre of the umbilicus.

How to find

From the centre of the umbilicus, measure 0.5 cun in a lateral direction. Located on the same level are → **Ren-8** (in the centre of the umbilicus), → **ST-25** (2 cun lateral to the midline), → **SP-15** (4 cun lateral to the midline) and → **G.B.-26** (on the level of umbilicus, on a vertical line drawn through the free end of the 11th rib).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy. Avoid the umbilicus.

Actions/Indications

- Regulates the Qi, regulates and warms the Stomach and Intestines

Special features

Meeting point with the *chong mai*

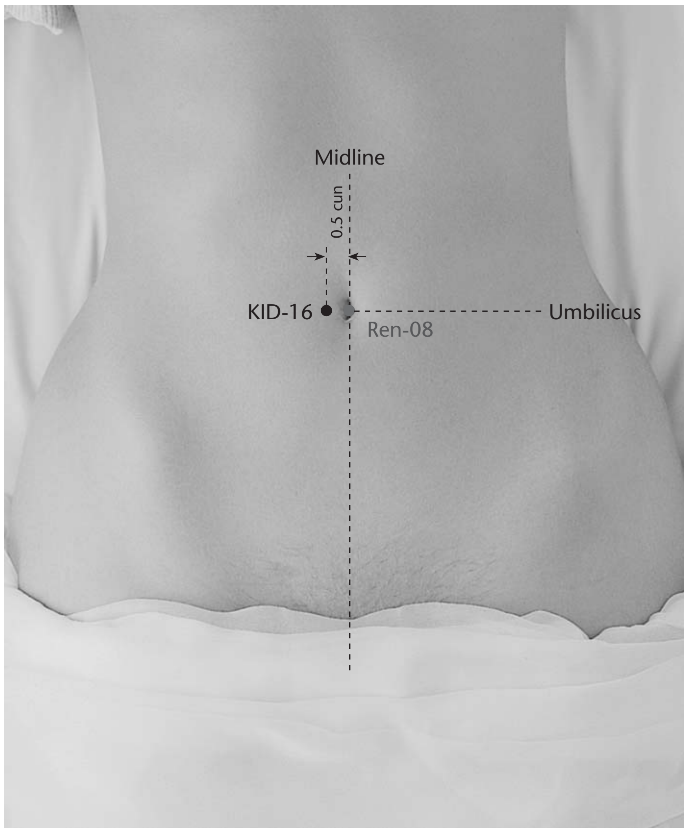
Midline
0.5 cun
KID-16
Umbilicus
Ren-08

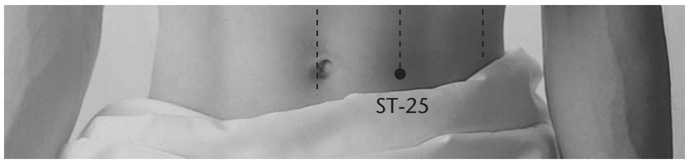

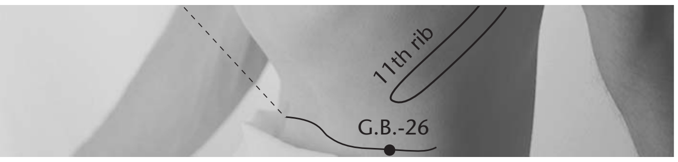
11th rib
G.B.-26

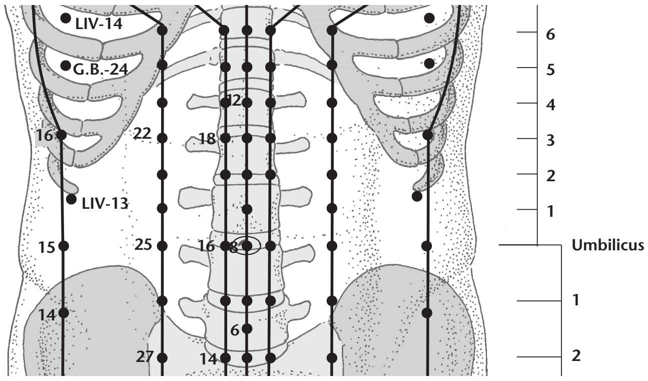
LIV-14
G.B.-24
16
22
12
18
LIV-13
15
25
16
8
14
27
6
14
6
5
4
3
2
1
Umbilicus
1
2

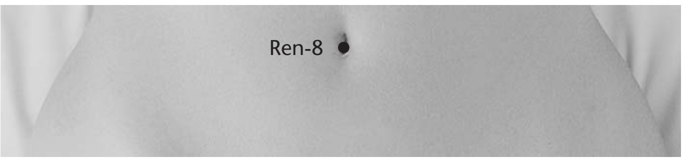

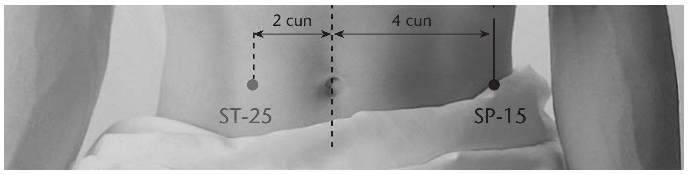
2 cun
4 cun
ST-25
SP-15

Shang Bend SHANGQU KID-17

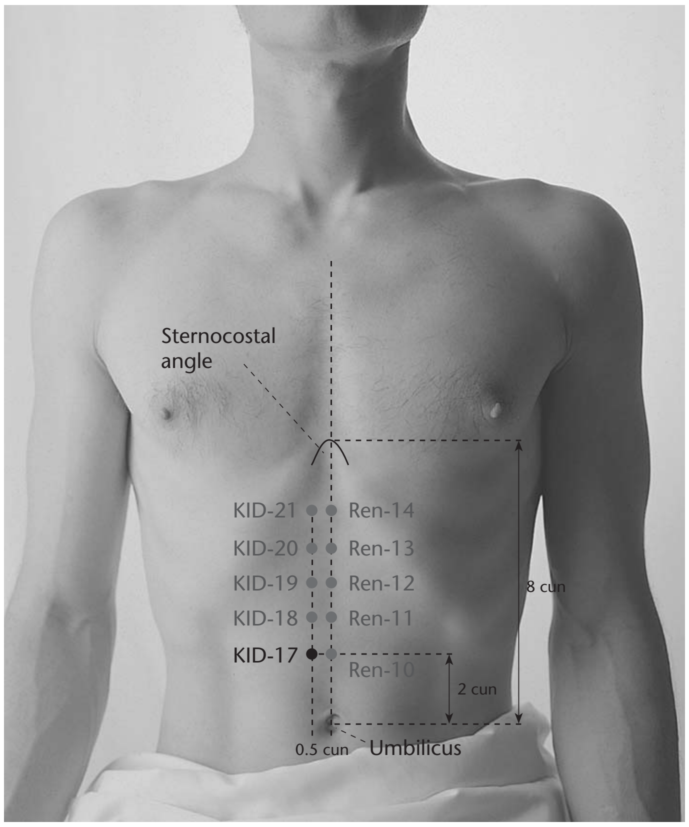
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

Location

2 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction and on this level locate **KID-17** 0.5 cun lateral to the midline.
Located on the same level are → **Ren-10** (on the midline), → **ST-23** (2 cun lateral to the midline) and **Ex-CA** (*weishang*, 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and Intestines, eliminates stagnation

Special features

Meeting point with the *chong mai*

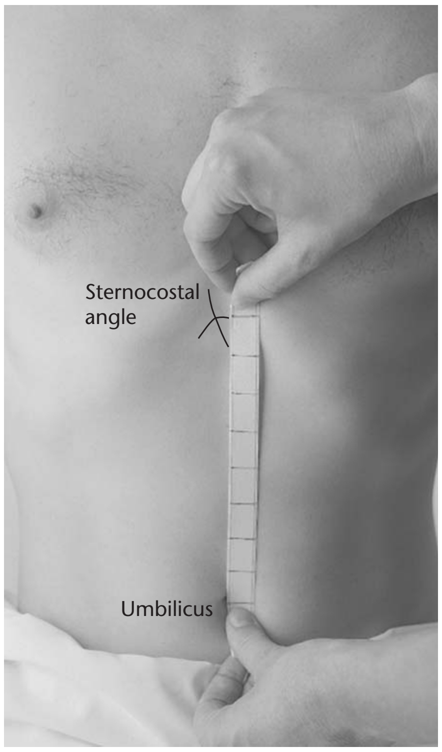
Sternocostal angle
Umbilicus

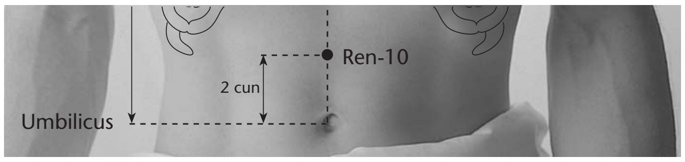
Ren-10
2 cun
Umbilicus

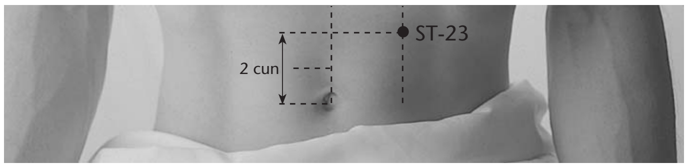

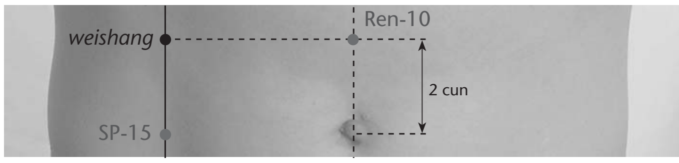
Ren-10
weishang
2 cun
SP-15

KID-18 *Stone Gate* **SHIGUAN**

Location

3 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (elastic tape → 2.3.1). From the centre of the umbilicus measure 3 cun in a superior direction and on this level locate **KID-18** 0.5 cun lateral to the midline.
Located on the same level are → **Ren-11** (on the midline), → **ST-22** (2 cun lateral to the midline) and → **SP-16** (on the mamillary line).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the Middle, regulates the Stomach and Intestines, removes Blood stagnation

Special features

Meeting point with the *chong mai*

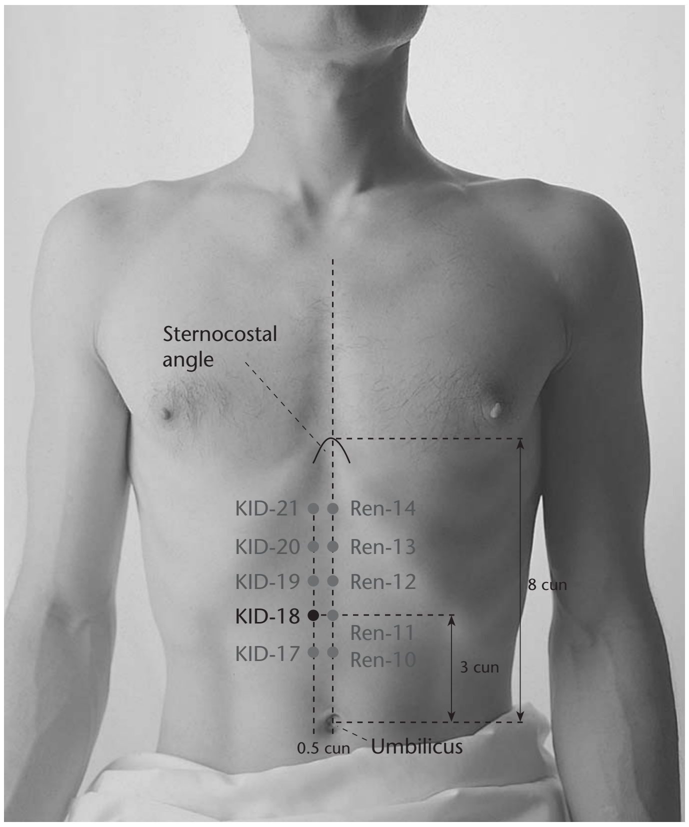
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
3 cun
0.5 cun
Umbilicus

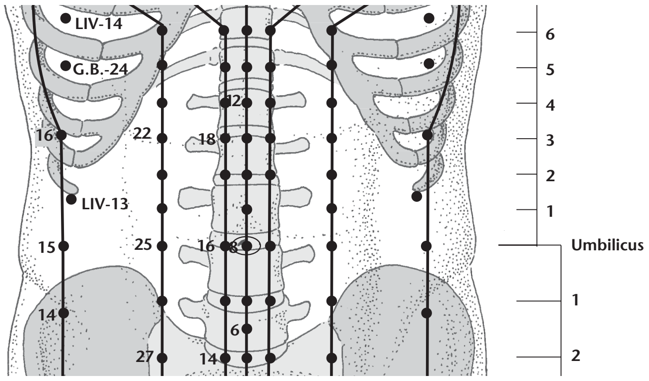
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
6
27
14
6
5
4
3
2
1
Umbilicus
1
2

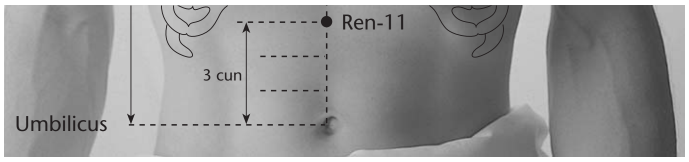
Ren-11
3 cun
Umbilicus

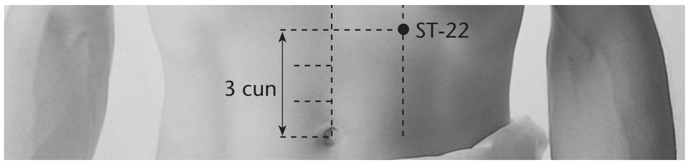

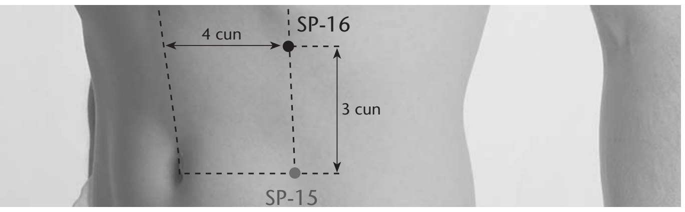
4 cun
SP-16
3 cun
SP-15

Yin Metropolis YINDU **KID-19**

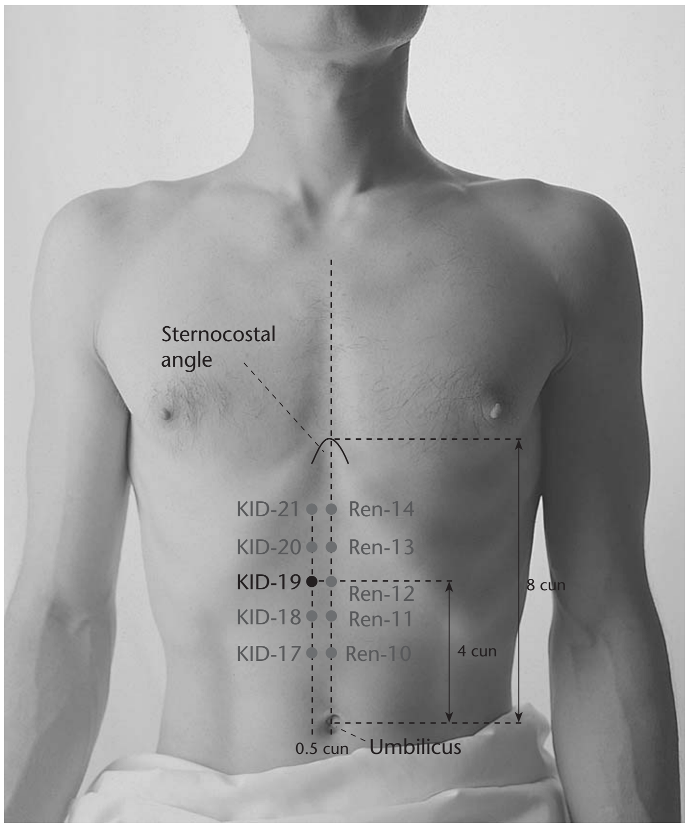
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
4 cun
0.5 cun
Umbilicus

Location

Midway between the sternocostal angle and the centre of the umbilicus, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). Spreading hands technique (→ 2.3.3): Place the little fingers on the centre of the umbilicus and the sternocostal angle respectively and join the thumbs at the midpoint of this distance by evenly spreading both hands. **KID-19** is located on this level, 0.5 cun lateral to the midline.

Located on the same level are → **Ren-12** (on the midline) and → **ST-21** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Qi, regulates the Stomach and Intestines
- Regulates counterflow Qi

Special features

Meeting point with the *chong mai*

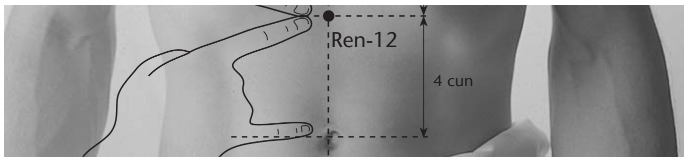

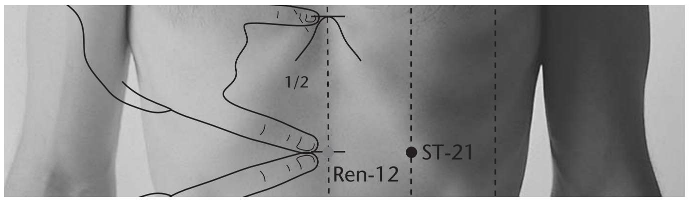
1/2
ST-21
Ren-12

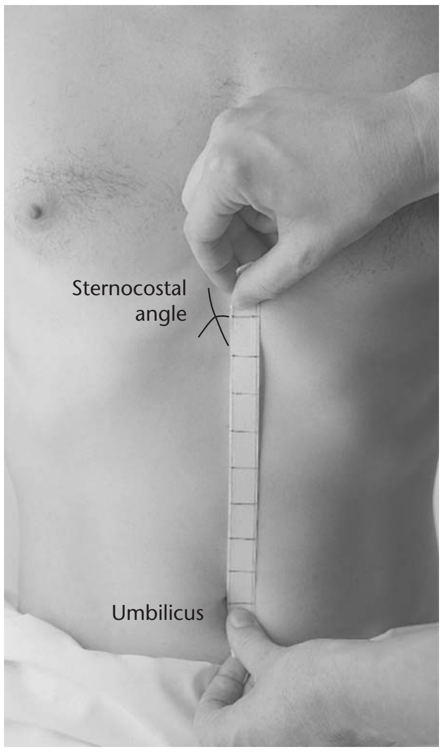
Sternocostal angle
Umbilicus

KID-20 *Abdomen Connecting Valley* FUTONGGU

Location

5 cun superior to the umbilicus or 3 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (➞ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (elastic tape ➞ 2.3.1). From the sternocostal angle, measure 3 cun in an inferior direction and on this level locate **KID-20** 0.5 cun lateral to the midline.

Located on the same level are ➞ **Ren-13** (on the midline), ➞ **ST-20** (2 cun lateral to the midline) and ➞ **G.B.-24** (on the mamillary line, in the 7th intercostal space).

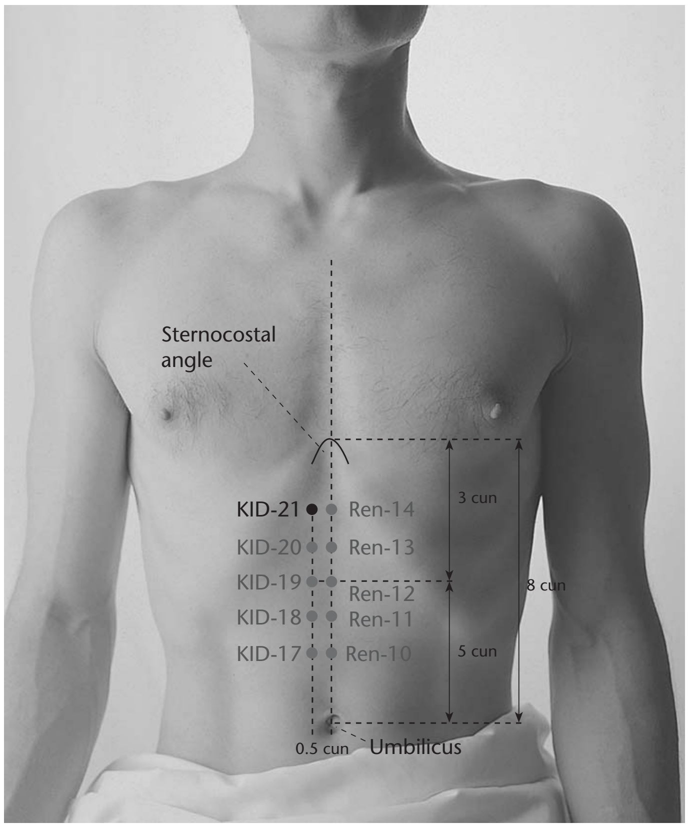
Sternocostal angle
KID-21
Ren-14
KID-20
Ren-13
KID-19
Ren-12
KID-18
Ren-11
KID-17
Ren-10
3 cun
8 cun
5 cun
0.5 cun
Umbilicus

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the Spleen, harmonises the Stomach and Intestines, regulates counterflow Qi

Special features

Meeting point with the *chong mai*

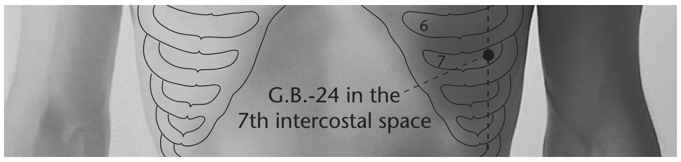
6
7
G.B.-24 in the
7th intercostal space

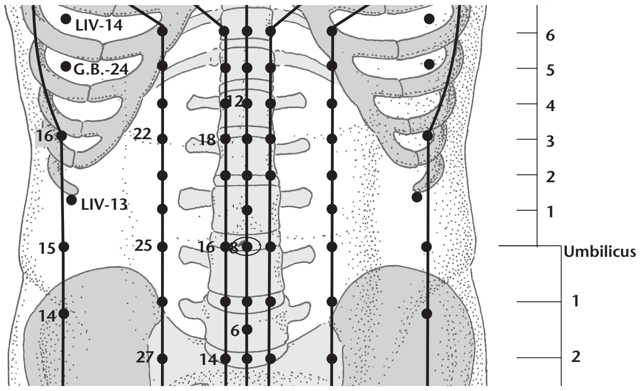
LIV-14
G.B.-24
21
16
22
18
LIV-13
15
25
16
8
14
6
27
14
6
5
4
3
2
1
Umbilicus
1
2

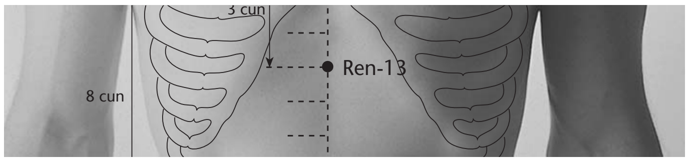
3 cun
Ren-13
8 cun

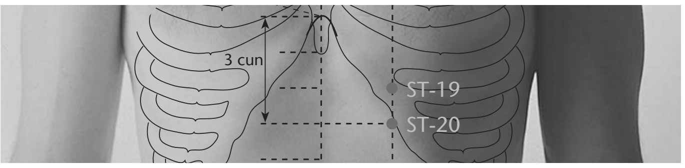
3 cun
ST-19
ST-20

Hidden Gate YOUMEN KID-21

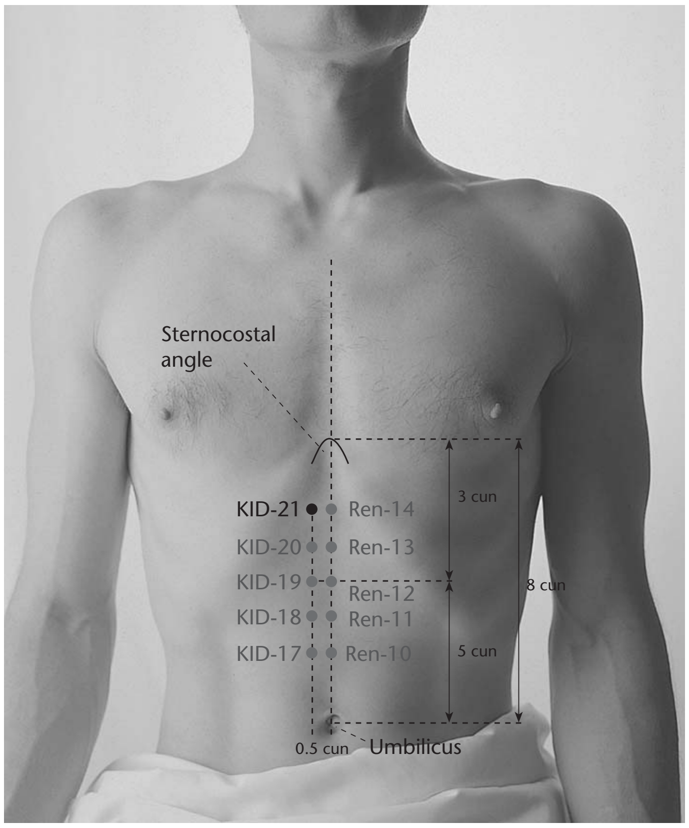
Sternocostal angle
KID-21
KID-20
KID-19
KID-18
KID-17
Ren-14
Ren-13
Ren-12
Ren-11
Ren-10
3 cun
8 cun
5 cun
0.5 cun
Umbilicus

Location

2 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun should be used (elastic tape → 2.3.1). From the sternocostal angle, measure 2 cun in an inferior direction and on this level locate **KID-21** 0.5 cun lateral to the midline.

Located on the same level are → **Ren-14** (on the midline), → **ST-19** (2 cun lateral to the midline) and → **LIV-14** (on the mamillary line, in the 6th intercostal space).

Needling

Vertically 0.5–1 cun. Caution: The liver is located on the right, the peritoneum on the left. Caution during pregnancy.

Actions/Indications

- Harmonises the Stomach, regulates counterflow Qi

Special features

Meeting point with the *chong mai*. Last point of the *chong mai* located on the Kidney channel.

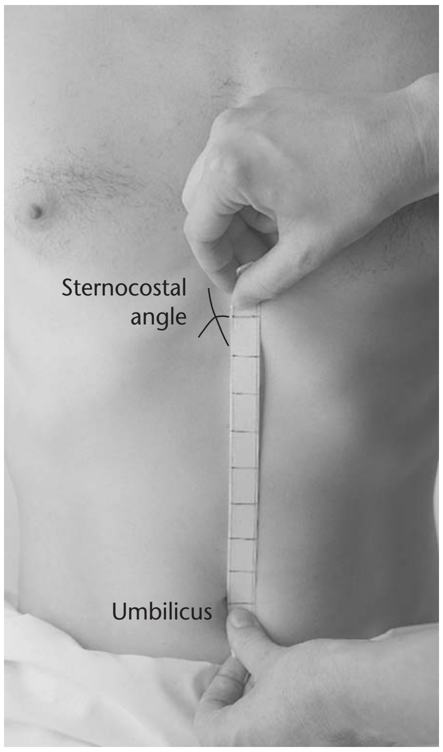
Sternocostal angle
Umbilicus

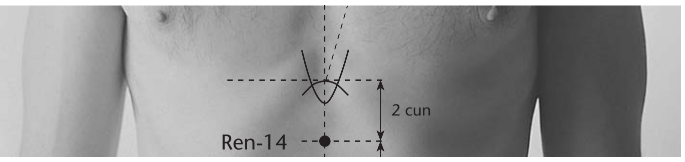

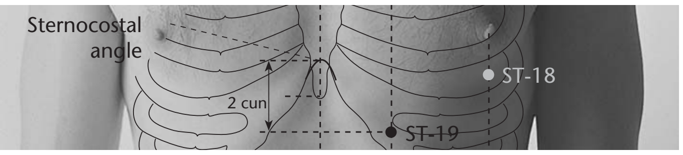
Sternocostal angle
2 cun
ST-18
ST-19

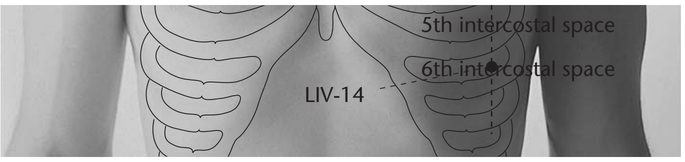
5th intercostal space
6th intercostal space
LIV-14

KID-22 *Walking Corridor* BULANG

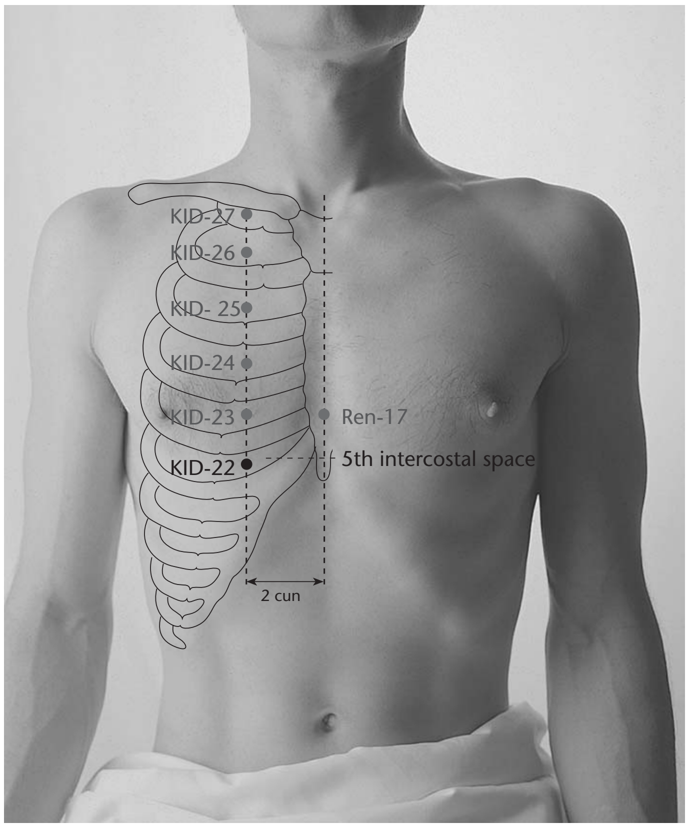
KID-27
KID-26
KID- 25
KID-24
KID-23
KID-22
Ren-17
5th intercostal space
2 cun

Location

In the 5th intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 5th ICS either by counting parasternally from the clavicle or from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-22** 2 cun lateral to the anterior midline. Located on the same level (5th ICS) are → **Ren-16** (on the midline), → **ST-18** (on the mamillary line) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) with or against the channel pathway. Caution: Avoid injuring the liver (on the right side), the heart (on the left side) and the pleura.

Actions/Indications

- Regulates the Lung Qi and counterflow Qi, unbinds the chest

Special features

Exit point. The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

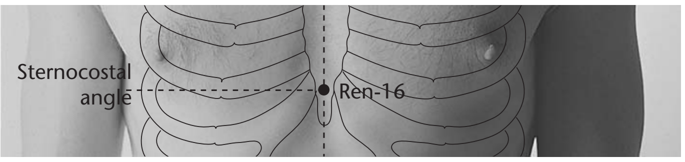
Sternocostal angle
Ren-16

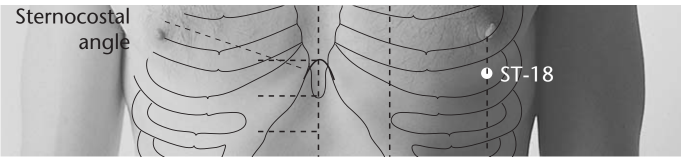
Sternocostal angle
ST-18

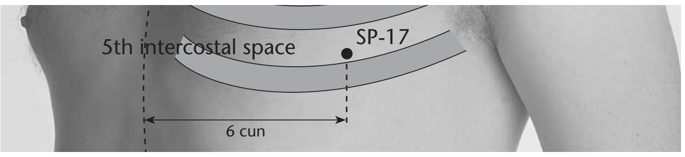
5th intercostal space
SP-17
6 cun

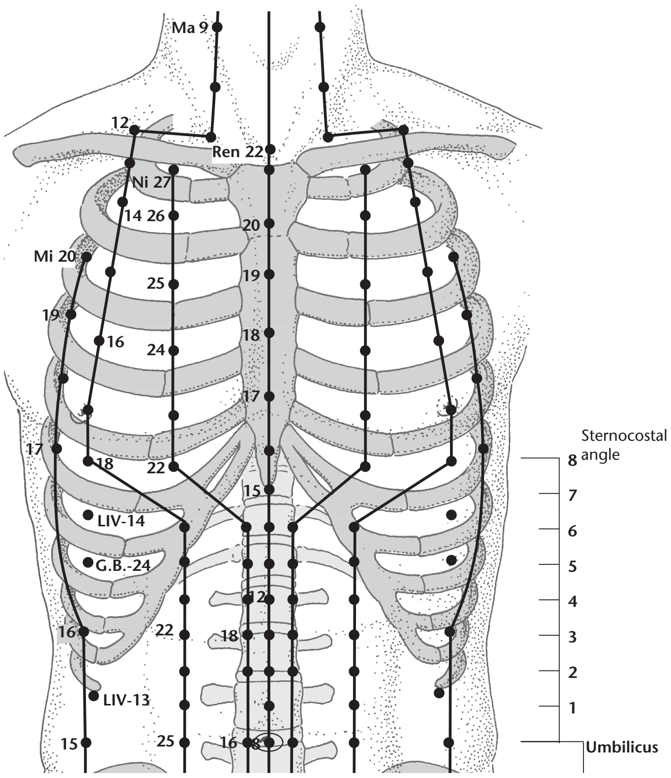
Ma 9
12
Ren 22
Ni 27
14 26
20
Mi 20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus

Spirit Seal SHENFENG KID-23

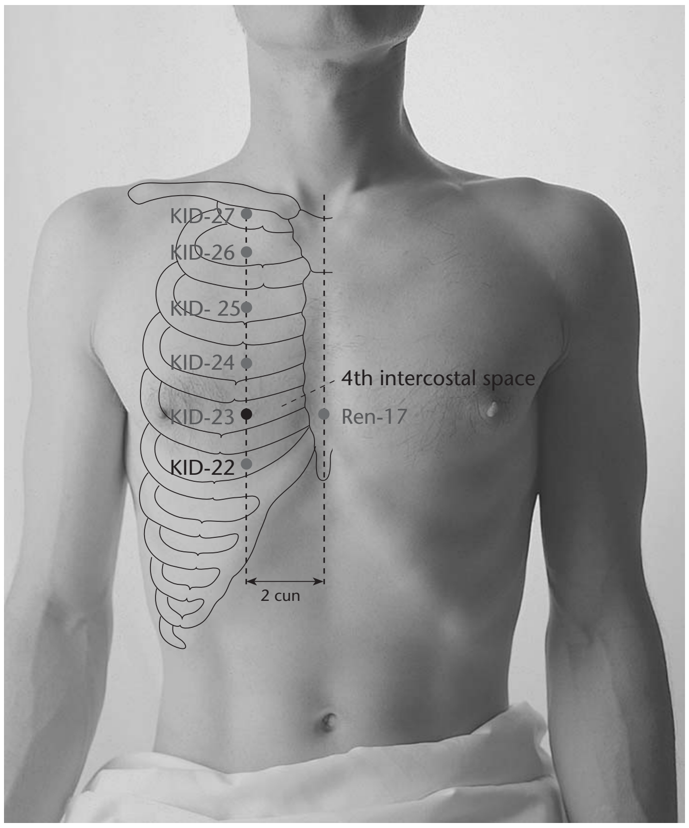
KID-27
KID-26
KID- 25
KID-24
KID-23
KID-22
4th intercostal space
Ren-17
2 cun

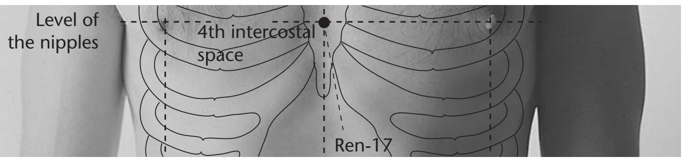
Level of the nipples
4th intercostal space
Ren-17

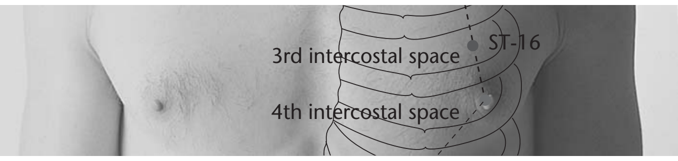
3rd intercostal space
ST-16
4th intercostal space

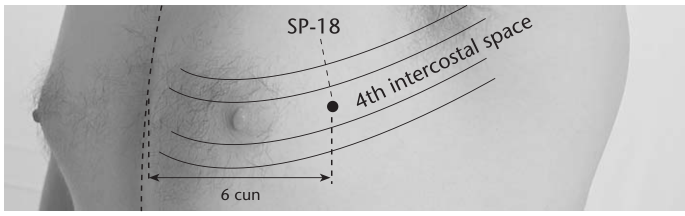
SP-18
4th intercostal space
6 cun

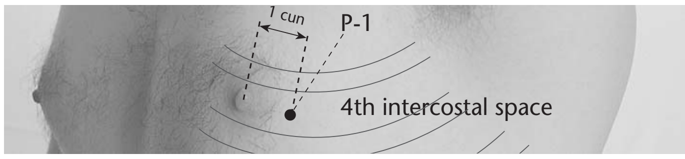
1 cun
P-1
4th intercostal space

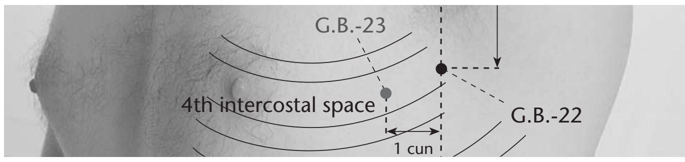
G.B.-23
4th intercostal space
G.B.-22
1 cun

Location

In the 4th intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 4th ICS either by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-23** 2 cun lateral to the anterior midline.

Located on the same level (4th ICS) are → **Ren-17** (on the midline), → **ST-17** (on the nipple), → **SP-18** (6 cun lateral to the midline), → **P-1** (1 cun lateral to the nipple) and → **G.B.-22** (3 cun inferior to the axilla) as well as → **G.B.-23** (1 cun anterior to **G.B.-22**).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Heart, pneumothorax.

Actions/Indications

- Regulates the Lung Qi and counterflow Qi, unbinds the chest
- Harmonises the Stomach
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

KID-24 *Spirit Ruin* LINGXU

Location

In the 3rd intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 3rd ICS by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-24** 2 cun lateral to the anterior midline. Located on the same level (→ 3rd ICS) are → **Ren-18** (on the midline), → **ST-16** (4 cun lateral to the midline) and → **SP-19** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung and Stomach Qi
- Unbinds the chest
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

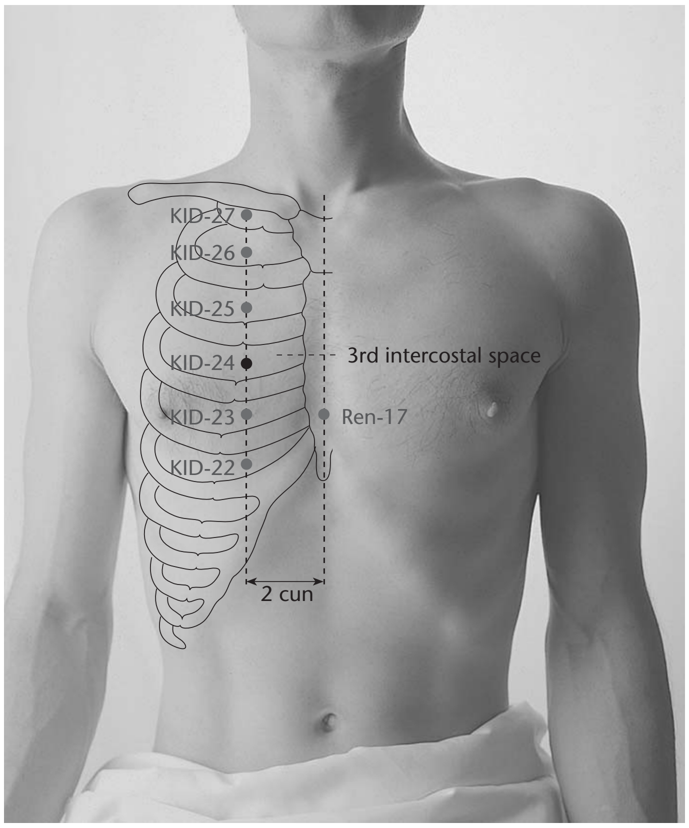
KID-27
KID-26
KID-25
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22
2 cun

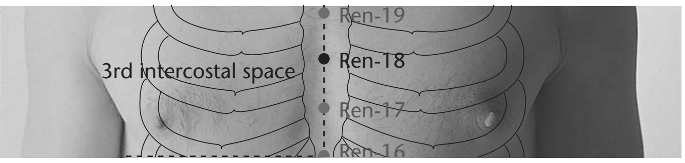
Ren-19
3rd intercostal space
Ren-18
Ren-17
Ren-16

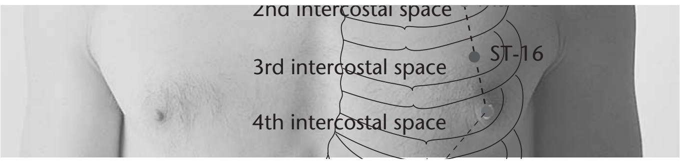
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

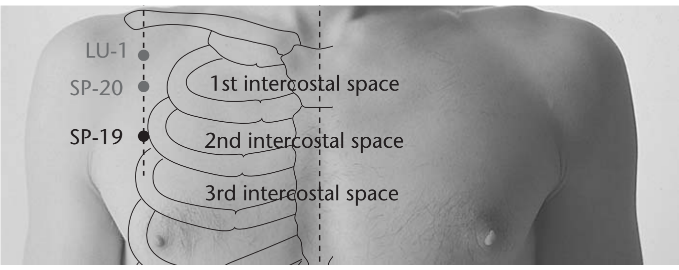
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

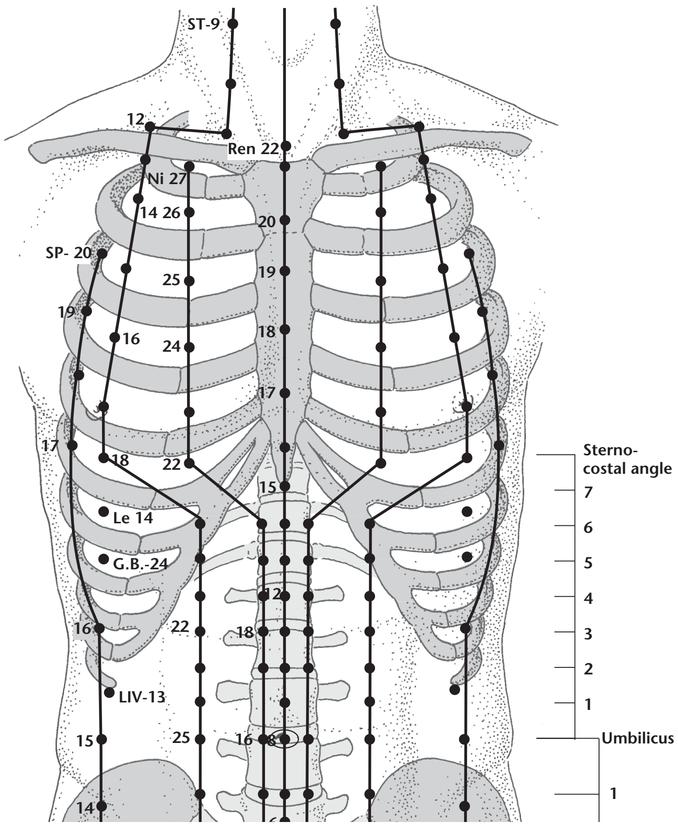
ST-9
12
Ren 22
Ni 27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
Sterno-costal angle
7
15
Le 14
6
G.B.-24
5
12
4
16
22
18
3
2
LIV-13
1
15
25
16
8
Umbilicus
14
1

Ch04.8-F10028.qxd 2/22/08 6:14 PM Page 345

Spirit Storehouse SHENCANG KID-25

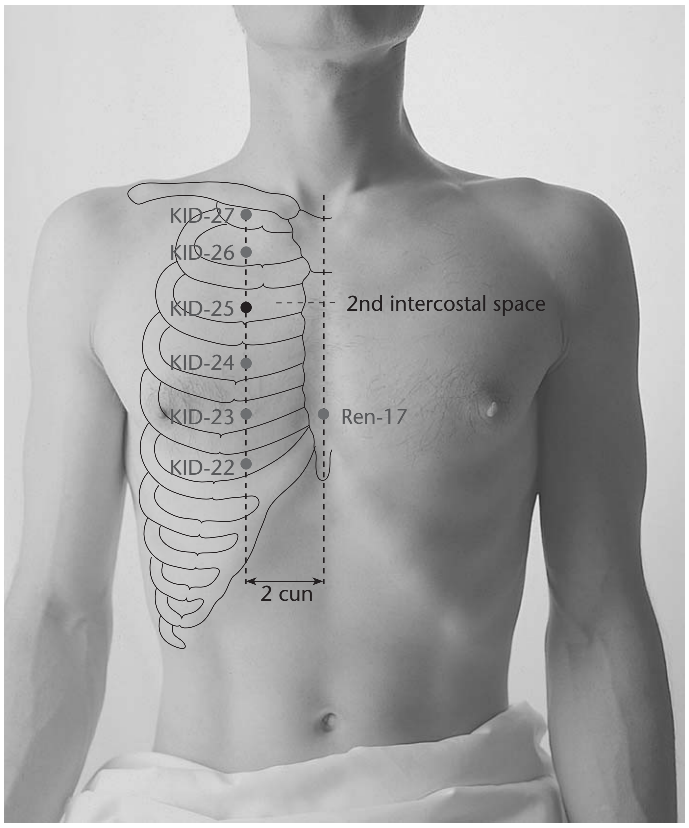
KID-27
KID-26
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17
KID-22
2 cun

Location

In the 2nd intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Locate the 2nd ICS by palpating parasternally from the clavicle or by counting from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5), then locate **KID-24** 2 cun lateral to the anterior midline. Located on the same level (→ 2nd ICS) are → **Ren-19** (on the midline), → **ST-15** (4 cun lateral to the midline) and → **SP-20** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi
- Unbinds the chest

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

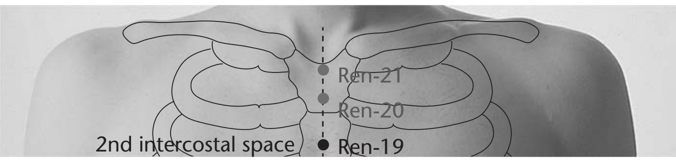
Ren-21
Ren-20
2nd intercostal space
Ren-19

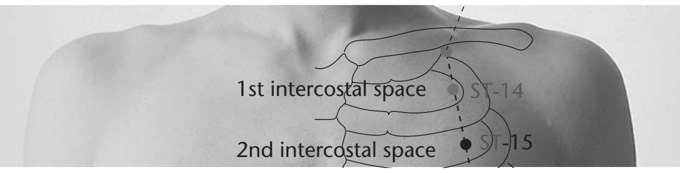
1st intercostal space
ST-14
2nd intercostal space
ST-15

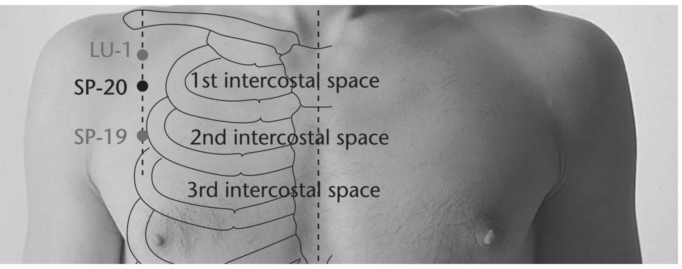
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

KID-26 *Comfortable Chest* YUZHONG

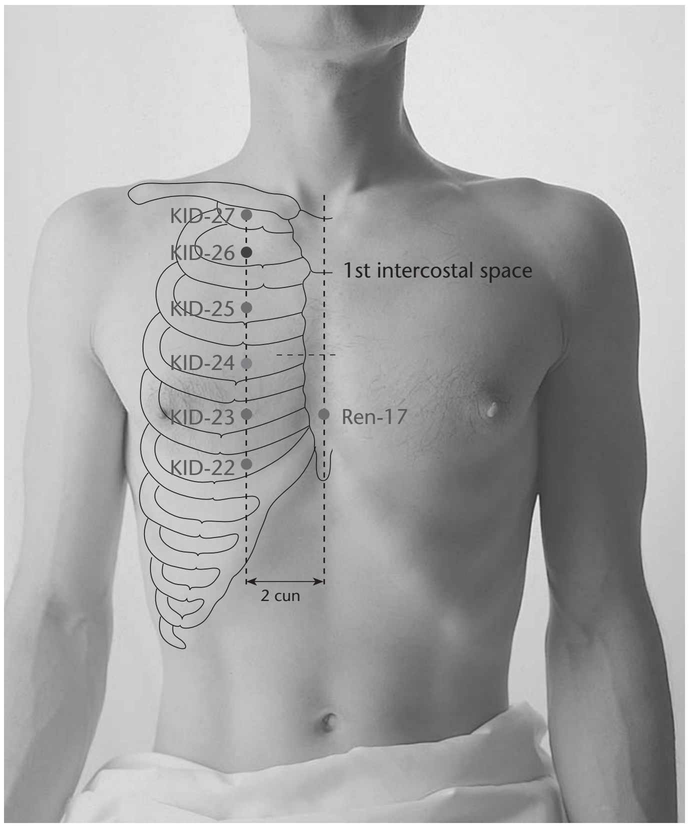
KID-27
KID-26
KID-25
KID-24
KID-23
KID-22
1st intercostal space
Ren-17
2 cun

Location

In the 1st intercostal space (ICS), 2 cun lateral to the anterior midline.

How to find

Palpating parasternally from superior to inferior, the 1st rib can be palpated directly below the clavicle; in some cases, it can be completely hidden under the clavicle. Locate **KID-26** below the 1st rib in the 1st ICS, 2 cun lateral to the midline. Located on the same level (1st ICS) are → **Ren-20** (on the midline), → **ST-14** (4 cun lateral to the midline) and → **LU-1** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi and transforms Phlegm
- Unbinds the chest
- Benefits the breasts

Special features

The *Su Wen* describes points **KID-22** to **KID-27** as *shu* points of the chest, indicated for difficult breathing and disorders of the thorax. They are particularly indicated for breathing difficulties due to 'excess above, deficiency below', which refers to the inability of the Kidneys to absorb the Qi that has descended from the Lung.

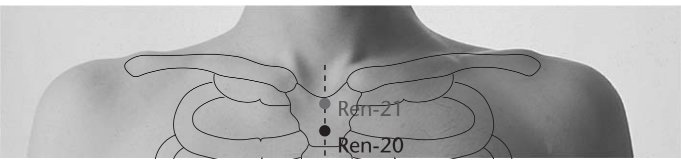
Ren-21
Ren-20

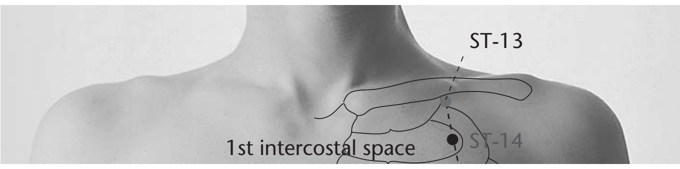
ST-13
1st intercostal space
ST-14

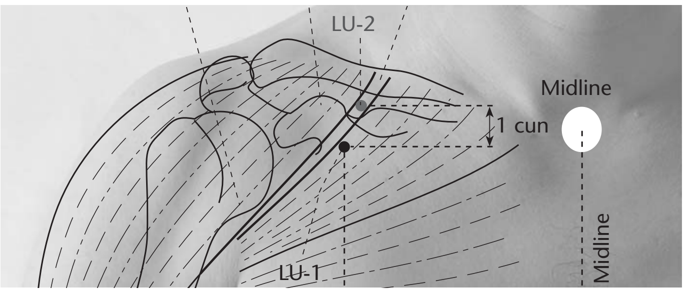
LU-2
Midline
1 cun
LU-1
Midline

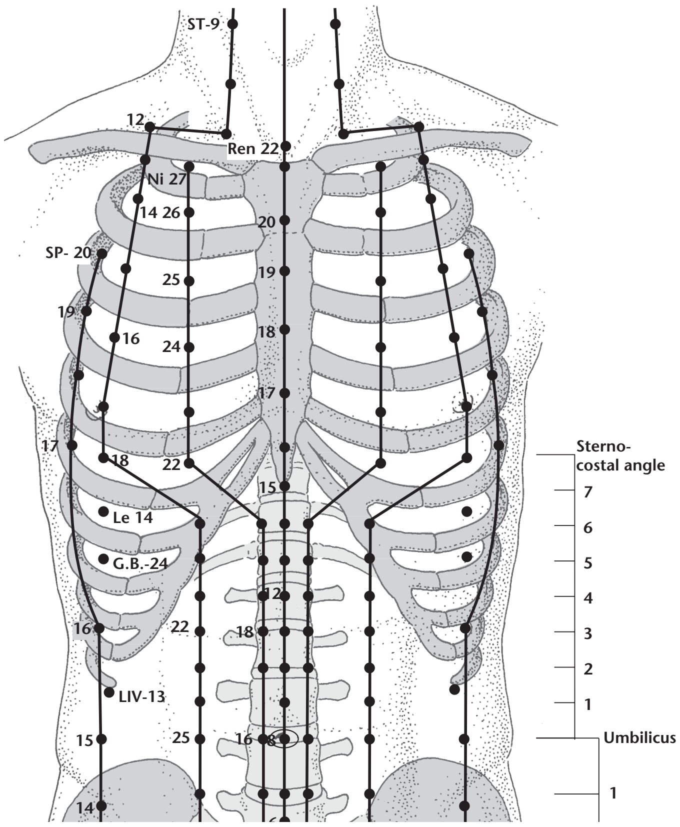
ST-9
12
Ren 22
Ni 27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
15
Le 14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
Sterno-costal angle
7
6
5
4
3
2
1
Umbilicus
1

Shu *Mansion* SHUFU KID-27

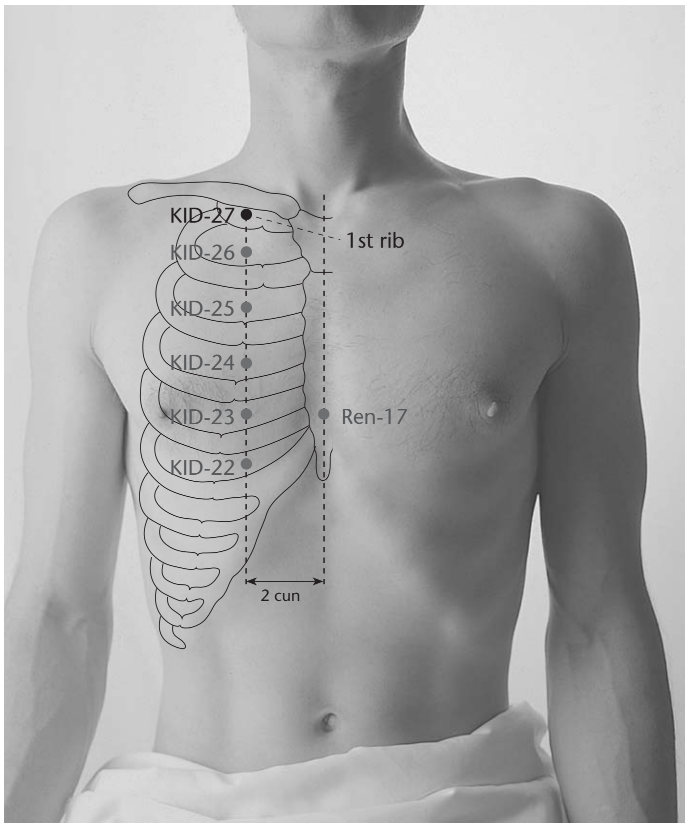
KID-27
KID-26
KID-25
KID-24
KID-23
KID-22
1st rib
Ren-17
2 cun

Location

At the lower border of the clavicle, 2 cun lateral to the anterior midline

How to find

When palpating parasternally from superior to inferior, the 1st rib can be felt directly below the clavicle; in some cases, it can be completely hidden under the clavicle. Therefore, **KID-27** will be located either on the 1st rib or, in some cases, at its lower border. → **ST-13** is located lateral to the short curvature of the 1st rib, below the midpoint of the clavicle.

Located on approximately the same level are → **Ren-21** (on the midline) and → **LU-2** (6 cun lateral to the midline).

Needling

0.5–0.8 cun obliquely or transversely (subcutaneously) along the course of the ICS or with or against the channel pathway. Caution: Pneumothorax.

Actions/Indications

- Regulates the Lung Qi and transforms Phlegm
- Unbinds the chest

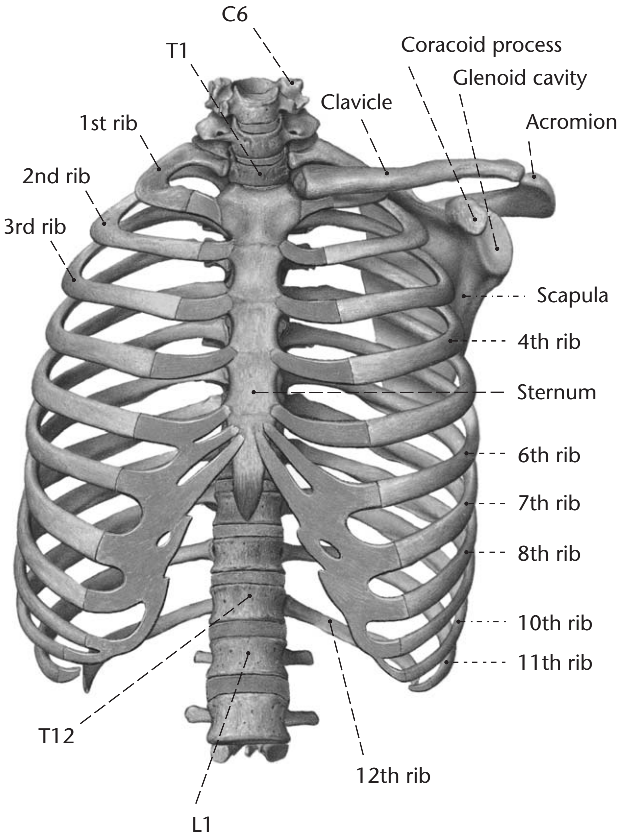
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

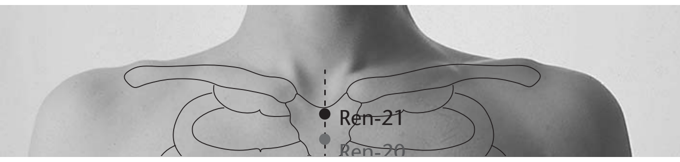
Ren-21
Ren-20

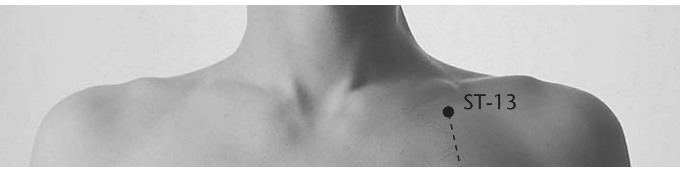

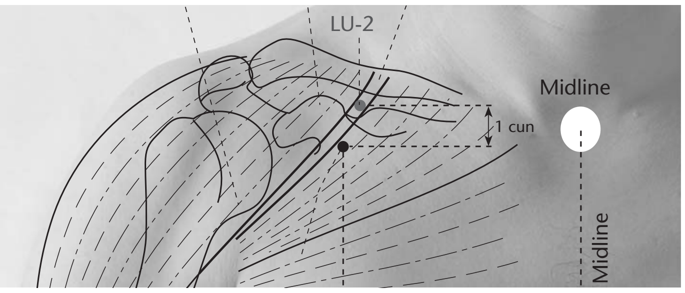
LU-2
Midline
1 cun
Midline

4.9 The Pericardium Channel System – Hand *jueyin* (*shou jue yin jing luo*)

4.9.1 Primary Pericardium Channel (*shou jue yin jing*)

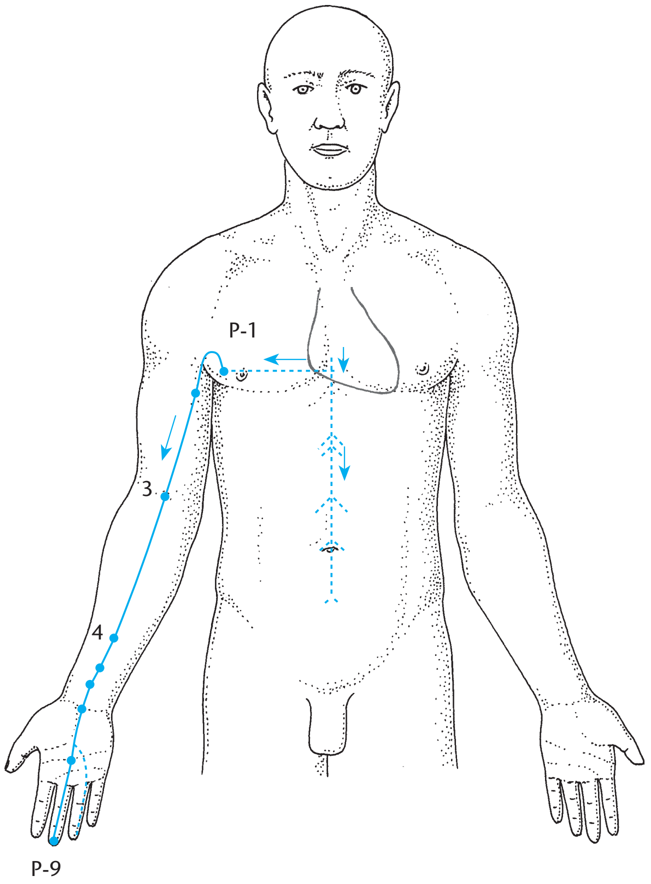

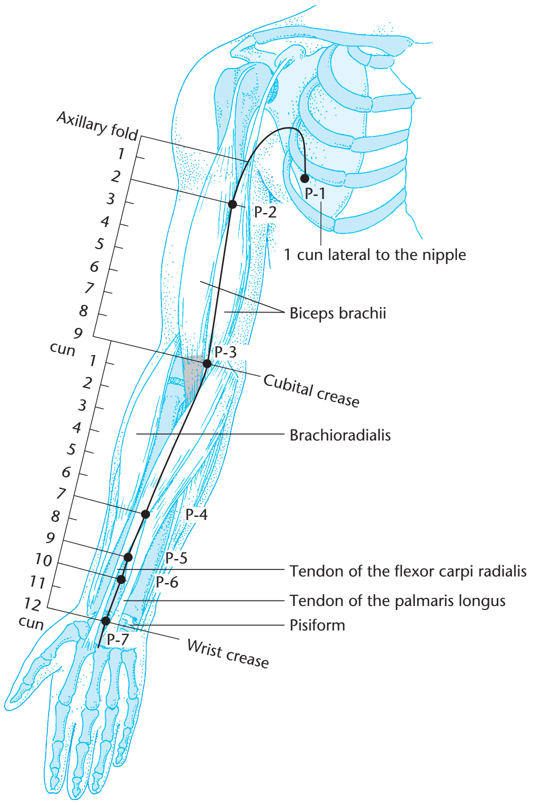
Axillary fold
1
2
3
4
5
6
7
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun
P-1
P-2
1 cun lateral to the nipple
Biceps brachii
P-3
Cubital crease
Brachioradialis
P-4
P-5
P-6
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Pisiform
P-7
Wrist crease

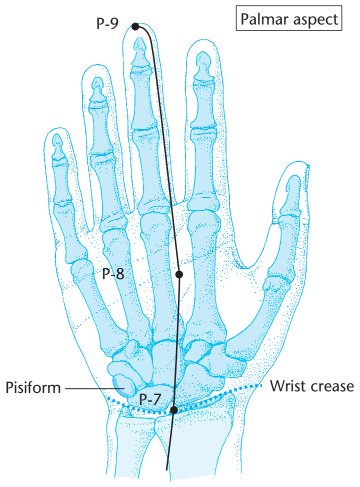
P-9
Palmar aspect
P-8
Pisiform
P-7
Wrist crease

Pathway

The Pericardium primary channel begins in the centre of the thorax on the level of → **Ren-17** (*shanzhong*), where an internal branch of the Kidney primary channel ends, having also connected with the Liver, Lung and Heart (deep Yin–Yin connection).

The Pericardium primary channel then divides into two branches:

- ➡ The **first branch** penetrates the diaphragm, descends to the abdomen and connects with the Upper, Middle and Lower Burner.
- ➡ A **further internal branch** also originates in the thoracic region, continues to the breast and emerges approximately 1 cun lateral to the nipple in the 4th intercostal space at **P-1** (*tianchi*). Beneath **P-1**, an internal branch of the Liver primary channel emerges from the Liver and joins the Pericardium primary channel (Yin axes of the third great circuit: *jueyin*).

The **external** pathway ascends to the axilla, follows the medial aspect of the upper arm between the Lung and Heart primary channels, passes the elbow and continues along the forearm between the tendons of the palmaris longus and flexor carpi ulnaris muscles, ending at the tip of the middle finger.

Another **internal** branch originates in the centre of the palm at **P-8** (*laogong*), runs along the ring finger and connects at its tip with the Triple Burner primary channel (hand Yin–Yang connection of the third circuit).

Clinical importance (→ 1.2)

Exterior *(biao)* signs and symptoms: Occipital stiffness, leg cramps, red complexion, eye pain, swellings in the subaxillary region, hypertonic musculature with limited range of motion of the arm and elbow, Heat sensations in the palms

Interior (*li*) or *zangfu*-Organ signs and symptoms: Delirious speech, confusion, irritability, fullness and distension in the thorax and lateral costal region, aphasia, palpitations

Connections of the Pericardium primary channel (→ 1.2)

Connections with other channels

Triple Burner primary channel (*shou shao yang jing*)

Connection: Hand Yin–Yang pairing of the third great circuit
Location: P-8 → T.B.-1 (on the hand)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Liver primary channel (*zu jue yin jing*)

Connection: Six-channel pairing of the third great circuit (Yin axes): *jue yin*
Location: LIV → P. An internal branch of the Liver primary channel emerges at the Liver, penetrates the diaphragm and connects with the Pericardium primary channel beneath **P-1** (*tianchi*).
Circulation: Non-circadian (**not** according to the Organ clock)
Importance: Above–below relationship

Kidney primary channel (*zu shao yin jing*)

Connection: Deep Yin–Yin connection
Location: KID → P (on the thorax). A branch of the internal pathway of the Kidney primary channel travels from the Kidney to the Liver, penetrates the diaphragm and spreads in the Lung (*fei*). From the Lung, an internal branch runs to the Heart (*xin*), where it joins the Pericardium primary channel (deep Yin–Yin connection) and further reaches → **Ren-17** (*shanzhong*).
Circulation: Circadian (according to the Organ clock)
Importance: The Pericardium primary channel receives Nutritive Qi (*ying qi*) from the Kidney primary channel (first circulation of the *ying qi* → 1.1.4).

Connections with *zangfu*-Organ systems

Pericardium (*xin bao*), Triple Burner (*san jiao*)

4.9.2 The Pericardium Divergent Channel (*shou jue yin jing bie*)

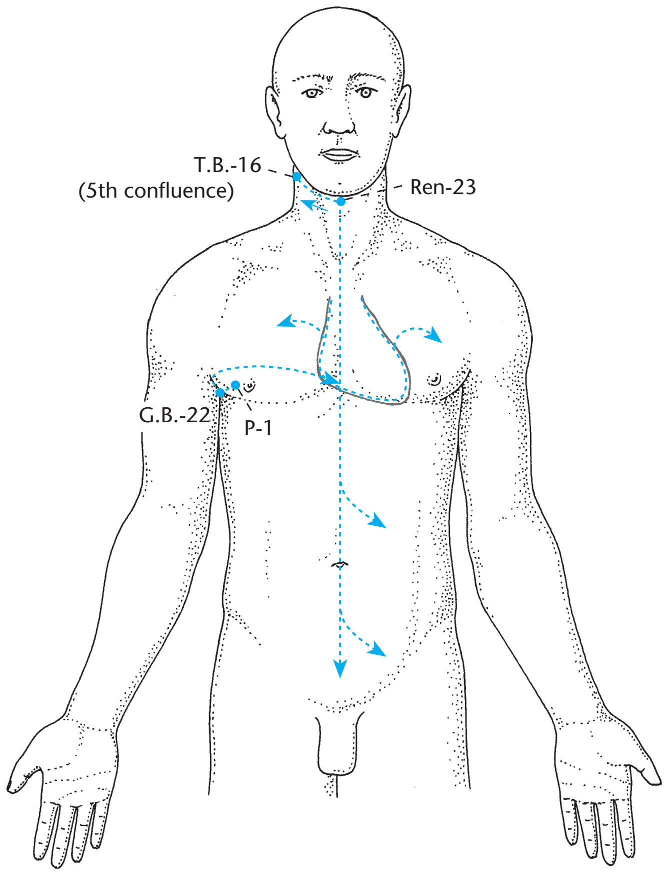

Pathway

The Pericardium divergent channel separates from the Pericardium primary channel near **P-1** (*tianchi*), 1 cun lateral to the nipple on the level of the 4th intercostal space. It

- runs in a lateral direction to → **G.B.-22** (*yuanye*)
- penetrates the thorax, continues to the Heart (*xin*) and divides into **two branches:**
  - **one branch** connects with the Upper, Middle and Lower Burner
  - the **other branch** ascends to the neck to → **Ren-23** (*lianquan*) and continues to the lateral aspect of the neck, emerges at → **T.B.-16** (*tianyou*) on the posterior border of the sternocleidomastoid muscle on the level of the mandibular angle, connecting with the Triple Burner primary channel and the Triple Burner divergent channel to form one of the 6 *he*-confluences (here: P/T.B. as 5th Confluence → 1.3)

Clinical importance

- Strengthens the relationship between the Pericardium and the Triple Burner (*zangfu*-Organ systems). Points on the Pericardium primary channel can therefore be used for disorders of the Triple Burner; vice versa, points on the Triple Burner primary channel can treat disorders of the Pericardium.

- In contrast to the Pericardium primary channel, the Pericardium divergent channel reaches the occiput and the neck. This explains why several points on the P channel can be used for disorders of the neck and ears, chronic throat pain, etc.
- By reaching the thorax, breast and hypochondrium, the Pericardium divergent channel supports the relationship between the Pericardium primary channel and these regions.

4.9.3 The Pericardium Sinew Channel (*shou jue yin jing jin*)

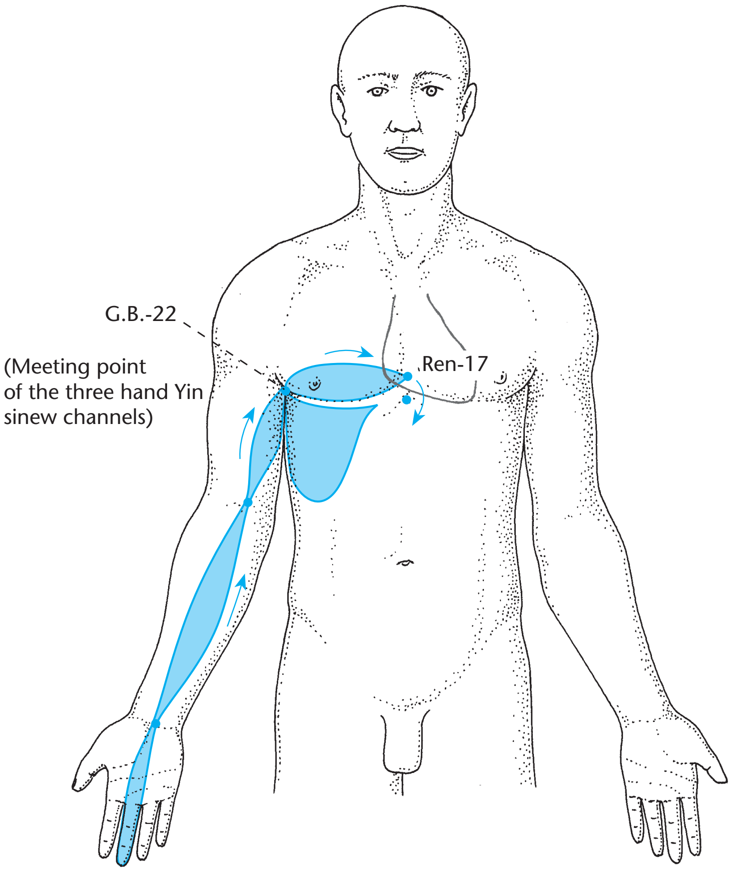

Pathway

The Pericardium sinew channel begins at the tip of the middle finger

➡ runs along the anterior aspect of the middle finger and along the 3rd metacarpal bone

➡ spreads in the centre of the palm and meets the Lung sinew channel

➡ binds (*jie*) in the centre of the wrist

➡ ascends the anterior aspect of the forearm as well as the anteromedial aspect of the upper arm to the axilla, binding (*jie*) at **P-3** (*quze*) on the elbow and at → **G.B.-22** (*yuanye*), where it meets the other hand Yin sinew channels.

At → **G.B.-22**, the channel separates into two branches:

- **one branch** disperses on the anterior, lateral and posterior aspects of the costal region
- **the other branch** penetrates the thorax below the axilla and reaches the midthoracic region on the level of → **Ren-17** (*shanzhong*), from where it spreads in the thoracic cavity to bind (*jie*) at the diaphragm.

Clinical importance

Pathology: Stiffness, pain, spasms and distending sensations along the pathway of the Pericardium sinew channel. Thoracic pain and tightness, *xi fen* syndrome (one of the five *ji*-syndromes, an ‘accumulation syndrome’) due to Lung Qi stagnation with accumulation of Phlegm and Heat manifesting with symptoms such as masses in the right hypochondriac region, possibly accompanied by thoracic or back pain, haemoptysis, chills and fever, shortness of breath and dyspnoea, etc.

Indication: Mainly for pain, spasms and distending sensations along the pathway of the sinew channel, limited range of motion along the medial aspect of the arm and elbow, swellings and pain in the axilla and the breast.

4.9.4 The Pericardium *luo*-Connecting Vessel System (*shou jue yin luo mai*)

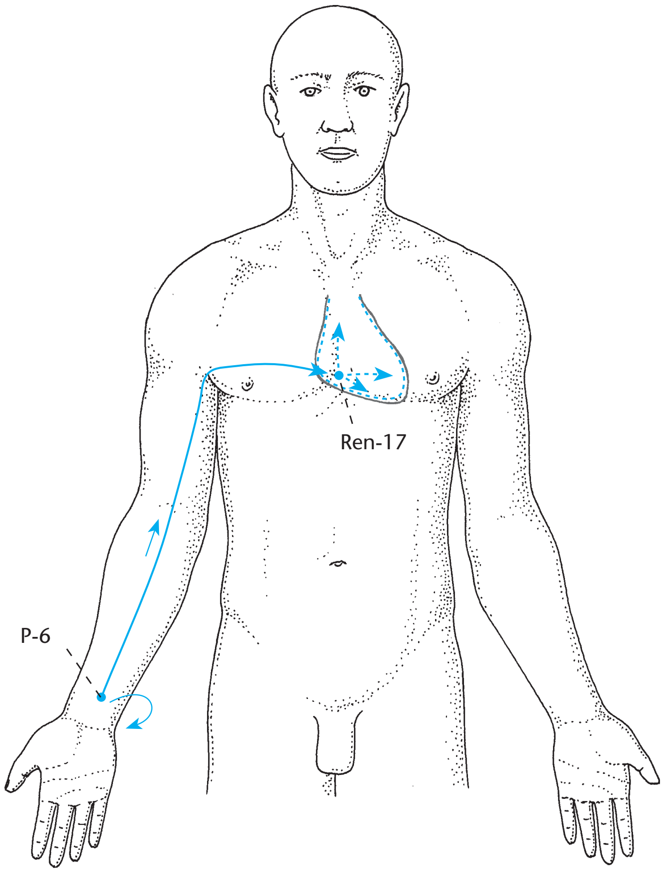

Pathway

The *luo*-connecting Pericardium vessel system separates from the Pericardium primary channel at its associated *luo*-connecting point **P-6** (*neiguan*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Triple Burner primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** Pericardium *luo*-connecting vessel to the *yuan*-source point → **T.B.-4** (*yangchi*).
- **A longitudinal division** follows the Pericardium primary channel along the anterior aspect of the arm, crosses → **Ren-17** (*shanzhong*) on the midline on the level of the 4th intercostal space, penetrates the thorax and terminates at the Pericardium (*xin bao*) and the Heart (*xin*).

Clinical importance (→ 8.1.2)

Pathology

Excess (*shi*): Heart pain, angina pain, thoracic pain
Deficiency (*xu*): Restlessness, irritability

4.9.5 Cutaneous Region (*jue yin pi bu*)

See description and figures → 1.6.

4.9.6 Points of the Primary Pericardium Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): P-7** (*daling*) 🟦■
- ***Luo*-connecting point (→ 8.1.2.): P-6** (*neiguan*) 🟦🟦
- ***Xi*-cleft point (→ 8.1.3): P-4** (*ximen*) 🟦■
- **Associated Back-*shu* point (→ 8.1.4): BL-14** (*jueyinshu*)
- **Associated Front-*mu* point (→ 8.1.5): Ren-17** (*shanzhong*) 🟦🟦
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood), tonification point: **P-9** (*zhongchong*) 🟦■
  ying-spring point (Fire), *ben* point (Five Phases): **P-8** (*laogong*) 🟦■
  shu-stream point (Earth), sedation point: **P-7** (*daling*) 🟦■
  jing-river point (Metal): **P-5** (*jianshi*) 🟦■
  he-sea point (Water): **P-3** (*quze*) 🟦■
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *yin wei mai:* P-6** (*neiguan*) 🟦🟦
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the LIV, T.B., G.B.* channels: **P-1** (*tianchi*)
  – of other channels with the Pericardium channel: –
- **Gao Wu command point (→ 8.1.11) for the thoracic region: P-6** (*neiguan*) 🟦🟦
- **Window of Heaven point (→ 8.1.12): P-1** (*tianchi*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): P-7** (*daling*) 🟦■, **P-8** (*laogong*) 🟦■
- **Other functional points:**
  – important point for eliminating Phlegm misting the orifices: **P-5** (*jianshi*) 🟦■
  – important point for calming the *shen*: **P-6** (*neiguan*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** hand – **P-7** (*daling*) 🟦■
- **Adjacent points (→ 8.2.1): –**
- **Distal points (→ 8.2.1):** tongue – **P-8** (*laogong*) 🟦■; Lung – **P-6** (*neiguan*) 🟦🟦; Heart – **P-4** (*ximen*) 🟦■; **P-6** (*neiguan*) 🟦🟦; Spleen – **P-6** (*neiguan*) 🟦🟦; thorax and epigastrium – **P-6** (*neiguan*) 🟦🟦; foot – **P-7** (*daling*) 🟦■

Specific points according to the channel pathway (in numeric order)

- **P-1** (*tianchi*) 🟦■: *jiaohui*-meeting point with the LIV, T.B., G.B.* channels (→ 8.1.10); Window of Heaven point (→ 8.1.12)
- **P-3** (*quze*) 🟦■: lower *he*-sea point (Water) (→ 8.1.6)
- **P-4** (*ximen*) 🟦■: *xi*-cleft point (→ 8.1.3); distal point for the Heart (→ 8.2.1)
- **P-5** (*jianshi*) 🟦■: *jing*-river point (Metal) (→ 8.1.6); important point for eliminating Phlegm misting the orifices
- **P-6** (*neiguan*) 🟦🟦: *luo*-connecting point (→ 8.1.2); Gao Wu command point (→ 8.1.11) for the thorax; Opening point (→ 8.1.8) of the *yin wei mai*; distal point for the Heart, Spleen, thorax and epigastrium (→ 8.2.1); important point for calming the *shen*
- **P-7** (*daling*) 🟦■: *yuan*-source point (→ 8.1.1); *shu*-stream point (Earth) (→ 8.1.6); sedation point; Sun Si Miao Ghost point (→ 8.1.15); local point for the hand (→ 8.2.1); distal point for the foot (→ 8.2.1)
- **P-8** (*laogong*) 🟦■: *ying*-spring point (Fire) (→ 8.1.6); *ben* point (Five Phases); Sun Si Miao Ghost point (→ 8.1.15); distal point for the tongue (→ 8.2.1)
- **P-9** (*zhongchong*) 🟦■: *jing*-well point (Wood) (→ 8.1.6); tonification point

---

* Mentioned by only some authors

Heavenly Pool TIANCHI P-1

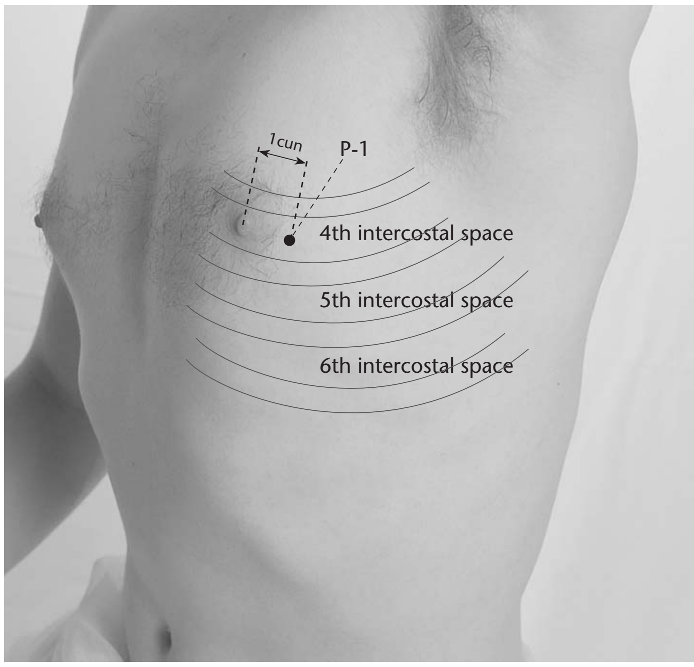
1cun
P-1
4th intercostal space
5th intercostal space
6th intercostal space

Location

In the 4th intercostal space (ICS), 1 cun lateral to the nipple.

How to find

Quick method, especially in male patients: The 4th ICS is located on the same level as the nipple. Locate **P-1** 1 cun lateral to the nipple in the 4th ICS (note: ascending course of the ICS towards lateral). **Or:** For better orientation in the intercostal region, first, locate the manubriosternal synchondrosis (→ 3.4.2.). This is on the level of the costal cartilage of the 2nd rib, followed by the 2nd ICS below it. Count down to ICS to the 4th ICS and measure 5 cun in a lateral direction. There, locate **P-1**.

Located on the same level (4th ICS) are → **Ren-17** (on the midline), → **KID-23/ST-17/SP-18** (2/4/6 cun lateral to the midline), as well as → **G.B.-22** and → **G.B.-23**.

Needling

0.5–1 cun transversely (subcutaneously) along the course of the ICS. Especially in women, transverse needling only, avoid puncturing the mamillary tissue. Caution: Pneumothorax.

Actions/Indications

- Descends counterflow Qi
- Eliminates excess (especially Heat and Phlegm) from the thorax
- Promotes lactation and eliminates local Qi stagnation

Special features

Meeting point with the LIV and T.B. channels, also with the G.B. channel according to some authors. Window of Heaven point, entry point.

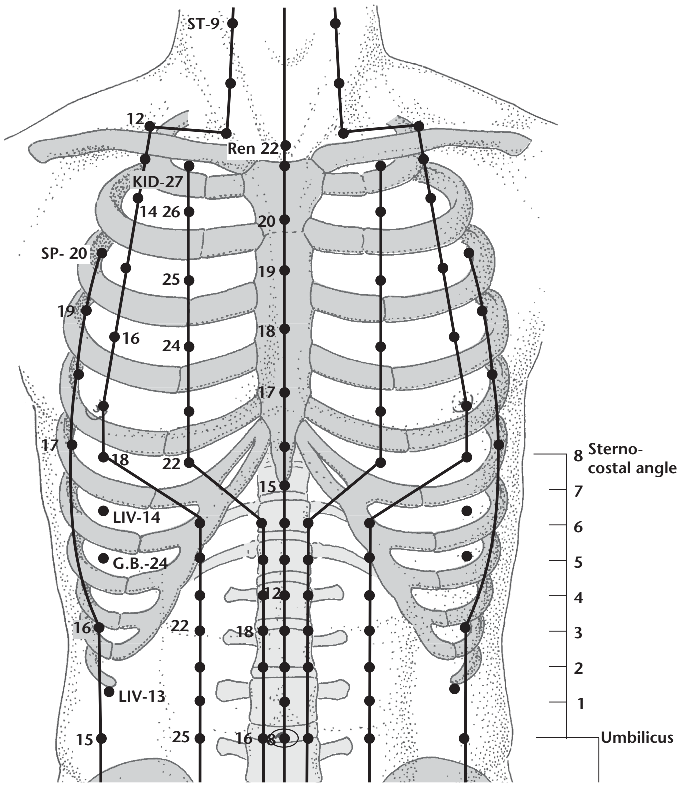
ST-9
12
Ren 22
KID-27
14 26
20
SP- 20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
8 Sterno-costal angle
7
6
5
4
3
2
1
Umbilicus

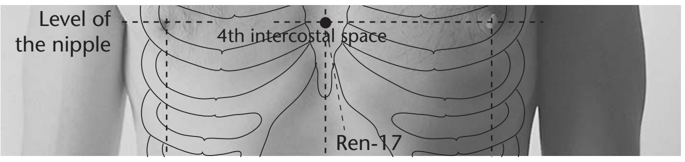
Level of the nipple
4th intercostal space
Ren-17

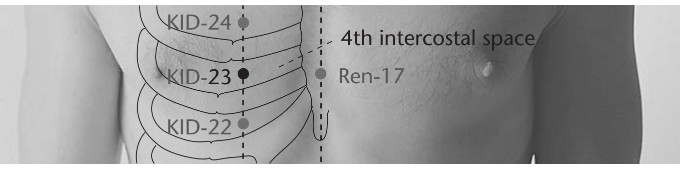
KID-24
KID-23
KID-22
4th intercostal space
Ren-17

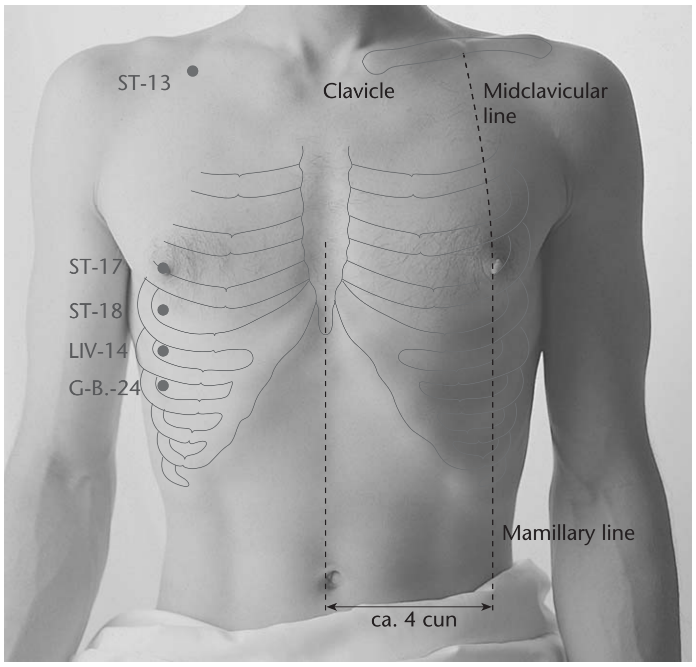
ST-13
Clavicle
Midclavicular line
ST-17
ST-18
LIV-14
G-B.-24
Mamillary line
ca. 4 cun

P-2 *Heavenly Spring* TIANQUAN

Location

Between the two heads of the biceps brachii muscle, 2 cun inferior to the axillary fold.

How to find

Locate the biceps brachii muscle through slight external rotation; in patients with less muscle definition, ask the patient to flex their elbow against resistance. **P-2** is located between the two heads of the biceps brachii muscle, 2 cun inferior to the upper end of the anterior axillary fold (→ 2.2).

→ **LU-3** is located radial to **P-2,** on the lateral aspect of the biceps brachii muscle and 1 cun more distally.

Needling

1–1.5 cun obliquely in a proximal or lateral direction.

Actions/Indications

- Unbinds the chest.
- Nourishes and invigorates the Blood
- Calms the *shen*

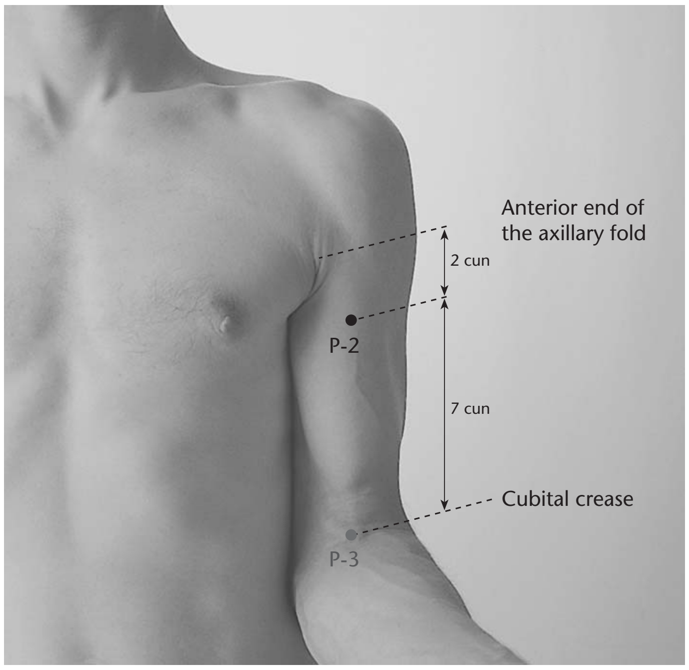
Anterior end of the axillary fold
2 cun
P-2
7 cun
Cubital crease
P-3

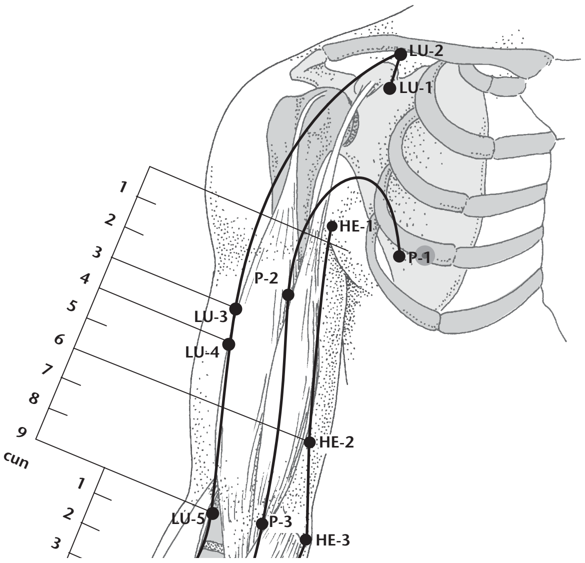
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
1
2
3
4
5
6
7
8
9
cun
1
2
3

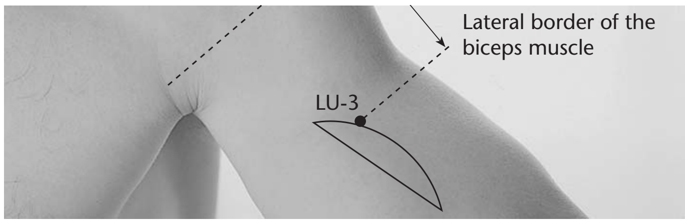
Lateral border of the biceps muscle
LU-3

Marsh at the Crook QUZE **P-3**

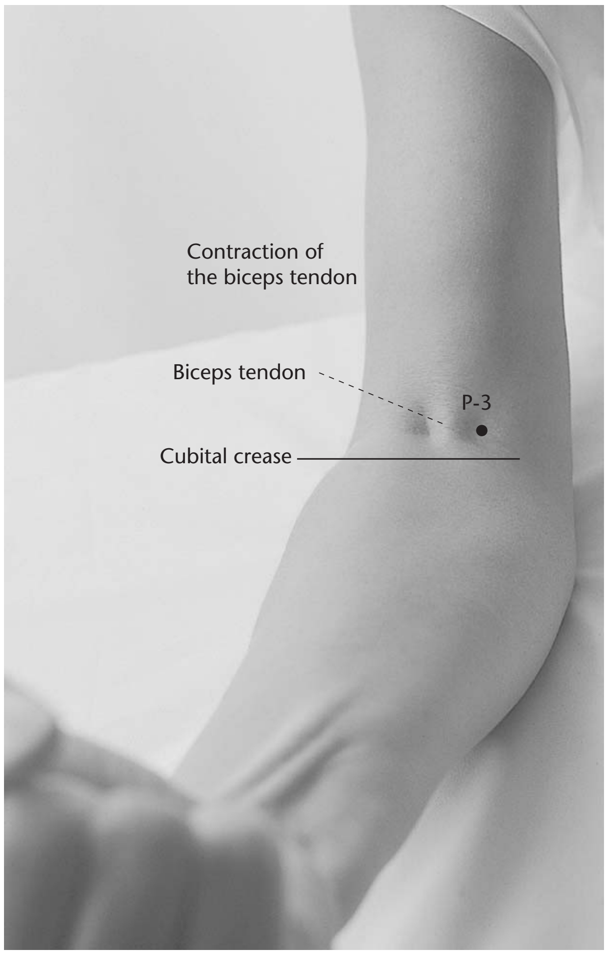
Contraction of the biceps tendon
Biceps tendon
P-3
Cubital crease

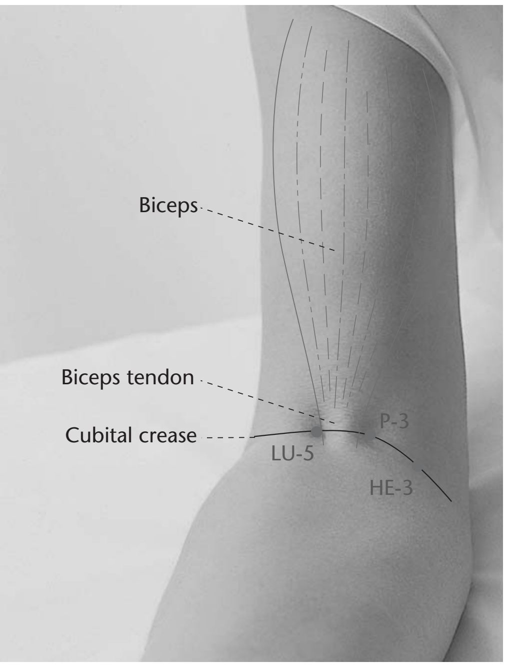
Biceps
Biceps tendon
P-3
Cubital crease
LU-5
HE-3

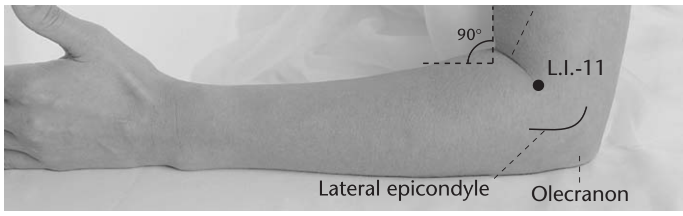
90°
L.I.-11
Lateral epicondyle
Olecranon

Location

On the cubital crease, on the ulnar side of the biceps tendon, between the tendon and the brachial artery.

How to find

Place the patient's arm in a relaxed supinated position. This point is best located with the elbow bent and the biceps flexed in order to make the tendon and elbow crease more clearly visible and palpable. **P-3** is located on the cubital crease, on the ulnar aspect of the tendon.

Also located in the area of the cubital crease are → **LU-5** (on the radial side of the tendon), → **L.I.-11** (between the radial end of the crease and the lateral epicondyle of the humerus) and → **HE-3** (with the elbow fully flexed, at the ulnar end of the cubital crease).

Needling

0.5–1 cun vertically or prick to cause bleeding of the vein. Caution: Brachial artery/vein. Palpate for the pulse before needling.

Actions/Indications

- Clears Heat, drains Fire
- Harmonises the Stomach and Intestines, stops vomiting
- Opens the channel, alleviates pain

Special features

He-sea point, Water point. Important local point and important point to drain Heat. For Heat conditions, apply bloodletting.

P-4 Xi *Cleft Gate* XIMEN

Location

5 cun proximal to the anterior space of the wrist joint (most distal wrist crease), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

Spreading hands technique (➞ 2.3.3): Place the little fingers in the centre of the anterior wrist joint space (➞ 3.3.3, ➞ **P-7**) and the cubital crease respectively (➞ **P-3**). From the midpoint of this distance, measure 1 cun in a distal direction. On this level, locate **P-4** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-4** is located on its ulnar aspect.

Needling

0.5–1 cun vertically or 1–1.5 cun obliquely in a proximal direction.

Actions/Indications

- Tonifies the Blood, dispels Blood stasis
- Calms the *shen*
- Cools the Blood
- Opens the channel

Special features

Xi-cleft point. Major point for acute, painful Blood stasis in the region of the thorax and the Heart

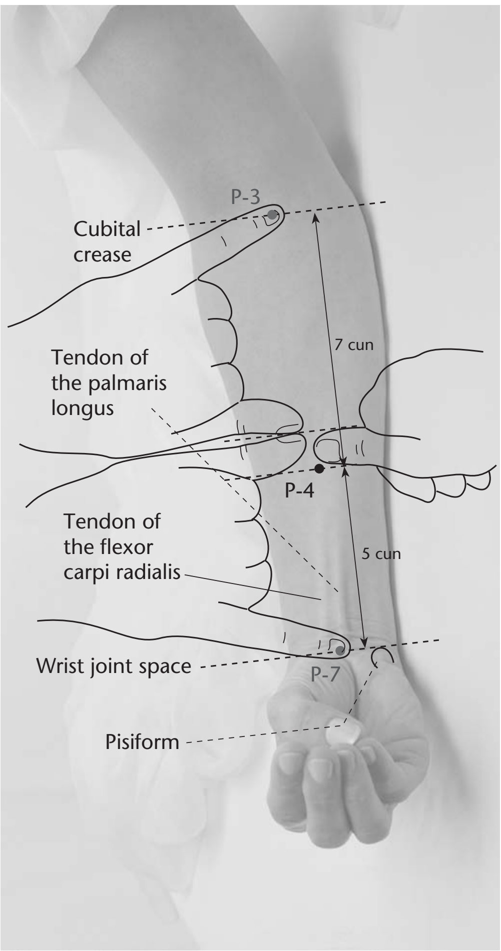
P-3
Cubital crease
7 cun
Tendon of the palmaris longus
P-4
Tendon of the flexor carpi radialis
5 cun
Wrist joint space
P-7
Pisiform

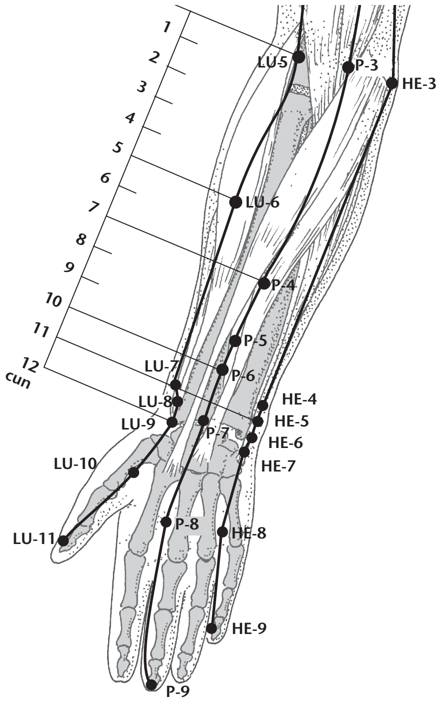
1
2
3
4
5
6
7
8
9
10
11
12 cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

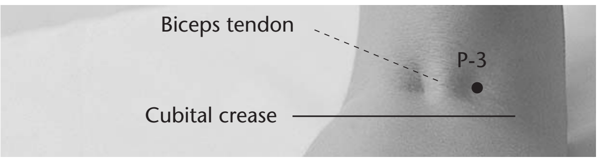
Biceps tendon
P-3
Cubital crease

Intermediary Messenger JIANSHI P-5

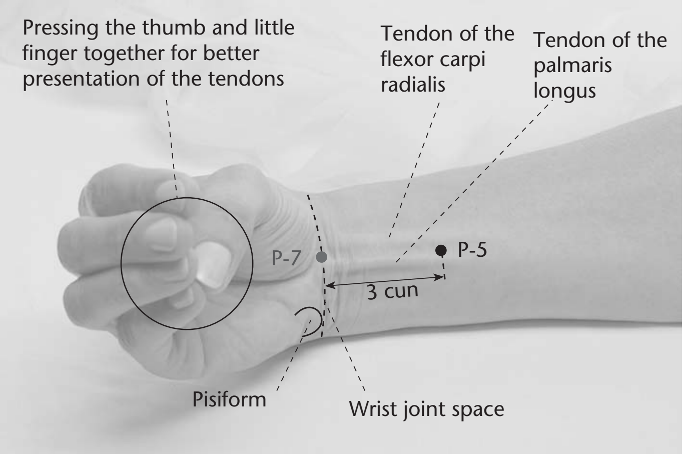
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
P-7
P-5
3 cun
Pisiform
Wrist joint space

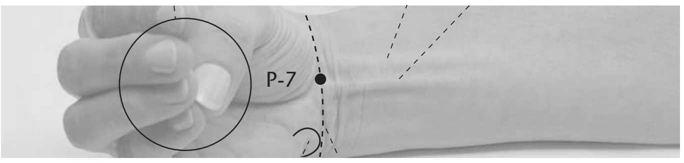

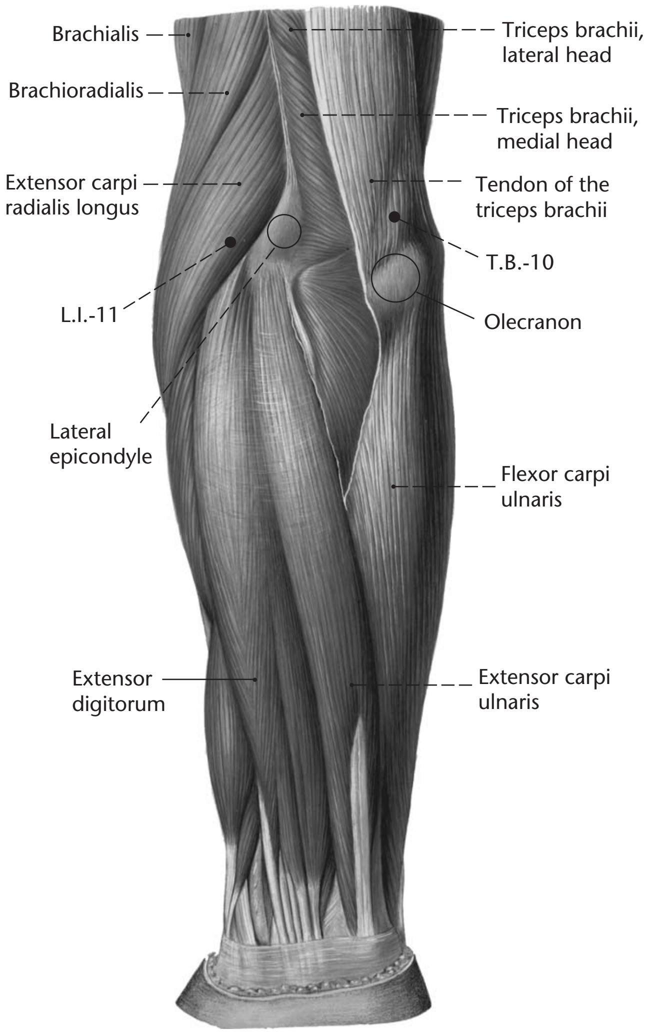
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Extensor digitorum
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris
Extensor carpi ulnaris

Location

3 cun proximal to the anterior wrist joint space (‘most distal wrist crease’), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). It can easily be palpated by moving the patient’s hand in a relaxed way. From the centre of the joint space (→ **P-7**), measure 3 cun in a proximal direction and locate **P-5** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-5** is located at its ulnar aspect.

Needling

0.5–1 cun vertically or obliquely in a proximal direction. Caution: The median nerve here is only covered by the radial head of the flexor digitorum superficialis. Vertical needling should not exceed 0.5 cun; no strong needle stimulation.

Actions/Indications

- Calms the *shen* and the Heart
- Harmonises the Middle Burner
- Clears Heart Fire
- Unbinds the chest
- Transforms invisible Phlegm

Special features

Jing-river point, Metal point. Important point for transforming Phlegm blocking the orifices.

P-6 *Inner Gate* NEIGUAN

Location

2 cun proximal to the anterior wrist joint space ('most distal wrist crease'), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). It can easily be palpated by moving the patient's hand in a relaxed way. From the centre of the wrist joint space (→ **P-7**), measure 2 cun in a proximal direction and locate **P-6** between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-6** is located on its ulnar aspect.
Or quick method: From the centre of the wrist joint space, glide in a proximal direction between the two tendons. The skin will begin to form a bulge in front of the sliding finger, which will come to a halt at **P-6.**
→ **T.B.-5** (2 cun proximal to the dorsal wrist joint space) is located opposite **P-6**.

Needling

0.5–1 cun vertically or obliquely in a proximal or lateral direction. Caution: The median nerve is located directly below this point.

Actions/Indications

- Calms the *shen*
- Unbinds the chest
- Harmonises the Stomach
- Opens the channel

Special features

Luo-connecting point, opening (master) point of the *yin wei mai*, Gao Wu command point for the thorax. Main point for nausea and vomiting. Very important point.

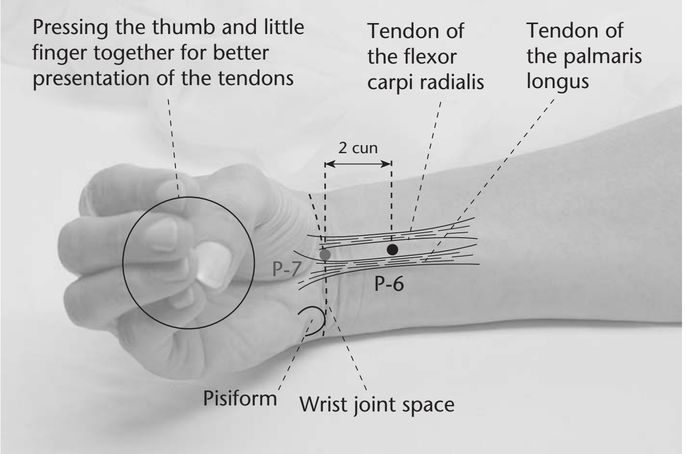
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
2 cun
P-7
P-6
Pisiform
Wrist joint space

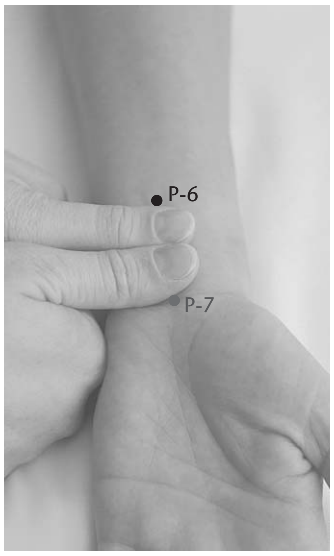

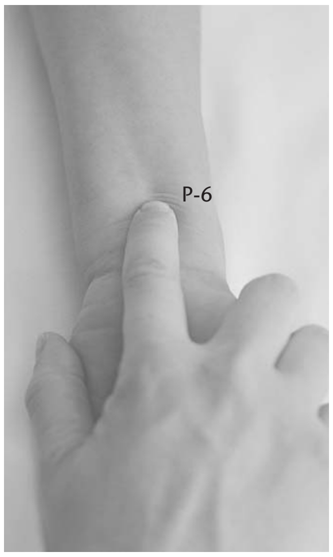

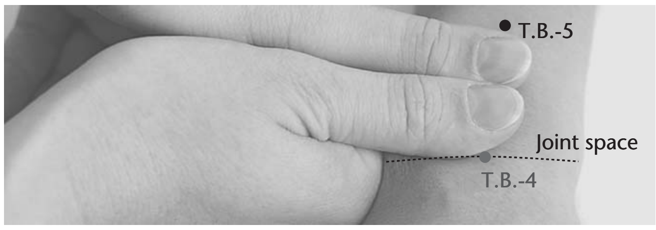
T.B.-5
Joint space
T.B.-4

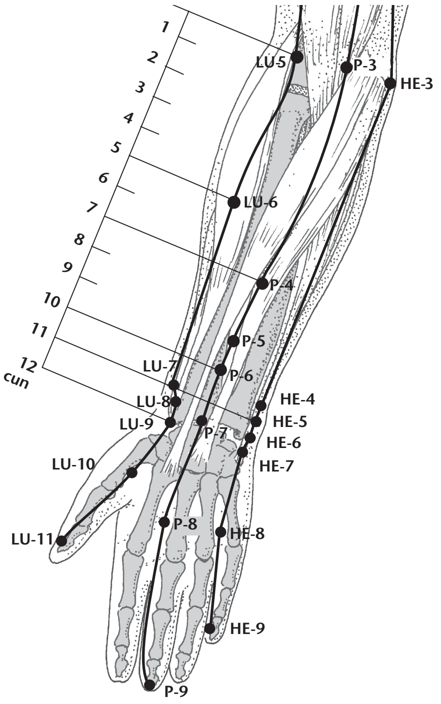
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
HE-4
LU-8
HE-5
LU-9
P-7
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

Great Mound DALING P-7

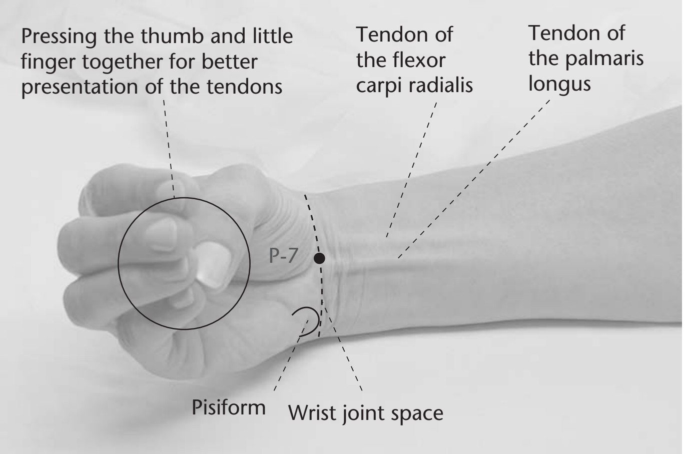
Pressing the thumb and little finger together for better presentation of the tendons
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
P-7
Pisiform
Wrist joint space

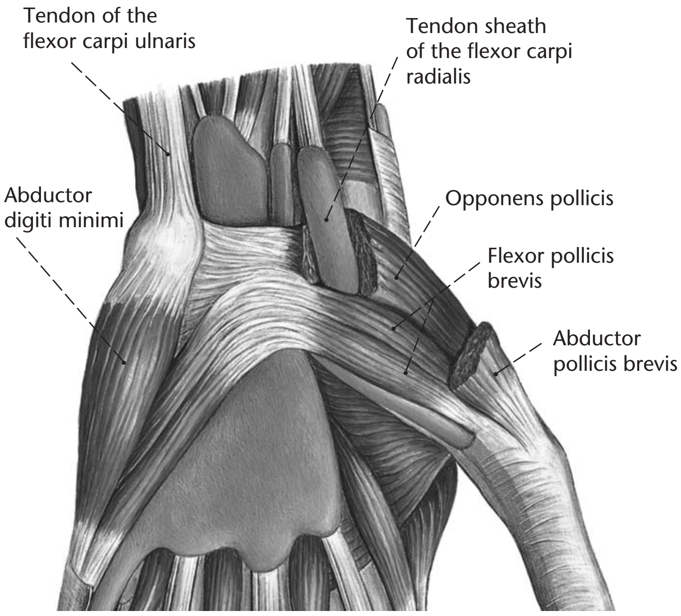
Tendon of the flexor carpi ulnaris
Tendon sheath of the flexor carpi radialis
Abductor digiti minimi
Opponens pollicis
Flexor pollicis brevis
Abductor pollicis brevis

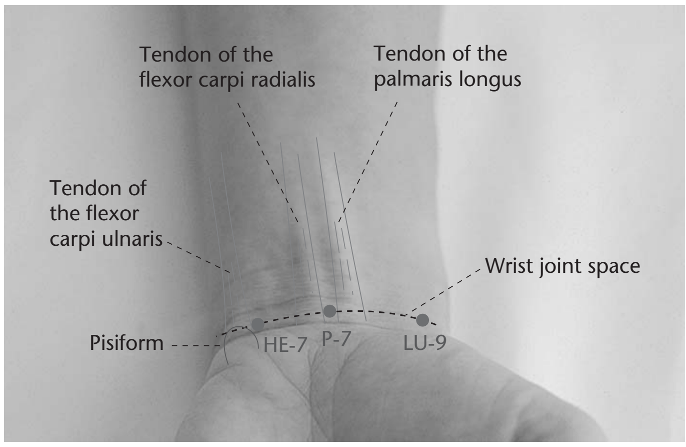
Tendon of the flexor carpi radialis
Tendon of the palmaris longus
Tendon of the flexor carpi ulnaris
Wrist joint space
Pisiform
HE-7
P-7
LU-9

Location

On the anterior aspect of the wrist joint space ('most distal wrist crease'), between the tendons of the palmaris longus and flexor carpi radialis muscles.

How to find

As the location of the wrist crease varies, the wrist joint space is a more reliable reference point (→ 3.3.3). By moving the patient's hand in a relaxed way, it can easily be palpated. Locate **P-7** on this level between the two tendons, which become more pronounced when the fingertips are pressed together. If only one tendon is visible, this will be the tendon of the flexor carpi radialis muscle. **P-7** is located on its ulnar aspect.

Located on the same level are → **HE-7** (on the ulnar aspect of the wrist joint, radial to the tendon of the flexor carpi ulnaris muscle) and → **LU-9** (on the radial aspect of the wrist joint, lateral to the radial artery.

Needling

Vertically 0.3–0.5 cun, for carpal tunnel syndrome also 0.5–1 cun obliquely in a lateral direction along the tunnel. Caution: The median nerve is located directly under this point, thus needling can cause a significant electric sensation; in this case, no further needle manipulation to avoid possible injury.

Actions/Indications

- Clears Heat from the Heart, calms the *shen*
- Harmonises the Stomach and Intestines
- Cools Heart Fire affecting the Bladder
- Unbinds the chest
- Cools the Blood
- Opens the channel, alleviates pain

Special features

Yuan-source point, *shu*-stream point, Earth point, sedation point, Sun Si Miao Ghost point. Important point for calming, main point for carpal tunnel syndrome.

P-8 *Palace of Toil* LAOGONG

Location

In the centre of the palm, between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone.

How to find

When making a loose fist, the tip of the middle finger will rest on **P-8,** between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone.

→ **HE-8** is located more radially, between the 4th and 5th metacarpal bones.

Needling

Vertically 0.3–0.5 cun. Thinner needles should be used as needling this point can be painful.

Actions/Indications

- Clears Heart Fire and calms the *shen*
- Harmonises the Middle Burner
- Clears the *ying*-nutritive level and cools the Blood

Special features

Ying-spring point, Fire point, *ben* point (Five Phases), Sun Si Miao Ghost point, exit point.

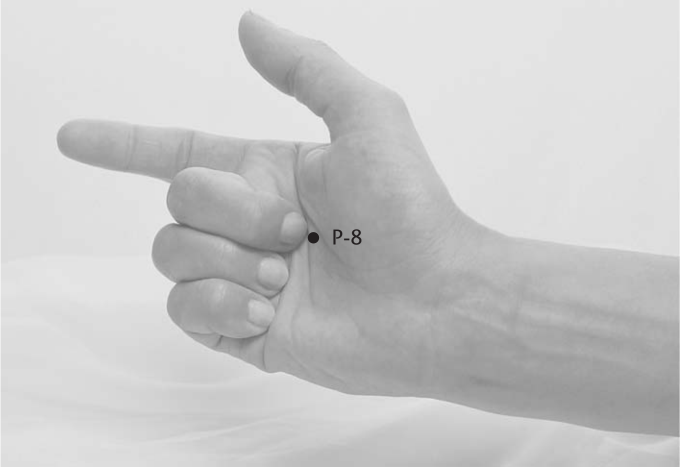

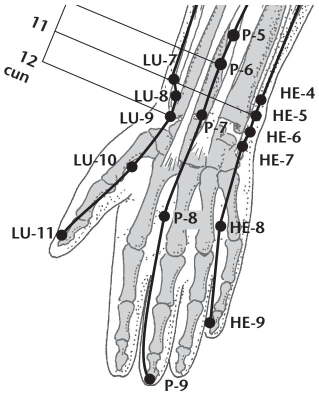

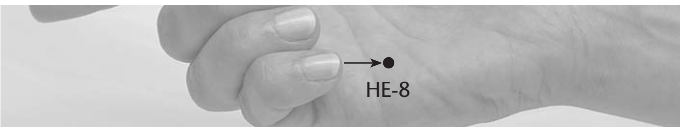

Middle Rushing ZHONGCHONG **P-9**

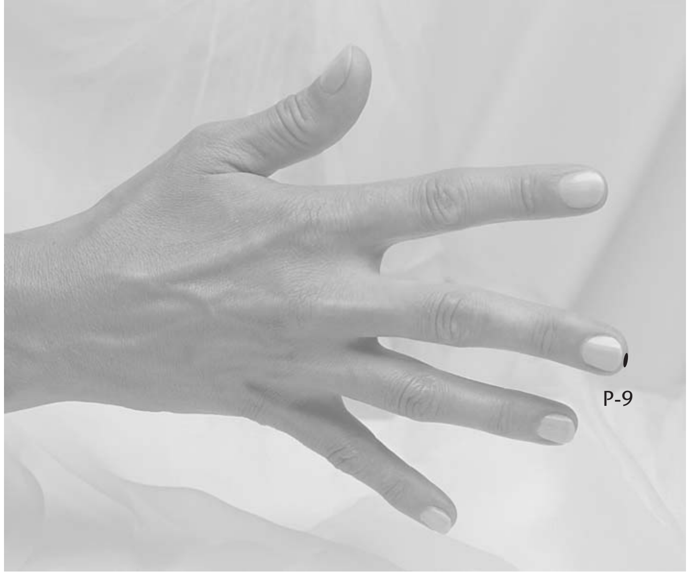

Location

On the most distal point of the middle finger.

How to find

Find the most distal point on the tip of the middle finger and there locate **P-9**.

Needling

Vertically 0.1–0.2 cun or prick to bleed. Caution: Painful point.

Actions/Indications

- Clears Heat, especially from the Heart and Pericardium
- Strengthens collapsed Yang
- Expels internal Wind

Special features

Jing-well point, Wood point, tonification point.

4.10 The Triple Burner Channel System – Hand *shaoyang* (*shou shao yang jing luo*)

4.10.1 Triple Burner Primary Channel (*shou shao yang jing*)

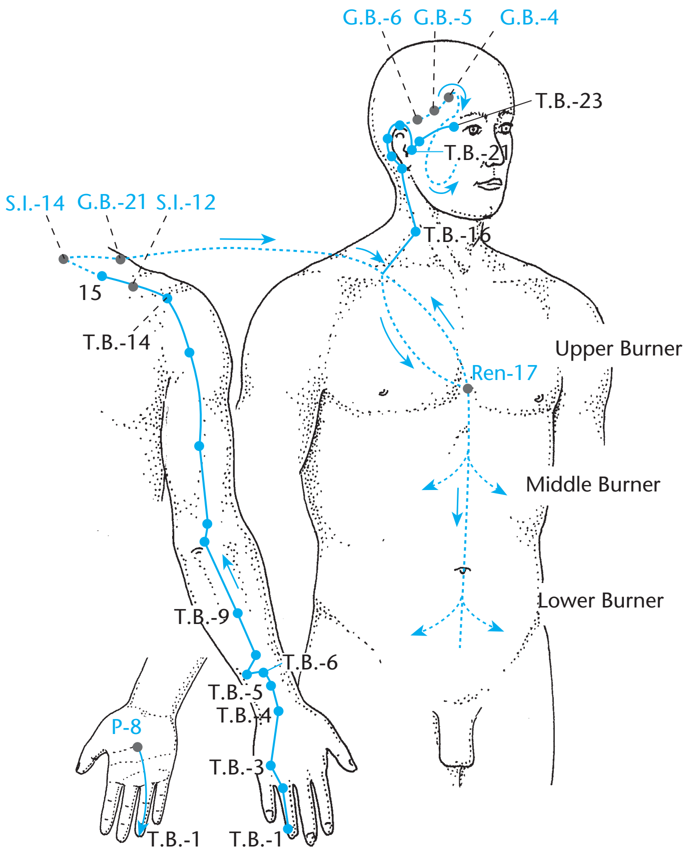

Pathway

The **external pathway** of the Triple Burner primary channel begins on the ring finger, at the ulnar corner of the nail at **T.B.-1** (*guanchong*). **T.B.-1** is reached by a branch of the Pericardium primary channel that separates on the palm of the hand at → **P-8** (*laogong*) (hand Yin–Yang connection of the third great circuit). The channel runs along the ulnar side of the ring finger

- continues between the 4th and 5th metacarpal bones
- travels along the posterior aspect of the forearm between the radius and the ulna towards the lateral aspect of the arm
- passes the olecranon and continues along the upper arm to the shoulder
- traverses the shoulder, intersects with → **S.I.-12** (*bingfeng*) and meets the other Yang primary channels at → **Du-14** (*dazhu*)
- runs to the anterior aspect of the body, crossing → **G.B.-21** (*jianjing*) and → **ST-12** (*quepen*), the latter in the supraclavicular fossa.

At **ST-12,** the channel enters the body to begin its **internal** section. It meets the Pericardium (*xin bao*) at → **Ren-17** (*shanzhong*), penetrates the diaphragm and enters the abdomen, connecting with the Upper, Middle and Lower Burners.

From the Lower Burner, an **internal branch** (according to some authors, the T.B. channel) descends to the popliteal fossa to → **BL-39** (*weiyang*), the lower *he*-sea point of the Triple Burner.

A further **inner branch** originates in the thorax at → **Ren-17** (*shanzhong*), emerges at the supraclavicular fossa and ascends to the posterior border of the sternocleidomastoid muscle to **T.B.-16**, follows via **T.B.-17, T.B.-18, T.B.-19** the temporal bone along its border behind the ear and reaches **T.B.-20** (*jiasun*) above the apex of the ear. It crosses the Gall Bladder primary channel at **G.B.-6** (*xuanli*), **G.B.-5** (*xuanlu*) and **G.B.-4** (*hanyan*), according to Deadman et al 1998) also at **G.B.-11** (*touqiaoyin*) and **G.B.-14** (*yangbai*), descends the cheek to the lower jaw and ascends again to intersect with → **S.I.-18** (*quanliao*) on the lower border of the zygomatic arch and continues to the infraorbital region.

From **T.B.-17** (*yifeng*), a branch penetrates the ear, emerges to the **Exterior** anterior to the tragus, intersects with → **S.I.-19** (*tinggong*), continues past **T.B.-21** (*ermen*) and **T.B.-22** (*erheliao*), meets → **G.B.-3** (*shangguan*), crosses its own cheek branch and ends at **T.B.-23** (*sizhukong*) on the frontozygomatic suture. From **T.B.-23**, or, according to some authors from **T.B.-22**, a small branch runs to → **G.B.-1** (*tongziliao*) at the lateral canthus of the eye (Yang axes of the third great circuit: *shao yang*).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Neck pain, cheek pain, red and painful eyes, deafness, pain behind the ears, pain on the posterior aspect of the shoulder and the upper arm

Interior (*li*) or *zangfu*-Organ signs and symptoms: Abdominal fullness and distension, difficult urination, skin oedema and skin swellings, enuresis

Connections of the Triple Burner primary channel (→ 1.2)

Connections with other channels

Pericardium primary channel (*shou jue yin jing*)

Connection: Hand Yin–Yang pairing of the third great circuit
Location: P-8 → **T.B.-1** (on the hand).
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Gall Bladder primary channel (*zu shao yang jing*)

Connection: Paired according to the six-channel theory (hand–foot pairing): *shao yang* (Yang axes of the third great circuit)
Location: T.B.-23 (or **T.B.-22**) → **G.B.-1** (on the head).
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with *zangfu*-Organ systems
Pericardium (*xin bao*), Triple Burner (*san jiao*)

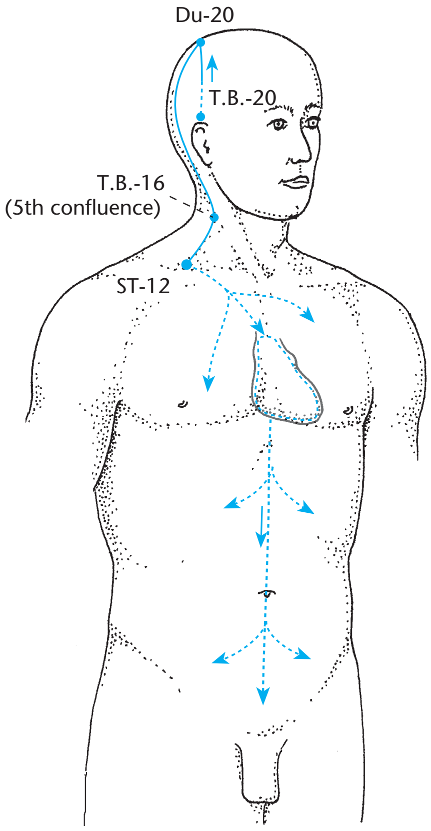

4.10.2 The Triple Burner Divergent Channel (*shou shao yang jing bie*)

Pathway

The Triple Burner divergent channel separates from the Triple Burner primary channel near **T.B.-20** (*jioasun*)

➡ ascends to → **Du-20** (*baihui*)

➡ descends along the posterior aspect of the ear to the neck to **T.B.-16** (*tianyou*) on the posterior border of the sternocleidomastoid muscle, where it connects with the Triple Burner primary channel and the Pericardium primary channel to form one of the six *he*-confluences (here: P/T.B. as fifth confluence → 1.3)

➡ The Triple Burner divergent channel further descends to → **ST-12** (*quepen*) in the supraclavicular fossa, penetrates the thorax, disperses in the Pericardium (*xin bao*) and reaches the Upper, Middle and Lower Burners.

Clinical importance

- Strengthens the relationship between the Upper, Middle and Lower Burners. T.B. points can be used for disorders affecting these areas and Organs.
- The Triple Burner divergent channel reaches the vertex at → **Du-20**. Therefore, T.B. points can be used for headaches, dizziness, etc, but also to raise the Yang Qi for prolapse, etc.
- The Triple Burner divergent channel spreads in the thorax, so that T.B. points can be used to support the treatment of thoracic problems such as pain, coughing etc.

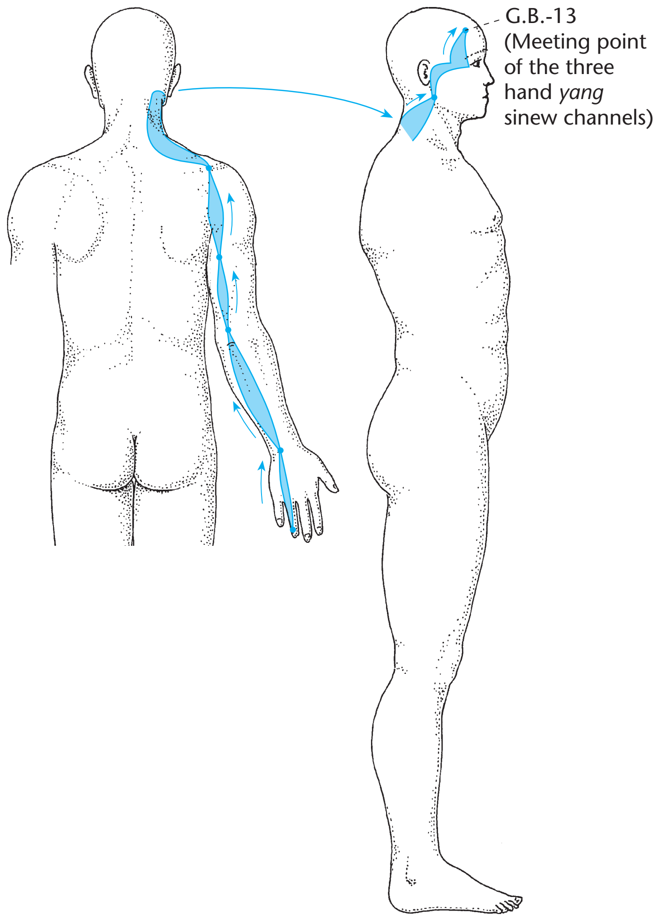

4.10.3 The Triple Burner Sinew Channel (*shou shao yang jing jin*)

Pathway

The Triple Burner sinew channel begins on the ring finger, at the ulnar corner of the nail at **T.B.-1** (*guanchong*)

➡ travels between the 4th and 5th metacarpal bones to the wrist

➡ binds (*jie*) at **T.B.-4** (*yangchi*)

➡ continues along the posterior aspect of the forearm between the Small Intestine and Large Intestine sinew channels, binds (*jie*) at the olecranon, ascends the posterior aspect of the upper arm to bind (*jie*) at the posterior corner of the acromion, traverses the posterior aspect of the shoulder and the lateral aspect of the occiput and there meets the Small Intestine sinew channel. It then reaches the mandibular angle, where it divides into **two branches**:

➡ **One branch** enters internally and penetrates the root of the tongue

- ➡ **The other branch** ascends in front of the ear, binds (*jie*) at the outer canthus of the eye, crosses the temple and terminates at → **G.B.-13** (*benshen*) in the frontoparietal region. There, it also meets the other hand Yang sinew channels.

Clinical importance

Pathology: Stiffness, swellings and distending sensations along the course of the sinew channel. Curling or contraction of the tongue.

Indication: Predominantly for pain, stiffness, spasms, tension and distending sensations along the course of the channel, including the lateral aspects of the head and the body as well as the lateral sections of the extremities. T.B. points can also be used for muscle tension caused by emotional disorders. Owing to the connection of the sinew channel with the tongue, some T.B. points are also indicated for motor disorders of the tongue.

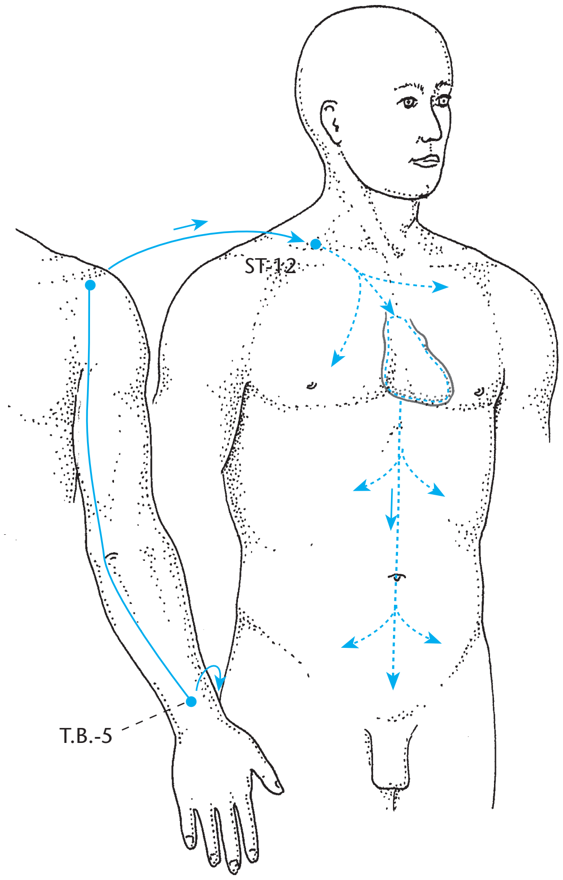

4.10.4 The Triple Burner *luo*-Connecting Vessel System (*shou shao yang luo mai*)

Pathway

The Triple Burner *luo*-connecting vessel system separates from the Triple Burner primary channel at its associated *luo*-connecting point **T.B.-5** (*waiguan*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and subbranches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- ➡ Horizontal divisions run to the Interiorly–Exteriorly paired Pericardium primary channel; according to some schools of thought (for example Ngyen Van Nghi → Appendix), they travel as a **transverse** Triple Burner *luo*-connecting vessel to the *yuan*-source point **P-7** (*daling*).
- ➡ **A longitudinal division** follows the posterior aspect of the arm, crosses the posterior shoulder region, continues to the front of the body and, according to some authors (for example, Solinas et al 1998), reaches → **ST-12** (*quepen*) in the supraclavicular fossa, penetrates the thorax and reaches the Pericardium primary channel, both dispersing in the Upper, Middle and Lower Burners.

Clinical importance

Pathology

Excess (*shi*): Tension of the elbow joint

Deficiency (*xu*): Muscle weakness in the arm, difficulty flexing the elbow

4.10.5 Cutaneous Region (*shao yang pi bu*)

See description and figures → 1.6.

4.10.6 Points of the Triple Burner Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): T.B.-4** (*yangchi*)
- ***Luo*-connecting point (→ 8.1.2.): T.B.-5** (*waiguan*) ■■
- ***Xi*-cleft point (→ 8.1.3): T.B.-7** (*huizong*)
- **Associated Back-*shu* point (→ 8.1.4): BL-22** (*sanjiaoshu*)
- **Associated Front-*mu* point (→ 8.1.5): Ren-5** (*shimen*)
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal): **T.B.-1** (*guanchong*)
  ying-spring point (Water): **T.B.-2** (*yemen*)
  shu-stream point (Wood), tonification point: **T.B.-3** (*zhongzhu*)
  jing-river point (Fire), *ben point* (Five Phases): **T.B.-6** (*zhigou*)
  he-sea point (Earth), sedation point: **T.B.-10** (*tianjing*)
- ***Hui*-meeting point (→ 8.1.7): –**
- **Opening point (→ 8.1.8) of the *yin wei mai*: T.B.-5** (*waiguan*) ■■
- **Lower *xiahe*-sea points (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - with the *yang wei mai* and the G.B. channel: **T.B.-13*** (*naohui*)
  - with the *yang wei mai* and the G.B. channel: **T.B.-15** (*tianliao*) ■■
  - with the G.B. channel: **T.B.-17** (*yifeng*) ■■
  - with the S.I.* and G.B. channel: **T.B.-20** (*jiaosun*)
  - with the S.I. and G.B. channels: **T.B.-22** (*erheliao*)

---

* Mentioned by only some authors

  - – of other channels with the T.B. channel: **S.I.-12, BL-11, (S.I.-18, S.I.-19*), Du-14, Ren-12, Ren-17, ST-12, G.B.-21, G.B.-14, G.B.-11, G.B.-6, G.B.-5, G.B.-4, G.B.-3, G.B.-1, (S.I.-9*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): T.B.-16** (*tianyou*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): –**
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points: –**

Points according to region

- **Local points (→ 8.2.1):** ears – **T.B.-21** (*ermen*) 🟦■; shoulder – **T.B.-14** (*jianliao*) 🟦■, **T.B.-15** (*tianliao*) 🟦■
- **Adjacent points (→ 8.2.1):** temporal aspect of the head – **T.B.-17** (*yifeng*) 🟦■; eyes – **T.B.-23** (*sizhukong*) 🟦■; ears – **T.B.-17** (*yifeng*) 🟦■; jaw – **T.B.-17** (*yifeng*) 🟦■; shoulder – **T.B.-13** (*naohui*), **T.B.-15** (*tianliao*) 🟦■
- **Distal points (→ 8.2.1):** temporal aspect of the head – **T.B.-3** (*zhongzhu*) 🟦■, **T.B.-5** (*waiguan*) 🟦🟦; ears – especially **T.B.-3** (*zhongzhu*), **T.B.-5** (*waiguan*) 🟦🟦; hypochondrium – **T.B.-5** (*waiguan*) 🟦🟦, **T.B.-6** (*zhigou*); cervical spine – **T.B.-8** (*sanyangluo*); cervical spine and shoulder – **T.B.-5** (*waiguan*) 🟦🟦; shoulder – **T.B.-1** (*guanchong*)

Specific points according to the channel pathway (in numeric order)

- **T.B.-1** (*guanchong*): *jing*-well point (Metal) (→ 8.1.6); distal point for the shoulder (→ 8.2.1)
- **T.B.-2** (*yemen*): *ying*-spring point (Water) (→ 8.1.6)
- **T.B.-3** (*zhongzhu*): *shu*-stream point (Wood) (→ 8.1.6); tonification point; distal point for the temporal region and the ears (→ 8.2.1); local point for the fingers (numbness and pain) (→ 8.2.1)
- **T.B.-4** (*yangchi*): *yuan*-source point (→ 8.1.1); local point for the hand (→ 8.2.1)
- **T.B.-5** (*waiguan*) 🟦🟦: *luo*-connecting point (→ 8.1.2); Opening point (→ 8.1.8) of the *yang wei mai*; distal point for the temporal region, the ears, the cervical spine, the shoulder, the elbow and the hypochondrium (→ 8.2.1); regional point for the hand and fingers (→ 8.2.1)
- **T.B.-6** (*zhigou*): *jing*-river point (Fire) (→ 8.1.6); *ben* point (Five Phases); distal point for the hypochondrium (→ 8.2.1)
- **T.B.-7** (*huizong*): *xi*-cleft point (→ 8.1.3)
- **T.B.-8** (*sanyangliao*): distal point for the cervical spine (→ 8.2.1)
- **T.B.-10** (*tianjing*): *he*-sea point (Earth) (→ 8.1.6); sedation point; local point for the elbow (→ 8.2.1)
- **T.B.-13** (*naohui*): *jiaohui*-meeting point (→ 8.1.10) of the *yang wei mai** and the G.B. channel*; adjacent point for the shoulder (→ 8.2.1)
- **T.B.-14** (*jianliao*) 🟦■: local point for the shoulder (→ 8.2.1)
- **T.B.-15** (*tianliao*) 🟦■: *jiaohui*-meeting point of the *yang wei mai* and the G.B. channel; local point for the shoulder (→ 8.2.1)
- **T.B.-16** (*tianyou*): Window of Heaven point (→ 8.1.12)
- **T.B.-17** (*yifeng*) 🟦■: *jiaohui*-meeting point of the G.B. channel; adjacent point for the temporal region, the ears and the jaw (→ 8.2.1)
- **T.B.-20** (*jiaosun*): *jiaohui*-meeting point of the S.I.* and G.B. channels
- **T.B.-21** (*ermen*) 🟦■: local point for the ears (→ 8.2.1)
- **T.B.-22** (*erheliao*): *jiaohui*-meeting point with the S.I. and G.B. channels
- **T.B.-23** (*sizhukong*) 🟦■: adjacent point for the eyes (→ 8.2.1)

---

*Mentioned only by some authors

Rushing Pass GUANCHONG T.B.-1

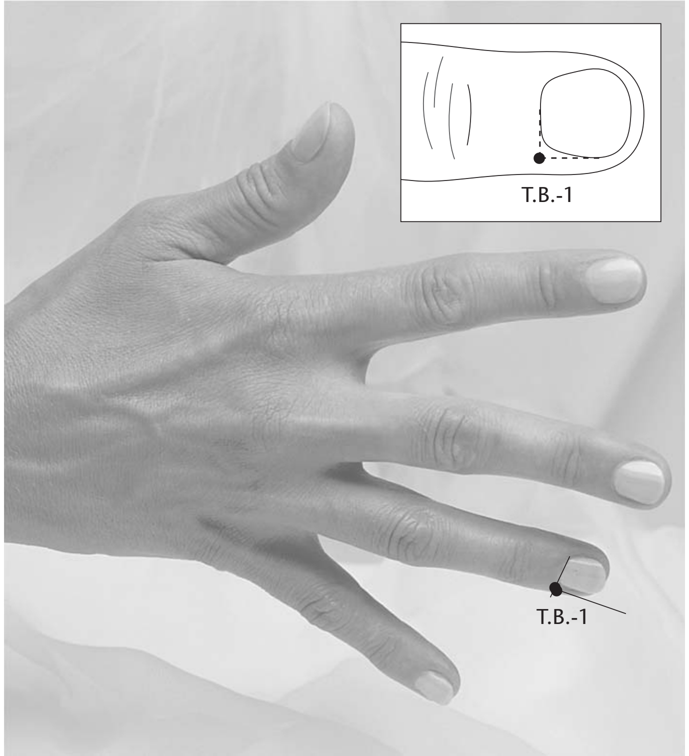

Location

On the ring finger, 0.1 cun from the ulnar corner of the nail.

How to find

This point is located at the junction of two tangents bordering the ulnar and proximal margins of the nail, approximately 0.1 cun from the corner of the actual nail.

Needling

Approximately 0.2 cun vertically or obliquely in a proximal direction or prick to bleed. Caution: Avoid needling too close to the nail; painful point.

Actions/Indications

- Clears Heat from the Upper Burner
- Benefits the ears and tongue
- Opens the channel and alleviates pain

Special features

Jing-well point, Metal point, entry point. Especially indicated for acute disorders.

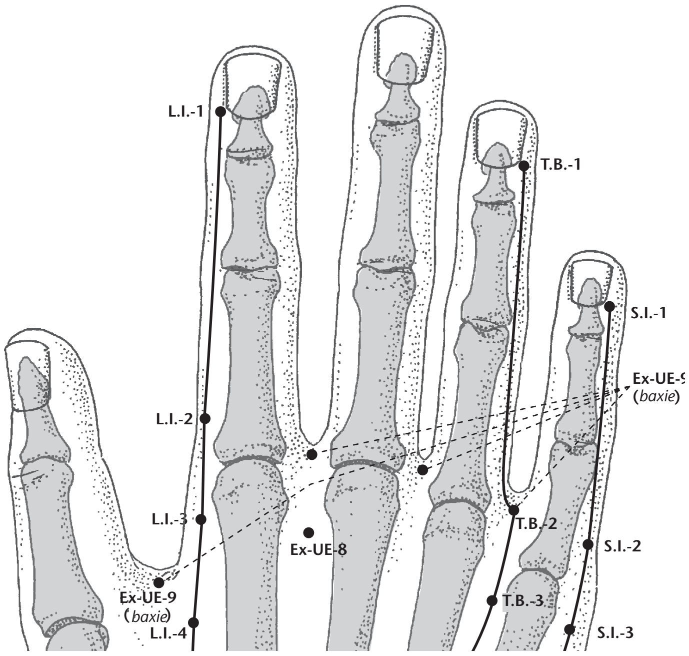

T.B.-2 *Fluid Gate* YEMEN

Location

Between the little finger and ring finger, proximal to the margin of the web.

How to find

This point is best located when making a loose fist. Locate the web between the little finger and ring finger (4th and 5th fingers) and locate **T.B.-2** proximal to its margin. **T.B.-2** is part of → **Ex-UE-9** (*baxie*: proximal to the webs between the fingers). Located in a comparable position on the foot is → **G.B.-43** (on the web between the 4th and 5th toes), which is also part of → **Ex-LE-10** (*bafeng*).

Needling

Vertically up to 0.5 cun

Actions/Indications

- Clears Heat from the Upper Burner
- Benefits the ears and calms the *shen*
- Opens the channel and alleviates pain

Special features

Ying-spring point, Water point. Especially indicated for psycho-neurological disorders caused by Heat and local channel problems.

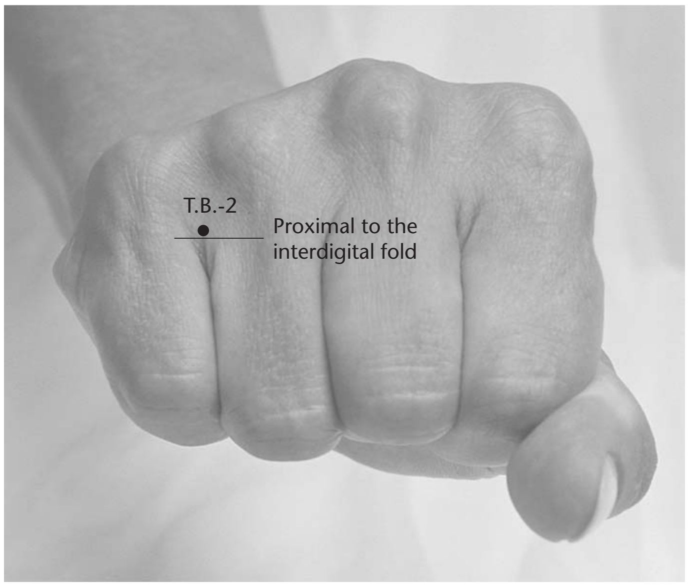
T.B.-2
Proximal to the interdigital fold

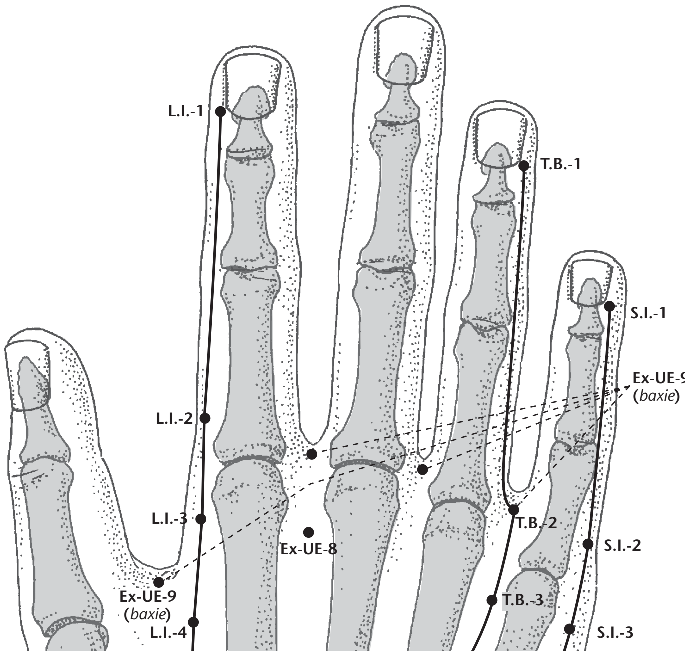
L.I.-1
T.B.-1
S.I.-1
Ex-UE-
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3

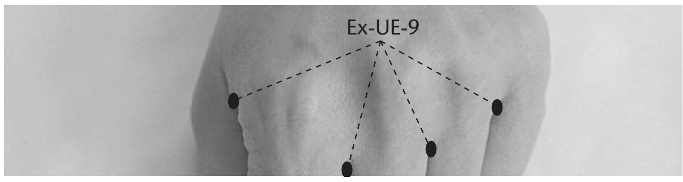

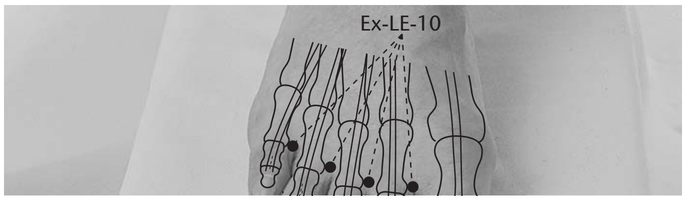

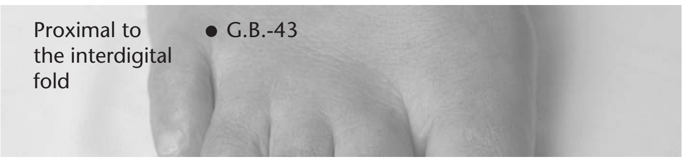
Proximal to the interdigital fold
G.B.-43

Central Islet ZHONGZHU T.B.-3

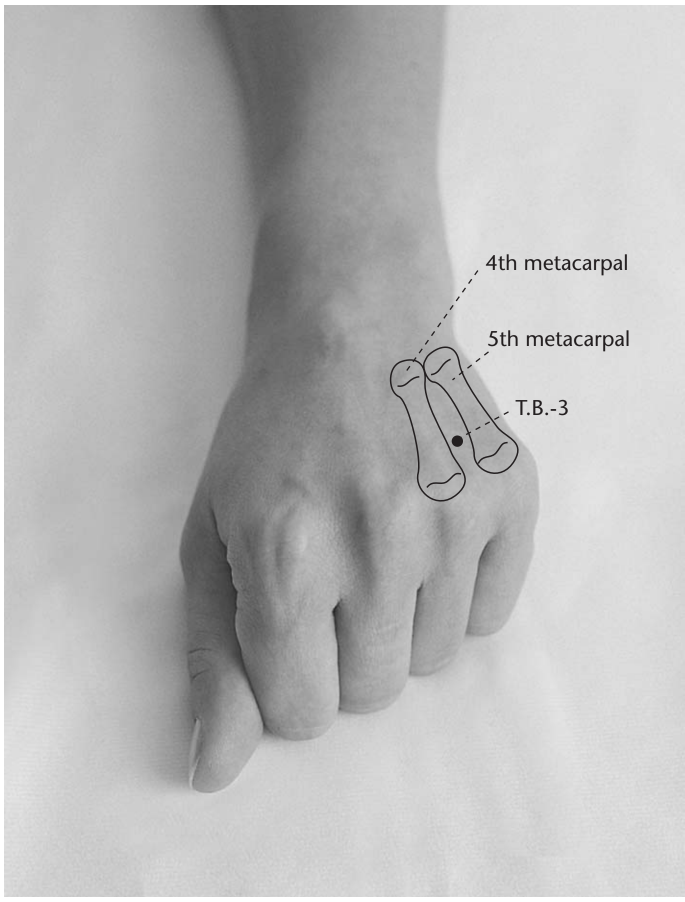

Location

On the dorsum of the hand, in a depression between the 4th and 5th metacarpal bones, proximal to the metacarpophalangeal joints. **T.B.-3** is located at the junction between the heads and shafts of the two metacarpal bones.

How to find

The hand should be relaxed or a loose fist should be made. Palpate from the metacarpophalangeal joints along the groove between the 4th and 5th metacarpal bones towards proximal. **T.B.-3** is located at the widest/deepest point of the groove, slightly distal to the carpometacarpal joints.

Located in comparable positions are → **Ex-UE-8** (*wailaogong*: between the 2nd and 3rd metacarpal bones), → **S.I.-3** (on the ulnar aspect of the 5th metacarpal bone) and → **L.I.-3** (on the radial aspect of the 2nd metacarpal bone).

Needling

0.5–1 cun vertically or obliquely in a proximal direction.

Actions/Indications

- Benefits the ears, clears Heat (and drains Fire), clears the head and the eyes
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

Shu-stream point, Wood point, tonification point. One of the most important distal points for any disorder of the ears.

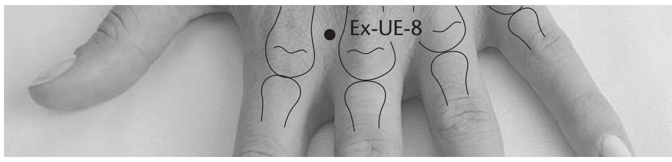

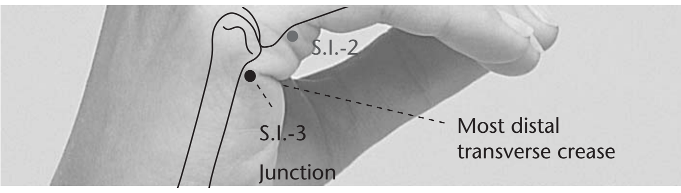

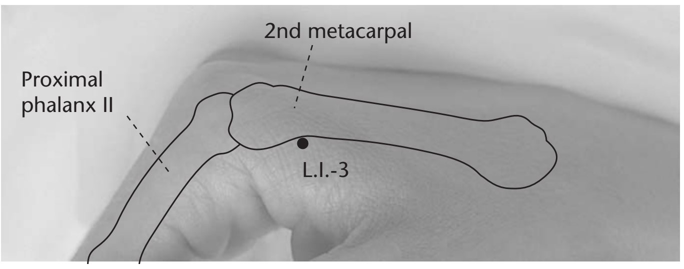

T.B.-4 *Yang Pool* YANGCHI

Location

On the dorsum of the wrist (above the wrist joint space, ‘dorsal wrist crease’), in the gap between the tendons of the extensor digitorum (on the ulnar aspect) and the extensor digiti minimi muscles (on the radial aspect).

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient’s wrist in a relaxed way, the joint space becomes well defined. **T.B.-4** is located slightly laterally to the midpoint of the joint space, in the depression between the tendons of the extensor digiti minimi muscle (continuing to the little finger) on its radial aspect and the tendon of the extensor digitorum muscle on its ulnar aspect. The tendons of the extensor digitorum muscle will become more pronounced by moving the three middle fingers.

Located on the same level but radial to the tendons of the extensor digitorum muscle is → **Ex-UE-3** (*zhongquan*).

Needling

0.3–0.5 cun vertically or slightly obliquely in a proximal direction or transversely (subcutaneously) deep to the tendons towards the radial aspect of the wrist.

Actions/Indications

- Clears Heat, opens the channel, relaxes the tendons. In Japanese acupuncture tradition, **T.B.-4**, in its function as *yuan*-source point, is used for deficiency syndromes.

Special features

Yuan-source point

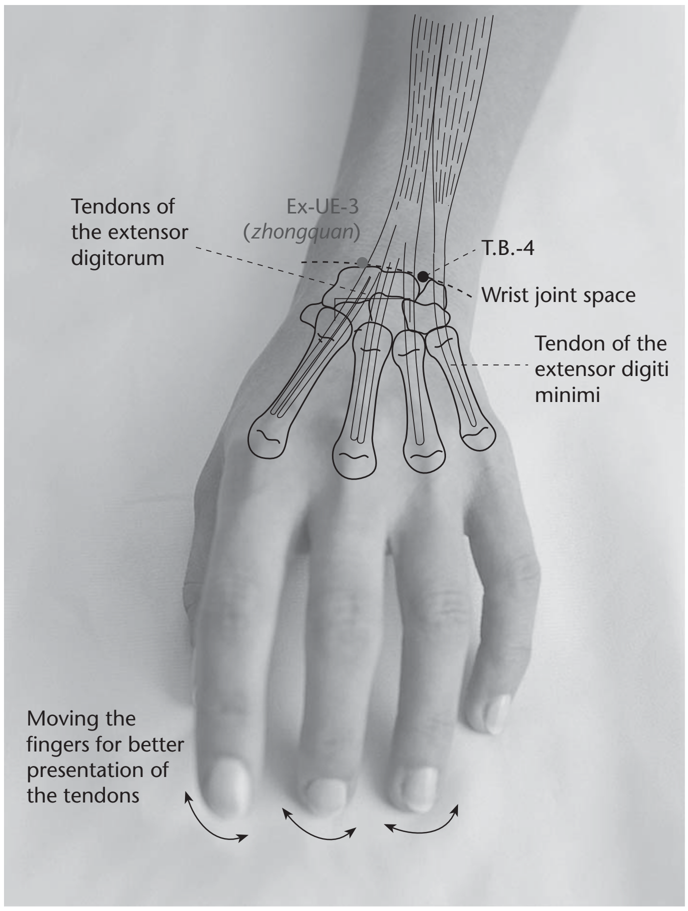
Tendons of the extensor digitorum
Ex-UE-3 (*zhongquan*)
T.B.-4
Wrist joint space
Tendon of the extensor digiti minimi
Moving the fingers for better presentation of the tendons

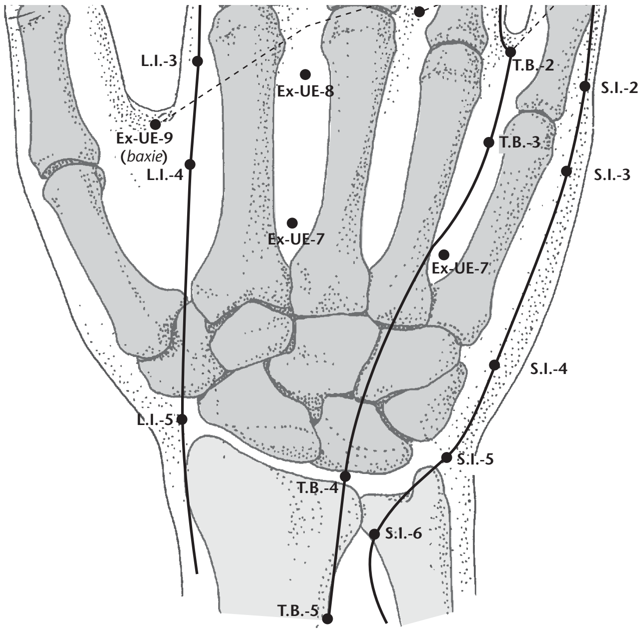
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9 (*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

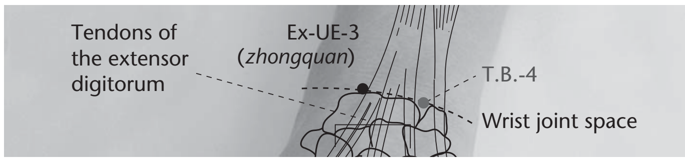
Tendons of the extensor digitorum
Ex-UE-3 (*zhongquan*)
T.B.-4
Wrist joint space

Outer Pass WAIGUAN T.B.-5

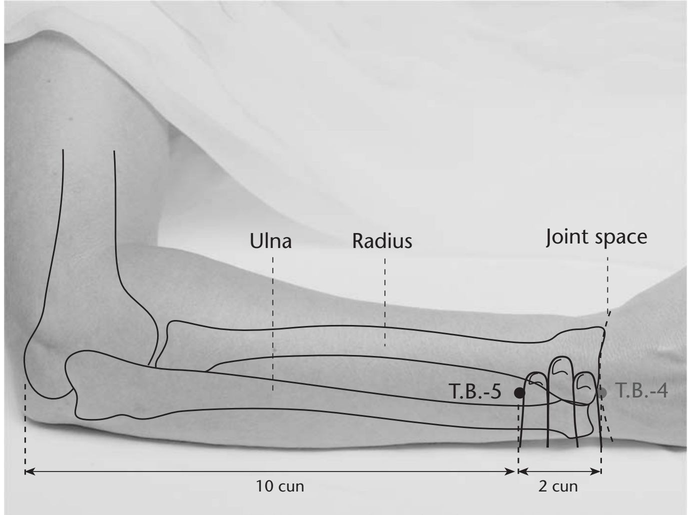

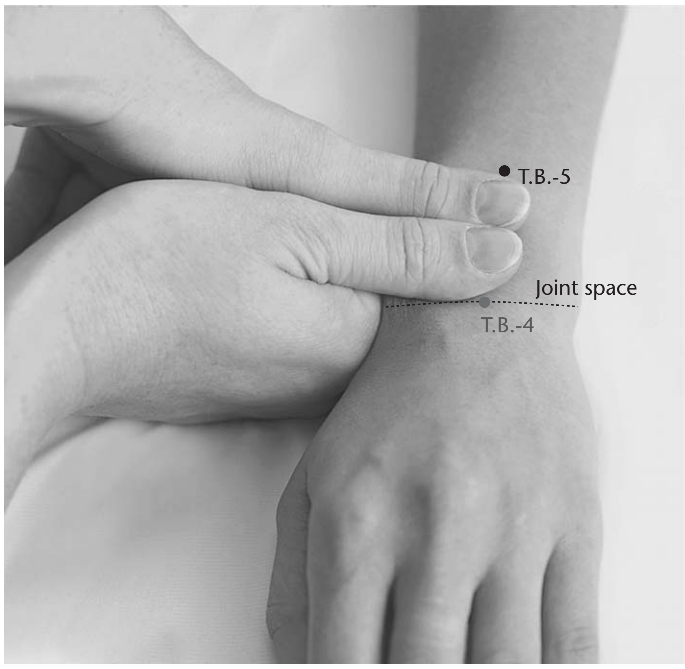

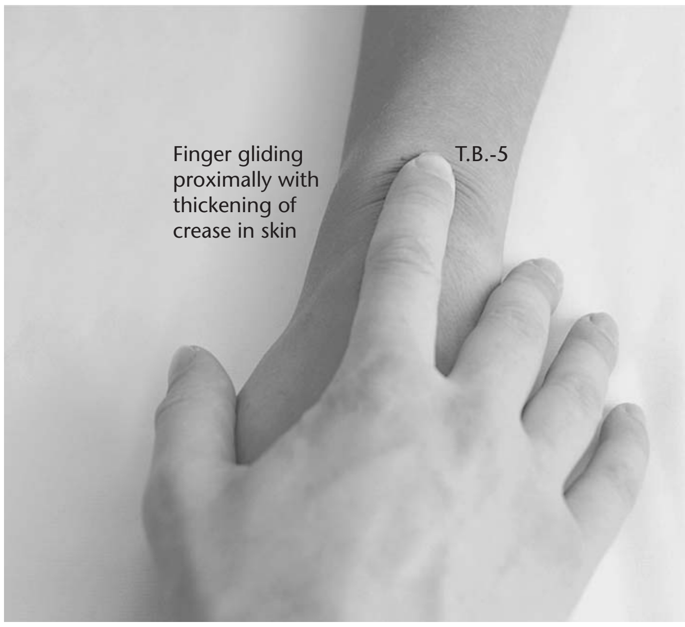

Location

2 cun proximal to the dorsal wrist joint space (‘dorsal wrist crease’), between the radius and the ulna.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (➞ 3.3.3). By moving the patient’s wrist in a relaxed way, the joint space becomes more easily palpable. **T.B.-5** is located 2 cun proximal to the midpoint of the joint space, in the depression midway between the radius and the ulna. **Or quick method:** With the palpating finger, slide with slight pressure from the dorsal aspect of the wrist joint in a proximal direction between the radius and the ulna, until the finger is brought to rest at **T.B.-5** by the increasing skin bulge building up in front of it.

➞ **P-6** is located opposite **T.B.-5,** on the anterior aspect of the forearm.

Needling

Vertically or obliquely 0.5–1.5 cun. Caution: Movement of the hand/arm during needle retention may cause bending of the needle.

Actions/Indications

- Expels Wind and benefits the head and ears, clears Heat, opens and regulates the *yang wei mai*
- Opens the channel and alleviates pain

Special features

Luo-connecting point, opening (master) point of the *yang wei mai*. Major point for expelling Wind-Heat, analgesic point for the upper extremity.

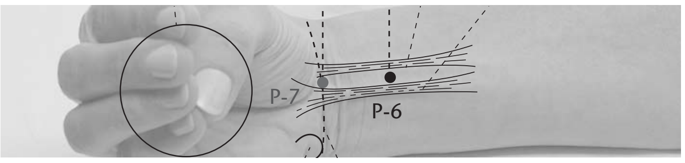

T.B.-6 *Branching Ditch* ZHIGOU

Location

3 cun proximal to the dorsal wrist joint space ('dorsal wrist crease'), in a depression between the radius and the ulna, radial to the tendon of the extensor digitorum communis muscle.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient's wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 3 cun in a proximal direction. At this point, the extensor digitorum communis muscle will often lie midway between the ulna and the radius. **T.B.-6** is located in a depression close to the border of the radius, on the radial side of the muscle.

Located on the same level is → **T.B.-7,** in a depression between the ulna and the extensor digitorum communis muscle. → **P-5** is also located on this level, but on the anterior aspect of the forearm.

Needling

0.5–1.5 cun slightly obliquely towards the ulna or obliquely in a proximal (elbow) or distal (wrist) direction. Or through-needling to **P-5** (*jianshi*). Caution: Movement of the hand/arm during needle retention may cause bending of the needle.

Actions/Indications

- Regulates the Qi, clears Heat in the Triple Burner, benefits the lateral costal region, promotes bowel movements
- Benefits the voice
- Opens the channel

Special features

Jing-river point, Fire point, *ben* point (Five Phases).

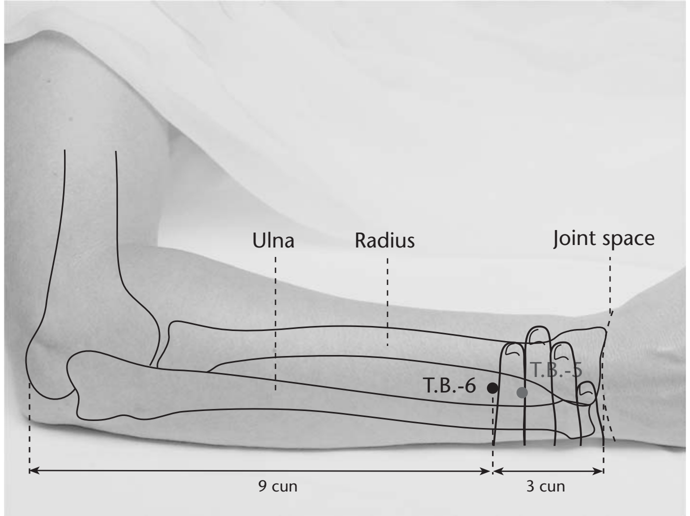

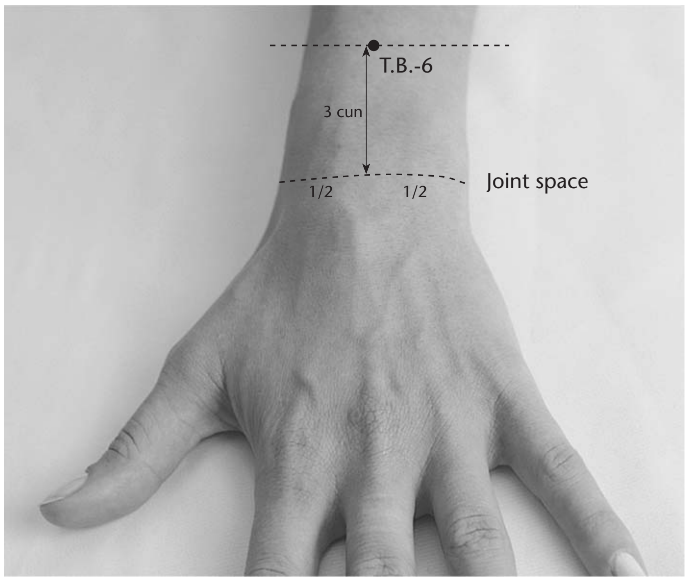

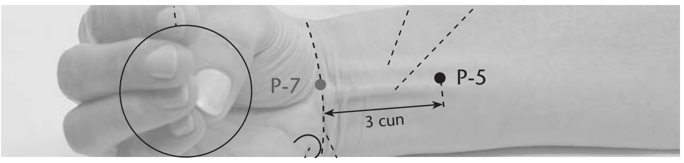

Ch04.10-F10028.qxd 2/23/08 2:27 PM Page 373

Ancestral Meeting HUIZONG T.B.-7

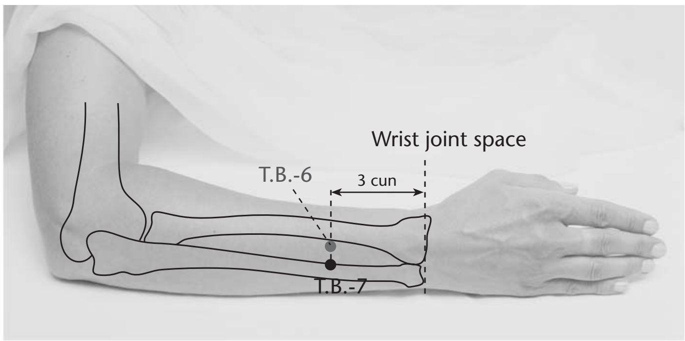

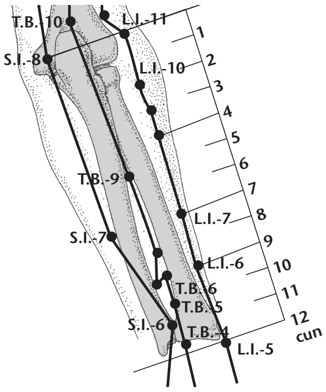

Location

3 cun proximal to the dorsal wrist joint space (‘dorsal wrist crease’) and 0.5 cun ulnar to the centre of the forearm.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (→ 3.3.3). By moving the patient’s wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 3 cun in a proximal direction and locate **T.B.-7** in a depression between the ulna and the extensor digitorum communis muscle. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the elbow crease and the wrist joint space respectively (this distance corresponds to 12 cun). Divide this distance into quarters and locate **T.B.-7** one quarter of the distance from the wrist joint and 0.5 cun ulnar to the centre of the forearm (where → **T.B.-6** is located) on the border of the ulna.

Located on the same level is → **T.B.-6** (0.5 cun towards the radius).

Needling

Vertically or obliquely up to 1.5 cun

Actions/Indications

- Opens the channel and alleviates pain
- Benefits the ears

Special features

Xi-cleft point

T.B.-8 *Three Yang Connection* SANYANGLUO

Location

4 cun proximal to the dorsal wrist joint space ('dorsal wrist crease'), between the radius and the ulna, radial to the tendon of the extensor digitorum communis muscle.

How to find

As the location of the dorsal wrist crease varies, the wrist joint space should be used for orientation instead (➞ 3.3.3). By moving the patient's wrist in a relaxed way, the joint space becomes more easily palpable. From the joint space, measure 4 cun in a proximal direction and locate **T.B.-8** in a depression between the ulna and radial to the extensor digitorum communis muscle. **Or:** Spreading hands technique (➞ 2.3.3): Divide the distance between the elbow crease and the wrist joint space (=12 cun) into thirds and locate **T.B.-8** one third of the distance from the wrist joint between the radius and the ulna.

Needling

Vertically or obliquely up to 1.5 cun

Actions/Indications

- Benefits the ears and the voice
- Opens the channel and the *luo*-connecting vessels, alleviates pain

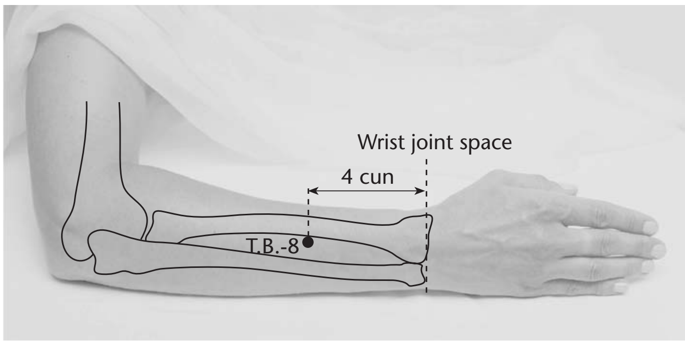

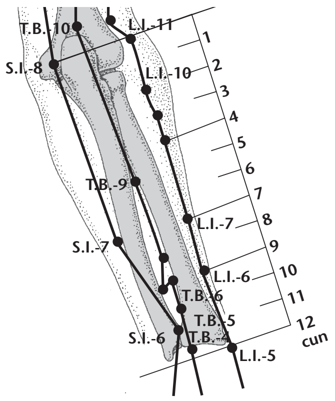

Four Rivers SIDU **T.B.-9**

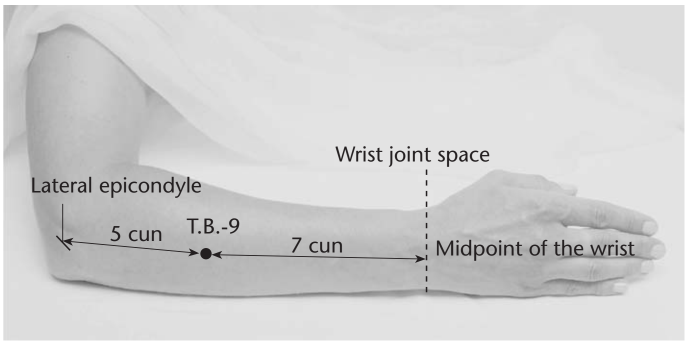

Location

7 cun proximal to the dorsal wrist joint space ('dorsal wrist crease'), between the radius and the ulna.

How to find

From the wrist joint space (→ 3.3.3, more accurate reference point than the variable dorsal wrist crease), measure 7 cun in a proximal direction and locate **T.B.-9** in a depression between the extensor digitorum communis muscle and the extensor carpi ulnaris muscle. Alternatively, measure 5 cun distally from the lateral epicondyle of the humerus. **T.B.-9** is located on a line connecting the midpoint of the dorsal wrist joint space and the lateral epicondyle of the humerus. **Or:** Spreading hands technique (→ 2.3.3): Place the little fingers on the cubital crease and the wrist joint space (→ 2.2) respectively and join the thumbs at the midpoint of this distance (=12 cun). From the midpoint of this line, measure 1 cun in a proximal direction and there locate **T.B.-9** between the radius and the ulna.

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and *luo*-connecting vessels
- Benefits the ears and throat

T.B.-10 *Heavenly Well* TIANJING

Location

On the lateral aspect of the upper arm, with the elbow flexed in a depression approximately 1 cun proximal to the olecranon.

How to find

This point is best located with the patient's elbow flexed to 90°, which will reveal an easily palpable depression 1 cun superior to the olecranon. **T.B.-10** is located in this depression, on the tendon of the triceps brachii muscle.

Needling

Up to 1 cun vertically or obliquely in a proximal direction.

Actions/Indications

- Transforms Phlegm and disperses accumulations
- Regulates and descends the Qi
- Calms the *shen*
- Clears Heat from the channel
- Opens the channel and *luo*-connecting vessels, alleviates pain

Special features

He-sea point, Earth point, sedation point.

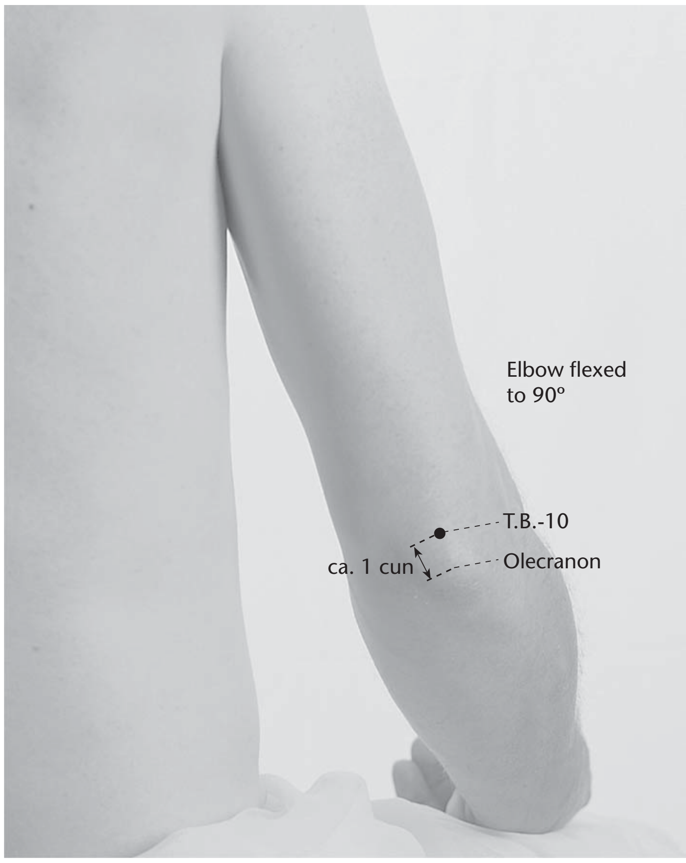
Elbow flexed to 90°
T.B.-10
ca. 1 cun
Olecranon

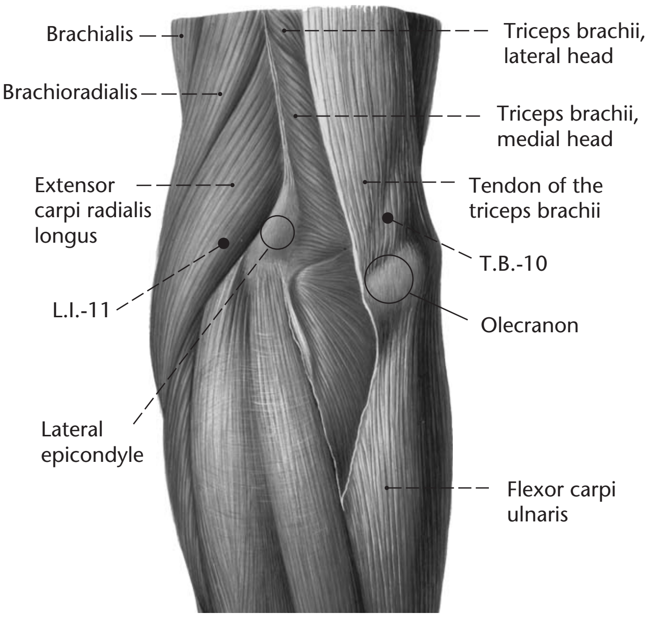
Brachialis
Brachioradialis
Extensor carpi radialis longus
L.I.-11
Lateral epicondyle
Triceps brachii, lateral head
Triceps brachii, medial head
Tendon of the triceps brachii
T.B.-10
Olecranon
Flexor carpi ulnaris

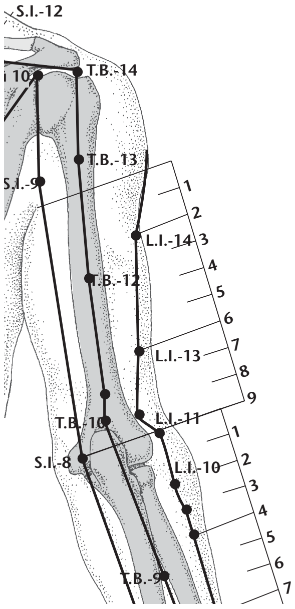
S.I.-12
i 10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
L.I.-13
7
8
9
T.B.-10
L.I.-11
1
S.I.-8
L.I.-10
2
3
4
5
T.B.-9
6
7

Clear Cold Abyss QINGLENGYUAN **T.B.-11**

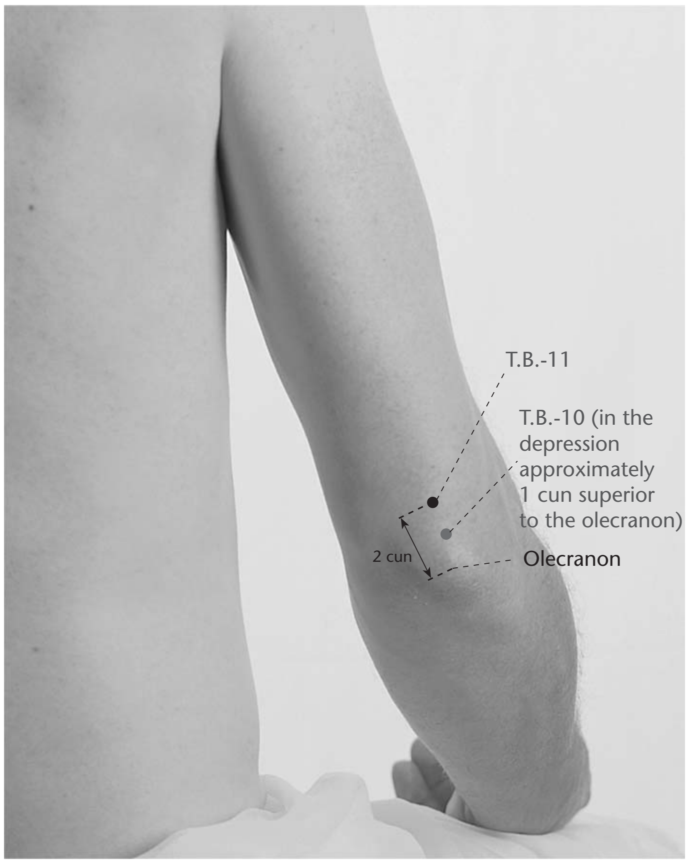
T.B.-11
T.B.-10 (in the depression approximately 1 cun superior to the olecranon)
2 cun
Olecranon

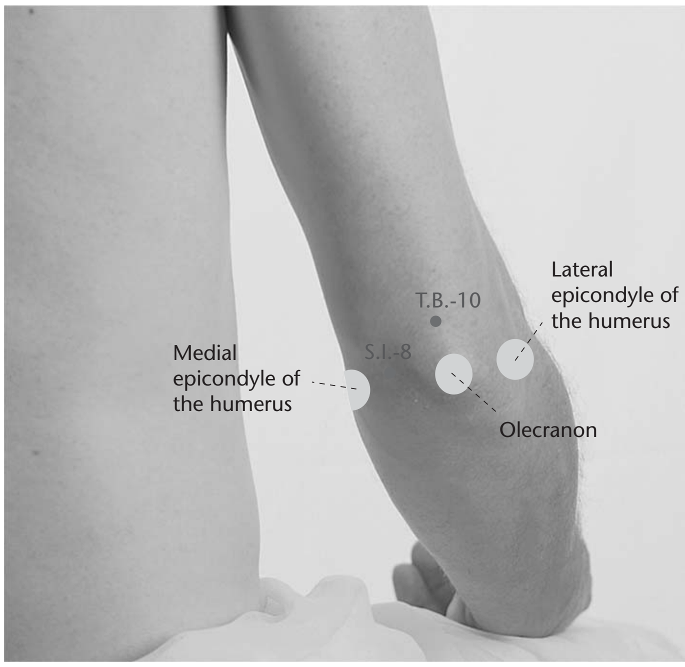
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

Location

On the lateral aspect of the upper arm, with the elbow flexed 1 cun proximal to → **T.B.-10** or 2 cun proximal to the olecranon, on the triceps brachii muscle.

How to find

This point is best located with the patient's elbow flexed to 90°, using → **T.B.-10** as a reference point. This is located in an easily palpable depression 1 cun superior to the olecranon. From → **T.B.-10**, palpate approximately 1 cun in a proximal direction and locate **T.B.-11** in a depression on the triceps brachii muscle.

Needling

Vertically up to 1 cun

Actions/Indications

- Dispels Wind-Damp
- Opens the channel and the *luo*-connecting vessels
- Clears Damp-Heat

T.B.-12 *Dispersing Luo River* XIAOLUO

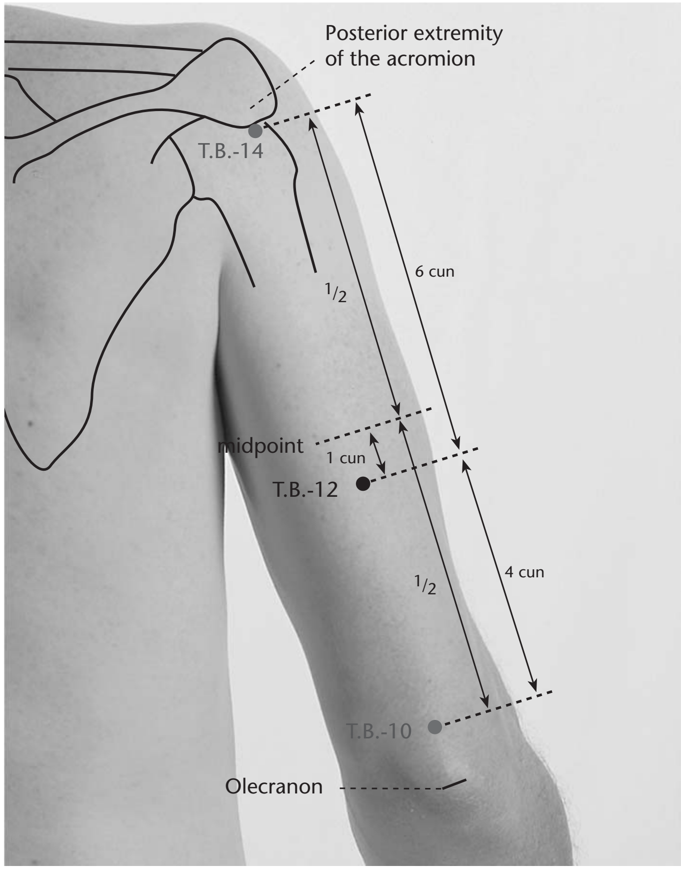
Posterior extremity
of the acromion
T.B.-14
1/2
6 cun
midpoint
1 cun
T.B.-12
1/2
4 cun
T.B.-10
Olecranon

Location

4 cun proximal to ➞ **T.B.-10** (with the elbow flexed in a depression superior to the olecranon) or 5 cun proximal to the olecranon on a line connecting the olecranon and the lateral extremity of the acromion (location of ➞ **T.B.-14**).

How to find

First, locate the acromion (➞ 3.3.1). When the patient's arm is in abduction, two depressions will form below the acromion, in the area of origin of the deltoid muscle. ➞ **T.B.-14** is located in the posterior depression, below the posterior aspect of the lateral extremity of the acromion. Next, on a line connecting ➞ **T.B.-14** and the olecranon, measure 4 cun from ➞ **T.B.-10** (1 cun proximal to the olecranon, in the depression that forms when the elbow is flexed) in a proximal direction and there locate **T.B.-12**. **Or quick method:** Spreading hands technique (➞ 2.3.3): Place the little fingers on ➞ **T.B.-14** and ➞ **T.B.-10** respectively and determine the midpoint of this distance with the thumbs. From the midpoint, measure 1 cun in a distal direction and there locate **T.B.-12. For orientation: T.B.-12** is midway between ➞ **T.B.-11** (1 cun proximal to **T.B.-10** or 2 cun proximal to the olecranon) and ➞ **T.B.-13** (3 cun distal to ➞ **T.B.-14**, on the margin of the deltoid muscle).

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and alleviates pain

Special features

T.B.-12 represents an area extending from 5–7 cun proximal to the tip of the elbow rather than a discrete acupuncture point. Determine the exact location by tenderness.

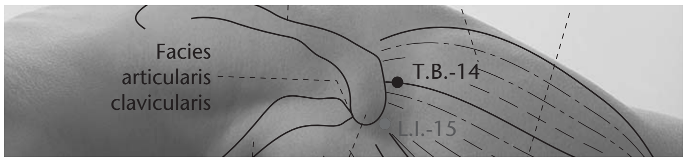
Facies
articularis
clavicularis
T.B.-14
L.I.-15

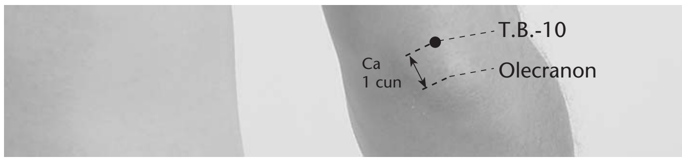
T.B.-10
Ca
1 cun
Olecranon

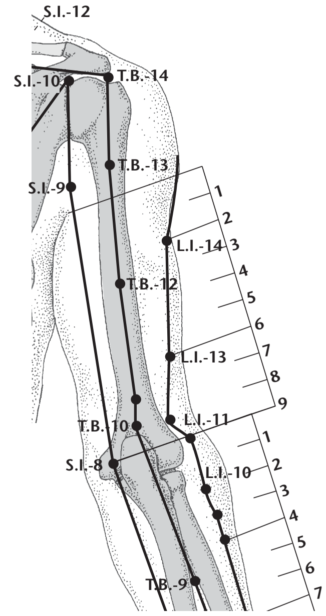
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
L.I.-13
7
8
9
L.I.-11
T.B.-10
1
S.I.-8
2
L.I.-10
3
4
5
6
T.B.-9
7

Upper Arm Meeting NAOHUI T.B.-13

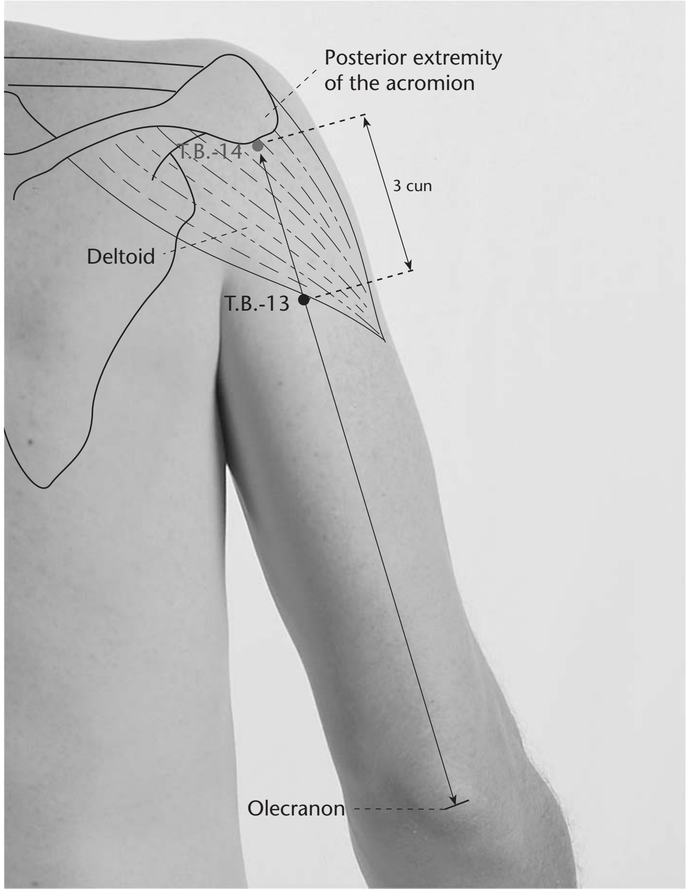
Posterior extremity of the acromion
T.B.-14
3 cun
Deltoid
T.B.-13
Olecranon

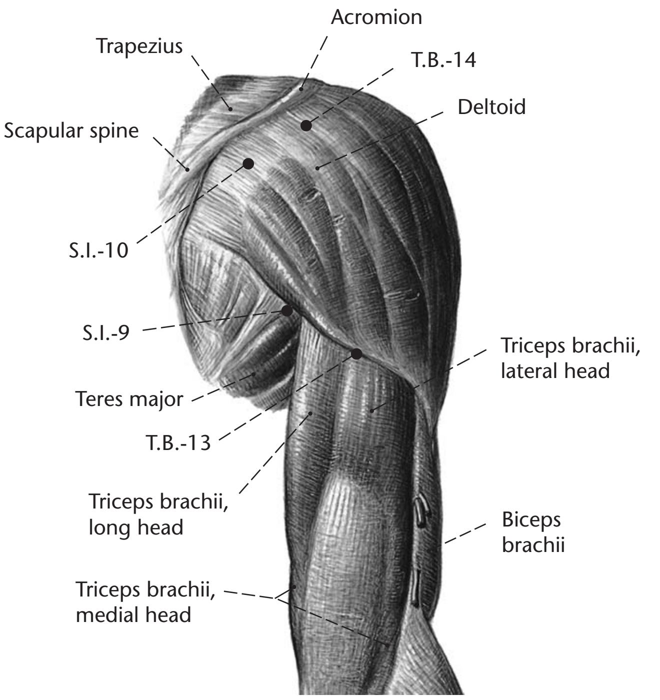
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

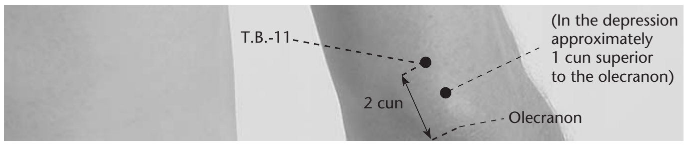
T.B.-11
(In the depression approximately 1 cun superior to the olecranon)
2 cun
Olecranon

Location

3 cun distal to the lateral extremity of the acromion (location of → **T.B.-14**), on a line connecting → **T.B.-14** and the olecranon, at the junction of this line with the margin of the deltoid muscle.

How to find

First, locate the acromion (→ 3.3.1). When the patient's arm is in abduction, two depressions will form inferior to the acromion, in the area of origin of the deltoid muscle. → **T.B.-14** is located in the posterior depression, below the dorsal aspect of the lateral extremity of the acromion. Next, locate **T.B.-13** 3 cun distal to → **T.B.-14**, on a line connecting → **T.B.-14** and the olecranon. **T.B.-13** is located at the junction of this line with the margin of the deltoid muscle, approximately on the level of the end of the posterior axillary fold.

Needling

Vertically or obliquely up to 2 cun

Actions/Indications

- Opens the channel and the *luo*-connecting vessels, alleviates pain
- Regulates the Qi flow and transforms Phlegm

Special features

According to some authors, meeting point with the *yang wei mai* and the G.B. channel

T.B.-14 *Shoulder Crevice* JIANLIAO

Location

Inferior to the lateral extremity of the acromion between the acromial and spinal portions of the deltoid muscle or, with the arm abducted, in the posterior of the two depressions on the shoulder joint.

How to find

Ask the patient to abduct their arm to a 90° angle, which will reveal two depressions anterior and posterior to the acromion. **T.B.-14** is located in the posterior depression, directly below the dorsal aspect of the lateral extremity of the acromion, between the acromial and spinal fibres of the deltoid muscle. The dorsal aspect of the lateral extremity of the acromion becomes more easily palpable when the upper arm is gently pulled (extension of the upper arm), allowing the head of the humerus to glide anteriorly.

→ **L.I.-15** is located in the depression anterior to **T.B.-14.**

Needling

With the arm adducted, 0.5–1 cun perpendicularly towards the centre of the axilla or 1–2 cun transversally to obliquely towards the elbow.

Actions/Indications

- Disperses Wind and Dampness, benefits the shoulder joint, opens the channel and the *luo*-connecting vessels, alleviates pain

Special features

Important local point for shoulder pain, especially for the dorsal aspect of the shouler

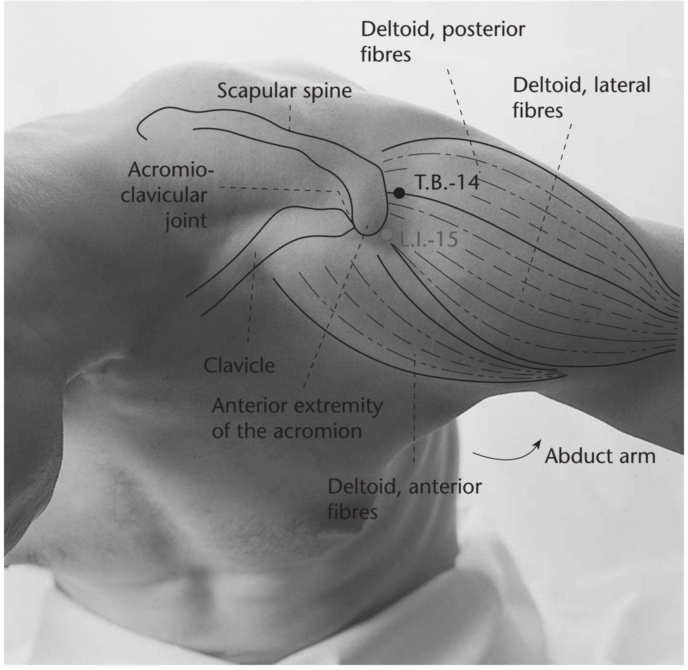
Deltoid, posterior fibres
Scapular spine
Deltoid, lateral fibres
Acromio-clavicular joint
T.B.-14
L.I.-15
Clavicle
Anterior extremity of the acromion
Abduct arm
Deltoid, anterior fibres

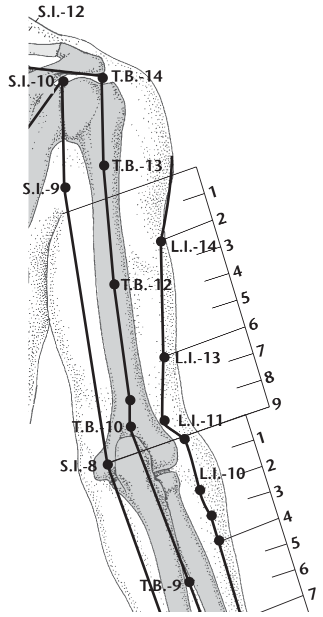
S.I.-12
S.I.-10
T.B.-14
T.B.-13
S.I.-9
1
2
L.I.-14
3
4
T.B.-12
5
6
L.I.-13
7
8
9
L.I.-11
T.B.-10
S.I.-8
1
L.I.-10
2
3
4
5
6
T.B.-9
7

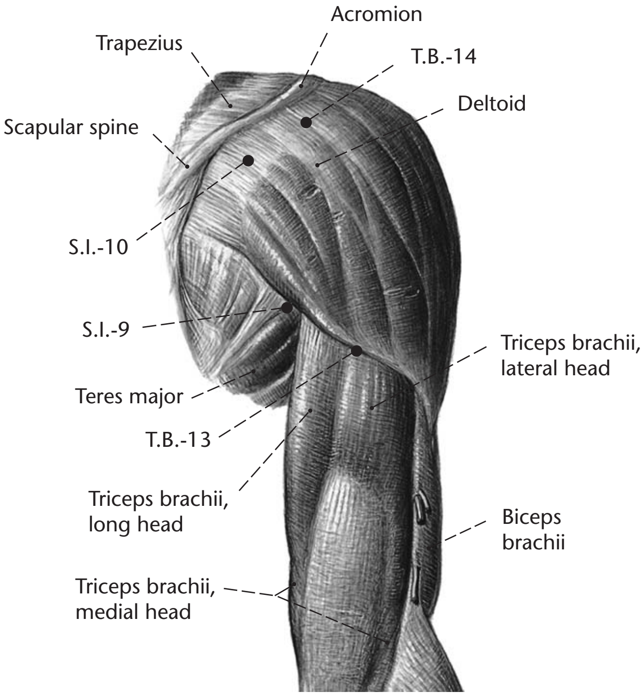
Acromion
Trapezius
T.B.-14
Deltoid
Scapular spine
S.I.-10
S.I.-9
Triceps brachii, lateral head
Teres major
T.B.-13
Triceps brachii, long head
Biceps brachii
Triceps brachii, medial head

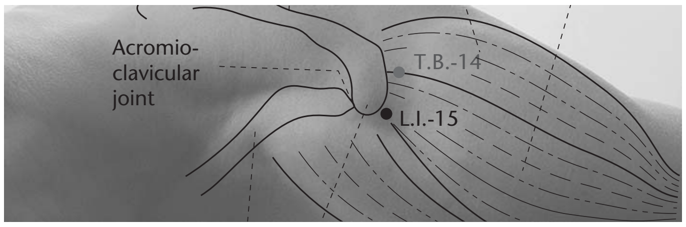
Acromio-clavicular joint
T.B.-14
L.I.-15

Heavenly Crevice TIANLIAO T.B.-15

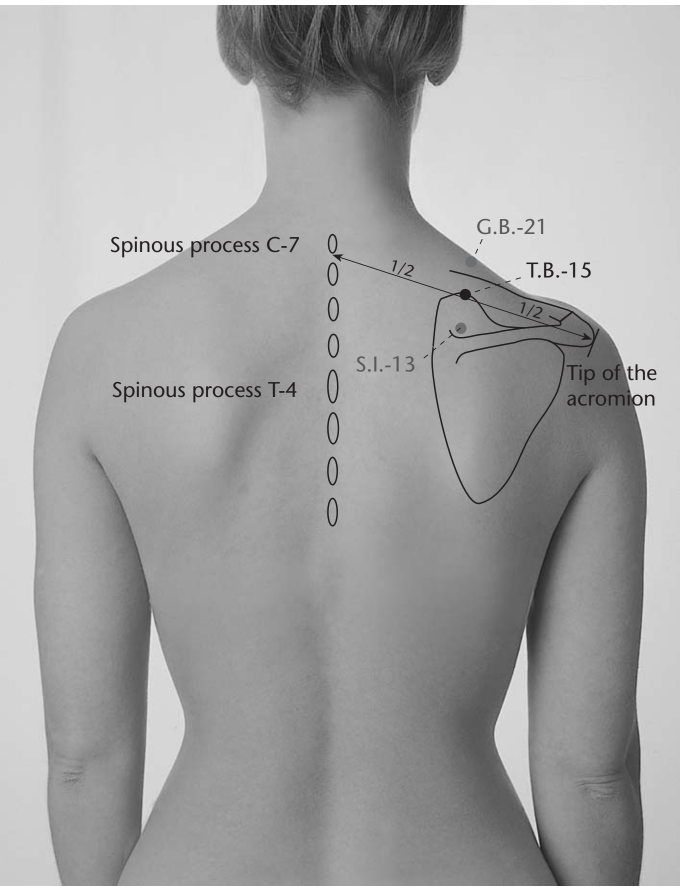
G.B.-21
Spinous process C-7
1/2
T.B.-15
1/2
S.I.-13
Tip of the acromion
Spinous process T-4

Location

At the midpoint of an imaginary line between the spinous process of the 7th cervical vertebra (C7) and the tip of the acromion (lateral extremity of the acromion).

How to find

First, locate C7 (→ 3.4.1) and the acromion (→ 3.3.1), which represents the lateral extension of the scapular spine (→ 3.3.1) and, at its most lateral end, forms a flat surface superior to the head of the humerus (acromial angle). Using, for example, the spreading hands technique (→ 2.3.3), determine the midpoint between C7 and the acromial angle. Located at the highest point of the shoulder is → **G.B.-21. T.B.-15** is located approximately 1 cun inferior to → **G.B.-21**, on the superior angle of the scapula.

→ **S.I.-13** is located inferior to **T.B.-15**, directly superior to the medial end of the scapular spine.

Needling

0.5–1 cun vertically or obliquely towards the neck or shoulder. Caution: Pneumothorax.

Actions/Indications

- Opens the channel and alleviates pain
- Dispels especially Wind-Damp, unbinds the chest

Special features

Meeting point with the *yang wei mai* and the G.B. channel

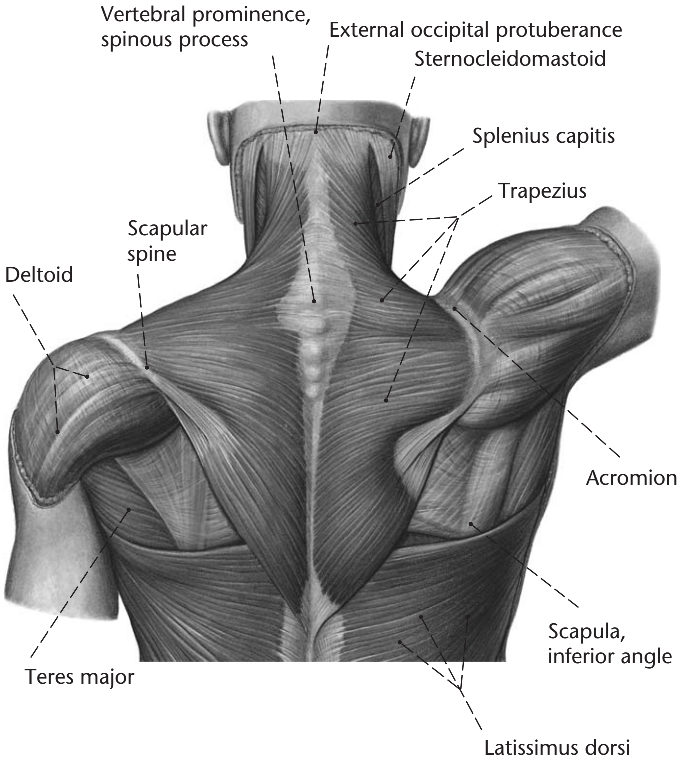
Vertebral prominence, spinous process
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius
Scapular spine
Deltoid
Acromion
Scapula, inferior angle
Teres major
Latissimus dorsi

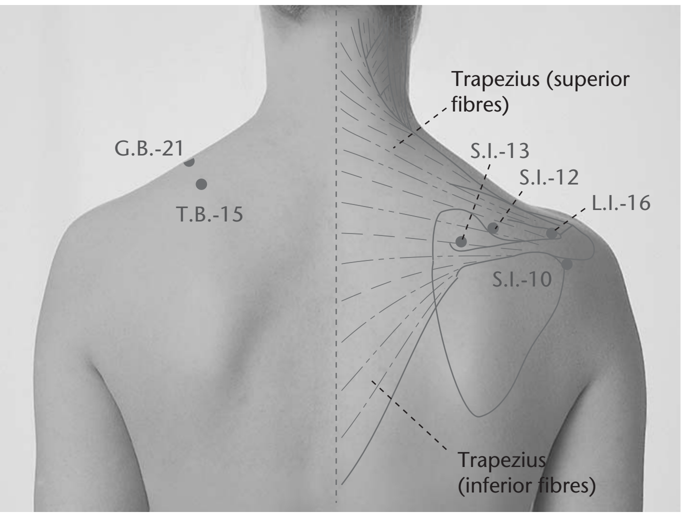
Trapezius (superior fibres)
G.B.-21
S.I.-13
S.I.-12
L.I.-16
T.B.-15
S.I.-10
Trapezius (inferior fibres)

T.B.-16 *Window of Heaven* TIANYOU

Location

At the posterior border of the sternocleidomastoid muscle, directly inferior to the mastoid process, on the level of the mandibular angle.

How to find

For a better definition of the sternocleidomastoid muscle, ask the patient to rotate their head against resistance towards the side to be needled. **T.B.-16** is located on the level of the mandibular angle, on the dorsal aspect of the sternocleidomastoid muscle. This point is located inferior to the posterior border of the mastoid (→ 3.1.4). → **G.B.-12** is located directly posterior and inferior to the mastoid; → **S.I.-17** is also located on the level of the mandibular angle, but on the anterior border of the sternocleidomastoid muscle.

Needling

Vertically 0.5–1 cun. Caution: Carotid artery.

Actions/Indications

- Benefits the head and sensory organs
- Descends the Qi, expels Wind
- Opens the channel and alleviates pain

Special features

Window of Heaven point

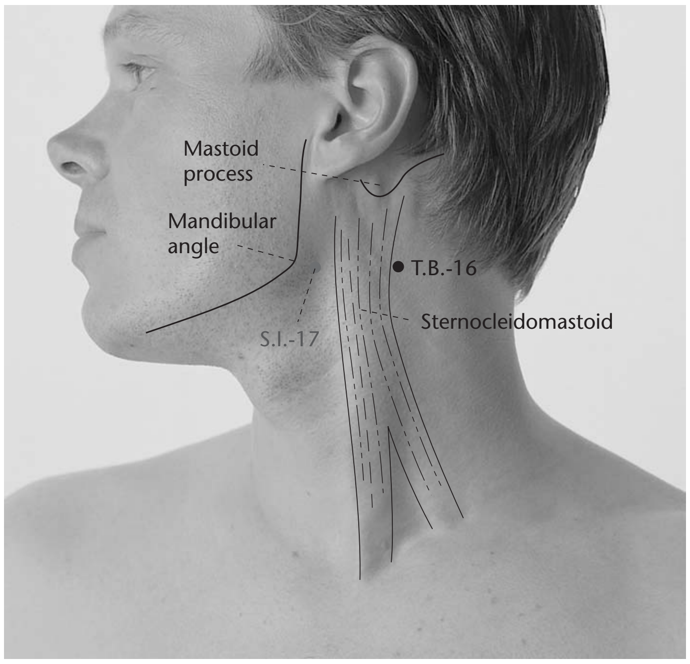
Mastoid process
Mandibular angle
T.B.-16
Sternocleidomastoid
S.I.-17

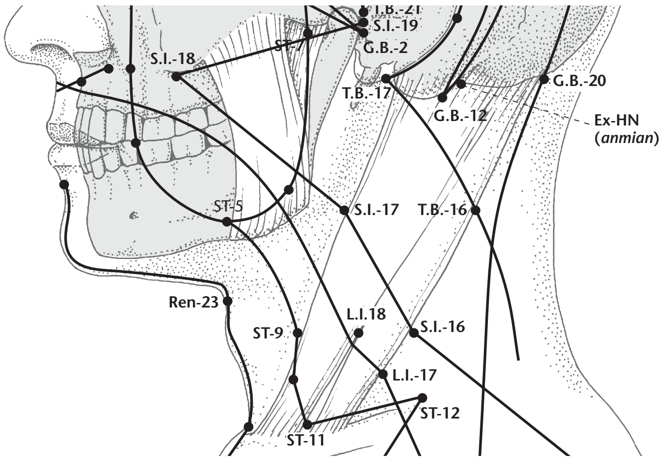
T.B.-21
S.I.-19
G.B.-2
ST-7
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN (*anmian*)
ST-5
S.I.-17
T.B.-16
Ren-23
L.I.18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

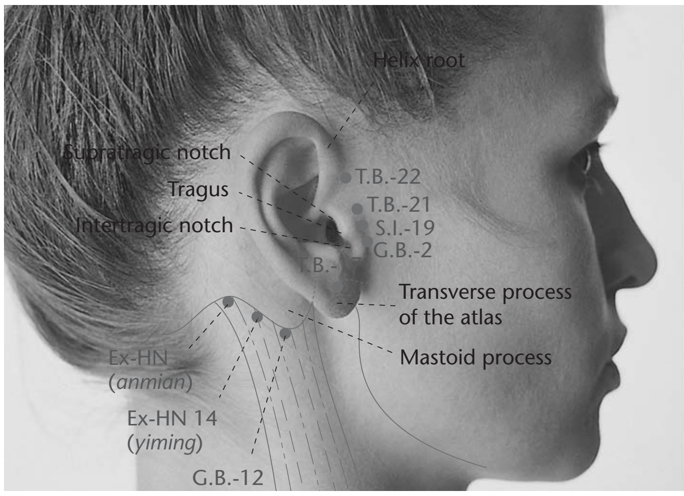
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-
Transverse process of the atlas
Mastoid process
Ex-HN (*anmian*)
Ex-HN 14 (*yiming*)
G.B.-12

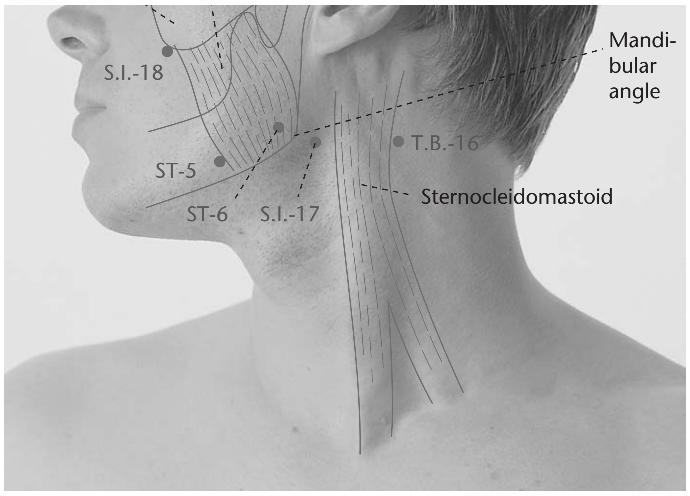
S.I.-18
Mandibular angle
T.B.-16
ST-5
Sternocleidomastoid
ST-6
S.I.-17

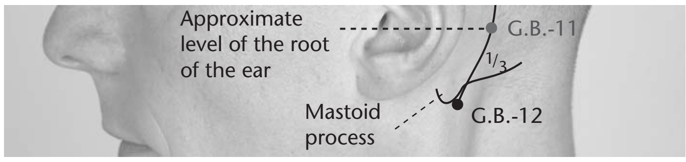
Approximate level of the root of the ear
G.B.-11
1/3
Mastoid process
G.B.-12

Wind Screen YIFENG T.B.-17

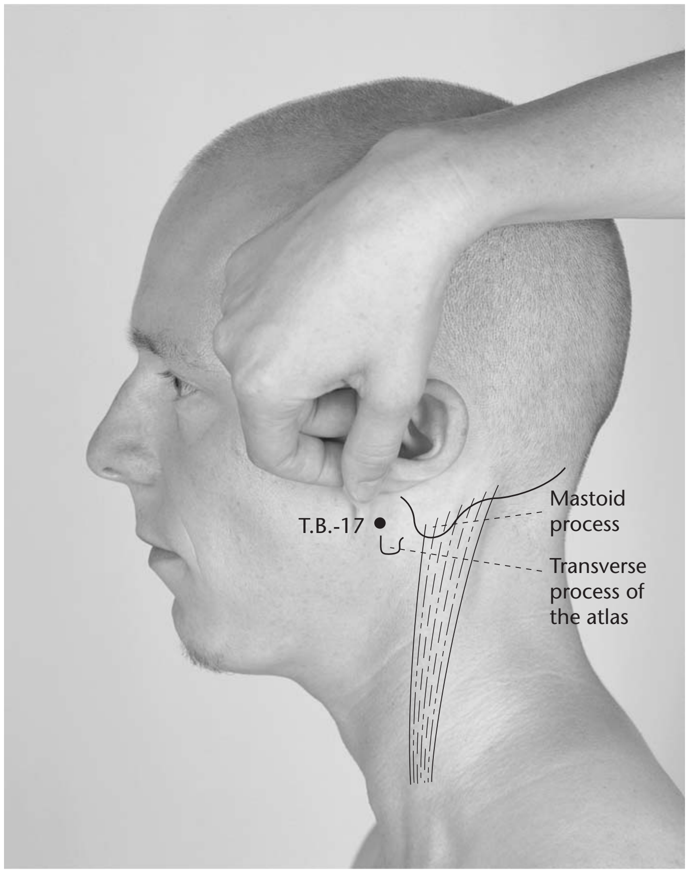
T.B.-17
Mastoid process
Transverse process of the atlas

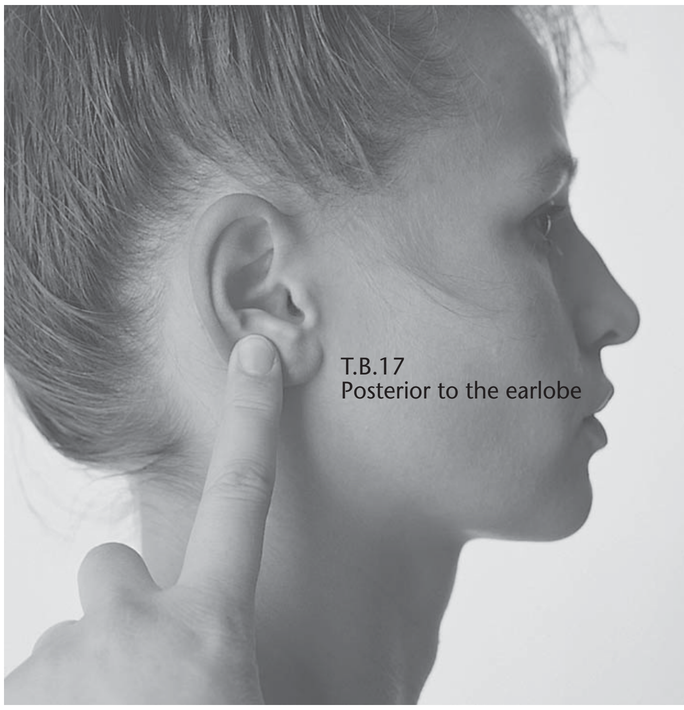
T.B.17
Posterior to the earlobe

Location

With the patient's mouth open, in the depression behind the earlobe, between the mastoid process and the mandible.

How to find

This point should be located with the patient's mouth open. **Quick method:** When pressing onto the ear lobe, it will touch **T.B.-17**. **Or:** Gently fold the earlobe anteriorly. **T.B.-17** is located posterior to the lobe, in a pressure-sensitive depression which can be palpated between the mandible and the mastoid process. The tip of the needle will get close to the transverse process of the 1st cervical vertebra (transverse process of the atlas, → 3.1.4), which, below the earlobe, can be palpated as a deep bony structure often sensitive to pressure.

Needling

This point should be needled with the patient's mouth slightly open. Caution: This point is close to the facial nerve, therefore no deep needling. Superior to the transverse process, the vertebral artery curves round posteriorly, therefore strictly needle 0.5–1 cun in an anterior direction.

Actions/Indications

- Expels (external) Wind, benefits the ears, clears Heat, opens the channel and *luo*-connecting vessels, alleviates pain

Special features

Meeting point with the G.B. channel. Important local point for disorders of the ears and the temporomandibular joint.

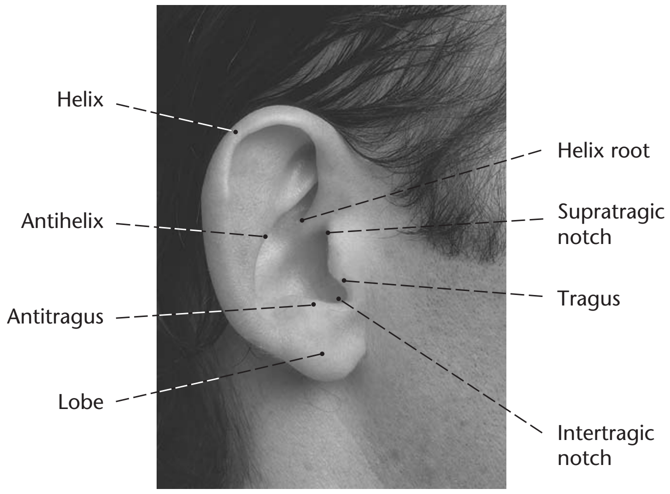
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

T.B.-18 *Spasm Vessel* QIMAI

Location

Posterior to the ear, in a well-defined depression in the centre of the mastoid.

How to find

On an imaginary clock face superimposed over the auricle (12 o'clock = apex of the ear, 6 o'clock = earlobe), you can find an easily palpable depression at 8 o'clock (on the right side) and at 4 o'clock (on the left side). It is located directly behind the margin of the auricle and is the location of **T.B.-18**.

→ **G.B.-11** is located slightly superior to **T.B.-18** and approximately 0.3 cun from the ear; → **G.B.-12** is located inferior to **T.B.-18**, directly posterior to the mastoid.

Needling

0.5 cun transversely (subcutaneously) in an inferior direction or prick to bleed.

Actions/Indications

- Dispels Wind
- Opens the channel, alleviates pain
- Benefits the ears

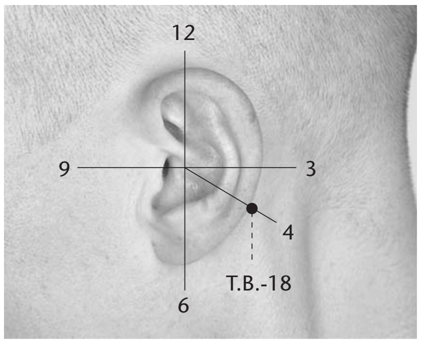
12
9
3
4
T.B.-18
6

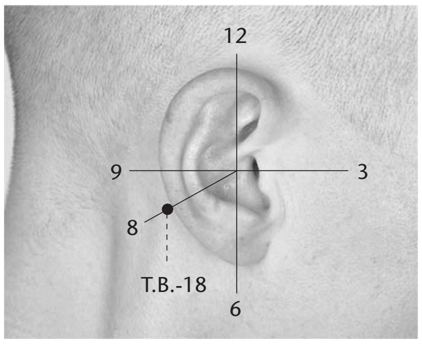
12
9
3
8
T.B.-18
6

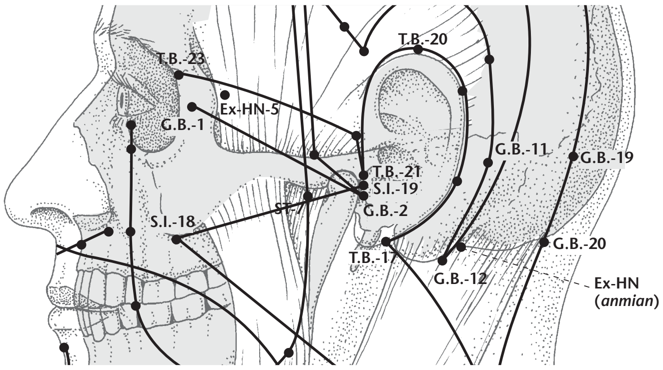
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
ST-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)

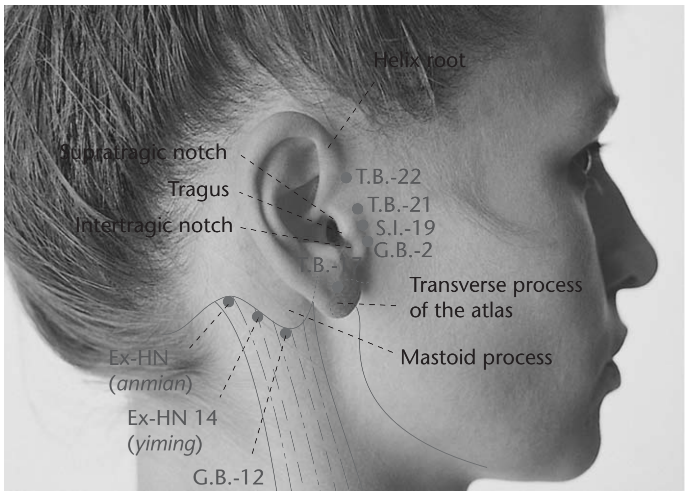
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process
of the atlas
Mastoid process
Ex-HN
(*anmian*)
Ex-HN 14
(*yiming*)
G.B.-12

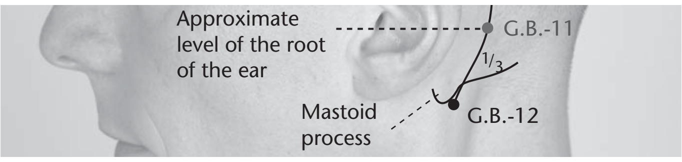
Approximate
level of the root
of the ear
G.B.-11
1/3
Mastoid
process
G.B.-12

Skull's Rest LUXI T.B.-19

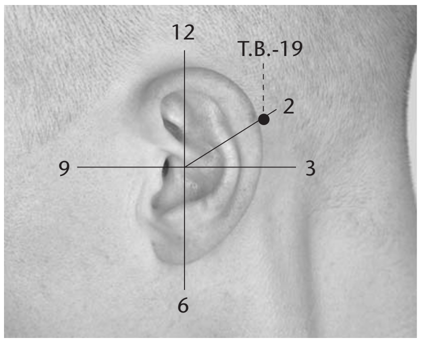

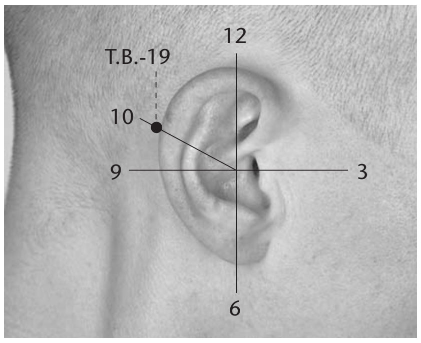

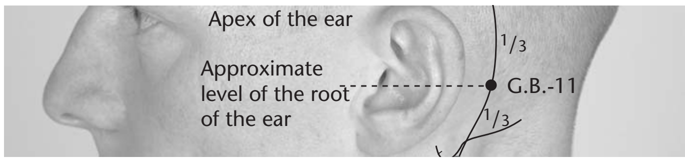

Location

Posterior to the ear, in a well-defined depression superior to the centre of the ear.

How to find

On an imaginary clock face superimposed over the auricle (12 o'clock = apex of the ear, 6 o'clock = earlobe), you can find an easily palpable depression at approximately 10 o'clock (on the right side) and at 2 o'clock (on the left side). It is located directly behind the margin of the auricle and is the location of **T.B.-19**.

→ **G.B.-11** is located slightly inferior to **T.B.-19** and approximately 0.3 cun from the margin of the ear.

Needling

Up to 0.5 cun transversely (subcutaneously) in an inferior direction or prick to bleed.

Actions/Indications

- Dispels Wind
- Opens the channel
- Benefits the ear, clears Heat

Ch04.10-F10028.qxd 2/23/08 2:27 PM Page 386

T.B.-20 *Minute Angle* JIAOSUN

Location

Directly superior to the apex of the ear.

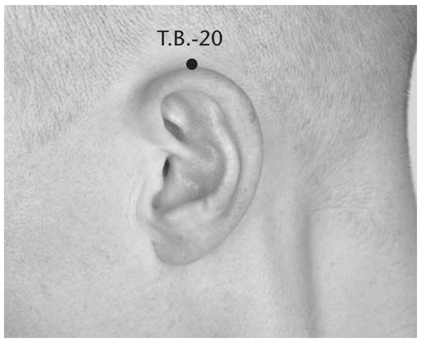

How to find

Directly superior to the apex of the ear. → **G.B.-7** is also located on the apex of the ear, but anterior to it, within the circumauricular hairline. → **G.B.-8** is located 1.5 cun superior to the apex of the ear.

Needling

Up to 1.5 cun transversely (subcutaneously) in the direction of the disorder.

Actions/Indications

- Benefits the ears and the eyes
- Clears Heat, especially from the mouth

Special features

Meeting point with the G.B. channel, also with the S.I. channel according to some authors.

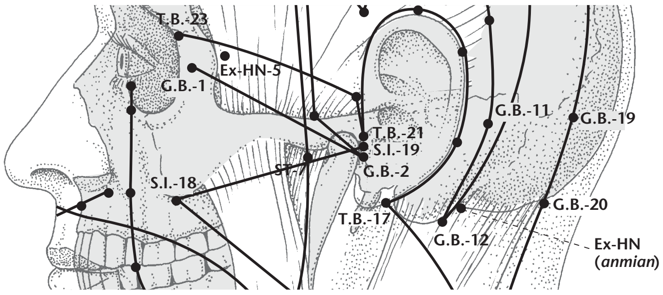
T.B.-23
Ex-HN-5
G.B.-1
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-11
G.B.-19
G.B.-20
G.B.-12
Ex-HN
(anmian)

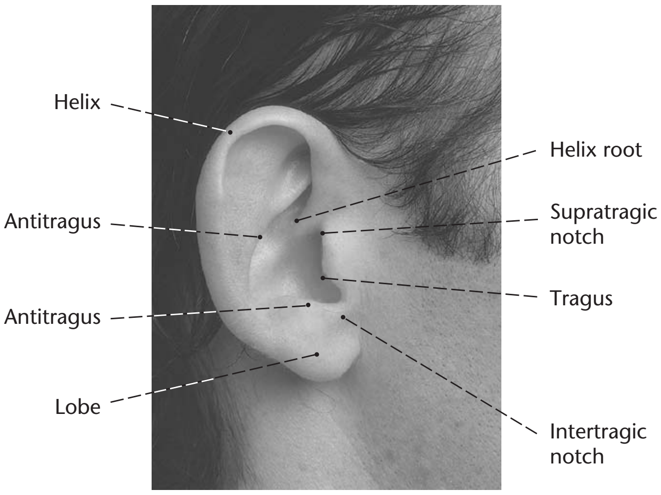
Helix
Antitragus
Antitragus
Lobe
Helix root
Supratragic notch
Tragus
Intertragic notch

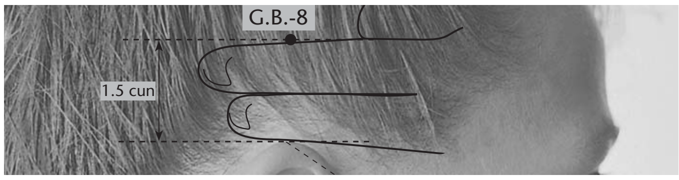

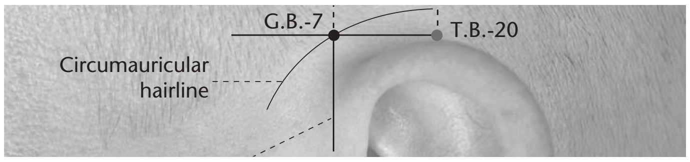
G.B.-7
T.B.-20
Circumauricular hairline

Ear Gate ERMEN T.B.-21

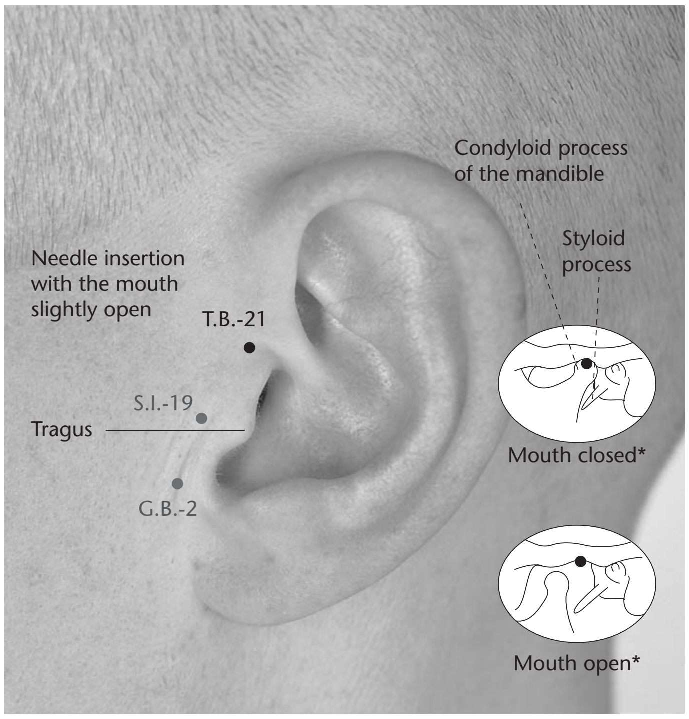
Condyloid process of the mandible
Styloid process
Needle insertion with the mouth slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed*
Mouth open*

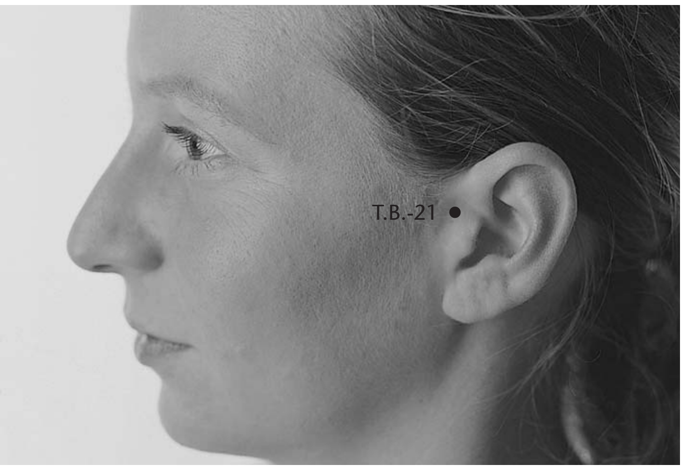

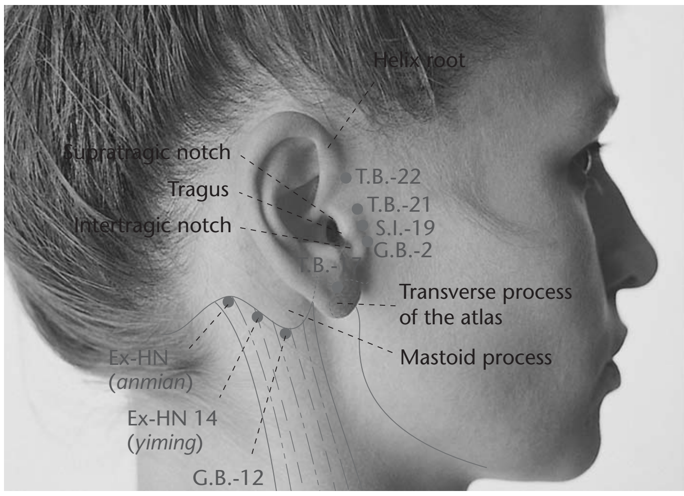
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)
Ex-HN 14 (yiming)
G.B.-12

Location

Anterior to the ear, with the patient's mouth open in the depression on the level of the supratragic notch and slightly superior to the condyloid process of the mandible.

How to find

This point should be both located and needled with the patient's mouth open. This allows the condyloid process of the mandible to slide anteriorly to reveal the depression where **T.B.-21** is located. Locate the sulcus between the auricle and the cheek anterior to the tragus (this can be more or less well defined). Then locate **T.B.-21** on the level of the supratragic notch on the sulcus. If the latter cannot be clearly identified (it becomes more pronounced with increasing age), the depression can be located with an (ear) point locator when the patient's mouth is opened. **T.B.-21** is the most superior of three points located anterior to the ear (**→ S.I.-19** and **→ G.B.-2** are located more distally).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the patient's mouth open (to avoid intra-articular insertion). After insertion, patients can close their mouth again. Caution: Like S.I.-19 and G.B.-2, this point is close to the superficial temporal artery and the auriculotemporal nerve.

Actions/Indications

- Clears Heat, benefits the ears

Special features

Important local point for disorders of the ears

---

* Accoring to Deadman et al 1998.

T.B.-22 *Ear Harmony Crevice* ERHELIAO

Location

In a depression at the border of the circumauricular temporal hairline, anterior to and on the level of the root of the auricle.

How to find

First, locate the anterior aspect of the root of the ear. From there, palpate approximately the breadth of the little finger in the direction of the eye. There, **T.B.-22** is located superior to the zygomatic arch (→ 3.1.2), which forms a bony ridge when sliding inferiorly from **T.B.-22**.
Located inferior to the zygomatic arch is → **T.B.-21**.

Needling

Transversely (subcutaneously) 0.5 cun. Caution: Superficial temporal artery.

Actions/Indications

- Dispels Wind
- Opens the channel and the *luo*-connecting vessels

Special features

Meeting point with the S.I. and G.B. channels, exit point according to some authors.

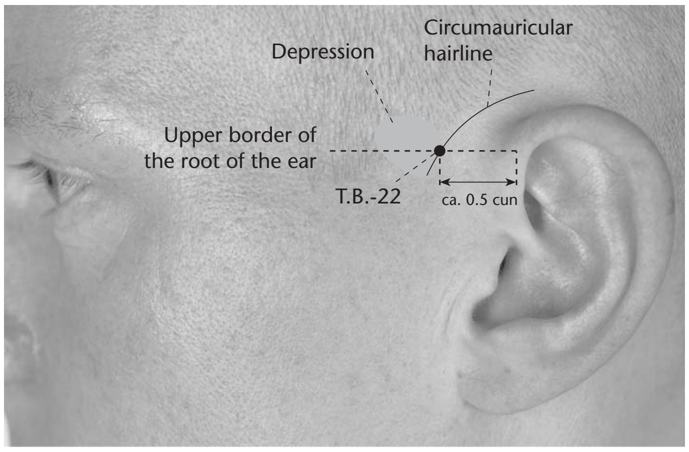
Depression
Circumauricular hairline
Upper border of the root of the ear
T.B.-22
ca. 0.5 cun

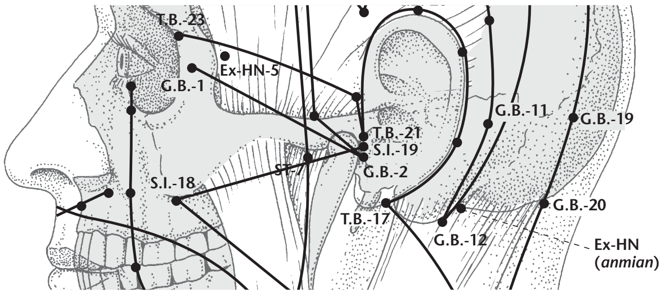
T.B.-23
Ex-HN-5
G.B.-1
T.B.-21
S.I.-19
G.B.-2
S.I.-18
ST-7
T.B.-17
G.B.-12
G.B.-11
G.B.-19
G.B.-20
Ex-HN (anmian)

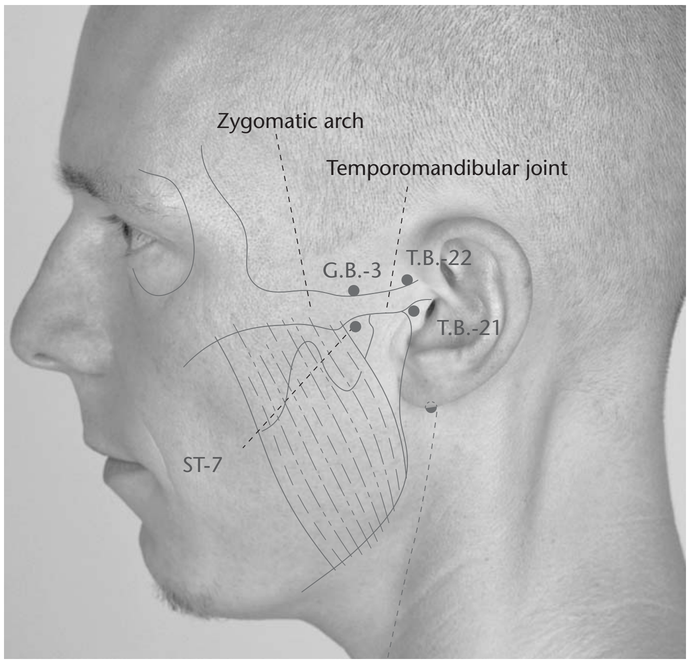
Zygomatic arch
Temporomandibular joint
G.B.-3
T.B.-22
T.B.-21
ST-7

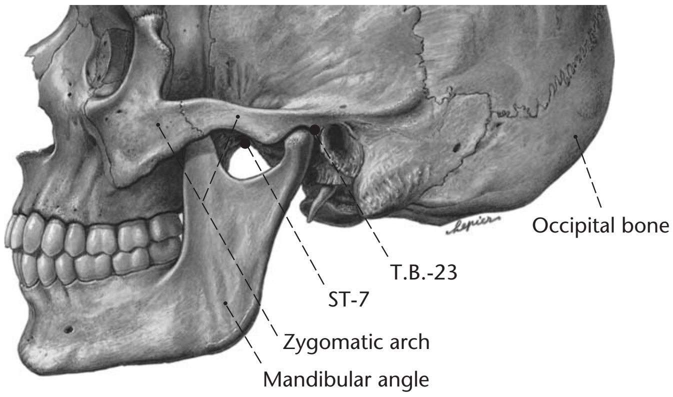
Occipital bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

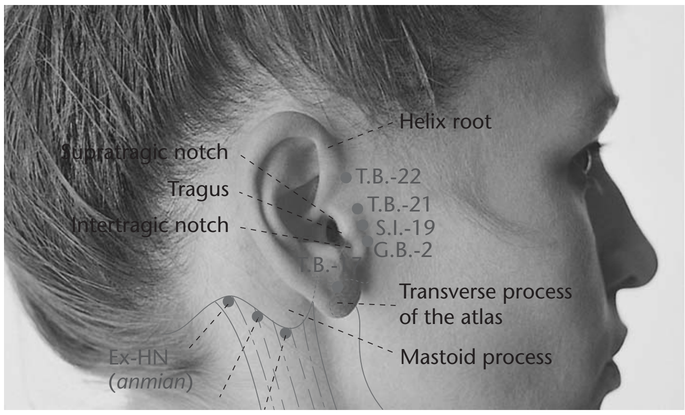
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
Transverse process of the atlas
Mastoid process
Ex-HN (anmian)

Silken Bamboo Hollow SIZHUKONG **T.B.-23**

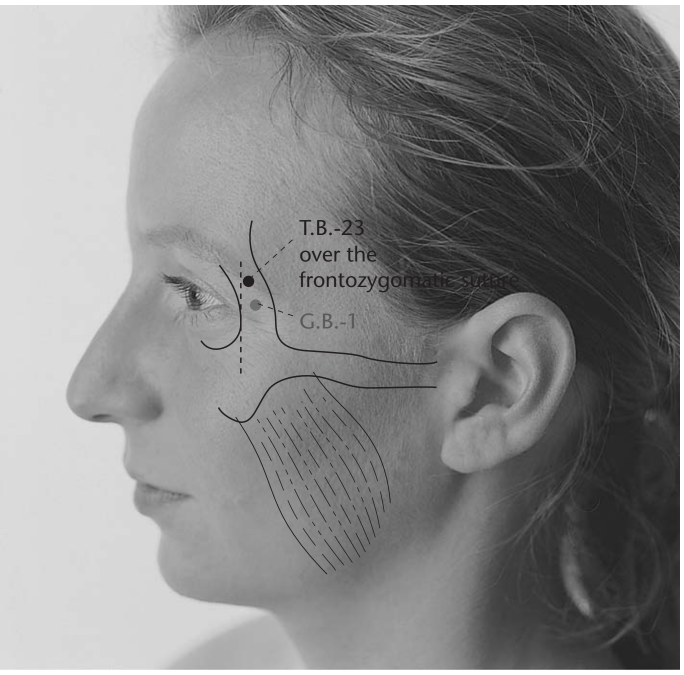
T.B.-23
over the
frontozygomatic suture
G.B.-1

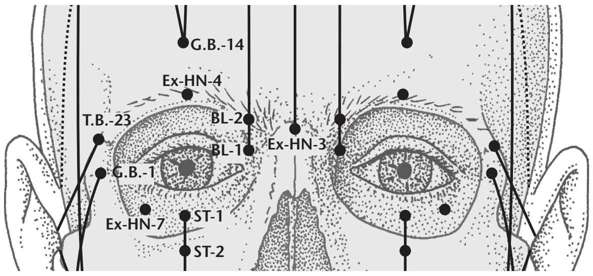
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2

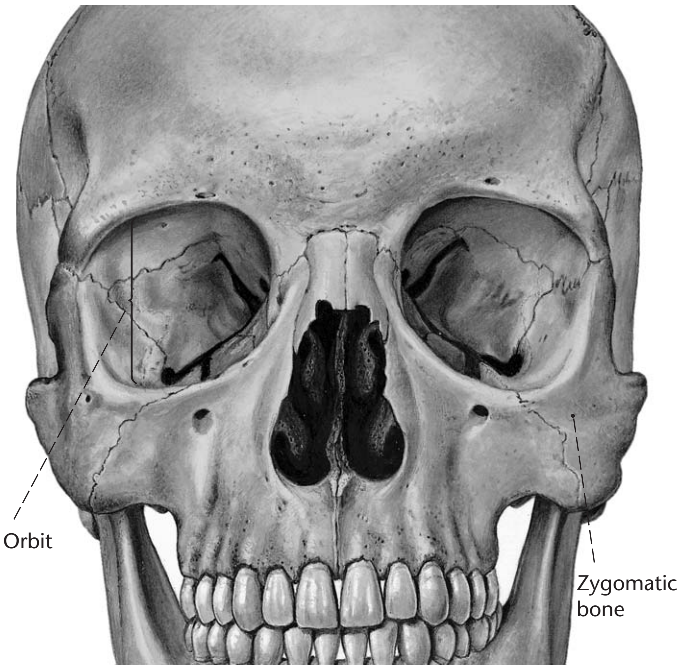
Orbit
Zygomatic
bone

Location

On the lateral end of the eyebrow, in the bony depression of the frontozygomatic suture, between the frontal and zygomatic bones.

How to find

The frontozygomatic suture (between the zygoma and the frontal bone) is generally located at the lateral end of the eyebrow. As the position of the latter can vary considerably, the suture is a more suitable reference point. In order to locate it, palpate from the outer canthus of the eye superiorly along the orbital margin, from its zygomatic to the frontal section, until you can feel a bony depression in the area of the suture. Locate **T.B.-23** in this mainly pressure-sensitive depression.

Needling

0.5–1 cun obliquely or transversely (subcutaneously) in a posterior direction towards **Ex-HN-5** (*taiyang*). In China, this point is also needled along the eyebrow to **Ex-HN-4** (*yuyao*: in the centre of the eyebrow). Moxibustion is contraindicated according to some classic texts.

Actions/Indications

- Benefits the eyes, alleviates pain, eliminates Wind

Special features

Important local point for headaches and disorders of the eyes

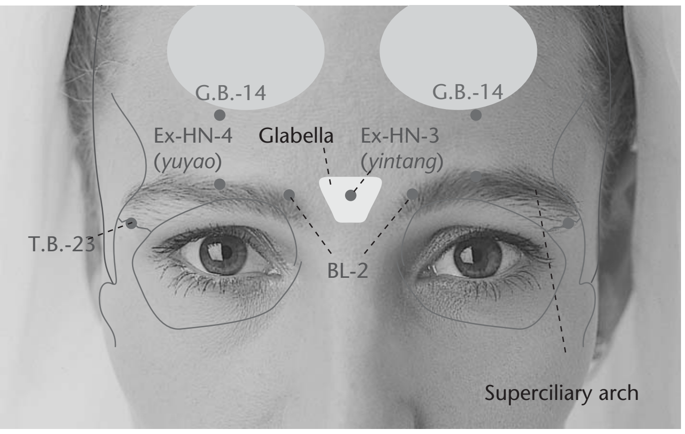
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2
Superciliary arch

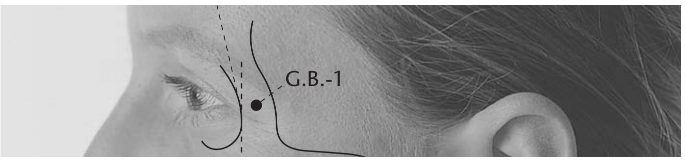

4.11 Gall Bladder Channel System – Foot *shaoyang* (*zu shao yang jing luo*)

4.11.1 The Gall Bladder Primary Channel (*zu shao yang jing*)

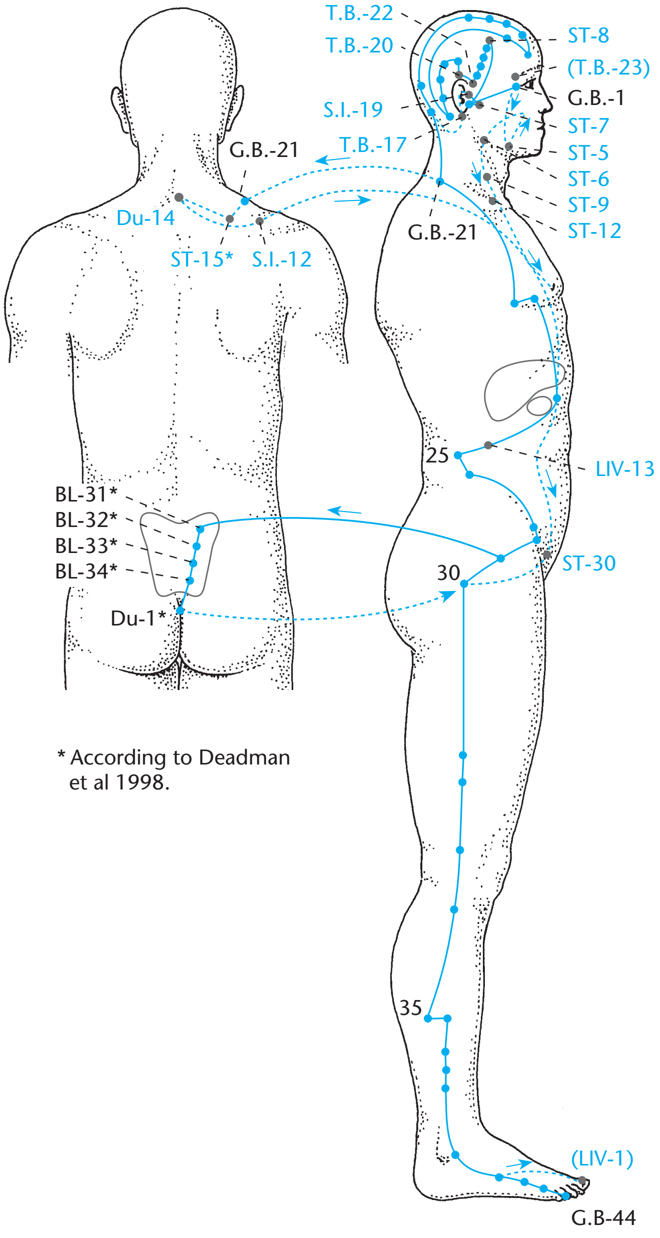
T.B.-22
T.B.-20
S.I.-19
T.B.-17
ST-8
(T.B.-23)
G.B.-1
ST-7
ST-5
ST-6
ST-9
ST-12
G.B.-21
G.B.-21
Du-14
ST-15*
S.I.-12
BL-31*
BL-32*
BL-33*
BL-34*
Du-1*
25
LIV-13
30
ST-30
35
(LIV-1)
G.B-44

* According to Deadman et al 1998.

Pathway

The Gall Bladder primary channel begins at **G.B.-1** (*tongziliao*) at the outer canthus of the eye. This point is reached by a **small branch** which separates from the Triple Burner primary channel at → **T.B.-23** (*sizhukong*) on the frontozygomatic suture, according to some authors at → **T.B.-22** (hand–foot pairing of the third great circuit, Yang axes: *shaoyang*).

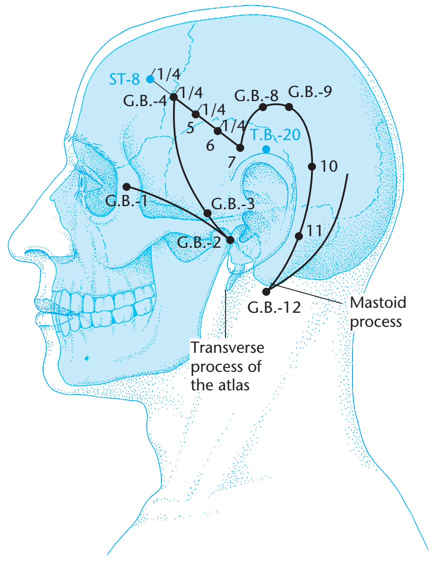
ST-8
1/4
G.B.-4
1/4
1/4
1/4
5
6
7
G.B.-8
G.B.-9
T.B.-20
10
G.B.-1
G.B.-3
11
G.B.-2
G.B.-12
Mastoid process
Transverse process of the atlas

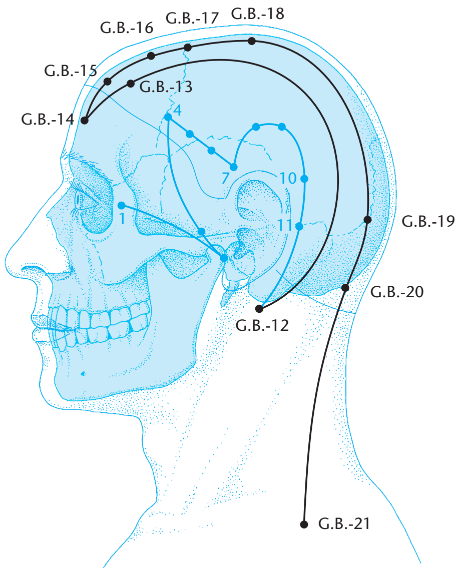
G.B.-16
G.B.-17
G.B.-18
G.B.-15
G.B.-13
G.B.-14
4
7
10
1
11
G.B.-19
G.B.-20
G.B.-12
G.B.-21

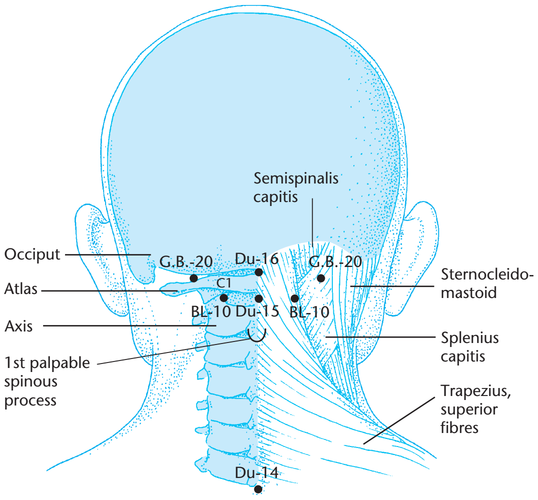
Semispinalis capitis
Occiput
Atlas
Axis
1st palpable spinous process
G.B.-20
Du-16
G.B.-20
C1
BL-10
Du-15
BL-10
Sternocleido-mastoid
Splenius capitis
Trapezius, superior fibres
Du-14

From **G.B.-1**, the **external pathway** runs to **G.B.-2** (*tinghui*) in front of the ear

➡ ascends to the forehead and **ST-8** (*touwei*)

➡ traverses the temporal region in a slight curve, passing

Ch04.11-F10028.qxd 2/23/08 11:23 AM Page 392

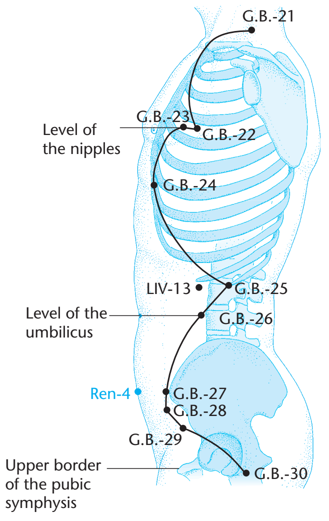
G.B.-21
Level of the nipples
G.B.-23
G.B.-22
G.B.-24
LIV-13
G.B.-25
Level of the umbilicus
G.B.-26
Ren-4
G.B.-27
G.B.-28
G.B.-29
Upper border of the pubic symphysis
G.B.-30

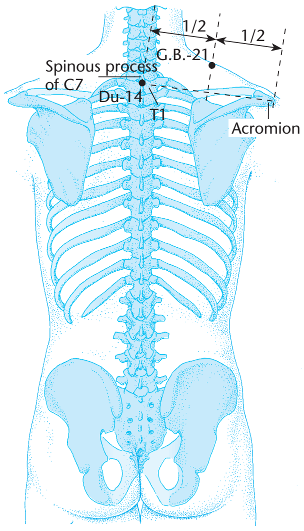
1/2
1/2
G.B.-21
Spinous process of C7
Du-14
T1
Acromion

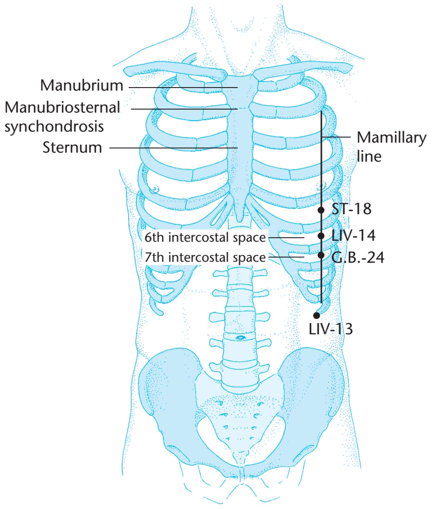
Manubrium
Manubriosternal synchondrosis
Sternum
Mamillary line
ST-18
6th intercostal space
LIV-14
7th intercostal space
G.B.-24
LIV-13

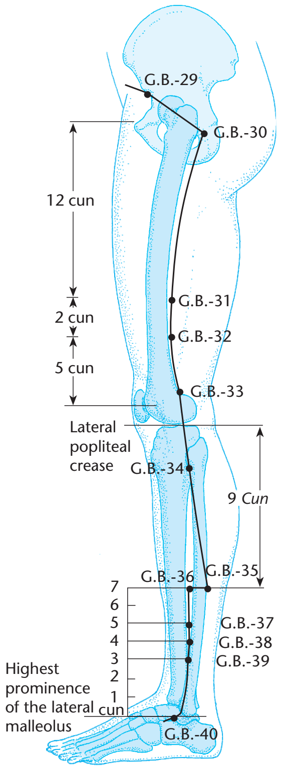
G.B.-29
G.B.-30
12 cun
G.B.-31
2 cun
G.B.-32
5 cun
G.B.-33
Lateral popliteal crease
G.B.-34
9 Cun
G.B.-36
G.B.-35
7
6
5
4
3
2
1
G.B.-37
G.B.-38
G.B.-39
Highest prominence of the lateral malleolus
cun
G.B.-40

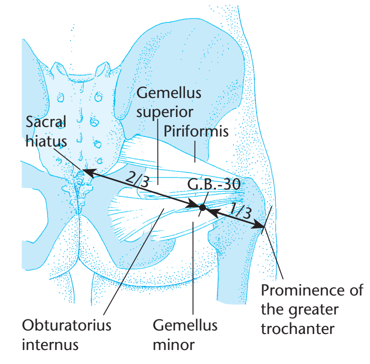
Sacral hiatus
Gemellus superior
Piriformis
2/3
G.B.-30
1/3
Prominence of the greater trochanter
Obturatorius internus
Gemellus minor

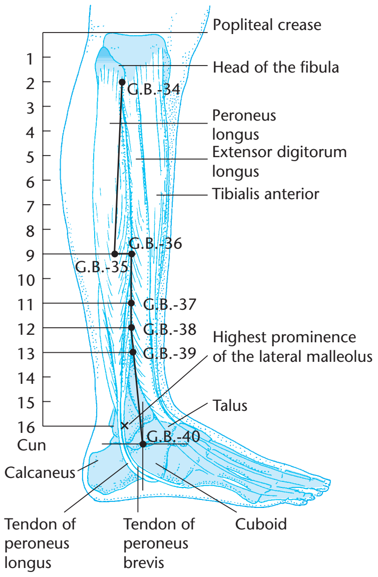
Popliteal crease
Head of the fibula
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Cun
G.B.-34
Peroneus longus
Extensor digitorum longus
Tibialis anterior
G.B.-36
G.B.-35
G.B.-37
G.B.-38
G.B.-39
Highest prominence of the lateral malleolus
Talus
G.B.-40
Calcaneus
Tendon of peroneus longus
Tendon of peroneus brevis
Cuboid

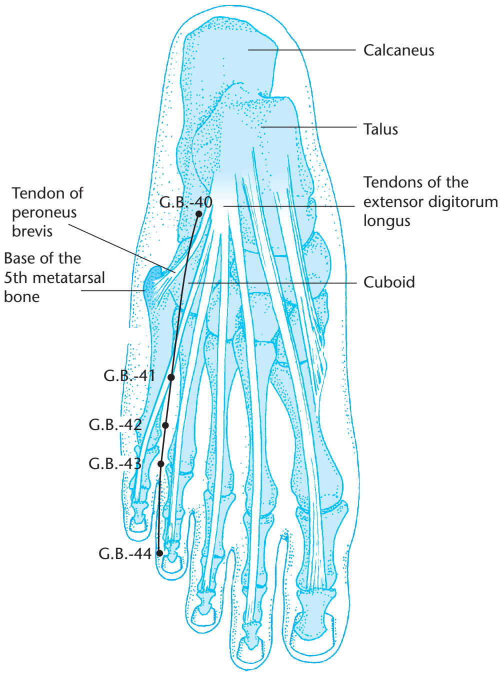
Calcaneus
Talus
Tendon of peroneus brevis
G.B.-40
Tendons of the extensor digitorum longus
Base of the 5th metatarsal bone
Cuboid
G.B.-41
G.B.-42
G.B.-43
G.B.-44

G.B.-4 to **G.B.-7**, and continues to → **T.B.-22** (*erheliao*) anterior to the root of the ear

- passes → **T.B.-20** (*jiaosun*) superior to the apex of the ear and curves posterior to the ear from **G.B.-8** to **G.B.-12**
- traverses the lateral aspect of the head to **G.B.-13** (*benshen*) and **G.B.-14** (*yangbai*) on the forehead
- again traverses the side of the head to **G.B.-20** (*fengchi*)
- crosses the superior aspect of the shoulder and passes **G.B.-21** (*jianjing*) and → **T.B.-15** (*tianliao*) to reach → **Du-14** (*dazhu*), where it meets the other Yang primary channels
- continues to **S.I.-12** (*bingfeng*), according to some authors (Deadman et al 1998) previously passing → **BL-11** (*dashu*) and **ST-12** (*quepen*) in the supraclavicular fossa.

One branch runs from **G.B.-20** to → **T.B.-17** (*yifeng*), enters the ear and travels to → **S.I.-19** (*tinggong*), passes → **ST-7** (*xiaguan*) and continues to **G.B.-1** (*tongziliao*) at the lateral canthus of the eye, descends to → **ST-5** (*daying*), again ascends to the infraorbital region, where it meets the Triple Burner primary channel; then descends to → **ST-6** (*jiache*) on the lower jaw, traverses the lateral aspect of the neck, passing → **ST-9** (*renying*) at the sternocleidomastoid, and meets the main branch again at

→ **ST-12** (*quepen*) in the supraclavicular fossa, where the latter divides into **2 branches**:

- ➡ **The inner branch** enters the thorax, meets the Pericardium primary channel at → **P-1** (*tianchi*), runs across the diaphragm, connects with the Liver (*gan*) and the Gall Bladder (*dan*), traverses the hypochondrium and continues to the lower abdomen, enters the lumbar region near the femoral artery, according to Solinas et al (1998) via → **ST-30** (*qichong*), and continues along the pubic hairline to the hip.
- ➡ **The outer branch** descends superficially from the supraclavicular fossa, crosses the midaxillary line and continues along the lateral thoracic wall, passing → **LIV-13** (*zhangmen*) at the free end of the 11th rib and continuing to the hip to **G.B.-29** (*juliao*). It then traverses the sacral region, passing → **BL-31** (*shangliao*), → **BL-32** (*ciliao*), → **BL-33** (*zhongliao*) and → **BL-34** (*xialiao*) over the sacral foramina and continues to → **Du-1** (*changqiang*). From there, it travels laterally to **G.B.-30** (*huantiao*), where it reconnects with the main branch (→ see comment below*).

The **external pathway** descends along the lateral aspect of the thigh to the knee and further along the anterior/posterior border of the fibula, passing the anterior aspect of the lateral malleolus and ending at the ulnar corner of the nail of the fourth toe.

A further branch originates at **G.B.-41** (*zulinqi*) on the dorsum of the foot and travels between the 1st and 2nd metatarsal bones to the tip of the big toe and **LIV-1** (*dadun*), where it joins the LIV channel.

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Alternating fevers and chills, headaches, eye pain, pain in the cheek and on the chin, subaxillary swellings, deafness, lateral knee and leg pain

Interior (*li*) or *zangfu*-Organ signs and symptoms: Pain in the lateral costal region, vomiting, bitter taste in the mouth, thoracic pain

Connections of the Gall Bladder primary channel (→ 1.2)

Connections with other channels (→ 1.2)

Liver primary channel (*zu jue yin jing*)

Connection: Foot Yin–Yang pair of the third great circuit
Location: G.B.-41 → LIV-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Triple Burner primary channel (*shou shao yang jing*)

Connection: Hand–foot pairing of the third great circuit: *shaoyang* (Yang axes)
Location: T.B.-23 (or **T.B.-22**, according to some authors) → **G.B.-1** (on the head)
Circulation: Circadian (according to the Organ clock)
Importance: Above–below relationship

Connections with *zangfu*-Organ systems

Liver (*gan*), Gall Bladder (*dan*)

4.11.2 The Gall Bladder Divergent Channel (*zu shao yang jing bie*)

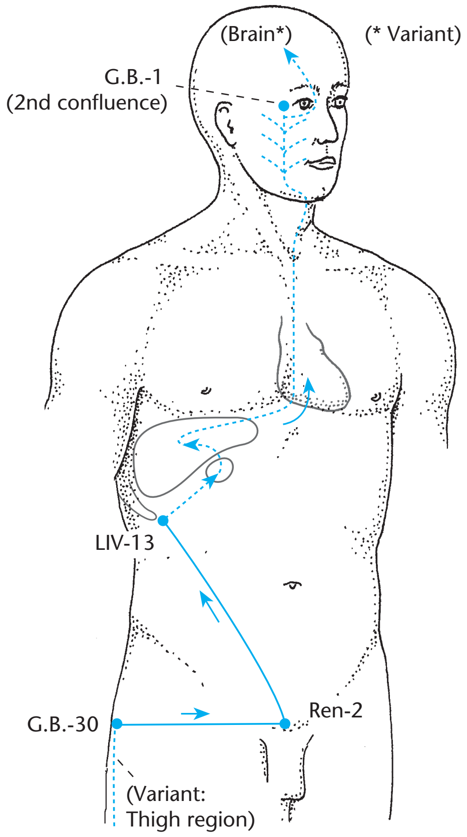

Pathway

The Gall Bladder Divergent channel separates from the Gall Bladder primary channel near **G.B.-30** (*huantiao*); according to some authors, it separates on the thigh (→ variant in the illustration)

- ➡ traverses the hip region to the anterior aspect of the body
- ➡ reaches the inguinal region and crosses the Liver divergent channel at → **Ren-2** (*qugu*)
- ➡ ascends the abdomen to **LIV-13** (*zhangmen*) below the free end of the 11th rib
- ➡ continues internally along the thoracic wall
- ➡ connects with the Gall Bladder (*dan*) and the Liver (*gan*) and passes the Heart (*xin*)
- ➡ ascends to the neck
- ➡ emerges on the lower jaw

---

*Not all authors describe the pathway of the G.B. channel as passing the points **BL-31** to **BL-34** and **Du-1** (see figure for variant pathway).

- disperses on the face
- connects with the Gall Bladder primary channel and the Liver divergent channel to form one of the six *he*-confluences (here: G.B./LIV as 2nd confluence → 1.3)
- continues to the eye system and ends in the Brain.

Clinical importance

- Strengthens the relationship between the Gall Bladder and the Liver (*zangfu*-Organ systems). Points on the Gall Bladder primary channel can therefore be used to treat disorders of the Liver and vice versa.
- Supports the relationship between the oesophagus and the Heart.
- The Gall Bladder Divergent channel covers the area around the eyes, thus supporting the action of some G.B. points for disorders of the eyes.

4.11.3 The Gall Bladder Sinew Channel (*zu shao yang jing jin*)

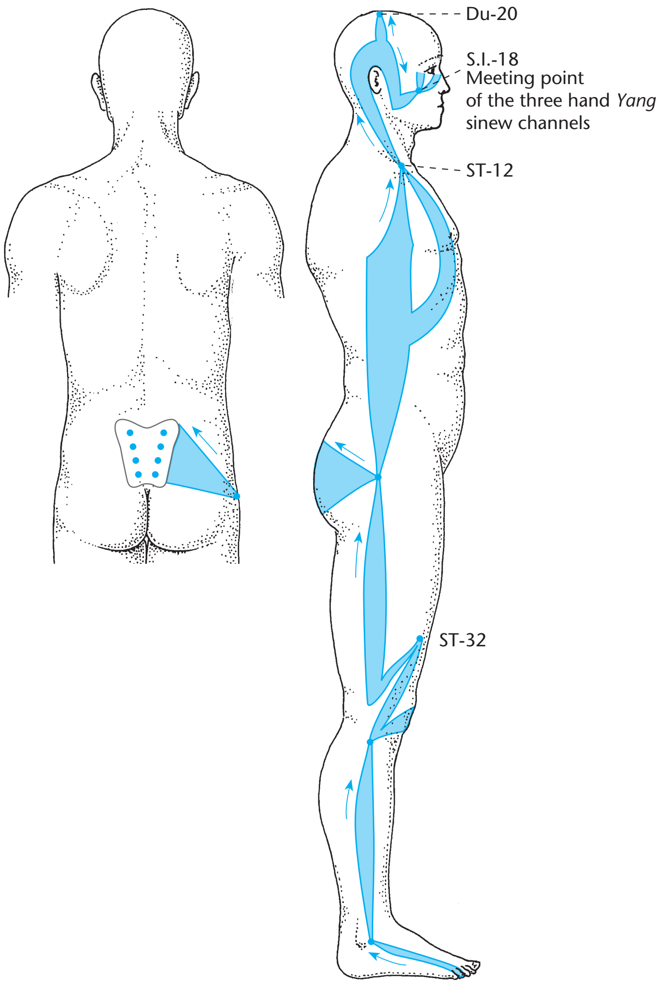

Pathway

The Gall Bladder sinew channel begins at the 4th toe, binds (*jie*) anterior and inferior to the lateral malleolus in the area of **G.B.-40** (*qiuxu*), follows the lateral aspect of the leg, binds (*jie*) at the fibula and the lateral aspect of the knee.

From the fibula, it continues to ascend the lateral aspect of the thigh

- while a **branch** runs obliquely to → **ST-32** (*futu*), where it binds (*jie*).

The main branch further ascends the leg to bind (*jie*) at the greater trochanter. Here, a **branch** separates and disperses over the gluteal and sacral regions.

From the hip, the channel continues to ascend the flanks to the lower costal region, where it divides into **two branches**:

- **one branch** traverses the lateral aspect of the thorax and ascends to bind (*jie*) at → **ST-12** (*quepen*)
- **the other branch** follows the midaxillary line to the thorax, meeting the other branch in the supraclavicular fossa.

From **ST-12**, the channel ascends the lateral aspect of the neck

- curves around the ear
- a branch continues to the apex of the ear and to → **Du-20** (*baihui*)
- from the temporal region, **a branch** descends and traverses the cheek to the lower jaw
- crosses the zygomatic arch to reach **S.I.-18** (*quanliao*), where it meets the other foot Yang sinew channels and divides into **two branches**. One branch travels laterally to the root of the nose, the other to the outer canthus of the eye.

Clinical importance

Pathology: Stiffness and distending sensations in the area of the 4th toe as well as on the lateral aspect of the knee. Limited range of motion of the knee joint. Pain, tension and distending sensations in the popliteal crease radiating to the thigh and the sacral region and vice versa. Pain and distending sensations in the sacral region radiating to the hypochondrium and vice versa. Pain and tension in the supraclavicular fossa, thorax, breast and neck region. Pain along the left side of the channel and inability to open the right eye and vice versa.

Indication: Predominantly for pain, muscle tension, stiff joints and distending sensations on the lateral aspect of the body. Used for disorders of the anterior, lateral and posterior aspects of the legs, especially those affecting the knees and the thighs. The Gall Bladder Sinew channel spreads in the thorax and the breasts so that points on the G.B. channel can be used for disorders in those areas. Also for headaches on the vertex and temporal regions. The channel also reaches the lateral aspect of the nose, which supports the relationship with the nose. Therefore, G.B. points can be used for chronic nasal disorders, chronic sinusitis, etc.

Ch04.11-F10028.qxd 2/23/08 11:23 AM Page 395

4.11.4 The Gall Bladder *luo*-Connecting Channel (*zu shao yang luo mai*)

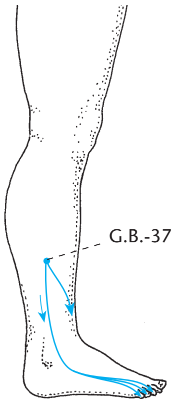

Pathway

The Gall Bladder *luo*-connecting channel separates from the Gall Bladder primary channel at its *luo*-connecting point **G.B.-37** (*guangming*) (→ 8.1.2). It forms a three-dimensional reticulate network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Liver primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** G.B. *luo* vessel to the *yuan*-source point **LIV-3** (*taichong*).
- A **longitudinal** division descends to the dorsum of the foot and spreads to the 3rd, 4th and 5th toes.

Clinical importance (→ 8.1.2)

Pathology

Excess (*shi*): Cold sensations in the feet

Deficiency (*xu*): Weakness, weak musculature of the foot, including difficulty in standing, paralysis of the lower extremity

4.11.5 Cutaneous Region (*shao yang pi bu*)

See description and figures → 1.6.

4.11.6 Points of the Gall Bladder Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): G.B.-40** (*qiuxu*) 🟦■
- ***Luo*-connecting point (→ 8.1.2.): G.B.-37** (*guangming*) 🟦■
- ***Xi*-cleft point (→ 8.1.3): G.B.-36** (*waiqiu*)
- **Associated Back-*shu* point (→ 8.1.4): BL-19** (*danshu*) 🟦■
- **Associated Front-*mu* point (→ 8.1.5): G.B.-24** (*riyue*) 🟦■
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Metal): **G.B.-44** (*zuqiaoyin*) 🟦■
  ying-spring point (Water), tonification point: **G.B.-43** (*xiaxi*)
  shu-stream point (Wood), *ben* point (Five Phases): **G.B.-41** (*zulinqi*) 🟦■
  jing-river point (Fire), sedation point: **G.B.-38** (*yangfu*)
  he-sea point (Earth): **G.B.-34** (*yanglingquan*) 🟦🟦
- ***Hui*-meeting point (→ 8.1.7)**
  – with the sinews: **G.B.-34** (*yanglingquan*) 🟦🟦
  – with the Marrow: **G.B.-39** (*xuanzhong*) 🟦■
- **Opening point (→ 8.1.8) of the *dai mai:* G.B.-41** (*zulinqi*) 🟦■
- **Lower *he*-sea point (→ 8.1.9) of the Gall Bladder: G.B.-34** (*yanglingquan*) 🟦🟦
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the S.I. and T.B. channels: **G.B.-1** (*tongziliao*)
  – with the T.B. and ST channels: **G.B.-3** (*shangguan*), **G.B.-4** (*hanyan*)
  – with the T.B., L.I. and ST channels: **G.B.-5** (*xuanlu*), **G.B.-6** (*xuanli*)
  – with the BL channel: **G.B.-7** (*qubin*); **G.B.-8** (*shuaigu*) 🟦■, **G.B.-9** (*tianchong*), **G.B.-10** (*fubai*)
  – with the BL, (S.I., T.B.*) channels: **G.B.-11** (*touqiaoyin*)
  – with the BL channel: **G.B.-12** (*wangu*) 🟦■
  – with the *yang wei mai*: **G.B.-13** (*benshen*) 🟦■
  – with the *yang wei mai*, (T.B., ST, L.I. channels*): **G.B.-14** (*yangbai*) 🟦■
  – with the *yang wei mai* and the BL channel: **G.B.-15** (*toulinqi*)
  – with the *yang wei mai*: **G.B.-16** (*muchuang*). **G.B.-17** (*zhengying*), **G.B.-18** (*chengling*), **G.B.-19** (*naokong*)
  – with the *yang wei mai, yang qiao mai* and (T.B. channel*): **G.B.-20** (*fengchi*) 🟦🟦
  – with the *yang wei mai*, T.B. and (ST*) channels: **G.B.-21** (*jianjing*) 🟦■
  – with the BL channel*: **G.B.-23** (*zhejin*)
  – with the SP channel and (*yang wei mai**): **G.B.-24** (*riyue*) 🟦■
  – with the *dai mai*: **G.B.-26** (*daimai*), **G.B.-27** (*wushu*), **G.B.-28** (*weidao*)
  – with the *yang qiao mai* and (*dai mai**): **G.B.-29** (*juliao*)
  – with the BL channel: **G.B.-30** (*huantiao*) 🟦■
  – with the *yang wei mai**: **G.B.-35** (*yangjiao*)
  – of other channels with the G.B. channel: **ST-7, ST-8, ST-9, ST-12, S.I.12, S.I.-19, (BL-1, BL-11*), BL-31–34, P-1, (T.B.-15*), T.B-17, T.B.-20, T.B.-22, LIV-13, Du-1, Du-14, (Du-20, ST-5, ST-6, ST-30*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): G.B.-9** (*tianchong*)
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star points (→ 8.1.14): G.B.-30** (*huantiao*) 🟦■, **G.B.-34** (*yanglingquan*) 🟦🟦

---

* Mentioned by only some authors.

- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - – main point for all disorders of the head, sensory organs, Brain: **G.B.-20** (*fengchi*) 🟦🟦
  - – front-*mu* point of the Kidneys (→ 8.1.5): **G.B.-25** (*jingmen*) 🟦■
  - – *xi*-cleft point (→ 8.1.3) of the *yang wei mai:* **G.B.-35** (*yangjiao*)

Points according to region

- **Local points (→ 8.2.1):** forehead – **G.B.-14** (*yangbai*) 🟦■; head, temporal region – **G.B.-8** (*shuaigu*) 🟦■; head, occipital region – **G.B.-20** (*fengchi*) 🟦🟦; ears – **G.B.-2** (*tingshui*) 🟦■; cervical spine – **G.B.-20** (*fengchi*) 🟦🟦; neck and shoulder region – **G.B.-20** (*fengchi*) 🟦■, **G.B.-21** (*jianjing*) 🟦■; Gall Bladder – **G.B.-24** (*riyue*) 🟦■; hypochondrium – **G.B.-25** (*jingmen*) 🟦■, **G.B.-26** (*daimai*), **G.B.-27** (*wushu*); hip – **G.B.-29** (*juliao*), **G.B.-30** (*huantiao*) 🟦■; knee – **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-35** (*yangjiao*); foot – **G.B.-40** (*qiuxu*) 🟦■
- **Adjacent points (→ 8.2.1):** head, temporal region – **G.B.-20** (*fengchi*) 🟦🟦; ears – **G.B.-8** (*shuaigu*) 🟦■, **G.B.-20** (*fengchi*) 🟦🟦; cervical spine and shoulders – **G.B.-21** (*jianjing*) 🟦■; lumbar region – **G.B.-25** (*jingmen*) 🟦■, **G.B.-30** (*huantiao*) 🟦■; hip – **G.B.-31** (*fengshi*); knee – **G.B.-33** (*xiyangguan*); foot – **G.B.-34** (*yanglingquan*) 🟦🟦; toes – **G.B.-38** (*yangfu*)
- **Distal points (→ 8.2.1):** head, temporal region – **G.B.-43** (*xiaxi*), **G.B.-41** (*zulinqi*) 🟦■; eyes – **G.B.-37** (*guangming*) 🟦■; ears – **G.B.-41** (*zulinqi*) 🟦■; Liver – **G.B.-34** (*yanglingquan*) 🟦🟦; Gall Bladder – **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-40** (*qiuxu*) 🟦■; hypochondrium – **G.B.-38** (*yangfu*), **G.B.-34** (*yanglingquan*) 🟦🟦, **G.B.-43** (*xiaxi*); cervical spine and hip – **G.B.-39** (*xuanzhong*) 🟦■; hip – **G.B.-41** (*zulinqi*) 🟦■

Specific points according to the channel pathway (in numeric order)

- **G.B.-1** (*tongziliao*): *jiaohui*-meeting point with the S.I. and T.B. channels (→ 8.1.10)
- **G.B.-2** (*tinghui*) 🟦■: local point for the ears (→ 8.2.1)
- **G.B.-3** (*shangguan*): *jiaohui*-meeting point with the T.B. and ST channels (→ 8.1.10)
- **G.B.-4** (*hanyan*): *jiaohui*-meeting point with the T.B. and ST channels (→ 8.1.10)
- **G.B.-5** (*xuanlu*): *jiaohui*-meeting point with the T.B., L.I. and ST channels (→ 8.1.10)
- **G.B.-6** (*xuanli*):): *jiaohui*-meeting point with the T.B., L.I. and ST channels (→ 8.1.10)
- **G.B.-7** (*qubin*):): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-8** (*shuaigu*) 🟦■: *jiaohui*-meeting point with the BL channel (→ 8.1.10); local point for the lateral aspect of the head (→ 8.2.1); adjacent point for the ears (→ 8.2.1)
- **G.B.-9** (*tianchong*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)*; Window of Heaven point (→ 8.1.12)*
- **G.B.-10** (*fubai*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-11** (*touqiaoyin*): *jiaohui*-meeting point with the BL, (S.I., T.B.)* channels (→ 8.1.10)
- **G.B.-12** (*wangu*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-13** (*benshen*) 🟦■: *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-14** (*yangbai*) 🟦■: *jiaohui*-meeting point with the *yang wei mai* and (T.B., ST, L.I.*) channels (→ 8.1.10); local point for the forehead (→ 8.2.1)
- **G.B.-15** (*toulinqi*): *jiaohui*-meeting point with the *yang wei mai* and BL channel (→ 8.1.10)
- **G.B.-16** (*muchuang*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-17** (*zhengying*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-18** (*chengling*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-19** (*naokong*): *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10)
- **G.B.-20** (*fengchi*) 🟦🟦: *jiaohui*-meeting point with the *yang wei mai, yang qiao mai* and (T.B. channel*) (→ 8.1.10); local point for the occiput (→ 8.2.1); local point for the neck and shoulders (→ 8.2.1); local point for all disorders of the head, sensory organs, Brain; adjacent point for the lateral aspect of the head and the ears (→ 8.2.1)
- **G.B.-21** (*jianjing*) 🟦■: *jiaohui*-meeting point with the *yang wei mai*, T.B. and (ST*) channels (→ 8.1.10); adjacent point for the neck and shoulders (→ 8.2.1)
- **G.B.-23** (*zhejin*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **G.B.-24** (*riyue*) 🟦■: associated Front-*mu* point (→ 8.1.5); *jiaohui*-meeting point with the SP channel and (*yang wei mai**) (→ 8.1.10); local point for the Gall Bladder (→ 8.2.1)
- **G.B.-25** (*jingmen*) 🟦■: Front-*mu* point of the Kidneys (→ 8.1.5); local point for the hypochondrium (→ 8.2.1); adjacent point for the lumbar region (→ 8.2.1)
- **G.B.-26** (*daimai*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10); local point for the hypochondrium (→ 8.2.1)
- **G.B.-27** (*wushu*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10); local point for the hypochondrium (→ 8.2.1)
- **G.B.-28** (*weidao*): *jiaohui*-meeting point with the *dai mai* (→ 8.1.10);
- **G.B.-29** (*juliao*): *jiaohui*-meeting point with the *yang qiao mai* and (*dai mai**) (→ 8.1.10); local point for the hip (→ 8.2.1)
- **G.B.-30** (*huantiao*) 🟦■: *jiaohui*-meeting point with the BL channel (→ 8.1.10); Ma Dan Yang Heavenly Star point (→ 8.1.14); adjacent point for the lumbar region and the hip (→ 8.2.1); local point for the hip (→ 8.2.1)
- **G.B.-31** (*fengshi*): adjacent point for the hip (→ 8.2.1)
- **G.B.-33** (*xiyangguan*): adjacent point for the knee (→ 8.2.1)

- **G.B.-34** (*yanglingquan*) 🟦🟦: *he*-sea point (Earth) (➞ 8.1.6); *hui*-meeting point with the sinews (➞ 8.1.7); lower *he*-sea point of the Gall Bladder (➞ 8.1.9); Ma Dan Yang Heavenly Star point (➞ 8.1.14); distal point for the Liver, Gall Bladder and hypochondrium (➞ 8.2.1); local point for the knee (➞ 8.2.1); adjacent point for the foot (➞ 8.2.1)
- **G.B.-35** (*yangjiao*): *xi*-cleft point of the *yang wei mai* (➞ 8.1.3); *jiaohui*-meeting point with the *yang wei mai** (➞ 8.1.10); local point for the knee (➞ 8.2.1)
- **G.B.-36** (*waiqiu*): *xi*-cleft point (➞ 8.1.3)
- **G.B.-37** (*guangming*) 🟦■: *luo*-connecting point (➞ 8.1.2); distal point for the eyes (➞ 8.2.1)
- **G.B.-38** (*yangfu*): *jing*-river point (Fire) (➞ 8.1.6); sedation point; distal point for the hypochondrium (➞ 8.2.1)
- **G.B.-39** (*xuanzhong*) 🟦■: *hui*-meeting point with the Marrow (➞ 8.1.7); distal point for the cervical spine and the hip (➞ 8.2.1)
- **G.B.-40** (*qiuxu*) 🟦■: *yuan*-source point (➞ 8.1.1); local point for the foot (➞ 8.2.1); distal point for the Gall Bladder (➞ 8.2.1)
- **G.B.-41** (*zulinqi*) 🟦■: *shu*-stream point (Wood); *ben* point (Five phases) (➞ 8.1.6); Opening point (➞ 8.1.8) of the *dai mai*; distal point for the temporal region and the ears (➞ 8.2.1); distal point for the hip (➞ 8.2.1)
- **G.B.-43** (*xiaxi*): *ying*-spring point (Water) (➞ 8.1.6); tonification point; distal point for the lateral aspect of the head and the hypochondrium (➞ 8.2.1)
- **G.B.-44** (*zuqiaoyin*) 🟦■: *jing*-well point (Metal) (➞ 8.1.6)

Pupil Crevice TONGZILIAO **G.B.-1**

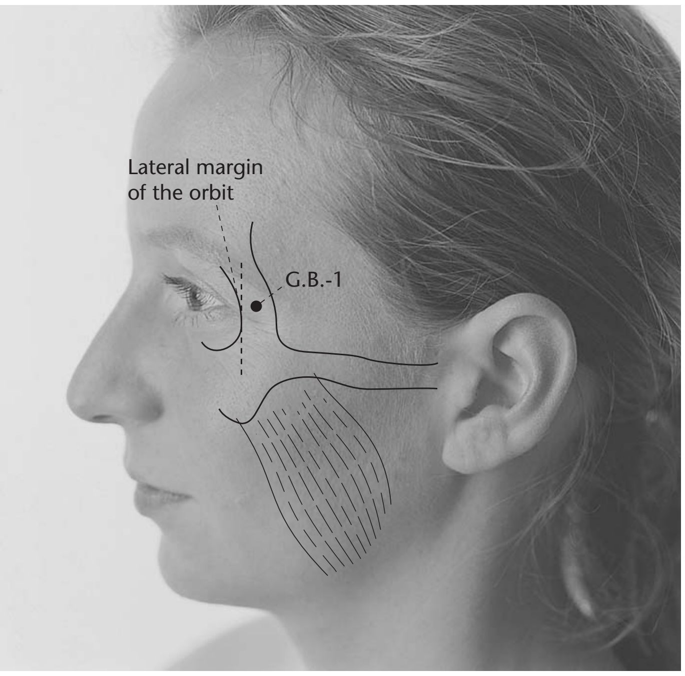
Lateral margin
of the orbit
G.B.-1

Location

In a bony depression on the lateral aspect of the orbital margin, on the level of the outer canthus of the eye.

How to find

From the outer canthus of the eye, palpate in a lateral direction. Locate **G.B.-1** on the outer aspect of the orbit (diagonally inferior to the temple), where a bony depression can be palpated. → **T.B.-23** is located more superiorly on the lateral end of the eyebrow, in a depression on the frontozygomatic suture. → **BL-1** is located at the inner canthus of the eye.

Needling

0.2–0.3 cun obliquely in a posterior direction or up to 1 cun transversely (subcutaneously) towards → **Ex-HN-5** (*taiyang*). According to some texts, moxibustion is contraindicated.

Actions/Indications

- Expels Wind and Heat from the eyes

Special features

Meeting point with the S.I. and T.B. channels, entry point.

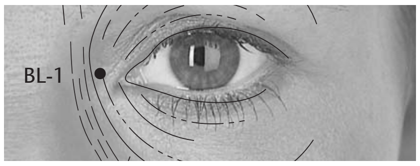

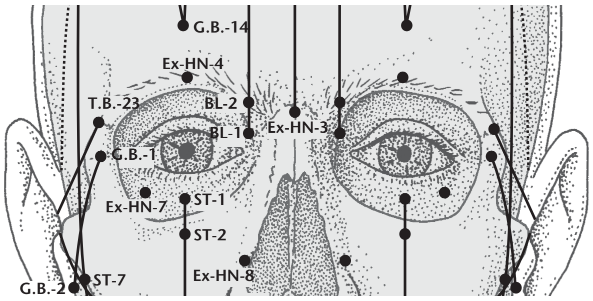
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7

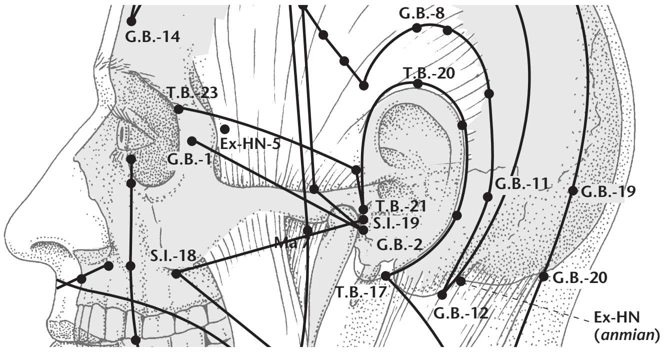
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma 7
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(anmian)

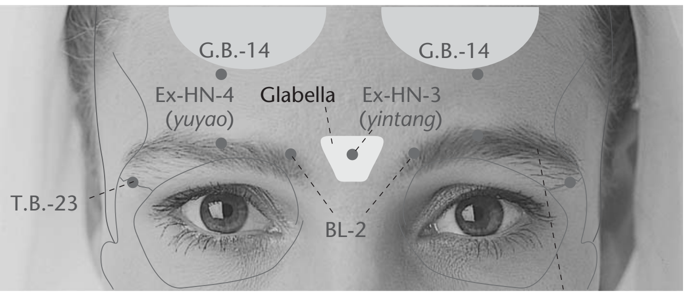
G.B.-14
G.B.-14
Ex-HN-4
(yuyao)
Glabella
Ex-HN-3
(yintang)
T.B.-23
BL-2

G.B.-2 *Meeting of Hearing* TINGHUI

Location

Anterior to the ear, with the patient's mouth open in the depression on the level of the intertragic notch, on the lower border of the condyloid process of the mandible.

How to find

Locate and needle with the patient's mouth open as this allows the condyloid process of the mandible to slide towards the anterior, revealing the depression where **G.B.-2** is located. Find the vertical sulcus anterior to the ear (at the ear/cheek junction), which may be more or less pronounced. Then locate **G.B.-2** on the sulcus, on the level of the intertragic notch. If the sulcus is not clearly defined (it becomes more clearly visible with increasing age), the depression can be located by using an (ear) point locator. **G.B.-2** is the most distal of three points located anterior to the ear (→ **S.I.-19** and → **T.B.-21** are both located more superiorly).

Needling

0.5–1 cun vertically or slightly obliquely in an inferior direction. Needle with the patient's mouth open to avoid intra-articular insertion. Caution: Like **T.B.-21** and **S.I.-19**, this point is located close to the superficial temporal artery and the auriculotemporal nerve.

Actions/Indications

- Eliminates Wind, clears Heat, benefits the ears and the temporomandibular joint
- Opens the channel and the *luo*-connecting vessel, alleviates pain

Special features

Important local point for disorders of the ears and the jaw. This point is often used alternately with **T.B.-21** and **S.I.-19**.

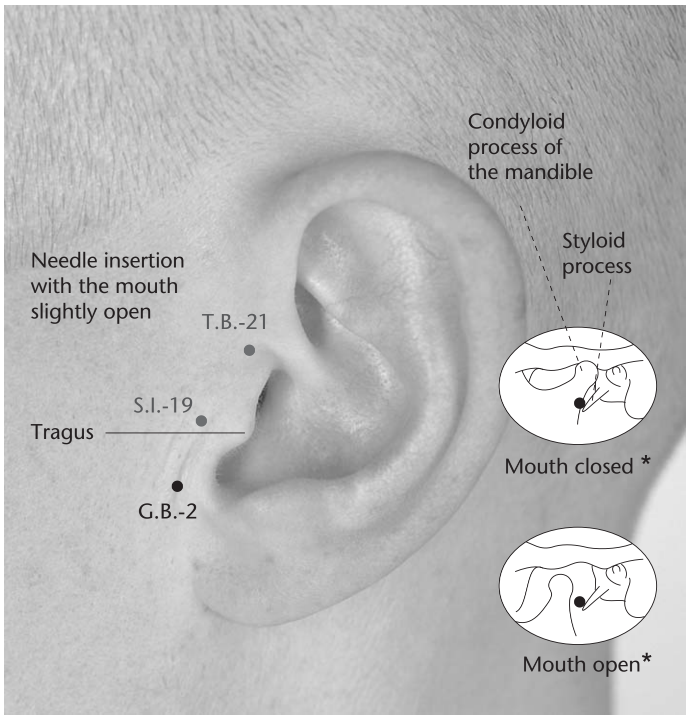
Condyloid
process of
the mandible
Styloid
process
Needle insertion
with the mouth
slightly open
T.B.-21
S.I.-19
Tragus
G.B.-2
Mouth closed *
Mouth open*

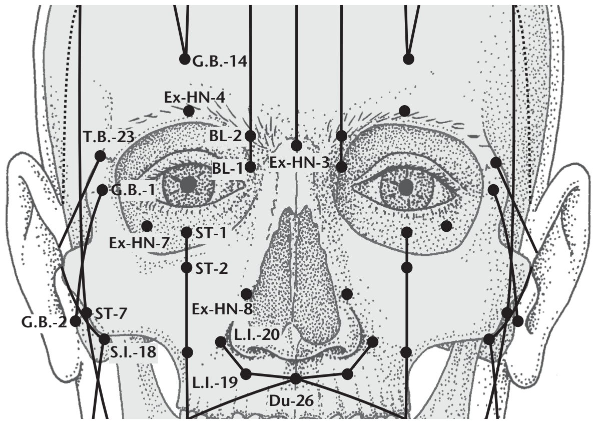
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8
G.B.-2
ST-7
L.I.-20
S.I.-18
L.I.-19
Du-26

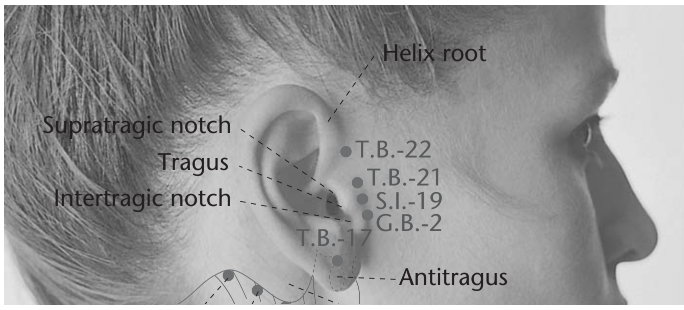
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Antitragus

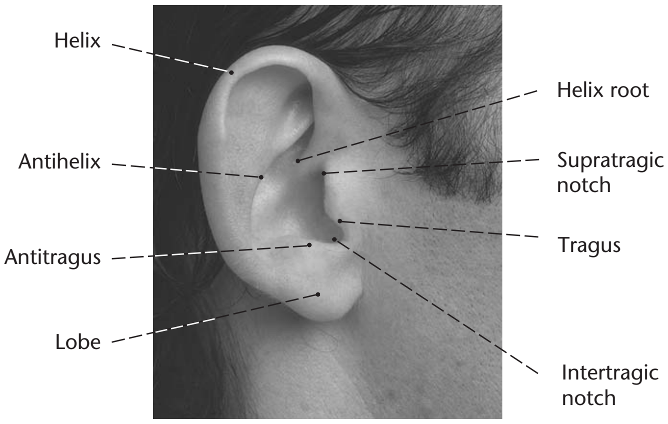
Helix
Antihelix
Antitragus
Lobe
Helix root
Supratragic
notch
Tragus
Intertragic
notch

---

* According to Deadman et al 1998.

Above the Joint SHANGGUAN G.B.-3

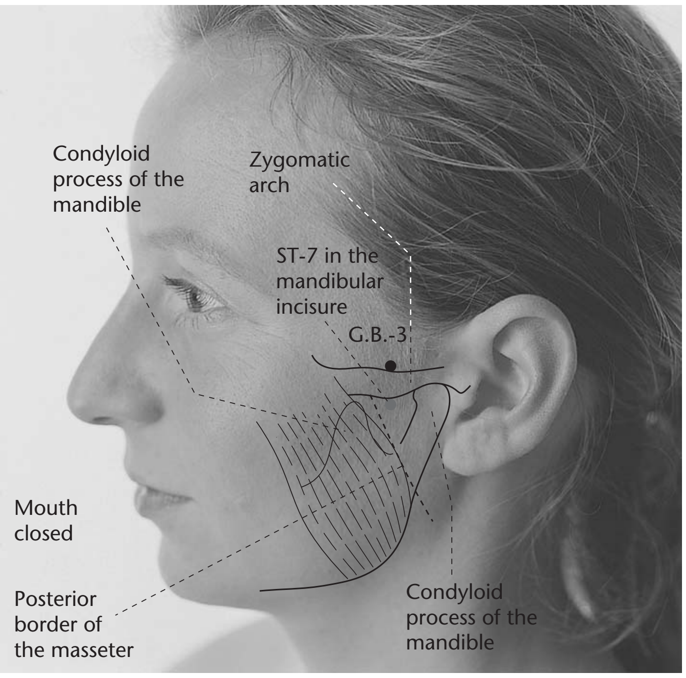
Condyloid
process of the
mandible
Zygomatic
arch
ST-7 in the
mandibular
incisure
G.B.-3
Mouth
closed
Posterior
border of
the masseter
Condyloid
process of the
mandible

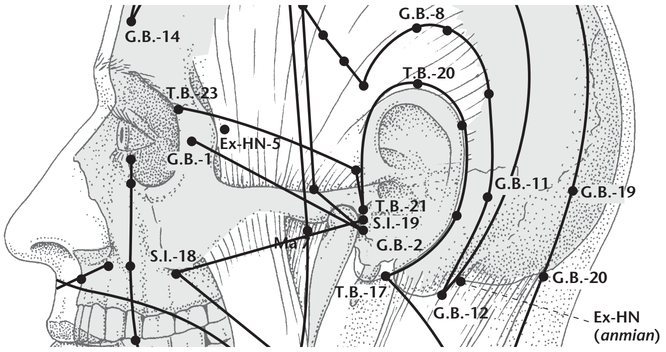
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma 7
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(*anmian*)

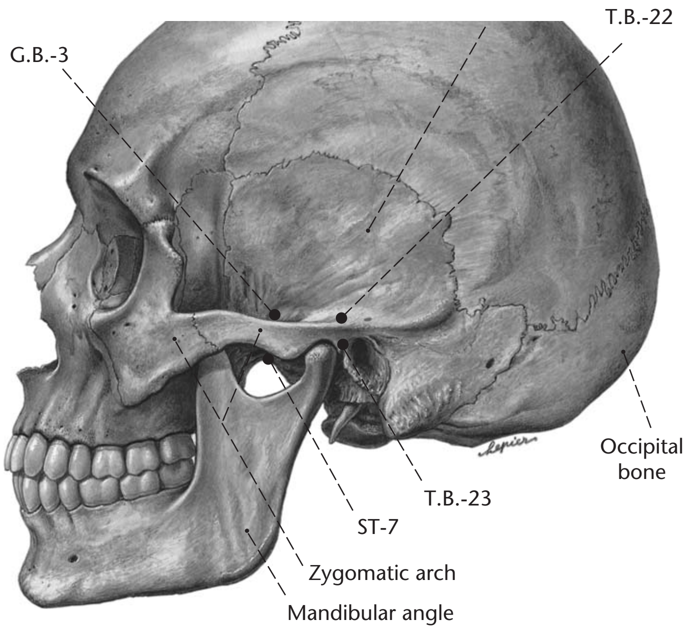
G.B.-3
T.B.-22
Occipital
bone
T.B.-23
ST-7
Zygomatic arch
Mandibular angle

Location

In a depression on the upper border of the zygomatic arch, approximately 1 cun anterior to the root of the ear, superior to → **ST-7**.

How to find

Locate the zygomatic arch (→ 3.1.2) by palpating from the root of the ear (root of the helix) approximately 1 cun towards the orbit. Follow its course by placing one finger above and the other finger below the arch. As soon as the lower finger has reached a clearly palpable depression anterior to the temporomandibular joint and posterior to the masseter muscle (→ **ST-7**), the upper finger will be resting on **G.B.-3**, which is located directly superior to → **ST-7** in a shallow depression on the superior border of the zygomatic arch.

Needling

Vertically 0.3–0.5 cun, no strong stimulation. Caution: Branches of the temporal, transverse facial and masseteric arteries! Traditionally, deep needling is prohibited.

Actions/Indications

- Opens the channel, alleviates pain
- Eliminates Wind
- Benefits the ears

Special features

Meeting point with the T.B. and ST channels

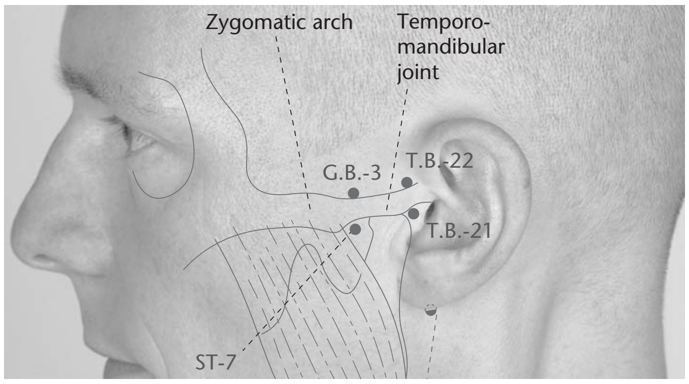
Zygomatic arch
Temporo-
mandibular
joint
G.B.-3
T.B.-22
T.B.-21
ST-7

G.B.-4 *Jaw Serenity* HANYAN

Location

At the junction of the upper quarter and the second quarter of a line connecting → **ST-8** and → **G.B.-7**.

How to find

First, locate the two reference points: → **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and → **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-4** at the junction of the upper quarter with the lower three quarters. Generally, **G.B.-4** is located on the temporal hairline and the anterior portion of the temporalis muscle, which can be felt when chewing.

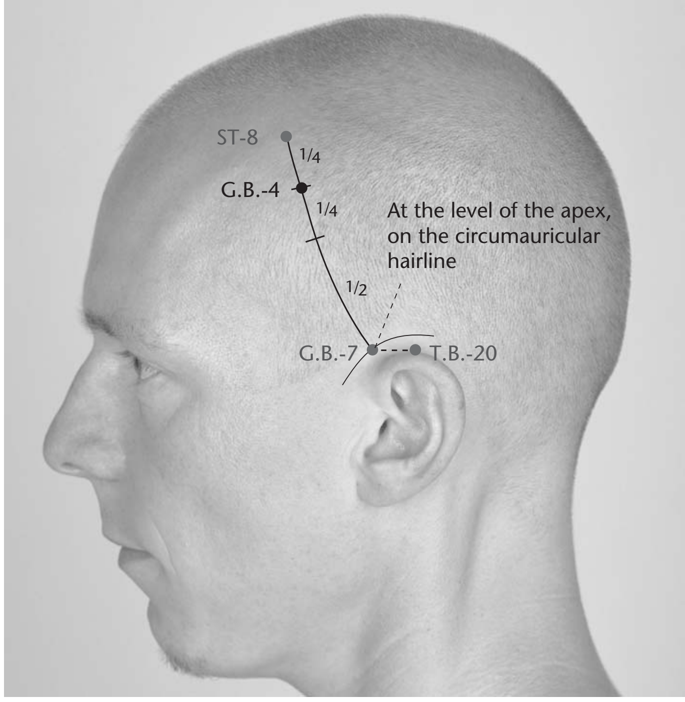
ST-8
1/4
G.B.-4
1/4
At the level of the apex, on the circumauricular hairline
1/2
G.B.-7
T.B.-20

Needling

0.3–1.5 cun transversely (subcutaneously), tangentially along the skull, in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Eliminates Wind and Heat
- Opens the channel

Special features

Meeting point with the T.B. and ST channel

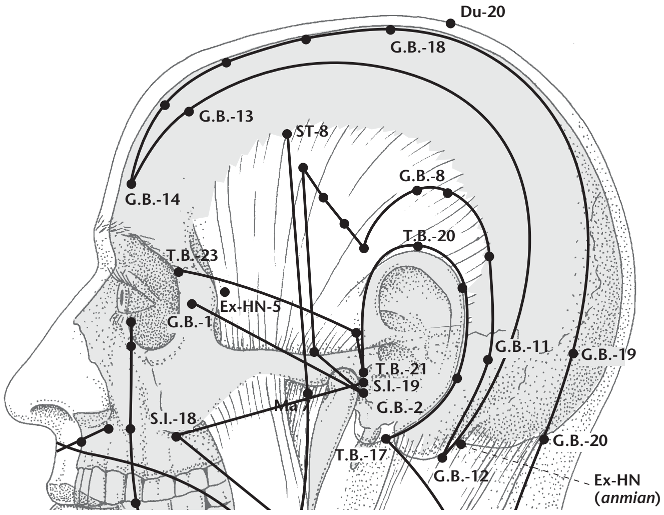
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma 7
G.B.-2
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN
(*anmian*)

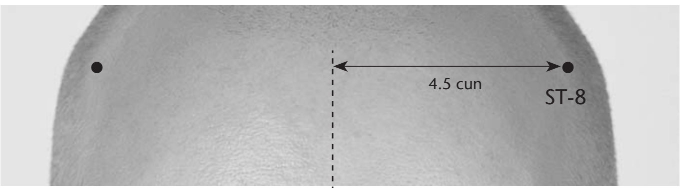

Suspended Skull **XUANLU** **G.B.-5**

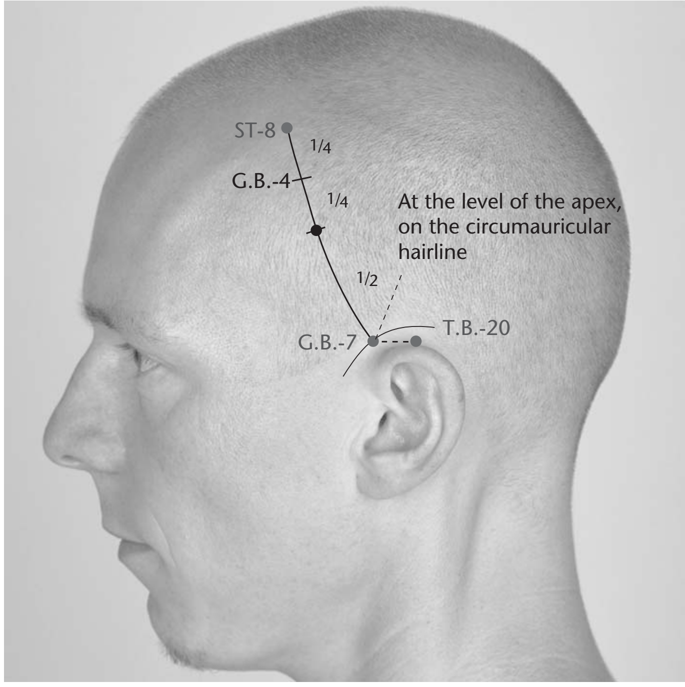
ST-8
1/4
G.B.-4
1/4
At the level of the apex, on the circumauricular hairline
1/2
G.B.-7
T.B.-20

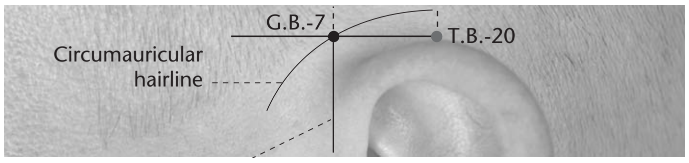
G.B.-7
T.B.-20
Circumauricular hairline

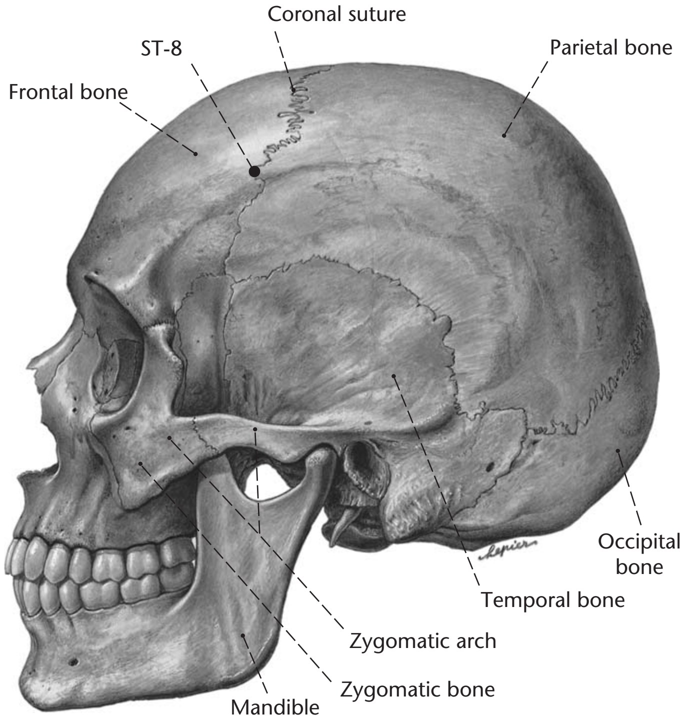
Coronal suture
ST-8
Parietal bone
Frontal bone
Occipital bone
Temporal bone
Zygomatic arch
Zygomatic bone
Mandible

Location

At the junction of the 2nd and 3rd quarter of an imaginary line connecting → **ST-8** and → **G.B.-7**.

How to find

First, locate the two reference points: → **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and → **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-5** at the junction of the upper two quarters with the lower two quarters. Generally, this point is located on the level of the parietal suture, just within the hairline.

Needling

0.3–1.5 cun transversely (subcutaneously), tangentially along the skull, in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Expels Wind and Heat
- Opens the channel

Special features

Meeting point with the T.B., L.I. and ST channels

G.B.-6 *Suspended Hair* XUANLI

Location

At the junction of the lower quarter and the upper three quarters of a line connecting → **ST-8** and → **G.B.-7**.

How to find

First, locate the two reference points: → **ST-8** (4.5 cun lateral to the midline and 0.5 cun within the anterior hairline, at the corner of the forehead) and → **G.B.-7** (in the depression on the level of the apex of the ear, within the circumauricular temporal hairline). Then divide the slightly curved line between these two points into quarters and locate **G.B.-6** at the junction of the upper three quarters with the lower quarter.

Needling

0.3–1.5 cun transversely (subcutaneously) in the direction of the occiput or towards the disorder/pain.

Actions/Indications

- Expels Wind and Heat
- Opens the channel

Special features

Meeting point with the L.I., ST and T.B. channels

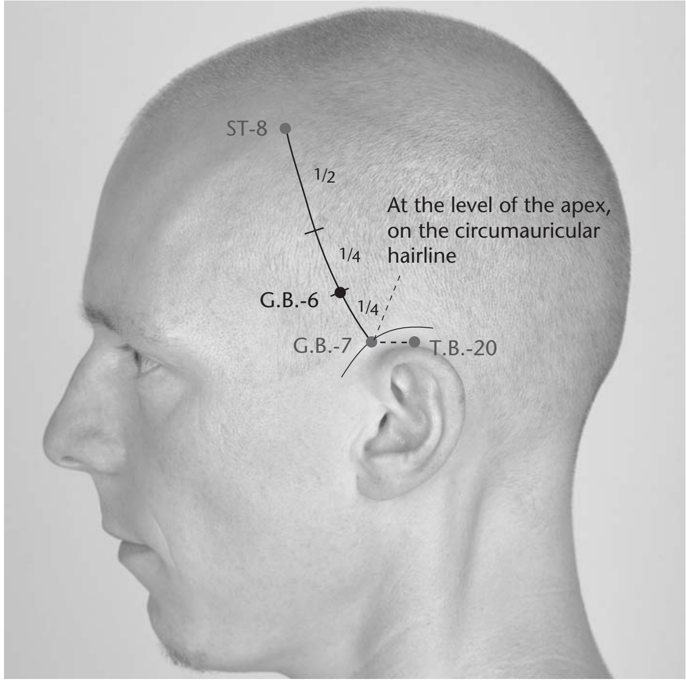
ST-8
1/2
1/4
1/4
G.B.-6
G.B.-7
T.B.-20
At the level of the apex, on the circumauricular hairline

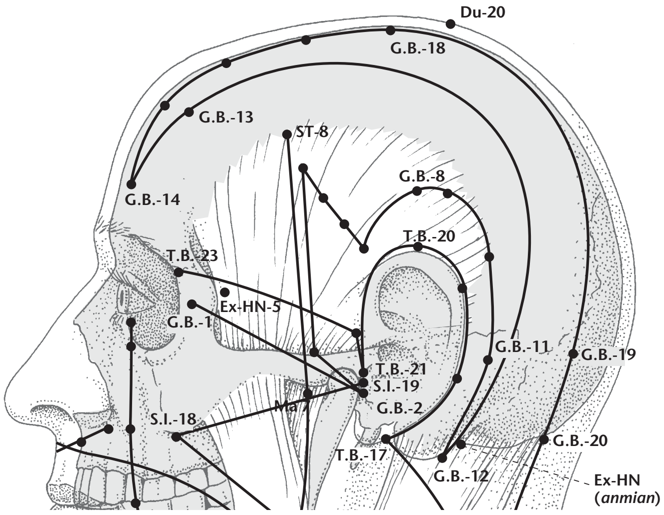
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)

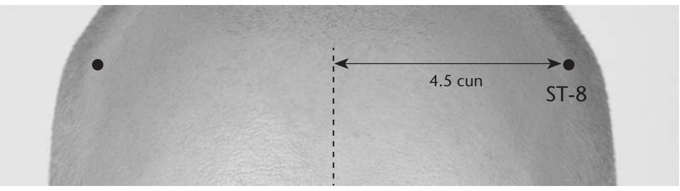

Crook of the Temple QUBIN G.B.-7

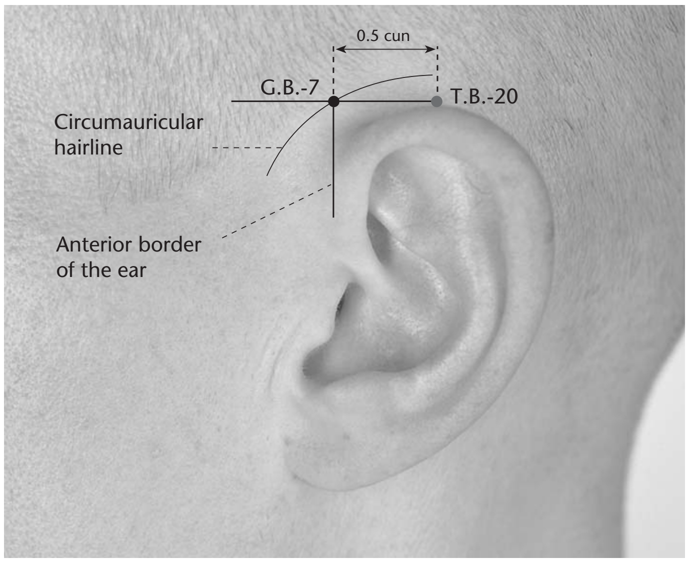

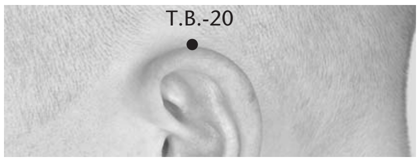

Location

In a depression on the level of the apex of the ear, within the circumauricular temporal hairline, approximately at the junction of a horizontal line through the apex of the ear and a vertical line along the posterior border of the temple anterior to the ear.

How to find

Locate **G.B.-7** by palpating for a small depression anterior to the apex of the ear at the circumauricular temporal hairline. The point is located approximately at a junction between a horizontal line through the apex of the ear and a vertical line along the posterior border of the temple anterior to the ear.

For orientation: → **T.B.-20** is located directly superior to the apex of the ear.

Needling

0.3–1.5 cun transversely (subcutaneously) in the direction of the occiput or towards the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Benefits the mouth and jaw

Special features

Meeting point with the BL channel

G.B.-8 *Leading Valley* SHUAIGU

Location

1.5 cun directly superior to the apex of the ear, in a depression on the upper border of the temporalis muscle.

How to find

Locate the apex of the ear, which becomes more clearly defined by folding the auricle towards the anterior so that the posterior part of the upper helix covers its anterior part. From the apex of the ear, measure 1.5 cun (2 fingerbreadths) in a superior direction. There, the palpating finger will glide into a bony depression (**G.B.-8**) which is often sensitive to pressure. For orientation: When a chewing movement is made, this can just about be felt at this point.
→ **G.B.-9** is located on the same level, 0.5 cun dorsal to **G.B.-8**.

Needling

0.3–1.5 cun transversely (subcutaneously), mostly from anterior to posterior or towards the site of the pain.

Actions/Indications

- Expels Wind, benefits the head and ears, alleviates pain, harmonises the diaphragm and Stomach

Special features

Meeting point with the BL channel. Important local point for parietal and temporal headaches.

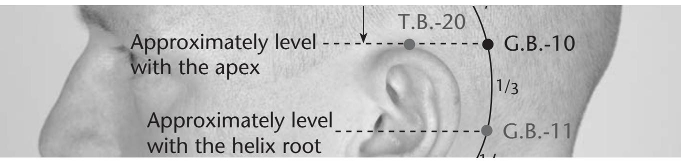
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11

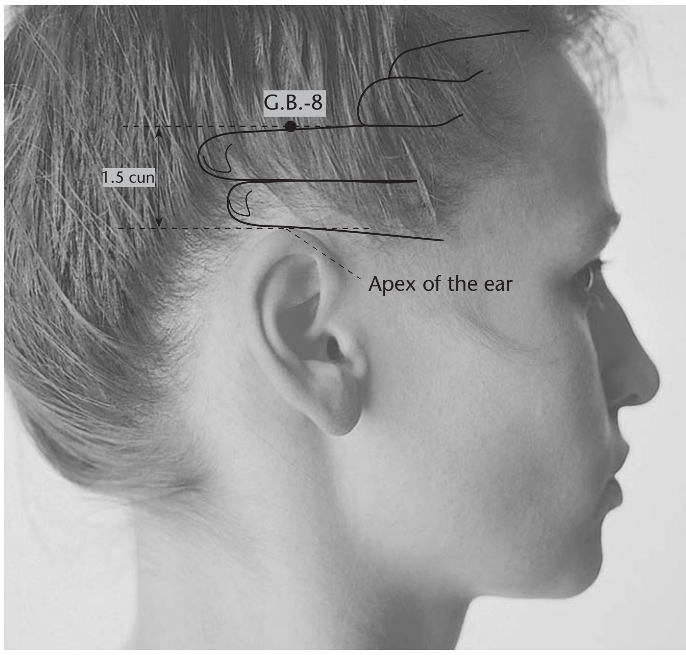
G.B.-8
1.5 cun
Apex of the ear

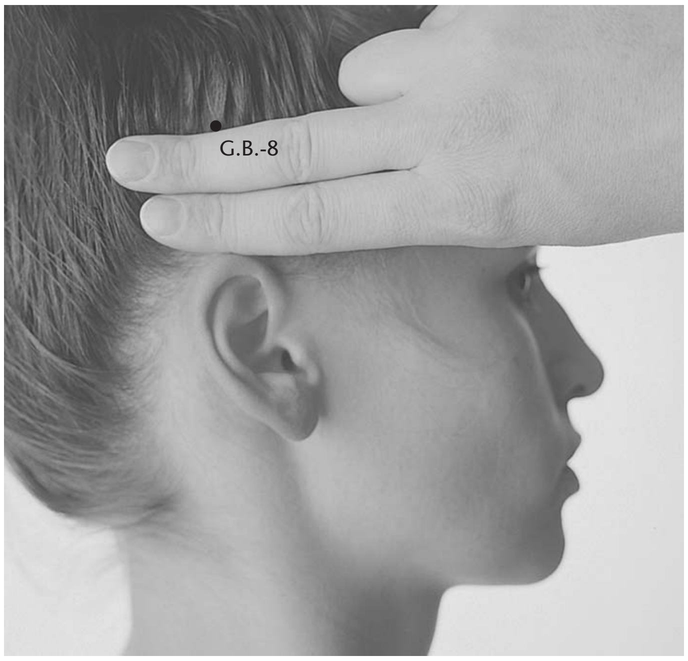

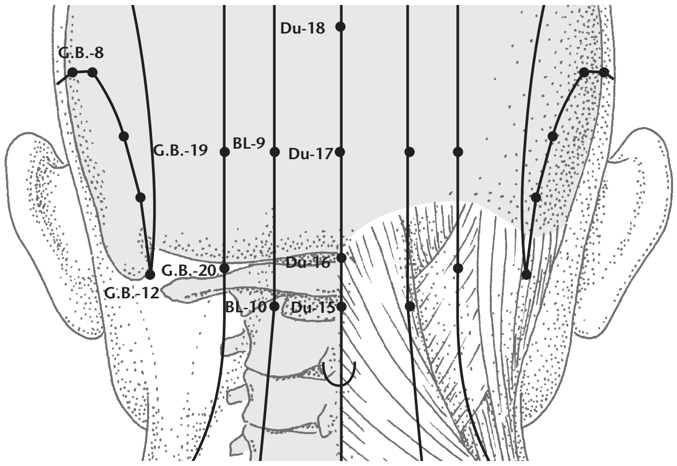
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

Heavenly Rushing TIANCHONG G.B.-9

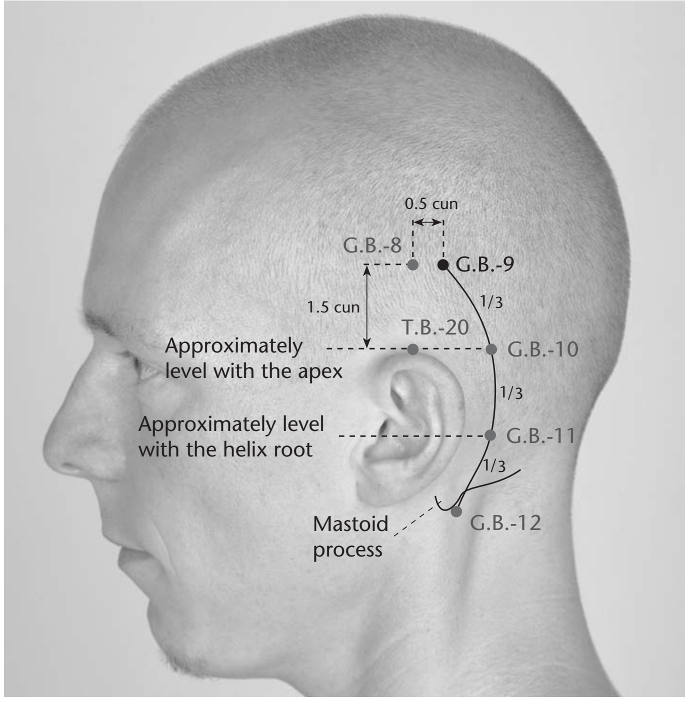
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
T.B.-20
Approximately level with the apex
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

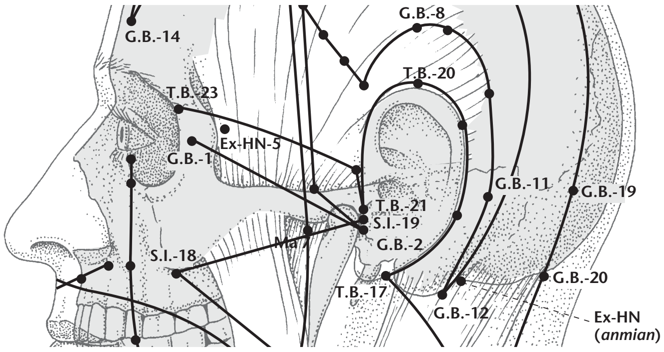
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma 7
T.B.-17
G.B.-12
G.B.-20
Ex-HN (*anmian*)

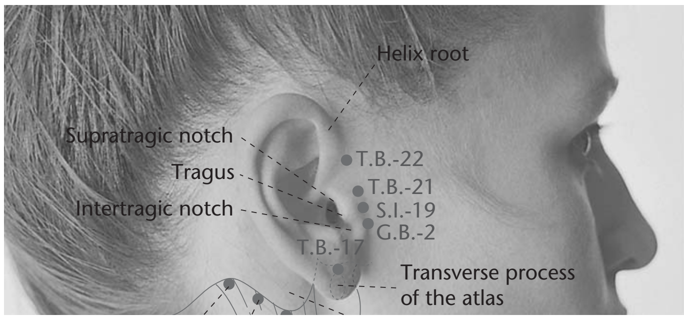
Helix root
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas

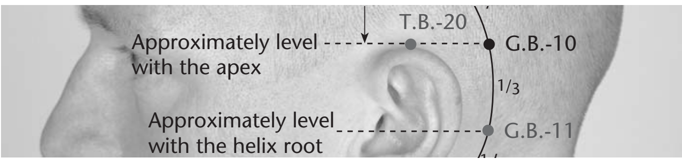
T.B.-20
Approximately level with the apex
G.B.-10
1/3
Approximately level with the helix root
G.B.-11

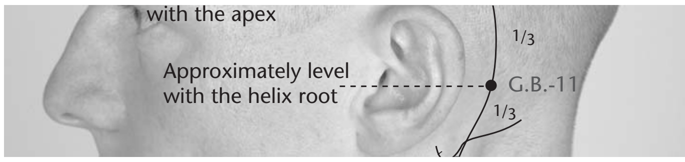
with the apex
1/3
Approximately level with the helix root
G.B.-11
1/3

Location

1.5 cun directly superior to the apex of the ear (→ **T.B.-20**) and 0.5 cun posterior to → **G.B.-8**. The point is approximately superior to the posterior margin of the ear.

How to find

First, locate → **G.B.-8** 1.5 cun superior to the apex of the ear. The latter becomes more clearly defined by folding the auricle towards the anterior so that the posterior part of the upper helix covers its anterior part. From → **G.B.-8**, measure 0.5 cun in a posterior direction and there locate **G.B.-9**. Similar to → **G.B.-8**, **G.B.-9** is located in a slight 'dip' in the bone.

Both **G.B.-9** and → **G.B.-12** (in the depression posterior and inferior to the mastoid process) are reference points for a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. When dividing this line into thirds, → **G.B.-10** and → **G.B.-11** are located at the junctions of the thirds.

Needling

0.5–1.5 cun transversely (subcutaneously) towards the occiput or the site of the pain.

Actions/Indications

- Expels Wind and Heat
- Opens the *luo*-connecting vessels
- Calms the *shen*

Special features

Meeting point with the BL channel; Window of Heaven point according to some authors.

G.B.-10 *Floating White* FUBAI

Location

Posterior to the ear, at the junction of the upper third with the two lower thirds of the curved line connecting → **G.B.-9** and → **G.B.-12**.

How to find

First, locate → **G.B.-9** (1.5 cun superior to the apex of the ear and 0.5 cun in a posterior direction). Then locate → **G.B.-12** in a depression posterior and inferior to the mastoid process (→ 3.1.4). These two points are the endpoints of a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. Divide this line into thirds and locate **G.B.-10** at the junction of the upper and the middle third, where often a small 'dip' can be felt on the bone. **G.B.-10** tends to be on the level of the apex of the ear.

→ **T.B.-20** is located directly above the apex of the ear.

Needling

Transversely (subcutaneously) 0.5–0.8 cun

Actions/Indications

- Expels Wind
- Opens the *luo*-connecting vessels

Special features

Meeting point with the BL channel

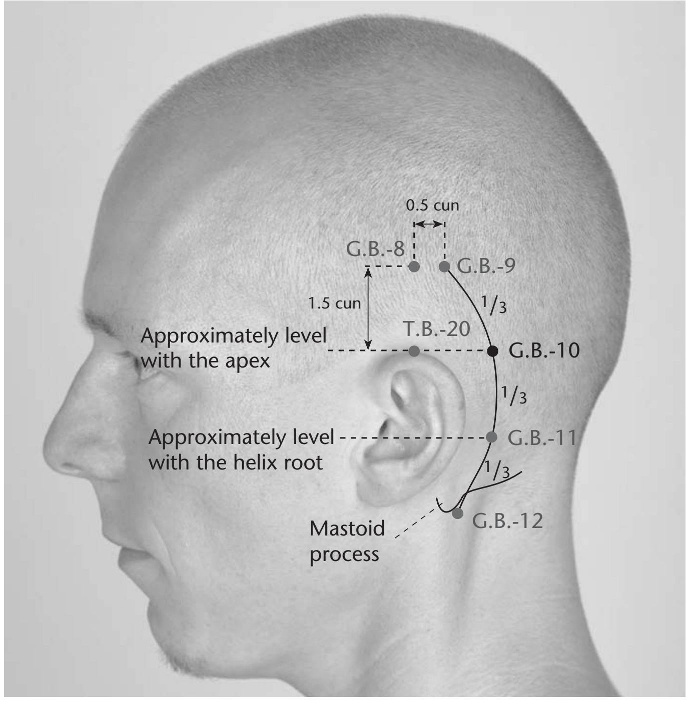
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

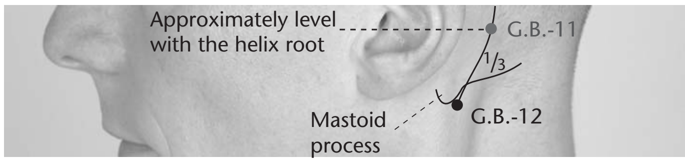
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

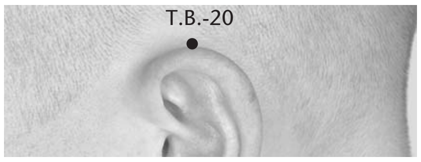

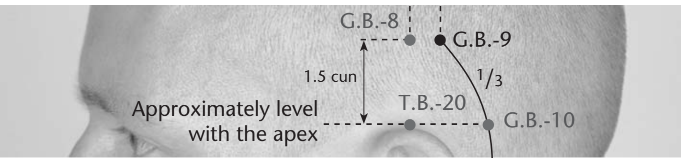
G.B.-8
G.B.-9
1.5 cun
1/3
Approximately level with the apex
T.B.-20
G.B.-10

Head Portal Yin TOUQIAOYIN G.B.-11

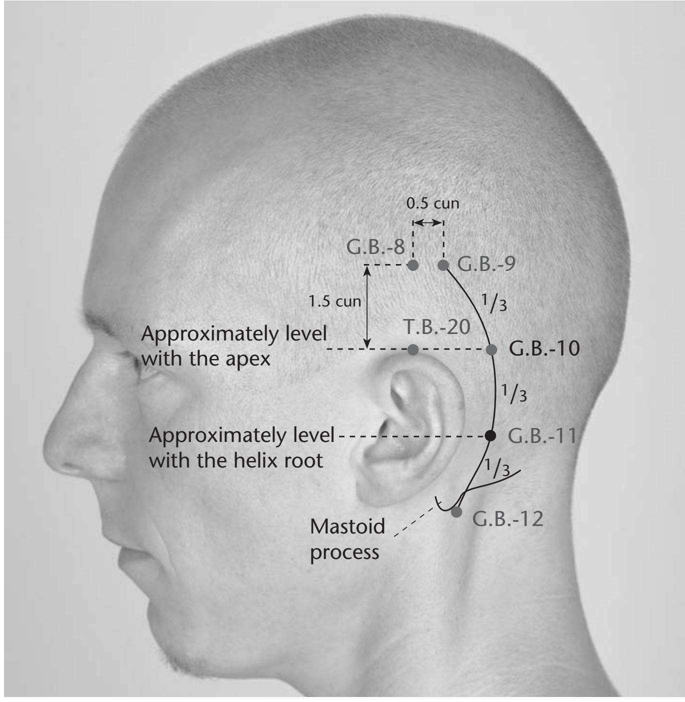
0.5 cun
G.B.-8
G.B.-9
1.5 cun
1/3
T.B.-20
Approximately level with the apex
G.B.-10
1/3
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

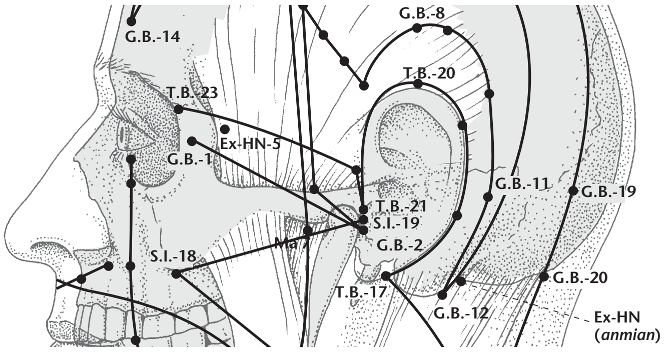
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
Ma 7
G.B.-2
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN (anmian)

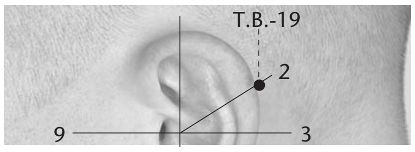

Location

Posterior to the ear, at the junction of the lower third with the two upper thirds of the curved line connecting → **G.B.-9** and → **G.B.-12**.

How to find

First, locate → **G.B.-9** (1.5 cun superior to the apex of the ear and 0.5 cun in a posterior direction). Then locate → **G.B.-12** in a depression posterior and inferior to the mastoid process (→ 3.1.4). These two points are the endpoints of a curved line that runs approximately parallel to the posterior margin of the ear within the hairline. Divide this line into thirds and locate **G.B.-11** at the junction of the middle and the lower third. At this point, a small 'dip' often can be felt in the bone. For reference: **G.B.-11** is located at the midpoint between a line connecting → **G.B.-10** and → **G.B.-12** and is generally on the level of the root of the helix.

→ **T.B.-19** is located on the same level, directly posterior to the helix.

Needling

Transversely (subcutaneously) 0.5–0.8 cun

Actions/Indications

- Expels Wind and Damp-Heat
- Opens the ears and the eyes
- Moves the (Liver) Qi and Blood

Special features

Meeting point with the BL channel, also with the S.I. and T.B. channels according to some authors.

G.B.-12 *Mastoid Process* WANGU

Location

In a depression directly posterior and inferior to the mastoid process.

How to find

First, locate the mastoid process posterior to the ear (→ 3.1.4). This can be palpated at the junction of the cranium and the neck as a cone-shaped, bony structure. With the palpating finger, find the inferior aspect of the mastoid process and locate **G.B.-12** on its lower border posterior to the tip of the process.

→ **T.B.-17** is located more anteriorly, in the depression posterior to the earlobe, between the mastoid process and the mandible.

Needling

0.5–1 cun obliquely in an inferior direction.

Actions/Indications

- Expels Wind, Heat and Dampness/Phlegm
- Opens the ears
- Calms the *shen*

Special features

Meeting point with the BL channel

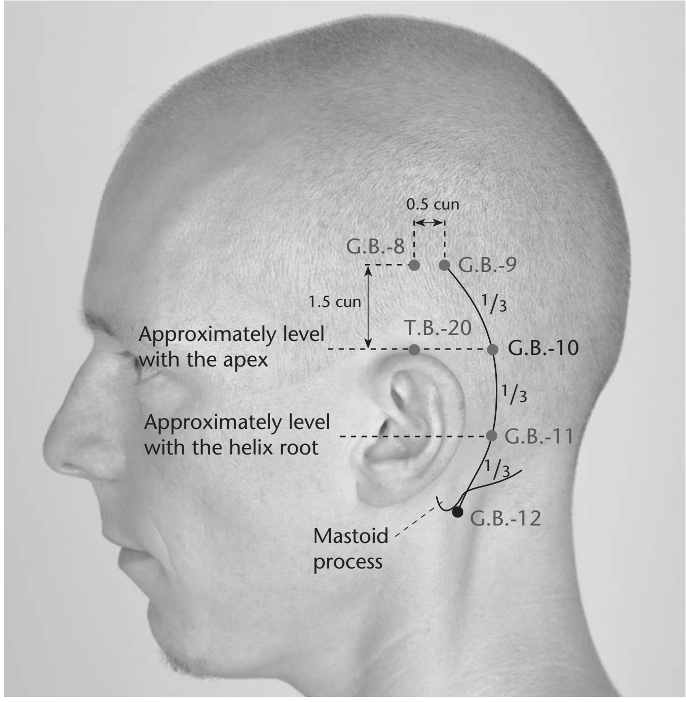
0.5 cun
G.B.-8
G.B.-9
1.5 cun
$^1/_3$
Approximately level with the apex
T.B.-20
G.B.-10
$^1/_3$
Approximately level with the helix root
G.B.-11
$^1/_3$
G.B.-12
Mastoid process

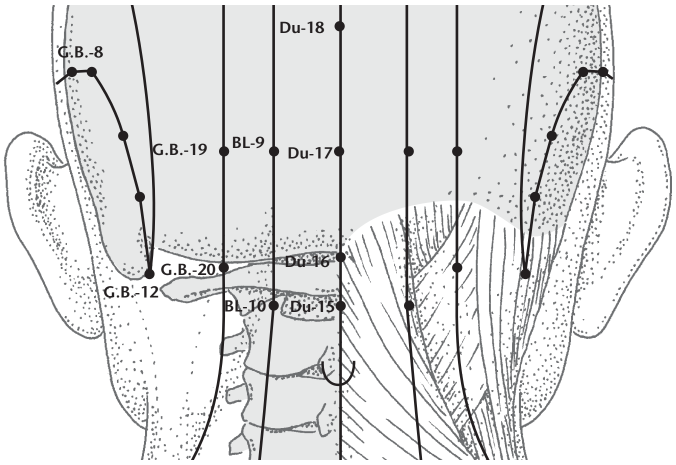
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

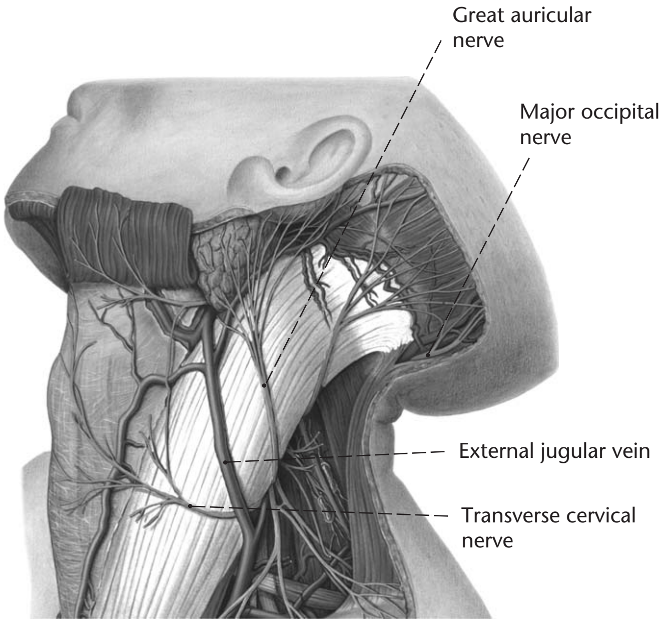
Great auricular nerve
Major occipital nerve
External jugular vein
Transverse cervical nerve

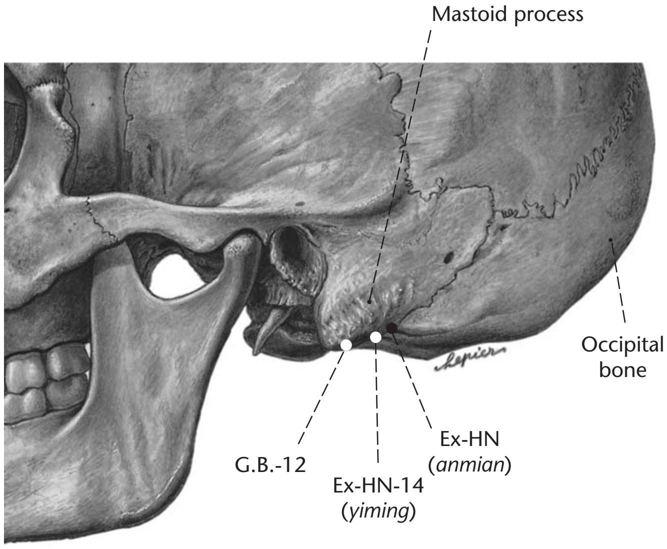
Mastoid process
Occipital bone
Ex-HN (*anmian*)
G.B.-12
Ex-HN-14 (*yiming*)

Spirit Root BENSHEN G.B.-13

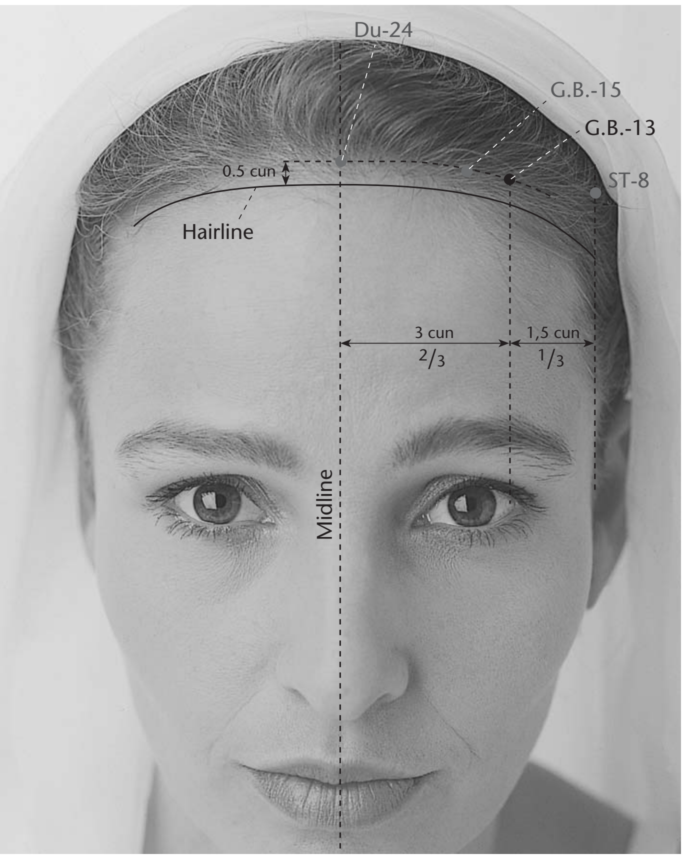
Du-24
G.B.-15
G.B.-13
0.5 cun
ST-8
Hairline
3 cun
2/3
1,5 cun
1/3
Midline

Location

3 cun lateral to → **Du-24** (on the midline, 0.5 cun superior to the anterior hairline).

How to find

First, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above) and there locate → **Du-24** on the midline and 0.5 cun within the hairline. **G.B.-13** is located on the same level, 3 cun lateral to the midline, measured in proportional cun based on the distance → **ST-8** to → **Du-24** (=4.5 cun, 2.2). Divide this distance into thirds and locate **G.B.-13** one third of the distance from **ST-8**. Located at the same level (0.5 cun within the anterior hairline) are → **Du-24/BL-3/BL-4/G.B.-15/ST-8** (on the midline/superior to the inner canthus of the eye/1.5 cun lateral to the midline/on the pupil line or 2.25 cun lateral to the midline/on the corner of the forehead).

Needling

0.5–0.8 cun transversely (subcutaneously) towards the occiput.

Actions/Indications

- Expels Wind
- Calms the *shen*
- Benefits the eyes

Special features

Meeting point with the *yang wei mai*

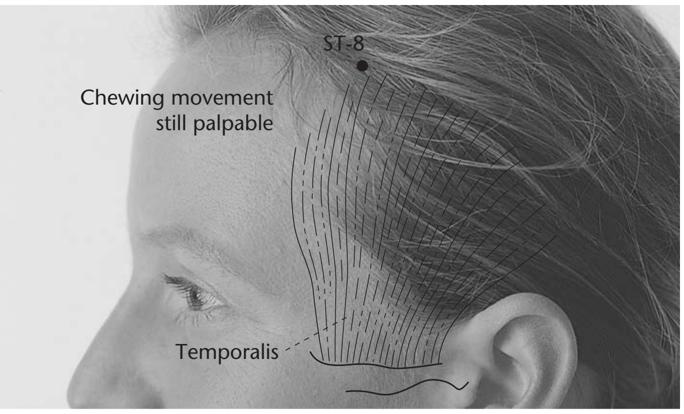
ST-8
Chewing movement
still palpable
Temporalis

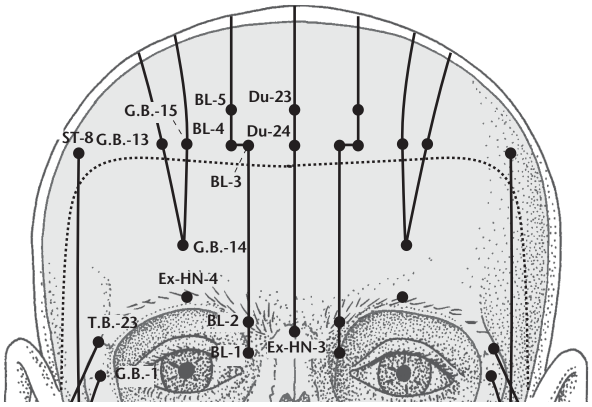
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1

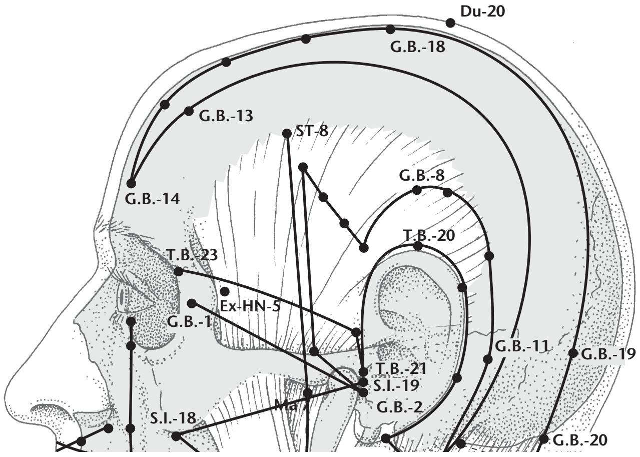
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma
S.I.-18
G.B.-20

G.B.-14 *Yang White* YANGBAI

Location

With the patient looking straight ahead, on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow, at the junction of the frontal eminence and the superciliary arch.

How to find

On the pupil line, palpate from the anterior hairline in an inferior direction, past the frontal eminence, and locate **G.B.-14** at the deepest point between the frontal eminence and the superciliary arch. The distance between the midpoint of the eyebrow and the anterior hairline (→ 3.1.1) is 3 proportional cun (→ 2.2). **G.B.-14** is located one third of the distance or 1 cun superior to the midpoint of the eyebrow. With the patient looking straight ahead, **G.B.-14** is located on the pupil line.

Needling

0.3–1 cun transversely (subcutaneously) towards the centre of the eyebrow or towards the site of the pain. The pinching-skin method might be used for needling: pinch the skin between the thumb and index finger so that a skin fold forms. Insert the needle subcutaneously into this fold, directing it towards the site of the pain.

Actions/Indications

- Expels (internal and external) Wind and Wind-Heat, benefits the head, alleviates pain
- Benefits the eyes

Special features

Meeting point with the *yang wei mai*, also with the T.B., L.I. and ST channels according to some authors. Important local point for frontal headache regardless of the pathology.

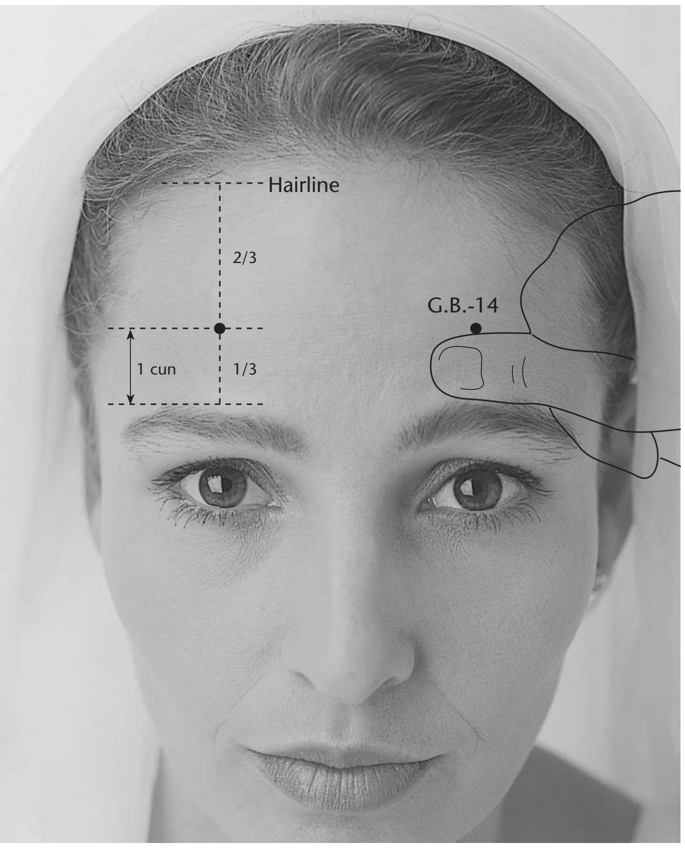
Hairline
2/3
G.B.-14
1 cun
1/3

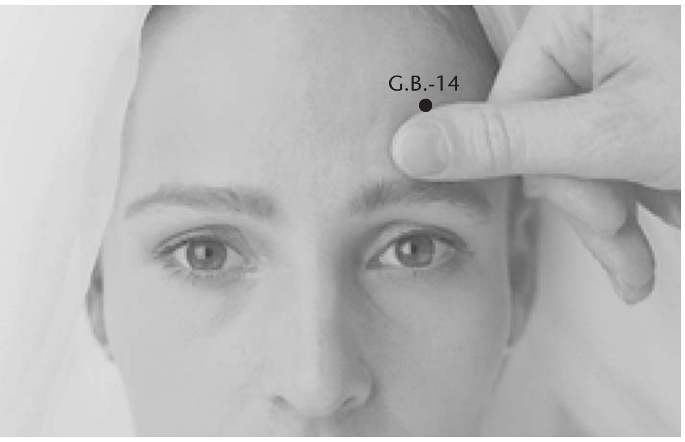

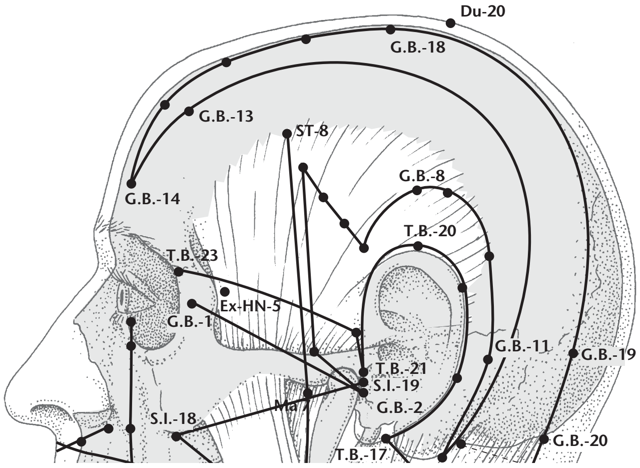
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
S.I.-18
Ma 7
T.B.-17
G.B.-20

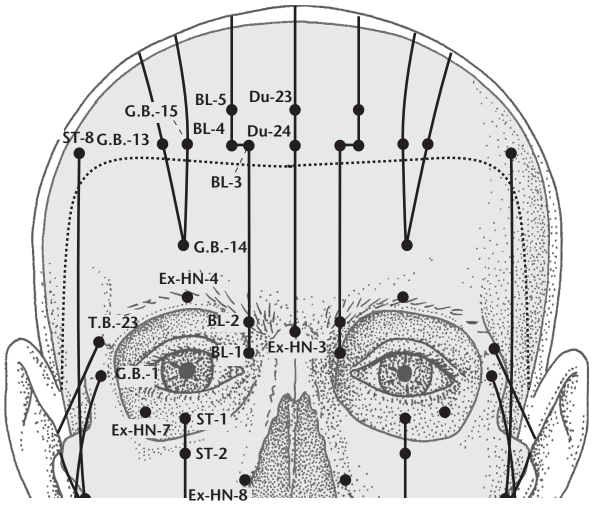
G.B.-15
BL-5
Du-23
BL-4
Du-24
ST-8
G.B.-13
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
Ex-HN-8

Head Governor of Tears TOULINQI G.B.-15

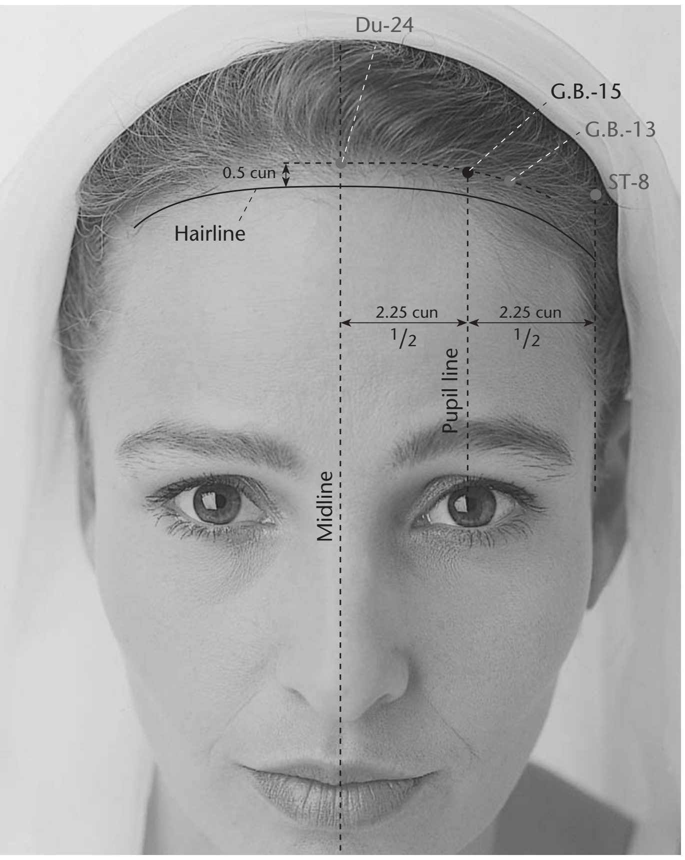
Du-24
G.B.-15
G.B.-13
ST-8
0.5 cun
Hairline
2.25 cun
1/2
2.25 cun
1/2
Pupil line
Midline

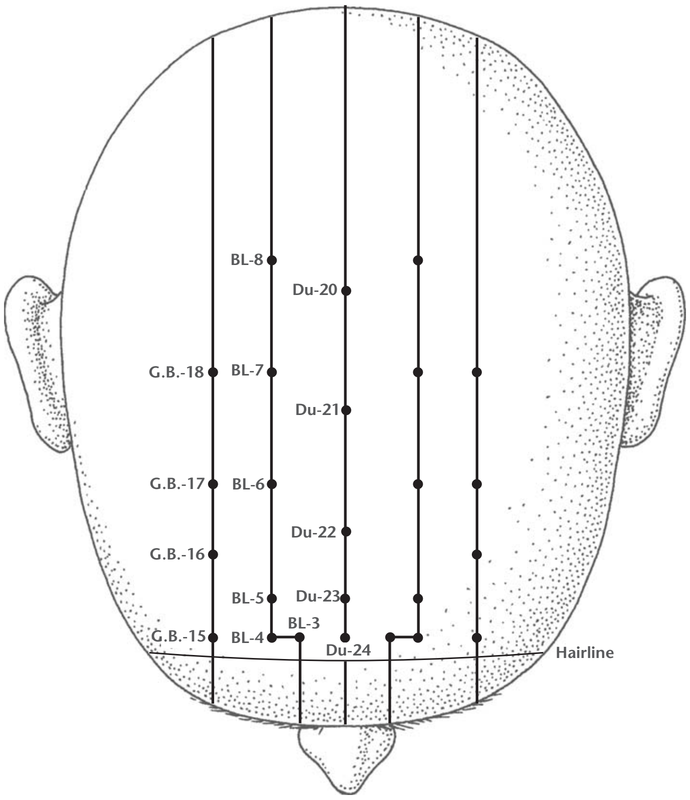
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Location

With the patient looking straight ahead, directly superior to the pupil and 0.5 cun superior to the anterior hairline.

How to find

First, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above). With the patient looking straight ahead, **G.B.-15** is located 0.5 cun superior to the anterior hairline and directly superior to the pupil. **G.B.-15** can also be described as being located midway between → **Du-24** (on the midline) and → **ST-8** (4.5 cun lateral to the midline, on the corner of the forehead).

Located on the same level (0.5 cun within the anterior hairline) are → **Du-24/BL-3/BL-4/G.B.-13/ST-8** (on the midline/superior to the inner canthus of the eye/1.5 cun lateral to the midline/3 cun lateral to the midline or one third of the distance from → **ST-8** (on the corner of the forehead) to → **Du-24**.

Needling

Transversely (subcutaneously) 0.3–0.5 cun

Actions/Indications

- Expels Wind
- Opens the eyes and nasal passages
- Calms the *shen*

Special features

Meeting point with the *yang wei mai*, also with the BL channel according to some authors.

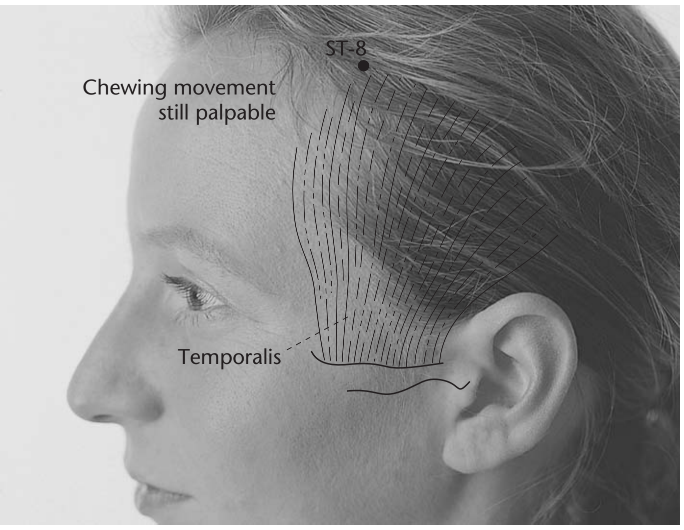
ST-8
Chewing movement still palpable
Temporalis

G.B.-16 *Window of the Eye* MUCHUANG

Location

1.5 cun superior to the anterior hairline, on the pupil line or 2.25 cun lateral to the midline (midway between → **Du-24** and → **ST-8**).

How to find

Locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above). With the patient looking straight ahead, **G.B.-16** is located 1.5 cun superior to the anterior hairline, on a vertical line through the centre of the pupil. This line is located 2.25 cun lateral to the midline (this corresponds to the midpoint of the distance between → **Du-24** and → **ST-8**). For further orientation: The distance between the anterior hairline and → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) is 5 cun. → **G.B.-16** is located 3.5 cun anterior to → **Du-20**, on the curved line connecting → **G.B.-15** and → **G.B.-20**.
→ **G.B.-15** is located on the same vertical line, but 0.5 cun superior to the anterior hairline.

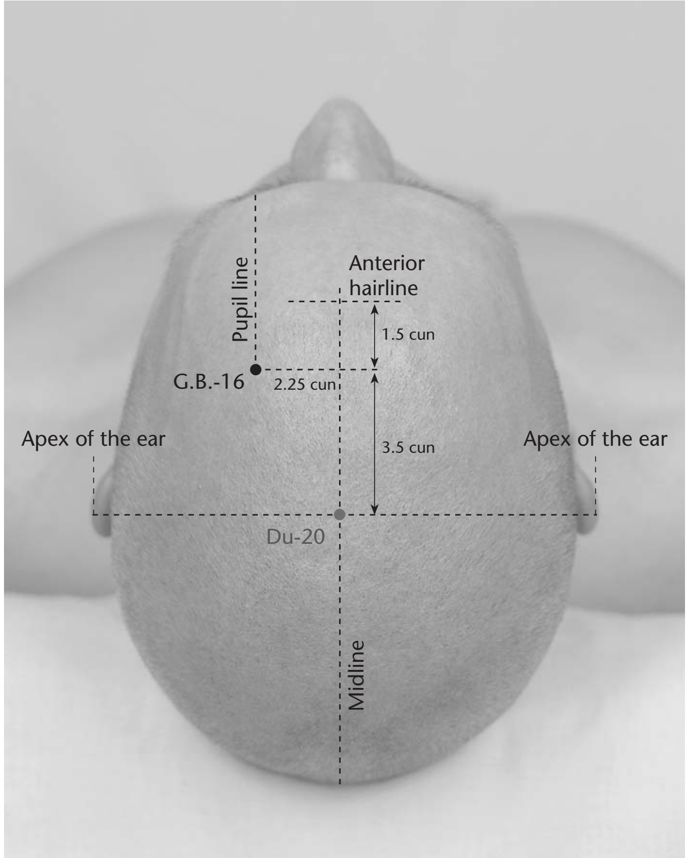
Pupil line
Anterior hairline
1.5 cun
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear
Du-20
Midline

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels pathogenic factors (especially Wind) from the head and eyes
- Opens the *luo*-connecting vessels

Special features

Meeting point with the *yang wei mai*

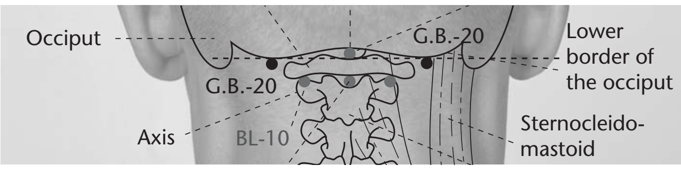
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

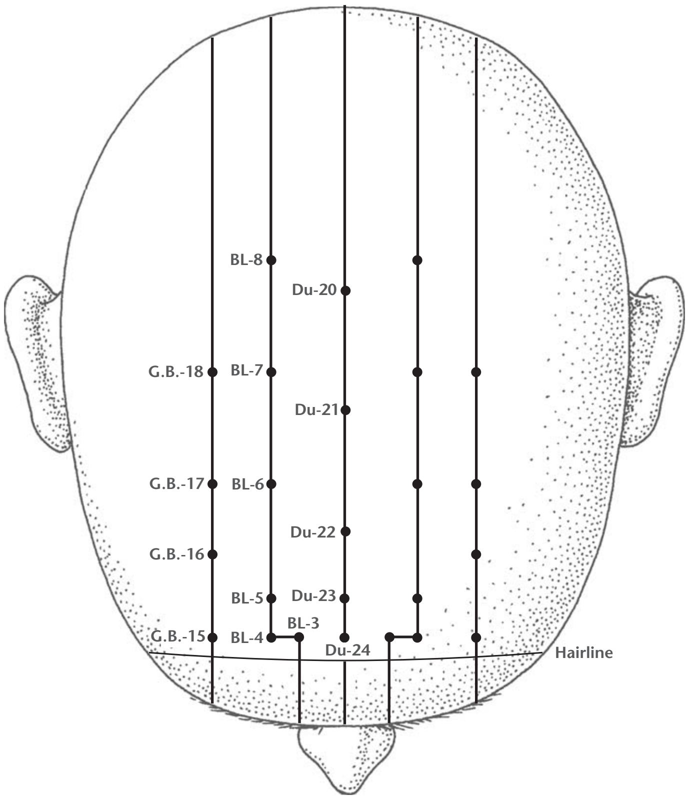
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Upright Nutrition ZHENGYING G.B.-17

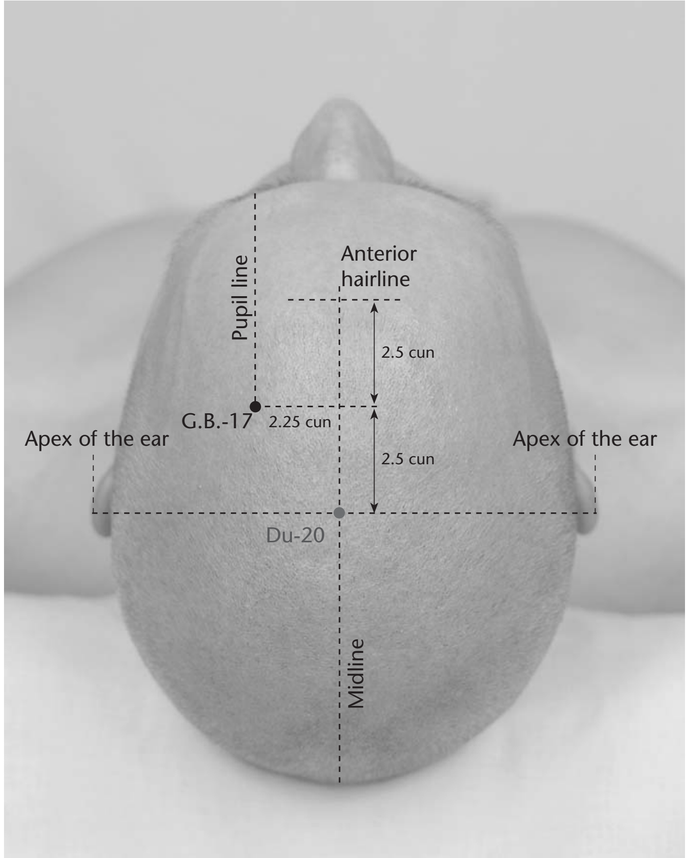
Pupil line
Anterior hairline
2.5 cun
G.B.-17
2.25 cun
Apex of the ear
2.5 cun
Apex of the ear
Du-20
Midline

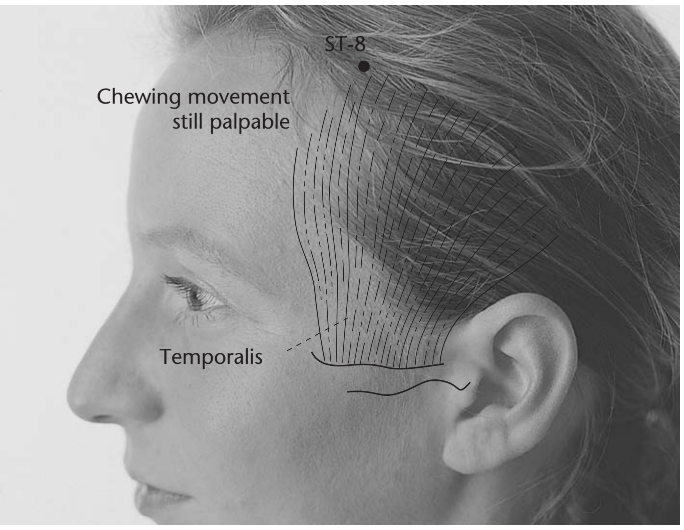
ST-8
Chewing movement still palpable
Temporalis

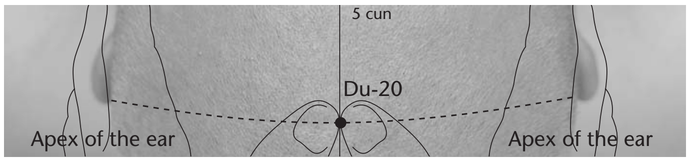
5 cun
Du-20
Apex of the ear
Apex of the ear

Location

2.5 cun superior to the anterior hairline and 2.25 cun lateral to the midline.

How to find

First, on the midline, locate the anterior hairline (→ 3.1.1, with hairloss: when the patient frowns, the anterior hairline is marked by the border between the creased forehead and the smooth skin above) and → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears). The distance between these two points is 5 proportional cun (→ 2.2). Locate the midpoint (for example by using the spreading hands technique → 2.3.3). **G.B.-17** is located on the curved line connecting → **G.B.-15** and → **G.B.-20** (an extension of the pupil line, 2.25 cun lateral to the midline).

→ **BL-6** is also located 2.5 cun anterior to → **Du-20**, but 1.5 cun lateral to the midline.

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Opens the *luo*-connecting vessels
- Descends counterflow Stomach Qi

Special features

Meeting point with the *yang wei mai*

Ch04.11-F10028.qxd 2/23/08 11:24 AM Page 416

G.B.-18 *Support Spirit* CHENGLING

Location

4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20** and 2.25 cun lateral to the midline.

How to find

First, locate → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears, distance to the anterior hairline = 5 cun). Next, locate **G.B.-18** 1 cun anterior to → **Du-20** on the curved line connecting → **G.B.-15** and → **G.B.-20** (= extension of the pupil line, midway between → **ST-8** and → **Du-24**).
→ **BL-7** is located on the same level (1 cun anterior to → **Du-20**), but 1.5 cun lateral to the midline.

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind, especially from the nose

Special features

Meeting point with the *yang wei mai*

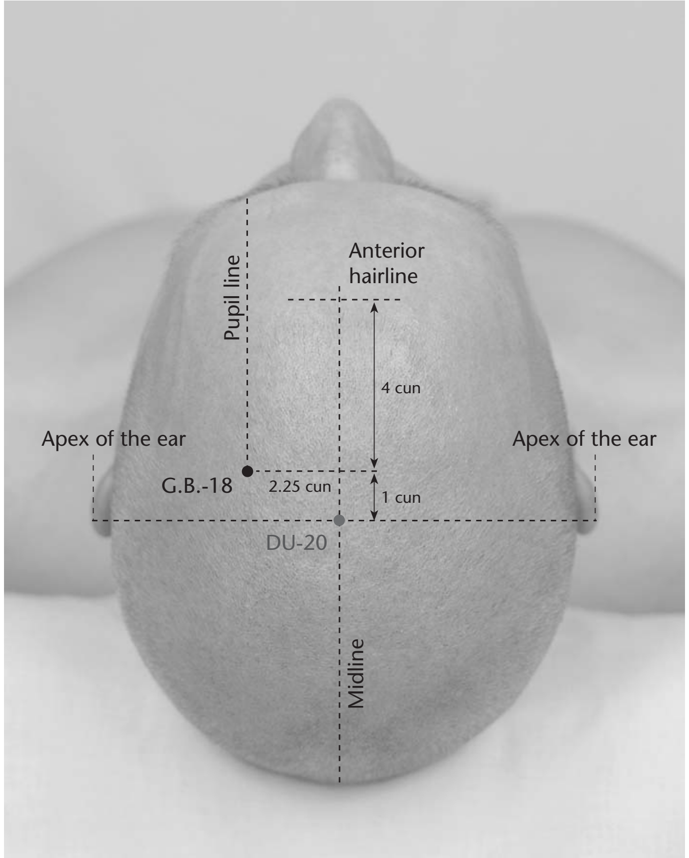
Pupil line
Anterior hairline
4 cun
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
DU-20
Midline

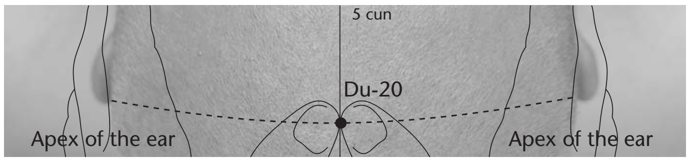
5 cun
Du-20
Apex of the ear
Apex of the ear

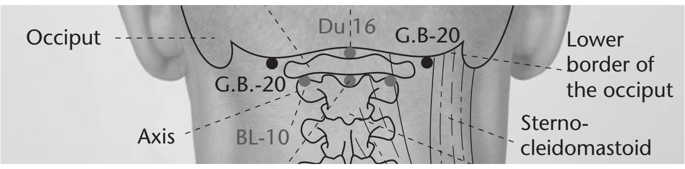
Occiput
Du 16
G.B-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sterno-cleidomastoid

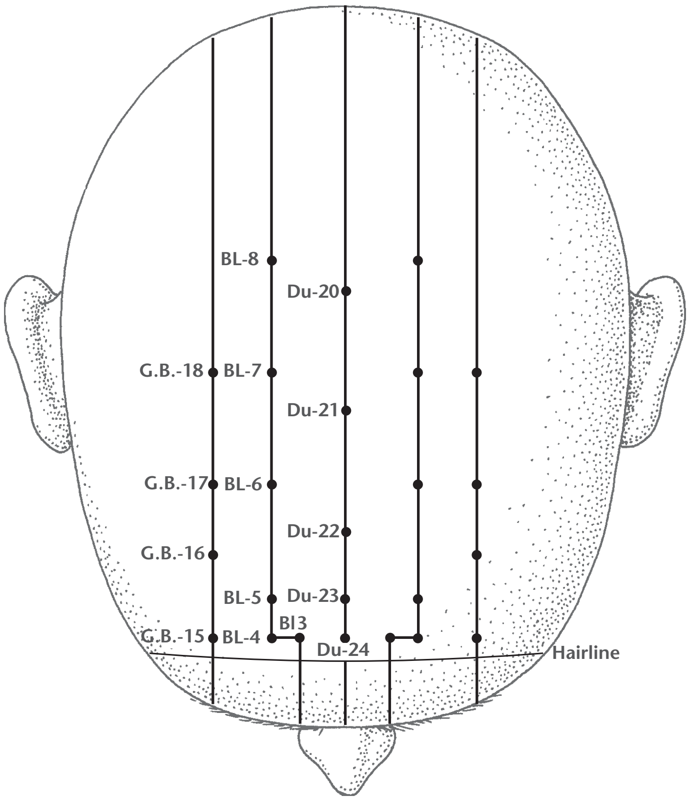
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
Bl 3
G.B.-15
BL-4
Du-24
Hairline

Brain Hollow NAOKONG G.B.-19

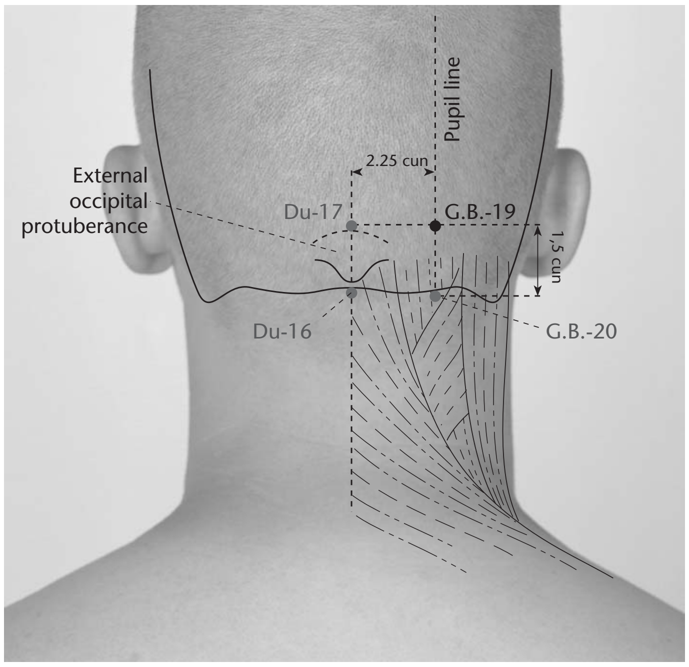
Pupil line
2.25 cun
External occipital protuberance
Du-17
G.B.-19
1.5 cun
Du-16
G.B.-20

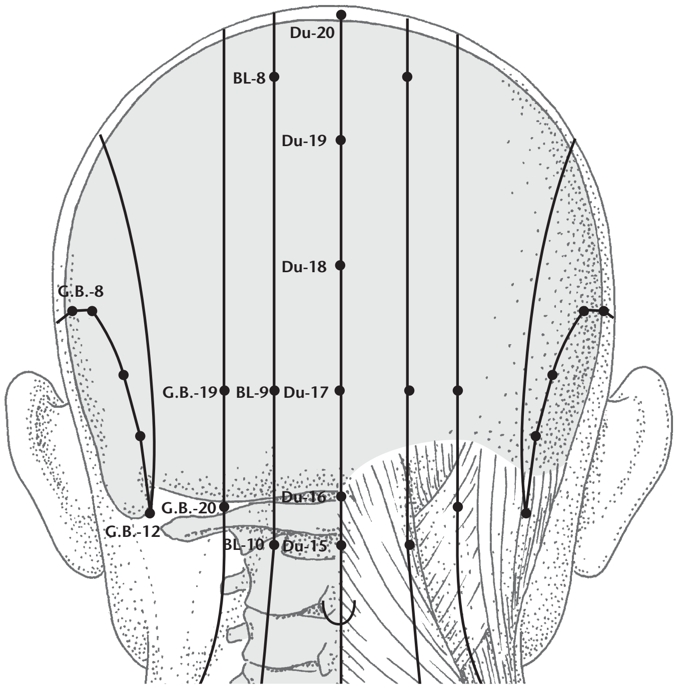
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

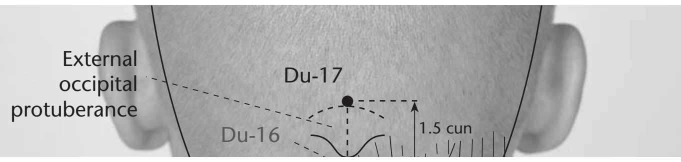
External occipital protuberance
Du-17
Du-16
1.5 cun

Location

On the posterior aspect of the head, at the upper border of the external occipital protuberance (→ **Du-17**) and 2.25 cun lateral to the midline.

How to find

First, locate the external occipital protuberance (→ 3.1.5), which can be palpated as a flat protruding area on the midline. Next, find → **Du-17** in the depressoin just superior to the protuberance. From there, measure 2.25 cun in a lateral direction and locate **G.B.-19** on the curved line connecting → **G.B.-15** and → **G.B.-20**, an extension of the pupil line (→ 2.25 cun lateral to the midline or midway between → **Du-24** and → **ST-8**). **For orientation: G.B.-19** is located approximately 2.5 cun superior to the posterior hairline (→ 3.1.5) and approximately 1.5 cun superior to → **G.B.-20** (on the lower border of the occiput, between the origins of the sternocleidomastoid and trapezius muscles).

Needling

0.3–1.5 cun transversely (subcutaneously) towards the occiput or the site of the disorder/pain.

Actions/Indications

- Expels Wind
- Clears the sensory orifices
- Opens the *luo*-connecting vessels

Special features

Meeting point with the *yang wei mai*

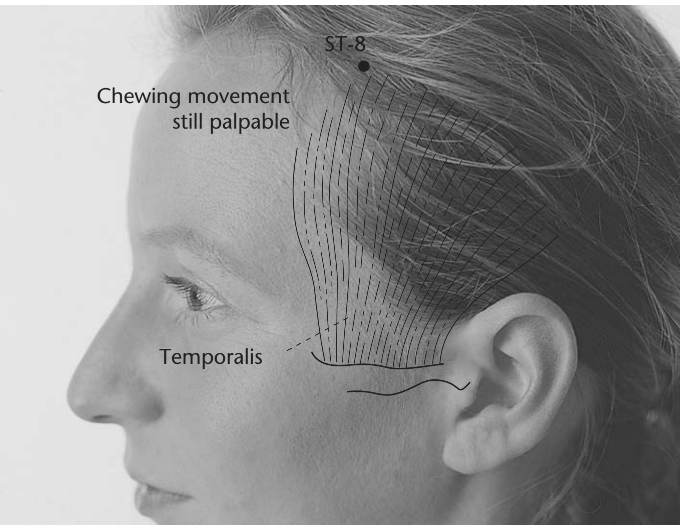
ST-8
Chewing movement still palpable
Temporalis

G.B.-20 *Wind Pool* FENGCHI

Location

At the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius muscles.

How to find

Patient's position: Prone, sitting or supine (supported by a pillow, so that the occipital region is accessible). Starting at the midline, glide with the palpating finger along the lower border of the occiput, crossing the bulge of the origin of the trapezius muscle, until you reach a depression the size of a finger pad. Locate **G.B.-20** in its centre.

Located on the same level is → **Du-16**. → **BL-10** is located more medially and inferiorly.

Needling

Needle with the patient's head bent forward and the tip of the needle pointing inferiorly. Depending on the position of the head, insertion towards the tip of the nose or the contralateral orbit, 0.5–1.2 cun. Caution: In slim patients, do not needle deeper than 2 cm (the vertebral artery is located at a depth of approximately 4 cm).

Actions/Indications

- Eliminates Wind, benefits the head, clears the sensory organs
- Opens the channel
- With tonifying needle technique: strengthens the Marrow and the Brain (according to Maciocia)

Special features

Meeting point with the T.B. channel*, *yang wei mai, yang qiao mai*. Major point for all 'Wind disorders', very important point for disorders of the head and eyes.

---

*According to Deadman et al 1998.

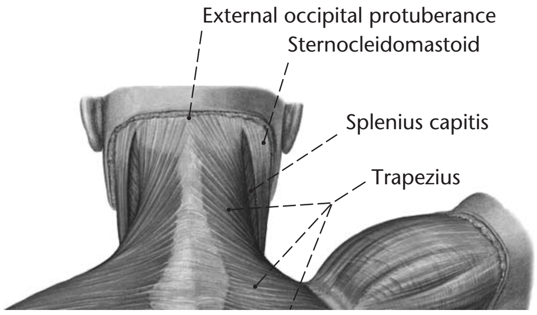
External occipital protuberance
Sternocleidomastoid
Splenius capitis
Trapezius

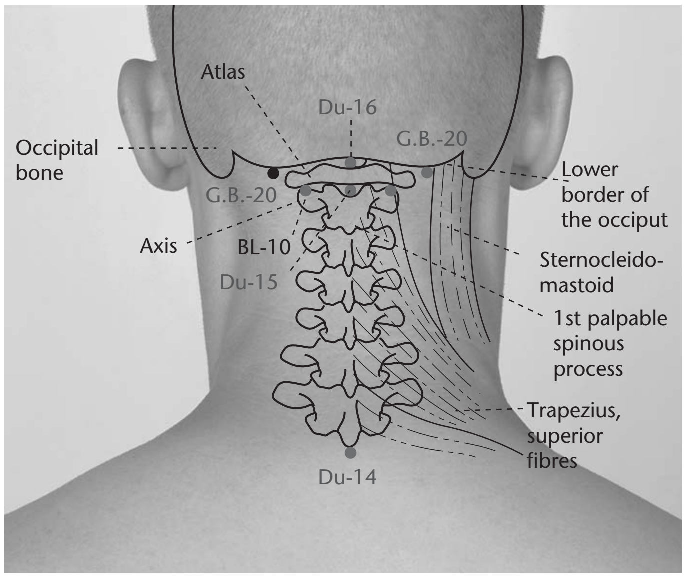
Atlas
Du-16
G.B.-20
Occipital bone
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

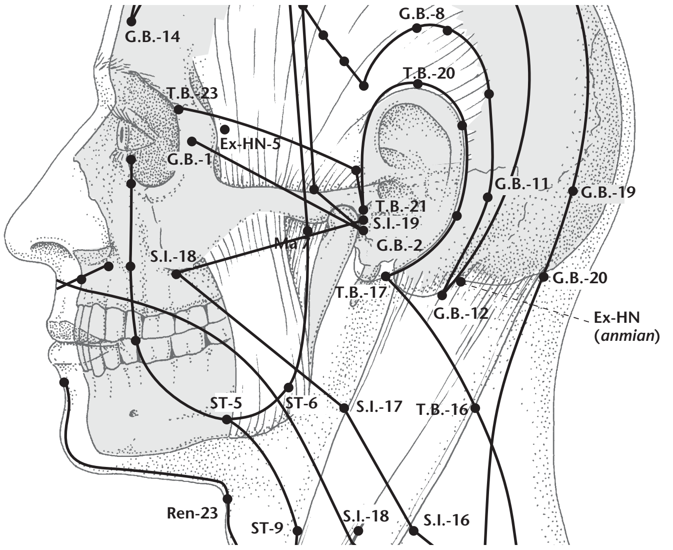
G.B.-8
G.B.-14
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
S.I.-18
S.I.-16

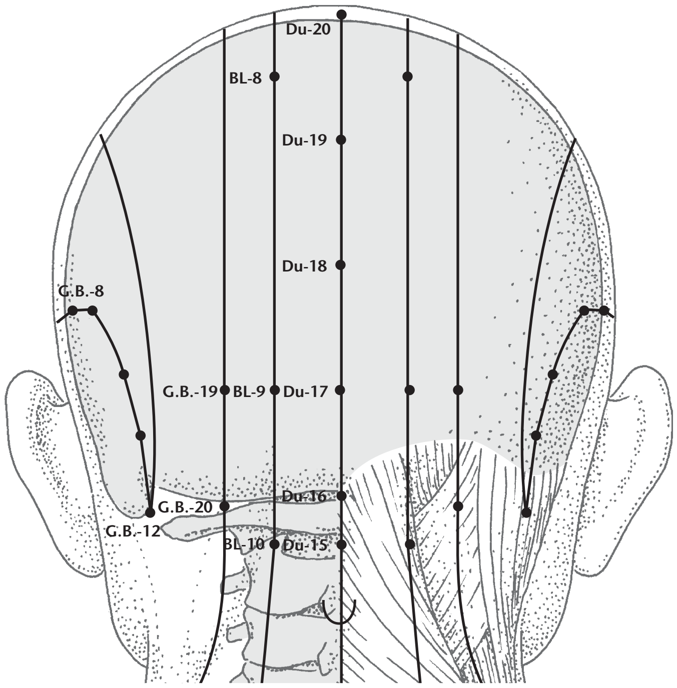
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

Shoulder Well JIANJING G.B.-21

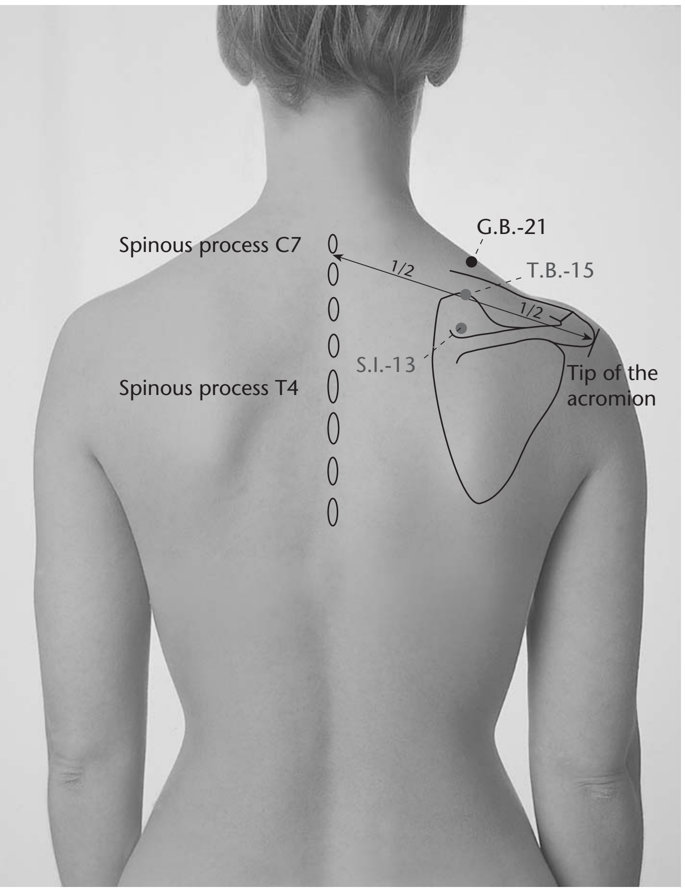

Location

At the highest point of the shoulder, at the midpoint of a line connecting the 7th cervical vertebra (C7) and the lateral extremity of the acromion.

How to find

First, locate C7 (→ 3.4.1) and the lateral extremity of the acromion (→ 3.3.1). Next, locate **G.B.-21** at the midpoint of a line connecting these two reference points, on the highest point of the trapezius muscle on the sagittal plane.
→ **T.B.-15** is located 1 cun inferior to **G.B.-21**, midway between **G.B.-21** and → **S.I.-13** (medial to the supraspinous fossa).

Needling

Vertically approximately 0.3–0.5 cun (lift the muscle). An alternative, safer method: Lift the muscle and insert the needle approximately 1 cun anteriorly or posteriorly into the muscle belly. Caution: Contraindicated during pregnancy! Pneumothorax.

Actions/Indications

- Opens the channel
- Descends the Qi
- Regulates the Qi flow, transforms Phlegm, dissipates nodules
- Promotes labour, benefits the breasts

Special features

Meeting point with the *yang wei mai* and the T.B. and ST* channels. Important local point. Trigger point for the shoulder with a distal effect on the Uterus.

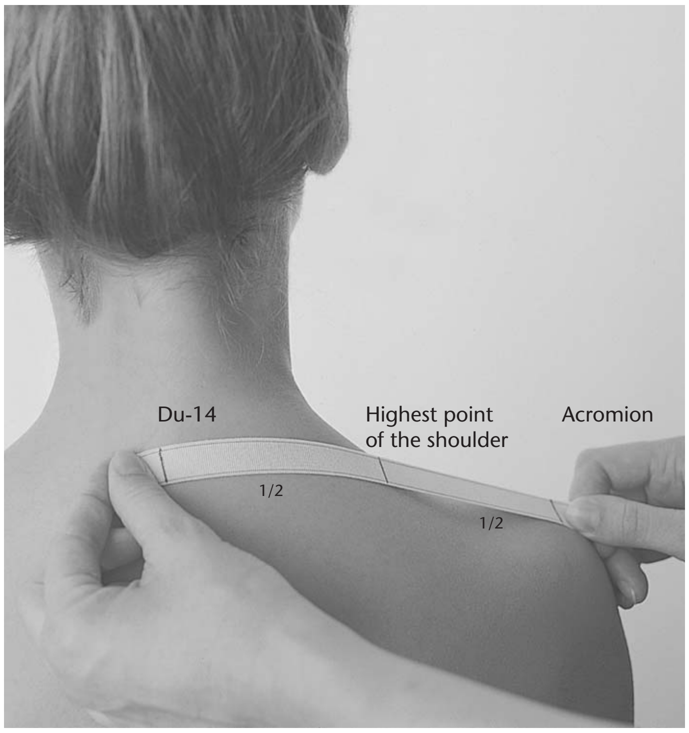

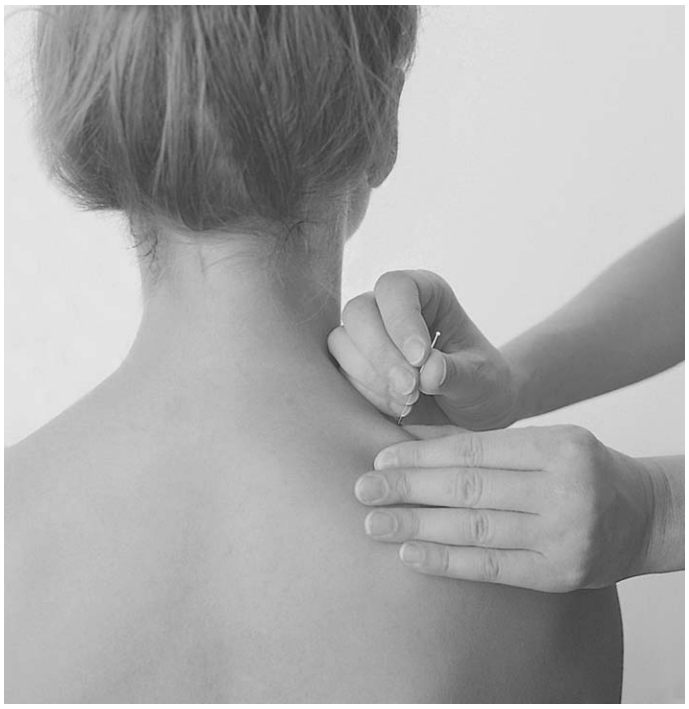

---

*According to Deadman et al 1998.

G.B.-22 *Armpit Abyss* YUANYE

Location

On the midaxillary line, with the arm abducted approximately 3 cun inferior to the apex of the axilla, in the 4th intercostal space (according to some authors, in the 5th intercostal space).

How to find

With the patient's arm slightly abducted, locate the apex of the axilla. Locate **G.B.-22** 3 cun inferior to it in the 4th intercostal space. **For orientation:** In men, the nipple is located in the 4th intercostal space; in women, in a supine position, its location may vary. Therefore, in women, the manubriosternal synchondrosis (→ 3.5) is a more reliable reference point for the costal region. Note: The intercostal space curves in a superior direction towards lateral.

Also located on the level of or in the 4th intercostal space, but more medially, are → **Ren-17/KID-23/ST-17/P-1/SP-18/G.B.-23** (on the midline/2 cun lateral to the midline/in the centre of the nipple or 4 cun lateral to the midline/1 cun lateral to the nipple or 5 cun lateral to the midline/6 cun lateral to the midline/1 cun anterior to **G.B.-22**).

Needling

0.5–1 cun obliquely or transversely (subcutaneously) along the intercostal space. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi in the Upper Burner
- Opens the *luo*-connecting vessels of the axilla

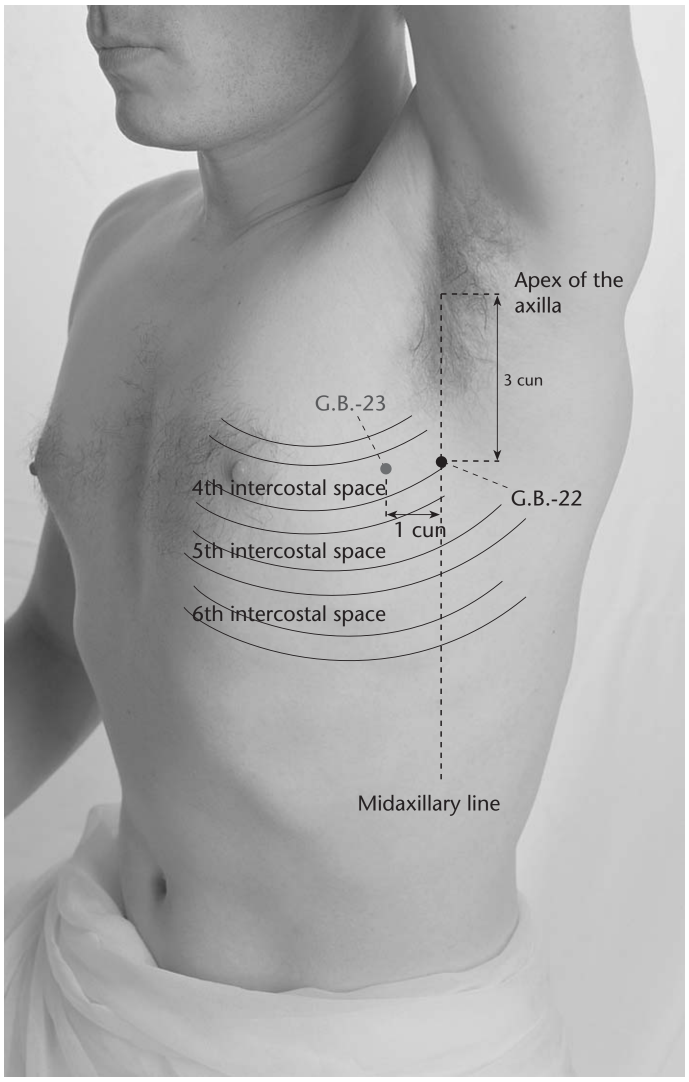
Apex of the axilla
3 cun
G.B.-23
4th intercostal space
G.B.-22
1 cun
5th intercostal space
6th intercostal space
Midaxillary line

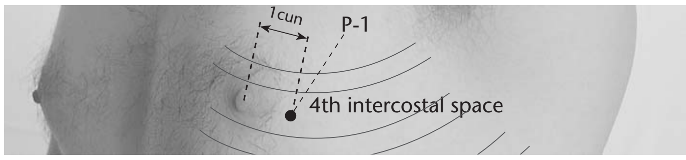
1cun
P-1
4th intercostal space

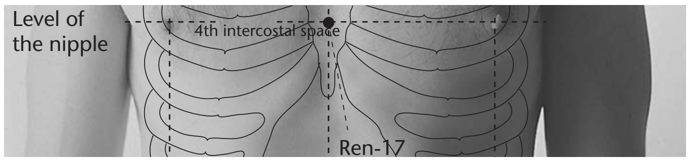
Level of the nipple
4th intercostal space
Ren-17

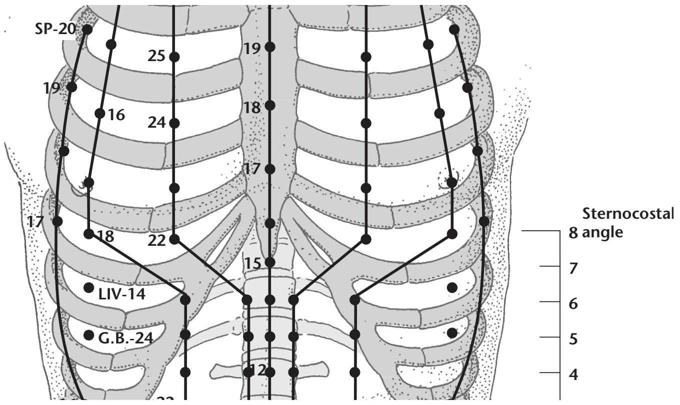
SP-20
19
25
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
Sternocostal angle
8
7
6
5
4

Flank Sinews ZHEJIN G.B.-23

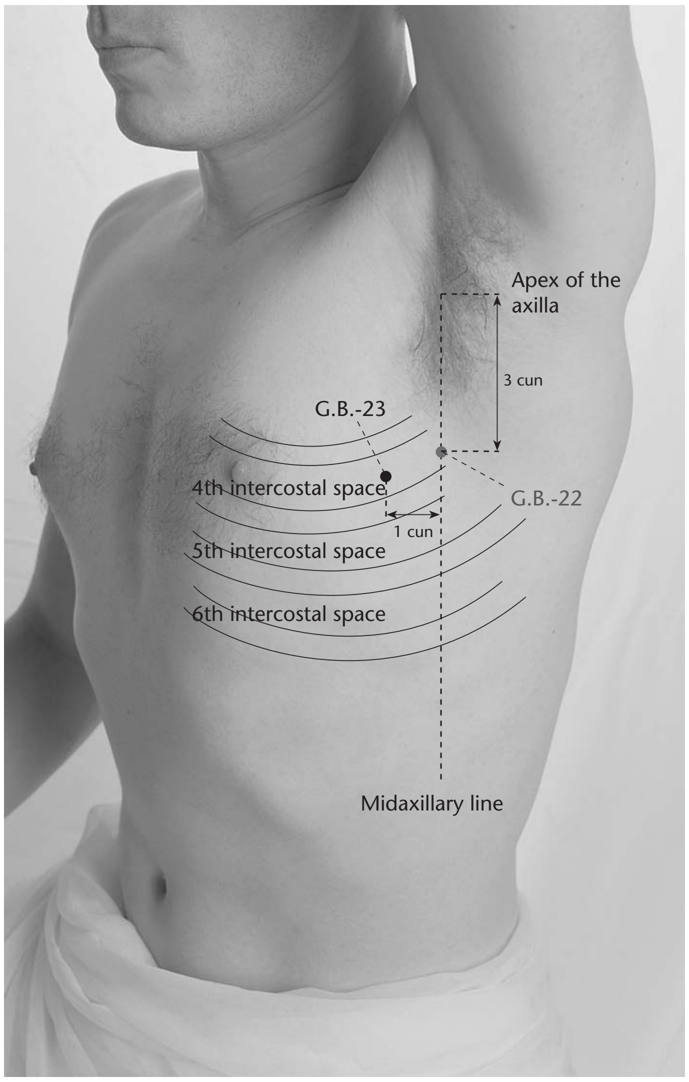
Apex of the axilla
3 cun
G.B.-23
4th intercostal space
1 cun
G.B.-22
5th intercostal space
6th intercostal space
Midaxillary line

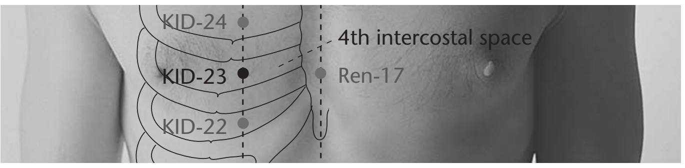
KID-24
KID-23
KID-22
4th intercostal space
Ren-17

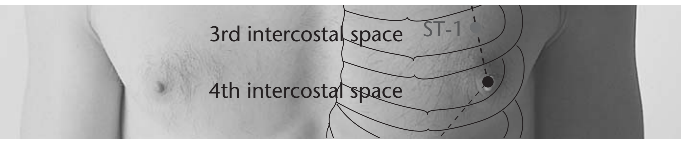
3rd intercostal space
ST-1
4th intercostal space

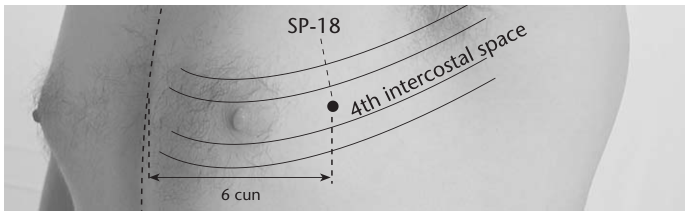
SP-18
4th intercostal space
6 cun

Location

1 cun anterior to → **G.B.-22** (on the midaxillary line, 3 cun inferior to the apex of the axilla in the 4th intercostal space).

How to find

With the patient's arm slightly abducted, locate the apex of the axilla. First, locate → **G.B.-22** 3 cun inferior to it in the 4th intercostal space, then locate **G.B.-23** 1 cun anterior to → **G.B.-22** and also in the 4th intercostal space. **For orientation:** In men, the nipple is located in the 4th intercostal space; in women, in a supine position, its location may vary. Therefore, in women, the manubriosternal synchondrosis (→ 3.5) is a more reliable reference point for the costal region. Note: The intercostal space curves in a superior direction towards lateral.

Also located on the level of or in the 4th intercostal space, but more medially, are → **Ren-17/KID-23/ST-17/P-1/SP-18** (on the midline/2 cun lateral to the midline/in the centre of the nipple or 4 cun lateral to the midline/1 cun lateral to the nipple or 5 cun lateral to the midline/6 cun lateral to the midline).

Needling

0.5–1 cun obliquely or transversely (subcutaneously) along the 4th intercostal space. Caution: Pneumothorax.

Actions/Indications

- Regulates the Qi between the Upper and Middle Burners
- Opens the channel

Special features

Meeting point with the BL channel according to some authors

G.B.-24 *Sun and Moon* RIYUE

Location

In the 7th intercostal space, on the mamillary line (4 cun lateral to the anterior midline).

How to find

The mamillary line, a vertical line 4 cun lateral to the midline, is used for reference in the thoracic region (→ 3.5). **Quick method, especially in men:** In men, the nipple is usually located on the level of the 4th intercostal space. From the nipple, count downward to the 7th intercostal space. **Or:** A more reliable reference, especially in women, is the manubriosternal synchondrosis, a horizontal bony structure on the sternum. The costal cartilage of the second rib is lateral to the synchondrosis, with the 2nd intercostal space below. From there, count downward to the 7th intercostal space and locate **G.B.-24** on the mamillary line. → **LIV-14** is located directly superior to **G.B.-24** in the 6th intercostal space.

Located approximately on the same level (1 cun inferior to the sternocostal angle) are → **Ren-13/KID-20/ST-20** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

0.3–0.8 cun obliquely in a lateral direction along the intercostal space. Caution: Pneumothorax.

Actions/Indications

- Benefits the Gall Bladder, eliminates Damp-Heat, regulates and spreads Liver Qi, descends counterflow Qi, harmonises the Middle Burner

Special features

Front-*mu* point of the Gall Bladder, meeting point with the SP channel, also with the *yang wei mai* according to some authors. Major point for disorders of the Gall Bladder.

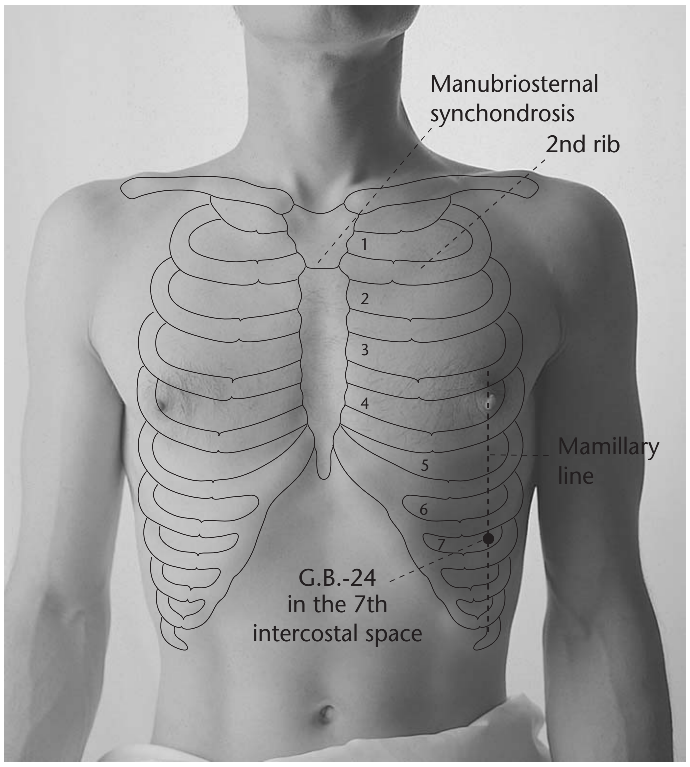
Manubriosternal synchondrosis
2nd rib
1
2
3
4
5
6
7
Mamillary line
G.B.-24 in the 7th intercostal space

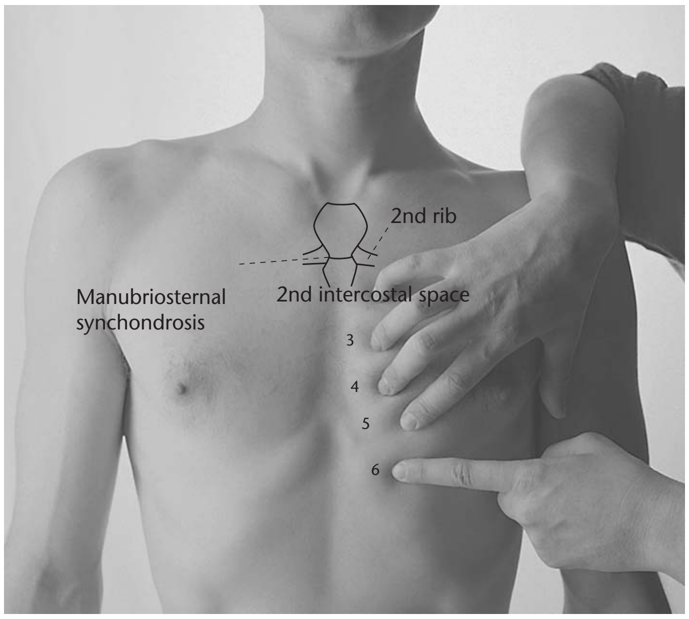
2nd rib
Manubriosternal synchondrosis
2nd intercostal space
3
4
5
6

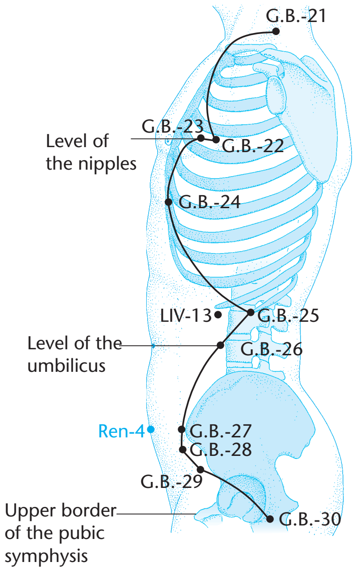
G.B.-21
G.B.-23
Level of the nipples
G.B.-22
G.B.-24
LIV-13
G.B.-25
Level of the umbilicus
G.B.-26
Ren-4
G.B.-27
G.B.-28
G.B.-29
Upper border of the pubic symphysis
G.B.-30

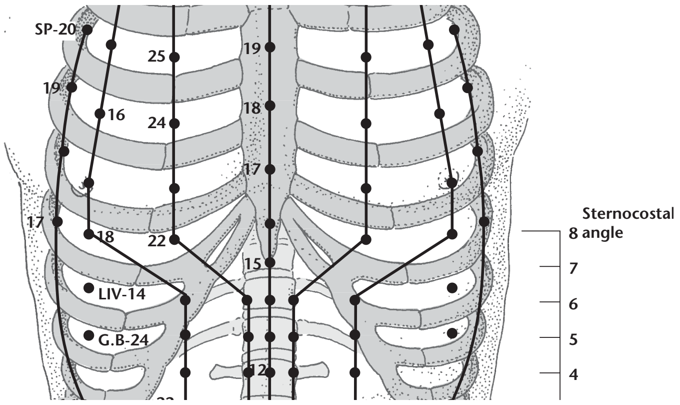
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B-24
12
Sternocostal angle
8
7
6
5
4

Capital Gate JINGMEN G.B.-25

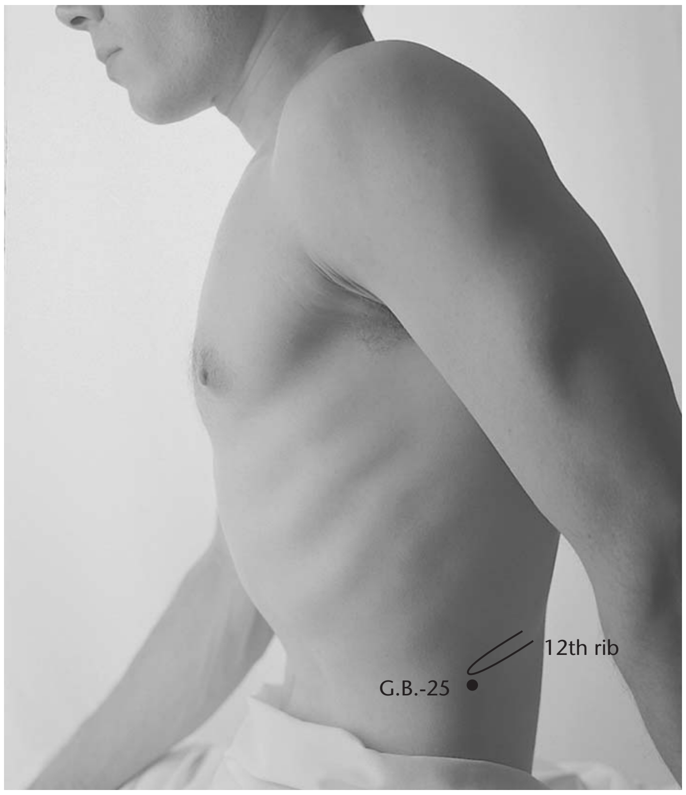

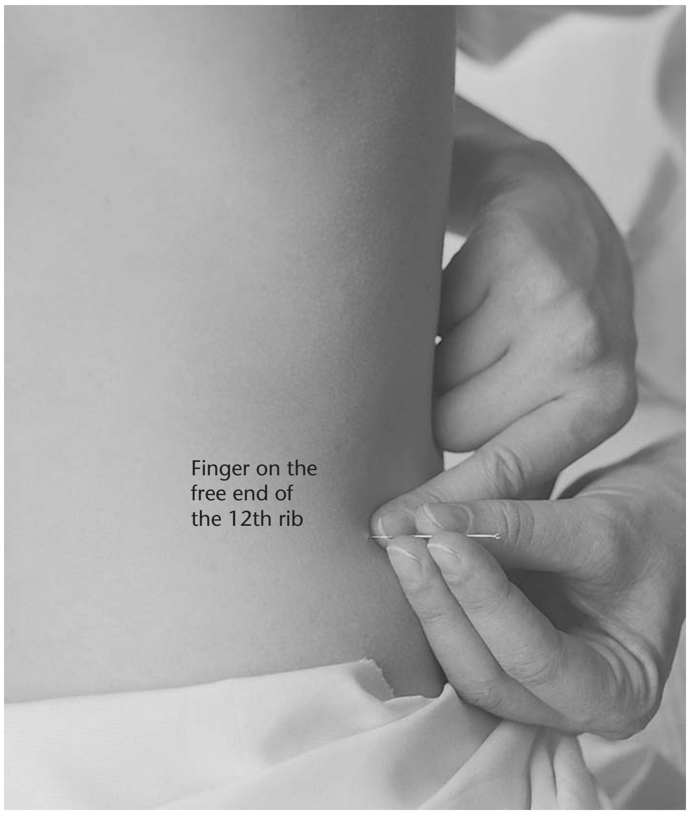

Location

On the lateral aspect of the ribcage, at the lower border of the free end of the 12th rib.

How to find

Place the hand on the upper abdomen and, with gentle pressure, glide along the lower border of the ribcage until you can feel the free end of the 11th rib (→ **LIV-13**) slightly superior to the umbilicus. By continuing to palpate along the lower border of the ribcage, you will feel the free end of the 12th rib on the lateral aspect of the waist. Locate **G.B.-25** on its lower border. **For orientation:** When pressing the flexed elbow onto the thorax, the tip of the olecranon will rest on the area around the free end of the 11th rib (→ **LIV-13**).

Needling

Vertically or obliquely 0.3–1 cun. Caution: Peritoneum; the needle should be inserted into the obliquus externus or internus muscles.

Actions/Indications

- Tonifies the Kidneys and regulates the water passages
- Strengthens the Spleen, regulates the Intestines
- Benefits the lumbar region

Special features

Front-*mu* point of the Kidneys

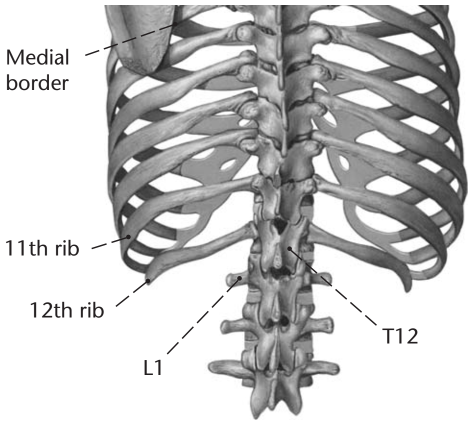

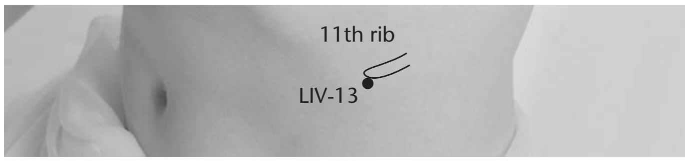

G.B.-26 *Girdling Vessel* DAIMAI

Location

On the lateral aspect of the waist, at the junction of a vertical line through the free end of the 11th rib and a horizontal line through the umbilicus, approximately 1.8 cun inferior to → **LIV-13**.

How to find

First, palpate the lower border of the ribcage to locate the free end of the 11th rib (→ **LIV-13**). Next, locate **G.B.-26** inferior to the free end of the 11th rib, level with the umbilicus. **Quick method for locating → LIV-13:** When pressing the flexed elbow onto the thorax, the tip of the olecranon will rest on the area around the free end of the 11th rib (→ **LIV-13**).

Located on the same level are → **Ren-8** (in the centre of the umbilicus), → **KID-16/ST-25/SP-15** (0.5/2/4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Be careful with slim patients.

Actions/Indications

- Regulates the *dai mai*
- Regulates the Uterus
- Drains Damp-Heat

Special features

Meeting point with the *dai mai*

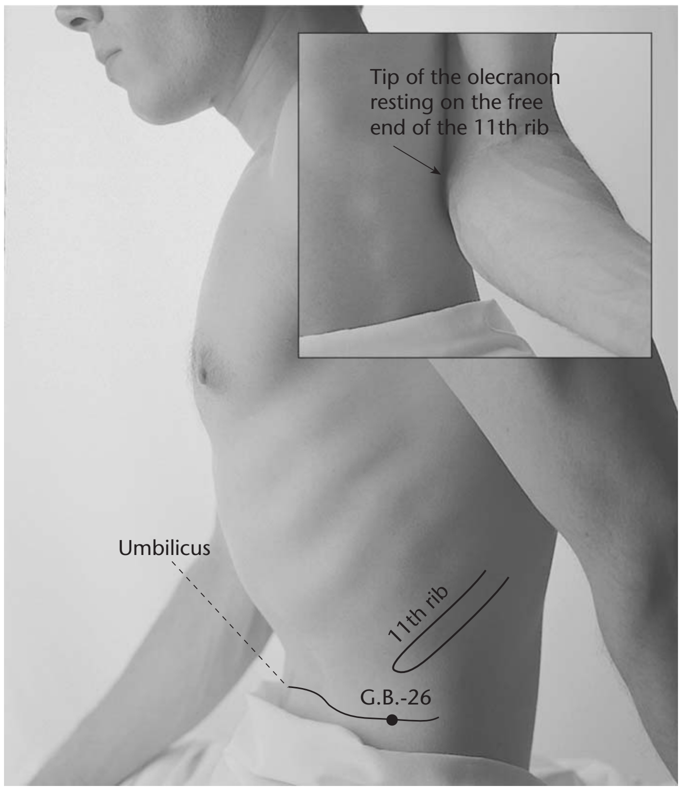
Tip of the olecranon resting on the free end of the 11th rib
Umbilicus
11th rib
G.B.-26

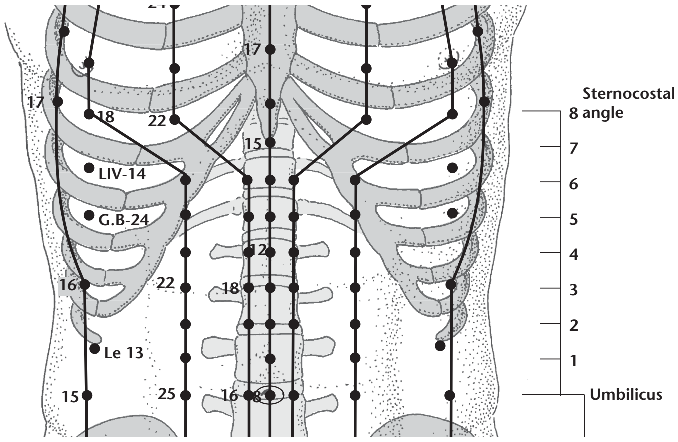
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus

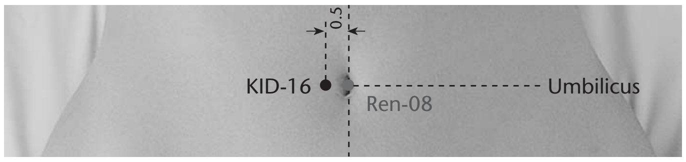
0.5
KID-16
Umbilicus
Ren-08

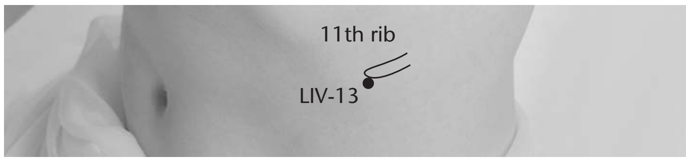
11th rib
LIV-13

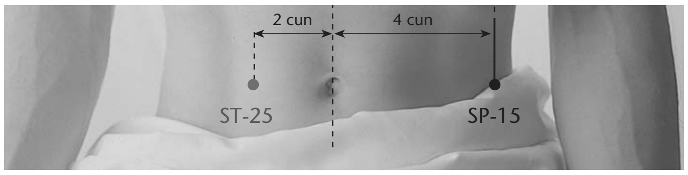
2 cun
4 cun
ST-25
SP-15

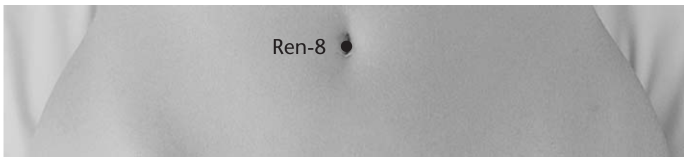

Fifth Pivot WUSHU G.B.-27

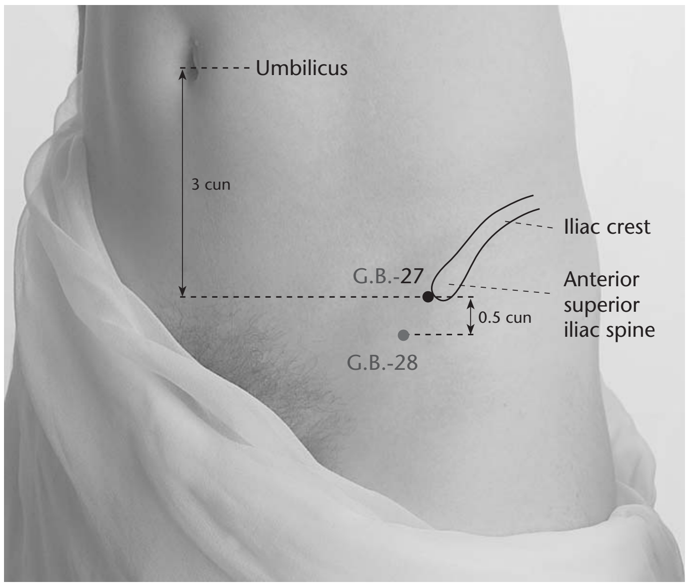
Umbilicus
3 cun
Iliac crest
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

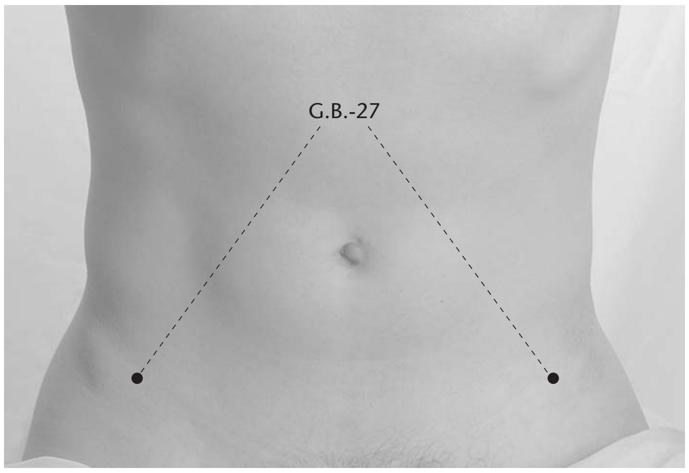

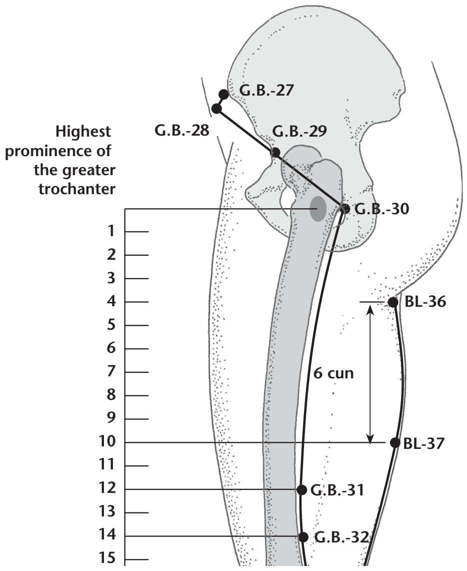
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Location

In the depression medial to the anterior superior iliac spine (ASIS), approximately 3 cun inferior to the umbilicus.

How to find

First, locate the anterior superior iliac spine (ASIS → 3.5) by palpating along the upper border of the iliac crest in an anterior and inferior direction. At its anterior end, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen. Locate **G.B.-27** in a depression anterior and medial to the ASIS.
→ **G.B.-28** is located approximately 0.5 cun inferior and medial to **G.B.-27**.
Located approximately on the same level (3 cun inferior to the umbilicus) are → **Ren-4/KID-13/ST-28** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

Vertically 1–1.5 cun. Caution during pregnancy.

Actions/Indications

- Strengthens the Kidneys
- Regulates the *dai mai*
- Regulates the Triple Burner

Special features

Meeting point with the *dai mai*

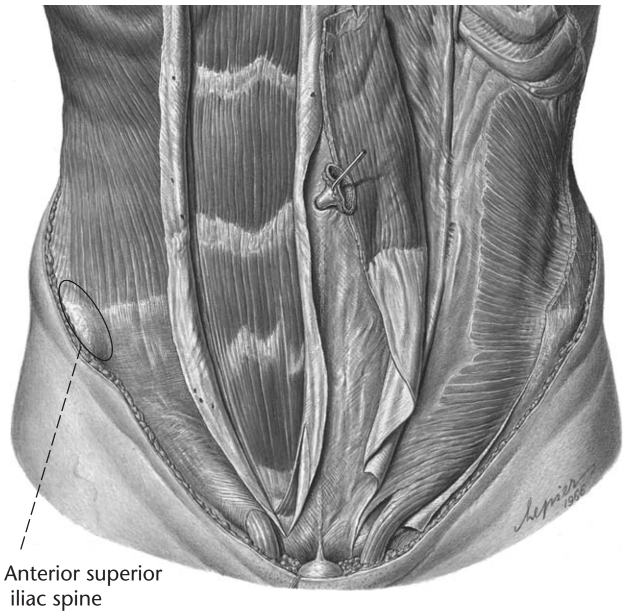
Anterior superior iliac spine

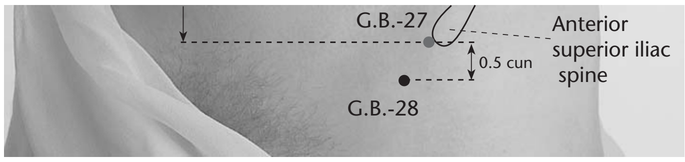
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-28 *Linking Path* WEIDAO

Location

On the lateral aspect of the abdomen, anterior and inferior to the anterior superior iliac spine (ASIS), approximately 0.5 cun anterior and inferior to → **G.B.-27**.

How to find

First, locate the ASIS (→ 3.5) by palpating along the upper border of the iliac crest in an anterior and inferior direction. On its anterior border, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen, with → **G.B.-27** located in a depression anterior and medial to the ASIS. Locate **G.B.-28** approximately 0.5 cun anterior and inferior to → **G.B.-27**.

Needling

Vertically 1–1.5 cun. Caution during pregnancy.

Actions/Indications

- Regulates the *dai mai*
- Regulates the Triple Burner
- Eliminates Dampness

Special features

Meeting point with the *dai mai*

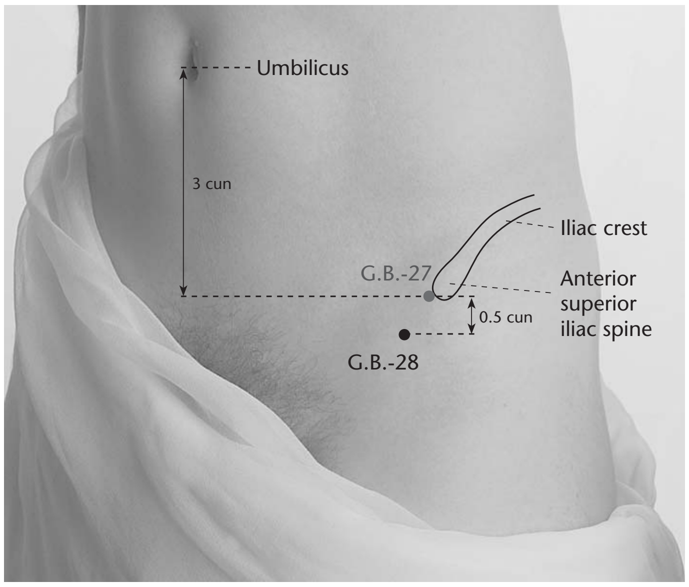
Umbilicus
3 cun
Iliac crest
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

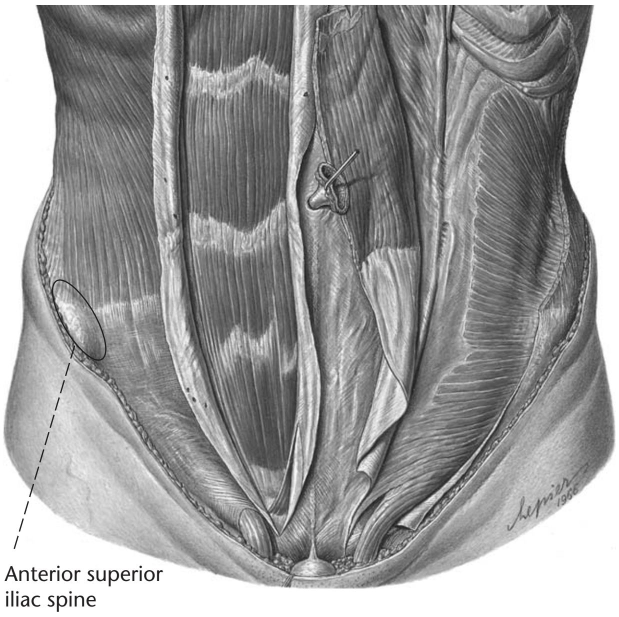
Anterior superior iliac spine

G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

Stationary Crevice JULIAO G.B.-29

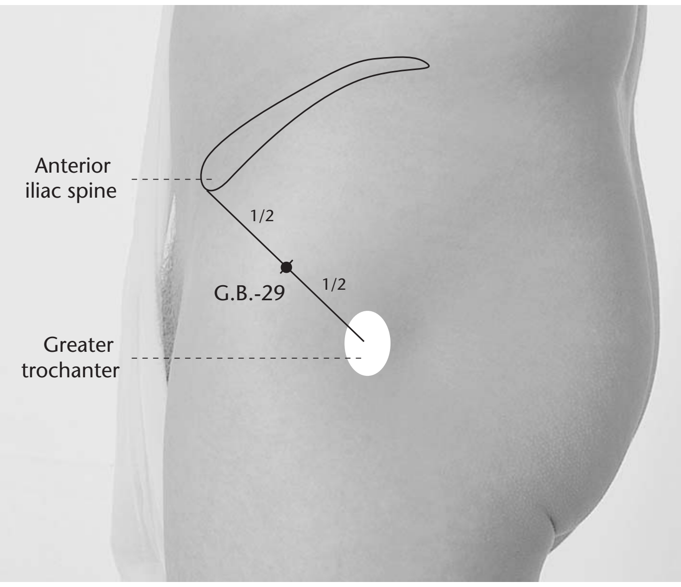
Anterior iliac spine
1/2
G.B.-29
1/2
Greater trochanter

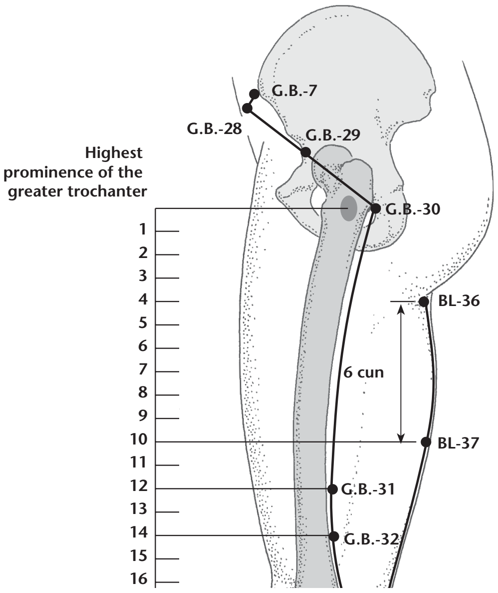
G.B.-7
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
BL-36
6 cun
BL-37
G.B.-31
G.B.-32

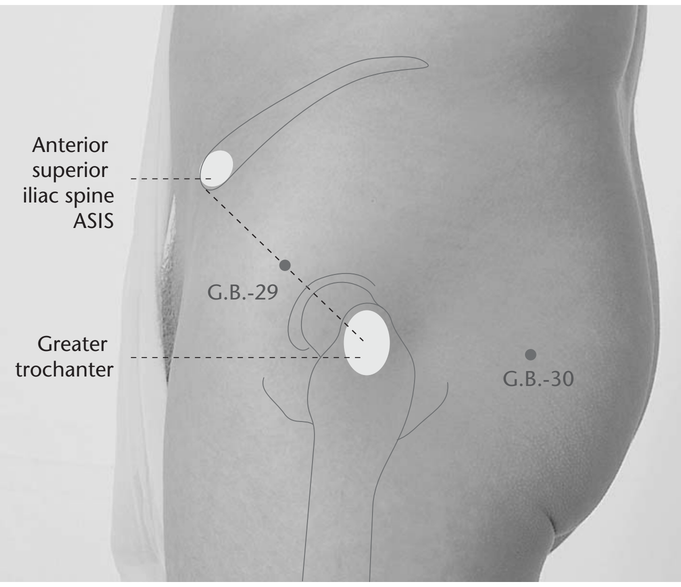
Anterior superior iliac spine ASIS
G.B.-29
Greater trochanter
G.B.-30

Location

At the midpoint of a line connecting the anterior superior iliac spine (ASIS) and the greater trochanter, at the anterior border of the iliac crest.

How to find

First, locate the ASIS (→ 3.5), the highest point on the anterior aspect of the iliac crest, by palpating along the upper border of the iliac crest in an anterior and inferior direction. At its anterior end, the ASIS can be palpated as a bony ridge on the lateral aspect of the lower abdomen. With the patient lying on their side and their leg slightly flexed, locate the greater trochanter (→ 3.6), a clearly marked bony structure in the region of the hip joint. **G.B.-29** is located at the midpoint of a line connecting these two reference points.

Needling

Vertically 1–2 cun

Actions/Indications

- Expels Cold, Dampness and Wind, opens the channel

Special features

Meeting point with the *yang qiao mai*, also with the *dai mai* according to some authors.

Ch04.11-F10028.qxd 2/23/08 11:25 AM Page 428

G.B.-30 *Jumping Circle* HUANTIAO

Location

With the patient lying on their side, the point is at the junction between the medial two-thirds and the lateral third of a line connecting the greater trochanter and the sacral hiatus.

How to find

Patient's position: Supine or, better, lying on their side, preferably with the hip and knee joints of the side to be needled flexed and the lower leg straight. Use pillows, etc, for a comfortable position. Reference points: the sacral hiatus (→ 3.4.4) and the lateral prominence of the greater trochanter (→ 3.6). **G.B.-30** is located on a line connecting these two points, one third of the distance from the greater trochanter.

Needling

1.5–3 cun vertically towards the genital region. The needle may reach the fascia of the obturator internus muscle as well as intermuscular connective tissue. Long 3 cun needles (50mm) should be used. Caution: Needling is often painful. Needling can result in an electric *de qi* sensation radiating to the toes. This sensation is particularly common with sciatic disorders and if **G.B.-30** is needled 1 cun inferior to its normal location.

Actions/Indications

- Opens the channel and the *luo*-connecting vessels, alleviates pain, benefits the hip and legs, eliminates Wind-Damp.

Special features

Meeting point with the BL channel, Ma Dan Yang Heavenly Star point. Important point for disorders of the hips.

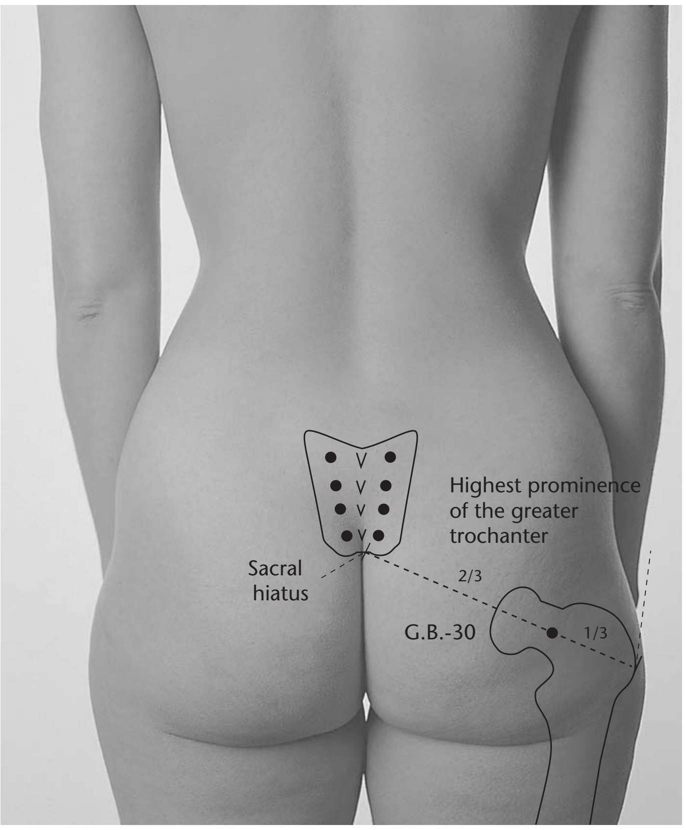

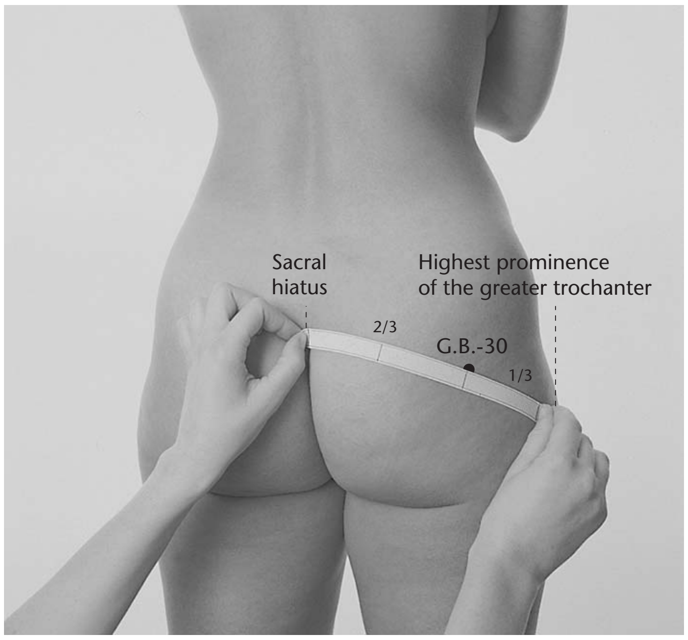

Wind Market FENGSHI G.B.-31

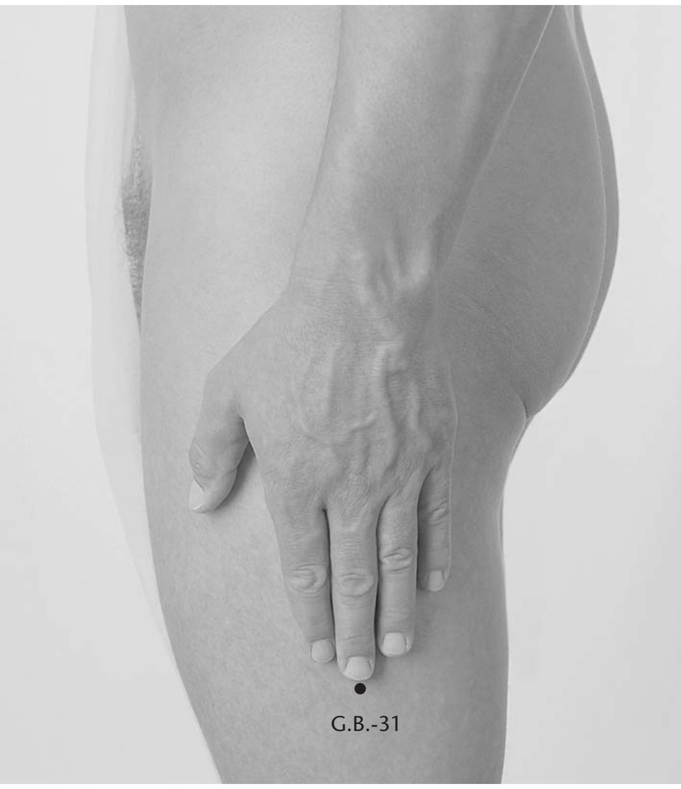

Location

On the lateral aspect of the thigh, inferior to the greater trochanter, approximately 7 cun proximal to the popliteal crease.

How to find

Ask the patient to place their hands on the imaginary seam of their trousers (this is best done with the patient standing). **G.B.-31** is located where the middle finger touches the lateral aspect of the thigh. **Or:** The distance between the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease is 19 cun (→ 2.2). Divide this distance into thirds and locate **G.B.-31** 1 cun proximal to a third of the distance from the popliteal crease. Sensitivity to pressure should help determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Eliminates Wind, Dampness and Heat
- Opens the channel and alleviates pain

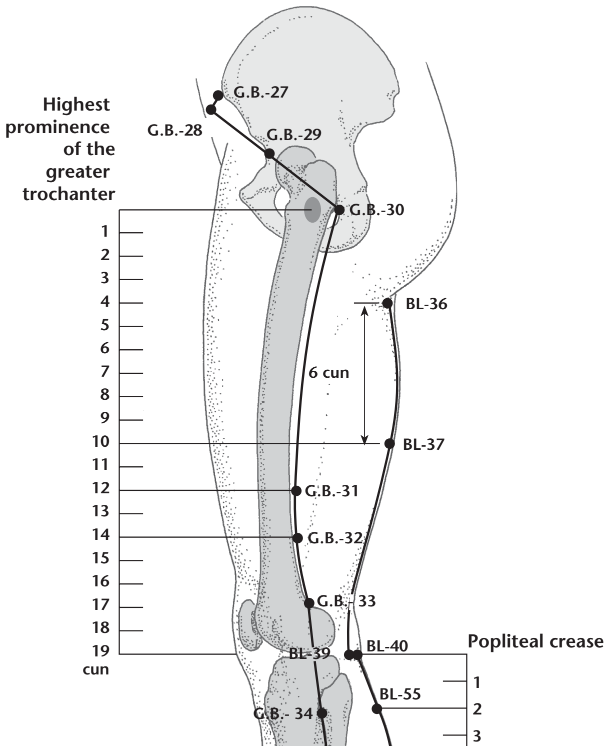
Highest prominence of the greater trochanter
G.B.-27
G.B.-28
G.B.-29
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-39
BL-40
Popliteal crease
BL-55
G.B.- 34
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
1
2
3

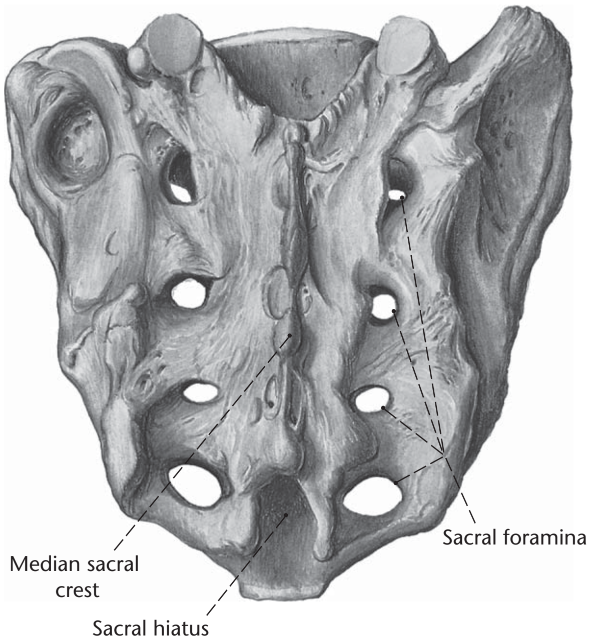
Median sacral crest
Sacral hiatus
Sacral foramina

G.B.-32 *Middle Ditch* ZHONGDU

Location

On the lateral aspect of the thigh, 5 cun proximal to the popliteal crease, between the vastus lateralis and biceps femoris muscles.

How to find

First, ask the patient to place their hands on the imaginary seam of their trousers (this is best done with the patient standing) and locate → **G.B.-31** where the middle finger touches the thigh. **G.B.-32** can now be located 2 cun distal to → **G.B.-31. Or:** The distance between the highest prominence of the greater trochanter (→ 3.6) to the popliteal crease is 19 cun (→ 2.2). Divide this distance into quarters (use an elastic tape or the spreading hands technique) and locate **G.B.-32** slightly proximal to a quarter of the distance from the popliteal crease. Sensitivity to pressure should help determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Eliminates Wind, Dampness and Cold, opens the channel

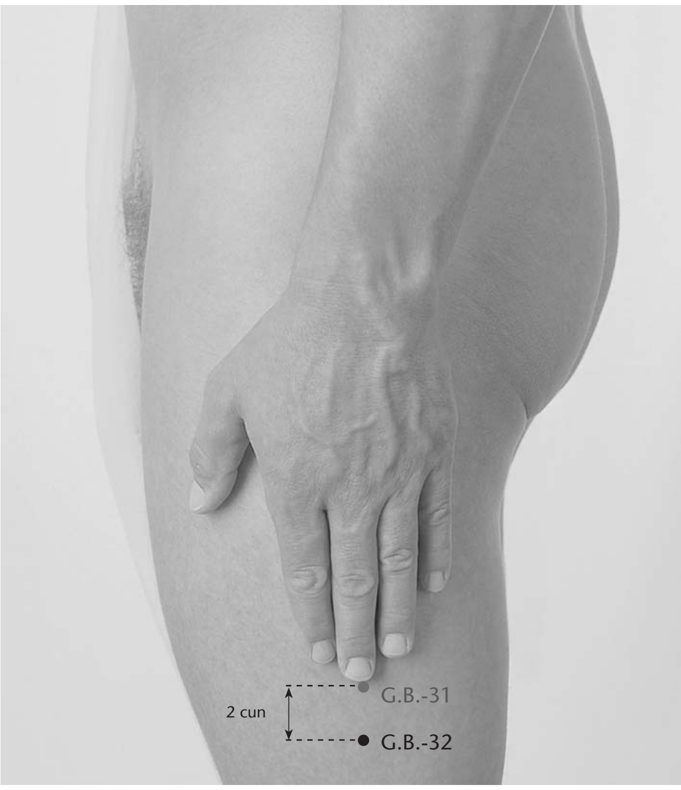
2 cun
G.B.-31
G.B.-32

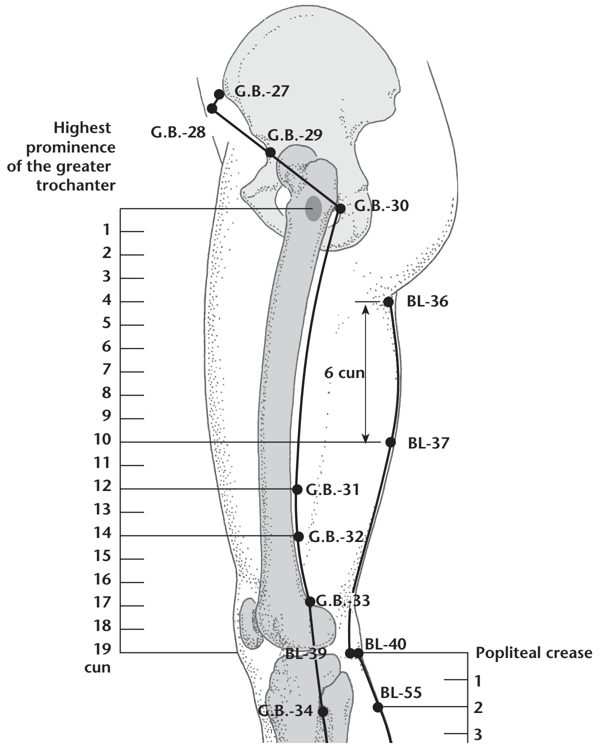
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
Popliteal crease
BL-39
BL-55
G.B.-34
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
1
2
3

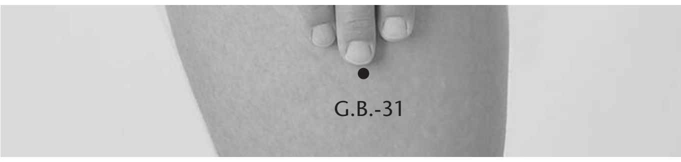

Knee Yang Gate XIYANGGUAN G.B.-33

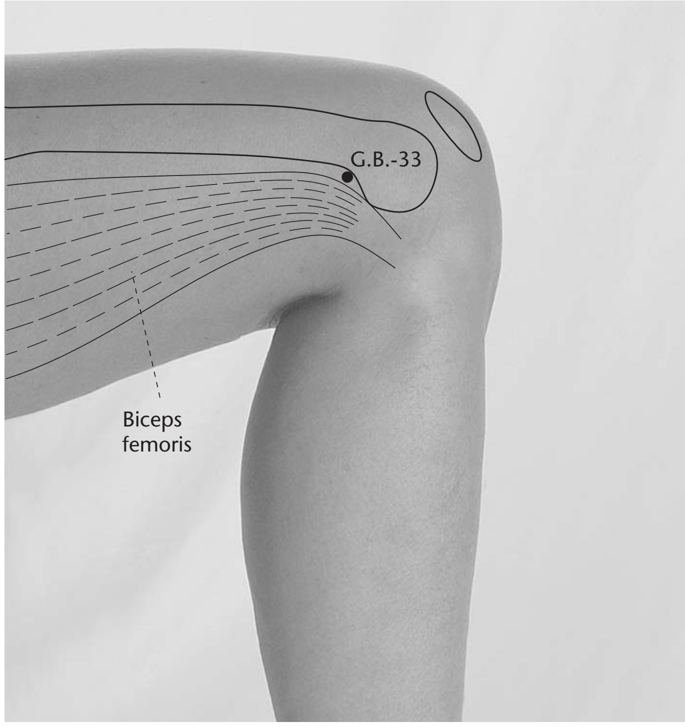

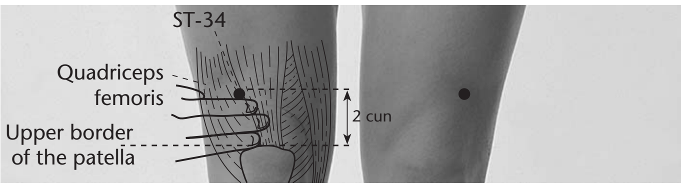

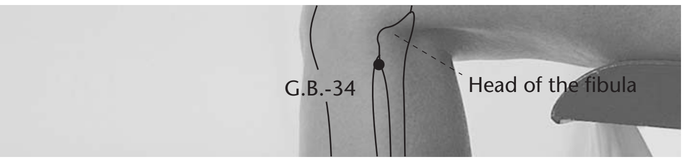

Location

On the lateral aspect of the knee. With the knee flexed, in the depression between the shaft and the lateral epicondyle and the tendon of the biceps femoris muscle, approximately 3 cun proximal to → **G.B.-34**.

How to find

This point is best located with the patient's knee flexed. On the level of the lateral upper border of the patella, palpate in a lateral direction towards the thigh and palpate for the distal end of the lateral epicondyle of the femur. Then locate **G.B.-33** in a clearly palpable depression between the epicondyle and the tendon of the biceps femoris muscle. The latter runs along the lateral aspect of the leg (along the line of an imaginary trouser seam) and attaches inferior to the knee joint at the head of the fibula.

→ **G.B.-34** is located 3 cun more distal, → **ST-34** is located 2 cun proximal to the lateral upper border of the patella.

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel
- Relaxes the tendons

G.B.-34 *Yang Mound Spring* YANGLINGQUAN

Location

In the depression anterior and inferior to the head of the fibula, between the peroneus longus and extensor digitorum longus muscles.

How to find

This point is best located with the patient's knee flexed (use knee support). Palpate for the head of the fibula on the lateral aspect of the leg and hold it with the index and middle fingers in a tweezer-like way. Gliding distally with both fingers, the more medial finger will drop into a depression directly anterior and inferior to the head of the fibula, the location of **G.B.-34**.

Located on the same level but on the medial aspect of the lower leg is → **SP-9** (at the junction of the shaft and the medial condyle of the tibia).

Needling

1–1.5 cun vertical insertion between the tibia and fibula towards the interosseous membrane. Caution: Deep peroneal nerve with deep needling, in some cases also common peroneal nerve. The needle may reach the interosseous membrane as well as the epineural tissue of the peroneal nerve.

Actions/Indications

- Benefits the tendons and joints
- Opens the channel, alleviates pain, benefits the lateral costal region
- Clears Damp-Heat from the Liver and Gall Bladder
- Spreads Liver Qi

Special features

He-sea point, Earth point, *hui*-meeting point of the tendons (coordination/movement), lower *he*-sea point of the Gall Bladder, Ma Dan Yang Heavenly Star point. Major point for disorders of the tendons and musculature.

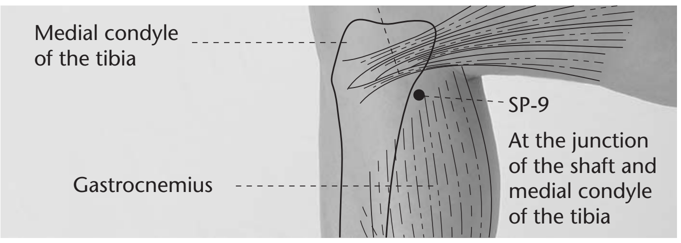
Medial condyle of the tibia
Gastrocnemius
SP-9
At the junction of the shaft and medial condyle of the tibia

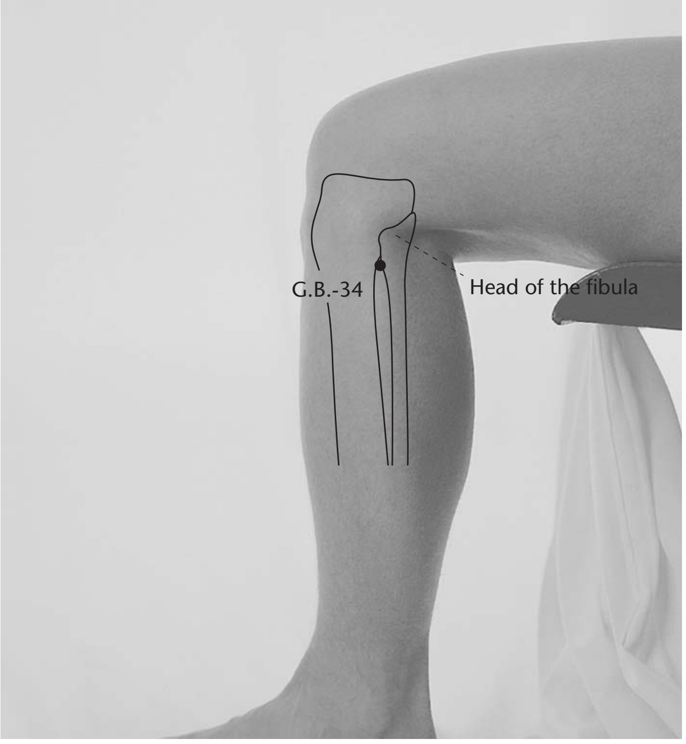
G.B.-34
Head of the fibula

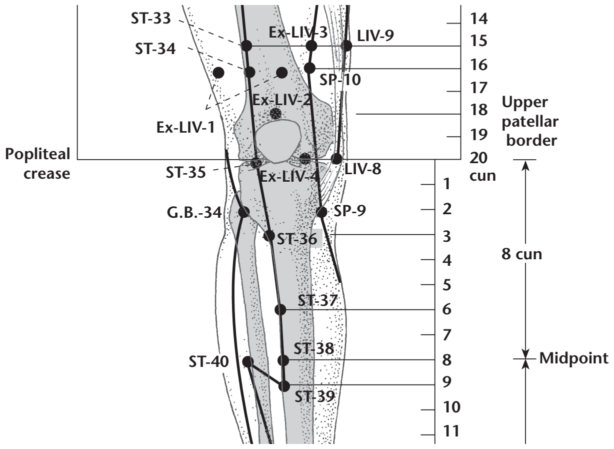
ST-33
ST-34
Ex-LIV-3
LIV-9
SP-10
Ex-LIV-2
Ex-LIV-1
Popliteal crease
ST-35
Ex-LIV-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
5
6
7
8
9
10
11
8 cun
Midpoint

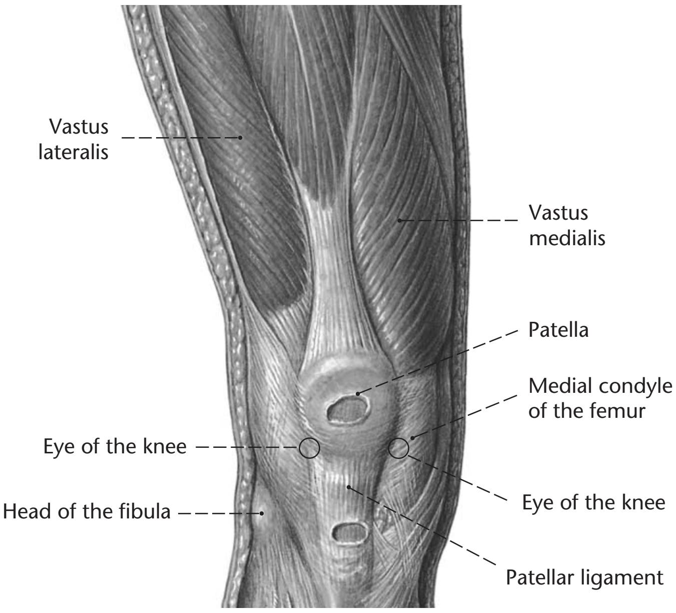
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
Eye of the knee
Eye of the knee
Head of the fibula
Patellar ligament

Yang Intersection YANGJIAO G.B.-35

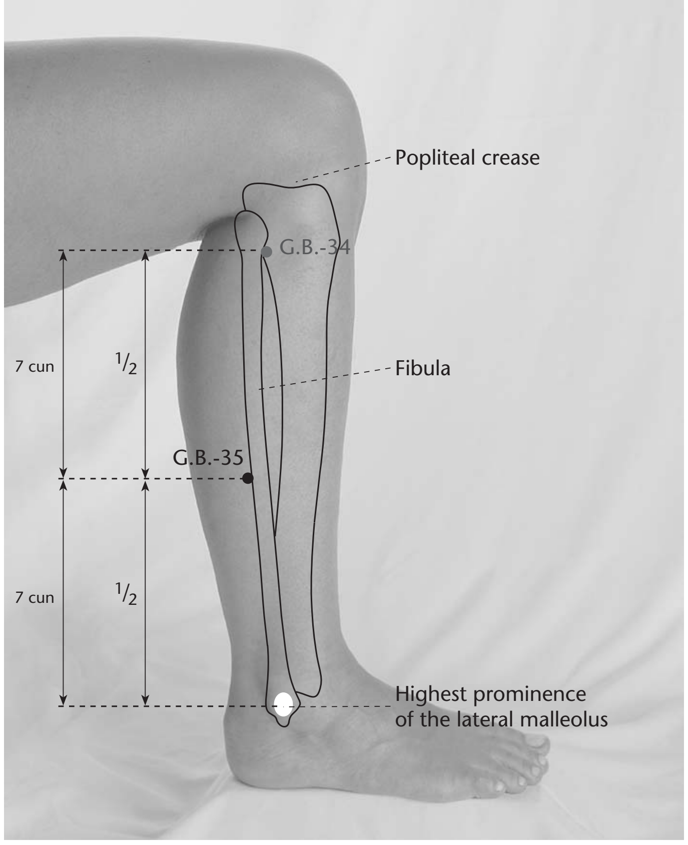
Popliteal crease
G.B.-34
7 cun
1/2
Fibula
G.B.-35
7 cun
1/2
Highest prominence
of the lateral malleolus

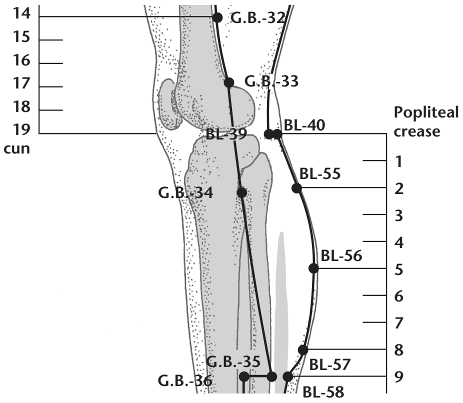
14
15
16
17
18
19
cun
G.B.-32
G.B.-33
BL-39
BL-40
Popliteal crease
1
BL-55
2
G.B.-34
3
4
BL-56
5
6
7
8
G.B.-35
BL-57
G.B.-36
9
BL-58

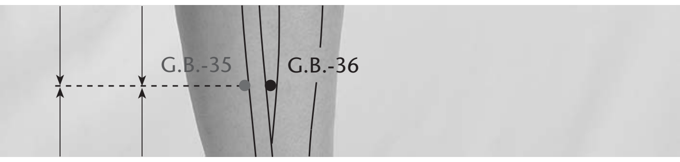

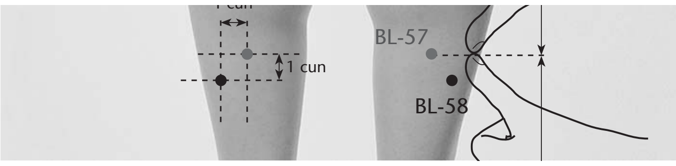
1 cun
BL-57
BL-58

Location

7 cun proximal to the highest prominence of the lateral malleolus, on the **posterior** border of the fibula.

How to find

Quick method: Spreading hands technique (→ 2.3.3): Locate **G.B.-35** at the midpoint of a line connecting → **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (→ 3.6.2), on the **posterior** border of the fibula (distance = 14 cun). **G.B.-35** is located 7 cun proximal to the prominence of the lateral malleolus. **Or:** Spreading hands technique (→ 2.3.3): Locate **G.B.-35** 1 cun distal to the midpoint of a line connecting the popliteal crease and the highest prominence of the lateral malleolus, on the **posterior** border of the fibula (distance = 16 cun, → 2.2).
For orientation: The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the posterior border of the fibula just superior to the lateral malleolus and then locate the point on an imaginary line running to the head of the fibula.
Located on the same level are → **G.B.-36** (on the **anterior** border of the fibula) and → **BL-58** (7 cun proximal to → **BL-60**).

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Clears Heat, opens the channel, relaxes the tendons
- Regulates Qi of the Gall Bladder
- Calms the *shen*

Special features

Xi-cleft point of the *yang wei mai*, meeting point with the *yang wei mai* according to some authors.

G.B.-36 *Outer Mound* WAIQIU

Location

7 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

Quick method: Spreading hands technique (→ 2.3.3): Locate → **G.B.-35** at the midpoint of a line connecting → **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (→ 3.6.2), on the **anterior** border of the fibula (distance = 14 cun). **G.B.-36** is located 7 cun proximal to the prominence of the lateral malleolus. **Or:** Spreading hands technique (→ 2.3.3): locate **G.B.-36** 1 cun distal to the midpoint of a line connecting the popliteal crease and the highest prominence of the lateral malleolus, on the **anterior** border of the fibula (distance = 16 cun, → 2.2).
For orientation: The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.
Located on the same level are → **G.B.-35** (on the **posterior** border of the fibula) and → **BL-58** (7 cun proximal to → **BL-60**).

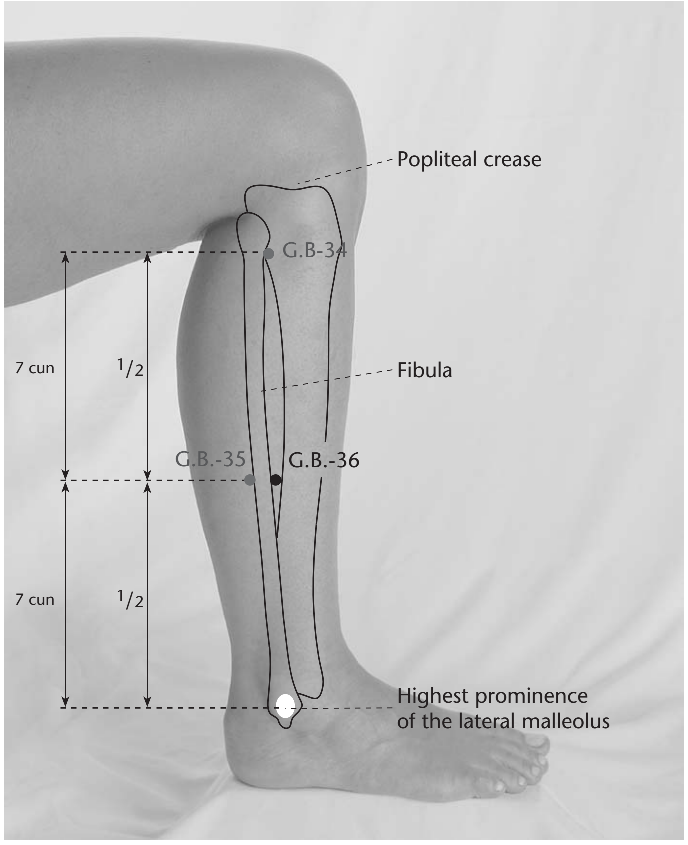

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Regulates Gall Bladder and Liver Qi
- Drains Damp-Heat
- Relaxes the tendons and muscles
- Calms the *shen*

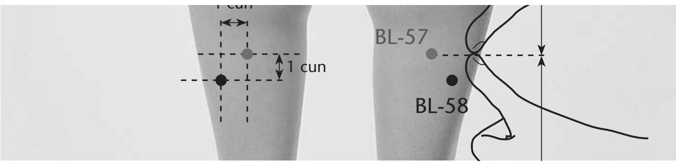

Special features

Xi-cleft point

Bright Light GUANGMING G.B.-37

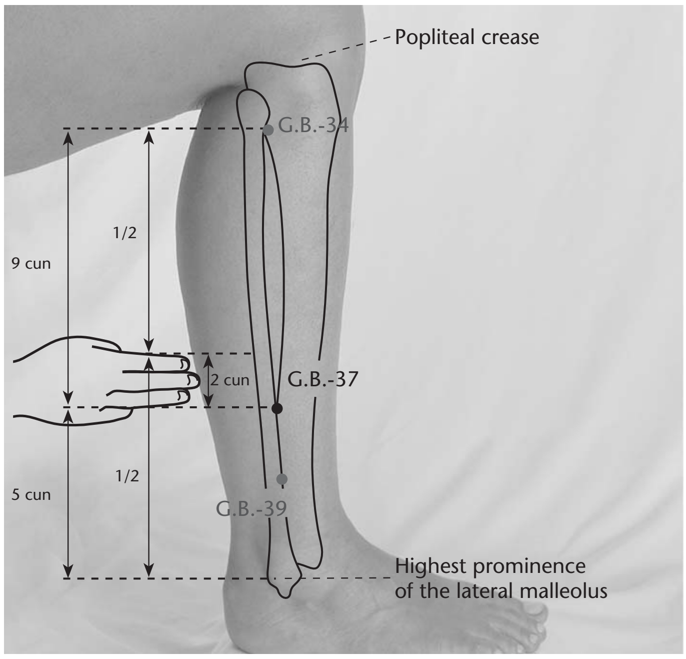
Popliteal crease
G.B.-34
1/2
9 cun
2 cun
G.B.-37
1/2
5 cun
G.B.-39
Highest prominence of the lateral malleolus

Location

5 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula, between the peroneus longus and extensor digitorum longus muscles.

How to find

Quick method: Spreading hands technique (→ 2.3.3): Place the little fingers on → **G.B.-34** (in the depression anterior and inferior to the head of the fibula) and the highest prominence of the lateral malleolus (→ 3.6.2). This distance is 14 cun. From the midpoint of this distance, measure 2 cun in a distal direction and locate **G.B.-37** on this level in a depression on the **anterior** border of the fibula. This depression is located 5 cun proximal to the prominence of the lateral malleolus (the width of 1 hand and 2 thumbs). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Benefits the eyes
- Eliminates Wind-Damp, opens the channel, alleviates pain

Special features

Luo-connecting point. Important distal point for disorders of the eyes.

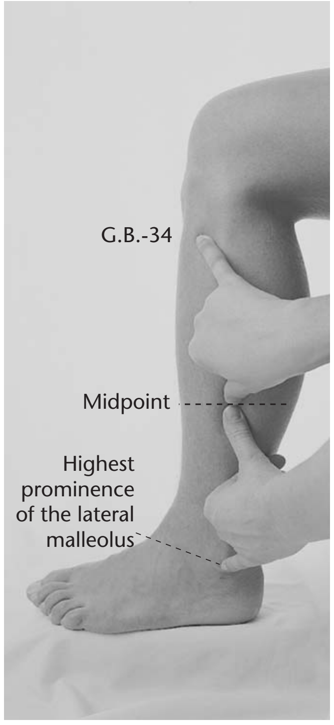
G.B.-34
Midpoint
Highest prominence of the lateral malleolus

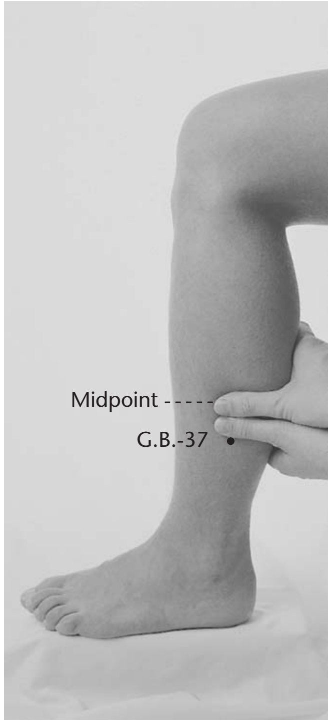
Midpoint
G.B.-37

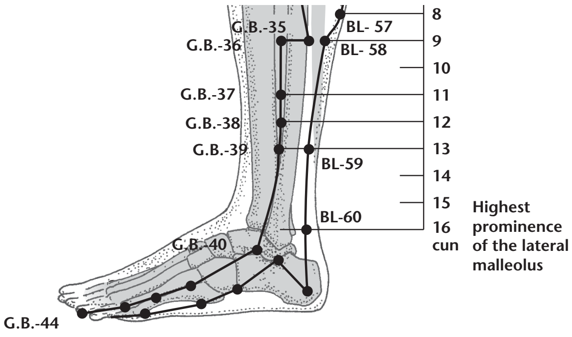
G.B.-35
G.B.-36
BL- 57
BL- 58
8
9
10
11
12
13
14
15
16
cun
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
Highest prominence of the lateral malleolus
G.B.-40
G.B.-44

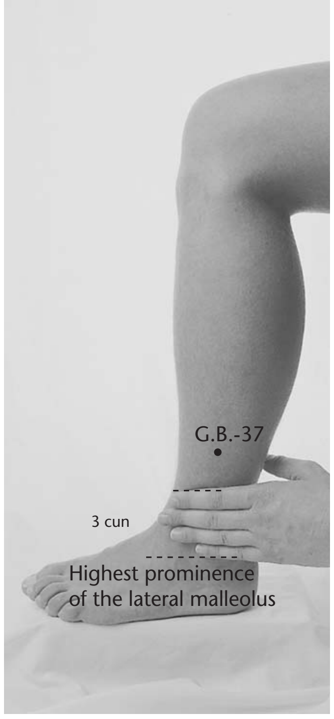
G.B.-37
3 cun
Highest prominence of the lateral malleolus

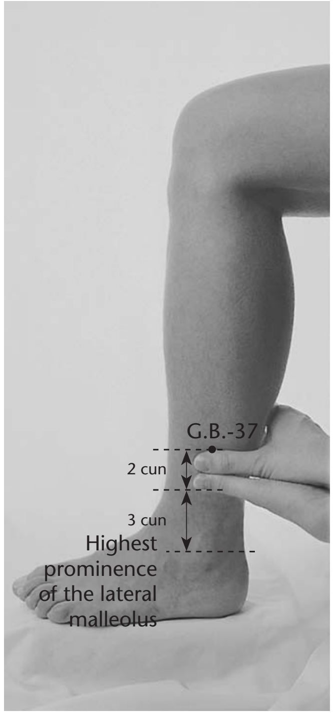
G.B.-37
2 cun
3 cun
Highest prominence of the lateral malleolus

G.B.-38 *Yang Assistance* YANGFU

Location

4 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

First, locate the prominence of the lateral malleolus (→ 3.6.2) and, from there, measure 4 cun in a proximal direction. There, locate **G.B.-38** on the **anterior** border of the fibula. **Or:** On the lateral aspect of the lower leg, the distance between the prominence of the lateral malleolus (→ 3.6.2) and the popliteal crease is 16 cun (→ 2.2). Divide this distance into quarters and locate **G.B.-38** one quarter of the distance from the lateral malleolus (aid: spreading hands technique or elastic tape → 2). **For orientation:** The borders of the fibula are deep to the peroneus brevis muscle and are often not easily palpable. For this reason, it is suggested to palpate the anterior border of the fibula just superior to the ankle and then locate the point on an imaginary line running to the head of the fibula.

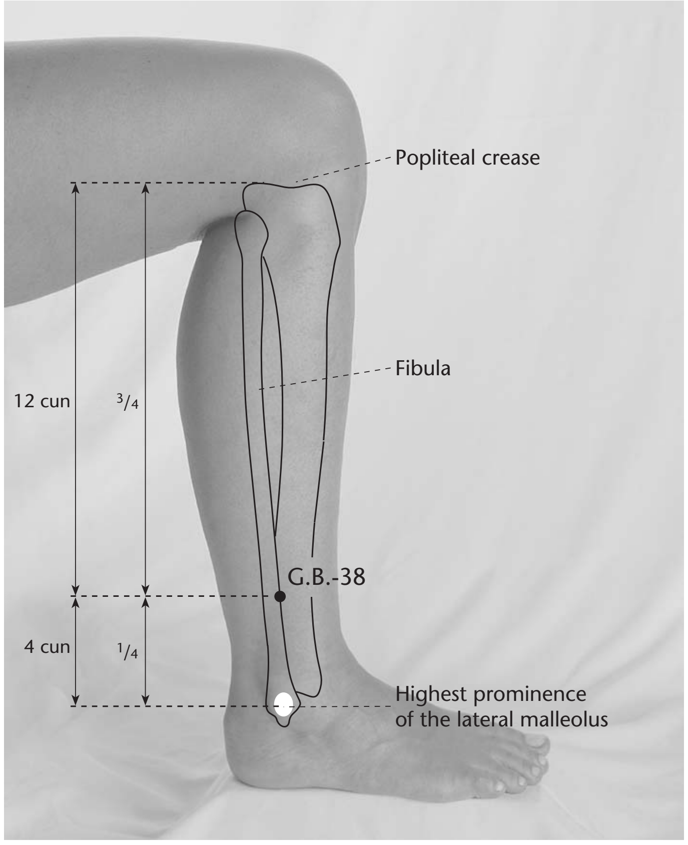
Popliteal crease
Fibula
12 cun
3/4
G.B.-38
4 cun
1/4
Highest prominence of the lateral malleolus

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Opens the channel
- Clears Wind and Heat

Special features

Jing-river point, Fire point, sedation point.

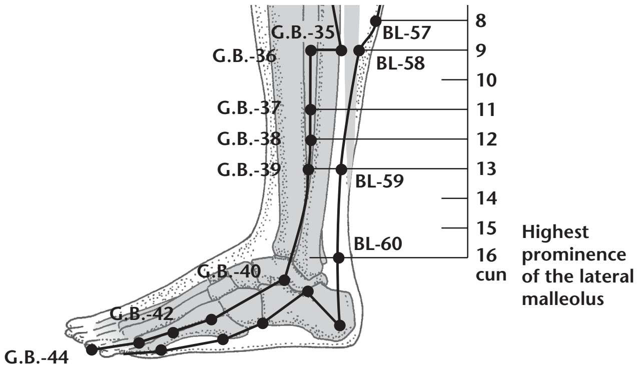
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

Hanging Bell XUANZHONG G.B.-39

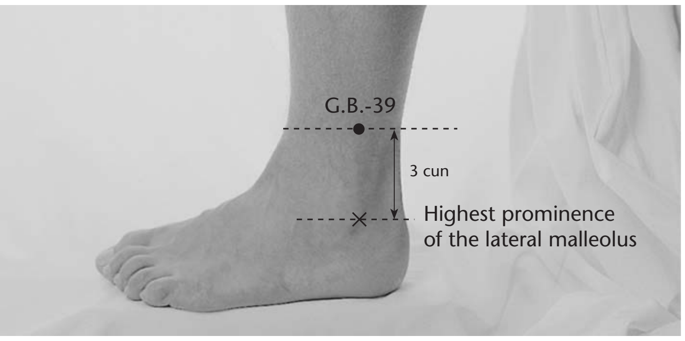
G.B.-39
3 cun
Highest prominence
of the lateral malleolus

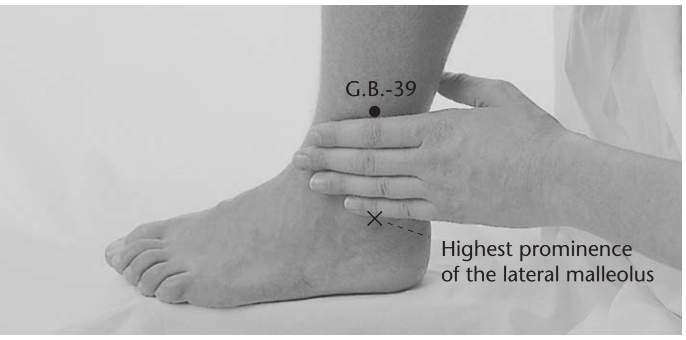
G.B.-39
Highest prominence
of the lateral malleolus

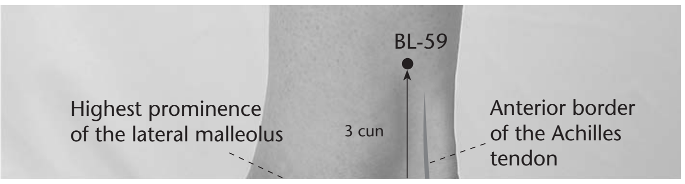
BL-59
Highest prominence
of the lateral malleolus
3 cun
Anterior border
of the Achilles
tendon

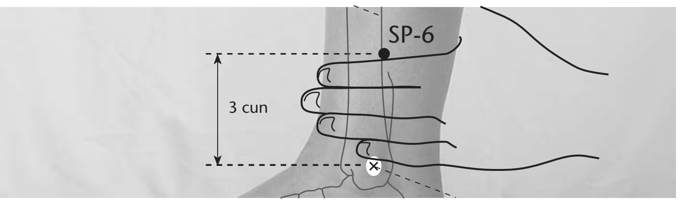

Location

3 cun proximal to the highest prominence of the lateral malleolus, on the **anterior** border of the fibula.

How to find

From the highest prominence of the lateral malleolus (→ 3.6.2), measure 3 cun (1 handbreadth) in a proximal direction and locate **G.B.-39** in a depression on the **anterior** border of the fibula. According to some authors, it is located between the posterior border of the fibula and the tendons of the peroneus longus and brevis musles. When in doubt, choose the more pressure-sensitive point.

Located on the same level but 3 cun proximal to → **BL-60** (depression between the malleolus and the Achilles tendon) is → **BL-59**. In a comparable position but on the medial aspect of the leg is → **SP-6** (3 cun proximal to the highest prominence of the medial malleolus, meeting point of the three leg Yin channels).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel, benefits the tendons and bones
- Benefits the Marrow, clears Wind-Dampness
- Clears Heat from the Gall Bladder

Special features

Hui-meeting point of the Marrow. Important distal point for disorders of the cervical spine.

G.B.-40 *Mound of Ruins* QIUXU

Location

In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of the extensor digitorum longus muscle.

How to find

For easier location, ask the patient to flex their ankle at a 90° angle. From the lateral malleolus (→ 3.6.2), let the palpating finger glide into a well-defined depression anterior and inferior to the malleolus. By lifting the toes, the tendons of the extensor digitorum longus muscle as well as the depression (**G.B.-40**) will become more pronounced. **For orientation: G.B.-40** is located at the junction of a vertical line along the anterior border of the lateral malleolus and a horizontal line along its lower border.

→ **SP-5** is located in a comparable position but on the medial aspect of the ankle (in the depression anterior and inferior to the medial malleolus). On a line connecting → **SP-5** and **G.B.-40**, → **LIV-4** is located medial to and → **ST-41** is located lateral to the tendon of the extensor hallucis longus muscle, which runs to the big toe.

Needling

0.5–1.5 cun vertically or slightly obliquely in the area of the fibulotarsal ligaments towards the medial arch of the foot.

Actions/Indications

- Opens the channel, benefits the joints
- Spreads Liver Qi, clears Heat and Dampness from the Gall Bladder

Special features

Yuan-source point

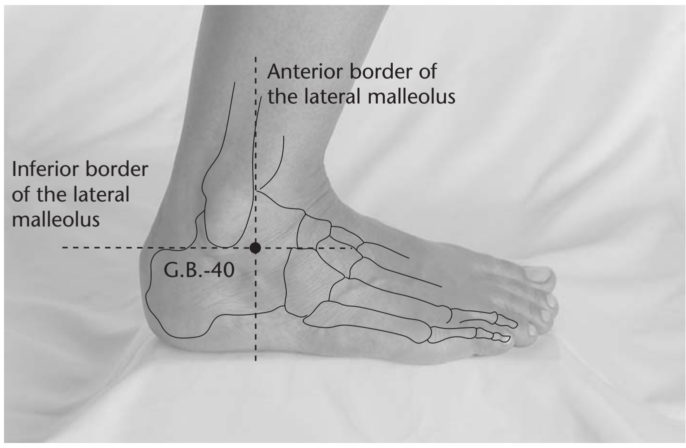
Anterior border of the lateral malleolus
Inferior border of the lateral malleolus
G.B.-40

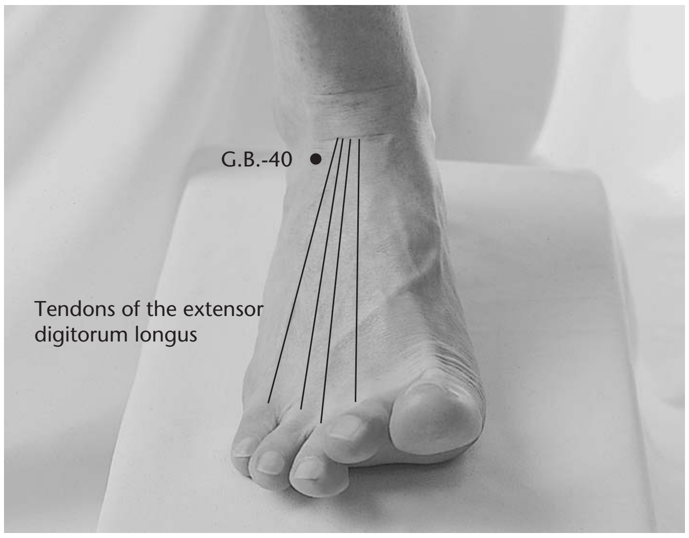
G.B.-40
Tendons of the extensor digitorum longus

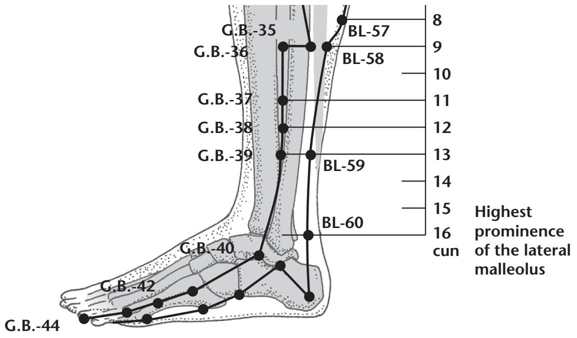
G.B.-35
G.B.-36
G.B.-37
G.B.-38
G.B.-39
G.B.-40
G.B.-42
G.B.-44
BL-57
BL-58
BL-59
BL-60
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

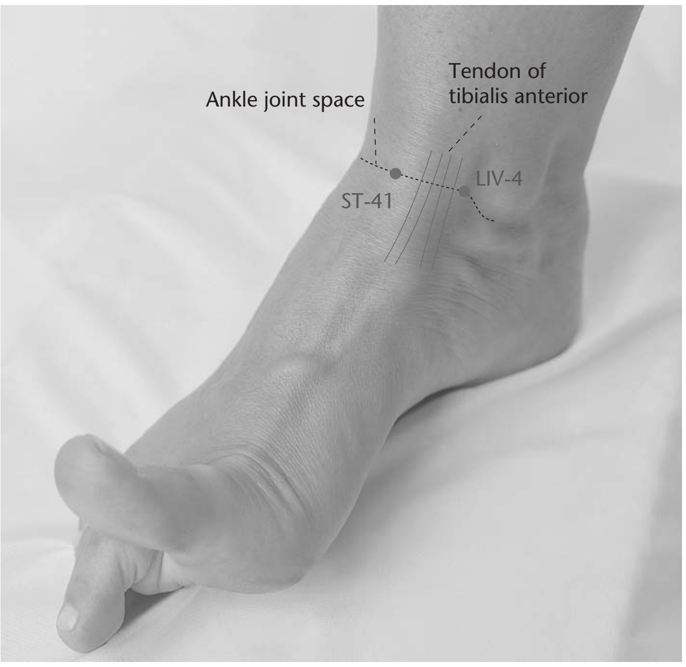
Ankle joint space
Tendon of tibialis anterior
LIV-4
ST-41

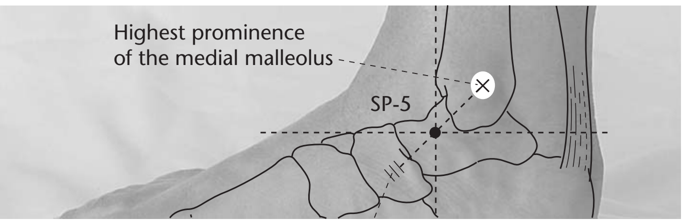
Highest prominence of the medial malleolus
SP-5

Foot Governor of Tears ZULINQI G.B.-41

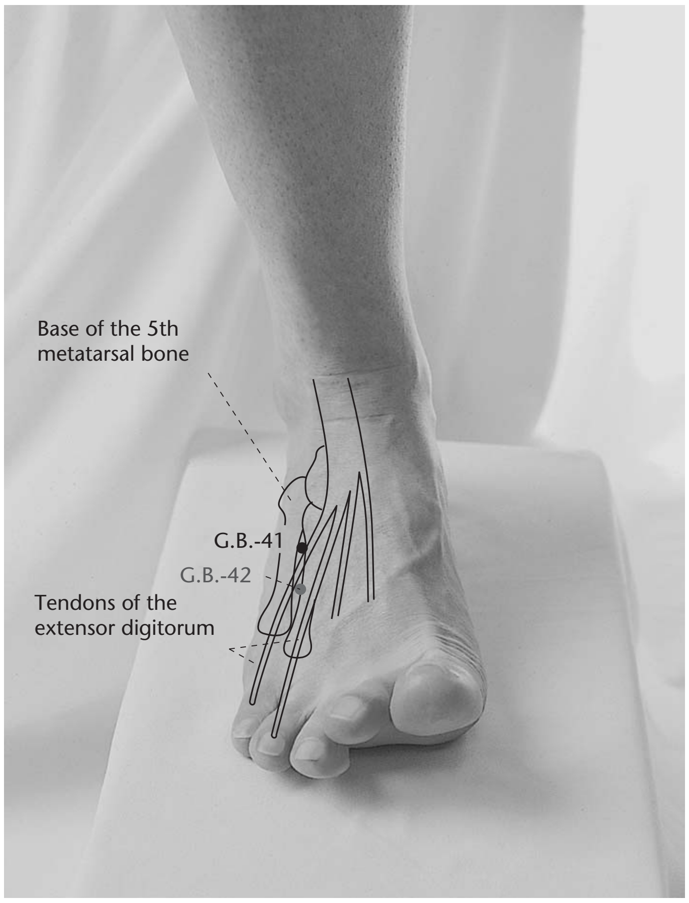
Base of the 5th metatarsal bone
G.B.-41
G.B.-42
Tendons of the extensor digitorum

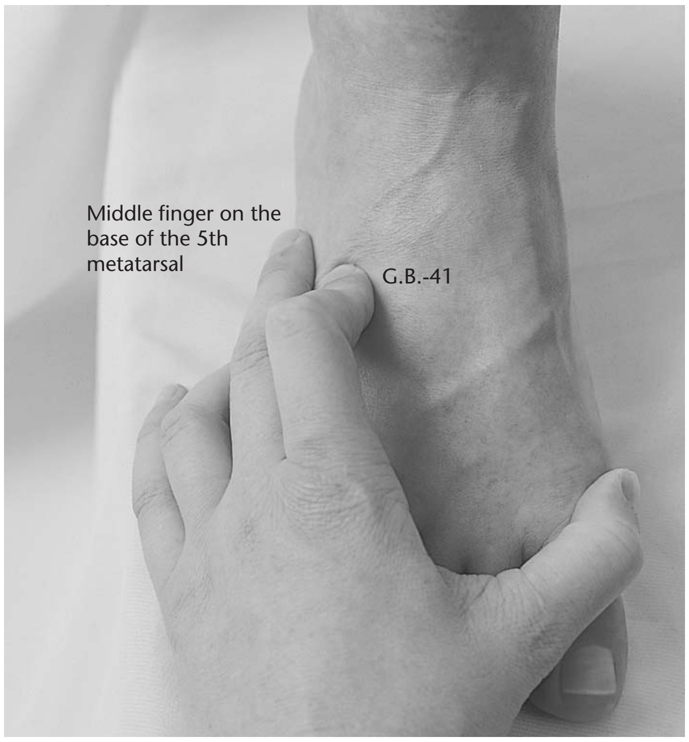
Middle finger on the base of the 5th metatarsal
G.B.-41

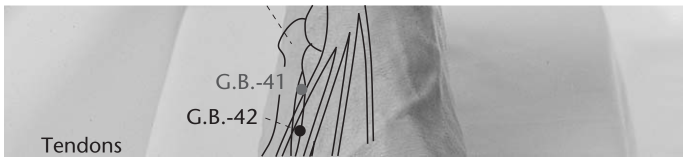
G.B.-41
G.B.-42
Tendons

Location

In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the tendon of the extensor digitorum longus muscle.

How to find

Palpate for the distinct tuberosity of the 5th metatarsal bone (→ 3.6.2) on the lateral aspect of the midfoot. With the palpating finger, glide on from the tuberosity to the dorsum of the foot, into the groove between the 4th and 5th metatarsal bones. There, palpate in a distal direction. **G.B.-41** is located at the junction of the heads and the shafts of the two bones. **Or:** Ask the patient to abduct their toes so that the branch of the tendon of the extensor digitorum longus muscle extending to the little toe becomes more pronounced. Then palpate in the groove between the 4th and 5th metatarsal bones from distal to proximal to where the tendon crosses the groove. → **G.B.-42** is located medial (or anterior) to it, while **G.B.-41** is located in the depression lateral (or posterior) to the tendon.

Needling

Perpendicularly or obliquely approximately 0.3–0.8 cun

Actions/Indications

- Spreads Liver Qi, benefits the thorax and the lateral costal region, transforms Phlegm, dissipates nodules, benefits the breasts
- Clears the head, benefits the eyes

Special features

Shu-stream point, Wood point, opening point of the *dai mai*, *ben* point (Five Phases), exit point.

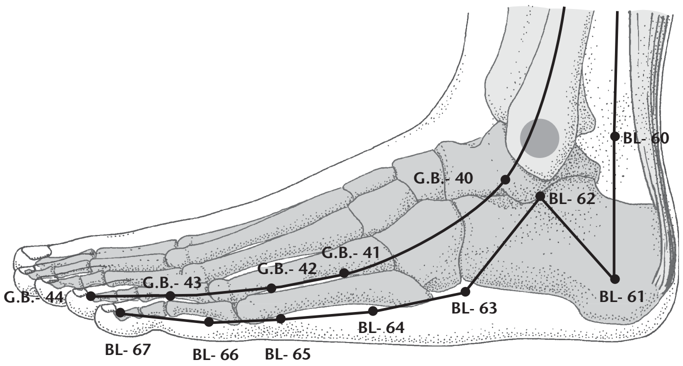
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 42
G.B.- 43
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

G.B.-42 *Earth Five Meetings* DIWUHUI

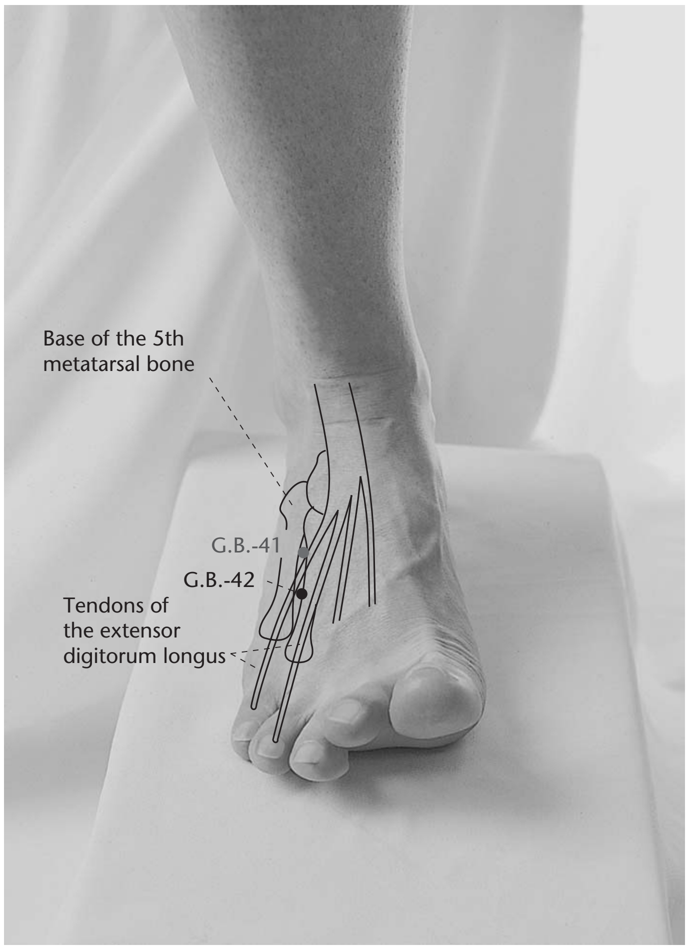
Base of the 5th metatarsal bone
G.B.-41
G.B.-42
Tendons of the extensor digitorum longus

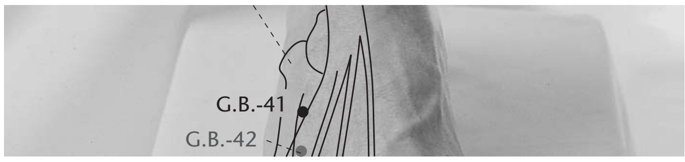

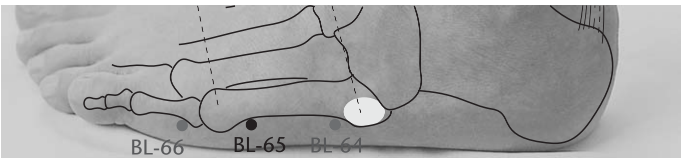

Location

Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of the extensor digitorum longus muscle.

How to find

From the space between the 4th and 5th toes, palpate towards the ankle. Locate **G.B.-42** in the groove between the two bones, proximal to the metatarsophalangeal joints (and proximal to the heads of the two metatarsal bones). **Or:** Ask the patient to abduct their toes so that the branch of the tendon of the extensor digitorum longus muscle extending to the little toe becomes more pronounced. Then palpate from the space between the toes along the groove between the 4th and 5th metatarsal bones from distal to proximal to where the tendon crosses the groove. **G.B.-42** is located medial (or anterior) to the tendon.

→ **G.B.-41** is located lateral (or posterior) to the tendon when continuing to palpate along the groove. → **BL-65** is located approximately on the same level on the lateral border of the foot, proximal to the head of the 5th metatarsal bone.

Needling

Vertically or obliquely 0.3–0.8 cun

Actions/Indications

- Moves Liver Qi
- Clears Heat from the Gall Bladder

Clamped Stream JIAXI/XIAXI G.B.-43

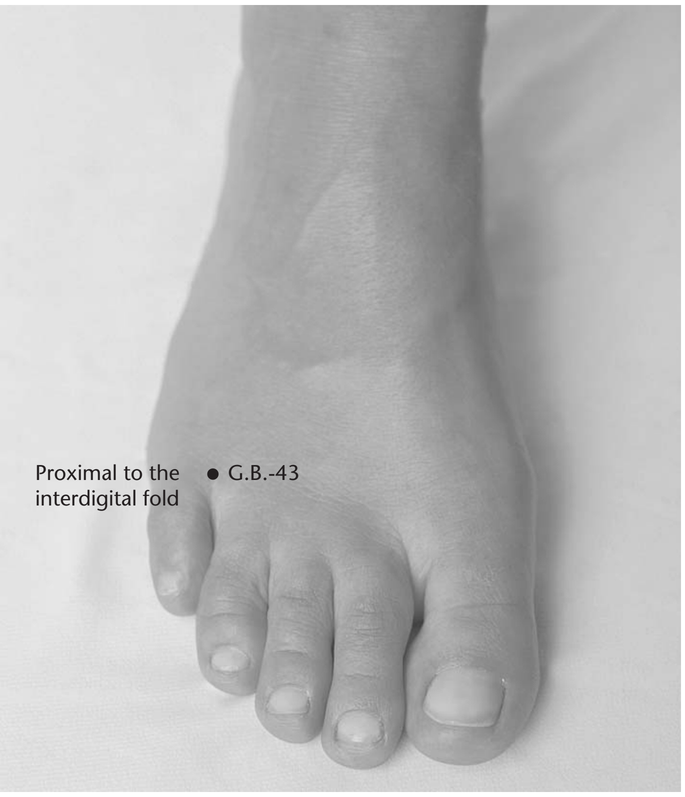
Proximal to the
interdigital fold
G.B.-43

Location

Between the 4th and 5th toes, proximal to the margin of the interdigital web.

How to find

Locate the interdigital web between the 4th and 5th toes. Then locate **G.B.-43** slightly proximal to the margin of the web. **G.B.-43** as well as → **LIV-2** and → **ST-44** are part of → **Ex-UE-10** (*bafeng*: proximal to the margins of the webs between the toes). → **T.B.-2** is located in a comparable position on the hand, between the ring finger and little finger. It is also part of → **Ex-UE-9** (*baxie*).

Needling

Up to 0.5 cun vertically or 1 cun obliquely in a proximal direction.

Actions/Indications

- Clears Heat, Damp-Heat and Wind, especially from the opposite end of the channel
- Calms Liver Yang

Special features

Ying-spring point, Water point, tonification point.

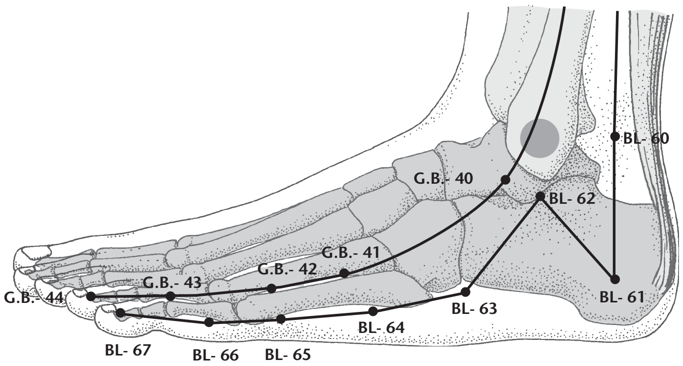
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 42
G.B.- 43
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

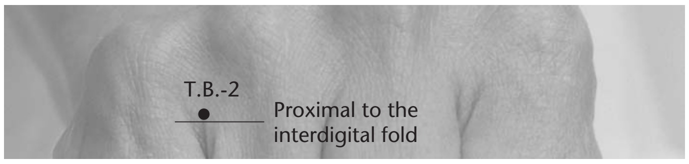
T.B.-2
Proximal to the
interdigital fold

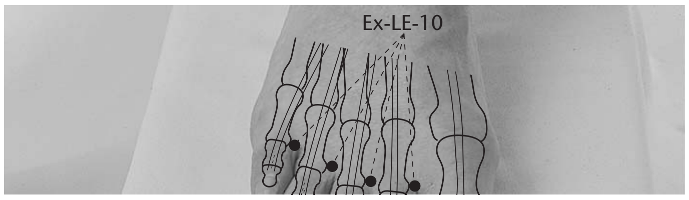

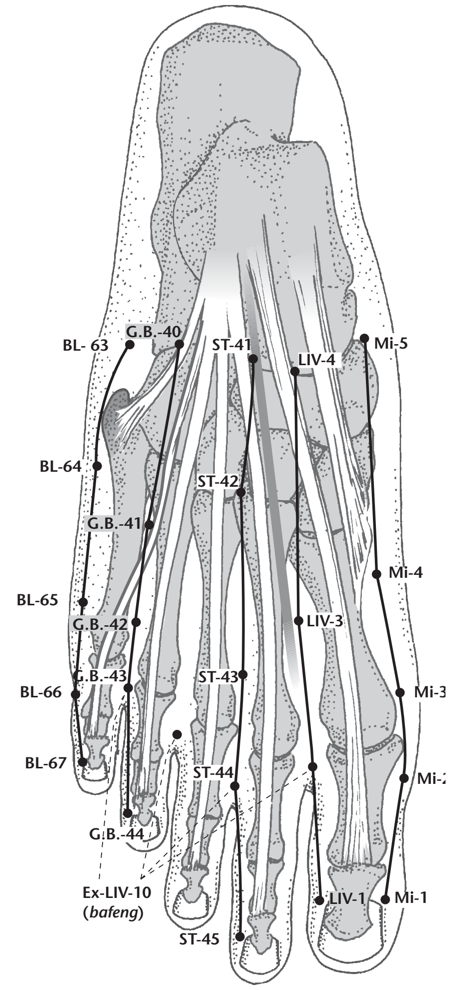
BL- 63
G.B.-40
ST-41
LIV-4
Mi-5
BL-64
ST-42
G.B.-41
Mi-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
Mi-3
BL-67
ST-44
Mi-
G.B.-44
Ex-LIV-10
(bafeng)
LIV-1
Mi-1
ST-45

G.B.-44 *Foot Portal Yin* ZUQIAOYIN

Location

On the 4th toe, 0.1 cun from the lateral corner of the nail.

How to find

G.B.-44 is located at the junction of two tangents along the proximal and lateral borders of the 4th toe.

→ **BL-67** is located at the lateral corner of the little toe.

Needling

0.1 cun vertically or 0.2 cun obliquely in a proximal direction or prick to bleed. Caution: Painful point.

Actions/Indications

- Clears Wind, Heat and Fire
- Harmonises the Liver and Gall Bladder
- Calms the *shen*

Special features

Jing-well point, Metal point

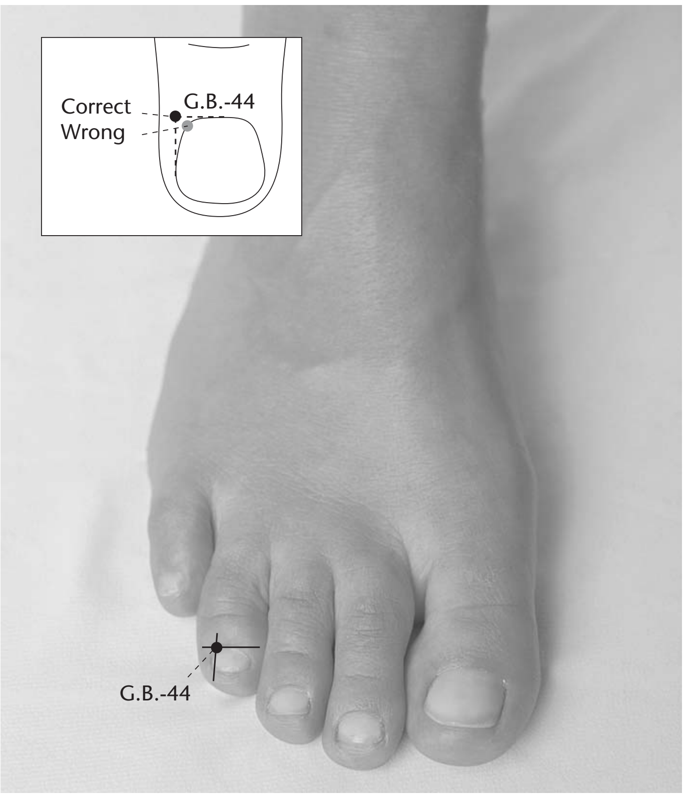
Correct
Wrong
G.B.-44
G.B.-44

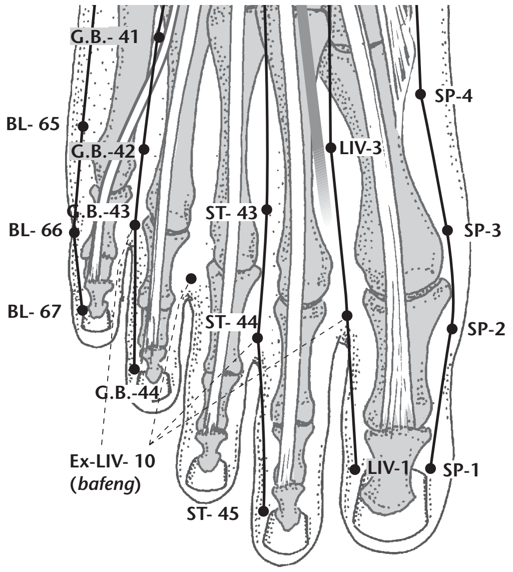
G.B.- 41
SP-4
BL- 65
G.B.-42
LIV-3
G.B.-43
ST- 43
BL- 66
SP-3
BL- 67
ST- 44
SP-2
G.B.-44
Ex-LIV- 10
(*bafeng*)
LIV-1
SP-1
ST- 45

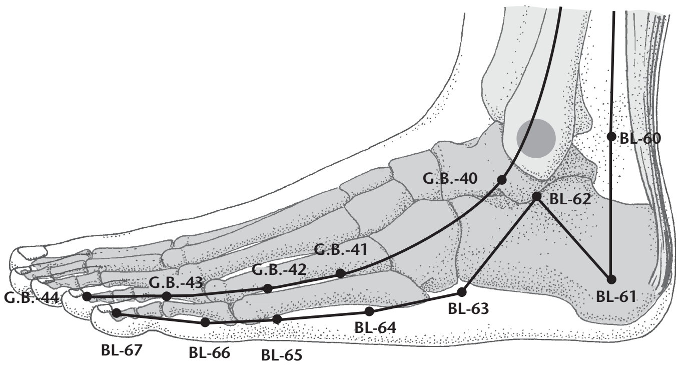
BL-60
G.B.-40
BL-62
G.B.-41
G.B.-42
G.B.-43
G.B.-44
BL-61
BL-63
BL-64
BL-67
BL-66
BL-65

4.12 The Liver Channel System – Foot *jueyin* (*zu jue yin jing luo*)

4.12.1 The Liver Primary Channel (*zu jue yin jing*)

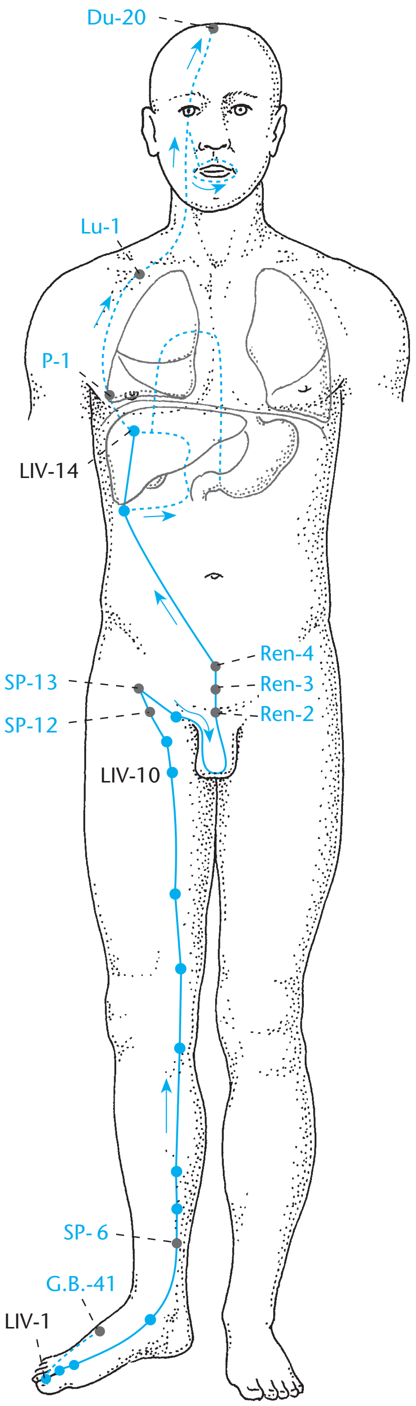
Du-20
Lu-1
P-1
LIV-14
Ren-4
SP-13
Ren-3
SP-12
Ren-2
LIV-10
SP- 6
G.B.-41
LIV-1

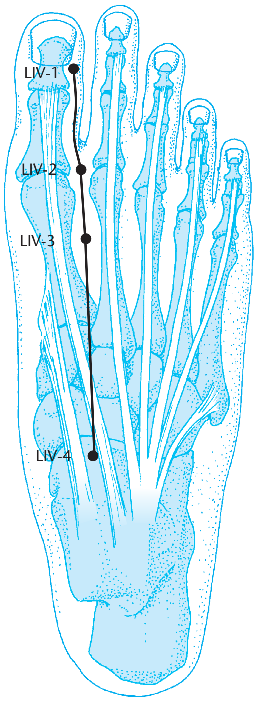
LIV-1
LIV-2
LIV-3
LIV-4

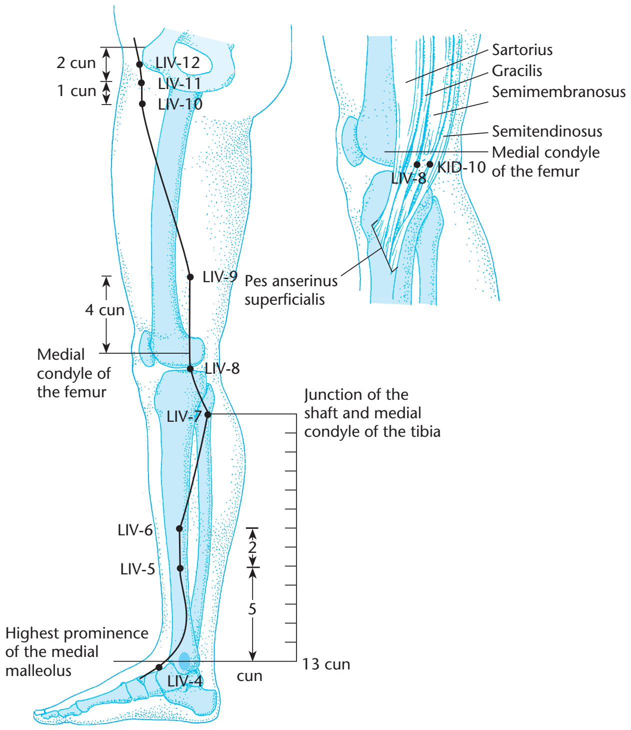
2 cun
1 cun
LIV-12
LIV-11
LIV-10
Sartorius
Gracilis
Semimembranosus
Semitendinosus
Medial condyle of the femur
KID-10
LIV-8
LIV-9
Pes anserinus superficialis
4 cun
Medial condyle of the femur
LIV-8
Junction of the shaft and medial condyle of the tibia
LIV-7
LIV-6
LIV-5
2
5
Highest prominence of the medial malleolus
LIV-4
cun
13 cun

Pathway

The Liver primary channel begins at **LIV-1** (*dadun*) on the lateral corner of the nail of the big toe. This point is reached by a small branch of the G.B. channel, which separates from the Gall Bladder primary channel on the dorsum of the foot at → **G.B.-41** (*zulinqi*) (foot Yin–Yang connection of the third great circuit).

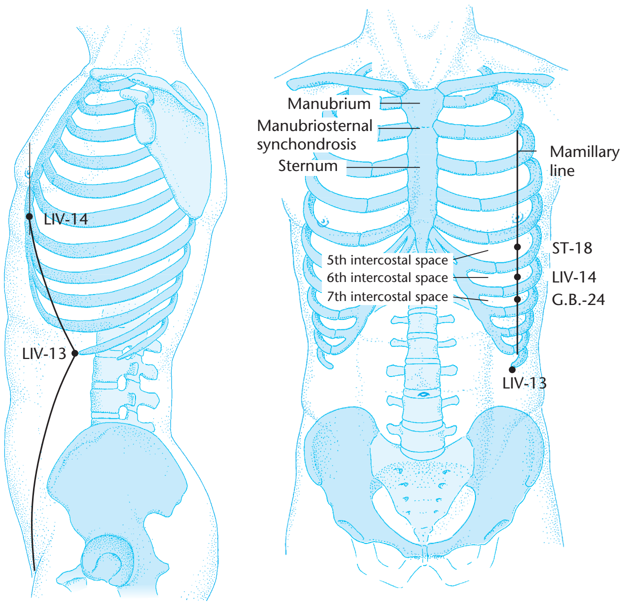

The **external pathway** runs proximally between the 1st and 2nd metatarsal bones

- ➡ ascends anterior to the medial malleolus
- ➡ crosses the SP channel at → **SP-6** (*yinlingquan*) on the medial aspect of the lower leg
- ➡ ascends the medial aspect of the leg towards the knee anterior to the SP channel
- ➡ continues along the medial aspect of the thigh to the pubic region, where it passes → **SP-12** (*chongmen*) and → **SP-13** (*fushe*)
- ➡ circles the external genitalia
- ➡ ascends the lower abdomen, passing → **Ren-2** (*qugu*), → **Ren-3** (*zhongji*) and → **Ren-4** (*guanyuan*)
- ➡ obliquely traverses the abdomen to **LIV-13** (*zhangmen*) at the free end of the 11th rib and to **LIV-14** (*qimen*), where the **external pathway** terminates.

At **LIV-13**, the channel enters the abdomen, marking the beginning of the **internal pathway**. It

- ➡ circles the Stomach (*wei*)
- ➡ connects with its pertaining *zang*-Organ, the Liver (*gan*) and its paired *fu*-Organ, the Gall Bladder (*dan*)
- ➡ penetrates the diaphragm
- ➡ spreads in the lateral hypochondrium and thoracic region.

The channel then ascends along the posterior aspect of the trachea to the throat and nasopharynx and connects with the eye system and the brain. It crosses the forehead and ascends to the vertex, where it connects with the extraordinary vessel *du mai* at → **Du-20** (*baihui*).

An **internal branch** descends from the maxillary sinus to the cheek and circles the **inner surface** of the lips.

An **internal branch** emerges from the Liver, penetrates the diaphragm, disperses in the Lung (*fei*) and meets the Lung primary channel (deep Yin–Yin connection), closing the first circuit of the Nutritive Qi (*ying qi*) (→ 1.1.4). This branch further connects with the Pericardium primary channel beneath → **P-1** (*tianchi*) (hand–foot pairing of the third great circuit: Yin axes, *jue yin*).

Clinical importance (→ 1.2)

Exterior (*biao*) signs and symptoms: Headaches, dizziness, blurred vision, tinnitus, fever, spastic extremities
Interior (*li*) or *zangfu*-Organ signs and symptoms: Distension, fullness and pain in the costal region, tightness and fullness in the chest, abdominal pain, vomiting, jaundice, diarrhoea, *shan* disorder, enuresis, urinary retention, yellow urine

Connections of the Liver primary channel (→ 1.2)

Connections with other channels

Gall Bladder primary channel (*zu shao yang jing*)

Connection: Foot Yin–Yang pairing of the third great circuit
Location: G.B.-41 → LIV-1 (on the foot)
Circulation: Circadian (according to the Organ clock)
Importance: Exterior–Interior relationship

Pericardium primary channel (*shou jue yin jing*)

Connection: Paired according to the six-channel theory (hand–foot pairing): *jue yin* (Yin axes of the third great circuit)
Location: LIV → P. An internal branch originating in the Liver penetrates the diaphragm and connects with the Pericardium primary channel beneath → P-1.
Circulation: Non-circadian (not according to the Organ clock)
Importance: Above–below relationship

Lung primary channel (*shou tai yin jing*)

Connection: Deep Yin–Yin connection
Location: LIV → LU (in the thorax). An internal branch originating in the Liver penetrates the diaphragm and disperses in the Lung to connect with the Lung primary channel.
Circulation: Circadian (according to the Organ clock)
Importance: The Lung primary channel receives part of its Nutritive Qi (*ying qi*) from the Liver primary channel (first circulation of the *ying qi* → 1.1.4).

Connections with zangfu-Organ systems

Stomach (*wei*), **Liver (*gan*)**, **Gall Bladder (*dan*)**, Lung (*fei*)

4.12.2 The Liver Divergent Channel (*zu jue yin jing bie*)

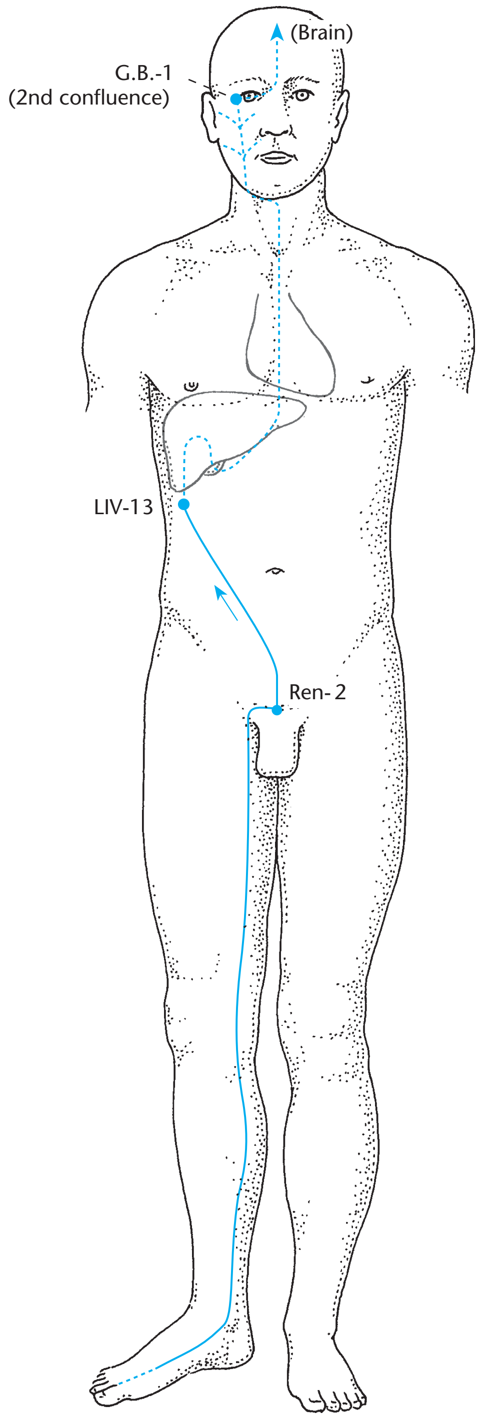

Pathway

The Liver divergent channel separates from the Liver primary channel on the dorsum of the foot

- ➡ ascends the medial aspect of the leg to the inguinal region
- ➡ meets the Gall Bladder divergent channel near → **Ren-2** (*qugu*)
- ➡ traverses the abdomen and enters internally at **LIV-13** (*zhangmen*)
- ➡ continues inside the thoracic cavity, spreads in the Liver (*gan*), connects with the Gall Bladder (*dan*) and runs to the Heart (*xin*)
- ➡ ascends to the neck
- ➡ emerges at the mandibular angle and disperses over the face
- ➡ connects with the Gall Bladder primary channel and the Gall Bladder divergent channel at → **G.B.-1** (*tongziliao*) on the outer canthus of the eye to form one of the six *he*-confluences (here: G.B./LIV as 2nd confluence → 1.3)
- ➡ continues to the eye system and ends in the Brain.

Clinical importance

- Strengthens the relationship between the Liver and the Gall Bladder (*zangfu*-Organ systems). Points on the Liver primary channel can therefore be used to treat disorders of the Gall Bladder and vice versa.
- Based on the pathways of the Gall Bladder and Liver divergent channels, points on both channels can be used for disorders of the hips and the lower extremities as well as for disorders of the eye system.

Ch04.12-F10028.qxd 2/23/08 11:31 AM Page 446

4.12.3 The Liver Sinew Channel (*zu jue yin jing jin*)

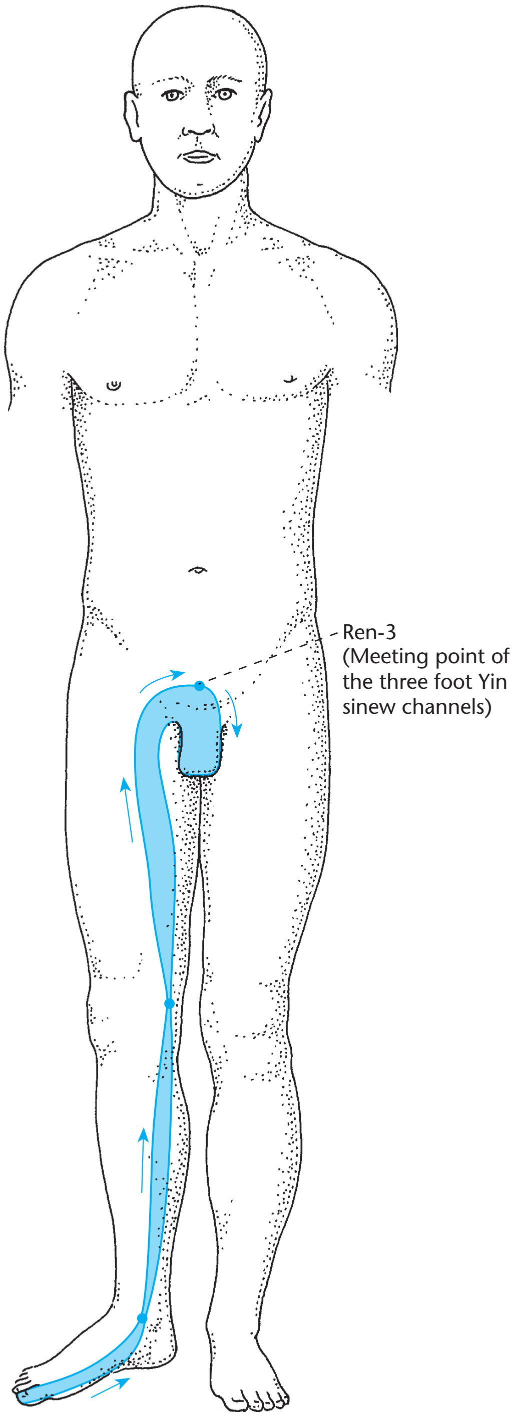

Pathway

The Liver sinew channel begins on the dorsum of the big toe

➡ follows the dorsolateral aspect of the big toe to the anterior aspect of the medial malleolus, where it binds (*jie*)

➡ ascends the medial aspect of the tibia and binds (*jie*) at the medial aspect of the knee

➡ further ascends the medial aspect of the thigh between the Kidney and Spleen sinew channels

➡ reaches the inguinal region, meeting the other foot Yin sinew channels at → **Ren-3** (*zhongji*) and continues to the genitalia.

Clinical importance

Pathology: Stiffness, tension and distending sensations along the big toe. Pain around the medial malleolus and medial aspect of the knee. Pain and muscular tension along the medial aspect of the thigh. Disorders of the genital region.

Indication: Pain, tension, muscle contractions and distending sensations along the pathway of the channel. Disorders of the genital region.

4.12.4 The Liver *luo*-Connecting Vessel System (*zu jue yin luo mai*)

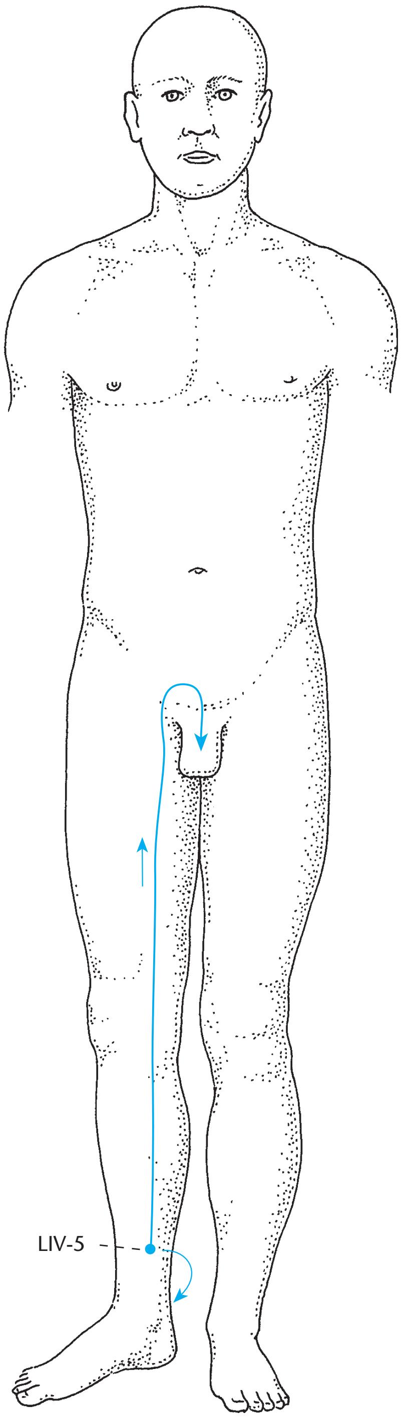

Pathway

The Liver *luo*-connecting vessel system separates from the Liver primary channel at its *luo* point **LIV-5** (*ligou*) (→ 8.1.2). It forms a three-dimensional reticular network, dividing into multiple branches and sub-branches (*sun luo, fu luo, xue luo* → 1.5) within the surrounding tissue.

- Horizontal divisions run to the Interiorly–Exteriorly paired Gall Bladder primary channel; according to some schools of thought (for example, Ngyen Van Nghi → Appendix), they travel as a **transverse** Liver *luo*-connecting vessel to the *yuan*-source point → **G.B.-40** (*qiuxu*).
- **A longitudinal division** ascends the medial aspect of the leg to the genital region, where it branches out.

Clinical importance (→ 8.1.2)

Pathology

Counterflow Qi: Swellings and pain of the testicles and scrotum, hernia

Excess (*shi*): Priapism

Deficiency (*xu*): Itching of the external genitalia

4.12.5 Cutaneous Region (*jue yin pi bu*)

See description and figures → 1.6.

4.12.6 Points of the Liver Primary Channel (Overview)

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): LIV-3** (*taichong*) 🟦🟦
- ***Luo*-connecting point (→ 8.1.2.): LIV-5** (*ligou*) 🟦⬛
- ***Xi*-cleft point (→ 8.1.3): LIV-6** (*zhongdu*) 🟦⬛
- **Associated Back-*shu* point (→ 8.1.4): BL-18** (*ganshu*) 🟦⬛
- **Associated Front-*mu* point (→ 8.1.5): LIV-14** (*qimen*) 🟦⬛
- **Five *shu*-transporting points (→ 8.1.6):**
  jing-well point (Wood): **LIV-1** (*dadun*) 🟦⬛, *ben* point (Five Phases)
  ying-spring point (Fire): **LIV-2** (*xingjian*) 🟦🟦
  shu-stream point (Earth): **LIV-3** (*taichong*) 🟦🟦
  jing-river point (Metal): **LIV-4** (*zhongfeng*)
  he-sea point (Water): **LIV-8** (*ququan*) 🟦⬛, tonification point
- ***Hui*-meeting point (→ 8.1.7) of the *zang*-Organs: LIV-13** (*zhangmen*) 🟦🟦
- **Opening point (→ 8.1.8): –**
- **Lower *he*-sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  – with the G.B. channel and the *dai mai**: **LIV-13** (*zhangmen*) 🟦🟦

---

*Mentioned by only some authors

  - with the *yin wei mai* and the ST channel: **LIV-14** (*qimen*) 🟦⬛
  - of other channels with the LIV channel: **SP-6, SP-12, SP-13, Ren-2, Ren-3, Ren-4, P-1, (Du-20*)**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): –**
- **Points of the Four Seas (→ 8.1.13): –**
- **Ma Dan Yang Heavenly Star point (→ 8.1.14): LIV-3** (*taichong*) 🟦🟦
- **Sun Si Miao Ghost point (→ 8.1.15): –**
- **Other functional points:**
  - Front-*mu* point of the Spleen: **LIV-13** (*zhangmen*) 🟦🟦

Points according to region

- **Local points** (→ 8.2.1): Liver – **LIV-13** (*zhangmen*) 🟦🟦; knee – **LIV-7** (*xiguan*), **LIV-8** (*ququan*) 🟦⬛; foot – **LIV-4** (*zhongfeng*)
- **Adjacent points (→ 8.2.1):** Stomach/Spleen – **LIV-13** (*zhangmen*) 🟦🟦; hypochondrium – **LIV-13** (*zhangmen*) 🟦🟦, **LIV-14** (*qimen*)
- **Distal points** (→ 8.2.1): vertex – **LIV-3** (*taichong*) 🟦🟦; eyes – **LIV-2** (*xingjian*) 🟦🟦, **LIV-3** (*taichong*) 🟦🟦; Liver – **LIV-3** (*taichong*) 🟦🟦; Gall Bladder – **LIV-3** (*taichong*) 🟦🟦; urogenital region – **LIV-3** (*taichong*) 🟦🟦; genital region – **LIV-5** (*ligou*) 🟦⬛; lower abdomen – **LIV-8** (*ququan*) 🟦⬛

Specific points according to the channel pathway (in numeric order)

- **LIV-1** (*dadun*) 🟦⬛: *jing*-well point (Wood), *ben* point (Five Phases) (→ 8.1.6)
- **LIV-2** (*xingjian*) 🟦🟦: *ying*-spring point (Fire) (→ 8.1.6), sedation point; distal point for the eyes (→ 8.2.1)
- **LIV-3** (*taichong*) 🟦🟦: *yuan*-source point (→ 8.1.1); *shu*-stream point (Earth) (→ 8.1.6); Ma Dan Yang Heavenly Star point (→ 8.1.4); distal point for the vertex, eyes, Liver, Gall Bladder and urogenital region (→ 8.2.1)
- **LIV-4** (*zhongfeng*): *jing*-river point (Metal) (→ 8.1.6); local point for the foot (→ 8.2.1)
- **LIV-5** (*ligou*) 🟦⬛: *luo*-connecting point (→ 8.1.2); distal point for the genital region (→ 8.2.1)
- **LIV-6** (*zhongdu*) 🟦⬛: *xi*-cleft point (→ 8.1.3)
- **LIV-7** (*xiguan*): local point for the knee (→ 8.2.1)
- **LIV-8** (*ququan*) 🟦⬛: *he*-sea point (Water) (→ 8.1.6); tonification point; distal point for the abdomen (→ 8.2.1); local point for the knee (→ 8.2.1)
- **LIV-13** (*zhangmen*) 🟦🟦: Front-*mu* point of the Spleen (→ 8.1.5); *hui*-meeting point of the *zang*-Organs (→ 8.1.7); *jiaohui*-meeting point with the G.B. channel and the (*dai mai**) (→ 8.1.10); adjacent point for the Spleen/Stomach and the hypochondrium (→ 8.2.1)
- **LIV-14** (*qimen*) 🟦⬛: associated Front-*mu* point(→ 8.1.5); *jiaohui*-meeting point with the *yin wei mai* and the SP channel (→ 8.1.10); regional point for the hypochondrium (→ 8.2.1)

---

* Mentioned by only some authors

Big Mound DADUN **LIV-1**

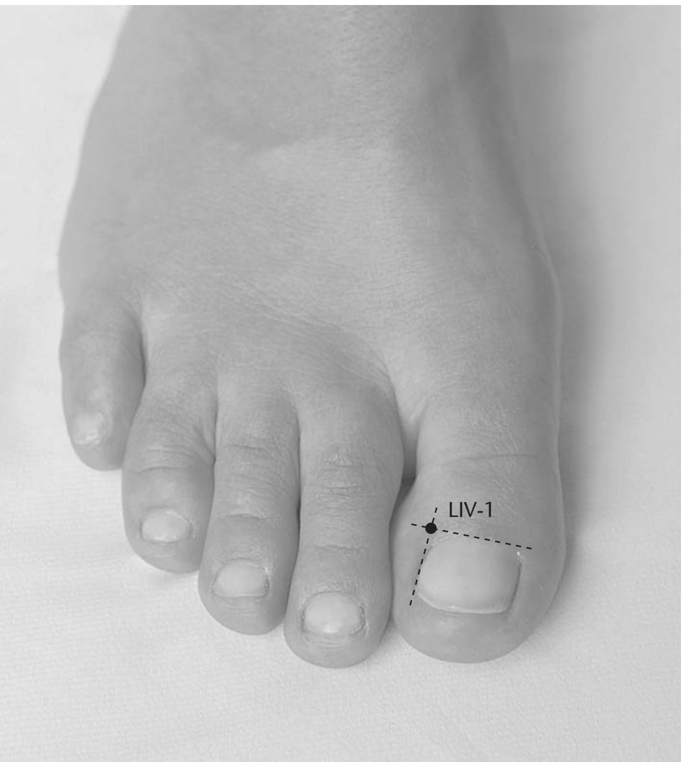

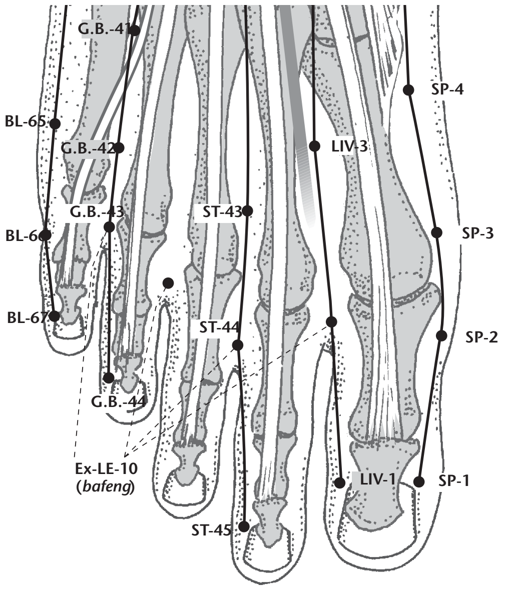
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

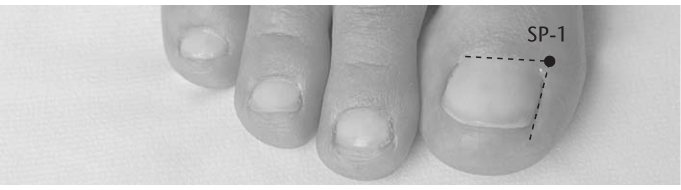

Location

On the lateral aspect of the big toe, 0.1 cun proximal and lateral to the corner of the nail.

How to find

LIV-1 is located at the junction of two tangents along the proximal and lateral borders of the big toe, 0.1 cun from the actual margin of the nail. → **SP-1** is located at the medial corner of the big toe.

Needling

Vertically or obliquely 0.1–0.2 cun. Avoid needling into the perionychium. For excess conditions, prick to bleed. Caution: Painful point.

Actions/Indications

- Regulates the Lower Burner, the genitals and Liver Qi
- Eliminates Dampness
- Opens the sensory orifices

Special features

Jing-well point, Wood point, *ben* point (Five Phases), entry point.

LIV-2 *Moving Between* XINGJIAN

Location

Between the 1st and 2nd toes, proximal to the margin of the interdigital web.

How to find

Locate **LIV-2** slightly proximal to the margin of the interdigital web between the 1st and 2nd toes.
Located in comparable positions are → **ST-44** (between the 2nd/3rd toes) and → **G.B.-43** (between the 4th/5th toes). → **LIV-2,** → **ST-44** and → **G.B.-43** are all part of the extra point → **Ex-LE-10** (*bafeng*). Its counterpart on the hand is the extra point → **Ex-UE-9** (*baxie*).

Needling

0.3–0.5 cun obliquely towards the heel or vertically.

Actions/Indications

- Clears Liver Fire, spreads Liver Qi, calms (internal) Liver Wind, clears Heat and Blood Heat, stops bleeding, benefits the Lower Burner

Special features

Ying-spring point, Fire point, sedation point. Major point for excess syndromes of the Liver (especially Liver Fire).

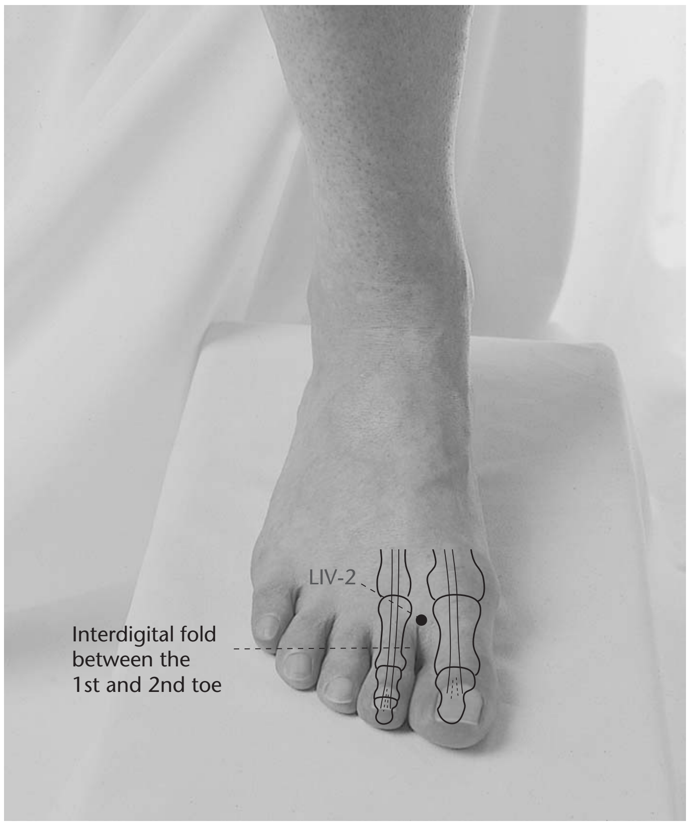

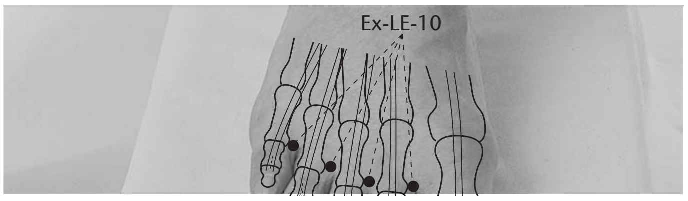

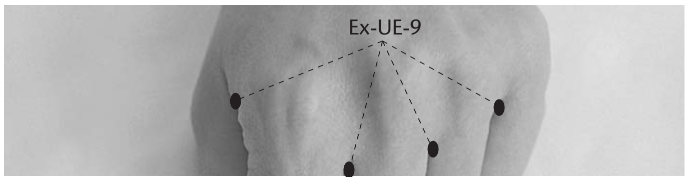

Ch04.12-F10028.qxd 2/23/08 11:32 AM Page 451

Great Rushing **TAICHONG** **LIV-3**

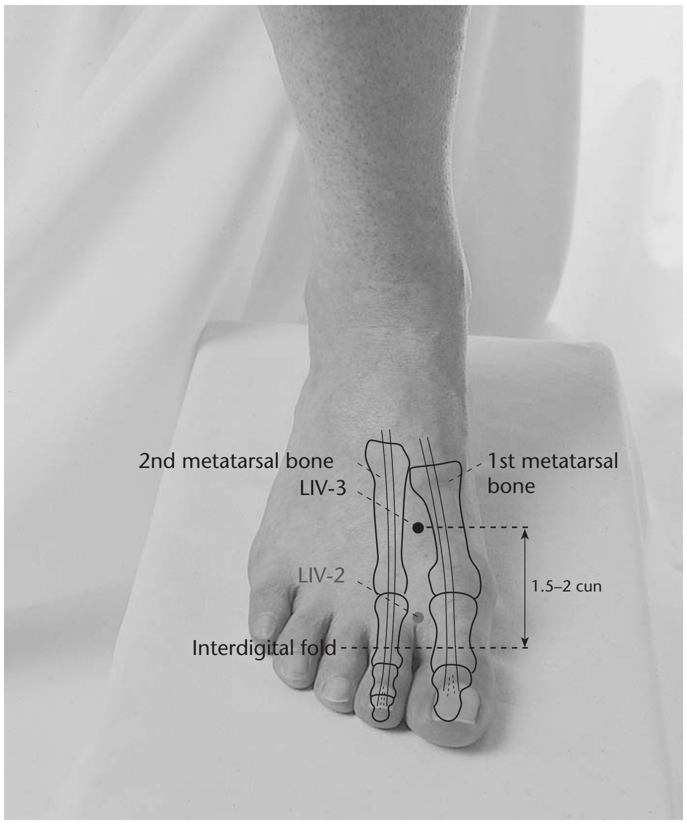
2nd metatarsal bone
LIV-3
1st metatarsal bone
LIV-2
1.5–2 cun
Interdigital fold

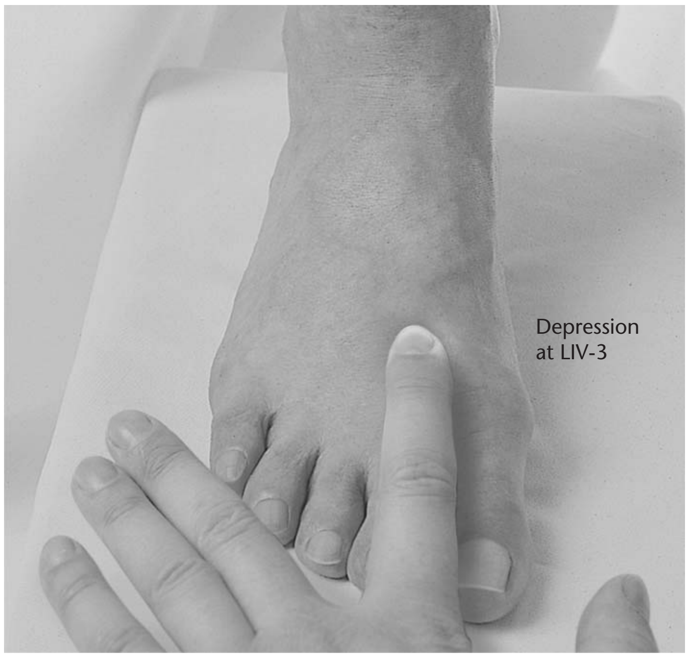
Depression at LIV-3

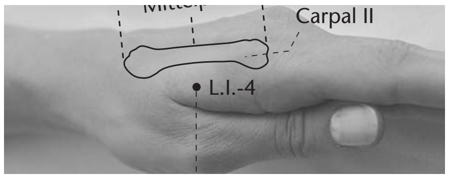
Carpal II
L.I.-4

Location

On the dorsum of the foot, between the 1st and 2nd metatarsal bones, in the depression proximal to the metatarsophalangeal joints and the proximal angle between the two bones.

How to find

From the web between the 1st and 2nd toes, palpate proximally past the metatarsophalangeal joints and along the groove between the 1st and 2nd metatarsal bones, until you reach the widest and deepest part of the groove. There, locate **LIV-3**, which is often sensitive to pressure. To double-check: When continuing to palpate along the groove, it will become more narrow and shallow again.

Located in a similar position on the hand is ➞ **L.I.-4** (between the 1st and 2nd metacarpal bones).

Needling

Vertically or slightly obliquely 0.5–1 cun. Caution during pregnancy.

Actions/Indications

- Spreads Liver Qi, clears the head, benefits the eyes, nourishes Liver Blood and Liver Yin, calms spasms, alleviates pain, regulates menstruation and the Lower Burner
- Eliminates (internal) Wind, pacifies Liver Yang

Special features

Yuan-source point, *shu*-stream point, Earth point, Ma Dan Yang Heavenly Star point (later added by Xu Feng). Major point for spreading Liver Qi.

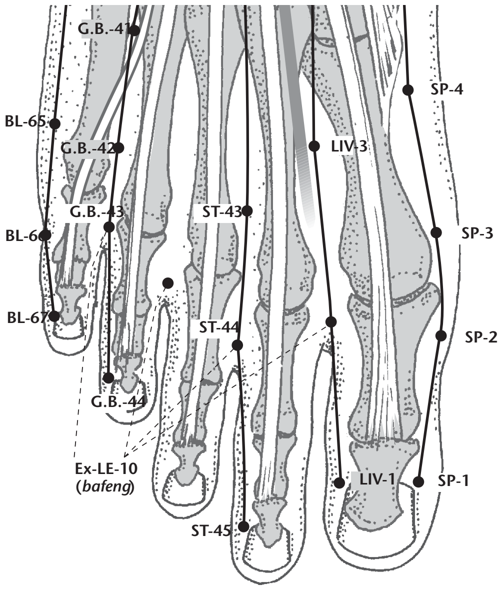
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(bafeng)
LIV-1
SP-1
ST-45

LIV-4 *Middle Seal* ZHONGFENG

Location

1 cun anterior to the prominence of the medial malleolus, medial to the tendon of the tibialis anterior muscle.

How to find

Even with slight dorsiflexion of the ankle, the tendon of the tibialis anterior muscle will become more pronounced on the medial aspect of the ankle. **LIV-4** is located between the tendon and the prominence of the medial malleolus, in a depression over the palpable joint space.

→ **ST-41** is also located over the joint space (on the midpoint of its anterior aspect); → **SP-5** lies only slightly distal to → **LIV-4** at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Opens the channel, promotes the flow of Liver Qi, regulates the Lower Burner

Special features

Jing-river point, Metal point.

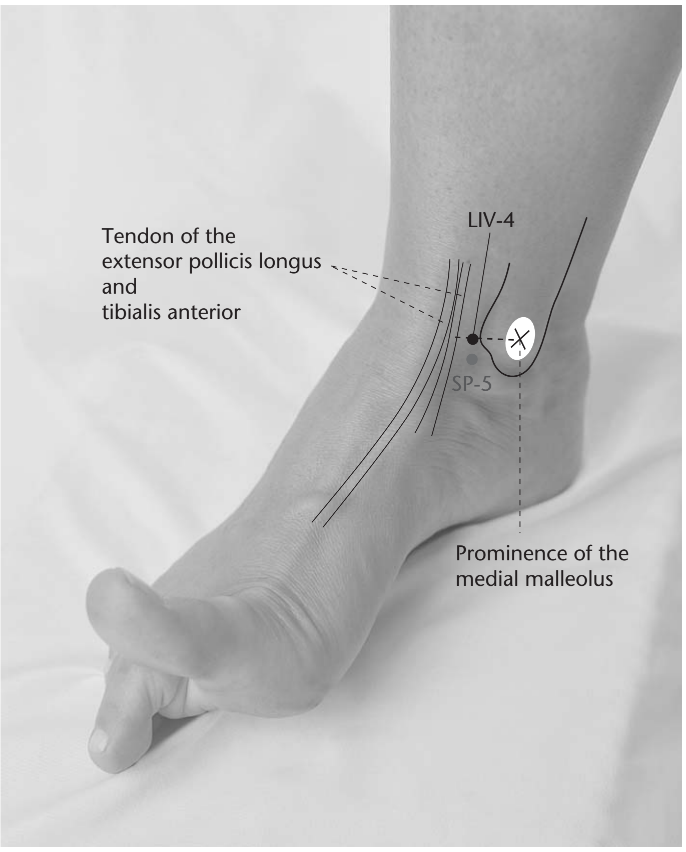
Tendon of the extensor pollicis longus and tibialis anterior
LIV-4
SP-5
Prominence of the medial malleolus

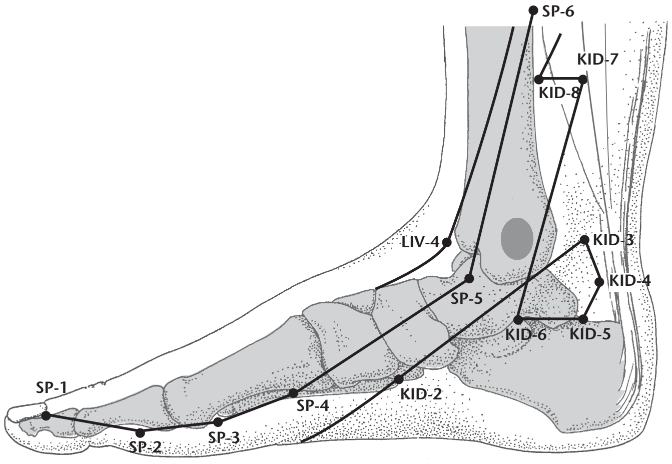
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

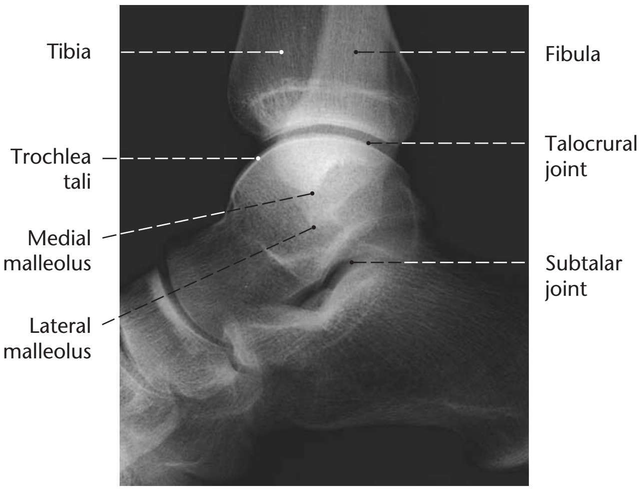
Tibia
Fibula
Trochlea tali
Talocrural joint
Medial malleolus
Subtalar joint
Lateral malleolus

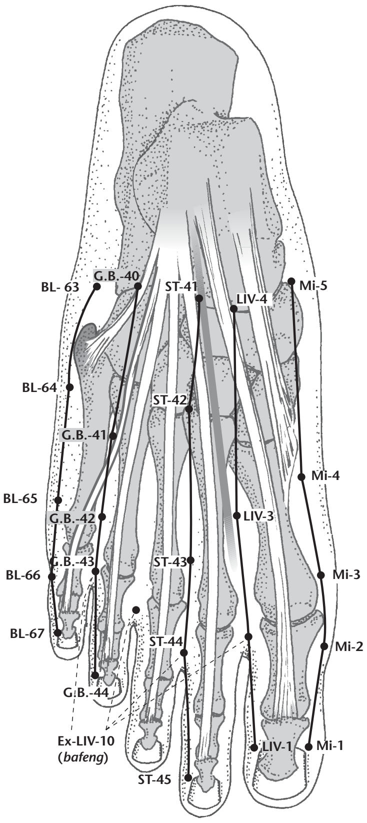
BL-63
G.B.-40
ST-41
LIV-4
Mi-5
BL-64
ST-42
G.B.-41
Mi-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
Mi-3
BL-66
BL-67
ST-44
Mi-2
G.B.-44
Ex-LIV-10 (*bafeng*)
LIV-1
Mi-1
ST-45

Woodworm Canal LIGOU LIV-5

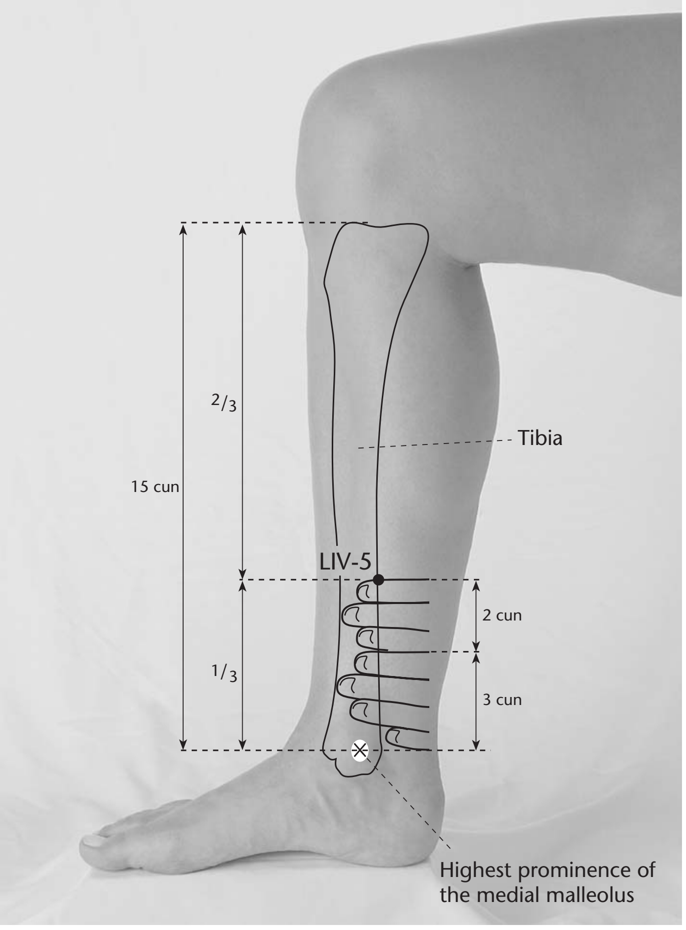

Location

On the medial aspect of the lower leg, 5 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia, between the crest of the tibia and the gastrocnemius muscle.

How to find

First, locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate 5 cun in a proximal direction and there locate **LIV-5** in a depression directly posterior to the medial crest of the tibia. **Or quick method:** Divide the distance between the medial end of the popliteal crease and the highest prominence of the medial malleolus (=approximately 15 cun, → 2.2) into thirds and locate **LIV-5** a third of the distance from the medial malleolus. Note: Some texts locate **LIV-5 on** the tibia, others on the **posterior** border of the tibia; sensitivity to pressure should be the determining factor. → **KID-9** is located on the same level but directly superior to → **KID-3** (between the medial malleolus and the Achilles tendon).

Needling

0.5–1 cun vertically or obliquely in a posterior direction towards the fibula or transversely (subcutaneously) along the tibia towards the abdomen.

Actions/Indications

- Regulates Liver Qi, benefits the genitals, clears Damp-Heat from the Lower Burner, regulates menstruation
- Eliminates 'plum-pit' Qi
- Moves Qi locally

Special features

Luo-connecting point. Important distal point for the urogenital region (especially with Liver Qi stagnation and Damp-Heat).

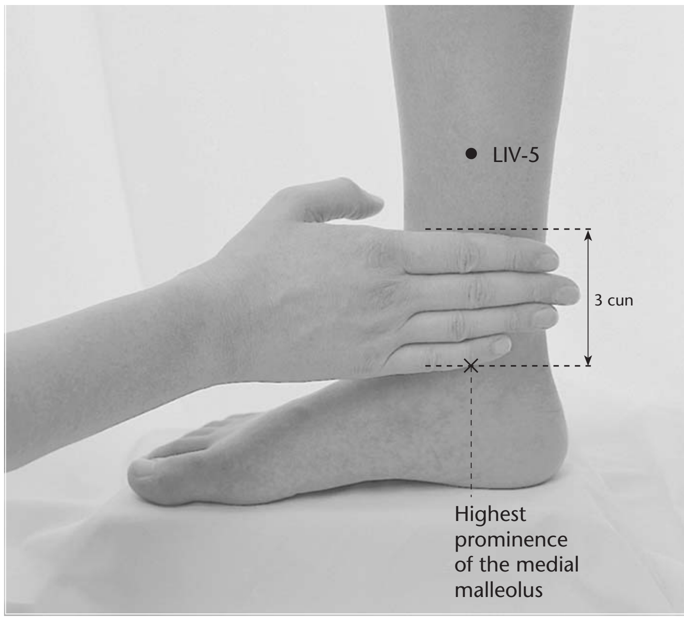

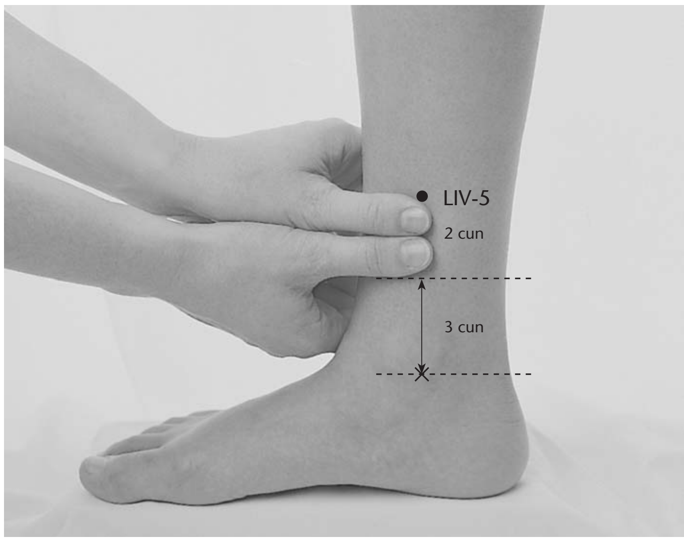

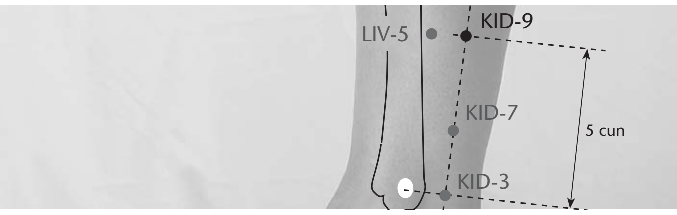

LIV-6 *Central Capital* ZHONGDU

Location

7 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia.

How to find

First, locate the highest prominence of the medial malleolus (→ 3.6.2). From there, palpate 7 cun (= the width of 2 hands and 1 thumb) in a proximal direction and there locate **LIV-6** in a depression directly posterior to the medial crest of the tibia. **Or quick method:** Spreading hands technique (→ 2.3.3): Place the little fingers on the medial end of the popliteal crease and the highest prominence of the medial malleolus respectively and join the thumbs on the posterior border of the tibia at the mid-point of this distance (= approximately 15 cun, 2.2). From there, measure 0.5 cun in a distal direction and here locate **LIV-6** posterior to the medial crest of the tibia. Note: Some texts locate **LIV-6 on** the tibia, others on the **posterior** border of the tibia; sensitivity to pressure should be the determining factor.

Needling

0.5–1 cun vertically or obliquely in a posterior direction towards the fibula or transversely (subcutaneously) in a proximal direction along the tibia.

Actions/Indications

- Spreads Liver Qi, regulates the Lower Burner, regulates Blood, eliminates Dampness

Special features

Xi-cleft point

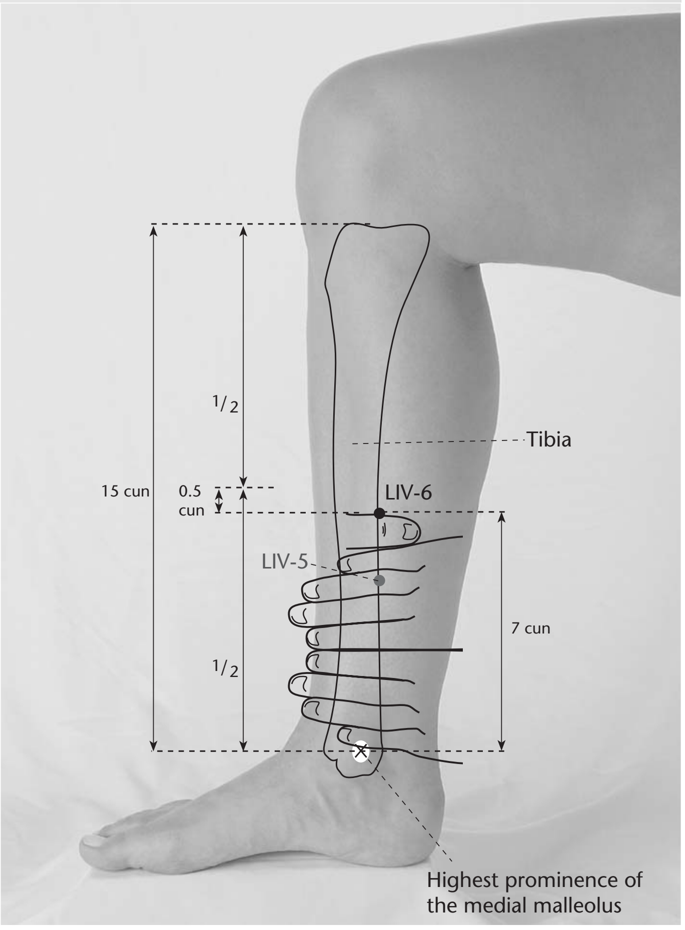

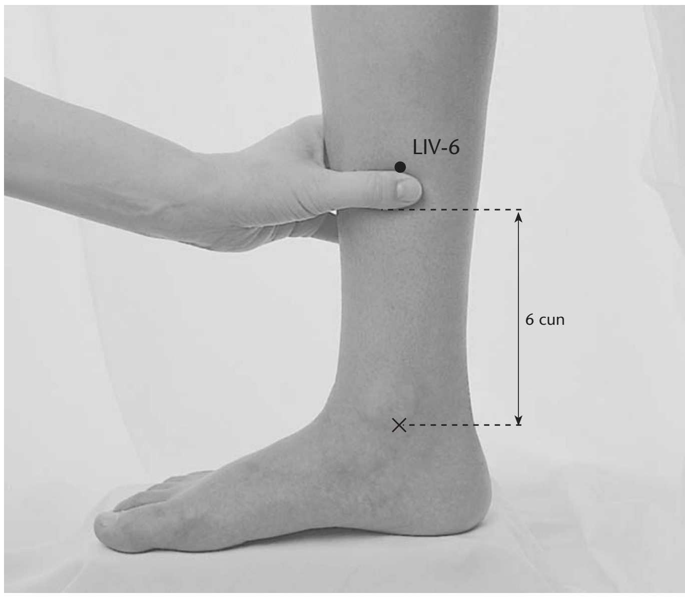

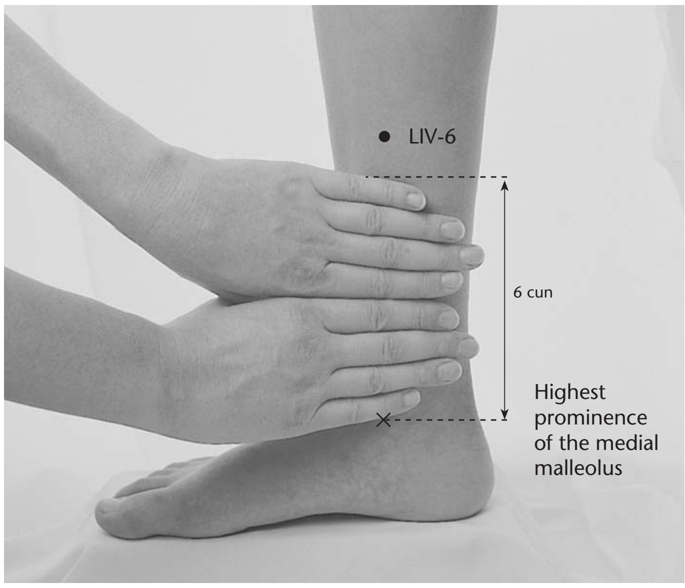

Knee Joint XIGUAN LIV-7

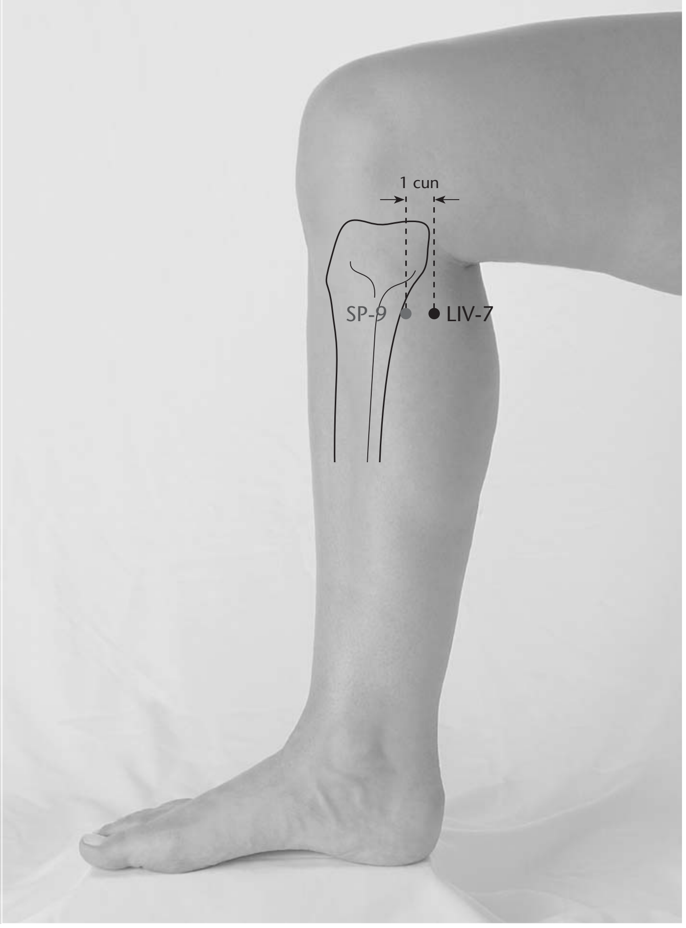
1 cun
SP-9
LIV-7

Location

At the junction of the shaft and the medial condyle of the tibia, 1 cun posterior to → **SP-9**.

How to find

The junction of the shaft and the medial condyle of the tibia can be easily palpated. There, locate → **SP-9** and **LIV-7** 1 cun posterior to it. Both points tend to be sensitive to pressure if indicated.

Needling

Vertically 1–2 cun

Actions/Indications

- Opens the channel
- Expels Wind and Dampness

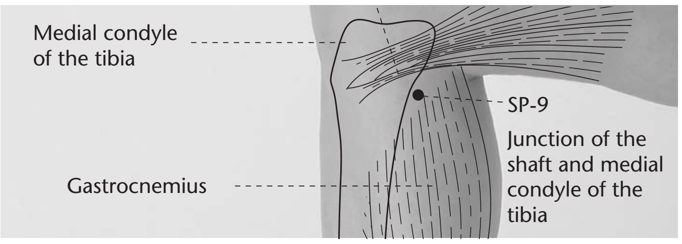
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

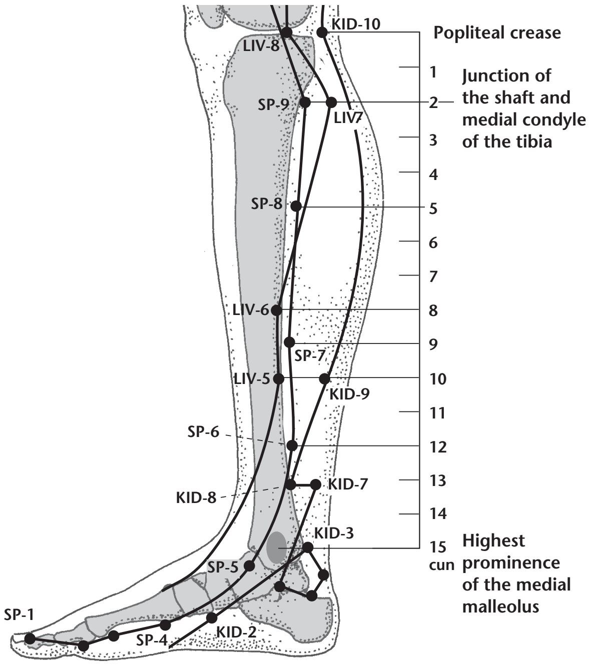
KID-10
Popliteal crease
LIV-8
1
Junction of the shaft and medial condyle of the tibia
SP-9
2
LIV7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-8
KID-7
13
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
SP-4
KID-2

LIV-8 *Spring at the Crook* QUQUAN

Location

With the knee flexed, **LIV-8** is located directly proximal to the medial end of the popliteal crease, in a depression anterior to the tendons of the semitendinosus and semimembranosus muscles.

How to find

LIV-8 is best located with the patient's knee flexed and slight external rotation of the hip. First, palpate the prominent, ropey tendon of the semitendinosus muscle, which becomes more pronounced with the knee flexed. The tendon of the semimembranosus muscle lies deep to the semitendinosus and is often more difficult to isolate. With the palpating finger, glide from the tendon of the semitendinosus towards the patella. Approximately after 1 cun, you can feel a depression between the muscle bellies. **LIV-8** is located anterior to both tendons. Some authors locate **LIV-8** 1 cun superior and anterior to → **KID-10**. Sensitivity to pressure should determine the location.

→ **KID-10** is located approximately 1 cun in the direction of the popliteal crease and, from a medial perspective, posterior to the tendon of the semimembranosus and anterior to the tendon of the semitendinosus.

Needling

Vertically 0.5–1.5 cun

Actions/Indications

- Clears and eliminates Dampness and Damp-Heat from the Lower Burner (major action), benefits the genitals and the Uterus
- Nourishes Blood and Yin
- Moves Qi locally

Special features

he-sea point, Water point, tonification point.

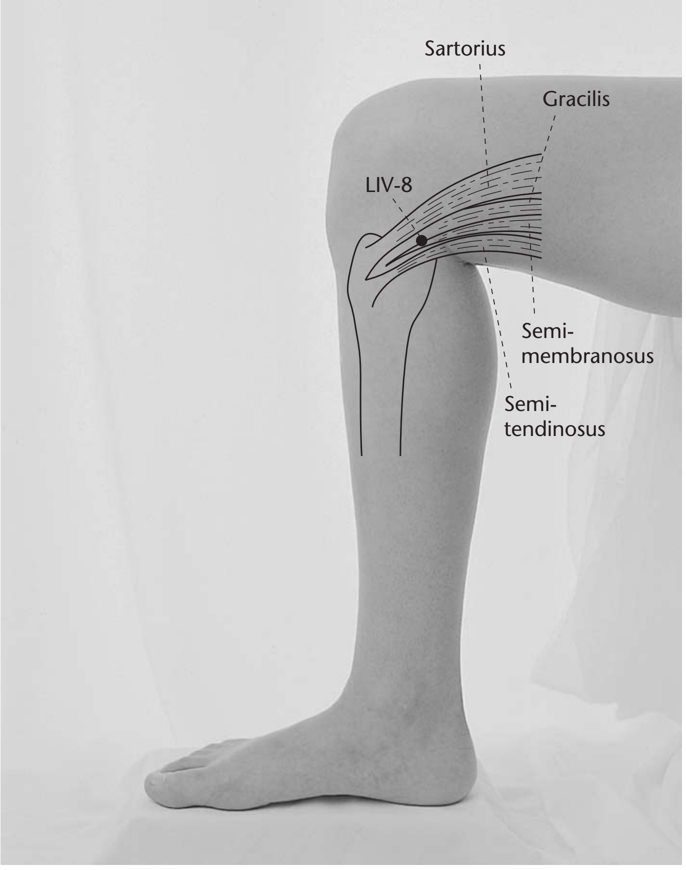

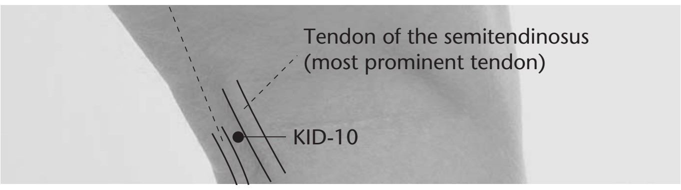

Yin Wrapping YINBAO **LIV-9**

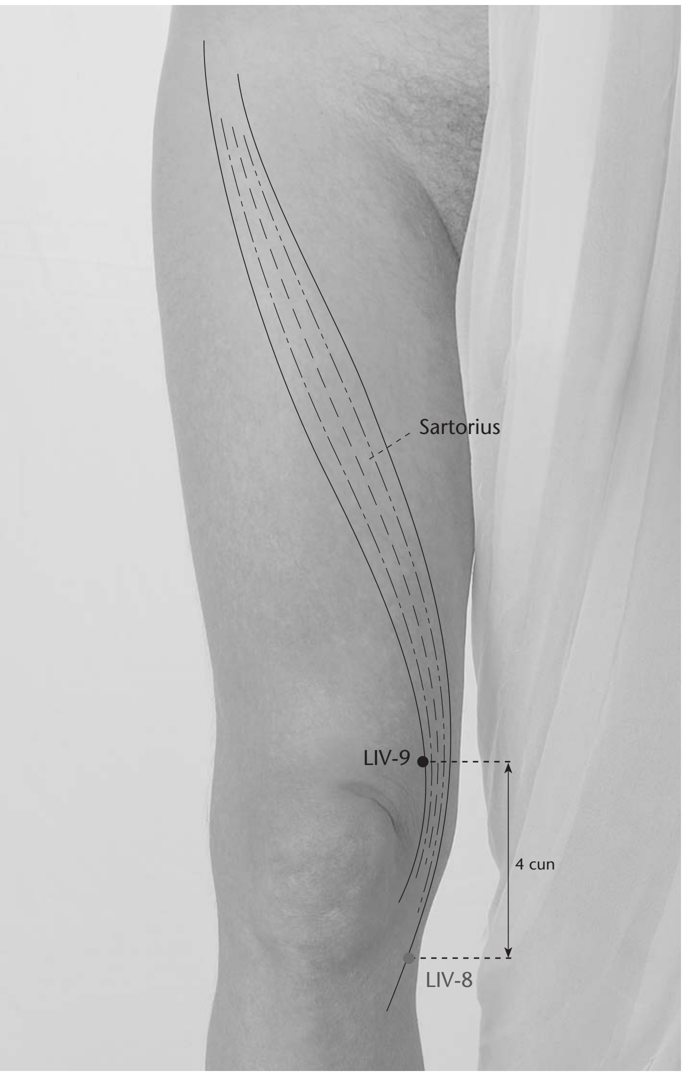
Sartorius
LIV-9
4 cun
LIV-8

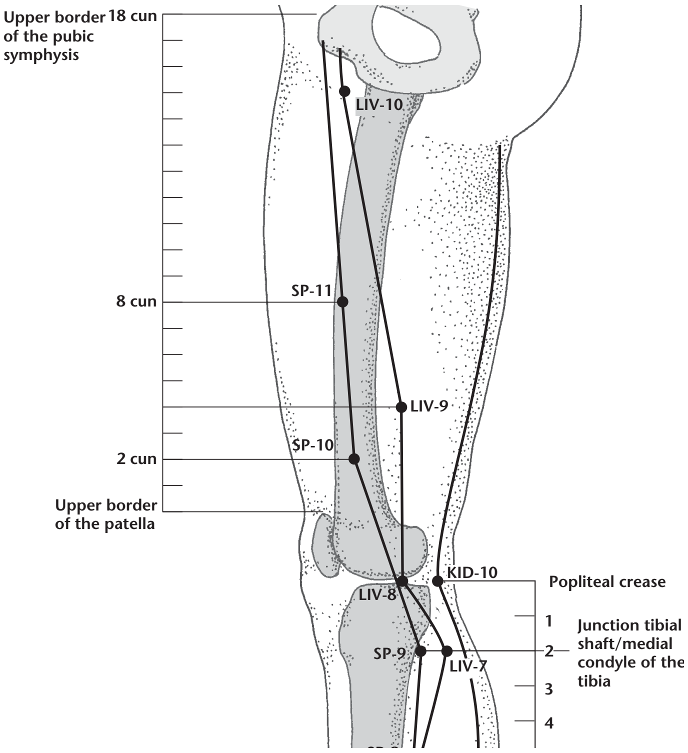
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
SP-10
2 cun
Upper border of the patella
KID-10
Popliteal crease
LIV-8
1
Junction tibial shaft/medial condyle of the tibia
SP-9
2
LIV-7
3
4

Location

4 cun proximal to the medial condyle of the femur, between the sartorius and vastus medialis muscles. Alternative locations: 4 cun proximal to ➞ **LIV-8** (for example, Deadman et al 1998, ➞ photo), 4 cun proximal to the junction of the shaft and the condyle of the femur (for example, Hecker et al ➞ see line drawing) or 4 cun proximal to the midpoint of the condyle of the femur (for example, Ogal and Stör 1999, ➞ illustration on p. 443).

How to find

From ➞ **LIV-8**, measure 4 cun in a proximal direction. There, locate **LIV-9** on the border between the more medial, narrow sartorius muscle and the anteromedial vastus medialis muscle. Sensitivity to pressure should determine the location of this point.

Needling

Vertically 1–2 cun

Actions/Indications

- Benefits the Lower Burner
- Opens the channel

Ch04.12-F10028.qxd 2/23/08 11:32 AM Page 458

LIV-10 *Leg Five Miles* **ZUWULI**

Location

On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of the adductor longus muscle.

How to find

Ask the patient to flex their adductors by pressing their feet against each other. **LIV-10** is located in a groove on the medial border of the adductor longus muscle, where it forms an angle with the sartorius muscle.

Needling

Vertically 0.5–1.5 cun. Caution: Great saphenous vein, femoral artery/nerve.

Actions/Indications

- Relaxes the tendons and muscles
- Clears Damp-Heat from the Lower Burner
- Regulates the water passages

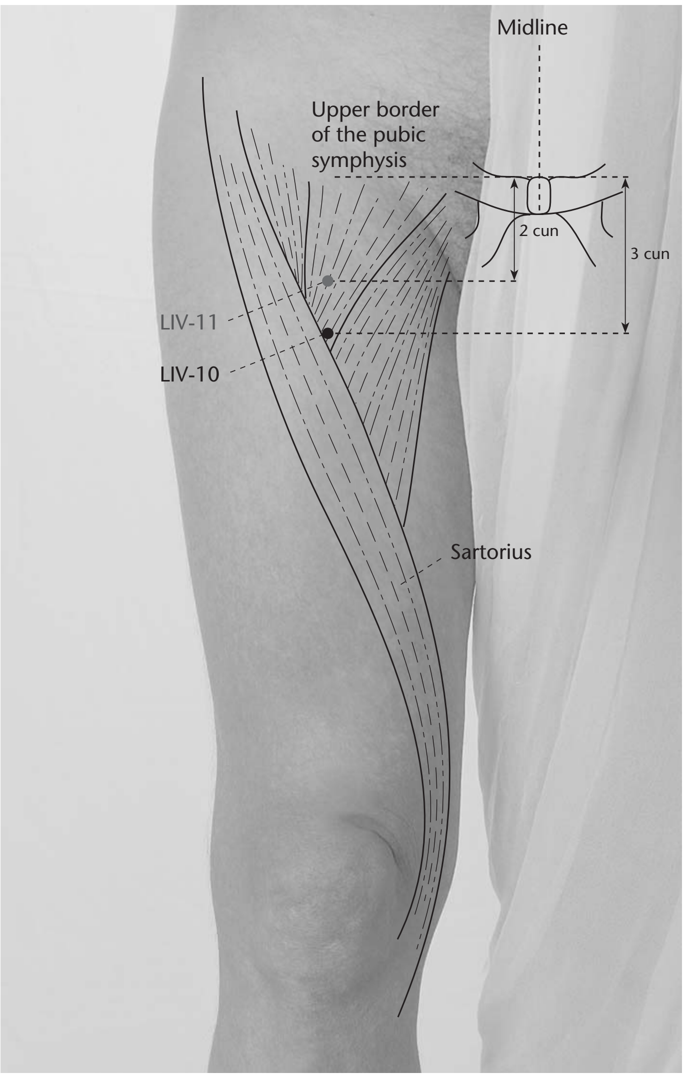
Midline
Upper border of the pubic symphysis
2 cun
3 cun
LIV-11
LIV-10
Sartorius

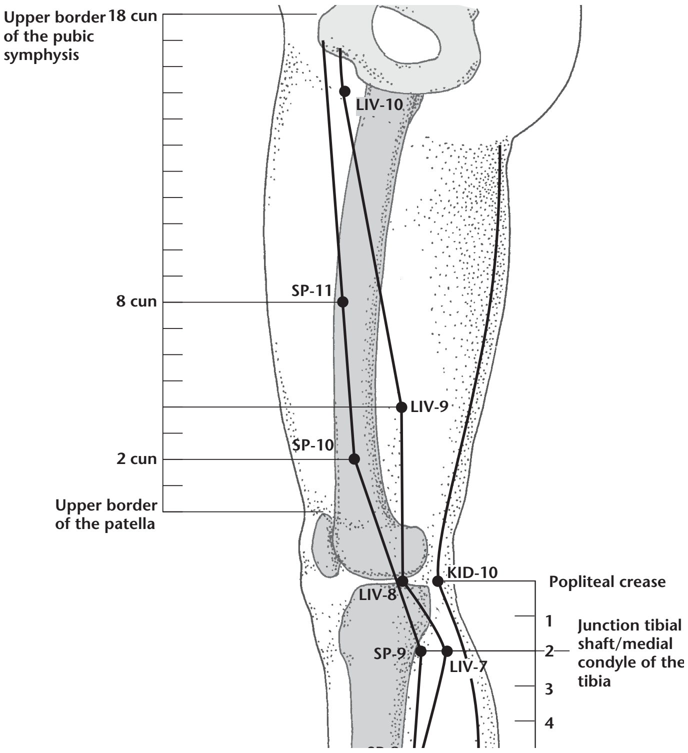
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
2 cun
SP-10
Upper border of the patella
KID-10
Popliteal crease
LIV-8
1
2
Junction tibial shaft/medial condyle of the tibia
SP-9
LIV-7
3
4

Yin Corner YINLIAN LIV-11

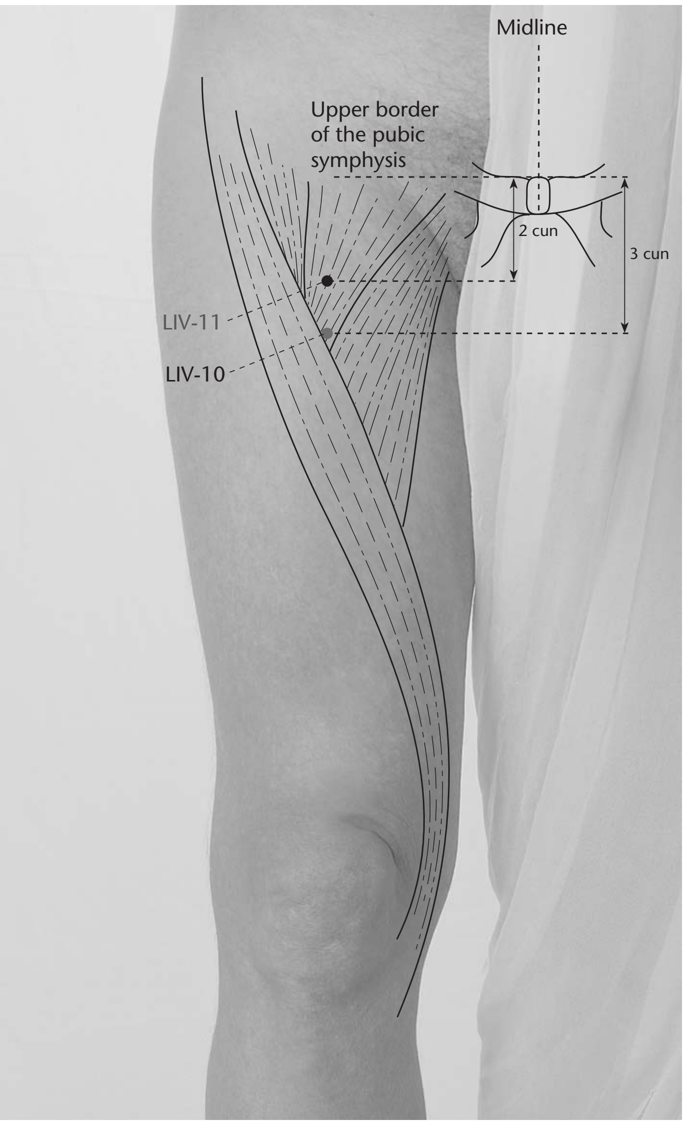
Midline
Upper border of the pubic symphysis
2 cun
3 cun
LIV-11
LIV-10

Location

On the anterior aspect of the thigh, 2 cun inferior to the upper border of the symphysis and on the lateral border of the adductor longus muscle.

How to find

Ask the patient to flex their adductors by pressing their feet against each other. **LIV-11** is located 2 cun inferior to the upper border of the pubic symphysis, in a groove on the medial border of the adductor longus muscle.

LIV-11 is located approximately 1 cun inferior to where the femoral artery passes under the inguinal ligament.

Needling

Vertically 0.5–1.5 cun. Caution: Great saphenous vein, femoral artery/vein/nerve. Classical texts recommend moxibustion for female sterility.

Actions/Indications

- Relaxes the tendons
- Benefits the Uterus and regulates menstruation
- For disorders of the lumbar spine

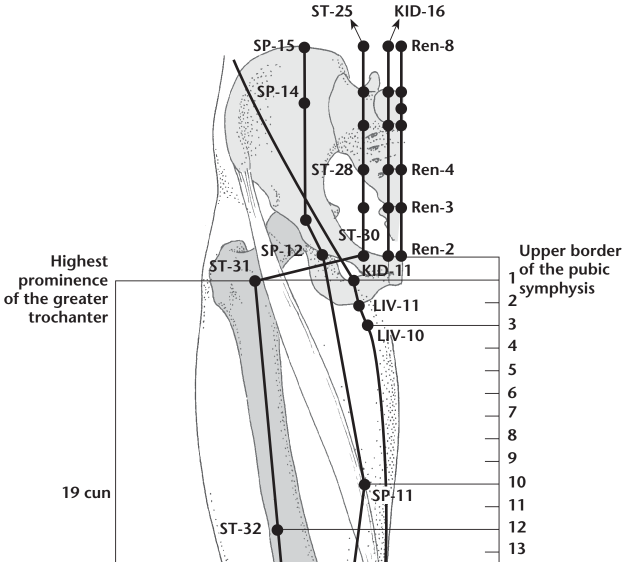
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
SP-12
Ren-2
Highest prominence of the greater trochanter
ST-31
KID-11
Upper border of the pubic symphysis
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
19 cun
SP-11
ST-32

LIV-12 *Urgent Pulse* JIMAI

Location

In the inguinal groove, 2.5 cun lateral to the anterior midline and 1 cun inferior to the upper border of the symphysis.

How to find

From the upper border of the pubic symphysis, measure 2.5 cun in a lateral and 1 cun in an inferior direction. In this area, you can palpate the pulse of the femoral artery. **LIV-12** is located medial to the pulsating of the artery. As the femoral vein also runs medial to the artery, **LIV-12** should be needled in a medial direction, at least 1 fingerbreadth from the artery.

→ **ST-30** is located 1 cun superior and 0.5 cun medial to **LIV-12**.

Needling

Slightly obliquely 0.5–0.8 cun. Caution: Femoral artery/vein. The femoral vein is located medial to the artery and is approximately the breadth of a finger. To avoid injury to the vein, **LIV-12** should not be needled medial to the artery. Owing to its tricky location, classical texts recommend moxibustion only, while in modern texts moxibustion is contraindicated for the same reason.

Actions/Indications

- Dispels Cold from the Liver channel and benefits the Lower Burner

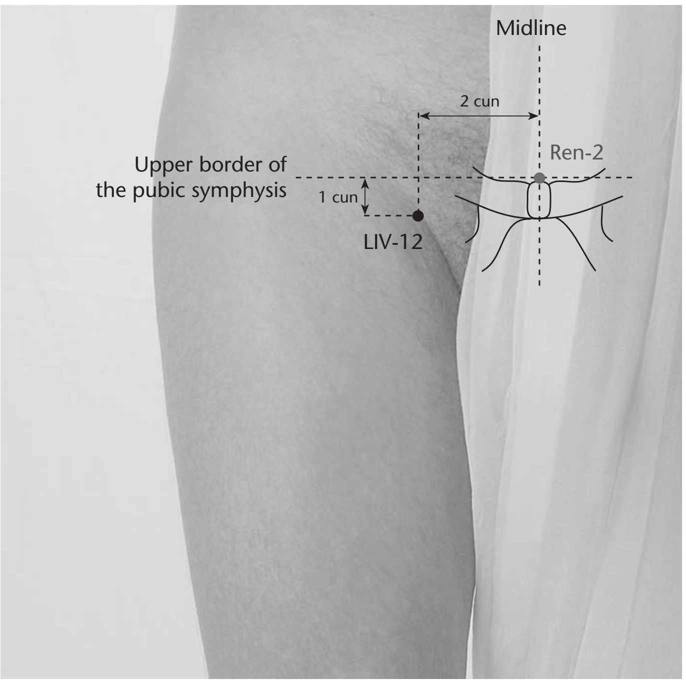
Midline
2 cun
Ren-2
Upper border of
the pubic symphysis
1 cun
LIV-12

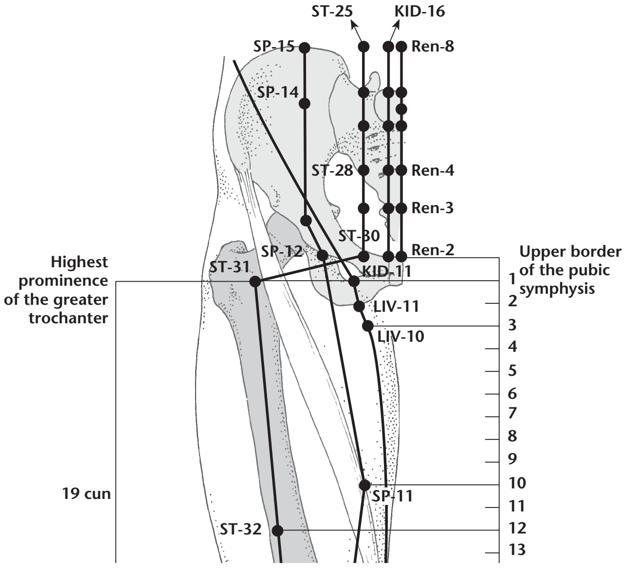
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
SP-12
Ren-2
ST-31
KID-11
Highest
prominence
of the greater
trochanter
Upper border
of the pubic
symphysis
LIV-11
LIV-10
1
2
3
4
5
6
7
8
9
10
11
12
13
19 cun
SP-11
ST-32

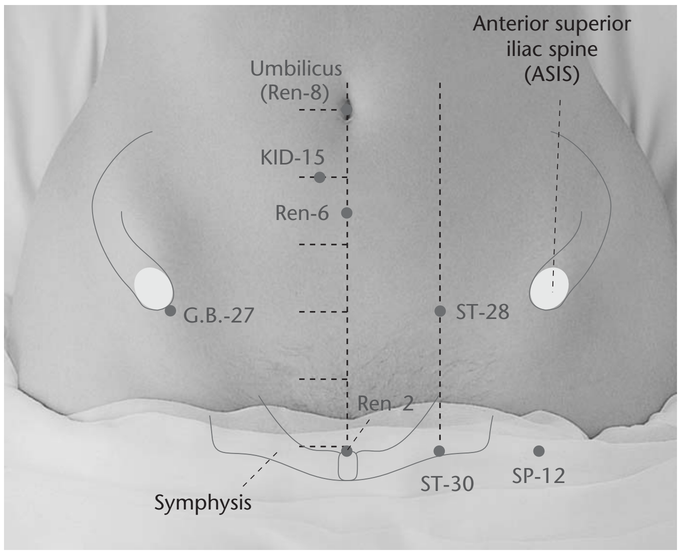
Anterior superior
iliac spine
(ASIS)
Umbilicus
(Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

Completion Gate ZHANGMEN LIV-13

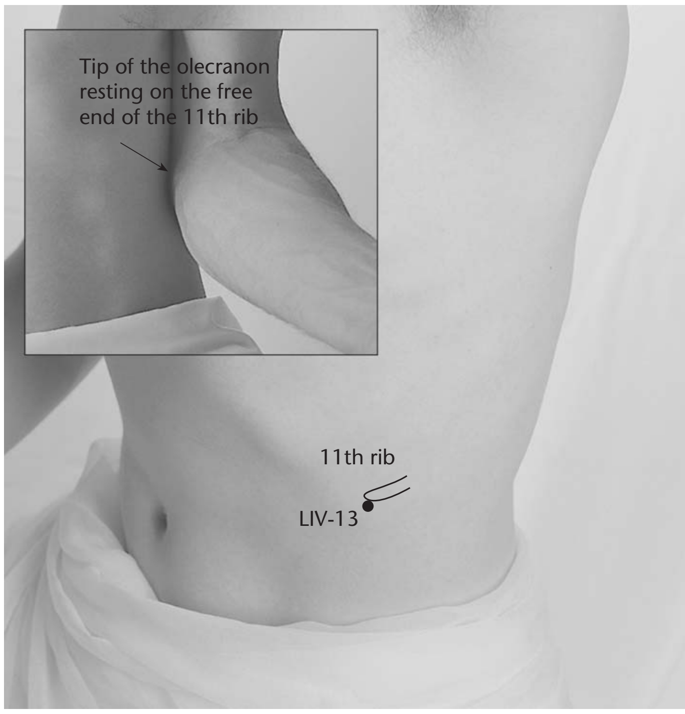
Tip of the olecranon resting on the free end of the 11th rib
11th rib
LIV-13

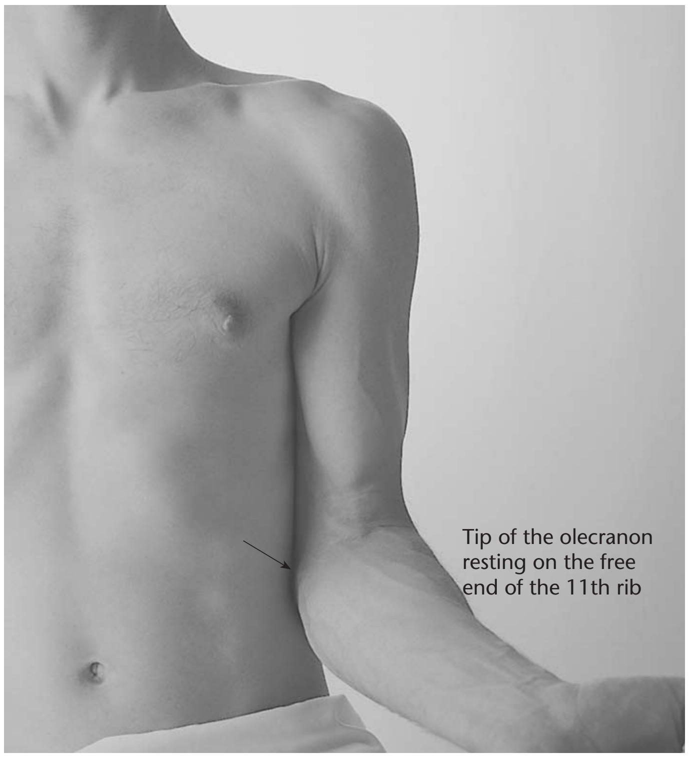
Tip of the olecranon resting on the free end of the 11th rib

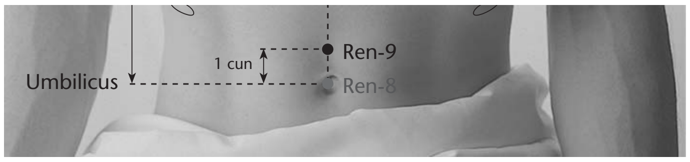
Umbilicus
1 cun
Ren-9
Ren-8

Location

Anterior and inferior to the free end of the 11th rib, approximately 1 cun superior to the umbilicus.

How to find

Quick method: Ask the patient to press their flexed elbow (90°) against the lateral aspect of the thorax and the tip of the olecranon will be pointing to the free end of the 11th rib. **LIV-13** can then be located on its anterior and inferior border. **Or:** Place the hand on the upper abdomen and, with gentle pressure, palpate along the lower border of the rib cage until you can feel the free end of the 11th rib just superior to the umbilicus. Locate **LIV-13** on its anterior and inferior border.

Located on approximately the same level (1 cun superior to the umbilicus) are → **Ren-9** (on the midline) and → **ST-24** (2 cun lateral to the midline). → **G.B.-25** is located more laterally and inferiorly, at the free end of the 12th rib.

Needling

0.5–1 cun vertically or better obliquely in a lateral direction. Caution: Peritoneum, hypertrophied organs (right: liver, left: spleen).

Actions/Indications

- Harmonises the Liver and the Spleen, regulates Liver Qi (especially in the Middle and Lower Burner), strengthens the Spleen
- Opens the channel

Special features

Front-*mu* point of the Spleen, meeting point with the *dai mai* and the G.B. channel. *Hui*-meeting point of the *zang*-Organs.

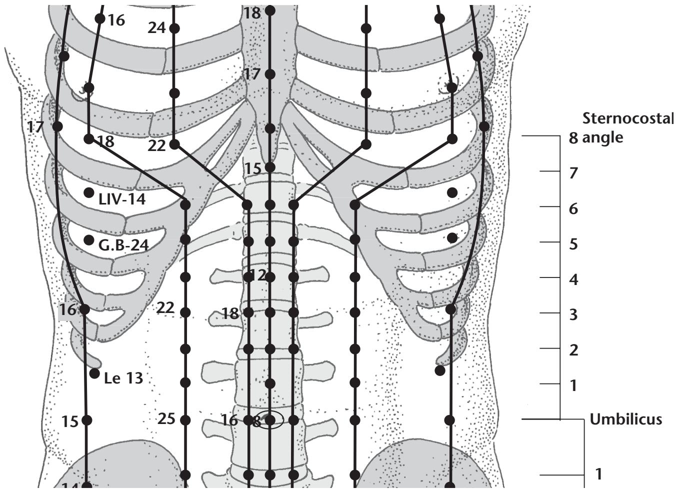
16
24
18
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
14
Sternocostal angle
8
7
6
5
4
3
2
1
Umbilicus
1

LIV-14 *Cycle Gate* QIMEN

Location

In the 6th intercostal space, on the mamillary line or 4 cun lateral to the midline.

How to find

Quick method, especially in men: The nipple is generally located on the level of the 4th intercostal space. From there, palpate downward 2 intercostal spaces and locate **LIV-14** in the 6th intercostal space. **Or:** For more precise orientation on the ribcage (→ 3.5), palpate for the manubriosternal synchondrosis, a well-defined horizontal bony structure on the sternum. Lateral to it, locate the 2nd rib and below it the 2nd intercostal space. From there, count downward 4 intercostal spaces to the 6th intercostal space and there locate **LIV-14** on the mamillary line (4 cun lateral to the midline).
→ **G.B.-24** is located in the 7th intercostal space. Located on the same level (2 cun inferior to the sternocostal angle → 3.5) are → **Ren-14** (on the midline), → **KID-21** (0.5 cun lateral to the midline) and → **ST-19** (2 cun lateral to the midline).

Needling

0.5–1 cun obliquely, either in a lateral or medial (along the course of the ICS) direction. The latter is recommended for disorders of the breasts. Caution: Pneumothorax.

Actions/Indications

- Regulates Liver Qi and Liver Blood (especially in the Upper and Middle Burner), cools the Blood, disperses accumulations, harmonises the Liver and the Stomach

Special features

Front-*mu* point of the Liver, meeting point with the *yin wei mai* and the Spleen channel, exit point.

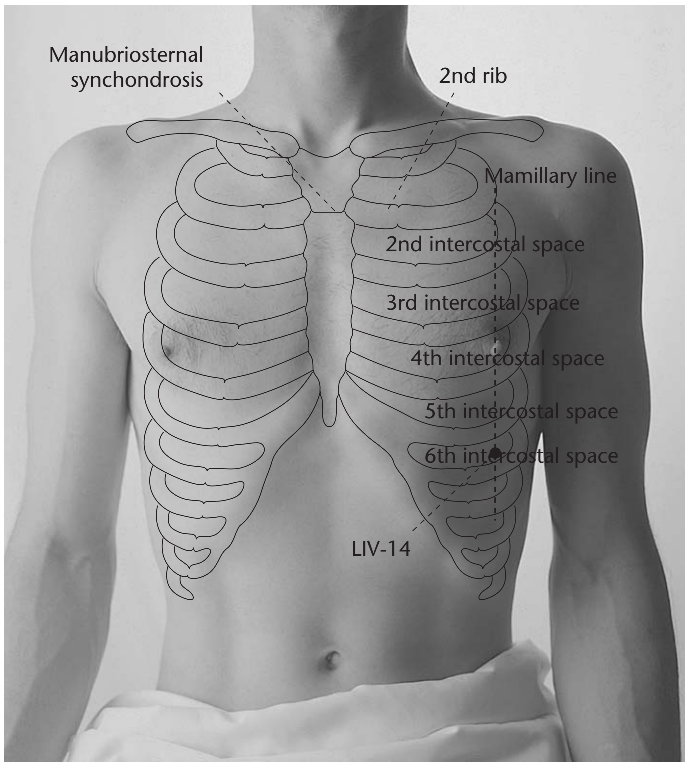
Manubriosternal synchondrosis
2nd rib
Mamillary line
2nd intercostal space
3rd intercostal space
4th intercostal space
5th intercostal space
6th intercostal space
LIV-14

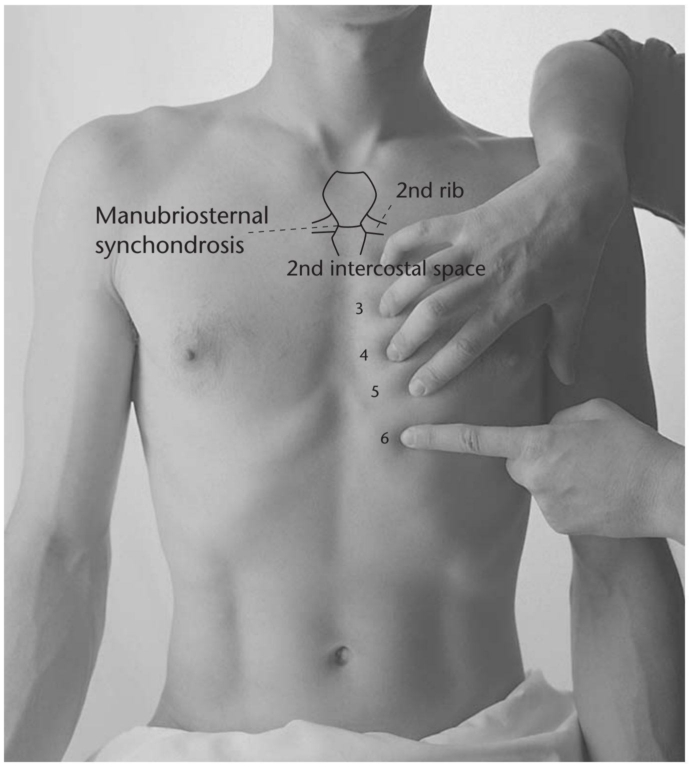
Manubriosternal synchondrosis
2nd rib
2nd intercostal space
3
4
5
6

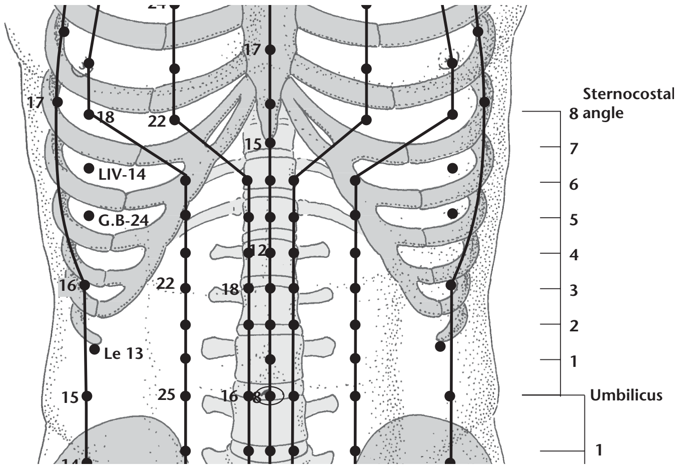
17
17
18
22
15
LIV-14
G.B-24
12
16
22
18
Le 13
15
25
16
8
Sternocostal
8 angle
7
6
5
4
3
2
1
Umbilicus
1

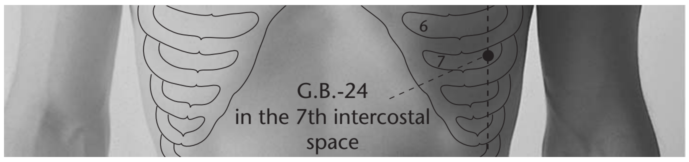
6
7
G.B.-24
in the 7th intercostal space

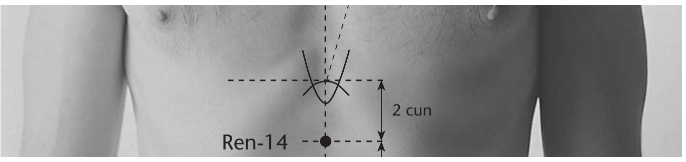

5 The Eight Extraordinary Vessels (*qi jing ba mai*) with Points

Claudia Focks, Ulrich März

For an overview of the theoretical background of the eight extraordinary vessels → 1.7.

5.1 *Chong mai*

Synonym: Penetrating Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *chong mai*/*dai mai*
- **Central/peripheral:** *chong mai*/*yin wei mai*. Regions supported by this pairing: Heart, thorax, Stomach. Opening point: SP-4 (*gongsun*), coupled point: P-6 (*neiguan*)

Pathway

The extraordinary vessel *chong mai* originates like the *du mai* and *ren mai* in the lower abdomen at *bao zhong* (in women in the Uterus) or, according to some authors, at the Kidneys. It crosses the genital area and divides into two branches:

- ➡ A **deep branch** reaches the perineum at → **Ren-1** (*huiyin*), travels to the spine, penetrates the spinal canal and ascends to the level of T1 and T2. (According to some authors the vessel ascends the anterior aspect of the spine.)
- ➡ The **other branch** emerges at → **ST-30** (*qichong*) to the surface of the body and divides into **two external branches**, one descending and one ascending branch:
- the **ascending** branch reaches → **KID-11** (*henggu*) 0.5 cun lateral to the midline at the level of the pubic symphysis and from there follows the KID channel to → **KID-21** (*youmen*). The KID channel initially ascends the abdomen at a distance of 0.5 cun from the midline, passing → **KID-12** (*dahe*), → **KID-13** (*qixue*), → **KID-14** (*siman*) and → **KID-15** (*zhongzhu*), runs to the midline, meeting → **Ren-7** (*yinjiao*) 1 cun inferior to the umbilicus, and again returns to its pathway 0.5 cun lateral to the midline, passing → **KID-16** (*huangshu*), → **KID-17** (*shangqu*), → **KID-18** (*shiguan*), → **KID-19** (*yindu*), → **KID-20** (*futonggu*) and → **KID-21** (*youmen*). The vessel then disperses into the intercostal spaces, ascends the neck, circles the inner surface of the lips and spreads in the upper pharynx and the nasal cavities.

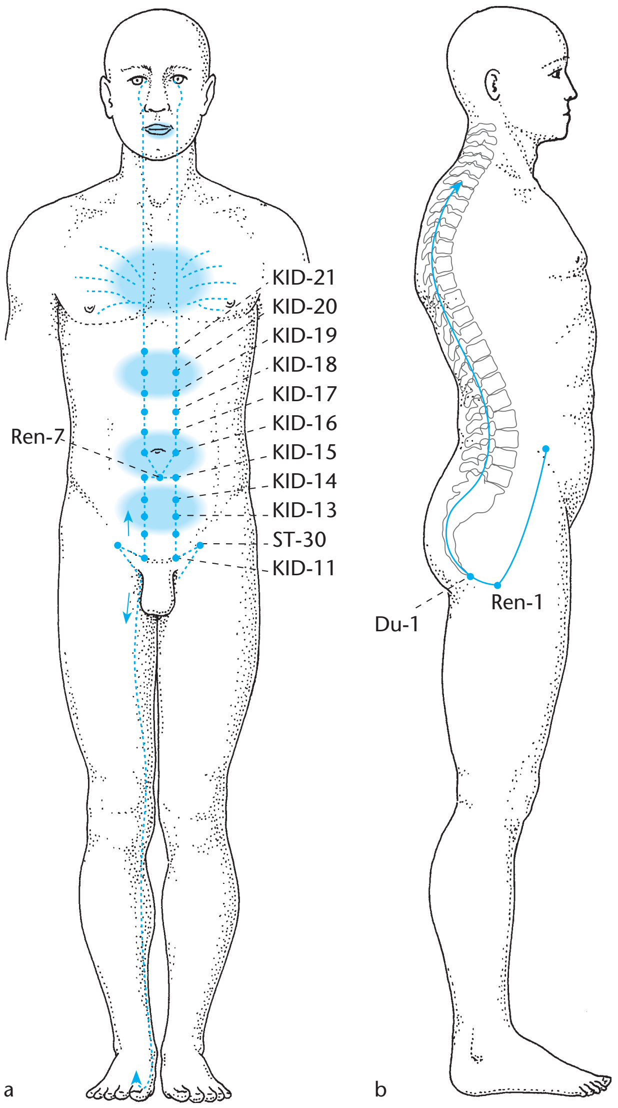

- the **descending branch** meets the Kidney primary channel and descends the medial aspect of the thigh (according to the *Ling Shu* it divides into the three foot Yin channels), penetrates the popliteal fossa, follows the medial border of the tibia and emerges posterior to the medial malleolus, where it divides into two sub-branches: one branch travels along the anteromedial aspect of the foot to the big toe, the other branch reaches the sole of the foot.

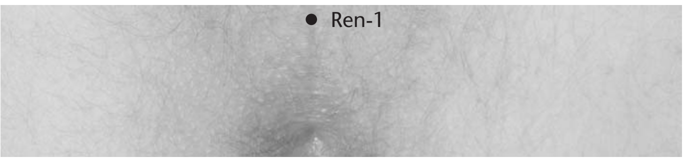

Ren-1 (*huiyin*): In the centre of the perineum.

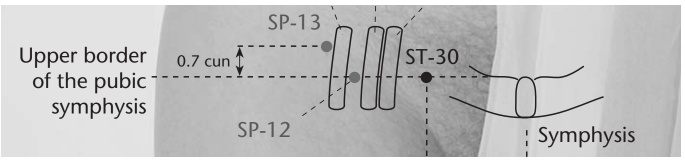
SP-13
Upper border of the pubic symphysis
0.7 cun
ST-30
SP-12
Symphysis

ST-30 (*qichong*): 2 cun lateral to the upper border of the pubic symphysis, medial to the femoral artery; at the level of Ren-2

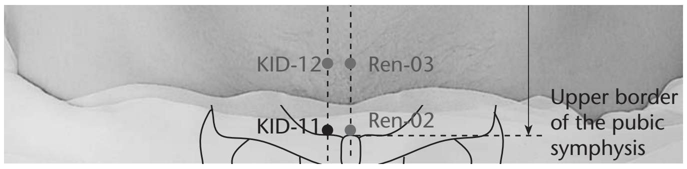
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

KID-11 (*henggu*): On the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

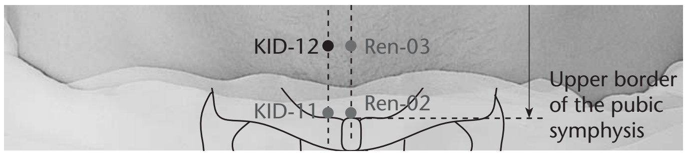
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

KID-12 (*dahe*): 1 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

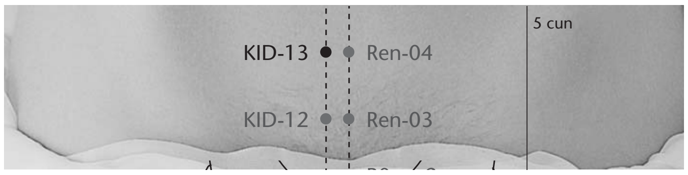
5 cun
KID-13
Ren-04
KID-12
Ren-03

KID-13 (*qixue*): 2 cun superior to the upper border of the pubic symphysis, 0.5 cun lateral to the anterior midline

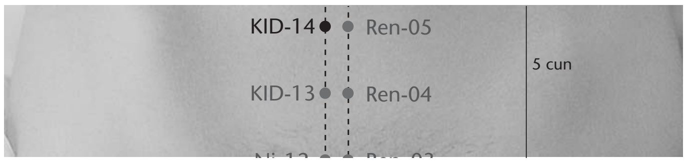
KID-14
Ren-05
5 cun
KID-13
Ren-04

KID-14 (*siman*): 2 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline

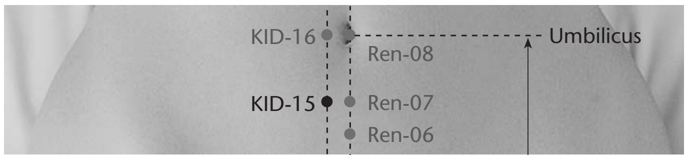
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

KID-15 (*zhongzhu*): 1 cun inferior to the umbilicus, 0.5 cun lateral to the anterior midline

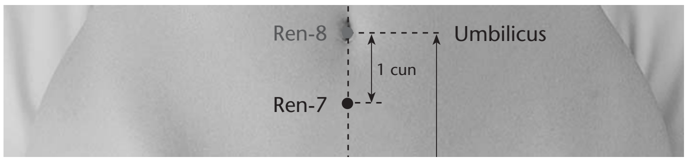
Ren-8
Umbilicus
1 cun
Ren-7

Ren-7 (*yinjiao*): On the anterior midline, 1 cun inferior to the umbilicus

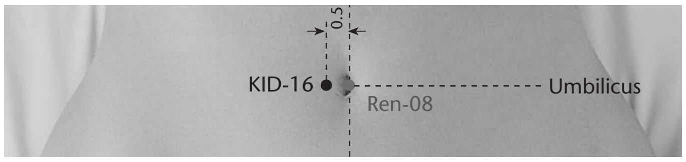
0.5
KID-16
Umbilicus
Ren-08

KID-16 (*huangshu*): 0.5 cun lateral to the centre of the umbilicus

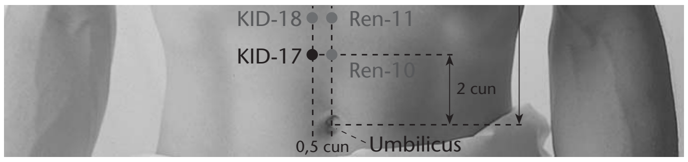
KID-18
Ren-11
KID-17
Ren-10
2 cun
0,5 cun
Umbilicus

KID-17 (*shangqu*): 2 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline

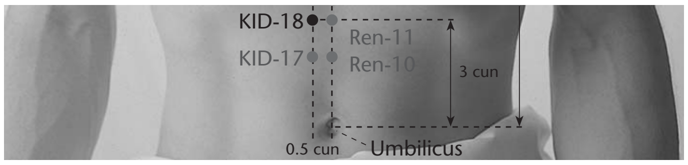
KID-18
Ren-11
KID-17
Ren-10
3 cun
0.5 cun
Umbilicus

KID-18 (*shiguan*): 3 cun superior to the umbilicus, 0.5 cun lateral to the anterior midline

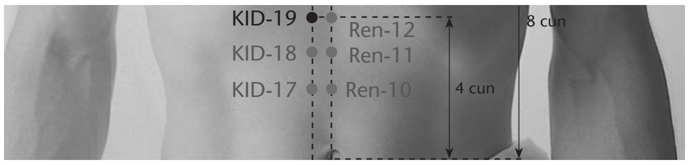
KID-19
Ren-12
8 cun
KID-18
Ren-11
KID-17
Ren-10
4 cun

KID-19 (*yindu*): Midway between the sternocostal angle and the centre of the umbilicus, 0.5 cun lateral to the anterior midline

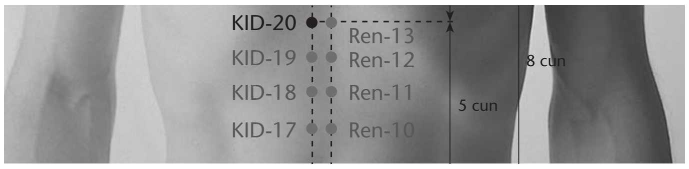

KID-20 (*futonggu*): 5 cun superior to the umbilicus or 3 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline

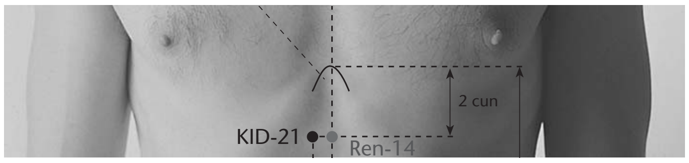

KID-21 (*youmen*): 2 cun inferior to the sternocostal angle, 0.5 cun lateral to the anterior midline

Connections with other channels and organs

- Kidney primary channel, *ren mai*, *du mai*, Stomach and Liver primary channels
- Kidney, Uterus

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of Blood**, strong connection to the various functions of the Blood
- Strong relationship to the Uterus, regulates the menses
- Regulates the 12 primary channels, effectively regulates the circulation of the Blood (*xue*) in the whole body
- Regulates counterflow Qi
- Connects the Stomach and Kidney primary channels and strengthens the relationship between the *ren mai* and *du mai*

5.2 *Ren mai*

Synonym: Conception Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *ren mai/du mai*
- **Central/peripheral:** *ren mai/yin qiao mai*. Regions supported by this pairing: face, throat, thorax, Lungs, diaphragm, abdomen. Opening point: **LU-7** (*lieque*); coupled point: **KID-6** (*zhaohai*)

Pathway

The extraordinary vessel *ren mai* originates like the *chong mai* and *du mai* in the lower abdomen (*bao zhong*) or, according to some authors, at the Kidneys, connects with the urogenital area and emerges at the perineum at **Ren-1** (*huiyin*). From there the **external** pathway ascends the anterior midline, passing the abdomen, the chest, and the throat. It terminates on the chin in the mentolabial groove at **Ren-24** (*chengjiang*), where it meets the Stomach primary channel and the extraordinary vessel *du mai*. From here the vessel continues internally, encircles the inner surface of the lips and connects with the *du mai* at → **Du-28** (*yinjiao*) beneath the upper frenulum. It divides into **two branches**, which both reach the infraorbital ridge at → **ST-1** (*chengqi*), where they cross the Stomach primary channel as well as the extraordinary vessel *yin qiao mai*, and terminate near the eyes.

- **An internal branch** runs from the perineum to the spine, penetrates the spinal canal at **Du-1** (*changqiang*) and ascends in the spine.

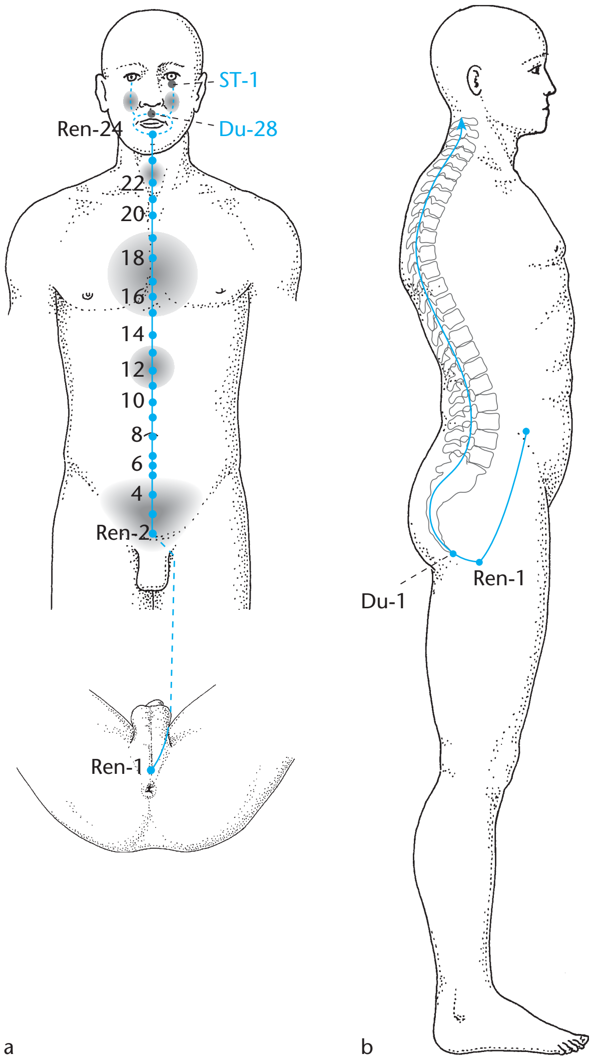

Meeting Points of other channels with the *ren mai*

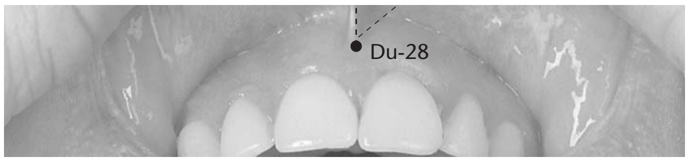

Du-28 (*yinjiao*): On the inside of the upper lip, at the junction of the frenulum and the upper gum

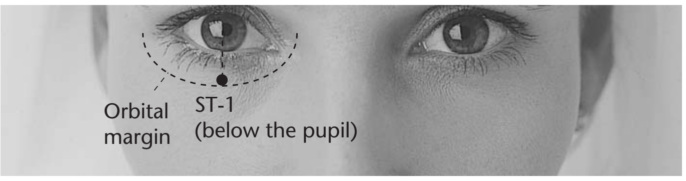

ST-1 (*chengqi*): With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

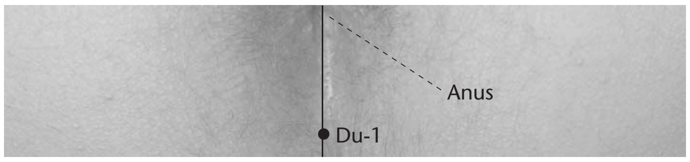

Du-1 (*changqiang*): On the midline, midway between the tip of the coccyx and the anus

Connections with other channels and organ systems

- *Du mai*, *chong mai*, Kidney and Stomach primary channels
- Uterus, Kidney

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of the Yin Qi**, regulates and circulates the Yin Qi (including the Blood, Essence (*jing*) and body fluids) in order to nourish and moisten the body
- Connects all Yin channels, especially those of the lower abdomen
- Controls the anterior thoracic region and the abdomen
- Close relationship to the functions of the abdominal organs, particularly the Kidney, Bladder, Uterus and Lung
- Regulates reproduction
- Stores the Essence Qi (*jing qi*) overflowing from the Yin channels
- Absorbs excess from the Yin channels, especially stagnant Qi and Blood

Notes about the *ren mai*

Both the *ren mai* and *du mai* have a unique position among the eight extraordinary vessels. They are the only two extraordinary vessels with their own specific points – all other extraordinary vessels can only be accessed through points on the primary channels. For this reason the *ren mai* and the *du mai* together with the primary channels are often categorised as 14 channels. However, from the point of view of channel energetics this is not correct, since the extraordinarily vessels occupy a special position, both evolutionary as well as functionally (→ 1.7).

The pathway of the *ren mai*, like that of the *du mai*, is limited to the torso and the head. While it does not course along the extremities and does not have any points on the limbs (neither general points nor specific points such as the five *shu*-transporting points, *yuan*-source point, etc.) it does have an opening point (or Master point) (→ 8.1.8). The *ren mai* is also referred to as the Sea of Yin and governs the Yin in the whole body. Along its pathway are located the anterior aspects of the energy centres (or chakras according to Yogi tradition). Daoist tradition describes three energy centres, the three cinnabar fields (*dantian*) on the anterior aspect of the body, which play an important role in *Qigong* practice as centres where the Qi pools and originates. While all three centres are located on the pathway of the *ren mai*, they do not correspond to specific points but rather to certain areas.

The *ren mai* runs along the Yin side of the body. It contains not only several Front-*mu* points acting directly on the *zangfu*-Organs, but also multiple meeting points with other channels. This clearly denotes the outstanding importance of the *ren mai*.

Specific points according to their function

- ***Yuan*-Source point (→ 8.1.1): –**
- ***Luo*-Connecting point (→ 8.1.2.): Ren-15** (*jiuwei*) 🟦■
- ***Xi*-Cleft point (→ 8.1.3): –**
- **Associated Back-*shu* point (→ 8.1.4): –**
- **Associated Front-*mu* point (→ 8.1.5): –**
- **Five *shu*-transporting points (→ 8.1.6): –**
- **Hui-meeting point (→ 8.1.7)**
  - – of the Qi: **Ren-17** (*danzhong*) 🟦🟦
  - – of the *fu*-Organs: **Ren-12** (*zhongwan*) 🟦■
- **Pertaining opening point (→ 8.1.8): LU-7** (*lieque*)
- **Pertaining coupled point (→ 8.1.8): KID-6** (*zhaohai*)
- **Lower *he*-Sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**
  - – with the *chong mai* and the *du mai*: **Ren-1** (*huiyin*)
  - – with the LIV channel: **Ren-2** (*qugu*) 🟦■
  - – with the SP, LIV, KID channels: **Ren-3** (*zhongji*) 🟦🟦, **Ren-4** (*guanyuan*) 🟦🟦
  - – with the *chong mai* and the KID channel*: **Ren-7** (*yinjiao*)
  - – with the SP and ST* channels: **Ren-10** (*xiawan*)
  - – with the S.I., T.B.* and ST channels: **Ren-12** (*zhongwan*) 🟦🟦
  - – with the ST and S.I. channels: **Ren-13** (*shangwan*)
  - – with the SP*, KID, S.I.* and T.B. channels: **Ren-17** (*danzhong*) 🟦🟦
  - – with the *yin wei mai*: **Ren-22** (*tiantu*) 🟦🟦, **Ren-23** (*lianquan*)

---

* Mentioned by only some authors

–468 with the *du mai*, the L.I. and ST channels: **Ren-24** (*chengjiang*) 🟦⬛
  - of other channels with the *ren mai*: **Du-28, ST-1, Du-1**
- **Gao Wu command point (→ 8.1.11):** –
- **Window of Heaven point (→ 8.1.12): Ren-22** (*tiantu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13) Qi: Ren-17** (*danzhong*) 🟦🟦
- **Ma Dan Yang Heavenly Star point (→ 8.1.14):** –
- **Sun Si Miao Ghost points (→ 8.1.15): Ren-1** (*huiyin*), **Ren-24** (*chengjiang*) 🟦⬛
- **Other functional points:**
  - Front-*mu* point of the Bladder: **Ren-3** (*zhongji*) 🟦🟦
  - Front-*mu* point of the Small Intestine: **Ren-4** (*guanyuan*) 🟦🟦
  - Front-*mu* point of the Triple Burner: **Ren-5** (*shimen*)
  - important, general tonification point: **Ren-6** (*qihai*) 🟦🟦
  - Front-*mu* point of the Stomach: **Ren-12** (*zhongwan*) 🟦🟦
  - Front-*mu* point of the Heart: **Ren-14** (*juque*) 🟦🟦
  - important calming point: **Ren-15** (*jiuwei*) 🟦⬛
  - Front-*mu* point of the Pericardium, point of the Sea of Qi: **Ren-17** (*danzhong*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** throat – **Ren-23** (*lianquan*); Lungs – **Ren-22** (*tiantu*) 🟦🟦, **Ren-17** (*danzhong*) 🟦🟦; Heart – **Ren-17** (*danzhong*) 🟦🟦; Spleen/Stomach – **Ren-12** (*zhongwan*) 🟦🟦; Small Intestine – **Ren-4** (*guanyuan*) 🟦🟦, **Ren-9** (*shuifen*); Bladder – **Ren-3** (*zhongji*) 🟦🟦; urogenital region – **Ren-3** (*zhongji*), **Ren-4** (*guanyuan*) 🟦🟦; epigastrium – **Ren-12** (*zhongwan*) 🟦🟦, **Ren-13** (*shangwan*); abdomen – **Ren-3** (*zhongji*) 🟦🟦, **Ren-4** (*guanyuan*) 🟦🟦, **Ren-6** (*qihai*) 🟦🟦
- **Adjacent points (→ 8.2.1):** throat – **Ren-22** (*tiantu*) 🟦🟦; tongue – **Ren-23** (*lianquan*); Heart – **Ren-14** (*juque*) 🟦🟦; Gall Bladder – **Ren-11** (*jianli*); Kidney – **Ren-4** (*guanyuan*) 🟦🟦; epigastrium – **Ren-8** (*shenque*)
- **Distal points (→ 8.2.1):** –

Specific points according to the vessel pathway (in numerical order)

- **Ren-1** (*huiyin*): *jiaohui*-meeting point with the *chong mai* and *du mai* (→ 8.1.10); Sun Si Miao Ghost point (→ 8.1.15)
- **Ren-2** (*qugu*) 🟦⬛: *jiaohui*-meeting point with the LIV channel (→ 8.1.10)
- **Ren-3** (*zhongji*) 🟦🟦: *jiaohui*-meeting point with the SP, LIV, KID channels (→ 8.1.10); Front-*mu* point of the Bladder; local point for the Bladder, urogenital region and lower abdomen (→ 8.2.1)
- **Ren-4** (*guanyuan*) 🟦🟦: Front-*mu* point of the Small Intestine (→ 8.1.5); *jiaohui*-meeting point with the SP, LIV, KID channels (→ 8.1.10); local point for the Small Intestine, urogenital region and lower abdomen (→ 8.2.1); adjacent point for the Kidneys
- **Ren-5** (*shimen*): Front-*mu* point of the Triple Burner
- **Ren-6** (*qihai*) 🟦🟦: important general tonification point; local point for the lower abdomen (→ 8.2.1); adjacent point for the Lung (→ 8.2.1)
- **Ren-7** (*yinjiao*): *jiaohui*-meeting point with the *chong mai* and the KID channel (→ 8.1.10)
- **Ren-8** (*shenque*): adjacent point for the epigastrium (→ 8.2.1)
- **Ren-9** (*shuifen*): local point for the Small Intestine (→ 8.2.1)
- **Ren-10** (*xiawan*): *jiaohui*-meeting point with the SP and ST* channels (→ 8.1.10)
- **Ren-11** (*jianli*): adjacent point for the Gall Bladder (→ 8.2.1)
- **Ren-12** (*zhongwan*) 🟦🟦: *hui*-meeting point (→ 8.1.7) of the *fu*-Organs; *jiaohui*-meeting point with the S.I., T.B.* and ST channels (→ 8.1.10); Front-*mu* point of the Stomach (→ 8.1.5); local point for the SP/ST and the epigastrium (→ 8.2.1)
- **Ren-13** (*shangwan*): *jiaohui*-meeting point with the ST and S.I. channels (→ 8.1.10); local point for the epigastrium (→ 8.2.1)
- **Ren-14** (*juque*) 🟦🟦: Front-*mu* point of the Heart (→ 8.1.5); adjacent point for the Heart
- **Ren-15** (*jiuwei*) 🟦⬛: *luo*-connecting point (→ 8.1.2); important point for calming
- **Ren-17** (*danzhong*) 🟦🟦: *hui*-meeting point (→ 8.1.7) of the Qi; *jiaohui*-meeting point with the SP, KID, S.I. and T.B. channels (→ 8.1.10); Front-*mu* point of the Pericardium, point of the Sea of Qi; local point for the Lung and Heart (→ 8.2.1)
- **Ren-22** (*tiantu*) 🟦🟦: *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10); Window of Heaven point (→ 8.1.12); local point for the Lungs (→ 8.2.1); adjacent point for the throat (→ 8.2.1)
- **Ren-23** (*lianquan*): *jiaohui*-meeting point with the *yin wei mai* (→ 8.1.10); local point for the throat (→ 8.2.1); adjacent point for the tongue (→ 8.2.1)
- **Ren-24** (*chengjiang*) 🟦⬛: *jiaohui*-meeting point with the *du mai*, the L.I. and ST channels (→ 8.1.10); Sun Si Miao Ghost point

---

* Mentioned by only some authors

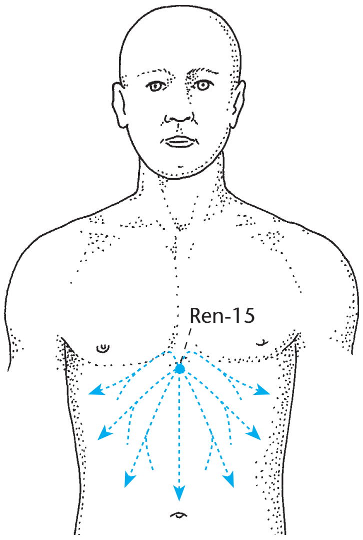

Luo-Connecting Vessel of the *Ren Mai* (→ 1.6)

Pathway

The *luo*-connecting vessel of the *ren mai* originates at **Ren-15** (*jiuwei*), below the xiphoid process (→ 3) and disperses over the skin of the abdomen.

Clinical importance

Pathology

Excess (*shi*): pain of the abdominal skin
Deficiency (*xu*): itching of the abdominal skin

Meeting of the Yin HUIYIN Ren-1

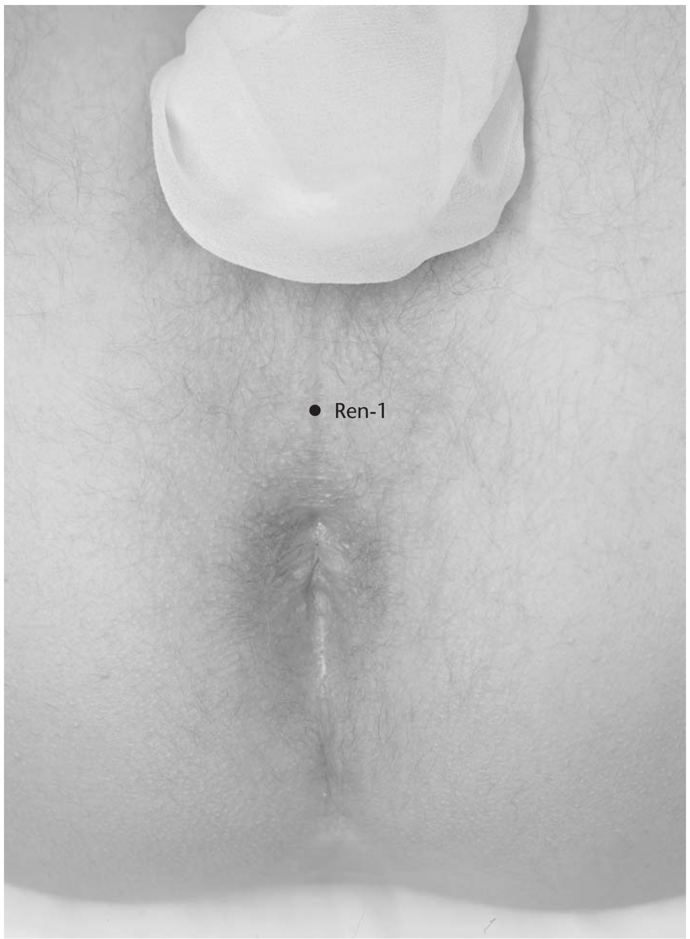

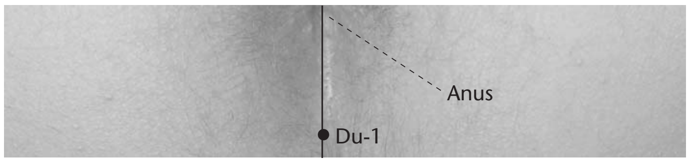

Location

In the centre of the perineum.

How to find

In women: midway between the anus and the posterior labial commissure. In men: midway between the anus and the scrotum.

Needling

Vertically 0.5–1 cun. Caution: Contraindicated during pregnancy. Some classic texts prohibit needling this point. Particular attention should be paid to the prevention of infection.

Actions/Indications

- Clears Heat, strengthens the Kidneys and drains Dampness
- Regulates menstruation

Special features

Meeting point with the *du mai* and the *chong mai*. As a Sun Si Miao Ghost point, **Ren-1** is traditionally indicated for treating epilepsy and psychoses.

Ren-2 *Curved Bone* QUGU

Location

On the anterior midline, on the upper border of the pubic symphysis.

How to find

In the pubic area, locate the upper border of the pubic symphysis. **Ren-2** is located directly superior to its upper border, on the anterior midline.

Located on the same level are → **KID-11** (0.5 cun lateral to the midline), → **ST-30** (2 cun lateral to the midline) and → **SP-12** (3.5 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum. Also caution with a full bladder and during pregnancy.

Actions/Indications

- Benefits urination, warms the Yang and strengthens the Kidneys
- Regulates the Lower Burner

Special features

Meeting point with the LIV channel

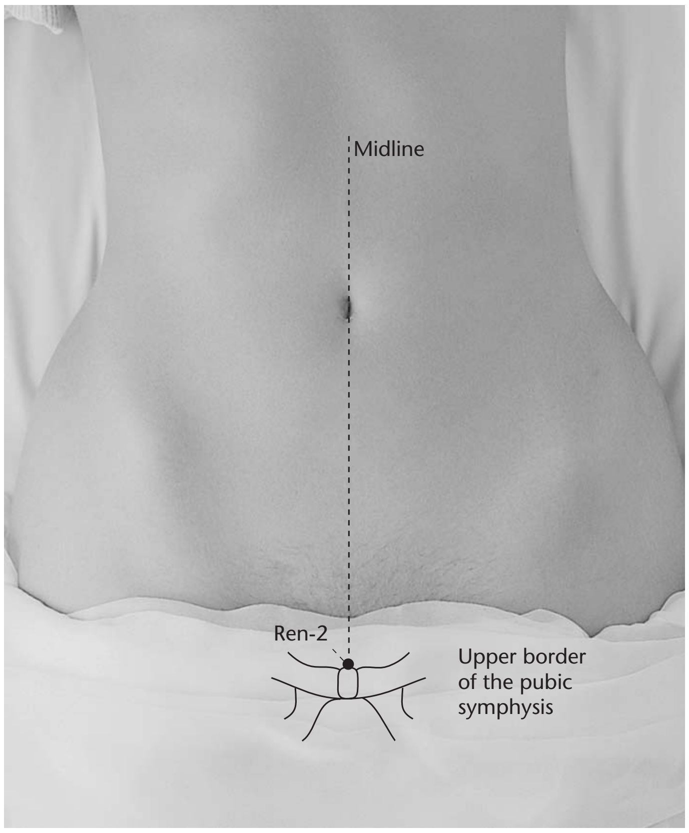
Midline
Ren-2
Upper border of the pubic symphysis

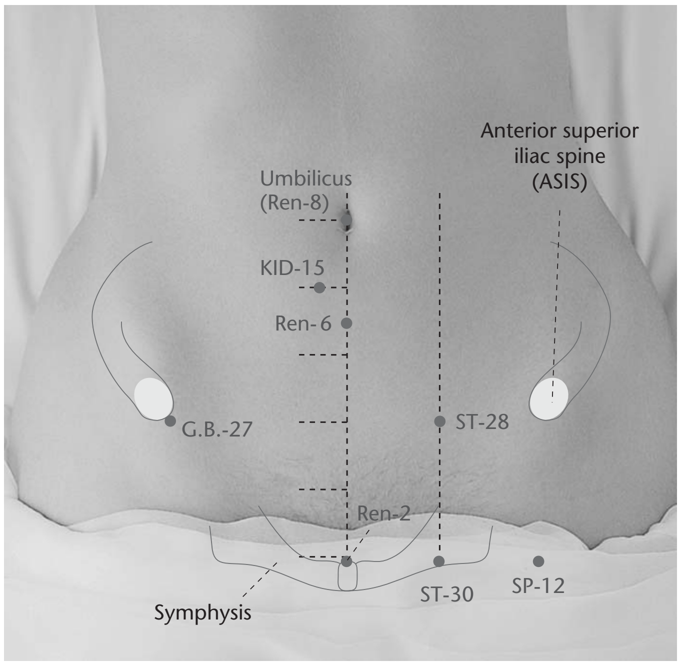
Anterior superior iliac spine (ASIS)
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
Symphysis
ST-30
SP-12

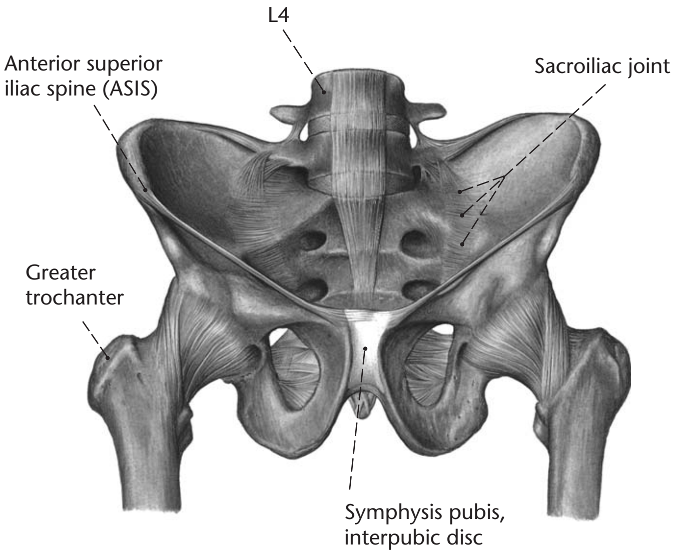
L4
Anterior superior iliac spine (ASIS)
Sacroiliac joint
Greater trochanter
Symphysis pubis, interpubic disc

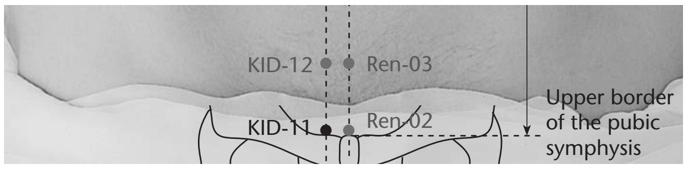
KID-12
Ren-03
KID-11
Ren-02
Upper border of the pubic symphysis

Middle Pole ZHONGJI Ren-3

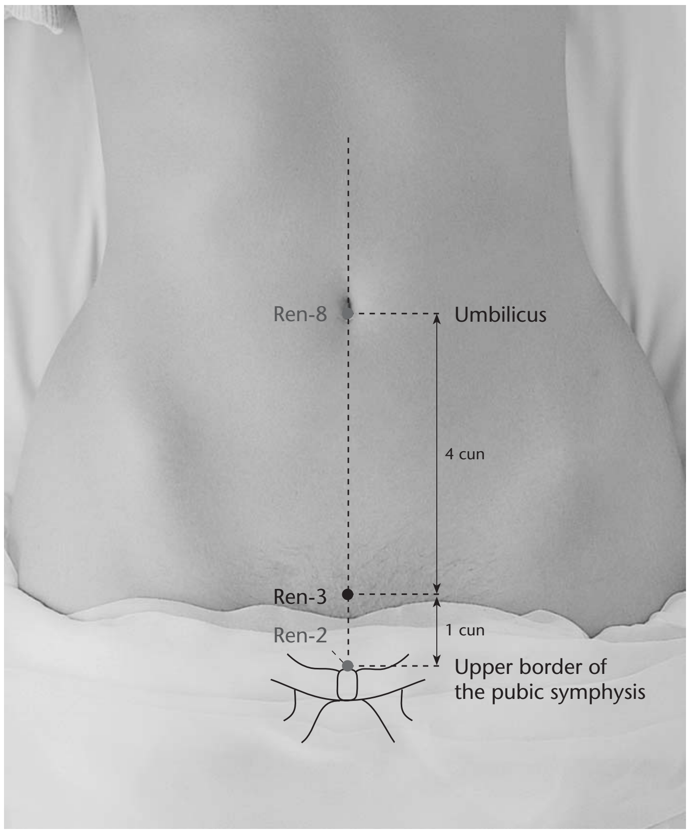
Ren-8
Umbilicus
4 cun
Ren-3
1 cun
Ren-2
Upper border of
the pubic symphysis

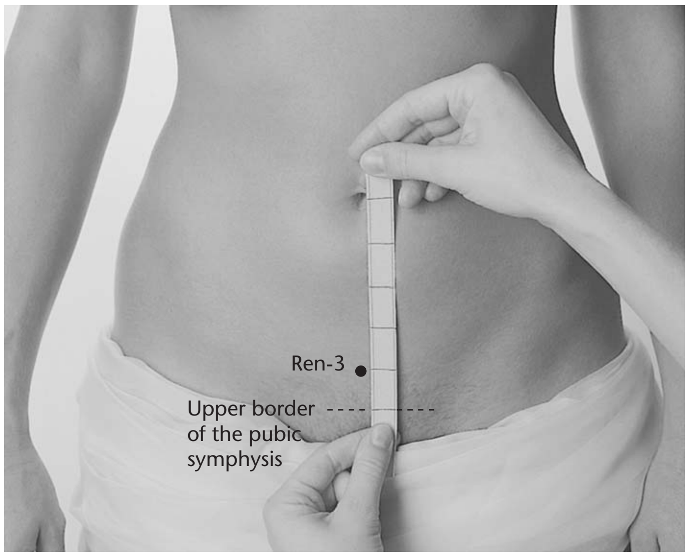
Ren-3
Upper border
of the pubic
symphysis

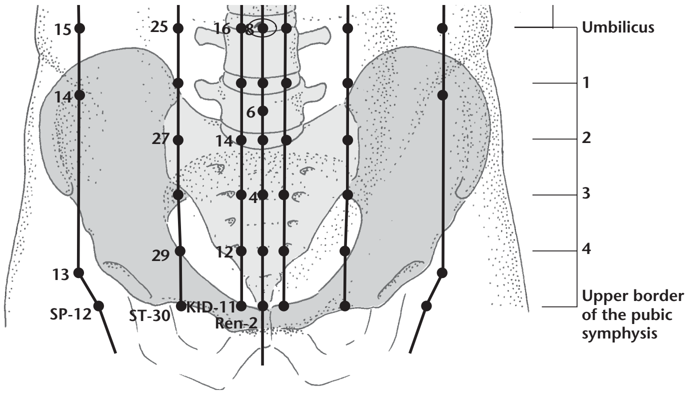
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

Location

On the anterior midline, 1 cun superior to the upper border of the pubic symphysis or 4 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun. As these can vary considerably from the patient's finger cun, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). Locate **Ren-3** on the midline, 1 cun superior to the midpoint of the upper border of the pubic symphysis.

Located on the same level are → **KID-12** (0.5 cun lateral to the midline) and → **ST-29** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, bladder. Contraindicated during pregnancy. Ask the patient to void their bladder before needling.

Actions/Indications

- Benefits the Bladder, drains Dampness and Damp-Heat, dispels stagnation, benefits the Lower Burner
- Regulates menstruation
- Strengthens the Kidneys (→ **Ren-4** is more commonly used for this purpose)

Special features

Front-*mu* point of the Bladder, meeting point with the LIV, KID and SP channels. Important point for disorders of the urogenital tract, with its main action on the Bladder (especially for acute cases).

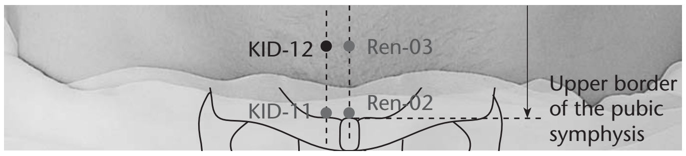
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

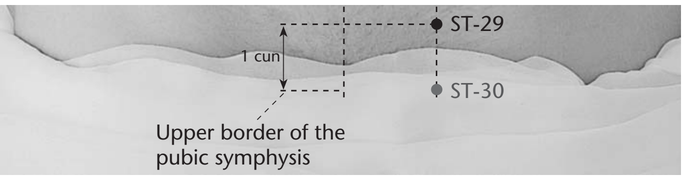
ST-29
1 cun
ST-30
Upper border of the
pubic symphysis

Ren-4 *Gate of the Source* GUANYUAN

Location

On the anterior midline, 2 cun superior to the upper border of the pubic symphysis or 3 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-4** is located on the midline, either 2 cun superior to the midpoint of the upper border of the pubic symphysis or 3 cun inferior to the umbilicus.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and three extra points → **Ex-CA** (*yijing*: 1 cun lateral to the midline; *qimen*: 3 cun lateral to the midline; *tituo*: 4 cun lateral to the midline). Also located on approximately this level is → **G.B.-27** (medial to the ASIS).

Needling

Vertically 0.5–1.2 cun. Caution: Peritoneum. Needling contraindicated during pregnancy.

Actions/Indications

- Tonifies the Original Qi (*yuan qi*), benefits the Essence (*jing*), strengthens and nourishes the Kidneys as well as the Blood and Yin
- Eliminates Cold and Dampness from the Lower Burner (especially with moxibustion)

Special features

Front-*mu* point of the Small Intestine, meeting point with the SP, KID and LIV channels. Important point for tonification. Major point for urogenital disorders (especially gynaecological disorders).

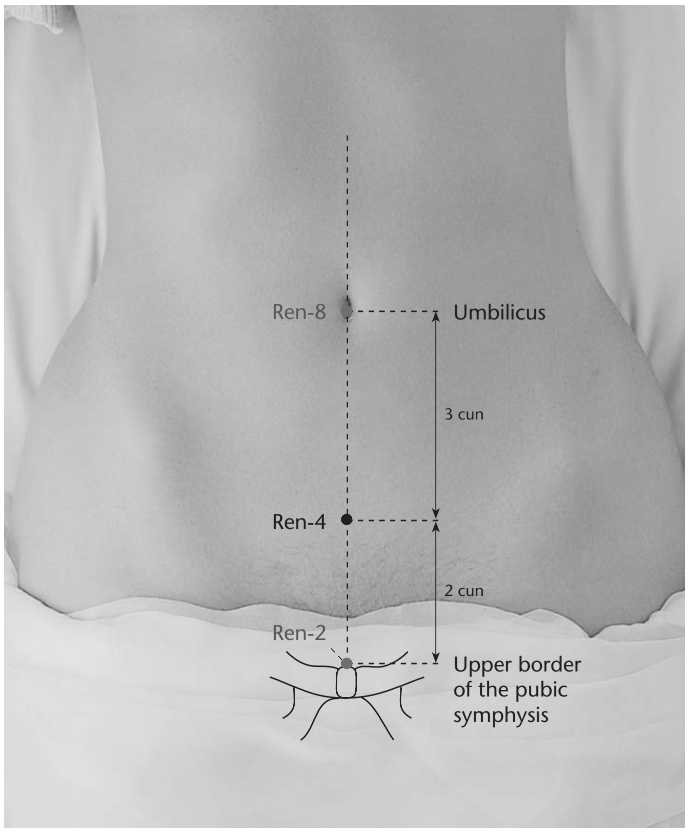
Ren-8
Umbilicus
3 cun
Ren-4
2 cun
Ren-2
Upper border of the pubic symphysis

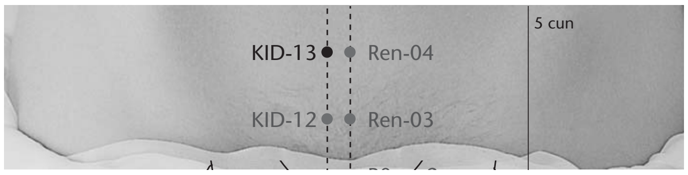
5 cun
KID-13
Ren-04
KID-12
Ren-03

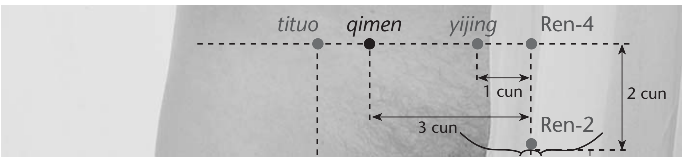
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

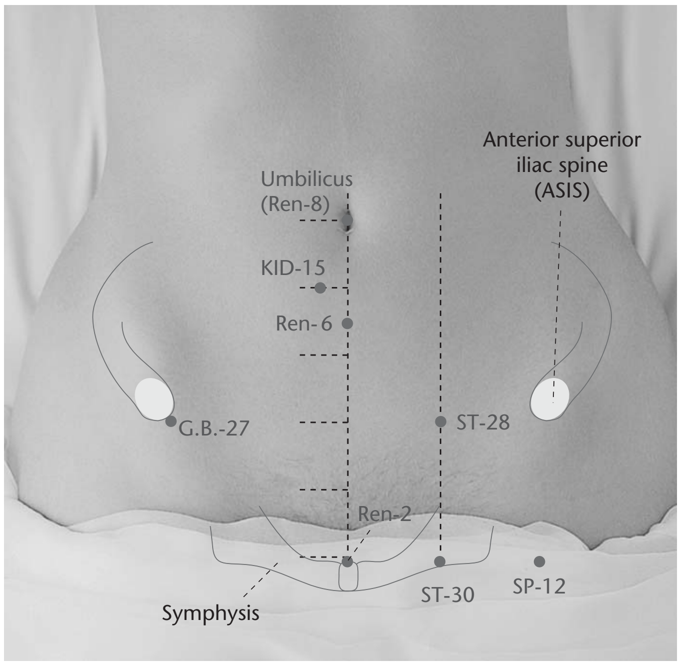
Anterior superior iliac spine (ASIS)
Umbilicus (Ren-8)
KID-15
Ren-6
G.B.-27
ST-28
Ren-2
ST-30
SP-12
Symphysis

Stone Gate SHIMEN **Ren-5**

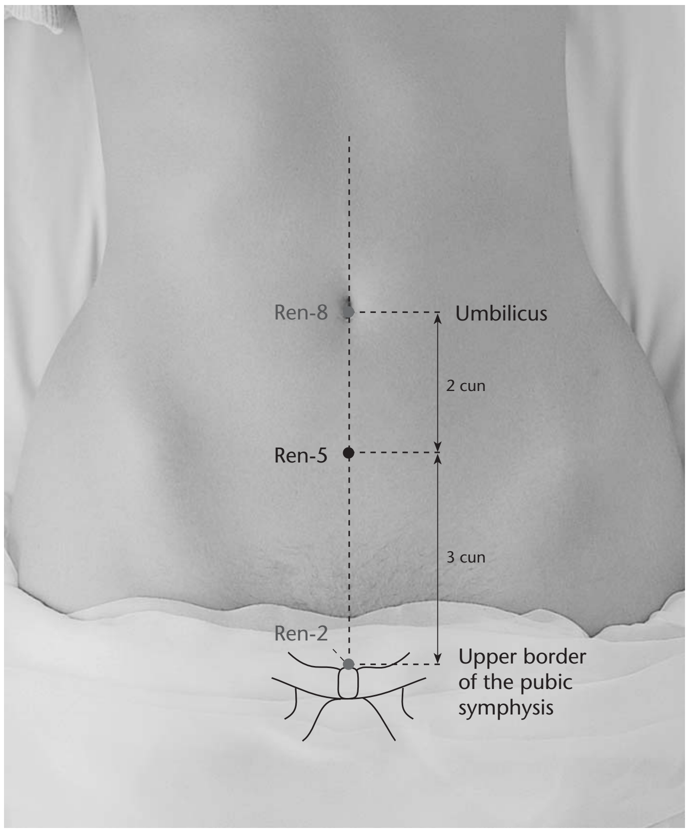
Ren-8
Umbilicus
2 cun
Ren-5
3 cun
Ren-2
Upper border
of the pubic
symphysis

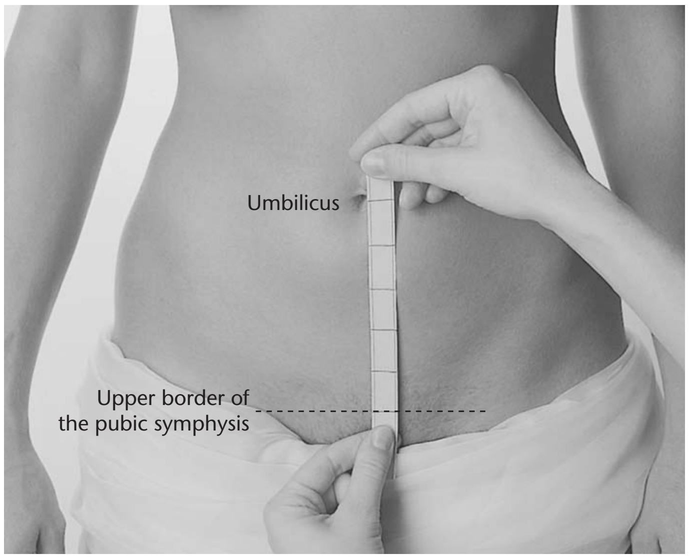
Umbilicus
Upper border of
the pubic symphysis

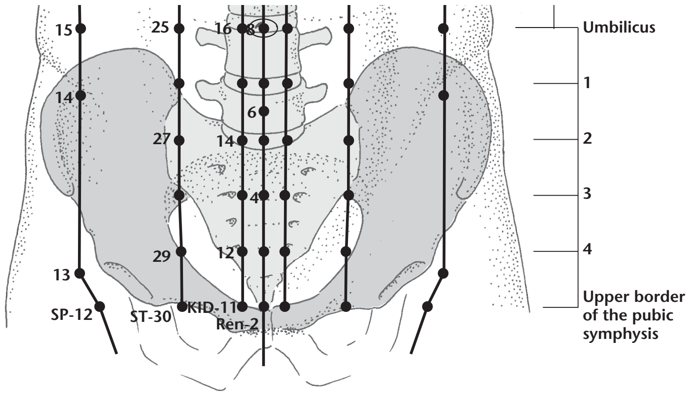
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

Location

On the anterior midline, 2 cun inferior to the umbilicus or 3 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape ➞ 2.3.1). **Ren-5** is located on the midline, 2 cun inferior to the umbilicus. Located on the same level are ➞ **KID-14** (0.5 cun lateral to the midline) and ➞ **ST-27** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Moves and benefits the water passages
- Regulates the Qi and Blood in the Lower Burner
- Regulates the Uterus

Special features

Front-*mu* point of the Triple Burner. Some classic texts warn that needling this point in women can lead to infertility (hence the name: infertile women were referred to as 'stone women'). There are no such warnings in modern texts.

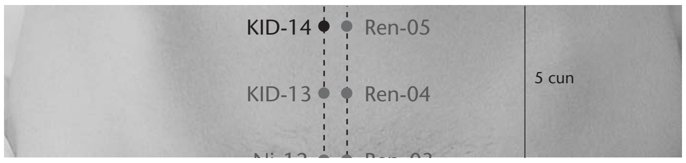
KID-14
Ren-05
KID-13
Ren-04
5 cun

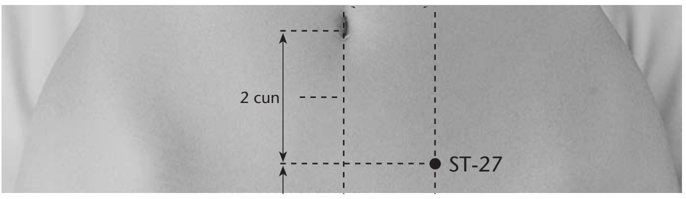

Ren-6 *Sea of Qi* QIHAI

Location

On the anterior midline, 1.5 cun inferior to the umbilicus or 3.5 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-6** is located on the midline, 1.5 cun inferior to the umbilicus or 3.5 cun superior to the upper border of the pubic symphysis.

Needling

Vertically 0.8–1.5 cun. Caution: Peritoneum, pregnancy. According to some authors, this point should never be needled with strong reducing techniques owing to its strong Qi tonifying action. Moxibustion and cupping are recommended for this point.

Actions/Indications

- Tonifies the Original Qi (*yuan qi*) as well as Qi in general, strengthens the Kidneys (especially Kidney Yang), regulates Qi, harmonises Blood and eliminates Dampness

Special features

Important tonification point for mental and physical exhaustion, strengthens especially the Qi and Yang with moxibustion. One of the most important acupuncture points.

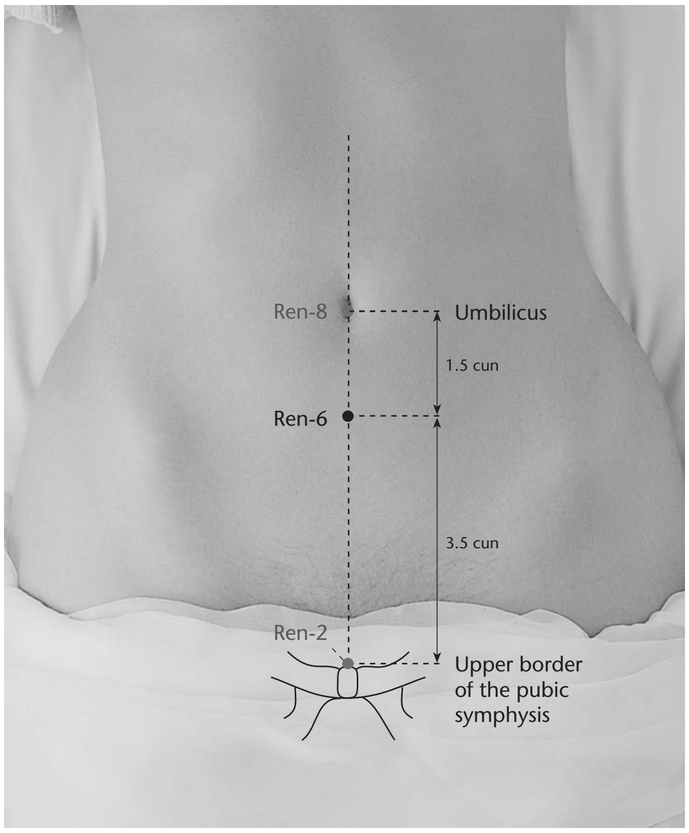
Ren-8
Umbilicus
1.5 cun
Ren-6
3.5 cun
Ren-2
Upper border of the pubic symphysis

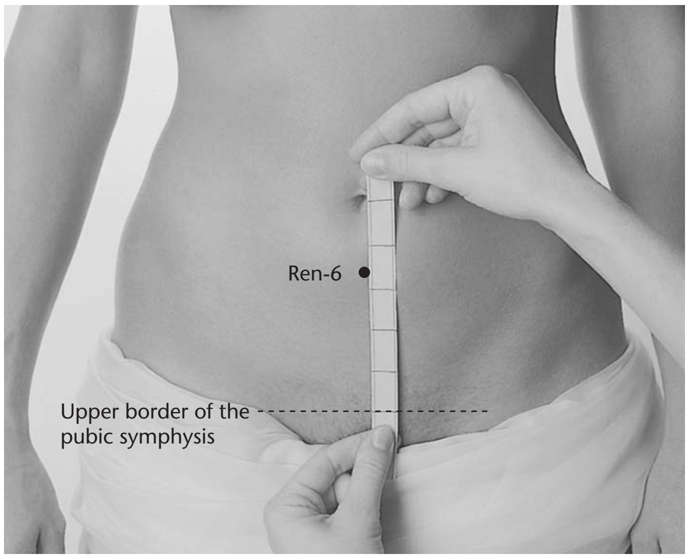
Ren-6
Upper border of the pubic symphysis

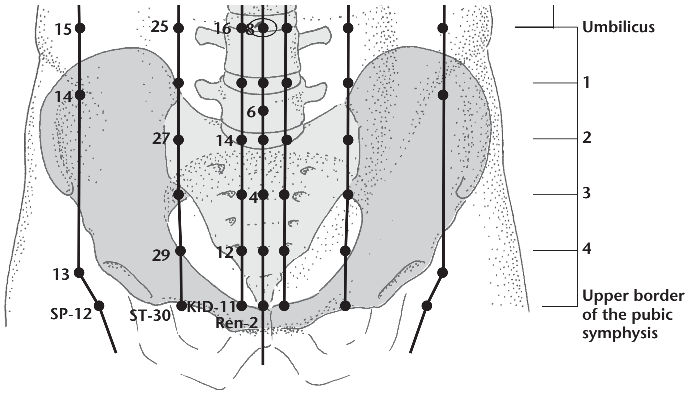
15
25
16
8
Umbilicus
1
14
6
27
14
2
4
3
29
12
4
13
SP-12
ST-30
KID-11
Ren-2
Upper border of the pubic symphysis

Yin Intersection YINJIAO Ren-7

Location

On the anterior midline, 1 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used here (helpful device: elastic tape → 2.3.1). **Ren-7** is located on the midline, 1 cun inferior to the umbilicus.

Located on the same level are → **KID-15** (0.5 cun lateral to the midline) and → **ST-26** (2 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates menstruation
- Regulates the Qi flow in the abdomen and the genital area

Special features

Meeting point with the *chong mai* and the KID channel

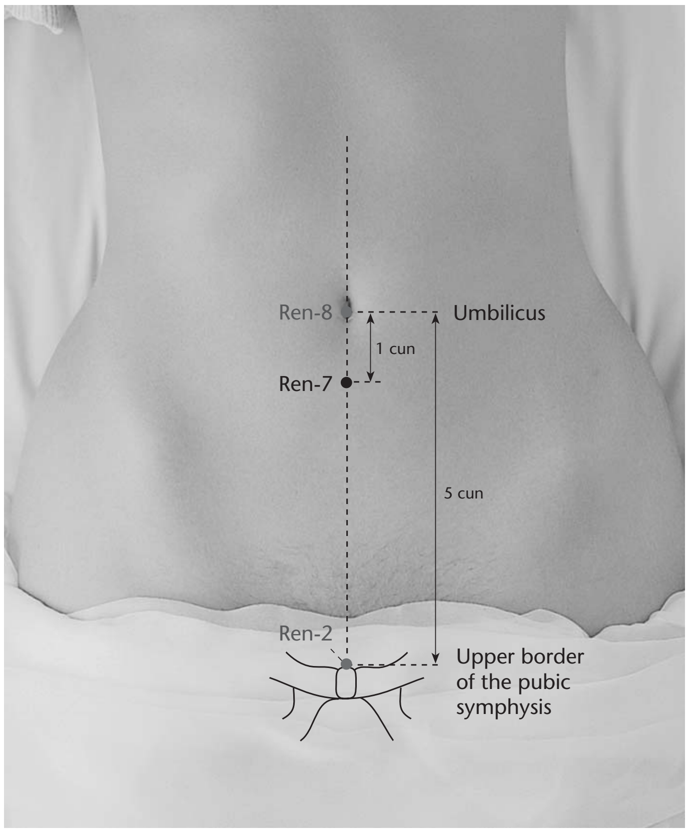
Ren-8
Umbilicus
1 cun
Ren-7
5 cun
Ren-2
Upper border
of the pubic
symphysis

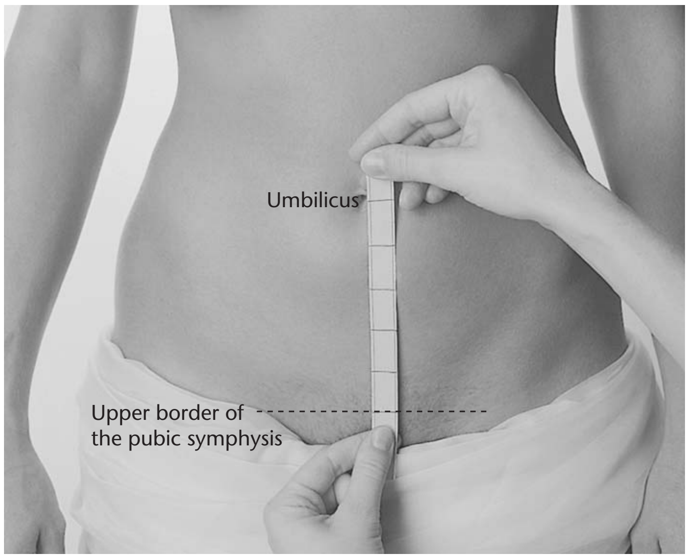
Umbilicus
Upper border of
the pubic symphysis

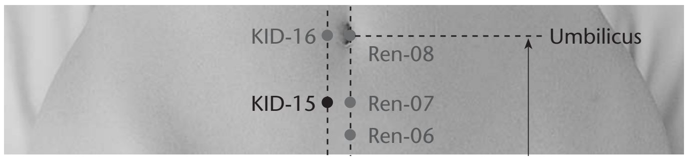
KID-16
Umbilicus
Ren-08
KID-15
Ren-07
Ren-06

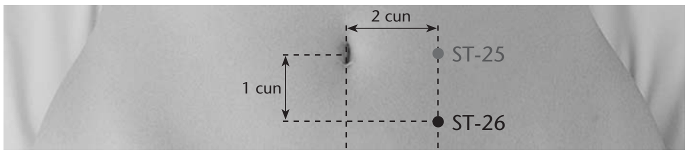
2 cun
ST-25
1 cun
ST-26

Ren-8 *Spirit Gateway* SHENQUE

Location

In the centre of the umbilicus.

How to find

Located on the same level are ➞ **KID-16** (0.5 cun lateral to the midline), ➞ **ST-25** (2 cun lateral to the midline), ➞ **SP-15** (4 cun lateral to the midline) as well as ➞ **G.B.-26** (on a vertical line through the free end of the 11th rib).

Needling

Needling contraindicated. Only indirect moxibustion (on salt or ginger, moxa pole, moxa box).

Actions/Indications

- Warms and stabilises the Yang and the Intestines

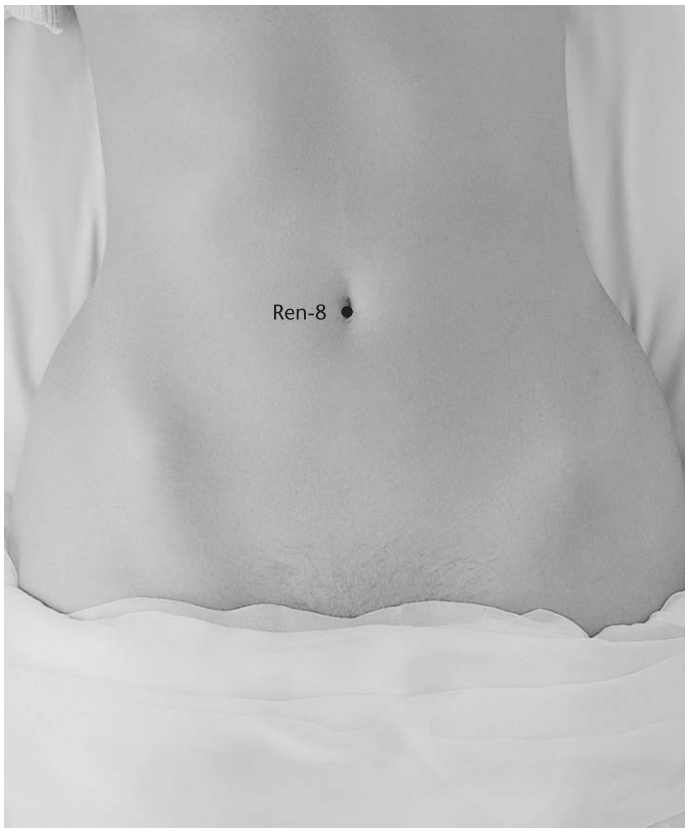

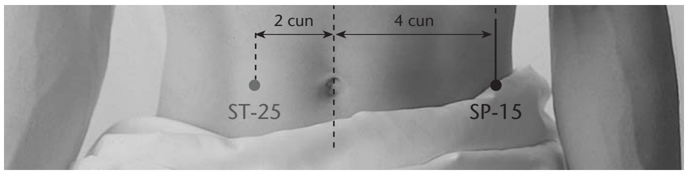
2 cun
4 cun
ST-25
SP-15

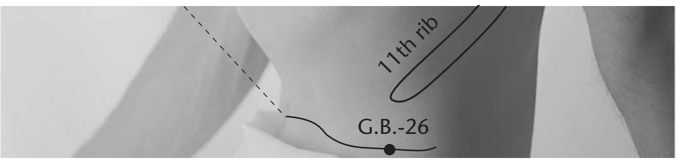
11th rib
G.B.-26

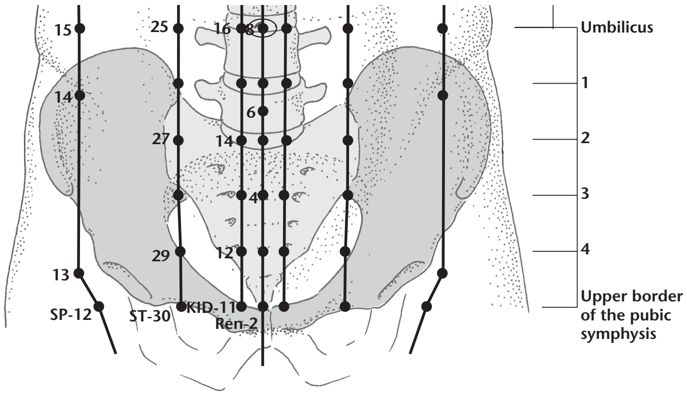
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

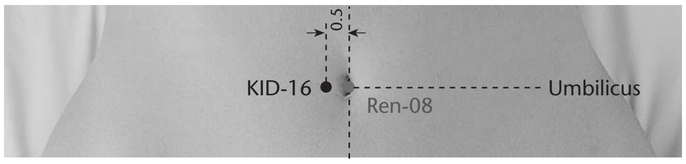
0.5
KID-16
Ren-08
Umbilicus

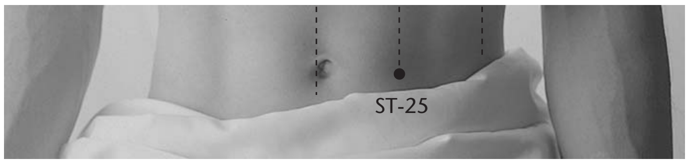

Water Separation SHUIFEN Ren-9

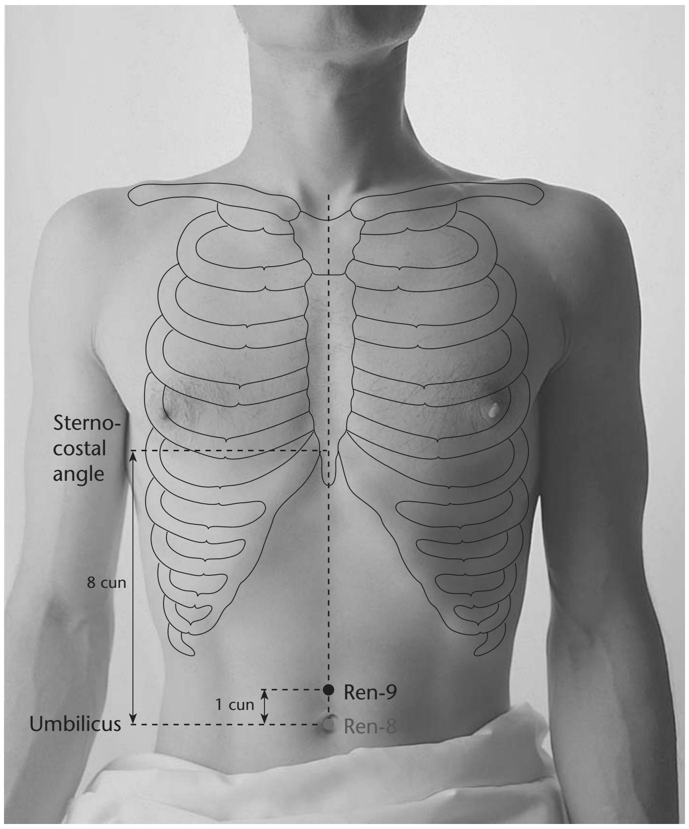
Sterno-
costal
angle
8 cun
1 cun
Ren-9
Umbilicus
Ren-8

Location

On the anterior midline, 1 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 1 cun in a superior direction and there locate **Ren-9** on the anterior midline.

Located on the same level is → **ST-24** (2 cun lateral to the midline). → **LIV-13** is also located on approximately this level (at the free end of the 11th rib).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy

Actions/Indications

- Eliminates water accumulations, drains Dampness and controls the water passages Regulates the Intestines and disperses accumulations
- Delayed closure of the fontanelles

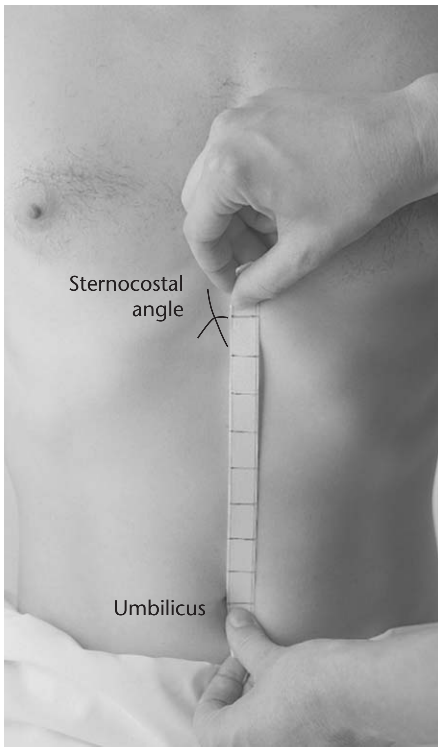
Sternocostal
angle
Umbilicus

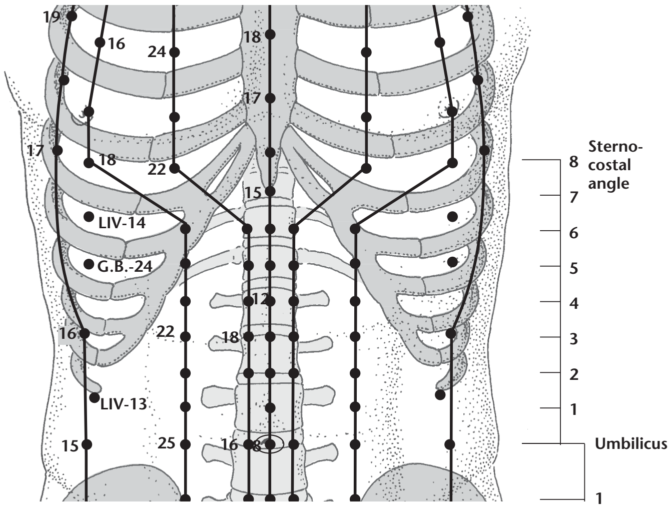
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

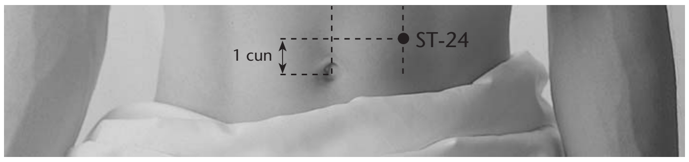

Ren-10 *Lower Cavity* XIAWAN

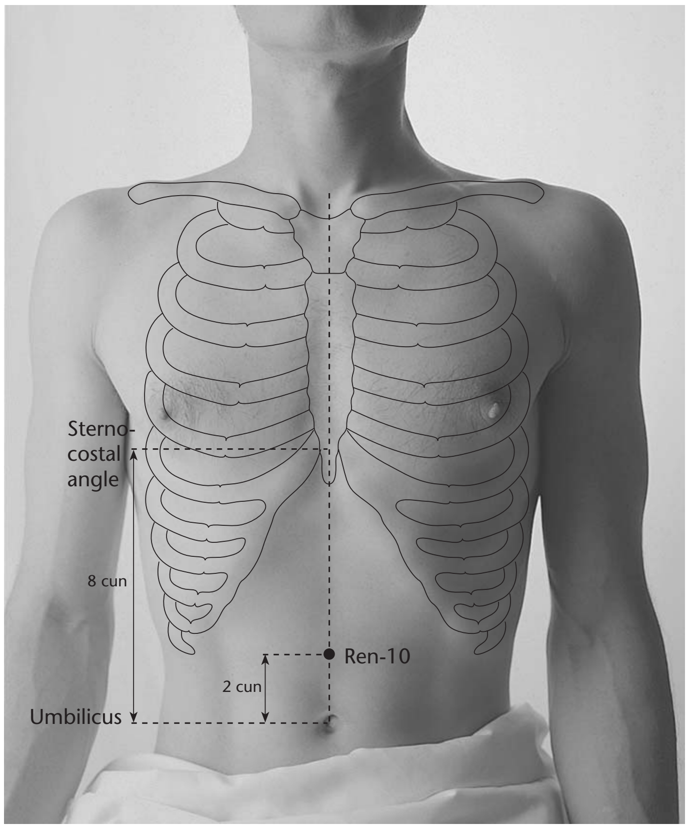
Sterno-
costal
angle
8 cun
Ren-10
2 cun
Umbilicus

Location

On the anterior midline, 2 cun superior to the centre of the umbilicus.

How to find

The distance between the sternocostal angle (➞ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore only proportional cun measurements should be applied here (helpful device: elastic tape ➞ 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction and there locate **Ren-10** on the anterior midline.

Located on the same level are ➞ **KID-17** (0.5 cun lateral to the midline), ➞ **ST-23** (2 cun lateral to the midline) and **Ex-CA** (*weishang*; 4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Tonifies and regulates the Spleen and Stomach Qi, dispels food stagnation

Special features

Meeting point with the SP channel and according to some authors with the ST channel. According to some classical texts, this point is contraindicated during pregnancy (for both needling and moxibustion).

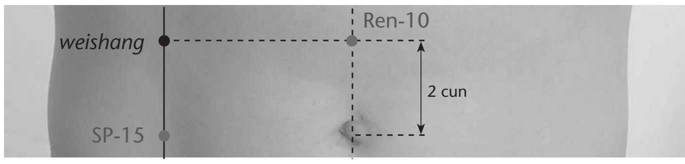
weishang
Ren-10
2 cun
SP-15

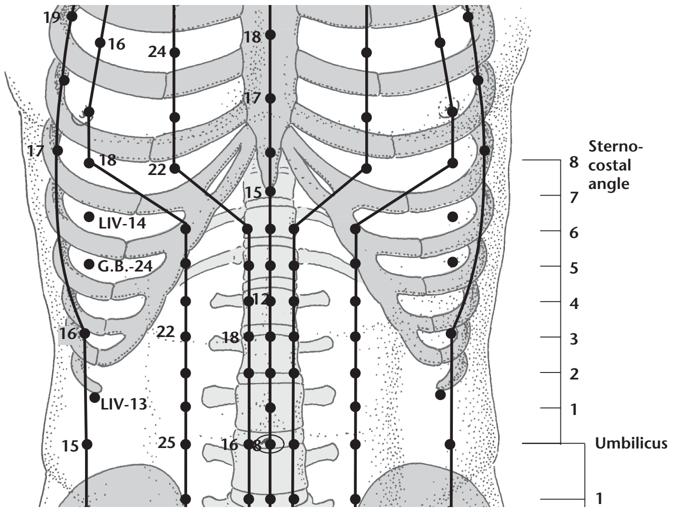
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
7
6
5
4
3
2
1
Umbilicus
1

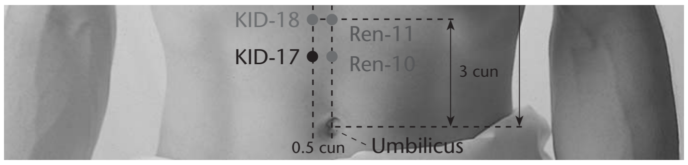
KID-18
KID-17
Ren-11
Ren-10
3 cun
0.5 cun
Umbilicus

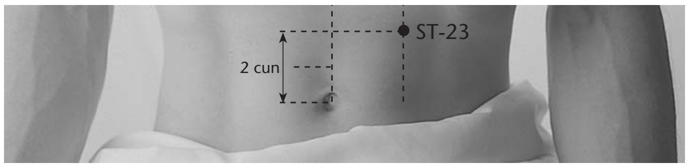

Interior Strengthening JIANLI **Ren-11**

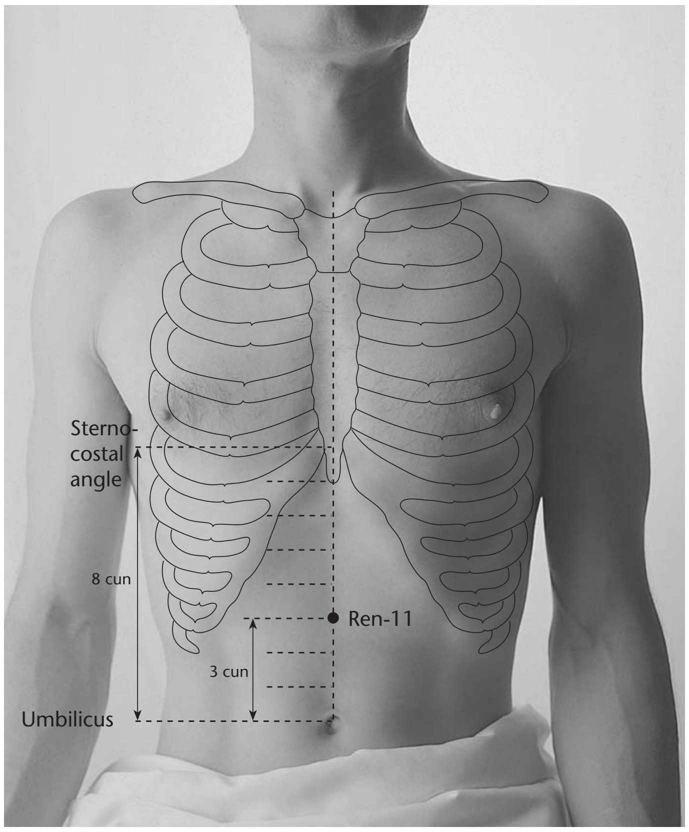
Sterno-
costal
angle
8 cun
Ren-11
3 cun
Umbilicus

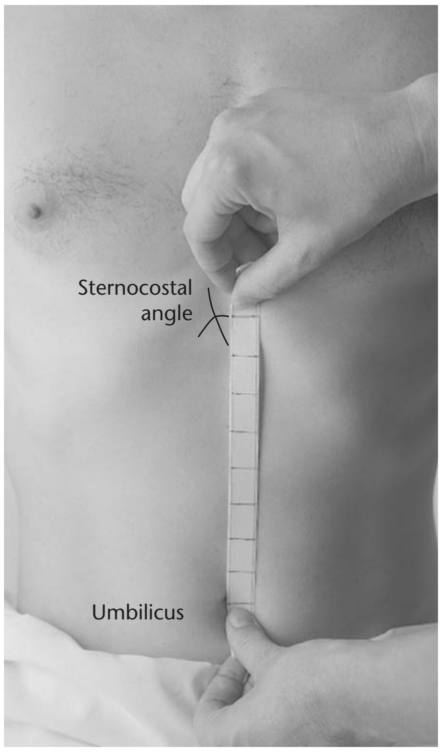
Sternocostal
angle
Umbilicus

Location

On the anterior midline, 3 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 3 cun in a superior direction and there locate **Ren-11** on the anterior midline. Located on the same level are → **KID-18** (0.5 cun lateral to the midline), → **ST-22** (2 cun lateral to the midline) and → **SP-16** (4 cun lateral to the midline).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Harmonises the Middle Burner, regulates the Qi flow

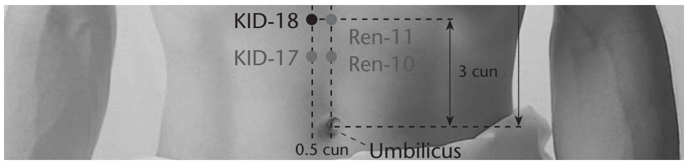
KID-18
KID-17
Ren-11
Ren-10
3 cun
0.5 cun
Umbilicus

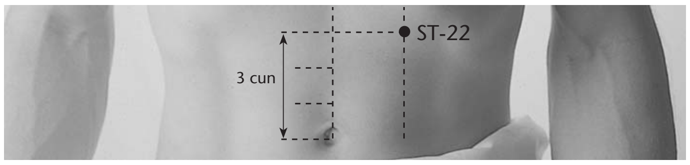

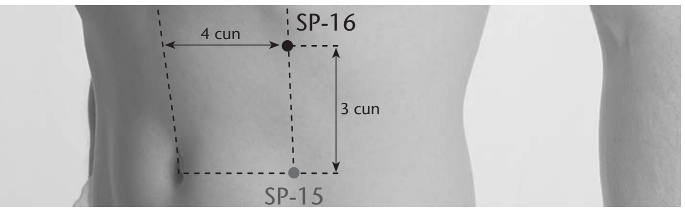
SP-16
4 cun
3 cun
SP-15

Ren-12 *Middle Cavity* ZHONGWAN

Location

On the anterior midline, 4 cun superior to the umbilicus or 4 cun inferior to the sternocostal angle.

How to find

The distance between the sternocostal angle (➞ 3.5) and the centre of the umbilicus can be divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape ➞ 2.3.1). From the centre of the umbilicus, measure either 4 cun in a superior direction or from the sternocostal angle 4 cun in an inferior direction. There, locate **Ren-12** on the anterior midline. **Or:** Spreading hands technique (➞ 2.3.3): Place the little fingers on the sternocostal angle and the centre of the umbilicus respectively and determine the midpoint (=**Ren-12**) of this distance by joining the index fingers. Located on the same level are ➞ **KID-19** (0.5 cun lateral to the midline) and ➞ **ST-21** (2 cun lateral to the midline).

Needling

0.8–1.5 cun vertically or obliquely towards surrounding points (➞ **ST-21**, **Ren-10**, **Ren-15**). Caution: Peritoneum, pregnancy. If indicated, moxa is recommended.

Actions/Indications

- Harmonises and tonifies the Middle Burner, descends counterflow Qi, regulates the Qi flow and alleviates pain, transforms Dampness

Special features

Front-*mu* point of the stomach, meeting point with the S.I., T.B.* and ST channels, *hui*-meeting point of the *fu*-Organs. Major point for (functional) stomach disorders.

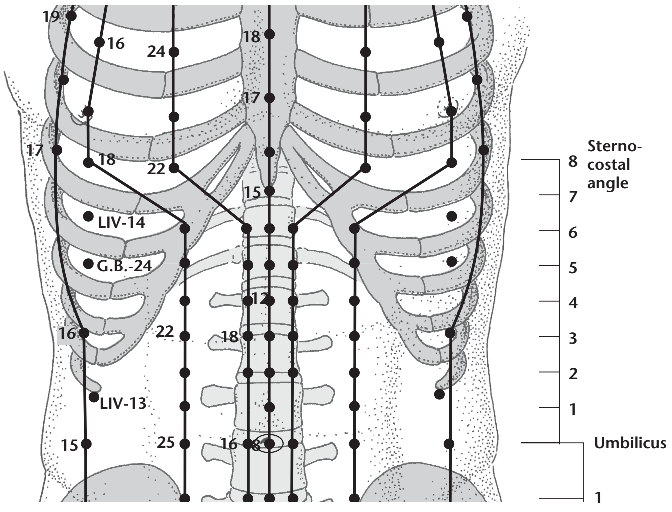
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

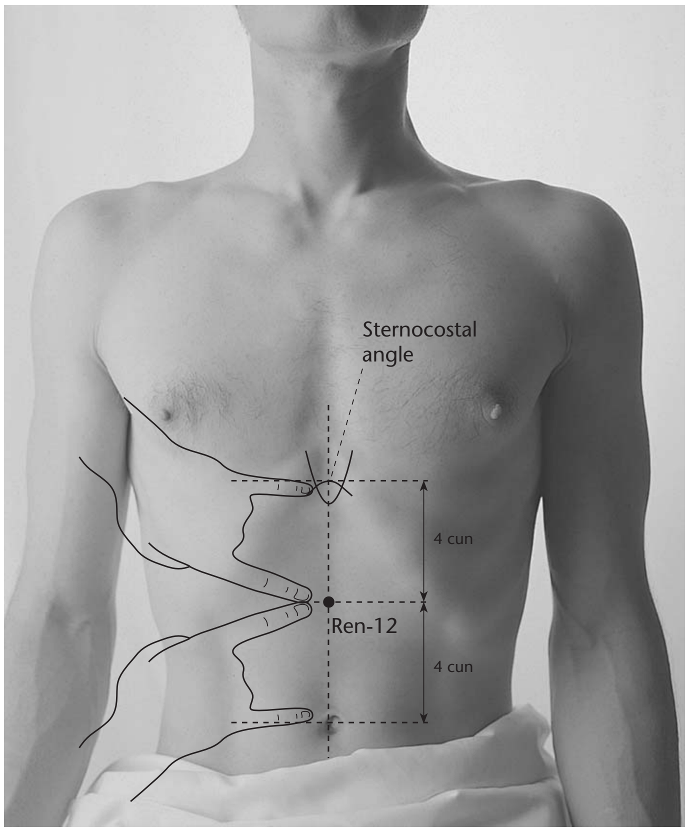
Sternocostal angle
4 cun
Ren-12
4 cun

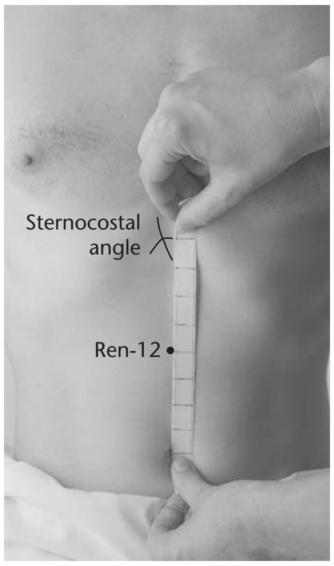
Sternocostal angle
Ren-12

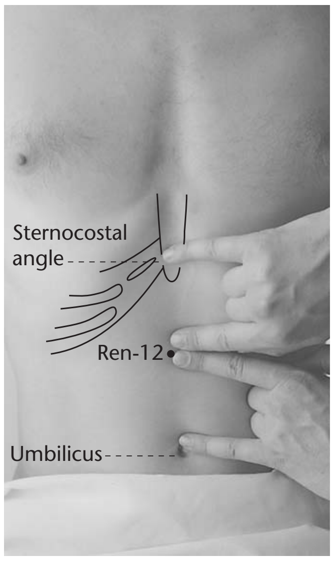
Sternocostal angle
Ren-12
Umbilicus

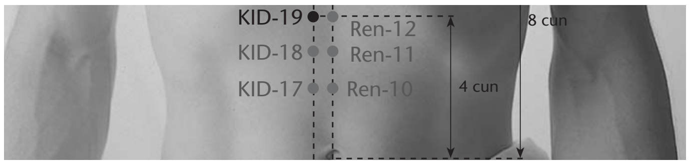
KID-19
Ren-12
KID-18
Ren-11
KID-17
Ren-10
8 cun
4 cun

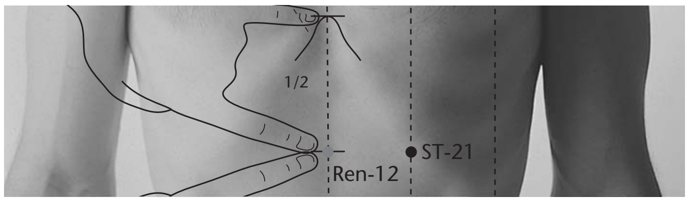
1/2
ST-21
Ren-12

---

* Mentioned by only some authors

Upper Cavity SHANGWAN **Ren-13**

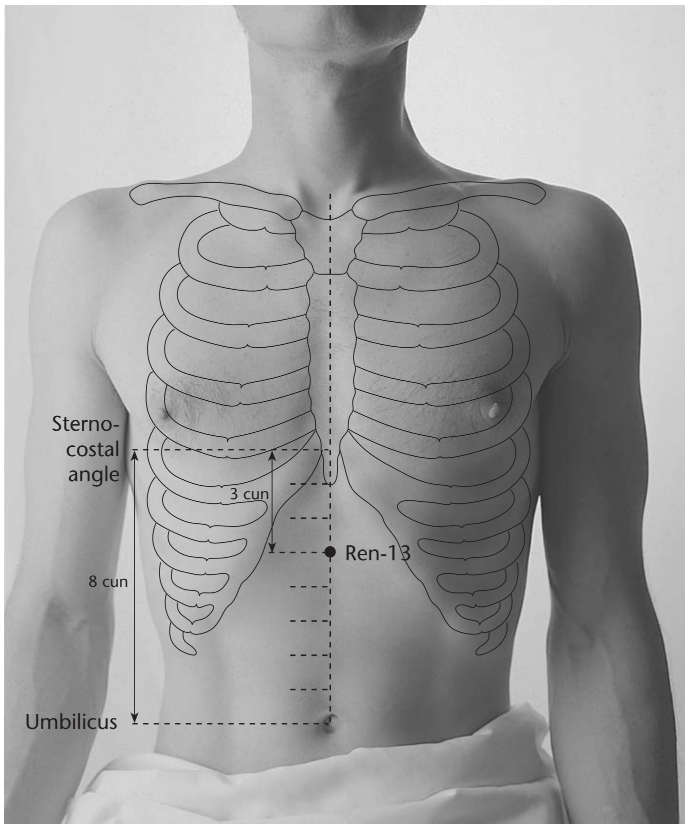
Sterno-
costal
angle
3 cun
Ren-13
8 cun
Umbilicus

Location

On the anterior midline, 3 cun inferior to the sternocostal angle.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the sternocostal angle, measure 3 cun in an inferior direction and there locate **Ren-13** on the anterior midline.

Located on the same level are → **KID-20** (0.5 cun lateral to the midline), → **ST-20** (2 cun lateral to the midline) and → **G.B.-24** (approximately on this level, in the 7th intercostal space, on the mamillary line).

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Regulates the Stomach and counterflow Qi
- Regulates the Heart

Special features

Meeting point with the ST and S.I. channel

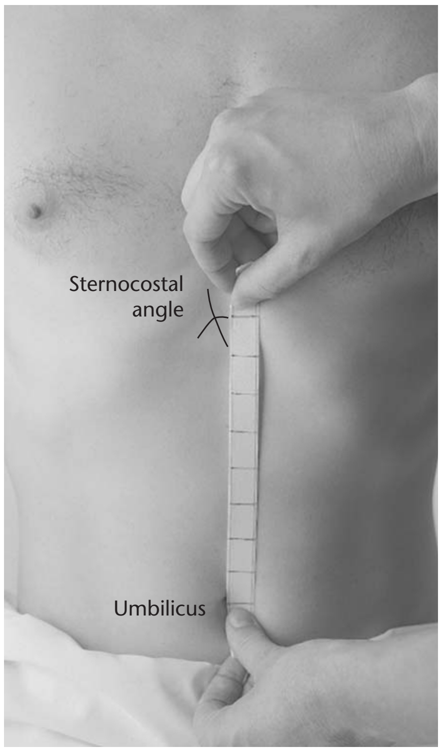
Sternocostal
angle
Umbilicus

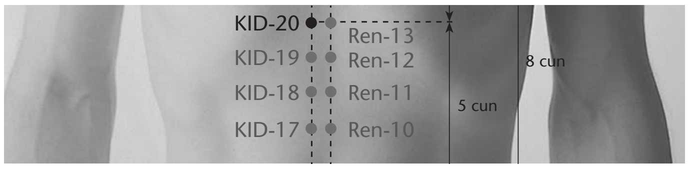
KID-20
Ren-13
KID-19
Ren-12
KID-18
Ren-11
KID-17
Ren-10
8 cun
5 cun

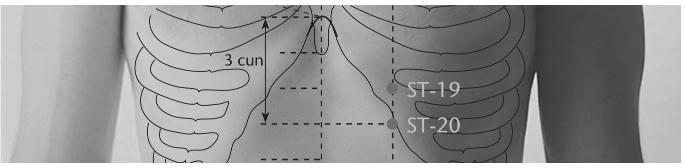
3 cun
ST-19
ST-20

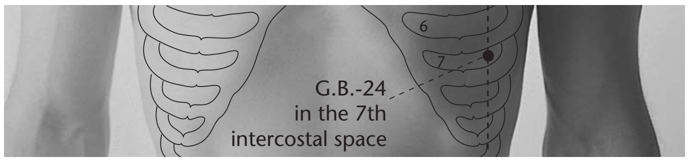
6
7
G.B.-24
in the 7th
intercostal space

Ren-14 *Great Gateway* JUQUE

Location

On the anterior midline, 2 cun inferior to the sternocostal angle or 6 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun (note: these are proportional cun → 2.2). **Ren-14** is located on the midline, 2 cun inferior to the sternocostal angle or 6 cun superior to the umbilicus. **Or:** Use spreading hands technique (→ Chapter 2): Place the little fingers on the sternocostal angle and the umbilicus respectively and determine the midpoint of this distance (→ **Ren-12**). Then locate **Ren-14** on the midpoint of the distance between → **Ren-12** and the sternocostal angle.

Located on the same level are → **KID-21** (0.5 cun lateral to the midline), → **ST-19** (2 cun lateral to the midline) and → **LIV-14** (approximately on this level, in the 6th intercostal space, on the mamillary line).

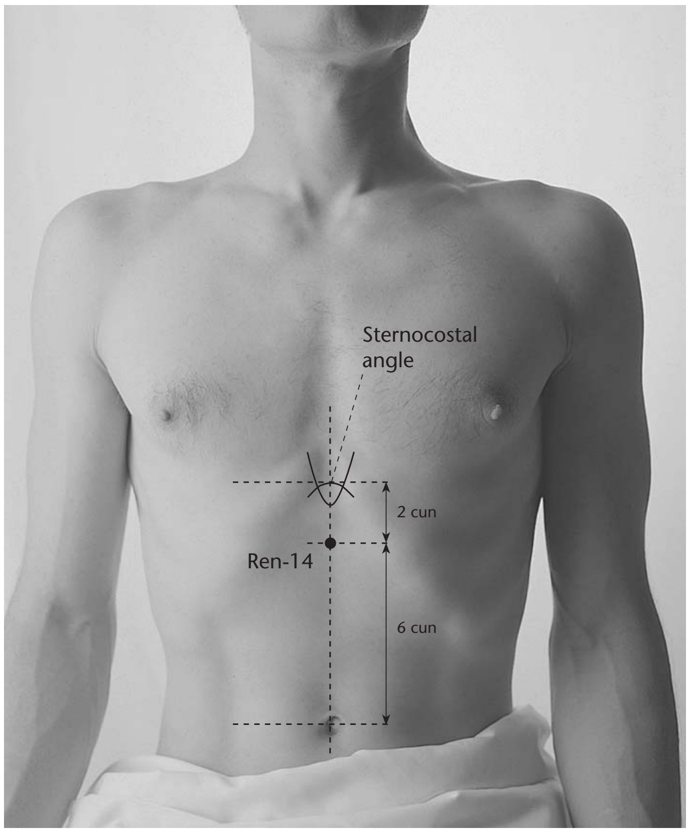
Sternocostal angle
2 cun
Ren-14
6 cun

Needling

0.3–0.8 cun vertically or often obliquely in an inferior direction or up to 1.5 cun transversely (subcutaneously). Depending on individual anatomy, this point can be located on the xiphoid process. Caution: Peritoneum, injury to hypertrophied organs (left: heart, right: liver) depending on the thickness of the abdominal wall.

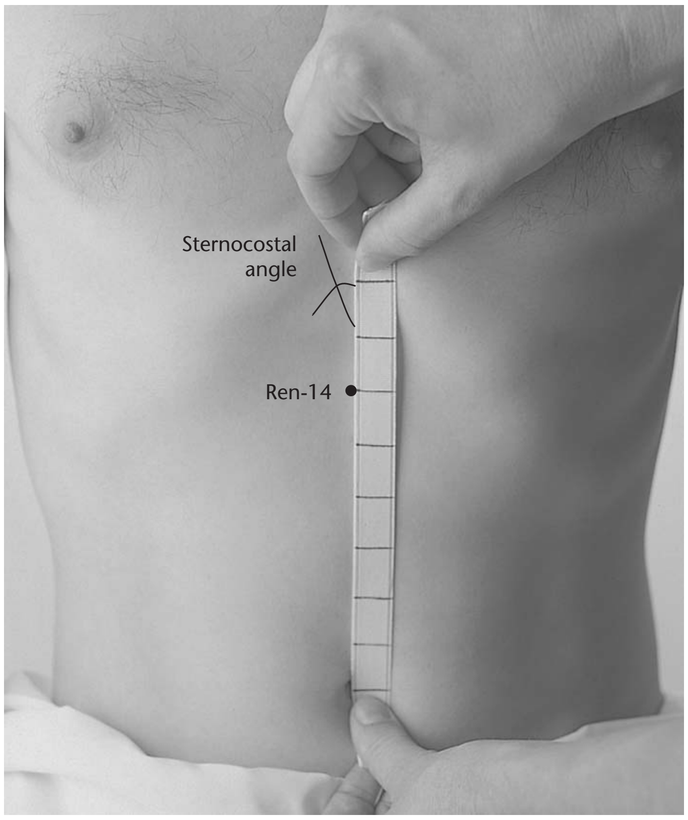
Sternocostal angle
Ren-14

Actions/Indications

- Regulates the Heart, disperses stagnant Phlegm, alleviates pain, unbinds the thorax, descends the Qi
- Harmonises the Stomach, descends Stomach Qi
- Transforms Phlegm, calms the *shen*

Special features

Front-*mu* point of the Heart. Major local point for thoracic pain caused by Blood stasis or Phlegm blockages.

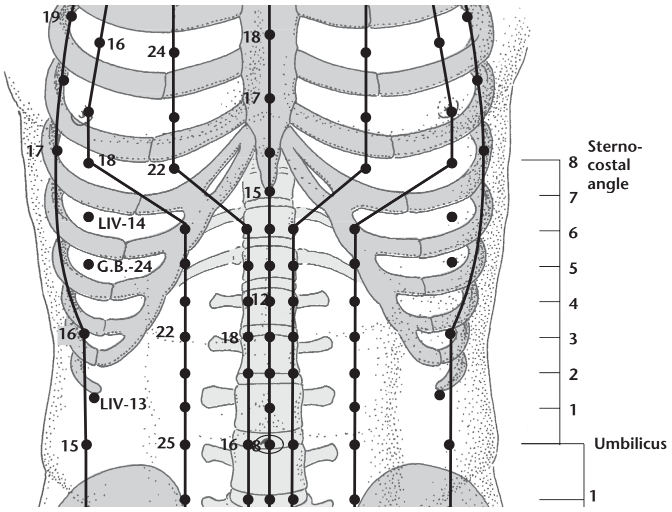
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus
1

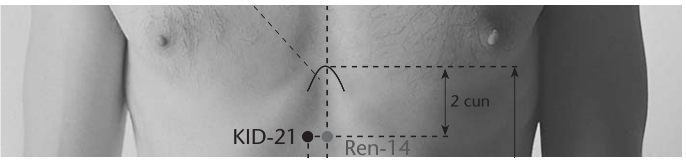
2 cun
KID-21
Ren-14

Turtledove Tail JIUWEI Ren-15

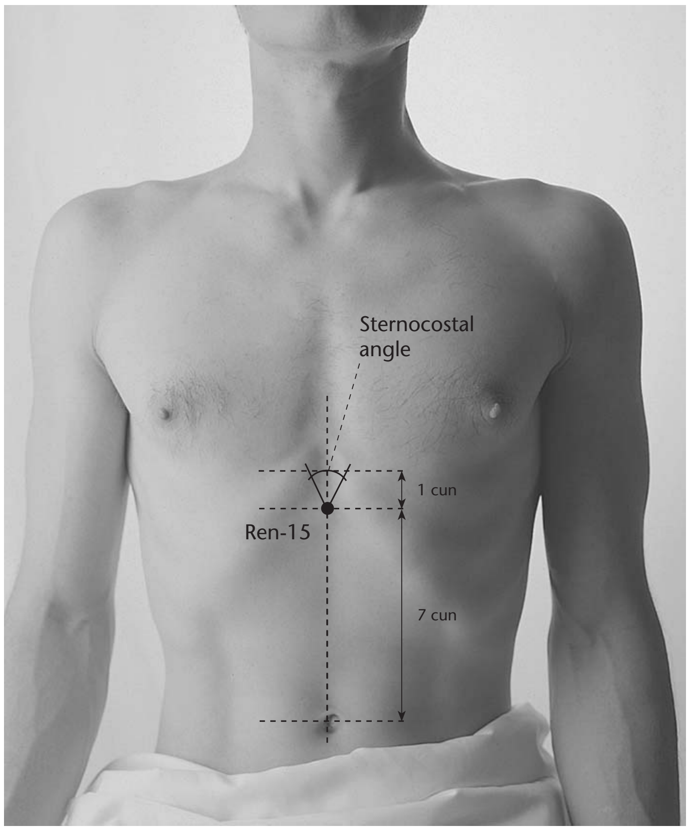
Sternocostal angle
1 cun
Ren-15
7 cun

Location

On the anterior midline, 1 cun inferior to the sternocostal angle or 7 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle (→ 3.5) and the centre of the umbilicus is divided into 8 cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be applied here (helpful device: elastic tape → 2.3.1). From the sternocostal angle, measure 1 cun in an inferior direction and there locate **Ren-15** on the anterior midline. It is often located directly below or, depending on individual anatomy, on the xiphoid process.

Needling

0.5–1 cun obliquely in an inferior direction. The patient should be in an extended position. Caution: Possible injury to organs with deep needling or oblique insertion in a superior direction, especially in patients with hypertrophied organs (left: Liver, right: Heart). According to some classical texts, moxibustion is contraindicated.

Actions/Indications

- Regulates the Heart, calms the *shen*
- Descends counterflow lung and stomach Qi, unbinds the chest
- Regulates the *luo*-connecting vessel of the *ren mai*

Special features

Luo-connecting point of the *ren mai*. Important point for calming.

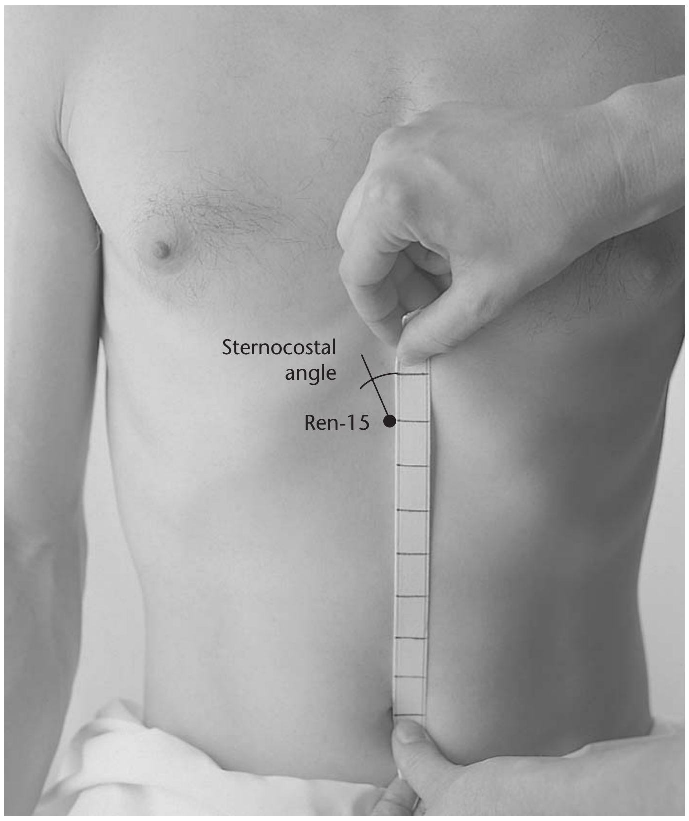
Sternocostal angle
Ren-15

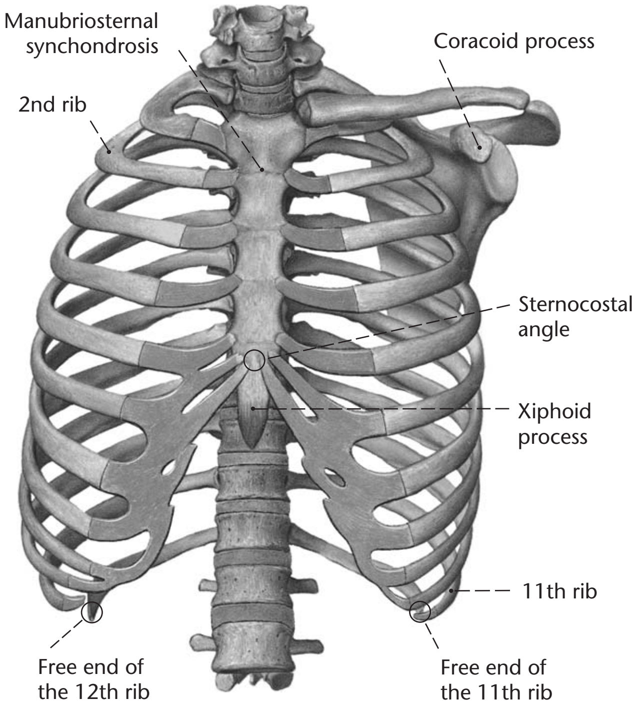
Manubriosternal synchondrosis
2nd rib
Coracoid process
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

Ren-16 *Central Courtyard* ZHONGTING

Location

On the anterior midline, on the level of the sternocostal angle.

How to find

Ren-16 is located on the level of the sternocostal angle, where the two lower borders of the ribcage (→ 3.5) meet. To locate this point, palpate upward along the lower borders of the ribcage with one hand on either side until the palpating fingers meet at the sternocostal angle. Here, locate **Ren-16** in a shallow depression between the sternum and the xiphoid process.

Located on the same level (approximately the 5th intercostal space) are → **KID-22** (2 cun lateral to the midline), → **ST-18** (4 cun lateral to the midline) and → **SP-17** (6 cun lateral to the midline).

Needling

0.5 cun transversely (subcutaneously) in a superior or inferior direction. Caution: Danger of heart injury in patients with a split xiphoid process or a sternal foramen. Therefore, only transverse needling is recommended.

Actions/Indications

- Unbinds the chest
- Regulates the Stomach and counterflow Stomach Qi

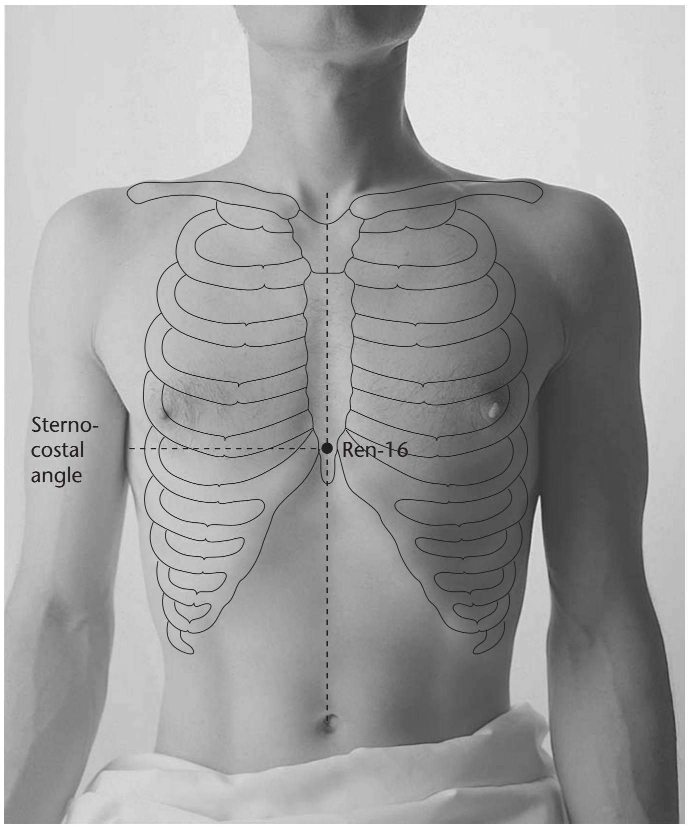
Sterno-
costal
angle
Ren-16

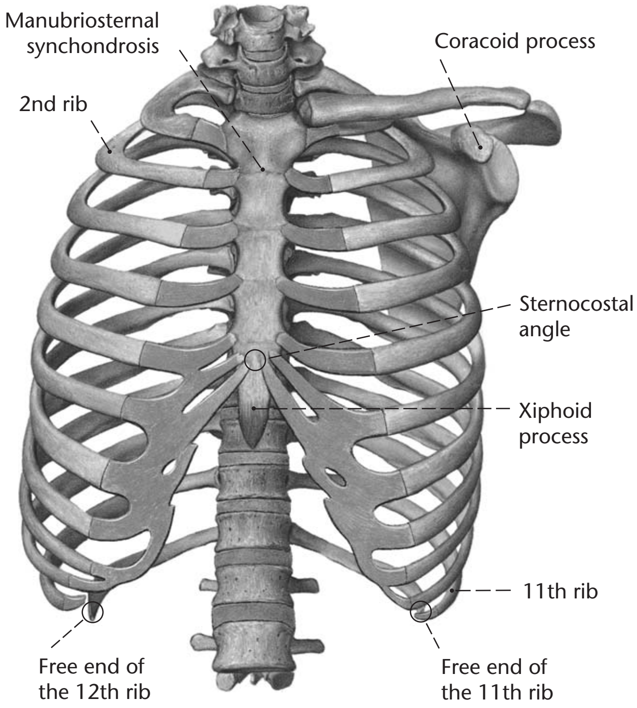
Manubriosternal synchondrosis
Coracoid process
2nd rib
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

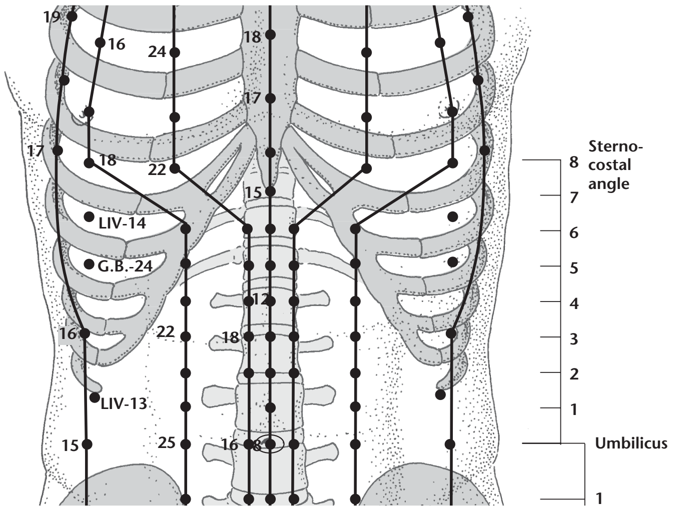
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

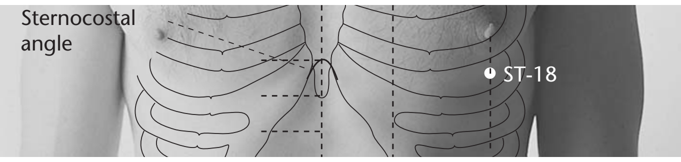
Sternocostal angle
ST-18

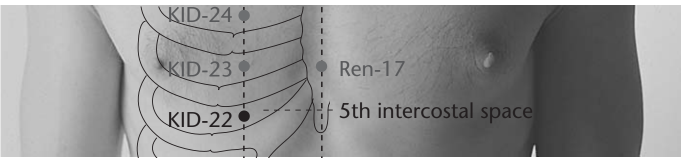
KID-24
KID-23
Ren-17
KID-22
5th intercostal space

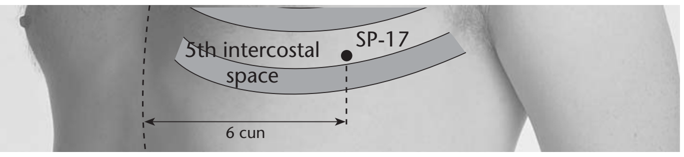
5th intercostal space
SP-17
6 cun

Ch05a-F10028.qxd 2/22/08 9:44 PM Page 487

Chest Centre DANZHONG Ren-17

Location

On the anterior midline, on the level of the 4th intercostal space.

How to find

Quick method, especially in men: Locate **Ren-17** on the sternum, between the two nipples. For more detailed orientation on the intercostal region (→ 3.5), first, locate the manubriosternal synchondrosis, which can be felt on the sternum as a well-defined horizontal bony structure on the level of the costal cartilage of the 2nd rib. The 2nd intercostal space is below it. From there, count in an inferior direction to the 4th intercostal space. Locate **Ren-17** on this level on the midline of the sternum. Located on the same level in the 4th intercostal space are **→ KID-23/ST-17/P-1/SP-18/G.B.-22/G.B.-23** (2 cun lateral to the midline/on the nipple/1 cun lateral to the nipple/6 cun lateral to the midline/on the axillary line/1 cun anterior to **→ G.B.-22**).

Needling

Depending on the indication, 0.5–1 cun transversely (subcutaneously) in a superior or inferior direction or towards the nipples (for mamillary disorders). Caution: Intracardiac needling, pneumothorax. The sternum can be very thin in this area owing to embryonic ossification disorders! Moxibustion is recommended in order to tonify the Qi, but caution with asthmatic patients.

Actions/Indications

- Regulates and benefits the Qi, unbinds the chest, descends counterflow Lung and Stomach Qi
- Benefits the breasts

Special features

Front-*mu* point of the Pericardium, meeting point with the SP, KID, S.I. and T.B. channels. *Hui*-meeting point of the Qi. Sea of Qi. Important point for respiratory disorders.

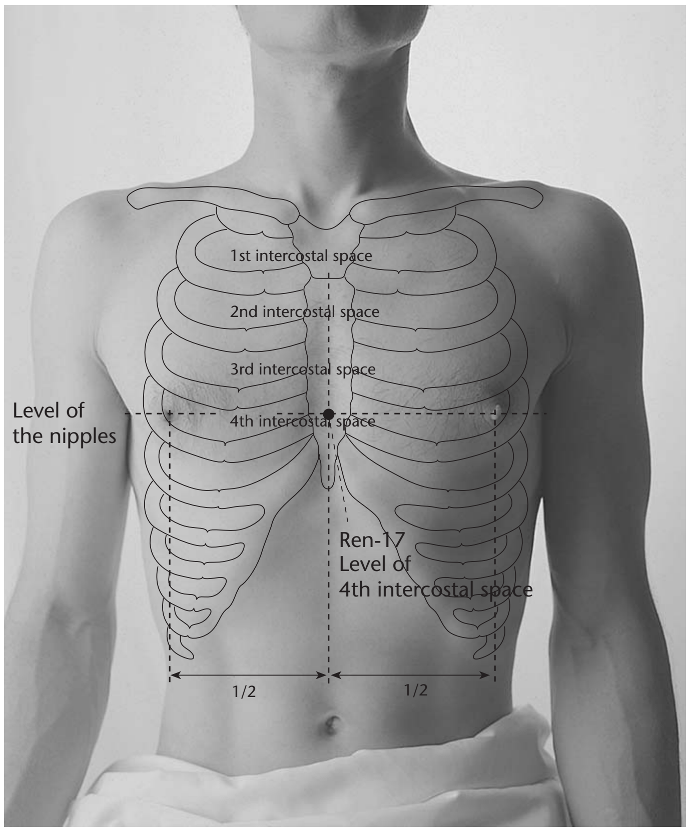
1st intercostal space
2nd intercostal space
3rd intercostal space
Level of the nipples
4th intercostal space
Ren-17
Level of
4th intercostal space
1/2
1/2

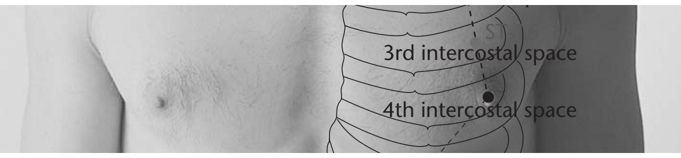
3rd intercostal space
4th intercostal space

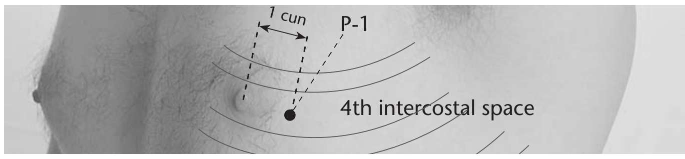
1 cun
P-1
4th intercostal space

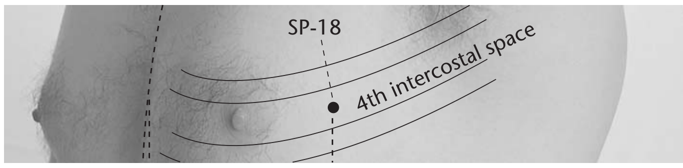
SP-18
4th intercostal space

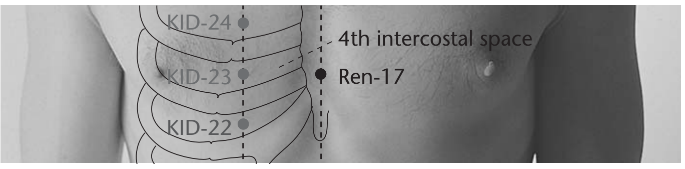
KID-24
4th intercostal space
KID-23
Ren-17
KID-22

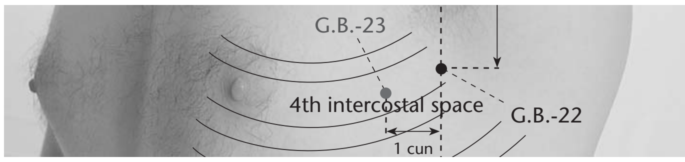
G.B.-23
4th intercostal space
G.B.-22
1 cun

Ren-18 *Jade Hall* YUTANG

Location

On the anterior midline, on the level of the 3rd intercostal space.

How to find

Locate the 3rd intercostal space either by parasternal palpation starting at the clavicle or by counting downward from the manubriosternal synchondrosis (➞ 2nd rib) (➞ 3.5).
Located on the same level (➞ 3rd intercostal space) are ➞ **KID-24/ST-16/SP-19** (2/4/6 cun lateral to the midline).

Needling

0.5–1 cun transversely (subcutaneously) in a superior or inferior direction. Caution: Danger of injury to the heart or the pleura in patients with a sternal foramen, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest, regulates and descends the Qi

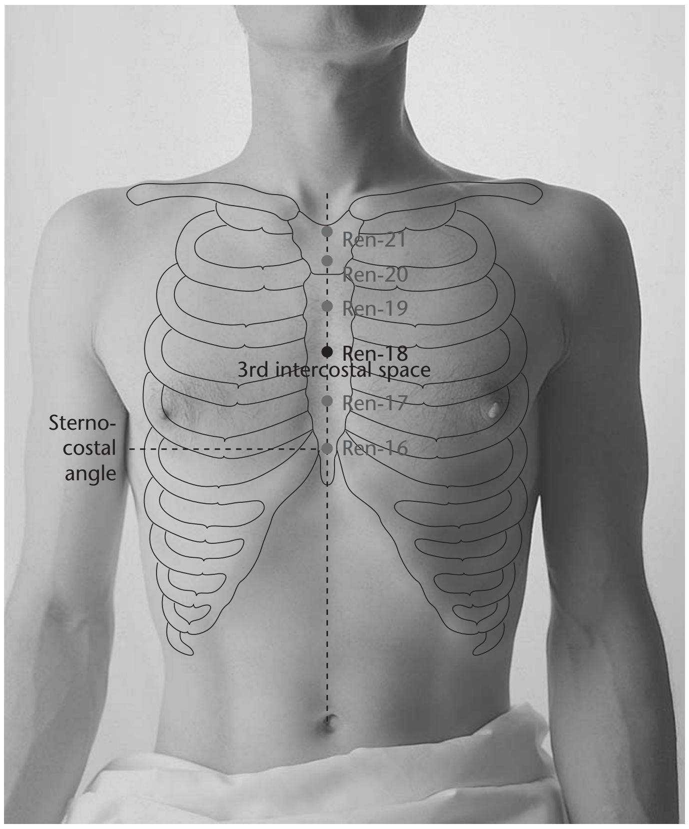
Ren-21
Ren-20
Ren-19
Ren-18
3rd intercostal space
Ren-17
Sterno-
costal
angle
Ren-16

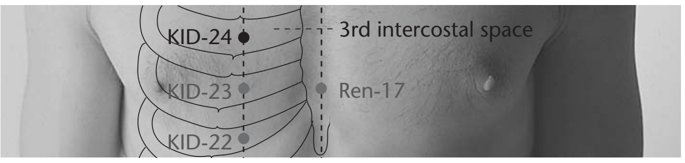
KID-24
3rd intercostal space
KID-23
Ren-17
KID-22

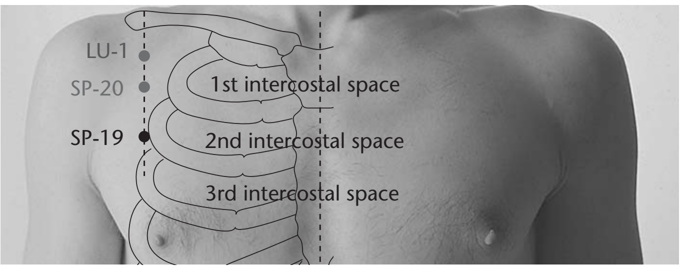
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

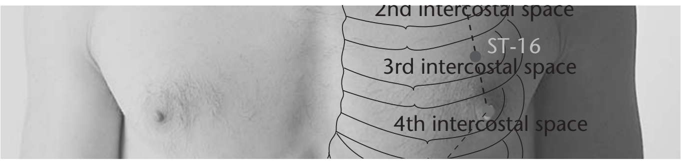
2nd intercostal space
ST-16
3rd intercostal space
4th intercostal space

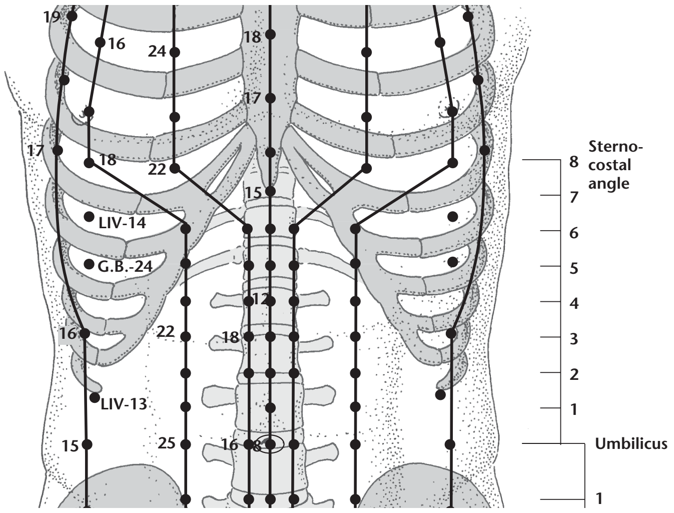
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-
costal
angle
8
7
6
5
4
3
2
1
Umbilicus
1

Purple Palace ZIGONG Ren-19

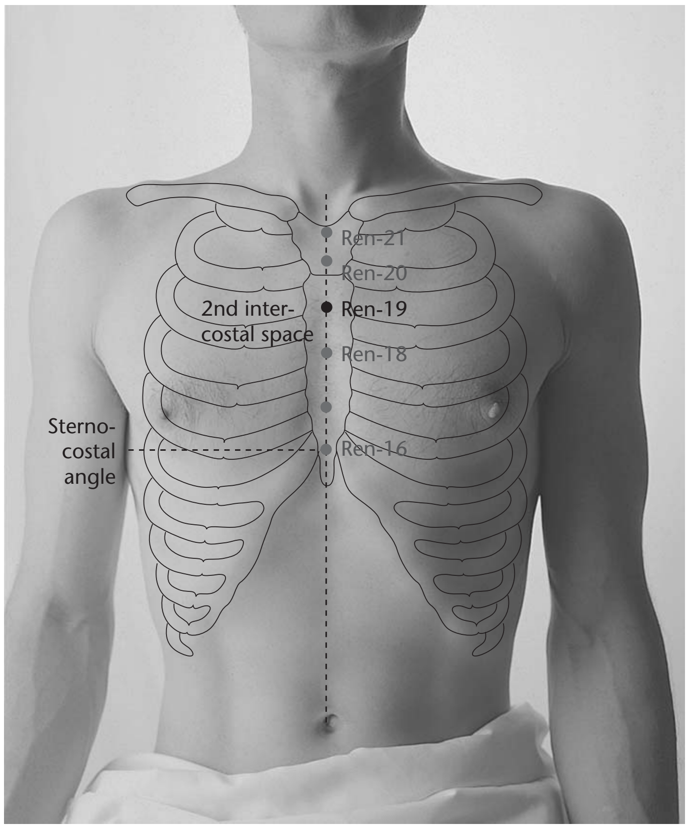
Ren-21
Ren-20
2nd inter-costal space
Ren-19
Ren-18
Sterno-costal angle
Ren-16

Location

On the anterior midline, on the level of the 2nd intercostal space.

How to find

Locate the 2nd intercostal space by parasternal palpation starting at the clavicle or by palpating from the manubriosternal synchondrosis (→ 2nd rib) (→ 3.5).

Located at the same level (2nd intercostal space) are → **KID-25/ST-15/SP-20** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Regulates and descends the Qi

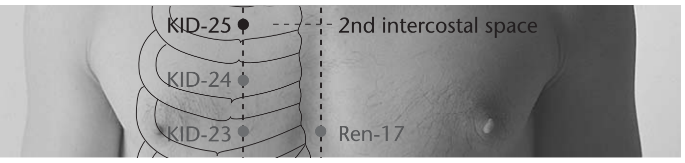
KID-25
2nd intercostal space
KID-24
KID-23
Ren-17

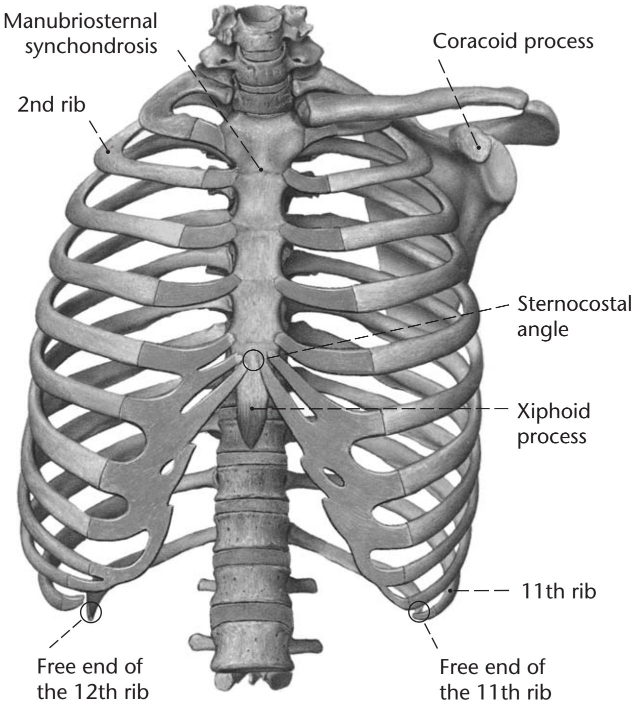
Manubriosternal synchondrosis
Coracoid process
2nd rib
Sternocostal angle
Xiphoid process
11th rib
Free end of the 12th rib
Free end of the 11th rib

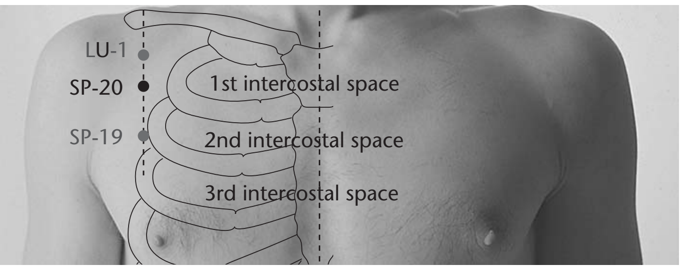
LU-1
SP-20
1st intercostal space
SP-19
2nd intercostal space
3rd intercostal space

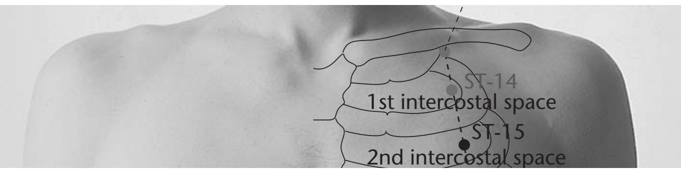
ST-14
1st intercostal space
ST-15
2nd intercostal space

Ren-20 *Magnificent Canopy* HUAGAI

Location

On the anterior midline, on the lower part of the manubrium sterni, on the level of the 1st intercostal space.

How to find

The manubriosternal synchondrosis generally forms a distinct horizontal bony structure on the upper part of the sternum. Locate **Ren-20** slightly superior to the synchondrosis, on the midline and on the level of the 1st intercostal space.

Located at the same level (1st intercostal space) are → **KID-26/ST-14/LU-1** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Regulates and descends the Qi

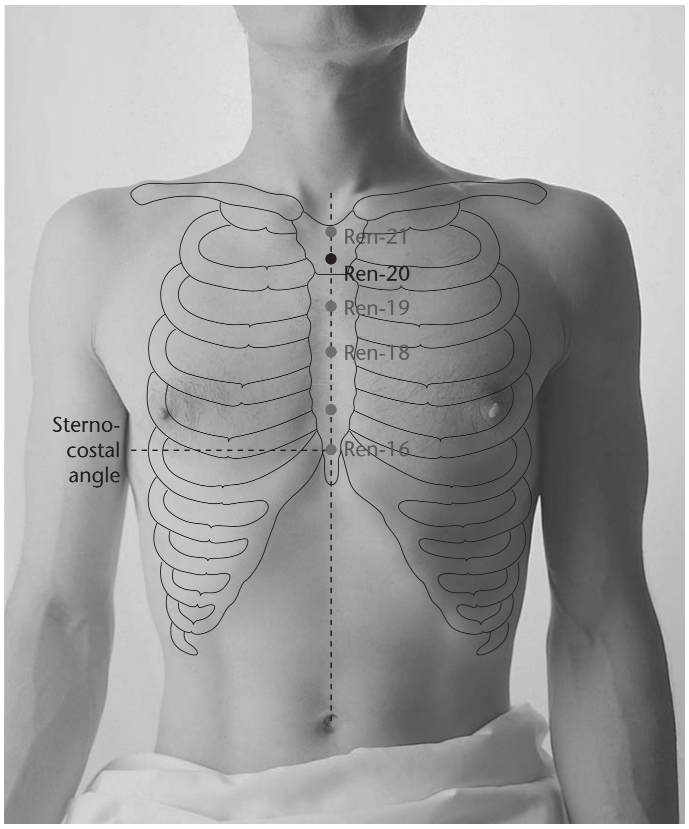
Ren-21
Ren-20
Ren-19
Ren-18
Sterno-costal angle
Ren-16

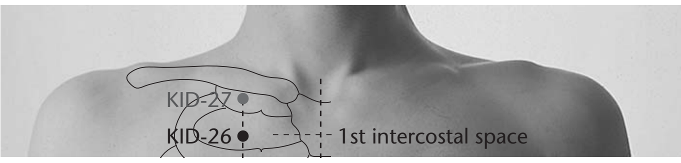
KID-27
KID-26
1st intercostal space

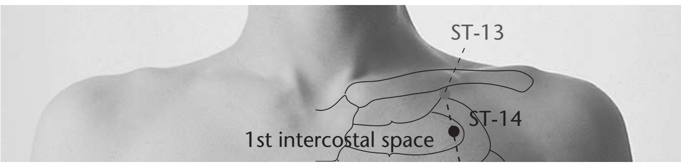
ST-13
ST-14
1st intercostal space

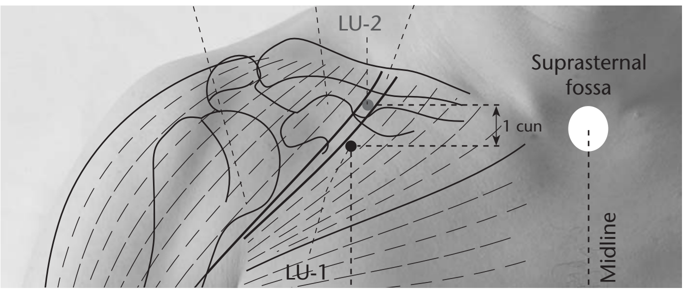
LU-2
Suprasternal fossa
1 cun
LU-1
Midline

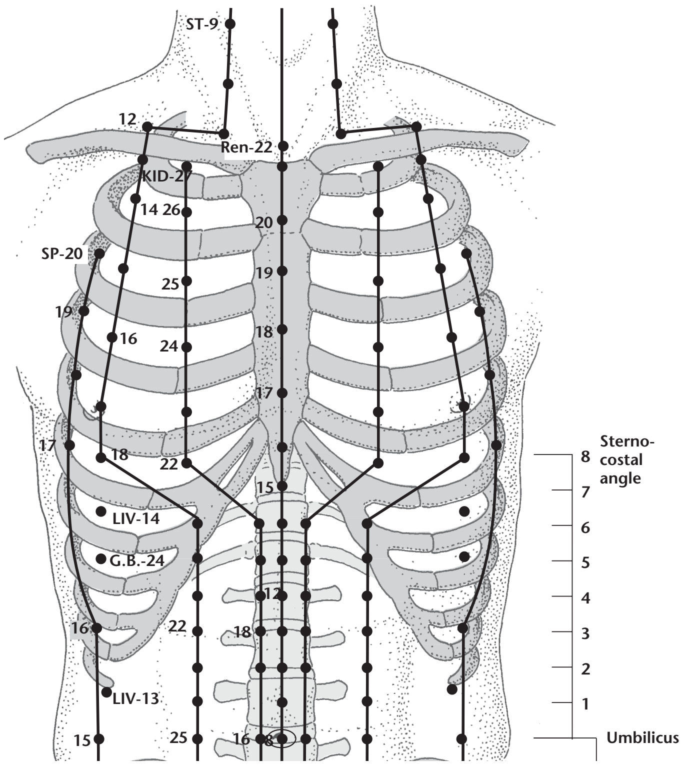
ST-9
12
Ren-22
KID-27
14
26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
Sterno-costal angle
8
7
6
5
4
3
2
1
Umbilicus

Jade Pivot XUANJI **Ren-21**

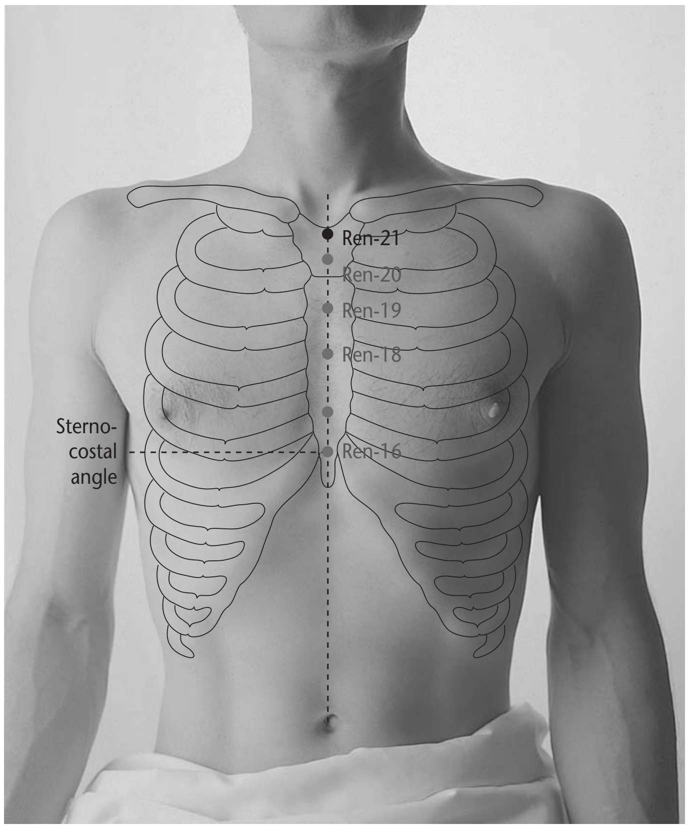
Ren-21
Ren-20
Ren-19
Ren-18
Ren-16
Sterno-
costal
angle

Location

On the anterior midline, below the upper border of the manubrium sterni.

How to find

The upper border of the manubrium sterni forms the lower border of the suprasternal fossa. Locate **Ren-21** on the midline, inferior to the bony margin of the sternum.
Located on approximately the same level (below the clavicle) are → **KID-27/ST-13/LU-2** (2/4/6 cun lateral to the midline).

Needling

0.5 cun transversely in a superior or inferior direction. Caution: In patients with a sternal foramen, danger of injury to the pleura or the mediastinum, therefore only transverse needling with or against the flow of the channel.

Actions/Indications

- Unbinds the chest
- Benefits the throat
- Descends counterflow Qi

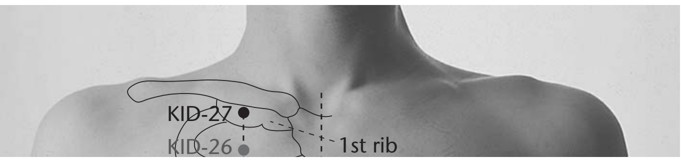
KID-27
KID-26
1st rib

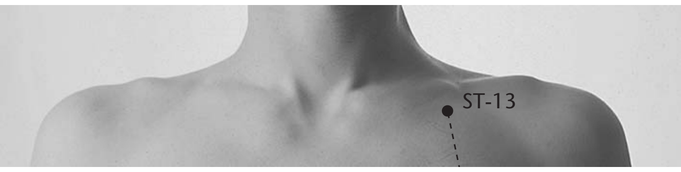

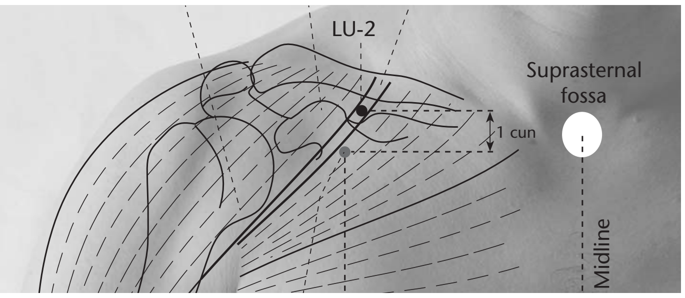
LU-2
Suprasternal
fossa
1 cun
Midline

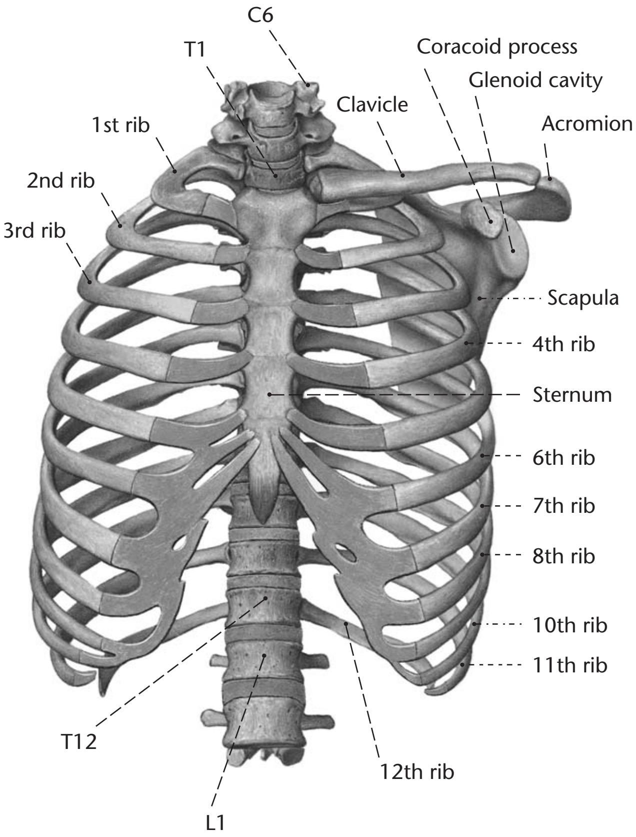
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Ren-22 *Heavenly Prominence* TIANTU

Location

0.5 cun superior to the sternum, in the centre of the suprasternal fossa.

How to find

This point is best located with the patient in a supine position and a pillow under the shoulders or seated with a comfortable and safe head support. Locate the centre of the suprasternal fossa, approximately 0.5 cun superior to the sternum and there locate **Ren-22**.

Needling

Initial vertical insertion 0.2 cun (in order to penetrate the skin), then direct the tip of the needle inferiorly and continue to insert the needle parallel to the posterior aspect of the sternum up to 0.5–1 cun. Caution: Dangerous point. Wrong insertion technique (for example, vertical or oblique retrosternal needling) may lead to injury of the big vessels and organs inside the mediastinum. This point should only be needled by experienced practitioners.

Actions/Indications

- Descends counterflow Lung Qi, alleviates coughing and dyspnoea, benefits the throat

Special features

Meeting point with the *yin wei mai*, Window of Heaven point. Important emergency point for acute asthma attacks.

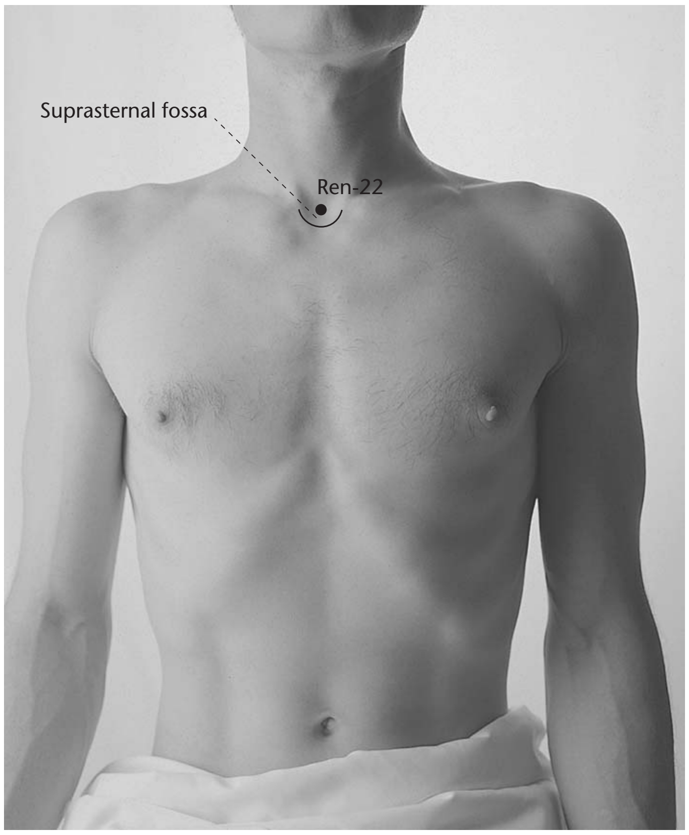
Suprasternal fossa
Ren-22

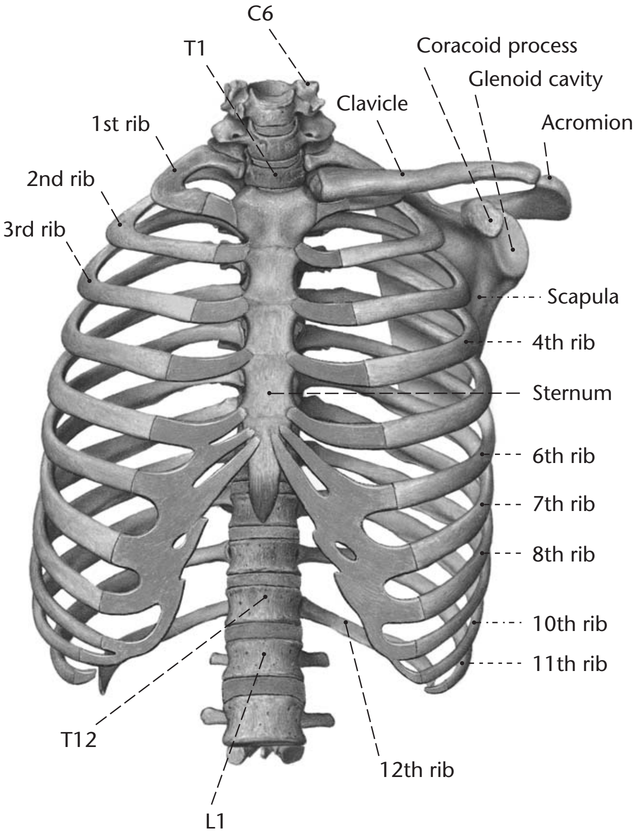
C6
T1
Coracoid process
Glenoid cavity
Clavicle
Acromion
1st rib
2nd rib
3rd rib
Scapula
4th rib
Sternum
6th rib
7th rib
8th rib
10th rib
11th rib
T12
12th rib
L1

Corner Spring LIANQUAN Ren-23

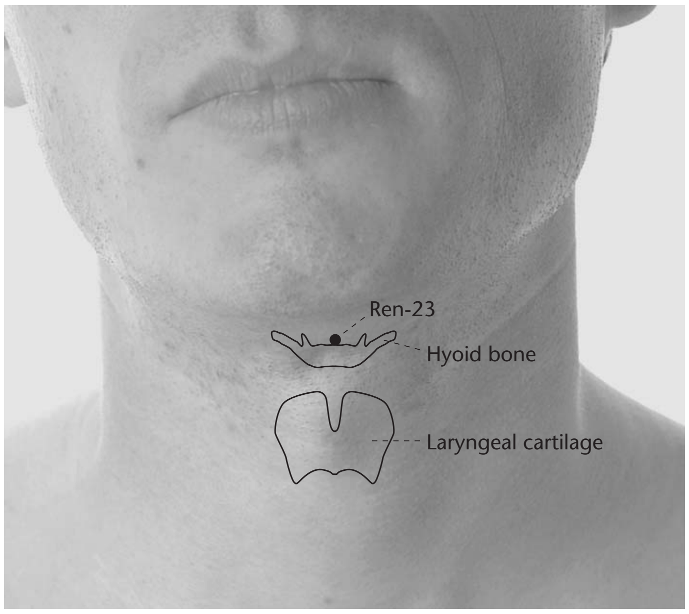
Ren-23
Hyoid bone
Laryngeal cartilage

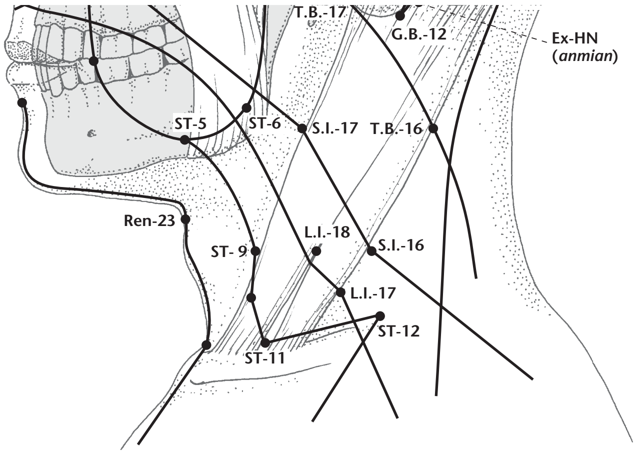
T.B.-17
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST- 9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

Location

On the anterior midline, superior to the upper border of the hyoid bone.

How to find

The hyoid bone can be palpated as a delicate bony structure slightly superior to the upper border of the laryngeal prominence. The patient should not recline their head too much, as this will cause the soft tissue to become stretched, making palpation of the bone difficult. **Ren-23** is located on the midline, on the upper border of the hyoid bone and at the junction of the vertical aspect of the neck to the horizontal floor of the mouth (with normal tissue distribution).

Needling

0.3–1 cun obliquely in a superior direction towards the vertex (→ **Du-20**). Moxibustion is contraindicated according to some texts.

Actions/Indications

- Benefits the tongue
- Benefits the throat, eliminates Phlegm and Heat
- Strengthens the voice

Special features

Meeting point with the *yin wei mai*

Ren-24 *Container of Fluids* CHENGJIANG

Location

On the anterior midline, below the lower lip, in the mentolabial groove.

How to find

On the chin, locate the mentolabial groove (→ 3.1.3), which forms a horizontal depression at the transition from the chin to the lower lip. Locate **Ren-24** in this groove, on the midline.

Needling

0.2–0.3 cun obliquely in a superior direction.

Actions/Indications

- Eliminates (external) Wind, opens the *luo*-connecting vessels, alleviates pain and swellings, regulates the *ren mai*
- As Sun Si Miao Ghost point for epilepsy and mania

Special features

Meeting point with the *du mai*, ST and L.I. channels, Sun Si Miao Ghost point. Important local point.

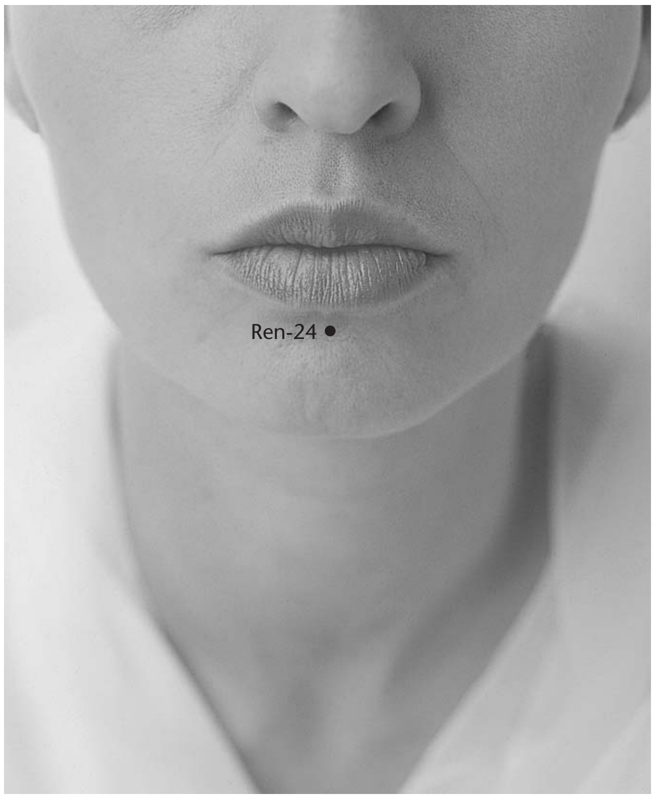

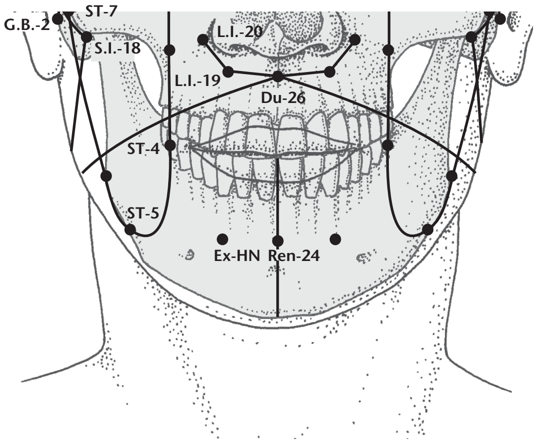

5.3 *Du mai*

Synonym: Governing Vessel

Relationships (→ 1.7.3)

- **Yin/Yang:** *ren mai*/*du mai*
- **Central/peripheral:** *du mai*/*yang qiao mai*. Regions supported by this pairing: inner canthus of the eye, occiput, shoulders, back, S.I. and BL channels, opening point: **S.I.-3** (*houxi*), coupled point: **BL-62** (*shenmai*)

Pathway

The extraordinary vessel *du mai* originates like the *chong mai* and *ren mai* in the lower abdomen (*bao zhong*), according to some authors at the Kidneys, emerges at the perineum and divides into several branches:

➡ The **main pathway** of the *du mai* begins at → **Ren-1** (*huiyin*) at the perineum, reaches **Du-1** (*changqiang*) between the anus and the coccyx and ascends the posterior midline along the spine. At **Du-12** (*shenzhu*), below the spinous process of T3, a branch separates on either side to reach → **BL-12** (*fengmen*), rejoining the main branch at **Du-13** (*taodao*) below the spinous process of T1. It then continues to **Du-16** (*fengfu*) on the lower border of the occipital bone. Here, an **internal branch** enters the Brain. The external pathway continues to ascend the head to **Du-20** (*baihui*), from there beginning its descent along the midline to the forehead and the nose to **Du-26** (*renzhong*) where it crosses the ST and L.I. channels. It terminates inside the mouth below the frenulum at **Du-28** (*yinjiao*) where it meets the ST channel and the extraordinary vessel *ren mai*.

➡ **one branch** of the *du mai* (the spinal branch) originates at → **Ren-1** (*huiyin*), courses to the tip of the coccyx where it meets the Kidney primary channel, ascends **within** the spinal canal to the space between L2 and L3 and enters the Kidneys.

➡ **The second branch** of the *du mai* (the abdominal branch) also originates at → **Ren-1** (*huiyin*), encircles the external genitalia and ascends to the umbilicus, passes through the Heart (*xin*), continues to the throat, circles the lips, travels to the cheek, passes → **ST-1** (*chengqi*) at the infraorbital ridge and reaches → **BL-1** (*jingming*).

➡ **The third branch** of the *du mai* travels bilaterally alongside the BL channel from → **BL-1** (*jingming*) to the vertex. From here, it branches out **internally** into the Brain. The **external** pathway follows the medial branches of the BL channel in an inferior direction and enters the body at the level of the lower border of the spinous process of L2 at → **BL-23** (*shenshu*) to terminate at the Kidneys (*shen*).

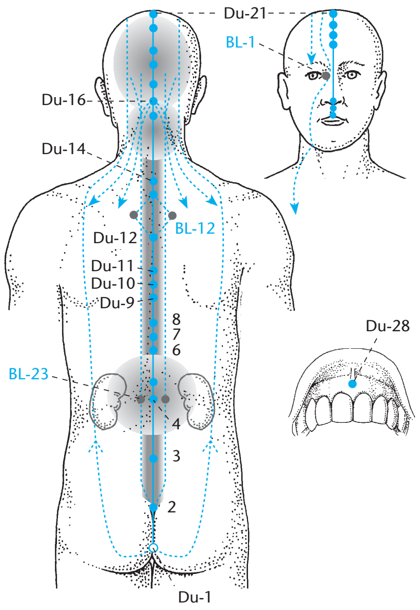

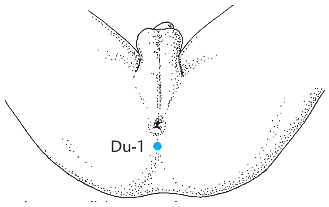

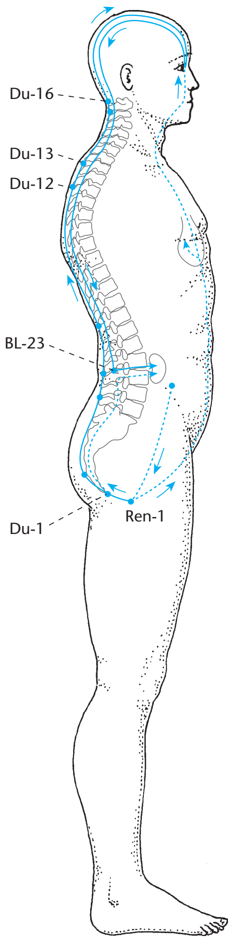

Meeting points of other channels with the *du mai*

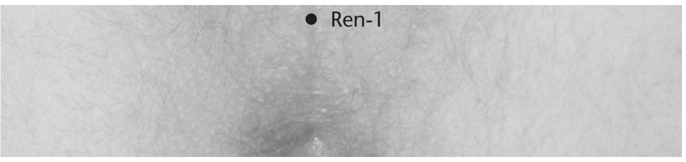

Ren-1 (*huiyin*): In the centre of the perineum

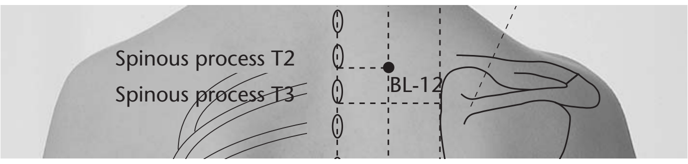

BL-12 (*fengmen*): 1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd thoracic vertebra (T2)

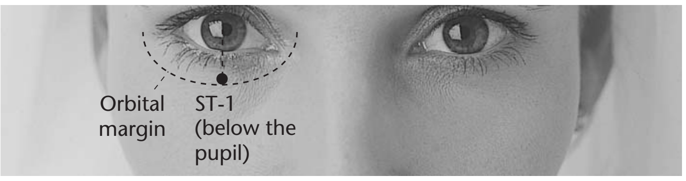

ST-1 (*chengqi*): With the eyes looking straight ahead, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

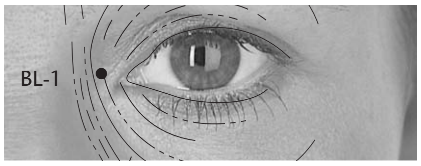

BL-1 (*jingming*): 0.1 cun superior and medial to the inner canthus of the eye, in a depression

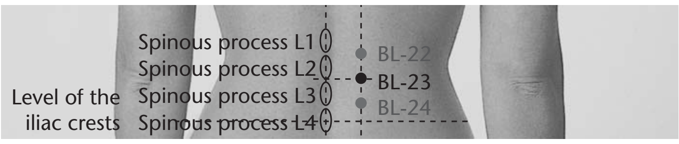

BL-23 (*shenshu*): 1.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 2nd lumbar vertebra (L2)

Connections with other channels/Organs

- *Ren mai*, *chong mai*, Bladder and Kidney primary channels
- Uterus, Kidneys, Brain

Clinical importance (→ 1.7.2, 1.7.3)

- **Sea of the Yang Qi**, supplies Yang Qi in order to warm the Organs and channels
- Supports and regulates all Yang channels, especially those of the upper back and head
- Influences and strengthens the functions of the Brain, Marrow and sensory organs
- Strengthens the body constitutionally and supports the immune system (circulation of the Defensive Qi (*wei qi*))
- Stores the Essence Qi (*jing qi*) overflowing from the Yang channels
- Absorbs excess from the Yang channels, especially excess Heat and internal Wind

Notes about the *du mai*

Both the *du mai* and *ren mai* have a unique position among the eight extraordinary vessels. They are the only two extraordinary vessels with their own specific points – all the other extraordinary vessels can only be accessed through points on the primary channels. For this reason the *ren mai* and the *du mai*, together with the primary channels, are often categorised as 14 channels. However, from the point of view of channel energetics this is not correct, since the extraordinary vessels occupy a special position, both evolutionarily and functionally (→ 1.7).

The pathway of the *du mai*, like that of the *ren mai*, is limited to the torso and the head. While it does not course along the extremities and does not have any points on the limbs (neither general points nor specific points such as the five *shu*-transporting points, *yuan*-source points, etc.) it does have a pertaining opening point (or master point) (→ 8.1.8).

The *du mai* is also referred to as the Sea of Yang and governs the Yang in the whole body. Located along its pathway are the posterior aspects of the energy centres (or chakras according to Yogi tradition) while their anterior aspects are located along the pathway of the *ren mai*. In Daoist tradition the *du mai* plays an important role in the cultivation of Qi. This concerns the development of a 'spirit child', which can be influenced by opening and activating energy centres located along the spine.

Furthermore, the *du mai* contains many meeting points with other channels. This clearly denotes the outstanding importance of the *du mai*.

Specific points according to their function

- ***Yuan*-source point (→ 8.1.1): –**
- ***Luo*-Connecting point (→ 8.1.2.): Du-1** (*changqiang*)
- ***Xi*-cleft point (→ 8.1.3): –**
- **Back-*shu* point (→ 8.1.4): –**
- **Front-*mu* point (→ 8.1.5): –**
- **Five *shu*-transporting points (→ 8.1.6): –**
- ***Hui*-meeting point (→ 8.1.7): –**
- **Pertaining opening point (→ 8.1.8): S.I.-3** (*houxi*)
- **Pertaining coupled point (→ 8.1.8): BL-62** (*shenmai*)
- **Lower *he*-Sea point (→ 8.1.9): –**
- ***Jiaohui*-meeting points (→ 8.1.10):**

  - with the G.B. and KID channels and the *ren mai*: **Du-1** (*changqiang*)
  - with the BL channel: **Du-13** (*taodao*)
  - with all Yang channels: **Du-14** (*dazhui*) 🟦🟦
  - with the *yang wei mai*: **Du-15** (*yamen*) 🟦■
  - with the *yang wei mai* and the *yang qiao mai**: **Du-16** (*fengfu*) 🟦🟦
  - with the BL channel: **Du-17** (*naohu*)
  - with the BL, G.B., T.B. and LIV channels: **Du-20** (*baihui*) 🟦🟦
  - with the BL and ST channels: **Du-24** (*shenting*) 🟦■
  - with the L.I. and ST channels: **Du-26** (*renzhong*) 🟦🟦
  - with the *ren mai* and ST channel*: **Du-28** (*yinjiao*)
  - of other channels with the *du mai*: **Ren-1, BL-12**, **ST-1, BL-1, BL-23**
- **Gao Wu command point (→ 8.1.11): –**
- **Window of Heaven point (→ 8.1.12): Du-16** (*fengfu*) 🟦🟦
- **Points of the Four Seas (→ 8.1.13):**
  - Sea of Qi: **Du-14** (*dazhui*) 🟦🟦, **Du-15** (*yamen*) 🟦■
  - Sea of Marrow: **Du-16** (*fengfu*) 🟦🟦; **Du-20** (*baihui*) 🟦🟦
- **Ma Dan Yang Heavenly Star point (→ 8.1.14):**
- **Sun Si Miao Ghost point (→ 8.1.15): Du-16** (*fengfu*) 🟦🟦, **Du-26** (*renzhong*) 🟦🟦
- **Other functional points:**
  - very important point for strengthening the Yang (with moxibustion): **Du-4** (*mingmen*) 🟦🟦
  - important emergency point: **Du-26** (*renzhong*) 🟦🟦

Points according to region

- **Local points (→ 8.2.1):** Temporal region – **Du-20** (*baihui*) 🟦🟦; nose – **Du-25** (*suliao*); rectum – **Du-1** (*changqiang*); lumbar spine – **Du-3** (*yaoyangguan*), **Du-4** (*mingmen*) 🟦🟦
- **Adjacent points (→ 8.2.1):** occipital region – **Du-14** (*dazhui*) 🟦🟦; temporal region, eyes – **Du-23** (*shangxing*); nose – **Du-23** (*shangxing*), **Du-24** (*shenting*); Lungs – **Du-14** (*dazhui*) 🟦🟦; Heart – **Du-11** (*shendao*); cervical spine – **Du-14** (*dazhui*) 🟦🟦
- **Distal points (→ 8.2.1):** tongue – **Du-15** (*yamen*) 🟦■; rectum – **Du-20** (*baihui*) 🟦🟦; lumbar spine – **Du-12** (*shenzhu*), – **Du-26** (*renzhong*) 🟦🟦

Specific points according to the vessel pathway (in numerical order)

- **Du-1** (*changqiang*): *luo*-connecting point (→ 8.1.2); *jiaohui*-meeting point with the G.B., KID channels and the *ren mai* (→ 8.1.10); local point for the rectum (→ 8.2.1)
- **Du-3** (*yaoyangguan*): local point for the lumbar spine (→ 8.2.1)
- **Du-4** (*mingmen*) 🟦🟦: very important point for strengthening the Yang (with moxibustion)
- **Du-11** (*shendao*): adjacent point for the Heart (→ 8.2.1)
- **Du-12** (*shenzhu*): distal point for the lumbar spine (→ 8.2.1)
- **Du-13** (taodao): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **Du-14** (*dazhui*) 🟦🟦: *jiaohui*-meeting point with all Yang channels (→ 8.1.10); point of the Sea of Qi (→ 8.1.13); adjacent point for the occiput and Lung (→ 8.2.1); adjacent point for the cervical spine (→ 8.2.1)
- **Du-15** (*yamen*) 🟦■: *jiaohui*-meeting point with the *yang wei mai* (→ 8.1.10); Sea of Qi point (→ 8.1.13); distal point for the tongue (→ 8.2.1)
- **Du-16** (*fengfu*) 🟦🟦: point of the Sea of Marrow (→ 8.1.13); Window of Heaven point (→ 8.1.12); Sun Si Miao Ghost point (→ 8.1.15); *jiaohui*-meeting point with the *yang wei mai* and the *yang qiao mai** (→ 8.1.10)
- **Du-17** (*naohu*): *jiaohui*-meeting point with the BL channel (→ 8.1.10)
- **Du-20** (*baihui*) 🟦🟦 *jiaohui*-meeting point with the BL, G.B., T.B. and LIV channels (→ 8.1.10); point of the Sea of Marrow (→ 8.1.13); local point for the temporal aspect of the head (→ 4.2.1); distal point for the rectum (→ 8.2.1)
- **Du-23** (*shangxing*): regional point for the temporal region, the eyes and the nose (→ 8.2.1)
- **Du-24** (*shenting*) 🟦■: *jiaohui*-meeting point with the BL and ST channels (→ 4.1.10); adjacent point for the nose (→ 8.2.1)
- **Du-25** (*suliao*): local point for the nose (→ 4.2.1)
- **Du-26** (*renzhong*) 🟦🟦: *jiaohui*-meeting point with the L.I. and ST channels (→ 8.1.10); Sun Si Miao Ghost point (→ 8.1.15); distal point for the lumbar spine (→ 8.2.1); important emergency point
- **Du-28** (*yinjiao*): *jiaohui*-meeting point with the *ren mai* and ST channel (→ 8.1.10)

---

* Mentioned by only some authors

luo-Connecting Vessel of the *Du Mai*

Pathway

The *luo*-connecting vessel of the *du mai* originates at **Du-1** (*changqiang*) between the anus and the tip of the coccyx, bilaterally ascends the spine to the neck and disperses in the occipital region, where it connects with the Bladder primary channel. Both course together to the shoulder region. There the *luo*-connecting vessel enters deeper into the body at the level of the scapula and spreads in the paravertebral musculature.

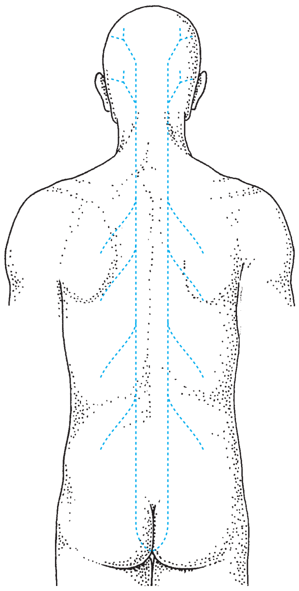

Clinical importance

Pathology

Excess (*shi*): limited range of motion and stiffness of the spine
Deficiency (*xu*): heavy sensations of the head, dizziness, balance disorders

Long Strong CHANGQIANG **Du-1**

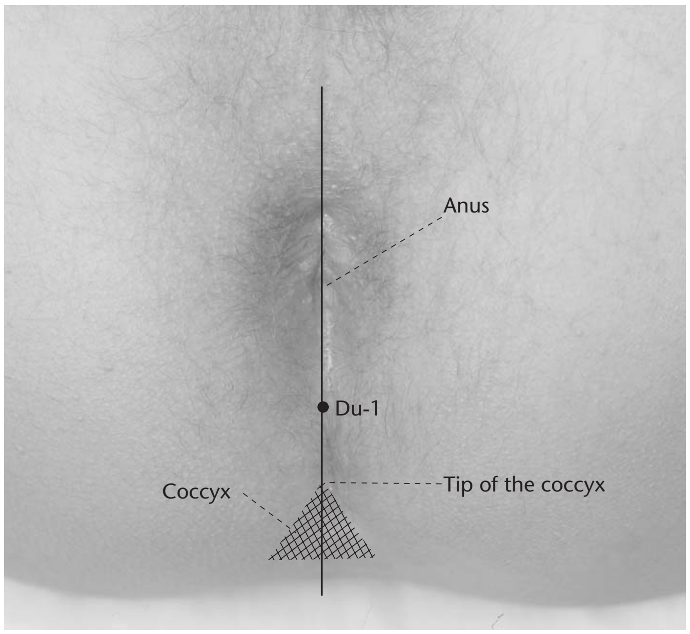
Anus
Du-1
Coccyx
Tip of the coccyx

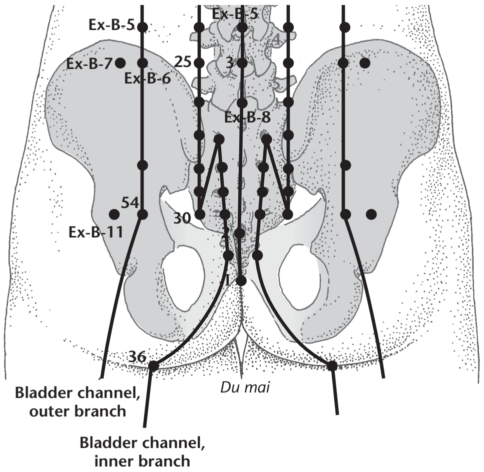
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Bladder channel, outer branch
Du mai
Bladder channel, inner branch

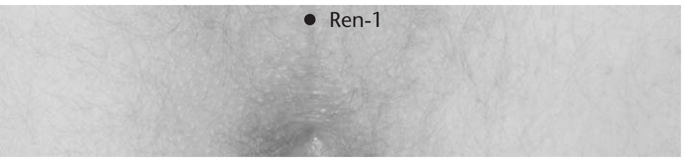

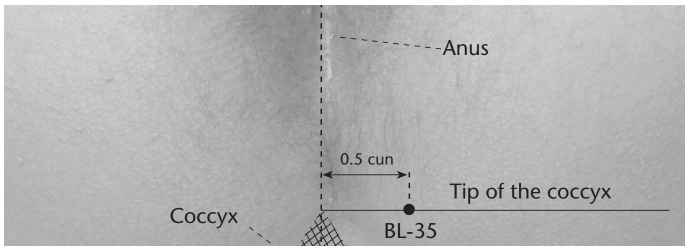
Anus
0.5 cun
Tip of the coccyx
Coccyx
BL-35

Location

On the midline, midway between the tip of the coccyx and the anus.

How to find

First, locate the coccyx superior to the anus, then palpate for its tip. From the tip of the coccyx, palpate with pressure towards the anus. **Du-1** can often be located based on a strong *de* Qi sensation.

→ **Ren-1** is located anterior to the anus, posterior to the genitals.

→ **BL-35** is located 0.5 cun lateral to the midline, at the level of the tip of the coccyx.

Needling

After thorough disinfection, vertically 0.5–1 cun. Avoid puncturing the rectum.

Actions/Indications

- Clears Damp-Heat, lifts the Qi
- Opens the channel and alleviates pain
- Calms the *shen*

Special features

Luo-connecting point, meeting point with the *ren mai*, G.B. and KID channels.

Du-2 *Lumbar Shu* YAOSHU

Location

On the midline, in the sacral hiatus.

How to find

Palpate for the sacral hiatus (→ 3.4.4) from the sacrum. It can be felt at the inferior end of the iliac median crest as a U-shaped depression opening towards the inferior. **Du-2** is located on the midline, directly below the hiatus.

→ **BL-35** is located slightly more inferiorly, 0.5 cun lateral to the midline, on the level of the tip of the coccyx.

Needling

Up to 1 cun vertically or obliquely in a superior direction.

Actions/Indications

- Strengthens the lumbar region and the legs
- Dispels Wind-Dampness
- Warms the Lower Burner

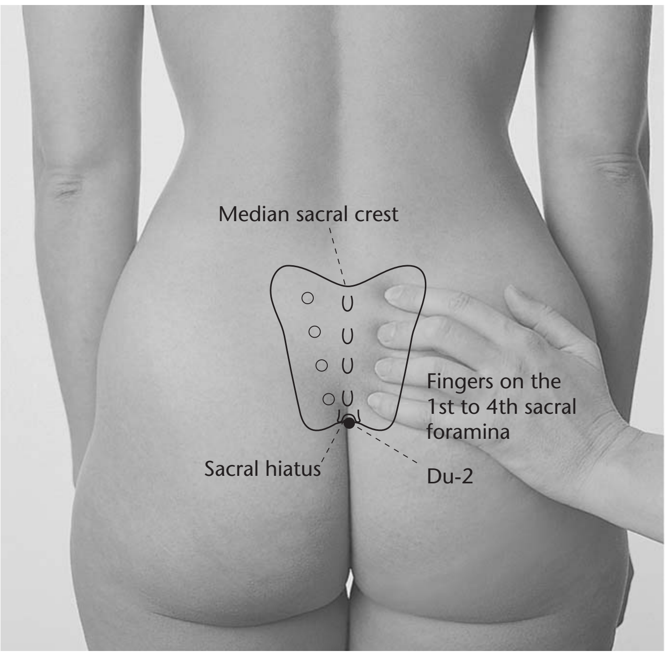
Median sacral crest
Fingers on the 1st to 4th sacral foramina
Sacral hiatus
Du-2

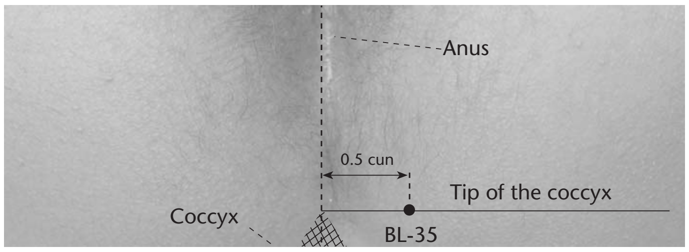
Anus
0.5 cun
Tip of the coccyx
Coccyx
BL-35

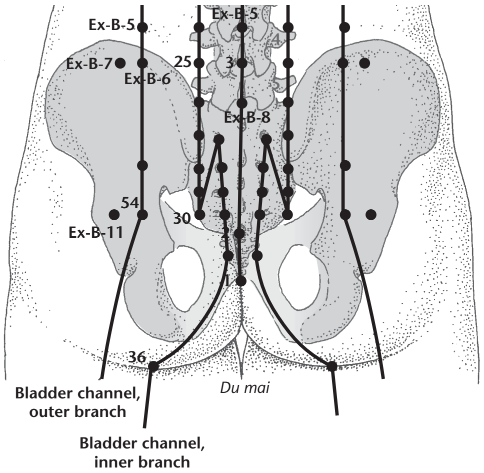
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Bladder channel, outer branch
Du mai
Bladder channel, inner branch

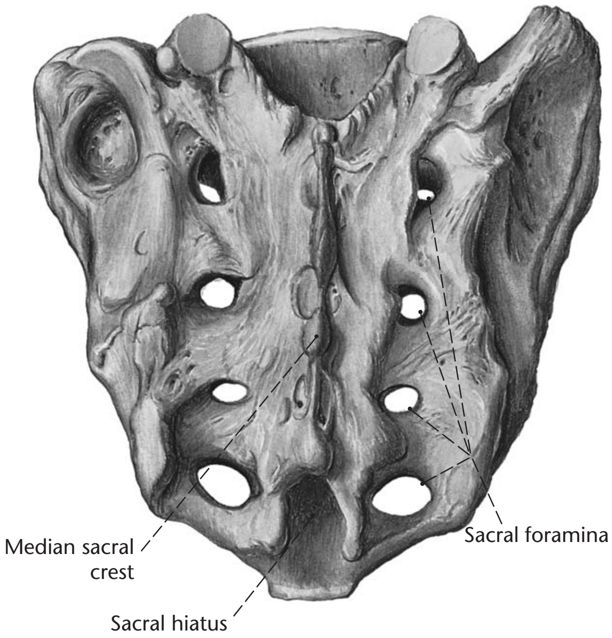
Median sacral crest
Sacral foramina
Sacral hiatus

Lumbar Yang Gate YAOYANGGUAN Du-3

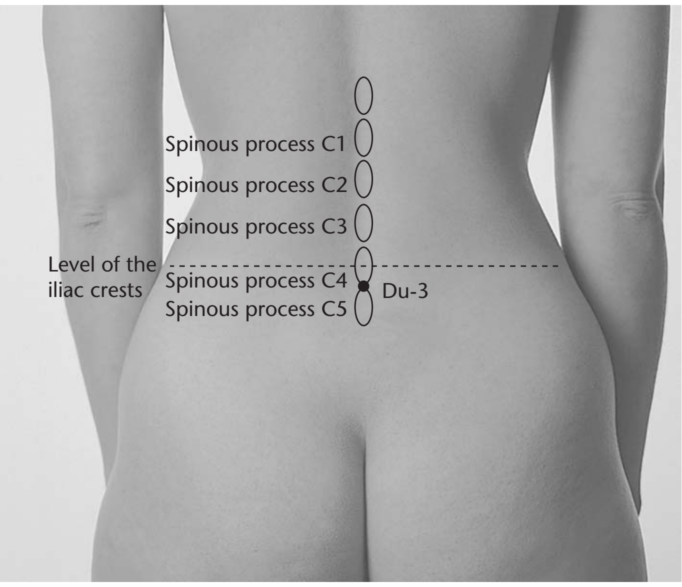
Spinous process C1
Spinous process C2
Spinous process C3
Level of the iliac crests
Spinous process C4
Du-3
Spinous process C5

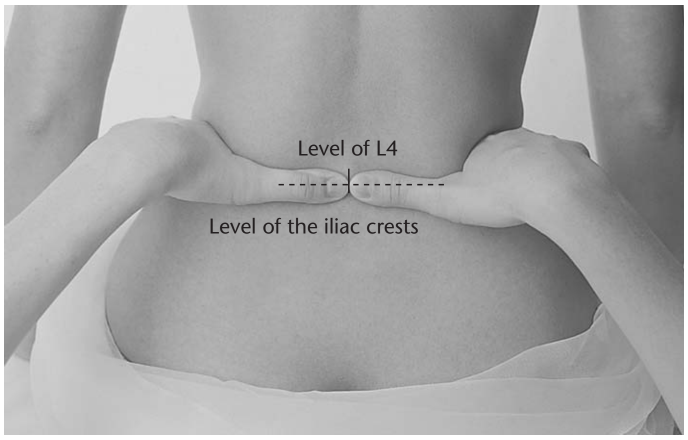
Level of L4
Level of the iliac crests

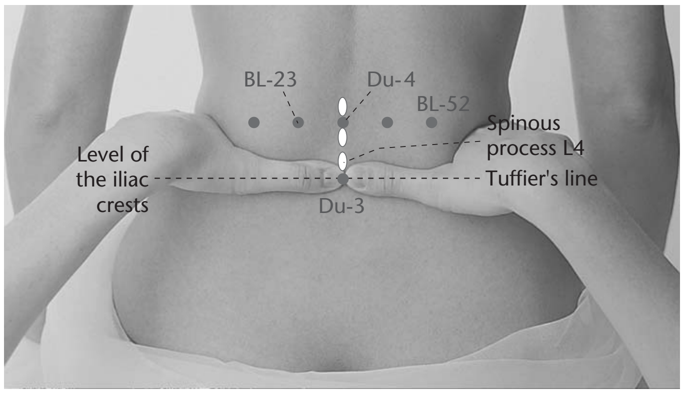
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Location

On the midline, below the spinous process of the 4th lumbar vertebra (L4).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line (a line connecting the two most superior points on the iliac crest), which generally intersects with the spinous process of L4. (Note: Point location in the lumbar region is dependent on the patient's position; for more detail on orientation → 3.4). Locate **Du-3** on the midline, below the spinous process of L4.

Located on the same level are a point of → **Ex-B-2/BL-25/Ex-B-6/Ex-B-7** (0.5/1.5/3/3.5 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction (generally, there is only danger of accidental puncture of the spinal canal from the level of L2 in a cranial direction).

Actions/Indications

- Dispels Wind, Cold and Dampness
- Regulates Qi in the Lower Burner, strengthens the Original Qi (*yuan qi*)

Special features

Important local point for combined back and leg pain

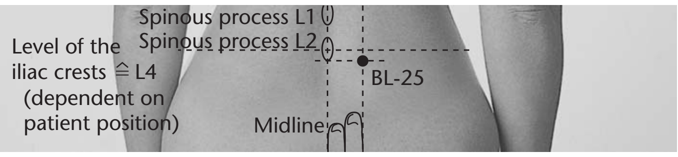
Spinous process L1
Spinous process L2
Level of the iliac crests ≙ L4 (dependent on patient position)
BL-25
Midline

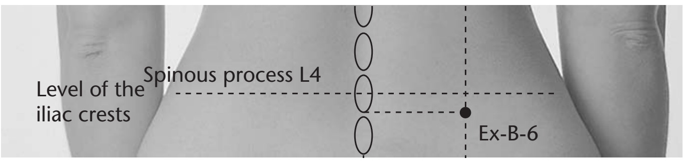
Spinous process L4
Level of the iliac crests
Ex-B-6

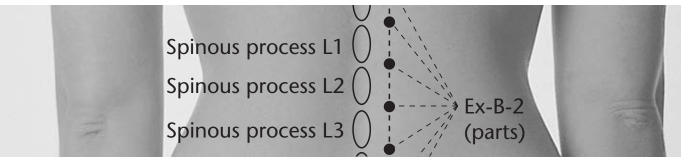
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

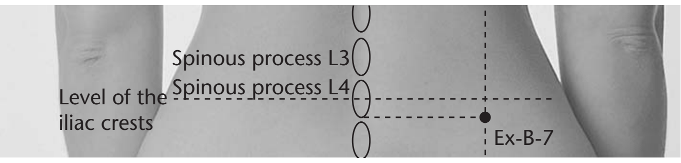
Spinous process L3
Spinous process L4
Level of the iliac crests
Ex-B-7

Du-4 *Gate of Life* MINGMEN

Location

On the posterior midline, below the spinous process of the 2nd lumbar vertebra (L2).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line by placing both hands directly superior and lateral to the two highest points on the iliac crest, joining the thumbs on the midline. The line will generally intersect with the spinous process of L4. (Note: Orientation on the lumbar region depends on the patient's position; for more detail on differential orientation → 3.4). From there, count up to the spinous process of L2 and locate **Du-4** on the midline, below the spinous process. Located on the same level are a point of → **Ex-B-2/BL-23/BL-52** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun vertically to slightly obliquely in an inferior direction, with the patient's back straight or flexed. The spinal cord generally ends between the 1st and 2nd lumbar vertebrae; still, there is hardly any mention in the literature of spinal punctures with deep needling and/or insertions in a superior direction.

Actions/Indications

- Tonifies Kidney Yang (especially with moxibustion), warms the *mingmen*, strengthens the Kidneys, regulates the *du mai*, benefits the lumbar region
- Pacifies Wind in the *du mai*

Special features

Major point for tonifying Yang, especially Kidney Yang

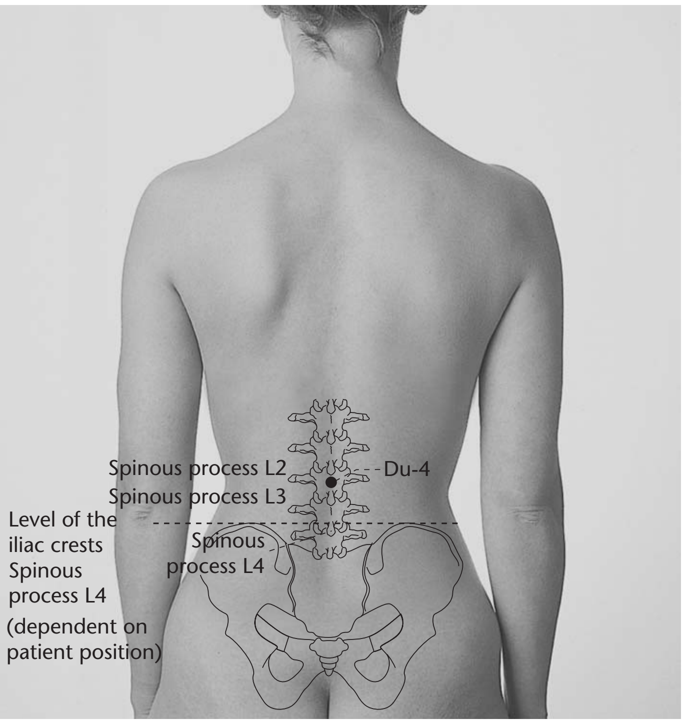
Spinous process L2
Du-4
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L4 (dependent on patient position)

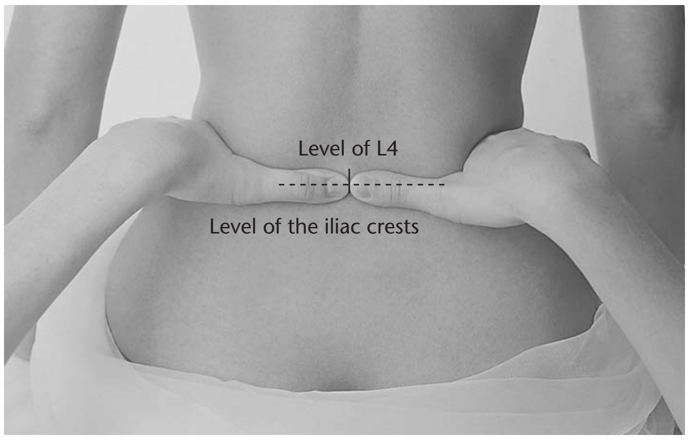
Level of L4
Level of the iliac crests

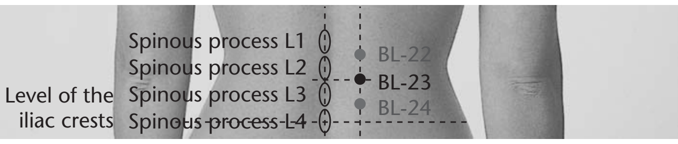
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Level of the iliac crests
BL-22
BL-23
BL-24

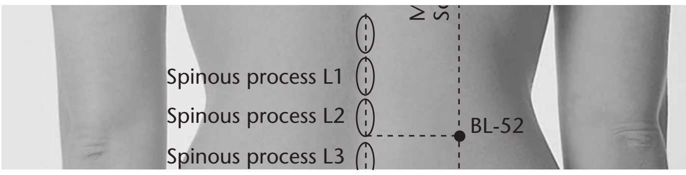
Spinous process L1
Spinous process L2
Spinous process L3
BL-52

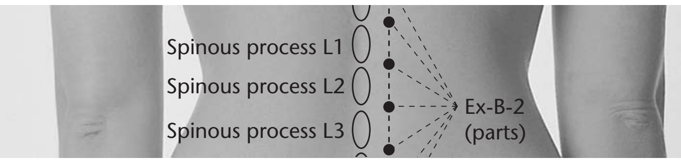
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

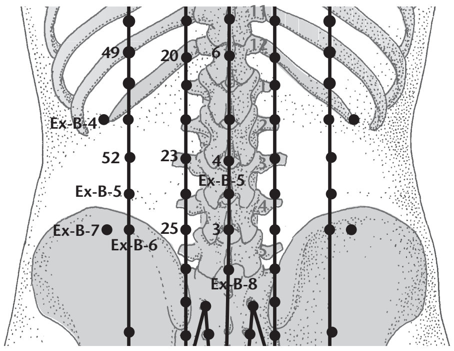
11
49
20
6
12
1
Ex-B-4
2
52
23
4
3
Ex-B-5
Ex-B-5
4
Ex-B-7
25
3
Ex-B-6
5
Ex-B-8

Suspended Pivot XUANSHU **Du-5**

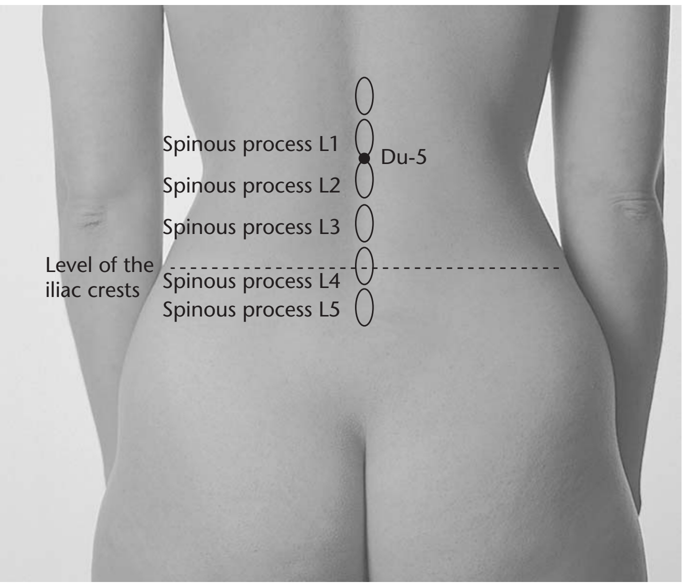
Spinous process L1
Du-5
Spinous process L2
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5

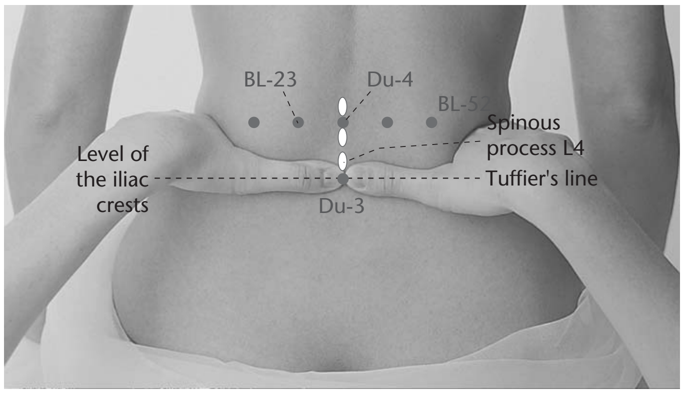
BL-23
Du-4
BL-5
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

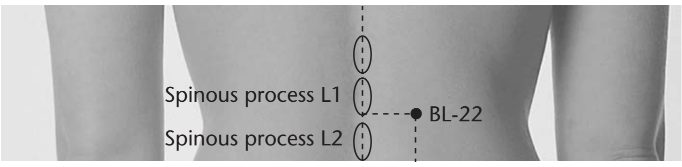
Spinous process L1
BL-22
Spinous process L2

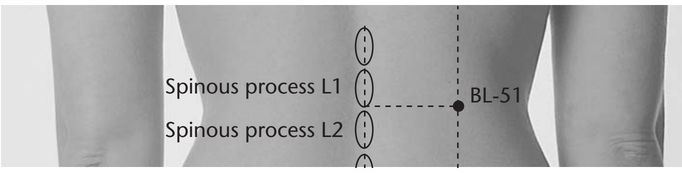
Spinous process L1
BL-51
Spinous process L2

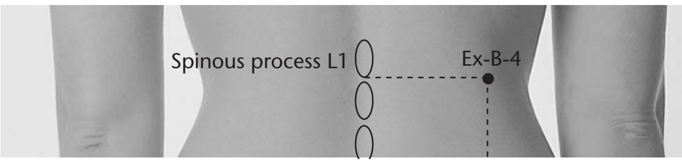
Spinous process L1
Ex-B-4

Location

On the midline, below the spinous process of the 1st lumbar vertebra (L1).

How to find

For orientation in the lumbar region (→ 3.4.3), first, locate the Tuffier's line by placing both hands directly superior and lateral to the two highest points on the iliac crest, joining the thumbs on the midline. The Tuffier's line generally intersects with the spinous process of L4. (Note: Anatomical orientation in the lumbar region is dependent on the patient's position; for more detail on differential orientation → 3.4.) From there, count up to the spinous process of L1 and locate **Du-5** on the midline, below its spinous process. **Or:** Use the spinous process of T7 (inferior angle of the scapula → 3.4.2) for reference.

Located on the same level are a point of → **Ex-B-2/BL-22/ BL-51/Ex-B-4** (0.5/1.5/3/3.5 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Opens the channel
- Regulates Qi in the Lower Burner

Du-6 *Centre of the Spine* JIZHONG

Location

On the midline, below the spinous process of the 11th thoracic vertebra (T11).

How to find

Locate T7 (➞ 3.4.2) and, from there, count down 4 spinous processes to T11. Locate **Du-6** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (➞ 3.4.3). Locate T11 by counting up 5 spinous processes from L4 and there locate **Du-6** on the midline, below the spinous process.
Located on the same level are a point of ➞ **Ex-B-2/BL-20/BL-49** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Strengthens the Spleen and its function of transforming Dampness
- Benefits the spine

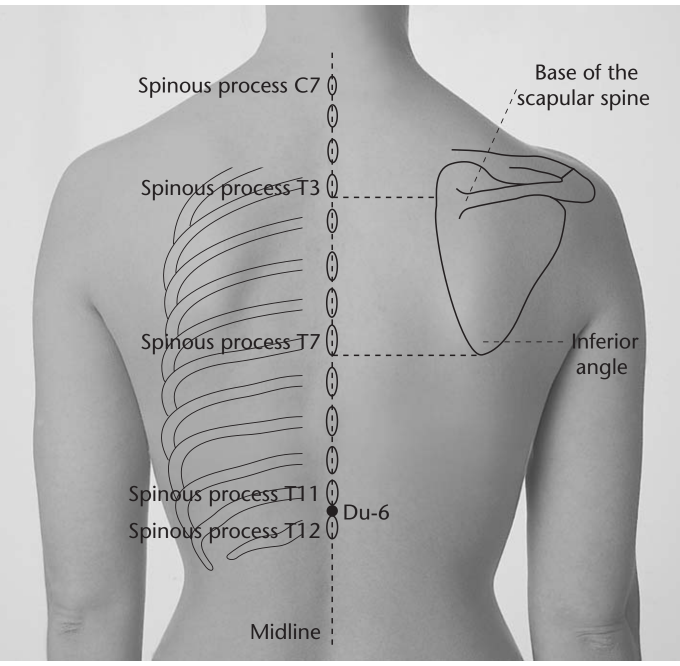
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T11
Du-6
Spinous process T12
Midline

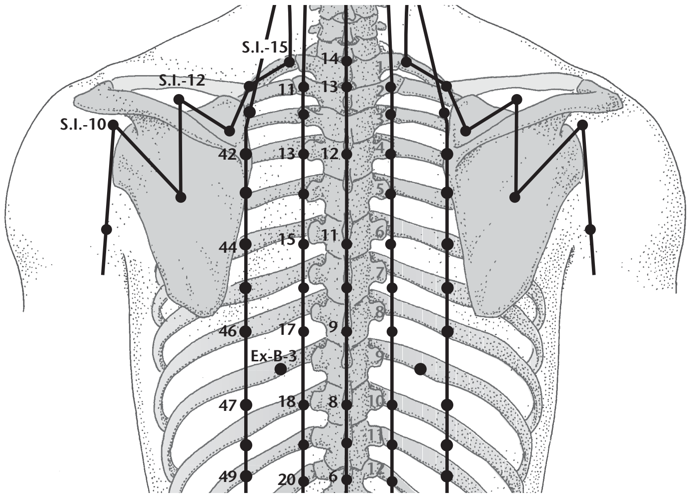
S.I.-15
S.I.-12
S.I.-10
14
13
11
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6

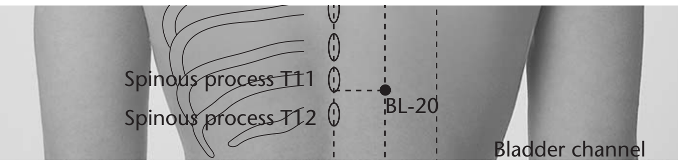
Spinous process T11
Spinous process T12
BL-20
Bladder channel

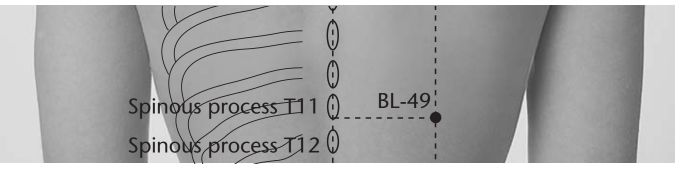
Spinous process T11
BL-49
Spinous process T12

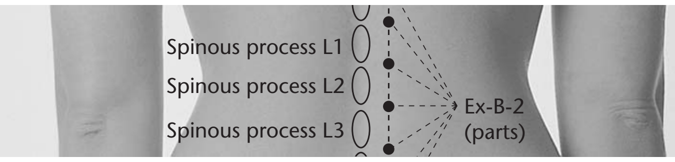
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2
(parts)

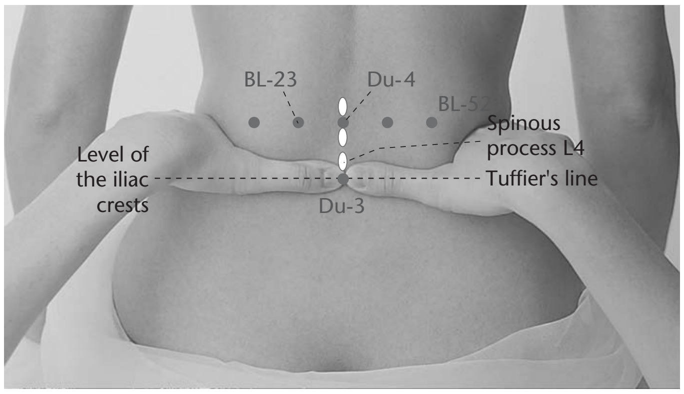
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Central Pivot ZHONGSHU **Du-7**

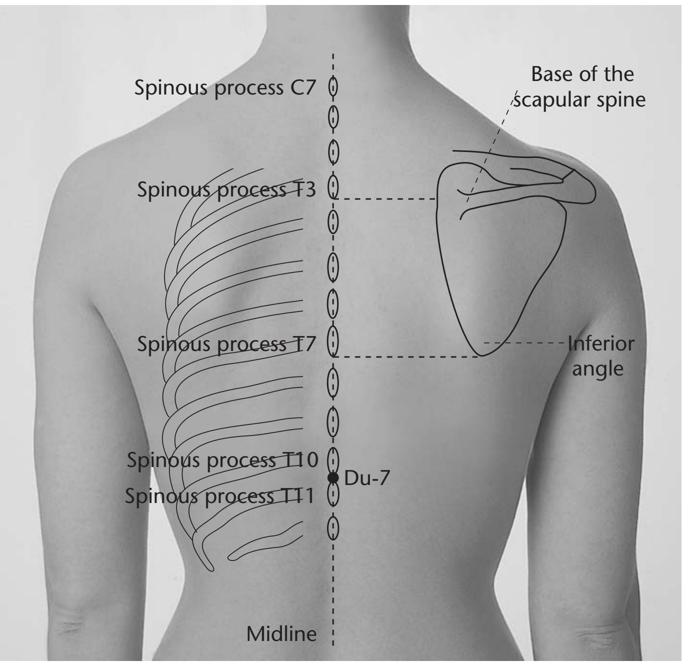
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T10
Du-7
Spinous process T11
Midline

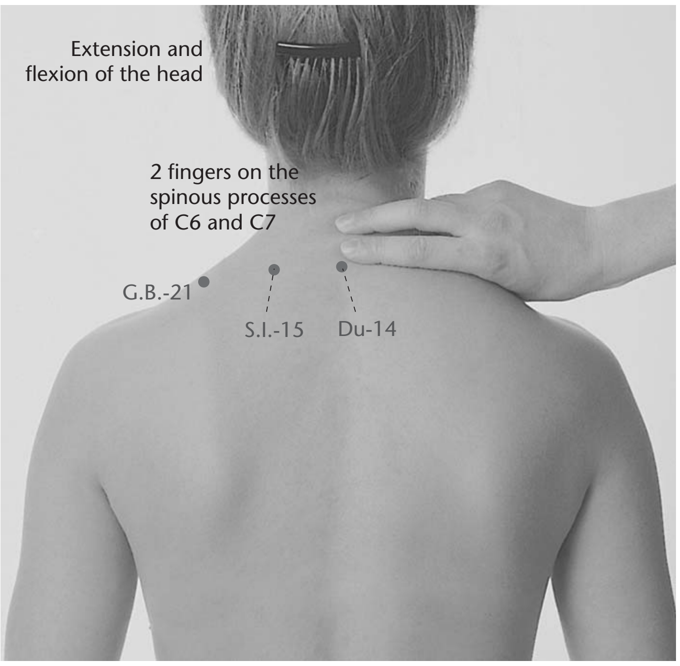
Extension and flexion of the head
2 fingers on the spinous processes of C6 and C7
G.B.-21
S.I.-15
Du-14

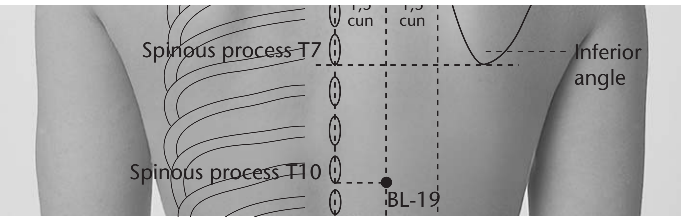
cun
cun
Spinous process T7
Inferior angle
Spinous process T10
BL-19

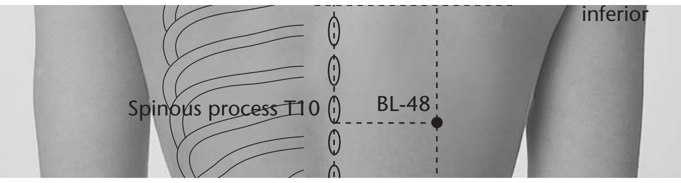
inferior
Spinous process T10
BL-48

Location

On the midline, below the spinous process of the 10th thoracic vertebra (T10).

How to find

Locate T7 (→ 3.4.2) and, from there, count down 3 spinous processes to T10. Locate **Du-7** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (→ 3.4.3). Locate T11 by counting up 5 spinous processes from L4 and there locate **Du-7** on the midline, below the spinous process.

Located on the same level are a point of → **Ex-B-2/BL-19/BL-48** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Tonifies the Middle Burner
- Benefits the spine

Du-8 *Sinew Contraction* JINSUO

Location

On the midline, below the spinous process of the 9th thoracic vertebra (T9).

How to find

Locate T7 (→ 3.4.2) and, from there, count down 2 spinous processes to T9. Locate **Du-8** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (→ 3.4.3). Locate T9 by counting up from L4 and there locate **Du-8** on the midline, below the spinous process.
Located on the same level are a point of → **Ex-B-2/BL-18/BL-47** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Soothes the Liver
- Dispels Wind
- Calms the *shen*

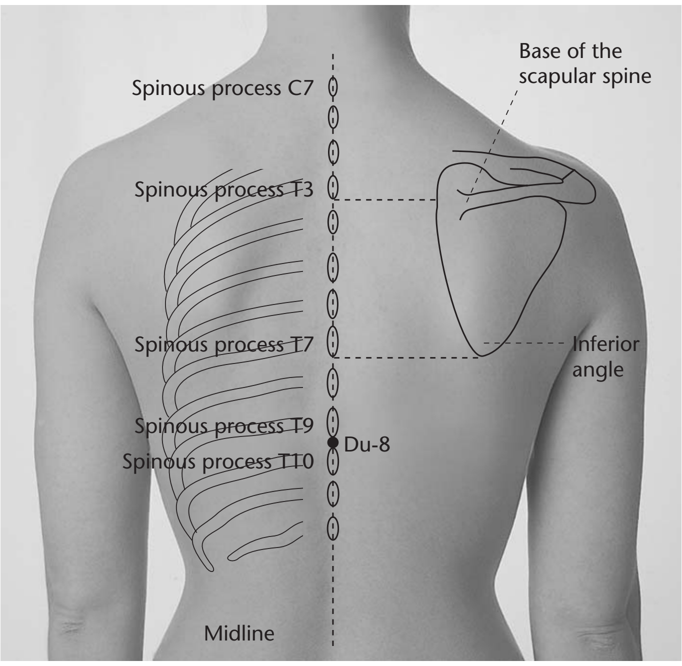
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T9
Du-8
Spinous process T10
Midline

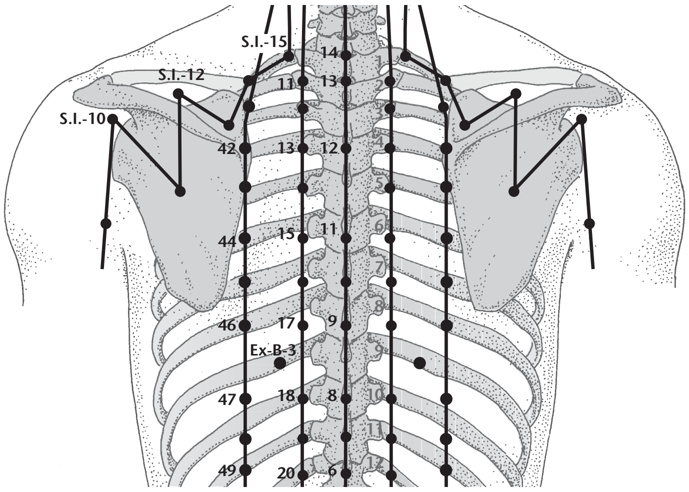
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6
1
2
3
4
5
6
7
8
9
10
11
12

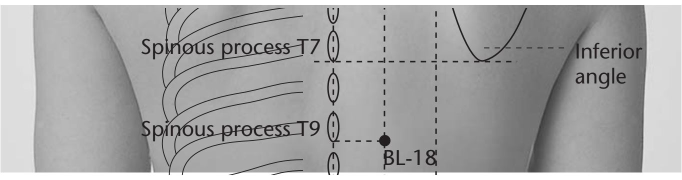
Spinous process T7
Inferior angle
Spinous process T9
BL-18

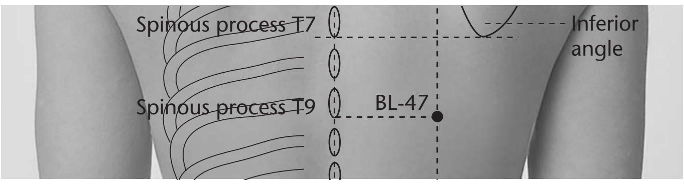
Spinous process T7
Inferior angle
Spinous process T9
BL-47

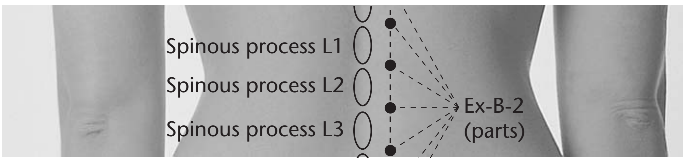
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ch05b-F10028.qxd 2/22/08 9:55 PM Page 507

Reaching Yang ZHIYANG **Du-9**

Location

On the midline, below the spinous process of the 7th thoracic vertebra (T7).

How to find

Find T7 (➞ 3.4.2) and locate **Du-9** on the midline, below its spinous process. **Or:** Use the Tuffier's line for reference (➞ 3.4.3). Locate T7 by counting up from L4 and there locate **Du-9** on the midline, below the spinous process.
Located on the same level are a point of ➞ **Ex-B-2/BL-17/BL-46** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Strengthens the Spleen and eliminates Dampness
- Clears Damp-Heat from the Middle Burner
- Harmonises the Upper Burner

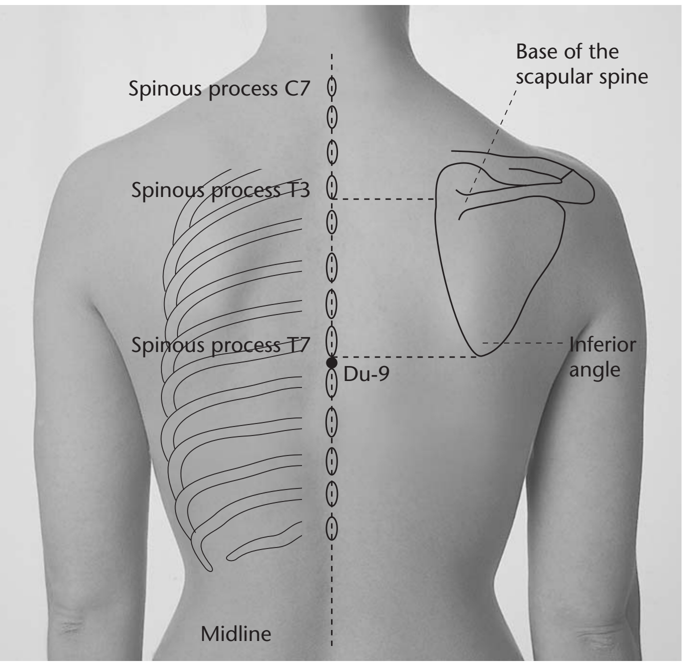
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T7
Du-9
Inferior angle
Midline

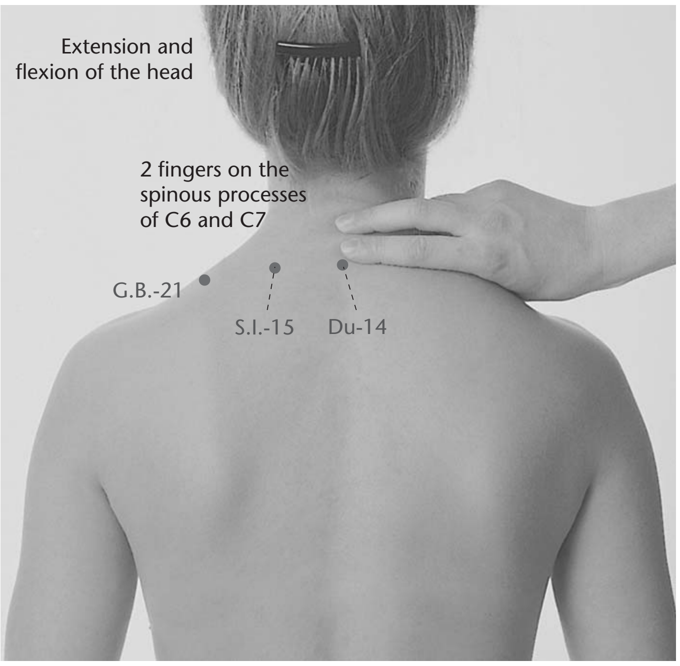
Extension and flexion of the head
2 fingers on the spinous processes of C6 and C7
G.B.-21
S.I.-15
Du-14

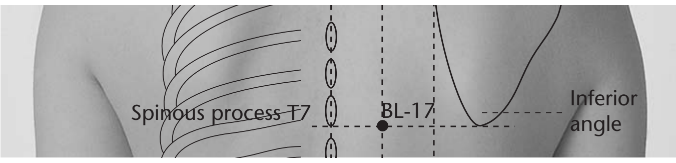
Spinous process T7
BL-17
Inferior angle

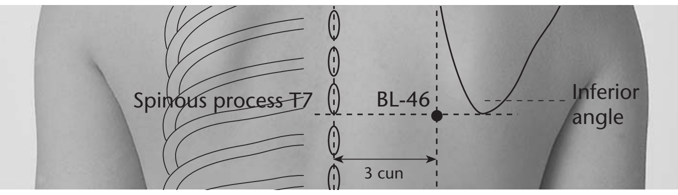
Spinous process T7
BL-46
Inferior angle
3 cun

Du-10 *Spirit Tower* LINGTAI

Location

On the midline, below the spinous process of the 6th thoracic vertebra (T6).

How to find

Locate T7 (→ 3.4.2). **Du-10** is located on the midline, below the spinous process of the T6. **Or:** Locate C7 (→ 3.4.1). **Du-10** can then be located by counting down 6 spinous processes (T1–T6). Located on the same level are a point of → **Ex-B-2/BL-16/ BL-45** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Harmonises the Upper Burner
- Clears Heat and Fire toxins

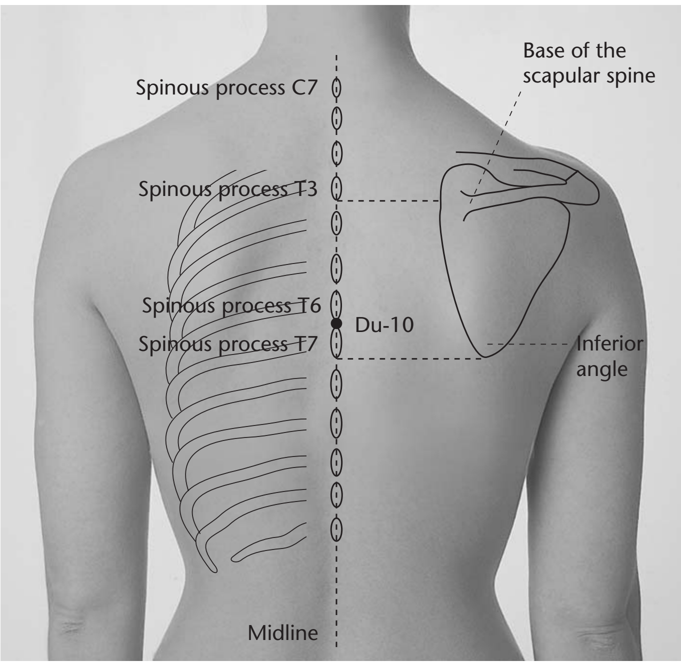
Spinous process C7
Base of the scapular spine
Spinous process T3
Spinous process T6
Du-10
Spinous process T7
Inferior angle
Midline

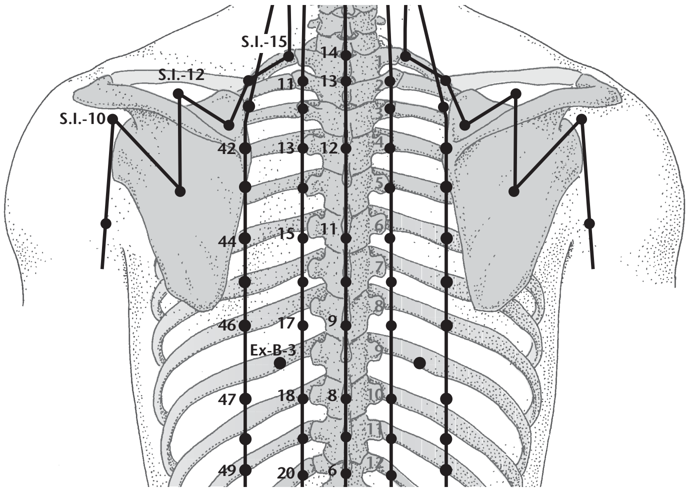
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
5
44
15
11
6
7
46
17
9
Ex-B-3
47
18
8
10
11
49
20
6
12

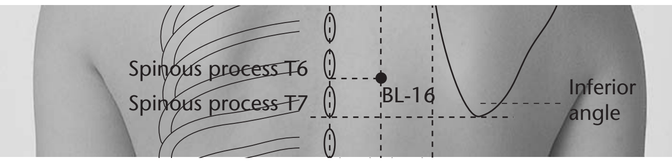
Spinous process T6
Spinous process T7
BL-16
Inferior angle

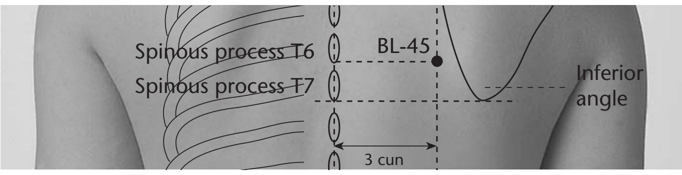
Spinous process T6
Spinous process T7
BL-45
Inferior angle
3 cun

Spirit Pathway SHENDAO Du-11

Location

On the midline, below the spinous process of the 5th thoracic vertebra (T5).

How to find

Locate T7 (➞ 3.4.2). From there, count up 2 spinous processes and find **Du-11** on the midline, below the spinous process of T5.
Or: Locate C7 (➞ 3.4.1). **Du-11** can then be located by counting down 5 spinous processes (T1–T5).
Located on the same level are a point of ➞ **Ex-B-2/BL-15/BL-44** (0.5/1.5/3 cun lateral to the midline).

Needling

0.5–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Tonifies the Heart and the Lungs
- Calms the *shen*
- Dispels Wind and Heat

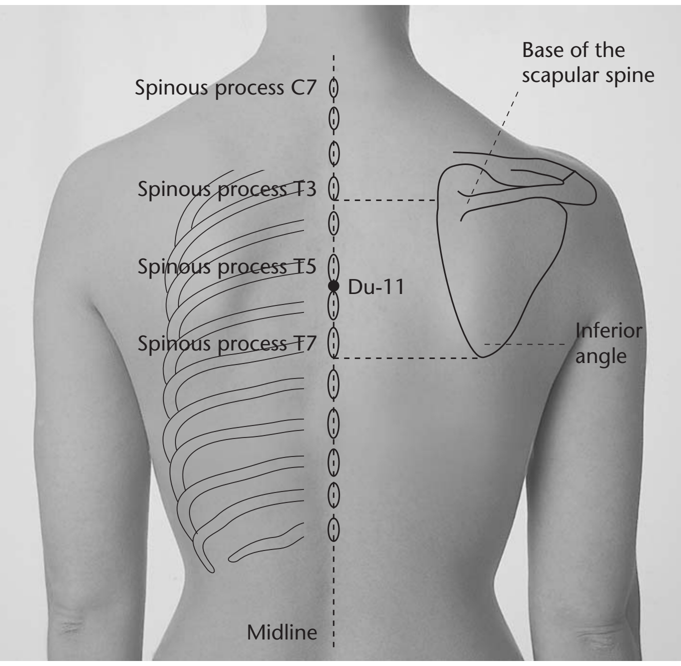
Base of the scapular spine
Spinous process C7
Spinous process T3
Spinous process T5
Du-11
Spinous process T7
Inferior angle
Midline

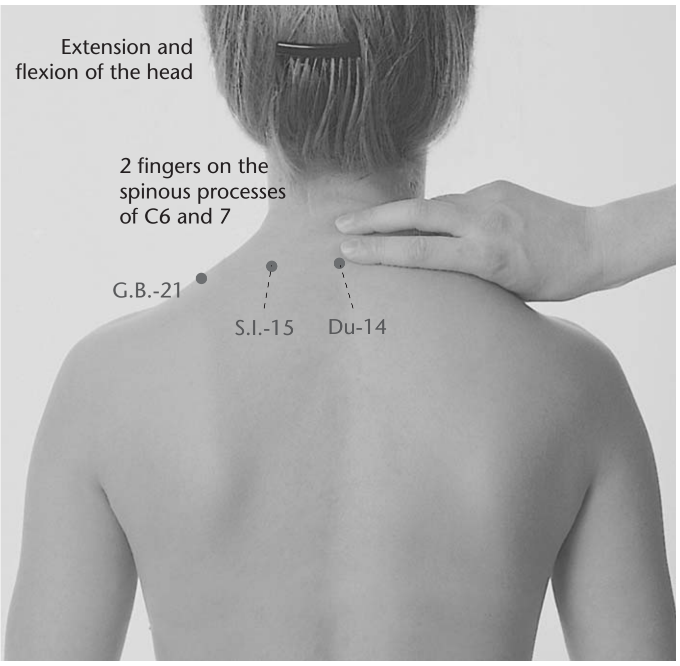
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

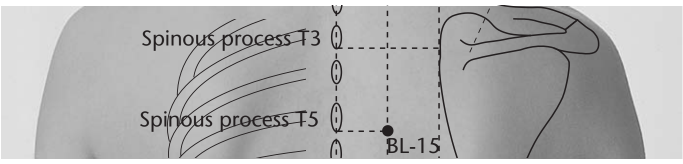
Spinous process T3
Spinous process T5
BL-15

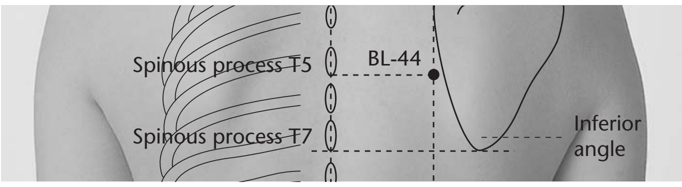
Spinous process T5
BL-44
Spinous process T7
Inferior angle

Du-12 *Body Pillar* SHENZHU

Location

On the midline, below the spinous process of the 3rd thoracic vertebra (T3).

How to find

Locate T7 (→ 3.4.2). From there, count up 4 spinous processes and locate **Du-12** on the midline, below the spinous process of T3. **Or:** Locate C7 (→ 3.4.1). **Du-12** can then be located by counting down 3 spinous processes (T1–T3).
Located on the same level are a point of → **Ex-B-2/BL-13/BL-42** (0.5/1.5/3 cun lateral to the midline).

Needling

0.3–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Dispels internal and external Wind
- Calms the *shen*
- Tonifies the Lungs

Special features

Meeting point with the BL channel

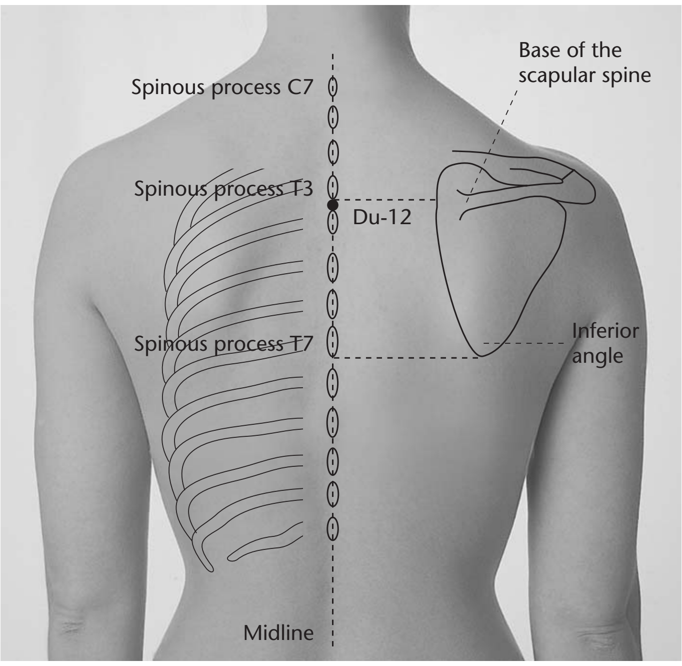
Spinous process C7
Base of the scapular spine
Spinous process T3
Du-12
Spinous process T7
Inferior angle
Midline

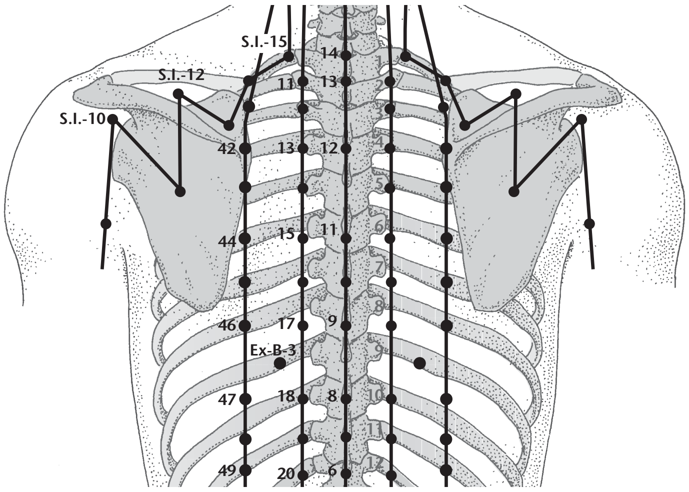
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6

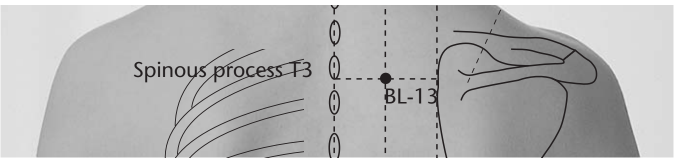
Spinous process T3
BL-13

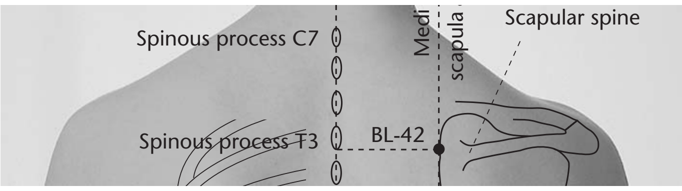
Spinous process C7
Medi
scapula
Scapular spine
Spinous process T3
BL-42

Way of Happiness TAODAO **Du-13**

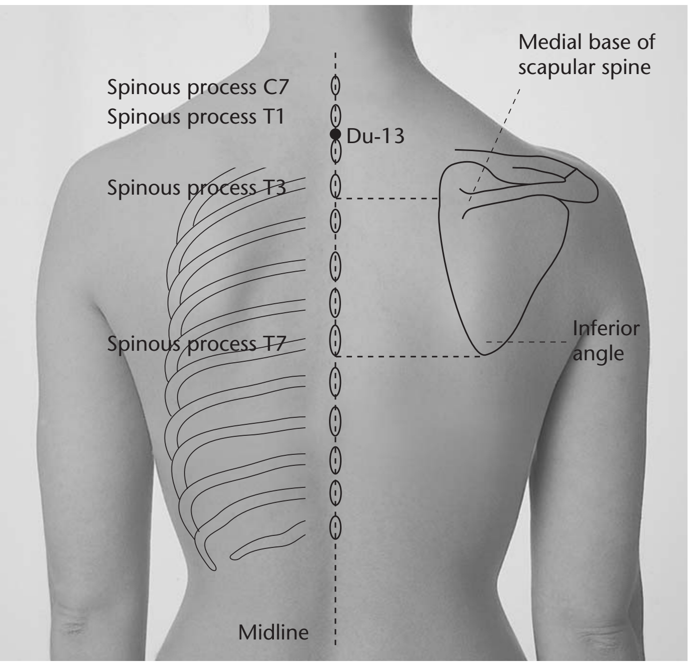
Medial base of scapular spine
Spinous process C7
Spinous process T1
Du-13
Spinous process T3
Spinous process T7
Inferior angle
Midline

Location

On the midline, below the spinous process of the 1st thoracic vertebra (T1).

How to find

Locate C7 (→ 3.4.2). Palpate downward to the next spinous process (T1) and locate **Du-13** on the midline, below its spinous process.

Located on the same level are a point of → **Ex-B-2/BL-11/S.I.-14** (0.5/1.5/3 cun lateral to the midline).

Needling

0.3–1 cun strictly perpendicularly to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun.

Actions/Indications

- Dispels pathogenic factors
- Opens the channel and *luo*-connecting vessels

Special features

Meeting point with the BL channel

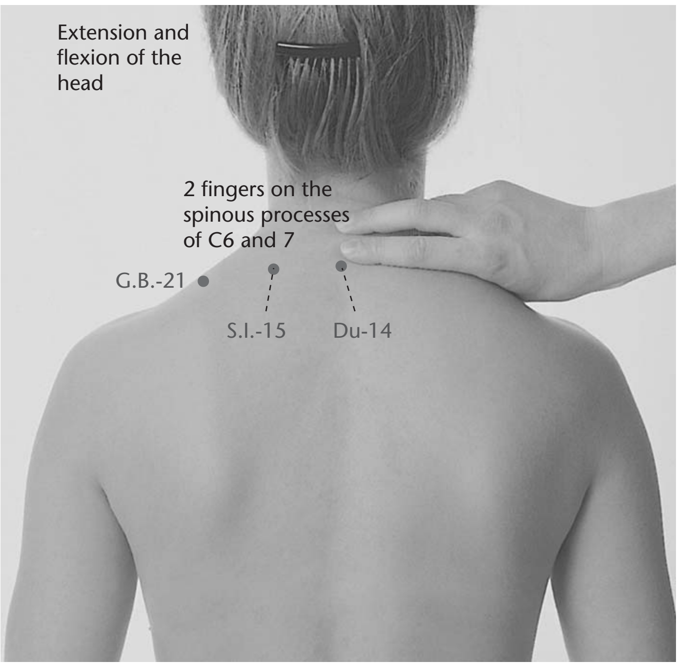
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

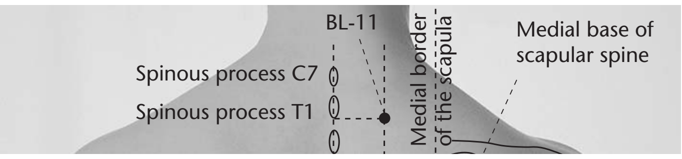
BL-11
Medial border of the scapula
Medial base of scapular spine
Spinous process C7
Spinous process T1

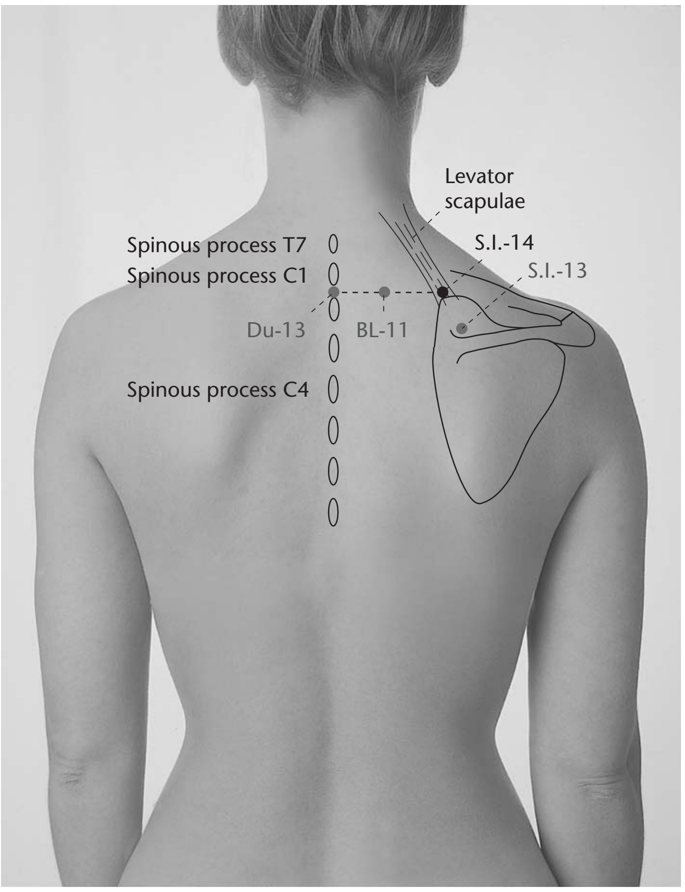
Levator scapulae
Spinous process T7
Spinous process C1
S.I.-14
S.I.-13
Du-13
BL-11
Spinous process C4

Du-14 *Great Vertebra* DAZHUI

Location

On the midline, below the spinous process of the 7th cervical vertebra (C7).

How to find

First, identify C7 (→ 3.4.1): Place 2 fingers on the spinous processes believed to belong to C6 and C7 and ask the patient to flex and extend their head. In a fully functional spine and with correct finger placement, C6 will glide anteriorly with neck extension, while C7 will remain fixed. However, if the vertebra below the upper finger does not move with neck extension, the fingers will most probably have been placed on C7 and T1. Locate **Du-14** below the spinous process of C7.

Located on the same level are → **Ex-B-1** (*dingchuan*)/ **S.I.-15/Ex-B** (*jiehexue*) (0.5/2/3.5 cun lateral to **Du-14**).

Needling

0.3–1 cun strictly vertically to the skin or obliquely in an inferior direction, with the patient's back straight or flexed in order to avoid puncturing the spinal canal. Oblique insertions should only be carried out by experienced practitioners, as in small persons (irrespective of their body weight), the spinal canal may be reached after only 1.25 cun. For persistent epistaxis, place a cold metal object (for example, a spoon) or a cold cloth on **Du-14**. Cupping on this point for acute colds.

Actions/Indications

- Clears Wind, firms and regulates the surface, dispels pathogenic factors
- Clears Heat
- Calms (internal) Wind and the *shen*
- Strengthens Qi and Yang in deficiency conditions
- Benefits the spine, especially the cervical spine

Special features

Meeting point with all Yang channels, Sea of Qi point.

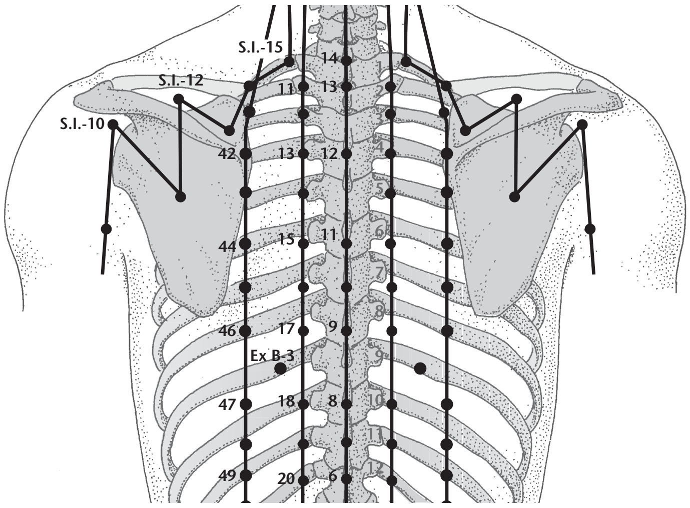
S.I.-15
S.I.-12
S.I.-10
14
11
13
42
13
12
44
15
11
46
17
9
Ex B-3
47
18
8
49
20
6

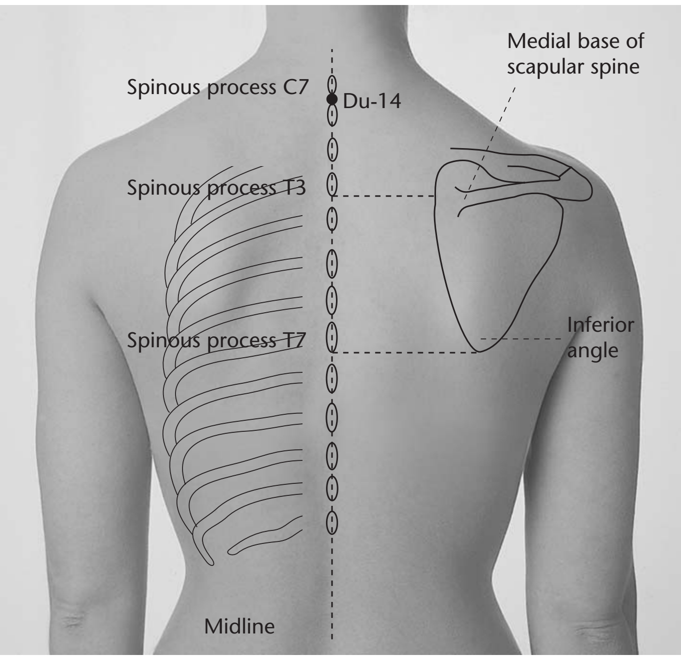
Medial base of scapular spine
Spinous process C7
Du-14
Spinous process T3
Spinous process T7
Inferior angle
Midline

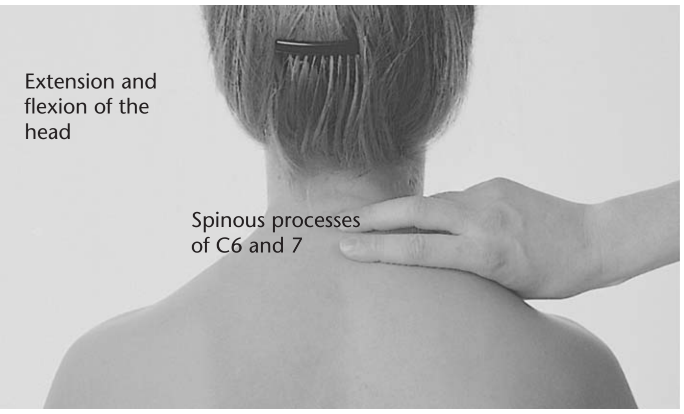
Extension and flexion of the head
Spinous processes of C6 and 7

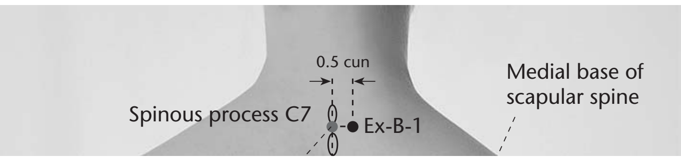
0.5 cun
Medial base of scapular spine
Spinous process C7
Ex-B-1

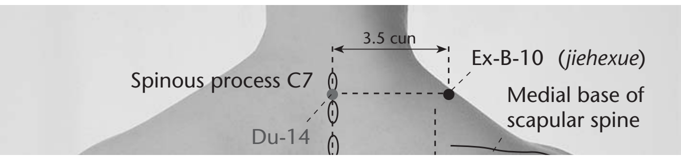
3.5 cun
Ex-B-10 (jiehexue)
Spinous process C7
Medial base of scapular spine
Du-14

Ch05b-F10028.qxd 2/22/08 9:55 PM Page 513

Gate of Muteness YAMEN **Du-15**

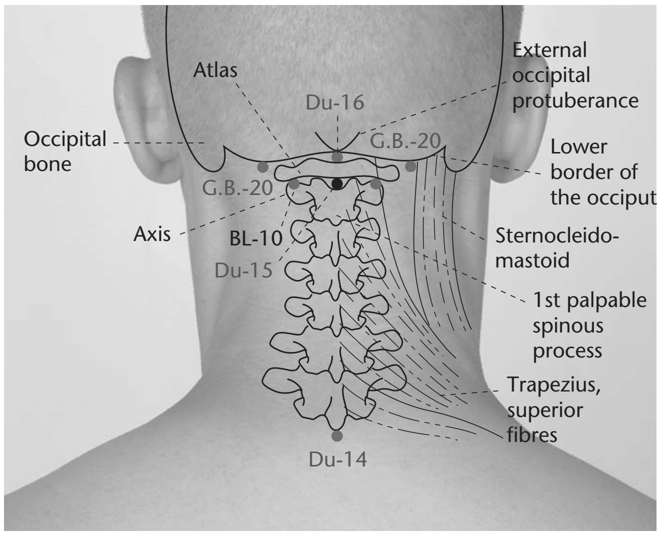
Atlas
External occipital protuberance
Du-16
G.B.-20
Occipital bone
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

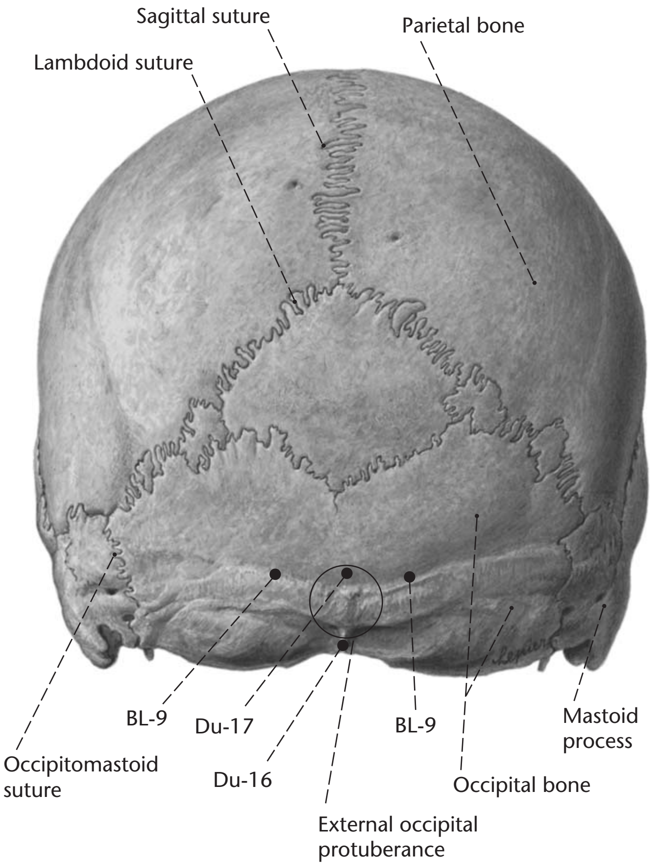
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

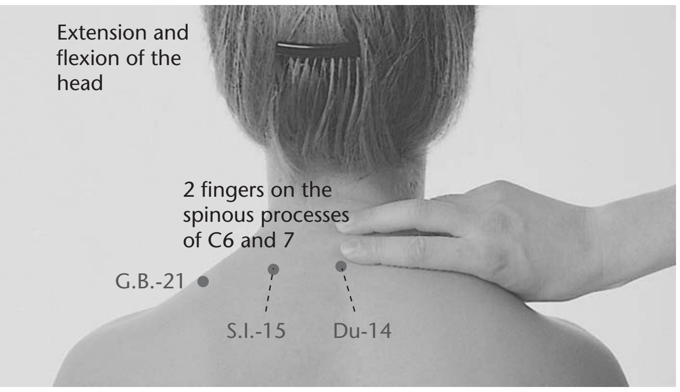
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

On the occiput, on the posterior midline, in the depression between the 1st (atlas) and 2nd (axis) cervical vertebrae, approximately 0.5 cun inferior to → **Du-16** (directly below the external occipital protuberance).

How to find

The external occipital protuberance (→ 3.1.5) is a hump-shaped projection on the posterior surface of the occipital bone, on the midline and slightly superior to the craniocervical junction (→ 3.1.5). → **Du-16** is located in a depression on the posterior midline that you can palpate directly inferior to the protuberance. From there, glide 0.5 cun in an inferior direction and locate **Du-15** superior to the first palpable spinous process (belonging to the axis; the atlas has no spinous process) and approximately 0.5 cun superior to the posteror hairline.
Located on the same level is → **BL-10** (on the lateral aspect of the trapezius muscle).

Needling

0.5–1 cun strictly perpendicularly in an inferior direction. Caution: Injury to the cervical marrow. According to classic texts, moxibustion is contraindicated.

Actions/Indications

- Benefits the tongue, ears, neck and spine
- Dispels Wind
- Clears Yang Heat/excess

Special features

Meeting point with the *yang wei mai*. Sea of Qi point.

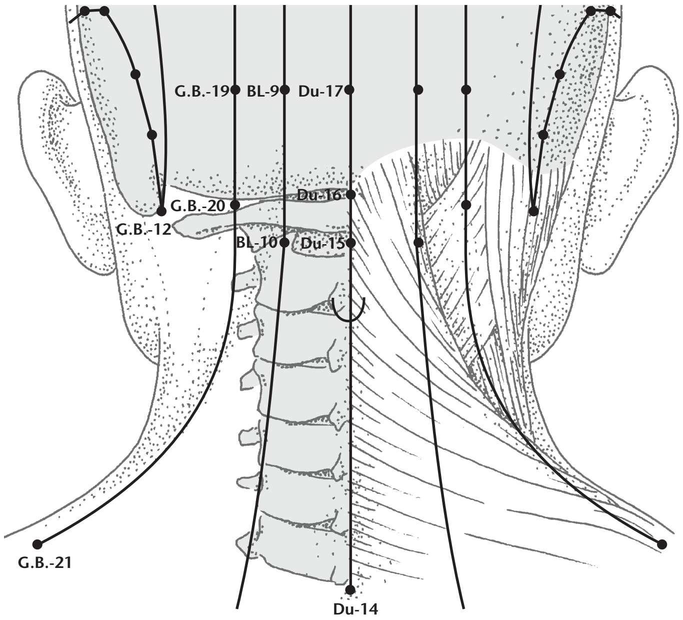
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

Du-16 *Palace of Wind* FENGFU

Location

On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius muscle.

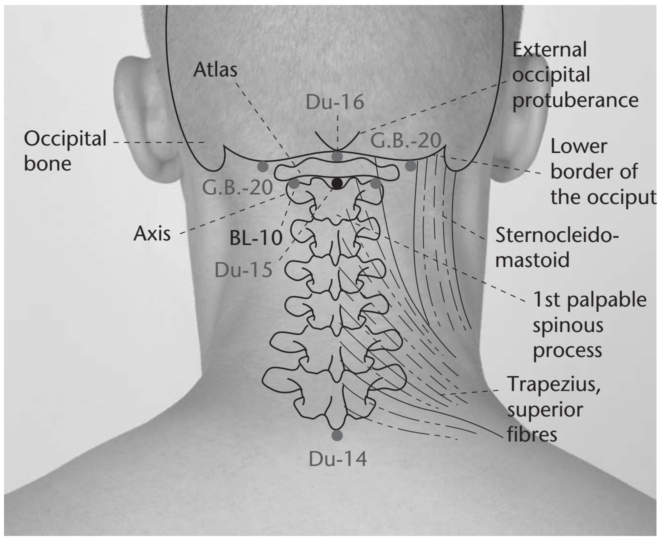
Atlas
External occipital protuberance
Du-16
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid
Du-15
1st palpable spinous process
Trapezius, superior fibres
Du-14

How to find

The external occipital protuberance (→ 3.1.5) is a hump-shaped projection on the posterior surface of the occipital bone, on the midline and slightly superior to the craniocervical junction (→ 3.1.5). Locate **Du-16** in a depression on the posterior midline that can be palpated directly inferior to the protuberance and between the two origins of the trapezius muscle. Generally, **Du-16** is located approximately 1 cun superior to the posterior hairline. → **G.B.-20** is located on the same level below the occiput, in a depression between the origins of the trapezius and sternocleidomastoid muscles.

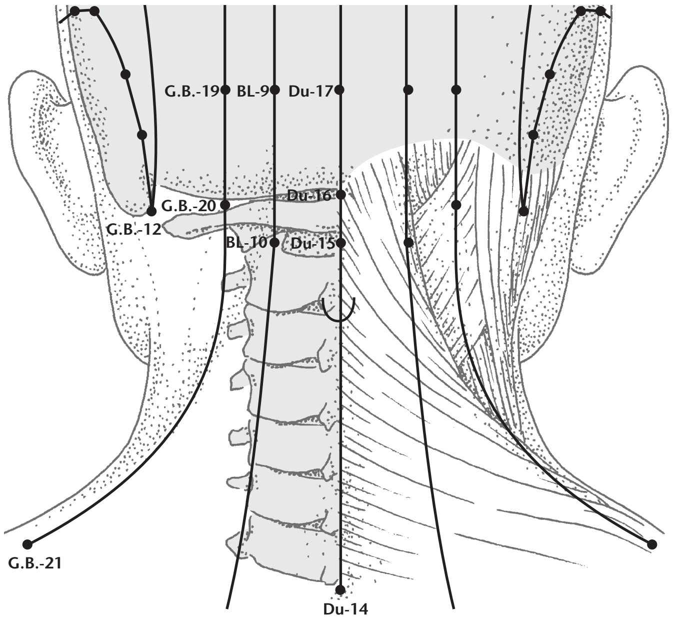
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

Needling

0.5–1 cun slightly obliquely in an inferior direction. Bend the patient's head slightly forward when needling. Caution: Dangerous point with deep needling! No stimulation. Strictly no needling in a superior direction, danger of puncturing the cerebellomedullary cistern (also site for suboccipital puncture). The needle should be inserted into the nuchal ligament. According to some authors, moxibustion is contraindicated.

Actions/Indications

- Dispels (external) Wind
- Calms internal Wind and the *shen*
- Nourishes the Sea of Marrow, lightens the *shen*

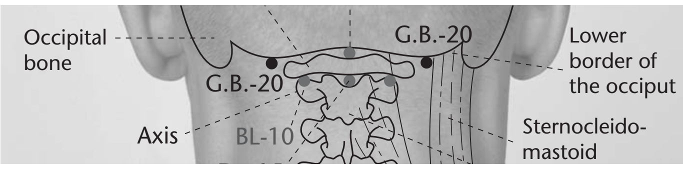
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid

Special features

Meeting point with the *yang wei mai* and, according to some authors, the *yang qiao mai*. Sea of Marrow point, Window of Heaven point, Sun Si Miao Ghost point. Used for all disorders which are caused by 'external or internal' Wind and which are characterised by sudden onset, changing location and/or intensity of the disorder, involuntary movements.

Brain's Door **NAOHU** **Du-17**

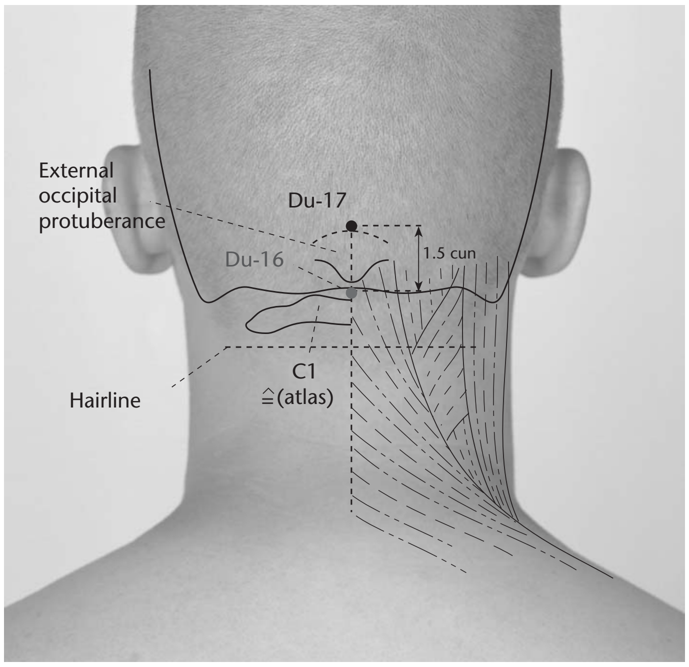
External
occipital
protuberance
Du-17
Du-16
1.5 cun
C1
≙(atlas)
Hairline

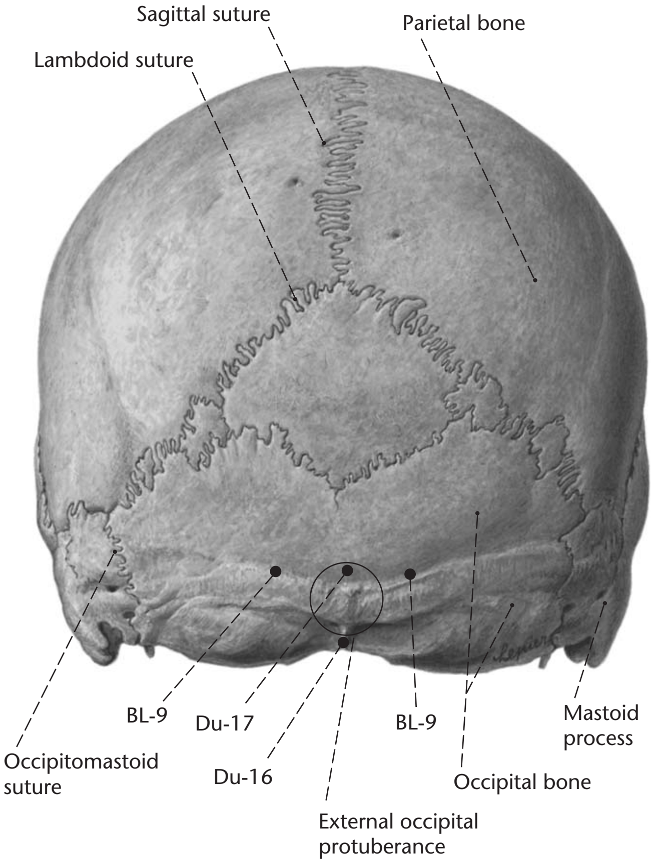
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid
process
Occipitomastoid
suture
Du-16
Occipital bone
External occipital
protuberance

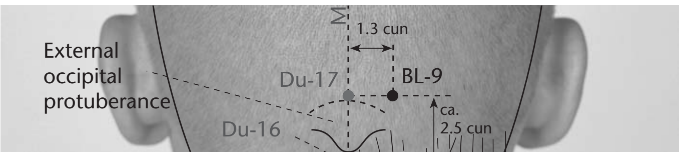
M
1.3 cun
External
occipital
protuberance
Du-17
BL-9
ca.
2.5 cun
Du-16

Location

In a depression superior to the external occipital protuberance, approximately 2.5 cun superior to the posterior hairline or 1.5 cun superior to ➞ **Du-16**.

How to find

First, locate the external occipital protuberance (➞ 3.1.5), which forms a hump-shaped projection on the posterior surface of the occipital bone. Next, locate **Du-17** on the posterior midline, in a depression directly superior to the protuberance. This corresponds to a distance of approximately 2.5 cun superior to the posterior hairline (➞ 3.1.5).

➞ **Du-16** is located directly inferior to the protuberance, whereas ➞ **BL-9** is located 1.3 cun lateral to **Du-17**.

Needling

0.5–1 cun transversely (subcutaneously) in an inferior or superior direction. Moxibustion controversial.

Actions/Indications

- Dispels Wind
- Opens the channel
- Benefits the eyes
- Calms the *shen*

Special features

Meeting point with the BL channel

Du-18 *Unyielding Space* QIANGJIAN

Location

On the posterior midline, 1.5 cun superior to **Du-17** (directly superior to the external occipital protuberance) or 3 cun inferior to → **Du-20** (on the vertex).

How to find

First, locate → **Du-17** in a depression directly superior to the external occipital protuberance (→ 3.1.5), which forms a hump-shaped projection on the posterior surface of the occipital bone. From there, palpate 1.5 cun in a superior direction and there locate **Du-18** on the posterior midline.
→ **Du-20** is located 3 cun in a superior direction, on the vertex.

Needling

Transversely (subcutaneously) 0.5–1 cun

Actions/Indications

- Dispels Wind, especially internal Wind
- Opens the channel
- Calms the *shen*

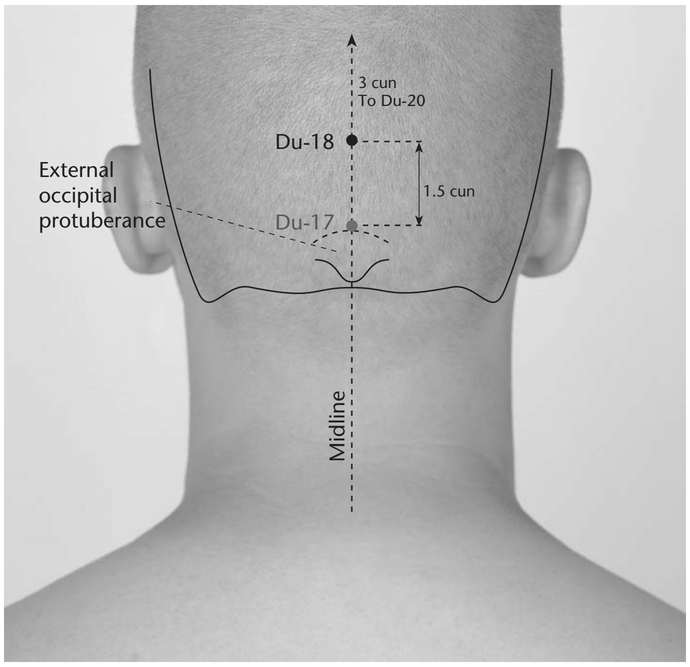
3 cun
To Du-20
Du-18
1.5 cun
External occipital protuberance
Du-17
Midline

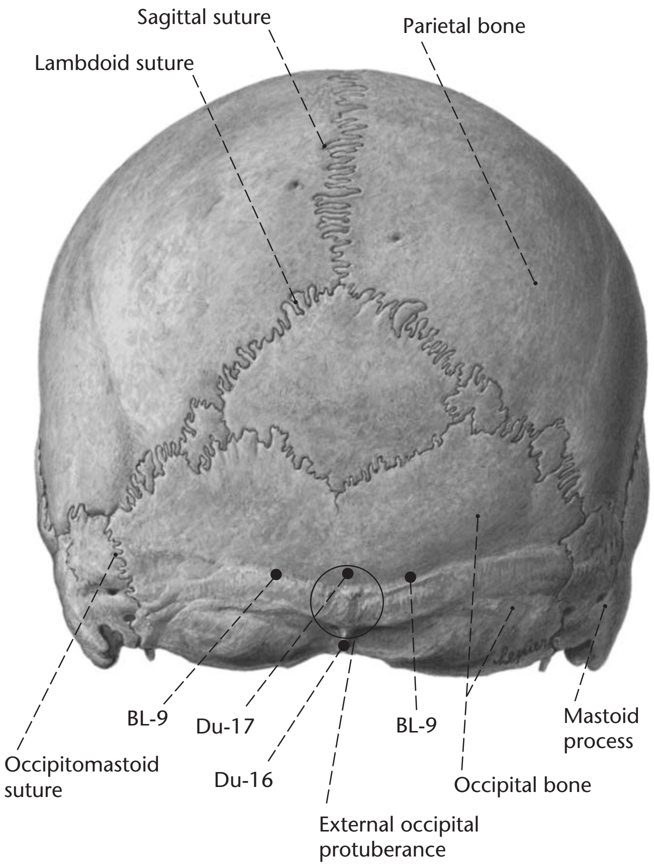
Sagittal suture
Parietal bone
Lambdoid suture
BL-9
Du-17
BL-9
Mastoid process
Occipitomastoid suture
Du-16
Occipital bone
External occipital protuberance

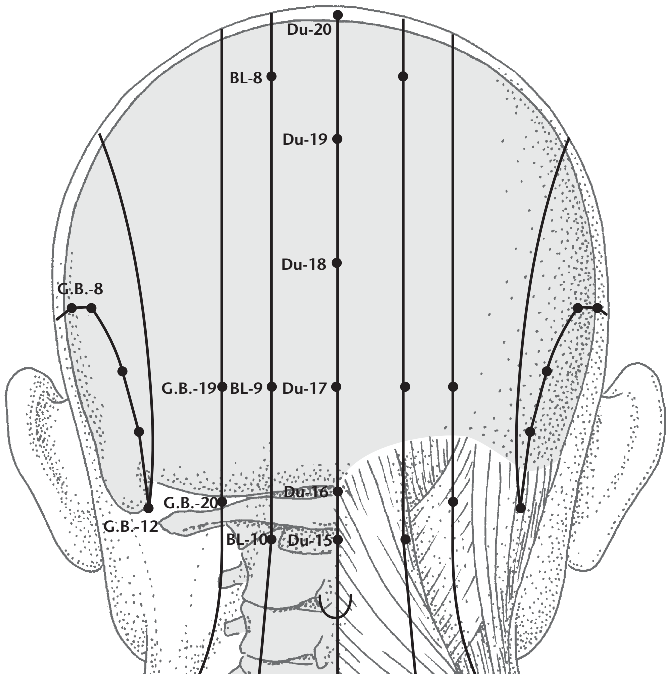
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15

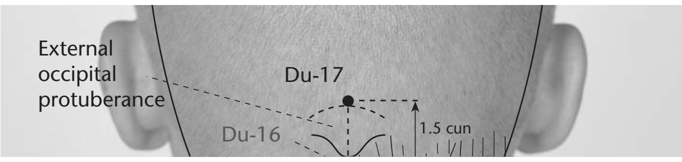
External occipital protuberance
Du-17
Du-16
1.5 cun

Behind the Crown HOUDING Du-19

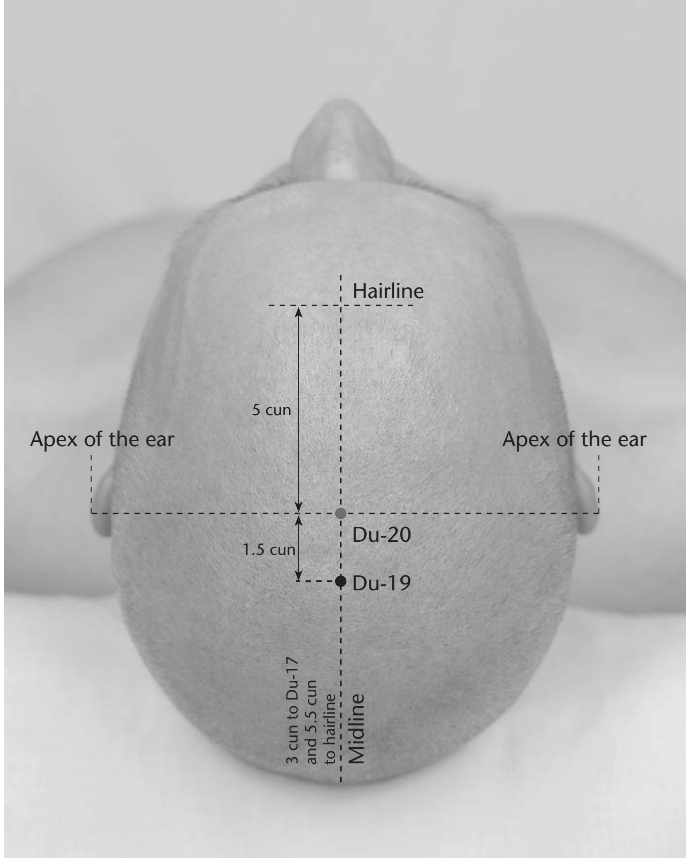
Hairline
5 cun
Apex of the ear
Apex of the ear
Du-20
1.5 cun
Du-19
3 cun to Du-17
and 5.5 cun
to hairline
Midline

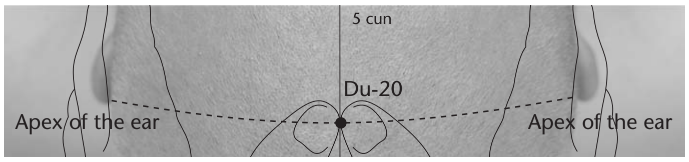
5 cun
Du-20
Apex of the ear
Apex of the ear

Location

On the midline, 3 cun superior to → **Du-17** (directly superior to the external occipital protuberance) or 1.5 cun posterior to → **Du-20**.

How to find

This point is best located by using → **Du-20** as a reference point (at the junction of the vertical midline and a line connecting the apices of the ears). **Du-19** can then be located by palpating 1.5 cun in a posterior direction from → **Du-20**. **Or:** First, locate → **Du-17** superior the external occipital protuberance (→ 3.1.5), then locate **Du-19** on the midline, 3 cun superior to → **Du-17**.

Needling

Transversely (subcutaneously) 0.5–1 cun.

Actions/Indications

- Dispels Wind
- Opens the channel
- Calms the *shen*

Du-20 *Hundred Meetings* BAIHUI

Location

At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively.

How to find

Spreading hands technique (→ 2.3.3): Place the hands on both sides of the head, with the little fingers touching the apices of the ears. Join the thumbs on the midline and locate **Du-20** in shallow depression on the vertex of the head (alternative location method: use an elastic tape, with the midpoint marked on it). The symbolic counterpart to **Du-20** (as the highest point on the body and in contact with heaven) is → **KID-1**, the lowest part of the body and in contact with the earth.

Needling

0.5–1 cun transversely (subcutaneously) towards the posterior (reducing effect) or the anterior (tonifying effect). Caution: Apply pressure to head points after needle removal to avoid bleeding. Remember to remove the needle after treatment.

Actions/Indications

- Calms Wind, pacifies Yang, benefits the Brain and sensory organs, calms the *shen*
- Nourishes the Sea of Marrow; applying moxibustion raises the Yang

Special features

Meeting point with the BL, G.B., T.B. and LIV channels; Sea of Marrow point. Important point for descending the Yang (reducing needle techniques) or raising the Yang (tonifying needle techniques, moxibustion).

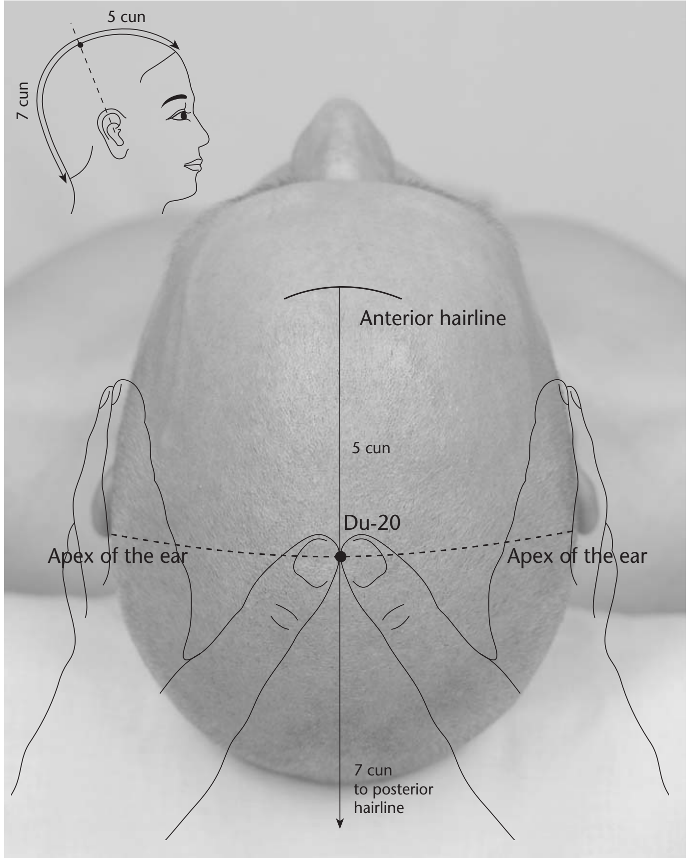
5 cun
7 cun
Anterior hairline
5 cun
Du-20
Apex of the ear
Apex of the ear
7 cun
to posterior
hairline

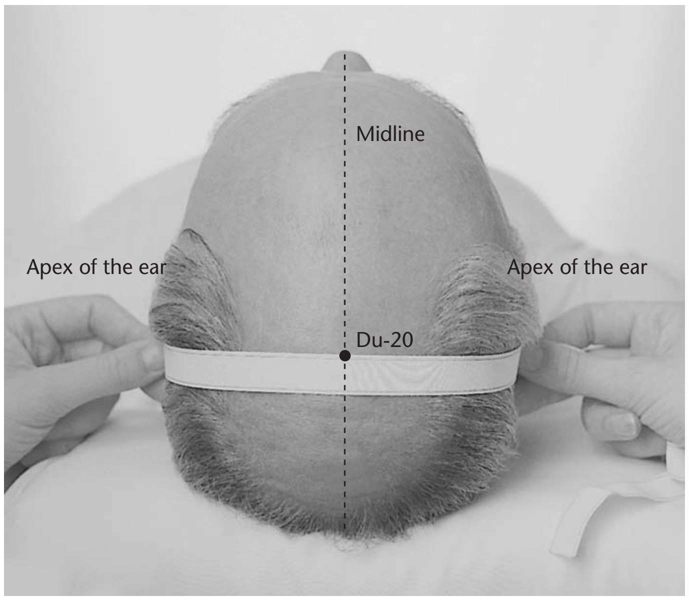
Midline
Apex of the ear
Apex of the ear
Du-20

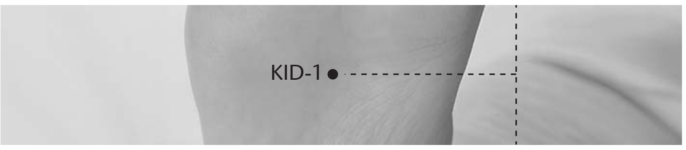

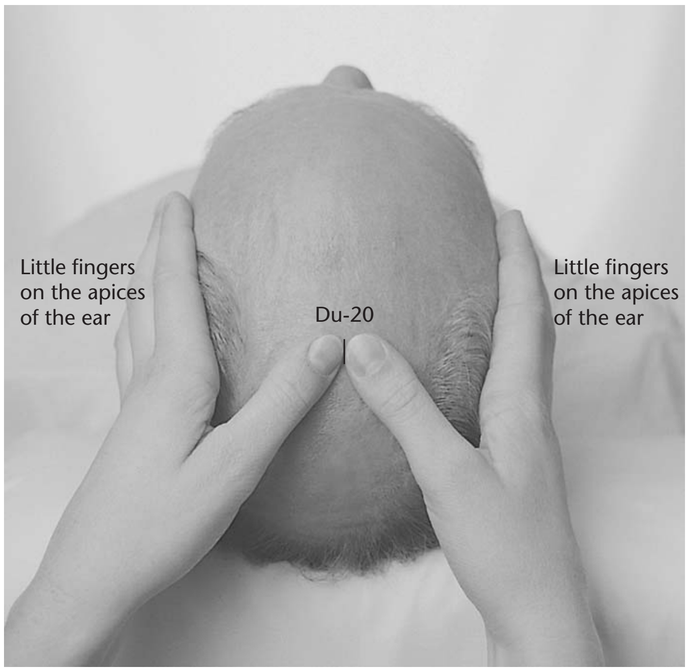
Little fingers
on the apices
of the ear
Du-20
Little fingers
on the apices
of the ear

In Front of the Crown QIANDING **Du-21**

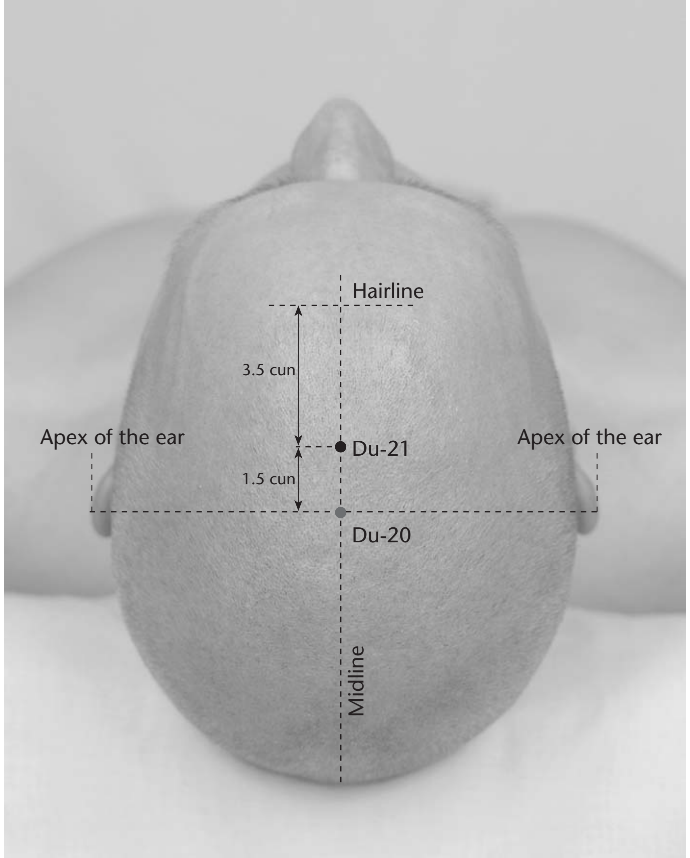
Hairline
3.5 cun
Apex of the ear
Du-21
Apex of the ear
1.5 cun
Du-20
Midline

Location

On the midline, 3.5 cun superior to the anterior hairline or 1.5 cun anterior to ➞ **Du-20**.

How to find

First, locate ➞ **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) and, from there, measure 1.5 cun in an anterior direction. **Or:** Locate the anterior hairline (➞ 3.1.1; the distance from the anterior hairline to ➞ **Du-20** is 5 cun) and palpate 3.5 cun in a superior/posterior direction. There, locate **Du-21** on the midline.

Needling

Transversely (subcutaneously) 0.5–1 cun. Caution with infants whose fontanelle has not yet closed. Moxibustion possible.

Actions/Indications

- Dispels Wind and Dampness
- Opens the channel
- Calms the *shen*

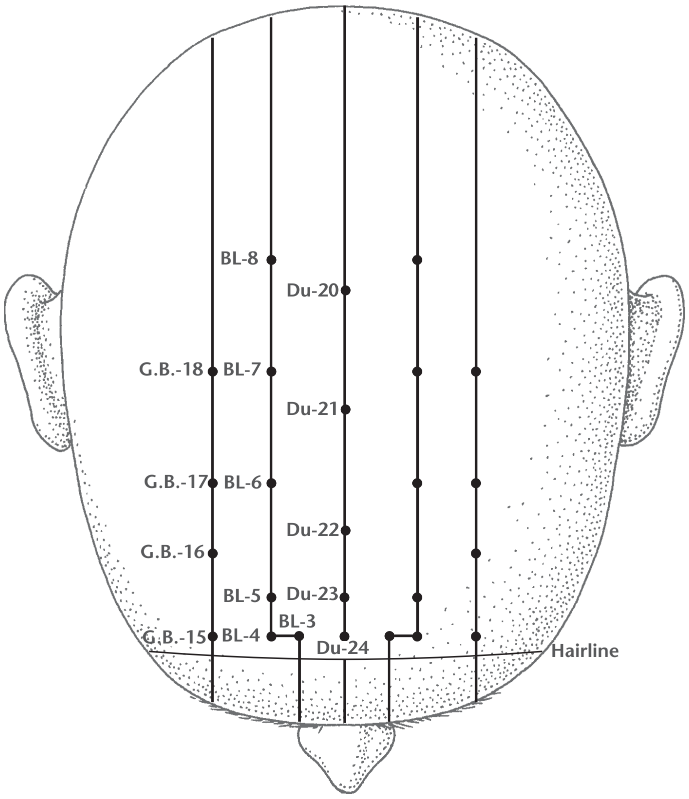
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

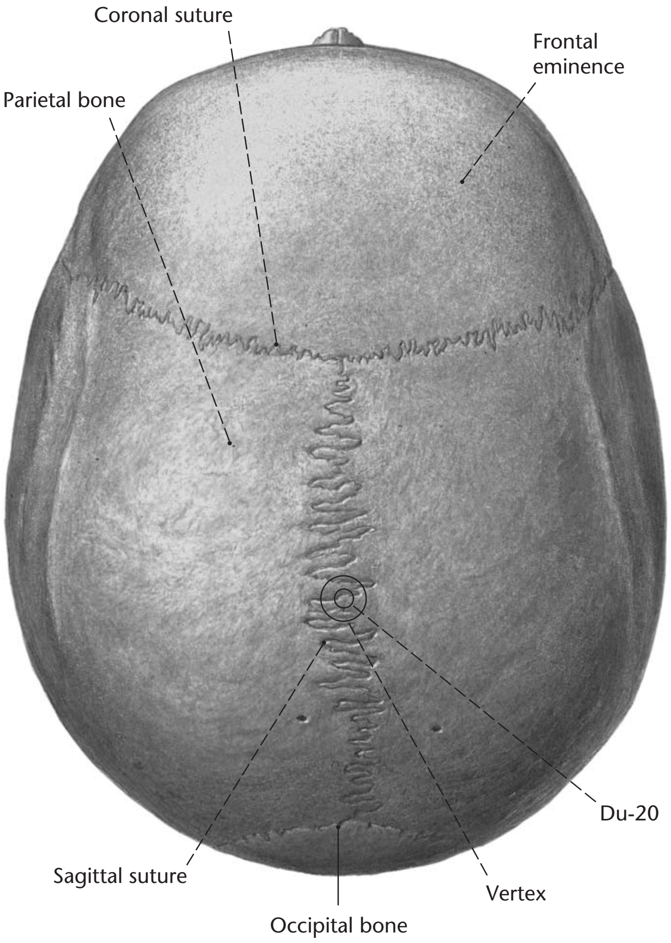
Coronal suture
Frontal eminence
Parietal bone
Du-20
Sagittal suture
Vertex
Occipital bone

Du-22 *Fontanelle Meeting* XINHUI

Location

On the midline, 2 cun superior to the anterior hairline.

How to find

First, locate the anterior hairline (→ 3.1.1) and, from there, palpate 2 cun in a superior direction. There, locate **Du-22** on the midline. For reference: The distance from the anterior hairline to → **Du-20** is 5 cun. **Or:** From → **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears), measure 3 cun in an anterior direction and there locate **Du-22** on the midline.

Needling

Transversely (subcutaneously) 0.5–1 cun. Caution with infants whose fontanelle has not yet closed (**Du-22** is located on the anterior border of the fontanelle). Moxibustion possible.

Actions/Indications

- Dispels Wind
- Calms the *shen*

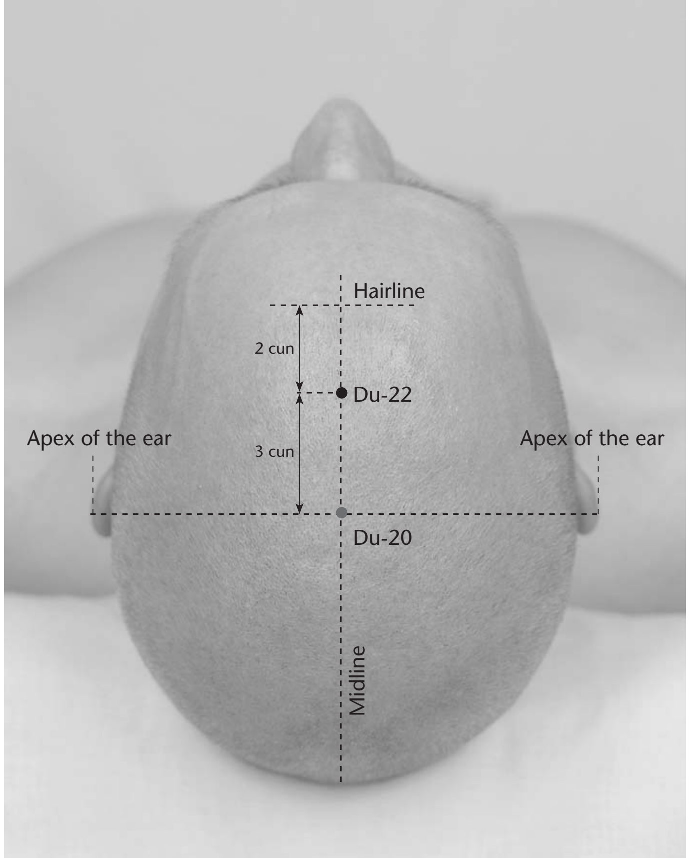
Hairline
2 cun
Du-22
3 cun
Apex of the ear
Apex of the ear
Du-20
Midline

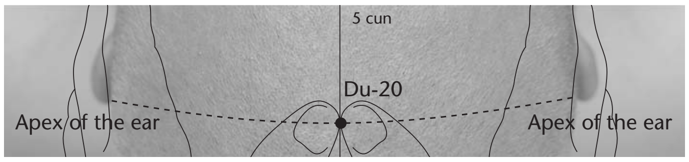
5 cun
Du-20
Apex of the ear
Apex of the ear

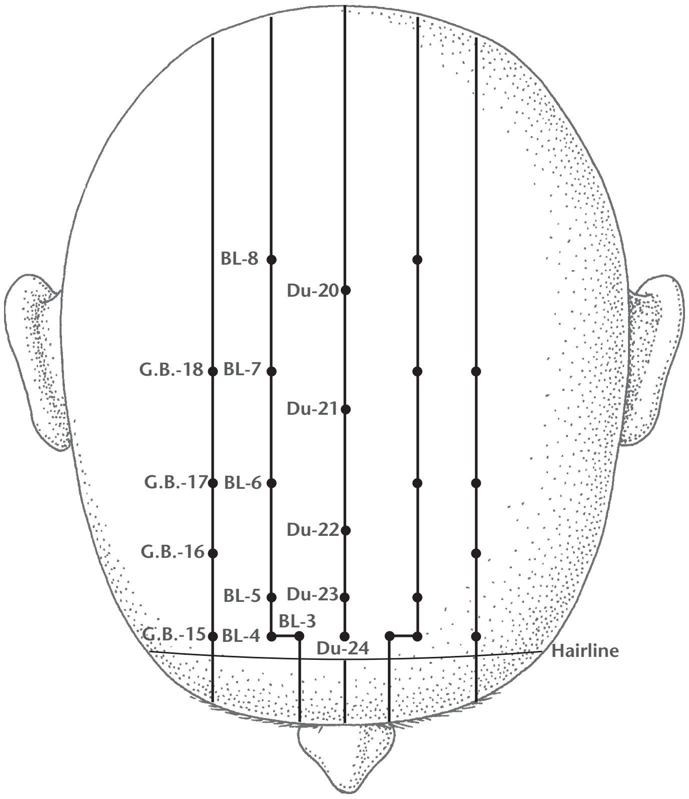
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

Upper Star SHANGXING Du-23

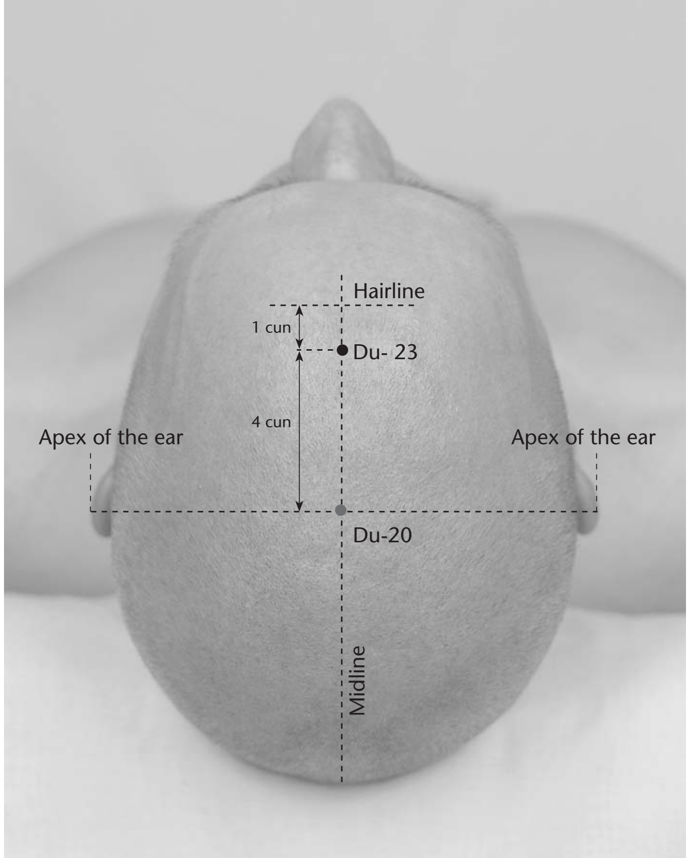
Hairline
1 cun
Du- 23
4 cun
Apex of the ear
Apex of the ear
Du-20
Midline

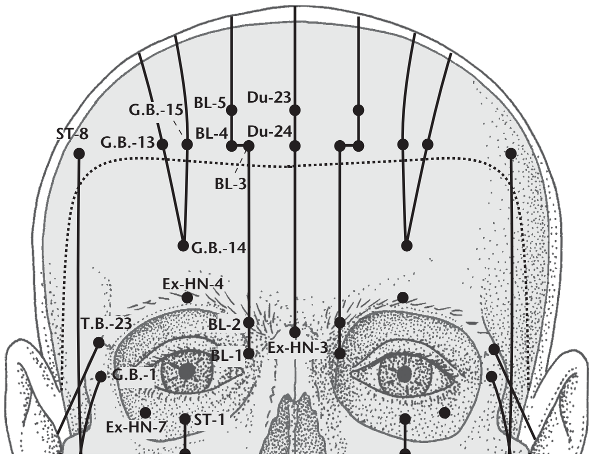
ST-8
G.B.-15
G.B.-13
BL-5
BL-4
Du-23
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1

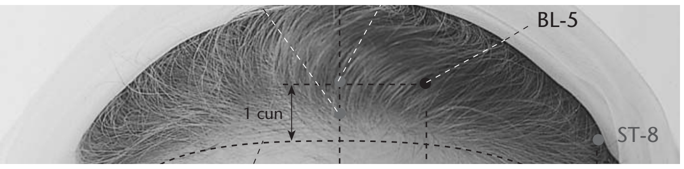
BL-5
1 cun
ST-8

Location

On the midline, 1 cun superior to the anterior hairline or 4 cun anterior to → **Du-20**.

How to find

First, locate the anterior hairline (→ 3.1.1) and, from there, palpate 1 cun in a superior direction. There, locate **Du-23** on the midline. For reference: The distance from the anterior hairline to → **Du-20** is 5 cun.

Located on the same level is → **BL-5** (1.5 cun lateral to the midline).

Needling

Transversely (subcutaneously) 0.3–0.5 cun. Moxibustion possible.

Actions/Indications

- Dispels pathogenic factors (especially Wind and Heat) from the nose
- Dispels pathogenic factors (especially Heat) from the face
- Calms the *shen*

Du-24 *Courtyard of the Spirit* SHENTING

Location

On the midline, 0.5 cun superior to the anterior hairline or 4.5 cun anterior to ➞ **Du-20**.

How to find

First, locate the anterior hairline (➞ 3.1.1) and, from there, palpate 0.5 cun in a superior direction. There, locate **Du-24** on the midline. For reference: The distance from the anterior hairline to ➞ **Du-20** (at the junction of the vertical midline and a line connecting the apices of the ears) is 5 cun.

Located on the same level (0.5 cun superior to the anterior hairline) are ➞ **BL-3/BL-4/G.B.-15/G.B.-13** (superior to the inner canthus of the eye/1.5 cun lateral to the midline/on the pupil line or 2.25 cun lateral to the midline/3 cun lateral to the midline). These distances refer to the proportional cun distance (➞ 2.2) between **Du-24** and ➞ **ST-8** (on the corner of the forehead), which equals 4.5 cun (1.5 cun lateral to the midline).

Needling

Transversely (subcutaneously) 0.3–0.5 cun. Moxibustion possible.

Actions/Indications

- Calms the *shen*
- Dispels (internal) Wind
- Clears pathogenic factors from the eyes and nose

Special features

Meeting point with the BL and ST channels. One of the major points for calming the *shen*.

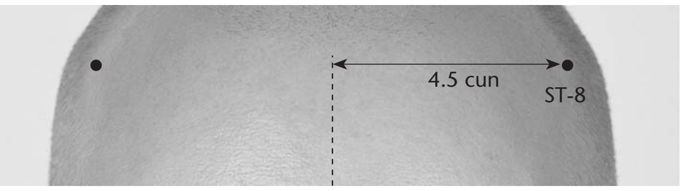

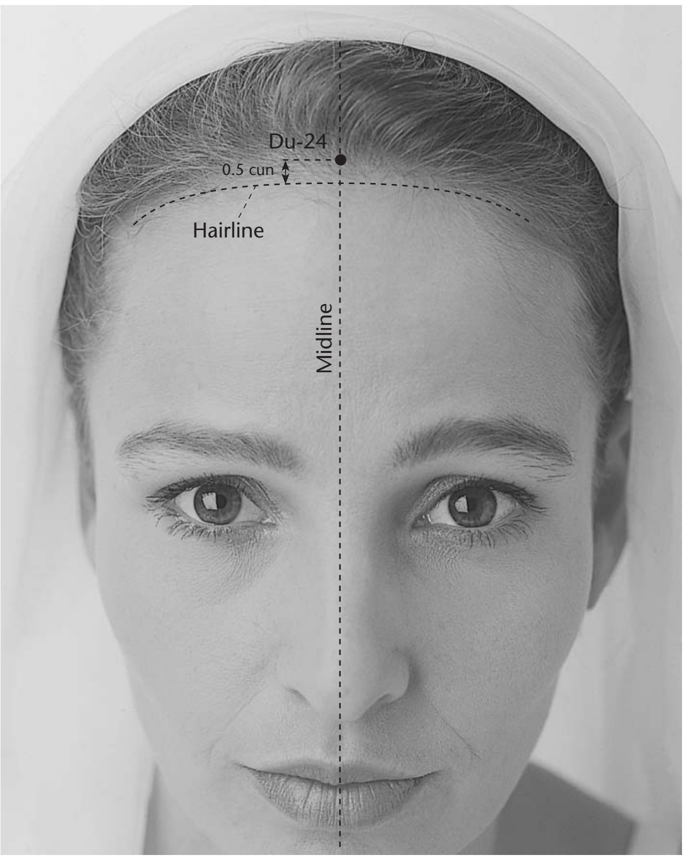
Du-24
0.5 cun
Hairline
Midline

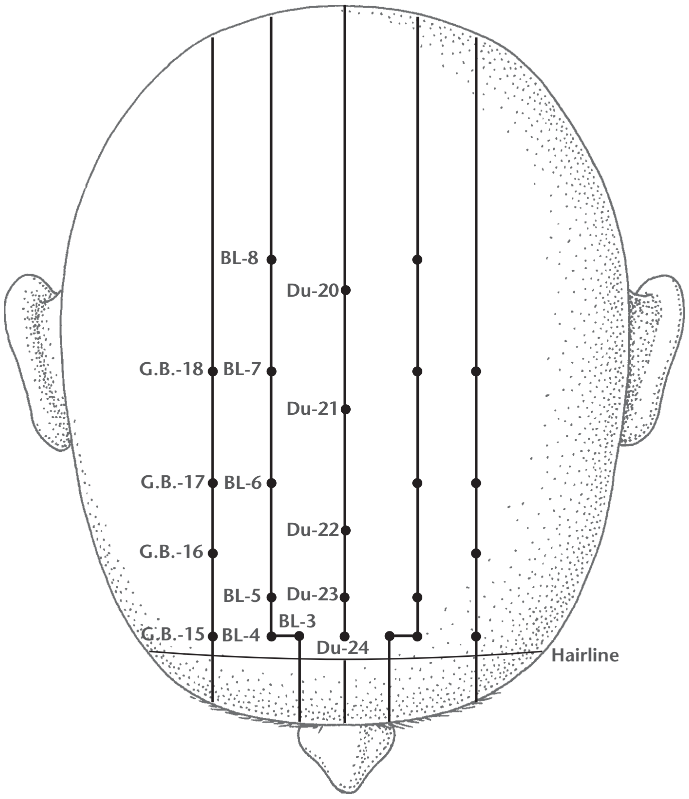
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

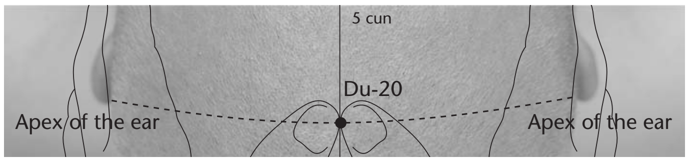
5 cun
Du-20
Apex of the ear
Apex of the ear

White Crevice SULIAO Du-25

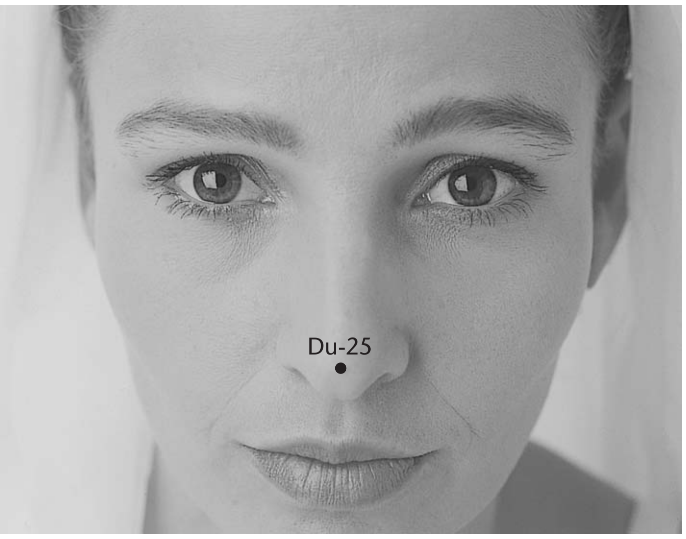

Location

In a depression on the tip of the nose.

How to find

As the name implies, **Du-25** can be located in a depression on the tip of the nose, which generally can be easily palpated.

Needling

0.2 cun vertically or transversely (subcutaneously) in a superior direction up to 1 cun. Or prick to bleed. Moxibustion controversial.

Actions/Indications

- Benefits the nose

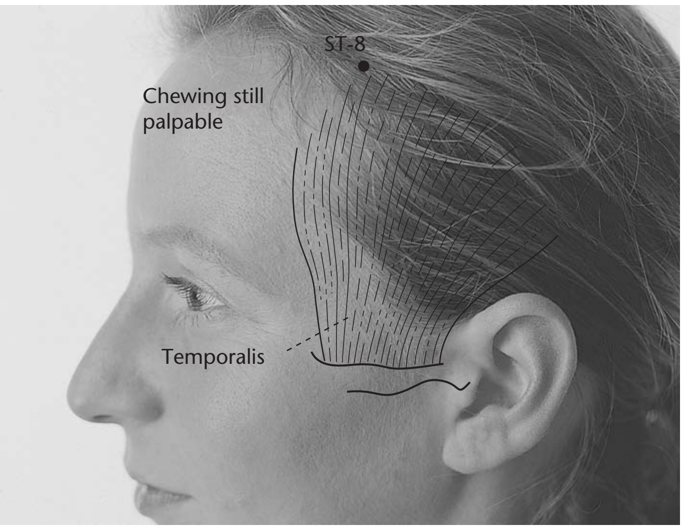
ST-8
Chewing still palpable
Temporalis

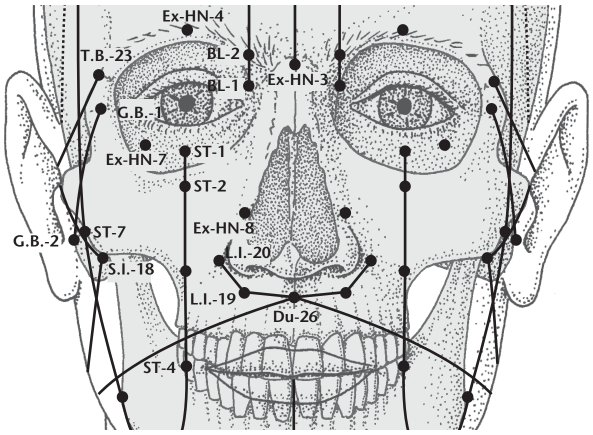
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4

Du-26 *Man's Middle* RENZHONG

Location

Below the nose, on the upper third of the philtrum. Note: *shuigou* (Water Grave) is an alternative name for this point.

How to find

Locate the philtrum, which forms a distinct groove on the midline between the root of the nose and the margin of the upper lip. Locate **Du-26** sligthly superior to the midpoint of the philtrum. → **L.I.-19** is located on the same level, 0.5 cun lateral to the midline.

Needling

0.3–0.5 cun obliquely in a superior direction. Caution: Needling this point may be painful.

Actions/Indications

- Revives consciousness
- Benefits the face and nose, eliminates (external) Wind
- Benefits the spine

Special features

Meeting point with the L.I. and ST channels; Sun Si Miao Ghost point. Major point for acute emergencies: Needle with strongly reducing technique. If no acupuncture needle is available, use a syringe or perform strong acupressure with the nail of the thumb or index finger, supporting the patient's chin with the same hand. Can be used in cases of needle collapse – first remove all other needles.

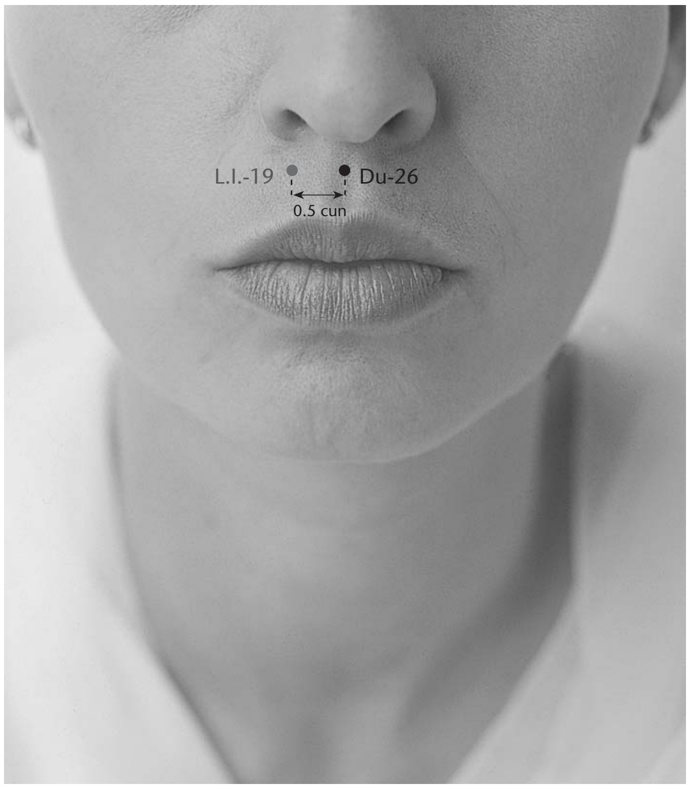

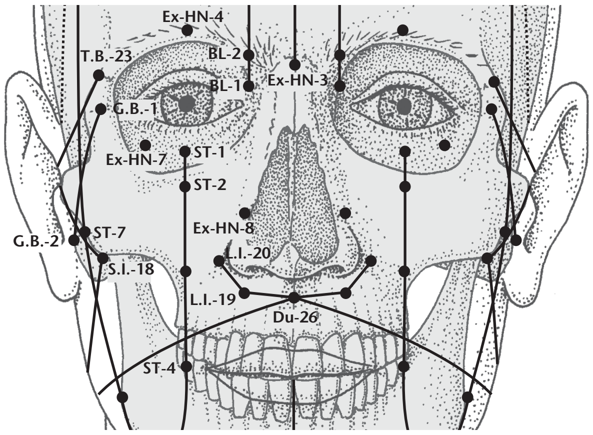

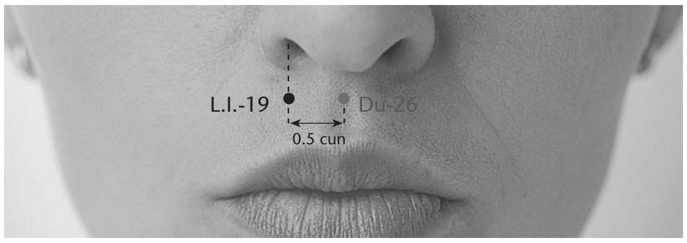

Extremity of the Mouth DUIDUAN Du-27

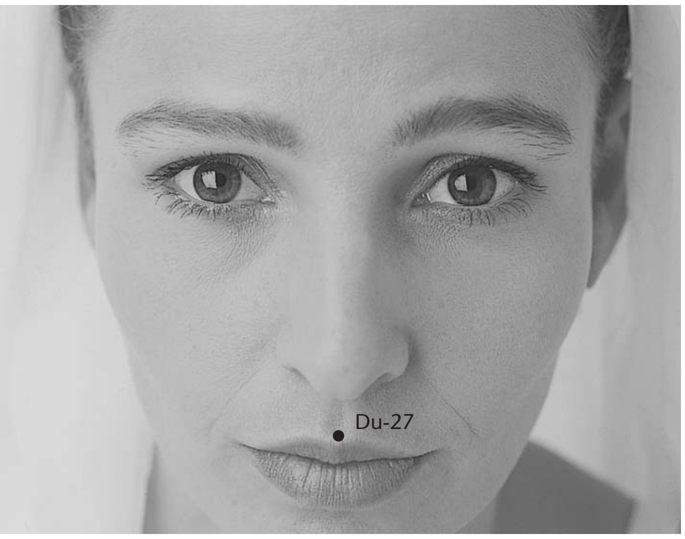

Location

On the midline, on the margin of the upper lip and the philtrum.

How to find

Location help: As the name implies, **Du-27** is located on the midline, at the junction of the upper lip and the philtrum.

Needling

Up to 0.3 cun obliquely in a superior direction. No moxibustion.

Actions/Indications

- Clears Heat
- Moistens the body
- Local point for the mouth
- Calms the *shen*

Du-28 *Gum Intersection* YINJIAO

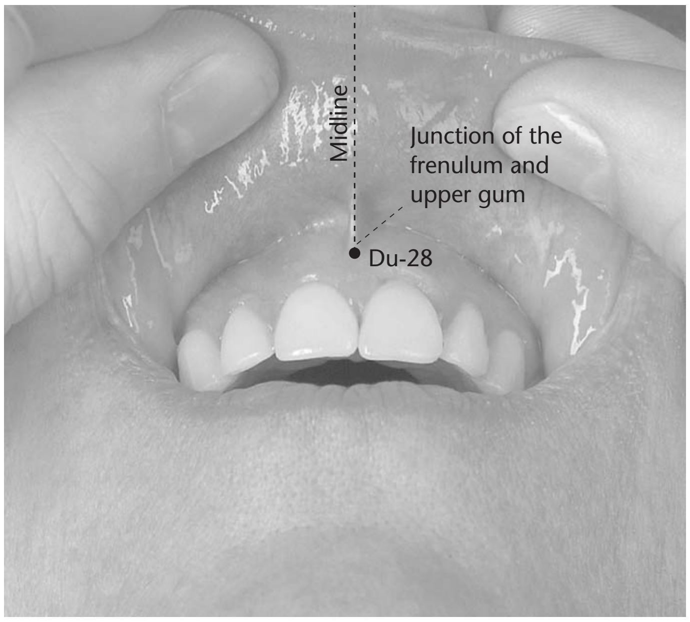

Location

On the inside of the upper lip, at the junction of the frenulum and the upper gum.

How to find

Hold the upper lip in a raised position in order to reveal the frenulum and the upper gum. Locate **Du-28** on the midline, at the junction of the frenulum with the upper gum.

Needling

0.1–0.2 cun obliquely in a superior direction or prick to bleed. Do not puncture the frenulum.

Actions/Indications

- Clears Heat (especially from the eyes and the mouth)

Special features

As the name implies, meeting point with the *ren mai* and the ST channel

5.4 *Dai mai*

Synonym: Girdling Vessel

The extraordinary vessel *dai mai* is the only vessel (or channel) with a horizontal pathway (except for *luo*-connecting vessels).

Relationships (→ 1.7.3)

- **Yin/Yang:** *chong mai*/*dai mai*
- **Central/peripheral:** *dai mai*/*yang wei mai*. Regions supported by this pairing: lateral eye region as well as temporal region, ears, cheeks, occiput and shoulders. Opening point: **G.B.-41** (*zulinqi*), coupled point: **T.B.-5** (*waiguan*).

Pathway

The extraordinary vessel *dai mai* originates in the hypochondrium at the level of L2; according to many other authors it originates at → **LIV-13** (*zhangmen*) at the free end of the 11th rib. It encircles the waist of the body like a belt. As it does so it passes → **G.B.-26** (*dai mai*) at the level of the umbilicus and below the free end of the 11th rib, → **G.B.-27** (*wushu*) and → **G.B.-28** (*weidao*).

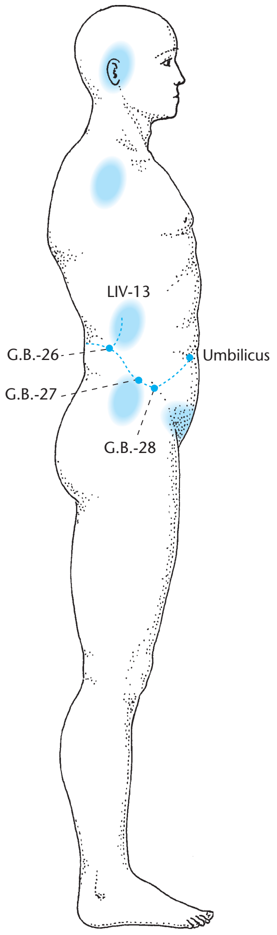

Meeting points with other channels

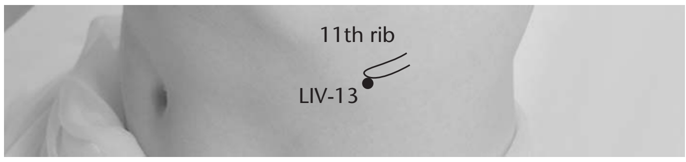

LIV-13 (*zhangmen*): Anterior and inferior to the free end of the 11th rib

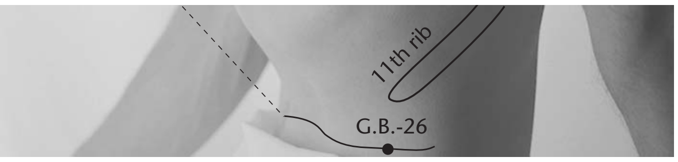

G.B.-26 (*daimai*): On the lateral aspect of the waist, at the junction of a vertical line through the free end of the 11th rib and a horizontal line through the umbilicus, approximately 1.8 cun inferior to → **LIV-13**

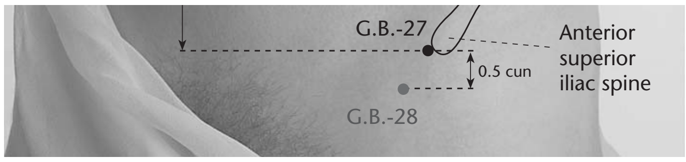

G.B.-27 (*wushu*): In the depression medial to the anterior superior iliac spine, approximately level with Ren-4

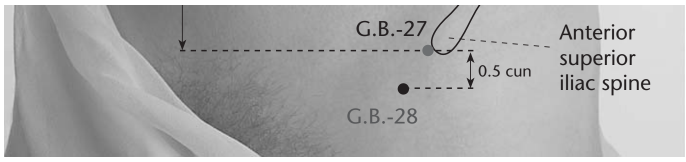

G.B.-28 (*weidao*): Anterior and inferior to the anterior superior iliac spine, 0.5 cun anterior and inferior to G.B.-27

Connections with other channels/Organs

- Gall Bladder and Liver primary channels, Kidney divergent channel
- Kidney, Uterus

Clinical importance (→ 1.7.2, 1.7.3)

- **Controls** and **encircles** the horizontal channels like a belt
- **Connects** the upper and lower halves of the body at the waist
- **Regulates** vaginal discharge
- **Eliminates** Damp-Heat from the Lower Burner, especially from the genitalia
- **Controls** the *shaoyang* channels, especially the G.B. channel

5.5 *Yin wei mai*

Synonym: Yin Motility Vessel (Deadman), **Yin Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

- **Yin/Yang:** *yin wei mai/yang wei mai*
- **Central/peripheral:** *chong mai/yin wei mai*. Regions supported by this pairing: Heart, thorax, Stomach. Opening point: **P-6** (*neiguan*), coupled point: **SP-4** (*gongsun*).

Pathway

The extraordinary vessel *yin wei mai* originates in the hypochondrium at → **KID-9** (*zhubin*), 5 cun proximal to the prominence of the medial malleolus, ascends the medial aspect of the leg to the inguinal region, follows the Spleen and Liver primary channels, passing the points → **SP-12** (*chongmen*) and → **SP-13** (*fushe*). It then passes → **SP-15** (*daheng*) and → **SP-16** (*fuai*) and again meets with the Spleen and Liver primary channels at → **LIV-14** (*qimen*). The vessel then traverses the thoracic region, passes → **Ren-22** (*tiantu*) in the suprasternal fossa and follows the extraordinary vessel *ren mai* to → **Ren-23** (*lianqiao*) where it terminates.

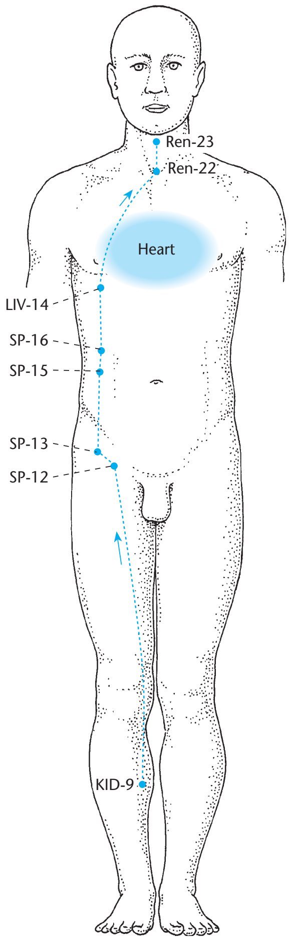
Ren-23
Ren-22
Heart
LIV-14
SP-16
SP-15
SP-13
SP-12
KID-9

Meeting points with other channels

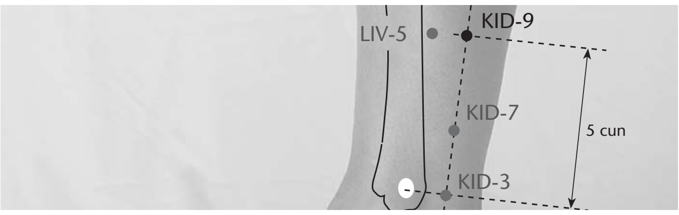
LIV-5
KID-9
KID-7
5 cun
KID-3

KID-9 (*zhubin*): 5 cun proximal to KID-3 and 2 cun posterior to the medial border of the tibia

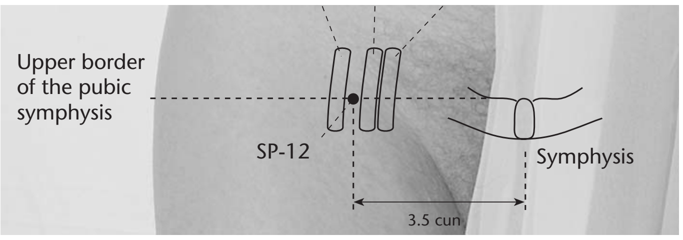
Upper border of the pubic symphysis
SP-12
Symphysis
3.5 cun

SP-12 (*chongmen*): 3.5 cun lateral to the midline, level with the upper border of the pubic symphysis, lateral to the femoral artery

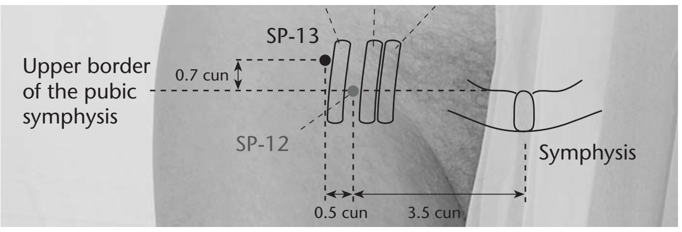
SP-13
Upper border of the pubic symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

SP-13 (*fushe*): 4 cun lateral to the anterior midline (mamillary line) and 0.7 cun superior to the upper border of the pubic symphysis

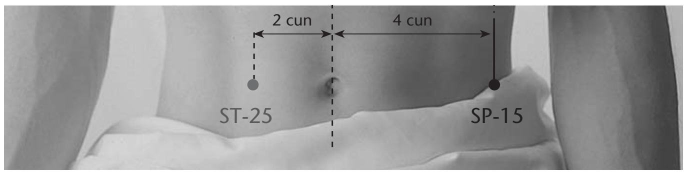
2 cun
4 cun
ST-25
SP-15

SP-15 (*daheng*): 4 cun lateral to the centre of the umbilicus

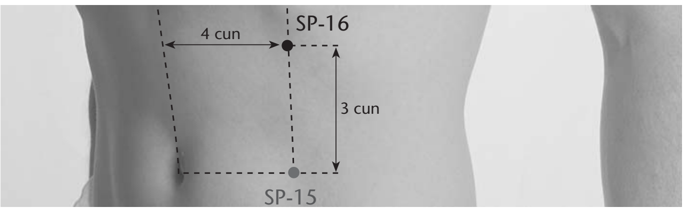
4 cun
SP-16
3 cun
SP-15

SP-16 (*fuai*): 3 cun superior to the centre of the umbilicus and 4 cun lateral to the anterior midline, on the mamillary line

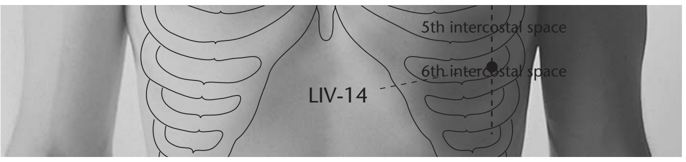
5th intercostal space
6th intercostal space
LIV-14

LIV-14 (*qimen*): In the 6th intercostal space, on the mamillary line, 4 cun lateral to the midline.

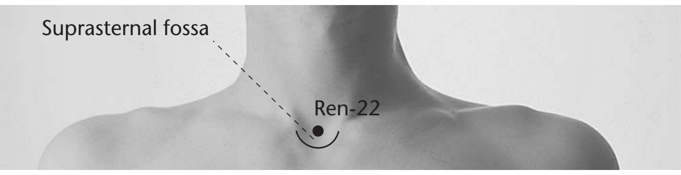
Suprasternal fossa
Ren-22

Ren-22 (*tiantu*): 0.5 cun superior to the sternum, in the centre of the suprasternal fossa

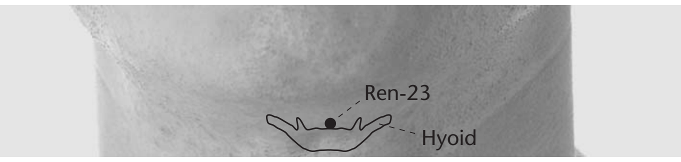
Ren-23
Hyoid

Ren-23 (*lianquan*): On the anterior midline, superior to the upper border of the hyoid bone

Connections with other channels/organs

- Kidney, Spleen and Liver primary channels, *ren mai*
- Internal organs

Clinical importance (→ 1.7.2, 1.7.3)

- **Connects** and **regulates** all **Yin channels** of the body. It connects the KID, SP and LIV channels and the *ren mai* and dominates the **Interior** of the body (the internal organs).
- **Strengthens** Yin and Heart-Blood, especially in women
- **Balances** the emotions

5.6 *Yang wei mai*

Synonym: Yang Motility Vessel (Deadman), **Yang Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

Yin/Yang: *yin wei mai*/*yang wei mai*
Central/peripheral: *dai mai*/*yang wei mai*. Regions supported by this pairing: Lateral eye region, temporal region, ears, cheeks, neck and shoulders. Opening point: **T.B.-5** (*waiguan*), coupled point: **G.B.-41** (*zulinqi*).

Pathway

The extraordinary vessel *yang wei mai* originates at → **BL-63** (*jinmen*) in the depression posterior to the tuberosity of the 5th metatarsal bone, runs anterior to the lateral malleolus, passes → **G.B.-35** (*yangjiao*), ascends the lateral aspect of the leg, passes the hip, ascends the posterior aspect of the costal region to the shoulder, passes → **S.I.-10** (*naoshu*), → **T.B.-15** (*tianliao*), → **G.B.-21** (*jianjing*) and ascends the neck. The vessel then ascends anterior to the ear (according to some authors, for example Solinas et al. (1998), posterior to the ear) to the forehead, passing → **ST-8** (*touwei*), → **G.B.-13** (*benshen*) and → **G.B.-14** (*yangbai*). It then runs with the G.B. channel from **G.B.-15** to **G.B.-20** on the occiput, passing → **G.B.-15** (*toulinqi*), → **G.B.-16** (*muchuang*), → **G.B.-17** (*zhengying*), → **G.B.-18** (*chengling*), → **G.B.-19** (*naokong*) and → **G.B.-20** (*fengchi*). The vessel then continues from **G.B.-20** to the posterior midline, passing → **Du-16** (*fengfu*) and → **Du-15** (*yamen*), where the vessel terminates.

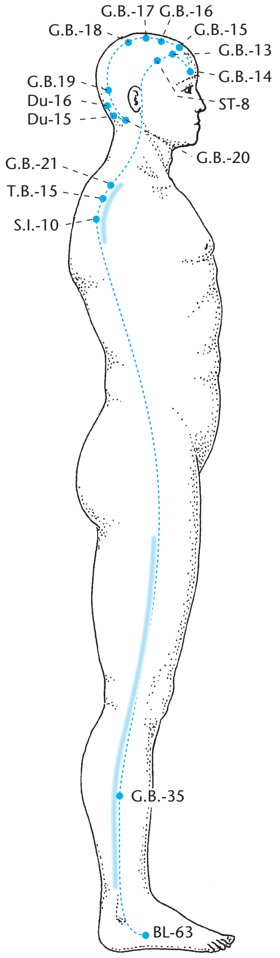

Meeting points with other channels

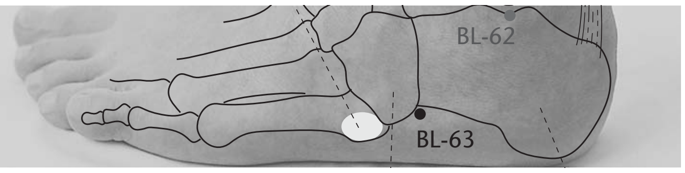

BL-63 (*jinmen*): In a depression anterior and inferior to → **BL-62** between the calcaneus and the cuboid bone, proximal to the tuberosity of the 5th metatarsal bone

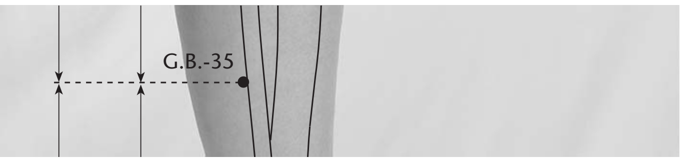

G.B.-35 (*yangjiao*): 7 cun proximal to the highest prominence of the lateral malleolus, on the posterior border of the fibula

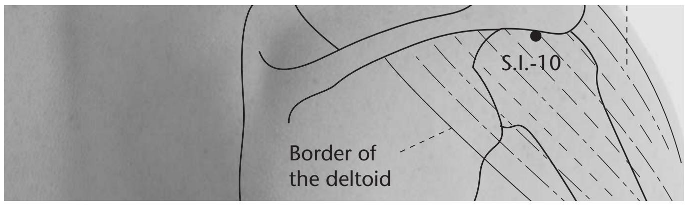
S.I.-10
Border of the deltoid

S.I.-10 (*naoshu*): With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

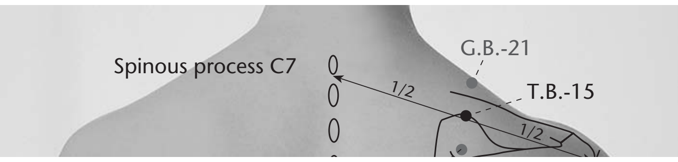
Spinous process C7
G.B.-21
1/2
T.B.-15
1/2

T.B.-15 (*tianliao*): At the midpoint of an imaginary line between the spinous process of the C7 and the lateral extremity of the acromion

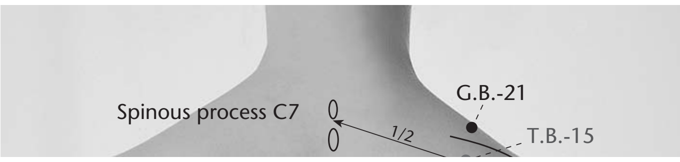
Spinous process C7
G.B.-21
1/2
T.B.-15

G.B.-21 (*jianjing*): At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

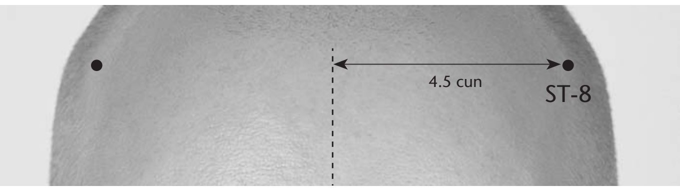

ST-8 (*touwei*): At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (Du-24)

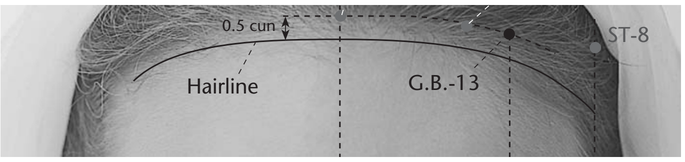
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*): 3 cun lateral to → **Du-24** (on the midline, 0.5 cun superior to the anterior hairline)

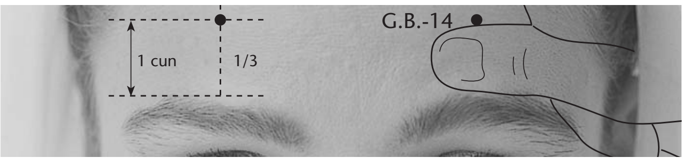
G.B.-14
1 cun
1/3

G.B.-14 (*yangbai*): When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow, at the junction of the eminence and the superciliary arch

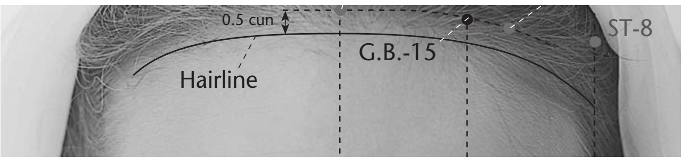
0.5 cun
ST-8
G.B.-15
Hairline

G.B.-15 (*toulinqi*): When looking straight ahead superior to the pupil, 0.5 cun superior to the anterior hairline

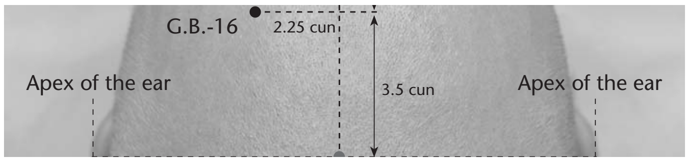
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear

G.B.-16 (*muchuang*): 1.5 cun superior to the anterior hairline, on the pupil line, or 2.25 cun lateral to the midline (= midway between → **Du-24** and → **ST-8**)

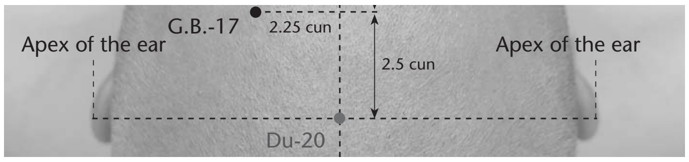
G.B.-17
2.25 cun
Apex of the ear
Apex of the ear
2.5 cun
Du-20

G.B.-17 (*zhengying*): 2.5 cun superior to the anterior hairline and 2.25 cun lateral to the midline

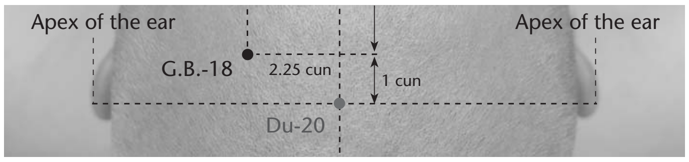
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*): 4 cun superior to the anterior hairline or 1 cun anterior to → **Du-20**, 2.25 cun lateral to the midline

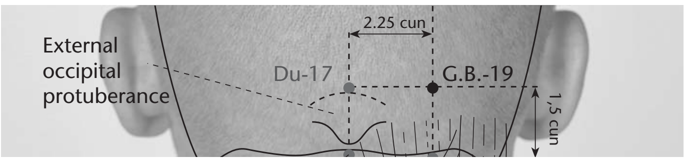
External
occipital
protuberance
2.25 cun
Du-17
G.B.-19
1.5 cun

G.B.-19 (*naokong*): On the occiput, level with the upper border of the external occipital protuberance (→ **Du-17**), 2.25 cun lateral to the midline

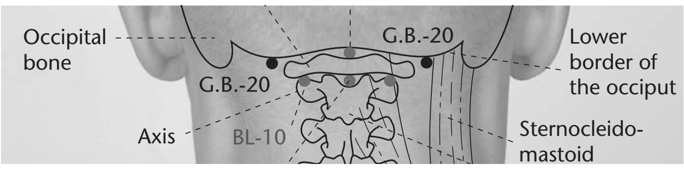
Occipital
bone
G.B.-20
G.B.-20
Lower
border of
the occiput
Axis
BL-10
Sternocleido-
mastoid

G.B.-20 (*fengchi*): On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

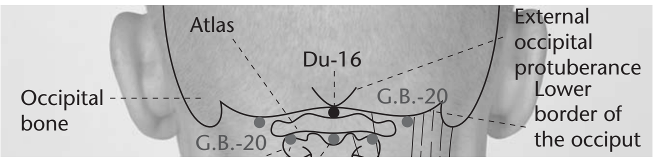
Atlas
Du-16
External
occipital
protuberance
Occipital
bone
G.B.-20
Lower
border of
the occiput
G.B.-20

Du-16 (*fengfu*): On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

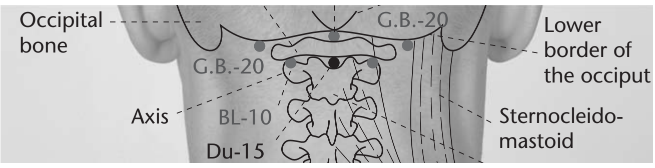
Occipital
bone
G.B.-20
Lower
border of
the occiput
G.B.-20
Axis
BL-10
Du-15
Sternocleido-
mastoid

Du-15 (*yamen*): On the occiput, on the posterior midline, in the depression between the 1st (atlas) and 2nd cervical vertebrae, approximately 0.5 cun below → **Du-16**

Connections with other channels/Organs

- BL, G.B., ST, S.I., T.B. and L.I. primary channels, *du mai*

Clinical importance

- **Connects** and **regulates** all **Yang channels** of the body. It connects the BL, G.B., ST, S.I., T.B. and L.I. channels as well as the *du mai* and controls the Exterior of the whole body (especially the *taiyang* and *shaoyang* channels)
- Harmonises the Nutritive Qi (*ying qi*) and the Defensive Qi (*wei qi*)
- Expels Exterior Wind-Cold, especially with *shaoyang*-syndrome.

5.7 *Yin qiao mai*

Synonym: Yin Motility Vessel (Deadman), **Yin Heel Vessel** (Maciocia)

Relationships (→ 1.7.2, 1.7.3)

- **Yin/Yang:** *yin qiao mai/yang qiao mai*
- **Central/peripheral:** *ren mai/yin qiao mai*. Regions supported by this pairing: Face, throat, thorax, Lung, diaphragm, abdomen. Opening point: **KID-6** (*zhaohai*), coupled point: **LU-7** (*lieque*).

Pathway

The extraordinary vessel *yin qiao mai* originates at → **KID-2** (*rangu*) inferior to the navicular tuberosity (according to some authors it only begins at → **KID-6**), runs to → **KID-6** (*zhaohai*) and → **KID-8** (*jiaoxin*), ascends the posteromedial aspect of the leg to the external genital region, ascends the abdomen to the thorax, crosses the supraclavicular fossa and continues to the throat and face. It then runs to the inner canthus of the eye where it connects with the Bladder primary channel and the *yang qiao mai* at → **BL-1** (*jingming*). It then ascends to enter the Brain.

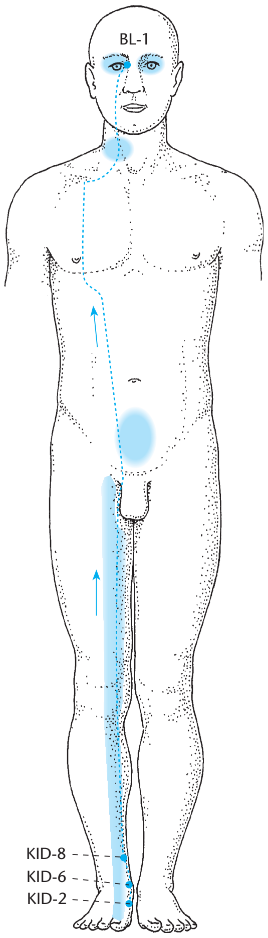

Meeting points with other channels

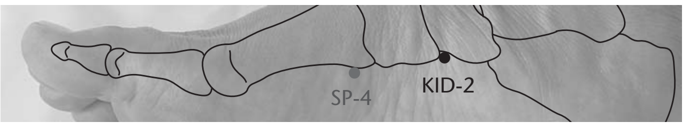

KID-2 (*rangu*): In a depression at the anterior/inferior border of the navicular bone, at the border of the 'red and white' skin

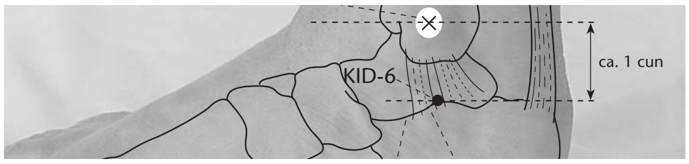

KID-6 (*zhaohai*): In the depression inferior to the highest prominence of the medial malleolus, over the joint space between the talus and calcaneus

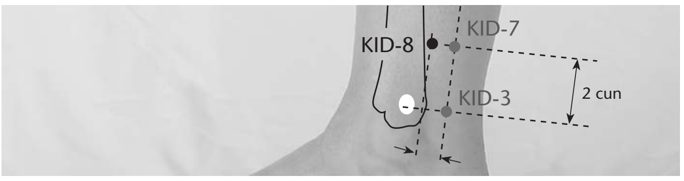

KID-8 (*jiaoxin*): 2 cun proximal to the highest prominence of the medial malleolus, posterior to the border of the tibia

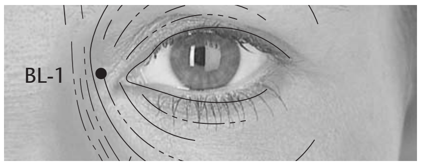

BL-1 (*jingming*): In a depression 0.1 cun superior and medial to the inner canthus of the eye

Connections with other channels/Organs

- Kidney and Bladder primary channels
- Brain

Clinical importance (→ 1.7.2, 1.7.3)

- Together with the *yang qiao mai* controls the musculature of the lower extremities
- Together with the *yang qiao mai* regulates the opening and closing of the eyes
- Regulates the functioning of the Brain
- Removes stagnation (of Qi, Blood or Dampness) in the Lower Burner, especially in the Uterus in women

5.8 *Yang qiao mai*

Synonym: Yang Motility Vessel (Deadman), **Yang Heel Vessel** (Maciocia)

Relationships (→ 1.7.3)

- **Yin/Yang:** *yin qiao mai/yang qiao mai*
- **Central/peripheral:** *du mai/yin qiao mai*. Regions supported by this pairing: Inner canthus of the eye, occiput, shoulders and back. Opening point: **BL-62** (*shenmai*), coupled point: **S.I.3** (*houxi*)).

Pathway

The extraordinary vessel *yang qiao mai* originates at → **BL-62** (*shenmai*) below the prominence of the lateral malleolus, curves briefly around the malleolus to → **BL-61** (*pucan*) and ascends anterior to the Achilles tendon to → **BL-59** (*fuyang*), travels up the lateral aspect of the leg, crosses the hip, passing → **G.B.-29** (*juliao*), ascends the flanks and the posterior aspect of the shoulders, passes → **S.I.10** (*naoshu*), continues to → **L.I.-15** (*jianyu*) and runs in a curve to → **L.I.-16** (*jugu*), crosses the supraclavicular fossa, ascends the throat, according to some authors passing → **ST-9** (*renying*), and reaches the face. Here the vessel passes the points → **ST-4** (*dicang*) and → **ST-3** (*juliao*); according to some authors also → **ST-2** (*sibai*) and → **ST-1** (*chengqi*) in the infraorbital region. It then reaches → **BL-1** (*jingming*), where it meets the extraordinary vessel *yin qiao mai*, ascends the forehead to the vertex and descends posteriorly to → **G.B.-20** (*fengchi*); according to the *Nan Jing* and the *Nei Jing* it also reaches → **Du-16** (*fengfu*), where it enters the Brain.

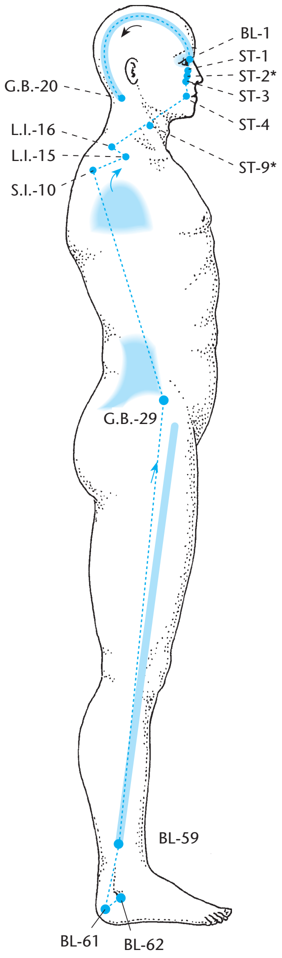

* According to Deadman et al 1998.

Intersection points with other channels

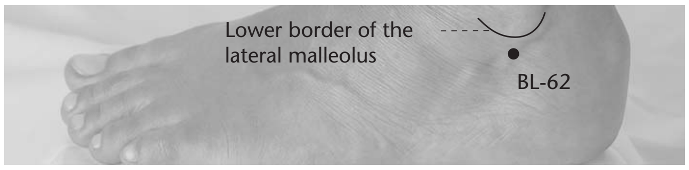

BL-62 (*shenmai*): In a depression directly inferior to the highest prominence of the lateral malleolus, over the joint space between the talus and calcaneus

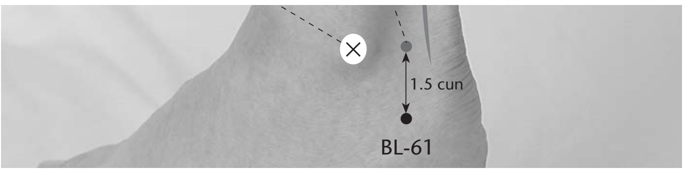

BL-61 (*pucan*): In a depression on the calcaneus, approximately 1.5 cun inferior to → **BL-60**

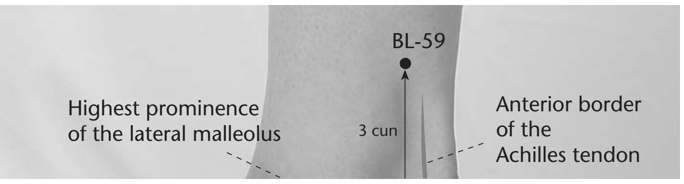
BL-59
Highest prominence of the lateral malleolus
3 cun
Anterior border of the Achilles tendon

BL-59 (*fuyang*): 3 cun superior to → BL-60 (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon)

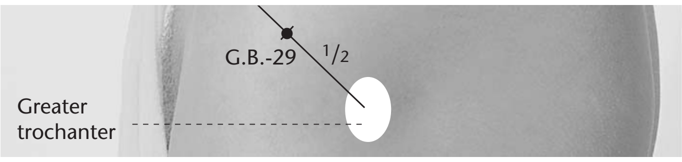
G.B.-29
1/2
Greater trochanter

G.B.-29 (*juliao*): At the midpoint of a line connecting the anterior superior iliac spine and the greater trochanter, at the anterior border of the iliac crest

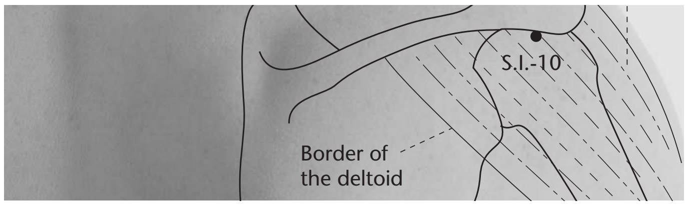
S.I.-10
Border of the deltoid

S.I.-10 (*naoshu*): With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

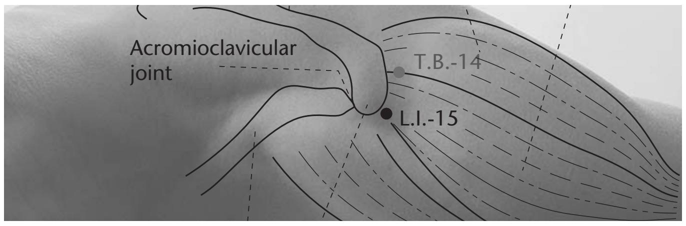
Acromioclavicular joint
T.B.-14
L.I.-15

L.I.-15 (*jianyu*): In the depression distal and anterior to the acromion, between the clavicular and acromial portions of the deltoid muscle

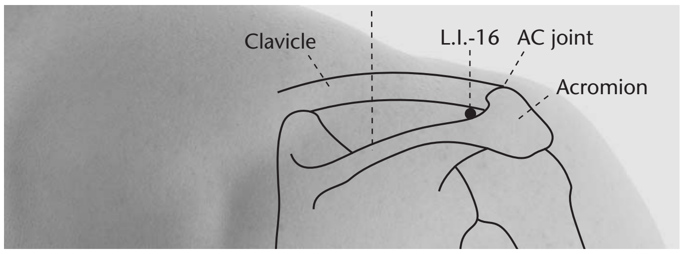
Clavicle
L.I.-16
AC joint
Acromion

L.I.-16 (*jugu*): In a depression between the acromial extremity of the clavicle and the junction of the scapular spine and the acromion

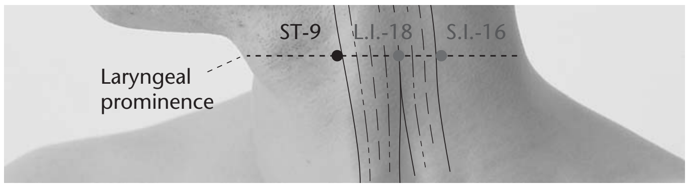
ST-9
L.I.-18
S.I.-16
Laryngeal prominence

ST-9 (*renying*): Approximately 1.5 cun lateral to the anterior midline, level with the laryngeal prominence, at the anterior border of the sternocleidomastoid

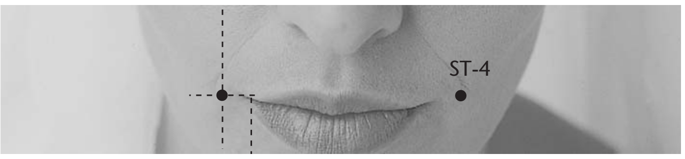

ST-4 (*dicang*): When looking straight ahead on the pupil line, approximately 0.4 cun lateral to the corner of the mouth

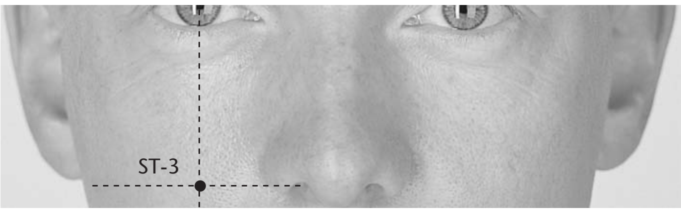

ST-3 (*juliao*): When looking straight ahead, directly below the centre of the pupil, level with the lower border of the ala nasi

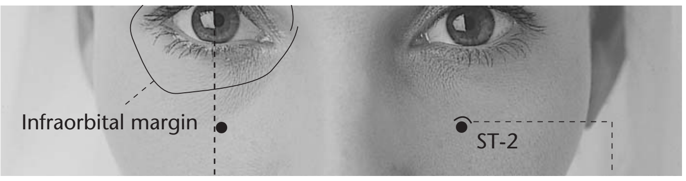
Infraorbital margin
ST-2

ST-2 (*sibai*): When looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen

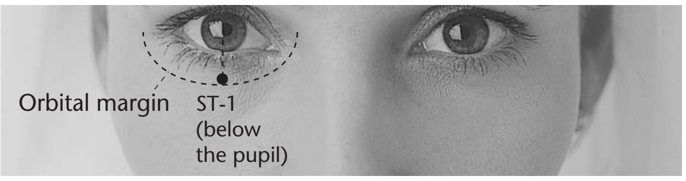
Orbital margin
ST-1
(below
the pupil)

ST-1 (*chengqi*): With the eyes looking straight forward, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

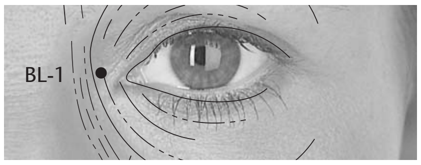

BL-1 (*jingming*): In a depression 0.1 cun superior and medial to the inner canthus of the eye

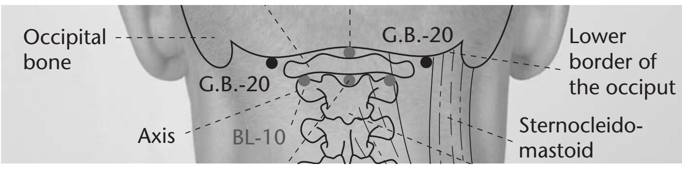
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*): On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

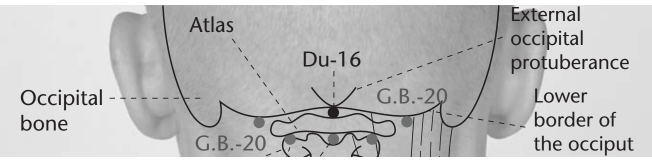
Atlas
Du-16
External occipital protuberance
Occipital bone
G.B.-20
Lower border of the occiput
G.B.-20

Du-16 (*fengfu*): On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

Connections with other channels/Organs

- Kidney, Bladder, Gall Bladder, Stomach, Large Intestine and Small Intestine primary channels
- Brain

Clinical importance (→ 1.7.2, 1.7.3)

- Together with the *yin qiao mai* controls the musculature of the lower extremities
- Together with the *yang qiao mai* regulates the opening and closing of the eyes
- Regulates the functioning of the Brain, eliminates internal and external Wind from the head
- Removes obstructions and stagnation from the spine, especially after traumatic injuries

6 Extra Points

Claudia Focks, Ulrich März

Besides the 361 classic acupuncture points located on the channels there are a number of **Ex**tra Points (**Ex**), located with a few exceptions outside the pathways of the primary channels. In 1991, a new nomenclature for these extra points was confirmed in the People's Republic of China, listing 48 points (*The Location of Acupoints, State Standard of the People's Republic of China*, Foreign Languages Press, Beijing, 1990).

In the past, authors have developed ways of identifying the extra points by names or numbers:

- Nguyen Van Nghi; König and Wancura: '**P**unkte **a**ußerhalb der **M**eridiane' (points not on meridians) (PaM) and '**N**eu-**P**unkte' (new points) (NP); Schnorrenberger (wall charts translated from the Chinese language) uses the same numbering system as Nguyen Van Nghi but refers to the **PaM** as '**Z**usatz**p**unkte' (additional points) (ZP) and also as '**N**eu-**P**unkte' (NP)
- Shanghai College for Traditional Medicine (*Acupuncture – a Comprehensive Text*, English translation by O'Connor J, Bensky D, which also forms the foundation for Deadman et al. 1998 and Ellis and Wiseman 1991): **M**iscellaneous (M) and **N**ew (N) Points
- Hempen 1995: '**E**xtrapunkte' (extra points) (Ex).

The following tables give an overview of the most commonly used extra points also listed in this atlas.

6.1 Extra Points: Head and Neck (EX-HN)

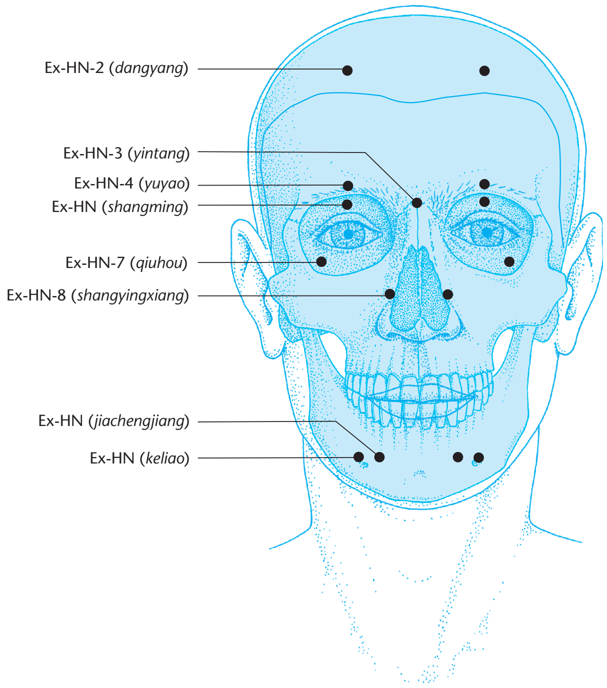
Ex-HN-2 (*dangyang*)
Ex-HN-3 (*yintang*)
Ex-HN-4 (*yuyao*)
Ex-HN (*shangming*)
Ex-HN-7 (*qiuhou*)
Ex-HN-8 (*shangyingxiang*)
Ex-HN (*jiachengjiang*)
Ex-HN (*keliao*)

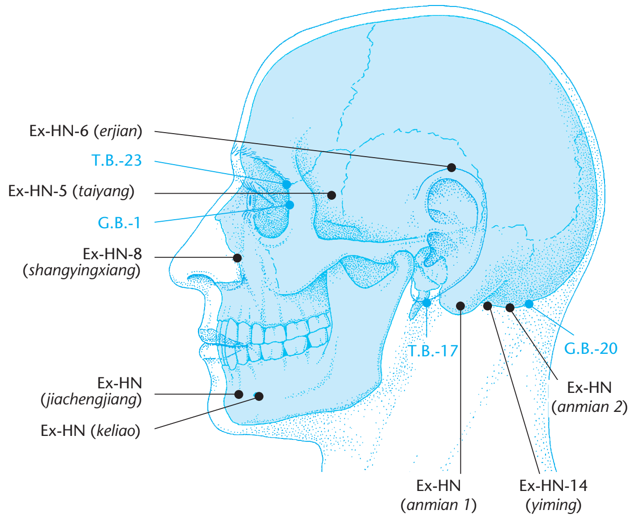
Ex-HN-6 (*erjian*)
T.B.-23
Ex-HN-5 (*taiyang*)
G.B.-1
Ex-HN-8 (*shangyingxiang*)
Ex-HN (*jiachengjiang*)
Ex-HN (*keliao*)
T.B.-17
G.B.-20
Ex-HN (*anmian 2*)
Ex-HN (*anmian 1*)
Ex-HN-14 (*yiming*)

English abbreviation (standard)	*Pinyin* name	Nguyen Van Nghi, König and Wancura, Schnorrenberger	Shanghai College	Ex (Hempen)
Ex-HN-1	*sishencong*	PaM or ZP1	M-HN-1	Ex6
Ex-HN-2	*dangyang*		M-HN	
Ex-HN-3	*yintang*	PaM or ZP3	M-HN-3	Ex1
Ex-HN-4	*yuyao*	PaM or ZP6	M-HN-6	
Ex-HN-5	*taiyang*	PaM or ZP9	M-HN-9	Ex2
Ex-HN-6	*erjian*	PaM or ZP10	M-HN-10	
Ex-HN-7	*qiuhou*	PaM or ZP8	M-HN-8	
Ex-HN-8	*shangyingxiang/bitong*	NP12 (PaM or ZP14*)	M-HN-14	Ex3
Ex-HN-9	*neiyingxiang*		M-HN-35	
Ex-HN-10	*juquan*		M-HN-36	
Ex-HN-11	*haiquan*		M-HN-37	
Ex-HN-12	*jinjin*†	PaM or ZP20	M-HN-20	
Ex-HN-13	*yuye*†	PaM or ZP20	M-HN-20	
Ex-HN-14	*yiming*	PaM or ZP13	M-HN-13	Ex4
Ex-HN-15	*jing(bailao)*	PaM or ZP30	M-HN-30	
Further extra points				
Ex-HN	*shangming*		N-HN-4	
Ex-HN	*anmian*‡	NP27 and 28 (*anmian*‡1 und 2)	N-HN-54 M-HN-54 (Deadman)	Ex5
Ex-HN	*jiabi*	PaM or ZP15		
Ex-HN	*jiachengjiang*	PaM or ZP18 (*heliao/keliao*)	M-HN-18	Ex7
Ex-HN	*chonggu/zhuidong*	PaM or ZP31	M-HN-31	
Ex-HN	*jingbi*		M-HN-41	

---

[*] Nguyen Van Nghi and Schnorrenberger describe a different location for the point with this *pinyin* name (location: 0.5 cun inferior to the inner canthus of the eye); the 'standard' location is described under NP12 as '*bitong'* or '*bicong'*.

[†] The Shanghai nomenclature and also Nguyen Van Nghi, König and Wancura and Schnorrenberger describe the extra points *jinjin* and *yuye* as a pair of points.

[‡] There exists differing information regarding the extra point *anmian*; for example the Shanghai College (and Wiseman) describe *anmian* as N-HN-54 (in Deadman: M-HN-54) with the following location: midway between G.B.-20 and T.B.-17; Nguyen Van Nghi and Schnorrenberger describe *anmian* as two new points with the following locations: *anmian* 1 is located midway between T.B.-17 and *yiming* (PaM13 or Ex-HN-14), *anmian* 2 midway between G.B.-20 and *yiming* (*anmian* according to the Shanghai College is shown on the point page; the location of *anmian* 1 and 2 according to Nguyen and Schnorrenberger in figure 6.1).

Four Alert Spirit SISHENCONG **Ex-HN-1**

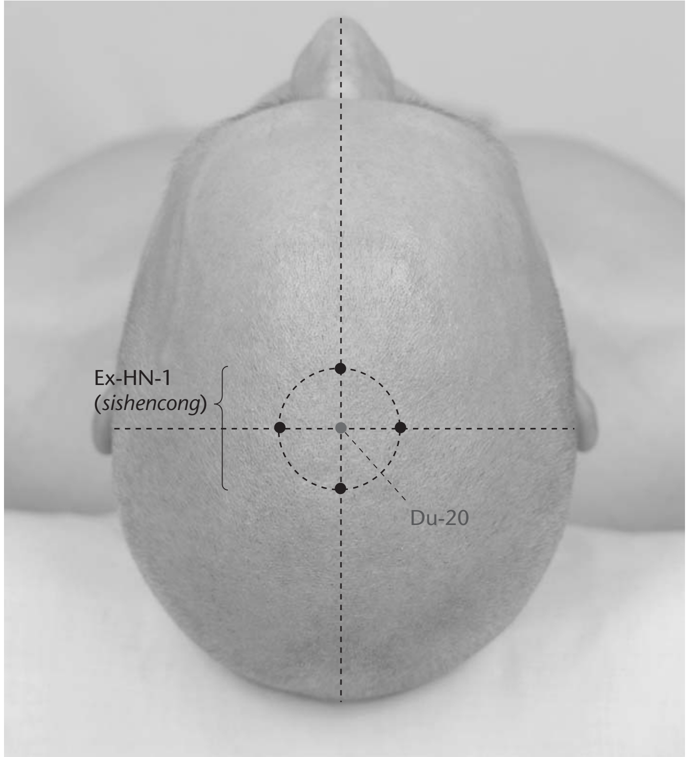
Ex-HN-1
(*sishencong*)
Du-20

Location

A group of four points, each located 1 cun from → **Du-20** (anterior, posterior and lateral) (→ 3.1.1).

How to find

First, locate → **Du-20** on the highest point of the head, at the junction of a line connecting the apices of the ears and the vertical midline. Alternative location method: Locate → **Du-20** 5 cun superior to the anterior and 7 cun superior to the posterior hairline. **Ex-HN-1** (*sishencong*) forms a star-shaped group of points around → **Du-20**: two points are located at a distance of 1 cun each on the *du mai*, whereas the other two are located 1 cun each from → **Du-20** in a lateral direction.

Needling

0.5–1 cun obliquely towards → **Du-20**

Actions/Indications

- Calm the *shen*
- Alleviate pain
- Calm Wind
- Benefit the eyes and ears

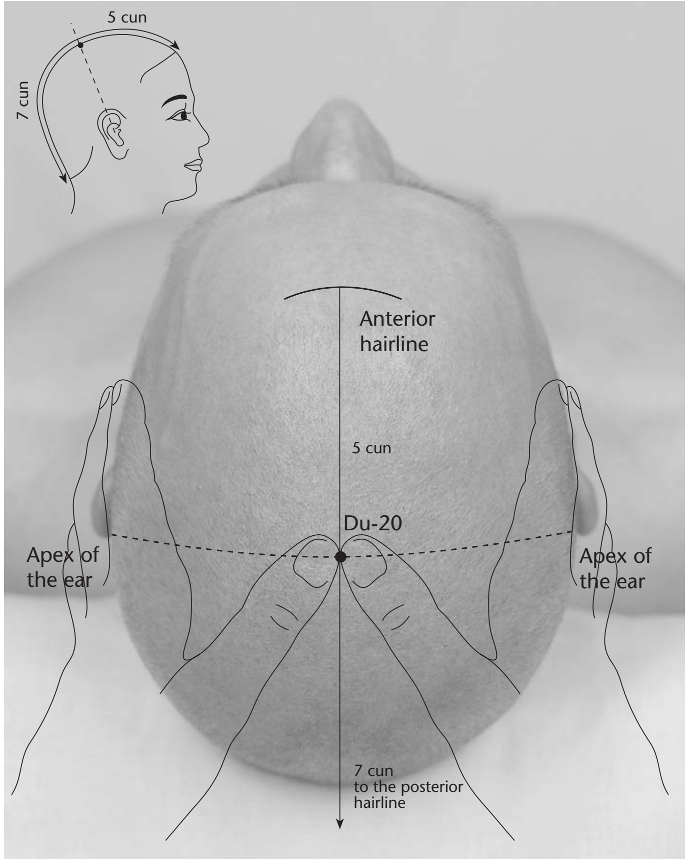
5 cun
7 cun
Anterior hairline
5 cun
Du-20
Apex of the ear
Apex of the ear
7 cun to the posterior hairline

Ex-HN-2 *Above the Yang* DANGYANG

Location

With the patient looking straight ahead, vertically above the pupil, 1 cun above the anterior hairline.

How to find

In patients with a receding hairline, the original hairline can often be recognized by a change in skin texture. Location help (➞ 3.1.1): Ask the patient to frown, as this often reveals the site of the original hairline. Ask the patient to look straight ahead, then locate **Ex-HN-2** by measuring 1 cun from the hairline, on a vertical line through the pupil.

➞ **G.B.-14** is also located on the pupil line (on the lower border of the frontal eminence). Located on the same level (1 cun above the hairline) are ➞ **BL-5** (more medial, 1.5 cun from the anterior midline) and ➞ **Du-23** (on the anterior midline).

Needling

0.5 cun transversely (subcutaneously) towards the site of the disorder/pain.

Actions/Indications

- Dispels Wind and Heat, alleviates pain

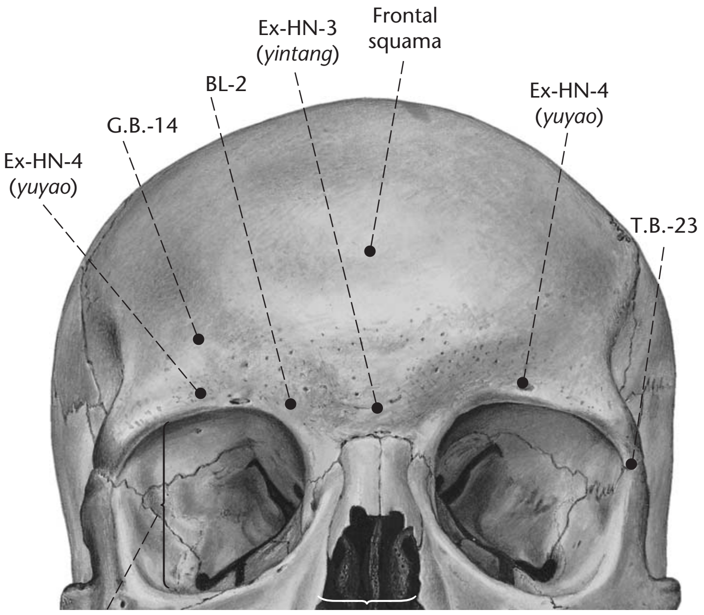
Ex-HN-3
(*yintang*)
Frontal
squama
BL-2
Ex-HN-4
(*yuyao*)
G.B.-14
Ex-HN-4
(*yuyao*)
T.B.-23

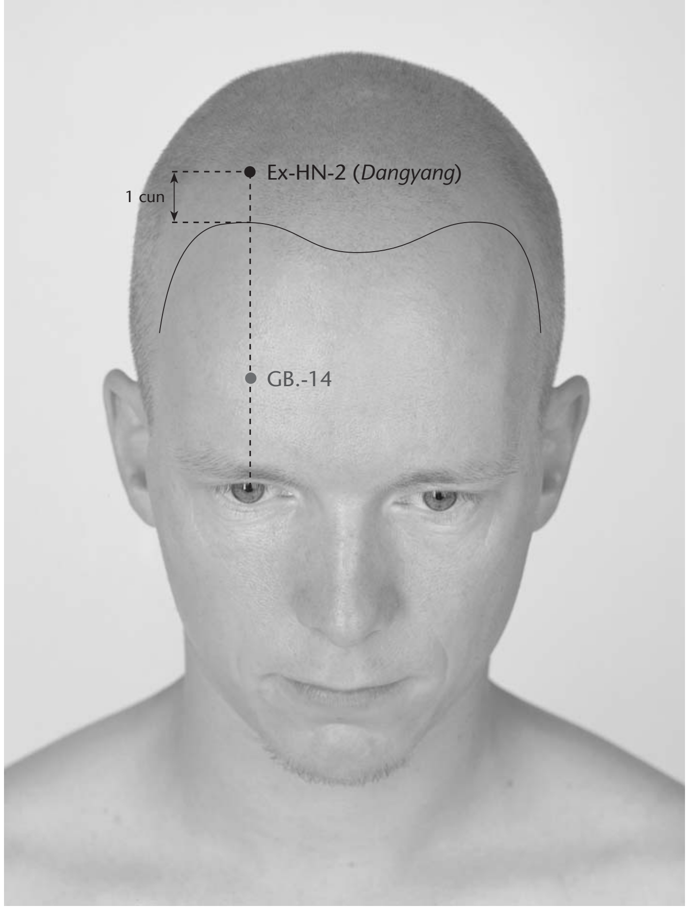
1 cun
Ex-HN-2 (*Dangyang*)
GB.-14

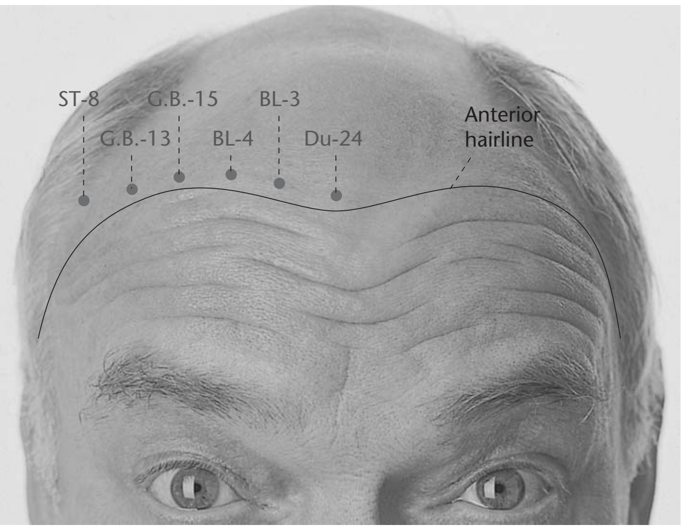
ST-8
G.B.-15
BL-3
G.B.-13
BL-4
Du-24
Anterior
hairline

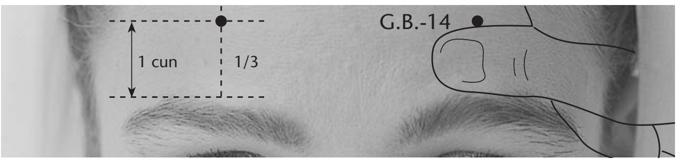
1 cun
1/3
G.B.-14

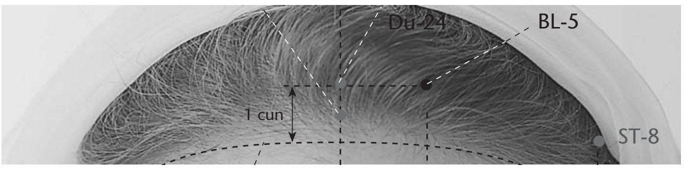
Du-24
BL-5
1 cun
ST-8

Hall of Impression YINTANG Ex-HN-3

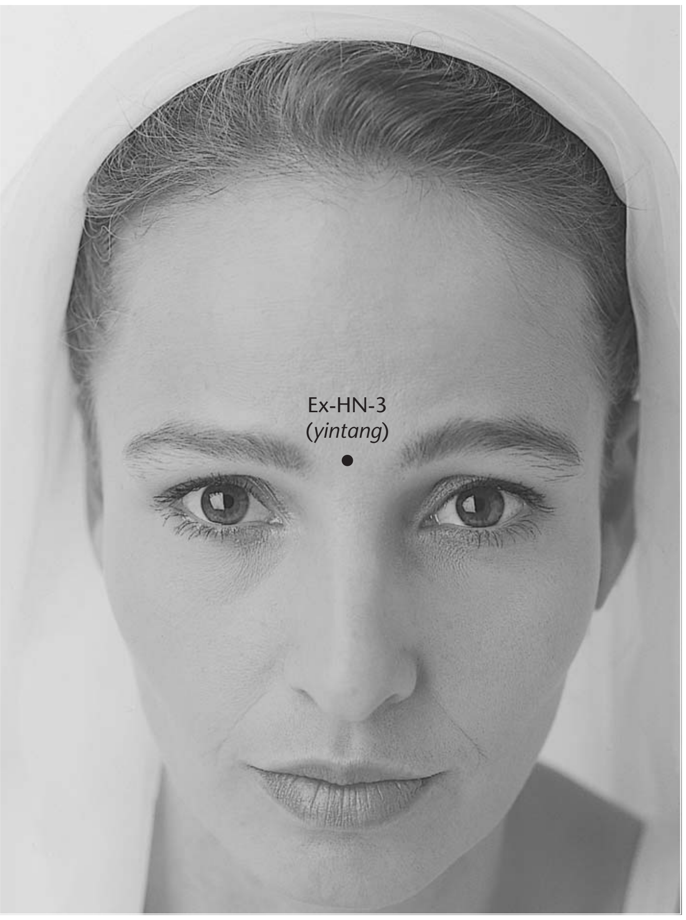
Ex-HN-3
(*yintang*)

Location

On the anterior midline (*du mai*), between the eyebrows.

How to find

Locate the glabella (→ 3.1.1), a smooth surface on the frontal bone, directly above the root of the nose, between the superciliary arches. Then locate **Ex-HN-3** (*yintang*) in its centre, on the midline and between the medial extremities of the eyebrows. Located along the superciliary arches (from medial to lateral) are → **BL-2,** → **Ex-HN-4** (*yuyao*) and → **T.B.-23**.

Needling

Pinching-skin method: With the thumb and index finger, form a skin fold above the point and insert the needle 0.3–0.5 cun transversely from superior to inferior into the fold, towards the root of the nose. Alternative methods: Oblique or transverse (subcutaneous) insertion towards the eyebrows (→ **BL-2**) or prick to bleed.

Actions/Indications

- Calms the *shen*
- Benefits the nose
- Eliminates (internal) Wind
- Opens the channel, alleviates pain

Special features

Commonly used extra point with a harmonising effect

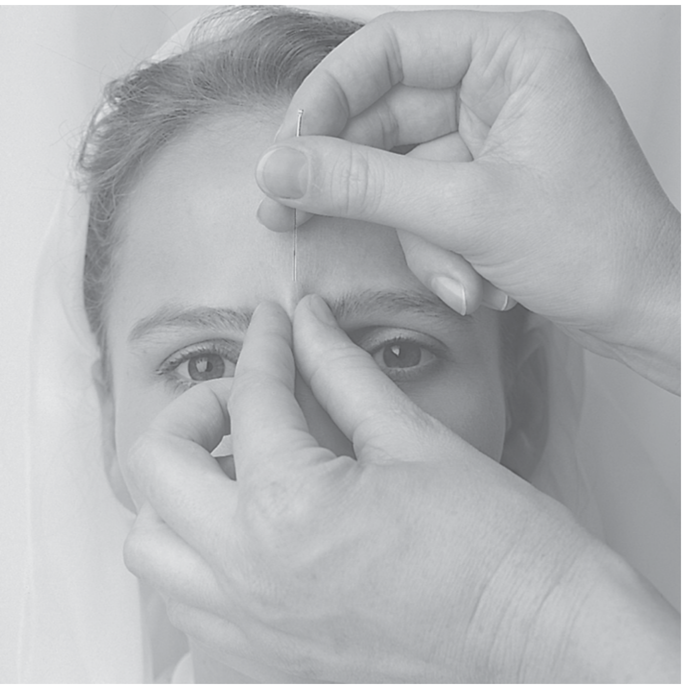

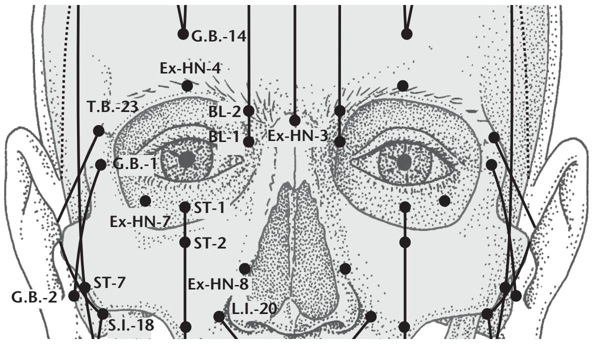
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20
S.I.-18

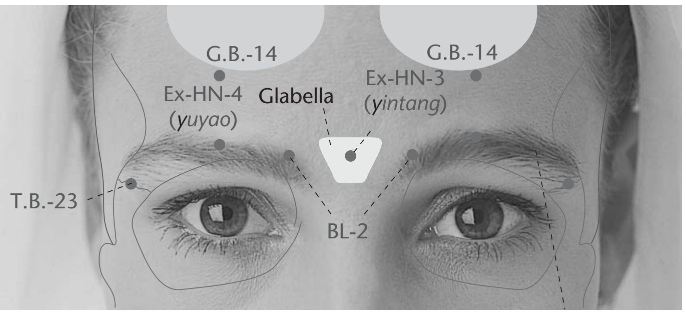
G.B.-14
G.B.-14
Ex-HN-3
(*yintang*)
Ex-HN-4
(*yuyao*)
Glabella
T.B.-23
BL-2

Ex-HN-4 *Fish Waist* YUYAO

Location

In the centre of the eyebrow, with the eyes looking straight ahead directly above the pupil.

How to find

With the patient looking straight ahead, palpate the eyebrow directly above the pupil, superior to the supraorbital ridge. There, locate **Ex-HN-4** (*yuyao*) in a small depression on the superciliary arch.

Needling

0.3–0.5 cun vertically or obliquely towards the eye.

Actions/Indications

- Benefits the eyes
- Dispels pain

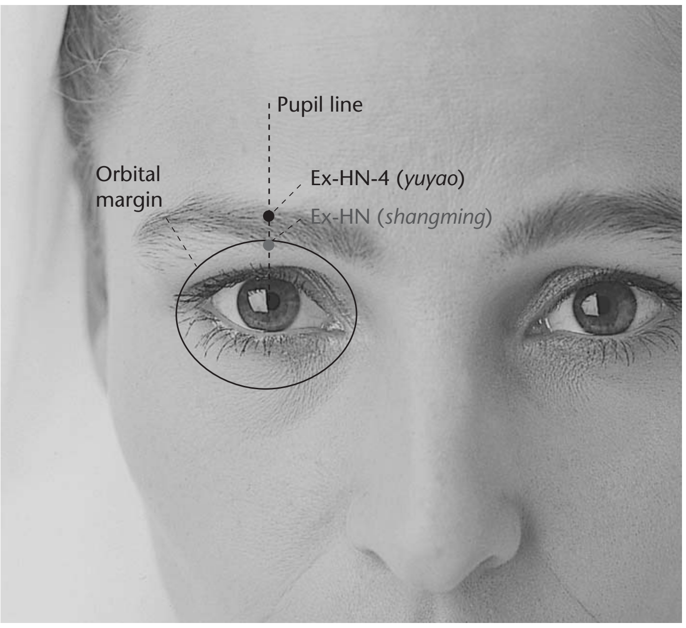
Pupil line
Orbital margin
Ex-HN-4 (*yuyao*)
Ex-HN (*shangming*)

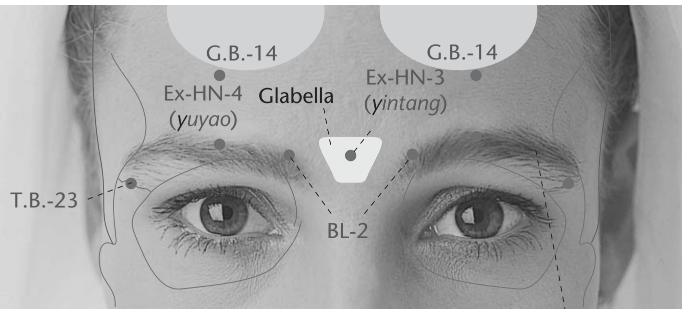
G.B.-14
G.B.-14
Ex-HN-3 (*yintang*)
Ex-HN-4 (*yuyao*)
Glabella
T.B.-23
BL-2

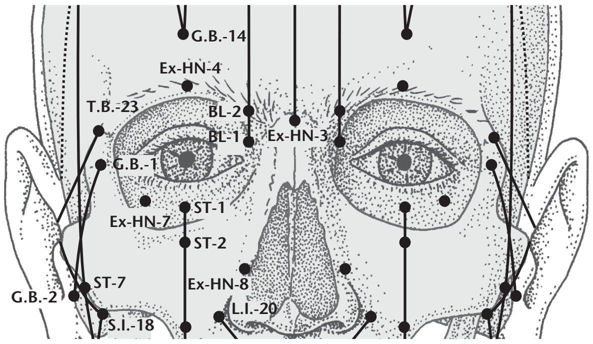
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18

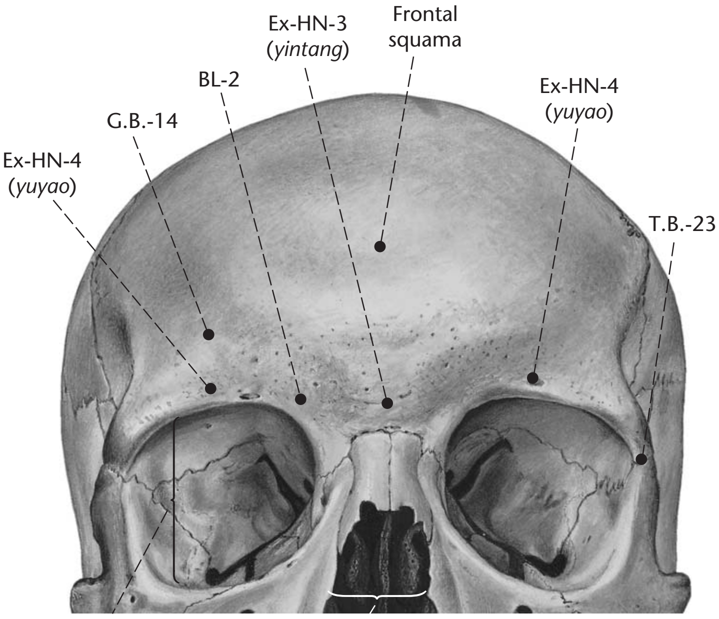
Ex-HN-3 (*yintang*)
Frontal squama
BL-2
Ex-HN-4 (*yuyao*)
G.B.-14
Ex-HN-4 (*yuyao*)
T.B.-23

Supreme Yang TAIYANG Ex-HN-5

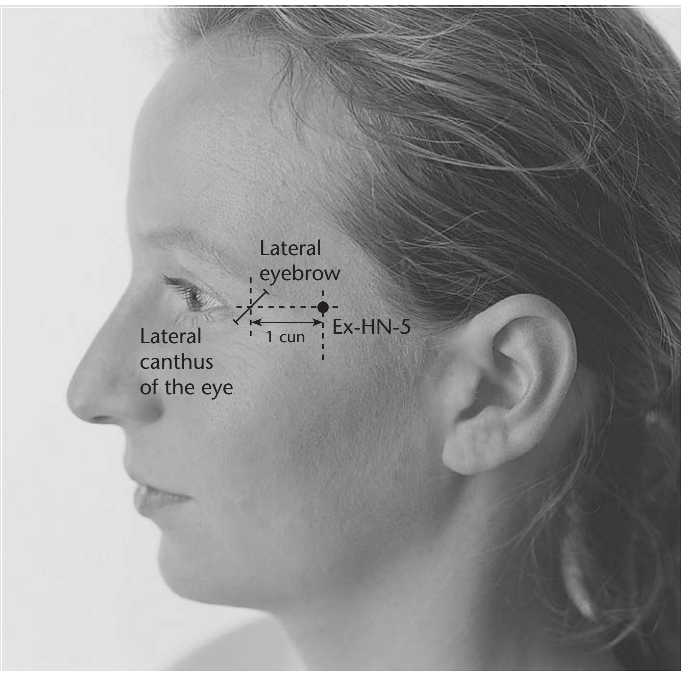
Lateral eyebrow
Lateral canthus of the eye
1 cun
Ex-HN-5

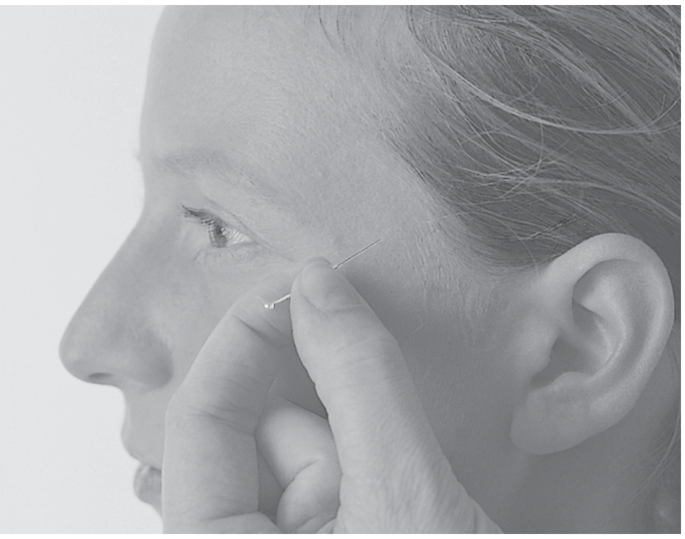

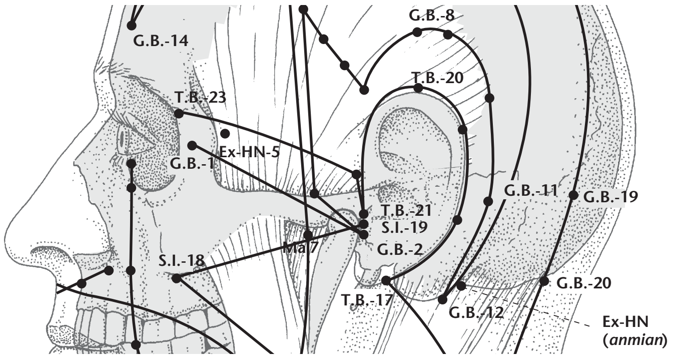
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
G.B.-2
Ma-7
S.I.-18
G.B.-20
T.B.-17
G.B.-12
Ex-HN (anmian)

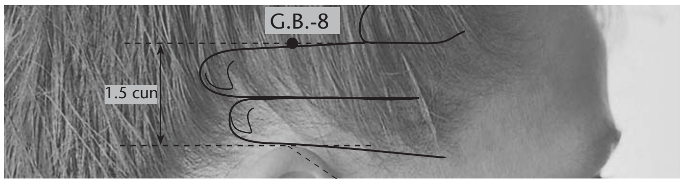

Location

On the temple, in a depression approximately 1 cun lateral to the midpoint of a line connecting the lateral extremity of the eyebrow and the outer canthus of the eye.

How to find

On the temple, first, draw a line between the lateral extremity of the eyebrow and the outer canthus of the eye. Next, palpate from the midpoint of this line in a lateral direction, until you can feel a distinct depression in the temporal bone. **Ex-HN-5** (*taiyang*) is located in the centre of this depression. This point tends to be pressure-sensitive, especially with temporal headaches. Patients often get relief by using acupressure on this point.

Needling

0.5–0.6 cun vertical insertion into the temporal muscle or obliquely in a lateral direction or transversely (subcutaneously) towards → **G.B.-8.** This point often bleeds when needled. Prick to bleed to drain Heat.

Actions/Indications

- Eliminates Wind, clears Heat, reduces swellings, opens the channel, alleviates pain

Special features

One of the most important and most commonly used extra points, especially with headaches

Ex-HN-6 *Tip of the Ear* ERJIAN

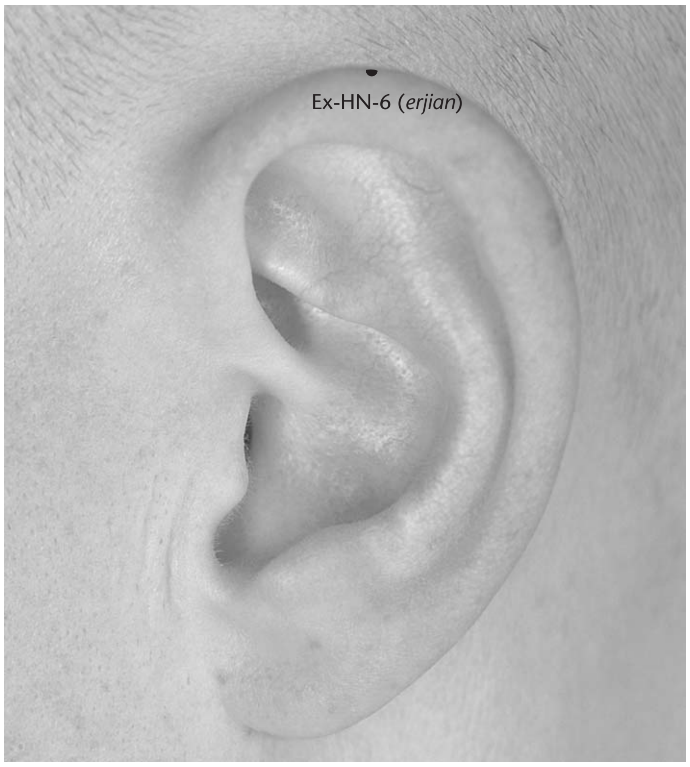

Location

On the apex of the ear, on the helix.

How to find

Fold the ear anteriorly, so that the posterior portion of the upper helix covers its anterior part. The point is located at the highest point of the fold. This is also the location of the auricular point **78**.

Needling

Vertically 0.1–0.2 cun. Some authors advise to bleed this point for Heat conditions, but this is not recommended owing to the danger of causing an otic haematoma. Indirect moxibustion for vision disorders.

Actions/Indications

- Clears Heat, alleviates pain, benefits the eyes and the throat

Behind the Ball QIUHOU Ex-HN-7

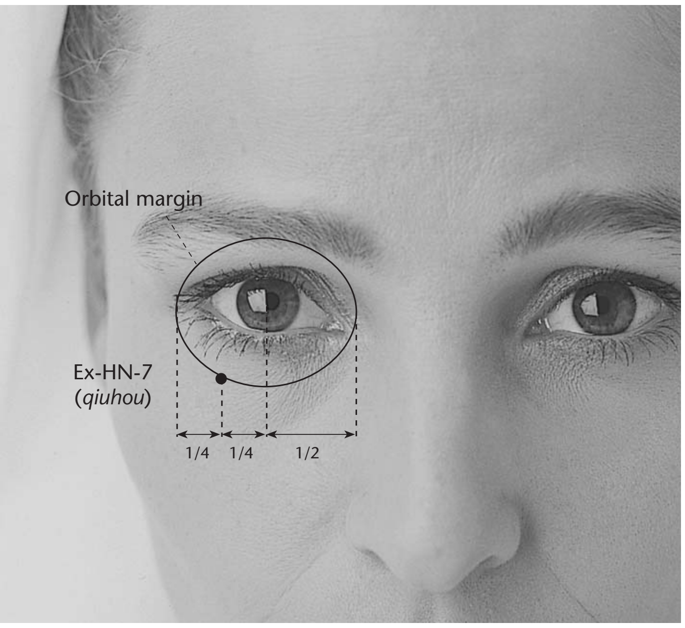
Orbital margin
Ex-HN-7
(*qiuhou*)
1/4
1/4
1/2

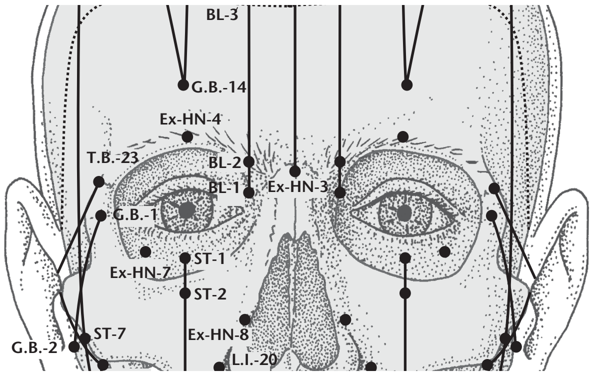
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
G.B.-2
Ex-HN-8
L.I.-20

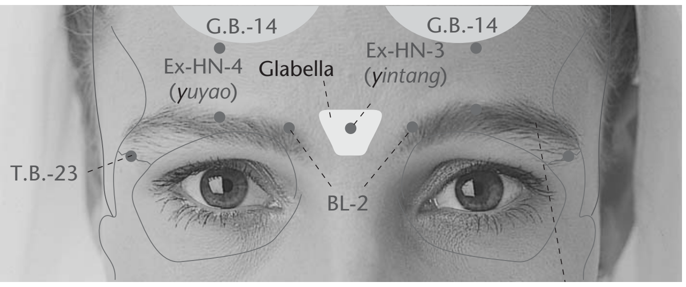
G.B.-14
G.B.-14
Ex-HN-3
(*yintang*)
Ex-HN-4
(*yuyao*)
Glabella
T.B.-23
BL-2

Location

On the lower border of the orbit, at the junction of the lateral quarter and the medial three quarters.

How to find

Divide the horizontal extension of the diameter of the orbit in quarters. Then locate the point at the junction of the first and second lateral quarters, slightly superior to the border of the orbit.

Needling

Ask the patient to look upward and gently push the eyeball upward, away from the lower eyelid. Slowly insert the needle 0.5–1 cun into the fatty tissue immediately above the bone (orbit). Caution: Do not injure the eyeball and the periosteum. Avoid the venous plexus and arteries! Pay attention to any pain from needling. No needle manipulation! After removal of the needle, compress the site for 10 minutes. Haematomas may still occur (inform the patient prior to needling). Moxibustion is contraindicated. Alternative, less difficult points for eye disorders: **BL-2, T.B.-23**, **G.B.-1**, **ST-2, Ex-HN-5** (*taiyang*), **Ex-HN-4** (*yuyao*).

Actions/Indications

- Disorders of the eyes

Ex-HN-8 *Upper* Yingxiang SHANGYINGXIANG

Location

At the upper end of the nasiolabial groove, at the junction of the maxilla and the nasal cavity. Note: *bitong* (Clear Nose) is an alternative name for this point.

How to find

At the upper end of the nasiolabial groove, palpate for the bony margin of the nasal cavity. Here, **Ex-HN-8** (*shangyingxiang* or *bitong*) is located at the junction of the nose and the cheek.

Needling

0.3–0.5 cun vertically towards the centre of the nasal cavity.

Actions/Indications

- Clears Heat, opens the nasal passages

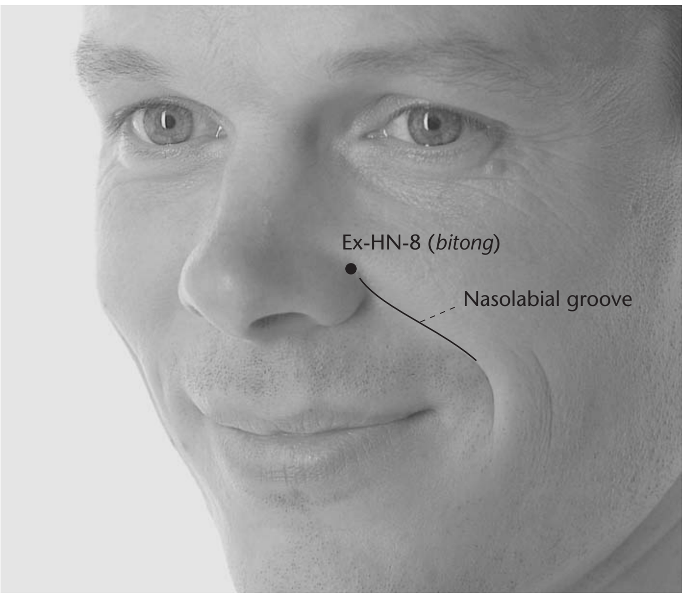
Ex-HN-8 (*bitong*)
Nasolabial groove

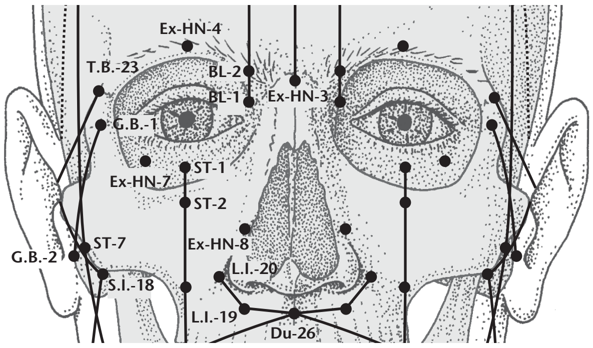
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26

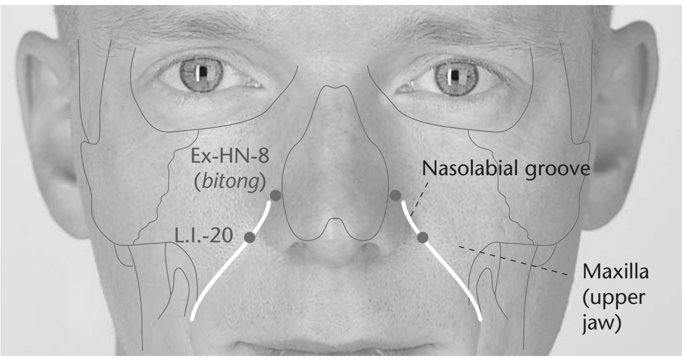
Ex-HN-8
(*bitong*)
Nasolabial groove
L.I.-20
Maxilla
(upper
jaw)

Ch06a-F10028.qxd 2/23/08 1:25 PM Page 547

Inner Yingxiang NEIYINGXIANG Ex-HN-9

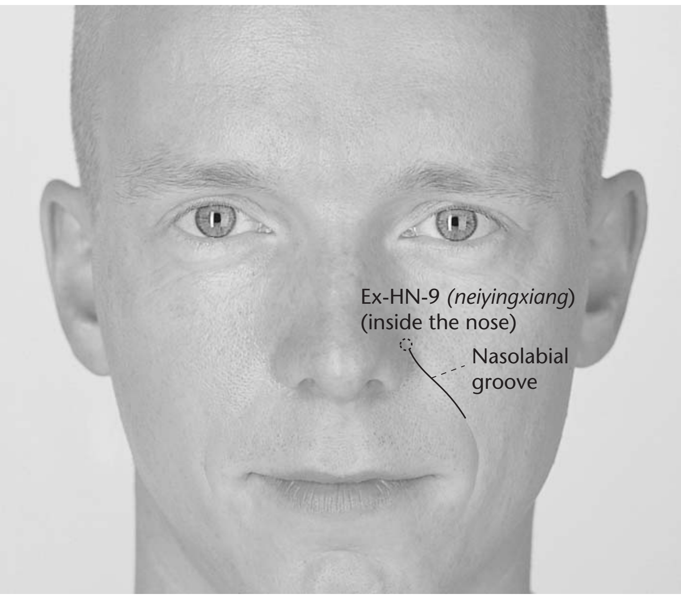
Ex-HN-9 *(neiyingxiang)*
(inside the nose)
Nasolabial groove

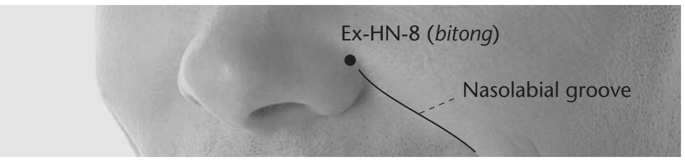
Ex-HN-8 *(bitong)*
Nasolabial groove

Location

In the nasal cavity, at the junction of the nasal bone and the nasal cartilage.

How to find

This point lies ‘opposite’ the external point ➞ **Ex-HN-8** (*shangyingxiang* or *bitong*), which is located at the upper end of the nasolabial groove.

Needling

Prick to bleed with a needle, lancet or three-edged needle. Caution: Contraindicated in patients with bleeding disorders (or taking anticoagulants). Needling may be painful!

Actions/Indications

- Clears Heat, drains Fire

Ex-HN-10 *Gathering Spring* JUQUAN

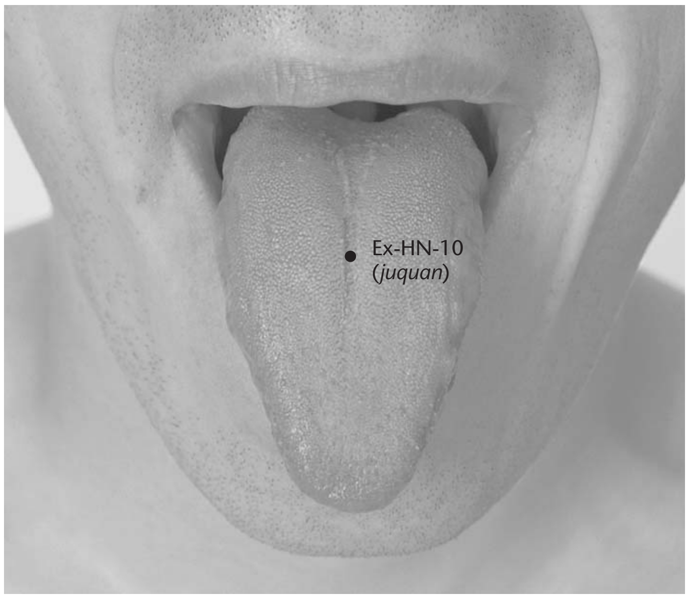

Location

With maximal extension of the tongue, in the centre of the tongue body.

How to find

With maximal extension of the tongue in the centre of the tongue body.

Needling

Vertically 0.2 cun. Needling may be painful!

Actions/Indications

- Deviation of the tongue, impaired mobility or atrophy of the musculature of the tongue: for example, after a stroke or with loss of sense of taste

Sea Spring HAIQUAN Ex-HN-11

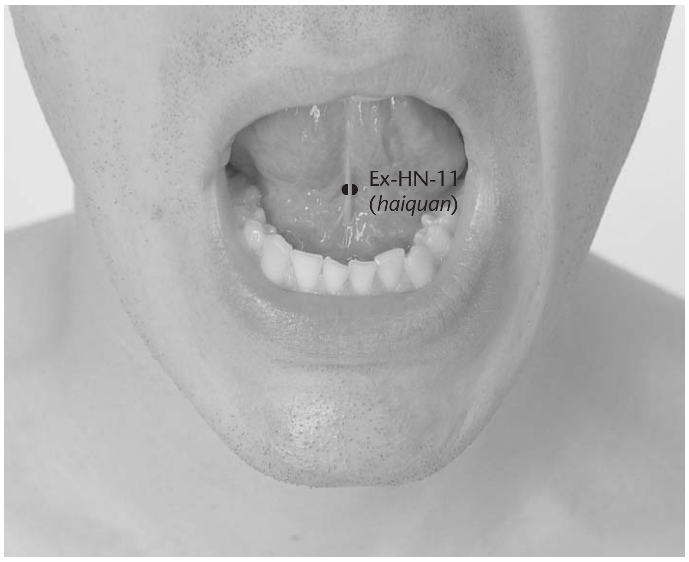
Ex-HN-11
(haiquan)

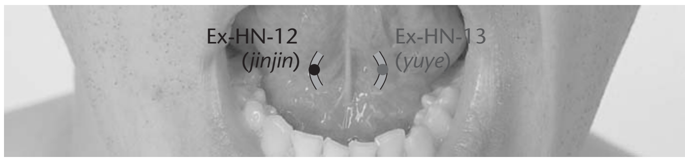
Ex-HN-12
(jinjin)
Ex-HN-13
(yuye)

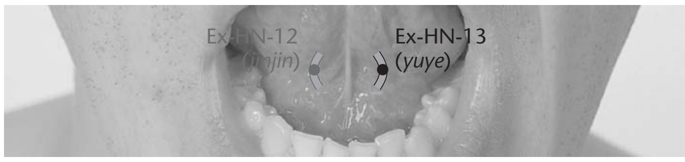
Ex-HN-12
(jinjin)
Ex-HN-13
(yuye)

Location

Below the tongue, in the centre of the frenulum, between the points ➞ **Ex-HN-12** (*jinjin*) and ➞ **Ex-HN-13** (*yuye*).

How to find

Ask the patient to roll their tongue upward to reveal the frenulum. To avoid injury to the frenulum, this point should be located at the base of the frenulum.

Needling

Vertically 0.2 cun. Remove needle immediately after short stimulation. Needling may be painful!

Actions/Indications

- Mouth and tongue ulcers, hiccups

Ex-HN-12 *Golden Liquid* JINJIN

Location

On the underside of the tongue, on the lingual vein to the left of the frenulum.

How to find

Ask the patient to roll their tongue upward to reveal the lingual veins.

Needling

Prick to bleed: Briefly puncture with a needle, lancet or three-edged needle and let bleed. Often used together with → **Ex-HN-13** (*yuye*). Caution: Contraindicated with bleeding disorders or in patients taking anticoagulant medication.

Actions/Indications

- Dispels Heat and Wind

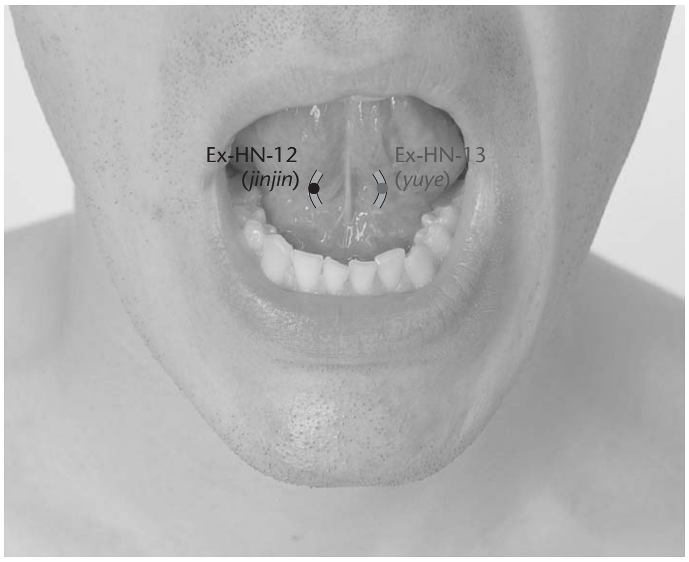

Jade Fluid YUYE Ex-HN-13

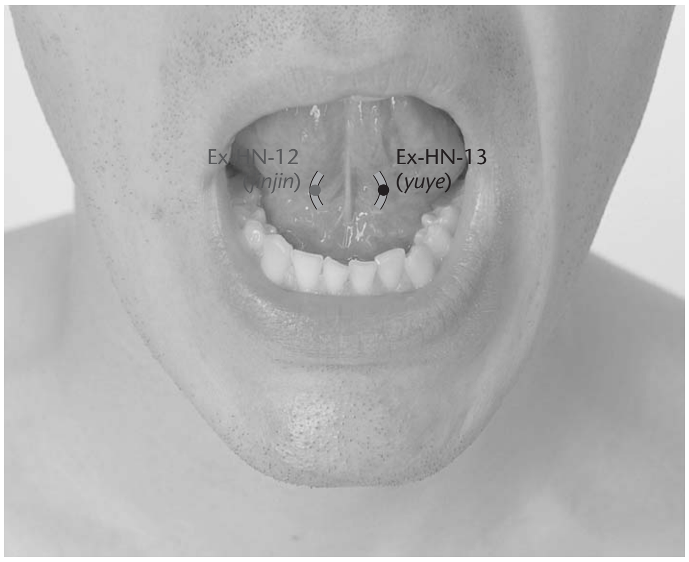

Location

On the underside of the tongue, on the great lingual vein to the right of the frenulum.

How to find

Ask the patient to roll their tongue upward to reveal the lingual veins.

Needling

Prick to bleed: Briefly puncture with a needle, lancet or three-edged needle and let bleed. Often used together with → **Ex-HN-12** (*jinjin*). Caution: Contraindicated with bleeding disorders or in patients taking anticoagulant medication.

Actions/Indications

- Dispels Heat and Wind

Ex-HN-14 *Eye Brightening* YIMING

Location

At the junction of the head and the occiput, posterior to the mastoid bone.

How to find

Ex-HN-14 (*yiming*) is located on a line connecting → **T.B.-17** (below the earlobe, between the maxilla and the mastoid) and → **G.B.-20** (at the junction of the occiput and the neck, in the centre of the dorsolateral depression). From → **T.B.-17**, measure 1 cun towards the posterior and there locate *yiming* posterior to the mastoid.

Located slightly superior and posterior to *yiming* is → **Ex-HN** (*anmian*), in the corner between the mastoid and the occiput.

Needling

Vertically 0.5–1 cun

Actions/Indications

- Disorders of the eyes and ears, dizziness, insomnia

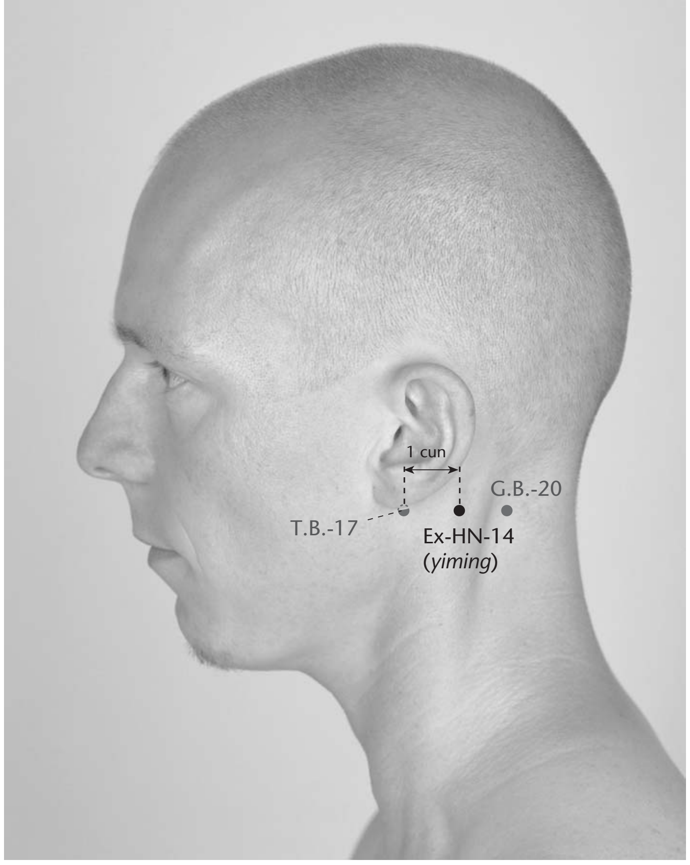
1 cun
G.B.-20
T.B.-17
Ex-HN-14
(*yiming*)

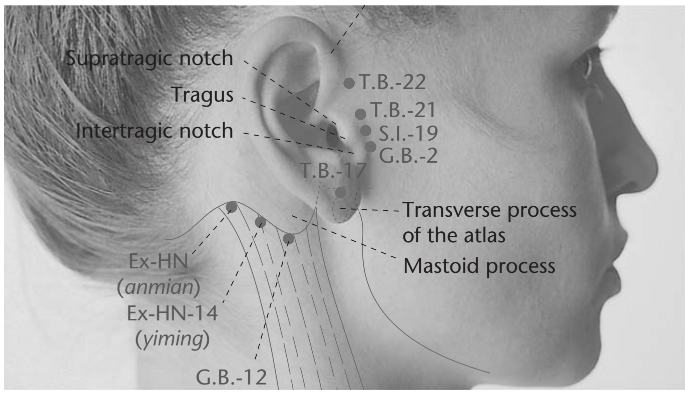
Supratragic notch
Tragus
Intertragic notch
T.B.-22
T.B.-21
S.I.-19
G.B.-2
T.B.-17
Transverse process of the atlas
Mastoid process
Ex-HN (*anmian*)
Ex-HN-14 (*yiming*)
G.B.-12

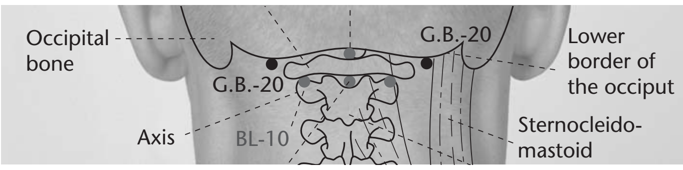
Occipital bone
G.B.-20
G.B.-20
Lower border of the occiput
Axis
BL-10
Sternocleido-mastoid

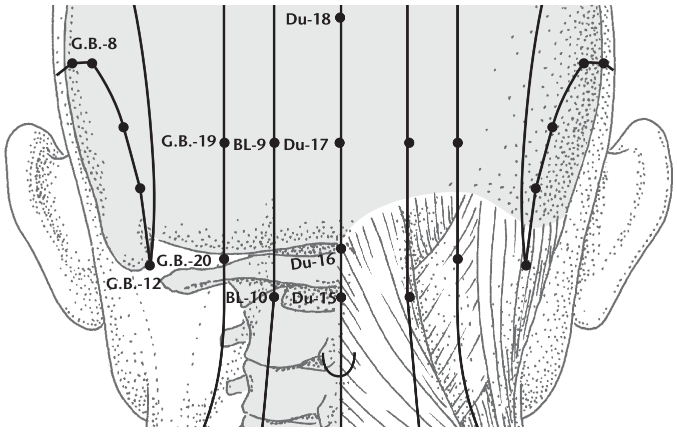
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
G.B.-20
Du-16
G.B.-12
BL-10
Du-15

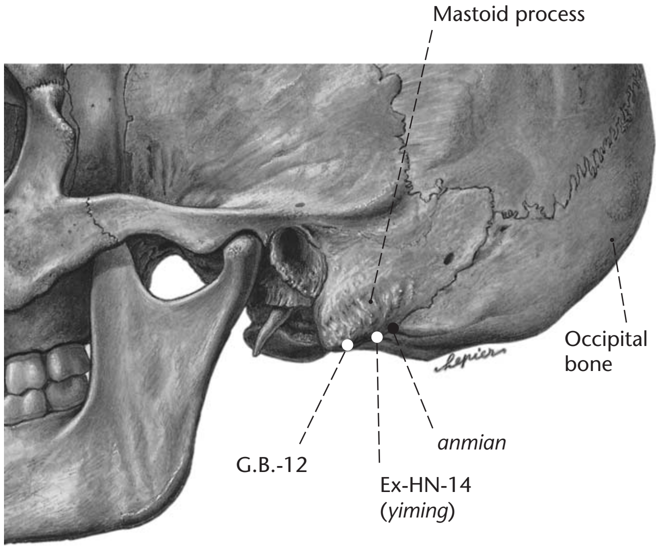
Mastoid process
Occipital bone
anmian
G.B.-12
Ex-HN-14
(*yiming*)

Hundred Taxations JINGBAILAO/BAILAO Ex-HN-15

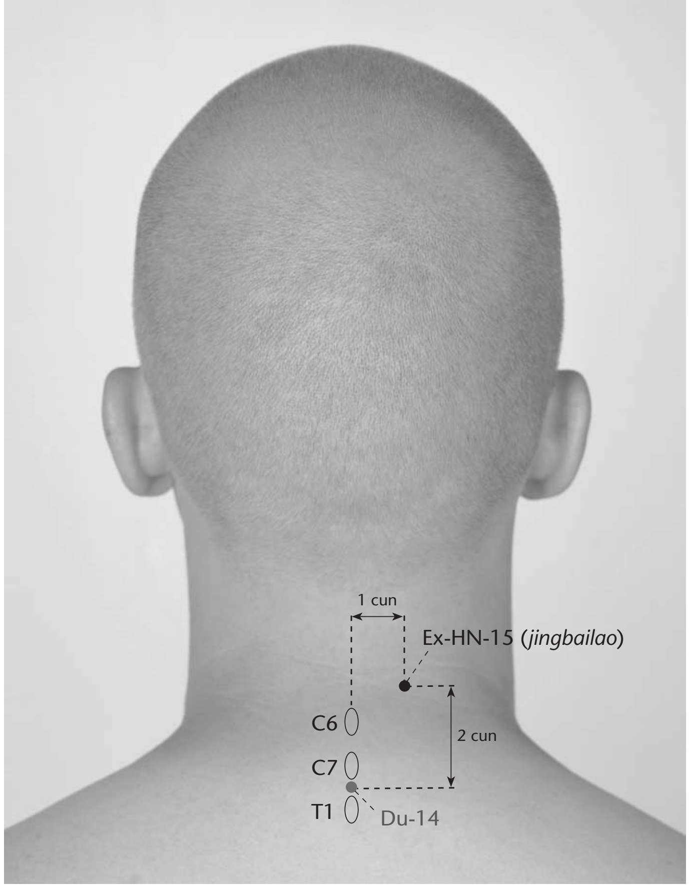
1 cun
Ex-HN-15 (*jingbailao*)
C6
2 cun
C7
T1
Du-14

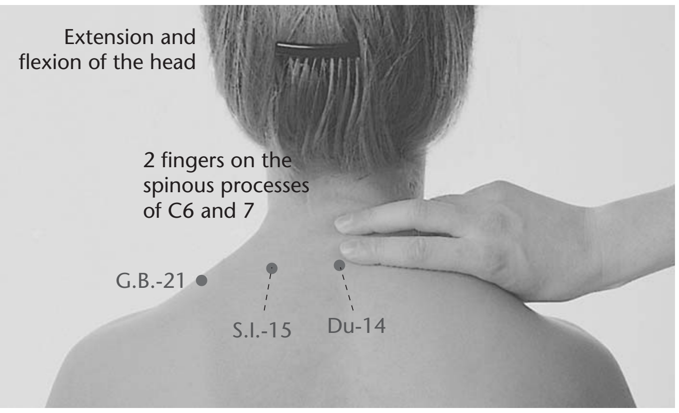
Extension and
flexion of the head
2 fingers on the
spinous processes
of C6 and 7
G.B.-21
S.I.-15
Du-14

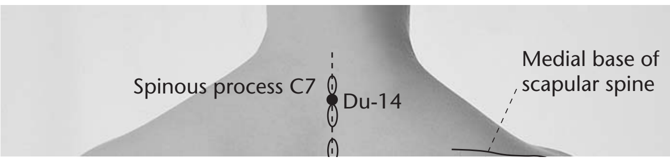
Spinous process C7
Du-14
Medial base of
scapular spine

Location

2 cun superior to the lower border of the spinous process of the 7th cervical vertebra (C7) and 1 cun lateral to the midline.

How to find

For locating C7, see → 3.4.1. → **Du-14** (*dazhui*) is located below the spinous process of C7. From there, measure 2 cun in a superior and 1 cun in a lateral direction.

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Dispels Wind and Dampness
- Opens the channel

Ex-HN *Upper Brightness* SHANGMING

Location

Directly superior to the pupil, below the margin of the orbit.

How to find

With the patient looking straight ahead, palpate for the margin of the orbit directly above the pupil. **Ex-HN** (*shangming*) is located between the upper margin of the orbit and the eyeball.

Needling

Gently push the eyeball downward. Slowly insert the needle 0.5–1 cun vertically into the fatty tissue immediately below the bone (orbit). Caution: Pay attention to any pain from needling. No manipulation! After removal of the needle, compress the site for 10 minutes. Haematomas may still occur (inform the patient prior to needling). Alternative points with less risk of complications are **BL-2, T.B.-23**, **G.B.-1, ST-2, Ex-HN-5** (*taiyang*), **Ex-HN-4** (*yuyao*).

Actions/Indications

- Disorders of the eyes

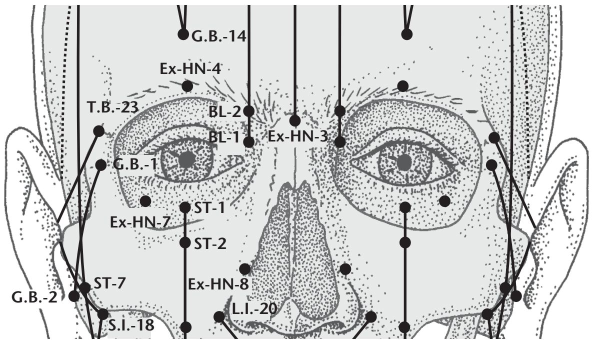
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18

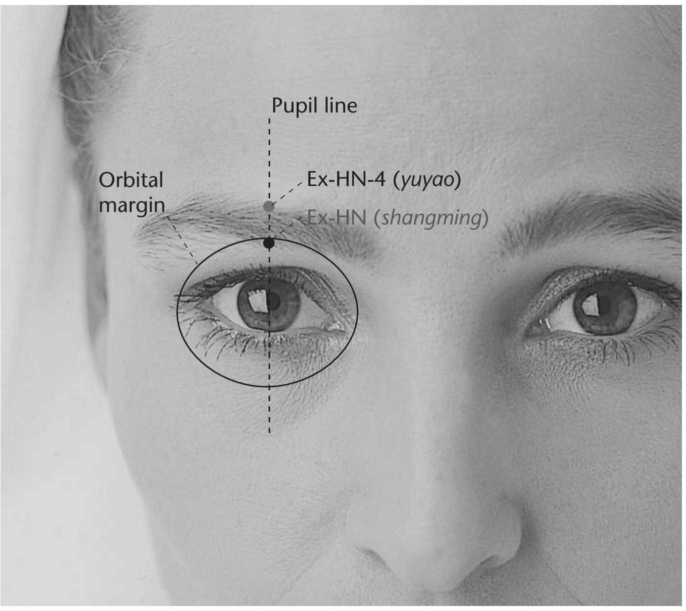
Pupil line
Orbital margin
Ex-HN-4 (yuyao)
Ex-HN (shangming)

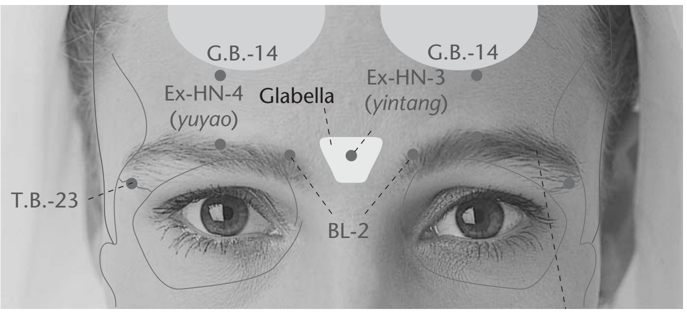
G.B.-14
G.B.-14
Ex-HN-4 (yuyao)
Glabella
Ex-HN-3 (yintang)
T.B.-23
BL-2

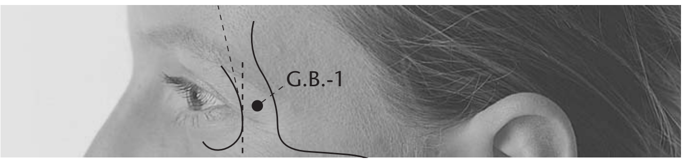

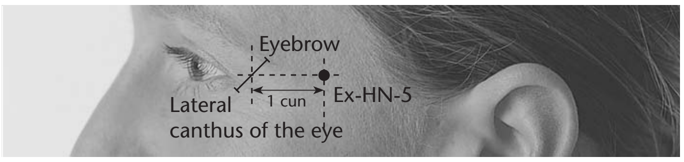
Eyebrow
Lateral canthus of the eye
1 cun
Ex-HN-5

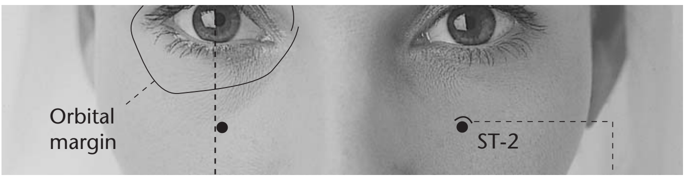
Orbital margin
ST-2

Peaceful Sleep ANMIAN Ex-HN

Location

Posterior to the ear, between ➞ **T.B.-17** and ➞ **G.B.-20**, posterior to the mastoid process.

How to find

First, locate ➞ **T.B.-17** (directly below the earlobe, in the depression between the lower jaw and the mastoid process (➞ 3.1) that forms when the mouth is open). Next, locate ➞ **G.B.-20** on the lower border of the occiput, in the depression between the origins of the sternocleidomastoid and trapezius muscles. Locate **Ex-HN** (*anmian*) approximately midway between these two points, slightly posterior to the mastoid bone and superior to ➞ **G.B.-12**, in a depression on the lower border of the occiput.

Needling

0.5–1 cun vertically or obliquely towards ➞ **T.B.-17** (*yifeng*) or ➞ **G.B.-20** (*fengchi*).

Actions/Indications

- Calms the *shen*

Special features

Important calming point for sleeping disorders. This extra point is not part of the WHO classification of extra points.

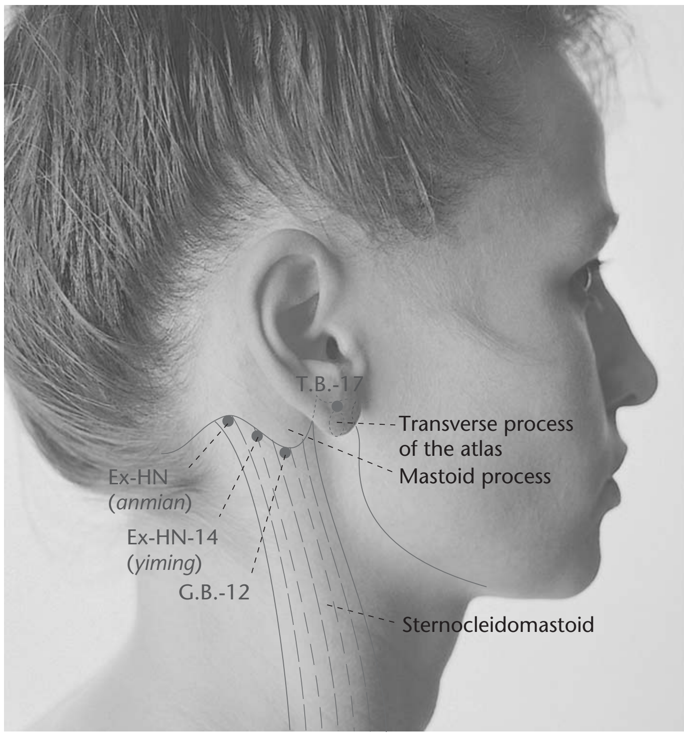
T.B.-17
Transverse process of the atlas
Mastoid process
Ex-HN (*anmian*)
Ex-HN-14 (*yiming*)
G.B.-12
Sternocleidomastoid

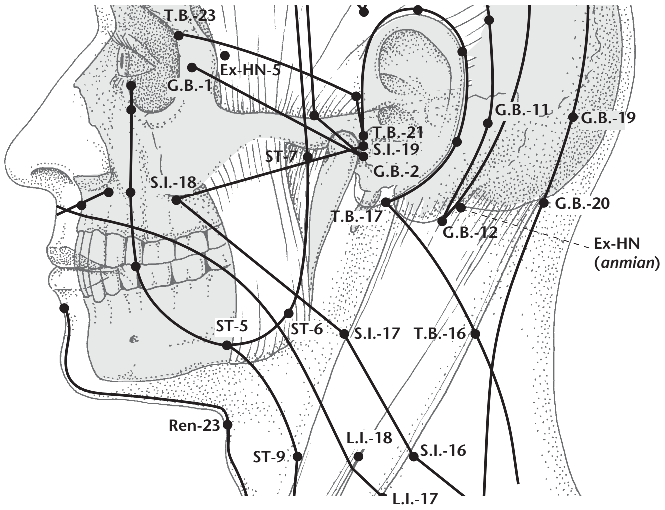
T.B.-23
Ex-HN-5
G.B.-1
T.B.-21
S.I.-19
G.B.-2
G.B.-11
G.B.-19
ST-7
S.I.-18
T.B.-17
G.B.-12
G.B.-20
Ex-HN (*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17

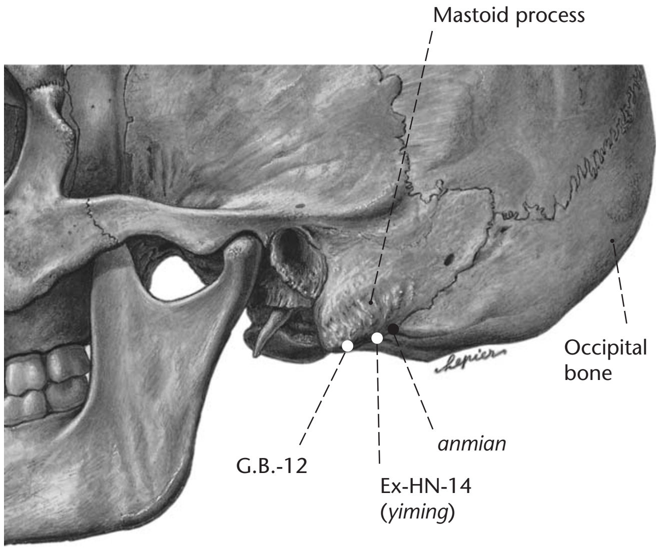
Mastoid process
Occipital bone
anmian
G.B.-12
Ex-HN-14 (*yiming*)

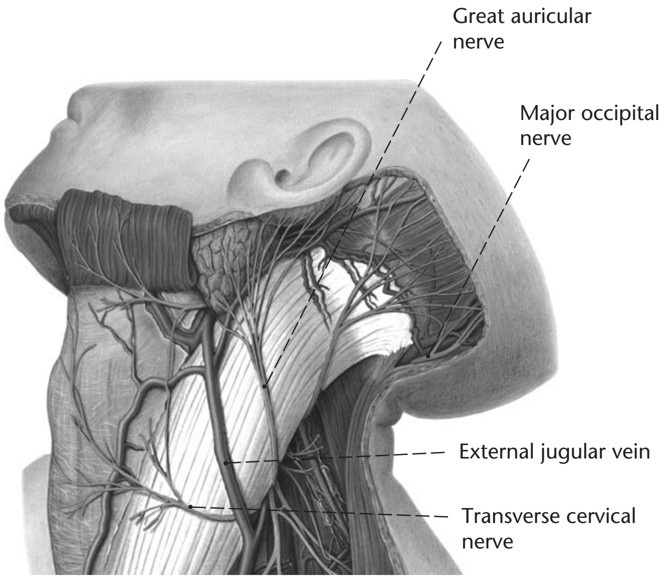
Great auricular nerve
Major occipital nerve
External jugular vein
Transverse cervical nerve

Ex-HN *Cheek Centre* JIALI

Location

Inside the mouth, on the mucosa of the cheek, 1 cun posterior to the corner of the mouth.

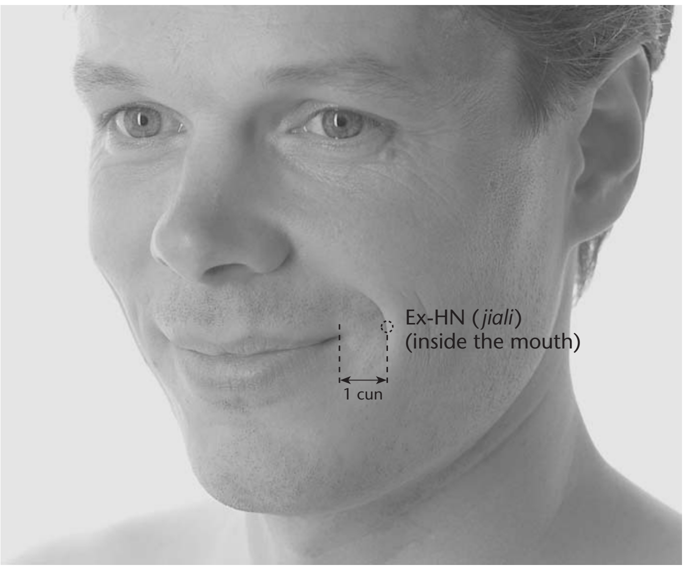

How to find

Locate this point on the mucosa of the cheek, 1 cun posterior to the corner of the mouth.

Needling

0.3–0.5 cun obliquely in a posterior direction, prick to bleed.

Actions/Indications

- Clears Heat
- For inflammations of the mouth and throat, for gastritis

Adjacent to Container of Fluids JIACHENGJIANG Ex-HN

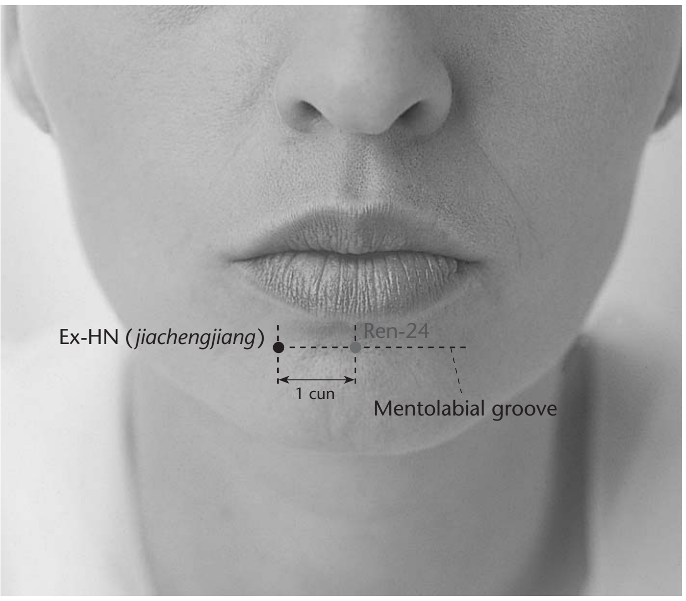
Ex-HN (*jiachengjiang*)
Ren-24
1 cun
Mentolabial groove

Location

1 cun lateral to the centre of the mentolabial groove.

How to find

This point is located on the mental foramen. → **Ren-24** (*chengjiang*) is located in the centre of the mentolabial groove.

Needling

0.3–0.5 cun vertically or up to 1 cun obliquely towards the mental foramen.

Actions/Indications

- Dispels Wind, opens the channel

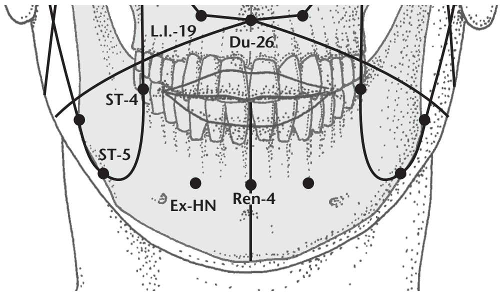
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-4

Ex-HN *Prominent Bone* CHONGGU ZHUIDONG

Location

Below the spinous process of the 6th cervical vertebra (C6).

How to find

This point is located on the posterior midline and thus on the *du mai*. In fact, the 'prominent bone' (vertebra prominens) inferior to *chonggu zhuidong* is the spinous process of either C7 or T1. → **Du-14** is located between the two. For the correct method of locating C6 and C7 → 3.4.1.

Needling

0.5–1 cun obliquely in a superior direction.

Actions/Indications

- Dispels external pathogenic factors
- Harmonises the *shen*

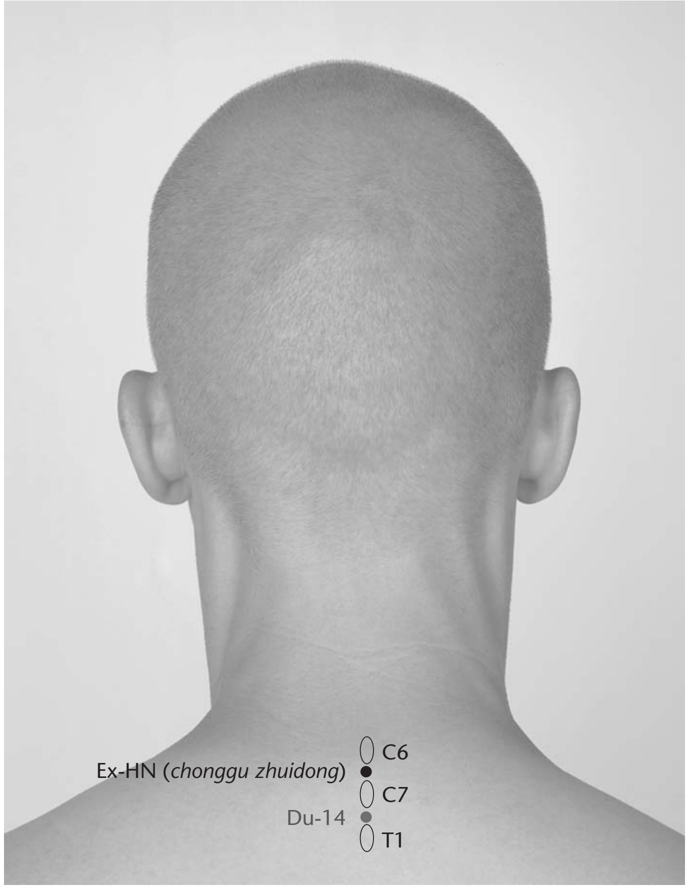
Ex-HN (*chonggu zhuidong*)
C6
C7
Du-14
T1

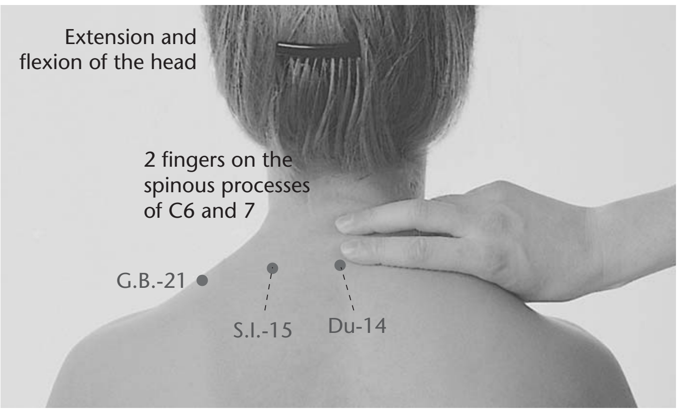
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Upper Arm JINGBI **Ex-HN**

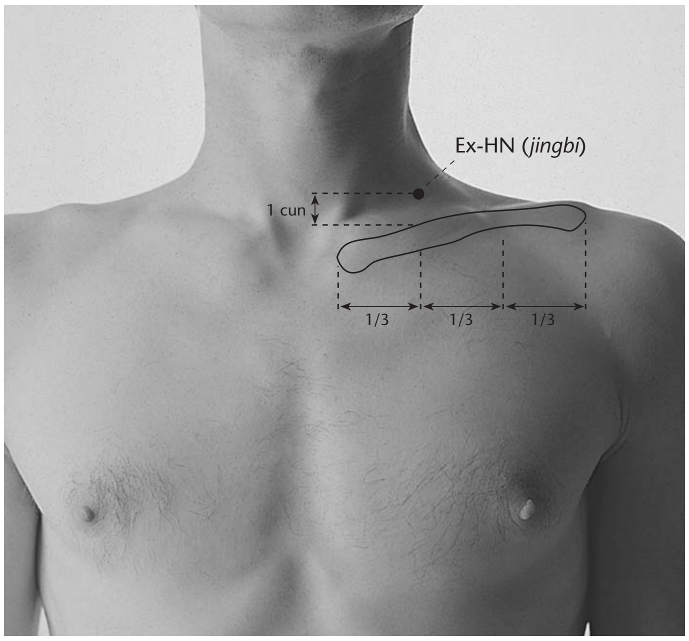
Ex-HN (*jingbi*)
1 cun
1/3
1/3
1/3

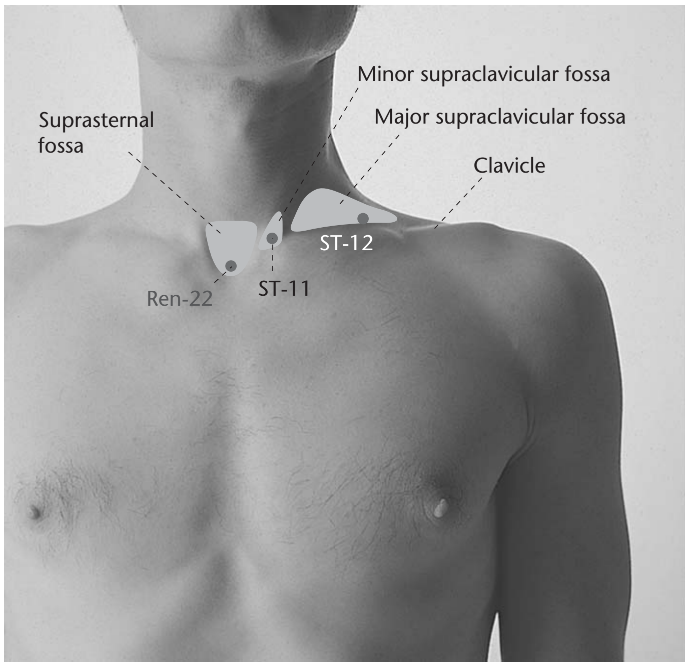
Minor supraclavicular fossa
Major supraclavicular fossa
Suprasternal fossa
Clavicle
ST-12
ST-11
Ren-22

Location

1 cun superior to the junction of the proximal and middle third of the clavicle.

How to find

Measure a third of the clavicle from its medial end, then measure 1 cun in a superior direction and there locate *jingbi* in the supraclavicular fossa. The point is located above the brachial plexus. → **ST-12** is very close to this point, directly superior to the midpoint of the clavicle.

Needling

Vertically 0.3–0.5 cun. During insertion or stimulation, a tingling or warm sensation should be felt radiating to the fingers. Caution: Pneumothorax.

Actions/Indications

- Paraesthesia and paralysis of the upper extremity

6.2 Extra Points: Chest and Abdomen (EX-CA)

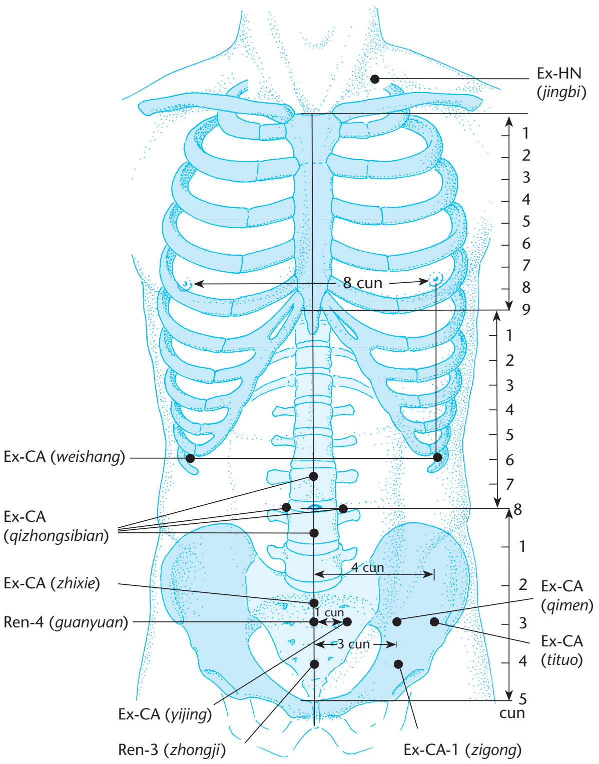
Ex-HN
(*jingbi*)
1
2
3
4
5
6
7
8
9
8 cun
1
2
3
4
5
6
7
8
Ex-CA (*weishang*)
Ex-CA
(*qizhongsibian*)
Ex-CA (*zhixie*)
Ren-4 (*guanyuan*)
4 cun
1 cun
3 cun
1
2
3
4
5
cun
Ex-CA
(*qimen*)
Ex-CA
(*tituo*)
Ex-CA (*yijing*)
Ren-3 (*zhongji*)
Ex-CA-1 (*zigong*)

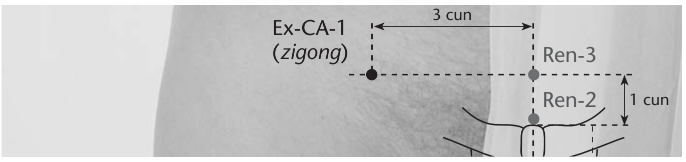
Ex-CA-1
(*zigong*)
3 cun
Ren-3
Ren-2
1 cun

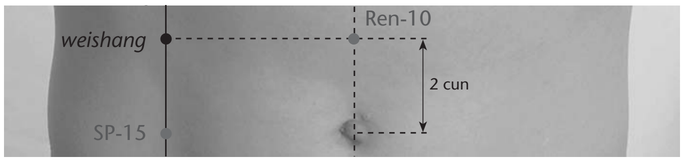
weishang
Ren-10
2 cun
SP-15

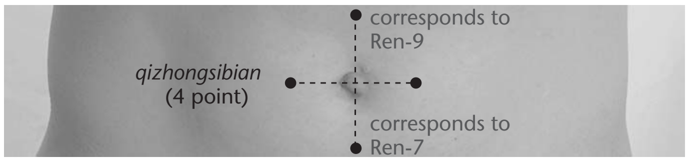
corresponds to
Ren-9
qizhongsibian
(4 point)
corresponds to
Ren-7

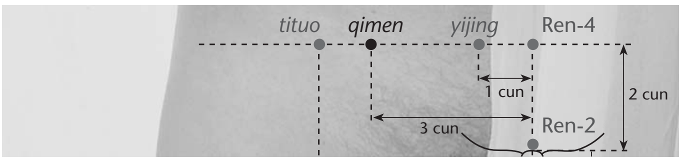
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

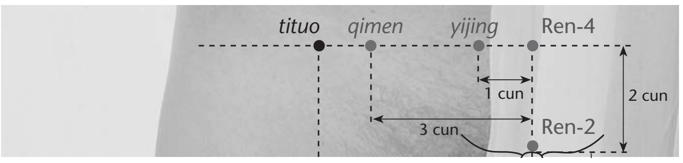
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

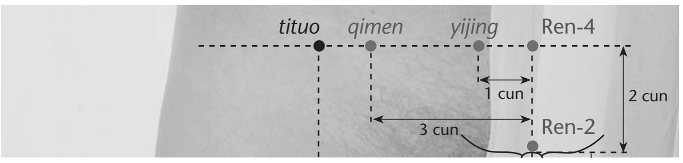
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-CA-1	*zigong*	PaM or ZP49	M-CA-18	Ex9
Further extra points				
Ex-CA	*weishang*			
Ex-CA	*qizhongsibian*			
Ex-CA	*yijing*	PaM or ZP45		
Ex-CA	*qimen*	PaM or ZP46		
Ex-CA	*tituo*	NP39	M-CA-4	
Ex-CA	*zhixie*	NP38	N-CA31	

Palace of the Child ZIGONG Ex-CA-1

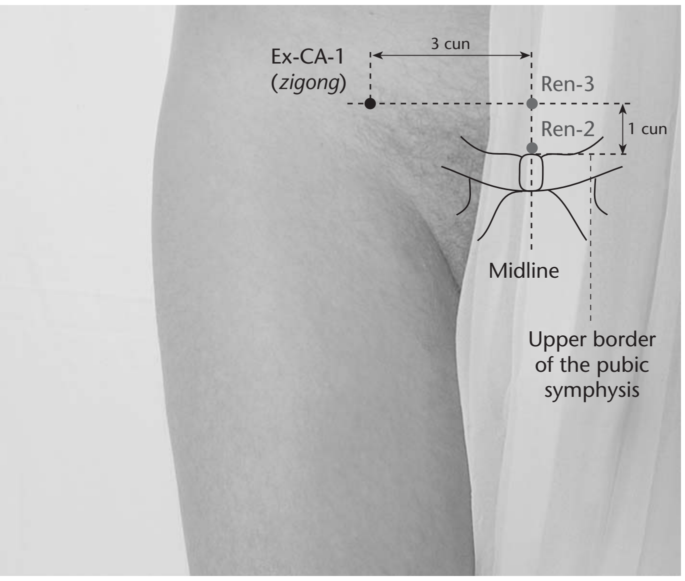
Ex-CA-1
(*zigong*)
3 cun
Ren-3
Ren-2
1 cun
Midline
Upper border
of the pubic
symphysis

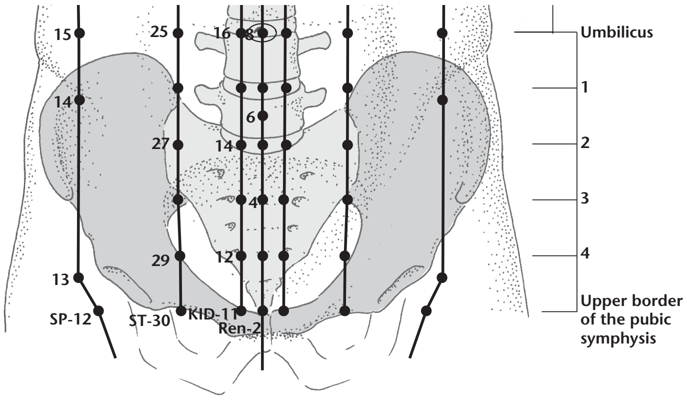
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

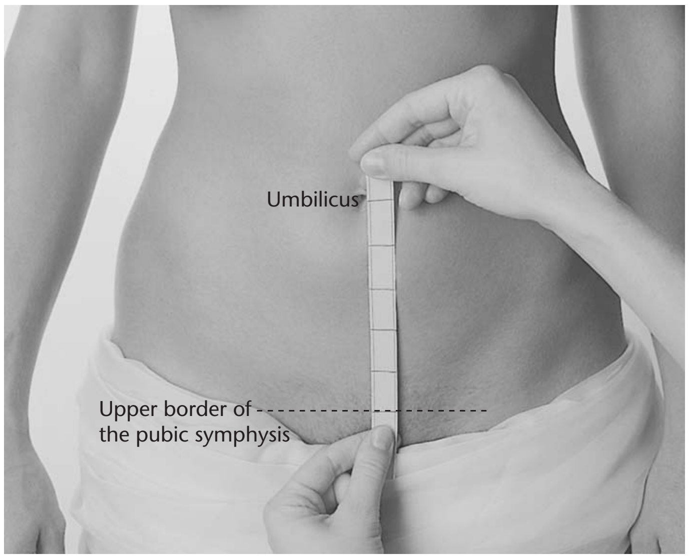
Umbilicus
Upper border of
the pubic symphysis

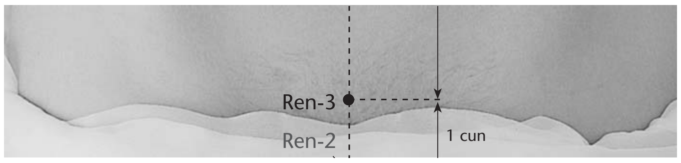
Ren-3
Ren-2
1 cun

Location

3 cun lateral to the anterior midline and 1 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 1 cun in a superior direction – this is the location of → **Ren-3**. From there, measure 3 cun in a lateral direction to locate *zigong*.

Located on the same level are → **KID-12** (0.5 cun from the midline) and → **ST-29** (2 cun lateral to the midline).

Needling

0.5–1 cun vertically or 1–2 cun obliquely towards the upper border of the pubic symphysis. Caution: Peritoneum, pregnancy, full bladder.

Actions/Indications

- Strengthens and raises Qi
- Regulates menstruation
- Alleviates pain in the abdomen and the lumbar region

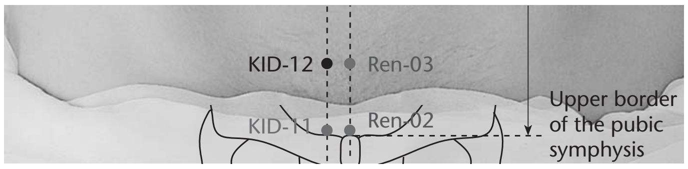
KID-12
Ren-03
KID-11
Ren-02
Upper border
of the pubic
symphysis

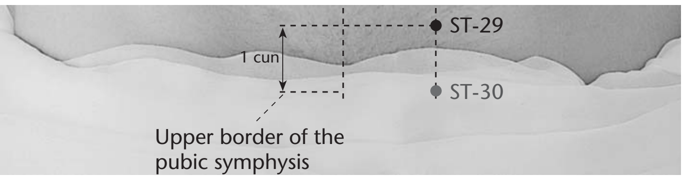
ST-29
1 cun
ST-30
Upper border of the
pubic symphysis

Ch06a-F10028.qxd 2/23/08 1:25 PM Page 564

Ex-CA *Lifting the Stomach* WEISHANG

Location

On the Spleen channel, 4 cun lateral and 2 cun superior to the umbilicus.

How to find

The distance between the sternocostal angle and the centre of the umbilicus is divided into 8 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the centre of the umbilicus, measure 2 cun in a superior direction (location of → **Ren-10**) and then locate *weishang* 4 cun lateral to the midline. 1 cun superior to → **Ex-CA** (*weishang*) is → **SP-16**.
Located on the same level are → **Ren-10/KID-17/ST-23** (on the midline/0.5 cun lateral to the midline/2 cun lateral to the midline).

Needling

2–3 cun obliquely in the direction of the umbilicus. Caution: Peritoneum, pregnancy.

Actions/Indications

- Gastroptosis, abdominal pain

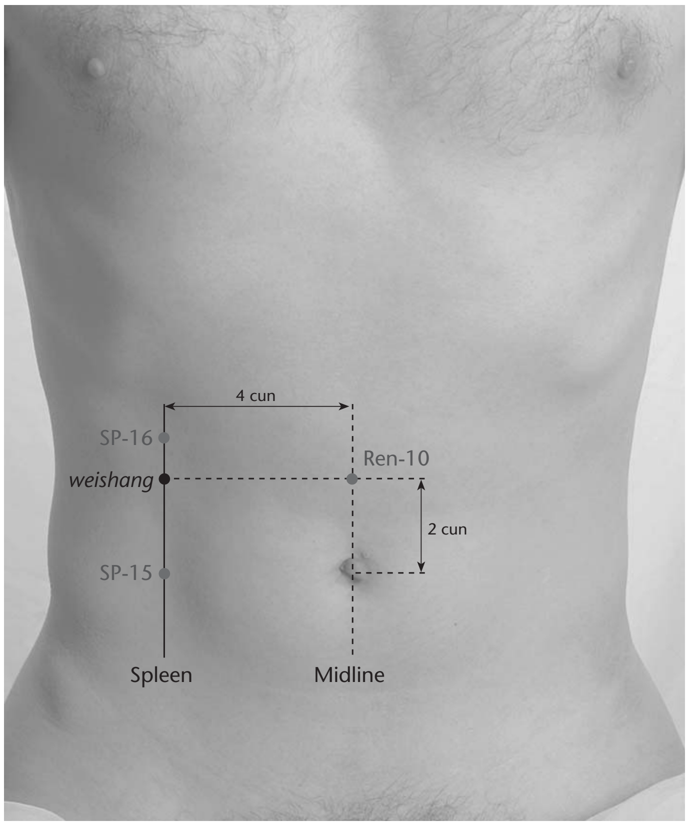
4 cun
SP-16
Ren-10
weishang
2 cun
SP-15
Spleen
Midline

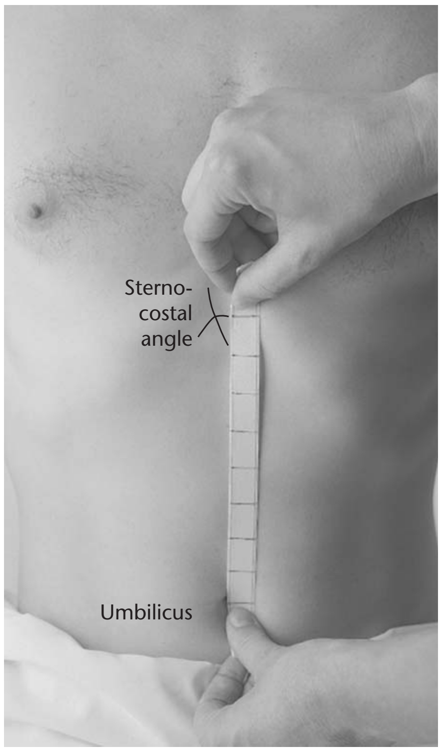
Sterno-
costal
angle
Umbilicus

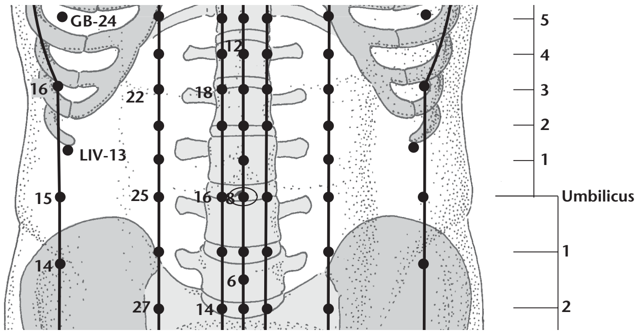
GB-24
12
16
22
18
LIV-13
15
25
16
8
14
6
27
14
5
4
3
2
1
Umbilicus
1
2

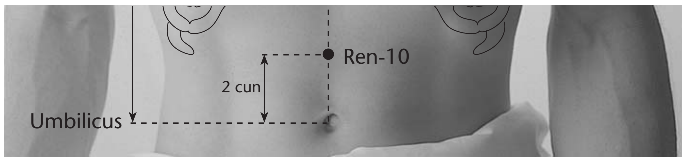
Ren-10
2 cun
Umbilicus

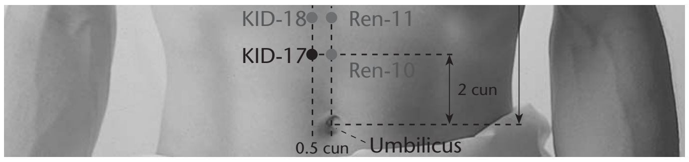
KID-18
Ren-11
KID-17
Ren-10
2 cun
0.5 cun
Umbilicus

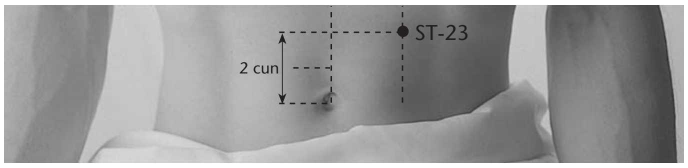

Four Points Around the Umbilicus QIZHONGSIBIAN Ex-CA

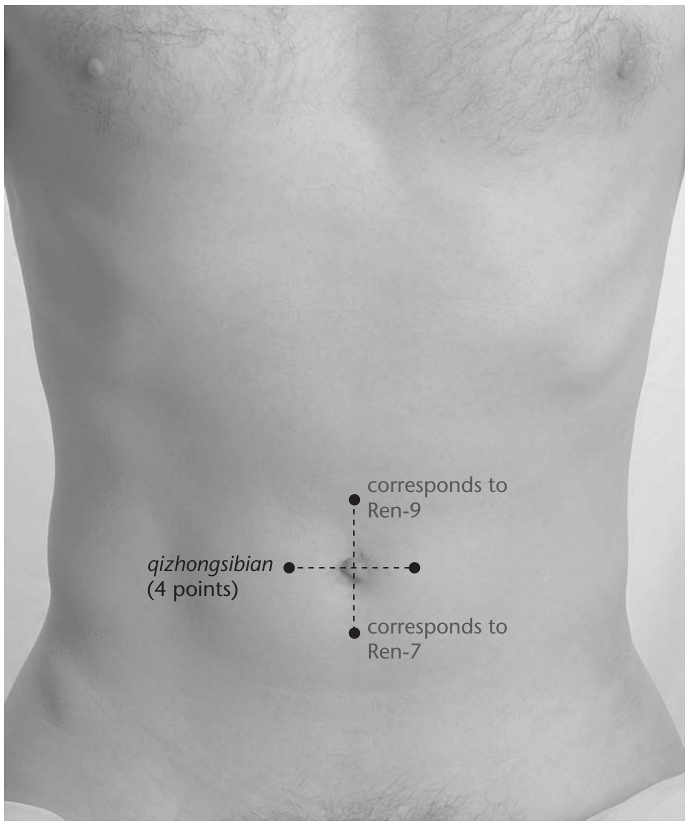
corresponds to
Ren-9
qizhongsibian
(4 points)
corresponds to
Ren-7

Location

Four points, 1 cun lateral, superior and inferior to the umbilicus.

How to find

These four points are arranged in a star shape around the umbilicus.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Distension, diarrhoea, dyspepsia, dysmenorrhoea

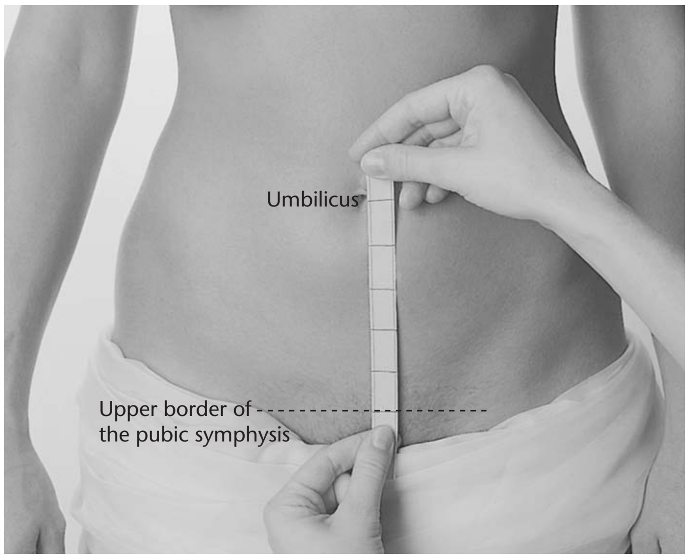
Umbilicus
Upper border of
the pubic symphysis

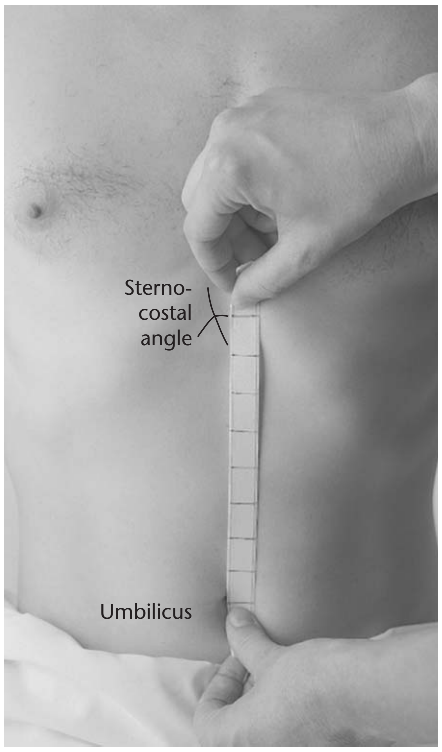
Sterno-
costal
angle
Umbilicus

Ex-CA *Loss of Semen* YIJING

Location

1 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and, from there, measure 1 cun in a lateral direction. There, locate *yijing*.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), **Ex-CA** (*qimen*, 3 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and **Ex-CA** (*tituo*, 4 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

Ejaculation disorders, impotence, scrotal eczema

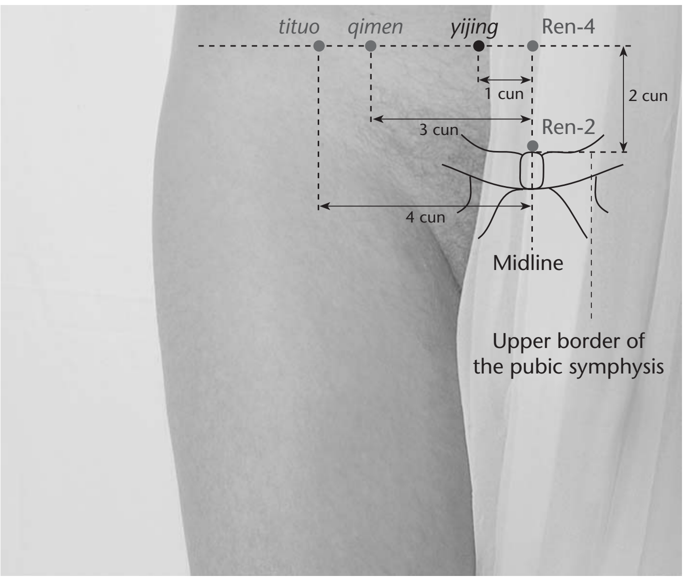
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of
the pubic symphysis

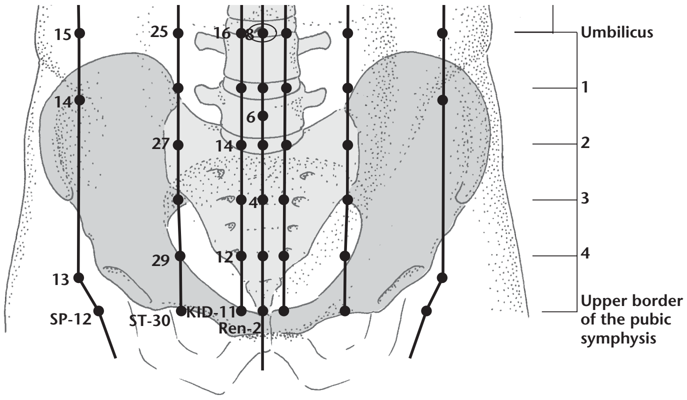
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border
of the pubic
symphysis

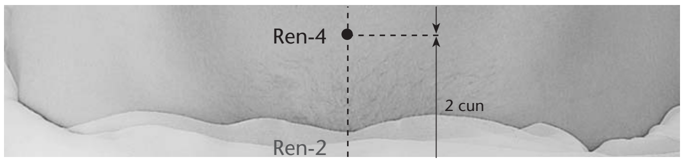
Ren-4
2 cun
Ren-2

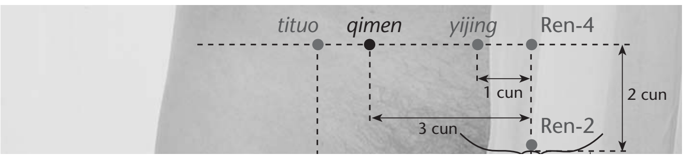
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2

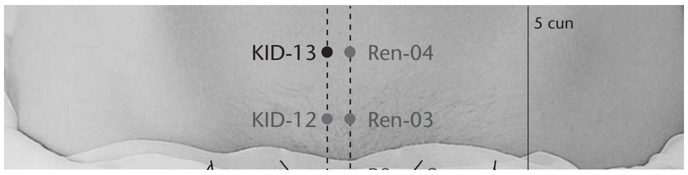
5 cun
KID-13
Ren-04
KID-12
Ren-03

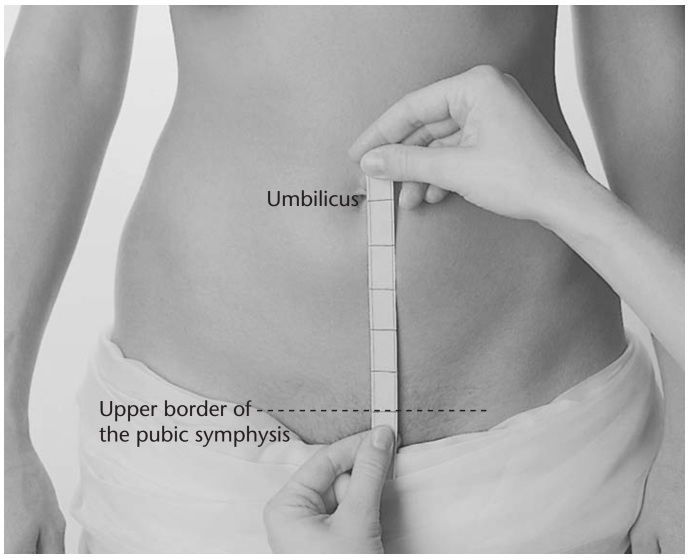
Umbilicus
Upper border of
the pubic symphysis

Qi Gate QIMEN **Ex-CA**

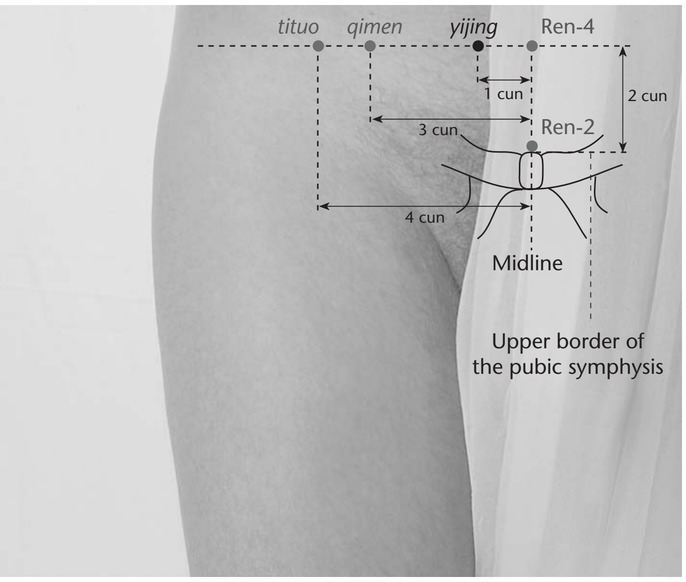
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

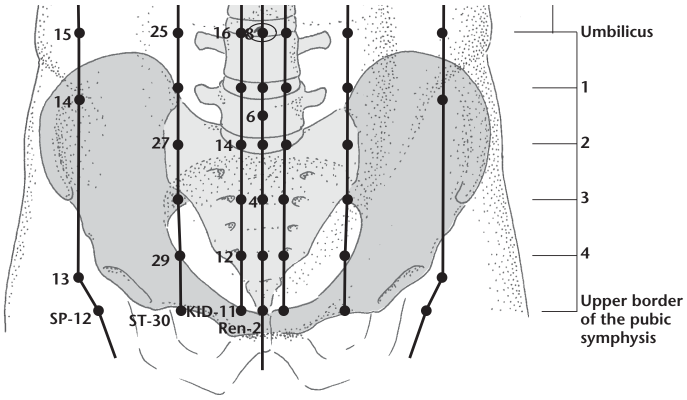
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

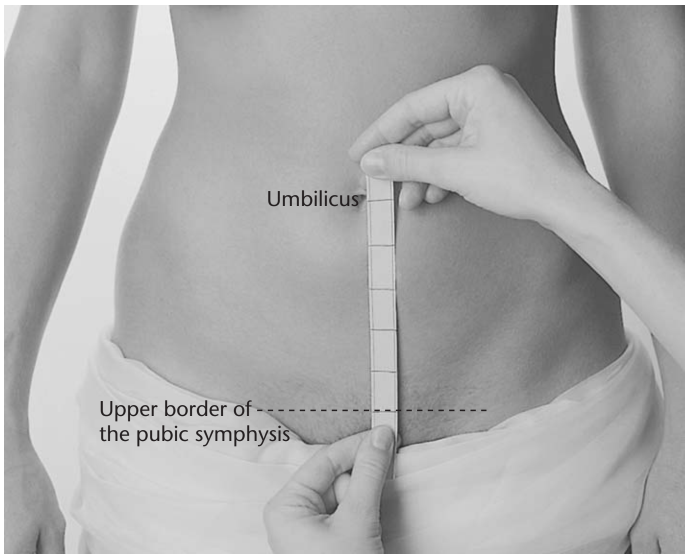
Umbilicus
Upper border of the pubic symphysis

Location

3 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and, from there, measure 3 cun in a lateral direction. There, locate *qimen*.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), **Ex-CA** (*yijing*, 1 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and **Ex-CA** (*tituo*, 4 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Metrorrhagia, female infertility, orchitis, urinary tract infections, persistent lochial discharge

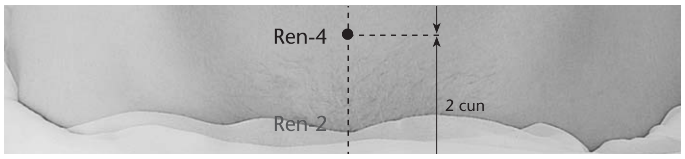
Ren-4
2 cun
Ren-2

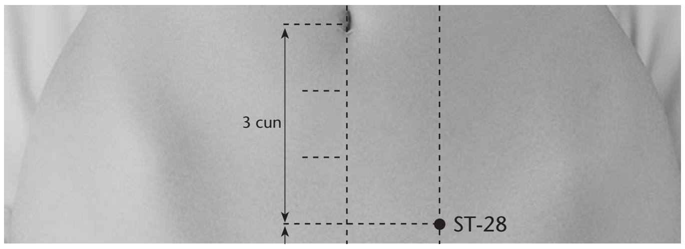

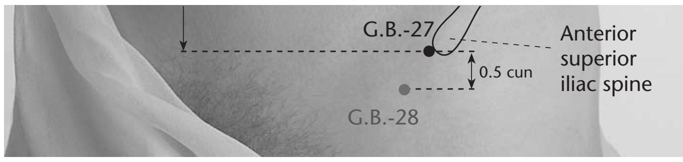
G.B.-27
0.5 cun
G.B.-28
Anterior superior iliac spine

Ex-CA Lift and Support TITUO

Location

4 cun lateral to the anterior midline and 2 cun superior to the upper border of the pubic symphysis.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape → 2.3.1). From the upper border of the pubic symphysis, measure 2 cun in a cranial direction (→ **Ren-4**) and from there measure 4 cun in a lateral direction. There, medial to the ASIS, locate *tituo*.

Located on the same level are → **KID-13** (0.5 cun lateral to the midline), → **Ex-CA** (*yijing*, 1 cun lateral to the midline), → **ST-28** (2 cun lateral to the midline) and → **Ex-CA** (*qimen*, 3 cun lateral to the midline). → **G.B.-27** is located medial to the ASIS, approximately on the same level.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, pregnancy.

Actions/Indications

- Strengthens the rising Qi and alleviates organ prolapse

Special features

In modern TCM, *tituo* is considered one of the most important points for prolapse of the uterus.

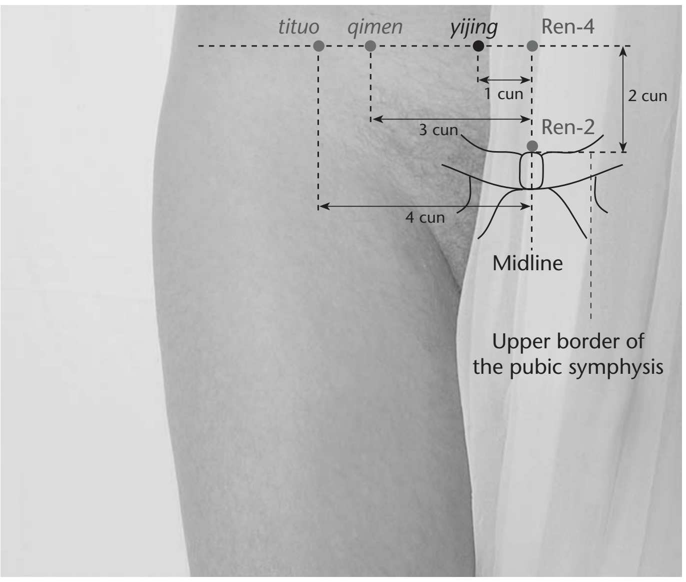
tituo
qimen
yijing
Ren-4
1 cun
2 cun
3 cun
Ren-2
4 cun
Midline
Upper border of the pubic symphysis

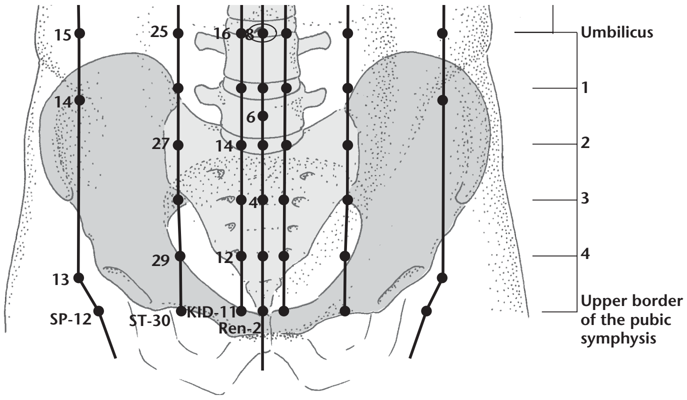
15
25
16
8
14
6
27
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

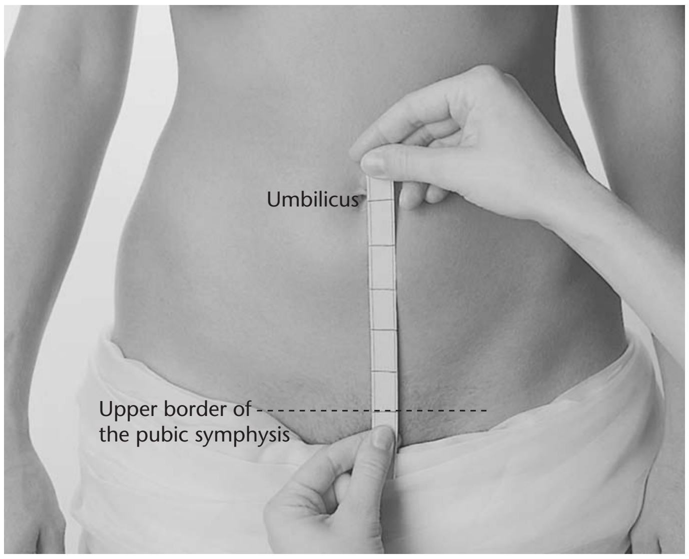
Umbilicus
Upper border of the pubic symphysis

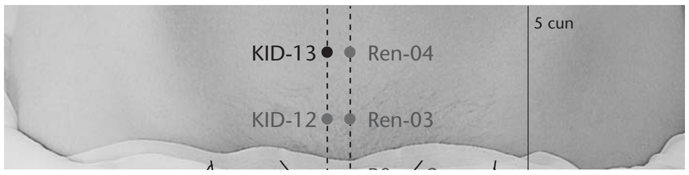
KID-13
Ren-04
5 cun
KID-12
Ren-03

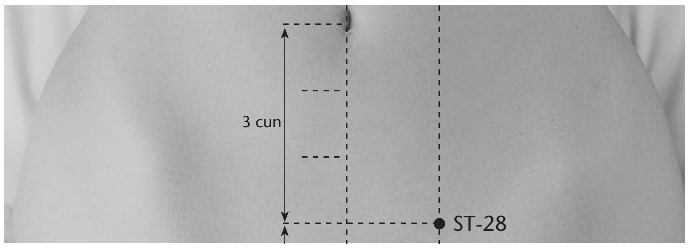

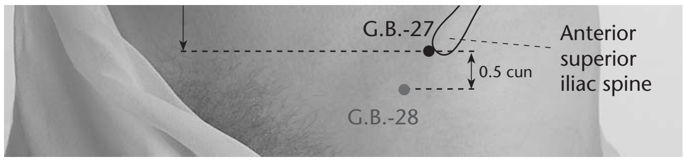
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

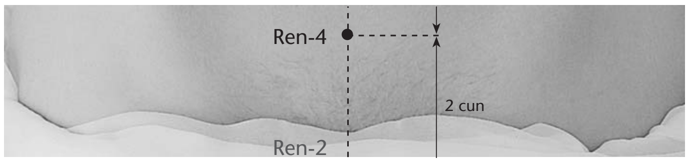
Ren-4
2 cun
Ren-2

End Diarrhoea ZHIXIE Ex-CA

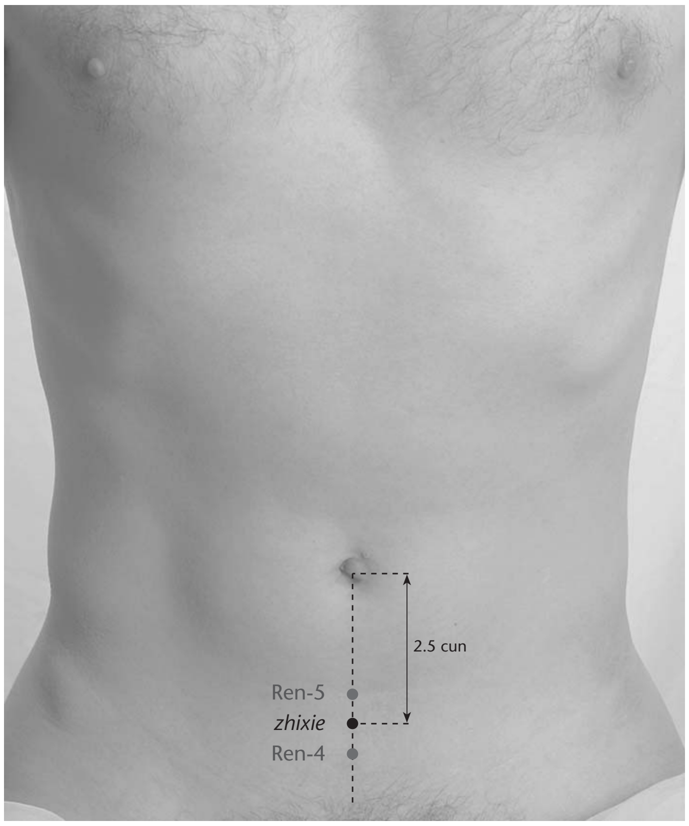

Location

On the anterior midline, 2.5 cun inferior to the umbilicus.

How to find

The distance between the centre of the umbilicus and the upper border of the pubic symphysis is divided into 5 proportional cun, which can vary considerably from the patient's finger cun. Therefore, only proportional cun measurements should be used (helpful device: elastic tape ➞ 2.3.1). *Zhixie* is located at the midpoint of this distance, between ➞ **Ren-4** (2 cun superior to the upper border of the pubic symphysis) and ➞ **Ren-5** (2 cun inferior to the umbilicus). **Or:** Spreading hands technique (➞ 2.3.3): Place the little fingers on the umbilicus and the upper border of the pubic symphysis respectively and spread the hands evenly, joining the thumbs at the midpoint of this distance.

Needling

Vertically 0.5–1 cun. Caution: Peritoneum, full bladder, pregnancy.

Actions/Indications

- Stops diarrhoea

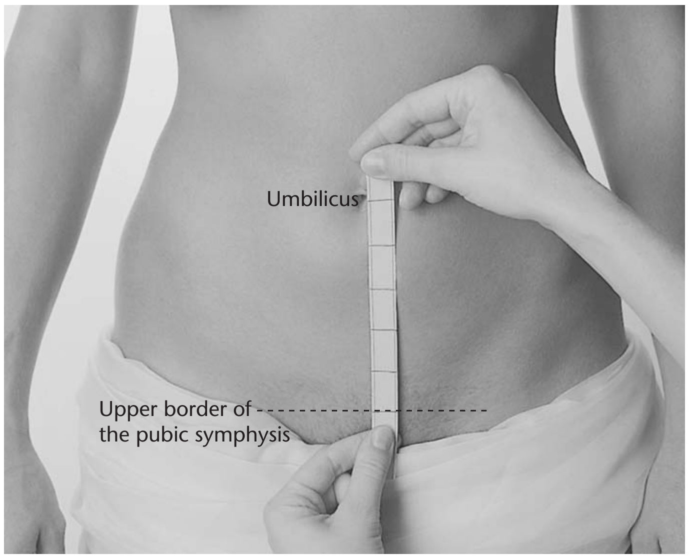

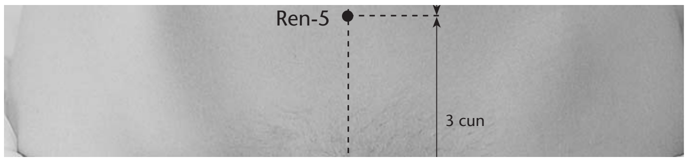

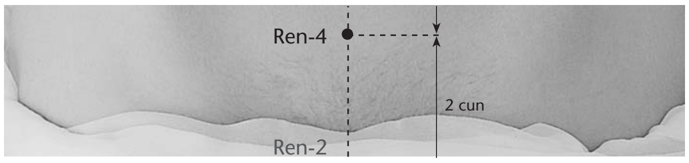

Ex-CA Triangle Moxibustion SANJIAOJIU

Location

These three points are located on the corners of an equilateral triangle, the apex of which is formed by the umbilicus, while the base forms a horizontal line on the abdomen. The sides are equal to the patient's smile.

How to find

It is helpful to explain the location of this point to the patient. Invariably, this will make them smile, allowing the practitioner to determine the length of the sides of the triangle.

Needling

Moxibustion only, no needling.

Actions/Indications

- Regulate Qi and stop diarrhoea.

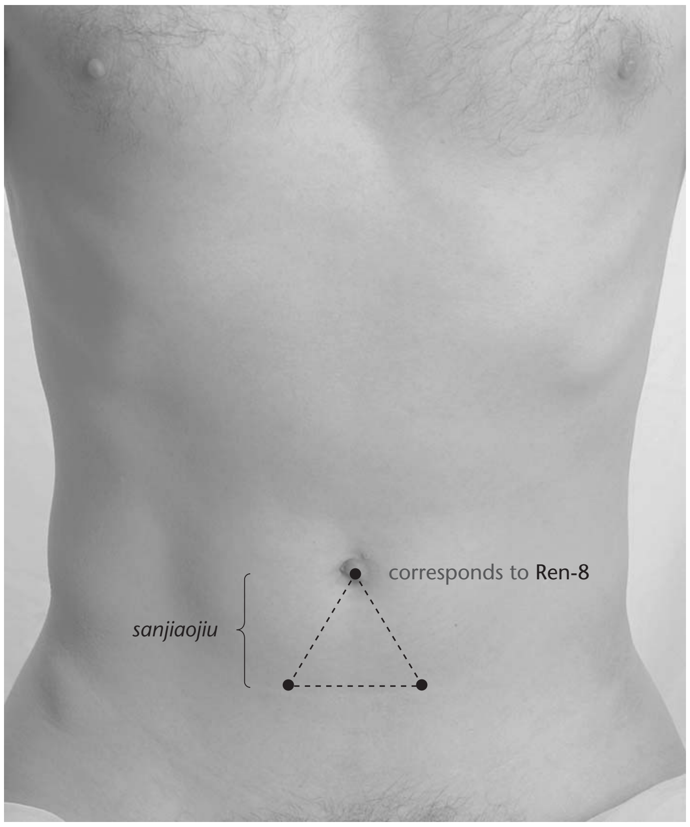

6.3 Extra Points: Back (EX-B)

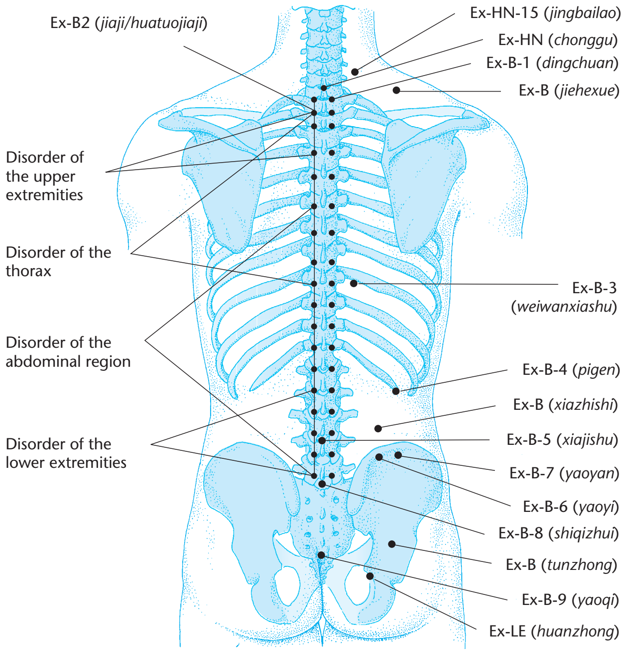
Ex-B2 (*jiaji/huatuojiaji*)
Disorder of the upper extremities
Disorder of the thorax
Disorder of the abdominal region
Disorder of the lower extremities
Ex-HN-15 (*jingbailao*)
Ex-HN (*chonggu*)
Ex-B-1 (*dingchuan*)
Ex-B (*jiehexue*)
Ex-B-3 (*weiwanxiashu*)
Ex-B-4 (*pigen*)
Ex-B (*xiazhishi*)
Ex-B-5 (*xiajishu*)
Ex-B-7 (*yaoyan*)
Ex-B-6 (*yaoyi*)
Ex-B-8 (*shiqizhui*)
Ex-B (*tunzhong*)
Ex-B-9 (*yaoqi*)
Ex-LE (*huanzhong*)

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-B-1	*dingchuan*	NP45	M-BW-1	Ex10
Ex-B-2	*huatuojiaji/jiaji*	PaM or ZP85	M-BW-35	Ex12
Ex-B-3	*weiwanxiashu/* *waiguanxiashu/* *bashu* or *yishu*	PaM or ZP62	M-BW-12	
Ex-B-4	*pigen*	PaM or ZP66	M-BW-16	
Ex-B-5	*xiajishu*	PaM or ZP71	M-BW-21	
Ex-B-6	*yaoyi*	PaM or ZP73	M-BW-23	
Ex-B-7	*yaoyan*	PaM or ZP74	M-BW-24	
Ex-B-8	*shiqizhui/shiqizhuixia*	PaM or ZP75	M-BW-25	Ex11
Ex-B-9	*yaoqi*	PaM or ZP79	M-BW-29	
Further extra points				
Ex-B	*jiehexue*	NP47	N-BW-6	
Ex-B	*tunzhong*	PaM or ZP83	M-BW-33	

Calm Dyspnoea DINGCHUAN **Ex-B-1**

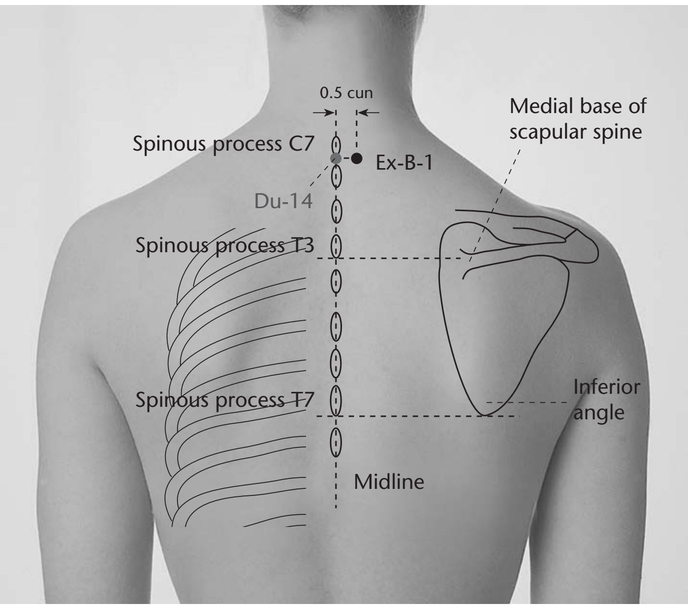
0.5 cun
Medial base of scapular spine
Spinous process C7
Ex-B-1
Du-14
Spinous process T3
Inferior angle
Spinous process T7
Midline

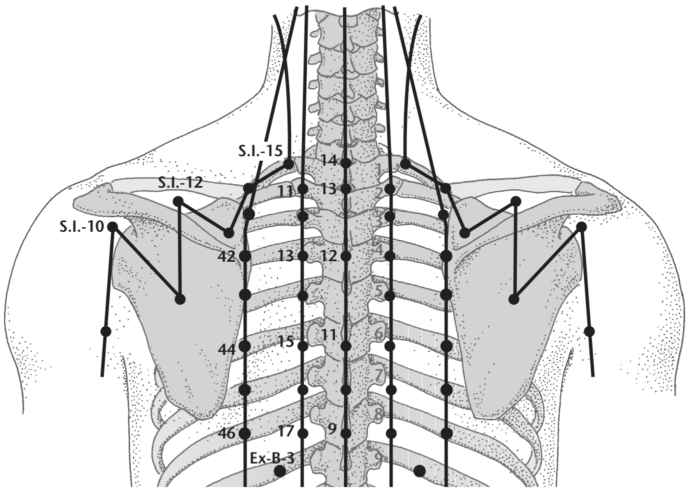
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3

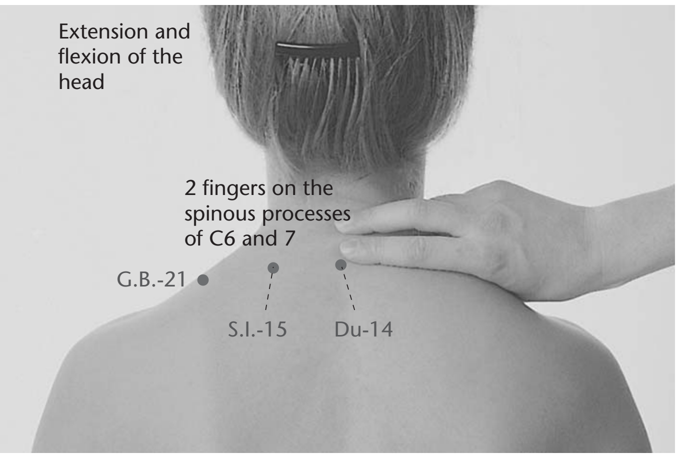
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

Location

0.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra (C7).

How to find

First, identify C7 (→ 3.4.1). Next, measure 0.5 cun from the lower border of its spinous process in a lateral direction and there locate **Ex-B-1.**

Located on the same level are → **Du-14** (on the midline), → **S.I.-15** (2 cun lateral to the midline) and **Ex-B** (*jiehexue*, 3.5 cun lateral to the midline).

Needling

0.5–1 cun in a medial direction.

Actions/Indications

- Harmonises Lung Qi
- Local point for the neck and shoulders

Special features

Modern main point for dyspnoea

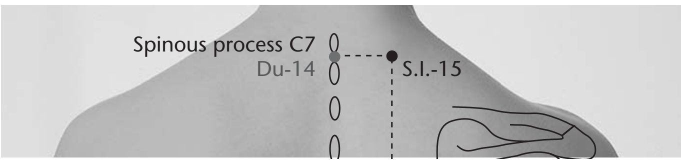
Spinous process C7
Du-14
S.I.-15

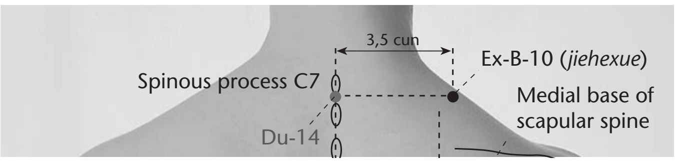
3,5 cun
Ex-B-10 (*jiehexue*)
Spinous process C7
Medial base of scapular spine
Du-14

Ex-B-2 *Hua Tuo's Paravertebral Points* HUATUOJIAJI

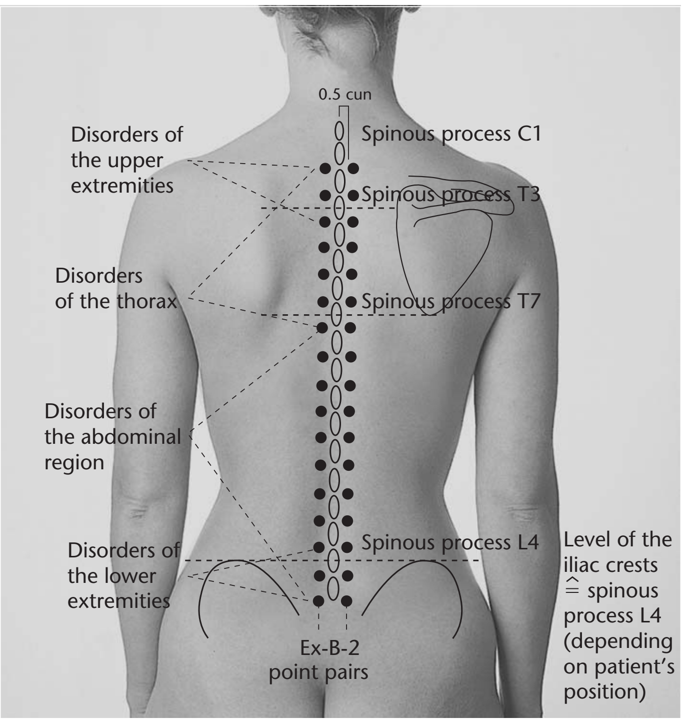
0.5 cun
Spinous process C1
Disorders of the upper extremities
Spinous process T3
Disorders of the thorax
Spinous process T7
Disorders of the abdominal region
Disorders of the lower extremities
Spinous process L4
Level of the iliac crests ≙ spinous process L4 (depending on patient's position)
Ex-B-2 point pairs

Location

17 point pairs, 0.5 cun lateral to the lower borders of the spinous processes, close to the spinal facet joints:

- 12 thoracic point pairs (*xiongjiaji*): between T1 and T12
- 5 lumbar point pairs (*yaojiaji*): between L1 and L5, Depending on the school of thought, corresponding points lateral to the cervical spine are described as 'additional *huatuojiaji*'.

How to find

Select points of **Ex-B-2** depending on the disorder. Determine the relevant spinous process and locate the point pairs 0.5 cun lateral to its lower border.

Needling

0.5–1 cun vertically or better obliquely in a medial direction towards the spine, up to 1.5 cun in the lumbar region. The purpose of needling these points is to stimulate the local nerve roots. This means that the insertion angle has to be adapted to the patient's anatomy. Never needle in a lateral direction. Once *deqi* is obtained, do not further stimulate the needle; it is possible to apply electrostimulation.

Actions/Indications

- Regulate and harmonise the five *zang* and six *fu* Organs, depending on the points selected

Special features

With disorders of the facet joints of the cervical spine, there are often tender points 0.5 cun lateral to the lower borders of the spinous processes. In China, these are needled for chronic disorders of the cervical spine. If oblique insertions (approximately 45°) in a medial direction are used for points on the inner branch of the Bladder channel, the tip of the needle will reach the area of the *huatuojiaji* points. This enhances the therapeutic effect.

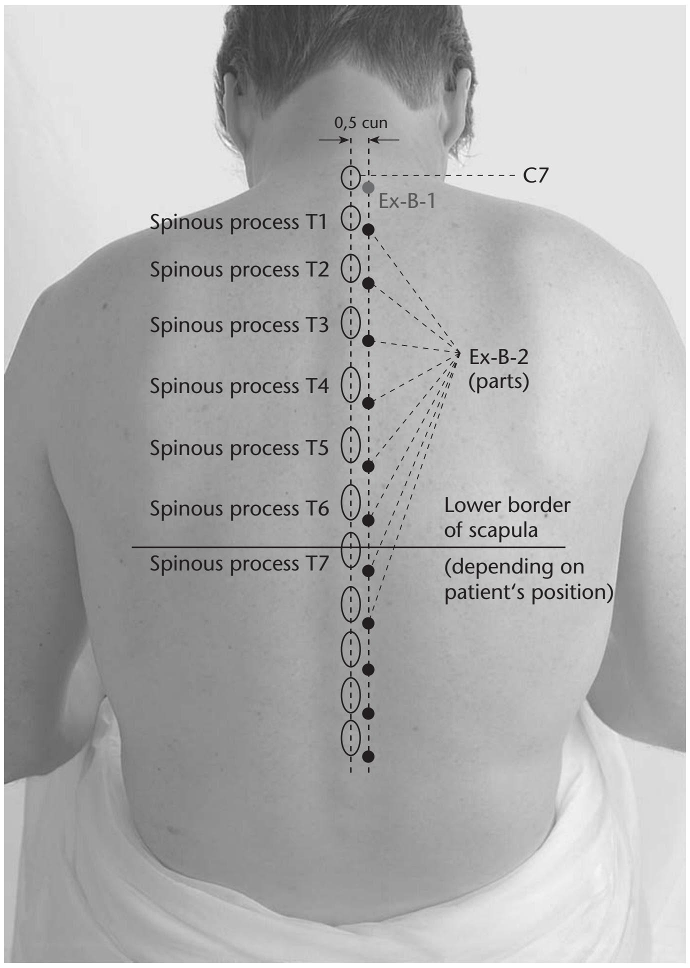
0,5 cun
C7
Ex-B-1
Spinous process T1
Spinous process T2
Spinous process T3
Ex-B-2 (parts)
Spinous process T4
Spinous process T5
Spinous process T6
Lower border of scapula
Spinous process T7
(depending on patient's position)

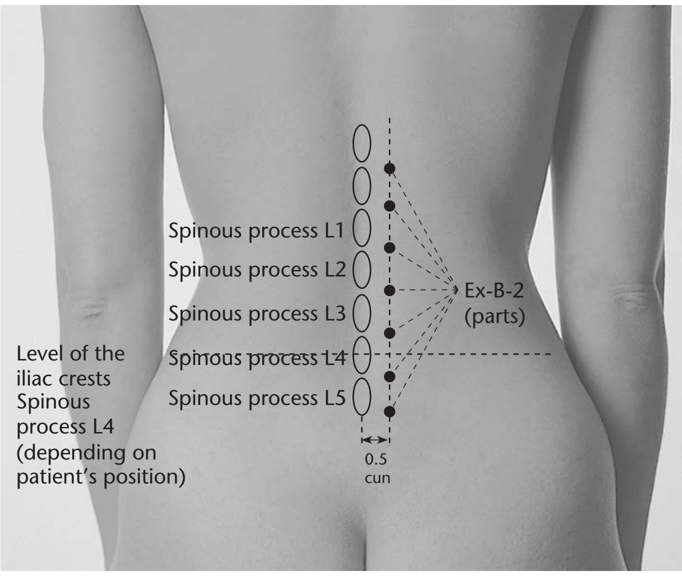
Spinous process L1
Spinous process L2
Ex-B-2 (parts)
Spinous process L3
Level of the iliac crests Spinous process L4 (depending on patient's position)
Spinous process L4
Spinous process L5
0.5 cun

Stomach Controller Lower Shu WEIWANXIASHU Ex-B-3

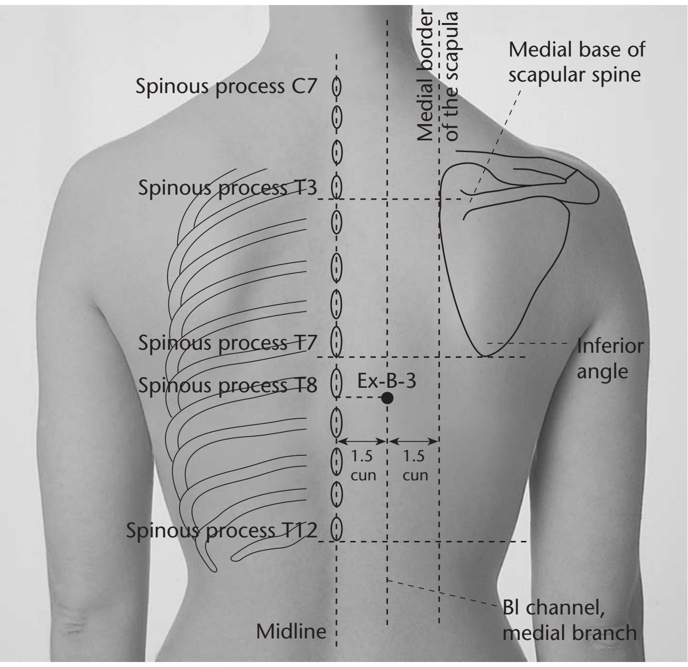
Spinous process C7
Medial border of the scapula
Medial base of scapular spine
Spinous process T3
Spinous process T7
Inferior angle
Spinous process T8
Ex-B-3
1.5 cun
1.5 cun
Spinous process T12
Midline
Bl channel, medial branch

Location

1.5 cun lateral to the lower border of the spinous process of the 8th thoracic vertebra (T8).

How to find

Locate T7 (→ 3.4.2). From there, count down one spinous process (to T8) and locate **Ex-B-3** 1.5 cun lateral from the midline. Use the Tuffier's line (L4, 3.4.3) to confirm the location of this point: from L4, count up the spinous processes of L3 to L1 and of T12 to T8.

Located on the same level is a point of → **Ex-B-2** (0.5 cun lateral to the midline).

Needling

0.5–1 cun vertically or up to 1.5 cun obliquely in a medial direction. Do not needle in a lateral direction. Caution: Pneumothorax.

Actions/Indications

- Opens the channel locally
- Moistens the body and clears Heat

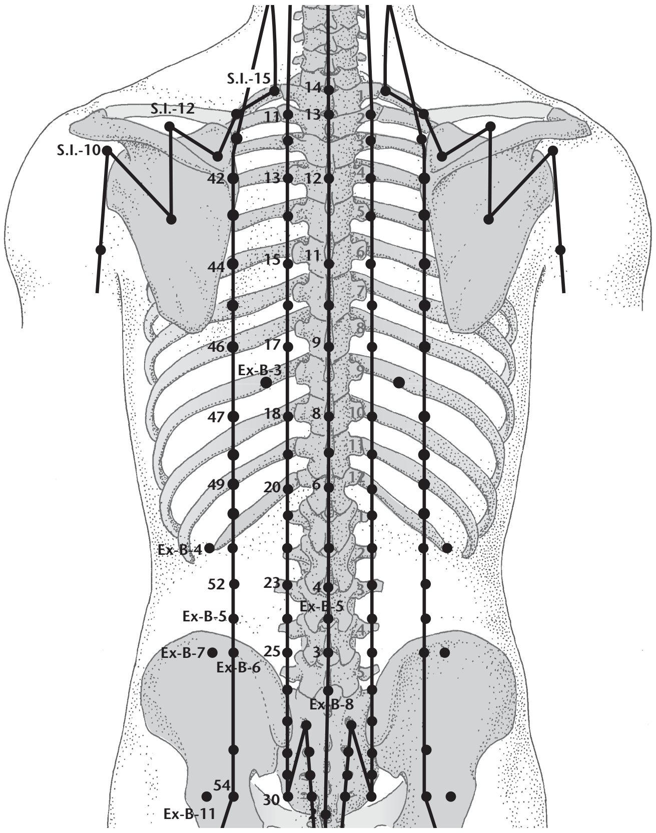
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11

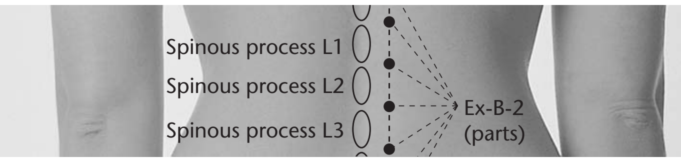
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ex-B-4 *Fullness Root* PIGEN

Location

3.5 cun lateral to the posterior midline, on the level of the lower border of the spinous process of the 1st lumbar vertebra (L1).

How to find

Locate the T7 (→ 3.4.2) and count down 6 spinous processes to L1. Locate **Ex-B-4** on the level of the lower border of its spinous process, 3.5 cun lateral to the midline. Use the Tuffier's line (L4, → 3.4.3) to confirm the location of this point: from L4, count up the spinous processes of L3 to L1.

Located on the same level are → **Du-5** (on the midline), a point of → **Ex-B-2/BL-22/BL-51** (0.5/1.5/3 cun lateral to the midline) as well as → **G.B.-25** (on the lower border of the free end of the 12th rib).

Needling

0.8–1 cun obliquely in a medial direction. Caution: Injury to the kidneys!

Actions/Indications

- Improves the Qi flow between the Upper and Middle Burner
- Eliminates local Qi and Blood stagnation

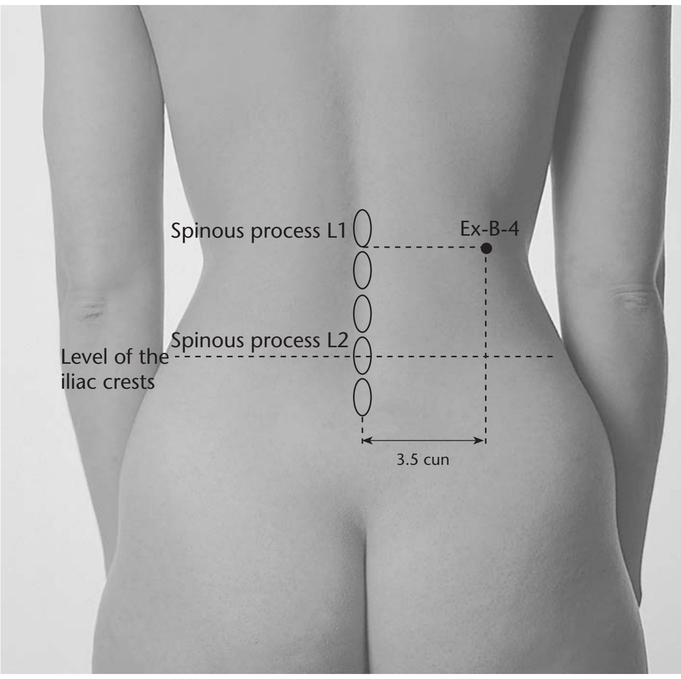
Spinous process L1
Ex-B-4
Spinous process L2
Level of the iliac crests
3.5 cun

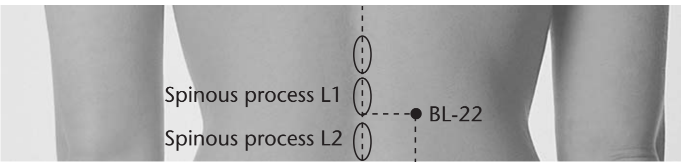
Spinous process L1
BL-22
Spinous process L2

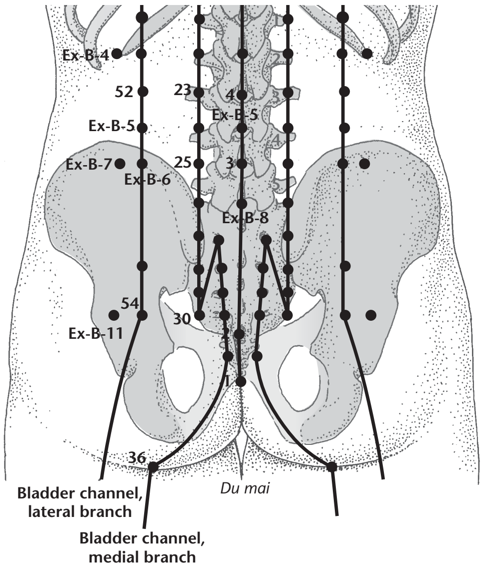
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
2
1
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

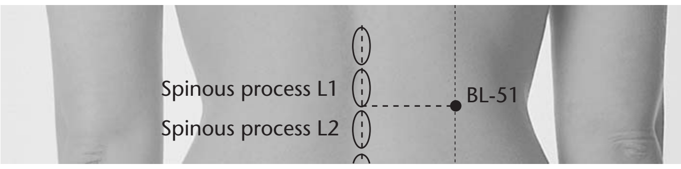
Spinous process L1
BL-51
Spinous process L2

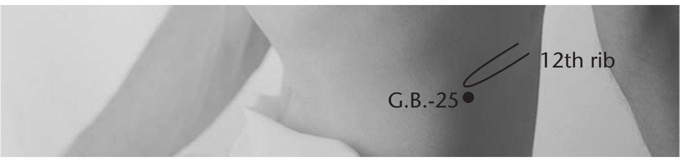
12th rib
G.B.-25

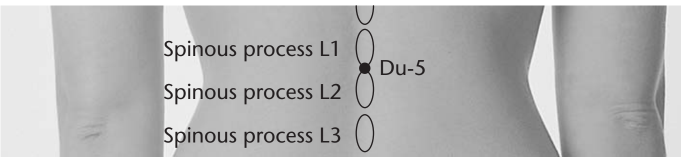
Spinous process L1
Du-5
Spinous process L2
Spinous process L3

Lower Shu Point XIAJISHU Ex-B-5

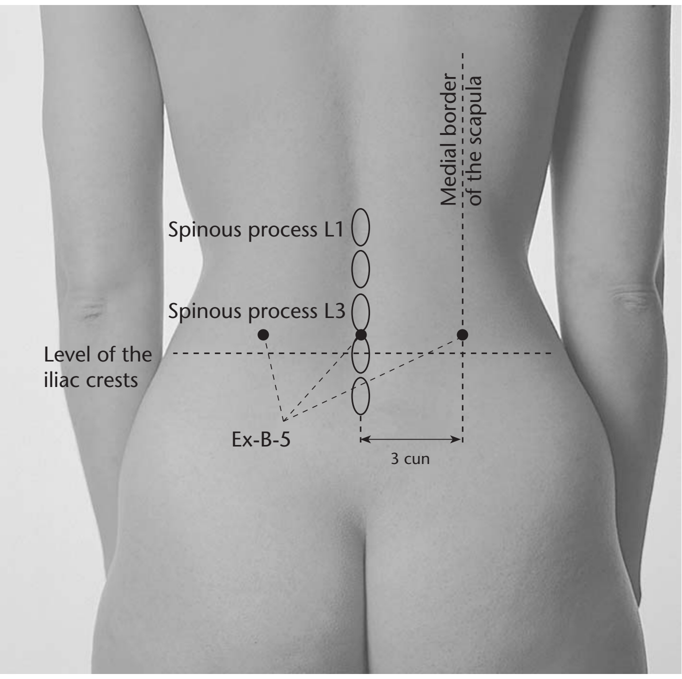
Medial border of the scapula
Spinous process L1
Spinous process L3
Level of the iliac crests
Ex-B-5
3 cun

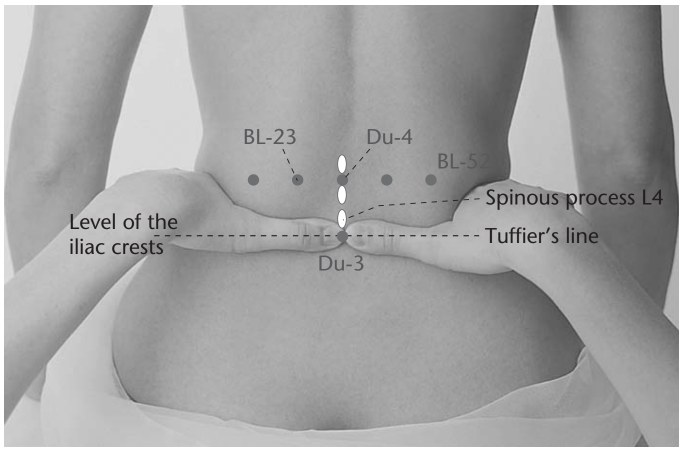
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

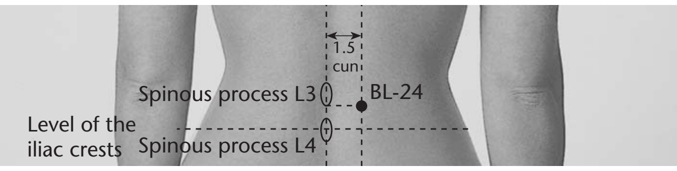
1.5 cun
Spinous process L3
BL-24
Level of the iliac crests
Spinous process L4

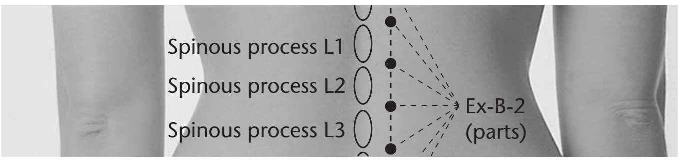
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Location

Some authors consider **Ex-B-5** (*xiajishu*) to comprise three points: one point on the midline below the spinous process of the 3rd lumbar vertebra (L3) (this is the more common and single location of this extra point), complemented by two lateral points 3 cun lateral to the centre point on the midline.

How to find

Locate the Tuffier's line and L4 (➞ 3.4.3). Locate **Ex-B-5** on the midline, on the lower border of the spinous process of L3. Then measure 3 cun in a lateral direction and there locate the two lateral points of **Ex-B-5.**

Located on the same level are a point of ➞ **Ex-B-2/BL-24** (0.5/1.5 cun lateral to the midline).

Needling

0.5–0.8 cun vertically to the skin or obliquely in an inferior direction. The patient's back should be straight or overextended to avoid a spinal puncture. Oblique insertions in a superior direction should only be carried out by experienced practitioners, as in small persons (regardless of their weight), the spinal canal can be reached after only 1.25 cun. Caution: In pregnant women about to go into labour, this point might have a labour-promoting effect.

Actions/Indications

- Tonifies Kidney Yang
- Opens the channel locally

Special features

Strictly speaking, this extra point should be considered as a single point and a pair of points with differing indications: depending on the school of thought and the clinical requirements, the lateral points are used for localised Qi and Blood stagnation, whereas the central point is indicated for Kidney Yang deficiency.

Ex-B-6 Back Pain Point YAOYI

Location

3 cun lateral to the midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). On the level of the lower border of the spinous process of L4, measure 3 cun in a lateral direction and there locate **Ex-B-6.** Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2/BL-25/ Ex-B-7** (0.5/1.5/3.5 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Opens the channel and *luo* vessels

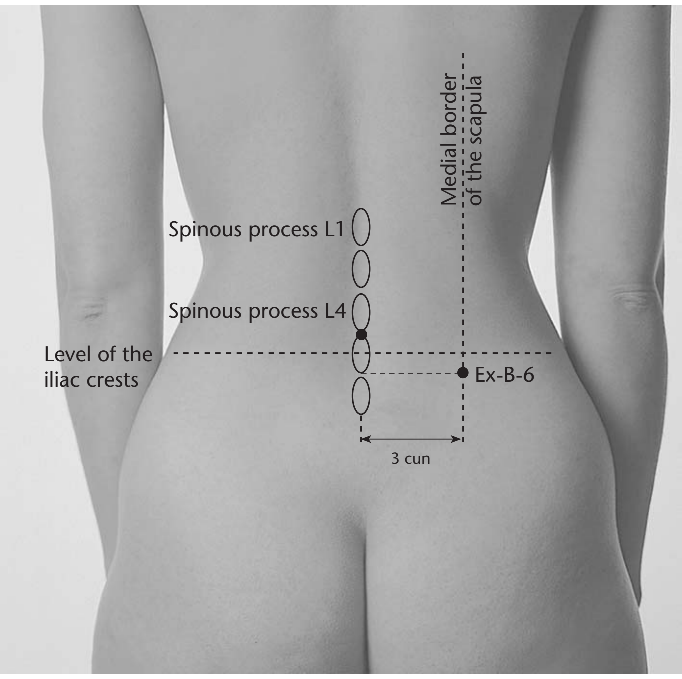
Medial border of the scapula
Spinous process L1
Spinous process L4
Level of the iliac crests
Ex-B-6
3 cun

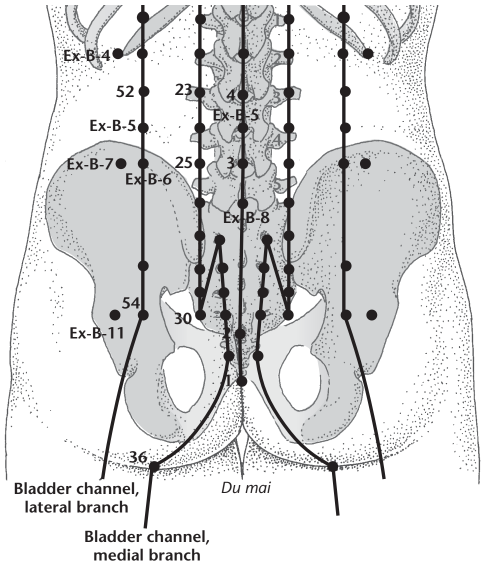
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Bladder channel, lateral branch
Du mai
Bladder channel, medial branch

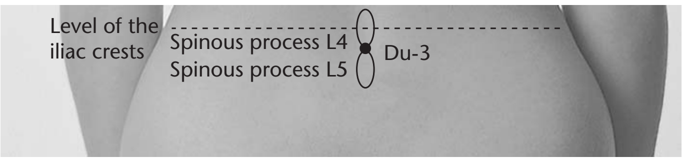
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

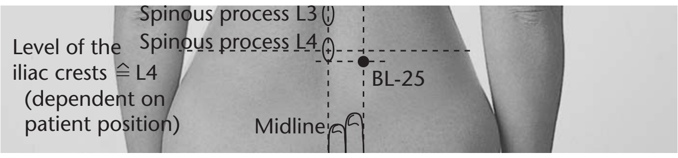
Spinous process L3
Spinous process L4
Level of the iliac crests ≙ L4 (dependent on patient position)
BL-25
Midline

Lumbar Eyes YAOYAN **Ex-B-7**

Location

3.5 cun lateral to the midline, on the level of the lower border of the spinous process of the 4th lumbar vertebra (L4).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). **Ex-B-7** is located 3.5 cun lateral to it. Located on the same level are → **Du-3** (on the midline), a point of → **Ex-B-2/BL-25/Ex-B-6** (0.5/1.5/3 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Opens the channel locally
- Tonifies the Kidneys

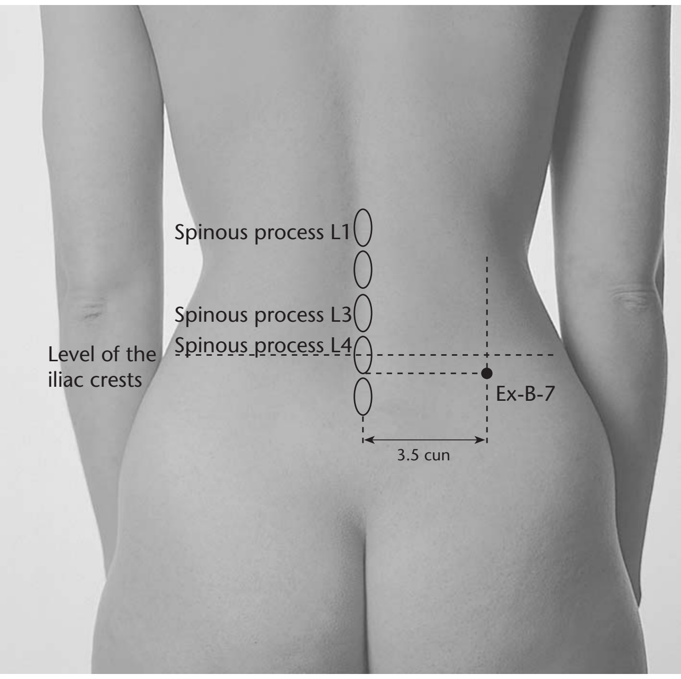
Spinous process L1
Spinous process L3
Spinous process L4
Level of the iliac crests
Ex-B-7
3.5 cun

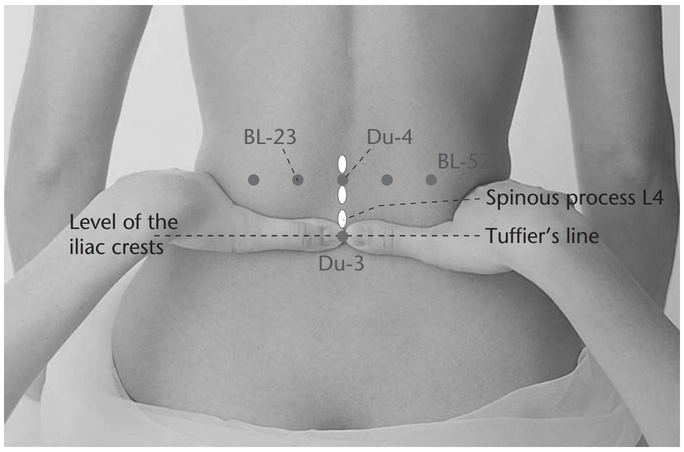
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

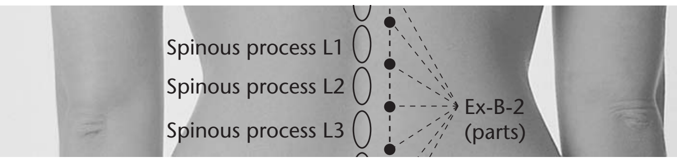
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

Ex-B-8 *Below the 17th Vertebra* SHIQIZHUI/SHIQIZHUIXIA

Location

On the midline, below the spinous process of the 5th lumbar vertebra (L5).

How to find

Locate the Tuffier's line and L4 (→ 3.4.3). Next, palpate in an inferior direction to the next vertebra (L5) and locate **Ex-B-8** below the border of its spinous process, at the lumbosacral joint. Located on the same level are a point of → **Ex-B-2/BL-26** (0.5/1.5 cun lateral to the midline).

Needling

0.5–1 cun vertically to the skin or obliquely in an inferior direction. The patient's back should be straight or slightly flexed to avoid a spinal puncture. Oblique insertions in a superior direction should only be carried out by experienced practitioners, as in small persons (regardless of their weight), the spinal canal can be reached after only 1.25 cun. Caution: In pregnant women about to go into labour, this point might have a labour-promoting effect.

Actions/Indications

- Tonifies the Kidneys
- Opens the channel locally

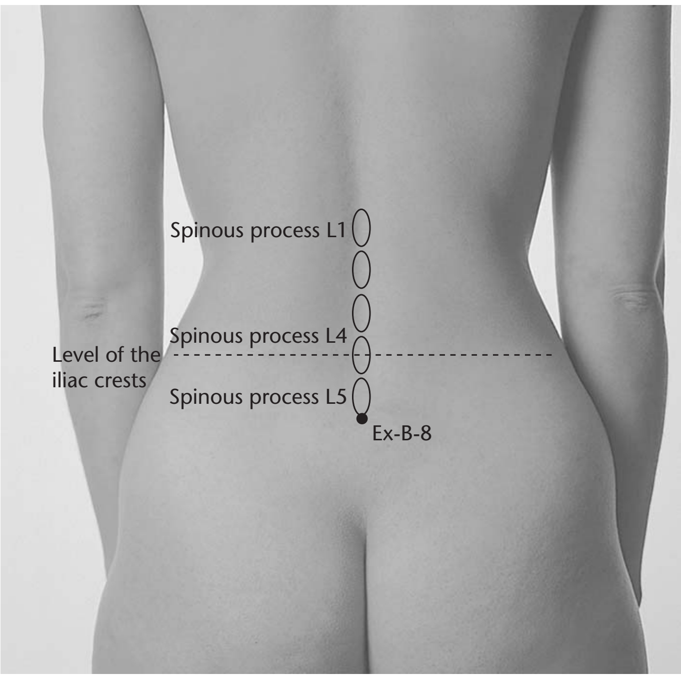
Spinous process L1
Spinous process L4
Level of the iliac crests
Spinous process L5
Ex-B-8

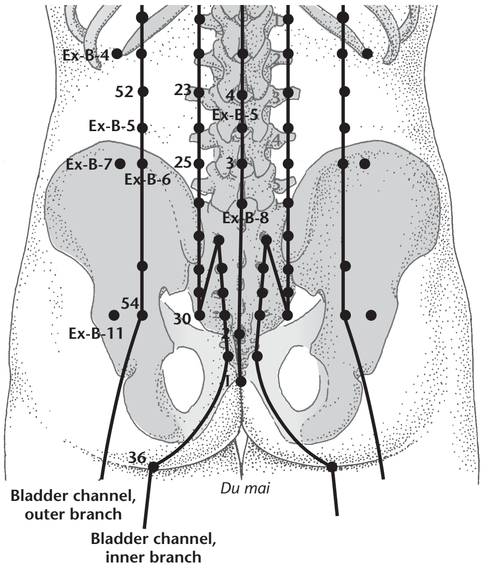
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
Ex-B-11
30
36
Du mai
Bladder channel, outer branch
Bladder channel, inner branch

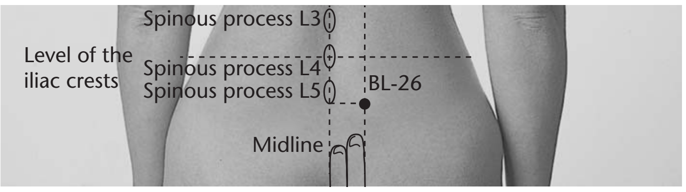
Spinous process L3
Level of the iliac crests
Spinous process L4
Spinous process L5
BL-26
Midline

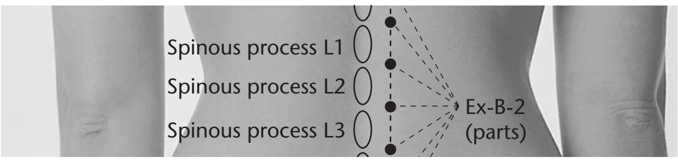
Spinous process L1
Spinous process L2
Spinous process L3
Ex-B-2 (parts)

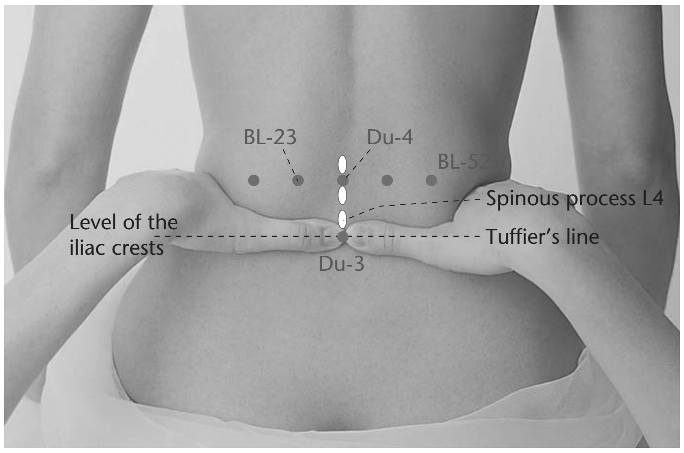
BL-23
Du-4
BL-52
Spinous process L4
Level of the iliac crests
Tuffier's line
Du-3

Miraculous Lumbar Point YAOQI **Ex-B-9**

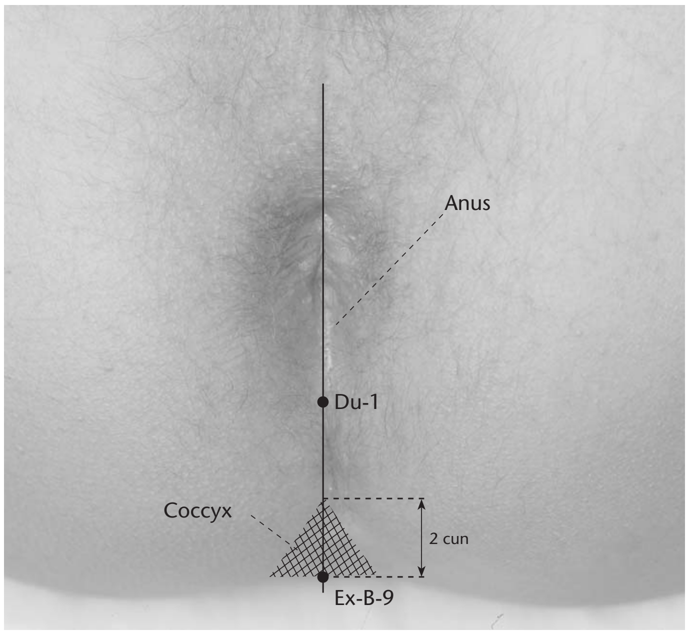

Location

2 cun superior to the tip (inferior end) of the coccyx.

How to find

Locate the tip of the coccyx superior to the anus and palpate for the sacral hiatus. From there, palpate 2 cun in a superior direction and locate **Ex-B-9** in a depression between the sacral processes.
→ **Du-1** is located on the midline, between the tip of the coccyx and the anus.

Needling

Up to 1.5 cun obliquely in a superior direction.

Actions/Indications

- Calms the *shen*

Ex-B-10 *Tuberculosis Point* JIEHEXUE

Location

3.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra (C7).

How to find

Locate the spinous process of C7 and, on this level, palpate 3.5 cun in a lateral direction. There, locate **Ex-B** (*jiehexue*).
Located on the same level are → **Du-14** (on the midline), → **Ex-B-2/S.I.-15** (0.5/2 cun lateral to the midline).

Needling

Vertically 0.5–0.8 cun. Caution: Pneumothorax.

Actions/Indications

- Tonifies the Lung
- Opens the channel locally

Special features

This is an unofficial extra point.

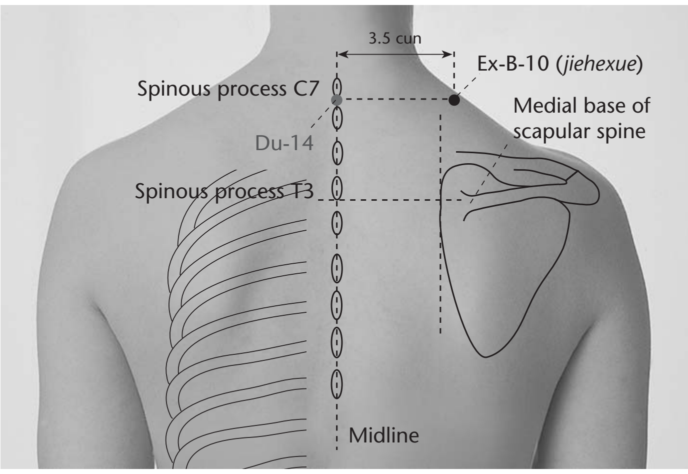
3.5 cun
Ex-B-10 (*jiehexue*)
Spinous process C7
Medial base of scapular spine
Du-14
Spinous process T3
Midline

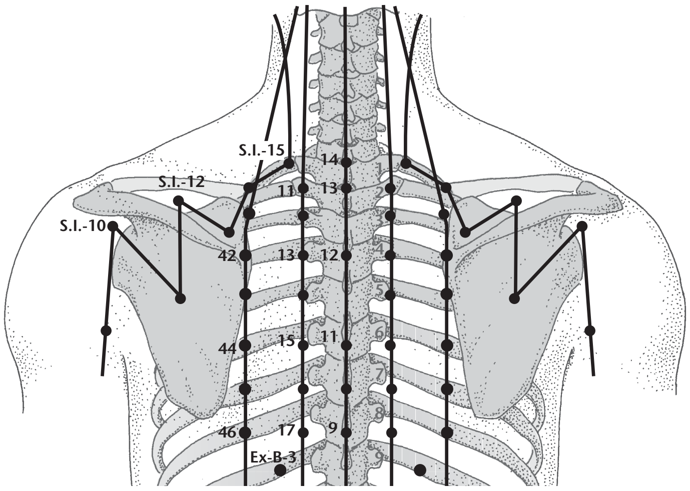
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
6
46
17
9
Ex-B-3

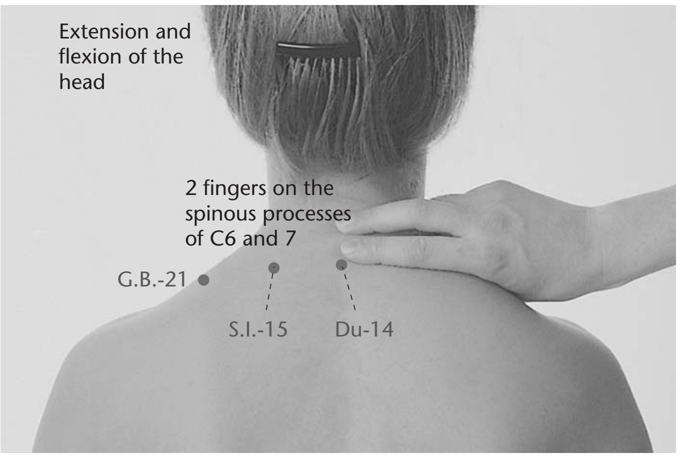
Extension and flexion of the head
2 fingers on the spinous processes of C6 and 7
G.B.-21
S.I.-15
Du-14

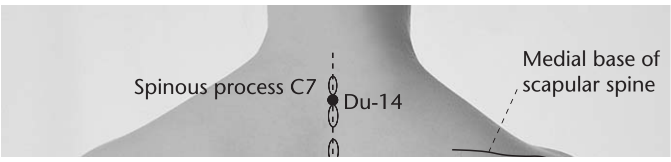
Spinous process C7
Du-14
Medial base of scapular spine

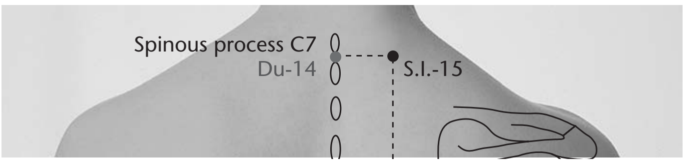
Spinous process C7
Du-14
S.I.-15

Buttock Centre TUNZHONG Ex-B

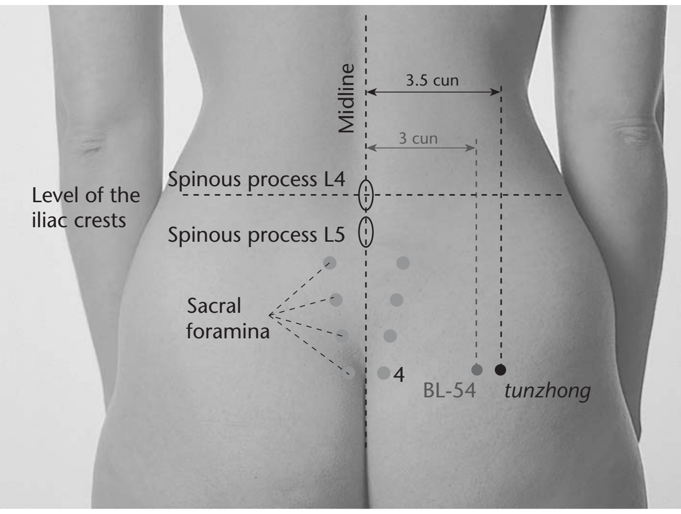
Midline
3.5 cun
3 cun
Level of the iliac crests
Spinous process L4
Spinous process L5
Sacral foramina
4
BL-54
tunzhong

Location

At the centre of the buttock, 3.5 cun lateral to the posterior midline, on the level of the 4th sacral foramen.

How to find

Locate the lumbosacral joint, which often forms an easily palpable 'kink' between the L5 and the sacrum. Alternatively, locate the Tuffier's line (→ 3.4.3) for reference. Then palpate the inferior aspect of the sacrum for the sacral hiatus (→ 3.4.4). Between these two reference points, the four pairs of sacral foramina are arranged on either side of the midline in a slight V-shaped form and can be palpated as shallow depressions. **Ex-B** (*tunzhong*) is located 3.5 cun lateral to the lowest depression (=4th foramen), approximately at the centre of the buttock.

Located on the same level are → **BL-34/BL-30/BL-54** (over the 4th foramen/1.5/3 cun lateral to the midline).

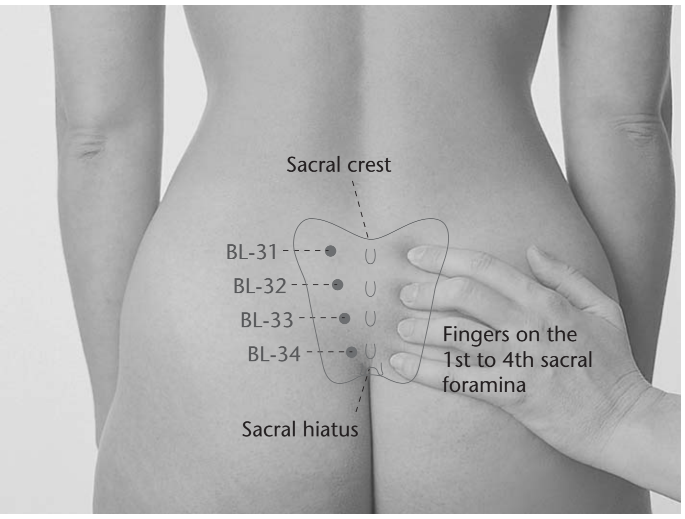
Sacral crest
BL-31
BL-32
BL-33
BL-34
Fingers on the 1st to 4th sacral foramina
Sacral hiatus

Needling

Vertically 2–3 cun

Actions/Indications

- Opens the channel and *luo* vessels locally

Special features

This is an unofficial extra point.

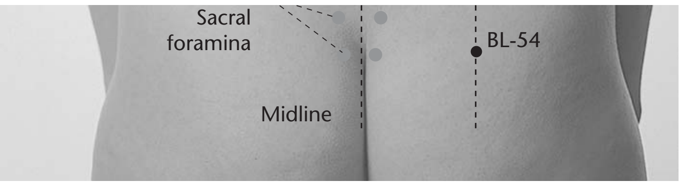
Sacral foramina
BL-54
Midline

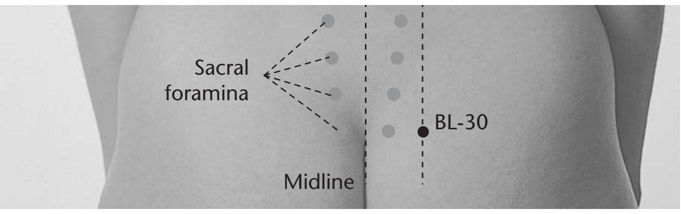
Sacral foramina
BL-30
Midline

6.4 Extra Points: Upper Extremities (EX-UE)

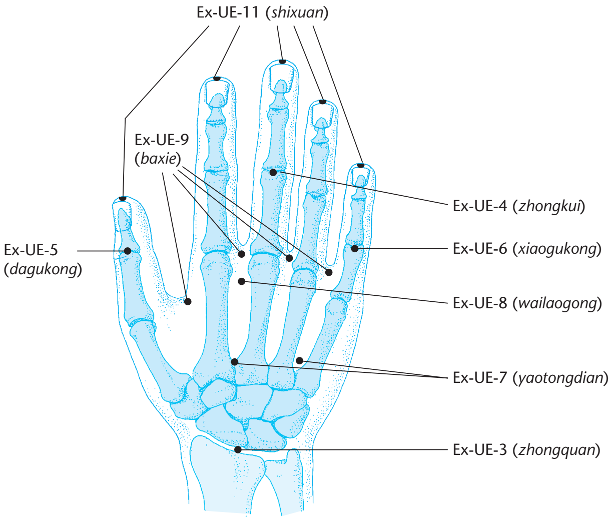
Ex-UE-11 (*shixuan*)
Ex-UE-9 (*baxie*)
Ex-UE-5 (*dagukong*)
Ex-UE-4 (*zhongkui*)
Ex-UE-6 (*xiaogukong*)
Ex-UE-8 (*wailaogong*)
Ex-UE-7 (*yaotongdian*)
Ex-UE-3 (*zhongquan*)

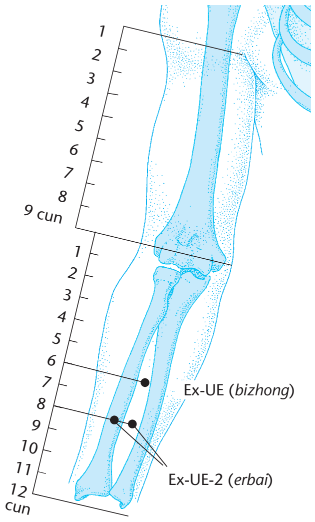
1
2
3
4
5
6
7
8
9 cun
1
2
3
4
5
6
7
8
9
10
11
12 cun
Ex-UE (*bizhong*)
Ex-UE-2 (*erbai*)

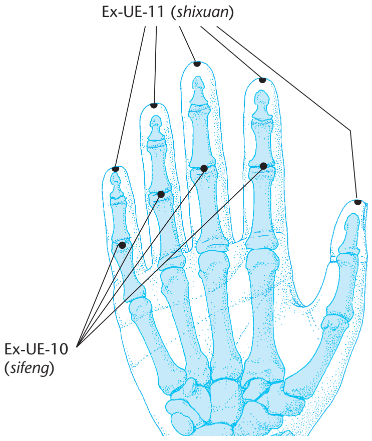
Ex-UE-11 (*shixuan*)
Ex-UE-10 (*sifeng*)

English abbreviation (standard)	*Pinyin* name	Nguyen Van Nghi, König and Wancura, Schnorrenberger	Shanghai College	Ex (Hempen)
Ex-UE-1	*zhoujian*		M-UE-46	
Ex-UE-2	*erbai*	PaM or ZP114	M-UE-29	
Ex-UE-3	*zhongquan*	PaM or ZP118	M-UE-33	
Ex-UE-4	*zhongkui*	PaM101	M-UE-16	
Ex-UE-5	*dagukong*	PaM or ZP100	M-UE-15	
Ex-UE-6	*xiaogukong*	PaM or ZP102	M-UE-17	
Ex-UE-7	*yaotongxue/yaotongdian*	Handpoint 2 PaM or ZP110/111[*]	M-UE	Ex18
Ex-UE-8	*wailaogong*[†]*/luozhen*	PaM or ZP108[†]	M-UE-24	Ex17
Ex-UE-9	*baxie*	PaM or ZP107	M-UE-22	Ex14
Ex-UE-10	*sifeng*		M-UE-9	Ex13
Ex-UE-11	*shixuan*	PaM or ZP86	M-UE-1	Ex15
Further Extra points				
Ex-UE	*jianqian/jianneiling*	Part of NP74[‡] (*sanjian*): *jianyu* (L.I.-15), *jianqian*, *jianhou*	M-UE-48[§]	Ex16
Ex-UE	*bizhong*	PaM or ZP115	M-UE-30	

---

[*] The 'standard' point Ex-UE-7 (*yaotongdian* or *yaotongxue*) is described by Nguyen Van Nghi and Schnorrenberger as two separate points with different *pinyin* names: PaM or ZP110 (*weiling*) and 111 (*jingling*). Their location also corresponds to handpoint 2.

[†] Nguyen Van Nghi and Schnorrenberger distinguish between *luozhen* (PaM or ZP108) and *wailaogong* (PaM or ZP109). They describe the location of *luozhen* as identical with the 'standard' location of Ex-UE-8 (*wailaogong*): on the dorsum of the hand, between the 2nd and 3rd metacarpal bones, approximately 0.5 cun proximal to the metacarpophalangeal joint. 'Their' *wailaogong* is located at the midpoint of a line connecting the wrist and the head of the 3rd metacarpal bone, between the metacarpals, directly opposite to the point *laogong* (P-8) – hence the name *wailaogong* ('outer' *laogong*).

[‡] **NP74** (*jiansanzhen*/Three pinpricks at the Shoulder) according to Schnorrenberger: a set of three points: the first is identical with **L.I.-15** (*jianyu*), the second (*jianquan*/ Anterior to the Shoulder) is located 1 cun superior to the end of the anterior axillary fold, and the third (*jianhou*/Posterior to the Shoulder) is located 1 cun superior to the end of the posterior axillary fold.

[§] According to Deadman et al. (1998), **M-UE-48** (*jianqian*) is located midway between the end of the anterior axillary fold and **L.I.-15** (this point is not mentioned by the Shanghai College). See the point page for the location of this point.

Ch06b-F10028.qxd 2/22/08 10:04 PM Page 587

Elbow Tip ZHOUJIAN **Ex-UE-1**

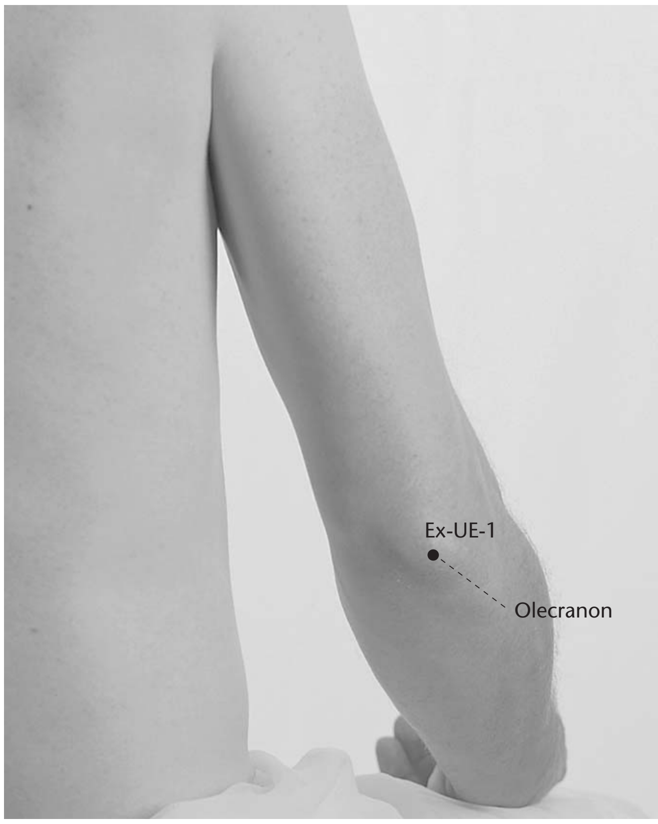
Ex-UE-1
Olecranon

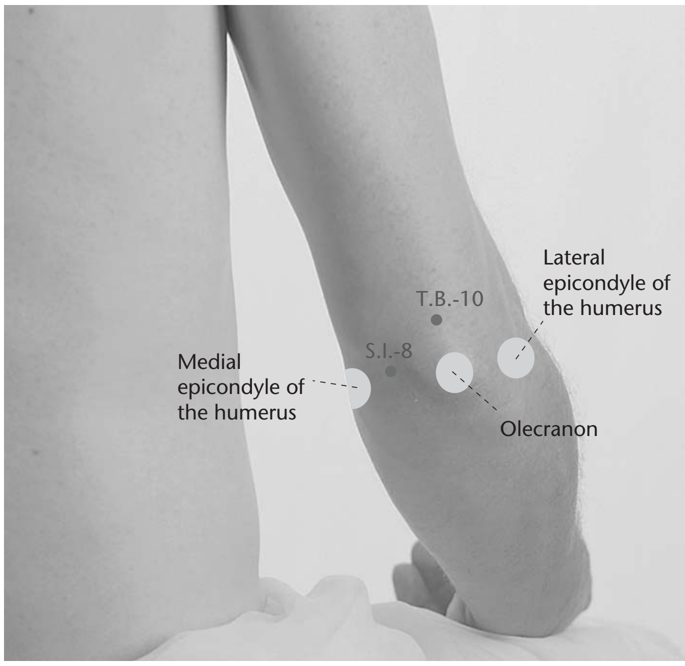
Lateral epicondyle of the humerus
T.B.-10
S.I.-8
Medial epicondyle of the humerus
Olecranon

Location

On the tip of the olecranon.

How to find

On the tip of the olecranon. This point is best located with the elbow flexed.

Needling

Moxibustion only!

Actions/Indications

- Opens the channel and *luo* vessels locally
- Disperses Phlegm locally

Ex-UE-2 *Two Whites* ERBAI

Location

A pair of points on the palmar aspect of the forearm, 4 cun proximal to the wrist joint space (most distal wrist crease), on either side of the tendon of the flexor carpi radialis muscle.

How to find

Place the patient's arm in a relaxed supinated position. As the location of the wrist crease can vary, the wrist joint space provides a more reliable point of reference (➞ 3.3.3). From the centre of the wrist joint space, measure 4 cun in a proximal direction and, on this level, locate one of the points lateral to and the other point medial to the tendon of the flexor carpi radialis muscle. **Or:** Divide the distance between the cubital crease (at ➞ **P-3**) and the wrist joint space (at ➞ **P-7**) into thirds, then locate ➞ **Ex-UE-2** one third of the distance from the wrist joint (helpful tool: elastic tape ➞ 2).

Needling

Up to 1 cun vertically or up to 1.5 cun obliquely towards proximal.

Actions/Indications

- Raises Qi to treat prolapse of the rectum and haemorrhoids

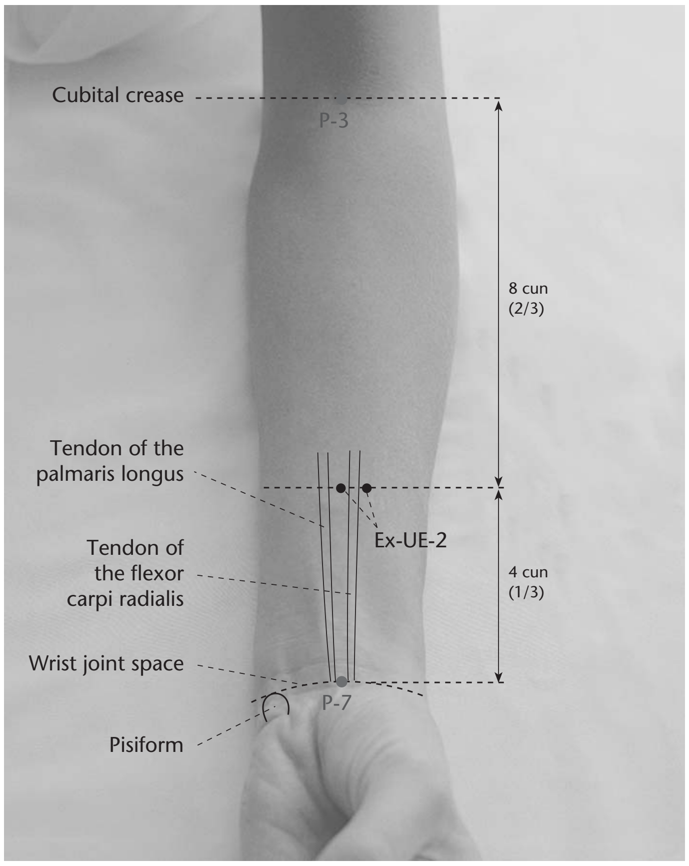
Cubital crease
P-3
8 cun
(2/3)
Tendon of the
palmaris longus
Ex-UE-2
Tendon of
the flexor
carpi radialis
4 cun
(1/3)
Wrist joint space
P-7
Pisiform

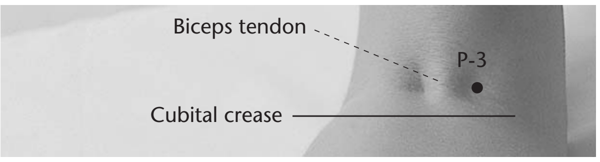
Biceps tendon
P-3
Cubital crease

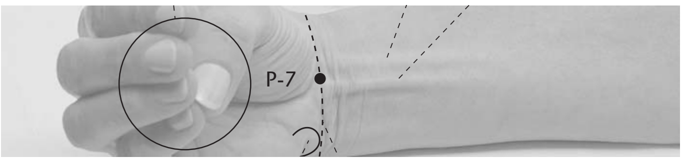

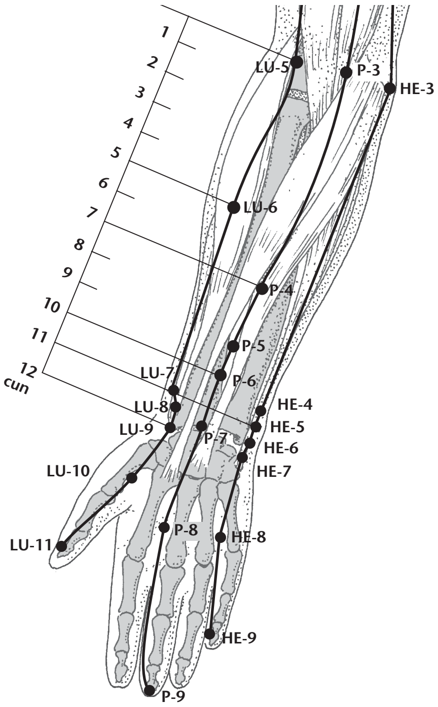
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
LU-7
P-6
LU-8
LU-9
HE-4
HE-5
P-7
HE-6
HE-7
LU-10
P-8
HE-8
LU-11
HE-9
P-9

Posterior Spring ZHONGQUAN **Ex-UE-3**

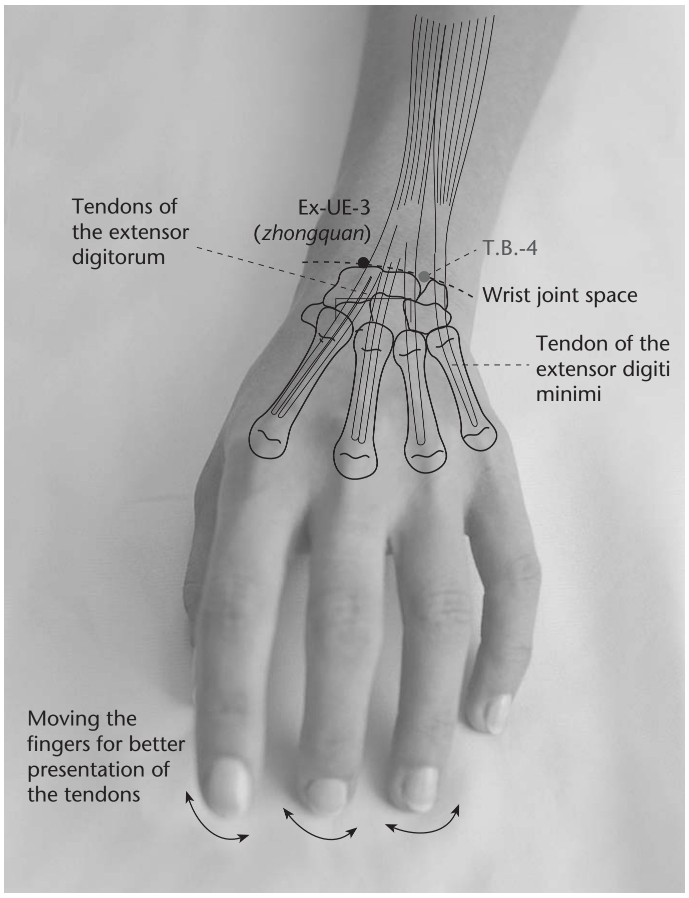
Tendons of the extensor digitorum
Ex-UE-3 (*zhongquan*)
T.B.-4
Wrist joint space
Tendon of the extensor digiti minimi
Moving the fingers for better presentation of the tendons

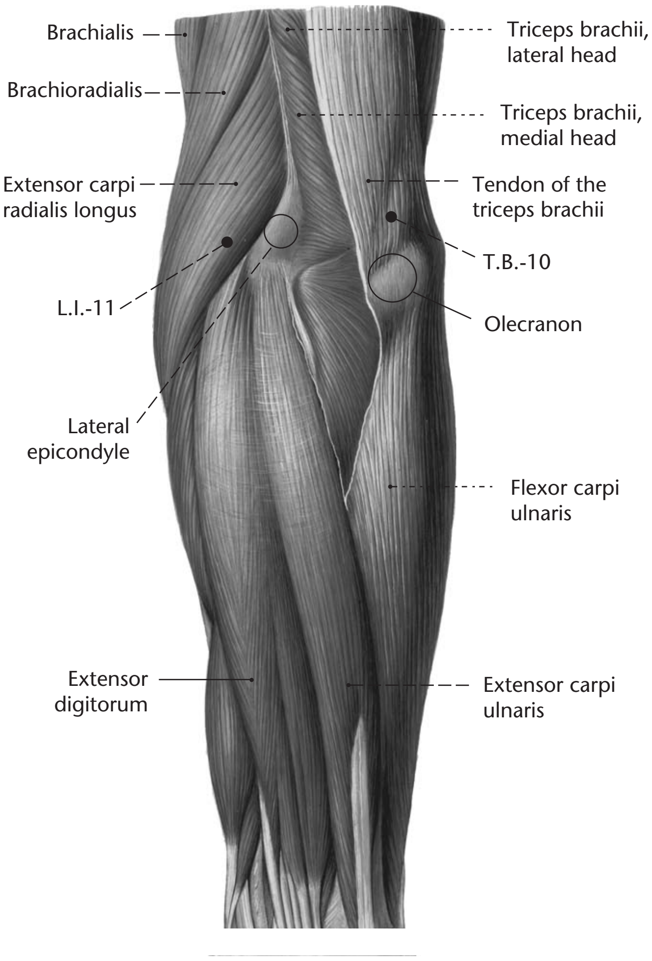
Brachialis
Triceps brachii, lateral head
Brachioradialis
Triceps brachii, medial head
Extensor carpi radialis longus
Tendon of the triceps brachii
T.B.-10
L.I.-11
Olecranon
Lateral epicondyle
Flexor carpi ulnaris
Extensor digitorum
Extensor carpi ulnaris

Location

On the dorsal aspect of the wrist joint space (dorsal wrist crease), radial to the tendon of the extensor digitorum communis muscle.

How to find

Place the patient's arm in a relaxed pronated position. As the location of the dorsal wrist crease can vary, the joint space between the proximal row of carpals and the radius/ulna provides a more reliable point of reference (→ 3.3.3). By moving the wrist in a relaxed way, you can easily palpate the joint space. Locate **Ex-UE-3** in the joint space, radial to the tendon of the extensor digitorum communis muscle. The tendons become more pronounced by moving and overextending the three medial fingers.
→ **T.B.-4** is located on the same level, ulnar to the tendon of the extensor digitorum communis muscle.

Needling

Vertically 0.3–0.5 cun

Actions/Indications

- Harmonises the Qi flow between the Upper and Middle Burner

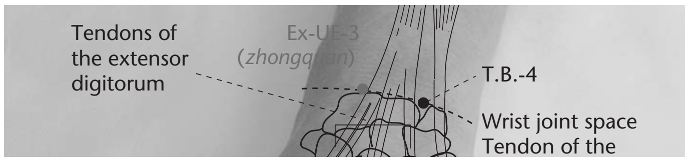
Tendons of the extensor digitorum
Ex-UE-3
(zhongquan)
T.B.-4
Wrist joint space
Tendon of the

Ex-UE-4 *Back of the Middle Finger* ZHONGKUI

Location

As the name implies, on the dorsal aspect of the middle finger, in the centre of the transverse creases of the proximal interphalangeal joint (PIP).

How to find

Slightly bend the PIP of the middle finger. The point is located at the vertex of the flexed joint.
Located in a comparable position on the little finger is → **Ex-UE-6** (*xiaokugong*, in the centre of the proximal PIP of the little finger).
Also located in a comparable position is → **Ex-UE-10** (*sifeng*, four points on the palmar aspect of the fingers of each hand, in the centre of the creases of the PIP joints of the 2nd to 5th fingers).

Needling

Prick to bleed or moxibustion.

Actions/Indications

- Descends Stomach Qi
- Clears Heat

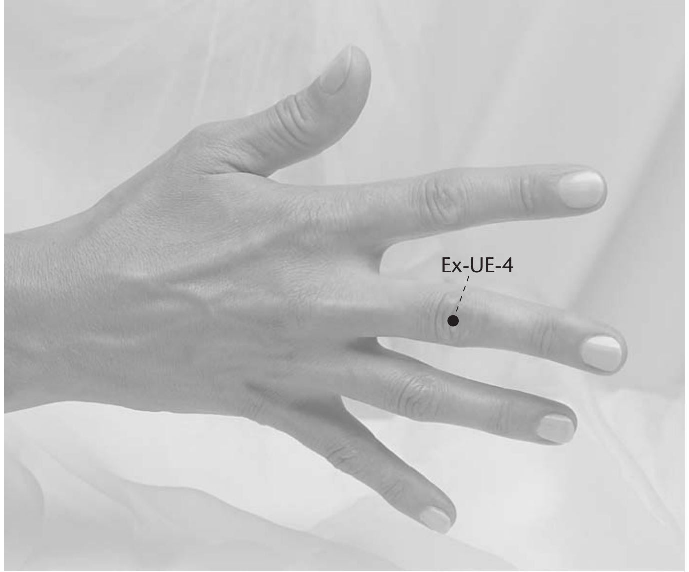

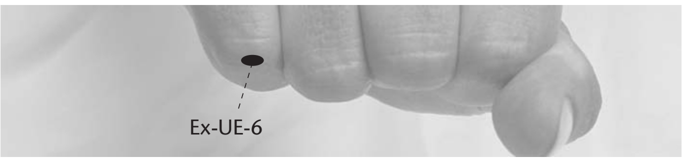

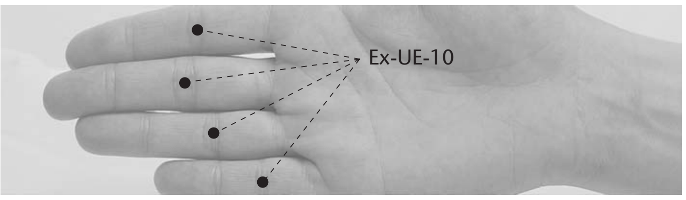

Ch06b-F10028.qxd 2/22/08 10:04 PM Page 591

Thumb Joint DAGUKONG **Ex-UE-5**

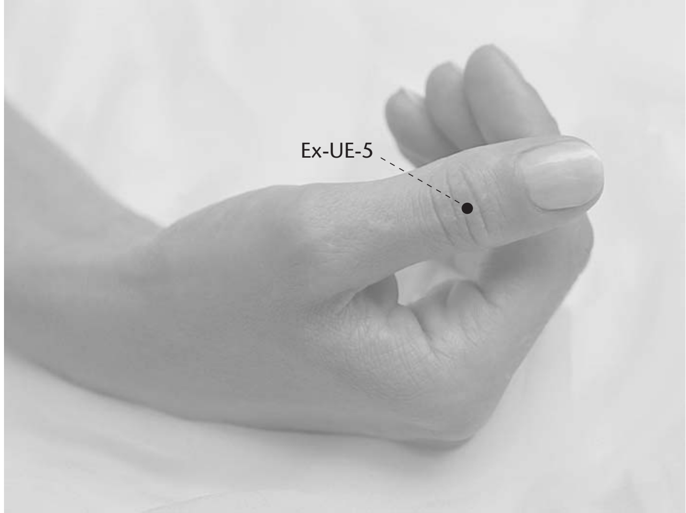

Location

As the name implies, on the dorsal aspect of the thumb, in the centre of the transverse creases of the interphalangeal joint.

How to find

Ask the patient to slightly flex the interphalangeal joint of the thumb. The point is located at the vertex of the bent joint.

Needling

Prick to bleed or moxibustion.

Actions/Indications

- Clears Heat
- Harmonises the Middle Burner

Ex-UE-6 *Little Finger Joint* XIAOGUKONG

Location

As the name implies, on the dorsal aspect of the little finger, in the centre of the transverse creases of the proximal interphalangeal joint (PIP).

How to find

Slightly bend the PIP of the little finger. The point is located at the vertex of the flexed joint.

Located in a comparable position on the middle finger are → **Ex-UE-4** (in the centre of the proximal PIP of the middle finger) and, on the palmar aspect of the little finger, → **Ex-UE-10** (four points on the palmar aspect of the fingers of each hand, in the centre of the creases of the PIP joints of the 2nd to 5th fingers).

Needling

Prick to bleed

Actions/Indications

- Clears Heat

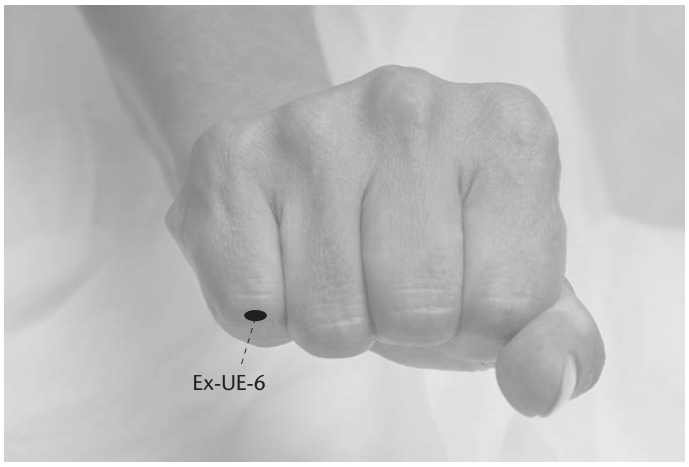

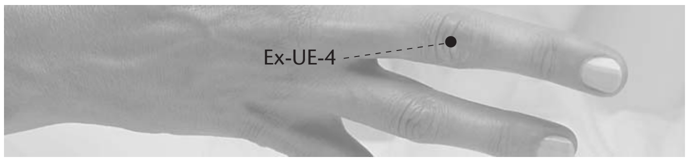

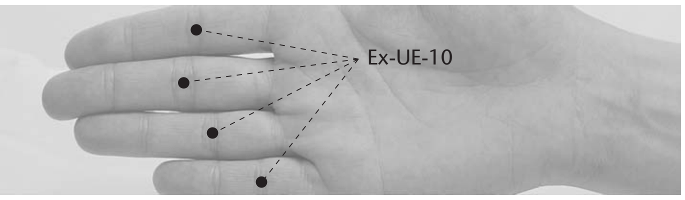

Lumbar Pair Point YAOTONGDIAN Ex-UE-7

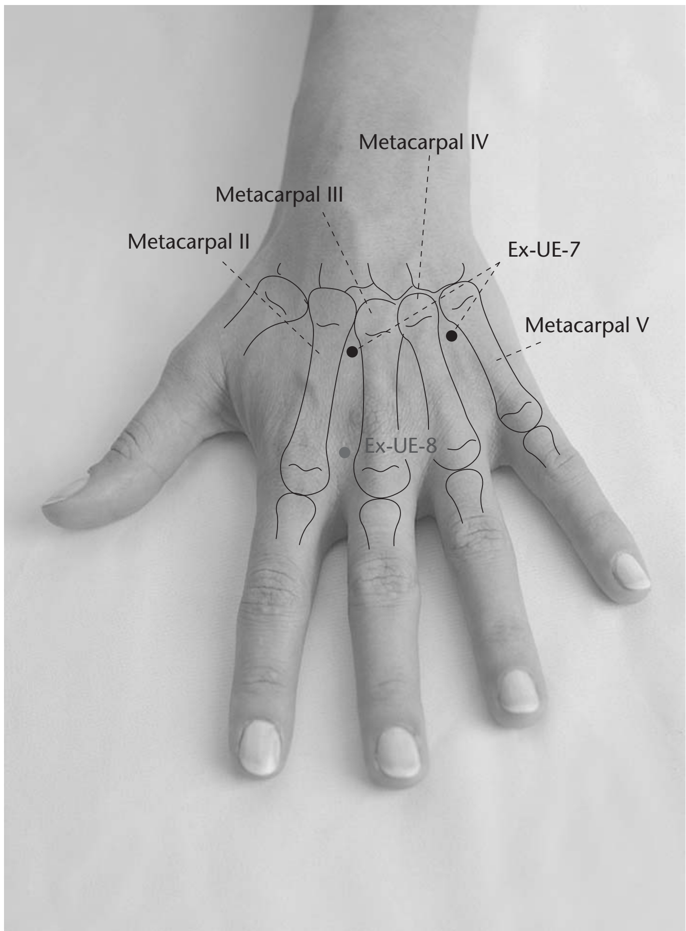
Metacarpal IV
Metacarpal III
Metacarpal II
Ex-UE-7
Metacarpal V
Ex-UE-8

Location

Two points on the dorsal aspect of the hand, between the 2nd/3rd and 4th/5th metacarpal bones, on the level of the junctions of the shaft and the base of the respective metacarpal bones.

How to find

On the dorsal aspect of the hand, glide with the palpating fingers in the grooves between the 2nd and 3rd as well as between the 4th and 5th metacarpal bones towards the wrist joint, until the fingers come to rest in the depressions just distal to the bases of the metacarpal bones. Here, locate the points of **Ex-UE-7** (*yaotongdian* or *yaotongxue*).

Needling

0.5–0.8 cun vertically or slightly obliquely towards the centre of a loose fist. Caution: Painful point! Do not stimulate these distal points in patients with a weak constitution: danger of needle collapse!

Actions/Indications

- Strengthen the Qi and Blood in the lumbar region, alleviate pain

Special features

Very effective distal point for acute disorders of the lumbar spine

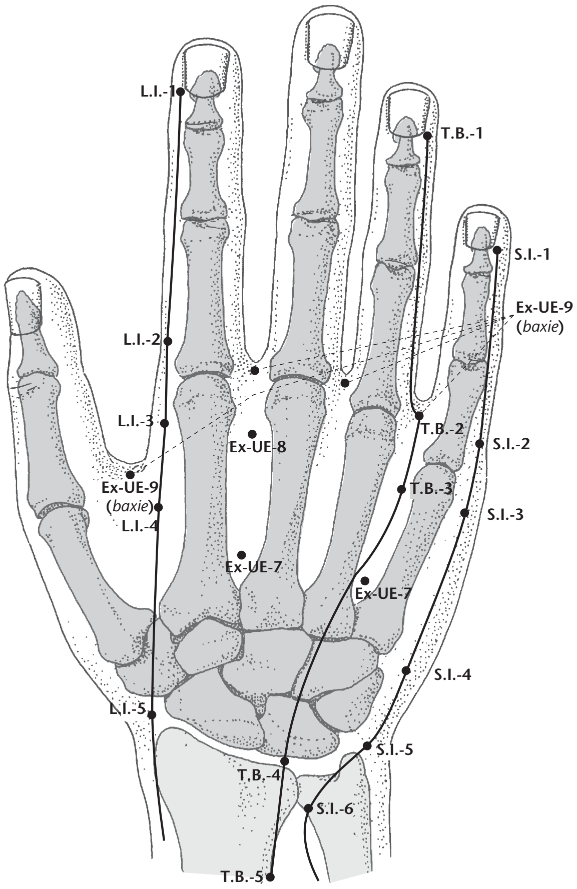
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

Ex-UE-8 *Stiff Neck* WAILAOGONG/LUOZHEN/XIANQIANG

Location

On the dorsal aspect of the hand, between the 2nd and 3rd metacarpal bones, proximal to the metacarpophalangeal joints, at the junction of the heads and the shafts of the metacarpal bones.

How to find

This point is best located with the patient making a loose fist. With the palpating finger, glide in the groove between the 2nd and 3rd metacarpal bones from the wrist joint towards the fingers until the finger comes to rest in the depression just proximal to the heads of the two metacarpal bones. Here, locate **Ex-UE-8** (*wailaogong*). Located in a comparable position between the 4th and 5th metacarpal bones is → **T.B.-3**.

Needling

0.5–1 cun vertically or slightly obliquely. Caution: Painful point! Do not stimulate this point in patients with a weak constitution: danger of needle collapse!

Actions/Indications

- Strengthens the Qi and Blood of the neck and alleviates pain

Special features

Very effective distal point for acute disorders of the cervical spine. Use strongly reducing needling techniques (for 1–2 minutes) on the affected or contralateral side while asking the patient to gently move their neck/cervical spine. The practitioner should keep the patient's hand in a fixed position and ensure that the hand and forearm are sufficiently supported.

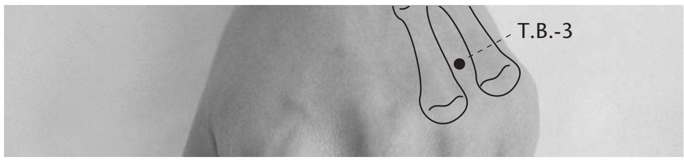

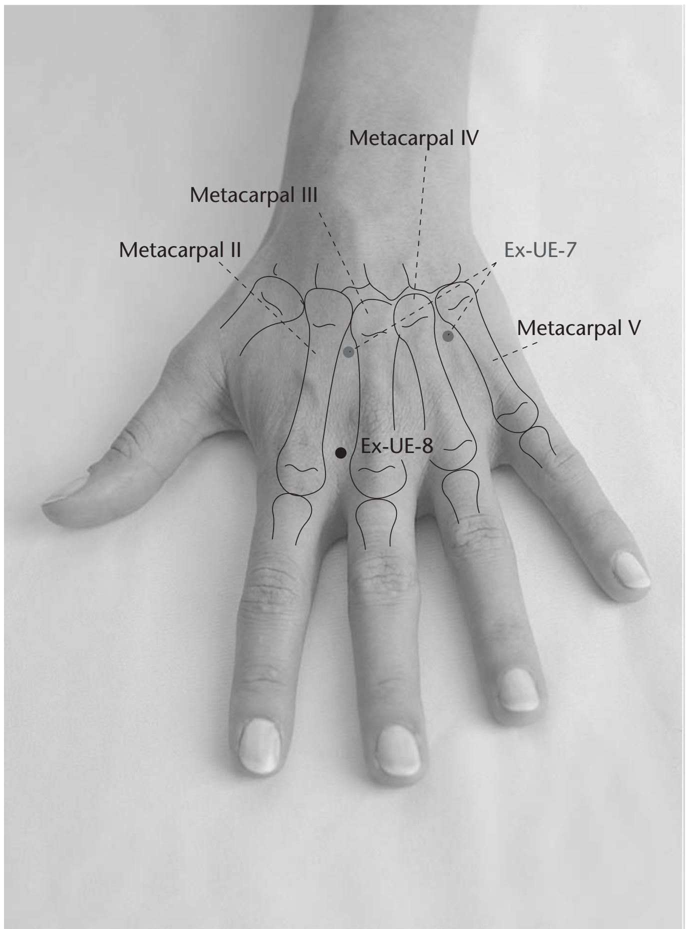
Metacarpal IV
Metacarpal III
Metacarpal II
Ex-UE-7
Metacarpal V
Ex-UE-8

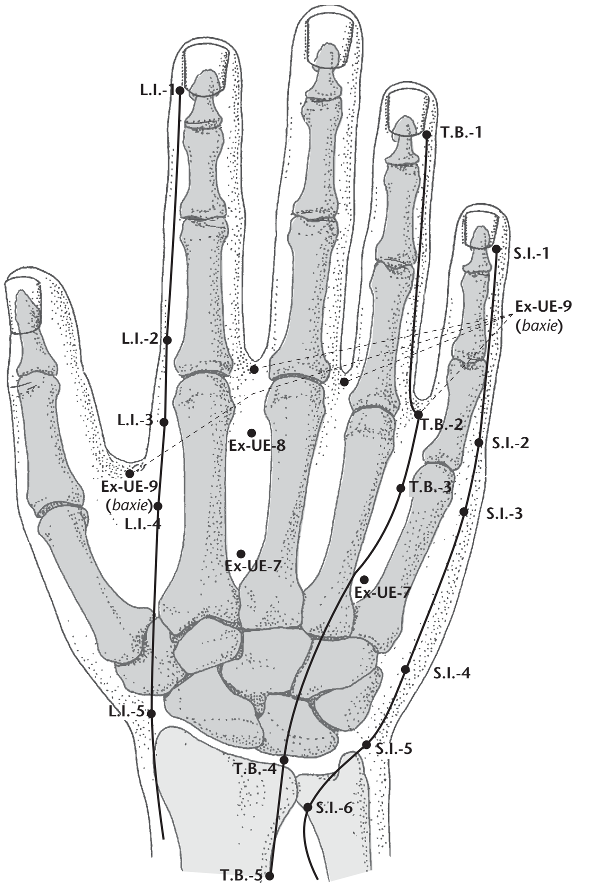
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
T.B.-2
Ex-UE-8
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

Eight Pathogens BAXIE Ex-UE-9

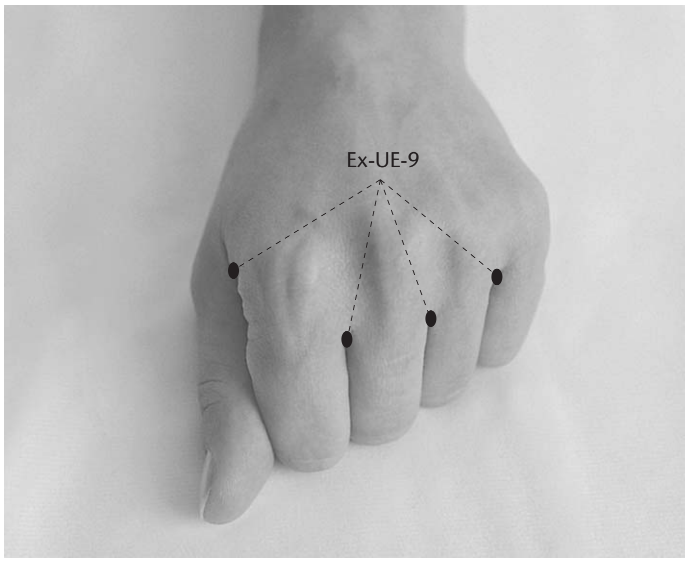

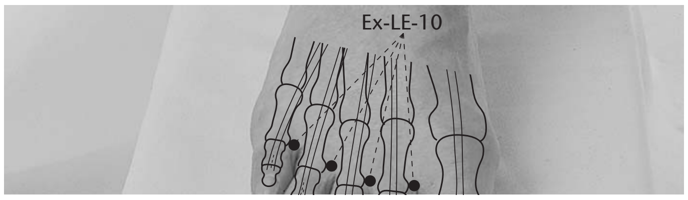

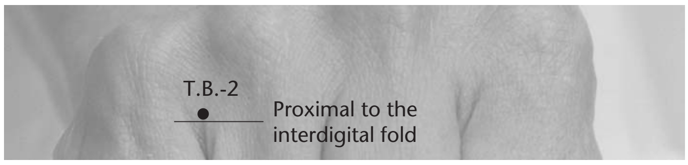

Location

Slightly proximal to the margins of the webs between the fingers, on the border of the red and white skin.

How to find

These points are best located when a loose fist is made. Locate the points on the border between the red and white skin. → **T.B.-2** (between the 4th and 5th fingers) is a point of **Ex-UE-9** (*baxie*). The points of → **Ex-LE-10** (*bafeng*) are located in a comparable position on the foot (proximal to the margins of the webs between the toes). → **LIV-2,** → **ST-44** and → **G.B.-43** are all part of → **Ex-LE-10**.

Needling

Up to 1 cun parallel to the metacarpal bones.

Actions/Indications

- Clears Heat
- Expels Cold and Dampness

Ex-UE-10 *Four Seams* SIFENG

Location

On the palmar aspect of the 2nd to 5th fingers, at the midpoint of the transverse creases of the proximal interphalangeal joints (PIP).

How to find

Locate these points on the palmar aspect of the 2nd to 5th fingers, at the midpoint of the transverse creases of the proximal interphalangeal joints (PIP).
Located in a similar position on the dorsal aspect of the little and middle fingers are → **Ex-UE-6** and → **Ex-UE-4** respectively.

Needling

Prick to bleed

Actions/Indications

- Harmonises the Qi flow between the Upper and Middle Burner

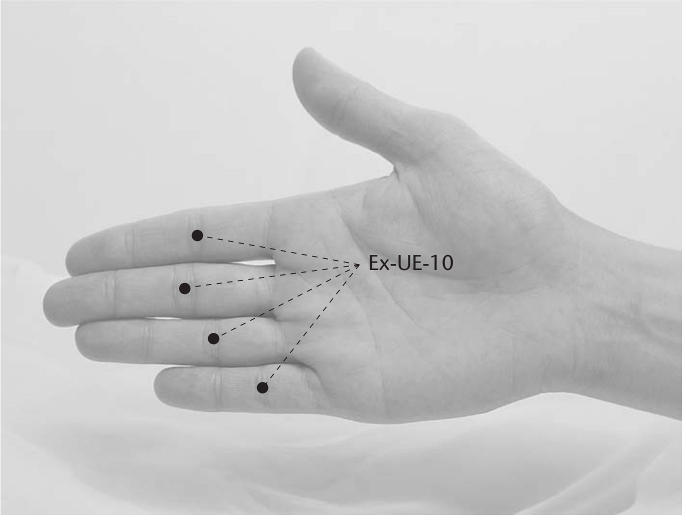

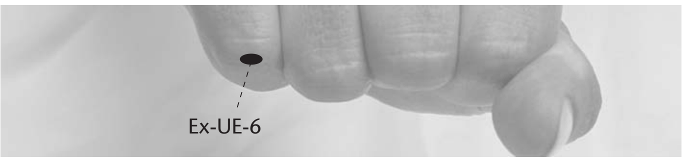

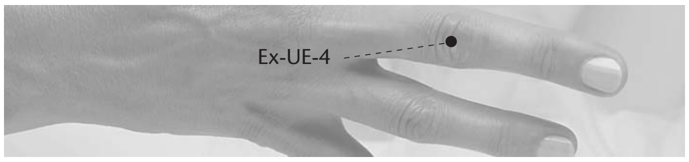

Ten Diffusions SHIXUAN Ex-UE-11

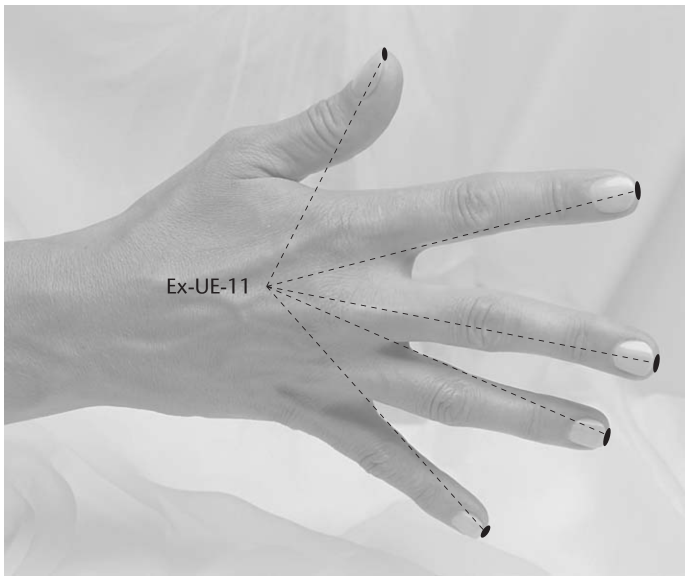

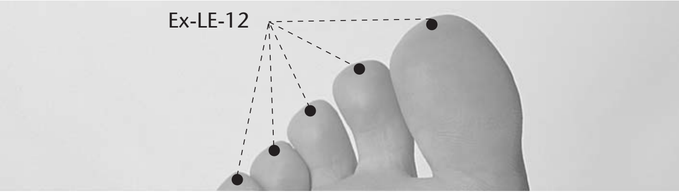

Location

On the tips of the 10 fingers.

How to find

These 10 points are located on the tips of the 10 fingers, approximately 0.1 cun from the free margin of the fingernail.
Located in a comparable position on the foot are the points of → **Ex-LE-12** (*qiduan*, on the tips of the 10 toes).

Needling

Prick to bleed

Actions/Indications

- Expels pathogenic factors

Ex-UE *Front of the Shoulder* JIANQIAN/JIANNEILING

Location

On the midpoint of a line connecting the end of the anterior axillary fold and → **L.I.-15**.

How to find

To locate → **L.I.-15**: Ask the patient to actively abduct their arm and locate **L.I.-15** in the depression anterior and inferior to the acromion (=ventrolateral extremity of the acromion). The end of the anterior axillary fold is formed by the easily palpable lower border of the pectoralis major muscle. *Jianqian* is located on the anterior aspect of the shoulder joint, midway between these two reference points.

Needling

Vertically up to 1.5 cun

Actions/Indications

- Opens the channel and *luo* vessels locally

Addendum: For more details, → Table 6.4

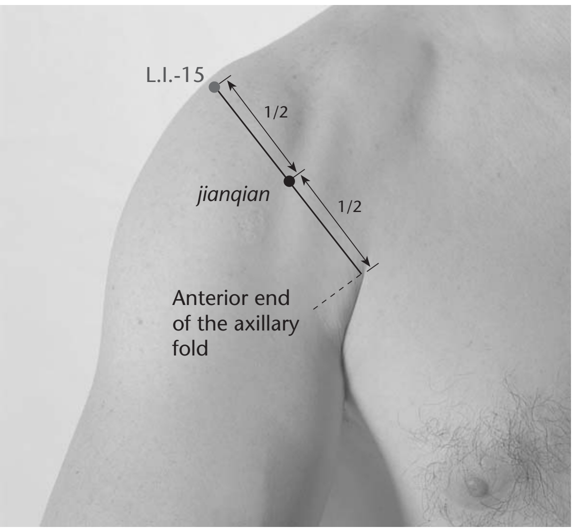
L.I.-15
1/2
jianqian
1/2
Anterior end of the axillary fold

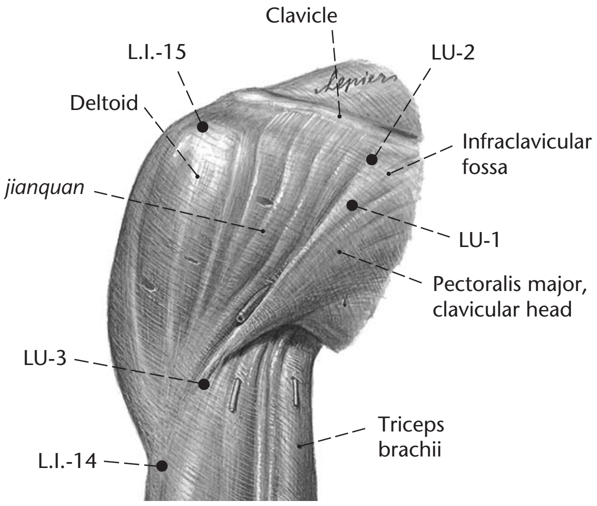
Clavicle
L.I.-15
LU-2
Deltoid
Infraclavicular fossa
jianquan
LU-1
Pectoralis major, clavicular head
LU-3
Triceps brachii
L.I.-14

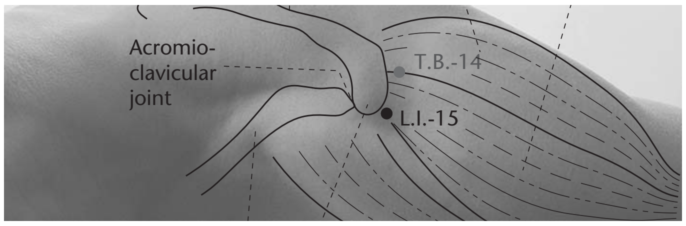
Acromio-clavicular joint
T.B.-14
L.I.-15

Arm Centre BIZHONG **Ex-UE**

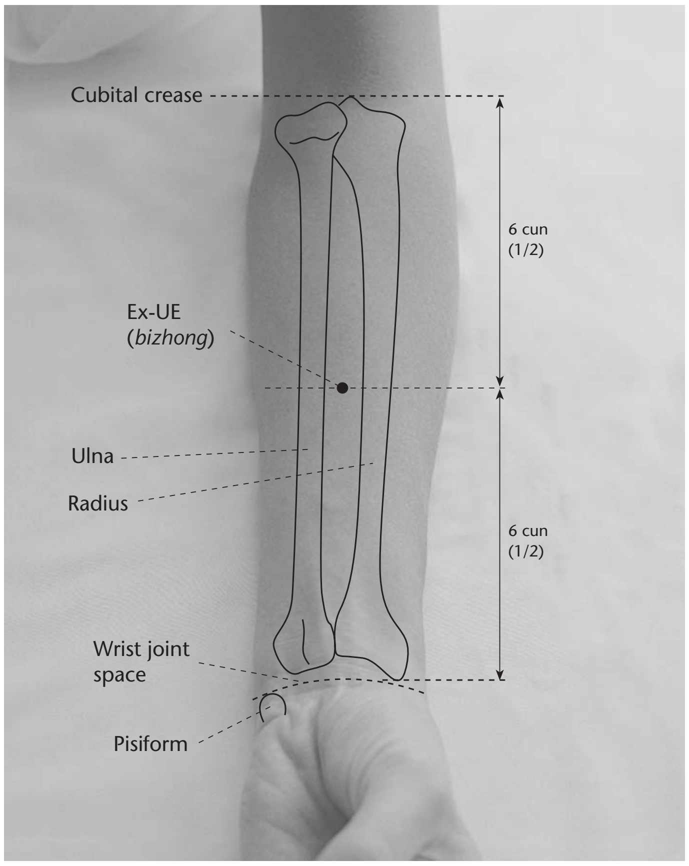
Cubital crease
6 cun
(1/2)
Ex-UE
(*bizhong*)
Ulna
Radius
6 cun
(1/2)
Wrist joint
space
Pisiform

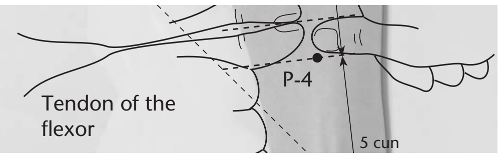
P-4
Tendon of the
flexor
5 cun

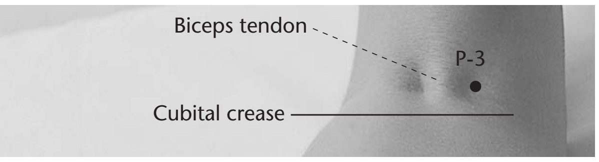
Biceps tendon
P-3
Cubital crease

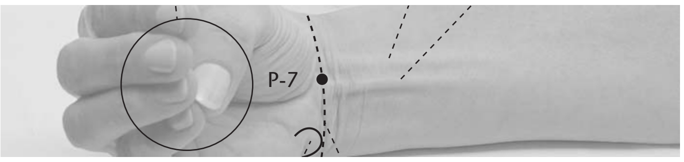

Location

On the palmar aspect of the forearm, midway between the cubital crease and the wrist joint space (‘most distal wrist crease’).

How to find

Place the patient’s forearm in a relaxed supinated position. As the location of the wrist crease may vary, the wrist joint space presents a more reliable reference point (→ 3.3.3). The spreading hands technique is the best method for locating this point (→ 2.3.3): Place the little fingers on the centre of the wrist joint (→ **P-7**) and the cubital crease (→ **P-3**) respectively and determine the midpoint of this distance by joining the thumbs. Locate **Ex-UE** (*bizhong*) on this level between the radius and the ulna.

→ **P-4** is located 1 cun distal to the midpoint of the distance between → **P-3** and → **P-7** or 5 cun proximal to → **P-7**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Opens the channel and the *luo* vessels locally

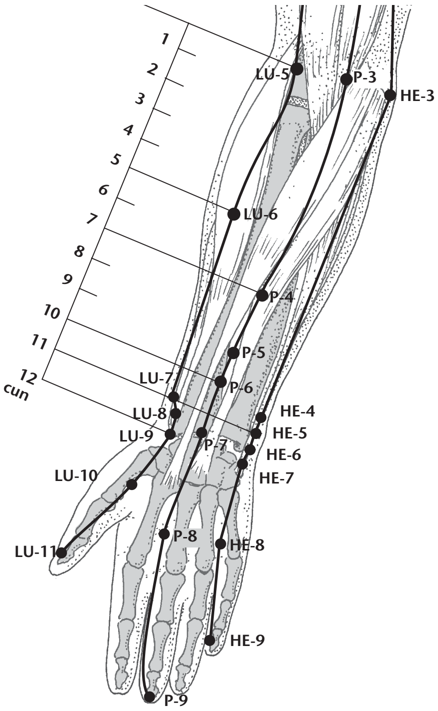
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

6.5 Extra Points: Lower Extremities (EX-LE)

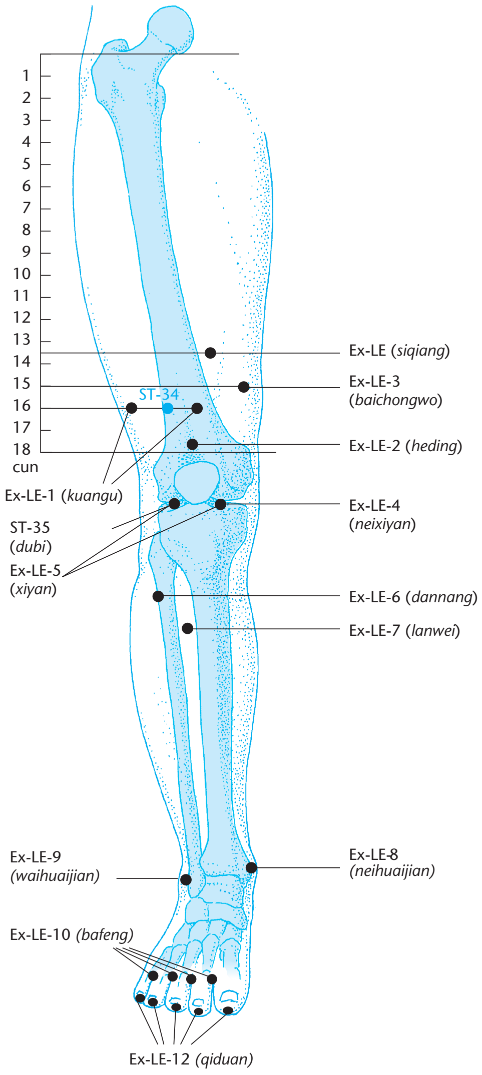
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
cun
ST-34
Ex-LE (*siqiang*)
Ex-LE-3 (*baichongwo*)
Ex-LE-2 (*heding*)
Ex-LE-1 (*kuangu*)
ST-35 (*dubi*)
Ex-LE-5 (*xiyan*)
Ex-LE-4 (*neixiyan*)
Ex-LE-6 (*dannang*)
Ex-LE-7 (*lanwei*)
Ex-LE-9 (*waihuaijian*)
Ex-LE-8 (*neihuaijian*)
Ex-LE-10 (*bafeng*)
Ex-LE-12 (*qiduan*)

English abbreviation (standard)	***Pinyin* name**	**Nguyen Van Nghi, König and Wancura, Schnorrenberger**	**Shanghai College**	**Ex (Hempen)**
Ex-LE-1	*kuangu*	PaM or ZP165 (*changgu*)	M-LE-28	
Ex-LE-2	*heding/xiding*	PaM or ZP156	M-LE-27	
Ex-LE-3	*baichongwo*	PaM or ZP163	M-LE-34	Ex21
Ex-LE-4	*neixiyan*	(PaM or ZP145)	M-LE	(Ex23)
Ex-LE-5	*xiyan*	PaM or ZP145	M-LE	Ex23
Ex-LE-6	*dannangxue*	PaM or ZP152	M-LE-23	
Ex-LE-7	*lanweixue*	PaM or ZP142	M-LE-13	Ex22
Ex-LE-8	*neihuaijian*	PaM or ZP146	M-LE-17	
Ex-LE-9	*waihuaijian*	PaM or ZP151	M-LE-22	
Ex-LE-10	*bafeng*	PaM or ZP137	M-LE-8	Ex19
Ex-LE-11	*duyin*		M-LE	
Ex-LE-12	*qiduan*		M-LE-6	
Further extra points				
Ex-LE	*huanzhong*	PaM or ZP84	M-**BW-34**	Ex20
Ex-LE	*siqiang*	NP94	**N**-LE-19	
Ex-LE	*lineiting*	PaM or ZP130	M-LE-1	

Hip Bone KUANGU **Ex-LE-1**

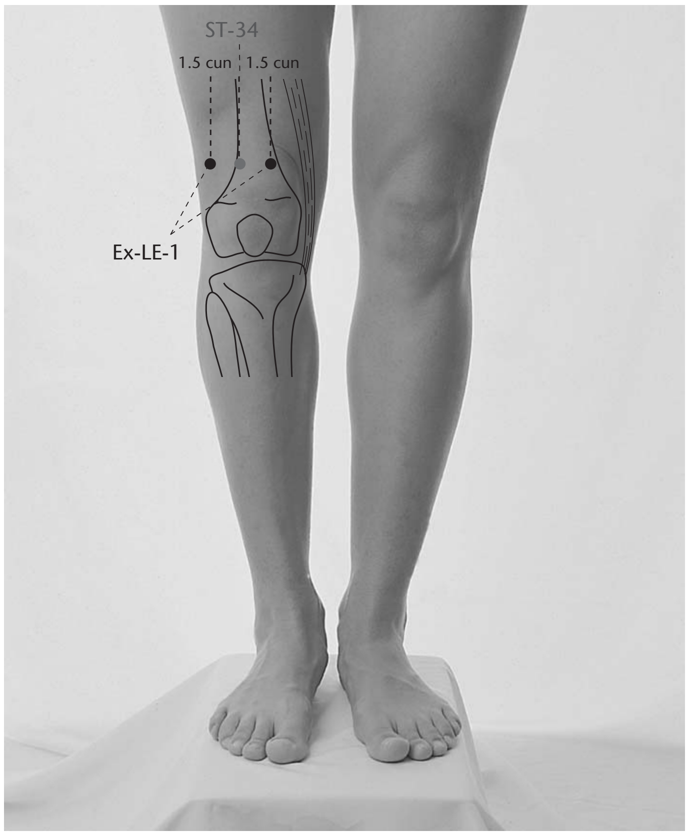
ST-34
1.5 cun
1.5 cun
Ex-LE-1

Location

A pair of points 2 cun superior to the patella and 1.5 cun lateral and medial to → **ST-34** (*liangqiu*).

How to find

First, locate → **ST-34**: measure 2 cun superiorly from the lateral upper border of the patella and there locate → **ST-34** in a depression on the vastus lateralis muscle. Then locate **Ex-LE-1** 1.5 cun lateral and medial to → **ST-34**.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Alleviate pain

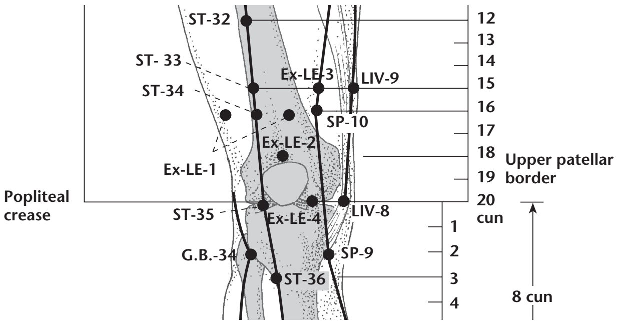
ST-32
ST- 33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
8 cun

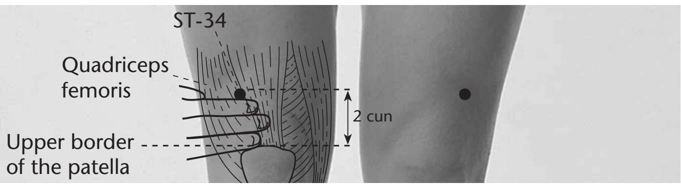
ST-34
Quadriceps femoris
2 cun
Upper border of the patella

Ex-LE-2 *Crane's Summit* HEDING/XIDING

Location

In the centre of the upper border of the patella.

How to find

This point is best located and needled with the patient's knee slightly flexed (knee support). Locate the midpoint of the upper patellar border and there locate **Ex-LE-2** (*heding*).

Needling

Vertically 0.5–0.8 cun

Actions/Indications

- Benefits the knee joint

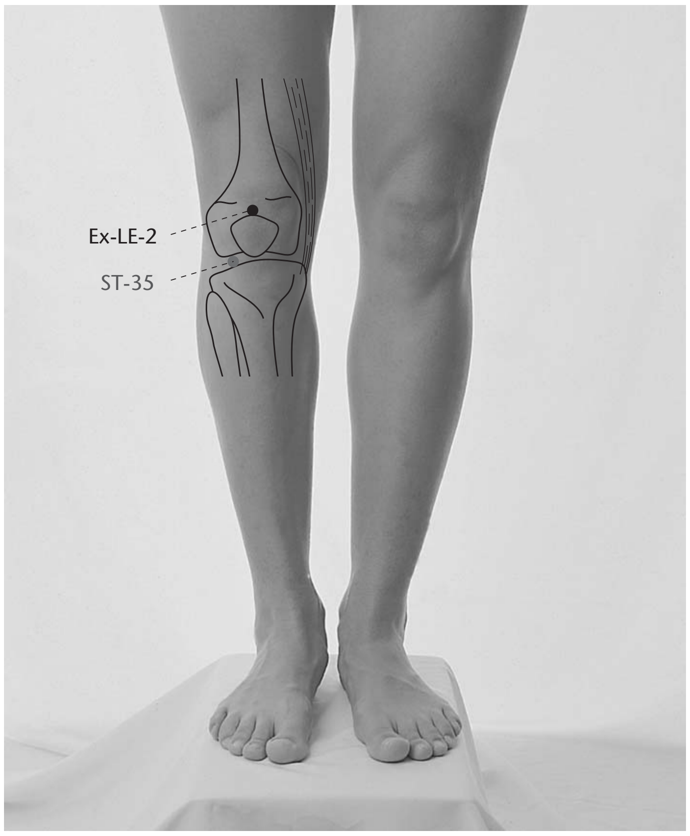

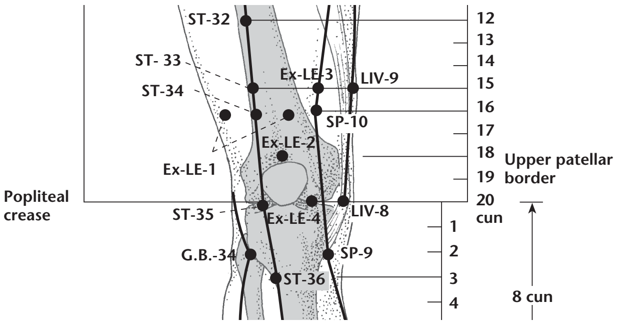
ST-32
ST- 33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20 cun
Upper patellar border
1
2
3
4
8 cun

Hundred Insect Burrow BAICHONGWO Ex-LE-3

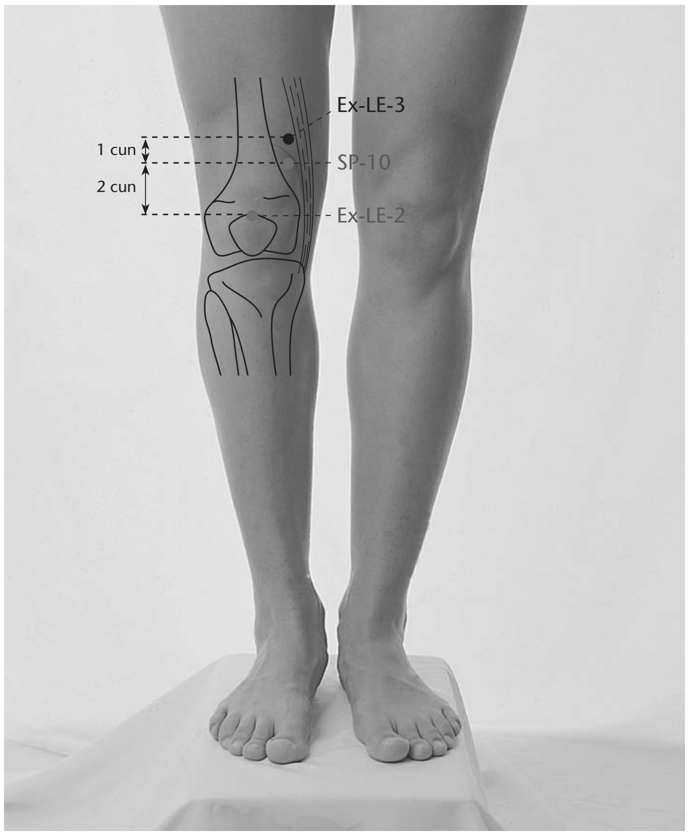

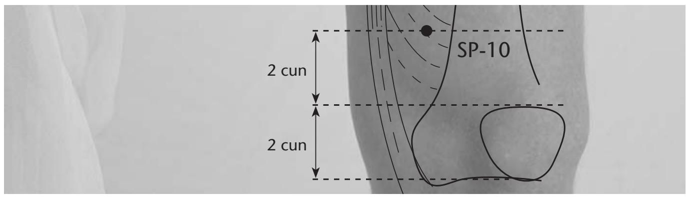

Location

3 cun superior and 1 cun medial to the upper medial border of the patella, in a small depression on the vastus medialis muscle or 1 cun superior to → **SP-10**.

How to find

The patient should be in a supine or preferably in a seated position, with their knee flexed. Locate the upper medial border of the patella and, from there, measure 3 cun (1 handsbreadth) towards proximal. There, locate **Ex-LE-3** in a small depression on the vastus lateralis muscle.

Needling

Vertically 1–2 cun

Actions/Indications

- Cools Blood-Heat, eliminates Wind, drains Dampness

Ex-LE-4 *Inner Eye of the Knee* NEIXIYAN

Location

With the knee flexed, inferior to the patella, in a depression medial to the patellar ligament.

How to find

This point is best located with the patient's knee flexed (use knee support). The medial eye of the knee can be located in a distinct depression on the level of the lower border of the patella, medial to the patellar ligament. The lateral eye of the knee corresponds to → **ST-35** (*dubi*). Together, these two points form the extra point → **Ex-LE-5** (*xiyan*).

Needling

0.5–1 cun vertically or obliquely towards the lateral eye of the knee. Caution: Knee joint.

Actions/Indications

- Disorders of the knee joint

Special features

Important local point

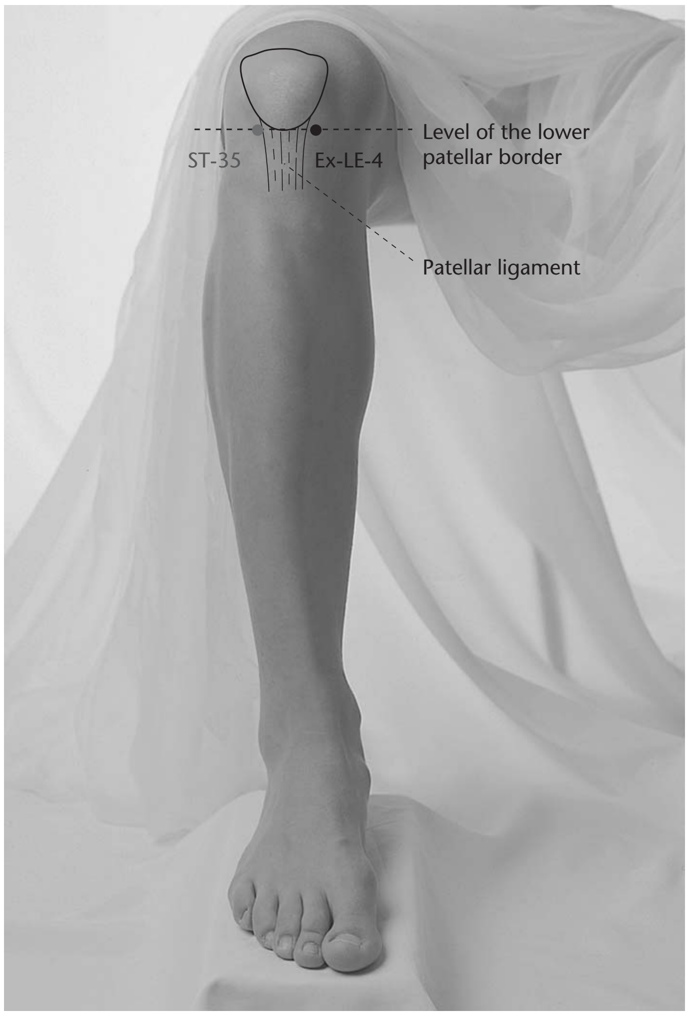
ST-35
Ex-LE-4
Level of the lower patellar border
Patellar ligament

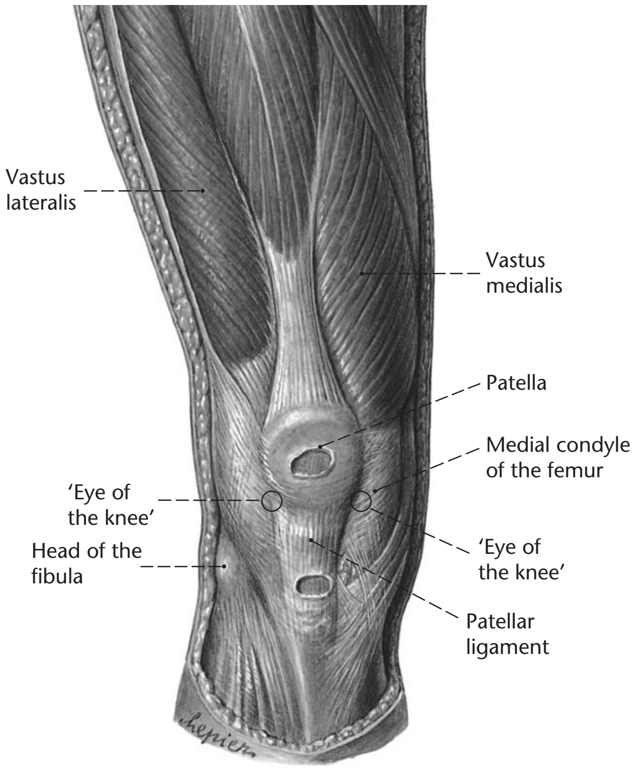
Vastus lateralis
Vastus medialis
Patella
Medial condyle of the femur
'Eye of the knee'
Head of the fibula
'Eye of the knee'
Patellar ligament

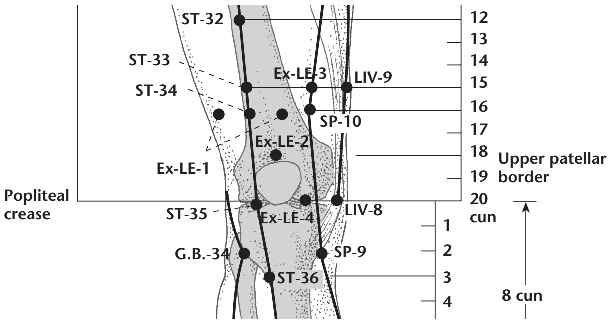
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
12
13
14
15
16
17
18
19
20
cun
Upper patellar border
1
2
3
4
8 cun

Eyes of the Knee XIYAN **Ex-LE-5**

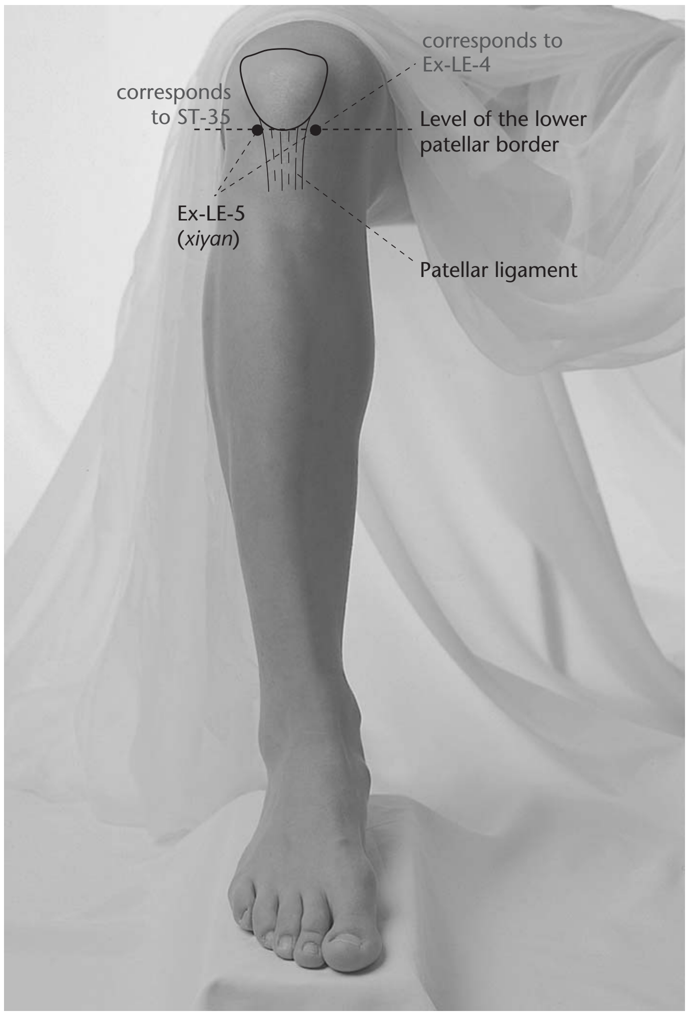

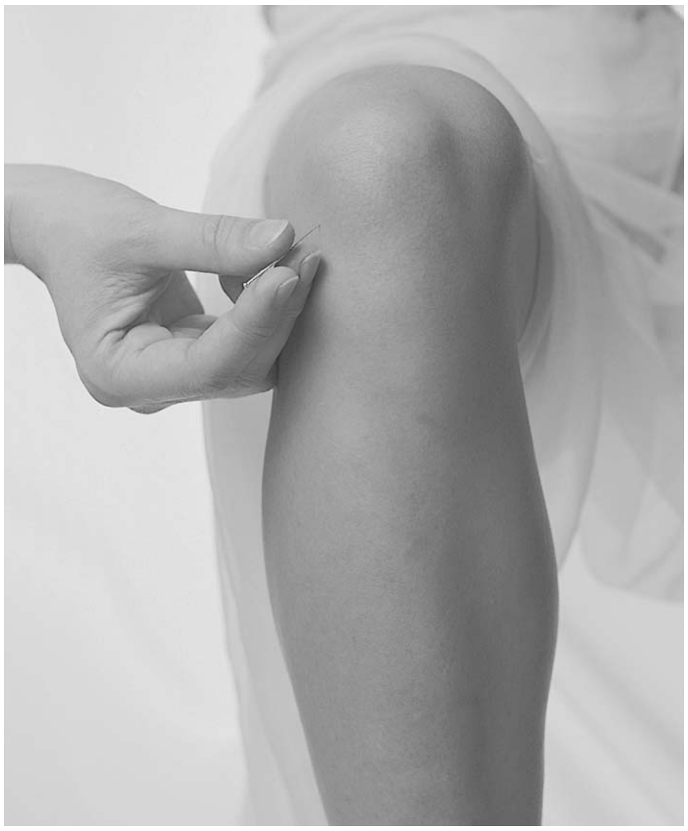

Location

With the knee flexed, this pair of points is located inferior to the patella, medial and lateral to the patellar ligament. **Ex-LE-5** (*xiyan*) includes two points: the medial eye of the knee corresponds to → **Ex-LE-4** (*neixiyan*), the lateral eye of the knee to **ST-35** (*dubi*).

How to find

This point is best located with the patient's knee flexed (use knee support). The eyes of the knee can be located on the level of the lower border of the patella, lateral and medial to the patellar ligament. The lateral eye of the knee corresponds to → **ST-35** (*dubi*), the medial eye of the knee to → **Ex-LE-4** (*neixiyan*). Together, these two points form the extra point **Ex-LE-5** (*xiyan*).

Needling

Medial eye of the knee: 0.5–1 cun vertically or obliquely towards the lateral eye of the knee; lateral eye of the knee: 0.5–1 cun vertically or obliquely towards the medial eye of the knee. Caution: Do not needle too deeply to avoid intra-articular puncture (these points correspond to the incision sites for knee arthroscopy).

Actions/Indications

- Eliminate Wind-Dampness, reduce swellings and alleviate pain

Special features

Important local points for disorders of the knee

Ex-LE-6 *Gall Bladder Point* DANNANG/DANNANGXUE/DANNANGDIAN

Location

On the G.B. channel of the **right** leg, approximately 1–2 cun distal to → **G.B.-34**, palpate for the most tender point.

How to find

First, locate → **G.B.-34** on the right leg: locate the head of the fibula on the lateral aspect of the leg (where the seam of the trousers would be) and hold it with the index and middle fingers (as with tweezers). By gliding distally with both fingers, the index finger will come to rest in a depression directly inferior and anterior to the head of the fibula (→ **G.B.-34**). From there, continue to palpate slowly along the G.B. channel, until you find the most tender point approximately 1–2 cun inferior to → **G.B.-34**. This is the location of **Ex-LE-6** (*dannangxue*).

Needling

Vertically 1–1.5 cun

Actions/Indications

- Clears Heat and drains Dampness for acute and chronic disorders of the Gall Blader such as cholecystitis and cholelithiasis (tenderness on this point is diagnostically significant).

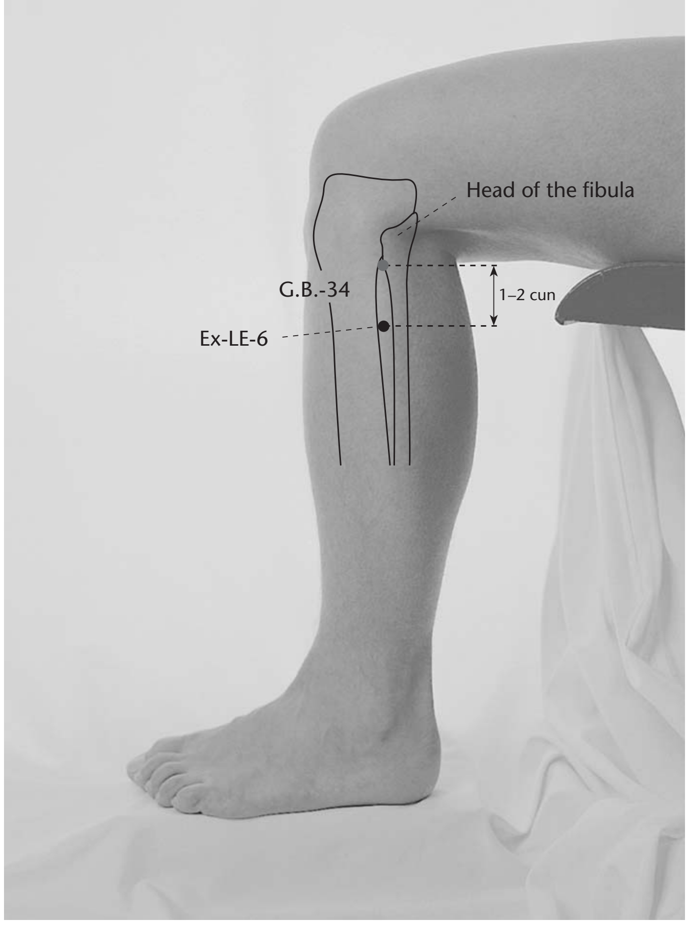

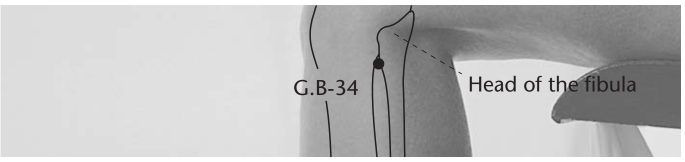

Appendix Point LANWEI/LANWEIXUE Ex-LE-7

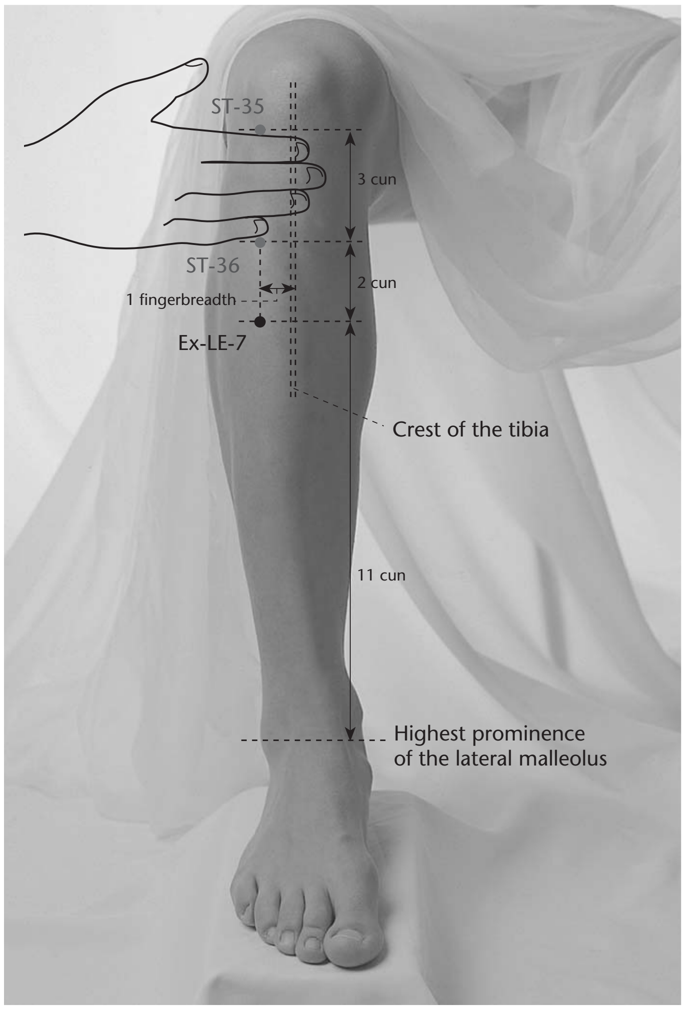
ST-35
3 cun
ST-36
2 cun
1 fingerbreadth
Ex-LE-7
Crest of the tibia
11 cun
Highest prominence
of the lateral malleolus

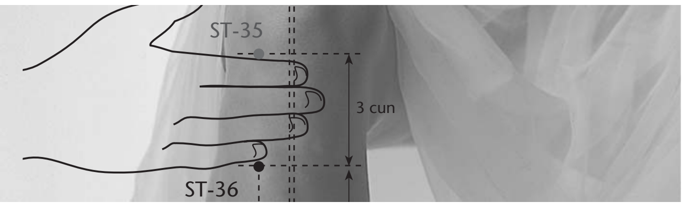
ST-35
3 cun
ST-36

Location

On the ST channel of the **right** leg, the most tender point approximately 2 cun distal to ➞ **ST-36**.

How to find

Locate this point on the right leg only (this corresponds to the unilateral location of the appendix on the abdomen). First, locate ➞ **ST-36** 3 cun inferior to the knee joint space (➞ **ST-35**) and 1 cun lateral to the anterior crest of the tibia. Then continue to palpate along the ST channel for approximately 2 cun (towards ➞ **ST-37**, 3 cun inferior to ➞ **ST-36**) and locate **Ex-LE-7** (*lanweixue*) on the most tender point.

Needling

Vertically 1–1.5 cun

Actions/Indications

- Activates the Qi and Blood flow in the Large Intestine, clears Heat and Heat toxins from the Large Intestine

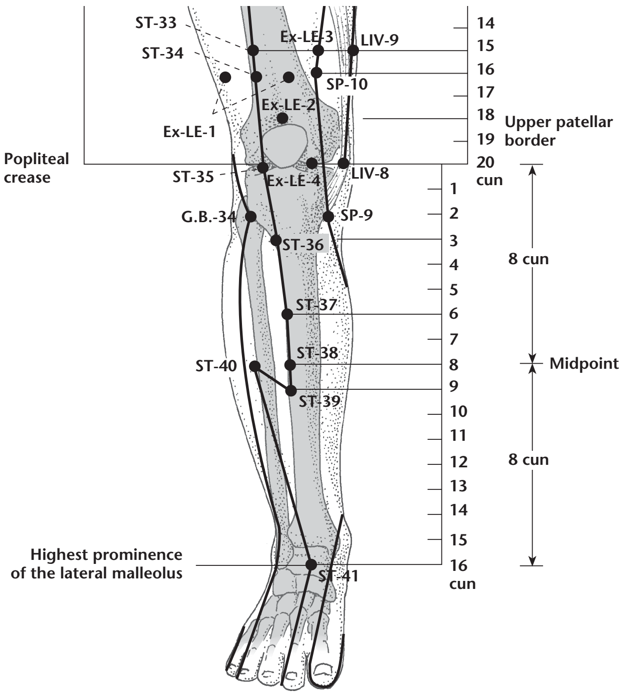
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
Highest prominence of the lateral malleolus
ST-41
14
15
16
17
18
19
20 cun
Upper patellar border
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16 cun
8 cun
Midpoint
8 cun

Ex-LE-8 *Medial Malleolus Tip* NEIHUAIJIAN

Location

On the highest point of the medial malleolus.

How to find

Find the highest prominence of the medial malleolus and there locate **Ex-LE-8** (*neihuaijian*). Located in a comparable position but on the lateral malleolus is → **Ex-LE-9** (*waihuaijian*).

Needling

0.1 cun transversely (subcutaneously) or prick to bleed.

Actions/Indications

- Pain in the medial ankle region

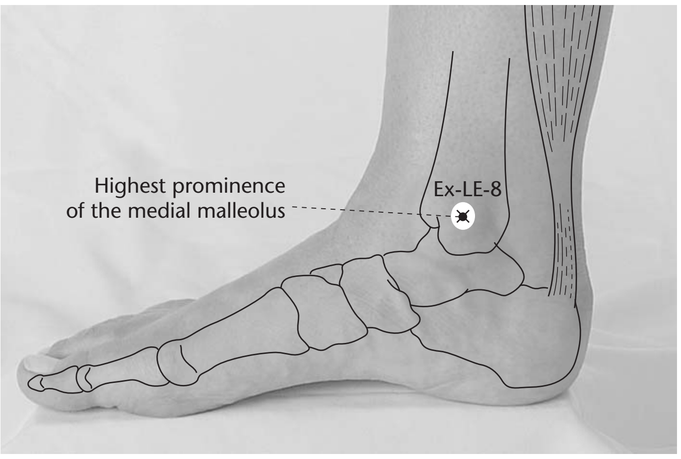
Highest prominence
of the medial malleolus
Ex-LE-8

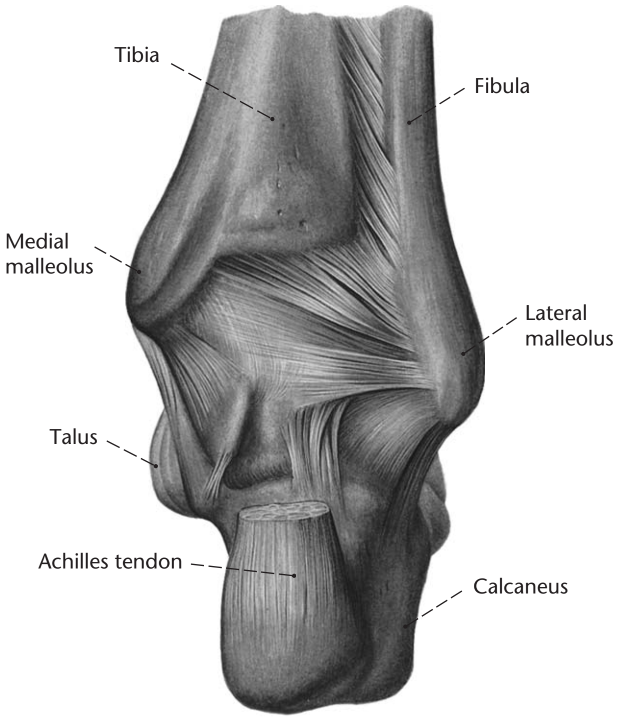
Tibia
Fibula
Medial
malleolus
Lateral
malleolus
Talus
Achilles tendon
Calcaneus

Outer Malleolus Tip WAIHUAIJIAN Ex-LE-9

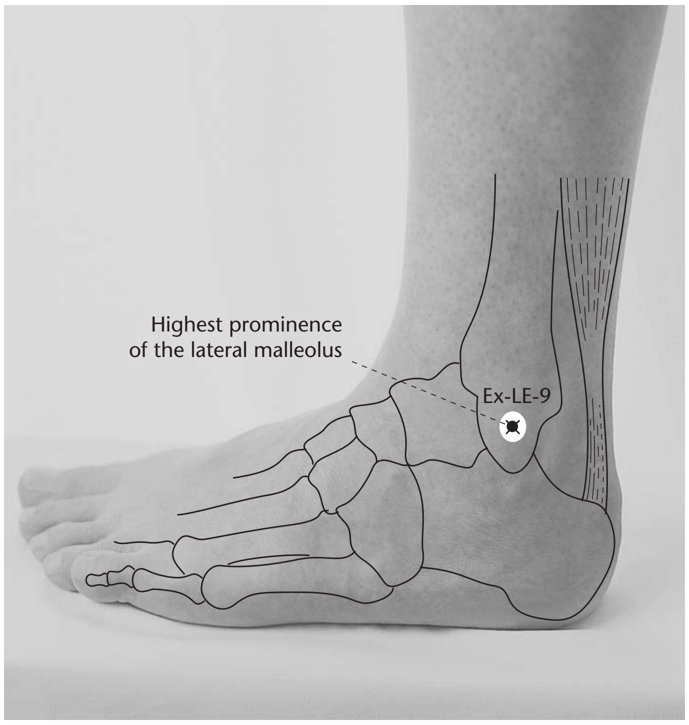

Location

On the highest prominence of the lateral malleolus

How to find

Find the highest prominence of the lateral malleolus and there locate **Ex-LE-9** (*waihuaijian*). Located in a comparable position but on the medial malleolus is → **Ex-LE-8** (*neihuaijian*).

Needling

0.1 cun transversely (subcutaneously) or prick to bleed.

Actions/Indications

- Pain in the lateral ankle region

Ex-LE-10 *Eight Winds* BAFENG

Location

Eight points located on the dorsum of the foot, slightly proximal to the margins of the interdigital webs.

How to find

Locate the points of **Ex-LE-10** (*bafeng*) on the dorsum of the foot, proximal to the margins of the webs between the toes, on the border of the red and white skin. → **LIV-2** (*xingjian*), → **ST-44** (*neiting*) and → **G.B.-43** (*xiabai*) are all part of the *bafeng* points. The extra points → **Ex-UE-9** (*baxie*) are located in a comparable position on the dorsum of the hand, proximal to the margins of the webs between the fingers (→ **T.B.-2** is part of the → **Ex-UE-9**).

Needling

0.3–1 cun obliquely in a proximal direction or prick to bleed.

Actions/Indications

- Clears Heat and reduces swelling

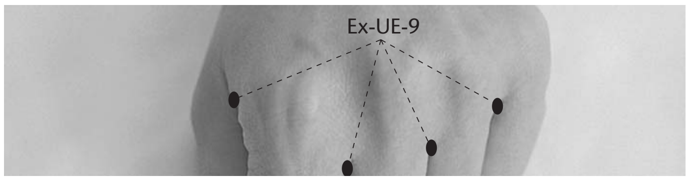

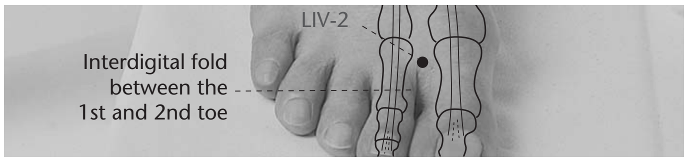
LIV-2
Interdigital fold between the 1st and 2nd toe

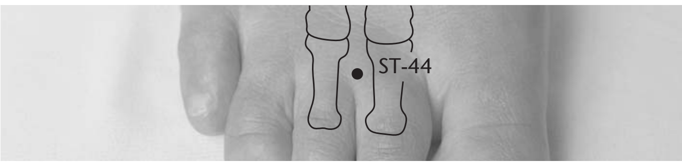

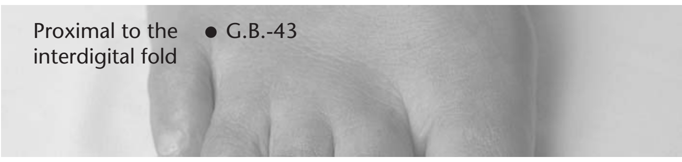
Proximal to the interdigital fold
G.B.-43

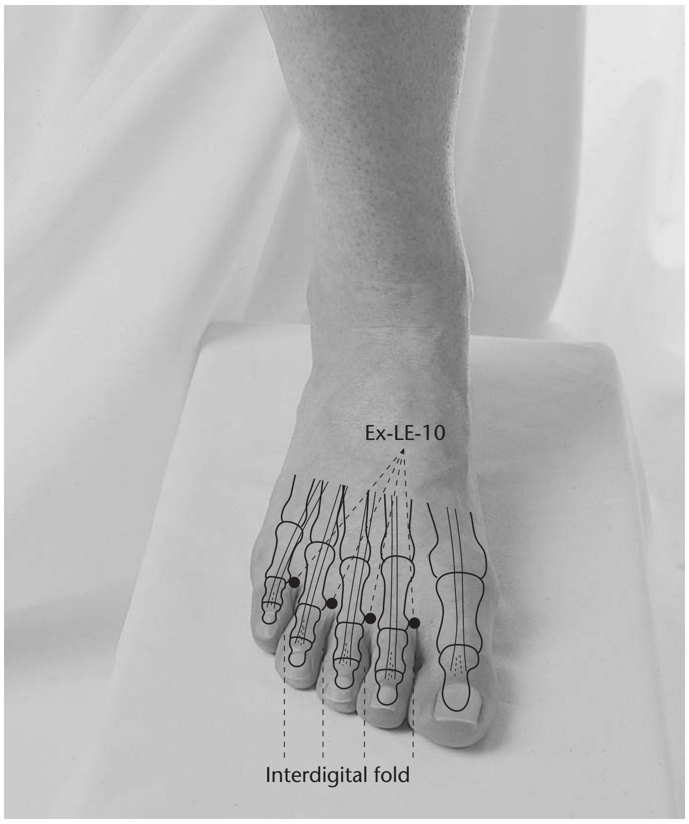
Ex-LE-10
Interdigital fold

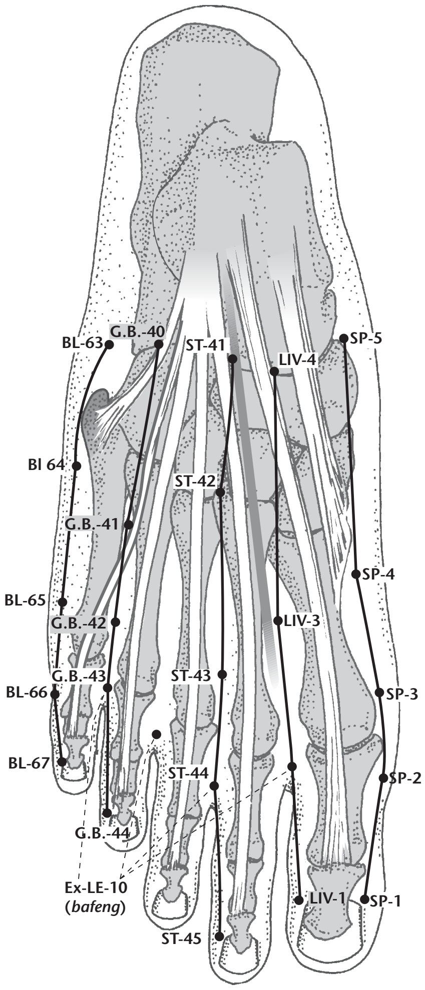
BL-63
G.B.-40
ST-41
LIV-4
SP-5
Bl 64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
G.B.-43
ST-43
BL-66
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10 (bafeng)
LIV-1
SP-1
ST-45

Solitary Yin DUYIN Ex-LE-11

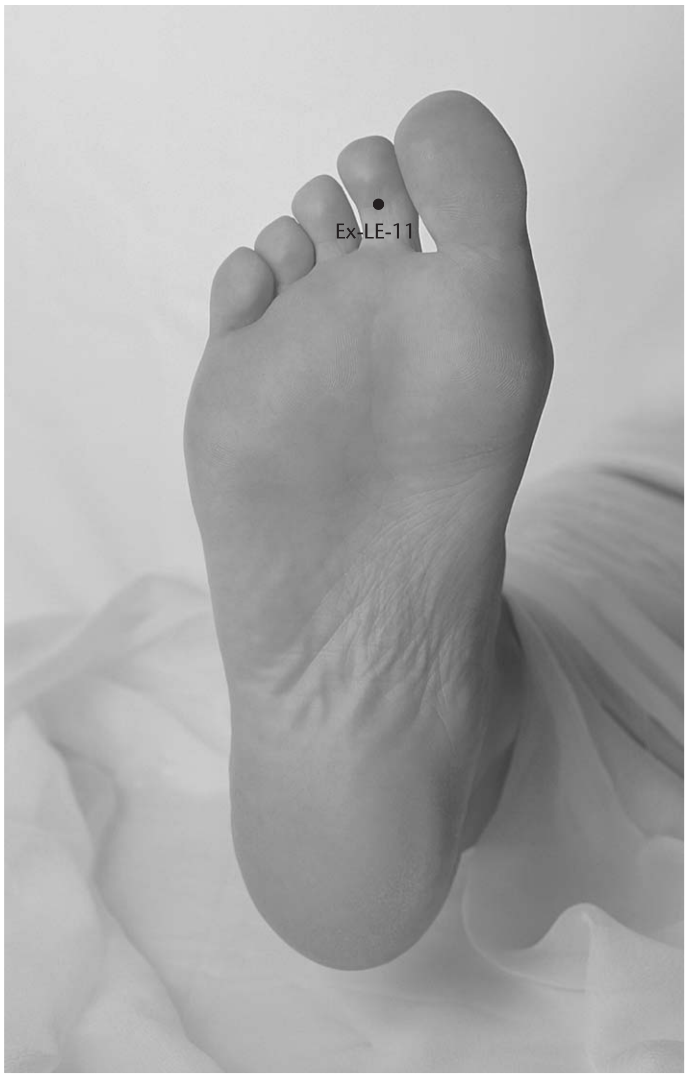

Location

On the plantar aspect of the 2nd toe, at the midpoint of the transverse crease of the distal interphalangeal joint.

How to find

Locate the plantar aspect of the 2nd toe. There, locate **Ex-LE-11** (*duyin*) at the midpoint of the transverse crease of the distal interphalangeal joint.

Needling

0.2–0.3 cun vertically or transversely (subcutaneously) or prick to bleed or moxibustion.

Actions/Indications

- Acute angina, thoracic and hypochondriac pain, nausea, vomiting, retention of the lochia, irregular menstruation, inguinal hernia

Ex-LE-12 *End of Qi* QIDUAN

Location

On the tips of the 10 toes.

How to find

Locate the points of **Ex-LE-12** (*qiduan*) on the tips of the 10 toes, 0.1 cun from the free margin of the nail.
The points of → **Ex-UE-11** (*shixuan*) are located in a comparable position on the hand, on the tips of the 10 fingers.

Needling

Prick to bleed

Actions/Indications

- Syncope, oedema of the feet, acute abdominal pain

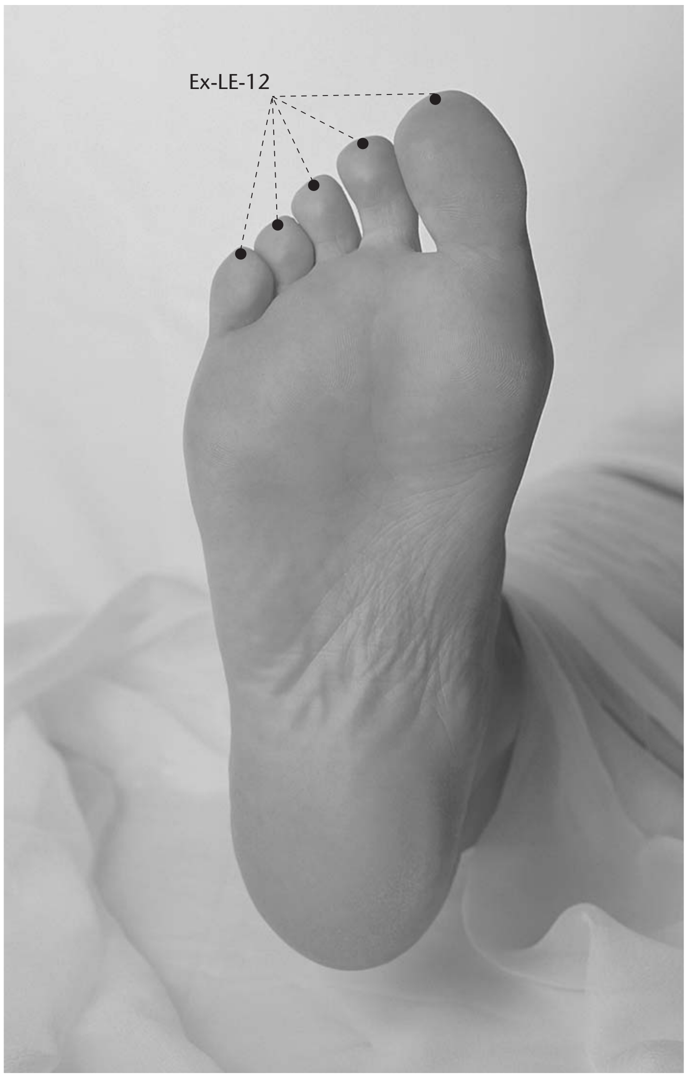

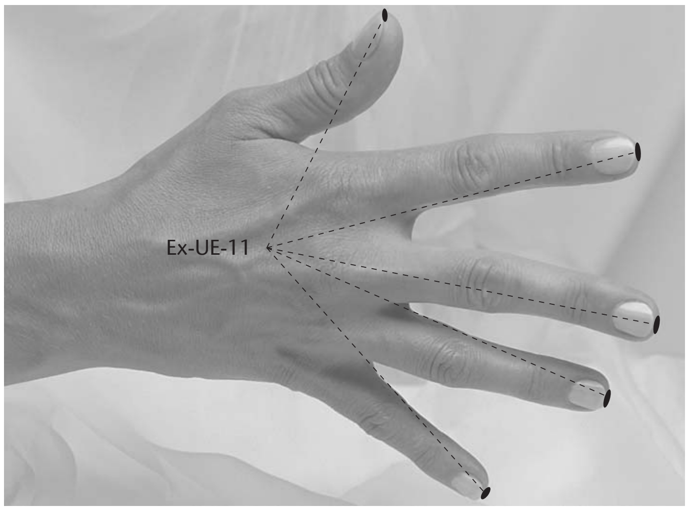

Circle Centre HUANZHONG Ex-LE

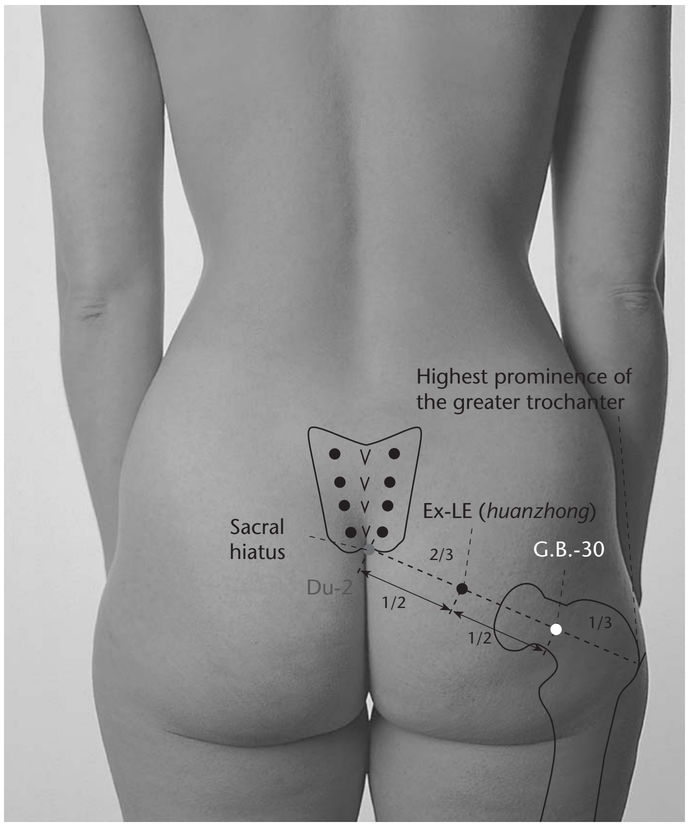
Highest prominence of
the greater trochanter
Ex-LE (*huanzhong*)
Sacral
hiatus
2/3
G.B.-30
Du-2
1/2
1/2
1/3

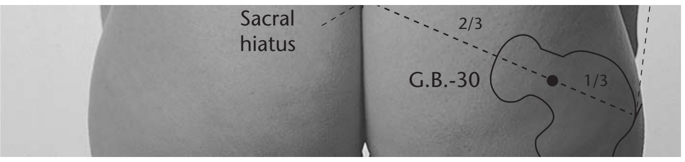
Sacral
hiatus
2/3
G.B.-30
1/3

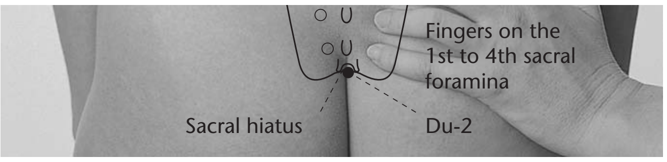
Fingers on the
1st to 4th sacral
foramina
Sacral hiatus
Du-2

Location

Midway between → **G.B.-30** (*huantiao*) and → **Du-2** (*yaoshu*).

How to find

First, locate → **G.B.-30**: divide the distance between the prominence of the greater trochanter and the sacral hiatus into thirds (→ 3.4.4) and locate → **G.B.-30** a third of the distance from the greater trochanter. Then locate → **Du-2** directly below the sacral hiatus. Next, locate **Ex-LE** (*huanzhong*) midway between these two points.

Needling

Vertically 2–2.5 cun

Actions/Indications

- Sciatica, urinary tract infection, haemorrhoids, paralysis of the lower extremities

Special features

- This point is not part of the WHO classification of extra points.

Ex-LE Four Muscles Strengthening Point SIQIANG

Location

4.5 cun superior to the centre of the upper patellar border.

How to find

Find the centre of the upper patellar border. From there, measure 4.5 cun (1 handbreadth and 2 fingerbreadths) in a superior direction. **Or:** The distance between the upper border of the pubic symphysis and the upper border of the patella is 18 cun (→ 2.2). Divide this distance into quarters (spreading hands technique, elastic tape → Chapter 2) and **Ex-LE** (*siqiang*) is located on the centre of the thigh, one quarter of the distance from the upper patellar border.

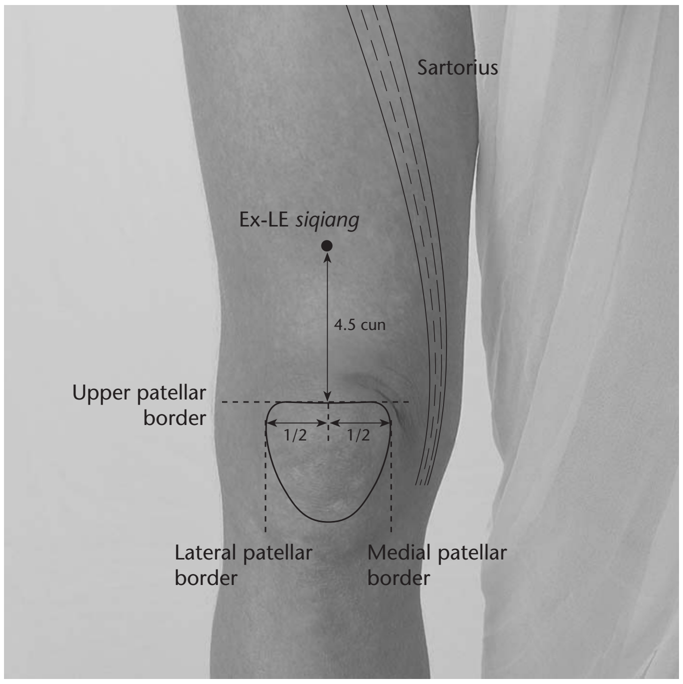
Sartorius
Ex-LE *siqiang*
4.5 cun
Upper patellar border
1/2
1/2
Lateral patellar border
Medial patellar border

Needling

Vertically 1–2 cun

Actions/Indications

- Paralysis and atrophy of the muscles of the lower extremities, especially of the quadriceps femoris muscle

Special features

'Four muscles' refers to the four portions of the quadriceps femoris muscle. This point is not part of the WHO classification of extra points.

Plantar Inner Space LINEITING Ex-LE

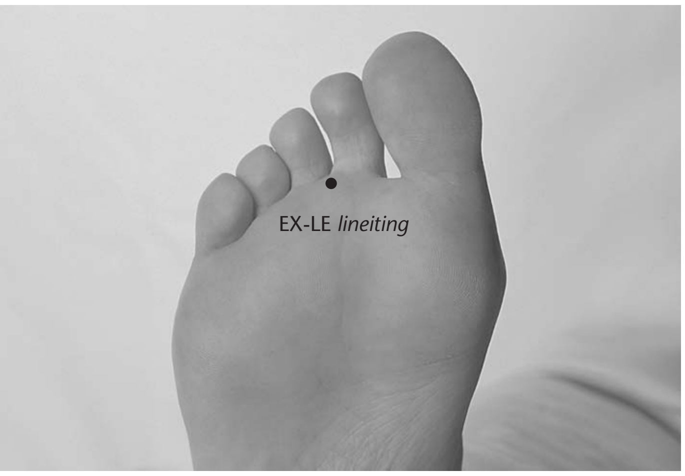

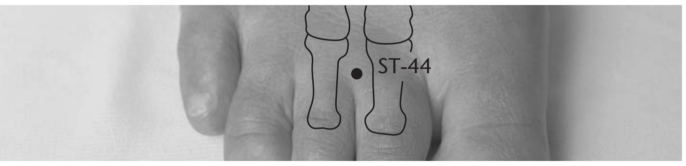

Location

On the plantar aspect of the foot, between the 2nd and 3rd metatarsal bones, opposite → **ST-44** (*neiting*).

How to find

Locate the space between the 2nd and 3rd toes on the plantar aspect of the foot. The extra point *lineiting* lies opposite → **ST-44** (on the dorsal aspect of the foot, proximal to the margin of the web between the 2nd and 3rd toe).

Needling

0.2–0.3 cun vertically towards → **ST-44**. Caution: Painful point.

Actions/Indications

- Acute epigastric pain, local pain, epilepsy, restlessness

Special features

This point is not part of the WHO classification of extra points.

7 Important Points According to Region

Claudia Focks

7.1 Anterior Aspect of the Head

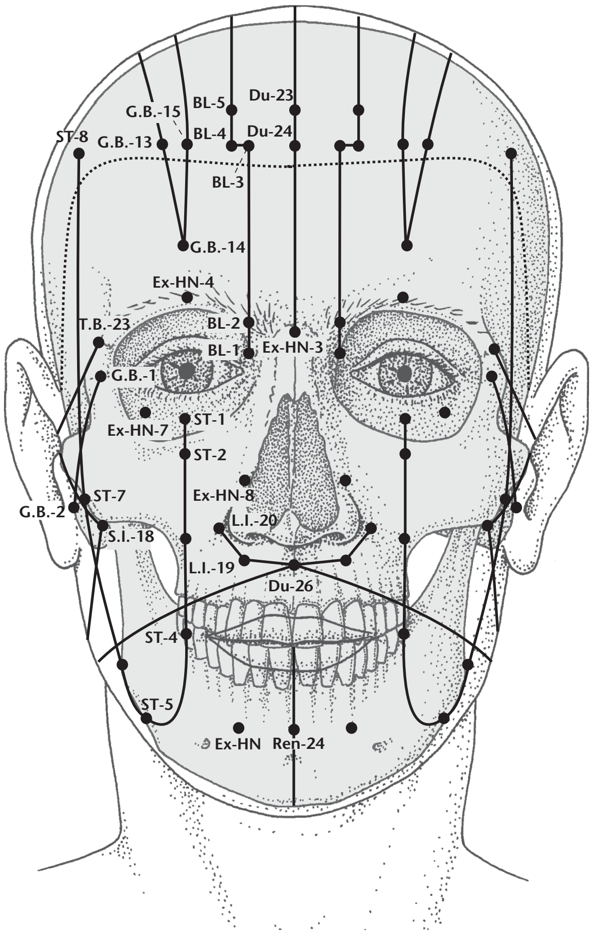
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
Ex-HN-7
ST-1
ST-2
ST-7
Ex-HN-8
G.B.-2
S.I.-18
L.I.-20
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

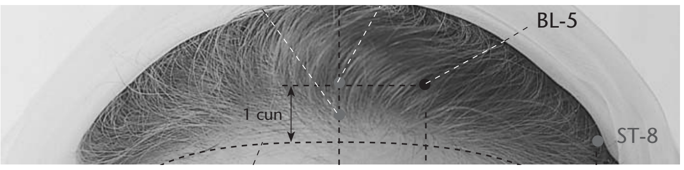
BL-5
1 cun
ST-8

BL-5 (*wuchu*) 'Fifth Place' 1 cun superior to the anterior hairline and 1.5 cun lateral to the midline (directly above **BL-4**, level with **Du-23**)

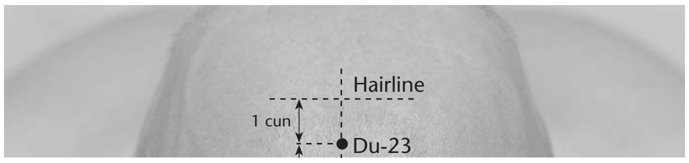
Hairline
1 cun
Du-23

Du-23 (*shangxing*) 'Upper Star' On the midline, 1 cun superior to the anterior hairline

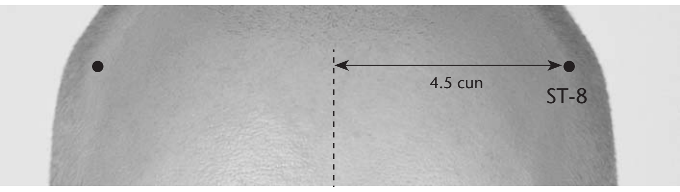

ST-8 (*touwei*) 'Head's Binding' At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (**Du-24**)

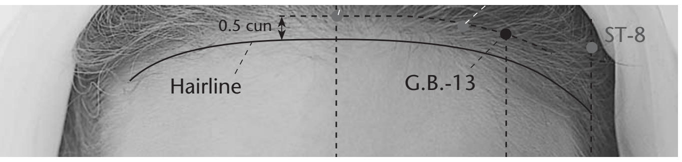
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*) 'Root of the Spirit' 0.5 cun superior to the anterior hairline and 3 cun lateral to the midline, on the line between **Du-24** and **ST-8** (= 4.5 cun) at the junction of the lateral and medial third

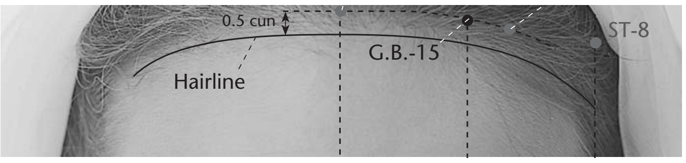
0.5 cun
ST-8
G.B.-15
Hairline

G.B.-15 (*toulinqi*) 'Head Governor of Tears' On the pupil line when looking straight ahead, 0.5 cun superior to the anterior hairline, at the midpoint of the line between **Du-24** and **ST-8** (= 4.5 cun)

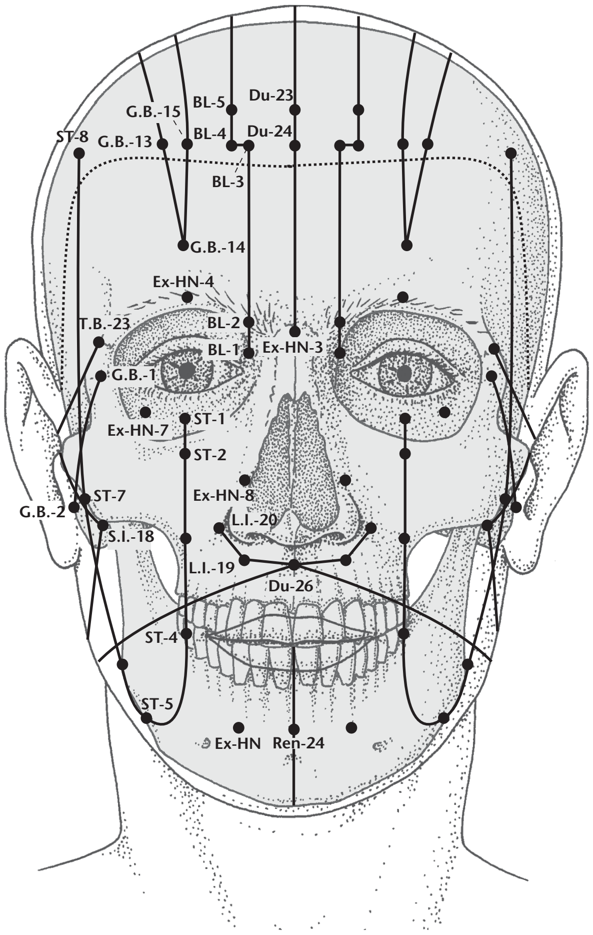
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
ST-7
Ex-HN-8
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

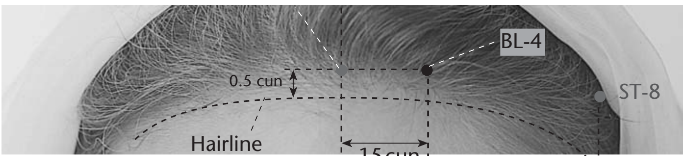
BL-4
ST-8
0.5 cun
Hairline
1.5 cun

BL-4 (*qucha*) 'Crooked Curve' 0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one third of the distance between **Du-24** and **ST-8** (= 4.5 cun)

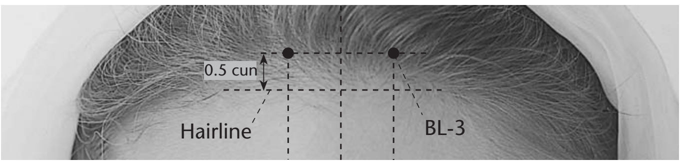
0.5 cun
Hairline
BL-3

BL-3 (*meichong*) 'Eyebrows' Pouring' 0.5 cun within the anterior hairline, superior to the medial canthus of the eye (**BL-1**)

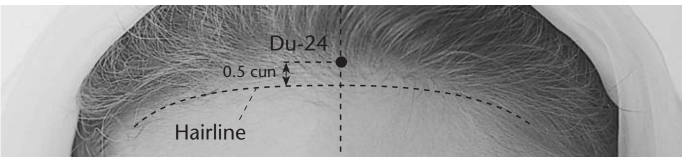
Du-24
0.5 cun
Hairline

Du-24 (*shenting*) 'Courtyard of the Spirit' On the midline, 0.5 cun superior to the anterior hairline

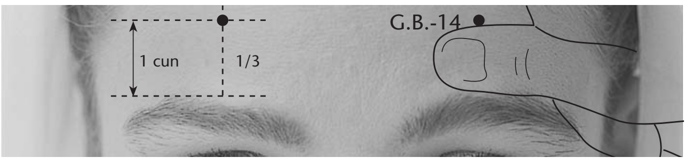
G.B.-14
1 cun
1/3

G.B.-14 (*yangbai*) 'Yang White' When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow (centre of the eyebrow to anterior hairline = 3 cun)

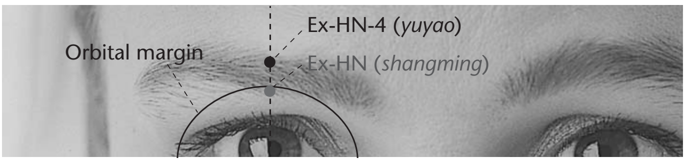
Ex-HN-4 (*yuyao*)
Orbital margin
Ex-HN (*shangming*)

Ex-HN-4 (*yuyao*) 'Fish Waist' In the centre of the eyebrow, with the eyes looking straight ahead directly above the pupil

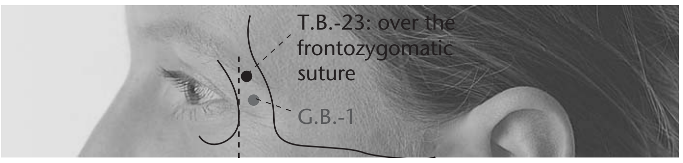
T.B.-23: over the frontozygomatic suture
G.B.-1

T.B.-23 (*sizhukong*) 'Silken Bamboo Hollow' In the depression (frontozygomatic suture) at the lateral end of the eyebrow

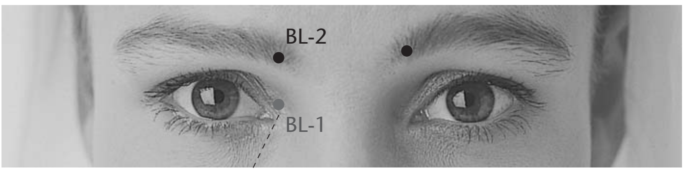

BL-2 (*zanzhu*) 'Gathered Bamboo' In a depression at the medial end of the eyebrow, directly superior to the inner canthus of the eye

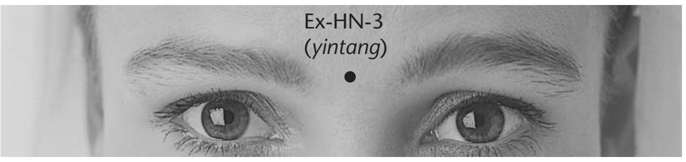
Ex-HN-3
(*yintang*)

Ex-HN-3 (*yintang*) 'Hall of Impression' On the anterior midline between the eyebrows

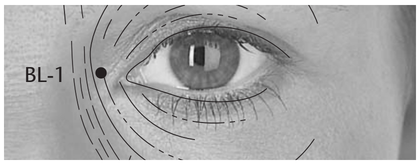

BL-1 (*jingming*) 'Bright Eyes' 0.1 cun superior to the inner canthus of the eye, where spectacles tend to rest

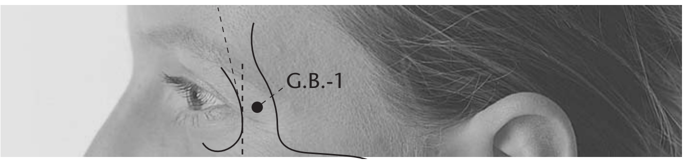

G.B.-1 (*tongziliao*) 'Pupil Crevice' 0.5 cun lateral to the outer canthus of the eye

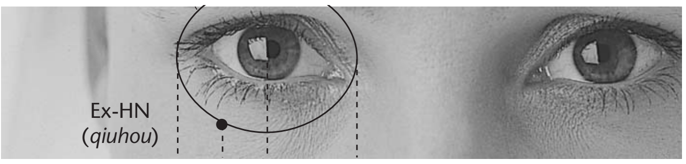

Ex-HN-7 (*qiuhou*) 'Behind the Ball' On the lower border of the orbit, at the junction of the lateral quarter and the medial three-quarters of the orbital margin

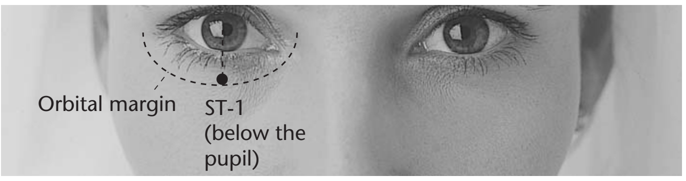

ST-1 (*chengqi*) 'Container of Tears' With the eyes looking straight forward, directly below the centre of the pupil, between the eyeball and the infraorbital ridge

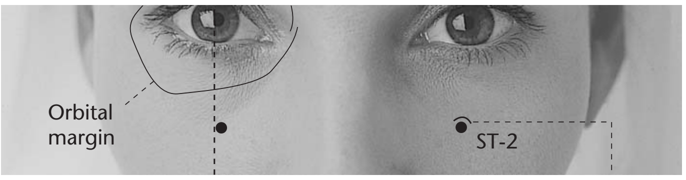

ST-2 (*sibai*) 'Four Whites' When looking straight ahead, directly below the centre of the pupil, in the depression at the infraorbital foramen

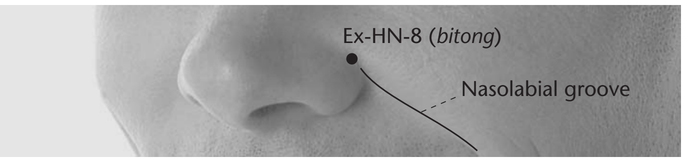

Ex-HN-8 (*shangyingxiang or bitong*) 'Upper Yingxiang' or 'Free Nose' At the upper end of the nasolabial groove, at the junction of the maxilla and the nasal cavity (at the border between the nasal bone and cartilage)

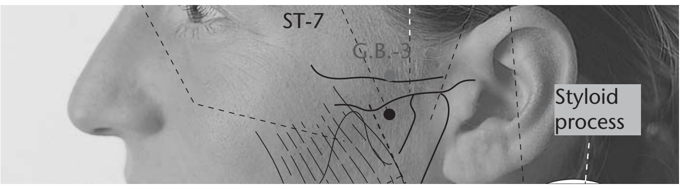

ST-7 (*xiaguan*) 'Below the Joint' With the mouth closed in the depression of the mandibular incisure, between the coronoid and condyloid process of the mandible

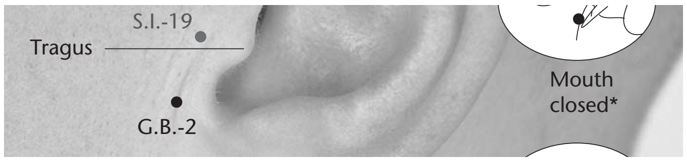

G.B.-2 (*tinghui*) 'Meeting of Hearing' Anterior to the ear, with the mouth open in the depression at the level of the intertragic notch, on the lower border of the condyloid process of the mandible.

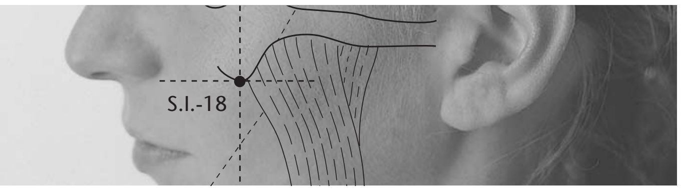

S.I.-18 (*quanliao*) 'Cheek Bone Crevice' In a depression at the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone

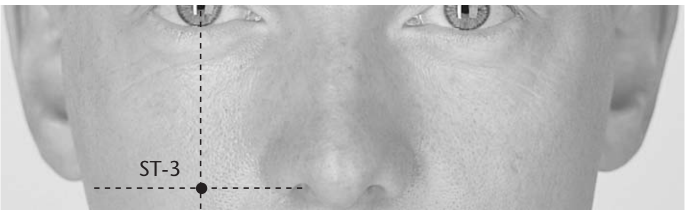

ST-3 (*juliao*) 'Great Crevice' When looking straight ahead, directly below the centre of the pupil, level with of lower border of the ala nasi.

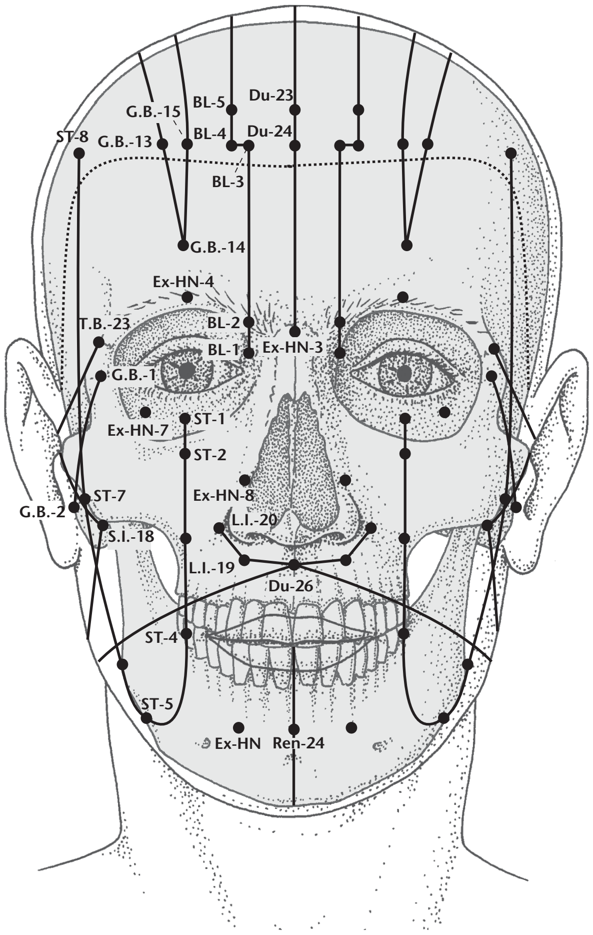
G.B.-15
BL-5
Du-23
ST-8
G.B.-13
BL-4
Du-24
BL-3
G.B.-14
Ex-HN-4
T.B.-23
BL-2
BL-1
Ex-HN-3
G.B.-1
ST-1
Ex-HN-7
ST-2
Ex-HN-8
ST-7
G.B.-2
L.I.-20
S.I.-18
L.I.-19
Du-26
ST-4
ST-5
Ex-HN
Ren-24

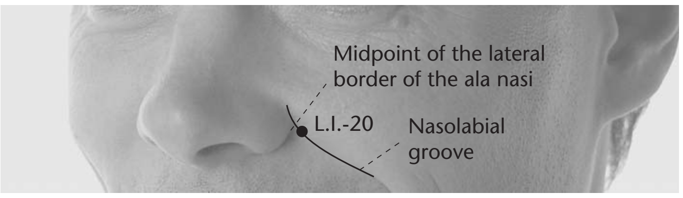
Midpoint of the lateral border of the ala nasi
L.I.-20
Nasolabial groove

L.I.-20 (*yingxiang*) 'Welcome Fragrance' In the nasolabial groove, at the level of the midpoint of the lateral border of the ala nasi.

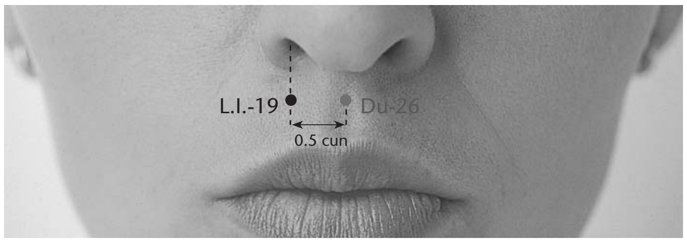
L.I.-19
Du-26
0.5 cun

L.I.-19 (*kouheliao*) 'Mouth Grain Crevice' Directly inferior to the lateral margin of the nostril, level with **Du-26**, which is located slightly superior to the midpoint of the philtrum (meeting point of the L.I. channel)

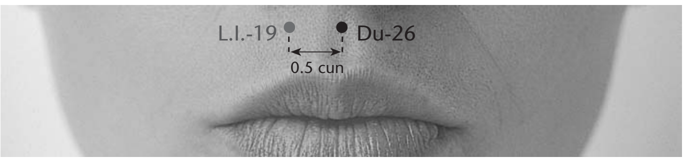
L.I.-19
Du-26
0.5 cun

Du-26 (*renzhong* or *shuigou*) 'Man's Middle' Below the nose, on the midline, at the junction of the upper and middle third of the philtrum

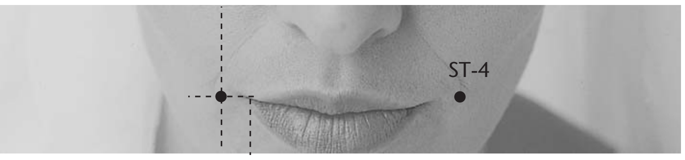

ST-4 (*dicang*) 'Earth Granary' When looking straight ahead on the pupil line, approximately 0.4 cun lateral to the corner of the mouth

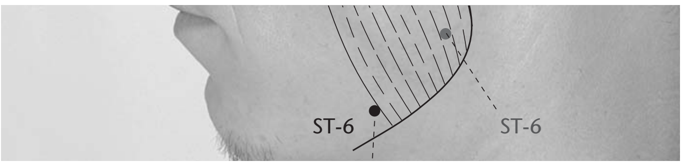

ST-5 (*daying*) 'Great Welcome' On the lateral mandible, anterior to the masseter (ask the patient to clench the teeth); a branch of the facial artery is palpable here

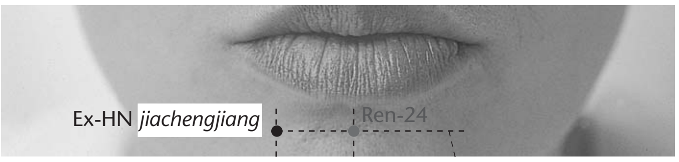
Ex-HN *jiachengjiang*
Ren-24

Ex-HN (*jiachengjiang*) 'Adjacent to Container of Fluids' At the same level and 1 cun lateral to **Ren-24** (chengjiang)

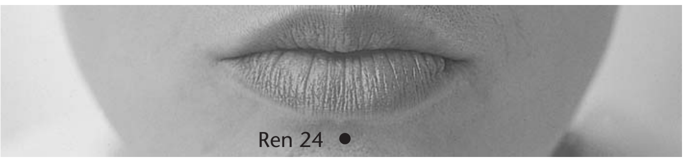

Ren-24 (*chengjiang*) 'Container of Fluids' On the midline, in the mentolabial groove below the lower lip

7.2 Head and Lateral Aspect of the Neck

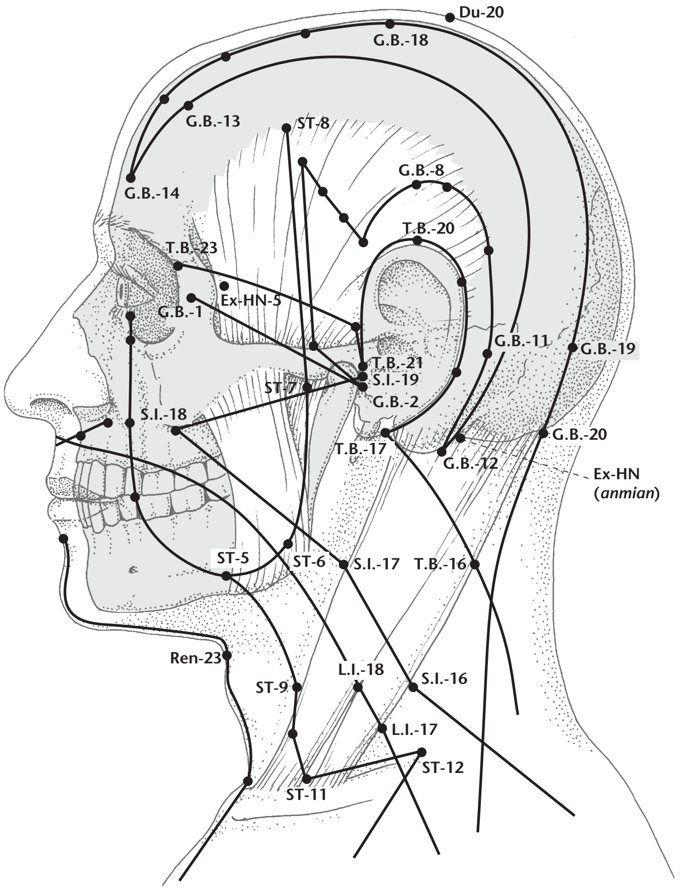
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(anmian)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

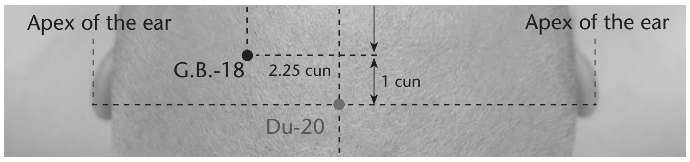
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*) 'Support Spirit' 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**, on the line connecting **G.B.-15** and **G.B.-20**, or 2.25 cun lateral to the midline (midpoint of the line connecting **ST-8** and **Du-24**)

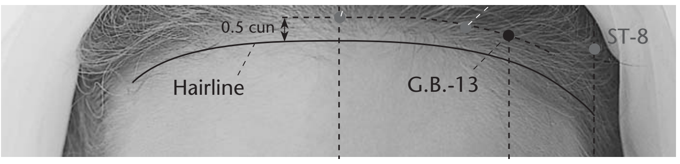
0.5 cun
ST-8
Hairline
G.B.-13

G.B.-13 (*benshen*) 'Root of the Spirit' 0.5 cun superior to the anterior hairline and 3 cun lateral to the midline, on the line between **Du-24** and **ST-8** (= 4.5 cun) at the junction of the lateral and medial third

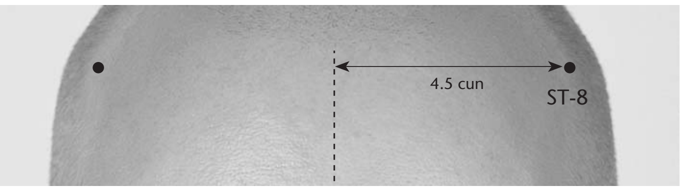

ST-8 (*touwei*) 'Head's Binding' At the temporal corner of the forehead, on the border of the temporalis muscle and 0.5 cun within the anterior hair line or 4.5 cun lateral to the anterior midline (**Du-24**)

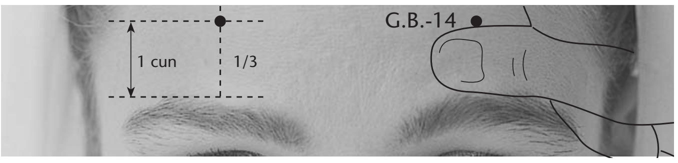
G.B.-14
1 cun
1/3

G.B.-14 (*yangbai*) 'Yang White' When looking straight ahead on the pupil line, approximately 1 cun superior to the midpoint of the eyebrow (centre of the eyebrow to the anterior hairline = 3 cun)

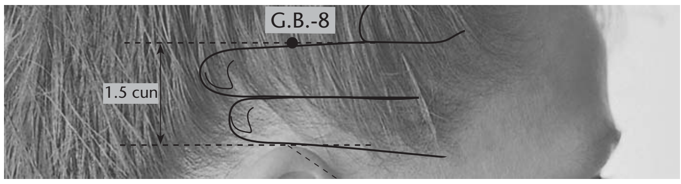

G.B.-8 (*shuaigu*) 'Leading Valley' 1.5 cun directly superior to the apex of the ear

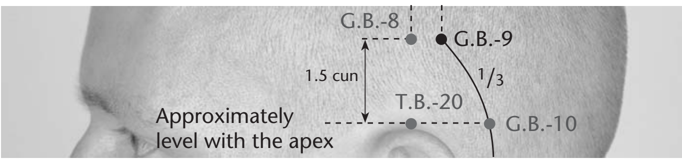
G.B.-8
G.B.-9
1.5 cun
1/3
T.B.-20
G.B.-10
Approximately
level with the apex

G.B.-9 (*tianchong*) 'Heavenly Rushing' 1.5 cun superior and 0.5 cun posterior to the apex of the ear or 0.5 cun posterior to **G.B.-8**

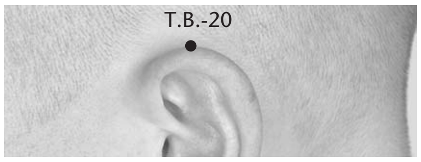

T.B.-20 (*jiaosun*) 'Minute Angle' Directly superior to the apex of the ear

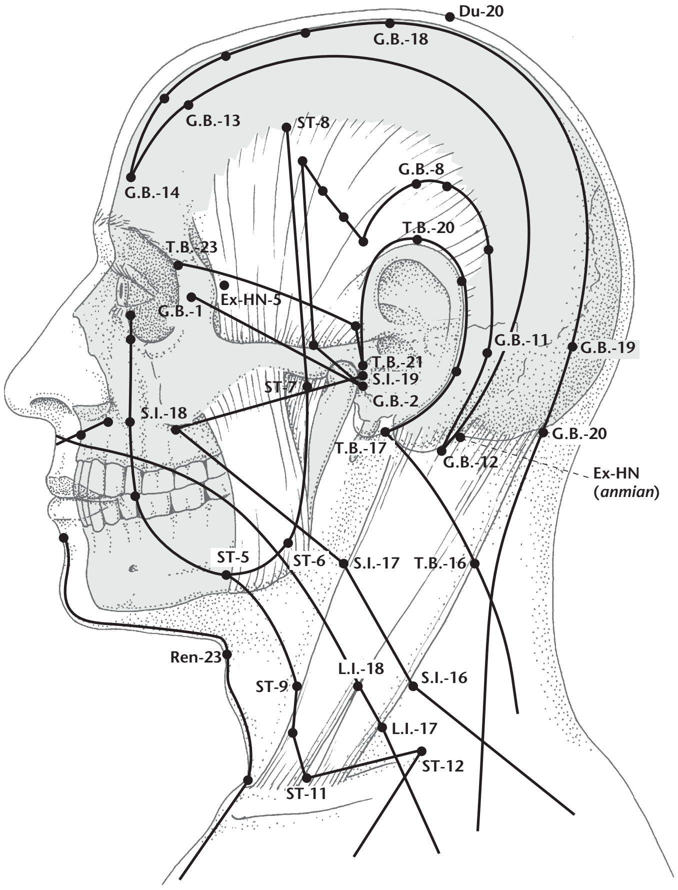
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
T.B.-17
G.B.-12
G.B.-20
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
L.I.-18
S.I.-16
ST-9
L.I.-17
ST-12
ST-11

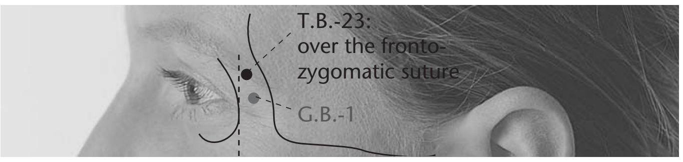
T.B.-23:
over the fronto-
zygomatic suture
G.B.-1

T.B.-23 (*sizhukong*) 'Silken Bamboo Hollow' In the depression (frontozygomatic suture) at the lateral end of the eyebrow

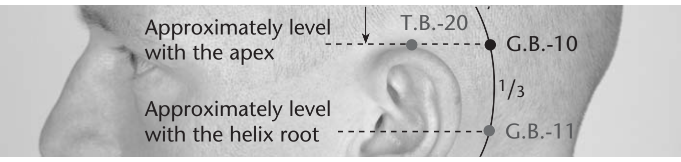
Approximately level
with the apex
T.B.-20
G.B.-10
1/3
Approximately level
with the helix root
G.B.-11

G.B.-10 (*fubai*) 'Floating White' At the junction of the upper third with the lower two-thirds of the curved line connecting **G.B.-9** and **G.B.-12**

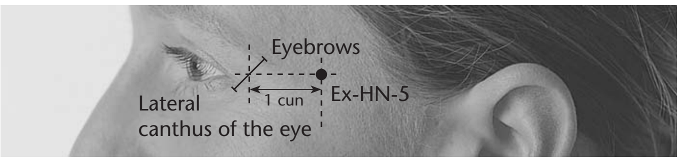
Eyebrows
Lateral
canthus of the eye
1 cun
Ex-HN-5

Ex-HN-5 (*taiyang*) 'Supreme Yang' In the depression approximately 1 cun posterior to the midpoint between the lateral end of the eyebrow and the outer canthus of the eye

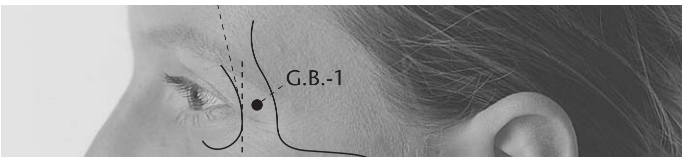

G.B.-1 (*tongziliao*) 'Pupil Crevice' 0.5 cun lateral to the outer canthus of the eye

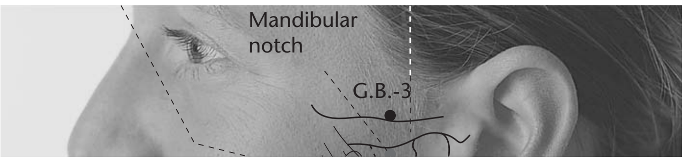
Mandibular
notch
G.B.-3

G.B.-3 (*shangguan*) 'Above the Joint' In a depression on the upper border of the zygomatic arch, approximately 1 cun anterior to the root of the ear, directly superior to **ST-7** (lower border of the zygomatic arch)

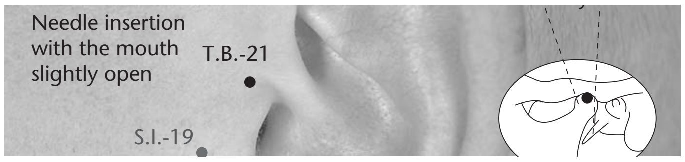
Needle insertion
with the mouth
slightly open
T.B.-21
S.I.-19

T.B.-21 (*ermen*) 'Ear Gate' Anterior to the ear, with the mouth open in the depression at the level of the supratragic notch, slightly superior to the condyloid process of the mandible

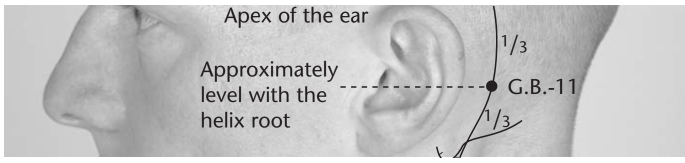
Apex of the ear
1/3
Approximately
level with the
helix root
G.B.-11
1/3

G.B.-11 (*touqiaoyin*) 'Yin Portals of the Head' At the junction of the lower third with the upper two-thirds of the curved line connecting **G.B.-9** and **G.B.-12**

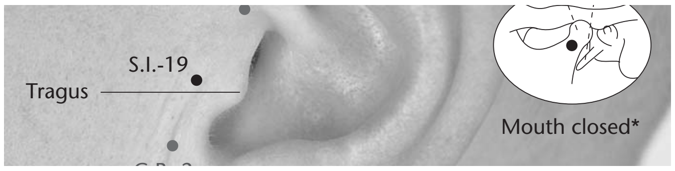
S.I.-19
Tragus
Mouth closed*

S.I.-19 (*tinggong*) 'Palace of Hearing' Anterior to the ear with the mouth open, at the level of the middle of the tragus, in a depression between the tragus and condyloid process of the mandible

---

* According to Deadman et al 1998.

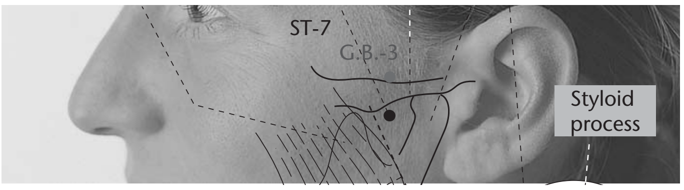
ST-7
G.B.-3
Styloid process

ST-7 (*xiaguan*) 'Below the Joint' With the mouth closed in the depression of the mandibular notch, between the coronoid and condyloid process of the mandible

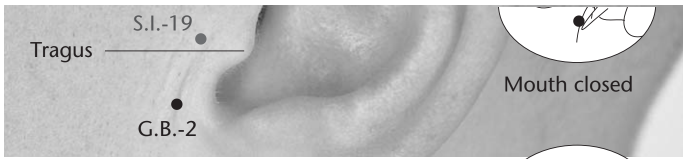
S.I.-19
Tragus
G.B.-2
Mouth closed

G.B.-2 (*tinghui*) 'Meeting of Hearing' Anterior to the ear with the mouth open, in the depression at the level of the intertragic notch, on the lower border of the condyloid process of the mandible

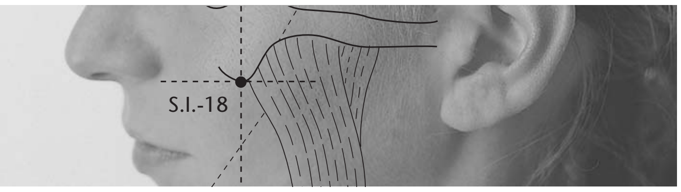

S.I.-18 (*quanliao*) 'Cheek Bone Crevice' In a depression at the intersection of a vertical line through the outer canthus of the eye with the lower border of the zygomatic bone

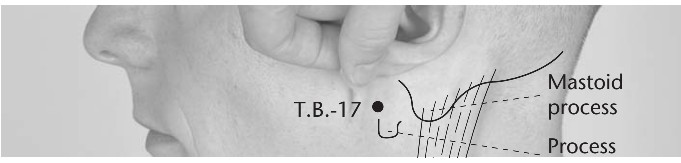
T.B.-17
Mastoid process
Process

T.B.-17 (*yifeng*) 'Wind Screen' With the mouth open in the depression behind the earlobe, between the mastoid process and the mandible

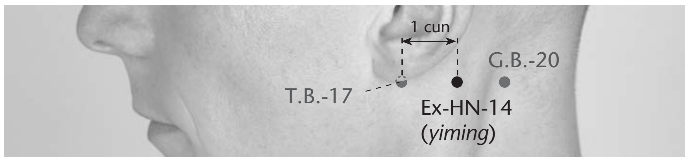
1 cun
G.B.-20
T.B.-17
Ex-HN-14
(*yiming*)

Ex-HN-14 (*yiming*) 'Eye Brightening' On the occiput, 1 cun posterior to **T.B.-17**, on a line connecting **T.B.-17** and **G.B.-20**

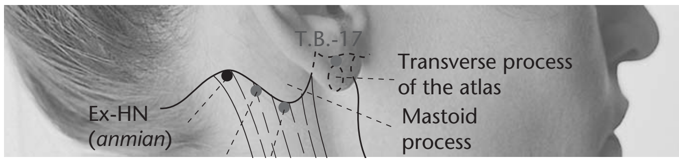
T.B.-17
Transverse process of the atlas
Ex-HN
(*anmian*)
Mastoid process

Ex-HN (*anmian*) 'Peaceful Sleep' Posterior to the ear, midway between **T.B.-17** and **G.B.-20**, posterior to the mastoid process

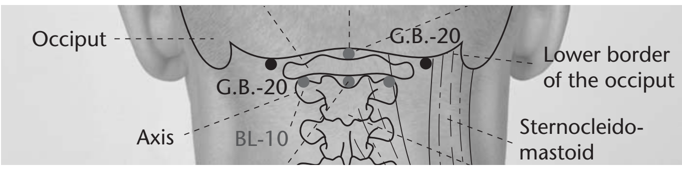
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*) 'Wind Pool' On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

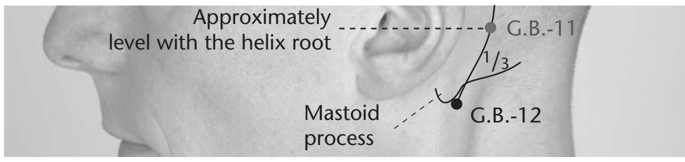
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

G.B.-12 (*wangu*) 'Mastoid Process' In a depression posterior and inferior to the mastoid process

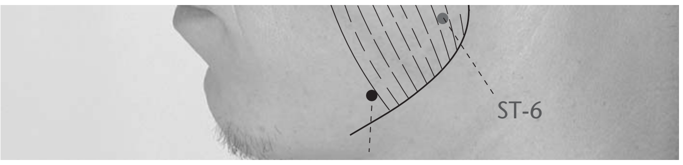

ST-6 (*jiache*) 'Jaw Bone' With the teeth clenched, on the highest prominence of the masseter, approximately one fingerbreadth anterior and superior to the mandibular angle

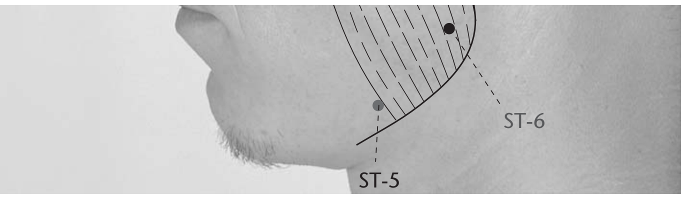

ST-5 (*daying*) 'Great Welcome' On the lateral mandible, anterior to the masseter (ask the patient to clench the teeth); a branch of the facial artery is palpable here

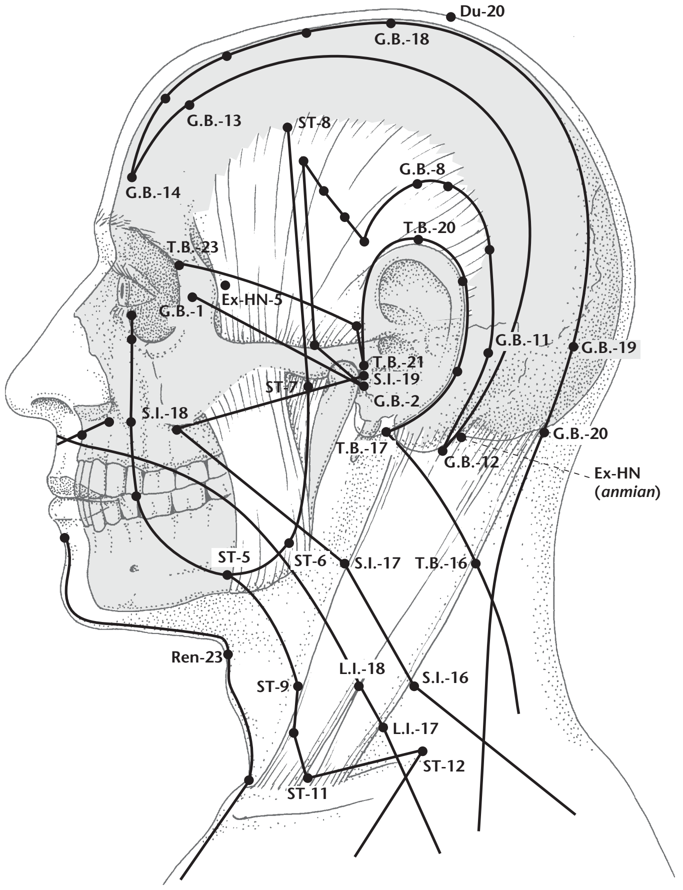
Du-20
G.B.-18
G.B.-13
ST-8
G.B.-14
G.B.-8
T.B.-20
T.B.-23
Ex-HN-5
G.B.-1
G.B.-11
G.B.-19
T.B.-21
S.I.-19
ST-7
G.B.-2
S.I.-18
T.B.-17
G.B.-20
G.B.-12
Ex-HN
(*anmian*)
ST-5
ST-6
S.I.-17
T.B.-16
Ren-23
ST-9
L.I.-18
S.I.-16
L.I.-17
ST-12
ST-11

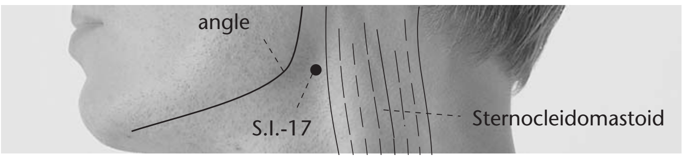
angle
S.I.-17
Sternocleidomastoid

S.I.-17 (*tianrong*) 'Heavenly Appearance' Level with and posterior to the mandibular angle, on the anterior border of the sternocleidomastoid

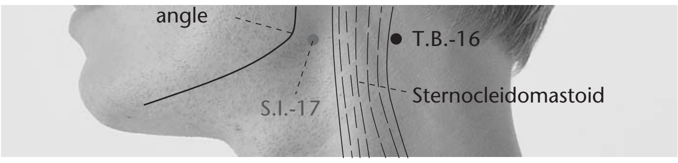
angle
T.B.-16
S.I.-17
Sternocleidomastoid

T.B.-16 (*tianyou*) 'Window of Heaven' posterior and inferior to the mastoid process, on the posterior border of the sternocleidomastoid, directly level with the mandibular angle

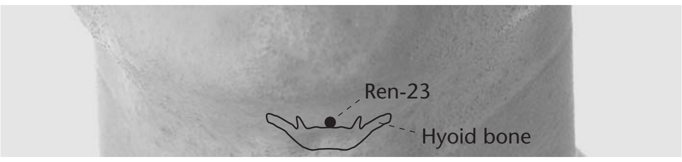
Ren-23
Hyoid bone

Ren-23 (*lianquan*) 'Corner Spring' On the anterior midline, superior to the upper border of the hyoid bone

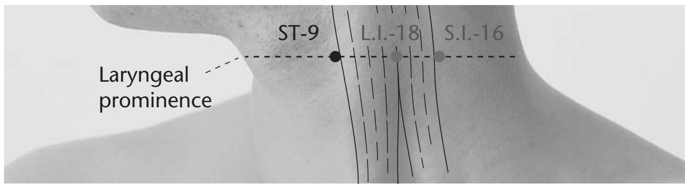
ST-9
L.I.-18
S.I.-16
Laryngeal prominence

ST-9 (*renying*) Man's Welcome' Approximately 1.5 cun lateral to the anterior midline, level with the laryngeal prominence, at the anterior border of the sternocleidomastoid

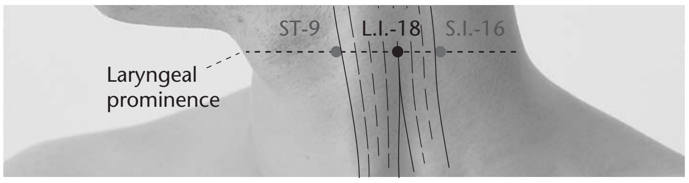
ST-9
L.I.-18
S.I.-16
Laryngeal prominence

L.I.-18 (*futu*) 'Support the Prominence' At the level of the laryngeal prominence, between the sternal and clavicular heads of the sternocleidomastoid muscle

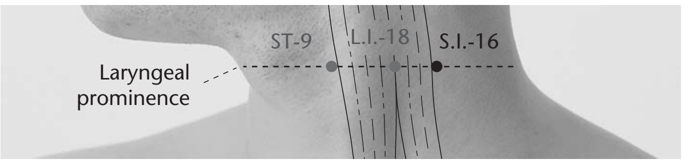
ST-9
L.I.-18
S.I.-16
Laryngeal prominence

S.I.-16 (*tianchuang*) 'Heavenly Window' Approximately 3.5 cun lateral to the anterior midline, level with the laryngeal prominence, on the posterior border of the sternocleidomastoid

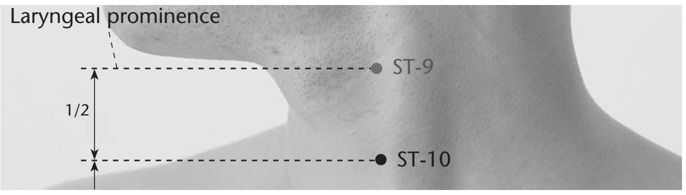
Laryngeal prominence
ST-9
1/2
ST-10

ST-10 (*shuitu*) 'Water Prominence' On the anterior border of the sternocleidomastoid, at the midpoint of a line connecting **ST-9** and **ST-11**

Ch07-F10028.qxd 2/22/08 10:08 PM Page 627

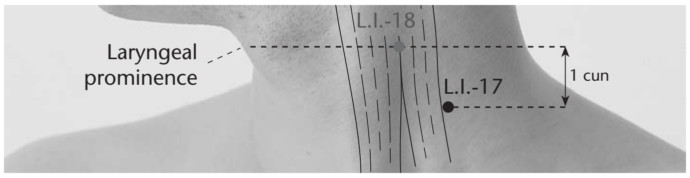
L.I.-18
Laryngeal prominence
L.I.-17
1 cun

L.I.-17 (*tianding*) 'Heaven's Tripod' 1 cun inferior to the laryngeal prominence, on the posterior border of the sternocleidomastoid

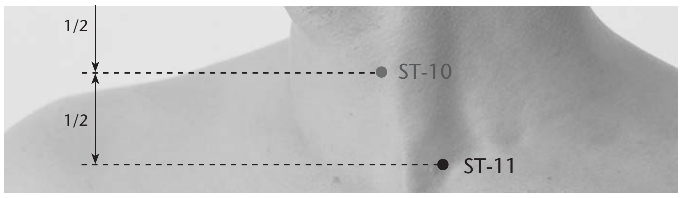
1/2
ST-10
1/2
ST-11

ST-11 (*qishe*) 'Abode of Qi' On the upper border of the clavicle, between the tendons of the sternal and clavicular heads of the sternocleidomastoid

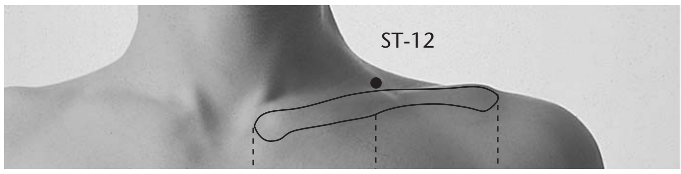

ST-12 (*quepen*) 'Empty Basin' In the supraclavicular fossa, superior to the midpoint of the clavicle, approximately 4 cun lateral to the midline

7.3 Occipital Region

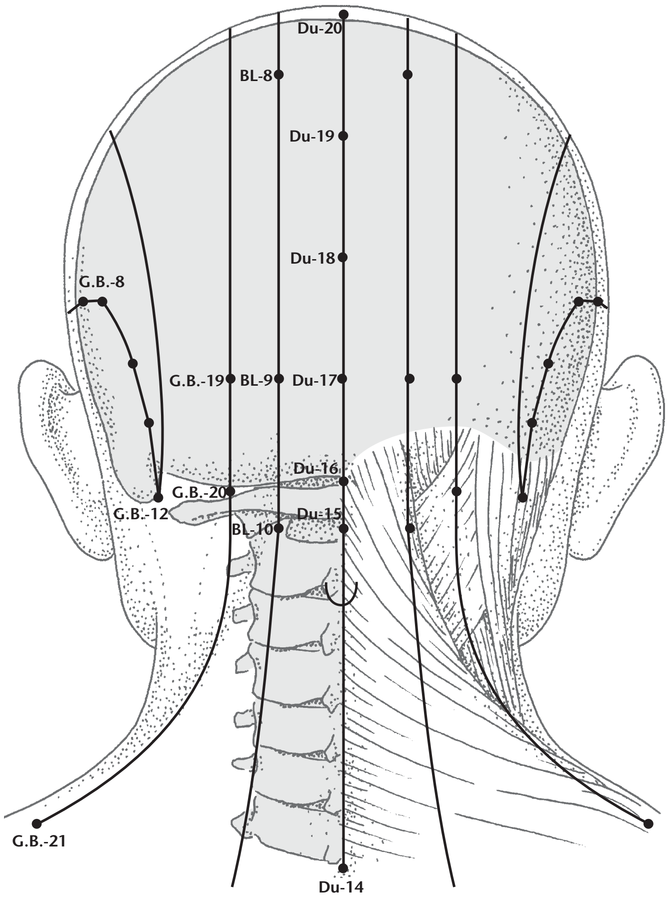
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
Du-15
BL-10
G.B.-21
Du-14

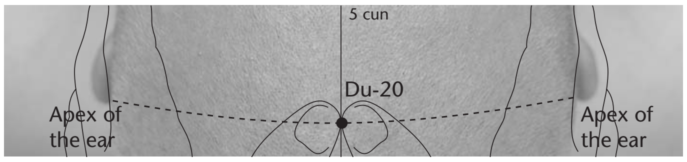
5 cun
Du-20
Apex of the ear
Apex of the ear

Du-20 (*baihui*) 'Hundred Meetings' At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively

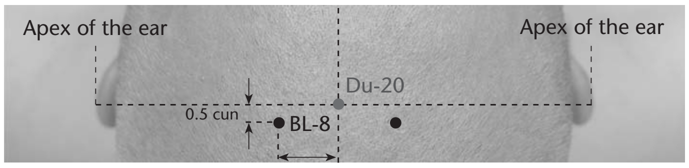
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8

BL-8 (*luoque*) 'Declining Connection' 1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline, or 0.5 cun posterior to **Du-20**

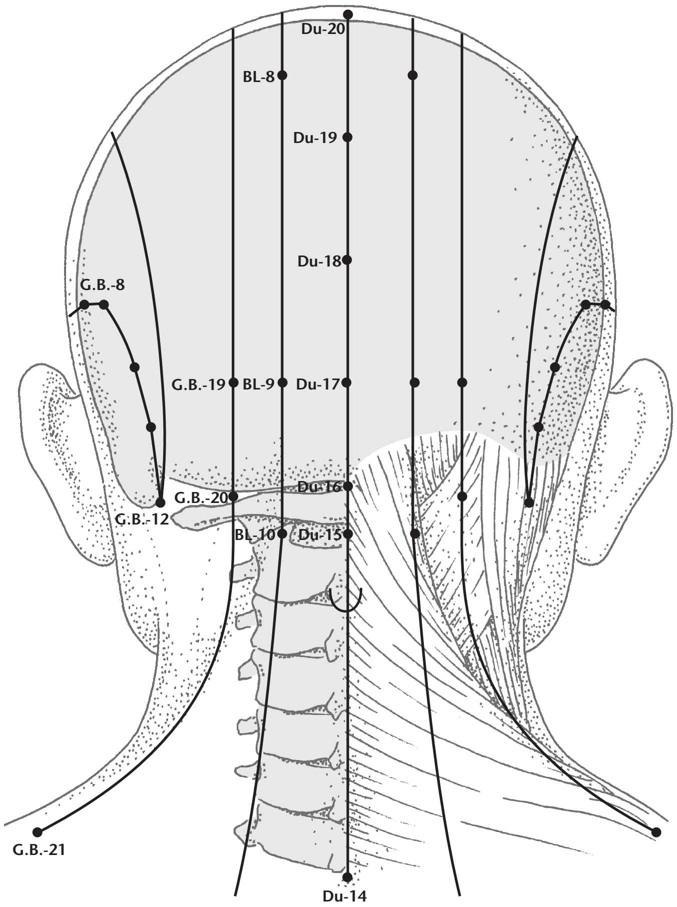
Du-20
BL-8
Du-19
Du-18
G.B.-8
G.B.-19
BL-9
Du-17
Du-16
G.B.-20
G.B.-12
BL-10
Du-15
G.B.-21
Du-14

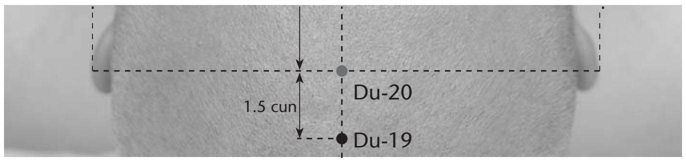
1.5 cun
Du-20
Du-19

Du-19 (*houding*) 'Behind the Crown' On the midline, 1.5 cun posterior to **Du-20** or 3 cun superior to **Du-17**

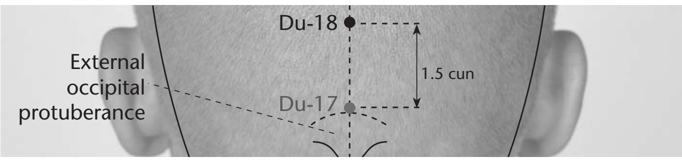
Du-18
External occipital protuberance
1.5 cun
Du-17

Du-18 (*qiangjian*) 'Unyielding Space' On the posterior midline, 1.5 cun superior to **Du-17** or 3 cun posterior to **Du-20**

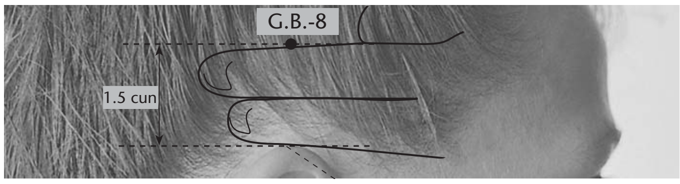

G.B.-8 (*shuaigu*) 'Leading Valley' 1.5 cun directly superior to the apex of the ear

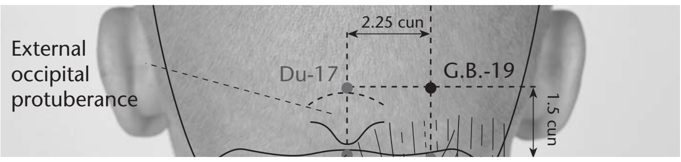
External occipital protuberance
2.25 cun
Du-17
G.B.-19
1.5 cun

G.B.-19 (*naokong*) 'Brain Hollow' Level with the upper border of the external occipital protuberance (**Du-17**), on the line connecting **G.B.-15** and **G.B.-20**, and 2.25 cun lateral to the midline

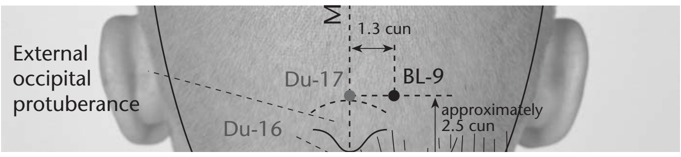
External occipital protuberance
1.3 cun
Du-17
BL-9
Du-16
approximately 2.5 cun

BL-9 (*yuzhen*) 'Jade Pillow' 2.5 cun superior to the posterior hairline and 1.3 cun lateral to the midline or **Du-17** (directly superior to the external occipital protuberance)

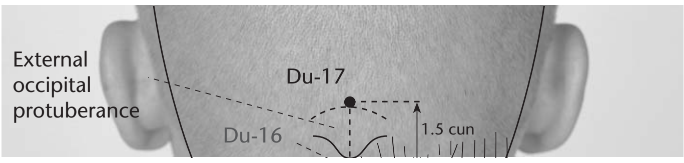
External occipital protuberance
Du-17
Du-16
1.5 cun

Du-17 (*naohu*) 'Brain's Door' On the midline, in a depression superior to the external occipital protuberance, approximately 2.5 cun superior to the posterior hairline, 1.5 cun superior to **Du-16**

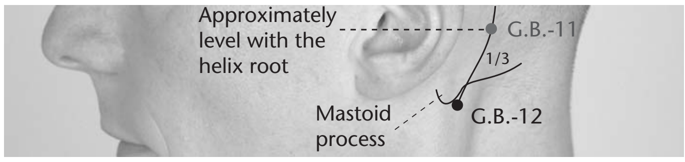
Approximately level with the helix root
G.B.-11
1/3
Mastoid process
G.B.-12

G.B.-12 (*wangu*) 'Mastoid Process' In a depression posterior and inferior to the mastoid process

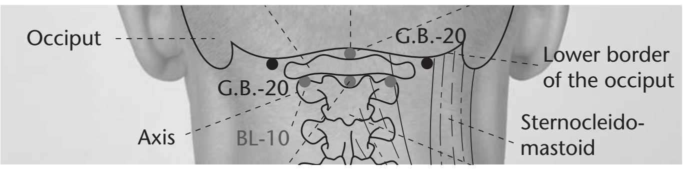
Occiput
G.B.-20
Lower border of the occiput
G.B.-20
Axis
BL-10
Sternocleido-mastoid

G.B.-20 (*fengchi*) 'Wind Pool' On the lower border of the occipital bone, in the depression between the origins of the sternocleidomastoid and trapezius

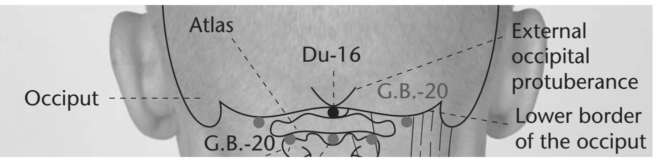
Atlas
Du-16
External occipital protuberance
Occiput
G.B.-20
Lower border of the occipit
G.B.-20

Du-16 (*fengfu*) 'Palace of Wind' On the posterior midline, directly below the external occipital protuberance, in the depression between the origins of the trapezius

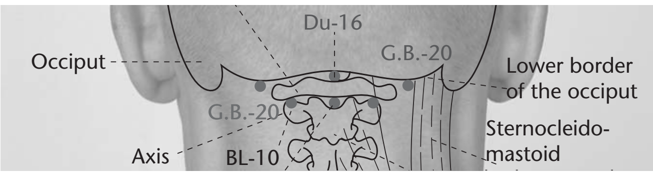
Du-16
Occiput
G.B.-20
Lower border of the occipit
G.B.-20
Sternocleido-mastoid
Axis
BL-10

BL-10 (*tianzhu*) 'Celestial Pillar' Below the lower border of the occiput, approximately 1.3 cun lateral to the midline (**Du-15**), at the insertion of the trapezius

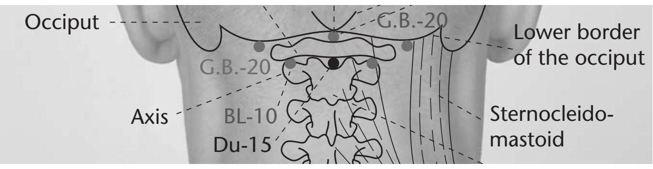
Occiput
G.B.-20
Lower border of the occipit
G.B.-20
Axis
BL-10
Sternocleido-mastoid
Du-15

Du-15 (*yamen*) 'Gate of Muteness' On the posterior midline, in the depression between the 1st (atlas) and 2nd cervical vertebrae (axis = first palpable spinous process), approximately 0.5 cun below **Du-16**

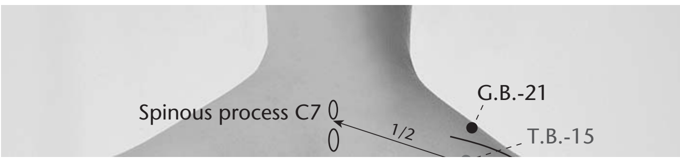
G.B.-21
Spinous process C7
1/2
T.B.-15

G.B.-21 (*jianjing*) 'Shoulder Well' At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

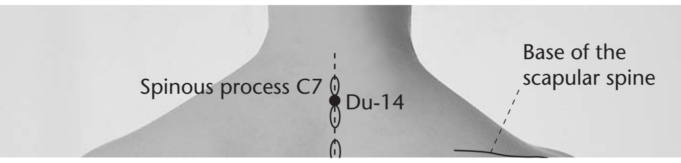
Base of the scapular spine
Spinous process C7
Du-14

Du-14 (*dazhui*) 'Great Vertebra' Below the spinous process of C7

7.4 Vertex

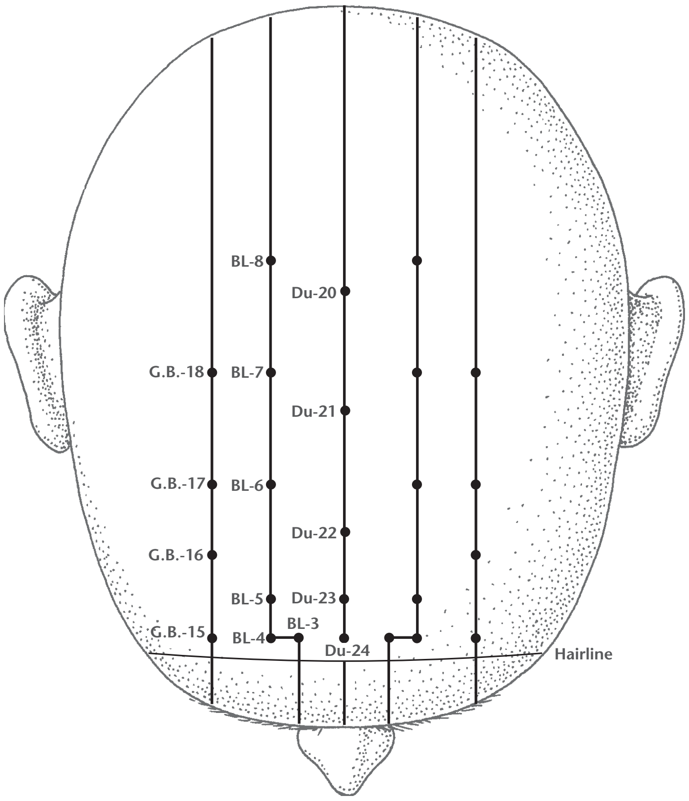
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

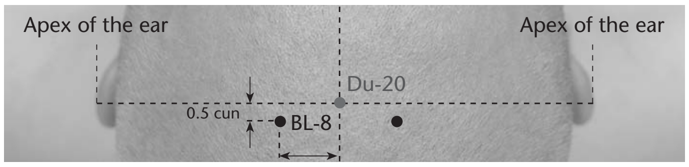
Apex of the ear
Apex of the ear
Du-20
0.5 cun
BL-8

BL-8 (*luoque*) 'Declining Connection' 1.5 cun lateral to the midline and 5.5 cun superior to the anterior hairline, or 0.5 cun posterior to **Du-20**

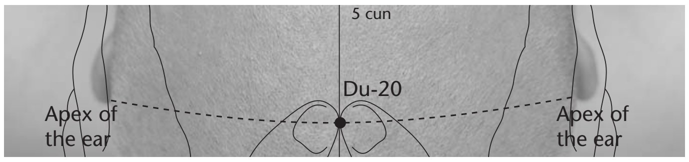
5 cun
Du-20
Apex of the ear
Apex of the ear

Du-20 (*baihui*) 'Hundred Meetings' At the junction of a line connecting the apices of the ears and the midline, 5 cun from the anterior or 7 cun from the posterior hairline respectively

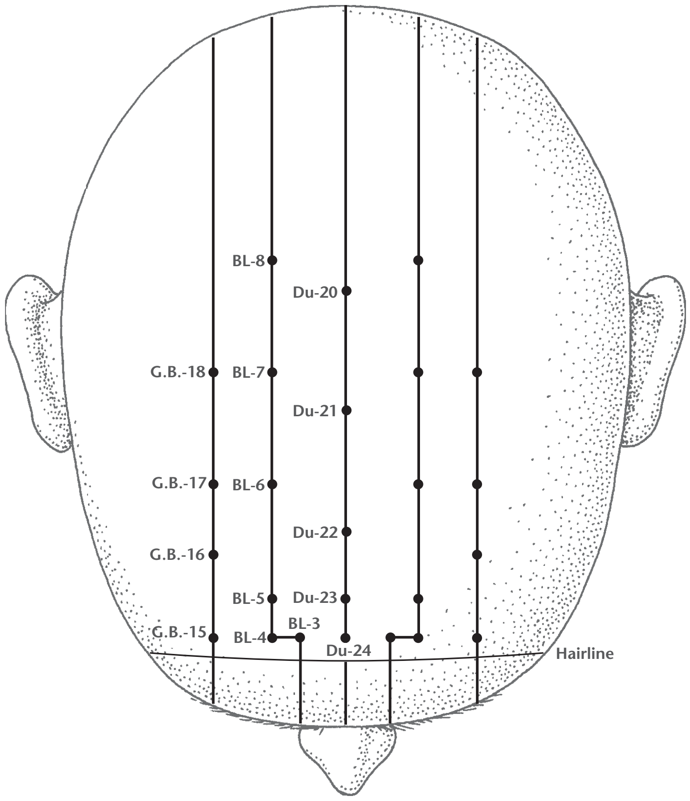
BL-8
Du-20
G.B.-18
BL-7
Du-21
G.B.-17
BL-6
Du-22
G.B.-16
BL-5
Du-23
BL-3
G.B.-15
BL-4
Du-24
Hairline

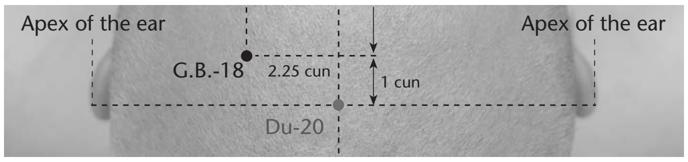
Apex of the ear
Apex of the ear
G.B.-18
2.25 cun
1 cun
Du-20

G.B.-18 (*chengling*) 'Support Spirit' 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**, on the line connecting **G.B.-15** with **G.B.-20**, or 2.25 cun lateral to the midline (midpoint of the line connecting **ST-8** and **Du-24**)

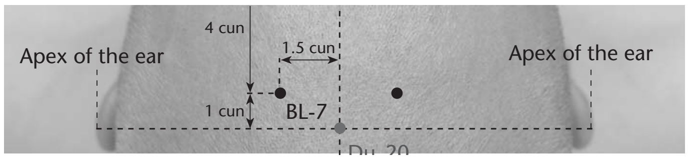
Apex of the ear
4 cun
1.5 cun
Apex of the ear
1 cun
BL-7
Du-20

BL-7 (*tongtian*) 'Heavenly Connection' 1.5 cun lateral to the midline and 4 cun superior to the anterior hairline or 1 cun anterior to **Du-20**

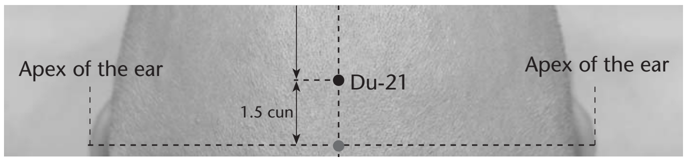
Apex of the ear
Du-21
Apex of the ear
1.5 cun

Du-21 (*qianding*) 'In Front of the Crown' On the midline, 1.5 cun anterior to **Du-20** or 3.5 cun above the anterior hairline

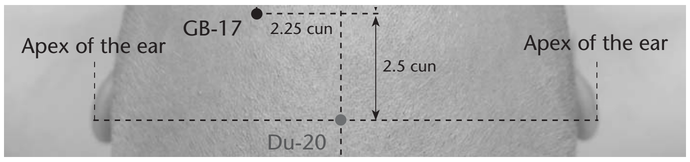
GB-17
2.25 cun
Apex of the ear
2.5 cun
Apex of the ear
Du-20

G.B.-17 (*zhengying*) 'Upright Nutrition' 2.5 cun superior to the anterior hairline, on the pupil line, on the line connecting **G.B.-15** and **G.B.-20** (= 2.25 cun lateral to the midline or midway between **Du-23** and **ST-8**)

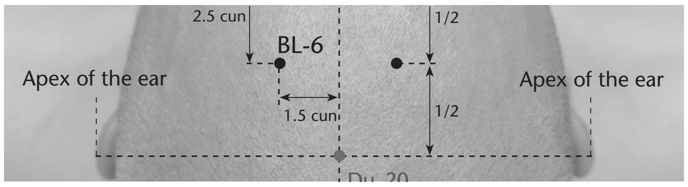
2.5 cun
1/2
BL-6
Apex of the ear
Apex of the ear
1.5 cun
1/2
Du-20

BL-6 (*chengguang*) 'Receiving Light' 2.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one-third of the distance between **Du-24** and **ST-8**

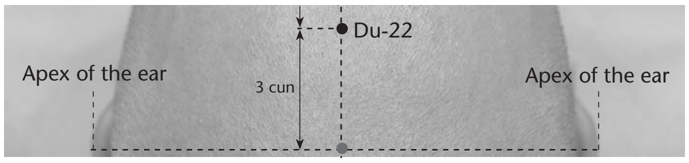
Du-22
Apex of the ear
3 cun
Apex of the ear

Du-22 (*xinhui*) 'Fontanelle Meeting' On the midline, 2 cun superior to the anterior hairline or 3 cun anterior to **Du-20**

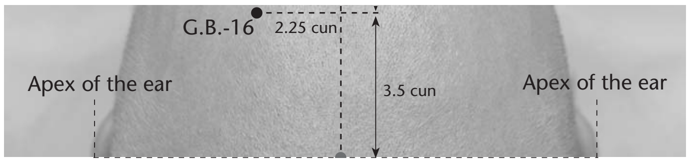
G.B.-16
2.25 cun
Apex of the ear
3.5 cun
Apex of the ear

G.B.-16 (*muchuang*) 'Window of the Eye' 1.5 cun superior to the anterior hairline, on the pupil line, on the line connecting **G.B.-15** and **G.B.-20** (= 2.25 cun lateral to the midline or midway between **Du-24** and **ST-8**)

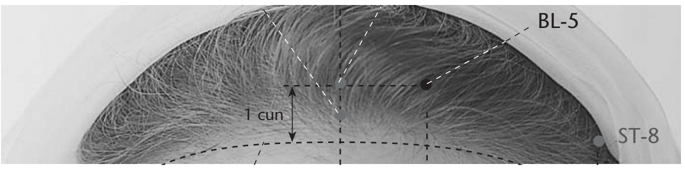
BL-5
1 cun
ST-8

BL-5 (*wuchu*) 'Fifth Place' 1 cun superior to the anterior hairline and 1.5 cun lateral to the midline (directly above **BL-4**, level with **Du-23**)

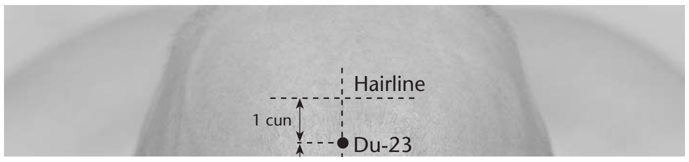
Hairline
1 cun
Du-23

Du-23 (*shangxing*) 'Upper Star' On the midline, 1 cun superior to the anterior hairline

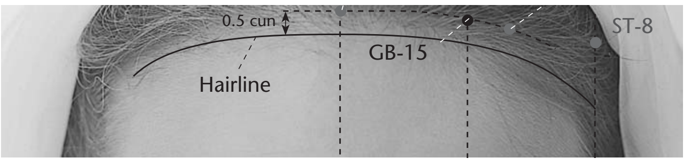
0.5 cun
ST-8
GB-15
Hairline

G.B.-15 (*toulinqi*) 'Head Governor of Tears' On the pupil line when looking straight ahead, 0.5 cun superior to the anterior hairline, at the midpoint of the line between **Du-24** and **ST-8** (= 4.5 cun)

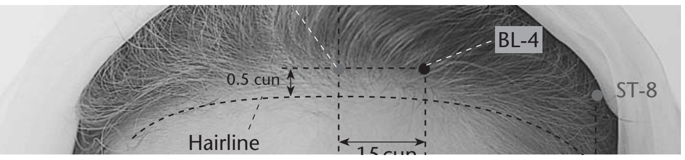
BL-4
0.5 cun
ST-8
Hairline
1.5 cun

BL-4 (*qucha*) 'Crooked Curve' 0.5 cun superior to the anterior hairline and 1.5 cun lateral to the midline, or one-third of the distance between **Du-24** and **ST-8** (= 4.5 cun)

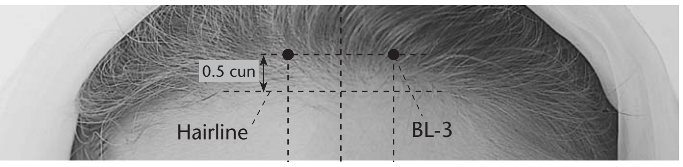
0.5 cun
Hairline
BL-3

BL-3 (*meichong*) 'Eyebrows' Pouring' 0.5 cun within the anterior hairline, superior to the medial canthus of the eye (**BL-1**)

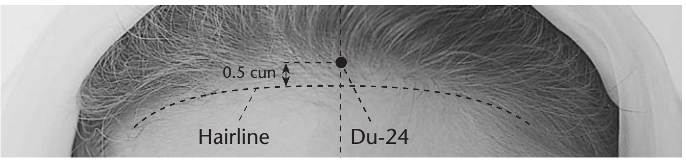
0.5 cun
Hairline
Du-24

Du-24 (*shenting*) 'Courtyard of the Spirit' On the midline, 0.5 cun superior to the anterior hairline

7.5 Back

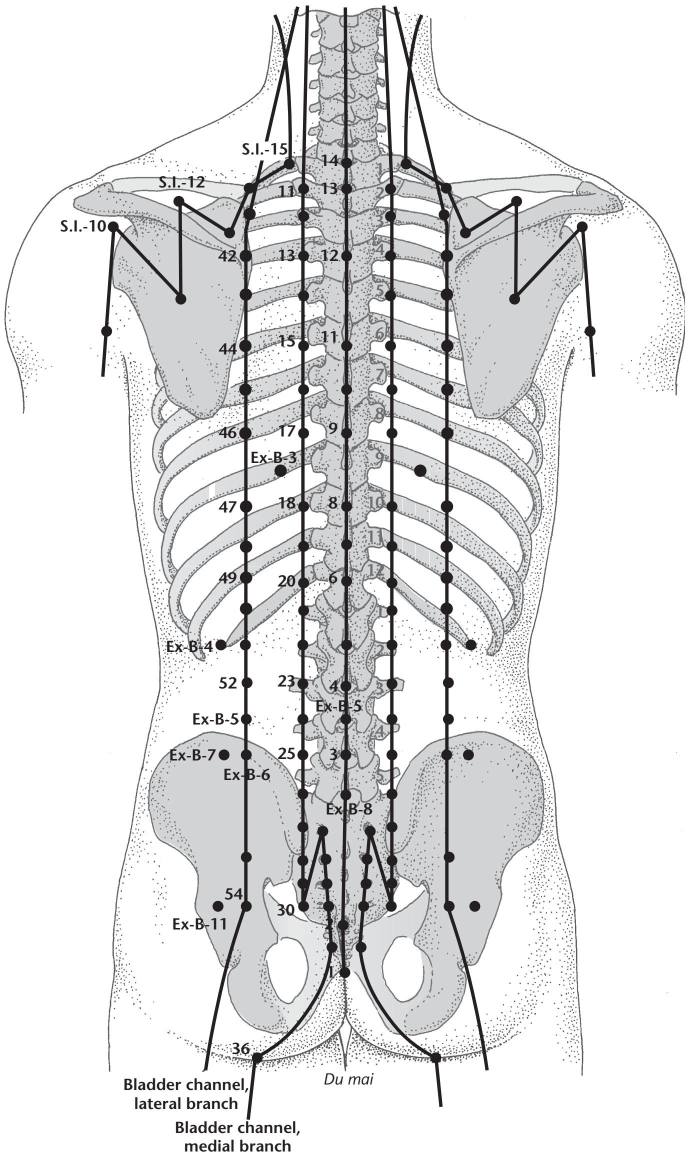
S.I.-15
14
S.I.-12
11
13
S.I.-10
42
13
12
44
15
11
46
17
9
Ex-B-3
47
18
8
49
20
6
Ex-B-4
52
23
4
Ex-B-5
Ex-B-5
Ex-B-7
25
3
Ex-B-6
Ex-B-8
54
30
Ex-B-11
36
Du mai
Bladder channel, lateral branch
Bladder channel, medial branch

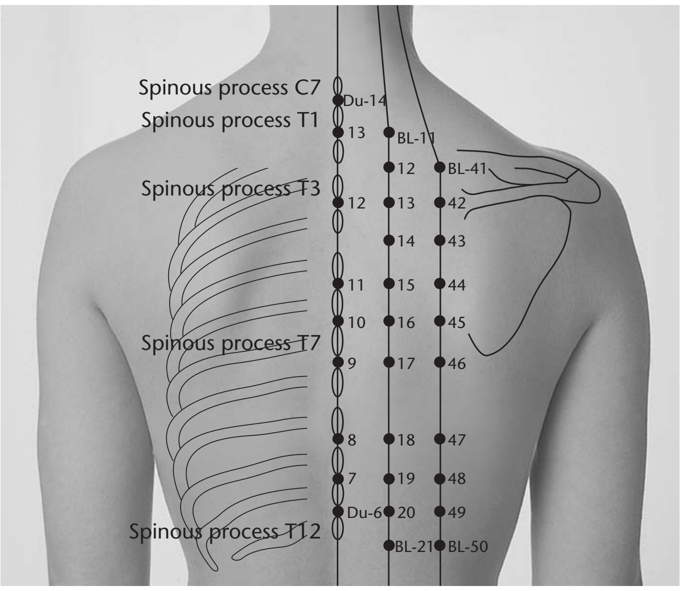
Spinous process C7
Spinous process T1
Spinous process T3
Spinous process T7
Spinous process T12
Du-14
13
BL-11
12
BL-41
12
13
42
14
43
11
15
44
10
16
45
9
17
46
8
18
47
7
19
48
Du-6
20
49
BL-21
BL-50

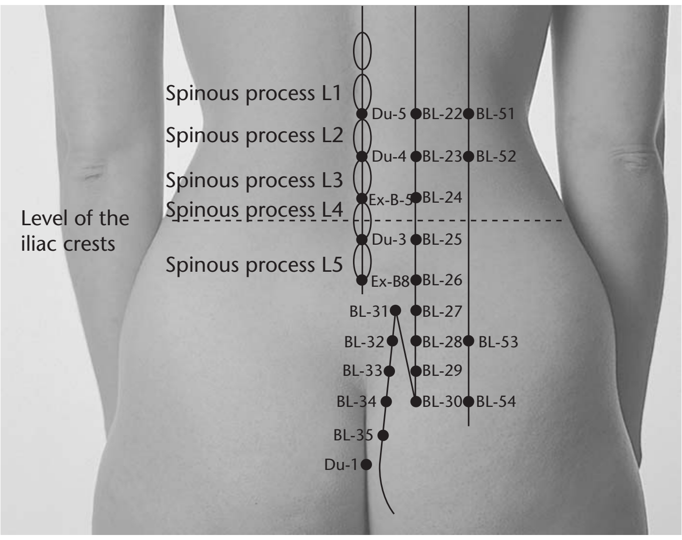
Spinous process L1
Spinous process L2
Spinous process L3
Spinous process L4
Spinous process L5
Level of the iliac crests
Du-5
BL-22
BL-51
Du-4
BL-23
BL-52
Ex-B-5
BL-24
Du-3
BL-25
Ex-B8
BL-26
BL-31
BL-27
BL-32
BL-28
BL-53
BL-33
BL-29
BL-34
BL-30
BL-54
BL-35
Du-1

Vertebra	**Below the spinous process or in the sacral foramen**	**0.5 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**	**1.5 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**	**3 cun lateral to the midline, level with the lower border of the spinous process or sacral foramen**
C7	Du-14	Ex-B-1 *(dingchuan)*	**2 cun lateral** S.I.-15	**3.5 cun lateral** Ex-B (*jiehexue*)
T1	Du-13	A point of Ex-B-2 (*huatuojiaji*)	BL-11	S.I.-14
T2		A point of Ex-B-2 (*huatuojiaji*)	BL-12	BL-41
T3	Du-12	A point of Ex-B-2 (*huatuojiaji*)	BL-13	BL-42
T4		A point of Ex-B-2 (*huatuojiaji*)	BL-14	BL-43
T5	Du-11	A point of Ex-B-2 (*huatuojiaji*)	BL-15	BL-44
T6	Du-10	A point of Ex-B-2 (*huatuojiaji*)	BL-16	BL-45
T7	Du-9	A point of Ex-B-2 (*huatuojiaji*)	BL-17	BL-46
T8		A point of Ex-B-2 (*huatuojiaji*)	Ex-B-3 (*weiwanxiashu*)	
T9	Du-8	A point of Ex-B-2 (*huatuojiaji*)	BL-18	BL-47
T10	Du-7	A point of Ex-B-2 (*huatuojiaji*)	BL-19	BL-48
T11	Du-6	A point of Ex-B-2 (*huatuojiaji*)	BL-20	BL-49
T12		A point of Ex-B-2 (*huatuojiaji*)	BL-21	BL-50
L1	Du-5	A point of Ex-B-2 (*huatuojiaji*)	BL-22	BL-51/ Ex-B-4 (**3.5 cun lateral**)
L2	Du-4	A point of Ex-B-2 (*huatuojiaji*)	BL-23	BL-52
L3	Ex-B-5 (*xiajishu*)	A point of Ex-B-2 (*huatuojiaji*)	BL-24	Ex-B-5 (according to WHO a set of 3 points)
L4	Du-3	A point of Ex-B-2 (*huatuojiaji*)	BL-25	Ex-B-6/Ex-B-7 (**3.5 cun lateral**)
L5	Ex-B-8	A point of Ex-B-2 (*huatuojiaji*)	BL-26	
1st sacral foramen	BL-31		BL 27	
2nd sacral foramen	BL-32		BL 28	BL-53
3rd sacral foramen	BL-33		BL 29	
4th sacral foramen	BL-34		BL 30	BL-54/Ex-B-11 (3.5 cun lateral)
Sacral hiatus	Du- 2			
		BL-35 (0.5 cun lateral and inferior to Du-2)		
Between the coccyx and the anus	Du-1			

Ch07-F10028.qxd 2/22/08 10:08 PM Page 634

7.6 Frontal and Lateral Thorax and Abdomen

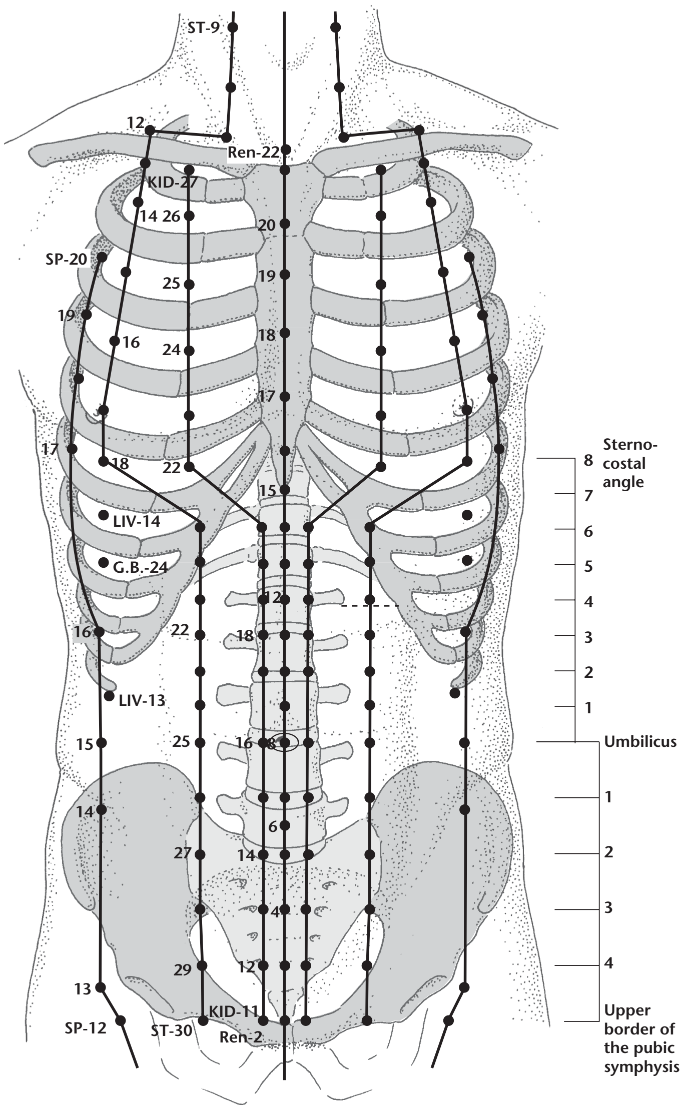
ST-9
12
Ren-22
KID-27
14 26
20
SP-20
25
19
19
16
24
18
17
17
18
22
15
LIV-14
G.B.-24
12
16
22
18
LIV-13
15
25
16
8
14
27
6
14
4
29
12
13
SP-12
ST-30
KID-11
Ren-2
8
7
6
5
4
3
2
1
Sterno-costal angle
Umbilicus
1
2
3
4
Upper border of the pubic symphysis

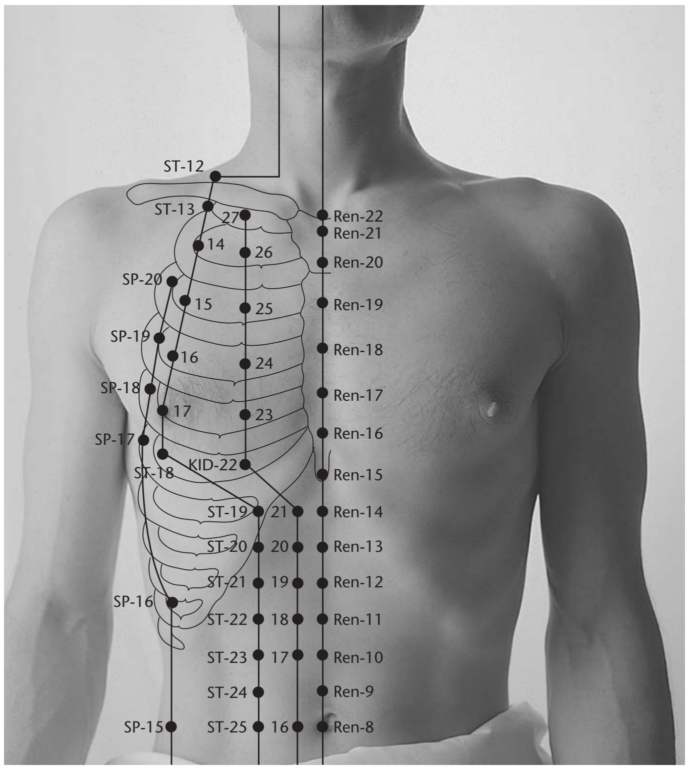
ST-12
ST-13
27
Ren-22
Ren-21
14
26
Ren-20
SP-20
15
25
Ren-19
SP-19
16
24
Ren-18
SP-18
Ren-17
17
23
Ren-16
SP-17
ST-18
KID-22
Ren-15
ST-19
21
Ren-14
ST-20
20
Ren-13
ST-21
19
Ren-12
SP-16
ST-22
18
Ren-11
ST-23
17
Ren-10
ST-24
Ren-9
SP-15
ST-25
16
Ren-8

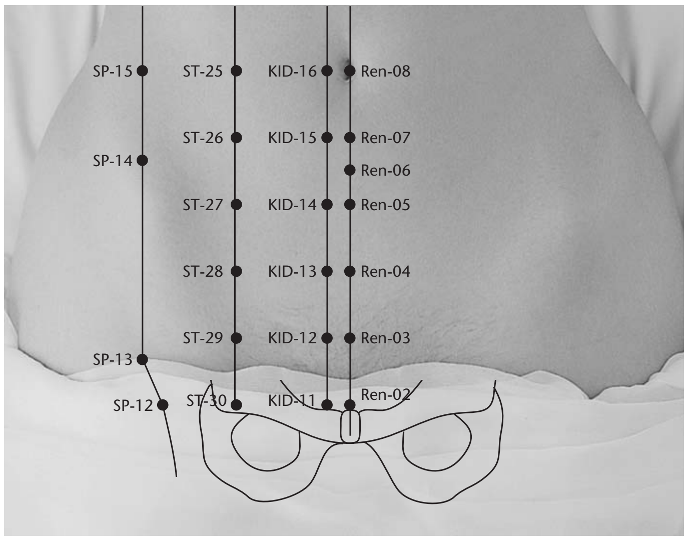
SP-15
ST-25
KID-16
Ren-08
ST-26
KID-15
Ren-07
SP-14
Ren-06
ST-27
KID-14
Ren-05
ST-28
KID-13
Ren-04
ST-29
KID-12
Ren-03
SP-13
SP-12
ST-30
KID-11
Ren-02

Anatomical level	***Ren Mai* (midline)**	**KID channel (2 cun lateral to the midline)**	**ST channel (4 cun lateral to the midline)**	**Other (6 cun lateral to the midline)**
Below the clavicle	≈ Ren-21	KID-27	ST-13	LU-2
1st ICS	Ren-20	KID-26	ST-14	LU-1
2nd ICS	Ren-19	KID-25	ST-15	SP-20
3rd ICS	Ren-18	KID-24	ST-16	SP-19
4th ICS	Ren-17	KID-23	ST-17 (nipple)	SP-18/P-1 (1 cun lateral to the nipple)/G.B.-22 (3 cun inferior to the apex of the axilla)/G.B.-23 (1 cun anterior to G.B.-22)
5th ICS	Ren 16	KID-22	ST-18	SP-17
Anatomical level	***Ren mai* (midline)**	**KID-channel (0.5 cun lateral to the midline)**	**ST channel (2 cun lateral to the midline)**	**Other (4 cun lateral to the midline)**
6 cun superior/6th ICS	Ren-14	KID-21	ST-19	LIV-14 (6th ICS)
5 cun superior/7th ICS	Ren-13	KID-20	ST-20	G.B.-24 (7th ICS)
4 cun superior	Ren-12	KID-19	ST-21	
3 cun superior	Ren-11	KID-18	ST-22	SP-16
2 cun superior	Ren-10	KID-17	ST-23	Ex-CA (*weishang*)
1 cun superior	Ren-9		ST-24	≈ LIV-13 (free end of the 11th rib)
Level of umbilicus	Ren-8	KID-16	ST-25	SP-15, G.B.-26 (inferior to the free end of the 12th rib, level with the umbilicus)
1 cun inferior	Ren-7	KID-15	ST-26	
(1.5 cun inferior)	Ren-6			
2 cun inferior	Ren-5	KID-14	ST-27	
(2.5 cun inferior)	Ex-CA (*zhishi*)			
3 cun inferior	Ren-4	KID-13	ST-28	Ex-CA (*tituo*)/Ex-CA (*qimen*)/ Ex-CA (*yijing*)/(4/3/1 cun lateral to the midline)/≈ G.B.-27 (anterior and medical to the ASIS → 3)
4 cun inferior	Ren-3	KID-12	ST-29	Ex-CA (*zigong*), 3 cun lateral to the midline
5 cun inferior **symphysis**	Ren-2	KID-11	ST-30	SP-12 (3.5 cun lateral to the midline)

Note: The intercostal spaces and the ribs curve upwards laterally!

7.7 Posterior Shoulder and Arm

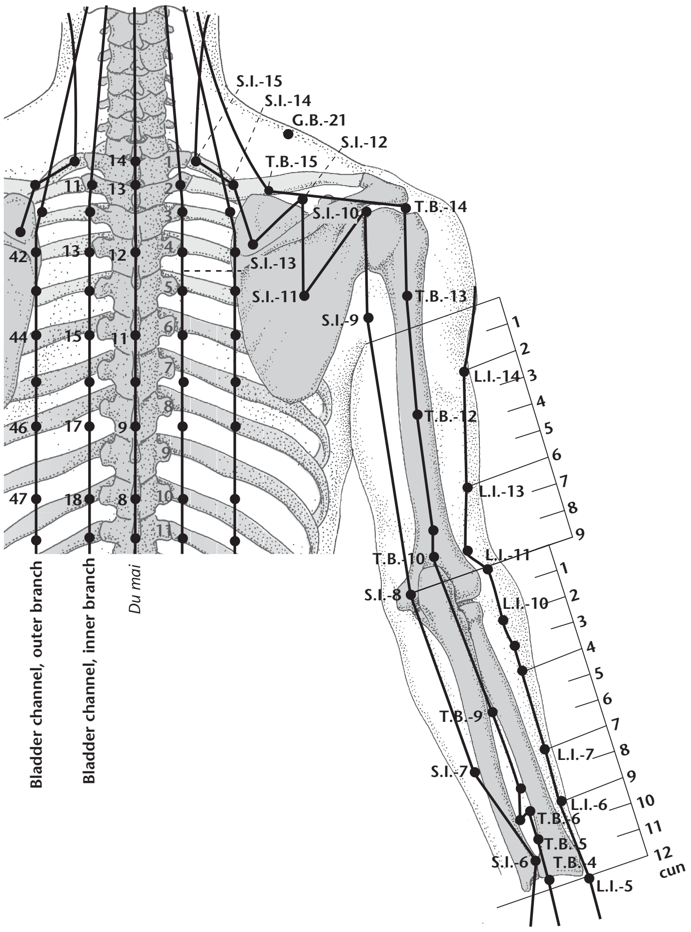
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
14
1
11
13
2
T.B.-14
S.I.-10
3
42
13
12
4
S.I.-13
S.I.-11
T.B.-13
5
S.I.-9
44
15
11
6
7
46
17
9
8
47
18
8
9
10
11
L.I.-14
T.B.-12
L.I.-13
1
2
3
4
5
6
7
8
9
T.B.-10
L.I.-11
S.I.-8
L.I.-10
1
2
3
4
5
6
7
8
9
10
11
12
cun
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
S.I.-6
T.B.-4
L.I.-5
Bladder channel, outer branch
Bladder channel, inner branch
Du mai

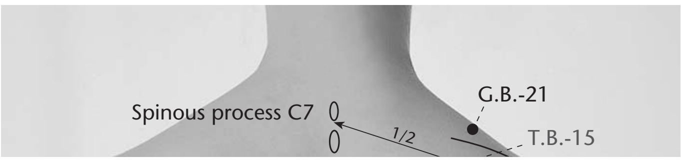
Spinous process C7
G.B.-21
1/2
T.B.-15

G.B.-21 (*jianjing*) 'Shoulder Well' At the highest point of the shoulder, midway between the lower border of the spinous process of C7 and the lateral extremity of the acromion

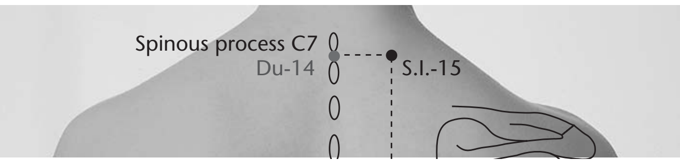
Spinous process C7
Du-14
S.I.-15

S.I.-15 (*jianzhongshu*) 'Middle Shoulder Shu' 2 cun lateral to the lower border of the spinous process of the T7 (**Du-14**)

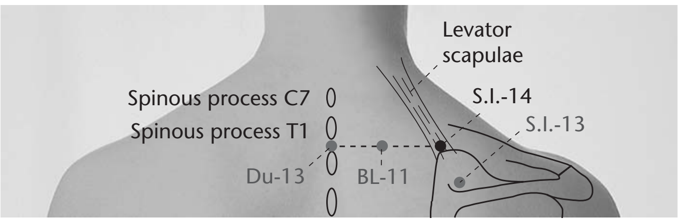
Levator
scapulae
Spinous process C7
Spinous process T1
S.I.-14
S.I.-13
Du-13
BL-11

S.I.-14 (*jianwaishu*) 'Outer Shoulder Shu' 3 cun lateral to the lower border of the spinous process of the T1 (**Du-13**), at the insertion of the levator scapulae

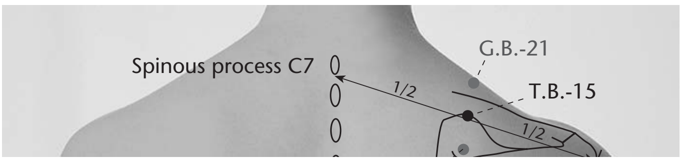
Spinous process C7
G.B.-21
1/2
T.B.-15
1/2

T.B.-15 (*tianliao*) 'Heavenly Crevice' At the midpoint of an imaginary line between the spinous process of the C7 and the lateral extremity of the acromion

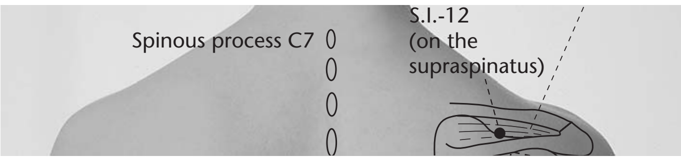
Spinous process C7
S.I.-12
(on the
supraspinatus)

S.I.-12 (*bingfeng*) 'Grasping the Wind' Directly superior to **S.I.-11**, in the centre of the supraspinous fossa

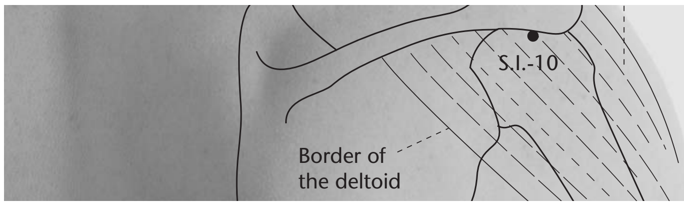
S.I.-10
Border of
the deltoid

S.I.-10 (*naoshu*) 'Upper Arm Shu' With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the scapular spine

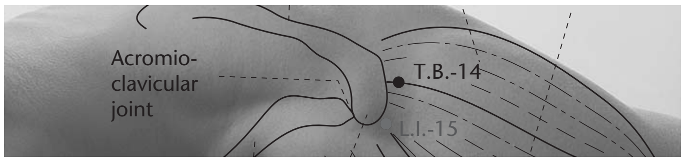
Acromio-
clavicular
joint
T.B.-14
L.I.-15

T.B.-14 (*jianliao*) 'Shoulder Crevice' Inferior to the lateral extremity of the acromion between the acromial and spinal portions of the deltoid, or with the arm abducted in the posterior of the two depressions on the shoulder joint

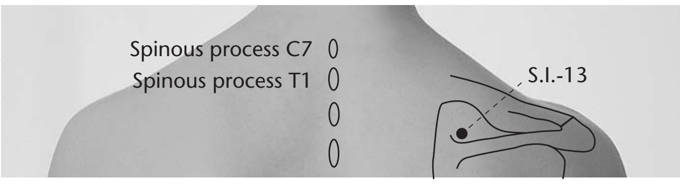
Spinous process C7
Spinous process T1
S.I.-13

S.I.-13 (*quyuan*) 'Crooked Wall' At the medial end of the supraspinous fossa

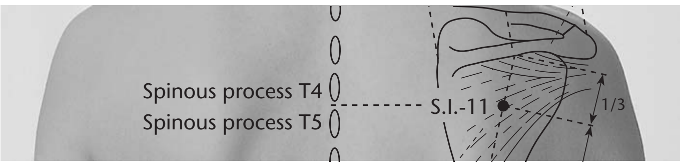
Spinous process T4
Spinous process T5
S.I.-11
1/3

S.I.-11 (*tianzong*) 'Heavenly Gathering' On the scapula, in a depression on the infraspinatus, one third of the distance from the midpoint of the scapular spine and the inferior angle of the scapula

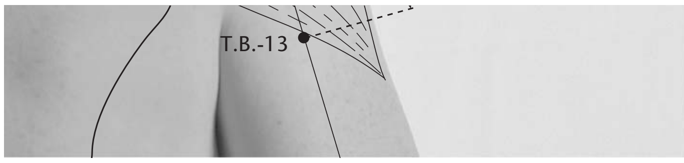

T.B.-13 (*naohui*) 'Upper Arm Meeting' 3 cun distal to the lateral extremity of the acromion (location of **T.B.-14**), on a line connecting **T.B.-14** and the olecranon, at the junction of this line with the margin of the deltoid

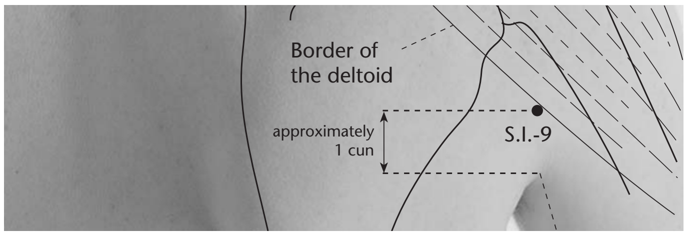
Border of
the deltoid
approximately
1 cun
S.I.-9

S.I.-9 (*jianzhen*) 'True Shoulder' With the arm adducted on a line extending in a superior direction from the posterior axillary fold, on the lower border of the deltoid, 1 cun superior to the posterior axillary fold

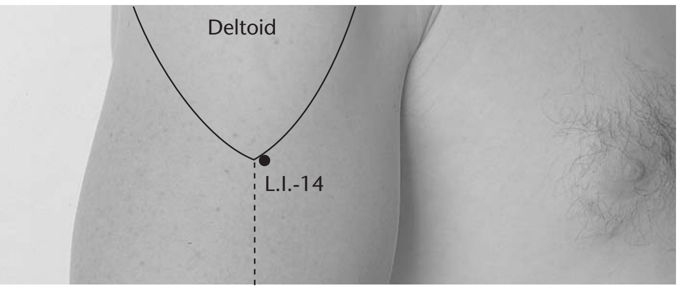
Deltoid
L.I.-14

L.I.-14 (*binao*) 'Upper Arm' 7 cun proximal to **L.I.-11**, at the distal end of the deltoid, on a line connecting **L.I.-11** and **L.I.-15**

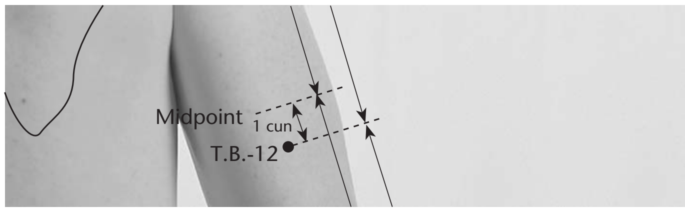
Midpoint
1 cun
T.B.-12

T.B.-12 (*xiaoluo*) 'Dispersing Luo River' 4 cun proximal to **T.B.-10** on a line connecting the olecranon and the lateral extremity of the acromion (location of **T.B.-14**)

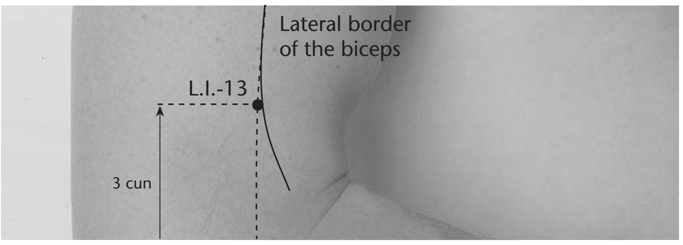
Lateral border
of the biceps
L.I.-13
3 cun

L.I.-13 (*shouwuli*) 'Arm Five Miles' On the lateral aspect of the upper arm, 3 cun proximal to **L.I.-11**

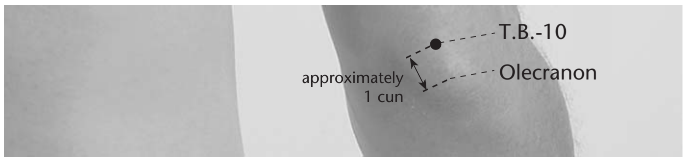
T.B.-10
Olecranon
approximately
1 cun

T.B.-10 (*tianjing*) 'Heavenly Well' With the elbow flexed in a depression approximately 1 cun proximal to the olecranon

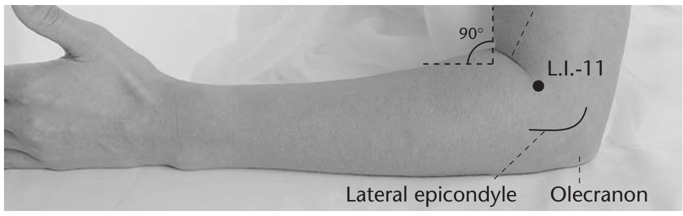
90°
L.I.-11
Lateral epicondyle
Olecranon

L.I.-11 (*quchi*) 'Pool at the Crook' With the elbow flexed, on the lateral end of the elbow crease, in a depression between the end of the crease and the lateral epicondyle of the humerus, on the extensor carpi radialis longus muscle

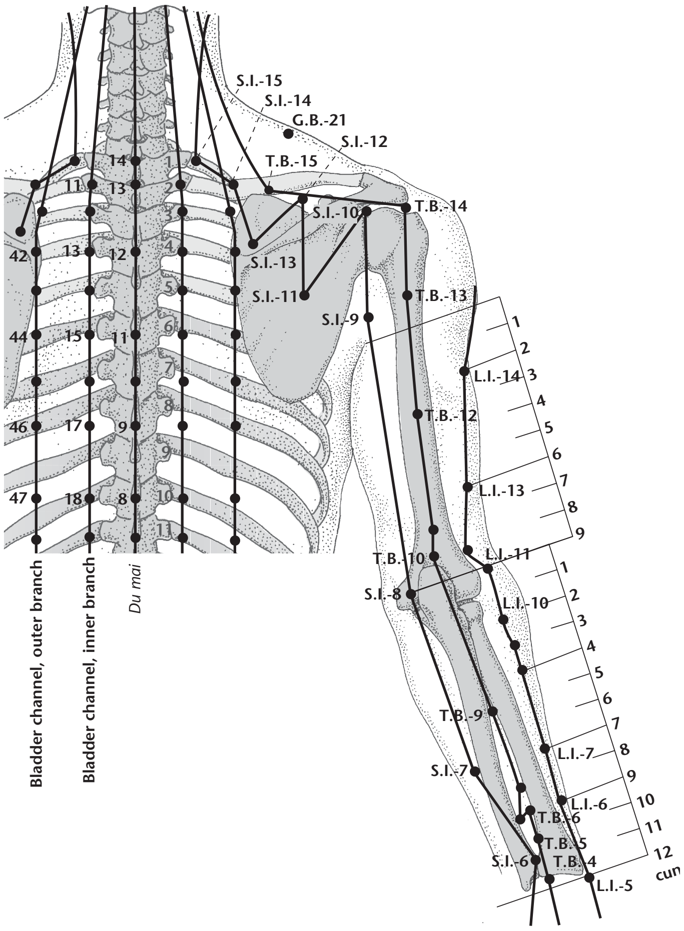
S.I.-15
S.I.-14
G.B.-21
S.I.-12
T.B.-15
T.B.-14
S.I.-10
S.I.-13
S.I.-11
T.B.-13
S.I.-9
L.I.-14
T.B.-12
L.I.-13
T.B.-10
L.I.-11
S.I.-8
L.I.-10
T.B.-9
L.I.-7
S.I.-7
L.I.-6
T.B.-6
T.B.-5
T.B.-4
S.I.-6
L.I.-5
14
13
12
11
9
8
11
13
15
17
18
42
44
46
47
1
2
3
4
5
6
7
8
9
10
11
1 2 3 4 5 6 7 8 9
1 2 3 4 5 6 7 8 9 10 11 12
cun
Bladder channel, outer branch
Bladder channel, inner branch
Du mai

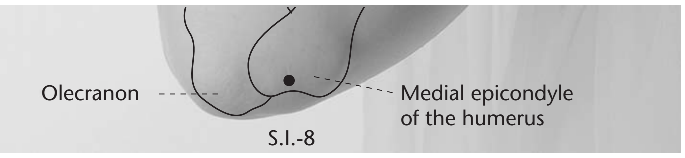
Olecranon
Medial epicondyle of the humerus
S.I.-8

S.I.-8 (*xiaohai*) 'Small Sea' With the elbow flexed in the depression between the olecranon process of the ulna and the medial epicondyle of the humerus

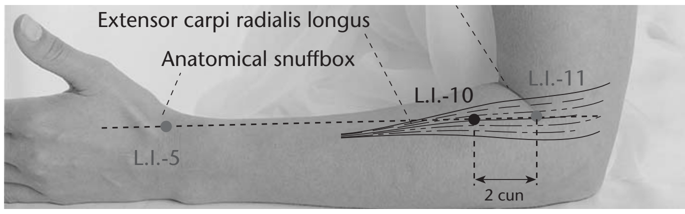
Extensor carpi radialis longus
Anatomical snuffbox
L.I.-11
L.I.-10
L.I.-5
2 cun

L.I.-10 (*shousanli*) 'Arm Three Miles' 2 cun distal to **L.I.-11**, on the line connecting **L.I.-5** and **L.I.-11**

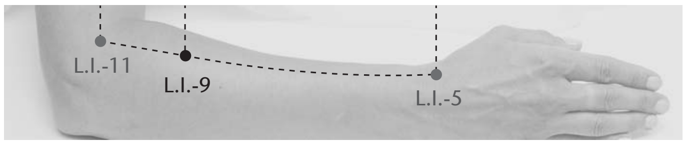
L.I.-11
L.I.-9
L.I.-5

L.I.-9 (*shanglian*) 'Upper Angle' 3 cun distal to **L.I.-11**, on the line connecting **L.I.-5** and **L.I.-11**

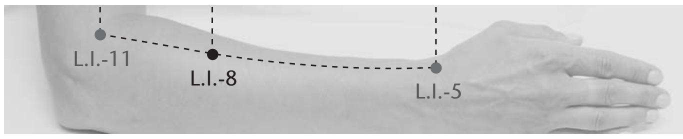
L.I.-11
L.I.-8
L.I.-5

L.I.-8 (*xialian*) 'Lower Angle' 4 cun distal L.I.-11, on the line connecting **L.I.-5** and **L.I.-11**

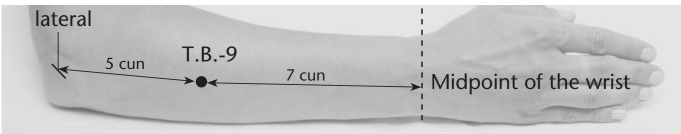
lateral
5 cun
T.B.-9
7 cun
Midpoint of the wrist

T.B.-9 (*sidu*) 'Four Rivers' 5 cun distal to the olecranon or 7 cun proximal to **T.B.-4**, between the radius and the ulna

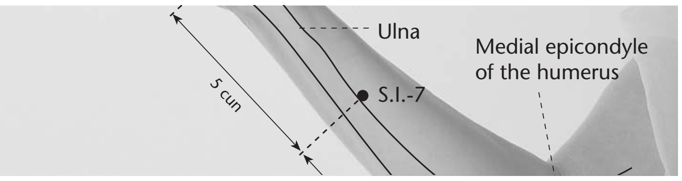
Ulna
Medial epicondyle of the humerus
5 cun
S.I.-7

S.I.-7 (*zhizheng*) 'Branch of the Upright' 5 cun proximal to the wrist joint space, on the line connecting **S.I.-5** and **S.I.-8**

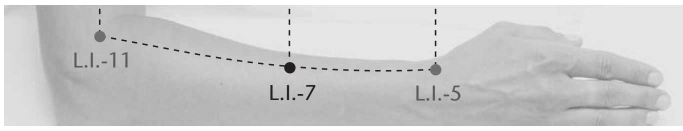
L.I.-11
L.I.-7
L.I.-5

L.I.-7 (*wenliu*) 'Warm Flow' 5 cun proximal to **L.I.-5** or midway between **L.I.-5** and **L.I.-11**

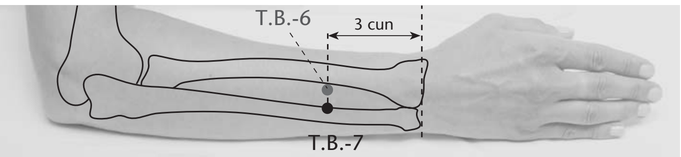
T.B.-6
3 cun
T.B.-7

T.B.-7 (*huizhong*) 'Ancestral Meeting' 3 cun proximal to **T.B.-4**, between radius and ulna, ulnar to extensor digitorum communis

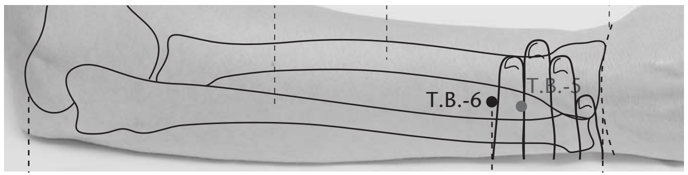

T.B.-6 (*zhigou*) 'Branch Ditch' 3 cun proximal to **T.B.-4**, between radius and ulna, radial to the tendon of extensor digitorum communis

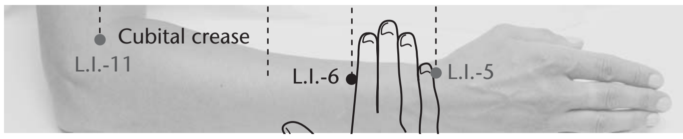
Cubital crease
L.I.-11
L.I.-6
L.I.-5

L.I.-6 (*pianli*) 'Veering Passage' 3 cun proximal to **L.I.-5**, on the line connecting **L.I.-5** and **L.I.-11**

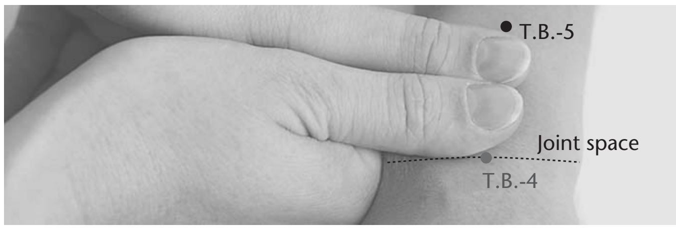
T.B.-5
Joint space
T.B.-4

T.B.-5 (*waiguan*) 'Outer Pass' 2 cun proximal to **T.B.-4** (on the dorsal wrist joint space, between radius and ulna.

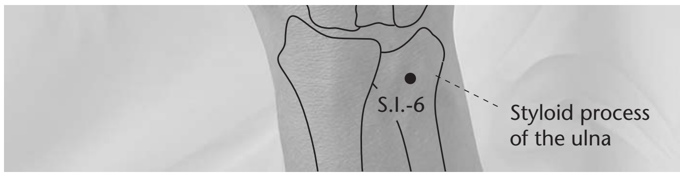
S.I.-6
Styloid process
of the ulna

S.I.-6 (*yanglao*) 'Support the Aged' In the depression radial and proximal to the styloid process of the ulna that forms when turning the hand from a pronated to a supinated position.

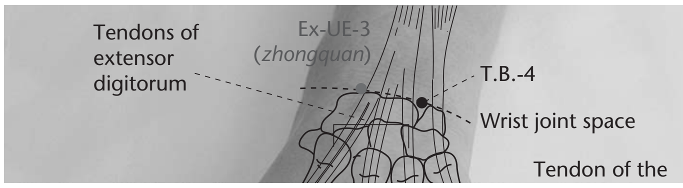
Tendons of
extensor
digitorum
Ex-UE-3
(zhongquan)
T.B.-4
Wrist joint space
Tendon of the

T.B.-4 (*yangchi*) 'Yang Pool' On the dorsum of the wrist, in the gap ulnar to the tendons of the extensor digitorum

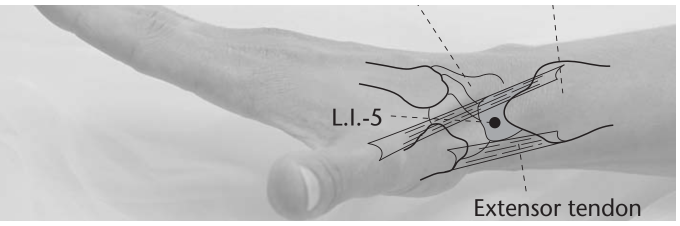
L.I.-5
Extensor tendon

L.I.-5 (*yangxi*) 'Yang Stream' With the thumb abducted in the depression between the tendons of the extensor pollicis longus and brevis muscles ('anatomical snuff box')

7.8 Anterior Shoulder and Arm

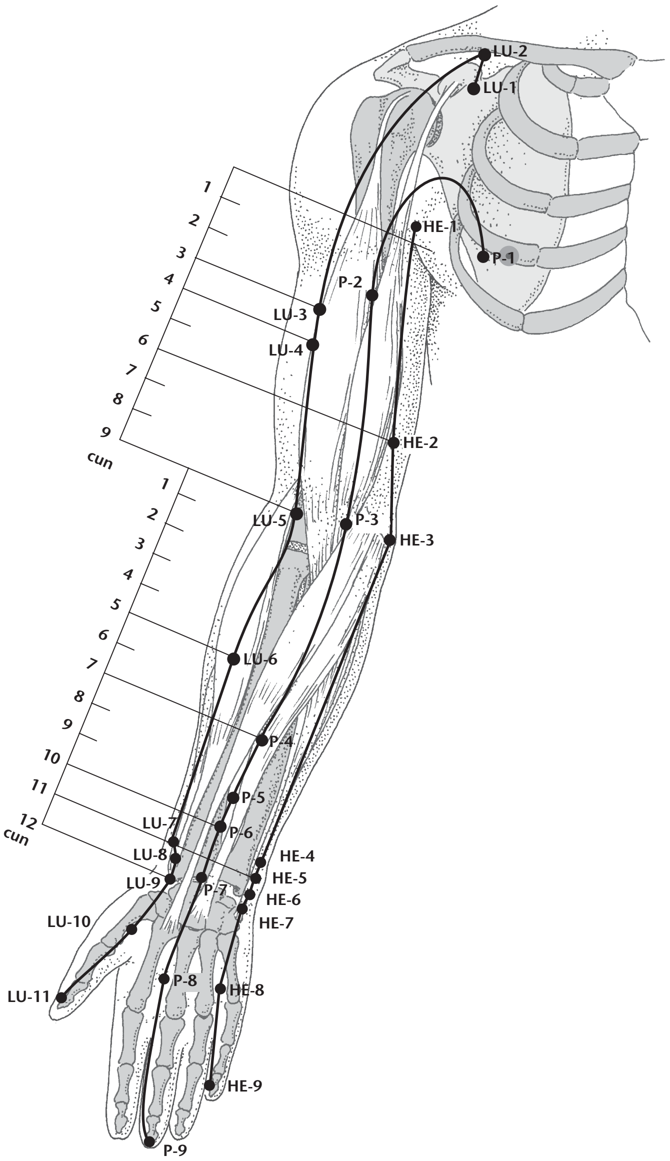
LU-2
LU-1
1
2
3
4
5
6
7
8
9
cun
HE-1
P-1
P-2
LU-3
LU-4
HE-2
1
2
3
4
5
6
7
8
9
10
11
12
cun
LU-5
P-3
HE-3
LU-6
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9

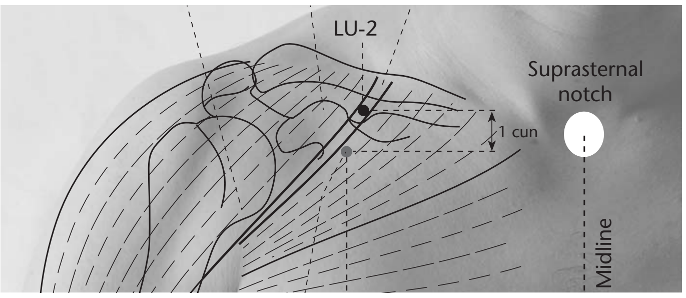
LU-2
Suprasternal notch
1 cun
Midline

LU-2 (*yunmen*) 'Cloud Gate' 6 cun lateral to the anterior midline, below the clavicle, in the centre of the deltopectoral triangle.

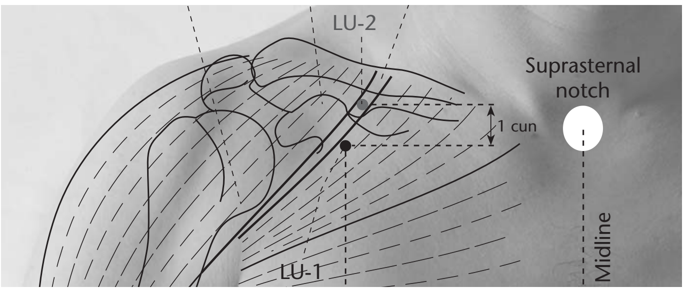
LU-2
Suprasternal notch
1 cun
LU-1
Midline

LU-1 (*zhongfu*) 'Middle Palace' 6 cun lateral to the anterior midline and approximately 1 cun inferior to **LU-2**, slightly medial to the lower border of the coracoid process

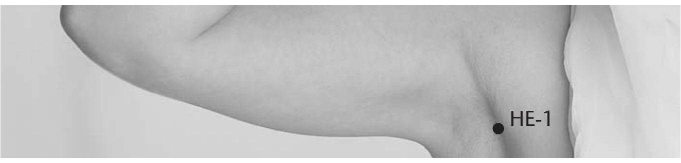

HE-1 (*jiquan*) 'Summit Spring' With the arm abducted, in the centre of the axilla, medial to the axillary artery

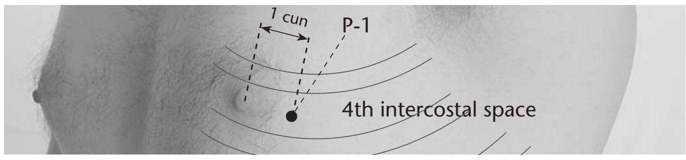
1 cun
P-1
4th intercostal space

P-1 (*tianchi*) 'Heavenly Pool' In the 4th ICS, 1 cun lateral to the nipple

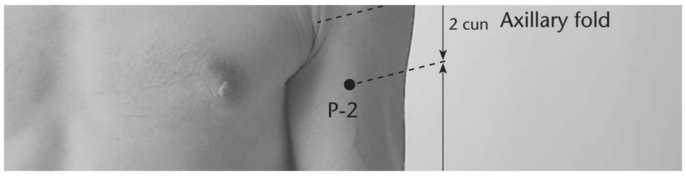
2 cun
Axillary fold
P-2

P-2 (*tianquan*) 'Heavenly Spring' Between the two heads of the biceps brachii muscle, 2 cun inferior to the axillary fold.

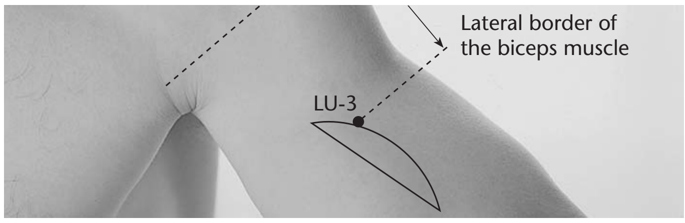
Lateral border of the biceps muscle
LU-3

LU-3 (*tianfu*) 'Palace of Heaven' 3 cun distal to the end of the anterior axillary fold, on the medial aspect of the upper arm, radial to the biceps brachialis

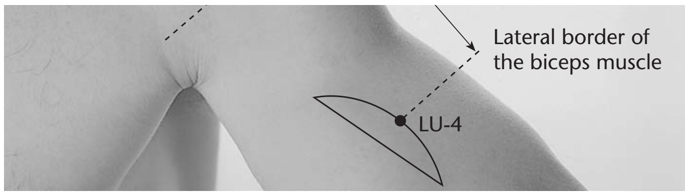
Lateral border of
the biceps muscle
LU-4

LU-4 (*xiabai*) 'Clasping the White' 4 cun distal to the end of the anterior axillary fold, on the medial aspect of the upper arm, radial to the biceps brachii

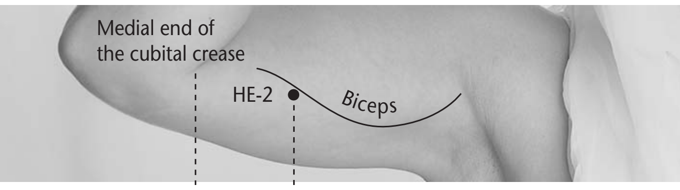
Medial end of
the cubital crease
HE-2
Biceps

HE-2 (*qingling*) 'Green Spirit' 3 cun proximal to the cubital crease, on the medial border of the biceps brachii

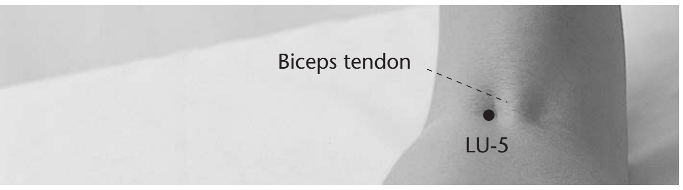
Biceps tendon
LU-5

LU-5 (*chize*) 'Cubit Marsh' In the cubital crease, radial to the tendon of the biceps brachii

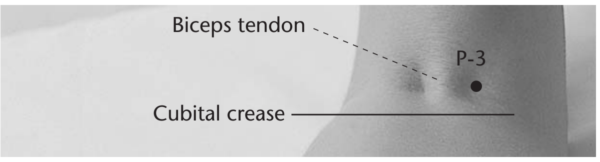
Biceps tendon
P-3
Cubital crease

P-3 (*quze*) 'Marsh at the Crook' In the cubital crease, ulnar to the biceps tendon

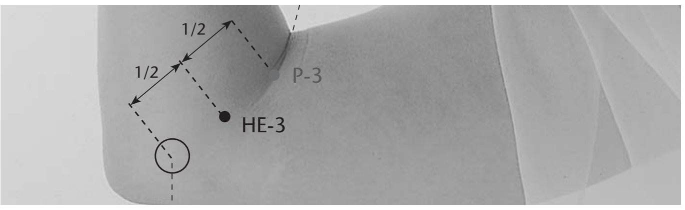

HE-3 (*shaohai*) 'Lesser Sea' With the elbow flexed, in the depression between the ulnar end of the cubital crease and the medial epicondyle of the humerus

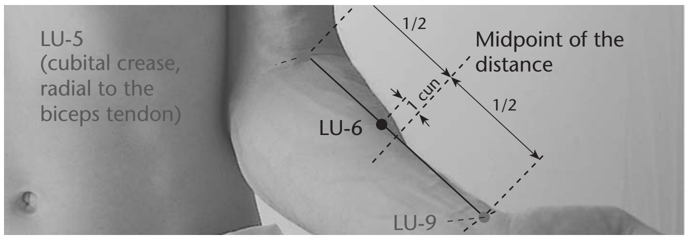
LU-5
(cubital crease,
radial to the
biceps tendon)
1/2
Midpoint of the
distance
1 cun
1/2
LU-6
LU-9

LU-6 (*kongzui*) 'Maximum Opening' 5 cun distal to **LU-5** or 7 cun proximal to the wrist joint space, on a line connecting **LU-5** and **LU-9**

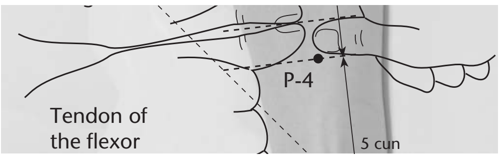
P-4
Tendon of
the flexor
5 cun

P-4 (*ximen*) 'Xi-Cleft Gate' 5 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis, or 1 cun distal to the midpoint of the line connecting **P-3** and **P-7**

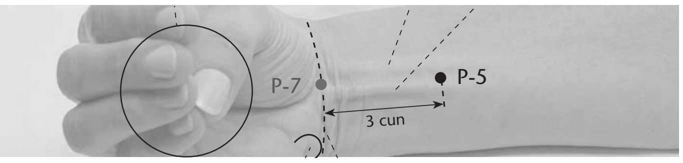

P-5 (*jianshi*) 'Intermediate Messenger' 3 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis

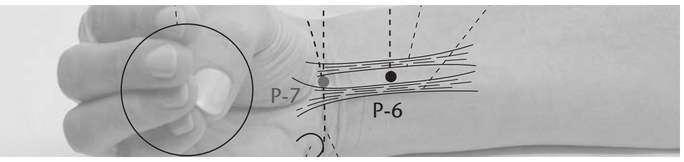

P-6 (*neiguan*) 'Inner Pass' 2 cun proximal to the palmar wrist joint space, between the tendons of palmaris longus and flexor carpi radialis

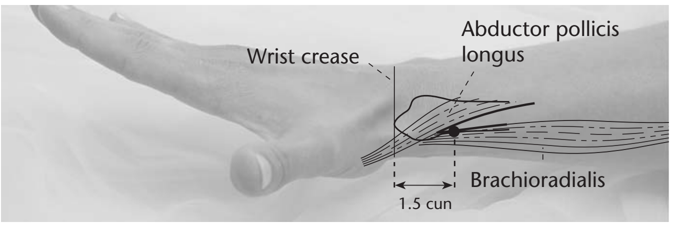
Wrist crease
Abductor pollicis
longus
Brachioradialis
1.5 cun

LU-7 (*lieque*) 'Broken Sequence' On the radial aspect of the forearm, immediately above the styloid process of the radius, approximately 1.5 cun proximal to the wrist joint space in a V-shaped groove

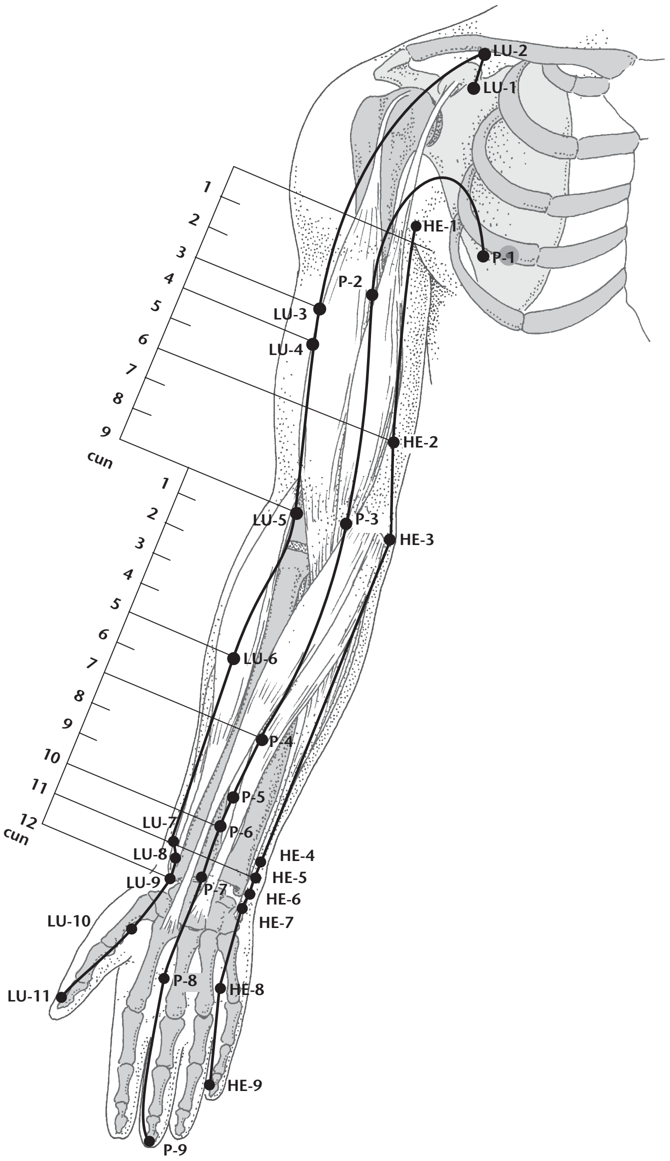
LU-2
LU-1
HE-1
P-1
P-2
LU-3
LU-4
HE-2
LU-5
P-3
HE-3
LU-6
P-4
P-5
P-6
LU-7
LU-8
LU-9
P-7
HE-4
HE-5
HE-6
HE-7
LU-10
LU-11
P-8
HE-8
HE-9
P-9
1
2
3
4
5
6
7
8
9
cun
1
2
3
4
5
6
7
8
9
10
11
12
cun

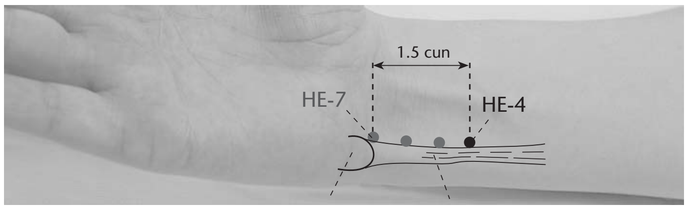
1.5 cun
HE-7
HE-4

HE-4 (*lingdao*) 'Spirit Path' 1.5 cun proximal to the palmar wrist joint space, radial to the flexor carpi ulnaris tendon

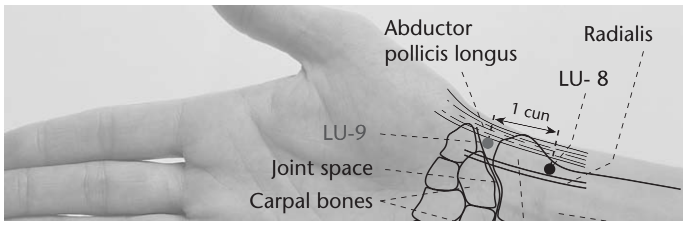
Abductor pollicis longus
Radialis
LU- 8
1 cun
LU-9
Joint space
Carpal bones

LU-8 (*jingqu*) 'Channel Gutter' 1 cun proximal to the palmar wrist joint space (**LU-9**), lateral to the radial artery

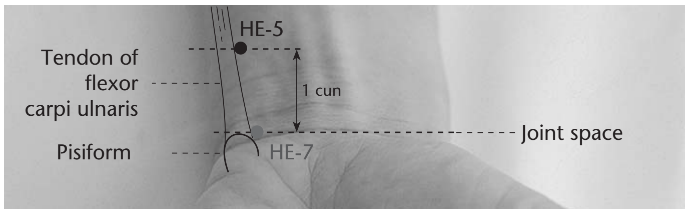
HE-5
Tendon of flexor carpi ulnaris
1 cun
Joint space
Pisiform
HE-7

HE-5 (*tongli*) 'Penetrating the Interior' 1 cun proximal to the palmar wrist joint space, on the radial side of the flexor carpi ulnaris tendon

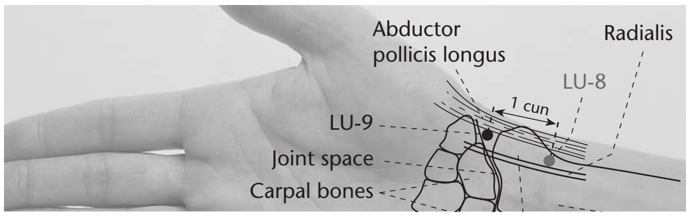
Abductor pollicis longus
Radialis
LU-8
1 cun
LU-9
Joint space
Carpal bones

LU-9 (*taiyuan*) 'Supreme Abyss' On the wrist joint space, lateral to the radial artery and ulnar to the tendon of abductor pollicis longus

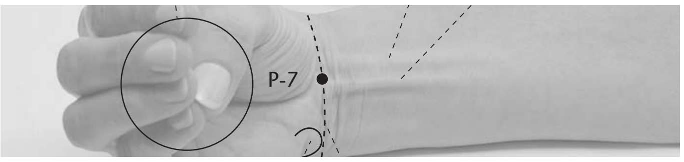

P-7 (*daling*) 'Great Mound' In the centre of the palmar aspect of the wrist joint space ('most distal wrist crease'), between the tendons of palmaris longus and flexor carpi radialis

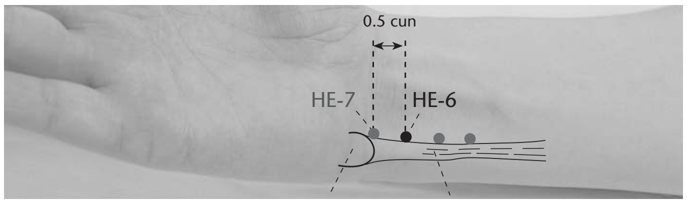
0.5 cun
HE-7
HE-6

HE-6 (*yinxi*) 'Yin Cleft' 0.5 cun proximal to the palmar wrist joint space, radial to the flexor carpi ulnaris tendon

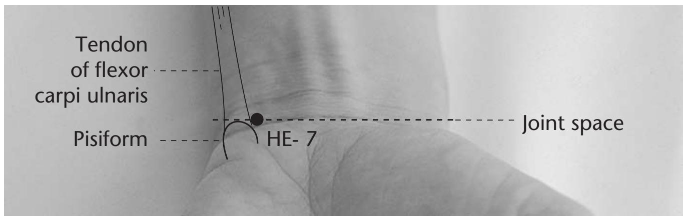
Tendon of flexor carpi ulnaris
Pisiform
HE- 7
Joint space

HE-7 (*shenmen*) 'Spirit Gate' On the palmar aspect of the wrist joint space ('most distal wrist crease'), in the depression radial to the insertion of the flexor carpi ulnaris tendon

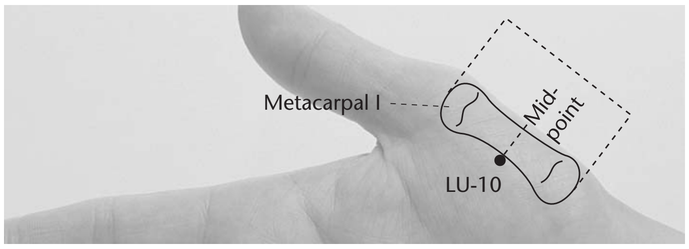
Metacarpal I
Mid-point
LU-10

LU-10 (*yuji*) 'Fish Border' At the midpoint of the palmar border of the 1st metacarpal bone

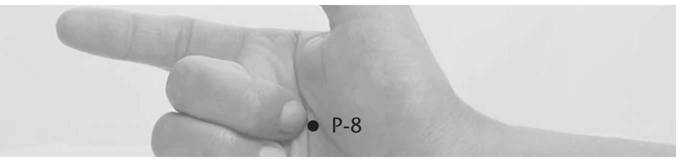

P-8 (*laogong*) 'Palace of Toil' In the centre of the palm, between the 2nd and 3rd metacarpal bones, slightly closer to the 3rd metacarpal bone. When a fist is made, the point is located beneath the tip of the middle finger

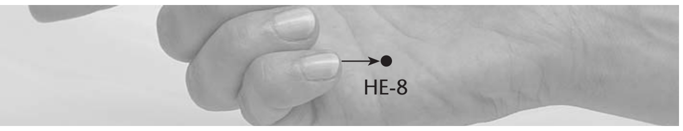

HE-8 (*shaofu*) 'Lesser Palace' When a fist is made, beneath the tip of the little finger, on the palm, between the 4th and 5th metacarpal bones

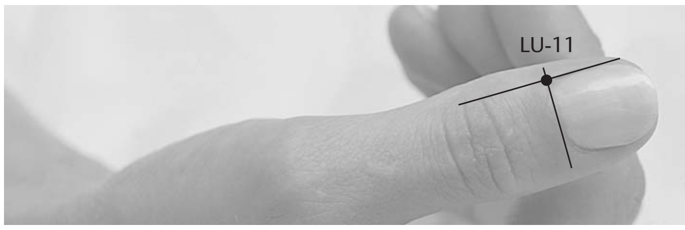

LU-11 (*shaoshang*) 'Lesser Shang' 0.1 cun from the radial corner of the thumb nail

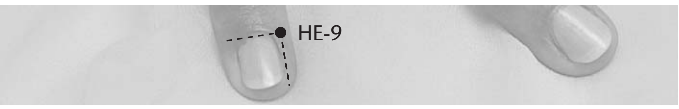

HE-9 (*shaochong*) 'Lesser Rushing' 0.1 cun lateral to the radial corner of the nail of the little finger

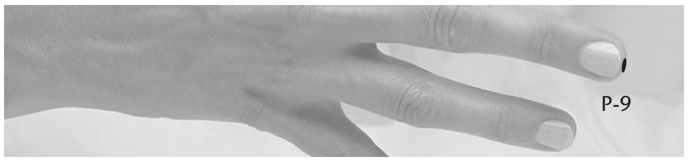

P-9 (*zhongchong*) 'Middle Rushing' On the most distal point of the middle finger

7.9 Dorsum of the Hand

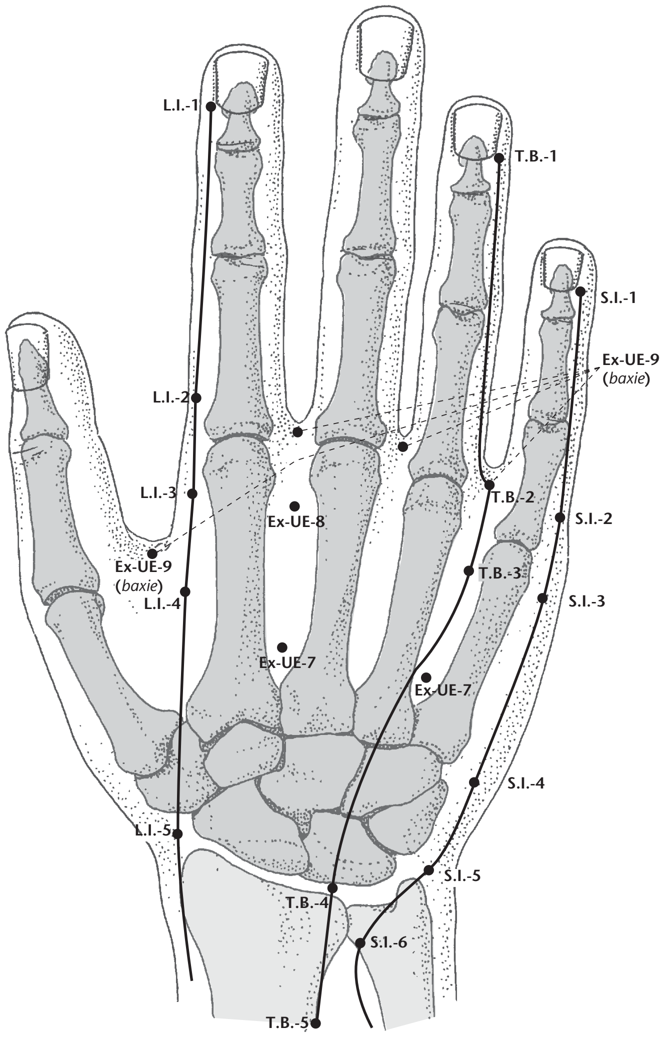
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(*baxie*)
L.I.-2
L.I.-3
T.B.-2
S.I.-2
Ex-UE-8
Ex-UE-9
(*baxie*)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

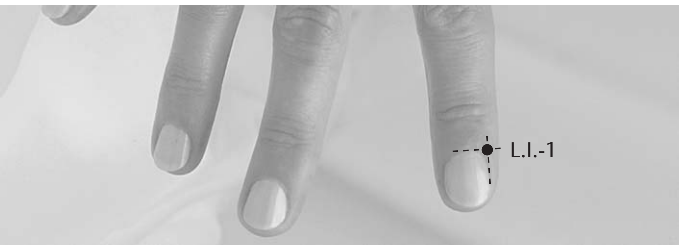

L.I.-1 (*shangyang*) 'Shang Yang' On the index finger, 0.1 cun from the radial corner of the nail

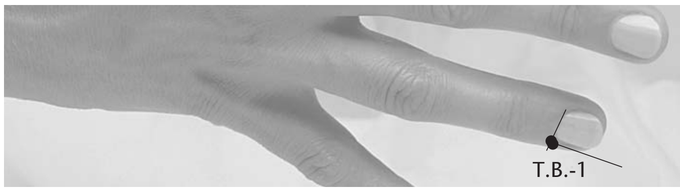

T.B.-1 (*guanchong*) 'Rushing Pass' On the ring finger (4th finger), 0.1 cun from the ulnar corner of the nail

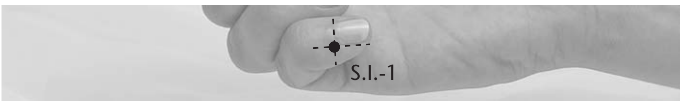

S.I.-1 (*shaoze*) 'Lesser Marsh' On the little finger, 0.1 cun from the ulnar corner of the nail

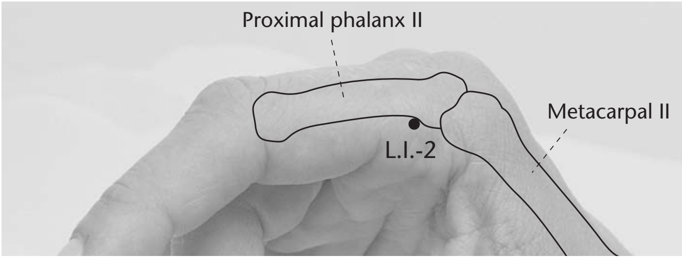
Proximal phalanx II
Metacarpal II
L.I.-2

L.I.-2 (*erjian*) 'Second Space' On the radial aspect of the index finger, distal to the metacarpophalangeal joint at the junction of the shaft and the basis of the proximal phalanx

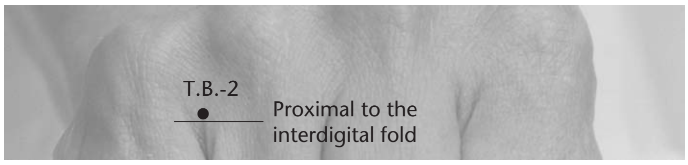
T.B.-2
Proximal to the interdigital fold

T.B.-2 (*yemen*) 'Fluid Gate' Between the little finger and ring finger, proximal to the margin of the web; locate with the patient making a fist

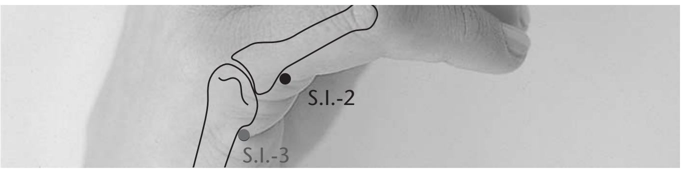

S.I.-2 (*qiangu*) 'Front Valley' On the ulnar aspect of the little finger, at the junction of the shaft and the base of the proximal phalanx

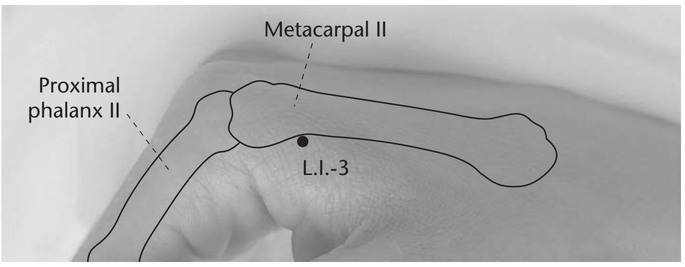
Metacarpal II
Proximal phalanx II
L.I.-3

L.I.-3 (*sanjian*) 'Third Space' On the radial aspect of the index finger, proximal to the metacarpophalangeal joint at the junction of the shaft and the head of the 2nd metacarpal bone

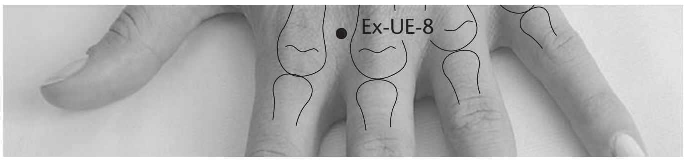

Ex-UE-8 (*wailaogong/luozhen/xianqiang*) 'Stiff Neck' On the dorsum of the hand, between the 2nd and 3rd metacarpal bones, approximately 0.5 cun proximal to the metacarpophalangeal joints

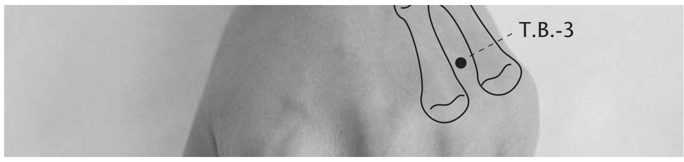

T.B.-3 (*zhongzhu*) 'Central Islet' On the dorsum of the hand, in the depression between the 4th and 5th metacarpal bones, proximal to the metacarpophalangeal joints. **T.B.-3** is located at the junction between the heads and shafts of the two metacarpals

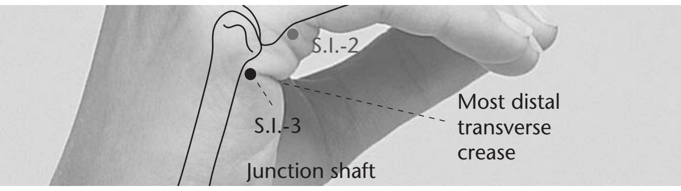
S.I.-2
S.I.-3
Most distal transverse crease
Junction shaft

S.I.-3 (*houxi*) 'Back Stream' On the ulnar border of the hand, in the depression proximal to the metacarpophalangeal joint, at the junction between the head and shaft of the 5th metacarpal

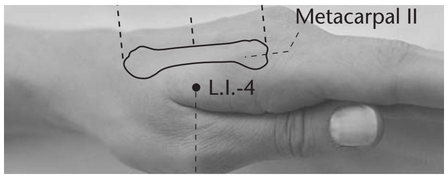
Metacarpal II
L.I.-4

L.I.-4 (*hegu*) 'Joining Valley' On the radial aspect of the hand, between the 1st and 2nd metacarpal bone, closer to the 2nd metacarpal bone and approximately at its midpoint

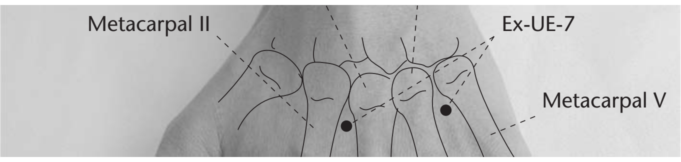
Metacarpal II
Ex-UE-7
Metacarpal V

Ex-UE-7 (*yaotongdian/yaotongxue*) 'Lumbar Pair Point' Two points on the dorsum of the hand, between the 2nd/3rd and 4th/5th metacarpal bones

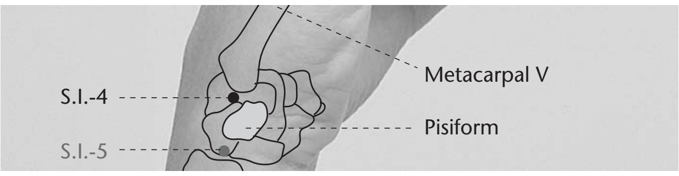
S.I.-4
S.I.-5
Metacarpal V
Pisiform

S.I.-4 (*wangu*) 'Wrist Bone' On the ulnar border of the hand, between the 5th metacarpal and the carpal bones, at the border of the red and white skin

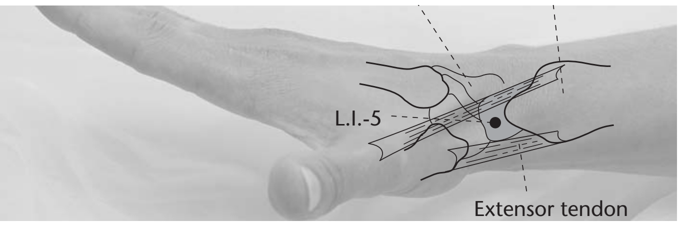
L.I.-5
Extensor tendon

L.I.-5 (*yangxi*) 'Yang Stream' With the thumb abducted in the depression between the tendons of the extensor pollicis longus and brevis muscles ('anatomical snuff box')

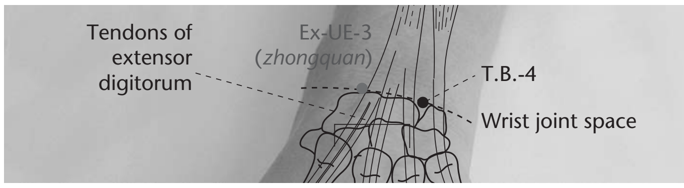
Tendons of extensor digitorum
Ex-UE-3 (zhongquan)
T.B.-4
Wrist joint space

T.B.-4 (*yangchi*) 'Yang Pool' On the dorsum of the wrist, in the gap ulnar to the tendons of extensor digitorum

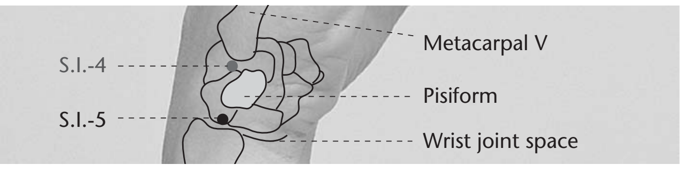
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Wrist joint space

S.I.-5 (*yanggu*) 'Yang Valley' On the ulnar aspect of the wrist, at the level of the lateral joint space

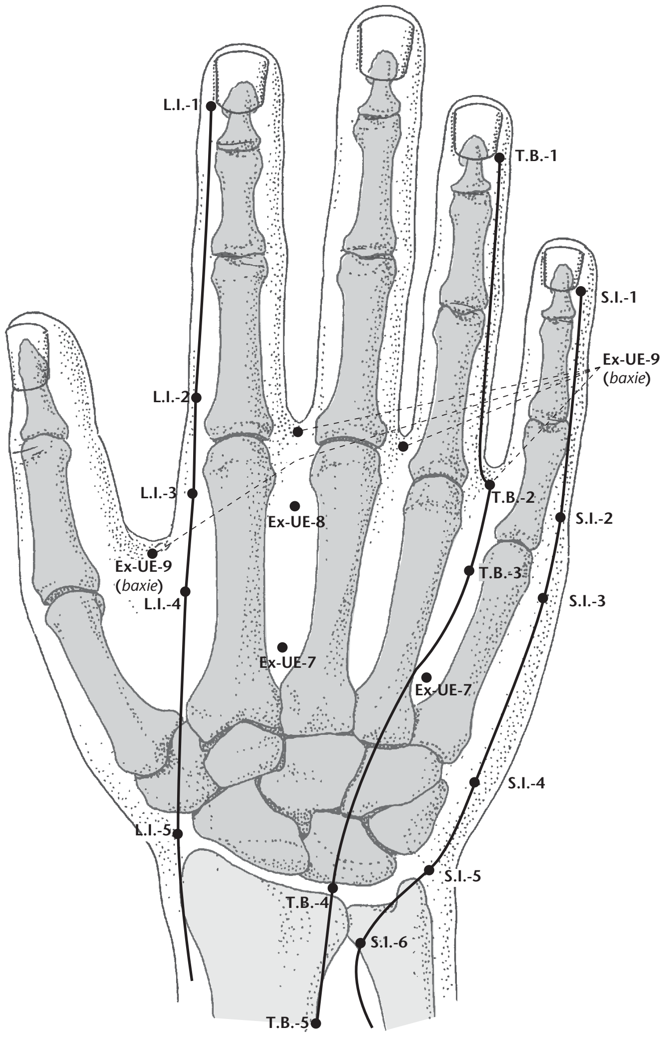
L.I.-1
T.B.-1
S.I.-1
Ex-UE-9
(baxie)
L.I.-2
L.I.-3
Ex-UE-8
T.B.-2
S.I.-2
Ex-UE-9
(baxie)
L.I.-4
T.B.-3
S.I.-3
Ex-UE-7
Ex-UE-7
S.I.-4
L.I.-5
S.I.-5
T.B.-4
S.I.-6
T.B.-5

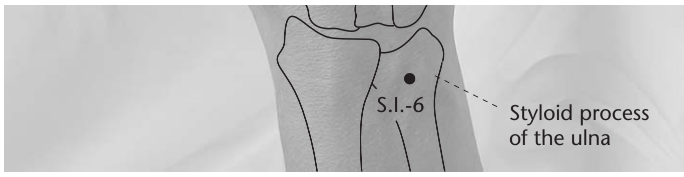
S.I.-6
Styloid process
of the ulna

S.I.-6 (*yanglao*) 'Nourishing the Old' In the depression radial and proximal to the styloid process of the ulna that forms when the hand is turned from a pronated to a supinated position

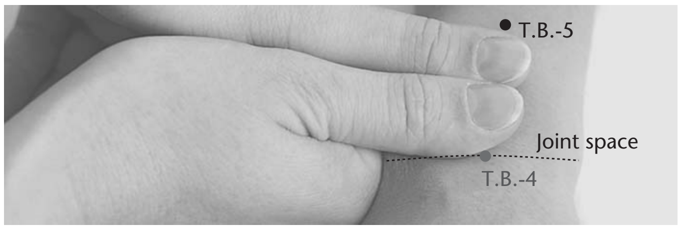
T.B.-5
Joint space
T.B.-4

T.B.-5 (*waiguan*) 'Outer Pass' 2 cun proximal to **T.B.-4** (on the dorsal wrist joint space, between radius and ulna)

7.10 Anterior Leg

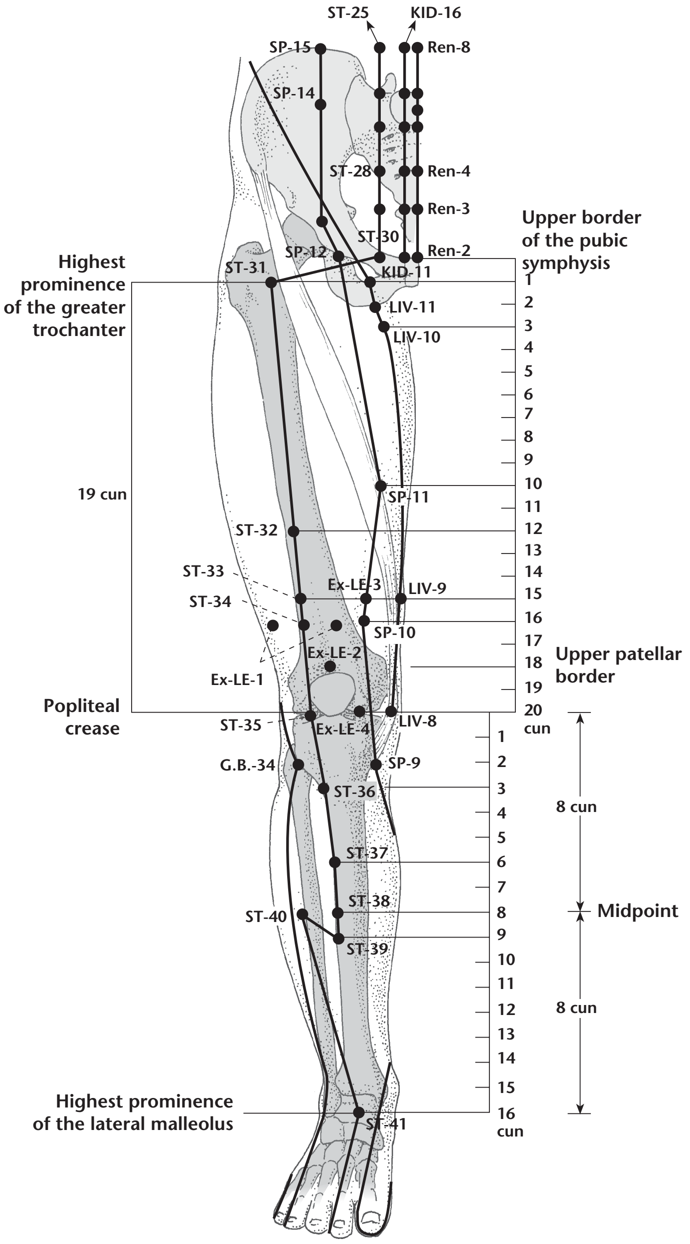
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
ST-30
SP-12
Ren-2
ST-31
KID-11
Upper border
of the pubic
symphysis
Highest
prominence
of the greater
trochanter
LIV-11
LIV-10
19 cun
SP-11
ST-32
ST-33
ST-34
Ex-LE-3
LIV-9
SP-10
Ex-LE-2
Ex-LE-1
Upper patellar
border
Popliteal
crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-39
ST-40
8 cun
Midpoint
8 cun
Highest prominence
of the lateral malleolus
ST-41
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun

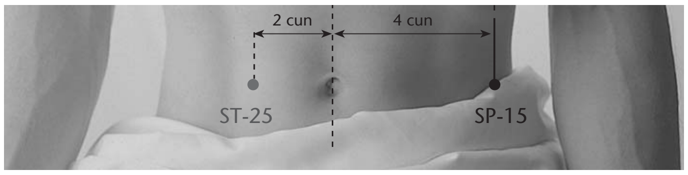
2 cun
4 cun
ST-25
SP-15

SP-15 (*daheng*) 'Great Horizontal' 4 cun lateral to the centre of the umbilicus, on the mamillary line

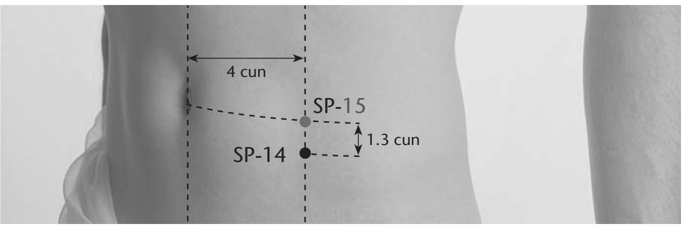
4 cun
SP-15
1.3 cun
SP-14

SP-14 (*fujie*) 'Abdomen Knot' 4 cun lateral to the anterior midline (mamillary line), 3 cun superior to **SP-13** or 1.3 cun inferior to **SP-15**

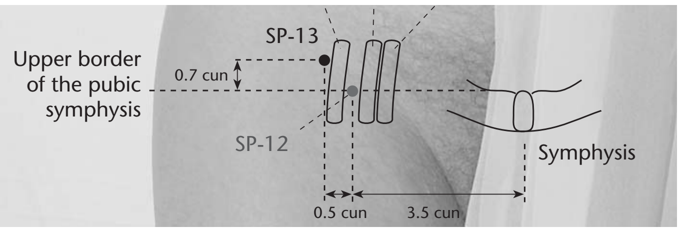
SP-13
Upper border
of the pubic
symphysis
0.7 cun
SP-12
Symphysis
0.5 cun
3.5 cun

SP-13 (*fushe*) 'Meeting of the Fu Organs' 4 cun lateral to the anterior midline (mamillary line) and 0.7 cun superior to the upper border of the pubic symphysis

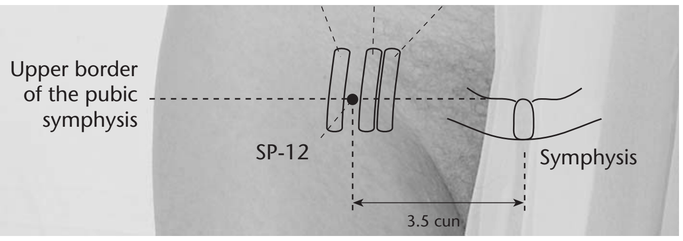
Upper border
of the pubic
symphysis
SP-12
Symphysis
3.5 cun

SP-12 (*chongmen*) 'Rushing Gate' 3.5 cun lateral to the midline, level with the upper border of the pubic symphysis, lateral to the femoral artery

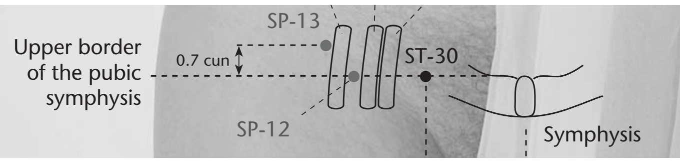
SP-13
Upper border
of the pubic
symphysis
0.7 cun
ST-30
SP-12
Symphysis

ST-30 (*qichong*) 'Rushing Qi' 2 cun lateral to the upper border of the pubic symphysis, medial to the femoral artery; at the level of **Ren-2**

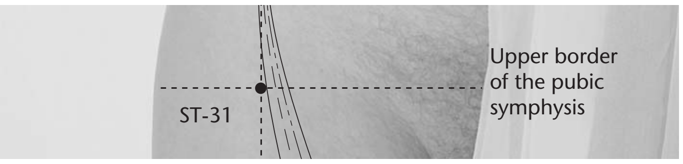
Upper border
of the pubic
symphysis
ST-31

ST-31 (*biguan*) 'Thigh Gate' With the hip joint flexed, in the depression below the anterior superior iliac spine, lateral to the sartorius

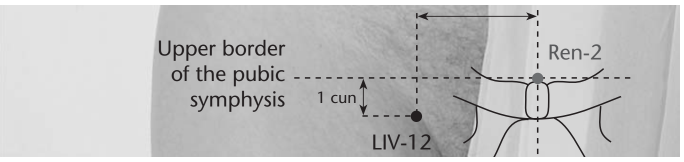
Upper border
of the pubic
symphysis
Ren-2
1 cun
LIV-12

LIV-12 (*jimai*) 'Urgent Pulse' In the inguinal groove, 2.5 cun lateral to the midline and 1 cun inferior to the upper border of the symphysis, inferior to **ST-30** and medial to the femoral artery

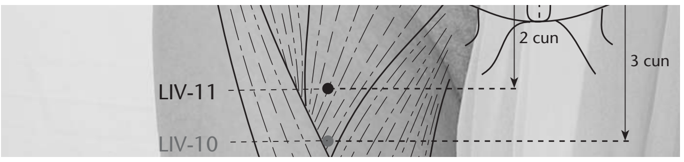
2 cun
3 cun
LIV-11
LIV-10

LIV-11 (*yinlian*) 'Yin Corner' 2 cun inferior to **ST-30** (level with the upper border of the pubic symphysis), on the lateral border of the adductor longus muscle or approximately 1 cun inferior to where the femoral artery passes the inguinal ligament

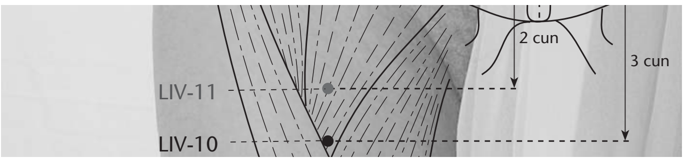
2 cun
3 cun
LIV-11
LIV-10

LIV-10 (*zuwuli*) 'Leg Five Miles' On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of adductor longus

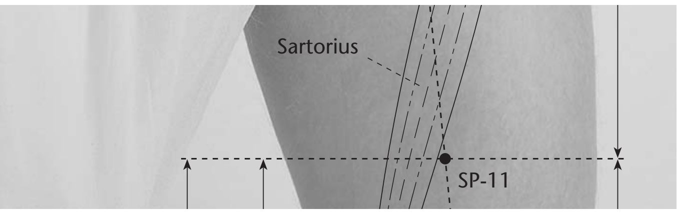
Sartorius
SP-11

SP-11 (*jimen*) 'Winnowing Gate' 6 cun proximal to **SP-10** at the midpoint of the femur, in the depression between the sartorius and vastus lateralis

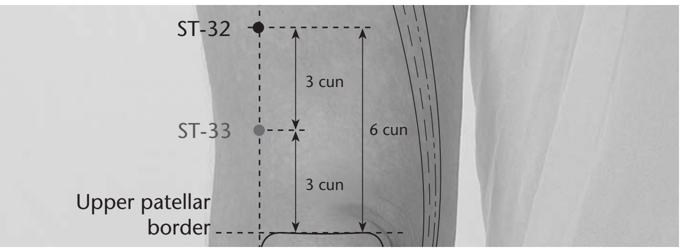
ST-32
3 cun
ST-33
6 cun
3 cun
Upper patellar
border

ST-32 (*futu*) 'Crouching Rabbit' 6 cun proximal to the lateral upper border of the patella, on a line connecting with the anterior superior iliac spine

Ch07-F10028.qxd 2/22/08 10:09 PM Page 648

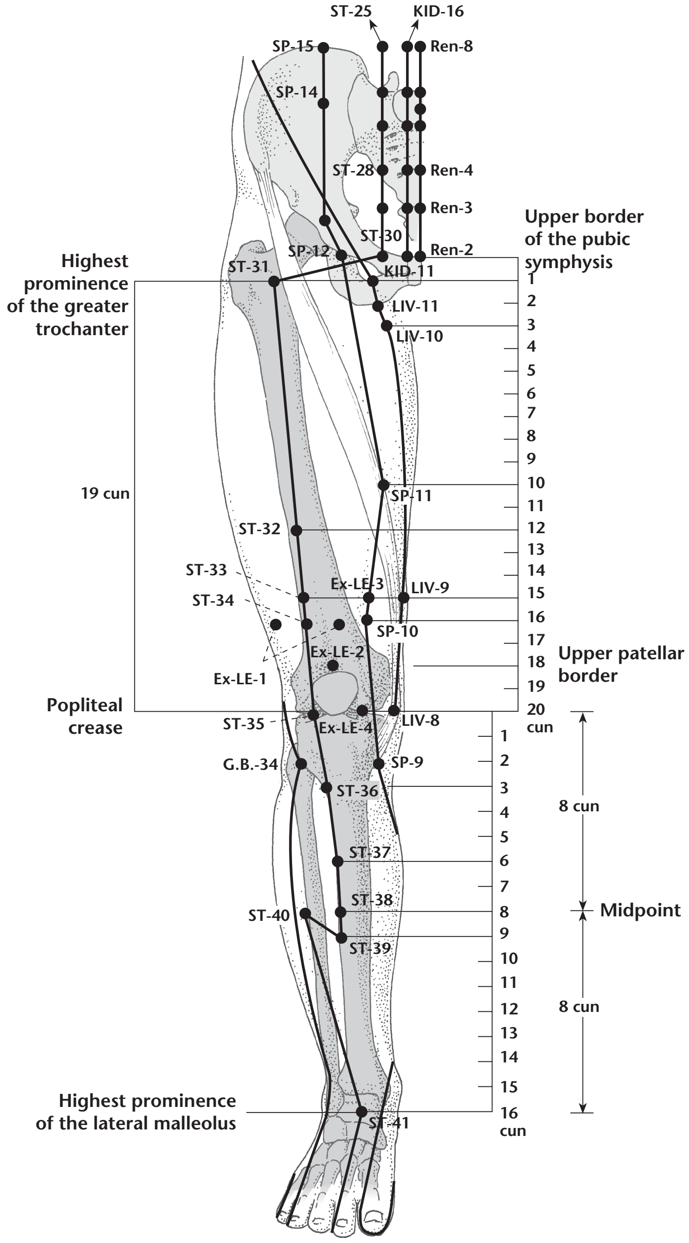
ST-25
KID-16
SP-15
Ren-8
SP-14
ST-28
Ren-4
Ren-3
Upper border of the pubic symphysis
ST-30
SP-12
Ren-2
Highest prominence of the greater trochanter
ST-31
KID-11
LIV-11
LIV-10
19 cun
SP-11
ST-32
ST-33
Ex-LE-3
LIV-9
ST-34
SP-10
Ex-LE-2
Upper patellar border
Ex-LE-1
Popliteal crease
ST-35
Ex-LE-4
LIV-8
G.B.-34
SP-9
ST-36
ST-37
ST-38
ST-40
ST-39
Highest prominence of the lateral malleolus
ST-41
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20 cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16 cun
8 cun
Midpoint
8 cun

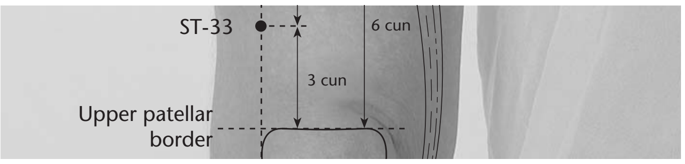
ST-33
6 cun
3 cun
Upper patellar border

ST-33 (*yinshi*) 'Yin Market' 3 cun superior to the upper lateral border of the patella

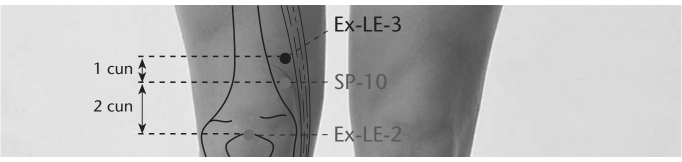
Ex-LE-3
1 cun
SP-10
2 cun
Ex-LE-2

Ex-LE-3 (*baichongwo*) 'Hundred Insect Burrow' 3 cun superior and 1 cun medial to the upper medial border of the patella, in a small depression on the vastus medialis muscle or 1 cun superior to **SP-10**

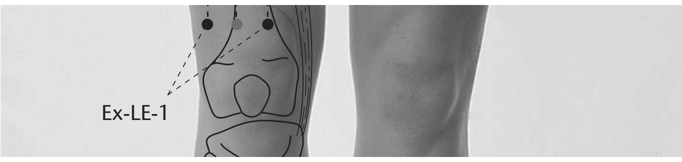

Ex-LE-1 (*kuangu*) 'Hip Bone' A pair of points 2 cun superior to the patella and 1.5 cun lateral and medial to **ST-34**

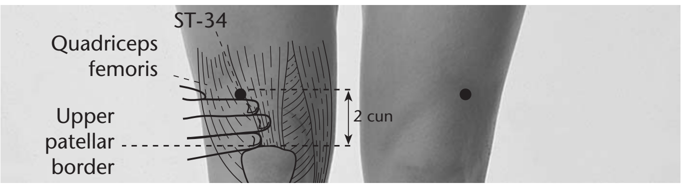
ST-34
Quadriceps femoris
Upper patellar border
2 cun

ST-34 (*liangqiu*) 'Ridge Mound' 2 cun proximal to the upper lateral border of the patella, in a depression on vastus lateralis

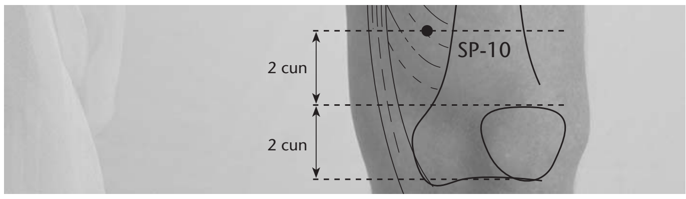
SP-10
2 cun
2 cun

SP-10 (*xuehai*) 'Sea of Blood' With the knee flexed, 2 cun proximal to the medial superior border of the patella. To find: place the right palm on the patient's left patella and locate the point at the tip of the thumb

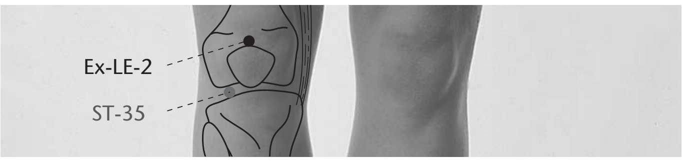

Ex-LE-2 (*heding/xiding*) 'Crane's Summit' With the knee flexed, at the midpoint of the upper patellar border

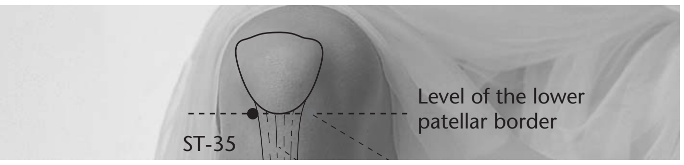
Level of the lower
patellar border
ST-35

ST-35 (*dubi*) 'Calf's Nose' With the knee flexed, in the depression inferior to the patella and lateral to the patellar ligament; corresponds to the lateral eye of the knee; **Ex-LE-5** (*xiyan*) is the combination of this point and **Ex-LE-4**

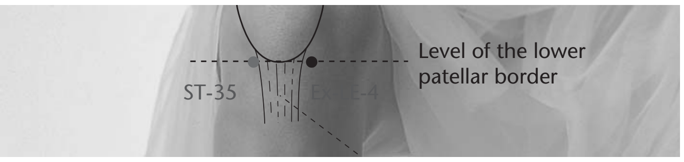
Level of the lower
patellar border
ST-35

Ex-LE-4 (*neixiyan*) 'Inner Eye of the Knee' Inferior to the patella, in a depression medial to the patellar ligament; **Ex-LE-5** (*xiyan*) is the combination of this point and **ST-35**

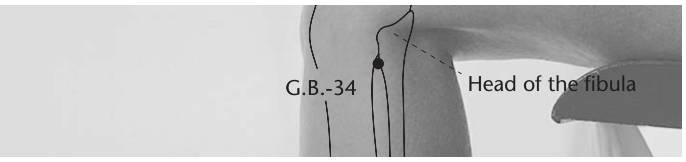
G.B.-34
Head of the fibula

G.B.-34 (*yanglingquan*) 'Yang Mound Spring' In the depression anterior and inferior to the head of the fibula

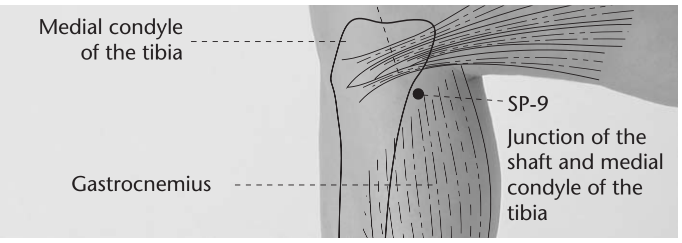
Medial condyle
of the tibia
SP-9
Junction of the
shaft and medial
condyle of the
tibia
Gastrocnemius

SP-9 (*yinlingquan*) 'Yin Mound Spring' With the knee flexed in a depression distal to the medial condyle of the tibia. To find: palpate from the medial posterior border of the tibia towards proximal, at roughly the same level as **G.B.-34**

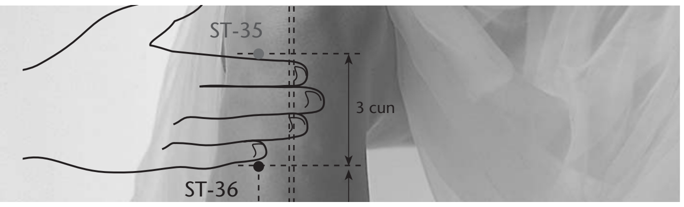
ST-35
3 cun
ST-36

ST-36 (*zusanli*) 'Leg Three Miles' one fingerbreadth (middle finger) lateral to the anterior crest of the tibia, 3 cun distal to **ST-35**

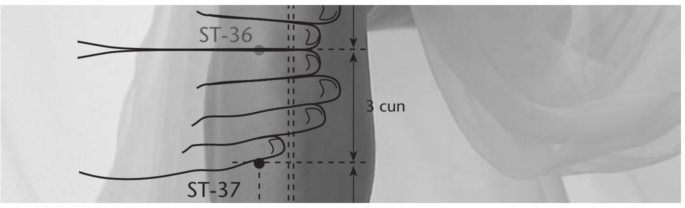
ST-36
3 cun
ST-37

ST-37 (*shangjuxu*) 'Upper Great Void' 1 fingerbreadth lateral to the anterior crest of the tibia, 6 cun distal to **ST-35** or 3 cun distal to **ST-36**

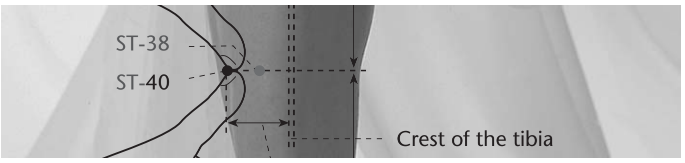
ST-38
ST-40
Crest of the tibia

ST-40 (*fenglong*) 'Abundant Bulge' 2 fingerbreadths lateral to the anterior crest of the tibia, at the midpoint of the line joining **ST-35** ('lateral eye of the knee', level with the knee joint space) and the highest prominence of the lateral malleolus

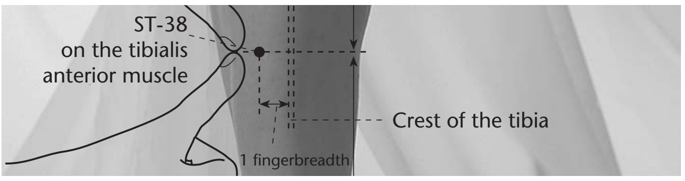
ST-38
on the tibialis
anterior muscle
Crest of the tibia
1 fingerbreadth

ST-38 (*tiaokou*) 'Lines Opening' 1 fingerbreadth lateral to the anterior crest of the tibia, at the midpoint of the line joining **ST-35** ('lateral eye of the knee', level with the knee joint space) and the highest prominence of the lateral malleolus

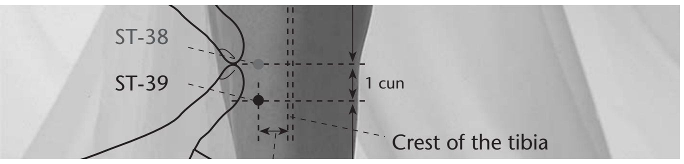
ST-38
ST-39
1 cun
Crest of the tibia

ST-39 (*xiajuxu*) 'Lower Great Void' 1 cun distal to **ST-38**

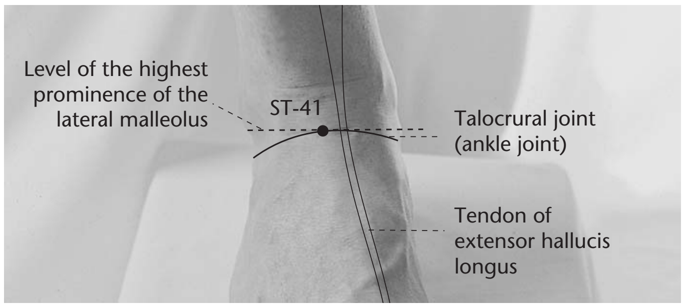
Level of the highest
prominence of the
lateral malleolus
ST-41
Talocrural joint
(ankle joint)
Tendon of
extensor hallucis
longus

ST-41 (*jiexi*) 'Stream Divide' On the ankle, between the tendons of the extensor digitorum and the extensor hallucis longus, at the level of the highest prominence of the lateral malleolus

7.11 Medial Leg

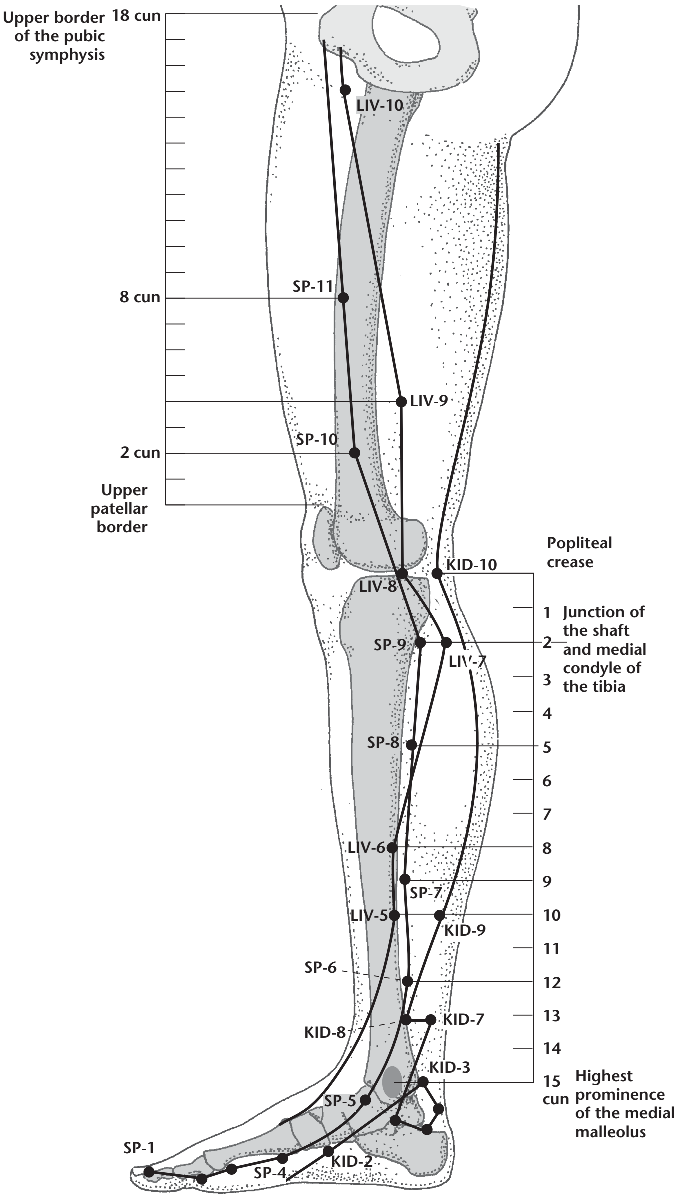
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
2 cun
SP-10
Upper patellar border
Popliteal crease
KID-10
LIV-8
SP-9
LIV-7
1 Junction of the shaft and medial condyle of the tibia
2
3
4
5
6
7
8
9
10
11
12
13
14
15 cun
SP-8
LIV-6
SP-7
LIV-5
KID-9
SP-6
KID-8
KID-7
KID-3
SP-5
Highest prominence of the medial malleolus
SP-1
KID-2
SP-4

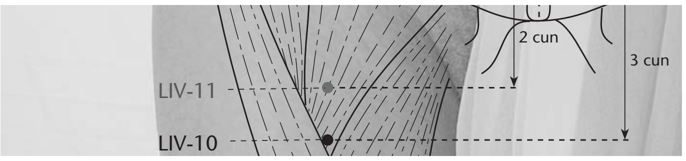
2 cun
3 cun
LIV-11
LIV-10

LIV-10 (*zuwuli*) 'Leg Five Miles' On the anterior aspect of the thigh, 3 cun inferior to the upper border of the symphysis, on the lateral border of the adductor longus

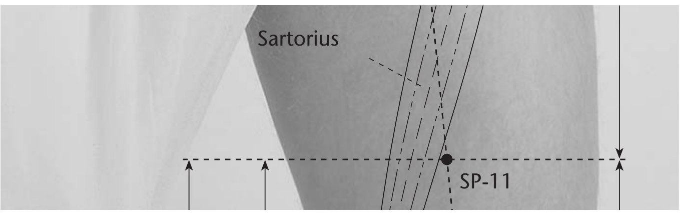
Sartorius
SP-11

SP-11 (*jimen*) 'Winnowing Gate' 6 cun proximal to **SP-10** at the midpoint of the femur, in the depression between the sartorius and vastus lateralis

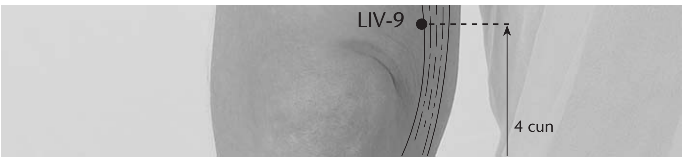

LIV-9 (*yinbao*) 'Yin Wrapping' 4 cun proximal to the medial condyle of the femur, between the sartorius and vastus medialis

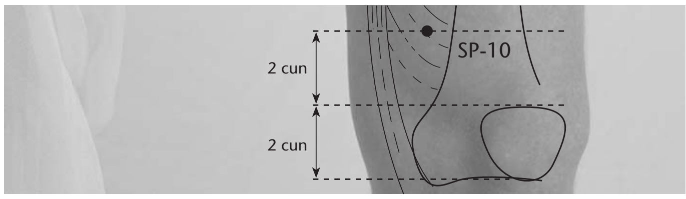
2 cun
SP-10
2 cun

SP-10 (*xuehai*) 'Sea of Blood' With the knee flexed, 2 cun proximal to the medial superior border of the patella. To find: place the right palm on the patient's left patella and locate the point at the tip of the thumb

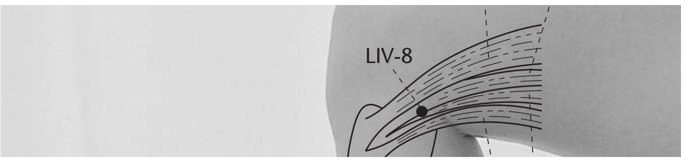

LIV-8 (*ququan*) 'Spring at the Crook' With the knee flexed, at the medial end of the popliteal crease, in the depression anterior to the tendons of semitendinosus and semimembranosus

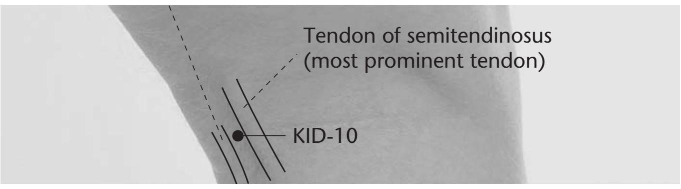
Tendon of semitendinosus (most prominent tendon)
KID-10

KID-10 (*yingu*) 'Yin Valley' At the medial end of the popliteal crease, between the tendons of semimembranosus and semitendinosus, at the level of the knee joint space

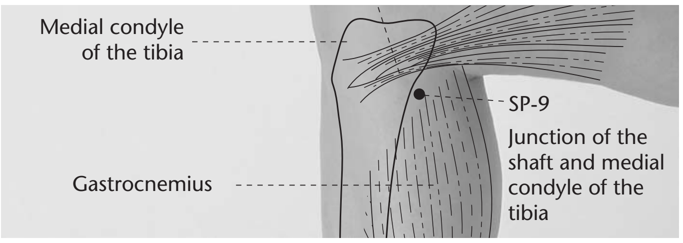
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

SP-9 (*yinlingquan*) 'Yin Mound Spring' With the knee flexed, in a depression distal to the medial condyle of the tibia. To find: palpate from the medial posterior border of the tibia towards proximal, at roughly the same level as **G.B.-34**

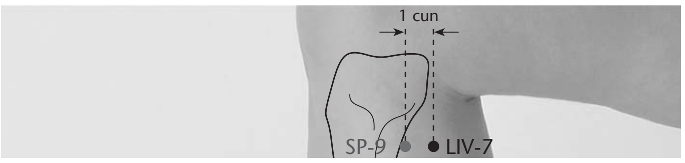
1 cun
SP-9
LIV-7

LIV-7 (*xiguan*) 'Knee Joint' At the junction of the shaft and the medial condyle of the tibia, 1 cun posterior to **SP-9**

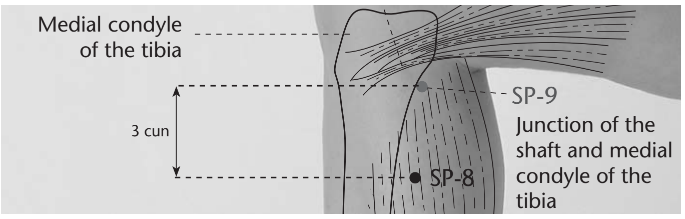
Medial condyle of the tibia
3 cun
SP-9
Junction of the shaft and medial condyle of the tibia
SP-8

SP-8 (*diji*) 'Earth Pivot' 3 cun distal to **SP-9** (at the junction of the shaft and the medial condyle of the tibia) on the posterior border of the tibia

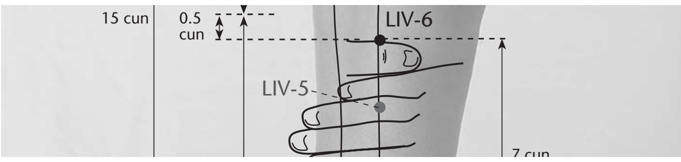
15 cun
0.5 cun
LIV-6
LIV-5
7 cun

LIV-6 (*zhongdu*) 'Central Capital' 7 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia

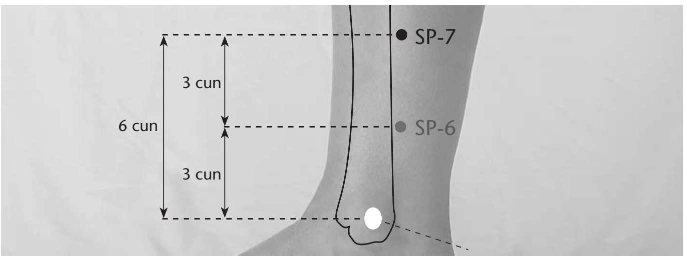
SP-7
3 cun
6 cun
SP-6
3 cun

SP-7 (*lougu*) 'Dripping Valley' 6 cun proximal to the highest prominence of the medial malleolus or 3 cun proximal to **SP-6**, on the medial posterior border of the tibia

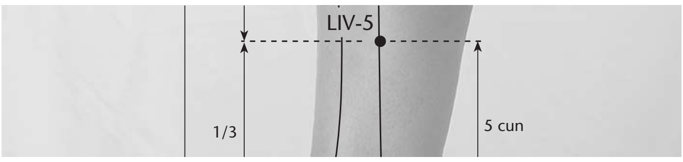
LIV-5
1/3
5 cun

LIV-5 (*ligou*) 'Woodworm Canal' 5 cun proximal to the highest prominence of the medial malleolus, just posterior to the medial crest of the tibia

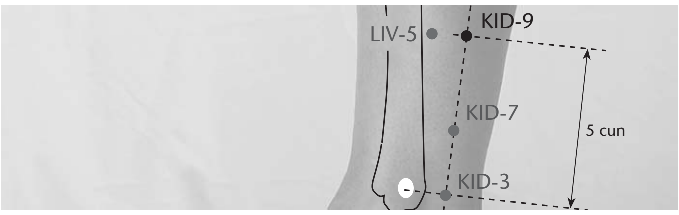
LIV-5
KID-9
KID-7
5 cun
KID-3

KID-9 (*zhubin*) 'Guest House' 5 cun proximal to KID-3 and 2 cun posterior to the medial border of the tibia

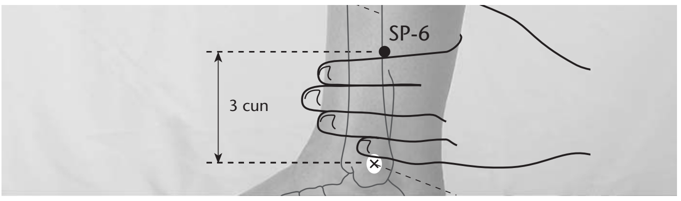

SP-6 (*sanyinjiao*) 'Three Yin Intersection' 3 cun proximal to the highest prominence of the medial malleolus, posterior border to the medial border of the tibia

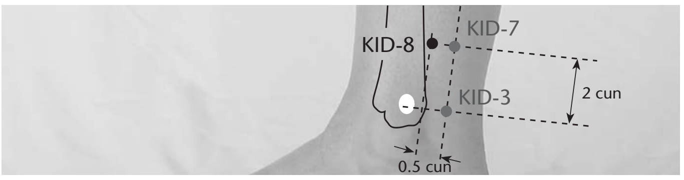
KID-8
KID-7
2 cun
KID-3
0.5 cun

KID-8 (*jiaoxin*) 'Exchange Belief' 2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia

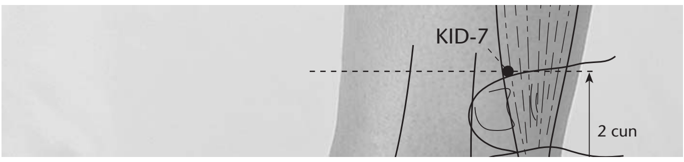

KID-7 (*fuliu*) 'Returning Current' 2 cun proximal to KID-3, in a depression on the anterior border of the Achilles tendon

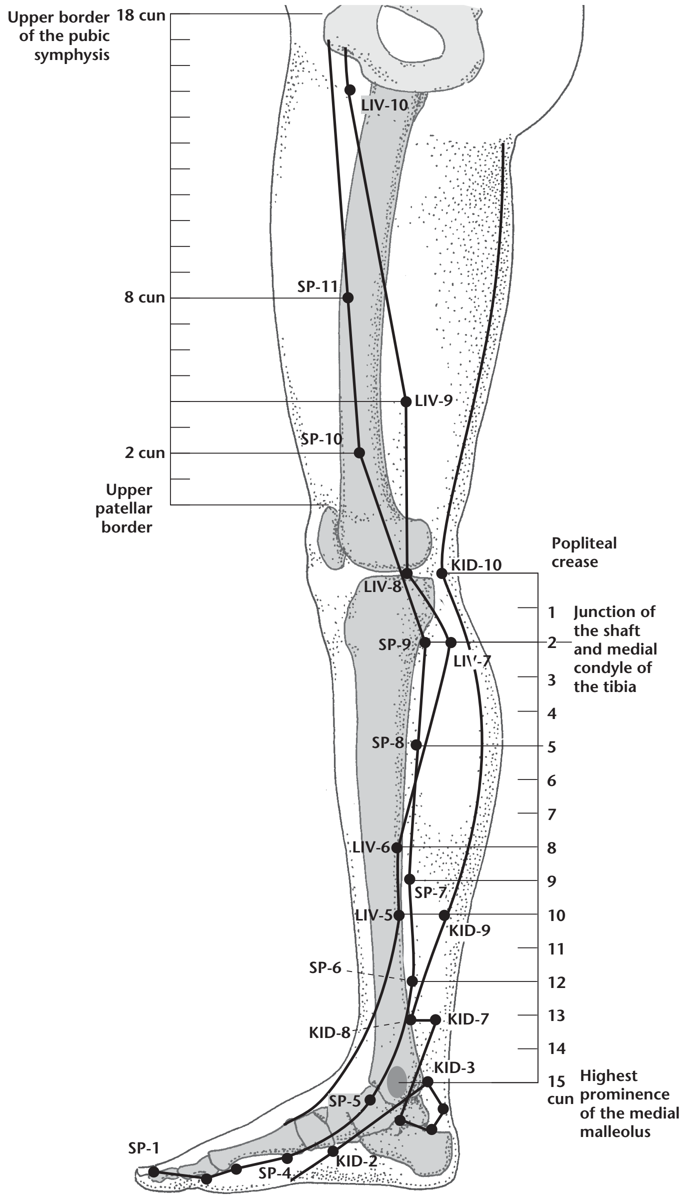
Upper border of the pubic symphysis
18 cun
LIV-10
8 cun
SP-11
LIV-9
2 cun
SP-10
Upper patellar border
Popliteal crease
KID-10
LIV-8
1
Junction of the shaft and medial condyle of the tibia
SP-9
2
LIV-7
3
4
SP-8
5
6
7
LIV-6
8
9
SP-7
LIV-5
10
KID-9
11
SP-6
12
KID-8
KID-7
13
14
KID-3
15 cun
Highest prominence of the medial malleolus
SP-5
SP-1
KID-2
SP-4

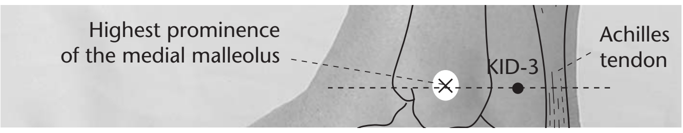
Highest prominence of the medial malleolus
KID-3
Achilles tendon

KID-3 (*taixi*) 'Supreme Stream' In the depression between the highest prominence of the medial malleolus and the Achilles tendon

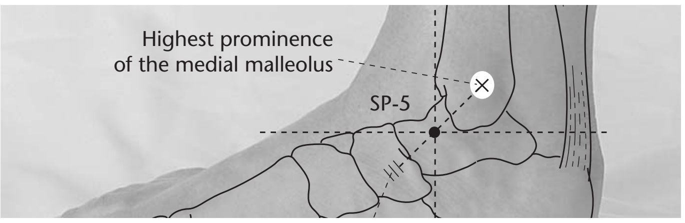
Highest prominence of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

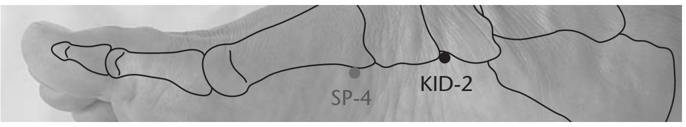

KID-2 (*rangu*) 'Blazing Valley' In a depression at the anterior inferior border of the navicular bone, at the border of the red and white skin

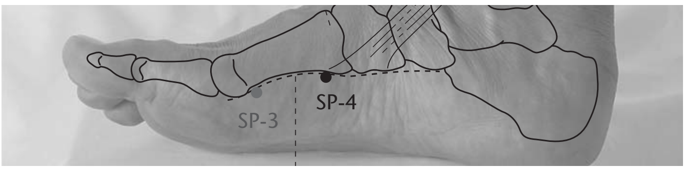

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

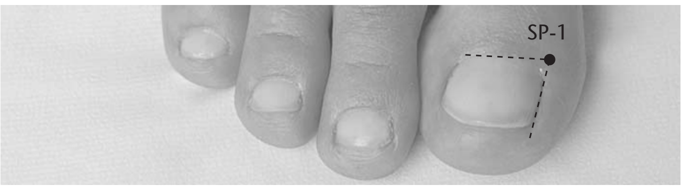

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

7.12 Lateral Leg

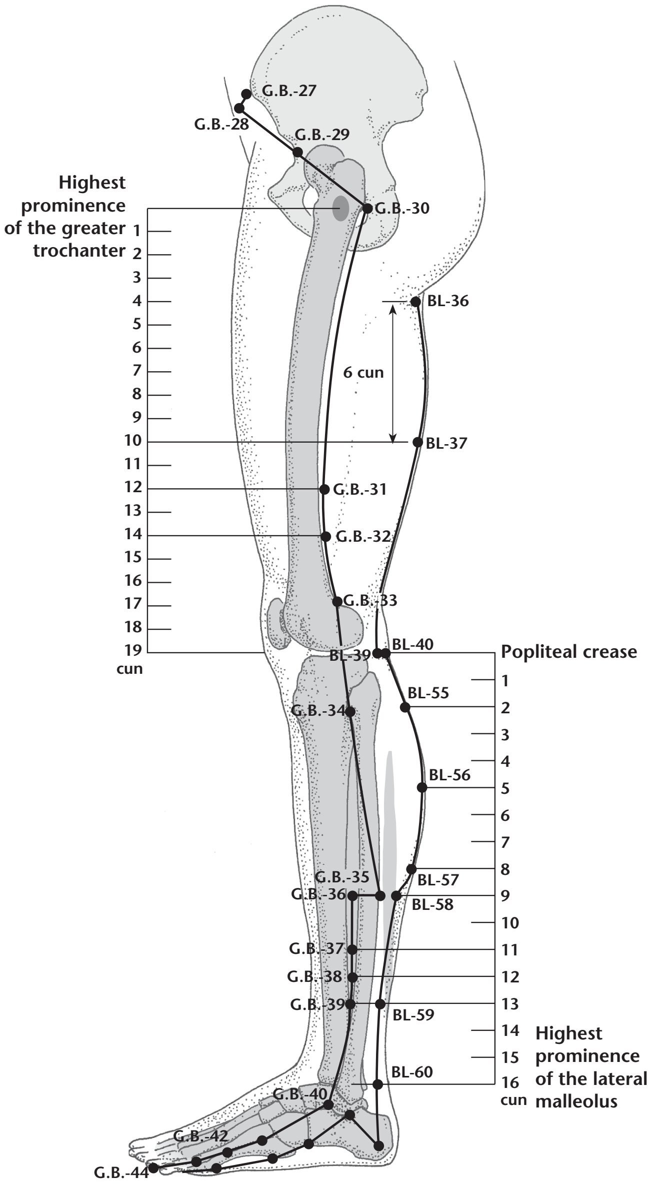
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
Popliteal crease
G.B.-34
BL-55
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
Highest prominence of the lateral malleolus
BL-60
G.B.-40
G.B.-42
G.B.-44
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun

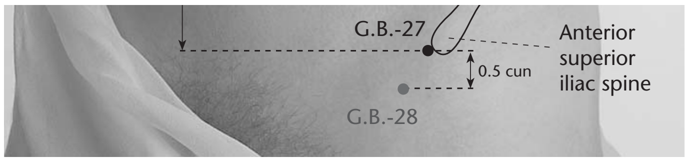
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-27 (*wushu*) 'Five Pivots' In the depression medial to the anterior superior iliac spine, approximately level with **Ren-4**

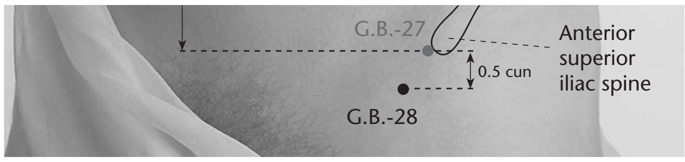
G.B.-27
Anterior superior iliac spine
0.5 cun
G.B.-28

G.B.-28 (*weidao*) 'Linking Path' Anterior and inferior to the anterior superior iliac spine, 0.5 cun anterior and inferior to **G.B.-27**

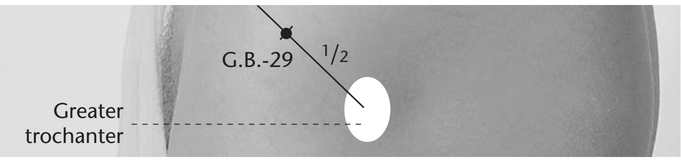
G.B.-29
1/2
Greater trochanter

G.B.-29 (*juliao*) 'Stationary Crevice' At the midpoint of a line connecting the anterior superior iliac spine and the greater trochanter, at the anterior border of the iliac crest

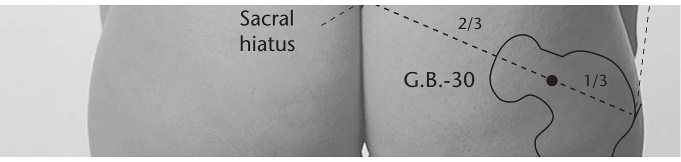
Sacral hiatus
2/3
G.B.-30
1/3

G.B.-30 (*huantiao*) 'Jumping Circle' With the patient lying on his/her side, at the junction between the medial two-thirds and the lateral third on a line connecting the greater trochanter and the sacral hiatus

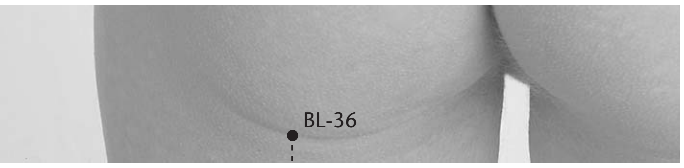

BL-36 (*chengfu*) 'Hold and Support' In the gluteal crease, superior to the midpoint of the popliteal crease

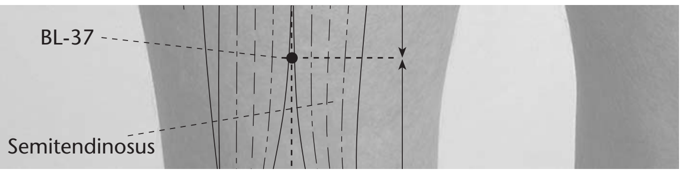
BL-37
Semitendinosus

BL-37 (*yinmen*) 'Gate of Abundance' 6 cun distal to **BL-36** (gluteal crease), on a line connecting BL-36 and BL-40 (centre of the popliteal crease), in a gap in the musculature

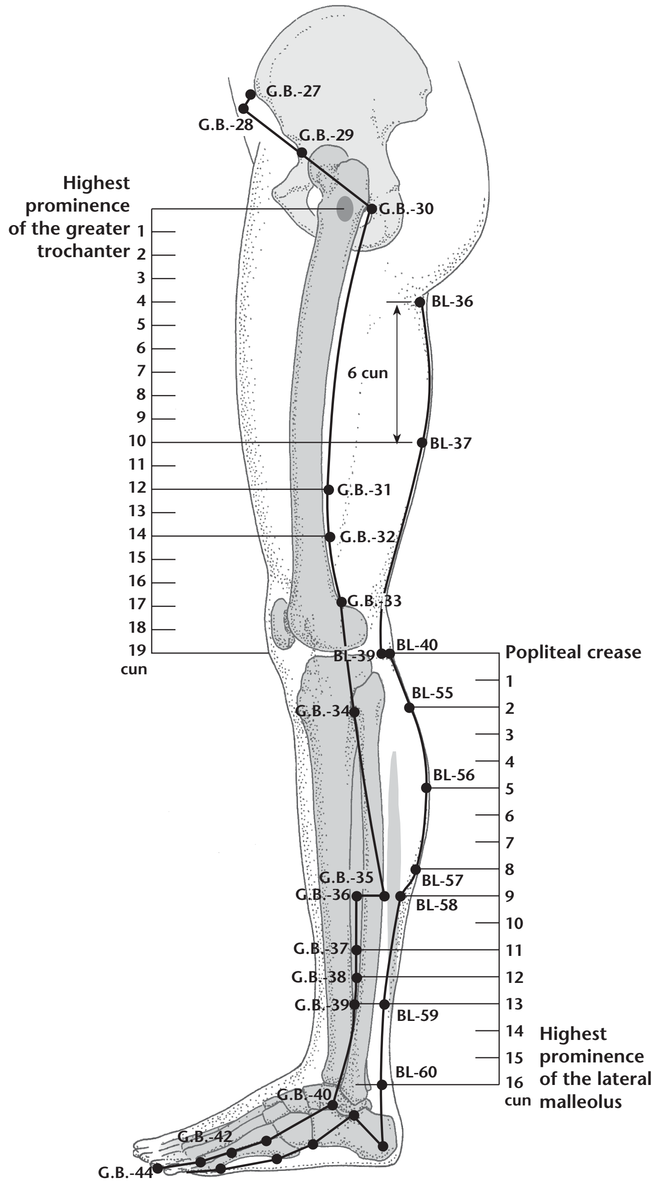
G.B.-27
G.B.-28
G.B.-29
Highest prominence of the greater trochanter
G.B.-30
BL-36
6 cun
BL-37
G.B.-31
G.B.-32
G.B.-33
BL-40
BL-39
Popliteal crease
BL-55
G.B.-34
BL-56
G.B.-35
G.B.-36
BL-57
BL-58
G.B.-37
G.B.-38
G.B.-39
BL-59
BL-60
G.B.-40
G.B.-42
G.B.-44
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
cun
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
cun
Highest prominence of the lateral malleolus

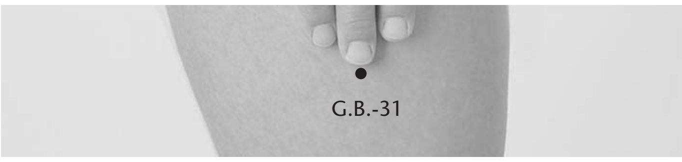

G.B.-31 (*fengshi*) 'Wind Market' On the lateral aspect of the thigh, approximately 7 cun proximal to the popliteal crease. With the patient standing upright, the point is where the patient's middle fingertip touches the thigh

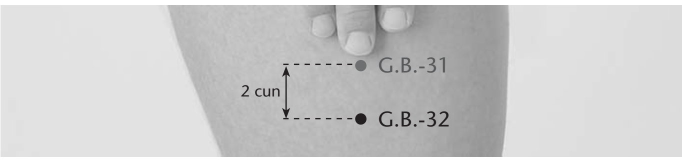
G.B.-31
2 cun
G.B.-32

G.B.-32 (*zhongdu*) 'Middle Ditch' On the lateral aspect of the thigh, 2 cun inferior to **G.B.-31** or 5 cun superior to the popliteal crease, between vastus lateralis and biceps femoris

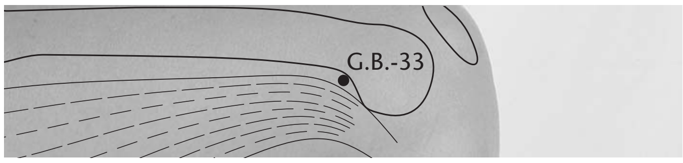

G.B.-33 (*xiyangguan*) 'Knee Yang Gate' On the lateral aspect of the knee, with the knee flexed, in the depression between the lateral epicondyle of the femur and the tendon of biceps femoris, approximately 3 cun proximal to **G.B.-34**

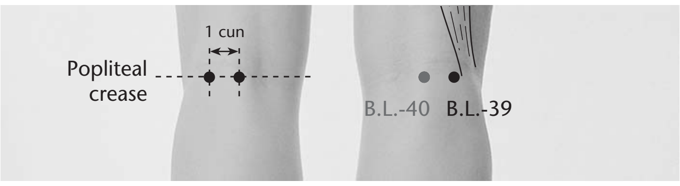
1 cun
Popliteal crease
B.L.-40
B.L.-39

BL-39 (*weiyang*) 'Outside of the Crook' At the lateral end of the popliteal crease, on the medial side of the tendon of the long head of the biceps femoris, 1 cun lateral to **BL-40** (in the centre of the popliteal crease).

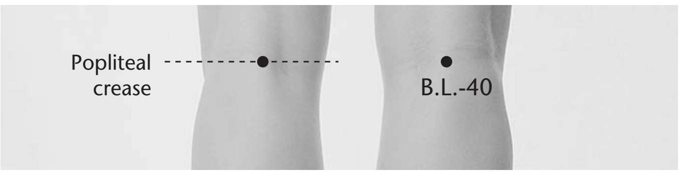
Popliteal crease
B.L.-40

BL-40 (*weizhong*) 'Middle of the Crook' In the centre of the popliteal crease, between the tendons of the biceps femoris and semitendinosus.

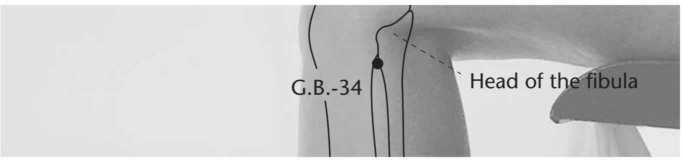
G.B.-34
Head of the fibula

G.B.-34 (*yanglingquan*) 'Yang Mound Spring' In the depression anterior and inferior to the head of the fibula

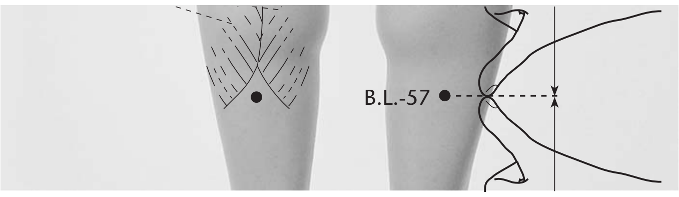

BL-57 (*chengshan*) 'Support of the Mountain' In the middle of the calf, between the two heads of the gastrocnemius, on a line connecting **BL-40** and **BL-60**, approximately 8 cun distal to **BL-40** (spreading hands technique → 2)

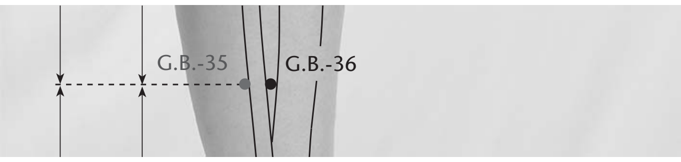

G.B.-36 (*waiqiu*) 'Outer Hill' 7 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula

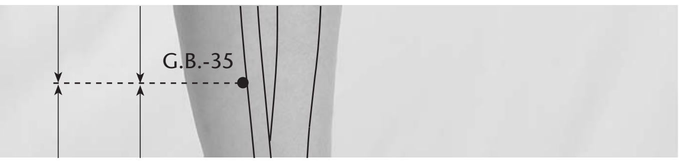

G.B.-35 (*yangjiao*) 'Yang Intersection' 7 cun proximal to the highest prominence of the lateral malleolus, on the posterior border of the fibula

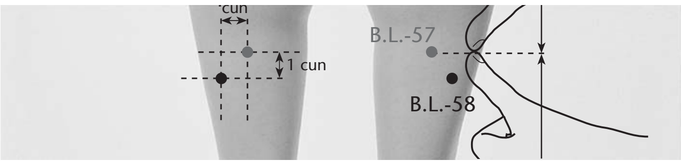
cun
1 cun
B.L.-57
B.L.-58

BL-58 (*feiyang*) 'Soaring Upwards' 1 cun distal and 1 cun lateral to **BL-57** or 7 cun proximal to **BL-60**, at the posterior border of the fibula and on the lower border of the gastrocnemius

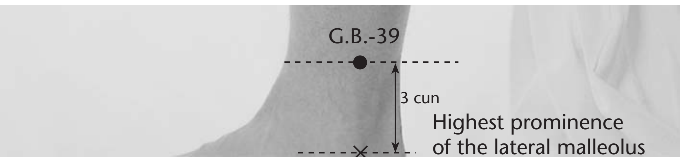
G.B.-39
3 cun
Highest prominence of the lateral malleolus

G.B.-39 (*xuanzhong*) 'Suspended Bell' 3 cun proximal to the highest prominence of the lateral malleolus, on the anterior border of the fibula

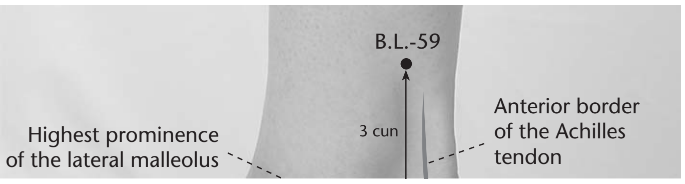
B.L.-59
3 cun
Highest prominence of the lateral malleolus
Anterior border of the Achilles tendon

BL-59 (*fuyang*) 'Instep Yang' 3 cun superior to **BL-60** (in the depression between the highest prominence of the lateral malleolus and the Achilles tendon)

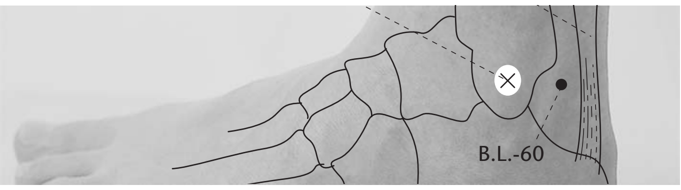

BL-60 (*kunlun*) 'Kunlun Mountains' In the depression between the Achilles tendon and the highest prominence of the lateral malleolus

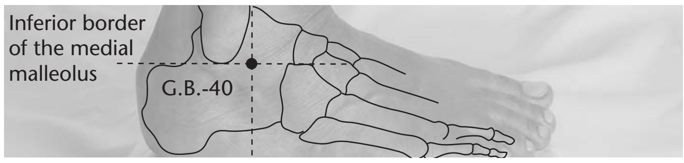
Inferior border of the medial malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of the extensor digitorum longus muscle.

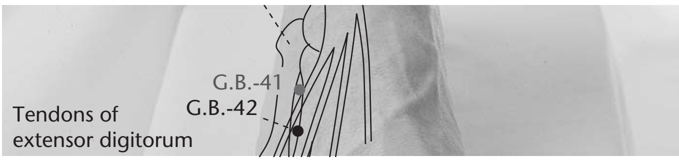
G.B.-41
G.B.-42
Tendons of extensor digitorum

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of extensor digitorum longus

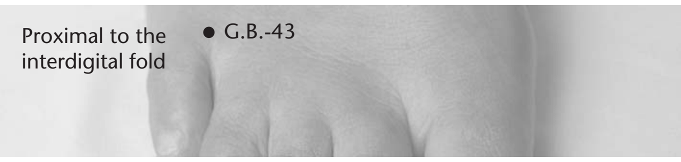
Proximal to the interdigital fold
G.B.-43

G.B.-43 (*jiaxi/xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

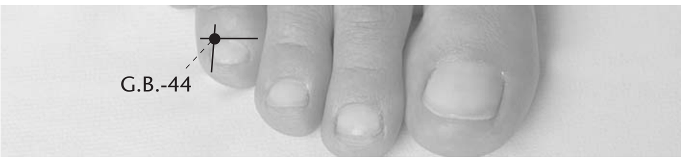

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

7.13 Dorsum of the Foot

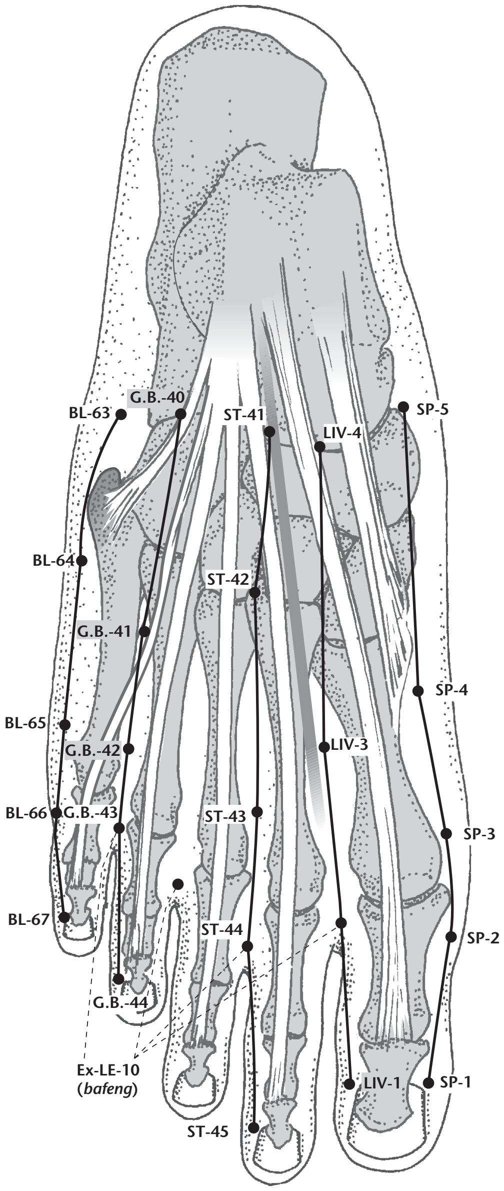
G.B.-40
BL-63
ST-41
SP-5
LIV-4
BL-64
ST-42
G.B.-41
SP-4
BL-65
LIV-3
G.B.-42
BL-66
G.B.-43
ST-43
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

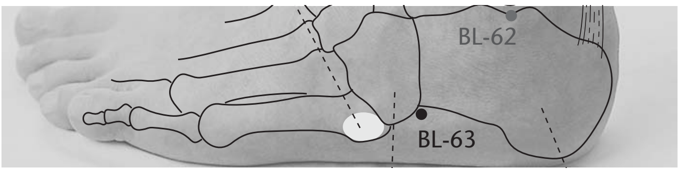

BL-63 (*jinmen*) 'Golden Gate' In a depression anterior and inferior to **BL-62** between the calcaneus and the cuboid bone, proximal to the tuberosity of the 5th metatarsal bone

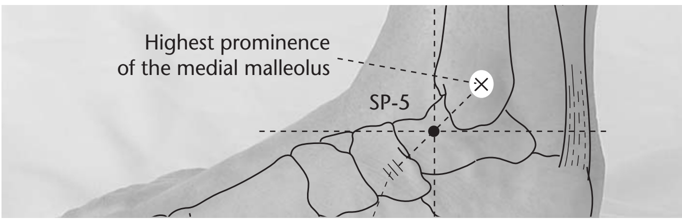
Highest prominence
of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

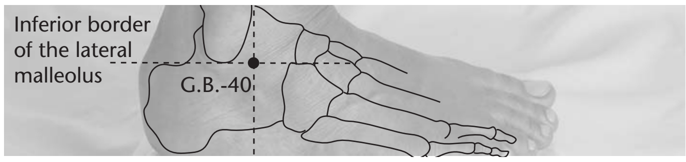
Inferior border
of the lateral
malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the lateral malleolus, lateral to the tendons of extensor digitorum longus muscle.

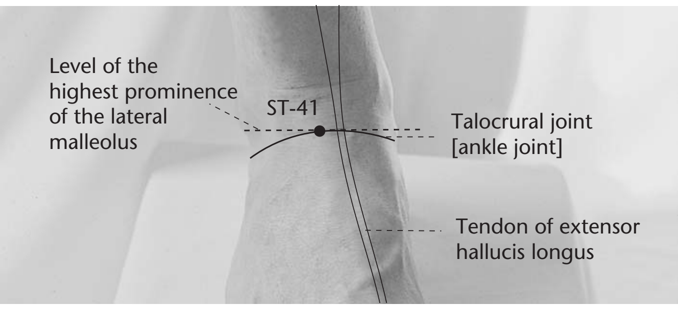
Level of the
highest prominence
of the lateral
malleolus
ST-41
Talocrural joint
[ankle joint]
Tendon of extensor
hallucis longus

ST-41 (*jiexi*) 'Stream Divide' On the ankle, between the tendons of extensor digitorum and extensor hallucis longus, at the level of the highest prominence of the lateral malleolus

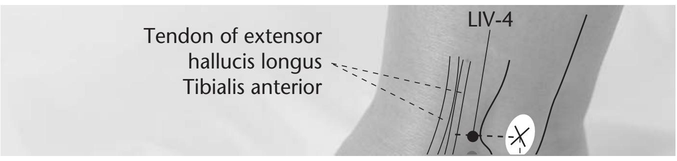
Tendon of extensor
hallucis longus
Tibialis anterior
LIV-4

LIV-4 (*zhongfeng*) 'Middle Seal' 1 cun anterior to the prominence of the medial malleolus, medial to the tendon of the tibialis anterior

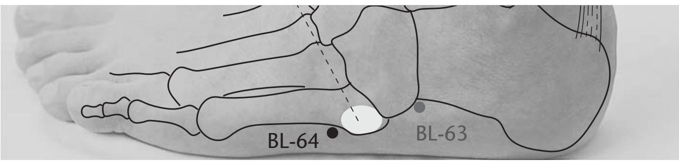

BL-64 (*jinggu*) 'Capital Bone' Distal to the tuberosity of the 5th metatarsal bone, at the border of the red and white skin

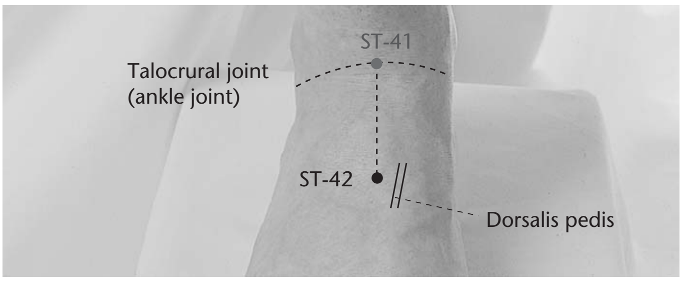
ST-41
Talocrural joint
(ankle joint)
ST-42
Dorsalis pedis

ST-42 (*chongyang*) 'Rushing Yang' On the highest point of the dorsum of the foot, between the tendons of extensor hallucis longus and extensor digitorum longus, directly lateral to where the dorsalis pedis artery may be palpated. The point is bordered proximally by the second and third metatarsal bones and distally by the second and third cuneiform bones. **Alternative location:** sometimes this point may be located lateral to the medial portion of the extensor digitorum longus tendon (joining the second toe)

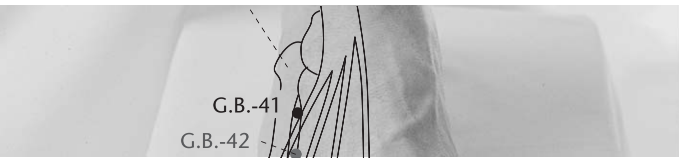

G.B.-41 (*zulinqi*) 'Foot Governor of Tears' In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the branch of the extensor digitorum longus tendon running to the little toe

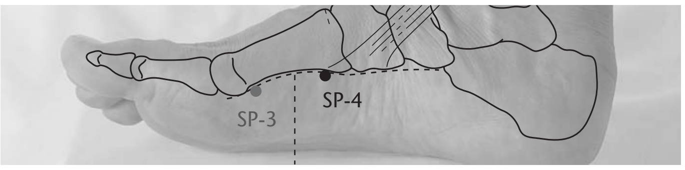

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

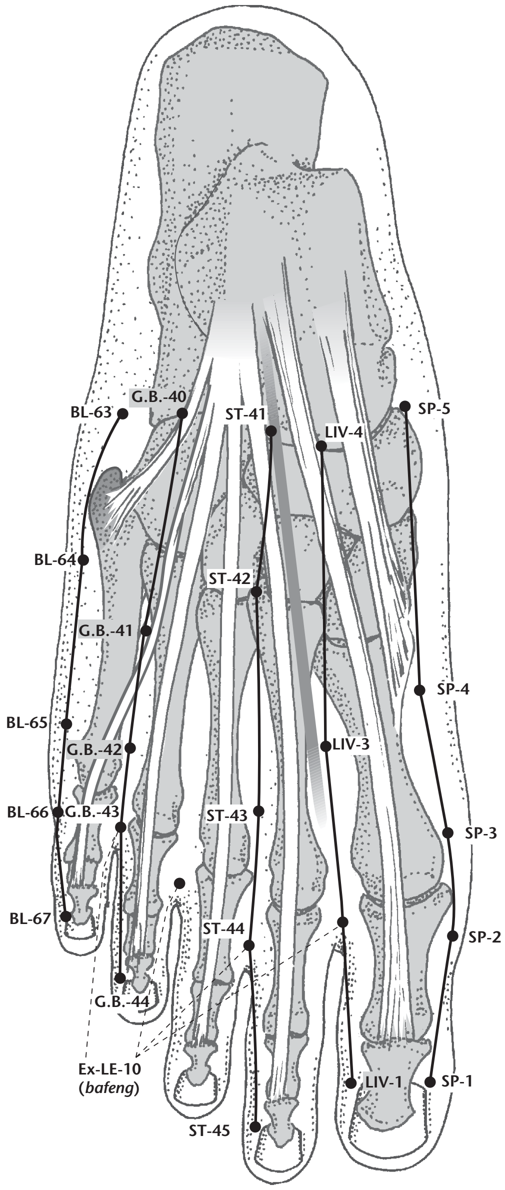
BL-63
G.B.-40
ST-41
LIV-4
SP-5
BL-64
ST-42
G.B.-41
SP-4
BL-65
G.B.-42
LIV-3
BL-66
G.B.-43
ST-43
SP-3
BL-67
ST-44
SP-2
G.B.-44
Ex-LE-10
(*bafeng*)
LIV-1
SP-1
ST-45

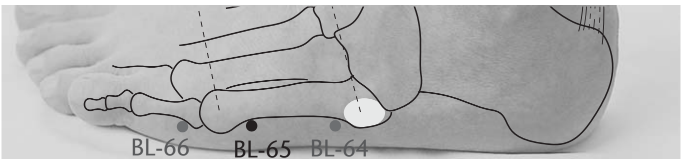
BL-66
BL-65
BL-64

BL-65 (*shugu*) 'Restraining Bone' In the depression proximal to the head of the 5th metatarsal bone, at the border of the red and white skin

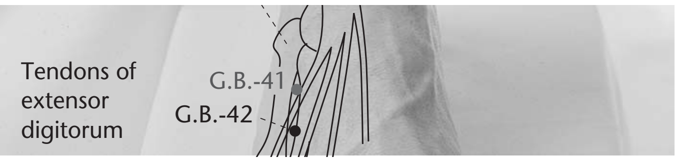
Tendons of extensor digitorum
G.B.-41
G.B.-42

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of the extensor digitorum longus

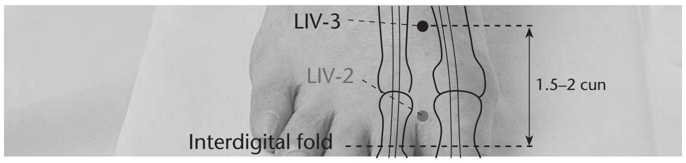
LIV-3
LIV-2
1.5–2 cun
Interdigital fold

LIV-3 (*taichong*) 'Great Rushing' On the dorsum of the foot, in the depression between the 1st and 2nd metatarsal bones, approximately 1.5–2 cun proximal to the interdigital web

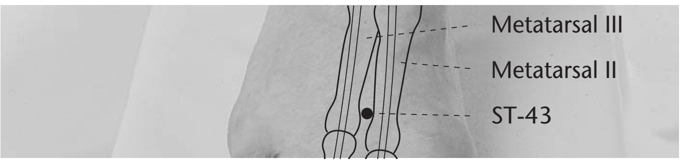
Metatarsal III
Metatarsal II
ST-43

ST-43 (*xiangu*) 'Sunken Valley' In the depression between the second and third metatarsal bones, at the junction of the shafts and the heads of these metatarsal bones

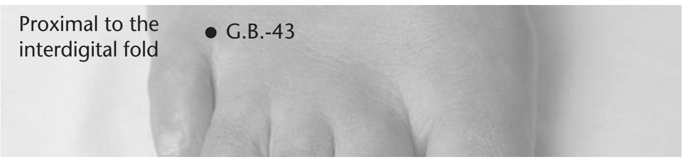
Proximal to the interdigital fold
G.B.-43

G.B.-43 (*xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

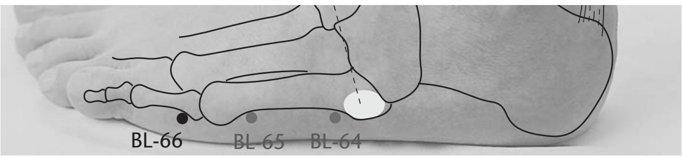
BL-66
BL-65
BL-64

BL-66 (*zutonggu*) 'Foot Connecting Valley' In a depression distal to the metatarsophalangeal joint of the little toe

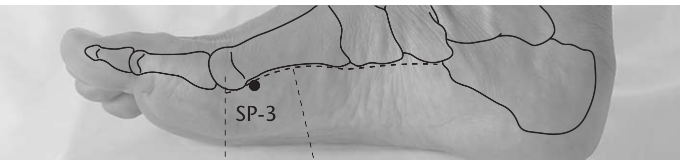

SP-3 (*taibai*) 'Supreme White' In the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin

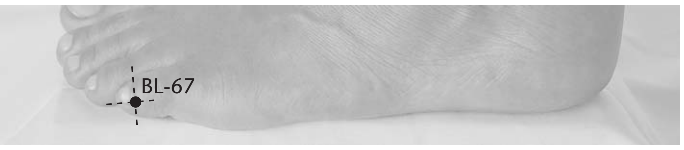

BL-67 (*zhiyin*) 'Reaching Yin' On the little toe, 0.1 cun from the lateral corner of the nail

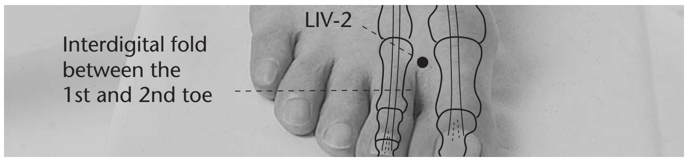

LIV-2 (*xingjian*) 'Moving Between' Between the 1st and 2nd toe, proximal to the margin of the interdigital web

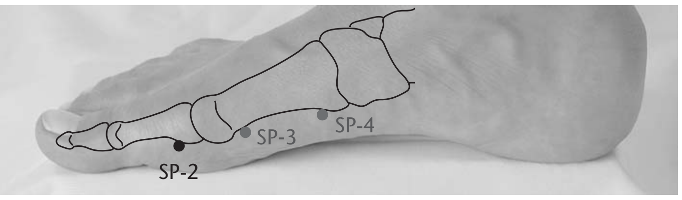

SP-2 (*dadu*) 'Great Metropolis' On the medial aspect of the big toe, in the depression distal to the first metatarsophalangeal joint, on the border of the red and white skin

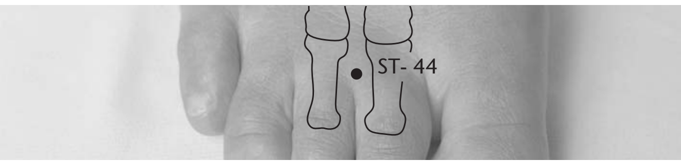

ST-44 (*neiting*) 'Inner Court' Between the 2nd and the 3rd toe, proximal to the interdigital fold

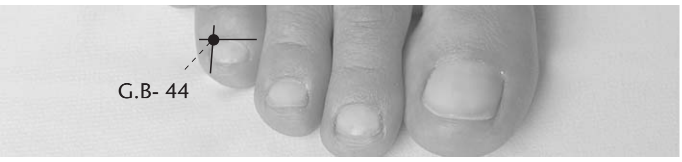

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

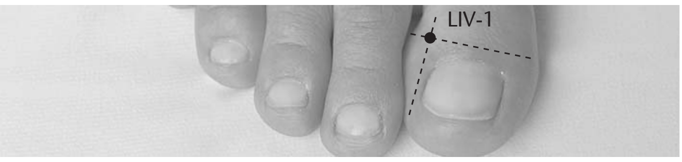

LIV-1 (*dadun*) 'Big Mound' On the lateral aspect of the big toe, 0.1 cun proximal and lateral to the corner of the nail

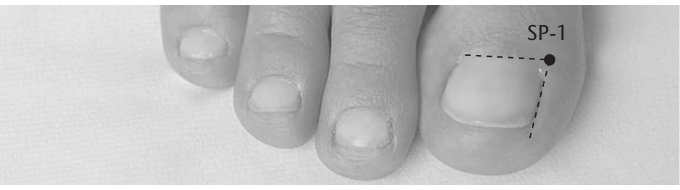

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

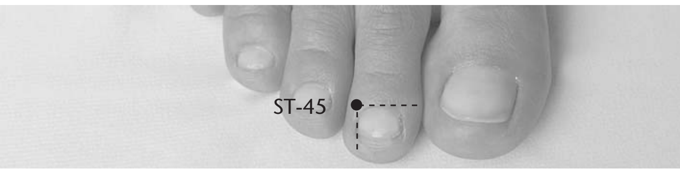

ST-45 (*lidui*) 'Strict Exchange' On the 2nd toe, 0.1 cun from the lateral corner of the nail

7.14 Medial Foot

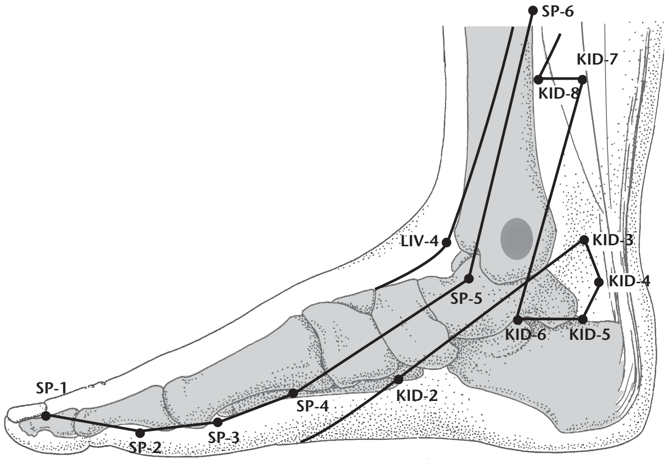
SP-6
KID-7
KID-8
LIV-4
KID-3
KID-4
SP-5
KID-6
KID-5
SP-1
KID-2
SP-4
SP-3
SP-2

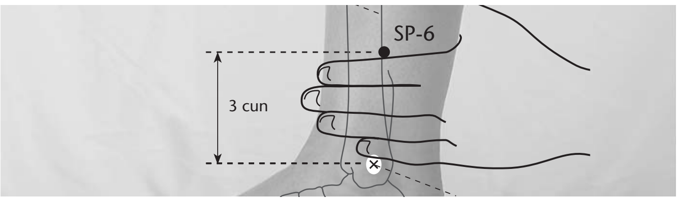

SP-6 (*sanyinjiao*) 'Three Yin Intersection' 3 cun proximal to the highest prominence of the medial malleolus, on the posterior border of the medial crest of the tibia

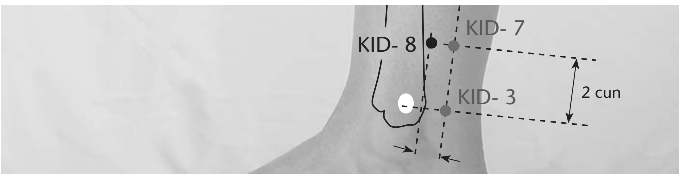
KID- 8
KID- 7
KID- 3
2 cun

KID-8 (*jiaoxin*) 'Exchange Belief' 2 cun proximal to the highest prominence of the medial malleolus, posterior to the medial border of the tibia.

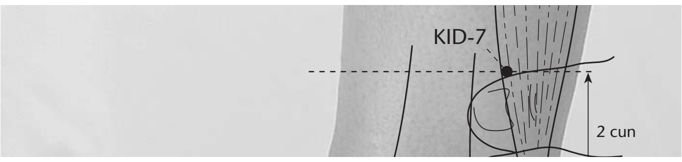

KID-7 (*fuliu*) 'Returning Current' 2 cun proximal to **KID-3**, in a depression on the anterior border of the Achilles tendon

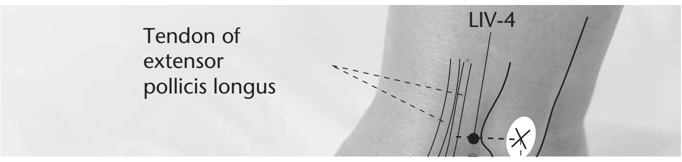
Tendon of
extensor
pollicis longus
LIV-4

LIV-4 (*zhongfeng*) 'Middle Seal' 1 cun anterior to the prominence of the medial malleolus, medial to the tendon of tibialis anterior

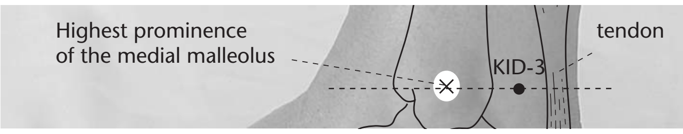
Highest prominence
of the medial malleolus
tendon
KID-3

KID-3 (*taixi*) 'Supreme Stream' In the depression between the highest prominence of the medial malleolus and the Achilles tendon

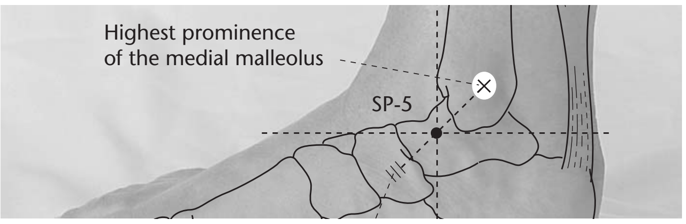
Highest prominence
of the medial malleolus
SP-5

SP-5 (*shangqiu*) 'Shang Mound' In the depression at the junction of a vertical line along the anterior border and a horizontal line along the lower border of the medial malleolus. Or: in the depression halfway between the highest prominence of the medial malleolus and the navicular tuberosity

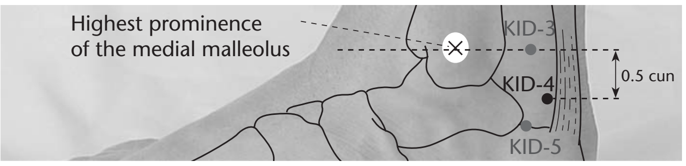
Highest prominence
of the medial malleolus
KID-3
KID-4
0.5 cun
KID-5

KID-4 (*dazhong*) 'Great Bell' Anterior to the medial border of the Achilles tendon, superior to its insertion at the calcaneus

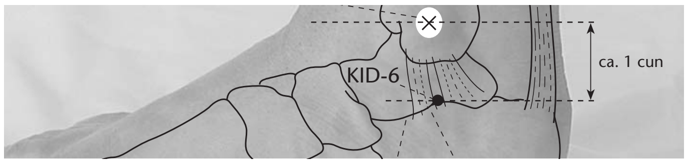
KID-6
ca. 1 cun

KID-6 (*zhaohai*) 'Shining Sea' In the depression inferior to the highest prominence of the medial malleolus, over the joint space between the talus and the calcaneus

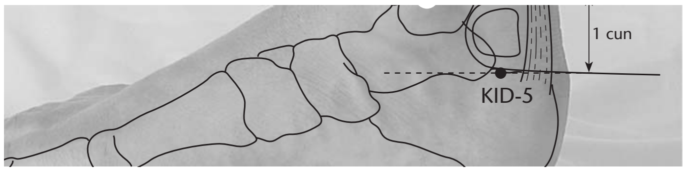

KID-5 (*shuiquan*) 'Water Spring' 1 cun distal to **KID-3**, in a depression over the joint space between the talus and calcaneus

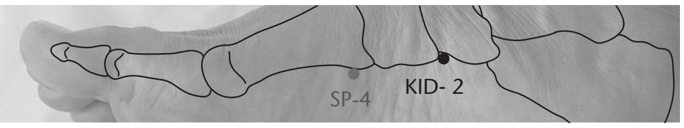

KID-2 (*rangu*) 'Blazing Valley' In a depression at the anterior inferior border of the navicular bone, at the border of the red and white skin

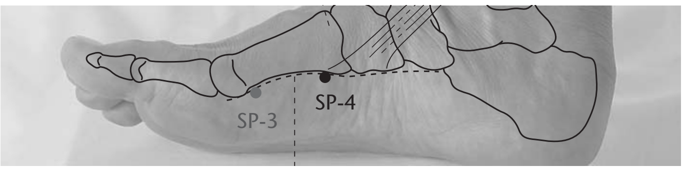

SP-4 (*gongsun*) 'Grandfather Grandson' In a depression distal and inferior to the base of the 1st metatarsal bone, at the border of the red and white skin

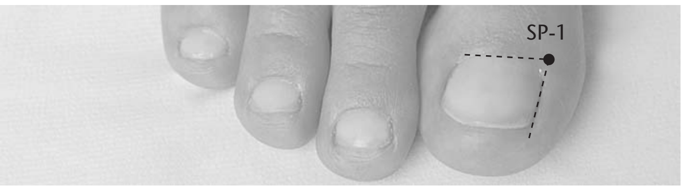

SP-1 (*yinbai*) 'Hidden White' On the big toe, 0.1 cun from the medial corner of the nail

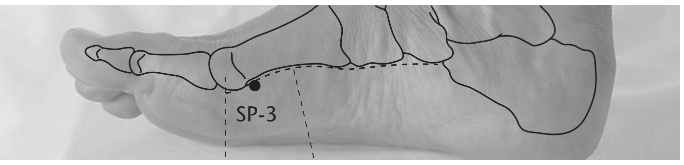

SP-3 (*taibai*) 'Supreme White' In the depression proximal to the head of the 1st metatarsal bone, at the border of the red and white skin

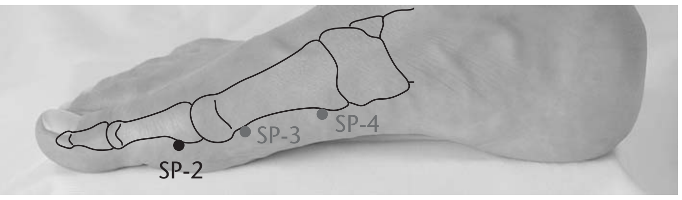

SP-2 (*dadu*) 'Great Metropolis' On the medial aspect of the big toe, in the depression distal to the first metatarsophalangeal joint, on the border of the red and white skin

7.15 Lateral Foot

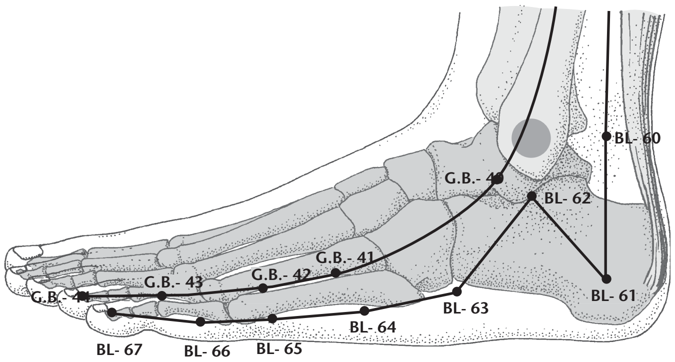
BL- 60
G.B.- 40
BL- 62
G.B.- 41
G.B.- 42
G.B.- 43
G.B.- 44
BL- 61
BL- 63
BL- 64
BL- 67
BL- 66
BL- 65

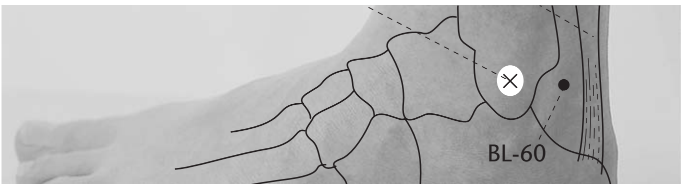

BL-60 (*kunlun*) 'Kunlun Mountains' In the depression between the Achilles tendon and the highest prominence of the lateral malleolus

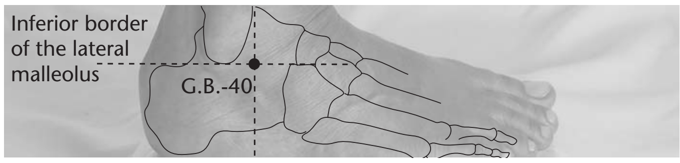
Inferior border
of the lateral
malleolus
G.B.-40

G.B.-40 (*qiuxu*) 'Large Mound' In the depression anterior and inferior to the medial malleolus, lateral to the tendons of extensor digitorum longus muscle.

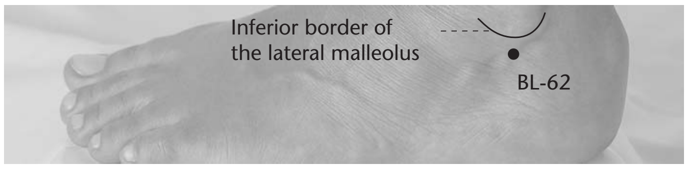
Inferior border of
the lateral malleolus
BL-62

BL-62 (*shenmai*) 'Extending Vessel' In a depression directly inferior to the highest prominence of the lateral malleolus, over the joint space between the talus and calcaneus

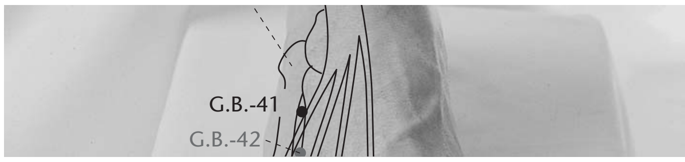

G.B.-41 (*zulinqi*) 'Foot Governor of Tears' In the depression at the junction of the shafts and the bases of the 4th and 5th metatarsal bones, lateral to the branch of the extensor digitorum longus tendon running to the little toe

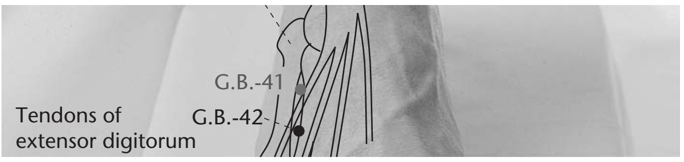
G.B.-41
Tendons of
extensor digitorum
G.B.-42

G.B.-42 (*diwuhui*) 'Earth Five Meetings' Between the 4th and 5th metatarsal bones, proximal to the metatarsophalangeal joints and medial to the tendon of extensor digitorum longus

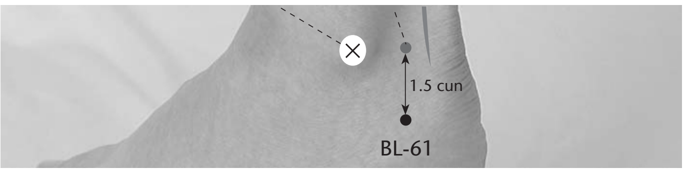

BL-61 (*pucan*) 'Servant's Respect' In a depression on the calcaneus, approximately 1.5 cun inferior to **BL-60**

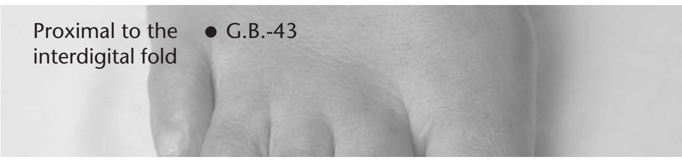
Proximal to the
interdigital fold
G.B.-43

G.B.-43 (*xiaxi*) 'Clamped Stream' Between the 4th and the 5th toes, proximal to the margin of the interdigital web

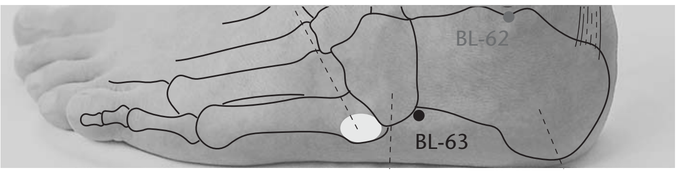

BL-63 (*jinmen*) 'Golden Gate' On the lateral aspect of the foot, proximal to the tuberosity of the 5th metatarsal bone, in a depression anterior and inferior to **BL-62** between the calcaneus and the cuboid bone

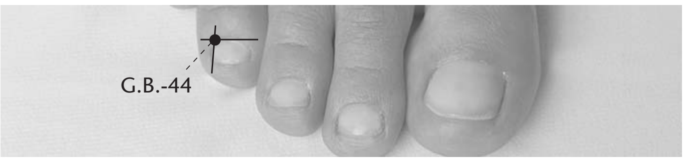

G.B.-44 (*zuqiaoyin*) 'Yin Portals of the Foot' On the 4th toe, 0.1 cun from the lateral corner of the nail

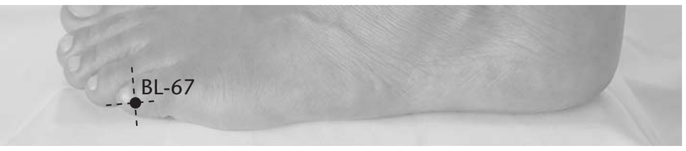

BL-67 (*zhiyin*) 'Reaching Yin' On the little toe, 0.1 cun from the lateral corner of the nail

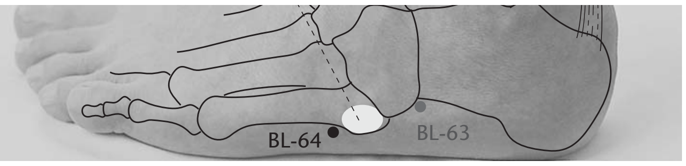

BL-64 (*jinggu*) 'Capital Bone' Distal to the tuberosity of the 5th metatarsal bone, at the border of the red and white skin

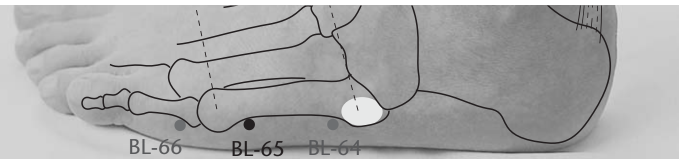

BL-65 (*shugu*) 'Restraining Bone' In the depression proximal to the head of the 5th metatarsal bone, at the border of the red and white skin

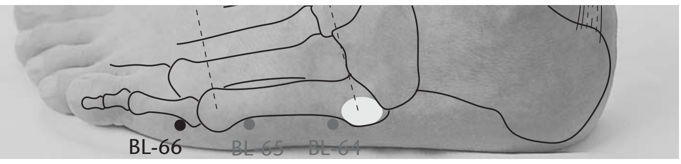

BL-66 (*zutonggu*) 'Foot Connecting Valley' In a depression distal to the metatarsophalangeal joint of the little toe

8 Point Categories and Point Selection

Claudia Focks

There are a total of 361 acupuncture points located on the channels (→ Chapter 1). Translated literally from the Chinese language, the meaning of the word for 'point' is rather 'opening' or 'hole'. Stimulation of the points can be achieved through massage (in acupressure) or through the insertion of needles (in acupuncture). Besides the regular acupuncture points, there exist a wide variety of so-called extra points, located outside the pathways of the channels. The most commonly used of these extra points are described in Chapter 6, as classified by the WHO.

The acupuncture points are effective

- **locally** (for example, points on the elbow treat disorders of the elbow)
- **regionally** (the channel pathway – for example, points on the arm treat disorders of the upper extremity)
- some points have an effect on their **pertaining *zangfu***-Organ
- some points have **additional actions** such as calming the *shen* or eliminating Heat.

The following table provides an overview for the general indications of the channel points (modified after Ellis and Wiseman 1991)

Indications of the channel points

Channel	Indications		
Hand-Yin channels			
LU	Lungs, throat		Disorders of the thoracic region
P	Heart, Stomach	Spirit (*shen*) disorders	
HE	Heart		
Hand-Yang channels			
L.I.	Face, mouth, occiput, teeth, nose		Eyes, disorders of the throat, febrile diseases
T.B.	Head/lateral costal region	Disorders of the ears	
S.I.	Occipital region, shoulder, scapula, *shen* disorders		
Foot-Yin channels			
SP	Spleen, Stomach, intestines		Urogenital disorders, menstrual disorders, vaginal discharge, spirit (*shen*) disorders
LIV	Liver, genitalia		
KID	Kidneys, Lungs, throat		
Foot-Yang channels			
ST	Frontal aspect of the head, face, mouth, teeth, throat, Stomach, intestines		Psychoemotional disorders (*shen* disorders), febrile colds
G.B.	Lateral aspect of the head, ears, lateral costal region	Disorders of the eyes	
BL	Posterior aspect of the head, lower back		

8.1 Point Categories

Certain acupuncture points or groups of points are categorised according to their actions or location on their respective channels. (For an overview of these points and point groups → inside back cover.)

8.1.1 *Yuan*-Source Points

At the *yuan*-source points, the Original *Qi* (*yuan qi* → 1.1.4) of the respective Organ system emerges to the surface of the body.
Location: near the hand or ankle.

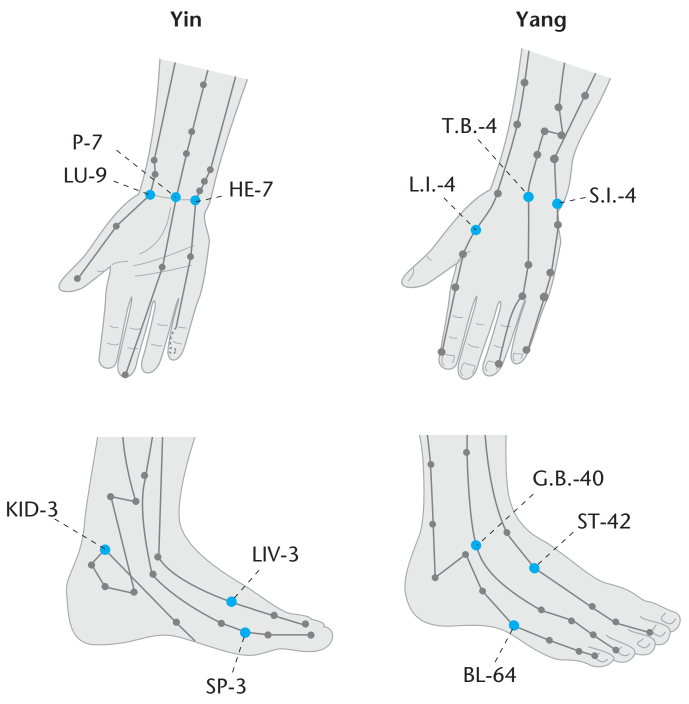

Points

Yuan-source points of the Yin channels

LU-9 (*taiyuan*)	**SP-3** (*taibai*)	**HE-7** (*shenmen*)
KID-3 (*taixi*)	**P-7** (*daling*)	**LIV-3** (*taichong*)

- On the Yin channels the *yuan*-source point is always the **3rd point** from the distal end of the channel
- They are identical with the *shu*-stream points and the Earth point of the channel

Function

- They tonify deficiency or weakness of their pertaining *zang*-Organ
- They regulate the Yin/Yang balance, restoring homeostasis
- As Earth points they have a 'grounding' effect on the body, the emotions and the mind.

Yuan-source points of the Yang channels

L.I.-4 (*hegu*)	**ST-42** (*chongyang*)	**S.I.-4** (*wangu*)
BL-64 (*jinggu*)	**T.B.-4** (*yangchi*)	**G.B.-40** (*qiuxu*)

- On the Yang channels the *yuan*-source point is always the **4th point** from the distal end of the channel, except for the G.B. channel (5th point)
- Regarding the energetics of the channel, they are all located between the *shu*-stream point and the *jing*-river point.

Function

- Their main function is to expel pathogenic factors in excess syndromes
- They tonify their pertaining *fu*-Organ.

Clinical application

Diagnosis

Disorders of a particular channel or its pertaining *zangfu*-Organ will often manifest themselves with tenderness, skin changes such as discolorations, swellings, redness, etc. in the area of the respective *yuan*-source point.

Therapy

- For deficiency of the relevant channel or Organ, use tonifying needling techniques; for excess, use reducing techniques
- Combine with the *luo*-connecting point of its Interiorly–Exteriorly paired channel (→ 1.2.2) in order to increase the therapeutic effect and to balance Yin and Yang (→ 8.3.3)
- Combine with the respective Back-*shu* point (→ 8.1.4) for disorders of the *zangfu*-Organs

8.1.2 *Luo*-Connecting Points

The *luo*-connecting points are located in those regions of the body where the *luo*-network vessels (→ 1.5) branch out, connecting the Interiorly–Exteriorly paired Yin and Yang channels (→ 1.2.2).

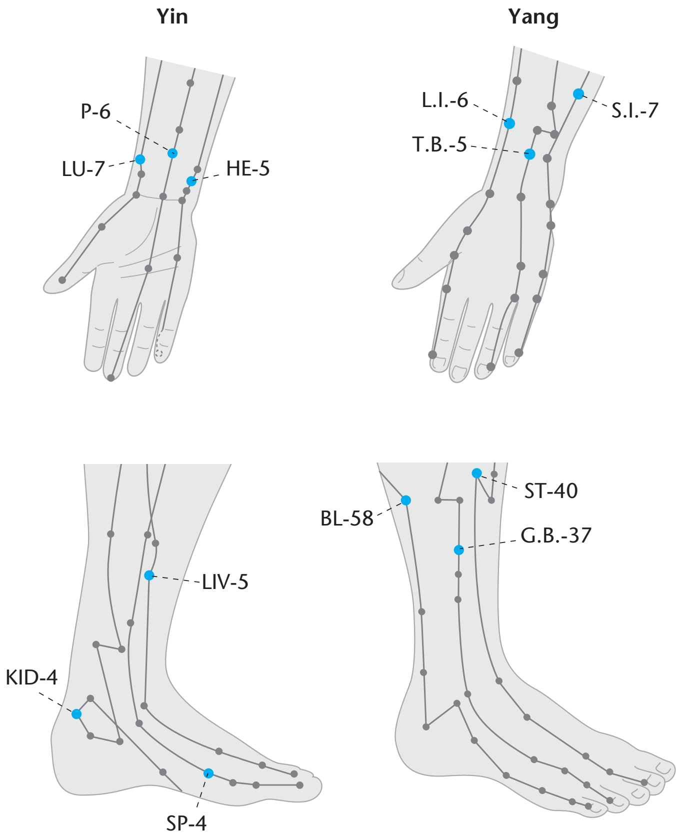

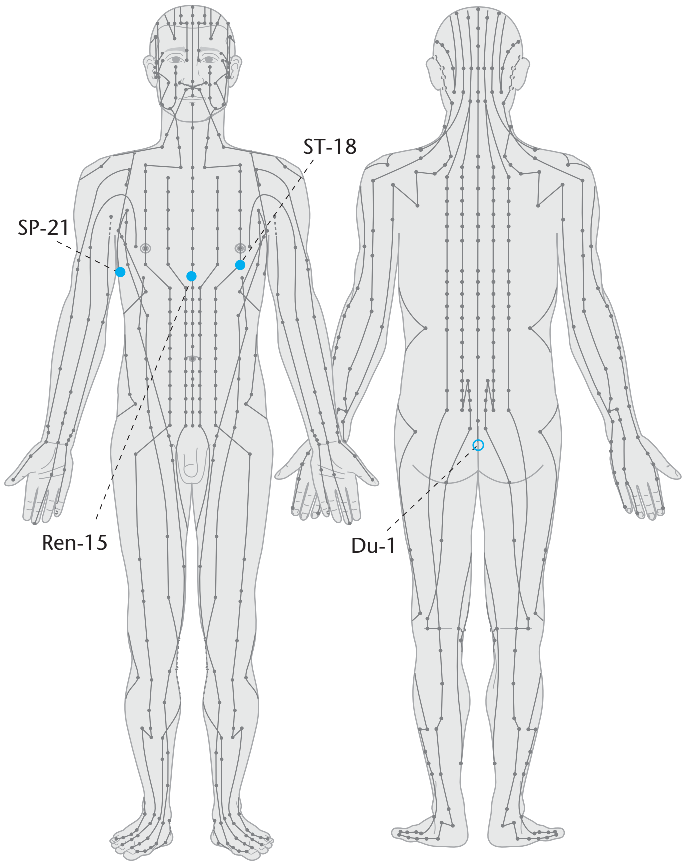

Points

The 12 *luo*-connecting points of the primary channels are located between the fingers and the elbow on the upper extremity and between the toes and the knee on the lower extremity.

- *Luo*-connecting points of the primary Yin channels
  - hand Yin channels: **LU-7** (*lieque*), **P-6** (*neiguan*), **HE-5** (*tongli*)
  - foot Yin channels: **SP-4** (*gongsun*), **LIV-5** (*ligou*), **KID-4** (*dazhong*)
- *Luo*-connecting points of the primary Yang channels
  - hand Yang channels: **L.I.-6** (*pianli*), **T.B.-5** (*waiguan*), **S.I.-7** (*zhizheng*)
  - foot Yang channels: **ST-40** (*fenglong*), **G.B.-37** (*guangming*), **BL-58** (*feiyang*)
- *Luo*-connecting points of the *ren mai:* **Ren-15** (*jiuwei*); action: regulates the *luo*-connecting vessels of the Yin channels; indication: abdominal diseases and disorders
- *Luo*-connecting points of the *du mai:* **Du-1** (*changqiang*); action: regulates the *luo*-connecting vessels of the Yang channels; indication: disorders of the back
- Great *luo*-connecting point of the Spleen: **SP-21** (*dabao*); action: connects with all *luo*-vessels and governs all *luo* and Blood vessels of the body; indication: disorders of the joint, general weakness and pain
- Great *luo*-connecting point of the Stomach: ***xu li****, near the left **ST-18**; moxibustion is contraindicated, only superficial needling is recommended. Indication: breathing and heart disorders such as asthma, dyspnoea, emphysema and bradycardia, also for shortness of breath due to anxiety attacks.

Clinical application

Diagnosis

- In excess syndromes the *luo* vessels often become visible near the *luo*-connecting points, for example skin discolorations and swellings may occur. With pain due to Cold the predominant colour may be bluish to greenish, with Heat syndromes reddish, with chronic stagnation and stasis dark to dark purple.
- In deficiency syndromes there are often no colour changes. However, in severe chronic cases (for example severe Qi deficiency) there will more likely be distinct muscular flaccidity, presenting as a dip or depression in the area of the point.

Therapy

- For disorders of the pertaining *zangfu*-Organ or channel
- For diseases of the Interiorly–Exteriorly paired Organ or channel
- For disorders due to the accumulation of pathogenic factors or substances, for example Qi stagnation, Blood stasis, retention of Dampness or Phlegm
- For psychoemotional problems, especially *luo*-connecting points of the Yin channels, for example classically **P-6** (*neiguan*) and **HE-5** (*tongli*)
- Bloodletting (three-edged needle, lancet or plum-blossom needle) on the *luo*-connecting point of the affected channel: this is particularly indicated for excess syndromes with Qi stagnation and Blood stasis. For more details on point selection ➞ 8.3.3

8.1.3 *Xi*-Cleft Points

Synonyms: Accumulation points (Maciocia)
The *xi*-cleft points are all located between the fingers and the elbow on the upper extremity and between the toes and the knee on the lower extremity. Only **ST-34** is located more proximally on the thigh.

---

* *Xu li* is only rarely mentioned in Western literature. The location and significance of this point is based on *Su Wen*, chapter 18: 'The geat *luo* vessel of the Stomach is called *xu li* and emerges at **ST-18** under the left breast, passes through the diaphragm and ascends to connect with the Lung. One can feel its pulsation with the hand. It is the place where the *zong qi* (➞ 1.1.4) of the channels resides.' (Ngyen Van Nghi, 1989). *Xu li* is the only *luo*-connecting vessel to directly emerge from a *fu*-Organ. It plays an important role in Korean acupuncture.

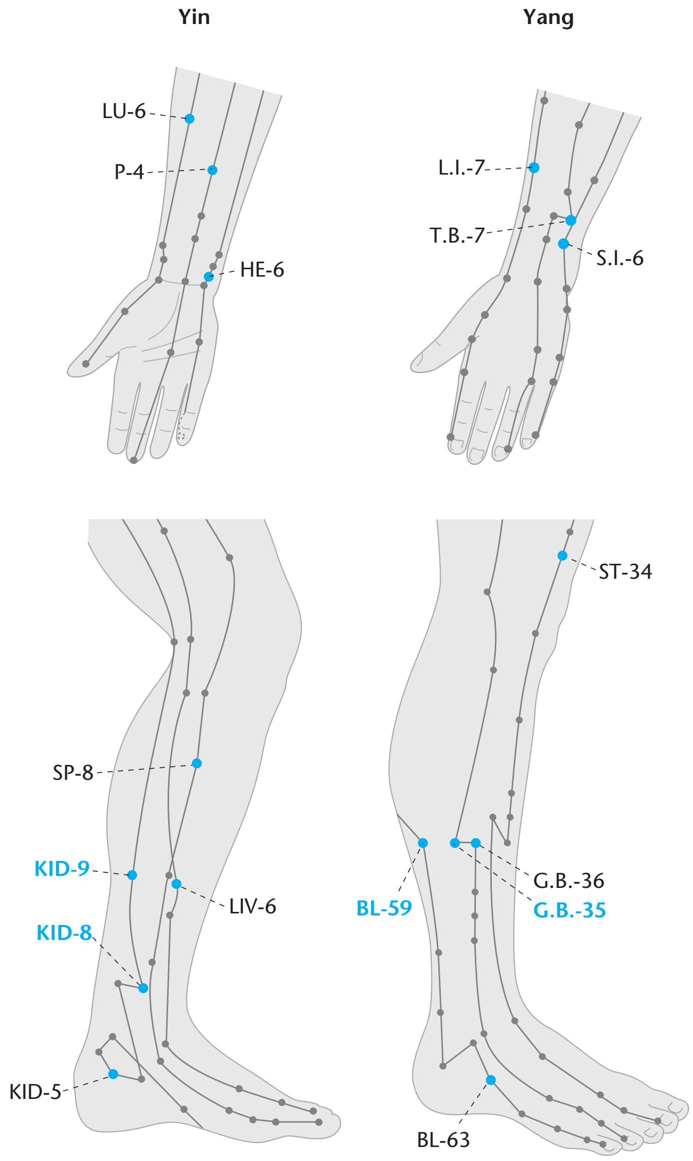
Yin
Yang
LU-6
P-4
HE-6
L.I.-7
T.B.-7
S.I.-6
ST-34
SP-8
KID-9
LIV-6
KID-8
KID-5
BL-59
G.B.-36
G.B.-35
BL-63

Points

Xi-cleft points of the primary Yin channels

Hand Yin channels

LU-6 (*kongzui*)	**HE-6** (*yinxi*)	**P-4** (*ximen*)

Foot Yin channels

SP-8 (*diji*)	**KID-5** (*shuiquan*)	**LIV-6** (*zhongdu*)

Xi-cleft points of the primary Yang channels

Hand Yang channels

L.I.-7 (*wenliu*)	**S.I.-6** (*yanglao*)	**T.B.-7** (*huizhong*)

Foot Yang channels

ST-34 (*liangqiu*)	**BL-63** (*jinmen*)	**G.B.-36** (*waiqiu*)

Xi-cleft points of the extraordinary vessels (point names printed in blue)

- *Xi*-cleft point of the *yang qiao mai*: **BL-59** (*fuyang*)
- *Xi*-cleft point of the *yin qiao mai*: **KID-8** (*jiaoxin*)
- *Xi*-cleft point of the *yang wei mai*: **G.B.-35** (*yangjiao*)
- *Xi*-cleft point of the *yin wei mai*: **KID-9** (*zhubin*)

Clinical application

Diagnosis

- They provide information about disorders affecting their pertaining channel or *zangfu*-Organ
  - Excess: acute, intense pain upon pressure or redness and swelling in the area of the point
  - Deficiency: dull, mild pain upon pressure or presence of a depression in the area of the point

Therapy

- Mainly used for acute and/or therapy-resistant disorders, especially for excess conditions and pain of the affected channel and/or pertaining *zangfu*-Organ. In patients with a robust constitution use reducing needling techniques.
- Add the *xi*-cleft points of the Yin channels for disorders of the Blood.
- In order to increase the therapeutic effect, the *xi*-cleft point may be combined with a relevant *hui*-meeting point (→ 8.1.7).

8.1.4 Back-*Shu* Points

Synonym: Back-transporting points

The Back-*shu* points are located on the inner branch of the BL channel at the level of their pertaining *zangfu*-Organ. They are mostly at the same level as the corresponding Front-*mu* point (→ 8.1.5).

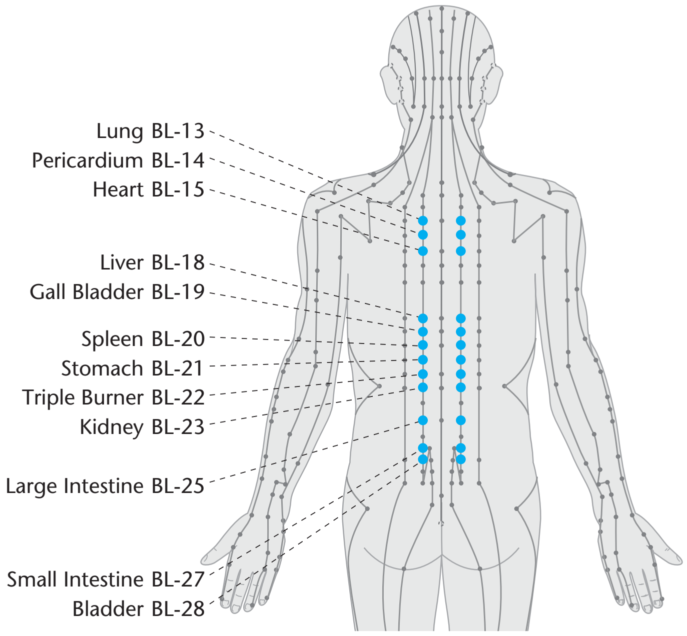
Lung BL-13
Pericardium BL-14
Heart BL-15
Liver BL-18
Gall Bladder BL-19
Spleen BL-20
Stomach BL-21
Triple Burner BL-22
Kidney BL-23
Large Intestine BL-25
Small Intestine BL-27
Bladder BL-28

Points

Organ	Back-*shu* point
Lung	**BL-13** (*feishu*)
Large Intestine	**BL-25** (*dachangshu*)
Stomach	**BL-21** (*weishu*)
Spleen	**BL-20** (*pishu*)
Heart	**BL-15** (*xinshu*)
Small Intestine	**BL-27** (*xiaochangshu*)
Bladder	**BL-28** (*pangguangshu*)
Kidney	**BL-23** (*shenshu*)
Pericardium	**BL-14** (*jueyinshu*)
Triple Burner	**BL-22** (*sanjiaoshu*)
Gall Bladder	**BL-19** (*danshu*)
Liver	**BL-18** (*ganshu*)

Clinical application

Diagnosis

- Often spontaneously tender and/or sensitive upon pressure in disorders of the pertaining *zangfu*-Organ

Therapy

- Mainly used for disorders of the corresponding *zangfu*-Organ. These points have a tonifying effect for chronic disorders and deficiency syndromes, but can also be used for acute disorders of their pertaining *zangfu*-Organ.
- For disorders of the sensory organs, use the Back-*shu* point of the pertaining *zangfu*-Organ.
  Example: for disorders of the eyes, needle the Back-*shu* point of the Liver **BL-18** (*ganshu*).

> For excess syndromes, the Back-*shu* points can be needled with reducing needling techniques; for deficiency syndromes, tonifying needling techniques should be applied. When Cold signs are present, moxibustion can also be used. The needles should not be retained longer than 10 minutes – longer needle retention can have a sedating effect and lead to tiredness.

8.1.5 Front-*Mu* Points

Synonym: Alarm points
The Front-*mu* points are located on the thorax and abdomen in the area of their pertaining *zangfu*-Organs.

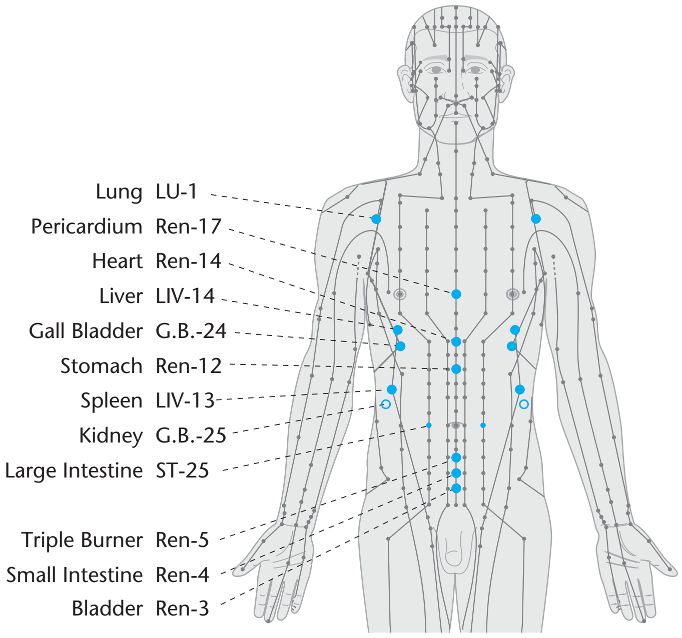

Points

Organ	Back-*shu* point
Lung	**LU-1** (*zhongfu*)
Large Intestine	**ST-25** (*tianshu*)
Stomach	**Ren-12** (*zhongwan*)
Spleen	**LIV-13** (*zhangmen*)
Heart	**Ren-14** (*juque*)
Small Intestine	**Ren-4** (*guanyuan*)
Bladder	**Ren-3** (*zhongji*)
Kidney	**G.B.-25** (*jingmen*)
Pericardium	**Ren-17** (*danzhong*)
Triple Burner	**Ren-5** (*shimen*)
Gall Bladder	**G.B.-24** (*riyue*)
Liver	**LIV-14** (*qimen*)

Clinical application

Diagnosis

- Often spontaneously tender or sensitive upon pressure in disorders of the pertaining *zangfu*-Organ

Therapy

- Used to regulate the pertaining *zangfu*-Organ. For excess syndromes use reducing needling techniques; for deficiency syndromes use tonifying methods.
- Commonly used point combinations:
  – *Shu–mu* method (→ 8.3.2)
  – For disorders of the *fu*-Organs: simultaneous needling of the lower *he*-sea point (→ 8.1.9) and the Front-*mu* point of the affected *fu*-Organ.

8.1.6 The Five *Shu*-Transporting Points

The five *shu-t*ransporting points are located **between the fingers and the elbow on the upper extremity, between the toes and the knee on the lower extremity**. The classification and application of these points is based on two theories: on the one hand they are applied according to the Qi flow in the channel from distal to proximal, on the other hand they are used according to the Five Phases (→ 8.2.5).

While the location of the five *shu-t*ransporting points is identical to the Five Phase points, their dynamics and range of application based on the Qi flow in the channels differs from or is even contrary to their actions as Five Phase points. The theoretical foundations of the Qi flow in the channels as applied in the instance of the *shu*-transporting point are based on a rather earlier circulatory concept (→ 1.1.1).

The therapeutic application of the Five Phase points will be explained in more detail in → 8.2.5, while the clinical application according to channel flow is described below.

Theoretical background of the Qi circulation in the channels

According to the more original centripetal model (→ 1.1.1), the Qi flows from the peripheral parts of the body to the elbows and knees, similar to a river coursing from its spring to its estuary and finally into the sea:

- The Qi flow begins at the anastomosis at the *jing*-well point. Here the water surfaces, it is dynamic and unstable (change of polarity from Yin to Yang)
- The water course then changes to a spring (*ying*) and a stream (*shu*) – both are at this point still quite dynamic features, the water flowing rapidly and superficially.
- The water then becomes increasingly deeper and slower, taking on the characteristics of a river (*jing*).
- Eventually it flows as a big river into the sea (*he*).

This image of a river system is often used to explain disorders caused by pathogenic factors (*xie qi*). Due to the superficial position of these points on the extremities, pathogens can here easily penetrate the body. By the same token, they can be removed again with relatively little effort. Barbara Kirschbaum (course material 1999) provides a useful visual image: she suggests imagining a ship carrying a cargo of pathogens that begins its journey at the tips of the extremities. The further a point is located peripherally in the body, the easier it is for the ship to get rid of its dangerous cargo (or to expel the pathogen according to the dynamics of the Qi flow). However, once the ship has reached the *he*-sea point, the ship and its cargo will disappear in the big, wide sea, having the potential to injure the internal *zangfu*-Organ. Thus it is the therapeutic goal to eliminate pathogens from the body as soon as possible.

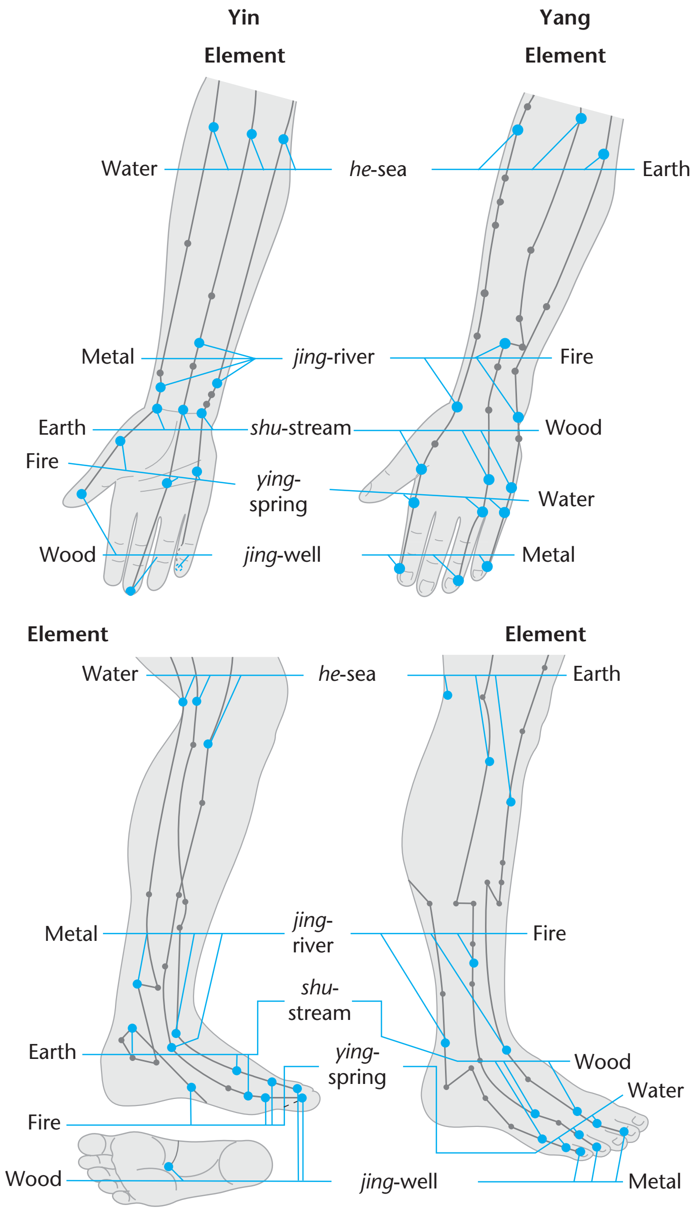

Overview: *shu*-transporting points

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41+	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

Clinical application of the *shu*-transporting points

Jing-well points (*jing xue*)

***Shu*-transporting Points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (→ 8.2.5) the *jing*-well points of the **Yin channels** correspond to the **Wood points**, while those of the **Yang channels** correspond to the **Metal points**.

Location: as the **last or first point of each channel**, they are located on the **tips of the fingers or the toes** (exception: **KID-1** of the sole of the foot). The course of the channel is here at its **most superficial**. It is also here that a change of polarity is taking place in the energy: from Yin to Yang or vice versa. **Caution**: needling these points can be quite intense and painful, presenting a strong stimulus to the body.

Actions

- Clear Heat, restore consciousness, eliminate Heat and excess from the upper end of the channel: commonly pricked to bleed.
- Treat fullness and Cold below the Heart
- Treat disorders of the spirit (*shen*) for psychological disorders

Indications

- For acute conditions and emergencies: the *jing*-well points can quickly expel Exterior pathogenic factors (*xie qi*), especially Heat. They have a strong effect on the psyche, for example for coma, collapse and shock as well as for severe irritability and restlessness.
- The points on the Yin channels also have an effect on their pertaining *zang*-Organs and can eliminate internal Wind.

Examples

- **LU-11**: prick to bleed for acute sore throat
- **KID-1**: for spasms and unconsciousness
- **HE-9**: for collapse
- **SP-1**: for uterine bleeding

For psychological disorders:

- **KID-1**: for severe restlessness
- **P-9** and **HE-9**: for extreme restlessness and insomnia

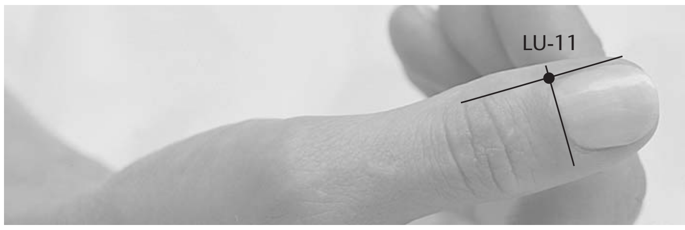

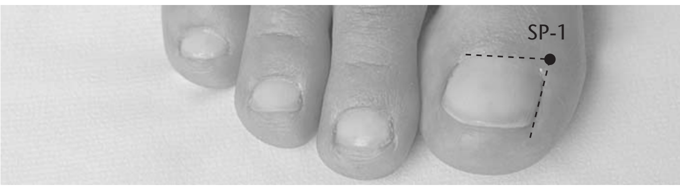

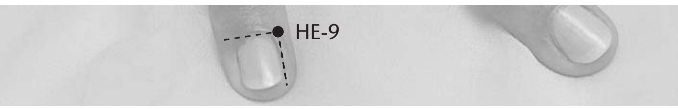

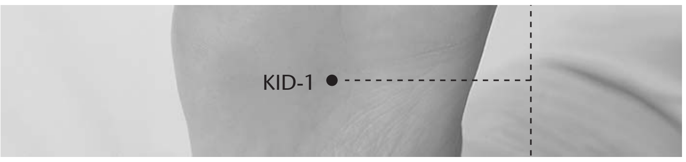

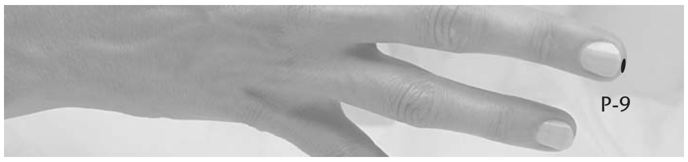

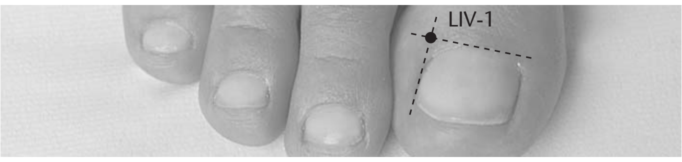

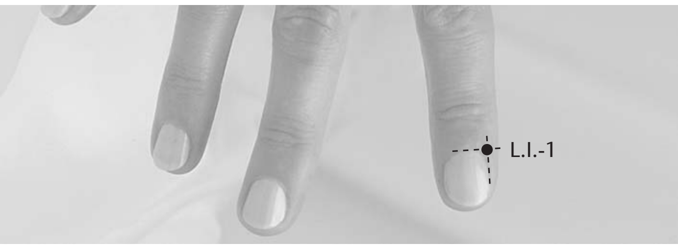

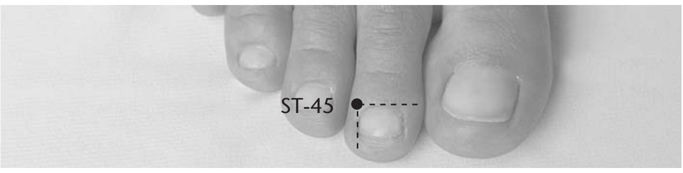

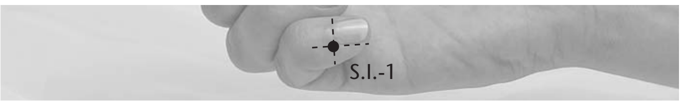

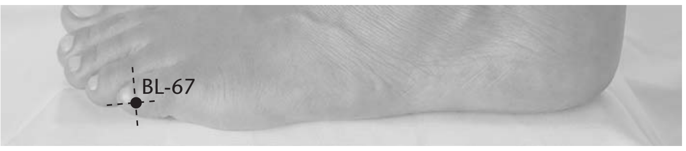

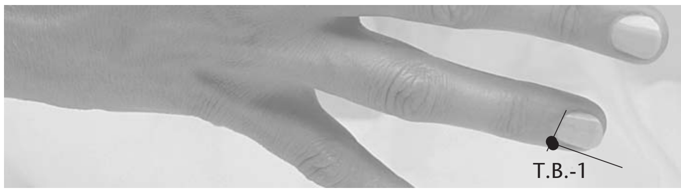

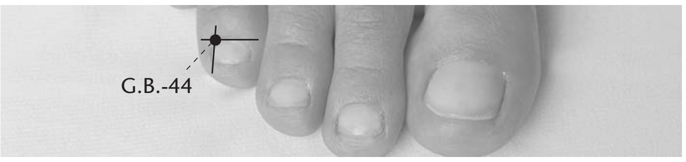

Ying-spring points (*ying xue*)

***Shu*-transporting points of the Yin channels**

Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**

Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43 +	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; −, sedation point

According to the Five Phases (→ 8.2.5) the *ying*-spring points of the Yin channels correspond to the Fire points, while those on the Yang channels correspond to the Water points. **Caution**: needling some of these points can be painful, for example **LU-10, P-8, HE-8**.

Location: as the second most distal point of each channel, they are located on the upper extremity between the phalanges and metacarpal bones and on the lower extremity between the phalanges with the exception of KID-2.

Actions

- They are very dynamic points that eliminate both external and internal pathogenic factors.
- They clear Heat from their pertaining *zangfu*-Organ and channel, especially in the upper half of the body.

Indications

- For example for febrile diseases: Exterior Heat syndromes are mainly treated using points on the Yang channels while Interior Heat or deficient Heat syndromes are treated with points on the Yin channels.

Examples

- **LU-10:** for acute inflammations of the throat
- **ST-44:** for acute frontal headaches or acute inflammations affecting the face
- **LIV-2:** for severe temporal headaches, for example due to blazing Liver Fire

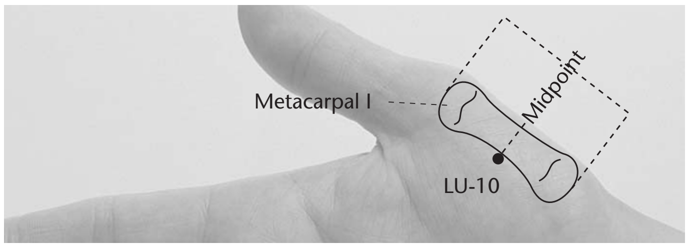

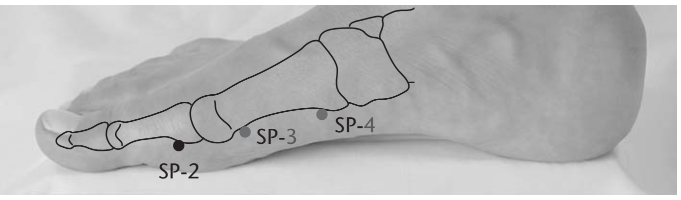

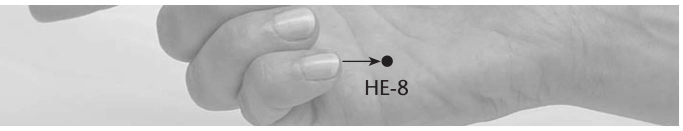

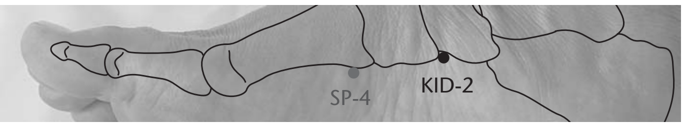

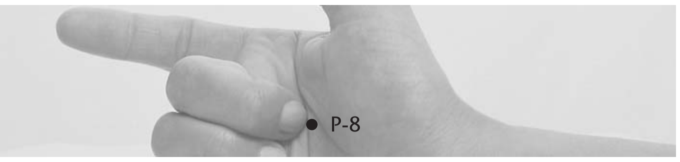

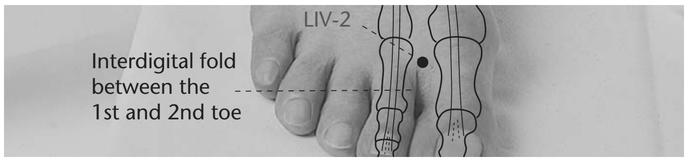

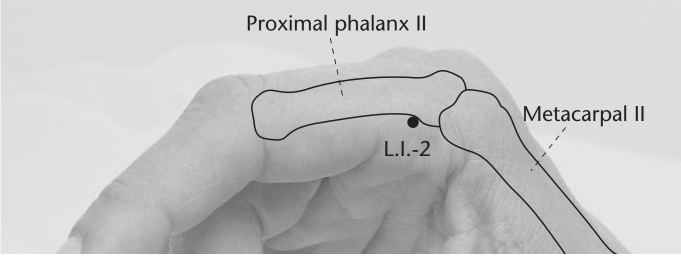

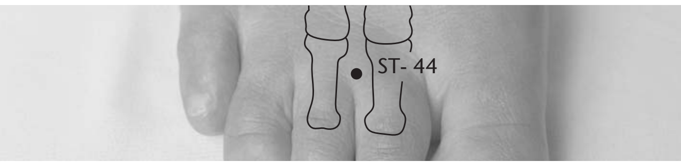

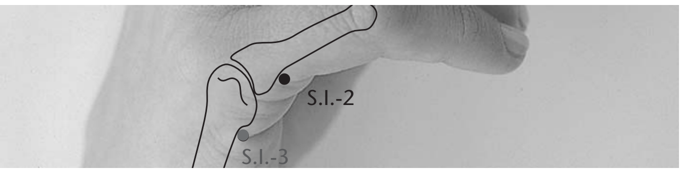

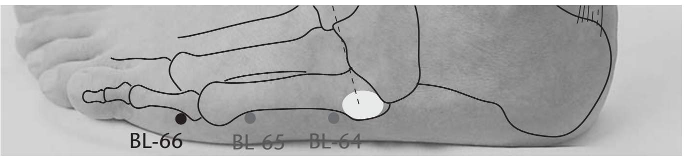
BL-66
BL-65
BL-64

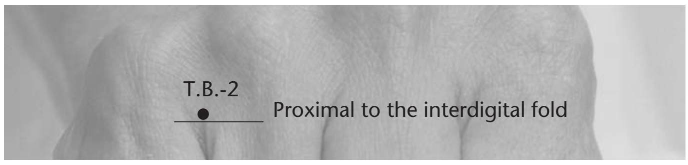
T.B.-2
Proximal to the interdigital fold

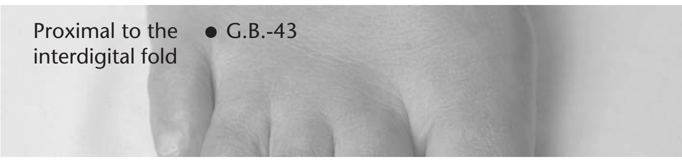
Proximal to the interdigital fold
G.B.-43

Shu-stream points (*shu xue*)

***Shu*-transporting points of the Yin channels**

Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2−	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**

Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11+
ST	ST-45 −	ST-44	ST-43	ST-41+	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; −, sedation point

According to the Five Phases (→ 8.2.5) the *shu*-stream points of the **Yin channels** correspond to the **Earth points** and ***yuan*-Source points** (→ 8.1.1), while those on the **Yang channels** correspond to the **Wood points**.

Location: counting from the periphery, they are the **third point of each channel**, the only exception being the G.B. channel (4th point on the channel). They are located proximal to the metacarpophalangeal or metatarsophalangeal joints, except for **KID-3** (posterior to the medial malleolus), **LIV-3** and **SP-3** (both located at the distal ends of the metatarsal bones) as well as **LU-9, HE-7** and **P-7** (all located on the wrist joint space).

Actions

They are considered to be the entry points for pathogenic factors; needling these points can strengthen the Defensive Qi (*wei qi*) (→ 1.1.4) and eliminate pathogenic factors from the respective channel. At the *shu*-stream points, the Qi flow is already somewhat deeper: the bed of the river is widening and deepening.

Indications

For *bi*-syndromes of the joints, especially due to Damp or pathogenic Wind. They are generally used for heaviness of the body as well as for *shaoyang* syndrome with intermittent symptoms.

- Points on the Yang channels are predominantly used for Exterior syndromes: they have a specific action on their respective channel.
- Points on the Yin channels are more commonly used for Interior syndromes; they strengthen and harmonise their respective *zang*-Organs (this is in part due to these points also being the *yuan*-source points).

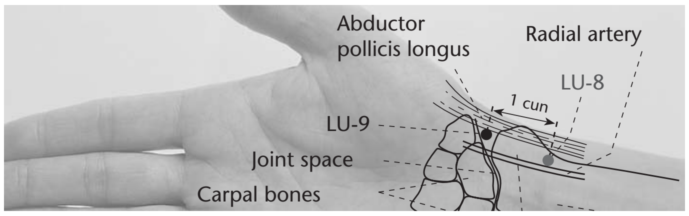
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

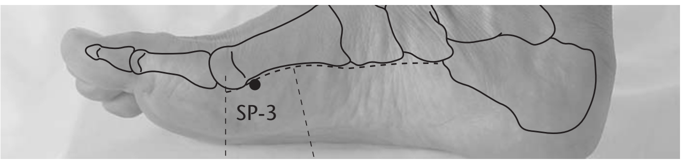

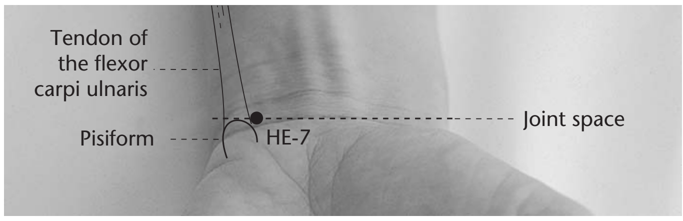
Tendon of the flexor carpi ulnaris
Joint space
Pisiform
HE-7

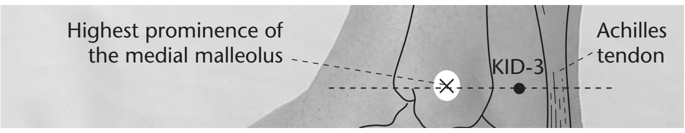
Highest prominence of the medial malleolus
Achilles tendon
KID-3

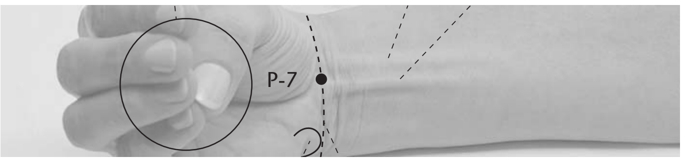

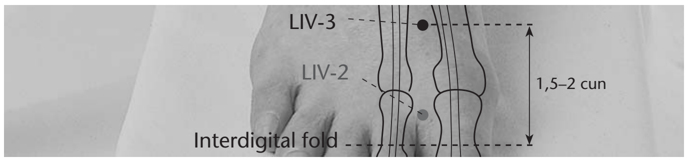
LIV-3
LIV-2
1,5–2 cun
Interdigital fold

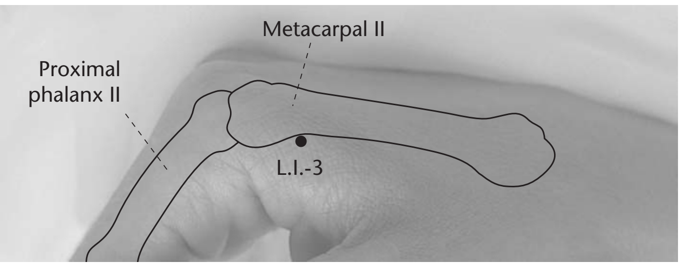

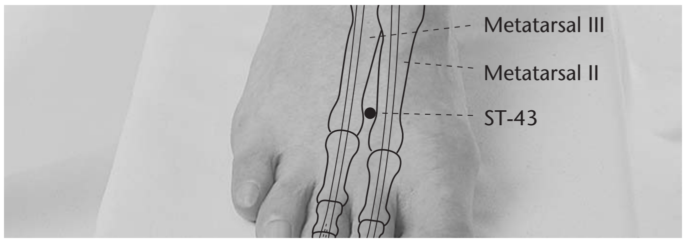

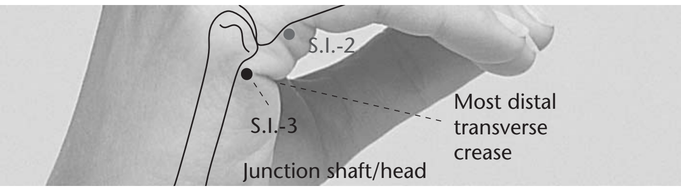

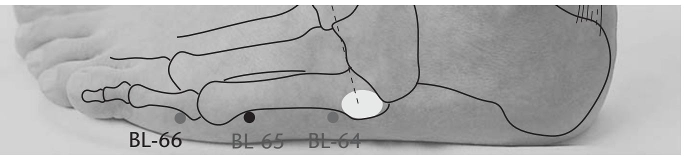

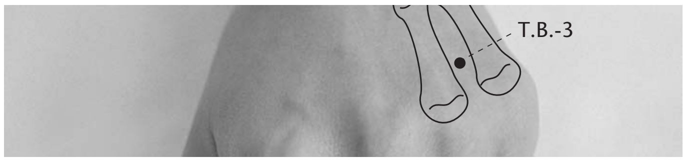

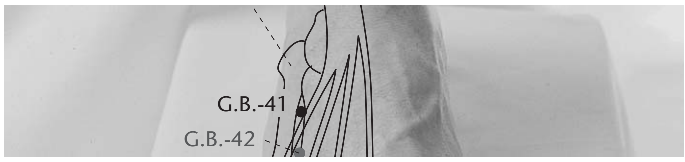

Jing-river points (*jing xue*)

***Shu*-transporting points of the Yin channels**					
Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 –
SP	SP-1	SP-2 +	SP-3	SP-5 –	SP-9
HE	HE-9 +	HE-8	HE-7 –	HE-4	HE-3
KID	KID-1 –	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7–	P-5	P-3
LIV	LIV-1	LIV-2 –	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point
***Shu*-transporting points of the Yang channels**					
Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 –	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 –	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 –
BL	BL-67 +	BL-66	BL-65 –	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 –
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38 –	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; – , sedation point

According to the Five Phases (→ 8.2.5), the *jing*-river points of the **Yin channels** correspond to the **Metal points**, while those on the **Yang channels** correspond to the **Fire points**.

Location: they are located **between** the ***shu*-stream points** and the ***he*-sea points**. On the upper extremity, they are located either at, or distal or proximal to the wrist, while on the lower extremity they are located either at, or distal or proximal to the ankle (see figure on p. 676).

Actions

They divert external pathogenic factors away from joints, bones and tendons towards the Exterior. At the *jing*-river points the Qi flow becomes increasingly wider, slower, deeper and larger – thus the actions of these points are also less dynamic compared to the first three, more peripheral, *shu*-transporting points. At the *jing*-river points external pathogens have already penetrated deeper into the body, especially into the joints, bones and tendons.

Indications

- Dyspnoea, disorders of the throat such as inflammations, cough, disorders manifesting in the voice.
- Alternating fever and chills (especially points on the Yin channels)
- Also for *bi*-syndromes such as painful joint and tendon disorders.

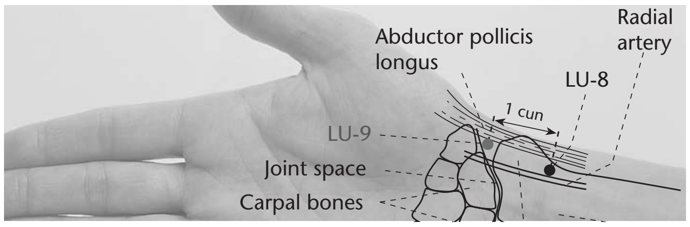
Abductor pollicis longus
Radial artery
LU-8
1 cun
LU-9
Joint space
Carpal bones

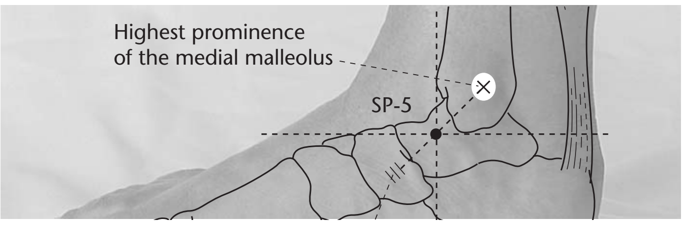
Highest prominence of the medial malleolus
SP-5

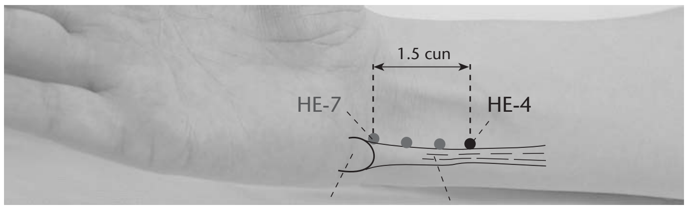
1.5 cun
HE-7
HE-4

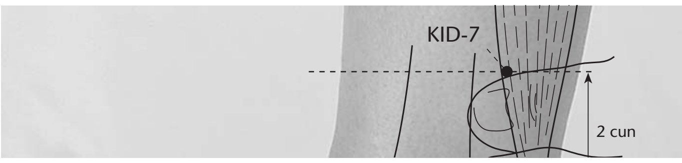

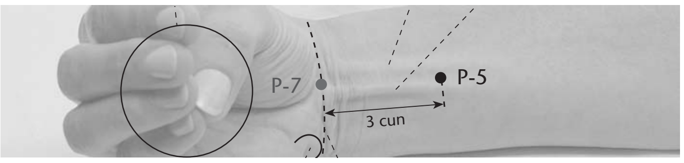

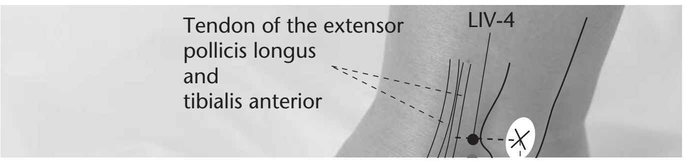
Tendon of the extensor pollicis longus and tibialis anterior
LIV-4

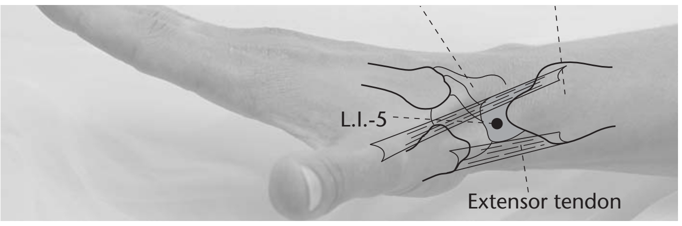
L.I.-5
Extensor tendon

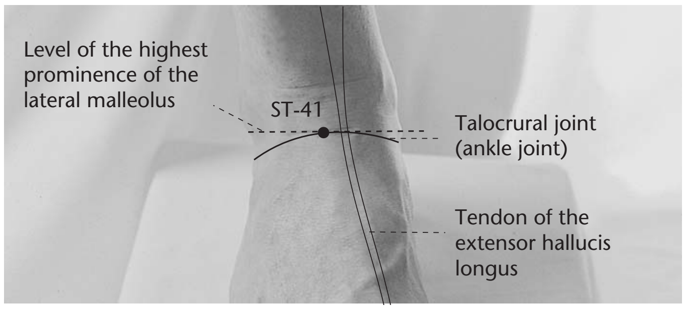
Level of the highest prominence of the lateral malleolus
ST-41
Talocrural joint (ankle joint)
Tendon of the extensor hallucis longus

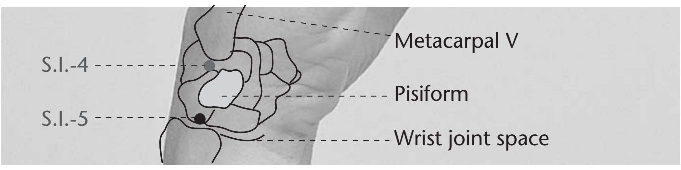
S.I.-4
S.I.-5
Metacarpal V
Pisiform
Wrist joint space

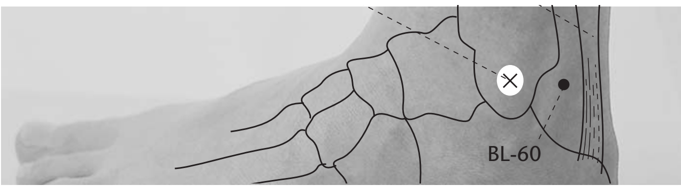

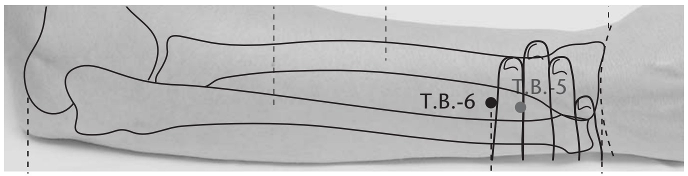

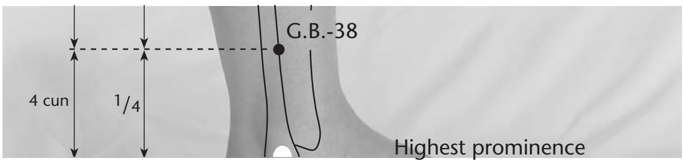
G.B.-38
4 cun
1/4
Highest prominence

He-sea points (*he xue*)

***Shu*-transporting points of the Yin channels**

Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7−	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9+	P-8	P-7−	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**

Phase/ Channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

According to the Five Phases (→ 8.2.5) the *he*-sea points of the **Yin channels** correspond to the **Water points**, while those on the **Yang channels** correspond to the **Earth points**.

Location: on the upper extremity these points are located near the **elbow**, on the lower extremity near the **knee**.

Actions

- They descend counterflow Qi and stop diarrhoea. At the *he*-sea points the Qi flow is very wide, deep and slow, 'terminating in the deep sea' where the 'toxic cargo' can 'disappear' into the body unless treated adequately. External pathogens (*xie qi*) which have not been eliminated from the channel beforehand can here penetrate the Interior of the body and injure the *zangfu*-Organs.
- Diseases of the *fu*-Organs (here especially points on the lower extremities are indicated).

Indications

- Disorders of the *fu*-Organs, especially gastro-intestinal disorders with symptoms such as vomiting and diarrhoea
- Some *he*-sea points are also indicated for skin diseases, for example **LI-11, BL-40**.

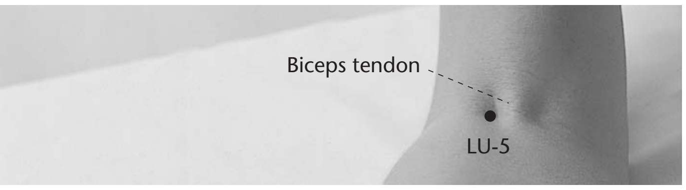
Biceps tendon
LU-5

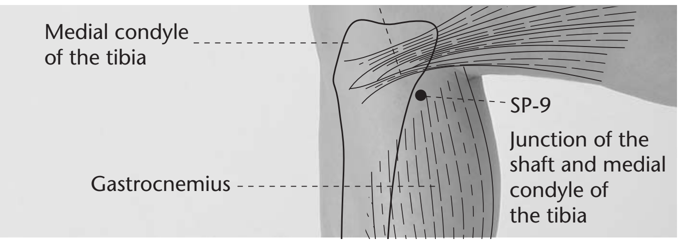
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

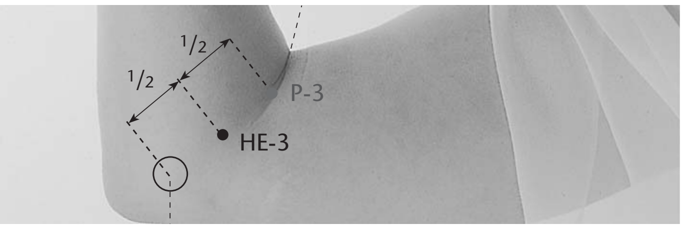

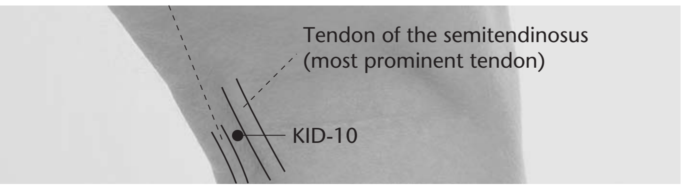
Tendon of the semitendinosus (most prominent tendon)
KID-10

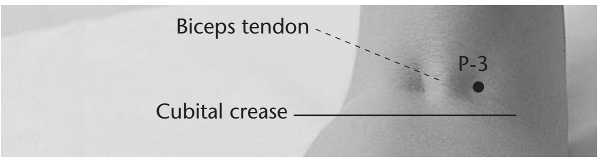
Biceps tendon
P-3
Cubital crease

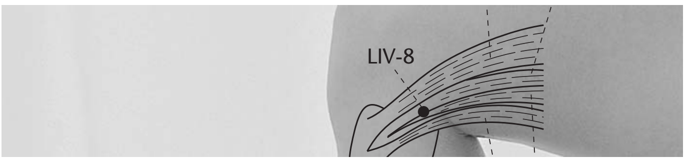

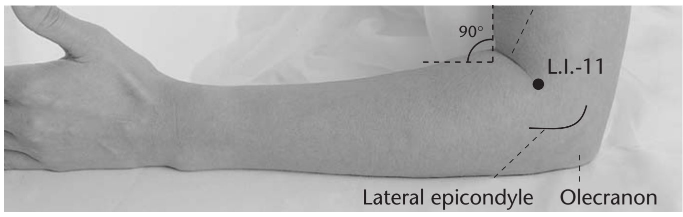
90°
L.I.-11
Lateral epicondyle
Olecranon

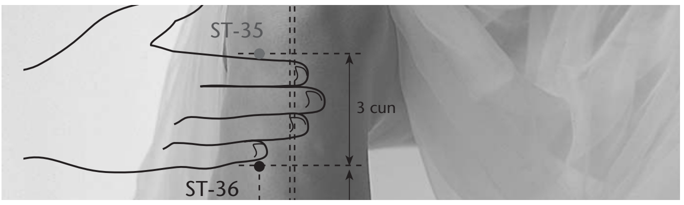
ST-35
3 cun
ST-36

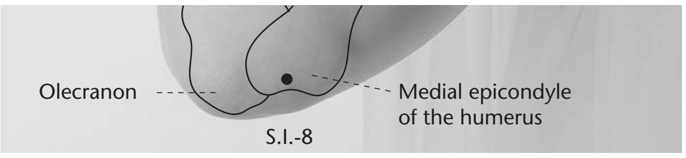
Olecranon
Medial epicondyle of the humerus
S.I.-8

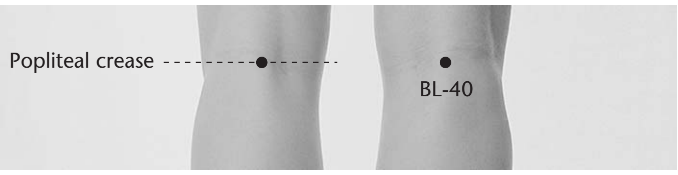
Popliteal crease
BL-40

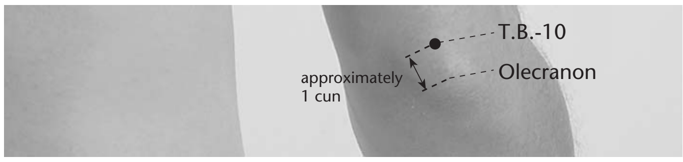
T.B.-10
approximately 1 cun
Olecranon

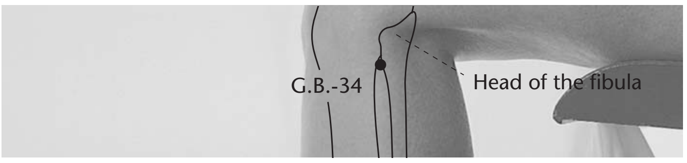
G.B.-34
Head of the fibula

8.1.7 The Eight *Hui*-Meeting Points

Synonym: Gathering Points (Maciocia)
The Qi of a particular tissue, Organ system or substance gathers at the *hui*-meeting points and can be influenced through these points.

Points and clinical application

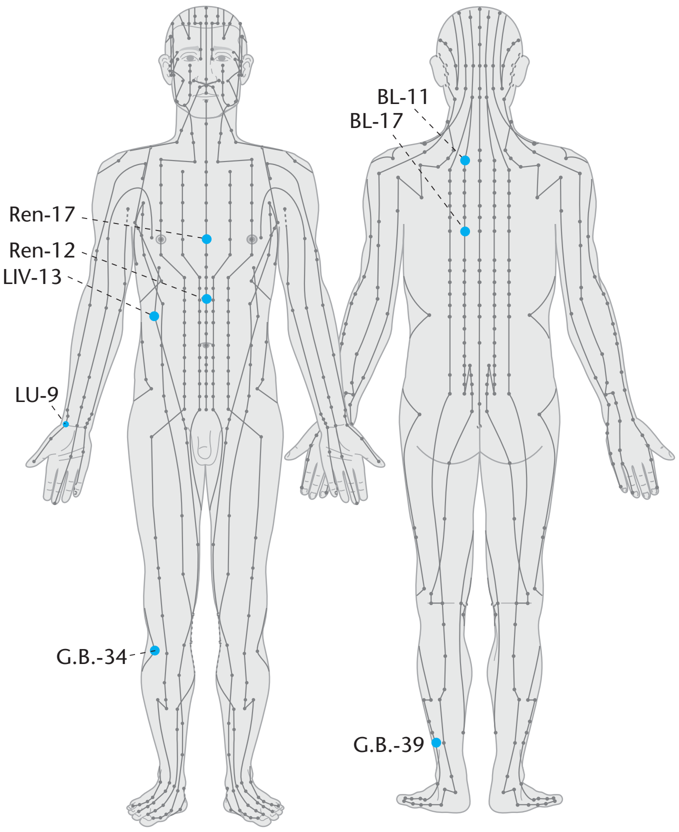
BL-11
BL-17
Ren-17
Ren-12
LIV-13
LU-9
G.B.-34
G.B.-39

The Eight *hui*-Meeting Points		
Point	Tissue/organ system/substance	Indications
LIV-13 (*zhangmen*)	*zang*-Organs	Disorders of the *zang*-Organs (tonifies the Spleen and indirectly all *zang*-Organs)
Ren-12 (*zhongwan*)	*fu*-Organs	Disorders of the *fu*-Organs, for example of the gastro-intestinal tract
Ren-17 (*danzhong or shanzhong*)	Qi, has a strong effect on the *zong qi*	Respiratory diseases and disorders of the Qi flow such as hiccups; controls the *zong qi* and therefore functioning of the Lungs and Heart
BL-17 (*geshu*)	Blood (*xue*)	Disorders of the Blood such as anaemia, Blood stasis, haemorrhagia and gynaecological disorders
G.B.-34 (*yangling-quan*)	Sinews	Disorders of the joints, sinews and muscles
LU-9 (*taiyuan*)	Blood vessels	Stimulates the peripheral Blood circulation for disorders of the blood vessels such as vasculitis and arteriosclerosis
BL-11 (*dazhu*)	Bones	Disorders of the Bones such as shoulder and spine problems, joint and bone pain
G.B.-39 (*xuanzhong*)	Marrow	Disorders of the Bones, the Marrow and Brain (nerves)

8.1.8 Opening Points of the Eight Extraordinary Vessels

Synonyms: Confluent points (Deadman et al 1998), Master points (Pirog 1996)
The table below shows the opening points of the eight extraordinary vessels. For more detail on their origin and pathways → 1.7, Chapter 5.

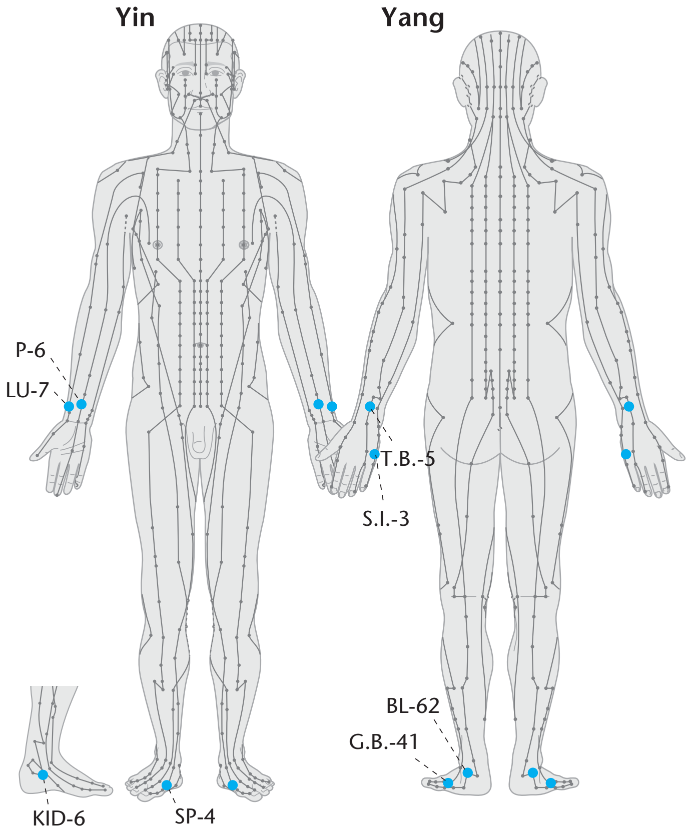
Yin
Yang
P-6
LU-7
T.B.-5
S.I.-3
BL-62
G.B.-41
KID-6
SP-4

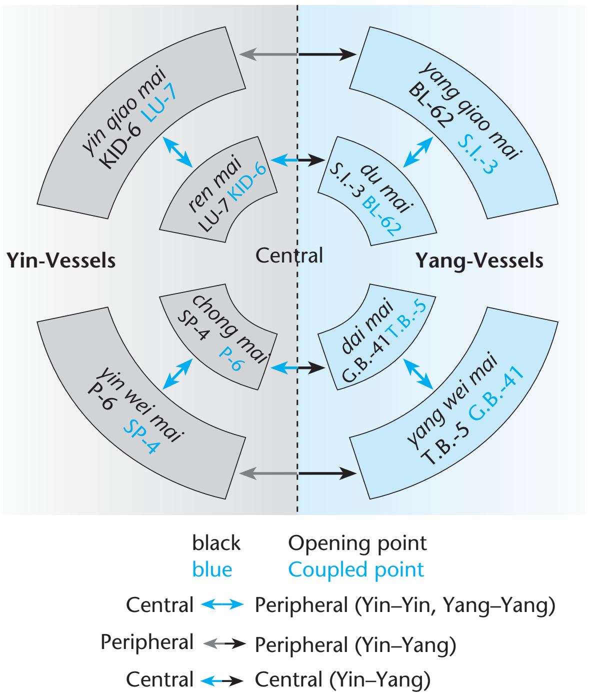
yin qiao mai
KID-6 LU-7
ren mai
LU-7 KID-6
yang qiao mai
BL-62 S.I.-3
du mai
S.I.-3 BL-62
Yin-Vessels
Central
Yang-Vessels
chong mai
SP-4 P-6
dai mai
G.B.-41 T.B.-5
yin wei mai
P-6 SP-4
yang wei mai
T.B.-5 G.B.-41
black Opening point
blue Coupled point
Central ↔ Peripheral (Yin–Yin, Yang–Yang)
Peripheral ↔ Peripheral (Yin–Yang)
Central ↔ Central (Yin–Yang)

Points

Pairs	Extraordinary Vessel	Opening point	Coupled point (opening point of the paired vessel)	Supported body regions
Pair 1	*chong mai*	**SP-4** (*gongsun*)	**P-6** (*neiguan*)	Heart, chest, Stomach
	yin wei mai	**P-6** (*neiguan*)	**SP-4** (*gongsun*)	
Pair 2	*du mai*	**S.I.-3** (*houxi*)	**BL-62** (*shenmai*)	Occiput, shoulder, back
	yang qiao mai	**BL-62** (*shenmai*)	**S.I.-3** (*houxi*)	
Pair 3	*dai mai*	**G.B.-41** (*zulinqi*)	**T.B.-5** (*waiguan*)	Temples, ears, lateral aspects of the body
	yang wei mai	**T.B.-5** (*waiguan*)	**G.B.-41** (*zulinqi*)	
Pair 4	*ren mai*	**LU-7** (*lieque*)	**KID-6** (*zhaohai*)	Face, throat, chest, Lungs, abdomen
	yin qiao mai	**KID-6** (*zhaohai*)	**LU-7** (*lieque*)	

Clinical application

Various approaches exist for needling the opening/coupled points of the extraordinary vessels. The following method **for 'opening' or regulating the extraordinary vessels** has been used with good therapeutic results in clinical practice:

- Needle the **opening point** of the relevant channel first, then needle the **contralateral coupled point** (opening point of the paired extraordinary vessel). This will mobilise and regulate the Qi of the extraordinary vessel, making it more dynamic.
- During a subsequent session, needle **either the *xi*-cleft point** of the previously treated extraordinary vessel (→ 8.1.3) or add **other channel points of the vessel** (see channel pathways, chapter 5) during the same or subsequent session. This will further support the dynamics of the Qi flow in the vessel.

In addition Maciocia (1989) and Kirschbaum (1995) suggest the following procedure for 'opening' the extraordinary vessels:

- **In women:** First needle the opening point on the *right* side with even method, then the coupled point on the *left* side with even method. Afterwards other points on the respective vessel may be added.
- **In men:** First needle the opening point on the *left* side with even method, then the coupled point on the *right* side with even method. Afterwards other points on the respective vessel may be added.
- **Needle retention:** the needles should be retained for 20–25 minutes and should be removed in the opposite order.

8.1.9 Lower *He*-Sea Points

The lower *he*-sea points are used as the main points for disorders of their pertaining *fu*-Organ.

Points and clinical application

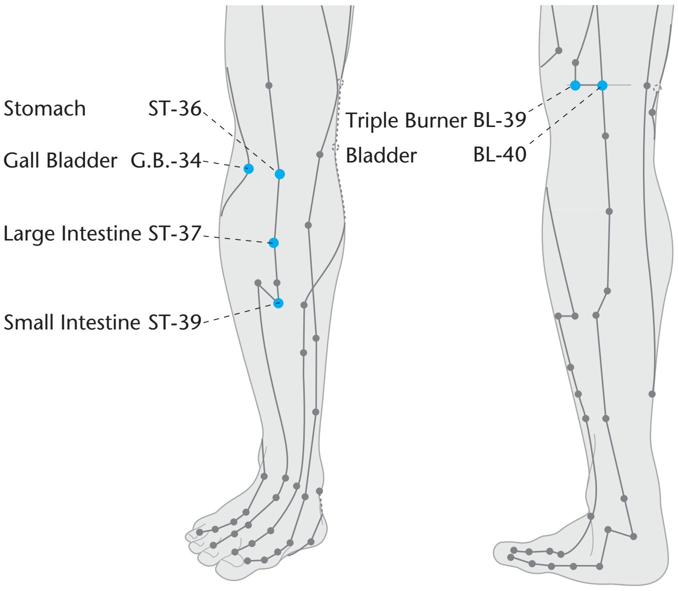

- **ST-36** (*zusanli*): Lower *he*-sea point of the Stomach; indication: loss of appetite, epigastric pain and fullness, acid reflux, abdominal pain, constipation, diarrhoea
- **ST-37** (*shangjuxu*): Lower *he*-sea point of the Large Intestine; indication: appendicitis, diarrhoea
- **ST-39** (*xiajuxu*): Lower *he*-sea point of the Small Intestine; indication: diarrhoea, abdominal pain
- **G.B.-34** (*yanglingquan*): Lower *he*-sea point of the Gall Bladder; indication: cholecystitis, vomiting
- **BL-40** (*weizhong*): Lower *he*-sea point of the Bladder; indication: urinary incontinence, urinary retention
- **BL-39** (*weiyang*): Lower *he*-sea point of the Triple Burner; indication: urinary incontinence, urinary retention

8.1.10 Meeting Points (*Jiaohui* Points)

Synonym: Intersecting points (Practical Dictionary)
Meeting points (*jiaohui* points) represent intersections between channels and vessels. Needling them allows more than one channel to be influenced and widens the range of actions of a given point. By the same token, the number of needles may be reduced while maintaining a good therapeutic effect.

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
LU 1	▒			■																
L.I. 14		░				▨	▨								▨					
L.I. 15		░															■			
L.I. 16		░															■			
L.I. 20		░	■																	
ST 1			░										▨	■			■			
ST 2			░														▨			
ST 3			░														■			
ST 4		■	░										▨	▨			■			
ST 5			░								▨									
ST 6			░								▨									
ST 7			░								■									
ST 8			░								■				■					
ST 9			░								■									
ST 12		■	░			■				■	■									
ST 30			░								▨								■	
SP 6				▒				■				■						▨		
SP 12				▒						▨		■				■				
SP 13				▒								■				■				
SP 15				▒												■				
SP 16				▒												■				
S.I. 10						░	■								■		■			
S.I. 12		■				░				■	■									
S.I. 18						░				■										
S.I. 19						░				■	■									
BL 1			■			■	░			▨	▨		■				■	■		
BL 11						■	░			▨	▨		▨							
BL 12							░						■							
BL 23							░						▨							
BL 31							░				■									
BL 32							░				■									
BL 33							░				■									
BL 34							░				■									
BL 41						■	░													
BL 59							░										▨			
BL 61							░										■			
BL 62							░										▨			
BL 63							░								■					
KID 2								▒										▨		
KID 6								▒										▨		
KID 8								▒										▨		
KID 9								▒								■				
KID 11								▒											■	
KID 12								▒											■	
KID 13								▒											■	
KID 14								▒											■	
KID 15								▒											■	
KID 16								▒											■	
KID 17								▒											■	
KID 18								▒											■	

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
KID 19								▒											■	
KID 20								▒											■	
KID 21								▒											■	
P 1									▒	▨	■	■								
T.B. 13										░					▨					
T.B. 15										░	■				■					
T.B. 17										░	■									
T.B. 20						▨				░	■									
T.B. 22						■				░	■									
G.B. 1						■				■	░									
G.B. 3			■							■	░									
G.B. 4			■							■	░									
G.B. 5		■	■							■	░									
G.B. 6		■	■							■	░									
G.B. 7							■				░									
G.B. 8							■				░									
G.B. 9							■				░									
G.B. 10							■				░									
G.B. 11						▨	■			▨	░									
G.B. 12							■				░									
G.B. 13											░				■					
G.B. 14		▨	▨							▨	░				■					
G.B. 15							■				░				■					
G.B. 16											░				■					
G.B. 17											░				■					
G.B. 18											░				■					
G.B. 19											░				■					
G.B. 20										▨	░				■		■			
G.B. 21			▨							■	░				■					
G.B. 23							▨				░									
G.B. 24				■							░				▨					
G.B. 26											░									■
G.B. 27											░									■
G.B. 28											░									■
G.B. 29											░						■			▨
G.B. 30							■				░									
G.B. 35											░						▨			
LIV 13											■	▒								▨
LIV 14				■								▒				■				
Du 1								■			■		░	■						
Du 13							■						░							
Du 14		■	■			■	■			■	■		░							
Du 15													░		■					
Du 16													░		■		▨			
Du 17							■						░							
Du 20							■			■	■	■	░							
Du 24			■				■						░							
Du 26		■	■										░							
Du 28			▨										░	■						
Ren 1													■	▒					■	

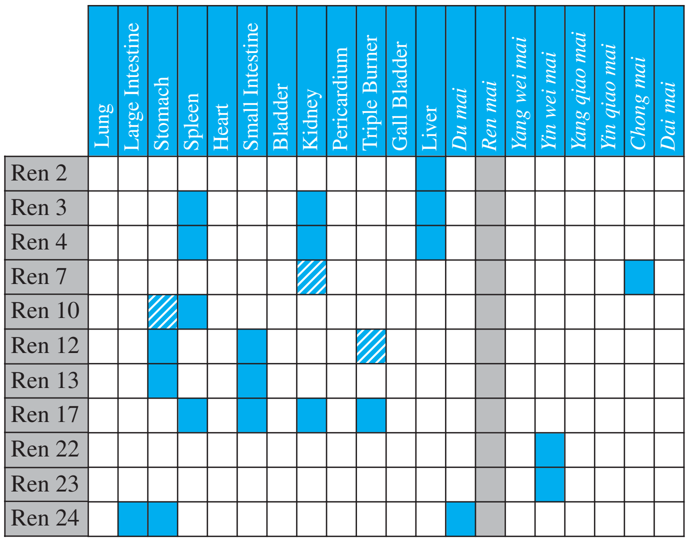

	Lung	Large Intestine	Stomach	Spleen	Heart	Small Intestine	Bladder	Kidney	Pericardium	Triple Burner	Gall Bladder	Liver	*Du mai*	*Ren mai*	*Yang wei mai*	*Yin wei mai*	*Yang qiao mai*	*Yin qiao mai*	*Chong mai*	*Dai mai*
Ren 2												■								
Ren 3				■				■				■								
Ren 4				■				■				■								
Ren 7								▨											■	
Ren 10			▨	■																
Ren 12			■			■				▨										
Ren 13			■			■														
Ren 17				■		■		■		■										
Ren 22																■				
Ren 23																■				
Ren 24		■	■										■							

▨ only mentioned by some authors

8.1.11 The Gao Wu Command Points

Synonym: none known

These points have the ability to affect certain regions of the body. By combining these points with local and other specific points they can increase the therapeutic effect in a particular area of the body.

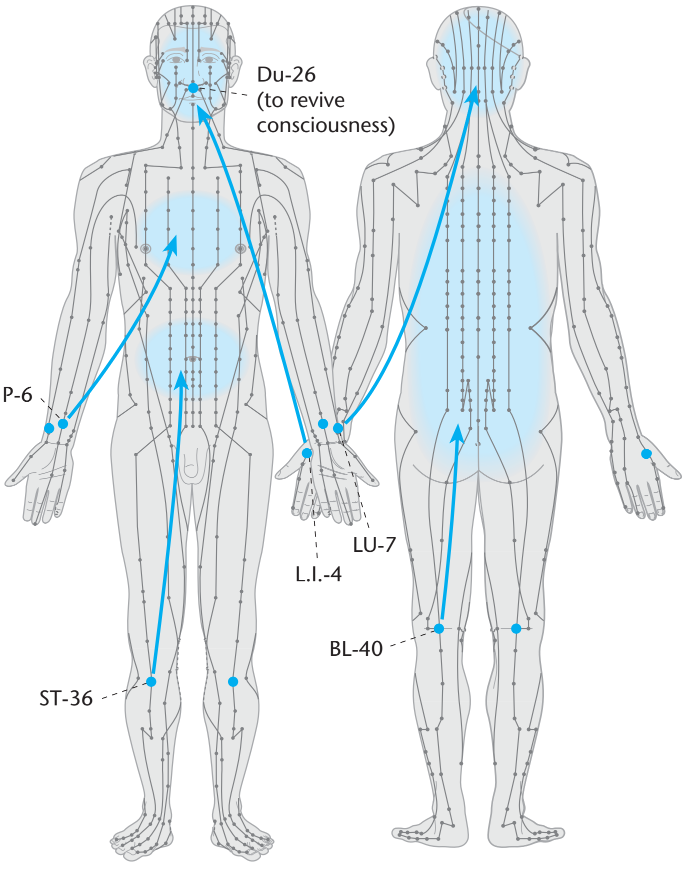

Points and clinical application

- **ST-36** (*zusanli*): for all disorders of the abdomen
- **L.I.-4** (*hegu*): for disorders of the face and mouth
- **LU-7** (*lieque*): for disorders of the occiput
- **BL-40** (*weizhong*): for disorders of the back and lumbar region
- **P-6** (*neiguan*): for disorders of the thorax
- **Du-26** (*renzhong*): for restoring consciousness (The last two points were added later, → Deadman et al 1998)

8.1.12 The Window of Heaven Points

In Chinese classical literature there are only a few references to these points. In modern times, both Ross (1995) and Deadman et al (1998) mention this group of points.

Points

LU-3 (*tianfu*)
T.B.-16 (*tianyou*)
ST-9 (*renying*)
Du-16 (*fengfu*)
L.I.-18 (*futu*)
S.I.-16 (*tianchuang*)
BL-10 (*tianzhu*)
P-1 (*tianchi*)
S.I.-17 (*tianrong*)
Ren-22 (*tiantu*)

Location: Eight of the Window of Heaven points are located in the neck region (exceptions are **LU-3** on the upper arm and **P-1** on the thorax). This might provide a clue about the action of these points in regulating the Qi flow between the head and the body.

Comment: According to Deadman, the Ming dynasty physician Ma Shi remarked that **S.I.-17** should be replaced by **G.B.-9** (*tianchong*). If that was the case all six Yang channels running to the head would be represented by a Window of Heaven point.

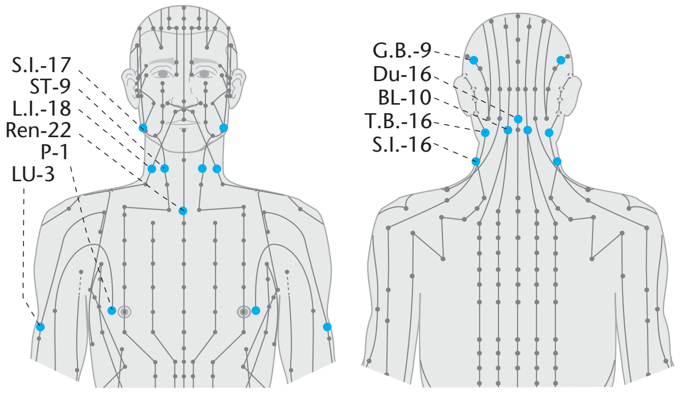

Clinical application

- For disorders of the Qi flow:
  - **LU-3, L.I.-18, Ren-22** for counterflow Qi causing cough, wheezing, etc.
  - **P-1** for counterflow Lung Qi with copious Phlegm
  - **ST-9** for counterflow Lung and Stomach Qi
  - **BL-10** treats dizziness, headaches and stiffness of the occiput, etc.
  - **Du-16** for internal Liver Wind

- For goitre, swellings, pain and Qi stagnation in the occiput and neck: especially indicated if these occur locally
- For disorders with an acute onset, for example
  - **L.I.-18** for acute aphonia
  - **ST-9** for acute diarrhoea
  - **S.I.-16** for acute aphonia, for example following a stroke
  - **BL-10** for epilepsy
  - **T.B.-16** for acute hearing loss
  - **Ren-22** for acute dyspnoea
  - **Du-16** for acute aphasia following a stroke
- For psychological disorders:
  - **LU-3** for somnolence, grief, disorientation, forgetfulness, insomnia
  - **S.I.-16** for manic agitation and bipolar disorder
  - **BL-10** for mania, hallucinations, epilepsy, childhood convulsions
  - **T.B.-16** for vivid dreaming
  - **Du-16** for mania, grief, anxiety with anxiety-induced palpitations
- For disorders of the sensory organs:
  - **LU-3** for nose bleeding, blurry vision, myopia
  - **ST-9** for blurry vision
  - **S.I.-16** for deafness, tinnitus, ear pain
  - **S.I.-17** for tinnitus and deafness
  - **BL-10** for eye pain, blurry vision, excessive tearing, speech disorders, blocked nasal passages, loss of sense of smell
  - **T.B.-16** for hearing disorders, visual disorders, eye pain, excessive tearing, loss of sense of smell, blocked nasal passages
  - **Du-16** for speech disorders (for example following a stroke), blurry vision, nosebleeds
  - **Ren-22** for speech disorders

8.1.13 Points of the Four Seas

These points have a particular, supportive effect on their respective 'Sea'.

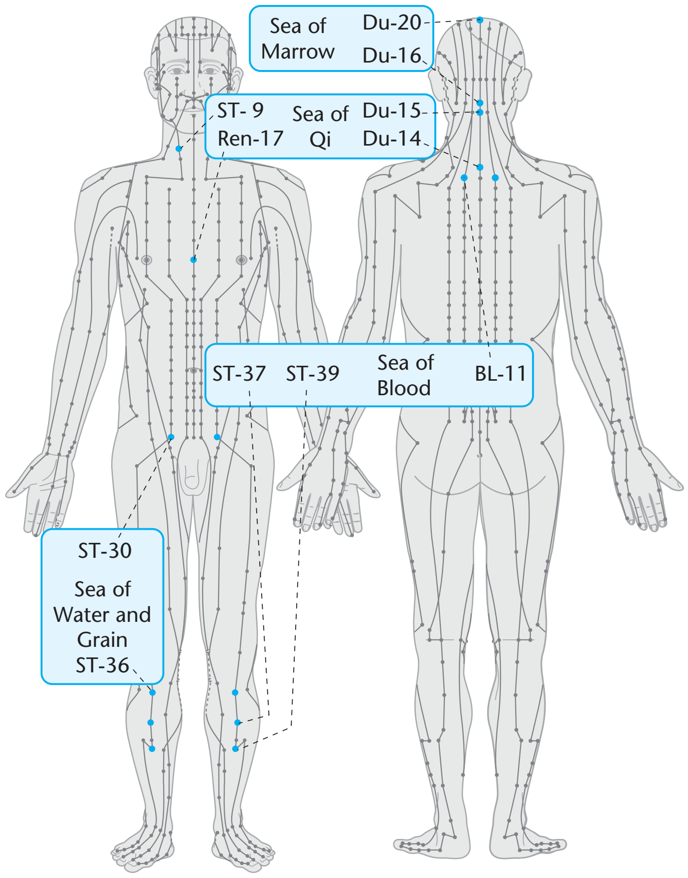

Sea of Qi	**ST-9** (*renying*), **Ren-17** (*danzhong or shanzhong*), **Du-15** (*yamen*), **Du-14** (*dazhui*)
*Ling Shu**: 'When the Sea of Qi is in excess there is fullness in the chest, urgent breathing and a red complexion. When the Sea of Qi is insufficient, there is scanty energy, insufficient for speech.'	
Sea of Blood	**BL-11** (*dazhu*), **ST-37** (*shangjuxu*), **ST-39** (*xiajuxu*)
*Ling Shu**: 'The Penetrating vessel is the sea of the twelve channels. In the upper it is conveyed to BL-11 (*dazhu*), and in the lower it emerges at both the ST-37 (*shangjuxu*) and ST-39 (*xiajuxu*) …When the Sea of Blood is in excess, then one has a sensation of the body being big; one feels disquiet, but does not know what the illness is; when the Sea of Blood is insufficient, one has the sensation of one's body being small; one feels reduced but does not know what the illness is.'	
Sea of Water and Grain	**ST-30** (*qichong*) as upper point, **ST-36** (*zusanli*) as lower point
*Ling Shu**: 'When the Sea of Water and Grain is in excess, there is abdominal fullness, and when it is deficient, there is hunger with inability to eat.'.	
Sea of Marrow	**Du-20** (*baihui*), **Du-16** (*fengfu*)
*Ling Shu**: 'When the Sea of Marrow is in excess, there is lightness of the body and much strength, and a person's feeling of 'self' exceeds the normal level; when the Sea of Marrow is insufficient there is a whirling sensation of the brain, dizziness, tinnitus, pain of the lower legs, impairment of vision, indolence and desire to sleep.'	

---

* Quotes from the *Ling Shu* according to Deadman et al. (1998).

Ch08-F10028.qxd 2/23/08 4:41 PM Page 683

8.1.14 The 12 Heavenly Star Points of Ma Dan Yang

Ma Dan Yang, a famous physician of the Jin dynasty, considered these initial 11 points as the most important acupuncture points. It was the physician Xu Feng who later added **LIV-3** as the 12th point.

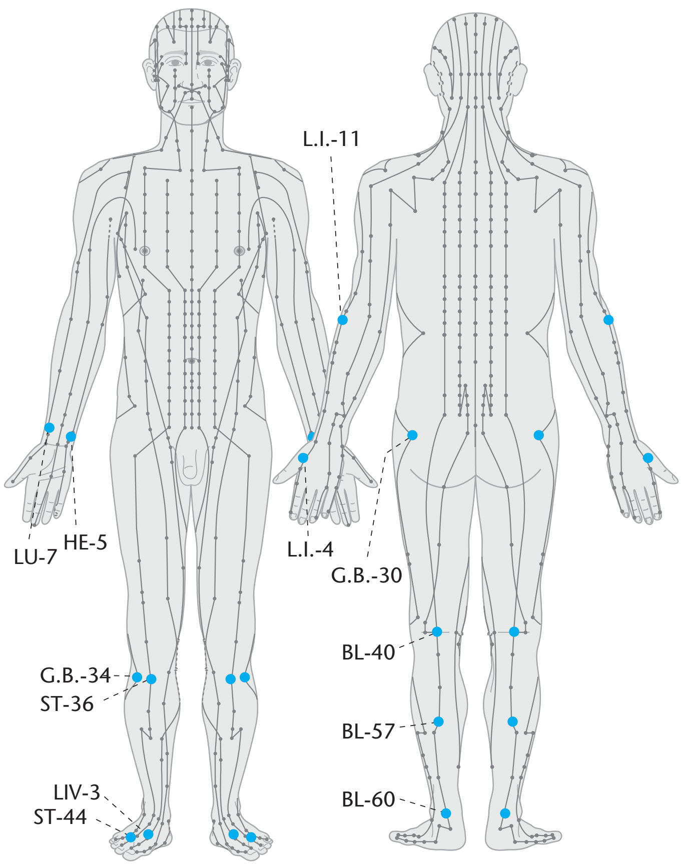

LU-7 (*lieque*)
ST-36 (*zusanli*)
BL-40 (*weizhong*)
G.B.-30 (*huantiao*)
L.I.-4 (*hegu*)
ST-44 (*neiting*)
BL-57 (*chengshan*)
G.B.-34 (*yanglingquan*)
L.I.-11 (*quchi*)
HE-5 (*tongli*)
BL-60 (*kunlun*)
LIV-3 (*taichong*)

8.1.15 The 13 Ghost Points of Sun Si Miao

Sun Si Miao, a famous physician of the Tang Dynasty, applied these 13 points (also called demon points) for treating disorders that today would be considered severe, manic, psychological disorders and/or epilepsy. Each *Sun Si Miao* Ghost point has an alternative Chinese name containing the component 'demon' or 'ghost' (see list according to Deadman et al. 2000). According to Deadman et al., some authors assumed that *guixin* is **LU-9** (*taiyuan*) rather than **P-7** (*daling*), while *guilu* is believed to be **P-5** (*jianshi*) or **P-8** (*laogong*) rather than **BL-62** (*shenmai*). There also exist different versions of this list of points; for example the physician Gao Wu omitted **BL-62, Du-23, Ren-1*** (in the original text of *Sun Si Miao: yumentou/yinxiafeng*) and **L.I.-11**, adding **Du-24, ST-17, LIV-2 and G.B.-34** instead.

	Name	Alternative name	Translation
Du-26	*renzhong*	*guigong*	Ghost Palace
Du-16	*fengfu*	*guizhen*	Ghost Pillow
Du-23	*shangxing*	*guitang*	Ghost Hall
Ren-24	*chengjiang*	*guishi*	Ghost Market
LU-11	*shaoshang*	*guixin*	Ghost Faith
L.I.-11	*quchi*	*guitui*	Ghost Leg
ST-6	*jiache*	*guichuang*	Ghost's Bed
SP-1	*yinbai*	*guilei*	Ghost Fortress
P-7	*daling*	*guixin*	Ghost Heart
P-8	*laogong*	*guicu*	Ghost Cave
BL-62	*shenmai*	*guilu*	Ghost Path
Ex-HN-11	*haiquan*	*guifeng*	Ghost Seal
Ren-1*	*huiyin*	*guicang*	Ghost Store

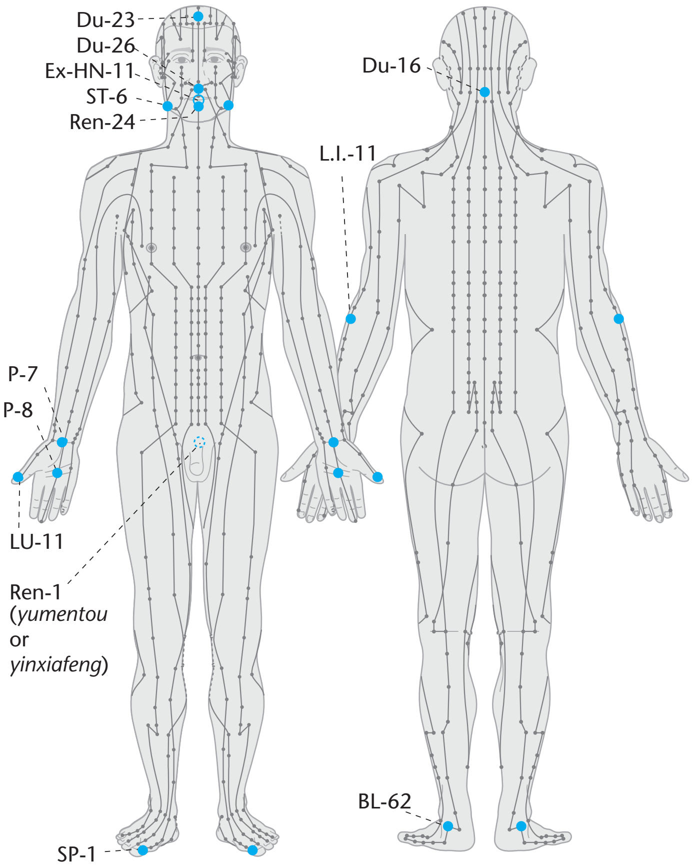

---

* *Yumentou* (in women) and *yinxiafeng* (in men): these two extra points approximately correspond to Ren-1 (*huiyin*/*guicang*) (Ghost Store).

8.1.16 Entry/Exit Points

While in modern Chinese literature entry/exit points are not mentioned as a separate point category, they are listed as such in some Western literature (for example Jarrett 2003, Hicks et al. 2004, Pirog 1996). However, Hicks et al. (→ 2004, p. 250) postulate that original Chinese sources exist for these points. The concept of entry/exit points is based on a continuous flow of Qi from one channel to the next, the sequence corresponding to the Organ clock (→ 1.1.4). Shunts are suggested in order to facilitate the flow of Qi from the end of one channel to the beginning of the next. At an entry point the Qi therefore flows from the preceding channel into the following one, while at the exit points it leaves the current channel in order to enter the next one. However, it should be noted that the entry/exit points do not always correspond to the first or last point on a channel. In the following table these points are marked by grey shading.

Channel (sequence according to Organ clock)	Time of maximum Qi flow		Entry/exit points in sequential order		
LU (LU-1–LU-11)	3–5am	↱	LU-1 (*zhongfu*)	➡	LU-7 (*lieque*)
					⬇
L.I. (L.I.-1–L.I.-20)	5–7am		L.I.-20 (*yingxiang*)	⬅	L.I.-4 (*hegxu*)
			⬇		
ST (ST-1–ST-45)	7–9am		ST-1 (*chengqi*)	➡	ST-42 (*chongyang*)
					⬇
SP (SP-1–SP-21)	9–11am		SP-21 (*dabao*)	⬅	SP-1 (*yinbai*)
			⬇		
HE (HE-1–HE-9)	11am–1pm		HE-1 (*jiquan*)	➡	HE-9 (*shaochong*)
					⬇
S.I. (S.I.-1–S.I.-19)	1–3pm		S.I.-19 (*tinggong*)	⬅	S.I.-1 (*shaoze*)
			⬇		
BL (BL-1–BL-67)	3–5pm		BL-1 (*jingming*)	➡	BL-67 (*zhiyin*)
					⬇
KID (KID-1–KID-27)	5–7pm		KID-22 (*bulang*)	⬅	KID-1 (*yongquan*)
			⬇		
P (P-1–P-9)	7–9pm		P-1 (*tianchi*)*	➡	P-8 (*laogong*)*
					⬇
T.B. (T.B.-1–T.B.-23)	9–11pm		T.B.-22 (*erheliao*)	⬅	T.B.-1 (*guanchong*)
			⬇		
G.B. (G.B.-1–G.B.-44)	11pm–1am		G.B.-1 (*tongziliao*)	➡	G.B.-41 (*zulinqi*)
					⬇
LIV (LIV-1–LIV-14)	1–3am	↳	LIV-14 (*qimen*)	⬅	LIV-1 (*dadun*)

---

* In women P-2 is often substituted for P-1 due to its anatomical location close to the breast.

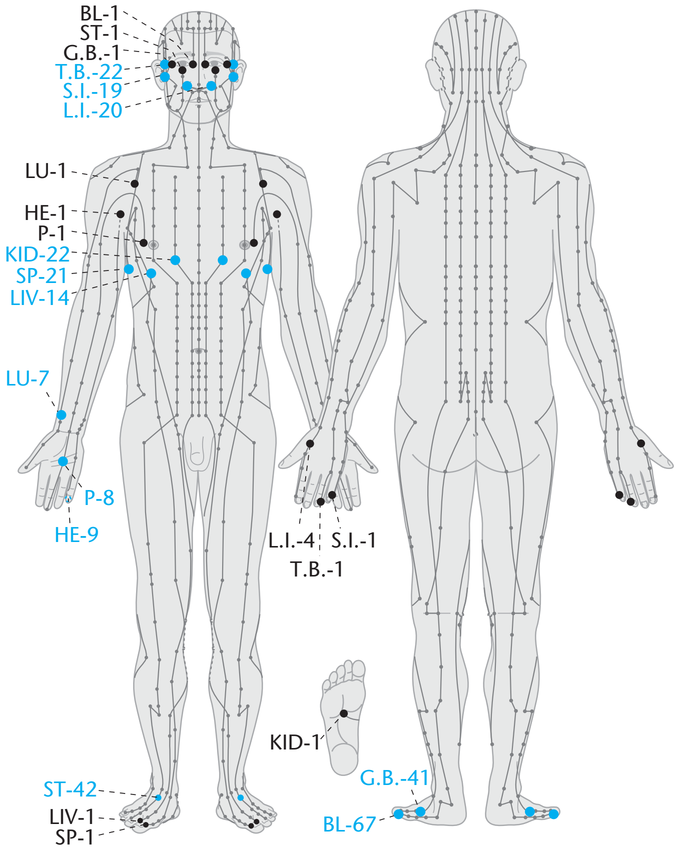
BL-1
ST-1
G.B.-1
T.B.-22
S.I.-19
L.I.-20
LU-1
HE-1
P-1
KID-22
SP-21
LIV-14
LU-7
P-8
HE-9
L.I.-4
S.I.-1
T.B.-1
KID-1
G.B.-41
ST-42
LIV-1
SP-1
BL-67

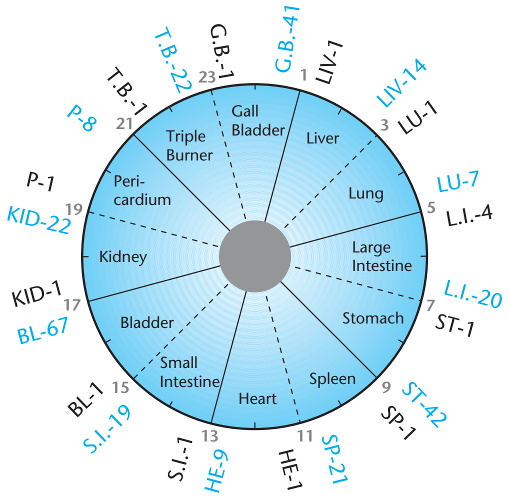
G.B.-1
G.B.-41
T.B.-22
LIV-1
LIV-14
T.B.-1
LU-1
P-8
23
1
21
3
Gall Bladder
Triple Burner
Liver
Peri-cardium
Lung
P-1
KID-22
19
LU-7
5
L.I.-4
Large Intestine
Kidney
KID-1
17
BL-67
L.I.-20
7
ST-1
Stomach
Bladder
Small Intestine
Spleen
Heart
BL-1
15
S.I.-19
9
ST-42
SP-1
13
11
S.I.-1
HE-9
HE-1
SP-21

Entry/exit block and therapy

Pulse diagnosis can reveal the partial or complete block between entry and exit points. An entry/exit block will inhibit the Qi flow from one channel to the next. In order to restore the blocked flow, the exit point of the blocked channel and the entry point of the following channel are needled at the same time. According to Hicks et al. (2004), further exit and entry points preceding or following the affected channel can be added in order to regulate the Qi flow over a greater distance. Needling should always be **bilateral**, even if the disorder occurs only on one side of the body. If it is not only the shunt between two channels, but the circulation in a channel as a whole that is affected, the flow in the whole channel can be strengthened and stimulated by simultaneous needling of both its entry point and its exit point.

Pirog (1996) compares the entry/exit points to input and output valves regulating the flow in a series of pipes (the channels) by 'opening' or 'closing' them.

- Tonifying needling techniques at the entry point (opening the input valve) will increase the flow into the channel. For example: tonifying **LU-1** opens the input valve. This will increase Qi flow in the LU channel for deficiency in this channel.
- Tonifying needling techniques at the exit point (opening the output valve) will decrease the flow in the channel. For example: Tonifying **LIV-14** opens the output valve: excess in the LIV channel will be drained.
- Reducing needling techniques at the entry point (closing the input valve) will reduce the flow into the channel. For example: Reducing **LU-1** closes the input valve and reduces the flow from the LIV channel into the LU channel, for example if there is pre-existing excess in the LU channel.
- Reducing needling techniques at the exit point (closing the output valve) will strengthen the flow in the channel. For example: Reducing **LIV-14** closes the output valve and prevents a further loss of Qi if there is a pre-existing deficiency in the LIV channel.

8.2 Strategies for Point Selection

8.2.1 Local, Adjacent and Distal Points

Local points

Local points are located in the immediate vicinity of the affected region. Each point that is tender with pressure (*ashi* point) can be considered to be a local point.

Adjacent points

Adjacent points are located near the affected or painful region.

Distal points

Distal points, despite being located at a distance from the affected area, still have a therapeutic effect, either directly or by being connected to the affected area by a channel (→ 1, also → 8.2.2, selection of distal points). The most effective and dynamic distal points are located distal to the knee (for the leg) or the elbow (for the arm). Generally the distal points of the foot channels tend to be more dynamic and effective than those on the arm channels.

Needling technique: In acute conditions or with severe pain the distal points should be needled with a reducing technique in order to activate the Qi flow more strongly in the blocked channels.

Stimulation of distal points: For a limited range of motion accompanied by pain the relevant distal point is strongly stimulated with a reducing technique while the patient moves the affected joint.

Enhancing the therapeutic effect: The therapeutic effect can be enhanced by combining distal points on the hand channels with those on the foot channels or by combining distal points with relevant local and adjacent points (→ 8.3.1).

8.2.2 Point Selection According to Affected Channel

Here the method of selecting points is based on the channel system (→ 1):

- Selecting points on the channel which **traverses the affected area**
- Selecting points on the channel that is **Interiorly–Exteriorly pertaining** (→ 1.2) to the affected channel
- Selecting points from the channel which is connected to the affected channel by **hand–foot pairing** (for example, *taiyang* axis, etc. → 1.2.3)

Differential selection of distal points

The following possibilities for selecting distal points are, among others, based on course material about the ‘one point therapy’ by R. Thambirajah (1990, 1991), R. Tan (2003) and my own clinical experience.

Selection of distal points according to corresponding areas

The distal points can be selected according to corresponding areas. The affected area is carefully examined, then the corresponding area is needled contralaterally (→ Fig.).

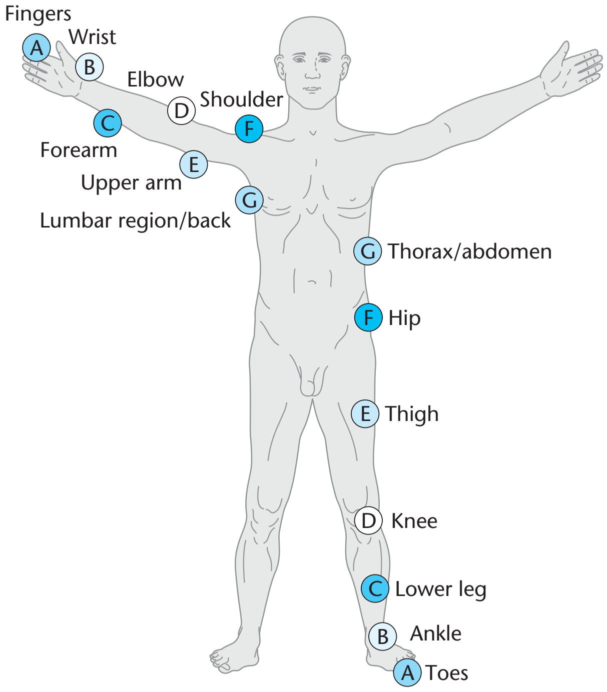

Selection of distal points according to Interiorly–Exteriorly connected channels

Here points are selected on the Interiorly–Exteriorly connected channel (Yin/Yang). The affected area is carefully located, then the corresponding point (mirror point) is needled with reducing technique on the contralateral side on the Interiorly–Exteriorly pertaining channel. Another specific example for this method of point selection is the *yuan–luo* combination (→ 8.3.3).

Example: For a disorder affecting the ST channel (foot *yangming*) select a point on the Interiorly–Exteriorly pertaining SP channel (foot *taiyin*). For example, for knee pain at **ST-35**, needle **SP-9** contralaterally (→ Fig.).

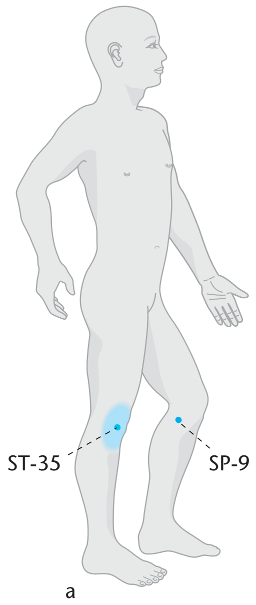

Selection of distal points according to the six great channels (*liu jing*, hand–foot pairing, → 1.2.3)

The point selection with this method is based on the six great channels (hand–foot pairing: *taiyin, taiyang*, etc., → 1.2.3). Disorders affecting the area of the hand channel are therefore treated with points on the corresponding foot channel and vice versa.

Example: Disorders in the area of the hand *taiyin* channel (LU channel) can be balanced by contralateral needling of points on the foot *taiyin* channel (SP channel). For example, shoulder pain with limited range of motion and maximum pain near **LU-1** (*zhongfu*) can be improved by contralateral needling of **SP-9** (*yinlingquan*) with reducing technique.

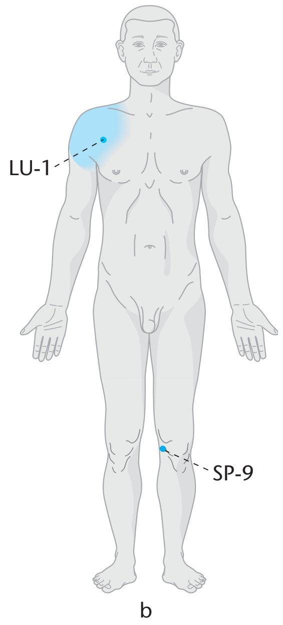

Selection of distal points according to the Organ clock (→ 1.4)

Distal points are selected on the channel that lies chronologically opposite to the affected channel on the Organ clock (→ 1.4). With this method, hand channels will treat disorders affecting the foot channels and vice versa.

Example: For a disorder affecting the G.B. channel (foot *shaoyang*), needle a point on the HE channel (hand *shaoyin*) located opposite the G.B. channel on the Organ clock. Needling can be contralateral or ipsilateral.

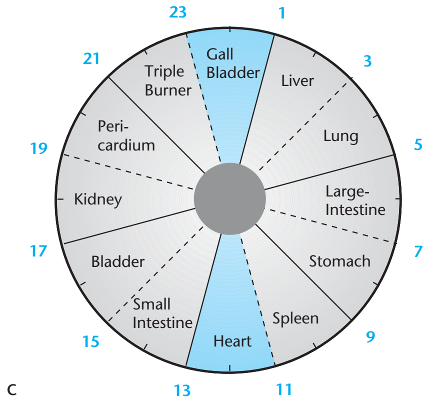

8.2.3 Point Selection Based on Symptoms

Points can also be selected based on traditional or empirical therapeutic experiences. While this method cannot replace a root treatment based on a differential diagnosis, it can be very helpful in situations where quick action is required.

Examples:

- **Du-14** or **L.I.-11** for fever
- **ST-36, Ren-12** for acute gastrointestinal disorders

8.2.4 Point Selection Based on the Qi Flow of the Organs

Descend counterflow Qi:

- Stomach: **Ren-10, Ren-13, ST-34, ST-44, ST-45, L.I.-4**
- Lung: **LU-1, LU-5, LU-7**
- Heart: **HE-5, HE-8, Ren-15**
- Kidney: **KID-1, KID-7, Ren-4**
- Liver: **LIV-2, LIV-3, LIV-1, LIV-14**

Raise Qi:

Spleen: **Ren-6, Ren-12, Du-20, BL-20**

Ch08-F10028.qxd 2/23/08 4:41 PM Page 688

8.2.5 Point Selection According to the Five Phases (Elements)

While the Five Phase points are identical to the *shu*-transporting points in location, their function and application differ from the latter (for more detail → 8.1.6). There exist various very specific methods regarding diagnosis, point selection and therapy based on the Five Phases, which are practised by several different schools of thought. Within the framework of this Atlas, only a simplified overview of the possible clinical applications is listed below.

***Shu*-transporting points of the Yin channels**

Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9 +	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2 −	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**

Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43 +	G.B.-41	G.B.-38 −	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; − , sedation point

Cycles

The Five Phases influence each other both physiologically (*sheng* and *ke* cycle) as well as pathologically (*cheng*, *wu* and to some extent also *sheng* cycle). If the balance between the Five Phases is disturbed, pathological syndromes will appear.

- *Sheng* cycle (generating cycle, mother–son cycle): a phase generates and nourishes the following phase. Each phase nourishes (as mother) and is nourished (as son) at the same time. Fire nourishes Earth, Earth nourishes Metal, Metal nourishes Water, Water nourishes Wood, and Wood nourishes Fire. In a pathological condition either the mother is too weak to nourish the son sufficiently or the son is too strong, draining his mother and weakening her.
- *Ke* cycle (controlling cycle): one phase controls another and is itself controlled by yet another phase.
- *Cheng* cycle (overacting cycle): The controlled phase is pathologically suppressed or weakened.
- *Wu* cycle (insulting cycle): A phase is pathologically stronger than its controlling phase.

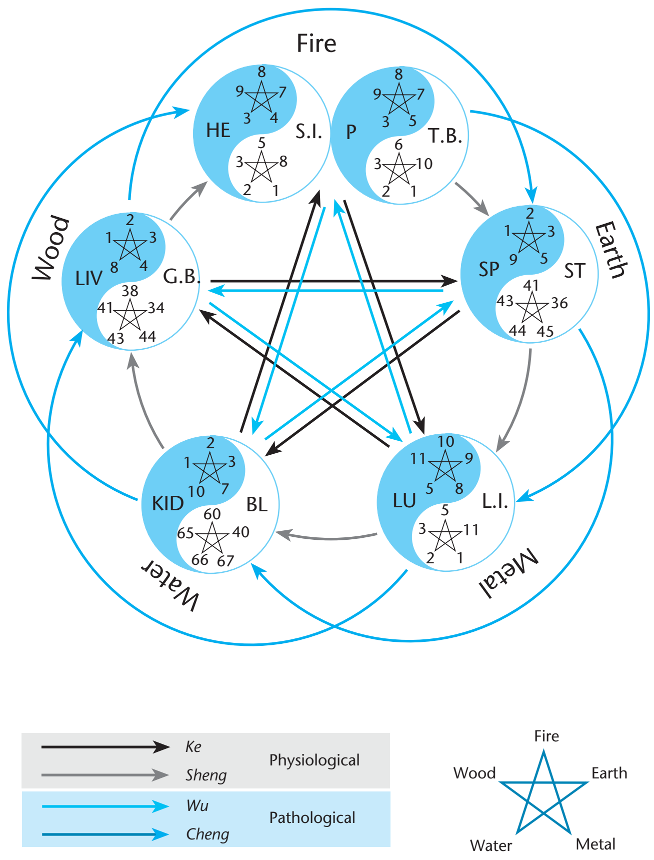

Ch08-F10028.qxd 2/23/08 4:41 PM Page 689

Simple point selection according to the *sheng* cycle

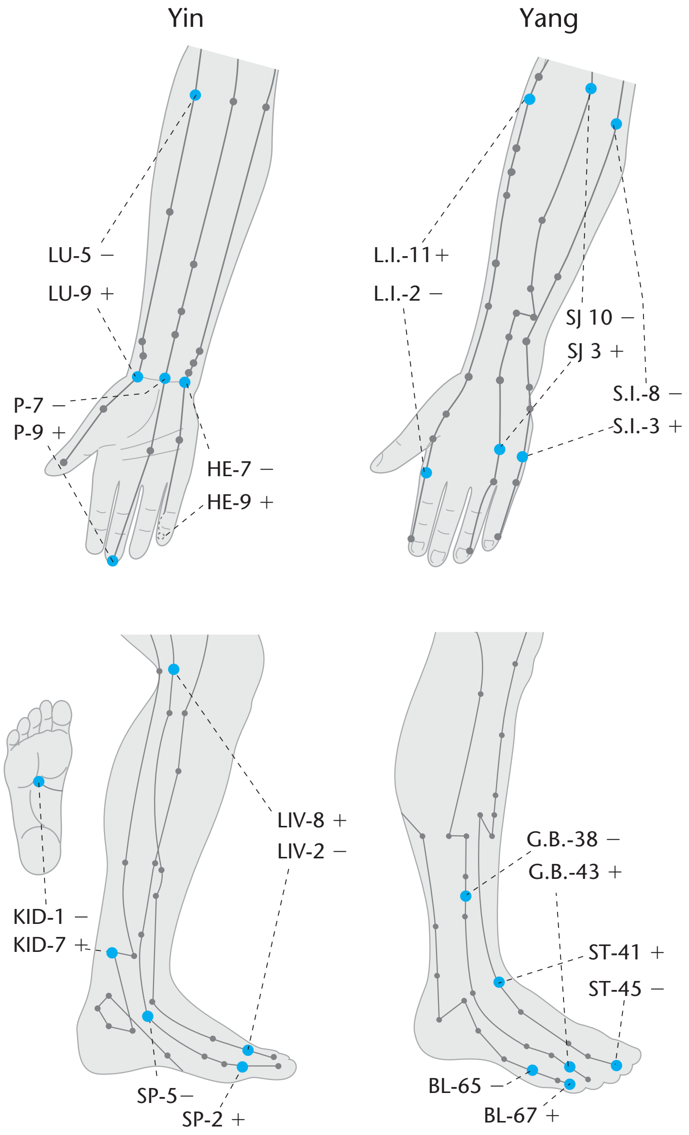
Yin
Yang
LU-5 –
LU-9 +
P-7 –
P-9 +
HE-7 –
HE-9 +
L.I.-11 +
L.I.-2 –
SJ 10 –
SJ 3 +
S.I.-8 –
S.I.-3 +
LIV-8 +
LIV-2 –
KID-1 –
KID-7 +
SP-5 –
SP-2 +
G.B.-38 –
G.B.-43 +
ST-41 +
ST-45 –
BL-65 –
BL-67 +

Treatment principle: For deficiency tonify the mother, for excess drain (sedate) the son. Mother–son points are identical to the tonification and sedation points of a channel. The correct needling technique is important in order to obtain the desired result:

- tonifying method for tonification points
- reducing method for sedation points

Example for the clinical application of the *sheng* cycle

For deficiency: In cases of deficiency of a particular channel/Organ, tonify that point on the affected channel that corresponds to the phase of the mother.

Example: Earth is the mother of Metal. If the Lung (Metal) is deficient, needle the Earth point (**LU-9**) on the Lung channel with tonifying technique.

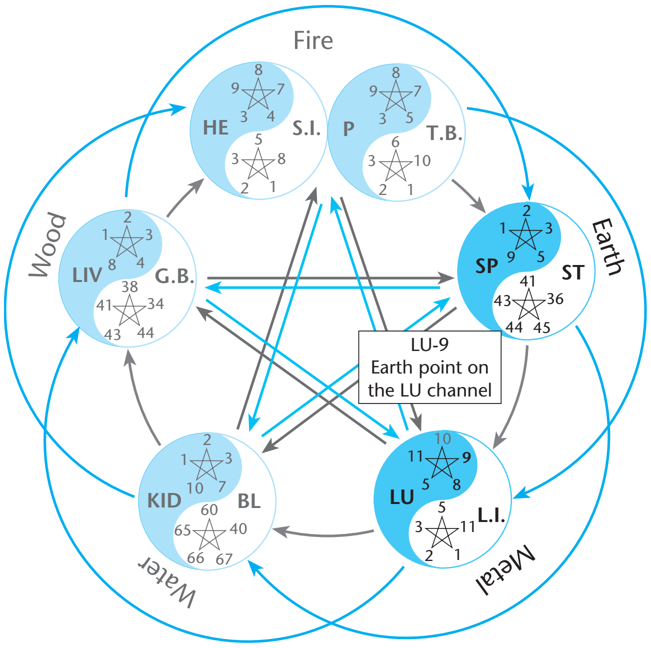
Fire
Earth
Metal
Water
Wood
HE
S.I.
P
T.B.
SP
ST
LU
L.I.
KID
BL
LIV
G.B.

LU-9
Earth point on
the LU channel

For excess: If there is excess in a particular channel/Organ, select that point on the channel which corresponds to the phase of the son and needle it with reducing technique.

Example: Water is the son of Metal. If the Lung (Metal) is in excess, needle the Water point (LU-5) on the Lung channel with reducing technique.

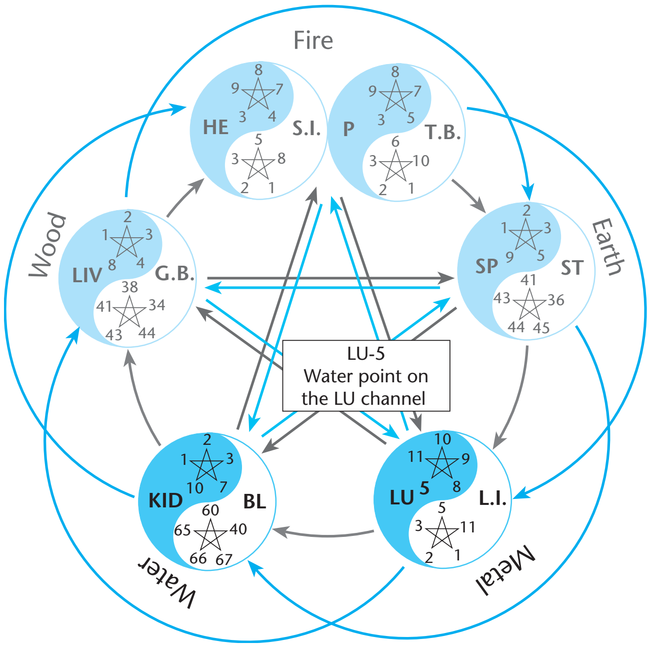
Fire
Earth
Metal
Water
Wood
HE
S.I.
P
T.B.
SP
ST
LU
L.I.
KID
BL
LIV
G.B.

LU-5
Water point on
the LU channel

***Addendum*:** The terms tonification/sedation points are somewhat controversial, since the properties of a point can often be overshadowed by its other features. For example, according to the Five Phase theory, **P-9** and **HE-9** are considered tonification points, since they correspond to the phase of the mother. However, in clinical practice they are much more frequently used as *jing-*well points for acute situations – thus they are considered from the point of view of the Qi flow in the channels, especially in order to drain Heat.

Application of the Five Phase points for the elimination of external pathogenic factors

A connection exists between the Five Phases and pathogenic factors (Maciocia 1994; according to Ross 1998, however, these points should only be used for internal disorders):

- Wood corresponds to Wind
- Fire corresponds to Heat
- Earth corresponds to Dampness
- Metal corresponds to Dryness. According to Maciocia this correlation presents an exception: the Metal point should not be needled to treat Dryness; a moistening treatment is here the preferred choice.
- Water corresponds to Cold.

In accordance with these interrelations, Five Phase points can be applied to eliminate pathogenic factors.

Example: Acute sore throat with fever, redness and swelling of the throat due to Wind-Heat

Affected phases: Wood (Wind), Fire (Heat)

Affected Organ/channel: Lungs

Therapy: Needling of the Wood point on the LU channel (**LU-11**) and the Fire point on the LU channel (**LU-10**).

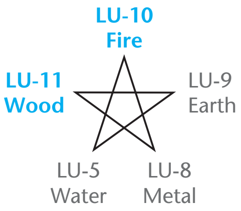

The application of the ***ke* cycle (controlling cycle)** allows a more differentiated approach in the therapy and elimination of pathogenic factors:

- On the Yin channel, needle the point corresponding to the pathogenic factor with reducing technique (for example, for Heat needle the Fire point)
- In addition, needle the point which corresponds to the controlling phase (according to the *ke* cycle/controlling cycle → 8.2.5) of the pathogenic factor on the paired Yang channel with tonifying technique.

Ben points

***Shu*-transporting points of the Yin channels**

Phase/ channel	Wood point	Fire point	Earth point	Metal point	Water point
LU	LU-11	LU-10	LU-9 +	LU-8	LU-5 −
SP	SP-1	SP-2 +	SP-3	SP-5 −	SP-9
HE	HE-9 +	HE-8	HE-7 −	HE-4	HE-3
KID	KID-1 −	KID-2	KID-3	KID-7 +	KID-10
P	P-9+	P-8	P-7 −	P-5	P-3
LIV	LIV-1	LIV-2−	LIV-3	LIV-4	LIV-8 +
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

***Shu*-transporting points of the Yang channels**

Phase/ channel	Metal point	Water point	Wood point	Fire point	Earth point
L.I.	L.I.-1	L.I.-2 −	L.I.-3	L.I.-5	L.I.-11 +
ST	ST-45 −	ST-44	ST-43	ST-41 +	ST-36
S.I.	S.I.-1	S.I.-2	S.I.-3 +	S.I.-5	S.I.-8 −
BL	BL-67 +	BL-66	BL-65 −	BL-60	BL-40
T.B.	T.B.-1	T.B.-2	T.B.-3 +	T.B.-6	T.B.-10 −
G.B.	G.B.-44	G.B.-43+	G.B.-41	G.B.-38−	G.B.-34
Channel Qi flow	*Jing*-well point	*Ying*-spring point	*Shu*-stream point	*Jing*-river point	*He*-sea point

+, tonification point; –, sedation point

Ben points (root points, element points) represent the same phase as its pertaining channel. **Example:** The Spleen corresponds to Earth; thus the Earth point on the Spleen channel is its *ben* point.

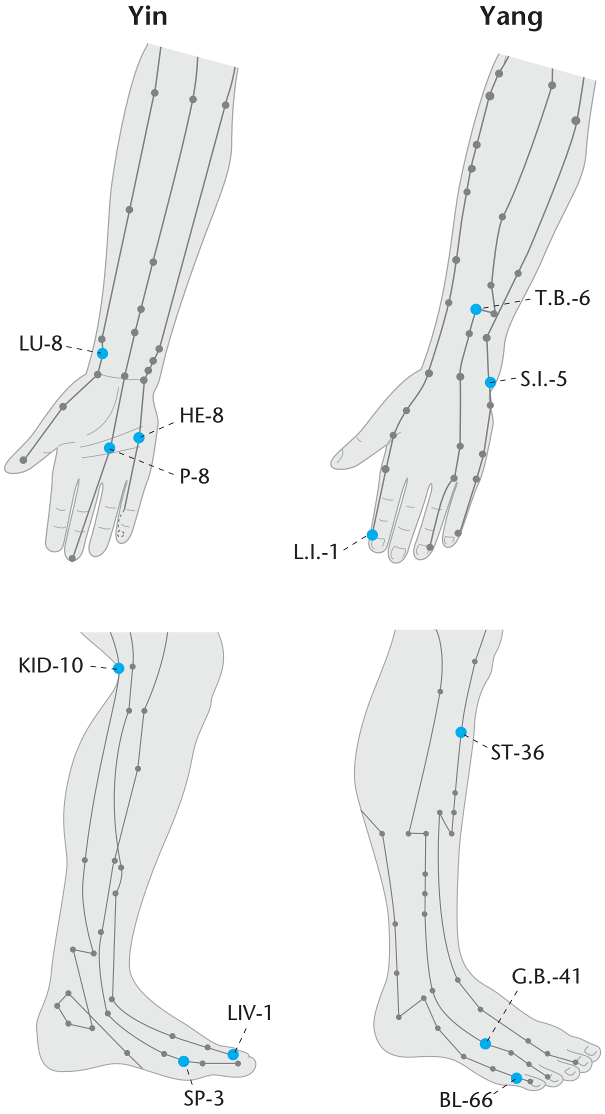

Clinical application

- **For deficiency conditions:** needling the *ben* point with tonifying technique will tonify its corresponding Organ or channel
- **For excess conditions:** needling the *ben* point with reducing technique will drain the excess from its corresponding Organ or channel
- **Spiritual aspect:** the *ben* points on the Yin channels also affect the spiritual aspect of their corresponding phases. Thus the LU *ben* point **LU-8** allows the *po* (corporeal soul) to unfold its potential. Similarly the HE *ben* point **HE-8** will empower the *shen* (spirit), the SP *ben* point **SP-3** the *yi* (thinking), the LIV *ben* point **LIV-1** the *hun* (ethereal soul) and the KID *ben* point **KID-10** the *zhi* (will).

8.3 Point Combinations

8.3.1 Combining Local and Distal Points

Point selection and combination can be carried out according to affected areas and channels (➞ 8.2.1, ➞ 8.2.2).

For disorders affecting the channel: For example, for *bi*-syndromes affecting the sinews and joints, distal points can be stimulated first. Then local and adjacent points can be selected based on tenderness. These are needled and/or subsequently cupped.

For disorders of the *zangfu*-Organs: Generally, only distal points are applied in acute conditions. Once the condition (for example, acute pain) has improved, local points can be selected. In chronic conditions the combination of local and distal points – and especially Back-*shu* and Front-*mu* points as adjacent points (➞ 8.3.2) – is commonly used. Another possibility is the combination of the opening point of an extraordinary vessel (➞ 8.1.8) with local points in the affected area.

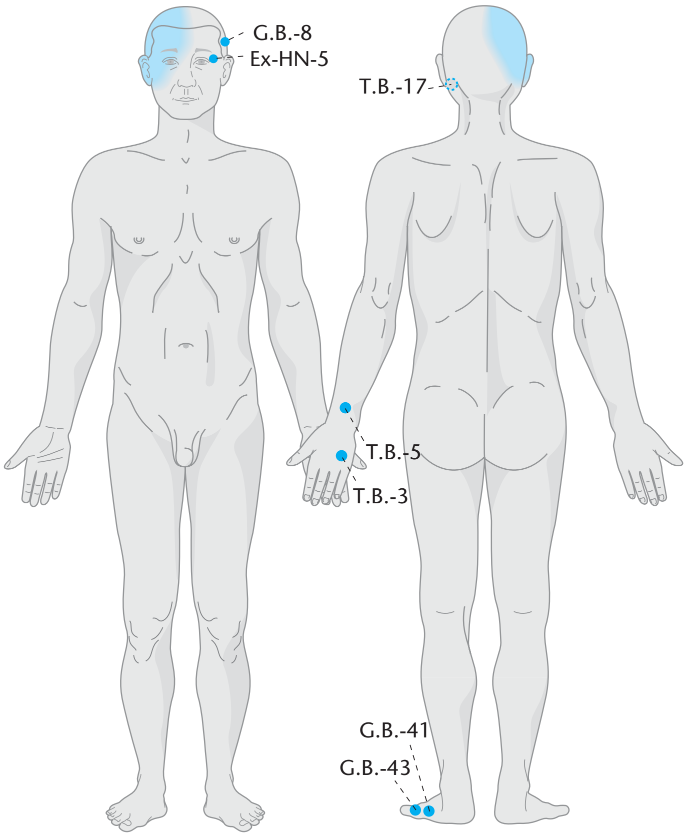

8.3.2 Combining Points on the Front and Back of the Body

Here, points on the anterior and posterior aspect of the body are combined:

- **Anterior points**: mainly located on the thorax and abdomen
- **Posterior points**: mainly located on the back and lumbar region

Back-*shu*/Front-*mu* method

The combination of the Back-*shu* point (→ 8.1.4) and the Front-*mu* point (→ 8.1.5) of the affected Organ is the most important way of combining points on the front and the back of the body. The Back-*shu*/Front-*mu* method enhances and extends the therapeutic effect, compared to using either of those points only. It has a strong Yin/Yang balancing effect and it is particularly beneficial for chronic disorders of the *zangfu*-Organs.

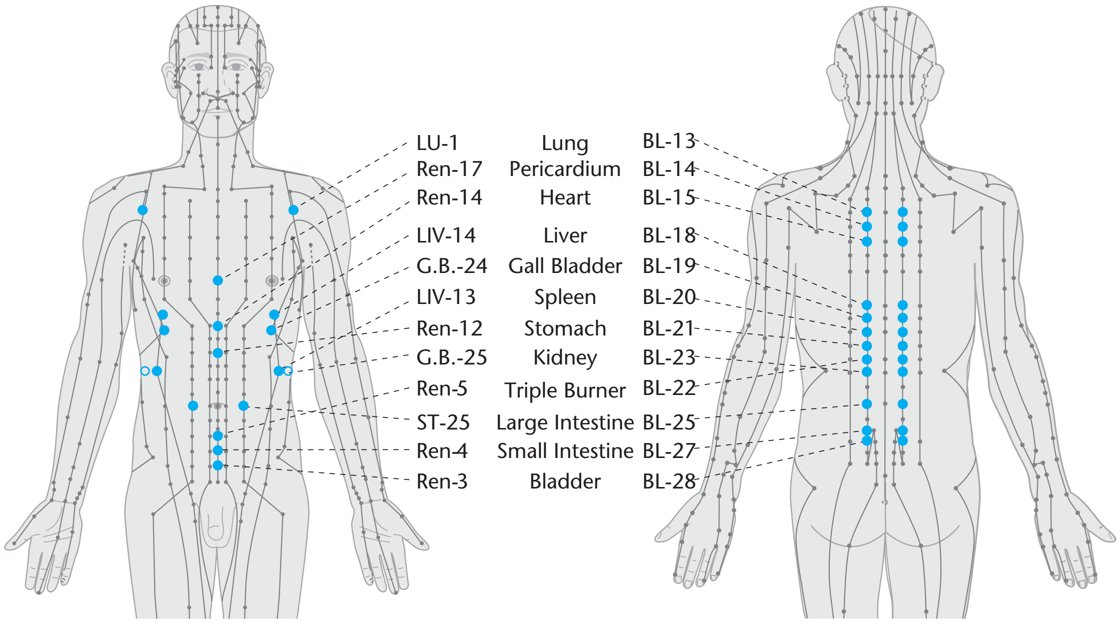
LU-1 Lung BL-13
Ren-17 Pericardium BL-14
Ren-14 Heart BL-15
LIV-14 Liver BL-18
G.B.-24 Gall Bladder BL-19
LIV-13 Spleen BL-20
Ren-12 Stomach BL-21
G.B.-25 Kidney BL-23
Ren-5 Triple Burner BL-22
ST-25 Large Intestine BL-25
Ren-4 Small Intestine BL-27
Ren-3 Bladder BL-28

Clinical application of the Back-*shu*/Front-*mu* method

- Needling of the Back-*shu* and Front-*mu* points during one session
- For repeated treatments within a short period of time, alternate needling of the Back-*shu* and Front-*mu* points (for example, during the first treatment select the relevant Back-*shu* point, in the next treatment select the corresponding Front-*mu* point, etc.)
- The Back-*shu*/Front-*mu* combination can also be applied to balance an incorrect treatment. For example, if the needles have been retained too long in the Back-*shu* points and the patient is tired, Front-mu points can be needled to restore balance.

Balancing the *du mai* and the *ren mai*

The points on the *ren mai* are located on the anterior aspect of the body, while those of the *du mai* are predominantly located on the posterior aspect. Combining points on the *ren mai* and *du mai* has a balancing affect on the Yin and Yang and can regulate the ascending and descending flow of Qi. This combination therefore not only balances the front and the back of the body, but also the Yin and the Yang. Furthermore, it has a strong effect on the psyche and, depending on the needling technique and point selection, will have a calming or stimulating effect (figure modified according to Ross 1998).

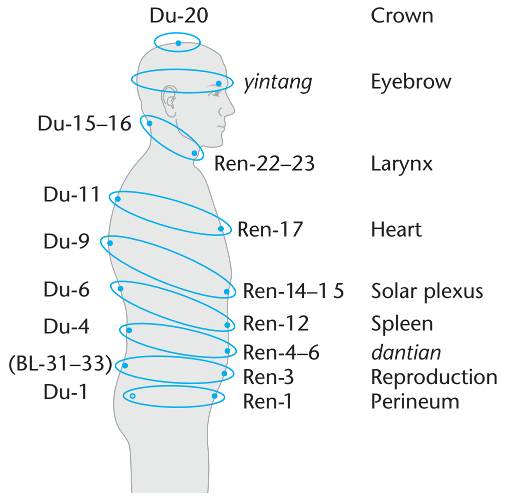

8.3.3 Yin–Yang Combination

Synonym: Interior/Exterior combination
The Yin–Yang combination balances the flow of Yin and Yang in the channels. It is based on the concepts of channel energetics (➞ channel circuits, Fig. 1.12).

Needling too many points on the Yang channels can make the patient nervous and restless. In this case, needling points on the Yin channels can restore balance and have a calming effect.
Needling too many points on the Yin channels can lead to tiredness. In this case, needling points on the Yang channels has a balancing, stimulating effect.

Balancing Yin and Yang in paired channels

- Points on the Interiorly–Exteriorly paired channels are combined in order to balance Yin and Yang. For example the LU and L.I. channels are the Interiorly–Exteriorly paired channels of the hand, while the ST and SP channels are the Interiorly–Exteriorly paired channels of the foot of the first circuit.
- This method increases the therapeutic effect compared to using only Yin or only Yang points.

Yuan/luo combination

Most important Yin–Yang combination (Synonym: host–guest combination): The *yuan*-source point (host) of the primarily affected channel/Organ is combined with the *luo*-connecting point of the Interiorly–Exteriorly paired channel.

Important *yuan/luo* combinations

- **L.I.-4** (*hegu*) and **LU-7** (*lieque*) support the descending function of the Lung Qi, expel pathogenic factors, tonify the Defensive Qi (*wei qi*) and calm the *shen*
- **SP-3** (*taibai*) and **ST-40** (*fenglong*) tonify the Spleen and transform Phlegm
- **LIV-3** (*taichong*) and **G.B.-37** (*guangming*) benefit the eyes for Liver syndromes
- **T.B.-4** (*yangchi*) and **P-6** (*neiguan*) regulate the Triple Burner, move Liver Qi, calm the *shen*, relax the muscles in the occiput and shoulders.

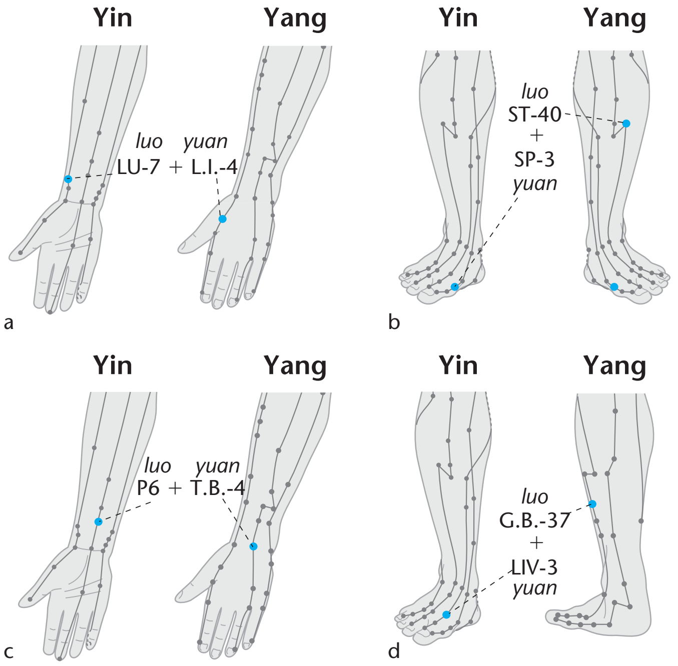

Further methods of combining points of Interiorly–Exteriorly paired channels and modified *yuan/luo* combinations

a Combining the *luo*-connecting points of Interiorly–Exteriorly paired channels tends to enhance the therapeutic effect; for example combine **ST-40** and **SP-4** for abdominal pain

b For chronic disorders, combine the *yuan*-source point and the *luo*-connecting point of the same Yin channel in order to increase the therapeutic effect; for example needle **LU-9** and **LU-7** for chronic cough.

c For disorders affecting a Yin channel, only needle the *yuan*-source point of the Interiorly–Exteriorly paired Yang channel. For example for an acute cold (a Lung syndrome), needle only **L.I.-4** (*yuan*-source point).

d For unilateral musculoskeletal disorders affecting a Yang channel: in addition to local and adjacent points on the affected channel, needle the contralateral *luo*-connecting point of the Interiorly–Exteriorly paired Yin channel. This method, which promotes balance between Yin and Yang, is mainly used for channel-pertaining disorders. For example: for unilateral arm or shoulder pain along the L.I. channel, needle points on the L.I. channel on the affected side as well as **LU-7** on the opposite side of the body.

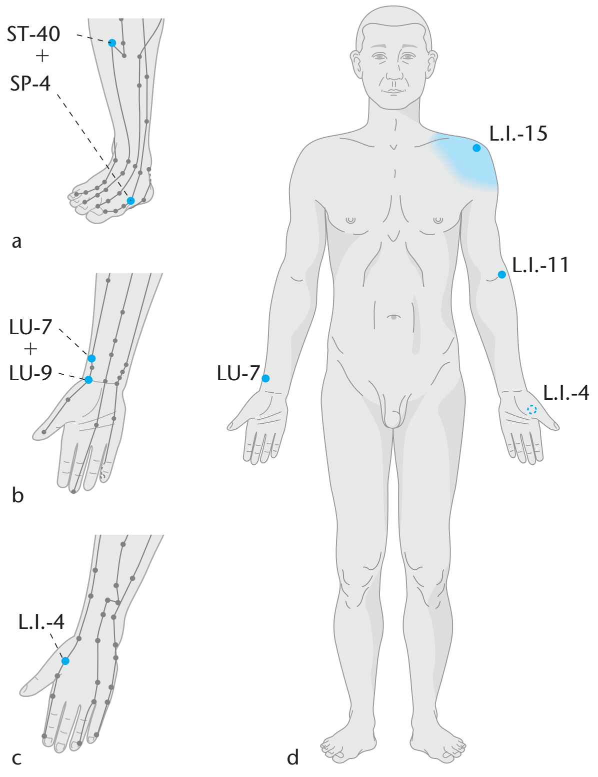

Yin–Yang balance for unpaired channels

Examples:

ST-36 + P-6	Harmonises the Middle Burner and descends counterflow Qi
L.I.-4 + LIV-3	Most important combination for regulating the Qi flow in the whole body

8.3.4 Combining Points from Above and Below

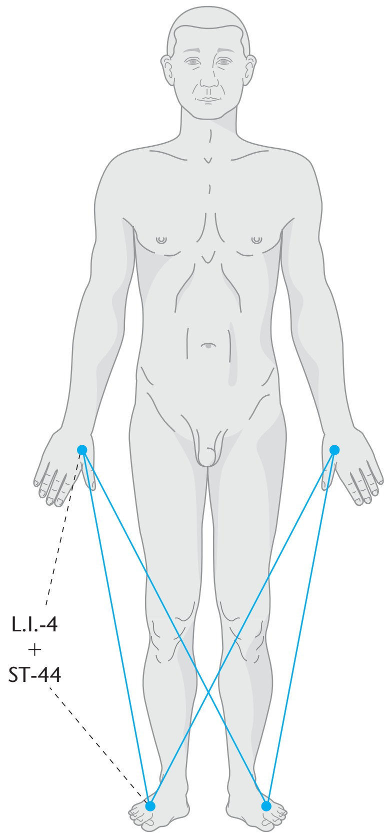

- A balance between the **upper and lower** half of the body is achieved by selecting points that are evenly distributed. For headaches, therefore, **L.I.-4** on the upper extremity is combined with **ST-44** on the lower extremity.
- Combining points above with points below **promotes the smooth flow of Qi**. For example: combining the opening point and coupled point of an extraordinary vessel (→ 8.1.8).

When points from above and below should not be combined

- for acute disorders of the joints and the back – here only distal points should be needled with the reducing method.
- if there is a pre-existing energetic disharmony between above and below. For example: for Heat sensations of the head (for example due to Kidney Yin deficiency with Empty Fire) needle **KID-1** in order to cause the Qi to descend. On the other hand, for a uterine prolapse use moxibustion **on Du-20** in order to raise the Qi.

8.3.5 Left–Right Combination

Often the balance between the right and left half of the body can be restored through bilateral needling of the relevant points. This will increase the therapeutic effect as opposed to unilateral needling.

Clinical application for Interiorly–Exteriorly paired channels (→ 1.2.3)

- For acute unilateral channel disorders caused by the invasion of pathogenic factors, combine **local points** on the affected side with the ***luo*-connecting point** on the contralateral side.
- For chronic unilateral channel disorders caused by the invasion of pathogenic factors, combine **local points** on the affected side with the *luo*-**connecting point** of the **Interiorly–Exteriorly paired channel** on the **contralateral** side (with tonifying technique).

Indications for unilateral and/or contralateral needling

- **For acute, painful disorders:** contralateral or diagonal needling (corresponding hand and foot channel, also see → 8.2.1 and → 8.2.2)
- **For chronic joint disorders:** ipsilateral needling of several points balanced by needling of contralateral points.
- **In paediatric acupuncture** (use fewer needles)
- **For ongoing treatment sessions:** occasional balancing contralateral needling of the healthy side of the body

8.3.6 Chain and Lock Point Association Method

Two, three or more points on the same channel are selected and needled one after the other (in a row). This method is predominantly applied for disorders of the musculoskeletal system or the nervous system.

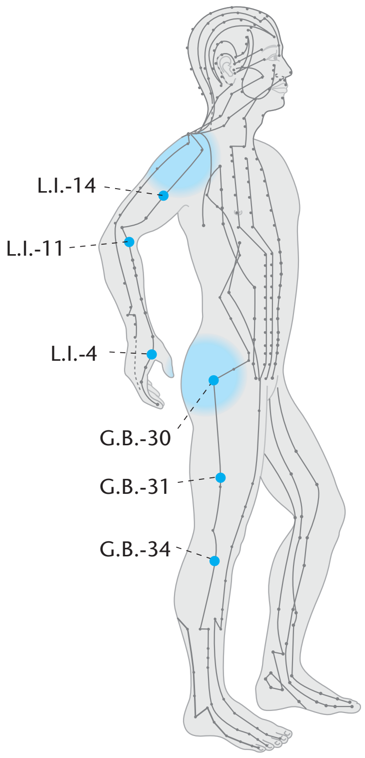

8.3.7 Point Selection According to the Organ Clock

Within a 24-hour cycle each channel experiences a two-hour period of maximum Qi flow (→ Fig. 1.17, Chinese Organ clock). Some modern (Western) schools extrapolate the circadian Qi flow to the Organs (internal channel pathways and divergent channels). If one Organ is at its maximum period, the Organ on the opposite side of the clock will be at its minimum period, in other words the Qi flow will at that time be at its lowest. Symptoms of a disorder occurring at a particular time may provide a hint about the Organ affected, which at that time will be during its **maximal period in excess syndromes** or during its **minimal period in deficiency syndromes**.

For example: Always waking up between 1 and 3 am can be a sign of a Liver syndrome. Thus the tonification point and/or *luo*-connecting point of the channel opposite to the Liver on the Organ clock could be selected, in this example **S.I.-3** (*houxi*, tonification point) or **S.I.-7** (*zhizheng*, *luo*-connecting point).

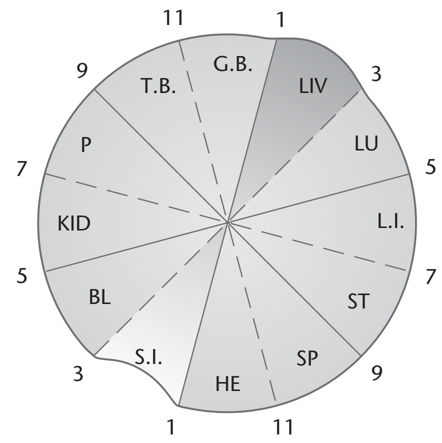

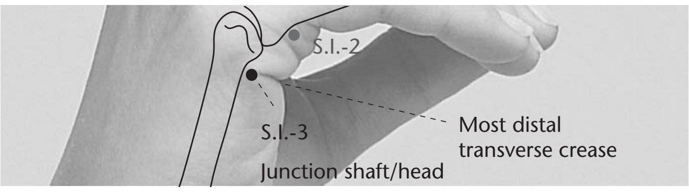

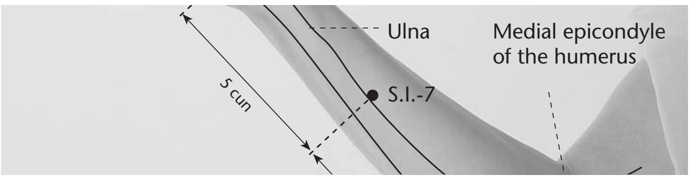

Addendum: There are some very specific methods for chrono-acupuncture, for exampel the *zi wu liu zhu* method and the *ling gui ba fa* method, where points are needled at specific times (→ Kubiena and Ramakers 2002).

8.3.8 Clinical Application

The following illustration shows the clinical application of several point selection methods (8.3.1 – 8.3.5) for the treatment of dysmenorrhoea.

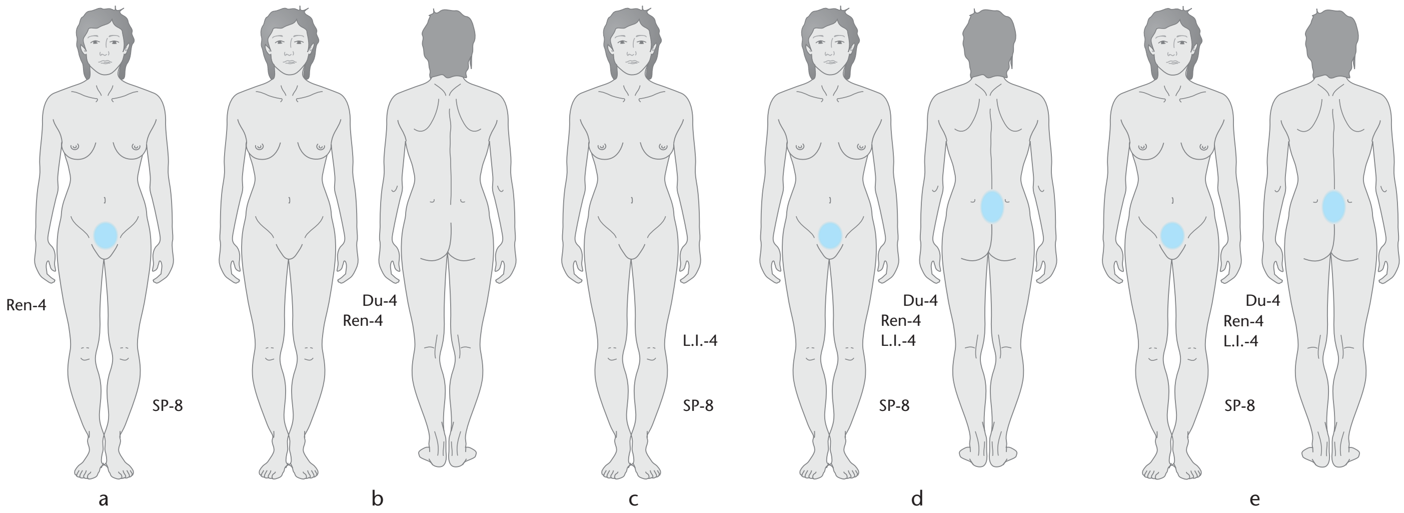

Possible point combinations for the treatment of dysmenorrhoea:

a. combining local and distal points (➞ 8.3.1)
b. combining points on the front and the back of the body (➞ 8.3.2) and Yin–Yang (➞ 8.3.3)
c. Yin–Yang combination (➞ 8.3.3) and above–below (➞ 8.3.4)
d. Combination of all of the above methods (➞ 8.3.1–8.3.4)
e. Additionally combining points on the left and right side of the body (➞ 8.3.5)

9 Scientific Research

Ingolf Hosbach

9.1 Introduction

The aim of this chapter is to give a comprehensive overview about the scientifically proven or disproved effectiveness of certain points. It is not the aim of this chapter to prove the effectiveness of acupuncture as such. Based on the latest research, and 1000 recent human studies with mostly positive results, it is assumed that acupuncture is effective, despite criticism. If applied properly it is a well-established, low-risk method with minimal side effects.

Accordingly, the scope of acupuncture research has developed from simple questions, such as whether acupuncture works, to more specific questions such as which acupuncture points can be used for which conditions and with which stimulation methods. Furthermore, studies are increasingly designed for precise diagnoses and treatments based on the principles of Traditional Chinese Medicine. By the same token, the quality of the biomathematical methods used for the design and analysis of the studies is increasing. The number of randomised, controlled studies (RCTs) is also on the rise (Fig. 1).

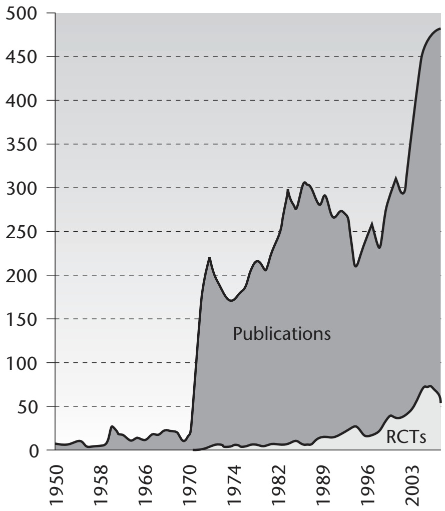

Fig. 1: Number of all publications and RCTs per year between 1954 and 2005. (Source: PubMed, US National Library of Medicine, Bethesda, Maryland)

The percentage of RCTs of all publications is nearly 7%, a figure that compares well with most conventional forms of therapy (for example: arthroscopy 5%, beta-blockers 7%, pain medication 7%, lowering cholesterol medication 8–11%).

Fundamental principles for scientific assessment

One of the prerequisites for including a study in this chapter was the mention of specific acupuncture points in the abstract. Out of 1092 studies listed in PubMed until May 2006 under the keyword ‘acupuncture’, 259 were therefore chosen for this chapter. In some exceptional cases, studies based on animal experiments were also considered. Regarding the presentation of the results, particular emphasis was placed on the role of biomathematical methods as a parameter for reliability. The terms which allow assessment of a study in terms of its biomathematical quality are explained below:

Case study: Usually a retrospective (=in hindsight) summary of results of a particular treatment method in patients with a particular disorder; no control group or control therapy. Most basic form of clinical trial.

Controlled study: Usually a prospective (=planned in advance) scientific comparison of a treatment method with a group of untreated patients or of patients treated with a placebo or conventional therapy.

Placebo: A form of ‘mock’ or sham therapy. This can take the form of a pill containing no active ingredient or a therapy without any authentic foundation. The administration of a placebo tries to eliminate the non-specific improvement of the researched disorder (such as the therapist’s care of the patient and expectations of the patient) compared to the study group receiving authentic treatment. While the possibilities for placebo-acupuncture are plentiful, none is free of disadvantages. The effectiveness of acupuncture as a holistic therapy can be seen as the sum of many psychological as well as non-specific and specific physiological effects. Depending on the choice of the acupuncture placebo, certain effects can be excluded. Based on the concept of the sinew channels, however, there is an ongoing debate whether non-acupuncture points exist at all. The highest level of placebo acupuncture is the so-called ‘sham acupuncture’ where actual deep needling takes place at non-acupuncture points. The value of a study increases considerably if it is controlled by a placebo therapy.

Randomisation: The random assignment of a patient to a treatment group, without any influence of either the patient or the therapist. Randomisation significantly increases the value of a study.

Waitlist: A control group formed by patients receiving no treatment as a form of control. While a patient is waiting for his/her treatment, all his/her data relevant to the treatment are collected and later compared to those data obtained during treatment. Since there is neither randomisation nor a placebo treatment, the value of a waitlist group as a control group is very low.

Cross-over study: A particular design of controlled study. Each patient will participate in two different courses of treatment (for example treatment 1: medication; treatment 2: acupuncture). The order in which the treatments are administered can differ from patient to patient and be random (=randomised cross-over study). However, any long-term effect of the first form of treatment poses a problem for the final result. Therefore so-called washout phases during which patients receive no treatments at all are implemented to eliminate these effects and prevent a distortion of the data to be collected during the second phase. Since the effects of some acupuncture treatments have been shown to last for up to a year, cross-over studies are not very suitable for acupuncture research and particularly non-significant results should be questioned.

The 'n of 1' study is another form of research: a patient will receive various different treatments in order to find out which form of treatment led to the best improvement for the relevant disorder. However, the same limitations as discussed above will affect this type of study.

Significance: In simple terms, the significance of a study indicates if the difference of a particular parameter which has been observed between two groups is caused by the treatment (and is therefore a true difference) or whether the difference is the result of random deviation. For example, if in a large number of treated patients the observed data differ 'significantly' from a group of untreated patients, it is **highly probable** that this difference is not due to a random deviation. This probability can be expressed mathematically. If, for example, the significance is expressed as $p = 0.03$ this means that, considering a random error of 3%, the difference noted between the two groups is statistically significant. The convention is to accept a random error of up to 5% as acceptable. The significance of a study is based on the number of patients treated and the extent of the difference in the treatment result (for example by assuming the same random error, the difference in treatment results has to be much larger in a group of 10 patients ($n = 10$) than if the same study was carried out with 100 patients ($n = 100$)). Only differences classified as 'significant' are accepted for research studies.

Power of the study: This refers to the ability of a study to be statistically significant. The higher the number of participants, the higher the power of the study. By international standards, 80% is an acceptable level of power. Depending on the difference between the two study groups, the number of patients required can therefore easily be in the 100s or 1000s.

Experimental studies: Through standardising and controlling conditions of the experiment strictly, every attempt is made to avoid random results, often leading to just a small number of participants. Generally these tend to be healthy volunteers, who are subjected to a clearly defined stimulus. In this case acupuncture as a form of therapy will achieve a high level of validity. More questionable are studies of healthy subjects without a well-defined stimulus, since in that case acupuncture as a balancing form of treatment is not used according to its principles (for example changes in peristalsis caused by acupuncture in healthy subjects).

Blinding: The gold standard of clinical trials are the so-called double blind studies in which both the patient and the therapist remain ignorant as to which treatment (for example the tablet with or without the active ingredient) is administered to the patient. However, blinding of an acupuncturist is not possible. Even in a single blind study (only the patient is 'blinded'), his/her behaviour will distort the result of the trial. This effect can be counterbalanced to some extent by blinding the assessor of the treatment results (for example changes in the range of motion) so that he/she does not know which treatment (authentic or sham acupuncture) was received by the patient.

Standardisation: In pharmaceutical trials the amount of the active ingredient per tablet can easily be stated (for example standardisation of a 10 mg tablet). However, acupuncture cannot be standardised. The acupuncturist's point selection will largely be influenced by experience, prior knowledge, specialisation and cultural background. Strictly speaking, acupuncture trials therefore do not investigate acupuncture per se but acupuncture as performed by an acupuncturist or a group of acupuncturists.

The fact that acupuncture is to a large extent centred on the therapist has in the past led to a paradoxical situation: experienced Chinese acupuncturists performing the 'best' acupuncture produced biomathematically 'bad' trials due to their lack of training, while 'bad' Western acupuncturists produced 'good' trials from a biomathematical point of view. 'Importing' Chinese acupuncturists into Western trials did not solve the problem either, since the Chinese clinicians were now treating Europeans with diseases not common in China and which reacted differently to acupuncture. Only as experience grows on both sides – which would have to be evaluated individually for each trial – will the value of acupuncture as a researched form of therapy increase accordingly.

9.2 Scientific Research According to Channels

Lung Channel

LU-5 → Ren-23; L.I.-10

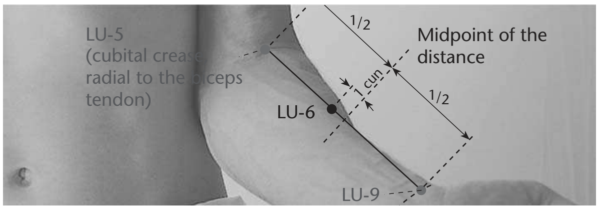

LU-6 192 patients with bronchial asthma were treated with **LU-6** and **LU-10**. 98.9% of the patients experienced an immediate improvement; the rate of clinical remission plus marked improvement was 76.5%. Best therapeutic results were achieved for allergic asthma according to Western medicine (asthma due to Cold according to TCM) and with longer needle retention (approximately 40 minutes). The lack of blinding and control groups reduce the value of this precise therapeutic case study. **Zang, J Tradit Chin Med 1990**

LU-7 → ST-8

LU-10 → LU-6

A Chinese prospective, randomised, controlled, non-blinded trial investigated the effect of bilateral bloodletting at **LU-10**, **P-8**, **HE-8**, **L.I.-2**, **T.B.-2** and **S.I.-3** on heart rate and consciousness in patients with **apoplexy** that had occurred within 3 days. Only patients with mild injuries showed improved consciousness while the heart rated increased independently of the size of injury. **Yi et al**, **J Tradit Chin Med 2005**

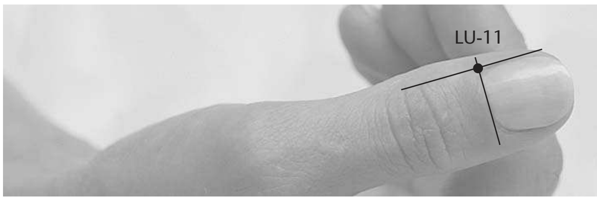

LU-11 Three Russian animal trials demonstrated the antipyretic action of **LU-11**. LU-11 can also strengthen the effect of hypothermic substances, suggesting a combination with antipyretic Chinese herbs. **Nezhenzev et al**, **Biull Eksp Bio Med 1992, and Fizion Zh SSSR Im I M Sechenova 1991**

A Taiwanese prospective, randomised, non-blinded, controlled trial ($n = 76$) investigated whether bloodletting at **LU-11** and **L.I.-1** can prevent or treat the occurrence of **laryngospasm** after tracheal extubation in children. The incidence of laryngospasm was significantly less in the bloodletting group (5% vs 24%). If laryngospasm developed, this could be relieved with the treatment of LU-11 or L.I.-1 within one minute in all patients. **Lee et al, Anaesthesia 1998**

Large Intestine Channel

L.I.-1 → LU-11

L.I.-2 → LU-10

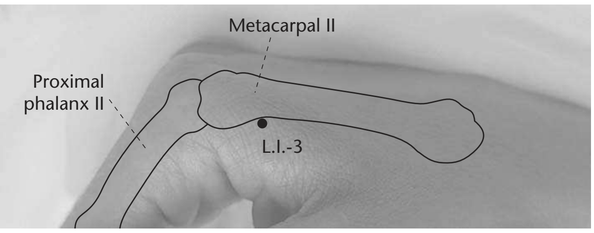

L.I.-3 A Chinese prospective, randomised, controlled study ($n = 210$) reports the significantly better effects of needling **L.I.-3** and ***lingxia*** in the treatment of **periarthritis of the humerus** compared to the control group treated with local points. Of interest is the application of *lingxia*, an unofficial extra point 2 cun inferior to G.B.-34. **Feng, J Tradit Chin Med 2003**

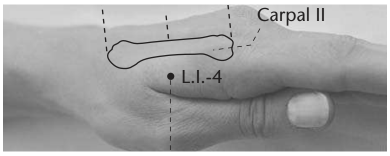

L.I.-4 → ST-36; P-5, P-6; ST-8; G.B.-43; Du-20; LIV-3; S.I.-18; SP-9; L.I.-11; ST-2

A Turkish prospective, controlled, non-blinded, group-comparison trial ($n = 55$) investigated the effect of electro-acupuncture (EA) at **L.I.-4**, **L.I.-11**, **ST-25**, **ST-36**, **ST-44** and **LIV-3** (combined with auricular points *shenmen* and *sanjiao*) on **obesity in women**. Treatments were given daily for 20 days. One control group received a 1425 kcal diet for 20 days while the other control group received no treatment. There was a 4.8% weight reduction in the EA group, whereas women on the restricted diet had a 2.5% weight loss. Both groups showed a significant reduction in total cholesterol and triglyceride levels. In the EA group only, LDL levels were also reduced significantly. **Cabioglu et Ergene**, **Am J Chin Med 2005**

A Chinese (Hong Kong) prospective, randomised, controlled, non-blinded study ($n = 29$) investigated the effect of **L.I.-4** and **P-6** on the heart rate variability (HRV) in normal subjects in **fatigue** and **non-fatigue states** respectively. Only fatigued subjects showed a significant change in HRV. The authors concluded that the effect of acupuncture depends not only on the points selected but also on the functional state of the subject, one of the principles of Chinese Medicine. **Li et al, Eur J Appl Physiol 2005**

An unusual veterinary case study investigated the effects of acupuncture on the **immune system**. According to the authors, immunostimulant points **include LI-4**, **LI-11**, **ST-36**, **GB-39**, **SP-6**, **GV-14**, **BL-11**, **BL-20**, **BL-23**, **BL-24**, **BL-25**, **BL-26**, **BL-27**, **BL-28,** and **Ren-12**. **BL-47** is considered to be immunosuppressive. Anti-febrile points include **ST-36** and **Du-14**. **Rogers et al, Probl Vet Med 1992**

A Chinese prospective, randomised, single-blinded, experimental trial ($n = 20$) compared the effect of transcutaneous electric nerve stimulation (TENS) at **L.I.-4** on the **heat pain threshold** and **vibration threshold** in humans with a placebo control point. In the study group, after TENS at **L.I.-4**, the pain threshold for heat was higher for up to half an hour compared to the control group. There was no effect on the threshold for feeling vibration. **Wang et al, Chin Med J (Engl) 2003**

A Korean prospective, randomised, placebo-controlled trial ($n = 35$) evaluated the efficacy of electro-acupuncture (EA) and moxibustion on **L.I.-4**, **L.I.-10**, **L.I.-11** and **T.B.-5** on **spasticity due to stroke**. The control group was treated with routine acupuncture. Only the EA group experienced a significant, immediate and lasting reduction in spasticity (using a modified Ashworth scale). **Moon et al, Am J Chin Med 2003**

A German prospective, randomised, controlled, non-blinded trial ($n = 29$) investigated if acupuncture at **L.I.-4**, **T.B.-5**, **ST-36**, **P-6**, **Du-20** and **Ex-UE-9** (*baxie*) had a vasospasmolytic effect on patients with **Raynaud's syndrome** during the winter season. There was no significant difference regarding the number and intensity of episodes, nor was there a significant effect on skin microcirculation between the acupuncture group and the placebo group before and after treatment. **Hahn et al, J Dtsch Dermatol Ges 2004**

A Japanese prospective, randomised, controlled study ($n = 21$) investigated the effect of **L.I.-4**, **ST-36 and SP-6** on objective and subjective parameters regarding **exhaustion** in elite female soccer players during a competition period. Only the subjects treated with acupuncture experienced a better rating for both subjective and objective parameters. **Akimoto et al, Med Sci Sports Exerc 2003**

A Chinese prospective, randomised, controlled trial ($n = 89$) researched the effect of **L.I.-4**, **ST-36**, **BL-18** and **BL-23** combined with CT positioning scalp circum-needling (SCN) and herbal therapy in patients with **poly-infarctional vascular dementia**. The control group received herbal treatment only. The treated group had a significantly better short-term result (96.5%) than the control group (75%) regarding changes of clinical symptoms, intelligence and haemorrheological characteristics. **Lun et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2003**

A Chinese prospective, randomised, controlled experimental trial ($n = 11$) investigated the changes in **brain activation patterns** evoked by stimulating **L.I.-4** manually versus stimulation with electro-acupuncture (EA). These were assessed by functional magnetic resonance imaging (fMRI). Results showed that EA mainly produced fMRI signal increases in precentral gyrus, postcentral gyrus/inferior parietal lobule and putamen/insula; in contrast, manual needle manipulation produced prominent decreases of fMRI signals in posterior cingulate, superior temporal gyrus, putamen/insula. These results suggest that different brain mechanisms may be recruited during manual and EA. **Kong et al, J Altern Complement Med 2002**

An American prospective, randomised, controlled, non-blinded experimental trial ($n = 13$) evaluated the effect of transcutaneous electric acupoint stimulation (TEAS) on **L.I.-4** and **P-6** on **tonic pain** induced with the cold-pressor test. While TEAS alone had a significant analgesic effect, this could be increased by combining TEAS analgesia with low-dose morphine. **Yuan et al, J Clin Pharmacol 2002**

A Taiwanese prospective, randomised, placebo-controlled experimental study ($n = 13$) investigated the effect of manual acupuncture as well as transcutaneous electrical nerve stimulation (TENS) with 2 Hz and 100 Hz respectively at **L.I.-4** on the spinal cord by measuring the **H-reflex**. While the application of TENS with both 2 Hz and 100 Hz changed the amplitude of the H-reflex, it was retained longer with the 100 Hz stimulation. These findings suggest that TENS enhances the excitability of the motoneuron pool in the spinal cord. **Chang et al, Acupunct Electrother Res 2001**

An Austrian prospective, randomised, controlled study ($n = 45$) evaluated whether acupuncture at **L.I.-4** and **SP-6** at term can influence cervical ripening, **induce labour** and thus reduce the need for postdate induction. Compared to the control group, the study group showed a significantly shorter time span (three days) between estimated date of confinement (EDC) and actual time of delivery. Medical induction of labour 10 days after EDC was also significantly lower in the acupuncture group (20%) than in the control group (35%). **Rabl et al, Wien Klin Wochenschr 2001**

An Israeli prospective, randomised, double-blinded controlled trial ($n = 25$) investigated the effect of acupuncture in the treatment of **irritable bowel syndrome**. True acupuncture was performed at **LI-4** and sham acupuncture at **BL-60** during a total of two sessions. There was no significant difference between the two groups. However, the choice of acupuncture points seems somewhat peculiar. While both points are Ma Dan Yang points (the 12 most important points), they have no specific action for intestinal disorders. Obviously the authors chose the points according to the channel (Large Intestine). Furthermore, only two treatments does not seem sufficient. **Fireman et al, Digestion 2000**

A British prospective experimental trial ($n = 14$) investigated the effect of acupuncture at **L.I.-4** on the **resting EEG** of healthy subjects. Acupuncture did not induce any changes in the resting EEG. **Rosted et al, Complement Ther Med 2001**

An Irish prospective, randomised, controlled trial ($n = 30$) applied transcutaneous electric nerve stimulation (TENS) either at **L.I.-4** or over the median nerve, in order to investigate **cutaneous blood flow**. The control group received no treatment. Only TENS over the median nerve showed a significant increase in cutaneous blood flow. **Cramp et al, Acupunct Electrother Res 2001**

A Japanese prospective, randomised, controlled, non-blinded trial ($n = 22$) investigated the effect of acupuncture at **L.I.-4**, **ST-6** and **ST-7** on pain after mandibular **wisdom tooth extraction**. Compared to the control group there was a significant reduction in pain in the treatment group. **Kitade et al, Acupunct Electrother Res 2000**

A Swedish prospective, randomised, controlled experimental trial ($n = 12$) investigated if acupuncture at **L.I.-4** induced **changes in the sympathetic and/or parasympathetic nervous system** in healthy subjects. For this purpose, regular and superficial needling of **L.I.-4** was compared. Changes were measured by using power spectral analysis and heart rate frequency analysis. Only deeper needling at L.I.-4 led to a significant activation of the sympathetic and parasympathetic nervous system during and after treatment as well as to a significant reduction of the heart rate frequency. **Haker et al, J Auton Nerv Syst 2000**

A US prospective, randomised, sham-controlled, single-blinded trial (*n* = 101) assessed the effect of varying intensities of transcutaneous electric acupoint stimulation (TEAS) at **L.I.-4** on the **postoperative patient-controlled analgesia** (PCA) requirement for **hydromorphine** (HM), the incidence of opioid-related side effects, and the recovery profile after lower abdominal surgery in women. In the sham-TEAS group HM-requirement was reduced by 23%, low-intensity TEAS produced a 34% decrease in the HM requirement while the high-intensity TEAS reduced requirement significantly by 65%, accompanied by a reduction in the occurrence of dizziness, nausea and pruritus. **Wang et al, Anesth Analg 1997**

A Chinese prospective, randomised, controlled trial (*n* = 110) investigated the effect of transcutaneous electric acupoint stimulation TEAS) at **L.I.-4**, **Ex-HN-4** ***(yuyao)*** and **G.B.-31** during enflurane anaesthesia in patients undergoing craniotomy. In comparison to the control group (anaesthesia maintained with enflurane only) the **minimum alveolar concentration** (MAC) of enflurane decreased in the TEAS-supplemented group by 38–47%. If anaesthesia was supplemented by TEAS plus scalp infiltration with procaine, there was an even higher reduction of 42–66%. The haemodynamic balance was also more stable during the operation, and postoperative recovery was faster. **Wang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1994**

A German prospective, waitlist-controlled trial (*n* = 6) investigated the effect of electro-acupuncture at **L.I.-4**, **ST-25**, **LIV-3** and **BL-25** on stool frequency and colonic transit time in **chronic constipation**. There were no significant differences in all the parameters between the control group and the waitlist group. **Klauser et al, Z Gastroenterol 1993**

A Japanese experimental trial researched the influence of **L.I.-4 on vibration induced finger flexion reflex**. Unilateral acupuncture at L.I.-4 suppressed the reflex in both hands. **Takakura et al, Am J Chin Med 1992**

A Chinese prospective, randomised, matched controlled trial (*n* = 12) investigated the analgesic effect of aqueous acupuncture (injection of glucose solution) at **L.I.-4** and **G.B.-34** in **postoperative pain control**. In comparison to the control group, the intensity of postoperative pain as well as the requirement for analgesic medication were significantly lower in the study group. **Chen et al, Gaoxiong Yi Xue Ke Xue Za Zhi 1991**

A German prospective, non-randomised, sham-controlled trial (*n* = 36) investigated the **analgesic effect** of acupuncture at **L.I.-4**, **P-6**, **ST-36** and **SP-4** prior to a **colonoscopy**. The pain experienced by the acupuncture patients was significantly lower compared to the group receiving sham acupuncture. The analgesics and sedatives required were also significantly less. **Li et al, Dtsch Med Wochenschr 1991**

A prospective, randomised, single-blinded experimental study (*n* = 39) investigated the effect of acupuncture and laser acupuncture at **L.I.-4** and **Ex-UE** ***jianqian*** on the **pain threshold**. Compared to laser acupuncture, acupuncture significantly increased the pain threshold. **Brockhaus et al, Pain 1990**

A British prospective, controlled experimental cross-over study investigated the effectiveness of electro-acupuncture at **L.I.-4** on experimental **tooth pain**. After 30 minutes of electric stimulation the pain threshold increased by 27%. This increase could be partially blocked by naloxone. **Ernst et al, Acupunct Electrother Res 1987**

A British prospective, randomised, placebo-controlled study (*n* = 51) investigated the effect of acupuncture at **L.I.-4** as an **analgesic for operative dentistry**. There were no significant differences between the treatment group and the placebo group. **Taub et al, Oral Surg Oral Med Oral Pathol 1979**

A prospective non-randomised, placebo-controlled experimental trial (*n* = 40) investigated the effect of acupuncture at **L.I.-4** on the **pain threshold of postoperative dental pain** compared to codeine. Both **L.I.-4** and codeine had a significantly better effect than the placebo. The combination of both had the strongest effect. **Sung et al, Anesth Analg 1977**

An Austrian prospective, sham-controlled, single-blinded experimental trial (*n* = 12) measured the effect of **L.I.-4** and **P-6** on **pain threshold and pain tolerance**. In contrast to sham acupuncture, true acupuncture significantly raised the pain threshold. **Stacher et al, Am J Chin Med 1975**

L.I.-5 → P-5

L.I.-6 → P-5

L.I.-7 → P-5

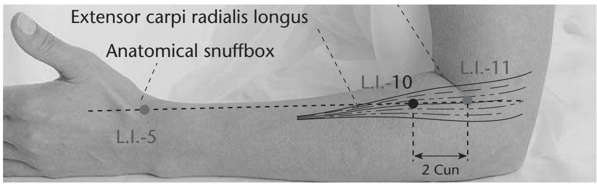

L.I.-10 → L.I.-4; Du-12; ST-36

A Swedish prospective, randomised, controlled, double-blinded trial (*n* = 49) investigated the effect of laser acupuncture on **L.I.-10**, **L.I.-11**, **L.I.-12**, **LU-5** and **T.B.-5** on **lateral humeral epicondalgia (tennis elbow)**. No significant difference occurred at a dose of 0.36 J/point. **Haker et al, Pain 1990**

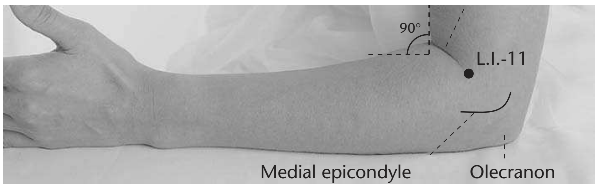

L.I.-11 → L.I.-4; ST-36; LIV-3; L.I.-10; L.I.-15

A Taiwanese prospective, randomised, placebo-controlled, single-blinded trial (*n* = 40) investigated the effect of acupuncture at **L.I.-11** on **uraemic pruritus**. The results of the pruritus score questionnaires demonstrated that only true acupuncture at **L.I.-11** resulted in a significantly reduced pruritus lasting for a period of three months. Uraemia levels remained unchanged. **Che-yi et al, Nephro Dial Transplant 2005**

A German prospective, randomised, placebo-controlled, single-blinded experimental trial ($n = 22$) investigated the effect of **L.I.-11** on the **phagocytic immune system**. Based on laboratory parameters, the **activity of neutrophils** was significantly higher in the study group. **Karst et al, Complement Ther Med 2003**

A Swedish prospective, non-randomised, placebo-controlled, single-blinded experimental longitudinal trial investigated the influence of electro-acupuncture at **L.I.-11** and **L.I.-4** on sympathetic nerve activity and the pain threshold. Only electro-acupuncture at L.I.-11 and L.I.-4 increased the pain threshold in combination with increased sympathetic nerve activity. **Knardahl et al, Pain 1998**

A Swedish prospective, randomised, placebo-controlled, single-blinded trial ($n = 58$) investigated the effect of laser acupuncture at **L.I.-11** and **L.I.-12** for **lateral humeral epicondalgia (tennis elbow)** compared to laser application on the painful area. There were no significant differences between the laser and placebo treatments after the treatment period. **Haker et al, Arch Phys Med Rehabil 1991**

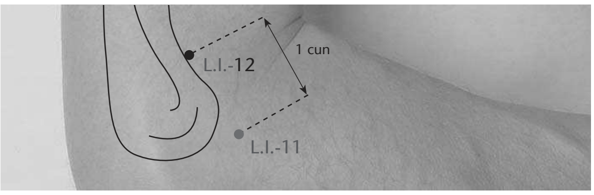

L.I.-12 → L.I.-11; L.I.-10

A Chinese prospective, randomised, controlled trial ($n = 64$) investigated the effect of electro-acupuncture at **L.I.-12**, **T.B.-5**, **G.B.-30** and **ST-36** accompanied by medication for **acute cerebral infarction** with regard to the plasma and CSF somatostatin levels. The control group received medication only. Both plasma and CSF somatostatin levels increased significantly in the acupuncture group. **Zhang et al, J Tradit Chin Med 1999**

A Chinese prospective, randomised, controlled trial ($n = 64$) investigated the effect of electro-acupuncture at **L.I.-12**, **T.B.-5**, **G.B.-30** and **ST-36** on the plasma and CSF levels of VIP (vasoactive intestinal peptide), somatostatin and pancreatic polypeptide in patients with **acute cerebral infarction**. The CSF VIP level dropped significantly while the plasma pancreatic polypeptide level increased significantly. In patients with a good response both plasma and CSF somatostatin levels increased significantly after acupuncture. **Zhang et al, Zhen Ci Yan Jiu 1996**

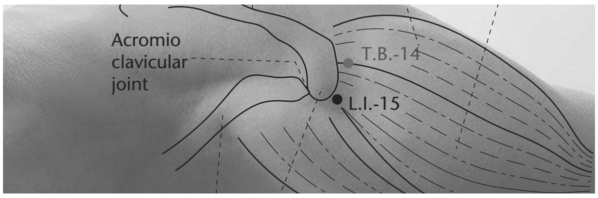

L.I.-15 A Taiwanese prospective, randomised, controlled, non-blinded trial ($n = 150$) investigated the effect of electro-acupuncture at **L.I.-15**, **G.B.-21** and **Ex-UE *jianneiling*** with or without regional nerve block (RNB) (stellate ganglion and suprascapular plexus) for **'frozen shoulder' (adhesive capsulitis)**. This was compared to RNB treatment alone. The combination of electro-acupuncture and RNB produced the most significant reduction of pain and duration of pain relief as well as improvement of range of motion. **Lin et al, Acta Anaesthesiol Sin 1994**

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 63$) investigated the effect of acupuncture at **L.I.-15**, **L.I.-11**, **L.I.-4**, **G.B.-30**, **G.B.-34**, **G.B.-37**, **LIV-3**, **G.B.-20**, **ST-25**, **ST-40**, **ST-36**, **SP-6** and **KID-3** on symptoms of **cerebral infarction** compared to the treatment with the calcium channel blocker verapamil. The effectiveness in the acupuncture group was 94%, in the verapamil group 84%. This difference is statistically significant. However, the choice of a calcium channel blocker seems unusual for this indication. **Zou et al, Zhong Xi Yi Jie He Za Zhi 1990**

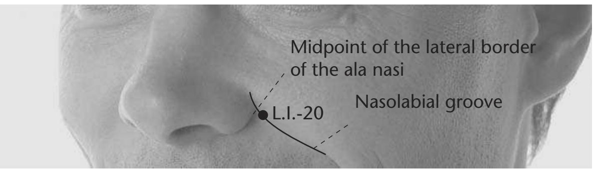

L.I.-20 → G.B.-14

A British prospective, randomised, controlled trial ($n = 20$) investigated the effect of a 30 second massage at **L.I.-20** on **nasal airflow** compared to an untreated control group. Symptoms improved significantly in the massage group. **Takeuchi et al, Am J Rhinol 1999**

Stomach channel

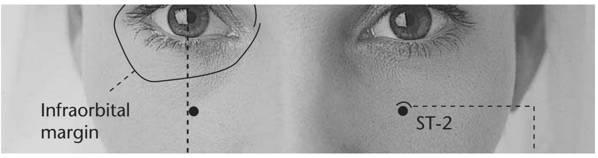

ST-2 → G.B.-14

A prospective, non-randomised experimental trial ($n = 33$) investigated the effect of acupuncture and electro-acupuncture at **ST-2**, **L.I.-4** and **ST-44** on experimentally induced **tooth pain**.

While both procedures increased the pain threshold, the increase was not sufficient to justify acupuncture as a means of pain control in conservative dentistry. **Bakke, Scand J Dent Res 1976**
ST-4 → G.B.-14
ST-6 → G.B.-14; L.I.-4
ST-7 → L.I.-4

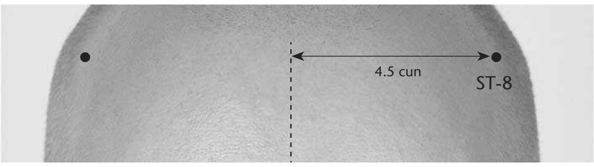

ST-8 A large prospective, randomised, multi-centre, group-comparison trial ($n = 120$) in patients suffering from **migraine without aura** demonstrated that acupuncture at **ST-8**, **G.B.-5**, **G.B.-20**, **Du-14** and **LU-7** was significantly more effective than conventional drug therapy, regarding both frequency and duration of the migraine attacks as well as number of sick days. The authors calculated that by treating migraine with acupuncture the Italian health system could save approximately 0.5 billion Euros per year. **Liguori et al., J Tradit Chin Med 2000**
A Bulgarian case study ($n = 35$) reports the successful treatment of headaches following caesarian sections with spinal anaesthesia with acupuncture at ST-8, L.I.-4, G.B.-11, G.B.-14, BL-10, Du-14 and Du-20 during 1–3 treatments. **Tsenov, Akush Ginekol (Sofia) 1996**
A Swedish prospective, placebo-controlled trial investigated the effect of intrasegmental (**ST-8**) and extrasegmental (**L.I.-4**) electro-acupuncture on sensory thresholds. Only the pain threshold could be increased, and this only with intrasegmental acupuncture at ST-8. **Lundeberg et al, Am J Chin Med 1989**
ST-18 → G.B.-14; Du-1
ST-20 → T.B.-8
ST-21 → Du-20

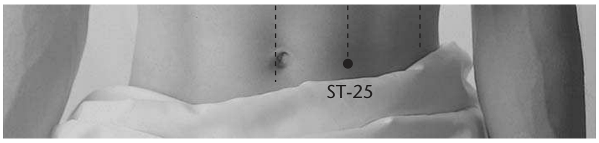

ST-25 → Du-1; Du-20; L.I.-4; L.I.-15
A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of acupuncture and moxibustion at ST-25 and Ren-4 for chronic colitis. This was compared to modern drug therapy. The acupuncture group showed the same therapeutic results but with fewer side effects. **Yang et al, J Tradit Chin Med 1999**

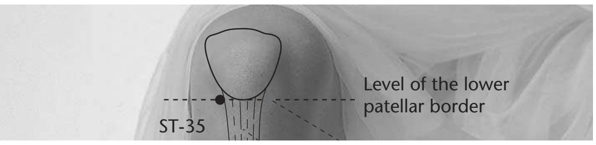

ST-32 → G.B.-31
ST-34 → ST-36; SP-9
ST-35 A Chinese prospective, randomised, single-blinded, controlled, long-term trial ($n = 24$) investigated the effect of electro-acupuncture (EA) and transcutaneous electrical nerve stimulation (TENS) at **ST-35** and **Ex-LE-4** (eight treatments over a period of two weeks) in older patients (average age: 85 years) suffering from **painful osteoarthritic knees**. This was compared to conventional therapy for this disorder. Both TENS and EA significantly reduced the knee pain. EA also had an effect on range of motion. **Ng et al, Altern Complement Med 2003**

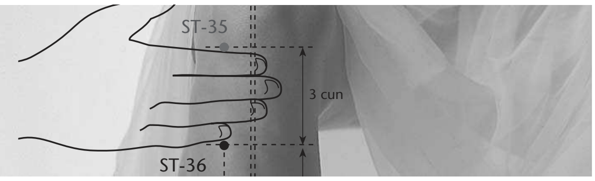

ST-36 → L.I.-4; BL-27; P-6; T.B.-8; G.B.-26; G.B.-31; Du-1, Ex-B-3; Du-4; Du-20; LIV-3; SP-9; BL-60; L.I.-12; L.I.-15; Ren-17
An Iranian prospective, randomised, placebo-controlled, single-blinded trial ($n = 51$) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** and **ST-38** on nitroglycerin-induced **hypotension under general anaesthesia with halothane**. With true acupuncture at the above points, the mean arterial blood pressure (MAP) was significantly lower (60 vs 66 mmHg). The effect of acupuncture was significantly greater with lower dosages of nitroglycerin compared with higher dosage. In addition, the time to reach steady state MAP was significantly shorter (10 vs 15.8 minutes) and the quality of operative ischaemia was excellent in 83.3% subjects compared to 25% in the sham group. **Saghaei M. et al, Acta Anaesthesiol Taiwan 2005**
A US prospective non-blinded experimental trial investigated the response to acupuncture at **ST-36** as evidenced by **fMRI**. The limbic and paralimbic structures of the cortical and subcortical regions in the telencephalon, diencephalon, brainstem and cerebellum were activated in response to acupuncture. **Hui et al, Neuroimage 2005**
A prospective, randomised placebo-controlled single-blinded trial ($n = 13$) investigated the effect of electro-acupuncture at **ST-36** on the **heart rate variability** (HRV) in healthy subjects. There was no acupuncture-specific effect on HRV. **Chang et al, Am J Chin Med 2005**
A German prospective randomised placebo-controlled single-blinded experimental crossover study ($n = 10$) investigated the effect of acupuncture at **ST-36**, **L.I.-11**, **SP-10 and Du-14** on

leukocyte circulation in healthy young subjects. There was a significant decrease in leukocyte and lymphocyte values in the acupuncture group while cortisol and norepinephrine plasma levels remained unchanged. **Kou et al, Brain Behav Immun 2005**

A US prospective, non-randomised, placebo-controlled, single-blinded trial ($n = 7$) investigated if electro-acupuncture at **ST-36, SP-6**, **ST-34** and **L.I.-4** before and during induction of general anaesthesia (desflurane) influenced the **anaesthetic requirement** compared to a control group. There was no reduction in pain in the treatment group. **Chernyak et al Anesth Analg 2005**

A Taiwanese prospective, randomised, placebo-controlled, single-blinded experimental cross-over study ($n = 15$) investigated if electro-acupuncture (EA) at **ST-36** can normalise **atropine-induced gastric dysrhythmia**. While there was a significant increase in the percentage of normal frequency, atropine-induced gastric dysrhythmia was not normalised by EA (no significant differences in the EGG). **Chang et al, Dig Dis Sci 2002**

A Taiwanese prospective, randomised, controlled experimental trial ($n = 15$) investigated the effect of **ST-36** on **gastric myoelectrical regularity**. There was a significant decrease in the tachygastric and bradygastric rhythm during electro-acupuncture on ST-36. **Chang et al, Digestion 2002**

A Taiwanese prospective, randomised, sham-controlled trial ($n = 100$) investigated the effect of electro-acupuncture at **ST-36** on **postoperative pain and opioid-related side effects**. The postoperative pain, the requirement for morphine (via PCA) and the incidence of nausea and dizziness were all significantly reduced. **Lin et al, Pain 2002**

A US prospective, randomised, double-blinded, sham-controlled experimental trial ($n = 14$) investigated if electro-acupuncture at **ST-36**, **G.B.-34** and **BL-60** could reduce the anaesthetic requirement (desflurane) for **experimental pain stimuli**. Electro-stimulation of these points did not reduce desflurane requirements compared to the control group. **Morioka et al, Anesth Analg 2002**

A Taiwanese prospective case study ($n = 15$) investigated the effect of bilateral electro-acupuncture **at ST-36** on gastric slow waves in **diabetic patients** with symptoms suggesting **gastric motor dysfunction**. Changes of gastric activity were measured by EGG. During and after acupuncture there was a significant increase in the percentages of normal frequency. In addition the percentage of tachygastric frequency was decreased significantly during and after acupuncture. **Chang et al, Digestion 2001**

A Chinese case study ($n = 104$) reports the successful injection-acupuncture at **ST-36** for **chemotherapy-induced leukopenia**. **Yin et al, J Tradit Chin Med 2001**

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 26$) investigated the effect of acupuncture at **ST-36** plus auricular plaster therapy on **peristalsis** for postoperative recovery of intestinal function after abdominal surgery. 92% of patients in the treatment group showed recovery of normal peristalsis within 72 hours, while in the control group only 46% patients recovered within the same time-period. **Wan et al, J Tradit Chin Med, 2000**

A US prospective, randomised, controlled pilot study ($n = 17$) investigated the effect of acupressure at **ST-36** and **P-6** on the intensity and frequency of **nausea** in patients undergoing **chemotherapy due to breast cancer**. Both the frequency and intensity of the nausea was significantly lower in the treatment group than in the control group. **Dibble et al, Oncol Nurs Forum 2000**

An Italian prospective, randomised, controlled experimental trial ($n = 120$) investigated the effect of acupuncture at **ST-36** and **L.I.-4** regarding changes in the levels of beta-endorphins and other parameters (VIP, lymphocyte subsets, NK cells and monocyte phagocytosis) in patients suffering from various painful disorders. These changes were compared to an untreated control group. Only in the acupuncture group were the endorphin levels, the CD3 and CD4 values and monocyte phagocytosis increased for at least 24 hours after treatment. At the same time there was an increase of the CD8 values. **Petti et al, J Tradit Chin Med 1998**

A Chinese retrospective case study reports good results in treating **leukopenia** with acupuncture at **ST-36**. **Wei, J Tradit Chin Med 1998**

A Taiwanese prospective, randomised, sham-controlled experimental trial ($n = 18$) investigated the effect of acupuncture at **ST-36** and **L.I.-4** on central nervous system pathways with functional MR imaging of the brain. Besides a significant reduction of the heart rate frequency, acupuncture at both points resulted in activation of the hypothalamus and nucleus accumbens (structures of the descending antinociceptive pathway) and deactivation of the rostral part of the anterior cingulate cortex, amygdala formation, and hippocampal complex (limbic areas of pain recognition). Therefore this study provides explanations for the **pain-reducing effect** of acupuncture. **Wu et al, Radiology 1999**

A prospective, sham-controlled, single-blinded experimental longitudinal trial ($n = 13$) investigated the effect of acupuncture at ST-36 only, compared to acupuncture at **ST-36 plus L.I.-10** on **auditory endogenous potentials** (P300). The effect of the different forms of acupuncture was identical. In contrast to the sham group, it led to a decrease of P300 amplitude in the acupuncture group. **Hsieh et al, Am J Chin Med 1998**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 100$) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** on **opioid analgesic requirement** by PCA following hysterectomy or myomectomy compared to peri-incisional dermatomal stimulation by TENS. TENS applied at the dermatomal level of the incision is as effective as stimulation at ST-36, and both were more effective than stimulation at a sham location. **Chen et al, Anesth Analg 1998**

A Chinese prospective, randomised, controlled trial ($n = 69$) compared the so-called immunotherapy at **ST-36** to conventional desensitisation therapy in patients suffering from **anaphylactic asthma**. Results in the acupuncture group were significantly better, regarding both clinical and immunohistochemical parameters. **Chen et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1996**

A Mongolian prospective, randomised, controlled study ($n = 48$) compared the effect of acupuncture at **ST-36** and **P-6** combined with epidural anaesthesia to simple epidural anaesthesia for subtotal gastrectomies. The acupuncture group, while requiring a lower dosage of anaesthetic, experienced a stronger analgesic effect, the abdominal muscles were more relaxed and the haemodynamic disturbances were smaller. **Sun, Zhen Ci Yan Jiu 1996**

A US prospective, randomised, controlled, experimental trial ($n = 11$) researched the effect of transcutaneous electric nerve stimulation (TENS) at **ST-36** and **ST-37** on **peripheral blood circulation** and **haemodynamics**. Twenty minutes of TENS on healthy subjects did not produce a significant change in the measured parameters. **Balogun et al, Disabil Rehabil 1996**

A Taiwanese prospective, randomised, placebo-controlled, single-blinded trial investigated the influence of acupuncture at **ST-36** on the **pulse spectrum**. The results indicated that acupuncture at ST-36 has a specific effect on the Fourier components of the pulse: C2 and C4 decreased while C5, C6, C8 and C9 increased. This specific frequency effect was not found when acupuncture was applied on a non-acupuncture point. **Wang et al, Am J Chin Med 1995**

A Chinese prospective, randomised, controlled trial ($n = 45$) investigated the effect of **ST-36**, **L.I.-11** and **KID-6** on the immunoactivity of natural killer cells and the interleukin-2 level in **malignant tumour** patients. After receiving one treatment of 30 minutes daily for 10 days, parameters in the treatment group had increased significantly compared to the control group. **Wu et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1994**

A Chinese prospective, randomised, controlled trial ($n = 80$) compared **anaesthesia with electro-acupuncture** at **ST-36** and **L.I.-4** to epidural anaesthesia in **appendectomies**. The operations were equally successful in both groups, without any significant differences regarding the success rate. **Sun et al, Zhen Ci Yan Jiu 1992**

A Chinese prospective, randomised, controlled trial ($n = 39$) investigated the effect of acupuncture at **ST-36** and **SP-6** on **bowel movements following abdominal surgery**. In the acupuncture group the first postoperative bowel movement occurred after approximately 58 hours, but only after 86 hours in the untreated control group. **Liu et al, Zhong Xi Yi Jie He Za Zhi 1991**

ST-37 → ST-36
ST-38 → G.B.-34; ST-36
ST-40 → L.I.-15
ST-44 → L.I.-4; ST-2

Spleen channel

SP-1 A US prospective experimental trial ($n = 13$) investigated the effect of electro-acupuncture (EA) at **SP-1** and **LIV-1** on **thermal pain thresholds**. The pain threshold increased significantly 30 seconds after onset of EA. The authors postulate that EA at the above points has an inhibitory effect on the C-fibre afferents; the analgesic benefit observed is most probably A-delta afferent mediated. **Leung et al, J Altern Complement Med 2005**

SP-4 → Du-4; P-6; L.I.-4

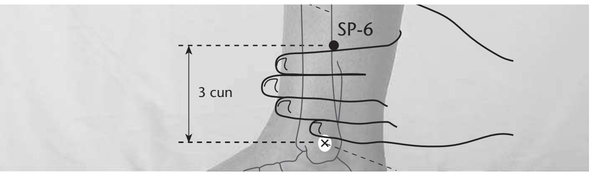

SP-6 → ST-36; L.I.-4; P-6; G.B.-26; Du-4; Du-20; LIV-3; L.I.-15

In a prospective, randomised, double-blinded group-comparison trial ($n = 56$) patients suffering from **interstitial cystitis** performed daily laser therapy on SP-6 at home for 30 seconds over a period of 12 weeks. There were no significant differences between the treatment and control cohorts. **O'Reilly et al, J Urol 2004**

An Australian prospective, randomised, placebo-controlled trial ($n = 20$) investigated the effect of transcutaneous electrical stimulation (TENS) at **SP-6** and **LIV-3** on **uterine contractions** in post-date pregnant women. A significant increase in frequency and strength of contractions was found in the TENS group compared with the placebo group. **Dunn et al, Obstet Gynecol 1989**

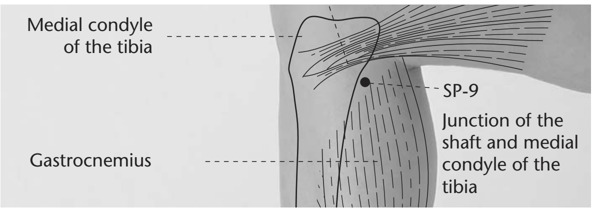
Medial condyle of the tibia
SP-9
Junction of the shaft and medial condyle of the tibia
Gastrocnemius

SP-9 A British prospective, randomised, controlled, assessor-blinded trial ($n = 44$) investigated the influence of **SP-9**, **SP-10**, **ST-34**, **ST-36** and **L.I.-4** on unilateral versus bilateral acupuncture in patients with **advanced osteoarthritis of the knee**. The symptoms improved significantly in both groups. This improvement was sustained for six months. There was no statistically significant difference between the groups. **Tillu et al, Acupunct Med 2001**

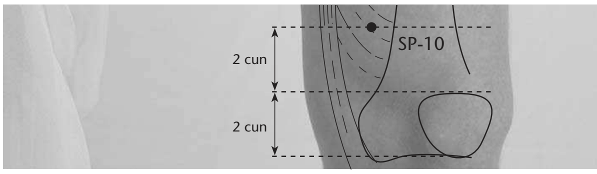

SP-10 → SP-9; ST-36; Ren-17

A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of Q-wave millimetre microwave application at **SP-10** and **BL-17** on **chemotherapy-induced leukopenia in gastrointestinal cancer patients**. One group received irradiation before the start of chemotherapy, the other group only after the onset of the chemotherapy-induced leukopenia. Beginning the irradiation before starting chemotherapy produced a significantly better result (86% versus 73% with later application). **Wu et al, Zhongguo Zhong Xi Yi Jie He za Zhi 1997**

SP-12 → Du-4

Heart channel

HE-3 → P-6
HE-5 → P-6

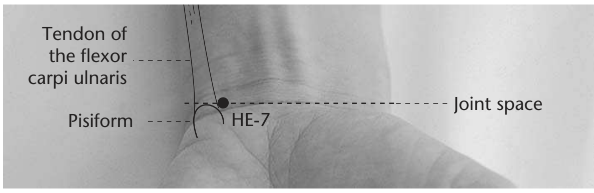

HE-7 → G.B.-26; Du-4; Du-20; S.I.-3; P-6

A German prospective, randomised, placebo-controlled, single-blinded trial ($n = 36$) investigated the effect of **H-7**, **P-6**, **Du-20**, **BL-62** and **Ex-HN-6** ***erjian*** on the cardiac autonomous nervous system in patients with **minor depression and anxiety disorders**. Only the acupuncture group demonstrated a relative increase of cardiovagal modulation of heart rate. **Agelink et al, Fortschr Neurol Psychiatr 2003**

A British pilot study ($n = 17$) investigated the effect of **HE-7** on certain **stress parameters**. There was an average decrease of 44% (based on the Edinburgh Postnatal Depression Scale (EPDS). **Chan et al, Acupunct Med 2002**

HE-8 → LU-10
HE-9 → Du-4

Small Intestine channel

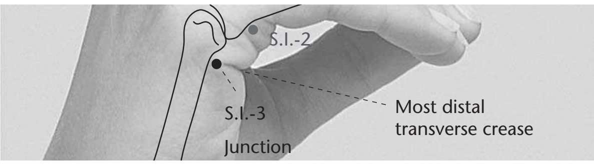

S.I.-3 → LU-10; Du-12

A larger-scale Taiwanese prospective, controlled, randomised trial ($n = 100$) demonstrated a significant improvement of neurological functioning in patients with **severe spinal cord injuries** (ASIA status A and B) after electro-acupuncture at **S.I.-3** and **BL-62** plus auricular acupuncture versus the control group (conventional and rehabilitation therapy only). **Wong et al, Am J Phys Med Rehabil 2003**

A non-controlled pilot study ($n = 18$) found that the treatment of **acute torticollis** with a single treatment of ipsilateral acupuncture at **S.I.-3** and **M-UE-8** resulted in a mean improvement of lateral head rotation of 53%. **Samuels N, Am J Chin Med 2003**

A Chinese prospective, randomised, controlled trial ($n = 62$) investigated the effect of acupuncture at **S.I.-3** and **HE-7** in patients with **cerebral traumatic dementia** compared to conventional physiotherapy. There were significant changes of MMSE scores and auditory P300 values in the acupuncture group only. **Zhang et al, Zhen Ci Yan Jiu 1996** and **Zhang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1995**

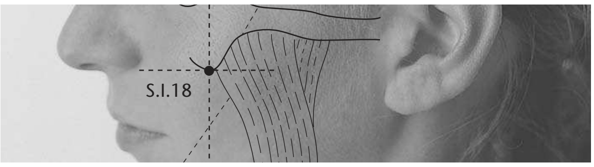

S.I.-18 A Chinese prospective, randomised, controlled, single-blinded trial ($n = 42$) investigated the effect of electro-acupuncture at **S.I.-18**, **Ex-HN-04**, **G.B.-20** and **L.I.-4** on controlled hypotension induced by isoflurane during a cerebral operation in order to avoid blood loss. The concentration of isoflurane necessary to achieve controlled hypotension was significantly reduced by 31–42% compared to the control group. **Wang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

Bladder channel

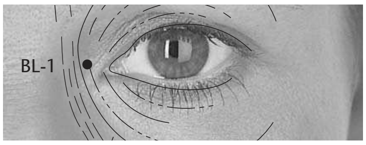

BL-1 → G.B.-14

A non-randomised group-comparison trial ($n = 34$) investigated the effect of warmed needle acupuncture at **BL-1** in patients suffering from **epiphora** due to dysfunction of the lacrimal duct. The effect in the treatment group was significantly better in comparison to the control group treated with lacrimal duct irrigation and norfloxacin eye drops (92% vs 54%). **Ni et al, J Tradit Chin Med 2002**

A Chinese case study reports the treatment of **epiphora** due to **insufficiency of lacrimal passage** with acupuncture at **BL-1**. Of the 68 eyes treated in 42 patients, 28 eyes were cured and 35 improved. 34 eyes improved after only one course of treatment. **Ni Y et al, J Tradit Chin Med 1999**

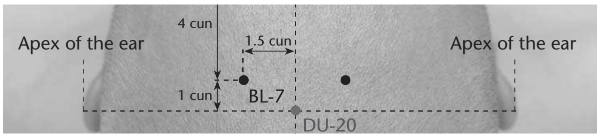

BL-7 A Chinese experimental trial carried out in 1988 was the first to discuss the possible effect of **BL-7**, **G.B.-6** and **Ex-HN-1** on **hemiplegia** through changes of the micro-circulation in the nail bed. **Shun et al, Zhen Ci Yan Jiu 1988**

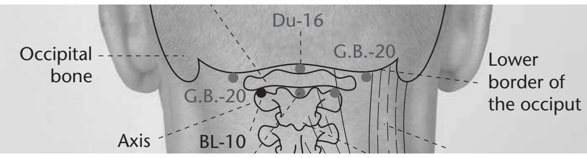

BL-10 → ST-8

A placebo-controlled, non-blinded group-comparison trial ($n = 65$) investigated the effect of treating children with an acuplaster at **BL-10**, **BL-11** and **G.B.-34** the night before undergoing **strabismus surgery,** for the prevention of **post-operative nausea and emesis**. Nausea and emesis were reduced by 50% during the early and later postoperative stage compared to the placebo-group. The well-researched point P-6 often failed to prevent nausea in this type of surgery. **Chu et al, Acta Anaesthesiol sin 1998**

BL-11 → Du-12; BL-10; L.I.-4; Ex-B-1

BL-12 → Ex-B-1

BL-13 → Ex-B-1; Du-14

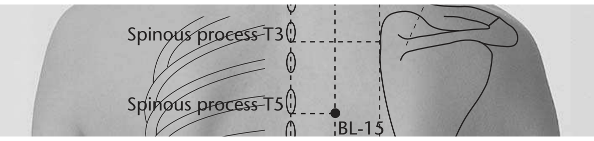

BL-15 → Ex-B-1; P-6

In an animal experiment, moxibustion at **BL-27** was able to reduce the renal excretion of Na^+ while at the same time decreasing urinary volume. Systolic blood pressure remained unchanged while plasma levels of aldosterone and ANP (atrial natriuretic peptide) decreased. In contrast, moxibustion at BL-15 increased urinary volume while reducing excretion of Na^+ and decreasing systolic blood pressure. Plasma levels of aldosterone and ANP decreased. These results suggest that BL-15 and BL-27 might be applied for certain forms of **hypertension**. **Lee et al, Am J Chin Med 1997**

BL-17 → SP-10

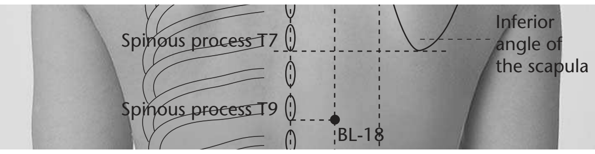

BL-18 → BL-23

A Japanese prospective, randomised, placebo-controlled, double-blinded trial ($n = 189$) investigated the effect of Japanese intradermal acupuncture at **BL-18**, **BL-19**, **BL-20**, **BL-21**, **BL-22**, **BL-23**, **BL-24**, **BL-25** and **BL-26** on **analgesic requirement, nausea and emesis** as well as stress markers after **abdominal surgery**. Starting from the recovery room and including the second postoperative day incisional pain at rest and during coughing was significantly lower in the verum group than in the control group. The requirement for epidural morphine was also significantly lower (only 50%). Nausea and emesis were also significantly lower (20–30%). Plasma cortisol and epinephrine concentrations were reduced by 30–50% in the acupuncture group during recovery and on the first postoperative day. This is one of the most valuable trials due to its excellent methodology and design. **Kotani et al, Anesthesiology 2001**

BL-19 → BL-18

BL-20 → L.I.-4; Du-1; BL-18; Du-14; P-6

BL-21 → BL-18

BL-22 → BL-18; Ex-B-2

BL-23 → L.I.-4; Du-4; Du-20; BL-18; Du-14

BL-24 → L.I.-4; BL-18

BL-25 → L.I.-4; BL-18; Ex-B-2

BL-26 → L.I.-4; BL-18

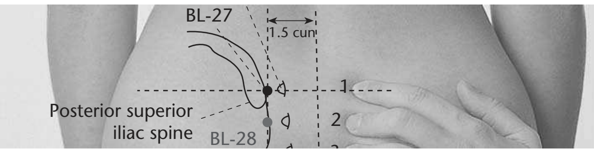

BL-27 → L.I.-4; BL-15

A German randomised, placebo-controlled experimental trial ($n = 42$) compared the effect of acupuncture at **BL-27** on pain reduction with a point of the Yamamoto New Scalp Acupuncture (YNSA). Experimental pain stimuli were set on the **upper calcaneus edge**. The difference in pain reduction between the acupuncture group and the scalp acupuncture group was highly significant. There are also highly significant differences concerning the verum and the placebo treatment. BL-27 could therefore be a useful point for treating calcaneal pain. **Ogal et al, Anasthesiol Intensivmed Notfallmed Schmerzther 2002**

An animal experiment in 15 healthy dogs investigated the effect of acupuncture on intestinal (duodenum) motility. Electro-acupuncture at BL-27 decreased the frequency of intestinal motility by 31% during stimulation, while showing an increase of 18% after stimulation. Electro-stimulation of ST-36 led to the opposite result: a 20% increase during stimulation followed by a decrease of 7% after stimulation. Based on this result, BL-27 could be applied for **paralytic ileus** or **constipation** while ST-36 could be used for treating **bloating, tenesmus and diarrhoea**. **Choi et al, J Vet Sci 2001**

BL-28 → L.I.-4; Du-4

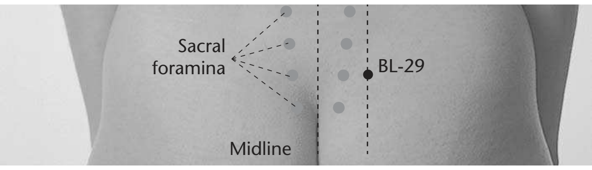

BL-29 There exists only a small case study describing **various disorders**. This is of educational value rather than anything else. **Chen Y., J Tradit Chin Med 2002**

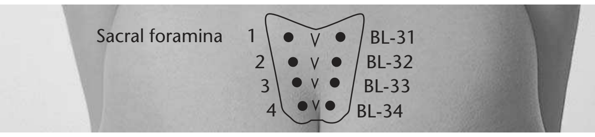
Sacral foramina
1
2
3
4
BL-31
BL-32
BL-33
BL-34

BL-32 → Du-4; Ren-3
A medium-scale Chinese case study ($n = 30$) reports the successful treatment of **urinary and faecal incontinence** due to nerve damage with electro-acupuncture at **BL-35** and **BL-32**. **Yang et al, J Tradit Chin Med 2003**

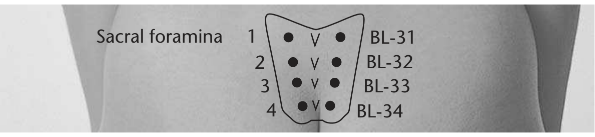
Sacral foramina
1
2
3
4
BL-31
BL-32
BL-33
BL-34

BL-33 A Japanese case study ($n = 10$) documents a significant reduction of pain for **chronic pelvic pain syndrome** with **intrapelvic venous congestion** by weekly treatments with bilateral acupuncture at **BL-33** (needle retention 10 minutes, manual rotation). **Honjo et al, Int J Urol 2004**

A Japanese prospective case study ($n = 13$) investigated the effect of bilateral acupuncture at **BL-33** on cystometry and **incontinence** in **spinal-cord-injured patients**. Incontinence completely disappeared in 15% of cases and decreased to 50% or less compared to baseline in a further 46% of patients. Maximum cystometric bladder capacity increased significantly from 75 ml to 148 ml. When cystometry was repeated in six patients one month after the last acupuncture treatment, bladder capacity showed a further average increase to 187 ml. **Honjo et al, Urol Int 2000**

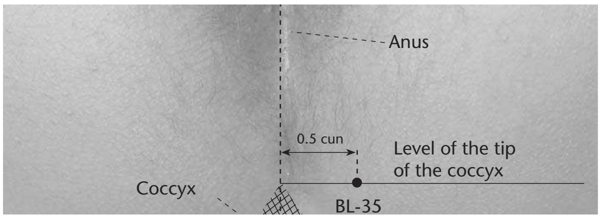
Anus
0.5 cun
Level of the tip
of the coccyx
Coccyx
BL-35

BL-35 → BL-29; BL-32
A larger-scale Chinese case study ($n = 103$) reports the successful treatment of **post-partum urinary retention** with classic acupuncture at **BL-35**. **Li et al, J Tradit Chin Med 1996**
BL-47 → L.I.-4

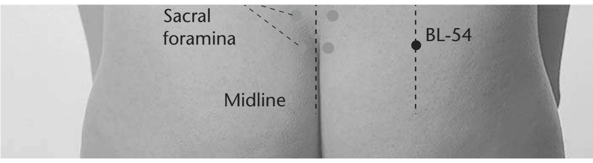
Sacral
foramina
BL-54
Midline

BL-54 → Du-3
A larger-scale Chinese case study ($n = 100$) reports the successful treatment of **lumbar pain** with acupuncture at **BL-54**. **Cui, J Tradit Chin Med 1992**

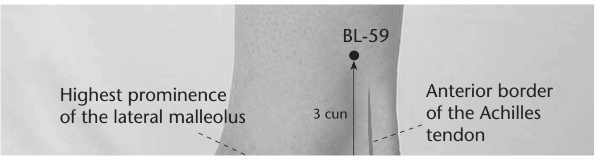
BL-59
Highest prominence
of the lateral malleolus
3 cun
Anterior border
of the Achilles
tendon

BL-59 A smaller-scale Chinese case study reports the treatment of **acute lumbar pain** with acupuncture at **BL-59**. **Hu, J Tradit Chin Med 1993**

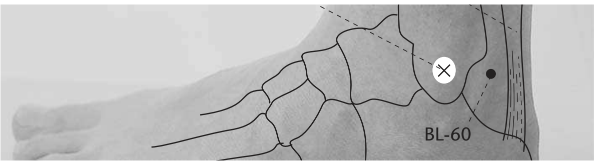

BL-60 → Du-3; ST-36; L.I.-4
A US prospective case study ($n = 7$) reports the positive subjective (questionnaire) and objective (tibial H-reflex) effect of non-invasive electro-acupuncture at **BL-60**, **ST-36**, **KID-1** and **LIV-3** on **HIV-related neuropathy**. Results showed a significant overall improvement in functional activities. **Galantino et al, J Altern Complement Med 1999**
BL-62 → S.I.-3; HE-7; Du-20

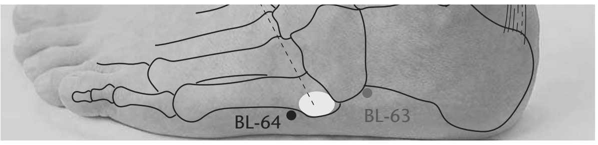

BL-64 An animal experiment demonstrated that electro-acupuncture at BL-64 and BL-65 increased the activity of nitric oxide synthase in the brain stem nuclei (nucleus gracilis). **Ma at al, Acupunct Electrother Res 2002**
BL-65 → BL-64

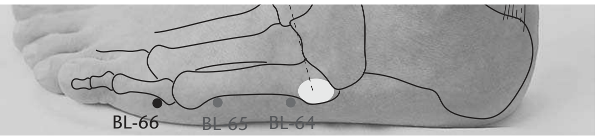

BL-66 An experimental study in rats investigated the antipyretic effect of manual acupuncture at **BL-66** on lipopolysaccharide-induced **fever**. Acupuncture at this point reduced the fever and also reduced the production of the pro-inflammatory cytokines interleukin-1 beta and interleukin-6 in the hypothalamus of the treated rats. **Son et al, Neurosci Lett 2002**

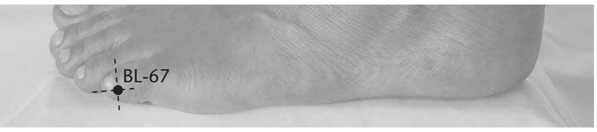

BL-67 An Italian prospective, randomised, controlled, single-blinded trial ($n = 123$) investigated the effect of moxibustion at **BL-67** on the conversion rate in **breech presentations**. The trial had to be terminated prematurely due to a high drop-out rate of participants. At that time, there were no significant differences between the moxibustion group and the control group. The authors conclude that there is too little acceptance for moxibustion treatment in non-Asian cultures. **Cardini et al, Br J Obstet Gynaecol 2005**

A randomised, controlled, prospective larger-scale trial ($n = 67$) demonstrated a higher rate of conversion of fetal **breech presentations** in a group treated with 30 minutes of manual acupuncture at **BL-67** (76.4%) compared to an untreated control group (45.4%). **Habek et al, Fetal Diagn Ther 2003**

A large-scale ($n = 260$), randomised, controlled, prospective trial demonstrated a higher rate of conversion of fetal **breech presentations** in women treated with moxibustion at **BL-67** for 7–14 days (75.4%) than the conventionally treated control group (47.7%). **Cardini et al, JAMA 1998**

A medium-scale case study ($n = 48$) with retrospective control groups demonstrated that electro-acupuncture at **BL-67** for the conversion of fetal **breech presentations** had a 81.3% success rate. **Li et al, J Tradit Chin Med 1996**

Acupuncture at **BL-67** resulted in a significant increase in the number of **c-Fos-positive cells in the primary visual cortex** of binocularly deprived rat pups to levels suggesting that BL-67 has an influence on the activity of the primary visual cortex. **Lee et al, Am J Chin Med 2002**

A recent Italian prospective, randomised trial investigated the effect of acupuncture and moxibustion at **BL-67** on the conversion of **breech presentations** in 240 pregnant women. At delivery, breech presentation was significantly lower in the active-treatment group (37%) than in the observation group (54%). Therefore the proportion of caesarean sections indicated for breech presentation was also significantly lower in the treatment group than in the observation group (52% vs 67%). **Neri et al, J Matern Neonatal Med 2004**

A Croatian prospective, randomised, controlled, non-blinded group-comparison trial ($n = 67$) assessed the value of acupuncture at **BL-67** for 30 minutes in the conversion of **breech presentation**. The success rate of the acupuncture correction of the breech presentation was 76% while the rate in the untreated control group was only 45%. This is a highly significant difference. **Habek et al, Fetal Diagn Ther 2003**

An Austrian prospective, randomised, placebo-controlled experimental trial ($n = 10$) investigated the effect of laser acupuncture at **BL-67** on the visual cortex using fMRI to monitor **neuroradiological changes**. Only verum acupuncture resulted in an activation in the cuneus (Brodmann Area BA 18) and the medial occipital gyrus (BA 19). These results confirm the traditional application of this point for ophthalmic disorders. **Siedentopf et al, Neurosci Lett 2002**

An Italian prospective, non-randomised, sham-controlled, single-blinded trial ($n = 12$ pregnant women with **breech presentation**) demonstrated that acupuncture at **BL-67** significantly reduced the fetal heart rate while increasing fetal movement and producing more accelerations. **Neri et al, J Soc Gynecol Invest 2002**

Kidney channel

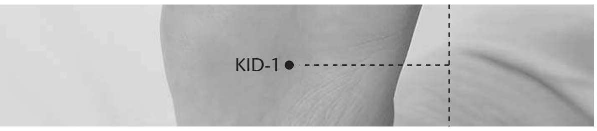

KID-1 → BL-60

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 150$) investigated the effect of acupuncture at **KID-1** in the treatment of **children with bronchial asthma** compared to a control group treated with conventional drugs. The clinical effectiveness of acupuncture at KID-1 was significantly higher than treatment with drugs (89% vs 64%). In addition, in the acupuncture group the level of eosinophils and IgE was highly significant and significantly lower respectively than in the control group. **Gao et Zhu, Zhongguo Zhong Xi Yi Jie He Za Zhi 2005**

A Chinese retrospective case study reports the successful local stimulation of **KID-1** by applying a paste in the treatment of **essential hypertension**. **Gong et al, J Tradit Chin Med 1995**

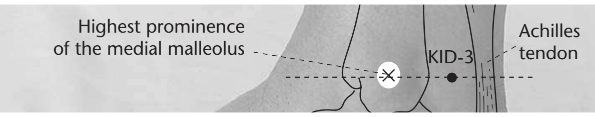

KID-3 → Du-4; L.I.-15

A Taiwanese prospective, randomised, controlled experimental trial ($n = 157$) compared the absorption of Tc-99 m (a radioactive technetium marker) at **KID-3** with non-acupuncture points. Absorption at KID-3 was significantly better than at non-acupuncture points. **Wu et al, Am J Chin Med 1994**

KID-4 → P-6
KID-5 → Du-4
KID-6 → ST-36

Pericardium channel

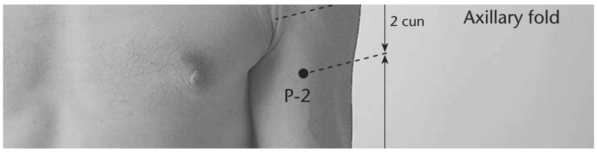

P-2 A Chinese experimental trial (*n* = 100) examined cardiac function before and during acupuncture at **P-2**, **P-3**, **P-4** and **P-6** as well as at control points (some arbitrary points and **G.B.-37**) in patients with **coronary heart disease**. In contrast to the control points the verum points had an effect – even if only a minor one – on the measured parameters (PEP, LVET, P/L, HI, SV, CO, ST segment and T wave of ECG). **You et al, Zhen Ci Yan Jiu 1993**
P-3 → P-2
P-4 → P-2

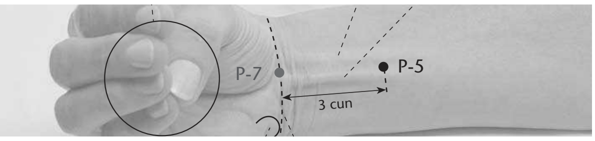

P-5 An earlier Chinese trial interpreted the synchronising effect of electro-acupuncture at **P-5** on EEG activity as having a **homeostatic effect on the whole body**. **Huang et al, Zhen Ci Yan Jiu 1990**
Electro-acupuncture at **P-5** improved the ST segment of ECG in rabbits with induced **acute myocardial infarction**. **Cao et al, Zhen Ci Yan Jiu 1990**
An earlier Chinese experimental trial (*n* = 10) found that electro-acupuncture at **P-5** and **P-6** shortened sino-atrial conduction in healthy subjects and increased frequency at the sinus node (**positive chronotropic and bathmotropic effect**). **Xi et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1993**
Previous work of the team led by Li and Tjen-A-Looi suggests that the inhibitory effect of electro-acupuncture (EA) on the pressor reflex induced by bradykinin (BK) applied to the gallbladder is in part due to the activation of opioid receptors most likely located in the rostral ventrolateral medulla (rVLM). This trial investigated the specific opioid receptor subtypes and neurotransmitters responsible for this inhibition. Therefore BK was applied to the gallbladder of anaesthetised cats to induce an increase of arterial blood pressure. It was found that applying EA at P-5 and P-6 activated the mu and delta opioid receptors located in the rVLM, thus preventing a rise of the arterial blood pressure for a longer period of time through the activating influence of the splanchnic nerve (for example due to stretching of the gallbladder or stomach). This experiment substantiated the positive effect of EA at P-5 and P-6, especially for myocardial ischaemia due to **coronary heart disease**. **Li et al, Auton Neurosci 2001; Li et al, Am J Physiol Regul Integr Comp Physiol 2002; Tjen-A-Looi et al, Auton Neurosci 2003**
In a US study 17 healthy subjects were subjected to weekly bicycle training for 3–4 weeks. Subjects were asked to perform the training with and without electro-acupuncture (EA) at **P-5**, **P-6**, **L.I.-4**, **L.I.-5**, **L.I.-6**, **L.I.-7**, **G.B.-37**, **G.B.-38** and **G.B.-39**. In 70% of subjects, EA at points on the Pericardium and Large Intestine channels led to an increase in maximal workload with decreased blood pressure. Points on the Gall Bladder channel had no effect. Further trials may investigate the application of these points in the treatment of hypertension induced by exercise stress. **Li et al, Clin Auton Res 2004**

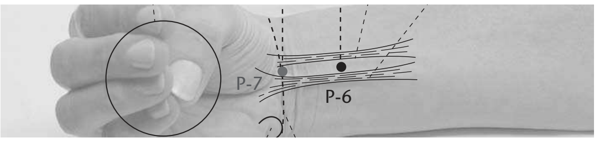

P-6 → P-2; P-5; Du-20; Du-14; HE-7; LIV-3; L.I.-4; ST-36
A Taiwanese prospective, randomised, placebo-controlled, double-blinded trial (*n* = 110) investigated the effect of acupressure bands at **P-6** on **nausea and vomiting during spinal anaesthesia for caesarean delivery**. The reduction and occurrence of nausea (64% vs 71%) and vomiting (22% vs 27%) was not significant. **Ho et al, Anesth Analg 2006**
A US prospective, randomised, placebo-controlled, single-blinded trial (*n* = 94) investigated the effect of acupressure bands at **P-6** on the occurrence of **nausea and vomiting during caesarean delivery under spinal anaesthesia**. There was no statistically significant difference between the active and sham control groups in the incidence of intraoperative nausea (30% vs 43%), postoperative nausea (23% vs 41%), intraoperative vomiting (13% vs 9%), postoperative vomiting (26% vs 34%), intraoperative antiemetic requirement (23% vs 18%), postoperative antiemetic requirement (34% vs 39%), complete intraoperative response (55% vs 57%) or complete postoperative response (51% vs 34%). These results did not reach the selected level of significance. There were also no significant differences between the groups in nausea scores, number of vomiting episodes and patient satisfaction. **Habib et al, Anesth. Analg 2006**
A Swedish prospective, randomised placebo-controlled experimental double-blinded trial (*n* = 60) investigated the effect of acupressure at **P-6** on **nausea induced by eccentric rotation**. Mean time to nausea was significantly the longest in the P-6 group (352 seconds), compared to the placebo-acupressure group (280 seconds) and the untreated control group (151 seconds). **Alkaissi et al, Can J Anaesth 2005.**
A Turkish prospective randomised controlled non-blinded trial (*n* = 90) investigated the effect of transcutaneous electrical acupoint stimulation (TEAS) at **P-6** and **Ren-13** compared to ondansetron on **postoperative nausea following paediatric**

tonsillectomy. There was the same incidence of nausea episodes in the TEAS group and in the ondansetron group, but significantly fewer episodes than in the untreated control group. Side-effects were significantly higher in the ondansetron group. **Kabalak et al, J Altern Complement Med 2005**

An Indian prospective, randomised, placebo-controlled, double-blinded trial (*n* = 120) investigated the effect of capsicum-plasters on **P-6** on **postoperative nausea and vomiting** (PONV) compared to conventional treatment with ondansetron. For the duration of the plaster application (6 hours after the operation) the incidence of PONV and the requirement for antiemetics were significantly lower in both groups. **Misra et al, Can J Anaesth 2005**

A Turkish prospective, randomised, placebo-controlled, single-blinded trial (*n* = 127) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **P-6** on the incidence and severity of **nausea and gagging during gastroscopies**. There were no significant differences between the treatment group and placebo group. **Tarcin et al, Turk J Gastroenterol 2004**

A Chinese prospective, non-blinded, randomised, group-comparison trial (*n* = 30) investigated the effect of acupuncture at **P-6** plus mexiletin on **intermittent ventricular extrasystole**. The treatment group had a total effective rate of 90% compared with the control group, who was administered mexiletin only (80% total effective rate). Only the acupuncture group showed no increase in extrasystole. **Zhang et al, J Tradit Chin Med 2004.**

A Chinese group-comparison trial investigated the effect of acupuncture at **P-6** in the treatment of patients with angina pectoris and **acute myocardial infarct**. The effectiveness of acupuncture was significantly better (91%) than conventional treatment with isosorbide dinitrate and nifedipine. **Meng et al, J Tradit Chin Med 2004**

A US prospective, controlled, group-comparison trial (*n* = 77) investigated the effect of stimulation at **P-6** on the symptoms of **motion sickness** during exposure to optokinetic drum rotation. Despite the extreme test conditions, stimulating P-6 delayed the onset of symptoms considerably. **Miller et al, Aviat Space Environ Med 2004**

A Croatian prospective, single-blinded group-comparison trial (*n* = 36) investigated the efficiency of acupuncture at **P-6** in the treatment of **hyperemesis gravidarum**. Treatment results were measured by the requirement for anti-emetic medication. Acupuncture at P-6 had an efficiency rate of 90%, acupressure at P-6 64%, placebo acupuncture 12.5% and placebo acupressure 0%. **Habek et al, Forsch Komplementarmed Klass Naturheilkd 2004**

In a German prospective, randomised, placebo-controlled double-blinded group-comparison trial acupuncture at **P-6** significantly reduced the incidence of **emesis after gynaecological and breast surgery** (40% placebo, 25% P-6). Postoperative emesis was only reduced for gynaecological surgery, not for breast surgery. **Streitberger et al, Anaesthesia 2004**

A Chinese prospective, randomised, single-blinded, placebo-controlled group-comparison trial (*n* = 41) investigated the effect of twice daily acupressure at **P-6**, **ST-36** and **SP-6** on **gastrointestinal motility following trans-abdominal hysterectomy**. In comparison to the control group receiving acupressure on sham points, gastrointestinal motility (measured with a multifunctional stethoscope) increased significantly in the acupressure group. **Chen et al, Am J Chin Med 2003**

A German prospective, randomised, placebo-controlled, single-blinded trial (*n* = 80) investigated the anti-emetic effect of acupuncture at **P-6** in addition to administration of ondansetron in patients with **chemotherapy-related nausea**. There was no significant difference between the treatment and control group. **Streitberger et al, Clin Cancer Res 2003**

A US prospective, randomised, placebo-controlled, double-blinded trial (*n* = 230) investigated the effect of TENS at **P-6** for the relief of **nausea and vomiting in pregnancy**. Based on the Rhodes Index, results in the verum group were significantly better than in the control group. **Rosen et al, Obstet Gynecol 2003**

A British prospective, partially randomised, partially blinded, placebo-controlled trial (*n* = 301 patients with **acute myocardial infarction**, 125 of whom acted as a non-randomised control group) compared the effect of wristband-acupressure at **P-6** with placebo-acupressure. P-6 led to a significant reduction in **nausea and vomiting** during the last 20 hours of the 24-hour study phase (18%), compared with the placebo group (32%) and the control group (43%). **Dent et al, Complement Ther Med 2003**

A Swedish prospective case study (*n* = 39) investigated the effect of **P-6** combined with ondansetron for **nausea and vomiting** associated with **cyclophosphamide chemotherapy**. The authors stated that, compared with ondansetron treatment alone, the combined acupuncture-ondansetron treatment was significantly more effective, but they do not mention a control group. **Josefson et al, Rheumatology (Oxford) 2003**

In a US prospective, randomised, single-blinded trial (*n* = 53) the application of 'minute sphere' acupressure at **P-6**, **ST-36**, **SP-6** and **SP-4** showed no decrease of **postoperative pain and morphine requirement after abdominal surgery**. **Sakurai et al**, **Anesth Analg 2003**

A Swedish prospective, randomised, placebo-controlled, double-blinded, multi-centre trial (*n* = 410) investigated the effect of **P-6** for the treatment of **postoperative nausea and vomiting after gynaecological surgery**. The incidence of nausea and vomiting was significantly lower in the acupuncture group (33%) than in the control group (46%). **Alkaissi et al, Can J Anaesth 2002**

A US prospective, randomised, placebo- and sham-controlled, double-blinded trial (*n* = 120) investigated the effect of the prophylactic application of a 'ReliefBand' at **P-6** on the incidence of **nausea and vomiting after plastic surgery,** compared to the application of the ReliefBand in addition to 4 mg ondansetron. The occurrence of nausea and vomiting as well as the need for antiemetic 'rescue' medication was significantly the lowest in the P-6 plus ondansetron group. **White et al, Anesthesiology 2002**

A Korean prospective, randomised, placebo-controlled, double-blinded trial (*n* = 160) investigated among other parameters the effect of a capsicum plaster applied at **P-6** on **postoperative nausea and vomiting**. The incidence of nausea and vomiting was significantly less (P-6: 26% within 24 hours, placebo: 57%). **Kim et al, Anesth Analg 2002**

A Japanese case study investigated the effect of different stimulation techniques at **P-6** on the coronary arteries in patients with **coronary heart disease**. The mean coronary dilation with acupuncture was 69% of that caused by isosorbide dinitrate. **Kurono et al, Am J Chin Med 2002**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 187$) investigated the effect of **P-6** on the incidence of **postoperative nausea and vomiting** in children. P-6 produced significantly better results than the sham points and was equally effective as administration of droperidol, but without the sedative, hypotensive side effects. **Wang et al, Anesthesiology 2002**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 27$) investigated the effect of **P-6** on the incidence of **nausea and vomiting during chemotherapy**. Compared to the control group, the anti-emetic requirement was significantly lower in the P-6 group. **Roscoe et al, Altern Ther Health Med 2002**

An Austrian prospective, randomised, placebo-controlled, experimental cross-over study ($n = 51$) investigated the effect of acupuncture at **P-6** on the **skin blood perfusion** measured by laser Doppler perfusion imaging. This was compared to needling a sham point. Changes in the skin blood perfusion occurred significantly earlier in the acupuncture group than in the sham group, with a basically more pronounced reduction in skin blood perfusion at the verum point. **Litscher et al, Lasers Med Sci 2002**

A large-scale Australian prospective, randomised, sham-controlled trial ($n = 593$) investigated the effect of acupuncture at **P-6** on the incidence **of nausea and vomiting in early pregnancy**. Women receiving traditional acupuncture reported significantly less nausea after the second treatment and significantly less dry retching after the third week compared with women in the control group. Women treated with sham acupuncture showed significantly less nausea and retching after the third treatment compared with untreated women. Individualised acupuncture resulted in significantly less nausea and retching already after the first treatment. In this study none of the treatments had a significant influence on the frequency of vomiting. **Smith et al, Birth 2002**

A US prospective, randomised, sham-controlled trial ($n = 120$) investigated the influence of electro-acupuncture (EA) at **P-6** on postoperative **nausea and vomiting** in paediatric patients who had undergone **ENT surgery** and were awake. EA at P-6 significantly decreased the incidence of nausea (P-6: 60%, sham: 85%, control group: 93%). While vomiting had the lowest incidence in the P-6 group, the difference from the other groups was not significant. **Rusy et al, Anesthesiology 2002**

A Swedish prospective, randomised, placebo-controlled pilot study ($n = 60$) investigated the effect of tuina (acupressure) at **P-6** on **pregnancy-related nausea and vomiting**. The frequency of both nausea and vomiting was significantly lower in the acupressure group compared to the sham acupressure group and untreated control group. **Werntoft et al, J Reprod Med 2001**

A US prospective, randomised, controlled trial ($n = 25$) investigated the effect of an acupressure wrist band at **P-6** on **motion sickness**. Both the symptoms of motion sickness and abnormal gastric activity as recorded via EEG were significantly lower in the group treated with wristbands at P-6. **Stern et al, Altern Ther Health Med 2001**

A US prospective, randomised, placebo-controlled, single-blinded trial investigated the effect of ‘sea-bands’ (acupressure wristbands) at **P-6** on the incidence of **postoperative nausea and vomiting**. There were no significant differences. **Windle et al, J Perianesth Nurs 2001**

A Swedish prospective, randomised, sham-controlled, single-blinded cross-over study ($n = 33$) investigated the effect of acupuncture at **P-6** on pregnant women with **hyperemesis gravidarum** (vomiting in pregnancy). The verum group showed a significantly faster reduction of the nausea, and also a higher number of the patients experienced no vomiting. **Carlsson et al, J Pain Symptom Manage 2000**

A Chinese prospective, controlled, randomised, group-comparison study ($n = 181$) and a Chinese case study ($n = 33$) demonstrated that electro-acupuncture at **P-6** and **P-8** as well as at **L.I.-4**, **T.B.-5**, **ST-36** and **SP-6** with de-escalating treatment frequency from four times daily to once weekly for a period of 15 and 28 days had a significant effect on the **withdrawal symptoms of heroin addicts**. **Zhang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000; Wu et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

A Taiwanese prospective, non-randomised, placebo-controlled, single-blinded experimental cross-over study ($n = 44$) investigated the effect of acupuncture at **P-6** on the left ventricular ejection fraction (LVEF) in 22 healthy subjects and 22 patients with **coronary artery disease** (CAD). There was no change of LVEF in the healthy subjects while LVEF significantly increased in the CAD patients. **Ho et al, Am J Chin Med 1999**

A US prospective, randomised, placebo-controlled, single-blinded trial ($n = 100$) investigated the effect of pre-operative acupressure (verum group) and intra-operative acupuncture (verum and control group) at **P-6** on **postoperative nausea and vomiting** after **tonsillectomy** in children. There were no significant differences. **Shenkman et al, Anesthesiology 1999**

A Swedish prospective, randomised, placebo-controlled, double-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **nausea and vomiting** after **minor gynaecological surgery**. Only the group receiving acupressure at P-6 experienced no vomiting and also required no anti-emetic medication. **Alkaissi et al, Acta Anaesthesiol Scand 1999**

An Austrian prospective, randomised, placebo-controlled, double-blinded trial investigated the effect of laser acupuncture at **P-6** on **nausea and vomiting in children undergoing strabismus surgery**. The laser acupuncture was administered 15 minutes before induction of anaesthesia and 15 min after arriving in the recovery room. In the laser stimulation group, the incidence of vomiting was significantly lower (25%) than that in the placebo group (85%). **Schlager et al, Br J Anaesth 1998**

A Chinese prospective, randomised, placebo-controlled, double-blinded trial ($n = 163$) investigated the effect of acupoint stickers at **P-6**, **Ren-8** and **T.B.-18** on **motion sickness**. All tested acupuncture points showed a significantly better result than the active control points (scopolamine plaster) and the placebo control (lactose plaster). T.B.-18 was the most effective point (symptoms 100% reduced), followed by Ren-8 (81%) and P-6 (46%). **Pei et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1998**

A Chinese prospective, randomised, controlled trial ($n = 50$) compared two forms of anaesthesia for the **anterior approach cervical discectomy**: local anaesthesia plus intravenous anaesthesia (IVA) in the control group and electro-acupuncture at **P-6** and **L.I.-4** plus IVA. The effect of the anaesthesia was the same in both groups. The authors therefore recommend electro-acupuncture at P-6 and L.I.-4 as a less risky alternative. **Li et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1997**

A British prospective, randomised, placebo-controlled, double-blinded trial ($n = 81$) investigated the effect of intra-operative acupuncture at **P-6** in the prevention of **postoperative nausea and vomiting** in patients undergoing **gynaecological laparoscopic surgery**. The use of acupuncture significantly reduced the incidence of postoperative nausea and vomiting by 30–38% compared with placebo. **Al-Sadi et al, Anaesthesia 1997**

A US prospective, randomised, placebo- and sham-controlled, double-blinded trial ($n = 75$) investigated the effect of acupressure at **P-6** on **nausea and vomiting during caesarean section under spinal anaesthesia** compared with administration of 10 mg metoclopramide. Acupressure at P-6 is as effective as 10 mg metoclopramide. **Stein et al, Anesth Analg 1997**

An Australian prospective, randomised, placebo-controlled, double-blinded trial ($n = 84$) investigated the effect of TENS at **P-6** and **L.I.-4** on **postoperative nausea and vomiting** in **paediatric patients** undergoing **minor urological surgery**. The differences between the treatment and control group were not statistically significant. **Schwager et al, Anaesth Intensive Care 1996**

A Taiwanese prospective, randomised, sham-controlled, single-blinded experimental trial ($n = 48$) evaluated the relationship between electro-acupuncture at **P-6** and **ST-36** and cardiopulmonary function in healthy subjects. In the P-6/ST-36 group **resting heart rate**, **carbon dioxide production** and **oxygen requirement** were significantly decreased, indicating that acupuncture can lower the metabolic rate. **Lin et al, Chin Med J (Engl) 1996**

A Taiwanese prospective, randomised, placebo-controlled, double-blinded trial ($n = 60$) investigated the effect of acupressure wristbands at **P-6** on the incidence of **nausea and vomiting after epidural morphine for post-caesarean section pain relief**. Compared to the control group, the incidence of nausea significantly decreased from 43% to 3%, vomiting significantly decreased from 27% to 0%. **Ho et al, Acta Anaesthesiol Scand 1996**

A US prospective, randomised, placebo-controlled, single-blinded cross-over study ($n = 9$) investigated the effect of acupressure wristbands at **P-6** on nausea and vomiting for **seasickness**. Depending on the time of application (earlier or later during the trip) subjects were either symptom-free or complained about nausea and vomiting. The difference was highly significant. **Bertolucci et al, Aviat Space Environ Med 1995**

A US prospective, randomised, placebo- and sham-controlled, single-blinded experimental trial ($n = 9$) investigated the effect of acupressure at **P-6** on the incidence of visually-induced **motion sickness**. Acupressure at P-6 significantly reduced the incidence of nausea and abnormal gastric myoelectric activity. **Hu et al, Aviat Space Environ Med 1995**

A Chinese prospective, randomised, controlled trial ($n = 40$) investigated the effect of acupuncture on **P-6**, **L.I.-4**, **ST-36** and **KID-4** on the regulation of cellular immune function in patients with **malignant tumours**. There was a highly significant increase of the CD3+ and CD4+ levels, an increase in the CD4+/CD8+ ratio, a higher endorphin level and a decreased level of soluble interleukin-2 receptor (SIL-2R). **Wu, Zhen Ci Yan Jiu 1995**

A British prospective, randomised, controlled trial ($n = 46$) investigated the effect of acupressure at **P-6** on **nausea and vomiting** following **laparoscopy** for **gynaecological surgery** and the requirement for anti-emetic therapy. There was a significant reduction in the requests for anti-emetic therapy in the P-6-acupuncture group. **Allen et al, Anaesth Intensive Care 1994**

A Chinese prospective paired trial ($n = 40$) investigated the effect of electro-acupuncture (EA) at **P-6** during different times of the day (*chen*-time: 7–9am, *xu*-time: 7–9pm) on left ventricular function (LVF) in patients with **coronary heart disease**. EA performed at *chen*-time improved LVF, while EA administered under the same conditions during *xu*-time led to an impairment of LVF. **Li et al, J Tradit Chin Med 1994**

A British prospective case study ($n = 27$) reports that wearing a special wrist band decreased **pregnancy-related nausea and vomiting** more than 50%. **Stainton et al, Health Care Women Int 1994**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **pregnancy-related nausea and vomiting**. There was a significant decrease of nausea in the P-6 acupressure group. The incidence of vomiting did not differ from the sham control group. **Belluomini et al, Obstet Gynecol 1994**

A German prospective, randomised, placebo-controlled, single-blinded trial ($n = 60$) investigated the effect of acupressure at **P-6** on **nausea** in patients undergoing **gynaecological operations of longer duration** (6–8 h). Nausea was reduced from 53% in the placebo group to 23% in the acupressure group. **Gieron et al, Anaesthesist 1993**

A Chinese prospective, randomised, placebo-controlled, single-blinded trial ($n = 15$) investigated the effect of acupuncture at **P-6**, **HE-7** and **HE-3** on the frequency and duration of **angina pectoris** during exercise. The control groups received either sham acupuncture or no treatment at all. In the verum group, the anginal attack occurred significantly later and the duration of the attack after stopping the exercise was significantly shorter than in the control groups. **Zhou et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1993**

A Taiwanese prospective, randomised, controlled trial ($n = 120$) investigated the effect of **P-6** acupoint injection with 0.2 ml 50% glucose in water compared to intravenous injection of 20 micrograms/kg droperidol for the prevention of **vomiting after gynaecological laparoscopy**. A non-controlled comparison-group received no treatment at all. The incidence of vomiting significantly decreased in both the P-6 group and the droperidol group. **Yang et al, Acta Anaesthesiol Scand 1993**

An Italian prospective, randomised, placebo-controlled, double-blinded cross-over study ($n = 60$) investigated the effect of unilateral and bilateral acupressure at **P-6** for the treatment of **morning sickness** compared to acupressure at a sham point. Compared to the placebo group both unilateral and bilateral acupressure at P-6 significantly reduced the frequency of morning sickness from 60% to 30%. **De Aloysio et al, Obstet Gynecol 1992**

A US ‘n of 1’ trial investigated the effect of **P-6** wristbands on **nausea and vomiting in hospice patients** ($n = 6$). There were no significant differences between the verum group, the placebo group and the group without wristbands. **Brown et al, Am J Hosp Palliat Care 1992**

A Canadian prospective, randomised, controlled, double-blinded trial ($n = 90$) investigated the effect of acupuncture at **P-6**, administered after induction of anaesthesia, on the frequency of **nausea and vomiting after strabismus surgery in children**. The control group received droperidol. There was no significant difference between the two groups. **Yentis et al, Can J Anaesth 1992**

A Canadian prospective, randomised, controlled, single-blinded trial ($n = 45$) investigated the effect of acupuncture at **P-6** administered after induction of anaesthesia on the frequency of **nausea and vomiting after tonsillectomy in children**. There was no significant difference between the verum group and the untreated control group. **Yentis et al, Br J Anaesth 1991**

A US prospective, randomised, controlled, double-blinded trial ($n = 66$) investigated the effect of acupuncture at **P-6** administered after induction of anaesthesia on the frequency of **nausea and vomiting after strabismus surgery in children**. There was no significant difference between the treatment group and placebo group. **Lewis et al, Br J Anaesth 1991**

An Irish case study ($n = 100$) reports that TENS at **P-6** reduced **chemotherapy-induced nausea** by 75%. **Dundee et al, J R Soc Med 1991**

A Swedish prospective, randomised, controlled cross-over study ($n = 21$) investigated the effect of acupuncture at **P-6**, **HE-5**, **BL-15**, **BL-20** and **ST-36** on the frequency of **angina pectoris** (AP) and the performance before onset of pain during exercise in patients suffering from coronary heart disease. The frequency of AP attacks decreased significantly from 11 to 6 attacks per week. Accordingly, the performance before onset of pain during exercise increased significantly from 82 W to 94 W. **Richter et al, Eur Heart J 1991**

A Chinese prospective, randomised, non-blinded, controlled trial ($n = 64$) investigated the effect of a 3 ml normal saline injection into **P-6** on the incidence of **postoperative vomiting**. Compared to the control group, vomiting was significantly reduced (2 vs 10 patients) in the treatment group. **Shyr et al, Ma Zui Xue Za Zhi 1990**

A British prospective, randomised, controlled experimental trial ($n = 18$) compared the effectiveness of acupressure bands at **P-6** on **motion sickness** to a placebo and 0.6 mg scopolamine. Only the subjects taking scopolamine showed a significant increase in tolerance to a laboratory nauseogenic cross-coupled motion challenge. **Bruce et al, Aviat Space Environ Med 1990**

A Taiwanese prospective, randomised, controlled trial ($n = 100$) compared the effect of electro-acupuncture (EA) and TENS at **P-6** on **nausea and vomiting after laparoscopy**. Compared to the control group (44%), nausea was significantly reduced in the EA and perchlorperazine group (12% each), but not in the TENS group (36%). **Ho et al, Anaesthesia 1990**

A British prospective, randomised, controlled trial ($n = 162$) compared the effect of acupressure wristbands at **P-6** on **postoperative nausea and vomiting** in the recovery room with conventional anti-emetic treatment. Acupressure at P-6 significantly reduced the intensity of the nausea. The reduced anti-emetic requirement and the reduced incidence of vomiting were not statistically significant. **Barsoum et al, J R Soc Med 1990**

A Chinese prospective, randomised, controlled trial ($n = 40$) compared the effect of electro-acupuncture (EA) at **L.I.-4** and **P-6** on intra-operative **analgesia** in patients undergoing **thyroidectomy** with superficial cervical plexus block. There was no significant difference between the two groups. **Ouyang et al, Zhen Ci Yan Jiu 1990**

An Irish prospective, randomised, sham-controlled, single-blinded trial investigated the effect of pre-operative electro-acupuncture and acupuncture at **P-6** on the frequency of **nausea and vomiting after minor gynaecological surgery** compared to sham acupuncture. The frequency of nausea and vomiting was significantly reduced in the P-6-group only. **Dundee et al, Br J Anaesth 1989**

A US prospective, randomised, controlled cross-over study ($n = 16$) investigated the efficacy of acupressure wristbands at **P-6** on **morning sickness**. Use of the acupressure wristbands significantly relieved morning sickness and also significantly reduced anxiety, depression and behavioural dysfunction. **Hyde, J Nurse Midwifery 1989**

An Irish prospective, randomised, sham-controlled cross-over study ($n = 10$) investigated the effect of electro-acupuncture on **P-6** on **chemotherapy-induced nausea**. In the verum group sickness was either completely absent or reduced considerably in 97% of patients and no side effects were encountered. The same research team confirmed these results in a larger-scale case study ($n = 105$). **Dundee et al, J R Soc Med 1988**

An Irish prospective, randomised, sham-controlled trial ($n = 350$) compared the effect of acupressure at P-6 on **morning sickness** to sham acupressure and a control group. Only acupressure at P-6 significantly relieved the symptoms of morning sickness. **Dundee et al, J R Soc Med 1988**

A prospective, randomised, placebo-controlled trial ($n = 75$) investigated the effect of acupuncture at **P-6** on **nausea and vomiting after minor gynaecological surgery**. Acupuncture at P-6 significantly reduced both nausea and vomiting. **Dundee et al, Br Med J (Clin Res Ed) 1986**

P-8 → P-6; LU-10

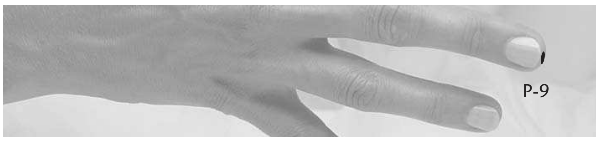

P-9 A larger-scale Chinese case study ($n = 100$) describes the positive effect of **P-9** on **infantile morbid night crying** in otherwise healthy children. **Zhao, J Tradit Chin Med 2002**

Triple Burner channel

T.B.-5 → P-6; L.I.-4; L.I.-12; L.I.-10

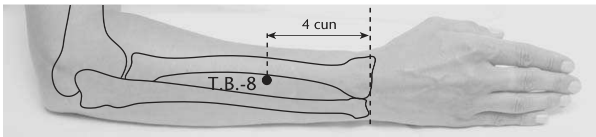

T.B.-8 Acupuncture at **T.B.-8** and **Du-15** activates areas important for **word generation** in the right inferior frontal gyrus, but not in the left inferior frontal gyrus. **Li et al, Hum Brain Mapp 2003**

A Russian case study reports the positive effects of acupuncture at **T.B.-8**, **T.B.-9**, **ST-20**, **ST-36**, **G.B.-14** and **Du-20** in 66 patients with **duodenal ulcer** accompanied by apparent shifts in psychovegetative correlations. **Kravtsova et al, Vopr Kurortol Fizioter Lech Fiz Kult 1994**

T.B.-9 → T.B.-8

T.B.-18 → P-6

Gall Bladder channel

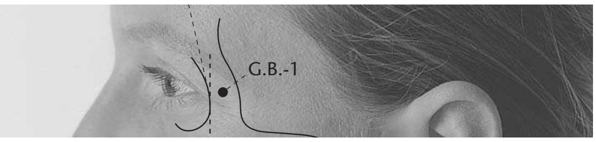

G.B.-1 An earlier Italian case study reports the successful application of **G.B.-1** for **heroin withdrawal** in nine male heroin addicts. After six months, 66% of these patients were classified as abstinent. Besides reservations regarding methodology of such a small-scale trial, the accompanying medication also poses problems in evaluating the effect of the acupuncture treatments. **Cocchi et al, Minerva Med 1979**

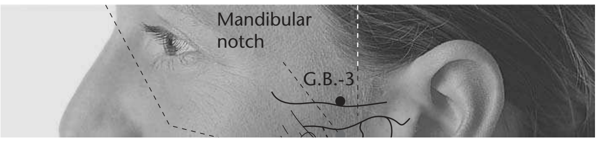

G.B.-3 A US prospective, randomised, double-blinded, experimental cross-over study ($n = 20$) investigated the effect of TENS at **G.B.-3** on the **anaesthetic requirement** (desflurane) to prevent purposeful movement of the extremities in response to noxious electrical stimulation. In comparison with the placebo group, bilateral application of TENS at G.B.-3 significantly reduced the desflurane requirement by 11%. **Greif et al, Anesthesiology 2002**

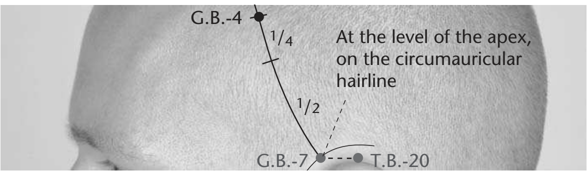

G.B.-4 Following the induction of an **ischaemic cerebral infarction** by occlusion of the middle cerebral artery in rats, animals treated with electro-acupuncture (EA) at **G.B.-4**, **G.B.-5**, **G.B.-6** and **G.B.-7** showed a significantly faster recovery than animals in the untreated control group or in the group treated with EA at **Du-20** and **Du-26**. Infarct volume was significantly reduced in the EA-treated groups. The authors discuss the acupuncture-induced production of vascular endothelial growth factor (VEGF) in astrocytes of the peri-infarct area as a causative factor. **Wang et al, Neurol Res 2003**

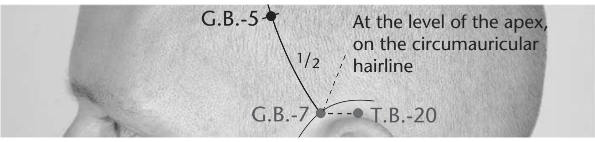

G.B.-5 → G.B.-4; ST-8

A Chinese case study recommends a particular needling technique (from **G.B.-5** or **Ex-HN-5** to **G.B.-8**) for the treatment of **migraine**. **Gan et al, J Tradit Chin Med 1986**

G.B.-6 → BL-7; G.B.-4

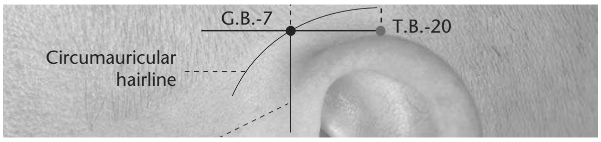

G.B.-7 → G.B.-4
A large-scale Chinese case study (*n* = 500) reports the successful application of acupuncture at **G.B.-7** in the treatment of **hemiplegia** after apoplexy. **Sun et al, J Tradit Chin Med 1985**

G.B.-8 → G.B.-5

G.B.-11 → ST-8

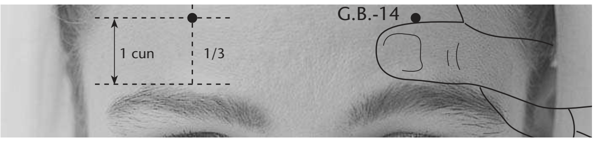

G.B.-14 → T.B.-8; ST-8; LIV-3
A case study reports that 63 patients with **facial paralysis** were treated with acupuncture at **G.B.-14**, **BL-1**, **ST-2**, **L.I.-20**, **ST-4**, **ST-6** and **ST-18** plus auxiliary points according to the underlying TCM pattern. Just over half of the patients were cured and all except one patient showed a marked improvement. **Liu et al, Zhen Ci Yan Jiu 1992**

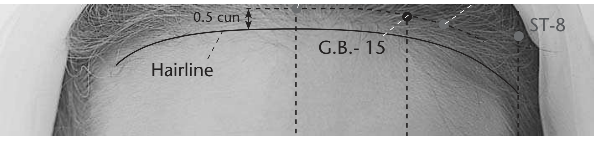

G.B.-15 An uncontrolled, non-randomised, group-comparison trial investigated the effect of acupuncture on **childhood diarrhoea**. Group I received scalp acupuncture (after Yamamoto) with points on line 3 starting at **G.B.-15**; Group II received body acupuncture; and Group III was treated with antibiotics. Group I showed the best results, followed by Group II and Group III, with results in Group I and II significantly better than in Group III. However, besides some methodological shortcomings, the authors did not seem to have a very firm grasp of their subject matter: while using the correct points for the correct indication, they refer to line 3 as line 2 and to G.B.-15 as BL-15! **Lin et al, J Tradit Chin Med 1993**

G.B.-20 → ST-8; LIV-3; S.I.-18; L.I.-15

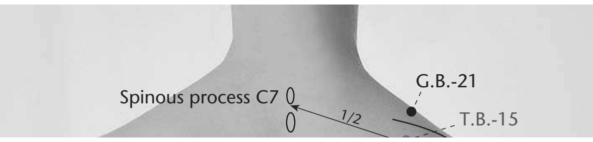

G.B.-21 → L.I.-15
A Chinese case study reports the successful treatment of **achalasia of the cardia** with acupuncture at **G.B.-21**. **Shi et al, J Tradit Chin Med 1994**

G.B.-24 → LIV-14

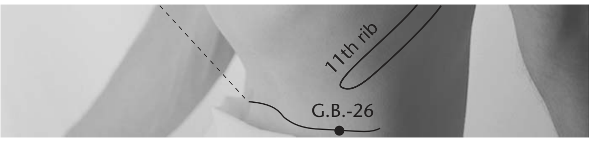

G.B.-26 An earlier prospective group-comparison trial investigated the **analgesic effect** of electro-acupuncture (EA) at **G.B.-26**, **ST-36**, **SP-6** and **HE-7** in patients who underwent hysterectomy by subumbilical midline incision. EA for 40 minutes had the same analgesic effect as 30 mg pentazocine. However, in contrast to pentazocine; EA also improved the vital capacity of the lung for 3–4 hours after the treatment. **Facco et al, Am J Chin Med 1981**

G.B.-30 → L.I.-12; L.I.-15

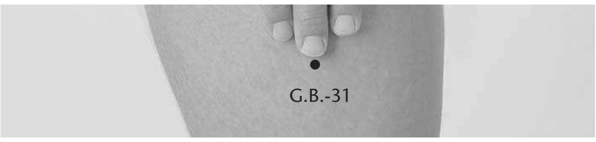

G.B.-31 → L.I.-4
Three publications by a Canadian research team led by Romita demonstrated that electro-acupuncture (EA) at **G.B.-31**, **ST-32** and **ST-36** resulted in a more pronounced inhibition of the **nociceptive tail withdrawal reflex** than classical acupuncture in slightly anaesthetised rats. **Romita et al, Brain Res 1997, Romita et al, Brain Res Bull 1997, Romita et al, Brain Res 1996**

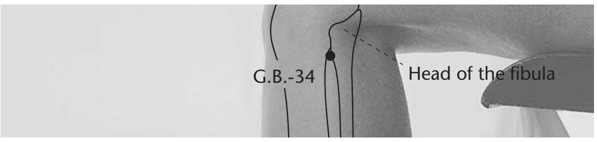

G.B.-34 → BL-10; Du-20; ST-36; L.I.-4; L.I.-15
A Korean prospective randomised placebo-controlled experimental trial (*n* = 10) investigated the effect of acupuncture at **G.B.-34** on motor cortex activities as evidenced by **fMRI**. Bilateral sensorimotor areas BA 3, 4, 6 and 7 were activated during acupuncture, providing a basis for future investigations regarding therapeutic interventions in **stroke** patients. **Jeun et al, Am J Chin Med 2005**

A Chinese prospective, randomised, controlled trial (*n* = 20) investigated the effectiveness of electro-acupuncture (EA) at **G.B.-34** and **ST-38** on **epicondylitis lateralis (tennis elbow)** compared to manual acupuncture at the same points. After six treatments over a period of two weeks, the EA group showed significantly better results in relation to pain relief (Pain visual analogue scale) and pain free hand grip strength (PFG). **Tsui et al, Acupunct Electrother 2002**

A British, prospective, randomised, sham-controlled, double-blinded experimental trial ($n = 240$) investigated the effect of transcutaneous electric nerve stimulation (TENS) at **G.B.-34** on mechanical **pain thresholds**. Low frequency, high intensity stimulation showed a significant analgesic effect that was sustained post-stimulation. **Chesterton et al, Pain 2002**

G.B.-37 → P-2; P-5; L.I.-15

G.B.-38 → P-5

G.B.-39 → LIV-3

G.B.-40 → LIV-3

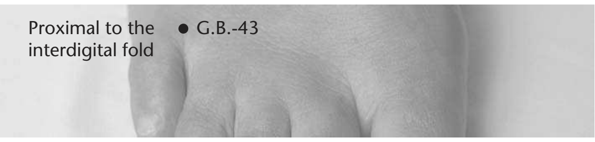

G.B.-43 An Austrian prospective, randomised, placebo-controlled, single-blinded trial investigated the effect of laser acupuncture at **G.B.-43** on cerebral cortical and subcortical activations as evidenced by **fMRI**. Only true laser acupuncture led to ipsilateral activations within the thalamus, nucleus subthalamicus, nucleus ruber, the brainstem, and the Brodmann areas 40 and 22. **Siedentopf et al, Laser Med Sci 2005**

A purely descriptive study investigated the effect of **G.B.-43** and **L.I.-4** on **somatosensory evoked potentials**. Particular attention was paid to the effect of the *deqi*-sensation. **Wu et al, Zhen Ci Yan Jiu 1993**

Liver channel

LIV-1 → SP-1

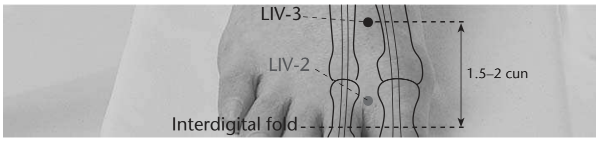

LIV-3 → Du-20; BL-60; L.I.-4; L.I.-15; SP-6

In a Chinese-German single-blinded experimental trial, acupuncture at **LIV-3** and **G.B.-40** resulted in the activation of the secondary somatosensory cortical areas, frontal areas, the right side of the thalamus and the left side of the cerebellum. Sham-points had no effect. **Fang et al, Neuroradiology 2004**

An Italian open, prospective, randomised group-comparison trial investigated the effect of acupuncture, laser acupuncture and TENS on **LIV-3**, **SP-6**, **L.I.-4**, **G.B.-20**, **Du-20** and **Ex-HN-5** on **transformed migraines**. The number of days with headache per month significantly decreased during treatment in all groups. Acupuncture showed the best effectiveness over time. **Allais et al, Neurol Sci 2003**

An Italian prospective, randomised, controlled, non-blinded trial ($n = 160$) investigated the effectiveness of acupuncture at **LIV-3**, **SP-6**, **ST-36**, **Ren-12**, **L.I.-4**, **P-6**, **G.B.-20**, **G.B.-14**, **Ex-HN-5** and **Du-20** on **migraine attacks** without aura compared to flunarizine. The number of attacks after two and four months of therapy was significantly lower in the acupuncture group than in the flunarizine group, and analgesic consumption was accordingly also significantly lower in the acupuncture group. Pain intensity was significantly reduced only by acupuncture treatment. **Allais et al, Headache 2002**

A British prospective, randomised, placebo-controlled cross-over study ($n = 56$) investigated the effect of **LIV-3** on **rheumatoid arthritis (RA)**. There were no significant changes between the treatment and placebo group. The treatment of only one point for a condition such as RA is unusual. **David et al, Rheumatology (Oxford) 1999**

A US prospective, randomised, sham-controlled, single-blinded trial ($n = 10$) investigated the effect of electro-acupuncture (EA) on **LIV-3**, **ST-36** and **L.I.-11** on lowering diastolic blood pressure in hypertensive subjects. Compared to the sham group, diastolic blood pressure showed a significant decrease in the EA-group. **Williams et al, Phys Ther 1991**

LIV-13 → Du-20

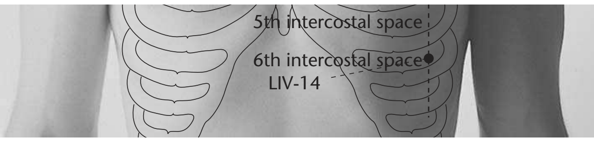

LIV-14 A Chinese prospective, non-randomised, controlled, non-blinded trial ($n = 56$ plus control group) investigated the effect of water injections at **LIV-14**, **G.B.-24** and **Ren-14** on biliary colic pain. The pain disappeared in 57% of cases and showed improvement in a further 39%. This result is significantly better compared to the control group receiving conventional treatment. **Jiang et al, J Tradit Chin Med 1995**

Du mai (governing vessel)

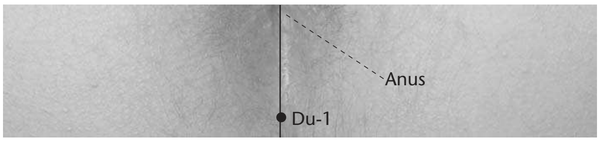

Du-1 A prospective, controlled, group-comparison trial in **hypogalactic postpartum sows** ($n = 42$) showed that apitherapy (acupuncture with bee venom) at **Du-1** and **ST-18** was superior to treatment with conventional drugs. **Choi et al, J Vet Sci 2001**

A prospective, controlled, group-comparison animal experiment showed that apitherapy at **Du-1** and **ST-25** in piglets with preweaning **diarrhoea** ($n = 91$) is as effective as conventional treatment with drugs. **Choi et al, Am J Chin Med 2003**

In an earlier prospective, controlled, group-comparison trial in piglets ($n = 44$) with induced **enteropathogenic *Escherichia coli*** diarrhoea, traditional acupuncture at **Du-1**, **ST-36**, **BL-20**, **Ren-12** and **ST-25** produced the best results in comparison to the groups treated with electro-acupuncture or antibiotics. There were too many groups given the number of cases. **Hwang et al, Am J Vet Res 1988**

Several animal experiments investigated the effect of acupuncture on signalling and modulating pain. Acupuncture at **Du-1** and **Du-2** inhibited the reaction on **experimental pain stimuli**. **Bing et al, Neuroscience 1991, Xu et al, Zhen Ci Yan Jiu 1989, Xu et al, Zhen Ci Yan Jiu 1989**

Du-2 → Du-1

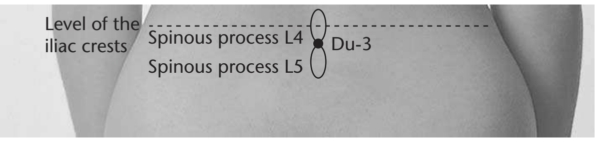
Level of the iliac crests
Spinous process L4
Spinous process L5
Du-3

Du-3 Rats treated with electro-acupuncture at **Du-3**, **BL-54** and **BL-60** 15 minutes after a standardised **spinal cord injury** at T8, showed a marked improvement (regarding both morphology and functioning) three days post-operatively compared to rats not treated with acupuncture. However, none of the beneficial effects occurred in rats given acupuncture treatment 24 hours after spinal cord injury. **Politis et al, Acupunct Electrother Res 1990**

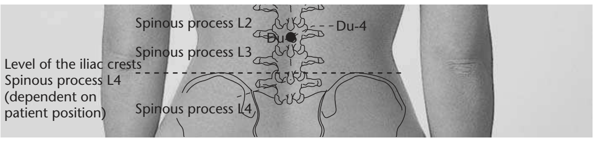
Spinous process L2
Du-4
Du-5
Spinous process L3
Level of the iliac crests
Spinous process L4
(dependent on
patient position)
Spinous process L4

Du-4 A Turkish prospective, randomised, non-blinded, controlled group-comparison trial ($n = 24$) investigated the efficacy of acupressure at **Du-4**, **Du-15**, **Du-20**, **BL-23**, **BL-28**, **BL-32**, **HE-7**, **HE-9**, **ST-36**, **SP-4**, **SP-6**, **SP-12**, **Ren-2**, **Ren-3**, **Ren-6**, **KID-3** and **KID-5** on **enuresis** compared to oxybutynin given to the control group. Acupressure was administered to 12 patients by the parents. While treatments in both groups were successful, complete recovery was significantly higher in the acupressure group (83.3%) compared to the control group (58.3%). **Yuksek et al, J Int Med Res 2003**

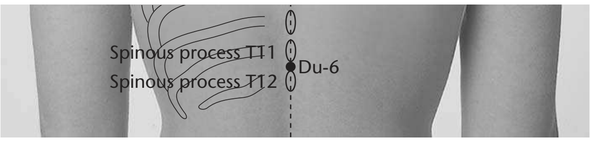
Spinous process T11
Spinous process T12
Du-6

Du-6 An earlier Chinese case study reports the application of **Du-6** for **acupuncture anaesthesia for hysterectomies. Authors' Collective, Chin Med J (Engl) 1978**

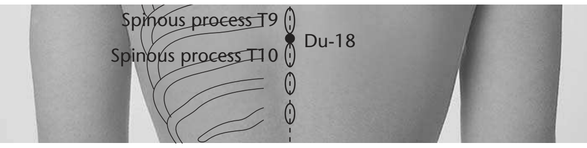
Spinous process T9
Spinous process T10
Du-18

Du-8 An experiment in rats supports the effect of **Du-8** on **seizures** (by releasing melatonin) and on **cerebral infarction** (reduction of the cerebral infarction volume due to release of taurine and decrease of aspartame). **Chao et al, Acupunct Electrother Res 2001, Zhao et al, Acupunct Electrother Res 1997, Yin et al, Zhen Ci Yan Jiu 1994**

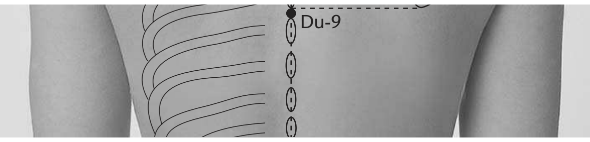

Du-9 The effect of electro-acupuncture at **Du-9** was investigated in 23 patients suffering from **coronary heart disease**. The results showed a slight dilation of the coronary arteries and a slight decrease of the heart rate frequency (monitored by coronary arteriography). **Yan et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1998**

Two earlier Chinese case studies report decreased symptoms in patients with coronary heart disease after acupuncture or tuina at Du-9. Wang et al, Zhong Xi Yi Jie He Za Zhi 1988, Wang et al, Zhong Xi Yi Jie He Za Zhi 1987

A Chinese case study reports the successful treatment of **angina pectoris** by implanting a micro-depressor at **Du-9**. **Wang et al, Zhong Xi Yi Jie He Za Zhi 1988**

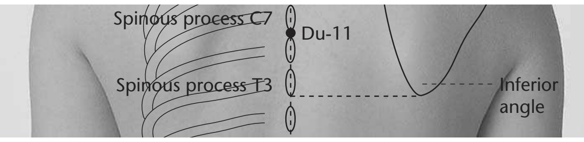
Spinous process C7
Du-11
Spinous process T3
Inferior
angle

Du-11 Electro-acupuncture at **Du-11** and **Du-16** prior to, during and after an experimentally induced **transient ischaemia** in gerbils suggests the protective influence of these points. They seem to suppress both glutamate release and reperfusion injury following the ischaemic insult. **Pang et al, Am J Chin Med 2003**

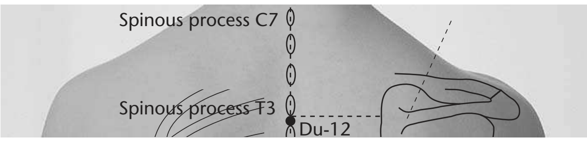
Spinous process C7
Spinous process T3
Du-12

Du-12 A Chinese case study reports that classic acupuncture at **Du-12**, **L.I.-10**, **BL-11** and **S.I.-3** with *deqi* sensation in the direction of the disorder in 55 patients with **cervical vertebra**

disease and **impaired cerebral blood flow capacity** led to a significant improvement of blood flow capacity. **Qie et al, Zhong Xi Yi Jie He Za Zhi 1991**

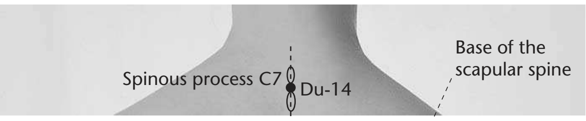

Du-14 → L.I.-4, ST-8; ST-36

A Chinese experimental trial ($n = 40$) investigated the effect of acupuncture at **Du-14** and **P-6** on driving performance during a three-hour driving simulator test. The acupuncture group showed a significantly better driving performance and a decrease in **driving fatigue**, compared to the untreated control group. **Li et al, Accid Anal Prev 2004**

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 40$) investigated the effect of acupuncture at **Du-14 and P-6** on heart rate and driving performance during a three-hour driving simulator test. The acupuncture group showed a significantly decreased **activation of the sympathetic nervous system** and increased **parasympathetic activity** compared to the untreated control group. **Li et al, Eur J Appl Physiol 2003**

A Chinese retrospective case study ($n = 25$) investigated the effect of cupping at **Du-14**, **BL-13**, **BL-20** and **BL-23** on **bronchial asthma** while reducing orally administered corticsone. The author reports that 56% of patients showed a marked improvement after 30 treatments. **Hu, J Tradit Chin Med 1998**

A Chinese prospective case study ($n = 50$) reports the successful treatment of **perennial allergic rhinitis** by applying 10% *Cantharides* extract at **Du-14** and **P-6**. The results showed an effective rate of 88%. The allergic nasal mucosa provocative test of the treated group improved after the treatment, the number of eosinophils and basophils in nasal secretion decreased and the serum total IgE also reduced significantly. **Tang et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 1995**

Du-15 → T.B.-8; Du-4

Du-16 → Du-11

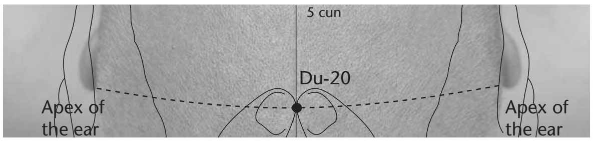

Du-20 → T.B.-8; G.B.-4, ST-8; Du-4; LIV-3; HE-7, L.I.-4

A Chinese prospective randomised controlled single-blinded group-comparison trial ($n = 50$) investigated the effect of a standardised acupuncture protocol, used either on its own or accompanied by needling of **Du-20**, **Du-26** and **HE-7** (either singly or in combination) on **vascular dementia** in hemiplegic patients. Clinical symptoms were observed and scales such as HDS-R, ADL and functional activities questionnaire (FAQ) used for assessment. The group additionally treated with **Du-20 and HE-7** showed marked improvement of memory, orientation and reaction. At the same time mental rigidity and trance decreased. **Du-20** on its own improved understanding, calculating and social adaptation, while **Du-26** was helpful in treating mental retardation, failure to perform daily activities, trance and poor memory. The combination of the three points showed overall the best results in improving intelligence and social adaptation. The particular design of this trial – combining a standard protocol used in both groups with the actually relevant points in the verum group – presents an interesting model for eliminating persistent problems regarding placebo methods and blinding. However, in this particular trial the number of participants (5 groups of 10 subjects) was too low to determine if the results were statistically significant. **Lai et Huang, Chin J Integr Med 2005**

A German prospective, non-blinded, pair-matched, controlled trial ($n = 121$ pregnant women between 30 and 39 weeks of gestation) investigated the effect of acupuncture at **Du-20** and **ST-36** on **fetal and maternal cardiotocographic parameters and maternal circulation**. The main difference was a transient increase of the Fisher score in the treatment group compared to the control group, with a persistent increase of the Fisher score. **Scharf et al, Z Geburtshilfe Neonatol 2003**

A Croatian prospective, randomised, placebo-controlled trial ($n = 57$) investigated the effectiveness of **Du-20**, **L.I.-4**, **Ren-3**, **Ren-4**, **Ren-6**, **G.B.-34**, **BL-23**, **SP-6** and **HE-7** on **primary dysmenorrhoea** compared to placebo-acupuncture. Subjects were observed for a period of two years. After one year, 93.3% of acupuncture patients were symptom-free and/or not taking medication compared to only 3.7% of the placebo patients. **Habek et al, Gynakol Geburtshilfliche Rundsch 2003**

A Polish prospective, randomised, controlled, non-blinded trial ($n = 69$) investigated the effect of acupuncture at **Du-20**, **Ren-12**, **ST-36**, **ST-21**, **ST-25**, **LIV-3**, **LIV-13**, **P-6** and **HE-7** in addition to a low calorie diet in the treatment of obesity (BMI approximately 33 kg/m^2). This diet consisted of a daily intake of about 1200–1500 kcal, which was about 1000 kcal below the daily energy requirement of the patients. The acupuncture group showed a significantly better result after only 12 treatments over a period of 6 weeks. The mean body weight decreased from 85.5 kg to 72.7 kg, while it only decreased to 77.1 kg in the diet group. **Wozniak et al, Ginekol Pol 2003**

A German prospective, randomised, placebo-controlled, double-blinded trial ($n = 56$) investigated the effect of acupuncture at **Du-20**, **Ex-HN-6**, **HE-7**, **P-6** and **BL-62** in the treatment of **depression** and **generalised anxiety**. Results were assessed using questionnaires. After 10 treatments, the verum acupuncture group showed a significantly larger clinical improvement compared to the placebo group (61% vs 21%). **Eich et al, Fortschr Neurol Psychiatr 2000**

Du-26 → G.B.-4; Du-20

Ren mai (conception vessel)

Ren-2 → Du-4

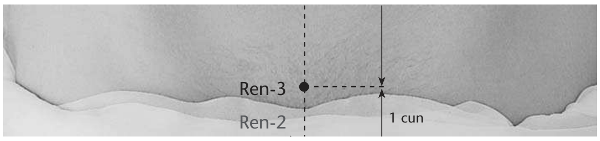

Ren-3 → Du-4; Du-20

A Taiwanese prospective, randomised, controlled trial ($n = 60$) investigated the influence of electro-acupuncture (EA) at **Ren-3**, **Ren-4** and **BL-32** on the **neurogenic bladder of spinal cord injured patients** compared to the conventional bladder training programme received by the control group. The acupuncture group took significantly less time to achieve balanced voiding than the control group (on average 28 days less). The best results were achieved in patients who received acupuncture within three weeks after injury (39 days less). **Cheng et al, Spinal Cord 1998**

Ren-4 → Du-20; ST-25; Ren-3

Ren-6 → Du-4; Du-20; Ren-17

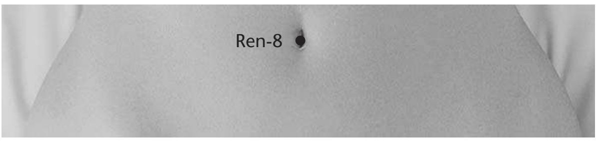

Ren-8 → P-6

A Chinese prospective, randomised, controlled, non-blinded trial ($n = 56$) investigated the effect of moxibustion at **Ren-8** in addition to chemotherapy for **nasopharyngeal carcinoma** (stage III and IV). There were significantly fewer toxic and side effects in the moxibustion group compared to the control group. The five-year survival rate in the two groups were 50.0% and 35.7% respectively. **Chen et al, Zhongguo Zhong Xi Yi Jie He Za Zhi 2000**

A Chinese prospective, randomised, controlled trial ($n = 140$) investigated the effect of applying a topical ointment at **Ren-8** on **epigastric pain**. Results in the treatment group were significantly better than in the control group (94% vs 53%). **Ba et al, J Tradit Chin Med 1999**

Ren-12 → L.I.-4; Du-1; Du-20; LIV-3; Ren-17

Ren-13 → P-6

Ren-14 → LIV-14

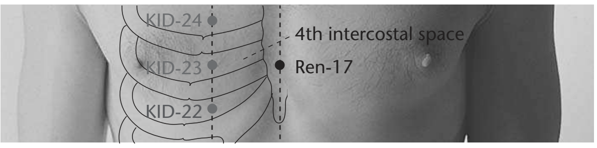

Ren-17 A Chinese prospective, randomised, controlled trial ($n = 60$) investigated the effect of **Ren-17**, **Ren-12**, **Ren-6**, **ST-36 and SP-10** on the symptoms of **vascular dementia** compared to conventional treatment. Both acupuncture and conventional treatment showed a highly significant increase of MMSE, HDS-R and ADL scores. However, the results of the acupuncture group were statistically more significant regarding MMSE, HDS-R and the total rate of effectiveness (80% vs 47%). **Yu et al, Neurol Res 2006**

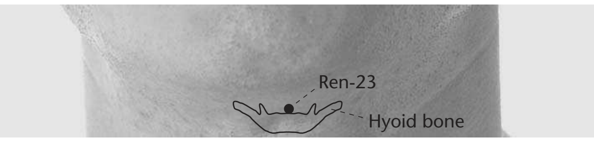

Ren-23 A Chinese retrospective case study ($n = 120$) reports the successful treatment of **pseudobulbar palsy** with acupuncture at **Ren-23** and **LU-5. Wang et al, J Tradit Chin Med 1998**

Ren-24 A German prospective, single-blinded, controlled trial ($n = 41$) reports that acupuncture at **Ren-24** significantly reduced the severity of the **gag reflex** during transoesophageal echocardiography compared to acupuncture at a sham point. **Rosler et al, J Altern Complement Med 2003**

Extra points

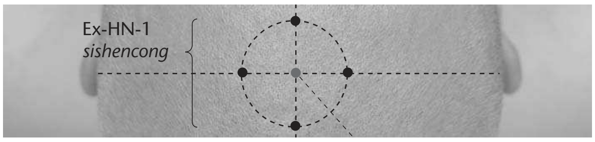

Ex-HN-1 → BL-7

A Taiwanese, prospective, randomised, placebo-controlled, single-blinded experimental trial ($n = 9$) investigated the effect of the *sishencong*-points (**Ex-HN-1**) on the variance of heart rate frequency. The authors observed an increase in the vagal activities but a suppression of the sympathetic activity. **Wang et al, Auton Neurosci 2002**

A Chinese case study ($n = 51$) reports the successful treatment of **hyperthyreosis** with acupuncture at Ex-HN-1. **Xie et al, J Tradit Chin Med 1994**

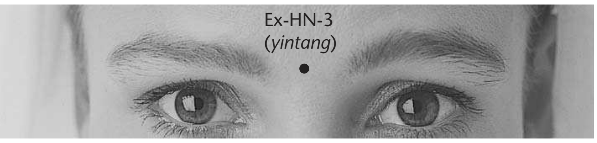

Ex-HN-3 An Austrian prospective, randomised, single-blinded experimental cross-over study ($n = 25$) showed that acupuncture (but not laser acupuncture) at *yintang* monitored by EEG had a sedating effect. Clinical application of this point, for example for **sleeping disorders**, requires further investigation. **Litscher et al, Eur J Anaesthesiol 2004**

Ex-HN-4 → S.I.-18; L.I.-4

Ex-HN-5 → G.B.-5; LIV-3

Ex-HN-6 → HE-7; Du-20

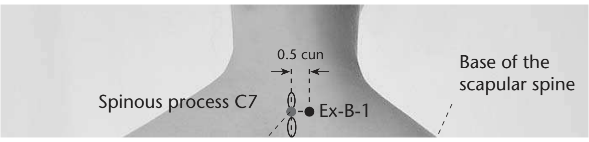

Ex-B-1 A Chinese prospective, randomised, controlled trial ($n = 84$) investigated the effect of a herbal plaster at **Ex-B-1**, **BL-11**, **BL-12**, **BL-13** and **BL-15** on **chronic bronchitis** compared to intramuscular injection therapy. In the plaster group the clinical total effective rate was 93.2% immediately after the treatment course and 91.5% after 18 months (control group: 80% and 80%). X-rays of the thorax showed the effectiveness of the plaster as 40.7% and 89.8% after 18 months (control group: 20% and 76%). **Huo et al, Zhongguo Zhong Xi Yi Jie He za Zhi 2001**

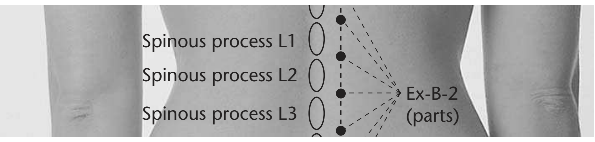

Ex-B-2 A Chinese case study ($n = 43$) reports the positive effects of treating **ulcerative colitis** with acupuncture at **Ex-B-2** as well as stimulation with a plum-blossom needle at **BL-22** and **BL-25. Yue and Zhenhui, J Tradit Chin Med 2005**

A Chinese retrospective case study ($n = 168$) reports the successful treatment of **sciatica** with acupuncture at the **Ex-B-2** points. **Pei et al, J Tradit Chin Med 1994**

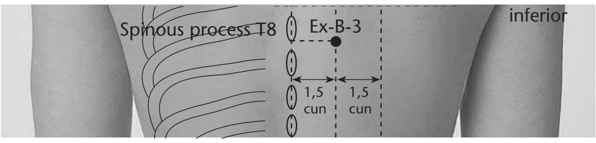

Ex-B-3 An experiment in diabetic rabbits demonstrated that electro-acupuncture at **Ex-B-3** decreased plasma glucose levels and inhibited the release of pancreatic glucagon. The effect was increased by adding **ST-36**. This latter point led to no significant changes when needled alone. **Zeng et al, J Tradit Chin Med 2002**

Ex-UE-9 → L.I.-4

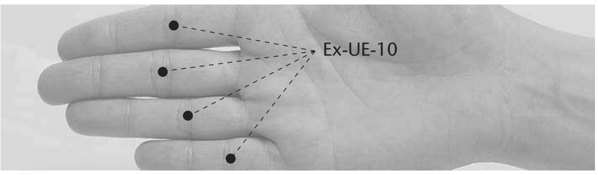

Ex-UE-10 A case study describes the successful treatment of **hiccups** with acupuncture at **Ex-UE-10**. **Qi, J Tradit Chin Med 1993**

Ex-UE (***JIANNEILING***) → L.I.-15

Ex-UE (***JIANQIAN***) → L.I.-4

Ex-LE-4 → ST-35

Ex-UE (***ZHONGPING***)

A Chinese retrospective case study ($n = 115$) reports the successful treatment of **shoulder and back pain** with acupuncture at ***zhongping***. The point *zhongping* is located midway between ST-36 and ST-37. **Wang et al, J Tradit Chin Med 1995**

9.3 Summary

The data presented clearly demonstrates that several actions of a number of acupuncture points are undoubtedly effective. These points (for example P-6) were tested according to the highest scientific standards, comparing standard acupuncture to deep acupuncture at non-acupuncture points (a methodology considered the most superior acupuncture placebo).

This proof is even more important since the results of the two largest acupuncture trials to date, ART (Acupuncture Randomized Trials) and GERAC (German Acupuncture Trial), may have led to the impression that, while acupuncture is indeed effective, its effect is independent of the location of the acupuncture points. The hitherto unsurpassed size of the trials and the undoubtedly intelligent and biomathematically impeccable design (especially of the GERAC trial) further contribute to this impression. Due to their flexible point prescriptions and their complexity, these studies have not been included in section 9.2.

But what are the basic results of these trials? (up to May 2006: **Witt et al, Lancet 2005; Linde et al, JAMA 2005; Melchart et al, Br Med J 2005; Brinkhaus, Arch Intern Med 2006**; **Diener et al, Lancet Neurol 2006**)

For both **gonarthrosis** and **lower back pain,** acupuncture had significantly better results than the untreated waitlist (ART) or conventional therapy (GERAC). For **migraine**, acupuncture is significantly better than the untreated waitlist (ART), and treatments with acupuncture for only six weeks had the same results as treatments with conventional medicine for six months (GERAC). For **tension headaches**, acupuncture provided significantly better results than the waitlist (ART). A comparison between acupuncture and conventional therapy in the treatment of tension headaches within the framework of GERAC was not possible, since all but two patients refused the stigmatised conventional treatment with amitriptyline due to the high incidence of side-effects. So far the results do not hold any surprises. Their importance lies in the comparison between true and placebo acupuncture (acupuncture at non-acupuncture points).

For the treatment of lower back pain, both studies (GERAC and ART) showed no significant advantage for true acupuncture; the effect of acupuncture therefore seems to be independent of the location of the acupuncture points. For the treatment of gonarthrosis, the ART trial observed a significant short-term effect in favour of true acupuncture. However, without consistent treatments the disorder will progress, further data showing no significant difference between true and sham acupuncture. Results of the GERAC trial confirm the significant short-term results. For the treatment of migraines and tension headaches, the relevant ART trials could not establish a significant difference between true and placebo acupuncture. The migraine section of the GERAC trial found a significantly reduced number of days with headache in subgroup analyses. Regarding tension headaches, up to now the GERAC reports show significant differences in favour of true acupuncture.

How can we explain the good results of placebo acupuncture, especially for the treatment of lower back pain? And how can the differing results between the ART and GERAC trials for gonarthrosis, migraines and tension headaches be explained? Shouldn't the trials, which are both designed according to the criteria listed in Section 9.1, be very reliable because of the high power of their design? The key to the answer – besides the already well-known deficits of the GERAC trials (by international standards, with only 140 hours – lowest standard – of training, questionable choice of true and sham acupuncture points) – lies in the so-called 'internal validity' of the collected data. The external validation, especially of the GERAC trials, was carried out in an exemplary manner: patients and physicians were separately questioned about the treatments by telephone interviewers. This filtered out a high number of patients receiving unacceptable additional therapies, thus increasing the validity of the data. However, it was hardly possible to control the internal validity of the trial. With treatments taking place in hundreds of clinics, no one was able to control if true or sham points were needled. One peculiarity of both trials suggests that physicians treating patients in the placebo group selected true acupuncture points: generally, the participating physicians weren't research scientists, but family doctors who, by participating in the trial, were able to give their patients treatments paid for by their obligatory health insurance, thus securing a part of their own income. Therefore, participation in the trials not only contributed to the income of these physicians, it also meant that the crucial performers of the trials had other goals in mind – patient satisfaction, their income – rather than the trial itself (i.e. collecting reliable data regarding the effectiveness of acupuncture). These conflicting goals particularly affected the placebo-acupuncture, which had a higher potential of negatively affecting patient satisfaction and the income of the participating physicians, and which could not be balanced by external quality assurance through telephone interviewers. The part of the GERAC trials investigating tension headaches provides another interesting insight. This section of the trials prohibited needling of points located on the head and it was here that the treatment protocol was disregarded most often (according to GERAC symposium, RUB 16 November 2005). And it was exactly in this externally validated part of the trial that GERAC observed significant differences between true and placebo acupuncture!

An important lesson to be learnt from the ART and GERAC trials is that size and adherence to biomathematical standards alone will not guarantee quality. If, in a trial above a certain size, the data collection is compromised due to conflicting interests of the initiators of the trial and the treating physicians, this will have a considerable adverse affect on the validity of the study.

The above reservations should result in a rather cautious interpretation of the results of large-scale trials such as GERAC and ART.

Even conventional medical procedures, considered as proven beyond doubt, recently failed spectacularly in controlled, double-blinded trials (**Moseley, N Engl J Med 2002**), which means that the efforts for scientific proof regarding specific details of acupuncture will be seen in a different light.

Appendix

Notes on Chapter 1

1: The *yin yang sho yi mai jiu jing* (Treatise on the Yin Yang Eleven Vessels Moxibustion), a manuscript found at an archaeological site at Mawangdui and dated to 169 BC, contains a simple description of the channel system. In this manuscript, the channels refer to regions of the body rather than to internal organs. It does not mention the Pericardium channel (hand *jueyin*), the channels are not divided into three Yin and three Yang channels, nor is the concept of the Five Phases mentioned.

According to chapter 17 of the *Ling Shu* and chapter 22 of the *Nanjing*, the channels have a centripetal course, originating at the fingers and the toes. They do not represent a closed circuit. Equally, chapter 5 of the *Ling Shu*, describing the 'roots' and 'knots' of the channels (→ 1.2.3) is based on a centripetal model.

In contrast, chapter 21 of the *Su Wen* contains references about the early conceptual stages of the channel theory, listing channel categories:

- The posterior of the body is governed by the *taiyang*
- The anterior of the body by the *yangming*
- The lateral aspect by the *shaoyang*
- The medial aspect by the *taiyin.*

2: Chapters 10 and 15 of the *Ling Shu* describe the channel system as a continuous circuit in the order LU → L.I. → ST etc. But only later texts such as the *Shi Si Jing Fa Hui* ('Commentary on the 14 Channels', 1341 AD) routinely mention the circulation model, first hinted at in chapter 15 of the *Ling Shu*. 'The Yang declines when the Yin increases' – according to Manaka et al (1995); this statement from the classics is at the root of the later theory that the Yin channels represent the ascending flow of Qi, running from the toes to the trunk and from the trunk to the fingers. It was further suggested that the Yang channels represent the descending flow of Qi, running from the fingers to the head and from the head to the toes.

THE BOY'S BOOK of CRYSTAL SETS

BY

W. J. MAY

⁂

BERNARDS (PUBLISHERS) LTD.
THE GRAMPIANS
WESTERN GATE
LONDON W.6

General Editor
WALTER J. MAY

First Published—March, 1954
Second Impression—June, 1954
Third Impression—April, 1955
Fourth Impression—January, 1956
Fifth Impression—February, 1957
Sixth Impression—August, 1957
Seventh Impression—March, 1958
Eighth Impression—October, 1958
Ninth Impression—April, 1959
Tenth Impression—December, 1959
Eleventh Impression—April, 1960
Twelfth Impression—October, 1960

⁂

Printed by V. Cooper & Partners, Ltd., Flitcroft St., W.C.2
for Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.

PREFACE

The Publishers make no apology for reviving interest in the crystal set. For too long the technical press has immersed itself in an orgy of complex superheterodyne receivers and television equipment, all of which is necessarily expensive. In the 1920's, crystal set construction was a fascinating and inexpensive pastime indulged in by the majority of boys of that day. Today this pleasurable hobby can be even more attractive, the introduction of the germanium crystal diode and high performance coils, has opened up new paths which will capture the interest of boys of all ages. Perhaps most important of all, not only can a receiver capable of impeccable reproduction be produced without any technical knowledge; but the cost is no greater than that of a few visits to the cinema

❧

CONTENTS

	Page
Preface - - - - -	4
Introduction - - - -	5
Receiver No. 1 - - - -	12
Receiver No. 2 - - - -	13
Receiver No. 3 - - - -	14
Receiver No. 4 - - - -	16
Receiver No. 5 - - - -	19
Receiver No. 6 - - - -	20
Receiver No. 7 - - - -	23
Receiver No. 8 - - - -	24
Receiver No. 9 - - - -	26
Receiver No. 10 - - - -	27
Receiver No. 11 - - - -	30
Receiver No. 12 - - - -	31

INTRODUCTION

Before any attempt is made to construct a receiver, it is necessary to examine the problems which surround the crystal set, so that the best can be obtained from any of the designs attempted.

Firstly, it must be understood that the crystal set as it is to-day does **not** provide any amplification. It relies entirely on what is fed into it via the aerial and earth system and gives a very faithful replica of the original transmission.

From this it will be obvious that the aerial and earth system must be as efficient as possible if the final results are to be in any way outstanding. This because these are the only means by which the signals are fed to the receiver.

Secondly it is necessary to understand the nature of the transmitted signal, then it will be easy to understand the working of the set and to appreciate the function of each of the components.

When crystal sets first became popular, very little information was generally available, at least, not in a form that school-boys could understand.

We all built sets of all shapes and sizes, with coil designs that had to be seen to be believed, but very few of us had much idea of how they worked.

I well remember the case of a cousin of mine who, in those days acquired a magnificent variable capacitor or condenser as it was then called, this instrument had a most impressively engraved dial of polished ebonite, brass vanes and nickel plated end plates. Having observed similar dials on several highly priced commercial receivers in the town, he at once ripped out the somewhat tattered coil from the family receiver and replaced it with this device. The profound silence which ensued caused considerable amazement and dismay until a better informed adult explained the mysteries of L and C to him. Readers of this manual, however, will be better informed and there is no risk of failure with any of the designs described provided the instructions are carefully followed.

The Signal

To commence, we will assume that an orchestra is playing in a broadcasting studio. Since the principle of radio transmission is electrical it is necessary to change the sound produced by the orchestra into an electrical equivalent. This is carried out by the microphone, which picks up the sound and changes it into minute electric currents.

As they appear at the output of the microphone they are too small to be of use and accordingly are passed through a high power amplifier. These amplified currents could now be transmitted, but unfortunately, as we shall learn, owing to the inherent nature of the signal in this state, transmission over any useful distance would be impractical.

When the music from the orchestra is transformed into electrical currents they are in the form of alternating currents, usually called A.C., that is they rise to a maximum in one direction, fall to a minimum, rise to a maximum in the opposite direction and then fall to minimum again. This process is repeated over and over again. One complete rise and fall in each direction is called a cycle and is drawn in Fig. 1. Every time a note is struck on a piano, vibrations are sent out which reach the ear enabling you to hear it. These vibrations are also spoken of as cycles, they rise and fall in intensity the same way as an alternating current. The number of cycles radiated by any given note over a period of one second are referred to as its frequency. Middie C on the piano sends out 261 cycles every second and is known as having a frequency of 261. The microphone also "hears" the note and in the case of middie C produces minute A.C. at 261 cycles. This can be drawn as in Fig. 2, the only difference between Fig. 1 and 2 is, that the time factor is given so that the frequency can be identified. The higher the pitch of a note the higher the frequency and the lower the pitch the lower the frequency. On a piano the frequency of the top note is 3515 cycles and that of the lower 27 cycles. Those of you who have listened to an organ in a concert hall will have noticed that when a very deep note was played, it sounded like a growl to the ear, but the vibrating fre-

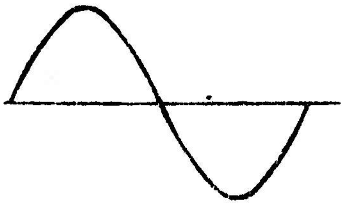

FIG 1

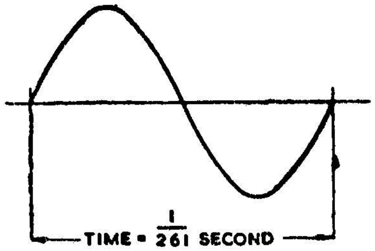

FIG 2

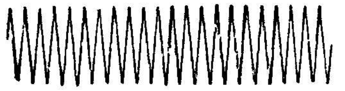

FIG 3

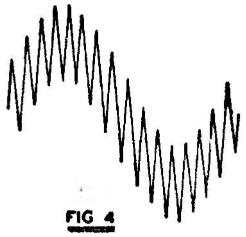

FIG 4

quency could be distinctly felt through the seat. Higher notes have too high a frequency to be observed in this manner.

The range of sounds which can be detected by the human ear are known as audio or low frequencies. From this you will understand that a low frequency amplifier is one which amplifies sound.

So that the transmitter will carry the programme over a useful distance it is necessary to radiate high frequencies. Now, as explained, the programme to be transmitted consists of low frequencies, and to overcome the difficulty, the transmitter generates A.C. of high frequency and combines it with the low frequencies. It will now be understood that the transmitted signal consists essentially of two different parts, a high and a low frequency content.

Fig. 3 gives a representation of the high frequency signal generated by the transmitter. In the case of the London Home Service, the frequency is 908000 cycles.

When referring to a high frequency signal on the medium or long wave-band it is usual to express the frequency in thousands of cycles, thus 908000 cycles becomes 908 kilo-cycles, which in turn may be abbreviated to 908 k/cs. It might at first be thought that by adding the low frequency or L.F. signal to the high frequency or H.F. carrier a form such as in Fig. 4 would result, such a combination is useless, and, so that the original L.F. content can be satisfactorily extracted by the receiver, the L.F. signal must vary the amplitude or output power of the H.F. signal as in Fig. 5.

It is in this form that the signal arrives at the receiving aerial. The aerial in itself is incapable of discriminating between one signal and another, and countless signals will be collected by the aerial at any one time. Many of these are too weak to be of use but the stronger ones must be sorted out since there is no point in receiving several programmes at once.

Fig. 6 shows the basic circuit of the input to a crystal. The coil L possesses a quality known as inductance, and the capacitor C, that of capacitance. If the coil had no capacitance whatever across it, all signals arriving at the aerial would be effectively short circuited to earth. As a matter of interest, it is impossible to obtain this state of affairs since even without any additional capa-

citor any coil **must** contain a certain amount of self capacitance.

By combining a coil and capacitor as in Fig. 6 a peculiar effect is observed, at one particular frequency, the signals are not short circuited to earth, but are developed across the coil. In other words the combined effect of L and C no longer provides a short circuit, **but only at one particular frequency.** If the value of C is altered the effect will be observed at a different frequency; likewise by altering L the frequency at which the effect will take place can be changed. There is a name for this phenomena, the frequency at which it occurs with any given L and C combination is known as the **resonant frequency.**

The values of the coils and capacitors shown in this manual have been carefully chosen so that resonance will be obtained at all frequencies where stations are broadcasting. Broadcasting stations work in bands of frequencies, those of major interest to crystal set constructors are the medium wave-band 1200 k/cs—600 k/cs and the long wave-band 300 kc/s—150 k/cs.

Usually a variable capacitor is used with a fixed inductance to cover one band and an additional coil switched in to increase the inductance to cover the other. In this way the L and C combination can be adjusted to provide resonance at the desired frequency of any given station. In other words you can **select the** station you want by varying C, that is turning the dial of the variable capacitor. This procedure is referred to as tuning. Having selected or tuned the required station it still remains necessary to change the form of the signal back to that of the original transmission. This process is called detection or de-modulation.

Examination of Fig. 5 will show that the signal has been duplicated, in other words, as it rises in one direction it also rises equally in the other. In this form the signal is useless since each half of the signal cancels the other, and if this signal is applied to a pair of headphones silence will result. Obviously some provision must be made to get rid of the unwanted half of the signal, and it is here that the crystal detector must be considered. This device will pass current in one direction only, ignoring any signal in the opposite direction, so that if the signal of Fig. 5 is passed through such a crystal, that of Fig. 7 will result. There is still the H.F. content to be reckoned

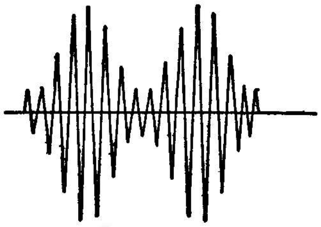

FIG 5

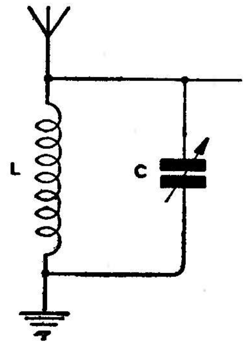

FIG 6

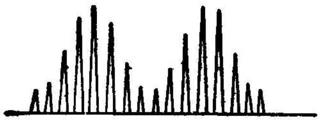

FIG 7

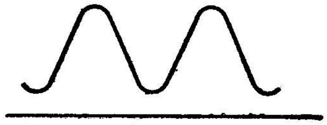

FIG 8

with, fortunately this is easily dealt with, a capacitor connected across the headphone terminals effectively disposes of this, leaving only the audio or L.F. content, as shown in Fig. 8. This audio content, which is a faithful replica of the original transmission is fed to the headphones. These in turn reverse the process of the microphone and transform the electrical currents into sound waves acceptable to the human ear.

Briefly then, your requirements are as follows:—

(1) A good aerial and earth installation, to make the most of the available signals.
(2) A receiver containing:—
  (a) Some form of coil and capacitor (L & C) combination to select or tune in the wanted station.
  (b) A crystal to get rid of the unwanted half of the signal (detection).
  (c) A fixed capacitor across the headphones to get rid of any remaining carrier.
(3) A pair of sensitive high resistance headphones.

Aerials

By this time the intending constructor will be able to appreciate the necessity of a good aerial. It is a point which cannot be over-emphasised. Assuming that you are in the fortunate position of being able to erect an outdoor aerial there are two main considerations, height and length.

One of the best that can be used is the inverted L shown diagramatically in Fig. 9.

It should be erected as high as is practical, every foot counts. The horizontal wire, that is the aerial proper, should have a minimum length of 60′ to which of course the length of the down lead is added. Where it is impossible to erect an aerial with an ideal horizontal length, a compromise must be effected.

Fig. 10 shows a three wire spreader aerial which gives quite a good effective length.

Suitable wire for a receiving aerial will not set any problems, stranded copper about 7/22 gauge is the best. 7/22 means that it consists of 7 strands of 22-gauge wire. This wire may be obtained covered, and for the present purpose is better than the plain or enamelled kind.

Note that insulators are used between the actual aerial wire and its anchoring supports, it is important that these are used, otherwise leakage will occur which will of course spoil its efficiency. Fig. 11 and 12 show how the wire may best be attached to two of the most common types of insulator available. Fig. 11 is of porcelain and is usually referred to as an egg insulator, whereas Fig. 12 shows a more modern (and more expensive) type in glass. If you are using the glass pattern, one is usually sufficient at each end of the aerial, but with the egg type two should be used. No doubt many readers will not be in a position to erect an out-door aerial, and must necessarily be content with an indoor installation.

The next best thing to a good out-door aerial is a replica constructed in a loft. If this form of construction is used, care must be taken when feeding the down lead to avoid close contact with the wall of the house. At the point where the lead feeds under the eaves, a length of rubber tubing can be used to cover the wire. The lead is fed through one of the small spaces left for ventilation purposes. A general idea is given by Fig. 13.

A less elaborate but quite effective aerial can be obtained by using a bed-spring. A length of 7/22 copper insulated wire is connected between the receiver and the spring. The spring should first be cleaned with emery cloth. Remove about 2″ of insulation from the wire and bind it tightly round the prepared spring. The joint may be covered with insulating tape which is obtainable from all electrical stores for a few pence.

Such aerials are quite popular since so many crystal sets are built for bedroom use.

When an aerial is required in the living room, the picture rail can be conveniently used. Insulated screw-eyes are fixed at intervals of about 3 feet along the rail, the wire is firmly anchored to one of them, and stretched right round the room until you arrive back at the starting point. By use of one of the insulated screw-eyes, the wire is secured and the down lead fed to the receiver. The idea is illustrated in Fig. 14.

Soldering

Before leaving the subject of aerials, a few words on soldering will not come amiss. Down leads on outdoor aerials should be soldered, and the same applies to loft types. Apart from aerial leads; earth leads and all the connections in the receiver will require soldering.

For the type of soldering necessary in radio construction an electric iron is the best solution. One of the small types marketed by Adcola or Henleys will be found admirable for the job. These are excellent for actual set construction but are hardly large enough for soldering the down lead to an outdoor type of aerial. Here the heat is dissipated much more quickly and a larger iron is required in order to get the solder to flow.

Ordinary irons which may be heated by a gas flame can be obtained very cheaply from most ironmongers' stores. Assuming the use of such an iron, first heat the iron until the copper bit is giving off a green coloured flame, the iron is now at the correct operating temperature. The bit is now discoloured or oxidised and its tip

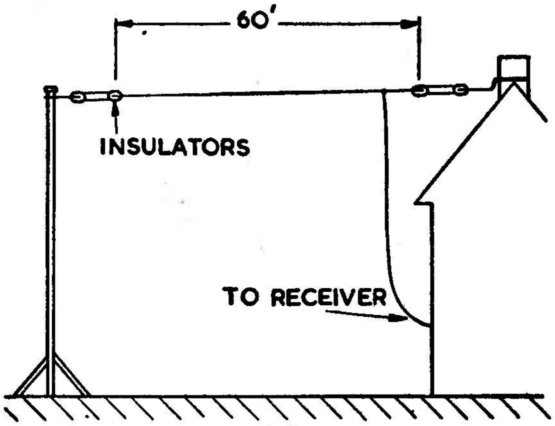
60'
INSULATORS
TO RECEIVER

INVERTED 'L' AERIAL

FIG. 9

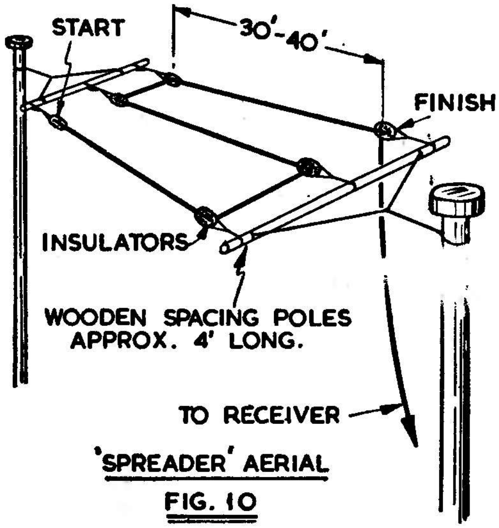
START
30'-40'
FINISH
INSULATORS
WOODEN SPACING POLES APPROX. 4' LONG.
TO RECEIVER

'SPREADER' AERIAL

FIG. 10

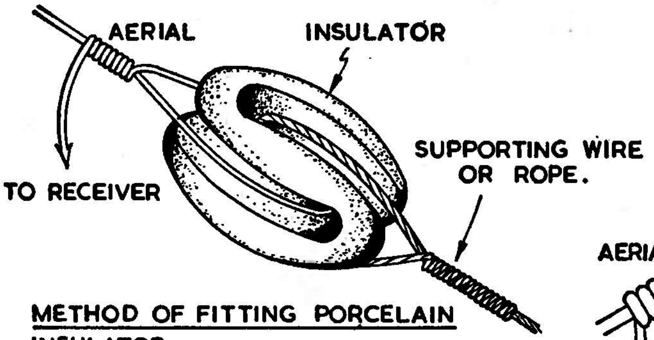
AERIAL
INSULATOR
TO RECEIVER
SUPPORTING WIRE OR ROPE.

METHOD OF FITTING PORCELAIN INSULATOR

FIG. 11

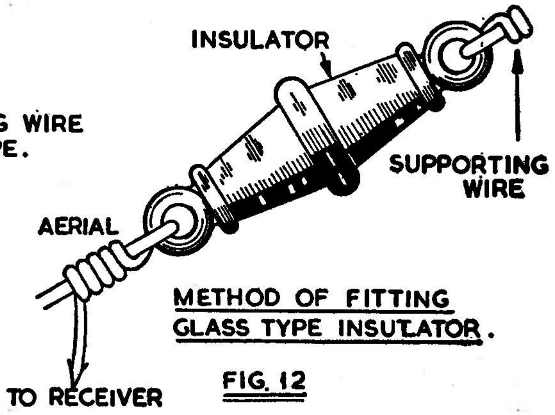
INSULATOR
SUPPORTING WIRE
AERIAL
TO RECEIVER

METHOD OF FITTING GLASS TYPE INSULATOR.

FIG. 12

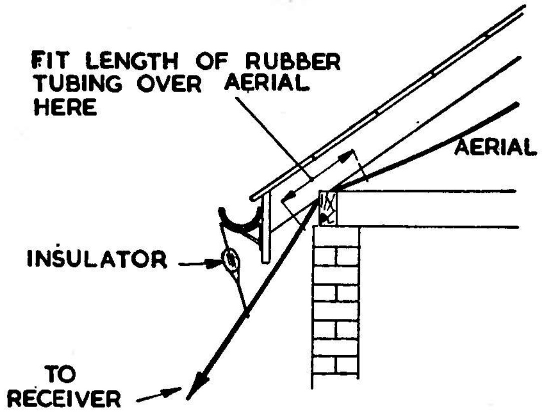
FIT LENGTH OF RUBBER TUBING OVER AERIAL HERE
AERIAL
INSULATOR
TO RECEIVER

LOFT AERIAL.

FIG. 13

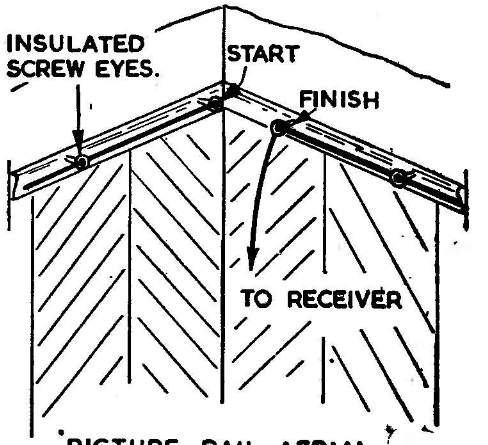
INSULATED SCREW EYES.
START
FINISH
TO RECEIVER

'PICTURE RAIL AERIAL.'

FIG. 14

should be quickly cleaned with an **old** file which should be kept specially for the purpose. A better idea is to obtain a small block of sal-ammoniac and rub the tip of the hot iron on it. This will clean the tip of the bit perfectly. Next take a length of cored solder such as Ersin Multicore, and melt a little on to the prepared tip faces, now smooth it evenly over the surface with a piece of old rag (be very careful not to burn your fingers) the iron is now " tinned " and ready for use. When heating the iron, be very careful not to let the bit overheat or get red hot, otherwise the tinned surface will be destroyed and the whole process will have to be gone through again.

An electric iron will not overheat, and the tinning will last much longer than with ordinary types, and since the heat does not deteriorate there is less likelihood of making faulty joints.

Having obtained a tinned iron, the process of soldering joints is remarkably simple and anyone with a little patience can acquire the art in a very short time.

To solder two copper wires together: clean the wire with emery cloth, apply a prepared iron and some cored solder to it, the solder will flow evenly over the wire thus tinning it. Repeat the process with the remaining wire and then twist the two together. Now apply the iron to the joint from the underside and the cored solder to the joint on top. Solder will flow evenly over the joint. Remove both iron and solder and allow to set. The solder will harden or set in a few seconds but during this period the joint must not be touched or moved, as otherwise the joint will be " dry " and quite useless mechanically or electrically. On radio components, tags for soldering are already tinned though if they are old, or discoloured it is best to re-tin them. Normally, however, it is only necessary to twist the connecting wire to it and apply the iron and solder as explained.

Remember: **never** apply the solder to the iron and then the iron to the joint, always apply the iron **and** the solder to the joint. It is however a good thing to apply a little solder to the iron tip even when it is perfectly tinned just before making a new joint.

One final " DON'T," you will have noticed that the solder referred to is " cored," that is, it contains resin and other substances through its centre, plain solder as used by electricians and plumbers will not do, as it is the resin or flux

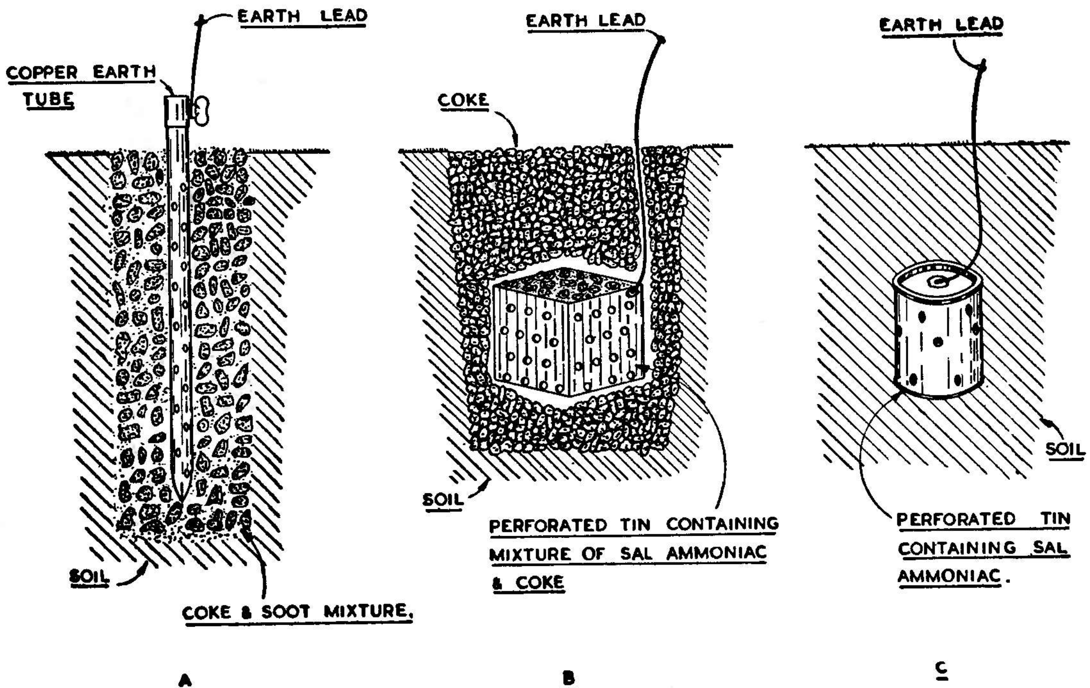
EARTH LEAD
COPPER EARTH TUBE
SOIL
COKE & SOOT MIXTURE.
A
EARTH LEAD
COKE
SOIL
PERFORATED TIN CONTAINING MIXTURE OF SAL AMMONIAC & COKE
B
EARTH LEAD
SOIL
PERFORATED TIN CONTAINING SAL AMMONIAC.
C

FIG. 15.

as it is called which makes the solder flow evenly and permits a good electrical joint.

Equally important do not attempt to use a separate flux or soldering fluid with plain solder or to " help " the cored solder; since these are almost certain to cause eventual corrosion and will destroy any components that have been contaminated.

Earths

The provision of a good earth is just as important as the rest of the installation.

Much disappointment would be avoided if this fact were not lost sight of, to avoid any slip up in this direction I am proposing to outline several well tried and efficient earth systems.

If it can be obtained, an earth rod, specially designed for the purpose provides the basis of a good earth connection. These were very popular some years ago when the majority of receivers were either crystal or battery operated, but in these days of modern mains driven sets they are not often used and consequently not always readily available.

Fig. 15 shows how it is used. First excavate a hole to a depth of three feet and fill with a mixture of soot and coke. Drive in the earth rod, which is a hollow copper tube, perforated. The lead to the receiver is connected at the top, and should be of covered 7/22 gauge, as used for the aerial system. Though not obvious it is important that this lead-in should be in covered wire, otherwise a number of indifferent earth contacts are likely to be made at various points along its length until it reaches the set. This is very undesirable and will spoil the efficiency of the system. It is essential to keep the soil surrounding the rod moist, which is one reason why the tube is hollow, and care must be taken to pour water into the tube at intervals.

A very efficient earth is the percolative type also popular at one time. Due to the chemicals used, it will extract moisture from its surroundings, thus maintaining a permanently moist earth.

It should be installed as in Fig. 15b.

The container is of copper or zinc, anything else will quickly rust away, again a good quantity of coke is used. Fill the container with a mixture of sal-ammoniac and coke, and then bury in coke as illustrated. This earth will not require further attention. It is possible to use powdered calcium chloride instead of sal-ammoniac but unless a chemist can be persuaded to make some up it is better to stick to sal-ammoniac.

If a zinc container is not available and you do not use or cannot easily get coke, Fig. 15c should be used. This makes a better earth than many so-called " earths " that I have come across. Obtain as large a tin as possible, make a number of holes as shown. Solder the lead in the bottom and fill it with sal-ammoniac. Replace the lid and bury in the ground. The tin will eventually rust away, but the replacement cost is negligible.

If it is quite impossible to make direct contact with the ground, a water pipe must be pressed into service. This should be a main pipe feeding straight to ground and not a hot water pipe or one fed from a tank. Scrape the pipe clean and twist the lead in tightly around it, a copper clip is even better. Do not attempt to solder on to the water pipe; since cold water is flowing through the pipe, it is extremely unlikely that your soldering iron will heat up the water supply sufficiently to allow a sound electrical joint, though you may spring a leak.

No attempt should be made to utilise gas pipes; the possibility of causing a fire is certainly very remote but, they make incredibly bad earth connections due to a number of joints made before true ground is reached. These joints are at best only semi-conductors, at least from an electrical standpoint.

Headphones

Since the late war there have been a large number of head-phone sets available on the surplus market. These may be roughly divided into two types, high-impedance and low-impedance. For the crystal sets detailed in this book high-impedance 'phones are required and the low-impedance pattern will not be suitable unless a matching transformer is used. As this is likely to cost more than the rest of the installation including the set, it will be as well to avoid them.

High-impedance types have an impedance of 2000Ω to 4000Ω whereas the low-impedance types are usually 600Ω.

The remaining consideration is weight; often cheap headphones are very heavy and uncomfortable to wear, every endeavour should be made to obtain 'phones as light in weight as possible.

Crystals

All the sets shown in this book have been designed to work with modern germanium crystals rather than the older galena crystal.

These germanium crystals require no adjustment, which, in itself, removes the main objection to this class of receiver. Suitable crystals are available from the following manufacturers and on the surplus market: Mullard, G.E.C., Brimar, Westinghouse and B.T.H. They are of robust construction, some in glass, others in ceramic or plastic, but must not be subjected to heavy knocks, otherwise the contact point may become dislodged.

From the information given, you should now be able to install an excellent aerial and earth system, and have some idea of how the crystal set works, so it is time to pass on to the actual receiver construction.

CONSTRUCTION 1

If you examine the following circuits you will find that each one is different. In most cases the difference lies in the coil design and/or the method by which the crystal and aerial is tapped into it. Each of these circuits has its own particular advantage to suit different conditions and the ideal circuit in some localities is not necessarily the best in others. It is not just a matter of a given circuit giving louder results than another, if it were there would be no point in showing more than one.

The main problem is to obtain adequate selectivity without reducing the volume level.

A receiver is said to be selective when it tunes sharply, a set with poor selectivity allows the stations to spread over the dial and when used near a transmitter will receive the local stations mixed together, which of course is useless.

Consider Fig. 16a, this is a very simple receiver, with no special attempt to provide any great amount of selectivity. In areas where signal

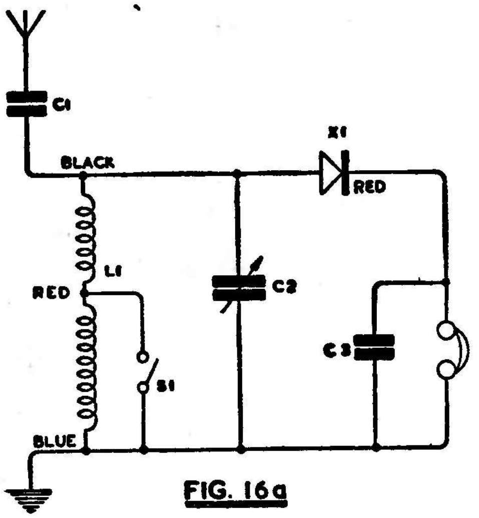

FIG. 16a

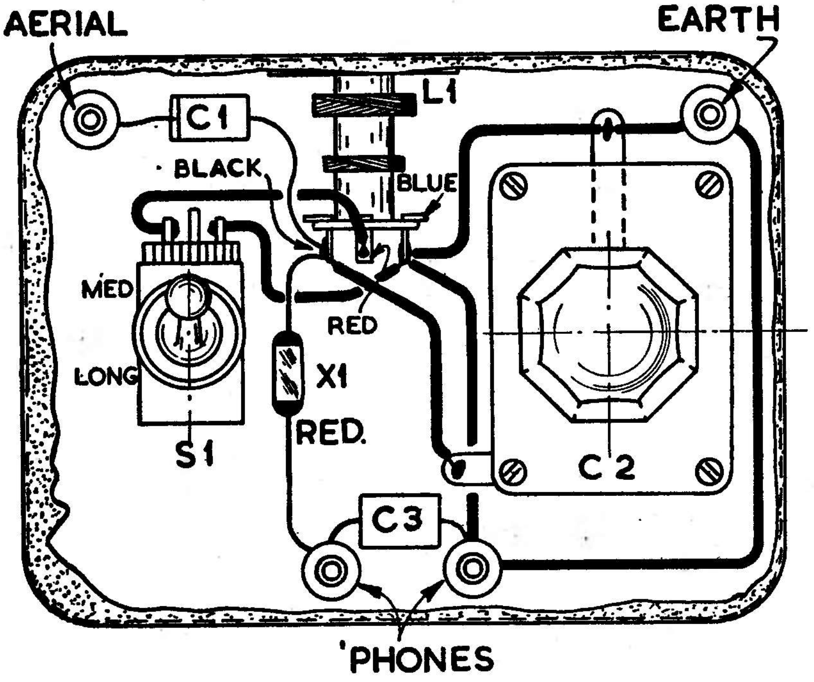

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG 16 b

strength is not high, or a short aerial is used, it will probably be ideal.

It would have been quite easy to increase the selectivity by providing a tap on the coil for the aerial as in Fig. 17a, but unfortunately, as the selectivity is increased overall volume is likely to decrease, so that unless you live close enough to the transmitter to have a large signal available and therefore need the selectivity, the circuit of Fig. 16a will be quite satisfactory.

Capacitor C1 is to prevent the aerial damping the circuit too heavily because this would flatten the tuning unnecessarily, however with small aerials it may be better to take the aerial direct to the black tag on L1. The signals are selected or tuned by L1 and C2, X1 is the crystal and C3 the capacitor across the 'phones to prevent unwanted carrier or R.F. reaching the 'phones.

A practical diagram (Fig. 16b) is provided showing the layout and all the wiring. You will require nuts and bolts to fix the coil, about ½", 4BA size will do, the other parts have locking nuts provided. An old 2-oz. tobacco tin makes a very good container and keeps the size down. Note that the metal box is connected to earth. With the exception of AC/DC receivers the metal work on any receiver or amplifier is connected to earth.

When S1 is open as in the diagram, the set will tune in long-wave stations, but when closed the medium wave-band will be received.

Components List, Fig. 16a

C1	100pF mica capacitor.
C2	500pF tuning capacitor, solid dielectric.
C3	1000pF mica capacitor.
X1	Germanium crystal.
L1	Crystal Set Coil. R.E.P.
S1	Single Pole toggle switch.
4	Insulated wander-plug sockets and plugs.
1	2-oz. tobacco tin (or similar container).

Make sure that the wander-plug sockets are of the **insulated** type, otherwise the metal case will join all the sockets together electrically.

Try to follow the theoretical diagram when wiring, a little practice will soon enable you to wire up a set without a practical diagram, which is a great advantage because often only the theoretical diagram is given when circuits are detailed in the technical press.

Fig 28 on page 34 gives a list of symbols used on the theoretical diagrams, so that you can readily identify the components.

2

Fig. 17a circuit is similar in many respects to that of Fig. 16a. The difference is purely one of selectivity, tuning will certainly be sharper, and even with comparatively inefficient aerials the design will put up a very good performance. There is no reason why the construction should not follow the same lines as the previous receiver but, by enlarging the set a little and using an air spaced tuning capacitor, the efficiency is improved. Observe that the colour coding on the coil, an R.E.P. Dual Range (Blue Box) is different from that of the Crystal set coil used on the previous design, and make sure it is correctly wired in.

C1 is a mica compression capacitor, often referred to as a "trimmer" and because it is adjustable, permits the set to be matched to aerials of varying lengths. Construction is carried out on a square panel of bakelite, perspex or wood, as shown in Fig. 17b. Perspex being clear like glass presents a most attractive finish, provided the set is neatly wired. It must be drilled slowly, however, otherwise the generated heat of the drill will make the hole wander, and spoil the panel.

After the set has been wired and tested a small wooden container can be made to house the completed set.

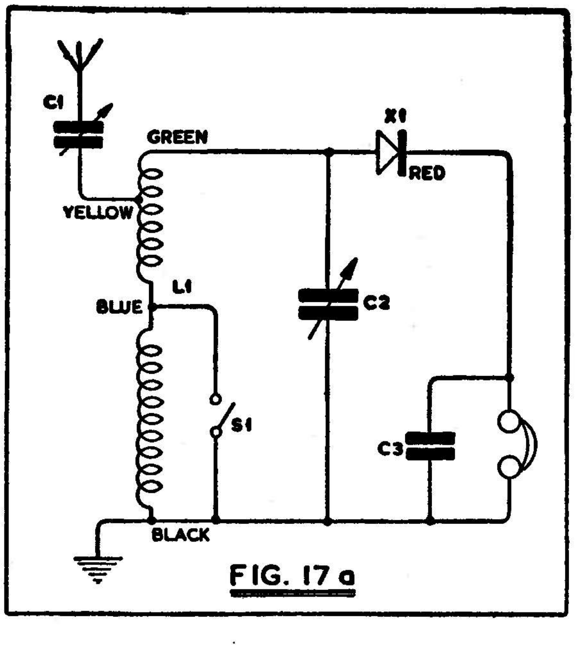

FIG. 17a

When carrying out reception tests, a little experiment is well worth while, to get the best out of the set. Try removing the germanium crystal from the green tag on L1 and connecting it to yellow, at the same time removing the lead from C1 to yellow and connecting it to green. Once the best arrangement has been found the wiring

can be left permanently in that position. Note that two tags on the coil are unused, this is intentional because on this circuit the extra winding connected to these tags is not required. With the switch S1 closed, the coil will cover the medium wave-band, and when open, long wave stations can be received. It is customary to abbreviate the expression medium wave-band to M.W. and long wave-band to L.W.

Components List, Fig. 17a

C1	100pF mica trimmer capacitor.
C2	500pF variable capacitor (air spaced).
C3	1000pF mica capacitor.
L1	Dual Range Coil R.E.P. (Blue Box).
S1	Single Pole toggle switch.
X1	Germanium Crystal.
4	Terminals or wander-plugs and sockets.
	Perspex, bakelite or wood for mounting panel.

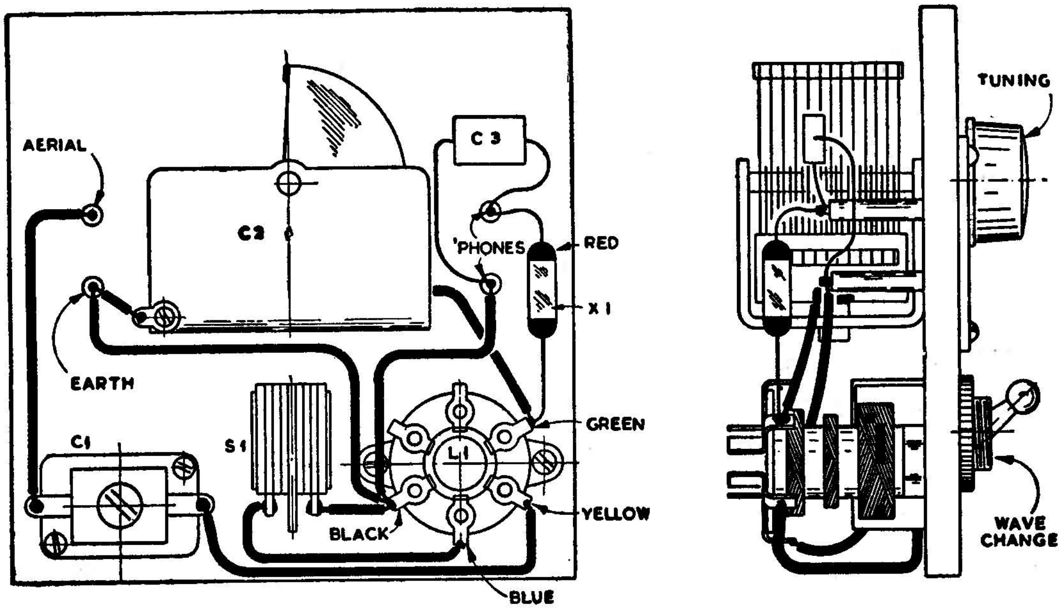

FIG. 17b

3

The design shown in Fig. 18a and 18b is more flexible than either of the preceding sets. That is, it can be varied to suit local conditions to a greater extent than the first two sets. This is made possible by the large number of taps provided on the coil.

You will notice that in this instance the coil is home-constructed on a cardboard former and is much larger than commercially produced coils. Coils wound in this manner are called **solenoids**, the turns are wound on side by side; in this case to a depth of some three inches. The diameter of the coil is made purposely large so as to obtain high efficiency. Modern commercial coils are invariably wave-wound and quite often are litzendraht which is the German for litz wire. This wire is made up of a number of strands of fine copper wire, each strand is enamelled to insulate it from the others, the whole is then silk covered. Litzendraht is more efficient than solid copper wire, and by using this and wave-winding, manufacturers can produce an efficient coil which is also small. Unfortunately wave-wound coils cannot be produced without a complex winding machine and home-constructed coils must take the solenoid form. It would be possible to use litzendraht but it is not easy to obtain by the reel, it is most expensive, and is difficult to handle. This last point is because at termination points, each strand must be cleaned of its enamel before a joint is made, and if one strand is broken, its advantage over plain copper wire is lost. How-

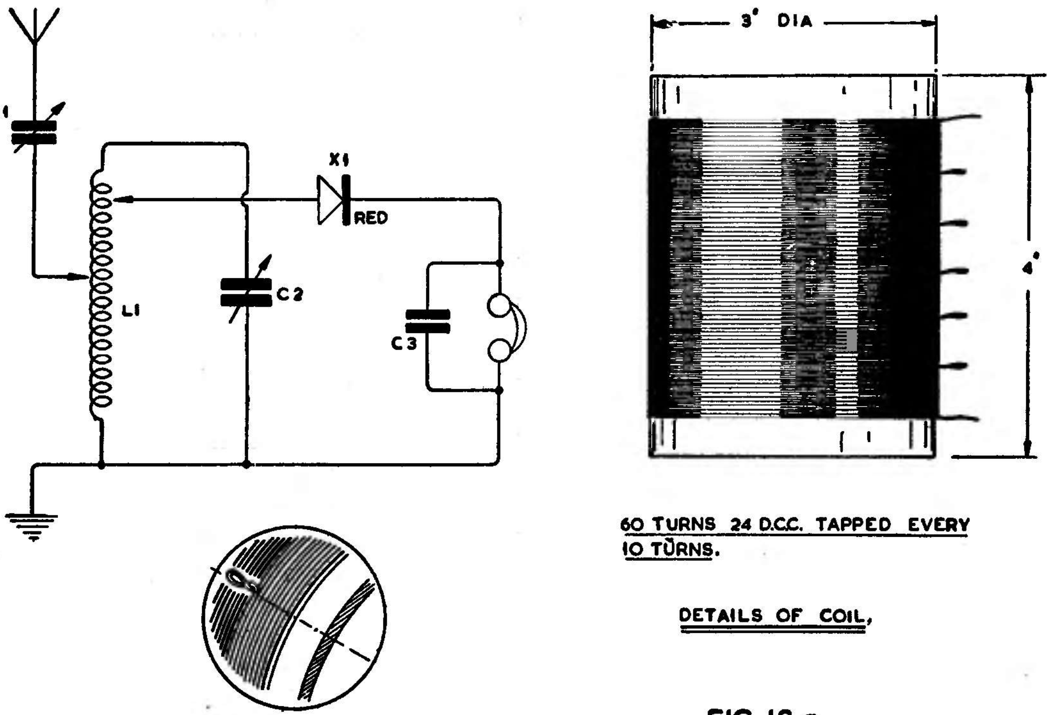

60 TURNS 24 D.C.C. TAPPED EVERY 10 TURNS.

DETAILS OF COIL.

INSET SHOWING METHOD OF TAPPING COIL.

FIG. 18 a

ever, by using a former of reasonably large dimensions an efficient coil can be produced cheaply using ordinary copper wire.

The wire must be firmly secured at the start and finish of the windings. Pierce three small holes about ¼" apart ½" from the end of the former. Pass the wire through the first from the outside, return it through the second and pass it back again through the third. Leave some 6" of wire at the end to make off the connection. It will now be possible to wind the turns on tightly without wire slipping.

Count on ten turns and make a loop 1" long. The method of preparing loops is shown in Fig. 18a, loops or taps should always be made in this manner, never by baring the wire and soldering a further length of wire to it. Carry on with the winding, making off the taps every ten turns as directed until the coil is complete.

Construction is carried out on a wooden base-board and front panel. Secure the coil to the base-board as shown on Fig. 18b, mount the remaining components and proceed with the wiring.

When testing out, it will be found that the further the aerial is tapped down the coil towards the earth end, the greater the selectivity. A position should be found which permits separation of local stations without excessive loss of sensitivity.

The crystal tap is adjusted for best results and different settings of C1 tried out, when choosing the best position for the aerial tap.

Components List, Fig. 18a

C1	450pF padder.
C2	500pF variable capacitor.
C3	1000pF mica capacitor.
L1	See text and Fig. 18a.
X1	Germanium Crystal.
2	Crocodile Clips (to connect leads to coil taps).
4	Wander-plugs and sockets.

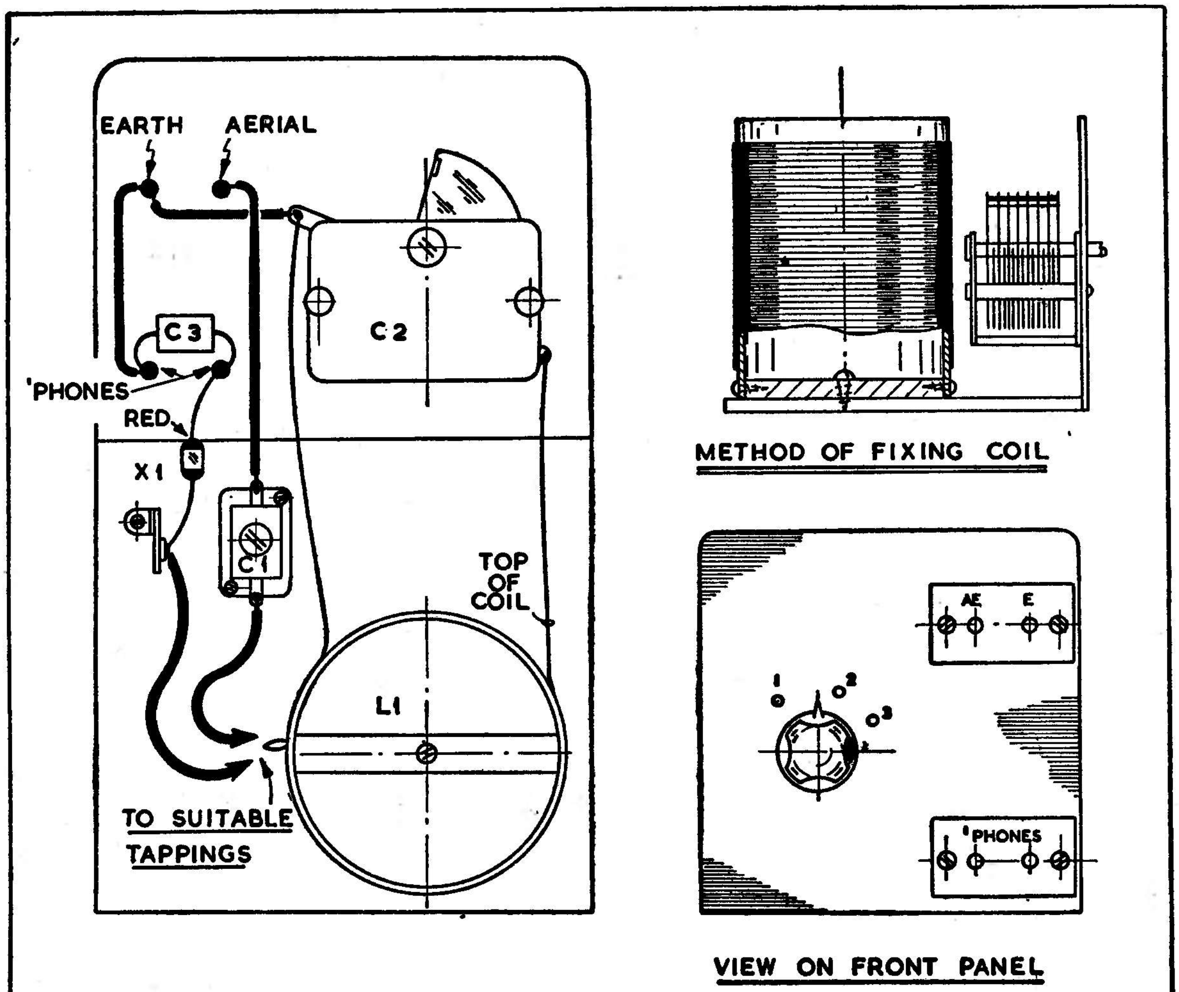

FIG. 18 b

4

A most unusual design is reproduced in Fig. 19a.

It first appeared in America about three years ago, and it certainly does offer some advantage over more conventional sets. The tuning circuits are duplicated and two crystals are used, so the circuit may be described as a full-wave receiver. Signals received will certainly be louder than with simple sets, but care must be taken with the coil winding, as with the other receivers using home-made coils, otherwise results will be disappointing. Be very careful to ensure that all three windings are in the same direction, this is very important.

The distance between each winding should be ¼". After the coil has been wound it is a good plan to warm it before a fire and paint the windings with "Durafix." Heating the coils makes the "Durafix" run freely. This substance sets quite hard and there will be no risk of the windings loosening. A small quantity of enamelled wire will cover the requirements of this coil, a 2 oz. reel will provide more than sufficient.

Before wiring in the germanium crystals, examine them carefully, note that one end is coloured red or in some cases marked with a positive sign thus +. You will notice, the sign is the same as the addition symbol used in arithmetic.

It is essential that both the red or positive ends are connected together, note that this is clearly marked on Fig. 19b. The receiver cannot work if

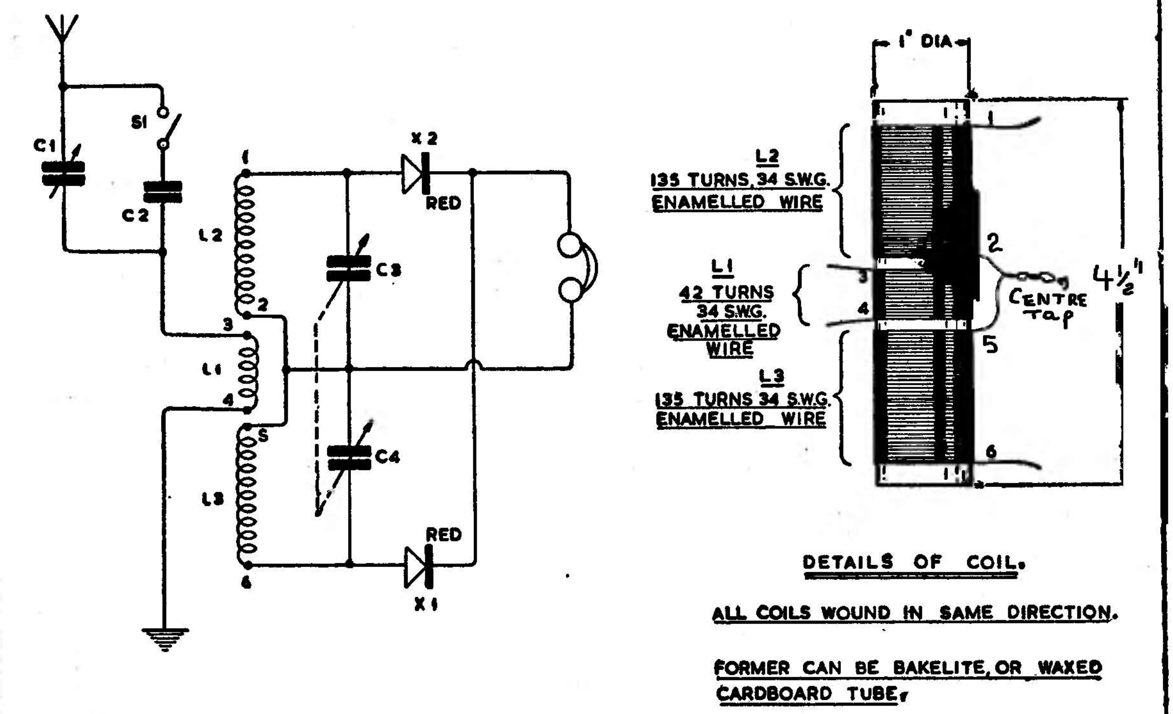

DETAILS OF COIL.

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE, OR WAXED CARDBOARD TUBE.

FIG. 19a

one of the crystals is connected the reverse way round.

A baseboard and panel form of assembly is used for this set which looks quite attractive if housed in a small polished cabinet. Dimensions are not at all critical but Fig. 19b gives a general idea of the layout that should be used. To avoid any unnecessary losses the coil should be mounted on perspex.

Obtain two strips of perspex 5½" long by ½" wide, drill fixing holes at both ends of each strip. It is best to clamp the two together when drilling these holes. Place one strip on the board in position, then put the coil and remaining strip over the first one and screw down. The coil will be clamped neatly and rigidly into position.

Twin socket bakelite strips are used to carry the aerial/earth and 'phone connections, if desired terminals mounted on pieces of bakelite could be used. These socket strips can be purchased from most shops selling components. If feet are not provided, drill two holes and screw into the edge of the base-board. Tag-strips are used to anchor some of the wires from the coil, and the germanium crystals. These also can be purchased for a few pence. At least two tags are necessary on one, and three on the other, though if the strips have more tags than required it is of no consequence. C3-C4, the ganged tuning capacitors have trimmers fitted, this should be stipulated when purchasing, and when first operating the set these small trimming capacitances are set about half-way. The idea of these trimmers is, that they compensate for any difference between the self-capacity of the tuning coils. With the switch S.1 open, that is in the OFF position, turn

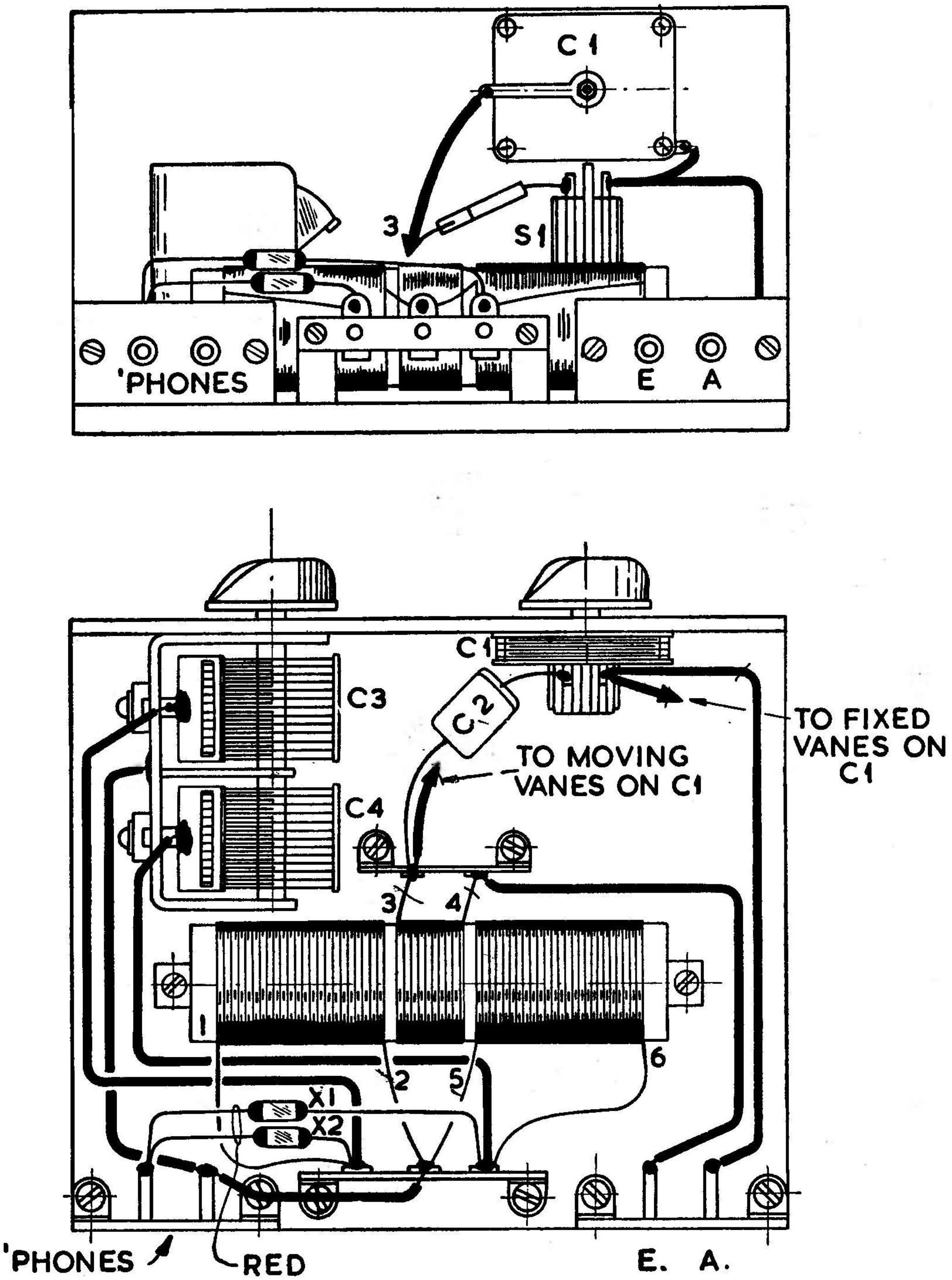
C1
3
S1
'PHONES
E
A
C1
C3
C2
TO FIXED VANES ON C1
TO MOVING VANES ON C1
C4
3
4
1
2
5
6
X1
X2
1
'PHONES
RED
E.
A.

FIG. 19 b

✓

in a station at the high-frequency end of the band, that is with the vanes of C3-C4 towards the disengaged position. Adjust C1 for maximum volume, without interference from other stations. Now adjust the trimmers for loudest headphone strength. It may be necessary to run over the adjustments several times for best results. Once they are correctly set, the trimmers require no further adjustment. The tuning coil covers the M.W. band, 1500 kc/s. to 600 kc/s., as the capacitor C3-C4 is advanced and the vanes start to mesh, the frequency to which the receiver tunes **decreases** so that with the vanes disengaged the set is tuned to 1500 kc/s. approximately and at full mesh, 600 kc/s. Switch S1 is used as follows, for stations **lower** in frequency than 850 kc/s it should be in the ON position with the contact closed, but for stations **higher** in frequency, it is in the OFF position. As an example both the London Home Service and London Light Programme transmitters on the M.W. band operate at a higher frequency than 850 kc/s. If any doubt exists as to the operating frequency of a given station, reference should be made to the Radio Times which quotes both wave-lengths and frequency. If you know the wave-length in metres, it is simple to find the frequency. Divide 300,000 by the wave-length in metres, the dividend equals the frequency in kilocycles, i.e., 300,000 ÷ 300 metres = 1,000 kc/s. In the same way, dividing 300,000 by the frequency in kilocycles will produce the wave-length in metres.

To use the receiver, tune in the signal by C3-C4, adjust C1 for maximum volume without allowing stations to overlap.

Components List, Fig. 19a

C1	300pF solid dielectric variable capacitor.
C2	1000pF mica capacitor.
C3-4	500pF twin gang variable capacitor.
L1-2-3	See text and Fig. 19a.
X1-2	Germanium Crystal.
2	Tag-strips.
2	Twin socket strips.
S1	Single Pole toggle switch.

5

Whenever crystal sets are discussed, constructors are apt to think in terms of medium-wave reception. When you consider that at least 98% of published circuits are designed for this band (sometimes with the long-wave band thrown in as an afterthought) it is understandable. However, if you can provide a good outdoor aerial (and an equally good earth) there is a lot of fun to be had listening to the short-wave bands. This receiver is designed specially for short-wave reception.

A metal chassis is used for construction, these can be obtained ready made in aluminium from most good supply houses. It can be quite small and on the original model a 6″ x 4″ was used. Low loss components are used so as to obtain greatest efficiency on the short wave bands. The coil is a commercial product which plugs into a 4-pin base. Best results were obtained on the 3mc/s to 7mc/s band though coils covering other bands are available.

Note that the tuning capacitor C1 is smaller than normally used and has a maximum capacitance of 140pF.

Coil type 706/R covers the 3mc/s to 7mc/s band, but the set is equally satisfactory on the M.W. band, and to cover this coil type 706P should be used. This coil has an iron dustcore which can be adjusted to make the coil cover the required band. The effect of the core is as if turns were being added or removed from an ordinary coil. When dealing with the short waves it is customary to refer to the frequencies in terms of mega-cycles (mc/s) rather than kilo-cycles. One mega-cycle is equivalent to 1000 kilo-cycles, that is 1,000,000 cycles.

Layout and wiring can be clearly followed from Fig. 20b and a theoretical diagram is shown in Fig. 20a. Care should be taken when arranging a mounting for the aerial terminal. The best plan is to cut a 1″ hole in the chassis, mount the aerial terminal on a piece of perspex and screw into the chassis. This avoids any unnecessary losses between aerial and chassis. The earth terminal is screwed directly into the chassis.

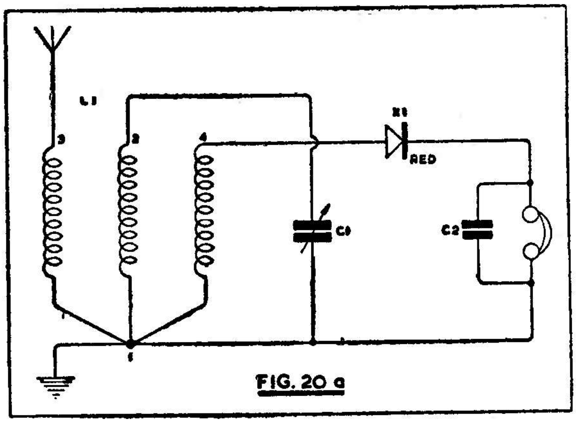

FIG. 20 a

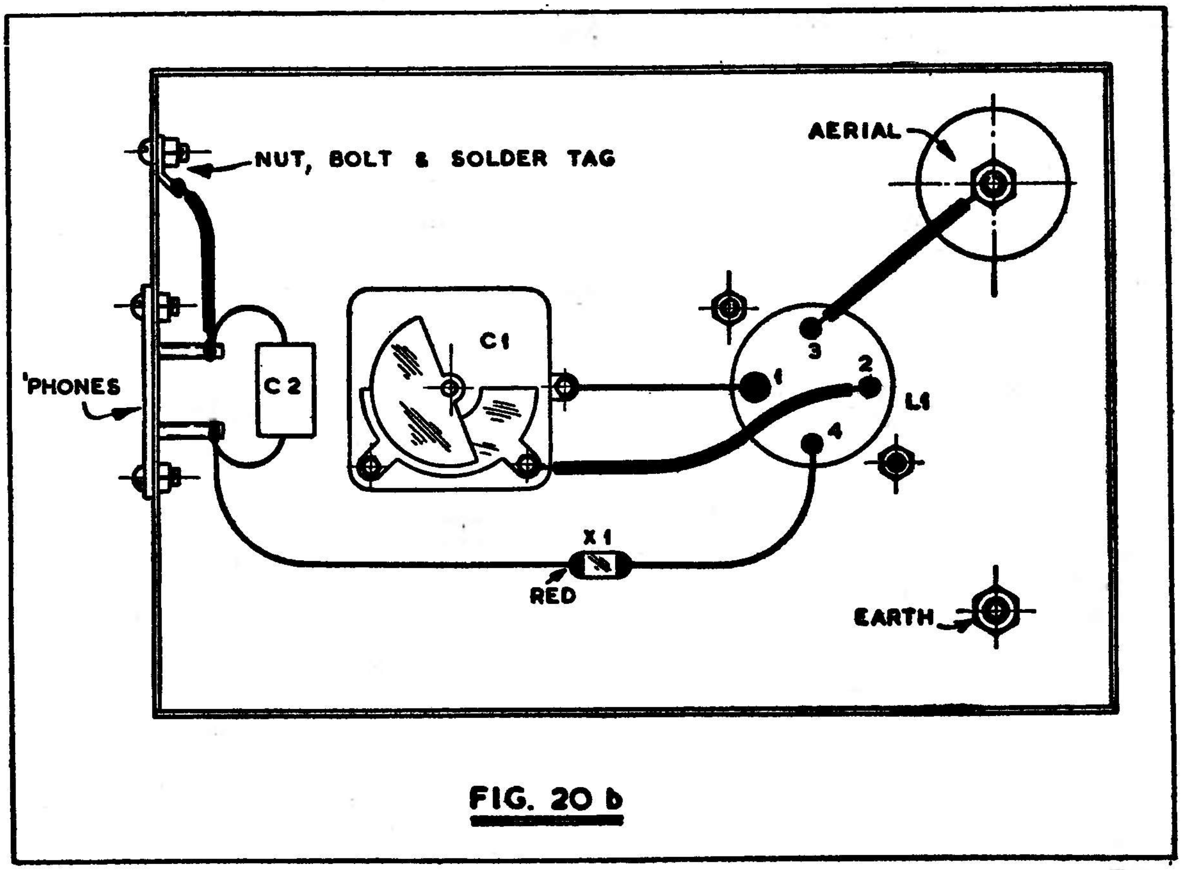

FIG. 20 b

Wiring as shown in Fig. 20a is likely to give the best results, but other arrangements are possible and to get the best out of the set they should be tried.

Refer to Fig. 20a. The aerial is connected to pin 3 on L1, C1 to pin 2 and X1 the crystal to pin 4. If you have a long aerial try this combination—Aerial to pin 4, C1 to pin 2, and the crystal to pin 3. Connections to pin 1 are not altered. In a few cases the following arrangement may prove best. Crystal and C1 to pin 2, aerial to pin 3, pin 4 left free and again pin 1 is unaltered.

Tuning on the short-wave bands is more critical than on the medium waves so tune very slowly over the band and remember signal strength is likely to vary from day to day.

Components List, Fig. 20a

C1	140pF variable capacitor Eddystone 586.
C2	1000pF mica capacitor.
1	Coil holder Eddystone 707.
L1	3m/c—7m/c coil Eddystone 706/R.
1	Engraved tuning dial.
1	6"x4"x2½" chassis.
2	Terminals (Aerial-Earth).
1	Twin socket strip ('phones).
X1	Germanium Crystal.

6

Fig. 21a-b is yet another design from America and appeared in Radio Craft some years ago, designed by Mr. W. J. Spain. The original used a silicon crystal which preceded the development of the germanium. Results are certainly very good though of course on the model built here, a germanium crystal was used. Selectivity can be adjusted to suit all conditions and even at short distances from local stations. Good volume can be obtained without the programmes overlapping. Home made coils are used and for best results a 75′ outdoor aerial is desirable.

Baseboard and panel construction is used, which can conveniently be housed in a small wooden cabinet.

Operation is a little more complicated than some of the other designs, but the results certainly merit the extra trouble taken. As with the other receivers much of the success obtainable is due to the coils and these must be carefully made otherwise results will be disappointing.

Layout and wiring can be followed quite easily from Fig. 21b and requires no special comment.

First make the coils. L1 is a tapped coil, the taps should be made in the same way as shown in Fig. 18b. It consists of 90 turns of 22 D.C.C. (double cotton covered) copper wire tapped at the following number of turns, 5, 10, 15, 25, 30, 40, 50, 60, 70 and 80. The former used has a diameter of 2″. L2 is also wound on a 2″ diameter former, this coil however is not tapped, it consists of 110 turns of 38 D.C.C. or enamelled copper wire. Great care must be taken when handling this wire, it is quite fine and in the hands of the inexperienced will easily break.

After wiring has been carried out the set can be put into operation.

Selectivity is controlled by S2, in position B tuning is broad but in position A it can be quite sharp.

When searching for a station switch S2 to B. Tune the station and switch S2 to A. The frequency range is controlled by the position of the tap on L1, a good plan is to start with the 5 turn tap on L1.

Components List, Fig. 21a

C1	500pF variable capacitor.
C2	500pF variable capacitor.
C3	1000pF Mica Capacitor.
X1	Germanium Crystal.
L1 & L2	See text.
S1	Single Pole toggle switch.
S2	Single Pole 2-way toggle switch.
2	Twin socket strips (Aerial/Earth and 'Phones).

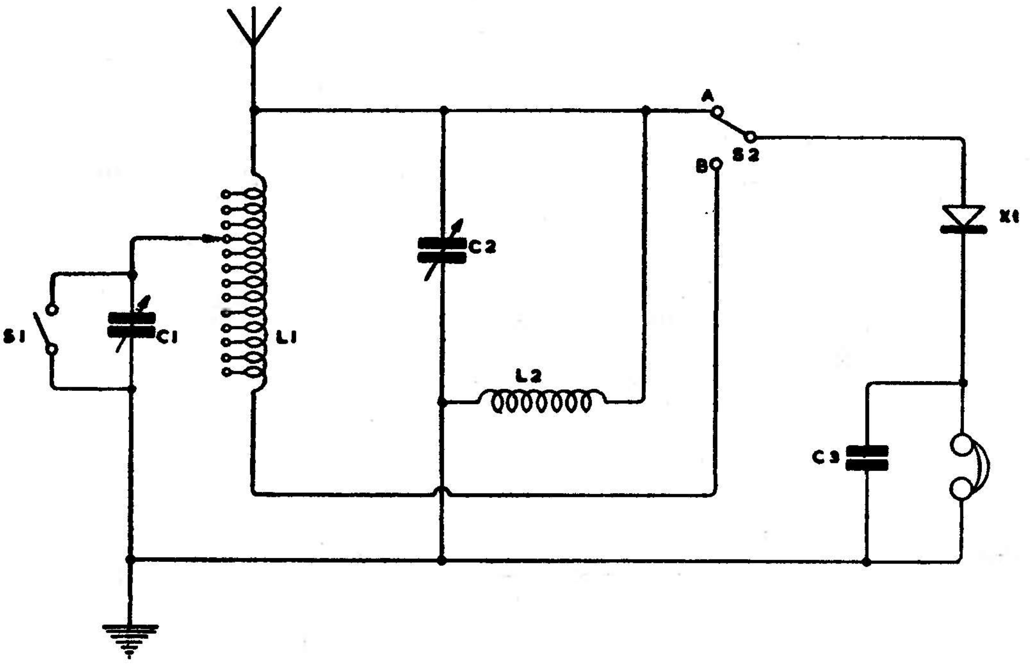

FIG. 21 a

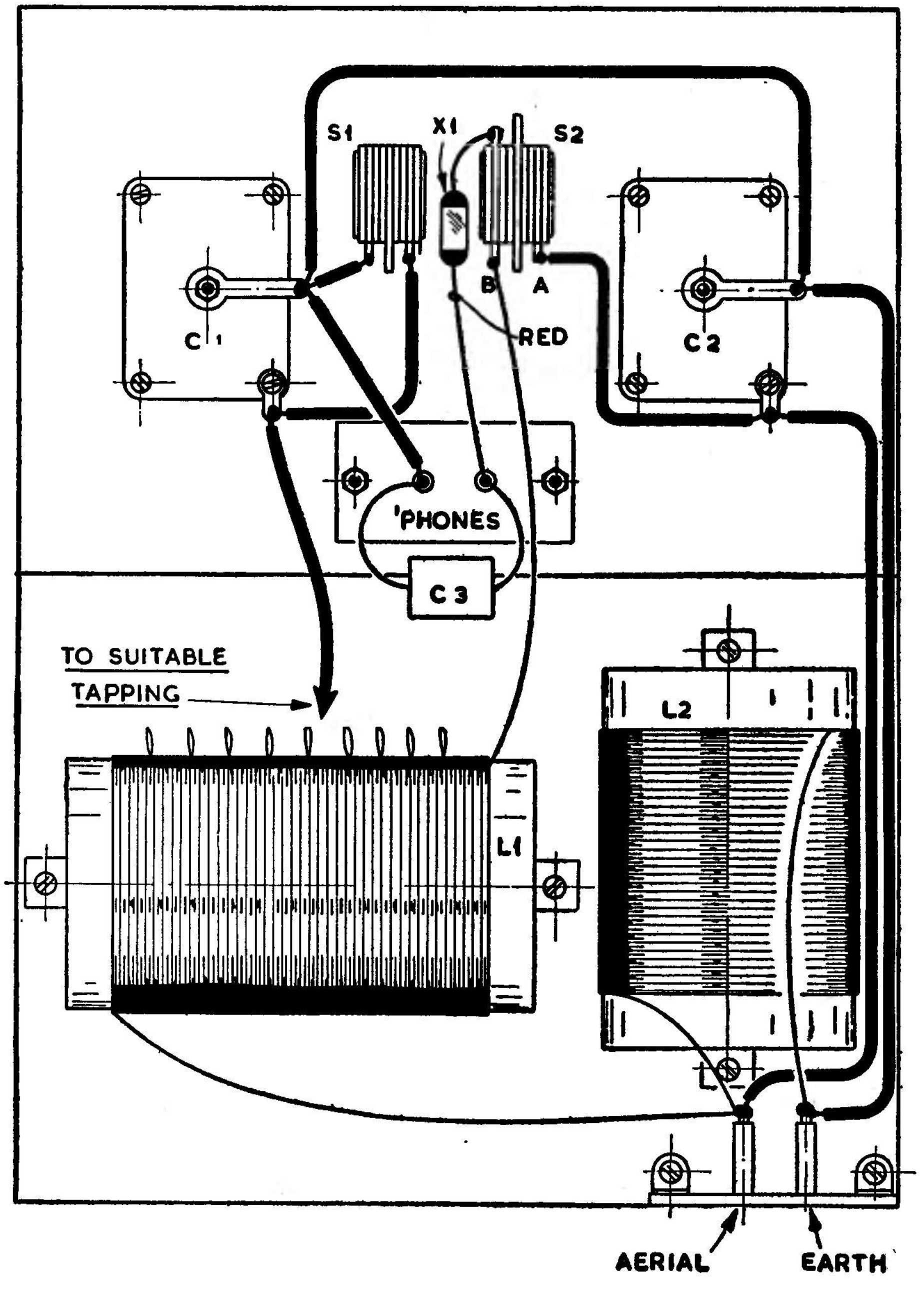
S1
X1
S2
B
A
RED
C 1
C 2
'PHONES
C 3
TO SUITABLE TAPPING
L2
L1
AERIAL
EARTH

FIG. 21 b

Keep S1 open and C1 for tuning, adjusting the tap on L1 for best results. Very powerful signals are best brought in with S1 closed. When using C1 and adjusting the taps on L1, switch S2 is kept in the A position.

It will be found that C2 acts as a fine control on C1. One division of the tuning dial on C1 is roughly equal to a movement of 20 divisions on C2.

Note that the coils are mounted at right angles to one another and that in this instance wood has been used as far as possible, an aluminium or other metal chassis is not suitable.

7

A very useful receiver for both medium and long-wave reception is possible with the circuit of Fig. 22a-b. A home-made coil is used. On the original a 1¼" diameter former was used but in fact this is not very critical and a 1½" former could be used. An interesting point is the construction of the long-wave coil section. Solenoid coils as used for the medium wave-band are not efficient if the length is too great relative to the diameter. Now as the long-wave section has 300 turns, if wound as a solenoid the winding length would be very great and a lot of efficiency would be lost. To overcome the difficulty the coil is pile wound. If the whole coil were to be wound in one pile it would still be inefficient because of the capacity formed in the winding, but by dividing it into five sections a coil of reasonable efficiency results. The same remarks concerning winding apply to this coil as to all the others in

Components List, Fig. 22a

C1 300pF Mica Capacitor.
C2 500pF Variable Capacitor.
C3 1000 Mica Capacitor.
X1 Germanium Crystal.
L1-2 See Text.
S1 Single Pole Toggle Switch.

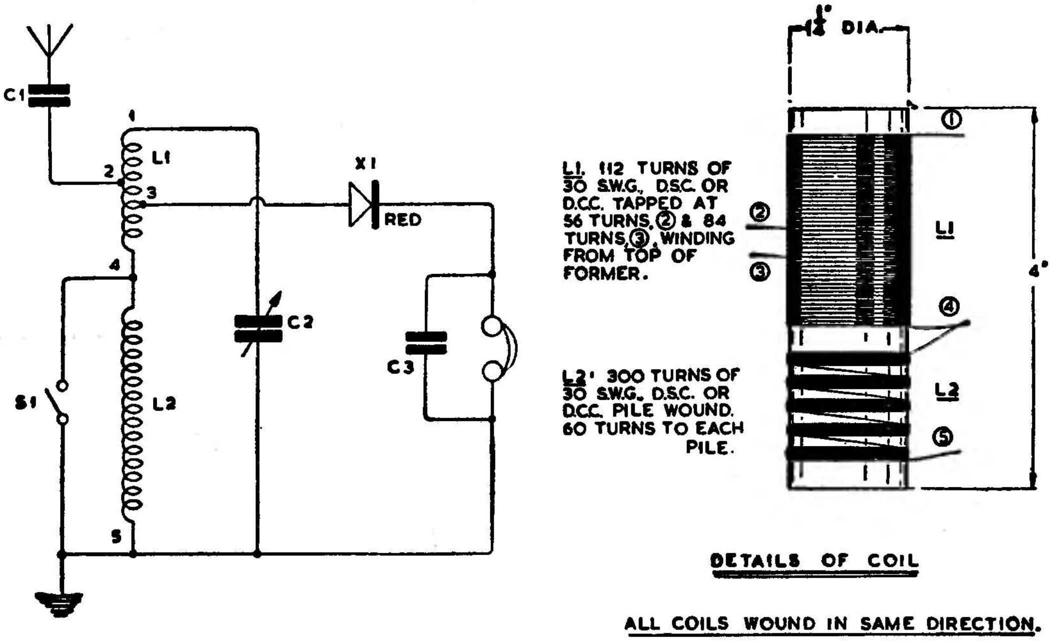
C1
1
2
3
L1
4
S1
L2
5
X1
RED
C2
C3
1¼" DIA.
L1. 112 TURNS OF 30 S.W.G., D.S.C. OR D.C.C. TAPPED AT 56 TURNS ② & 84 TURNS ③ , WINDING FROM TOP OF FORMER.
L2' 300 TURNS OF 30 S.W.G., D.S.C. OR D.C.C. PILE WOUND. 60 TURNS TO EACH PILE.
①
②
③
④
⑤
L1
L2
4"
DETAILS OF COIL

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE OR WAXED CARDBOARD TUBE.

FIG. 22a

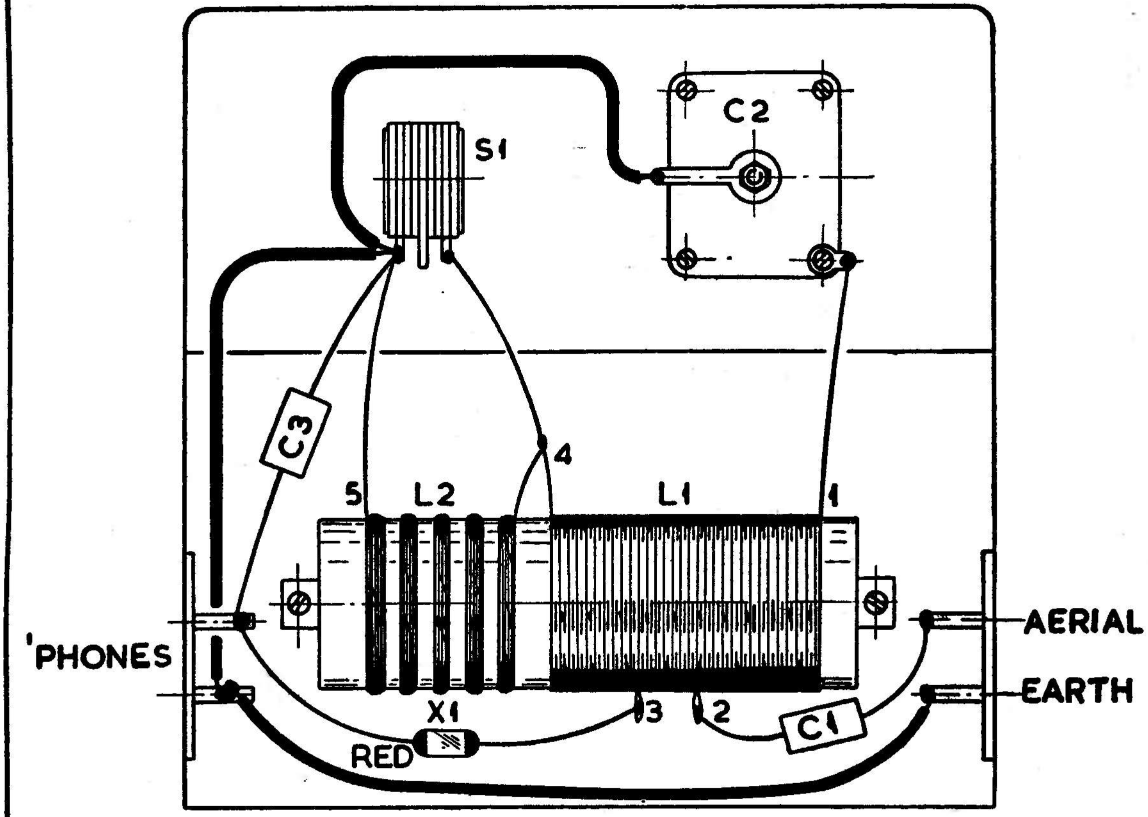

FIG. 22 b

this book, and readers should not, by this time, encounter any difficulties.

The design provides good selectivity with good sensitivity, and provided the set is not required to operate within a few miles of a powerful transmitter, no trouble will be encountered from interference between adjacent stations.

A wooden base-board and front panel is best for this design, and the wiring and layout can be clearly followed from Fig. 22b. Switch S1 is open for long-wave reception and closed for the medium-wave band.

Coil mounting can be on the same principle as for the full-wave design Fig. 19a-b.

8

If you look at all the other circuits shown in this book you will find that in each case the tuning capacitor and tuning coil are in parallel, this is by no means essential and the circuit of Fig. 23a shows a series arrangement that is the tuning capacitor is between the coil and earth instead of being wired up across it. This design also appeared in **Radio Craft** a few years ago.

An interesting point is that the coil is a conversion adapted from another type of component. To make this coil you must obtain an old I.F. transformer as used in superheterodyne receivers.

You will require one designed to work anywhere between 450 and 470 kc/s. Some of the

very early ones worked at 110 kcs/ but these are unsuitable.

The I.F. transformer must also be of the type tuned by an iron dust core at each end.

Dismantle the transformer, remove the parallel capacitors and mounting wires so that you are left with two coil bobbins and their former. Be very careful when un-soldering the wires from the supports because the coils are usually wound in litzendraht and must remain intact. Remove 25% (approximately) of the turns from one coil which we will call the secondary, and completely remove the iron dust slug from this coil, now cut the former in half, and mount as in Fig. 23a. Warm the coils with heat from your soldering iron or from a fire and gently slide the coils into position at the ends of the former as in Fig. 23a. The wax will set again holding the coils in their new position.

Note that the wires from the coils are connected to tags on the mountings. These are pieces of bakelite, and usually the original end pieces used on the I.F. transformer can be used.

A word about soldering litzendraht. Do not attempt to use emery paper. Prepare the end of wire by removing the outer silk covering and dip the end in methylated spirits. Ignite with a match and after about five seconds wipe with a cloth held between finger and thumb. This will quench the flame and will remove all traces of enamel from the wire. It is a good plan to experiment with the odd length of wire removed from the secondary coil. If the coils prove to be constructed of ordinary copper wire, it can of course be prepared with fine emery cloth in the normal manner.

Components List, Fig. 23a

C1	500pF variable capacitor.
X1	Germanium Crystal.
L1	See Text.
4	Plugs and sockets or terminals.

Wiring is straight forward as can be seen from Fig. 23b and requires no explanation.

Operation is as follows, selectivity is controlled by varying the distance between coils with the wing nut, and to an extent by the position of the primary iron dust core. Remember that the aerial and earth installations are as important with this set as with any other crystal receiver. No parallel capacitance appears necessary across the 'phones, the set works quite happily relying on the self-capacitance of the head-phones.

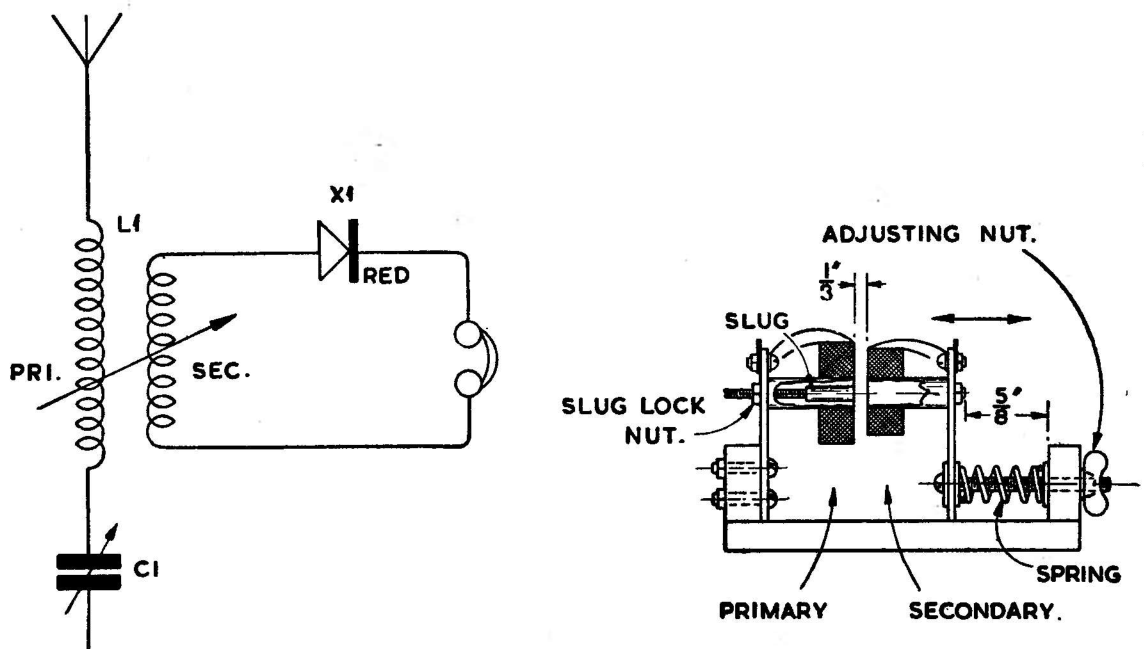

FIG. 23 a

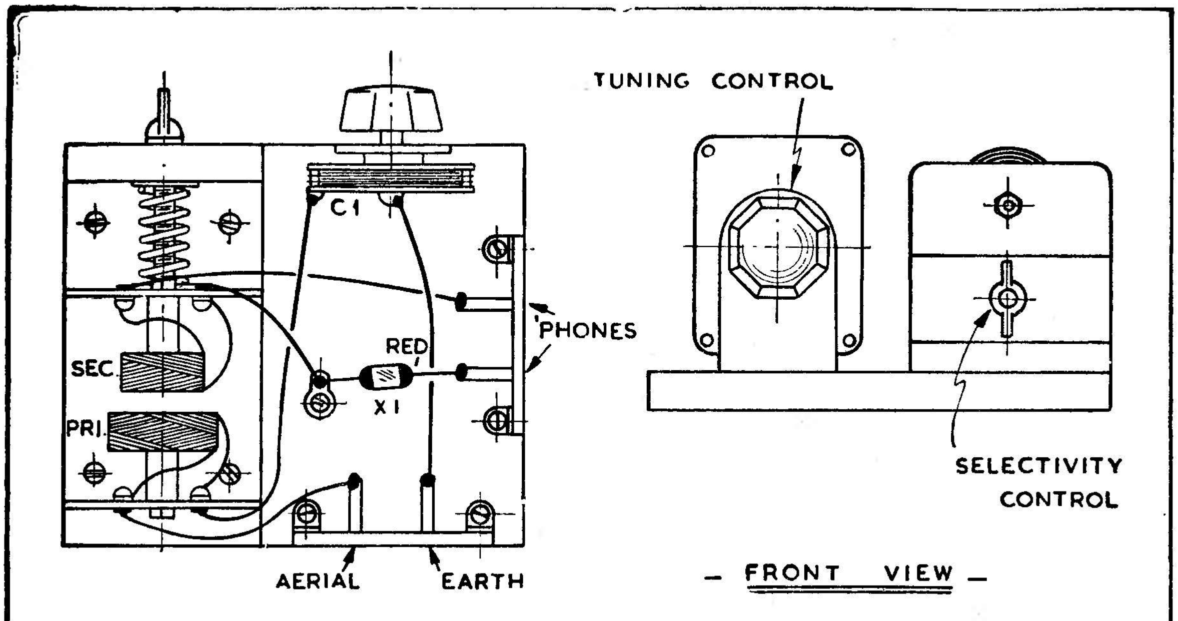
TUNING CONTROL
C1
PHONES
RED
SEC.
PRI.
X1
AERIAL
EARTH
SELECTIVITY CONTROL
— FRONT VIEW —

FIG. 23 b

9

So as to give a truly varied selection of circuits, some of the sets have been designed around commercial coils whilst others use coils which must be home-constructed.

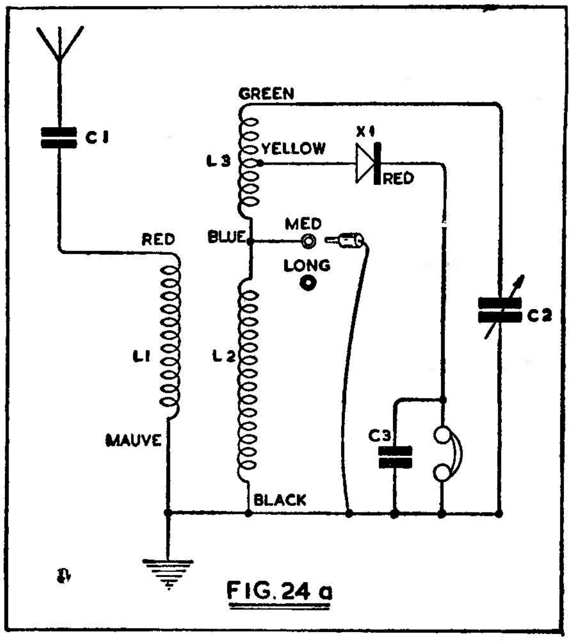
C1
GREEN
X1
L3
YELLOW
RED
MED
RED
BLUE
LONG
C2
L1
L2
MAUVE
C3
BLACK
FIG. 24 a

Most enthusiasts like to build their own coils but when it is a question of something really small combined with efficiency a commercial coil is the obvious choice.

Fig. 24a-b uses a tobacco tin or a similar container just as the design shown in Fig. 16a.

This receiver however is a little more selective and is more suitable than the earlier design, if you are near a transmitter or have a very long aerial. Take great care to connect the coil according to the colours shown on the diagrams. Sockets are used for wave-changing, this cheapens the cost a little and they are just as efficient.

Components List, Fig. 24a

C1	100pF Mica capacitor.
C2	500pF variable capacitor solid dielectric.
C3	1000pF Mica capacitor.
X1	Germanium crystal.
L1-2	R.E.P. dual range coil. Blue box.
6	Insulated wander-plug sockets.
5	Wander Plugs.
2oz.	Tobacco tin or similar container.

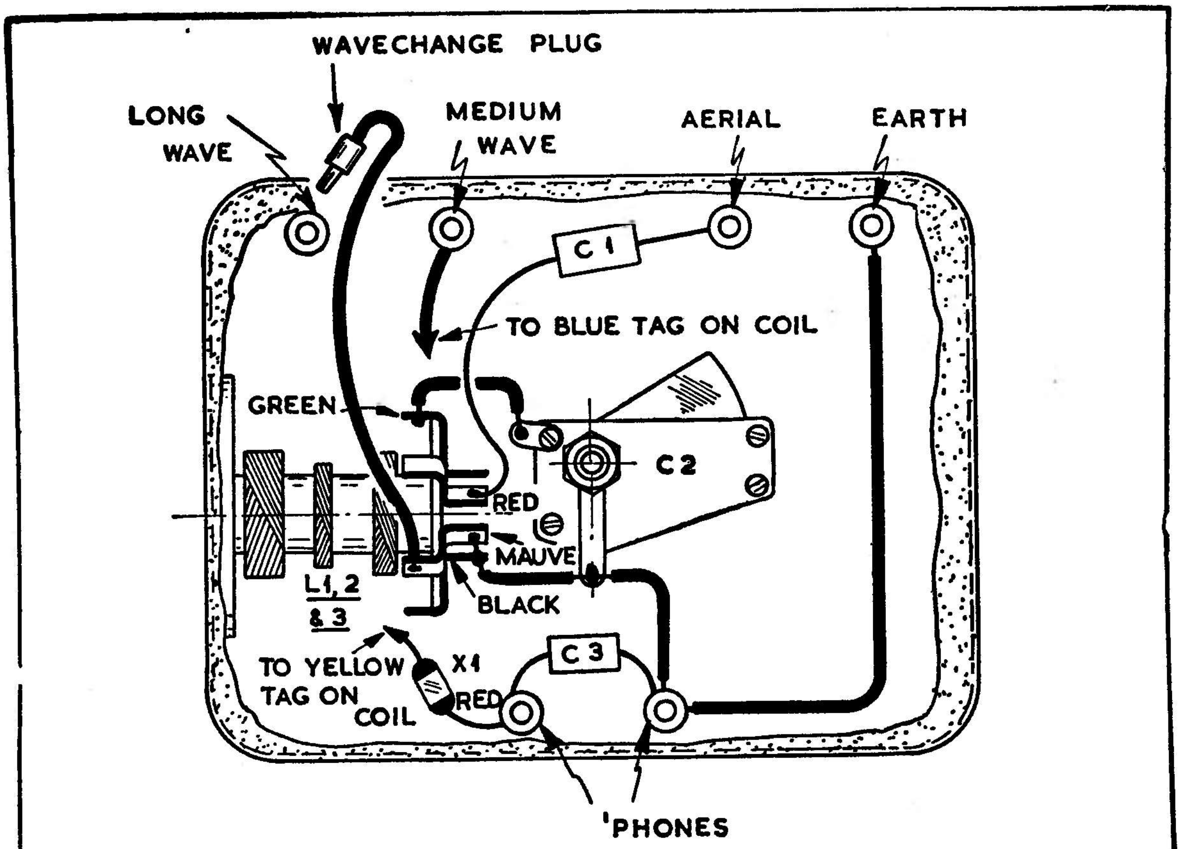

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG. 24 b

10

In certain cases, signals from the local transmitter are too powerful to permit a crystal set with a normal coil to be of much use. I myself live in Hertfordshire a few miles from the London transmitters, and find a more elaborate circuit is necessary. Now in radio circles it is a well-known fact that the greater the number of tuned circuits, the greater the overall selectivity. Look at Fig. 25a, you will see that the usual L and C tuning arrangements have been duplicated. This type of circuit is known as a Band-Pass Filter. There are many kinds of such filters, this particular type is an inductively-coupled filter, note that energy from the first half (L2) is conveyed to L3 by way of two small inductors L1 and L4.

Tuning is quite sharp and the "feel" of the set is quite different from that of single coil designs. A small chassis is ideal for construction. 6"x4"x2½" was used on the original, though it could have been smaller.

Fig. 25b gives the layout and wiring.

Take particular notice that L1-2-3 is mounted horizontally and at right angles to L4-5-6, this is to prevent the coupling from being too "tight" which would ruin the idea of the filter. A good aerial and earth is of course necessary. It may be found on test that the selectivity is too great

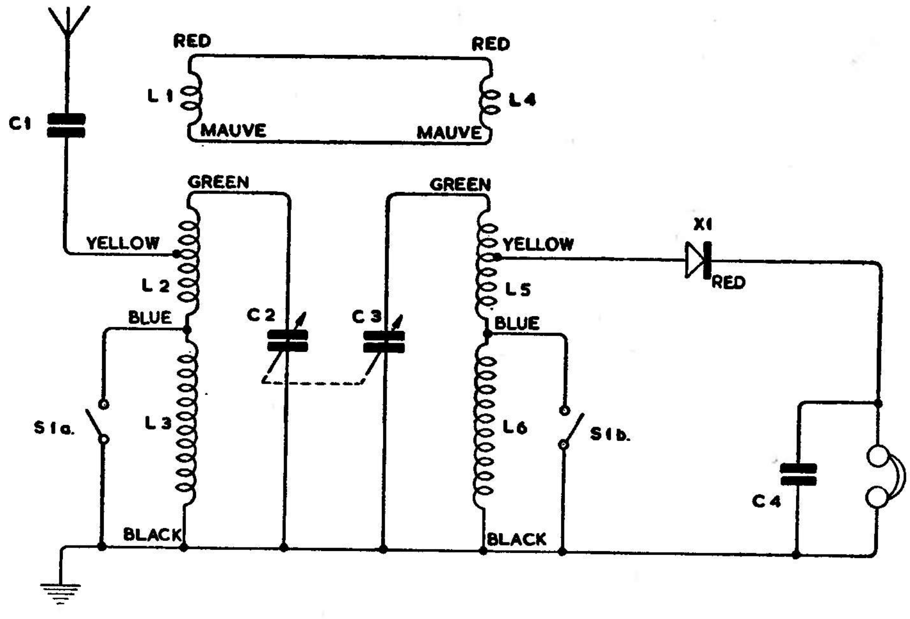

FIG. 25 a.

and that volume has suffered. This can be ingeniously overcome by connecting a very small capacitance between the two Green tags on the coils. The simplest way is to connect it across the fixed vane connections on C2-C3. Values for this extra capacitance are a matter for experiment usually a value between 10pF and 47pF is sufficient. A small 50pF trimmer could be used and adjusted to suit. The ideal setting is to enable powerful adjacent stations to be received just short of overlap so that as much volume as possible is obtained.

There is one point to be observed with the ganged capacitor. It should be of the type fitted with trimmers though of course, there is no reason why you should not fit them yourself. These are to allow for differences between coils to be balanced out, and consist of a small variable capacitance across each main section.

When setting up the receiver, first set each trimmer at half way and tune in a station near the high-frequency end of the band (vanes nearly out) adjust the trimmers for maximum volume. If extra coupling capacity is added as described, try adjusting them further.

Once properly set up, no further adjustment is necessary.

Components List, Fig. 25a

C1 100pF Mica capacitor.
C2-3 2 x 500pF variable capacitor (see text).
C4 1000pF Mica capacitor.
L1-2-3 Dual Range Coil R.E.P. (Blue Box).
L4-5-6 Dual Range Coil R.E.P. (Blue Box).
X1 Germanium Crystal.
S1a-b 2 pole single throw toggle switch.
4 Terminals or plugs and sockets.

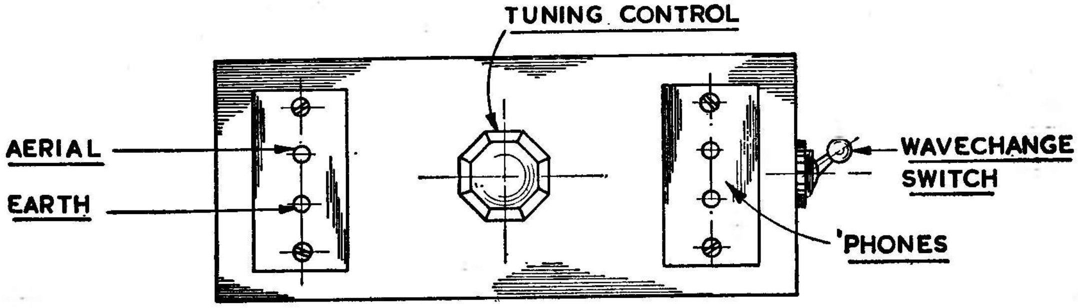
TUNING CONTROL
AERIAL
EARTH
WAVECHANGE SWITCH
'PHONES

FRONT OF COMPLETED RECEIVER.

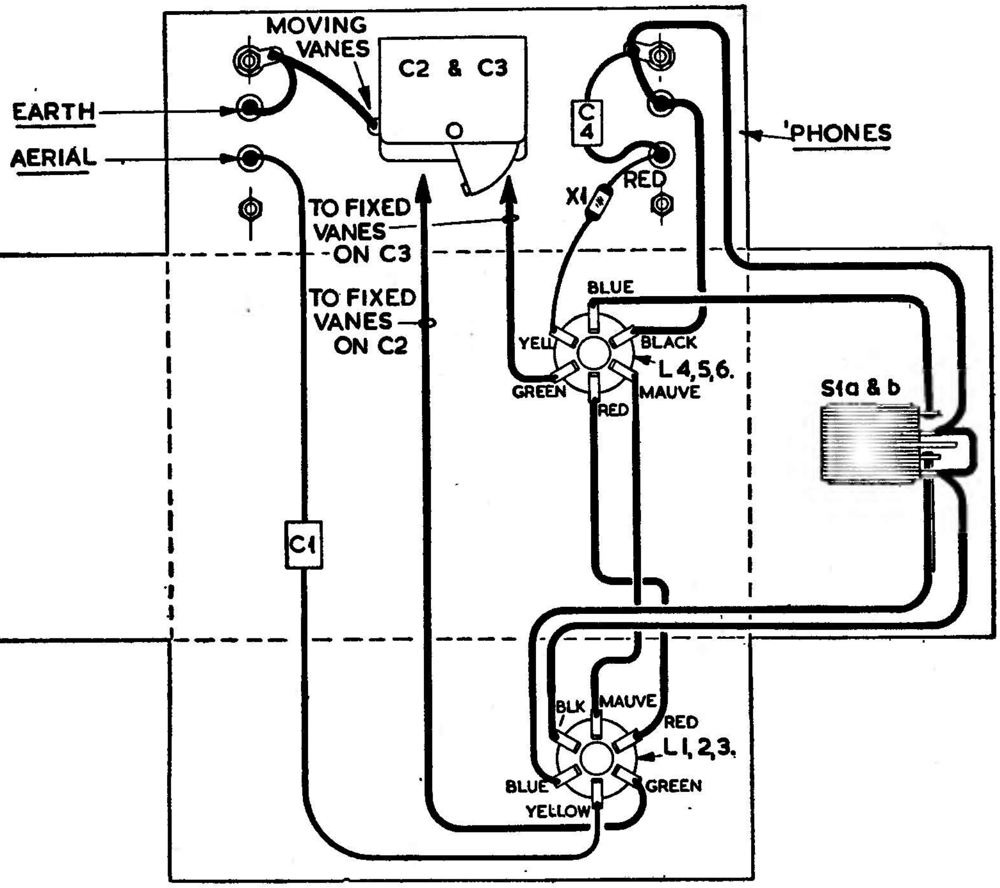
MOVING VANES
C2 & C3
EARTH
AERIAL
C4
'PHONES
RED
X1
TO FIXED VANES ON C3
TO FIXED VANES ON C2
BLUE
YELL
BLACK
L 4,5,6.
GREEN
RED
MAUVE
S1a & b
C1
BLK
MAUVE
RED
L1,2,3.
BLUE
YELLOW
GREEN

EXPLODED UNDER-CHASSIS VIEW

FIG. 25 b

11

Another band-pass circuit is shown in Fig. 26a. This time home-made coils are used. This filter is not inductively coupled but entirely capacitive, relying on C3. As in the previous circuit it must be adjusted to give just sufficient " broadness " to the tuning to provide adequate volume without station overlap. To avoid unwanted coupling a screen is mounted between the coils, this is necessary because due to the size of the coils they would be bound to couple quite tightly without it. Such precautions were not necessary with the circuit of Fig. 25a, as the commercial coils are much smaller physically and the risk of unwanted coupling reduced.

The coils for this receiver are for medium wave reception only, this keeps the size down. Coil winding procedure has already been fully explained, and covers all the requirements of these coils. Base-board and panel construction is used, and for more accurate tuning since the coils are home-made, separate tuning capacitors are recommended. Fig. 26b gives the layout and practical wiring.

Components List, Fig. 26a

C1	100pF Mica capacitor.
C2	500pF variable capacitor.
C3	50pF Mica trimmer.
C4	500pF variable capacitor.
C5	1000pF mica capacitor.
L1-2	See text.
X1	Germanium Crystal.
4	Terminals or plugs and sockets.

For those who do not wish to bother with these a ganged capacitor can be used, but tuning may not be as good as with separate ones. The screen mounted between the coils is of aluminium or copper and should be about 1″ higher than the coils and the same length as the baseboard depth.

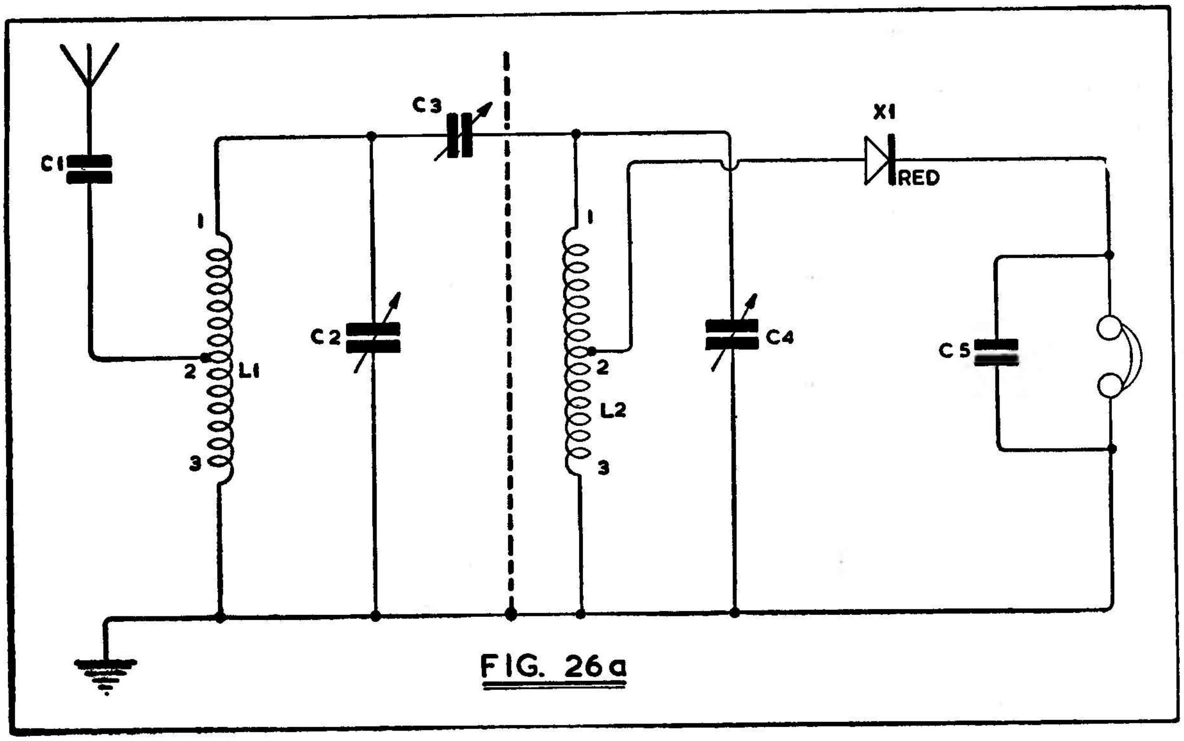

FIG. 26a

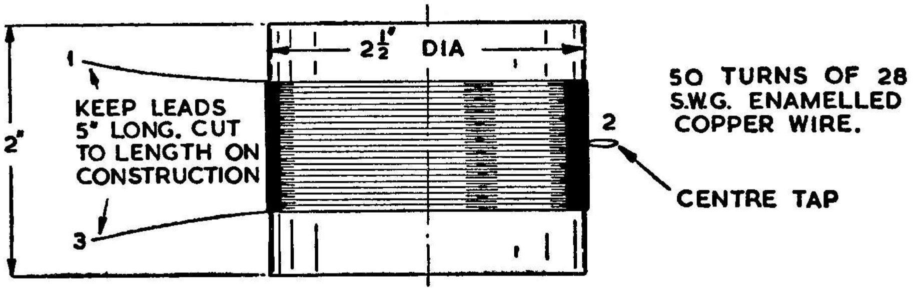
2½" DIA
1
KEEP LEADS 5" LONG. CUT TO LENGTH ON CONSTRUCTION
2"
3
2
50 TURNS OF 28 S.W.G. ENAMELLED COPPER WIRE.
CENTRE TAP

COIL DETAILS.

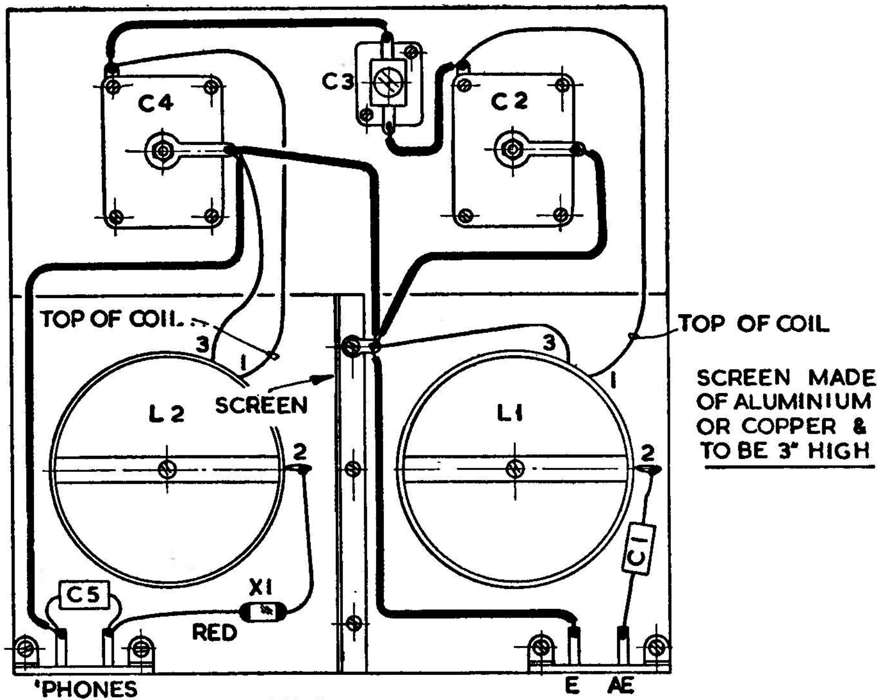
C4
C3
C2
TOP OF COIL
3
1
L2
SCREEN
2
3
TOP OF COIL
1
L1
2
SCREEN MADE OF ALUMINIUM OR COPPER & TO BE 3" HIGH
C1
C5
X1
RED
'PHONES
E
AE

FIG. 26 b

12

Fig. 27a-b uses a band-pass filter for tuning, and the coils are home-constructed.

With this design a ganged capacitor is quite suitable as the coils tune quite accurately. It is essential that the coupling condenser C3 is **non-inductive,** there will be no difficulty in obtaining a new component of this pattern but it is as well to avoid old components which may not employ this form of construction.

Receivers using this type of filter are used to the best advantage when situated a short distance from a powerful transmitter and if selectivity is not important this particular design is not the most suitable.

Both medium-waves and long-waves are catered for: consider L1-2-3, this is the first half of the filter.

L1 is the aerial coupling coil, with S1 closed,

the medium-waves are received. This coil is inductively coupled to L2 and L3. Again with S2 open L2 and L3 combined cover the long-waves and when S2 is closed the medium waves are covered.

Tuning for this half of the filter is by means of the capacitor C1, which is one half of the ganged capacitor. A screen is erected between the two sets of coils comprising the filter, and coupling is effected by C3, an earth return for the coils is provided by R1.

L4 and L5 operated in a similar manner to L2 and L3, tuning is by means of C2, the remaining section of the ganged capacitor. You will realise that the coils must be accurately wound otherwise tuning will not remain constant between the two sections over the band.

The ganged capacitor should be fitted with trimmer capacitors when purchased so that the two circuits can be balanced.

You will have noticed that this circuit uses three sets of switch contacts. A three-pole two-way wafer switch is the best choice otherwise either two or even three separate switches are necessary. All the necessary coil details are given in the diagrams, the tap on L4 is made in the same way as shown for other coils in the manual. One other component used only on this design, is the resistor R1, its size is not important because the current flowing is so small it can be ignored.

A ¼ watt rating is adequate, its precise value is not critical so that a 20% tolerance component is quite suitable.

Resistors are coded by colours and the 1000Ω type you require will bear the following sequence. At one end will be painted three coloured rings.

Components List, Fig. 27a

C1-2	500pF 2 gang, variable capacitor.
C3	0.05mfd. non-inductive paper capacitor.
C4	1000pF mica capacitor.
R1	1000 ohm resistor.
X1	Germanium Crystal.
S1-2-3	3-pole 2-way wafer switch.
L1-2-3 & L4-5	See text.
4	Terminals (Aerial, Earth and 'Phones).
	Sheet copper or aluminium for screen.

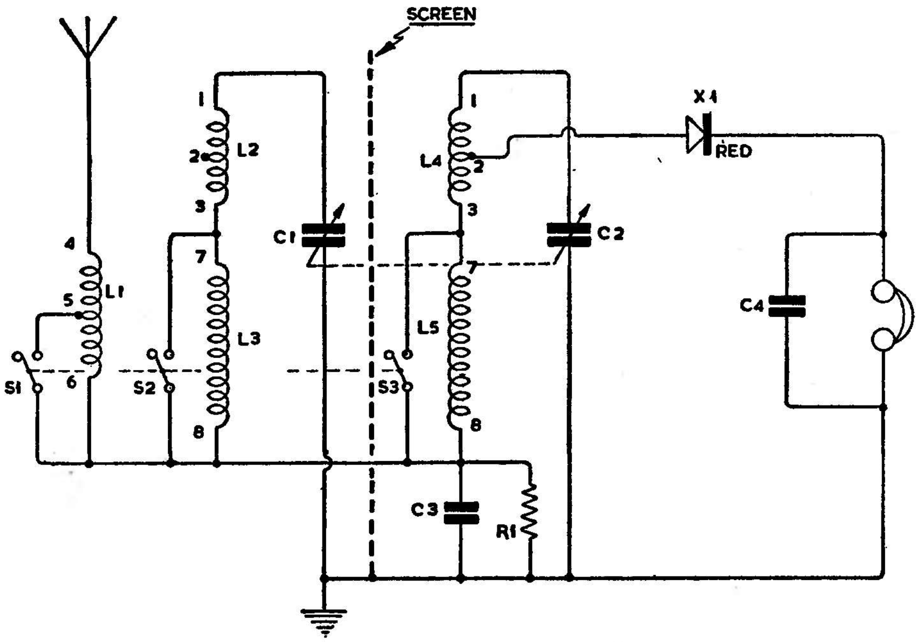

FIG. 27a

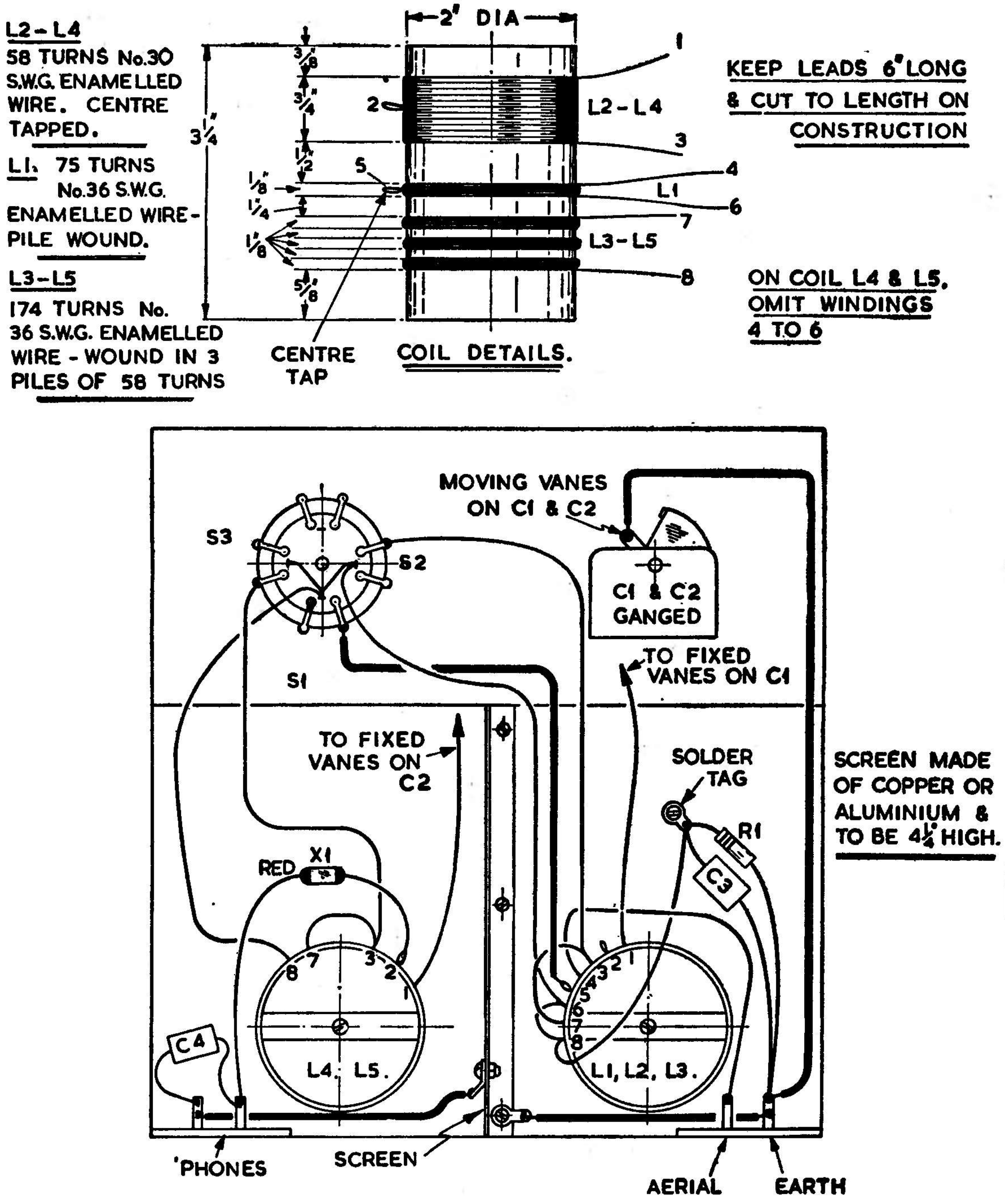

L2-L4 58 TURNS No.30 S.W.G. ENAMELLED WIRE. CENTRE TAPPED.

L1. 75 TURNS No.36 S.W.G. ENAMELLED WIRE-PILE WOUND.

L3-L5 174 TURNS No. 36 S.W.G. ENAMELLED WIRE-WOUND IN 3 PILES OF 58 TURNS

COIL DETAILS.

KEEP LEADS 6" LONG & CUT TO LENGTH ON CONSTRUCTION

ON COIL L4 & L5, OMIT WINDINGS 4 TO 6

SCREEN MADE OF COPPER OR ALUMINIUM & TO BE 4¼" HIGH.

2" DIA
3¼"
⅜"
¾"
½"
⅛"
¼"
⅛"
⅝"
1
2
3
4
5
6
7
8
L2-L4
L1
L3-L5
CENTRE TAP
MOVING VANES ON C1 & C2
C1 & C2 GANGED
S3
S2
S1
TO FIXED VANES ON C1
TO FIXED VANES ON C2
SOLDER TAG
R1
C3
RED
X1
C4
L4, L5.
L1, L2, L3.
'PHONES
SCREEN
AERIAL
EARTH

FIG. 27 b

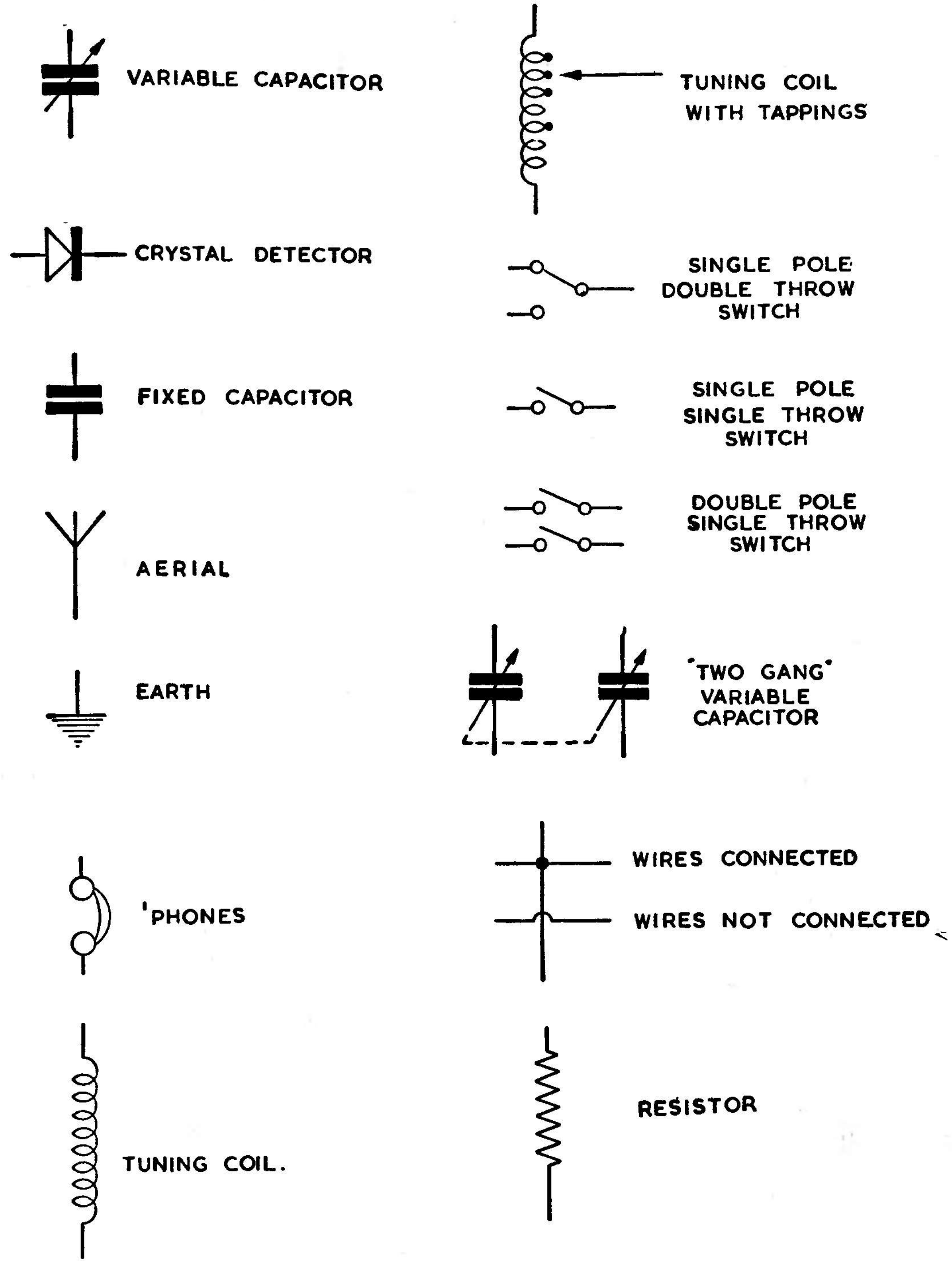
VARIABLE CAPACITOR
TUNING COIL WITH TAPPINGS
CRYSTAL DETECTOR
SINGLE POLE DOUBLE THROW SWITCH
FIXED CAPACITOR
SINGLE POLE SINGLE THROW SWITCH
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
EARTH
"TWO GANG" VARIABLE CAPACITOR
WIRES CONNECTED
WIRES NOT CONNECTED
'PHONES
RESISTOR
TUNING COIL.

SYMBOLS USED IN THE CIRCUITS DESCRIBED.

FIG. 28

The first one is BROWN which represents 1, the first figure of the value, the second BLACK, this indicates the second numeral is nought, and the third colour is RED which tells us that two further ciphers or noughts are added to give the total value. From this you can see, we have 1 plus a nought equalling 10, plus two further noughts which total 1000. Some earlier types of resistor were painted differently, the whole body colour represented the first figure, one tip was coloured to represent the second figure, and a painted dot in the centre gave the number of ciphers to be added.

There are no other special points regarding construction except perhaps that if a long aerial is used, a small 100pF capacitor might be tried in the aerial lead to prevent damping the first tuned circuit.

When setting up the receiver, tune to a station at the high-frequency end of the M.W. band and set the trimmers for loudest volume, no other adjustments are necessary.

CONCLUSION

The twelve receivers described are representative of the best crystal set designs available to-day. Their construction will provide many hours of useful enjoyment, and the results will give lasting pleasure.

A final word about components, if no actual maker is specified, any good class component can be used. Switches, fixed and variable capacitors and crystals, are available from numerous manufacturers all of which are invariably of excellent quality. Where a particular manufacturer's product is called for, the specification should be adhered to. The commercial coils specified are freely available from most supply houses specialising in components for constructors, but in case of difficulty write to Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.6, who will, on receipt of a S.A.E. be pleased to supply the address of your nearest stockist. Coil formers are not always easy to obtain and in this case advice should be sought from Post Radio Supplies, 395, Queensbridge Road, London, E.8, who will be pleased to help in any way possible.

* *

Crystal Set Mobile

By Al Klase – N3FRQ – December 2006

Back in the 1990's there was a Yahoo Internet "club", devoted to crystal radio development called "The Ferrite Core." One of the perennial topics of conversation was the difference between operating a crystal set near high-power stations as opposed to a more rural setting. Someone, who had read Aesop's Fables likened this to "The City Mouse and the Country Mouse." The country mouse needed a sensitive radio hooked to a large antenna, but could faintly hear distant stations without a lot of interference. The city mouse's crystal set could drive a loudspeaker using only the bed springs for an antenna, but he couldn't hear much beyond the city limits.

I've spent most of my life as a country mouse, but recently I've been frequently finding myself in Jersey City. Indeed, crystal set operation is very different there. Short antennas work pretty well, and even my nearly-deaf rocket radio picks up multiple stations. It's even possible to operate a properly designed set with the antenna terminal connected to ground with the headphone cord and listener serving as the actual antenna.

I decide to apply what I learned down town to a couple of radios I was building as Christmas presents. The circuit is very similar to the NJARC Pretty-Good Crystal Set, but the coil is wound on a ferrite toroid to save space, and there are additional antenna taps on the "high end" of the coil to better accommodate short antennas.

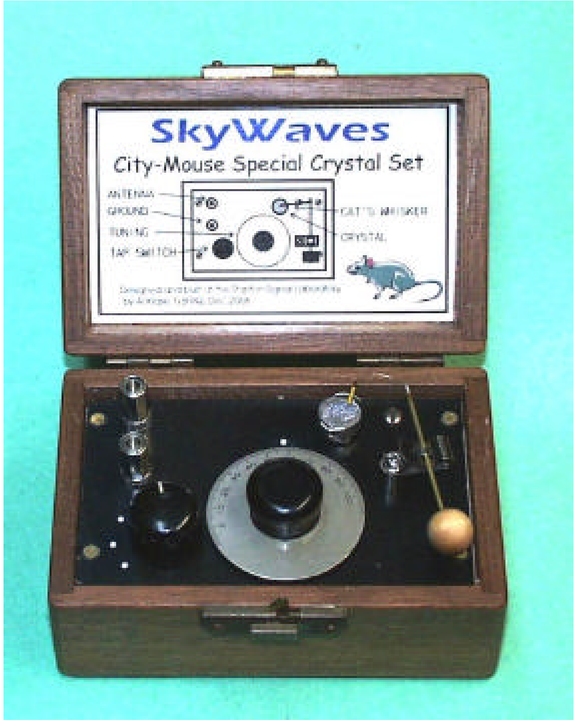

I recently found myself in Maywood, near Hackensack, with some time to kill while Peggy rehearsed before a Sunday-afternoon concert. I took the city mouse for a walk in a nearby park. With 20 feet of wire, 5 feet in the air, and the ground wire clipped to a metal park bench, I could comfortably choose from five or six stations. I could still hear the Jets game on WABC, with just the ground connected to the antenna terminal.

Still having some time to waste, I picked up a four-foot stick as I returned to the car. I grounded the set to the ignition key, and extended the antenna wire out the sunroof with the stick. Sure enough, the Jets were still ahead. I wound the rest of the wire around the stick as a loading coil.

The ugly stick

After an enjoyable concert (The Four Seasons and a Brandenburg Concerto w/harpsichord), I listened to the crystal set down the turnpike through the Meadow Lands, past all the towers, all the way home. I think most were on night-time power by now, as none were obnoxiously loud. I need to try that again, in broad daylight, with a speaker.

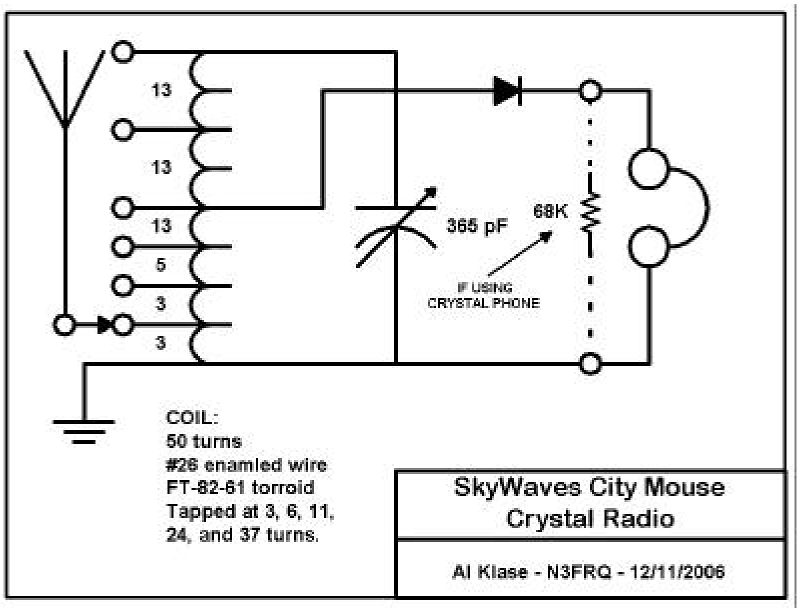
13
13
13
5
3
3
365 pF
68K
IF USING
CRYSTAL PHONE
COIL:
50 turns
#26 enamled wire
FT-82-61 torroid
Tapped at 3, 6, 11,
24, and 37 turns.
SkyWaves City Mouse
Crystal Radio
Al Klase - N3FRQ - 12/11/2006

all about
CRYSTAL SETS

by
Charles Green

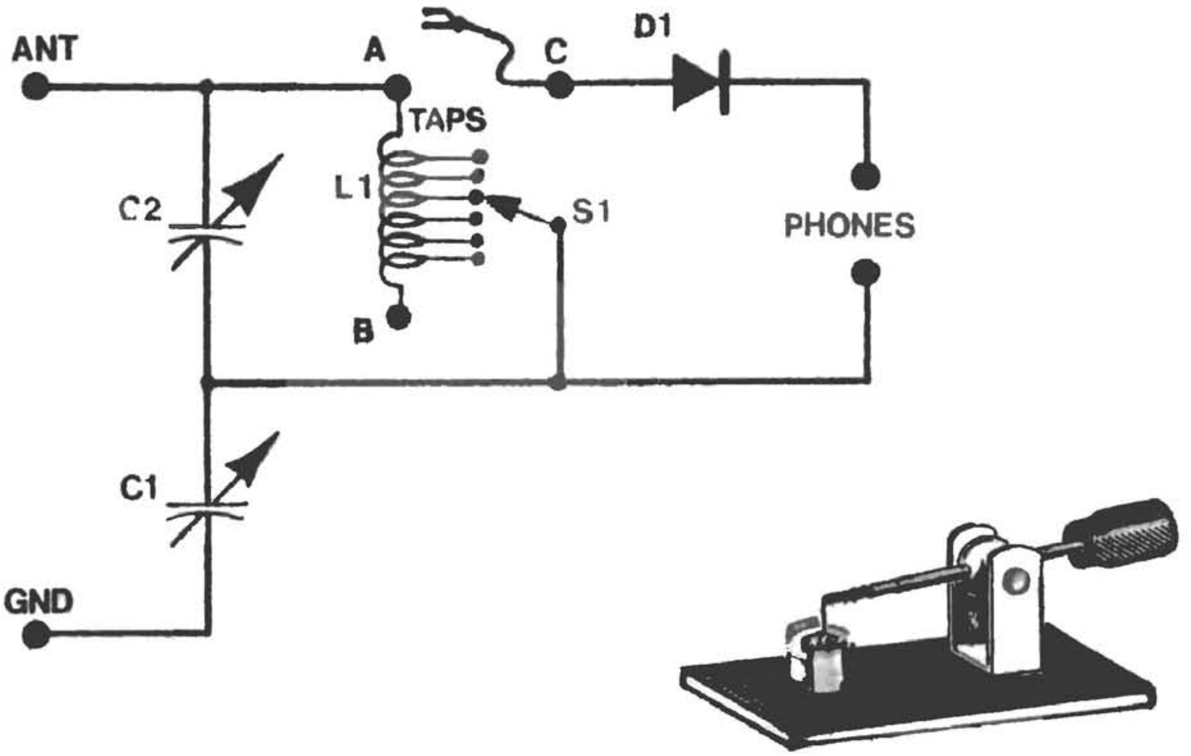
ANT
A
C
D1
TAPS
L1
C2
S1
PHONES
B
C1
GND

How-To-Build Simple
Crystal Set Radios

621.3841
G795A
c.1

allabout books no. 101

all about
CRYSTAL SETS

by

Charles Green

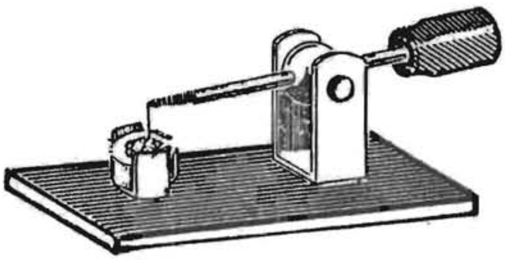

How - To - Build Simple
Crystal Set Radios

FIRST EDITION

All rights reserved, including the
right of reproduction in whole or
in part in any form.

Copyright © 1984 by Charles Green

Published by Allabout Books
P.O. Box 22366
San Diego, CA 92192-2366

ISBN 0-930003-00-4

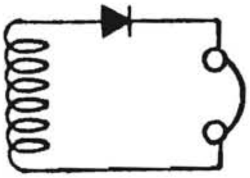

PREFACE

In the beginning of this century, when radio was called "Wireless", most of the early experimenters used a crystal set to receive radio signals.

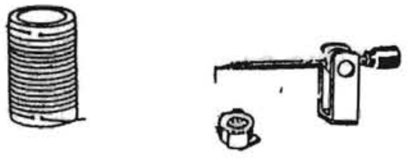

The crystal set is still the simplest type of radio receiver. A catwhisker (made from a thin wire) touching a sensitive spot on a piece of galena or silicon mineral crystal and a tuning coil, form the basic parts of this radio.

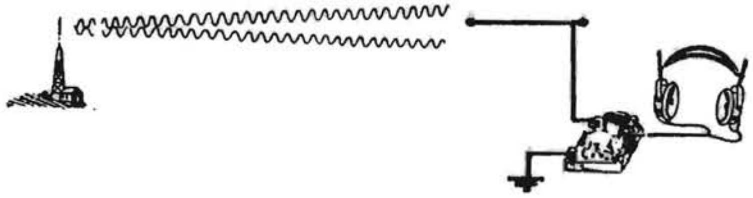

Radio signals from a nearby station are picked up via antenna and ground connections to the crystal set. The galena or silicon crystal detects the signals and the resultant audio is heard with headphones. No battery or electrical power is required.

This book tells you about crystal sets and shows you how to build them.

Charles Green

CONTENTS

1

THE CRYSTAL DETECTOR....................................4
How-it-Works, 4
Detection of Radio Waves, 6
Tuning the Crystal Set, 7
Capacitor, 8
RF Coil, 8
Tuned Circuit, 9
Antenna and Ground, 10
Headphones, 11

2

SIMPLE RECEIVER PROJECTS.............................12
Simple Crystal Set, 12
Slider Tuned Crystal Set, 17
Ferrite Coil Crystal Set, 23
Spiderweb Coil Crystal Set, 27

3

ADVANCED RECEIVER PROJECTS...........................32
Dual-Tuned Crystal Set, 32
Selective Crystal Set, 37
Carborundum Crystal Set, 43

4

COMPONENT PROJECTS...................................48
Book Tuning Capacitor, 48
Rotary Switch, 53

5

WORKSHOP...54
Soldering, 54
Project Construction Tips, 55
Mounting Minerals, 56

INDEX..57

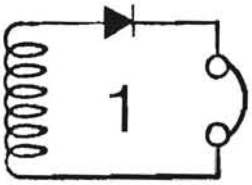

THE CRYSTAL DETECTOR

A crystal detector of radio waves can be either a traditional catwhisker and natural mineral crystal, or a modern semiconductor diode.

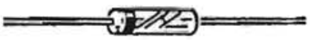

SEMICONDUCTOR DIODE

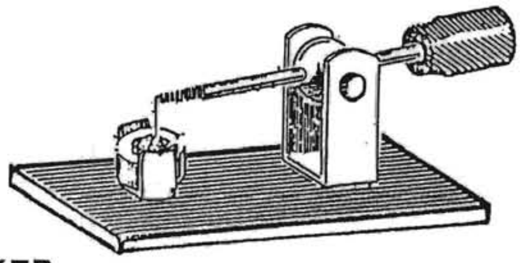

CATWHISKER DETECTOR

Certain natural minerals such as galena (lead sulphide) and silicon have electrically sensitive surface spots. In the beginning of this century, experimenters discovered that electrical current would flow into the sensitive spot on a mineral crystal from a catwhisker, but would not pass readily (high resistance) from the crystal to the catwhisker. This unique feature was used to detect (rectify) radio waves and extract the audio modulation.

HOW-IT-WORKS

There are two general types of crystal detctors;

- . Point contact diode
- . Junction diode

Both types use semiconductor material such as silicon, germanium, or galena. The basic semiconductor junction diode is made of two portions of the material "doped" in manufacturing by adding controlled amounts of chemicals. The doped portions have opposite electrical properties:

"P" TYPE, with an excess of positive electrical charges (holes).
"N" TYPE, with an excess of negative electrical charges (electrons).

Junction Diode

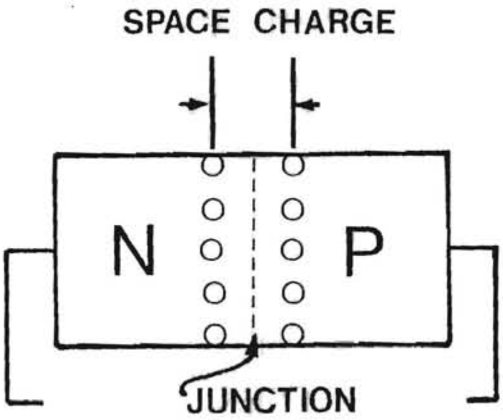

The boundary between the two doped portions is called a PN junction. The area on both sides of the PN junction is called the space charge region. When a positive voltage is applied to the "P" material (and negative voltage to the "N" material), the space charge region becomes narrow and electron current flows through the PN junction.

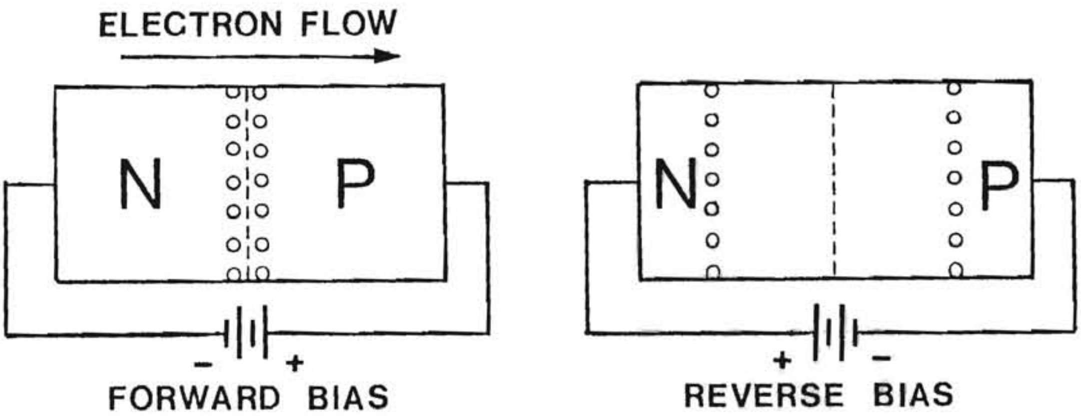

When the voltage polarities are reversed, negative to "P" and positive to "N", the space charge region widens (increasing the effective resistance of the PN junction) and the electron flow is very small (back current). If an ac voltage is applied to the PN junction, it will act as a rectifier (or radio wave detector).

Point Contact Diode

The point contact diode uses a "catwhisker" of thin wire contacting the surface of doped "N" semiconductor material. The pointed end of the wire produces a high intensity electric field in the area of the semiconductor surrounding the wire. When an electric current flows from the wire to the semiconductor material, a small region of "P" material forms around the area of the point contact. Thus a PN junction is formed which operates in the same way as the junction diode.

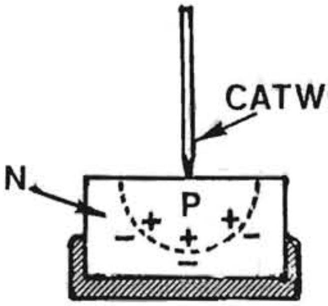

The sensitive spots of galena and silicon crystals used in traditional catwhisker detectors have naturally doped areas of semiconductor material, and operate in the same way as the point contact detector.

DETECTION OF RADIO WAVES

At the transmitting radio station, a microphone converts sound waves into electrical audio signals. The audio signals then modulate the transmitted radio waves.

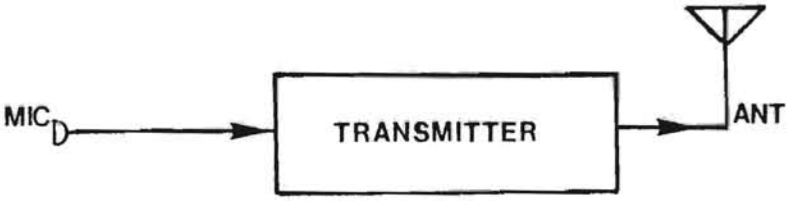

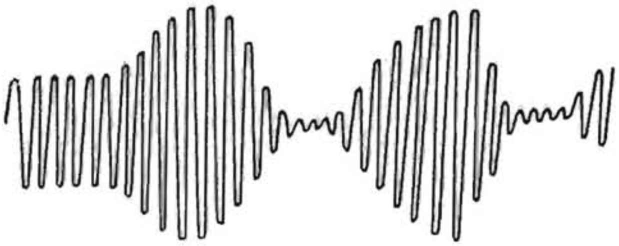

MODULATED RADIO WAVES ⟶

the crystal detector

The transmitted radio waves are intercepted by the receiving antenna and small rf currents are induced in the antenna wire. The rf currents (high frequency ac) flow through the detector to ground. As previously explained, the crystal detector will pass current in only one direction, therefore, pulsating dc voltages (detected audio signals) are developed by the crystal equivalent to the audio modulation of the transmitted radio waves.

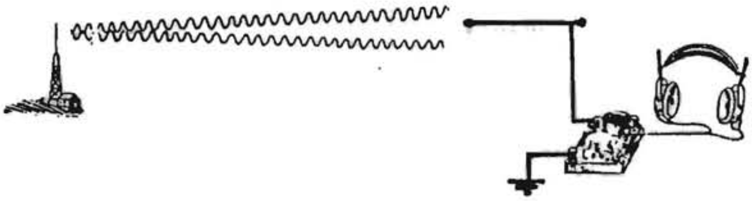

The detected audio signals flow through the headphone coils and affect the intensity of a magnetic field. This causes the steel discs of the headphones to vibrate and produce audible sound waves.

TUNING THE CRYSTAL SET

The antenna intercepts all radio waves in the area, and the signals will be heard jumbled up in the earphones. A nearby radio station may be received with the other stations heard in the background.

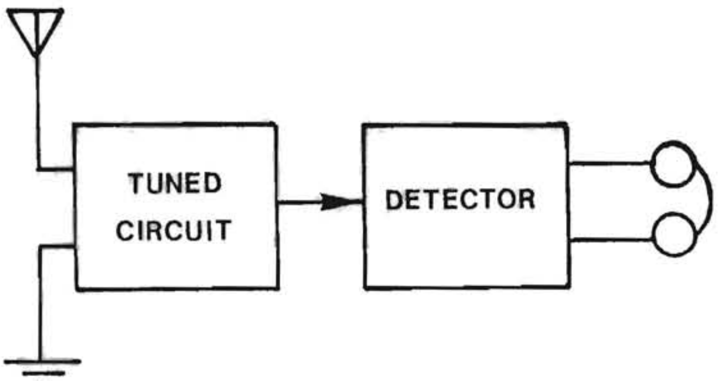

In order to receive a radio station properly, the crystal set must be tuned to the frequency of the transmitter. A tuned circuit composed of an rf coil (inductance) and a capacitor, is connected between the antenna/ground circuit and the detector.

Capacitor

The modern capacitor is evolved from the old Leyden jar. It consists of two or more metal sections separated by a dielectric (air, paper, or plastic). The capacitor stores energy in the form of an electrostatic charge in its dielectric. A variable capacitor (used in a tuned circuit) has movable metal sections that adjust the amount of capacity.

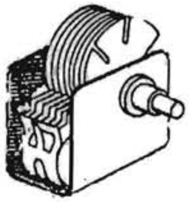

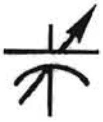

A capacitor will allow pulsating dc current or ac current to flow through it, but will block steady dc current. A capacitor has an oppositon or "reactance" to ac that varies with frequency. The higher the frequency, the lower the reactance of a given value of capacity. Capacity is measured in micro-farads (uF) or micro-micro-farads (picofarads) (uuF or pF).

RF Coil

An rf coil is made of insulated copper wire wound on a plastic or cardboard tube. When electric current flows through the coil, a magnetic field is created. This magnetic field resists any change in electric current flow. This opposition or "reactance" to ac varies with frequency. The higher the frequency, the higher the reactance of a given value of rf coil inductance. Inductance is measured in millihenrys (mH) and microhenrys (uH).

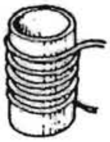

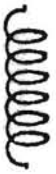

Tuned Circuit

A tuned circuit is composed of a capacitor and an rf coil. As previously described, the reactances of a capacitor and an rf coil change with frequency (but in opposite ways). At a particular frequency, the two reactances will cancel each other out. This condition is called "resonance". The impedance of a parallel tuned circuit is at its highest at resonance. Radio signals at the resonant frequency will be greatly magnified, while signals at other frequencies will be shunted to ground (rejected).

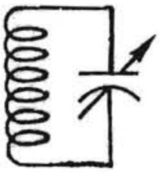

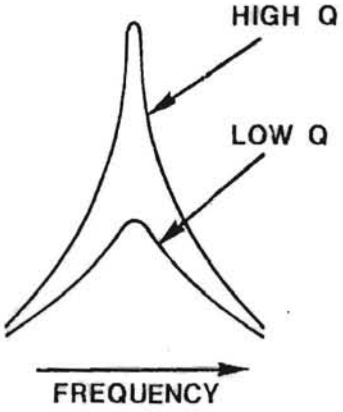

The resonant frequency of a tuned circuit can be changed by adjusting the capacitance (using a variable capacitor), or by varying the inductance (using a slider or taps on the rf coil). A ferrite core can also be used to change the inductance of a coil. The selectivity of a tuned circuit is dependent upon the quality (Q) of the components and the inductance /capacity ratio (L/C ratio). A high-Q tuned circuit is more selective than a low-Q circuit. A tuned circuit with a high L/C ratio (large inductance and small capacity) is more selective than a low L/C ratio circuit.

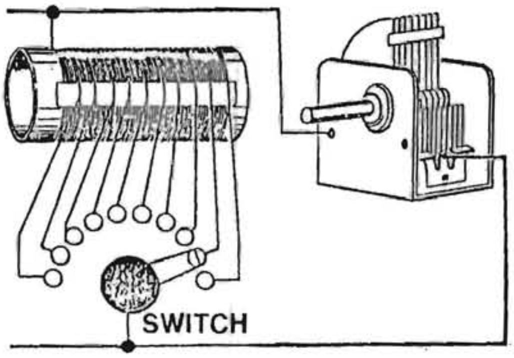

ANTENNA AND GROUND

A simple crystal set depends upon the strength of the radio waves received by the antenna to produce audio in the earphones. If you live in a strong signal area (a nearby radio transmitter), a crystal set can produce loud audio in earphones with only an inside antenna. But in lower signal areas, a good outside antenna with a ground connection is required.

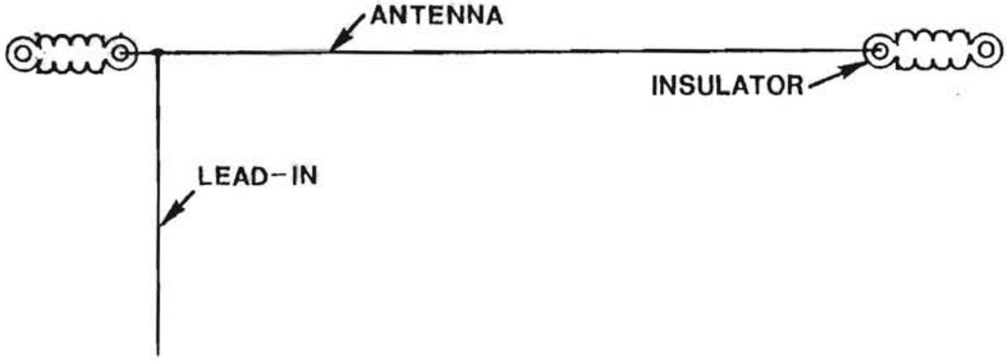

If you are fortunate enough to be able to put up an outside antenna, the basic rule of thumb is the higher, the better reception. Check your local electronics store for an antenna kit. You can construct an antenna using a glass or plastic insulator at each end of a length of copper wire (insulated or bare). Make the antenna as long as possible. It does not have to be in a straight line to work.

CAUTION

ANTENNA INSTALLATION
NEAR ELECTRICAL POWER
LINES IS DANGEROUS.

Make sure that your antenna
is far enough away from any
power line, so that it will
not touch a wire if it falls.

Connect a length of insulated wire between the end of the antenna and your crystal set antenna terminal. This antenna lead-in wire can be fed into your house through a window.

If you can not erect an outside antenna, run a length of insulated wire as long and high as possible through your house or apartment. It will work, but not as well as an outside antenna because of the shielding affect of the house walls.

Connect a ground lead to a cold water pipe. Make sure that it is a metal pipe. Or, you can buy a ground rod from your local electrical supply store and drive it into the ground near your house. Then connect your crystal set ground lead to the rod. For safety install a lightening arrestor between the antenna lead and ground (see the instructions with the device).

HEADPHONES

Headphones (also called earphones, or headset) are required to hear the received audio from a crystal set. The standard headphones are made with dual coils of fine wire connected in series in each phone. Horseshoe magnets are built into each phone case to provide a magnetic field. A disk of thin sheet iron (diaphragm) is mounted near the top of the coils. The phones are connected in series with a set of flexible leads terminated in metal "phone tips".

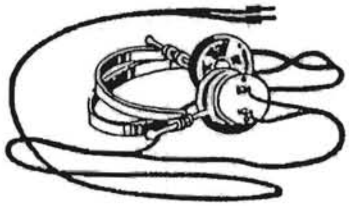

Detected audio signals flow through the coils and affect the intensity of the magnetic field from the horseshoe magnet. This changes the magnetic attraction of the diaphragm, and causes it to vibrate and produce audible sounds.

The total resistance of the series headphone coils is used as a measure of audio sensitivity. A resistance of 2000 ohms is sufficient. More sensitive headphones have a resistance of over 5000 ohms and are used to hear very weak signals.

The 2000 ohm headphones can be obtained at electronic parts mail order houses.

2

SIMPLE RECEIVER PROJECTS

SIMPLE CRYSTAL SET

Look at the schematic diagram. The crystal set uses a tunable ferrite coil L1, and a crystal diode D1. This type of tuning coil is called a "loopstick". The ferrite core is screwed into the coil form and increases the inductance of L1, thus tuning the receiver to a lower frequency.

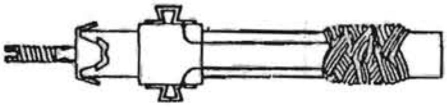

CONSTRUCTION

As shown in the illustration, the crystal set components are mounted on a 5 1/2-inch by 2 1/2-inch by 3/4-inch wood base. Layout and install the crystal set parts in the locations shown. The loopstick L1, is mounted with a metal strip supplied with it. Push the snap-in end of L1 into the large hole of the metal strip. Then bend 1/2-inch of the other end of the strip to form a bracket and mount it on the base with a wood screw. If the metal strip is not available, make one out of sheet metal approximately 2-inches long by 1/2-inch wide. Then drill a hole to fit L1 at one end, and a small hole to fit a wood screw at the other end.

Mount the four fahnestock clips and solder lugs with wood screws in the locations shown in the illustration.

WIRING

Use hookup wire to connect the crystal set components as shown in the schematic and illustration. The loopstick L1 may have three terminals. One of the terminals is a tap, which is not used in this circuit. Make sure that you connect only to the terminals at the start and end of the loopstick coil.

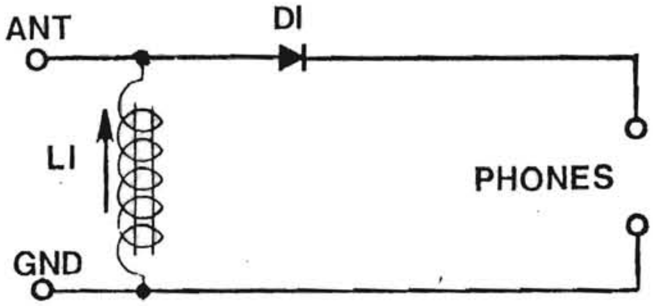
ANT
DI
LI
GND
PHONES

SIMPLE CRYSTAL SET

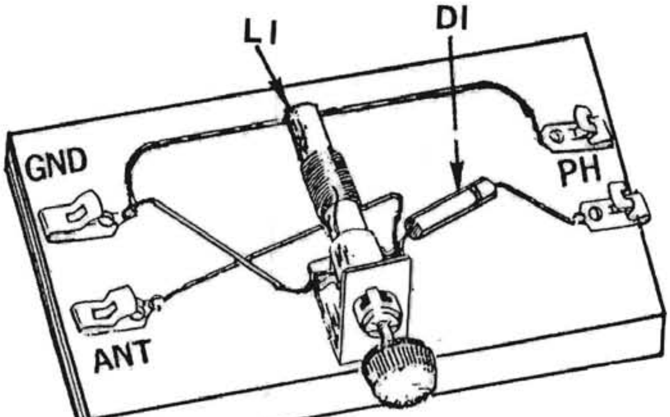
LI
DI
GND
PH
ANT

List of Materials

D1 - Germanium diode
(1N34A or equiv.)
L1 - Loopstick
(see text)
4 - Fahnestock clips
BASE - 5 1/2 X 2 1/2 X
3/4-inch wood section
PHONES - 2000-ohm headphones
MISC. - Solder lugs, wood screws,
knob and bracket to fit L1,
hook-up wire.

Install a small knob on the end of the L1 adjustment screw. Make sure that the screw turns freely with the knob. Position the 1N34A diode as shown before soldering the leads.

Experiment 1

Connect the antenna, ground, and 2000-ohm headphones to the crystal set terminals. Rotate the L1 knob counter-clockwise as far as it will go. Do not turn past this point, or you may break the ferrite rod inside the coil form. The ferrite is now outside of the coil winding. This is the highest frequency point of the tuning range of L1.

Slowly turn the L1 knob clockwise and listen for received radio stations. When the L1 knob is turned fully clockwise, the ferrite will be completly inside the coil (increasing the inductance) and it will be at the lowest frequency point of the tuning range.

Experiment 2

Connect capacitors, with values from 50 mmf to 300 mmf, in parallel with L1. Then tune L1 and note how the frequency range has changed.

AMPLIFIER CONNECTION

The simple Crystal Set (and the other receiver projects) can be connected to an amplifier in place of earphones. This amplifier can be a single channel input of your home stereo.

Parts values are not critical; C1 is approximately .005 mfd, and the volume control R1 can be any value from 1000 ohms to 10,000 ohms. Use shielded wire between R1 and the amplifier input jack.

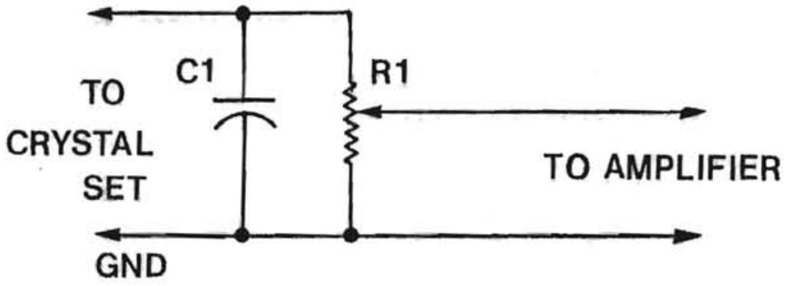
TO
CRYSTAL
SET
GND
C1
R1
TO AMPLIFIER

Amplifier Connection

Schematic Symbols

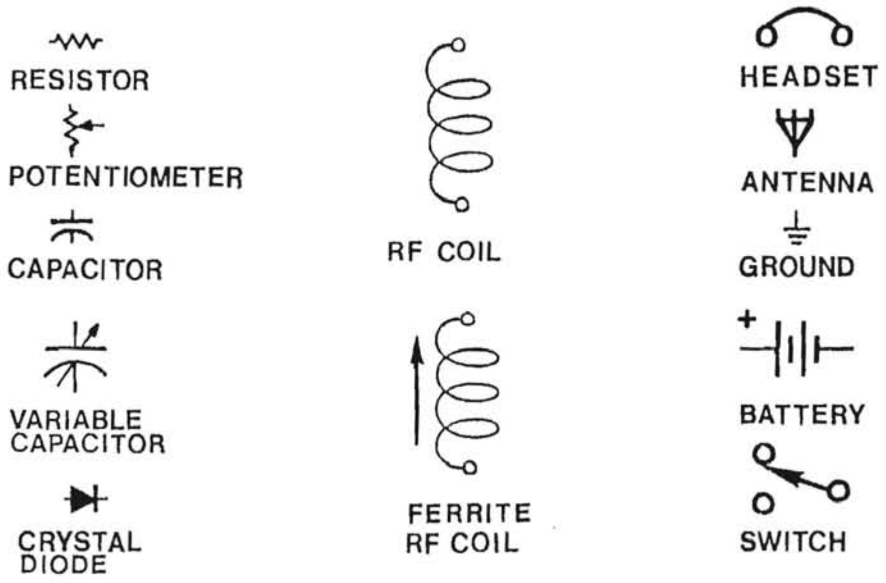
RESISTOR
POTENTIOMETER
CAPACITOR
VARIABLE CAPACITOR
CRYSTAL DIODE
RF COIL
FERRITE RF COIL
HEADSET
ANTENNA
GROUND
+
BATTERY
SWITCH

List of Materials

D1 - Germanium diode, type 1N34A, 1N60, 1N82A, or equiv.
L1 - Slider-tuned coil (see text)
4 - Fahnestock clips (terminals)
BASE - 6 x 3 x 1/2-inch section of wood (see text)
PHONES - 2000 ohms headphones
MISC. - Coil form (see text), No. 24 enameled copper wire, 3/32 inch dia. brass rod, brass ball slider (see text), solder lugs, two small brackets, hook-up wire, wood screws, machine screws and nuts.

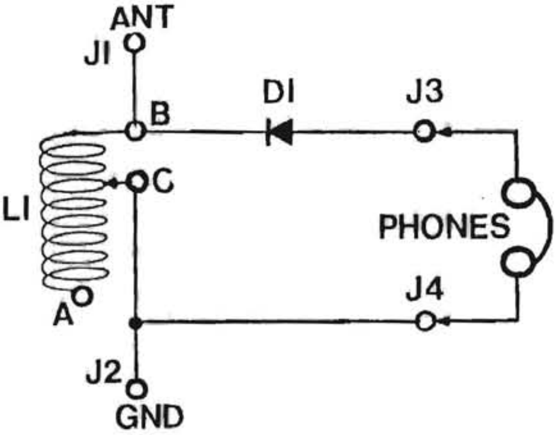
ANT
JI
B
DI
J3
C
LI
PHONES
A
J4
J2
GND

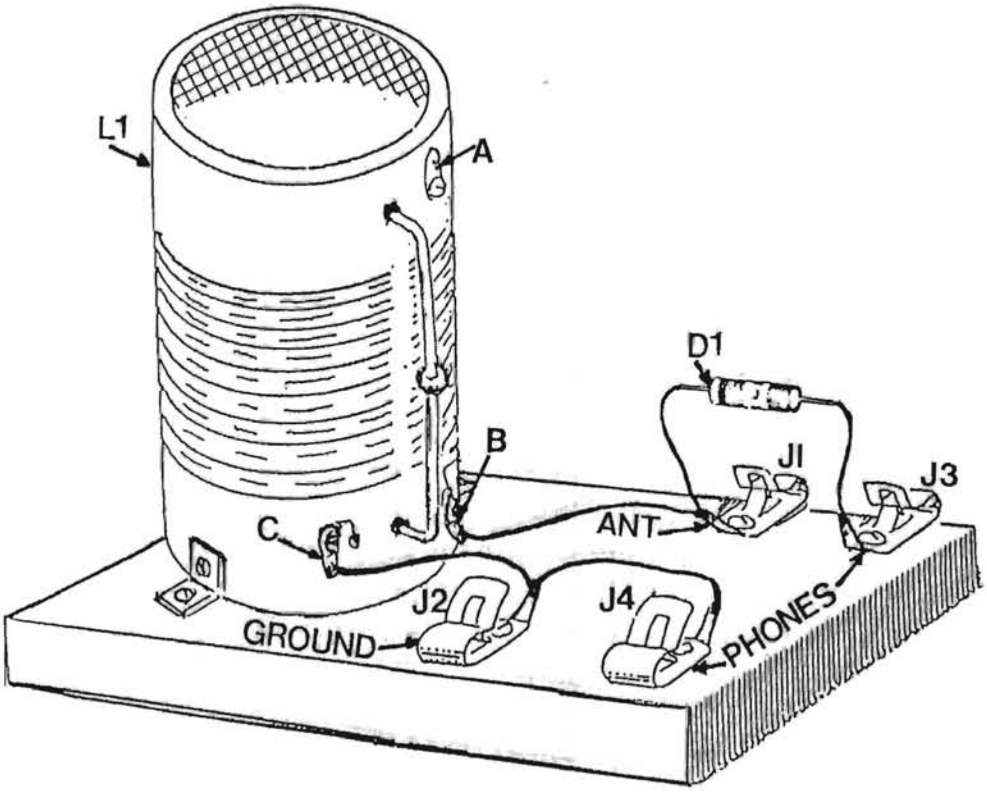
L1
A
D1
B
JI
J3
C
ANT
J2
J4
GROUND
PHONES

SLIDER TUNED CRYSTAL SET

SLIDER TUNED CRYSTAL SET

In the early days of radio, slider tuned coils were used in crystal sets. This method of tuning the inductance uses an adjustable metal contact (slider) that moves over the coil. The slider makes contact with individual coil wires to tune in the radio stations.

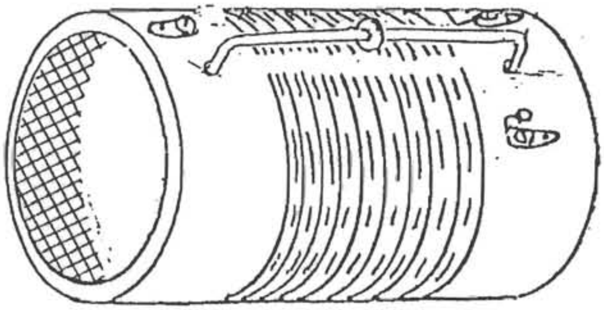

The enamel insulation of the coil wires in the slider path is removed for good electrical contact.

You can experiment with a slider tuned coil by building this crystal set. As shown in the illustration, the coil is mounted upright on a 6-inch by 3 1/2-inch by 1/2-inch wood base. A semiconductor diode is used for a detector.

TUNING COIL CONSTRUCTION

Look at the drawing of the tuning coil L1 construction details on the next page. The tuning coil is wound on a cardboard or plastic mailing tube section, 2 5/8-inches in diameter and 5-inches long. Make sure that you select a rigid, thick-walled tube for a coil form. The tube used in our model has a 3/16-inch wall. The coil form must be rigid to hold the slider ball and rod assembly.

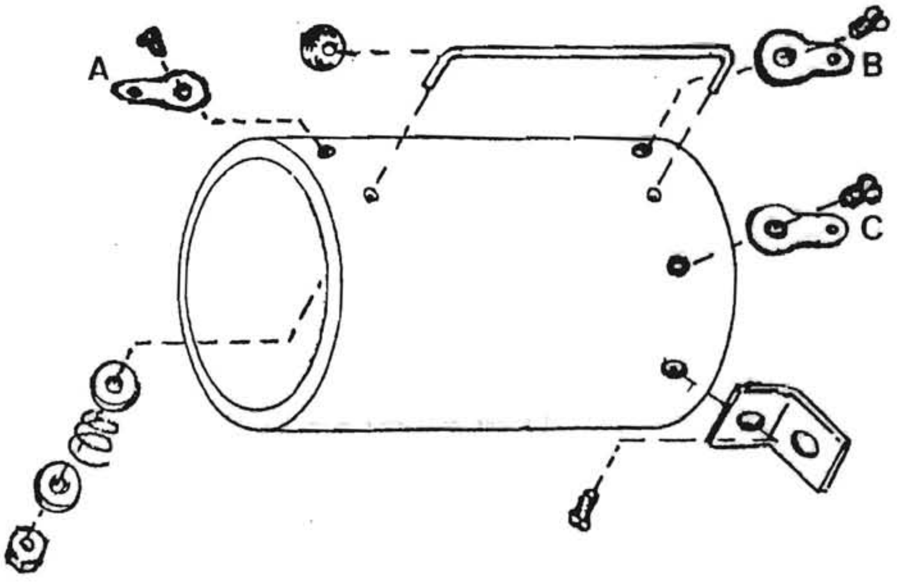

Before winding the coil, drill all the holes in the form as shown in the drawing, and then install the solder lugs. Position the solder lug ends toward the coil form edges.

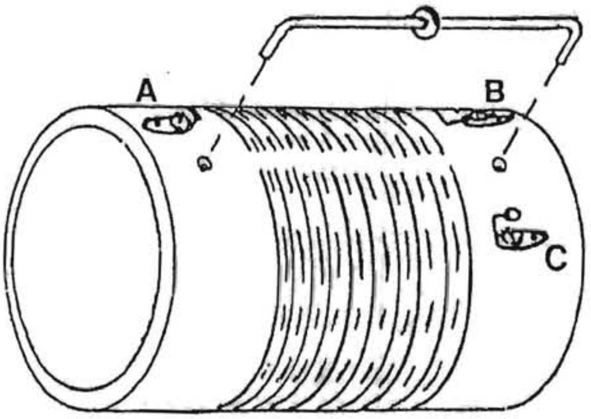

The coil is wound with No. 24 enameled copper wire. Remove the enamel from the wire end with sandpaper and then solder the wire to the lug. Start winding the coil, keeping the turns as tight and close together as possible. Wind the wire to the front of the bottom solder lug B. Remove the enamel and solder the wire end to the lug.

A brass ball approximately 3/8-inches in diameter is used as the moving contact slider. This brass ball can be found at your local electrical supply store. It is a type of ball that is used as a top nut for small table lamps. Drill a hole through the ball so that it will slide over a length of 3/32-inch diameter brass rod. Bend the brass rod at each end (after placing the ball in the center of the rod), to fit through the holes in the coil form as shown in the drawing. Cut the rod 1-inch down from the bend at each end.

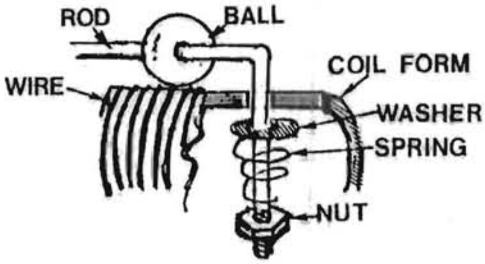

Insert the brass rod with the slider ball into the coil form holes. Then place a small washer and spring on one end of the rod (inside the coil form). Solder a nut on the end of the rod to hold the spring in place. Do the same to the other end of the rod. After soldering, slide the ball over the wires and check that it slides smoothly. Use sandpaper to remove the enamel insulation of the wires in the path of the ball slider. Connect a wire between the soldered nut on the bottom end of the brass rod and the solder lug near the rod.

BASEBOARD

All of the crystal set components are mounted on a wood baseboard as shown in the drawings. Cut the baseboard to size; a 6-inch by 3 1/2-inch section of 1/2-inch thick wood (maple is used in our model). Sand the edges after cutting. It is not necessary to paint or stain the baseboard.

Mount the four terminals and solder lugs as shown in the drawing. Label the four terminals with either rub-on lettering or cemented paper tags. Install two brackets on opposite sides of the bottom of the coil, and mount it on the baseboard with the lugs toward the terminals.

WIRING

Bare No.22 wire was used to wire our model, but any kind of hookup wire can be used. Connect the wires between the solder lugs as shown in the drawing, except do not connect the antenna terminal lug wire to the coil terminal at this time. Connect the crystal diode between the solder lugs as shown.

Experiment 1

This experiment will show how an untuned crystal detector functions.

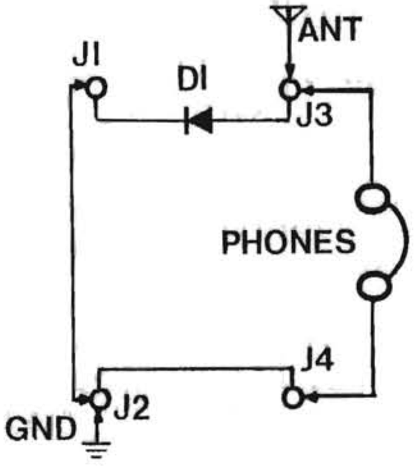

Connect the untuned detector circuit as shown in the drawing, and then hook-up your antenna and ground leads. Connect 2000-ohm headphones to the circuit and listen to the detected radio waves.

You will probably hear a strong local radio station, with other, weaker radio station transmissions all jumbled together.

Experiment 2

Reconnect the baseboard wires to the circuit of the Slider-Tuned Crystal Set as shown in the drawing. Listen to the detected radio waves as you tune the coil by sliding the brass ball over the sanded wire path. You should be able to receive the radio stations much better than the untuned detector circuit of Experiment 1.

LOADING COIL

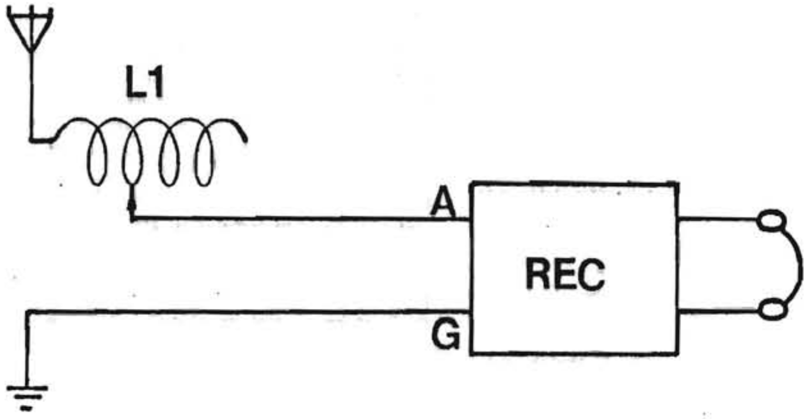

An additional slider tuned coil, L1, can be used as a Loading Coil to increase the electrical length of your antenna. This may allow better reception of the lower broadcast band frequencies. Connect L1 in series with the antenna lead to the receiver, and tune it for best reception.

WAVE TRAP

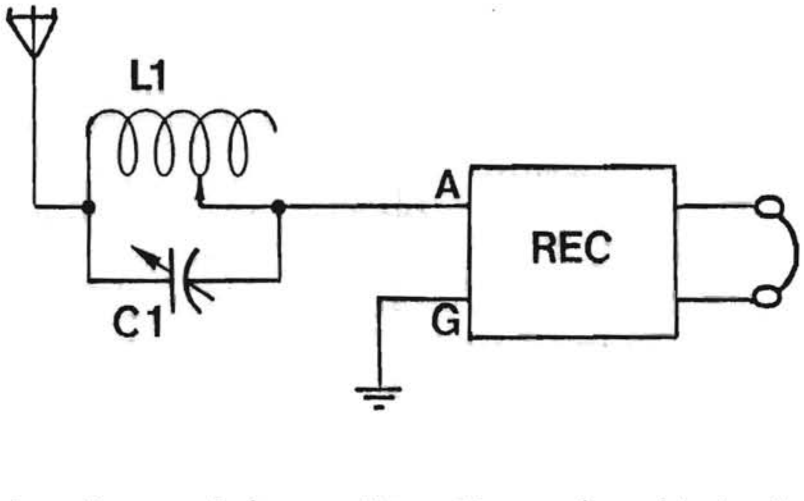

L1 can also be used in a Wave Trap circuit to tune out a strong local signal, and allow easier tuning of weaker stations. C1 is a 365 mmf variable capacitor connected in parallel with L1. Adjust L1 and C1 so that the interfering station is at minimum.

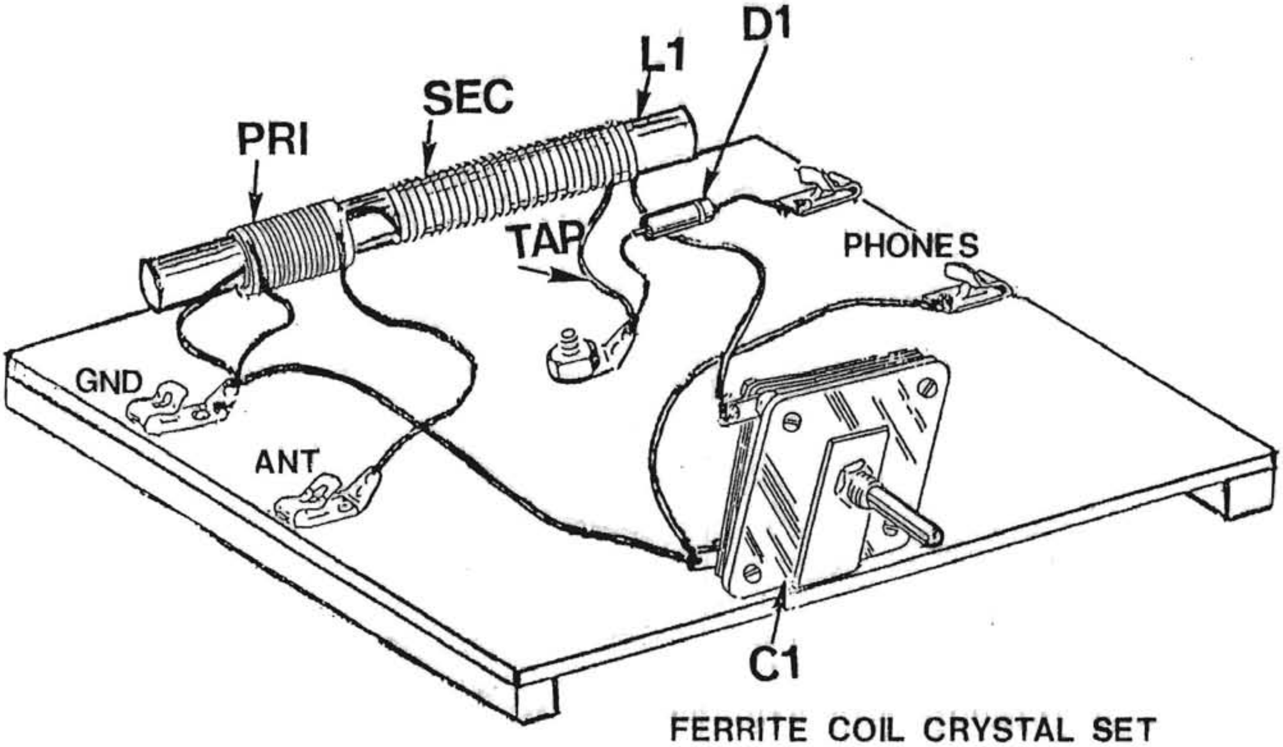
D1
L1
SEC
PRI
TAP
PHONES
GND
ANT
C1

FERRITE COIL CRYSTAL SET

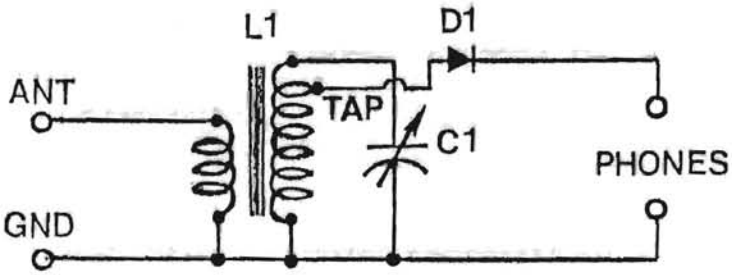
L1
D1
ANT
TAP
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Germanium diode (1N34A or equiv.).
L1 - Ferrite coil (see text).
4 - Fahnestock clips.
2 - wood sections; 6 X 1/2-inch square (see text).
BASE - 8 1/2 X 6-inch sheet plastic (see text).
PHONES - 2000-ohm headphones
MISC. - plastic tube section for L1 primary winding (see text), solder lugs, machine screws and nuts, wood screws, bracket for C1 (see text), knob for C1, hookup wire.

FERRITE COIL CRYSTAL SET

This crystal set has a large ferrite core coil L1 of a type that is normally used for a built-in antenna in portable radios. This type of coil has a very high "Q" and may be more selective than an air-core coil. As you can see in the schematic, a primary coil is wound around the ferrite coil. The original winding is used as the secondary.

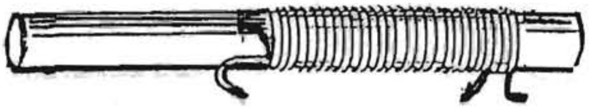

The primary coil (connected to the antenna and ground) can be moved close to, or away from, the secondary (original winding connected to D1). This variable coupling is used to control the selectivity of the crystal set.

PRIMARY COIL

The primary coil is wound on a 1-inch section of plastic tube that can fit loosely over the ferrite rod. A 1/2-inch diameter plastic tube is used on the receiver shown in the illustration. Wind 12-turns of hookup wire evenly over the tube and twist the ends together to prevent movement. Leave 4-inch leads from the primary coil winding. Slide the primary coil over the center of L1.

BASEBOARD

The crystal set components are mounted on a 8 1/2-inch by 6-inch section of sheet plastic. The thickness of the plastic is not critical, but it must be rigid enough to hold the parts without bending. A 6-inch long 1/2-inch square wood section is installed under each end of the baseboard. A section of wood can be used for the baseboard with wood screws to mount the parts, in place of machine screws and nuts.

The model shown uses a plastic variable tuning capacitor, but an air type capacitor can be used for C1 instead. Make a metal bracket approximately 1-inch wide by 1 1/4-inch high and a 1/2-inch right angle mounting foot from sheet aluminum. Drill a hole in the top of the bracket to fit C1 and two holes in the mounting foot to fit small machine screws. Locate the components on the baseboard as shown in the illustration and then drill mounting holes. If you have a cardboard or plastic mounting supplied with ferrite coil L1, you can mount it on the baseboard. If not, L1 can be taped loosely on top of the baseboard. Th ferrite coil L1 in our crystal set was obtained from an old radio. Install the components on the board with solder lugs, machine screws, and nuts.

WIRING

Connect the parts with hookup wire as shown in the illustration. Use the schematic diagram as a guide. Make sure that the ground lead from L1 secondary passes inside the primary coil. Also check that the primary coil slides freely over the L1 winding. Install a knob on the C1 shaft, and mark the terminal designations on the baseboard.

Experiment 1

Hookup the antenna, ground and 2000-ohm headphones to the crystal set terminals. Slide the primary coil over the center of the L1 winding. Tune C1 for radio stations. Then slide the primary coil towards the ground end of L1 while tuning C1. Note the difference in selectivity.

Experiment 2

Change the number of turns of the primary winding and repeat Experiment 1.

Operating Tips

After operating a crystal set for some time, you may notice a difference between day, evening, and night reception of radio stations. This is due to the two ways that radio signals travel between the transmitting antenna and a receiving antenna.

(a) The ground wave; which moves along the surface of the earth.
(b) The sky wave; which travels upward and is reflected by the ionosphere.

The reception of broadcast stations is dependent upon the time of day. During daylight hours, stations are received primarily by ground wave transmission. At night, the broadcast stations will be received via sky wave transmission. Also, some radio stations change to a lower power at night. Reception in the evening may be more difficult as the fading ground wave is interfered with by reflected waves. But later, it may be possible to hear far-away stations transmitted by sky wave.

Your geographical location to radio stations is very important for reception, as a crystal set receiver does not amplifiy signals. It is entirely dependent upon the strength of the radio waves. If you are in a deep valley surrounded by high mountains, reception will be difficult. If you live in a flat plain, with no obstacles to radio transmission, reception may be very good.

An outside antenna is necessary for for best reception. If possible, try erecting an extra antenna at a right angle to the original one, and switch it in to provide a directional effect for increased signal strength. Keeping a log of received stations for a week may be helpful in determining the radio reception conditions in your area.

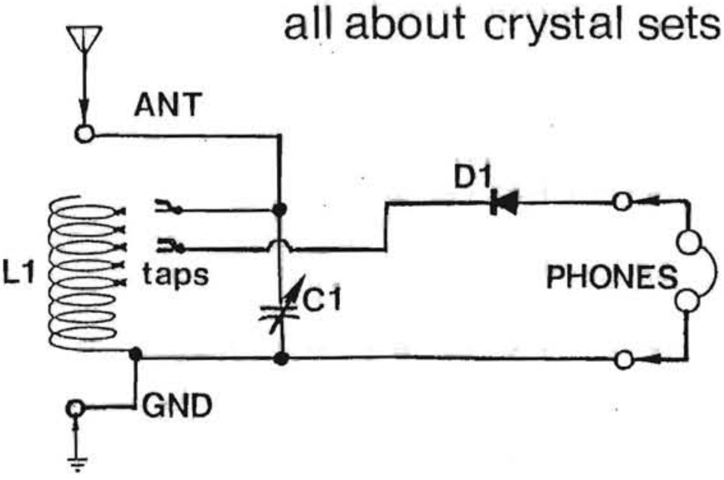
ANT
D1
L1
taps
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector (or 1N34A or equiv. diode).
L1 - Spiderweb coil (see text).
4 - Fahnestock clips.
2 - small alligator clips
BASE - 8 X 6 X 1/2-inch section of wood (see text).
MTG. POST - 5 x 3/4 x 3/8-inch wood section (see text).
PHONES - 2000 ohms headphones.
MISC. - Coil form (see text), No. 22 enameled copper wire, solder lugs, hook-up wire, wood screws, machine screws, and nuts.

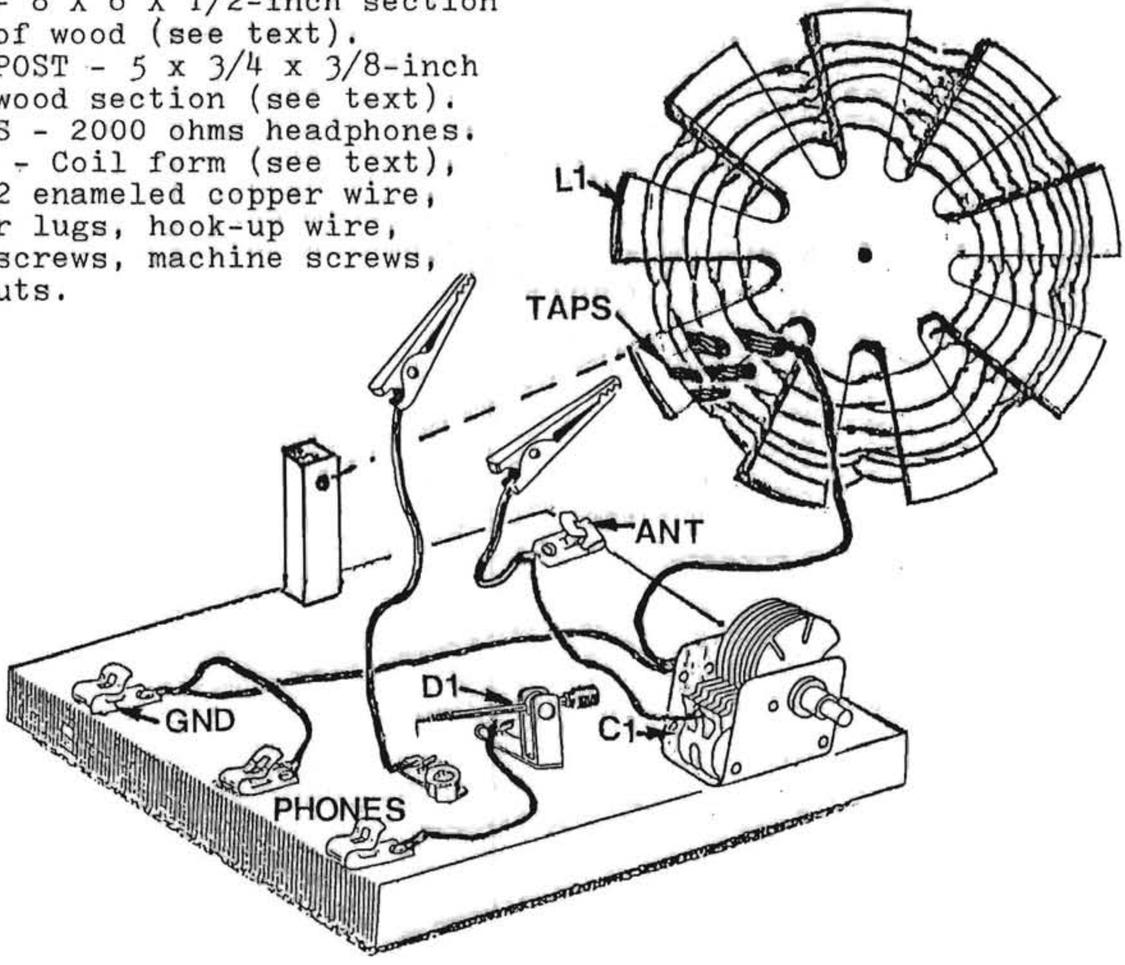
L1
TAPS
ANT
D1
C1
GND
PHONES

SPIDERWEB COIL CRYSTAL SET

SPIDERWEB COIL CRYSTAL SET

The spiderweb coil is a tuning coil that is wound on a flat wheel form that has an uneven amount of spokes. This type of coil was also used as a loop antenna in early radios.

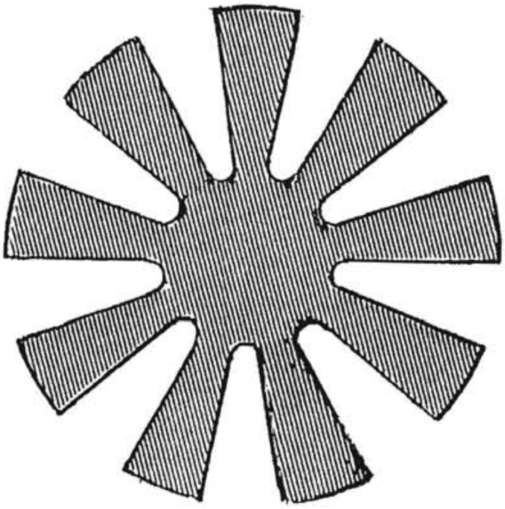

Our receiver uses a tapped spiderweb coil that is tuned by a variable capacitor. As shown in the illustration, the spiderweb coil is mounted at the rear of a 8-inch by 6-inch by 3/4-inch wood base. A catwhisker crystal detector and a variable capacitor are mounted near the front of the wood base.

SPIDERWEB

The spiderweb coil is wound on a 6-inch diameter flat form that has nine vanes. Make a tracing of the spiderweb coil form template on the next page. Then use carbon paper to transfer the tracing to a section of 1/8-inch fibreboard or sheet plastic. Cut the coil form out with a jeweler's saw, and drill a center hole to fit a small wood screw.

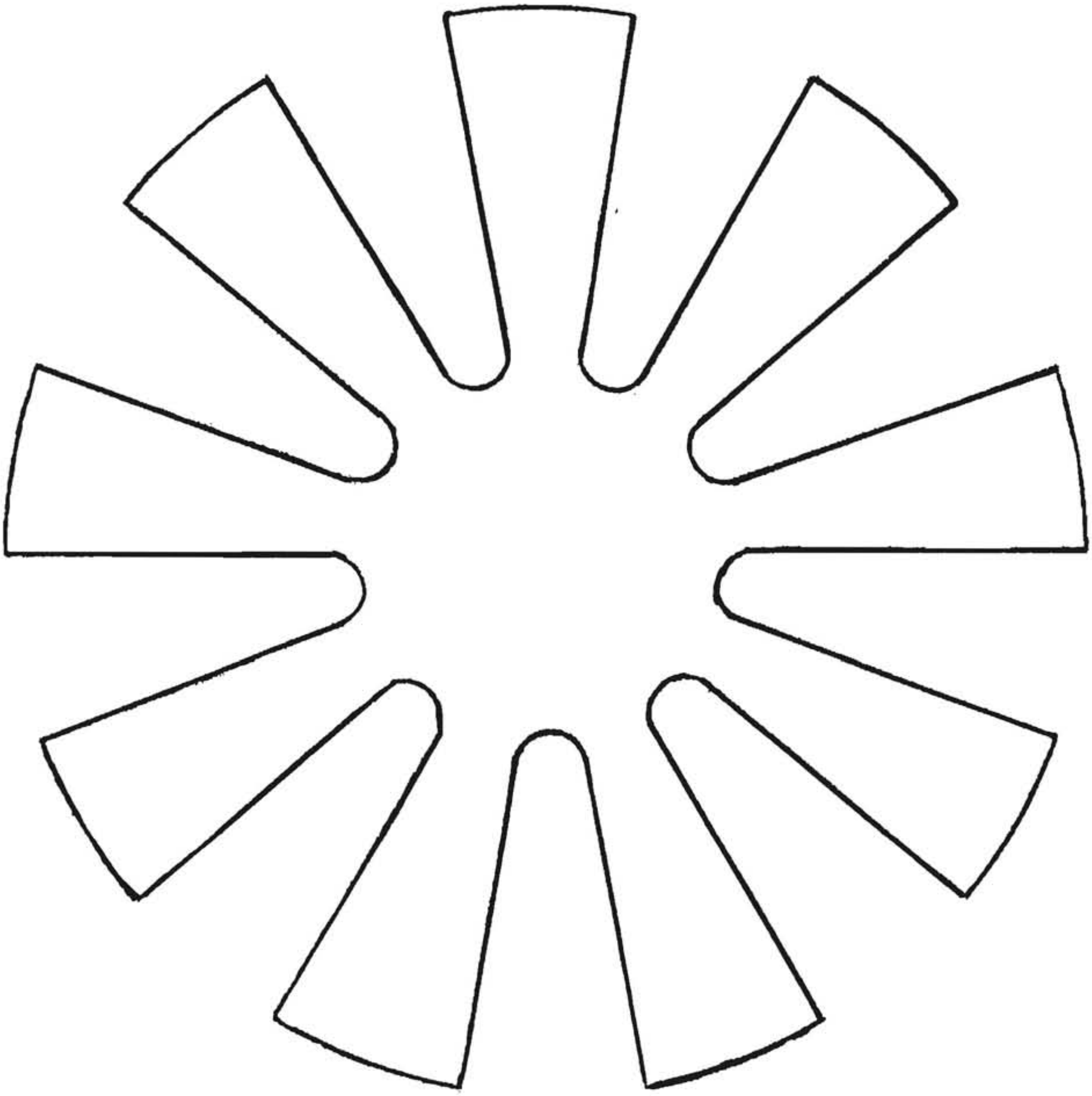

Spiderweb Coil Form Template

Wind the coil with No. 22 enameled copper wire, starting from the center and winding outward. Leave about 5-inches of the starting end of the wire free. It will be used as a connecting lead. Feed the wire behind a vane and over the front of the next vane as you wind. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Wind the wire to about 1/4-inch from the end of a form vane. Then drill a small hole near the vane end, and bend a solder lug through it. Solder the wire end to the lug.

On one of the vanes, count down five wires and then carefully lift it outward. Remove about 1/4-inch of the enamel with a section of sandpaper and then bend the end of a solder lug aound the bared wire. Solder the lug to the wire, then cut off remaining portion of the lug (away from the bentover end). Repeat this every five turns until you have a total of six lug taps on the coil. Stagger the lugs away from each other to keep them from shorting.

BASEBOARD

As shown in the illustration, the Spiderweb Coil Crystal Set parts are mounted on a wood baseboard. Cut the 8-inch by 6-inch baseboard from a section of 1/2-inch thick wood. White pine was used in our model. Sand the surface of the wood after cutting.

Make the spiderweb coil mounting post by cutting out a 5-inch by 3/4-inch by 3/8-inch wood section. Then cut a notch in the center of the rear of the baseboard to fit the wood section. Install the wood section in the notch with small wood screws.

Mount the parts on the baseboard as shown in the illustration with small wood screws. If the mounting holes of the tuning capacitor C1 are on the bottom, drill matching holes in the baseboard and use flathead machine screws to hold the capacitor.

Install solder lugs with the four terminals, and mount the spiderweb coil to the mounting post with a wood screw through a center hole in the coil form. The coil should be positioned with the lugs facing forward (on the left side of the coil form).

WIRING

Connect the crystal set components with hookup wire as shown in the illustration. Use 6-inch lengths of insulated stranded wire to the two small alligator clips.

Experiment 1

Connect antenna, ground, and 2000-ohm headphones to the crystal set. Then connect the D1 alligater clip lead to the antenna terminal. Set the tuning capacitor C1 to minimum capacity (rotor blades all the way out). Connect the antenna clip lead to the tap on L1 near the center of the form.

Adjust the crystal detector D1 catwhisker to a sensitive spot and tune C1 for a radio station. It may be easier to tune for stations if you temporarily connect a crystal diode (1N34A or equiv.) in place of the catwhisker crystal detector.

After a station is heard, you can disconnect the diode and probe for a sensistive spot on the crystal with the catwhisker. Change the antenna clip lead to other L1 taps and continue tuning C1 for radio stations.

Experiment 2

Disconnect the D1 clip lead and reconnect it to a tap on L1 close to the center of the coil. This will load L1 more lightly than if the detector was connected to the top of the coil. Repeat the tuning procedure of Experiment 1 and note the difference in selectivity and volume of the received radio stations. Then tune in a radio station and move the D1 clip lead to other taps on L1. There will be a particular tap point for optimum selectivity and volume for each received station.

STATION LOG

DATE	TIME	DIAL	STATION	NOTES

3

ADVANCED RECEIVER PROJECTS

DUAL TUNED CRYSTAL SET

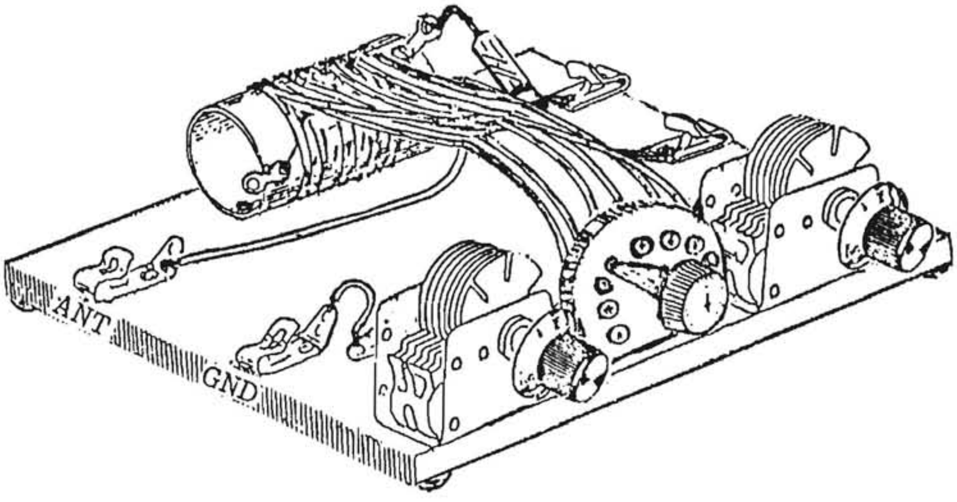

This crystal set circuit uses two variable capacitors C1 and C2. As you can see in the schematic drawing, C1 tunes the ground circuit, and C2 tunes a tapped coil L1. The taps of L1 are connected via the rotary switch S1.

TUNING COIL

Look at the drawing of the coil form construction details on page 34. The tuning coil L1 is wound on a section of 1/8-inch wall plastic tube, 2-inches in diameter, and approximately 4-inches long.

Drill holes in the form and install three solder lugs with machine screws and nuts. Position the solder lug ends towards the coil form ends as shown.

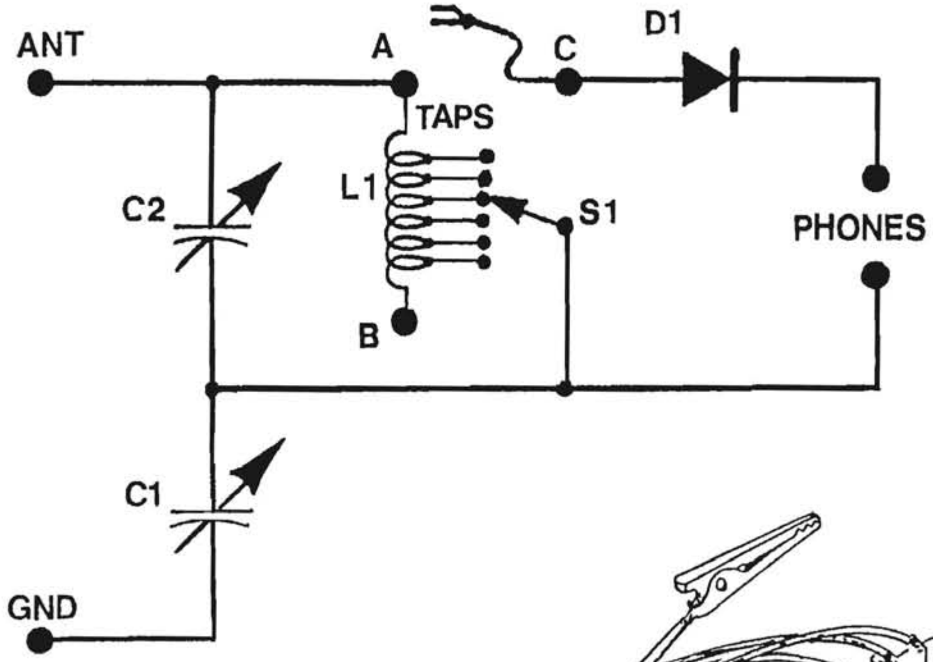
ANT
A
C
D1
TAPS
L1
C2
S1
PHONES
B
C1
GND

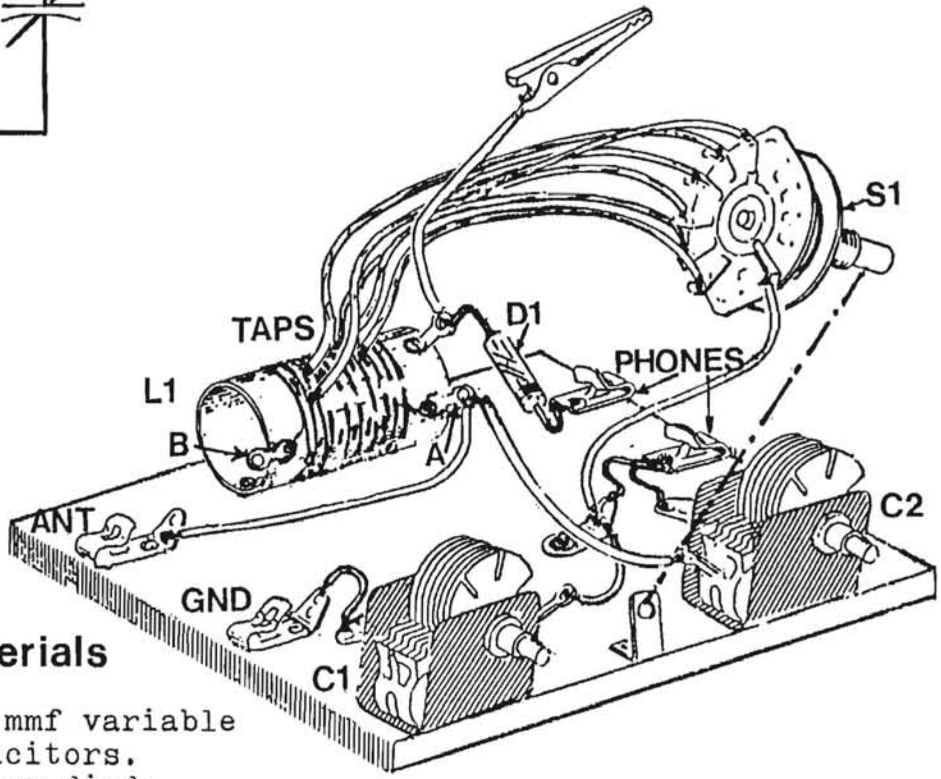
S1
TAPS
D1
PHONES
L1
B
A
ANT
C2
GND
C1

List of Materials

C1,C2 - 365 mmf variable capacitors.
D1 - Germanium diode (1N34A or equiv.).
L1 - Tuning coil (see text).
S1 - 1-pole, 10-position rotary switch (see text).
4 - Fahnestock clips.
1 - Small alligator clip.
BASEBOARD - 5 x 8 x 3/8-inch wood section (see text).
PHONES - 2000-ohm headphones.
MISC. - Coil form (see text), No. 24 enameled copper wire, solder lugs, hook-up wire, wood screws and nuts, knobs, bracket for S1.

DUAL TUNED CRYSTAL SET

The coil L1 is wound with No. 24 enameled copper wire, and is tapped every five turns for a total winding of 75 turns.

Drill a small hole in the coil form approximately 1 1/4-inch from the "A" lug. Remove the enamel from the end of the No. 24 wire and then feed the wire end through the hole. Solder the wire to lug "A".

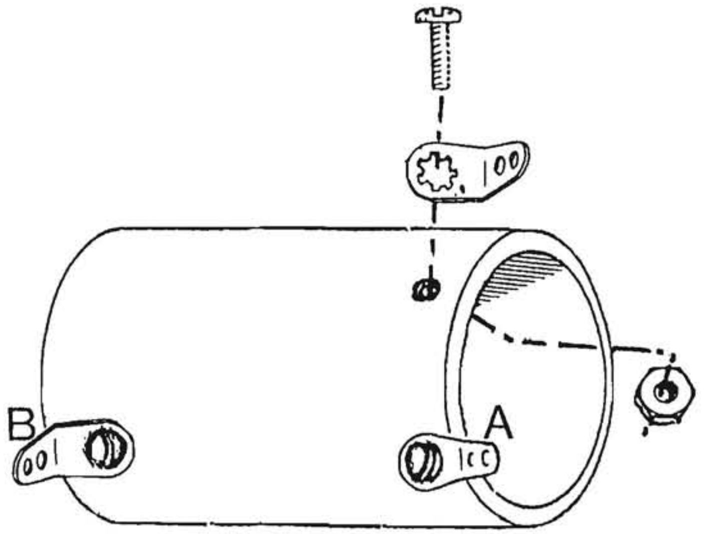

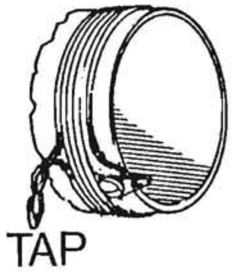

Wind five closely spaced tight turns around the coil form. Then bend back and twist together 1/2-inch of the wire to make a tap. Position the tap outward and then continue winding the coil, keeping the turns as tight and close together as possible. Make a tap every five turns and then drill a small hole in the coil form after the 75th turn. Feed the wire through the hole, remove enamel, and solder the end to lug "B".

BASEBOARD

The Dual-Tuned Crystal Set is built on a wood baseboard as shown in the illustration. Cut the 5-inch by 8-inch baseboard from a section of 1/4-inch thick plywood. Sand the edges and surface of the baseboard to remove any splinters. Drill holes in the baseboard and install the parts with small machine screws and nuts in the locations shown in the illustration. Drill mounting holes at each end of the tuning coil L1 and position it on the rear of the baseboard with the coil taps toward the front. The illustration shows a commercial rotary switch, but a wooden rotary switch can be used for S1 instead. (See page 53 for construction details.)

Install the four fahnestock clips (used as antenna, ground and headphone terminals) with solder lugs. Mount an additional solder lug on the baseboard midway between S1 and L1.

WIRING

Connect the crystal set parts with hookup wire as shown in the pictorial illustration and the schematic diagram. Use the center solder lug as a junction point for leads from C1, C2 and S1 as shown in the illustration. Connect the leads from S1 to the taps on L1 in numerical sequence; every 5 turns for 7 taps. Then connect the last three leads every 10 taps. Insert the lead ends into the taps and hold them with a drop of solder. This will make it easier to move leads to other L1 taps. Use a 5-inch length of stranded hookup wire for the alligator clip lead connected to D1 at solder lug C.

Experiment 1

Connect 2000-ohm headphones, antenna and ground leads, to the crystal set terminals. Then connect the D1 clip lead to the L1 "A" solder lug.

Set S1 to each tap position in turn and tune C1 and C2 for received radio stations. If all of the local stations can not be heard, experiment with different connections to the L1 taps. More turns may have to be added to L1 (with additional five-turn taps) to receive stations at the lower frequency end of the broadcast band.

Experiment 2

Remove the D1 clip lead from "A" and then connect it to the "B" solder lug on L1. Repeat the tuning method of Experiment 1 and note if the crystal set is more selective. This is possible, because L1 is now connected as an rf transformer. You can also connect the D1 clip lead to each of the S1 taps and note if there has been any improvement in reception.

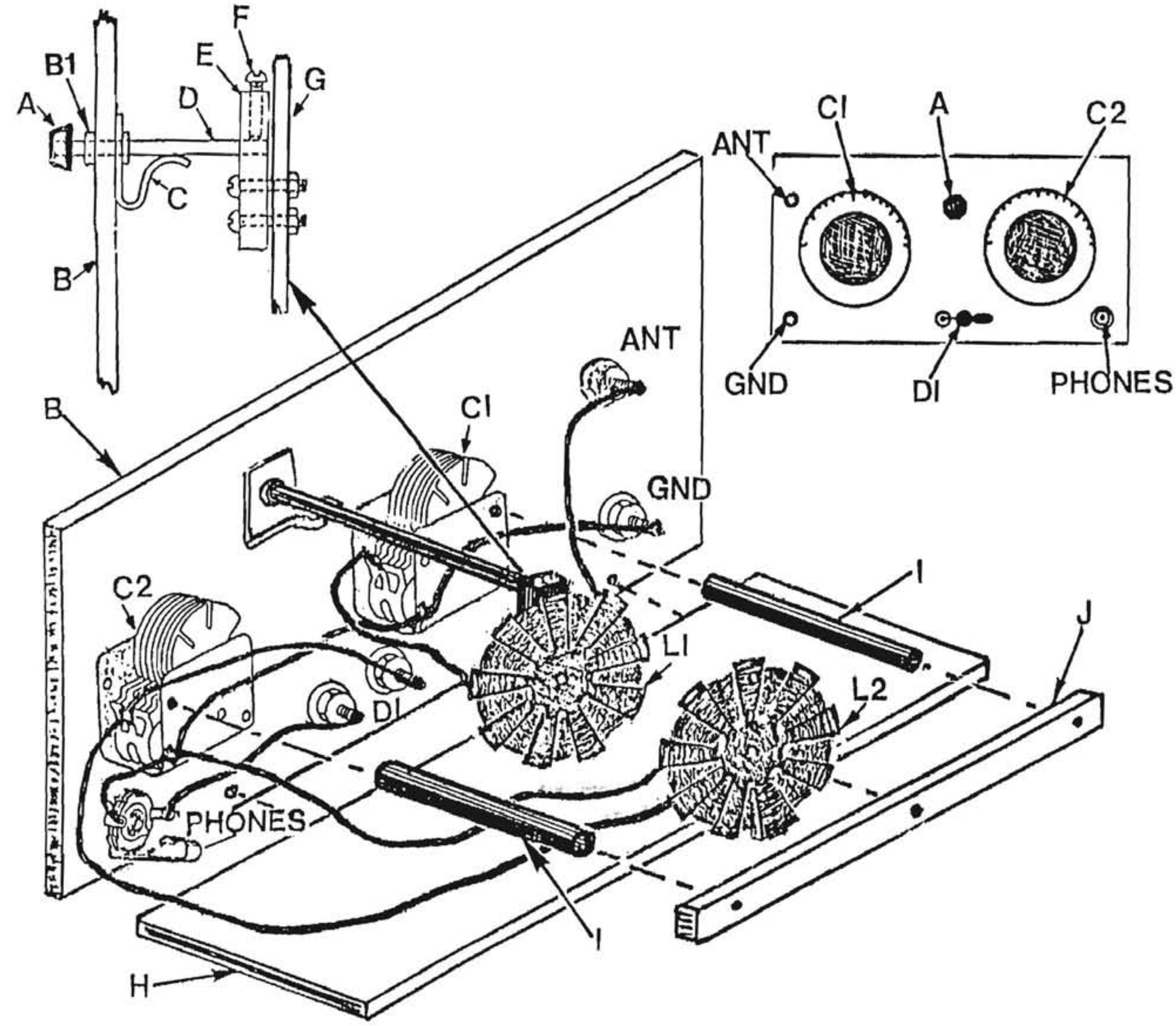
F
B1
E
G
A
D
C
B
ANT
CI
A
C2
GND
DI
PHONES
B
ANT
CI
GND
C2
I
J
LI
L2
DI
PHONES
I
H

SELECTIVE CRYSTAL SET

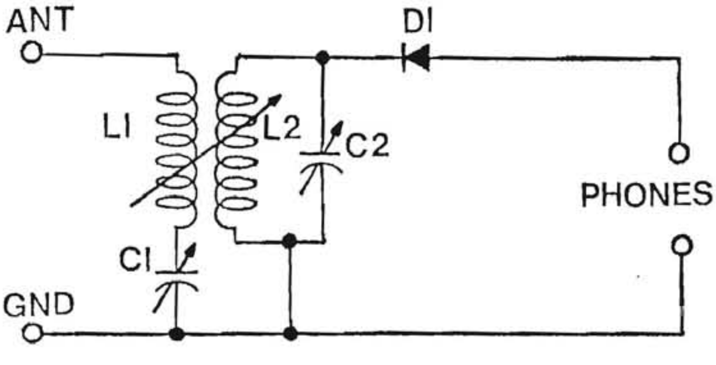
ANT
DI
LI
L2
C2
PHONES
CI
GND

SELECTIVE CRYSTAL SET

This receiver uses two spiderweb coils in a two-circuit tuner for good selectivity and sensitivity. As you can see in the schematic diagram, the antenna-ground rf input circuit is tuned by L1-C1 for maximum signal gain. L1 acts as the primary winding of an rf transformer, with L2-C2 as the tuned secondary. Coupling between L1 and L2 is variable for best selectivity. L2 is connected to a catwhisker crystal detector and the audio signals are fed through the phone jack to 2000-ohm headphones.

As shown in the illustration, the crystal set parts are mounted on the rear of a black plastic panel. This type of layout is similar to the type of construction of the old-time crystal sets.

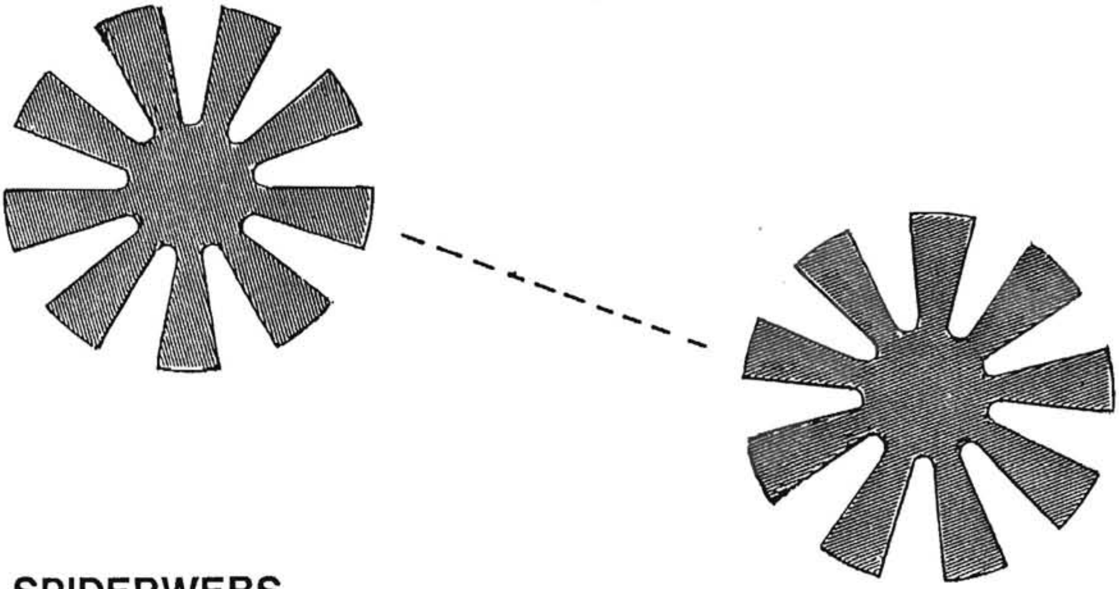

SPIDERWEBS

Both L1 and L2 spiderweb coils are made exactly the same way. L1 and L2 are wound on 17-vane coil forms 3 5/8-inches in diameter. The coil forms in our receiver are made from 1/16-inch sheet plastic of the same type used for printed circuit boards (but without the copper foil). Any kind of rigid sheet plastic should be suitable for the coil forms.

Make a tracing of the spiderweb coil template. Then use carbon paper to transfer the tracing on a section of 1/16-inch thick sheet plastic. Cut out the vanes with a jeweler's saw. Drill two holes as shown, and install the two solder lugs.

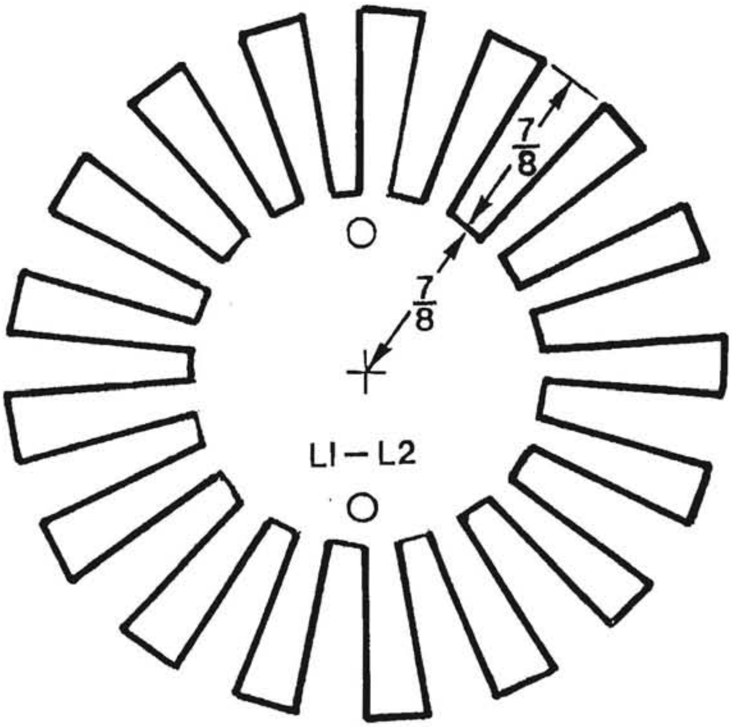

COIL FORM TEMPLATE FOR L1–L2

Wind as much No.22 enameled wire as possible on each coil form. Solder the wire to a solder lug before winding a coil. Then wind the wire over a vane and under the next vane. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Solder the wire end to the remaining solder lug

CONSTRUCTION

The panel used in our receiver is a 7 3/4-inch by 4 1/2-inch section of 1/4-inch thick black acrylic plastic. The size is not critical, but it should be large enough to mount all of the parts as shown in the illustration.

Layout and drill the mounting holes of the components on the panel. Make sure that C1 and C2 are positioned so that the edges of the tuning knob dials are more than 1/2-inch apart. Install the catwhisker crystal detector assembly, antenna and ground terminals, and the phone jack in the locations shown.

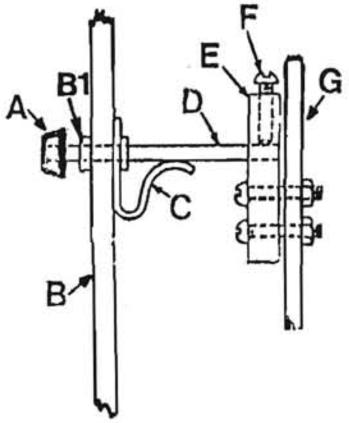

As you can see in the illustration, L1 is installed on the end of a variable coupling metal rod. This 1/4-inch diameter rod is 3-inches long with a flat side. Drill and tap a small hole close to the end of the rod. The rod used in our model came from a discarded volume control. Cut a 2-inch long by 1/2-inch wide by 3/8-inch thick wood section and drill holes to fit the rod, two mounting screws, and a hole for the rod set screw. Locate and mount the wood section in the center of a spiderweb coil L1 so that the rod will be 3/4-inch down from the edge of the top vane.

Hold L1 behind the panel and using the free end of the rod, locate the panel hole. Position L1 midway between C1 and C2, and make sure that the top edge of the coil is below the top of the panel. Drill a hole to fit the rod bushing in the panel hole location. The panel bushing in our model was removed from a discarded volume control.

Drill a hole to fit the bushing in one end of a 2-inch by 1/2-inch section of sheet aluminum. Bend the other end of the section to fit against the flat side of the rod. Install the metal section and the bushing in the panel hole.

Insert the rod into the bushing and make sure that the metal section end fits close to the flat side to prevent rotation of the rod. Install a knob on the rod end and make sure that the rod can be pushed in and out easily. If necessary, put a small amount of grease on the flat side of the rod.

Install the remaining spiderweb coil L2 on a wood section that is mounted on the rear of the panel with two spacers. The spacers should be long enough for L2 to be located 1/4-inch behind L1 with the rod pushed all the way in. The 5 1/2-inch by 1-inch by 1/4-inch wood section in our model is supported with screws installed through 1 1/2-inch metal spacers into tapped holes in the rear of C1 and C2. The screws and bushings can be made longer and mount into panel holes in your receiver.

Install a wood section on the bottom rear of the panel to keep the receiver upright. A 5-inch by 3-inch section of 1/4-inch thick wood is used in our receiver.

SET WIRING

Connect the receiver components with hookup wire as shown in the illustration and schematic diagram. Use flexible stranded wire for connections to L1. Make sure that the L1 leads do not touch any component as the variable coupling rod is moved in and out.

Install knobs on C1, C2, and the variable coupling rod. Then place the crystal in the detector assembly.

OPERATION

Connect antenna and ground leads to the receiver terminals. Plug in a set of 2000-ohm headphones into the panel phone jack. For easier initial operation of the receiver, temporarily connect a germanium diode (1N34A or equivalent) across the D1 connecting lugs. Make sure that the catwhisker is not touching the crystal at this time.

Push the variable coupling rod half way in and tune C1 and C2 for a station. Adjust the variable coupling rod and retune C1-C2 for best reception. Check operation over the entire broadcast band. It may be necessary to add or subtract turns of wire on L1-L2 for complete coverage. After checking operation with the germanium diode, remove it and try using the catwhisker crystal detector.

List of Materials

C1,C2 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector assembly.
L1,L2 - Spiderweb coils (see text).
2 - Terminals.
1 - Phone jack.
PHONES - 2000-ohm headphones.
MISC.- Plastic sheet for spiderweb coils, solder lugs, machine screws and nuts, wood screws, hookup wire.
CONSTRUCTION MATERIALS (See Text).
A - knob
B - Panel
B1- Bushing
C - Aluminum section
D - Rod
E - Wood section
F - Setscrew
G - L1 coil form
H - Wood section
I - Spacers (2)
J - Wood section

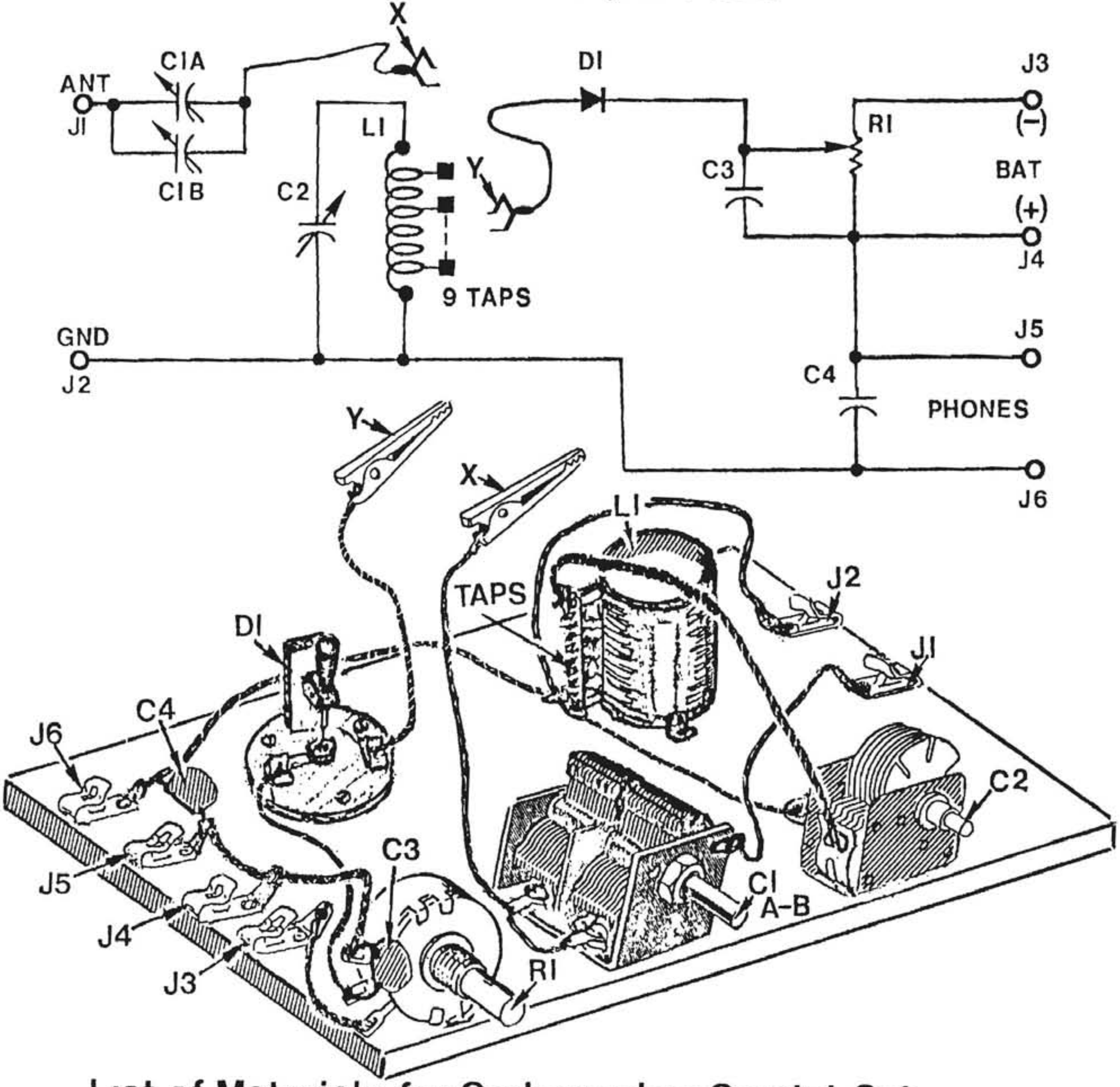
X
C1A
ANT
J1
C1B
C2
L1
Y
9 TAPS
D1
C3
R1
J3
(−)
BAT
(+)
J4
J5
C4
PHONES
J6
GND
J2
Y
X
L1
TAPS
J2
J1
D1
C4
J6
C2
C3
J5
C1
A-B
J4
R1
J3

List of Materials for Carborundum Crystal Set

C1A-C1B - Dual 365 mmf variable capacitor.
C2 - 365 mmf variable capacitor.
C3 - 0.1 mf capacitor.
C4 - 1000 mmf capacitor.
D1 - Carborundum crystal and catwhisker detector assembly (see text).
J1 to J6 - Fahnestock clips.
L1 - Tuning coil (see text).
R1 - 5,000-ohm potentiometer (linear taper).
BAT - 6-volt battery
PHONES - 2000-ohm headphones.
MISC. Coil form (see text), No.24 enameled copper wire, 1/2-inch long spacers, perfboard strip (see text), push-in clips, solder lugs, knobs, hook-up wire, 1N34A diode (or equiv.), wood for base and D1 detector assembly (see text), wood screws, machine screws and nuts, alligator clips X-Y, brackets for L1 and R1.

CARBORUNDUM CRYSTAL SET

Materials other than galena, silicon or germanium have the property of detection. In the beginning of radio, carborundum (the name given to a compound of silicon carbide) was also used as a detector. The required heavy catwhisker pressure made it suitable for the early radio stations on ships, as it did not fall out of adjustment. It would also handle the strong signal energy from nearby spark transmitters without burning out. But what makes this type of crystal detector different, is the necessity for a bias battery.

A crystal detector has a high current flow with voltage applied so that it conducts in the forward direction (catwhisker to crystal), and a very low current flow in the reverse direction. The amount of current flow in the forward direction depends upon the characteristics of the crystal material and the applied forward voltage.

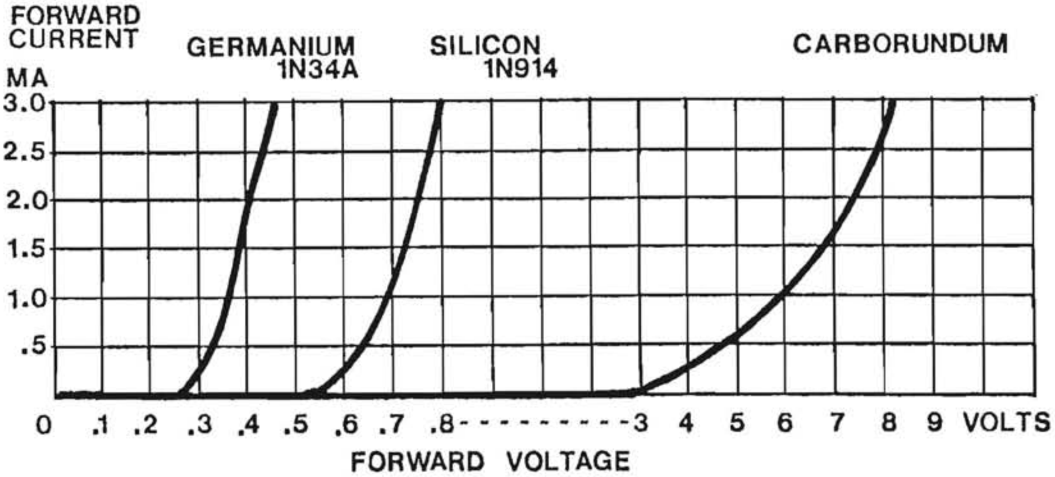

TYPICAL CRYSTAL DIODE CURRENT

The graph compares carborundum with germanium and silicon crystal detectors. Germanium minimum conduction voltage is about 0.3 V. Silicon is 0.6 V, and carborundum is 3 V. The high carborundum minimum conduction voltage is the reason a bias battery is used to move the threshold down so that weak radio signal voltages can be detected.

TUNING COIL

As shown in the illustration, the tuning coil L1 is wound on a cardboard (or plastic) mailing tube section 2-inches in diameter and 2 3/4-inches long.

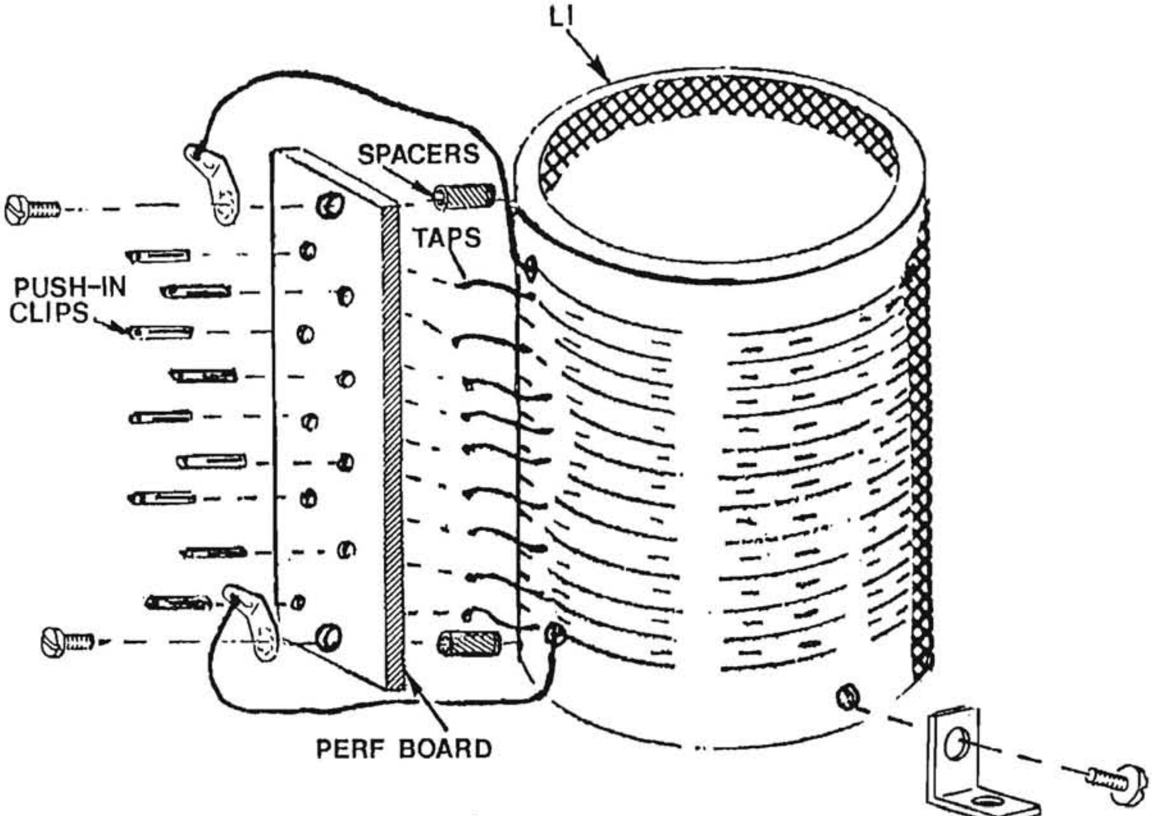

TUNING COIL ASSEMBLY

Locate and drill a hole 1/4-inch from each coil form end for machine screws to fit the two spacers. Also drill holes for two mounting brackets at one end. Then drill a small hole near each spacer hole to feed the wire on the form before you start winding.

Wind the coil with No. 24 enameled copper wire. Tap the coil every 10 turns for a total of 9 taps. An easy way to make the taps is to twist the wire together for a half inch and position the free end out. Make all the coil taps in the area between the two spacer holes. Thread the free end of the wire through the small hole at the end of the form near the spacer hole. Make sure that all of the turns are tightly wound and close together. Use a section of sandpaper to remove the enamel from the tap wire ends. Then tin the ends.

Mount 9 push-in clips in a 1/2-inch by 2 1/4-inch perf board section. Locate and drill holes on the board and install it on the coil form with two solder lugs, 1/2-inch spacers, machine screws, and nuts. Solder the coil taps to the ends of the push-in clips. Connect the coil leads to the solder lugs and mount the two 1/2-inch brackets on the coil form.

DETECTOR

The crystal detector assembly is vertical, (as shown in the illustration) instead of horizontal as in most catwhisker detectors. Begin construction by cutting the 2 1/2-inch diameter base and 2-inch by 4-inch vertical section from 1/4-inch thick wood.

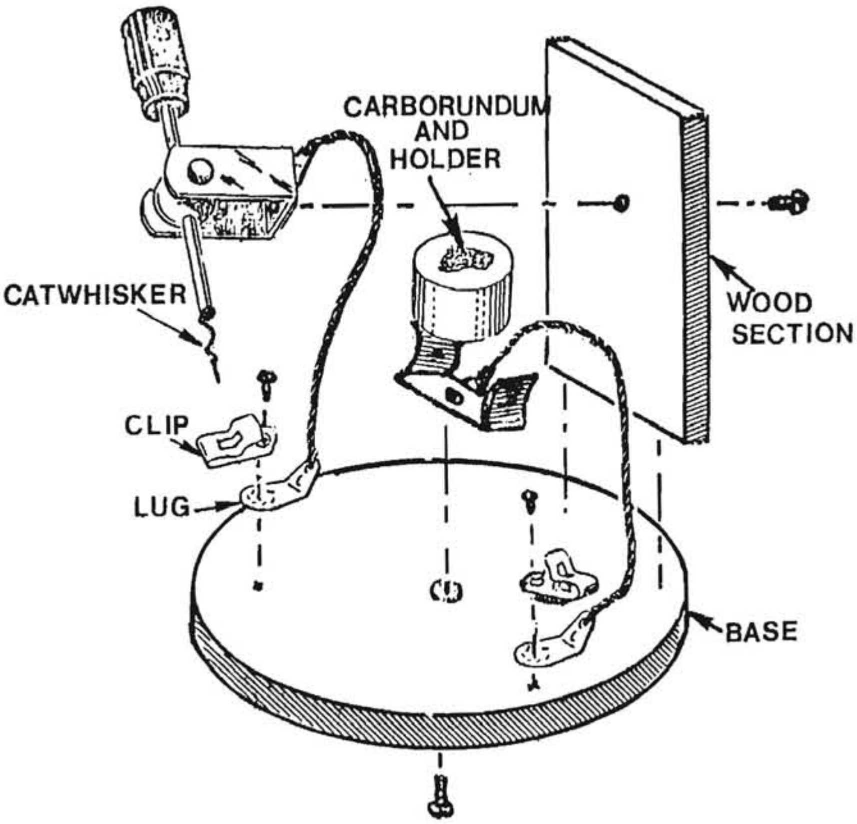

DETECTOR ASSEMBLY

Locate and install the two fahnestock clips on the base with solder lugs and small wood screws. Mount the vertical section on the edge of the base with wood screws (or cement). Locate and install the catwhisker holder with a solder lug, machine screw and nut near the top of the vertical section.

Temporarily position the crystal holder on the base so that the catwhisker will touch the center of the crystal. Locate and mount the crystal holder in this position with a solder lug and machine screw and nut. Connect the detector assembly with hookup wire as shown in the illustration. Solder a short section of coiled steel spring wire to the end of the positioning rod to serve as a catwhisker. Cut the end of the wire diagonally to give it a sharp point.

BASEBOARD

The receiver parts are installed on a 8-inch by 7-inch by 3/4-inch wood base. Wherever poossible, small wood screws are used to hold the components on the base.

Begin construction by laying out and installing the parts in the positions shown in the illustration. It will be necessary to make a metal mounting bracket to fit R1. C1A-B and C2 are mounted with machine screws in the tuning capacitor frames through countersunk holes in the base. Install solder lugs in the rear frame of both tuning capacitors. Also mount solder lugs with all of the terminals. Install knobs on the tuning capacitors and R1.

WIRING

Use hookup wire to connect the crystal set parts as shown in the schematic and illustration. Connect 5-inch leads of flexible stranded insulated wire to each alligator clip. Solder one lead to the stator lugs of C1A-B and connect the other lead to D1. Connect C3 and C4 as shown and keep their leads as short and direct as possible.

Experiment 1

Connect your antenna to J1, ground to J2, and 2000-ohm headphones to J5 and J6. Temporarily connect a germanium diode (1N34A or equiv.) in place of the carborundum crystal detector D1. The diode will make it easier to check out the receiver.

Do not attach the battery
to J3-J4 at this time, or
the diode may be damaged.

Connect **X** C1A-B clip lead to the top tap of L1, and **Y** D1 clip lead to the second tap from the top. Tune in a radio station with C2 and then adjust C1A-B for best volume. Change the position of the clip leads and retune C1A-B and C2 for best reception. Check operation of the receiver over the entire broadcast band.

Experiment 2

Tune the receiver to a radio station and then disconnect the germanium diode. Do not disturb the receiver tuning. Install a carborundum crystal into the D1 assembly. Set R1 to mid range, and connect a 6-volt battery to the receiver; negative lead to J3, positive lead to J4.

Adjust the D1 catwhisker for a sensitive spot on the carborundum crystal. Then set R1 for best volume of received signal. Retune C1A-B and C2 and change the clip leads on the L1 taps as necessary for reception of radio stations.

Note

Carborundum that will work as a detector may be hard to find. Grinding wheels are not suitable as a source, because the carborundum is usually in fine particles mixed with other material. Best bet is to contact collectors of antique radio parts. But you can use this receiver to experiment with minerals for semiconductor operation. See the Workshop Section for information on mounting mineral sections for detector experiments.

4 COMPONENT PROJECTS

BOOK TUNING CAPACITOR

During the roaring 20's,the Crosly Radio Corporation used a type of tuning capacitor that had two plates hinged together on one edge. The dielectric was a sheet of mica between the two plates. This type of tuning capacitor was called a Book Condenser.

The tuning capacitor shown in the illustration is similar to the Crosly Book Condenser. It uses aluminum foil plates cemented to fixed and movable wood sections, and a plastic sheet dielectric (cut from thin transparent plastic notebook protector material).

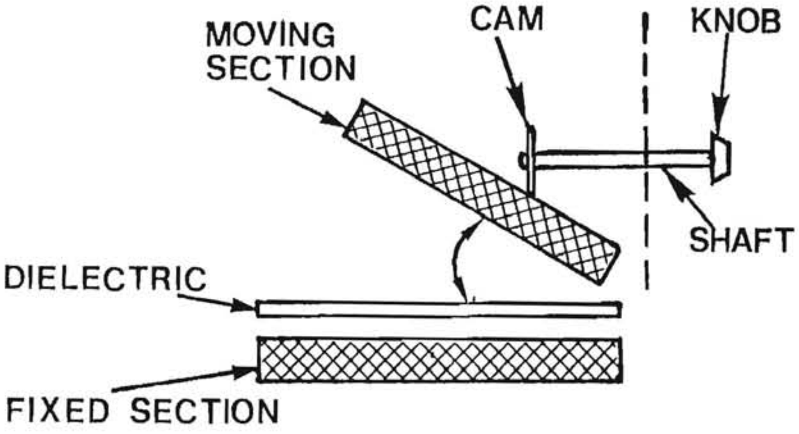

Adjusting the tuning knob causes a metal cam to rotate and control the movement of a spring loaded wood section. This varies the capacity between the two foil plates. As the plates are brought closer together, the capacity increases. As they are separated, the capacity decreases.

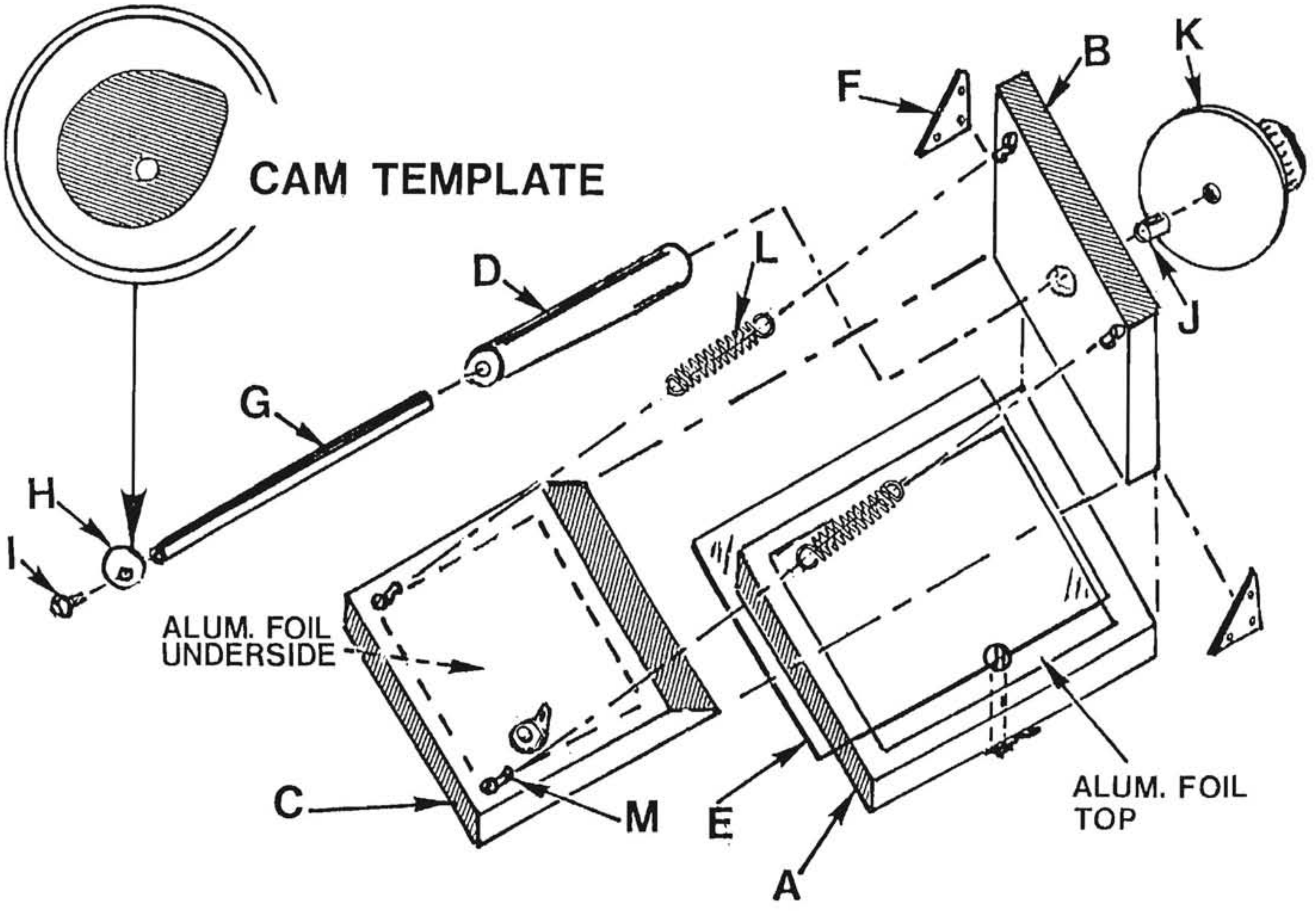

BOOK TUNING CAPACITOR

List of Materials

A - 3 x 3 1/2 X 3/8-inch wood section
B - 3 x 1 3/4 x 3/8-inch wood section
C - 3 x 3 1/8 x 3/8-inch wood section (one end beveled 45 degrees - see text).
D - 3/4 - inch diameter wood dowel, 2 1/4 inches long.
E - 3 1/2 x 3 -inch sheet plastic (See text).
F - 2 triangular 1-inch metal sections
G - 1/4-inch metal rod, 3 -inches long.
H - Metal cam (see text).
I - Machine screw (see text)
J - Bushing 3/8-inch long (to fit rod G).
K - Tuning knob (to fit rod G).
L - 2 springs 1/4-inch diameter by 1 1/4-inch long.
M - 4 solder lugs.
MISC. Small wood screws, nails, aluminum foil, machine screws, nuts, solder lugs and washers.

CONSTRUCTION

Start construction by cutting the wood sections from 3/8-thick wood to the sizes shown in the drawing. Section C has a 45 degree bevel at one end. This acts as a pivot point for the movable section. Drill and countersink a hole to fit a flathead machine screw in the places shown on sections A and C. Do not install the screws at this time.

Cement aluminum foil on the top of section A and on the bottom of section C. Then cut and remove a 1/2-inch strip of foil from the front edge of section A. Cut a 1/8-inch strip from the remaining edges. Also cut away a 1/8-inch strip from all the edges of the foil on section C.

Carefully slit the foil on top of the machine screw hole on section A and press the foil gently inside the hole. Install the flathead machine screw into the hole, making sure that it is in contact with the foil. Fasten the screw with a washer and nut. Make sure that the screw does not turn and rip the foil. Install a solder lug and nut on the free end of the screw. Repeat this operation with section C.

Cut the sheet plastic E to size and place it on top of section A. Then place section B on top of the plastic sheet and install it on the end of section A with wood screws. Install the two metal sections F with small nails.

Saw the metal rod G to size and tap one end to fit a machine screw. Cut the dowel D to size and center drill it to fit the rod G. Cut a flat on one side of the dowel.

Make a tracing of the cam H template and cement it on a section of heavy gauge sheet aluminumm. Cut the cam to size and drill a hole to fit the rod machine screw in the location shown. Install the cam with a lock washer on rod G. Insert the free end of the rod into dowel D. Place section C on top of the plastic sheet E on section A. Then temporarily position dowel D with rod G in the center of section B. Rotate the cam so that its long end touches the top of section C. Mark the placement of the dowel on section B. Drill a hole in section B to fit dowel D. Cement the dowel in place.

Locate and install solder lugs with wood screws on sections A and C as shown in the drawing. Bend over the ends of the solder lugs and install two springs L in the lugs.

Mount tuning knob K and bushing J on the free end of rod G. Adjust the knob and check that section C moves in and out with the cam rotation. changing the capacity between the foil plates.

You may have to experiment with the cam diameter for best operation of section C. File the cam edges in small increments only and recheck the operation often. Use very light springs to keep section C from pressing against the cam and moving it.

Reverse the cam on rod G to alter the direction of capacity (clockwise or counterclockwise). The capacitor can be installed vertically on a receiver project with a small metal bracket mounted on the bottom side of section A.

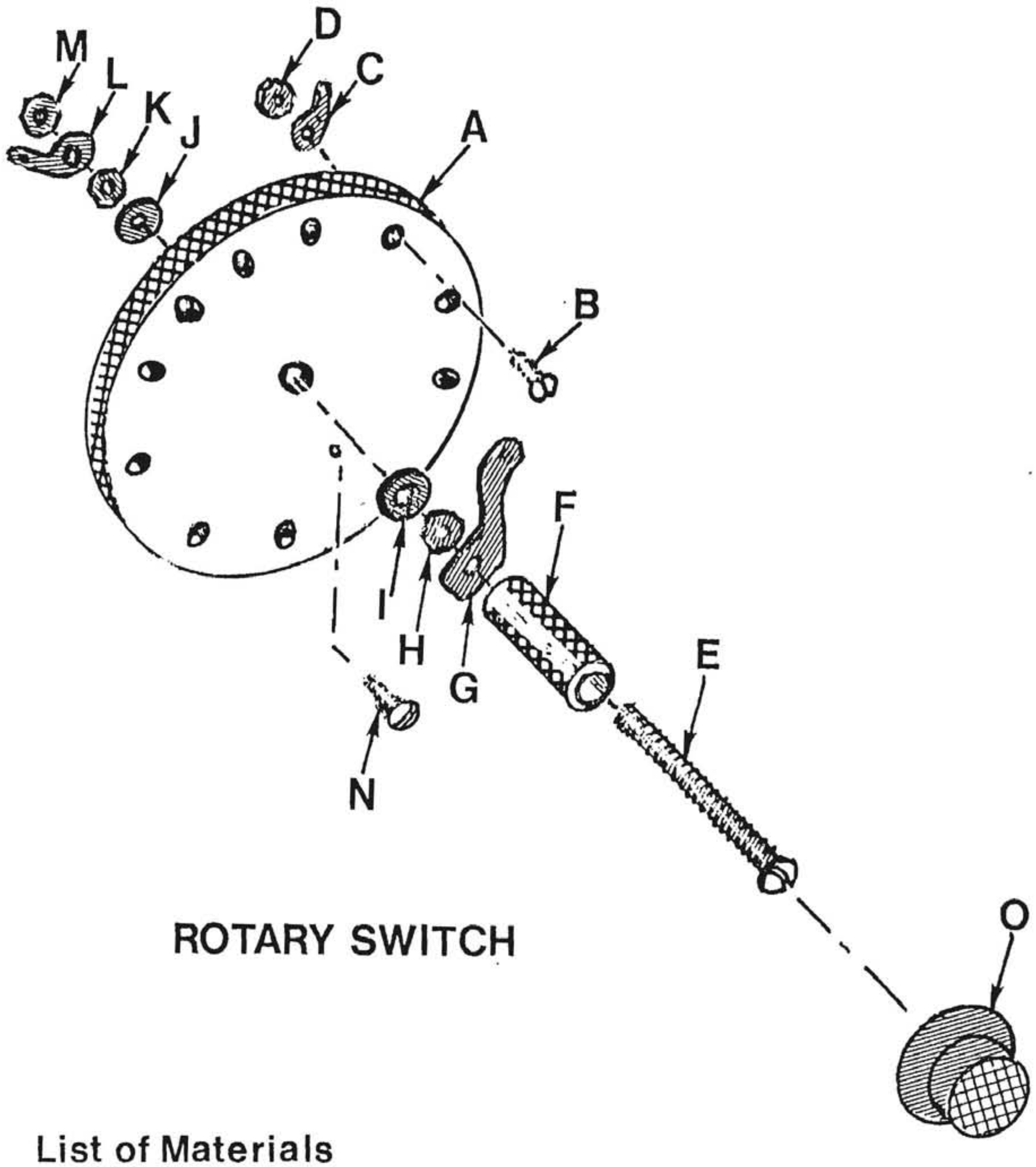

ROTARY SWITCH

List of Materials

A - 2 1/4 diameter x 3/8-inch wood section.
B - Flat head machine screws.
C,L - Solder lugs.
D - Nuts to fit B.
E - Machine screw 1 1/4-inches long.
F - 1/4 diameter x 1/2-inch long bushing (to fit E).
G - Switch arm 1 x 1/4-inch (sheet brass, see text).
H,K,M - Nuts to fit E.
I,J - Washers to fit B.
N - Small woodscrew.

ROTARY SWITCH

In the early days, radio experimenters could not buy parts that are readily available today. They had to build their radio components from wood sections and common hardware.The rotary switch ilustrated is similar to a type constructed for use with tapped coils of crystal sets.

The rotary switch shown uses flat head screws for contact points that are mounted on a wood section. The switch arm is cut from sheet brass and is installed on a machine screw.

CONSTRUCTION

Refer to the drawing of the rotary switch. Begin construction by cutting the 2 1/4-inch diameter section A from 3/8-inch thick wood. Locate, drill,and countersink the holes for the flathead machine screws B. The rotary switch shown has 10 machine screws, but you can space the hole locations for the number of switch contact points desired. Allow a space between the starting and ending contact points for the woodscrew N. This woodscrew is used as a stop for the rotary arm G. Locate and drill a center hole for the machine screw E. Install the screws B in section A with solder lugs C and nuts D. Install the woodscrew N.

Measure and cut the rotary arm G from a section of sheet brass. The arm should be long enough to reach the contacts. Make sure that it is not too wide, or it may short out to a nearby contact. Make a dent or "dimple" on the contact end, and drill a hole for screw E on the other end. Bend the rotary arm as shown in the drawing. Install screw E into bushing F, and arm G, and tighten the assembly with nut H. Place washer I on the screw and install screw E into the center hole of A. loosely fasten the screw with washer J and nut K. Then rotate the bushing F and arm G. Adjust the tightness of nut K as required to allow rotation of G with only a slight "drag". Install solder lug L on screw E with nut M, and bend the lug towards the screw end. Install knob O on bushing F, and recheck operation of the arm G. It should touch all of the contacts B and stop at screw N.

The rotary switch is used as shown in the Dual-Tuned Crystal Set Construction Project. Make sure that extra lead length is allowed for connection to the solder lug L. Bend the lead into a "pig-tail," so that it will not break as the solder lug L is rotated.

5 WORKSHOP

The construction projects in this book can be built with common hand tools found in an average home workshop.

Useful Hand Tools

Safety glasses or face shield
Steel rule
Square
Hack saw
Wood saw
Screwdrivers
Pliers
Wirecutters
Tin snips
Center punch
Files
Hammer
Soldering Iron

Wear safety glasses or face shield while working. Observe common sense safety precautions.

SOLDERING

Make sure that the part to be soldered is clean. Use only rosin core solder. Prepare the soldering iron tip as follows:

File each surface of the tip to remove oxidation. Plug in the iron and let it heat until you can melt solder on the tip (before the copper tip turns blue or bronze color). Melt enough solder until all sides of the tip are coated. Then wipe off excess solder with a damp sponge or cloth.

Make sure that the bare wire lead to be soldered is free of oxide. It is best to tin it first by melting a bit of rosin core solder on the lead end. Crimp the lead in the part hole to keep it from moving during soldering. Hold the iron tip under the part until it is hot enough to melt solder. Apply only enough rosin core solder to cover the surface of the part. Remove the iron. Do not move the part or wire while it is cooling.

PROJECT CONSTRUCTION TIPS

Tape a sheet of graph paper over a project wood section. Mark locations of the parts on the paper. Center punch mounting hole centers through the paper. Use a compass to draw circles the same size as the holes. Then drill holes through the paper.

Lettering on the projects can be done with a labelmaker that prints on strips of adhesive coated plastic. You can also use decals or dry transfer lettering sheets.

Small metal brackets can be made from heavy gauge sheet aluminum. Layout and drill the mounting holes and then cut it to size with a tin snips. Form the bracket by using a heavy mallet and a vise.

Tuning Capacitors, Knobs, and Dials

Old radios are a good source of components. The radios generally use two tuning capacitor sections ganged together. The receivers previously described will operate with one of the capacitor sections (365 mmf). You can cut out cardboard discs and cement them to radio knobs and use them as dials with the tuning capacitors. Mark the dials with dry transfer lettering, or black drawing ink.

Panels and Cabinets

The layout and construction of crystal sets is not critical. They can be fitted into many different types of wood cabinets. You can use almost any type of non-metallic box, but metal boxes are not recommended as sheet metal close to the tuning coil may lower its "Q". Black sheet plastic used for the front panel with white dry transfer lettering will give your receiver project an antique look.

MOUNTING MINERALS FOR CRYSTAL DETECTORS

Minerals (galena, etc) for experimentation as crystal detectors, will work best when mounted in a metallic base. For temporary mounting (while testing), wrap aluminum foil tightly around the sides and bottom of the mineral. Then place it into a sheet metal ring or cup that will fit into the detector assembly.

After testing, the mineral can be permanently mounted in a metal slug. Drill a shallow hole of a diameter equal to the detector assembly crystal in a block of wood. Fill the hole with melted solder. While the solder is still molten, carefully place your mineral section into the center. Hold the mineral in place with tweezers or long nose pliers until the solder cools. Then pry the metal slug out of the wood. Be sure that the top of the mineral section is free of solder.

INDEX

Advanced receiver projects,32
Amplifier connection,14,15
Antenna lead-in,10
Antenna,10

Book tuning capacitor,48
Boundary,5
Bias,5

Capacitor,8
Capacity,8
Carborundum crystal set,43
Carborundum,43
Catwhisker detector,4,6
Catwhisker,6
Component projects,48
Construction tips,55
Crystal detector assembly,45
Crystal detector,4

Detected audio signals,7
Detection,4,6
Diode electron flow,5
Dual spiderweb coils,37
Dual-tuned crystal set,32

Earphones,11
Enamel insulation,17
Electron current flow,5

Ferrite coil crystal set,23
Ferrite coil,12
Ferrite rod,23

Galena,4
Germanium,4
Ground lead,10
Ground wave,25
Ground,10

Hand tools,54
Headphones,7,11
Headset,11

Junction diode,4,5

Knobs and dials,55

Lettering,55
Loading coil,21
Loopstick,12

Modulated radio waves,6
Mounting minerals,56

N type material,4
Natural crystal,4

Operating tips,25

P type material,4
PN junction,5
Panels and cabinets,5
Parallel tuned circuit,9
Point contact diode,4,6
Primary coil,23

Reactance,8
Receiving radio waves,7
Resonant frequency,9
Rf coil,8
Rotary switch project,53

Schematic symbols,15
Secondary coil,23
Selective crystal set,37
Selectivity,9
Semiconductor diode,4
Silicon,4
Simple crystal set,12
Simple receivers,12
Sky wave,25
Slider tuned coil,17
Slider tuned set,17
Soldering,54
Space charge region,5
Spiderweb coil set,27
Station log,31

Tapped coil,32,44
Tapped spiderweb,27
Transmitted waves,6,7
Transmitting station,6
Tuned circuit,7,9
Tuning crystal set,7

Untuned detector,20

Variable capacitor,8
Variable coupling,39

Wave trap,21
Workshop,54

About the Author

Charles Green has been involved in the design and construction of electronic equipment for industry and the U.S. Government for 30 years. Much of his experience has been as a technical writer preparing operation and maintenance manuals. His continuing interest is in constructing radios and he has had over a hundred articles published in hobbyist magazines over the past two decades.

A sun-powered transistor amplifier makes this a—

Space-Age Crystal Set

By D. S. Halacy Jr.

LIGHT-POWERED RADIO plays indoors under artificial light, too. Performance is boosted by connecting antenna to TV aerial or telephone-dial finger stop; a short length of wire, with alligator clips at the ends, helps.

THIS tiny radio puts satellite technology in the palm of your hand. Like radios in artificial moons, it's powered by the sun—or any other light.

Basically, it's a 1960-model crystal set. A rugged sealed-in-glass diode has replaced the shaky catwhisker and galena crystal. The usually faint crystal-set signal is magnified many times by a transistor, in turn powered by a high-efficiency silicon solar cell.

This unit—and other parts—can probably be obtained locally for less than five dollars. All are standard parts, stocked by most retail and mail-order shops. In a pinch you can substitute the less-efficient International Rectifier B2M selenium cell for the quarter-round section of a Hoffman silicon solar battery.

First steps. Best place to begin construction is with the case for the radio—a small plastic box. You may find one in the medicine chest or at the druggist's. Start the necessary holes with a hot ice pick, then file or ream to size Keep try-

Author Halacy is an expert on the. sun. A couple of years ago he wrote Fabulous Fireball, *a book that tells the story of the sun and solar energy. Last fall. Macmillan also published his* Fun with the Sun. *a book describing how to make seven solar-energy projects at home. The present, article is an outgrowth of his research for the latter book.*

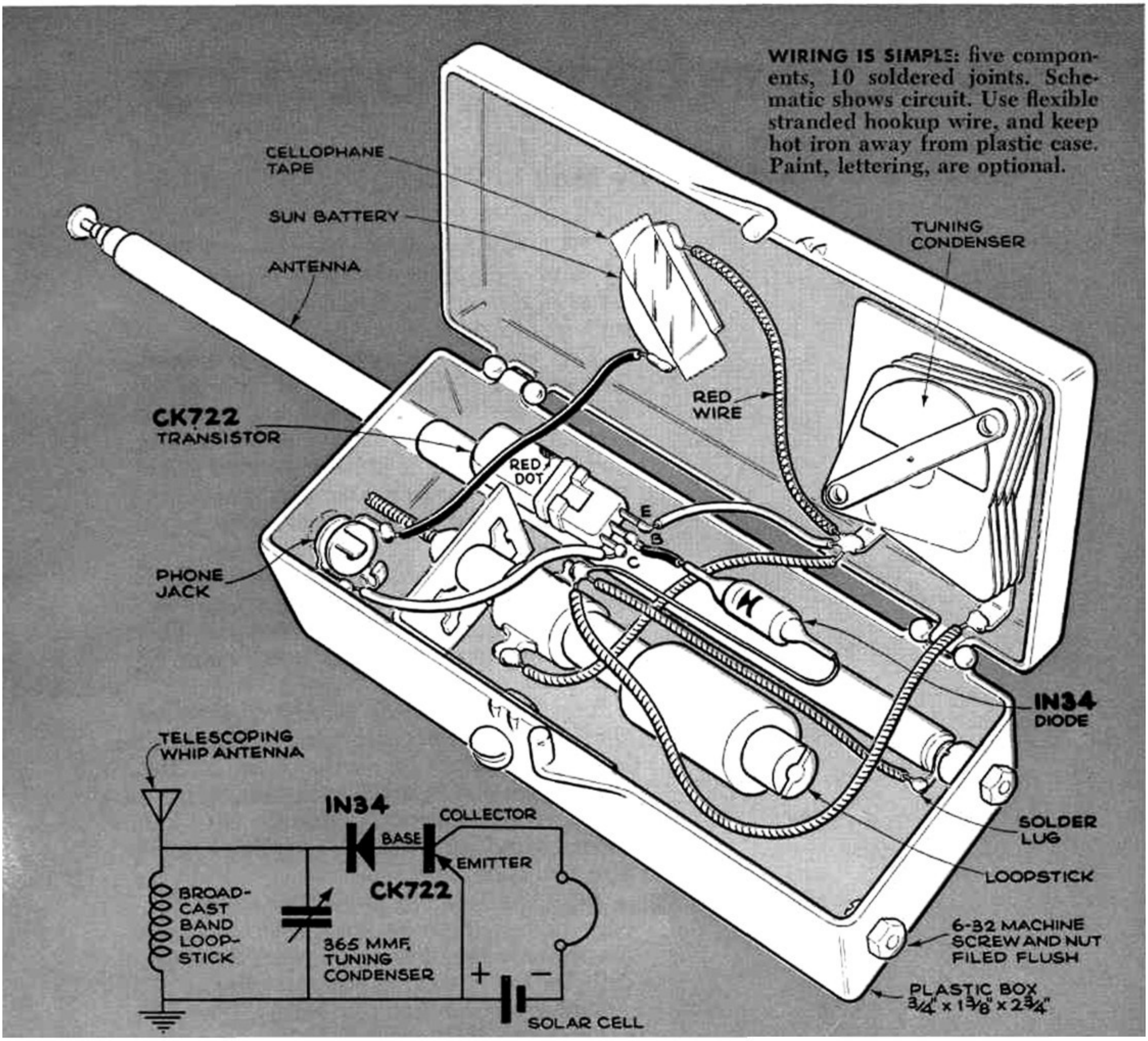
CELLOPHANE TAPE
SUN BATTERY
ANTENNA
CK722 TRANSISTOR
RED DOT
E
B
C
PHONE JACK
RED WIRE
TUNING CONDENSER
IN34 DIODE
SOLDER LUG
LOOPSTICK
6-32 MACHINE SCREW AND NUT FILED FLUSH
PLASTIC BOX 3/4" x 1 3/8" x 2 3/4"
TELESCOPING WHIP ANTENNA
IN34
BASE
COLLECTOR
EMITTER
CK722
BROAD-CAST BAND LOOP-STICK
365 MMF. TUNING CONDENSER
+
−
SOLAR CELL

WIRING IS SIMPLE: five components, 10 soldered joints. Schematic shows circuit. Use flexible stranded hookup wire, and keep hot iron away from plastic case. Paint, lettering, are optional.

ing the part as you enlarge the hole, to insure a snug fit.

The case can be painted or left transparent. The one pictured has a finish of Day-Glo spray paint, applied to the outside of the case. The paint softened the plastic slightly and created a wrinkled finish.

The hinged top is left clear and a piece of painted bond paper cut to fit inside. This paper was lettered, and cut to accept the dial shaft and solar battery. The lettering is pure whim—there's no trademark or commercial name.

After painting, go ahead with the assembly. Set the antenna mounting base through the bottom hole, put on the nut, cut off excess threads, and file smooth. Add a 6-32 screw and nut to the other side if you expect the radio to stand up.

Condenser and battery. Install the tuning condenser in the hinged top by first removing the dial and the small flathead screws. Remove both hex nuts; use one on the outside to hold the condenser in place. Then mount the dial and twist the silver center screw thumb tight. Place the solar battery in position over the opening in the bond paper, and fasten it in place with a strip of cellophane tape.

Bend the loopstick mounting bracket as shown, snap the loopstick in place, and fasten the bracket with a 6-32 screw and

[Continued on page 230]

Space-Age Crystal Set

[Continued from page 171]

nut. Install the phone jack, and the radio is ready for wiring.

Use stranded hookup wire; it can take the flexing imposed by opening and closing the case.

Pitfalls for amateurs. The wiring shown in the drawing is straightforward and follows the circuit schematic. If you're familiar with electronic wiring, you'll probably find some short cuts. If you're a beginner, stick to the diagram. If you run into trouble, don't hesitate to remove the parts, do the soldering outside, and then replace them.

Use care when soldering the diode and transistor—excess heat will damage them. Sidestep part of the risk by soldering to the lugs of a transistor socket, then slipping the transistor in place. You'll have to clip the transistor's leads to make it fit.

Bend the diode's leads to fit before soldering it between the base of the transistor and the loopstick.

Close the case when you've finished soldering, plug in the earphone, and listen. There's no on-off switch: The battery never wears out.

You should pick up stations at their proper position on the dial as indicated by the point of the battery or a dot on the lid. If the calibration is too far off, make a hole in the end of the case in line with the slotted screw on the loopstick. Turning this screw one way or the other will shift the stations enough to align the dial.

Outdoors, you'll find you don't have to point the Solar Mite at the sun—there is plenty of incident light for satisfactory operation. Keep the whip upright for maximum volume.

For stronger signals. Touching the silver center of the tuning dial may give you a stronger signal, since your body acts as a ground. Also try wetting your finger before touching the dial.

While the Solar Mite will play on the whip antenna, for peak performance connect the whip to your TV antenna. If you don't have an outdoor antenna of any kind, clip a lead on the finger stop of a telephone dial. You'll find that the radio works well enough to bring in a strong station even in the dark—acting as a crystal set without the help of its transistor amplifier.

No. 96

CRYSTAL SET CONSTRUCTION

SIMPLE INSTRUCTIONS FOR MAKING NINE CRYSTAL SETS

by B. B. BABANI

There is to-day a marked revival of interest in crystal receivers, and this book has been specially prepared to introduce the home constructor to the design and building of the various types. The text is simple and free from technicalities, so that even the young reader may confidently embark on the construction of any of the receivers shown.

CONTENTS

1. AERIALS.
2. EARTHS.
3. CRYSTAL DETECTORS.
4. COILS.
5. THE BEGINNERS' CRYSTAL RECEIVER.
6. FRAME AERIAL RECEIVER.
7. LONG DISTANCE RECEIVER.
8. HIGH GAIN RECEIVER.
9. HIGH SELECTIVITY RECEIVER.
10. BATTERY AIDED CRYSTAL RECEIVER.
11. MEDIUM AND LONG BAND RECEIVER.
12. ULTRA SENSITIVE ALL WAVE RECEIVER.
13. AN ADVANCED CRYSTAL DIODE RECEIVER.

THERE ARE 12 DIAGRAMS

LONDON: BERNARDS (PUBLISHERS) LIMITED

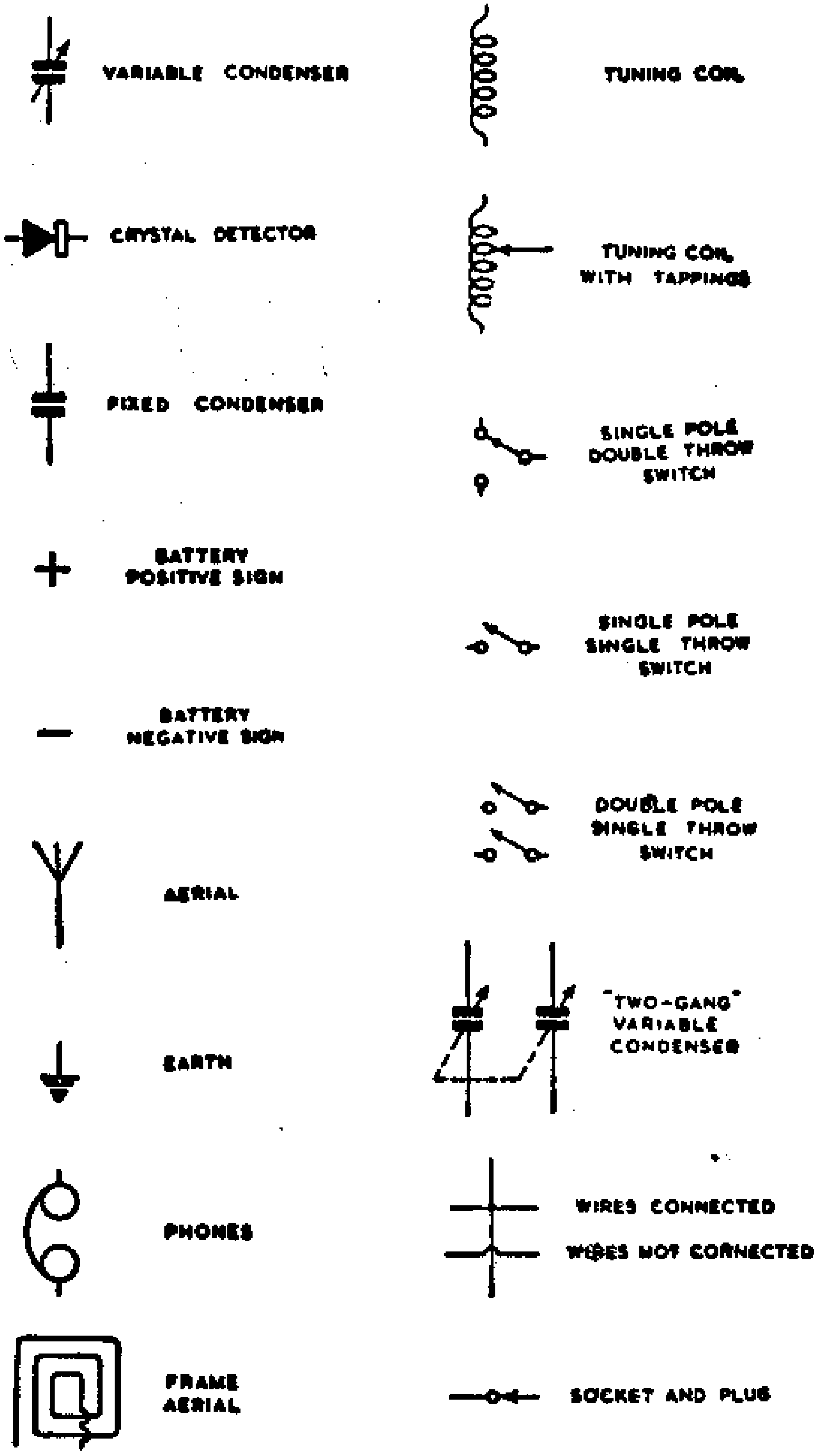
VARIABLE CONDENSER
CRYSTAL DETECTOR
FIXED CONDENSER
BATTERY POSITIVE SIGN
BATTERY NEGATIVE SIGN
AERIAL
EARTH
PHONES
FRAME AERIAL
TUNING COIL
TUNING COIL WITH TAPPINGS
SINGLE POLE DOUBLE THROW SWITCH
SINGLE POLE SINGLE THROW SWITCH
DOUBLE POLE SINGLE THROW SWITCH
"TWO-GANG" VARIABLE CONDENSER
WIRES CONNECTED
WIRES NOT CONNECTED
SOCKET AND PLUG

FIG. I SYMBOLS USED IN THE CIRCUITS DESCRIBED.

1. Aerials.

One of the essentials for satisfactory reception with crystal sets is an efficient aerial. Four different types are suggested and are shown in Fig. 2.

There are several important points to be noted when erecting an aerial for a crystal set.

(1) See that it is suspended as high as possible and is not screened by tall buildings, obstructions, etc.

(2) The aerial itself and the lead-in from the aerial must be adequately insulated. Be sure that the porcelain or plastic insulators do not permit the aerial or lead-in to touch the building, tree, or post from which they are suspended. The lead-in should be brought into the house through an insulated rod.

(3) It is exceedingly important to use insulated copper wire, either single or multi-stranded, for both the aerial and the lead-in.

2. Earths.

An adequate and efficient earth contributes to the good performance of any well-designed crystal set and Fig. 3 shows the best method of obtaining an efficient earth where there is a garden or patch of ground available at a short distance from the receiver.

Obtain a 3-ft. length of ½-in. galvanised iron pipe and at one end drill a hole right through. Insert a bolt with spring washer and locking nut in the hole. Connect the earth lead-in wire to the bolt; the other end of the lead-in to the earth connection on the receiver. Leave both ends of the pipe open. Insert the end opposite the bolt into the ground vertically until only 3-in. are showing above the surface of the surrounding earth.

It is important to see that the surrounding ground is kept damp, if necessary by occasionally pouring a pint or two of water down the open projecting end of the pipe.

Ready-made copper earthing rods with a pointed end for easy hammering into the earth are sometimes obtainable from radio dealers.

Should a garden not be available, an alternative earth is a cold water pipe coming from a rising main. Simply file very lightly the pipe at the back of the tap and wind the copper wire earth lead-in round the filed part and cover the joint with insulating tape.

In no circumstances should a gas pipe be used for an earth.

3. Crystal Detectors.

Most radio dealers can supply very efficient permanent or semi-permanent crystal detectors of proprietary makes. However, there are always available quite a number of ex-Government surplus units which are called crystal valves and, though these were originally designed for a very different use in service equipment, they are easily adaptable and will give extremely efficient results when used in circuits designed for them. To enable the constructor to purchase suitable types for this purpose—and they cost very little—the following list gives the Government reference numbers.
CV 101, CV 102, CV 103, CV 111, CV 112, CV 113, CV 226, CV 241, CV 246, CV 247, CV 253, CV 291, CV 351, CV 361, CV 364, CV 367, CV 727, CV 749, and CV 1785.

One of the satisfactory proprietary makes of this class of crystal detector is the BTH CS7A, which, incidentally, is the commercial equivalent of the CV 253.

For all these Government types, one connection should be made to each end by arranging a holder to fix them into or by carefully soldering the appropriate connecting wires to each end.

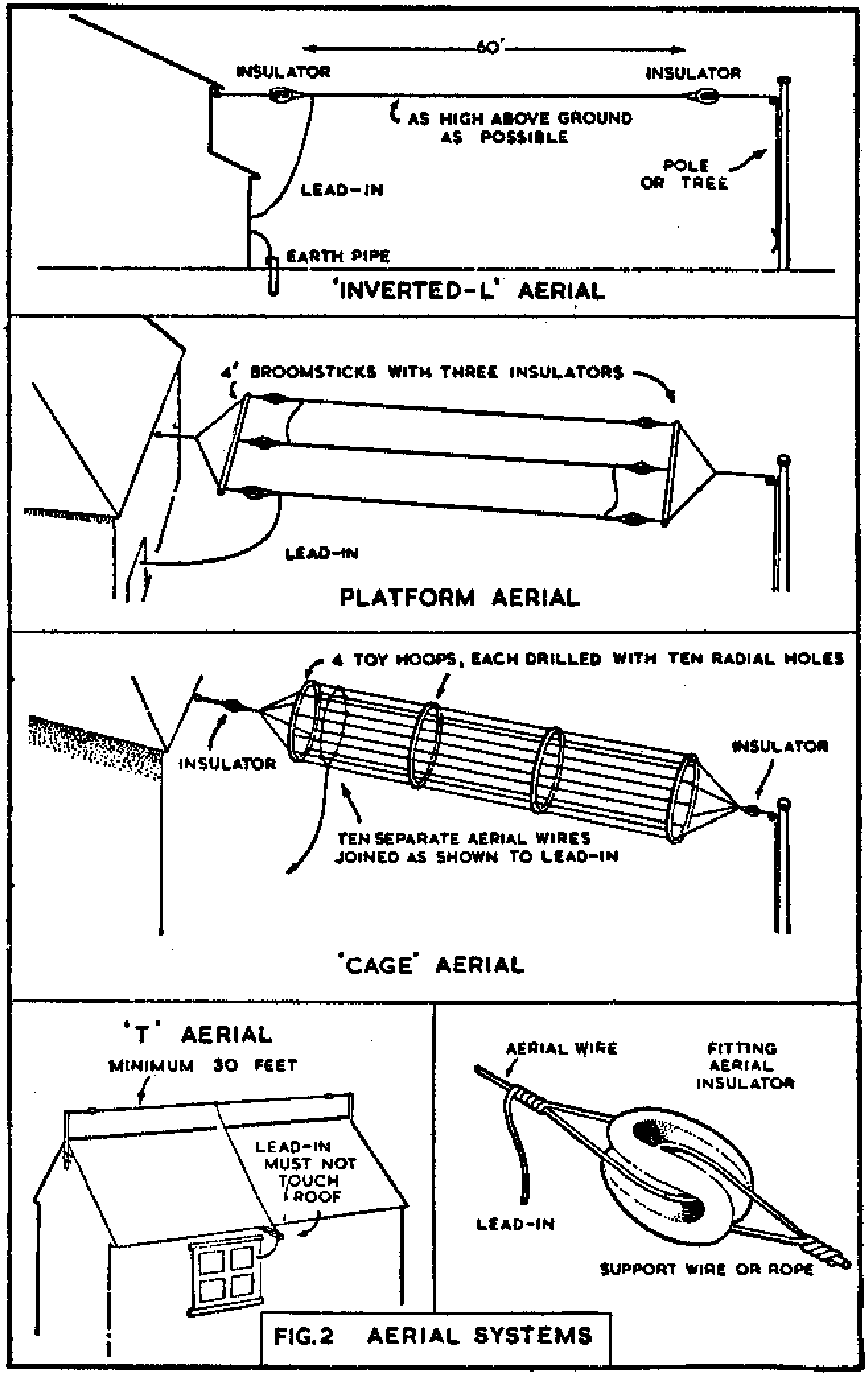
60'
INSULATOR
INSULATOR
AS HIGH ABOVE GROUND AS POSSIBLE
POLE OR TREE
LEAD-IN
EARTH PIPE
4' BROOMSTICKS WITH THREE INSULATORS
LEAD-IN
4 TOY HOOPS, EACH DRILLED WITH TEN RADIAL HOLES
INSULATOR
INSULATOR
TEN SEPARATE AERIAL WIRES JOINED AS SHOWN TO LEAD-IN
MINIMUM 30 FEET
LEAD-IN MUST NOT TOUCH ROOF
AERIAL WIRE
FITTING AERIAL INSULATOR
LEAD-IN
SUPPORT WIRE OR ROPE

'INVERTED-L' AERIAL

PLATFORM AERIAL

'CAGE' AERIAL

'T' AERIAL

FIG. 2 AERIAL SYSTEMS

4. Coils.

In Fig. 3 two methods are shown of fixing the coil to the chassis. However, the ingenious constructor may evolve another arrangement.

The important thing to remember is that the coil should be fixed so that it is completely insulated from the metal chassis that may be used.

It is also important to remember that if a metal chassis is used, one must see that all components and wiring, with the exception of earth connections, are adequately insulated from this chassis so as to make sure that no short circuits take place. It is recommended that all crystal sets should be built on a panel, and that both panel and chassis should be of bakelite, paxolin or perspex, as these three materials are efficient insulators in themselves.

5. The Beginners' Crystal Receiver.

The popularity of the crystal receiver among beginners is undoubtedly due to the very modest outlay necessary, and the ease with which such receivers may be built. It is not always appreciated that, to obtain worth while results,

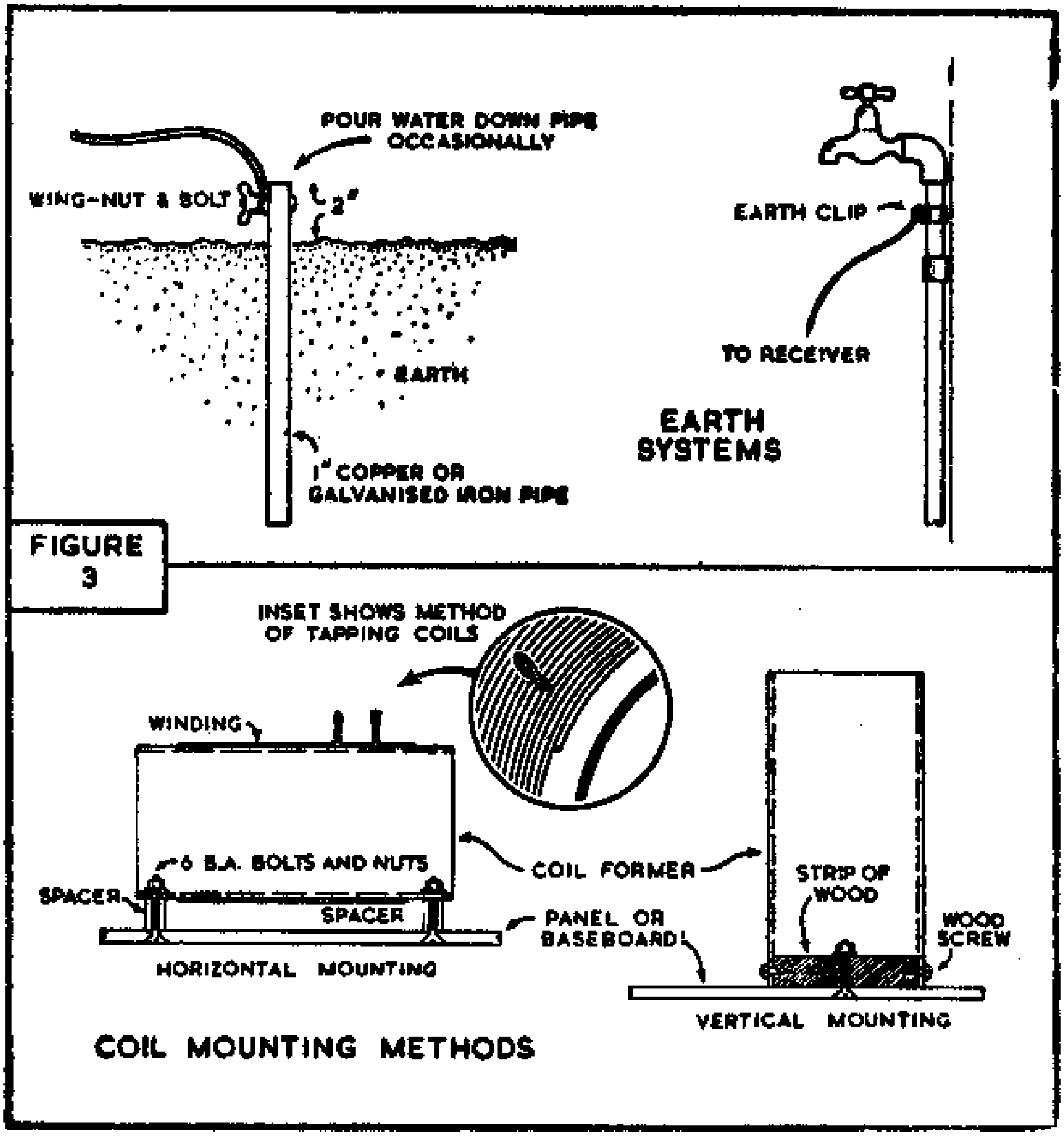
POUR WATER DOWN PIPE OCCASIONALLY
WING-NUT & BOLT
2"
EARTH
1" COPPER OR GALVANISED IRON PIPE
EARTH CLIP
TO RECEIVER
INSET SHOWS METHOD OF TAPPING COILS
WINDING
6 B.A. BOLTS AND NUTS
SPACER
SPACER
COIL FORMER
PANEL OR BASEBOARD
STRIP OF WOOD
WOOD SCREW
HORIZONTAL MOUNTING
VERTICAL MOUNTING

EARTH SYSTEMS

FIGURE 3

COIL MOUNTING METHODS

far more care in avoiding losses is necessary than with a three or four valve receiver. One reason for this is the fact that the crystal receiver can offer no help by way of amplification and it is therefore dependent on an efficient aerial/earth system, low loss design within the receiver itself, and the use of sensitive headphones. The experience gained is well worth the time expended since, when more ambitious receivers are attempted, the constructor will remember the benefits gained by efficient construction and installation with the result that each receiver attempted will be an instrument capable of giving a first class performance.

Losses within the receiver itself may be considerably reduced by efficient coil design. Unfortunately, selectivity is a factor which must not be overlooked; it is this factor which governs the sharpness of the tuning and permits the separation of adjacent powerful stations. The greater the selectivity, the greater the ease with which signals may be separated; but at maximum selectivity the received signals are at minimum strength. In Fig. 4, the coil is tapped at four points, which enables the best point between selectivity and sensitivity to be chosen. The nearer the aerial tap to the earth end of the coil, the greater the selectivity. The pre-set condenser C1 also aids selectivity; at minimum capacity the selectivity is high, and *vice versa*.

Used in conjunction with the coil taps, the constructor should have no difficulty in selecting a condenser setting and tapping point which permit the reception of strong signals and at the same time effect the required separation of powerful stations.

The coil is constructed on a former 3-in. diameter × 2-in., and consists of 50 turns 36 S.W.G. enamelled wire close wound and tapped at every tenth turn. These taps are brought out in the form of loops twisted firmly so that there is no actual break in the winding continuity. After completion the winding may be painted with Durafix to prevent any movement, and the taps and two free ends cleaned with emery cloth to remove the enamel. The layout and wiring is shown in Fig. 4. The headphones should have a resistance of some 4,000 ohms or more and should be of reputable manufacture. Any of the aerials previously described are suitable for use with this receiver.

C2 .0005 variable condenser.
C1 .0002 pre-set condenser.
C3 .001 mica condenser.
1 Crystal detector, semi-permanent pattern.
2 oz. 30 S.W.G. enamelled copper wire.
1 Coil former 3-in. dia. × 2-in.
1 Ebonite or Bakelite panel 6-in. × 5-in. × ¼-in.
4 terminals, connecting wire, etc.

6. Frame Aerial Receiver for Local Station use on Medium wave bands.

This receiver was specially designed for constructors who are within 5 to 10 miles of a powerful station and who have no facilities for constructing a good outdoor aerial.

The frame itself consists of two ¾-in. × ¾-in. square section wooden rods each 36-in. in length. These are placed together at their centre line to form a cross; they are fastened by a nut and bolt and these in turn pass through an upright piece of wood ½-in. × 1-in. section 2-ft. in length. This upright support is then screwed into a baseboard 8-in. × 8-in. × 1-in. thick. At each of the arms of the 36-in. cross a 4-in. strip of ½-in. square section wood is placed. Reference to Fig. 5B will give the details of the construction of this aerial quite clearly.

It will be noted that approximately 85-ft. of 20 S.W.G. double cotton-covered wire are required for the inductance to cover the medium wave broadcast band. The wire should be wound around the four crossbars, thereby creating

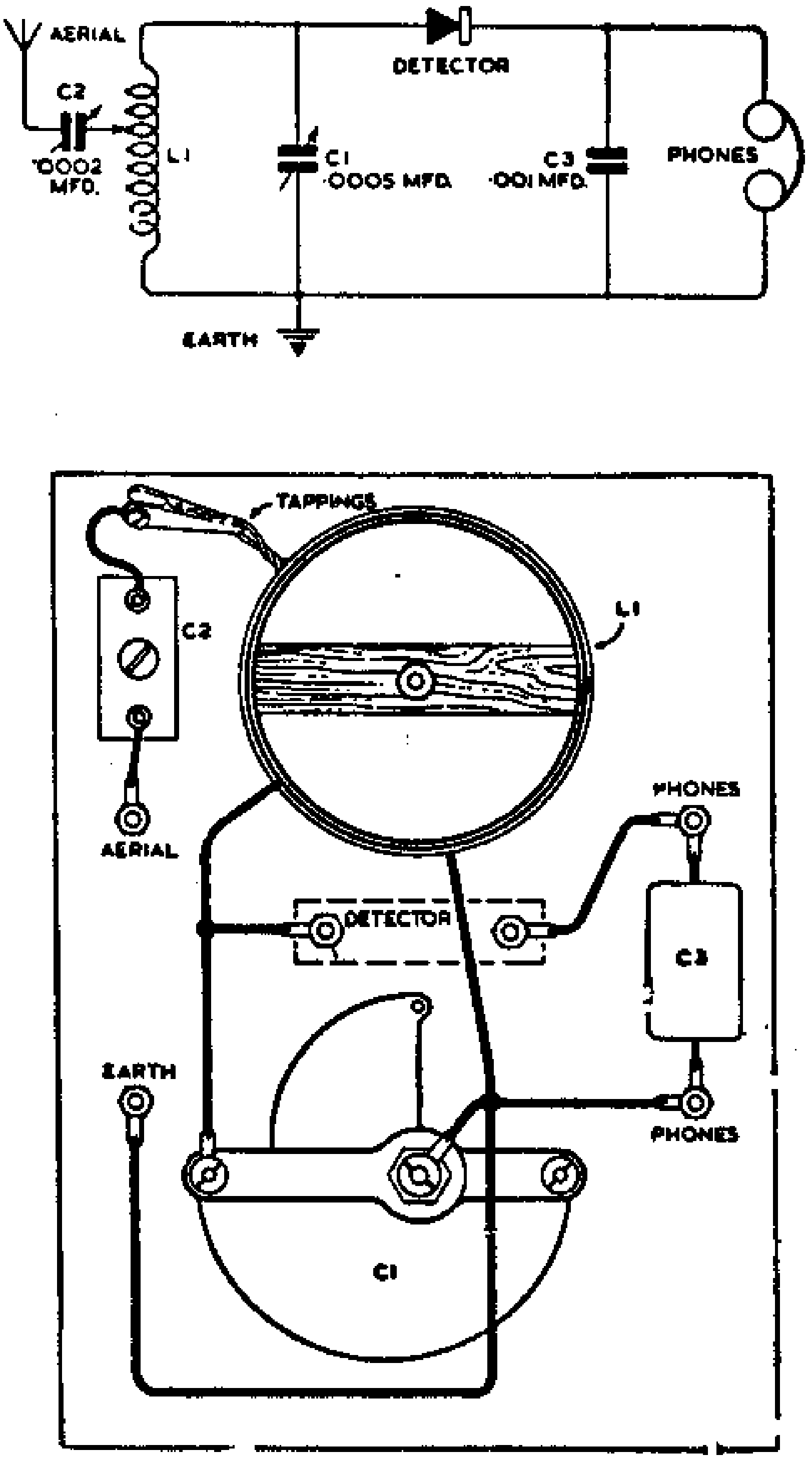
AERIAL
C2
·0002 MFD.
L1
C1
·0005 MFD.
DETECTOR
C3
·001 MFD.
PHONES
EARTH
TAPPINGS
C2
L1
AERIAL
PHONES
DETECTOR
C3
EARTH
PHONES
C1

FIG. 4 THE BEGINNER'S CRYSTAL RECEIVER.

a square wire frame, and each turn of the wire should be spaced approximately ⅛-in. The best way to effect this spacing accurately is to file on the 4-in. crossbars a number of small nicks at ⅛-in. intervals. The two ends of the wire should then be brought down on the two terminals which may be screwed into the square baseboard and two leads taken from there to the crystal set itself.

The actual set may be constructed on a small ebonite, paxolin or perspex panel 8-in. × 5-in. which may be screwed to one side of this base supporting piece of wood, thereby making the set virtually self-contained except for the earphones. Fig. 5A shows the circuit of the receiver.

To obtain the best results with this receiver, the aerial should be rotated by hand until the signals become loudest.

1 Semi-permanent crystal detector.
C1 1 .0005 mfd. variable condenser.
C2 1 .0005 mfd. fixed condenser.
1 condenser dial marked 0 to 100 or 0 to 180.
1 pair of high sensitivity earphones.
85-ft. to 90-ft. of 20 S.W.G. double cotton covered wire. (8oz.)
Sufficient connecting wire for wiring purposes.

7. Long Distance Receiver.

This receiver has been specially designed to achieve high sensitivity. With a sufficiently good outdoor aerial effective reception has been obtained up to 150 miles.

The circuit is quite straightforward, as can be seen from Fig. 6 and the only special point is in winding the coil L1.

This coil consists of a 2½-in. diameter former on which are wound a total of 51 turns using 24 S.W.G. double cotton wire. Starting at one end, tap every two turns until there are eight taps; wind a further 15 turns and tap; then a further 10 turns and tap; finally, add 10 turns to end the winding. The wire should be close wound, *i.e.*, with turns touching.

To obtain the best results, the two crocodile clips A and B should be tried on the various coil taps until the loudest reception is obtained. It may be found in use that the loudest reception of any particular station causes another station to be heard at the same time; care should, therefore, be taken to select the appropriate tapping point to separate any two powerful stations that may interfere with one another.

4 oz. 24 S.W.G. DCC copper wire.
Coil former 2½-in. dia. × 3-in.
C1-2 single variable condensers .00035 mfd.
2 Condenser dials marked from 0 to 100 or 0 to 180.
1 Permanent crystal detector.
C3 .001 fixed condenser.
1 pair of high sensitivity earphones (preferably between 4,000 ohms and 8,000 ohms impedance).
2 Crocodile clips.

8. High Gain Receiver.

This receiver has been designed to give good reception up to 50 miles distance from any normal power broadcasting station in the medium band. It is a very selective circuit and will therefore enable the user to separate closely situated stations.

It will be noticed in the circuit (Fig. 7) that the variable condenser C1, has an on/off switch S1 to throw it out of circuit if necessary. It is used to increase the selectivity of the receiver and, in operating, it should be tried with this condenser either in or out of circuit on the various tappings on the coil L1.

L1 consists of 90 turns of 22 S.W.G. double cotton covered wire wound on a 2-in. diameter former and tapped every ten turns. L2 separately wound on a 2-in. former, consists of 90 turns of 22 S.W.G. double cotton covered wire, also tapped as L1. The components required are given on page 9

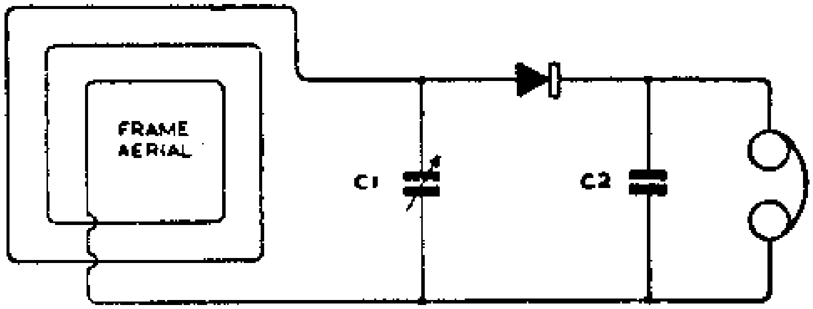
FRAME AERIAL
C1
C2

FIG. 5A FRAME AERIAL CRYSTAL RECEIVER

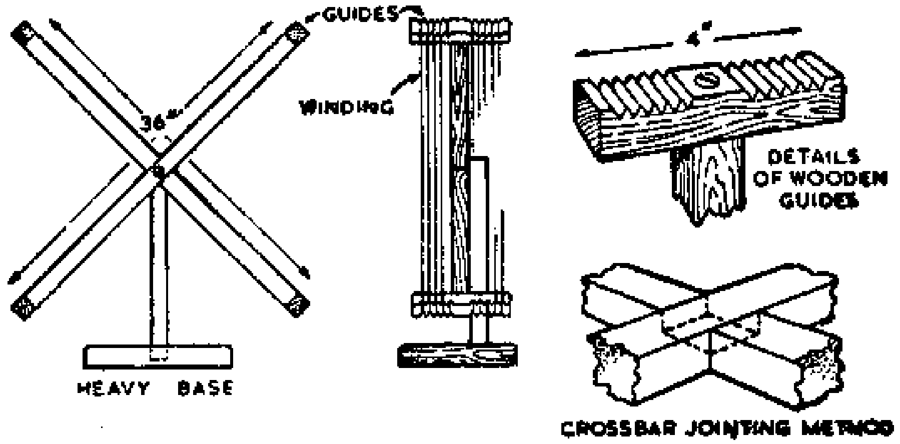
GUIDES
36"
WINDING
4"
DETAILS OF WOODEN GUIDES
HEAVY BASE
CROSSBAR JOINTING METHOD

FIG. 5B. DETAILS OF FRAME AERIAL

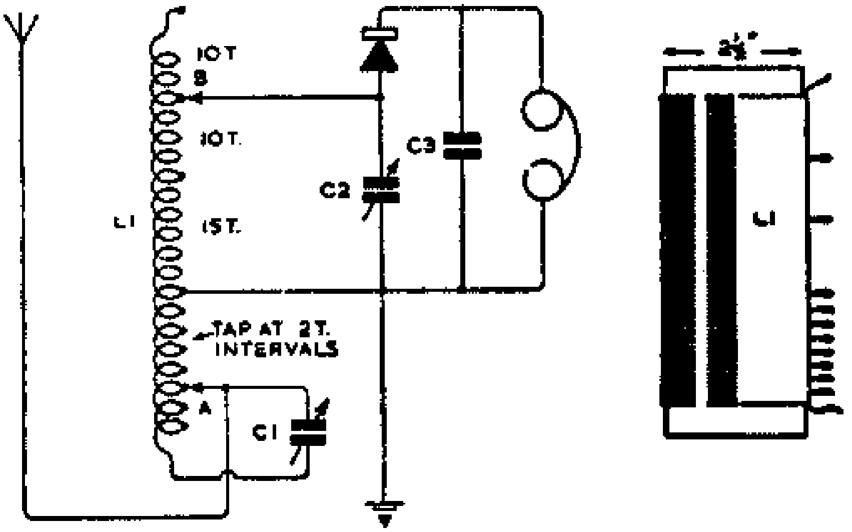
10T
B
10T.
15T.
L1
C2
C3
TAP AT 2T. INTERVALS
A
C1
2½"
L1

FIG. 6 LONG-DISTANCE CRYSTAL RECEIVER

C1-2 Single variable condensers .00035 mfd.
S1 Single pole single throw switch.
C5 .0001 preset condenser.
1 Permanent crystal detector.
C3-4 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
8 oz. 22 S.W.G. DCC.
2 Dials for variable condenser marked 0 to 100 or 0 to 180.

9. High Selectivity Receiver.

This circuit (see Fig. 8) is for use in high saturation strength areas where it is desired to separate powerful stations. The circuit is not complicated and is extremely effective.

It will be noted that C1 and C2, each .00035 mfd. capacity, are 2 gang unit.

The only special point in this set is the design of coils L1, 2 and 3, which are constructed as follows: L1, 60 turns of 22 S.W.G. double cotton covered wire on a 3 × 4 inches former tapped every five turns; L2, 15 turns of 22 S.W.G. double cotton covered wire on a 3 × 6 former; and L3, 85 turns of 22 S.W.G. double cotton covered wire wound on the same former as L2, the two windings being separated by approximately ¼-in. (see Fig. 8).

Note that the aerial lead is brought down to one of the tappings on wire L1. This should be tried on each of the tappings and the one giving the required selectivity should be used.

C1-2 2-gang variable condenser .00035 mfd.
2 3-in. formers for coils L1-2-3.
8 oz. 22 S.W.G. DCC.
C3 .0005 fixed condenser.
1 Semi-permanent crystal detector.

10. Battery-aided Crystal Receiver.

Fig. 9 shows a design which utilises a 9-volt grid bias battery to improve performance.

The charge built up in the .02 mfd. fixed condenser is changed to a positive current when it passes through the coil L1. This, in turn, is superimposed on the positive charge that is commencing to be rectified by the crystal detector and thereby increases the signal strength.

Coil L1 consists of a 4-in. diameter former wound with 85 turns of 24 S.W.G. double cotton covered wire tapped at 5, 25, 45 and 65 turns from the earth end. The earth end and these four taps are brought out to five small sockets and the connection from the earth is taken to a small socket. The plug should be tried in each of the five positions to find which one gives the greatest signal strength.

The 9-volt battery is connected across the .02 mfd. condenser and special care should be taken to see that the battery is connected correctly, *i.e.*, negative pole to earth.

1 Permanent crystal detector.
1 4-in. dia. × 3-in. former.
8 oz. 20 S.W.G. DCC copper wire.
C1 Variable condenser .0005 mfd.
C2 Fixed condenser .005 mfd.
C3 Fixed condenser .02 mfd.
1 pair of high sensitivity headphones.
1 Clix plug.
5 Clix sockets.
1 9-volt grid bias battery.

11. Medium and Long wave band Receiver.

This set has been designed for reception on two wave bands.

Care must be taken in winding the coils L1, 2, 3 and 4. These are wound on a single 3-in. diameter former. L1 consists of 30 turns of 28 S.W.G. DCC tapped at 8, 16 and 24 turns. L2, 60 turns of 32 S.W.G. DCC. L3, 40 turns of 28 S.W.G. DCC, and L4, 80 turns of 32 S.W.G. DCC. All are close wound and are separated from each other as follows: between L1 and L2, 3/16-in.;

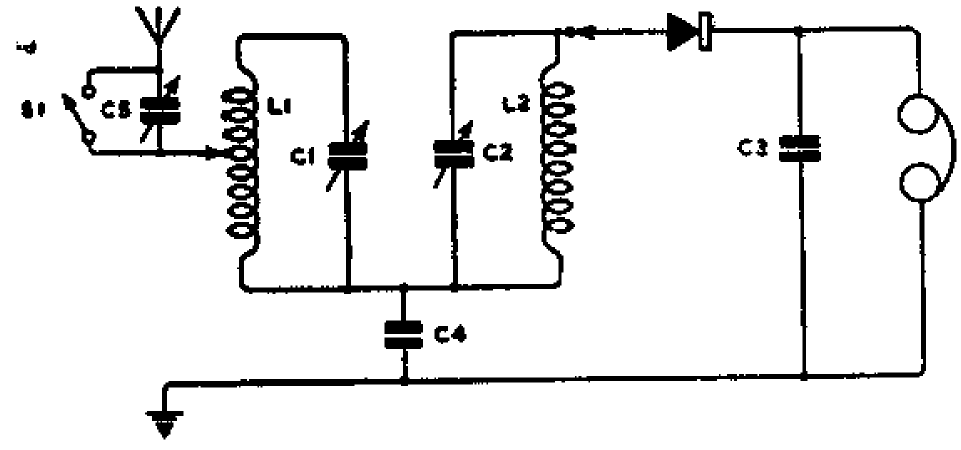
S1
C5
L1
C1
C2
L2
C3
C4

FIG. 7 HIGH-GAIN CRYSTAL RECEIVER

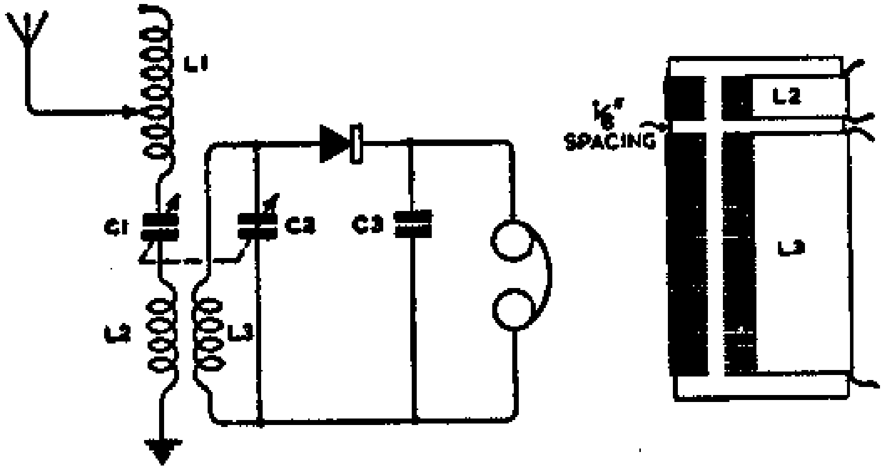
L1
C1
C2
C3
L2
L3
1/8" SPACING
L2
L3

FIG. 8 HIGH SELECTIVITY CRYSTAL RECEIVER

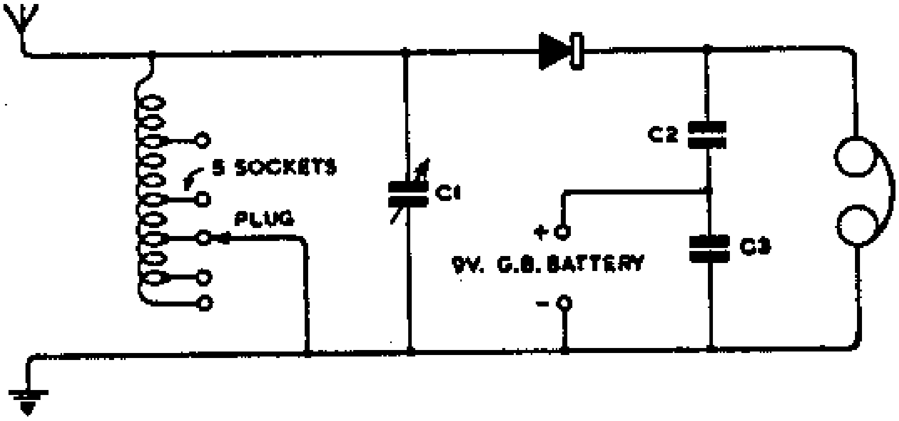
C2
5 SOCKETS
PLUG
C1
+
9V. G.B. BATTERY
−
C3

FIG. 9 BATTERY-AIDED CRYSTAL RECEIVER

between L2 and L3, 1-in.; and between L3 and L4, 3/16-in. See Fig. 10.

C1-2 single variable condensers .0005 mfd.
2 Variable condenser dials marked 0 to 100 or 0 to 180.
1 Clix plug.
3 Clix sockets.
S1 Single pole double throw switch.
C3 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
1 3-in. dia. × 8-in. former.
2oz of both 28 and 32 S.W.G. double cotton covered wire to make these 4 coils.
1 Permanent crystal detector.

In use, the plug should be tried in each of the three sockets to find the point which gives the best results. The switch is used as a wave band change switch.

12. Ultra Sensitive All Wave Receiver.

The receiver shown in Fig. 11 is for long distance reception. It is a novel circuit and, though apparently complicated in construction, it is really very simple to assemble with the minimum of tools and labour.

It will be seen from the circuit diagram that there are two banks of sockets, five black and six red. When using this set the appropriate black and red plugs should be inserted in whichever of the similar coloured sockets give the best results. The coil is quite easy to construct and consists of a primary winding which is inserted at the earth end of the secondary winding. For the medium wave band, the secondary winding L2 consists of 54 turns of 20 S.W.G. enamelled or cotton covered wire. This coil has four taps in addition to the starting point and the end connection, and these taps are taken at six turns, 14 turns, 27 turns and 40 turns from the earth end. The taps are taken out and a connection is made to an appropriate socket, as shown in the diagram.

The coil is wound on a 3-in. diameter former which may be a paper tube, bakelite, paxolin or perspex, and spacing is approximately 17 turns to the inch.

For short wave work down to about 30 metres, L2, the secondary coil, will consist of the same diameter former wound with 15 turns of the same gauge wire, the four taps being taken at 3, 6, 9 and 12 turns from the earth end.

L1, the primary coil, is used for both wave bands and consists of 11 turns of similar gauge wire to that used for L2, wound on a 2-in. × 1-in. former, the winding to be spaced the same as L2. The coil is then inserted at the earth end of whichever secondary coil L2 is being used.

The crystal detector may be one of the semi-permanent types freely available to-day; the old-fashioned cat's whisker pattern; or one of the silicon permanent detectors which are sold by many surplus dealers.

The 'phones should be of a standard type 4,000 ohms impedance although 'phones of other values will probably operate quite satisfactorily with this powerful receiver. It should, however, be stressed that the better the headhones used, the more satisfactory the reception.

Coils for other wave bands may be experimentally designed by the constructor, remembering that four tappings are advisable for the secondary coil L2.

A long aerial is very desirable with this receiver, placed as high as possible. It is recommended that a minimum length of 70-ft. be allowed, to include the aerial and lead-in.

The values of the components are shown in the following components list. The constructor should experience no difficulty in obtaining excellent results from this cleverly designed receiver.

C1 .0005 variable condenser.
C2 .002 fixed condenser.
1 semi-permanent crystal detector.
1 coil former 3-in. dia. × 4½-in.
1 coil former 2-in. dia. × 1-in.
½ lb. 20 S.W.G. enamelled or DCC copper wire.
6 red wander plug sockets.
6 black wander plug sockets.
1 red and 1 black wander plug.
Headphones, terminals, connecting wire

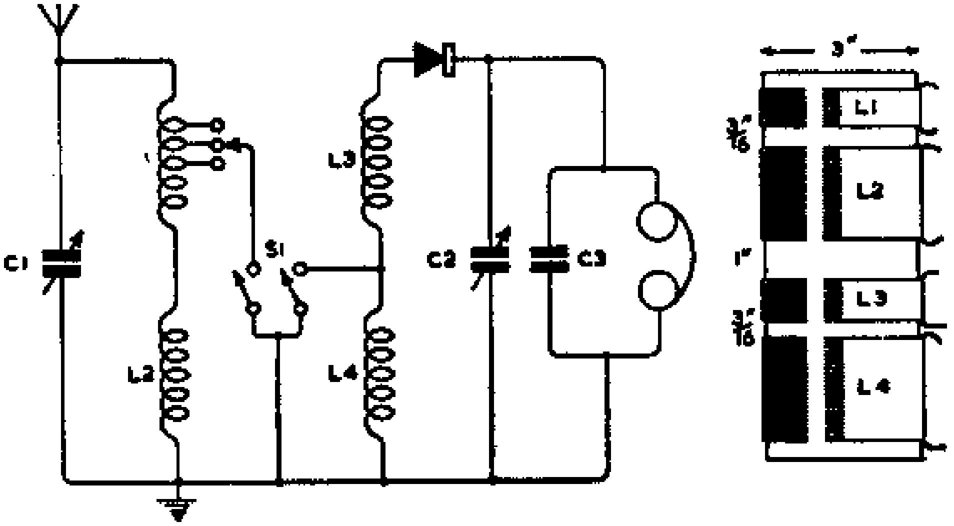
C1
L2
S1
L3
L4
C2
C3
3"
3/16"
L1
L2
1"
L3
3/16"
L4

FIG. 10 MEDIUM AND LONG-WAVE CRYSTAL RECEIVER

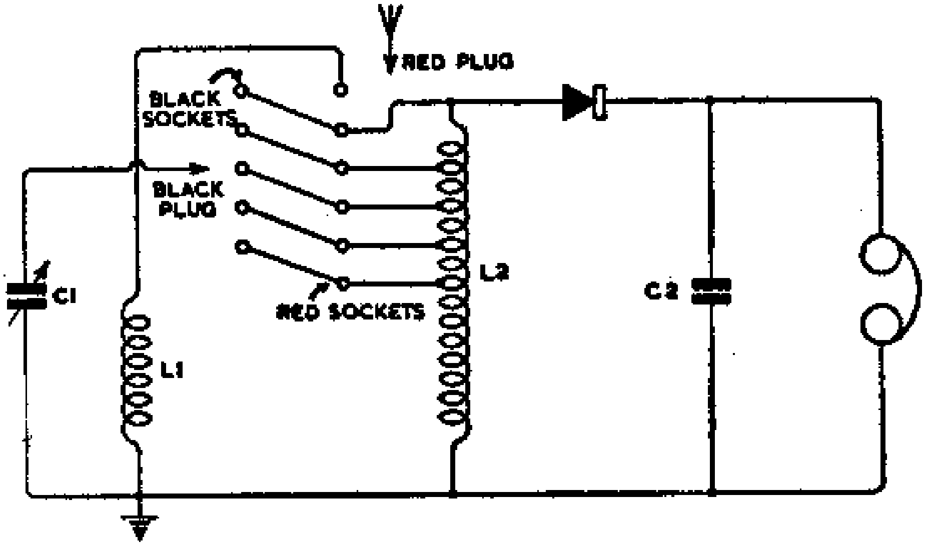
RED PLUG
BLACK SOCKETS
BLACK PLUG
RED SOCKETS
C1
L1
L2
C2

FIG. 11 ULTRA-SENSITIVE ALL-WAVE CRYSTAL RECEIVER

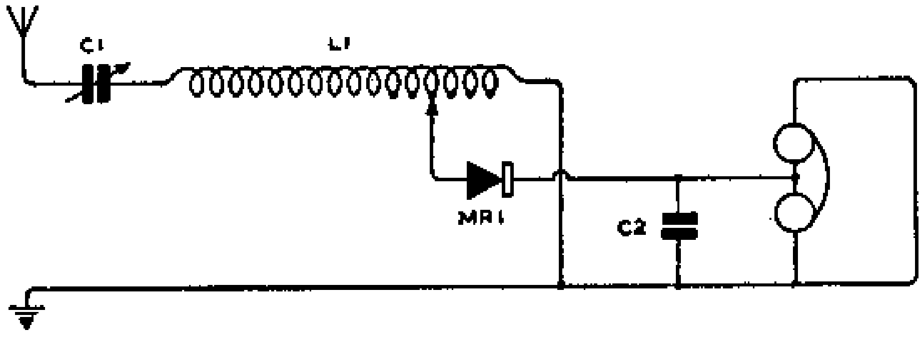

FIG. 12 ADVANCED CRYSTAL DIODE RECEIVER

13. An Advanced Crystal Diode Receiver.

The modern crystal diode of the radar type does not require any adjustment and thus has a distinct advantage over the ordinary crystal detector which needs delicate adjustment—even the semi-permanent type.

Such crystals are now obtainable on both the regular and the surplus markets, and a popular type is the B.T.H. CS7A silicon crystal (also coded as CV 253).

A rather different circuit from the usual crystal set arrangement is needed to suit the characteristics of a crystal diode. The diode must be tapped on to the tuning coil, and it is found that the tuning circuit itself gives best results if a series-tuned acceptor circuit is used. The circuit of a radar crystal receiver is shown in Fig. 12. The type of aerial used with this circuit has a very great bearing on the behaviour of the receiver for, in effect, the tuned circuit, with a series resistance equivalent to the reflected crystal load resistance, is in series with the capacitance of the aerial to earth and the aerial's effective series resistance.

At resonance—when the combination is tuned to any particular signal—the inductance resonates with the capacitances of the tuning condenser and the aerial in series, and the final effect is that the reflected load of the crystal, in series with the coil's R.F. resistance, is paralleled across the aerial's series resistance.

For maximum power transfer, the crystal load resistance must be made equal to the sum of the aerial series resistance and the coil R.F. resistance, and so the method in which the crystal is tapped into the tuning coil, and the exact capacitance required to tune the receiver to any required signal, must depend to a very great extent on the aerial itself.

At the same time the crystal resistance varies with the signal strength, the resistance being high for weak signals and dropping by as much as 50 per cent. and more for strong signals, so that this effect also has a bearing on the correct coil tap. The output impedance of the crystal also varies similarly, affecting the matching of the headphones into the crystal diode, and so for any set of conditions, the receiver requires to be matched up to both the aerial and the signal being received for best results. This would mean a series of coil taps and (theoretically, not practically) a matching transformer between the diode and the headphones; but in practice it will be found that a receiver using standard parts can be built up to give very good results under various conditions.

It has been shown already that the tuning of the receiver depends to a great extent on the characteristics of the aerial, and while a .0005 mfd. variable capacitor is shown in Fig. 12 as the tuner, the constructor must be prepared to experiment with different capacitance values until the required station is tuned in. The range of reception given by the receiver is quite good, if a really long and high aerial, and a good earth connection, are used; but no more than the local station signal can be expected, and the tuning therefore must be adjusted to suit the station frequency.

It must be mentioned that the headphones are shown parallel connected. High resistance headphones of the 4,000 ohm type must be used, and if two of these are connected in parallel rather than in the more usual series method, they will provide a roughly accurate match to the crystal. If more than one pair of headphones are to be connected in, then the pairs of headphones may be left series connected in the usual manner, the sets of headphones being connected in parallel.

The capacitor across the headphone terminals completes the crystal R.F. circuit, and any value between about 0.001 and 0.005 mfd. will serve. The higher capacitance will, of course, by-pass some of the higher audio frequencies, so that if good steady reception is obtained some experiment with this condenser is also worth while.

C1, .0005 mfd. variable condenser.
C2, .005 mfd. fixed condenser.
4oz. 26 S.W.G. enamelled copper wire.
M.R.1 CV 253 Crystal diode.
Coil former 1½-in. dia. × 4½-in.

Coil details are as follows: 150 turns of 26 S.W.G. tapped at 20, 25, 30, 35, 40 and 45 turns from the earth end. At each tapping-point the wire should be twisted up into a loop and the winding then continued without breaking the wire; when the coil is completed and the ends anchored the tapping loops can be bared.

BERNARDS RADIO BOOKS

No.		Price
56.	Radio Aerial Handbook	2/6
57.	Ultra-Shortwave Handbook	2/6
58.	Radio Hints Manual	2/6
64.	Sound Equipment Manual	2/6
68.	Frequency Modulation Receivers' Manual	2/6
73.	Radio Test Equipment Manual	2/6
83.	Radio Instruments and their Construction	2/6
96.	Crystal Set Construction	1/–
99.	One Valve Receivers	1/6
100.	A Comprehensive Radio Valve Guide, Book 1	5/–
103.	"Radiofolder" A. The Master Colour Code Index for Radio and T.V.	1/6
104.	Three Valve Receivers	1/6
107.	Four Valve Circuits	1/6
108.	Five Valve Circuits	2/6
121.	A Comprehensive Radio Valve Guide, Book 2	5/–
123.	"Radiofolder" F. The Beginners' Push-Pull Amplifier	1/6
126.	Boys' Book of Crystal Sets and Simple Circuits	2/6
129.	Universal Gram Motor Speed Indicator	1/–
134.	F.M. Tuner Construction	2/6
135.	All Dry Battery Portable Construction	2/6
138.	How to Make F.M. and T.V. Aerials, Bands 1, 2 and 3	2/6
141.	Radio Servicing for Amateurs	3/6
143.	A Comprehensive Radio Valve Guide, Book 3	5/–
145.	Handbook of AM/FM Circuits and Components	2/–
146.	High Fidelity Loudspeaker Enclosures	5/–
147.	Practical Tape Recording Handbook	5/–
148.	Practical Transistor Receivers, Book 1	5/–
149.	Practical Stereo Handbook	3/6
150.	Practical Radio Inside Out	3/6
151.	Transistor Superhet Receivers	7/6
155.	Portable Transistor Radio and Radiogram	2/6
156.	Transistor Circuits Manual, No. 1	2/6
157.	A Comprehensive Radio Valve Guide, Book 4	5/–
158.	Radio, Television, Industrial Tubes, Semiconductor and Diodes Equivalents Handbook (208 pages)	9/6
159.	Realistic High Fidelity	5/–
160.	Coil Design and Construction Manual	5/–
161.	Radio, Television and Electronics Data Manual	3/6
162.	High Fidelity Stereo Gramophone	5/–
163.	Transistor Circuits Manual, No. 2	2/6
164.	High Fidelity Tape Recorder for the Home Constructor	2/6
165.	Hartley on Hi-Fi, Book 1, Radio Tuners	5/–
166.	Public Address Systems	2/6
167.	Transistor Circuits Manual, No. 3	2/6
168.	Transistor Circuits Manual, No. 4	2/6
169.	Hi-Fi Transistor F.M. Tuner for the Home Constructor	3/6
170.	Transistor Circuits for Radio Controlled Models	7/6
171.	Super Sensitive Transistorised Pocket Set Construction	3/6
172.	International Radio Stations List	2/6
173.	Practical Transistor Audio Amplifiers	3/6
174.	Transistor Subminiature Receivers Handbook	5/–
175.	Transistorised Test Equipment and Servicing Manual	3/6
176.	Transistor Audio Amplifiers Manual	6/–
177.	Modern Transistor Circuits for Beginners	7/6
178.	Comprehensive Radio Valve Guide, Book 5	6/–
179.	Transistor Circuits Manual No. 5	5/–
180.	British Semiconductor Survey	15/–
181.	22 Tested Circuits Using Micro Alloy Transistors	5/–
182.	"At A Glance" Radio Valve & T.V. Tube Equivalents	3/6
183.	How to Receive Foreign T.V. Programmes on your Television Set by simple Modifications	5/–
	Resistor Colour Code Disc Calculator	1/6
	Engineers' Reference Tables	1/6
	International Radio Tube Encyclopaedia—3rd Edition	63/–

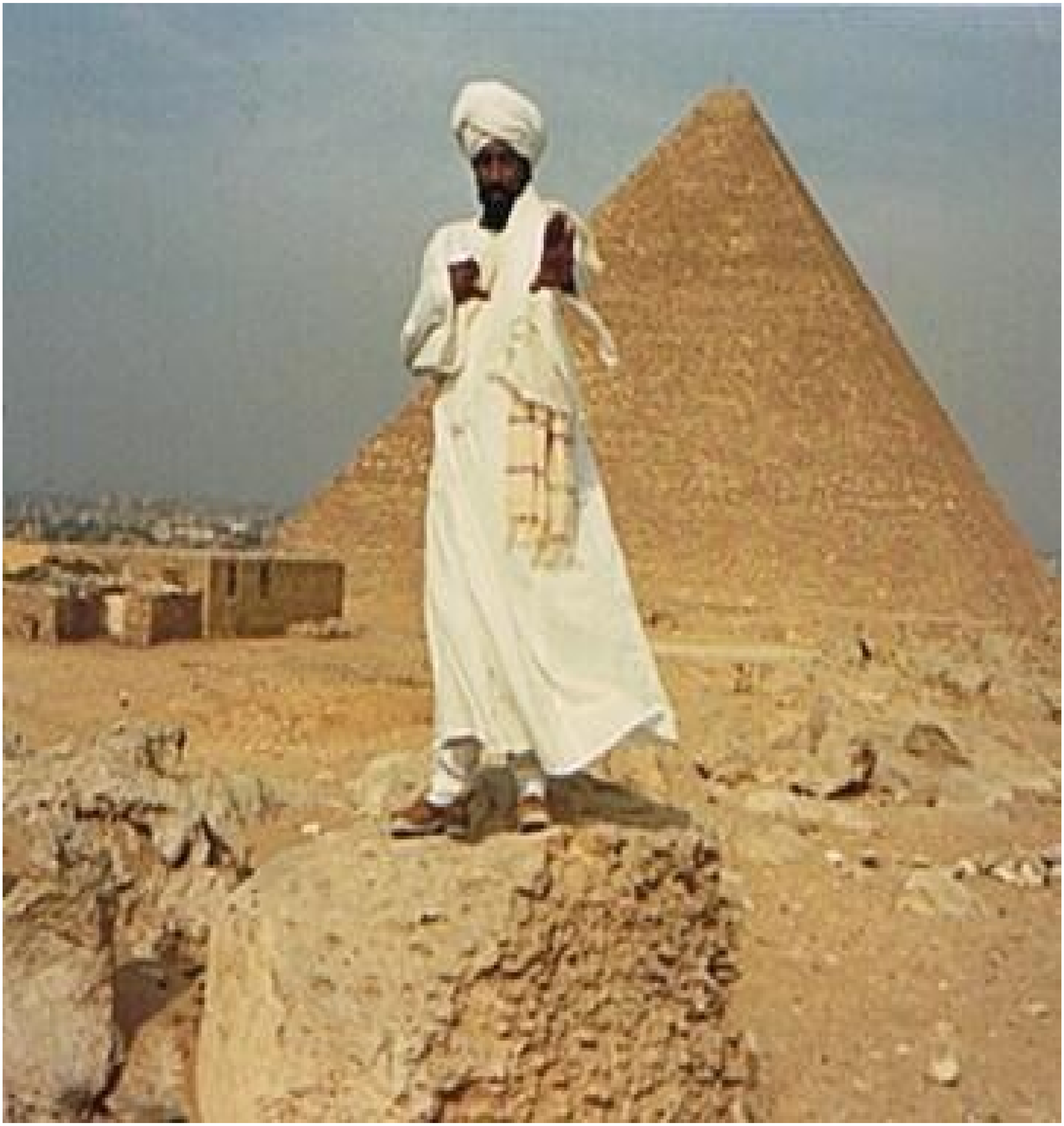

The Master Teacher The
Imperial Grand Potentate, Noble: Rev. Dr. Malachi Z. York 33°/720°.
Defines Success"

QUESTION: What is Success?

ANSWER: Success is the progressive realization of a worthy idea. If you have a goal and you are working towards that goal, then you're on the path to success.

"A 30 Day Program "

All One Has to Do Is Apply The Method To Their Life And Your Life Will Never Be The Same Again!

~~~~~~``

SAY This for 30 days straight

I plant positive thoughts in my mind

I am practicing this for 30 days straight.

I am doing a 30 day practical method.

I am applying this method for 30 days straight.

This is going to be easy.

This will alter my life for the better and for ever.

I think and I relax.

I am writing down my want list on a card.

I am imagining and concentrating on my want list for 30 days straight and 3 times a day.

I am looking at this want list for 30 days straight and 3 times a day.

SAY This for 30 days straight

I want (state wish or want)____?

This is my reality the moment I think about it

____
~~~~~~

I create positive habits.

I always think positive.

I take control of my mind.

My mind will only think of what I allow it to think.

I Do more.

I think more.
I give myself more then I ever did before.

I don't worry about how I am going to achieve my goals.

All The answers will come to me.

I think of all the reasons I can be successful.

I write out the persons that I love to become

I am like ____
I can not live no other way after this right knowledge.

I my whole life to live in a superior way.

I am super-strong mentally.

I am a super-strong physical .

I am super-strong 9 Ethereally

I am cured of all mental diseases, namely ignorance.

I am a Supreme Being of the pure radiant Green Light.

I change my mental environment and I change my life.

I am a superior thinker.

I am bred to be superior in mind.

I am programmed by my superior mental environment

I am educated In a Superior mental environment.

Named Nuwaubu/Wu-Nuwaubu

I am physically and spiritually free.

I am a supreme being.

I am goal-minded.

I achieve anything I want to achieve in my life.

I am financially independent.

I know how to use my superior knowledge to become super-successful.

I get everything I want in my life.

I have the super-magic touch.

Anything I touch turns Into success.

I continue to become more & more successful in my life.

I always think big and successful.

I am a super-successful supreme being.

I am progressive In the realization of my worthy ideas.

I do everything in any field I choose to.
I am successful in anything I choose to do, because that's what I want to do.

I propagate the 'Message of the Most High'.

I set goals.

I am going to ....(state action)____.

I am going to ....(state action)____.

I am going to ....(state action) _______________________________________

(set date for deadline).

____

I will succeed.

I will reach all of the goals that I set for myself.

I set out to achieve all of them.

I want to become a "Supreme Being "

I am a "Supreme Being "

Success is anybody that's doing a pre-determined job.

I will (state action)____.

I am (Determine)____.

I posses a burning desire to be what I think about all the time.

I create all my opportunities in my life.

I know where I am going.
I have complete plans.
I know my directions.
I know my destination.
I know where I am going all the time.
I have the superior guidance.

I am a superior sales man.
I have no limits in ALL

I always achieve.
I sell anything, because I want to.
I sell everything.
I believe in what I am selling.

I don't hurt anybody.

I don't displease anybody.

I make everything I do marketable to all.

I am superior competitor.
I am a super-winner.
I super-succeed in my life.
I really overstand this principle completely.

My life is altered for super-success

The Bes things are always attracted to me.

Everything I want, I get.

From now on, I won't have any problems.

I won't have any worries.

I don't have any doubt, nor any fears in my life.

~~~~~~
"THIS IS THE KEY!!"
SAY This for 30 days straight

I am what I think about all the time.

(say what you want to think about All the time)

<<<here>>>

____ ________

Everything will come if I only be patient and wait for it.

I am completing my purpose.

I alter my life for The Bes.

I alter my attitude to a superior state of mind.

I am a superior mind.

I care enough for super results.

I will achieve it.

I achieve all

I am super-rich.
I am super wealthy
I am super-healthy.
I am super-intelligent.
I am super- knowledgeable on all subjects on earth and the whole boundless universe.
~~~~~~

Anything I think in my mind, I am it.
Anything I feel in my heart I am it.

My life is what my thoughts make it.

I posses positive, crystal clear and vivid pictures in my mind of what I want to do.

I posses positive, crystal clear and vivid pictures in my mind of what I want to become

I know my goals and purposes In my life.

I have set a target before me.

I am prepared.

I practice.

I am talented.

I am skillful.

I have the complete idea.

I achieve all of my goals when I think about them.

I am helping to save positive agreeable Nubian people

who don't believe in themselves.

Possible Nine-mind revision:
(I am helping to save positive agreeable Nubian people who don't KNOW themselves.)

I totally believe in myself.

Possible Nine-mind revision:
(I totally trust in my abilities within myself.)

I will instill this superior mind-concept in all of my children.

I am the greatest of the great.

My children are the greatest of the great.

I am a chosen child of Anu.

I have the superior-faith

I have super-determination.

I will succeed.

I succeed

I know all things are possible to me.

I think in only positive terms.
I get only positive things sent to me.

I believe and I succeed.

Possible Nine-mind revision:
(I am conscious of my ability to achieve success)

I know and I succeed.

I have no doubt in myself.

I attempt all things I want & wish for.

I make everything happen for me in my life

I become what I think about.

I think about positive, concrete, worth-while goals and ideas.

I will always become what I think about.

I think about all the positive things I want to become in my life.

I become what I think about.

Anything I plant in my mind I become it.

I plant super-success in my mind and everything I want and wish for in my life becomes reality.

I concrete on only worth-while goals.

I always value my mind, my body, my soul and my spirit.

I don't take my mind, my body, my soul and my spirit for granted.

I plan my goals in my mind.

I am making the most important decision in my life.

I plan all my goals in my mind and I achieve them.

I picture myself as a Supreme Being.

I see myself as having already achieving all goals in my life.

I am outstanding in everybody's eyes.

I am a great, supreme being.

I find everything I need in my life to help all others.

I am super-rich in personality, fame and respect.

I plant any thought In my mind and I easily achieve it.

I care and work towards achieving my goals in life and everything

I plant in my mind becomes a reality.

That is Universal law.

This always works, because this is Universal law.

I am where I want to be.

I am guided by my Inner Being

I think, I think and I think

I am learnt in this superior science.

I overstand this Supreme science.

I know this supreme science works with the forces of nature

put your want and wish in the lines

I am _________________________.

I believe in_____________________________.

Possible Nine-mind revision:
(I focus on) _____________________________.

I have faith in_______________________________.

Possible Nine-mind revision:
(I strive towards) _______________________________.

I am in control of_____________________________.

I do the job that I love doing.

I only do super-great things in my life

I am A Supreme Being of the pure radiant GREEN LIGHT

Stainer and Barrett's Dictionary of musical terms

Sir John Stainer,
William Alexander
Barrett

Digitized by Google

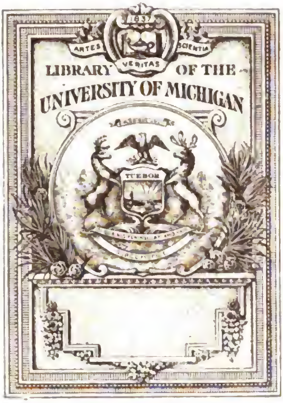
1837
ARTES
VERITAS
SCIENTIA
LIBRARY OF THE
UNIVERSITY OF MICHIGAN
TUEBOR

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

NEW AND REVISED EDITION.

STAINER AND BARRETT'S
DICTIONARY
OF
MUSICAL TERMS

REVISED AND EDITED BY

SIR J. STAINER,

M.A., Mus.D., Oxford; Hon. D.C.L. & Mus. Doc., Durham.

* * *

London: NOVELLO AND COMPANY, Limited.
New York: THE H. W. GRAY CO., Sole Agents for the U.S.A.
1898.

Digitized by Google

Music
ML
1C8
.S78
1898

LONDON:
NOVELLO AND COMPANY, LIMITED,
PRINTERS.

Digitized by Google

Transfer to
Music
11-9-65

Transf.. To
GL-Stacks
6-15-61

PREFACE.

The only way in which the Editor of a book of reference can show his gratitude to the public for demanding new editions on the exhaustion of the old, is by doing his best to keep its contents up to date. The reviser has done so; many articles have been entirely re-written or enlarged—*e.g.*, Copyright and Licensing, Madrigal, Mass, Opera, Pianoforte, Pitch, and others of less dimensions; also, many of those little slips which show such persistent survival in works of this kind have been removed. The Editor is grateful to many, whose names are among the authors and helpers, for their share in this last revision.

Oxford, *July*, 1898.

Class 12-9-30 T E.N.

207599

Digitized by Google

The following have contributed articles, or have in other ways rendered valuable assistance to the Editor :—

BOSANQUET, R. H. M., *M.A.*, Fellow of St. John's College, Oxford.

CHAMPNEYS, F. H., *M.A.*, *M.D.* (Oxford), *F.R.C.P.*, *F.L.S.*

CASSELL, Messrs.

CHAPPELL, W., *F.S.A.*

CHAPPELL, Messrs.

CLAYTON, HENRY R., *M.A.*, *LL.B.*

DONKIN, A. E., *M.A.*, *F.R.A.S.*, formerly Fellow of Exeter College, Oxford.

ELLIS, A. J., *B.A.*, *F.R.S.*, *F.S.A.*, *F.C.P.S.*, *F.C.P.*

GADSBY, HENRY, Esq.

HELMORE, Rev. T., *M.A.*

HIPKINS, A. J., *F.S.A.*

HULLAH, J., *LL.D.*

HUSK, W. H., Esq.

MCNAUGHT, W. G., *Mus. Doc.*, *F.R.A.M.*

MONK, W. H., *Mus. Doc.*

MURRAY, JOHN, Esq., *M.A.*

STAINER, Miss C.

STAINER, J. F. R., *B.C.L.*, *M.A.*

TROUTBECK, Rev. J., *D.D.*, *M.A.*, Precentor of Westminster.

WILLIS, HENRY, Esq.

Digitized by Google

DICTIONARY

OF

MUSICAL TERMS.

A. (1) The note called *Proslambanomenos* in the greater perfect system of the Greeks. The letter-name of *Mese*, the highest note of the middle tetrachord; and of *Nete*, the highest note of the upper tetrachord. [Greek Music.]

— (2) The first note of (*a*) the Hypo-Dorian mode, or church-scale, commencing four notes below the Dorian; (*b*) the Hyper-Phrygian mode, or church-scale, commencing four notes above the Phrygian; (*c*) the Eolian mode. [Greek Music.]

— (3) The next note above Gamma Ut, in the Grave Hexachord of the Guidonian system, where it is *A re*. Also, the first note of the acute and super-acute Hexachords, in which it is *a la mi re*. [Notation.]

— (4) The normal minor scale of modern music, so called because it is the relative minor of C. It is sometimes also named the *natural* minor scale, because no sharps or flats are required in its signature.

— (5) The normal sound (*Ger. Normalton*), because the instruments of an orchestra tune to this note, as given by the Oboe or Organ.

— (6) The key-note of the major scale which has three sharps for its signature.

— (7) The name given to a string tuned to the sound A. The A-string of a violin is its *second* string; of a viola, its *first* string; of a violoncello, its *first* string; but on this instrument it is one octave lower in pitch than the A-string of a violin or viola; of a double-bass, its *third* string, which is two octaves lower in pitch than the first string of a violoncello; of a guitar, its *fifth* string. The string tuned highest in pitch is called the first string of an instrument; the next below it, the second; and so on.

— (8) The actual sound is in some systems represented by A (capital letter), while AA represents the note one octave below that sound, and AAA the note two octaves below it. Proceeding upwards, the note one octave above A is represented by *a* (italic); that two octaves above it by a (once underlined); that three octaves above it, by a (twice underlined), and so on. [Pitch.]

A (*It.*) At, by, in, for, to, with, towards, &c.; *a prima vista*, at first sight; *a battuta*, lit. by the beat, in strict time; *a tempo*, in time, &c. &c.; *q.v. sub vocibus.*

A (*Fr.*) For, by, on; *à deux mains*, for two hands, &c.; *q.v. sub vocibus.*

Abacus (*Lat.*) ἄβαξ (*Gk.*) Any square tablet; hence, a board on which calculations were made; hence, in music, a diagram of the notes with their names.

A ballata (*It.*) (1) In the ballad style. Applied to any song, duet, or chorus, constructed in the ballad or narrative form. Also to a song with a refrain in chorus, as is found in many old and modern English songs; for example, the song of which a portion runs:

> "Ich have house and land in Kent,
> And if you love me, love me now;
> Two pence half penny is my rent,
> Ich cannot come everich day to woo."
>
> *Chorus.*
>
> "Two pence half penny is his rent,
> He cannot come everich day to woo"

—is a song, *a ballata*.

(2) The term is also used to describe a vocal piece, that may be danced to. [Ballad.]

Abbandonatamente (*It.*) With self-abandonment, despondingly.

Abbandonarsi (*It.*) To abandon oneself (to the influence of the music)

Abbandono, con (*It.*) With self-abandonment, passionately.

Abbassamento (*It.*) A lowering. *Abbassamento di mano*, a lowering of the hand; *abbassamento di voce*, a lowering of the voice.

A battuta (*It.*) In strict time. This term is usually employed when a break in the time of a movement has occurred, and it is desirable to resume the original pace "by the beat."

WEBER.

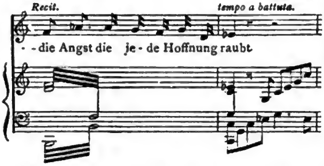

Abbellire (*It.*) To overload a plain melody with ornaments.

Abbellitura (*It.*) Ornament, embellishment.

Abblasen (*Ger.*) To sound a trumpet call.

Abbreviare (*It.*) To shorten.

Abbreviation. A system frequently employed in music, by which a portion of a technical term is made to stand for the whole. The following is a list of the abbreviations in most common use; the explanation of each term may be found on reference to the word itself in its proper place:

Accel. ... Accel^{o} ...	Accelerando
Acc. ... Accom ... Accomp.	Accompaniment
Accres......	Accrescimento
Adg^{o} or ad^{o}.	Adagio
Ad l...... Ad lib ...	Ad libitum
Affett^{o} ...	Affettuoso
Affrett^{o} ...	Affrettando
Ag^{o} ...... Agit^{o} ...	Agitato
All^{o}	Allegro
Allgtt^{o} ...	Allegretto
All'ott... All'8va...	All' ottava
Al seg	Al segno
And^{no}	Andantino
And^{te}	Andante
Anim^{o}	Animato
Arc	Coll' Arco, or Arcato
Arp^{o}.........	Arpeggio
Ard^{o}.........	Ardito
A t. ...... A tem ... A temp ..	A tempo
Aug	By augmentation.
B..........	Bass (voice) Bassoon Contra basso
B.C..........	Basso continuo
Brill.........	Brillante
C.B.	Col Basso
C.D.	Colla destra
C.S.	Col^{la} sinistra
Cad	Cadence
Cal	Calando
Can	Cantoris
Cant	Canto
Cantab ...	Cantabile
Cello	Violoncello
Cemb	Cembalo
Ch.	Choir organ
Chal	Chalumeau
Clar	Clarinet
Clar^{tto} ...	Clarinetto
Clar	Clarino
Co. So. ...	Come sopra
Col C. ...	Col canto
Col ott^{a} ...	Coll' ottava
Col Vo. ...	Colla voce
Con. esp.	Con espressione
Cor.	Cornet or horn
Cres., Cres^{o} Cresc	Crescendo
C.S.	Colla sinistra
C. 8va	Coll' ottava
C^{o} 1mo	Canto primo
Co. 1mo ...	Come primo
C^{to}	Concerto
D.	Destra, droite, right
D.C.	Da capo
Dec	Decani
Decres ...	Decrescendo
Delic	Delicatamente
Dest	Destra
Diap	Diapasons
Dim	By diminution
Dim	Diminuendo.
Div	Divisi
Dol	Dolce
Dolcis	Dolcissimo
Dopp. ped..	Doppio pedale
D.S.	Dal segno
Energ	Energicamente
Espr ... Espres ...	Espressivo
F. or for ...	Forte
Fag	Fagotto
Falset	Falsetto
Ff. or Fff...	Fortissimo
Fl.	Flauto
F.O. ... F. Org. ..	Full Organ
Fz. ...... Forz. ...	Forzando
G	Gauche
G. O. ... G. Org... Gt.	Great Organ
Grand^{o} ...	Grandioso
Graz^{o}	Grazioso
Gr.	Grand
Hauptw. Hptw. ... H^{k}	Hauptwerk
Haut.	Hautboy
H.C.	Haute Contre
Intro.	Introduction
Inv	Inversion
L	Left
Leg	Legato
Legg^{o}	Leggiero
L.H.........	Left hand
Lo.	Loco
Luo.	Luogo
Lusing. ...	Lusingando
M. ...... Main. ... Mano. ...	Manual
Maest^{o} ...	Maestoso
Magg	Maggiore
Man.........	Manuals
Manc ... Man^{do} ...	Mancando
Marc.	Marcato
M.D......	Mano diritta Main droite Manu dextrâ
M.G.........	Main gauche
M.M. ...	Maelzel's Metronome
M.M. ♩ = 92 ...	The beat of a crotchet is equal to the pulse of the pendulum of the Metronome said to be Maelzel's, with the weight set at 92.
M.P.	Mezzo piano
MS. ...	Manuscript Mano sinistra
Mus.Bac	Bachelor of Music
Men.	Meno
Mez.	Mezzo
M.V.	Mezzo voce
Mf. or Mff..	Mezzo forte
Mod^{to} ...	Moderato
Mus. Doc..	Doctor of Music
Ob.	Oboe, or Hautbois
Obb.	Obbligato
Oberst. ...	Oberstimme
Oberw. .. Obw. ...	Oberwerk
Oh. Ped....	Ohne Pedal
Org.	Organ
8va ...... 8^{a}	Ottava
8va alta ...	Ottava alta
8va bas. ...	Ottava bassa
P.	Piano
Ped..........	Pedal
Perd.	Perdendosi
P.F.	Più forte
Piang.	Piangendo
Pianiss. ...	Pianissimo
Pizz.	Pizzicato
Pmo. ... Pp. ...	Pianissimo
Ppp. ... Pppp....	Pianississimo
Prin........	Principal
1ma.........	Prima
1mo.........	Primo
4^{tte}	Quartet
5^{tte}	Quintet
Rall.	Rallentando
Raddol. ...	Raddolcendo
Recit.	Recitative
Rf., rfz., or rinf.	Rinforzando
R.H.	Right hand
Ritar.	Ritardando
Riten.	Ritenuto
S. or Sen...	Senza
𝄋	Segno
Scherz. ...	Scherzando
2da	Seconda
2do	Secondo
Seg.	Segue
Sem. ... Semp. ...	Sempre
7^{tt}	Septet
6^{tt}	Sestet
Sfz. or Sf.	Sforzando
Smorz. ...	Smorzando
Sinf.	Sinfonia
S. Int....	Senza interruzione
S.S. ... S. Sord .	Senza Sordini
Sos. ..... Sos^{t}	Sostenuto
Spir	Spiritoso
S.T.	Senza tempo
Stacc.	Staccato
St. Diap.	Stopped Diapason
String.......	Stringendo
Sw.	Swell organ
Sym.	Symphony
T.	Tenor, tutti, tempo, tendre
T.C..	Tre corde
Tem.	Tempo
Tem. 1^{o} ...	Tempo primo
Ten.	Tenuto
Timb.	Timballes
Timp.	Timpani
Tr.	Trillo

Trem.	Tremolando
3º	Trio
Tromb.	Trombi
Tromb.	Tromboni
T.S.	Tasto solo
U.	Una
U.C.	Una corda
Unis.	Unisoni
V.	Voce
V.	Volti
Va.	Viola
Var.	Variation
Vcllo	Violoncello
Viv.	Vivace
Vo. **Vno.** **Violº**	Violino
V.S.	Volti subito
Vni **V.V.**	Violini

There are other abbreviations employed in manuscript or printed music, the chief of which are as follow:

In time, a dash with a figure above signifies the length of the pause in bars, *e.g.*:

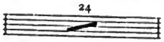

In notes, the trouble of writing a passage in full is saved by the use of abbreviations, *e.g.*:

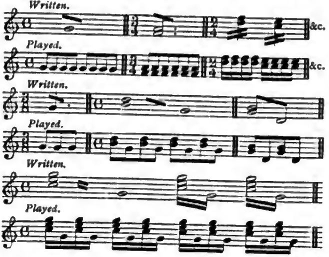

Repetition phrases are thus shortened:

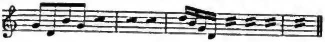

Abbreviations, by signs, of musical graces:

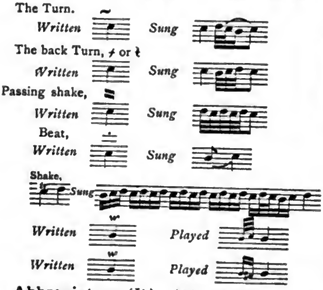

Abbreviature (*It.*) Abbreviations.

Abbreviazione (*It.*) An abbreviation.

Abcidiren (*Ger.*) (1) To use a system of solmisation. A series of exercises in which the names of the notes are used instead of words. (2) A method of instruction for sight singing.

Abbellare (*It.*) To decorate, ornament, or embellish.

Abbellimento (*It.*) A decoration, ornament, or embellishment.

Abendglocke (*Ger.*) The Curfew, evening bell.

A bene placito (*It.*) At pleasure. The singer or performer may alter the time, introduce ornaments, cadenzas, &c., according to fancy, or may use certain instruments specified, or not, without detriment to the effect required. [Chamber Music.]

Abenteuerlich (*Ger.*) Strange and uncouth. The music of the new German school is so called by the unthinking.

Abgestossen (*Ger.*) Struck off. Staccato.

AbgeleiteterAkkord (*Ger.*) An inversion of a chord.

Ab initio (*Lat.*) From the beginning. [Da capo.]

Abkürzungen (*Ger.*) Abbreviations.

Abnehmend (*Ger.*) *Lit.* taking away, *decres.*

Abrégé (*Fr.*) Shortened.

Abréger (*Fr.*) To shorten, curtail, abridge.

Abreissung (*Ger.*) A sudden pause.

Abrupt cadence. An interrupted cadence.

Abrupt modulation. A sudden change of key for which no preparation has been made. [Modulation.]

Absatz (*Ger.*) Cadence. [Cadence.]

Abschnitt (*Ger.*) Section. [Form.]

Absetzen (*Ger.*) } To render music '*staccato.*'
Abstossen (*Ger.*) }

Abub. [Ambubajæ.]

Abwechselnd (Ger.) Alternating, *mit abwechselnden Manualen*, alternately from one manual to another.

Académie de Musique (*Fr.*) An academy of music.

Académie Royale de Musique (*Fr.*) The opera house in Paris.

Académie Spirituelle (*Fr.*) A concert or performance of sacred music.

Academy of Music. A name given to an organised society of performers and teachers of music; originally applied to the Royal Academy of Music, founded 1824 in London.

A cappella, *or,* Alla cappella (*It.*) (1) In the church style; vocal pieces unaccompanied, especially those of the Italian school, because the music sung in the Sistine Chapel was never accompanied by instruments.

(2) Church music in a duple time (two or four minims in each bar), this being for a long period considered more ecclesiastical than triple measure.

Digitized by Google

A capriccio (*It.*) At will, according to individual fancy.

Acathistus. A hymn of praise, sung in the Greek Church in honour of the Blessed Virgin, upon the Saturday in the fifth week in Lent, by the whole congregation standing.

Accarezzevole (*It.*) (Lit. flatteringly.) Agreeable, pleasing; used occasionally to describe the anticipation of notes.

Accarezzevolmente (*It.*) In a caressing style.

Accelerando *or* **Accelerato** (*It.*) Gradually increasing the pace.

Accent (*Fr.*) A sign used in old French music for the Harpsichord:

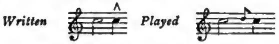
Written
Played

The custom of employing a variety of ornaments in harpsichord music, arose from the fact that the instrument was not capable of sustaining tone without the use of repeated touches.

Accent.—In its ancient and widest sense, a sign placed over a syllable to indicate the elevation of the voice when pronouncing it. Hence, the term came to imply a raising upwards of the voice in the scale series from the monotone or note of recitation, to a sound of higher pitch. By using various forms of accents, different elevations of the voice were obtained, until a rude sort of chant resulted. The most ancient known accents, those of the Hebrews as found in the Pentateuch, Psalms, and Book of Job, were interpreted only by tradition, not according to definite rule. Unfortunately therefore in comparing the musical rendering of them as taken down in modern notation by authors in various parts of Europe, it will be found that their original force is now quite unknown, and that the various "foliations" which are supposed to represent their meaning, are the growth of many surrounding musical or other indirect influences.

2. In early Greek Church music, the accents had to a great extent, a plain and definite intention, and as they were fixed and adapted to various poems by John of Damascus in the 8th century, and their use has been kept up in its purity to this day in Byzantine music, a very good insight into early forms and principles of notation can be obtained by their study.

3. In mediæval music the term *accent* was also applied to musical notation, the first two and most common of the signs being the *accentus acutus*, and *accentus gravis*. Afterwards the term came to signify the system generally, and so became synonymous with *neuma*, under which head information as to mediæval notation will be found. [Neuma.] [Old systems of *notation*.]

4. In plain song, the term accent, or *accentus ecclesiasticus*, was used to designate that system of movement of the voice, by learning the principles of which (*modus legendi choraliter*), a chanter could read collects, epistles, gospels, &c., from an un-noted book. Hence, it resolved itself into a series of rules relating to the inflexions or intonations of the voice on reaching a comma, semicolon, colon, full stop, and also a note of interrogation. But perfect uniformity is not to be found in these regulations regarding the *puncta*. According to its position in the sentence, or the interval covered by the movement of the voice, accent was said to be (1) immutabilis, (2) medius, (3) gravis, (4) acutus, (5) moderatus, (6) interrogativus, (7) finalis. The following are examples of these different species:

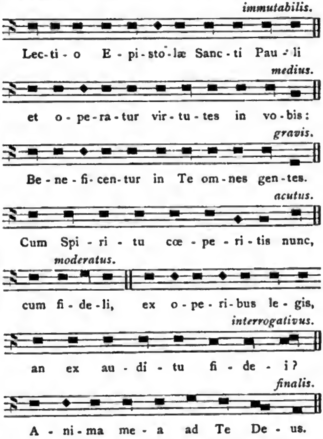
immutabilis.
Lec-ti-o E-pi-sto-læ Sanc-ti Pau-li
medius.
et o-pe-ra-tur vir-tu-tes in vo-bis:
gravis.
Be-ne-fi-cen-tur in Te om-nes gen-tes.
acutus.
Cum Spi-ri-tu cœ-pe-ri-tis nunc,
moderatus.
cum fi-de-li, ex o-pe-ri-bus le-gis,
interrogativus.
an ex au-di-tu fi-de-i?
finalis.
A-ni-ma me-a ad Te De-us.

But according to some authors, the epistle should be on monotone, except at a point of interrogation, *e.g.*:

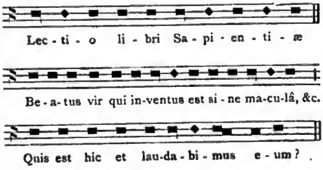
Lec-ti-o li-bri Sa-pi-en-ti-æ
Be-a-tus vir qui in-ventus est si-ne ma-cu-lâ, &c.
Quis est hic et lau-da-bi-mus e-um?

But in some countries, the epistle is chanted with the greatest elaboration, the note above the reciting note being introduced before the full stop, and the whole of an interrogative sentence being recited on a note below the *ut*. But as these uses differ not only in various places, but according to the Church seasons, an exhaustive account is impossible.

In chanting the gospel, an *accentus medius* takes place at the fourth syllable from a full stop, or thereabouts, and also the *accentus interrogativus*.

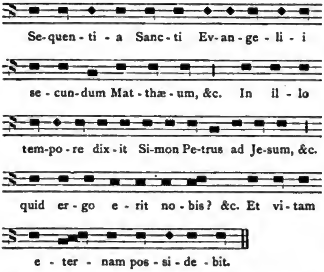
Se-quen-ti-a Sanc-ti Ev-an-ge-li-i
se-cun-dum Mat-thæ-um, &c. In il-lo
tem-po-re dix-it Si-mon Pe-trus ad Je-sum, &c.
quid er-go e-rit no-bis? &c. Et vi-tam
e-ter-nam pos-si-de-bit.

In the chanting of collects, a fall from Ut to La, or from Fa to Re takes place at a *punctum principale*, and from Ut to Si, or Fa to Mi once only at a *semipunctum*.

The accentus ecclesiasticus of lections and prayers must not be confused with those inflexions which tradition assigns to other parts of the service, such as confessions, proper prefaces, and lections of the Passion; all of which are to be found noted in authorised books. It should be remarked that the Belgian and French uses often differ much from that of the Romans, although uniformity in such things is without doubt desirable.

5. In modern music, accent is the stress which recurs at regular intervals of time. By the proper grouping of a series of accents, rhythm is produced. When music was indissoluble from poetry, a sign for marking accent was not needed, but it is necessary to point out in order to avoid confusion, that our word accent corresponds more to the ancient *ictus* than to *arsis*. For although the latter represented a raising of the voice, it did not necessarily occur on a long syllable, whereas it is considered a fault in modern music, if a short syllable occurs on an accented note. The position of the accent is plainly and simply indicated by upright strokes called bars [Bar], it being understood that the first note inside a bar is without exception accented. The measurement of the whole duration of the notes between these accents, is recorded at the commencement of a movement, and constitutes what is called the time-signature. In bars containing more than one group of notes, as is the case in compound times, other accents occur on the first note of each group, but they are not so strong as that falling on the first note of the first group. The latter therefore is called the primary or principal accent; the former secondary or subordinate. Besides these normal positions of accents, there are others which can be produced at any point by the use of a sign > or *sf*. An accent can also be displaced for a time from its usual seat by binding an unaccented chord to a like chord at a point of accent, and so preventing its repercussion; or, by both combined.

Beethoven's Sym., No. 3.

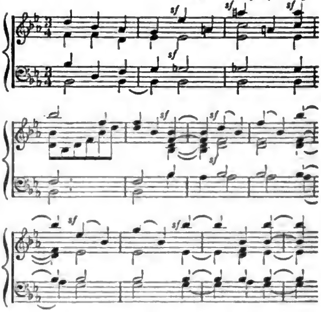

The throwing of the accent on to an unaccented part of the bar is called syncopation. A similar effect can be produced by a process the converse of the above, that is by making rests fall where an ictus is expected, *e.g.*:

Beethoven's Sym. No. 3.

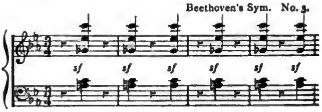

This intentional upsetting of our accepted notions of the expected position of accent is capable of a most remarkable and powerful effect. Heard by a musician just two centuries ago, its effect would probably not have been so striking, as he would have supposed the writer to have changed from triple to duple time, a constant habit in those days. Such accents are sometimes called *cross*, or *false*.

Digitized by Google

Accentuare (*It.*) To accent.

Accentuation. The act of accenting, or giving to certain notes their due emphasis.

Accessory stops and movements. Stops and movements acting only on the mechanism of an organ, not having pipes in connection with them, as, Couplers, Tremolo, Signal to the blowers; Composition pedals, &c.

Acciaccatura (*It.*) A short appoggiatura, *e.g.*:

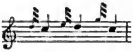

lightly yet clearly to be sung, or played.

Accidentals. Sharps, flats, or naturals, introduced into a piece of music, beyond those already in the signature.

A cinque (*It*). In five parts.

Accolade (*Fr.*) A brace, uniting several staves, as in pianoforte or organ music, or in a score.

Accommodare (*It.*) To tune an instrument in agreement with another.

Accompagnamento (*It.*) } [Accompaniment.]
Accompagnement (*Fr.*) }

Accompaniment *ad libitum. Accompaniment at will.* That is, one which can be played or omitted without injury to the harmonic construction of a composition, c.f. *obbligato.*

Accompaniment *obbligato.* [Obbligato.]

Accompaniment. A separate part or parts, for voices or instruments, added to a solo or concerted piece. Accompaniment may consist of a single simple instrument, such as a violin or flute, or a single compound instrument as an organ or pianoforte, a combination of selected orchestral instruments, or a whole band, or of voices in harmony. Accompaniments are those portions of a composition which are independent of the principal parts, and which are added to support, or to produce such effects as would be otherwise unattainable. Accompaniment should always be subservient to the chief part, so as not to overload or obscure it, and should be so constructed, that the voice or voices or solo instruments should be made to appear to the most favourable advantage. In the earliest records we possess, it is found that some sort of accompaniment was generally employed either to assist the voice, or mark the time or rhythm of the songs sung: of these many examples could be quoted if it were necessary. In the Bible, instances are mentioned in which singing is accompanied by musical instruments; but of the nature of these accompaniments we can form no definite idea, beyond the fact that, from a comparative knowledge of the instruments spoken of, it is presumed that little, if any attempt was made to gain independent effects. The ancient Greek dance, and the poetry of the tragedy, was always accompanied, at first with the lyre, afterwards with the flute.

Donaldson, in his "Theatre of the Greeks," says, that the Ancient Dorian Choral song, the Pœan, was originally accompanied with the Harp (φόρμιγξ.)

According to modern views, an accompaniment implies some construction in harmony, but the ambiguity with which the terms melody and harmony are employed by the ancient Greek and Latin, as well as by the more modern writers, has given rise to a doubt as to whether they had any knowledge of the art of combining certain concords according to such rules as we now possess. The science of harmony is of relatively modern growth, and the art of employing instruments in combination, as accompaniments to vocal music, is more recent than the growth of harmony.

It has been stated that dancing is of greater antiquity than singing, and that singing was generally the usual accompaniment to dancing. The same root supplying many words referring alike to dancing, singing, and playing, suggests a common origin for dancing and its accompaniment. Most of the words of ancient ballads are set to tunes that were danced to, and the practice is still observed of arranging words to melodies, that were originally intended to accompany the dance. So that the "new" idea of singing waltzes or other dance measures, is but a revival of the ancient practice. Nothing seems more natural than that those not actually engaged in dancing, but as interested spectators, upon the recurrence of a musical phrase should accompany it with the voice. This would form such a burden or chorus as that alluded to by Shakespeare:

> "Foot it featly here and there,
> And let the rest the burden bear."

Some simple instrument would be employed to play the tune, either with or without some means of marking the time or rhythm; a combination like that of the pipe and tabour, would form an adequate accompaniment to the dance, as the burden would to the song. As most of the mediæval learning and accomplishments came through the Church, it is very reasonable to assume that the methods of the Church would be reflected in the practices of every-day life. And as it is known that the organs employed in the service were so constructed as to be adapted only for the purposes of melody, it is scarcely probable that harmony should have been in use as an accompaniment to secular songs and tunes. Although instruments were employed as ac-

Digitized by Google

companiments to the voice, there is no reason for supposing that they were not at times engaged in the performance of purely instrumental pieces, but on the contrary, the numerous instances in which musical instruments are mentioned in ancient English poetry, show that the performers were not only accustomed to accompany singing with their several instruments, but they also imply that they were able to play independent pieces. [Mediæval musical instruments.]

In some instances we find that the accompanist was distinct from the singer, for example: "In Alwyni episcopi . . . et durante pietancia in aula conventus, sex ministralli, cum quatuor citharisatoribus, faciebant ministralcias suas." Regis. Prior: S. Swithini Winton (c. 1374). It cannot be ascertained whether the harpers did any more than play the same melody to which the poems were recited, or by the constant repetition of certain notes serve any other purpose but that of maintaining the pitch or rhythm.

The following quotation from David Lyndsay's poem, "The Dreme," 1579, implies the performance of a purely instrumental piece:

> "Thay beir ane ald stok-image throu the toun,
> With talbrone, trumpet, shalme, and clarione."

Whether this combination produced concord, or unison, it is impossible now to decide, for, as before stated, the loose manner with which the terms melody and harmony are employed helps to confuse, rather than to make clear all conjecture. For example:

> "And all above there lay a gay sautrie,
> On which he made on nightis melodye." *Chaucer.*

Again, in the "Lyfe of Saint Werburge," printed by Pynson in 1521, we read:

> "Certayne at each course of service in the Hall
> Trumpettes blewe up, shalmes, and claryouns
> Shewynge theyr melody, with toynes musycall"

and again in the same poem:

> "A singuler Mynstrell, all other ferre passynge
> Toyned his instrument in pleasaunt armony."

In the two first quoted examples, that which is called melody might well be harmony, and in the third, that which is called harmony might well be melody.

In nearly every list of instruments of minstrelsy, there are one or two pulsatile instruments spoken of — "Tymphans, tabours, nacaires," whose use was to mark the rhythm, sung, played, or danced to; but where these regular instruments were wanting, the clapping of hands, the beating of a stick upon a shovel, or the clashing of two sticks together, or the "ancient natural instruments of our islands," as Dr. Burney calls the tongs, marrow-bones and cleavers, salt-box and rolling-pin, with the hurdy-gurdy, were employed as accompaniments to rustic songs or dances. These "natural instruments" required little skill to use, beyond a correct sense of rhythm, though there are instances on record where some degree of science was brought to bear in their employment. The butchers of Clare Market, in the parish of St. Clement Danes in London, were at one time noted for their aptitude in playing tunes with marrow-bones upon cleavers of various sizes capable of sounding a scale of notes, in a manner somewhat similar to the hand-bell ringing in the North of England, but their performances were ultimately made independent, and therefore could scarcely be considered as accompaniments. Addison's description of a Burlesque musician, and his cultivation of strange instruments for accompaniment, may be read with interest in reference to this subject (Spectator, No. 570.) Skelton, in his description of RIOT, speaks of one, who

> "Counter he coulde, O LUX, upon a potte,"

probably meaning that while he sang the melody of the ancient hymn, "O Lux, beata Trinitas," he would beat upon the vessel he had been drinking out of, a part, in derisive imitation of the florid counterpoint sung by the monks during service. The poets and authors of the middle ages give no lucid information concerning the abilities of the musicians of their times, probably on the presumption, that because the method of their performances was well known and understood by every one, no particular description was needed. All our knowledge on the subject is derived from inference, and so it is assumed that the minstrels and later musicians, previously to the 16th century, had little, if any, knowledge of harmony; or if they had, they probably left the practice of it to the Church.

There is a further reason for supposing that the ancient minstrels accompanied their songs by playing the same melody which was sung, in the character of the majority of the instruments alluded to in old writings:—

> "Harpys, fythales, and eke rotys
> Lutys, ribibles, and geternes,
> Orguys, cytolis, monacordys
> . . . . trumpes, and trumpettes
> Lowde shaluys, and doncettes."

The drone of the bagpipe and the fundamental sound of the drum might suggest the formation of harmony; the character and construction of the earliest piece of harmony we possess, "Sumer is icumen in" offering curious confirmation of this notion.

All writers are agreed as to this being the oldest song with musical notes extant. Warton believed it to belong to the 15th century, but Sir Frederick Madden showed it to be at least two hundred years older; and judging by the character in which it is written, and

Digitized by Google

other evidence, fixed its date at about the year 1250. It is among the Harleian MSS. in the British Museum. The piece is arranged as a continuous melody; but by commencing that melody at certain indicated places, it forms a canon in the unison in four parts, with a "pes," a foot, or burden for two other parts.

The first attempts at accompaniment in harmony were arranged for the voice, in a manner that to our ears would be simply barbarous [Diaphony], [Descant], [Fa burden]. It was only by slow degrees that it was discovered that certain instruments were capable of producing complicated sounds forming harmony, and were therefore especially valuable for accompaniment. The chief among these instruments was the Lute, which appears to have been a favourite instrument in Chaucer's time, and to have continued in favour in one form or another, until the virginals and spinets, being of more convenient form and less trouble to keep in order, completely superseded it. [Tablature.]

The opportunity the virginals and spinets gave for the employment of both hands, tended considerably towards the improvement of accompaniments, as well as developing the power of execution for solo purposes. "Consorts of viols" were at first used only "for Cantilenas and tunes for dancing," though compositions were occasionally so arranged as "to be apt for viols or voyces," and when instruments were employed with voices they generally played the same notes as the voices, a practice observed until the time of Handel. No doubt an organist accompanying some of the services and anthems in Church, occasionally indulged in a little license in the matter of fingering, and introduced flourishes and cadences according to fancy, while the choir were singing the plain vocal harmonies. A copy of a portion of Gibbons's service in F [1583-1624] found among the old music in Magdalen College, Oxford, in which such variations are written down, is still preserved. Although the MS. belongs to the middle of the 17th century, there is no doubt but that it records a continuation of a custom of long standing. The character of this accompaniment may be seen by the following quotation:

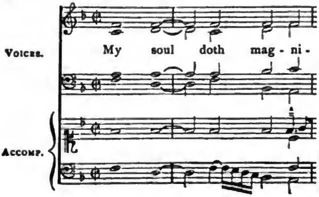

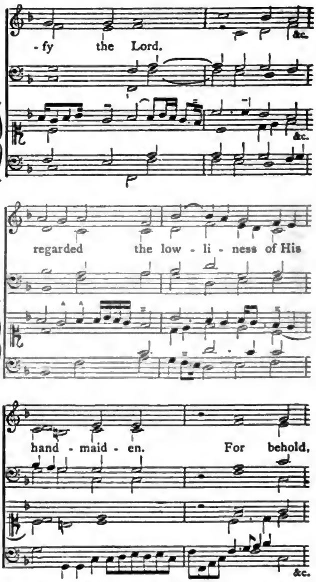
- fy the Lord.
&c.
&c.
regarded the low - li - ness of His
hand - maid - en. For behold,
&c.

A comparison of the vocal with the organ score, here shows that the ornaments introduced are such as involve little, if any departure from the rules of harmony, while they impart a distinct character to the accompaniment, such a character as most of the spinet or harpsichord music of the period possessed. The peculiarity of all old key-board stringed instruments, their feebleness of tone and their lack of sustaining power, probably suggested to the skilful player the necessity of breaking up the accompaniments to vocal music, ornaments and graces being considered perfectly legitimate so long as they did not interfere with the essential notes of the harmony as represented by the figured bass. But as in many cases it was doubtless deemed unwise to attempt display in the accompaniments during the singing, therefore all exhibition of skill on the part of the accompanist was reserved for the *ritornelli*, with which songs of the 17th century abound.

It was Monteverde [1568-1643] who, among other of his art-benefiting inventions, conceived the idea of constructing independent accompaniments for instruments, breaking up long notes into effective repetitions, and so imparting novel rhythms and striking instrumental figures. From his conception arose the Italian school of accompaniment—a school which influenced all musical teaching for more than a hundred years, and only decayed with the growth of a distinct style of instrumentation, the result of Italian ideas implanted in the German mind, after which the Symphony and Sonata, retaining names derived from their connection with vocal music, became independent and distinct, and accompaniment once more was lifeless and uninteresting, a mere adventitious aid. Indifference with regard to the part accompaniment should play, marked most of the music produced for a long period, but the power of the genius of Mendelssohn aroused new thoughts and new deeds, and care as great as that involved in the production of a vocal piece was employed in accompaniments, with true artistic effect. The followers of Wagner, in imitation of him, are striving to impart a new form to accompaniment, by giving to every instrument employed, a certain amount of independent work to do—a practice at once laudable and ingenious, but neither novel nor needful. It is laudable, for the reason that it is as well to interest the performer; it is ingenious, as no common amount of thought is involved in its production; it is not novel, for it was the practice of the Italian writers, and it is not needful, as accompaniment should always be subservient to the thing accompanied.

Purcell was among the first of the musicians in England who attempted to give colour to the accompaniments in the scores of his operas, but only occasionally introduced variety in the organ parts of his Anthems. The works of the musicians of the latter portion of the 17th and the commencement of the 18th centuries, show a desire to depart from the habitual rule, by giving independent melodies to the accompanying parts, specimens of which may be traced in Blow's Amphion, and contemporary works. The scores of J. S. Bach's "Passions" and other of his compositions, contain some excellent specimens of free instrumental parts in the accompaniments, and many of Handel's *obbligati* foreshadow the true use of orchestral colouring, a shadow to which substance was given by Haydn, Mozart, and Beethoven.

The influence of conventionality may be observed in accompaniments from the earliest period to the present day. Because it was the practice in olden time to accompany recitative upon the "cembalo," composers rarely thought of setting down more than a figured bass to these parts of their scores, excepting when the recitative took something of the form of the "soliloquæ;" and because the "maestro di cembalo" became an obsolete office in the orchestra, as soon as conductors considered it to be undignified to do other than direct with the bâton, it was necessary that some mode of accompanying recitative should be devised, and the figured bass was interpreted by a violoncello and double bass. Trumpets and drums were generally employed together, as much for rhythmical, as for any other purpose; and many other instances of the thoughtless practice of taking things for granted, and doing that which was held to be correct, because it was the custom, are to be found in well-known works of the lesser lights in music, the great thinkers constantly despising ordinary forms, and inventing new ones, which in their turn became models for imitation, and therefore standards of conventionality.

The bagpipe, fiddle, lute, cittern, virginals, spinett, harpsichord, pianoforte, harp, and guitar have each had their turns of favour and appreciation at several times.

The use of the guitar for an accompaniment became exceedingly popular during the latter part of the last century, to the injury of the makers of pianofortes, until Kirkman gave away a number of cheap guitars to milliners' apprentice girls, and so made the instrument unfashionable. The portability of both harp and guitar rendered them useful for the purposes of accompaniment, more especially at a time when the pianoforte was less frequently found in dwelling-houses than it is now; but improved mechanism and tone, together with smallness of cost, have made the pianoforte the most available accompanying instrument in private as well as in public; and the fact that one is to be found in nearly every house has contributed greatly towards the neglect of more portable means of accompanying songs and other household music.

Accompanist. The player who accompanies. The qualities necessary to form a good accompanist are, (1) that he possess a knowledge of reading music at sight, and of harmony; (2) that he should be acquainted with the style of the music performed; (3) that he should know the characteristics of those performers whom he is called upon to accompany; (4) that while playing with firmness and decision, he should not attempt to lead.

Accoppiato (*It.*) Joined or connected.

Accord (*Fr.*) (1) The series of notes to which an instrument is tuned, *e.g.*, *Accord du Violon*,

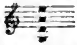

hence, *Accord à l'ouvert*, open strings, *q.v.*

(2) A chord. Concord. hence, *d'accord*, in tune.

Digitized by Google

Accord de sixte ajoutée (*Fr.*) The chord of the Added Sixth. [Added Sixth.]

Accordamento (*It.*)
Accordanza (*It.*) } Agreement in tune.

Accordando (*It.*) Tuning.

Accordare (*It.*) To tune.

Accordato (*It.*) Tuned.

Accordatura (*It.*) [Accord.]

Accorder (*Fr.*) To tune an instrument.

Accordeur (*Fr.*) A tuner.

Accordion. A simple musical instrument, of oblong form, invented by Damian, of Vienna, in 1829. The tone is produced by the inspiration and respiration of a pair of bellows acting upon metallic reeds or tongues. [Free reed.]

The first instruments had only four buttons, or keys, each of which acted on two reeds, making the compass one octave of diatonic scale, but with a separate arrangement, by which these notes might be accompanied with a tonic and dominant harmony. At first it was used only as a toy, but the introduction of a chromatic scale made the Accordion more capable of producing a varied melody and harmony, although the awkwardness of the form was always a hindrance to its use. The German Accordion or Concertina (*q.v.*) of hexagonal form made the principle of the accordion more acceptable. The English concertina (*q.v.*)and the harmonium (*q.v.*) are superior instruments constructed upon similar principles.

Accordo (*It.*) Concord, agreement, harmony.

Accordoir (*Fr.*) A tuning key or hammer.

Accresciuto (*It.*) Increased. Augmented as applied to intervals.

Acetabulum. An ancient instrument, originally made of earthenware, afterwards of metal, which, when struck with a rod, produced a sweet sound. [See an allusion to it in Boet. de Inst. Mus., Lib. i. cap. xi.]

Achromatic. Not chromatic.

Achtelnote (*Ger.*) A quaver ♪ The eighth part of a semibreve.

Achtelpause (*Ger.*) A quaver rest 𝄾

A chula (*Port.*) A dance similar to the Fandango, *q.v.*

Acoustics. The science which treats of the nature and laws of sound.

2. The sensation of sound consists in the communication of a vibratory motion to the tympanic membrane of the ear, through slight and rapid changes in the pressure of the air on its outer surface.

3. The mode of propagation of sound in air may be explained in the following manner. Suppose a small particle of fulminating silver to be exploded in free air; the air particles immediately contiguous are driven outwards in all directions by the explosion, their motion is almost instantaneously communicated to the adjacent ones, those first agitated coming at the same time to rest; the adjacent ones pass on the impulse in the same way to those at a greater distance, and so on; thus the explosion gives rise to what may be looked on as a rapidly expanding shell of constant thickness, containing at any instant between its exterior and interior surfaces a stratum of agitated air particles each one of which performs a single vibration to and fro during the passage of the shell over it; in other words the exterior and interior surfaces of the shell are at any time the *loci* of all those points at which the particles at that instant come under the influence of the impulse, and are left at rest by it respectively, so that its thickness depends both on the rapidity of their vibration and the rate at which they pass on the impulse, one to another.* Let us suppose now that immediately after the first explosion a second were to take place; then, in exactly the same way there would be a second pulse propagated in all directions. If a series of explosions at regular intervals were to take place, there would be a regular series of these expanding shells; and if the intervals were sufficiently small, the alternate changes of pressure, due to the successive collisions of the air particles against the tympanic membrane of an ear in the neighbourhood of the explosions would convey to the brain a sensation of a continuous note. Exactly the same thing occurs if, for a series of explosions, are substituted the vibrations of an elastic body; and it is, in general, by the latter means that all sounds, and especially musical ones, are produced. The motion of a sound wave must not be confounded with the motion of the particles which transmit the wave. In the passage of a single wave each particle over which it passes makes only a small excursion to and fro, the semi-length of which is called the *amplitude* of the vibration, the time occupied during one vibration being called its *period*.

4. The intensity of a sound is proportional to the square of the maximum velocity of the vibrating particles. It also approximately varies inversely as the square of the distance from the origin of the sound; for, supposing the latter to be produced at a uniform loudness, the same amount of energy has to be communicated to the particles contained within the external and internal surfaces of shells of the same thickness but of different *radii*. For

---

* The word vibration must be taken in its full sense, viz.: as meaning the whole motion of the particle during the time that elapses from the instant it sets off from its original position to the instant when it next regains that position, and is ready to start again over the same path.

Digitized by Google

example, if we take a shell of air whose internal radius is one foot, one of the same thickness whose radius is two feet will contain four times the quantity of matter; one whose radius is three feet, nine times the quantity, and so on. Thus the amount of matter over which a given quantity of energy has to be distributed augments as the square of the distance from the origin of sound, and therefore the amount of energy, or, what comes to the same thing, the intensity of the sound, diminishes in the same ratio.

5. At a temperature of *zero Centigrade* sound is propagated at the rate of about 1090 feet per second, and this speed augments about two feet per second for every additional degree of temperature; thus at 15° C. the rate of propagation would be about 1120 feet per second. The velocity of sound in air depends on the elasticity of the air in relation to its density. It is also directly proportional to the square root of the elasticity, and inversely proportional to the square root of the density. Now for a constant temperature the elasticity varies as the density, hence in this case they neutralise one another, and the velocity of the sound is independent of the density of the air.

6. One sound differs from another not only in quantity, but also in quality and pitch.* The pitch of a sound depends on the number of vibrations per second by which it is caused: the greater this number is the higher is the sound, and *vice versâ;* thus pitch is a more or less relative term, and it is therefore necessary to have some standard to which different sounds may be referred. This standard is so chosen that the middle C of the pianoforte shall be produced by 264 vibrations per second.†

7. Knowing the velocity of sound in air we can estimate the different wave lengths corresponding to notes of different pitch in the following manner. The wave length is the distance through which the sound travels while any particle over which it passes describes a complete vibration; hence, if we know the number of vibrations the particle performs per second, the required wave length will be found by dividing the number of feet over which the sound travels per second, by that number. Now, by means of an instrument invented by Cagniard de la Tour, and by him named the *syren*, the number of vibrations corresponding to a note of any given pitch can be determined very exactly. For a detailed account of this instrument and of its improvements by Helmholtz, the reader is referred to Tyndall's Lectures on Sound, p. 64; but to describe it shortly it may be said in its original form to consist of two equal discs, one forming the top of a hollow fixed cylinder into which air can be driven, the other capable of revolving concentrically upon it with the smallest possible amount of friction. A circle of small holes equidistant from each other is bored upon each disc and concentric with it; those in the upper disc being inclined slantwise to its plane, those in the lower being slantwise also but in the opposite direction; there are also arrangements both for driving a constant supply of air into the hollow cylinder, and for registering the number of revolutions the upper disc performs in a minute; thus, when the upper disc is so turned that its holes coincide with those of the lower, and air is forced into the cylinder, it will pass out through the perforations, and by reason of their obliquity will cause the moveable disc to revolve with a rapidity corresponding to the pressure; and each time that the holes of the former coincide with those of the latter a number of little puffs of air get through simultaneously, giving rise to an agitation in the surrounding atmosphere which spreads round in all directions in the way before described, and if the pressure of the air in the cylinder is sufficient, the series of impulses thus given will link themselves together, forming a continuous note.‡

Hence, to determine the number of vibrations per second, corresponding to a sound of given pitch, we have only to maintain such a pressure of air in the syren as will cause it to produce the same sound for the space of a minute, and note the number of revolutions registered in that time. Now, for every revolution of the upper disc, the same number of sound waves are propagated around as there are perforations, hence the whole number propagated in a second will be the product of the number of holes and number of revolutions per minute divided by 60; and this result will evidently be the required number of vibrations per second caused by the given sound.

To apply this to find the wave length corresponding to the note given by the open C string of the violoncello, we should adjust the supply of air to the syren till it gives a note of the same pitch. Supposing the number of holes in each disc to be 18, the number of revolutions per minute would be found to be 220. Hence the number of vibrations per second of the string, and therefore of the surrounding particles of air, would be $\frac{220 \times 18}{60}=66$. Supposing the temperature were 16° C the velocity of sound would be about 1122 feet per second, and the quotient obtained by dividing this number by 66 gives the wave length corresponding to that number of vibrations per second; that is, just 17 feet; the sound then will travel through this distance during the time the string takes to perform one complete vibration.

---

* For the cause of the different qualities of sound see § 16.

† That is, according to German pitch; at present there is no definitely fixed standard in general use in England.

‡ It should be remarked that the pitch of the sound would be exactly the same if there were only one perforation in the revolving disc, the number of holes merely serving to increase its intensity; if the number of holes in the revolving disc is less than the number in the lower one, those of the former must be situated so as all to coincide simultaneously with an equal number of the latter.

8. If the number of vibrations per second be increased, the pitch of the sound caused by them is raised, and *vice versâ*, as can easily be illustrated by driving more or less air into the syren, and observing the sound it produces. Dr. Wollaston has shown (Phil. Trans. 1820, p. 336) that if the number be increased, beyond a certain limit the sound becomes inaudible, although this limit is not the same for all ears, some persons being perfectly sensible of sounds inaudible to others. In general it is probable that no sound is heard when the number of vibrations per second exceeds 40,000; while on the other hand the perception of pitch appears to begin when the number of vibrations is somewhere between 8 and 32, the wave length being in the former case about 0·03 of an inch—in the latter ranging from 140 feet to 35 feet.

9. Sounds are primarily divided into two classes, musical and unmusical; the former being defined as those produced by regular or periodic vibrations, the latter by such as are irregular or non-periodic. These definitions require some explanation, since, by sounding together a sufficient number of notes sufficiently near in pitch, it is plain that we could produce as unmusical a sound as we pleased, although the components would be themselves due to periodic vibrations, and would be therefore musical. The answer to this is found in the fact that when two or more sets of sound waves impinge on the ear at the same instant, since each one cannot impress its own particular vibration on the tympanum contemporaneously with those of the others, the motion of the latter membrane must be in some way the *sum* of all the different motions which the different sets of waves would have separately caused it to follow; and this is what in fact does happen, *i.e.*, the vibrations due to each set combine and throw the tympanum into a complicated state of vibration, causing the sensation of the consonance or combination of the different sounds from which the sets of sound waves proceed.

Now the unassisted ear is only able to distinguish the separate notes out of a number sounded at once up to a certain point; beyond this it fails to distinguish them individually, and is conscious only of a confused mixture of sounds which approaches the more nearly to the character of noise the more components there are, or the nearer they lie to one another. A noise, then, may be defined as a sound so complicated that the ear is unable to resolve or analyse it into its original constituents.

10. As the character of a sound depends upon that of the vibrations by which it is caused, it is important to know of what kind the latter must be in order that they may give the sensation of a perfectly simple tone, *i.e.*, one which the ear cannot resolve into any others. Such a vibration is perhaps best realised by comparison with that of the pendulum of a clock when it is swinging only a little to and fro. Under these circumstances it is performing what are called harmonic vibrations, and when the air particles in the neighbourhood of the ear are caused by any means to vibrate according to the same law as that which the pendulum follows, and also with sufficient rapidity, a perfectly simple tone is the result. Such a tone is, however, rarely heard except when produced by means specially contrived for the purpose. If a note on the pianoforte is struck, the impact of the hammer on the string throws it into a state of vibration which, though periodic, is not really harmonic; consequently we do not hear a perfectly simple tone, but one which is in reality a mixture of several higher simple tones with that one which corresponds to the actual length of the string. The former are, however, generally faint, and become associated by habit with the latter, appearing to form with it a single note of determinate pitch. These higher tones are the *harmonics* of the string, and are produced by vibrations whose numbers per second are respectively twice, three times, four times, &c., as great as those of the fundamental tone of the string (§ 13). The same may be said of the notes of all instruments, including the human voice, which are usually employed for the production of musical sounds.

11. Since the consonance of two or more such simple tones always gives a more or less musical sound, and since also the ear is always more or less capable of resolving the latter into its components, the question naturally arises whether all sounds are not, theoretically at least, resolvable into simple tones. The answer to this is contained in a celebrated theorem due to the French mathematician Fourier. He has shown that any periodic vibration is the result of combining together a certain number of simple harmonic

Digitized by Google

vibrations whose periods are aliquot parts of that of the former; and we have conclusive reasons for supposing that, in the same way as a compound periodic vibration gives rise to a compound sound (§ 9), so the simple tones into which the ear resolves the latter are respectively due to the simple harmonic vibrations which, as the above mentioned theorem proves, make up the former.*

12. The theorem of Fourier referred to in the preceding article is of such great importance in all questions connected with acoustics that a few words illustrative of it may not be out of place.†

If a peg is fixed into the rim of a wheel capable of revolving about a fixed centre, and at right angles to the plane of the wheel, and if the latter is caused to rotate uniformly and is looked at edgeways the peg will appear to move up and down in a straight line, its velocity being the greatest at the middle of its course, and diminishing as it approaches each end. Under these circumstances the peg appears to perform harmonic vibrations.

Now suppose a second wheel, also furnished with a peg in its rim, is made to revolve about the peg of the first as an axis. If the latter is at rest the peg of the second will appear, looked at as above, to perform harmonic vibrations; but if the former is also caused to revolve these vibrations are no longer harmonic, but are the result of adding together the separate harmonic vibrations of the two pegs, in other words of *superposing* the harmonic vibrations which the second peg performs if the first wheel is at rest, upon those which the first peg performs when it is itself in motion. Now it is evident that by continuing this process indefinitely, and by giving the wheels different *radii*, and different uniform velocities of rotation, the final motion of the last peg looked at sideways as before, would be an exceedingly complicated one, and that an infinite number of different vibrations could be produced by varying the number, position at starting, *radii*, and velocities of the wheels, though it could not be assumed without proof that *every possible variety* could be so produced. This however is what Fourier's theorem asserts, provided that the velocities of rotation of the several wheels of the series are in the proportion of 1, 2, 3, 4, &c.‡ In other words, every periodic vibration is the resultant of a certain number of harmonic vibrations whose periods are one-half, one-third, one-fourth, &c., &c., that of the former.

13. A harmonic scale is formed by taking a series of notes produced by vibrations whose numbers in a given time are respectively as 1, 2, 3, 4, &c.

If we take as fundamental tone the open C string of the violoncello, the series of tones which with it form a harmonic scale will be as follows:—

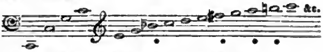

The notes marked with an asterisk do not exactly represent the corresponding tones; but are the nearest representatives which the modern notation supplies. All the notes of the harmonic scale can theoretically be produced by either a single string, or by a simple tube used as a trumpet. If we lightly touch the string of a violin, without causing it to come in contact with the finger board, at any one of a series of points dividing it into a number of equal parts, and excite it by means of a bow, it no longer vibrates as a whole, but separates into the number of equal vibrating segments which is the least possible consistent with that point forming one of their points of division; the latter remain stationary, or very nearly so, and are called nodes, their number being evidently just one less than that of the segments. It is plain that if the point of application of the bow be one of a series of nodes, no sound will be produced, provided, of course, the finger remains on any other of the same series, and this may serve to explain why it is sometimes difficult to bring out the higher harmonics of a violin, as the bow may, unconsciously to the performer, be passing exactly over one of the corresponding nodes. The first harmonic, as it is called, of the open string is produced by touching it while in a state of vibration at its middle point, and thereby dividing it into two equal portions, both of which vibrate twice as fast as the whole, and accordingly give the octave. The second harmonic, or the twelfth of the fundamental, corresponds to a division of the string into three equal portions, and so on. And generally, in order to produce the n^{th} harmonic the finger should touch the string at any one of the series of points which divide it into n equal portions.§ In

---

* A periodic vibration is any movement which recurs after equal intervals or periods of time, such as that of a uniformly working punching machine, or of the hammer of a clock bell when it is striking, and so on. It should be observed that though all harmonic vibrations are periodic, it is by no means the case that all periodic vibrations are harmonic. See foot note to § 3.

† For a complete discussion and demonstration of the theorem, the reader is referred to the work on Acoustics by the late Professor Donkin, published in the Clarendon Press series.

‡ The *order* in which the wheels are arranged with respect to their velocities is quite arbitrary.

§ That is supposing n to be a prime number (*i.e.* having no divisors). If such is not the case, it is plain that some points of the series when touched would give harmonics of lower pitch.

Digitized by Google

practice, however, the finger should always touch the string at the point of division adjacent to either end.

14. The harmonics of a simple tube used as a trumpet are the same as those of a vibrating string, *viz.*, the octave, twelfth, fifteenth, &c., and are produced by modifications of the breath and lips; but there is a great difference between the nature of the vibrations which produce sound, in the case of strings and pipes. In the former case the vibrations are executed at right angles to the length of the string, that is, are *lateral*, while in the latter they are in the direction of the pipe, or *longitudinal*, and are the vibrations of the air itself within it.

15. When an open organ pipe is sounding its fundamental tone, the particles of the column of air within it are all, more or less, in a state of vibration parallel to the length of the pipe, of which the intensity is at its maximum at the two ends, growing less and less towards the middle, where there is a node, that is, a point of no disturbance. The harmonics of an open pipe follow the same law as those of a simple trumpet, or vibrating string.

The fundamental note of a stopped organ pipe is an octave below the fundamental note of an open one of the same length. When it is sounding this note there is no node, and the first harmonic is a fifth above the octave, the second a major sixth above the first, the third a diminished fifth above the second, and so on. Or, more simply, the successive tones of the harmonic scale of an open pipe are produced by vibrations which are as 1, 2, 3, 4, &c., those of a stopped pipe by vibrations which are as 1, 3, 5, 7, &c.

16. It was stated (§ 10) that the sound of a vibrating string was in general compounded of a number of simple tones, and a well trained ear can detect a considerable number of them. If it were not for these harmonic components the tones of strings, pipes, of the human voice, or in short, of every instrument most generally used for the production of sound, would be flat and uninteresting like pure water. Each harmonic component is by itself a simple tone, and is due to the vibration of the corresponding segment of the string superposed upon that of the whole. The same statement applies, *mutatis mutandis* to pipes, whether open or stopped. That the harmonics of different instruments greatly influence their several characters is observable in the difference of the tones of a flute, and clarinet. A flute is an open pipe, a clarinet a stopped one; in the former, therefore, the harmonics follow the order of the natural numbers 1, 2, 3, 4, and in the latter the order 1, 3, 5, 7; —the intermediate notes being supplied by opening the lateral orifices of the instrument.

17. When two simple tones, that is (as explained above), notes deprived of all the harmonic components which under ordinary circumstances accompany them, are sounded together very nearly in unison, there are heard what are called *beats* succeeding one another at regular intervals, their rapidity depending inversely on the smallness of the interval between the two tones. Their origin may be explained thus: Suppose the tones to be produced by vibrations numbering 500 and 501 per second respectively, then every 500th sound wave of the former will strike on the tympanum at exactly the same instant as every 501st of the latter and will reinforce it; while at the 250th of the first the corresponding wave of the other will be just half a period in front of it. Now a sound wave consists of a condensed and rarefied stratum of air particles, and therefore the condensed portion of one wave here coincides with the rarefied portion of the other and neutralises it. Thus there will be an alternate reinforcement and diminution of sound, every second, from the maximum intensity when both waves impinge on the tympanum at the same instant to the minimum when they counteract each other as much as possible and *vice versâ*.

In the above case it was supposed that the number of vibrations of one tone were only *one* more per second than those of the other; but if the difference of the numbers had been *two*, for instance, then in one second the first tone would have gained two vibrations on the other, and there would have been two beats; and in general the number of beats per second is always equal to the difference between the two rates of vibrations per second.

18. In the preceding section, the cause of beats due to two simple tones of nearly the same pitch was explained, and it was seen that the number of beats per second was always equal to the difference of the numbers of vibrations per second of each tone; so that as the interval between them increased so would the number of beats increase in a given time. Hence it is obvious that if the interval became sufficiently large, the beats would succeed each other so rapidly as to become undistinguished. For instance, in the case of the fifth whose lower and upper tones are produced by vibrations numbering 264 and 396 per second respectively, the number of beats per second would be 132 and would therefore be undistinguishable—and still more so supposing the upper tone to have 397 or more vibrations per second; but, on the other hand it is a well-known fact, that if an imperfect fifth, octave, or any other tolerably simple interval is played on a violin or violoncello,

Digitized by Google

the beats are most distinctly heard succeeding each other at perceptible intervals—whereas according to what was said above they should occur so rapidly as not to be heard at all. Two explanations of this phenomenon have been given, of which by far the most simple is due to Helmholtz—and which here follows. It appears that when the tones are simple and at a sufficiently large interval the beats should occur too rapidly to be heard, whereas when the interval is played on a violin they are easily distinguishable. The reason of this fact is that in the latter case the tones are no longer simple but compound—and the beats which are heard are not due to the fundamental tones themselves but arise from two of their harmonic components which are nearly in unison. Suppose the ratio of the interval between the fundamental tones to be $\frac{m}{n}$, that is, let $\frac{m}{n}$ be the fraction, reduced to its lowest terms, which is formed by putting in the numerator the number of vibrations per second of the upper tone, and in the denominator those of the lower. Then it is plain that the n^{th} harmonic component of the tone m, will be of the same pitch as the m^{th} harmonic component of the tone n; for they will each have exactly mn vibrations per second. Now let $\frac{M}{N}$ be the ratio, expressed in the same way, of another interval, nearly, but not quite, equal to $\frac{m}{n}$; then the n^{th} harmonic component of M will have Mn vibrations per second, while the m^{th} component of N will have Nm. Now since $\frac{M}{N}$ is nearly equal to $\frac{m}{n}$, the difference between Mn and Nm will be a small number; and when the two notes are sounded together the number of beats per second will be equal to that difference.

For example, let $\frac{m}{n}$ be the ratio of a fifth, that is the fraction $\frac{3}{2}$, and let $\frac{M}{N}$ represent very nearly the same interval, say $\frac{397}{264}$; then the difference between Mn and Nm, or 794 and 792, is 2; hence if two strings tuned apart at an interval represented by $\frac{397}{264}$ are sounded simultaneously there will be two beats heard per second.

19. When the vibrations of the air due to a number of different sounds which co-exist at the same time are infinitely small, they are merely superposed one on another, so that each separate sound passes through the air as if it alone were present; and this law of superposition holds, though only approximately, until the vibrations have increased up to a certain limit, beyond which it is no longer true. Vibrations which give rise to a large amount of disturbance produce secondary waves; and it is to these that the phenomena of resultant tones are due.

Thus if two notes a fifth apart, for instance, are forcibly sounded together, a third tone is heard an octave below the lower of the two, and this ceases to be perceptible when the loudness of the concord diminishes. In general the resultant tone of any combination of two notes is produced by a number of vibrations per second equal to the difference of the numbers per second of the notes. This fact formerly led to the supposition that the resultant tone was produced by the beats due to the consonance, which, when they occurred with sufficient rapidity, linked themselves together so as to form a continuous musical note. If this were so it is clear that the resultant ought to be heard when the original notes are sounded gently as well as forcibly; and it was the failure of this condition that led Helmholtz to the re-investigation of their origin. These resultant tones have been named by him *difference tones;* he has also discovered the existence of resultant tones formed by the sum of the numbers of vibrations of the primaries. These *summation tones* as they are called cannot be explained on the old theory.

20. The theory of beats explains the law that the smaller the two numbers are, which express the ratio of their vibrations, the smoother is the combination of any two tones. When two simple tones are sounded together whose rates of vibration per second differ by more than 132, the beats, according to Helmholtz, totally disappear. As the difference grows less the beats become more and more audible, the interval meanwhile growing proportionately dissonant, till they number 33 per second, at which point the dissonance of the interval is at its maximum.

This, however, depends upon the position of the interval as regards its pitch. For it should be remembered that though the *ratio* of any given interval remains the same whatever the absolute pitch of its tones may be, yet the difference of the actual numbers of their vibrations, and therefore the number of beats due to their consonance, alters with it. And *vice versâ*, if the difference of the number of vibrations remains constant, the interval must diminish as its pitch rises. For instance, either of the following combinations

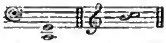

would give rise to 33 beats per second, since the numbers of vibrations of their tones per second, are 99·66, and 528·495, respectively. Now it is obvious that in the latter case the dissonance would be far greater than in the former.

Digitized by Google

The above explanation of the cause of dissonance is also due to Helmholtz, and completely solves a question which had remained unanswered since the time of Pythagoras, although that philosopher made the important discovery that the simpler the ratio of the two parts into which a vibrating string was divided, the more perfect was the consonance of the two sounds.

21. The sound of the piano, violin, &c., is only in a small measure due to the actual vibration of the strings themselves. The latter communicate their own motion to the sound board of the piano, and to the front, back, and enclosed air of the violin. In the latter instrument communication is made to the surrounding air from that within it by means of the *f* holes.

If a string were merely stretched between two pegs firmly fixed in a stone wall and caused to vibrate, scarcely any sound would be heard at all, owing to the mass and rigidity of the wall, which would refuse to be thrown into vibration by so small an amount of energy as that which the string would possess. On the other hand, the sound board of a piano readily answers to the vibrations imposed on it when the string is struck, and having a large surface in contact with the air, every point of which originates a system of waves, it causes a full and powerful sound.

22. The vibrations of straight rods may be either longitudinal or transversal. The former have not been generally employed for the production of musical sounds; the latter are such as take place when a tuning fork is struck, or when a musical box or triangle is played. In the case of a curved rod the vibrations are more complicated, but there is one interesting case, namely, that in which the curved rod takes the form of a circular ring. In this case the fundamental tone is obtained by suspending it horizontally by four strings attached at equidistant points in the circumference, and by lightly tapping it midway between any two. If the number of vibrations then given be $2n$ per second, those of the successive harmonics are proportional to $3n\sqrt{6}$, $4n\sqrt{13}$, $5n\sqrt{22}$, &c.

23. The nature of the vibrations of a bell may be partly inferred from those of a ring, as the bell may be considered as consisting of a connected series of rings of different diameters all vibrating simultaneously; thus the fundamental tone of a bell would cause it to divide itself longitudinally into four equal segments, corresponding to the four quadrants into which the suspended ring divides. The period of its vibrations could not, however, be similarly inferred.

24. The vibration of plates is not, musically speaking, a subject of much interest, as the only instruments which depend upon it directly for the production of their sounds, are gongs and cymbals, and the same may be said of membranes. Chladni was the first to show the positions of the lines of nodes on a plate, by clamping it horizontally in a vice, and causing it to vibrate by passing a violin bow over one edge, having previously sprinkled it with a little sand. The lines of nodes being those parts of the plate which, like the nodes of a string (§ 13), are not thrown into vibration, remain covered with the sand which collects there from the vibrating portions, and in this way very curious and interesting figures are produced.

Act (Acte, *Fr.;* Akt, *Ger.;* Atto, *It.*) A distinct division in the plot or design of a drama or opera, forming an incident complete in itself, but bearing reference to the general idea of the whole. Every dramatic plot naturally divides itself into three portions: the exposition, the development, and the conclusion; and this division would seem to point to the separation of a dramatic design into three acts, but where the piece is in four, five, or more acts, it will be found on examination that the tripartite division is essentially the same, greater prominence or care in detail being given to one or more of the sections. Thus the exposition may be spread over two or three acts, the development over one or two, and the conclusion or unravelling, reserved for the final act.

The classical trilogies—groups of three tragedies—were most frequently united by a common idea, each forming a complete incident, connected by a bond of sympathy, sentiment, or subject with the grand design.

Bartholomé Torres Naharro, of Torre in Spain, who wrote at the commencement of the 16th century, is said to have been the first who suggested the division of plots into acts, or *jornados*, although Cervantes claims the invention for himself. It is certain that Naharro's printed dramas are not so divided.

Donaldson, in speaking of the Æschylean Trilogy (the Agamemnon, Choephoræ, and the Eumenides), says, that the three plays mutually cohere, is plain; and as they were actually brought on the stage in sequence, they may be regarded as so many acts of one grand heroic drama. This is mentioned, in order to vindicate the practice of Shakspeare and other modern dramatists, in compressing into one drama an extensive cycle of human destinies; because the very objection that has been made to the practice is the alleged example of the ancients to the contrary.

Wagner's *Nibelungen* trilogy, though stated to be "new from end to end" in idea and design, bears a close affinity to the ancient Greek drama. For the subject is mythical

Digitized by Google

and "the mythical subject has a plastic unity; it is perfectly simple and easily comprehensible, and it does not stand in need of the numberless small details, which a modern playwright is obliged to introduce to make some historical occurrence intelligible. It is divided into a few important and decisive scenes, in each of which the action arises spontaneously from out of the emotions of the actors; which emotions, by reason of the small number of such scenes, can be presented in a most complete and exhaustive manner." In many modern operas, the division of the work into acts is made, less with reference to dramatic principles, than to the requirements of the stage-manager.

Act Music in Oxford. Cantatas composed by the Professor of Music, to words written by the Professor of Poetry, and performed at grand commemorations in the University.

Act tunes. [Playhouse tunes.]

Acte de Cadence (*Fr.*) Certain chords by means of which the final cadence is introduced.

Action. The mechanism of an organ or pianoforte, or other compound instruments.

Acuta. The accent attached to certain letters in the Greek system of musical notation, thus, M´

Acuta (*Lat.*) An organ stop of high pitch.

——— (*It.*) High, sharp.

Acutæ claves, acuta loca, acutæ voces. Those keys, places, and sounds which lie between 'alamire acutum' and 'alamire superacutum' of the hexachords, that is between little a

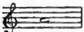

and A

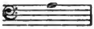

Acute. High as to pitch; opposed to grave.

Adagietto (*It.*) A diminutive of Adagio.

Adagio (*It.*) Slowly; also a name given to a movement written in that time.

Adagio assai
——— **di molto** } Very slowly.
——— **cantabile.** Very slow, and sustained, as if being sung.
——— **patetico.** Slow and with pathos.
——— **pesante.** Slow and weighty.
——— **sostenuto.** Slow and sustained.

Adagisissimo (*It.*) (superlative of Adagio). More than usually slow, very slow indeed.

Added Sixth, Chord of the. This dissonant combination of sounds is so called because it has the appearance of a common chord of the Subdominant of the key in which it occurs, with the addition of a sixth from the bass note, *e.g.*:

Ex. 1.

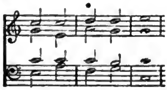

The above example being in the key of C, the Subdominant of the scale is F, and a common chord of F consists of F, A, and C, to which is found added at * the sixth of the bass note: namely, D. Although as a mere name, the expression *added sixth* may not be without value, it is by many authors considered very doubtful whether the notes, F, A, C are really the constituents of a Subdominant common chord; and the fact that the apparent fifth of this chord (C) is nearly always treated as a discord, and made to descend, is rightly cited as the cause of their doubt.

This naturally leads to the second explanation of the chord, which is, that it is an inversion of the chord of the seventh on the supertonic, *e.g.*:

Ex. 2.

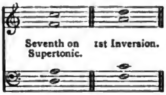

But, the system of constructing chords on every degree of the scale, though once much adopted, is daily losing ground; and justly, because by it, the particular progression of each component note of a chord, either has to be ignored, or else treated of with an amount of detail which is puzzling to the student, owing to the impossibility of laying down several laws as to the usual progression of stated intervals.

A third explanation is, that it is a *dominant* chord—consisting of the fifth, seventh, ninth, and eleventh from that root, *e.g.*:

Ex. 3.

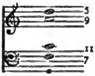

The objection to this is, that the bass note of the chord (F), in nine cases out of every ten, *ascends* in the resolution, as seen in Ex. 1. But, on the other hand, it will be found that the ninth (A), and the eleventh (C), descend properly, and it is a fact well known to careful analysts of harmony, that when several discords are heard simultaneously, the *regular* resolution of part of them often completely satisfies the ear; and also, that when the root or generator of a chord is omitted, there is more license in its treatment.

A fourth explanation has been offered: it is to the effect that the chord contains two minor sevenths, namely, F and C, derived from two roots, G and D respectively:

Ex. 4.

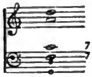

If both sevenths were properly resolved, consecutive fifths would ensue, therefore one (and generally the lower one) is made to ascend, on the principle just mentioned above. It is also urged that the lower seventh, F, is practically resolved, if the progression be followed to the cadence in C. In favour of this view, it is also stated that chords built upon the dominant of a scale *and* its fifth are not uncommon, as for instance chords of the augmented sixth, &c.

Whatever explanation be accepted, it is manifest that as this chord is made up of four notes, it can occur in as many positions, each note forming it being in turn placed in the bass, *e.g.*:

Ex. 5.

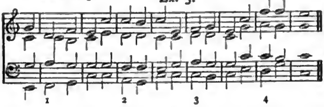

It is *resolved* generally as in the above example, but many other resolutions are occasionally met with, *e.g.*:

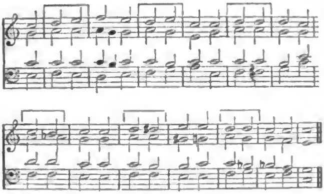

When the other positions of the chord are used, greater scope for varied resolution will be found.

The chord of the Added Sixth is also to be found in the minor series of chords amongst those authors who accept, though perhaps under protest, the exigencies of the tempered scale, *e.g.*:

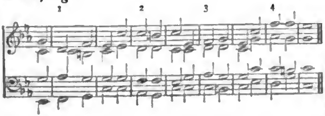

From this source a vast number of resolutions of the chord of the Added Sixth will be attainable, the tracing of which in the works of great masters will be found as instructive as interesting.

Additato (*It.*) Fingered; having signs pointing out what fingers are to be used for certain passages. [Fingering.]

Addition. The old name for a dot or point.

Additional accompaniments. Parts not in an original score, but added by another hand. Such additions may be made for the following reasons: 1, because the author accidentally left his score in an incomplete state: 2, for the supposed purpose of beautifying the original, by supplying parts for instruments either unknown or imperfectly known in the author's time: 3, to enable modern performers to play such parts as were intended for instruments now obsolete, or those of a similar tone now in use: 4, in order to compensate for the altered constitution of the orchestra, in which the number of stringed instruments is now larger in proportion to the number of reed wind-instruments, than formerly: 5, that, when for the sake of adding to the power or volume of tone other instruments must be added, they should be of varied qualities of tone instead of a mere numerical reinforcement of those already used in the original.

Additional keys. Keys added to enlarge the compass of any instrument.

Addolorato (*It.*) In an afflicted manner, sorrowfully.

A demi jeu (*Fr.*) With half the power of the instrument.

A demi voix (*Fr.*) With half strength of voice. [Mezzo voce.]

A deux (*Fr.*)
A due (*It.*)
For two voices or instruments. When the parts of two instruments are written on one line the portion to which this term is prefixed is intended to be performed by both in unison; the opposite term is *divisi q.v.*

A deux temps (*Fr.*) In common time of two in a bar.

A deux valse (*Fr.*) [Valse.]

Adirato (*It.*) In an angry manner.

Adiaphonon. An instrument of the Pianoforte class, not liable to get out of tune. Invented by Schuster of Vienna in 1820.

Adjunct notes. Short notes, not essential to the harmony, occurring on unaccented parts of a bar. [*c.f.* Auxiliary notes. Passing notes.]

Ad libitum (*Lat.*) At will. (1) In passages so marked, the time may be altered at the will of the performer.

Digitized by Google

(2) A cadenza *ad libitum* is a cadenza, the construction of which is left to the performer.

(3) Accompaniments *ad libitum* are additions to a piece, which may be performed, or not, at discretion.

(4) The word is also used to indicate the point at which a cadenza may be introduced in a concerto.

Adornamento (*It.*) An ornament, or grace.

Ad placitum (*Lat.*) At pleasure. A free part. A part added to a strict Canon, which does not come under the laws which govern that class of composition.

A due corde (*It.*) On two strings. [A deux.]

—— stromenti (*It.*) For two instruments.

A due voci (*It.*) For two voices.

A dur (*Ger.*) The key requiring three sharps to complete the major scale. The key of A major.

Ad videndum (*Lat.*) A species of counterpoint, which was written down or noted, as opposed to that which was *alla mente* or improvised.

Æolian harp. [Eolian.]

Æolian mode. [Eolian.]

Æolian piano. A piano having wooden bars, instead of strings, which, when struck by the hammers, produced a tone of peculiar quality.

Æolodicon. Æolodion. A musical instrument, the sounds of which are produced by the striking of steel springs by hammers set in motion by an ordinary key-board.

Æolomelodicon, called also a Choraleon; an Æolodicon having brass tubes over the metal springs, for the purpose of giving more power to the tone.

Æolopantalon, a pianoforte in connection with the Æolodicon.

Æquisonæ voces (*Lat.*) Equal sounds, but not unison; that is, such a consonant combination as, a note and its octave; or a note and its super-octave.

Aeusserst (*Ger.*) Extreme. *Äusserst rasch.* Extremely quick.

AEVIA. The vowels in the word Alleluia, used as an abbreviation of that word in Antiphons and endings of chants.

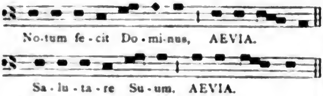

cf. EVOVÆ.

Affabile (*It.*) In a pleasing kindly manner.

Affannato (*It.*) In a distressed manner.

Affannosamente (*It.*) Restlessly.

Affannoso (*It.*) Mournfully.

Affetto, con; **Affettuosamente**; **Affettuoso** (*It.*) With affection.

Affinity. Connection by relation. Keys of affinity. [Relative Keys.]

Afflitto, *or* **con afflizione** (*It.*) Afflictedly, with sadness.

Affrettando (*It.*) **Affrettato** (*It.*) **Affrettore** (*It.*)	Hastening the time.

A fofa (*Port.*) A dance, like the Fandango, *q.v.*

Agevole (*It.*) **Agevolezza** (*It.*)	With facility and lightness.

Aggraver la fugue (*Fr.*) To augment the subject in a fugue.

Agilità, con (*It.*) With sprightliness.

Agilité (*Fr.*) Lightness and freedom in playing or singing.

Agilmente (*It.*) **Agilmento** (*It.*)	Cheerfully, in a lively manner.

Agitamento (*It.*) Restlessness.

Agitato (*It.*) An agitated or restless style of playing or singing, in which the time and expression are broken and hurried.

Agitazione, con. (*It.*) With agitation.

Agnus Dei (*Lat.*) [Mass.]

Agoge (*Gk.*), ἀγωγή. (*Lat.* ductus; *It.* conducimento.) The name of one of the subdivisions of Melopœia, *q.v.*, among the Greeks. The order in which successive notes of the scale followed each other, with regard to their pitch, in a melody. It is thus defined by Aristoxenus: *ἀγωγὴ μὲν οὖν, ἐστιν ἡ διὰ τῶν ἑξῆς φθόγγων ὁδὸς τοῦ μέλους.* There were three kinds of Agoge: 1st, *εὐθεια* (ductus rectus), when the melody proceeded from a grave to a higher sound by single degrees, *εὐθεῖα μὲν καλεῖται ἡ ἀπὸ βαρύτητος εἰς ὀξύτητα* (Aristides, Quin); 2nd, *ἀνακάμπτουσα* (ductus revertens), when a higher sound was followed by a lower; 3rd, *περιφερής* (ductus circumcurrens), when a modulation was introduced in an ascending, and afterwards, descending succession of notes, by making one of the notes, which was flattened in ascending, sharp on descending; or, *vice versâ*, *e.g.*:

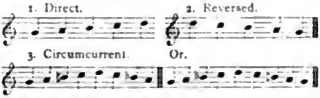

Agoge rhythmica. The succession of melodic sounds viewed with regard to their *accent*, and *rhythm*.

A grand chœur (*Fr.*) For the full choir or chorus.

A grand orchestre (*Fr.*) For the full orchestra.

Agrémens (*Fr.*) Turns, graces, and embellishments in harpsichord music.

Aigu (*Fr.*) Acute, high.

Air. In its modern sense, a tune, or the tune. The word air was formerly used to

Digitized by Google

describe dance tunes, as "Court Ayres, Pavins, Corants and Sarabands," also melodies for instruments; for before the invention of the Sonata, the music for concerts (*concertos*) of violins, consisted altogether of *airs* in three and sometimes four parts. The word air (*aria*), first used by Italian writers in the 16th century, was, when translated into English, represented by the word "fancy." Lord Bacon in his essay on "Beauty," uses the word *air*, and perhaps unintentionally describes its character thus—"the sweetest airs in music are made by a kind of felicity, and not by rule." The air was formerly assigned to the middle voice part or *medius*, corresponding to our tenor. The practice of giving the air to the soprano, or upper part, arose from the custom of the Italian theatres, where the "musico," while being supposed to sing the air in the tenor, really sang it in the soprano range. It was afterwards adopted by the composers of instrumental music, and the habit of giving the principal melody to the highest voice or instrument has continued until now. [Ballad.] [Sonata.] [Song.]

Ais (*Ger.*) The note A sharp.

Akkord (*Ger.*) A chord, as *Nonenakkord*, chord of the ninth, &c.

A la, Al, All', Alla (*It.* and *Fr.*) Like, in, at, in the style of.

A la même; à la mesure (*Fr.*) *A tempo.*

A la mi re. The name of the note *a* in the acute and super-acute hexachords of the Guidonian system. [Notation.]

Alamoth (*Heb.*) This word occurs in Ps. lxviii. 25. "First go the *sharim* (singers), then follow the *neginim* (kinnors), in the midst are *alamoth* (damsels playing on the timbrels)." Gesenius and others understand the word to signify treble music, "vox clara et acuta, quasi virginum." But, on the other hand, in 1 Chron. xv. 20 the names of *men* are given as players of "nebels on alamoth." It is one of the many obscure musical terms which are met with in the Bible. It however seems to have been associated with *nebels*, much as the expression *sheminith* is with *kinnors*, and may therefore be supposed to refer to the pitch or method of playing on those instruments.

Alarum, All'armi (*It.*) A call to arms.

> "Alarums sounded, and ordnance shot off."
> (*Shakspeare.*)

Originally a general shout; afterwards, a recognised signal by trumpets and drums.

Alberti Bass. A bass consisting of *arpeggios* or *broken harmony*, *e.g.*:

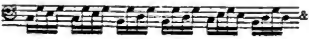

&c.

so-called after its reputed inventor, Domenico Alberti, who died in 1739.

Alcuna licenza, con (*It.*) With a little license; that is, the power of altering the time at will.

Aliquot tones. Overtones or harmonics. [Acoustics, § 9.]

A livre ouvert (*Fr.*) At sight.

Alla breve (*It.*) A direction that the notes are to be made shorter; that is, the pace taken quicker than usual. It is generally found attached to movements having four or eight minims in a bar, and is expressed in the signature by 𝄵. The following unusual sign for *alla breve* is found in the signature of one of the exercises "pour le Clavecin," by J. S. Bach, as published about 1760:

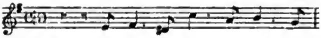

[Time.]

Alla caccia (*It.*) In the hunting style.

—— **camera** (*It.*) In the style of chamber music, *q.v.*

—— **cappella** (*It.*) [A Cappella.]

—— **diritta.** By direct intervals.

—— **hanacca.** In the style of the *hanaise*, a sort of polka, or polacca polonaise.

—— **marcia.** In the style of a March.

Alla mente (*It.*) A barbarous species of counterpoint in thirds and fifths, improvised upon the plain song, called in France "Chant sur le livre," and in England "Fa burden." This peculiar harmony is said to have had its rise in the 12th century, but it is probably older, as Hucbald, who was living about the year 880, describes it, as also did Odo, Abbot of Cluny, in his Enchiridion [c. 920.] While the use of this hideous harmony was encouraged by the church, musicians of feeling never failed to protest against it, therefore we find after Hucbald, Odo, and Franco of Cologne—Johannis de Muris, and others complaining of its use, and suggesting various measures for its reform, and by degrees paving the way to the modern system of counterpoint. The decree of the Pope John XXII, dated at Avignon 1322, had some effect in checking its use, but did not entirely suppress it, as it was sung as late as the middle of the 15th century.

Alla militare. (*It.*) In a military manner.

—— **moderna** (*It.*) In the modern method.

All' antico (*It.*) In the old style.

Alla polacca (*It.*) Like a polonaise.

—— **quinta** (*It.*) At the fifth.

—— **russe** (*It.*) In the Russian style.

—— **scozzese** (*It.*) In the Scotch style.

—— **siciliana** (*It.*) A species of melody in $\frac{6}{8}$, $\frac{12}{8}$, or $\frac{6}{4}$, having the longest note at the accented pulse. Handel's Pastoral Symphony is *alla siciliana*.

Alla stretta (*It.*) Bringing closer and closer, alike as to subject and movement.

Alla zoppa (*It.*) Lamely, halting, against time, syncopation.

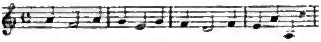

Allegramente (*It.*) Joyfully.

Allegretto (*It.*) (Diminutive of *allegro.*) (1) Slower than allegro. (2) A movement in this time.

Allegrettino (*It.*) (Diminutive of allegretto.) (1) Not so fast as *allegretto*. (2) A short allegretto movement.

Allegro (*It.*) (*Lit.* joyful.) Quick, lively. The word is occasionally employed to describe a whole movement of a quartett, sonata, or symphony. In music it is sometimes qualified as :—

Allegro agitato	(*It.*)	Quick and in an excited manner.
,, assai	,,	(*Lit.*) Fast enough. A quicker motion than simple allegro.
,, commodo or comodo	,,	An easy, graceful allegro.
,, con brio	,,	Quickly and with spirit.
,, con fuoco	,,	Rapidly and with fire.
,, con moto	,,	With sustained joyfulness.
,, con spirito	,,	Joyfully and with spirit.
,, di bravura	,,	A movement full of executive difficulties intended to exhibit the capacity of the singer or player.
,, di molto	,,	Exceedingly quick.
,, furioso	,,	Rapidly and with fury.
,, giusto	,,	In quick but steady time.
,, ma grazioso	,,	Lively and with graceful motion.
,, ma non presto	,,	Rapidly, but not too fast.
,, ma non tanto	,,	Quickly, but not too much so.
,, ma non troppo	,,	Lively, but not too fast.
,, moderato	,,	Moderately quick.
,, molto	,,	Very quick.
,, risoluto	,,	Lively and with firmness and decision.
,, veloce	,,	Lively and with speed.
,, vivace	,,	Lively and brisk.
,, vivo	,,	Quick and lively.

Allein (*Ger.*) Alone, as *Sanfte Stimmen allein*, soft stops only.

Alleluia. Latin for Hallelujah. (*Heb.*) Praise ye the Lord. An invitation to praise, used in every Christian community with varying regularity. St. Augustine says that the African Church used it between the feasts of Easter and Pentecost, and it is at that time that its use is more prevalent in the ritual of the Eastern and Western churches than at other seasons of the Christian year. In the Roman Catholic Church it is not employed from Septuagesima to Easter, but in the Anglican Church it is said twice at least every day throughout the whole year, the English form, "Praise ye the Lord," being substituted in the present Prayer-book for the ancient Hebrew word which was inserted in the first Prayer-book of Edward VI. The word Hallelujah being a short, musical, and rhythmical word, is frequently used by many anthem-writers—apparently without a just understanding of its meaning—to eke out an idea in music; but at other times the word has been set to music with sublime effect, as by Beethoven in his "Mount of Olives," and by Handel in his "Messiah." Carl Engel, in "The Music of the Most Ancient Nations," gives examples of melodies sung by the Copts, and the women of Syria, Arabia, and Persia, to the word.

Allemande (*Fr.*) Alemain, Allemaigne, Almain. A dance in duple time, said to have been invented by the French in the reign of Louis XIV. as a symbolical allusion to the newly-acquired German provinces. It was revived and frequently performed at the theatres during the time of the First Napoleon, during whose rule it became exceedingly popular. The measure was slow, and the steps were made in a rapid sliding manner as in the modern waltz, but there was no turning, only a peculiar entwining and unloosening of the arms of the dancers in the various steps. It is said by some that the Allemande was invented in the lesser provinces of Germany or Switzerland. Scarlatti, Corelli, Bach, Handel, and other composers of the period they represent, incorporated the measure of this dance in their Suites, Sonatas, and Lessons, in which it was written in common time of four crotchets in a bar.

The tradition concerning the origin of its invention, mentioned above, is very picturesque, but it is not founded in fact. Louis XIV. took Strasbourg by surprise in the time of peace, in the year 1681, and the treaty of Ryswick which confirmed his possession was not made until 1697. It is quite possible to believe that the fulsome flattery of his flippant court would magnify an act of rapine into a worthy victory, and that empty-headed adulators would find heels light enough to dance their joy at the same; but unfortunately for the story, there were Allemandes in existence before the time of the illegal seizure. In England, Almaines as musical compositions were published in 1662, in a book called "Courtly Masquing Ayres," and there were those by Lully, issued in France, bearing date 1670; the subjoined one, by Dumont, is even earlier, as it is to be found in "Meslanges à 2, 3. 4 et 5 parties, avec la basse continue, contenante plusiers chansons, motets, Magnificat, préludes et ALLEMANDES pour l'orgue et pour les violes; livre 1^{mer}; Paris, Robert Ballard, 1649." It is not even possible to refer the origin of the dance to the recovery of the towns in the Low Countries, as Turenne did not win them until six years after Dumont's publication.

The mention of the word as a dance is to be found in earlier writings than either of the above-mentioned, for there is a passage in Ben Jonson's play, "The Devil is an Ass," first acted in 1610, which proves the dance to have been known in his time:

> "He may, perchance, in tail of a sheriff's dinner,
> Skip with a rhyme on the table, from New-nothing,
> And take his *Almain*-leap into a custard,
> Shall make my lady mayoress and her sisters
> Laugh all their hoods over their shoulders."

The old Almains probably had leaping steps, as the foregoing passage, and another from George Chapman's "Alphonsus, Emperor of Germany," would imply:

> "We Germans have no changes in our dances,
> An Almain and an Up-Spring—that is all."

This play was printed in 1599.

ALLEMANDE.

Henri Dumont, 1610-1684.

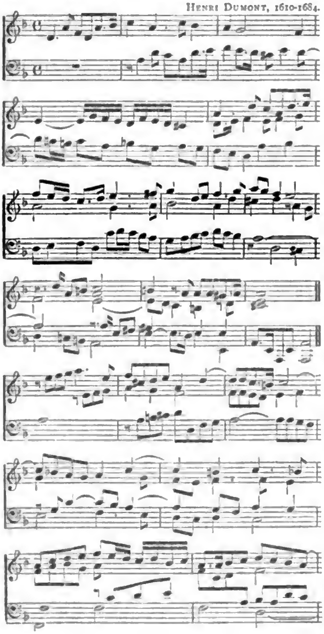

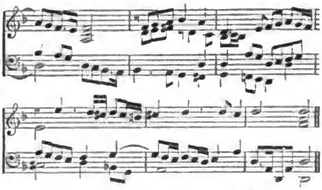

Allentamento (*It.*)
Allentato (*It.*) Yielding, slackening the time, giving way.

All' improvvista (*It.*) Without preparation, extemporaneously.

Al loco (*It.*) (*Lit. in the place.*) (1) A term which is used to counter-order a previous direction to play an octave higher or lower.

(2) A direction to a violinist to return from a shift to his previous position. [Shift.]

Allonger l'archet (*Fr.*) To prolong the stroke of the bow.

All' ottava (*It.*) At the octave, a direction to play an octave higher, or lower, than is written.

All' unisono (*It.*) In unison or octaves.

Alphorn. Alpine horn. A long tube of fir-wood played by the herdsmen of the Alps. It has the same natural series of sounds as a trumpet, but does not possess any means of reducing the tenth harmonic to the ordinary pitch of our scale. [Ranz des Vaches.]

Al piacere (*It.*) At pleasure.

Al rigore di tempo (*It.*)
Al rigore del tempo (*It.*) In strict time.

Al riverso (*It.*)
Al rovescio (*It.*) By contrary motion, that is, answering an ascending interval by one descending a like distance, *e.g.*:

Handel.

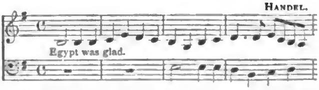

2. A phrase or piece which can be played backwards. [Per recte et retro.]

Al segno (*It.*) To the sign. *Dal segno* (*It.*) from the sign 𝄋. Directions that the performer must return to that portion of the piece marked with the sign 𝄋, and conclude with the first double bar which follows, or go on to the word *Fine*, or to the pause 𝄐.

Alt (*Ger.*) Alto voice, part, or instrument.

Alta (*It.*) High or higher. As, 8va *alta*, an octave higher.

Alterata (*Lat.*) A name given to those scales into which notes were introduced foreign to the old church modes.

Alteratio (*Lat.*) In mediæval music, the *doubling* of the value of a note. It was only the note falling on the second beat of a ternary measure that could be so doubled, and then only when the succeeding note was of the next higher denomination, and the ternary measure could not be completed in any other way. Thus the phrase

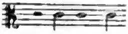

occurring in tempus imperfectum, prolatio major, would now be written—

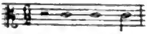

Alterato (*It.*)
Altéré (*Fr.*) Altered, augmented (with reference to intervals).

Alternamente (*It.*) Changing by turns.
Alternando (*It.*)
Alternativo (*It.*) To choose one of two ways of performing a passage.

Altgeige (*Ger.*) The tenor violin, the viola. [Viola.]

Altieramente (*It.*) Proudly, grandly, haughtily.

Altisono (*It.*) Sonorous, ringing.

Altissimo (*It.*) The highest. [Pitch.]

Altista (*It.*) A name formerly given to an alto singer.

Alto-basso (*It.*) An ancient Venetian stringed instrument, a Hackebrett, *q.v.* It was formed of a square box of pine-wood, supported on legs and strung with catgut. The player struck the strings with a sort of bow, which was held in the left hand, the right hand being engaged in holding a sort of flute or flageolet with which a melody was performed. The instrument was only used by the lower class of people, and is now obsolete.

Alto clef. The C clef, placed upon the third line of the stave, in order that the notes proper to the Alto voice may be conveniently represented:

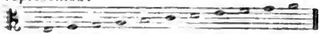

The Alto clef is used for the tenor violin or viola and the alto trombone. [Clef.]

Alto viola (*It.*) The tenor instrument of the violin family, called Alto, Tenor, or Viola. (*Ger.*) Bratsche or Altgeige. [Viola.]

Alto voice. Called also counter-tenor, when used by men, and counter-alto or contralto, when used by women. It is the deepest tone of voice among women and boys, to whom it may be said to be natural, and it is called the highest voice among men for lack of a better term to describe it. Properly speaking, the tenor voice is the highest man's voice, the alto or counter-tenor voice being entirely an artificial production, and simply a development of the *falsetto*. The register usually written for this voice lies between tenor G and treble C,

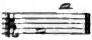

As the best notes of the alto voice are within the octave from B flat, those notes are most generally employed, for the higher notes are harsh and discordant, and the lower of small musical quality, and therefore ineffective. The alto voice in man is mostly formed upon an indifferent bass voice, and there is always a break between the chest and the head voice; this break varies between C and E 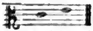 and the careful union of the chest and head qualities of voice, and the judicious employment of the "*messa voce*" are characteristic of every good alto singer. The alto voice is almost peculiar to English singers, not one of the continental nations possessing the capability of producing the quality or of appreciating it when produced; the consequence is, that there is no music written for this voice by any but English composers, and the majority of writers of the present day forming their style upon the foreign model, neglect and ignore the voice, disregarding its claim to usefulness, in places and at times, when and where female voices are unavailable. The value of the voice, its flexibility, sympathetic quality, and harmonious power, when carefully cultivated, are well displayed in cathedral music, and glee singing: a great number of melodious compositions by the most noted English writers, depend upon the alto voice for their proper effect. Many of the songs in Handel's oratorios were assigned to this voice, which are now, in consequence of the heightened pitch at present employed, sung by females: for instance, the part of Solomon in the oratorio of that name; of Barak and Sisera in the oratorio of "Deborah;" and of Daniel in "Belshazzar;" are each given to an alto voice. As this practice is of quite recent growth, it is but reasonable to conclude that Handel intended the music of the wisest king, and that of the two brave warriors to be sung by men altos, rather than by women, for the sake of appearance, if for no more powerful reason. The fact before alluded to, of the non-recognition of the voice by foreigners, has given an advantage to English musical literature not enjoyed by any other people, in the cultivation and sole possession of the Glee and the Anthem.

As many of the principal effects are obtained in these two species of composition through the medium of the alto voice, if only for the sake of the performance of the many noble

Digitized by Google

specimens of art in these two styles, the alto voice will always be cultivated in England until such time as the Glee and Anthem cease to exist. The cultivation of the Part-song has almost superseded the use of the alto voice in modern music, for the upper part in this class of composition is given to tenor voices, and the difficulty in producing the notes of the higher register so far influences the character of the music written, that many of the part-songs for male voices are of a bold, boisterous style, entirely different to that of the glee, which by reason of the peculiarity of the alto voice is of a more quiet character, depending in a great measure for its effect upon delicate and expressive singing. Many composers of eminence have completely ignored the alto voice, whether male or female, a quantity of music for Church use being written for treble, tenor, and bass, as by Cherubini and others.

In quality and power of expression the female alto voice is peculiar, and unlike any other voice. Its character is grave, tender, spiritual, and moving, and is admirably adapted to express emotions of dignity, grandeur, and piety. The male alto being an artificial voice, its usefulness is of limited duration, for when the singer is past fifty years of age the voice becomes harsh, reedy, nasal, and the break is painfully apparent.

Altposaune (*Ger.*) The alto trombone.

Altra, *fem.*, **Altri**, *pl.*, **Altro**, *mas.* (*It.*) Other, another, others.

Altschlüssel (*Ger.*) Alto clef.

Altviole (*Ger.*) [Alto Viola.]

Altzeichen (*Ger.*) The alto clef,

Alzamento di mano (*It.*) Raising the hand in conducting.

Alzando (*It.*) Lifting up, raising, elevating.

Amabile (*It.*) Lovely, gentle, tender.

Amabilità, con (*It.*) With gentleness, tenderness.

Amarevole (*It.*) Sad, bitter.

Amarezza, con (*It.*) With sadness.

Amarissimamente (*It.*)
Amarissimo (*It.*) } Very bitterly. In a mournful, sorrowful style.

Amateur (*Fr.*) A lover of music; one who pursues the practice of any art for pure love, in distinction to one who is engaged in its employment for pay.

Ambira (*Afr.*) A kind of drum or pulsatile instrument, made of wood, in cylindrical form, upon which a series of tongues of iron, cane, or wood are so arranged that they may be made to vibrate upon pressure. The Ambira is used by the negroes of Senegambia and Guinea.

Ambitus (*Lat.*) (*lit.* circuit.) The *compass* of an ancient church tone. The word is, however, used sometimes in a more extended sense than our word *compass*, as it is made to signify the proper steps which lie between the extreme limits of the tone—"Toni debitus ascensus et descensus." The rules (regulæ) which govern the *ambitus* depend upon the position of the *final* of the tone, and although much elaborated in mediæval treatises, their force seems to be that the proper *ambitus* should (regulariter) not exceed the octave (diapason) included between the highest note of the mode above the final and the lowest below it, except "by licence" (licentialiter). These dispensations only allow the introduction of three notes outside the *ambitus* in each direction, an authentic mode being allowed a descent to the lowest note of its corresponding plagal tone; a plagal tone being allowed to ascend to the highest note of its corresponding authentic. This fact, simple enough in itself, is cleverly put into a shape as unintelligible as possible in the following lines, framed for the supposed assistance of the student:

> "Undenis gradibus vult juste vadere prothus.
> Per sex et quinas claves vult deuterus ire.
> Octo tribus gradibus vult juste cepere tritus.
> Per sex quinque gradus juste capit ire tetrardus."

Ambo *or* **Ambon** (*Gk.*, *ἄμβων* from *ἀναβαίνω.*) A desk or pulpit. The raised platform in Eastern churches, on which the singers mounted when they sang. A canon of the Council of Laodicea (A.D. 360-370) decreed that no one should sing in the church besides the regular singers (*πλὴν τῶν κανονικῶν ψαλτῶν*), who ascended the *ambo* and sang from the parchment (*ἀπο διφθέρας*). This early attempt to bring about the separate performance of trained choir-singers did not obtain any favour in the Western churches of that period, and with the introduction of congregational song, the *ambo* became disused.

Ambrosian Chant. The system of church-song introduced by Ambrose, Bishop of Milan, in the fourth century. It formed the basis of the Gregorian system. [Plain Song.]

Ambrosian Hymn. The "Te Deum" so-called, because, by some, its authorship is attributed to Bishop Ambrose.

Ambrosian Te Deum. A musical setting of the Te Deum in plain-song, called after Bishop Ambrose because of its antiquity and the possibility of its having been the first setting of the Hymn after the introduction of properly-regulated song into Christian worship. Marbecke, who adapted it to English words, chose a simple setting, and a comparison between his version and others will be found interesting. Meibomius (in his well-known work, Ant. Mus. Auct. Septem. 1652), published it with Greek notation, as an exponent of his own principles, not as a copy of any authorized edition. [Plain Song.]

Ambubajæ (*Lat.*) Companies of immoral Syrian women, who attended festivals and gatherings among the Romans as minstrels. Their instruments were called *abub*, or *ambub*, whence their name.

Ambulant (*Fr.*) Wandering. Applied to strolling musicians.

Ame (*Fr.*) The sound-post of a violin, or other stringed instrument of its kind.

American Organ. An instrument having one or more manuals, and registers which control series of free reeds. It is, in its principles of construction, diametrically opposed to the Harmonium, for whereas in this, air is forced through the reeds from a bellows, — in the American Organ the action of the treadle exerts *suction*. When it is required to shut off a row of reeds, the putting in of the stop-handle places a thick air-tight covering of *felt* over the outside of the row of reeds, so that the air cannot be sucked through them,—the drawing of the stop, by removing this obstruction, allows the free passage of the air through the reeds. In short, in the Harmonium, air is blown from the bellows through the reeds; but in the American Organ, air is sucked through the reeds into the bellows. There are, however, other differences. The reeds of the American Organ are thinner than those of the Harmonium, and are slightly curved in shape, by which greater quickness of *speech* is insured. A very beautiful *undulating* tone is obtained by setting in motion a rotating *fan*, immediately above one of the rows of vibrators. This, by producing a variable pressure of air just outside the reeds, affects also their vibration as they draw it in. The mechanism which sets the fan in motion is called the *Vox Humana* stop. When there is an *Expression* stop it gives the player some control over the pressure of suction, and thus it is the converse of the stop of the same name in the Harmonium; but the effects which can be produced by its use are more striking in the latter instrument. The tone of American Organs is extremely melodious and sweet, but it does not *travel* well. For chamber music these instruments are eminently qualified, not only because of the character of their tone, but because they possess the enormous advantage of remaining for a longer period and under varying circumstances—*in tune*.

A mezza aria (*It.*) A compromise between Air and recitative. [Aria parlante.]

A mezza voce (*It.*) With half the strength of the voice. (2) The quality between the chest and head voice. (3) The subdued tone of instruments.

A moll (*Ger.*) A minor, the tonality of the relative minor to the key of C.

A monocorde (*Fr.*) On one string.

Amore, con (*It.*) With love, affection, tenderness, ardour.

Amorevole (*It.*) Affectionately.

Amorevolmente (*It.*) Tenderly.

A Moresco (*It.*) In the Moorish style.

Amorosamente (*It.*) Lovingly.

Amoroso (*It.*) In a loving style.

Amor-Schall. A horn of peculiar construction invented in the year 1760, by Kölbel, one of the musicians of the Emperor of Russia. It was intended as an improvement upon the ordinary French horn, but the introduction of the cylinder and valve system led to the disuse of the Amor-Schall. A duet for these instruments was composed by Cherubini, and dedicated to Lord Cowper.

Amphibrachys. A metrical foot consisting of a long between two short syllables, ⏑ – ⏑. [Metre.]

Amphimacer. A foot consisting of a short between two long syllables, – ⏑ –. [Metre.]

Amplitude of vibration. The distance from the point of rest of a particle, to either end of its journey, when a sound-wave passes over it. [Acoustics, § 3.]

Ampollosamente (*It.*) } In a bombastic,
Ampolloso (*It.*) } inflated style.

Ampoulé (*Fr.*) Bombastic.

Amusement (*Fr.*) A short and lively piece of music. (*It.*) *Divertimento.*

Anabasis (*Gk.*), ἀνάβασις, a succession of ascending sounds.

Anabathmi (*Gk.*), ἀναβαθμοί, the name given to certain antiphons in the Greek Church; so called because their words were selected from the Psalms called in the Septuagint ᾠδαὶ τῶν ἀναβαθμῶν (Ps. 120-134 in Eng. version), "Songs of degrees," the *gradual* Psalms of the Roman use.

Anakampsis (*Gk.*), ἀνάκαμψις, a succession of descending sounds.

Anakamptos. (*Gk.*) [Anakampsis.]

Anakara (*Gk.*) The ancient kettledrum. A larger sort was used for battle purposes, and there was a smaller drum which a woman could hold with one hand and beat with the other. [Drum.] [Nacchera.]

Anakarista (*Gk.*) A drum player.

Anapæst. A foot consisting of a long preceded by two short syllables, ⏑ ⏑ –. [Metre.]

Anaploke (*Gr.*) ἀναπλοκή. A combination of notes ascending the scale; opposed to καταπλοκή, a descending series of combined sounds.

Anche (*Fr.*) The reed in the mouth-piece of a hautboy, bassoon, &c., the name also applied to a reed in an organ. [Reed.]

Ancia (*It.*) [Anche.]

Digitized by Google

Ancora (*It.*) Again, once more, encore.

Ancor piu mosso (*It.*) Still quicker, more motion yet.

Andacht (*Ger.*) Devotion.

Andächtig (*Ger.*) Devotionally, devoutly.

Andamento (*It.*) A fugue subject of an extended nature, consisting of two contrasted portions:—

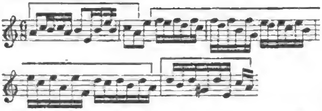

A portion of the latter half of such a subject is often used afterwards as an *attacco* or point of imitation which may be developed.

Andante. (1) (*Lit.*, walking.) At a moderate pace. (2) A complete movement, either separate, or part of a symphony or sonata.

Andante affettuoso. Slow, and in an easy pathetic style.

Andante cantabile. Slow, and in a singing style.

Andante con moto. With energy or emotion. Faster than andante.

Andante grazioso. Slow, and in graceful motion.

Andante maestoso. Slow, and with majesty.

Andante non troppo. Moderately, but not too slow.

Andante pastorale. Slow, graceful, and with pastoral simplicity.

Andantemente (*It.*) Easily, fluently, without interruption, in the manner of an Andante.

Andantino (*It.*) A diminutive of Andante, by some understood to mean *not so slow* as andante; by others, *rather slower* than andante.

Andante sostenuto. Moderately slow, and very smoothly.

Andar diritto (*It.*) Go straight on.

Andare in tempo (*It.*) To go in time. Keep to the time.

Anelantemente (*It.*) Ardently, eagerly, pantingly.

Anelanza
Anelito } (*It.*) Shortness of breath.

Anemochord. A variety of the Eolian harp, made by Jacob Schnell, in Paris, 1789. [Eolian harp.]

Anemometer. [Wind-gauge.]

Anesis (*Gk.*) ἄνεσις, from ἀνίημι, to loosen. (1) The progression from a high sound to one lower in pitch. (2) The tuning of strings to a lower pitch (ἄνεσις χορδῶν.)

Anfangsritornell (*Ger.*) Introductory symphony. [Symphony, § 4.]

Anfangsgründe (*Ger.*) Rudiments, principles, beginnings.

Angehaltene Cadenz (*Ger.*) An organ or pedal point. [Fugue.]

Angelica (*Ger.*)
Angélique (*Fr.*) } [Vox angelica.]

Angenehm (*Ger.*) Pleasing, agreeable.

Anglaise (*Fr.*)
Anglico (*It.*) } The English country dance, *q.v.*

Anglican Chant. [Chant.]

Angore (*It.*) Anguish, grief, distress, passion.

Angosciamente (*It.*)
Angosciamento (*It.*) } Sorrowfully, anxiously.

Angosciosissimamente (*It.*) With extreme sorrow.

Angoscioso (*It.*) Anxious, painful.

Anhang (*Ger.*) A coda. [Coda.]

Anima, con (*It.*) With animation, spirit.

Animato (*It.*) Lively, animated.

Animazione (*It.*) Liveliness, animation.

Animo, con (*It.*) With courage, spirit, dash, and fire.

Animo Corde (*It.*) [Anemochord.]

Animosamente (*It.*) Spiritedly, energetically.

Animosissimamente (*It.*)
Animosissimo (*It.*) } Exceedingly bold, energetic and spirited.

Animoso (*It.*) Lively, energetic.

Anklang (*Ger.*) Tune, harmony, accord.

Anklingeln (*Ger.*) To sound or ring a bell

Anklingen (*Ger.*) To accord in sound, to be in tune.

Anlage (*Ger.*) Indication of talent: the sketch of a musical thought; also the plan or design of a composition.

Anlaufen (*Ger.*) To increase or swell in sound.

Anleitung (*Ger.*) Instruction, guidance, direction, preface.

Anmuth (*Ger.*) Charm, sweetness, grace, agreeableness.

Anonner (*Fr.*) To stutter, to hesitate. To stumble in performing, to play in an unskilful style.

Anpfeifen (*Ger.*) To whistle at, to hiss at; in music, to condemn. [Fiasco.]

Ansatz (*Ger.*) (1) Attack, *q.v.* (2) The adjustment of the mouth to the position required for the production of the voice in singing. (3) The adjustment of the lips necessary for the proper production of the tone of wind instruments, as in French, "embouchure," and in English, "lipping," *q.v.*

Anschlag (*Ger.*) (1) Touch, or the production of tone upon such keyed instruments as the organ, pianoforte, or harmonium.

(2) The clash of a discord before resolution. [Percussion.]

Ansingen (*Ger.*) To welcome with song.

Ansprechen (*Ger.*)
Anstimmen (*Ger.*) } To sound, to sing, to give out tone.

Anstimmung (*Ger.*) Intonation, sounding, singing.

Antecedent. (1) A phrase or point proposed for Imitation. (2) Any passage which is answered. (3) The subject of a Fugue.

Anthem.—A composition for voices, with or without organ or other instrumental accompaniment, enjoined by the Ritual of the Anglican Church to be sung at Morning and Evening service, "in choirs and places where they sing." The words are generally selected from the Psalms, or other portions of the Bible, but paraphrases of Scripture, and words in prose and metre, of less authority, are sometimes used. It is the one ornament of the Service, reserved for the Choir, in which the congregation takes no part.

2. Anthems may be divided into various kinds, according to the character of the words; but with this division it is not our province to deal. The form of the music suggests four divisions, namely: the Full, the Full with verses, the Verse, and the Solo. When Anthems were accompanied with instruments other than the organ, they were formerly called Instrumental Anthems.

3. A Full Anthem, which is the earliest model, consists entirely of chorus, with or without Organ accompaniment. A Full and Verse Anthem is one in which certain parts are assigned to voices *soli*, with choruses to commence and conclude. A Verse Anthem is one that *begins* with portions intended to be sung by a single voice to a part, the word *verse* probably meaning a turn of thought to be forcibly or clearly expressed, a change of treatment or sentiment properly echoed in the style of the music. The words of the *verse* are often chosen from portions of Scripture other than the main body of the Anthem, by way of gloss. The chief voices on one side, Decani or Cantoris, usually sing the Verse, and the whole choir, both sides, the chorus or Full part. The character of the Solo Anthem is sufficiently obvious, through its title; in every case there is, however, a concluding chorus, even if it be only the word *Amen* once sung. An Instrumental Anthem may partake of either or all the characteristics of the Anthems above described. At the end of the 17th and the beginning of the 18th centuries the Instrumental Anthem was in frequent use at the Chapel Royal, St. James's; and, until thirty years ago, the whole of the music sung at the Festival of the Sons of the Clergy, at St. Paul's Cathedral, was regularly given with the accompaniment of a full orchestra. This custom was revived on the like occasion in 1873 with fine effect.

The Anthem is especially an English production, a development of the Motett; but the *Antifona* of the Italians, the *Antienne* of the French, and the *Wechselgesang* of the Germans, preserve to this day the character of the same prototype. The Antiphon was a special selection of words intended to be suitable to the service for the day, and was invariable, but the words of the Anthem in the Anglican Church were chosen by the composer, with or without reference to any particular Season, the early composers very rarely mentioning the special season for which they intend their music. The ignorance of Præcentors and other rulers of the choir, or their partiality for some one class of composition has often led them into having Anthems performed that are ludicrously inappropriate to the Season.

4. The history of the Anthem may be comprised within a period of little more than three centuries, and falls into three divisions, namely: the Motett period, the Verse period, and the Modern period. The Motett period lasted from the time of the Reformation to the death of Henry Lawes, say from 1550 to 1650. During the troublous times of the Commonwealth, the Anthem, in common with nearly all other Church music, excepting hymn tunes, had little or no life or character. The Verse period existed from 1670 to about 1777, the time of the death of the elder Hayes. To this succeeded another lapse of more than forty years, during which time Church needs in this matter were supplied by a series of adaptations from Oratorios and Masses, which were greatly favoured,—even Madrigals were laid under forced contribution. The absence of proper encouragement to original composers prevented many able writers, the elder Samuel Wesley among others, from employing their talents towards relieving the want of the Church. That Wesley was a writer of no mean order of genius the existence of his Latin Motetts "Omnia Vanitas," "In Exitu Israel," "Exultate Deo," sufficiently proves.

The modern period commenced with Thomas Attwood, and was continued by the younger Samuel Sebastian Wesley, and John Goss. The earliest composers of music for the Reformed Church have left no examples of either solo or verse Anthems, their contributions to this order of music being similar in character and construction with the Motett of the Italian Church. The greater portion of the Anthems, by the early English writers, were adaptations of English words to music formerly set to Latin words, a proceeding both useful and needful in the shifting period immediately succeeding the Reformation. The first music set to English words for the service of the Church—exclusive of Marbecke's plain-song—was the work of Thomas Tallis, organist to the Court, in the reigns of Henry VIII., Edward VI., Queens Mary and Elizabeth; and much of his

Digitized by Google

music, which is still extant, is a mere collection of dry contrapuntal exercises without much attempt at musical or religious expression, although some of it exhibits great originality and an agreement with the character of the words. The next writer of importance was William Byrde (1543–1623.) His Anthems, set to Latin words, and first published in 1589 under the title of "Cantiones Sacræ," were afterwards adapted, probably by himself, to English words of like character. One of them, still frequently in use, "Bow Thine ear," or "Be not wroth, very sore," for there are two versions of the same composition, was originally sung to the words, "Civitas sancti Tui," the second part of the Motett, "Ne irascaris;" and the music beautifully expresses the sentiment of the text. It is only occasionally that such a happy combination is found in the works of Anthem writers up to the period of the restoration of Charles II.; for, although the compositions of Tye, Tallis, Farrant, Allison, Morley, Hooper, Byrde, Bull, and Gibbons, are models of constructive skill, there is little that could be fairly called musical expression to be found in any one of them. Neither was there any distinction of style between sacred and secular music at this period. For example, the madrigal by Gibbons, "The Silver Swan," and his Anthem, "Hosanna," might change places, so that the madrigal might be made an anthem, and the anthem a madrigal, without any violation of character, and yet each would be counted a noble specimen of its class. It was not until men's manners and modes of thought had undergone the change brought about by the emancipation necessarily succeeding a reformation in religion, that a special and marked difference was made between the style of music used for the Church and that for secular purposes. For nearly twenty years, that is, from the death of Lawes in 1645, to the time when Pelham Humphreys was writing, Church music was represented by such writers as Child and Rogers, the best of whose compositions are but pale reflections of old styles. The pause in church matters, during the Commonwealth, had its bad effect upon Church music, until the new interest aroused by the works of foreign writers produced fresh vitality. When Humphreys began to supply the want in Church music caused by the revival of the service according to the Ritual of the Prayer-book, some degree of difficulty arose, for it was impossible to pursue the practice formerly in vogue, of making little, if any, difference in the style of sacred and secular music, for secular music had now assumed a character unfitted for the dignity and solemnity of Divine worship. To meet this difficulty a sort of compromise was effected; the secular style of a preceding generation was adopted as the prevailing standard character for Church music, a practice which has continued in use to this day. Now, as then, recently acquired ideas were used in combination with old fashioned notions, and at the period of the history of the Anthem, now being treated of, the novelty introduced was the Verse, modified by French and Italian influence upon English music. The best Verse and Solo Anthems are those by Humphreys, Purcell, Wise, Weldon, Blow, Croft, and Greene; and, with the last named writer, the verse Anthem proper culminated, and then decayed, for the poor productions of Nares, Kent, Pring, and others, although popular in their day, simply lumber the shelves upon which they are placed. Boyce and the elder Hayes were more successful in their Full, than in their Verse Anthems, some of which are models of beauty and effective writing. It is a singular fact that for many years there was a hiatus in the supply of original Anthems, the exigencies of the Church service being supplied by a series of bad arrangements, for, counting the single contributions furnished by men of genius like Battishill, who were living between the time of Boyce and Wesley, the majority of these contributions were the weak repetition of themes that had been better treated before. Thomas Attwood, a pupil of Mozart, and organist at St. Paul's Cathedral, was the first who made the laudable endeavour to supersede bad arrangements by attempting to give some adequate and connected expression to the words set as Anthems; and, although his works are to a great extent valueless as Church music, his intention should be mentioned with respect, especially as his writings and mode of thought aroused the emulation of a worthy series of followers.

It is interesting in reviewing the history of the Anthem to notice to what an extent organ accompaniment has developed and expanded. The Anthems of the first period are as effective without organ support as with it, and in those choirs in which an unaccompanied service is sometimes performed, they form the repertory from which selections are made. The organ part to Anthems of the second period is almost indispensable, by reason of the frequency with which *ritornelli*, and solos and duets are introduced. In the Anthems of the more modern period the organ is exalted almost to the dignity of a solo instrument, many Anthems being written less for vocal than for instrumental effect. The variety of stops, improved mechanism of the organ, and the advanced skill of cathedral organists form a combination too tempting to the composer,

Digitized by Google

who is, in most cases, himself an organist. The tendency of most of the music written for the organ is to treat it as an imitation of an orchestra; this improper use of the instrument is influencing the character of the Anthems of the present day; and, unless composers are wise in time, the Church music of the latter part of the 19th, will be as feeble and as useless to future generations as that of the latter part of the 18th century.

5. The first published collection of Anthems in score was that made in 1724 by Dr. W. Croft, of his own compositions, the only piece of Church music which had been previously issued in this style, being a service by Henry Purcell. The old practice of printing each part separately not only led to the loss of the several parts, but also increased the difficulty of a correct understanding of the effect, for want of a score. Of Barnard's Church music, printed in this manner, no perfect copy is known to exist, as even the parts intended for the several voices on the Decani and Cantoris sides of the choir were published in separate and distinct books.

6. Many suggestions have been made as to the derivation of the word Anthem, of which the following are the chief:—

(*a*). From the word Antiphon, it being understood that the Anthem was the successor of the more ancient Antiphon. If it was the intention of the framers of the Prayer-book to continue the use of the Antiphon it would probably have been expressed, and a table of the Antiphons proper to the Church Seasons would have been compiled and inserted among the directions for order of the service, in like manner with the table of lessons and the division of the Psalms. If, however, this direction was omitted, the tradition of their retention would doubtless remain, and so it would be found that the early Anthem writers would have been saved the trouble of making their own selection of words, and would have set to music the words of the Antiphons, all of which were taken from Scripture, and therefore in conformity with the principle of the Reformation; but this was not done, as a reference to the Anthems in Barnard's Collections, and to the words in Clifford's Anthem-book, sufficiently shows.

(*b*). From ἀντίφωνος, or according to some, the mediæval ἀνθυμνός, on the supposition that the Anthems or Antiphons were sung from side to side of the church. The choir being still divided according to ancient custom, the practice of antiphonal singing is apparently maintained, if by such a term alternation is implied. But Antiphony means classically, singing at the octave, and anthem singing is the reverse of such antiphonal singing. Durandus gives another meaning to the word, when he says, that the sentences which precede the Psalms and Canticles are called Antiphons "non quia alternatim a diversis choris cantentur; sed quia sicut claves et indices, ad quorum modulationem ac sonum, sequens canticum psalmusque alternatim cantatur. Tonus enim totius psalmi ex tono antiphonæ sumitur"—"not because they are sung by two choirs alternately, but because they are as keys and indices to the tone and mode, to which the Canticle or Psalm following ought to be chanted antiphonally. For the tone of the whole Psalm is taken from that of the Antiphon." When the Præcentor of ancient times started the Tone for the psalms some sort of antiphonal singing was practised between him and the Choir, in singing the Antiphons; and the connection of the word Anthem with *responds* and *invitatories* in the preface to the Prayer-book, would seem to imply alternate singing. The word Anthem is used three times in the Prayer-book, to the Venite, to the portions of Scripture appointed to be sung in place of the Venite on Easter-day, and in the Rubric after the third Collect. In two out of these three cases the word is used in the same sense as the old Antiphon. An antiphonal character (in the sense of alternating) is implied by the use of the words in the Office for the Communion of the Sick in the Liturgy of King Edward VI., A.D. 1549, but as the office does not comprise singing, it may be inferred that the words Anthem and Antiphon in this case refer to other than a musical meaning. It may be here mentioned that the word Anthem was at one time applied to texts of Scripture—Bishop Scory's text for his sermons preached at St. Paul's Cross, being called his Anthem—and also to secular compositions as well as sacred. In the Prioress's Tale, of Chaucer, the words—

> "—— bad me for to synge
> This antym veraily in my deying."

refer to the ancient Hymn, "Alma redemptoris." Shakspeare makes Falstaff use the word with a very different application; and the "Boar's Head Carol" annually sung at Queen's College in Oxford, was until recently, called an Anthem in the printed copies.

(*c*). A derivation from the word ἀνάθεμα has been suggested, under the impression that an Anthem is an invocation; and it is curious to note in confirmation of this conjecture, that in the original edition published in 1663, of Clifford's words of Anthems, the first book of its kind ever issued, the majority of the selected verses set as Anthems—144 out of 167—are of an invocatory character.

Digitized by Google

(*d*). Lastly, the word has been derived from ἄνθημα (a flower), from whence the word anthology. The ancient and still existing name for the book containing the words and notes of the anthems or antiphons, is Anthologium or Antiphonarium, probably from the idea that a collection of such words might reasonably be considered as a series of choice flowers from Holy writ, as the Anthem was formerly held to be the flower of the service, by those who attended church for the sake of listening to it only, and who left in an unseemly manner at its close,—a vicious habit which is scarcely yet extinct.

Anthema (*Gk.*) ἄνθεμα, short poetic form of ἀνάθημα, literally anything set up (ἀνατίθημι), hence ornaments or apparatus of a feast; hence *music and dancing*.

Anthologium (*Gk.*) A collection of antiphons set to music. A term used in imitation of the word ἀνθολογίαι, collections of small Greek poems, selected and made up, as it were, into a nosegay, from ἄνθος, a flower, λέγω, to pick. [Antiphonarium.]

Anthropoglossa (*Gk.*) [Vox Humana.]

Antibacchius. A foot consisting of two long syllables, followed by one short, - - ˘.

Anticipation. The introduction of notes before the time in which they are naturally expected in the harmony, *e.g.*:

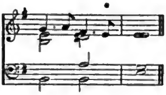

Antico (*It.*) Ancient.

Antienne (*Fr.*) }
Antifona (*It.*) } Antiphon. [Anthem.]

Antifonario (*It.*) }
Antiphonaire (*Fr.*) } [Antiphonary.]

Antiphon. (1) In ancient Greek music antiphony (ἀντιφωνή, or ἀντίφωνοι) meant "sounds in octaves" as being responsive to, or over against each other. The relation between sounds at the interval of an octave was thus implied by ἀντιφωνή, while the actual interval of an octave was called *diapason*.

(2) From the above meaning of the word it came in time to be applied to the alternate singing of choirs, as being similar in some respects to the ἐπίρρημα and ἀντεπίρρημα of the old Tragedy. The word Antiphony is the more appropriate for recitation *alternatim* if it be remembered that in the earliest public services of Christianity a choir of women and children was often responded to by a choir of men: δύο γίνονται τὸ πρῶτον χοροί, ὁ μὲν ἀνδρῶν, ὁ δὲ γυναικῶν (Philo de Vita, cont.). The custom of antiphonal singing seems to have been first introduced at Antioch by Diodorus and Flavian: οὗτοι πρῶτοι διχῇ διελόντες τοὺς τῶν ψαλλόντων χοροὺς ἐκ παραδοχῆς ᾄδειν τὴν Δαυϊδικὴν ἐδίδαξαν μελῳδίαν· καὶ τοῦτο ἐν Ἀντιοχίᾳ πρῶτον ἀρξάμενον παντόσε διέδραμε καὶ κατέλαβε τῆς οἰκουμένης τέρματα (Theod. Hist. Eccl.). Afterwards, through the musical ability of Ambrose, the same system of psalm-singing became general in the Western church. Although Christian authors give accounts of antiphonal singing as something new, there can be no doubt of its great antiquity. There are sufficient allusions to it in the Old Testament to show that it was well known to the Jews; and the very structure of many of the Psalms implies its existence. In our own times, the choirs of cathedrals and churches are usually (when seated in the chancel,—not in a west gallery) divided into two sides, one called Decani, from the fact that it is on the Dean's side of the choir (south); the other, Cantoris, because it is placed on the Precentor's or Succentor's side (north). The Psalms are sung by alternate verses, from side to side, both sides joining in the Gloria Patri. In some foreign churches, the Antiphony consists of the chanting of one verse by a single voice, the next being sung by the full choir, in response. This is not a correct system, ritualistically speaking, although, in the Church of England, in those rare places where they do *not* sing, the minister and congregation go through an exactly similar process. The Antiphony, which was once common in this country, between the minister and parish-clerk, who rivalled each other in the uncommon phases of meaning which their particular method of "saying" the Psalms often rendered prominent, is now happily almost obsolete. Miraculous stories of the introduction of responsive choir-singing are not wanting. Socrates, in his ecclesiastical history (Book vi. chap. viii), says: "Ignatius, third Bishop of Antioch in Syria from the Apostle Peter, who had also conversed familiarly with the Apostles themselves, saw a vision of angels hymning in alternate chants the Holy Trinity; after which he introduced the mode of singing he had observed in the vision into the Antiochian Churches, whence it was transmitted by tradition to all the other churches."

(3) A short sentence, generally from Holy Scripture, sung before and after the Psalms for the day, or the Canticles, selected for its appropriateness to the Church season in which it is sung. As an example, one of the Antiphons used on the fourth Sunday in Advent is here given:

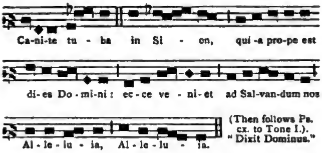
Ca-ni-te tu - ba in Si - on, qui -a pro-pe est
di-es Do-mi-ni: ec-ce ve - ni-et ad Sal-van-dum nos
Al-le-lu - ia, Al-le-lu - ia.

(Then follows Ps. cx. to Tone I.). "Dixit Dominus."

The use of the Antiphon in this manner, has no doubt grown out of the frequent recurrence of Alleluia, and other devout exclamations as found in the Psalms, which have ever been used in the Church as "Responsoria." The Gallican "Liturgy," which may with probability be ascribed to the second century, commences with an Antiphon. But the word Antiphon is used in many other senses, sometimes even given to a complete set of Versicles and Responses; thus Augustine and his followers are said to have entered Canterbury singing as an Anthem (Antiphona) one of the Litanies. Certain of the hymns sung at the end of Compline are also called Antiphons.

(4) The greater Antiphons (Antiphonæ majores) are sung on the eight days preceding Christmas-day, before the Magnificat. The first of them commences with the words, "O Sapientia," which is still found in the calendar of the English Prayer-book, on December 16th. It will be found interesting to compare the text of these greater Antiphons as found in the Sarum use, and in the Vesperal now issued from Mechlin.

(5) In the early Greek Church, in the services of which, hymns and canticles of all kinds were sung by two alternating choirs, the word Antiphon was specially applied to the three canticles which preceded the lesser Introit.

(6) Antiphona (*Lat.*) An Anthem. The English word is supposed to have been corrupted from the Latin. But several other explanations, which have been brought forward from time to time, will be found *sub voce.* [Anthem.]

Antiphonarium (*Lat.*)
Antiphonaire. Antiphonier (*Fr.*) } [Antiphonary.]

Antiphonary. } A service book of the **Antiphoner.** } Roman Church, which contained originally the *antiphons* sung in the services of the Hours, properly arranged and *noted*, to which, from time to time, other portions of music and words were added, such as Invitatories, Hymns, Responses, &c. The advantage, perhaps necessity, of referring to ancient copies of service-books for the true restoration of plain-song, which has ever had a tendency to vary in its character by unauthorized additions, or *foliations*, was felt as much one thousand years ago as it is now. For we read that the good lessons in plain-chant given to the French clergy, when Pope Stephen II. was the guest of Pepin, King of France, were soon forgotten; and that in the time of his son, Charlemagne, the church-song had become exceedingly corrupt. Charlemagne, for the purpose of remedying this, obtained the services and help of Theodore and Benoît, who carried with them from Rome a copy of the *Antiphoner* of Gregory, which the Pope himself (Adrian) had *noted*. Other accounts are to be found of these reformations of plain-chant, all of which however point to the importance of the preservation of, and reference to, old antiphoners. And later on, in the 12th century, we find St. Bernard the Abbot making efforts to stem the tide of innovations, by publishing his tract—"De Cantu seu correctione Antiphonarii"—in which he says, "take the antiphonary used at Rheims, and compare it with that of Beauvais, or Amiens, or Soissons, which are almost at your doors, and see if they are the same, or even like each other."

The number of service-books seems to have rapidly increased; for, by the constitutions of Archbishop Winchelsey (A.D. 1305) it was required that every church in the province of Canterbury should be provided with a Legend, an Antiphonary, a Grail, a Psalter, a Troper, an Ordinal, a Missal, and a Manual. In 1549, when all such books were abolished to make way for the "Booke of Common Praier," they appear to have been still more numerous: being described as "Antiphoners, Missals, Grayles, Processionals, Manuals, Legends, Pies, Portuasses, Primers in Latin or English, Couchers, Journals, and Ordinals." The Grayle, or Gradual, contained tracts, sequences, hallelujahs, creeds, offertories, the sanctus, and the office of sprinkling with holy water. Legends, or Lectionaries, contained the Lessons, which were not in the Antiphonary. It is unnecessary here to enter into an explanation of all these terms; suffice it to say, that the copying of choir-books was a matter of great labour, and that the books themselves were in consequence very costly. It is related by Spelman that two antiphonaries cost the Monastery of Crabhuse, in Norfolk, twenty-six marks in the year 1424; and it is also related that a common Missal cost five marks—a year's income of a cleric at that time. Upon the dissolution of monasteries, valuable books of this sort were dispersed throughout the country, and, from carelessness or wanton waste, destroyed in large numbers.

Antispastus. A foot, consisting of two long between two short syllables, ⏑ – – ⏑.

Digitized by Google

Antistrophe. [Strophe.]

Anwachsend (*Ger.*) Swelling, crescendo.

Aperto (*It.*) Open. The use of the damper-pedal in pianoforte music.

Apfelregal (*Ger.*) A reed stop in the organ now no longer made; the pipes, which were small, had a round hollow nob at the top like an apple, whence the name.

Antode (*Gk.*) ἀντῳδή. Responsive singing.

Aoidoi (*Gk.*), plural of ἀοιδός. Minstrels, bards. (*Lat.*) *Vates.* [Bard.]

A piacere (*It.*)
A piacimento (*It.*) } (1) At pleasure. Not strictly in time, *ad libitum.* (2) The introduction of a cadenza.

Aplomb (*Fr.*) Steadiness, self-possession.

A poco a poco (*It.*) More and more. By degrees. Applied to the increase of time or expression.

A poco più lento (*It.*) A little slower.

A poco più mosso (*It.*) Somewhat faster.

Apollo-lyra. [Psalmmelodicon.]

Apollonicon. An organ, invented in 1800 by John Henry Völler, of Hesse Darmstadt, and manufactured in London by Messrs. Flight and Robson in 1828; it consisted of about 1900 pipes, with six sets of keys, so that half a dozen performers might play simultaneously. The action was so arranged, that it might be performed upon by six players in the ordinary manner, or the various effects might be elicited by the revolution of certain cylinders which set the wind in motion, and regulated the stops according to the character of the music played. An imitation of an orchestra, with the usual instruments, including kettle-drums, was the object sought to be gained by the invention.

Apolutikion (*Gk.*) ἀπολυτίκιον. A hymn sung at the close of Vespers in certain seasons of the Greek Church. The word is probably derived from the opening sentence of the Nunc Dimittis, "νῦν ἀπολύεις τὸν δοῦλόν σου," and signifies a hymn of dismissal.

Apopemptic Song. An ancient farewell or parting hymn usually sung to a stranger about to return to his own land. Apopemtic strains were sung to the gods on certain days on which it was believed that the several deities returned to their original countries.

Apotome (*Gk.*) Ἀποτομή. A major semitone. "Major pars toni: quæ semitonium majus vulgariter dicitur" (Tinctor.) "Id quod vere semitonium nuncupatur, pars toni minor est quam dimidia. Reliqua igitur pars, quæ major est, apotome nuncupatur a Græcis, a nobis vero potest vocari *decisio.*" (Boethius De Inst. Mus., Lib. ii., cap. 29 et 30.)

Appassionato (*It.*) With feeling, passion, or affection.

Appassionamento (*It.*) With passion, love.

Appassionatamente (*It.*) Passionately.

Appenato (*It.*) With an expression of suffering, with bitterness or grief.

Applause. Praise or approbation expressed by clapping the hands, stamping the feet or the utterance of certain cries, as *bravo, encore.* In the ancient Greek theatre, Donaldson says, that "the conduct of the audience was much the same as that of the spectators in our own theatres, and they seem to have had little scruple in expressing their approbation or disapprobation, as well to the poet as to the actors. Their mode of doing this was sometimes very violent, and even in the time of Machon it was customary to pelt a bad performer with stones."

Hissing, as an expression of disapproval or contempt, is of very ancient use, and it was the custom to augment the power of the hiss, by blowing through reeds and whistles, a custom not altogether unrepresented in later times, when cat-calls, introduced into an English theatre, gave Addison a subject for an amusing paper (No. 361) in the Spectator. The hollow pipe of a key serves the purpose of the ancient *calamus* or *fistula*, in modern Italy, and the frequency with which indifferent operas are received "colla chiave," proves that the spirit of old times still lives and is active. [Fiasco.]

History shows us that applause was not confined to secular performances, but was allowed and even looked for in churches as well as theatres. Hone, in his "Ancient Mysteries described," quotes the following passage relative to this custom: Jerome desired Gregory Nazianzen to explain to him what was meant by *the second Sabbath after the first*, in St. Luke vi. 1. Gregory answered, "I will teach you that at Church, where, *when all the people shall applaud me*, you will be forced to know, what you do not know; for if you only keep silence, you will be looked upon as a fool."

At one time encores were not permitted in France, neither were calls allowed for the author of a piece which had given pleasure. When reforms were taking place, opportunities were found to break through this rule, Jean Baptiste Lemoine or Moyne, in 1789, being the first composer called upon the stage in France after the performance of his opera, "Nephté." A few years later in Italy, Paisiello was the means of removing the prohibition on the audience from applauding at all in San Carlo, for he induced the King to set the example of the change, by applauding an aria sung by Carlo Raino, in the opera "Papirius," produced in the year 1805.

It has been happily said that "Il più grand 'omaggia alla musica sta nel silenzio," and, influenced by some such principle, the better sort among a mixed audience refrain

Digitized by Google

from indiscriminate applause, encores intended as compliments often becoming an oppressive tax, levied by the unthinking, or those who care little for true art. The general opinion of the reasonable on this subject, is expressed in the following epigram:—

> "The 'sovereign people' rule all things,
> So levellers would say;
> But all 'encores' in concert-rooms,
> The 'shilling people' sway."

[Claque.]

Applicatur (*Ger.*) (1) The art of using the fingers freely upon a musical instrument of any kind, (2) shifting, *q.v.*, and recovering the original position.

Appoggiando (*It.*) Drawing out, lengthening, leaning upon.

Appoggiato (*It.*) Supported. Appoggiato notes are those notes which suspend the resolution, or that supply gaps in passages of intervals. See also PASSING NOTE, SUSPENSION, and SYNCOPATION.

Appoggiatura (*It.*) A note leant upon in singing or playing, applied to beats and grace notes, *q.v.*

Apprestare (*It.*) To make ready, to prepare for playing, to set in tune.

A première vue (*Fr.*)
A prima vista (*It.*) } At first sight.

A punta d'arco (*It.*) With the end of the bow near the point.

A punto (*It.*) In exact time, precise, strict, accurate.

Apycni (*Gk.*) (ἄπυκνοι). The notes Proslambanomenos (προσλαμβανόμενος), nete synemmenon (νήτη συνημμένων), and nete hyperboleon (νήτη ὑπερβολαίων) of the Greek system of music. The notes are so named because of their remoteness from each other (from ἄπυκνος, not close, not dense).

A quattro mani (*It.*) *À quatre mains* (*Fr.*) For four hands; as a duet for two performers on the pianoforte or organ.

A quatre seuls (*Fr.*)
A quattro soli (*It.*) } For four soloists.

A quattro parti (*It.*) In four parts.

A quattro voci (*It.*) *À quatre voix* (*Fr.*) For four voices in harmony.

Arbitrio (*It.*) Will, pleasure, *a suo arbitrio*, at his pleasure.

Arcato (*It.*) With the bow, as opposed to *pizzicato*, plucked with the finger. *Coll' arco* is a direction to the same effect.

Arched viall. An instrument somewhat in fashion like a hurdy-gurdy, invented about A.D. 1664, and thus described by Pepys in his Diary, under the date October 5th, in that year: "To the Musique meeting at the Post Office, where I was once before. And thither anon came all the Gresham College, and a great deal of noble company; and the new instrument was brought, called the Arched Viall, where being tuned with Lute-strings, and played on with kees like an organ, a piece of parchment is always kept moving; and the strings, which by the kees are pressed down upon it, are grated in imitation of a bow, by the parchment; and so it is intended to resemble several vyalls played on with one bow, but so basely and so harshly, that it will never do. But after three hours' stay it could not be fixed in tune; and so they were fain to go to some other music of instruments." Pepys had probably no design in writing the word viall in the manner in which he has done; but, in doing so, he has intimated a connection with ancient vielle or hurdy-gurdy, which the Arched viall somewhat resembled; the parchment was doubtless "always kept moving" by means of a wheel.

Archeggiamento (*It.*) (1) The same as *arcato*, or *coll' arco*. (2) The use of the bow.

Archet (*Fr.*) **Arco** (*It.*) The bow with which stringed instruments are played. [Bow.]

Archicembalo (*It.*) *Archicembalum* (*Lat.*) A cembalo with an enharmonic scale, supposed to have been invented about the year 1537 in Italy, described by Salinas as having each tone divided into parts, of which three were given to the greater semi-tone and two to the less, the whole octave being divided into thirty-one parts.

Archlute, Arciliuto (*It.*) **Archiluth** (*Fr.*) [Theorbo.]

Arco (*It.*) The bow. *Coll' arco*, with the bow, as opposed to *Pizzicato*, pinched by the finger.

Ardente (*It.*) (*Fr.*) Ardently, with fire.

Arditezza, con (*It.*) With boldness, energy.

Ardito (*It.*) Bold and energetic.

Aretinian syllables. The names Ut, Re, Mi, Fa, Sol, La, given to the Hexachord by Guido d'Arezzo (Guido Aretinus). These syllables happen to occur in consecutive notes of the scale, in an ancient hymn to S. John Baptist.

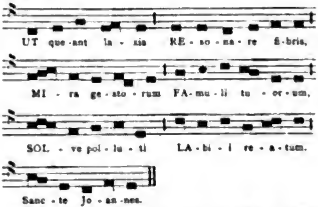
UT que-ant la-xis RE-so-na-re fi-bris,
MI-ra ge-sto-rum FA-mu-li tu-o-rum,
SOL-ve pol-lu-ti LA-bi-i re-a-tum.
Sanc-te Jo-an-nes.

From the system of over-lapping Hexachords, arise the compound names of notes such as E-la-mi, A-la-mi-re, &c., which are explained under Notation.

Arghool. A simply constructed wind instrument, now used in Egypt. It is made of common cane, and is played by mouth-pieces containing reeds. There are two species of *arghool*; the first (Fig. 1) consists of two tubes both pierced with holes, so that the performer may play in thirds and sixths; the second (Fig. 2) consists also of two tubes, but one only is pierced with holes, the other being longer and used as a *drone*. The pitch of the drone can be altered by the addition of extra pieces, which are attached to the instrument, as are also the mouth-pieces, by waxed thread.

Fig. 1.

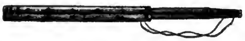

Fig. 2.

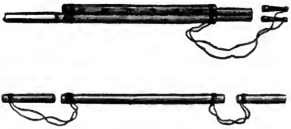

Aria (*It.*) An air, tune, song, or melody in rhythmical proportion, now understood to mean a movement for a single voice or instrument, with an accompaniment. [Air.]

Aria d'abilità (*It.*) A song of difficult execution, requiring great skill in its proper and satisfactory performance.

Aria buffa (*It.*) A song with some degree of humour in the words, or in the treatment of the music.

Aria cantabile (*It.*) An air in a graceful, flowing style, capable of much musical expression.

Aria concertante (*It.*) An air in the concert style, that is a melody for a single voice, accompanied by instruments having *obbligato* or solo passages assigned to them.

Aria di bravura (*It.*) A melody with florid, bold, and energetic passages and phrases for the voice. An *aria di bravura* is more or less an *aria d'abilità.*

Aria fugata (*It.*) A song or air, in which the accompaniment is written in fugal style, or in imitation. The difficulty of expressing dramatic emotions in this species of composition led to its ultimate disuse, though at one time it was greatly in favour. The subjoined specimen, perhaps one of the most expressive of its class, is said to be the work of Bononcini, usually called the rival of Handel, but it may be the production of either of the other writers concerned in the opera of Thomyris, from whence it was taken. This opera, produced under the direction of Heidegger, at the "King's Theater, in yᵉ Haymarket" in 1709, was a pasticcio of melodies and compositions selected from the works of Albinoni, Gasparini, Steffano, Scarlatti, and Bononcini. The opera was called English though the singers delivered some portions in Italian, and others in English. The libretto of the opera was by no means of a high poetical order.

"Aria fugata" out of the Opera of Thomyris.

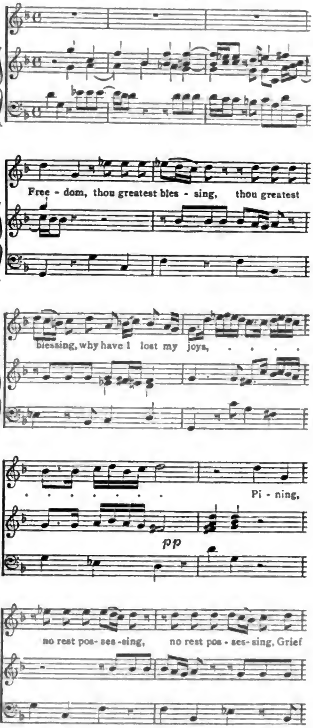
Free - dom, thou greatest bles - sing, thou greatest
blessing, why have I lost my joys, ...
. . . .
Pi - ning,
pp
no rest pos- ses -sing, no rest pos - ses- sing, Grief

Digitized by Google

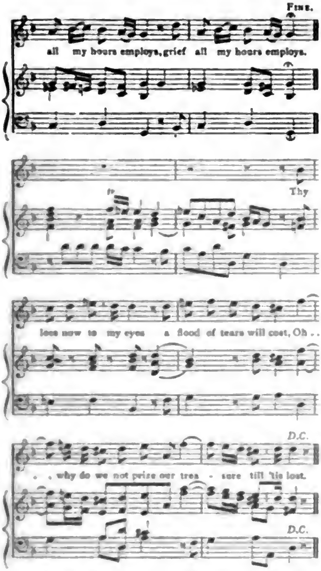
Fine.
all my hours employs, grief all my hours employs.
tr
Thy
loss now to my eyes a flood of tears will cost, Oh . .
D.C.
. . why do we not prize our trea - sure till 'tis lost.
D.C.

Aria parlante (*It.*) (1) Vocal music suitable to, and designed for, a proper declamation of the words.

(2) A style of song-writing invented towards the close of the 16th century by those Florentine dilettanti who, imbued with the spirit of Renaissance which had already revolutionized other arts, turned their attention to the necessity of ridding music of cold formalities and restoring it to its proper function, which indeed it held among the Greeks, of being a just vehicle of the ever-varying emotions which poetry calls forth. Monteverde, Peri, Corsi, and Caccini, were the musicians who made the first attempts at *aria parlante*, several operas being composed by them individually or in combination, to words by Rinuccini, in which the *aria parlante* occupied an important position. The *aria parlante* was not a *recitative*, but was sung in strict time. The latter, however, very soon grew out of the former, and assumed a separate existence in the works of Carissimi.

In the preface to the first Opera printed with the music, "Le Musiche de Jacopo Peri, Nobil Fiorentino, sopra L'Euridice de Signor Ottavio Rinuccini, Rappresentate nello Sponsalizio della Christianissima Maria Medici, Regina di Francia e di Navarra. In Venetia, MDCVIII," the author states that the groundwork of the imitation proposed "usassero un armonia, che avanzado quella del parlare ordinario."

The character of this harmony, which was intended to be a medium between common speech and singing, will be seen in the following Aria parlante for Pluto, in answer to Orpheus seeking Euridice:

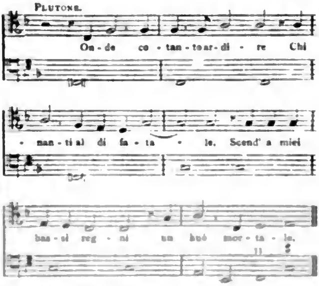
Plutone.
On - de co - tan - to ar - di - re Chi
- nan - ti al di fa - ta - le. Scend' a miei
bas - si reg - ni un huò mor - ta - le.

[Opera.] [Recitative.]

Aria Tedesca (*It.*) An air in the German style—that is to say, in which the accompaniment is inseparable from the melody.

Arie aggiunte (*It.*) Supplementary songs introduced into a work after the first performance or representation.

Arietta (*It.*) The diminutive of aria; a short air or melody. Ariette (*Fr.*).

Arigot (l'arigot). [Larigot.]

Arioso (*It.*) In the style of an air. (1) A direction that the music to which it refers is to be performed tunefully, sweetly.

(2) An intimation that recitative form has more or less been incorporated into, or perhaps superseded by, a smooth and melodious treatment of the words. The *Arioso* is found in its perfection in the works of Mendelssohn, but examples of it are not wanting in older writers, although they are simply called recitative—*e.g.*, "Behold and see" ("Messiah") and in Nos. 17 and 74 of Bach's Passion (St. Matthew).

Aristoxenians. The followers of the musical system of Aristoxenus. *c.f.* Pythagoreans.

Armer la clef (*Fr.*) To indicate the key by the number of sharps or flats in the signature.

Armoneggiare (*It.*) To harmonise, to sound in chords.

Armonia (*It.*) [Harmony.]

Digitized by Google

Armonica (*It.*) Harmonica, Armonicon, Harmonicon. (1) The musical glasses, a series of glass cups of various sizes and thicknesses, capable of producing the different notes of the diatonic scale by friction upon the edges. The name armonica was given to this instrument by Benjamin Franklin, to whom also the credit of the invention is sometimes given, but the idea was suggested by a Mr. Pickeridge, an Irish gentleman, and first carried out by M. Delaval, and was in use long before the name *armonica* was given to it by Franklin.

(2) An instrument now used by children, consisting of a flat oblong box, containing free reeds so arranged that when applied to the mouth, inspiration and respiration through the orifices in the side, produce different sounds of the scale, in a series. [Cheng.] [Harmonium.]

Arpa (*It.*) [Harp.]

Arpa doppia (*It.*) A double harp. [Harp.]

Arpanetta *or* **arpanella** (*It.*) A small harp.

Arpége (*Fr.*) [Arpeggio.]

Arpeggiando (*It.*) Playing arpeggio, *q.v.*

Arpeggiare (*It.*)
Arpeggiato (*It.*) To strike the notes of a chord in succession in the manner of harp playing.

Arpeggiatura (*It.*) [Arpeggio.]

Arpeggio (*It.*) In the style of a harp. A term applied to the notes of a chord when they are struck consecutively, instead of simultaneously.

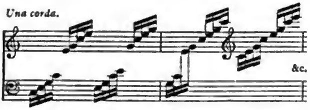

In pianoforte music a waved line is written beside a chord intended to be played *arpeggio:*

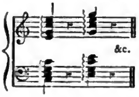

Arrangement. A selection or adaptation of the parts of a composition, to fit them for performance by other voices or instruments than those originally designed.

There are very few examples existing of acknowledged arrangements in the earliest musical publications; for few ever thought of tampering with an author's compositions so far as to divert them from their original intentions. The adaptation of new words, although it occasionally involved a slight alteration in the time-value of some of the notes, was a matter of small importance; and the musician who undertook such a matter rarely gave himself the credit of having done so clever a thing as modern arrangers would have us believe such an alteration to be. Thus Nicolas Yonge, in his collection of Madrigals, "Musica Transalpina," London 1588, leaves it to the judgment of the reader to infer the part he took in giving his book to the world, calling his work "Madrigales translated of four, fiue, and sixe parts, chosen out of diuers excellent Authors, vvith the first and second part of La Verginella, made by Maister Byrd, out of two Stanz's of Ariosto, and brought to Speak English with the rest. Published by N. Yonge in favour of such as take pleasure in Musick of voices." For a somewhat lengthened period arrangements were described as "brought to light," "framed," "figured," "fitted," "made proper," and "newly set forth," for example: "Lessons for Consort, made by sundry excellent authors, and set to sixe severall instruments, Namely, the Treble Lute, Treble Violl, Base Violl, Bandora, Citterne, and the Flute. Now newly *set forth* by Philip Rossetor, 1609."

Richard Alison, in his book "An Howres Recreation in Musicke, apt for Instruments and Voyces" (1606), describes his arrangement as "*Framed* for the delight of gentlemen and others which are well affected to that qualitie." A Dutch edition of Gastoldi's ballets for "5 en 6 stemmen, te singen of speelen," 1648, is "gestelt"—that is, arranged or accommodated — "of 3 en 4 stimmen," and this is perhaps one of the earliest instances of an alteration of an original design. Arrangements such as these could only be called into existence by the desire to possess condensations of larger works.

The "Modulorum Hortus ab excellentissimæ Musicæ auctoribus" is described as being merely collected by R. Floridus, Rome, 1647,—"in lucem curavit edendam." A few years later arrangements are described as "transpositions." Thus, in the Mercurius Musicus for 1699, the "New teaching songs, compos'd . . . . With a Thorow Bass for the Harpsichord or Spinett," we have the further intimation of "The songs being *Transpos'd* for the Flute at the end of the Book." In the "Orpheus Britannicus, a collection of all the choicest songs for One, Two, and Three Voices, composed by Mr. Henry Purcell; together with such Symphonies for Violins or Flutes as were by him designed for any of them, and a Thorow-bass to each song, figur'd for the Organ, Harpsichord, or Theorbo Lute (1698-1702); also, in "Suits of the most Celebrated Lessons for Viols, collected and *fitted*

Digitized by Google

to the Harpsichord or Spinett, by William Babell (1702);" and in "A choice Collection of Lessons, being excellently *Sett* to the Harpsichord, viz. Old Simon the King, Moteley's Maggot, Mortlack's Grounds, and several others (by Blow and Purcell) 1705." In "A Collection of the Newest Minuets, Rigadoons, and French dances perform'd att Court and Publick entertainments," 1716. The tunes are made "*proper* for the Violin, Hoboy, or Flute," and in "Six Setts of Choice Opera Songs or Arietts, with their Symphonys *fitted* for two Flutes. The Second Parts being compleat and airy as the first, not thin and heavy as Second Trebles usually are;" in both parts their proper Variations for the Humour of the Flute (1712.)

A little later in date, we find, "Song in the Opera of Flora, with the Humorous Scenes of Hob, designed by y^e celebrated Mr. Gravelot, and engrav'd by G. Bickham, Junr. The Musick proper for y^e Violin, German and Common Flute, Harpsichord, or Spinet, with a New Base, and thoro' Base to each Song" (1737). The business of arrangement, that is to say, of altering music intended for one purpose, so that it might serve another—more or less hinted at in the preceding collections—arose with the popularity of Handel's works; thus, copies of "favourite Choruses" out of Mr. Handel's celebrated oratorios "adapted for the Harpsichord or organ and a single voice," began to appear soon after his death. Such "arrangements" being part and parcel of the system of piracy which was most ingeniously and unblushingly carried on during the last century. A chorus arranged for a single voice ceasing of course to be a chorus; but, as a double security, many of these pieces were made cleverly incorrect. Thus Pitt, organist of Worcester, evaded all copyright that might have existed by arranging his "Church music" from the sacred works of Handel, by a system of dove-tailing and occasional alteration of key.

"The beauties of Handel, consisting of his most favourite Songs, Duets, and Trios; arranged with a separate accompaniment for the pianoforte, and figured from the MS. scores of the author, by Jos. Corfe" (c. 1782) is perhaps one of the earliest collections of confessedly "arranged" music. In 1795, J. W. Holder, Mus. Doc., Oxon, one of the most talented pianoforte players of his time, published an arrangement of the choruses of Handel for four hands, which were the standard pieces of their kind for many years, being frequently played by two performers on one organ also. Giambattista Cimador (1750-1810) was probably the first who was employed by the publishers of London on purpose to make arrangements of large works for the pianoforte, or small bands, his arrangement of Twelve Symphonies by Mozart, as sestetts with a seventh part, *ad libitum*, being considered at the time they were made as of more than ordinary excellence. These were undertaken by Cimador out of pure love for Mozart's works, and a desire to communicate that love to the musicians of his time who thought "Mozart's symphonies too arduous and difficult."

About the same time J. S. C. Possin (1755-1822), a musician of such singular modesty that he never would have his name printed with his works, arranged for Salomon the twelve symphonies of Haydn, known as the "Salomon set," for the pianoforte, in "an admirable manner;" indeed, says his biographer, "they were the first adaptations of orchestra music worthy of notice." From that time to the present "adaptations, arrangements, and transcriptions," have been issued in unlimited quantities, of more or less value.

Arranger (*Fr.*) To arrange a piece of music. [Arrangement.]

Arrangiren (*Ger.*) To arrange a piece of music. [Arrangement.]

Arsis (*Gk.*) *ἄρσις* (from *αἴρω*), a raising, an elevation, as opposed to thesis (*θέσις*, from *τίθημι*), a depression or lowering.

There are two kinds of Arsis. (1) of accent; (2) of metre.

The former of these does not perhaps call for special attention from musicians, unless it be looked upon as a subject into which their educated *ear* qualifies them to enter; or unless it be considered (as it undoubtedly was by the Greeks) as an essential part of the education of those who attempt to set *words* to *music*. The latter has been explained from two opposite points of view, both of which, however, are closely connected with the former,—a slight sketch of the whole subject is therefore subjoined:

(1) Though not accepted without dispute, the following facts seem generally to be admitted; first, that in speaking, the voice is constantly varying slightly in pitch, that is, is not absolutely on *monotone;* next, that the component syllables of polysyllabic words are not exactly of the same duration; lastly, that there is an emphasis on particular syllables, which is independent alike of the raising or depression of voice, and of the length of time during which any syllable is held.

The elevation, or pitch of the voice is classically termed Accent (from *ad* and *cantus*, just as *προσῳδία* is from *πρὸς* and *ᾠδή*); the duration of syllables is called Quantity; and the metrical emphasis is called Ictus.

Accent is of two kinds, vocabular and

oratorical. The former is that method of pronunciation which a word receives if it stands alone in a vocabulary or dictionary; the latter that which it receives in consideration of its position in a sentence, words being of course influenced by the meaning to be expressed in a sentence of prose, or by their metrical position in verse.

There can be no doubt that a nice ear and appreciation of pitch is required before accent and quantity can be distinguished from each other in modern languages. With regard to ancient languages the same difficulty does not exist, because, putting aside the question of the correctness of our pronunciation of them, quantity is governed by either known laws of syllabic structure and position, or by the actual shape of the letters. Hence, many have thought that *quantity* does not exist in modern languages, and all that we possess is accent (elevation of the voice) and emphasis, and that these two always coincide, and are commonly included in the one term *accent*. But as a matter of fact, the pitch of voice is in modern languages quite independent of quantity, *e.g.*, *precarious*, *request*, &c., in which the voice is high for the short syllable, drops in pitch for the long, yet no one can doubt that there are long syllables in these words, just as much as in such others, as *probable*, *symmetry*, *pendant*, &c. The pronunciation of English in the common conversation will give but a very slight clue to the intricacies of our language in this respect. For, in addition to the acute accent already spoken of, we certainly have a *flat* accent corresponding to the Greek `, *e.g.*, cumbersòme, where the voice drops a little below what might be termed its key-note. We have also the up-and-down slide indicated by the Greek circumflex, *e.g.*, feârful, loâthsome. But with us these are always oratorical, never vocabular.

Having said thus much as to elevation and depression of the voice, it is now time to show how *arsis* is used as a musical term. According to Scaliger, when the voice is raised on a syllable it is called *arsis*, when it returns to its original position it is called *thesis*. Priscian (see Foster on Accent, p. 81, note) not only says the same thing, but gives as an example the word *natura*, pointing out that there is an *arsis* at the syllable *tu*, and thesis on *ra*. In this sense *arsis* is evidently the accent, or elevation of the voice, which has been already spoken of.

(2) But *arsis* and *thesis* are not only applied to the elevation and depression of voice, but also to the *strong* and *weak* parts of metrical scansion. But unfortunately, scholars have used these terms in two ways. For instance, Tate says (see Donaldson, Theatre of the Greeks, p. 371), "those syllables which have the metrical *ictus* are said to be in *arsi;* those which have it not, in *thesi* . . . . . the latter is sometimes called the *debilis positio.*" In this he follows Bentley, who makes *ictus* (or *percussio*), *elevatio*, and *arsis* synonymous.

To this other scholars object, and say truly that a syllable often is in *arsi* as regards *metre*, when it is in *thesi* as regards accent (elevation of voice). Also, Victorinus says distinctly that "*arsis* and *thesis*, as used by the Greeks, refer to the movement of the foot (significant pedis motum), and that the former is 'the elevation of the foot' *without sound*, the latter the 'lowering of the foot' to the ground, *with a sound*, the *sounds* marking the metrical ictus. To this Foster (on Accent, p. 166) agrees. With these authors, therefore, *ictus* and *thesis* are synonymous.

Hence, musicians who agree with the former of these opinions and make *arsis* and *ictus* synonymous are justified in saying that there is an *arsis* on the down-beat of every bar, and its up-beats are in *thesi*, for if *thesis* is *debilis positio*, it would be absurd to say that this occurs on the down-beat, except in some rare cases of syncopation.

Those musicians, on the other hand, who believe that *ictus* and *thesis* coincide, because the *thesis* of the foot marked the *ictus* of the metre, have a perfect right to say that the down-beat of a bar is in *thesi*, and an up-beat in *arsi*.

Inasmuch as the confusion among musicians in using these terms has resulted from the disagreement of scholars as to their proper application, it is much to be hoped that they will be allowed to sink into disuse. The expressions, *strong position* and *weak position* of the bar, imply all that is understood by *arsis* and *thesis*, without the risk, by their use, of calling forth absolutely contradictory opinions as to their meaning.

Art (*Ger.*) Species, kind, sort, as *auf polnische Art*, a sort of *polonaise*, &c.

Articulation. (1) In singing, the art of distinct pronunciation. (2) In instrumental music, the art of producing proper tone by a right adjustment of the fingers, or the lips. The latter application of the term is less commonly met with than the former.

Artist. One who possesses in a high degree that appreciation of the beautiful and that refined temperament, which, when duly trained and educated, become active faculties, and render their owner an able and influential exponent of Art.

As (*Ger.*) The note A♭.

Asamentata, Assamenta, or **Axamenta** (*Lat.*) The songs or hymns sung by the Salii, *q.v.*

Ascaules (*Gk.*) ἀσκαύλης, a player on the ascaulos.

Ascaulos (*Gk.*) ἀσκαυλος, a bagpipe, from ἀσκός, a leathern bag, and αὐλός, a pipe. [Bagpipe.]

As dur (*Ger.*) The key of A flat major.

Ashantee Trumpet. An instrument formed of the tusk of an elephant carefully hollowed. Its peculiarity consists in the fact that the embrochure is not at the end, but in the side, a short distance from it.

As moll (*Ger.*) The key of A flat minor.

Asor. [Azor.]

Asosra (*Heb.*) [Chatzozerah.]

Aspiration (*Fr.*) (1) The sign ' for shortening the duration of a note. [Spiccato.]

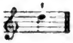

(2) A former name for an appoggiatura.

Aspirare (*It.*) To take breath audibly, bad management of the breath in singing.

Asprezza (*It.*) Harshness, severity.

Assai (*It.*) Very. *Allegro assai*, very fast.

Assemblage (*Fr.*) (1) A series of rapid passages executed on wind instruments. (2) Double tongueing on the flute or cornet.

Assez (*Fr.*) Enough, very; as, *assez lent*, rather slow.

Assonance. Agreement of tone, consonance.

A string. [A § 7.]

A suo arbitrio (*It.*) At one's pleasure or judgment.

A suo comodo (*It.*) At one's leisure.

A suo bene placito (*It.*) At one's pleasure.

A suo luogo (*It.*) At one's position or place.

Atabal. A Moorish tambour.

A table sec (*Fr.*) The performance of vocal exercises without the accompaniment of an instrument. *Sec.* lit. *dry*, *cf.* Lat. *assa vox*, an unaccompanied voice, and *assæ tibiæ*, flutes used without a voice accompaniment.

A tempo (*It.*) In time. [A battuta.]

A tempo comodo (*It.*) In a convenient, easy, moderate time.

A tempo di Gavotta (*It.*) In the time of the Gavotte, *q.v.*

A tempo di Minuetto (*It.*) In the time of the Minuet, *q.v.*

A tempo giusto (*It.*) At a just pace. (1) In general, an indication that the movement should be taken at a moderate tempo. (2) A direction (in older writers) to return to strict time after irregular declamation.

A tempo ordinario (*It.*) At an ordinary pace.

A tempo primo (*It.*) In the time first given.

A tempo rubato (*It.*) Robbed time; time made slightly irregular for the sake of expression.

A tre (*It.*) For three voices, instruments, or parts.

A tre mani (*It.*) For three hands upon an organ or pianoforte.

A tre parti (*It.*) For three parts.

A tre soli (*It.*) For three principals, either vocal or instrumental performers.

A tre stromente (*It.*) For three instruments.

A tre voci (*It.*) For three voices, or parts.

Attacca (*It.*) Commence at once, without a pause.

Attacca subito (*It.*) [Attacca.]

Attaccato subito (*It.*) To be begun at once. Go on.

Attack. (1) A vigorous entry of voices or instruments at a leading point. (2) A courageous rendering.

Attacco (*It.*) (*Lit.*, a binding together.) A "point" of imitation—that is, any short passage or figure proposed for treatment by imitation. In fugues the *attacco* is often made from a fragment of the subject or of the *andamento*, but is sometimes quite independent. [Andamento.] In some polyphonic vocal music (*e.g.*, some madrigals or masses) a considerable number of *attacci* are used, as a method of sustaining interest in a movement not in modern form.

Attendant Keys. Relative keys, keys of affinity. Attendant keys in a scale are the relative minor or major, the dominant and subdominant, and their relative minors or majors. [Relative Key.]

Atto [*It.*] An act in an opera. [Act.]

Attore or **Attrice** (*It*). An actor or actress, the chief singers in an opera.

Aubade (*Fr.*) (1) An open air morning concert, the antithesis of a serenade. (2) The word is derived from *aube*, day-break, and was similar in character to the English "Hunts up" (*q.v.*) Sometimes unmusical noises were made for an aubade, and so the word came to be employed as a term for an insult. The *Aubades de Calène* occupied in France the position of the Waits (*q.v.*) in England, as they were performed in the evening for a month or so before Christmas. Although doubtless of religious origin, the performers gradually introduced secular melodies. The players, like the Waits, were officially licensed. The word *Calène* is a French provincial form of the word Calendes, Christmas Day being formerly called "le jour des Calendes."

Audace, con (*It*). With vigour, boldness.

Auditory nerve. [Ear.]

Auferions (*Old Eng.*) Wire strings.

Aufgeweckt (*Ger.*) Brisk, lively, sprightly, cheerful.

Aufgewecktheit (*Ger.*) Sprightliness, liveliness.

Aufhalten (*Ger.*) To stop, to keep back, retard.

Aufhaltung (*Ger.*) Suspension. [Harmony.]

Auflösung (*Ger.*) Resolution of a discord.

Aufschlag (*Ger.*) Unaccented beat.

Aufstrich (*Ger*). An up bow in violin playing.

Auftakt (*Ger.*) The unaccented part of a bar. The up-beat.

Augmentatio (*Med. Lat.*) The lengthening of a note by the addition of half its length, thus corresponding to the use of the modern dot.

Augmentation. The introduction of the subject of a fugue or canon, in the course of its progress, in notes of longer duration than those in which it was first proposed. [Fugue.]

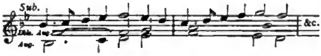

Augmented interval. [Interval.]

Augmented subject. [Augmentation.]

Auletes (*Gk.*) Αὐλητης. A player on the Aulos or Flute. [Aulos.]

Auletrides (*Gk.*) Plural of αὐλητρίς. Female players on the Aulos or Flute, *q.v.*

Aulæum (*Lat.*), αὐλαία (*Gk.*) The curtain of a theatre.

Aulos (*Gk.*) ἀυλός, derived from ἄημι, to blow, as *flute* is from the Lat. *flo*. The most important wind instrument of the Greeks. The aulos was sometimes double, the two tubes being called *dextra* and *sinistra*, and sometimes *male* and *female*. Though generally rendered *flute*, there is much reason for supposing that it was a reed-instrument, or, at the least, that the term, used generally, included instruments of the *oboe* family. The fact that the two tubes were often of different lengths (*impares*) has been explained by saying that they were tuned in different modes. But it is far more probable that they were constructed like the *arghool*, and that the longer tube gave out a drone. The double flute was not unknown to the ancient Egyptians and Assyrians, as shown in figs. 1 and 2, but they were divergent, or perhaps actually separate from each other. Fig. 3 represents two ancient Greek flutes, preserved in the British Museum.

Fig. 1.

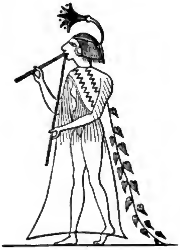

Fig. 2.

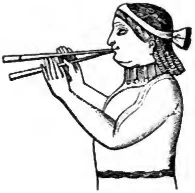

Fig. 3.

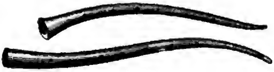

A una corda (*It.*) With, or on, one string. A direction (1) in pianoforte music, to use the soft pedal. (2) In music for stringed instruments to play the passage so marked on one string only, by the shift. [Shift.]

Ausarbeitung (*Ger.*) The working out of a theme, the climax of a composition.

Ausdruck (*Ger.*) Expression, *q.v.*

Ausführung (*Ger.*) (1) Performance or execution. (2) The working out of a subject in composition.

Aushaltung (*Ger.*) The time a note occupies in sounding, the duration of sound, sustaining a sound.

Aushaltungszeichen (*Ger.*) A pause.

Äussere Stimmen (*Ger.*) [Extreme parts.]

Ausweichung (*Ger.*) Change, modulation.

Authentic cadence. A final close, in which the common chord of the Tonic is immediately preceded by the common chord of the dominant. [Cadence.]

Authentic mode. The name given to those modes on which were afterwards constructed other modes called Plagal, by an alteration of the final to a fourth below. [Plain Song.]

Authentic part of the Scale, in Counterpoint and Fugue, is that which lies between a note and its Dominant, whilst that which lies between the Dominant and its superior Tonic is termed Plagal. The terms are used chiefly in connection with Subject and Answer. [Fugue.]

Autos Sacramentales (*Sp.*) One of the early forms of Spanish drama, similar in some respects to the mysteries and moralities in England, but in which music and dancing formed an important part. The *Autos* had reference to the adminstration of the Sacraments according to the ideas received by the people.

Auxiliary Notes. Notes not essential to the harmony, introduced for the sake of breaking monotony, or of giving freedom of motion to one or more of the parts. They may occur on either the accented or the unaccented part of the bar, and if introduced below the melody should be only a semitone from the proper note of that melody, but if above they may be either a tone or a semitone as the position in the scale would warrant, or taste suggest.

BEETHOVEN. No. 4 Symphony.

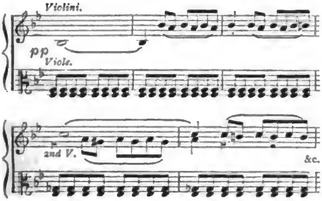

Examples of extended auxiliary notes, and of auxiliary notes, on the accented part of the bar.

VERDI. Coro "Vedi le fosche" (Trovatore).

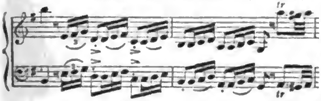

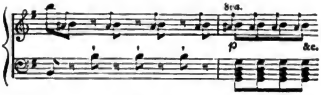

ROSSINI. "Guillaume Tell" Overture.

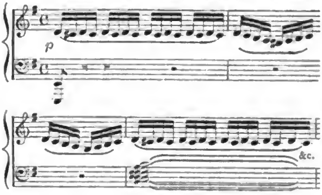

AUBER. Coro "En bons militaires buvons" (Fra Diavolo).

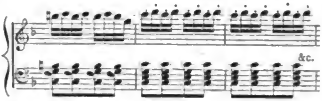

Auxiliary Scales. The scales of relative or attendant keys, *q.v.*

Ave Maria (*Lat.*) (Hail! Mary.) The angel's salutation of the Blessed Virgin Mary, used in the Roman Catholic Church as an Antiphon, &c.

Avena (*Lat.*), lit. *oats.* An oaten pipe, hence (1) any simple reed used as a shepherd's pipe—"est modulatus avena Carmen," Tibull, 2, 1, 53. (2) The syrinx or pan-pipes, reeds joined together with wax—"pastor junctis pice cantat avenis," Ovid, Tristia 5, 10, 25. [Pan's pipes.]

A vista (*It.*) At sight; at first sight.

Away. A direction in Mace's Musicks Monument, published in 1676, signifying a return to the original time.

Azione Sacra (*It.*) Sacred dramas. [Autos Sacramentales. Oratorio. Passion Music.]

Azor (*Heb.*) This word which occurs in the Book of Psalms and elsewhere, is variously rendered according to the view which is taken of its association with *nebel.* In Psalm xxxiii., 2, "Sing unto him with a *nebel* and *azor*" some drop the "and" and understand *azor* as qualifying *nebel*, making the compound word to signify a "ten-stringed nebel" (psalterium decem chordarum). Whether the *azor* was a distinct instrument, or not, it is impossible to say, although Engel, Fétis, and some other authors have so considered it, and have ventured to assign to it a definite number of strings.

Digitized by Google

B.

B. (1) The name of the note above Proslambanos, in the greater perfect system of the Greeks. The first note of the lowest Tetrachord (Hypaton). [Greek Music.]

(2) The third note of the grave hexachord of the Guidonian system, in which it is B mi. [Notation.]

(3) The seventh note of the normal scale C, the note *Si* [Si], in Tonic Sol-fa system *Te*.

(4) The major scale having five sharps in its signature.

(5) The note B♭ in Germany, where B♮ is known as H, whence the possibility of making the letters B, A, C, H, into a fugue subject,

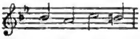

as has been done by Bach, Schumann, Liszt, and others.

(6) In old solmizations this note was called a *Mi*. [Solfeggio.]

There is no *authentic* church-mode commencing on this note, owing to the imperfection of its fifth when unraised by the signature.

B. Abbreviation of *Bass* voice, *Bassoon*, and *Double-bass*.

Baarpÿp. The name of a stop in some of the Dutch organs; (*lit.*) the Bear-pipe, written also Bärpfeife and Bärenpfeife, so called from the instrument played as an accompaniment to dancing bears.

Baas or Base Dance. A dance or slow movement, similar to the Measure, *q.v.*, or the Minuet, so called probably in contradistinction to the vaulting dances in which greater agility was displayed.

> "And then came downe the l prince and the lady Cecill, and daunced two baas daunces, and departed up againe; the l. prince to the King, and the lady Cecill to the Queene."—*Wright's Provincial Dictionary.*

Baccalaureus Musicæ (*Lat.*) Bachelor in Music.

Bacchanalian Song. (1) Songs sung in procession during the worship of Bacchus. (2) Any song in praise or defence of wine drinking, of which there are numbers belonging to the 18th century.

Bacchia. Kamschatdale dance, in $\frac{3}{4}$ time.

Bacchius. A metrical foot consisting of one short and two long syllables. [Metre.]

Bacciocolo (*It.*) Tuscan musical instrument of the guitar kind.

Bach Society, The. To this Society, founded in 1849, is due the credit of giving the first performance in England of Bach's "Matthew" Passion, on April 6, 1854, under the direction of W. Sterndale Bennett. Among many well-known musicians associated with this Society were Robert Barnet, George Cooper, E. J. Hopkins, Hullah, Henry and Sir George Smart, C. Potter, and Dr. Steggall. The Society was dissolved in 1870.

Back Choir Organ. The English equivalent of Rückpositiv, *q.v.*

Backfall. A Turn in Lute or Harpsichord music, *written* thus 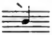 *played* 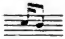

Back fall. [Organ, § 10.]

Badinage (*Fr.*) Playfulness.

Bagana. The ten-stringed lyre of the Abyssinians. It has only five different notes, but each note has its octave-string.

Bagatelles (*Fr.*) Sketches, short pieces, trifles.

Bagpipe. The *ascaulus* (ἄσκαυλος) of the Greeks (from ἀσκός, a leathern bag, and αὐλός, a pipe); the *tibiæ utriculariæ* or *utricularium* of the Romans; *sampogna* or *zampogna* of Italy; the *cornemuse* of France; the *chifonie* or *symphony* of the middle ages; the *souggarah* or *souggarah* of the Arabians. An ancient wind instrument of almost universal adoption, formerly in common use in every part of Europe, but now only found in parts of Italy, Sicily, Calabria, Brittany, Poland, and Scotland, in form more or less varied; in Ireland the bagpipes under the name of the Union pipes, are yet to be met with, but as a musical instrument among the English it has completely disappeared, in consequence of the advance in musical taste. A form of bagpipe is probably meant by the word *symphonia* (Dan. iii. 15) translated in the Italian version of the Bible *zampogna*. In its general construction the bagpipe consists of a leathern bag frequently formed of the whole skin of a kid or other small animal, which contains the wind conveyed from the mouth of the player through a tube, a small valve preventing its rapid escape. The sound comes from four

Digitized by Google

pipes, three of which united are called the *drone*, and are capable of producing only one note each, subject to tuning. These notes are heard throughout the performance. The fourth pipe, the *chanter*, furnished with a reed, is bored with six or eight holes which are stopped by the ends of the fingers of the performer. The scale of some of the Scotch bagpipes, with eight ventages, is in the minor mode with the seventh flat:

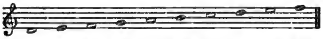

in others with six holes, the fourth and seventh are omitted:

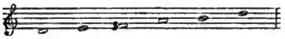

that of the Calabrian bagpipe is the diatonic scale:

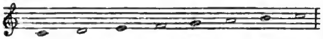

The bagpipe was known to the Anglo-Saxons, and that it was at one time in England a popular instrument, may be inferred from the frequent mention made of it in mediæval times. Strutt quotes a MS. recording many payments made to bagpipers in the reign of Edward III., about 1335, both for their personal performance and as an allowance to enable them to visit the foreign minstrel schools. The same authority also records a payment to another bagpiper in 1494. The manner in which the instrument is mentioned by Chaucer, and other poets, shows it to have been exceedingly popular and of frequent use in England in their days; and a large number of tunes quoted or alluded to in William Chappell's "Popular Music" bear evidence of having been of bagpipe character. There is no proof that the bagpipe is a national Scottish instrument, for its introduction into Scotland only dates from the time it began to be disused in England.

There is a tradition that bagpipes were used at the Battle of Bannockburn, and there is a tune, "Hey taitti, taittie," said to be the identical march played by them. Ritson, in his preface to a collection of Scottish songs, doubts whether the Scots had any martial music, and quotes Froissart's account of each soldier in the army wearing a little horn, on which, at the onset they would make such a horrible noise "as if all the devils in hell had been let loose." He further notes that as these horns are the only instruments mentioned by Barbour the Scottish chronicler, it must remain a moot point whether Bruce's army was ever cheered by the sound of a bagpipe.

The earliest mention of the bagpipe as forming part of the military music of the Scotch was at the Battle of Balrinnes (1594), though the oldest known pibroch is called the "Battle of Harlaw," but it could not be contemporary with the event (1411). There is mention of trumpets and drums in the old ballad relating to the battle, but none of the bagpipe:

> "The armies met, the trumpet sounds,
> The dandring drums alloud did touk."

The Irish or Union pipes are furnished with a pair of bellows (worked with the elbow) with which to inflate the bag. There are three drones, two tuned in unison, and one an octave below; most pipes have a valve by means of which the drone can be silenced, and there is also a contrivance for sounding at will the common chord of the key note in which the pipes are set. The quality of the chanter is more like that of the clarinet than the oboe, and the general tone of the Irish pipes is softer and less piercing than the Scottish bagpipe. The native Irish pipers call the instrument "ullan piobe," the pipes of the elbow. Shakespeare's mention of "woollen pipes" in the "Merchant of Venice," Act. iv. sc. 1:

> "Why he, a harmless necessary cat ·
> Why he, a *woollen* bagpipe,"

refers probably to the "ullan pipes;" and the word "union," as applied at the present day, may be only a modern substitute for the right word, for it is difficult to see the force of the application of the term "union" to bagpipes, unless the word be a corruption of a proper term.

It is supposed that the bagpipe came originally from the East; it is still to be met with in use among many Eastern nations. In India, China, Persia, and Egypt, it is the subject of frequent mention by many travellers.

Baguettes (*Fr.*) Drumsticks.

Baisser (*Fr.*) To lower.

Balafo. A musical instrument popular among the negroes of Senegambia. It is made of a series of graduated pieces of wood, placed over gourds, which act as resonance-boxes, is struck with hammers, and has a scale of two octaves, sometimes tuned in accordance with the white notes of a pianoforte.

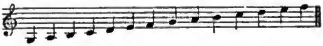

Balalaika (*Russ.*) A Russian instrument, in form like a guitar, but narrower and of less depth; it has two strings. With it the Russian Moujiks accompany their popular songs.

Balancement (*Fr.*) Tremolo.

Balcken or **Balken** (*Ger.*) The bar under the belly of a violin.

Digitized by Google

Balg (*Ger.*) Bellows, wind-chest.

Bälgetreter (*Ger.*) The bellows-treader. In old organs the blower worked the bellows by standing on them in turns.

Ballad. A song designed to suit a popular audience. A varied derivation has been claimed for the term, which doubtless meant originally a dance song. Hence its connection with the Mediæval Latin word *ballare*, (*βάλλω, βαλλίζω*). As a poem, the ballad has undergone so many transmutations that it is difficult to describe it properly, many pieces to which the term is applied having little or nothing in common with the primitive form, and poems of exactly similar character being described at one time as romances, at another ballads, at another lyric-epics. The Italians—among other writers, Dante—gave the title *ballata* to short lyrical pieces of inartistic construction allied to the sonnet or madrigal. It was against the French equivalent for these *ballate* that Molière wrote. The Spanish romances, erroneously called ballads, belong to epic poetry. The ballad, as we now understand its meaning and application, is confined to the people of Northern Europe, the Germans, following Bürger, the creator of the modern ballad, have given it an artificial character by the introduction of reflections arising out of the incidents. A ballad, properly speaking, is a simple narrative of one or more events, told without gloss, commentary, or deduction, set to a tune sufficiently rhythmical to act as one of the original purposes of a ballad, namely, a dance tune. The old ballad tunes still existing are nearly all of this character. In fact, the majority of the melodies have been recovered from having been preserved in collections of them made by dancing masters at various periods. The title of Ballet or Ballad—says Warton—was often applied to poems of considerable length, of various subjects, sometimes to prose compositions, sometimes to plays or interludes, sometimes to religious verses or discourses.

Ballad (Old English). The English have ever been a ballad-loving people, and although the taste was more widely diffused among all classes in former days than now, yet there is no present sign that it will soon die away. Ballads were embodied into our earliest histories, because the bards or minstrels—called Scopes in the language of the country[*]—were the earliest of our historians. The Scope was both poet and musician. He recorded deeds of ancient valour, and enlarged upon them in order to stimulate the warlike spirit of his hearers. He adopted ancient stories of adventures, and re-applied them to some more recent hero, in order to give greater interest in them to those who were assembled around him. In the Anglo-Saxon Chronicle we have about a dozen fragments of historical ballads, but these commence only from Athelstan's victory over the Danes. When William of Malmesbury was writing the history of King Edward, son of Alfred the Great, he said: "Thus far I have written from trustworthy testimony—that which follows I have learnt more from old ballads, popular through succeeding times, than from books written expressly for the information of posterity. I have subjoined them, not to defend their veracity, but to put the reader in possession of all I know."[†]

Again, after recounting the pride of King Edgar in having compelled subject kings to be his oarsmen, while he sat at the prow, William says: "For this he is justly blamed by history, but the other imputations which I shall mention hereafter, have rather been cast upon him by ballads."[‡]

It may be asked, "what kind of music had these ballads?" The answer will be that, although we have no existing specimen of ballad music of such early dates, yet we have hymns to Latin words, some of which have more tune in them than would be expected, and that they are our only existing means of forming a judgment. It was not mere natural song with indefinable sounds, but with regulated notes upon the diatonic scale. In the year 951 the double organ at Winchester Cathedral had 400 pipes and required two organists. It was intended to be heard all over Winchester, in honour of Saint Peter, to whom the Cathedral was dedicated. Wolstan, or rather Wulfstan, of Winchester, who describes it fully in his Life of Saint Swithun, was himself the author of a treatise on Harmony (*De tonorum Harmonia*), which was a standard book, and remained in use 200 years after it had been issued. William of Malmesbury, writing after 1100, describes this book as "very useful" (*valde utile*). It is quoted (or else some second treatise on music by the same author) as the *Breviloquium Wolstani*, at the end of the 13th century. That our nation had a considerable supply of hymn tunes and ritual music peculiar to itself in the latter half of the tenth and the early part of the eleventh centuries is proved by the rich collection of manuscripts in the Bodleian and other libraries. It was for some time claimed that the English made use of a Staff of four lines and spaces prior to the appearance of Guido d'Arezzo, on the grounds that a staff of this kind is found in the Winchester Troper over a litany of the time of Æthelred (A.D. 979—1016). But a careful examination of the manuscript showed beyond doubt that these lines and spaces had been added by a later hand. This Troper contains also some tunes whose names prove them to have been popular melodies: "Berta vetula" (The old woman Bertha), "Candidi planctus cygni" (The lament of the white swan), the melodies of which, it is hoped, may some day be recovered. (See "The Winchester Troper," by the Rev. W. H. Frere, and also Mr. E. W. B. Nicholson's Preface to "Bodleian Fac-similes," Vol. I.) As to secular music, we find in the *Gesta Herwardi*, or the Life of Hereward, who was son of Leofric, Earl of Mercia, and the Lady Ædiva (the Lady Godiva of popular fable), that he seized the harp and sang "with correct musical intervals" (for that is the meaning of "per discrimina vocum"[*]), sometimes alone, and at other times in three parts with his companions after the manner of the Gyrwians.[†] These Gyrwians were the inhabitants of the fenny districts between East Anglia and Mercia, including Peterborough (then called Medeshamstede) in the north, and the Isle of Ely in the south. This was not a district likely to be in advance of the rest of England, and yet, even here, we read of singing in three parts as customary.

---

[*] Anglo-Saxon "Scop" or "Sceop."

[†] Sequentia magis cantilenis per successiones temporum detritis, quam libris ad instructiones posteriorum elucubratis, didicerim." (*De Gestis Regum Anglorum*, Lib. 2, cap. 6.)

[‡] "Inde merito, jureque, culpant eum literæ; nam ceteras infamias, quas post dicam, magis resperserunt cantilenæ." (*De Gestis Regum Anglor.*, Lib. 2, cap. 8.)

When Archbishop Alfric wished to translate the Latin word "Concentor" for his vocabulary of Latin and English words, he rendered it by "mid-singend," which seems sufficiently to express singing in three parts, for there would be no middle in two or four. It might perhaps have been taken to mean "singing all together," or in "chorus," if Alfric had not also given two different translations of "Chorus,"[‡] besides others for "Song," "Duet," "Tune for an instrument alone," "Harmony," "Discord,"[§] &c., all which, taken together, are sufficient to prove the very early cultivation of music in England.

These notices of early music may not appear, at first sight, to be immediately connected with our text of "Old English Ballads," but the arts of music and poetry were then united, all poetry being intended to be sung; and there is so much new matter to be adduced in the history of music, especially in that of our own country, of which Dr. Burney's account is most inaccurate, that it is difficult to avoid the temptation of referring to the subject. If we desire to prove that music was cultivated by the working classes as well as by those above them, we may quote the fact of the Watermen of London having made a round for three voices, in honour of Sir John Norman, Lord Mayor of London, who, in 1453, commenced the custom, which became afterwards established, of going to Westminster in his barge to be sworn into his office of Lord Mayor, instead of riding both to and fro with a procession on horseback as before. The music of the Round[‖] is like the chiming of bells from one church steeple to another, and might be sung by hundreds of men together to the words,

> "Heave and ho, rumbelow,
> Row the boat, Norman, row,
> Row to thy Leman."

The idea of representing the taking charge of the City of London, as a "leman" or "loved one," was quite watermanic.

The people were then fond of singing, and altogether more cheerful than after the advent of that severe puritanism which told them, in the words of Prynne, to "go about chattering like cranes, and cooing like doves for their own *and others'* sins."[¶] Solomon thought that there was a time for everything—"a time to dance, a time to sing, and a time to play;" but these wise-acres did not. They put down the Maypoles and the dances on the village green, and thus reduced the people to drinking and to earnest politics as the only excitements left to them.

The character of "Merry England" will compare favourably with that of "Old England"—for England had not the title of "Old" until a "New England" had been planted in America, and puritanism had become both rampant and dominant at home.

"The merry, free, and frank disposition of the Old English," says Camden, "was thus described by Alfred of Beverley" (who died A.D. 1136): "England, *full of sports*, a free people, delighting in jokes."[**] In the same

---

[*] It is a quotation from Virgil's "septem discrimina vocum," and one frequently employed to express the seven intervals of the diatonic scale, viz., A, B, C, D, E, F, G.

[†] "Multipliciter cum ea [cythara] canendo, et per discrimina vocum, nunc solitarie, et nunc tripliciter cum suis sociis, more Girwiorum, cantavit." (From a photographic copy of the Peterborough Manuscript.)

[‡] "Singende heap" and "Hluddra sang."

[§] "Sang," "twegra sang," "answege sang," "geð—waere sang," and "ungeswege sang." (*Vocabularies* edited by T. Wright, F.S.A., p. 28, privately printed by Joseph Mayer, Esq., of Liverpool, F.S.A., &c)

[‖] The easy music of this little Round is printed in *Popular Music of the Olden Time*, Vol. II. p. 783.

[¶] Prynne was parodying Hezekiah's words when he thought himself dying: "Like a crane or a swallow, so did I chatter: I did mourn as a dove:" but Hezekiah was mourning for his sickness, and not for the sins of others.

[**] "Anglia, plena jocis, gens libera, et apta jocari."—*Camden's Remaines.*

Digitized by Google

strain runs William of Malmesbury, referring to the Norfolk and Suffolk men, or East Anglians: "they are a merry, pleasant, jovial race, but apt to carry their jokes to an irritating excess."[*] For a third testimony we may take an extreme part of England: "Merry Michael, the Cornish poet, piped this upon his oaten pipe, for Merry England," says Camden:

"For money, dinners, varied drinks, no land will e'er be found
Like England, famous England, where the fertile soil is crown'd
With countless flocks and herds, and where all social joys abound." [†]

We know from another source that there was no lack of tunes when the Normans came, for Thomas, the first Norman Archbishop of York (1070), set about collecting those which he heard from the minstrels, and wrote hymns to them.[‡] Richard de Ledrede, a Londoner, who was Bishop of Ossory, from 1318 to 1360, did the same thing; but carried the tunes to Ireland with him. We know the names of the ballads, because they are written in the Red Book of Ossory over his Latin hymns. Among them are "Sweetest of all, sing!" "How should I with that old man?" "Do, do, nightingale, sing full merry," and "Good day! my leman dear." Thus he anticipated the Rev. Rowland Hill, or whoever else may have said, that "the Devil should not have all the pretty tunes." And yet there was some danger from this appropriation of secular words, lest they might become so fixed in the memory as to crop up unexpectedly and unwittingly. Giraldus Cambrensis relates a case that should have been a warning. It is of a priest in Worcestershire, who had been listening to choral singing and dancing near the church during the night, and who, in pure forgetfulness, sang one of these popular burdens in the morning, instead of greeting the people with "Dominus vobiscum."

As to London, the first good description of the city and of its customs was written in 1174 by Fitz-Stephen (Stephanides), the friend and biographer of Thomas Becket. He says that "in summer evenings the young people danced till dark, to the sound of the harp (or cittern), and that some of the maidens acted as the musicians."[§] Also that, on festival days, the boys of the London schools attached to the three principal churches "contended with each other in verse," and wound up their contests "by recitations of epigrams, *ballads*, and rhymes, in which the foibles and frailties of their fellows were sarcastically exposed, without naming the individuals." At this "the auditors, who were prepared to enter into the jest, shook the assembly with peals of laughter."

These are gayer pictures and of more content, than are common now. Examples might be continued to the extent of a volume, but one more, from Oxford in the reign of Queen Elizabeth, will suffice. It was written by the learned Dr. John Case, whose *Speculum moralium Quæstionum in universam Ethicen Aristotelis*, was the first book printed at the new press at Oxford in 1585. The extract is from *The Praise of Musicke*, printed at Oxford by John Barnes, in the following year: "Every troublesome and laborious occupation hath musick for a solace and recreation, and hence it is that the wayfaring men solace themselves with songs, and ease the wearisomness of their journey; considering that musicke, as a pleasant companion, is unto them instead of a waggon on the way. And hence it is that manual labourers and mechanical artificers of all sorts keep such a chanting and singing in their shoppes—the tailor on his bulk, the shoemaker at his last, the mason at his wall, the shipboy at his oar, the tinker at his pan, and the tiler on the house-tops." Even the proverbially merry cobler has now almost ceased to sing, and tailors seek only to mend the State. A tuneless tailor, in former days, was such a *rara avis* as to become at once an object of suspicion. "Never trust a tailor that does not sing at his work," says Fletcher, "for his mind is of nothing but filching." The treatment of the poor was perhaps less considerate than now; but the people having their amusements were certainly more content. The number of ballads left for entry at Stationers' Hall at the end of year 1560 was 796, and only 44 books.

We have still a large number of extant ballads, such as were printed on one side of a sheet of coarse paper, to be sung about the streets and villages in the 16th and 17th centuries. Their tunes are also to be found, being included in early collections of country dances. Ball and ballad are words derived from the same root, and when the people danced country-dances they accompanied them with song. Nearly every old ballad

---

[*] "Gens læta et lepida, facetaque festivitate jocorum ad petulantiam pronior." (*Gesta Reg. Anglor.*, Book 2, cap. 13.)

[†] "Nobilis Anglia pocula, prandia donat, et æra,
Terra juvabilis et sociabilis, agmine plena:
Omnibus utilis, Anglia fertilis est et amœna."—
Camden's Remaines.

[‡] "Si quis in auditu ejus arte jocularia, aliquid vocale, sonaret, statim illud in divinas laudes effigiabat."—*W. Malmesbury.*

[§] "Puellarum cithara choros ducit usque imminente luna, et pede libero pulsatur tellus." (*Descrip. Lond.*, ed. T. Pegge.)

Digitized by Google

has the name of the tune printed upon it for which it was intended, and it has been owing to this combination of circumstances that so many of our national airs have been recoverable, and that words and tune could be re-fitted together in authentic forms. Mere tradition is the frailest of guides in music, for hardly do any two untaught singers sing an air alike, and they often vary the tune between one stanza and another.

Captain Cox, the Coventry mason, is the first recorded collector of old printed ballads. He is mentioned by Laneham in his letter from Kenilworth in 1575. The next in order of date is the learned Selden. He lent his collection to Samuel Pepys, the amusing diarist, who did not return it. We are, in all probability, indebted to that circumstance for its preservation; for Pepys left his library, including his collection of ballads and those borrowed from Selden, to Magdalene College, Cambridge, where they remain under the strictest custody, owing to the terms of the bequest. Old Pepys took the greatest care to prevent others from indulging in his own little habit of filching.

The united collections of Selden and Pepys (or of Pepys including those ballads that he borrowed from Selden) are bound in five folio volumes containing 1785 ballads, mostly with second parts. Sometimes two ballads were printed upon one page, and these would enlarge the above named number. In addition to the broadsides there are three volumes, lettered "Penny Merriments," which were also collected by Pepys, and which include a large number of "Garlands," that are in themselves collections of ballads, but printed in octavo or other small size, instead of "in broadside," *i.e.*, on one side of a folio page.

A second great collection of broadside ballads is that which is now called the Roxburghe Collection. John, Duke of Roxburghe, was only one of several proprietors through whose hands the collection successively passed; but his name became especially connected with it, owing to the notoriety of the comparatively large price it produced at the sale of his library. The collection had been purchased for the Duke at the auction of Mayor Thomas Pearson's library in 1788 for £36 14s. 6d., and was resold in 1813, with the duke's additions, for £477 15s. It was originally formed by Robert Harley, who was raised to the peerage as Earl of Oxford and Mortimer in 1711, the same whose magnificent collection of manuscripts, known as the Harleian Collection, is one of the wonders of the British Museum.

This collection consists of four volumes, containing 2133 pages of ballads. Sometimes one ballad takes two pages including its second part, and sometimes two ballads are printed on the same page. It was purchased for the British Museum at the sale of the library of the late Benjamin Heywood Bright, M.P., in 1845.

Another important collection of ballads in the British Museum is that formed by Bagford, who was agent in purchasing for Harley, Earl of Oxford. It is bound in three volumes. There are also many minor collections in the same library, and a large number of political ballads and songs among "The King's Pamphlets."

For early date there are no extant collections to compare with those of Mr. Henry Huth, Mr. S. Christie-Miller, and of the Society of Antiquaries of London. These are unrivalled for rarity, but they are not of so large an extent as some others.

The Bodleian Library at Oxford is particularly rich in ballads, and the Public Library at Cambridge particularly poor, if, indeed, it possess any collection at all. Oxford can boast of the Douce collection, which is, perhaps, next in extent to the Roxburghe and Pepys, but rather later, as to average date, than either. It contains 877 ballads bound in 4 vols., the fourth volume being later than the rest. The Bodleian also possesses Anthony Wood's famous collections, both in print and in manuscript, as well as a smaller number of printed ballads collected by Rawlinson. Wood's printed collection is of 279, and the Rawlinson of 218 black-letter ballads, and (as a rough guide to the number of duplicates to be found in the great public libraries) it may be stated that although the Roxburghe Collection contains about ten times the number of the Rawlinson, yet the latter includes 130 ballads, of which no edition whatever is to be found in the Roxburghe. Yet they are generally of coeval dates.

The Cheetham Library, Manchester, possesses an extensive collection of ballads presented by James Orchard Halliwell, F.R.S. Of other collections in private hands, it may be sufficient to name first, that of the late W. Ewing, F.S.A., Scot., which, according to the printed catalogue, contains 408 ballads; a collection at Osterley Park; and a rare collection formed by Mr. J. Payne Collier, and now in the possession of Frederic Ouvry, Esq., Treasurer to the Society of Antiquaries.

There are, no doubt, many more collections in private hands, as well as many ballads scattered in collections of pamphlets, both in public and private libraries; but even in those already named, the number of extant English ballads dating from the reign of Henry VIII. to the year 1700 cannot be computed at less than ten thousand. It would be much larger, if ballads printed with music were taken into

Digitized by Google

account, or even if manuscripts, like the Percy folio, and Wood's collection, were included in the calculation. The tunes for some thousands of them have been traced, and many are printed in the "History of Popular Music of the Olden Time." "In a word," says an old writer, "scarce a cat can look out of a gutter, but up starts a halfpenny chronicler, and presently *a proper new ballad of a strange sight* is indited."

Ballade (*Ger.*) A dance, also a ballad.

Ballata (*It.*) The melody of any song which may furnish a tune for dancing. [Ballet.]

Ballatetta (*It.*) Diminutive of Ballata.

Ballematia.
Ballistia. } Songs in dance-style.

Ballet. A Madrigalian part-song with a *fa la* chorus. The "Ballets" or "fa las" of Giovanni Gastoldi [1532-1598] the reputed originator of this form of vocal music, are in most cases in simple counterpoint — note against note—but the rhythm, strongly marked and well defined, is admirably suited to the purposes of the dance which these vocal harmonies were intended to accompany. There are many examples of Ballets to be found in the writings of the Elizabethan madrigal composers.

Ballet (*Fr.*) A representation in dancing and gesticulation, of some story, without words.

The rise of the Ballet is almost coeval with dancing itself, for it is difficult to believe that any number of dancers could have so disported themselves as to give delight to the spectators, if there had not been some definite and organised arrangement. The dances described as having been led by Miriam, David and Jephtha's daughter, the Emmelœia, the Pyrrhic dances, the Motions of the Mimes, Minstrels, and Joculators, and the homely dances popular among the peasantry, besides the more stately measures favoured by people of high degree, were all ballets—in which certain motions were made to the sound of music, and whose gestures and actions had meanings and intentions that were commonly understood. When these dances were transferred from home circles to the stage, the gestures and actions made were such that could be readily interpreted by the lookers-on, and even when great skill was acquired by the performers in following ages, the old conventional signs, attitudes, and motions were retained, that all who chose might understand. The first ballets on the stage were those that were introduced into the oratorios, masques, and comedies, each being a development of portions of certain entertainments, from which they arose in common.

The oratorio and the drama arose from the ancient sacred and classical plays and the mediæval mysteries and moralities, and circumstances gave importance to particular parts of those productions, so that from forming a continued or dependent whole, they became detached and separate, and made what seemed in later years distinct things of those that had a common origin. The splendour of the Court Masques, the glory of the unity of the genius of the poet, architect, painter, and musician, are matters of history. Give prominence to the music, let the poetry fade away from inanity, retain the skill and genius of the machinist and scene painter, and you have opera. Let your poet write prose, have as little music as possible, respect the scenic effects and mechanical means, and drama is the result. Dispense with poetry or words of any kind, make music subservient, but do all that can be done with scenery and machinery, and make the dramatis personæ bound, caper, and gesticulate, and ballet is the product.

The Ballet had its origin in the Masques, which were written for and often performed by princes and other distinguished personages. In England, Italy, and France, it arose almost simultaneously out of the remains of the Masque. Count Aglio, at Turin, invented pieces that were at the same time pastoral, mythological, allegorical, and fantastic, in which the princes of the Court took part. In France Louis XIII. danced in a ballet, and his successor, Louis XIV., did the same in his turn, these ballets being portions of spectacles that were operatic, dramatic, and terpsichorean by turns or in combinations. Antoine de la Motte improved the ballet, and made it distinct, and independent of other means for explanation and elucidation; he also encouraged the introduction of female dancers, till then almost unknown in Europe, and from that time the ballet gradually sank from importance and consideration, and became a mere exhibition of artificial agility and natural comeliness. It is not many years since it was a very considerable item in the evening's entertainment at the opera in England, rivalling in spectacular splendour the famous ballets of Milan, the absence of vigour and intellectual power in the operas produced being counterbalanced by the so-called glory of the ballet. But as people began to be alive to the fact that contortions, dislocations, and indecent postures were the reverse of elevating or instructive, and not really amusing, the patronage of the ballet as a distinct entertainment fell away and finally ceased altogether, and an attempt to revive it apart from and out of the course of the situations of an opera, during the season of 1871, met with so little encouragement that it was silently abandoned. The ballet was once a poem and a power: kings did not scorn to exhibit trained and practised personal skill for the edification

Digitized by Google

of their loyal subjects, and the exaltation of the exercise in which they indulged; but the "improvements" of De la Motte introduced an element which was at once the cause of its glory and of its shame, its culmination and contempt. [Dance.]

Balletto (*It.*) A ballet, a dance.

Ballet-master. One to whom is entrusted the direction of the motions of the ballet, and the order of the performers.

Balli Inglesi (*It.*) English dances.

Balli della Stiria (*It.*) Styrian dances.

Balli Ungaresi (*It.*) Hungarian dances. Dances in the Hungarian style.

Ballo (*It.*) A dance, a ball.

Ballonchio (*It.*) [Passamezzo.]

Ballonzare (*It.*) To dance wildly, recklessly, without rule.

Band (*Ger.*) A part, a volume, any thing sewn together.

Band. Instrumentalists collected together for the performance of music.

(1) Brass Band. A collection of players on brass wind-instruments.

(2) String Band. (*a*) That portion of an orchestra which consists of players on stringed-instruments of the violin family. (*b*) A band consisting only of instruments played with a bow.

(3) Wind Band. *Stromenti di fiato* (*It.*) That portion of an orchestra which consists of players on flutes, oboes, clarinets, bassoons, and horns; but not on trumpets, trombones, and other loud brass instruments, these being included only under the sign "tutti."

(4) Wood Band. The players on the wood wind-instruments, flutes, oboes, clarinets, bassoons; but not on the serpent, which is usually classified with brass instruments.

(5) Military Band. A number of musicians belonging to a regiment in the service of the King or Queen of a country. In England, those who perform upon instruments provided by the officers of the corps to which they belong, the military regulations only recognising side drums, fifes, bugles, and trumpets, as necessary; these are supplemented by clarinets, flutes, cornets, bassoons, horns, trombones, ophicleides, bombardons, triangles, cymbals, big drums, &c., and the combination is known as a military band. The band is conducted by some one skilled in directing and arranging, who is usually a civilian, but the bandsmen over whom he presides are in every other respect the same as private soldiers. Military bands are sometimes used upon the opera stage, for the purpose of gaining an increased effect.

Banda (*It.*) A military band. [Band § 5.]

Bandora (*Fr.*) [Bandore.]

Bandore. An English form of the ancient **Greek** Pandoura, having twelve strings of steel-wire. The Bandore is said to have been invented by John Rose, of London, in 1561.

Bandurria (*Sp.*) A form of guitar, strung with wire instead of cat-gut.

Banja or **Banjo.** A stringed instrument of supposed African origin, popular with the negroes of America, and one of the most important musical instruments employed by troupes of fictitious negroes. The instrument consists of a handle, which, running the whole length, serves at once as finger-board, as support for the hoop, over which a skin, acting as sounding board, is stretched, and also as a hold for the pegs which tighten and keep the strings in tune. The banjo is strung with five strings, so arranged that they may be *stopped* in the ordinary way to vary the melody or harmony, together with an octave string which is never stopped. The tuning, which may be in any key, is generally according to the following plan:

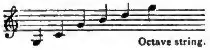

The character of banjo music is sprightly and well adapted for dancing, for which it is as often used as it is for accompanying the voice.

Bar. A line drawn from the top to the bottom of the stave to denote the division of the time in a piece of music, and the place of the strong accent. Each portion comprised within two of these lines is also called a bar. In mediæval music the bar, also called the *lesser* bar, to distinguish it from the greater or double bar, was often used solely for the purpose of showing the end of a line or sentence of the words, hence it was said "to give time for the whole choir together to draw breath" (Nievers, sur le Chant Gregorien); whereas the two great bars or the double bar is "the most efficacious contrivance that can be thought on to remedy all the cacophonies and contrarieties in the voices of the singers, who without them could not guess when to rest." (*Ibid.*) Mr. Chappell ("Popular Music of the Olden Time") remarks that the Tunes in the "Dancing Master," printed in 1651 in only a single part, have no bars, but that the score of the moral play, "The four elements" (to which Dr. Dibdin has ascribed the date 1510), is barred. He further adds, that so far as he has observed, all music in the ordinary notation, even for one voice or one instrument, was barred after 1660. It is probable, however, that the regular barring of music had its origin in the system of Tablature, in which its efficacy as a means of pointing out the position of *accent* must have been generally observed. But for a considerable period after the introduction of the bar, its use in ecclesiastical music was very irregular, two, four, six, or even eight minims being included in a bar which, by the signature, should contain only four. In modern music-printing and engraving, care is taken that the bars in the separate lines forming a score shall stand exactly over one another. The neglect of this in early publications in score adds greatly to the difficulty of reading them.

Barbet. [Barbiton.]

Barbiton. An ancient Greek instrument said to have been invented by Anacreon; it was in the form of a lyre, and had seven strings. The name was applied to instruments of the violin class in the 16th and 17th centuries.

Barcarolle (*Fr.*) A simple melody, composed in imitation of the songs of the Venetian gondoliers, many of which are of striking beauty. Apolloni Salvadore, a Venetian barber and fiddler in 1720, is named as the composer of several popular tunes of this class.

Bard. A name given to hereditary poets and minstrels by all the Celtic nations. In their songs and poems the bards recorded the deeds and prowess of the warriors, kings, and people, at festive and social gatherings; and at religious assemblies they celebrated the acts and fame of the gods and heroes, accompanying their songs with the harp and crowd or crwth. The power, reputation, and influence of the bards were very great, and the favour of kings, princes, and nobles was accorded to them. They, like the Aoidoi of ancient Greece, were the historians, poets, and chroniclers of their time; they incited their armies to courage in the hour of battle, and by their heroic strains roused the fury and valour of the warriors. In time of peace they were ambassadors, heralds, and the depositaries of all historical tradition, and of much of the learning that was at that time possessed by the community. As an institution they kept longest influence in England, Scotland, and Wales. In the last-named country their privileges were fixed by King Howel Dha, A.D. 940, and a century and a half later Griffith ap Conan revised and reformed the whole system. The Eisteddvodau, as the congregations of Welsh bards are called, were held from time to time until the conquest of Wales by Edward I., in 1284, when the bards were persecuted, and as some authors declare, were put to death. Although the power of the bards was broken, still their Eisteddvodau were encouraged by the rulers of succeeding generations, until the time of Queen Elizabeth, who was the last monarch who made any concessions to the bards of Wales. The preservation of such of their literature that has survived them is owing to several learned societies who have made this object their peculiar study. An Eisteddfod as now understood is a curious compound of heterogeneous matter, for although it is held for the purpose of encouraging national art in music and poetry, its judges appear to be satisfied with the least worthy effusions common to London concert-rooms, provided they are given through the medium of the Cambrian tongue. Little is known of the bards in Scotland beyond the fact that they were similar in constitution to the bards in Ireland, who, like those of Wales, were a hereditary community. The Irish bards were divided into three classes, the Filhedha, the Braitheamain, and the Senachaidhe. The first sang the sacred and heroic songs, and were employed as heralds and counsellors, the second recited and expounded the laws, and the third were the chroniclers and recorders. They were endowed with many privileges and had great influence, and their power over the minds of the people was so strong that severe measures had frequently to be resorted to, to check their sway. In England bardism took a more refined and less exciting form than that which was acceptable to those nations of more strictly Celtic origin. [Minstrel.] [Ballad.]

Bardone. [Viola di Bordone.]

Barem (*Ger.*) An organ-stop, consisting of closed flute-pipes of 8 ft. or 16 ft. pitch, of a soft character of tone.

Barginet, Berginet, Bargaret, or **Bergeret.** Shepherd's songs, to accompany dances. Songs relating to pastoral matters.

> "A bargaret in praising the daisie.
> For, as methought, among her notes swete,
> She said, 'Si douce est le Margarite.'"
> CHAUCER.—*Floure and Leafe.*

Baribasso. A deep bass voice.

Bariolage (*Fr.*) A medley. A cadenza, or series of cadenzas, whose appearance forms a design upon the music paper, "a waistcoat pattern" as it is called by performers.

Baritenor. A deep tenor.

Baritone. A brass instrument of deep tone. [Metal Wind-Instruments.]

Bariton, or **Baryton.** [Viola di Bordone.]

Bariton Clarinet. An instrument used in military bands, the tone of which is between the clarinet and bassoon.

Baritone Clef. The F clef placed upon the third line of the stave.

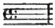

It is not now used, but was frequently employed in vocal music of the 16th and 17th centuries. Purcell's Song, "Let the dreadful engines," was originally written in this clef, and it was also used occasionally for horn parts by Handel, Cooke, and other writers in the 18th century.

Bariton (*Fr.*)
Baritono (*It.*) The barytone voice.

Baritone voice. [Barytone.]

Barocco (*It.*)
Barock (*Ger.*)
Baroque (*Fr.*) Unusual, singular, eccentric, whimsical, irregular. Applied to a composition with over-chromatic harmonies, or unrhythmical melodic phrasing.

Bärpfeife (*Ger.*) [Baarpÿp.]

Barquarde (*Fr.*) An obsolete term for Barcarolle, *q.v.*

Barré (*Fr.*) In guitar or lute playing, the pressing of the fore-finger of the left hand across all the strings, so as to alter temporarily the pitch of the instrument, the remaining fingers being at the same time engaged in forming a chord. The first finger, therefore, performs the duties of a capotasto, *q.v.*

Barre de luth (*Fr.*) The bridge of the lute.

Barre de mesure (*Fr.*) [Bar.]

Barre de répétition (*Fr.*) A double bar with points, marking a repeat.

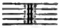

Barrel. A revolving cylinder of wood or metal.

(1) Barrel-organ. An organ in which a wooden cylinder furnished with pegs or staples, when turned round, opens a series of valves to admit a current of air to a set of pipes, producing a tune either in melody or harmony. The barrels are sometimes made moveable, in order to obtain a variety in the tunes, as the capability of a single barrel is necessarily limited. Barrel organs furnished with hymn and psalm tunes, or even voluntaries, were sometimes used in places of worship, but the increased knowledge of music, even in remote places, has led to the introduction of the harmonium, which has superseded the use of barrel organs to a great extent. The tone of barrel organs is incapable of expression or variety, and has consequently been found seriously monotonous. The only advantage belonging to the instrument is its portability, and this renders it available for street musicians, who generally hire one at a small charge, the cost of the instrument (from £20 to £70) being beyond their means. Many of the poor hirers are cruelly used by the Padrone from whom they obtain their instruments. The barrel-organ, as a street entertainment in London, dates from about the year 1790. The stops in a barrel-organ generally consist of a stopped diapason and flute or principal, to which is sometimes added a reed stop of coarse quality. The compass rarely exceeds two octaves and a half.

(2) Barrel of a musical box is constructed in a manner somewhat similar to that of an organ, but is of metal, and instead of opening a series of valves, the pegs and staples set in vibration the teeth of a steel comb, which produce the sounds. [Musical box.]

Barypycni (*med. Lat.*, from *Gk. βαρὺς deep*, and *πυκνὸς close*). (1) Lowest strings of tetrachords in the chromatic or enharmonic scale. (2) In ecclesiastical music, those modes which have the *pycnon* or semitone at the bottom of the tetrachord, *e.g.*:

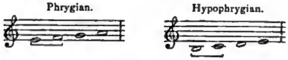

see mesopycni, oxypycni.

Barytone voice. A voice of fuller quality than a tenor and lighter than a bass, having a compass partly included in both, namely,

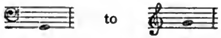

This is the extreme compass, and both limits are rarely reached.

This voice has only been distinguished by name, as being of a separate character, within the present century. Early writers indicate its existence by the use of its special clef. The term Barytone is unmeaning, unless it be looked upon as a corruption of Barytenor, but it is quite possible it was borrowed from the instrument Barytone or Bordone, which occupied a place between the tenor and bass viols. Rousseau calls this voice *Basse-chantante*, or *Basse-taille;* and Shield, in his "Introduction to Harmony," having used the word Barytone, thinks it necessary to explain in a foot-note that it is "a voice between a tenor and a bass."

Bas-dessus (*Fr.*) Mezzo soprano, or second treble.

Base. Old form of the word Bass.

Bass. Low, as bass trombone, bass viol, bass voice, &c.

Bassa ottava (*It.*) At the lower octave.

Basse (*Fr.*) Bass.

Basse chantante (*Fr.*) A barytone voice.

Basse chiffrée (*Fr.*)
Basse continuée (*Fr.*) A figured bass, a bass part, the accompanying harmonies of which are expressed by numbers.

Basse contrainte (*Fr*). [Ground bass.]

Basse contre (*Fr.*) A deep bass voice, capable of singing below the ordinary bass part.

Basse de cremone (*Fr.*) The bassoon.

Basse de hautbois (*Fr.*) Corno inglese.

Basse d'harmonie (*Fr.*) The ophicleide.

Basse de viole (*Fr.*) The violoncello

Basse de violon (*Fr.*) Double-bass.

Basse double (*Fr.*) Large double-bass.

Basse figurée (*Fr.*) Figured bass.

Basse fondamentale (*Fr.*) Root-bass or generator. [Harmony.]

Digitized by Google

Basse recitante (*Fr.*) [Basse chantante.]

Basse-taille (*Fr.*) The Barytone voice.

Basset-horn, Corno di Bassetto (*It.*) A transposing instrument of the clarinet order, of a beautiful, soft, and rich quality, invented in Passau about the year 1770, and improved by Lotz of Presburg twelve years later. In form like a long clarinet, with a curved and bell-shaped metal end. The compass extends from F below Gamut to C in Alt.

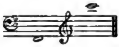

with all the intermediate semitones, except the F sharp and A flat in the lower range.

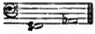

The music is written for it in the bass and treble clefs a fifth higher than the real sound. Mozart has written with brilliant effect for the basset-horn in his "Nozze di Figaro," in "Clemenza di Tito," and in the "Requiem."

Bassetto (*It.*) (1) The diminutive of Basso. A name sometimes given to the tenor violin. (2) A reed stop in the organ of 8 ft. or 16 ft. in length.

Bass Flute. The lowest in pitch of instruments of the flute family, now obsolete. Its compass was 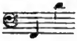. It was a *flûte à-bec*, not a *flauto traverso;* that is, it was blown at the end (like a flageolet), not at a hole in its side. In order to enable the player to reach the remote holes with his fingers, a bent tube turning upwards conveyed the air from his lips to the mouthpiece of the instrument.

Bassgeige (*Ger.*) Bass Viol.

Bass Horn. (1) Deep B♭ Horn. (2) An instrument which was a precursor of the Ophicleide.

Basslaute (*Ger.*) [Bass Lute.]

Bass Lute. [Theorbo.]

Basso (*It.*) A bass singer, also the double-bass, and the bass part.

Basso buffo (*It.*) A comic singer, with a bass voice.

Basso cantante (*It.*) [Basse chantante.]

Basso concertante (*It.*) The principal bass, that which accompanies solos and recitatives.

Basso continuo (*It.*) A bass part figured for the organ or pianoforte.

Basso figurato (*It.*) (1) Basso continuo. (2) A bass part, with running passages.

Basso fondamentale (*It.*) The fundamental ground bass, or root.

Basson (*Fr.*) [Bassoon.]

Basson quinte (*Fr.*) A bassoon, the pitch of which is five notes higher than that of the common bassoon. The part given to it must therefore be written five notes lower than the actual sounds required. Its *written* compass is 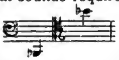 sounding 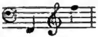 including all the intervening semitones. Its tone is more powerful, but less sympathetic, than that of the *corno inglese.*

Basso numerato (*It.*) A bass, the accompanying harmonies to which, are indicated by numbers.

Bassoon. Basson (*Fr.*) **Fagotto** (*It.*) A reed wind-instrument of deep pitch, with a compass of more than three octaves from low B flat.

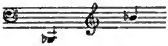

This compass includes all the intermediate semitones, with the exception of

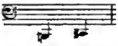

which are as yet to be obtained only from instruments of improved construction. Some performers can produce three notes higher than the B flat, but for all common orchestral purposes they are unnecessary. The bassoon ordinarily forms the bass or deepest tone among wood wind-instruments, and is capable of excellent independent effects, among which the grotesque ought not to be forgotten, as in Beethoven's "Pastoral Symphony," and the "Clown's March," in the music to the "Midsummer Night's Dream," by Mendelssohn.

It is customary to write for the Bassoon in the Bass clef, and as the instrument is usually employed in pairs, one stave serves for the two parts. The tenor clef is often employed for the higher notes of the register of the Bassoon, sometimes in a separate stave.

Some writers assert that the Bassoon is the invention of Afranio of Ferrara in 1540, and that he gave it the name Fagotto from its resemblance to a bundle or fagot of sticks (his instrument being made of several pieces laid together), but it was known long before under the name of Buisine, Buzaine, Courtal, Bombard, or Wait.

There is reason to believe that the Bassoon is of Eastern origin, introduced into western Europe in the twelfth century, and that it is an improvement of the drone-pipe of the Bagpipe. The Egyptian word for a pipe of deep tone, and for the drone of the Bagpipe is, according to E. W. Lane ("Modern Egyptians"), *Zummarah-bi-soan*, and the manner in which the word *Buzaine*, *Buisine*, is used in mediæval MSS., shows a possible connection with

Digitized by Google

this origin. The instrument was introduced into the orchestra about the commencement of the 18th century; for a long time it was employed to strengthen the voice parts only. Handel generally makes it double the bass voice part, or treats it as a bass to the oboe; he has, however, made excellent use of it as a solo instrument in the scene of Saul and the Witch of Endor, in his oratorio of "Saul."

Bass-Bar. A piece of wood fixed inside the belly of violins, &c., to support the pressure of the left foot of the bridge.

Basso ostinato (*It.*) [Ground bass.]

Basso ripieno (*It.*) The bass of the full or chorus parts.

Basspommer, called also **Bassetpommer** (*Ger.*) A deep-toned instrument of the oboe family, precursor of the Bassoon.

Bassposaune (*Ger.*) Bass trombone. [Trombone.]

Bassschlüssel (*Ger.*) The bass clef.

Bassthema (*Ger.*) [Ground bass.]

Bass Trumpet. An old instrument, now superseded by the trombone.

Bass Tuba. A brass instrument, a species of bombardon, not capable of such rapid execution as a bass ophicleide, but producing a much finer quality of tone. The name now given to the lowest of the saxhorns; there are two kinds, one is in E♭, another in B♭; the former shaped like a large euphonium, the latter circular and passing round the neck of the performer. The lowest note of the E♭ instrument is the E♭ in the 16 ft. octave; that of the B♭, a fourth lower.

Bass Viol. (1) A familiar name for the violoncello. (2) The largest and deepest in tone among a chest of viols, which had five and sometimes six strings, and a fretted finger-board. The manner of tuning the open strings varied according to the music to be played.

Playford (Introduction to the Skill of Music) mentions three sorts of Bass viols "as there is three manner of ways in playing." "First, a *Bass viol* for *consort* must be one of the largest size, and the strings proportionable. Secondly, a *Bass viol* for *divisions* must be of a less size, and the strings according. Thirdly, a *Bass viol* to play *Lyra-way*, that is by *Tableture*, must be somewhat less than the two former, and strung proportionably."

The common *accordatura* of the six-string instrument was as follows:

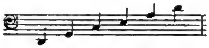

Bass Voice. The lowest register of the human voice, having a compass ranging between two octaves from lower D:

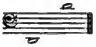

The whole of the bass voice should be produced from the chest, and the most useful notes, and those generally written are between G and tenor C:

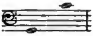

A bass voice rarely reaches full perfection of quality or sonorousness before the possessor is thirty years of age, and a true bass voice has seldom much flexibility.

Batillus. An instrument formerly employed by the Armenians in their Church service to supply the place of bells, which they were forbidden to use. A board struck with a hammer.

Bâton (*Fr.*) (1) A stick used in beating time. (2) The method of a conductor is called his *bâton*. (3) A pause of two or more bars is also so named, *e.g.* 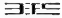 a *bâton* of five measures or bars.

Battement (*Fr.*) An ornament in singing, opposed to the Cadence (*Fr.*) *e.g.*:

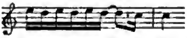

is called a *cadence*, whereas the following

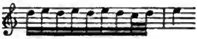

is a *battement*. [Beat.]

Battere, il (*It.*) The down-stroke in beating time.

Batterie (*Fr.*) A roll upon the side drum.

Battery. An effect in harpsichord music, written 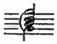 and played 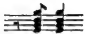

Battimento (*It.*) [Battement.]

Battuta (*It.*) (1) In correct time. (2) A bar.

Bau (*Ger.*) The structure of musical instruments.

Bauernleyer (*Ger.*) [Hurdy-gurdy.]

Bauerpfeife (*Ger.*) An organ stop of 8 ft. length of a small scale.

Baxoncillo (*Sp.*) An organ stop like an open diapason.

Bayadères. Dancing girls attached to a Hindu temple.

Bayles (*Sp.*) Comic dancing songs, in the Spanish gipsy dialect. [Ballad.]

B cancellatum (*Lat.*) The cancelled B. The note B♭ as altered by means of a ♮ or ♯ in old music. Up to the middle of the 18th century the ♯ frequently had the force of the ♮ as now used.

Digitized by Google

B dur (*Ger.*) The key of B flat major.

B durum (*Lat.*) B natural.

Bearbeitet (*Ger.*) Arranged, adapted.

Bearings. Those few notes which a tuner accurately tunes or *lays down* before proceeding to adjust the whole compass of the instrument.

Beat. (1) A short shake, or transient grace note, played or sung before the note it is desired to embellish. The beat is always a semitone lower than the ornamented note.

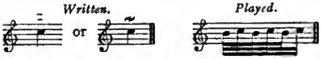

(2) The portion of a bar of music occupied by the movement or supposed movement of the hand in counting time. Thus, a beat in $\frac{6}{8}$ time is equal to three quavers; a beat in $\frac{3}{2}$ time is equal to a minim.

(3) The peculiar "throbbing" heard when sounds not quite identical in pitch are sounded together. [Acoustics, § 17.]

Bebung (*Ger.*) The tremolo stop in an organ. A repeated note in pianoforte music.

Bebisation. A series of syllables recommended by Daniel Hitzler, a Fleming, in 1630, as a means of teaching the notes. He proposed to substitute the syllables la, be, ce, de, mi, fe, gi, for ut, re, mi, fa, sol, la, si, already in use.

Bécarre (*Fr.*) The sign ♮.

Bec (*Fr.*) } A mouth-piece, *lit.*, a beak.
Becco (*It.*) } [Flûte-à-bec.]

Becco polacco (*It.*) A large bag-pipe.

Becken (*Ger.*) A cymbal.

Bedon (*Fr.*) An obsolete term for a drum, or tambour.

Begeisterung (*Ger.*) Spirit, excitement, enthusiasm.

Begleiten (*Ger.*) To accompany.

Begleitende Stimmen (*Ger.*) Accompanying voices or parts.

Begleitete Fuge (*Ger.*) A Fugue with free parts. [Free parts.]

Begleitung (*Ger.*) Accompaniment.

Beisser (*Ger.*) [Mordente.]

Belebt, Belebend (*Ger.*) *Animato, animando*

Bell. The lower termination of any tubular musical instrument which by the outward turning of the rim assumes the form of a bell. Fr. *Pavillon*, Ger. *Schallstück.*

Bell diapason. An organ stop consisting of open metal pipes with bell mouths. Its tone is more reedy and powerful than that of an ordinary open diapason. Generally of 8ft. length.

Bellezza (*It.*) Beauty of expression and tone in playing and singing.

Bell Gamba. An organ-stop, the pipes of which are conical and surmounted by a bell. It was introduced by Mr. Hill, organ-builder, of London. Its tone is remarkably sweet, not unlike that of a stringed instrument, though somewhat more reedy. The pipes speak rapidly.

Bellicosamente (*It.*) } Warlike, martial.
Bellicoso (*It.*) }

Bell metronome. A metronome in which the recurrence of a set number of beats is marked by the sound of a bell. [Metronome.]

Bell Open Diapason. [Bell Diapason.]

Bellows. In the harmonium, organ, concertina, &c., that contrivance by means of which wind is supplied to the pipes, tongues, or reeds. [Organ.]

Bell Piano. [Glockenspiel.]

Bells. 1. Musical instruments of percussion, consisting of a series of metal basins or cups, the outline of which has from time to time been modified. The materials of which bells are usually made are copper and tin, the proportions varying in several countries and even among the manufacturers.

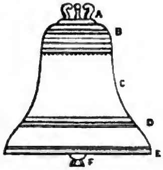

The various parts of the bells are A, the Canons; B, the Shoulder; C, the Waist; the thick part between D and E, the Sound Bow; E, the Rim or lip; F, the Clapper.

The following analyses of English and some foreign bells, will give a correct idea of the composition of the ancient bells.

ENGLISH BELLS.	
Copper	80·0
Tin	10·1
Zinc	5·6
Lead	4·3

ROUEN BELLS.	
Copper	72·0
Tin	25·0
Zinc	1·8
Lead	1.2

PARIS BELLS.	
Copper	72·9
Tin	25·56
Iron	1·54

SWISS HOUR BELLS.	
Copper	75·0
Tin	25·0

Mr. Denison recommends on theoretical grounds the following proportion

Copper	76·5
Tin	23·5

2. The use of bells to call worshippers together is supposed to be of Christian origin, but it is said that the feast of Osiris in Egypt was announced by the ringing of bells. Aaron and the Jewish high priests had bells attached to their vestments, and

Plutarch says that small bells were used in the mysteries of Bacchus, and the priests of Cybele at Athens employed bells in their rites. The Greeks sounded bells in their camps, and the Romans indicated the hours of bathing and business by the tintinnabulum. It is also said, that in some places large gongs were suspended in the air, and as the wind brought them together, so was the character of the sounds made, interpreted as an unfavourable or favourable augury. Trumpets were employed among the Jews to call the faithful to worship (Exodus xx., 13; Numbers x., 2; Joel ii., 15). Plates of iron are still used in the Levant, and a plank of wood is occasionally employed for the same purpose that we use bells in some of the old Wallachian monasteries. In the East the call to prayer is made by the Muĕddin of each mosque, who, having ascended the gallery of the mad'neh or minaret, chants the "hadan" or call to prayer, apparently in opposition to the Christian use of bells. [Hadan.]

The introduction of bells into churches is attributed to Paulinus, Bishop of Nola, in Campania, about the year 400, but there is an epistle of that bishop still extant in which he describes his church, but makes no mention of either tower or bells; indeed, it is believed that towers were not constructed until two centuries later. Yet it is not a little remarkable that the general name for bells was Nolæ or Campanæ, and hence the words *knoll* as meaning the sound of a single bell, and *campanile* a bell tower. Sabianus, who was Pope in 604, ordered the bells to ring the *horæ canonicæ* at the proper times during the day, and Benedict, Abbot of Wearmouth, brought his bells from Italy about the year 680. Bells were hung in towers in the East in the 9th, and in Germany in the 11th century. Those that were in use before are supposed to be hand bells; several examples, as old as the 6th century, are still preserved in some parts of Europe and the United Kingdom. St. Patrick's bell, St. Ninian's bell, St. Gall's bell, and others are plates of iron rivetted together. St. Gall's bell (about 646) is still shown in the monastery of the city called by his name in Switzerland. In the 13th century larger bells were cast, but it was not until the end of the 15th century that they began to assume great proportions. St. Dunstan, in the 10th century, seems to have the credit of having established the first foundry in England, Glastonbury, Malmesbury, and other places having been furnished with bells by him. Bells were rung not only to indicate the commencement of certain services, but also were tolled to mark certain stages in those services. Thus we find mention made of the Saints or Sanctus bell, the Compline bell, the Judas bell, the Pardon or *Ave* bell, the Passing bell, the last tolled to warn all "Christen soules" to pray for the parting soul of the dying. Bells, being thus intimately connected with the services of the Church, have been supposed to possess a certain sacred character. They were founded with religious ceremonies, consecrated, baptised, and were anointed with holy oil (see Schiller's "Lay"). St. Colomba, in the 6th century, made use of a bell whose name was "Dia Dioghaltus," or "God's vengeance," to test the truth of assertions made, as it was believed that the wrath of God would speedily overtake any who swore falsely by it. Pious inscriptions are frequent on bells of the middle ages, and inscriptions, not always pious, are found on those of later date. Bells were often rung to allay storms, there being a special endowment belonging to Old St. Paul's, "for ringinge the hallowed belle in great tempestes and lightninges." The curfew bell, still sounded in many parts of England and Scotland, is of more ancient practice than the period usually assigned as its commencement, the reign of William the Norman; and there are many social practices announced by the ringing or tolling of the church bells.

3. Change ringing, or campanology, is frequently practised when there are more than three bells, such changes being known by the names of bob-majors, bob triples, Norwich court bobs, grandsire bob-triples, and caters. The number of changes a set of bells is capable of, may be known by in-multiplying the numbers of the set. Thus, three bells may ring six changes, 123, 132, 213, 231, 321, 312; four bells will give 24 changes; 5 bells, 120 changes; 6 bells, 720 changes; 7 bells, 5,040 changes; 8 bells, 40,320 changes; 9 bells, 362,880 changes; 10 bells, 3,628,800 changes; 11 bells, 39,916,800 changes; 12 bells, 479,001,600 changes. To ring the changes that 12 bells are capable of, would take 91 years at two strokes per second, while a peal of 24 bells can make so many changes that it would occupy 117,000 billions of years to ring them all.

The technical terms for the various peals, on sets of bells of different numbers, are the following:

Rounds	On	three	bells.
Changes or singles	,,	four	,,
Doubles or grandsires	,,	five	,,
(Bobs) Minor	,,	six	,,
Triples	,,	seven	,,
(Bobs) Major	,,	eight	,,
Caters	,,	nine	,,
(Bobs) Royal	,,	ten	,,
Cinques	,,	eleven	,,
(Bobs) Maximus	,,	twelve	,,

4. A bell is said to be "set" when she is mouth upwards, at "hand stroke" when the "sallie" or tuft on the rope has to be pulled.

at "back stroke" when the ringer has to pull the end of the rope. A bell is said to be "going up" when she moves her position in the change from "treble" towards that of "tenor," and "down" when she is changing her position from that of "tenor" towards that of "treble." A bell is said to be "behind" when she is the last of the changing bells, and at "lead" when she is the first. Thus the progress from "lead" to behind is said to be "going up," and from behind to lead is called "going down." "Dodging" is moving a place backwards out of the ordinary hunting course. A bell is said to be "hunting up" when she is pulled after the one which previously pulled after her. A bell is said to "make a place" when she strikes two blows in succession at any one place. To "lie a whole pull" is synonymous with "making a place." Two blows at "lead" and "behind" are a part of "hunting," in making these therefore a bell is not said to be "making a place." "Bob" and "singles" are words used to produce a certain series of changes by disturbing the ordinary system of "hunting." The full knowledge of the meaning of these and many other technical terms used in ringing can only be learnt in the belfry. The method of Doubles named after Stedman (1640) is, in principle, as follows: while three of the bells are ringing changes, the other two are dodging behind, but at the completion of each set of six changes one bell comes down from behind to take part in the changes, one, of course, at the same time going up behind to take part in the dodging.

5. Bells are occasionally employed as orchestral instruments—small bells, tuned to a certain scale, being most favoured—as in Victor Masse's "Les noces de Jeannette," a whole peal of small bells being used with great effect. These, as in Mozart's "Magic Flute," are so arranged as to be played with keys, like a pianoforte. [Glockenspiel.]

Auber employs a single bell in the finale to "Fra Diavolo." Rossini has introduced a bell sounding

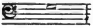

in the opening of the second act of "William Tell." Donizetti also, in the finale to "Lucia di Lammermoor," has written for a bell tuned to the same note. Meyerbeer, in his "Huguenots," employs a bell in

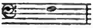

with clarinets and bassoons. In "Dinorah," in what is popularly known as the "Goat Trio," a bell with the note

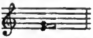

is used. Ambroise Thomas has a series of clever harmonies for the orchestra in his opera "Hamlet," while a deep-toned bell strikes the midnight hour. Flotow, in "Martha," uses a bell, as does Gounod in "Jeanne d'Arc," tuned to the following note:

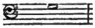

and there are numerous other instances where bells of all grades of tone have been used with skill and effect.

Bell founding. The shape and proportions of the intended bell having been decided upon according to a certain scale, the first part of the process of casting is commenced, by constructing an inner mould called the *core*, by which the form of the inside of the bell is determined. This *core* has a foundation of rough brickwork or iron, hollow in the centre, afterwards plastered over with loam or soft clay. A guage of wood, called a *crook*, made to revolve or sweep round on a central pivot by the hand of a workman, gives the clay the exact form required. This process will be at once understood on reference to the following diagram. A is the *core*, B the *crook*, which is fastened to C, the *pivot* on which it revolves:

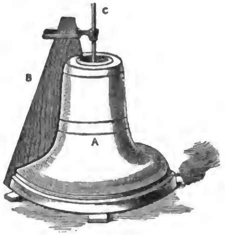

The *core* is hardened by a fire made in its hollow, and when it is sufficiently "set," it is covered with grease and tan, over which is placed a coating of haybands and loam, of the thickness of the intended bell, and upon this the *cope* or outer mould is shaped. When this is dried it is removed, the thickening of haybands and loam which represented the shape of the bell to be cast, is destroyed, and the two moulds, the *core* and the *cope*, are examined and finished.

The *core* is sometimes made on an iron foundation, instead of brickwork, in which case it can be dried in a furnace, instead of by the fire in its hollow. The *cope* having been carefully adjusted over the *core*, the head and the staple to hold the clapper are then fitted on,

Digitized by Google

and the whole mould is firmly imbedded in the earth, leaving only the holes at the top visible.

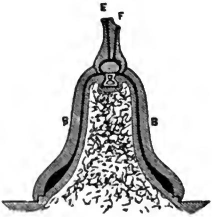

The above diagram shows the position of mould ready for the metal. A is the *core*, B the *cope*, F the channel for the metal to run in, E the hole for the air and gases to escape during the casting, and the thick black line the section of the bell. When the metal is quite ready, the furnace-door is opened, and the molten mass rushes down a channel, previously prepared, into the moulds sunk in the pits, and excepting mishaps, from insecure "bedding," the splitting of the *cope*, or other accidents, the bell is cast, and, when cold, is dug from the pit, the clay mould destroyed, and the bell is ready for the next process, that of tuning. The tuning is effected by means of a lathe and some simple machinery. If the bell requires sharpening, the diameter is lessened in proportion to its substance, if it is too sharp, the sound-bow is thinned by the same means; but, as a rule, bells are now so accurately cast, that little if any tuning is necessary after the bell leaves the mould. It is stated in "Knight's Encyclopædia, 1854," that the German bell-founders made the various dimensions of the bell to bear certain ratios to each other. The thickest part where the hammer strikes is called the "Sound Bow." If this thickest be called one, then the diameter of the mouth equals 15, the diameter of the top or shoulder 7½, the height equals 12, and the weight of the clapper $\frac{1}{40}$ of the weight of the bell.

Denison recommends that the sound bow of the three or four larger bells of a peal should be of the thickness of a thirteenth of the diameter, and that the smaller bells may gradually increase in thickness up to the twelfth in a peal of six, the eleventh in a peal of eight, and to the tenth in a peal of ten or twelve, greater thickness impeding the freedom of the sound.

The bells of the Cathedral at Exeter, one of the largest peal of bells in England, the greater number of which were cast in 1676, have the following weights, diameters, and tones;—

WEIGHT.			DIAMETER.		TONE.
Cwt.	qr.	lb.	Ft.	in.	
67	1	20	5	11½	B flat.
46	3	14	5	4½	C.
38	1	16	4	11	D.
30	1	12	4	7	E flat.
21	0	0	4	1	F.
15	0	0	3	10	G.
12	2	0	3	4½	A.
10	1	2	3	1½	B flat.
9	3	20	3	0	C.
8	3	20	2	9½	D.

The relative diameters of a peal of eight tuneable bells should be according to the following proportion: 60, 53½, 48, 45, 40, 36, 32, 30. The relative weights being generally in the proportion, 100, 70·23, 51·2, 42·2, 29·63, 21·6, 15·18, 12·5.

Belly. The upper plate of the resonance-box. In instruments of the Violin and Guitar family the strings are stretched over the belly, and the bridge across which they pass is so placed as to set the belly, and by its means, the air contained in the resonance-box, into vibration. In instruments of the Pianoforte class, the belly is that thin plate of fir-wood which, placed behind the strings, acts as a sound-board. Instead of an upper plate of wood, the guitars of many of the less civilized nations have a stretched parchment. The belly thus formed answers all the purposes of resonance for which it is intended; the Kissar of Nubia, the Banjo of the American negroes, the Nanga or Negro Harp which shares the combined designs of a guitar and harp, may be quoted as examples.

Bemes or Beemes. Saxon Trumpets or Bugles.

> "Of brass they broughten beemes and box,
> Of horn and bone, in which they blew and pouped,
> And therewithal they shriked and they houped;
> It seemed as that the heven shulde falle."
>
> *Chaucer.* "Nonne preestes tale." See also the "Romaunt of the Rose."

Bémol (*Fr.*), **Bemolle** (*It.*) The note B♭.

Ben (*It.*) Well. *Ben marcato*, well and clearly marked. *Ben sostenuto*, or *ben tenuto*, well-sustained.

Benedictus (*Lat.*) A movement of the Mass, immediately following the Sanctus. [Mass.]

Bene placito (*It.*) At pleasure, *ad libitum*, *e.g.*, "Bassani's Ballate corrente, Gighe, e Sarabande, a violino, e violone, overo spinetta, con il secondo violino *à bene placito*" (1684).

B quadro (*It.*) The square B or ♮, that is, B durum or *natural*, as opposed to the ♭, soft B, B molle, or rounded b which, in its slightly altered outline, is now known as a flat. That the note B was the first note altered by an accidental, accounts foi the fact that signs of ♭ and ♮ are of general application.

Bequadro (*It.*)
Béquarre (*Fr.*) } The sign ♮.

Berceuse (*Fr.*) A cradle song.

Bergomask. Burgomask. Bergamasca. A lively dance in triple time, for two clumsy performers, in imitation of the dances of the country people of Bergamo, who were considered the least graceful of the Italians.

> "Will it please you to see our epilogue, or to hear a Bergomask dance, between two of our company?"
> *Shakespeare.*

Bes (*Ger.*) The note B double flat.

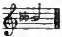

Besaiten (*Ger.*) To string an instrument.

Bestimmt (*Ger.*) With decision.

Bewegung (*Ger.*) Motion.

Bhat. A Hindu Bard.

Bianca (*It.*) A minim, 𝅗𝅥. The *white* note, as opposed to the ♩, or black note (*nero*).

Bichord. Having two strings to each note.

Bicinium (*Lat.*) A duet, from *bis* and *cano*. "Cum duo canunt, *bicinium* appellatur; cum multi, *chorus*."

Bifara (*Lat.*) An organ-stop, with two pipes to each note, producing a tremulant effect. [Vox Angelica.]

Bilancojel. An Indian flute with seven holes, played by a mouthpiece.

Bimmolle (*It.*) The note B♭. [See Quadro.]

Bina, or Vina. An Indian guitar, with a long finger board, and a gourd attached to each end. Seven strings or wires wound round pegs in the usual way are attached to the finger board, four on the surface, and three at the sides. There are about twenty frets, some standing up as high as an inch from the finger board; these are fastened with wax, and the performer regulates the positions of them at his pleasure. In the performance one gourd is rested on the left shoulder, and the other on the right hip. Its scale consists of a series of small intervals lying between a note and its octave, in the bass-stave.

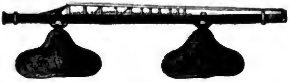

Binary Form. The form of a movement which is founded on two principal themes or subjects. [Sonata Form.]

Binary Measure. Common time. [Tonic-Sol-fa.]

Bind. (1) A curved line, ⁀, a sign which, when placed over two notes of the same name or same pitch (enharmonically changed), directs that the two are to be sustained as one. It is of frequent occurrence at points of Syncopation and Suspension.

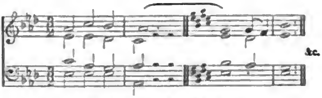

Enharmonic change, or Modulation.

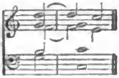

Suspension.

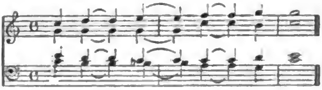

Syncopation.

When a curved line is placed over two notes, not of the same name or pitch, it is called a *Slur*, and directs that they are to be played smoothly, *e.g.*:

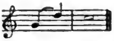

It is to be regretted that the horizontal line introduced by Sir W. S. Bennett as a *Bind*, so that no confusion can exist between the *Bind* and *Slur*, has not been generally adopted, *e.g.*:

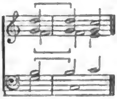

(2) A Brace (*Fr.*, *Accolade*) which binds together the separate parts of a score.

Binde (*Ger.*) [Bind.]

Bindebogen (*Ger.*) The bind-bow. [Bind.]

Bindung (*Ger.*) Syncopation, suspension, so called because the notes forming it are *bound*, or at least might be so written.

Bindungszeichen (*Ger.*) A slur or bind.

Birne (*Ger.*) That portion of a clarionet or similar instrument in which the mouth-piece is inserted, so called from its *pear*-like shape.

Bis (*Lat.*) Twice. (1) A direction that the passage over which it is placed shall be twice played or sung. Its use is generally limited to short passages, marks of *repeat*

Digitized by Google

being written for a long repetition. It is placed under or over a slur, *e.g.*:

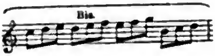

(2) Again. *Encore.*

Bischero (*It.*) The peg, or pin, with which the strings of an instrument are secured.

Biscroma (*It.*) } A semiquaver, 𝅘𝅥𝅯.
Biscrome (*Fr.*) }

Bisdiapason. The interval of a double octave, or fifteenth.

Bissex. A kind of guitar with twelve strings (*Ger.* Zwölfsaiter), invented by Vanhecke in 1770. Of the twelve strings six were over the finger-board, six below, hence the name *twice-six*. Its compass was three and a half octaves.

Bis unca (*Lat.*) A semiquaver 𝅘𝅥𝅯, or note with *two hooks*.

Bit. A small piece of tube, generally furnished with two raised *ears*. It is used for supplementing the crook of a trumpet, cornet-a-piston, &c., so as to adapt the instrument to a slight difference of pitch.

Bizzarramente	(*It.*)	Fantastically, wonderfully.
Bizzarria	(*It.*)	Singularity, affectation.
Bizzarro	(*It.*)	Odd, droll.

Blanche (*Fr.*) A minim 𝅗𝅥 [Bianca.]

Blanche pointée (*Fr.*) A dotted minim.

Blasbalg (*Ger.*) The bellows of an organ. Saxon, *bles belg*, a wind-bag.

Blasinstrument (*Ger.*) Wind instrument. Flute, oboe, bassoon, cornet, trumpet, trombone, &c.

Blasmusik (*Ger.*) Music for wind instruments.

Blatt (*Ger.*) A vibrating tongue or blade. [Reed.]

Blechinstrumente (*Ger.*) [Metal wind instruments.]

B moll (*Ger.*) The key B♭ minor.

B molle (*Lat.*) The note B♭, *cf.* B quadratum.

Bobibation or **Bocedisation.** Solfeggi taught by Huberto Walraent at the end of the 16th century for scale practice, which were bo, ce, di, ga, la, mi, ni. [See Bebisation.] Bobibation or Bocedisation, in which the syllables bo, cé, di, ga, lo, ma, ni were substituted for those attributed to Guido, was introduced and taught in many schools in Flanders, and so this peculiar use came to be called Belgian solmisation. Walraent's method was adopted in Italy in 1599 by Henri de la Putte, who wrote an elaborate Latin treatise in defence of it; and a few years later Calwitz, ignoring its invention and taking the credit to himself, introduced it into Germany. In Spain and France the method was proposed by Pietro de Ureno and John Lemaire, but without success. To the last-named musician the addition of the syllable Si for the leading note is attributed. Bobibation was accepted by some musicians and rejected by others, and the result was a petty war, which lasted until the commencement of the 18th century. Hitzler, a few years later, suggested the use of the syllables la, be, ce, de, mi, fe, gi—this system he called *bebisation* or *labecedation;* and Graun recommended da, me, ni, po, tu, la, be, from which his plan was called *damenisation*.

Bob major. Bob maximus, Bobs. [Bells.]

Bocal (*Fr.*) The mouth-piece of the horn, serpent, trombone, &c.

Bocca (*It.*) The mouth. *Con bocca chiusa*, with closed mouth, humming.

Bocca ridente (*It.*) *Smiling mouth*, the position of the mouth needful for the production of pleasing tone.

Bocchino (*It.*) A mouth-piece of wind instruments.

Bockpfeife (*Ger.*) Bagpipe.

Bockstriller (*Ger.*) A goat-like tremolo upon one note, a bad shake. [Vibrato.]

Boden (*Ger.*) [Body.]

Body. The resonance box of a string instrument. That part of a wind instrument which remains after the removal of mouth-piece, crooks, and bell.

Bogen (*Ger.*) Bow.

Bogenclavier (*Ger.*) [Tetrachordon.]

Bogenfuhrung (*Ger.*) The art of playing with a bow upon stringed instruments.

Bogenstrich (*Ger.*) The stroke of a bow.

Bolero (*Sp.*) A Spanish dance in triple measure with strongly-marked accent, also called Cachucha. It is accompanied with singing and castanets, and the performer assumes in the course of the dance all the various feelings supposed to be excited by love, from the greatest shyness to the highest ecstasy. [Chica.]

Bombard. Bombarde (*Fr.*) A reed stop on the organ, usually among the pedal registers, of large scale, rich tone, and often on a heavy pressure of wind.

Bombardino (*It.*) A small bombardo, *q.v.*

Bombardo (*It.*) A mediæval wind instrument, the precursor of the oboe, of which it was no doubt a large and coarse species. The word *Pommer*, applied to these instruments, was a corruption of the Italian name. The *chalumeau* or *shawm* (Ger. *Schalmey*) was the smallest of this class, the *bombardone* the largest. Clarinets, oboes, and bassoons, now so clearly defined, grew out of one common parentage. The clarinet has but one vibrating

Digitized by Google

reed; the oboe and bassoon double vibrating reeds.

Bombardon. A brass instrument, in tone not unlike an ophicleide. It is not capable of rapid execution. The compass is

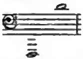

Music for it is written without transposition, although it is in F.

Bombaulius, *βομβαύλιος* (*Gk.*) A facetious name for a bag-piper. A pun on the words ἀυλητῆς, a flute-player, and *βομβύλιος*, a buzzing insect, whence our word, bumble-bee; *c.f.*, Latin, *Bombus*.

Bombyx (*Gk.*) *βόμβυξ*. A Greek flute, perhaps so named from its supposed resemblance to the silk-worm. It was probably a reed-instrument of powerful tone. The following illustration is given by Burney from a sarcophagus in the Capitoline Museum, at Rome.

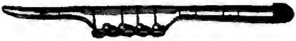

Bones. Four pieces of the ribs of horses or oxen, held in the hands and struck together for the purpose of marking time, in accompaniment to the voice or an instrument. The bones are of ancient use in England, and are alluded to by Shakespeare in the fourth act of "A Midsummer Night's Dream," as forming one means of rustic music. And in figures designed by Inigo Jones for the Court Masques one is represented playing upon knicky-knackers of bone or wood. The word knicky-knackers, by which the bones are known to the country people, may have its origin from the word "nakeres." In Strutt's "Sports and Pastimes," a payment is recorded as being made to Janino le Nakerer, among the minstrels of King Edward II. The nakerer was probably the drummer; but, as the minstrels frequently indulged in burlesque music, Janino may have been the performer on that primitive or rustic instrument, the bones.

Bon temps de la mesure (*Fr.*) The accented portion of the bar.

Bordone (*It.*) [Bourdon.] [Viola di Bordone.]

Boulou. A harp used by the negroes of Senegambia and Guinea, in shape like the Oriental harp. Its strings are of fibre.

Bourdon (*Fr.*) (1) A drone bass, a burden such as that produced by a bagpipe, or a hurdy-gurdy. [Burden.] (2) An organ stop, consisting of stopped wooden pipes, generally of 16 ft. tone. Sometimes, but rarely, the upper part is of metal. It is found on manuals as a "double" stop, and also on the pedal organ as a soft foundation-stop. It was formerly made of a large scale in England, but from a better knowledge of scientific principles organ builders are now able to produce a strong and pure body of tone from a pipe of moderate scale. As a rule, it is important that it should be free from a preponderance of harmonics or over-tones, but sometimes they are purposely produced with the fundamental note, in which case the stop is called *quintaton*, because the first harmonic or over-tone of a stopped pipe is its twelfth, or octave fifth. Hence a Bourdon was sometimes said to be *fifthy*.

Bourrée (*Fr.*) A dance tune in common time, said by Hawkins to come from Auvergne. Other writers give Biscay as its birthplace. The earliest mention of it is probably about 1580. It is still popular with the peasants of Lower Brittany. It often forms one of the movements of the earlier Sonata.

Boutade (*Fr.*) A dance which was supposed to be impromptu.

Bow. An instrument of wood and horse-hair, employed to set the strings of the violin, &c., in vibration. The bow, originally curved, as its name implies, has been subject to many changes of shape from time to time, from a large curve to an almost flat form.

Fig. 1.

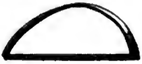

Fig. 2.

Fig. 3.

Fig. 4.

Fig. 5.

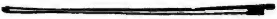

The bow shown in fig. 1 is that still used with the *rebab* of Algeria. Fig. 2 is given by Gerber from a MS. of the 8th or 9th century. The bow now used for the violin is about 29 inches in length (almost straight, but with a slight curve *inwards*, not outwards, as in the older bows), the violoncello bow being a little shorter, fig. 5. Christopher Simpson (1676) says that twenty-seven inches was the length of the bow in his time, the "sonata bow," fig. 3, according to tradition, being only twenty-four inches, the common "fiddle-stick" being still shorter. The form of the bow, which was anciently employed for all stringed instruments of the violin kind, known now as the "Corelli bow," fig. 4, is to a certain extent preserved in the double-bass bow as at present employed in England.

Most of our stringed-instruments can be traced to an Eastern source, but as the earliest figure of the bow is found in MSS. relating to this country, it has been supposed that it is of British origin. There are many representations of it in MSS. as early as the Anglo-Saxon period (see Sandy's "History of the Violin," pp. 52, 53, &c.), and later through successive generations, besides existing specimens of actual ancient bows, all of which partake of the *bowed* character, as seen in fig. 1. The curved bow was still in fashion until the death of Handel, if any trust is to be placed in contemporary pictures and drawings.

The little that is known of the early bows gives the notion that they were incapable of producing anything like delicacy of tone; and it was not until the early part of the 17th century, when violin-playing began to be cultivated, that we find that any attempt was made to improve either the form of the bow or the means of stretching the horse-hair, so that an even pressure might be obtained. The alteration from the bowed form is said to have been made after a suggestion by Tartini [1692-1770]. There is every reason to believe that the improvement in violin-playing due to him could only be effected by means of a better bow than the clumsy one of his time. The earliest improvement was made when a metal band, with teeth-like edges, was introduced, with the design of regulating the position and tension of the hair at or near the handle. This helped to prevent the hair assuming the cord-like form which players found to be awkward and clumsy. But it was reserved for François Tourte [1747-1835] to devise the plan of keeping "the hair flat by means of a clasp." Tourte also introduced the screw and button for slackening or tightening the hair at pleasure, and was the first to choose Brazil-wood as a material in bow-making. It was the father of Tourte who attempted the first improvements in bow-making, but it was the son above-mentioned who introduced the most valuable inventions. The bows of the younger Tourte are almost as much sought as those of the elder Dodd [1705-1810], who lived to the great age of 105 years, but the works of the last-named are most highly valued. Panormo, Tubbs, Eury, and Adam are counted among the chief of modern bow-makers.

Bow-hand. The term is employed by violinists to describe the power and skill with which a player produces the tone of his instrument.

Bowing. The art of managing the bow, so as not only to bring out the best tone the instrument is capable of, but also so to phrase the passages played that the best possible character may be imparted to the music. The importance both to a violinist, and a composer of music for the violin, of a thorough knowledge of the art of bowing, cannot be overrated. By varying the system of bowing, a simple musical sentence may be changed in its character, almost indefinitely. Formerly, very little attention was paid to this subject, the system of bowing being left very much to the discretion of the players, who only occasionally had such general directions as *legato* or *staccato*. There always, however, existed certain traditional rules, *e.g.* that the down-bow should be used at the first beat of a bar, or where any great emphasis was required (as in some cases of syncopation); also, that where no directions are given, the passage should be *bowed*, that is, the notes should be alternately played by an *up* and *down* bow. But it is evident that in simple music, of triple measure, these rules will clash, for, alternate bowing will lead to the recurrence of an up-bow on every alternate down-beat. Hence, even if an excellent band is playing music up to the date of, and including that by Handel, it must have often been observed that the bowing is far from being uniform. In modern music, every direction is given to the performers which is requisite for the production of absolute uniformity, and more than this, the various effects which are capable of production by the different systems of bowing are used as part of a composer's material. The prominent features in modern bowing are the more frequent antithesis between *legato* and *staccato*, and the use of at least three kinds of the latter. When notes have the ordinary dot placed above them they are *bowed* staccato; when the dots are under a slur, they are played with *one bow* (that is by the movement of the bow in one direction) the short length of the sounds being brought about by keeping the bow always *lying on the string*, so that any movement of the bow which has produced a sound shall be followed by absolute silence. The third kind of staccato is produced by holding the bow as lightly as possible and allowing it almost to dance upon the string. In this manner rapid passages may be played either by *one* bow up or down, or by an alternate bowing, during which the movement of the bow at right angles to the string is so slight that it seems to rise and fall almost perpendicularly. A favourite division of four rapid notes is to make two

Digitized by Google

legato and the following two staccato. The well-known Var. II. of Beethoven's Sonata for Violin, known as the "Kreutzer," is a good illustration of this:

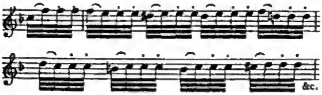

The effect which results from moving the bow on an unaccented part of the bar is most striking and beautiful, *e.g.* (from the same Sonata):

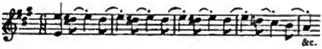

and in the following (Beethoven Symphony, No. 9):

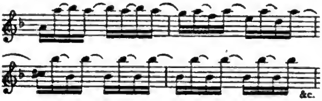

Groups of three notes are often divided into two legato and one staccato, *e.g.*:

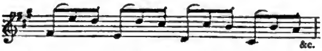

less often into one staccato and two legato, *e.g.*:

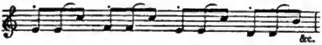

Two notes out of eight are often made legato, *e.g.*:

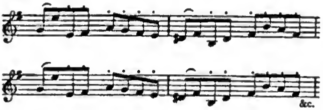

The above few examples may serve to give some idea of the inexhaustible resources of the art of bowing. It will of course be understood that what is here said of the violin applies equally to the viola and violoncello. But, in consequence of the thickness of the strings, the double-bass is not so capable of rapid contrasts of bowing as the rest of its family. In studies and exercises it is usual now to direct a down-bow by the sign ⊓; and an up-bow by ʌ. The French terms corresponding to these signs are *tiré* (draw); *poussé* (push), sometimes abbreviated by *p* and *t*. The quality of tone produced depends not only on the nature and quantity of pressure exercised by the bow upon the strings, but also upon the position of the point of impact. Thus, if played very close to the bridge (*sul' ponticello*), the tone is of peculiar brightness and shrillness; as the bow is used further from the bridge, the tone passes through a stage of great purity and strength, until, at close proximity to the finger-board, it becomes soft and somewhat dull. The practised performer chooses that part of the string capable of producing the tone best suited to the passage he is playing, and he draws it forth with that part of bow most suitable for the purpose. As a general rule, from the heel to about the middle of the bow, is the part naturally used for *forte* or *sforzando* passages; and from the middle to the point for those of a more delicate character. But actual experience is the only method of learning the intricacies and beauties of the art of bowing.

Boyau (*Fr.*) Cat-gut strings.

B quadratum (*Lat.*)
B quadrum (*Lat.*)
B quarre (*Fr.*)
B natural. The name of B quadratum or B square was given by reason of its shape, which was originally that of a gothic B. [B cancellatum.]

Brabançonne. The Belgian patriotic song first sung at the time of the revolution of 1830. The words were written by an actor named Jenneval, at that time engaged at the theatre at Brussels, the music being set by a singer named Campenhout. Upon the death of Jenneval at Berghem his mother was allowed a pension of 2400 fr. Campenhout was appointed director of the Royal Chapel.

Brace. (1) A mark connecting two or more staves together.

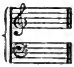

(2) The leather slides upon the cords of a drum, used for raising or lowering the tone by tightening or loosening the head.

Branle, Bransle (*Fr.*); **Brawl** (*Eng.*) This is an old round dance in which the performers joined hands in a circle. A country dance.

> "Then first of all he doth demonstrate plain
> The motions seven that are in nature found,
> Upward and downward, forth, and back again,
> To this side, and to that, and turning round;
> Whereof a thousand brawls he doth compound,
> Which he doth teach unto the multitude,
> And ever with a turn they must conclude."
> Sir John Davies. *Orchestra*, 1607.

Digitized by Google

The brawl and the bransle were the same dance. Douce gives an account of "le branle du bouquet,' from "Deux dialogues du nouveau langage François, Italianize," Anvers, 1579, in which, kissing the whole of the ladies, by each of the gentlemen in turn, seems to have been one of the chief features in the dance.

The following Branle is from Arbeau's "Orchesographie," 1589:

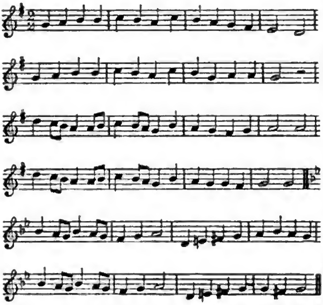

Bratsche (*Ger.*) The Tenor Violin, Alto, or Viola. So called from the *Viola da braccio*, or viola held on the arm, as distinguished from the *viola da gamba*, or viola held between the legs, the precursor of the *violoncello*.

Brava, *fem.* (*It.*)
Bravi, *pl.* (*It.*) Well, or bravely, done.
Bravo, *mas.* (*It.*)

Bravura (*It.*) Dash, brilliancy. *Con bravura*, with dash or brilliancy. *Aria di bravura*, an air, distinguished from a simple melody by the introduction of florid passages. [Aria.]

Break. (1) The point of junction in the quality of tenor, soprano, and alto voices. A genuine bass voice has no break. The lower range is called *voce di petto*, or chest voice; the upper, *voce di testa*, or head voice; and the place of junction is called the break. A properly-cultivated voice should have the break so under control, that the union of the two qualities should be imperceptible. [Larynx.]

(2) In the clarinet the break in the tone of the instrument occurs between B flat and B natural,

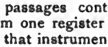

Hence, rapid passages containing frequent transitions from one register to the other are impossible on that instrument. In trumpets and horns, when from imperfect lipping, the note produced is other than the tone intended, such note is called a break. A similar result often occurs in imperfectly formed or unset voices.

(3) Break, in an organ stop, is the sudden alteration of the proper scale-series of the pipes by returning to those of an octave lower in pitch. A break becomes necessary in the smaller compound stops, for, when proceeding far upwards in pitch, it is found that the pipes would be so small as to be inaudible. As mixtures, sesquialteras, and other stops of the same class, are generally formed of several ranks of high harmonics, breaks in them are frequent. The break generally takes place between C and C♯, or F and F♯; but organ-builders do not act with uniformity, either as to the position of the break, or the exact series of sounds to be produced on the return.

Breast. (*Old Eng.*) The voice.

"Trulye two degrees of men, which have the highest offices under the Kinge in all this realme, shall greatly lacke the use of singinge, preachers, and lawyers, because they shall not, withoute this, be able to rule their breastes for everye purpose."—*Ascham's Toxophilite.*

"By my troth, the fool has an excellent breast."—SHAKESPEARE, *Twelfth Night*, ii., 3.

"The better brest, the lesser rest,
To serve the queer, now there, now heer,
For time so spent, I may repent
And sorrowe make."

TUSSER (*Five hundred points of good husbandry*, 1540).

Breit (*Ger.*) Broadly, *largamente*.

Breve. In Mediæval music, the note equal to half the value of a *long*, and one quarter of the double *long* or maxima. Its shape was ■. "Quandocunque punctus quadratus invenitur, qui caret omni tractu *brevis* dicitur" (Franco, of Paris.) Breves like other notes in the early attempts at mensurable music, affected the length of other notes, and were in turn affected by other notes according to their relative position. A *long* "in modo perfecto" was reduced by one third or made imperfect by having a *breve* next to it on either side. A breve "in tempore perfecto" was made imperfect, that is, was reduced from three to two beats, by juxtaposition with a semibreve. A breve, when imperfect, was also subject to "alteratio," which doubled its value (= four beats). When so altered it was called *alterata* or shortly *altera*. All these laws and many others of a like character were drawn up by writers in the fifteenth century, in the dawn of "mensurable" music. Having discovered the utility of showing the relative length of notes by their shape, authors seem to have revelled in constructing new complications. These were gradually dropped by succeeding writers, until the breve became the unit of duration, a position which it held for nearly two centuries. The Semibreve is now our recognized unit, the Breve being a *double* note and of rare occurrence. But these changes have been a slow growth, not sudden alterations of existing opinions or practices.

Bretgeige (*Ger.*) A pocket fiddle; hence, Fr. *pochette*, Ger. *Taschengeige*, It. *sordino*, from the small quantity of tone it is capable of producing. Eng. *kit*.

Bridge. A piece of wood which, on instruments having a resonance-box, performs the double duty of raising the strings above the belly, and of terminating at one end their vibrating portion. In instruments played with the bow, the bridge is arched, in order to allow the bow to impinge upon any one string without touching others. In instruments, such as the guitar and pianoforte, its upper edge runs parallel to the belly. In violins, the material and adjustment of the bridge are of great importance. Some instruments require a bridge made of coarse-grained wood, others of close-grained. It stands on two legs; that on the right hand should rest on the belly at a short distance behind the sound post. The legs should lie flat on the surface of the belly, in order that the vibrations of the strings should be duly transferred to the resonance-box. The tone of an instrument is largely influenced by the position of the bridge, and only great experience and nice handling can discover where it is best set up.

Brillante (*It.* and *Fr.*) Brilliant, in a showy sparkling style.

Brillenbässe (*Ger.*) "Spectacle basses," music for the drum, so called from its resemblance to a pair of spectacles. 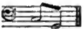

Brindisi (*It.*) A drinking song. Often of a florid character, so arranged as to exhibit the change from the chest to the head voice in rapid succession, something similar to the German *jödl*, *q.v.* The air "Libiamo" in Verdi's "La Traviata," is called a *brindisi*.

Brio, con (*It.*) With spirit, vigour, and force.

Brioso (*It.*) Joyfully, vigorously, forcibly.

Brisé (*Fr.*) [Broken chords, arpeggios.]

Brisk. Lively. A term frequently used by writers of the last century before the general adoption of the term *vivace*.

Broderies (*Fr.*) Ornaments with which it was the fashion in a past age to cover any simple melody; these were generally left to the caprice of the performer, until Rossini set the fashion of writing those ornaments which he wished his music to bear.

Broken cadence. An interrupted cadence.

Broken chords. [Arpeggio.]

Broken music. Probably music played on harps, guitars, or lutes, because the sounds of these instruments cannot be sustained at will. Shakespeare, "Troilus and Cressida," Act iii. sc. 1:

> "Fair prince, here is good broken music."

Also "Henry V," Act. ii. sc. 1; "As you like it," Act i. sc. 2.

Brontium (*Lat.*) βροντεῖον. A contrivance for imitating thunder, used in the Greek theatre. Sheets of copper were laid out in the hyposcenium over which were rolled bladders filled with pebbles.

B rotundum (*Lat.*) B flat. 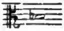 See B cancellatum.

Brummeisen or **Maultrommel** (*Ger.*) Jew's-harp. From *Maul*, the mouth. [Jew's harp.]

Bruscamente (*It.*) Coarsely, roughly, strongly accented.

Buccina. Bucina (*Lat.*) A crooked horn or trumpet, *tuba* being the straight trumpet. It was used as a signal for changing the night-watches, hence the expressions *ad primam bucinam*, *secundam*, &c., at the first and second watches. Public assemblies were also summoned by it in early Roman times. Poets and sculptors have represented Triton as blowing through a *bucina*, from *bucinum*, a shell called the sea-trumpet.

Buccinator. A muscle situated in the fleshy part of the cheeks. It is so called because, when the cheeks are filled with air, the contraction of the buccinator muscles forces it out. It derives its name from *buccinare*, Lat., to blow a trumpet.

Buccolica (*It.*) } Rustic, *à la bucolique*,
Bucolique (*Fr.*) } in a rustic manner.

Büchse (*Ger.*) The boot or foot of an organ pipe.

Buffa (*It.*) *fem.* } Comic. *Aria buffa*,
Buffo (*It.*) *mas.* } a humorous melody
opera buffa, a comic opera.

Buffare (*It.*) To trifle, joke, to play the fool.

Buffet (*Fr.*) An organ case. Key-board case.

Buffone (*It.*) A comic singer in the opera.

Buffonescamente (*It.*) In a burlesque or humorous style.

Bugle. (1) A hunting-horn of a straight or curved form. (2) A copper instrument of the horn quality of tone, but of less compass, furnished with keys. The tone is sweet, powerful, and distinct; it has rarely been employed in the orchestra. There are bugle horns in C, B flat, and E flat, each capable of producing its generator and 7 harmonics. The ventil-horn is an improvement upon the bugle. The word Bugle, from the Anglo-Saxon *buge*, to *bend* or *curve*, was anciently applied to many things of a curved shape, thus, the head of a bishop's crozier was called

Digitized by Google

the *bugle*, and the crozier itself the *bugle-rod*. The handle of a kettle, basket handles, and a peculiar sort of elongated glass-bead are each called by the name *bugle*. Some writers derive the word from *bowgle* or *bougle*, a bull, on the ground that the earliest horns were bull's horns, and that the earliest representations of hunting horns are in shape like bull's horn. [Metal wind instruments.]

Buonaccordo (*It.*) A small triangular spinet for the use and amusement of children, the notes of which were made small to suit the length of their span.

Buonamente (*It.*) Justly, truly.

Buona nota (*It.*) Accented note.

Buon gusto (*It.*) In good taste.

Burden. (1) The chorus or refrain of a song. [Ballad.] (2) The drone of the bagpipe. (3) The tune sung as an accompaniment to a dance when there were no instruments.

> "Foot it featly here and there,
> And let the rest the burden bear."

> "Belike it hath some *burden* then."
> "And clap us into *Light o' love*, that goes without a burden.
> Do you sing it, and I'll dance it."—*Shakespeare.*

> "This sompnour bear to him a stiff bordoune
> Was never trompe of half so gret a soun."
> *Chaucer.*

Burla (*It.*) A jest.
Burlando. **Burlescamente.** } [Buffo.]

Burlesca (*It.*) A jest, a movement in a jocular style, *c.f.* scherzo.

Burletta (*It.*) A comic operetta, a farce interspersed with songs.

Busaun. **Busain.** **Buzain.** A reed-stop on the organ. Generally of 16 ft. length, and on the pedal organ. Its quality of tone is soft. It is not improbable that this word is connected with *bassoon*.

Bussone (*It.*) An obsolete wind-instrument, *cf.* bassoon.

Button. (1) A small round piece of leather which, when screwed on the tapped wire of a tracker, prevents it from jumping out of place. [Organ.] (2) The keys of the first-made accordions. [Accordion.]

Buxus. **Buxea tibia** (*Lat.*) A flute made of boxwood.

> "Tympana vos buxusque vocant Berecyntia matris
> Idææ."—Virg. Æ. ix. 619.

Byssynge songes (*early Eng.*) Lullabyes, cradle songs.

C.

C. (1) The note *Ut* in the Guidonian system and in modern French and Italian nomenclature.

— (2) The letter whose original form was afterwards modified into the C clef.

— (3) The first note of the Hypo-Eolian mode. The first note of the Ionian mode.

— (4) The first, or key note of the modern normal scale, so called because if it be desired to write down the scale now used, C is the only note from which the series can start unless sharps or flats be added.

— (5) A capital letter C signifies the note in the second space of the bass stave (Tenor C). A small *c* signifies the note one octave above this, *middle C*. [Pitch.]

Cabaletta (*Sp.*) (*Lit. a little horse.*) A melody in rondo form, at first sung simply, afterwards with variations, probably so called because accompaniments to cabalettas were in triplet form like the noise made by a horse cantering.

Cabinet d'orgue (*Fr.*) Organ case.

Cabinet Pianoforte. An old-fashioned upright pianoforte, about six feet in height. [Pianoforte.]

Cabiscola (*Lat.*) A corruption of the words caput scholæ. The precentor in a choir (Precentor). In Narbonne and many parts of Italy, the office of capischol was held by the Dean.

Caccia (*It.*) Hunting. (1) Music accompanied by horns, or in praise of field sports, is said to be *alla caccia*, in the hunting style. (2) Instruments used in hunting are called *da caccia*, as *oboe da caccia*, hunting oboe, a large kind of oboe; *corno da caccia*, hunting horn.

Cachucha (*Sp.*) A Spanish dance. [Bolero.]

Cacophony (*Gk.*)
Cacophonie (*Fr.*)
Cacofonia (*It.*)
Harsh sounding music—not necessarily incorrect, but often treated as though it were, because of its unusual appeal to imperfect judgment. The word is, however, generally used in a bad sense.

Cadence. (1) A vocal or instrumental shake or trill, run or division, introduced as an ending, or as a means of return to the first subject.

(2) The end of a phrase, formerly called a fall, either in melody or harmony.

> "That strain again
> It had a dying fall."—*Shakespeare.*

(3) There are four principal forms of cadence in harmony, the whole, or authentic, the half, the interrupted, and the plagal cadence. When the last chord—the major or minor chord of the key note—is preceded by the major chord of the dominant, such cadence is called whole or perfect. If the last chord is the dominant and is preceded by the chord of the tonic, the cadence is called half or imperfect. When the last chord of the phrase is other than the tonic chord and is preceded by that of the dominant, the cadence is said to be interrupted, false, or deceptive. The cadence, called plagal, is that in which the chord of the tonic is preceded by the major or minor chord of the subdominant. The whole cadence is used to conclude most modern music; the half and the interrupted cadence in the progress of a harmonised melody. The plagal cadence was frequently employed as a close by the old contrapuntal writers.

Whole or Perfect Cadences. Half or Imperfect Cadences.

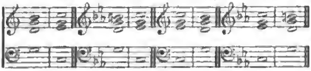

Interrupted, False or Deceptive Cadences.

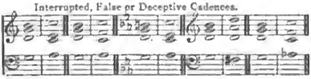

Plagal Cadences.

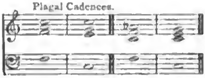

By some authors, cadences are divided into two kinds, perfect and imperfect; the authentic and plagal being considered perfect; all other cadences, imperfect. From another point of view cadences have been divided into simple

Digitized by Google

and compound; a cadence being simple when both the penultimate and final chords which form it are plain common chords; and compound when suspensions or other devices are introduced, *e.g.*:

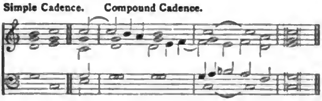

A series of cadences can be constructed by making any one of the *relative chords* (or its inversions) precede the final tonic chord, a relative chord being a common chord which can be made up out of the notes of any given scale. The relative chords of C are therefore D minor, E minor, F major, G major, A minor, but not B major or minor, as there is no F♯ in the key of C. We shall, therefore, get this series:

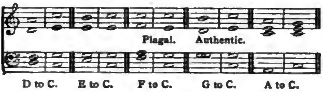

And from the relative chords of the minor scale:

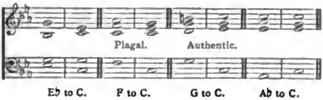

It will be observed that there is no relative chord on the second or seventh degree of the minor mode owing to the imperfection of the interval of the fifth. For further information see Stainer's "Treatise on Harmony."

Cadence imparfaite (*Fr.*) An imperfect cadence.

Cadence parfaite (*Fr.*) A perfect cadence.

Cadence perlée (*Fr.*) A brilliant cadence.

Cadence rompue (*Fr.*) A broken or interrupted cadence.

Cadenz (*Ger.*) [Cadence.]

Cadenza (*It.*) (1) A passage introduced towards the close of the first or last movement of a concerto, either actually extempore or of an impromptu character. (2) A running passage at the conclusion of a vocal piece. Solo performers in times past were accustomed to display their skill and invention in a final flourish, apparently extempore, but often the result of careful study and preparation. In some cases, however, the attainment of the performer was the object of the display, and hence the added cadenza was often so inappropriate and incongruous, especially in vocal music, that composers felt bound to write down all the ornaments or embellishments they considered their music capable of bearing. Purcell is said to have so acted with regard to many of his songs; and it has been supposed that the runs or divisions so common in music of the 18th century were introduced as concessions to the custom of the time of ornamenting a plain melody. Every performer considered himself at liberty to alter an air to suit his own peculiarities, and singers were estimated according to their vocal agility. An extract from a diary kept at Rome in 1697, by a young Scotch gentleman, speaking of Corelli and his playing, says: "This is his manner in adagios, to which he adds innumerable graces; not crowded in confusion as some do, but gentle, easy and sliding, and suited withal to the composition of the other parts, which no man but he who has taste and knowledge of the composition, can perform." When a great master in art sets the example, followers are always found, and what is pleasing in a man of genius becomes the reverse when imitated. Although cadenzas were fashionable in the time of Handel, instances of fully written examples by him are rare. The conclusion of the duet, "O lovely peace," in "Judas Maccabæus," is one of the few specimens he has left. Many of the songs in his oratorios were constantly so changed by the singers by means of graces, notes, and turns, that their form was completely disguised. But while the custom existed there were not wanting some musicians who constantly protested against what they considered the ill-usage of an author's ideas, forgetting that the composer, probably knowing the bad habit of his singers, had constructed his melodies so that they might not suffer by the overlading of *fioritura*. Rubinelli the singer, on his first appearance in England, was censured for embellishing and changing his airs. On his second appearance in this country, he determined to sing without introducing a single ornament not written, and so fickle was the taste of the time (c. 1780), that when he sang "Return, O God of Hosts," from "Samson," in Westminster Abbey, his hearers thought the song and his style of singing alike insipid.

The omission of ornaments in a musical performance was a matter for surprise a century earlier than that just named. Richard Lygon, in his "History of Barbadoes, 1687," describes his satisfaction at hearing a minstrel sing a song, "savouring much of antiquity — no *graces*, *double relishes*, *trillos*, *grupos*, or *piano-fortes*, but plain as a pack-

Digitized by Google

staff; his lute, too, was but of ten strings, so that the rarity of this antique piece pleased me beyond measure."

Cathedral chants, services, and anthems —even psalm and hymn tunes—were written with every possible florid turn, as shown by existing examples of the Church compositions of the latter part of the last century.

The story told of the elder Dubourg and Handel's comment upon his cadenza is well known; and there is another anecdote illustrating the absurdity of a misplaced cadenza, told concerning the trombone player at the first performance of Mendelssohn's "Lobgesang." The composer met the player and asked him if he had looked at his part, as he had given him plenty of important work to do. "O yes, Herr Director, I have studied it carefully." The astonishment of Mendelssohn may be imagined when at the rehearsal he heard the result of the *careful study* of the trombonist in the announcement of the initial phrase of the symphony as follows:

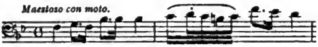

Mendelssohn, in his sweet manner, told the performer that he would rather have the phrase played as he had written it. Other musicians of less agreeable dispositions have sarcastically thanked performers for taking "so much trouble to sing or play notes that were not written;" and it is on record that Beethoven repeatedly quarrelled with vocalists for not adhering to his text, and it is also well known that Rossini wrote every cadenza out in full, "for he thought it better so to do than to trust to the *gaucheries* of conceited vocalists."

In instrumental compositions the habit of leaving a space for the *ad libitum* fancies of the performer opened a door for the admission of eccentricities and absurdities, which the better sort of musicians have sought from time to time to remedy, by the composition of suitable cadenzas as much as possible in accordance with the original composer's design. Clementi wrote cadenzas for the whole of Mozart's concertos, and Dr. Hiller and others have done like things for other works in which spaces have been left.

The cadenza has been made the vehicle for the expression of musical humour, as by Mendelssohn in the Music to "A Midsummer Night's Dream," and of quaintness in instrumentation, as in Beethoven's No. 5 Concerto, and in other works needless to particularise. "It is usual," says Jousse in his "Dictionary of Music," "to commence a cadenza with a plain note or chord sung or held out, so that the accompanying performers may know when it has been begun; and it is also customary to make a long shake at the end of the cadenza, as a signal that the accompaniment is to be resumed."

Cadenza d'inganno (*It.*) A deceptive cadence.

Cadenza fioritura (*It.*) An ornamental cadence.

Caisse (*Fr.*) A drum.

Caisses claires (*Fr.*) Snare-drums; *grosse caisse* (*Fr.*), big drum.

Caisse roulante (*Fr.*) Tenor-drum, larger than a snare-drum.

Calamus. (Gk. *κάλαμος*) A reed-flute. Probably a simple rustic instrument like our oaten-pipe. But some suppose it to have been similar in construction to the *syrinx*, or pan's-pipes, and to have been synonymous with *arundo*. From *calamus* is derived the post-classical *calamaulos*, a flute made of reed, whence calamaulis (*καλαμαύλης* and *καλαμαυλήτης*) a player on reed-pipes; hence too, *chalumeau*, *schalmey*, *shawm*, the precursor of the modern clarinet, one of the registers of which is still said to be of *chalumeau* tone.

Calando (*It.*) (*Calare.* To descend, decrease.) A passage marked *calando* is to be sung or played with decreasing volume of tone and slackening pace.

Calandrone (*It.*) (Calandra, a woodlark.) A small reed instrument of the shawm or clarinet character, with two holes, much used by the Italian peasantry.

Calascione. [Colascione.]

Calata (*It.*) An Italian dance in $\frac{2}{4}$ time, of a sprightly character.

Calcando (*It.*) Hurrying, pressing the time.

Calcant (from *Lat.* Calcare.) Treading. The bellows-treader (Bälgetreter) of the old German organs.

Calcanten-Glocke (*Ger.*) Bells sounded by means of pedals.

Call. A military term for the variations of certain musical notes played on a trumpet or bugle, or a special sort of beat upon the drum, each call being the signal for a definite duty.

Call. A toy instrument made by winding a narrow tape round two small oblong pieces of tin, so that one fold of the tape may be set in vibration when blown through. The call is used by men who *work* the drama of "Punch and Judy."

Ça ira (*Fr.*) 'That will do.' The refrain of a song popular during the revolution in France in 1793. The melody to which it was sung was a favourite with the unhappy Queen

Marie Antoinette. The song was called the "Carillon national."

Le refrain.

Ah! ça ira, ça ira, ça ira,
Le peuple en ce jour sans cesse répète
Ah! ça ira, ça ira, ça ira,
Malgré les mutins, tout réussira.

Calma, con (*It.*) With calmness.

Calmato (*It.*) Calmed, quieted, appeased.

Calore, con (*It.*) With heat, warmth.

Caloroso (*It.*) Warmly, full of passionate feeling.

Cambiare (*It.*) To turn, change, alter.

Cambiata (*It.*) [Nota.]

Camera, musica di (*It.*) [Chamber Music.]

Camminando (*It.*) Walking, flowing, *andante*.

Campana (*It.*) A bell. [Bells.]

Campanella,-o, (*It.*) A small bell.

Campanellino (*It.*). A very small bell.

Campanista (*It.*) A bell-ringer.

Campanology. The knowledge of the construction and use of bells. [Bells.]

Campanetta (*It.*) A set of bells tuned to a scale, and played with hammers or keys. [Glockenspiel.]

Canaries. A dance probably of English invention. The melody was a lively air of two phrases. Purcell introduced a Canaries tune in his opera of "Dioclesian."

The following example (from Delaborde) shows the rhythm of this dance:

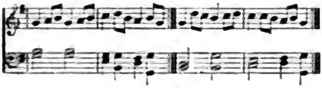

Cancrizans. [Canon Cancrizans.]

Canon (*Gk.* κανών). A rule,—a term applied to the measurement of the ratios of intervals by means of the monochord, hence the system of Pythagoras was called the *canon* of Pythagoras; that of Euclid, the *canon* of Euclid. Hence, too, the science of calculating musical intervals is called *canonik*. *Sectio canonis* (*Lat.*), a division of a string, or monochord, formed by a moveable bridge or frets.

Canon. Owing to the various forms which canons assume it is almost impossible to give a general definition which will be intelligible. The essence of a canon is this, that the music sung by one part shall, after a short rest, be sung by another part note for note. The simplest form is when there are only two parts, *e.g.*:

Ex. 1.

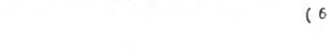

The above is called a canon 2 in 1 at the octave, because two parts are singing one thing at that interval. The part which commences is called the *subject* or *antecedent* (guida); that which follows, the *answer* or *consequent* (consequenza). The above is also an *infinite* canon, because, anyone having such a remarkable desire as to play it for ever, could do so. The pause shows where it may be concluded.

Ex. 2.

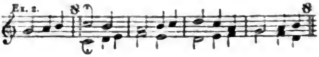

The above (Ex. 2) is also 2 in 1, but at the under-fifth.

Ex. 3.

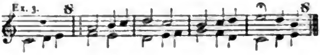

Ex. 3 is a canon 2 in 1 at the upper sixth, the upper part being the consequent.

Ex. 4. SCHUBERT. "Song of Miriam."

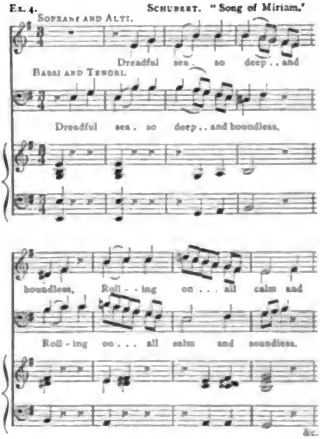
SOPRANI AND ALTI.
Dreadful sea, so deep .. and
BASSI AND TENORI.
Dreadful sea, so deep .. and boundless,
boundless, Roll - - ing on . . . all calm and
Roll - ing on . . . all calm and boundless.
&c.

The above example (4) shows a canon 2 in 1 at the octave, with a *free* accompaniment. Any part of a canon which is not an antecedent or consequent is said to be a *free part* (*ad placitum*). It is also *finite* because there is no repeat, the canon being dropped at the close of the theme. The same description will apply to the next example (5).

Ex. 5. **Beethoven.** Symphony, No. 4.

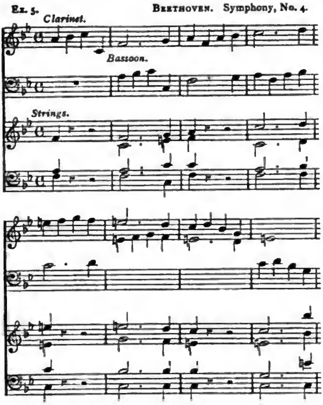
Clarinet.
Bassoon.
Strings.

Ex. 6. (Transposed.)

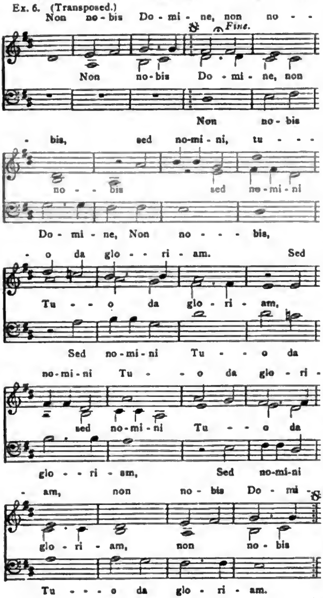
Non no-bis Do-mi-ne, non no-
Fine.
Non no-bis Do-mi-ne, non
Non no-bis
-bis, sed no-mi-ni, tu-
no-bis sed no-mi-ni
Do-mi-ne, Non no-bis,
-o da glo-ri-am. Sed
Tu-o da glo-ri-am,
Sed no-mi-ni Tu-o da
no-mi-ni Tu-o da glo-ri-
sed no-mi-ni Tu-o da
glo-ri-am, Sed no-mi-ni
-am, non no-bis Do-mi-
glo-ri-am, non no-bis
Tu-o da glo-ri-am.

The above well-known canon by Byrde is 3 in 1, because there is only one theme which all the three parts sing.

Ex. 7. J. S. Bach. Mass in B minor.

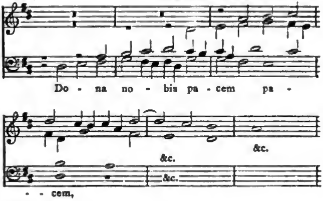
Do-na no-bis pa-cem pa-
-cem,
&c.
&c.
&c.

The above example (7) is a finite canon 4 in 1.

Ex. 8. Attwood. Service in F.

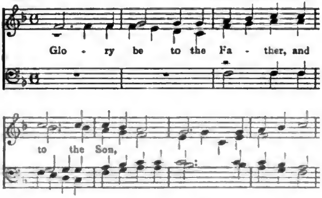
Glo-ry be to the Fa-ther, and
to the Son,

The above example (8) shows a canon 4 in 2 because it is in four parts and there are two themes. Enough has been given to show the exact meaning of the numerical descriptions of canons; the first number giving the number of the parts in which it is composed; the second number, the number of themes sung by them, thus 16 in 4 signifies that 16 parts have 4 subjects; 8 in 1 that 8 parts sing in turn the same theme, &c. A canon by *augmentation* is when the consequent is double the length of the antecedent, *e.g.*:

Ex. 9. From Cherubini.

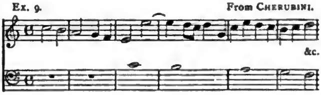

A canon by diminution is when the consequent is half the value of the antecedent, *e.g.*:

Ex. 10. From Cherubini.

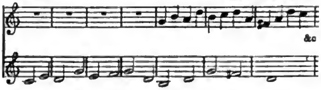

Fragments of canon by *augmentation* and diminution are not uncommon in fugal writing, *e.g.*:

Ex. 11.

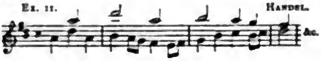

A canon by triple or quadruple augmentation is when the three or four parts of which it is composed are each twice the time-value of its predecessor. A canon is said to be *strict* when the consequent follows the antecedent at an exact interval (say a major fifth or fourth, &c.) regardless of key tonality. The canon in Ex. 3 is not therefore strict. If it were so, the consequent must be in the *key* of the sixth above, which would be impossible. A canon by *inversion* is when the consequent follows the *inverted intervals* of the antecedent, *e.g.*:

Ex. 12.

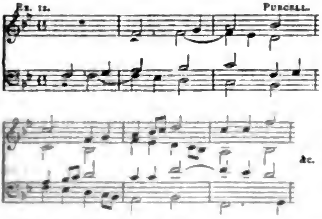

The above (Ex. 12) is a canon 4 in 1, because there is only one antecedent. The part appearing like a second antecedent being only the inversion of the first. A canon by *retrogression* is when the parts forming it (generally, only one is antecedent and one consequent) sing each other's notes backwards. An example will be found under "canon cancrizans."

Originally canons were a kind of musical riddle, the antecedent, and the number of parts, only being given; and the student being required to solve the problem. Thus, Ex. 1 would be put forth:

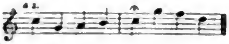

And Ex. 2:

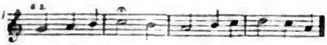

"Non nobis Domine" would be given thus:

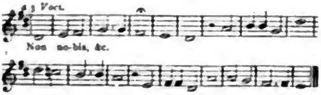

From this method of enunciating canons, the name is probably derived, as the reader had to discover the *rule* or *canon* on which the composition was constructed. A canon written out in full was called *canone aperto*, and one written in riddle-form *canone chiuso*. A canon at the unison becomes a *round*, if the antecedent has a cadence before the entry of the consequent. Thus every round is a canon at the unison, although a canon at the unison is not necessarily a round. [Round.] Some of the early writers have left canons of the greatest ingenuity. Some very good specimens are to be found in Hawkins. It had been well if the labour and perseverance which must have been requisite for their production could have been more profitably directed. The constant study of canon-writing is much to be deprecated, as it checks the inventive faculty, and at most only teaches the student how to force themes into cohesion. It is probable that much of the ugly and crabbed part-writing of the 17th and early part of the 18th century is due to the over estimation of canons. Canonical imitation with free accompaniment is, however, capable of very beautiful effects. Specimens of this style have already been given in Ex. 4 and 5, and the fine example in Mendelssohn's 95th Psalm may be studied with advantage. The highly dramatic effect of the canon in two parts, afterwards breaking into four, at the words, "And the sea was upheaved," in No. 34 of Mendelssohn's "Elijah," is so well known that it need not be quoted here.

Canone al sospiro (*It.*) A canon, the subject of which is answered at one beat of time:

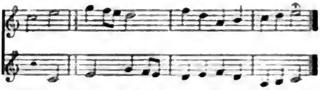

The answer to any subject is said to be *close* when it enters shortly after the subject. A canon *al sospiro* is therefore the most *close* of all canons, as it is impossible to answer at less time than the beat.

Canone aperto (*It.*) A canon written out in full.

Canon cancrizans. A canon by retrogression. A canon practically consisting of two parts in double counterpoint, that is, parts which are grammatically interchangeable, so constructed that they may read actually backwards, hence probably the derivation of cancrizans, *walking backward like a crab*. The following example will be found to consist only of four bars, at the close of which, having exchanged lines, the parts proceed backwards. A canon cancrizans may of course be accompanied by free parts:

From André's "Lehrbuch der Tonkunst," 1832.

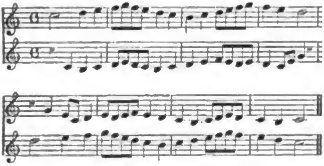

The following is a canon cancrizans with a bass part *per recte et retro:*

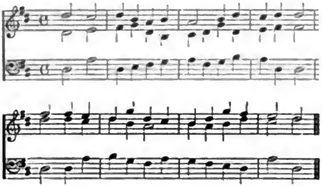

Productions of this class are utterly valueless as contributions to art.

Canone chiuso (*It.*) A close canon. [Canon.]

Canone infinito or **perpetuo** (*It.*) Never ending canon. Infinite canon.

Canone sciolto (*It.*) A free canon, not strict.

Canonici. A name given to followers of the Pythagorean system of music, as opposed to Musici, the followers of the Aristoxenian system. [Pythagoreans.]

Cantabile (*It.*) In a singing style.

Cantadour (*Old Fr.*) } A street singer.
Cantambanco (*It.*) } A mountebank.

Cantamento (*It.*) The air or melody of a phrase.

Cantando (*It.*) [Cantabile.]

Cantante (*It.*) A singer.

Cantare (*It.*) To sing.

Cantare a aria (*It.*) To sing with a certain amount of improvisation. [Penillion singing.]

Cantare a orecchio (*It.*) To sing by ear.

Cantare di maniera (*It.*)
Cantare di manierata(*It.*) } To sing in a florid or ornamental style.

Cantata (*It.*) A cantata consisted originally of a mixture of recitative and melody, and was given to a single voice, but the introduction of choruses altered the first character of the cantata, and gave rise to some confusion in the manner of describing it. So that it has been variously defined as "an elegant and passionate species of vocal composition for a single voice," "a long vocal composition, the text of which is Italian," "a kind of short oratorio, or opera not intended for the stage," "a short piece of vocal music of a pathetic character," "one of the Psalms or portions of Scripture set to music for voices and instruments," according to the work the describer had in his mind at the time, but a cantata is now understood as a short work in the musical form of an oratorio, but without *dramatis personæ.*

Cantatilla
Cantatina (*It.*) } The diminutive of Cantata.

Cantatore (*It.*) A male singer.

Cantatrice (*It.*) A female singer.

Cantatorium. A music book.

Cantellerando (*It.*) Singing in a subdued voice, trilling.

Canti carnascialesci
Canti carnivali (*It.*) } Songs sung during the carnival.

Cantici (*It.*) Another name for the Laudi spirituali, or songs sung in the old Romish church in praise of God, the Blessed Virgin and Saints, and Martyrs.

Canticle (1) A song or hymn in honour of God, or of some special sacred event.

(2) The word is also applied to certain detached psalms and hymns used in the service of the Anglican Church, such as the *Venite exultemus*, *Te Deum laudamus*, *Benedicite omnia opera*, *Benedictus*, *Jubilate Deo*, *Magnificat*, *Cantate Domino*, *Nunc dimittis*, *Deus misereatur*, and the verses used instead of the Venite on Easter-day.

Canticum (*Lat.*) (1) A song. (2) A song in the Roman comedy, accompanied by music and dancing. Sometimes one person sang the song while another went through the appropriate gesticulation.

Cantilena (*It.*) (1) An oft-repeated, old song. (2) In mediæval music, singing exercises, in which were introduced all the intervals of the scale, &c. (3) In old church-song the plain-song or canto-fermo sung in unison by one or more persons to an organ accompaniment. (4) A ballad.

Cantilenare (*It.*) To sing without accompaniment.

Cantilenaccia (*It.*) Bad singing.

Cantillatio (*Lat.*) Declamation in a singing style, applied to a method of reading the Epistles and Gospels in the church. [Accentus Ecclesiasticus.]

Cantino (*It.*) The smallest string upon the violin. The E string. (*Fr.*) *chanterelle.*

Cantique (*Fr.*) A sacred song or melody, a canticle.

Canto (*It.*) The upper voice-part in concerted music, so called because it has the melody or air. [Air.]

Canto a cappella (*It.*) Sacred music; *cantore di cappella*, the precentor.

Digitized by Google

Canto armonico (*It.*) A part-song.

Canto cromatico (*It.*) A scale or song in chromatic style.

Canto fermo (*It.*) [Cantus firmus.]

Canto figurato (*It.*) Florid melody, or melody varied. [Cantus figuratus.]

Canto Gregoriano (*It.*) Gregorian chant.

Canto llano (*Sp.*) Plain chant.

Canto plano (*It.*) Plain chant.

Canto primo (*It.*) First soprano.

Canto recitativo (*It.*) Declamatory singing, recitative.

Canto ripieno (*It.*) Additional soprano chorus-parts. [Ripieno.]

Canto secondo (*It.*) Second soprano.

Cantor. [Precentor.]

Cantor choralis (*Lat.*) Chorus master.

Cantore (*It.*) A general name for a singer.

Cantoris (*Lat.*). (From the word Cantor.) The cantoris side in a cathedral choir is the side upon which the Precentor sits, usually the north side, opposite to Decaní.

Cantus Ambrosianus (*Lat.*) Ambrosian chant. [Plain-song.]

Cantus coronatus (*Lat.*) [Cantus fractus.]

Cantus durus (*Lat.*) Music which modulated into a key having one or more sharps in its scale. Such keys were at one period strictly proscribed by church-musicians.

Cantus ecclesiasticus (*Lat.*) (1) In a general sense, plain-song and other early church-melodies. (2) The method of *singing* as opposed to *saying* Lections, Collects, Gospels, and special offices, such as the Improperia, &c. See Accentus ecclesiasticus under Accent § 4, and "Passion Music."

Cantus figuratus (*Lat.*) Florid church song, that is, in which more than one note of music was sung to a syllable. The purest system of ancient church-song prescribed only one note to each syllable. [Plain song.]

Cantus firmus (*Lat.*) (1) The *tenor* or chief melody, originally sung by the tenor-voices, afterwards transferred to the treble-part, hence called *Canto*. (2) A fragment of plain-song, to which counterpoint has been added. (3) Any subject chosen for contrapuntal treatment, generally a short diatonic passage of semibreves or other long notes.

Cantus fractus (*Lat.*) A broken melody, a term applied to a tune which proceeded either by perfect or imperfect consonances. When accompanied by a Faburden, or Fauxbourdon, it was called Cantus coronatus.

Cantus Gregorianus (*Lat.*) The Gregorian system of church-song. [Plain song.]

Cantus mensurabilis (*Lat.*) Mensurable-song. The very name of this art explains at once its scope and the probable date of its birth. The indissoluble association of music and poetry, or of music, poetry, and dancing, in ancient times, rendered a system of notation, by which the comparative duration of sounds could be exhibited to the eye, unnecessary. If the metre of the poetry were duly appreciated, the length of the musical notes to which the poetry was set would be undoubted. If dancing accompanied the music and poetry, it would be, of course, impossible to sing to any other rhythm than that prescribed by the movement of the feet. As long as music of this kind was unisonous, or, at most, consisted of a series of chords, the component parts of which were of equal length, no difficulty or doubt as to the length of notes could occur. But when prose-writing was set to music, and still more when, in polyphonous compositions, it was desired that a particular voice should sing two or more notes to one note of another, it became an absolute necessity that the signs used should be so formed as to direct the performer, without a chance of doubt, as to how long he should hold any note with reference to that held in another part. Hence, the formation of Cantus mensurabilis. As to the date of its invention, learned and reliable authors differ much in their opinion. Having been ascribed to Johannes de Muris (circ. 1330) for many centuries by writers who have been but too ready to copy from each other, asking no questions, it seems that the laurel must be taken from his brow, and that the credit is due to authors who lived—some say a few years, others two centuries at least—before him. It is, however, certain that Robert de Handlo wrote on the subject before Johannes de Muris, and equally certain that Robert de Handlo had the benefit of the labours of Franco. But here a new difficulty arises: not only was Franco so common a name that many learned Francos existed at the same date, but at least three of this name were musicians—Franco of Paris, Franco of Cologne, Franco of Liége. Nor is this all—two distinct dates are attributed to the Franco who wrote on Cantus mensurabilis, which differ by about 200 years! The reader who cares to enter deeply into this question may refer to Fétis, Kiesewetter, Hawkins, Burney, Forkel, and Coussemaker, all of whom have bestowed much thought on the subject; having done so, he will find that he is still in ignorance. The truth is, that mensurable music, like many other highly important ingredients of our intellectual life, was a *growth*, not a sudden invention. There are evidences that in the twelfth century a proportionate

Digitized by Google

subdivision of the length of sounds was reached after, and naturally enough, the first step was, that two sounds might be sung to one, hence the *long* and *short*, or *long* and *breve*, as they were called. The shortest note or *minim* found its way into use, probably, in the thirteenth century, and was in time followed by other subdivisions. Then followed the triple division of notes, a threefold division being called *perfect* on theological grounds; then rapidly followed, in the 14th and 15th centuries, a complication of mensurable signs, which now baffles the most enthusiastic interpreter of music of that period,—the value of notes varying according to their position with regard to other notes; or, according to the position of the tails, if up or down, or on the right or left sides; or, as to the complete blackness or open outline (*evacuatio*) of the notes; or as to the manner in which consecutive sounds to one syllable were written in *continuous lines*, forming *ligatures*. Happily, from the 16th century a genuine taste for part-music led to an unremarked disuse of these utterly useless conceits, a full account of which can only be found in ancient learned treatises, where any one having more taste for music than antiquities, will do well to leave them.

Cantus planus (*Lat.*) Plain song.

Cantus Romanus (*Lat.*) Roman chant or song. (1) The Gregorian system of music. (2) The early attempts at harmonizing a melody known as the *organum*.

Canun or **Kanoon** (*Turkish*). An instrument strung with cat-gut, in form like a dulcimer, with which the women in the harems accompany their singing. The sound is brought out by means of plectra—thimbles made of tortoiseshell pointed with cocoa-nut wood, and worn upon the ends of the fingers.

Canzona, Canzone (*It.*) (1) A short song, the music being more important than the words. It is one of the ancient forms of measured melody, and when the older writers employed it, it was usually made the vehicle for the display of skill and contrivance in the treatment of the phrases in fugal imitation. A secondary meaning of the word, *scoffing* or *banter*, perhaps accounts for the use of a *form* in which a musical imitation or *mocking* was shown.

(2) In the early part of the last century the word was used to describe an instrumental composition, similar to the sonata as then known.

(3) It was also understood to mean the same as allegro, "for it denotes that the movement of the part to which it is fixed ought to be after a lively, brisk, or gay manner."

CANZONA.*

Girolamo Frescobaldi (1591—1640).

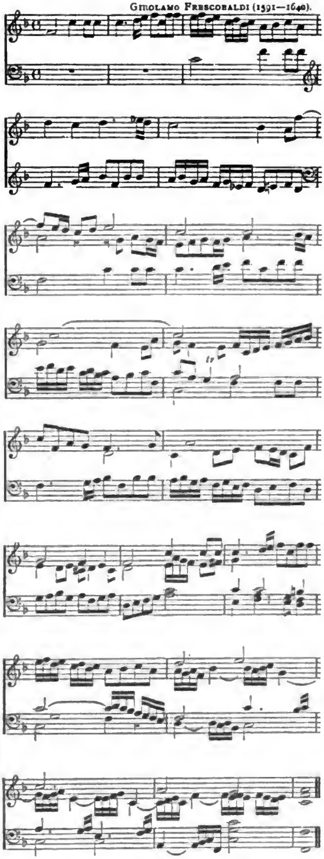

Canzonaccia (*It.*) A commonplace song.
Canzoncino (*It.*) A short poem or air.

---

* From "Il secondo libro di Toccate, Canzone versi d'Hinni Magnificat, Gagliarde, Correnti, et Altri Partite, di Cembalo et Organo." Rome, 1637.

Canzonet, Canzonetta (*It.*) A diminutive of canzona, "denoting a little short song, tune, cantata, or suonata." Originally applied to a short song in parts. Luca Marenzio, Giovanni Ferreti, and Horatio Vecchi are said to have excelled in this species of composition. The title was also employed by poets to describe verses either of a trifling character or subject; and musicians, when they set such words, repeated the poet's title without reference to the musical meaning of the word. Brossard, "Dictionnaire de Musique, 1703," speaks of two sorts of canzonets—the Neapolitan, with two phrases, and the Sicilian, a sort of jig in $\frac{12}{8}$ or $\frac{6}{8}$ time, each in rondo form. Thomas Morley (1597) describes a series of madrigals as "Canzonets, or Little Short Songs to Four Voyces; celected out of the best and approued Italian Authors;" and Haydn's use of the word with reference to his well-known examples will be familiar.

Canzoniere (*It.*) A lyric poem or song.

Caoinan (*Irish.*) A funeral song (Keeners).

Capellmeister (*Ger.*) Maestro di Cappella (*It.*) (1) The musical director of a church or chapel. A post of considerable honour, especially when connected with a royal or ducal chapel. The list of eminent musicians, from Palestrina to Mendelssohn, who have held such offices is very large, and the fact that men of general musical ability have thus been necessarily brought into contact with sacred music, has probably greatly influenced the character of the compositions of the 16th, 17th, and 18th centuries. There is no post in the English Church or at our Court which exactly corresponds to that of Capellmeister, including as it does the duties—as circumstances may require—of conductor, accompanist, choir-trainer, and composer. The *choir-master*—an office lately instituted or revived in this country—is perhaps the nearest approach to the Capellmeister. By the combination, which not unfrequently took place, of the offices of "Composer to his (or her) Majesty" and "Master of the Children of the Royal Chapel," a veritable Capellmeister was created. In our cathedrals the precentor and organist practically divide the duties of this post.

(2) The title has sometimes been applied to a conductor of a band or an opera.

Capellmeister-Musik (*Ger.*) A term of contempt for music *made* and not inspired.

Capiscolus (Precentor) Cabiscola.

Capistrum (*Lat.*) A muzzle. A sort of bandage wound round the head and face of the ancient trumpeters, to protect the cheeks while playing their instruments, on account of the unusual exertion necessary for the proper production of tone.

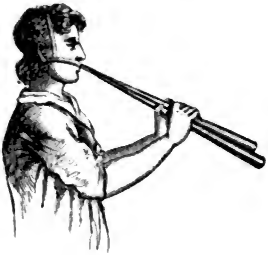

Cappella, alla (*It.*) In the ecclesiastical style. In duple time. [A cappella.]

Capo (*It.*) Head, commencement.

Capo, da (*It.*) A direction to return to the first or other indicated movement.

Capo d'opera (*It.*) (1) The principal song or piece in an opera. (2) A *chef d'œuvre*.

Capotasto (*It.*) (Lit. head-stop.) A mechanical arrangement by which the pitch of the whole of the strings of a guitar is raised at once. The capotasto, or capodastro as it is sometimes called, is screwed over the strings on to the finger-board and forms a temporary nut, *e.g.*, *con capotasto sulla 3a Pos.*

Capriccietto (*It.*) A little caprice, or fancy.

Capriccio (*It.*) A freak, whim, fancy. A composition irregular in form.

Caprice (*Fr.*) [Capriccio.]

Capriccioso (*It.*) Whimsical, humorous.

Caractères de musique (*Fr.*) The signs used in music. [Notation.]

Caral (*old Eng.*) Kyrriole (*Ang.-Sax.*) [Carol.]

Carattere (*It.*) Character, dignity, quality.

Carezzando (*It.*) } Caressingly, singing or
Carezzevole (*It.*) } playing with a frequent introduction of notes of anticipation or *appogiatura*.

Caricato (*It.*) Loaded, over displayed.

Carillon. A set of bells so arranged as to be played by hand or by machinery. The word has by some authors been connected with (*Fr.*) *clarine*, a little bell, which is pro-

Digitized by Google

bably connected with (*Lat.*) *clarisonus;* but others derive it from the word *quadrille*, or *quadriglio*, on the ground that this dance was popular, and probably "set" to bells, in the 16th century. There can be no doubt as to the antiquity of thus using *small* bells. They were probably graduated in size so as to produce a diatonic scale, and were called a Tintinnabulum.

Fig. 1.

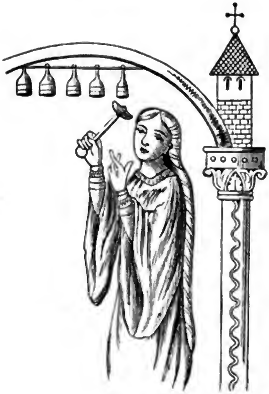

Fig. 2.

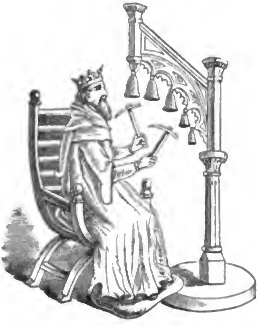

Fig. 3.

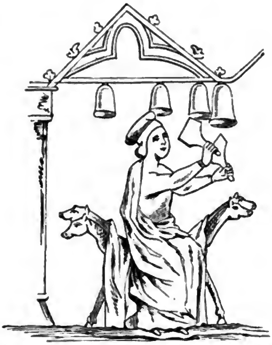

Fig. 1 is given by M. Coussemaker as being from a MS. probably of the 9th century. Fig. 2 is from an ancient Psalter in the British Museum. Fig. 3 is from a MS. in the Royal Library of Brussels. Five seems to have been the number of bells usually employed in earliest times, but they were afterwards increased to six or seven. It is to the bell-founders of the Low Countries we owe the perfecting of the art of bell-founding and the construction of carillons, during the 15th, 16th, 17th, and 18th centuries. Pre-eminent among them stands the Van den Gheyn family, whose works are to be found in almost every Belgian belfry. Originally of Mechlin, they afterwards removed to Louvain, where Matthias Van den Gheyn (b. 1721) deservedly attained the highest fame, as organist, composer, carillon-maker, and carilloneur. The brothers Von Aerscholdt, the great bell-founders, now living in Louvain, are lineal descendants of Matthias Van den Gheyn. The finest carillons, namely those at Antwerp, Mechlin, Bruges, Ghent, and Namur, consist of about forty bells, extending from huge specimens of several tons in weight up to little bells weighing only a few pounds.*

---

* The fine chimes in Mechlin consist of 45 bells, the largest of which weighs between 9 and 10 tons. This rich-toned bell was cast by Aerscholdt in 1844. At Ghent there are 48 bells (44 above and the 4 heaviest in the lower storey), the largest of which was cast by Du Mery, 1744, and weighs about 5½ tons. At Antwerp there are in reality two carillons—one connected to the machinery, and in use, the other disused. That in use consists of 48 bells, the largest of which weighs about 7 tons. At Bruges there are 48 bells, the largest nearly 10 tons. At Namur there are about 50 bells, the largest about 4 tons. . Many of the bells in the Belgian chimes are found to be of Dutch make, and (by their inscriptions) have been issued from old foundries in Amsterdam, Rotterdam, Zutphen, and elsewhere.

They are in most cases arranged as follows: the smaller bells are fixed to strong timbers and arranged in rows, according to size, the largest being nearest to the floor—the bells and framework thus representing the outline of a pyramid. Where there are many specially large bells, these are generally placed in a lower storey, not uncommonly below the chiming machinery. To each bell is attached one or more hammers on the outer side, and a clapper in the inside. To the lever-end of the hammers thick wires are attached, which pass down to long iron rods. The lever-end of these rest on the *tambour*, or barrel, on which are arranged projecting staples. When the barrel is turned (which is done by ordinary clockwork) the staple forces up the end of the iron rod, the other end at the same time pulling down the wire and raising the hammer. When the barrel *releases* the iron-rod, it drops suddenly and causes the hammer to strike the bell. Some time is of course required for the raising of the larger hammers, hence the necessity of having several hammers to some of the bells, so that if a quick repetition of the sound is required, one hammer shall be ready to strike while another is being brought into position. There are, therefore, always a larger number of staples on the barrel than there are bells in the carillon. The clapper, before-mentioned as being in every bell, is held by a wire-loop, within an inch or two of the side of the bell; this wire passes down to the *clavier*, or keyboard—a series of small round sticks, arranged in an order similar to that of the black and white keys of a pianoforte, but separated from each other by a sufficient distance to allow each one to be struck with the *fist* without fear of that on either side of it being also struck. The clappers of the heaviest bells are, owing to their weight, generally attached to a pedal-board, and the carilloneur usually guards his hand with a thick glove when playing.

It will be understood from this short description that the mechanism by which these beautiful bells are chimed and played is of the roughest description. Vast improvements have, however, been lately made, chiefly in England; and Messrs. Gillett and Bland have invented an ingenious piece of mechanism, by which the hammers are held up constantly, and only have to be *released* by the action of the barrel. This insures a regularity in the striking which cannot on the old system be attained, and does away with the necessity for multiplying hammers to a single bell.

The higher octaves contain generally a complete *chromatic* scale. But the heavier bells, owing to their great cost and the large amount of room they occupy, are limited to such important fundamental basses as tonic, subdominant, and dominant; or, at most, to the first five degrees of the *diatonic* scale. A short "flourish" is played at the half-quarter, a slightly longer phrase at each quarter, a tune at each half hour and hour. It is to be regretted that we in England are but just beginning to appreciate the beauty of the effect produced by carillon - music. But, on the other hand, nowhere but in England can genuine change-ringing be heard, in which, the tone produced by the bells as they swing completely round is totally different in character from that obtained by the dead stroke of a hammer. But bells can be easily arranged so as to do the double duty of chiming and change-ringing, and it is to be hoped that they will often in future be so arranged.

Carillonneur (*Fr.*) Bell-player. [Carillon.]

Carità, con (*It.*) With tenderness.

Carmagnole. A dance accompanied by singing, named from Carmagnola in Piedmont. Many of the wildest excesses of the French revolution of 1792 were associated with this dance. It was afterwards applied to the bombastic reports of the French successes in battle. The song commenced with "Madame Véto avait promis," and each verse ended with the burden "Dansons la carmagnole, vive le son du canon."

Carol. To sing or warble, to celebrate in song.

Carol. A song of praise, applied to a species of songs sung at Christmas-tide. It originally meant a song accompanied with dancing, in which sense it is frequently used by the old poets (perhaps connected with *choraula*). It appears to have been danced by many performers, by taking hands, forming a ring, and singing as they went round. It will be readily imagined that a dance of this character would lead to a certain wildness if not rudeness of behaviour, so that the warning contained in the following verse addressed to those of gentle blood who indulged in the exercise, might not be altogether unnecessary:

"Fille quant ferez en karolle
Dancez gentiment par mesure
Car, quant fille se desmesure
Tel la voit qui la tient par folle."

Bishop Taylor says that the oldest carol was that sung by the heavenly host when the birth of the Saviour was announced to the Shepherds on the plains of Bethlehem. It is probable that the practice of singing carols at Christmas-tide arose in imitation of this, as the majority of the carols declared the good tidings of great joy; and the title of Noels, nowells, or novelles, applied to carols, would seem to bear out this idea.

Carol singing is of great antiquity among Christian communities, as the carol by Aurelius Prudentius, of the 4th century, will show.

Digitized by Google

This poem contains twenty-nine stanzas, commencing:—

"Quid est quod arctum circulum
Sol jam recurrens deserit,
Christusne terris nascitur,
Qui lucis auget tramitem?"

Carols were both serious and humorous in the 14th and 15th centuries. Mr. Chappell quotes a tune that might be sung to words of either character, but bearing reference to the observances of the season of Christmas. (Popular Music, i. 42.)

In later times carols were written of a more sober character, and we find in 1630 the publication of "Certaine of David's Psalmes intended for Christmas carols fitted to the most sollempne tunes everywhere familiarlie used, by William Slayter, printed by Robert Yong." Upon a copy of the later edition (1642), preserved in the British Museum, a former possessor has written the names of some of these tunes; for example, Psalm 6, to the tune of Jane Shore, Psalm 9 to Bara Forster's Dreame, Psalm 43 to Crimson Velvet, Psalm 47 to Garden Greene, &c. Shakspeare alludes to the Puritan practice of adapting religious words to secular melody in his "Winter's Tale:"

"There is but one puritan among them and he sings psalms to hornpipes."

After the Restoration, carols of the old kind became again popular, and from that time to the present the singing of carols at Christmas became steadily encouraged.

Warton supposes the religious carol to have been introduced by the Puritans, but this is a mistake, as a reference to Mr. Wright's collection, made for the Percy Society, will show.

The earliest printed collection was made by Wynkyn de Worde, 1521, but all these are of a convivial character.

Many of the old carols had scraps of Latin intermixed with English, as—

"Puer nobis natus est de Maria Virgine
Be glad lordynges, be the more or lesse
I bring you tydinges of gladnesse
As Gabriel me bereth witnesse."

Compare also "In dulci jubilo," in which Latin and German were used.

Carola (*It.*) A dance accompanied by singing, which grew into unenviable notoriety during the Republic of 1792 in France. *cf.* Carmagnole.

Cartellone (*It.*) The prospectus of an operatic season.

Carnyx (*Gk.*) An ancient Greek trumpet of a shrill tone, known afterwards to the Celts and Gauls. κάρνυξ (*Gk.*)

Cassagrande (*It.*) The big drum.

Cassatio. [Gassatio.]

Castagnette (*It.*) **Castagnettes** (*Fr.*)
Castañuelas (*Sp.*) Castanets.

Castanets. A musical instrument of percussion introduced into Spain by the Moors. The castanets were originally dried chestnut husks, from whence their name is derived, but were afterwards made of hard wood, by which means the tone was rendered more defined. The ancient κρόταλον, was a species of castanet (knicky-knackers). When a guitar is available it is used in addition to castanets as an accompaniment to national Spanish dances; those who have no castanets join in the rhythm by clapping their hands. [Bones.]

Catch. A species of canon or round for three or four voices, in which the words are so contrived that by the union of the voices a different meaning is given by the singers *catching* at each other's words. Poems of a trivial character, similar in style to nursery-rhyme doggrels, were also called catches. For example, there is a poem by "the learned clarke, Lewis Wager," printed in 1567," beginning:

"I have a pretty titmouse
Come pecking on my toe;"

and one of John Lyly's songs from "Endymion," 1591, is distinguished by the title of "a catch." The musical catch originated about the early part of the 17th century, the first collection of catches being made by Ravenscroft in 1609, under the title of "Pammelia, Musicks miscellanie, or mixed varieties of pleasant Roundelays and delightfull Catches of 3, 4, 5, 6, 7, 8, 9, 10 parts in one; none so ordinarie as musicall; none so musicall as not to be all very pleasing and acceptable." These, and others contained in later publications are little else than rounds, without the humour, so called, of the catch as it was afterwards accepted. William Jackson, of Exeter, says that "they are three parts obscenity and one part music. If they are not indecent, they are nothing. There is no particular object in them, but they are a species of musical false wit." Of the few catches which may be yet sung in a mixed company, "Would you know my Celia's charms?" by S. Webbe, and "Have you read Sir John Hawkins's History?" and "Ah! how Sophia," by Callcott, are the most favourable specimens. The words of the first are as follows:

"Would you know my Celia's charms,
Which now excite my fierce alarms?
I'm sure she has fortitude and truth
To gain the heart of every youth.

She's only thirty lovers now,
The rest are gone I can't tell how;
No longer Celia ought to strive,
For certainly she's fifty-five."

The humour of this catch consists in the emphasis placed upon the words *fortitude*, *thirty*, and *fifty-five*, by which it appears to the hearers that each singer is contending in

Digitized by Google

turn to uphold his notion of the age of the lady. In the second, the words "Burney's History" are made to sound like "burn his history," and in the third, one voice cries, "a house a fire," another, "go fetch the engines," while one apparently indifferent exclaims, "I'm but a lodger," from the following words:

> 'Ah! how, Sophia, could you leave
> Your lover, and of hope bereave,
> Go fetch the Indian's borrowed plume,
> But richer far than that you bloom.
> I'm but a lodger in her heart,
> Where more than me I fear have part."

There were formerly a number of clubs supported for the purpose of encouraging the production and performance of this species of musical trifle, only one or two of which are at present in existence, a better feeling having diverted the main object of these societies into the encouragement and execution of glees, part-songs, &c. [Round, canon.]

Catena di trilli (*It*). A chain, or succession, of short vocal or instrumental shakes.

Catgut. Boyau (*Fr.*), Minugia (*It.*), Darm (*Ger.*) The name given to the material of which the strings of many musical instruments are formed; it is made from the intestines of the sheep, and sometimes from those of the horse, but never from those of the cat.

Cathedral Music. A term applied to that music which has been composed to suit the form of service used in our cathedrals since the Reformation. It includes settings of canticles and also of anthems. The first writers of this class of music were Marbecke, Tallis, Tye, and Byrd, and the works of the two last named especially illustrate the state of cathedral music at the period in which they lived, for they employed Latin and English words to the same music, so that it might be available whether the service was according to the ancient or reformed usage. The style of the earliest cathedral music was formed on the model of the Italian motets and other sacred compositions, and with the exception of a difference in the words was identical with the secular music of the period.

It was feared that the Commissioners appointed by the Statute 27 Henry VIII. to compile a body of ecclesiastical laws "as should in future be observed throughout the realm," taking into consideration the abuse of music in the Church, would forbid its use altogether. As the King was fond of music they deemed it politic to retain it for the service of the Church, but they implied a return to simple forms, in directing certain parts of the service to be sung by the "ministers and clarkes" in a plain, distinct, and audible manner. The rubric of the First Book of Edward VI. prescribes the saying or *singing* of "mattens and evensong;" and in the ministration of the Communion that the clerks shall sing in English for the office or "Introite as it is called," a psalm appointed for that day. And again it directs that the clerks shall sing one or many of the sentences therein mentioned, according to the length and shortness of the time that the people be offering. In John Marbecke's "Booke of Common Praier noted" 1550, it will be seen that the whole of the service was sung either to some general kind of recitation or intonation with small inflections, to an adaptation of the ancient cantus or *accentus ecclesiasticus*, or to some modification of the old use by Marbecke himself. Queen Elizabeth in her injunctions concerning the clergy and laity of this realm, published in the first year of her reign, 1559, desired the "continuance of syngynge in the Churche" and "that there bee a modeste and destyncte song so used in all parts of the Common Prayers in the Churche, that the same may be as playnely understanded as if it were read without syngynge."

Notwithstanding this injunction the use of singing and of organs in the Church was only maintained by a majority of one in the Lower House of Convocation, a strong objection to Cathedral music existing even in that early period. After Marbecke's book, which has music in one part only, John Day (1560) published a service in four parts, adding five years later, those "offices" which had been omitted in the former collection. These preserved to a certain extent the prescribed or adopted "use" in those parts of the service which were always intended to be performed simply, the publication also indicated the places where a more elaborate musical setting might be allowed, and composers taking advantage of the licence wrote original music for the Venite, Te Deum, Benedictus, Jubilate, Communion Service, the Canticles used at evensong, and "such godly praiers and psalmes in the like form to the honor and praise of God,"—"so they may be songe as anthems." The use of organs and singing in the Church was nevertheless a sore grievance to the Puritans; they did not, however, object to metrical psalms, and employed them whenever and wherever possible; but the cathedrals always objected to their introduction, as not being cathedral music properly so called; it is within the last half-century that hymns or psalm tunes have been sung in cathedrals as an integral part of the ordinary service. In the time of the Commonwealth metrical psalms were the only things sung in the churches, but they were also sung at other times, and it was not until the Restoration of Charles II. that Cathedral service was resumed, this time in a considerably altered form. The Communion Service or Mass, in

Digitized by Google

times past held to be the most important act of worship, was placed in the back-ground, and was, when celebrated, given without the aid of music. Church composers did not take the trouble to set those parts of the service which were never performed, and consequently there is not a single "Gloria in excelsis" produced by any cathedral writer between 1660 and 1840, other than as an anthem. When so set, it was considered allowable to omit some sentences and add others at discretion, so that it would be scarcely available for the Communion Service. The "Sanctus" was set to music, as it became the habit to sing it in the place of the Introit, a fashion not yet dead in many cathedrals. When the Communion Service was restored to its true importance about twenty-five years since, adaptations of Marbecke's arrangements were freely and properly used, until a new generation of composers employed their talents to supply the deficiency. At the time of the Restoration, the character of cathedral music also underwent a change. The influence of the French school may be traced in the writings of Purcell, Humphries, Blow, Wise, Weldon, and others. It is not a little strange that while most temporary influences can be seen in the various periods of cathedral music, there are few instances of any church composers copying Handel's style, and none in which it was done with success. Dr. Greene, his contemporary, has a special character of his own, Dr. Boyce has also *his* individuality, and the elder Hayes shows no leaning towards the great oratorio writer. Handel's oratorios, though not written for the cathedral, are often laid under contribution, whereas the anthems composed by him for the service of the Church are comparatively neglected and unknown.

At the latter part of the last century and the beginning of the present, cathedral music was at its weakest point; adaptations, arrangements, florid melodies, with paltry accompaniments, chants of a gay and undignified style, and all music used in the service, showing the influence of a general indifference and carelessness, which, to a certain extent, still exists, though happily in no strong degree, for a more reverent feeling abounds and is nourished. Cathedral music, like every other branch of art, should increase and be progressive, should take advantage of every new discovery or admitted truth in music. All styles should be fairly represented, and no one style should be considered as indicative of special doctrinal views. There are few who seriously object to a building in which successive styles are seen, but on the contrary think that all that is good should be retained. The many who have spoken in music in past ages should have their sayings preserved when they are worthy of being kept, but it would be folly to insist upon the retention of all that could be gathered of the works of a writer, because he has said one happy and lasting thing. It is not given to men to be wise at all times, and the best of cathedral musicians have written unworthy stuff. Taste and good sense, free from prejudice, will guide to a proper and useful selection, so that cathedral music for ordinary purposes may include the thoughts uttered under all influences in many ages.

The small number of voices considered sufficient for the usual services of our cathedrals is a bar to grand effects. This has been felt by composers, who have been compelled so to arrange their music that it may produce adequate effects from the usual small choirs. Probably with a prophetic view of the future augmentation of the musical staff of a cathedral, many modern composers have so constructed their works, that while they are not ineffective with a small body, they are nobly grand when given by increased numbers. So that there is reason to believe that in the days of the future, when cathedral choirs shall be in numbers and skill worthy of the service to which they minister, cathedral composers will be equal to the task of writing music suitable to the time and place. The grand effect produced by a large body of voices in a cathedral during the performance of an oratorio upon the occasion of a festival is never without some influence in turning men's minds to higher things. Music is the handmaid of religion, and there can be no reasonable objection to the introduction of oratorios and other extensive sacred compositions, with all the effects that a trained choir and orchestra can produce, provided always, that such performances are made an integral part of an act of worship. In the metropolis such performances have been given with the most satisfactory results at stated times, and the day may not be very far distant, when they may be made of more frequent occurrence, and so, our cathedrals, by calling into requisition all musical talent, inventive or executive, will become again what they once were, the nurseries and centres of musical culture and knowledge.

Catlings. The smallest sized lute-strings.

Cauda (*Lat.*) The tail of a note.

Cavaletta (*It.*) [Cabaletta.]

Cavaletto (*It.*) (1) A little bridge. (2) The break in the voice.

Cavalquet (*Fr.*) A trumpet-signal to cavalry.

Cavata (*It.*) [Cavatina.]

Cavatina (*It.*) A melody of a more simple form than the *aria*. A song without a second part and a "*Da capo*." The term is, however, applied with less strictness to airs of other kinds. (See "Be thou faithful," in Mendelssohn's "St. Paul," and "Salve dimora," in Gounod's "Faust," &c.)

C.B. Abbreviation for Contra-basso.

C barré (*Fr.*) The term for the time indicator C, with a dash through it, ₵.

C clef. The clef showing the position of middle C, in which are written the alto, tenor, and (in old music) other parts.

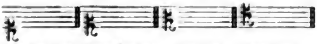

Clef.]

C dur (*Ger.*) C major.

Cebell. The name of an air or theme in common time of four bar phrases, forming a subject upon which to execute "divisions" upon the lute or violin. This style of air, although frequently found in books for the violin in the 17th century, is now obsolete; its principal feature was the alternation of grave and acute notes which formed the several strains. The following are examples:

THO. MACE, 1676.

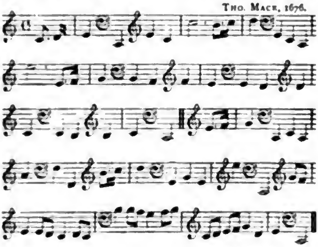

H. PURCELL, b. 1658.

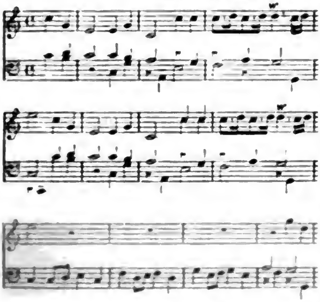

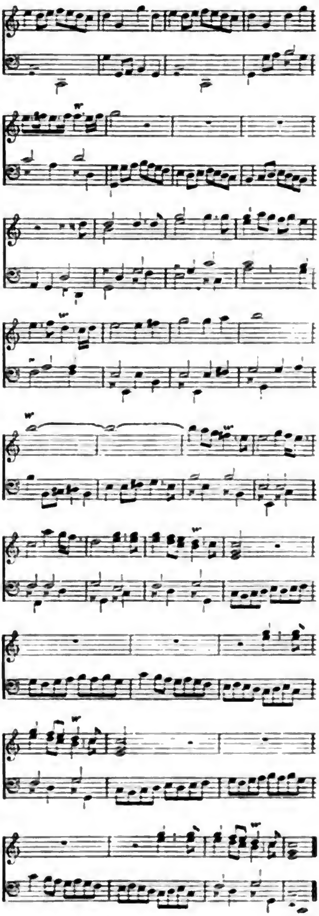

Celere (*It.*) Quick, swift.

Celerità, con (*It.*) With speed, haste. Quickly.

Céleste (*Fr.*) A direction for the use of the soft pedal.

Céleste, voix (*Fr.*) A stop on the organ or harmonium. [Vox Angelica.]

Celeusma (*Gk.*) *κέλευσμα*, or *κέλευμα* (from *κελεύω*, to urge on, to command). The word or sing-song of the *κελευστής* (fugle-man or leader), by which oarsmen were encouraged to row rhythmically, and by which, to this day, sailors pull uniformly and simultaneously at a rope.

Celli. *Abb.* of violoncelli.

Cello. *Abb.* of violoncello.

Cembalista (*It.*) A pianoforte player.

Cembalo. Clave-cembalo, cimbalo. A harpsichord. [Pianoforte.]

Cembanella (*It.*) [Cennamella.]

Cennamella (*It.*) A pipe, or flute.

Cento (*Lat.*) (In Greek *κέντρων*.)
Centone (*It.*) Patch-work. A musical work made up of extracts from an author's compositions, as a *cento* was from an author's poems. *cf.* pasticcio.

Cercar la nota (*It.*) To feel for a note, to reach it by slurring.

Cervalet or **Cervelat** (perhaps *dim.* of cervus, signifying a little stag-horn). An ancient wind-instrument of a small size, from which, by means of a reed, tones similar in character to those of the bassoon could be produced.

Ces (*Ger.*) C flat.

Cetera (*It.*) A citara or guitar.

Chacona (*Sp.*)
Ciaccona (*It.*)
Chaconne (*Fr.*) A slow dance in ¾ time, frequently constructed upon a ground bass, and sometimes formerly introduced as a movement of a sonata. [Chica.]

It is usually stated that the *chaconne* is in the major mode, and that the *passacaille*, which is somewhat similar to it in rhythm, is in the minor. This is not the case, as the following theme, on which Bach's celebrated *ciaccona* for violin solo is founded, will show:

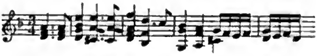

Chair organ. A name given to the Prestant or choir-organ, from a notion that it formed the seat of the organist, when placed behind him.

Chalumeau (*Fr.*) Stem, or straw-pipe, from the Latin *calamus*, a reed. The lower register of the clarinet and the basset-horn is called the *chalumeau* tone, from the obsolete instrument *shawm*, *schalmey*, precursor of the oboe and clarinet.

Chamber music. Kammermusik (*Ger.*) **Musica di camera** (*It.*) Vocal or instrumental compositions suitable for performance in a chamber, as opposed to a concert-room. The performance in private upon single instruments of any class constituted the first chamber music properly so-called. Strictly speaking, any music vocal or instrumental played in private is chamber music; but the term is now applied not only to performances upon a single instrument, with or without accompaniment, but also to any combination of different instruments, with only one player to each part—duets, trios, quartetts, &c., for voices or instruments. It is probable that the first chamber music constructed as such was entirely vocal, and not of much earlier date than the end of the 15th or the beginning of the 16th centuries—the Scolia of the Greeks, the music of the minstrels, and of public and private musicians of later date, including among the former "the waits," "noises," and other private bands, not being of a character that could fairly be called by the title chamber music. Therefore the Madrigal will be regarded as among the first specimens of chamber music. The titles of more than one collection, for example, "Madrigali di Tavolina," "Madrigali di Camera," "Madrigali Concertati," "Madrigali et Arie per sonare et cantare," and so forth, together with the peculiar style in which many of the early books are printed,—two parts on one page intended to be read by two persons seated opposite to each other at the same table—would show conclusively that they were intended as chamber music.

The addition of instrumental accompaniments to madrigals probably arose out of a desire to support the voices and keep them in tune, as well as to give employment to those who could play and not sing, but who were desirous of taking part in that which was going on. This practice—at first a mere conciliation to the instrumentalists—suggested the use of instruments alone for the purposes of concert. Thus we find attached to the early productions, instructions to the effect that they are "apt for Instrumentes and Voyces," as in Alison's "An Howres Recreation in Musicke," or as in Bonaffino's "Madrigali," that they are available "per cantar e sonar nel Clave cimbalo, Chitarrone o altro simile Instrumento," or as in the later editions of Byrd's Psalmes, Songs, and Sonnets, framed to be "fit for Voyces or Viols."

Doubtless from such small beginnings the writers of the time were induced to compose "Consort lessons," "Ayres," "Fancies," "Canzone da Sonare," and the like, often written in six parts, the number of viols in a "chest." These compositions at first differed very little in point of form and treatment from the madrigals from whence they were derived, until the demand arose for pieces of less dignity, in obedience to which demand we find dance tunes, "Almaines, Ayres, Corants, Sarabands, Moriscoes, Jiggs, &c.," hitherto only set for a single instrument, arranged in parts for "Viols or Violins;" and these and other dance-tunes issued in *suites* made into the first sonatas, and the symmetrical shape in which each was necessarily written for the purposes of the dance gave rise to that which is known as Form. The word Sonata, at first applied to pieces for a solo instrument, as well as to those for several, became gradually to be used as a term for compositions of a certain character for a single compound instrument, as the organ, harpsichord, or pianoforte.

The most important era in the history of chamber music was the final quarter of the last and the first of the present centuries; the labours of Boccherini, whose trios, quartetts, and quintetts are form-like, easy, and graceful, as well as those of Fiorillo, Giardini, Pugnani, and Viotti, leading to the foundation of the school in which Pleyel, Haydn, Mozart, and Beethoven were such apt pupils and masters.

Changeable chant. A single or double chant which can be sung either in the major or minor mode without other alteration than the substitution of the minor third and sixth of the scale for those of the corresponding major.

W. TURNER.

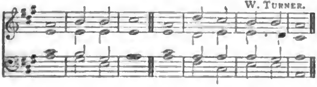

W. TURNER.

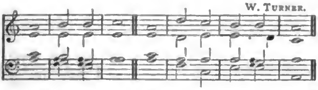

Change of voice. [Larynx.]

Change ringing. [Bells.]

Changer de jeu (*Fr.*) To alter the stops on an organ or harmonium.

Changes. The altered melodies produced by varying the sounds of a peal of bells. [Bells.]

Changing notes. Passing notes or discords which occur on the accented parts of a bar.

Chanson (*Fr.*) (1) A song. (2) A national melody. (3) A part-song.

Chansonnette (*Fr.*) A little song.

Chant. A short musical composition to which the Canticles and the prose version of the Psalms are sung, either in unison or in four-part harmony. There are two kinds of chant in common use—the Anglican and the Gregorian.

(1) A Gregorian chant consists of five parts; the intonation; the first reciting note or dominant; the mediation; the second reciting note or dominant; the ending, *e.g.*:

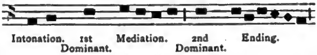

The intonation is used generally to every verse of a canticle, but only to the first verse of a psalm, unless a special psalm be used on a solemn occasion, as for instance the Miserere (Psalm li.) during Lent.

With regard to the pointing of the Prayer-book version of the Psalms, several important facts have to be considered. The undoubted object of the chants as originally used in the Roman Church was to enable, as far as possible, a pure *syllabic* recitation of the words, so many of the words of a verse being recited on the dominants as would leave one syllable only to each note of the mediation and ending. As these chants were in use many centuries before the invention of *cantus mensurabilis*, it is quite impossible that they were ever sung rhythmically at the close of each recitation. But there is a growing tendency to treat the Gregorian chants Anglican-wise, and either by accents or bars, to definitely shape out their rhythm. If any proof were wanting of this fact, it is only necessary to give the following:

Ex. 1.

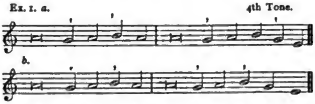

The ending of the first of these is evidently intended to be in triple measure, that of the second in duple. But to force the ending into either one of these measures is to wilfully cast aside the invaluable property it possesses of bearing an accent on any note, as the words require. A like desire for modern chant-form has led to the following differences of accent:

Ex. 2.

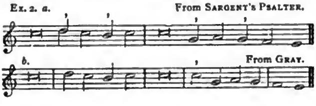

Ex. 3.

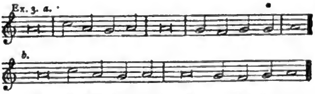

It will be seen from the above, that in proportion to the adoption of strict time in the ending, the true use of chants for syllabic treatment becomes lost. In short, "Gregorians," as used for the most part in England at the present time, are nothing more than ordinary chants, not, however, having an uniform number of bars of music. Hence the same difficulties present themselves which will be explained below in the account of Anglican pointing, *e.g.*:

Ex. 4. *a.*

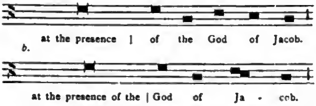
b.
at the presence | of the God of Jacob.
at the presence of the | God of Ja - cob.

It is generally understood that when the number of notes exceeds the number of syllables, the notes not required may be omitted, *e.g.*:

Ex. 5.

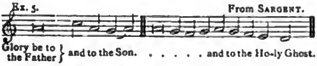
From Sargent.

This is a negative proof of the original syllabic tendency of Gregorian pointing, it being merely a corollary of the law that there should be "one syllable to one note," to say, "if only a few syllables are left, let the notes not wanted be omitted." This rule is now generally neglected; and, even in canticles with such short verses as the Te Deum, the syllables are *slurred* to the superfluous notes, lest the hearers' notion of "the tune" should be disturbed.

In 1843, the Rev. F. Oakeley published his Gregorian Psalter, carrying out, in its integrity, the principle of the syllabic system, *e.g.*:

Ex. 6.

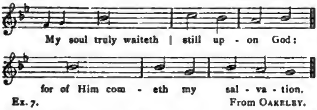
From Oakeley.
My soul truly waiteth | still up - on God:
for of Him com - eth my sal - va - tion.

Ex. 7.

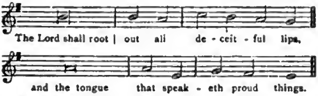
From Oakeley.
The Lord shall root | out all de - ceit - ful lips,
and the tongue that speak - eth proud things.

Notwithstanding its merits, this Psalter seems not to have been largely used, and where used has been superseded by others in which the tones have been "anglicanized" and made more palatable by the unjustifiable introduction of fixed accent and rhythm.

Another danger which presents itself to the advocates of Gregorian chants is their limited number. It is out of the question that new Gregorian chants should be "expressly composed" for Psalters, but it is absolutely necessary to provide a variety of chants to avoid the monotony of over-repetition. Hence it is, (1) that *endings* heard on the Continent, whatever be their modern growth or their incompatibility with the Gregorian scale, are greedily seized and made use of in this country; and (2) that *foliations* or the vicious introduction of auxiliary notes, above or below the genuine notes of the chant, are as readily welcomed by Gregorian editors, *e.g.*:

Ex. 8.

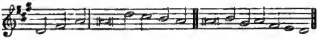

Ex. 9.

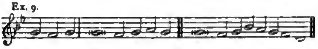

Ex. 10.

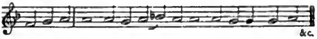
&c.

Ex. 8 is called a form of the 5th tone, No. 9 a form of the 4th.

Ex. 10 shows a foliated form of the mediation of the 1st tone. The division of the plain tones and foliated tones into ferial and festival does credit to the ingenuity, but not to the historical integrity, of Psalm-pointers.

The French and Belgians have ever been celebrated as clever adulterators of plain-song; and as their manuals now form the chief textbooks of English Gregorianizers, it is not difficult to prophecy a general decadence of the art of Gregorian chanting in this country.

(2) An Anglican chant is of two sorts, single and double. A single chant is in two strains, the first of three, and the second of four bars in length·

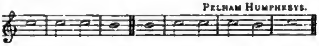
Pelham Humphreys.

A double chant has the length of two single ones:*

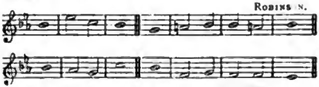
Robinson.

---

* It has been stated, and the statement is often repeated, that the double chant was suggested by the accidental performance of two single chants in succession by a nameless pupil of Hine, who was organist of Gloucester Cathedral between the years 1710 and 1730. In "Boyce's Cathedral Music," published 1760—1778, is a double-chant by John Robinson, who was organist of Westminster Abbey from 1727 to 1762, dying at the ripe age of eighty; and as there exists a MS. copy of the same chant in the handwriting ot Dr. Turner, the father-in-law of Robinson, with the date 1706, in one of the old MS. service-books belonging to St. Paul's Cathedral, there is reason for questioning the story concerning the accepted origin of the double chant.

The two strains are also called *halves;* one half is sung to that part of a verse of the Prayer-book version of the Psalms on each side of the colon, whether the number of words be many or few, whether the sentence is complete or not; as:

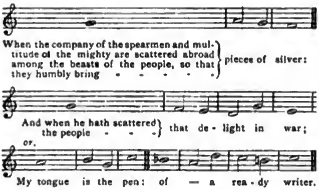
When the company of the spearmen and multitude of the mighty are scattered abroad among the beasts of the people, so that they humbly bring - - - - -
pieces of silver:
And when he hath scattered the people - - -
that de - light in war;
or.
My tongue is the pen: of — a rea - dy writer.

The opening chord of a chant, and also the first chord after each double bar, may be sustained at will, to accommodate the number of syllables contained in each part of the verse. These chords are called *reciting notes*, those which follow are called the *inflections;* or, according to some, the first half of the chant is the *mediation*, and the second the *cadence*. The fitting of the words to the music is called *pointing*. The pointing of the Psalms and Canticles is a matter concerning which there are diversities of opinion. The principal object to be aimed at in pointing is "the apportioning out of the emphasis of the words to be sung, after the manner that an eloquent speaker would recite them;" but as sentences are capable of as many accents more or less sensible as there are words, the diversity of opinion on the subject is not to be wondered at. The words are divided in the Prayer-book not always in the best manner as regards their complete grammatical sense; and as it is at present deemed unwise to adopt any plan but the one therein suggested, difference of opinion will exist until a change is made in its system of stops. The varieties of pointing arise from the desire to unite an oratorical with a musical accent; and the many ways in which this is attempted will be best seen by the following quotations from pointed Psalters in frequent use.

[The lines after each set of words indicate the place of the bar in the chant.]

Psalm cxxxvii.

No. 1. Dr. Wesley's Psalter.

For they that led us
away captive required
of us then a sóng and | melody in our | heaviness.

No. 2. The Cathedral Psalter.

For they that led us
away captive required
of us then a song and
mélody | in our | heaviness.

No. 3. The Congregational Psalter.

For they that led us
away captive required
of us then a song and
melodỳ in | our | heaviness.

No. 4. Monk and Ouseley's Psalter.

For they that led us
away captive required
of us then a song and
melody in our | heavi- | -ness.

The *form* of the chant has been the real cause of the difficulties of pointing. An ordinary melodic sentence consists of two, four, or eight bars, but the chant has first three, then four bars. This peculiarity does not, however, offend the ear so much as the eye, for in reciting, the rhythmical cadence is to a certain extent completed.

Various theories have been put forth to account for the 7-bar or twice 7-bar form of the Anglican Chant, all writers being agreed that a 7-bar phrase is not actually presented to the ear in the process of chanting. The theorists may be divided into two classes—those who would add a bar to the commencement of the chant, that is, to the reciting note; and those who would add a bar at the half cadence and whole cadence. The following is the method in which the former would write out Robinson's Chant:

Ex. 1.

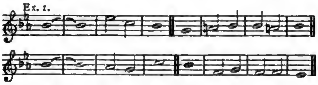

Those who lean to the latter opinion would write it thus:

Ex. 2

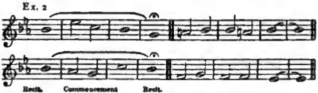

In opposition to the first view taken, it may be urged that in music the chords of cadence *precede* the final chord, and in Ex. 1, they fall on the final accent, as is shown by doubling the bars, *e.g.*:

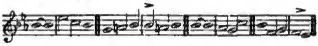

Except in rare forms of dance-tunes the above rhythm would be unbearable. In favour of the second form (Ex. 2), it may be stated that in all the best pointed Psalters an accent, a larger fount of type, or a bar, marks the close of the Recitation and commencement of the musical rhythm, and that the syllable or syllables so made prominent only occupy *one* bar of time. In opposition to the second form, it may be remarked that the final bar of both halves of a double chant is not *in practice* held out for the length of two bars. One or the other of these theories may be true, and the reader is left to decide on their respective merits.

Certain writers have assumed that the Anglican Chant is a highly tractable collection of sounds, bound by no laws of rhythm; and acting on this notion, have attempted to *unbar* some modern chants. This view has led to a system of pointing by which as many words as possible are collected on the reciting note, *e.g.*:

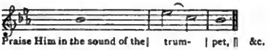

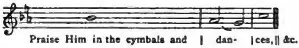

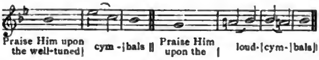
Praise Him upon the well-tuned | cym - | bals || Praise Him upon the | loud- | cym- | bals ||

The above system (known as the "Sudbury") is said to be smooth, but the number of *slurs* involved would produce this effect, although opposed to the true principles of chanting.

It has been said "that the best practical solution of the difficulty of chanting would be offered by selecting a set of the most appropriate chants, whose melodies, within the range of all voices, would not suffer by being sung by a whole congregation, and to have every word set to a note of relative length, so as to ensure evenness of tone and accuracy of accent." Some of the early church composers have left examples of the Venite set to distinct music, often chant-like, so that the thing here suggested would not be so great a novelty. But it would be difficult to make such a plan general, for, leaving out of the question the additional time such a service would occupy, none but educated choirs could perform it, and the ordinary chant is so easy that there is little if any trouble needed to teach it to unskilful choristers. The chant at present in use might be retained, and if elocution is the main object of chanting, a different system of pointing might be devised, by employing the present authorised division of the verses only when convenient. Alterations might be made in a verse (1) when the sense is incomplete in it, (2) when a verse contains two distinct subjects, (3) when the present colon interrupts the logical sequence, *e.g.*:

(1) Psalm xvii.

8. Keep me as the apple of an eye: hide me under the shadow of thy wings, (9) from the ungodly that trouble me.

(2) Psalm lxxxix.

49. Remember Lord the rebukes that thy servants have: and how I do bear in my bosom the rebukes of many people; (50) wherewith thine enemies have blasphemed thee, and slandered the footsteps of thine anointed.

Praised be the Lord for evermore: Amen and Amen.

(3) Psalm xiv.

11. Who shall give salvation unto Israel out of Sion? (:) when the Lord turneth the captivity of his people, then shall Jacob rejoice, and Israel shall be glad.

In some of the numerous editions (issued between the years 1655 and 1730) of Playford's "Introduction to the Skill of Musick," there is an appendix containing the "order of singing the Divine Service in Cathedrals" In these it is said that "the Venite is begun by one of the choir, then sung by sides, observing to make the like break or close in the middle of every verse, according as it is shorter or longer." The use or tune for each day in the week is given to the first verse of the Venite, and these tunes are such as are now called Gregorian. There are two others —"Canterbury tune" and "Imperial tune" —"proper for Choir; to sing the Psalms, Te Deum, Benedictus, or Jubilate, to the organ or sometime without it." The manner in which the words are disposed will be seen by the following copy of the first-named of these tunes:

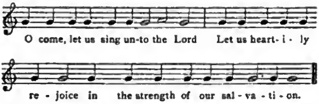
O come, let us sing un-to the Lord Let us heart- i - ly
re - joice in the strength of our sal - va - ti - on.

Dr. Turner, 1706, gives the pointing of the same verse as follows:

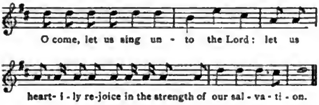
O come, let us sing un - to the Lord: let us
heart- i - ly re-joice in the strength of our sal - va - ti - on.

It will be seen that this chant contains only five complete bars of four crotchets each. The bars in it do not indicate the place of accent, or even the best division of the chant for the purpose of pointing. The earliest printed copy of a now well-known chant by the same author, is given in the following form (from "Fifty double and single chants being the most Favourite as performed at St. Paul's, Westminster, and most of the Cathedrals in England. London: Printed for C. and S. Thompson, at No. 75, St. Paul's Churchyard").

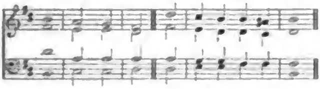

In the following example (from the "Harmony of Sion") the chant is compressed into four bars.

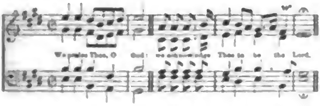

And Dr. Boyce, in his "Cathedral Music," writes the Venite to the chant ascribed to Tallis thus:

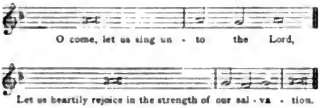
O come, let us sing un - to the Lord,
Let us heartily rejoice in the strength of our sal - va - tion.

The method of chanting the Psalms adopted in the present day is for each of the two divisions of the choir to sing a verse alternately. In some places where a double chant is used the whole of the chant is sung to two verses by each side in turn. In one cathedral (Oxford) each side sings one half of a verse only. In consequence of this custom of the alternate chanting of the Psalms, it is supposed that wherever alternate singing is mentioned in ancient records, chanting is meant. The objections entertained against chanting by the followers of Wickliffe, and in later time by those of Calvin, were expressed in violent terms, not altogether necessary to repeat here. It is true these reformers approved of the people joining "with one voice in a plain tune, but not of tossing the Psalms from one side to the other with intermingling of organs." Less moderate in their deeds than in their words, the Puritans, when in power during the Commonwealth, destroyed all organs and every music-book they could lay their hands upon. Metrical psalmody supplied the place of chanting, the Psalms were rarely if ever sung, and so, in contradistinction to Sternhold and Hopkins, or Tate and Brady, were called the "Reading Psalms," a practice continued as lately as the year 1873, when the Psalms for the day sung at the meeting of the Charity Children in St. Paul's Cathedral were so called. Chanting was regarded as essentially Popish, and alternate singing an abomination even among church people. It was never heard in parish churches in the last century, the cathedrals alone retaining the traditional use. By degrees a change was effected; the unclean thing was handled without any alarming effect, and even Dissenters changed their opinions upon the subject. Instead of condemning chanting, they adopted it. Dr. Channing thought it "the most purely Protestant music;" Mr. Newman Hall considered it "a homage to the Bible, calculated to make the Word of God better known, appreciated, and loved;" others would "by no means have it banished;" and the preface to the fourth edition of a little book called "Euphonia" (1870), designed to familiarise Non-conformists with the principles of chanting, states that "the objections entertained by many to the ancient practice of chanting having been much diminished, there is now a growing feeling in favour of singing portions of the Bible in the very words of Scripture, rather than through the medium of metrical versions exclusively." This book, which is historically valuable, contains one hundred portions of Scripture pointed for chanting, together with a selection of familiar, if not good, Anglican chants; the principle guiding the choice being liveliness of melody and general tunefulness. These facts tend to show that the usefulness of chanting is in process of general recognition by "all who profess and call themselves Christians;" that it is no longer held to be the type of a peculiarity of religious opinion; that there is a mutual interchange of means towards a spiritual end; that whereas one side does not disdain to encourage the use of metrical psalmody in its services, the other borrows chanting, defending it as "a simple but impressive mode of worship."

Chants have been found convenient means of rendering hymns of irregular metre, or any hymn of which a simple musical treatment is required.

Chant (*Fr.*) (1) Song, tune. (2) The voice part or melody.

Chantant (*Fr.*) Singing, musical, as *café chantant*, a musical coffee-house.

Chant en ison, or **chant égal** (*Fr.*) (1) The name of a species of chant, consisting of two sounds only, which was adopted by many of the old religious orders. (2) Monotone.

Chanter. (1) A name given to the singing priest on duty. (2) A lay vicar.

Chanter à livre ouvert (*Fr.*) To sing at sight.

Chanterelle (*Fr.*) (1) The first or highest string upon instruments played with a bow. The E string of the violin, and the A of the viola and violoncello. (2) The highest string of a guitar or lute.

Chanterres (*Fr.*) A name given to ballad or poem singers in mediæval times, originally applied to the Provençal Cantadours.

Chanteur (*Fr.*) A male singer.
Chanteuse (*Fr.*) A female singer.

Chant Grégorien (*Fr.*) Plain song.

Chanties. Songs sung by sailors when at work, often set to "nonsense" words.

Chant pastoral (*Fr.*) A shepherd's song, or melody in imitation of one.

Chantries. Endowed foundations in the Romish Church, instituted for the due performances of requiem masses for the repose of the soul of the founder and his family. Chantries were attached to existing parish churches, or more frequently to monastic establishments and cathedrals. At the Reformation the practice of soul-masses and the chantries became disused, and their revenues absorbed.

Chant-royal (*Fr.*) A certain form of early French poetry set to music. Pasquier describes it as a song in honour of God, the Virgin, or the saints, or any other "argument of dignity, especially if coupled with distress." The *chant-royal* was written in heroic stanzas, and closed with a *L'envoy* or stanza containing a dedication, recapitulation, or moral.

Chantry priest. A chaplain or singing priest attached to a chantry, whose duty it was to sing masses for the dead.

Chant sur le livre (*Fr.*) A system of descant by which the part sung by one voice, as written in the open book, could be accompanied by another voice in counterpoint, more or less free, according to the movement of the *canto fermo*, which was sung generally by a bass voice, the *déchant* being taken by a tenor or other high voice. It was necessary that the singer of the *canto fermo* or plainsong should render it *à la rigueur*, that is, should not make those slight changes of the length of the notes which would be justifiable and usual when singing alone; nor could he hold out the rests (*tenere punctum*) as he otherwise would, lest the *déchanteur* should be upset in his calculations. The full rules of this system are to be found in early treatises. It was called in Italy *contrappunto di mente*, or *alla mente*.

Chapeau Chinois (*Fr.*) A set of small bells arranged in the form of a Chinese hat. *Pavillon chinois*.

Characteristischer Ton (*Ger.*) The leading note. (*Fr.*) *Note sensible*.

Characters. A general name for the signs employed in music, such as brace, bind, bar, sharp, flat, natural, clef, stave, shake, turn, beat, and the signs of words indicating time and expression, *e.g.* < > C ₵, &c.

Characterstücke (*Ger.*) Pieces of music written with the intention of describing certain impressions by means of sound. Beethoven's Pastoral Symphony, Mendelssohn's Reformation Symphony, and the overture and music to "A Midsummer Night's Dream," are specimens of this style of composition.

Charivari (*Fr.*) Mock music, clatter.

Chasse (*Fr.*) Hunting; *à la chasse*, in the hunting style.

Chatzozerah, Khatsotsrah (*Heb.*) This is generally thought to have been a straight trumpet, with a bell or "pavillon" as it is termed. Moses received specific directions as to making them. "Make thee two trumpets of silver; of a whole piece shalt thou make them: that thou mayest use them for the calling of the assembly, and for the journeying of the camps." In Ps. xcviii. 6, the *chatzozerah* and *shophar* are brought into juxtaposition: "With *chatzozerah* and sound of *shophar* make a joyful noise before the Lord the King:" or, as it incorrectly stands in the Prayer-book version, "With trumpets also and shawms, &c." In this passage the Septuagint has it, Ἐν σάλπιγξιν ἐλαταῖς, και φωνῇ σάλπιγγος κερατίνης, "With ductile trumpets, and the sound of horn-trumpets." So, too, the Vulgate: "In tubis ductilibus et voce tubæ corneæ." The word *mikshah*, which is applied to the description of the *chatzozerah* in Num. x. 2, which means "rounded" or "turned," may either apply to a complete twist in the tube of the instrument, or, what is more probable, to the rounded outline of the *bell*. But if the former is the real interpretation of the epithet, it would make it more like a trombone, and similar in form to that depicted on the Arch of Titus. But, on the other hand, the account given by Josephus points out the latter characteristic of shape. He says, "Moses invented a kind of trumpet of silver; in length it was little less than a cubit, and it was somewhat thicker than a pipe; its opening was oblong, so as to permit blowing on it with the mouth; at the lower end it had the form of a bell, like a horn." It seems chiefly to have been brought into use in the Hebrew ritual, but was also occasionally a battle-call, and blown on other warlike occasions.

Check-action. [Pianoforte.]

Check-spring. A small spring added for the assistance of any weakness in the return of action in the mechanism of an organ.

Chef d'attaque (*Fr.*) The leader of an orchestra, or chorus.

Chef-d'œuvre (*Fr.*) The master-work of any composer.

Chef d'orchestre (*Fr.*) (1) The leader. (2) Conductor of an orchestra.

Digitized by Google

Chelidonizing (from the *Gk.* χελιδονίζω, to twitter like a swallow). Singing the swallow-song (χελιδόνισμα), a popular song sung by Rhodian boys in the month Boëdromian, on the return of the swallows, and made into an opportunity of begging. A similar song survives in modern Greece. A *crow* was also carried about by begging boys who sang; whence *Gk.* κορωνίζω. Examples of both songs are given by Athenæus. Pamphilicus of Alexandria, in his chapter on names, calls the men making collections for the crow, *coronistæ*, and their songs, *coronismata*. There was a similar custom in Ireland on St. Stephen's day. A number of young men carried a furze-bush on which a wren was tied, and stopping before the houses of the gentry, repeated the following lines:

> "The wren, the wren is the king of all birds,
> Was caught on St. Stephen's day in the furze,
> Although he's little, his family's great,
> Then pray, kind gentle folks, give him a treat."

In England and Scotland there are many customs of a like character, as for example, "going a gooding" on St. Thomas's day; singing the *Hagmena* on the three days preceding Christmas day; the children's May-day march, when they carry garlands of spring-flowers and boughs, and stopping at the doors of people of the better sort, sing a long song, one verse of which runs:

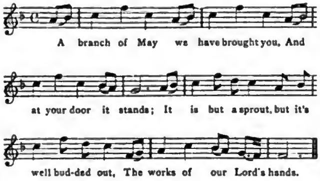
A branch of May we have brought you, And
at your door it stands; It is but a sprout, but it's
well bud-ded out, The works of our Lord's hands.

Chelys. (*Gk.*) χέλυς, lit. *a tortoise* (*Lat. testudo*). (1) The lyre of Mercury, supposed to have been formed by strings stretched across a tortoiseshell. (2) In the 16th and 17th centuries a bass-viol and division-viol were each called *chelys*.

Cheng. The Chinese organ, which consists of a series of tubes having free reeds. It is held in the hand and blown by the mouth. The introduction of this instrument into Europe led to the invention of the accordion and harmonium. Kratzenstein, an organ-builder of St. Petersburg, having become possessed of one, conceived the idea of applying the principle to organ-stops. The tone of free reeds is enforced by tubes, as in the cheng and in certain organ-stops, but the tubes can be dispensed with, as is the case in a harmonium.

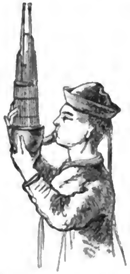

Cherubical hymn. The *ter sanctus*, or *trisagion* in the service of the Holy Communion, "Holy, holy, holy," &c.

Chest of viols. An expression signifying a set of instruments necessary for a "consort of viols." They were six in number, namely two trebles, two tenors, and two basses. A chest of viols, with a harpsichord or organ, with an occasional hautboy or flageolet, formed an ordinary orchestra in the early part of the 17th century.

Chevalet (*Fr.*) The bridge of a stringed instrument.

Cheville (*Fr.*) A peg for a violin, guitar, lute, &c.

Chevroter (*Fr.*) To skip, quiver, to sing with uncertain tone, after the manner of goats *Alla vibrato*.

Chiara (*It.*) Clear, distinct, pure, *e.g.*, *chiara voce*, clear voice; *chiara quarta*, a perfect fourth.

Chiaramente (*It.*) Clearly, purely, distinctly.

Chiarezza, con (*It.*) With brightness, clearness.

Chiarina (*It.*) A clarion or trumpet.

Chiave (*It.*) (1) Key or clef. (2) A failure. [Fiasco.]

Chica. The name of a dance popular among the Spaniards and the South American settlers descended from them. It is said to have been introduced by the Moors, and to have been the origin of the Fandango, which some writers declare to be the Chica under a more decent form. It is of a similar character with the dance of the Angrismene performed at the festivals of Venus, and still popular among the modern Greeks. The English jig is said to be one form of the Chica. It is not a little singular that the word came into use soon after a free intercourse with Spain was opened. The words Chaconne

Digitized by Google

(*Fr.*), Ciaccona (*It.*), Cachucha (*Sp.*), Czardasch (*Hungarian*), describe modern modifications of the Chica. [Bolero.] [Country Dance.]

Chiesa (*It.*) Church. *Sonata di Chiesa*, a sacred sonata.

Chiffres (*Fr.*) Figures, *basse chiffrée*, figured bass.

Chime. (1) To play a tune on bells, either by machinery or by hand, by means of hammers, or swinging the clappers, the bell remaining unmoved. It is opposed to *ringing* in which the bells are *raised*, that is, swung round. (2) A carillon.

Chirimia (*Sp.*) An oboe (from Chirimoya, a pear), the portion of the oboe in which the mouth-piece is inserted, called in German *Birne*, a pear.

Chirogymnast. Finger-trainer. A contrivance for strengthening the fingers, consisting of a cross-bar, from which are suspended rings attached to springs. The term is also applied to any apparatus designed for a like object.

Chironomy. *Gk. χειρονομία.* (1) Gesticulation by the use of the hands. (2) Directions given by movements of the hand, especially to a chorus. In the early church of the West such a system was much in vogue; and some have maintained that the signs of sounds, as then written, were merely pictorial representations of the movement of the hand.

Chiroplast. Finger-former. An instrument invented by Logier in 1810, to facilitate the proper method of playing the pianoforte. It consisted of a position-frame, finger-guides, and a wrist-guide. The position-frame consisted of two parallel rails extending from one extremity of the keys to the other, and fastened to the pianoforte. This frame served as a line upon which the finger-guides travelled; these guides were two moveable brass frames, with five divisions for the fingers, and to each guide was attached a brass wire with a regulator, called the wrist-guide, by which the position of the wrist was preserved from inclination outwards. With the instructions for the use of the chiroplast, progressive lessons on the pianoforte were given; and in the success attending the use of the hand-guide, these lessons, which were cleverly designed, had doubtless as much to do as the machine itself, which, however, soon fell into disuse.

Chitarra (*It.*) A guitar.

Chitarra col arco (*It.*) A violin with sides gently curved, as in a guitar; without corners, as in an ordinary violin.

Chitarrina (*It.*) A small Neapolitan guitar.

Chiudendo (*It.*) Closing, ending. The word is generally employed in connection with another, *chiudendo colla prima strofe*, ending with the first verse.

Chiuso (*It.*) Close, hidden, concealed, *e.g.*, *canone chiuso*, a close canon, [Canon]; *con bocca chiusa*, with the mouth closed humming.

Chœur (*Fr.*) [Chorus.]

Choir. (1) A part of the building in a cathedral or collegiate chapel set apart for the performance of the ordinary daily service. The choir is generally situated at the eastern end of the building, and is frequently enclosed by a screen, upon which the organ is placed.

(2) The minor canons, choral vicars, and choristers, or other singers taken collectively, are spoken of as the choir. The choral body is usually divided into two sets of voices, the one sitting on the north and the other on the south side of the chancel, and are known by the respective titles of Cantoris and Decani from their nearness to the Cantor (or Precentor) and to the Decanus (or Dean). In most cathedrals and collegiate chapels, the Decani side is held to be the side of honour, the best voices are placed there, and all the "*verses*" or *soli* parts, if not otherwise directed, are sung by that side, which is also considered the "first choir" (*coro primo*) in eight-part music.

Choir-man. An adult member of a choir.

Choir Organ. [Organ, § 1.]

Chor (*Ger.*) Chorus. Choir of a church or concert room.

Choragus. (*Lat.*) (1) The leader of the chorus in the ancient Greek drama. [Chorus.]

(2) The title of a musical official at Oxford University, whose duties are described in the Statutes.

Choral. (1) Of, or belonging to the choir, concert, or chorus. Choral service, a service with music. (2) A hymn or psalm tune.

Choräle (*Ger.*) [Hymn tunes.]

Choraliter (*Ger.*) In a choral form.

Choralmässig (*Ger.*) [Choraliter.]

Choral Music. Vocal music in parts, as opposed to instrumental.

Choral Service. A service of song; a service is said to be *partly* choral, when only canticles, hymns, &c., are sung; *wholly* choral, when in addition to these, the versicles, responses, &c., are sung.

Choral Vicars. [Lay Vicars.]

Choramt (*Ger.*) Choral service. Cathedral service.

Choraules (*Gk. χοραύλης* from *χορός* and *αὐλέω.*) (1) A player on the flute in the Greek Theatre. (2) One who keeps a chorus and plays in it himself.

Chord. A combination of musical sounds, consonant or dissonant. [Harmony.]

Chord. A string.

Chorda characteristica. A chord of the 7th in which a leading note appears.

Chordæ essentiales (*late Lat.*) The tonic and its 3rd and 5th. The key-chord.

Digitized by Google

Chordaulodion. A self-acting musical instrument, invented by Kauffmann, of Dresden, in 1812.

Chordienst (*Ger.*) [Choramt.]

Chordometer. A gauge for measuring the thickness of strings.

Chords étouffés (*Fr.*) (1) Chords played on the pianoforte with the *sordino* pedal held down. (2) Chords on the harp, lute, guitar, or dulcimer, *damped* by placing the hand gently on the strings.

Choriambus A metrical foot consisting of two short between two long syllables.

Chorister. A member of a choir whether juvenile or adult. At the present day the *children* of the choir of a church or cathedral, are those usually distinguished by the term, but so recently as the commencement of the present century, all who were engaged in taking part in the musical portion of the service, were called choristers. The word derived from χορός, by metonomy came to signify a band of singers or dancers, or any member of such a band, and hence the term is often applied to a singer in a chorus not necessarily belonging to a church, just as *choir* is applied to the place in which church singers sit, as also to any body of singers of sacred or secular music. For example, *Les enfans de chœur*, children of the choir or chorus; *Domchor*, cathedral choir or chorus, and *Coro della chiesa*, church choir or chorus; Choristers, or boy singers, called "clerks of the third form," in some places, are attached to every cathedral in England, and receive advantages of more or less value, in exchange for their services as members of the choir. In addition to necessary instruction in music, they have an education in other matters, varying in many places according to the construction put upon the Statutes by the deans and chapters of the cathedrals. The interpretation of these Statutes has been the subject of grave dispute, as the advantages accruing to the choristers have been from time to time most shamefully ignored. In days past, the children have been shut out from the enjoyment of preferential privileges made concerning them, and their education and moral training has been so little cared for, that many a child who in early years was familiarised with the most sacred matters, has acquired for them the proverbial result of familiarity. A better state of things is now being brought into existence with a result which cannot be considered other than hopeful.

In some places private instructors have been engaged to teach cathedral choristers a few matters besides music; in others they are admitted into the chief grammar schools of the several cities. The course of instruction also varies, for in some cases they are taught the simple elements of reading and writing, in others they learn as much of the higher branches of education as is possible in addition to the duties of their profession. In many instances their musical instructors impart no more than is absolutely needful for the exercise of cathedral duty, and in some music is taught scientifically as well as practically, not only in connection with the immediate work in hand, but also with reference to future use. In very few instances are the boys boarded and lodged within the precincts of the cathedral, or placed under the immediate care of the cathedral authorities out of the hours devoted to duty—a matter of much regret. The organist is sometimes music-master of the choristers, sometimes the office is distinct, and is held independently of the organist. In many cathedrals a sum of money as apprentice fee is paid to a chorister on leaving the choir; this is instead of the money at one time set apart for the maintenance of the chorister as a student at the universities. For instance, in the Statutes of Stoke College, in Suffolk, founded by Parker, Archbishop of Canterbury, are these words: "of which said queristers, after their breasts be changed (their voices broken) we will the most apt of wit and capacity be holpen with exhibition of forty shillings, the rest with lesser summe."

In olden times, choristers were privileged to demand a fee from every newly installed officer of the church, and to levy "spur money" from all who attended the service in riding gear. In the former case the fee varied according to the position of the installed officer, and was paid without conditions being imposed in return; in the latter, the wearer of spurs could require the youthful tax-gatherer to repeat his "gamut" perfectly; if he hesitated, he lost his spur-money. The boys of the Chapel Royal were the last to keep up the custom which has now fallen into disuse with many others equally absurd. For example: the choristers in many cathedrals and collegiate establishments were permitted to rule over their superiors for a short period once a year, generally from December 6th, the Feast of St. Nicholas (the patron saint of sailors, parish clerks, thieves, and boys) until Innocents-day, December 28th. From the aptitude acquired in these ludicrous ceremonies, the choristers gained such a skill in acting that they were selected to perform in the mystery plays of old time, and later to represent the masterly conceptions of such writers as Shakespeare, Ben Jonson, and others. It was not alone to the choir boys attached to unimportant establishments that these matters were entrusted, but also to the children of "Powle's churche" and of the "Chapel Royale" of Her Majesty Queen

Elizabeth. The possession of such powers and privileges may have been exceedingly pleasing while it lasted, but the ill sorted union of the theological and the theatrical is happily dissolved, it is hoped for ever.

The life of a chorister in these remote days was, however, not all bliss, as the owner of a good voice would probably find to his cost, more especially if he was not fortunate enough to belong to St. Paul's Cathedral, or the Chapel Royal, for there were officers armed with the awful warrant of the Royal Court, empowering them to roam the country, to visit all churches and cathedrals of the lesser sort in which choral singing was practised, and to select and take away all boys "with good breasts," that is to say, all with voices of more than ordinary excellence, for the service of the privileged choirs. The "placard" or warrant was often used illegally, and children were impressed for choirs other than those above mentioned. There is reason for believing that choristers so gained were in general badly used, if we may trust Thomas Tusser (1523-1580), the author of "Five hundred points of good husbandry," for he speaks of his good fortune in having been assigned to John Redford, organist of St. Paul's, in terms which prove that choristers were not so kindly used in other places. His situation at Wallingford, from whence he was impressed, he laments in the words:

> "O shameful time! for every crime
> What toosed ears, like baited beares.
> What bobbed lippes, what yerkes, what nips.
> What hellish toies!
> What robes, how bare, what colledge fare,
> What bread, how stale; what penny ale,
> Then Wallingford, how wert thou abhor'd
> Of silly boies."

In another verse he contrasts his treatment:

> "But marke the chance, myself to vance,
> By friendships lot to Pauls I got,
> So found I grace a certayn space
> Still to remaine
> With Redford there, the like no where,
> For cunning such, and vertue much.
> By whom some part, of musicke art
> So did I gain."

On the Continent choristers are attached to many cathedrals, but their duties and education are based upon a different system to that in general use in Great Britain at the present time. Before the time of the dissolution of the monasteries the position of choristers was much the same as that enjoyed abroad by them, and it was no uncommon thing to find "the children of the choir" in after life occupying stations of eminence and trust in both Church and State. The venerable Bede, St. Swithun, St. Hugh of Lincoln, William of Wykeham, William Wainfleet, Erasmus, and his friend, Dean Colet, the founder of St. Paul's School, and scores of other distinguished men were choristers. The musicians who have gained the first knowledge of their art within the walls of a church are many, and comprise among others the names of Palestrina, Frescobaldi, Orlando di Lasso (impressed from Hainault into Italy as a child), Padre Martini, John Sebastian Bach, Haydn, William Byrd, Tallis, Dr. Bull, Dr. Rogers, Dr Blow, Elias Ashmole, Henry Purcell, Dr. Croft, Pelham Humphreys, Dr. Greene, Battishill, Dr. Burney, Attwood, and many famous living musicians whose names it is not necessary here to catalogue.

Chorton (*Ger.*) (1) The ancient ecclesiastical pitch in Germany. It was supposed to be higher than that employed for secular music by about a tone. The terms Kammerton and Chorton were used to signify the difference between a high and a low pitch for the same denominated sound. (2) The melody of a hymn or psalm tune.

Chorus. Chor (*Ger.*) Chœur (*Fr.*) Coro (*It.*) (1) A band of singers and dancers employed on certain occasions in the ancient Greek theatres, and other public places. It was the custom for the whole population of a city to meet on stated occasions, and to offer thanksgivings to the gods for any special advantages obtained, by singing hymns accompanied with dances. Donaldson derives the word from χορος, the name of the place where these exercises were performed in Sparta, and shows the connection between the civil and religious ceremonies of the ancient Greeks, saying that music and dancing were the basis of the religious, political, and military organisation of the Dorian States. The choral songs were always written in the Doric dialect, and the choral dances were Dorian also. In course of time, as the fine arts became more cultivated, the duties of the chorus as a branch of worship devolved upon a few, and ultimately upon one, who bore the whole expenses, when paid dancers were employed. This person was called the choragus and it was his business to provide the chorus in all plays, whether tragic or comic. His first duty after collecting his chorus was to find and pay a teacher (χοροδιδάσκαλος) who instructed them in the songs and dances which they had to perform. The choragus was allowed to press children, if their parents did not give them up of their own accord. He lodged and maintained the chorus until the time of performance and provided them with such aliments as conduce to strengthen the voice, he had also to find masks and dresses. The honour was much coveted among the wealthy Athenians. The choragus who exhibited the best theatrical entertainment generally received a tripod as a reward of praise.

Digitized by Google

The choral dance reached its perfection in the χορὸς κυκλικός at Athens. This chorus consisted of 50 persons. The number of the chorus varied in later times according to the performance. The χορὸς τραγικός consisting of from 12 to 15, the χορὸς κομικός of 24, and the χορός σατυρικός of the same number as the τραγικός. The chorus in the time of the Attic tragedy consisted of a group of persons, male and female, who remained in the theatre as witnesses as well as spectators. When they spoke, it was to offer reflections on the scene passing before them, taking part with or against the *dramatis personæ* by offering advice, comfort, exhortation, or dissuasion. At times the chorus was divided and spoke antiphonally. These divisions moved according to a pre-arranged order, which movement probably originated the naming of the stanzas which were called *strophe*, *antistrophe* and *epode*. When not engaged in singing, the chorus grouped itself upon a platform called the Thymele, which was in the centre of the building, and from whence all measurements were made, the semicircle of the amphitheatre being described from it as its centre. Of the exact part music played, whether elaborate compositions were employed or not, little is now known. It is supposed that a simple rhythmical declamation analogous to chanting was used. The accompaniment of flutes in unison was made use of for the choruses. The chorus declined with the ancient tragedy, and the few attempts made by modern writers to revive the manner of the ancients, as in Schiller's "Braut von Messina" have not been successful. The well-known Antigone and Œdipus Colonæus of Mendelssohn can scarcely be regarded as a reproduction of the ancient Greek chorus, owing to the insuperable difficulty of adapting modern instrumentation to the spirit and observances of the older customs.

(2) An ancient instrument variously described by different writers. A bagpipe must have been signified when the word was used in the 10th century, as a chorus or corus is described as "pellis simplex cum duobus cicutis." The word is supposed to be connected with *cornemuse*, as it is sometimes written *cormusa* and *corusa*. In the Promptorium Parvulorum, 15th century, the word is used to describe "a crowde, an instrument of musyke";—the drone of the bagpipe and the unstoppable strings of the "crowde" bearing a sort of burden or chorus to the melody played on the other pipes or strings. Busby in his "Dictionary of Music," 1810, says that the word is the old Scottish name for a trumpet of loud tone.

(3) A personage in some of Shakespeare's plays, who between the acts utters reflections upon scenes that are past, and describes scenes to come.

(4) A composition for a number of singers, with or without accompaniment, intended as the expression of the united sentiments of a multitude. A chorus may be independent and complete in itself, or may be a portion of a large work either sacred or secular. It may contain opposed sentiments interwoven, as in the Kermesse scene in Gounod's "Faust;" in Meyerbeer's "Huguenots," and "L'Etoile du Nord;" Wagner's "Tannhäuser" and "Lohengrin;" according to the purposes of the drama. Choruses with opposed subjects are not infrequent in oratorios, as in Handel's chorus, "Fixed in his everlasting seat," in "Samson." The union of independent themes may be traced by the student with advantage, in the choruses of such compositions as Bach's Passion music, &c. Double, triple, or even quadruple choruses are often found in the works of the old Italian church writers, as well as in the sacred compositions of Bach, Handel, and later musicians.

The choruses in the early Italian operas were devoid of dramatic character, and in fact, were often independent of the action of the opera in which they were inserted. As they contained occasional reflections on passing events they were in some sort connected with the ancient Greek chorus, the stage directions enjoining the dancers to accompany the singing with motions and gestures, also formed another link binding them to their ancient model. The invention of the operatic chorus, or rather the introduction of combined voices as a necessary part of the dramatic action is claimed by the French. Many of the lesser musical dramas, burlettas, interludes, &c., had no choruses properly speaking, a glee or some concerted piece for the principals being all that is found in them. The choruses of Lully are not very dramatic, and those of Rameau are very badly constructed, and often incorrect as to their harmony, so that the improvements introduced by these two masters were not extended to the chorus. Among the followers of Lully, Campra (1660-1744) is the most distinguished; he treated his choruses in a more advanced manner than his model, not only in the development of harmonic effects, but also by the introduction of novel rhythms. Gluck invented *morceaux d'ensemble*, grand indeed, when compared with the choral effects by other composers of his own and preceding times; Spontini added new instrumental colouring; Cherubini employed the graces of form to clothe the musical outlines suggested by his predecessors; Rossini did as little comparatively for the chorus from a dramatic point of view, as Meyerbeer did much. Some of Bellini's choruses are conceived in fine dra-

Digitized by Google

matic spirit; Verdi's, though occasionally vulgar in detail, are by no means wanting in general force and appropriateness; the faults they exhibit are due to the influence of the fetters of tradition, and a wearying use of simple tonic and dominant harmonies; Gounod's are often admirable, but as the expression of the voices and opinions of a multitude, Wagner's choruses are nearest the true ideal. In many of the early English operas the chorus is an inconsiderable item. Exception must, however, be made in favour of those by Henry Purcell, whose works of this class exist, while many of the other portions of his operas have fallen out of memory. Neither Lampe, Arne, nor Storace paid much attention to the development of chorus, but its improvement in smaller dramatic works is due to Sir Henry Bishop. As conductor of the music at Covent Garden Theatre for many years, he produced a series of compositions of more or less value, some of which live, while the dramas and plays for which they were written have fallen completely into oblivion. His earlier choruses have the *glee* attached, the part for the multitude of voices being as easy as possible. It is presumed that operatic chorus singing was not in its highest state of perfection in the days of Mozart or Beethoven, for neither of these composers has given the chorus much that is difficult or important in their operas—a matter of necessity perhaps, as it was hard to find a body of singers in those days, who would be so far content with the gifts they possessed as to accept an inferior position. Chorus singing was very little cultivated in London so recently as the time when the Sacred Harmonic Society was established, for it was found necessary to invite a contingent of singers from the North of England to take up a residence in the Metropolis, employment being found for them for the hours when their vocal services were not required.

If it is necessary to account for the slow growth of operatic chorus, when compared with the progress made in other portions of the musical drama, a very simple reason can be found, in the fact of the difficulty attending the first performance of a variety of works when the singers have to commit the whole of their parts to memory. Straightforward as many of Handel's choruses are, it was found necessary to make some slight alterations in the choruses of such a work as "Acis and Galatea" when given on the stage. It is not difficult to gain the most sublime effects from chorus singing when the performers have the copies before their eyes, as the performance of oratorios and similar works by large choral societies can sufficiently testify.

In the Oratorio, the chorus is of the greatest importance, and the number of voices to a part is generally larger than it was in the time of Handel, though some writers question the advantage gained by multiplying the forces, as it is supposed by them that a greater number of voices does not necessarily produce a proportionate power of tone, and the difficulty of moving a large body in concert may involve a change of *tempi*, but this need not be the case with a trained body well acquainted with the works performed. It is said that "a chorus of thirty-five voices from the Pope's chapel who sang at the Coronation of Napoleon I., in the Cathedral of Notre-Dame, Paris, produced a far greater and more wonderful effect when they entered singing the *Tu es Petrus*, than another chorus of hundreds of voices, and eighty harps, that had been assembled and trained for the same occasion, in expectation of surpassing all that man could imagine." As the knowledge of music is more general in the present day, this objection cannot with reason be entertained now, otherwise the choruses at the Handel Festivals could not be held to be the most attractive features of such gatherings.

(5) The union of a number of voices for the joint performance of a composition. The whole of the male and female singers other than the principals whether in the oratorio, drama, or at a concert.

(6) The refrain or burden of a song whether sung by one or by many voices. As for example:

(*a*) I lov'd a lass, a fair one,
As fair as e'er was seen;
She was indeed a rare one,
Another Sheba Queen.
But fool, as then I was
I thought she lov'd me too,
But now, alas! she's left me.
Chorus. Falero, lero, loo.

George Wither

(*b*) Lisette, dont l'empire
S'étend jusqu'à mon vin,
J'éprouve la martyre
D'en demander en vain.

Pour souffrir qu'à mon âge
Les coups me soient comptés,
Ai-je compté, volage,
Tes infidelités?

Chœur. Lisette, ma Lisette,
Tu m'as trompé toujours,
Mais vive la grisette
Je veux, Lisette,
Boire à nos amours.

Béranger

(*c*) Lasst tanzen uns und springen,
Hier, wo die Blumen stehn;
Und frohe Lieder singen,
Im Freien klingt es schön.
Chor. Und frohe Lieder, &c.

Uhland.

Digitized by Google

(7) The name given to the mixture and compound stops in an organ.

Christe eleison (*Gk.*) A portion of the Kyrie in the Mass. [Mass.]

Christmas Carol. [Carol.]

Christmas Music. (1) Cantatas, the words of which are suitable to Christmas-tide. (2) Music played by waits. [Waits.]

Chroma (*Gk.* χρῶμα, colour or complexion). The name of one of the modifications of the Greek musical scale. The principal chromatic scale of the Greeks was called *χρῶμα τοναιον*; its chief characteristic is the omission of the 4th and 7th. [Greek Music.]

Chroma duplex (*Lat.*) (1) A semiquaver. (2) A double sharp.

Chromatic. That which includes notes not belonging to a diatonic scale.

(1) A chromatic chord is that which contains a note or notes foreign to diatonic progression, *e.g.*:

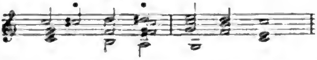

(2) Chromatic harmony is that which is made up of chromatic chords.

(3) A chromatic interval is that which is augmented or diminished, *e.g.*:

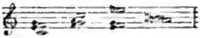

(4) Chromatic modulation is a passing into an extreme key, by means of chromatic harmony.

(5) A chromatic scale is one which consists of a succession of semitones.

Chromatique (*Fr.*) } Chromatic.
Chromatisch (*Ger.*) }

Chrotta. [Crowd.]

Church modes. [Plain Song.]

Ciaccona (*It.*) [Chaconne.]

Cicogna (*It.*) The mouth piece of a wind instrument; *lit.* a stork.

Cicuta (*Lat.*) A flute or pan-pipes, made from the stalks of the hemlock plant.

Cimbalo (*It.*) (1) Harpsichord. (2) *cimbali*, cymbals. (3) A tambourine.

Cimbel (*Ger.*) A mixture-stop in German organs.

Cimbelstern (*Ger.*) *Lit.* cymbal-star. A mechanical contrivance in some German organs, consisting of star-shaped cymbals attached to a wheel set in motion by a pedal.

Cink (*Ger.*) A reed organ stop.

Cinque (*It.*) A fifth part in concerted music.

Cinyra. An old term for a harp.

Cipher. The sounding of a note on an organ caused by mechanical derangement.

Circular Canon. A canon so constructed that it closes in the key one semitone above that in which it commences. As, at each repeat, it begins, not at the *original* pitch, but at the pitch at which it closed, it is evident that twelve repetitions would take it through all the known keys.

Circulus (*Lat.*) A circle. One of the time signatures of early music. It was only applied to *tempus perfectum* and *tempus imperfectum*, that is to the division of a breve into semibreves. When the breve was divided into three semibreves (tempus perfectum) a complete circle was placed at the signature "quia forma rotunda perfecta est." When the breve was divided into two semibreves (tempus imperfectum) a broken circle or semicircle was used. This last sign is now corrupted into a C which is erroneously said to stand for *common* time. [Time.]

Cis (*Ger.*) The note C sharp.

Ciscis (*Ger.*) The note C double sharp.

Cis dur (*Ger.*) The key of C sharp major.

Cis moll (*Ger.*) The key of C sharp minor.

Cistella (*Lat.*) A dulcimer, *lit.* a little box. [Citole.]

Cistre (*Fr.*) [Cittern.]

Cistrum. [Sistrum.]

Citara (*It.*) A cittern, guitar, cither.

Cithara (*Gk.* κιθάρα). The ancient lute. It probably differed from the lyre in having something behind the upper part of the strings, whereas the lyre-strings were open on both sides. [Guitar.]

Cithara bijuga. A guitar or lute having a *double neck*. Some of the strings passed along the side of the finger-board and could not therefore be stopped, and some of the higher strings were tuned in pairs in unison. Sometimes it was strung with *wire* and played with a plectrum like a cithara; at others, it was strung with cat-gut and played like a lute.

Citharœdus. One who sings whilst playing the cithara, whereas a *citharista* only played.

Citole. An instrument similar in form to the dulcimer. The name is supposed by some to be derived from *cistella* a little box; by others, from *cithara*, a guitar. It is frequently mentioned by early poets, apparently with various meanings. William Guiart (1248) says:—

> "Qui le roi de France à celle crée
> Enveloppa si de parolles
> Plus douces que son de citoles"

Chaucer's allusion to it would imply that it was a kind of guitar:—"A citole in hir right hand had sche."

Cittern. A lute-shaped instrument strung with wire instead of gut. It had eight strings tuned to four notes *g*, *b*, *d*, and *e*, or corresponding intervals. The instrument was at one time very popular, a cittern being part of the furniture of a barber's shop, the customers amusing themselves with it

Digitized by Google

while waiting. The music for the cittern was written in Tablature. There were several instruction books for the instrument issued, while it was still in use. The title of the earliest known is "The Cittharn Schoole, by Antony Holborne, Gentleman, and servant to Her Most Excellent Maiestie. Hereunto are added six short Airs, Neapolitan like to three voyces without the Instrument: done by his brother William Holborne. London, 1597." [Guitar.]

Civetteria, con (*It.*) In a coquettish manner.

Clairon (*Fr.*) [Clarin.]

Clang. (1) *Timbre* (*Fr.*) *Klang* (*Ger.*) Quality of tone. (2) The peculiar "ringing" noise or *din* produced by the clash of metals, or the blast of loud wind instruments. Lat. *clangor*. The Gk. κλαγγή is also applied to the *whiz* or *twang* caused by the discharge of an arrow, &c.

Clapper. [Bells.] [Bones.]

Claque (*Fr.*) A body of hired applause-makers, openly employed in France and sometimes secretly resorted to in England. The "claque" in France is divided into several ranks; *rieurs, pleureurs, chatouilleurs, bisseurs*, and so forth. These officers distributed in several parts of the theatre, laugh, weep, gossip with their neighbours, cry *encore*, &c., under the direction of a fugle-man whose business it is to study the work produced, and after consultation with the author, the performers, and the stage manager, to direct and regulate the reception of certain portions of the entertainment.

Claque-bois (*Fr.*) A gigelira. [Ligneum Psalterium.]

Clarabella. An organ stop consisting of open wood pipes, invented by Bishop. It is of a soft and sweet quality of tone. It is generally merged into a stopped diapason below middle C, as the larger pipes do not produce a clear tone. It is usually of 8 ft. pitch.

Clara voce (*It.*) A clear voice.

Claribel Flute. An organ stop of similar construction to the clarabella, but generally of 4 ft. pitch.

Clarichord. A stringed instrument of mediæval times, by some writers presumed to be identical with the clavichord, the precursor of the spinet, harpsichord and pianoforte.

The earliest stringed instrument with a keyboard for the fingers, is said to have been invented about the year 1300, in Italy, and to have been called Clavycytherium, the cithara or harp with keys. In the Musurgia of Luscinius, printed in 1536, a picture of this early instrument is given, but as there is no statement that it had been drawn from an existing example, and moreover, as it appeared nearly two centuries and a half after the supposed invention, its likeness is at the least rather doubtful. Trustworthy representations of keyed instruments are very rare before the latter of the above dates, and a consequent amount of confusion exists, as well with regard to forms as to names. In the few early English glossaries extant, no clear or lucid description of the majority of instruments in existence can be obtained, and many modern writers, misled by varied spelling, and the number of different names for the same thing, have exercised their wits in finding a variety of forms, shapes, and uses for that which was after all but of one character, and probably the same thing described variously. To speak of the older names of instruments of the kind now under consideration, the words clavichord, monochord, manichord, clavycymbal, cembalo, clavecin, all have the same meaning, a stringed instrument with keys played by hand. Some writers explain the words clavichord to be so called, because the strings were wrested in tune with a *key* (clavis);—this description would also apply to the **clarichord,** called also claricols, clarigold, &c., for that was kept in tune by a wrest, as William Cornishe in his poem, "A treatise betwene Trouthe and Enformacion." 1500, tells us. He says:

> "The clarichorde hath a tunely kinde
> As the wyre is wrested hye and lowe
> So it tuneth to the players mynde.
> For as it wrested so must it nede showe.
> Any instrument mystuned shall hurt a trewe song,
> Yet blame not the clarichord the wrester doth wrong"

By this it would appear that the clarichord was a kind of harp, tuned as it was required to be used, and it is somewhat singular that as the words clarichord, clarigols, clarigold, &c., are only used in ancient English writings, no form of them being found in old French or Latin, the derivation of the word from an ancient British etymon, such as *clar*, to grip or bend, would be applicable to the harp as an instrument whose strings were plucked or gripped, the modern Irish word for playing on harps is **claꞃeꝺıꞅeꞃ** (clarediser), and the harp itself is called **claꞃꞅcaꞇ** (clarscat), gripped or clawed string. The word **ꞅcaꞇ** might be compared with the German Saite.

Dr. Rimbault, after quoting the definitions of the word clarichord given by a few of the lexicographers of the 17th century, observes that they make "no distinction between the terms clarichord and clavichord, but the one can hardly be a corruption of the other." He adds that the "words suggest a totally different etymology," and he then proposes that it might be from the French word *claire*, denoting a transparent tone; but in the earliest musical dictionaries clarichord is said to be "called also the Dumb spinnet, on account of the cords being covered with pieces of cloth." The Clavichord or clavicembalo is said to signify a harpsichord. In every instance in which the word clarichord is employed before the 16th century it might fairly be translated harp.

At the marriage of James of Scotland with the Princess Margaret, in the year 1503, "the king began before hyr to play of the clarychordes and after of the lute. And upon the said clarychorde Sir Edward Stanley played a ballade and sange therewith." (Warton, "History of English Poetry.") It is quite possible that the similarity of the two letters v and r in ancient MSS. might have led the transcribers to mistake one for the other, and by writing clarichord or clavichord indiscriminately they might innocently cause controversy. All modern authorities on the subject declare that the words are of separate and distinct origin. [Pianoforte.]

Clarin (*Ger.*) A species of trumpet, a clarion, also an organ reed stop of 4 ft. pitch. *Clarinblasen*, the sound of a trumpet. Sometimes the word is applied to the soft tones produced on this instrument.

Clarinet
Clarinette (*Fr.*)
Clarinetto (*It.*)
An important wind instrument said to have been invented by John Christopher Denner, who was born in Leipsic, in 1654, but it was in reality only a modification and improvement of the more ancient *shawm* or *chalumeau*. Most authors relate that Denner invented the instrument in 1659, at which date he was four years old; but it was made by him after his residence in Nuremberg, in 1690. The name clarinet, or *clarionet* was probably attached to it on account of its pure and brilliant tone, not unlike that of the clarion or trumpet. In modern instruments of this class, the tone has been rendered far purer and sweeter than that originally produced. The difference between the tone of the hautboy and that of the clarinet, is due to the circumstance that the one has a double, and the other a single reed. The difference in the nature of the scale arises from the fact that the hautboy is conical, while the clarinet is a cylinder, the series of harmonics in the hautboy following each other in the ratio 1, 2, 3, 4, &c., those of the clarinet 1, 3, 5, 7, &c., hence, that whereas the first *overtone* of the former is its octave, the first *overtone* of the latter is its twelfth. On this fact depends the difficulty of making shakes and of playing rapid passages on certain parts of the clarinet. All sudden changes from the end of the first range of twelve notes to the commencement of the second series, are difficult, some impossible.

The compass of the instrument is about three octaves and a half from tenor E, including all the intermediate semitones. The clarinet being of the nature of a stopped pipe, as to its harmonics, can be played from its lowest note E up to twelve notes higher without a break by means of its keys. At this point the player has to increase the pressure of wind, and commence a new series of sounds, the transition between these two registers forming the chief difficulty in "clarinet blowing." The registers are four in number, and are as follows:

1. The low includes all notes between

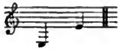

2. The second between

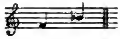

3. The third between

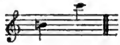

4. The fourth comprises all remaining notes from

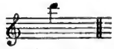

The first two registers are called the "chalumeau part," and when this is employed for any continued time it is written an octave higher, and marked *Chal*, the original pitch being restored by *Clar*. There are three lengths of tubing employed for the clarinet, by which means the instrument may be made to sound three different scales according as the tube is short or long. The longer tube is used for the A clarinet, a medium for the B flat, and the shorter for the C.

As the fingering is in each case the same in each instrument it has been found convenient to adopt C as the normal scale, so that a piece of music apparently the same to the eye is different to the ear, according to the clarinet employed. Thus the passage written as follows:

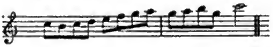

When played upon a C clarinet would sound as it stands, upon an A clarinet would sound:

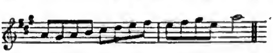

And upon a B flat clarinet, as:

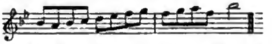

It is, however, easy to make each instrument give out the same notes by employing a change in the signature, thus the passage for the A clarinet should be written:

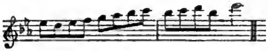

And for the B flat clarinet:

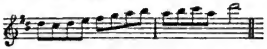

when they will each give the sounds as written in No. 1.

The kind of clarinet required is usually stated at the commencement, as clarinet in A, B flat, or C, and whenever a change is needed during the progress of a piece, the same is indicated during a period of rest for the instrument, by the words change to A, B, or C, as the case may be.

The advantage of a change of clarinet is that complicated scales upon one instrument become easy upon another. For instance the scale of F sharp major which is very difficult on a C clarinet, when played upon an A clarinet is fingered as A major, the real sounds produced being those of the scale of F sharp major. Similarly the scale of D flat major would become the scale of E flat major on a B flat clarinet. This accounts for the fact that the clarinet part in a full score is sometimes in a flat key while the movement is in a sharp key. For example a piece in the key of E minor (one sharp) not uncommonly has a part for an A clarinet written, of course, in G minor with two flats.

The favourite (because easy) keys of the clarinet are the keys of C, F, and G, B flat, E flat, A flat and D with their relative minors. Hence the skill of the composer is shown in writing for that clarinet capable of producing the best effects in certain keys.

Clarinets are usually employed in pairs, and the parts are ordinarily written on one stave. They, in conjunction with the two bassoons similarly written, form a grand basis or support for all the other wind instruments. The small E flat clarinet (playing a minor third above the notes actually written) is used in military bands. Its tone is shrill and piercing. The introduction of the clarinet as a regular instrument in the orchestra of the opera is due to J. Christian Bach, who wrote special parts for a pair of clarinets in his opera "Orione, ossia Diana vendicata," which was produced in London in 1763.

Clarinettista (*It.*) Clarinettiste (*Fr.*) A performer on the clarinet.

Clarinetto (*It.*) [Clarinet.]

Clarino (*It.*) (1) A trumpet. (2) An organ stop, consisting of reed pipes, generally of 4-ft. pitch.

Clausula (*Lat.*) A close or cadence, *e.g.*: *clausula falsa*, a false cadence; *clausula finalis*, a final cadence, &c.

Clavecin (*Fr.*) (1) A harpsichord. (2) The keys by means of which the *carillonneur* plays upon the bells. [Pianoforte.]

Claviatur (*Ger.*) (1) The key board of an organ or pianoforte. (2) Fingering.

Clavi-cembalum (*Lat.*) Clavicembalo (*It.*) [Pianoforte.]

Clavichord. [Clarichord.]

Clavicylinder. An instrument in the form of tubes or cylinders of glass, invented by Chladni. There was another instrument with the same name made of plates of glass of graduated lengths, the tone of which was produced by hammers set in motion by a key-board.

Clavicytherium. [Clarichord.]

Claviglissando. An instrument with a key-board, invented by C. W. Le Jeune, which is intended to combine the properties of the violin and harmonium—of the violin in obtaining a slide or portamento, and the harmonium in the capability of imitating the tones of various wind instruments.

Clavier (*Ger.*)} (1) The pianoforte. (2) A
Clavier (*Fr.*) } row of keys on an organ.

Clavierauszug (*Ger.*) A pianoforte score, as opposed to *Partitur*, a full score.

Clef (*Lat. clavis.*) The sign placed at the commencement of a staff or stave, showing the absolute pitch, the lines without it showing only the *relative* distances of sounds. When it was found that *neumes* could be better interpreted by the use of lines, a red and a yellow line were used, it being understood that the former bore the note F, the latter C. It is easy to see, that the fact having been once established that lines could represent notes, it would be found much easier to attach the letter itself to the commencement of the lines, than to *colour* the line. The coloured lines were invented by Guido, but Walter Odington (13th century) used one of the letters of the musical notes as a clef to his stave of four lines. At this period, it was not usual to employ leger lines, but if the voice exceeded in compass the limit of the stave, the position of the clef was altered; a practice still retained in plain-song books. [Notation.] The letters C and F were most commonly used in all ecclesiastical music up to the time of Palestrina, after which other clefs were introduced.

There were five sorts of clefs in use in the 16th century, namely, the gamut Γ, from the Greek gamma, the F, C, g and d clefs. These were ultimately reduced to three, the gamut and the d clefs being found to be unnecessary. The position of the clefs was held to represent a certain pitch, and as it was supposed that the scale was incapable of extension beyond the notes indicated by the

Digitized by Google

clefs gamut and d, their places marked the boundary of ecclesiastical compass. The other clefs might be made moveable if needed, for reasons already intimated, but whatever the number of lines above or below the clefs, each clef represented a particular sound. Thus, the F clefs indicated F *finale*, the C clef, *acutum*, and the G clef, G *superacutum*, &c. Many of the musical treatises of the 16th century contain a chapter "De Clavibus signatis," which is interesting as showing the form of the clefs as then employed.

All the writers of these tracts distinguish between the clefs proper to plain-song, and those employed for figurate or mensurable music. In the following copies of these arrangements of clefs, those for plain-song are on the left, and those for figurate music on the right. No. 1, from Finck's "Practica Musica," 1556:

Signa clavium in utroque cantu.

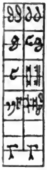

Et ponuntur omnes in lineali situ, quædam tamen sunt magis familiares, utpote F et C, g rariuscule. Γ vero et dd rarissime utimur. Unde, Linea signatas sustentat scilicet omnes. Et distant inter se mutuo per diapentem. F tamen γάμμα distinguat septima quamvis.

No. 2, from "Erotemata Musices Practicæ," by Ambrosius Wilphlingsederus, 1563:

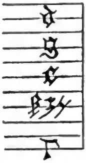

In chorali cantu simpliciter prescribuntur ita.

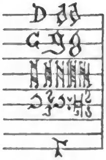

In mensurali vero hoc modo.

No. 3, from "Erotemata Musicæ," by Lucas Lossius, 1570:

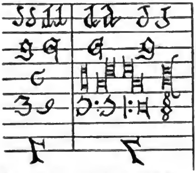

In cantu chorali. In cantu figurali.

In later times three clefs F, C and G were found sufficient for all purposes. The C clef appears upon all lines but the fifth.

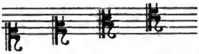

The first is called the *Soprano*, and is most frequently found in ecclesiastical music, though it not uncommonly indicates the treble voice part in modern full scores. The second is called the *Mezzo Soprano*, and is assigned to second treble, and sometimes alto voices, and in music of the early part of the 18th century is often used for the tenor violin. The third is the *Alto*, and the fourth the *Tenor* clef; the former being used for alto voices, violas, and the highest trombone, the latter by tenor voices, trombones and the upper register of the bassoon and violoncello. The F clef is placed upon the fourth line of the stave and is used for all bass voices and instruments. When it is found upon the third line as in some old music it is called the baritone clef:

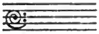

The following quotations from music books of various dates will show the forms through which the several clefs have passed. The "Compendium Musicæ," by Lampadius, 1537, is supposed to be the earliest printed book in which the G clef is used in a shape nearly similar to that now employed:

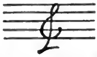

In Lully's and other French scores it is sometimes placed upon the *first* line:

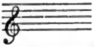

In this position it was called the "French violin clef;" and in an earlier work by Christopher Demantius, "Isagoge Artis Musicæ," 1656, it appears upon the third line:

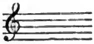

In "Ayres and Dialogues for one, two, and three voyces," by Henry Lawes, 1653, the forms of the clefs are as follows:

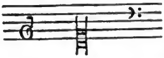

In Christopher Simpson's "Compendium of

Digitized by Google

Practical Musick," 1678, the clefs are in this shape:

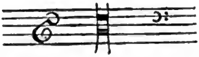

Playford's Psalms, third edition, 1697, the following are the forms:

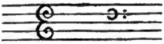

Matthew Wilkins's "Book of Psalmody," 1699:

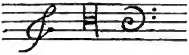

Dr. Croft's "Thirty Select Anthems," 1724:

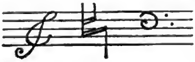

Godfrey Keller's "Rules for playing a thorow-bass," 1731:

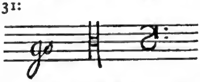

Emanuel Bach's "Sechs leichte Clavier-Sonaten," 1766:

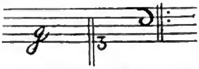

Malcolm, "A Treatise of Music," 1779:

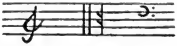

Shield's "Introduction to Harmony," 1800:

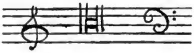

In many modern French music books the F clef is written thus:

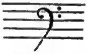

In this a resemblance to the letter F, the ancient clef sign for this pitch, may be traced. Many writers have maintained that clefs create difficulties in the way of a "right understanding of music," and have therefore suggested their removal or the substitution of simpler signs.

Thomas Salmon, 1676, proposed the use of the letter T, for the treble clef, M for the mean or C clef, and B for the bass clef. His proposal led him into a controversy with Matthew Locke, which was maintained on both sides in language not very creditable to either. Francis Delafond, in 1725, suggested the use of one clef only, the F, or bass clef. A century later, Miss Glover, in a pamphlet explaining her views on what she called the Tetrachordal System, proposed to abolish all clefs, a proposition which has since been carried out in the Tonic-Sol-fa method of teaching singing. The use of the treble clef for all purposes has also been recently advocated, but with little success.

Clivus (*Lat.*) [Neumes.]

Clocca (*Med. Lat.*) A bell. Irish, *clog*, a small bell. Fr. *cloche*. Ger. *Glocke*.

Cloche (*Fr.*) A bell.

Clochettes (*Fr.*) Hand-bells.

Clock, to. To set a bell in vibration, by attaching a rope to the clapper, and swinging it to and fro till it strikes the side of the bell which remains stationary. It is an undesirable practice, as many valuable bells have in this manner been cracked.

Clokerre (*Old Eng.*) *Clocherre* (*Old Fr.*) A belfry. In low Latin, *clocherium*.

Close harmony. Harmony produced by drawing the parts which form it closely together.

Close play. A direction in lute playing. The following explanation of the term is from Barley's Lute book: "Thou shalt not neede but to remoove those fingers which thou shalt be forced, which manner of handling we call *close* or *covert* play." It would appear to correspond to the *smooth style* sometimes adopted on the organ or other keyed instruments.

Clynke-bell. [Chime.]

C moll (*Ger.*) C minor.

Coalottino. [Concertino.]

Coda (*It.*) (1) The tail of a note. (2) The bars occasionally added to a contrapuntal movement after the close or finish of the *canto fermo*. (3) The few chords or bars attached to an infinite canon in order to render it finite; or a few chords *not in canon*, added to a finite canon for the sake of obtaining a more harmonious conclusion. (4) That closing adjunct of any movement, or piece, specially intended to enforce a feeling of completeness and finality.

Codetta (*It.*) *dim.* of coda. A short coda.

Codon (*Gk.* κώδων.) (1) A small bell, such as those attached to the trappings of horses. (2) A crier's bell. (3) The bell of a trumpet (Fr. *pavillon*). (4) A trumpet with a bell-mouth.

Cogli stromenti (*It.*) With the instruments.

Digitized by Google

Coi bassi (*It.*) With the basses.

Coi violini. With the violins.

Col (*It.*) With the; *e.g.*:

Col arco, with the bow.
Col basso, with the bass.
Col canto, with the melody.
Colla destra, with the right hand.
Coll' arco, with the bow.
Colla parte, with the principal part.
Colla punta dell' arco, with the point of the bow.
Colla sinistra, with the left hand.
Colla voce, with the voice.

College of Organists (**Royal**). Founded in 1864, and Incorporated by Royal Charter in 1893, having for its object the granting of diplomas after examination or *honoris causâ*, and otherwise benefiting organists.

College Youths. A Society of bell ringers established in 1637, at the church connected with the college of St. Spirit and Mary, Upper Thames Street, London.

Col legno (*It.*) With the wood. A direction to strike the strings of a violin with the back or wood of the bow.

Collet de violon (*Fr.*) Neck of a violin.

Collinet. (*Fr.*) A flageolet, so called after a celebrated player of that name.

Colophony. Colofonia (*It.*) Colophonium (*Lat.*) Colophane (*Fr.*) Resin. The gum used for making the hair of bows rough, so as to set the strings freely into vibration. So called from Colophon in Greece (κολοφωνία, and ῥητίνη, gum.)

Color (*Lat.*) Colour. A term variously employed in mediæval treatises on music to represent: a repetition of a sound in part music (repititio ejusdem vocis); purity of tone (pulchritudo soni); a movement of the voice from the part (florificatio vocis); an alteration of rhythm by different voices (idem sonus repetitus in tempore diverso a diversis vocibus); a discord purposely introduced for the sake of variety (aliquando unus eorum ponitur in discordantiam propter *colorem* musicæ). Some have gathered from the definition—"Repetitio diversæ vocis est idem sonus repetitus in tempore diverso a diversis vocibus," that a musical *canon* is meant to be described. (2) The coloured lines first used for the purpose of rendering *neumes* more intelligible. "Quamvis perfecta sit positura neumarum, cæca omnino est et nihil valet sine adjunctione literarum vel *colorum*" (Guido). [See Clef and Notation.]

Coloratura (*It.*) Divisions, runs, trills, cadenzas, and other florid passages in vocal music.

Coloscione or **Colachon.** A species of guitar, called also Bichordon or Trichordon, according as it was strung.

Come (*It.*) As, like; *e.g.*: *Come prima*, as at first; *Come sta*, as it stands; *Come sopra*, as above.

Comes (*Lat.*) An answer. [Dux.]

Comic Opera. An opera the incidents and dialogue of which are humorous. Comic musical stage-performances date back to the Middle Ages, and probably have an ancient origin; but the comic opera and *opera buffa* first appear in the latter half of the seventeenth century, France being their birthplace. *Opéra comique* was by no means identical with *comic opera*, it was one branch of evolution, the *grand opera* being the other; the former being chiefly distinguished from the latter by having spoken dialogue, it was thus nearly allied to the German *Singspiel*, and often was full of tragical incident. [Opera.]

Comic Song. A song developing in humorous verse some ludicrous idea or incident set to a tune already popular, or to a melody easy to be caught up by the hearers. Some have a chorus or refrain in which all the party joined; others have short *spoken* recitations between the stanzas or lines. Although humorous songs are found in nearly all old English song-books, they were not collected into "comic song-books" until the latter part of the eighteenth century. They offered a wide field for the caricaturist, and some of the best efforts of the Cruickshanks and of Th. Jones illustrate them.

Comma. The small interval between a major and a minor tone, that is between a tone whose ratio is 8:9 and one whose ratio is 9:10. The ratio of a comma is therefore 80:81. A Pythagorean comma is the difference between the note produced by taking 7 octaves upwards and 12 fifths.

Common chord. A note accompanied by its major or minor 3rd and perfect 5th. [Harmony.] In thorough bass, the figure 3, a sharp, flat or natural, as the case may be, or the absence of any letter, character, or figure, denotes the common chord of the bass note. When there is more than one chord on the same bass note, the common chord is figured $\substack{5\\3}$.

Common or **Duple time.** Time with two beats in a bar, or any multiple of two beats in a bar. The beats may be of the value of any note or rest or compound of notes and rests, providing the sum required by the time sign be exactly contained in each bar. Common time is of two kinds, simple and compound. Simple common time is that which includes four beats in a bar, or any division of that number, or square of the number or its divisions. The signs used to express simple common time are the following: $\substack{2\\1}$, $\substack{2\\2}$, $\substack{2\\4}$, $\substack{4\\4}$, $\substack{4\\8}$, and the characters 𝄴 and 𝄵. In these signs the upper figure denotes the quantity of notes required in the bar, and the lower figure the quality of the

Digitized by Google

notes, 1 signifying a semibreve, 2 a minim, 4 a crotchet, 8 a quaver, and so on, each figure showing the relative proportionate value to the semibreve which is now reckoned as the time-standard. The sign (C) is called the sign of alla cappella time, and is usually followed by four minims in a bar, played or sung in slow time; the sign (₵) is called the sign of alla breve time, and has also four or eight minims in a bar played or sung in a shorter time, as its title implies. The use of words directing the pace in which pieces of music are intended to be taken, has created a certain amount of confusion in the use and meaning of all the time signs descriptive of form in a bar [Expression, Time]. Compound common-time is expressed by the signs $\frac{6}{4}$, $\frac{6}{8}$, $\frac{12}{8}$, such signs meaning two or four beats of three crotchets or quavers to each beat.

In mediæval music a circle O was used to indicate what was called perfect time (tempus perfectum), a portion of the circle being omitted C showed that the time was imperfect, a line through the latter sign ₵ meant that every note should have half the value of the same in C. When these signs were reversed they implied that double diminution should take place; every note a quarter of C. Thus the degree of rapidity would be shown by the time signs arranged as follows: O C Ɔ ₵ Ɔ̸

Comodamente (*It.*) *Lit.* in a convenient manner. Easily, quietly.

Comodo (*It.*) Easily, at will, without haste.

Compagnia del gonfalone (*It.*) An ancient society of mystery or miracle play actors established at Rome, in 1264, who illustrated their dramatic performances of sacred subjects with music. They took their name from the banner (gonfalone) which they bore. Their performances are supposed by some writers to have suggested the Oratorio.

Company of Musicians. The Musicians' Company of the City of London was established by letters patent under the great seal of England, on April 24, in the ninth year of the reign of Edward IV. (1472-3). The company was instituted as a perpetual Guild, or Fraternity and Sisterhood of Minstrels—a minstrel being a musician *qualified* to sing or play in public. A new charter was granted by James I., on July 8, 1604. The Musicians' is the only city company for the exercise of a profession.

Compass. The whole range of sounds capable of being produced by a voice or instrument.

Compiacevole (*It.*) Pleasant, agreeable, charming.

Complement. The interval which must be added to any other interval, so that the whole shall be equal to an octave; *e.g.*, the complement of a 3rd is a 6th; that of a 4th, a 5th; of a 5th, a 4th; and so on. It will be seen that the intervals are always considered as overlapping.

Compline (from the Lat. completorium). The short evening service which *completes* the day-hours.

Composer. (1) An author of music. One who "finds out musical tunes." (2) An inventor and arranger of a series of changes in bell ringing.

Composition. (1) A piece of music, for voices or instruments, or a combination of both effects, constructed according to the rules of art. (2) The art of composing music, guided by scientific rules. (3) In an organ, the particular combination of sounds which form a compound stop. (4) A mechanical arrangement on the organ by which certain combinations of stops may be employed or not, at the wish of the performer, upon his opening or closing a valve, or by using a pedal which acts upon the sliders.

Composizione (*It.*) A composition.

Composizione di Tavolino (*It.*) Table music. Convivial compositions, *c.f.* Ger. *Liedertafel.* [Chamber music.]

Compound intervals. Intervals greater than an octave, as opposed to *simple* intervals which are less than an octave.

Compound Stops. Organ stops having more than one rank of pipes.

Compound Times. Times in which the bar is divided into two or more *groups* of notes, *e.g.*, $\frac{6}{8}$ which consists of two groups of three notes; $\frac{9}{8}$ which consists of three groups of three, &c. Compound Times are classified as duple or triple, according to the number of groups in each bar, not according to the number of notes in each group; *e.g.*, $\frac{6}{8}$ is a duple time; $\frac{9}{8}$ a triple time; $\frac{12}{8}$ (four groups of three) a duple time, &c. The principal accent falls on the first note in each bar, and a subordinate accent on the first note of each group.

Comus (*Gk.* κῶμος). A revel, carousal, merry making with music and dancing. The revellers paraded the street crowned, carrying torches, and sang verses in praise of the gods or the victors in the games.

Comus. (*Gk.* κομμός). A mournful song sung in alternate verses by an actor and a chorus in the Attic drama.

Con (*It.*) With; *e.g. con amore*, with affection; *con moto*, with spirited movement; *con sordini*, with the mutes on, &c. (See the words to which it is prefixed.)

Concento (*It.*) Harmony.

Concentus (*Lat.*) (concinno). Musical

Digitized by Google

harmony. Part music; *e.g., concentus vocis Lyræque.* Consonance; *e.g., concentus tubarum ac cornuum.*

Concert. (1) A performance of music in which several executants are employed. Concerts of music, to which the general public is admitted by payment, are of comparatively recent origin in the history of music. Public musical performances, more or less connected with state or religion, were anciently given from time to time, on occasions of importance. Kings, nobles, and civic officials, employed musicians in their trains, but their performances could scarcely be considered in the light of concerts. Organized bands of musicians who performed in the houses of the great and wealthy; "waits" and "noises" are frequently mentioned in old records, but concerts of music in hired houses, assembly rooms of taverns, &c., apart and distinct from the "entertainment" ordinarily provided at hostelries are rare before the time of Charles II. Pepys, in his Diary, speaks of "musick meetings" and "concerts," but they were private affairs, and therefore not within the meaning of the term as now understood. The first public concert in England was given at Oxford, in the year 1670, the first in London two years later. After which a periodical concert was established in Aylesbury Street, Clerkenwell, over the shop of Thomas Britton, the musical small-coalman. Before this time musicians roved from tavern to tavern, instruments in hand, waiting the pleasure of the guests "if they were willing to heare any musick." These bands of fiddlers played by the hour together such popular tunes as were best calculated to delight audiences gathered *impromptu.* But from all that can be learned, their performance was not scientific; "for the most part it was that of violins, hautboys, and trumpets, without any diversity of parts, and consequently in the unison."

The advertisement of the first London concert is still extant, and runs as follows:—

"These are to give notice, that at Mr. John Banister's house (now called the Musick School) over against the George Tavern, in White Fryers, this present Monday, will be music performed by excellent masters, beginning precisely at 4 of the clock in the afternoon, and every afternoon for the future, precisely at the same hour. *London Gazette,* Dec. 30th, 1672." From this time forward concerts of all kinds, vocal and instrumental, given not only "by excellent masters" but also by those who cannot with justice be called either "masters" or "excellent," become common enough. It would be both tedious and unnecessary to trace the history of concerts step by step, neither is it to the present purpose to describe in detail the several sorts of concerts which have taken place since that given "over against the George Tavern." It may not, however, be uninteresting to state that the word has been applied to the performance of oratorios in church as the following quotation will show:

"The Oratorios for the opening of the elegant Organ now erected in the Minster, at Beverley, will be on the 20th, 21st, and 22nd of September, 1769, viz:

"On Wednesday the 20th, the Sacred Oratorio of the Messiah.

"On Thursday, the 21st, the Oratorio of Judas Maccabæus.

"On Friday the 22nd, the Oratorio of Samson; and that being the Anniversary of the King's Coronation, the Performance will conclude with Mr. Handel's grand Coronation Anthem.

"The first violin by Mr. Giardini. The principal voices by Mrs. Hudson, of York, Miss Radcliffe, Mr. Norris, and Mr. Matthews, both of Oxford. The remainder of the band will be numerous, and will consist of the best performers, vocal and instrumental, that can be procured.

"Tickets for the Great Aisle at 5s. each. Galleries at 2s. 6d. each, to be had of Mr. Hawdon, organist, of Mr. Norris, at the Bell and of Mrs. Todd, at the Tyger, in Beverley; of Mr. Forster, carver, in Salthouse Lane, and of Mr. Ferraby, Bookseller, in the Butchery, in Hull. Of whom may be had Books of the Oratorios, with Mr. Handell's Alterations and Additions, as they will be performed at Beverley. Price 4d. each. The North doors will be open'd at Ten in the Morning, and the Concert to begin at Eleven. The Great Aisle will be fill'd with Benches. And to add to the solemnity of the performance, the singers will be dress'd in surplices.

"All tickets transferable. No Money taken at the Door.

"A Concerto upon the Organ each Day. And Mr. Giardini will oblige the company with a Solo.

☞ "An Assembly on Wednesday and Friday."

(2) (*Ger.*) A concerto. [Concerto.]

Concertante (*It.*) (1) A composition suitable for performance at a concert. (2) A composition in which several of the parts are in turn brought into prominence.

Concerted Music. Music for two or more performers, either vocal or instrumental, as opposed to vocal or instrumental solo, with or without accompaniment by a single instrument.

Concertina (*Eng.*) A portable musical instrument of hexagonal form, invented by Professor Wheatstone, consisting of a series of vibrating metal reeds acted upon and set

Digitized by Google

in motion by the current of air, caused by a bellows placed in the body of the instrument connecting the two ends in which the metal tongues or springs are fixed and worked by the player, both hands being in such a position that the wrists move the bellows while the fingers are free to press the stops or keys which cause it to sound. The compass of the concertina is of three and a half octaves with intermediate semitones from fiddle G:

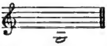

Each note in this scale is double, that is to say, is capable of being produced by the inspiration or respiration of the bellows. [Accordion.]

Concertina (*Ger.*) An instrument of shape similar to the English concertina, but of less finished appearance and more limited compass. The bellows excites the vibration of the free metallic reeds as in the English concertina, but the scale instead of being double is single, that is, the respiratory note is different to the inspiratory note, and has only those chromatic notes necessary for the modulation of melody into the tonic or dominant of the scale in which the instrument is tuned. There is also an escape valve to allow the passage of superfluous air, a contrivance not necessary on the English concertina, where the notes are of double sound. The German concertina is capable of being performed only in the one key in which it is tuned, the English concertina can be played in any key.

Concertino (*It.*) (1) The principal instrument in a concerto, as *violino concertino.* (2) The diminutive of concerto.

Concerto (*It.*), **Concert** (*Fr., Ger.*) A name first given to a vocal composition [concerto da chiesa], but afterwards to one in several movements (generally three) for a group of instruments, string or wind, or both. In its earliest form it was more what we should call a concerted piece, than one intended for a soloist accompanied by a band. This special view of it seems to have gradually grown out of the earlier, and then the predominating instrument was named, *e.g.*, *oboe concerto*, *organ concerto*, &c. The earliest use of the word in this sense is in a work by Scipio Bargaglia, "Trattimenti ossia divertimenti da sonare" (Venice, 1587), but the evolution of the concerto proper did not take place for another century. Torelli (1683–1708) must have the credit of leading the way for the *concerti grossi* of Corelli, Geminiani, Vivaldi, and others. The *concerti* by Handel and Bach must be placed, with certain limitations, under this category. It was Mozart to whom we owe the concerto as now understood. Though essentially symphonic, it exhibits modifications of this type, both in outline and detail. [See under **Form**, p. 181.] In many important respects Beethoven varied the conventional form, by giving more importance to the orchestral portion, and dropping the opening orchestral enunciation of both principal subjects, &c. By allowing the soloist opportunities of introducing *extempore* passages, Handel may have led the way to the cadenza so characteristic of the modern concerto.

Concerto da chiesa. (1) A church, or sacred concerto. Viadana (1612-1620) gave the name to a composition for voice *solo* with thorough bass, and also to vocal compositions in two, three, or four parts with organ accompaniment, to which other instruments were added later. (2) An instrumental piece in a more serious style than a "Concerto da camera" and consisting of one movement.

Concerto spirituale (*It.*) Concert spirituel (*Fr.*) A concert formed of a miscellaneous selection of vocal and instrumental pieces with words of a sacred character. The concerts spirituels in Paris were founded in the year 1725.

Concertmeister (*Ger.*) The leader of the band, the conductor.

Concertspieler (*Ger.*) A performer; a solo-player; the player of a concerto.

Concertstück (*Ger.*) A concert piece, a concerto.

Concha (*Lat.*) A trumpet in the conventional form of a shell fish; Triton's horn; a conch.

Concitato (*It.*) Moved, disturbed, agitated.

Concord. [Harmony.]

Conductor. (1) A director or leader of an orchestra or chorus. It is supposed that a leader or a fugleman was employed by the Assyrians, to regulate the rhythm of the songs or dances; he was armed with two sticks, one of which he beat against the other, and so marked the time or accent.

Among the Greeks the Coryphœus or exarchus led the dance, and in everything requiring united action, a leader or conductor by his voice or certain understood gestures secured the desired result. The word in connection with music has several applications. It signifies one who directs with a bâton the performance of a band of players. It is also applied to one who accompanies vocal or instrumental pieces on the pianoforte. A conductor, as an independent time beater, was not known until the end of the last century. The player who sat at the harpsichord gave the time to the leader of the band, who, directing his subordinates, was called conductor. [Orchestra.]

(2) The inventor or leader of a chime, or

change in bell ringing, is also known as the conductor or composer.

Conductus (*Lat.*) The name given to a certain vocal composition in parts, in the 13th and 14th centuries. It has been variously described as a composition having descant on an *original* melody (qui vult facere conductum, primum cantum debet invenire pulchriorem quam potest, &c., *Franco of Cologne*); on an original or borrowed theme (conducti sunt compositi explicabilibus canticis decoris cognitis vel inventis, &c., *Walter Odington*). The definition of John of Garland points to an elaborate construction: "conductus autem est super unum metrum, multiplex consonans cantus qui etiam secundarias recipit consonantias." . . . "In florificatione vocis fit color ut commixtio in conductis simplicibus." Conducti were sometimes sung without words. They were called simple, double, triple, or quadruple, but the real distinction between the different kinds cannot be clearly learnt from the old treatises. All the information which can be brought together will be found in Coussemaker's *L'Art harmonique aux XII^e^ et XIII^e^ siècles.*

Cone Gamba. [Bell Gamba.]

Confrérie de St. Julien. A Society of Musicians in France, at one period possessing great power. At the end of the 13th century, the troubadours ceased to exist, the "courts of love" were closed, the sentiment of the troubadour poetry was no longer enjoyed, and the excesses of the singers and poets were the things for which they were best or worst remembered. Their followers, no longer having poetry as a veil and excuse for their peculiarities, became disreputable, and led a wandering, careless, and shifty life; the sins of their masters were visited upon their devoted heads, and they became Ishmaelites against whom every man's hand was raised, and who had inclination but not power to lift hand against every man. They were no longer welcome for their skill at all times as heretofore. They were no longer free to enter the houses of the great and wealthy "without leave and license previously had and obtained," as they had been when in attendance upon some poet prince or troubadour sovereign. It was, however, necessary that they should live, and those that did not become openly dishonest, "robbers on the king's highway with a gallows at the end of it," sought the means of livelihood in another sphere than that to which they had previously been always welcomed in. The love for music still existed among a lower class of people, and these gladly received the musicians and performers "whose strains had made many a gentle heart beat quickly," as a proof of their own elevation of taste in matters of courtesy and refinement. Gathering courage by the patronage bestowed upon them, and deeming it needful to "sort their humours" to those of the people for whose amusement they exercised their calling, they made a change in their programme and spoke out openly that which had hitherto been conveyed only through a delicate innuendo. Their success was complete, and to save the effect of troubles which might arise from an ungodly enterprise, they became apparently religious, placed themselves under the protection of a patron saint, Julien, Archbishop of Toledo, who after having led a life of vicissitude and vagabondage, died in 662, and became the tutelar protector of all vagabonds except thieves—who were committed to the care of St. Nicholas. There was also another Saint who divided the honours of their devotions, Genesius, a comedian, who was martyred for his Christianity at the end of the 4th century. Under the care of these two patrons, the musicians flourished, and in the year 1330, settled themselves in the good city of Paris, and formed themselves into a guild for mutual protection and support. The title by which they were enrolled was that of the "Compagnons, jongleurs, menestreux or menestriers," and this title sufficiently indicated their position of *companions*, *yokefellows*, and *servants* to the former troubadours. They had sufficient interest left with the friends and connexions of their old masters to obtain "sealed letters" in November 27th, 1331, recognising their position; and they lived together in one street, hence, called *St. Julien des ménétriers.* To this quarter all had to come who desired their services, and as the minstrels became further encouraged, and increased in numbers they assumed a line of conduct which caused William de Germont, Prefect of Paris, to place various restrictions upon them, which were continued with additions and modifications by his successors.

By a decree issued in the year 1393, the pain of imprisonment was visited upon all members of the guild who offended by reciting scurrilous and scandalous verses either in the streets or in the houses of those who hired them. This check caused a division of the society, the one part devoting itself to the practice of tumbling and rope-dancing—these were called *baceleurs*—the other carrying on the music-entertaining business—took to viol playing, and marked a certain progress in their art, by the introduction of bass instruments hitherto not used by them. They distinguished themselves by the title of "*Ménestrels joueurs d'instrumens tant haut que bas.*" By this name they were recognised by Charles VI. who gave them letters patent dated April 14th 1401. Armed with this document they elected a chief called "Roi des ménestrels," and they

Digitized by Google

built and endowed a chapel in the Rue St. Martin, as much to conciliate the ecclesiastical powers, as to mark their attachment to religious forms. They had a monopoly of all music in France, especially in Paris; no one could learn an instrument without employing a member of the confraternity, no one could give a banquet of music without the leave and license of the Confrérie de St. Julien. Even the king, not to speak of the mayor of Paris, was compelled to be indebted to them for the after dinner amusement of his guests, as well as for the means of giving brilliancy to pageants, processions, and other state businesses. The Confrérie were all-important in the matter of vocal and instrumental music; they had the monopoly of the court and municipal music of Paris until the reign of Louis XIV., who in 1658 was weak enough to confirm their charter and privileges. These privileges must have been great, for they allowed the title of musician to be possessed by those who had funds sufficient to purchase fellowship in the guild: musical skill was of no import, for neither city or court cared for the performances of the Confrérie, though they were compelled to pay handsomely for that which they would rather have dispensed with. If the cultivation of true musical science had been the object in the maintenance of the guild, musical composition from the 14th to the 17th centuries would have been less of a puzzle and annoyance than it is; but unfortunately for the body, in course of time, pressure from without was brought to bear, and the whole thing collapsed, though not without a struggle. From the date when attention was drawn to it for its inefficiency and incompetency, to the day when it finally ceased to exist, no less a period than 100 years elapsed. Louis XIV. in 1660 happened to hear a piece of music by a rising composer called Jean Baptiste Lully, and thinking that it was exactly the sort of lullaby or réveil he should like to hear performed in his own palace, desired to have it executed by his own court band, but, alas! they were musicians only in name, and this little request gave them trouble. But they plucked up courage; the king was informed that they held their places by prescriptive right, and as it had not been the custom for the "Musiciens du Roi," to exercise the art which they professed and were paid for, for many generations, the king's request was unreasonable and unconstitutional. The confraternity would still enjoy their privileges and emoluments and the king must go without his music. This he was not inclined to do; he was therefore pleased to command and ordain that Lully should organise a band, himself at the head, and in order to avoid collision with the patent place-holders and privileged inefficients, this new band of practical musicians, these four and twenty were to be called "Petits violons du Roi," the king's little fiddlers, instead of the king's musicians, a very nice and comforting distinction. By degrees the income arising from the property possessed by the confraternity of St. Julien was applied to a proper purpose, the king's chamber music was executed by legitimate performers, and the perverted association was finally suppressed in 1761, after four hundred years of profitable but comparatively useless existence.

Congregational music. Music in which the people or congregation take part, as opposed to that which is sung by the trained choir alone. The plain-song of the Responses, Creeds, and of the Lord's Prayer; and the melody of psalm and hymn tunes are congregational music, but services and anthems are specially set aside for performance by the choir, acting as it were as the skilled representatives of the listening and meditating people.

Conjunct. (1) One of the Greek systems of music. [Greek Music.] (2) Conjunct motion, a succession of sounds proceeding by single degrees.

Consecutives. A forbidden progression of parallel fifths or octaves, *e.g.*:

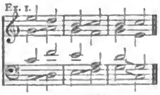

Consecutive fifths.

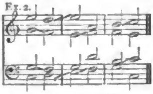

Consecutive octaves.

Consecutives are considered to be *saved*, if they do not occur between the *same two parts*, as shown in the following (Ex. 3), which embodies the harmonic progressions of Ex. 1:

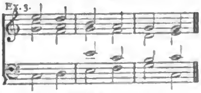

That consecutive fifths are often productive of an ugly and distressing effect, is not to be denied; but their use when not objectionable seems to have been somewhat thoughtlessly forbidden by musical law-givers. The great masters not unfrequently use them with good results, as will be seen by the following examples. (Stainer's "Treatise on Harmony.")

BACH. "Motett," No. 2.

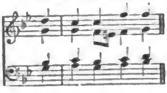

HANDEL'S "Solomon." "Almighty power."

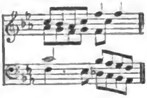

Mendelssohn's "St. Paul."
"To God on high."

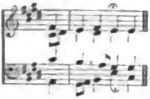

Mendelssohn's "St. Paul."
"To Thee, O Lord."

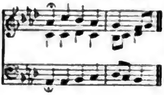

Spohr's "The Last Judgment."
Introduction to Part III.

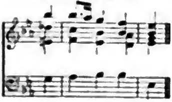

Haydn's Symphony, No. 4.

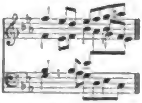

Mozart's Symphony, No 4.

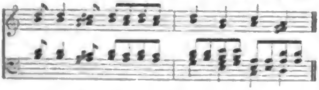

Beethoven's Pianoforte Sonata, No. 1, Op. 29.

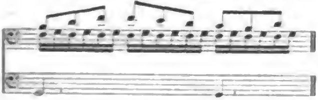

It having been found by experience that the early attempts at harmony known as *diaphony* and *organum*, which consisted entirely of consecutive quarts, quints, and octaves, were remarkably unpleasant and barbarous, it is probable a reaction took place which led to the indiscriminate condemnation of consecutive fifths.

No satisfactory reason has yet been brought forward for the unpleasant effect of consecutive fifths, but it is easy to see why consecutive *octaves* should not be allowed in pure part-writing. If in a duet, or trio, for instance, one part occasionally moves in unison or octaves with another, such a part is for the time *wasted*, and loses its power of forming a distinct melody or adding a real contribution to the harmony. But, on the other hand, any phrase, however short, may be legitimately enforced or strengthened by doubling at the unison or octave. Thus, the bass may be doubled by the left hand or the melody by the right hand, in pianoforte or organ music, and voices may join in unison whenever it is thought desirable. So, too, in writing for a full band, a theme or motive may be enforced by the combination of any instruments, at the unison, octave, or double octave; but, instruments playing in parts are not allowed to combine in such a manner unless for such a purpose.

Consecutives are said to be ***hidden*** when the progression of two parts gives the impression that they have occurred, although they have not actually been written, *e.g.*:

Ex. 4.

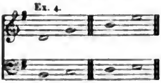

Hidden fifths. Hidden octaves.

They are to be discovered by filling up the interval of one of the parts with the intermediate scale-series, *e.g.*:

Ex. 5.

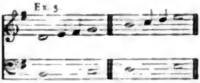

The law against hidden fifths and octaves has never been directed against the movement of inner parts, but only against that of *extreme* parts. But even this has been greatly relaxed in modern music, *e.g.*:

Ex. 6.

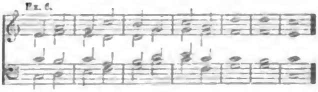

Consecutives are said to be "by contrary motion" when the parts forming them proceed in opposite directions, *e.g.*:

Schumann's "Luck of Edenhall."

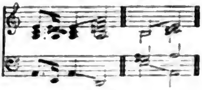

Consequent. *Consequenza (It.)* The answer to a fugue-subject or any subject proposed for imitation. [Dux.] [Guida.]

Conservatorium (*Lat.*) **Conservatorio** (*It.*) **Conservatoire de musique** (*Fr.*) A public music school.

The Italian academies or conservatories are the most ancient, and were formerly attached to hospitals and other benevolent institutions, and were intended for the education of the poor and fatherless, or orphans. Education, board and clothing, were dispensed without cost to both male and female pupils. The conservatorio at Milan was founded in 1808, and the advantages of the instruction of the professors is sought and obtained by many who are not eligible for the foundation, but who can obtain the benefit of a superior musical education at a small cost.

Digitized by Google

The first school of the kind was established in France, in 1784, with the title of "L'école royale de chant et de déclamation." A national institute was started during the period of the Revolution, to supply the want of musicians in the army; this institute became the present Conservatoire de musique, for the support of which the government is charged with an annual sum of 140,000 francs. Many of the masters have been, and still are, men famed throughout the world for their practical skill and their success in teaching, and the text books used at the conservatoire are among the standard works of reference in their department. There are also establishments of a similar character at Brussels, Cologne, Prague, Warsaw, Vienna and Leipzig (the last named established under the direction of Mendelssohn), Berlin, Paris, St. Petersburg.

Consolante (*It.*) In a consoling, comforting manner.

Consonance. Consonanz (*Ger.*) [Harmony.]

Consonant. Concordant.

Consonant interval. [Interval.]

Con sordini (*It.*) (1) With the mutes on. (2) With the soft pedal at the pianoforte held down.

Consort. (1) A consort of viols was a complete set, the number contained in a chest, usually six. [Chest of Viols.] (2) The sounds produced by the union of instrumental tone.

Consort. (1) To sound together, to form agreeable sounds by combination. (2) To form a concord.

Con stromenti (*It.*) With the instruments.

Continued Bass. [Figured Bass.]

Continuo (*It.*) [Figured Bass.]

Contours. Conteurs (*Fr.*) [Trouveur.]

Contra (*It.*) Against. In compound words this signifies an octave below, *e.g.*: *Contragamba*, a 16 ft. gamba; *Contrabbasso*, a double bass; *Contrafagotto*, a double bassoon.

Contra-bassist. A double-bass player.

Contrabbasso (*It.*) [Double-bass.]

Contraddanza (*It.*) [Country-dance.]

Contrafagotto (*It.*) [Double bassoon.]

Contralto voice. The voice of deepest tone in females. It is of a quality allied to the tenor voice in men, and the usual compass is within two octaves. The best notes of the range are between G or A flat below, and C or D above:

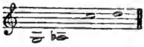

The notes above these are of a somewhat harsh and forced character, those below of little musical value. In most contralto voices there is a break varying between C sharp and A flat in the lower part of the register, and the careful adjustment of the two qualities of tone above or below this break is one of the chief qualities of good contralto singing. [Alto Voice.]

Contraposaune. An organ stop of 16 ft. and 32 ft. pitch.

Contrappuntista (*It.*) A writer on, or a composer of counterpoint.

Contrappunto (*It.*) [Counterpoint.]

Contrappunto alla mente (*It.*) Impromptu counterpoint. [Alla Mente.] [Chant sur le livre.]

Contrappunto doppio (*It.*) Double counterpoint. [Counterpoint.]

Contrapuntal. Belonging to counterpoint

Contrapuntist. A writer on, or a composer of counterpoint.

Contr'arco (*It.*) False or incorrect bowing on the violin, &c.

Contrary motion. Melodies or chords proceeding in opposite directions. [Motion.]

Contrassoggetto (*It.*) [Counter subject.]

Contra tempo (*It.*) Against time. (1) The part progressing slowly while another is moving rapidly:

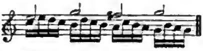

(2) Syncopation.

Contra-tenor. [Alto.]

Contratöne (*Ger.*) Deep tones of the bass voice.

Contraviolone (*It.*) Double-bass.

Contre-basse (*Fr.*) Double-bass.

Contredanse (*Fr.*) [Country dance.]

Contre-partie (*Fr.*) Counterpart, opposite. The entry of a second voice with a different melody, making harmony with the first.

Contre-point (*Fr.*) Counterpoint.

Contre-pointiste (*Fr.*) Contrapuntist.

Contre-sujet (*Fr.*) [Counter subject.]

Contre-temps (*Fr.*) Against time. Syncopation.

Conversio (*Lat.*) Inversion.

Convict of music. An institution for musical instruction in Leipsic; from Lat. *convictus (convivere)*, living together, social intercourse. [Conservatorio.]

Coperto (*It.*) Covered, concealed. *Timpani coperti*, muffled drums; *quinti coperti*, concealed fifths, hidden fifths.

Copula (*Lat.*) In mediæval music a free use of slurred running notes in descant.

Copyright. *Definition.*—The word itself conveys the best definition of the right which it denotes. Copyright is the "right to copy," and includes the right to prevent others from copying. The Act of Parliament (5 & 6 Vict. c. 45), which is the fountain-head of the present law on the subject, defines the right to be "the sole and exclusive liberty of

Digitized by Google

printing or otherwise multiplying copies of any subject to which the said word is herein applied." Two points are to be noticed in this definition—one affecting *the object* and the other *the method* of the multiplication of copies:

i. The mere multiplication, whatever be the object in view, is, if unauthorised, an infringement of the right. It matters not whether the multiplication be for the private use of the individual or, let us suppose, for the general use of a choir or choral society, or whether it be for purposes of sale. The "sole and exclusive" right of the proprietor of the copyright is equally prejudiced, whatever be the reason for the creation of the unlawfully produced copies. (Novello & Co. *v.* Sudlow.)

ii. Any method of multiplication of copies, if unauthorised, is an infringement of the right; and therefore the copying in manuscript of any copyright music, or the writing of it on a blackboard is unlawful, and frequently causes serious damage to valuable property. The transposition of copyright music into a different key is an unlawful copying. The melody protects all keys. On the question of manuscript copying the reader is referred to an article in the *Musical Times* of September, 1889.

Copyright is of two kinds—(i.) Copyright at Common Law and (ii.) Statutory Copyright.

i. *Copyright at Common Law* relates solely to unpublished works, and is more properly an incident of the general law of property which protects a man in the enjoyment of what is his own. So long as an author keeps his works to himself unpublished they remain his absolutely and for ever; but directly he publishes his work his Common Law rights cease, and thenceforth what protection he has is conferred upon him by Statute.

ii. *Statutory Copyright.* In return for the special privileges and protection afforded by Statute to copyright property the period of its duration is limited by the Legislature.

The duration of Copyright in Books and Music. The Statutory period is an alternative one—viz., forty-two years from the date of publication, or for the life of the author and seven years after his death, whichever period happens, in each particular case, to be the longer. That is to say, copyright *must* last for forty-two years, and it *may* last for a longer period if the author happens to live for more than thirty-five years after his work was published. This longer period of course cannot be determined until the author is dead. To take an example: A composer publishes his first work when he is 20 years of age and dies aged 80. His first work will enjoy copyright for sixty-seven years, and the period of copyright enjoyed by his works of later date will diminish, as shown in the following table, until the minimum period of forty-two years is arrived at:

TABLE SHOWING THE PERIOD OF COPYRIGHT ENJOYED BY A COMPOSER WHO DIES AGED 80 YEARS.

Age at date of publication.	Period during which copyright will last.
20 years	67 years.
21 „	66 „
22 „	65 „
23 „	64 „
24 „	63 „
⋮	⋮
44 „	43 „
45 „	42 „
46 „	42 „
47 „	42 „
48 „	42 „
49 „	42 „
⋮	⋮
80 „	42 „

N.B.—All posthumous works enjoy copyright for forty-two years from date of publication.

Commencement of the right. Statutory copyright dates from the first publication of the work, and when once the period has commenced to run nothing can stop it until, by expiration of the limited time, it dies a natural death. Each new edition of a copyright work, however, provided it contains new matter, or alterations, or re-arrangement, constitutes a new work, and in its revised form is treated as a newly published work in all respects for copyright purposes. The copyright of the work *in its original form* is obviously not prolonged by any additions to it or by any alterations or re-arrangement.

Publication. What constitutes publication has never yet been settled by Statute; but for the purposes of this article it may be defined as any method of offering a work to the public in the form of multiplied copies. In the vast majority of cases this is effected by offering the copies for sale, but a gratuitous distribution of copies, even amongst friends, may amount to publication, although no copies may have been, or will be, offered for sale.

Nature of the property. Copyright is personal property, and can only be assigned in three ways: 1. By writing (not necessarily by deed); 2. By transfer at the Registry at Stationers' Hall as regards works already entered there; 3. By operation of law—*e.g.*, by will, intestacy, bankruptcy, &c. In one case only can copyright be assigned by act of parties without a written document of some kind, and that is the case where one party engages another to write for him on the understanding that the work produced is to

Digitized by Google

be the copyright of the employer. It is essential, however, that the party employed be actually paid for the work. The copyright does not pass till the money is paid. Even in this case, however, a written document is practically necessary, if only for the purpose of registering and rendering certain the terms of the employment: but, apart from any question of a contract of employment, a purely verbal agreement for the transfer of copyright property is of no value whatever.

It is very desirable, if not necessary, that all licences granted by a copyright proprietor for the reproduction of copyright property should be in writing.

As regards evidence, the title to copyright property is similar to the title to land—that is to say, the holder for the time being must be in possession of each document relating to the property, from the original assignment by the author to the ultimate assignment to himself.

Entry at Stationers' Hall. It is commonly but erroneously supposed that entry at Stationers' Hall alone can confer copyright. This, however, is not the case. Statutory copyright in the British Dominions exists as the natural consequence of the mere act of publication, and no entry at Stationers' Hall is imperative, or incidental to the origin of the right. Indeed, a work cannot be entered at Stationers' Hall until it is already copyright—*i.e.*, published. Entry at Stationers' Hall is effected by entering under appropriate columns in a book kept for the purpose the following particulars of the work entered:—1. Title of work. 2. Name of publisher (*i.e.*, the first publisher) and place of publication. 3. Name and place of abode of the proprietor of the copyright. 4. Date of first publication. 5. Date of making the entry. A fee of five shillings has to be paid for each work entered. Entry at Stationers' Hall is merely a formality which must be complied with before any legal proceedings to protect a copyright can be commenced. It therefore concerns not copyright as a legal and statutory creation, but rather the legal protection of that creation.

Copyright may be either (i.) *British* or (ii.) *International*. The former relates to all works published for the first time in British Dominions, the latter to works published for the first time in such countries as have entered into treaties on the subject with Great Britain.

(i.) To obtain *British copyright* it is essential that the first publication take place on British soil. Until 1886 a work first published in Great Britain was by the act of publication made copyright throughout all British Dominions, and even a foreigner by first publishing in Great Britain could obtain British copyright quite irrespectively of any international treaty. But a work first published in a British Colony could enjoy no copyright except locally under the Copyright Laws, if any, of the Colony where the publication took place. It could claim no copyright in Great Britain. In 1886, however, an Act (49 & 50 Vict. c. 33) was passed which remedied this defect, and since that date publication in any colony or portion of the British Dominions has been sufficient to give the author a copyright which is protected throughout the entire British Dominions. It is even now a matter of some doubt whether it is not essential for the acquisition of British copyright that the author of the work be actually within British Dominions at the moment of publication.

Among the privileges enjoyed by the copyright owner may be enumerated the following:

1. The right to reproduce his work himself, and to prevent anyone else from reproducing it in any form or for any purpose.
2. The right to assign his copyright and to grant licences for its reproduction without parting with the copyright.
3. The right to sub-divide his copyright and to assign different portions of it for different colonies or internationally for different countries.
4. The right to transpose or arrange in any form an original melody, and to prevent anyone else from transposing or using his melody in any form.
5. The right to arrange any non-copyright melody and to prevent anyone else from using or reproducing his arrangement of it; although of course he cannot prevent anyone else from making an independent arrangement of the same non-copyright melody.
6. The right to prevent the importation into Great Britain of any piratical copies printed in any portion of the British Dominions, or foreign copies printed outside of British Dominions.
7. The right to prevent the sale of such copies so unlawfully imported.

Translation. Assuming the right to translate, there can be no question that the Law confers every copyright privilege upon the work produced by the translator. But it is very doubtful whether the exclusive right to translate a copyright work is one of the privileges enjoyed by the author of the original work: in other words, whether he has the right to *prevent others* from translating his work. No statute deals with the matter, nor has the point been definitely settled by Judicial decision. There are, however, dicta of the Judges which rather militate against the existence of an exclusive right of translation, although it seems monstrous to suppose that an author's

Digitized by Google

rights are limited to the actual words employed by him in giving expression to his ideas, and that the subject matter is common property which may be lawfully reproduced by anybody in any other form or language. Yet on this very ground American decisions have settled the point against the author of the original work; although more recently the American Statute of 1891 has expressly conferred upon authors of copyright works the *exclusive right* of translating those works.

The question of course is seldom likely to arise here as regards works published in the English language: but internationally it is a point of great importance, and internationally the matter is now definitely settled. On March 7, 1898, an order in Council was issued * giving effect to an "Additional Act" to the Berne Convention concluded on September 9, 1897, by eleven out of the thirteen parties to that Convention for the purpose of varying some of its provisions. Clause 3 of the "Additional Act" alters Article V. of the Convention which now reads as follows:—

"Authors belonging to any one of the Countries of the Union . . . shall enjoy in the other Countries the exclusive right of making or authorising the translation of their works during the entire period of their right over the original work. Nevertheless the exclusive right of translation shall cease to exist if the author shall not have availed himself of it, during a period of ten years from the date of the first publication of the original work, by publishing or causing to be published in one of the Countries of the Union a translation in the language for which protection is to be claimed." Prior to this alteration of the Berne Convention there was considerable doubt whether the exclusive right of translation was not limited to ten years from the date of the original publication, *whether an authorised translation had or had not been issued.* This doubt arose through the impossibility of reconciling Article V. of the Convention with Sec. 5 of the English Act of 1886 which gave effect to the Convention. The "Additional Act" adopts the view expressed in the English Act of Parliament.

Remedies for infringement. A Copyright owner has the following remedies:

1. To bring an action for damages.
2. To restrain infringements by injunction of the Court.
3. To bring an action for penalties.
4. To bring an action for delivery to him of the unlawful plates and copies wherever they may be found, even when they may be in the possession of innocent parties.
5. To cause the officers of the Customs to seize and destroy all copies unlawfully imported (see 5 and 6 Vict. c. 45, sec. 17, and the Customs Consolidation Act, 39 and 40 Vict. c. 36, secs. 42 and 152).

It is doubtful whether the proprietor of a copyright can prevent the importation into Great Britain of copies *lawfully* produced in a British Colony—*e.g.*, suppose he has subdivided his right and has assigned his Canadian rights to a Canadian publisher and his Australian rights to an Australian publisher. It is obvious that copies printed in either Colony are lawfully produced. But his right to prevent their importation into Great Britain is far from clear.

The same doubt would arise where a British copyright proprietor, having taken out an American copyright also, sells his American copyright to an American publisher. Of course there is no international copyright treaty between England and America, but internationally the doubt has been removed. In the recent case of Pitts *v.* George, the Court of Appeal decided that *internationally* copies lawfully printed abroad in the country of origin may not, in the case of a divided copyright, be imported into this country.

Limitation of Actions. All actions and claims in respect of any offences committed against the Copyright Law must be brought within twelve calendar months next after the offence was committed. But, of course, each fresh offence constitutes a fresh cause of action, and where the infringement is of a continuous nature, as in the case of a constant sale of unlawful copies, a fresh offence is committed daily. Still, damages cannot be claimed in respect of any sales effected more than twelve months prior to the commencement of the legal proceedings.

Notification of Copyright. The Law of Great Britain does not require any notification of the fact that a work is copyright to be printed upon the copies. Everything is presumed to be copyright until the contrary is ascertained. (Compare *American Law* and *Performing right.* See below.)

Titles. Our Law does not recognise any copyright in the title of a publication. For the sake, however, of protecting the public, the courts will interfere, at the instance of a copyright owner, to prevent his title being used in a way calculated to mislead the public. This protection is given not on the ground of copyright or in the interest of the copyright owner, but on general grounds for the suppression of fraud and in the interests of the public.

British Museum and Public Libraries. A copy of every published work must be delivered on behalf of the publisher at the

---

* See *London Gazette*, March 8, 1898.

Digitized by Google

British Museum within one month of publication in London, or within three months of publication elsewhere in the United Kingdom. This copy must be one of the best edition issued, and must be "bound, sewed, or stitched together." Similarly a copy must be delivered to the officer of the Stationers' Company for the use of each of the following libraries:—The Bodleian at Oxford, the Public Library at Cambridge, the Advocates' Library at Edinburgh, and the Library of Trinity College, Dublin. As regards these libraries, a "best" edition is not required and the copy need not be delivered at all unless demanded in writing within twelve months of the day of publication. If demanded, the copy must be delivered within one month of the receipt of the demand.

It was formerly incumbent upon the publisher of any book first published in any British possession, *but outside the limits of the United Kingdom*, to deliver a copy of the book at the British Museum, and at the privileged Libraries if demanded, within twelve months of publication. But since 1886 this is no longer necessary as regards books published outside the United Kingdom. 49 & 50 Vict. c. 33, sec. 8 (i.), Proviso (*b*).

(ii.) *International Copyright*. In 1838 this subject was first dealt with by the Legislature. It was generally recognised about that time throughout Europe that the growth of International intercourse demanded that copyright, like other species of property, should receive international acknowledgment and protection. Accordingly an Act was passed enabling Her Majesty by Order in Council to grant copyright protection throughout British Dominions to works published for the first time in the country named in the Order in Council. This Act only applied to such countries as had granted or might grant a reciprocal privilege to British subjects. The rights conferred were to be the same as those enjoyed by British subjects, and the period of copyright, to be defined by each Order in Council, was not to exceed the period of copyright then allowed by our law to British subjects. In 1844 this Act was repealed, and a new Act was passed for the purpose of extending the term of copyright so as to assimilate it to the extended period enjoyed by British subjects under the present law (1842) and to the remedies of that Act, and for the additional purpose of extending performing rights to foreign publications. Registry at Stationers' Hall within a certain time after the first publication abroad was an essential condition of copyright under the Act of 1844 (compare registration under 5 & 6 Vict. c. 45). A large number of foreign countries thereupon effected Copyright Treaties with Great Britain and obtained the benefits of our Act of 1844. It is still important to remember the provisions of that Act, because, notwithstanding the Berne Convention, all works published abroad and copyrighted in Great Britain prior to December 6, 1887, are still subject to the provisions and conditions of that Act and the Orders in Council issued thereunder, see the "Additional Article" to the Berne Convention; moreover, fresh International Copyright Treaties may still be the subjects of Orders in Council under the Act of 1844.

The Berne Convention of 1886 now regulates international copyright amongst the following countries: Belgium, France, Germany, Great Britain, Hayti, Italy, Spain, Switzerland, Tunis, Luxembourg, Monaco, Montenegro, and Norway. Publication in any one of these countries is sufficient to confer copyright, without registration or further ceremony, in all the other enumerated countries. The country in which the work was first published is called the "country of origin," and the rights conferred are in each case the rights enjoyed by the natives of each country; but the period of copyright enjoyed by a work of Foreign origin cannot exceed in any country the period enjoyed in the country of origin, although, of course, it may be less if the period enjoyed in any given country is less than the period allowed by the country of origin.

The English Act of 1886 (49 & 50 Vict. c. 33) enabled Her Majesty by Order in Council to give effect to the Berne Convention in British Dominions. On November 28, 1887, Her Majesty issued such an Order in Council, which came into force on December 6, 1887, and all works published in any one of the agreed countries since the last mentioned date are regulated by the provisions of the Berne Convention read in conjunction with the Act of 1886 and the Order in Council of November 28, 1887.

By Section 6 of the Act of 1886 the Berne Convention is given a retrospective force—that is to say, all foreign published works issued in any one of the agreed countries which were non-copyright in Great Britain on December 6, 1887, but which would have been copyright had the Berne Convention been in force at the date of their first production abroad, are copyright under the Berne Convention "provided that where any person has before the date of the publication of an Order in Council lawfully produced any work in the United Kingdom, nothing in this section shall diminish or prejudice any rights or interests arising from or in connection with such production which are subsisting and valuable at the said date."

Digitized by Google

By virtue of this section many foreign published works, which might have been copyrighted under the old 1844 Act, but which for want of registration or other cause had become non-copyright in Great Britain, suddenly became copyright in this country, and after having so become copyright, they enjoyed copyright as against everybody except as against anyone who had chanced to reproduce them whilst they were still non-copyright here. For instance, the French publisher of Gounod's "Messe Solennelle" omitted to copyright that work under the Act of 1844, and it consequently became public property in Great Britain. Prior to November 28, 1887, Messrs. Novello & Co. issued an edition of the "Messe" of their own publication, and prior to that date anyone was free to do the same. But on that day the Order in Council was published, and on December 6 following the French publisher became possessed of an English copyright in the work. Consequently from that day no new editions of the "Messe" can be lawfully issued in British Dominions, either by Messrs. Novello & Co. or by anyone else; but Messrs. Novello & Co. had "valuable rights and interests" in their edition, seeing that they had invested a considerable sum in producing it, and *on that ground* they may continue the sale of their copies and may continue to use their plates, and generally may exercise any other rights or interests which arose from or in connection with their original re-production of the work. One of these rights or interests is the right to authorise the public performance of any work from copies protected by the proviso of the section (Moul *v.* Groenings).

On March 7, 1898, the Queen, by an Order in Council, gave her sanction to an "Additional Act" to the Berne Convention for the purpose of varying certain of its provisions. The details of the alterations made are contained in a schedule to the "Additional Act," and are set out at length in the *London Gazette* of March 8, 1898. All the thirteen countries parties to the Berne Convention ratified the alterations except Hayti and Norway. The "Additional Act" is to be construed as if it formed part of the original Convention of 1886.

Although Austria-Hungary is not a party to the Berne Convention, a treaty on practically identical lines with that Convention was entered into between Great Britain and Austria-Hungary on April 24, 1893. The treaty was ratified on April 14, 1894, and an Order in Council giving effect to the treaty was made on April 30, 1894. The Order in Council came into force on May 11, 1894. See *The London Gazette*, May 1, 1894.

American Copyright. Prior to July 1, 1891, no one could hold a copyright in the United States unless he were an American citizen or resident in America. Therefore, no British publications and no works published in America by non-resident British subjects could enjoy copyright in the States. This was a serious drawback to British authors and composers, especially having regard to the common language spoken in both countries. On March 3, 1891, however, the American Legislature altered their law on the subject, and, by removing the restriction as regards citizenship or residence, gave their law an international character by enabling certain foreigners to enjoy its benefits. The Act came into force on July 1, 1891. It only applies, as far as foreigners are concerned, to citizens or subjects of those foreign countries which grant to Americans the benefit of copyright on substantially the same basis as to their own citizens. The existence of those conditions was to be determined by the President of the United States by proclamation. Belgium, France, Great Britain, and Switzerland were at once (July 1, 1891) proclaimed, and later on Germany (on April 15, 1892), Italy (on October 31, 1892), Denmark (on May 8, 1893), Portugal (on July 20, 1893), Spain (on July 10, 1895), Mexico (on February 27, 1896), and Chili (on May 25, 1896).

Under the present law copyright in America lasts for twenty-eight years from the date of recording the title of the work, but the author, if he be living, *or his widow or children* if he be dead, is entitled to have the right continued for a further term of fourteen years, provided all the original formalities are again observed within six months before the expiration of the twenty-eight years, "and such person shall within two months from the date of the renewal cause a copy of the record thereof to be published in one or more newspapers printed in the United States for the period of four weeks." Several questions arise upon these provisions for the extension of the period of copyright. The privilege is reserved for the author if he be living or for the widow or children if he be dead. An assignee of the copyright is not mentioned, and it seems clear that he cannot renew the copyright in his own name. But, in the absence of a special agreement between the assignor and the assignee, it is doubtful whether, assuming the copyright to have been renewed by the author or his widow or children, the assignee can claim the benefit of the renewal: it also seems very doubtful whether he can oblige the author or his widow or children to renew the copyright. And his position would be very awkward if the author died unmarried or if author, widow, and children all died before the expiration of the first period of twenty-eight years. It will be observed that "such persons"—*i.e.*, the author or his widow or children—are to publish a copy of the record in the newspapers, so that, whatever his rights may be, the assignee would be powerless without the assistance of the author or his family.

The formalities prescribed by the American law for securing copyright are:

1. On or before the day of publication in America or in any foreign country a printed or type-written copy of the title of the work must be delivered at the office of the Librarian of Congress at Washington, or deposited in the mail within the United States addressed to the Librarian of Congress.
2. Not later than the day of publication in America or in any foreign country two copies of the work to be copyrighted must be delivered or deposited as in the case of the title.
3. All works the subject of copyright, *except books, photographs, chromos, or lithographs*, may be manufactured outside the limits of the United States, but as regards the four exceptions they must be manufactured and printed within the limits of the United States; and here it should be noted that although under the English copyright law the word "book" includes music, and although under the general copyright law of America it has received the same interpretation, yet under the American Act of 1891 there is a distinction drawn, by the words of the Statute, between "book" and "musical composition," and it has been held (in Novello & Co. *v.* The Oliver Ditson Co., fully reported in the *Musical Times* of September, 1894, and June, 1895) that for the purposes of the Act of 1891 there is no necessity for music to be manufactured in the United States as a condition of obtaining copyright there.
4. A notification that a work is copyright, the year in which it was copyrighted, and the name of the original copyright owner must be printed in the case of books "on the title-page or the page immediately following" and in other cases including musical compositions "upon some visible portion thereof" thus: "Entered according to Act of Congress in the year —— by —— in the Office of the Librarian of Congress at Washington," or, at the option of the person entering the copyright, the words "Copyright 18—— by ——." The shorter form is nearly always used.
5. The following fees must be paid: For recording the title, fifty cents for a citizen or a resident and one dollar for anyone else. For a certified copy of the register, fifty cents.
6. Although the Statute does not specifically require it, yet it is an essential condition of American Copyright Law that the first publication must be American. It has already been remarked that the law of Great Britain requires (as far as British copyright is concerned, as opposed to international) the first publication to take place in British territory. At first sight this would appear to make it impossible to secure copyright for a work both in Great Britain and in America. The law of each country, however, recognises simultaneous publication—*i.e.*, publication on the same day in both countries—as being equivalent to first publication in its own country.
7. Publication must follow the entry of the title within a reasonable time. In Boucicault *v.* Hart it appeared that a title was entered on October 24, and the work was still unpublished in the following February when an action was commenced. The Court held that the publication was made too late to secure copyright.

These seven formalities are *collectively* the condition of American Copyright, which cannot exist unless they are all carefully observed. As has already been stated, publication alone secures copyright in British Dominions; nothing, however, can obtain copyright protection in America until *each* condition has been complied with. Hence the copyrighting of a song in its original key does not apparently protect copies of the song published in other keys unless they also are entered at the Librarian's Office. The song in the transposed key is obviously a new publication, and it is the publication and not the song as an artistic creation which is the subject matter of American copyright. Suppose then that the song is issued in several keys and that one key only is copyrighted, the melody would appear to become public property in America, seeing that it is contained in copies which have been allowed to become non-copyright. Anyone would, it is submitted, be at liberty to reprint the transposed and non-copyright versions, and apparently to make fresh arrangements of the non-copyright melody.

Performing right. This is an independent right and totally distinct from copyright. It is the right to control the public performance of dramatic works and musical compositions as opposed to the right of multiplying copies of such works. It had no existence at Common Law, and was first created, in respect of dramatic works only, by 3 & 4 Will. IV. c. 15. The right was subsequently extended to musical compositions in 1842 by 5 & 6 Vict. c. 45, sec. 20.

The period of duration of performing right is exactly the same as that of copyright. The period commences to run from the date of the first public performance; but, as it is an absolutely distinct right from copyright, it may commence at a different time, and consequently may expire at a different time, and the two rights may belong to different owners. An assignment of copyright simply does not operate as a conveyance of performing right unless the latter right is expressly included in the conveyance. A first performance abroad (unless protected by international treaty) destroys the performing right in this country, but there is no reason to suppose that a first performance abroad destroys *copyright* in this country. The point has never been expressly decided, but, as the two rights are quite distinct, and a first performance in this country has never been held equivalent to publication, it seems difficult to conclude that a first performance abroad, which undoubtedly destroys performing right here, can have any effect upon copyright. Until August 10, 1882, it was not necessary that the reservation of the performing right of a musical publication should be notified in any form whatever, but in that year an Act, 45 & 46 Vict. c. 40, was passed which requires the proprietor of any musical composition *first published after the passing of the Act* who wishes to reserve the performing right to notify the fact by printing on the title-page of every published copy a notice that the right of public representation and performance is reserved. The Act, however, does not attach any penalty to the omission of the notice; consequently the performing right remains unaffected, although it would be difficult to maintain an action to protect it in cases where the requirement of the Act has not been observed. If the performing right and the copyright are held by different persons, provision is made by which the owner of the performing right can compel, under penalties in case of default, the owner of the copyright to notify the fact that the performing right is reserved.

By 3 & 4 Will. IV. c. 15, sec. 2, a penalty of not less than 40/- (or the full advantage gained by the unlawful performance, or the loss suffered by the performing right owner, whichever is the greatest) is imposed for each unauthorised performance of a dramatic work, and by 5 & 6 Vict. c. 45, sec. 21, these remedies are extended to musical compositions. Notwithstanding these enactments, the Act of 1882, although it does not take away the remedies, provides that where the owner of the performing right of *a musical composition* in an action for penalties or damages for an unauthorised performance does not recover more than 40/- as penalty or damages, the costs of the action are to be in the discretion of the Court or judge who tried the action (sec. 4). The object being to put a stop to frivolous actions for the recovery of penalties by throwing the costs of the litigation upon the plaintiff who had suffered no real injury.

By an Act of 1888 (51 & 52 Vict. c. 17) the last-quoted section is repealed, but the Act provides that in actions for penalties or damages in such cases the judge may award less than 40/-, or a merely nominal penalty or damages, and that, *whatever sum be awarded*, the costs are to be in the absolute discretion of the judge. The proprietor of the place where the unauthorised performance took place is exempt from penalties and damages unless he wilfully caused or permitted the performance to take place knowing it was unauthorised. The Act does not apply to operas or stage plays given in a duly licensed theatre or place of public entertainment. (See *Licensing*.)

The American Law does not recognise any performing right in musical compositions except in such as are dramatic in character (see Act of March 3, 1891). The right is there known as "Playright," and music when protected is only protected as being an integral portion of a play.

Performing right is acknowledged and protected internationally by the Berne Convention on exactly the same principles as is copyright, but the author must expressly declare on the title-page or commencement of the work, that he forbids the public performance. The Act of 1886 (sec. 6), which gave its retrospective force to the Berne Convention as to copyright, makes it retrospective also as regards performing right.

Cor (*Fr.*) A horn.

Corale (*It.*) Chorale, hymn or psalm tune. [Hymn Tunes.]

Cor Anglais (*Fr.*) *Corno Inglese* (*It.*) English horn. A reed instrument of the hautboy character, possessing a compass of like extent but of lower pitch. Its scale is two octaves and a fifth from tenor E with the intermediate semitones:

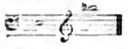

these being the actual sounds produced. The music for the cor Anglais is written in the treble clef, and the instrument transposes the sound a fifth below. Gluck introduced the instrument in his "Orfeo," Meyerbeer has made frequent use of it, and Rossini produces a fine effect in the overture to "William Tell" by means of its tone, but Beethoven only once employed it,—Mozart and Weber never.

Coranach, **Coranich**, **Coronach**, **Cronach** (*Gaelic.*) The word for a funeral song among the Scotch Highlanders; it is said to be derived from *corah-rainach* a crying together. [Keeners.]

Coranto (*It.*) *Courante* (*Fr.*) Current traverse (*Old Eng.*) (1) An Italian form of the country dance. (2) A movement in a suite or sonata of the early writers. The following is given as an early specimen:

Girolamo Frescobaldi, 1591-1640.

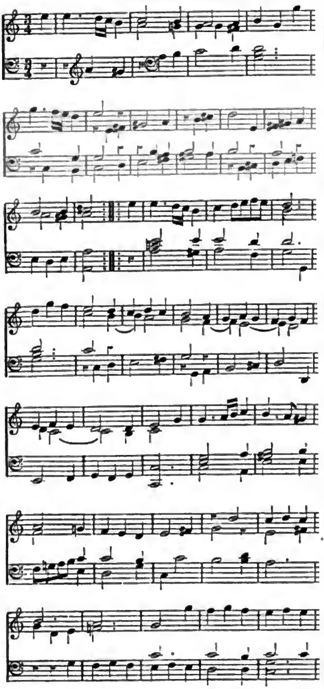

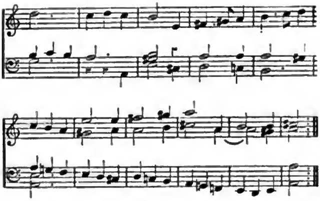

Corda, sopra una (*It.*) *Sur une corde* (*Fr.*) A direction that the passage is to be played on one string. [A una corda.]

Cordatura (*It.*) [Accordatura.]

Cordax (*Lat.*) Κόρδαξ (*Gk.*) An ancient Greek dance of a wanton character, in the old comedy; but sometimes danced off the stage by drunkards.

Corde à jour } (*Fr.*) An open string.
Corde à vide }

Cor de chasse (*Fr.*) A hunting horn.

Corde fausse (*Fr.*) A false string. [String.]

Cor de signal (*Fr.*) A bugle.

Cor de vaches (*Fr.*) Cow-horn, used in many places abroad to call the cattle home, and formerly employed in England to rouse the labourers to their work.

> "No more shall the horn
> Call me up in the morn."

Corifeo (*It.*) [Coryphæus (1).]

Cormuse. [Bagpipe.]

Cornamusa (*It.*) *Cornemuse* (*Fr.*) [Bagpipe.]

Cornare (*It.*) *Corner* (*Fr.*) To sound a horn.

Cornet. *Cornetto* (*It.*) *Zinken* (*Ger.*) An obsolete reed wind-instrument not unlike a hautboy, but larger and of a coarser quality of tone. In this country they were of three kinds, treble, tenor, and bass. The tubes gradually increased in diameter from the mouthpiece to the end, and their outline was gently curved, hence the Italian name *cornetto curvo*. In Germany, as in England, they were once in common use for sacred and secular purposes. They were often made of wood neatly covered with dark leather. [Waits.]

2. A reed stop on the pedals of some German organs, of 4 or 2 feet in length.

3. *Mounted cornet.* A solo stop on old organs, so called because it was placed on a separate sound board, and raised a few feet above the surrounding pipes, for the purpose of giving its tone special prominence. It consisted of several ranks of pipes, generally of five, namely, an open or stopped diapason (usually the latter), a principal, 12th, 15th, and tierce. Thus, if the stop were drawn, and

the finger held on middle C, the following sounds would be heard simultaneously:

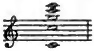

Although these would of course combine into *one*, and not be audible as separate and distinct sounds, yet it may be supposed that such a combination of loud harmonics with a comparatively soft ground-note would produce a most disagreeable and nasal tone. But, notwithstanding its unpleasant *timbre* it was a favourite stop in the last and in the early part of this century, and its general introduction into the best organs gave rise to a vicious and trumpery literature of "cornet voluntaries." The characteristic of these was, that while the left hand held down a soft chord on the choir organ, the right was engaged in *passages*, *turns*, *shakes*, and other musical capers, on the cornet stop of the great organ. The usual compass of the stop was from middle C upwards, but sometimes it commenced at tenor C. A large number of cornet stops were removed to make way for the clarabella when first invented by Bishop, and better taste has so far ejected them that a specimen in good playable condition may be looked upon as a curiosity.

4. *Echo cornet*. A stop often found in swell organs. Originally it consisted of the same series of ranks of pipes as the mounted cornet, but was always of a very small scale. But the name is now often applied to any small-scale sesquialtera or mixture enclosed in the swell box.

5. *Cornet-à-pistons*. A modern brass instrument of the trumpet family, but having valves or pistons by means of which a complete chromatic scale can be produced. In proportion to the number of valves introduced into tube-instruments, the quality of their tone is deteriorated, but notwithstanding this loss of purity and brilliancy, the cornet is most useful and valuable for many purposes. It has been brought into discredit by being unwisely used in some orchestras as a substitute for its parent, the trumpet, with the grandeur of which it cannot compete. [Metal Wind-instruments.]

Cornetto (*It.*) [Cornet.]

Corno (*It.*) [Horn.]

Corno alto (*It.*) High horn in B.

Corno basso (*It.*) Low horn in B.

Corno di bassetto (*It.*) [Basset-horn.]

Corno di caccia (*It.*) [Caccia.]

Corno Inglese (*It.*) [Cor Anglais.]

Cornopean. [Cornet, § 5.]

Coro (*It.*) [Chorus.]

Cor omnitonique (*Fr.*) A horn on which, by the use of valves, a chromatic scale could be played.

Corona (*It.*) A pause. 𝄐

Coronach. [Coranach.]

Corps de voix (*Fr.*) The quality or the fulness of the voice.

Corrente (*It.*) [Coranto.]

Corrépétiteur (*Fr.*) *Correpetitore* (*It.*) The instructor of the chorus; one who teaches the choral body to sing their several parts by ear.

Corti's organ. [Ear.]

Coryphæus (*Lat.*) Κορυφαῖος (*Gk.*) (1) A leader or conductor of the dances or chorus. (2) An officer in the University of Oxford, whose duty it is to give instruction in music.

Coryphée (*Fr.*) (1) A leader of the groups of dancers. (2) A female dancer.

Cotillon (*Fr.*) *lit.* under-petticoat. A lively, spirited dance, originally performed by a male and a female, in which the latter alternately attracted and repulsed her partner. It was first called cotillon in the reign of Louis XIV. was expanded in its design by the French in the last century, and arranged for eight persons. It is now danced with any step by an unlimited number of dancers. When it is possible, chairs are placed round the room for the performers. "Each gentleman places his partner on his right hand. There is no rule that any particular figure shall be danced. The selection is left to the determination of the leading couple who commence the figure, which the other couples repeat in succession. In large parties of twenty-four or thirty couples, it is customary for two or more couples to perform the same figure at the same time. The constant variety of the figures enables each gentleman to dance with almost every lady." The figures from which a selection is made are called the pyramid; the two flowers; the great bound and pass-under; the cushion; the round; the basket, ring, and flower; the two lines of six; the coquette; la gracieuse; the mirror; the handkerchief; the star; the cards; the double moulinet; the deceived lady; the quadrille; the two chairs; the rounds multiplied; the lancers; the three chairs, &c.: the whole being more or less allied to the old-fashioned country dance. It is not at all improbable that the tune "Petticoat loose" given in the article, "Country dance," furnished the title to the Cotillon.

Couac (*It.*) An onomatopœic word for the sound made by bad blowing on the clarinet, oboe, or bassoon. The quacking sound, the goose note.

Couched harp. [Spinet.]

Coulé (*Fr.*) A glide. (1) Slurred notes. (2) A *slide* in dancing. (3) An ornament in harpsichord music; *e.g.*:

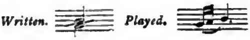

Counterpoint. The term "counterpoint" in its broadest sense may be defined as "the art of adding one or more parts to a given melody;" in its more limited sense as, "the art of harmonising a theme by adding parts which shall be in themselves melodious." The terms subject, melody, canto fermo, and theme, are synonymous. The common definition of counterpoint as the "art of combining melodies" is not strictly logical, unless the word "melody" has a definition not generally accepted; because, distinct melodies are never given to the student to be combined by him unless they have been previously proved capable of combination; and if a composer should attempt to combine two distinct melodies in accordance with the laws of strict counterpoint he will probably find it necessary to eliminate so much of one or both of his subjects that little real musical melody is left. The contrapuntist's notion of a melody is—a succession of sounds which does not infringe certain theoretical laws. No wonder then that authors who have bound themselves by the commands of counterpoint seem to have trodden in one almost identical path and to have added little that is valuable to the literature of counterpoint. Those masters who have exceptionally combined great genius with a deep study of the art of counterpoint, such as Bach, Cherubini, and Mozart, exhibit in their works more than any other authors do, with what beneficial results the laws of counterpoint may be purposely broken, for it cannot be denied that the first fact which startles, and shakes the faith of the student of counterpoint, is that the preaching and practice of contrapuntists are so thoroughly inconsistent. Their books consist of rules, their compositions of exceptions. But it would be dishonest to blink the fact that much good was for a time done by counterpoint, by eliminating crudities in harmony, by introducing an interesting rhythmical correlation of parts, and by opening to ingenious writers a large field for imitative construction of music at a time when the resources of key, modulation, form, and variety of tone in instruments, were greatly limited. Whether a course of study in counterpoint is not more interesting to the lover of musical history than beneficial to the gifted young composer, the reader may perhaps be able to judge for himself after reading the rules of the art and seeing the examples of its scope given below.

Counterpoint is simple or double. There are five species of simple counterpoint.

1, when the added part is note against note of the subject; 2, when the added part is two notes to one of the subject; 3, when the added part is four notes to one of the subject; 4, when the added part is in syncopation to each note of the subject; 5 when the added part is free, or has a florid accompaniment to each note of the subject.

In the first species, note against note, in two parts, the following rules and regulations are to be observed:

1. No discords are allowed.
2. More than three consecutive 3rds or 6ths are forbidden.
3. Consecutive 5ths and 8ths are forbidden.
4. The fourth is to be considered a discord.
5. No augmented or diminished intervals are to be used in the progression of the subject or counterpoint. The major 6th, major 7th, and minor 7th are similarly disallowed.
6. A tritone (or augmented 4th) should be avoided, between the component notes of a chord and that which immediately follows it.
7. False relations are forbidden.
8. Of the three kinds of motion—similar, oblique, and contrary—contrary motion is to be preferred.
9. Hidden fifths and octaves are forbidden.
10. Unison between subject and counterpoint is forbidden.

The first rule requires no explanation. The second is given to insure the independence of the counterpoint, as it is evident that if one part constantly follows another at the interval of the 3rd or 6th, it cannot possibly be said to form a separate melody. The third and fourth rules need no explanation. The fifth rule is saddled with many exceptions, as might be expected; the major 6th and augmented 4th in ascending, and the diminished 7th in descending, are tolerated. The origin of the sixth rule is to be traced in the difficulty of making dominant and subdominant harmony succeed each other with good effect, *e.g.*:

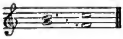

is certainly unpleasant to the ear, whereas

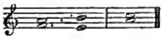

is certainly not so unpleasant, though equally forbidden by strict contrapuntists. The rule against false relations (7) is necessary in two-part writing, as it is impossible to introduce them with good effect. Such progressions as the following are palpably inadmissible:

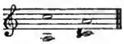

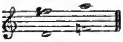

Rule 8 may be proved necessary on the same grounds as given in explanation of rule 2. Two parts cannot be forming separate *melodies*

Digitized by Google

when moving in similar motion, much less can they be doing so in oblique motion, in which, one part stands still: therefore, contrary motion is preferable as probably leading to more variety. The law against the use of hidden fifths and octaves, includes under it the well-known contrapuntal rule: "do not proceed from an imperfect to a perfect interval by similar motion," the fifth and octave being the only perfect intervals admissible in two-part counterpoint of the first species. The presence of hidden fifths or octaves is discovered by filling up the intervals between the consecutive notes of each part with the intermediate degrees of the scale, thus:

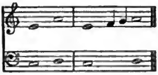
Hidden fifths.

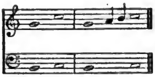
Hidden octaves.

or where both parts move by a skip:

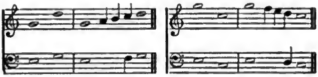

The examples given throughout this article are taken from Fux, "Gradus ad Parnassum," 1725, a work from which all later authors have borrowed largely, Cherubini and Ouseley not excepted: it is fair therefore to suppose that these specimens of counterpoint meet with the approval, if not the admiration, of modern expositors of the art.

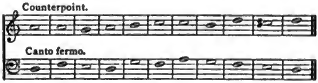

The above is said to be "contrappunto sopra il soggetto." The following example has the same subject in the upper part, "contrappunto sotto il soggetto."

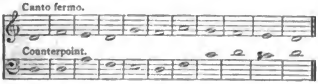

The second species of simple counterpoint is subject to the following rules and regulations:

1. Of the two notes in the counterpoint, the first must be a concord, the second may be a concord or passing discord.
2. Consecutive fifths or octaves on successive down-beats are forbidden. Some authors however admit the latter of these progressions if the skip to the second note is greater than a third.
3. Scale passages are preferable to broken harmony.
4. The counterpoint may commence on the up-beat of the first bar.
5. The cadence of the subject should be harmonised by contrary motion.
6. The interval of a fourth may occasionally be used on the down-beat.
7. A false relation is not avoided by the introduction of a passing note or passing discord.

A passing discord is a discord having a degree of the scale on each side of it, *e.g.*:

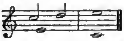

The first rule therefore forbids a discord to be a skip. The second rule is to prevent the use of such progressions as the following:

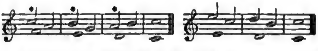

The third rule strives to enforce independence of motion in the counterpoint, *e.g.*:

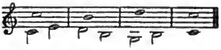

In this example, the upper part is practically harmonised by the lower one, the first bar representing the chord of C, the next two bars a chord of G. Rules 4, 5, and 6, require no explanation. Rule 7 is to prevent the admission of such passages as the following :

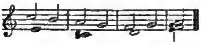

The following are specimens of this species:

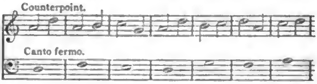

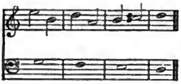

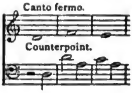

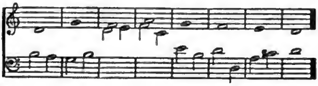

If counterpoint in triple time is used, the first minim must be a concord, the other two may be concords or passing discords as may be found desirable.

The third species of counterpoint in two parts in which there are four notes to each note of the canto fermo, is subject to the following rules and regulations:

1. The first note must be a concord, the second and fourth may be passing discords.
2. The third note may sometimes be a passing discord, but should be generally a concord.
3. The first bar may commence with a crotchet rest, if the note immediately after the rest is a concord.
4. The cadence should be by contrary motion.
5. There may be unison between counterpoint and subject, provided it does not occur on the first note of the bar.
6. The tritone is to be avoided between any four notes of the counterpoint, unless they occur as an integral part of the scale, that is, having the next note of the scale on each side of them.
7. Octaves and fifths between counterpoint and subject should not occur on successive down beats; or between the third crotchet of one bar, and the first crotchet of the following bar.

The cases in which the third note may be a discord (see 2) are of the following kind:

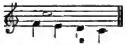

The sixth rule is to prevent the use of such passages as the following:

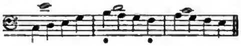

It will, however, be noticed that a passage almost similar to the above, occurs in the second of the two models given below from Fux. Rule 7 is constantly broken. Fux himself gives the following as a specimen of a cadence in this species:

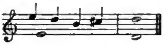

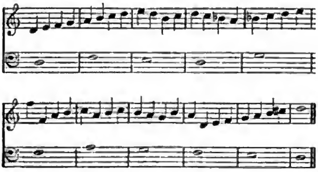

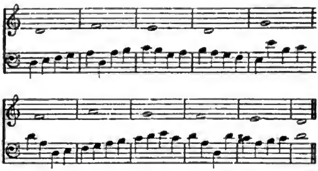

The fourth species of counterpoint in two parts, is that in which the counterpoint though containing practically note against note of the subject, has each note bound into the following bar, or, syncopated.

The following rules and regulations are to be observed:

1. That which is incorrect "sine ligatura" is incorrect when "cum ligatura."
2. It is necessary to begin on the up-beat.
3. Syncopations may be concordant or discordant, a concordant syncopation being one that is heard in both bars (half of each) as a concord: a discordant syncopation one that is a concord on the up-beat, but forms a discord on the down-beat.
4. If necessary, the syncopations may be relinquished for the space of two minims.
5. The best cadence is formed by the suspension 7 6 on the supertonic.

Rule 1 is directed against such progressions as the following:

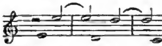

For, tested by the omission of the ligatures, it appears thus:

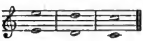

According to this rule the following passage is correct:

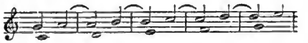

Although it seems to infringe rule 2 of the second species, and rule 7 of the third species, but without ligatures it becomes merely a succession of 6ths:

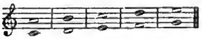

The following are specimens of counterpoint of this species:

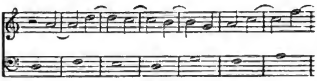

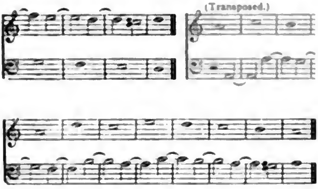

When this species is used in triple time the second note of the bar may be a concord or passing discord, the third must be a concord bound into the next bar and forming a syncopated concord or suspended discord.

The fifth species of counterpoint, florid or figurate counterpoint, consists of a mixture of the various kinds just given, and so far as it proceeds in any one species, is subject to the laws and regulations of that species. Shorter notes may occasionally be used. The following are examples:

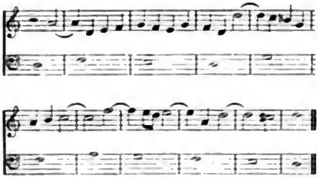

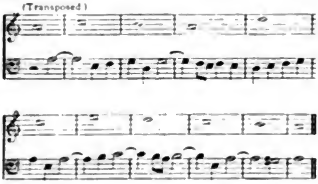

Counterpoint in three parts is, generally speaking, bound by the rules of its corresponding species in two parts. The additional part, however, makes the following rules necessary in the first species, note against note.

1. Every chord should be a common chord, if possible. When not possible the chord $^{6}_{3}$ may be used.
2. The third of the common chord should not be doubled.
3. The term "chord $^{6}_{3}$" includes under it the chord $^{6}_{3}$ on the supertonic, that is, the second inversion of the chord of the minor seventh, the root being omitted. As the old masters did not consider this chord a discord, the seventh of the root (third of the chord) is frequently resolved upwards by them. It will not be necessary to give specimens of counterpoint in every form of which it is capable, one example of each species will suffice, if the reader will remember that the canto fermo may appear in upper, lower, or middle parts. The student of counterpoint should refer to Fux, "Gradus ad Parnassum" Vienna, 1725, or to an English translation called "Practical rules for learning composition," printed by Welcker, Gerrard Street, Soho (at the end of the last century), or to Cherubini's work (Novello, London).

The second species of counterpoint in three parts contains one part having two notes to each note of the subject.

1. A syncopation is allowed to take place immediately before the cadence.
2. The third of the common chord should not be doubled.

The following is a specimen of counterpoint of this species:

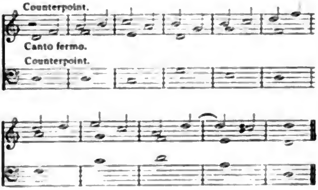

The third species of counterpoint in three parts contains one part having four notes to each note of the subject. A syncopation before the cadence is not permitted in this species:

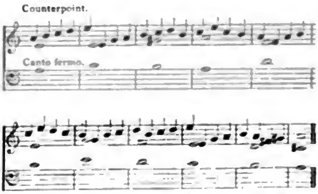

A mixed kind of counterpoint, containing one part having two notes to each note of the subject, and another having four notes to each note of the subject, may be classified under this species:

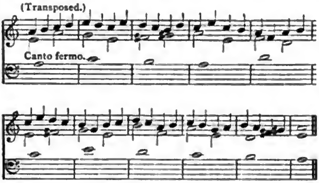

The fourth species of counterpoint in three parts contains one syncopated part:

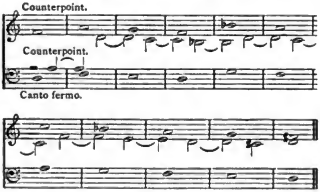

The fifth species of counterpoint in three parts contains one figurate or florid part. It is unnecessary to give an example of this species.

In counterpoint of four parts, the rules of two-part and three-part counterpoint are necessarily relaxed to some extent in the case of the inner parts, unless the inner parts consist of the canto fermo and the counterpoint specially characteristic of the particular species to which the example belongs.

The following rules and regulations apply to four-part counterpoint generally:

1. In the first species, only common chords should be used, but the chord $\frac{6}{3}$ may occasionally be used.
2. In every species, the different parts should be as much as possible equidistant.
3. Two parts may occasionally cross each other.
4. The laws against hidden fifths and octaves do not bind inner parts, and consecutive fifths by contrary motion are sometimes permitted.

It will be sufficient if two examples of four-part counterpoint are given:

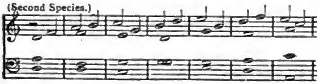

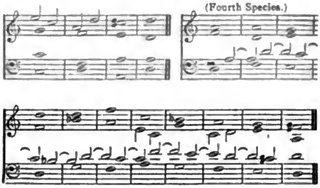

Counterpoint may be in 5, 6, 7, 8 or even 16 parts, but enough has been said to give the reader an insight into its principles.

Counterpoint (Double) has been well described as a "kind of artificial composition where the parts are inverted in such a manner that the uppermost becomes the lowermost, and *vice versa*." Or, in other words "the art of making melodies grammatically convertible at certain intervals."

If the melodies are interchanged at the interval of an octave, the double counterpoint is said to be "at the octave," but if the inverted melody is transposed one note, the other melody remaining untransposed, the double counterpoint is said to be at the 9th. Similarly, the double counterpoint may be at the 10th, 11th, 12th, 13th or 14th. But double counterpoint at some of these intervals imposes such difficulties in the construction of the component melodies, that it is rarely met with. Double counterpoint at the 8th, 10th and 12th, are the kinds most commonly used, and shall be explained in order. It will be seen if the following passage:

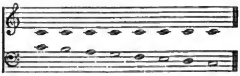

be inverted by playing the lowest line an octave higher, and the highest an octave lower, thus:

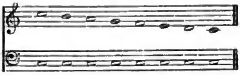

that the intervals between the two parts have undergone an entire change, with the exception of the octave which has become a unison.

Thus the 2nd has become a 7th.
,, 3rd ,, 6th.
,, 4th ,, 5th.
,, 5th ,, 4th.
,, 6th ,, 3rd.
,, 7th ,, 2nd
,, 8th ,, 1st

We have above, then, a complete scheme of

Digitized by Google

the changes intervals undergo by inversion at the octave. It is evident that the following rules must be observed:

1. As the 5th becomes a 4th, and the 4th is a discord, the 5th must be treated as a discord, if used at all. As a discord, it may be prepared, or treated as a passing discord.
2. The octave must be approached by a single degree, not a skip, in strict style.
3. As it is usual to place the upper melody an octave lower, leaving the lower unchanged, the interval of an octave between the two parts must not be exceeded, otherwise the object of inversion will be lost, *e.g.*:

Ex. 1.

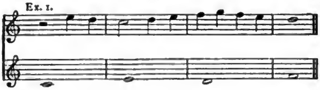

will become:

Ex. 2.

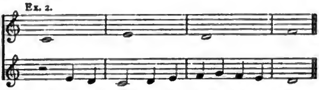

In the first and third bars of Ex. 2, no inversion has taken place. If melodies are framed with the intention of altering the pitch of both when inverting them, this rule does not of course hold good, *e.g.*:

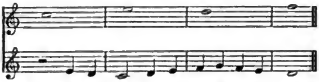

4. The melodies should be different in style, and one should commence on the up-beat. The following is an example of double counterpoint at the octave, from Fux:

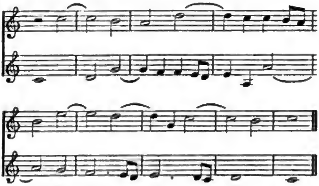

Inversion (transposed).

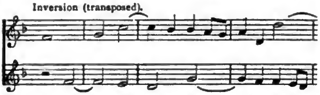

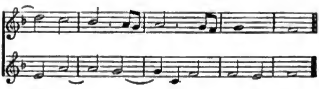

If we wish to discover the changes intervals undergo by double counterpoint at the 10th, we can, as before, write them out thus:

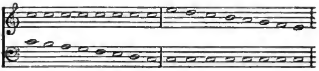

1. Consecutive 10ths become consecutive unions, *e.g.*:

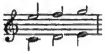

Will become when inverted:

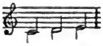

and consecutive 3rds become consecutive octaves; *e.g.*:

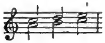

Will become when inverted:

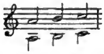

Both consecutive 10ths and 3rds must therefore be avoided.

2. Consecutive sixths become consecutive fifths; they therefore must be avoided, *e.g.*:

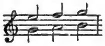

Will become when inverted:

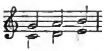

3. The suspension 4 3 becomes a 7th resolved wrongly; *e.g.*:

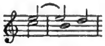

Will become when inverted:

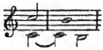

4. The interval of a tenth between the upper and lower melodies should not be exceeded, for the reason given in rule 3 of double counterpoint at the octave.

The following example is from Cherubini:

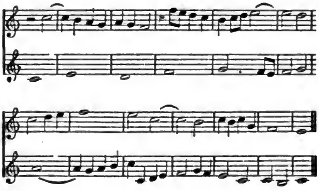

Inversion.

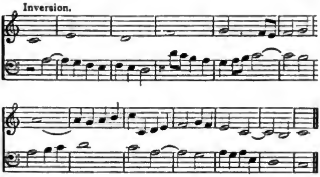

The subject of the above may also be written in the third above and the counterpoint in the octave below, throughout; or again, the counterpoint may be written in the third below, and the subject in the octave below. The following example, from Fux, shows how the same counterpoint may be used at the same time, at the octave and the tenth, each counterpoint being correct when taken separately:

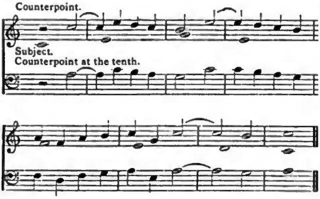

Double counterpoint at the 12th is much less hampered by the change of intervals than many other species.

1. The 6th becomes the 7th, if therefore introduced, it must be as a discord in the lowest part, *e.g.*:

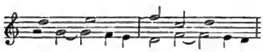

will become when inverted:

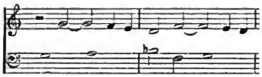

or, a sequence of prepared sevenths.

2. The final cadence will require special care in its treatment.

The following is an example of this species:

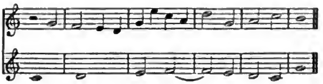

Inversion.

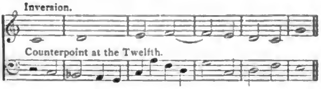

Interesting examples of combinations of counterpoints at the 10th and 12th are to be found in Fux.

Counterpoints, Triple and Quadruple, as their names show, are the due construction of three or four melodies respectively, in such a manner that they can be interchangeable without involving the infringement of the laws of musical grammar. It will be evident, on consideration, that the octave is the only feasible interval at which counterpoints of this class can be made, unless indeed one or more *free* parts, that is, parts not forming interchangeable melodies, are added. The following is an example of triple counterpoint, written out in full. Of course three sentences at least will be required for the exposition of *triple* counterpoint; four sentences for that of quadruple:

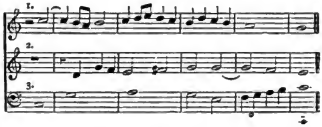

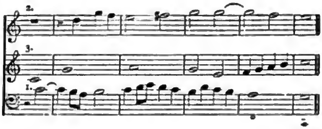

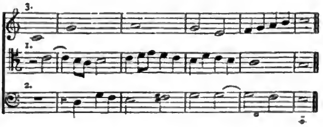

The following example of quadruple counterpoint is from Zimmerman:

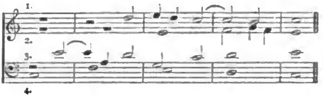

Digitized by Google

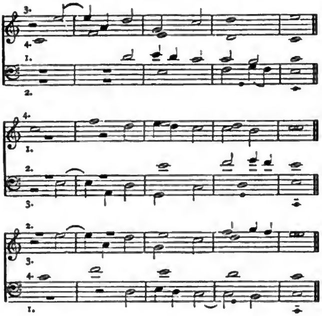

It is perhaps necessary to warn the lay reader against the confusion likely to arise between the terms two-part, three-part, four-part, counterpoint; and double, triple, and quadruple counterpoint. The former refers only to the number of parts added to a given subject, and such parts need not necessarily be interchangeable; whereas, the essence of the latter is that in each case all the parts must be capable of substitution one for the other.

Occasionally, specimens of quintuple counterpoint are to be met with, but they may be looked upon more as curiosities than as substantial additions to the musical art.

Counter subject. [Fugue.]

Counter tenor clef. The C clef placed upon the third line of the stave for the use of counter tenor or alto voices, the viola, &c.:

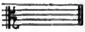

Counter tenor voice. The old name for the alto voice. [Alto voice.]

Country Dance. *Contre-danse* (*Fr.*) *Contraddanza* (*It.*) A rustic dance, of English origin, in which performers were arranged face to face, "one set against another," and performed certain prescribed figures. The old method of dancing the "country dance" was to place the ladies and the gentlemen in two parallel lines, the former on the left, the latter on the right, facing their partners. All advance, then retreat, during the first four bars of the music, then cross to opposite places, then advance and retreat, and then re-cross to original places. Each of these movements should occupy the time of four bars of music. The lady who stands at the top, and the gentleman whose place is at the bottom, advance towards each other, *courtesy* and bow, and return to their places. The gentleman at the top and the lady at the bottom do the same. Then the first named couple advance once more, give right hands and swing quickly round each other back to places. This figure is repeated by the other couple. The lady at the top then advances, gives her *right* hand to her opposite partner, and passes behind the two gentlemen standing in the places next to him: then, through the line and across it, giving her *left* hand this time to her partner, who meets her half way between the two lines, having passed behind the two ladies next to his partner's place. The lady then passes behind the two ladies next in the line, the gentleman moving in the like figure behind the two gentlemen next lowest, and so on, all down the line. At the bottom the lady gives her left hand to her partner, and they *promenade* back to their former places. Then the top couple come forward, *courtesy* and bow, the lady turns to the right, the gentleman to the left, each followed by the rest of her or his line. Top couple meeting at the bottom join hands and raise their arms to form an arch for the other couples to pass under, until all have reached their places except the top couple; these having become the bottom couple, repeat the figure from the beginning until they have worked back to their original places at the top of the lines, and then the dance is ended. Such is a general description of a dance which under various titles has been popular in England for centuries, has been adopted by other nations, and revived from time to time with a few modifications under the several titles applied to it by the people from whom it was last taken. Thus it has been called "contre-danse" and is erroneously said to be French; and when it has been named "coranto" it has been supposed to be Italian.

But few books had such a successful career as the "Dancing Master," first issued in 1651, by John Playford. Seventeen editions are known, extending to 1721-8. The following are taken from the first edition, probably the earliest collection of English country dances.

"STINGO; OR, THE OYLE OF BARLY."

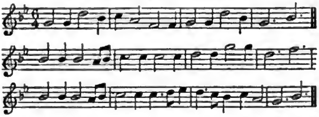

"MILL-FIELD."

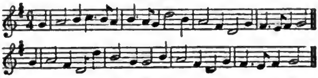

"JACK PUDDING."

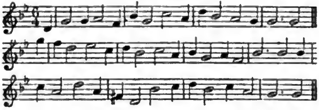

"THE MILKE-MAYDS BOBB."

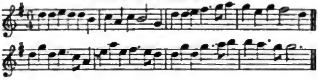

"PAUL'S WHARFE."

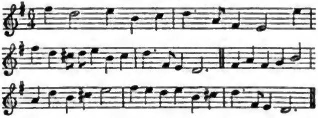

It is unnecessary here to enlarge upon the popularity of dancing, throughout all ages; a reference to Strutt's "Sports and Pastimes," and to Chappell's "Popular Music in the Olden Time," will satisfy those who are curious as to details. It may be here stated that the old poets and dramatists, from the time of Chaucer and later, have frequent allusions to the custom, and make mention of many dances by name without giving descriptions, so that it may be inferred that their allusions point to practices in their time so popular that particular description was not deemed needful.

One of the old English names for rustic dances was *hey digyes* or *rounds.*

> "While some the rings of bells and some the bagpipes play,
> Dance many a merry round, and many a *hy degy.*"
> DRAYTON'S "Polyolbion," Song xxv.

The "Hay," or "Raye" as it is also called, is probably the same as the "hey digyes." It was danced by many, forming a line or a circle, and the direction was to "wind round *handing* in passing until you come to your places."

"THE HAYE," a Countrie Dance, 1678.

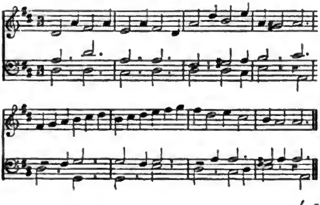

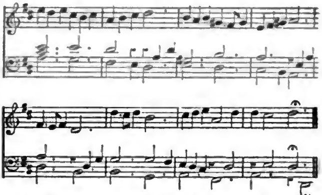

An ancient country dance called Dargason has survived to the present day, although it has gone through considerable modification. The origin of the name is unknown; if it is true that Dargason was the name of a dwarf in some ancient story, it is quite possible that it may have been "The Dwarf's dance." The following is the version which appears in the Dancing Master, 1651:—

"SEDANNY OR DARGASON."

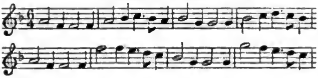

This version of the last line shows a curious attempt to make the tune "circular":—

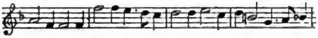

The same character of tune which suited the country dance was also used for the reel, the round, the morris-dance, the jig, and hornpipe, all of which are offshoots from the one original stem. Those among these dances now performed by one or at most two dancers, were not always so done, the reel was often "four or eight handed or even general," the jig and hornpipe were also dances for many. The two last named probably derived their title from the instruments employed as accompaniments, the usual accompaniment to most country dances was anciently the fiddle, in German *Geige;* or, pipe and tabour. There were many other names given to the country dance in successive ages, and the variety of the titles has led many writers into the belief that there were as many dances as names. The allusions found in the writings of the poets and dramatists have, to a certain extent, increased the confusion in the minds of readers, and commentators not deeming the subject worthy of the consideration it deserves, have often by wrongly directed notes and glosses, made matters in a worse condition than that arising from original error. Sir John Davies (1570-1626) in his poem "Orchestra" is clearer than other authors on the subject. He identifies rounds, corantos, measures, &c., with country dances. His description of a

Digitized by Google

country dance, to be found under "Brawl," is almost the same as that given above, which is the process of performing the country dance to this day. He calls a "measure" "a round dance for ever wheeling," and implies that "as men more civill grew, they did more grave and solemn measures frame" out of the primitive country dance. The "galliard" "a swift and wandering dance with passages uncertain to and fro, yet with a certain answer and consent." The coranto or "current traverses" in which he says of the dancer:

> "Everywhere he wantonly must range
> And turn with unexpected change"

All these forms are but slight variations of the simple original, and as it is admitted that "no rules have ever been laid down for the composition of a country dance, nor is it indeed confined to any particular measure; so that any common song, or tune, if sufficiently rhythmical may by adoption be made a country dance," the diversities of *tempo* in the several melodies, of the coranto, rondo, galliard, and measure ought not to be taken as a proof of a distinct character of dance.

In the rustic dances the motion was rapid, but when people of less humble condition deigned to adopt them, they varied the figures, made the motion more dignified, and giving a new title to the old diversion, created a certain amount of confusion in the minds of interested posterity. The "stately measure, the graceful minuet, and the courtly quadrille" are each and all country dances, and people of all conditions have indulged in the pastime they offer. Mr. William Chappell ("Popular Music in the Olden Time," p. 626) shows that country dances were popular at court in the reign of Queen Elizabeth, and succeeding reigns. The custom of dancing the ancient English country dance was kept up at court during the reign of George III, as many newspapers and other records show. Thus, in the "Universal Magazine," for June, 1784, we read: "June 4th, the anniversary of the King's birth-day, the drawing-room broke up about half-past five, when their majesties returned to the Queen's palace to dinner; and at about nine in the evening there was a grand ball, which was opened by the Prince of Wales, who walked the first minuets with the Princess Augusta . . . . The country dances began a little before twelve, and continued till past one."

It has been mentioned above that the rhythm of country dance tunes is various, some are in triple and some in duple measure. Among the most popular airs employed for the dance, those called "Sir Roger de Coverley," the "Tank," the "Triumph," "Gee ho, Dobbin," "Merrily danced the Quaker's Wife," "Petticoat Loose," "Gossip Joan," "The Devil among the Tailors," "Moll in the Wad," and the "Wind that shakes the Barley," are still popular: these are all different in accent and measure, yet all serve the purpose of the dance. It matters not whether the time be $\frac{9}{8}$, $\frac{12}{8}$, $\frac{6}{8}$, or $\frac{4}{4}$; all that is necessary is that the strains should be in four or eight bar phrases to accompany the several movements, and every need is satisfied.

"ROGER OF COVERLY."
(As published 1725.)

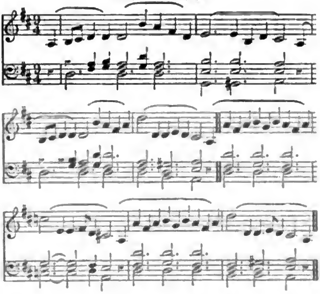

"THE TANK."

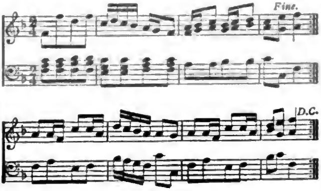

"THE TRIUMPH."

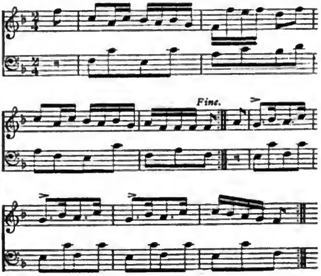

Digitized by Google

"Gee ho, Dobbin."

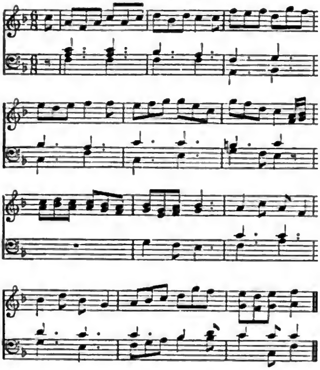

"Merrily danced the Quaker's Wife."

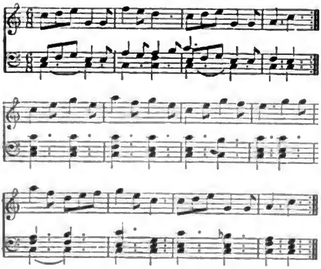

"Petticoat Loose." (Macfarren's Harmony.)

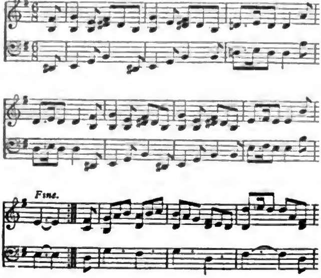

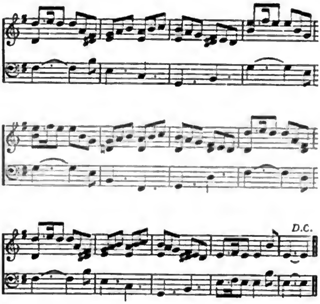

"Gossip Joan."

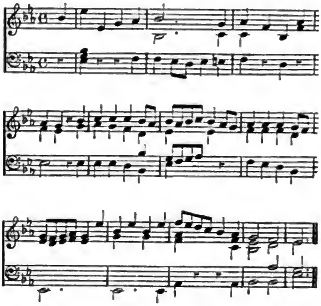

"The Devil among the Tailors."

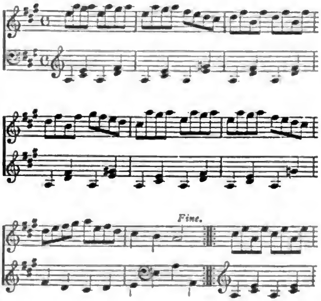

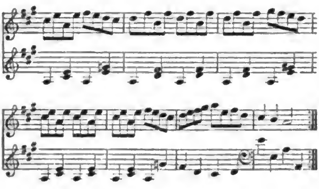

"Moll in the Wad."

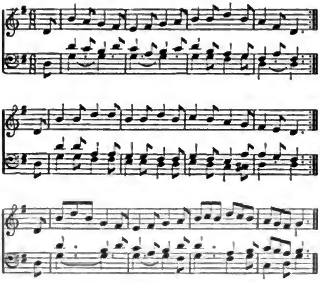

"Rolling in the Dew."

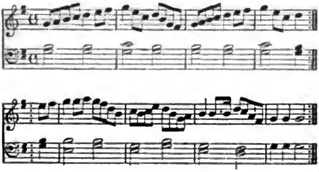

Country dances when imported into other nations have become as popular as at home. The Italians, in 1740, were said to be "fond to a degree" of them, and about the same period in Paris, "no kind of dance was received with so much favour as they." Dancing masters vied with each other in devising new combinations of figures, and musicians of the common order provided original or borrowed tunes for the dance, many of which were published in single sheets with such titles as "La N$^{\text{elle.}}$ Chartres, countre danse par Mr.——, M$^{\text{tre.}}$ de Danse, prix 4 s. la feuille; à Paris, ches M. de la Chevardiere, M$^{\text{d.}}$ de Musique, rue du Roule a la Croix d'or; M$^{\text{lle.}}$ Castagniere, rue des Prouvaires, avec privilege du Roy."

These publications consisted of four pages, the first occupied with the title, as above; the second containing a description "des figures de la contre danse;" the third diagram-plans of the said figures, and the fourth the music, which in the instance quoted above, is as follows:

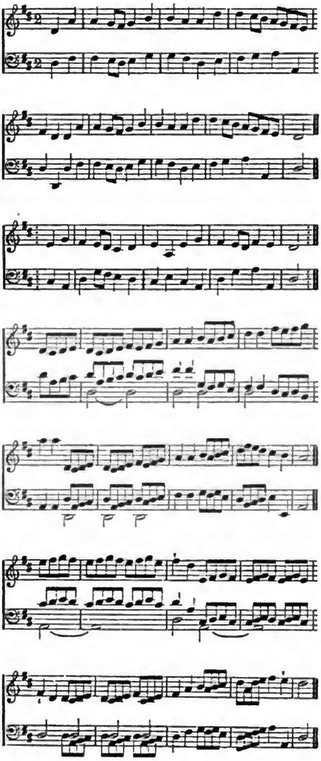

The following diagram from the above work, shows the various figures of another dance, the black marks representing the position of the men; the white marks, that of the women; the arrow-heads, the direction in which they move.

Digitized by Google

Description des Figures de la Contredanse.

1. *La grande chaine, un demi tour.*

2, 3. *La poussette dessus, et dessous, et la piroüette à chaque bout.*

4. *Autre ½ tour de chaine, la poussette, et la piroüette au bout.*

5, 6, 7, 8. *Gager d'une place sur le côté; deux se tenant les mains passent en dedans et les deux autres en dehors, continuant à passer de cette maniere 4 fois, jusqu'à ses places, faisant deux balancé à chaque place.*

9. *La chaine des dames sur les côtés.*

10. *Un tour allemande en piroüettant.*

11. *Refaire la chaine des dames.*

12. *Un tour d'allemande.* La Main.

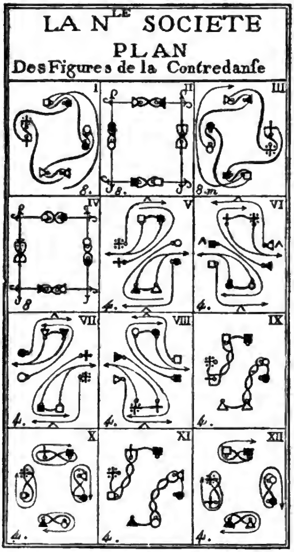

The subjoined is the melody to which it was danced:

"La Nouvelle Societe," Contre-danse Allemande.

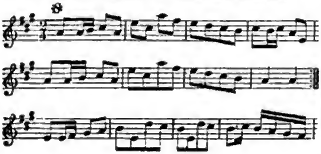

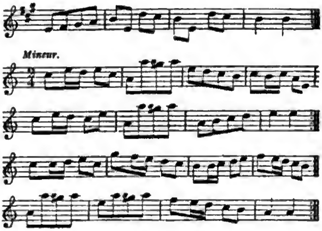

2 fois la rep. du majeur en rondeau, et deux fois chaque rep. du mineur.

Coup d'archet (*Fr.*) Stroke of a bow.

Couper le sujet (*Fr.*) To abbreviate or curtail a musical subject or theme.

Coupler. The mechanism which connects pedals with the manuals; or, different manuals together. [Organ.]

Couplet. (1) Two lines in rhyme, which contain a complete sentence. (2) A verse of a song. (3) Two notes occupying the time of three, *e.g.*:

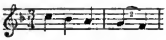

Courante (*Fr.*) [Coranto.]

Couronne (*Fr.*) The name for the sign of a pause 𝄐.

Courtaut, Cortaud, Corthal. An ancient instrument of the bassoon kind. [Bassoon.]

Covered consecutives. Hidden consecutives. [Consecutives.]

Covered strings. Strings of silk, wire, or gut, covered with a fine wire by means of a machine, by a process technically termed string-spinning. Covered strings are used for pianofortes, violins, violoncellos, guitars, &c., the wire covering, by adding weight and strength to the string, makes it slower of vibration, while, on the other hand, it is more elastic than an uncovered string of the same diameter.

Cownterynge yn songe (*Old Eng.*) In Lat. *concentus* or *accentus*, singing an accompaniment to a tune. [Descant.]

Cracovienne. [Polacca.]

Crackle. A direction in lute playing, thus explained by "Maister" Thomas Mace, 1676: "To crackle such three part stops is only to divide each stop, with your thumb and two fingers, so as not to loose time, but give each crochet its due quantity." [Arpeggio.]

Credo (*Lat.*) One of the movements in a mass. [Mass.]

Crembalum. [Jew's Harp.]

Cremona. (1) A violin made in the town of Cremona. (2) A reed stop in the organ. A corruption of the word Krummhorn.

Cremorne (*Fr.*) [Krummhorn.]

Crepitaculum or **Crepundia** (*Lat.*) An ancient instrument of a character like the castanets, but with sound produced more by friction than striking. [Castanets, Bones, Knicky-knackers.]

Crescendo (*It.*) Increasing, a gradual increase in the force of sound expressed by the sign <, or the abbreviation *cres.* The sign was first employed in England by Matthew Locke, in 1676.

Crescendo Zug (*Ger.*) The swell box in the organ. [Organ.]

Creticus (*Lat.*) A metrical foot consisting of one short syllable between two long – ⏑ –

Croche (*Fr.*) A quaver ♪, the *hooked* note. [Nomenclature.]

Croma (*It.*) A quaver ♪

Cromatico (*It.*) Chromatic; as, *fuga cromatica*, a chromatic fugue; *fantasia cromatica*, a chromatic fantasia, &c.

Crom horn. [Krummhorn.]

Crooks. Short tubes, either straight or curved, adapted for insertion between the mouthpiece and the body of the horn, trumpet, or cornet-à-pistons, for the purpose of altering the key. [Metal wind instruments.]

Crotalum (*Lat.*) κρόταλον (*Gk.*) A rattle, or clapper, used sometimes to mark the rhythm of dancing, in the worship of Cybele. They were generally made of wood, having a loose piece hinged about midway, so that, when shaken in the hand a clattering noise was produced, called by the Greeks πλαταγή.

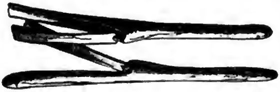

Instruments of this kind were in use among the ancient Egyptians, as the following illustration shows:

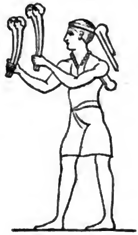

Crotchet. A note ♩ one-fourth of the value of a semibreve. [Nomenclature.]

Croupeza, κρούπεζα, κρούπαζα, κρούπτενα κρούπετα (*Gk.* from κρούω, to knock, strike) High wooden shoes worn by flute-players or others, with which the time was marked by striking with the foot; *c.f. Lat.* scrupedæ, women who wore high-heeled boots.

Crowd, Crwth. An ancient instrument, like a violin, with six strings, four of which were played upon by a bow, and the other two played, or plucked with the thumb, as an accompaniment. The neck had a hole, through which the player thrust his hand, so that he could only command the notes lying under his fingers. [Violin.]

Crowle. An early form of the word *corolla* (*Lat.*) a crowd, *q.v.*

C-Schlüssel (*Ger.*) The C clef.

Cum sancto (*Lat.*) A portion of the Gloria in the mass. [Mass.]

Cue. A catch word or phrase. The last notes or words of other parts inserted as a guide to singers or players who have to make an entry after rests:

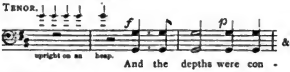

Cupo (*It.*) Darkly, mysteriously.

Currende (*Ger.*) Children carol-singers.

Cushion-dance. An old English round dance, in which each woman selected her partner by placing a cushion before him. Taylor, the Water Poet, calls it a "pretty little provocatory dance," for that reason. There was a dialogue carried on, according to the description given in the "Dancing Master" of 1686; and the note appended to the same description points—perhaps unwittingly—to the probable origin of the dance: "*Note.*—That the woman is kiss'd by all the men in the ring at her coming in and going out, and the like of the man by the women." This general direction for kissing makes it more than probable that the "Kissing Dance" was the "Cushion Dance." The following is taken from the Dancing Master, 1686. In the 1725 and 1735 editions the first note of the second bar is made natural; and in the 1725 edition the notes F in the last strain are made sharp. [Country Dance.]

"Joan Sanderson; or, The Cushion Dance" (1686).
A Round Dance.

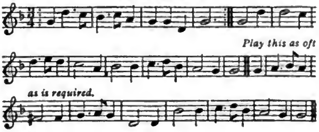

Custos (*Lat.*) A *direct*, the sign w or √ at end of a line showing first note of next line.

Cyclische Formen. Rondo forms. [Form.]

Cymbalista. A cymbal player.

Cymbals. *Cymbalum (Lat.) κύμβαλον (Gk).* Musical instruments of percussion, consisting of two metallic basins, which are set in vibration by being clashed together. The shape of cymbals varies, from that of the actual form of a cup or basin to an almost flat plate. The following illustration shows those used by the Assyrians. It will be remarked that the lower basin is held in a stationary position, while the upper one is dashed on it.

Fig. 1.

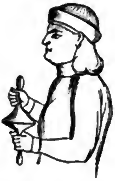

The Hebrews had two kinds of cymbals, mentioned by name in Psalm cl. 5, "Praise Him upon the loud cymbals; praise Him upon the high-sounding cymbals." The Arabians have two sorts at the present time, the larger they use in their religious ceremonies, but the smaller are rarely used but for the purpose of accompanying the dance. In India cymbals are used called *talan*, and a smaller sort called *kintal*. An illustration of Indian cymbals is given:

Fig. 2.

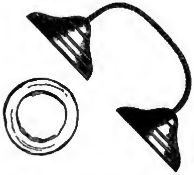

The Burmese instruments of this class are of the true basin shape, as shewn in the following:

Fig. 3.

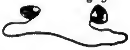

A pair of ancient Egyptian cymbals are in existence; they are about five inches in diameter, and are made of a mixture of copper and silver, and in outline are identical with those now used by modern Egyptians.

As has been the case with other musical instruments, the name cymbal has been applied in various ways. At one period the Italians called a *tambourine* by this name, and at another a *dulcimer*. As the harpsichord was the actual outgrowth of the dulcimer, the harpsichord came to be called *cembalo*, a word still to be found occasionally affixed to the *pianoforte* part of full scores. It is probable that the peculiar clang produced by striking the wire strings of a dulcimer with a wooden hammer gave rise to the association of the name cymbal with dulcimer.

In modern military bands cymbals are used in the ancient manner. One plate is held in each hand of the performer, and the sound is produced by clashing the plates together. In the orchestra of the concert-room, one plate of a cymbal is attached to the upper side of the rim of the big (upright) drum, and the other held in the left hand of the drummer. The tone produced by the beating of these is largely increased in power and depth by the connection with the drum. Very small cymbals were introduced by Berlioz, tuned a fifth apart, as an orchestral instrument, but have not come into common use. Small cymbals are sometimes attached to the fingers and are hence called finger-cymbals:

Fig. 4.

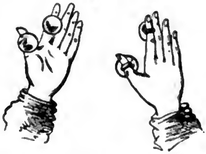

These naturally became associated with castanets; and they have also found their way into the rim of the tambourine, of which instrument they form an important element.

It should be stated that cymbals are not struck together, actually face to face, for by so doing not only would the free vibration of the plates be very much arrested, but they would in all probability be split by the blow. Turkey is still celebrated for its manufacture of cymbals and other instruments of percussion, and exports them in large quantities to all parts of the world. The exact composition of the metal used in Turkey is not known to the manufacturers in other countries.

Cypher-system. [Notation.]

Czakan. A flute made of cane or bamboo.

Czardasch (*Hung.*) [Chica.]

Czimken (*Polish*). A dance similar to the country dance.

Digitized by Google

D.

D. (1) The first note of the Phrygian, afterwards called Dorian, mode.

(2) The second note of the normal scale C.

(3) The scale having two sharps in its signature.

(4) The name given to a string tuned to D, *e.g.*, the third string of the violin, the second of the viola and of the violoncello.

(5) The name of a clef in old mensurable music, *D excellens*. [Clef.]

(6) *Abb.* for *Discantus*, *Dessus*, *Destra*, &c.

Da ballo (*It.*) In dance style.

Da camera (*It.*) For chamber use. In the style of chamber music.

Da cappella (*It.*) In the church style.

Da capo (*It.*) From the beginning. An expression first used by Scarlatti in his "Theodora," signifying that the performer must recommence the piece, and conclude at the double bar marked "Fine."

Da capo al fine (*It.*) From the beginning to the mark *Fine.*

Da capo al segno (*It.*) From the beginning to the sign (𝄋).

D'accord (*Fr.*) In tune.

Dach (*Ger.*) Sounding-board. Resonance-body of an instrument.

Da chiesa (*It.*) For the church. In the church style.

Dachschweller (*Ger.*) Swell-box.

Dactyl. A metrical foot, consisting of a long syllable followed by two short syllables.

Dactylion (*Gk.*) An instrument invented by Henri Herz, for strengthening the fingers for pianoforte playing. [*c.f.* Chiroplast.]

Daina or **Dainos.** A term given to some little Lithuanian love-songs.

Daire (*Turkish*). A tambourine.

Da lontano (*It.*) In the distance, *e. g.*, *corni da lontano*, horns heard in the distance.

Dal segno (*It.*) From the sign (𝄋).

Damenisation. The syllables da, me, ni, po, tu, la, be, which Graun employed for the notes of the scale in his vocal exercises. [Solmization.]

Damp, to. (1) On instruments played by plucking the strings, as the harp, guitar, &c., to check the vibrations by placing the hand lightly on the strings. (2) To apply mechanical dampers.

Damper. (1) Certain moveable pieces of mechanism in a pianoforte, made of wood covered with cloth, which, after the finger has struck the key and left it, immediately check the vibrations of the strings, and prevent that confusion of sound which would result if they were allowed to continue in vibration. (2) The mute of a horn, and other brass wind-instruments.

Dämpfer (*Ger.*)(1)Damper.(2)Violin-mute.

Dancing. A graceful movement of the feet or body, intended as an expression of various emotions; with or without the accompaniment of music to regulate its rhythm.

Dancing is mentioned by the earliest writers, both sacred and profane, as a constituent part of religious ceremonies. There are many instances named in the Bible, needless here to particularize, and the ancient Greek poets have abundant allusions to the practice in their writings. Homer mentions dancing and music at social entertainments; Aristotle tells of dancers who were able to express manners, passions, and deeds in rhythmical gestures; Herodotus, Pindar, Athenæus, and others of later date refer to the practice. Donaldson says that all ancient dancing was "either gymnastic or mimetic; it was gymnastic when intended merely as an exercise, it was mimetic when it was designed to express some mental feeling, or to represent by corresponding gestures the words of the accompanying chorus sung."

Athenæus speaks of three divisions of the Greek dance: the Emmeleia (ἐμμέλεια), the Sicinnis (σίκιννις), and the Cordax (κόρδαξ); the first named from the melody played to it, the second from its inventor Sicinnos, and the third probably for the reason hereinafter explained. The Emmeleia, the tragic dance, was a kind of slow dignified movement or ballet. The Sicinnis was of a grotesque character, and was performed with a peculiar shaking of the body and violent motion of the limbs. The Cordax was less decent in style than the last named. It was introduced into comedies, and was performed by actors assuming to be under the influence of wine. In addition to these there were the Pyrrhic or war dances, expressive of the pursuit and encounter of an enemy.

Digitized by Google

The Roman dances, at first connected with religious observances, became by degrees separated from them, and perhaps degenerated, as it was considered disgraceful for a free citizen to dance, excepting during devotional exercises.

The Almée, or dancing and singing girls of Egypt, the Nautch girls of India, perform, at feasts and solemn occasions, certain dances akin to those which formed part of the ancient observances.

There is ground for the belief that dancing was not discouraged among the early Christians, and there are records showing continuance of the custom among the less orthodox sects at different periods of the history of the Church.

Dancing and pantomimic actions formed part of the amusements sometimes offered by the jongleurs, a body of the minstrel class; their dancing often included acrobatic performances. The common dances, popular among the people in various European countries, vary more in name than in character; and as they are unquestionable legacies of heathen days, have been condemned from time to time by the more serious-minded.

The force and original meaning of dancing is now lost sight of. It is not now regarded in the light of an act of worship, but is encouraged only as a means of social enjoyment. The rude forms of dancing have been softened and polished during successive generations, their character changed, and their identity or connection with their origin disguised under modified motions. Each country in which dancing is practised has considered itself free to change the steps, arrangement, and significance of the dance, or to give preference to one portion of a complicated whole; and such alterations have been accepted as new dances, when they are not really so. The German waltz, the French cancan, the English country-dance, the Spanish bolero, the South American chica, the Italian saltarello, the Hungarian czardasch, are all forms traceable to one source. The allemande, the brawl, the coranto, the fandango, the forlana, the gavotte, the hornpipe, the jota, the kalamaika, the loure, the measure, the minuet, the passecaille, the quadrille, the ringeltanz, the saraband, the tarantella, trenchmore, zapateado, &c., are only different names of the several motions of that called in England the country-dance, with such variations in melody and rhythm as would arise from the use of accompanying musical instruments more or less perfect in their construction, or on account of the speed at which they were danced, by which means a rapid triple measure may be made to seem duple measure. The advancing and retreating in the various figures; the embracing and unloosing, the stamping, shrieking, and singing in some dances; the "grand chain," or the galopade which generally marks the concluding figure of a quadrille, are merely mild versions of some of the several peculiarities of the ancient prototype.

The Italians of the 16th century are credited with the distinction of having invented that form of dancing known by the general term of ballet: they arranged the motions and gestures of the body in an expressive pantomime, and reduced the various actions to a series of well-defined and understood rules, so that the performers were able to impart to the spectators a perfect story without the aid of words; but their claim cannot be upheld, as the like thing had been done by the Greeks ages before.

The rhythm of the more important dances will be found described under their respective titles.

Darabooka or **Darabukkeh.** An Arabian drum; the body, to which is attached a handle, is of hollowed wood. There are various kinds of this instrument.

Darmsaiten (*Ger.*) Strings of catgut. [String.]

Dash. (1) A line drawn through a figure in thorough-bass, showing that the interval must be raised one semitone, *e.g.*:

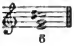

(2) A line drawn through the duple time-sign, *e.g.*, ₵, implying a division either of measurement or of pace.

(3) A short stroke placed above notes or chords, directing that they are to be played *staccato*.

(4) In harpsichord music, a dash passing between two notes, called a slur, or coulé:

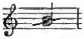

was thus played;

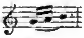

Da teatro (*It.*) In the theatrical style.

Dauer (*Ger.*) Duration or continuance of notes or sound.

Daumen (*Ger.*) The thumb.

D dur (*Ger.*) D major.

Début (*Fr.*) A first appearance.

Débutant, e (*Fr.*) A performer who appears for the first time.

Dec., *abb. of* Decani.

Dec., *abb. of* Decrescendo.

Decachordon (*Gk.*) An instrument with ten strings.

Decani (*Lat.*) A term used in cathedral music, to signify that the part so distinguished is to be sung by the singers on the dean's, or south side of the choir, in contradistinction

Digitized by Google

to "cantoris" the cantor's or præcentor's side. [Cathedral Music.]

Deceptive cadence. [Cadence.]

Décidé (*Fr.*) Firmly, with decision.

Decima (*Lat.*) A 10th, an interval of a 10th; *decima plena de tonis*, a major 10th; *decima non plena de tonis*, a minor 10th; *decima quarta*, a 14th or octave of the 7th; *decima quinta*, a 15th or double octave; *decima tertia*, a 13th or octave of the 6th.

Decimole. [Decuplet.]

Decisio. [Apotome.]

Deciso (*It.*) Determined, decided, with firmness.

Decke (*Ger.*) (1) Cover, an upper or lower plate of a resonance box. (2) The cover of stopped metal organ pipes; *e.g.*, *leiblich Gedackt*, the sweet toned stopped-diapason.

Declamando (*It.*) In a declamatory style.

Declamation. The proper rhetorical rendering of words set to music. [Recitative.]

Décomposé (*Fr.*) Unconnected, incoherent.

Décoration (*Fr.*) Signature of a piece of music.

Decres., *abb.* of Decrescendo.

Decrescendo (*It.*) Decreasing gradually the volume of tone. Indicated in music by the abbreviations Dec., Decres., or the sign >

Decuplet. A group of ten notes played in the time of eight or four.

Dedication. An address or inscription to a patron or friend, prefixed to a work.

Dedications frequently form a valuable guide to the historian, as by them it can be ascertained whether the author designed to honour any special individual, or, in the case of early works, whether a production was issued at the "cost and charges" of any particular patron.

Before the time when an author could command a large sale amongst the general public, it was not an uncommon practice to dedicate a book to one who had borne the chief expense in the production, and the ingenuity of the author was exercised in finding expressions sufficiently flattering in return for money expended or presented. The character of these addresses became at one time somewhat fulsome, as may be seen by the following, prefixed to Clifford's "Divine Services and Anthems:" London, 1663; the first book of its kind printed in England:

"To the Reverend Walter Jones, Doctor in Divinity and Sub-Dean of his Majesties Chappel-Royal, &c.:

"Sir,—Under your able patronage I have presumed to shelter this my weak endeavor, which if for no other reason than the wel-meaning devotion thereof, I was sure would not be unacceptable or troublesome to you. Be pleased therefore to intermit awhile those seraphical raptures, in the excellency whereof, and your thereto tuned piety, you are so famously happy. And vouchsafe an eare to the mean addresse of these rudiments (as it were) of Church Musick, which, like other perfections, hath suffer'd meerly through the peoples ignorance. To you therefore more especially doe I dedicate this essay, whose alone competent skill and judgement in the highest mysteries of this divine science, if it shall please you to descend and deign a favourable approbation thereunto, cannot but comand reception from others: since my knowledge at Oxford (improved further at London) of your eminency this way, cannot so far disoblige the world as not to believe you have the supreme mastery in religious musick; by which, as you charm the soul, and all its affections, no doubt you can prevail upon and perswade publick acceptance.

"I submit this piece in this (howsoever rude) manner to your judgement, having attempted, I hope something of tendency to the churches peace and harmony, whereof though I am a smal and an unworthy member, yet a mite even from such is justly expected: For higher works God hath fitted and prepared your most artfull hand, and hath placed you in an orb from whence your melody (as of the spheres) of holiness and constant goodness in and for the church is universally heard with joy and delight. In which happiness, God Almighty long continue you here and late translate you to the angelical choire: So prayes, Reverend Sir, your most devoted and obedient Servant, JAMES CLIFFORD."

Master Thomas Mace, in a more manly mood, dedicated his famous book called "Musick's Monument" (1676) to a higher power than a sub-dean, on the principle that a man's work should be "dedicated" to God, and only "inscribed" to a fellow-man. His "Epistle Dedicatory" runs as follows:—

"To Thee, One-Only-Oneness, I direct my weak desires, and works; please to protect both them, and me; for Thou alone art able, (and none but Thee) to make us acceptable unto the world.

"I am not of that Catholick belief,
(I mean the Roman's faith) who seek relief
(At th' second hand) from saints; but I thus take
My freedom, and (sans complement) thus make
My seeming-bold-address: not judging it
A crime with Thee; but rather count it fit;
Part of my duty call'd for, which I owe
Unto Thy goodness; therefore thus it show.
I've wonder'd much, to see what great ado
Men make, to dedicate their works, unto
High mortals, who themselves can no way save
From the sland'rous tongues, of every envious knave.
Thou (only) art the able-true-protector:
Oh be my shield, defender, and director
Then sure we shall be safe

Digitized by Google

Thou know'st, (O searcher of all hearts) how I,
With right-downright-sincere-sincerity,
Have longed long, to do some little good
(According to the best I understood),
With Thy rich tallent, though by me made poor;
For which I grieve, and will do so no more,
By Thy good grace assisting, which I do
Most humbly beg for: Oh adjoyn it, to
My longing-ardent-soul; and have respect
To this my weak endeavour; and accept
(In Thy great mercy) both of it, and me,
Ev'n as we dedicate our selves to Thee."

This is followed by "An epistle to all divine readers, especially those of the discenting ministry, or clergy, who want not only skill, but good-will to this most excelling-part of divine-service, viz., singing of psalms, hymns, and spiritual songs, to the praise of the Almighty, in the publick assemblies of His saints; and yet more particularly to all great and high persons, supervisors, masters, or governors of the Church, (if any such should be) wanting skill, or good-will thereunto."

In 1713 Mattheson published a sonata "dedicated to the person who will best perform it," and if it were necessary, many curious instances of remarkable dedications might be quoted to swell the list, but one only must suffice. There is extant a composition by Samuel Wesley, containing a series of intended violations of musical grammar, all of which are duly pointed out, and the whole is dedicated "without permission to William Horsley, Esqre., Mus. Bac., fifth and eighth catcher in ordinary and extraordinary to the Royal Society of Musicians."

Deductio (*Lat.*) The succession of notes as they appear in their proper places in the hexachords, which are in consequence called prima *deductio*, secunda, &c., up to *septima*. [Notation.]

Deficiendo (*It.*) Gradually dying away.

Degré (*Fr.*) Degree of a scale.

Degree of a scale. A step in the tone-ladder; it may consist of a semitone, a tone, or (in the minor scale) of an augmented tone.

Degree in music. The rank or title conferred by an University on a candidate who has matriculated, and passed through the necessary examinations. They are of two kinds, Bachelor in (or of) Music, and Doctor of Music. The latter is generally taken by bachelors of several years standing, but in special cases candidates are allowed (by a grace) to accumulate, that is, take both degrees at the same time.

Dehnung (*Ger.*) Expansion, extension.

Dehnungstriche (*Ger.*) A long stroke with the bow.

Délassement (*Fr.*) A light trifling entertainment.

Deliberatamente (*It.*) Deliberately.

Deliberato (*It.*) Deliberate.

Del, della, delle, dello (*It.*) Of the; *e.g.*, *sopra il soggetto della fuga seguente*, on the subject of the fugue which follows.

Delicato, delicatamente (*It.*) Delicately; *delicatissimo*, very delicately; *con delicatezza*, with delicacy.

Délicatesse (*Fr.*) Delicacy of performance.

Delirio, con (*It.*) With excitement, with frenzy.

Delyn (*Welsh.*) The harp.

Démancher (*Fr.*) To cross hands, in pianoforte playing. To shift, in violin playing.

Demande (*Fr.*) The subject, *dux*, or proposition of a fugue.

Demi-bâton (*Fr.*) A semi-breve rest.

Demi-cadence (*Fr.*) A half cadence, or the cadence on the dominant. [Cadence.]

Demi-jeu (*Fr.*) Half power. *Mezzo forte*, applied to organ or harmonium playing.

Demi-mesure, demi-pause (*Fr.*) A minim rest.

Demi-quart de soupir (*Fr.*) A demisemiquaver rest.

Demi-semiquaver. A note of the value of one-fourth of a quaver

Demi-soupir (*Fr.*) A quaver rest.

Demi-ton (*Fr.*) A semitone.

Demoiselle (*Fr.*) A coupler in the organ.

Denis d'or. An instrument having a finger board like a piano, and pedals like an organ, capable of producing a vast number of different qualities of sound. It was invented in 1762 by Procopius Divis, in Moravia.

Derivative. (1) The actual or supposed root or generator, from the harmonics of which a chord is derived. (2) A chord derived from another, that is, in an inverted state. An inversion.

Des (*Ger.*) D flat.

Descant, Discantus (*Lat.*) The addition of a part or parts to a tenor or subject. This art, the forerunner of modern counterpoint and harmony, grew out of the still earlier art of diaphony or the organum, of which it is necessary to give a slight sketch.

Diaphony (διαφωνία) signified in Greek music discordant sounds or dissonance (voces discrepantes vel dissonæ), as opposed to symphony (συμφωνία) consonance. But the term came afterwards to be applied to those first attempts at the harmonic combination of voices, and polyphony, which may be looked upon as the first life-pulse of modern harmony. It is indeed strange that the term *diaphony* should have been selected for these early efforts, for, crude and painful as they are to our ears, they gave undoubted pleasure to those who first listened to them, who speak of their "melodiæ suavitas" and "dulcis concentus;" moreover, diaphony was well known to signify *dissonance*, intervals being

Digitized by Google

divided into symphonic and diaphonic, the former including 4ths, 5ths, and octaves (and their compounds); the latter 2nds, 3rds, 6ths, and 7ths. That they should not have called it "harmony" would not surprise us if they were *only* cognizant of the exact force of the Greek *ἁρμονία*, but it happens that Isidorus of Seville (in the 6th century), gives the following definition of harmony, "consonantia plurimorum sonorum et coaptatio," a definition which so far diverges from that of the Greek, that it might have included all the efforts of the diaphonists, and indeed might almost pass muster as a definition of *modern* harmony. Why diaphony or the *diaphonia cantilena* was also called *organum* it is difficult to say. A later explanation, namely, that it was because of a supposed similarity to the music of the instrument "organ" is plainly untenable (apte similitudinem exprimat instrumenti, quod organum vocatur)—J. Cotton in Gerbertus Script. II. 263. In time of Charlemagne the art of diaphony must have reached some degree of perfection, as it is certain that Roman *cantors* were called upon to teach certain French chanters the *ars organandi*, and *pueri symphoniaci* were part of the musical staff of Vitalian at the end of the 7th century. The earliest forms of diaphony were of four kinds; when the organum was added to the "principal" or subject *throughout* at the interval (1) of an octave, (2) of a fifth (3) of a fourth, (4) at an octave above *and* below.

These species were combined in three or four-part music, thus presenting to the ear a simple succession of consecutives. It is unnecessary to give examples of each of the above: the following (from Gerbertus) will suffice as illustrations:—

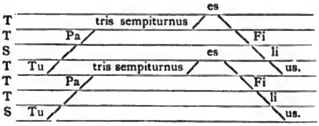

In the above S stands for semitone, T for tone. It is thus rendered by Gerbertus (De Cantu, &c., I. 112.):

Ex. 1.

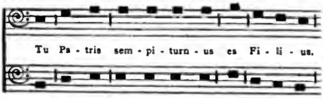

He also gives the following, but in old notes and clefs:

Ex. 2.

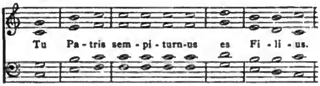

The next example is from Kiesewetter (Hist. of Modern Music):

Ex 3.

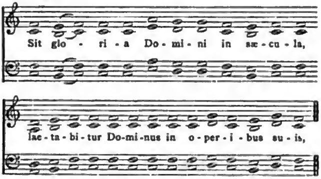
Sit glo - ri - a Do - mi - ni in sæ - cu - la,
læ - ta - bi - tur Do-mi-nus in o - per - i - bus su - is,

In another kind of organum, one voice held on a note like a drone while a second sang the tune.

But the first step towards harmony was to allow the *organizing* voices to have a choice of intervals, instead of compelling them, as seen in the examples above, to move in parallel quarts, quints, and octaves. The following example (No. 4) is from Gerbertus:

Ex. 4.

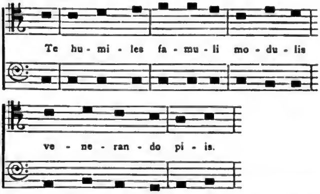
Te hu - mi - les fa - mu - li mo - du - lis
ve - ne - ran - do pi - is.

In the above it will be observed that several 3*rds* are introduced; the admission of this interval (both minor and major) into the list of consonances represents another forward movement in the art of music. But 3rds (and afterwards 6ths) were not allowed to be *perfect*, but classified as *imperfect* consonances, a title which to this day adheres to these beautiful intervals, although their superiority to 4ths and 5ths has compelled modern writers to call a 4th a *discord*, and to forbid consecutive 5ths as an *abomination*. Franco and Garland have a division of intervals which is interesting as marking the transition from the treatment of 3rds and 6ths as discords, and their present position in music. These authors (see Coussemaker, p. 49) classify the unison and octave as perfect consonances; the 4th and 5th as middle consonances; the major and minor third as imperfect consonances.

The next important step in the progress of harmony seems to have been the giving due consideration to what we now term the relative *motion of the parts*. In the early examples (1 to 3) it must have been observed that there is nothing but *similar motion;* in Ex. 4, *oblique* motion is mixed with *similar;* and next we find authors boldly laying down the law that when the principal (or melody) ascends, the added organum should descend, and *vice versâ* (ubi in recta modulatione est elevatio, ibi in organica fiat depositio et e converso). This acknowledgment of the beauty and value of *contrary motion* must have given a new impulse to the art. Thus step by step did it grope its way, till in the 11th century important treatises on it were produced, and there can be also traced that sure sign of a healthy circulation of thought, a marked partisanship of different and opposed systems. After this period, instead of the "principal" and its "organum," we begin to read of a "tenor" and its "descant," and by almost imperceptible degrees the old system dies away as the new is grafted and feeds on it. It must not, however, be supposed that the successive changes in diaphony, above sketched, occurred in the exact order in which they are presented to the reader. Systems of art are never of sudden growth, they overlap each other; they perhaps grow side by side for years, perhaps for centuries, until those which have in them the smallest power of development decay, and leave the less-matured but better-constituted systems to survive, with fresh opportunities of thriving.

Descant, *Discantus* (Lat.), may be said to have come into existence at the end of the 11th or beginning of the 12th century. The word itself is thought by some to be merely a latinized synonym of diaphony; others, among them Franco, considered it to be connected with *de cantu*, something framed on or growing out of a melody. Originally, as had been previously the case with diaphony, it consisted of two parts only, but later in its life developed into motetts and various other forms of composition. The real difference between diaphony and descant seems to have been that the former was rarely, if ever, more complicated than *note against note*, whereas descant made use of the various proportionate value of notes—"Discantus est aliquorum diversorum cantuum consonantia, in quo illi diversi cantus per voces longas, breves, vel semibreves proportionaliter adæquantur, et in scripto per debitas figuras proportionari ad invicem designantur" (Johan. de Moravia). It also included notes altered semitonally by accidentals, under which circumstances it was called musica ficta vel falsa, feigned or false music. This fact adds interest to a discussion which took place between M. Fetis and M. Coussemaker. The former taxed the latter with having misinterpreted his published specimens of diaphony, because, in it were introduced the tritone and the lesser 5th. M. Coussemaker's answer seems conclusive—it is to the effect that if the parallel motion of diaphony were consistently preserved, either such intervals must occur, or accidentally altered notes, outside the scale, must have been introduced; and such has not been proved to have taken place.

The rules of descant are numerous, and they provided for the regular addition of one part to any other part according to the movement of the tenor. The particular interval by which the *tenor* proceeded, dictated to the *descant* its own progression. It would be useless to give the rules here, as they can only be mastered by the student who carefully reads the many treatises on the subject, so elegantly reprinted by the indefatigable M. Coussemaker (Script. de Musica Medii Ævi). The scope of these rules was from time to time expanded, and in a treatise of the 14th century by Theinred, reprinted by W. Chappell (in the "Choir" newspaper of April 9, 1870), the divergences of the later from the earlier systems are apparent.

At this time, or even earlier, such expressions as *cantus firmus* and *contrapunctus* begin from time to time to be used, giving the first intimation of the art of adding *counterpoint* to a *canto fermo*, which was soon to supersede descant, as descant had superseded diaphony.

Opinions have been divided as to whether descant was merely a form of regulated improvisation, or whether it was a written art. In all probability, it grew from one state into the other. At first, without doubt, its rules were intended to direct a musician how to add "contrappunto alla mente;" afterwards, when an interesting and successful descant had been framed, and perhaps often repeated, it would indeed be strange if the author had made no attempt to lengthen its existence by committing it to paper; indeed, one sentence from Johannes de Muris substantiates the fact beyond dispute. He says, "Nihil enim prohibet in duobus cantibus simul esse cantantes plures, tam in tenore quam in discantu." It is inconceivable that a number of voices could be found to add descants *impromptu* without serious discrepancies. Descants were sometimes sung without words.

Descending. Passing from a higher degree of pitch to a lower.

Des dur (*Ger.*) The key of D♭ major.

Des moll (*Ger.*) The key of D♭ minor.

Dessin (*Fr.*) The design or plan of a composition.

Dessus (*Fr.*) One of the old names for the treble or *upper* part in vocal music.

Desto (*It.*) Sprightly.

Destra (*It.*) The right, as *mano destra*, the right hand.

Détaché (*Fr.*) Detached, or staccato notes.

Determinato (*It.*) Resolutely, definitely.

Détonation (*Fr.*) False intonation.

Détonner (*Fr.*) To sing out of tune: to sing harshly or coarsely.

Detto (*It.*) The same; as, *il detto voce*, the same voice.

Deutsche Flöte (*Ger.*) The German flute. [Flute.]

Deutscher Bass (*Ger.*) An instrument of the viol kind, with five or six gut-strings, midway in size between a violoncello and a double-bass.

Deuxième position (*Fr.*) (1) The second position or half-shift on the violin. [Shift.] (2) The second fret on a guitar.

Development of a subject. The elaboration of a given theme according to the rules of art. [Sonata.]

Devoto, Devozione, con (*It.*) With devotion, affection.

Dextra (*Lat.*) The right, *e.g.*, *dextra manu*, with the right hand.

Dextræ tibiæ (*Lat.*) Pipes held in the right hand; generally, the shorter of the *tibiæ impares*. Hence, *tibiæ dextræ* seems to imply a pair of *treble* pipes; *tibiæ sinistræ*, a pair of lower-toned or bass pipes. [Aulos.]

Di (*It.*) By, of, for, with. *Di grado*, by degrees; *stromenti di fiato*, wind instruments; *di chiesa*, for the church; *di bravura*, with *bravura* passages.

Diagonal Bellows. An old form of organ bellows, the construction of which may be thus explained.

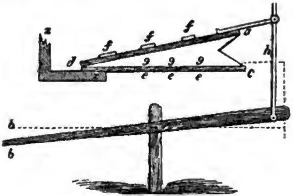

When the bellows are empty, the top (*a*), which is moveable (being hinged at *d*), lies close to the bottom (*c*) which is a fixture, and the handle (*b*) with its levers are in the position described by the dotted lines. Starting thus, if the handle be pressed down, as it leaves the dotted line, the top (*a*) will ascend, and air will enter the bellows through the apertures *e*, *e*, *e*. When the handle, having reached its maximum depression, is released, the surface weights (*f*, *f*, *f*) exercise their influence to restore the top (*a*) to its horizontal position; but in the mean time, the valves (*g*, *g*, *g*) inside the bellows have fallen over the apertures (*e*, *e*, *e*), and prevented the egress of the air through them; the air, therefore, is necessarily forced through *z*, the only exit left, into the sound boards below the pipes. It is evident from this, that during the time the handle is being pushed down, no air is being forced into the sound boards, because, the handle while being depressed negatives the effect of the surface weights. Hence, two diagonal bellows, at least, were absolutely necessary in every organ, whilst some had as many as 10, 12, or 14. The organ in St. Paul's Cathedral had originally 4 large diagonal bellows, measuring 8 feet by 4; and that in St. Sulpice, Paris, 14 diagonal bellows. This large number of bellows was sometimes arranged in a row, side by side, sometimes in two rows, one of which was placed over the other. In the latter case, ropes attached to the handles of those in the upper row allowed them to be blown from the same level as those below. Sometimes diagonal bellows were inflated by treadles, so arranged that the blower could easily step from one to another. whence the Ger. "Bälgetreter," "Calcant." One of the chief defects in the diagonal bellows was its inability to supply wind of an uniform pressure. This arose from two causes: the first, because the sides of the folds as they turned inwards, changing from an obtuse to an acute angle, gave more pressure as the top of the bellows gradually fell; the second, because the surface weights would exercise more and more pressure as the top, starting from an inclined plane, approached the horizontal. This defect was in time remedied by attaching to the rod (*h*) or end of the handle (*b*) a spring, whose tension was greatest when the top (*a*) was highest, and when therefore, as explained above, the air was least compressed. [For an account of Horizontal Bellows, see Organ, § 2.]

Dialogue. A duet.

Diana (*It.*), **Diane** (*Fr.*) An aubade, huntsup.

Diapason (*Gk.*) (1) An octave. [Greek music.] (2) The name given in this country to the most important foundation stops of an organ, termed in other countries more properly *Principal*. There are two kinds of diapasons,

Digitized by Google

the open and stopped. Open diapasons on the manuals are nearly always of metal, but on the pedals are often of wood. Stopped diapasons were formerly, in most cases, of wood, but now are frequently made of metal. When two or more open diapasons are on the same manual, they are of different scales. (3) Fixed pitch; *normal diapason*, a recognised standard of pitch. [Pitch.]

Diapason cum diapente. The interval of a 12th.

Diapason cum diatessaron. The interval of an 11th.

Diapente (*Gk.*) The interval of a 5th.

Diapentissare (*med. Lat.*) To descant at the interval of a 5th.

Diaphony. [See under Descant.]

Diaschisma (*Gk.*) *διάσχισμα*. An approximate half of a *limma*.

Diastema (*Gk.*) *διάστημα*. An interval.

Diatessaron (*Gk.*) The interval of a fourth.

Diatonic. (1) One of the three *genera* of music among the Greeks, the other two being the chromatic and enharmonic. [Greek Music.] (2) The modern major and minor scales. (3) Chords, intervals, and melodic progressions, &c., belonging to one key-scale. A diatonic chord is one having no note chromatically altered. A diatonic interval is one formed by two notes of a diatonic scale unaltered by accidentals. A diatonic melody is one not including notes belonging to more than one scale. A diatonic modulation is one by which a key is changed to another closely *related* to it.

Diaulion (*Gk.*) *διαύλιον*. An air played upon the aulos or flute during an interval in the choral song.

Diazeuxis (*Gk.*) *διάζευξις*. The separation of two tetrachords by a tone: opposed to synaphe (*συναφή*), or the overlapping of tetrachords. [Greek Music.]

Dichord. (1) An instrument having two strings. (2) An instrument having two strings to each note. [Bichord.]

Di colpo (*It.*) Suddenly, at once.

Diecetto (*It.*) A composition for ten instruments.

Diesare (*It.*), **Diéser** (*Fr.*) To sharpen.

Dièse (*Fr.*) A sharp, ♯

Diesis (*Gk.*) *δίεσις*. Originally the name of a *semitone*, called afterwards a limma. In later writings, applied to a third or quarter of a tone in the enharmonic and chromatic scales. The modern enharmonic diesis is the interval represented by 125 : 128; that is, the difference between three true major thirds and one octave.

Dièze (*Fr.*) A sharp, ♯

Diezeugmenon. [Greek Music.]

Difficile (*It.*) Difficult.

Di gala (*It.*) Merrily, cheerfully.

Digitorium. The name of a small portable dumb instrument, invented by M. Marks, for the purpose of strengthening and giving flexibility to the fingers for pianoforte playing. It consists of a key-board with five keys, kept in their places by springs of metal.

Di grado (*It.*) By conjunct intervals.

Dilettante (*It.*) [Amateur.]

Diludium. An interlude.

Diluendo (*It.*) Wasting away, diminishing; *decrescendo.*

Diminished. Made less. (1) Diminished intervals are those made less than minor, *e.g.*: G♯ to F♮ is a diminished 7th, because G to F being a minor 7th, G♯ to F contains one semitone less than the minor interval. Some authors, however, apply this term in a manner liable to lead to much confusion, namely, to a perfect interval when made smaller by *one* semitone, and to an imperfect interval when made less by *two* semitones; thus, according to them, C to G♭ is a diminished 5th, but C to E♭♭ or C♯ to E♭ a diminished 3rd. [Interval.] (2) Diminished subjects or counter-subjects are subjects or counter-subjects introduced with notes half the value of those in which they were first enunciated. (3) A diminished triad is the chord consisting of two thirds on the subtonic, *e.g.*, B, D, F in the key of C.

Diminué (*Fr.*), **Diminuito** (*It.*) Diminished.

Diminuendo (*It.*) Decreasing in power of sound.

Diminution. [Canon.]

Di molto (*It.*) Very much; as, *allegro di molto*, very fast.

Din-din. An Indian instrument of the cymbal class.

Di nuovo (*It.*) Anew, again.

Dioxia (*Gk.*) The interval of a 5th. This term was afterwards superseded by *diapente*.

Direct. A sign (*w*) placed at the end of a line or of a page of music, to indicate the note next to be sung or played:

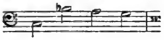

Directeur (*Fr.*), *Direttore* (*It.*) Director. Manager, guide, conductor of an orchestra.

Direct motion. [Motion.]

Dirge. A solemn piece of music, of a funereal or memorial character, so called from the first word of the Antiphon, "Dirige, Domine Deus meus, in conspectu tuo, viam meam." The office of burial of the dead was called in the Primer (cir. 1400) *Placebo* and *Dirige*, and in the Primer of Henry VIII. (1545) is called The Dirige.

Diretta, alla (*It.*) In direct motion.

Dis (*Ger.*) D sharp.

Di salto (*It.*) By a leap, spoken of melody progressing by skips.

Digitized by Google

Discant. [Descant.]

Discantgeige (*Ger.*) An old term for the violin.

Discantschlüssel (*Ger.*) The soprano clef. Descant clef. [Clef.]

Discord. A chord which when struck or sung requires to be resolved into a concord. [Harmony.]

Discreto, con discrezione (*It.*) Prudently, discreetly, with judgment.

Disdiapason (*Gk.*) An interval of two octaves; a 15th.

Disinvolto (*It.*) Free, unfettered, naturally.

Disjunct motion. [Motion.]

Dis moll (*Ger.*) D sharp minor.

Disperato, con disperazione (*It.*) Despairing, with desperation.

Dispersed harmony. Harmony in which the notes composing the chord are at wide intervals from each other:

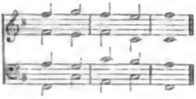

Disposition. Arrangement (1) of the parts of a chord, with regard to the intervals between them; (2) of the parts of a score, with regard to their relative order; (3) of voices and instruments with a view to their greatest efficiency or to the convenience of their positions; (4) of the groups of pipes in an organ, or of the registers or stops bringing them under control.

Dissonance. Discord. [Harmony.]

Dissonare (*It.*) To jar, to make discord.

Distanza (*It.*) Distance, an interval.

Distinto (*It.*) Clear, distinct.

Dithyrambus (*Gk.*) A song in honour of Bacchus, from which arose the first dramatic representations in Athens. The choruses to the early tragedies were in dithyrambic form.

Dito (*It.*) A finger.

Dito grosso (*It.*) The thumb.

Ditone. An interval of two major tones. This interval exceeds the major third, which consists of a major and minor tone, and is discordant.

Ditonus (*Lat.*) [Ditone.]

Ditty. A short, simple air, implying or containing a moral application. The word is said to be derived from the Latin word dictum, and signified a saying or sentence, not always connected with rhythm or music. [Ballad.]

Divertimento (*It.*) (1) An instrumental composition in several movements (six or seven) like a *serenade* or *cassation*. (2) A Pot-pourri.

Divertissement (*Fr.*) [Divertimento.]

Divisi (*It.*) Divided. A direction that instruments playing from one line of music are to separate and play in two parts. The reunion of the parts into unison is directed by the words *a due*, *e.g.*:

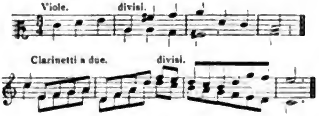

Division. An elaborate variation for voices or instruments upon a simple theme; a course of notes so connected that they form one series. Divisions for the voice are intended to be sung in one breath to one syllable. The performance of this style of music is called *running* a division:

HANDEL.

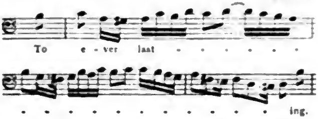
To e - ver last . . . . ing.

Division viol. A violin with frets upon the finger-board. [Violin.]

Divotamente, divoto (*It.*) *Divozione, con* (*It.*) Devoutly, devotedly, with devotion.

Dixième (*Fr.*) The interval of a 10th.

D moll (*Ger.*) The key of D minor.

Do. The first of the syllables used for the solfeggio of the scale. The note c, to which it is applied, was originally called Ut [Aretinian syllables], and is still called so in France. Its introduction dates from the 17th century. Lorenzo Penna in his "Albori Musicale," 1672, uses *do* for *ut*, and speaks of it as a recent practice. [Solfeggio.] [Notation.]

Doctor of, or in, music. [Degree.]

Dodecachordon (*Gk.*) (1.) An instrument with twelve strings. (2.) The title of Glareanus' great work on music, 1547.

Dodecuplet. A collection or group of twelve notes to be played in the time of eight.

Doigté (*Fr.*) (from Doigter, to finger). Marking by signs or numerals the manner in which a piece of music should be played by the fingers. [Fingering.]

Dolcan. [Dulciana.]

Dolce. A soft-toned 8-ft. organ stop.

Dolce (*It.*) Sweet; *dolce maniera*, in a sweet style.

Dolcemente, Dolcezza, con (*It.*) With softness and sweetness.

Dolciano, Dolcino (*It.*), *Dulcan* (*Ger.*), *Dulzaginas* (*Sp.*) [Dulciana.]

Digitized by Google

Dolcissimo (*It.*) With the utmost degree of sweetness.

Dolente, dolentemente, dolentissimo, con dolore, con duolo, doloroso (*It.*) In a plaintive, sorrowful style; with sadness.

Dolzflöte (*Ger.*) The old German flute, with seven ventages and one key.

Domchor (*Ger.*) The choir or body of singers in a cathedral church, usually consisting of boys and men.

Dominant. (1) The fifth degree of the scale. [Harmony.] (2) The reciting note of Gregorian chants. [Chant.]

Dominante (*Fr.*) Dominant.

Donna, prima (*It.*) The principal female singer in an opera.

Doppelbee (*Ger.*) A double flat, ♭♭.

Doppelflöte (*Ger.*) An organ stop, consisting of wood pipes having each two mouths.

Doppelfuge (*Ger.*) A double fugue; a fugue with two subjects. [Fugue.]

Doppelgeige (*Ger.*) One of the names by which the viole d'amour, *q.v.*, is known in Germany.

Doppelgriffe (*Ger.*) Double-stopping on a violin; playing on two strings at once.

Doppelkreuz (*Ger.*) A double sharp ×.

Doppelschlag (*Ger.*) A double beat or grace note. [Beat.]

Doppio (*It.*) Double, *e.g.*, *doppio movimento*, at double the pace; *doppio pedale*, the pedal part in octaves

Dopo (*It.*) After.

Dorian mode. [Greek Music.] [Plain Song.]

Dorien (*Fr.*) Dorian

Dot. (1) A point added to a note, or rest, which lengthens its value by one-half, *e.g.*: 𝅝 . is equal to 𝅗𝅥 𝅗𝅥 𝅗𝅥; 𝄽 . is equal to 𝄾 𝄾 𝄾 When a second dot follows the first (when the note or rest is *doubly dotted*), the second dot adds one-half of the value of the previous dot *e.g.*: 𝅝 . . is equal to 𝅝 𝅗𝅥 ♩; 𝄽 . . is equal to 𝄽 𝄾 𝄿 A dot was called the *point* of addition (punctus), hence a dotted note was called formerly a *pricked note;* this expression must not, however, be connected with *prick-song*, which signifies written music, as opposed to music sung by ear.

(2) When placed over a note, the dot is a direction that the note is to be played or sung *staccato*.

(3) When two or four dots are placed in the spaces of the stave, on either side of two double bars, they are a direction to repeat so much of the music as is enclosed between them.

(4) When placed under a slur, dots are a direction to play *spiccato*, that is, in violin playing, played by the same *bow*, but the bow must remain stationary between each sound. From violin music the term has been transferred to that for the pianoforte, and sometimes for the voice.

(5) A system of Tablature for wind instruments, the *dot system*. [Tablature.]

(6) Dots were formerly placed over a note to show its subdivision into lesser repeated notes, *e.g.*, would be equal to 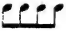

Double (*Fr.*) A turn:

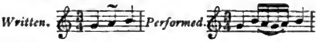

Double. (1) An old term for a variation. In some of Handel's harpsichord lessons, the variations of a theme are marked Double 1, Double 2, &c. A variation on a dance tune is called a double. (2) The repetition of words in singing was also called the "doubles or ingeminations thereof." (3) An artist who understudies a part in an opera, that is, who prepares a part on the chance of the accidenta absence of the principal. (4) That which is an octave below the unison in pitch, *e.g.*, *double-bass*, an instrument whose sounds are an octave below those of the violoncello; *double-bassoon*, an instrument similarly sounding an octave below the bassoon; *double-diapason*, an organ stop of 16-ft. pitch.

Double action. [Harp.]

Double backfall. An ornament in old music, *e.g.*:

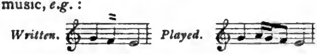

Double bar. A sign formed of two single bars showing (1) the end of a piece, (2) the end of a movement of a work, (3) the end of a portion to be repeated, (4) the commencement of a change of key, (5) the commencement of a change of time, (6) the end of a line of words set to music, as in a hymn tune [Bar.]

Double-bass. *Violone* (*It.*) *Contre-basse* (*Fr.*) The largest of the stringed instruments played with a bow. The strings are usually tuned a fourth apart to the following notes when three strings are employed:

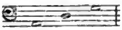

with the addition of the lower E:

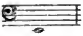

when there are four strings. The compass generally written for the instrument extends to the upper F:

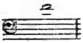

with every intermediate semitone from its lowest note. The actual sounds produced are an octave lower than written, hence the double bass is sometimes called a transposing instrument. The four-stringed double bass is more common abroad than in England, so it is not unusual to find passages written below the lower A, which a three-stringed bass performs on the octave above. Double notes are possible upon the instrument, but are rarely employed as they are ineffective. Continuous rapid passages are best divided between two instruments, but short quick runs are very telling especially when in unison or at the octave with the violoncello. Beethoven in the Pastoral Symphony takes his double bass down to C, an octave below the violoncello:

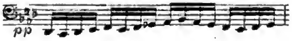

a passage which is impracticable upon the instrument as now generally constructed.

Beethoven also wrote passages for the instrument which, in his time, were considered too difficult for performance, and it was the custom for the players to "simplify" whenever his works were performed. But the improved skill of the players of the present time has justified the composer's foresight, and all the so-called difficult passages are given with ease and distinctness, even by the least distinguished double-bass player in the orchestra. The bow employed is the only representative now in use of one of the primitive forms, and although it has the advantage of producing a thick, heavy quality of sound in slow movements, it is not always successful in eliciting an even tone in quick passages.

The harmonics on the double bass are of a beautiful flute-like character, and have been made available by solo players in exciting wonder and admiration.

The mute is rarely, if ever, employed, but the *pizzicato* on the instrument has a very fine effect, as in the overture to "Der Freischütz," and elsewhere.

The story of the development of the double bass from earlier deep-toned instruments is involved in obscurity. It is impossible to say definitely at what period or by whom the doubling at the 8va was introduced, although Gasparo da Salò (1542-1609) was an early maker. The introduction of the instrument into the orchestra is due to Michael Monteclare, about the year 1696. Before this time the Bass-viol or Viola da Gamba was the deepest-toned stringed instrument employed. The "Contra Basso di Viola," mentioned in the score of Jacopo Peri's "Eurydice," is held to have been a larger sort of tenor violin, less in size than a bass-viol, and not a double bass.

Double Bassoon. The deepest-toned instrument of the Bassoon family. It stands in the same relation to the bassoon as a double bass does to the violoncello, that is to say, its sounds are actually an octave below those written. Its compass is

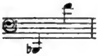

that is to say, from the B♭ below CCC to tenor F. It forms, in the orchestra, a magnificent support to the wind band, but good players are not commonly to be met with, partly because the large size of the instrument renders it very unwieldy, partly on account of the fatigue which the performer necessarily must undergo. The common habit of replacing it by an ophicleide should be discouraged, as the quality of the two instruments differs greatly.

Double beat. An ornament of old music, consisting of a beat repeated.

Double bourdon. An organ stop of 32-ft. tone. On the manuals it rarely goes below middle C; on the pedals it extends of course through the whole compass. It consists of stopped wood-pipes. It is found difficult to produce a pure tone in the longer pipes, as the first harmonic has a strong tendency to assert itself. [Bourdon.]

Double chant. [Chant.]

Double chorus. A chorus for two separate choirs: the several themes may be distinct, or so constructed that united they form one harmony. [Chorus.]

Double counterpoint. [Counterpoint.]

Double croche (*Fr.*) A semiquaver.

Double demisemiquaver. A note whose value is one half of a demisemiquaver.

Double diapason. [Double, § 4.]

Double dièze (*Fr.*) A double sharp.

Double drum. A drum with two heads, used in the bands of foot regiments, and being suspended from the neck of the player is struck with drumsticks held in the right and left hands. [Drum.]

Double flageolet. A flageolet having two tubes and one mouth-piece, admitting of the performance of simple music in thirds and sixths, &c.

Double flat. A sign (♭♭) used in music before a note already flattened in the signature, which depresses the note before which it is placed another half tone. It is contradicted by a natural and a flat.

Double octave. The interval of a 15th.

Double pedal point. A portion of a fugue or melody in which two notes are long sustained, generally the tonic and dominant. [Fugue.] [Sustained note.]

Double quartet. A composition for two sets of four voices or instruments *soli*.

Double reed. (1) The vibrating reed of instruments of the oboe class. (2) A reed-stop on an organ of 16-ft. pitch.

Double relish. An ornament in old music:

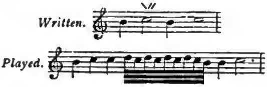

Double root. [Extreme sixth.]

Double sharp. A sign (×) used before a note already sharp, to indicate that it is desired to raise the pitch by a semitone. It is contradicted by a natural and a sharp.

Double sonata. A sonata for two solo instruments, as pianoforte and violin, or two pianofortes, &c.

Double stopped diapason. [Bourdon.]

Double-stopping. The stopping of two strings simultaneously with the fingers in violin playing. The practice was first suggested by John Francis Henry Biber in 1681, in a set of solos for a violin and a bass: one of these pieces is written in three staves, two for the violin playing in double-stopping, and the third for the bass. He also in the same work suggests a varied tuning in fourths and fifths for the purpose of making the double-stopping easy.

Double-tongueing. A peculiar action of the tongue against the roof of the mouth used by flute players, to ensure a brilliant and spirited articulation of staccato notes. The term is sometimes applied also to the rapid repetition of notes in cornet playing.

Double travale. A direction in tambourine playing. [Tambourine.]

Down beat. The first beat in each bar is so called, because in counting time the hand or conducting stick is allowed to fall at that place.

Down bow. The bow drawn over the strings from the heel or holding part of the bow to the point; the greatest power of tone in the strings is elicited by the down bow. [Bowing.]

Double trumpet. An organ reed-stop similar in tone and scale to, but an octave lower in pitch than, the 8-ft. trumpet.

Doublette (*Fr.*) A compound organ-stop consisting of two ranks, generally a twelfth and fifteenth.

Doucement (*Fr.*) Softly, sweetly.

Doux (*Fr.*) Soft, sweet.

Douzième (*Fr.*) A twelfth.

Doxology (*Gk.*) The hymn or song of praise—the Gloria Patri—used at the end of the Psalms in the Christian church; also any metrical form of the same.

Doxologia magna (*Lat.*) The version of the angels' hymn, "Gloria in excelsis Deo," sung at the celebration of the Holy Eucharist. The greater doxology.

Doxologia parva (*Lat.*) [Doxology.]

Drag. (1) An ornament consisting of descending notes in lute-music. (2) A *rallentando* (in old English church music).

Dramma lyrica, or per musica (*It.*) [Opera.]

Drammaticamente (*It.*) } In a dramatic
Drammatico (*It.*) } style.

Dreichörig (*Ger.*) The triple stringed grand pianoforte. A trichord.

Dreiklang (*Ger.*) A chord of three sounds. [Triad.]

Dreistimmig (*Ger.*) Music in three parts.

Dritta (*It.*) Right; *mano dritta*, right hand.

Driving notes. Syncopated notes. Notes driven through the ensuing accent.

Droite (*Fr.*) Right; as *main droite*, right hand.

Drone. (1) The monotonous bass produced from the larger of the three tubes of bag-pipes. As there are no governing holes in the drone the sound it gives forth serves as a continuous bass to any melody; the pipe second in size is tuned to give out the fifth above the drone; and the smaller pipe, called the chanter, has ventages by which the melody is made. [Bagpipes.] (2) The chorus or burden of a song.

Druckbalg (*Ger.*) A reservoir of wind, as in an organ, &c.

Drum. An instrument of percussion, of cylindrical form, having discs of vellum or parchment at each end, so made that the discs can be tightened or slackened at pleasure by means of braces acted upon by sliding knots of leather, or by the later application of screws. There are many kinds of drums: (1) The long drum, with two heads, held laterally and played on both ends with stuffed-nob drumsticks held in the hands of the performer. (2) The side-drum having two heads, the upper one only being played upon by two sticks of wood; the lower head has occasionally strings of catgut stretched over its surface, and then it is called a *snare* drum. (3) The kettle drum, always employed in pairs. (4) The bass-drum or grosse-caisse which is played with two sticks unless the player has to strike cymbals with his left hand.

The drum is an instrument of Eastern origin; it was employed by the Hebrews, Romans, Parthians, and other nations in religious dances, and as signals of war, and was probably first brought to Western Europe by the Crusaders or their followers. [Naker, nakeres.] [Kettledrum.]

Drum-major. The name of an officer in the British army who is responsible for the instruction of drummers in the various roll-calls, and for the invention and construction

Digitized by Google

of new beats, communicated by order of the major of the regiment to the drummers. The office does not appear to be older than the time of Charles II. There was formerly an officer in the Royal Household called the drum-major general, who granted licenses to other than the royal troops for the use of drums in their regiments.

Drum slade. An old term for a drummer.

D string The 3rd open string on violins, the 2nd on tenors, violoncellos, and three-stringed double basses, the 4th on the guitar.

Duan (*Gaelic.*) A verse, stanza.

Ductus (*Lat.*) [Agoge.]

Due, a (*It.*) [Divisi.]

Due corde (*It.*) (1) Two strings. A direction that the same note is to be played simultaneously on two strings of a violin or other instrument of its class. (2) A direction to cease holding down the soft pedal of a pianoforte.

Duet. A composition for two voices or instruments, or for two performers upon one instrument.

Duetto (*It.*) A duet.

Duettino (*It.*) A little duet.

Due volte (*It.*) Twice.

Dulçaynas (*Sp.*) The name of a larger sort of oboe, or small bassoon, “Se usa un genera de Dulçaynas que parecen nuestras Chirimias.”—*Don Quixote.* As it is supposed that the instrument was brought into Spain by the Moors, the word may be derived from the same root as the Egyptian Dalzimr, both instruments being of the oboe or reed kind.

Dulcian, or **dulcino** (*It.*) The name of a species of small bassoon. [Bassoon.] [Dulciana.] [Dulçaynas.]

Dulciana. A word now applied, in this country, solely to a soft and delicate-toned organ stop consisting of very *small-scale* flue pipes. Originally, a dulciana (dulcan, dulcian, dolcan, dolcin, or dulzain) was a kind of *hautboy* [Waits], and these terms are still found on some foreign organs as the names of soft *reed-stops*, as at Rotterdam, the Hague, and elsewhere, but in some cases the stop is not actually *reed*, but the pipes by their peculiar shape, narrow at the mouth and widening gradually towards the top, produce a reedy quality of tone. The dulciana stop was introduced into this country, or perhaps invented, by the celebrated organ-builder Snetzler. The first known specimen was included by him in the specification of the organ of St. Margaret's Church, Lynn, in 1754. Stops of this class are universally used, and are of great utility. They are most commonly found on the *Choir* organ.

Dulcimer. One of the most ancient musical instruments, used by various nations in almost all parts of the world, and which, in shape and construction, has probably undergone fewer changes than any other instrument. In its earliest and simplest form, it consisted of a flat piece of wood, on which were fastened two converging strips of wood, across which strings were stretched tuned to the national scale. The only improvements since made on this type are the addition of a series of pegs, or pins, to regulate the tension of the strings; and the use of *two* flat pieces of wood formed into a resonance-box, for the body. The word dulcimer is probably connected with *dolce*, sweet, through the intermediate word *dolcimela;* but the German name, *Hackebrett* (*chopping-board*), points to the manner in which it is played, the wires being struck by two hammers, one held in each hand of the performer. Perhaps the greatest divergence of form is to be seen in the Japanese *goto, or koto*, an illustration of which is now given :

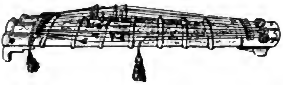

The next figure shows a dulcimer of Georgia :

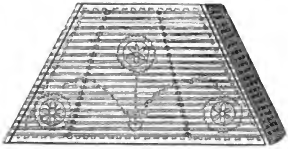

The Italians, who have ever been noteworthy for combining beauty with utility, have not failed to improve upon the original simplicity of the dulcimer, as the following illustration will shew :

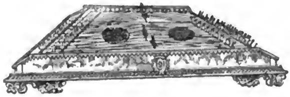

The form of the instrument given in the next figure, the dulcimer of Benares, suggests that it is not placed, for use, in the ordinary position:

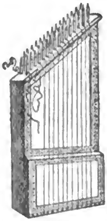

But, the fact which makes the dulcimer of the greatest interest to musicians is, that it is the undoubted forefather of our pianoforte. A modern grand pianoforte is, in reality, nothing more than a huge dulcimer, the wires of which are set in vibration, not by hammers held in the pianist's hands, but by keys: it is a *keyed-dulcimer.*

It is remarkable that in the immediate forerunners of the pianoforte (the spinet, harpsichord, &c.) the strings were plucked, so, the invention of "hammers" which constitutes the real difference between a pianoforte and a harpsichord, was in truth a return to a primitive type.

At one period the dulcimer came to be called in Italy a *cembalo*, possibly from its "ringing" cymbal-like tone, hence the same term was afterwards bestowed upon a *harpsichord* (*clavi-cembalo*). In full scores it is not even now an unusual thing to find the pianoforte part marked *cembalo.* The dulcimer is much less commonly met with in England than formerly, but it is still to be heard in some rural districts, as the musical accompaniment of a puppet-show.

The following Swiss dance of the Canton of Appenzell, as arranged for a violin, dulcimer, and bass (from a collection of Swiss songs, &c., Berne, 1826), will give a fair notion of the capabilities of the dulcimer:

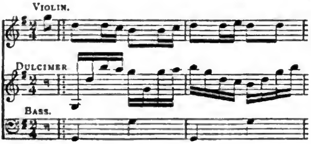

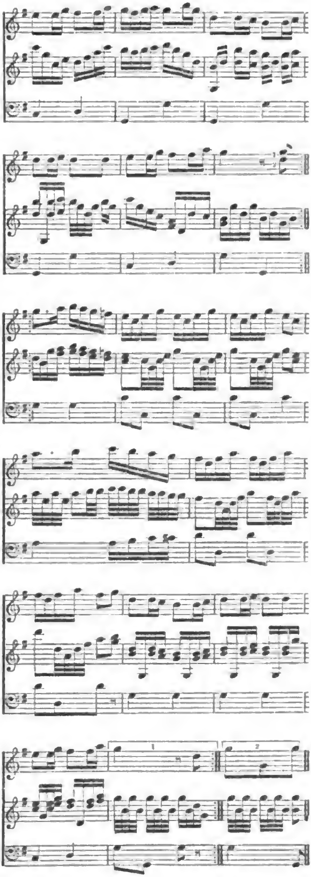

Dump or **dumpe.** The name of an old dance in slow time with a peculiar rhythm. It is doubtful whether it was entirely "dull and heavy," or merely the slowness of the measure that made the title of the dance synonymous with wearisomeness, for Shakespeare makes Peter, in "Romeo and Juliet," say, "O play me some merry dump," which may either have been descriptive of the character of the dance, or it may be a humorous contradiction in terms. Some authors have supposed that the dance is called *dump* from a trick of lute players who struck the open strings with the fist at certain marked intervals of the rhythm.

"My Ladye Carey's Dumpe." (cir. 1600.)

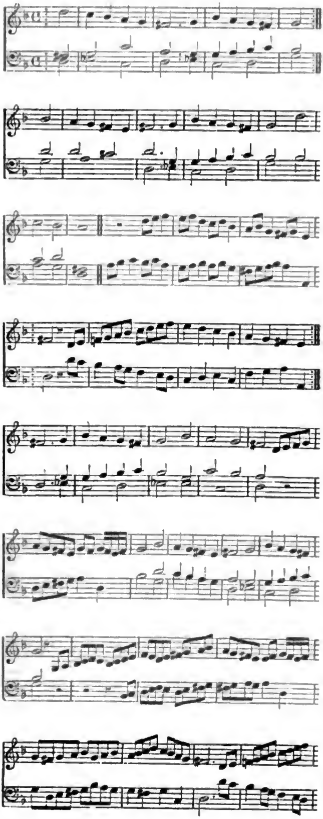

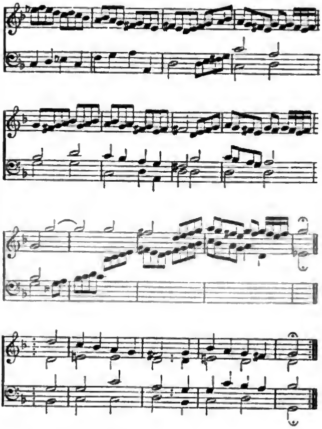

Duo (*It.*) A duet.

Duodecimole. A group of twelve notes.

Duodecimo (*It.*) The interval of a twelfth.

Duodramma (*It.*) A dramatic piece for two performers only.

Duolo, con (*It.*) With grief, sadness, pathos.

Duple. [Time-signature, § 2.]

Dur (*Ger.*) Major, as *C dur*, C major.

Dur (*Fr.*) Hard, coarse.

Durate, duramente, duro (*It.*) With harshness, roughly.

Durchführung (*Ger.*) The development of a theme or subject. [Form.]

Durchgehend (*Ger.*) Passing, transient.

Durezza (*It.*) Rigour, harshness.

Dutch concert. A so-called concert in which every man sings his own song at the same time that his neighbour is also singing his, a practice not necessarily so national as convivial. There is another form of Dutch concert, in which each person present sings in turn one verse of any song he pleases, some well-known chorus being used as a burden after each verse. When every person has sung his song, all sing their respective songs simultaneously as a grand *finale*.

Dux (*Lat.*) The proposition, theme, or subject of a fugue, the answer being called *Comes*.

Dystonic. With false intonation; discordant.

E.

E. (1) The note Hypate in Greek music. [Greek music.]

(2) The key-note of the Church mode called Phrygian.

(3) The note Elami in the system of Hexachords. [Notation.]

(4) The E above tenor C, the octave above it being represented by *e*, the octave below it by EE.

(5) The key having four sharps in its signature.

Ear. The Ear is the organ of hearing, in other words, the organ for the appreciation of sound, *i.e.*, of vibrations of the air or water. All that is necessary to form an ear is a nerve-mass capable of appreciating these vibrations. Its simplest actual expression is a sac, filled with fluids, containing "otoliths" (*οὖς, ὠτὸς*, an ear, and *λίθος*, a stone), and supplied with a nerve, a condition best exemplified in the sub-kingdom of Mollusca, represented familiarly by the oysters, the mussels, snails both terrestrial and aquatic, and the octopus. The "otoliths" are masses of carbonate of lime, as may readily be seen by placing one of them dissected out from any of the above mentioned animals (*e.g.* a snail) on a glass slide, covering it with an object glass with sufficient water to fill the interspace between the two, and adding at the side of the covering-glass a drop of any acid (acetic acid, or indeed ordinary vinegar will do very well) while the experimenter observes it through the microscope. Air bubbles—really bubbles of carbonic acid gas—will be seen to pour out from the otolith, and when these have ceased, that body will have entirely disappeared.

It is true that we are suspicious of a specialised organ of hearing even in such animals as possess no specialised nerve system. This is somewhat apparently of a paradox, for it may be asked, How can an animal without nerves feel at all? and is it not highly improbable that if no nerve system exists, any special sense-organ can be developed? The answer to the first of these objections is very plain: all animals, even those who possess no specialised organs whatever manifest the simple phenomena of sensation; all, even the Infusoria and the Amœbœ, of which we hear so much now-a-days, and which are very little more than simple masses of protoplasm, manifest this faculty. Any one who has observed these occupants of almost any drop of water, with a microscope of low power, will have seen enough to convince himself of this. As to the second objection, the fact remains that in some animals which have no undoubted nerves, in some medusæ[*] or jelly-fish,[†] we find in the mass forming their body crystals of carbonate of lime, which substance must intensify the vibrations of the water in which they live, and must, when put into a state of motion as a whole, or into one which affects its particles *inter se*, cause by its relative density greater disturbance of the soft matter in which it lies, than would be the case if it were absent.

But leaving this as somewhat problematical, and taking the hearing organ of molluscs as the type, we shall find this type essentially adhered to in the higher animals in spite of endless complications. Let us propound the bold paradox without fear of contradiction, that a man as well as a snail hears in *water*, and that the essential parts of his marvellous hearing apparatus are a sac containing fluid in which are otoliths, and round which are distributed the ultimate filaments of a nerve.

The typical physiology, as well as the typical anatomy of hearing, is very simple. These "ear-stones," by the vibrations conducted to them, are made to rattle in this bag containing fluid, and, by beating against its sides, cause more disturbance to the nerve filaments there distributed than would be caused by the same vibrations if they acted directly on the nerve.

The power of hearing must be very widely if not universally distributed through the animal kingdom, though the hearing organ is not always easy to find. Von Siebold[‡] has however discovered this organ in the European field-cricket, situated in the front legs of both sexes. That the lower animals have the power of making music very widely distributed among them is a fact of which any one who reads Mr. Darwin's "Descent of Man," part ii., can very easily convince himself; the references to this work would be too numerous to give with completeness, and extracts would be impossible, seeing that the subject occupies a considerable portion of the whole work. It is, however, sure that this power is possessed by some spiders (vol. i. p. 339); by many insects, as *e.g.* of the order Homoptera (p. 350—352), including the Cicados; of the order Orthoptera (352—360), including crickets and grasshoppers; of the order Hymenoptera (366), including bees and wasps; of the order Coleoptera (378—385), or beetles; of the order Lepidoptera (387), or moths and butterflies. All these animals possessing the power of music, which they use principally for attracting the other sex (the male being generally the musician), must possess also organs capable of appreciating such music—auditory organs.

---

[*] These medusæ are by some said to possess a ring of nerves, but this is disputed.

[†] "Gegenbaur, Grundzüge der Vergleichenden Anatomie," p. 129, fig. 15 *c*, and p. 131.

[‡] "Lehrbuch von Vergleichender Anatomie," von Siebold und Stannius. Berlin, 1848. Erster Theil, p. 582.

When we reach the sub-kingdom of Vertebrata we find the same type throughout, gradually becoming more complicated as we proceed from the lowest fish to the highest mammal, man; and it will be convenient to reverse the philosophical order and to describe the human ear first, since after the description of this it will be easy to allude to deficiencies or modifications in the less complicated organs of the lower vertebrata.

The *Human Ear* may be divided into three parts—the external, middle, and internal ear. The two former have the function of conveying vibrations to the latter which appreciates them.

Fig. 1

LEFT EAR.

(Transverse section).

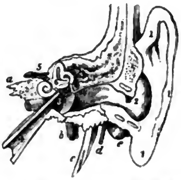

1. Pinna; 2. external auditory meatus or canal; 3. cavity of tympanum and membrane of tympanum (between 3 and 6 chain of small bones); 4. eustachian tube; 5. internal auditory canal or meatus giving passage to the auditory and facial nerves; 6. bony labyrinth (above fenestra ovalis); *a.* apex of petrous bone; *b.* internal carotid artery; *c.* styloid process; *d.* facial nerve passing to supply muscles of face; *e.* mastoid process.

The external ear (see fig. 1.) consists of two parts, the "pinna," or auricle (1), and the external auditory canal or "meatus" (2)

The pinna (1) is that part which is quite external, and which we unscientifically call "the ear," as when we say that a certain person has large or small "ears." It is composed almost entirely of cartilage or gristle, and has complicated foldings, to all of which names have been given, but which it is not our business to give here in detail. The general shape is that of an irregular funnel, having its apex in the auditory canal. The only part of which we shall speak particularly is the "helix," or the margin which is folded in. At the upper and posterior part of this is to be found, in many individuals, a small point or process, generally folded in like the rest of the helix, but sometimes projecting outwards. This is considered by Mr. Darwin to be a strong fact in determining the genealogy of man ("Descent of Man," vol. i. p. 22, fig. 2), for this point is well marked in many of the lower monkeys, as baboons and some species of macacus, and in them is not folded inwards, but stands erect. He thus considers it a reminiscence of some pointed-eared progenitors. He remarks (p. 23, note 26): "This rudiment apparently is somewhat larger in negroes and Australians than in Europeans (see Carl Vogt, "Lectures on Man," Eng. transl., p. 129), these races being confessedly lower and more like the lower animals than Europeans." The pinna is furnished with nine muscles, three of which are called extrinsic and move it as a whole, while the remaining six would, if they contracted, move its parts on one another, and are called intrinsic.

The extrinsic muscles are situated in front, above and behind, and move the pinna therefore forwards, upwards and backwards respectively. The anterior muscle is called the "attrahens aurem," the superior is called the "attollens aurem," and the posterior the "retrahens aurem." These muscles are in man without any function whatever; they do not have any effect on his hearing powers; they are capable of moving the ear only in some individuals, and are only rarely moveable at will. The "retrahens aurem," which is the strongest, is also the most commonly moveable; in some people it is contracted involuntarily in fright, just as in a timid horse who throws back his ears; it is much more rarely under the control of the will, but is less seldom so than the other muscles. Mr. Darwin ("Descent of Man," vol. i. p. 20) says: "I have seen one man who could draw his ears forwards, and another who could draw them backwards;" but this gives too great an idea of the rarity of such an accom-

Digitized by Google

plishment. A case has been observed in which a boy was able to move his scalp from side to side by alternately contracting the "attollens aurem" of either side, this muscle rising from the "aponeurosis" (or flat tendon) of the "occipito-frontalis" muscle, which has the power of moving the scalp forwards and backwards.

The intrinsic muscles are situated, four on the outer, two on the inner side of the pinna. They have never been known to contract. In some animals, however, they are functional, as any one can see who carefully watches the pinna of the ear of a cat when intent on some sound.

In spite of the complicated structure of the pinna, it is nearly sure that it possesses no effect on our powers of hearing, either by collecting sound to a focus or by conducting it along its substance;* that it is in fact of no use as far as hearing is concerned. This is the result of the observations of no less an authority than Mr. Toynbee.† It seems also that the nearly precisely similar ears of the orang and chimpanzee are equally functionless ("Descent of Man," vol. i. p. 21).

The external auditory canal (fig. I. 2) is about 1¼ in. long; rather less than the external half is formed of cartilage or gristle, the remainder of bone. Its direction is not directly inwards, but slightly forwards also. It is closed at its inner end by the "membrana tympani" (fig. I. 3), or membrane of the drum. The glands which secrete the wax (ceruminous glands) are situated in the cartilaginous part of the canal, and agree in their structure with the sweat-glands.

Fig. II.

SMALL BONES OF TYMPANUM OR OSSICULA AUDITUS.

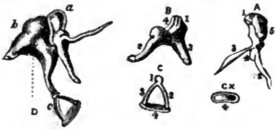

A. Malleus.—1. head; 2. handle; 3. processus gracilis; 4. short process. B. Incus.—1. body; 2. long process with orbicular process; 3. short or posterior process; 4. articular surface receiving head of malleus. C. Stapes.—1. head; 2. posterior leg; 3. anterior leg; 4. base. C*. Base of stapes. D. Bones in natural mutual relations.

The middle ear or tympanum (fig. I. 3) is separated from the external ear by the membrana tympani (fig. III. *m′*, and fig. I. 3), which inclines outwards, making an angle of 45° with the floor. It is a cavity which is not shut off from the air, for the "Eustachian Tube" (fig. I. 4) forms a communication between it and the "pharynx," the upper part of the cavity of the throat.

Fig. III.

OSSICULA AUDITUS AND MEMBRANA TYMPANI IN SITU.

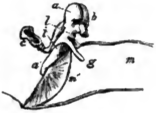

a. Malleus; *a′*. tip of handle of malleus, the letter lies in the cavity of the tympanum; *b*. incus; *c*. stapes; *m*. external auditory canal; *m′*. membrane of tympanum; *t*. line of tension of tensor tympani muscle pulling lower part of malleus inwards; *l*. line of tension of laxator tympani muscle pulling upper part of malleus inwards; *g*. slender process of malleus. The axis round which the chain of bones rotates passes through its base.

In the tympanum are situated three small bones (fig. I. 3 to 6, and figs. II. and III.), the Malleus (A), Incus (B), and Stapes (C), the names being derived from their shape. The malleus (hammer) has a round head (1), and a handle (2), and from the base of the head, a thin spike of bone, the "processus gracilis" (3) projects. The Incus (anvil), B, is more like a tooth with two fangs, a long and a short one. The long process (2) carries a knob or tubercle which is originally a separate bone, as it remains in some animals through life. The stapes (stirrup), C, is just like a stirrup. It is very difficult to understand the arrangement of these little bones from description or even from drawings, a model or the actual objects being almost necessary. It may, however, give a general idea of their position to say that the handle of the malleus and the long process of the incus are directed vertically downwards; the slender process of the malleus and the short process of the incus are horizontal, the former being directed forwards, the latter backwards; the stapes is also horizontal, but with its base horizontal and directed inwards. The malleus is anterior in position to the incus. The head of the malleus fits on to the body of the incus (the crown of the tooth), the long process of the incus fits on to the head of the stapes (the part where the stirrup-leather would be attached), and the base of the stapes fits loosely into the "Fenestra Ovalis," an oval window in the bony wall of the internal ear or labyrinth. The whole chain of bones turns round an axis formed by the slender process of the malleus and the short process of the incus. The handle of the malleus is firmly fixed to the membrana tympani on its inner aspect a little below the middle. There are three muscles in connection with this chain of bones, two are attached to the malleus, one to the stapes. The two attached to the malleus are the tensor and laxator tympani. The former is attached just below, the latter just above the origin of the processus gracilis (see fig. III. *t*, *l*), which, as we observed, is the axis of this chain of bones. The stapedius muscle is attached to the neck of the stapes. These bones are exceedingly important from the point of view of comparative anatomy, since their homologies play very different parts in the lower animals. The homologue of the malleus, for instance, in fish, amphibia, reptiles, and birds, is the "os quadratum" which suspends the lower jaw; and the stapes is in batrachia, reptiles, and birds, the "columella," a long bone, shaped somewhat like a straight post-horn, or stethoscope, which alone discharges the function which these three bones discharge in mammalia. Other interesting points will be related later on in connection with development.

---

* In using a stethoscope, however, we use the conducting power of the cartilage of the ear.

† "The Diseases of the Ear," by J. Toynbee, F.R.S. 1860, p. 12.

The cavity of the tympanum is practically enlarged by communicating with the "Mastoid Cells," air cavities which occupy the mastoid process of the temporal bone, that process of bone which may be felt behind and below the pinna, and is supposed by phrenologists to be the residence of "Pugnacity," though they have never explained the connection between that propensity and the function which these air-cells really discharge, that of increasing the tympanic cavity.

The internal ear or labyrinth (fig. IV.) is the essential part of the organ. It consists of two parts, a bony cavity enclosed in the thickness of the base of the skull, and a membranous sac within this.

Fig. IV.

RIGHT BONY LABYRINTH

(Smaller figure real size).

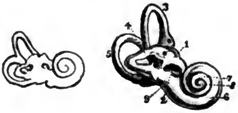

1. Vestibule; 2. fenestra ovalis; 3. superior semi-circular canal; 4. horizontal or external semi-circular canal; 5. posterior semi-circular canal; 6. first turn of cochlea; 7. second turn; 8. apex; 9. fenestra rotunda; * ampullæ of semi-circular canals.

The bony labyrinth may be briefly described as a chamber, the "Vestibule" (fig. IV. 1), which sends one prolongation forward (the "cochlea," 6, 7, 8), three others backwards ("semi-circular canals," 3, 4, 5), and has its outer and inner walls perforated, the outer by the fenestra ovalis (2), in which lies the base of the stapes, and by a round hole closed by membrane, and called the fenestra rotunda (9); the inner by a series of holes in a depression called the "Fovea hemispherica," which transmit branches of the auditory nerve from the internal auditory meatus in which lie the auditory and facial nerves. By these two lateral perforations it communicates with the cavity of the tympanum externally, and with that of the cranium internally. Close behind the "Fovea hemispherica" is a small canal, the "Aquæductus Vestibuli," to which reference will be given later, in describing the course of the development of the ear.

The bony semi-circular canals (fig. IV. 3, 4, 5) are three tubes bent so as to form about two-thirds of a circle. They are situated at the upper and back part of the vestibule with which they communicate by five openings, one end of the superior having an opening common also to the posterior semi-circular canal. Each tube at one end has an expansion, called an "Ampulla" (fig. IV.*) These canals are called from their position, superior, posterior, and external. The superior canal is vertical and transverse, the posterior is vertical and longitudinal, and the external is horizontal. The directions of these canals, or the planes in which they lie, will be best understood by placing a book with the two covers at right angles to one another, upright on end on a table, so that one of the covers faces the reader, the other being at right angles to the side of the table at which he is seated. Then the reader will be on the outer or tympanic side, the side of the table opposite to him will be the side of the cranial cavity. The plane of the table will represent the plane of the external or horizontal canal, the plane of the cover opposite to the reader the posterior, and that at right angles to the side of the table at which he is seated the superior canal, which is also the most anteriorly placed of the three. Thus it will be seen that the planes of these three canals are the three planes of a cube, a fact to which allusion will be made hereafter.

---

* With regard to the terms anterior, posterior, external, and internal, it may be necessary to explain that anterior means on the side towards the face; posterior on the side towards the back of the head, and external and internal remote from or near to an antero-posterior axis drawn from the face to the back of the head.

Digitized by Google

Fig. v.

VERTICAL SECTION OF COCHLEA, SHOWING ITS THREE PASSAGES OR SCALÆ.

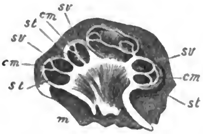

sv. Scala vestibuli; *cm.* canalis membranea, or canalis cochleæ, or ductus cochlearis; *st.* scala tympani; *m.* modiolus.

The cavity of the vestibule is prolonged anteriorly by the cochlea (fig. IV. 6, 7, 8), so-called from its likeness to the shell of a snail. As a whole, it forms a blunt cone with its apex outwards; this cone is formed by a gradually tapering spiral tube, the first curve having its concavity upwards; it is coiled 2½ times round a central column or "Modiolus" (fig. v. *m*), which sends an incomplete partition into the cavity of the tube (fig. VI. 3). This partition is called "Lamina spiralis ossea," and winds in the cavity of the spiral cochlea like the thread of a screw or the staircase in a turret; it is wanting at the apex of the tube. This lamina is completed by two membranes, that nearer the apex of the cochlea called the "Membrane of Reissner" (fig. VI. 1), that nearer the base, the "Membrana basilaris" (fig. VI. *mb*), so that three canals are formed, that on the side of the apex of the cochlea being called the "Scala vestibuli" (fig. v. and VI. *sv.*), that next the base called the "Scala tympani" (fig. v. and VI. *st*), and the intermediate one, belonging to the membranous labyrinth (here on the outer wall of the cochlea not lying free), called the "Canalis membranacea, vel Cochleæ," or "Ductus Cochlearis" (fig. v. *cm* VI. *a*). The scala vestibuli and scala tympani communicate at the apex of the cochlea, for the lamina spiralis does not extend quite to the apex, the scala vestibuli communicates below with the cavity of the vestibule as its name implies, the scala tympani would communicate below with the cavity of the tympanum through the fenestra rotunda, but that this is closed by a membrane. Thus it would be possible to get through the fenestra ovalis into the vestibule, thence enter into the semi-circular canals posteriorly, or anteriorly through the scala vestibuli to the apex of the cochlea, there into the scala tympani, through it to the fenestra rotunda, and through it again into the tympanum. Most books describe the cochlea as divided into two passages, the scala vestibuli and scala tympani, but it is both more accurate and plainer to describe three passages from the first, otherwise it is impossible to account for the canalis membranacea in the description of the membranous labyrinth. The difficulty in understanding this part consists, as will be presently seen, in the fact that whereas all other parts of the membranous labyrinth lie freely in the bony labyrinth, the canalis membranacea is not free at its outer side, the side farthest from the modiolus, but is there attached to the bony labyrinth.

The membranous labyrinth lies, except in one part already alluded to, freely in the cavity of the bony labyrinth, and corresponds almost exactly with it. Between the two is a fluid, the "perilymph," or "liquor Cotunnii;" and within the membranous labyrinth is the "endolymph," another fluid. The membranous labyrinth is the part of the internal ear which is essential to hearing, the bony labyrinth serving to enclose and protect it. The membranous vestibule is divided by a constriction into two halves, which do not communicate. The posterior and larger is called the "common sinus," or "utricle;" with it communicate the membranous semicircular canals, which correspond in arrangement with their bony cases. The anterior and smaller chamber is called the "saccule;" it becomes constricted anteriorly into a narrow canal, called the "canalis reuniens," which opens into the "canalis membranacea" of the cochlea. This latter canal is, as above described, interposed between the scala vestibuli and scala tympani. It ends blindly above at the apex of the cochlea. "Otoliths," or "otoconia" (ear-dust, *Gr.* οὖς, ὠτός, an ear, and κονία, dust), are found in the common sinus or utricle, in the saccule, and in the ampullæ of the semicircular canals; and besides them, the ampullæ are lined with long, stiff, hair-like filaments, called "fila acustica." They are six-sided crystals of carbonate of lime, with pointed ends, and lie in the walls of these parts of the membranous labyrinth. They are occasionally absent. In these parts we also find pigment cells, which seem in some mysterious manner to be essential to the sensitive parts of nearly all the special-sense organs; for they are present in the olfactory region of the nose, as well as in the globe of the eye, and only in the latter is their function known. It is a well known fact, that white cats (cats which have no pigment) are deaf.

Within the canalis membranacea cochleæ, and separated from it by a membrane called the "membrana tectoria," lies an assemblage

Digitized by Google

of structures known as the "organ[*] of Corti," after its describer. This is one of the most beautiful, as well as marvellously complicated, of all the structures in the body.

Fig. vi.

TRANSVERSE SECTION OF ONE COMPLETE PASSAGE OF COCHLEA.

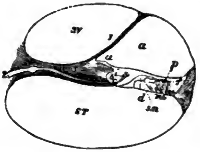

1. membrane of Reissner: 2. auditory nerve; 3. lamina spiralis ossea; 4. spiral ligament; *sv.* scala vestibuli; *st.* scala tympani; *a.* canalis membranacea or ductus cochlearis, or canalis cochleæ; *bc.* sulcus spiralis; *p.* membrana tectoria; *d.* rods of Corti; *f.* cells of Corti and Deiters; *i.* cells of Claudius; *mb.* membrana basilaris; *sm.* scala media.

The organ of Corti, then, has a floor called the "membrana basilaris," attached at the inner side to the lower lip of the free edge of the lamina spiralis ossea, and at the outer side to the circumferential wall of the cochlea. This membrane separates it from the scala tympani. Its roof is a membrane attached on the inner side to the upper lip of the free edge of the lamina spiralis ossea, and at the outer side to the circumferential wall of the cochlea. This membrane separates it from the canalis membranacea vel cochleæ, or ductus cochlearis. The essential part of the organ of Corti is a double series of rods, whose bases are separated by some distance, while their upper ends meet at an angle, the continuous series of rods forming a sort of spiral gabled roof, gradually diminishing as it follows the spiral course of the cochlea. The regularity of their arrangement, seen from above, suggests the key-board of a pianoforte. They have been estimated by Kölliker as about 3,000 in number, and are composed of a dense material. The inner series are more closely set and more numerous than the outer, which they overlap. Both series are enlarged at their bases and heads, especially the latter. The space between their bases and below their junction is called the "scala media" of the cochlea. Thus there are four canals in the cochlea, though this last belongs really to the canalis membranacea. Besides these rods, there are other bodies called "Cells of the organ of Corti." Some are placed between the inner series of rods and the free edge of the lamina spiralis ossea, and are called the "Cells of Claudius." They stand in a single line; their upper ends are provided with stiff bristle-like prolongations, or "cilia." The outer set are like them, but are placed between the outer series of rods and the circumferential wall of the cochlea. They are called the "pedunculated cells of Corti," and are set three deep, their cilia projecting through holes in a membrane extending from the junction of the rods of Corti over the outer cells of the organ of Corti to the circumferential wall of the cochlea, and called by Kölliker the "membrana reticularis." Between the rows of the pedunculated cells of Corti are interposed the "spindle-shaped cells of Deiters," which are fusiform, as their name implies. All of the basilar membrane not otherwise covered, is covered by six-sided cells of epithelium.

The auditory nerve, by which we appreciate sounds, does not rise from the brain, properly so called, but from the medulla oblongata, or that uppermost part of the spinal marrow which is enclosed in the skull, in company with the facial nerve, which supplies all the muscles of the face with motor power. The fibres of the auditory nerve can be traced to the floor of the fourth ventricle, *i.e.*, the expanded upper end of the tube which the spinal marrow really forms; and across this floor, to the fissure or furrow which separates it into two halves. Here they are found to rise from a mass of gray matter, *i.e.*, an assemblage of nerve cells called the "auditory nucleus." Other inconsiderable fibres are added from nerve centres in the neighbourhood, and some fibres are said to be connected with the sensory roots of the fifth or trigeminal nerve, the nerve of facial sensation.[†] It is connected by a small filament with the facial nerve. These two nerves run together along the floor of the cranial cavity outwards and forwards, and leave that cavity by the "internal auditory meatus," a hole in the petrous part of the temporal bone. The facial nerve traverses a canal in that bone, and leaves the bone to emerge by a hole just behind the socket of the lower jaw. While in the temporal bone, it gives off a slender branch called the "chorda tympani," which crosses the membrana tympani and handle of the malleus, and gives a twig to the laxator tympani muscle. It subsequently joins the gustatory nerve, the nerve of taste. Besides this, the facial nerve gives a twig to the stapedius muscle. In the internal auditory meatus, the auditory nerve divides into two portions, both of which contain nerve cells. One division is supplied to the cochlea, the other to the vestibule. The division which goes to the cochlea pierces the bony wall of the internal auditory meatus, not by one but by many foramina in the centre of the base of the cochlea. The central foramen is larger than the rest, and contains a nerve destined for the last half-turn of the lamina spiralis. The rest of the nerve fibres surround this, and ascending in the substance of the modiolus, they are distributed to the rest of the spiral lamina, piercing its substance, and running outwards to the scala media. These nerves form a continuous spiral ganglion, *i.e.*, a plexus or network of nerve fibres, with the addition of nerve cells near the edge of the spiral lamina. It is probable, but not proven, that their ultimate ends are connected with the organ of Corti.

---

[*] Zeitschrift für Wissenschaftliche Zoologie, tom. iii.

[†] Hirschfeld and Leveillé, pl. 15. fig. v. 5.

The vestibular division of the auditory nerve is distributed to the saccule, utricle, and ampullæ of the semi-circular canals, *i.e.*, the parts containing otoliths.

The *development of the Ear* is conducted, in its first stages, like that of the eye or nose. The skin becomes bulged in on the side of the head, and forms a pit which sinks deeper and deeper, while the opening gradually narrows, and at last closes. Thus a closed cavity is formed, and this becomes the membranous labyrinth. Towards it the auditory nerve grows from the developing medulla oblongata. Next a prolongation grows upwards and backwards, persistent in the lower vertebrata but not in mammals, except as the rudiment called the "Aquæductus vestibuli." Next, three portions of the sac get pinched up in an elongated form, these ridges rise higher, and by-and-by the middle or most elevated portion of each becomes separated from the under-lying sac, by its walls coalescing in this part. Thus three tubes are formed open at each end, and one end of each becomes dilated. These are the three membranous semi-circular canals. Another prolongation grows forward and gradually becomes spiral. This is the canalis membranacea cochleæ. The bony case becomes developed round these from cartilage or gristle, from three originally separate pieces called "proötic," "opisthotic," "epiotic" respectively; *i.e.*, anterior, posterior, and superior ear-bones or cartilages. These are separate in cold-blooded vertebrata throughout life. The space between the membranous labyrinth and its bony case is filled up by connective tissue which gradually liquefies and forms the perilymph.

The cavity of the tympanum, the external passage and the eustachian tube are developed out of what is called "the first branchial cleft;" for in the embryo there are formed four arches which lie on the sides of the neck exactly like the arches of the gills of a fish, and between them are similar slits so that the cavity of the gullet here communicates with the outer world. These all close except the first, which remains as the external ear, cavity of the tympanum, and eustachian tube. Some of the clefts however remain occasionally during life. Out of the first arch are developed in order from above downwards, the malleus, a piece of cartilage called "Meckel's cartilage," and the lower jaw, an arrangement persistent in fish; out of the second arch the stapes, stapedius muscle, incus, besides other structures which indirectly suspend the larynx, viz., the styloid process, stylo-hyoid ligament, and part of the hyoid bone. During the whole of fœtal life, according to Kölliker, the tympanic cavity is filled with connective tissue which embeds the little bones, and this only becomes absorbed after the child begins to breathe. The pinna is a lappet developed behind the first branchial cleft.

Propagation of Sound.—In order correctly to understand the sense of hearing we must have acquaintance with the principal laws of acoustics involved. Sound travels through air at about the rate of 1050 ft. a second, in water at about four times this velocity, and in very elastic solid bodies eighteen times as rapidly. In passing from solids to water the velocity is diminished, and from solids to air still more so; the passage from water to solids is easy, but that from water to air or from air to water, very difficult. Vibrations lose much of their intensity in passing from air to solids. The cases of passage therefore from the medium of least to that of greatest density, *i.e.*, from air to solid, or from the medium of least to that of considerably greater density, *i.e.*, from air to water, are the cases of greatest difficulty in the transmission of vibration. A dry stretched membrane easily receives and transmits vibrations of the air; and such a membrane placed on the surface of water overcomes in a great degree the difficulty of the passage between air and water. This assistance is enhanced when the membrane is combined with some solid body. Any membrane conducts sounds well when only in water. Sounds, like light, are liable to be reflected whether travelling in water or air.

Certain terms require explanation. Sounds are "communicated" when they are merely conveyed from one sounding body to another, and this can take place in a noise as well as a musical sound. Sounds are "excited"* under two circumstances: when the body which is sounding and that to be excited have the same note and the vibration of one produces sympathetic vibration of the other, the bodies are mutually called "reciprocating," while if the vibration of one produces its harmonics in the other, the latter is said, with regard to the exciting body, to be "resonant." According to Helmholtz, "timbre" or "quality" depends on definite combinations of certain secondary sounds or harmonics with a primary or fundamental sound, and such combinations he calls "sound colours."

---

* This property has been utilized in such instruments as the viola d'amore.

Hearing.—Sounds may reach the auditory nerve either through the combination of specialized structures lying between the tympanum and the filaments of that nerve, or through the bones of the skull. In the normal state the latter road is so much less efficacious that it may be disregarded; but when the other route is obstructed or rendered impervious, it then becomes the medium of the communication of sound. That sounds do however reach the auditory nerve in health by this way, anyone may learn by closing his ears and then speaking or singing. Under some circumstances the bones of the skull are the better conductors of the two; a tuning-fork held between the teeth gives a distinctly-audible note long after its vibrations have become inaudible through the air. Sounds are heard under water by this means.

The External Ear.—The *pinna* or *auricle* is said by some authors to help us to hear by reflecting sound into the meatus and by propagating it through its substance to the bony part of the meatus and thence to the membrana tympani. Reflection can only be helped by the large hollow behind the meatus called the concha, and by the point in front of the meatus called the tragus, the concha reflecting the vibrations on to the tragus, and this reflecting them in turn into the meatus. The other parts of the pinna have been supposed to assist sound by conduction, their various folds having the function, according to this view, of receiving vibrations in various planes perpendicularly and thus most favourably for propagation. Another view regards these folds as instrumental in neutralizing conflicting sound waves, that the principal vibrations may be able to enter the meatus without interruption. All these are mere speculations. An animal with moveable ears, such as a horse, turns his ears to the source of sound, but we have no such power, the extrinsic muscles of our ear are generally quite functionless, and never, in any case, possess this power, the only one which would help us to utilize our pinna. Mr. Toynbee believed the pinna to be quite functionless in man (see above).

The *External Meatus* is undoubtedly functional in conducting sounds, its closure will instantly prove this. Its curved course proves that the vibrations must reach the tympanum after manifold reflection from its walls and not directly. It serves to conduct vibrations without dispersion to the tympanic membrane. The column of air which it contains increases the strength of the vibrations which reach it, and by lengthening the tube of the meatus by adding a tube externally, and thus lengthening the column of air, the sounds are much increased in intensity. Its walls must conduct vibration to the membrane of the tympanum, but this function is so inconsiderable that we may practically neglect it.

The *Middle Ear* or *Tympanum.*—The *Membrana Tympani* serves to conduct vibrations received from the external air to the three small bones, the malleus, incus, and stapes, and thus to the internal ear. It is usually in a state of moderate relaxation, and is made more tense by the action of the tensor tympani muscle, and less tense probably by that of the stapedius and perhaps the laxator tympani. The vibrations which it receives are derived from the air in the external meatus, and perhaps also from the bony ring in which it is set.

The state of moderate relaxation which is usual to it, is the most favourable state for vibrating in sympathy with sounds of a wide range. The membrane vibrates reciprocally as a whole if the sound is in unison with the note to which it is (so to say) tuned by the muscles of the small bones, *i.e.* its fundamental; or in resonance in divisions, if the note sounded is higher than this, one of its harmonics. That it is not always tuned to the very note sounded is obvious, when we consider that this can only be the case when one note only is sounded. It cannot, of course, vibrate reciprocally to a note lower than its fundamental. A membrane has a large power of vibrating sympathetically since its harmonics are very numerous. The effect of increasing the tension of the membrane may be easily tested by closing the nose and mouth and either blowing air out from the lungs or drawing it in. By the former we blow air through the eustachian tube into the cavity of the tympanum, and force the membrane outwards, by the latter we decrease the pressure in the tympanic cavity, and the external air forces the membrane still more inwards than is naturally the case. The result is in either case the same, the sense of hearing is on the whole impaired, though very high sounds are heard better than before. We have stretched the membrane, raised its fundamental note, and diminished its power of vibrating in sympathy with low notes, though we have at the same time increased its range of sympathy upwards. Still it is chiefly im-

Digitized by Google

proved for *reciprocal* vibrations, for a lax membrane divides itself far more readily into segments which vibrate in sympathy with harmonics, the strength of such vibrations being increased by the number of the segments into which it divides itself.

The tensor tympani (perhaps the laxator tympani, though some deny that this is a muscle at all) and the stapedius are the muscles which regulate the tension of the membrane. These two muscles are generally considered antagonists, and are supplied by different nerves; the former renders the membrane tenser, the latter, more lax.* They simply tune the drum of the ear, making the membrane tenser for high, laxer for low sounds; there being practically a degree of tenseness which is most fit for perceiving vibrations of a certain average pitch. For exceedingly loud noises, as explosions, the probability is that the membrane is made tense, since in this state it cannot vibrate so freely. Thus is probably explained the sense of effort which we feel when *expecting* a loud noise which may never occur.

The *ossicula auditus* or small bones of the tympanum, that is, the malleus, incus and stapes, move as one piece, though they are not so tightly joined together but that they can play on one another. It is possible that their particles may also vibrate, but they are not adapted for this, seeing that much vibration must be lost at the joints between them. The direction of the application of the force to the malleus is the same as that in which it acts through the stapes; a line perpendicular to the membrane is parallel with the long axis of the stapes, or perpendicular to the membrane closing the fenestra ovalis. Each time the membrane of the tympanum is bent inwards the base of the stapes is driven more deeply into the fenestra ovalis. The axis of the chain of bones is described above. Since the tensor tympani is attached to the malleus below, the laxator tympani above this axis, as far as their action on the membrane of the drum is concerned, the malleus is worked by the former muscle as a lever of the third, by the latter as a lever of the first order. These bones are covered with mucous membrane which must insulate them and tend to prevent the propagation of vibrations from them to the air in the tympanum.

The *Eustachian Tube* serves to equalize the pressure on the outer and inner sides of the membrane of the tympanum. It is naturally closed, being only open during swallowing or yawning, when the muscles of the palate, inserted on opposite sides of the tube are put into action and pull its sides apart, thus temporarily opening it. During a bad cold in the head we often become deaf, especially after blowing the nose. We simply perform the experiment mentioned above of blowing air through the eustachian tube into the tympanum, forcing the membrane outwards, and also rendering it too tense. The eustachian tube, however, is often swollen during a cold, and is pervious only to great pressure of air; it therefore collapses and imprisons this extra amount of air. Some people know that their best chance of relieving this uncomfortable state is to swallow or yawn, though they do not know the reason. This sometimes opens the eustachian tube, and the much compressed air escapes, the hearing being at once regained. If this does not succeed it is sometimes necessary to pass a catheter, a tube appropriately bent, along the floor of the nose and into the eustachian tube, thus opening it. A very ingenious method was invented by Politzer for this purpose; the patient, whose eustachian tube is impervious, is given a glass of water, and the surgeon, having closed one of his nostrils with one hand, inserts a tube into the other; the patient is then told to drink. As he does so, the surgeon blows through the tube, and while the muscles of the palate open, or tend to open the tube, the additional pressure dislodges the plug of mucus or whatever was closing the tube; it becomes pervious, and hearing is at once restored.

The *air of the tympanic cavity* probably plays little or no part in the production of sound, though some effect must theoretically be produced through it on the membrane closing the fenestra rotunda, and through it in turn on the labyrinth, especially the scala tympani of the cochlea, by means of the perilymph. That this effect must, however, be very small is shown by experiment; for vibrations are very ill-conducted from the moist side of a membrane to air, and from this air to water through a membrane stretched on its surface, *i.e.* from the inner moist side of the membrana tympani, to the air of the tympanic cavity, and from it to the perilymph by means of the membrane of the fenestra rotunda; whereas they travel with remarkable intensity between air and water when conducted from the first membrane vibrating in air to the second membrane stretched over water through a chain of insulated solid bodies capable of vibrating as a whole, the last of which communicates with a solid body in close apposition with the second membrane, *i.e.* from the membrana tympani to the membrane of the fenestra ovalis through the ossicula, covered with moist mucous membrane, the last of which, the stapes, has its base in close apposition with the membrane of the fenestra ovalis, being in fact imbedded in it. This is, therefore, the principal route of the vibrations in their passage through the tympanum.

---

* Some, however, consider both as *tensors* of the membrane.

The *membranous labyrinth* is, as above said, the essential part of the auditory apparatus, and hearing remains even if all the structures between it and the external air are disorganized.

As to the special function of its constituent parts, we know nothing certainly; suppositions have been made, but the theories propounded on the question have not advanced beyond the region of hypothesis. The vestibule is probably the most essential part, inasmuch as it is not only the first part of the ear to be developed in man, but is also the first part to appear in the series of vertebrate animals, being present in the lowest fishes, except the amphioxus, which has no distinct organ of hearing.

The *semi-circular canals* are supposed to help us to determine the *direction* of sounds, since they would, if prolonged, intercept vibrations in any direction, being in the three planes of a cube, and this arrangement is found in nearly all cases where they are present at all. But whether they actually fulfil this function is quite unproved. M. Flourens * has experimented on the subject by cutting one or other of these canals, but has not removed the difficulty.

The *otoliths* are supposed to intensify sound by striking against the fine endings of the auditory nerve as they vibrate.

The *cochlea*, by far the most complicated part of the ear, is involved in the same obscurity as regards its functions as the rest of the ear. The complicated structures of the scala media have been supposed merely to deaden vibrations after they have produced their effect on the auditory nerve, thus preventing confusion. But it is by far more likely that they have a higher office to fill, and it is now generally believed that they serve to distinguish pitch. The rods of corti especially seem adapted to this function, arranged as they are in regular graduated series, dense as they are in structure, and elaborate as is the distribution of the cochleal nerve in their neighbourhood. They are supposed to vibrate each in sympathy with one note, and to transmit the vibrations to the special twig of the auditory nerve with which each is supposed to be connected. Not only would they thus appreciate pitch, but since "timbre" or "quality" depends on the definite combination of harmonics with a fundamental note (as Helmholtz has shown) they would thus convey what he well calls "sound-colours" to the sensorium, these sound-colours being combinations of a fundamental tone with harmonics, various both in pitch and relative intensity. Dugès, who first propounded the theory that the cochlea was the organ by which we appreciate "pitch," called attention to the concomitant variations in the evolution of the cochlea and the range of the voice in the three classes of mammals, birds and reptiles; the former having the largest, the latter the smallest development of both cochlea and vocal range.

A "musical ear" consists in the power of appreciating and distinguishing aerial vibrations both simple and compound, just as "the good eye for colour" consists in the power of appreciating and distinguishing the simple and compound vibrations of light.

When we hear a sound all that is proven is that particular filaments of the auditory nerve have been excited, not necessarily that there has been any external cause for the sensation. Aural delusions occur, though not so commonly as optical delusions. The singing in the ears which people often hear when they are out of health, overworked, when the blood-vessels of the head are congested, when blood is extravasated, when they are under the influence of a narcotic poison, when they are about to faint; all belong to this category. Many people are painfully conscious, even for many hours after a long railway journey, of the note to which their carriage has been vibrating.

Comparative Anatomy.—The lowest sub-kingdom in which we find any specialised hearing organ is that of the cœlenterata, the familiar representatives of which are the jelly-fish and sea-anemones. In them the ear is simply a sac filled with fluid, in which are crystals of carbonate of lime, the whole called a "lithocyst" or "stone-sac," by which the vibrations are intensified. This is analogous with the primitive auditory vesicle of man.

A similar structure is found in the sub-kingdom "Vermes," or worms, in certain marine worms called "Turbellaria," a familiar representative of which is the worm often seen on the sea-shore and called the

---

* Solucha. Pflüger's Archiv, vol. viii., quoted in the *London Medical Record*, Feb. 11, 1874. Solucha has made further experiments which make this still more probable. On cutting one or more of these canals the animal executes certain disorderly movements. These are probably due to the fact that the animal has lost proper conceptions as to the position of its head, since somewhat similar movements follow from merely fixing the head unsymmetrically. These canals, probably, possess the function of informing the animal, by a series of unconscious impressions as to the exact position of its head in space, and each canal has an exact relation to a dimension of space

"sea-man's bootlace." Many of the higher worms, or annelids, represented by the leech and earthworm, have a pair of such organs in the head, connected by a nerve with the nerve-ring surrounding the gullet.

Arthropoda, the sub-kingdom containing crustacea (crabs, lobsters, shrimps, prawns, and the wood-louse); Insecta; Arachnida (spiders and scorpions); and Myriapoda (centipedes), have not all of them distinct hearing organs. In *crustacea* both closed and open hearing organs are found. In the higher crustacea they are found at the base of the inner or smaller pair of antennæ or feelers, and in them they are open. In another creature called Mysis they are placed in the tail, and are composed of an otolithic sac lined with hair-like bodies, reminding us of the human "fila acustica," which are, like them, connected with the endings of a nerve. In those which have open hearing sacs, the particles of sand which are washed in are utilized for otoliths, being fastened in regular order to certain of these hairs. This variety of hearing sacs among crustacea is very instructive; whether closed or open they are closely connected with the integument, and the fact of their being sometimes open, sometimes closed, reminds us of the gradual development of the human ear, which is at first merely a pit in the integument, and afterwards becomes a closed sac (Gegenbaur, loc. cit. p. 388).

Among *Insects* the power of hearing must be almost universal, since music is so widely distributed among them, but the organs themselves have not been satisfactorily made out in many cases. Some have thought them to be represented by a tight membrane near the base of the feelers, others by the feelers themselves; among the grasshoppers and crickets by a sac filled with fluid, connected with a nerve, enlarged as it spreads over the sac, the whole sac being placed below a delicate membrane forming the floor of a pit on both sides of the first abdominal ring. In some locusts it is placed on the basal division of the front pair of legs, and is composed of a vibrating membrane like the tympanum, in the neighbourhood of which is an air-chamber connected with one of the tracheæ or air-tubes which pervade the body.[*]

It may not be uninteresting to insert a detailed description of a very elaborate hearing organ which is found in some orthoptera. The passage is translated from Von Siebold and Stannius, "Lehrbuch der Vergleichenden Anatomie," part i. p. 582:

"Only in certain orthoptera has a paired organ been successfully discovered which seems provided with the necessary apparatus of a sense of hearing. This organ is represented in the Acrididæ by a depression or pinna surrounded by a horny ring and more or less vaulted over, in the bottom of which a tympaniform membrane is spread out; on the inner surface of the latter a pair of horn-like appendages rise, between which a vesicle filled with clear fluid, extremely delicate, is fastened as a membranous labyrinth. Connected with this is a special auditory nerve, coming from the third thoracic ganglion, which swells to a ganglion on the tympanic membrane, and ends in the immediate neighbourhood of the labyrinth with a number of little and extremely slender club-shaped rods; loosely surrounded with ganglion cells (? are these endings primitive nerve fibres). The locusts and achetidæ (grasshoppers) possess a similar hearing organ in the 'shins' of both forelegs. Some of the locusts possess on both sides of the 'foreshins,' close under the knee-joint, a depression, while others of this family of the orthoptera are provided on the same spot with two more or less spacious hollows, opening forwards by an aperture (auditory capsules). In these pits and also in the hollows of both anterior tibiæ of the locusts an oval tympanic membrane is fitted. Between the two tympanic membranes the main tracheal stem of the forelegs forms a vesicular swelling, on the superior end of which the auditory nerve rising from the first thoracic ganglion and running down with the main crural nerve swells to a ganglion. From this ganglionic enlargement a band-like nerve mass runs down on the gently excavated anterior side of the tracheal vesicle, upon which nerve mass a linear series of vesicles with watery contents rise which again contain those remarkable club-shaped and slender rods (? primitive nerve fibres). The two great tracheal trunks of the forelegs open with two wide funnel-shaped openings at the posterior edge of the Prothorax, so that here also part of the tracheal system as in acrididæ allows of a comparison with a tuba eustachii. In the achetidæ an opening closed by a silvery membrane (membrana tympani) may be seen on the external side of both forelegs close under the knee-joint, behind which a similar hearing organ is concealed.

"*Note.*—In acheta achatina and italica an equally large tympanic membrane is situated also on the inner side of the forelegs which, in acheta sylvestris, domestica, and campestris is only feeblv indicated on this inner side of the tibiæ."

Hearing organs are also found in cockchafers in the root of the posterior wings. In the larva of crabs Heusen has described an ear consisting of an otolithic sac, in the adult the otoliths disappear and the sac is curiously composed of three demi-canals. In the *diptera* or flies they are situated in the rudimentary wings called halteres.

Auditory organs are found in all classes of *molluscs* either connected with the nerve ganglia in the "foot" or motor organ, as in the lower forms (oysters, muscles, snails, and slugs), or with the ganglia below the œsophagus, as in the higher forms. In the cephalopoda (octopus, &c.) they are placed in the mass of gristle which composes the head, and in them the sac is complicated. In all, the type is essentially the same, viz., a membranous sac containing fluid and an otolith or otoliths and supplied by a nerve. In certain mollusca called brachiopoda, hearing organs are found only in the larval state. In all molluscs the otolithic sac seems to be lined with ciliated epithelium, *i.e.*, epithelium which is furnished with eyelash-like hairs which continually lash the fluid in which they live.*

---

[*] Gegenbaur, loc. cit. p. 389

Vertebrate Animals (Fish).—No hearing organ has been found in the lowest fish, the amphioxus or lancelot. In all other fish it is present, and it is very interesting to trace its gradual evolution as we proceed towards the higher representatives. In the myxinoids the vestibule, the only part of the ear constant in fishes, consists of a simple ring-shaped tube lined with cilia, and lies freely at the sides of the head, like the primitive auditory vesicle of the human embryo. In the lamprey the ear consists of a vestibule with two semi-circular canals each of which has an ampulla. In all the higher fishes the labyrinth is enclosed in the bony or cartilaginous skull, and consists of three semi-circular canals, the vestibule being divided into two divisions as in man, with "endolymph" and "otoliths." The otoliths are often very large, as in the cod tribe. In the whiting, for instance, they may easily be found as two porcelain-like bodies, somewhat crescent-shaped and grooved transversely. In the rays the vestibule is prolonged by a tube which opens on the upper flat surface of the head. No fish has an external ear, tympanum, membrana tympani, or cochlea, but the labyrinth is often connected with the air-bladder, either by a tube or by a chain of bones. The air-bladder is thus pressed into the function of assisting the hearing (Weber, "De aure et auditu," p. 1245). In typical fishes we thus get the representatives of the vestibule, saccule and utricle, each with its otolith, and three semi-circular canals of man.

Amphibia.—In those amphibians which retain their gills through life, such as the newt and salamander, we have very little more than we found in fishes. They possess an internal ear only, which consists of a vestibule, three semi-circular canals; and as an addition to the ear of fishes, a "fenestra ovalis," with a small plate closing it, representing the base of the stapes. In one called the axolotl (that animal on which Cortez fed his army) this plate is connected with a little bone, but none of these creatures have a middle or external ear.

In those amphibians which when adult have lost their gills, we find the following additions. The labyrinth or internal ear has an otolithic saccule; in addition to the plate of cartilage closing the fenestra ovalis, which has a small muscle to move it (like the stapedius in man), we find a long thin bone "columella" running through a tympanic cavity and connected with a third member, a small cartilage which is attached to a tympanic membrane, and has another muscle attached to it (like the tensor tympani in man). In these animals the middle ear or tympanum first appears; this cavity is filled with air and communicates by an eustachian tube with the cavity of the mouth. The tympanic membrane is on the level of the surface of the body—there is no external passage or meatus.

Reptiles.—We divide reptiles into ophidia or serpents, lacertilia or lizards, chelonia, or turtles and tortoises, and crocodilia, crocodiles and alligators.

In serpents we seem to have an ear inferior to that of frogs; they have no eustachian tube, the tympanum does not contain air, but a sort of packing material called cellular or connective tissue, the tympanic membrane, as such, is absent, the tympanum being closed externally by skin. This substitution of cellular tissue for air is very interesting when we remember that in the human embryo the reverse change takes place. With this exception the ear is the same as that of frogs.

In lizards we again find a tympanic membrane and cavity, a eustachian tube, and in some the commencement of an external ear. In the iguana, for instance, there is a slight fold of skin beyond the tympanic membrane, and this is again instructive, for in the human embryo the tympanic membrane is at first on the level of the skin; and the external ear, both the bony and cartilaginous parts of the meatus and the pinna, are subsequent additions.

In turtles and tortoises we find the tympanic cavity divided into two by a bony septum or partition, which, however, is incomplete. The cochlea makes its first appearance as a slight conical bud, as it does in the human embryo, and there is a fenestra rotunda.†

In crocodiles the cochlea becomes bent and divided into two scalæ. The tympanic membrane is placed at the bottom of a deep fissure, and protected by a flap of integument containing cartilage and capable of closing the slit by muscles attached to it; thus we have an external ear. The tympanic cavity communicates with air-cells in many of the bones of the head as in birds. Indeed, to the comparative anatomists to whom the striking difference between feathers and scales is the least difference in the world and quite unimportant, the crocodile is nothing but a bird with certain practically unimportant distinctions.

---

* Gegenbaur, loc. cit. p. 513.

† There is one lizard of very exceptional structure which possesses a cochlea with an indication of the spiral curve which afterwards produces the form which we find in man. Its name is Hatteria, but it is indeed so full of anomalies that it is best mentioned in a note, and not as a representative of lacertilia.

The ear of *birds* is composed of an internal ear consisting of a vestibule with a foramen rotundum, and a foramen ovale, a cochlea with an incipient spiral turn, three semi-circular canals, and two cartilaginous bands representing the lamina spiralis ossea; of a middle ear consisting of a cavity, filled with air, communicating with air-cells in most of the bones of the head, provided with a membrana tympani and eustachian tube, and with a columella or stapes. This columella which we have seen from amphibia upwards, is, as we have said, the stapes of man. It is shaped like a long post-horn, or like a stethoscope.

The external ear consists of an external auditory meatus, and an indication of a pinna in the form of a fold of skin just in front of the meatus; this is largest in the owls. In some birds as the bustards the meatus is surrounded by a ring of specialized feathers, which perhaps serve to reflect vibrations towards the tympanum.

The *Ear of Mammals* is in the main so like that of man that it will be sufficient to mention such differences as we find in different classes.

In the internal ear we find that the cochlea has a very variable number of turns. The hedgehog has one-and-a-half turns, the seal two, many ruminants somewhat more, next the camel, horse, and elephant, and many bruta (ant-eaters, sloths, &c.); the bats, apes, and man, two-and-a-half; most carnivora three, the pigs nearly four; the guinea-pig and agouti quite four; and the paca (a rodent) five; marsupials have a very varying number of turns, the kangaroo two-and-a-half (like ruminants which they represent among marsupials), and the opossum nearly five.*

The otoliths are not universally found among mammals. The labyrinth has many variations, for which reference must be made to the larger treatises. The "ossicula auditus" are very variable in shape; in the lowest order of mammals, the monotremata, including the ornithorhynchus (duck-billed platypus), and echidna (Australian ant-eater), the stapes is shaped like that of amphibia, reptiles, and birds; it is a long and thin bone without any division—a "columella," in fact. In cetacea or whales, dolphins and porpoises, the ear is very remarkable; the external auditory meatus is almost obliterated, in one dolphin hardly admitting a pig's bristle. It is probable that the vibrations of sound are communicated in them to the auditory nerve, not by the meatus but by the bones of the head, as in fish, and in them the tympanic bone, forming the wall of the tympanum and supporting the drum, is very dense and hangs almost independently, reminding one of the large otoliths of fish (Owen), though whether it really fulfils the same office it is difficult to say.

The pinna is absent in most seals, the mole, cetacea, and the ornithorhynchus, in most diving animals it is very small. In some bats it is enormously developed, and has vibratile movements by which it seems to act as a sort of tactile organ "relating to the perception of atmospheric impulses rebounding from surfaces near which the bat approaches in flight."† Spallanzani says that a bat, after being deprived of the power of sight, hearing, and smell, by having the eyes put out, and the ears and nostrils plugged, was still able to avoid obstacles and to pass through openings only just large enough to admit its body.

Some animals have the power of voluntarily closing their external meatus—the elephant and the water shrew, for instance.

Ecbole (*Gk.*) The terms *eclysis* (ἔκλυσις) and *ecbole* (ἐκβολή) refer to the *flattening* and *sharpening* of sounds to adapt them to a change of key-note.

Eccedente (*It.*) Exceeding. *augmented*, a term applied to intervals.

Ecclesiastical modes. [Plain song.]

Echeion. ἠχεῖον (*Gk.*) (1) A hollow vessel, generally of metal (χαλκεῖον) used as a drum or gong. (2) Metallic vases so arranged behind the seats of the ancient theatre as to reinforce the sound of the actors' voices. An account of them is to be found in Vitruvius. (3) The resonance box of a lyre.

Echelle (*Fr.*) A scale; as, *échelle chromatique*, chromatic scale; *échelle diatonique*, diatonic scale.

Echo. A sound produced by reverberation; an imitation of a sound so produced. (1) In old organ music the use of this term signified that a passage so marked was to be played upon the echo-organ, a set of pipes enclosed in a box, by which a soft and distant effect was produced, incapable however of so great expression as that obtained by the use of the *swell*, which is an improvement upon the echo-organ. (2) Echo-stop on a harpsichord was a contrivance for obtaining a soft and distant effect.

Eclisses (*Fr.*) The sides of a lute, guitar or violin.

Eclysis. [Ecbole.]

Ecole (*Fr.*) A school or style of music.

Ecossaise (*Fr.*) In the Scotch style.

E dur (*Ger.*) The key of E major, the key having four sharps in its signature.

---

* Gegenbaur, p. 773.

† Owen, Anat. of Vertebrates, vol. iii. p. 189.

Digitized by Google

Effect. *Effet* (*Fr.*) *Effetto* (*It.*) The mental impression produced by the performance of music, arising from the genius of the composer in the novel invention of pleasing or striking melodies, or telling harmonies, and the happy fitness of choice of certain passages, vocal or instrumental, in certain understood situations; or the clever interpretation of those passages by the performers.

Eguaglianza (*It.*) Equality, evenness.

Eguale (*It.*) Equal, as *voci eguali*, equal voices.

Egualmente (*It.*) Equally, evenly

Eighth. The interval of an octave.

Einfach (*Ger.*) Simple; as, *einfache Intervalle*, simple intervals; *einfacher Contrapunkt*, simple counterpoint.

Einfalt (*Ger.*) Simplicity; as, *mit Einfalt und Würde*, with simplicity and dignity.

Eingang (*Ger.*) Introduction, as, *Eingangs-Schlüssel*, introductory key.

Eingestrichen (*Ger.*) Having one stroke, as *c'*, *d'*, &c. [Pitch.]

Einheit (*Ger.*) Unity.

Einleitungssatz (*Ger.*) An opening phrase, or introduction; an overture.

Einschlafen (*Ger.*) To slacken pace and diminish the power.

Einschlagend (*Ger.*) *Lit.* striking inwards, as is the case with a *percussion* reed; whereas *aufschlagend* is used with reference to a *free* reed. [Reed.]

Einschnitt (*Ger.*) An incomplete musical sentence or motive.

Eis (*Ger.*) E sharp.

Eisenvioline (*Ger.*) *Lit.*, iron fiddle. A nail violin, an instrument the sounds of which are produced from pointed pieces of iron. *Ger.*, *Nagelgeige.*

Eisteddfod (*Welsh*). "A sitting of learned men" for the election of Chief Bard, creation of Pencerdds, and conferring degrees. They are of great antiquity; as early as the seventh century King Cadwaladr is said to have presided over one. The great Eisteddfod of Carmarthen was held in 1450. Their present object is the encouragement of native poetry, music, and patriotism.

Eklysis or **Eclysis** (*Gk.*) [See Ecbole.]

Ela. The highest note in the Guidonian hexachords. 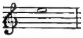 But it was more properly designated by *ee la*, double letters being used for the five upper notes.

Electric organ. An organ in which electricity is used as a portion of the mechanism, having generally for its object the placing of the *console* at a distance from the pipes.

Elégant (*Fr.*) *Elegantemente* (*It.*) *Eleganza, con* (*It.*) Elegantly, with elegance of style.

Elegiac In the style of an elegy; of a mournful character.

Elegy. ἐλεγεῖον (*Gk.*) (1) A distich consisting of an hexameter and pentameter. (2) A poem in elegiacs. (3) A composition of a mournful and commemorative character.

Elevatio (*Lat.*) (1) Arsis, *q.v.* (2) A motet sung at the elevation of the Host. (3) The raising of a mode beyond its *ambitus.*

Elevato (*It.*) Raised, exalted.

Elevazione (*It.*) A composition founded upon a special theme, as *Elevazione sopra il Pange lingua.*

Elève (*Fr.*) A pupil

Eleventh. The interval of an octave and a fourth. A compound fourth.

Embouchure (*Fr.*) The mouth-piece of a wind instrument.

Emmeleia, ἐμμέλεια (*Gk.*) (1) Consonance, concord in musical sounds. (2) A tragic dance accompanied by music. (3) The music of the tragic dance.

E moll (*Ger.*) The key of E minor. The relative minor of G major.

Empâter les sons (*Fr.*) To sing legato, or with a *portamento.*

Empfindung (*Ger.*) Emotion, passion, feeling.

Emphasis. Accent. [Arsis.] [Accent.]

Emporté (*Fr.*) Passionate, hurried.

Empressé (*Fr.*) Eager, hurried.

Enarmonico (*It.*) Enharmonic.

En badinant (*Fr.*) Scherzando. [Scherzo.]

Encore (*Fr.*) Again, more. A word used in England when a repetition of a piece is desired. It is used both as a noun and as a verb in common writing; as, *an encore. to encore.*

Encænia. Dedication festivals; in old English, chyrche-holy; Anglo-Sax., cyrichalgung, church hallowing.

Energia, con; energicamente; energico (*It.*) With energy, forcibly.

Enfant de chœur (*Fr.*) A chorister-boy.

Enfatico (*It.*) With emphasis, earnestly.

Enfler (*Fr.*) To swell, to increase in sound.

Enfasi, con (*It.*) With emphasis.

Enge (*Ger.*) Narrow, close, straight. A term used in reference to the small *scale* of organ pipes, or to the closeness of subject and answer in a *stretto.*

Engelstimme (*Ger.*) [Vox Angelica.]

Engraving of music. [Printing of music.]

Enharmonic. (1) One of the three genera of Greek music, the other two being the Diatonic and Chromatic. (2) Having intervals less than a semitone, *e.g.*, an enharmonic organ or harmonium is an instrument having more than twelve divisions in the octave, and capable, therefore, of producing two distinct sounds where, on the ordinary instrument, one only exists, as, for instance, G♯ and A♭, &c. An enharmonic *scale* is one containing intervals less than a semitone. (3) An enharmonic modulation is a change as to notation, but not as to sound, *e.g.* :

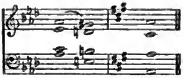

It is important to notice that an enharmonic modulation is not so termed in strict propriety, because, it is only feasible on an ordinary keyed-instrument by actually ignoring the existence of intervals smaller than a semitone.

Enoplius. ἐνόπλιος (*Gk.*) Warlike music. Music of the war-dance.

Ensemble (*Fr.*) Together. The whole. (1) The general effect of a musical performance. (2) The union of the whole company of performers in a concerted piece.

Entr'acte (*Fr.*) Music played between the acts or divisions of an opera, drama, or other stage performance.

Entrata (*It.*) *Entrée* (*Fr.*) Entry, introduction, or prelude. *Scena d'entrata*, the first *scena* allotted to a vocalist in an opera. (*Fr.*) *scène d'entrée.*

Entrechats (*Fr.*) The peculiar bounds with which a dancer leaps across the stage on entering.

Entremese (*Span.*) A short musical interlude, in one or two scenes, played by a few actors, rarely more than four. Entremeses were mostly of a burlesque character, and when performed between the preludes and the plays, Autos, or Loas, made an interlude of a nature peculiarly acceptable to the Spanish mind. The subjects were chosen from possible events of a droll character in common life, and were mostly written in verse. They cannot be traced to a higher antiquity than the 17th century, and are still popular in remote parts of Spain. When more than ordinary prominence is given to the music, the name Saynetes (the Spanish for dainties) is given to them.

Entremets (*Fr.*) Short dramatic or allegorical entertainments. A remote antiquity is claimed for this species of diversion, which some writers declare to be the origin of the opera and drama. The date of their invention has been fixed at an epoch during the reign of Saint Louis (1226-1270).

The king desired to re-awaken the enthusiasm of his nobles and warriors that they might join him in the endeavour to wrest the Holy Land from the hands of the infidels. He sought the aid of the Duke of Burgundy, who, flattered by the preference shown to him in being selected as the king's agent in the matter, sought every means to carry his wishes into effect. He gave a series of banquets and entertainments to the nobility, who at that time were noted for their luxury, not to say licentiousness; and in the course of these feasts certain allegorical poems, commemorating the deeds of the old warriors and kings, their ancestors, were recited or sung. Appeals were made to the chivalric spirit still supposed to exist in the breast of the scions of a warlike stock, and they were implored to unite in aiding the defenders of their religion in a crusade against the unbelieving occupants of the sacred cities. The stratagem succeeded, the nobles and princes joined with the Duke of Burgundy and bound themselves by oath to follow and support him.

The Entremets thus originated were continued on great occasions, and ultimately became diverted from their primary intention. The performers, *mimes*, *farceurs*, *baladins*, *ménétriers*, &c., as they are variously called, followed the fortunes of their lords, and in course of time invented new *entremets*, no longer confining themselves either to patriotic or religious subjects. These entremets suggested more extended performances, and what can now be gathered of their character has led many, not unreasonably, to assume, that in them was the germ of the modern opera.

In later times when acting was better understood, a further change was made in the style of the subjects selected, and the entremets were almost always of a humorous character, though heroic subjects were sometimes chosen. In 1237, upon the occasion of a marriage, Alberic, in his chronicle, speaks of the entremets then and there performed, and also adds that "Illi qui dicuntur ministrelli in spectaculo vanitatis multa ibi fecerunt, sicut ille qui in equo super cordam in aere equitaret, et sicut illi qui duos boves de scarlate vestitos equitabant cornitantes ad singula fercula quæ apponebantur regi in mensâ." In 1378, at a feast given by Charles V. to his uncle in the castle of St. Germain, two entremets representing the conquest of Jerusalem,, by Godfrey de Bouillon, were performed. And at the marriage of Charles IX. in 1572, the entremets were on the subject of the destruction of Troy. Jean Antoine de Bäif in 1573 published "Mimes, entremets, enseignmens et Proverbes," which were simply epigrams, and the change of opinion with regard to the use of the word in De Bäif's mind, shows that even then a gradual alteration in the meaning and force of the entremets was taking place. The word is now

Digitized by Google

employed to signify any small entertainment between two greater ones. *cf. Entremese.*

Entusiasmo. Entusiastico (*It.*) With enthusiasm.

Entwurf (*Ger.*) A sketch.

Eolian Harp. A musical instrument made of a long narrow box of thin even-grained deal, about five or six inches deep, having a circle of small holes drilled in the centre of the upper side. On this side the strings, six or more, are stretched in parallel lines over bridges fixed at each end, the tension being preserved by means of screwpins. The strings must be tuned in unison, and the box placed in a free current of air. A delicate combination of sounds is then produced, somewhat resembling the effect of a full orchestra, without instruments of percussion, when heard at a distance, the sound increasing or decreasing in power with the force of the wind. The usual method of using the instrument is by placing it on the ledge of a half-opened window; but the tone is best produced when the box is made of the exact length of the window opening, and the lower sash of the window closed as far as the box will allow. The Eolian harp is the invention of an Englishman of the name of Pope, and was improved by Kircher, a German (1670). The harmonics heard are due to the overtones of the strings.

> "Behoves no more,
> But sidelong, to the gently waving wind
> To lay the well-tuned instrument reclined;
> From which, with airy flying fingers light,
> Beyond each mortal touch the most refined,
> The God of Winds drew sounds of deep delight."
> THOMSON'S *Castle of Indolence.*

Eolian mode. The fifth of the authentic Gregorian modes. It consists of the natural notes La, Si, Do, Re, Mi, Fa, Sol.

Epicedion. ἐπικήδειον (*Gk.*) A dirge, elegy.

Epigoneion (*Gk.*), ἐπιγόνειον. A musical instrument, named after its inventor, Epigonus, the date of whose existence is matter of doubt, and the character of the instrument is also somewhat uncertain. It is described as having forty strings, but the method of tuning is absolutely unknown.

Epilenia, ἐπιλήνια (*Gk.*) Vintage songs.

Epinette (*Fr.*) [Spinet.]

Epinicium. ἐπινίκιον (*Gk.*) A song of victory.

Episode. A term in fugue writing, applied to those phrases which are supplemental to the main subjects or their answers. [Fugue.]

Epitasis (*Gk.*) ἐπίτασις. (1) The raising of the voice from a low to a higher pitch. (2) The tightening of the strings of an instrument, as opposed to ἄνεσις.

Epithalamium, ἐπιθαλάμιον (*Gk.*) A nuptial song.

Epode, ἐπῳδός (*Gk.*) (1) An after song, the strain of a lyric song after the strophe and antistrophe. (2) A burden or refrain.

E poi (*It.*) And then, after; as, *e poi la coda*, then go to the coda.

Equabilmente (*It.*) Equally, similarly.

Equale. A term recently brought into prominence by the revival of Beethoven's Equali for four trombones. It probably was originally applied to a piece written for instruments of the same family, though differing in pitch, much as we use the expression, "Equal voices" (*q.v.*). The term may perhaps be hereafter limited to short pieces written for instruments of the same timbre, suitable for use at commemorative services.

Equal voices. Voices of men only (A.T.B.), or of women or boys only, without regard to the difference of their compass or register. When the two are combined the piece is said to be for mixed voices. Strictly speaking, equal voices should signify voices of identical compass.

Equivoca (*Lat.*) Equivocal, or doubtful, *nota equivoca* was, in mediæval music, a note whose value varied according to the length of the notes on either side.

Equivocal or **doubtful chords.** A name given to combinations of sounds which are common to two or more distinct keys, and which, when heard, make the listener doubtful as to the particular key-tonality into which they are about to be resolved.

The simplest form of chords of this class is to be found in the so-called diminished triad, *e.g.*:

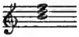

The above chord may be resolved into the keys of C major, C minor, A minor, B minor, or B major (the last two by means of an enharmonic change), thus:

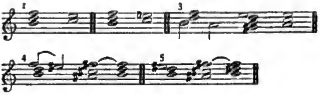

The inversions of this chord give, as might be expected, greater scope for varied progressions than can be obtained from its original position.

The next important doubtful chord is the diminished seventh, but in this case it will be noticed that the numerous resolutions are the result of its possible enharmonic change, whereas, in examples 1, 2, and 3 above, no alteration has been made in the notation of the chord:

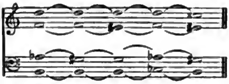

If this chord be struck and held down while the eye traces the various changes of notation through which it is capable of passing, it will be found that an impression of an entirely new key is given at each successive change.

Composers have not been slow to avail themselves of the sudden flights into remote keys, which such combinations suggest, *every key* being easily and naturally reached by a judicious treatment of this chord in the position given above or in its other three positions. Another class of chords are used as doubtful chords, though less frequently than those just described, namely, the chords of the extreme sixth (called also sharp or augmented sixth), *e.g.*:

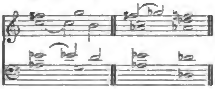

The use of doubtful chords is only to be traced in modern authors; old writers used them rarely, and then only to produce some startling effect, justified by the character of the words in vocal music, or by the professed drift of a piece of orchestral sound-painting.

Erhöhung (*Ger.*) Elevating, enhancing, raising; as, *Erhöhungszeichen*, the sign of chromatic elevation, a sharp or natural.

Erniedrigung (*Ger.*) Lowering, depressing; as, *Erniedrigungszeichen*, the sign for chromatic depression, a flat or natural.

Erst (*Ger.*) First; as, *erster Satz*, first part.

Ersterben (*Ger.*) To die away, *morendo*.

Erweitert (*Ger.*) Extended, augmented, amplified.

Es (*Ger.*) E flat.

Esatto (*It.*) Strict, exact; as, *esatto intonazione*, just intonation.

Es dur (*Ger.*) The key of E flat major.

Esecuzione (*It.*) Execution.

Eses (*Ger.*) E double flat.

Es moll (*Ger.*) The key of E flat minor.

Espace (*Fr.*) A space of the stave.

Espagnuolo, a (*It.*) In the Spanish style.

Espirando (*It.*) Dying away; gasping.

Espressione, con (*It.*) With expression.

Espressivo (*It.*) Expressive.

Essential harmony. Harmony independent of grace, auxiliary, passing, syncopated, anticipating, or pedal notes.

Essential notes. Notes belonging to a key-chord. The essential notes of the chord of F major are F, A, C. [Chordæ essentiales.]

Estinguendo. Estinto (*It.*) Dying away, gradually reducing both power and pace.

Estravaganza (*It.*) A work fanciful and far-fetched in composition or execution.

Estremamente (*It.*) Extremely.

Estro poetico (*It.*) Poetic rage, or fervency.

Etendue (*Fr.*) Extended.

Etouffé (*Fr.*) *Lit.* stifled. *Damped*, by means of pedal, mute, or palm of the hand. [Damp.]

Etouffoirs (*Fr.*) Dampers.

Etude (*Fr.*) A study, exercise, or lesson.

Et vitam. One of the movements of the Mass. A part of the Credo. [Mass.]

Etwas (*Ger.*) Somewhat; as, *etwas langsam*, rather slow, &c.

Euphonia (*Lat.*) *Eufonia* (*It.*) (1) Sweet sound, *suavitas vocis*. (2) A consonant combination of sounds.

Euphonium. A brass bass instrument, properly belonging to a military band, but sometimes introduced into the orchestra as a substitute for the third or bass trombone, to the tone of which the sound of the Euphonium has not the slightest affinity. [Metal wind instruments.]

Euphony. Sweet sound. An agreeable combination of sounds.

Evacuant (*Ger.*) An exhaust-valve, in an organ or other wind instrument.

Evacuatio (*Lat.*) (*Lit.*, an emptying.) In the black-headed notation, which was in general use before 1400 A.D., red notes, and afterwards open-headed notes (notæ evacuatæ) were occasionally introduced to enable the composer to write 3 against 2, three such open-headed notes being equal in value to two black-headed notes of the same denomination; when the open-headed notation came into vogue (soon after 1400) these functions were reversed, three black notes being equal to two white.

Eveillé (*Fr.*) Sprightly, quick, lively.

Evolutio (*Lat.*) The working out or development of a subject.

Evovae. The vowels of the words "seculorum amen" at the end of the Gloria Patri. Hence used as a name of the *endings* of Gregorian tones, *e.g.*, the following are the evovae of the fourth tone:

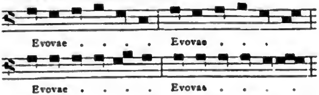

Exercise. (1) Preparatory practice in order to obtain skill. (2) A composition intended for the improvement of the singer or player. (3) A composition or thesis, required of candidates for degrees in music in the universities.

Expression. The power or act of rendering music so as to make it the vehicle of deep and pure emotion; the *spirit* of music, as opposed to the mere mechanical production of sound. In rendering works of a high class, a true expression involves the merging of the artist's personality in an enthusiastic effort to carry out to the highest extent, the fullest meaning of the composer. Hence the difficulty of giving a reading of classical works which shall satisfy those critics who have formed their own ideal of the author's conceptions. Compositions of a low order, often achieve great popularity owing to their clever treatment by practised artists, who know how to create an artificial interest in such a work, which its internal merit does not warrant.

Marks of expression are of comparatively modern use. It is said that Locke (c. 1677) was the first Englishman who used signs for *crescendo* and *diminuendo*, but there can be no doubt that an expressive treatment of music has at all times been known and appreciated, although the signs or directions for this expression were unwritten. This remark applies equally to solo and concerted music; in the latter, whether vocal or instrumental, sacred, or secular, the proper treatment of certain passages would, if not traditionally received, be suggested by the leading musicians among the performers. If this be true, directors of modern choirs or orchestras are to some extent justified in adding marks of expression to unmarked works to be performed, it being a fact that, where none exist, singers and players now-a-days sink into an uninteresting dead-level of production.

The absence of such marks gives the greatest latitude to the artist who renders music, and allows him to stamp his reading more with his own individuality than where the expression required is definitely indicated. Thus, some of the old simple songs or tunes depend entirely upon the performer for their true expression; whereas modern music is so full of directions that any intelligent reader may see the drift of the author's meaning. But the fictitious expression obtained only by a strict attention to orders, is vastly different from that true expression which is the offspring of sympathetic genius, which will ever remain the real test of the taste, culture, and ability of an artist.

Expression-stop. In a harmonium the expression stop when drawn, closes the waste-valve of the bellows. Any alteration of the pressure of the feet on the wind-pedals, causes therefore a corresponding alteration of the power of the tone produced. Hence, by a proper sympathy between the pressure of the foot, and the force of sound required, the most delicate contrasts of light and shade can be obtained.

Extempore. Musical improvisation. The art, or rather *gift*, of creating melody and harmony without premeditation. The ancient Greeks were said to have possessed the talent of poetical improvisation; and the gift is found in many races in which the imagination is free and vivid, such as the Arabs, and some tribes of Negroes. Among the former, the extempore effusions relating to small customs and superstitions such as those contained in the Sonnah, among the latter, hymns, religious poems and songs (generally with some rude kind of vocal and instrumental accompaniment), form the themes improvised upon. Some of their songs, originally extempore, afterwards remembered and made traditional, are not without a savage kind of beauty, but like most productions of the class to which they belong, do not appear to produce the effect upon paper it is known they do when sung by an excited body of singers.

In Europe the Italians, above all other nations, cultivate the gift of reciting extempore verses, which are not always mere simple effusions of a few stanzas, but are sometimes marked by extraordinary talent, and are extended to the length of an epic poem. Even tragedies and comedies have been made on the spur of the moment. It is said that the people of Tuscany and of the Venetian territories possess the gift in the strongest degree, and that females as well as males have exhibited powers of this sort.

Petrarch is said to have introduced the custom of singing extempore verses to the lute, and many names of eminent improvvisatori are preserved, one of the greatest being Metastasio, who, however, gave up the art at an early period of his long career. Among musicians, the gift of performing extempore upon an instrument is more remarkable than the power of making verses; for it not only requires a special aptitude, but also demands an extensive knowledge of art at ready command. John Stanley, the blind organist, contemporary with Handel, was an extraordinary impromptu player, capable of clothing any suddenly suggested theme with every resource of art. The stories told of J. S. Bach, in this respect, would be incredible, if his works did not show how great and free was his command over the technicalities of composition. The list of eminent musicians who have excelled as extempore performers might be swelled to a large extent, if it were necessary; but it will be sufficient for the present purpose to name only one or two, who may be said to represent the historical sequence of the existence of the faculty of performing at a moment's notice a subject arranged according to any form that might for the time be selected. Mozart possessed the power in no mean degree, for there are records of the fact of his having performed a concerto with only blank sheets of paper before him, he having been either too idle or too busy to write out more than the accompanying parts. Clementi, Moscheles, and Cramer, were famed for this gift, in their day, and the elder Samuel Wesley also was noted for his skill. Secular or trifling melodies have frequently been made themes for improvisation of an amusing or grotesque character.

One of the greatest pleasures Mendelssohn gave to his friends was that of listening to his extempore playing, and many living musicians of eminence have also displayed their powers as well in public as in private.

It is a singular fact that many performers highly gifted as extempore players, have failed, where it might appear at first sight they were eminently qualified to shine; a good extempore player often proving an indifferent, if not wholly bad accompanist or composer, and the reverse. The union of the different qualities in one and the same individual is rare.

Extemporize. To play *extempore*.

Extended compass. A range beyond the ordinary limit of a voice or instrument. A pianoforte was formerly said to be of extended compass, when a few notes more than the old five octaves were employed; now, a pianoforte is not considered of extended compass if it has less than seven octaves.

Extended harmony. [Dispersed harmony.]

Extraneous modulation. A modulation to an extreme or unrelated key. [Modulation.]

Extreme. (1) Outside; as, *extreme parts*, the highest and lowest parts in part-music. (2) Expanded to its furthest limit; as, *extreme intervals*, intervals greater than major or normal; *e.g.*, C to G♯ an extreme fifth. Such intervals are called also augmented, superfluous, or sharp. (3) Not closely related; a modulation into an extreme key is one into any key, other than, its own relative minor, its dominant, and sub-dominant, and their relative minors. (4) An old term for any key having more than three sharps or flats.

Extreme sixth, chord of the. A chord of modern growth, so called because the interval of an extreme or augmented sixth is contained in it, either directly or by inversion. It exists in three principal forms:

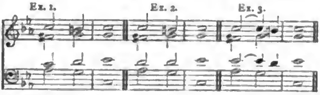

It will be noticed that this chord occurs on the sixth degree of the minor scale, but like many other chords originally formed of notes in the minor scale, it is as frequently resolved into the major key of the tonic, as into the minor; *e.g.*:

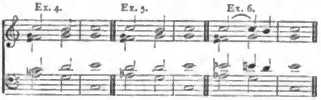

The dominant chord G, B♮, D, which is common to both C major and C minor, forms the connection between the resolutions given in Ex. 1, 2, 3, and those in Ex. 4, 5, and 6.

Various explanations of the origin of this chord have been suggested. Some consider it merely a chord of $^{6}_{3}$ on the sixth of the minor scale, or the first inversion of the subdominant common chord with the sixth chromatically raised (Ex. 1). Others look upon it as a chord of $^{6}_{3}$ on the sixth of the major scale, or the first inversion of the subdominant common chord, with the bass-note flattened (Ex. 4). These two opinions obtain favour in proportion to the supposed major or minor tonality of the chord.

Some authors find a much more complicated solution, namely, that it contains the minor ninth of the dominant, combined with the major third, seventh, and other notes of the fifth above the dominant. Hence it is called a *double-root-chord*, and $\frac{D}{G}$ would be given as its derivative in all the above examples.

This chord, as constructed in Ex. 1 and 4, is sometimes known as the Italian Sixth; as constructed in Ex. 2 and 5 as the French Sixth; as constructed in Ex. 3 and 6 as the German Sixth.

The component notes of these are often converted and form different inversions or positions, *e.g.*:

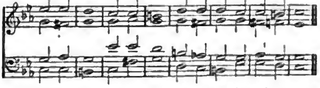

Chords of this kind are occasionally met with in the works of Bach and Handel, but are not of frequent occurrence at that date. Among modern authors, Spohr makes most use of them, and they form an important ingredient of his flowing chromatic progressions

Digitized by Google

F.

F. (1) The note called parhypate in the Greater Perfect system of the Greeks. The letter-name of Trite in the upper tetrachord.

(2) The first note of the Eolian mode, or church scale, commencing four notes above the hypo-Eolian. [Greek music.]

(3) The note called "Fa ut" in the hexachord system. [Notation.]

(4) The key-note of the major scale requiring one flat in the signature; and the key-note of the minor scale related to A flat.

Fa. The syllable used in solmisation for F. [Aretinian Syllables.]

Fa bémol (*Fr.*) F flat.

Fablier (*Provençal.*) [Trouveur.]

Faburden, Falso-bordone (*It.*) *Faux-bourdon* (*Fr.*) One of the early systems of harmonising a given portion of plain-song, or a *canto fermo*. As the word implies, *to faburden* signified originally to hold a *drone* (*bordonizare*). It was afterwards used as a term for a sort of harmony consisting of thirds and sixths added to a canto fermo. It will be remembered that the *organum* was similar in construction, being only *note against note*, but consisted of *fourths*, *fifths*, and *octaves*. [See Descant.] But when *counterpoint* had superseded both diaphony and descant, the term *faburden* still was retained, and applied to certain species of counterpoint, sometimes (but not always) note against note.

The following examples of **Falso-bordone,** by Bernabei (middle of 17th century), (from Proske's *Musica Divina*) are specially interesting as showing that composers, even at that time, ventured to alter church song when it suited their convenience. The introduction of the F♯ in the tenor at the close of Ex. 1, to secure a good cadence, disturbs the *mode* of the second tone; and in Ex. 2, the introduction of G♯ is equally fatal to the tonality of the fourth tone.

Ex. 1.

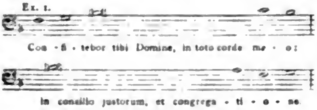
Con - fi - tebor tibi Domine, in toto corde me - o:
in consilio justorum, et congrega - ti - o - ne.

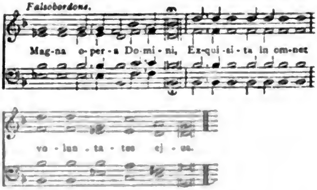
Falsobordone.
Mag-na o-per-a Do-mi-ni, Ex-qui-si-ta in om-nes
vo - lun - ta - tes ej - us.

Ex. 2.

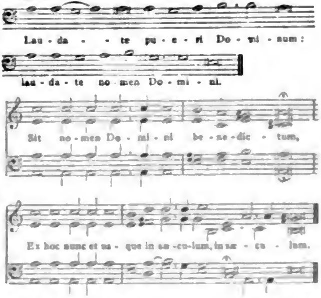
Lau - da - te pu - e - ri Do - mi - num:
lau - da - te no - men Do - mi - ni.
Sit no - men Do - mi - ni be - ne - dic - tum,
Ex hoc nunc et us - que in sæ - cu - lum, in sæ - cu - lum.

Faces d'un accord (*Fr.*) The positions or inversions of a chord.

Fach (*Ger.*) A rank of pipes, as in an organ.

Facile (*Fr.*) Easy.

Facilement (*Fr.*), **Facilmente** (*It.*) Easily, with ease.

Facilità (*It.*), **Facilité** (*Fr.*) Facility, readiness of execution.

Facilité (*Fr.*) Made easy. An easy arrangement of a difficult passage.

Facture (*Fr.*), **Fattura** (*It.*) (1) The construction of a piece of music. (2) The measurement, dimension, or scale of organ pipes.

Fa dièse (*Fr.*) F sharp.

Fagottista (*It.*) A bassoon player.

Fagotto (*It.*) [Bassoon.]

Fagottone (*It.*) A large bassoon [Double bassoon.]

Fall (*Old Eng.*) A cadence.

Fal las. Short songs with the syllables *fal la* at the end of each line or strain. Morley (c. 1580), who composed some, speaks of them as being a kind of *ballet*. The fal las of Hilton (c. 1600) are held in highest estimation for the freedom of their construction and the beauty of their melodies. Gastoldi is the reputed inventor of fal las.

Falsa musica (*Lat.*), called also *musica ficta*. False or feigned music was that in which notes were altered by the use of accidentals. "Falsa musica est quando de tono facimus semitonium et e converso" (Johannes de Garland). [Musica ficta.]

False cadence. [Cadence.]

False fifth. A fifth when not perfect.

False intonation. (1) The production of an unnatural or improper quality of tone. (2) Singing or playing out of tune.

False relation. The separation of a chromatic semitone between two parts.

False string. A badly woven string, which produces an uncertain and untrue tone.

Falsetto (*It.*) The artificial or supplementing tones of the voice, higher than the chest or natural voice. Falsetto is present in every voice with more or less power or quality. The similarity of the character of the natural and artificial voice in boys or females renders the two tones less distinct; but the chest voice and head voice in the man being of two qualities, the falsetto has a special character. The control of the falsetto requires great skill. The voce d'evirato is not falsetto, although high in pitch. [Larynx.]

Falso-bordone (*It.*) [Faburden.]

Fancies. (1) An old name for compositions in an impromptu style; a fantasy. (2) Short pieces of music without words.

> "And sing those tunes to the over-scutched huswives that he heard the carmen whistle, and sware—they were his *fancies*, or his good-nights."—(SHAKSPEARE.)

Fandango (*Sp.*) A lively Spanish dance in triple time, derived from the Moors. It is a mild form of the Chica, *q.v.* It is danced by two persons, male and female, and accompanied by the sounds of a guitar. The dancers have castanets, which they beat in time to the measure, though sometimes the male dancer beats a tambourine.

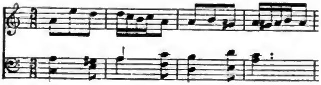

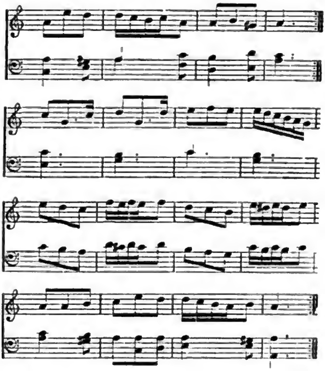

Fanfare (*Fr.*) A flourish of trumpets, a *call*.

Fantaisie (*Fr.*) [Fantasia.]

Fantasia (*It.*) *Fantasie* (*Ger.*) A composition in a style in which form is subservient to fancy. [Form.]

Fantasiren (*Ger.*) To play as fancy directs; to improvise.

Fantasticamente (*It.*) *Fantastico* (*It.*) *Fantastique* (*Fr.*) Fantastically, in a grotesque manner.

Farandola (*It.*) *Farandoule* (*Fr.*) A dance popular among the peasants of the South of France and the neighbouring part of Italy. It is performed by men and women taking hands, and forming a long line, and winding in and out with a waving motion. The manner of taking hands is peculiar. The men and women are placed alternately, each man's right hand is held by a woman's right hand, and his left by the left hand of another woman, so that along the line, when seen from the front of the row, there is a woman's face and a man's back, and the reverse. The dance is sometimes made the means of fanning popular excitement. A recent traveller, describing his experience of it, says:—"As the night wore on all the roughs in the town turned out, and began dancing the *farandole*—a kind of exciting dance peculiar to the south: men and women, hand in hand, form a long chain, and to a very quick step turn and twist along the various thoroughfares. This dance has the same effect on the fiery Southerners that the scalp dance has on the Red Indians, and makes them quite wild. It was after they had thus worked themselves up to a proper state of

Digitized by Google

excitement that the mob of Avignon massacred Marshal Brune in 1815."

The figures of the *Farandola* by the name of the "Spanish dance," were well known in English ball-rooms thirty years ago.

Farsa(*Lat.*) (*Lit.*, stuffed.) A term applied to interpolations between the sentences of a Kyrie, Gloria, &c., as specially exemplified in Tropes.

Fascia (*It.*) (1) A bind or tie. (2) The sides of a fiddle.

Fastoso, fastosamente (*It.*) Proudly, haughtily.

Fattura (*It.*) [Facture.]

Fausse corde (*Fr.*) [False string.]

Fausset (*Fr.*) [Falsetto.]

F clef. [Clef.]

F dur. (*Ger.*) The key of F major.

Federclavier (*Ger.*) Spinet.

Feier (*Ger.*) A festival. ***Feierlich***, in a festival style, grandly.

Feld (*Ger.*) (1) The disposition of pipes in an organ. (2) ***Feldflöte***, a rustic flute or pipe. (3) ***Feldmusik***, military music. (4) ***Feldton***, the key of E flat, in which military instruments are often set.

Ferial. Non-festal; as, ***ferial use***, music for use on ordinary days.

Fermamente, fermato (*It.*) Firmly, with decision.

Fermata (*It.*) A pause (from ***fermare***, to stay, or stop).

Fermo (*It.*) Firm, fast; as, *canto fermo*, the subject or part held firmly, while descant or counterpoint moved about it.

Feroce, con ferocità (*It.*) Wildly, fiercely.

Fertig (*Ger.*) Quick, dexterous.

Fervente, ferventemente (*It.*) Fervently, vehemently.

Fes (*Ger.*) The note F flat.

Fest (*Ger.*) A festival; as. ***Festgesang***, a festival cantata.

Fest (*Ger.*) Firm; as, *fester Gesang*, canto fermo. [Fermo.]

Festivamente (*It.*) Solemnly, pleasantly.

Festività, con (*It.*) With joyfulness.

Festivo (*It.*) Festive, solemn.

Festoso (*It.*) Joyous, gay.

FF. ff., *abb. of fortissimo.* Very loud.

Fiacco (*It.*) Weak, weary, faint.

F holes. The openings in the upper plate of a violin or other instrument having a resonance-body, so called from their common shape *f*.

Fiasco (*It.*) *lit.* a flask or bottle. A term applied to a failure in singing, playing, or representation. The *fistula pastoricia* was blown by the Romans to signify their dissatisfaction, and it is possible that the present use of the term arose from the similarity between the shape of a flageolet (*flaschinet*) and a flask. The Italians now blow sometimes into the pipe of a key, whence the expression *colla chiave*.

Fiato (*It.*) (1) Wind; as, ***stromenti di fiato***, wind instruments. (2) Breath, in singing; as ***in un fiato***, in one breath, a long note or passage performed with one respiration.

Ficta musica (*Lat.*) [Musica ficta.]

Fiddle. [Violin.]

Fidicen (*Lat.*) (From *fides* and *cano.*) A lute or harp player.

Fiedel (*Ger.*) Fiddle.

Fier (*Fr.*) *Fiero* (*It.*) Proud, fierce.

Fieramente, fiero (*It.*) Proudly, fiercely, boldly.

Fife. *Fifre* (*Fr.*), *Querpfeife* (*Ger.*), *Piffera* (*It.*) An ancient musical instrument, the name being cognate with pipe. The compass is two octaves from D:

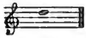

A combination of fifes and drums is the only music officially allowed in the British army and navy. Although of ancient use in England for military purposes, it was discontinued in the reign of James I., and was not restored until the siege of Maestricht in 1747. The fife in the orchestra is called flauto piccolo.

Fife. An organ stop. A piccolo, generally of two feet in length.

Fifre. [Fife.]

Fifteenth. The interval of a double octave. Bis-diapason.

Fifteenth. An organ stop of two feet in length on the manuals and four feet on the pedals, consisting of open metal pipes.

Fifth. A diatonic interval of five notes. Its ratio is 2 : 3, the *diapente* of the ancients.

Figura (*Lat.*) A note. ***Figura simplex***, a note standing by itself. ***Figura ligata***, a ligature, or a series of notes with contiguous sides.

Figure. A form of melody or accompaniment maintained throughout the phrase in which it is suggested. In a melody, figure is called sequence. In harmony a figure relates to the rhythmical observance of a certain form in all the accompanying chords to the melody. (2) A musical phrase. (3) A florid melody.

Figurato (*It.*) *Figuré* (*Fr.*) Figured.

Figured Bass. A bass having the accompanying chords suggested by certain numbers above or below the notes. It is at present the most satisfactory system of musical shorthand. The whole of the notes are not always indicated by a corresponding number of figures, because one number generally implies two or more to complete the chord. When there is no figure, it is understood that the common chord of such a note is to be used as its harmony. The following table will show the manner in which figures are used:

The figure 2 implies a 4th and 6th.
,, ,, 3 ,, 5th perfect, or diminished, according to the position of the note in the key.

The figure 4 implies a 5th, or 5th and 8th.
,, ,, 5 ,, 3rd and 8th.
,, ,, 6 ,, 3rd.
,, ,, 7 ,, 5th and 3rd.
,, ,, 8 ,, 3rd and 5th.
,, ,, 9 ,, 3rd and 5th.

A stroke through a figure directs the raising of the interval by a natural or sharp, as the case may be.

An accidental standing alone implies a corresponding alteration of the 3rd of the chord. Horizontal lines direct the continuance of the harmony of the previous chord. If there are no figures under the previous chord, the line or lines direct the continuance of the common chord of the first note under which they were placed.

Filar la voce (*It.*), *Filer le son* (*Fr.*) To prolong a sound, swelling and diminishing the tone by degrees.

Fin (*Fr.*) The end.

Finale (*It.*) The last movement of a concerted piece, sonata, or symphony; the last piece of an act of an opera; the last piece in a programme.

Fine (*It.*) The end; used to show the end of a piece or movement, after a repeat, or partial repeat.

Finger-board. *Fingerbrett* (*Ger.*) (1) The flat or slightly rounded piece of wood attached to the neck of instruments of the violin and guitar class, on to which the strings are pressed when *stopped* by the fingers. (2) A manual or clavier.

Finger cymbals. [Cymbals.]

Fingering. *Applicatura* (*It.*), *Application* (*Fr.*), *Doigter* (*Fr.*), *Applicatur* (*Ger.*), *Fingersetzung* (*Ger.*) The art of placing and using the fingers properly in performing upon a musical instrument.

(1) When instruments were for the first time constructed so that the leverage of their keys was light and admitted of rapid and ready motion, musicians soon formed rules for the employment of the fingers in such a manner as to give the greatest facility to the player. These rules were properly improved and extended by each master who taught the use of a keyed instrument, and there is reason to suppose that they were kept more or less secret by each teacher, long before it was deemed expedient to set forth the methods in a general publication.

Comparing the earliest published methods of fingering with the musical compositions belonging to the same period, it is difficult to conceive that a clear, distinct, and rapid performance could ever be attained by those methods, for they are of a cramped, stiff and awkward character, while the compositions are of an opposite nature, considering the state of the art at the time.

Such a thing as acquiring a knowledge of a keyed instrument without a master was out of the question, and it is not at all unlikely that this was contemplated by the authors or compilers of the books of instruction.

One of the earliest printed books in which rules for fingering are laid down was Eusebius Ammerbach's "Orgel oder Instrument-Tablatur," Leipsic 1571, where in the fourth chapter a scale is fingered in the following manner:—

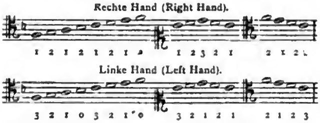
Rechte Hand (Right Hand).
Linke Hand (Left Hand).

0 stands for the thumb, 1 for the forefinger, 2 for the middle finger, and so on.

The thumb of the right hand was never used in scale passages, that of the left hand only occasionally, the little fingers were only used with the thumbs in spanning chords.

The rules for fingering in striking or playing chords are thus stated: 3rds in either hand were to be struck with the first and third fingers; 4ths, 5ths, and 6ths with the first and fourth; 7ths, octaves, 9ths and 10ths with the thumb and third finger, and only occasionally with the little finger and thumb.

Before saying anything further of other works on the subject, it may be as well to refer to a private MS. book of lessons preserved in the British Museum, and quoted by Stafford Smith in his Musica Antiqua, in which the master has marked the fingering of a passage in a very different manner to that laid down by the German author. It is true that there is a difference of nearly thirty years in the dates between Ammerbach's book and this, but as treatises published later preserve and teach the like awkward system of fingering, a curious confirmation of the conjecture that there was a considerable difference between the theory and the practice cannot fail to strike the thoughtful reader. The MS. referred to bears the date 1599, and the first lesson with the fingering (1 being thumb, 2 forefinger, &c.) is as follows:—

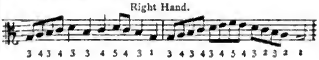
Right Hand.

Now in this fingering all the fingers are brought into play, and though it is somewhat clumsy according to modern views, it is less unhandy than that of Ammerbach, and as we have no ground for assuming that the English teachers had better principles to guide them than their German neighbours, and as it is known moreover, by tradition, that the most skilful players only imparted the secret of their power to favoured pupils, not only at that time but in later years, the assumption is not made without reasonable support.

A modern player would find it a very difficult task to perform the following extract from a "Gagliardo by Orlando Gibbons," printed in "Parthenia," 1611, if he confined himself to the use of three of his fingers in either hand:

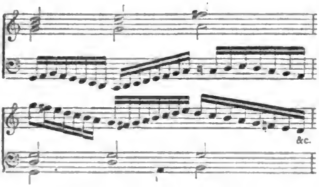

or to execute with ease any such a piece as that by Frescobaldi quoted in the article Canzona, belonging to about this period. And yet not only was the peculiarity of fingering as set forth by Ammerbach taught at that time, but was also continued to a later date.

There is a lapse of more than 100 years before the next important book made its appearance, "Das Musikalische Kleeblatt" of Daniel Speer, 1697, in which the improvements suggested are few, the chief of which was the more frequent use of the thumb of the left hand, as will be seen in the following scale fingered according to his directions:

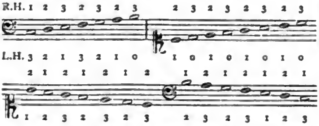
R.H. 1 2 3 2 3 2 3 2 3 2 3 2 3 2 3
L.H. 3 2 1 3 2 1 0 1 0 1 0 1 0 1 0
2 1 2 1 2 1 2 1 2 1 2 1 2 1
1 2 3 2 3 2 3 2 3 2 3 1 2 3

That there were differences of opinion on the subject with regard to fingering in the published books of about the same period, is shown by the subjoined quotation from "Kurtzer jedoch gründlicher Wegweiser, vermittelst welches man aus dem Grund die Kunst die Orgel recht zu Schlagen," Augsburg 1698, in which the scale of C is thus fingered; the thumb being indicated by the circle:

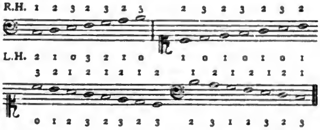
R.H. 1 2 3 2 3 2 3 2 3 2 3 2 3 2
L.H. 2 1 0 3 2 1 0 1 0 1 0 1 0 1
3 2 1 2 1 2 1 2 1 2 1 2 1 2 1
0 1 2 3 2 3 2 3 2 3 1 2 3 2 3

The well-known book, published by Walsh, under the title of "The Harpsichord Master, containing Plain and Easy Instructions for Learners on the Harpsichord or Spinnet," 1734, which, passing through many editions, may be fairly considered as correctly representing the method of that period, gives the following directions for fingering:—"Observe in y^{e} fingering of y^{r} right hand, y^{r} thumb is y^{e} 1st, so on to y^{e} 5th, and y^{r} left hand y^{r} little finger is y^{e} 1st, and so on, y^{e} fingers to ascend are y^{e} 3rd and 4th to descend y^{e} 3rd and 2nd."

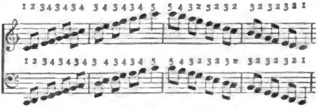

Mattheson ("Kleine Generalbasschule," 1735) and Maier ("Musiksaal," 1741) agree in their fingering, their method being as follows, little or no advance or alteration having been devised in the meanwhile in other publications.

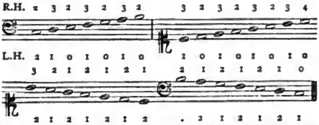
R.H. 2 3 2 3 2 3 2 3 2 3 2 3 2 3 4
L.H. 2 1 0 1 0 1 0 1 0 1 0 1 0 1 0
3 2 1 2 1 2 1 2 1 2 1 2 1 0
2 1 2 1 2 1 2 . 2 1 2 1 2 1

It is stated that J. S. Bach disregarded the ordinary principles taught in instruction books, and employed both thumb and little finger as frequently as the other fingers, whereby a greater power was gained, and the performer was able to move rapidly in extreme keys. Some writers claim the suggestion for the free use of all the fingers, for François Couperin, who, in his work, "L'art de toucher le Clavecin," 1716, describes the method of fingering practised and taught by Bach, and consequently they say that the German is indebted to the Frenchman for his ideas on the subject. But Couperin's fingering is somewhat different from that of Bach, the only similarity in their methods being the constant employment of the thumb.

Couperin, in the work alluded to above, calls the fingers of either hand 1, 2, 3, 4, 5, commencing with the thumb:

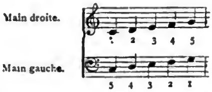

And in order to make his fingering acceptable, gives instances of the old style of playing contrasted with his own improvements. In some cases his suggestions are good, in others there is little if any help out of long-standing awkwardness, as the following fingered scale will show:

Progrès d'octaves.

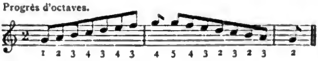

His next improved example is better, and more in accordance with modern methods:

"Manière ancienne de faire plusieurs tierces de suite."

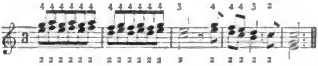

"Façon moderne pour couler les mêmes tierces."

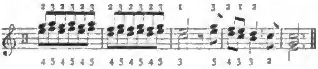

Bach never published his method, but it was made known through Forkel, who acquired it from Emanuel Bach. The peculiar methods of fingering shown in the quotations from the German books seem to have been confined to the country in which they were first given forth, the style adopted in England and France appearing to be borrowed from Italy. Many of the harpsichord instruction books printed in England in the early part of the last century profess to be based upon the "Italian method of fingering," which permitted the constant employment of all the fingers.

The invention of the pianoforte called a new set of instruction book writers into existence, but it required the genius of Cramer, Clementi, and Dussek to liberate learners from the trammels of the harpsichord fingering. Clementi's "Introduction to the Art of Playing on the Pianoforte" was written soon after his return to England, in 1784, and went through many editions. In 1810 he greatly improved the work, and added an appendix to it. Cramer made but few, if any, advances on the method of fingering suggested by Clementi in his "Introduction," but Dussek, in his "Book of Instructions," published about 1798, when he began business as a music publisher, laid down a code of rules which have served as the basis of teaching fingering from his time onward.

He recommends the pupil not to play his scales after the old method of fingering, by suffering the fingers to cross each other unnaturally, as in this example, right hand:

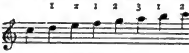

thereby implying that the old clumsy methods were still taught; but "to set it down as an invariable rule, and without any exception, that whether ascending or descending, the assistance, or rather, the displacing the *thumb* only is sufficient to effect the necessary change of position in the other fingers, the thumb being naturally formed to glide easily under them, without displacing the proper position of the hand."

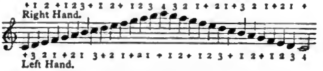

The free use of all the fingers is now common, and the old rules with regard to the sparing of the thumb so little observed, that there is a tendency to go the opposite extreme in employing it more often than is absolutely necessary. Space cannot here be allowed, or the whole of Dussek's rules for fingering might be reprinted with advantage as a set-off against modern extravagance, but those who have leisure and inclination might study them with profit, for no better code of rules for pianoforte playing have as yet been given to the world.

(2) In violin playing the fingers serve as stops shortening the length of the vibrating portion of the several strings as they are required, a like practice being followed in the performance upon such fingerboard instruments as the lute, mandoline, or guitar, as are plucked with the right hand. Upon the guitar the places of the several degrees are marked on the fingerboard by frets, which, when the string is pressed—always a little behind the one required—serve as a temporary nut.

(3) This is not the place to enter deeply into the anatomy of the hand, which can be found in any text-book on anatomy, and more especially in the elaborate Treatise of Sir Charles Bell. The bones of the hand are joined together by various sorts of joints, some of them fixed, some moveable. The joint by which the thumb joins the bone on which it plays is a remarkable one, admitting of a

most complex series of movements, and since the brain of man is indebted to the hand of man as much as the hand to the brain, each rendering higher development in the other possible, and thus helping to constitute a couple of mutually perfecting factors—by far the greater part of this potentiality on the side of the hand is owing to this articulation of the thumb. The thumb of man is pre-eminent in the animal series.

From a musical point of view another subject is of interest. If we straighten our fingers we shall see that no two of them are of the same length. Now let us bend the fingers on themselves, so as to bring the finger-tips into the middle of the palm, and we shall find that the finger-tips form an almost straight line with one another. The convenience of this to the musician is obvious, in playing keyed instruments such as the pianoforte or organ; or any wind instrument like the oboe, flute, or clarinet, we have a series of evenly arranged finger-tips ready to the notes. This is a point of what we may call accidental convenience, for the hand of man was not especially adapted for playing instruments. The question now arises how this curious effect is produced, and a few words will explain it. If any one will take the trouble to measure as accurately as he can the length of the first and of the third joint of each finger (the "proximal and distal phalanges") omitting the middle joint, he will find that by subtracting the length of the end (distal) joint which carries the nail from the length of the nearest (proximal) joint which joins the "knuckle," he will get a constant number, in other words, if the proximal phalanx is long, and would carry the finger-tip far beyond the middle of the palm, the distal phalanx is proportionately long and in a bent position of the finger carries the finger-tip proportionately far back towards the palm, or in yet other terms, the number of linear units obtained by subtracting the length of the distal phalanx of each finger from the length of the proximal phalanx of the same finger is constant in all the fingers. This is not *absolutely* correct, for the angles at which the phalanges are bent are not exactly right angles, so that the middle phalanx cannot be quite neglected, but the principle nevertheless we believe to be correct. The muscles which move the fingers and thumb are situated some in the forearm, some in the hand. If any one grasps his forearm, and moves the fingers, he will feel the muscles moving under his grasp. The thumb has the greatest number of muscles attached to it, and therefore takes the precedence in importance; the first or index-finger has a special muscle to extend or straighten it called the extensor indicis or "indicator." This muscle was said to be absent in all the apes, and there was an old saying, "no ape can point," which was quoted in favour of human superiority—this, however, is not true (Journal of Anat. and Phys., vol. vi. p. 185); we trust, notwithstanding, that the pre-eminence of man in the animal kingdom rests on somewhat more important foundations, and need not be ceded, however much apes may point. The little finger has a special muscle to extend or straighten it which the old anatomists called the "auricularis," because it is used to straighten the little finger when the extremity of that member is inserted into the ear. Thus did these ingenious men infuse an element of humour into the driest and apparently least comic of all subjects.

PALM OF HAND. Fig. 1.

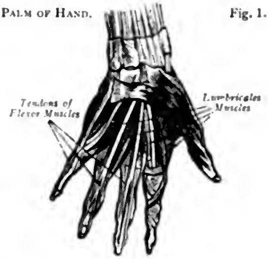

There are a set of muscles called "lumbricales" (from "lumbricus," a worm, because they are not unlike an earth-worm in size and shape), which are of great importance to musical performers, and in pianoforte players become very largely developed. They are attached to the flexor tendons (which bend the fingers) as they pass along the palm. They help to bend the fingers, but their individual action is somewhat complicated, viz., to bend the finger at the knuckle-joint, but to *straighten* the finger.

But there is one point in the anatomy of the hand which is of great interest to musical performers, especially to pianoforte and organ players. If any one will place the tips of his fingers on a table in a bent position, as if playing the pianoforte, and then try to raise them one at a time, he will find that he can raise his thumb easily, say four or five inches, the forefinger is also very moveable, and can be raised so that the tip is about three inches from the table, the middle finger about two-thirds of that height, and the little finger rather higher. Now let him try to raise the ring finger, keeping the middle and little fingers down, and he will find a peculiar difficulty in doing so. If he is a pianoforte or organ-player he will probably call to mind many exercises which he has toiled at, all of them easy enough but for that unlucky ring finger. In fact, all good writers of exercises have the education of the ring finger very prominently in view. What is the cause of this? It has been asserted popularly that this finger has a tendon too little, but this is just the opposite of the fact.

Back of Hand. Fig. 2

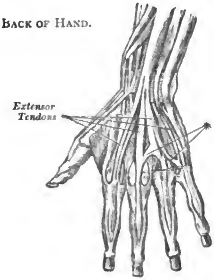

If we look at figure 2. we shall see the extensor tendons (the tendons which extend or straighten the fingers) running to the fingers, but, in addition, we shall see that the tendon which goes to the ring-finger gives off a small slip (*) on each side, one of which joins the extensor tendon of the little finger, the other that of the middle finger. Now we observed that none of these three fingers is as moveable as the forefinger. If we observe, also, the direction of these slips we shall see that they do not run straight across from the ring finger to the two on each side, but run at the same time a little towards the fingers. Now if we see what happens when we raise the middle finger alone, we shall observe that the slip will become tight at a certain point, but not until it has passed through a considerable space, the first action being to *relax* the tendinous slip; the same with the little finger. But if we raise the ring finger alone, the slips almost immediately become tight, and prevent its being raised, and the exercises before alluded to have for their purpose the stretching of the tendinous slips, especially in youth, when growth and change are active in the tissues. Now these slips must have some purpose, though it is not at all certain that we know it. It is probable that their purpose is to make the grasp firmer, the three fingers being so associated together that each as it were assists the other, and it is hard to detach one without the rest. They are present in monkeys (see paper before alluded to), and in them would be useful in climbing, in which exercise, strength of grasp is of course of great importance. The forefinger is not included in this association, as its value depends largely on its freedom of independent movement, especially on its capability of being opposed accurately to the thumb.

These slips have an historical interest. Robert Schumann, painfully aware of their presence, and acquainted with the cause of the difficulty connected with the ring finger, tied back that finger for a long time. Not being successful in his operation, he irrecoverably damaged his hand, and injured his pianoforte playing most seriously, and in consequence turned his attention to writing, to which fact we are largely indebted for the number of his masterly compositions. What was a loss to him and to his contemporaries has become a rich gain to posterity. A case is recorded in which an accidental wound to the back of the hand partially divided one of these tendinous slips, and the ring-finger gained a considerably increased degree of motion. The subject of the accident noticed the change on playing the pianoforte after the wound was healed. With regard to the feasibility of this as a definite operation, there is no doubt that if it could be done safely it would be worth every pianist's while to have these slips divided.

The experiment of Schumann, which led to the above suggestion, has resulted in a series of fourteen operations performed by Dr. Forbes, and described by him in an interesting article in the Boston Medical and Surgical Journal, Dec. 25, 1884. He reports that all were successful, and that no ill consequences followed. The remaining doubt is the permanency of the results. Who can say when this small operation may become the fashion?

Fingerleiter. (*Ger.*) [Chiroplast.]

Fingersatz, **Fingersetzung** (*Ger.*) [Fingering.]

Finite canon. [Canon.]

Finito (*It.*) Finished, ended.

Finto (*It.*) A feint, a term applied to deceptive cadences. [Cadence.]

Fiochetto (*It.*) Slightly hoarse.

Fioco (*It.*) Hoarse.

Fioriscente, **Fiorito** (*It.*) Ornamented, florid.

Fioriture (*It.*) Ornaments, cadenzas, florid passages in a melody or an accompaniment.

Fis (*Ger.*) F sharp.

Fisfis or **Fisis** (*Ger.*) F double sharp.

Fis dur (*Ger.*) F sharp major.

Fis moll (*Ger.*) F sharp minor.

Fistula (*Lat.*) A pipe. *Fistula dulcis*, a *flûte-à-bec. Fistula. cui semper decresci* *arundinis ordo*, pan-pipes. *Fistula eburneola*, the ivory pitch-pipe, from which an orator took the pitch for his voice. *Fistula pastoricia*, the shepherd's pipe, sometimes blown in the theatre as a sign of dissatisfaction. [Flute.]

Fithele (*old Eng.*) The ancient name of the fiddle, probably derived from *fidicula*, a small stringed instrument of the cithara class.

Flageolet. (1) A small pipe with a mouth-piece inserted in a bulb (hence the derivation of the name from the same root from which the word flagon comes), producing a shrill sound, similar, but much softer in quality than that produced from the flauto piccolo. It is an instrument of English invention, and was formerly employed in the orchestra. The obbligato in the song, "O, ruddier than the cherry," in Handel's "Acis and Galatea," is for a flageolet. (2) The tone produced from a violin by lightly pressing the bow near the bridge upon lightly touched strings, is called flageolet or flute tone.

Flageolettöne (*Ger.*) [Flageolet tones.]

Flageolet tones: The natural harmonics of stringed instruments, so called from their pure flute-like quality of tone.

Flaschinet (*Ger.*) The flageolet.

Flat. (1) The sign ♭, which directs the lowering of the note to which it is prefixed by one semitone. Its shape is derived from the ancient *b*. [B quadratum.] (2) Singing or playing is said to be *flat* when the sounds produced fail to reach the true pitch. (3) Minor; as, a flat 3rd, a flat 5th, &c.

Flatter la corde (*Fr.*) To play expressively upon a stringed instrument with a bow.

Flat tuning. One of the varieties of tuning on the lute; called also French tuning, or French flat tuning, because the French pitch was formerly lower than that used elsewhere. Hence the German term *Franzton* for a low pitch.

Flautando, flautato (*It.*) Like a flute; a direction to produce the flageolet tones on the violin, &c. [Flageolet tones.]

Flautino (*It.*) (1) An instrument of the accordion kind. (2) A little flute, piccolo. or flageolet. (3) [Flautando.]

Flauto (*It.*) [Flute.]

Flauto amabile (*It.*), *flûte d'amour* (*Fr.*) An organ stop, consisting of sweet-toned closed, or sometimes open, pipes. It is generally of 4 ft. pitch.

Flauto dolce (*It.*) [Flute.]

Flautone (*It.*) [Bass flute.]

Flauto piccolo (*It.*) [Piccolo flute.]

Flauto traverso (*It.*) The German flute held laterally, flutes having been formerly played with a mouth-piece, whence they were called flutes *à bec*. [Flute.]

Flebile, Flebilmente (*It.*, In a doleful, tearful manner.

Flessibilità (*It.*) [Flexibility.]

Flexibility. The power of free and rapid execution, in vocal or instrumental music.

Fling. A dance performed by Scottish Highlanders to a tune in common time. [Reel.]

F-Löcher (*Ger.*) [F holes.]

Florid counterpoint. The Fifth species, so called because notes of various lengths are used in it. It has its own special laws designed to enforce variety and interest. [Counterpoint.]

Florid music. Music in which the melody and accompanying parts are of an ornamental and embellished style.

Flöte (*Ger.*) [Flute.]

Flourish. (1) The execution of profuse but unmeaning ornamentation in music. (2) The old English name for a call, fanfare, or prelude for trumpets or other instruments together or alone. (3) The preparatory cadenza for "tuning the voice," in which singers formerly indulged just before commencing their song.

Flüchtig (*Ger.*) Light, rapid.

Flügel (*Ger.*) A grand pianoforte or harpsichord, so called because of the wing-like shape of the top.

Flügelhorn (*Ger.*, A bugle. A valve-horn.

Flute. (1) One of the most widely used of ancient musical instruments, and at this day one of the most important instruments in an orchestra. It has been remarked in speaking of the aulos, that the general idea of a "flute," probably included anciently, not only open tubes, but also instruments having a reed, such for instance, as the oboe. But the word has for many centuries been used only in the former sense.

Of tubes without reeds there are only two kinds—the flute played by a mouth-piece, and that played by placing the lips close against a hole on one side. The former kind was formerly called *flûte-à-bec;* the latter, *flûte traversière*, or *flauto traverso*, the cross flute. The *flageolet*, which still is in use, is a familiar example of a *flûte-à-bec*, but it is the smallest of its kind, for these instruments were at one time made sufficiently large to be called "tenor" and "bass" flutes; and complete four-part harmony could be obtained from a set [Bass flute]. The larger kinds only exist now as curiosities. The *flûte-à-bec* was used so commonly in England that it was called on the continent *flûte d'Angleterre.* They came to be called *beak*-flutes, because

Digitized by Google

of the similarity of the mouthpiece, through which the wind is directed against a sharp edge, to the beak of a bird. *Flûtes-à-bec* were single and double. Such double flutes were familiar both to Egyptians and Assyrians, and illustrations of them will be found on p. 40 (Aulos). The following illustration, from Boissard's Roman Antiquities, is interesting from its great likeness to the modern double-flageolet :

Fig. 1.

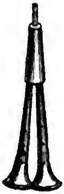

The Romans gave various names to their flutes : *calamus* from the material (reed) of which it was made ; *tibia*, because anciently flutes were formed out of a leg-bone, as shown in the following illustration :

Fig. 2.

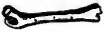

It is a remarkable fact that flutes of this barbarous construction are to this day used in many parts of Asia. The word *fistula* seems to have been applied both to flageolets and pan's-pipes. *Flûtes-à-bec* have at all times been a favourite object for ornamentation, and the next illustration shows a very beautiful example in carved ivory in the Kensington Museum :

Fig. 3.

The ancients possessed cross-flutes, and it is strange that their real value should have been found out, and their use made general, after so long a period of disuse, that on their re-introduction they were called German flutes as opposed to the old English beak-flute. They were used by the Egyptians, as the following illustration shows :

Fig. 4.

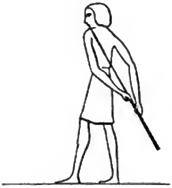

Fétis having obtained the exact measurements of an ancient Egyptian flute preserved in Florence, caused a flute to be made of the like dimensions and shape. The following figure shows it :

Fig. 5.

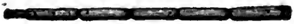

He found the lowest note it was capable of producing, to be the A below *middle C*. But if Egyptian artists are to be trusted, the flute in Fig. 4 must have been of even graver pitch.

Cross-flutes were known to the Greeks by the name plagiaulos (πλαγίαυλος), and to the Romans as *tibia obliqua*, both of these terms leave no doubt as to their nature. By the Romans the cross-flute was sometimes called also *tibia vasca*, the meaning of which is very doubtful.

It may be necessary to say, that although the *tibiæ* represented flutes of all kinds, yet if a real tibia or shinbone be made into a flute, it is held crossways, and the player blows into a hole in the side.

(2) The "German flute," so popular in England during the last century, has entirely superseded the old English *flûte-à-bec* in our orchestra, and is now known as *the flute*. Its construction has, from time to time, been improved, until it has now a compass of three octaves.

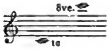

But as the lowest note is very soft, and three or four of the highest notes are exceedingly shrill, it is safer to consider it as possessing a chromatic scale of about two octaves and a half. Notwithstanding the vast improvements in the key work of flutes (of which, by the way, the ancients seem to have been entirely ignorant), there are shakes on certain notes which are almost impossible, and others which are of excessive difficulty ; a list of the shakes most difficult to execute will be found in

Prout's Instrumentation Primer, page 57, and other standard works on orchestration. The lovely effect of this instrument in an orchestra must be so well known to all as to render it unnecessary to quote special examples of its use.

(3) The piccolo-flute has the same extent of compass as the ordinary flute, but is one octave higher in pitch. When used in an orchestra with moderation and skill, it is capable of producing delightful effects; but, unfortunately, it is so commonly abused that it has got an undeserved bad character. The lower portion of its notes are bright and joyous, but in the upper part of its compass it is so shrill as to only justify its use when rare and special effects are required.

(4) As all open organ pipes of the *flue* class are made on the same principle as the *flûte à bec*, it will be easily understood that *flutes* are one of the most essential class of organ stops. They are of two kinds—open and stopped, and are equally common in metal and wood. The construction of the stopped flute, so far as the mouthpiece (foot) and *lips* are concerned, is identical also with that of the *flûte-à-bec*, only, of course, its first harmonic will be the twelfth, not the octave, of the primary sound. When organ builders describe some of their flute-stops as *flauto traverso*, or *flûte douce* (another name for the cross-flute), it must be understood that they have only imitated the quality of tone, not the construction of that instrument. By slight modifications of the shape of the different parts of a flute pipe, an almost endless variety of tone may be produced, and organ builders avail themselves of this fact to coin an endless variety of names. If the names so chosen carried with them a hint as to the special construction of each register, it would be unfair to complain of their multiplication; but, with a very few exceptions, this is not the case.

The following are some of the titles appended to flute-stops on English and foreign organs:

(1) Describing their material, as *wood flute*, *woud*, and *woude-fluit* (in Holland). Metal flute.

(2) Stating whether the pipes are open or closed: as, open flute, *flûte ouverte* (*Fr.*); stopped flute, *Gedacktflöte* (*Ger.*)

(3) Showing the pitch of the stop; as, *bass flute* (16 ft. and 8 ft.) *Flautone* (16 ft.) *Flauto grave* (16 ft.) *Flute principal* (8 ft.) *Flute major* (8 ft.) *Unison flute* (8 ft.) *Flute minor* (4 ft.) *Flûte octaviente* (4 ft.) *Quintflöte* (5⅓ ft.) *Quintaton* (sounding unison and twelfth). Piccolo flute (2 ft.) Flautino (2 ft.) *Kleinflöte* (2 ft.) *Terzflöte* (1⅗ ft.) *Flute discant*. *Flûte dessus* (treble flute).

(4) Describing the shape of the pipes, as *Doppelflöte* (with two mouths). Pyramid flute (having pipes larger at the top than at the mouth). *Flag fluit* (*Dutch*). *Flachflöte* (*Ger.*) (having flat lips). *Spitzflöte* (*Ger.*) and *Flûte pointue* (*Fr.*) (having pipes smaller at the top than at the mouth). *Rohrflöte* (*Ger.*) and *Flûte-à-cheminée* (*Fr.*) (having a chimney in the stopper).

(5) Intimating their quality of tone, as full flute. *Hohlflöte* (*Ger.*) and *Flûte creuse* (*Fr.*) (hollow toned). Clear flute. *Hellflöte* (*Ger.*) *Lieblichflöte* (*Ger.*) (lovely toned). *Zartflöte* (*Ger.*) (delicately voiced). *Flûte douce* (*Fr.*) Dulcet. *Flauto dolce* (*It.*) (sweet-toned). Oboe-flute. Clarinet-flute (slightly reedy in tone). *Sifflöte* (1 ft.) (whistle-flute).

(6) After their supposed nationality, as German flute, *flauto tedesca* or *allemande*. *Flûte-à-bec*, or English flute. *Suabe flute*. *Schweizerflöte* (Swiss flute, the German name for what was called in England the *German* flute). *Flauto Francese*. *Flute Ravena*. *Czakan flute*.

(7) Implying that the quality of tone is similar to the modern flute, more powerful than the *flûte-à-bec*, as orchestral flute, *flauto traverso* (*It.*), *flûte traversière* (*Fr.*), and *Traversflöte*, *Querflöte* (*Ger.*) (cross-flute), concert-flute.

(8) Names which are merely *fancy* titles, as *flûte d'amour*, *jubal flute*, *portunal flute*, old flute, recorder (*flûte-à-bec*), *Wald-* and *Bauer-Flöte* (*Ger.*) (pastoral pipe), echo flute (soft-toned), *flute tacet*, *cordedain*, &c.

It would be an advantage alike to organ builders and organists if some definite system of nomenclature of flute stops could be devised and universally adhered to.

Flûte-à-bec (*Fr.*) [Flute.]

Flûte d'amour (*Fr.*) A low-toned flute, an A flute, sounding a minor third below the notes actually written. It is now obsolete.

Flûte douce (*Fr.*) An organ stop. [Flute.]

Flûte traversière (*Fr.*) The German flute.

Fly. A hinged board which covers the keys of the pianoforte or organ when not in use.

F moll (*Ger.*) The key of F minor.

Foco (*It.*) Fire, spirit.

Focoso (*It.*) With spirit, ardently.

Foglietto (*It.*) A first violin part; the *leader's* part, which contains cues, &c., used by a conductor in the absence of a full score.

Fois (*Fr.*) Time, as *première fois*, first time; *dernière fois*, last time (of repeating), &c

Folia. A Spanish dance, similar to the fandango. The tune of a folia was sometimes written on a ground bass, as was also the Chaconne and Passacaille. [Follia.]

Foliated. A melody or portion of plain-song is said to be *foliated* when ornamental notes have been added above or below those of which it originally consisted.

Follia (*Sp.* and *It.*) Variations upon an air or melody, in which the ingenuity was held of more value than beauty. The name "Follias de España" became applied to laborious trifling in other matters besides music.

Fondamentale (*Fr.* and *It.*) Fundamental. *Basse fondamentale, basso fondamentale*, fundamental bass.

Fondamento (*It.*) (1) Fundamental bass. (2) The root or generator of a chord.

Fonds d'orgue (*Fr.*) The foundation stops, the diapasons and 8 ft. flutes on English organs, the *principals* of foreign instruments. In general, all flue stops of 8 ft. pitch, except solo stops of peculiar quality of tone.

Foot. (1) A metrical measure, *pes* (*Lat.*) (2) A drone bass. (3) The chorus of a song. (4) The part of an organ pipe below the mouth. (5) To foot, to dance.

Forlana (*It.*), *fourlane* (*Fr.*) A dance much in favour with the gondoliers of Venice. The tune is a lively measure in 6-8 time, and is similar to the Tarantella, but not so varied in its motions. It is said to have been first danced in Friuli, and to take its name from that fact.

La Fourlane Venetienne ou La Barcariuole

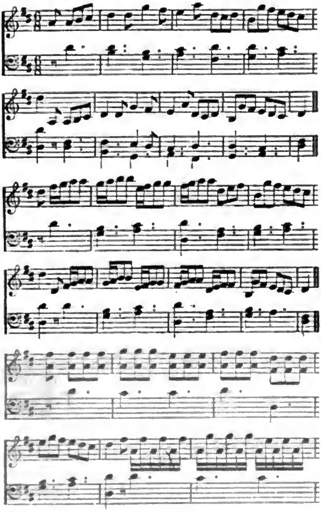

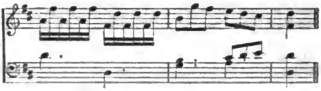

Form. The shape and order in which musical ideas are presented.

This definition is, perhaps, the nearest that can be given of a word of such general meaning. Form has been divided into harmonic and melodic. By harmonic form is meant the key-tonality of chords, such, for instance, as would be illustrated by a comparison of a composition by Palestrina with one by Spohr. But this question of the key relationship of chords is now generally made subordinate to the study of harmony, and is taken from the domain of form. By melodic form is meant the proper grouping of the successive sounds which form a tune. This, again, is made almost foreign to the higher meaning of form, and is held to be subordinate to the laws of rhythm. In its highest sense, form has relation more to the development than to the details of a composition.

In attempting to classify and give names to the portions of music which, by their combination or succession, go to make up a composition or movement, it will be necessary to say at once, that there is no settled or conventional usage of the terms employed, and all that is here done is to bring together those most commonly known, and as to whose meaning but little difference of opinion exists.

The component parts of melodic forms may be arranged according to the following order, (*a*) Motive or Figure; (*b*) Section; (*c*) Phrase; (*d*) Sentence; (*e*) Subject.

The word Motive has several meanings; primarily it signified the rhythmic effect produced by a Figure, or group of notes producing a complete and distinct impression. When used in this sense, a single bar might contain a simple motive, and the following two bars would be called a compound motive:

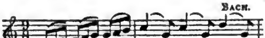

Before the existence of definite sonata form, the nature of the Motive was of the first importance, as it generally left its mark on the texture of the whole movement of which it was the initial conception. But now that distribution and contrast of key-tonalities have become such an important element in Form, the development of the "figure" or Motive (in its old sense) is relegated to a portion of a movement intentionally set aside for this purpose. Hence it will be easily seen how a whole subject (in its modern sense) came to be called the *motive*, as being something complete in itself as an initial conception, and as containing also what was required for future treatment by the composer. Next to the *figure*, or smallest sub-division, we have the *phrase*, formed by a group of figures; and above that, the *sentence*, formed of two or more phrases; and higher still, a complete *subject*, which may be limited to one sentence, or be formed of a group of sentences.* Subjects are either Principal or Subordinate. A principal subject may be described as one of the essential portions of the construction; it is a text proposed as the basis of a discourse, whereas a subordinate subject is analogous to a digression, however necessary or important it may be. The number, character, disposition, and treatment of subjects give to a *movement* its special name. A number of movements in succession, according to their different characters, constitute a complete *work*, or *opus*, as oratorio, opera, symphony, sonata, suite, &c.

But modern music of the higher type, that is to say, purely instrumental music, is so dependent for its beauty on the laws of sonata form that the word "form" is being gradually limited in its meaning to its special symphonic character. The story of the birth and growth of modern form takes us back to the period when instrumental music began to have a recognised existence independent of its function as an accompaniment of voices. On the one hand, the unisonous string accompaniment of madrigals and other polyphonic vocal compositions suggested that the instrumental parts might be dissociated and played separately; on the other hand, dance tunes, when they had developed from a single part for one instrument into a short harmonised movement for instruments of various types, became grouped into sets, the dances having a conventional mode of succession. The former of these two branches of the flowing stream of the art of music soon sank into disuetude, fragments of imitative counterpoint strung together without any method or plan when played on instruments, were found to be insufferably dull. But the dance tunes, already played freely on the rapidly-improving keyed instruments in the homes of the educated classes, offered a channel of interest which the then rudimentary orchestra was not slow to utilise. It is perhaps worthy of remark that as the orchestral *suite* had its counterpart for home use in the *suite de pièces* for the clavecin or harpsichord, so also has the orchestral symphony had its corresponding *sonata* for chamber use, and it may not be pushing the analogy too far to say that the counterpart of the madrigal for many voices may possibly be found in the glee for voices *soli*. For an account of the history of form, the reader is referred to Parry's "Art of Music," and for the same, stated in a more concise manner, to the earlier portion of Hadow's Primer on Sonata Form, from which he will obtain clear proof that the higher types of form are all to be found imbedded, in a rudimentary state, in the earliest attempts to create short national or popular tunes in a form which should create interest and give pleasure. This points to the fact that the pleasure we derive from contrast of key-tonality, and the pitch and character of themes, is part of our mental organization, and comes under psychological laws.

The study of form is most important to the composer. Without an adherence to its rules compositions are liable to become incoherent, unintelligible, and amorphous, especially in these days, when there is a great tendency, arising from ignorance or mistaken intention, to create music without much or any regard to form; which is as much a necessity to a musical composition, as it is to the design of a picture, a building, or a piece of sculpture. In classical compositions the sonata form is the basis upon which is constructed the Symphony, the Concerto, the Overture and the class of work from which it derives its title.

The sonata may consist of three movements in contrasted tempi and varied forms, but the first movement must be written according to given rules, which will be shown in detail after a general description of the ordinary arrangement of the sonata. Of the three movements the first should be an "allegro," with or without an introduction in slow *tempo*, though sometimes this is dispensed with. It may here be remarked, that whatever key their first movement is written in is the key by which the symphony is known, and all the other movements must be written in keys akin to it, but the last one must be the same as the first. The second movement marked with any *tempo* from andante to adagio, is usually called the "slow movement; the last movement is usually an allegro, and may be written on the plan of the first movement, or in what is called rondo form. As the symphony is the most important work in which the sonata form is employed, a description of the usual method of constructing it is subjoined, on the principle that the greater includes the lesser. In addition to

---

* This use of the word subject must not be confused with the subject (*canto fermo*) of a piece of counterpoint, or the subject of a fugue.

Digitized by Google

the number of movements in the sonata proper, the symphony has a minuet and trio or a scherzo, movements which are not necessary in the sonata.

The ~~minuet or the~~ scherzo—the latter most usual since Beethoven's day—ordinarily occupies the third place in the order of the movements in the symphony, but occasionally the adagio and scherzo change places, as in Mendelssohn's 3rd Symphony. Sometimes, as in Beethoven's Symphonies, Nos. 7 and 8, an allegretto appears instead of the andante or slow movement, but as there is no fixed rule for the order of the intermediate movements, a composer is at liberty to make such changes as he pleases.

Taking a symphony as a standard, the following is the usual order or form of each movement. The first of these, as has already been said, may begin with an introduction of a slow or moderately slow tempo. This in its design may foreshadow what is to appear in the succeeding allegro. If this is so written, it would give a coherence to the entire first movement;—and, on account of the frequently mysterious nature of this kind of instrumental music, coherence and intelligibility is a thing much to be aimed at. The length ~~of~~ the introduction may be left to the discretion of the composer; care being taken to lead well into the allegro or first movement proper. ~~For an~~ admirable example of this sort of treatment, the student is referred to Beethoven's Symphony No. 4. The allegro must contain two principal subjects: varied, and well contrasted, and written so as to give opportunity for good instrumental effect. The first of these is given in the tonic, and when the key of the symphony is well established the composer should prepare the introduction of his second principal theme. The old-fashioned way of doing this, if the symphony were in a major key, was to work up to a major chord on the supertonic of the original scale, sometimes with the 7th added, and by that means to glide into the second subject in the dominant; thus, if the key of the symphony were B♭ the movement would eventually arrive at a cadence on the chord of C with major third, the second principal subject being then heard in the key of F, or if the symphony were written in a minor key the composer worked up to a cadence on the dominant of the relative major—thus in a movement in C minor, the second subject always appeared in E♭ major. ~~But the~~ student is warned against blindly following this rule; it should be his aim to make his second theme grow out of the first; thereby avoiding the angularity of what may be termed the cut-and-dried school. The many ways of doing this must be left to the choice, guided by the ingenuity of the composer, but the chief thing to be borne in mind is to make the subjects melodious, striking, and workable. There may now be introduced one or two subsidiary or episodal subjects, growing out of, and in keeping with either the first or second principal subjects. With this ends what is called the first part of the first movement; a double bar is usually made here, and a repeat marked to the beginning of the allegro. This repeat in the first part of the first movement of a symphony is a *sine quâ non*, as without it the movement loses its distinctive form. The composer must end this portion of the work in the dominant if the key of the symphony is major; or in the relative major if it is written in a minor key. The most interesting part of this portion of a symphony now follows; it is technically known as the "free fantasia." Here the composer's imagination may be employed to its utmost limit; but he must bear in mind only to use material already brought forward in the subjects of the first part of the movement. A disregard of this injunction is the great mistake which composers (especially young ones) make now-a-days, and leads to diffuseness. Again, this portion of the symphony may be spoilt by the composer having to develop subjects which are not clear, well defined, and interesting; therefore to avoid this he should be careful to write subjects in the first part of the movement which will admit of much varied treatment. Having done this, he is at liberty to produce with them any legitimate effects: at the same time he should never allow his music to resemble a vague sort of improvisation, such as is frequently heard in many modern compositions, by so doing, all beauty of form vanishes, and the composer betrays a weakness and want of control over his subjects. These remarks apply with equal force to each portion of the entire work. Not more than a third of the movement should be devoted to the free fantasia; and when it is finished, the first principal subject is reintroduced in the original key, which in due course should be followed by the second principal subject, heard this time in the tonic. Then a coda may follow, after which the movement may come to a close; but it is suggested, for the sake of coherence, that the coda should be formed out of the material already employed, and it should not be too prolonged.

The second, or slow movement, may be similar in form to that of the first, but its character is entirely opposite; and further, no repeat is made at the end of the first part. The prevailing character should be that of pathos and repose; but though the two principal subjects should partake more or less of

Digitized by Google

this, the rhythm of each should be arranged so as to form a striking contrast to the other. In this movement the themes are usually much more elaborately treated than those of the Allegro, and sometimes variations are made upon the first of the principal subjects; examples of this latter style of treatment will be found in the slow movements of Beethoven's Symphonies, Nos. 5 and 9. The key should not be the same as that of the opening Allegro, such an arrangement would very likely produce monotony—the exception to this rule is when the original key is minor, the second movement could then appear in the same key; but it should be in the major mode, or *vice versâ*; any key may be chosen that has some kinship to that of the preceding movement.

If the minuet is selected for the third part, the composer must write it in accordance with the form of the dance of that name. But should the Scherzo be decided upon, he will find he has a much wider field for the expansion of his thoughts. The form of this movement may partake of that of the minuet, only the time should be generally double or three times as fast; it may also be in $\frac{2}{4}$ instead of $\frac{3}{4}$ time. Its character is usually light, fantastic, and even humorous. It is sometimes written with two trios, sometimes without any, as in the case in the Scherzo of Mendelssohn's 3rd Symphony. The key is usually the same as that of the first movement.

The last movement is an Allegro; and is generally somewhat lighter in style than the first movement; though its form may be the same. Rondo form may be adopted for this movement; or it may take the shape of an air with variations, as in the Finale to Beethoven's Symphony, No. 3. It is not necessary to repeat the first part of the finale, if it is written in the form of the first movement; and the free fantasia need not be elaborated to the same extent as in the opening Allegro. The key of the first and last movements must of necessity be the same.

The form of the Concerto is somewhat similar to that of the Symphony; but it differs from it in these respects: (1) The opening movement never appears with an introduction ~~placed before it~~. (2) The orchestra usually plays both the leading themes in the tonic before they are heard on the solo instrument in the usual symphonic form. (3) The repeat of the first part is not a necessity. (4) The Concerto never contains a Scherzo, and therefore consists but of three movements—the first of which should be an Allegro, the second an Adagio, and the third an Allegro. (5) The movements do not require to be developed at such great length as those of the symphony. (6) A cadenza is usually introduced towards the close of either the first or the third movement, sometimes written by the composer, sometimes left to the performers' improvisation. In all other respects the form of the concerto is identical with that of the symphony.

An improvement in the form of the concerto may yet be made in two particulars. The old plan of beginning the first movement by the orchestra playing the leading subjects in the tonic, before the entrance of the solo instrument, might be dispensed with as unnecessary, because the themes must appear towards the close of the movement in that key. Mendelssohn's Violin Concerto, and that for the Pianoforte in G minor, are examples of the advantage gained by beginning with the solo instrument at once, or after a few bars of orchestral prelude. The other suggestion for improvement is the abolition of all cadenzas, as being redundant. If they are written for the display of the performer's skill, surely the composer should give him sufficient opportunity for this during the movement. If they are to show that the subjects can be treated in a different manner, they should be included in that treatment in the body of either allegro or finale.

The modern overture in strict form should be written in one movement, usually an Allegro, with or without an introduction in a slower *tempo*, and partakes of the nature of the first movement of a symphony, without the repeat of the first part. The subjects of an overture may be lighter in character than those of a symphony; but they must appear in the same order, and be worked out in the same manner; greater importance being given to the coda in the overture than in the symphony. The introduction and allegro must, of course, be in the same key.

The form of the sonata is, as already shown, identical with that of the symphony in all points; the only exception being that a minuet or scherzo is not necessary. The development of the subjects of a sonata ought not to be of the same extent as those of the symphony, concerto, or overture; for the obvious reason that the tone-colour is much less varied; sonatas never being written for more than two instruments. Trios and quartets, &c., for strings and pianofortes, though written in this same form need not be developed to the same extent. Nor should the attempt be made to introduce grand symphonic effects in these works. A warning may be given to composers by pointing out how incomparably finer in effect Mendelssohn's D minor trio is to that of his trio in C minor, for the reason that in the latter work, especially in the last movement, passages are given to the violin and violoncello

Digitized by Google

which never produce the effect they are intended to convey: but nothing of the kind is found in the trio in D minor. Here everything has the character and form of chamber music; and composers should bear in mind not to write symphonic music when they are composing *musica di camera;* though the form of the work and its movements may be the same as those of a symphony. For a model pianoforte quartet the student is referred to that in G minor, from the pen of Mozart.

The string quartet for two violins, viola, and violoncello, is composed in the same form as the symphony; but also should be written without any striving after symphonic effect or development, care being taken to give each instrument, as much as possible, its equal share of work and independent motion. The quintet, the sestet, septet, and octet should all be written in sonata or symphonic form, each instrument having proper but not obtrusive prominence and independence. The student is reminded that the whole of these remarks are not to be considered absolutely final; since a composer may some day arise, who may alter the present forms of classical instrumental music as much as Haydn and Mozart did in their day. However, as such an one does not exist who has shown us any new form which is better than the old one, the student is exhorted to abide by the rules herein laid down, which are based upon precedent founded by the great masters.

Rondo form differs from sonata or symphonic form, in that the first part is not marked for repeat. The original subject does not modulate, but reappears in its key-chord at the close of the first period, and again after the modulation of the second subject, so that it must be heard three times.

The arrangement of a movement in rondo form is after the following order:

The first subject enters, sometimes without introduction, and remains in its original key. Then follows an episode, modulating into the relative major if the key be minor, or into the dominant if the key be major; after which comes the second subject in the dominant or relative major, as the case may be, followed by a modulation into the original key, to bring back the first subject. This ends the first part, which is not marked for repetition. The second subject, modulating into distant keys, commences the second half. This is followed by the first subject; then an episode, preparing the way for the second subject, and a final episode and coda, generally in the original key, with slight passing modulations. It must be understood that this general outline is sometimes varied by the genius of the composer; but, as a rule, the order indicated above is followed.

Fortsetzung (*Ger.*) Continuation. Further development or expansion of an idea.

Fort (*Fr.*) *Forte* (*It.*) Loud; expressed in music by the abbreviations *for.* or *f.*

Fortemente (*It.*) Loudly, vigorously, with force.

Forte-piano or **fp.** (1) Loud, then soft· strongly accented. (2) The pianoforte.

Forte possibile (*It.*) As loud as possible.

Fortissimo (*It.*) *Lit.* the loudest. Very loud. The letters *ff* or *ffor* are used as abbreviations of the word.

Forza, con (*It.*) With emphasis.

Forzando (*It.*) *Lit.*, Forcing. Emphasis or musical accent upon specified notes or passages, marked by the signs *fz*, *sf* or >

Forzato (*It.*) [Forzando.]

Fourchette tonique (*Fr.*) Tuning-fork.

Fourier's theorem. [Acoustics, § 11.]

Fourniture (*Fr.*) A *mixture* stop on an organ.

Fourth. An interval of four notes.

Fourth flute. [Quartflöte.]

Française (*Fr.*) A dance in triple measure, similar in character to the country-dance.

Franchezza (*It.*), **Franchise** (*Fr.*) Freedom, confidence.

Franculus (*Lat.*) A mediæval sign or neume for an ascending *brevis plicata* [Neumes.]

Franzton (*Ger.*) French pitch; lower than the recognised English concert pitch.

Frasi (*It.*) Phrases.

Freddamente, con freddezza (*It.*) With coldness, indifference.

Fredon (*Fr.*) (1) Vocal ornaments at the will of the performer; a tremolo or quavering upon every note. (2) The *humming* of a tune.

Free chant is a form of recitative music for the Psalms and Canticles, in which a phrase, consisting of two chords only, is applied to each hemistich of the words. The author of the form, Mr. John Crowdy, in his "Free Chant Cadences," claims for it that it removes all difficulties in dividing the words, and enables the unskilled worshipper to join confidently in the chanting, without the assistance of any marks beyond the colons provided for the purpose in the Prayer Book.

Free fugue. A fugue in which the answer and general treatment are not according to strict rules. [Fugue.]

Free parts. Additional parts to a canon or fugue, having independent melodies, in order to strengthen or complete the harmony. [Canon.]

Free reed. [Reed.]

Free style. Composition not absolutely according to the strict rules of counterpoint.

Digitized by Google

Fregiatura (*It.*) An ornament, embellishment.

French horn. [Horn.]

French sixth. [Extreme sixth.]

French flat-tuning. [Flat-tuning.]

French violin clef. The G clef, placed upon the first line of the stave. [Clef.]

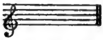

Frets. Small pieces of wood or ivory placed upon the finger-board of certain stringed instruments, to regulate the pitch of the notes produced. By pressing the string down to the finger-board behind a fret, only so much of the string can be set in vibration as lies between the fret and the bridge. Frets are, therefore, nothing more or less than little bridges; hence the word μαγάς came to signify a bridge or a fret. The Egyptian lutes had frets made of camel-gut, tied or glued round the finger-board. All the viols contained in a chest had frets, and some of the early forms of the *violin* were even furnished with them. But not only do they prevent the rapid fingering of difficult passages, but entirely deprive the violin of one of its most charming qualities, that of *slurring* or *portamento*, an attempt to produce which will, on a fretted instrument, result in a well-defined chromatic scale. Another reason for the abandonment of fretted violins was that, in extreme keys, the intervals could not be *tempered*.

Fretta, con (*It.*) With speed, haste, hurry.

Freie Schreibart (*Ger.*) Free writing; composition in a free style.

Frisch (*Ger.*) Lively.

Fröhlich (*Ger.*) Joyous, cheerful, gay.

Frosch (*Ger.*) The nut of a violin bow, into which the lower end of the hairs is fixed, and which, when moved up or down by means of the screw, tightens or slackens their tension.

Frottola (*It.*) A ballad.

F-Schlüssel (*Ger.*) The F or bass clef.

Fuga (*Lat.*) A fugue, *æqualis motus*, a real fugue; *authentica*, a fugue with a subject in the authentic part of the scale; *canonica*, a fugue in canon; *contraria*, a fugue by inversion; *impropria*, or *irregularis*, a free or irregular fugue; *in contrario tempore*, a fugue, the answer of which is differently *accented* to the subject; *libera* or *soluta*, a free or irregular fugue; *per arsin et thesin*, by inversion (1) of rhythm, (2) of interval; *retrograda*, a fugue by contrary motion; *obstinata*, a fugue in which a definite figure is maintained, Fuga (sixteenth century), a canon.

Fuga (*It.*) A fugue; as, *fuga doppia*, a double fugue; *fuga ostinata*, a fugue in which a definite figure is maintained; *fuga ricercata* (1) highly scientific fugue; (2) a fugue without episodes; *fuga sciolta*, a free fugue. [Fugue.]

Fugato (*It.*) In the fugue style; a composition containing fugal imitation, but which is not in strict fugue form.

Fuge (*Ger.*) A fugue.

Fughetta (*It.*)
Fughette (*Ger.*) } A short fugue.

Fugue. A polyphonic composition constructed on one or more short subjects or themes, which are harmonized according to the laws of counterpoint, and introduced from time to time with various contrapuntal devices; the interest in these frequently heard themes being sustained by diminishing the interval of time at which they follow each other (the *stretto*), and monotony being avoided by the occasional use of episodes, or passages open to free treatment.

So varied are fugues in their character, that it is impossible to give any definition which shall include all kinds, but from what has just been said above, it will be at once seen, that they differ from all other formal or set compositions (the canon only excepted) in that each component part (which might in other works be only a means of harmony) must stand in important relationship to every other part, sometimes even to the extent of being interchangeable with any one of them. The key-relationship, and also the rhythmical form of the sections and phrases of a fugue, have always been modified by contemporary art, and by this means a fugue of an early period may be easily distinguished from one of a later date, until in some modern examples the influence even of the sonata form is plainly discernible. As the growth of the splendid form now known as a fugue has been gradual, having extended over more than three centuries, it is not surprising that the name should, from time to time, have borne various meanings. In old writers it is sometimes used to signify a short theme, the measure or figure of which is to be frequently repeated; at other times, a canon, because herein one part enunciates a subject and then as it were, takes to flight (*fuga*), while the other, or others, pursue it closely note for note. Canons often formed an important ingredient of early fugues.

It is easy to trace the germ of the fugue in the higher developments of counterpoint. When music in two parts was written in conformity with the laws of double counterpoint, each performer found himself setting forth the theme proposed by the other, and the good effect thus produced would naturally suggest a repetition of the theme at other intervals (as in counterpoint), and also in other keys. The two elements of a fugue which separate it from the higher forms of

Digitized by Google

counterpoint are, first, the enunciation of the subject by itself, without harmony; next the *stretto* or drawing of subjects and answers more closely together. The former is not properly included in counterpoint, as *point* (or note) is no longer *against point* when one part is heard alone; nor is the latter traceable among the various devices of the art of counterpoint proper.

Fugues have been divided into many classes according to the point from which they have been regarded.

(1) By number of parts; as, a fugue in two, three, four parts, &c. (*a* 2, *a* 3, *a* 4, &c.)

(2) By number of subjects; as a *double fugue*, having two subjects; a *triple* fugue, three subjects, &c.

(3) By the relation of subject and answer; as a *fugue by inversion*, when the answer moves by the intervals of the *inverted* subject; by *augmentation* or *diminution*, when the answer has notes of double, or half the length of those of the subject respectively.

(4) By the scale-relation of subject and answer; as a *tonal fugue*, when the answer is modified according to prescribed rules, so that it shall remain within a given compass, or, within a given key; a *real fugue* when the answer is at a measured interval to the subject note for note.

(5) By its adhesion to, or neglect of, the laws of fugue form; as a *free fugue*, a fugue in which *strict* form is occasionally, or for the most part disregarded.

(6) By its scale, or the scale which predominates in it; as a *Doric fugue*, when the subject, and perhaps also development, is in the Doric mode; a *diatonic fugue*, in which diatonic harmony prevails; a *chromatic fugue*, when chromatic passages abound.

The chief elements of a fugue are:—

(1) The subject. *Dux*, *propositus*, (*Lat.*) *guida*, (*It.*) *antecedent*, &c.

(2) The counter-subject, or, contrapuntal harmonization of the answer by the part which has finished the enunciation of the subject.

(3) The answer. *Comes*, *Responsio* (*Lat.*); *consequenza* (*It.*); *consequent*, &c.

(4) Episodes.

(5) The stretto.

(6) The pedal point—*point d'orgue* (*Fr.*); *Orgelpunkt* (*Ger.*)

The whole of these are bound together into perfect unity, from the fact that the answer is either identical with, or a prescribed imitation of the subject; the counter-subject or fragments of it are of frequent use as the material of episodes; the stretto is usually founded on the subject or counter-subject; and the pedal point forms the basis of ingenious treatments of the subject or answers, and sometimes even as the basis of the stretto. The first giving out of subjects and answers is called the Exposition; and when repeated with a different arrangement of the parts, the Counter-exposition.

It can be seen from the above, that handled by a genius, fugue may be infinitely plastic in regard to form. But it must not be forgotten that in its earliest existence it was wretchedly mechanical, as the following directions how to compose a fugue in two parts from Fux (Welcker's English Translation) will prove. "First choose a subject suitable to the key you intend to compose in, and write down your subject in that part wherewith you intend to begin. This done, and having first examined your subject whether it be comformable to your key; if so, repeat the same notes in the second part, either in the fourth or fifth, and whilst the second part imitates the first wherewith you have begun, put such notes in the first part as will agree with your imitating part according to the directions given in the figurate or florid counterpoint, and after having continued your melody for some bars, regulate the parts thus, that the first cadence may be made in the fifth of the key. Then resume your subject mostly in the same part you have begun with, but by another interval, after having first put a rest of a whole or half bar, which however may be omitted in case there should happen to be a great skip instead of it. After this, endeavour to bring in your second part after some rest, and that before the subject of the first part draws towards a conclusion, and having carried on your subject a little longer, make your second cadence in the third of the key. Lastly, introduce your subject again in either part, and contrive it so that one part may imitate the other sooner than at first, and, if possible, after the first bar, whereupon both parts are to be united, and the fugue finished by a final cadence."

The musical example which he then gives as embodying the result of all this learning, is as follows:

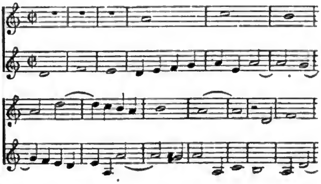

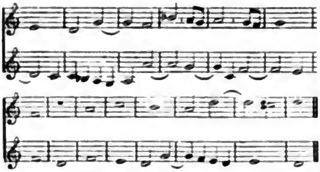

But the art of fugue was not long to remain thus lifeless. The successive improvements made by great masters have exalted it to the highest perfection, and have made it one of the noblest walks of the art of music.

The best way of showing the construction of a fugue will be to describe in detail the nature of the six constituent parts just now enumerated.

(1) The subject should not be very long if it does not contain any modulation, because a lack of interest may result. On the other hand, if it be very short, its treatment in the stretto will be difficult. It generally commences on the tonic or dominant of the scale.

Subjects may be broadly divided into diatonic and chromatic. Of course, a vast number of fugue-subjects lie between these two boundaries, but by a diatonic subject must be understood one on which an author intends to construct a fugue whose interest shall arise from genuine contrapuntal treatment and device, and simple modulations from key to key. By a chromatic subject is meant one which a composer takes with the avowed intention of constructing a fugue whose interest shall result from a complicated interweaving or frequent contrasting of changing key-tonality, with ordinary development of the subjects. The simplest form of diatonic fugue-subject is that which lies in a compass of a fifth. *e.g.*:

Ex. 1. J. S. Bach.

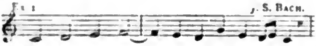

Ex. 2. Eberlin.

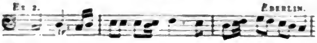

Ex. 3. Graun.

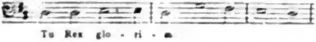

Or, when it reaches the compass of the sixth, *e.g.*:

Ex. 4. Haydn.

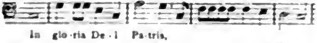

Ex. 5. J. S. Bach.

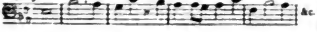

Ex. 6. Mozart.

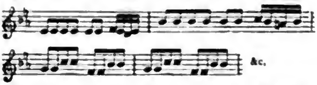

Ex. 7. J. S. Bach.

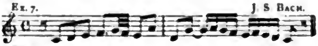

Ex. 8. Mozart.

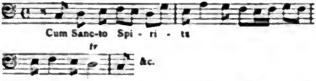

The following is given in order to show a grand subject in this compass, although not strictly worked out:

Ex. 9. Beethoven.

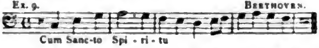

Diatonic subjects may, however, reach a very extended compass, especially in instrumental music:

Ex. 10. Handel.

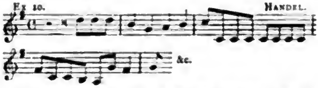

The following is remarkable both for its extended compass and length:

Ex. 11. (*Fuga a tre voci con licenza.*) Beethoven.

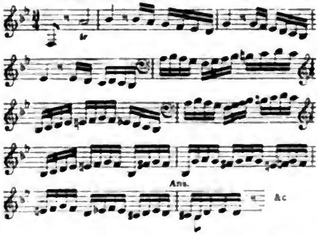

Scale passages, or such as move up or down an octave, have always been largely used as subjects, both in the major, *e.g.*:

Ex. 12. Benevoli.

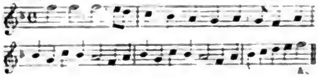

And in the minor, *e.g.*:

J. S. Bach.

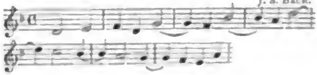

Sometimes the octave compass of a subject lies between the fifth above and fourth below the tonic, both in the major, *e.g.*:

Ex. 13. HANDEL.

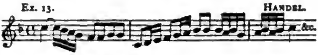

And also in the minor, *e.g.*:

Ex. 14. BACH.

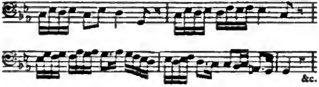

Chromatic subjects are also of varied extent and difficulty. The following, which is capable of much contrapuntal treatment, is commonly met with:

Ex. 15. SALA.

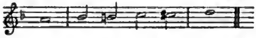

More elaborate chromatic subjects are often found, *e.g.*:

Ex. 16. EBERLIN.

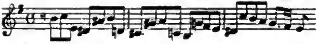

Ex. 17. J. S. BACH.

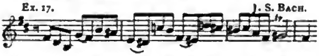

Sometimes both diatonic and chromatic passages are included in the subject, *e.g.*:

Ex. 18. J. S. BACH.

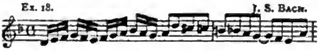

The interval of a diminished seventh has always been a favourite element of fugue subjects, *e.g.*:

Ex. 19. BACH.

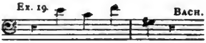

Ex. 20. HANDEL.

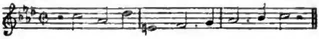

Ex. 21. J. S. BACH.

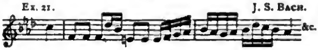

Ex. 22. MOZART.

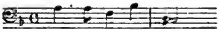

Ex. 23. J. S. BACH.

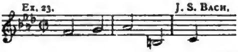

Ex. 24. EBERLIN.

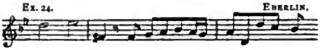

Subjects most commonly begin the key-note or its fifth, but there are exceptions to this rule, *e.g.*:

Ex. 25. (On the Second of the Scale.) BACH.

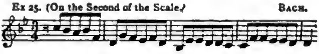

Ex. 26. (On the Third.) LEUTHARD.

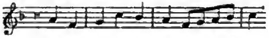

Ex. 27. (On the Fourth.)

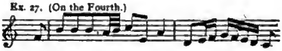

Ex 28. (On the Sixth.) MATTHESON.

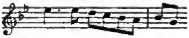

Ex. 29. (On the Seventh.) BACH.

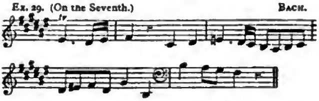

As a rule, the answer enters before the subject is finished, but exceptions are frequent. "He trusted in God" (Handel) may be cited as a well-known case. Sometimes after the subject has finished, a few notes are introduced to link it to the answer. These few notes are called a subject-coda or *codetta*. The name is also applied to the short passage sometimes connecting the answer and counter-subject with the re-introduction of the original subject (see Exs. 43 and 68).

(2) The counter-subject is primarily an accompaniment of the answer, and in a secondary sense, of the subject; but as such, must be according to the laws of strict counterpoint. It is usually written according to the laws of *double* counterpoint, in order that it may be used both above and below the subject or answer. Of course, the counter-subject may be in any species of counterpoint, but it most commonly is figurate, or florid, *e.g.*:

Ex. 30. MENDELSSOHN.

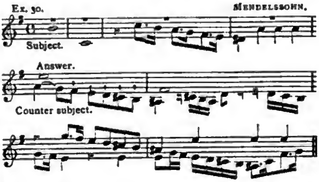

But when the counter-subject is in simple counterpoint it generally happens that it is in notes of greater length than those of the subject in a rapid fugue, *e.g.*:

Ex. 31. Bach.

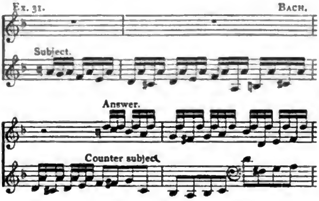

And *vice versâ*, shorter notes of counter-subject to those of the subject:

Ex. 32. J. S. Bach.

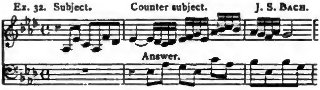

But something more is required of the counter-subject than to be a mere accessory to the subject and answer; it is very often used as an episodal theme, either just as it stands, or in a slightly modified form. In the following example the counter-subject of Bach's beautiful E major fugue is given:

Ex. 33.

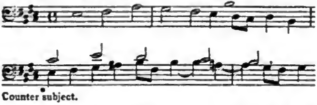

Before the development of the Fugue has proceeded far, this (from *) is used as the subject of an episode, *e.g.*:

Ex. 34.

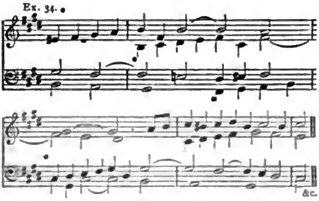

If, therefore, the counter-subject is intended for separate use and treatment, it is necessary that it should be melodious in itself, as well as capable of forming good counterpoint in combination with the answer. In speaking of a tonal fugue it will be shown that the counter-subject sometimes has to undergo a change in order to suit both subject and answer.

The term counter-subject is often applied in a manner which leads to much confusion. If in a fugue with two subjects the second subject is given out at the same time as the first, thus forming an accompaniment to it, it is by some called the counter-subject, instead of the second subject, *e.g.*:

Ex. 35. Mozart.

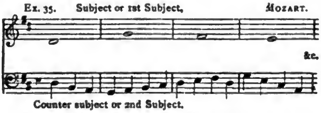

It were well if this use of the word counter-subject for second subject could be dispensed with, the former being limited to the signification of that counterpoint added to an answer or subject by a part which *has already gone through the subject or answer.* Nothing is gained by limiting the use of the words *second* and *third* subjects to such as are introduced separately.

(3) The answer of a fugue is one of the most important parts of its construction. If the subject be wrongly answered, the effect and success of the whole composition is marred. For not only does a wrong answer compel the construction of a false counter-subject, inasmuch as the counter-subject must be the accompaniment of the answer whether it be right or wrong, but also, it overthrows those episodes founded on the counter-subject, and not unfrequently the stretto too, as being founded on a close combination of subject and answer.

The large class of fugues called *tonal*, are so termed because the answer undergoes some slight modification in order to prevent a departure from the key-tonality of the subject. Roughly speaking, all answers are a fifth above or a fourth below the subject. If this relation of answer to subject were strictly carried out, a modulation in every answer would be inevitable. The answer is, however, often purposely made at strict intervals to the subject; in which case, the fugue is called strict or *real*. These two classes of fugues must be considered carefully.

First as to tonal fugues.

The idea which underlies all tonal treatment of answers is, that the scale is equally divided into two parts, namely, from the tonic up to the dominant, one part; from the dominant up to the octave-tonic, the other. But as a matter of fact the first half, tonic to dominant contains five diatonic steps, while the second, dominant to tonic, only contains four. This is the source of the whole difficulty of making a correct answer to any given subject, *e.g.*:

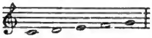

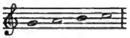

Now when the subject proceeds from tonic to dominant direct, the answer must proceed from dominant to tonic, *e.g.*:

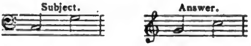
Subject.
Answer.

and *vice versa*, *e.g.*:

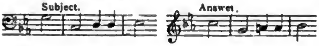
Subject.
Answer.

But if notes lying between the tonic and dominant are introduced, or if the subject exceeds the compass of a fifth, it is not easy to say that any uniform principle governs the relation of answer to subject, except that two notes must be represented by one, *e.g.*:

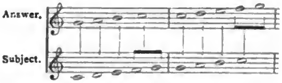
Answer.
Subject.

The following examples show how tonic is answered by dominant, and dominant by tonic:

Ex. 36.

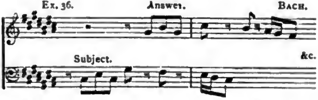
Answer.
BACH.
Subject.
&c.

Ex. 37.

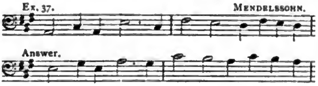
MENDELSSOHN.
Answer.

and, by the next it will be seen, that the *sub*-dominant also is answered by the tonic:

Ex. 38.

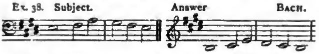
Subject.
Answer
BACH.

Ex. 39.

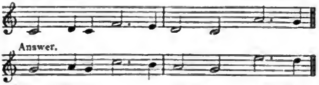
Subject.
MENDELSSOHN.
Answer.

Ex. 40.

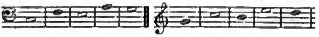
Subject.
Answer.
MOZART.

The following illustrates the application of the same principle in a case where the subject proceeds down to the dominant and then passes above the tonic:

Ex. 41.

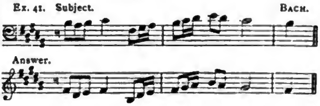
Subject.
BACH.
Answer.

The application of this principle to subjects in the minor key is much less easy than in the major. When, as formerly, fugues were often composed in the church modes, the position of the (so-called) *dominant* and *final* of the mode largely influenced the relation of answer to subject, but, although these complications do not lie in the path of the modern student, yet there is still much uncertainty and discrepancy as to the particular treatment of the *sub-tonic* of the minor. This arises from the fact that several sorts of minor scale are still in use, and the composer naturally frames his answer in accordance either to that kind of scale most congenial to him, or to that most capable of bringing into prominence the melodic form of his subject. Minor subjects are often too very chromatic, a fact which adds to the difficulty of forming a correct answer.

In its simple state, a minor subject is in effect answered by a modulation into the *minor* key a fifth above (or fourth below), *e.g.*:

Ex. 42.

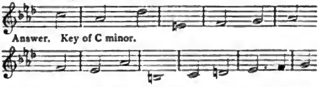
Subject. Key of F minor.
Answer. Key of C minor.

and the following shows the method of return to the original key, sometimes called a *codetta*·

Ex. 43.

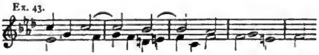

Another instance is here given:

Ex 44.

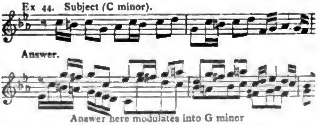
Subject (C minor).
Answer.
Answer here modulates into G minor

It is not an easy thing to connect two minor keys a fifth apart in a pleasing manner; it takes some little time to accustom the ear to the sound of the minor third of the new key; but the leading note of the old key must, of necessity, be discarded before the re-entry of the subject. In the fugue just quoted Bach inserts two bars between the close in G minor, as above, and the re-entry of the subject, namely:

Ex. 45. (Codetta.)

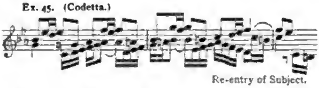

Re-entry of Subject.

As minor subjects naturally lead to a modulation in the answer, it happens, as might be expected, that the minor fugues are chiefly *real*: for if a modulation must take place at all, it may as well include the whole of the subject as its closing portion, tonal alterations of the answer are therefore rendered to a certain extent unnecessary.

But in strictly *chromatic* fugues the tonal answer is very common, as the following examples will show, although it will be observed that, in some cases, the answer, before many notes are past, becomes a mere transcript at a fifth above or fourth below of the original subject.

The next example is purely tonal:

Ex. 46.

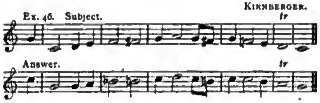

The following is mixed, being partly tonal, partly at a strict interval:

Ex 47.

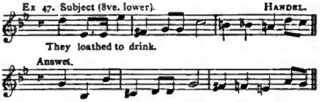

The tonal alteration of the answer to minor subjects often extends no further than the first note, *e.g.*:

Ex. 48.

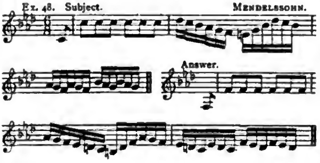

In major subjects this also happens, *e. g.*:

Ex. 49.

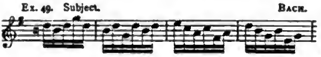

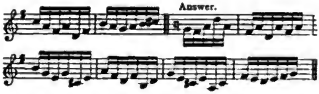

This is more noticeable in short chromatic subjects, *e.g.*:

Ex. 50.

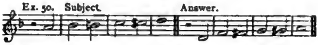

Ex. 51

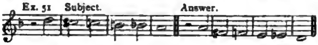

Enough has been said to show the general force of the laws of tonal answer. Study and experience are the only means of cultivating a true perception of this peculiar relation of subject to answer. Many writers have attempted to draw up a regular code of laws, but the exceptions which persistently come forward render them almost useless.

A real or strict fugue is one in which the answer is *throughout* at the interval of a fifth above or fourth below the subject, *e.g.*:

Ex. 52.

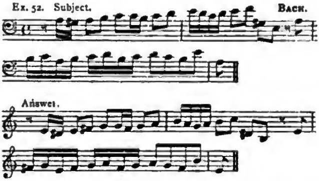

Ex. 53.

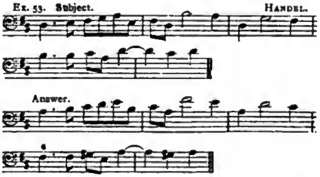

It is unnecessary to give more examples of this exact and constant distance between subject and answer.

The question naturally arises, How is it to be known when an answer ought to be tonal or real? It is only possible to answer this in the most general way. If a subject has one or more direct melodic progressions from tonic to dominant or dominant to tonic it is difficult to make the answer *real* or *strict* without

---

* Made one octave lower to suit the compass of ordinary tenor voices.

Digitized by Google

giving an unpleasant effect of unnecessary change of key; whereas, if the subject consist of a series of grades of the scale, it is difficult to make the answer *tonal*, without producing the effect of unnecessary alteration of melody (inasmuch as two notes have to do duty for one, and *vice versâ*). This is all that can be said, except that special prominence of the *subdominant* in the subject seems to demand a strict answer, *e.g.*:

Ex. 54.

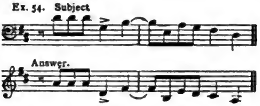

In cases where a fugue has more than one subject, if the second subject partakes of the tonality of the expected answer, and is introduced in the position ordinarily occupied by the answer, the answers of both first and second subjects may take place at the octave, sometimes without any alteration of the position of the parts, *e.g.*:

Ex. 55.

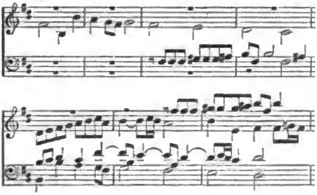
BACH.

The subjects are as often answered in the octave, but in *inverted positions*, *e.g.*:

Ex. 56.

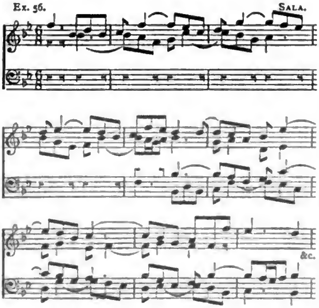
SALA.

It has already been stated that the counter-subject is often written in double counterpoint, so that it may be used without grammatical error, both above and below the subject. What has been said of a counter-subject applies with equal force to a *second* subject, as the above example (56) shows.

But the first and second subject are given out very frequently, each at its own proper tonal distance, *e.g.*:

Ex. 57.

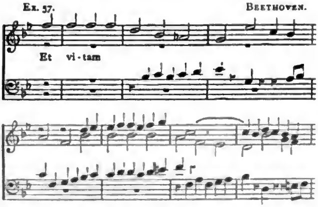
BEETHOVEN.

Sometimes the second subject appears after the first subject, but is not answered in the position expected, an answer to the *first* subject taking its place.

Ex 58.

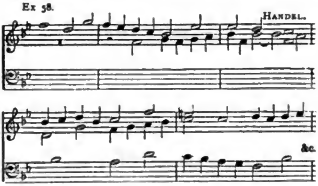
HANDEL.

In other cases, the first subject is, after its first enunciation, set aside for a lengthy treatment of the second subject, the first being reintroduced when the development has been proceeded with.

The above examples of fugues with two subjects, have tonal answers; but this is not always the case, as the following example of strict answer shows:

Ex. 59.

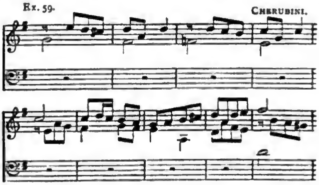
CHERUBINI.

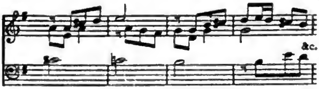

In fugues having three or more subjects, there seems to be no rule whatever as to the order or position of their entry. Sometimes they are enunciated in their order immediately after, or overlapping, each other; *e.g.*:

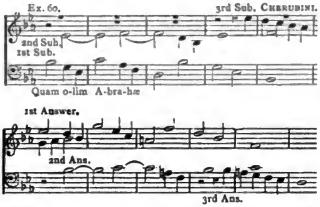
Ex. 60. CHERUBINI.
3rd Sub.
2nd Sub.
1st Sub.
Quam o-lim A-bra-hæ
1st Answer.
2nd Ans.
3rd Ans.

They are not unfrequently introduced and developed separately, at long intervals of time, and only brought together towards the close of the work; in this case, of course, the answers are each true to the subject as delivered, and the fugue has the form of two or more separate developments which are capable of coalition; *e.g.*:

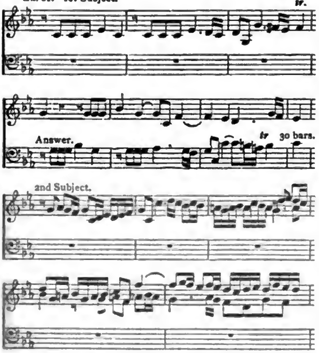
Ex. 61. BACH.
1st Subject.
tr.
Answer.
tr
30 bars.
2nd Subject.

In some cases the subjects are given out one by one, without any development being attempted till all have been heard, as in the well-known movement, "Let old Timotheus," from Handel's "Alexander's Feast," in which the four following subjects are given out successively by solo voices:

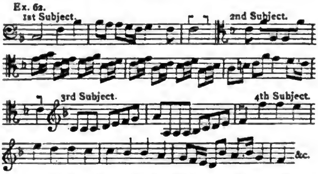
Ex. 62.
1st Subject.
2nd Subject.
3rd Subject.
4th Subject.
&c.

Some fugues have what is called a *free* part; that is a part whose sole function it is to support or supplement those constructed in the prescribed fugue form. Such examples generally take the shape of a regularly constructed fugue, accompanied by and built upon a separate bass, or *basso continuo*:

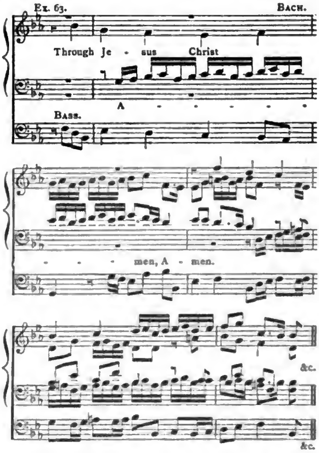
Ex. 63. BACH.
Through Je - sus Christ
A - - - - - -
Bass.
- - - - men, A - men.
&c.
&c.

Other orchestral parts sometimes accompany a fugue, *e.g.*:

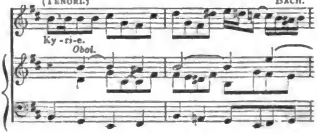
Ex. 64. BACH.
(Tenori.)
Ky-ri-e.
Oboi.

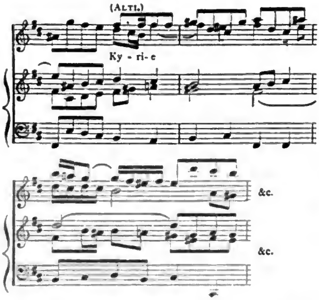

From the two preceding examples, it will be seen that free accompaniment may take place, whether the fugue is tonal or real, or whether it has but one, or more than one, subject.

A chorale is not unfrequently introduced into a fugue as a free part, as the well-known chorus, "But our Lord," in Mendelssohn's "St. Paul." There can be little doubt that the freedom of writing which distinguishes later fugues, was largely brought about by the habit of writing free parts on a ground-bass.

When the answer of a fugue is an inversion of the subject, the fugue is said to be *al rovescio*, *e.g.*:

Ex. 65. HANDEL.

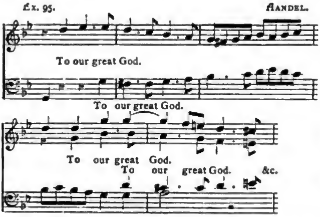

Fugues "by inversion," like all other kinds of fugues, are sometimes accompanied by free parts, *e.g.*:

Ex. 66. BACH.

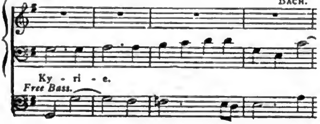

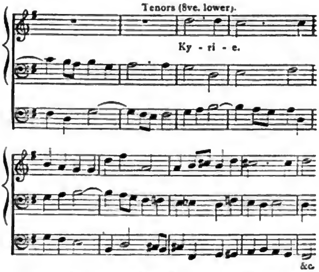

Episodes are passages introduced into a fugue, in which the actual development of the subject or subjects is for a time suspended in order to give some variety and relief to the ear. But it is important that episodes, while affording variety, should not disturb the character of the fugue of which they form part. To effect this object, they are generally made up of free or imitative treatments of a fragment of the subject, or of part of the countersubject, or of the coda connecting subject and answer, or of some new subject not dissimilar in style to one of these. An example (34) has already been given, of an episode formed on the countersubject. The following is an episode formed on a fragment of the *subject* given in Ex. 49:

Ex. 67. BACH.

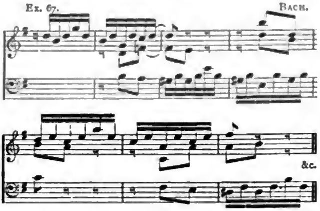

The following is an example of an episode, founded on a *coda:*

Ex. 68. Subject. BACH.

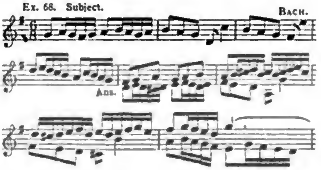

Coda connecting close of answer with re-entry of subject:

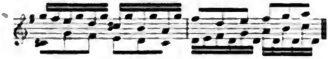

Episode founded on the figure of the above coda:

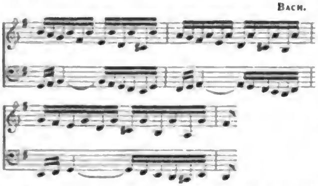

Episodes founded on a theme not part of subject, countersubject, or coda, are very commonly met with; in the following example the theme is quite congruous with the subject:

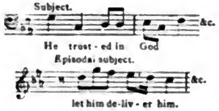

A fine example of episodes of a totally different character to the fugue into which they are introduced is to be found in Bach's organ-fugue:

Ex. 69.

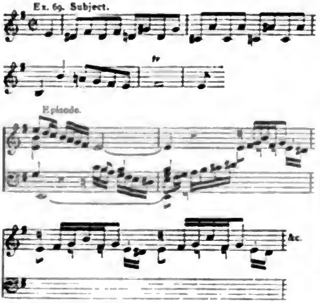

The stretto is that part of a fugue in which the subject and answer are drawn or pressed closer together (Lat. *strictus* from *stringere*). In nearly all carefully constructed fugues the entry of the subject and answer is brought closer together from time to time, as the development proceeds, but the word *stretto* is only applied to that special passage in which the whole of the parts, or as many as possible, take up the subject at as short an interval of time as possible. The simplest illustration of a stretto will be found in the treatment of some simple diatonic subject such as the following:

Ex. 70.

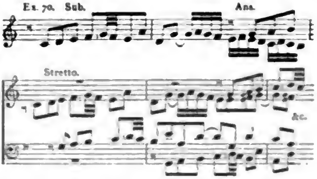

It not unfrequently happens that the subject and answer cannot be brought closer together than their original distance, owing to their harmonic inaptitude for such treatment. In such cases, an altered form of the subject, or a part of the coda, or part of the countersubject, or even an entirely new subject, may form the theme of the stretto. In many fugues there is no one passage which can be pointed out as *the stretto*, but the interest of the development is sustained by various other contrapuntal devices, or modulations.

The *masterly stretto* (stretto maestrale) is formed in strict canon. If the subject is not capable of such treatment it may be slightly altered for the purpose.

It is not necessary that the same intervals should be observed between the parts forming a stretto as are absolutely necessary in the enunciation of subject and answer, *e.g.*:

Ex. 71.

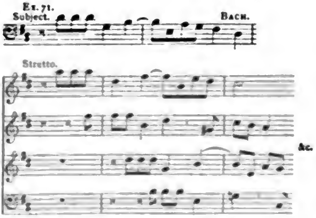

Strettos are often constructed on the pedal-point, an example of which will shortly be given.

Stretto, by augmentation or diminution, is when the subject, or subject and answer, are simultaneously introduced in notes of longer or shorter length, *e.g.*:

Ex. 72.

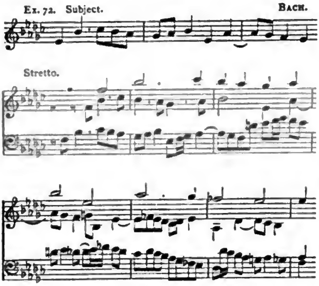

A pedal-point is a long-sustained note, generally the dominant, on which imitation, subject and answer (simply or by augmentation and diminution), or even the stretto itself, are constructed. It is not always found as an essential part of a fugue, a vast number of fugues, especially for the pianoforte, are without it. But in vocal fugues with accompaniment, and in fugues for the organ, it can always be introduced with fine effect. Modern composers have not neglected this interesting element of the art of fugue, as the next example shows:

Ex. 73.

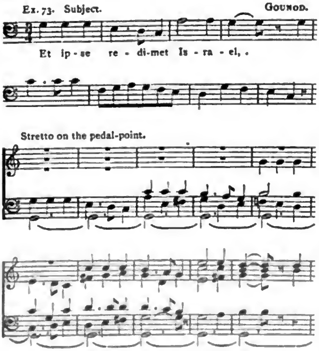

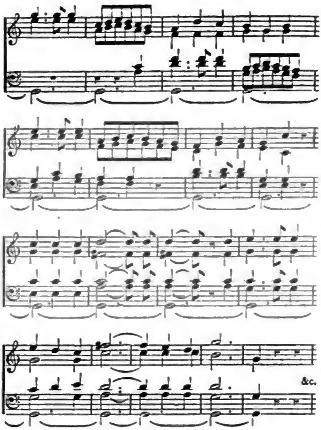

Sometimes the stretto precedes the pedal-point, sometimes it follows it. An episode introduced after the last close-imitation, or after the stretto, is called a *coda* to the fugue. It often is introduced on a tonic pedal, or worked into a prolonged plagal cadence.

A few general remarks are necessary in conclusion. The whole structure of a fugue points out that it is a work intended to be of constantly increasing interest, from the first exposition of subject and answer to the final bar. Out of this fact grow all the common rules for its formation, such, for instance, as, "no perfect cadence shall be heard till the end." A perfect cadence in *any* key gives a certain feeling of repose, and this feeling is alien to the spirit of the work. The only exception to this rule is when a fugue, with more than one subject, is broadly divided into two parts, as in Schumann's fugue No. 6 on the name Bach. *Half-closes* are not uncommon under similar circumstances, as in "Egypt was glad" (Handel).

Again "contrapuntal devices should be introduced in the order of their interest or ingenuity, beginning with the simplest, and the most complicated being introduced last."

Enough modulation should be introduced into a fugue to make it pleasing, and to avoid the tame effect of one continuous key-tonality; but, on the other hand, too much modulation would lead the hearers to believe that the work was intended to be made interesting as a *specimen of modulation*, and so take away their attention from the treatment of subject and answer. Hence, fugue-modulation is in a general way limited to *related keys*. The same object is kept in view by the rule, "if there is a tonic pedal-point, it should never be heard *before* the dominant pedal-point." Of course a breach of this rule would entirely undo the wonderful effect which the massive imitations or stretto have, when heard over the dominant, for dominant harmony always causes a yearning for tonic-harmony, and when the tonic is at length reached, then it is time to add yet more to the delay by multiplying superposed tonic harmonies.

A glance at the subjects of fugues given in the examples will show that there is not much room for originality left to modern writers. The more the vast literature of fugues which has come down to us is studied, the more apparent does this fact become. It is indeed almost impossible to write a short diatonic subject, capable of easy handling, which shall be in any sense *original*. The true lesson to learn from this is, that the modern *treatment* of fugue subjects should at least be original, and the composer who now sets about writing a fugue, should feel himself compelled, as an artist, to make use of all the freshness and novelty which modern chords, key-relationship, and rhythm are capable of producing.

It may be objected, that such a modernized fugue ceases to be a fugue at all. But the history of fugue clears away such objections. Starting from the early time when fugue had barely commenced a separate existence from counterpoint, the word fugue meant nothing more than the *subject*, hence *fuga composita*, or *fuga recta* was, when the subject moved about by single degrees, or in conjunct motion; and *fuga incomposita* was when the subject had skips in it, or proceeded by disjunct motion. Again, when the subject went upwards from the tonic it was called *fuga authentica;* when it went downwards from the tonic it was called *fuga plagalis*. Such expressions point out a very elementary stage in the art of fugue. What would now be almost distressing to us, namely, a fugue without any episodes, one in which subject and answer never cease to be heard, was at one period considered the perfection of a fugue. It was called *fuga ricercata*. Again, it is easy to trace the gradual introduction of episodes, and modulation, and the discarding of the complicated laws which bound subjects and answers to the tonality of the ancient church modes.

Then, again, an extension both of the *compass* and *length* of subject gave new scope to composers, while "licences" in counterpoint became of more frequent occurrence. In short, the fugue has gradually developed from an unartistic music-puzzle into a noble and splendid form, and it behoves modern composers to add their special share to its possible future development. It is quite true that a very large number of fugues, more or less in the old style, are at this time issued by so called scientific composers, and are considered clever, and favourably received by those who are not familiar with any music but that of the 19th century; but, were it worth the labour, such modern-antique fugues, could be proved to be mere rescripts and collections of what has been written long ago, not only once or twice, but scores of times. Having carefully examined the various periods in the life and history of fugue, and having accustomed himself to treat with respect the rules which fence in its earlier rudimentary forms, the student who reads aright will unhesitatingly endeavour to make fugue-form the handmaid of modern music, and so avoid the too common error of wilfully casting aside all that accumulation of experience and progressive improvement, which we happily possess, and should learn how to use.

Fugue renversée (*Fr.*) An inverted fugue.

Führer (*Ger.*) (1) Subject of a fugue. (2) A leader, director.

Full anthem. An anthem in which there is neither solo nor verses. [Anthem.]

Full cadence. A perfect cadence. [Cadence.]

Full chord. (1) A chord, some of the essential notes of which are doubled. (2) A chord for the full power of an instrument, orchestra, or voices.

Full score. A score in which all the parts for voices and instruments are displayed. [Score.]

Full service. (1) A setting of the Canticles for voices in chorus, with or without organ accompaniment. (2) An Office in which music is used to the fullest extent allowed by the rubrics.

Füllstimmen (*Ger.*) Additional chorus parts—*remplissage* (*Fr.*), *ripieni* (*It.*)—either of voices or instruments.

Full stop. (1) In lute playing, a full chord followed by a pause. (2) A chord in which all available fingers are occupied in stopping the strings.

Fundamental bass. [Harmony.]

Fundamental tones. The tones from which harmonics are generated. [Acoustics, § 10.]

Funèbre (*Fr.*) } Funereal, mournful, in the
Funerale (*It.*) } style of a dirge; as, *marche funèbre*, a funeral march.

Fünffach (*Ger.*) Five-fold. Often applied to a mixture stop of an organ--*having five ranks*.

Digitized by Google

Fünfstimmig (*Ger.*) In five parts.

Funzioni (*It.*) Functions, duty. The general title for services, oratorios, and other musical compositions performed in the Roman church.

Fuoco, con; fuocoso (*It.*) With fire, spirit, dash.

Furia, con; furibondo, furiosamente, furioso (*It.*) With fury, energy, vehemence.

Furlano (*It.*) A dance. [Forlana.]

Furniture. The name of one of the mixture stops in an organ.

Furore, con (*It.*) With fury, passion, enthusiasm.

Fusa (*Lat.*) A quaver, ♪

Fusée (*Fr.*) Rapid division, shake. or roulade.

Fusella (*Lat.*) A semiquaver, 𝅘𝅥𝅯

Fuss (*Ger.*) Foot. (1) The part of an organ pipe below the mouth. (2) The measure by which the pitch of organ stops is determined; as, 8 *füssig*, of 8 ft., or unison pitch.

Fz. Abbreviation for forzando.

Digitized by Google

G.

G. (1) The note Lichanos in Greek music. The third, or fore-finger string of the lyre. [Greek Music.]

(2) The first note of the church mode called Mixo-Lydian. [Plain-Song.]

(3) The lowest note of the grave hexachord; in the Guidonian system, *gamma ut*.

(4) The fifth note of the normal scale of C, called Sol.

(5) The lowest or fourth string of a violin, the third of the viola and violoncello.

(6) The key-note of the major scale, having one sharp in the signature.

(7) The letter-name of the treble clef.

G. *abb.* for *gauche* (*Fr.*) *Left;* as, *m.g.*, with the left hand.

Ga. The fourth syllable in the system of Bobibation.

Gabel (*Ger.*) A fork; *Stimmgabel*, a tuning-fork; *Gabelton*, the note A, as given for the pitch.

Gagliarda (*It.*) [Galliard.]

Gai (*Fr.*), **Gajo** (*It.*) Lively, merry, gay.

Gaillarde (*Fr.*) [Galliard.]

Gaiment (*Fr.*), **Gajamente** (*It.*) Gaily, cheerily, merrily.

Galantemente (*It.*) Gracefully, in good taste, bravely.

Galliard, Gaillarde (*Fr.*), **Gagliarda** (*It.*) An ancient dance, so called because of its *gay* rhythm and motion. It is said by some to have been similar in character to the Cushion dance, and is described by Sir John Davis as:

> —— "A swift and wandering dance,
> With passages uncertain to and fro,
>
> . . . . . .
>
> With lofty turns and caprioles in the air,
> Which to the lusty tunes accordeth fair."

Like the minuet, of which it was probably the parent, the galliard was danced by a lady and gentleman. If more than one couple performed the dance, they did so independently of other dancers.

The tune was generally written in triple time. "Hence," says Butler ("Principles of Musick," 1636), "the triple is oft called galliard time; and the duple, pavan time."

Brawls, corantos, and galliards were danced at court from the reign of Queen Elizabeth to that of Charles I., as country dances and minuets were in later time. Dowland's beautiful and well-known melody, "Now, O now," published, with words, in the "First Booke of Songes or Ayres, of foure parts," 1597, had been known before that date as a dance tune, under the name of the "Frog's Galliard." It is usually written in $\frac{6}{4}$ time; but as it is of slow pace, the subsidiary accent might be made a primary one, and so bring it within the general character of the measure of the galliard.

The composers of the early part of the 17th century frequently employed the rhythm of the galliard as a vehicle for "fancies," with florid passages for the virginals. A good example of this form of writing may be seen in "Parthenia, or the Maydenhead of the first musick that ever was printed for the Virginalls," 1611. The following tune, by Girolamo Frescobaldi, 1637, will show the measure of the dance:

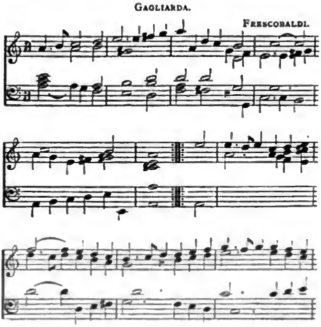

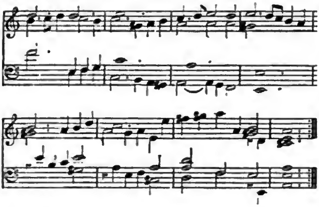

Sir John Hawkins says that "the tune for the galliard consists of five paces or bars in the first strain, and is therefore called *Cinque pace*," but the existing galliards do not justify this description.

Some writers say that the dance came in fashion about the year 1540. It had a reign of popularity extending over a hundred years, after which time composers ceased to employ the title. [Country Dance.] [Pavan.]

Galop, galopade (*Fr.*) A lively dance in 2-4 time, originally a separate and independent dance, but now also forming a portion of a set of quadrilles.

Galoubet (*Fr.*) A small flute of a primitive character with three holes, similar to the Picco pipe.

Gamba, Viola da (*It.*) A stringed instrument of the viol sort, with six strings, weaker in tone and smaller in size than the violoncello, so called because it was held between the knees of the player. [Viola da Gamba.]

Gamba or Viola da Gamba. An organ-stop, the pipes of which are, in continental organs, generally cylindrical, of small scale, and well cut up; but sometimes are conical in shape. Its tone is pungent and not unlike that of a violon or violoncello. In England the Bell Gamba is more commonly met with. [Bell Gamba.]

Gamme (*Fr.*) Gamut. [Notation.]

Ganz (*Ger.*) (1) Entire, whole; *Ganzer Ton*, a whole tone; *Ganze Note*, a semibreve, &c. (2) Very, as *ganz langsam*, very slow.

Garbo, con (*It.*) With grace, politely.

Garlands. A general name for collections of ballads, and other inferior literature upon given subjects. [Ballad.]

Garrire (*It.*) To warble, to chirp, to chatter.

Gassatio. A word of varied meaning. Some writers use it to describe a street serenade, "hergeleitet von dem Herumspazieren auf den Gassen um den Jungfern ein Ständichen oder Hoferecht zu machen;" others say it is a "familiar expression for instrumental compositions generally, symphonies as well as quartetts." Others assuming that the word comes from the Italian *Cassatio* or *Cassazione*, describe it as meaning a farewell or final piece, whether in a programme or as part of a whole composition. The word *Gassaten* or *Cassatio* is often made to do duty in describing the whole thing of which it only forms part, and thus it is that Suites or Sonatas of the earlier composers are sometimes called *Cassazioni* when final movements ought only to be so called. [Suite.] [Serenata.]

Gassenhauer. (*Ger.*) The name given to one of the dance tunes in Wolf. Heckel's "Lautten Buch," Strasbourg, 1562, which contains many songs and pieces; "Auch vil faltige Newe Tentz, sampt mancherley Fantaseyen, Recercari, Pavana, Saltarelli, und Gassenhawer, &c." The word is collaterally related to *Passacaille*, which is from the Spanish *Passa calle* ("*qui court les rues*"). The subjoined is a translation into modern notation, from the Tablature in the book referred to above, of the melody of this dance, which it may be perceived is capable of bearing a ground bass, like the *chaconnes* or *passacailles*, to which class it unquestionably belongs.

Wolf. Heckel's "Lautten Buch," 1562.

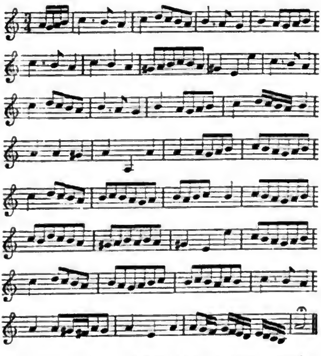

Gauche (*Fr.*) Left; as, *la main gauche*, the left hand.

Gavot, gavotte, or gavote (*Fr.*) *Gavotta* (*It.*) A dance tune of a lively yet dignified character, said to be of French origin, and to take its name from the Gavots, "Peuples montagnards du pays de Gap, ont donné le nom à cette danse que nous appellons gavote.' The description of the dance, "a brisk round for as many as will," identifies it with the country dance, and the form of the tune sup

Digitized by Google

ports this resemblance. The gavot seems to have been more popular as an instrumental piece than as a dance, and to have been a favourite movement in suites, lessons, and sonatas from the latter part of the 17th century, the time when the word appears to have been brought into use.

The descriptions of the measure and rhythm of the dance are many, and slightly different, one writer maintaining that it should begin "with two crotchets, or the half of a bar, with a rise of the hand in beating, ending also with two crotchets that begin the last bar." Another says, "It may begin with an odd quaver, as that in the 9th of Corelli's concerto does; or with a whole bar, as the same composer shows us in Sonata 1, Op. 2."

Hawkins says that the dance is in triple time, of two strains of four and eight bars respectively, the first ending in the key of the dominant; and quotes Walther, who states that the first strain should have its cadence in the third or fifth of the key; "for that if it be in the key-note itself, it is not a gavot, but a rondeau." It would be easy to produce numerous examples of gavots by well-known composers, in which the conditions mentioned above are not present. The following examples, selected originally with the intention of showing some early specimens of this dance, will also be interesting as bearing upon the question:

GAVOT IN GAMUT, BY DR. JOHN BLOW, 1700.

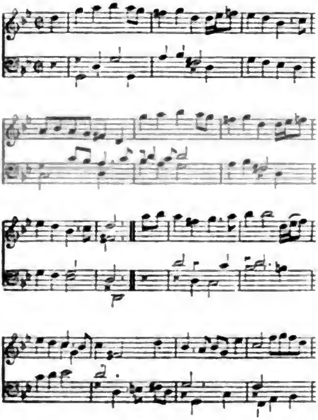

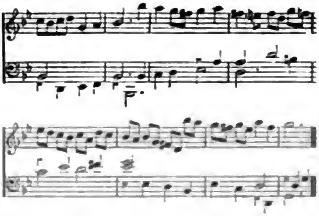

JEAN PHILIPPE RAMEAU, 1716.

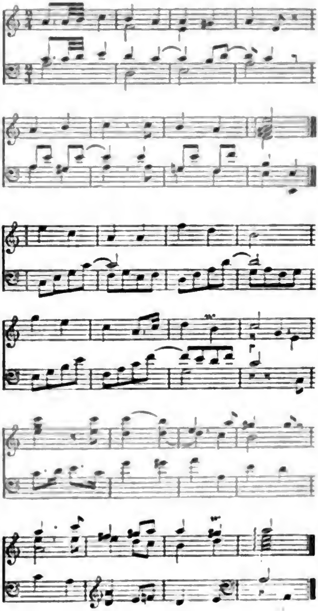

FRANÇOIS COUPERIN, b. 1668, d. 1733.

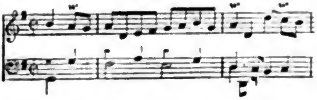

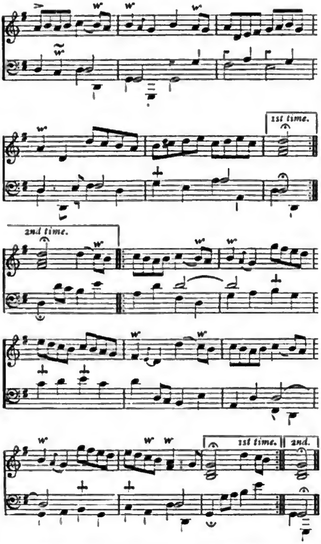

Corelli, Bach, Handel, and others who have written gavots, do not always adhere to the so-called rules of this form of composition, but display some remarkable deviations from it, which those interested in the subject may discover for themselves in the works of those writers. Like the galliard, the gavot, as an instrumental composition, had a limited period of popularity, for there are very few examples of later date than about the year 1760 to be found in the sonatas and suites. As a dance, the gavot was taught until a few years back, but the tune employed was different to those found in the compositions of the last century. Many of the old gavots are being restored to favour at the present time, and the composers of school-music are exercising their imitative powers in writing pieces after the manner of the old composers, to supply the demand made in consequence of the revived popularity of the melody and rhythm of this form of dance.

Gavotta (*It.*) *Gavotte* (*Fr.*) [Gavot.]

G clef. The treble clef, a character placed at the beginning of a stave to indicate the pitch of the notes. [Clef.]

G dur (*Ger.*) The key of G major.

Gebrochene Akkorde (*Ger.*) Distributed harmony, or arpeggio. [Arpeggio.]

Gedackt or **Gedeckt** (*Ger.*) Covered with a lid; closed. [Decke, § 2.]

Gefährte (*Ger.*) The answer to a fugue subject (Führer). [Fugue.]

Gefühl, mit (*Ger.*) With feeling, expression.

Gegenbewegung (*Ger.*) Contrary motion.

Gegengesang (*Ger.*) Antiphonal music.

Gegenpunkt (*Ger.*) Counterpoint.

Gegensatz (*Ger.*) Counter-subject.

Gehalten (*Ger.*) Held, *sostenuto*.

Gehend (*Ger.*) *Andante* (*It.*) *Lit.*, Going; at a moderate pace; *etwas gehend*, andantino.

Geige (*Ger.*) A violin. *Geigenblatt*, the finger-board; *Geigenbogen*, the bow; *Geigenharz*, resin; *Geigensaite*, fiddle-string; *Geigensattel*, bridge; *Geigenwirbel*, peg.

Geist (*Ger.*) Spirit, genius, soul.

Gelassen (*Ger.*) Calm, tranquil.

Gemessen (*Ger.*) Measured, *i.e.* not too fast.

Gemshorn (*Ger.*) (1) An instrument made of the horn of the chamois goat. (2) An organ stop, of conically-shaped pipes of tin, narrow at the open end, with ears at the broad end or mouth, to regulate the tuning. The tone is peculiar and pleasant. It is generally of 8-ft. tone, though sometimes of 4, and in the pedal organ of 16.

Generalbass (*Ger.*) Thorough bass; *basso continuo* (*It.*)

Generator. A ground note, fundamental bass, root, derivative.

Genere (*It.*) *Genre* (*Fr.*) (1) Manner or style. (2) Kind or class (of scales); as, *diatonico*, *cromatico*, *enarmonico*.

Generoso (*It.*) Nobly, with dignity.

Gentile; gentilezza, con (*It.*) Noble, with dignity.

Genus (*Lat.*) Sort or class, especially used with reference to scales; as, the *diatonic*, *chromatic*, and *enharmonic* genera. [Greek Music.]

Gerade Bewegung (*Ger.*) Similar motion. [Motion.]

Gerade Taktart (*Ger.*) Common time.

German flute. [Flute.]

German sixth. [Extreme sixth, chord of.]

Ges (*Ger.*) The note G flat.

Gesang (*Ger.*) Singing, song, cantata, hymn, &c.

Geschwind (*Ger.*) Quick, rapid.

Ges dur (*Ger.*) The key of G flat major.

Gestossen (*Ger.*) Staccato.

Getern, Getron (*Old Eng.*) Guitar.

Getragen (*Ger.*) *Lit.*, drawn, *legato*.

Ghazel (*Arab.*) A term used by Dr. Hiller to describe a piece in which a simple theme is constantly recurring. The name is suggested by those Eastern poems in which a word or sentence either forms the ending or

Digitized by Google

commencement of the lines. The following is a short "ghazel," written by Dr. Hiller on the theme, G, A, B:

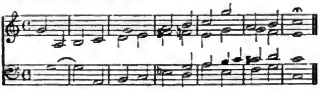

Ghiribizzi (*It.*) Fantastic devices.

Giga (*It.*), **Gigue** (*Fr.*) Jig.

Gigelira (*It.*) Giga vel lira. A name given to the Strohfiedel. [Ligneum Psalterium.]

Gingras. A small ancient flute, of Phœnician origin, afterwards adopted by some European nations.

Gingrina (*Lat.*) [Gingras.]

Giochevole (*It.*) Merry, jocose.

Giocondamente (*It.*) Joyfully, merrily.

Giocondezza (*It.*) Mirth, jocundity.

Giocondato (*It.*) Happy, joyful.

Giocondo (*It.*) Jocund.

Giocosamente, giocoso (*It.*) Sportively, playfully.

Giojante, giojosamente, giojoso (*It.*) With mirth, joyfully.

Gioviale (*It.*) Jovial, pleasant.

Giovialità, con (*It.*) With jollity.

Giraffe. An ancient form of the spinnet. [Pianoforte.]

Gis (*Ger.*) The note G sharp.

Gis moll (*Ger.*) The key of G sharp minor.

Gittern. [Guitar.]

Gittith (*Heb.*) This word, which is found in the titles of Ps. viii., lxxxi., lxxxiv., is by some supposed to signify a musical instrument (perhaps as used at Gath); by others, a vintage-song, or well-known tune, to which the Psalm could be sung. Various other explanations have been offered, which it is unnecessary to give here.

Giubiloso (*It.*) Jubilant.

Giustamente (*It.*) Strictly, accurately.

Giusto (*It.*) Strict, correct, moderate; *a tempo giusto*, at a moderate pace.

Glass. Musical instruments of this material are of two kinds, percussion and friction; the first consists of a series of small plates of graduated sizes, supported on tapes secured in a wooden box, the several tones being regulated by the size of the glass: this is a mere toy. For a description of the best of the second class see Armonica. Another form of a glass friction instrument is made of a number of tubes of various lengths, and as the tone is brought out by stroking the length of the several tubes with flannel or india-rubber, it is only capable of producing slow melodies.

Glee. A composition for voices in harmony, consisting of two or more contrasted movements, with the parts so contrived that they may be termed a series of interwoven melodies. It may be written for three or more voices, either equal or mixed; but it is necessary that there should be only one voice to a part. It may be designed with or without instrumental accompaniment, and set to words in any style—amatory, bacchanalian, pastoral, didactic, comic, or serious. As a composition, the glee appears to have historically followed the catch, and to have had its origin at the time when part-singing began to be revived. But when musical skill was at a very low ebb, a satisfactory performance of existing vocal compositions for combined voices was neither possible nor desirable: not possible; because the madrigal, to be effective, required many voices to a part; and not desirable, because the words set to the catches, the other sort of secular part-music, were not of a character which fitted them for the ears of decent folk. The earliest glees, so called, were set to words of a pastoral character. One of the first, if not the very first, printed composition for voices to which the title was attached, was "Turn, Amaryllis, to thy swain," by Thomas Brewer, included in the second book of Hilton's "Catch that catch can," 1652. The most ancient collection in which glees are specially mentioned, was published by Playford. It is called, "The Musical Companion, in two books: the First Book containing Catches and Rounds for Thiee Voyces; the Second Book containing Dialogues, Glees, Ayres, and Songs for two, three, and Four Voyces," 1673. The compositions contained in these books can only be regarded as exhibiting the qualities of preliminary attempts to fix and form the style, which afterwards became known as the "glee style." Many other species of musical works have grown to their present proportions by slow degrees; but the glee seems to have started into existence in its modern form all at once, and not to have been the result of a series of developments. From the time when Playford's book was published until between the year 1760 and 1770, the specimens of part-writing to which the authors attached the word "glee," are somewhat rare, the terms "ode," or "three, four, or five-part song," being preferred for vocal compositions in harmony.

Sir John Hawkins does not mention the word once in his "History of Music," published 1776, although institutions for the encouragement of glee-writing were already established in his time.

The period of the existence of the glee, as we now understand it, was about seventy years, namely, from 1760 to 1830; the most successful of the glee-writers during that time were S. Webbe, Dr. Cooke, Dr. Callcott, R. J. S. Stevens, Reginald Spofforth, J. Stafford Smith, W. Horsley, Sir Henry Bishop, Charles Evans, and to this list must be added, Sir John Goss.

The compositions of these writers, with a few by their contemporaries, form the whole literature of this class. The so-called German glees are, for the most part, simply harmonised melodies, and belong to the order of part-songs rather than to that of glees. The application of the term to this class of composition is correct philologically, but not formally. The old word *glee* meant harmony or combination; and, therefore, all compositions for voices in harmony may be rightly designated by the term. But the word is understood to signify a special sort of vocal harmony, and if the pieces so called do not fulfil the conditions of the character, already described above, they ought not to be called by the term.

The glee, like the anthem, is of English growth, and has never been successfully imitated by foreign writers. The increase of musical taste has led to the formation of large choral societies, by whom the master-works of the great composers are given with effect; but it has also led to the neglect of private social musical gatherings, and, consequently, to the disuse of one of the most delightful musical pleasures, the performance of the glee. Glee-singing is almost a lost art in England. The tradition has not been properly maintained, and we are in the somewhat anomalous position of a people in the possession of a special literature, which we cannot rightly interpret or appreciate.

A few remarks upon the origin and meaning of the word Glee may not be considered uninteresting or out of place here. The word comes from the Anglo-Saxon "gle," meaning music, or the performance of music. For example, the "Story of Genesis," written about 1250, and reprinted by the Early English Text Society, has the following words:

> "Jobal is broðer song and glew,
> Wit of music well he knew."

Chaucer, in his "Troilus and Creseide," uses the word with a like meaning:

> "For though that the best harpair upon live
> Would on the bestè sounid jolly harpe
> That evir was with all his fingers five
> Touch aie o string, or aie o warble harp
> Were his nailis poincted nevir so sharpe,
> It shuldè makin every wight to dull
> To here his gle and of his strokis full."

In the "Promptorium Parvulorum" (1440), the same word, spelt *glu*, probably in accordance with the provincial pronunciation of the writer, is translated *armonia*, *minstrelsy*. Some modern writers suppose the word to come from gligg, the Anglo-Saxon term for joy or merriment; or from *gleek*, which signifies to scoff, sneer, or banter. Neither of these derivations point to the musical use of the word; for the majority of glees are of a character too serious to be called merry, and too earnest to be called bantering.

The early writers of glees frequently used a qualifying term with the word, as "serious glee," "chearfull glee," &c., a practice which might be considered superfluous if the word only meant merriment. It may be, therefore, gathered, that they attached a meaning to the word similar to that found in the writings of the early poets and others, namely, combination. That glee meant consort or harmony is implied in the following extract from a poem by Robert Manning, of Lincolnshire, in the reign of Edward I., c. 1303:

> "Yn harpe and tabour and symphan gle
> Worship God in trumpes ant sautre;
> Yn cordes, yn organes, and bells ringying,
> Yn all these worship the hevene Kyng."

and in Davie's poem, the Life of Alexander. [temp. Ed. II.]

> "Orgues, chymbes, uche maner gle • • • • • • •
> Withouten the toums murey;"
> (Organs, chimes, all manner of harmony
> Outside the town's wall).

[Catch.] [Madrigal.] [Part-song.]

Gleek. [Glee.]

Gli (*It.*) The; as *gli stromenti*, the instruments.

Glissando, glissato, glissicato, glissicando (*It.*) (1) Playing a rapid passage in pianoforte music, by sliding the tips of the fingers along the keys instead of striking each note with a separate finger. (2) A rapid slur in violin playing.

Glisser (*Fr.*) To slide. [Glissando.]

Glockenspiel (*Ger.*) An instrument made of bells tuned diatonically and struck with hammers, or by levers acted upon by a keyboard. It is occasionally employed in the orchestra, notably by Mozart in his opera, "The Magic Flute." [Bells.] [Carillon.] (2) An organ stop of two ranks.

Gloria (*Lat.*) [Mass.]

Glottis. [Larynx.]

Glottis (*Gk.*, *γλωττις*); *Lat.*, *Lingula.* The reed used in some of the ancient flutes. These reeds were moveable, and were carried about in a little box called *γλωσσοκομεῖον*.

G moll (*Ger.*) The key of G minor.

Gnaccare (*It.*) [Castanets.]

Gong. (1) A pulsatile metal instrument, generally circular and shallow, formed of copper and tin mixed in the proportions 80 : 20. (2) A rod of metal bent into a plane spiral, used instead of a bell. (3) Flat bars of metal placed over resonators. (4) Bell-

Digitized by Google

shaped metal vessels with the mouth partially closed.

Gong Drum. A bass drum shaped like a tambourine.

Gorgheggi, Gorgheggiare (*It.*) Trills, quaverings, warblings.

Gosba. An Arabian flute. There are two sorts of the gosba, the one with three holes in the lower extremity, producing four sounds which with their harmonics at the fifth complete the octave. The instrument is employed to guide the voice of a singer. The other gosba is larger and pierced with six holes, with a double hole at the back.

Grace notes. [Graces.]

Graces. A general term for ornamental notes or short passages, introduced as embellishments into vocal or instrumental music, not actually essential to its harmony or melody. In former times, in vocal music, the selection of graces was left to the judgment of the performer to a great extent, but in instrumental music numerous signs have from time to time been used, explanations of which will be found under their distinctive names. Harpsichord and lute music was always lavishly ornamented, and in lesson books for these instruments, much care and space is often given to a full explanation of their force and meaning. [Harpsichord.] Music for viols was also *graced* in various ways, but never to so great an extent as that above named. As all these instruments are now obsolete it is unnecessary to enter further into the subject. In our own time a reaction has taken place against the absurd embellishments indulged in by our forefathers, and it has become fashionable to sing and play music just as it is written. This is perhaps to be regretted, as those who are rendering music should carefully consider whether the writer wished ornaments to be excluded, or, omitted to write them under a belief that they would certainly be introduced in performance. [See Cadenza.] [Accompaniment.]

Gracieux (*Fr.*) Graceful; in graceful style.

Gracile (*It.*) Small, thin; as, *voce gracile*, a thin voice.

Grad (*Ger.*) A degree or step of a scale.

Gradation (*Fr.*) *Gradazione* (*It.*) Gradation, by degrees of the scale.

Gradevole, gradevolmente (*It.*) Grateful, gratefully.

Graditissimo (*It.*) Most grateful.

Gradleiter (*Ger.*) A scale.

Grado (*It.*) Degree or step of a scale; as, *di grado*, by conjunct motion, as opposed to *di salto*, by a skip.

Graduale (*Lat.*) (1) A gradual. A piece of music performed between the reading of the Epistle and Gospel in the Roman Church. (2) The book containing the music of the Introit, Gradual, Tract, Sequence, &c., for every day in the year, and all festivals. It is essentially for the use of the choir; the Missal contains the music for the celebrant.

Gran cassa (*It.*) *Grosse caisse* (*Fr.*) The big drum.

Grand (*Fr.*) *Grande* (*It.*) Large, great, complete; as, *grand bourdon*, a double bourdon; *à grande orchestre*, for a complete band.

Grand barré (*Fr.*) A position in guitar playing, the object being to alter the pitch of the instrument by making a temporary nut of the forefinger laid lengthwise across the strings. [Guitar.]

Grande mesure à deux temps (*Fr.*) Common measure of two beats.

Grandezza (*It.*) Grandeur.

Grandioso (*It.*) Grand, in a lofty manner.

Grandisonante (*It.*) Loud, sonorous.

Grand jeu (*Fr.*) The power obtained by the use of the whole of the stops in an organ, or by the employment of a stop so called in the harmonium which calls into use the whole of the available registers.

Gran gusto (*It.*) Elevated taste or expression.

Grande orgue (*Fr.*) (1) Full organ. (2) The great organ.

Grand pianoforte. [Pianoforte.]

Gran tamburo (*It.*) The big drum.

Grappa (*It.*) *Lit.*, a stem; a brace which connects staves.

Grave (*Lat.*, *It.*, *Fr.*, *Eng.*) (1) Deep in pitch; as, *grave hexachord*, the lowest hexachord in the Guidonian system. (2) Slow in pace, solemnly.

Gravecembalum (*Lat.*), **Gravicembalo** (*It.*) [Harpsichord.]

Gravement (*Fr.*) Slow, and in a solemn style.

Gravità, con (*It.*) With dignity, weight, majesty.

Grazia, con (*It.*) With grace, elegance.

Graziosamente, grazioso (*It.*) Gracefully, elegantly.

Greater. Belonging to the major scale; as, *a greater third*, a major third, as C to E; *greater sixth*, a major sixth, as C to A. A piece of music, said by the old writers to be in any key with the *greater third*, was in the major mode: with the *lesser third*, in the minor mode.

Great octave. The sounds lying between

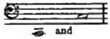

represented, according to one system, by single capitals, C, D, E, &c.; in another, by double capitals, as CC, DD, EE, &c. [Pitch.]

Great organ. [Organ.]

Greek Music (Systems of ancient).[*] From the time of Homer down to that of Terpander, who seems to have flourished some 300 years after Homer, the lyres of the Greeks had but four strings. At that early date the instrument could only have been used for the purposes of a pitch pipe, just as orators subsequently employed it to regulate the pitch of the voice. No tune could be drawn from four notes.

Terpander raised the number of strings from four to seven, for the service of the Gods. The following two lines, from one of his hymns, are preserved in the *Introductio Harmonica*, ascribed to, but evidently not written by, Euclid.[†]

"*Ἡμεῖς τοὶ τετράγηρυν ἀποστέρξαντες ἀοιδὴν,*
Ἑπτατόνῳ φόρμιγγι νέους κελαδήσομεν ὕμνους."[‡]
—(p. 19, edit. Meibom.)

This scale of seven notes was formed by connecting the first tetrachord, or series of four notes, with a second series of four, by one sound common to both. To represent these sounds in modern notes, they would be as E, F, G, A, and A, B flat, C D united by the A in the middle, which was the key note to the two. The Greeks had the same number of perfect fourths in a scale that we have, but when they formed their scales by tetrachords, or fourths, they selected that position of the fourth, in which the semitone came between the lowest two strings—as E, F, G, A. The ORIGINAL SEVEN-STRINGED SYSTEM OF THE LYRE was then as follows:—

Upper Tetrachord.	*d.* NETE (shortest string, giving the highest sound).
	c. PARANETE (beside Nete).
	b flat. PARAMESE (beside Mese).
	a. MESE (middle string and key note, connecting the two fourths).
Lower Tetrachord.	*G.* LICHANOS (forefinger string).
	F. PARHYPATE (beside Hypate).
	E. HYPATE (longest string, giving the lowest sound).

The above are names of the strings of the lyres, and not of notes of a fixed pitch. The same names would have been retained if the lyre had been tuned one, two, or three notes higher. The longest string was called Hypate, although it gave the lowest sound. If pitch had counted for height instead of mere length of string, the order of Nete and Hypate would have been reversed.

The lower four strings of the lyre were played by the thumb and three fingers of the left hand, the string that fell under the forefinger being called *lichanos* (the licking-up finger), and the thumb upon the key note or *Mese*. The three treble strings were played upon by a plectrum, which was a piece of ivory, ebony, horn, or any hard wood. This was held in the right hand, and its use being only occasional, the right hand was in a measure left free for action in addressing the auditors.

The next improvement in Greek music is connected with the most important of all dates in Grecian History—that at which Egypt was thrown open to the Greeks by Psammetichus the First, King of Egypt. From that event sprang the rapid advances of the Greeks in science, in art, and in literature. Philosophers, law-givers, historians, astronomers, mathematicians, musicians, architects, physicians, and alchemists—indeed all who were intent upon the acquisition of learning—sought it in that world of ancient civilization. It was there that Thales learnt to measure the height of a temple or of a pyramid by the length of its shadow—there to divide the year into 365 days. It was there that one of the philosophical re-discoveries of the last and of the present century, viz., that sounds may be both too high and too low to reach the human ear, was known thousands of years ago.

Until the reign of Psammetichus the Greeks had been going on a wrong road to music. The seven strings could produce nothing worthy of the name of tune with such a scale as they had; at least so long as the middle string remained the key note of that scale. All the ancient fables of Orpheus and Amphion must rest upon their skill in poetical recitation, which was one branch of music in the Greek sense. As to Amphion, he, no doubt, sang in such lively rhythm as to expedite the builders in order to keep time to it, and hence the fable of his having raised the walls of Thebes by his lyre.

Psammetichus I. began his reign in 664, B.C.[§] He was the first of the Pharaohs who

---

[*] "The Systems of Ancient Greek Music compared with Modern Music," abbreviated from Chappell's History of Music.

[†] Two treatises on music are ascribed to Euclid, the *Introductio Harmonica* and the *Sectio Canonis*. The second is a mathematical treatise quoted by Porphyrius as Euclid's, but the first is an Aristoxenian or practical musician's treatise in a different school. (It is none the less valuable, whoever may have been its author.) With this reservation both will hereafter be quoted as Euclid's, to abbreviate references. Proclus says only that Euclid wrote on the elements of music. (Κατὰ μουσικὴν στοιχειώσεις.)

[‡] But we, loving no more the four-toned song,
Will sing aloud new hymns to a seven-toned lyre.

[§] In Dr. W. Smith's *Dictionary of Greek and Roman Biography*, the date of Psammetichus I. is given as 671, B.C., and to this is added a note that Boeckh dates his reign as 654, B.C. As Egyptian dates can be carried back with tolerable certainty from the conquest of Egypt by Cambyses, and the above is rather vague for a matter of such importance (especially when early

Digitized by Google

cultivated the friendship of the Greeks; he invited them as settlers, and engaged Carian and Greek mercenaries in his army. It was the change in Egyptian policy that enabled Pythagoras to go to Egypt, where he is said to have lived 22 years. He is the reputed discoverer of the octave system of music, which was certainly known in Egypt at least a thousand years before his visit.

The popular myth of the Egyptian Hermes and the lyre, is that, when walking by the banks of the Nile, he accidentally kicked the shell of a dried tortoise, in which there was nothing remaining but dried sinews, and that it emitted musical sounds, and thus suggested to him the idea of forming it into a musical instrument. The Egyptian name of the God was Thoth. The instrument was the Egyptian *nefer*, in hieroglyphics *nfr* (without the vowels), so sometimes translated *nofra* or *nefru*. This musical instrument is found in hieroglyphics at as early a date as the building of the second pyramid. The meaning expressed by the hieroglyphic is "good."

The difference between the Greek lyre and the Egyptian lute was that the former had no neck, against which the strings could be pressed. A lyre with an open back could give but one sound from each string, but when the same string was pressed against a finger-board it would produce notes in every variety, according as the vibrating part of the string was shorter or longer. The first lesson to be learnt from it was that the half of a string would produce the sound of the octave above the whole length; next, that, by stopping one-third of it, the remaining two-thirds would sound the musical interval of a fifth above the whole length; and that by stopping a fourth part, the remaining three-fourths would sound the interval of a fourth above the whole. In this way the Greeks learnt to produce every note of a scale, as well as the relation between geometrical proportions and musical sounds.

At the time of this discovery Greek lyres had only been made, on the Terpander model, to carry seven strings, so that, on learning the octave system, which required eight, they were obliged to leave out one of the notes. After new lyres had been made to carry eight strings the entire octave was included upon the instrument. The old system of tuning the lyre was then called *Synaphe* or Conjunction (*συναφή*),* and the new, or octave, system was called Harmonia (*ἁρμονία*), the "fitting in" system, because it fitted in the lesser consonances of the fourth and fifth into the greater consonance of the octave. (Verb. *ἁρμόζειν*, and the participle, used as an adjective, *ἡρμοσμένος*.) When the principle was fully established, harmonia became a synonym for music, in our sense of the word, for the Greek word, *Mousikè*, embraced all the arts and sciences over which the Muses presided.

The old dissonant seventh (from E to d) was made into an octave (from E to e) by the interposition of a tone between the two, previously, conjoined fourths or tetrachords. This interposed tone was called the diazeuktic tone, or tone of disjunction (*τόνος διαζευκτικός*.)

The following is the scale for the seven, and for the eight stringed lyre upon THE EGYPTIAN OR OCTAVE SYSTEM. It is here printed in the Greek "common" musical scale—our A minor with a minor seventh:—

SEVEN STRINGED LYRE.

Upper Tetrachord.
e. Nete.
d. Paranete.
c. (*omitted.*)
b. Paramese or Trite.

Lower Tetrachord.
a Mese (key note).
G. Lichanos.
F. Parhypate.
E. Hypate

EIGHT STRINGED LYRE.

Upper Tetrachord.
e. Nete.
d. Paranete.
c. Trite.
b. Paramese.

The tone of Disjunction or Diazeuktic tone.

Lower Tetrachord.
a. Mese (key note).
g. Lichanos (or Diatonos).
F. Parhypate.
E. Hypate.

In the eight stringed lyre Paramese and Trite were no longer the same string. Paramese took its proper place, next to Mese, and Trite was third from the top, as its name indicates.

By this system the player had a fifth upwards from his key note and a fourth below it, so as to allow scope for recitation both

---

authorities differ also by ten years in the length of his reign), the assistance of Samuel Birch, Esq., LL.D., F.S.A., keeper of the Antiquities in the British Museum, was sought for to decide between the discrepancies, and most kindly given in the following words:—"The highest monumental date known of Psammetichus I. is 54 years, according to the Apis tablets of the Serapeium, which agrees with the statement of Herodotus. The date of 664, B.C., is the lowest probable date of the accession of Psammetichus, which *might* be a year or two higher, and Boeckh's date is inadmissible."

* Bacchius, Senr., p. 20.

Digitized by Google

above and below. He could then produce something more like to a tune than was possible upon the old scale. Before that time any Greek chant would have sounded to modern ears as never ending, for their key note would be to us as the third of the key, because we have a major scale (which the Greeks had not), and we could associate such recitations with the key of F major through the B flat. The scales of the Greeks were all in minor keys, and the nearest approach they had to a major scale was one of five tones, which might be extracted from the chromatic scale, of which hereaftet.

The reason why Pythagoras preferred to omit C, which was a major third from the top on the seven stringed lyre, and a minor third above the key note (a), was because, at that time (but not after Didymus, Claudius Ptolemy, and other mathematicians had revised the scale) Greek thirds were not only esteemed as discords, but were so. They were so because, at first, the Greeks used only major tones in their scales, and there are less than six major tones in an octave. It is not proposed here to enter upon the mathematical divisions of scales, which, after the introduction of "equal temperament" (*alias*, equal putting out of tune), few, especially they who have their pianofortes tuned for them, seem to care about. But still a very easy experiment may be recommended to prove the case.

A major tone is the difference by which a fifth overlaps a fourth. Therefore, tune a perfect fourth from C down to G (or ask the tuner to do it), and then a perfect fifth up from G to D. There will be a major tone from C to D. Repeat the same, but beginning from D, a perfect fourth down to A, and a perfect fifth up to E. There will be another major tone from D to E, and the two will form an old Greek ditone, or third, from C to E. Try it by the ear, and it will be understood at once why the Greeks and the early writers upon church music (who had the worst of Greek divisions of a scale through the imperfect treatise of Boethius) for a long time treated thirds as discords.

Harmonia, which thus had primarily the meaning of "The Octave System of Music," came to signify "The Science of Music" and "Music" generally, because Pythagoras had limited the doctrines of the science to the sounds which are included in an octave,[*] so Harmonia and the later word Harmonica (*ἁρμονική*) had the same meaning. The Pythagorean writers on music were called Harmonici (*ἁρμονικοί*),[†] and some of them, before the time of Aristoxenus, had given such exclusive preference to the seven-stringed system[‡] coupling with it the enharmonic division of the octave, and calling this enharmonic branch of the system "harmonia," that the word was not infrequently used, instead of enharmonia, for a long time after. This, however, was not the original meaning, as the following extract from Philolaos, who first published the Pythagorean doctrines, will show. It refers to the seven-stringed octave, so *Trite* is B, not C, which it became on the eight-stringed lyre.

[§] *Ἁρμονίας δὲ μέγεθός ἐντι συλλαβὰ καὶ δι'ὀξειᾶν · τὸ δὲ δι'ὀξειᾶν μεῖζον τᾶς συλλαβᾶς ἐπογδόῳ · ἔστι γὰρ ἀπὸ ὑπάτας ἐς μέσαν συλλαβὰ, ἀπὸ δὲ μέσας ποτὶ νεάταν δι'ὀξειᾶν · ἀπὸ δὲ νεάτας ἐς τρίταν συλλαβὰ · ἀπὸ δὲ τρίτασ ἐς ὑπάταν δι'ὀξειᾶν · τὸ δ'ἐν μέσῳ μέσας καὶ τρίτασ ἐπόγδοον. Ἁ δὲ συλλαβὰ ἐπίτριτον · τὸ δὲ δι'ὀξειᾶν ἡμιόλιον · τὸ διὰ πασᾶν δὲ διπλόον. Οὕτως ἁρμονία πέντε ἐπόγδοα καὶ δύο διέσιες, δι'ὀξειᾶν δὲ τρί'ἐπόγδοα καὶ δίεσις, συλλαβὰ δὲ δύ'ἐπόγδοα καὶ δίεσις.*[||]

In the above extract the distinction between Harmonia (the octave system) and Diapason (the octave) is clearly drawn. The use of the word *diesis* for the interval which Aristoxenians, or practical musicians, called a semitone, or rather hemitone (*ἡμιτόνιον*), and Pythagoreans, or mathematical musicians, more accurately called a limma or remnant (*λεῖμμα*), proves that Philolaos refers to the diatonic scale, or scale of tones and semitones, and not to the enharmonic scale which had quarters of tones instead. When the enharmonic system came more into use the word "diesis" was transferred to its quarter tones, and to thirds of tones (when they were used) in one of the chromatic scales. The semitone was then no longer called diesis.

---

[*] *Πυθαγορασ δ'ὁ σεμνὸς . . . τῇ δ'ἀναλογικῇ ἁρμονίᾳ · αὔταρκές τ'ἐνόμιζε μέχρι τοῦ διὰ πασῶν στῆσαι τὴν τῆς μουσικῆς ἐπίγνωσιν.*—Plutarch, *De Musica*, cap. 37.

[†] Οἱ *καλούμενοι ἁρμονικοί*, says Aristoxenus contemptuously, p. 40 and again p. 37.

[‡] *ἀλλὰ περὶ αὐτῶν μόνον τῶν ἑπταχόρδων, ἃ ἐκάλουν ἁρμονίασ, τὴν ἐπίσκεψιν ἐποιοῦντο* (p. 27.)

[§] "The extent of the octave system is a fourth and a fifth; but the fifth is greater than the fourth by a tone of the proportion of 9 to 8; for [the interval] from the lowest [string E] to *Mese* [the key note A] is a fourth, but from *Mese* to *Nete* [the upper e] is a fifth; from *Nete* [down] to the third string is a fourth; from the third to the lowest is a fifth; between the key note and the third string is a tone of 9 to 8. The fourth is in the ratio of 4 to 3; the fifth in that of 3 to 2; and the octave [diapason] of 2 to 1. Thus the octave system, Harmonia [contains] five tones and two limmas [or semitones], the fifth [contains] three tones and a semitone, and the fourth two tones and a semitone.

[||] Philolaos, edit., Boeckh, p. 66, 8vo.; Berlin, 1819.

Mese, which means middle, had also the office of key note at the time when it only connected the two fourths in the old seven-stringed system. It retained the name in the latter sense, as the centre and turning point of the system, when the lyre had eight or ten strings, and consequently no middle; for "eight has no middle," says Aristotle, referring to it.[*] "Systems without mutation," (ἀμετάβολα) says Aristides Quintilianus (in other words, "Systems in one key") are those with one Mese: Mutable systems (μεταβαλλόμενα) have several Meses."[†] No lyre could have several middle strings, so he can only mean key note. An endless number of quotations might be given to the same effect. Mese was not only the key note to all Greek scales; it is to this day the key note of our minor scales, which we derived wholly through the Greeks.

Although the Greeks had now arrived at the only true system of music, yet their old one was not allowed to die away. There were, no doubt, ancient hymns to the Gods upon that system, and so they continued to use it. Terpander, who first added a second tetrachord to the lyre, was a hymnologist, and more than 200 years after him, Ion of Chios, another hymnologist, added a third conjoined tetrachord, and so increased the number of strings from seven to ten. The following extract from one of Ion's hymns is preserved by Euclid:—[‡]

—Τὴν δεκαβάμονα τάξιν ἔχουσα
Τὰς συμφωνούσας ἁρμονίας τριόδους.
Πρὶν μὲν σ'ἑπτάτονον ψάλλον διὰ τέσσαρα πάντες
Ἕλληνες, σπανίαν μοῦσαν ἀειράμενοι.[§]

Ion produced his first tragedy (according to Suidas), B.C. 453, and died B.C. 421. It is clear that he here refers to the old system, and not to the new, by naming only three harmonies or concords (συμφωνίας) from ten strings, and from their meeting or conjunction; for the middle tetrachord of the conjunct system was united by its extremes to the other two. This system of Ion's was called Episynaphe, or Conjunction upon Conjunction [ἐπισυναφή].[‖]

After the time of Ion, the original Greek scale received only one more string, the eleventh, which was added at the base to make an octave to *Mese*, thus borrowing from the octave system. It was a great improvement, for in this form it gave an octave of the Hypo-Dorian or common Greek scale (our A minor with a minor seventh) from A to a, and an octave of the Dorian scale (our D minor with a minor seventh) from D to d—the last through having the b flat in the upper part of the scale.

In this, its completed form, it became "The lesser perfect system" of the Greeks, until Claudius Ptolemy disputed the claim of such a scale to be called "perfect." The defect he saw in it was that it did not include the twelfth or fifteenth.

THE LESSER PERFECT SYSTEM.

The Synemmenōn Tetrachord (συνημμένων) or Conjunct Fourth.	d.	Nete	Synemmenon.
	c.	Paranete	id.
	b♭.	Trite	id.
The Mesōn Tetrachord (μέσων) or Middle Fourth.	a.	Mese	
	G.	Lichanos	Mesōn.
	F.	Parhypate	id.
	E.	Hypate	id.
The Hypatōn Tetrachord (ὑπατῶν) or Lowest Fourth (added by Ion).	D.	Lichanos	Hypatōn.
	C.	Parhypate	id.
	B.	Hypate	id.
The acquired note not included in any Tetrachord	A.	Proslambanomenos.	

The original seven strings had seven different names, but no new ones were given to those added by Ion, so it became necessary to distinguish between the new and the old by adding the name of the tetrachord, or fourth, to which they belonged. Thus the original Hypate (E) became Hypate-Meson (*i.e.* lowest of the middle tetrachord) and the new Hypate (B) became Hypate-Hypaton, or lowest of the lowest tetrachord. And now, to quit the lesser perfect system and revert to the greater, and more important one.

---

[*] Problems, 25 and 44 of Sect. 19.

[†] Arist. Quint., p. 17.

[‡] Euclid, *Int. Harm.*, p. 19.

[§] Having the ten-note scale
With three musical consonances conjoined
Till now with seven-stringed fourths all the Greeks hymned thee,
Upraising stinted song.

[‖] Bacchius, Senr., p. 21.

Many have written of Greek music without distinguishing between the two systems, and as one instance, we may name Dr. Burney. He mixes the two into one as "the great, the perfect, the immutable system . . . composed of *five* tetrachords" (p. 3), and then says: "after ascending regularly thus, up to D, by three conjoint tetrachords, the fourth in the great system is begun by *descending* a minor third to B natural . . . Something of this *dodging* kind is to be found in the scale of Guido" (note to p. 5, vol. 1). This is altogether a mistake, there is no "dodging." The d, to which he refers, is the highest note of the lesser system, which was perfectly distinct from the greater.

Another difficulty of Dr. Burney's and of other writers, has been to understand Greek octaves. It was very natural to suppose that a Greek octave scale would begin and end like one of two octaves, viz., upon the key note; but it was not so. The Greek octave scale took from the middle of the two-octave scale, and began a fourth below the key note and ended a fifth above it. In other words, when the octave scale was increased to two octaves, it was by the addition of a new tetrachord or fourth at each extreme, and then joining on at the base, the "acquired tone" to make an octave to the key note, Mese. So that, whether they had a one octave or a two octave lyre, the key note was in or near the middle, and a Greek could recite or sing at least a few notes above, as well as a few notes below it.

The following is the "disjunct," two octave system complete.

THE GREATER PERFECT SYSTEM.

(*συστήμα τελείον*)

The extreme, or Hyperboleōn. (*ὑπερβολαίων*) Tetrachord.	a. Nete Hyperboleōn.
	g. Paranete id.
	f. Trite id.
The disjunct, or Diezeugmenōn (*διεζευγμένων*) Tetrachord.	e. Nete
	d. Paranete Diezeugmenōn.
	c. Trite id.
	b. Paramese id.
The tone of disjunction, or Diazeuktic tone (*τόνος διαζευκτικός*)	
The middle, or Mesōn (*μέσων*) Tetrachord.	a. Mese (key note).
	G. Lichanos (or Diatonos) Mesōn.
	F. Parhypate id.
The lowest, or Hypatōn (*ὑπατῶν*) Tetrachord.	E. Hypate id.
	D. Lichanos or (Diatonos) Hypatōn.
	C. Parhypate id.
	B. Hypate id.
The acquired tone not belonging to any Tetrachord.	A. Proslambanomenos.

This two-octave scale is at least as old as the fourth century B.C., and it was a sliding scale, to be taken to the extent of an octave higher. Aristoxenus speaks of the highest of the above tetrachords in one of his extant fragments, as well as of the art of writing down music (pp. 39, 40). He also enumerates the six different modes of tuning the lyre, viz., two diatonic, three chromatic, and one enharmonic (p. 50 *et seq.*). Of these hereafter.

When the Greeks changed from one genus, or kind of scale (*γένος*) to another, they never altered the tuning of more than the two inner strings of each tetrachord. The Lichanoses and Parhypates of the lower octave, and the Trites and Paranetes of the upper were alone moveable (*κινουμένοί* or *φερομένοι*). Of these it was only in the enharmonic genus that both second and third string of each tetrachord were tuned differently. In the Chromatic the third from the top (Trites or Parhypates) remained as they were. The extremes of tetrachords and the "acquired tone" (Proslambanomenos) were fixed sounds (ἑστῶτες). This did not prevent the re-tuning of the whole lyre to any other pitch.

A comparison of the greater with the lesser system will show that the lower octave is the same in both. It is only from the key note upwards that any change is made. In the lesser system, after *a*, it goes to *b* flat, *c*, *d*, and stops; while the greater system carries up a second octave of the same kind as the lower one.

The Greeks had in all fifteen Diatonic scales, viz., five Principal scales, Dorian, Iastian or Ionian, Phrygian, Æolian, and Lydian. Each of these had its attendant Hypo and Hyper, or Dominant and Sub-Dominant. The Hypos were a fourth below their principals (which gives the same scale as the fifth above) and the Hypers were a fourth above.

When they modulated from one key to another they did it as we do, by some sound common to both, and the greater the connection between the two scales, the better was the modulation esteemed.*

THE FIFTEEN SCALES OF ALYPIUS are:

PRINCIPALS.

(A.) Hypo-Dorian.	(D.) Dorian.	(G.) Hyper-Dorian or Mixo-Lydian
(B♭.) Hypo-Ionian	(E♭.) Ionian.	(A♭.) Hyper-Ionian.
(B.) Hypo-Phrygian.	(E.) Phrygian.	(A.) Hyper-Phrygian.
(C.) Hypo-Æolian.	(F.) Æolian.	(B♭.) Hyper-Æolian
(C♯.) Hypo-Lydian.	(F♯.) Lydian.	(B.) Hyper-Lydian.

It will be observed that the classical Lydian was F sharp, and not F, as in church scales. The true Lydian was a tone above the Phrygian.†

In Pindar's time the Hypo-Dorian scale was called Æolian; the above arrangement of intervals between scales is therefore less ancient than his date.

The Greeks had no fixed pitch—neither have we at this present time. The only directions about it are to tune the lyre from the lowest distinctly audible tone of the voice,‡ and every man had a different voice. Instruments made to be played together would necessarily be at one pitch; but there was no fixed rule for them.

The Greeks had, in the fifteen scales, one beginning upon every semitone of the octave and two beginning beyond it. The five Hypos extended from A to C sharp, the five principals from D to F sharp, and the five Hypers from G to b.

The three highest Hypers were therefore the same scales as the three lowest Hypos, only taken an octave higher. These double names for the same keys were unnecessary, except in relation to their principals.

Dr. Burney says "That the ancients had no G sharp or E flat" (p. 26, vol. 1); but at p. 41 of the same volume he shows by a table of the modes that they had both. This curious instance of self-contradiction remains in his second edition.

As all Greek scales were tuned with perfect fourth, fifth, and octave, and all (till about the birth of Christ) with major tones only, there could not possibly be any musical difference, other than that of relative pitch, between one scale and another, if the lyre was tuned for each scale. Differences of character between one key and another arise from one key being less perfectly in tune than another. But inasmuch as certain metres were associated with particular scales, and the character of the music would correspond with the spirit of the verse, there might be as much difference between them as between a hymn and a march. The difficulty is that Greek authors were not agreed upon the character of any scale but the Dorian. That was to be severe, grave, and manly. But as to Phrygian, while Plato esteemed it as smooth and fit for prayer, Aristotle speaks of it as enthusiastic and bacchic. These contradictory estimates have been collected by Boeckh in his Metres of Pindar (lib. iii. c. 8.)

The usual way of tuning the lyre was to the Dorian, the central scale of the seven, and esteemed as the true Greek system. This preference for the Dorian is proved by all the accounts of the Greek octaves. The seven principal scales are therefore presented in that form, showing what notes would come upon the octave lyre (within the cross lines) and upon the two-octave lyre.

---

* Euclid, p. 21.

† All Greek writers are agreed upon this; see, for instance, Bacchius, p. 12.

‡ Gaudentius, p. 22, &c.

Scales for the Lyre.

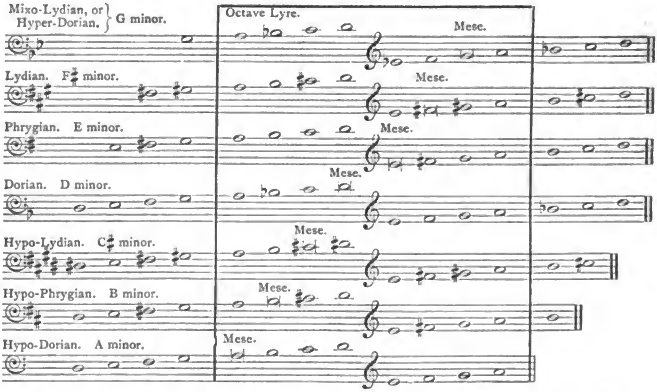

In the above diagram the sharps and flats are marked to the notes (as well as at the signature) only for the purpose of showing to the eye, at a glance, which of the strings must be retuned to change from one key into another. The Dorian, being the centre scale, has its entire fifteen notes; but the three scales above it want one, two, or three of their upper notes, while the three below it want one, two, or three of their lower. The octave lyre has its series complete.

Supposing a Greek singer to begin in the Dorian, and to wish to take in the Hypo and Hyper (or Dominant and Sub-Dominant), he would require either to re-tune one string for each, or else to have a ten-stringed lyre. All the other strings serve for the three connected keys, and it would be the same in any other key. Thus, in the key of C we require but F sharp and B flat for its Dominant and Sub-Dominant. A ten-stringed lyre would include the principal and its two connected scales. Hence the importance of a ten-stringed lyre, or a ten-stringed psaltery, such as we read of in the psalms. It was not the mere addition of an upper note or two, which was a great objection, in the public eye; as likely to lead to extravagances in declamation. After the Greeks had once discovered the octave system, they might have added another octave with the same facility as another string.

The preceding (as well as the following) diagram will explain that most ancient puzzle, the Greek octaves. The root of the difficulty has been this. Although the Greeks had different signs (σημεία) for writing down musical notes, and they wrote down music in the 4th century B.C., they had no fixed name for any note. Some readers may remember that there was an old plan of teaching singing in England (which has been partially revived), in which the key note was always called *Do*, and consequently every modulation or change of key made another *Do*. Just so with the Greeks, only instead of *Do* read *Mese*. Every string was tuned to Mese, and if a Greek knew the Mese he could tell the distance of any other note. So, when Euclid and others* say that the Mixo-Lydian octave begins upon Hypate Hypaton, they mean that it begins upon the lowest note but one of *its own* scale, just as it does in the preceding example. The key is G minor, with a minor seventh, and the octave lyre begins upon A. The great mistake has been to take the names of the strings for fixed sounds, and so to make a Mixo-Lydian octave in a Hypo-Dorian scale, instead of in its own scale. This error underlies all the old music called Gregorian (although in the time of S. Ambrose and S. Gregory there was no such peculiar music), and in consequence of this misapprehension "Gregorian tones or scales" have wrong key notes.

When Bacchius asks "What are the names of the three scales, if only three are used?" he answers for himself, commencing with the scale of highest pitch, "Lydian, Phrygian,

---

* Euclid, p. 15; Gaudentius, p. 19; Bacchius, p. 19; &c

Dorian." And "when seven?" "Mixo-Lydian, Lydian, Phrygian, Dorian, Hypo-Lydian, Hypo-Phrygian, and Hypo-Dorian." These are the seven in the preceding example.

Claudius Ptolemy proposed to reduce the entire number of fifteen scales to the above seven, thinking them sufficient, and he proposed another very desirable change, viz., to transpose them all a fourth lower for the lyre, so as to bring them all within the reach of ordinary voices. Dr. Burney says that many persons imagined Ptolemy to have proposed to raise them a fifth higher (*History*, vol. I., p. 45, line 4). That would have made them impossible for men. They were decidedly very high for men at the ancient historical pitch.

Ptolemy gives precisely the same rule for transposing these scales that any musician would give to-day, and the following is the result:—

Scales for the Lyre transposed a fourth lower, by Claudius Ptolemy.

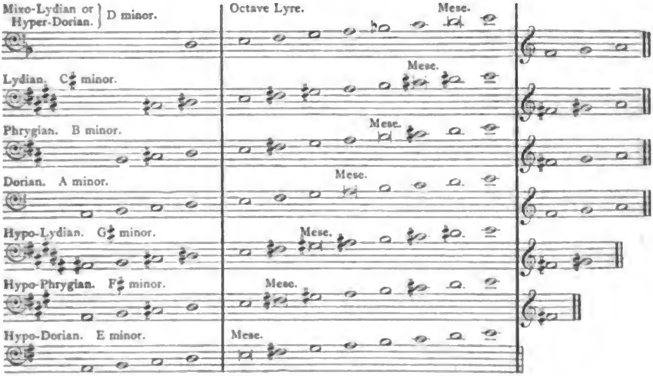

It will be observed that the key notes occupy the same positions as before; therefore the succession of intervals must be the same, for, as with us, the key note determines the succession.

And now, quitting the diatonic scale which is by far the most important of all, and the only one which the Romans adopted, we turn to the Chromatic Scale.

The Greeks had three kinds of Chromatic scale, of which only one was much used. Aristoxenus calls it the *Chroma tonaion* (χρῶμα τοναῖον). It ascended the tetrachord by semitone, semitone, and minor third, as below.

On the Octave Lyre A being the key note it stood thus—

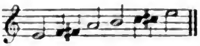

Or in our octave scale, thus—

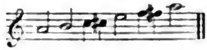

The outside notes of tetrachords are here marked in minims, and the inner notes in crotchets, only to be more readily distinguishable. Of the inner notes it is only the higher of each two that differs from the diatonic scale.

This chromatic scale is of interest in the history of music as being the first approach to a major scale among the ancients that has yet been discovered. It enabled them to play five tones in minor and to change them to five in major, but we have no proof that they ever made that use of the scale. There are the necessary F sharp and the C sharp for the key of A major, and, as the seventh of the scale is altogether omitted, the G sharp, which would be required in a complete scale, is not called into question. Five of the tones make a major scale, wanting the fourth and seventh of the key—in other words it is a scale of the five tones without the two semitones. If the major scale were played in the Lydian mode, beginning on F sharp, the succession of notes would be the same as the five short (and usually black) keys of the pianoforte. An enthusiastic Irishman or Scotchman might think this sufficient evidence that the five-toned Irish and Scotch tunes (we might add English, for there are many of them) are to be traced back to ancient Egypt. Proof would be wanting, but imagination sometimes goes a long way as a substitute. Divide the scale into major and minor and it runs thus:

Key of A minor.

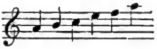

Key of A major.

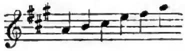

There could not be a complete major scale among the Greeks, because they had a musical law that the seventh of the scale must be at least a tone below the key note. It might be more, but it could not be less.

This chromatic scale was of very simple formation on the lyre. It required but to lower the forefinger string (lichanos), and such others as occupied the like position in the upper tetrachords, half a tone; and so to make a skip of a minor third down, instead of only a tone, between it and the highest string of the tetrachord, as from A to F sharp instead of from A to G.

The Enharmonic Scale was of the same kind as the chromatic, but made a skip down of a major third, as from A to F, instead of the minor third from A to F sharp, as in the chromatic. But the whole tetrachord only extended one semitone below F, viz. to E, and as there was a string already on E, and that a fixed sound, which could not be altered, the otherwise useless intermediate string was tuned to a quarter tone between E and F, and was occasionally used as a grace note. Such was the simple origin of quarter-tones in a Greek scale. They could not be harmonized.

Olympus, who seems to have flourished a short time after Terpander,* is said by Plutarch, on the authority of a lost work of Aristoxenus, to have discovered the enharmonic scale by merely passing over the lichanos, or forefinger string, in preluding—but that he did not use the quarter-tones. It was a later idea to utilize the unemployed string. The enharmonic of Olympus might have been played upon any lyre which had the ordinary tuning, for all the notes it required were common both to the Diatonic and to Chromatic scales.

The quarter tones were sometimes employed both in and before the time of Aristoxenus for he says that a singer could neither sing them with certainty nor the hearer judge of them.† He also says that no one could sing three quarter-tones in succession.‡

Other scales but little used.

Euclid, at the commencement of his treatise, (p. 3) names only the preceding principal scales, but afterwards recapitulates them together with others less used (p. 10). Although he gives but the tuning of one tetrachord of each he thereby explains the entire octave, because the octave (as we view it, *i.e.*, beginning from the key note) was made up first by the diazeuktic tone, or tone of disjunction (next above the key note), and then of two conjoined tetrachords above it. It would have been the same if begun from Proslambanomenos, the octave below this diazeuktic tone. A tetrachord, or fourth, consists, in Aristoxenian phrase, or roughly speaking, of two tones and a semitone. To show the divisions of the tetrachord we adopt Claudius Ptolemy's plan of explaining them (lib. i. c. 13) in preference to that of Aristoxenus and of Euclid (p. 11 and 12).

Aristoxenus and Euclid represent the semitone by 6 and the tone by 12, making the whole tetrachord 30. Ptolemy counts a diesis or quarter-tone for 6, a semitone for 12, a tone for 24, and the entire tetrachord as 60. Thus he avoids fractions. The following is the Complete List of Greek Scales:—

1. The Tonal Diatonic (already given)	*διάτονον σύντονον* ...	12, 24, 24 = 60.
2. The Soft Diatonic	*διάτονον μαλακόν* ...	12, 18, 30 = 60.
3. The Semitonic Chromatic (already given) ...	*χρῶμα τοναῖον* ...	12, 12, 36 60.
4. The Soft Chromatic	*χρῶμα μαλακόν* ...	8, 8, 44 = 60.
5. The Sesquialteral Chromatic	*χρῶμα ἡμιόλιον* ...	9, 9, 42 = 60.
6. The one and only Enharmonic (already given)	...	6, 6, 48 = 60.

---

* "Olympus must have flourished a short time after Terpander." Mueller's *Literature of Greece*, p. 202. M. Fétis most amusingly attributes this invention to an imaginary Olympus, who is said to have "lived about two centuries before the siege of Troy." The learned writer is as liberal with his thousands of years for the Greeks as if they were but *mille francs*. (*Histoire Generale de la Musique*, I. 131, 8vo., 1869.)

† Aristox., p. 14, l. 20.

‡ Aristox., p. 28.

Aristides Quintilianus describes also six enharmonic modes which, according to him, are of "very ancient" origin.* These scales are not mentioned by any other writer on music, neither is there any kind of allusion to the use of any second description of enharmonic scale elsewhere. Even Aristides himself says that the enharmonic scale is indivisible (at p. 133), and it must have been indivisible because the quarter-tone was the smallest interval employed in Greek music. The only two moveable sounds were already quarter-tones. These "very ancient" scales can therefore be nothing more than mixed scales.

The version of them given by Meibomius, who first published the treatise of Aristides, has been hitherto accepted without question.† The text that Meibomius followed was undoubtedly very faulty, but, when he attempted to amend it, he patched it in the wrong places. Scales were a great trouble to him, and he even failed to give the conjunct scale of three tetrachords correctly.‡

The following are the scales as printed by Meibomius The figures 1, 2, 3, relate to tones, and the ¼ to quarter-tones.

Lydian	¼	2	1	¼	¼	2	¼	..
Dorian	1	¼	¼	2	1	¼	¼	2
Phrygian . . .	1	¼	¼	2	1	¼	¼	1
Iastian	¼	¼	2	1½	1	...	...	...
Mixo-Lydian . .	¼	¼	1	1	¼	¼	3	..
Syntono-Lydian(*sic*)	¼	¼	2	1½	2	...	...	...

In the above, the key note of the Dorian is in its right place as the central scale, and, in that, it agrees with the manuscript. It has the diazeuktic tone next above it. The Phrygian, however, is in the wrong place. It ought to be a string above the Dorian. Meibomius has added a quarter-tone to this scale, to make it agree with the quantity stated in another line of the text; but he should have placed the added quarter-tone on the left instead of on the right of the key note. The figure 2 must be found wherever the key note or Mese is placed; but as it now stands, Dorian and Phrygian have their key note on the same string, which was impossible.

The scale above called "Mixo-Lydian" proves that these are mixed scales. There could be no interval of three tones without omitting both the key note and the diazeuktic tone—two fixed sounds—and there could not be a sequence of one tone after another in the enharmonic scale.

As to the scale here called Syntono-Lydian it is simply Hypo-Lydian, and nothing else. This is proved by its having its key note on the third string, and in that it accords with the manuscript. There was no such enharmonic scale as *Syntono*-Lydian, nor could such a prefix as Syntono be applied to any enharmonic scale whatever.§

The true positions of the key notes will be best exemplified by subjoining the enharmonic scales in their proper order on the octave lyre.

---

* *Αἷς καὶ οἱ πάνυ παλαιότατοι πρὸς τὰς ἁρμονίας κέχρηνται* (p. 21.)

† As by Boeckh in his Metres of Pindar, and by the late learned writer of the article "Musica" in the *Dictionary of Greek and Roman Antiquities*, edited by Wm. Smith, LL.D.

‡ See his notes upon Euclid, p, 63. He has omitted Hypate at the bottom of the lowest tetrachord (but Hypate was a standing sound, and could not be omitted), and commenced it upon Parhypate, which is one of the inner moveable sounds. All his tetrachords are therefore wrong. The origin of this mistake is that he has made Trite and Paramese into two strings, whereas in the conjunct scale they were never other than one, B flat. There should have been no Paramese in the scale, because it is the conjunct scale of Ion, and so dating more than 200 years after the octave system had been in use, Paramese then belonged only to the octave system. At some long previous time the names of Paramese and Trite had been indiscriminately applied to B flat in the conjunct scale, but never as separate strings in *that* system. It is singular that an editor of the Greek writers upon music should not have been able to write out a scale, and perhaps equally singular that such errors should have passed for 200 years unobserved. It is desirable to draw attention to them because these are by no means solitary errors in Meibomius. One instance more: in his notes upon Aristides Quintilianus (p. 209, column 1, line 6) he says that the two most ancient tetrachords were joined together by a string common to both, and that that string was called *hypate meson*. This is an entire mistake; the string was Mese. It is singular that he did not know it, even by the position of the tone of disjunction. How strange that so eminent a man should have edited the Greek authors upon music and yet not have mastered the scale system.

§ *συντεινω*, to draw together. In the enharmonic scale the moveable sounds, instead of being drawn tight, were *relaxed* to the extreme. "*Συντονωτάτη διάτονός ἐστιν*," says Aristoxenus, p. 25, line 11, and again p. 26. Further he says, "*καθόλου γὰρ βαρυτάται μὲν αἱ ἐναρμόνιοι λιχανοὶ ἦσαν, ἐχόμεναι δὲ αἱ χρωματικαί, συντονώταται δὲ αἱ διάτονοι*," p. 24, lines 22 to 25. The enharmonic was *πυκνοτάτον*, and its lichanos *βαρυτάτον* the *very opposite* to *συντονώτατον*—*πυκνὸν δ᾽ ἔσται τὸ ἐκ δύο διέσεων ἐναρμονίων καὶ χρωματικῶν ἐλαχίστων*. Aristox. p. 24, lines 17 to 19.

THE GREEK ENHARMONIC SCALE.

							Mese	
Mixo-Lydian . .	1	¼	¼	2	¼	¼	2	1
Lydian	¼	¼	2	¼	¼	2	1	¼
Phrygian . . .	¼	2	¼	¼	2	1	¼	¼
Dorian	2	¼	¼	2	1	¼	¼	2
Hypo-Lydian . .	¼	¼	2	1	¼	¼	2	¼
Hypo-Phrygian .	¼	2	1	¼	¼	2	¼	¼
Hypo-Dorian . .	2	1	¼	¼	2	¼	¼	2

The diagonal line shows the key note of each scale with its tone of disjunction next above it. The other figures of 2 are the highest notes of other tetrachords. Iastian (or Ionian) has no place in these scales, because it would require the position of one of the above seven, and it was for such reasons Claudius Ptolemy proposed to reduce the number of scales to seven. As to the Syntono-Lydian of the manuscript, it is clearly a mistake for Hypo-Lydian. It may seem strange that I should have to correct a Greek writer, but this is an unmistakable case, and one in which he was only speaking of something "very ancient," of which he had no intimate knowledge. The date of the writer has been clearly over-rated, and the manuscript of his treatise is exceedingly corrupt. In order to make one line of the text agree with another, Meibomius twice changed the word "tone" into "ditone" in the Lydian scale; he added a diesis, or quarter-tone to the Phrygian, the same to the Mixo-Lydian, and the final ditone to what is called Syntono-Lydian. All these additions and alterations will be seen by comparing his Latin translation with the Greek text (p. 21), and they are admitted and justified in Meibomius's notes.

Meibomius was the first to publish the work of Aristides Quintilianus, and he seems to have been desirous of magnifying such an acquisition to literature, by ascribing to the author as remote a date as he could guess. Meibomius overlooks his having copied from Claudius Ptolemy (the numerical estimate of 60 for a tetrachord is taken from Ptolemy) and thinks that Martianus Capella (who lived about the end of the fifth century), copied from Aristides. The numberless petty differences between the two upon a common subject seem greatly to militate against the theory, and Meibomius's own notes afford the evidence. Aristides had evidently studied Latin, because he quotes Cicero; he passes judgment upon Spaniards, Celts, and Thracians. They are either wild, brutal, or drunken, but Greeks are every thing that is good (pp. 72 and 73). Clearly he was a Greek. But even if he lived under the Roman empire, the Romans used no other scale than the one diatonic, so that all others were matters of history. We may meet in society a man of large general information, and yet if we ask him to define a "sackbut," he may fail; although he has probably seen one every time he went to opera or concert, and the name of the instrument was only changed in the last century. Or, we might even ask an accomplished musician to define a scale of Chaucer's time, and he might fail; so, likewise, may Aristides Quintilianus have failed. The corrections are supplied by his own ancestors. The passage referred to by Aristides is in the third book of the Republic of Plato.[*] The prefix of "Syntono" is usually unnecessary, because it means the *ordinary* Lydian, and therefore is rarely expressed, but Plato employs it, because he wishes to distinguish it from the Malakon (or laxly tuned) Lydian, to which also he adverts in the text.

Aristides Quintilianus is not the only Greek writer of comparatively late date, whose works require testing by those of his ancestors. The two Lexicographers, Hesychius and Suidas, explain musical terms indifferently, and the difficulties that have been found in translating certain passages of Plato and Aristotle, are in a measure due to reliance upon their imperfect definitions of technical words relating to the art.

Gregorianischer Gesang (*Ger.*) Gregorian chant. [Plain song.]

Gregorian. [Plain song.]

Griffbrett (*Ger.*) Finger-board of a stringed instrument.

Groppo (*It.*) A bunch or group of notes.

Gros Fa. The square notation used in old church music.

Grosse (*Ger.*) (1) Major, applied to intervals. (2) Grand, or great; as *grosse Sonate*, grand sonata. (3) Double in pitch; as, *grosse Nazard*, a quint, an organ stop, an octave below the twelfth; *grosse Quinte*, a pedal stop of 10⅔ ft. in length.

Grosse caisse (*Fr.*) The big drum.

Grosso (*It.*) Great, full, grand; as *concerti grossi*, grand concertos.

Gros tambour (*Fr.*) The big drum.

Grottesco (*It.*) Grotesque, comic, humorous.

Ground Bass. **Basse contrainte** (*Fr.*) *Basso ostinato* (*It.*) *Bassthema* (*Ger.*) A bass passage of four or eight bars in length, constantly repeated, each successive time accompanied with a varied melody and harmony.

---

[*] *Μιξολυδιστί, ἔφη, καὶ συντονολυδιστὶ. . . . Τίνες οὖν μαλακαί τε καὶ συμποτικαὶ τῶν ἁρμονιῶν; Ἰαστί, ἦν δ' ὅς, καὶ λυδιστί.* Lib. 3, p. 399a.

The idea of this peculiar form of composition was probably suggested by the practice of singing a varied descant upon a given plain song.

The old writers contemporary with, and immediately succeeding Palestrina, frequently made use of the church-tones as themes for counterpoint, but did not always give those themes to any one particular voice, but assigned them to all by turns. This form of writing was called by them "Falso-bordone." Monteverde in his opera "Orfeo," represented at Mantua 1607, has a Moresca, a dance written upon a sort of ground bass in a form that may be said to be the connecting link between one form of the falso-bordone and the basso ostinato. Strictly speaking, in this case, it is an imitated bass, as it appears each successive time in a new key, but the intervals are imitated throughout. It is repeated four times. The first time it is in G major, the second in C major, the third in A minor, the fourth time in D minor, each section ending with a major chord.

MONTEVERDE.

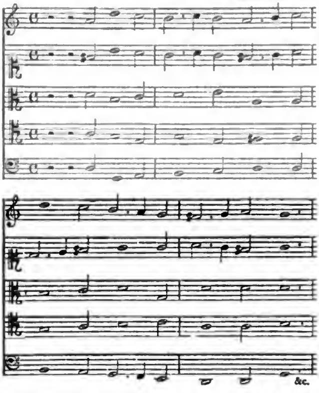

The employment of a ground bass as a regular musical device became more general later in the same century. In the works of the composers of that period, pieces with ground basses are frequently found, either strictly continued or with short digressions. They were employed for compositions in all styles, for the church, for the stage, or for the chamber; for movements in suites, arias, dance tunes, &c. Many famous musicians furnished "grounds" for the purpose of *extempore* performance. Well known "grounds" were often selected by composers for a species of writing called *folias* or *follias*. Thus Farinelli's or Fardinel's ground was used by Corelli, Vivaldi, and others.

Grounds by Purcell, Tollit, Moteley, Pepusch, and others, are often found as the bases of many compositions published, even so late as the latter part of the last century. The practice of performing upon a given ground bass gave rise to treatises professing to instruct the ambitious in the art. One of the most famous of these works, "Chelys Minutionem, or the Division Viol," by Christopher Simpson, 1665, gives, as far as possible, all the necessary rules, with many examples. His description of the method of performing "division on a ground," is interesting, and may not be out of place here. He says that "Diminution or division to a ground, is the breaking either of the bass, or of any higher part that is applicable thereto. The manner of expressing it is thus—a ground, subject, or bass, call it what you please, is prick'd down on two several papers; one for him who is to play the ground upon an organ, harpsichord, or whatever instrument may be fit for that purpose; the other for him that plays upon the viol, who, having the said ground before his eyes as a theme or subject, plays such variety of descant or division in accordance thereto as his skill and present invention do then suggest unto him." As the "grounds" given were very short, and the compass of the viol was in those days limited, this quaint sort of extempore descant was perhaps not difficult, neither could it have been very varied or interesting, except perhaps, to those immediately concerned.

The kind of ground bass given in the "Chelys" may be seen by the subjoined:

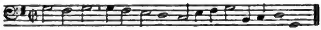

The opera "Dido and Æneas," written by Purcell in his 22nd year, contains a very good example of a song written on a ground bass, which is, curiously enough, almost identical with that chosen by Bach in his B minor Mass as the ground of the movement *Crucifixus*. The following example is from Lulli's opera "Persée." The ground bass only consists of four bars, which recur six times:

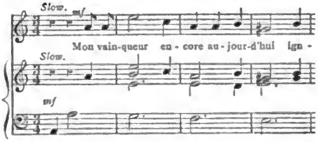

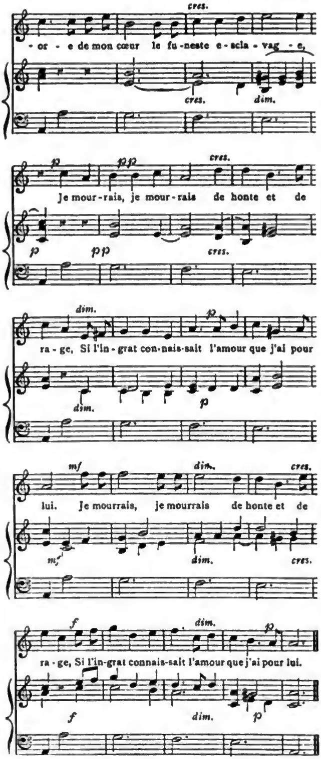
cres.
- or - e de mon cœur le fu-neste e-scla-vag - e,
cres.
dim.
p
pp
cres.
Je mour-rais, je mour-rais de honte et de
p
pp
cres.
dim.
p
ra-ge, Si l'in-grat con-nais-sait l'amour que j'ai pour
dim.
p
mf
dim.
cres.
lui. Je mourrais, je mourrais de honte et de
mf
dim.
cres.
f
dim.
p
ra-ge, Si l'in-grat connais-sait l'amour que j'ai pour lui.
f
dim.
p

Chaconnes, and Passacailles or Passacaglios, were generally written on ground basses, and the ingenuity and skill displayed in many existing examples are both interesting and instructive. There are Passacailles and Chaconnes by Couperin and Rameau, Bach, Handel, and others, too long to quote here. Handel, whose sixty-two masterly variations on a ground bass in his famous Chaconne are well known, has shown also how the like artifice may be effectively employed in choruses, as those in "Saul," "Susanna," and other works sufficiently prove. A fine modern example will be found in Brahms's Variations on a theme by Haydn.

Group. (1) A series of notes, of small time-value, grouped together; a division or run. (2) The method of setting out band parts in score.

Grundstimme (*Ger.*) The bass part.

Grundton (*Ger.*) (1) The bass note. (2) Fundamental bass.

Gruppetto, gruppo (*It.*) A series of notes grouped as a cadenza, division, or ornament. Playford (Introduction to the Skill of Musick) gives the name *Double relish* to the gruppo and the following directions for its performance:

Gruppo, or double Relish.

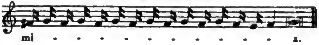

by which it would appear to have been similar to the grace now called a "shake." He writes the latter as follows:

Trill, or plain shake.

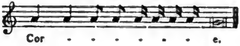

G string. The name of the first string on the double bass, the third on the violoncello, viola, and guitar, and the fourth on the violin.

G-Schlüssel (*Ger.*) The G or treble clef.

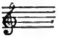

Guaracha. A lively Spanish dance in $\frac{3}{8}$ or $\frac{3}{4}$ time, usually accompanied on the guitar by the dancer himself.

Guddok (*Russ.*) A Russian fiddle. [Violin.]

Guerriero (*It.*) Warlike, martial.

Guet (*Fr.*) A flourish of trumpets.

Guida (*It.*) (1) A guide, a direct. (2) The subject of a fugue. [Fugue.] [Direct.]

Guide-main (*Fr.*) A hand-guide, a mechanical contrivance for regulating the position of the wrist in pianoforte playing, invented by Kalkbrenner.

Guidonian syllables. [Aretinian syllables.]

Guidonian system. [Notation.]

Guimbarde (*Fr.*) [Jew's Harp.]

Guitar. *Guitare* (*Fr.*) *Chitarra* (*It.*) *Guitarra* (*Sp.*) A stringed instrument, played by plucking or twitching the strings with the right hand while the left is engaged in forming the notes by "stopping" or pressing the strings against the frets on the finger board.

The modern, or Spanish guitar as it is called, has six strings, the three highest of gut, the three lowest of silk, covered

Digitized by Google

with a fine wire. The *accordatura* is as follows:

(Sounding an octave lower.)

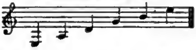

The guitar is but little used now in England, though at one time it was very fashionable. Other nations who still employ it, call it by several names, most of which will be described hereafter. The guitar is rarely, if ever, employed as an orchestral instrument, but is very valuable as a portable means of accompaniment. The existence of frets upon the guitar limits the number of modulations capable of being performed in the normal tuning. When it is desired to make a complete change of key the *capotasto* screwed over the finger-board alters the tuning at any desired point, or a temporary change is made by the *grande barré*, that is by laying the forefinger of the left hand completely over the strings, the remaining fingers being engaged in stopping a chord. In the classification of musical instruments it is convenient to speak of three general sorts, wind, string, and pulsatile. The guitar belongs to the second kind, and may be said to represent a very large family universally distributed, bearing a variety of names according to the tongue of the nation by which it is used. All instruments may be considered as belonging to the guitar family, which possess a resonance body or sound box, together with a finger-board, against which the strings with which they are furnished may be pressed or stopped.

Following the course of history, we find that instruments of the guitar kind are of great antiquity, as well as of general use by people of all nations.

The kinnor and nebel, mentioned in the Bible, were stringed instruments, of the guitar or harp family, but of their exact nature it must be confessed little is known, though much is conjectured.

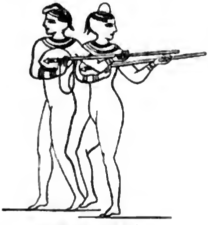

Egyptian Nefer.

The Egyptian frescoes and other paintings, valuable as showing the frequent use of musical instruments, include several specimens of the harp and guitar family. The Nefer, one of the latter class, had a neck, sometimes with a carved head, and was furnished with three strings, and had a resonance box. Upon the neck, or finger-board, frets were tied or fastened, as in the modern guitar. Each string is said to have been able to produce two octaves.

The three strings were supposed to correspond with the seasons of the Egyptian year.

Grecian writers, describing Egyptian instruments, do not afford much real information concerning them, and all attempts to reconcile their statements only lead to confusion; for conjecture is not conviction. Too much trust has been placed in the accuracy of sculptured and painted images, and various theories have been founded upon the character of musical instruments as deduced from their represented forms. As with ancient, so with modern musical instruments of far away countries, travellers' tales have too often been trusted, and their statements received as conclusive, when in the majority of instances they are confessedly ignorant of the subject upon which they give "authoritative judgment."

Philology does not, after all, furnish the best assistance towards determining relationships in this matter, and, as a rule, the picture of an instrument offers but a little help or guide in the matter. References to musical instruments by the poets of several ages offer no aid whatever, but on the contrary, often tend to mislead. If they were trustworthy, it might reasonably be assumed that no other instruments but the lyre and harp were ever employed to "assist the muse." But colloquial terms—often despised by classical poets—are of most value to the historian, and it is therefore found that the common names applied to a stringed instrument with a finger-board, kissar, cittern, zither, kitra, kithara, geytarah, guitar, point to a common origin.

There is no question but that the guitar was introduced into Europe after the Crusades. The name, purely Eastern, has been adopted with only such a variation in spelling as European use demands. The modern Egyptians call it "gytarah barbaryeh," the guitar of the Berbers, the people who are the direct descendants of the ancient race of the country; and as names and words in the East vary in the course of ages less than those in the West, it is likely that the word is of high antiquity.

This "gytarah," or kissar is of the following form :

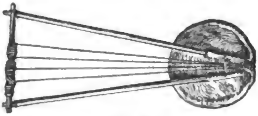

Nubian *Kissar.*

It is usually mounted with four strings tuned, according to Engel, as below :

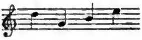

Its form is not unlike some of the instruments represented on ancient Egyptian and Assyrian monuments, and although the name would imply some connection with the modern guitar, its shape would identify it rather as belonging to the lyre kind. But there are other stringed instruments used by Eastern people more in the form of the modern guitar.

The kitra or kuitra popular in Morocco sometimes has the resonance-box or body made of a tortoise shell, after the manner that Hermes is said to have constructed his lyre.

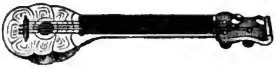

Tunisian Kitra.

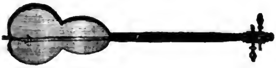

Kitra or Gunibry.

The resonance-body of the Gunibry is made of a bottle-pumpkin cut longwise, and covered with sheep skin. Its two strings are of catgut.

The sitar, choutarah, or tamboura of Hindostan had originally but three strings of wire—as the name *sitar* implies—which were afterwards increased to four or five. The body is made of a gourd, the neck of cocoa wood furnished with pegs. The strings are played with a plectrum of twisted wire called by the name of *mizrab*, worn on the forefinger of the right hand. There is another form of Hindoo guitar of a somewhat peculiar construction called Vina or Bina, which has a gourd at each end. [Bina].

Hindoo Guitar.

The Chinese, though a people of a different stock, have an instrument called Yue-kin or moon guitar, having four silken strings arranged in pairs, each pair being tuned in unison, and the two pairs a fifth apart. The instrument has been called by travellers following the method of pronouncing the name in Canton, *gut-kum*, which may or may not be philologically related to guitar. The *gut-kum, yue-kin*, or moon guitar, has inside its resonance box some pieces of loose metal which are occasionally shaken during performance.

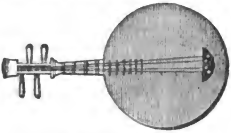

Chinese Yue-kin or Moon-guitar.

The lute, another member of this family, also comes from the East, the name is the European method of spelling its title "el 'ood." The pandore, bandore, pandoura, and mandoline are simply other names for a lute or guitar, arising from fancy or accident. The mandola or mandoline, for example, derives its title from the almond shape of the resonance-body. The Italian word for almond is mandola. Variety of names for the same thing, together with slight differences in form, often tend to confuse the enquirer. The method of performance, the shape, the mounting, the material of which they are constructed, and various other causes, are often taken into consideration in the naming of instruments. If these reasons are lost sight of, a certain amount of confusion naturally arises in the classification of musical names and titles; and things are treated and spoken of as dissimilar, which are really closely connected. If, for example, we were five thousand years older, and no specimen of a pianoforte or of its musical literature existed, and we were left to judge of the form and use of the instruments called by the several names applied to it, we might say with good authority out of existing documents, that our ancient English ancestors were accustomed to listen for hours to a performance upon a broad wood, probably an extensive forest or a wide plank, as the acute future critic would say. Further, it might be inferred that our German contemporaries were enraptured with the skill of one who was able to produce similar effects from a flügel, the wing of a bird. The connection between a forest and a bird's wing might suggest some very ingenious comments. In cases where titles are given independently of those already applied to certain things, and detailed descriptions are wanting, the difference becomes apparently wider each successive age until all ends in chaos. If the ancient Eastern title *geytarah* had not been adopted with the instrument by the Spaniards, and by other nations following them; the changes in the form of the instrument might have been held as indicating many origins. There seems to be no connection between the words nefer, nebel, pandoura, lyre, and kithara, still it is not unlikely that they had a common start-point.

An instrument of a form like the Egyptian Nefer is found depicted upon Assyrian monuments, but strange to say there are no representations of a finger-board instrument among the Greek antiquities. That the Greeks knew of the pandoura is evident from the fact that it is mentioned by Nicomachus, and subsequent historians. The modern conclusion that they preferred their own instruments without necks, "although they adopted the system of the Egyptians for the subdivision and measurement of strings," is very doubtful, for the one could not have been done without the aid of the other. For if we are to believe that the strings of the Egyptian *Nefer* had a compass of two octaves each, those two octaves must have been obtained by means of a finger-board, and if the Greek instruments were without finger-boards, and the strings were open from end to end, without a backing along their lengths, how were they stopped, or how could a string be subdivided? It is therefore probable if Greek music was as perfect as it is said to have been, that finger-boards, fretted or otherwise, were known and used, and poets and sculptors, disdaining to employ common forms, gave fancy shapes to musical instruments.

The ancient Greek kithara (*κιθάρα*) is admitted to have been portable; and Mr. Chappell ("History of Music," p. 37) says, "the lower strings of the kithara were played by the fingers of the left hand, and the higher strings by the plectrum held in the right hand." And again, that the instrument "was held on the left side of the body, with the left arm behind the instrument, for the purpose of reaching the base strings which were furthest from the player." Now, the difficulty of performing upon an instrument of the lyre or harp kind under the conditions above set forth, must have been great. If the description be amended thus: "The lower *part of the* strings of the kithara were *stopped* by the fingers of the left, and the higher *part of the* strings *played* by the plectrum," &c., the whole matter becomes perfectly clear, and the kithara shows its relationship to the guitar, as well as its power of sounding octaves on each string like its Egyptian prototype.

When the drawings of ancient performers on stringed instruments are examined, it will be found that if, as they are represented in the majority of cases, a modern player were to hold his instrument in a similar fashion, he would be unable to support and play it at the same time. As the human form appears to have been pretty much the same in old time as it is at the present day, it is more than likely that the artists "evolved" the representations out of their "inner consciousness," and, therefore, that they are not to be confidently trusted.

Plato, quoted by Hawkins (p. 91, Novello's Ed.), "advises to train up children to use the right and left hand indifferently." In some things, says he, "we can do it very well; as when we use the lyre with the left hand, and the stick with the right." Unless some other occupation than that of merely holding the instrument were intended, such a piece of advice would be superfluous. The cithara is mentioned by Ovid, Horace, Virgil, and other Latin authors, with but little reference to the manner of performing upon it, other than that it was held in one hand, while the other struck the strings with a plectrum.

The number of strings upon the guitar has been varied from time to time in Europe; and since its introduction, the instrument has been more or less popular. By the name of gittern, gittron, gitteron, &c., it is spoken of by the mediæval poets; and as the lute, it was familiar during the 16th and 17th centuries.

In France, Spain, and Italy, the guitar is employed as an accompaniment for the dance as well as for the voice; and at one time during the last century, it was so popular in England that the sale of pianofortes was interrupted, until an ingenious maker bethought him of a plan by which to weaken and ultimately to destroy its popularity. There is an extensive literature of guitar music, called into existence by the revival in favour of the instrument, brought about by the number of Spanish refugees resident in England during the Carlist rebellion of 1834-1839. But in the present day, the instrument is but little cultivated; in fact, it may be said to have become undeservedly neglected.

Gusto, con; gustoso (*It.*) With taste and expression.

Guttural. Tones produced in the throat.

H.

H. The note B natural in the German system of nomenclature, the letter B being used only for B flat.

Hackebrett (*Ger.*) [Dulcimer.]

Hadán or **hadan** (*Egyptian*). The call to prayer sung by the mueddins from the towers or minarets of the mosques, thus given by Lane:

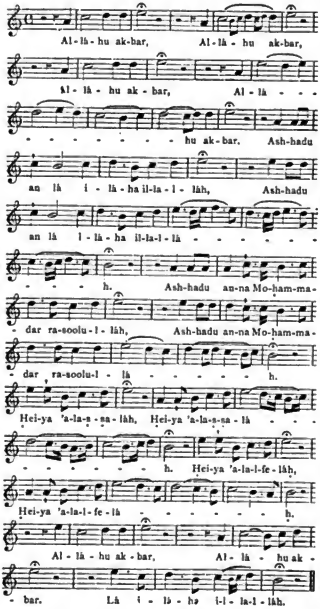

Al-lâ-hu ak-bar, Al-lâ-hu ak-bar,
Al-lâ-hu ak-bar, Al-lâ - - - - - - - - - - - hu ak-bar. Ash-hadu
an lâ i-lâ-ha il-la-l-lâh, Ash-hadu
an lâ i-lâ-ha il-la-l-lâ - - - - - - - - - - h.
Ash-hadu an-na Mo-ham-ma-dar ra-soolu-l-lâh,
Ash-hadu an-na Mo-ham-ma-dar ra-soolu-l-lâ - - - - - - - h.
Hei-ya 'a-la-s-sa-lâh, Hei-ya 'a-la-s-sa-lâ - - - - - - - - - - h.
Hei-ya 'a-la-l-fe-lâh, Hei-ya 'a-la-l-fe-lâ - - - - - - - - h.
Al-lâ-hu ak-bar, Al-lâ-hu ak-bar.
Lâ i-lâ-ha i-l-la-l-lâh.

which means:—God is most Great! (repeated four times). I testify that there is not a deity but God! (twice). I testify that Mohammed is God's Apostle! (twice). Come to prayer! (twice). Come to security! (twice). God is most great! (twice). There is no deity but God!

Halbcadenz (*Ger.*) Half-cadence, or half-close. [Cadence.]

Halbnote (*Ger.*) A minim.

Halbton (*Ger.*) A semitone.

Half cadence. [Cadence.]

Half-note. (1) A minim. (2) A semitone.

Half shift. A position of the hand in playing on instruments of the violin family. It lies between the open position and the first shift. [Shift.]

Hallelujah. [Alleluia.]

Halling. A Norwegian dance, somewhat of the character of a country dance.

A HALLING.

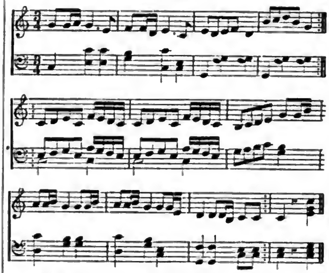

Hals (*Ger.*) The neck of an instrument.

Hammer. (1) A piece of wood having a padded end or a nob, with which strings are struck. In the case of the dulcimer the hammers are held in the hand; in keyed instruments the hammer is acted upon by leverage from the end of the key. [Dulcimer.] [Pianoforte.] (2) The iron or wood striker of a bell. According to Denison, the weight of the hammer should be a fortieth part of a bell whose diameter is equal to twelve times the thickness of the sound bow. But the distance the hammer is made to rise for the blow must of course influence the weight of the hammer.

Handguide. [Guide-main.]

Hardiment (*Fr.*) Boldly, daringly.

Harfe (*Ger.*) [Harp.]

Harmonica. An instrument, the tones of which are produced by striking rods or plates of glass with hammers, either held in the hand or acted upon by keys. It has a compass of about two octaves from middle C or D upwards.

Harmonica. A name sometimes given to a mixture stop on foreign organs.

Harmonic Flute. [Harmonic stops.]

Harmonichord. An instrument played like a pianoforte, but sounding like a violin. The tone is produced by the pressure of the keys, which sets a revolving cylinder of wood, covered with leather, and charged with rosin, in action over the strings. It has also been called piano-violin, violin-piano, tetrachordon, &c.

Harmonici. The followers of the Pythagorean system of music as opposed to that taught by Aristoxenus. They were also called Canonici. The Aristoxenians viewed music as an art governed by appeal to the ear; the Pythagoreans, as a science founded on physical laws.

Harmonicon. An instrument only used as a toy, which consists of free reeds enclosed in a box in such a way that inspiration produces one set of sounds, respiration another.

Harmoni-Cor. An instrument invented by Jaulin of Paris, consisting of a series of free reeds similar to those used in the Harmonium, placed in a tube shaped like a clarinet. The compass of the instrument is two octaves with intermediate semitones, the keys are arranged in a manner similar to those of a pianoforte, that is to say, all the notes of the normal scale are in one row and the chromatic notes in another. The wind is supplied by means of a mouth-piece.

Harmonics. The sounds produced by a vibrating string or column of air, when it is subdivided into its aliquot parts. [Acoustics, §10.]

Harmonic scale. The scale formed by a series of natural harmonics. [Acoustics, §10.]

Harmonic stops. Organ stops, both flute and reed, having tubes twice the normal length, but pierced with a small hole in the middle. Harmonic flute stops are of great purity and brilliancy, they are of 8 ft. or 4 ft. pitch. Harmonic piccolos are of 2 ft. pitch. Harmonic reed stops (tromba, tuba, trumpet, &c.,) are generally on a high pressure of wind, one of the great advantages of all harmonic stops being that they will take a very strong pressure of wind without overblowing. The fact is, that the harmonic-tube, having two synchronous vibrating columns of air, partakes of the nature of a pipe already overblown to its first harmonic, the octave.

Harmoniemusik (*Ger.*) Music for wind instruments.

Harmonique. (*Fr.*) Harmonic.

Harmonist. One who can sing or play in harmony.

Harmonium. A keyed wind instrument whose tones are produced by the forcing of air through free-reeds. The better class of harmoniums have several sets of vibrators of different pitch and of various qualities of tone. The stop called *expression* is a mechanical contrivance by which the waste-valve of the bellows is closed, so that the pressure of the foot has direct influence on the intensity of the sounds produced. A tremolo is produced by causing the wind to quaver as it passes through the reeds. The Vox Angelica gives a delicate undulating tone which is produced by two sets of vibrators to each note tuned slightly apart. [Reed.]

Harmony. In its earliest sense among the Greeks this word seems to have been a general term for music, a sense in which our own poets often use it. But from its meaning of "fitting together" it came to be applied to the proper arrangement of sounds in a scale, or, as we should say, to "systems of tuning." Whatever opinions may be held as to the antiquity of harmony in the sense of *symphony* or "sounds in combination," it is quite certain that among the ancients the art of harmony never advanced beyond the use of *accompanying chords.* Treatises on music, which we in these days call on "harmony," dealt (among the Greeks) with the following subjects:—The divisions of the monochord, the three genera, the sounds proper to the different modes, the shape and position of the letters representing musical sounds, and, to a limited extent, the art of tune-making, about which, however, but little is known. Boethius, who turned into the Latin tongue all the most important elements of Greek music, writes, on the rudiments such as sound, interval, consonance; on the ratios of intervals; on letter-notation; on the modes; on the discussions arising from the use of the monochord. When this work was written (in the early part of the sixth century) there had already been growing up for a considerable period a school of church music, probably started by Ambrose in the fourth century, whose function it was to form a practical school of music rather than scientific. But notwithstanding this fact, writers thought it either fashionable or necessary to found all their works on the then defunct Greek system. Even in the eleventh and following centuries, when the Hexachord system had rendered a study of the Greek scales practically useless, their discussion formed an important part of every treatise. Under the word Descant will

Digitized by Google

be found a short description of the different stages through which early forms of harmony passed. But although a treatise on Descant and Counterpoint in one sense is a "treatise on harmony," yet, a very different meaning is carried by the word in its more modern sense—it signifies, in fact, a statement of the system of forming chords with an account of their proper movement or progression according to key-relationship.

The authors of the earliest treatises upon harmony, in the sense just described, seem to have laboured to reconcile the old teaching with new discoveries, and though ostensibly treating with harmony, they began with ratios and proceeded into counterpoint, avoiding harmony in its proper sense altogether. In some cases the writers quietly and cleverly avoided the main question, giving apparently elaborate descriptions of the subject, which after all amounted to nothing. Thus Bateman, who in 1582 published an edition of Trevisa's translation of "Bartholomæus de proprietatibus rerum," gives to posterity the accepted signification of the word harmony in his own time, for although he is supposed only to have translated into more modern English the work of Trevisa, written in 1400, he actually altered and added to the text in such a fashion that his changes are as valuable as the original:

> "De Armonya.
>
> "Armonya Rithmica is a sownynge melodye, and comyth of smything of stringes, and of tynklyng other ryngynge of metalle. And dyverse instrumentis seruyth to this manere armonye, as Tabour, and Tymbre, Harpe, and Sawtry, and Nakyres, and also Sistrum.
>
> "And the melodye of musyk is nemdnyd and callyd by the names of the nombres. Dyatesseron, Dyapente, and Dyapason have names of the nombres whyche precedeth and gooth tofore the begynnynge of those sayd names. And the proporcion of theyr sownes is founde and had in those said nombres, and is not founde, nother had, in none other nombres."

One of the earliest books printed by its author on the subject of harmony was the "Theorica Musicæ" of Franchinus Gaffurius, Milan, 1480. The doctrines taught are essentially the same as those of Boethius, and as this has been described already nothing more need be said. It may here be noted, that the examples of harmony by Franchinus are more modern in style than those of other writers contemporary with or previous to him.

The next writer of any note among the scores of authors who wrote upon music was Andreas Ornithoparcus, whose "Micrologus," written in Latin and printed at Cologne in 1535, was translated into English by John Douland, and published in 1609. His chapters on harmony (concentus) show no advance of thought notwithstanding the fact that some modern writers claim a place for him in that part of the Temple of Fame devoted to discoverers. The majority of the treatises of the 16th century were very learned and doubtless very clever, but they add nothing to literature not already known. Neither Salinas "De Musica," Salamanca, 1577; Calvisius "Melopeian," Erfurth, 1595; Zarlino, "Institutione Harmoniche," Venice, 1592; Valerio Bona, "Regole di Contrappunto," Milan, 1595; Zacconi, "Prattica di Musica," Venice, 1596; Bottrigari, "Il Melone," Ferrara, 1602; Cerone, "El Melopeia," Naples, 1613; nor the multitude of lesser writers of the period in which the above named authors existed, do anything but repeat the received theories, in a more or less wordy manner, rather increasing the confusion into which the knowledge of the science of harmony had fallen by their controversies and partisanships. While the writers of Italy, Spain, and Germany were adding to the "learned ignorance," our countryman, Thomas Morley, was not a whit more far-seeing than his contemporaries. If the state of music could be judged by the treatises alone, it might reasonably be considered that no new musical discovery had been made for centuries, and the constitution of music was such that new discoveries were impossible. But while the theorists sought to confine all music to certain "proportions and ratios," practical musicians were quietly finding out new and forbidden combinations of chords, to the utter confusion of the theorists, who gradually had the conviction forced upon them in such wise that they were compelled, reluctantly however, to confess that, "Il senso d'udito è stato e sarà sempre il solo legislatore dell' arte musica." The more daring musicians were supported by a large, and to a certain extent, an influential crowd of admirers who delighted in the unscientific pleasures the new music brought to the ear. We read of Claudio Monteverde, in the year 1600, being engaged in dispute with "some of the ablest musicians of his time," in consequence of his use of certain dissonances which were employed in an unprecedented manner. Also of Ludovico Viadana, about the same time, inventing or rather perfecting the system of musical shorthand now known as "thorough bass." And in later years Frescobaldi, following the path laid open by Monteverde, boldly introduced a series of progressions in his works, which were against all accepted rules. By this time—the first quarter of the 17th century—writers on theory silently abandoned the "proportions and ratios." Though it is easy to trace the effects of the old teaching, even when it is not expressed openly.

The principles of harmony or composition, as set forth by Dr. Thomas Campion, 1620,

Digitized by Google

and later by Christopher Simpson, 1678, treat of nothing more terrible than such common chords as could be formed out of the unaltered notes of the scale. Throughout the whole of the works by these authors the chord of the seventh on the dominant is introduced but sparingly, and when it is used, its "discordance is softened as much as possible." There is little, if anything, entirely new in these books, nothing at all to account for the new chords that musicians were occasionally introducing into their compositions. As Simpson's "Compendium" had a continued and steady sale, running through many editions, teaching the same old and worn-out principles, while Purcell was writing with all the hardihood of enthusiastic youth such chords and progressions as those found in the song quoted in the article "Ground Bass" *theory* running in a widely different direction from *practice*, did not hesitate to condemn these inventions; and, moreover, to wonder "how any judge of correct and pure harmony could tolerate such licences." It was probably owing to the fact of this diversity of opinion between the laws that were made for composers and those they made for themselves, that there are but few treatises on harmony belonging to the latter part of the 17th and the early part of the 18th centuries. The many editions of Simpson's "Compendium," Elway Bevin's "Briefe and Short Introduction," Morley's "Plaine and Easie Introduction," show how little improvement had been effected, and how rare a thing a new idea on the subject was. This state of things continued for some years more, when the publication of Rameau's "Traité d'Harmonie," Paris, 1722, turned all thinking minds into the direction suggested by that book. Looking at it by the improved light of the 19th century, it is difficult at first to trace the cause of the enormous popularity of the work, and its powerful influence on all classes of theorists. The plan is very simple but incomplete, accounting for a part, but not attempting to grapple with the whole science of harmony. The eagerness with which it was received and imitated, goes to prove how ripe men's minds were for the subject, and how willing they were to receive that which attempted to account for some, if not for all the intricacies of their art. The marrow of the whole work, the plot of the whole scheme proposed by Rameau, is comprised in the following words, according to the English edition of 1752:—"All notes that carry the perfect or common chords may be deemed key-notes, and all notes that carry the chord of the seventh may be deemed governing notes, with this difference, that the governing note of the key must have a sharp third. These two are, as it were, the only chords in harmony."

The main principle of this book, the referring all harmony to a regular fundamental bass on which chords are constructed after the above plan, led its author into some errors, which were copied and increased by subsequent writers. In order to make the theory fit into and agree with practice, Rameau considered suspensions as essential chords, chords by *supposition* as he calls them. His desire to derive all the chords of modern harmony from a construction of third upon third was a praiseworthy one, and nearer the truth than many other theories. He avoids the chord of the ninth, calling it a secondary seventh, speaks of the fourth as an eleventh, admits that the judicious use of discords gives greater liberty to composers, states F and D to be the fundamental bass of the same chord, the chord now known by the name of the added sixth, and adds one or two other peculiar notions needless here to particularise.

The following is a figured bass passage with the fundamental bass, as given by Rameau:

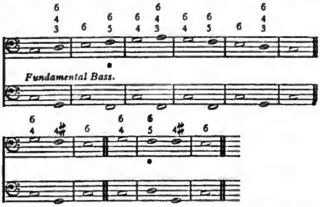

The chords marked *, though really the same, are derived from different roots by Rameau. Godfrey Keller ("Rules for a Thorough Bass," London, 1731) proceeds upon a somewhat different plan to Rameau. He makes no mention of the fundamental bass, employs many new chords, including those now known by the names of the major and minor ninth, *e.g.*:

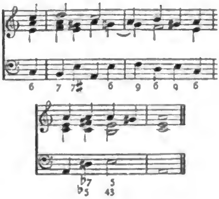

Digitized by Google

He allows that "the composer (especially in few parts) may compose as many sixes either ascending or descending by degrees as he thinks fit," but he does not lay claim to novelty in suggestion, though he might have done so safely.

The next writer on harmony, Roussier ("Traité des Accords," Paris, 1764), simply copies Rameau without acknowledgment. Following a rough chronological order, Marpurg next demands attention. In his "Handbuch bey dem General-Bass," Berlin, 1755, he professedly agrees with Rameau, but differs from him throughout by making the added thirds, which are to form chords by supposition, fundamental notes contrary to Rameau's teaching. Tartini, in his "Trattato di Musica," Padua, 1754, and "De' Principii dell'Armonia Musicale," Padua, 1767; Kirnberger, in his works, "Die wahren Grundsätze" and "Die Kunst des reinen Satzes," published in Berlin in 1773 and 1774, considered suspensions and all interruptions as no real part of a system, but, as the last named writer regarded them, "as clouds may be considered in astronomy, viz., as occasional occurrences, which must be tolerated when there, but which are passed by as if they had not been there at all." A very convenient, if not a satisfactory, method of getting over a difficulty. Twenty years later, Kollman ("Essay on Musical Harmony," London, 1796) endeavoured to found an entirely new system or theory; the gist of his arguments is that "No interval, or chord. ought to be judged of or treated according to its individual appearance alone, but according to the proofs of a regular connection," which is almost the same idea as that propounded by Rameau seventy years before. Our own countryman, William Shield ("Introduction to Harmony," 1800), does not progress a single step in a new direction, but Weber ("Versuch einer geordneten Theorie der Tonsetzkunst," Mainz, 1817) does advance, for he says that "The harmonical truths are by no means (as many have thought, or affected to think), as in a philosophical science, deducible from one independent superior principle, and subordinate one to another, as it were in a tabular manner," and therefore every combination is explained according to its individual appearance.

Harp. *Harpe (Fr.), Arpa (It.), Harfe (Ger.)* A stringed instrument of triangular form, furnished with gut strings. It has a compass varying from three to six octaves and a half, according to the size of the instrument. There are several kinds of harps still in use:

(1) The triple or Welsh harp, with three rows of strings, two rows tuned diatonically in unisons or octaves, the third or inner row arranged to supply the accidentals, sharps or flats. The strings are thin, and the tone is consequently consonant with the character of the strings. "It is simply impossible to modulate upon this instrument farther than to touch an occasional accidental from among the inner row of strings." This harp is derived from, and is almost identical with, the Irish harp, of which more will be said presently.

(2) The double harp (*arpa doppia*) with two rows of strings is less inconvenient but equally imperfect; all alterations of the pitch of the strings having to be made with the thumb.

(3) The single-action pedal harp, with one row of strings, containing a compass of nearly six octaves:

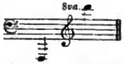

in the scale of E flat. There were seven pedals which altered the pitch of the note to which each pedal belonged, a semitone. The imperfection of the mechanism of the pedals involved the player in many difficulties, and rendered some keys perfectly useless.

(4) The double-action pedal harp, the invention of Erard. The compass of this instrument:

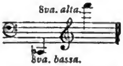

six octaves and a quarter, and the power of the pedals to change the pitch of each note two semitones, made it almost equal to the pianoforte in facility of modulation. This last-named harp is the one now generally employed for concert or orchestral purposes.

The invention of pedals to the harp has been variously attributed to Hochbrucker, of Donauwörth, in 1720, and to Paul Velter, of Nürnberg, in 1730. The harp in use before those times was generally tuned according to the key of the piece it was required in.

In the single harp passing modulations were made by stopping certain strings with the thumb, for continued change of key, all the notes in the new scale not in the ordinary tuning of the harp were made by turning the wrest pins during the progress of the music of all the notes required to be altered. Music for the harp is written on two staves. The instrument is capable of playing a melody with accompanying harmonies similar in character to pianoforte music. Successions of staccato chords have a fine effect upon the instrument, but broken chords (arpeggio) are better and more characteristic. The harp has been introduced into the orchestra with good effect, by several modern composers, Wagner and Gounod more especially. There is at present but little variety in the manner in which it has been used, arpeggios accompanied by violins playing on the high register *tremolo* with or without mutes, whenever angelic voices or exalted or heavenly ideas are intended to be presented.

The harp, like the guitar, may be traced to a very remote ancestry; it has, likewise, been subject to many vicissitudes of fortune. A favourite instrument with kings, it has also been the companion of beggars. Inspired strains have been sung to its strings, and it has accompanied verses neither pious nor inspired. By turns cultivated and neglected, it has never been wholly without a witness in the several ages of the history of music.

Engel, in "The Music of the Most Ancient Nations," says, that many Eastern nations have harps of different sorts, names and methods of stringing. The Burmese harp, called *saun*, has thirteen silken strings wound round a curved bar at one end in a way which admits of their being pushed up or down to be tuned. The harp is called *chang* in Persian, and *junk* in Arabic. The Negroes in Senegambia and Guinea call it *boulou* or *ombi*, and use strings made of fibre.

The harp in its primitive form is supposed to have been suggested by the warrior's bow. Many barbarous tribes preserve this form with some slight degree of variation. The "Nanga," or Negro harp is of this kind, as will be seen in the subjoined figure.

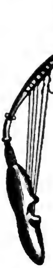

Nanga or Negro Harp.

It is said that some savage tribes still use their bows in times of peace as musical instruments. The harp of the Ossetes and Indo-Germanic tribe of the Central Caucasus have an instrument which supplies the connecting link between the form of the *Nanga* and the harps represented on ancient monuments. It is furnished with twelve strings of horse-hair, each string composed of six or eight hairs. It is about two feet in height, and has no forepillar. A specimen of this kind of harp is preserved in the South Kensington Museum.

The harp is mentioned in the authorised version of the Bible, Gen. iv. 21, "The father of all such as handle the harp and organ." The word in the original, "*kinnor*," appears in thirty-six other places in the Old Testament: in every case it is translated "harp." But while there is reason for believing that the Hebrews were acquainted with the harp, it is not certain that the word *kinnor* really means harp. The form of the *kinnor* is a matter of much uncertainty.

Kalkbrenner, who is considered as an authority on the subject of Hebrew music, in his Histoire de la Musique, Paris, 1802, makes no attempt to describe it, but merely contents himself with a commonplace remark after quoting the early writers who speak of it. "Le *kinnor*, en grec *kinnyra*, qui d'après la description de Saint Jerome, avoit la figure d'un △ et qui étoit monté de 24 cordes; l'autre description hébraïque du livre Schilte Haggeborin, donne au kinnor 32 cordes, tandis que l'historien *Joseph* et beaucoup d'autres ne lui en attribuent que dix; quelles contradictions!" These contradictions may be reconciled by supposing that the *kinnor*, like the Egyptian *bounà* (both of which words are translated "harp"), was of various sizes, variously strung.

There are numerous representations of harps upon the Assyrian bas-reliefs, from which it would appear that the instrument consisted of a slightly curved frame acting as a resonance body and as a stay for the strings at one end, while at the other they were secured to a horizontal bar. There was apparently a sort of tassel, supposed to be made of the unstretched ends of the strings.

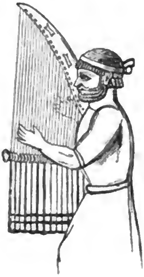

A slight difference of form may be seen in the various sculptures, but the general character of the Assyrian harp is the same as that shown above. It will be seen that there is no front pillar such as modern harps possess, by which it may be assumed that these harps were similar in construction to the Egyptian harp. The ancient Egyptian harps were of many sizes, and according to the representations preserved were strung with various numbers of strings, from three upwards. The player stood, knelt, or supported his instrument upon a stand while performing. Of the scale to which these harps were tuned, or of the manner in which they were kept in tune, no reasonable idea can be formed.

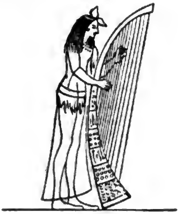

Egyptian Harp.

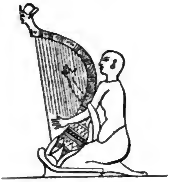

Egyptian Harp.

If they were kept in tension by the simple process of twisting with the finger and thumb round the horizontal bar, it is difficult to realise the idea that anything like a prearranged melody could be performed upon them. It is, moreover, probable that the greater part of the time of the performer was engaged in tightening the strings which his gripping had stretched. These views are grounded upon the assumption that the Egyptian and Babylonian artists have been as accurate in their delineations as they are said to be. But it was shown in the article "Guitar," that ancient artists are not always trustworthy as regards their pictures of lutes, for the drawings and sculptures often delineate a performer actively engaged in playing upon an instrument in a position in which it cannot possibly be supported. But admitting them to be correct in form, and acknowledging that the artist may have omitted to show the means by which the performer steadied his instrument, it is doubtful whether the ancient harps had either power or sweetness; for the ornament which overloads many of the depicted musical instruments must have greatly interfered with the tone.

The Greeks, who borrowed all their musical instruments, and, as some say, even the very names from foreign nations, must have known the harp, but they do not appear to have chosen it as a subject for representation, either in their paintings, sculpture, or pottery. The famous vase now in the Munich museum, dating from the time of Alexander the Great, offers one of the few instances in which it is depicted. The instrument is in form like the Assyrian harp.

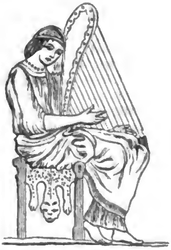

Grecian Harp.

The most favoured instrument in Grecian art was the lyre, which like the harp is described, in Greek literature, as having different numbers of strings, and to have been mounted occasionally on stands. The Pektis, the Simekion, and the Epigoneion, are supposed to have been actual harps. The first had only two strings, the others thirty-five and forty respectively. There was another instrument, the Phorminx, usually understood to be a lyre, which is sometimes translated harp.

The monuments of ancient Rome show very few examples of the harp in the forms familiar in those of Assyria and Egypt. The word "cithara," which is translated, harp, lyre, lute, guitar, &c., indifferently, offers no help in clearing away the cloud obscuring all knowledge of this matter. The claim of the Irish, asserted by Vincentis Galilei, in his "Dialogue Della Musica," 1581, to the invention of the harp cannot now be supported

Digitized by Google

since the discovery of the Assyrian remains, but there is no doubt that they were acquainted with it at a very early period of the world's history. They had four kinds of harp—(1) the *clar-seth*, *clar-scat*, or *clar-seach;* (2) the *Keir-nine;* (3) the *cionar-cruit;* (4) the *creamtine cruit.*

The first was brought to Ireland by the Celto-Phœnicians, the second was similar to the Eastern Kanoon, and like it, was played with plectra. The third had ten strings stretched over a resonance-body; and the fourth was the same as the Crwth, had six strings, four of which passed over a finger board, and could be stopped at the will of the player, the two others formed a drone. The clar-seth, also called teadloin (pronounced Tealoin, or Telin), was the instrument of the ancient Celts. The word *telin* is derived from *tel*, what is straight, stretched, or drawn tight.

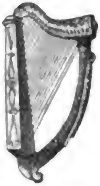

Irish Harp.

Vincentino Galilei (p. 143 of his "Dialogo della Musica Antica e Moderna," Florence, 1602), speaking of ancient instruments, says: "Ci è prima mente l'Harpa, la quale non è altro che un' antica Cithara di molte corde; se bene di forma in alcuna cosa differente, non da altro cagionatagli dagli artefici di quei tempi, che dalla quantita di esse corde & dalla loro intensezza; contenendo l'estreme graui con l'estreme acute piu di tre ottaue. Su portato d'Irlanda a noi questo antichissimo strumento (commemorato da Dante) doue si lauorano in eccellenza & copiososamente; gli habitatori della quale isola si esercitano molti & molti secoli sono in essa, &c." By which it will be seen that the Welsh were not the only people who adopted the harp from the Irish. The old Italian harp had two rows of strings, arranged after the manner of the old Irish harp, from which it is supposed to be derived. An instrument of this sort, an "Arpa doppia" is mentioned as one of the accompanying instruments in Monteverde's "Orfeo," 1607.

The harp was a favourite instrument among the ancient Britons. The old laws of Wales, the Triads, if they are to be trusted, specify the use of the harp as one of the three things necessary to distinguish a freeman or gentleman from a slave. Pretenders were discovered by their unskilfulness in "playing of the harp." The same laws forbade a slave to touch a harp, either out of curiosity, or to acquire a knowledge of it, and none but the king, his musicians, and other gentlemen were permitted to possess one. The harp was exempt from seizure for debt, as it was presumed that he who had no harp lost his position, and was degraded to the condition of a slave.

The harp was a familiar instrument with the Anglo-Saxons, as many references in existing chronicles prove conclusively. The harp-player was respected for his skill, whether he was known or unknown. Harp in hand he might wander scot and scathe free even in the camp of an enemy. Colgrin, the son of Ella, when besieged in York, about 495, received assistance from his brother, who went through the camp of the besiegers disguised as a harper; an artifice also adopted by King Alfred four hundred years later.

Bede states that it was the custom at festive meetings to hand the harp round for each of the company to sing and play in turn, and mentions the fact that Caedmon, the poet, had so neglected his studies in this respect in the pursuit of more serious knowledge, that at an assembly where he was present, and the harp was sent round, he being unable to play, rose from the table in shame, and went home to his house. "Surgebat a medea cœna, et egressus, ad suam domum repedebat," or as King Alfred has rendered it into Saxon, "Thonne aras he for sceome fram tham symle, and ham yode to his huse."

Performers upon the harp of special skill were notable before the reign of William the First. The services rendered by Taillefer, the harper, on the battle field are well known matters of history, as also is the foundation of the priory and hospital of St. Bartholomew in Smithfield, by Rahere, harper to King Henry the First.

Musicians were courted and respected in olden times, the harp was a sure passport everywhere, and a warrant of welcome in every society from the highest to the lowest, among kings or churls, among friends or foes. Where other men failed, the minstrel succeeded; admission to a house or castle was granted to a "gleeman," which was denied to all else. The existence of this privilege is frequently taken advantage of by writers of the old ballads, romances, &c., who often describe a luckless lover or chivalrous adventurer, gaining access to some giant's fortress or enchanted castle, in the garb of a minstrel or harper:

> "Horn sette him abenche
> Is harpe he gan clenche
> He made Rymenild a lay."
>
> *The Romance of Horn and Rymenild.*

Well-known historians and poets, besides the anonymous ballad writers of mediæval times, make frequent mention of the harp, thus showing its continued popularity. The introduction of the guitar made the harp less cultivated by private people, and as lutes, viols, and other more easily portable instruments became known, the harp was more and more disused. That which the guitar and viol had begun, the clarichord, virginals, and harpsichord completed, and the harp became rarely seen and still more rarely used. It never became wholly silent, but was to be heard in rural districts played in the same style, strung in the same fashion, and tuned after the same mode, as when it belonged to a more remote generation. It was a sort of a Rip van Winkle among instruments, living through many changes, yet unconscious of them. Older than all of the existing members of the great family of musical instruments by which it was surrounded, but uninfluenced and untouched by the progress towards perfection which all else were making. Before Hochbrucker's invention, the harp had been unchanged for centuries, remaining in the condition it had been left by the ancestors of the various peoples with whom it was found. When Handel's oratorio "Esther" (written in 1720) was produced at the request of the Duke of Chandos, harp parts to one of the choruses, for two Welsh players of the name of Powell were inserted by the composer. The harp they used was the old Welsh harp already described. Even after Hochbrucker and Velter had made their improvements, and rendered the harp more available as an orchestral instrument, it was rarely employed as such, all harp effects being made by *pizzicato* playing on the violin.

Hoyle ("Dictionarium Musicæ," 1770) describes the harp as having, in his day, "three rows of strings, which in all make seventy-eight; the first row contains twenty-nine, which makes four octaves, the second row makes the half turn: the third is unison with the first row. It's musick is like that of a spinet, all the strings going by semitones, and is played on with both hands, by pinching them in the same manner together. Some give it the name of the inverted spinet." As this style of stringing is nearly the same as the Irish method of tuning in use ages before, it is difficult to believe Hoyle's statement that the "moderns have much improved" the harp. The name of "inverted spinet" is singular, as one of the names for the spinet was the "Couched harp." The harp was not popular in those days, it required closer application to master its difficulties than the musical young people of the time cared to devote to it. One instrument of the harp kind, called the "Bell-harp," was in constant use in the time that Hoyle wrote, and has not completely disappeared in the present day. No great skill is required to perform upon it, all that is necessary is prehensile power and strength of wrist. The instrument is a wooden box a little more than two feet long, all closed except over the wrest pins, and a space near the opposite end. It is strung with steel wire, with eight strings or more, tuned with the lowest notes outside:

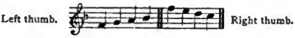

The fingers of each hand grasp the body of the box, leaving the thumbs free to strike the strings. The player swings the instrument as he strikes, producing the effect of the sound of a peal of bells borne on the wind.

To return to the main subject; when, in 1820, Erard improved the harp, it became suddenly and widely fashionable; nearly all the music published was "arranged for the pianoforte or harp." The instrument was as indispensable in the drawing-room of those calling themselves "musical," as the pianoforte is now. It was chiefly played by ladies, especially by those who possessed the advantages of a pretty hand and arm, not to mention a "neat foot," all these motive powers being well shown in harp playing. As the fair performers grew old, the charms of the harp decayed, and although the instrument is still played and taught, it is not cultivated to the extent which its merits might seem to warrant.

The derivation of the word is a matter of doubt, none of the earlier terms supplying the least etymological link. Du Cange, in his "Glossary," asserts that the harp takes its name from the Arpi, a people of Italy who were the first who invented it. In this he is not quite correct, as will be gathered from the foregoing account. Arpi was a very ancient city, without doubt, it was of a higher antiquity than Livy, who mentions it as an Apulian city. It is not at all improbable that the people of Arpi may have been the first to adopt the instrument upon its introduction into Italy, and as it is not at all unlikely that the Northern people adopted it from the Romans, the name may have been transmitted in its present form from the Arpi, the people

Digitized by Google

whom the Romans may have believed to have been the inventors. That the word has a common origin, its similarity in most modern tongues may tend to prove. There is evidence enough in music to show that the name of a people may give a permanent title to a composition, and also to an instrument. There is no apparent connection between the words Kinnor, Bouni, Saun, Chang, Junk, Boulou, Ombi, Nanga, Pektis, Simekion, Epigoneion, Phorminx, Cithara, Clarseth, Telin, and Harp, the names by which the instrument has been called by various peoples; but the words Earpe, Harp, Harpe, Harfe, and Arpa point conclusively to one origin, and Du Cange may be right with regard to one part of his statement, for it is not at all unlikely that the harp, when brought to and adopted by the more northern nations, would be called by the name given to it by the people from whom it was received.

Harpe (*Fr.*) A harp.

Harpist (*Eng.* and *Ger.*) A player upon the harp.

Harp-Lute. An instrument of the last century with 12 strings, never much used.

Harpsecol. Harpsichord.

Harpsichord. *Clavicembalo*, *Cembalo* (*It.*) *Clavecin* (*Fr.*) *Arpichord* (*Ger.*) A stringed instrument with a key-board, similar in form to a Modern grand pianoforte. As the pressure of the fingers upon the keys, when heavy or light, made no difference in the quantity of tone produced, the harpsichord sometimes had two key-boards, one for the loud, the other for the soft tones. There were also stops in some instruments, by means of which the tone could be modified by connecting the mechanism with or detaching it from the three or even four strings with which each tone was furnished. The keys were attached to levers, which at their ends had slips of wood, called "jacks," furnished with plectra of crow-quill or hard leather; these struck or twanged the strings and produced the tone, which has been likened to "a scratch with a sound at the end of it." [Clarichord.] [Spinet.] [Pianoforte.]

Harpsichord graces. Certain turns and ornaments employed in playing upon the harpsichord, introduced for the most part as compensation for the lack of sustaining power in the instrument. They were called by the general term of *agrémens* in French, and *Manieren* in German. The chief of these ornaments were the following, other signs are explained according to alphabetical order.

Turn or *Doublé* (*Fr.*), *Doppelschlag* (*Ger.*)

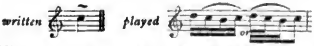

When the sign was placed upright it was understood to signify that the turn was to be commenced a note lower than the one over which it was written:

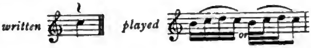

When the double was preceded by a small note or notes on the same line or space, the turn was to commence from that note:

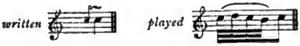

When the marked note was tied, the turn was to be commenced from the preceding note:

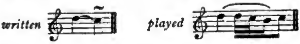

Battery, *Cadence* (*Fr.*), indicated by the sign +. When the cadence was written thus:

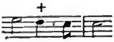

it was played:

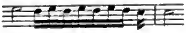

and was called a full Cadence or *Cadence pleine;* when it was written:

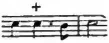

it was played:

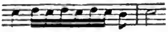

and was called *cadence brisée*, or broken cadence.

Sliding trill, *Flatté* (*Fr.*), *Schleifer* (*Ger.*):

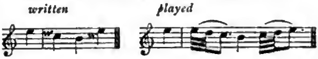

Or occasionally:

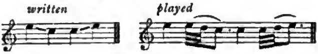

Trill; *Tremblement* (*Fr.*), *Pralltriller* (*Ger.*), expressed by the signs *tr*, ✓, ⁓ ∿. It was in effect a trill without the final turn, unless altered by other signs. There were six sorts of *tremblements;* namely, (1) the simple tremblement (*tremblement simple*)

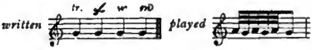

(2) The doubled or turned trill (*tremblement doublé*):

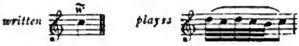

Digitized by Google

(3) The detached trill (*tremblement détaché*), which was performed when the note which should begin the trill had a place in the melody preceding the marked note:

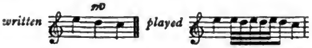

(4) The tied trill (*tremblement lié*) was almost the same at the preceding, only that the first two notes were tied:

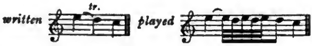

(5) The prepared trill (*tremblement appuyé* or *préparé*) was when a slight pause was made before commencing the shake, and its speed gradually increased:

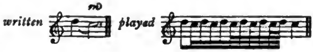

(6) A slided trill (*tremblement coulé*) when the shake was preceded by a slide:

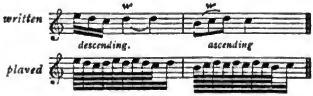

Haupt (*Ger.*) Principal, chief, head; as *Hauptnote*, the essential note in a turn or shake, &c.

Hauptmanual (*Ger.*) The great organ.

Hauptnote (*Ger.*) An essential note.

Hauptperiode (*Ger.*) Principal subject.

Hauptsatz (*Ger.*) Principal theme or subject.

Hauptschluss (*Ger.*) A perfect cadence.

Hauptstimme (*Ger.*) Principal part.

Hauptwerk (*Ger.*) The great organ.

Hausse (*Fr.*) The nut of a violin bow.

Hausser (*Fr.*) To raise, lift, to sharpen.

Haut (*Fr.*) High; as, *haute-contre*, the alto part; *haut-dessus*, treble part; *haute-taille*, first or high tenor.

Hautbois (*Fr.*) [Oboe.]

Hautbois d'amour (*Fr.*) A small species of oboe, now obsolete. Music written for it can be played on the ordinary instrument.

Hautboy. [Oboe.]

H dur. (*Ger.*) The key of B major.

Head. (1) The membrane stretched upon a drum. (2) That part of a violin or other stringed instrument in which the pegs are inserted. (3) The portion of a note which determines its position upon the stave, and to which the tail is annexed.

Head-stall. A head-band or *Capistrum* q. v. The use of the word *stall* in the sense of *bandage* is still to be found in our word *finger-stall*.

Head-voice. The sounds produced above the chest-register, but not in falsetto.

Heftig (*Ger.*) Boisterous, impetuous.

Heiss (*Ger.*) Hot, ardent.

Heiter (*Ger.*) Clear, calm. *Feierlich heiter*, solemn and calm.

Helicon (*Gk.*) ἑλικών. An instrument used by the Greeks in the calculation of ratios.

Hemidiapente. An imperfect fifth.

Hemiditone. The lesser or minor third.

Hemiolios (*Gk.*) ἡμιόλιος. (1) The ratio 3 to 2. The same as the *sesquialtera* in Latin treatises on music. (2) A kind of metre. A verse consisting of a foot and a half.

Hemiopus (*Gk.*) ἡμίοπος. Having a small number of holes. ἡμίοποι αὐλοί, small flutes with three ventages.

Hemitone. A semitone.

Heptachord. (1) A series of seven notes. A diatonic octave without the upper note. (2) An instrument with seven strings.

Herabstrich (*Ger.*) The down stroke of the bow in playing stringed instruments. A down-bow.

Herstrich. [Herabstrich.]

Hexachord. A series of six sounds. [Notation.]

Hey de Guise. A country dance.

> "In our antique hey de guise we go beyond all nations."
> *Middleton.*

[Country dance.]

Hidden Fifths, or **Octaves.** [Consecutives.]

Hinaufstrich
Hinstrich } (*Ger.*) An up-bow.

His (*Ger.*) The note B sharp.

Hissing. A manner of showing dissatisfaction. [Applause.] [Fiasco.]

H moll (*Ger.*) The key of B minor.

Hoboe. [Oboe.]

Hoboy. [Oboe.] [Waits.]

Hocket, Hoket, Ochetus. A species of mediæval music in which two parts were so arranged that while one had a note, the other had a rest. "Discantus species in quâ dum unus cantat, alter tacet, et e contrario." From the manner in which these parts were broken up, this device was also called *truncatio* ("truncatio, vel oketus, quod idem est"). It was in common use in the 14th century.

From a MS. in the Bodleian, circ. A.D. 1380.

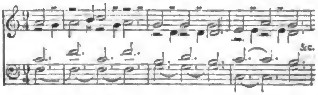

Digitized by Google

This term, and the Form it designates, are now obsolete.

Hochzeitsmarsch (*Ger.*) A wedding march.

Hohlflöte (*Ger.*) [Flute.]

Hold. An old English name for the sign of a pause 𝄐.

Holding note. A note sustained in one part, while the others are in motion.

Homophony (*Gk.*) Unison of voices or instruments of the same character. ὁμόφωνος.

Hopper. [Pianoforte.]

Hopser or **Hopstanz** (*Ger.*) A country dance.

Horæ Canonicæ. The canonical hours at which religious services are held. In time of persecution a night-service was held called *Nocturns*, which was, however, at a later period merged into *Lauds*—the thanksgiving for the dawn of day, and the whole was called Matins. But when Matins, or Nocturns, are retained they take place before *Lauds*. The following is the order of the Canonical hours: *Lauds*, at daybreak; *Prime*, or first hour, a later service; *Tierce*, or third hour, at nine a.m.; *Sext*, or sixth hour, at noon; *Nones*, or ninth hour, at three p.m.; *Vespers*, or evening service; *Compline*, or final service, at bed time. Each of these has fixed Psalms, except Vespers, which has certain Psalms read in course, and a Canticle. The Anglo-Saxon names of these services were, Uhtsang, Primesang, Undersang, Middaysang, Noonsang, Evensang, and Nightsang. The book for these Offices came to be called Breviary towards the end of the 11th century. It was sometimes called in England the Portifory.

Horn (Cor, *Fr.;* Corno, *It.;* Horn, *Ger.*) A metal wind instrument, formed of a continuous tube twisted into a curved shape for the convenience of holding. It is furnished with a mouthpiece and a bell. The mouthpiece is moveable, so as to allow additional pieces of tubing called crooks, to be added to its length in order to alter the pitch, and the bell is sufficiently wide to admit the hand of the player.

The horn sounds the harmonics of the fundamental note of its tube. That is to say, a horn in C, or sounding C as a generator has the following compass:

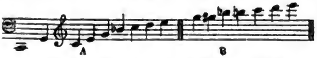

Notes written in the bass clef sound as they are written, those in the treble clef, an octave lower.

The notes in the section marked B are more difficult to produce than those in the section marked A.

The whole of the above notes are called the "open" notes; that is to say, they can be produced by simple blowing, and certain modifications of the form of the lips. There are other sounds called "closed," which are obtained by closing the bell in a certain manner with the hand.

The closed notes are less free in quality than the open notes, and they are therefore only employed when especial effects are required. Horns are generally employed in pairs, and the parts are usually written in the treble clef in the key of C. The most useful parts of the compass of the horn are those marked in section A. The "crooks"—alluded to above—serve to alter the pitch by lengthening the sounding tube, and the composer indicates the horn or horns required at the beginning of a piece of music thus: Corni in C; Corni in B flat; Corni in A, &c., though the parts are written as though they were to be played in C, the presumed normal key of the horn. All horns other than those in C are called transposing instruments, because the actual sounds produced are different to those represented on paper. And strictly speaking, the C horn is also a transposing instrument, because the notes are written an octave higher than they are played Thus the passage:

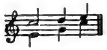

would sound differently according to the horn employed:

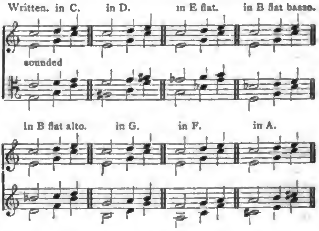

If in D, it would be a seventh; in E flat a major sixth; in E a minor sixth; in F a fifth; in G a fourth; in A a minor third; in B flat alto a second; in B flat basso, a ninth lower than the notation.

In movements in a major key, in which four horns are used, two are generally in the key of the tonic, two in that of the dominant Sometimes, two are in the tonic key, two in that of the mediant.

When horns are required for minor keys, one is set in the key of the relative major, the

Digitized by Google

other in the tonic major. Thus in G minor, one horn would be in B flat, another in G. Composers, however, vary in their use of horns in minor keys, so that no definite rule can be laid down.

There are other kinds of horns used in the orchestra, which, by means of valves, pistons, or ventils, are capable of producing the "open" and "closed" sounds in a scale of semitones throughout the compass, of very nearly even quality, and with comparatively little difficulty to the player.

The horn, sometimes called the French horn, to distinguish it from the English horn (cor Anglais), which is altogether a different instrument, was introduced into the orchestra towards the end of the 17th century, and as it became very fashionable and popular in the early part of the 18th century, it is frequently found in the scores of orchestral compositions, and is also employed as an accompaniment to hunting songs. A horn-player, named Spandau, was one of the first who discovered the art of producing the closed or stuffed notes (*sons étouffés*) on the instrument. Sir John Hawkins makes a note in his "History of the performance of a Concerto in 1773," of this player, "part whereof was in the key of C with the minor third." The "improvement was effected by putting his right hand into the bottom or bell of the instrument, and attempering the sounds by the application of his fingers in the different parts of the tube;" by which means "the intervals seemed to be as perfect as in any wind-instrument."

The earliest attempts to make the scale of the horn perfect are attributed to Hampel, of Dresden, in 1748. Twelve years later Kölbel invented his amor-schall, which was a horn with *ventils*. Claggett, a Londoner, attempted further to improve these plans; but all these inventions were forgotten or laid aside, in favour of Müller's system of ventils, first proposed about the year 1830. These in time were superseded by the inventions of J. G. Saxe and others. Although the horn has been so far improved, that the difficulties of producing the tones intermediate to the harmonics have been considerably lessened, it cannot be said that any great improvement has been effected in the quality of the added sounds, and the old French horn, the plain tube, whose sounds are "attempered by the fingers of the player," is still unsurpassed in brilliancy of tone by any of the modern inventions.

Hörner (*Ger.*), **Corni** (*It.*) Horns; *Hörner in Es*, horns in E flat.

Hornpipe. (1) The name of an old wind instrument of the *shawm* or *waits* character, receiving its name from the fact that the "bell" or open end was sometimes made of horn. In Wales, Ireland, and Cornwall, in Brittany, it was called pib-corn. "Pib" or "piob," meaning pipe, and "corn," horn. Chaucer uses the word to describe an instrument of music in his translation of the "Romaunt of the Rose."

> "Controve he wolde, and foule fayle
> With hornepipes of Cornewaile."

The original (written in 1260) is:

> "Une hore dit lés et descors
> Et sonnez doux de controvaille
> As Estives de Cornoaille."

The word "controve" comes from the same root as that from which "Trovere" or "Troubadour" is derived, and means to invent, or compose, so that the passage translated into modern English would be:

> "He would compose and play songs (lés)
> To the accompaniment of the hornpipe."

That the hornpipe was harsh-sounding may be gathered from the context in Chaucer when he says in the preceding lines:

> "—— yit would he lye
> Discordaunt ever fro armonye
> And distoned from melodie."

In some editions of the "Romaunt of the Rose," the words "chalemeaux de Cornouaille," stand instead of "estives de Cornoaille." That the Waits or Shawms had reeds is well known, and that the hornpipe, its rustic relation, was not likely to have a softer tone may be assumed. Robert Bell supposes that Chaucer "wrote corn-pipes, which," says he, "would be the proper translation of Chalemeaux." Thus Virgil:

> "Musam meditaris avenâ."

And Shakespeare:

> "And shepherds pipe on oaten straws.'

Others imagine the word "estive" to come from Æstas, the hot season, whence our word *Estival* for summer, and the word "Chalumeau" employed as a synonym for "estive" to be derived from "chaleur," heat. That the "estive," "Chalumeau," or "hornpipe," is "a shepherd's pipe made of corn which, growing in summer, furnishes the material for the instrument." These theories are, to say the least, ingenious. It is, however, much more in accordance with probability to infer that the word "Chalumeau," or "Chalemeaux," comes from *Calamus*, a reed; which is only similar to a straw in being hollow, and that the word "corn" has no connection with either reeds, straws, or oats. The word "corn" comes from "κέρας" through "cornu," and as the hard "c" is only one form of the guttural sound sometimes represented by the letter "h," the words "corn" and "horn" are the same when applied to a musical instrument. It may not necessarily be formed

Digitized by Google

of horn or have horn in its composition, but may receive its name from being bent or curved as a horn.

The horn-pipe was probably the same kind of instrument as that now known as the "Cor Anglais" or English horn. The words corn, cornet, cornamusa, &c., are applied to different sorts of wooden wind instruments with reeds. The *hoboy*, *shalm*, *shawn*, *wait*, or *reed-pipe* were common instruments in days gone by, and often furnished the melodies to which dancing was performed. Therefore the supposition that the "hornpipe" gave the name to the dance so called may not be very far from the truth.

(2) A dance of English origin, so-called from the instrument which played the tune to which it was danced. The measure or rhythm of the Hornpipe is as varied as that of the tunes for country dances, therefore all descriptions of the dance, which are founded upon the supposition that there is only one form of rhythm, are misleading. One writer says, "It is in compound triple time of nine crotchets in a bar, six down and three up;" another says, "It is in triple time of six crotchets in a bar, four whereof are to be beat with a down, and two with an up hand;" another says, that it is in "common time of four crotchets played lively;" and another says, "It is always in double triple time of two parts of four and eight bars each with repeats." These descriptions tally with the character of the measure of the dance which was called hornpipe at the time the description was written. But the hornpipe is, and has been, often danced to tunes called by other names, such as Brawl, Coranto, Canaries, Fandango, Jig, Loure, Reel, Matelotte, &c., the steps being varied to suit the several measures. The hornpipe now-a-days is danced by a single performer, to a tune in duple measure. The performer generally stands still with arms folded during the time the first phrase of the melody is played, and then commences the dance by running round in a circular direction; and when dancing a hornpipe called by some special title, such as "The Sailor's," "The Jockeys," "The Countryman's," or "The Monkey's" hornpipe, generally introduces some gestures characteristic of the distinguishing title.

The following are some specimens of the various sorts of melodies of hornpipes:

HORNPIPE.

Mr. Jerry Clark, 1700.

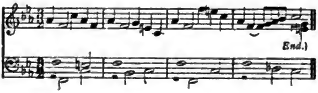

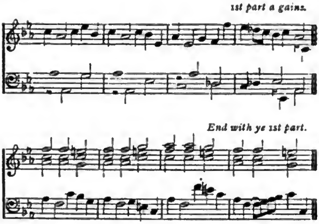

RAVENSCROFT'S HORNPIPE (1730).

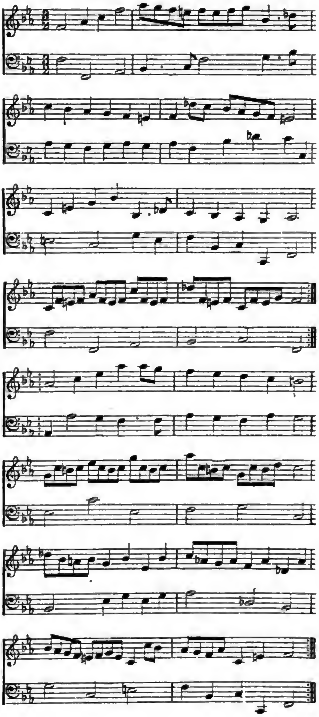

THE COLLEGE HORNPIPE.

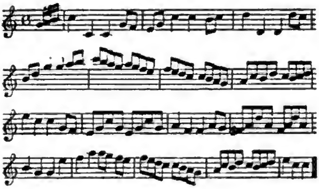

MISS CATTLEY'S HORNPIPE.

PERFORMED AT SADDLER'S WELLS, 1780.

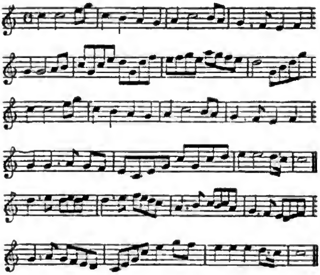

JACK AT GREENWICH.

C. DIBDIN, 1800.

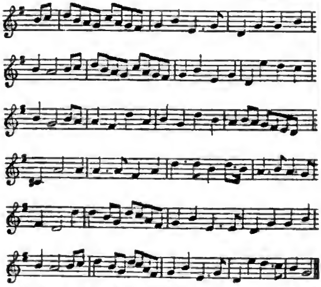

Hosanna. A part of the Sanctus in the Mass. [Mass.]

Huchet (*Fr.*) A postman's horn.

Hülfslinien (*Ger.*) Leger lines.

Hülfsnote (*Ger.*) An auxiliary note.

Hümmelchen (*Ger.*) A kind of small bagpipe. A drone. The word *Hummel* is cognate with our *bumble;* see Bombaulius, a bagpiper.

Hunting Horn. *Cor de chasse* (*Fr.*) *Corno di caccia* (*It.*) *Hiefhorn* (*Ger.*) [Horn.]

Hunting. A term in change-ringing. There are two kinds, *hunting up*, and *hunting down.* The first is performed on any number of bells, "by pulling after the bell which pulled last after you;" the second is accomplished by looking out for which bell strikes first of the set, and alter his number in the set at each change until he has gone through the full number of the bells, and then return to his original place. Thus, if he be first, he will next be second, then third, then fourth, fifth, sixth, and so on, according to the number of bells, and return backwards in counting sixth, fifth, fourth, &c.

Hunts-up. A morning song, of a noisy boisterous character, afterwards applied to any unseemly clamour at any time of the day

> "I never was yet
> At such a hunts up."
> MASSINGER, *The Duke of Milan.*

There is an old song of the time of Henry VIII. called "The King's Hunts up," the tune of which is lively and spirited. The term "hunts up" may have been derived from the noise of horses, dogs, and men setting out on a hunting expedition. Butler, in his "Principles of Musick" (1636), defines a "*hunts up*" as morning music; as does Cotgrave (1660), "Resveil, a hunts-up, a morning song for a new made wife." This last is an allusion to the custom of performing an *aubade* before a bride's window, thus mentioned by Gay:

> "Here rows of drummers stand in marshal file,
> And with their vellum thunder shake the pile,
> To greet the new made bride."

Hurdy gurdy. An instrument of ancient origin, popular among most of the European nations under the names of *Leyer* or *Bauernleyer.* It consists of a flat oblong sounding board, upon which are stretched four gut strings, two of which are tuned a fifth apart to form a drone bass, and placed where they cannot be acted upon by the ten or twelve keys fixed upon one side of the belly of the instrument; the other two are tuned in unison, and are so arranged that they may be shortened by the pressure of the keys. The instrument is thus capable of sounding a scale of ten or twelve notes. The strings are set in vibration by the friction of a wooden wheel charged with rosin and turned by means of a handle at one end. The hurdy gurdy is only adapted to the production of melodies

Digitized by Google

of the simplest character. The old name for the hurdy gurdy (vielle) has misled many authors with regard to the derivation and description of the instrument. Dr. Burney says, "That the instrument called a *rote*, so frequently mentioned by Chaucer, as well as by the old French poets, was the same as the modern *vielle*, and had its name from *rota*, the wheel with which its tones were produced."

The word *rota* is simply a Latinised form of the word *crwth*, *chrotta*, *rotta*, which was originally a sort of cithara or guitar, played by plucking the strings with the fingers.

In after time when an instrument of similar shape was played with a bow, it still retained the ancient name *crwth* or *crowd*.

Hurry. The technical, theatrical, or stage name for a tremolo passage on the violin, or a roll on the drum. The "hurry" is generally played as a preparation for the culminating point of a dramatic incident, the leading to a "picture," during stage struggles or like exciting actions.

Hurtig (*Ger.*) Nimble, quick, agile. *Allegro.*

Hydraulic Organ. [Organ.)

Hymenaion, ὑμέναιος (*Gk.*) An epithalamium or marriage song.

Hymn, ὕμνος (*Gk.*), *Inno* (*It.*), *Kirchenlied*, *Kirchengesang* (*Ger.*) A lyric for use in worship.

Hymn tunes. The keen attention paid, in our time, to all forms of Church music, particularly to such as come under the term "Congregational," renders doubly interesting the distinction of style which can be traced in the "Corale," as we find it at different periods, and in the different churches of Christendom. We shall probably, from year to year, become better informed of the historical relations between sacred and secular music, especially as our method of translating into modern notation the neumas of the manuscripts of the 10th and 11th centuries becomes more certain and less open to doubt. There is in the library of the "Ecole de Medicine," of Montpelier, a MS. "Horace," in which the ode to Phillis is noted to the melody assigned by Guy d'Arezzo to the Hymn "Ut queant laxis." "Was the melody," asks a modern critic, "written for the ode, or for the hymn?" In close connection with the reply to this question is the inquiry of which it may be considered a part, "To what extent did the growth of musical measure or rhythm influence the manner in which the mediæval hymns were recited or sung?" It was the only part of the ancient church office in which this characteristic of "measure" was felt, and the form of such melodies as the following (both from the "Directory" of Guidetti) seems to point the "style of transition" from unmeasured Plain Song to measured music.

No. 1.

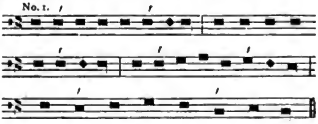

No. 2.

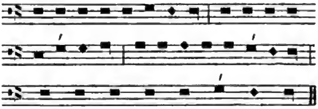

In No. 1 there is nothing but the "monotone" of the ancient chant in phrases 1 and 2, while every phrase of No. 2 commences with the same, and ends with the simplest inflection by way of melody. The strong "measure" or accent of both is noticeable—the latter would fall naturally into the following chant form:

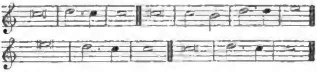

The Latin hymn tune, dating from about the 4th till the end of the 12th century, and as it appears in the office books in present use, must be considered as the first distinct "style" for the attention of the student. The distinctive characteristics of it are: 1. The scales to which its melodies belong—not the modern major or minor modes of the diatonic scale, but the so-called "modes" of the Church [Plain Song]. 2. The peculiarities of their rhythm, partly derived from the metrical flow of the verse itself, partly from the continually advancing sense of accent, or the alternation of "heavy and light."

It is a matter of conjecture whether in the great majority of cases the author of the hymn was the composer of the melody—a point which it will always be impossible to verify or to deny. It is certain that many of them are truly beautiful; as *e.g.* the following, from the Salisbury Hymnal:

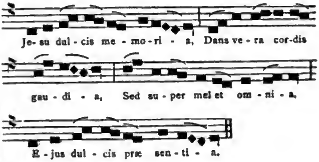
Je-su dul-cis me-mo-ri-a, Dans ve-ra cor-dis
gau-di-a, Sed su-per mel et om-ni-a,
E-jus dul-cis præ-sen-ti-a.

Or the following, from Guidetti :

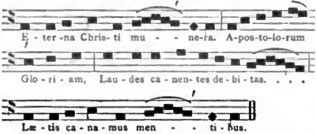
Æ - ter - na Chris - ti mu - - ne-ra. A-pos-to-lo-rum
Glo - ri - am, Lau - des ca - nen - tes de - bi - tas. . . .
Læ - tis ca - na - mus men - - ti - bus.

The recent publications of the Benedictines of the Abbey of Solesmes have thrown much new light on this subject. The student should read the *Paléographie Musicale* with its interesting collection of fac-similes and learned dissertations thereon. From the same source have been issued the Gradual, Antiphonary (2 vols.), and other works, all of which are of value. It is highly probable that ancient Latin melodies will be more largely used in our churches than formerly; it is therefore more important than ever that their most authentic forms should be discovered.

But we have already in our hymnals a large number of old Latin tunes which are often not recognised as such, owing to the regularity of their rhythm, *e.g., Conditor alme, Stabat Mater, Veni Creator,* and others. English adaptations are not satisfactory.

The period of the Reformation was one of great musical activity. For some time previous a quiet and half unconscious preparation had been going on, in the circulation among the people of popular hymns and their tunes in the vernacular, and of course only sung in private; and when Luther furnished his followers with metrical versions of the Creed, the Lord's Prayer, and other portions of the old Office Books, there were no bounds to the popularity of both hymns and tunes. Edition followed edition of books under the title of "Geistliche Lieder," &c., and in their multiplication many slight changes occurred in the original melodies; even in some that by the popular voice are accredited to Luther himself, and even during his lifetime. As it has become the custom in Germany for each diocese to be provided, by authority, with its own "Gesang-Buch," the *variations* of these popular melodies have themselves become stereotyped, as it were, in different localities, so that the original is often difficult to trace. The great Sebastian Bach has often harmonized different forms of the same melody, and has even (so it seems) not hesitated to alter it, if it suited the counterpoint of his accompanying "parts" to do so. The place occupied to this day by these melodies of the 16th century in the heart of the German people is perhaps only just less in reverence than that due to Holy Writ. Some of them were intimately associated with the occurrences of Luther's life—the following, for example, "Ein feste Burg ist unser Gott," which is probably his composition :

EIN FESTE BURG IST UNSER GOTT.*

M. LUTHER, 1529.

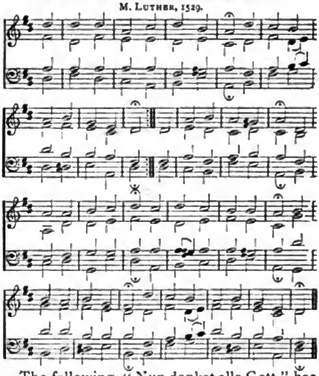

The following, "Nun danket alle Gott," has become the "National Hymn," and on occasions of popular rejoicing is often sung, in the open air, accompanied by all available force of military bands, and by the discharge of a cannon on each note of the melody. It was the composition of John Crüger, from 1622 to 1662 Music-Director in the St. Nicholas Church, Berlin.

NUN DANKET ALLE GOTT.

JOH. CRÜGER, 1649.

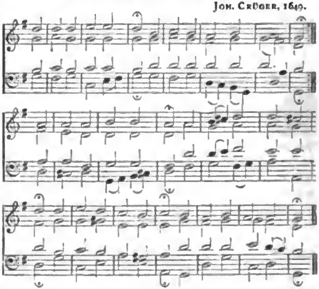

As an example of the alterations in melodies referred to, attention may be called to the various readings here given of the melody known in England as "Luther's Hymn;" with the authorities for each.

---

* This is the melody treated so variously, and with such consummate mastership, in Meyerbeer's "Huguenots."

VARIOUS READINGS OF THE MELODY KNOWN IN ENGLAND AS "LUTHER'S HYMN," WITH THE AUTHORITIES FOR EACH.

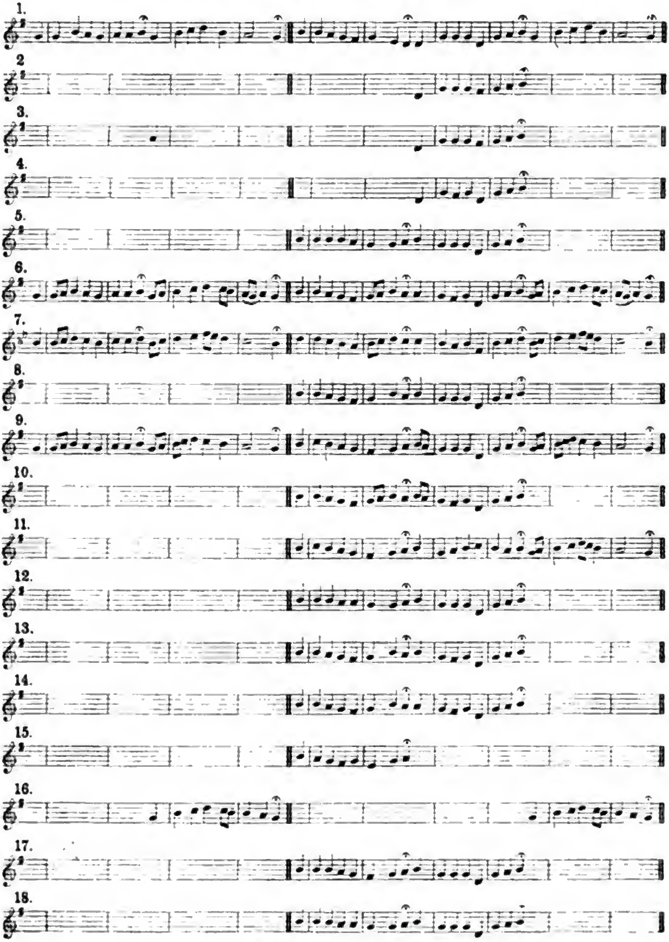
1.
2
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.

Digitized by Google

19.

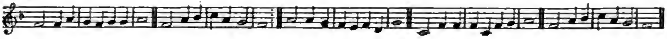

20.

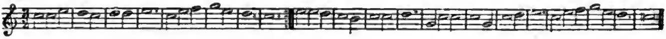

21.

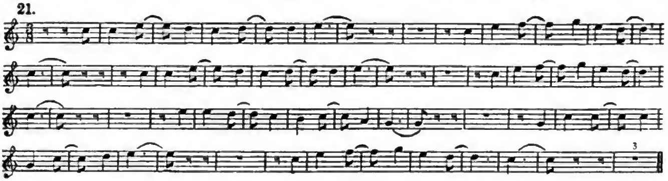

1. Is the original melody, ascribed to Luther, from the "Klug'schen Gesangbuch," 1535.

2. "Geistliche Lieder," Wittemberg, 1545 (Luther died 1546). "Geistliche Lieder," Leipzig, 1557. "Geistliche Lieder D. Martin Luther," Nuremberg, 1570. "Kern des Deutschen Kirchengesangs," Lairiz, Nördlingen, 1854. "Zion's Harpe," Kocker, 1855. Tucker, Stutgart, 1854. D. W. Volkmar, Erfurt and Leipzig, 1853. Winterfeld's edition of "Luther's Corales," 1840.

3. "Geystliche Lieder," Nürmberg, 1558.

4. "Ein schon geistliche Gesangbuch, darinnen Kirchen Gesange und geistliche Lieder, D. M. Lutheri und anderer frommen Christen, so in den Christlichen Gemeynden zu singen gebrauchlich begriffen." Durch Melchiorem Vulpium, Cantorem in Weymar, 1609.

5. From Graun's Oratorio, "Der Tod Jesu." Followed by Sir Henry R. Bishop, in his "Twelve Corales," as performed from time to time under his direction at the "Ancient Concerts," dedicated to H. R. H. Prince Albert.

6, 7. From J. S. Bach's, 371 Vierstimmige Choralgesänge: Ed. C. F. Becker, Organist an der Petri Kirche, Leipzig, 1831. No. 6 is also to be found in his "Choralvorspiele," Griepenkerl's Ed., Peters, Leipzig, in the key of G (vol. vi., p. 36). See also J. S. Bach's "Choralgesänge und Geistliche Arien, herausgegeben von Ludwig Erk." Peters, 1850.

8. "Evangelische Gesang Buch," Elberfeld, 1843. "Chorale Book," Heinrich Wächter, organist in Wiburg; Halle, 1856. "Chorale Book," Adolphe Hesse, Breslau, 1840.

9. "Chorale Gesang Buch," C. H. Strube, of Wolfenbüttel, 1841. "Wirtemberg Choral Book," 1841. "Stutgart Choral Book," 1857.

10. "Chorale Book." John Augustus Bechel, Leipzig, 1847. "Sixty-one Choral Melodies to Paul Gerhard's Hymns;" Becker, Leipzig, 1851.

11. "Chorale Book," Adolphe Hesse, König. Preus. Musick-direktor und Ober-Organist, Op. 69; Breslau, 1851.

12. "Evangelische Chorale Buch." Adolph Bernard Marx, Dr. and Prof. in the University of Berlin. 1832. "Chorale Book," dedicated to the Queen of Prussia, D. H. Engel, Berlin, 1844.

13. "460 Choral-Melodien." C. Karow, of Bunzlau-in-Schlesien; Dorpat, 1848. "Alte und newere Chorale-Melodien," C. G. Schramm, of Halle; Leipzig.

14. "Evangelische Choral Melodien Buch." A. G. Fischer, Musik-Direktor und Dom-Organist zu Merseburg; 1846-7.

15. From Martini's "Ecole d'Orgue," dedicated to the Empress Josephine; Paris, 1804. (?)

16. The modern form of the melody: supposed to have been introduced by the tenor singer Braham, who also interpolated an imitation of the trumpet of the "Dies Iræ." This form is also known in France (Chants Chrétiens; Paris, 1857, *et ante*).

17. From "Choral-Melodien for Hanover, Lüneburg and Hildesheim:" Heinrich Enckhausen: Adolph Nagel, Hanover (Modern).

18. Choral-buch, von August Haupt, Organist der Parochial Kirche, in Berlin. Berlin: Schlesingerschen Buch und Musikhandlung: Robert Lienau.

19, 20. Both said to be of 1535, from Winterfeld's "Der Evangelische Kirchengesang," 3 vols. quarto; Leipzig, 1843—a work to which reference may be made on the whole subject of German Hymnody.

21. From the "Introduction" to the "Woman of Samaria," by Sir Sterndale Bennett. Here inserted as an example of the Melody, really in *Common Tune*, driven

against and in combination with a movement in *Triple Measure.*

Among the early labourers in this field should be mentioned two French composers, Claude Goudimel and Claude Le Jeune, of whom the former published "Les pseaumes mis en rime française, par Clément Marot et Théodore de Bèze. Mis en Musique à quatre parties par Claude Goudimel." Without name of place, but with these words at the foot of frontispiece: "Par les heritiers de François Jaqui, 1565," 12mo. The melody is in the tenor (see below.) A second edition was published at Geneva, in the same year; a third at the same place in 1580; and another at Charenton, in 1607.

By the latter, 1. Do de ca corde contenant douze pseaumes de David, mis en musique selon les douze modes approuves des meilleurs autheurs anciens et modernes, à 2, 3, 4, 5, 6 et 7 voix, par Claude Le Jeune, compositeur de la Musique de la Chambre du roy. A La Rochelle par Hierome Haultin, 1598, 6 vol., in 4° obl. The words are from Clément Marot. The form of composition is that of the motets of the early Italian masters. Other editions of this work appeared at Paris in 1608 and in 1618. 2. Premier livre contenant cinquante Pseaumes de David, mis en Musique à 3 parties, par Claude Le Jeune, natif de Valenciennes, compositeur en Musique de la Chambre du roy, Paris, 1607, 3 vol., in 4° obl. The second and third books were published in 1608. No second edition of the work is known. 3. Les psaumes de Marot et de Théodore de Bèze, mis en Musique a 4 et 5 parties, par Cl. Lejeune, natif de Valenciennes. La Rochelle 1608, 4°. Première edition, publiée par Cécile Lejeune, sœur du compositeur, et dediée au duc de Bouillon, prince de Sedan. This edition is very rare. A second was published at Paris, 1613; a third at Geneva, in 1627; another at Amsterdam, in 1629; one at Paris, from which the Psalms in 5 parts had disappeared; one at Amsterdam, in 1633; and one at Leyden, in 1635. The music was also printed with a Dutch translation of the words at Schiedam, in 1664.

The first portion of the Church Service printed in the English language was the Litany, 1544; the celebrated letter of Cranmer to Henry VIII., in October of that year, mentions also his effort at translating the Hymn "Salve festa dies," to which he had "put the Latin note" as being "sober and distinct." (See Jenkyn's Cranmer, i. 315.) Within the next few years the custom of psalm-singing had become common enough to warrant the "Church Tunes" being referred to as "usual" or "wonted tunes." In 1562, the version of the Psalter, now known as the "old" version, the joint production of Sternhold and Hopkins, was first printed with these melodies; and in the year after, under the editorship of John Day, "The whole Psalms, in foure partes, which may be sung to al musical instruments," &c., &c., in which Thomas Tallis, William Parsons, Thomas Causton, Richard Edwards, and three others were the harmonizers. This was followed in 1579 by another publication by Day, the music arranged by William Damon, and in 1585 and 1591 other contributions of a like nature appeared; but inferior in their pretensions and interest to the collection published in 1592 by Thomas Est or Este, "the assigné of William Byrd, dwelling in Aldersgate Street, at the signe of the Black Horse." The arrangers of the tunes were Richard Allison, E. Blancks, Michael Cavendish, Wm. Cobbold, John Dowland, John Farmer, Giles Farnaby (Bac. Mus., Oxford, 1592), Edmund Hooper (Organist and Gentleman of the Chapel Royal), Edward Johnson (Mus. Bac., Cantab.), and George Kirbye. Of this work, a second edition appeared, in 1594, and a third, in 1604, and during the next twenty years more than twenty editions of the *old version* with Church tunes were published by the Company of Stationers, into whose hands the matter had fallen.

In 1621 appeared "The whole Book of Psalmes, with the hymnes evangelicall and songs spirituall, composed into 4 parts by sundry authors, to such severall tunes as have beene and are usually sung in England, Scotland, Wales, Germany, Italy, France, and the Netherlands, never as yet before in one vol. published," &c., &c. As it is also said to be "newly corrected and enlarged," by Thomas Ravenscroft, Bachelor of Musicke; it is possible this may not have been the original edition. Every Psalm of the old version is printed in full, with its tune, in four parts; some tunes, however, being several times repeated, so that the total number is ninety-eight, of which forty are new compositions.

Of these Psalters, two particulars should be noted: 1st, that many of the tunes, though not all, are distinguished by names—"Norwich," "Martyrs," &c., a custom which was English and Scotch exclusively; and 2nd, that the arrangement of voices places the tune in the *tenor part*, not in the treble, as is the modern custom. Though the vocal parts in Ravenscroft are much more difficult than in Este, it is doubtful whether either became a manual for use in Church, or whether it was ever intended to be so. If it were, the skill necessary for the correct execution of these parts must have been much greater then than is common in any but the *best* Church choirs now.

The following arrangement of the "Old Hundredth," by John Dowland, is from Este, and will give a fair idea of his book:

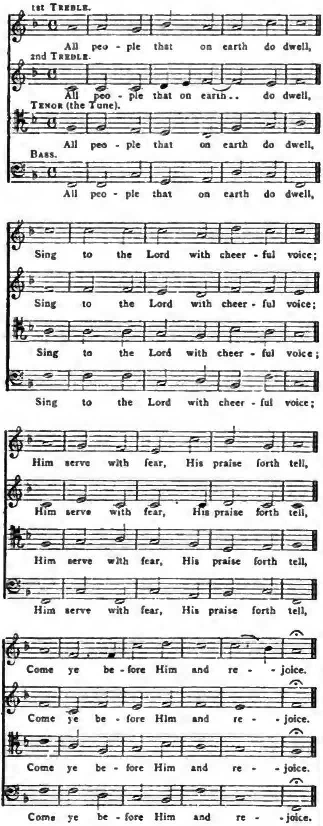
1st Treble.
All peo - ple that on earth do dwell,
2nd Treble.
All peo - ple that on earth.. do dwell,
Tenor (the Tune).
All peo - ple that on earth do dwell,
Bass.
All peo - ple that on earth do dwell,
Sing to the Lord with cheer - ful voice;
Sing to the Lord with cheer - ful voice;
Sing to the Lord with cheer - ful voice;
Sing to the Lord with cheer - ful voice;
Him serve with fear, His praise forth tell,
Him serve with fear, His praise forth tell,
Him serve with fear, His praise forth tell,
Him serve with fear, His praise forth tell,
Come ye be - fore Him and re - - joice.
Come ye be - fore Him and re - - joice.
Come ye be - fore Him and re - - joice.
Come ye be - fore Him and re - - joice.

The following is Ravenscroft's Harmony of the old 137th; the melody, as before, in the tenor. It is thought better for the general reader to reduce this to "short score." The frequent occurrence of syncopation in the accompanying parts should be noticed; the correct execution of this harmony would not be easy, even for an accomplished church choir:

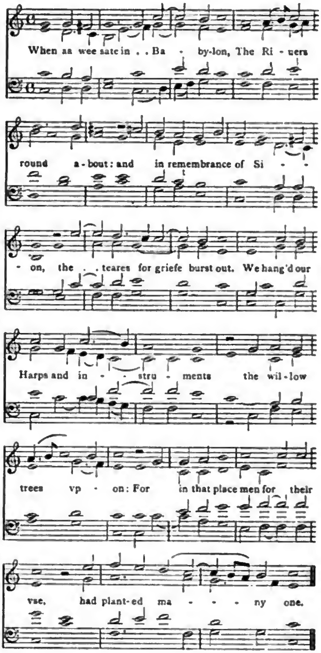
When as wee sate in . . Ba - by-lon, The Ri - uers
round a - bout: and in remembrance of Si - -
- on, the . . teares for griefe burst out. We hang'd our
Harps and in - - stru - ments the wil - low
trees vp - on: For in that place men for their
vse, had plant - ed ma - - - ny one.

In 1728, a professedly corrected edition of "Ravenscroft," edited by William Turner, was published by William Pearson; and another edition by J. Buckland in 1746. The "Harmonia Perfecta" of 1730, which professed to give the chief of Ravenscroft's tunes, avoided most of his syncopations, and spoiled many of his harmonies.

In 1615 appeared Andrew Hart's "Scotch Psalter; the cl. Psalms of David in prose and metre; with their usual notes and tunes, newly corrected and amended." Edinburgh, printed by Andrew Hart. In this edition the melody only appears, the harmonized edition being dated 1635. It contains a tune for each Psalm.

That the versifications of Tate and Brady, known as the "New Version" of the Psalms, had any beneficial influence on the culture of psalmody has never been pretended; but with the composition of hymns by Wesley and his followers came a new poetic material of which musicians were not slow to avail themselves. New compositions, in the new measures, and with the greater warmth of feeling suggested by the words, and differing entirely, both in melody and harmonization from the older school, follow naturally. The tune books and hymnals with which the press abounds, present every variety of these—from the very time of Wesley to the present—and these lie open before the student. In both melody and harmony they reflect not unhappily the style of the time in which they were written.

In the following example the employment of the voices is akin to that in a glee, and short passages are employed "a la fugue." It is from "Melodia Sacra; or, The Devout Psalmist's New Musical Companion," &c., &c. By William Tans'ur, Senior. "Musico Theorico," &c., London, G. Bigg, 1772. The method employed has long been acknowledged as utterly unsuited to any good purpose:

WESTERHAM. C.M.

As Ps. lxxxi. Old or New Version.

In Four Parts. W. T.

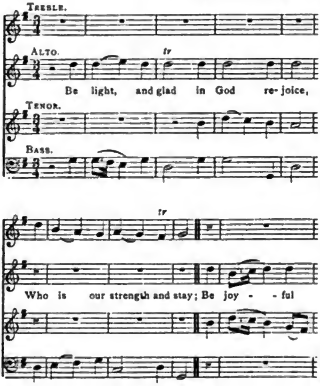
Treble.
Alto.
tr
Be light, and glad in God re-joice,
Tenor.
Bass.
tr
Who is our strength and stay; Be joy - - ful

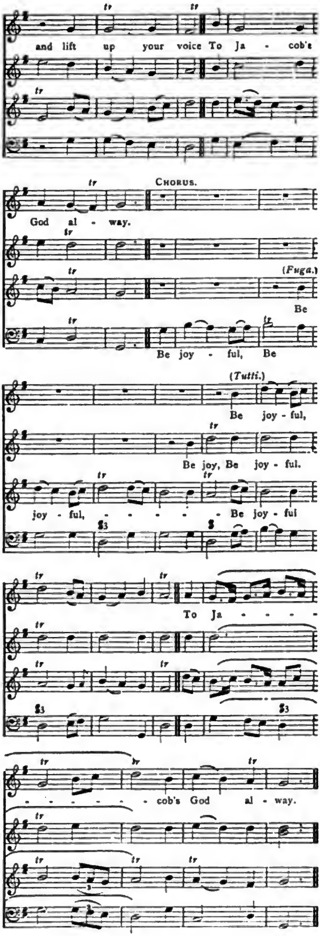
tr
tr
and lift up your voice To Ja - cob's
tr
tr
Chorus.
God al - way.
tr
tr
(Fuga.)
Be
tr
Be joy - ful, Be
(Tutti.)
Be joy - ful,
tr
Be joy, Be joy - ful.
tr
tr
joy - ful, - - - - - Be joy - ful
tr
tr
To Ja - - - -
tr
tr
tr
tr
tr
tr
- - - - - - - cob's God al - way.
tr
tr
tr
tr

The following is from "Parochial Psalmody," dedicated to His Majesty George the Fourth, by Thomas Greatorex, Conductor of His Majesty's Concert of Ancient Music, and Organist of Westminster Abbey; London, 182 . . (?) Many good examples of this style of writing may be found in the works of composers of the same date:

WESTMINSTER. C.M.

Psalm cv. New Version. "Sing praise to God."

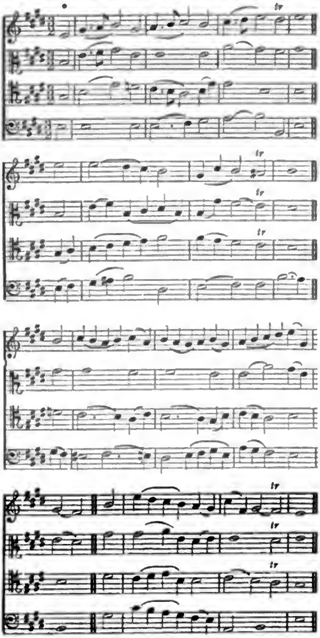

Remark in both the above (though separated by an interval of fifty years, and the latter about fifty years only from our own time), the frequent occurrence of the "tr," even on the four notes of the same chord.*

The following are examples of a style "to be avoided." They are reprinted without words, and in melody only, to show how unlike any accepted or acceptable form of music suited for Divine worship they are:

8.7.8.7. (DOUBLE.)

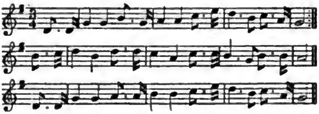

7.6.7.6. (TREBLE.)

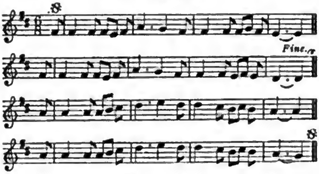

Hyper, ὑπέρ (*Gk.*) Above. A prefix to the names of modes one fourth *above* the *authentic;* as, Hyper-Eolian, Hyper-Dorian, Hyper-Iastian or Ionian, Hyper-Lydian, Hyper-Phrygian. [Greek Music.]

Hyper-diapason. Super-octave.

Hypo, ὑπό (*Gk.*) Below. A prefix to the names of modes commencing one fourth *below* the authentic, otherwise called Plagal, as, Hypo-Dorian, Hypo-Eolian, Hypo-Iastian or Ionian, Hypo-Lydian, Hypo-Phrygian [Greek Music.] [Plain Song.]

Hyporchema (*Gk.*) ὑπόρχημα. A religious hymn and dance connected with the worship of Apollo from the earliest times.

---

* A quotation from an older melody.

* For some interesting remarks as to the *trillo* as an integral part of a composition, see Preface to Hullah's edition of Pergolesi's "Stabat Mater."

I.

Iambus. A metrical foot consisting of a short syllable and a long. [Metre.]

Hymns are roughly said to be *Iambic* when they consist of such feet, the music commencing on the upbeat, *e.g.*, "All praise to Thee, my God, this night."

Iastian. Ionic. [Greek Music.]

Idyll (*Gk.*) *εἰδύλλιον*. *Lit.*, a small form. Hence, a beautiful form. (1) A poem of elegant structure, frequently, though not always, on a pastoral subject. (2) Music set to a poem of this character.

Il doppio movimento (*It.*) Double time, twice as fast.

Il fine (*It.*) The end.

Il più (*It.*) *Lit.*, the most, as, *il più piano possibile*, as soft as possible, &c.

Imboccatura (*It.*), *Embouchure* (*Fr.*) *Birne* (*Ger.*) The mouth-piece of a wind-instrument.

Imitando (*It.*) Imitating, as, *imitando la voce*, imitating the voice part, &c.

Imitation. The repetition of a short subject by another part. The subject proposed is sometimes called the *antecedent*, and the passage which afterwards imitates it the *consequent*. If the consequent is at a fixed interval from the antecedent throughout, the imitation is said to be *strict*, *e.g.*:

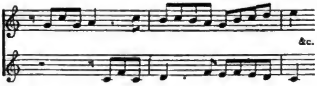

If the consequent only roughly follows the antecedent, the imitation is said to be *free*, *e.g.*:

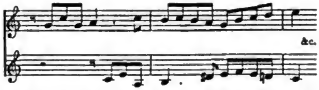

Imitation may be at any interval, and may be supported by any number of parts, and though it often forms an important element of fugue, it is unrestricted by the laws of subject and answer which are binding in that form of composition. The following are examples of imitation at the second:

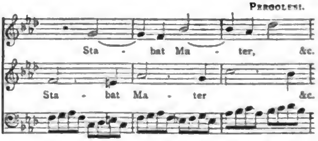

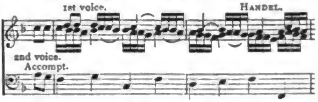

This subject may be treated in three parts as follows, including imitation at the third and sixth:

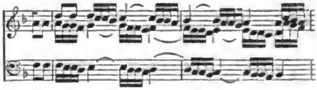

Examples of imitation at various intervals and numbers of parts might be multiplied to any extent. Imitation *by augmentation* is when the consequent is in notes twice the length of those of the antecedent, *e.g.*:

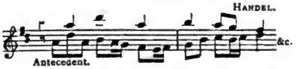

Imitation *by diminution* is when the consequent is in notes half the length of those of the antecedent, *e.g.*:

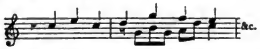

Imitation *by inversion* is when the intervals of the antecedent are *inverted* in order to form the consequent:

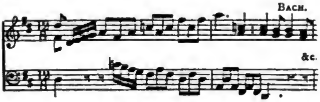

Imitation is said to be *convertible* when antecedent and consequent are interchangeable. If strict imitation be continued for any length of time it is said to be *canonical.*

Immer (*Ger.*) Always, as, *immer langsam*, slow throughout.

Impaziente, impazientemente (*It.*) Impatient, impatiently.

Imperfect. Not perfect. (1) An imperfect interval is one which is a semitone less than perfect. (2) The imperfect consonances are the third and sixth, as opposed to the fourth and fifth. (3) An imperfect cadence is one which does not give complete rest in key. All cadences not having a dominant or subdominant penultimate are said to be imperfect. [Cadence.] (4) An imperfect stop on an organ is one, the pipes of which do not extend through the whole compass of the manual. A *short* stop. (5) Duple measure was formerly called imperfect, as opposed to *tempus perfectum*,triple time.

Imperioso (*It.*) With grandeur, dignity, imperiously.

Impeto, con (*It.*)
Impetuosamente (*It.*) } Impetuously.
Impetuoso (*It.*)

Implied discord. A discord, the actual percussion of which is not found unless other parts be added to the chord, *e.g.*:

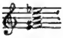

The intervals forming the above chord are practically minor thirds, which are not in themselves discordant.

Implied intervals. Intervals not expressed in thorough bass figuring. [Figured bass.]

Imponente (*It.*) In an emphatic, pompous style.

Impresario (*It.*) A designer, conductor, or manager of a concert or opera party. This term is often applied to a man who trains singers, or obtains them for public performances.

Impromptu (*It.*) (1) A piece of music written or played without previous preparation of the subject. An *extempore* performance. [Extempore.] (2) A composition not in any set form, having the character of an extempore piece.

Improvvisamente (*It.*) Unprepared, unpremeditated, *extempore.*

Improvvisare (*It.*) To found an *extempore* piece upon a suddenly suggested subject

Improvvisatore (*It.*) *Improvisateur* (*Fr.*) One who has the gift of improvising. [Extempore.]

In alt (*It.*) All notes in the first octave beyond the range of the treble stave. [Pitch.]

In altissimo (*It.*) All notes beyond the range of the first octave in alt. [Pitch.]

Incalzando (*It.*) Pressing on.

Incantation. The utterance of a charm or spell in a singing, monotonous tone.

Inchoatio (*Lat.*) The *Intonation* or introductory notes of a plain-song chant.

Indeciso (*It.*) Undetermined. In an undecided manner.

Indifferente
Indifferentemente } (*It.*) With indifference, or unconcern.
Indifferenza, con

Inferior. (*Lat.*) *Lit.*, lower. At an interval *below*, as, *Inversio in octavam inferiorem*, inversion at the octave below, &c.

Infinite Canon. [Canon.]

Inflexion. A departure from the monotone in chanting. [Accent, § 4.]

In fretta (*It.*) In haste. Hurriedly.

Inganno (*It.*) *Lit.*, deception, as, *cadenza d'inganno*, a deceptive cadence.

Inner parts. Those portions of the harmony that are not at the top or bottom.

Inner pedal. A sustained note in one of the inner parts. [Sustained note.]

Innig (*Ger.*) With expression and feeling.

Inno (*It.*) A hymn.

Innocentemente
Innocente } (*It.*) Innocently. In a child-like artless manner.
Innocenza, con

In nomine. (1) The name given to a *free* fugue as opposed to one constructed in strict accordance with the rules of fugue-form. (2) A piece in fugal style, that is, consisting largely of imitation and of contrapuntal devices, sometimes vocal, sometimes for instruments alone. When used in this sense the name has been by some traced to the words forming part of the Benedictus—*in nomine Domini*, and by others from the words of the Antiphon "in festo sanctissimi nominis Jesu," commencing "*in nomine* Jesu omne genu flectatur." Others again trace it to verse 5 of Psalm xx. "We will rejoice in Thy salvation, and triumph *in the name* (in nomine) of the Lord our God." The *in nomines* which exist are chiefly the production of composers of the 16th century.

Inquieto (*It.*) Restless, agitated.

Insensibile (*It.*)
Insensibilmente (*It.*) } Imperceptibly, by degrees.

Instrument à l'archet (*Fr.*) An instrument played with a bow.

Instrument à cordes (*Fr.*)
Instrumento a corda (*It.*) } A stringed instrument.

Instrument. Any mechanical contrivance for the production of sound: The musical instruments employed are divided into the following classes. Stringed, wind, and pulsatile. The stringed instruments are

Digitized by Google

the pianoforte and older instruments of its kind which are played by means of a clavier or key-board; the guitar and others whose strings are struck or plucked by the fingers; and the violin class played with a bow. Wind instruments are of wood or metal; those that are of wood are the flute, piccolo, hautboy, cor anglais, clarionet, basset horn, and bassoon; those that are of metal are the horn, trumpet, cornet-à-pistons, trombone, ophicleide, saxhorn, bombardon, &c. The pulsatile or percussion instruments are the kettle-drums, great drum, side drum, triangle, cymbals, and tambourine. A description of each will be found under its proper head.

Instrumentation. The art of using several musical instruments in combination; also, the style or treatment of orchestral instruments with a view to the production of special effects. [Orchestra.]

Instrument à vent (*Fr.*)
Instrumento da fiato (*It.*) } A wind instrument.

Instrumentirung (*Ger.*) Instrumentation.

Intavolare (*It.*) To set in notes, to score, to copy.

In tempo (*It.*) or *a tempo*, in strict time. A direction to follow strict time after any passage in which has occurred a *rall.*, &c.

Interlude, interludium (*Lat.*) A piece of music either impromptu or prepared, played between the acts of a drama, the verses of a canticle or hymn, or between certain portions of a church service.

Intermède (*Fr.*) **Intermedio** (*It.*) An interlude.

Intermedietto (*It.*) A short interlude.

Intermezzo (*It.*) An interlude.

Interrupted Cadence. [Cadence.]

Interruzione, senza (*It.*) Without hindrance, without pause.

Interval. The distance between any two sounds. In discussing all subjects relating to melody, or to the construction of chords, it is necessary to find names for the various kinds of intervals. Hence, in the earliest treatises on music they are divided into classes. The old axiom that consonance depends on simplicity of ratio naturally led authors to draw a line at the point at which two sounds ceased to be consonant and became dissonant. Among the Greeks, the unison, octave, fiftn, and fourth, were considered more perfect than the third and sixth and other intervals. In mediæval treatises an interesting division of intervals into perfect, medium, and imperfect is found; the unison and octave belonged to the first class, the fourth and fifth to the second, the third and sixth to the last. In later works appeared a division which is to this day followed by many writers, namely, into Perfect and Imperfect. This division runs thus:

Intervals
- Consonant
  - Perfect 4th, 5th, 8th.
  - Imperfect 3rd, 6th.
- Dissonant
  - 2nd, 7th
  - All augmented and diminished intervals.

So far, this division seems plain enough. But modern music requires that intervals shall be designated according to their *scale* value.

Hence a perfect interval when accidentally flattened becomes imperfect, and a major interval minor. But the application of the word *imperfect* to a perfect interval made smaller led to serious confusion, because, an *imperfect fifth* came as a subdivision of perfect intervals, while *imperfect* intervals of themselves formed a separate class. To avoid this cross-division, theorists have in the last few years used the name *diminished* for perfect fifths reduced by one semitone. But it is important to note that by this change of name a new difficulty arises, for *diminished* has almost by universal consent been applied to certain intervals when made *less than minor*, for example, C♯ to B♭ is a diminished seventh because C♯ to B♯ is a major seventh, C♯ to B♮ a minor seventh, C♯ to B♭ one semitone less than minor. The word *diminished* when applied by these authors to a fourth or fifth signifies that it is reduced from its normal state *by one semitone*, but the same word "diminished" when applied to a seventh signifies that it is reduced *by two semitones*. That utter confusion should result from such an undigested system is absolutely inevitable; and it will be found that professors and teachers of music to this day are unable to talk intelligibly to each other on the simple subject of "Intervals." For many years a system has been taught in Germany which makes the whole matter perfectly plain. It is this:—

1. Intervals are reckoned upwards, inclusively, and by the number of *names* of notes they contain.

2. Intervals are in their normal state when reckoned from the first note of the major scale. The lowest note of the interval being considered for the time as a *tonic*.

3. Normal intervals are *major*. Thus, D to F♯ is a major third, because F♯ is the third degree of the scale of D; F to B♭ is a major fourth, because B♭ is the fourth of the scale of F; B to F♯ is a major fifth, because F♯ is the fifth of the scale of B; and so on. In other words, *all* the intervals of any major scale reckoning up from the tonic respectively are major.

4. Intervals one semitone less than major are *minor*. Thus, C to D♭ is a minor second

because it is one semitone less than the normal D in the scale of C; B to F♮ is a minor fifth because one semitone less than the normal F♯ in the scale of B; and so on.

5. Intervals one semitone greater than major are *augmented.* Thus, C to D♯ is an augmented second, because it is one semitone greater than C to D, the normal interval; C to G♯, an augmented fifth; and so on.

6. Intervals one semitone less than minor are *diminished.* Thus, F♯ to E♭ is a diminished seventh; C♯ to G♭, a diminished fifth; C♯ to E♭, a diminished third; and so on.

The simplicity of this system is already apparent. The following diagram shows it at a glance:

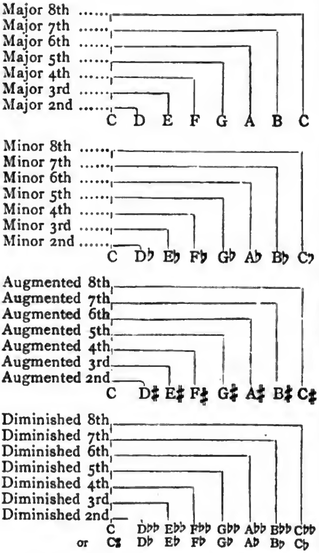
Major 8th
Major 7th
Major 6th
Major 5th
Major 4th
Major 3rd
Major 2nd
C D E F G A B C
Minor 8th
Minor 7th
Minor 6th
Minor 5th
Minor 4th
Minor 3rd
Minor 2nd
C D♭ E♭ F♭ G♭ A♭ B♭ C♭
Augmented 8th
Augmented 7th
Augmented 6th
Augmented 5th
Augmented 4th
Augmented 3rd
Augmented 2nd
C D♯ E♯ F♯ G♯ A♯ B♯ C♯
Diminished 8th
Diminished 7th
Diminished 6th
Diminished 5th
Diminished 4th
Diminished 3rd
Diminished 2nd
C D♭♭ E♭♭ F♭♭ G♭♭ A♭♭ B♭♭ C♭♭
or C♯ D♭ E♭ F♭ G♭ A♭ B♭ C♭

It will be seen from this last that diminished intervals are produced in two ways—either by making the upper note of a minor interval flatter, or by making the lower note of a minor interval sharper. Hence it is that some intervals have to be calculated from notes not having a diatonic scale of their own; thus, B♯ to A♮. In such cases, the nature of the interval is readily found by temporarily reducing the lower note; thus, B♮ to A♮ is a minor seventh, therefore B♯ to A♮, being a semitone less than minor, must be a *diminished* seventh.

The following is the simplest form of stating the rule for naming intervals according to this system. "When asked the nature of any interval, bear in mind the major scale of the lower note, then, if the upper note is higher by a semitone than it would be in the major scale of the lower note, it is *augmented;* if the upper note is actually a note of that major scale it is *major;* if it is less than major by one semitone it is minor; if it is less than minor by one semitone it is *diminished.* When the lower note is a sharpened note, which has no scale of its own, consider it as one semitone lower, when the interval can be easily determined, *e.g.*, C𝄪 to G♯ is a minor fifth, because the interval is one semitone less than C♯ to G♯, the normal fifth. When the lower note is a flattened note, which has no scale of its own, consider it as temporarily raised before determining the nature of the interval, thus C♭♭ to G♭ is an augmented fifth, because it is one semitone greater than the normal fifth C♭ to G♭."

The only obstacle to the general adoption of this excellent method of tabulating intervals is to be found in the pertinacity with which professors adhere to the expression *perfect* fifth and *perfect* fourth, and abhor the term *major* fifth and *major* fourth. This absurd prejudice, which arguments drawn from the history and science of music seem to have no power to remove, must be allowed to die of old age. If there is any real distinction between the perfection of a fifth and the imperfection of a third, it might even then be allowed to students to call fourths and fifths major, on the understanding that they also possessed a remarkable perfection which no other intervals possess.

It has been suggested that intervals should be called, instead of (1) minor, (2) major, and (3) augmented; (1) minor, (2) normal, and (3) major. This would certainly be a better system than any yet devised; but it is to be feared that it would be impossible to disturb the universally accepted meaning of the word major.

Intervals greater than major or normal have been termed (besides *augmented*) extreme, sharp, superfluous, pluperfect, &c.

Intimo (*It.*) With inward emotion.

Intonare (*It.*) To tune, to sing.

Intonation. (1) The method of producing sound from a voice or an instrument. (2) Correctness of pitch; *e.g.*, *just intonation*, singing or playing in *perfect tune.* (3) The method of chanting certain portions of the church services. [Accent, § 4.] (4) The notes which precede the reciting-note in a Gregorian chant.

Intonator. A monochord, or single string stretched across a flat sound-board. Below the string is a diagram of the exact divisions of the monochord necessary for the production of the true musical scale. By means of a moveable bridge, the student is able to sound the notes represented on the diagram, and so, to educate his ear to a true sense of relative pitch.

Intonazione (*It.*) Intonation.

Intoning. The practice of chanting in the Anglican church service, which includes the delivery of the prayers in monotone, and the precenting or leading of the plain song of the Psalms, Credo, Canticles, &c.

Intrada (*It.*) (From *intra*, between). An interlude or *entr'acte*.

Intrepidamente
Intrepidezza, con } (*It.*) Boldly, daringly, with courage.
Intrepido

Introduction. A preparation for, or preamble of, a movement or series of movements. Introductions were formerly employed only before larger musical works such as oratorios and operas, but are now used as preparations to works of nearly every kind from a cantata to a waltz.

Introducimento (*It.*) An introduction.

Introduzione (*It.*) Introduction.

Introit, Introito (*It.*), **Introitus** (*Lat.*, from intro-eo). An antiphon sung while the priest proceeds to the altar to celebrate mass. In the Anglican Church, a short anthem, psalm, or hymn, sung while the minister proceeds to the table to administer the Holy Communion. Formerly, in some English cathedrals, the Sanctus was sung as an Introit. This practice arose probably from the fact that the Communion Service soon after the Reformation ceased to be performed *chorally*, a proof of which is found in the fact that for nearly two centuries—namely, from 1660 to 1840—the Sanctus was never set to music except as an Introit, nor was the *Gloria in excelsis* set, but as an anthem. [Cathedral music.]

Invention. (1) The gift of finding new melodic phrases, or new combinations in harmony. The possessor of such a gift is frequently the founder of a school or style in music. (2) Suites de pièces. [Suite.] (3) The name of a piece of music of a fanciful character.

Invenzione (*It.*) Invention.

Inversion. The transposition of certain phrases having a common root. There are three kinds of inversions in music—(1) of chords; (2) of intervals; (3) of subjects.

(1) The inversion of a chord is effected by making one of the inner notes act as a bass note, and by this means as many inversions can be made, as there are actual notes in the chord, not counting the root. In such inversions the harmony remains the same, although the order of component parts is changed.

(2) Intervals are inverted by making that which was the upper note the lower, and the reverse. The inversion of an interval within the octave may readily be found in the difference between the figure 9 and the interval known; thus an interval of a second becomes a seventh by inversion, &c.

(3) The inversion of a subject is produced by inverting the intervals of which it consists. [Al rovescio.]

Invitatorium (*Lat.*) A sort of introductory sentence or antiphon immediately preceding one of the offices of the Church; such, for instance, as parts of the *Venite exultemus*, or other words inviting the presence or attention of the congregation.

Invitatory. [Invitatorium.]

Ionian mode. The Church mode commencing on the note C.

Ira, con
Iratamente } (*It.*) Wrathfully, with anger.
Irato

Irish bagpipe. [Bagpipe.]

Irish harp. [Harp.]

Irlandais (*Fr.*) In the Irish style.

Ironicamente
Ironico } (*It.*) Ironically.

Irresoluto (*It.*) Uncertain, without settled purpose, in an undecided manner.

Isochronism. Periodical recurrence. [Acoustics.]

Ison (*Gk.*) In the music of the early Greek Church, the ison was the key-note or tonic of a chant. The ison was *moveable.*

Istesso (*It.*) The same; as, *l'istesso tempo*, at the same pace.

Istrumento. An old form of *stromento*, an instrument.

Italian sixth. The name ot a chord containing a bass note accompanied by a major third and a sharp sixth.

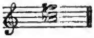

[Extreme sixth.]

Italian strings. Catgut strings for violins and similar instruments, prepared in a peculiar manner, so that they are more transparent in appearance, and less liable to become false in use than ordinary strings. The majority of real Italian strings are made in Rome. [Catgut.]

Italienne, a l' (*Fr.*)
Italiano (*It.*) } In the Italian style.

Ite missa est (*Lat.*) The concluding words of the Mass in the Romish Church, from whence the name of that service (Mass) is derived.

Digitized by Google

J.

Jack. The name of the hoppers attached to the keys of a harpsichord; the end of the jack was furnished with a quill, as plectrum, to pluck the strings. [Pianoforte.]

Jägerchor (*Ger.*) Hunting chorus.

Jaleo. *Zaleo* (*Sp.*) A national dance of Spain.

Janitscharenmusik (*Ger.*) Janissary music, noisy music on instruments of percussion.

Jeu (*Fr.*) A stop on the organ, or harmonium. *Grand jeu*, the full power.

Jeu d'anche (*Fr.*) Reed stop.

Jeu d'ange (*Fr.*) The *vox angelica* stop.

Jeu de flûte (*Fr.*) Flute stop.

Jeu d'orgue (*Fr.*) An organ stop.

Jeux doux (*Fr.*) Sweet or soft stops.

Jeux forts (*Fr.*) Loud stops.

Jew's harp. *Jew's trump, trompe de béarn, Guimbarde* (*Fr.*), *Maultrommel, Brummeisen* (*Ger.*) An instrument held between the lips, the sound coming from the vibrations of a tongue of metal, bent at a right angle, which is set in motion by being twitched with the forefinger. The sound is increased in intensity by the breath, and altered in pitch by the shape of the cavity of the mouth, which acts as a reflector. This name some derive from *jeu, play*, from the fact of its being a toy; others find the derivation of the word in the manner in which the instrument is used, *Jaw's* trump; the German word *Maultrommel* seems to confirm the latter derivation. Insignificant as the instrument seems, it is capable of producing the most beautiful and pleasing sounds of a melancholy character.

A common soldier, named Koch, who served in the army of Frederick the Great, first raised the Jew's harp from a toy to a solo instrument, and a M. Eulenstein gave concerts in London some fifty years ago, during which he performed upon sixteen instruments, by which means he obtained a compass of four octaves. The Jew's harp is an old invention, being mentioned under the name of Crembalum by Prætorius in his Organographia, in the year 1619. It is a favourite instrument with the Hottentots, at the Cape, who load the end of the elastic spring or vibrator with small portions of hard wax to deepen the tone. The inhabitants of the island of St. Kilda were formerly great proficients upon this, which was their only instrument of music.

Jig. Gigue (*Fr*). **Giga** (*It.*) (1) A lively dance which may be performed by one or more dancers. It is popular among many nations, is distinguished by various titles, and has a certain amount of difference in the steps according to the habits and customs of the people by whom it is adopted. With some it is a sober, steady, jog-trot sort of a country dance, with others it is a wild, savage exercise, without point or meaning. With some it is made a means of displaying the agility of the lower limbs of a combined company of dancers; with others it is a terpsichorean drama for two performers, in which all the emotions excited by love are represented by gestures and monosyllabic cries. The Hornpipe, La Matelotte, La Seguidilla, the Czardasch, the Reel, the Rant, are portions of the country dance; the Bolero and the Chica are only jigs in various forms.

The term has been derived from the word chica, the name of a dance identical with the jig in some of its characteristics. But the similarity of the title of the instrument "Geige," a fiddle, with which the dance was usually accompanied, with the name of the dance, would seem to point to an origin similar to that of "hornpipe," where the character of the instrument gave the name to the tunes played upon it.

The *jig, gigue, gighe*, or *giga*, as it is variously spelt, was one of the instruments used by the musicians of the 12th and 13th centuries, and this is generally understood to have been a sort of fiddle.

The "Geige," or "jigg" may have received its name from the "up and down" motion with which it is played. A fiddlestick is still called in the west of England a "jigger," and there is a sort of pump used by brewers which has the same name. In the east of England a sieve is called a "jiggin," and there are many other words of cognate origin, which it would scarcely serve the present purpose to quote.

Songs sung by clowns in the interludes to the plays were called *jigs*, and any ambling rhyme was also called a "jig."

Digitized by Google

In the romance of "Gawaine and the Green Knight," the word Bauderyk is used. It is explained in the "Promptorium Parvulorum" as meaning "guige," a transverse strap, by which the shield was suspended round the neck. The use of the word with such a meaning may account for the fact that the *vielle* or *hurdy-gurdy* is called *geige*, *gigelira* or *giga vel lira* from having been slung round the body.

Jig-tunes or fiddle-tunes are frequently found in old MS. and printed collections from the 16th century onwards. They are all of lively character though of no special or distinctive rhythm. Jigs, though now danced by one, or by two at the most, were formerly "round dances," in which many did indulge.

> —— "the mad lads
> And country lasses, every mother's child
> With nosegays and bridle-laces in their hats
> Dance all their country measures, rounds,
> and jigs."
> HEYWOOD, "*A woman killed with kindness.*"

Mr. Chappell (Popular Music, p. 792), says that "the jig is now associated in the public mind with Ireland," but he further says that he has "not found one called Irish before the latter part of the 17th century." Scotch jigs were noticed before Irish jigs, as for example, by Shakespear in "Much Ado about Nothing," comparing "wooing, wedding, and repenting" with a Scotch jig, a measure and a cinque-pace; and this comparison implies that the Scotch jig was of a lively character.

(2) As a movement in a "suite," the jig is found in works produced towards the latter part of the 17th century, and onwards to the time of Haydn. At first the phrases were short, and of no more variety than was needed for the purposes of the dance, for the jig was occasionally one of the figures of the country dance. But later it was made the vehicle for display in harpsichord playing, and was lengthened and elaborated and became the origin of the last movement of the sonata. It was written in $\frac{3}{4}$, $\frac{3}{8}$, $\frac{6}{4}$, $\frac{6}{8}$, $\frac{9}{4}$, $\frac{9}{8}$, and $\frac{12}{8}$ time; the peculiarity of the rhythm of triplets was nearly always preserved if not insisted upon.

Sir John Hawkins says that the characteristic of the jig "is duple time, thus marked $\frac{6}{8}$ or $\frac{12}{8}$," and "that the air consists of two strains, undetermined as to the number of bars;" others say it is "of compounds of threes in time," and many of the "gigues" belonging to the "suites" are written to show the prominence of the triplets. But the following "jigg" out of "Compositions for Broken and Whole Consorts of two, three, ffower, ffive and six parts, made by Matthew Locke, composer in ordinary to His Majesty Charles the II.," written in 1672, is an example of duple measure and unusual rhythm.

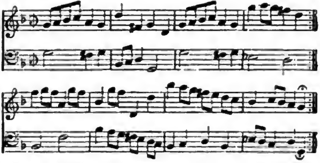

The subjoined specimens of jig tunes will give an idea of the progress in treatment from the earliest time when the dance was adopted as a harpsichord piece, until it began to be disused.

A JEGG.

BEN ROGERS, of Windsor, 1678.

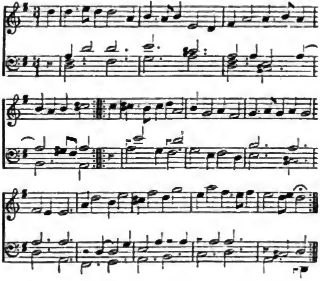

JIG.*

MR. ECCLES, c. 1690.

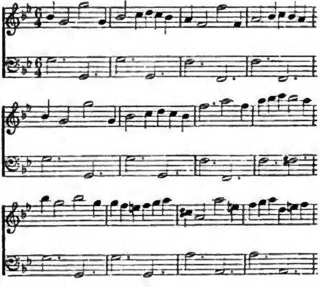

---

* A modification of this tune is still employed for the so-called Irish Jig.

Digitized by Google

JIG.

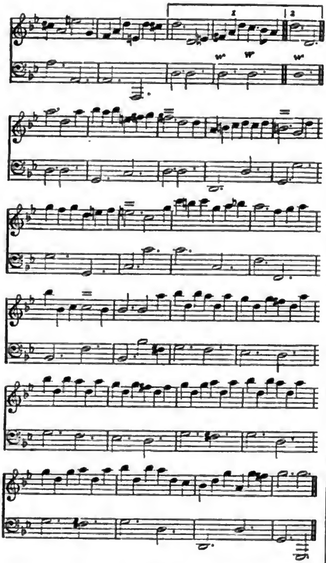

GIGUE.

J. Mattheson, 1681—1722.

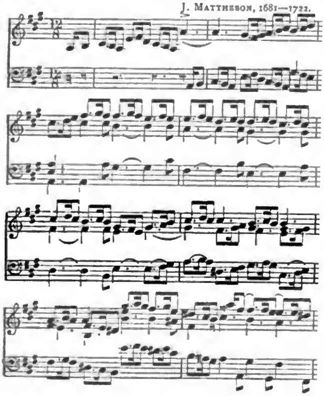

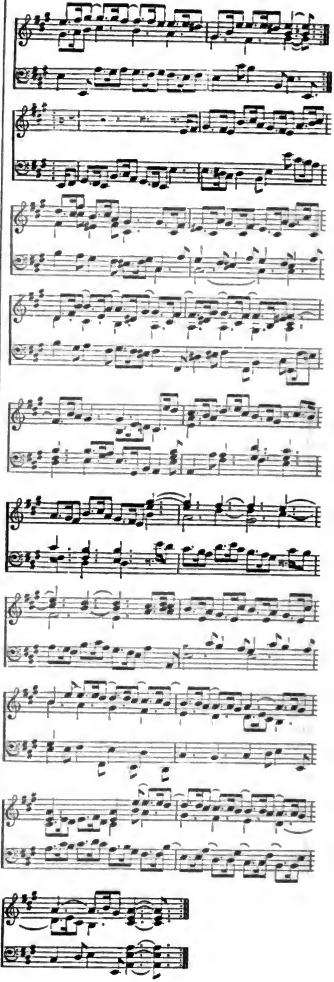

Jingles. Discs of tin, brass, or bell metal, fastened at intervals round a tambourine. [Tambourine.]

Jobel (*Heb.*) A word applied in the Holy Scriptures to certain *trumpets* or *horns*. It is probably equal to the affix *jubilee*; but, *jubilee-horns* were used for other purposes besides that of proclaiming of jubilees.

Joculator. [Troubadour.]

Jodeln (*Ger.*) **Jodle, Jödl.** A peculiar method of singing adopted by the Swiss and Tyrolese, by the rapid alternation in melodic progressions, of the natural and falsetto voice.

The following is a specimen of a *jödl* at the close of a national Swiss song.

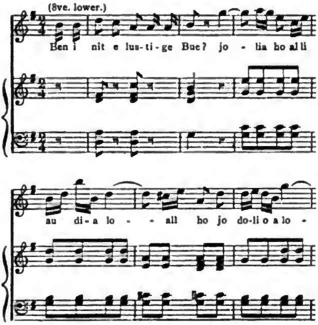
(8ve. lower.)
Ben i nit e lus-ti-ge Bue? jo - lia ho al li
au di-a lo - - all ho jo do-li o a lo -

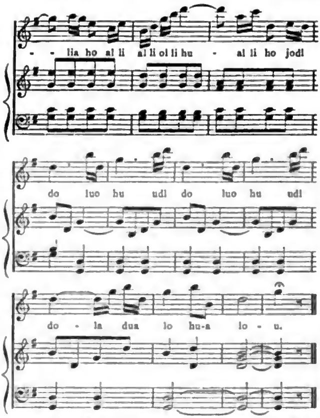
- - lia ho al li al li ol li hu - al li ho jodl
do luo hu udl do luo hu udl
do - la dua lo hu-a lo - u.

Jongleurs. [Confrérie de St. Julien.] [Troubadour.]

Jump. Another name for the dance called a dumpe. [Dump.]

Just intonation. The correct sounding of intervals in singing or playing. [Intonation.]

Juste (*Fr.*) *Lit.*, just. In tune.

Justesse (*Fr.*) Equality, purity, and correctness; as *justesse de la voix*, purity of intonation. *Justesse de l'oreille*, correctness of ear.

K.

Kabaro. A small drum used by the Egyptians and Abyssinians.

Kalamaika. A Hungarian dance.

Kammer (*Ger.*) Chamber, *e.g.*, *Kammerconcert*, chamber-concert; *Kammermusik*, chamber-music (*q.v.*); *Kammerton*, concert pitch. [Chorton.]

Kandele or **Kantele.** The ancient minstrel's harp of the Finns. The name is also given to a species of dulcimer, having five strings, in use among the same people.

Kanon. [Canon.]

Kanonik. The art of Canon.

Kanoon (*Turk.*) A dulcimer. [Canun.]

Kapellmeister (*Ger.*) The leader or conductor of a band of music. [Capellmeister.]

Keckheit (*Ger.*) Boldness, audacity, vigour.

Keeners. Irish singing mourners, who, in olden times, were hired to howl at funerals, in perpetuation of a heathen custom derived from a Phœnician ancestry. The duty of professional keeners was undertaken by the females of the family, or acquaintance of the deceased. After the body had been dressed in grave-clothes, adorned with flowers, and placed upon a bier, the chief keener arranged his followers at the foot and head of the corpse, and recited at the first the *Caoinan*, in which were described the virtues and possessions of the departed. This eulogy sung softly, and accompanied upon the harp, was followed by a full chorus of all assembled; the *ullalu*, or *hubbaboo* as it was sometimes called, next followed, joined in by every one in a louder voice; and after this was ended, the *goul*, or lament, followed, in which the sounds were often raised to howls and yells. The process was repeated during the wake or vigil before the burial, the keeners sometimes following the corpse to the grave. When the body was laid down for the purpose of resting the bearers, or the procession came in sight of a wayside cross or chapel, the keening was suspended, and all present knelt to say in silence an *Ave* or a *Pater noster*, a practice which may be held to denote that keening was felt to be incongruous with Christian usage.

Kent Bugle. [Metal wind instruments.]

Kemangeh. An Arabian instrument of the fiddle class. [Violin.]

Keras (*Gk.*) Κεράς, a horn. Originally, any instrument made out of the horn of an animal. *Lat.*, *cornu*.

Kerana or **Kerrena.** The name of the Persian horn which is sounded at sunset and at midnight.

Keraulophon. An organ stop, invented by Gray and Davison. Its pipes are of small scale, and are surmounted by a moveable ring of metal. Its tone is soft, delicate, and reedy

Keren. A Hebrew trumpet. The word is sometimes used in the Bible as synonymous with *shophar*, and to it is sometimes affixed *jobel*, rendered in the English version *rams'-horns*.

Kern (*Ger.*) The *language* of an organ-pipe.

Kessel (*Ger.*) [Mouthpiece.]

Kettle-drums. Timbales (*Fr.*); Timpani (*It.*); Pauken (*Ger.*) Instruments formed of shells of copper or brass, over the top of which parchment is stretched. Parchment is considered best when most transparent.

There are two drums usually employed, the tuning of the larger being limited to the notes between

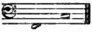

And of the smaller between

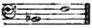

So that the two drums overlap each other in compass by a tone.

The tuning is effected by tightening or loosening the head or skin by means of a ring of metal moved by screws turned by a key. Several plans have been invented by which they may be tuned from one screw. Cornelius Ward's plan was by means of an endless cord passing into the interior of the drum. Potter's is by curved rods outside the shell and meeting in the centre at the bottom. Drums are usually tuned to the tonic and dominant of the piece in which they are employed, but other tunings are found in some

Digitized by Google

scores, as, for example, in Mendelssohn's Rondo in B♮, in which the drums are in D and E. Beethoven, the first who elevated the drum from a mere noise-producing machine into an orchestral instrument, has his drums tuned in various ways in his scores, as a reference to his symphonies will show. In No. 7, 1st and 2nd movements, they are in fifths played as written; 3rd movement, in *sixths* as written. In No. 8, 1st and 3rd movements, in fifths played as written; 4th movement, in octaves. In No. 9, 1st movement, in fourths as written; 2nd movement, in octaves; 3rd movement, in fifths; 4th, in fourths.

Drums for all modern scores are of 8ft. tone, giving a definite and accurate note. When the drums are required simply for marking rhythmical measures, the 16ft. tone is best. For this reason modern writers add the long drum (grosse caisse, *Fr.;* gran cassa, *It.;* Trommel, *Ger.*) to their scores. The bass clef is usually employed in writing for the kettle-drums. Sometimes the notes required are written in their proper places on the stave, but the notes to which they are to be tuned is always stated, as:

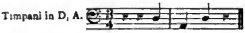

Neither sharps nor flats are ever written as signatures for the drums. If the sounds required are the tonic and dominant of the key, the drum part may be written in C. If other notes are needed, the rule is to write the actual notes, but without sharps or flats:

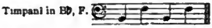

Drums were formerly used in combination with trumpets [Dettingen Te Deum, Mozart's No. 6 Symphony], Beethoven being the first to employ them as a means of gaining special solo effects (see his Symphony in F, No. 8, and Scherzo of No. 9).

Snare, or side-drums, are employed in the same manner that the long-drum is, that is to say for rhythmical, not for harmonic effects. [See Rossini's overture to "La Gazza ladra," and the "Benedizione de pugnali" scene in Meyerbeer's "Huguenots."]

The bass-drum being sometimes played with one stick, the player is frequently called upon to perform upon the cymbals at the same time, not with the best effect with regard to the cymbals, for cymbals ought to be slided one over the other and not struck flat together.

The best sort of sticks for the kettle-drums are those having whalebone handles with a wooden button covered by a piece of sponge; by the use of these the finest gradations of tone may be gained.

A long roll upon the drum is thus indicated:

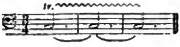

A short roll thus:

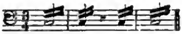

Kettle-drums are said to have been introduced into the orchestra by Handel, who employed a pair taken as part of the spoil at the battle of Dettingen, in the score of the Te Deum, written in celebration of that event; but the drum was first employed by Lully to strengthen the tutti parts of his overtures and choruses.

Among many remarkable passages for the drum may be noted that in A in the slow part of the overture to "Der Freischütz," also one note C *piano* when the first subject (syncopated) of the middle movement returns. Four notes in Violin Concerto, Beethoven. In Haydn's Mass No. 2, and of "Agnus Dei," on the words "Dona nobis pacem," some soft notes of the drum are introduced which at first seem inappropriate.

Key. (1) A mechanical contrivance for closing or opening ventages, as in flutes, clarinets, ophicleides, &c. By means of keys on such instruments, apertures too remote to be reached by the outstretched fingers are brought under control of the player. (2) A lever which brings the *pallets* of an organ under the control of the hand or foot of an organist. (3) A lever which controls the striking apparatus of a key-stringed instrument. In the harpsichord it acted on the *jack*, in the pianoforte it acts on the *hammer*. (4) The *wrest* or key used for tuning instruments having *metal* pegs. Its end is hollowed out, so as to fit over the four-sided end of the peg, and the crossbar with which it is surmounted gives leverage to the hand of the tuner, so that he is enabled to tighten or loosen a string, or (in the case of a drum) slacken or strain a parchment. (5) The sign placed at the commencement of the musical stave which shows the pitch of the notes, was originally called a *clavis* or *key*. This sign is called in modern music a *clef*. [Clef.] (6) *Key*, in its modern sense, is the starting point of the definite series of sounds which form the recognised scale. Different starting-points require the relative proportion of the steps of the scale to be maintained by means of sharps or flats in the signature. The key of C requires no flats or sharps for this purpose, hence it is called the *normal* key.

Key-board. *Clavier* (*Ger.*) The range of keys upon a pianoforte or organ. Keys

Digitized by Google

played by the fingers are called *manuals;* those by the feet are called *pedals.*

Key-bugle. [Metal wind instruments.]

Key-chord. The common chord of the tonic, *e.g.*: C, E, G is the key-chord of C.

Key-note. The note which, according to the signature, forms the starting point of the scale. The tonic. The *doh.*

Key-trumpet. [Metal wind instruments.]

Khalil, Chalil, or **Halil** (*Heb.*) The flute of the Hebrews. As the word is traced to a root signifying *bored through*,it is quite possible that it may have been like the *aulus*, used to describe either a flute or an oboe.

Kin. A Chinese musical instrument, the *scholar's lute*, a sort of *dulcimer.*

King. A Chinese instrument of percussion, consisting of metal plates, which are struck with a hammer.

Kink. A twist in a catgut string from close laying, which, by uncoiling and weakening that part of the string in which it occurs, frequently makes it useless for the instrument for which it is intended.

Kinnor. One of the most ancient of the Hebrew string-instruments, being the first mentioned in the Bible; Jubal "was the father of all such as handle the harp (*kinnor*) and organ (*ugab*)" (Genesis iv. 21). The German version has it, "from him descended *fiddlers* (Geiger) and *pipers*" Pfeiffer). Several theories have been brought forward as to the nature of the kinnor, but it was probably closely allied to the Nubian *Kissar*, perhaps identical with it (see p. 218).

Kirche (*Ger.*) Church, as *Kirchencomponist*, church composer; *Kirchenmusik*, church music; *Kirchenstyl*, church style, &c.

Kit. *Pochette* (*Fr.*) *Bretgeige* (*Ger.*) A small violin, about sixteen inches long, played with a bow of nearly the same length, used by dancing masters because of its convenience for carrying in the kit or pocket. It should be stated that some have derived the word from the Persian kitar, cithara, guitar. The Italian name of kit, *sordino*, is given to it because of its small *stifled* tone.

Klang (*Ger.*) Sound. Quality of sound, *timbre* (*Fr.*)

Klangboden (*Ger.*) Sound board. Resonance box.

Klangfarbe (*Ger.*) *Lit., sound-colour.* Quality of sound, *timbre* (*Fr.*)

Klanggeschlecht (*Ger.*) A kind or genus of sounds, as *diatonisches, chromatisches*, und *enharmonisches Klanggeschlecht*, the diatonic, chromatic, and enharmonic *genera.*

Klangstufe (*Ger.*) A degree of a scale. [Tonstufe.]

Klappe (*Ger.*) The key of any wind instrument. A valve.

Klappenflügelhorn (*Ger.*) Keyed bugle.

Klappentrompete (*Ger.*) Keyed trumpet.

Klein (*Ger.*) Small. (1) Minor, as *kleiner Halbton*, a minor semi-tone; *kleiner Nonenakkord*, the chord of the minor ninth, &c. (2) Small, as *Kleinbass*, or *kleine Bassgeige*, a violoncello; *klein Gedackt*, a small stopped diapason, a stopped flute.

Klingel (*Ger.*) A small bell.

Knee stop. A mechanical contrivance on harmoniums, by which certain shutters are made to open gradually when the knees are pressed against levers conveniently placed for the purpose. A *crescendo* and *diminuendo* are thus produced.

Knell. A stroke upon a bell made at periodic intervals, at the time of a death or funeral.

> "When thou dost hear a toll or knell,
> Then think upon thy passing bell."
> *Old Proverb.*

A single stroke is made for a child, two strokes for a woman, and three for a man.

Knicky-knackers. The common instrument of percussion known as *bones.*

> "Then let our armies join and sing,"
> And pit-a-pat make our knackers ring."
> MIDDLETON. *The Gipsy Rout*, 1623.

Kniegeige (*Ger.*) A *Viola da gamba.*

Knoll. [Knell.]

Kopfstimme (*Ger.*) Head voice. *Voce di testa* (*It.*)

Koppeln (*Ger.*) Couplers. [Organ.]

Kraft (*Ger.*) Energy, vigour. *Kräftig*, energetically.

Kreislieder (*Ger.*) A circle or series of songs. [Liederkreis.]

Kreuz (*Ger.*) The sign for a sharp. *Doppelkreuz*, a double sharp.

Kriegslied, *or* **Kriegsgesang** (*Ger.*) A battle song.

Krome (*Ger.*) A quaver ♪ (*Lat. chroma*).

Krotalon. [Crotalum.]

Kroumatic. Instrumental. κρουματική μουσική (*Gk.*), instrumental music. Sounds produced by *striking.*

Krummhorn (*Ger.*) *Lit.*, crooked horn. *Cornetto curvo* (*It.*) *Corno storto* (*It.*) An organ stop, consisting of reed pipes. It is of 8ft. pitch. This name has been corrupted to *Cremona* on English organs.

Krustische Instrumente (*Ger.*) Instruments of percussion (from Gk. κρούω, to strike). [Kroumatic].

Kuhhorn (*Ger.*) **Cor de vaches** (*Fr.*) A cow horn. The common Alpine horn.

Kurz (*Ger.*) Short.

Kyrie eleison (*Gk.*) [Mass.]

Digitized by Google

L.

L. The letter employed as the abbreviation of the word *left* or *linke* (*Ger.*), as L. H. *left hand* or *linke Hand*.

La. (1) The solfeggio name for the sixth degree of the scale. (2) The key-note of the minor scale without a signature.

La bémol (*Fr.*) The note A flat.

La bémol majeur (*Fr.*) The key of A flat major.

La bémol mineur (*Fr.*) The key of A flat minor.

Labial. Organ pipes with lips. *Flue* pipes.

Labialstimmen (*Ger.*) Stops belonging to the *flue* work. Lip-stops.

Labium. The lip of an organ pipe.

Lacrimoso (*It.*) Mournfully.

Lade (*Ger.*) Wind-chest of an organ.

La destra (*It.*) The right hand.

La dièse (*Fr.*) The note A sharp.

Lagrimando (*It.*)
Lagrimoso (*It.*)
} Mournfully, sadly, tearfully.

Lai (*Fr.*) A lay, song. [Lied.]

Lament. The name given to some Celtic melodies of a mournful character, or to tunes set to words of a melancholy cast.

Lamentabile (*It.*)
Lamentabilmente (*It.*)
Lamentando (*It.*)
Lamentevolmente (*It.*)
Lamentevole (*It.*)
Lamentoso (*It.*)
} Mournfully, plaintively.

Lampons (*Fr.*) Drinking songs, from the word *lamper*, to gulp down, whence, probably our word *lampoon*, a libellous ballad.

Lancers. The name of one of the arrangements of sets of dances.

Ländler (*Ger.*) The name given to a dance popular among the Styrian peasants. The tune is of a lively, graceful character in $\frac{3}{4}$ or $\frac{3}{8}$ time. Many of the melodies proper to this dance are of a beautiful and simple style.

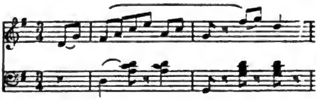

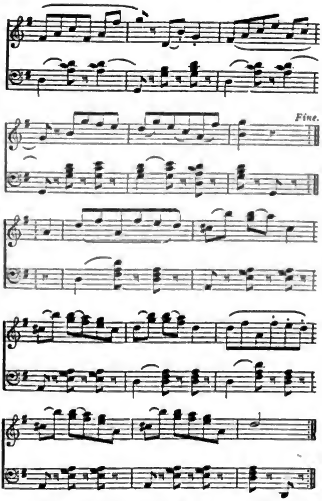

Langsam (*Ger.*) Slow; *etwas langsam*, rather slow; *ziemlich langsam*, moderately slow; *sehr langsam*, very slow.

Language of organ pipes. [Organ.]

Languemente (*It.*)
Languendo (*It.*)
Languente (*It.*)
} In a languishing style.

Languettes (*Fr.*) (1) The vibrating tongues of brass or other metal attached to the reed-pipes of an organ or harmonium. (2) The keys or valves of wind instruments generally.

Languid. Language of organ pipes.

Lantum. A name of uncertain derivation given to a modern instrument in form not unlike the hurdy-gurdy, but larger, and having metallic reeds or tongues similar to those of

Digitized by Google

the accordion, concertina, or harmonium. The air is supplied by a rotatory bellows, the handle of which is at the bottom of the instruments and the series of small square buttons or keys which are in front are played upon by the fingers.

Largamente (*It.*) Slowly, widely, freely, fully.

Large. A note formerly in use, of the value of two longs. [Notation.] [Nota.]

Largement (*Fr.*) [Largamente.]

Larghetto (*It.*) *Lit.*, rather broadly, widely. At a slow pace, but not so slow as *largo*.

Larghissimo (*It.*) Exceedingly slow, as slow as possible.

Largo (*It.*) Slow, broadly.

Largo assai (*It.*) Quite slow.

Largo di molto (*It.*) Very slow.

Largo ma non troppo (*It.*) Slow, but not too slow.

Larigot (*Fr.*) (1) An old name for the flageolet. (2) An organ stop of 16 inches pitch.

Larynx is the organ of voice; the organ by means of which we produce vocal sounds. It is situated at the top of the trachea or wind-pipe, of which it forms a continuation, and its position is known popularly by that of the "Adam's apple," the prominence of one of the cartilages or masses of gristle which form it.

Fig. 1.

FRONT VIEW OF HYOID BONE, LARYNX, TRACHEA, AND BRONCHI IN THEIR MUTUAL RELATIONS.

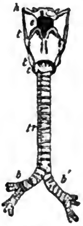

h. Hyoid bone; *e*. epiglottis; *t*. thyroid cartilage; *c*. cricoid cartilage; *tr*. trachea; *b*. right bronchus; *b'*. left bronchus.

The essential parts are two semi-lunar membranes, placed with their flat sides next to one another, and called the "vocal cords," and the cartilages which support them. The accessory portions are certain other cartilages, the muscles which move these cartilages, and the nerves and blood-vessels by which the whole structure is nourished and furnished with sensation and motion.

The cartilages are nine in number, six paired, three unpaired. The unpaired cartilages are the thyroid, cricoid, and epiglottis; the paired are the two arytenoids, the two "cornicula laryngis" or cartilages of Santorini, and the cuneiform cartilages of Wrisberg.

Fig. 2.

BACK VIEW OF LARYNGEAL CARTILAGES

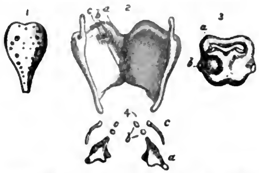

1. Epiglottis. 2. Thyroid cartilage: *a*. attachment of epiglottis; *b*. of false vocal chords; *c*. of true vocal chords. 3. Cricoid cartilage: *a*. facet for arytenoid cartilage; *b*. facet for lower bone of thyroid cartilage. 4. Arytenoid and accessory cartilages: *a*. arytenoid cartilages; *b*. cornicula laryngis or cartilages of Santorini; *c*. cuneiform cartilage or cartilage of Wrisberg.

The *cricoid* cartilage is ring-shaped, as its name (κρίκος, a ring) implies; it is the base of the support of the other cartilages. It is not equally deep in all dimensions, but is much deeper, from above downwards, behind than in front. At the hindermost part are two small smooth elevations, and lower down are two other smooth surfaces. The former pair of these smooth surfaces serve as the articulations or joints for the arytenoid cartilages which are perched on them and work with great freedom; the latter pair are the joints for part of the thyroid cartilages. The movements of the latter pair are not so free as those of the former, but are limited to a movement round an imaginary axis running across and through the cricoid cartilage, being thus a horizontal and transverse axis.

The *arytenoid cartilages* (ἀρύταινα, a pitcher) are perched on the two smooth elevations on the hinder part of the upper border of the cricoid cartilage. They are more or less pyramids of three sides in shape, but of the sides which form their base, two, viz., the posterior and internal, are at right angles to one another. The anterior tips are called the "processus vocales," or vocal processes, and to them are attached the vocal cords, of which we shall speak further on. On the summit of each of these pyramids are placed two other small cartilages, the "cornicula laryngis" (little horns of the larynx) or "cartilages of Santorini."

The *thyroid cartilage* (shield-like, from θυρεός, a shield) is the largest of the cartilages of the larynx. It is of a somewhat complicated shape. Take a thin book—a note-book will do very well—open it in the middle, place it upright, so that the two

Digitized by Google

covers shall be at an angle somewhat more acute than a right angle; cut away the upper third of the back, or junction of the covers, so as to leave a notch above, and to each of the corners stick a piece of thin wood, such as a pencil, the two upper pencils pointing upwards, the lower downwards, and you will have a fair idea of the thyroid cartilage. The two lower horns articulate or join with the two lower smooth surfaces or "facets" on the cricoid cartilage; to the two upper horns, ligaments are attached, which suspend the thyroid cartilage to the hyoid bone, a horseshoe shaped bone, with its convexity forward, forming the base of the tongue. The anterior part of the thyroid cartilage projects much more in adult men than in women or children, and the upper part of the anterior ridge, with the notch, form the projection known as the "pomum Adami," or Adam's apple, from an old tradition that the forbidden fruit stuck in his throat. The thyroid cartilage at the passage from youth to adult age grows very suddenly larger in men, and as suddenly but to a less degree in women. On this relative size and prominence of the thyroid cartilage depends the deepness of the voice, as will be hereafter mentioned.

Fig. 3.

BACK VIEW OF HYOID BONE, LARYNX, AND TRACHEA IN THEIR MUTUAL RELATIONS

(Only cartilages and bones shown).

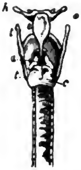

h. Hyoid bone; *e.* epiglottis; *t.* thyroid cartilage; *a.* arytenoid cartilages; *c.* cricoid cartilage.

The *epiglottis* (ἐπὶ on, and γλῶττα the tongue) is a soft cartilage situated at the back of the tongue. In shape it resembles what is called technically in botany an "ovate" or "obcordate" leaf, *i.e.*, it is oval above, and has a tail narrowing gradually from its lower end. This tail is prolonged by fibrous tissue and attached to the thyroid cartilage in the angle between its two sides, just below the notch. The front surface is free in its upper part, but below it is attached by an elastic ligament to the back of the hyoid bone. Its lateral or side borders are free at the uppermost part, but somewhat lower down a fold of mucous membrane (the skin lining any interior cavity is called by this name) runs backwards and joins the summit of the arytenoid cartilage of each side. This fold is called the glosso-epiglottidean fold. In these folds lie two small conical cartilages, the *cuneiform cartilages* or the "cartilages of Wrisberg." Below the level of the hyoid bone on the posterior surface is placed a pad of fat and mucous membrane called "the cushion of the epiglottis," functionally a very important structure.

Fig. 4.

INTERIOR OF LARYNX FROM BELOW.

(About life-size.)

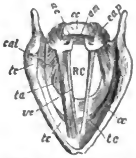

cc. Cricoid cartilage; *ac.* arytenoid cartilage; *tc.* thyroid cartilage; *am.* arytenoid muscle; *cap.* crico-arytenoideus posticus muscle; *vc.* vocal cord; *ta.* thyro-arytenoideus muscle; *cal.* crico-arytenoideus lateralis muscle.

The *true vocal cords*, otherwise called the "vocal cords," are two semi-lunar membranous folds which project on each side towards the middle line; they are attached behind to the tips of the "processus vocales" of the arytenoid cartilages, and in front, close together, to the angle between the two sides of the thyroid cartilage just below the notch. They are somewhat complex in structure, the basis of them is formed by two elastic bands near their edges, and both along these bands, the "inferior thyro-arytenoid ligaments," and more laterally, are muscles running in the same direction, which will be described later. The whole is covered by mucous membrane.

Fig. 5.

VERTICAL TRANSVERSE SECTION OF LARYNX.

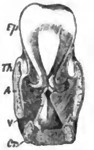

ep. Epiglottis; *th.* thyroid cartilage; *a.* ventricle of larynx; *v.* vocal cord and its free edge; *cr.* cricoid cartilage.

The *false vocal cords*, called also the "superior vocal cords" are two folds of mucous membrane of the same general shape and direction as the former; they are attached behind to the arytenoid cartilages above, and more laterally than the former, and in front to the thyroid cartilages just above the insertion of the true vocal cords. They enclose a ligament, the "superior thyro-arytenoid ligament." Their inner edges do not approach the middle line so nearly as the true vocal cords. Between the upper and lower vocal cord of each side is the opening to the saccule or ventricle of the larynx, a small saccule or cavity, as its name implies, containing some gland cells for secreting mucus. It only remains to say that all the structures are covered with mucous membrane.

Muscles. The larynx is covered with many muscles, but only those which are concerned in the production of the voice will here be mentioned.

Fig. 6.

SIDE VIEW OF LARYNX FROM RIGHT SIDE.

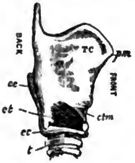

tc. Thyroid cartilage; *pm*. pomum Adami; *cc*. cricoid cartilage; *ct*. crico-thyroid muscle; *ctm*. crico-thyroid membrane; *t*. trachea.

The *crico-thyroid* rises from the fore part of the outside of the cricoid cartilage and running back, spreads itself on nearly the whole of the lower part of the thyroid cartilage.

The *posterior crico-arytenoid* rises from the back part of the outside of the cricoid cartilage and is inserted into the outer or lateral angle of the base of the arytenoid cartilage, sometimes called the "processus muscularis," it also extends some distance along the back part of the base of the arytenoid cartilage.

The *lateral crico-arytenoid* rises from the upper border of the side of the cricoid cartilage, and is inserted into the front side of the lateral projecting part of the base of the arytenoid cartilage nearly as far as the "processus vocalis."

The *arytenoid* is a single muscle, while all the other laryngeal muscles are paired; it runs across at the back of the arytenoid cartilages between them and joins them.

The *aryteno-epiglottidean* muscles rise from the lower and outer angles of the back of the arytenoid cartilages, run across to the upper and outer part of the base of the opposite arytenoid cartilages, crossing each other as they do so (like a pair of braces), and some fibres run no farther but are attached there; the rest pass forwards in the aryteno-epiglottidean fold and are inserted into the sides of the epiglottis.

The *thyro-epiglottideus* is composed of two parts, rising from the inner aspect of the fore part of the thyroid cartilage and having the following course, some (thyro-epiglottideus major) curve outside the saccule of the larynx and are inserted into the side of the epiglottis, others run straight up and are attached to the epiglottis on each side of the cushion and below the insertion of the larger muscle. This is called the "thyro-epiglottideus minor."

The *thyro-arytenoid* is likewise divided into two portions. Part of it (thyro-arytenoideus internus) runs just along the outer border of the true vocal cord. The other part (externus) runs from the arytenoid cartilage more externally, and before reaching the thyroid cartilage it surrounds the saccule of the larynx, it is therefore above as well as external to the former portion. Some of its fibres seem to lie in the false vocal cord.

The *nerves* of the larynx are two in number, called respectively the superior, and the inferior or recurrent laryngeal. Both of these are branches of the great vagus or pneumogastric nerve which rises from a special nucleus or mass of nerve cells in the medulla oblongata or upper portion of the spinal cord within the skull where this expands to form what is known as the "fourth ventricle." It leaves the cavity of the skull by the same opening as the internal jugular vein, which takes the greater part of the blood from the brain to the heart, and communicates with most of the large nerves in this situation. The pneumogastric nerve gives nerve-supply to the pharynx, gullet, stomach, liver, spleen, larynx, windpipe, lungs, and heart, and extends down as far as the middle of the abdomen.

The *superior laryngeal* nerve rises from the pneumogastric nerve just below the exit from the skull, and divides into two divisions called external and internal respectively. The *external laryngeal* nerve sends branches to some of the muscles which preside over swallowing, and ends in the crico-thyroid muscle. The *internal laryngeal nerve* gains the inside of the larynx above the thyroid cartilage, supplies the lining membrane of the larynx, and the arytenoid muscle, and sends a branch which joins the recurrent laryngeal nerve.

The *inferior or recurrent laryngeal nerve* has a most remarkable course. We have mentioned that the pneumogastric nerve supplies the lungs and heart; just after reaching the cavity of the chest, while on a level with, and in front of the large blood-vessels coming from the heart, it gives off on each side a large branch, the recurrent

Digitized by Google

laryngeal nerve. This nerve dives underneath the arch of the aorta (the main blood-vessel rising from the heart) on the left side, and under the subclavian artery (supplying the arm) on the right side, and then runs upwards to the larynx. Thus the fibres of the recurrent laryngeal nerve, which come down in the pneumogastric nerve from the medulla oblongata (for each separate strand in the smallest nerve has a separate connection with its nervous centre, brain, spinal cord or ganglion, a nerve being like an electric cable, seemingly one rope, but really composed of innumerable insulated wires) curve back and run upwards again. Thus every separate movement of the muscles of the larynx is due to a nervous impression which travels from the medulla oblongata down into the chest and then up to the larynx again. This seemingly purposeless length of course will be explained when the development of the larynx is spoken of. On their way to the larynx, the recurrent laryngeal nerves give twigs to the gullet and wind-pipe, and when they have reached the larynx they supply all the muscles except the crico-thyroid.

Epithelium. The mucous membrane of the larynx is lined with a layer of small cells called epithelium (which is the name which it bears on all internal parts of the body; on the skin it is called epidermis). Each of these cells has a little hair-like filament which continually keeps lashing upwards, so as to drive any mucus towards the mouth. This kind of epithelium is found on the air-passages generally as well as in some other parts. Above the vocal cords it gradually loses these hair-like filaments from which its name, ciliated (provided with eyelashes), is derived. These filaments are absent over the true vocal cords, as if to indicate that their function is special (Quain, Vol. I. Fig. 245, p. 326).

Fig. 7.

CILIATED EPITHELIUM OF RESPIRATORY MUCOUS MEMBRANE.

A, vertical section of epithelial lining of human trachea magnified 350 times.

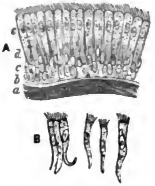

a. b. Subjacent membrane; *c.* lowest or spheroidal cells; *d.* middle or oval cells; *e.* superficial, elongated, and ciliated cells; B. separate columnar and ciliated cells.

Development. The larynx is essentially an apparatus for closing the wind-pipe, and in some of the lower animals has this function alone. The muscles, except the crico-thyroid and posterior crico-arytenoid, which lie outside of the calibre of the tube, are mere differentiations of a circle of muscular fibres called a "sphincter," such as surround, in some form or other, all the entrances and exits of the body. The cricoid and thyroid cartilages are probably developed in two separate halves.

Fig. 8.

DIAGRAM OF AORTIC OR BRONCHIAL VASCULAR ARCHES OF MAMMAL, ACCORDING TO RATHKE.

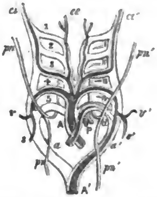

"A. P. primitive arterial stem or aortic bulb, now divided into A. the ascending part of the aortic arch, and P. the pulmonary; *a.* the right, *a'.* the left aortic root; A'. the descending aorta. On the right side 1, 2, 3, 4, 5, indicate the five bronchial primitive arterial arches; on the left side I. II. III. IV., the four bronchial clefts, which for the sake of clearness have been omitted from the right side. It will be observed that while the 4th and 5th pairs of arches rise from the part of the aortic bulb or stem, which is at first undivided, the 1st and 2nd and 3rd pairs are branches above *c*, of a secondary stem on each side. The permanent systemic vessels are represented in deep shade, the pulmonary arteries lighter; the parts of the primitive arches which have only a temporary existence are drawn in outline only. *c.* Placed between the permanent common carotid arteries; *ce.* the external carotid arteries; *ci. ci'.* the right and left internal carotid arteries; *s.* the right subclavian rising from the right aortic root beyond the fifth arch; *v.* the right vertebral, rising from the same spot opposite the fourth arch; *v'. s'.* the left vertebral and subclavian arteries, rising together from the left or permanent aortic root opposite the fourth arch; P. the pulmonary arteries rising together from the left fifth arch; *d.* the outer or back part of the left fifth arch forming the ductus arteriosus; *pn*, *pn'*, the right and left pneumogastric nerves descending in front of the aortic arches, with their recurrent branches represented diagrammatically as passing behind, with a view to illustrate the relations of these nerves respectively to the right subclavian artery, and the arch of the aorta and ductus arteriosus."

As to the recurrent laryngeal nerves, we must premise, that firstly the early embryo has no *neck*, the head abutting on the trunk, thus the heart is placed nearer the head; secondly, the arrangement of the blood-vessels corresponds strikingly with that which is permanent in fish. In them the heart sends off one large vessel which runs along the base of the gills and sends off branches on each side which run along the gills, and at the opposite extremity of them turn down and join in a common trunk which runs along the body and supplies it with blood, purified by having been aërated in passing through the gills. This is substantially the same in the human embryo, which possesses not true gills indeed, but slits in the neck just like the slits between the gills of a fish; along the bridges between these slits (five in all), corresponding to the gills, run the branches of the great blood-vessel. The further changes need not here be mentioned, but it will be sufficient to add that the last or hindmost of these arched blood-vessels of the branchial or gill-like bridges on the left side subsequently becomes the arch of the aorta, while the same arch on the right side is obliterated, the last but one becoming the subclavian artery. The larynx at the early stage is nearly or quite on a level with the lowest of these arches, the pneumogastric nerves run in front of them, and the recurrent laryngeal nerve, in its course to the larynx, runs directly, or nearly directly backwards, *i.e.*, towards the spine. As the embryo grows, the head becomes further separated from the body, the larynx is drawn up with it, the neck appears, the slits are filled up; but in the upward movement of the larynx the recurrent laryngeal nerves are drawn up too, and as they are hooked round the lowest arch on each side they are drawn out to a prodigious extent round the arch of the lowest vessel on each side. But the lowest arch on the right side is obliterated and disappears, so the recurrent laryngeal nerve on the right side is caught, so to say, by the last arch but one, the right subclavian artery. The gradual growth of the neck takes place, not only in the development of the embryos of the higher animals, but we can trace it in the progress from the lowest to the highest animals. A fish has no neck, a frog hardly any, a reptile rather more, birds and mammals have well-developed necks.

Growth. During childhood the larynx hardly grows at all, and is of the same size in both sexes; both have equally high-pitched voices, and in neither is that prominence of the "Adam's apple" (which depends on the acuteness of the angle between the two halves of the thyroid cartilage, and on the absolute size of the whole cartilage) observable. At the age of puberty, the passage to adult life, a sudden growth takes place in both sexes, but much greater in males than females; the whole larynx grows much, and the two halves of the thyroid cartilage are set at an acute angle, forming the prominence of the "Adam's apple," the notch between them becoming deeper. The larynx nearly doubles its size in males, becoming about one-third larger in females. It will readily be seen, that the growth of the thyroid cartilage, and especially its increase in prominence, implies a corresponding lengthening of the vocal cords, hence the increased deepness of voice at puberty which is always noticeable in males, sometimes in females—for a contralto voice in a child is a thing never heard; this is acquired at puberty. In old age a bone-like deposit is apt to form in all the cartilages of the larynx except the epiglottis, first or most often in the thyroid, next in the cricoid, then in the arytenoid. To this is due the cracked quality of the voice of old age, the quality of voice depending of course largely on the quality of the sounding bodies.

Physiology. The *crico-thyroid* muscles rotate the front of the thyroid cartilage on the cricoid, forwards and downwards. Since the pivots on which the thyroid works are below the attachment of the vocal cords, to the arytenoid, and therefore practically to the cricoid cartilage, any forward and downward movement of the thyroid will tend to lengthen the vocal cords. This will be evident from the diagram (Huxley's El. Phys. Fig. 50). If the finger be placed on the space in front, between the thyroid and cricoid cartilages, and a note and then a considerably higher one be sung, the space will be found to be considerably diminished; in other words, the crico-thyroid muscles will have pulled downwards and forwards the thyroid on the cricoid, and in so doing stretched the vocal cords, the increased tension producing the higher note. As before said, the external laryngeal nerve supplies the crico-thyroid muscles.

The *thyro-arytenoid* is the opponent of the crico-thyroid; it rotates the front of the thyroid cartilage upwards and forwards, and in so doing relaxes the vocal cords. Besides this action as a whole, the outer and upper part presses on the ventricle of the larynx, and tends to empty it of any contents; the inner portion, lying in the vocal cord (the thyro-arytenoideus internus) renders straight and tense the very edge of the vocal cord after the thyroid and arytenoid cartilages have been fixed by other muscles, and the length of the vocal cord thereby determined. By the contraction of its lower fibres, lying below the level of the vocal cord, the column of air might be narrowed, and the tone raised; for this is the result of such a narrowing of the column of air before reaching the resonant membrane. It is by this muscle that the peculiar quality of the notes of the "head-voice," as opposed to the "falsetto voice," is probably produced, the edge of the vocal cord being very tight and capable of very

Digitized by Google

rapid vibrations; the comparative thinness of quality is owing to the smallness of the bulk of the vibrating body. The "crack" in the voice which sometimes occurs during the attempt to sing a high note, and is so annoying to performer and audience, is perhaps due to the sudden paralysis of this muscle, which strikes work when overstrained. But it may also be due to the vocal cords *touching* each other, nodes being instantly produced; this is more probable. This muscle is supplied by the recurrent laryngeal nerve.

The *arytenoid* muscle draws the arytenoid cartilages together, and also tends to prevent their rotation. It is supplied by the superior and recurrent laryngeal nerves.

The *aryteno-epiglottidean* muscles, since between them they embrace the whole orifice of the air-tube, act as a sphincter, and in so doing draw the arytenoid cartilages together and forward, the aryteno-epiglottidean folds together towards the middle line, and the epiglottis downwards and backwards.

The *thyro-epiglottidean muscles* depress the epiglottis, the greater and upper one the upper part, the smaller and lower one the "cushion"; by their action the epiglottis is pulled forcibly on to the top of the vocal cords, completely stopping the passage, as takes place in straining or in swallowing. They are supplied by the recurrent laryngeal nerve.

Before describing the action of the lateral and posterior crico-arytenoid muscles, we must premise, first, that the pivot on which the arytenoid cartilages move is a very loose joint, so that the arytenoid cartilages are able to be moved bodily in any direction *without* rotation.

The *posterior crico-arytenoid* muscles, acting alone, pull backwards and downwards the outer angles of the arytenoid cartilages. By so doing they rotate these cartilages (*a a* in following Figs.) on their pivots (*p p*) and separate the front angles, or "processus vocales," separating therefore the vocal cords. The dotted line represents their action.

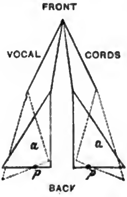

Thus it will be seen that the "glottis" or aperture of the larynx takes a diamond-shape.

But if the posterior arytenoid acts with them it brings the hinder angles of the cartilages together and prevents them from being separated as they would be in rotation, so that the whole cartilages are moved bodily backward and the vocal cords are stretched.

The *lateral crico-arytenoid muscles* have an opposite effect. Acting alone, they pull forward the outer angles of the arytenoid cartilages. By so doing, they rotate these cartilages (*a a*) on their pivots (*p p*) and approximate the "processus vocales," bringing the vocal cords together. The dotted line represents their action:

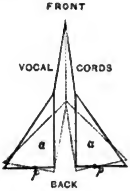

It will be seen that the posterior part of the glottis, between the cartilages, remains open, forming a triangular space with the apex forwards. This was formerly called the "glottis respiratoria" on the false idea that the position of the parts in this state was that of breathing. The space between the vocal cords in front was called the "glottis vocalis"; this name is correct, for it is only this part which assists in producing sound; but the "glottis respiratoria" is only an applicable name so far as it implies that this part is used in breathing, but not in producing sound. The *whole* glottis, however, is "glottis respiratoria" properly, as will be hereafter explained.

By the action of the arytenoid muscle this triangular space is obliterated. There are two other muscles which, though they do not properly belong to the larynx, yet assist in vocalization. These are the sterno-thyroid and the thyro-hyoid.

The *sterno-thyroid* muscles arise from the upper part of the sternum or breast-bone, and are inserted into the side of the thyroid cartilage. Their action is to pull down the thyroid cartilage, assisting the crico-thyroid muscles and helping to stretch the vocal cords. They are supplied by a nerve called "descendens noni," being a descending branch of the ninth or hypoglossal nerve which supplies the muscles of the tongue with motor power.

The *thyro-hyoid* arise from the hyoid bone, which forms the base of the tongue, and are inserted into the side of the thyroid cartilage. Their action is to pull upwards the thyroid cartilage, rotating it and helping the thyro-arytenoidei to relax the vocal cords. They are supplied by a branch of the ninth or hypo-glossal nerve.

The action of the sterno-thyroid muscle may be felt by placing a finger in the middle line of the neck just above the breast bone: in singing a high note, the muscles will be felt to contract.

TABLE OF MUSCLES.

Stretchers of vocal cords. Crico-thyroid, sterno-thyroid (posterior crico-arytenoid + posterior arytenoid), thyro-arytenoideus internus (stretches edge of vocal cords when arytenoid cartilages are fixed).

Relaxers of vocal cords. Thyro-arytenoid (when arytenoid cartilages are not fixed); thyro-hyoid. *Openers of glottis.* Posterior crico-arytenoid (alone). *Closers of glottis.* Lateral crico-arytenoid, posterior arytenoid.

Physiology of the voice. The following remarks are taken from Czermak's book, "Der Kehlkopfspiegel," which contains the best observations that have been made with the laryngoscope.

During ordinary quiet breathing the epiglottis touches the back of the pharynx, so as only to leave one aperture behind and one at each side. Through the aperture behind and in the middle, a deeper view is obtained: one here sees the elevations of the cartilages of Santorini or cornicula laryngis, which surmount the tips of the arytenoid cartilages. In pronouncing the vowel a as in "fate," the epiglottis raises itself and discloses the laryngeal cavity, even the front wall of the windpipe being visible. This is the best position in which to observe the preparations for making any vocal sound.

As soon as one prepares to make a sound with one's voice the arytenoid-cartilages rise and approach each other with surprising rapidity, the "processus vocales" are brought together and with them the vocal* cords by the lateral crico-arytenoid and arytenoid muscles. By practice this can be performed slowly, or even stopped half way, in which case one gets a position of the parts, as in Fig. 3.

It is impossible to actually observe the production of the lowest chest notes, for as soon as we really sing such a note the arytenoid cartilages approach and touch each other, and bend under the edge of the overhanging epiglottis (Fig. 4), which gives us much the same appearance as in quiet breathing.

During the production of the higher chest-notes, especially when sung sforzando, and still more in using the head-voice, it is easy to see into the larynx. It is from these that we infer the position of the parts in the deepest chest notes.

In uttering a piercing cry a free view is afforded into the larynx (Fig. 5). In the middle are the vocal cords closely opposed; then on each side of these a slit, the opening of the ventricle or saccule of the larynx,† between the true and false vocal cords; still farther outside the false vocal cords. The whole arytenoid cartilages, "processus vocales" and all, are closely opposed; and these erected as they are, together with the epiglottis and the intervening aryteno-epiglottidean folds, in which are seen the prominences of the cartilages of Wrisberg, form a sort of additional tube on the top of the larynx, all the parts being (as one's sensations alone would tell one) in extreme tension. On the back of the epiglottis one sees the cushion which touches the insertion of the vocal cords.

On taking breath (Fig. 6) the "processus vocales" are turned outward (posterior crico-arytenoid alone) and we get the diamond-shaped glottis (diagram 1). Sometimes, however, we get the triangular "glottis respiratoria" with a second inverted triangular space in the "glottis vocalis" (perhaps from the action of the lateral crico-arytenoid muscles; the posterior arytenoid and thyro-arytenoideus internus being much relaxed). Fig. 8. In these conditions all the parts are very relaxed.

It is seen that the upper opening of the larynx is formed by a fold of mucous membrane supported by seven cartilages, epiglottis, two cartilages of Wrisberg, two of Santorini, two arytenoid. (Fig. 2.)

In deeper quiet breathing the diamond-shaped glottis is exaggerated, as in Fig. 7.

During quiet breathing the parts of the larynx do not move, and in some cases the glottis is wide enough to admit a finger with ease.

During deep laboured breathing the arytenoid cartilages are brought somewhat together at each expiration, and are thrown into disorderly vibrations like the nostrils under similar circumstances; at each inspiration they are separated to their extreme limits.

---

* It is plain that the glottis is smaller in singing than in breathing, and smaller in singing a high than a low note. This is easily proved by comparing the time taken to empty the lungs in each case. The larger the aperture, the shorter the time. No sound is produced when the vocal cords are more than one-tenth of an inch apart.

† Its function seems to be to ensure a free space for the vibration of the vocal cords, and to keep them lubricated.

It is possible, under favourable circumstances, to see as far down the wind-pipe as its division into the two bronchial tubes, that is as far into the chest as the third dorsal or rib-carrying vertebra. (Fig. 7.)

In the shrillest cries the cushion of the epiglottis appears to be pressed down on the front part of the vocal cords, and to shorten their vibrating portion just as a string is stopped by the fingers on the finger-board of a stringed-instrument.

In air-tight closure of the larynx, as in straining (which is best observed in passing from the position of Fig. 5, that of uttering a piercing cry): 1st, The arytenoid cartilages—"processus vocales" and all—are closely pressed together, and with them the vocal cords. 2nd, The false vocal cords apply themselves to each other and to the true vocal cords closely, so as to conceal the openings of the ventricles of the larynx. 3rd, The epiglottis presses its cushion tightly on the top of everything.

(It is curious that whereas these parts of the larynx resent being touched so strongly that the smallest touch of them produces at least a violent fit of coughing, they suffer being touched by one another and by the epiglottis quietly. Czermak asks whether this is due to the *kind* of touch which they suffer, or to the peculiarity of the nervous apparatus through which the reflex action, which a foreign body excites, is produced.)

These changes are partly simultaneous, partly so quickly following one another that we require the greatest watchfulness to observe them. For instance, the false vocal cords cannot be *seen* to touch one another, for the epiglottis bends down and hides them while they are still approaching each other; but it is inferred that they do so, for on suddenly opening the closed glottis they are seen to project as far towards the middle line as the true vocal cords which are known to be closely apposed to each other.

If we stop the closure of the larynx half way, we get the appearance presented in Fig. 9. In complete closure we get the appearance presented in Fig. 10. Later, however, the free edge of the epiglottis sinks, and we get the same appearance as in Fig. 4, the appearance during the production of the lowest chest-notes.

By this complex apparatus we can understand how the glottis can withstand the tremendous pressure put on it by the expiratory muscles in straining, &c. If these are put into action the whole of the parts closing the glottis are lifted up and become convex or dome-shaped upwards, resuming their position when the pressure is relaxed.

In clearing the throat the glottis is firmly closed, and then the expiratory muscles exerting their force, the parts closing the glottis are suddenly relaxed, the epiglottis being, so to say, blown violently up, and all the parts are thrown into violent disorderly vibrations so as to remove any foreign body or mucus.

In swallowing, the glottis is also closely shut. In increasing the strength of a note the cords are slightly relaxed to compensate for the increased pressure of wind which would otherwise raise the note.*

EXPLANATION OF CZERMAK'S FIG. (TAF. III.)

z. Root of tongue.
h.ph. Hinder wall of the pharynx.
œ. Opening of gullet.
e. Epiglottis.
a. Arytenoid cartilages.
e.w. Cushion of epiglottis.
a.e. Aryteno-Epiglottidean fold.
w. Cartilage of Wrisberg.
g.a. Elevation between cartilages of Santorini and Wrisberg.
s. Cartilages of Santorini.
p.v. Processus vocales.
u.st. True vocal cords.
o.st. False vocal cords.
v.m. Ventricles of the larynx.
v.t. Front wall of wind-pipe.
h.t. Hind wall of wind-pipe.
b.d. Right bronchial tube.
b.s. Left bronchial tube.

Fig. 1.*

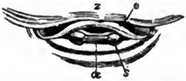

Quiet breathing. Wide glottis. Arytenoids apart and depressed. Epiglottis falling back so as to obscure the view into the larynx.

Fig. 2.*

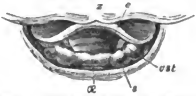

The same, but epiglottis raised by the pronunciation of ā as in "fate," or ee as in "green," but the actual sounding of the latter makes the tongue rise so high, as to obscure the view.

Fig. 3.*

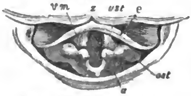

The preparations for sounding the voice after quiet breathing, the process stopped halfway. The arytenoids project and approach one another with free and rapid movements. The glottis is narrowed.

---

* The parts to the right of the middle line of these figures, obtained by a mirror, necessarily correspond to those on the left side of the larynx, and *vice versa.* That which is situated above in the drawing in reality exists in front, that which is below is situated behind.

Fig. 4.*

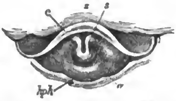

Position during a deep chest-note. The epiglottis lies back and obscures the view of the vocal cords.

Fig. 5.*

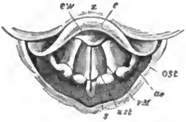

Position during a very high note. Glottis very narrow, all the parts very tense, arytenoid cartilages, aryteno-epiglottidean folds, and epiglottis, forming a sort of additional tube above the floor of the larynx. In the highest possible notes, the epiglottis-cushion is pressed on the front insertion of the vocal cords, shortening their vibrating length.

Fig. 6.*

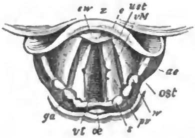

Position of parts on taking a deep breath after singing a high note. All the parts are relaxed and appear thicker, the arytenoid cartilages move apart, the "processus vocales" are turned out, the glottis is larger and diamond shaped.

Fig. 7.*

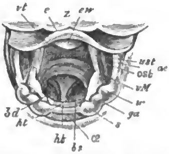

Position in quiet breathing. The same as Fig. 6, but exaggerated. The glottis still larger, large enough to easily admit a finger. The parts do not move during quiet breathing with inspiration or expiration.

Fig. 8.*

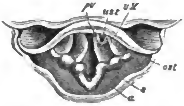

Position during whispering. The arytenoids are seen near together, but not so near as the "processus vocales;" these last are however too far apart to cause a vocal sound. The "processus vocales," being closer together than the rest of the arytenoids, produces a form of glottis, the opposite of that shown in fig. 6, namely, one approaching to that of two isosceles triangles, with their apices apposed; the whisper becomes louder as the "processus vocales" approach one another, until at last all that remains is a triangular space (the hinder of the two triangles), between the arytenoids. In the louder hoarser whisper, the cushion of the epiglottis presses on the front part of the vocal cords and additionally prevents their being thrown into vibrations, though while any chink remains this cannot happen.

Fig. 9.*

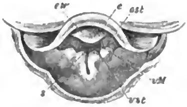

Position in air-tight closure of the glottis; the process stopped half way. The arytenoids and the vocal cords are firmly apposed, the false vocal cords are being approximated, the epiglottis with its cushion is being pressed down on the glottis.

Fig. 10.*

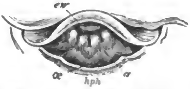

The state of complete closure. The epiglottis pressed firmly on the glottis. The false vocal cords probably, the true vocal cords certainly, closely apposed. When the epiglottis is still further pressed back, we have a view similar to that during the sounding of a deep chest-note (fig. 4), except that a small space exists in the latter case, between the epiglottis and arytenoids, for the passage of the air.

Fig. 11.[*]

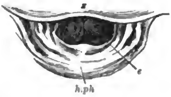

Position at the commencement of the act of swallowing.

The laws governing the vibrations. It has been naturally attempted to identify the vocal cords with some type of sounding bodies, in order to examine the laws governing their vibrations. They have a considerable superficial resemblance to strings, the higher sounds are as in them produced (other things equal) by increased tension. But 1st. no string as short as the vocal cords could give out so low a note, 2nd. the number of vibrations does *not* (as in strings) vary directly as the square root of the tension. Therefore they are not strings.

The human vocal organs are not a flue-pipe, there is nothing to lead us to think that they are, and besides the sound written

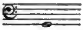

requires an open pipe, nearly 6 feet long, and the case is clearly not one of a *stopped* pipe.

Is it a Reed instrument? The question is best answered by adducing some of the laws which govern them. 1st. The pitch of a reed may be lowered, but cannot be raised, by joining it to a tube, 2nd. it can never be lowered more than an octave, 3rd. after being thus lowered, the note is again raised by a further lengthening of the tube, and again lowered by a still further lengthening, 4th. the length of tube necessary to lower the pitch of the instrument to a given point, depends on the relation between the frequency of vibrations of the tongue of the reed, and those of the column of air taken separately. If a wind instrument depends for its note entirely on the length of its pipe, it must be a flue-pipe, if the range to which it can be altered depends only partly on the length of the tube, and if the range capable of being produced by lengthening the tube is never more than an octave, (the embouchure being constant,) it is a reed instrument.

This latter is the case with the larynx. The reed in the voice corresponds with the "free reeds," since it vibrates freely backwards and forwards, and does not "beat" against an aperture, but unlike those which are in use (Concertina, Harmonium, &c.,) it is formed *with* a tube.[*]

The reed of the voice differs from ordinary reeds in not being a stiff lamina fixed at one end, freely vibrating at another; it is a stretched membrane. Membranous tongues made elastic by tension may have three forms, 1st. a stretched band in an interval between two firm plates, leaving a chink on each side, 2nd. it may be stretched over part of the end of a tube, the other part being occupied by a solid plate, a narrow chink being left between the free edges, 3rd. two elastic membranes may be stretched over the mouth of a tube leaving a chink between them. The last is obviously the case in point. But if the membranes are prolonged in a direction parallel with the current of air, not their edges only, but their whole surfaces are thrown into vibrations. This resembles the larynx still more closely. An instrument on these principles has been constructed, and corresponds very closely in its behaviour with that of the larynx. In such an apparatus, pitch depends on the length, tension, and thickness of the membranes, and though their edges must be close together to produce sounds, the size of the chink has nothing to do with the pitch. A lower note is formed from a pair of such membranes than from one, their pitch is heightened by increasing the strength of the current of air, thus they differ from rigid reeds in which the note is lowered by a similar proceeding.

Their behaviour when furnished with pipes of various lengths, corresponds in the main with that of metallic reeds, but it is not so precisely determinable, as might be imagined. A certain length of pipe suits a certain tone best. A tube *prefixed* to the reed, has the power of altering the tone, but not so definitely as when *affixed*.

Thus the human vocal apparatus is a pair of membranous free reeds, with a prefixed tube and complex affixed resonance tube.

---

[*] Helmholtz in his book "Die Lehre von den Tonempfindungen," 1862, divides membranous reeds into two classes. (1) Those in which the aperture is *closed* by the shock of the air (German "einschlagend") —"beating" reeds; and (2) Those which are *opened* by the shock of the air (German "aufschlagend"—lit. "striking open")—"free" reeds. The beating reeds have the free edges of the limbs of the reed pointing *against* the current of air, the free reeds in the *direction* of the current. The beating reeds give a *lower* tone than they would do if vibrating freely, the free reeds a *higher* tone. The reeds of organ pipes and wooden wind instruments are beating reeds; the human lips in playing brass instruments are *opened* by striking, and are therefore free reeds (aufschlagend). They consequently give higher tones (apart from the question of the selection of a particular harmonic by the resonance tube of the instrument) than when vibrating alone. The Larynx is also a free reed, as will be seen by Fig. 5, v.

All the air-passages above the larynx, the pharynx, mouth, nose, and even the cells in the frontal bone, form part of this resonance tube. These tubes, forming the resonance tube, have an important part to play in vocalisation, though their complexity makes it difficult to determine that part accurately, still that they do play a part is evident from the following facts. In ascending from a low to a high note the whole larynx is lifted towards the skull, the thyroid cartilage being drawn up within the hyoid bone so as even to press on the epiglottis; the space between the thyroid and cricoid cartilages is diminished, the soft palate is depressed and curved forward, the tonsils approach one another.

The aperture of the resonance tube is an important point, the smaller the aperture the lower the tone. In producing low notes the mouth is less open than in high notes. There is probably a certain length of both affixed and prefixed tubes and a certain size of the opening of the affixed tube which is most suitable for each given note, and these matters are partly managed automatically, and are partly the result of practice and education. In producing the lowest notes we retract and depress the tongue with the epiglottis so as to partly cover the lower aperture of the resonance tube.

The *head-voice* is probably produced by the vibration of the free edges of the vocal cords only. Helmholtz thinks this is produced by drawing aside the mucous coat below the cords, the edge of the cords being thus rendered sharper, the weight of the vibrating part less, the elasticity being constant.

The *falsetto voice* is a problem not yet cleared up. Many physiologists seem to have quite lost sight of the fact that it is different from the "head-voice." There are various theories on the subject.

Some believe that the vocal cords only vibrate in part of their length. We have seen this to be possible by means of the epiglottis-cushion. And it has thus much in its favour that some adventitious aid (if we may use the term) seems to be called in, for it is much easier to sing a very high note in the falsetto than in the chest register, the former depending on sheer muscular exertion as we know.

But on the other hand anyone can sing falsetto with his tongue out of his mouth, and consequently with the epiglottis-cushion far away from apposition with the vocal cords.

Others have said that the cords vibrate in segments and give harmonics instead of fundamental notes. This is possible, for some of the fibres of the thyro-arytenoideus internus muscle end at various points along the vocal cord and might favour such a splitting into segments.

Others say that the falsetto differs from the other registers in that only the edges of the vocal cords are thrown into vibration, instead of the whole of them. This is possible, the only objection being that the extreme tension of the very edge of the vocal cords seems to be an explanation involving far more labour than our sensations tell us is expended in the production of the falsetto voice. This seems much more likely to correspond with the production of the "head-voice."

In passing from the highest chest note to the lowest falsetto note, the space between the thyroid and cricoid cartilages in front which was closed suddenly opens, implying *relaxation* of the vocal cords. This is consistent with the hypothesis that they only vibrate in part of their length, for this would require less tension for a shorter membrane. In passing from the chest to the falsetto register, the larynx which was elevated descends and rises again as the falsetto voice rises.

Others have thought that the falsetto voice was the result of a flue-pipe arrangement, the glottis being the embouchure, the air passage the tubes. We have already dismissed the flue theory as untenable.

The person on whom most reliance is to be placed, M. Garcia, himself a singing-master, and the practical inventor of the laryngoscope gives a long account of the falsetto voice. He says that in the falsetto as in the chest register the lower tones are formed by the "processus vocales," as well as by the vocal cords, that these processes (which he calls the "apophyses") gradually approach one another in the middle line, that the vibrating length of the glottis is consequently diminished *pari passu*, that when the vibrating glottis is composed of the vocal cords and only the tips of the processus vocales, the weak uncomfortable tones well known to singers are formed, that when the processus vocales are so closely apposed throughout that only the vocal cords vibrate, the head voice is formed. "If we compare the two registers in these movements, we shall find some analogies in them; the sides of the glottis, formed at first by the apophyses and their ligaments, become shorter by degrees, and end by consisting only of the ligaments. The chest register is divided into two parts, corresponding to these two states of the glottis. The register of falsetto-head presents a complete similarity, and in a still more striking manner. On other points, on the contrary, these same registers are very unlike. The length of the glottis necessary to form a falsetto note, always exceeds that

Digitized by Google

which produces the unison of the chest. The movements which agitate the sides of the glottis are also augmented, and keep the vibrating orifice continually half opened, which naturally produces a great waste of air. A last trait of difference is in the increased extent of that elliptic surface (the glottis). All these circumstances show in the mechanism of the falsetto a state of relaxation which we do not find in the same degree in the chest register.

"When the external fibres of the lateral crico-arytenoid muscle remain inactive, we produce the falsetto. The lips of the glottis, stretched by the horizontal bundle of the thyro-arytenoid, come in contact by their edge alone, formed at once by the ligament and the apophyses, and offer little resistance to the air. Hence arises the great loss of this agent, and the general weakness of the sound produced here. During the chest register the vocal ligaments are stretched, and are in contact to an extent corresponding with the depth of the anterior apophyses of the arytenoid, whilst in the falsetto the edges alone of the ligaments are stretched and apposed."

In reviewing these various theories, we shall be assisted in arriving at some sort of an opinion by eliminating such of them as seem to us obviously untenable. The next theory to be discussed is that which states that only the edges as distinct from the whole of the vocal cords vibrate; this we shall put aside as accounting more satisfactorily for the "head voice," which is quite a different thing, especially when we remember the muscular effort necessary for producing a high head note as compared with the relative ease with which the same note can be produced in the falsetto voice, and also when we keep in view the fact above mentioned, that in passing from chest to falsetto register the crico-thyroid space enlarges, implying *relaxation* of the vocal cords. Let us now review the theory which sets forth that in producing the falsetto voice we set the vocal cords vibrating not in one segment, but in many segments or nodes. In the first place, this is not inconceivable, but we must freely admit that we do not know the mechanism by which it would be produced. Garcia states that the length of the vocal cords necessary to form a given falsetto note always exceeds that which produces the unison of that note in the chest register, and adds that the falsetto note is accompanied by a *relaxed* state of the glottis; now if the length is increased while the cords are relaxed, we have two circumstances which would render the pitch *deeper* if the cords vibrated in their entirety, but both these factors would be favourable to vibration in *segments*. But there is another reason for supposing this, which seems to us worthy of consideration, though it does not depend on direct observation. When the voice "cracks" in singing a high note it flies up not into the head, but the falsetto register. Now we know that this "crack" is produced by the vocal cords coming into contact and splitting into "nodes," just as a violin string will sound harmonics if lightly touched with the finger (this crack is due to the fact that a high note requires that the vocal cords should be all but parallel to each other and all but touching, but either owing to a want of delicate muscular coordination or to their being swelled by a cold or other similar cause, they *touch* in some part of their length and nodes are at once formed). Thus we have two modes of production of the falsetto voice, one of which we know. Is it not likely that we produce the same effect (the falsetto voice) by a similar method in both cases (a splitting into vibrating segments or nodes)? We think, therefore, that it is most probable that the falsetto voice is produced by the vibration of the vocal cords not in their entirety but in nodes, though the mechanism by which this is brought about is unknown to us. In Müller's Physiology by Baly, 1842, p. 1013, the nodal theory of falsetto is denied, "the cords can be seen to vibrate in their whole length." But a string sounding harmonics vibrates in its whole length (except at the nodes), though in segments. Also compare ibidem, Appendix, chapter ii. p. 19: "shrieking harmonic tones produced by the opposite points of the cords touching." Oertel, of Munich, has shown, by the use of revolving mirrors that, while in the chest voice the vocal cords vibrate *en masse*, in falsetto they vibrate in segments divided by longitudinal nodes running parallel to their edges, the number of segments varying with the pitch. —(Centralblatt med. Wiss. Nos. 5 and 6, 1878. Med. Times & Gazette, March 30, 1878, p. 342.)

Timbre of voice depends partly on the conformation, partly on the quality of the structures composing the larynx. Garcia also says that the epiglottis plays a very important part, for every time that it lowers itself and nearly closes the orifice of the larynx, the voice gains in brilliancy; and when, on the other hand, it is drawn up, the voice immediately becomes veiled.

Range of voice in man. Garcia gives the following table of the full extent of the human voice:

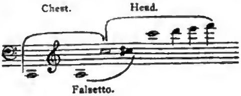

He says, "Let us here observe, that three registers of voice are generally admitted—chest, falsetto, and head. The first begins lower in a man's voice than in a woman's; *the second extends equally in both voices;* the third reaches higher in the female voice." Part of this statement is remarkable, and would seem to imply some essential difference in the physiology of the voice in both sexes, a difference beyond mere dimensions. I must leave it to musicians to criticise his statement, which does not commend itself to me.

The whole of his paper is worth the careful study of every musician. It is found in "Proceedings of the Royal Society," Vol. VII., No. 13, pp. 399-410.

In speech the pitch seldom varies more than a fifth.

The *nasal tone* is produced by bringing the arches of the palate together, and bodily raising the larynx and also the tongue; thus the air is prevented from passing freely through the nostrils, that part of the "resonance tube" being cut off. Some have stated that the air is prevented by this means from passing freely through the mouth, and passes along the nose, the cavity of the nostrils alone vibrating freely; but this is seen to be incorrect when we remember that the nasal tone cannot be produced when the mouth is closed and all the air is driven through the nostrils, while it is produced when the nostrils are held closed, or when they are obstructed by mucus during a cold. Thus the expression "talking or singing through the nose," is physiologically quite incorrect.

The power of musical intonation depends on the power of accurate adaptation of the muscular parts concerned in the production of the voice to a state known to be capable of producing the desired sound. This state is at first recognised by the effect on the ear, and afterwards directly through the "muscular sensibility" of the muscles concerned. The previous experience has thus been acquired through the ear and also from the muscles, and an "idea" of the sound required precedes the production of the sound. Thus eventually an idea of the sound and of the necessary muscular state are presented simultaneously to the mind.

Let us enlarge somewhat on this. A sound is emitted and appreciated by the ear as having a certain relative number of vibrations, and this experience is stored up as a memory. Secondly, attempts are made to reproduce this sound; when the desired sound is produced, the ear recognises it as being the same as that before produced, by means of the memory; and at the same time the sensorium takes cognisance of the state of the muscles of the larynx by means of "muscular sensibility," that sense by which we learn the state of our muscles, by which we know accurately the position of our limbs, &c., unassisted by our sight, and which is essential to a proper performance of any muscular action, especially of those beyond the regulating power of sight. Thus we have two memories stored up in intimate connection, viz., the memory of a definite sound and the memory of a definite state of muscles. This process is repeated for each note and fraction of a note, for each possible sound. By practice and frequent repetition of the process we get the phenomena reproduced in an inverted order, for whereas, in the first instance, hearing came first, muscular action with muscular sensibility second, we now have muscular action with muscular sensibility as the first executors or reproducers of a definite memory, continually checked, secondly, by the hearing. After a time this process becomes automatic, and is still more obviously so when a person is singing from notes, when we naturally have an additional factor or association, viz., that of a symbol with a definite musical sound.

Instrumental music can be expressed in precisely similar terms; it is true that at first muscular sensibility has less to do, for the guiding power of the eyes is possible, but this is only occasionally used as an additional sense when the performer becomes proficient. Again some instruments, viz., those in which the notes are already formed (all instruments but the stringed instruments and the trombone), require less guidance from the ear than those in which the production of each note, both in pitch and timbre, depends entirely on the performer. But these are only quantitative differences after all.

Thus we come back to what we said at first that all correct intonation, whether vocal or instrumental, is checked and regulated by two sensibilities, the acoustic and the muscular, in varying proportions. It is plain from what we have previously said, that musical intonation may be imperfect from various causes; of these the most serious is defect in the appreciation of sounds, since this is first in importance as in time,—a "bad ear" is irremediable; but a person may also play or sing out of tune from several other causes. First this may occur in the performance of a difficult passage. In this instance the muscular adaptability is deficient. Again a person may play or sing a single note out of tune from deficient muscular power, either the note is too high or too low or in other ways too difficult of production, or the performer may be tired with previous singing or playing. But when a person habitually plays or sings out of tune, or fails to keep in

Digitized by Google

tune in performing an easy passage, we at once say correctly "he has no musical ear," that is, his sensorium is incapable of appreciating the delicate relations of musical sounds.

It may not be out of place here to venture a few remarks as to the reason why it is much more irritating to those with good musical ears to hear a person sing sharp than to hear him sing flat. One is, of course, as incorrect as the other, but there is something particularly exasperating in hearing sharp singing. First of all it would be interesting to know whether persons possessing good musical ears, without any knowledge of music, would find sharp singing more painful than flat singing; we should imagine that this would not be the case. On analysing our feelings when a person is singing sharp we fancy we detect in them a certain feeling of indignation at the singer doing something *unnecessary*. This is quite correct, he is using unnecessary muscular force, and we have a natural though not purely rational idea that if he took less pains, exerted himself less, he would sing in better tune—we feel he is committing a "presumptuous" musical sin. On the contrary, when a person sings flat, his muscular exertions are inadequate, and we feel a sort of pity mingled with our annoyance, we feel that he is doing his best, but is not strong enough to sing correctly; in both cases, however, it is generally not actual force which is deficient, but the "guiding sensation," the musical ear; but there is this much correctness in our feelings that a person with ever so good an ear may sing flat from sheer weariness, just as he may be too tired to perform any other muscular exercise properly, while there is not the same excuse for singing sharp.

Comparative Anatomy.—No invertebrate animal possesses a larynx. In *Fish* we have only one or two instances of a larynx, and in them it is connected with the entrance of the duct of the air bladder; these fish are otherwise remarkable, being the Polypterus and Lepidosiren. In the latter the glottis is supported by a sort of laryngeal cartilage. Another fish, named Trigla, is capable of emitting sounds.

In *Amphibia* we find a better developed larynx. In those which keep their gills through life it is extremely simple, as is the whole of the trachea and bronchi. The glottis is a simple slit each side, strengthened by a cartilage, which is sometimes subdivided. Below the glottis is a membranous chamber, the laryngeal cavity strengthened by cartilages, and from this chamber the lungs diverge either immediately or after a short trachea.

In the Frogs the larynx is well-developed, and is present in all but two frogs. The vocal cords are stretched transversely; above and below them is a pouch, and sometimes a cartilage between them. The muscles are briefly classified as constrictors or dilators of the glottis, and one stretches the vocal cords. In the Toad the vocal cords are thin elastic membranes, and consist of two pairs. In many frogs we find pouches attached to the larynx or the mouth, and these serve either as resounding sacs or as reservoirs of the wind supply. Since all Amphibia and Reptiles do not respire as we do, but *swallow* the air, this, which is a laborious process, is economised by an elastic sac, in which the air, having passed through the larynx, is retained, and which, in contracting again, by its elasticity drives the air easily back into the lungs; by this means a continuous croaking is produced.

Reptiles.—The Serpents have no vocal cords, and nothing beyond a hissing can be produced by the passage of the air through their simple glottis.

The Lizards have a better developed larynx. Most have vocal cords, but in many of these they cannot be apposed or stretched.

In Turtles and Tortoises we first find the cartilages divided into cricoid, thyroid, and arytenoid. Some are capable of vocal sounds, some only of a hiss.

The Crocodiles have, again, only two cartilages, a crico-thyroid and two arytenoids; the vocal cords are capable of producing a sort of bellowing tone.

Birds.—Most Birds have *two* larynges, an upper and a lower. The upper larynx is situated as in us, at the top of the wind-pipe. There are several bony and cartilaginous supports, from two to ten in number. One of these represents the thyroid cartilage. The cricoid consists of three bony pieces, the two arytenoids are also bony. With respect to this, we may observe that many more parts are ossified in birds than in other animals. The glottis is simple, and composed of two rigid lips, which do not admit of being stretched, but only approximated. The principal action of the muscles is to close and open this glottis simply, which they can do with great accuracy. The function of this upper larynx is either simply to guard the opening of the windpipe, or at most to modify, and not in the first place to create the voice.

The lower larynx is placed at the bottom of the wind-pipe, where this divides into the two bronchial tubes, and is the true organ of voice. The purpose of its position here is probably to throw the weight which necessarily accompanies a complex apparatus of muscles and cartilages near the centre of gravity, instead of at the end of a long lever

Digitized by Google

like the neck. In the same way the masticating apparatus of birds is not a set of teeth placed at the end of this lever, but a grinding machine, the gizzard. placed at the centre of gravity.

In singing birds the apparatus consists of a double glottis, produced by a bony bar called "pessulus" or "os transversale," which runs across the lower end of the wind-pipe from before backwards, and supports a thin membrane which projects into the calibre of the windpipe, and ends in a free concave edge; it is called the "Membrana Semilunaris."

The muscular apparatus is complex, and in some singing birds consists of five pairs of muscles. In some birds, as the Ostrich and the Vulture, the lower larynx is absent.

In some birds, as the Stork, Crane, Capercailzie and Wild Swan, the windpipe is several times folded on itself, being very long.

The Parrot tribe has a single glottis, with a vibrating membrane on each side, thus resembling ours. The adjacent sides of these are concave, and there are muscles which stretch, approximate or separate these vocal cords. It is perhaps due to this peculiarity, as well as to the fleshy tongue of the Parrots, that they are able to imitate human speech.

Mammals.—The larynx of Mammals in most particulars resembles that of man. In the Kangaroos the vocal cords are feebly developed, and incapable of being stretched. Most Marsupials have little or no voice.

Some Rodents have fairly developed vocal cords, as the Hare and Rabbit. Some, as the Porcupine, have no vocal cords, and it is only at the breeding season that they have any voice, when the male makes a low grunt.

The Sloths have no false vocal cords, the Armadillos have no false vocal cords or voice.

In the Whales and Porpoises, as in many of the Marsupials, the upper opening of the windpipe is modified into a conical projection which rises up and is embraced by the muscles of the soft palate, so that a continuous air-tube is formed from the nose to the lungs, on each side of which fluid or solid food can continually pass to the gullet without getting into the wind-pipe. The purpose of this is obvious. In the Whales the large volume of water which they constantly receive into their wide open mouth, from which they extract their food, and which they eject through their blow-hole, is prevented from getting into their air-passages; in the young Marsupial while in the pouch the milk which it is constantly sucking is similarly prevented from going "the wrong way." Marsupials are born, as is well known, very immature, and while in the pouch are continually suckled; they are by this arrangement saved from the necessity of constantly closing their glottis in the act of swallowing, as is the case with other animals. In those aquatic Mammals called the Sirenia (as the Dugong and Manatee, the herbivorous Whales) the glottis is very small, and T shaped, the transverse slit being above the longitudinal slit.

In the Ass the "bray" is produced by alternate inspirations and expirations, both producing sounds, and is assisted by two large air sacs between the vocal cords and the thyroid cartilage.

The Giraffe only has a voice at the breeding season. The male Deer's larynx enlarges greatly at each breeding season and this growth is associated with a large growth of a gland below it called the thyroid gland, the two producing that graceful prominence in the throat which one observes at that time.

The purring of the Cat is produced by the vibrations of the false vocal cords, which are well developed; the true vocal cords are small and have no membranous part.

Many animals have large air sacs in connection with the larynx, either opening into the sacculus laryngis or placed in front between the thyroid cartilage and epiglottis, and opening between the epiglottis and true vocal cords; or opening between the thyroid and cricoid cartilages. In one monkey—the Mycetes or howling monkey—there are a pair of pouches lodged in cavities in the cricoid cartilage, another pair similarly lodged in the thyroid and extending on each side between the thyroid bone and epiglottis, and into a huge sac in the thyroid bone, and another pair between the glottis and arytenoid cartilages, three pairs in all. Most of the Old World monkeys have a sac in the hyoid bone, but smaller than that of the howling monkey.

The Gibbon has a well-developed larynx, and alone of all the apes can sing a complete octave; moreover the quality of its notes is decidedly musical.

In the Ouran-outan the air sacs in the male extend down over the fore-part of the neck and upper part of the chest, being subdivided into several sacs.

Castration in all mammals which possess a definite voice arrests the sudden development which takes place at puberty. Compare the voice of the ox and of the bull.

W. C. Linnæus Martin, in his work, *A general introduction to the Natural History of Mammiferous Animals*, 1841., *p*. 431, says—

"*Hylobates agilis, Sumatra.*—The voice of this gibbon is extraordinary, not only for its power and volume, but for the succession of graduated tones in which its cry is uttered.

In a room, it is overpowering and deafening: it consists of a repetition of the syllables oo-ah, oo-ah, at first distinctly repeated, and ascending in the scale, but at last ending in a shake, consisting of a quick vibratory series of notes, during which the whole of the animal's frame quivers with the effort to produce them; after this, she appears to be greatly excited, and violently shakes the netting or branch to which she may be clinging; which action being finished she again traverses her cage, uttering the preliminary syllables oo-ah, oo-ah, till the shake again concludes the series. It is principally in the morning that the animal thus exerts this modulated cry, which is, probably, its natural call to its mate, and which, from its strength, is well calculated for resounding through the vast forests. The following observations on the voice of the animal were obligingly presented to the author by Mr. Waterhouse:—

"'I should endeavour to give an idea of the whooping of the gibbon (as far as the music is concerned, but not as regards the quality of sound), by comparing it to the tuning of a harp; first beginning with an E string, and repeating it at short intervals; then being altogether silent for a little time, and then beginning again; next, two strings, as it were, are struck, E and E♯ (or F♮): the second string is then screwed up, by half-notes, until it reaches the octave; the E and F natural, E and F sharp, E and G natural, &c., being struck nearly together. It must be observed that before the upper note arrives at the octave, the animal amuses herself by occasionally descending a few semitones, then ascending again, and so on. But when the octave is once gained, and has been sounded a few times, we may imagine the upper string to be very rapidly let down by semitones; the lower note remaining the same as at first, and the two strings being always struck together. [*Note*: It appears, all through this rapid chromatic passage, as if the animal emitted two notes at a time, as in the music; but this is the effect of the rapid transition from the lower note to the upper.] The rapidity of the descending passage is equal to that of an extremely brilliant shake. The animal then remains quiet for a short time; after which follow two barks, each composed of the low and high E, sounded nearly together.

"'It appeared to me that, in ascending and descending the scale, the intervals were always exactly half-tones; and I am sure that the highest note was the exact octave to the lowest. In this passage the lips were engaged, and rapidly vibrated during its execution.

"'The quality of the notes is very musical; and I do not doubt that a good violinist would be able to give a correct idea of the gibbon's composition, excepting as regards its loudness. The gibbon's voice is certainly much more powerful than that of any singer I ever heard.

"'One more fact I noticed, viz.: That the gibbon is usually a long time before she comes to the rapidly-descending passage; but when she has given it once she soon runs through the preliminary part of her composition, and again comes to the descending passage.'

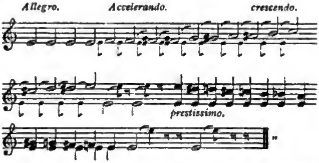

To this we have only to add that the quality is like that of a very powerful male alto, and that the sounds are produced, alternately by inspiration and expiration—the high notes by the former. This alternation, of course, makes it possible for the animal to sing continuously what would otherwise be an impossible passage, exceeding any possible capacity of lungs.

Laryngoscope. *History of the Invention*—Taken principally from Dr. Morell Mackenzie's works.—The use of mirrors for examining the teeth is of exceedingly ancient date. Such mirrors were used by dentists in the Augustan age, and the tubular instrument for examining the various orifices of the body, called a speculum, is also a very old invention; one was found at Pompeii. The use of a mirror is obviously to reflect light into a dark cavity; the use of a speculum is to render patent a tube which is, under ordinary circumstances, collapsed, and to allow the entrance of light and, at the same time, unimpeded vision. Both these principles have been applied variously in the gradual development of the laryngoscope.

In 1743, M. Levret, a French accoucheur, invented a mirror as a means of illuminating the nostrils, throat, ears, &c., for the removal of growths from those parts, but he considered it only an accessory of his method of ligaturing these growths, and he did not value it as a means of diagnosing diseases of the larynx.

In 1804, Dr. Bozzini, of Frankfort-on Main, invented an apparatus for examining the various orifices of the body. It consisted of a lantern and a number of specula; and the speculum intended for examining the

Digitized by Google

throat was curved at a right angle and provided with a mirror in the angle. The speculum was divided by a vertical partition into two passages, one to convey the light to the larynx, another to convey the image from it. This division of the tube was, of course, quite unnecessary; but Bozzini appreciated the things necessary for a laryngoscopic examination, though he over-stated them. This invention was decried by the medical profession, while the public had an almost superstitious regard for it, imagining that it would render possible a direct examination not only of the orifices but also of the cavities of the body. This instrument was soon forgotten.

In 1825, M. Cagniard de Latour tried to examine the Larynx, with a small mirror in the Pharynx and another mirror to reflect the light on to the first, but failed.

In 1827, Dr. Senn made a similar unsuccessful attempt, but without trying any means of illumination.

In 1829, Dr. Benjamin Guy Babington invented an instrument closely resembling the present Laryngoscope. A small mirror on a stem at an angle of about 120° with the stem, was introduced into the Pharynx; to the stem was fastened a spatula (a sort of flat spoon) which depressed the tongue; this spatula he afterwards abandoned. A hand-mirror was used to throw light on to the Pharyngeal mirror. This instrument he called a "Glottiscope."

In 1832, Selligue, a mechanic in Paris who was himself suffering from disease of the Larynx, invented for his own case an apparatus like that of Bozzini, except that instead of employing one tube divided by a partition, he used two tubes. With this Dr. Bennati of Paris professed to be able to see the vocal cords.

In 1838, M. Baumès exhibited a mirror for examining the Larynx and posterior nares.

In 1840, Liston used a mirror on a stem for examining the throat, but apparently never thought of seeing as far as the vocal cords. He preferred the sense of touch to that of sight.

In 1844, Dr. Warden of Edinburgh, who had used a prism for seeing the Tympanum, thought of using one for seeing the Larynx. He used a powerful Argand lamp for illumination, with a second prism attached.

In 1844, also, Mr. Avery of London, tried to use a speculum and reflector for the same purpose; and besides, invented a circular mirror worn on the forehead and perforated in the centre; the mirror reflected the light into the mouth, and the observer looked at the same time through its central hole. The defects of this were that the small Pharyngeal mirror was in a speculum, instead of at the end of a stem, and that a lamp (like Bozzini's) was attached to the mirror on the forehead, instead of being placed in any convenient position behind or at the side of the patient.

In 1854, M. Manuel Garcia, a singing-master still well known, thought of employing mirrors for studying the interior of the Larynx during singing. He made his observations on himself and succeeded admirably. He used a small mirror on a long stem suitably bent, and introduced this into the Pharynx. He directed the person experimented on to turn to the sun, so that the rays might be reflected into the Larynx, but adds that if the observer experiments on himself he should reflect the rays of the sun into his Pharynx with a second mirror, and, therefore, stand with his back to the sun. This second mirror served also for the person experimenting on himself to see the image of his own Larynx. Little notice was taken of Garcia's observations.

In 1857, Dr. Türck, of Vienna, who had read Garcia's paper, tried to use his mirror for clinical purposes. He used only one mirror, and employed no artificial light or second mirror for illumination. He did not succeed, and gave up his attempts after a few months.

Later, in 1857, Professor Czermak, of Pesth, borrowed from Dr. Türck the mirrors which he had discarded, and overcame all difficulties. He introduced artificial light, and the large circular perforated mirror worn on the forehead. He was the first to render the laryngoscope a practically useful instrument. He was favoured by nature for examining his own Larynx, since his Pharynx and Larynx were very large, his tonsils and uvula very small His demonstrations of his own larynx convinced the world of the usefulness of the laryngoscope. Since this time no real improvement has been made, though the subject has been worked continually.

Fig. 1.

THE LARYNGOSCOPE

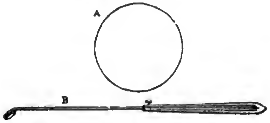

a Shape of mirror, *b* mirror and holder in profile much diminished in size.

The present apparatus consists of a mirror on a stem, bent at a suitable angle (about 120°), which is placed against the back of the pharynx; of a circular mirror, either flat or

Digitized by Google

slightly concave, perforated at the centre, which is worn on the observer's forehead in some way so that his eye comes opposite the central hole, or it may be fixed to the illuminating lamp.

Fig. 2.

LAMP WITH CONVEX LENS FOR LIGHTING.

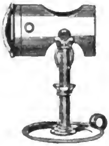

The third necessary part of the apparatus is a lamp, which may or may not be furnished with a double convex lens for concentrating the rays of light. The sunlight on a bright day may be used instead of a lamp. It is possible by casting a strong light on the neck from outside to illuminate the interior of the larynx, so that in this case the larynx is lighted *through* its walls.

The principles employed in making a laryngoscopic examination are the following; Dr. Morell Mackenzie thus well explains them :—

Fig. 3.

(TO ILLUSTRATE EQUALITY OF ANGLES OF INCIDENCE AND REFLECTION.)

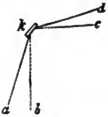

k. Mirror in profile, *ak*. is reflected as *kd*., *bk*. as *kc*., and *vice versa*.

"When rays of light fall on a plane surface, the angle of reflection is equal to the angle of incidence. A small mirror is placed at the back of the throat, at such an inclination that luminous rays falling on it are projected into the cavity of the larynx; at the same time the image of the interior of the larynx (lighted up by the luminous rays) is formed on the mirror, and seen by the observer. The mirror is held obliquely, so that it forms an angle of rather more than 45° with the horizon. The plane of the laryngeal aperture (bounded by the epiglottis, the ary-epiglottidean folds, and the arytenoid cartilages) is also oblique, the epiglottis being higher than the apex of the arytenoid cartilages.

Fig. 4.

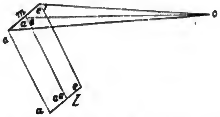

"The above diagram shows the position of the different parts, and explains their reflection. Let *m* represent the plane of the laryngeal mirror, *l* the plane of the upper opening of the larynx, and *o* the observer. In the plane of the larynx, *a* represents the arytenoid cartilages, *ae* the ary-epiglottidean folds, and *e* the epiglottis; the rays from these parts impinge on the mirror, at *á*, *aé*, and *é*, and are thence reflected to the observer at *o*. Thus the epiglottis, which is really the highest in the throat, appears at the upper part of the mirror, the ary-epiglottidean folds appear rather lower and at each side of the mirror, while at the lowest part of the mirror are the arytenoid cartilages. These remarks apply to the vertical reflection. The only inversion which takes place in the formation of the image is in the antero-posterior direction; the part which in reality is nearest to the observer, the anterior commissure of the vocal cords (ac. in B. fig. 5) becomes furthest in the image (ac. in A. fig. 5), and the posterior or inter-arytenoid commissure (pc. in B. fig. 5), which in reality is farthest from the observer, becomes nearest in the image (pc. in A. fig. 5). The symmetrical character of the image which makes it impossible to judge of right and left, and the antero-posterior inversion which actually takes place, often leads people to form erroneous opinions concerning the two sides of the larynx. The lateral relation of parts in the image must now be considered. The mirror being placed above and behind the laryngeal aperture, the rays of light proceeding from the larynx pass directly upwards and backwards, and the patient's right vocal cord is seen on the left side of the mirror, and the left vocal cord on the right side of the mirror (just as the

Fig. 5.

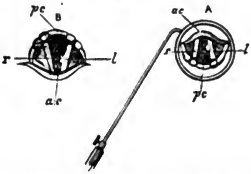

patient's right hand is opposite the observer's left, and his left hand opposite the observer's right)."

Fig. 6.

A LARYNGOSCOPIC SITTING.

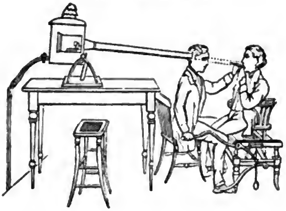

In making a laryngoscopic examination, "the patient should sit upright facing the observer, with his head inclined very slightly backwards. The observer's eyes should be about one foot distant from the patient's mouth; and a lamp burning with a strong clear light should be placed on a table at the side of the patient, the flame of the lamp being on a level with the patient's eyes. The observer now puts on the reflector, fixed generally on the forehead, and tries to throw a disk of light on to the fauces, so that the centre of the disk corresponds with the base of the uvula. He then passes the laryngeal mirror to the back of the throat, so that the posterior surface rests on the uvula, which should be pushed rather upwards and backwards towards the posterior nares. In this position the light is thrown from the mirror into the larynx, while the image of the larynx is observed on the mirror."

Auto-Laryngoscopy. The practice of Laryngoscopy in self-observation requires a second mirror to reflect the image from the laryngoscope to the observer. An ordinary toilet glass does very well.

Lateral Vibrations. [Acoustics, § 14.]

Laud (*Sp.*) Lute.

Laudi spirituali (*It.*) [Oratorio.]

Lauds. [Horæ canonicæ.]

Lauf (*Ger.*) (1) The peg-box in guitars or violins. (2) A run or division in singing or playing. A roulade.

Lauftanz (*Ger.*) *Lit.*, a running dance, a *coranto.*

Laut (*Ger.*) (1) Loud, *forte.* (2) Sound.

Laute (*Ger.*) [Lute.]

Lautenist (*Ger.*) A performer upon the lute.

Lavolta (*It.*) An old Italian dance, much practised in the days of Queen Elizabeth; many of the old dramatists allude to the dance, as Massinger in his play, "The Great Duke of Florence:"

"dance
A light lavolta with her."

It seems to have been the precursor of the modern waltz, as it is thus described in a poem by Sir John Davies, called "Orchestra," 1622:

"Yet there is one the most delightful kind
A lofty jumping or a leaping round,
Where arm in arm two dancers are entwined,
And whirl themselves with strict embracements bound.
And still their feet an anapæst do sound,
An anapæst in all their music's song,
Whose first two feet are short, the third is long."

Lay. A song (*Ger. Lied.*) (*Fr. lai.*) A ballad.

Lay vicars. *Lay clerks. Clerk vicars. Vicars choral. Secular clerks. Chanters. Songmen. Secular vicars. Secundarii.* The officers of a cathedral whose duty it is to sing that portion of the music of the services which can be performed by laymen or men in minor orders. In some of the old cathedrals they formed a corporation, often jointly with the priest vicars. In many cathedrals the vicars choral were formerly in priest's orders. With certain exceptions, in the new cathedrals lay vicars are not in holy orders and are merely stipendiary singers.

Lead. A point or short passage which has to be given out by one particular part. When the word is used as a direction, it calls attention to the importance of that point.

Leader. The name of the principal first violinist in an orchestra; of the chief clarinettist in a military band; and of the chief cornet-player in a brass band. Before the introduction of a separate conductor, the leader of an orchestra was its director, and gave the *tempo* with his fiddle-bow, a custom which has led to the use of a fiddle-bow as a bâton in France. [Conductor.]

Leading note. *Note sensible* (*Fr.*) *Characteristischer Ton* (*Ger.*) *Leitton* (*Ger.*) *Nota characteristica* (*Lat.*) *Subsemitonium modi* (*Lat.*) *Subtonic.* The seventh degree of the ascending major scale. It is called *leading* because of its tendency to rise or lead up to the tonic. It is called sensitive and characteristic, because it forms the essential difference between the modern scale and ancient modes. The Iastian or Ionic mode was the only church scale having a leading note. In consequence of the leading note forming part of the upper of the two tetrachords of which the modern scale is formed, that tetrachord is by some called *characteristic.*

Lebhaft (*Ger.*) Lively, *vivace* (*It.*)

Leçon (*Fr.*) A piece of music intended to develope the taste, skill and power of the performer. A study.

Ledger lines. [Leger lines.]

Legabile, legando (*It.*) Tied, connected, smoothly.

Legatissimo (*It.*) Exceedingly smooth, close, and connected.

Legato (*It.*) Bound, close, connected. A piece of music so marked is intended to be sung or played in an even, smooth, and gliding manner. It is opposed to *staccato*.

Legatura (*It.*) A bind, brace, or tie. *Accolade* (*Fr.*) *Bindungszeichen* (*Ger.*)

Leger or **ledger lines.** *Lit.*, light lines, *léger* (*Fr.*) Short lines drawn above or below the ordinary stave at the relative distances at which the whole lines would be placed.

On and between these lines, notes belonging to passages beyond the extent of the stave are placed. The use of leger lines is comparatively modern in musical notation, for it was anciently supposed that the stave, with a certain clef prefixed, was sufficient for the compass of the voice or instrument using such clef. When the compass was extended, the clef was shifted so that the music might be still expressed within the limits of a stave. To avoid the frequent shifting of clefs, Frescobaldi, in his "Toccate d'Intavolatura di Cimbalo et Organo," Rome, 1637, employs two staves, comprising fourteen lines, as may be seen in the following fac-simile of the opening of the composition quoted in the article "Galliard."

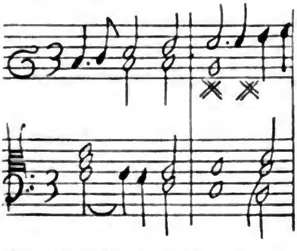

The name *ledger* is also given to account-books because of the light lines drawn in the margin of the pages.

Légèrement (*Fr.*) **Leggieramente** (*It.*) Lightly, easily, gracefully, rapidly.

Légèreté (*Fr.*) **Leggerezza** (*It.*) Lightness, rapidity.

Leggiadramente (*It.*) Beautifully, delightfully, charmingly.

Leggiadro (*It.*) Handsome, beautiful.

Leggierezza, con (*It.*) With much lightness.

Leggiere (*It.*) Very lightly, rapidly.

Leggieramente (*It.*) Lightly, swiftly.

Leggiero (*It.*) Light, easy, swift.

Leggierucolo (*It.*) Rather lightly.

Legno, col (*It.*) With the wood. A direction to strike the strings of violins with the back of the bow.

Leicht (*Ger.*) Easy, light.

Leidenschaftlich (*Ger.*) With passion.

Leise (*Ger.*) Quietly, *piano*.

Leitakkord (*Ger.*) A guiding chord. A chord which suggests an immediate resolution into another, as the chord of the dominant seventh, &c.

Leiter (*Ger.*) *Scala* (*It.*) *Lit.*, a ladder. The scale.

Leitereigene Akkorde (*Ger.*) Chords proper to the scale, that is, chords made up of the notes of any particular diatonic scale, as for instance, the triads on consecutive notes which are major on the fourth and fifth degrees, minor on the second, third, and sixth, and imperfect on the seventh.

Leiterfremd (*Ger.*) Notes foreign to the scale.

Leitmotiv (*Ger.*) A guiding theme associated with a special person, act, or sentiment in an extended composition.

Leno (*It.*) Faint, weak, dull, flexible, pliant.

Lent (*Fr.*) Slow, *lento*.

Lentamente (*It.*) Slowly.

Lentando (*It.*) Becoming slower by degrees; slackening the time.

Lentement (*Fr.*) Slowly.

Lenteur, avec (*Fr.*), **Lentezza, con** (*It.*) Slowly.

Lento (*It.*) Slow.

Lesser. Minor, as: *with the lesser third*, in the minor key; *lesser sixth*, a minor sixth.

Lesson. An assigned task, an exercise or tune for the voice or an instrument. The word formerly was applied to exercises (for the harpsichord or other instruments) of the character now known as Suites or Sonatas.

Lesto (*It.*) Light, lively, cheerful, gay.

License. Permission to break one of the rules or supposed rules of the art of music.

When the gradual growth of music in the church, and among wandering musicians, had proved that the loss of the art as practised by the Greeks was only temporary, attention seems to have been much more directed to the scientific side of music than to the artistic. No doubt the progress of music was considerably retarded by this, but the germ of harmony as found in crude diaphony was not to be stifled, and descant, counterpoint, canon, and fugue, came bit by bit into existence. But writers on music still continued to long for a perfect exactitude in the construction of music which should make mechanical skill of primary importance,

Digitized by Google

and the inventive faculty of secondary value At each stage, therefore, in its development, strong efforts were made by learned professors to build it up on a series of irrefragable laws, and fence it round with sundry warnings to the ignorant or unbelieving. The true artistic spirit can never submit to such restraint, and must ever be searching for new forms of the beautiful. Hence, these laws were from time to time broken, and as professors could but allow that the result was in many cases good, they satisfied their consciences by calling these breakings of the law "licenses," or in other words, they gave *dispensations.*

Any student fairly conversant with the history of music must have been convinced of this fact by the strongest evidence.

The rules of descant were framed as if any further growth of the art were not only impossible, but undesirable if conceivably possible. So too are the rules of counterpoint as first laid down.

In treatises of the fourteenth and fifteenth centuries, the novelties introduced by the then "moderni" are attacked with all the asperity and prejudice which some are directing at the present day against the so-called "music of the future."

A well-known living writer on harmony recommends the student to avoid certain counterpoints, "*although they have the sanction of great classical composers.*" The necessity of granting dispensations is thus made evident, for it is here admitted that *great classical composers* have practically repealed certain rules, yet, say the rule-makers, "the laws are true, so we will say they are broken *by license.*" Examples of this obstructive tendency in most writers on musical subjects might be multiplied to any extent, but it will only be necessary to refer shortly to one more case. "Discords must be prepared" was a law which was enforced until public opinion asserted that many could be delightfully used without preparation. Then, and not till then, students were allowed to use certain dominant discords without preparation *by license.* This dispensation is happily not enforced in later treatises, and dominant discords "may be used without preparation." One author, more honest, though less logical, than his contemporaries, has said that fundamental discords (a name given for some unexplained cause to dominant discords) are called fundamental, "*because* they can be used without preparation."

Of course art *has* its laws, and every composer and painter writes and paints under the direction of innumerable rules gathered by him from broad generalization and experience. He feels their force and obeys them, but whether they can ever be written down and codified is a question of great doubt. Even if music had now reached its climax of perfection, no power could succeed in converting it into an exact science, and rules should be sparingly made where "license" will certainly be craved in order to disregard them. Nor can it be said that there is any moral fault in breaking rules, for music is unlike morality, inasmuch as the artist should first do what his imagination prompts, and afterwards enquire whether it was right or wrong, and moreover, speaking generally, the end will justify the means. On the whole, the history of musical "license" is not very creditable to those to whom has been intrusted from time to time the training of enthusiastic musical youth; and even in these very days it would be well if critics and teachers could be made to understand that many innovations are legitimate growths, and not faults, pardonable only "by license."

Licensing (*musical and dramatic*). The laws relating to the licensing of musical and dramatic performances are so closely connected that it will be convenient to consider them in a single article. In early times there were no such laws, but in 1572 players were declared by statute (14 Eliz. c. 5) to be rogues and vagabonds, unless they were acting as the servants of some baron or person of higher degree. The licensing laws, however, properly so called, take their origin from the revival of theatrical entertainments after the Restoration. At that period, the irregularities of the stage became so great, and continued, notwithstanding the efforts of the Lord Chamberlain and his subordinate officer, the Master of the Revels, so uncontrollable that it was at length thought expedient to include *all* common players of interludes in the Vagrant Act (12 Anne, stat. 2, c. 23). The evil, however, still prevailed, and was heightened by an attempt to enforce the law against a player who was a housekeeper. In 1735, the increased number of theatres (*six* in all!) in which the actors played without any legal authority, and the loose and scandalous nature of the performances, induced Sir John Barnard to bring in a bill "to restrain the number of houses for playing of interludes, for the better regulating of common players." But a clause being proposed to enlarge, or at least to confirm the power of the Lord Chamberlain in licensing plays, and it being understood that the Act could not pass without such addition, Sir John Barnard, who disapproved of the Lord Chamberlain's jurisdiction, abandoned the bill. Two years afterwards Sir Robert Walpole introduced and carried the statute 10 Geo. II., c. 28, to explain and amend the statute of Anne, im-

Digitized by Google

posing penalties for acting without patent from the king or licence from the Lord Chamberlain.

Prior to this enactment the powers of the Lord Chamberlain were limited to the prevention of the continuance of a play actually in course of performance, and he had no *locus standi* to prohibit the first performance of any particular drama. The Act of 1737 (10 Geo. II. c. 28) for the first time conferred upon him this latter power, and to that end it required that all plays were to be sent to him for his examination and approval (sec. 3). The fifth section of this Act, which prohibited the granting of a patent or licence to act elsewhere than in Westminster or in places where the king might be resident, and during such residence only, is said to have been inserted at the instigation of Sir John Barnard. The statute 28 Geo. III. c. 30, empowered justices in general or quarter sessions to grant licences for occasional performances in places where they could not be authorised under the Act of 1737. Licences however under the new Act were not to extend beyond a period of sixty days.

By the Theatres Act, 1843 (6 & 7 Vict. c. 68), all previous enactments are repealed, and it is provided that all theatres for the public performance of stage plays must be licensed by justices, except in London, Westminster, Finsbury, Marylebone, the Tower Hamlets, Lambeth, Southwark, New Windsor, and Brighton, and except in such places as the Sovereign is in the habit of residing in. Such licences were not to be in force during the residence there of Her Majesty, her heirs or successors. In the excepted places the authority of the Lord Chamberlain is preserved (sec. 3), and he may suspend a licence or order a theatre to be closed in case of riot or misbehaviour taking place in a theatre licensed by him, or in a theatre holding a patent from the Crown (sec. 8). He may also forbid the acting of any particular play (sec. 14). Justices' licences are granted in special sessions of which seven days' notice must be given by their clerk to each justice within the division, district, or place. The licences are given under the hands and seals of four or more justices (sec. 5). Justices may make rules for the preservation of order in theatres licensed by them, and may, in case of riot or breach of the rules, order the theatre to be closed for such time as they may think fit (sec. 9).

The University towns of Oxford and Cambridge are the subjects of special exceptions in favour of their respective Chancellors or Vice-Chancellors (sec. 10). Anyone having or keeping any house or place of public resort for the public performance of stage plays, without the necessary licence, is liable to a penalty not exceeding twenty pounds for every day on which such house or place is kept open (sec. 2). By sec. 11 it is provided that every person who, for hire, shall act or present, or cause, permit, or suffer to be acted or presented, any part in any stage play in any place not being a patent theatre or duly licensed as a theatre, shall forfeit a sum not exceeding ten pounds for every day on which he shall so offend. The second section obviously applies to *buildings* used *habitually* for theatrical performances, while the eleventh section applies only to *persons* taking part in plays in unlicensed buildings used only *occasionally* for theatrical performances. It has been held that a booth is not a place of public resort under sec. 2, but a licence is required under sec. 11, or the players will incur the penalty imposed by that section.

New plays and additions to old ones must be submitted to the Lord Chamberlain for his approval (sec. 12). His fee for examination is not to exceed two guineas (sec. 13), and the maximum penalty for performing a play before it has been allowed, or after it has been disallowed, is fifty pounds, with forfeiture of the licence (sec. 15). Licences are to be granted only to the actual and responsible manager of the theatre, whether granted by the Lord Chamberlain or by justices (sec. 7). The fee for a Lord Chamberlain's licence is ten shillings per month, and for a justices' licence, five shillings per month, so long as the theatre remains open (secs. 4 & 6).

The word "stage play" includes every "tragedy, comedy, farce, opera, burletta, interlude, melodrama, pantomime, or other entertainment of the stage, or any part thereof"; but it does not include any theatrical representation in any booth or show which the justices or other authority have sanctioned "in any lawful fair, feast, or customary meeting of the kind" (sec. 23). It seems that whether a certain performance be or be not a stage play within the meaning of this Act is a question of fact and not of law; and that a ballet out of which any regular story could be constructed would probably be held to be a stage play requiring a licence (*see* Wigan v. Strange, *L.R.* 1, *C.P.*, 175, & 35 *L.J.*, *Magistrates' Cases*, 31).

The power of granting theatrical licences as regards London and a radius of twenty miles, was in 1878 vested in the Metropolitan Board of Works (41 & 42 Vict. c. 32, sec. 11 & 12). By the Local Government Act (51 & 52 Vict. c. 41, sec. 7) the licensing of houses or places for the public performance of stage plays was transferred to the County Council. This power or duty may be delegated to justices of the county sitting in petty sessions (sec. 28).

Digitized by Google

Subject to these changes as to the authority which is to grant the licences all the provisions of the Theatres Act, 1843, remain in force.

On July 5, 1892, the London County Council issued a number of rules to regulate the granting of theatrical licences in their district.

Music does not appear at any time to have been so apprehensively watched by the legislature as was the drama, although wandering musicians were ordered by the Vagrant Act of 1744 (17 Geo. II. c. 5) to be regarded as rogues and vagabonds. And the reason of this may be pretty clearly inferred from the preamble to sec. 2 of what is commonly called "The Music Hall Act of 1752," which runs as follows: "And whereas the multitude of places of entertainment for the lower sort of people is another great cause of thefts and robberies, as they are thereby tempted to spend their small substance in riotous pleasures and in consequence are put on unlawful methods of supplying their wants and renewing their pleasures. In order, therefore, to prevent the said temptation to thefts and robberies, and to correct as far as may be the habit of idleness which is become too general over the whole kingdom, and is productive of much mischief and inconvenience. Be it enacted, &c. . . . ." Although many of the dangers referred to in the above preamble are now happily minimised, and the places of entertainment which the legislature of the time appeared to regard with considerable anxiety are no longer monopolised by "the lower sort of people," the Act of 1752 is still in force, and at the present day regulates the licensing of all houses, rooms, gardens, or other places kept for public dancing, music, or other public entertainment of the like kind in the Cities of London and Westminster, or within twenty miles thereof.

The present law on the subject of dancing and music licences may be divided into four branches:—

1. As regards the Cities of London and Westminster, and within twenty miles thereof.
2. As regards the administrative County of Middlesex.
3. As regards local authorities generally which have adopted Part iv. of the Public Health Acts Amendment Act, 1890 (53 & 54 Vict. c. 59, sec. 51).
4. As regards all other districts.

1. As regards the Cities of London and Westminster, and within twenty miles thereof. The present Music Hall Act, 1752 (25 Geo. II. c. 36), was passed for a period of three years, to deal with music, dancing, or similar public entertainments. Its provisions were made perpetual by 28 Geo. II. c. 19, sec. 1. The Act provides that all places kept for public dancing, music, or other public entertainment in the Cities of London and Westminster, or within twenty miles thereof, not licensed by the last preceding Michaelmas Quarter Sessions, are to be considered as disorderly houses, and persons keeping them are to be liable to a penalty of £100, to such persons as shall sue for the same within six calendar months after the commission of the offence. Every licence must be signed and sealed by four or more justices in open court, and must afterwards be publicly read by the Clerk of the Peace, together with the names of the justices subscribing the licence. No such licence can be granted at an adjourned sessions, and no fee or reward is to be taken for the licence. Any constable or other person, authorized by warrant signed and sealed by one or more of the justices, may enter any such place and seize every person who may be found therein, in order that they may be dealt with according to law. Sec. 3 provides that, in order to give public notice of the places that are licensed pursuant to the Act, there shall be affixed and kept up in some notorious place over the door or entrance of every such place so licensed, an inscription in large capital letters in the following words: "Licensed pursuant to Act of Parliament of the Twenty-fifth of King George the Second"; and that no such place, although licensed, shall be open for any of the purposes contemplated by the Act before the hour of five in the afternoon. The affixing and keeping up of the prescribed inscription, and the limitation or restriction as to the time of opening are to be inserted and made conditions of every licence to be granted under the Act, and in case of any breach of either of the conditions, the licence is to be forfeited and revoked by the Justices of the Peace in their next general or quarter sessions, and is not to be renewed; and no new licence may be granted to any person who has once forfeited a licence, or to any other person on his behalf. By sec. 4, nothing in the Act contained is to apply "to the Theatres Royal in Drury Lane, or Covent Garden, or to the theatre commonly called the King's Theatre, Haymarket, or any of them, or to such performances and public entertainments as are or shall be lawfully exercised or carried on under or by virtue of letters patent, or licence of the Crown, or the licence of the Lord Chamberlain." By the Public Entertainments Act, 1875 (38 Vict. c. 21, sec. 1), "Noon" is substituted for "Five in the afternoon" in sec. 3 of the Act of 1752.

Digitized by Google

By the Local Government Act, 1888 (51 & 52 Vict. c. 41, sec. 3), the licensing of houses for music and dancing within the Metropolitan area was transferred to the London County Council, which on July 5, 1892, included music and dancing licences in the rules they then issued as notified above with reference to theatre licences.

2. The adminstrative County of Middlesex is controlled in these matters by its own special Act of 1894 (57 & 58 Vict. c. 15). This Act came into force on January 1, 1895, and deals with "places *ordinarily used* for public dancing or music or other public entertainment of the like kind." By sec. 2 (i.) no house, room, or place is to be kept or used for public dancing, *singing*, music or other similar public entertainment without a licence obtained from the County Council of Middlesex and payment of a fee of five shillings: the fee is not payable in the case of charitable entertainments. The County Council may grant such licences at any meeting convened with fourteen days' previous notice, or at any adjourned meeting (sec. 2, ii.), and the licence lasts for one year or such shorter period as the County Council may decide. The County Council may transfer a licence (sec. 2, iii). Any house, room, &c., kept without a licence is deemed a disorderly house and the occupier is liable on summary conviction to a penalty not exceeding five pounds for every day on which the house is kept or used for any such purpose. Any constable armed with a warrant may enter and apprehend any person found upon the unlicensed premises. The fact that the house is licensed must be notified in the usual way over the door or entrance, and the licence is to specify the hours during which the house may be kept open, but no such house is to be opened before noon or after midnight unless a special occasional licence has been granted under sec. 29 of the Licensing Act, 1872. In case of any breach of the conditions of the licence, the holder is liable on summary conviction to a penalty not exceeding twenty pounds and, in case the offence is continued after conviction, to a daily penalty of five pounds, and the County Council may revoke the licence (sec. 2, ix.). The County Council may grant occasional licences to last not longer than fourteen days, even though the usual fourteen days' notice required by sec. 2 (iv.) has not been given. As regards the administrative County of London, secs. 2 and 3 of the Act of 1752 (25 Geo. II. c. 36), and the whole of the Public Entertainment Act, 1875 (38 & 39 Vict. c. 21), are repealed. The powers conferred by this Act are in addition to and not in derogation of the licensing powers already vested in the County Council.

3. As regards local authorities generally which have adopted Part iv. of the Public Health Acts Amendment Act, 1890 (53 and 54 Vict. c. 59, sec. 51). This area includes places more than twenty miles from London or Westminster, which are outside the administrative County of Middlesex. Originally these licences, when necessary, were granted under the clauses of the various local Acts of Parliament passed for the purpose. Many of these Acts were in force as regards some of the larger provincial cities and towns. The Act of 1890, which, like the Act of 1894, deals only with "places *ordinarily used* for public dancing or music or other public entertainment of the like kind," provides that after the expiration of six months from the adoption by the local authority of Part iv. of the Act (*i.e.*, the part relating to Music and Dancing) all licences for public dancing, *singing*, music, or other public entertainment were to be granted by the Licensing Justices of the licensing district. The fee for each licence is five shillings as in the case of the Middlesex Act, 1894, but there is no exemption from payment of the fee in case of charitable entertainments. The other provisions of this Act are almost identical with those of the Middlesex Act, 1894, the Licensing Justices being throughout substituted for the Middlesex County Council. The Act of 1890, however, contains no provision to enable a constable to enter unlicensed premises and to apprehend persons found there. Part iv. of the Act does not apply within twenty miles of the Cities of London and Westminster (xii.).

4. Dancing and music licences which are granted in any place more than twenty miles from London and Westminster and not included in the administrative County of Middlesex and licences granted in districts which have not adopted Part iv. of the Public Health Acts Amendment Act, 1890, are granted by the County Council of the district under sec. 3 of the Local Government Act (51 and 52 Vict. c. 41).

It is important to remember that *in all cases* these licensing Acts only apply to rooms and places which are habitually used for the purposes contemplated by the Legislature. Occasional concerts require no licences, nor does a room used temporarily for dancing. Separate licences are granted for each specific purpose, and a licence granted for music does not authorise dancing, and *vice versâ*. Singing is referred to as distinct from music in the Acts of 1890 and 1894, consequently under these Acts singing licences can be granted without including music generally.

A few words may be added about *street music*, or the law regulating itinerant performers in the public thoroughfares of

Digitized by Google

the metropolis. By sec. 1 of statute 27 & 28 Vict. c. 55, a previous enactment on the subject (sec. 57 of 2 & 3 Vict. c. 47) is repealed, and in lieu thereof it is enacted that any householder within the metropolitan police district, personally or by his servant, or by any police constable, may require any street musician or street-singer to depart from the neighbourhood of the house of such householder, on account of the illness, or on account of the interruption of the ordinary occupations or pursuits of any inmate of such house, or for other reasonable or sufficient cause. And every person who shall sound or play upon any musical instrument, or shall sing in any thoroughfare or public place near any such house after being so required to depart, shall be liable to a penalty of not more than forty shillings, or, in the discretion of the magistrate before whom he shall be convicted, may be imprisoned for any time not more than three days. And it is lawful for any metropolitan police constable to take into custody without warrant any person so offending. The offender must, however, be given into custody by the person making the charge, and the latter must accompany the prisoner to the station-house and there sign the charge-sheet kept for that purpose. Whenever any person charged with such an offence shall be brought to any station-house during the time when the police-court is shut, the constable in charge may require the person making the charge to enter into a recognisance to appear to prosecute, and upon the refusal of such person to do so, the constable in charge may discharge the party from custody.

It has recently been decided that a householder requiring any street musician or street singer to depart from and out of hearing of the house of such holder on account of the illness or an account of the interruption of the ordinary occupations or pursuits of any inmate of such house, or for other reasonable or sufficient cause, must inform the musician or singer of the cause which justifies the householder in requiring him to depart.

Lichanos. [Greek Music.]

Lié (*Fr.*) Tied, bound.

Lieblich (*Ger.*) Lovely, sweet-toned; as, *lieblich Gedackt*, lovely stopped diapason.

Lied (*Ger.*) A song; but the word is specially applied to that class of song which owes its origin and evolution to the German temperament, poetic leanings, and national tastes. Starting from the outpourings of the Minnesinger on love, nature, and devotional subjects, which covered the earliest period up to the fourteenth century, the *Lied* was carried on by the Meistersinger, the church modes showing, up to this time, a constant influence on their tonality. The *Volkslied*, more rhythmical and truly popular, succeeded in turn, branching, in the sixteenth century, into two well-defined courses, the chorale and the secular songs of the people. Then the power of the advancing art showed itself in the *volksthümliches Lied* and *Kunstlied*, and Germany has ever since maintained a distinctive school of song writers, of which the more recent, such as Schumann, Franz, and Brahms, are the most eminent.

Liederkreis, Liedercyclus (*Ger.*) A cycle or set of songs which, though capable of being used separately, are bound together by the unity of thought in the various stanzas of a poem. Each song should, of course, reflect the spirit of the stanza or stanzas to which it is set, but the contrast between them should not be such as to disturb the harmony of the whole poem. For the purpose of binding the separate songs together, a motive or theme is sometimes repeated, either in the voice-part or the accompaniment.

Liedertafel (*Ger.*) A society meeting for the practice of part-songs for men's voices.

Ligaturæ (*Lat.*) In plain-song, the system of connecting notes (notulæ) for purposes of singing. Notulæ, says Franchinus (lib. i., cap. ii.) are of three kinds, simplex, composita, mediocris. A "Notula simplex" is one which is not joined to another, and is shaped thus, ■, or with a stem (virgula) thus, ■. A "Notula composita" is one that is attached to another, and is of two kinds. If the ligature ascends, that is, if the second note is higher than the first, the first note has no stem, *e.g.*:

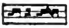

but if the ligature descends, the first note has a stem on its left hand side, *e.g.*:

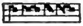

A "notula mediocris" has the form of neither of the preceding, but appears like ordinary "mensurable" descending semibreves, of which there are not less than two in succession, *e.g.*:

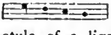

The final notula of a ligature may be written in three ways, 1st, lower than the penultimate, thus:

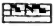

2. Immediately over the penultimate, with a foursided body, thus:

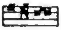

3. Or over the penultimate, on the right

Digitized by Google

side, with a foursided body and a stem on the right side, thus:

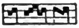

It is wrong to write the ultimate of a ligature thus , but two notes only may be written thus 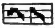; but an ascending oblique body may not be written upwards , at such a thing "usus perhorrescit."

To the middle notes of a ligature no special form is attached, they are either foursided, or oblique. Only one syllable is to be sung to all the notes included in one ligature.

Ligatures [Ligaturæ]. [Notation.]

Ligneum Psalterium (*Lat.*) A series of pieces of wood of graduated lengths, so arranged as to produce the various notes of a scale. These pieces are supported on two strings, running the length of the instrument, which is sometimes made in the shape of a boat. The tone is produced by striking. The instrument is very ancient and universal, modifications of its form being found in every country. It is also called Xylophone (*Gk.* wood-sound), *Strohfiedel*, *Gigelira*, wooden-laughter, &c.

Lilt. An Irish dance accompanied with singing.

> "While the chanter with his merry pipes
> Struck up a lilt so gaily O."

[Ballad.]

Limma (*Gk.*) [Greek Music.]

Linea (*Lat.*) *Linie* (*Ger.*) A line.

Liniensystem (*Ger.*) The stave of five lines.

Linke Hand (*Ger.*) Left hand

Lip, to. To adjust the lips so as to produce the proper tone of wind-instruments played by the mouth.

Lipping. [To lip.]

Lira (*It.*) A lyre. This word was formerly applied to many instruments of the viol class or others having a resonance-box, as, *lira da braccio*, a sort of large tenor-violin; *lira da gamba*, or *perfetta*, a sort of violoncello; *lira rustica* or *tedesca*, or *mendicorum*, a hurdy-gurdy (*Bauernleyer*). *Lira Barberina* was a bowed instrument, invented by John Baptist Doni, so named in honour of his patron, Cardinal Barberini.

Lire (*Fr.*) A lyre or harp.

Liressa (*It.*) An inferior lyre or harp.

Liscio (*It.*) Polished, smooth.

L'istesso (*It.*) The same, as *l'istesso movimento*, the same movement; *l'istesso tempo*, the same time.

Litany, λιτανεία (*Gk.*) A prayer, rogation, or supplication; in particular, that early form of prayer in which a minister recites a petition, and the people answer "Lord, have mercy." In the fourth century litanies were sung in solemn processions by the Eastern Church, but they were not employed by the Western Church till about a century later. Until the time of Mamertus they were only used on special occasions, but he, at the close of the fifth century caused them to be recited on fixed days. *Kyrie eleison, Christe eleison, Kyrie eleison*, is called the Lesser Litany.

Lituus (*Lat.*) A crooked trumpet, used chiefly for giving military signals.

Liuto (*It.*) A lute.

Lo. (1) Abbreviation for *Loco*. (2) One of the syllables used in Bocedisation.

Lobgesang (*Ger.*) A hymn of praise.

Loco (*It.*) In its proper place. A direction to return to the proper pitch after having played an octave higher.

Locrian (*Gk.*) A name sometimes applied to the Hypodorian mode.

Logeum (*Lat.*) λογεῖον (*Gk.*) [Pulpitum.]

Long. An ancient musical character equal to two or three breves. [Notation.]

Long Drum. [Drum.]

Longitudinal vibrations. [Acoustics, § 14.]

Lontano (*It.*) Distant. *Tromba da lontano*, a trumpet played in the distance.

Lo stesso tempo (*It.*) At the same pace.

Loure or **Louvre** (*Fr.*) A dance adapted to an air called "*L'aimable vainqueur*," said to have been a favourite of Louis XIV. Some authors, however, consider it to have been a kind of jig, or a waltz. The name is perhaps derived from the word *lourer*, to bind notes together, to slide.

Lugubre (*Fr.*) Dismal, doleful, lugubrious.

Lu-lu (*Chinese*). The Chinese official collection of treatises on the art of music.

Lunga pausa (*It.*) A long pause.

Luogo (*It.*) [Loco.]

Lures. Ancient Scandinavian trumpets. Some specimens discovered in Denmark would, if straightened, have been six feet in length; in their curved form they were three and a-half feet long. Notwithstanding their great antiquity, they were considerably ornamented and in sufficiently good preservation to give out musical sounds.

Lusingando (*It.*) **Lusingante** (*It.*) **Lusinghevole** (*It.*) **Lusinghevolmente** (*It.*)	Caressingly, in a coaxing manner.
Lusinghiere (*It.*) **Lusinghiero** (*It.*)	Caressing, coaxing.

Lute. *Luth* (*Fr.*), *Laute* (*Ger.*), *Liuto* (*It.*) *Laud* (*Sp.*) An instrument of the guitar family, formerly very popular in Europe. It was used for accompaniments and the performance of solos, duets, &c. It had four to six pairs of strings, each pair tuned in unisons or octaves. The *accordatura* was as follows:

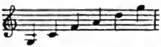

The lower G was omitted on the five-stringed, and both G's on the four-stringed Lute.

The German lutenists called the six strings,

Digitized by Google

counting from the lowest, *Grossbrummer*, *Mittelbrummer*, *Kleinbrummer* *Grossangsait*, *Kleinsangsait*, and *Quintsangsait*, or the five strings without the lower G, *Prime*, *Secund*, *Terz*, *Quart*, and *Quint*. In France, Italy, and Flanders, they were similarly named according to the language of the land, the upper string, upon which the melody was usually made being called *Soprano*, *il Canto*, *Chanterelle*, as well as *Quint*. In England the lowest string was called the *Bass*, the next the *Tenor*, the next the *Counter tenor*, the next the *Great mean*, the next the *Small mean*, and the highest the *Treble*. Gut strings were generally used, for covered strings were not known when the Lute was in common use. Thomas Mace ("Musicks Monument," 1676) describes the strings of his time in these words: "Be careful to get good strings which would be of three sorts, viz., minikins, Venice catlins, and Lyons (for basses).

Music for the Lute was written in Tablature, *q.v.*

The word has been variously derived, from the Latin *ludere*, to play, from the Greek ἀλιευτικός, from the lyre which was a dessicated tortoise, and from the Saxon *hlud* or *lud*, sonorous, but it is most probably from the Arabic *el'ood*, as the instruments came into Europe from the Moors through the Spaniards, who still call it *laud*.

LUTE.

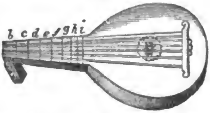

The several frets of the lute were distinguished by the letters of the alphabet, "one for each fret as many as there may be." The frets divided the strings into semitones.

The Orpharion Lute had a larger number of strings than the common lute, and its strings were of wire, instead of gut.

ORPHARION LUTE.

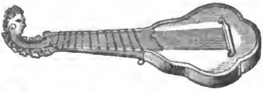

Lutenist. A performer upon the lute. The office of lutenist still exists in the Chapel Royal, but it has been a sinecure since the disuse of the instrument. The revival of the office was made in favour of Dr. Nares, who was appointed in 1780.

Luth (*Fr.*) A lute.

Luttuosamente (*It.*) } Mournfully, sadly.
Luttuoso (*It.*) }

Lydian mode. [Greek Music.]

Lyon strings. Thick spun strings for the basses of lutes or viols. [Lute.]

Lyra mendicorum (*Lat.*) [Hurdy-Gurdy.]

Lyra viol. An obsolete instrument, in form like the ordinary viol, having six strings and seven frets.

Lyre. One of the most ancient stringed instruments. The word lyre (λύρα) does not occur in Homer; he speaks only of the citharis (κίθαρις) and phorminx (φόρμιγξ). The distinction between a citharis (or *guitar*), and a lyre, is that the neck of the former runs behind the upper part of the strings, while the strings of the latter are free on both sides. A reference to the article "Guitar" will show the nature of the cithara. The following illustration of a Greek lyre exhibits the characteristic of the instrument just alluded to.

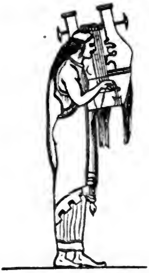

The lyre was also called *testudo* or chelys (χέλυς), because the back was of tortoiseshell, as shown in the next illustration.

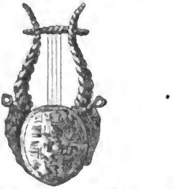

The cross bar at the top was called the yoke ζύγον, Lat. *transtillum*, a little *transtrum* or cross-bar. The uprights were called the horns πήχεις.

The history of the progress of the lyre as a musical instrument, is the history of Greek music itself, and the reader is referred to the article on that subject for information. [Guitar], [Lute], [Greek Music.]

Lyric. Poetry or blank-verse intended to be set to music and sung.

Lyric Stage. A term applied to operatic representations.

M.

M. Abbreviation of *mezzo*, *mano*, *main*, *manual*. M.M. stands for *Maelzel's metronome*, *q.v.*

Ma (*It.*) But; as, *vivace ma non troppo*, quick, but not too quick.

Machalath, *or* **Mahalath** (*Heb.*) This word occurs in the title of Psalms liii. and lxxxviii.; the former is inscribed to the "chief musician upon Mahalath," the latter to the "chief musician upon Mahalath Leannoth." Mahalath is by some authors traced (like Machol), to a root meaning *pierced* or *bored*, hence it is thought these Psalms were accompanied by flutes. It is generally thought that the term *leannoth* refers to antiphonal singing. Other writers consider the titles of these and several other Psalms to be a reference to well known tunes to which they were to be sung.

Machine-head. An arrangement of rack and pinion for the purpose of tightening and keeping in tension the strings of the double-bass, and the guitar, as the ordinary pegs employed to stretch the strings are of unequal leverage.

Machicot (*Fr.*) An obsolete term for one of the *chori ministri minores* of a cathedral, who in singing, added passing-notes between intervals of the plain-song; or, according to others, added a part to the plain-song at an interval of a third or fourth, thus forming a sort of *organum* or *diaphony*. The music thus sung was called *machicotage*.

Machol, *or* **Mahhol** (*Heb.*) A word often found in the Old Testament, associated with "toph" (timbrel), and almost always rendered in the English version by *dances* or *dancing*. But some authorities trace the word to a root meaning *pierced* or *bored*, and therefore consider it to have been a flute. It is not improbable that *Machol* and *toph* may mean "pipe and tabor," but as these two instruments are often associated with dancing, our version, and others which follow it, cannot in any case be said to be incorrect.

Machwerk (*Ger.*) Composition, construction, the result of labour rather than spontaneity.

Madriale (*It.*) A word derived from madrigal, and as in the early operas madrigals were performed between the acts, without necessarily having any connection with them, the word came to be applied to any species of *intermezzo*.

Madrigal. All that can be safely said as to the derivation of this musical term is that a madrigal was originally a lyric poem, and when such poems were set to music the old name was given to the combined music and words. Among the most prominent suggestions which have been made as to its possible derivation are the following: (1) from *madre* (*It.*), mother, a poem or song addressed to the Virgin Mary; (2) from μάνδρα (*Gk.*), mandra (*Lat.* and *It.*), a sheepfold, a pastoral ditty; (3) madrugada (*Sp.*), dawn, a morning song, a *mattinata* (*It.*); Madrigal, a town in Castile. The madrigal as a form of composition was the natural outcome of the division of vocal part-music into two streams, sacred and secular. The separation was slow and not always well defined, especially when composers of madrigals made use of the Church Modes. But the compositions of Dufay and his contemporaries (say from 1400 to 1450) contain characteristics and elements which were afterwards to be of the essence of the true madrigal, and in the Ashmole MS. of the Bodleian Library are two thoroughly genuine madrigals of about the date 1490, "Si je vous eslonge de l'œil" and "Helas! de vous je me doy bien complaindre," containing imitation, independent motion of parts throughout, and an unmistakable effort to make the music a constant reflection of the spirit of the words. They are probably of Franco-Belgian or Flemish origin. The madrigal style, though it continued to thrive for a century at least in the Flemish school, was, as to publication, in very early days transplanted into Italy, France, and England.

Three collections were issued by the inventor of music type, Ottaviano di Petrucci, in 1501-1503, containing compositions by Busnois, Ockenheim, Tinctor, Obrecht, Josquin des Prés, and others. "Trente et une Chansons Musicales" (Attaignant) were published in Paris in 1529, a collection of genuine madrigals of a very high order of merit, by Claudin de Sermisy, Courtoys, Deslouges, Dulot, Gasconge, Ianequin, Jacotin (a volume of whose chansons were issued in the same year), Lombart, and others. In 1555, Palestrina began to issue madrigals, and continued to do so till 1594. In 1556, Orazio Vecchi published in Venice a collection of madrigals; in 1570, Costello published his volume in Paris; in 1588 was published

Digitized by Google

Young's collection of madrigals, "Musica Transalpina," in London, containing madrigals by Byrd, Faignaut, Donato, Palestrina, Filippo di Monte, Luca Marenzio, Orlando di Lasso, Feretti, Conversi, and others; in 1593, Thomas Morley issued his "Canzonets, or Little Short Songs to Three Voyces"; in 1595, his "First Booke of Ballets to Five Voyces"; in 1588, Ferrabosco's madrigals were published; in 1589, Byrd's "Songs of Sundrie Natures"; in 1597, Wilbye's "First Set" of Madrigals; in 1598, Weelkes's "Ballets and Madrigals"; and in 1601 Morley issued the splendid collection of madrigals in honour of Queen Elizabeth, "The Triumphs of Oriana," containing compositions by Benet, Hilton, Tomkins, Wilbye, Weelkes, Kirby, and others.

This outline of madrigal literature, though so scanty, is, however, sufficient to show with what extraordinary rapidity this form of composition took root and flourished in all parts of civilised Europe. What the motet was to the Church, the madrigal was to the chamber, before the existence of concerts or public performances; and their comparison is of the utmost interest. Whilst this extraordinary productiveness proves a marvellously high standard of ability amongst European musicians, the daily performance of concerted vocal music in home life, as a matter of ordinary amusement, gives evidence of the advanced practical musical skill of what we should call "society" in the 16th and early part of the 17th centuries. Everyone is familiar with the passage in Morley's "Plaine and easie introduction to practicall Musicke" (1597), in which *Philomathes* describes the astonishment of a "society-party" when he was compelled to avow his ignorance of the art of music;—"Supper being ended, and Musicke bookes, according to the custome, being brought to the table; the mistresse of the house presented mee with a part, earnestly requesting mee to sing. But when after manie excuses, I protested unfainedly that I could not: everie one began to wonder. Yea, some whispered to others, demaunding how I was brought up: so that upon shame of mine ignorance I go nowe to seeke out mine olde frinde master *Gnorimus*, to make my selfe his scholler." And when it is remembered that such things as "vocal scores" had no existence, all having to sing from separate parts without bars or cues, and with frequent omission of the words when they were repeated, our astonishment and admiration is increased. The story of *accompanied* madrigals is of equal interest and importance. There can be little doubt that instruments of all sorts, though chiefly string-instruments, were of much more common use as a support for the voices than is generally supposed. There can be no question as to this custom in the epoch of Dufay and his contemporaries (15th century), and it seems quite possible that such accompaniments, though not mentioned, were used as a matter of course in the early madrigalian period. When the accompaniment began to be disused, some composers reminded the public that such help *could* be made use of if found necessary by stating that their madrigals were "apt for voyces and viols." It is fully recognised that concerted instrumental music was, in its earliest stages, a close imitation of contrapuntal madrigals, so much so that some of it might almost be described as *madrigals without words;* bearing this in mind, the extraordinary influence of this form of composition can be fully realised. It will be found that the term madrigal includes several forms varying greatly in style; the simply harmonised melody (note against note), such as Festa's "Down in a flowery vale"; the contrapuntal style (closely allied to the motet), as Gibbons's "O that the learned poets"; those in which the voices are often responsive, as Marenzio's "Lady, see on every side"; and those in dance rhythms, by Gastoldi, Morley, and others.

It has been asserted that the madrigal was *invented* by Willaert. But as a matter of fact the madrigal was a natural evolution in the life and progress of vocal part-music; and it is necessary to adjudge a composition to be a madrigal by its character and style, notwithstanding the designation under which it was issued. Many of the Chansons of Claudin de Sermisy, the Gesänge of Hassler, the Ballets of Morley, and Songs of sundry natures of Byrd (1589), are pure madrigals—*e.g.*, "C'est une dure departie," by Claudin (1529), "Vor freuden will ich singen," from Hassler's *Lustgarten* (1601), Morley's "Dainty fine nymph" (1595), and Byrd's "When I was otherwise than now I am." As might be expected, the decadence of the madrigal style corresponds with the decadence of the school of pure vocal Masses, and it is probable that neither of them can ever be revived. Praiseworthy and successful efforts are made to keep alive the taste for, and knowledge of, the literature of madrigals by the Madrigal Society of London (founded 1741), the Western Madrigal Society, the Bristol Madrigal Society, and other bodies. But it cannot be said that their efforts to revive the lost style (by means of prizes, &c.), have been equally successful. The reason is obvious. Whereas instruments of music have evolved in a most extraordinary way since the 16th century, thereby giving composers new means of expression, by their greater variety of tones, greater compass of sounds, and greater facility of execution, the human voice has remained in a stationary condition, and is incapable of achieving more

Digitized by Google

than it formerly did, either as to compass, elasticity, or power of endurance. Hence, the material at the disposal of a modern composer of pure vocal part-music is exactly what it formerly was, and anyone who has taken the trouble to arrange and analyse the melodic phrases, points of imitation, and harmonic progressions of madrigals of the best period, will inevitably come to the conclusion that the material has practically been exhausted, and that to make a new old-madrigal with any hope of originality is almost hopeless. One or two instances of the unexpected appearance of modern madrigals should be recorded, amongst them Walmisley's "Sweet flowers" and "Slow fresh fount" and Pearsall's "Great God of love" and "Lay a garland." But, as a rule, the composer who flatters himself he has written a new "old-madrigal" will be able, if he cares to hunt in the pages of the old masters, to discover that he has been forestalled in every respect, and the public and critics who applaud such imitations and shams only reveal their profound ignorance of the vast literature of *genuine* old madrigals waiting to be traversed.

The student of madrigalian literature will find much untrodden ground in tracing the gradual development of the modern scale. Many of the older madrigals are purely modal, many are evidently in a transitional state; and editors of re-issues are often much puzzled as to when an accidentally raised leading-note should be assumed, or when a whole tone of the scale should be allowed to remain immediately below the key-note. In some cases, where one voice-part seems to require a leading-note, one other of the voice parts makes its introduction impossible.

Madrigale (*It.*) [Madrigal.]

Madrigaletto *or* **Madrialletto** (*It.*) A short madrigal.

Maestà, con
Maestade, con
Maestevole
Maestevolmente
Maestoso
(*It.*) With dignity, majesty.

Maestrale (*It.*) or *magistrale*, a term sometimes applied to the stretto of a fugue when in canon.

Maestri secolari (*It.*) Teachers of secular music.

Maestria (*It.*) Skill, address, authority.

Maestro del coro, or di cappella (*It.*) Choir master, leader, or conductor. [Capellmeister.]

Magadis (*Gk.*) μαγάδις. An instrument of twenty strings, on which music could be played in octaves.

Magadize, to (*Gk.*) μαγαδίζω. (1.) To play upon the magadis. (2.) To play in octaves.

Magas (*Gk.*) μαγάς. (1.) The bridge (Lat. *pons*) of a cithara. (2.) A fret.

Maggiolata (*It.*) A May song. A song sung in celebration of the month of May.

Maggiore (*It.*) Major.

Maggot. One of the later names given to fancies, airs, and pieces of an impromptu character. The most celebrated of these fancies or *whimsical* airs was that by Moteley or Motley, which is subjoined.

From "A choice collection of lessons, being excellently sett for the harpsichord," 1705.

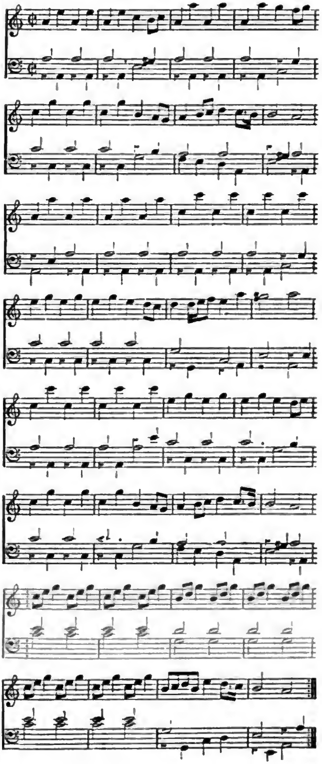

Magrepha. An organ mentioned in the Talmud as having been in existence in the second century. It had ten ventages, each of which communicated with ten pipes, and it was played upon by means of a clavier.

Main (*Fr.*) The hand, as *main droite*, the right hand; *main gauche*, the left hand, abbreviated thus, *m. d.* and *m. g.*

Maître de chapelle (*Fr.*) Choir master. [Capellmeister.]

Majeur (*Fr.*) Major.

Major (*Lat.*) Greater. A major third consists of four semitones, a minor third of three. A major tone is the whole tone having the ratio 8:9; a minor tone, that having the ratio 9:10. Intervals have had the term major applied to them in a conflicting manner. [See Interval.]

Major mode. The ordinary diatonic scale, having semitones between the third and fourth, and seventh and eighth degrees.

Mal (*Ger.*) Time, as, *zum ersten Mal*, for the first time.

Malinconia, con
Malinconicamente
Malinconico
Malinconioso
Malinconoso
(*It.*) With sadness, sorrow, or melancholy.

Mammets. Puppet-shows, usually accompanied with music on a dulcimer. [Dulcimer.)

Mancando (*It.*) Decreasing, dying away.

Manche (*Fr.*) The neck, of a violin or guitar, &c.

Mandola (*It.*)
Mandoline (*Eng.*)
Mandolina (*It.*)
An Italian fretted guitar, so called from its almond shape. There are several varieties, each with different tunings. The Neapolitan, considered the most perfect, has four strings tuned like the violin, *i.e.*, G, D, A, E. The Milanese, next in favour, has five double strings tuned G, C, A, D, E. A plectrum is used in the right hand, and the left is employed in stopping the strings. Mozart in Don Giovanni, wrote an accompaniment to "Deh Vieni" for this instrument.

Mandora (*It.*) A kind of guitar.

Manica (*It.*) Fingering.

Manichord. [Clarichord.]

Manico (*It.*) The neck, of a violin or guitar, &c. *Manche* (*Fr.*)

Maniera (*It.*) Manner, style, method; as, *maniera affettata*, an affected style, *maniera languida*, a languid, lifeless style.

Manière (*Fr.*) [Maniera.]

Manieren (*Ger.*) [Harpsichord graces.]

Männergesangverein (*Ger.*) A society formed for the performance of music for men's voices.

Mano (*It.*) Hand, as *mano destra*, *diritta*, right hand; *mano sinistra*, left hand.

Manual. [Key-board.]

Manuductor. The man who beat time by striking the left hand with the right. Oyster-shells or bones were sometimes used as accessories.

Marcando (*It.*) Marking the time or expression.

Marcatissimo (*It.*) Very marked.

Marcato (*It.*) Marked, or emphasized.

March. A musical composition so arranged as to be suitable for accompanying troops in walking. There are quick and slow marches in duple and triple time, besides marches peculiar to certain nationalities.

Marche (*Fr.*) (1) A march. (2) A sequence, as, *marche des accords*, a sequence of chords.

Marche redoublée (*Fr.*) A double quick march.

Marche triomphale (*Fr.*) A triumphal march.

Marziale (*It.*) Martial, warlike.

Marimba. [Balafo.]

Marionette (*It.*) A puppet.

Marks of Expression. Certain words or signs used in music to regulate the degrees of accent, power, time, or tone, required by the composer to produce the proper effect of his composition. The employment of elaborate marks of expression is a practice of comparatively modern times, none of the earliest writers making use of any beyond the *time signs*, these being all that were then needed, as they suggested a certain pace generally understood. Most of the early printed music had no other directions than *soft* and *loud*, and these only sparingly. Thus, in Croft's "Musica Sacra or Select Anthems in score" 1724, the "First Essay of Publishing Church Musick in England," printed from engraved metal plates, all the directions are in English, as *loud*, *soft*, *fast*, *slow*, *grave*, *brisk* and *lively*, the only foreign words employed being *solo* and *ritornello*. The introduction of Italian music into this country in the eighteenth century necessitated a knowledge of the Italian terms of expression, which led to the gradual employment of a few of the terms so acquired, and Italian words or their abbreviations ultimately superseded the English words to such an extent that very few native terms for expression are to be found in music of modern times. The Italian words originally introduced were in their simplest forms, but as composers grew more fastidious and exacting, the phrases became complicated by the employment of diminutives, expansions, and compounds of terms, so that it became extremely difficult to determine the exact value of the many marks. The doubts that presented themselves to the minds of the performers as to *pace* necessitated the employment of a method by which the time of a piece should be taken, and a string divided into inches was at first used as a standard to

Digitized by Google

regulate the time value of the notes in a bar. This gave rise to the invention of the metronome. [Expression.]

Marqué (*Fr.*), **Markirt** (*Ger.*) [Marcato.]

Marsch (*Ger.*) A march.

Marseillaise (*Fr.*) A song written by Rouget de Lisle, an officer of artillery in the garrison of Strasbourg in 1792. It received its title from having been sung by a party of the Marseillaise club as they entered Paris on the invitation of Madame Roland; the song, though less sanguinary in sentiment than most of the songs of the revolution, was employed as accompaniment to many of the horrible deeds of that, and of later periods, and by association became dangerous enough to be included among the songs prohibited to be sung in France. The tune to which it is set by the author of the words, contains progressions so unusual in popular songs, that it is difficult to account for its general adoption.

Martelé (*Fr.*)
Martellato (*It.*) } *Lit.*, hammered.

Martellare (*It.*) *Lit.*, to hammer. A term applied to *staccato* bowing on the violin, and to the strong percussion of notes of a pianoforte.

Masque. A species of dramatic entertainment in which originally the performers wore masks of peculiar forms suggestive of the allegorical characters assumed. In many instances the masque had no definite design or plot, but depended for its success upon the occasion for which it was written, the wit of the poet who furnished the words, the skill of the musician who supplied the music, and the ingenuity of the machinist and scene painter by whom the stage effects were produced. The early masques were simply acted pageants, but by degrees the genius of such writers as Fletcher and Ben Jonson furnished the poetical groundwork of many masques acted at Court by the children of His Majesty's Chapel Royal and St. Paul's Cathedral. The most beautiful work of this class is the "Comus" of Milton, acted at Ludlow in 1634, and although produced at a time when the taste for this class of entertainment had fallen off, it has always been held to be the most perfect specimen of a masque. [Ballet, Opera, Oratorio.]

Mass. Missa (*Lat.*), **Messa** (*It.*), **Messe** (*Fr.*), **Messe** (*Ger.*).

A name given to the Communion Office, Eucharistic Service, or Liturgy, from the words *Ite missa est* (Depart, the congregation is dismissed), sung by the deacon just before the end of the service. The portions of the Office which constitute the music of the Mass (apart from that chanted by the celebrant and clergy) are the *Kyrie*, *Gloria*, *Credo*, *Sanctus*, *Benedictus*, and *Agnus Dei*. The earliest known settings of these are of that simple, modal, monophonic, essentially ecclesiastical type known as *Plain-song*. Their origin is lost in obscurity; but some portions, such as the *Credo*, are known to have been in existence in the 6th century, and some of the plain-songs of the *Kyrie*, *Sanctus*, *Agnus Dei*, and *Gloria* are probably of a much earlier date. This type of music existed up to the 11th or 12th century. Many of these Masses survive and are still in use. The introduction of polyphonic settings probably took place as soon as the use of Faux Bourdon had taught men the interest and pleasure derived from part-music. Crude at first, consisting only of progression in similar motion, this addition of vocal parts above and below plain-song developed slowly and gradually into distinct harmonisation towards the end of the 13th century, and later on (say at the close of the 14th and beginning of the 15th centuries) the plain-song had become a veritable *canto fermo*, and the harmonisation had arrived at that stage of independent combination of parts known as counterpoint. This once started, it became habitual in the 15th century to drop the plain-song, or even a fragment of plain-song, as a *canto fermo*, and to use in its place any melody equally well known to the congregation, the parts added in counterpoint remaining still in one of the Church Modes.

This class of Mass probably had its origin in the Franco-Belgian school, which afterwards became the well-defined Flemish school, the great Dufay being credited as one of the earliest who dragged profane songs into the sanctuary as *canti fermi*. The names of the popular songs selected for this purpose show how depraved the taste and feelings of musicians had become—"L'homme armé," "Se la face ay pale," "Tant je me deduis," "Las! bel ami," "O Venus la belle," "Adieu nos amours," and others. The custom was so long in vogue that even Palestrina founded the music of one of his Masses on "L'homme armé." But Dufay's contemporaries or immediate successors, Obrecht, Caron, Gaspar, Josquin des Prés, and others, were not always guilty of such bad taste, though revelling in a pedantic use of counterpoint and canon. Still further improvement can be traced in the writers of the early part of the 16th century—Morales (b. *circ.* 1515), Cipriano di Rore (b. 1516), Ruffo (b. *circ.* 1520), Goudimel (b. *circ.* 1510).

We are now approaching that landmark in the history of the music of the Mass, the period of what may well be termed Palestrina's reformation. The composers just named were contemporaries of Palestrina (b. 1524?), but none possessed a tithe of his genius. It is unnecessary here to repeat the well known story of the danger in which polyphonic music stood, owing to the just objection of the clergy to its then highly artificial and often debased style; how Palestrina, carrying with him to Rome the purest vocal writing combined with contrapuntal skill, to which he also added a thoroughly religious sentiment, produced numerous Masses, and Church music of all sorts, and thus not only saved the art from exclusion from sacred edifices, but founded also a School which, in its own style, has never been equalled. Following his example, composers of various other schools, Vittoria, Anerio, Gabrielli, Croce, Orlando di Lasso, and others, produced works full of breadth and dignity. Tye, Byrd, and a few other English composers seem to have preceded Palestrina in their compositions of pure vocal Masses, Tye's Mass, "Euge, bone serve," a work of remarkable beauty and strength, having probably been composed when Palestrina was about ten years of age. But the Reformation then in progress in England and in Germany naturally checked the production of Masses in those two countries. The direct influence of Palestrina lasted elsewhere for about a century. Church compositions then began to be more or less coloured by secular influence, caused no doubt by the collateral growth of opera during the 17th century. It is, however, too much the custom to depreciate the School which succeeded Palestrina, simply on the supposed ground that it represents a period of decadence in art. Yet, Carissimi, Colonna, Benevoli, Durante, and others produced Masses of a very high order, which ought not and cannot be criticised or judged from a standpoint one century earlier. But it cannot be denied that the tendency to forsake the more severe traditions brought about eventually not only an operatic style of Mass, but also the wholesale introduction into the sanctuary of the operatic corps of female and male singers and orchestra. But, nevertheless, Leo, Pergolesi, Jomelli, Haydn, Mozart, Cherubini, Weber, Hummel, working on the lines which the evolution of art lays down peremptorily for every artist (whether he recognises the fact or not), added liberally to the splendid treasure of masterpieces which had been accumulating for so many centuries. The contributions to Mass music by more recent composers are comparatively few, but the names of Rossini, Berlioz, Gounod, Verdi cannot be passed over. Two Masses stand alone in the history of music, the Mass in B minor by Bach and the Mass in D by Beethoven. In neither case could the composer have hoped or expected his work to be commonly performed at Mass, the whole construction was on too gigantic a scale.

The comparison of these two works coming, with the interval of nearly a century between them, from the pens of the two greatest musical geniuses the world has yet seen, is deeply interesting. The contrast between the natural temperament of the two men, their training, their aspirations, their positions in the history of art, shows itself so clearly in these two great works. But during the last half of the 19th century there has been a distinct reaction in the Roman Church against the use of florid Masses. The female singers and opera company are gradually disappearing from the West gallery of Cathedrals, and a return to a more severe style, possibly to the pure, unaccompanied, polyphonic Masses, seems to be inevitable. It is, however, much to be hoped that this justifiable reaction will not be carried to such an extent as to check the production of Masses by modern composers. So consistently have the best writers contributed to the literature of Mass-music, from the days of Dunstable and Dufay to the present day, that the history of the music of the Mass would be practically the history of music itself; and no one can deny that the close association of the life of Art with the life of the Church is a desirable and important thing. But even if the severance of composers from Church-Art is not brought about by the return to the use of music of an older and more severe type, there remains another danger, not less injurious to Art—namely, musicians may be tempted to supply the sanctuary with modern *imitations* of old music. Such a result would be deplorable.

The various movements of the Mass received treatment from great composers in the most solemn, but most advanced style reached at their time. The *Kyrie*, for example, was, like all other portions, sung to ancient plain-song; but as the art of part-writing developed, it became a regular movement, generally in three divisions, following the three divisions *Kyrie eleison*, *Christe eleison*, *Kyrie eleison*. Sometimes it was founded on a plain-song, sometimes an independent and spontaneous movement. But composers of the best period seemed always to favour the use of some external initial idea or motive; hence we find them using the first line or phrase of Hymns, Antiphons, and other portions of the Liturgy—as, *Æterna Christi munera*, *Veni Creator*, *Dixit Dominus*, *Iste Confessor*, &c.; sometimes, as in Hassler's Mass, *Dixit Maria*, the theme is from an existing motet.

When such a theme was chosen, it was

Digitized by Google

usual to make it heard at the commencement of every movement. As this custom gradually died out, composers began to use the resources of modern thematic or fugal development. At a later period, when Masses were supported by orchestral or organ accompaniment, portions were more frequently than formerly allotted to *solo*, or voices *soli*. In the *Gloria* this often took place in the division *Qui tollis*. At the closing words, *Cum sancto spiritu*, &c., pure fugal treatment predominated, and some of the finest choral fugues in existence are settings of these words, or of the final clause of the *Credo*, *Et vitam venturi seculi. Amen.* The *Sanctus* was naturally slow and solemn, the *Osanna* being in strong contrast to it, both as to rhythm and style; the *Benedictus*, which immediately follows, has always called forth the highest powers of composers in the realm of sweet melody and smooth-flowing voice parts, either *soli* or in chorus, or both, in contrast or combination. The *Credo* afforded an opportunity of more spontaneous and individual treatment of the words, according to the composer's temperament. But the clause *Et incarnatus* is generally given to voices, *soli*. At the words "*Et ascendit*" there is hardly an instance of the theme or "point" being anything but an *ascending* passage. There are curious cases of words of the Creed being omitted by composers. In some 16th century Masses, the fact that the same, or nearly the same, clauses are omitted by different composers, leaves the only possible conclusion that this curtailment was not objected to by those in authority; it being remembered that under all circumstances the celebrant would say the whole of the Creed, while the choir sang the music. In Masses of more recent date, omissions of words are occasionally found, but it is most probable that they were the result of the composer's carelessness rather than a recognised method of shortening the choral work. The *Agnus Dei* is generally, as might be expected, a movement of a tender, almost pastoral character, a notable exception however being the opening of this movement in Beethoven's Mass in C, where the initial symphony leads up to a strikingly dramatic entry of the voices.

As choral celebrations of Holy Communion are becoming more and more common in the English Church, we may hope that our young composers will find in setting these solemn words scope for the expression of their highest artistic aspirations.

Not only is encouragement given to living composers to compose English Masses, but greater scope is given to their efforts by the occasional use of a full orchestra as accompaniment, as in St. Paul's Cathedral.

Mass (*Ger.*) Measure, time.

Mässig (*Ger.*) Moderate, temperate *Mässig langsam*, moderately slow.

Massima (*It.*) A semi breve.

Masternote. An old term for the sensible or leading note.

Masure, Masurek, Masurka (*Ger.*) [Mazurka.]

Matelote (*Fr.*) A sailors' dance. [Hornpipe.]

Matinée (*Fr.*) A morning concert.

Mattins. [Horæ Canonicæ.]

Maultrommel (*Ger.*) [Jew's-harp.]

Maxima (*Lat.*) [Nota.]

Mazourk
Mazurek } (*Ger.*) [Mazurka.]

Mazurka. A Polish dance of lively grotesque character, the music of which is in $\frac{3}{8}$ or $\frac{3}{4}$ time, with a peculiar rhythm. It is generally performed by four or eight pairs of dancers, is popular in Germany as well as in Poland, and has been successfully introduced into England. The mazurka, like the waltz, has been treated in a classical manner, notably by Chopin.

M. D. Abbreviation of (1.) *Mano destra* or *main droite*, the right hand; (2.) Doctor of music.

Mean. The name formerly given to the *tenor* part as being the mean in pitch between the bass and treble. The middle strings of instruments were also called *mean*.

Mean clef. The C clef.

Measure. (1.) A general name for a slow and stately dance, supposed to be like the minuet—

> "But after then as men more civile grew
> He did more grave and solemn measures frame."
> *Davies' Orchestra.*

Reed says that "the measures were performed at court and at public entertainments of the societies of law and equity, at their balls on particular occasions. It was formerly not deemed inconsistent with propriety, even for the gravest persons to join in them; and accordingly at the revels, which were celebrated at the Inns of Court, it has not been unusual for the first characters in the law to become performers in treading the measures."

(2.) Time, pace. (3.) Rhythm. (4.) The contents of a bar.

Mechanism. (1.) That part of an instrument which forms the connection between the player and the sound-producing portion. (2) The physical power of performance, as distinguished from the intellect or taste which directs it.

Medesimo movimento (*It.*) The same movement.

Medesimo tempo. The same time.

Mediæval musical instruments. The

Digitized by Google

musical instruments known in the middle ages appear to have been flutes, flageolets, bugles, and trumpets, among simple tubes; waits, cornets, bassoons, bagpipes, &c., among reed instruments; the harp, the guitar, the crowd, the violin or fiddle among stringed instruments; drums and tambourines, cymbals and bells among instruments of percussion, but all of these differ more or less from the instruments with the same names in use at the present day. There is a curious catalogue of musical instruments of the fourteenth century, which may be read with interest in reference to this subject, in a poem by Guillaume de Machault, entitled *Le Temps Pastour*. The description is headed *Comment l'amant fut au disner de sa dame*:—

> Mais qui veist, apres mangier
> Vener menestraux, sans dangier
> Pignez et miz en pure corps.
> Là feurent meints divers acors
> Car je vis là, tout en un cerne
> Viole, rubebe, guiterne
> L'enmovache, le mica mon
> Citole, et le psalterion
> Harpes, tabours, trompes, nacaires
> Orgues, cornes, plus de dix paires
> Cornemuse, flajos et chevrettes,
> Douceines, symboles, clochettes
> Tymberes, la flauste brehaingue
> Et le grand cornet d'Allemaingue
> Flajos de sans, fistule, pipe
> Muse d'Aussay, trompe petite,
> Buisines, èles, monocorde
> Ou il n'a quime seule corde
> Et muse de blet, tout ensemble
> Et certainement, il me semble
> Qu' oncques mais tele melodie
> Ne fut oncques veues ne oye
> Car chascun d'eux, selon l'acort
> De son instrument, sans descort
> Viole, guiterne, citole
> Harpe, trompe, corne, flajole,
> Pipe, souffle, muse, naquaire,
> Tabour, et quanque on put faire
> De dois, de pennes, et d'archet
> Ois et vis en ce parchet.

If each name here mentioned represented a different and distinct instrument, it might be supposed that there was a far greater variety in use in the fourteenth century than in the present time, with all our increase of knowledge and improved skill in construction, but the truth is that many of the names are simply synonyms. For example the line:—

> "Cornemuse, flajos et chevrettes"

would seem to imply three sorts of music making machines, but the names refer to one and the same thing, neither more nor less than the bagpipes, and the "muse d'Aussay," "muse de blet," and "muse" simply, are varieties of the same instrument. The "viole and rubebe," refer to the fiddle, the "flajole pipe, souffle, and fistule," to the flageolets; "la flauste brehaingue" to the English flute; "le grand cornet d'Allemaingue, buisines, douceines," were a species of bassoon; "tymberes, tabour, and naquaires," were drums; the "citole and psalterion," were dulcimers; "the guiterne" a guitar or cittern; "èles," signified harps, from their wing-like shapes; "L'enmovache," was also a harp of a larger kind; the "mica mon" is probably "maca man," a hand-drum or tambourine, and the rest are sufficiently obvious. Guillaume de Machault, the author of this poem, was probably not a practical musician, but simply made use of a list of instruments with whose use he was probably unacquainted, hence his repetition of synonyms. It is no uncommon thing to find a special ignorance of the use and qualities of musical instruments among modern poets, but when such a lack of knowledge is displayed by a poet writing in the palmy days of minstrelsy, it is not unfair to imagine that Guillaume de Machault was more of a singer than a savant. [Troubadour.]

Mediant. **Médiante** (*Fr.*) The third degree of any scale.

Mediation. That part of a chant which lies between the reciting-note and the next close. [Chant].

Medius (*Lat.*) The tenor part. [Mean.]

Mehrstimmig (*Ger.*) In many parts. Polyphonic.

Meisterfuge. (*Ger.*) A *fuga ricercata*, that is, a learned, elaborate fugue; also one in which the subject or answer is constantly heard.

Meistersinger. **Meistersänger** (*Ger.*) *Maître Chanteur* (*Fr.*) A title given to the most renowned musician of a township or district in Germany, during the middle ages.

When the Minnesingers or Troubadours, formed from among the ranks of the aristocracy, began to die away, the Meistersingers arose, and at the beginning of the fourteenth century formed themselves in the towns of Germany into guilds or trading companies, and agreed to uphold certain fanciful and arbitrary laws of rhythm.

Adepts in the poetic art, Master-singers, appear to have existed at the period when the Minnesingers were at their prime, and it is supposed that the title was only more universally and distinctively bestowed when the spirit of pedantry had become general and when the "Songschools" or "guilds" became in Germany what the "consistories" and "academies" were in other countries during the then general decay of art. Wandering orators in connection with the "Chambers of Rhetoric" who recited a miserable doggerel composed by themselves or others, afterwards formed themselves in Flanders and Brabant, into societies by the name of Chambers of Rhetoricians (Kamers der Rhetorÿkers or

Rederykers) and offered prizes for the most meritorious poems written according to their barbarous rules.

The ornaments and illustrations of this class of poetry were drawn from mean sources. Instead of princes, nobles, and knights, we have clerks, schoolmasters, and mechanics. Some of the poems, it is true, are not unworthy of notice and survive in one form or another to the present day; others have sunk into an oblivion not altogether undeserved.

The "Masters" were always anxious to clothe themselves with the ancient glory of the "Minnesingers." They were fond of tracing the origin of their school to a very remote antiquity; and the most celebrated names were placed by all sorts of anachronisms among the supposed united band of founders.

Wherever the "hoch-deutsch" was spoken, there the Master-singers founded a colony and indulged the vulgar propensity of giving importance to bombast, parade, and external distinction to their hearts' content. A candidate was admitted into the school of the "Master-singers" with great form and ceremony. Four "Merker" sat behind a silken curtain to pass judgment upon his qualifications. One of these had to decide whether the diction of the novice was pure and his grammar accurate, the others attended to the rhyme and metre of the composition and the melody to which it was sung. If they united in declaring that the candidate had complied with the statutes and regulations, he was decorated with a silver chain and badge—the latter representing King David playing on the harp—and he was honourably admitted into the Society. The poems of the "Master-singers" were always lyrical and sung to music. The entire poem was called a "Bar" and it was divided into five or more stanzas, or "Gesätzen," and each "Gesatz" also fell into three portions, the first of which was a "Stolle," the second an "Abgesang," and the third a "Stolle" like the first. The rhymes were classed into "stumpfe Reime," klingende Reime," "stumpfe Schlagreime," and "klingende Schlagreime," and other denominations needless to particularise. The poets, singers, and Merkers counted the syllables on their fingers, and if there was the proper number it was of no consequence whether they were long or short. The length of the verse, the number of the lines, and the order of the rhymes in each "Stolle" or "Abgesang" were variable and consequently the poems were susceptible of a great variety of forms, which were called tunes or "Weisen." The invention of a new "Weise" was considered the test of a Master-singer's abilities. There were some hundreds of these "Weisen," all named after their inventors; as Hans Tindeisen's rosemary Weise; Joseph Schmeirer's flowery-paradise Weise; Hans Fogel's fresh Weise; Henry Frauenlob's yellow Weise; his blue Weise; his frog Weise, and his looking-glass Weise. The code of criticism to which the Master-singers were subjected was contained in the rules or "Tabulatur" of the societies, some of which rules were certainly severe. They were prohibited from employing "sentences which nobody could understand," or "words wherein no meaning could be discovered," interdictions which would completely disqualify many a so-called poet in the present day for membership.

Mélange *(Fr.)* A medley.

Melisma *(Gk.)* *μέλισμα.* (1) A song, tune, melody. In 1611, Ravenscroft published "Melismata; Musical Phansies fitting the Court, citie, and country humours." (2) A grace, roulade. *Fioritura.*

Melismatik. (*Ger.*) The art of florid vocalisation.

Melodia *(It.)* Melody.

Melodia *(Gk.)* *μελῳδία.* (1) The singing of an ode to a *melos.* (2) The tune to which lyric poetry was sung.

Melodicon. An instrument made of steel bars in different lengths, tuned to the diatonic scale, struck with hammers held in the hand.

Melodico, melodicoso (*It.*) Melodiously, sweetly.

Mélodie (*Fr.*) Melody, tune.

Melodrama. A dramatic piece in which the interest is heightened by the character of the vocal or instrumental music accompanying certain situations. The melodrama is of French invention, and was introduced into England at the end of the last century; the subjects are generally of a romantic character, illustrated with picturesque costumes and scenery, and having serious and sensational incidents, and effective and striking tableaux. Although sometimes confounded with the opera, it differs from that higher class of work insomuch that the action is carried on in speaking and not in recitative and aria. [Opera.]

Melodrame (*Fr.*) **Melodramma** (*It.*) [Melodrama.]

Melody (*melos Gk. μέλος*). An agreeable succession of simple sounds, produced by a single voice or instrument, and so regulated as to give a pleasing effect, or to be expressive of some kind of sentiment. It is often founded on relative harmonies and yet is completely distinguished from harmony by not needing the addition of parts to make it perfect. The construction of melody is guided by certain rules in rhythm, and by the capabilities of the voice or instrument for which it is intended. Melos had a very general meaning among the Greeks, for it is said by

Digitized by Google

Plato to consist of "speech, music, and rhythm." In its more limited sense it signified any succession of musical sounds.

Melody Organ *or* **Harmonium.** A harmonium so constructed that the upper note of the chords played is louder than the rest of the sounds.

Melograph. An instrument invented for the purpose of writing down melodies when played upon a pianoforte. It has not yet been brought into use, as its action is imperfect.

Melopœïa (*Gk.*) *μελοποιΐα.* (1.) Music in general. (2.) The art or system of making a tune or *melos.* In this sense it is said, by Aristides Quintilianus, (Ed. Meibom. p. 28) to be of three kinds with reference to the pitch, namely, hypatoeides, mesoeides, and netoeides. He then divides Melopœïa into three parts, Sumptio (λῆψις), Mistio (μίξις), Usus (χρῆσις).

Sumptio, or the "taking," is the settlement of the pitch from which the "system" or scale is to be reckoned. Mistio, the mixing, is the co-ordinating or fitting together the sounds with regard to each other, or the modulations of the voice, or the key-systems. Usus, the use, is of three kinds, Ductus (ἀγωγή). Petteia (πεττεία), Nexus (πλοκή). Ductus is of three kinds, direct, reverse, circumcurrent (see Agoge). Petteia is the choosing of sounds that are suitable, and rejecting those that are unsuitable. Nexus is the interweaving of intervals or movements by skips. Quintilianus sums up by giving the following plain statement of the different kinds of melopœïa.

Melopœïa is of different kinds as to
- Genus .... Enharmonic. Chromatic. Diatonic.
- System ...... Hypatoid. Mesoid. Netoid.
- Scale (tonos) Dorian. Phrygian. Lydian.
- Mode (tropos) Gnomic. Dithyrambic. Tragic.
- Manner or character (ethos)
  - Systaltic. (Producing sadness.)
  - Middle. (Producing calmness.)
  - Diastaltic. (Producing excitement.)

Melopiano. An invention by which sustained sounds can be produced on a pianoforte. It consists of a series of small hammers set into very rapid vibration by the winding up of a spring. When a note is struck and held down, the constant repetition of the blows of the hammer causes a continuous vibration of the string which is of a most charming character. A beautiful crescendo is obtained by the ingenious device of raising the hammers gradually further from the string, causing them to descend with more force. It was invented by Caldera of Turin in 1870.

Melos (*Gk.*) (μέλος). A succession of musical sounds as opposed to *noises.* A tune. A song. [Melody.]

Même mouvement (*Fr.*) The same movement. *L'istesso tempo* (*It.*)

Men, meno (*It.*) Less; as *meno forte,* not so loud; *meno piano,* not so soft; *meno presto,* less rapid; *meno vivo,* not so quick.

Menaaneim (*Heb*). This word occurs once in Holy Scriptures, in 2 Sam. vi. 5, where it is improperly translated *cymbals.* Its derivation points to a root meaning to swing to and fro, to vibrate. It is probable, therefore, that it was a *sistrum.* [Sistrum.]

Ménestrels (*Fr.*) Minstrels. [Troubadours.] [Confrérie de St. Julien.]

Meno (*It.*) [Men].

Menschenstimme (*Ger.*) (1.) A man's voice. (2.) Vox humana.

Mensur (*Ger.*) (1.) The measurement of intervals on stringed instruments. (2.) The scale of organ pipes.

Mensurable Music. [Cantus mensurabilis.]

Mensuralgesang (*Ger.*) [Cantus mensurabilis.]

Mensuralnotenschrift (*Ger.*) Notation of time measurement.

Menuet (*Fr.*)
Minuetto (*It.*) } [Minuet.]

Mescolanza (*It.*) Cacophony.

Mese (*Gk.*) [Greek Music.]

Messa (*It.*) **Messe** (*Fr.*) **Messe** (*Ger.*) Mass.

Messa di Voce (*It.*) The swelling and diminishing of the sound of the voice upon a holding note.

Mesochorus (μεσόχορος.) [Coryphæus.]

Mesopycni (*Mediæval Lat., from Gk.* μέσος, *middle, and* πυκνὸς *close.*) Church modes having the close interval or semitone in the middle of their primary tetrachord, *e.g.*—

Dorian.

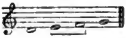

Hypodorian.

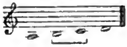

cf. Barypycni. Oxypycni.

Mesto, mestoso (*It.*) Sad, pensive.

Mesure (*Fr.*) Measure, time; as, *mesure à deux temps* or *trois temps,* duple or triple time.

Metallo (*It.*) *lit. metal.* The *ringing* quality of tone, as *bel metallo di voce.*

Metal pipes. [Organ.]

Metal plate of pianoforte. [Pianoforte.]

Metal wind instruments. Instruments formed of various lengths of brass tubing gradually enlarged towards one end called the

Digitized by Google

"bell" and with a mouthpiece at the other end. The tubing is folded into a size convenient for carrying about, either into an oblong or circular form. A single length of tubing is capable of sounding the primary harmonics of its tonic with that note, in the following order, according to the power of lip possessed by the player, the key note being taken as C:—

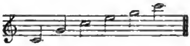

A double length of tubing enables the player to produce another octave with some additional harmonics, thus:—

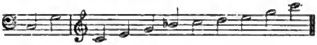

and sometimes

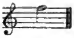

as in the French Horn and Trumpet. These are called the open notes, the intermediate tones—

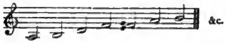

are made by inserting the hand in the "bell" of the Horn as well as the Trumpet, these are called closed notes and are less full and beautiful in character than the open tones, and being also difficult of production, are rarely written.

Upon the application of the slide to the trumpet, an application probably suggested by the trombone, the uncertainties of intonation were made somewhat less. The first instrument of the tube kind in which an improvement was sought to be effected was the Bugle, which was formerly without keys, like the horn and the trumpet. The bandmaster of an Irish regiment, named James Halliday, about the year 1814 or 1815, introduced a keyed bugle of his own contrivance which he called the Kent Bugle out of compliment to the Duke of Kent, the father of Her Majesty Queen Victoria, who was colonel of the regiment. The instrument became exceedingly popular, in consequence of the excellent performance of the elder Distin, who introduced it into the orchestra of Her Majesty's Theatre in 1830; and the French composer, Adolphe Adam, wrote several solos for it.

A bugle player named Macfarlane claims to have been the first who applied pistons or valves to the Cornopean. He called his contrivance "Russian valves," and he furnished his model with two of them, the first for the whole tone and the second for the half tone. As about that time (the year 1834) the celebrated player, Koenig, made his appearance from Paris with an instrument of similar make which he declared was by a French inventor, there was some controversy at the time, and as no one suspected the honesty of Macfarlane, it is possible that the same idea may have occurred simultaneously to the two claimants. The application of the third valve to the cornet or cornopean was claimed by a horn player named Kilback, but, on what grounds, it is difficult now to determine. Valves or pistons were employed in Germany before the introduction of the Russian valve (so-called) into England, these were the double cylinder and the rotatory valve, and the instruments played by Kalozdy's Hungarian band were on this latter principle.

The principle of the *Sax-valve* was a German invention, and was introduced to the public by a man named Stölzel. The *tuning slides* for the piston notes, and the improvement of Stölzel's valve are due to the maker whose name is connected with the invention. Since the time that the Sax-horns were introduced into England by the Distin family, many modifications and improvements have been effected, the object of each alteration being to obtain the easiest and most accurate means of lengthening and shortening the tube at will, so as to produce chromatic intonation.

The metal valved-instruments now in use may be thus classified:—

Soprano Compass.—Trumpet in F and E flat, both with slides and valves; soprano cornets in E flat and C, with 3 valves or pistons; soprano saxhorn in E flat, with 3 valves or pistons; soprano flügel horns in E flat and C, with 3 and 4 cylinder valves; bugle in C and B flat, with keys (rarely used now).

Alto or Contralto Compass.—Cornet in B flat, with crooks for A, A flat, and G; alto saxhorn in B flat, with 3 valves; alto flügel horn in B flat, with 3 and 4 cylinder valves; alto trombone in E flat, both slide and valved.

Tenor Compass.—French horn, with and without valves, with 10 crooks; tenor saxhorn, with 3 valves; tenor flügel-horn, with 3 or 4 cylinder valves; tenor trombone, both slide and valved.

Bariton Compass.—Bariton sax-horn, with 3 valves; bariton flügel-horn, with 3 and 4 cylinder valves.

Bass Compass.—Trombone in F or G, both slide and valved; euphonium in C or B flat with 3 or 4 valves; trombone in E flat.

Contra-basso.—Bombardon in F or E flat, with 3 and 4 valves; bombardon, circular shaped; contra-bombardon in BB flat, with 3 pistons or valves; contra-bombardon, circular shaped; ophicleide.

Most metal wind instruments are formed from a sheet of brass cut into the proper size. The sheet of metal is turned in the shape of a pipe until the edges meet correctly. The edges are welded together in a brazing forge and the tube is "pickled" in a chemical bath to remove the traces of brazing. It is then made perfectly cylindrical by means of a "draw bench," and afterwards is bent into the required curve, the tube being filled with molten pitch, for slight curves, or with lead for "U" curves. As both pitch and lead will melt at a lower temperature than brass, there is no difficulty in removing the interior "filling" by heat, leaving the tube in the required form without breaking or flattening it. The "bell" is made by careful hammering on a "mandril" of the required size and form, and is fitted to the tube by means of solder and the blow pipe. Shanks, crooks and the other parts are then put together; the instrument is tested and polished, and is ready for use.

Méthode (*Fr.*), *Metodo* (*It.*) Method. (1) System of teaching; as, *Kalkbrenner's method*, *Crivelli's method*, &c. (2) School or style of music; as, the *Italian method*.

Metre. A term used with various significations; (1.) A foot, as a subdivision of a bar or measure; (2.) The relation between two feet having the same subdivisions of time-units, but in a different order of succession; (3.) The proper grouping of a number of consecutive feet.

Authors who use the term in this last sense, consider it as equivalent to rhythm and divide it thus:

Metre (rhythm).
|
Measures (bars with accents).
|
Feet (groups of time-units).
|
Units of time (short and long).

But it will be seen further on that this division is not good; and also that the incorporation into music of the terms of prosody is not desirable. Before entering on this discussion, it is necessary to give a list of the names and nature of the most important feet.

Dissyllables have two units of time capable of a four-fold arrangement; *e.g.*,

Ex. 1.

– – Spondee.

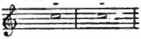

Ex. 2.

⏑ ⏑ Pyrrhic.

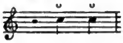

Ex. 3.

Mozart. Symp. No. 1.

– ⏑ Trochee.

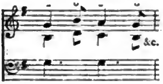

Ex. 4. Beethoven. Symp. No. 1.

(or)

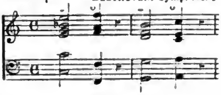

Ex. 5.

⏑ – Iambic.

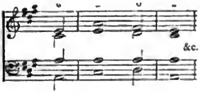

Ex. 6. Schumann

(or)

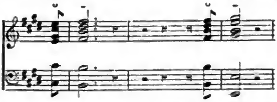

Trisyllables or feet consisting of three time-units are of eight kinds.

⏑ ⏑ ⏑ Tribrach.

Ex. 7

Beethoven.
Symphony No. 2.

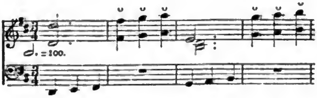

⏑ – – Bacchic.

Ex. 8.

Schumann.
"Paradise and the Peri."

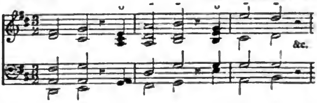

⏑ – ⏑ Amphibrach.

Ex. 9.

Beethoven.
Symph. No. 6.

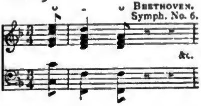

Ex. 10.

⏑ ⏑ – Anapæst.

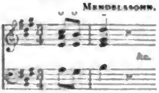

Mendelssohn.

Ex. 11.

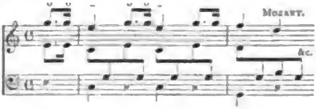

Mozart.

– – Molossus.

Ex. 12.

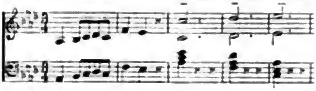

Weber. Mass in E♭.

– ⏑ ⏑ Dactyl.

Ex. 13.

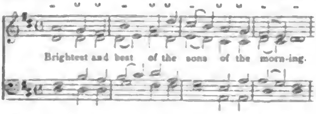

F. C.

Ex. 14.

– ⏑ – Cretic.

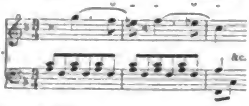

Beethoven. Symph. No. 3.

– – ⏑ Palimbacchic.

Ex. 15.

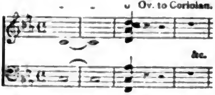

Beethoven. Ov. to Coriolan.

There are also tetrasyllables, sixteen in number, which, of course, consist of the possible positions of four syllables of different lengths.

⏑ ⏑ ⏑ ⏑	Proceleusmatic (Double Pyrrhic).
⏑ ⏑ – ⏑	Third Pæon (Pyrrhic & Trochee).
⏑ ⏑ ⏑ –	Fourth Pæon (Pyrrhic & Iambic).
⏑ ⏑ – –	Ionic a minore (Pyrrhic & Spondee).
⏑ – ⏑ –	Diambic (Double Iambic).
⏑ – ⏑ ⏑	Second Pæon (Iambic & Pyrrhic).
⏑ – – ⏑	Antispast (Iambic & Trochee).
⏑ – – –	First Epitrite (Iambic & Spondee).
– – – –	Dispondee (Double Spondee).
– – ⏑ ⏑	Ionic a majore (Spondee & Pyrrhic).
– – ⏑ –	Third Epitrite (Spondee & Iambic).
– – – ⏑	Fourth Epitrite (Spondee & Trochee).
– ⏑ – ⏑	Dichoree (Double Trochee).
– ⏑ ⏑ –	Choriambic (Trochee & Iambus).
– ⏑ ⏑ ⏑	First Pæon (Trochee & Pyrrhic).
– ⏑ – –	Second Epitrite (Trochee & Spondee).

Feet of more than four "times" or syllables, are, strictly speaking, merely compounded of dissyllables and trisyllables. Musical examples of all the above could be easily given if space allowed.

On examination of the musical dissyllables and trisyllables (Exs. 1 to 15) many important questions present themselves.

1. It will be noticed that long syllables almost invariably fall on the accented part of a bar. Can we conceive of a long syllable in music as *quantity* without *stress?* This question has been discussed over and over again with reference to modern languages. Some authorities say that stress and non-stress have ejected *quantity* from our poetry, and bring forward such lines as the following in proof of their assertion:

> "Th' infernal Serpent; he it was whose guile."

in which four out of the five syllables, short as regards *stress*, are by nature *long* in quantity.

Let us apply this to music, and ask whether we really reckon our feet by stress or quantity. If we reckon them by the latter then the following is correctly written:

Ex. 16.

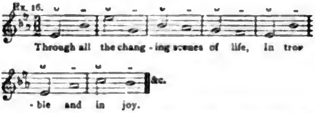
Through all the chang-ing scenes of life, In trou-ble and in joy.
&c.

Now this is *not* written correctly *musically*, because, although an iambus in poetry has only the relation of *short to long* in quantity, it has in music the relation *weak to strong* in emphasis. We therefore write it thus:

Ex. 17.

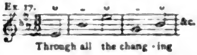

practically disposing of the first short syllable, and then presenting the succession of feet to the ear as *trochees;* for no one would venture to analyse this tune as consisting of the following groups:

Ex. 18.

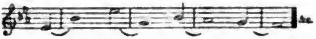

an analysis which is not only allowable but *necessary* if the analogy between poetry and music is to be sustained.

Again, in the following, the same *length* is given to the short syllables of the words as to the long:

Ex. 19.

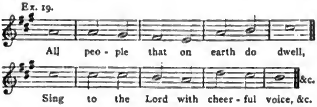
All peo - ple that on earth do dwell,
Sing to the Lord with cheer - ful voice, &c.

One author, enthusiastic on the subject of quantity, actually proposes that the Old Hundredth Psalm should be thus sung:

Ex. 20.

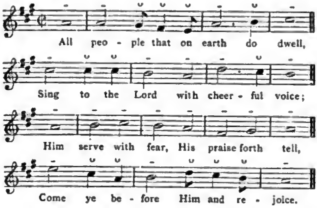
All peo - ple that on earth do dwell,
Sing to the Lord with cheer - ful voice;
Him serve with fear, His praise forth tell,
Come ye be - fore Him and re - joice.

If musical feet were governed by quantity, and not by stress, the above would not only be correct but pleasant, whereas, we feel, that just in proportion as the words receive their due quantity, the true rhythm of the music is lost.

On the ground, then, that musical metre is governed more by stress than quantity, the analogy between ancient poetry and modern music cannot be maintained.

2. If the reader will refer to Ex. 4, which gives a musical trochee, he will observe that the short syllable is followed by a rest. This example was given, not because it was felt to be correct, but because similar examples are to be found elsewhere; but, as a matter of fact, *rests* in music form *part of rhythm*, and therefore, if two minims in a bar (being equal to each other in length) are considered as two long syllables, Ex. 4 is not a trochee but a spondee; and, what is more important still, iambics and trochaics can only be represented by *triple* time. Not only, then, is Ex. 19 incorrect, but *all* duple-measure psalm tunes set to hymns ordinarily spoken of as iambics are bad. On the ground, therefore, that when there are two notes of equal length in a bar we not only *may*, but *do* often consider the accented note long and the non-accented note short, we are again justified in saying that the analogy between poetical metre and musical rhythm does not hold good. With regard to the inclusion of silence or "rest" as a constituent part of rhythm, music diverges widely from poetry, and stands on a vantage ground peculiarly its own. It is true that *cæsura* in verse bears some similitude to a "rest" in music, in the effect it produces, but the likeness is only apparent, and the musician has the power of suggesting *length* in quantity, by a *short* note if it is a *stress-note*, *e.g.*:

BEETHOVEN.
Ex. 21. Ov. Men of Prometheus.

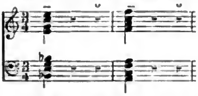

In the above example, each of the two bars is a trochee, the long syllable of which is represented by the chord and first rest, the short syllable by the final crotchet rest in each bar.

3. It will be seen that in Ex. 13, bars 3 and 4, two consecutive notes in each bar are grouped under one long syllable. This compression of two syllables into one foot, is a common occurrence both in ancient and modern poetry, as for example in the junction of consecutive vowels at the end of one word and commencement of another. In English poetry we have *mā ny a*, *glō rious*, &c., which exactly correspond to the musical junction of two notes under one metrical time-unit. This leads to the important question of the exact time relation of *short* and *long* in metrical feet. The general answer is that a long syllable occupies the time of two short syllables (—"Syllaba brevis unius est temporis, longa vero duorum"). But ancient writers relax this rule and say that all long syllables are not the same length, nor do they stand in the same relation to the short. Some allow that a short syllable may be a third of a long, or even less. How does this bear upon music? If we are to adopt the old system and names of metrical feet, are we only allowed to divide one long time-unit into two, three, four, or six parts? Our time table teaches that a semibreve is equal to thirty-two demisemiquavers, and this may be again multiplied by binding two stationary semibreves together into one "double" time-unit, in which case demisemiquavers will stand to it in the quantitative relation of sixty-four to one. The classical names of feet are therefore quite inadequate to represent the immense number of possible relations between the syllables of musical feet, and if we are to have musical metre on the model of classical metre, an entirely new system of nomenclature will be absolutely necessary.

After all, such names could only apply to a single melody, or a simple harmonised tune, whereas, we know that one of the highest forms of modern music is harmony which consists of a combination of many melodies.

4. Lastly, is it of any practical use to adopt names and symbols of metre, if metre and musical rhythm are contradictory? That they may be so is shown by a comparison of Exs. 16 and 17, the former of which is perfectly correct metrically, perfectly ridiculous musically. Scholars explain this by saying that whereas metre depends on quantity alone, rhythm depends upon a combination of quantity and emphasis. Be it so, but in modern music, quantity without emphasis is inconceivable, therefore any nomenclature founded on quantity only is of no practical value.

It may be thought strange that so much space has been given to this subject of metre, solely for the purpose of depreciating its value and discouraging its use. But the classical names of feet are largely used as musical terms by certain writers, more perhaps from a desire to appear learned at a little cost, than from a knowledge of their real bearing on the art.

Metrometer. [Metronome.]

Metronome. An instrument said to have been invented in 1815 by Maelzel, for the purpose of measuring the relative duration of the notes in a piece of music. The machinery is of clockwork, and the various grades of time are measured on a balance-rod serving the purpose of a pendulum, the speed being regulated by a shifting or sliding weight. To be correct the metronome should beat seconds when set at 60.

Maelzel's claim to the invention rests on very doubtful authority, that is to say, only upon his own word; for in the "Journal für Deutschland" 1796 is a description of a time measurer in the form of a chronometer, J. S. Stoeckel, of Burg, claiming its invention; but this being of large and inconvenient size, Maelzel suggested to Winkel the desirability of a new contrivance, which Winkel succeeded in producing. Maelzel exhibited the instrument as his own invention, and his name is always associated with it, though all the part he really had in the matter was in dividing the balance-rod to agree with the varieties of musical *tempi*. The claim to this portion of the invention was allowed before the Dutch commission appointed to enquire into the subject at the request of Winkel. But neither of these has the right to be considered as the inventors of the machine, although Stoeckel and Winkel acted in good faith, and in ignorance probably of what had been done before them. A pendulum for measuring time, independent of clock-work machinery, was common in England before the metronome was introduced. This was a tape like that used by surveyors, and as the length was in actual inches, it was necessary to have so great a length that its general use was inconvenient. Many of the glees of Callcott, Horsley, Webbe, and some of the compositions of Crotch and other writers of the latter part of the last century have the time indicated by such a pendulum. Twelve years before the date of Stoeckel's invention, namely, in 1784, an amateur violinist and composer of Paris, Jean Baptiste D'Avaux, wrote to the "Journal Encyclopédique," describing a newly-invented pendulum for measuring time in music:—"Lettre sur un instrument ou pendule nouveau qui a pour but de déterminer avec la plus grande exactitude les differentes degrés de vitesse, depuis le *prestissimo* jusqu'au *largo*, avec les nuances imperceptibles d'un degré à l'autre." This could hardly have been the invention of three-quarters of a century earlier, namely, in 1710, alluded to in the sixth edition of a work by Michel D'Affilard, published under the title of "Principes très-faciles, pour bien apprendre la musique, qui conduiront promptement ceux qui ont du naturel pour le chant jusqu'au point de chanter toute sorte de musique et à livre ouvert," in which the time of the exercises or airs is regulated by a "pendule." Taking the date into consideration, this instrument was probably the metronome or sonometer of Etienne Loulié, invented in 1696, described as a "chronomètre ou instrument de nouvelle invention, par le moyen duquel les compositeurs de musique pourront desormais marquer le veritable mouvement de leurs compositions, et leurs ouvrages, marquez par rapport à cet instrument, se pourront, executer en leur absence s'ils en battaient euxmêmes la mesure." If this had been simply a cord or tape it would have indicated the time only while the motion lasted; but the contrivance was of clock-work, as "l'estampe représentant le chronomètre" shows. In 1759 Henri Louis Choquel also advocated the use of a chronometer, probably that of Loulié; and in 1807 Jeanne Etienne Despréaux published a chart or table of the different measures in music, under the title of "Chronomètre musical établi sur les bases du pendule astronomique," the machine employed being a modification of Loulié's invention. All that Maelzel did was to suggest the construction of an instrument more portable than any already in use, but the principle he worked upon was that which had been carried out nearly 100 years before he was born.

Mette (*Ger.*) Matins.

Metzilloth, Metzillthaim, Tzeitzelim (*Heb.*) These words are rendered in our version as *cymbals*, except in Zech. xiv. 20, where they are translated "*bells* of the horses," which is substantially correct, as little cymbals were formerly used in the trappings of horses. The cymbals used by the Hebrews were probably similar in form to those of the ancient Egyptians, some existing examples of which are here depicted. They were found in the tomb of a certain musician-priest named Ankhapê, lying close to his side. They are very small compared to modern instruments of the same class, being not more than five inches in diameter. [Cymbals.]

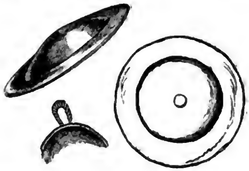

Mezzo, Mezza (*It.*) Half or medium, as, *mezza bravura*, semi bravura style; *mezza manica*, half shift; *mezza orchestra*, with half the orchestra; *mezza voce*, half the power of the voice; *mezzo carattere*, with a moderate degree of expression or execution; *mezzo forte*, moderately loud; *mezzo piano*, moderately soft; *mezzo soprano*, a voice lower in range than a soprano and higher than a contralto; *mezzo staccato*, slightly detached; *mezzo tenore*, a voice of tenor quality and baritone range; *mezzo tuono*, half the power of the voice; &c.

M. F. [Mezzo forte.]

M. G. Abbreviation of *main gauche* (*Fr.*), the left hand.

Mi. (1) A syllable used to indicate the third note in the major scale [Aretinian syllables]. (2) In old English solmisation Mi always indicated the leading note. [Sol-faing.]

Mi bémol (*Fr.*) The note E flat.

Mi bémol majeur or **mineur.** The key of E flat major or minor.

Mi contra fa. The name given by the old contrapuntists to the tritone, which was always to be avoided—"mi contra fa est diabolus." It is not as some suppose, the simultaneous sounding of the tonic with the leading note; but, as the notes of the scale of C in old solfaing were named—fa, sol, la, fa sol, la, mi, fa—the union or succession of the fa and mi would form the tritone, the leading note in the old scales always being called *mi*. [Sol-faing.]

Middle C. The note standing on the first leger line above the bass stave, and the first leger line below the treble stave. The note *c'* [Stave.]

Mi dièse (*Fr.*) The note E sharp.

Militairement (*Fr.*), **Militarmente** (*It.*) In a military style.

Military Band. [Band, § 5.]

Minaccevole
Minaccevolmente
Minacciando
Minacciosamente
Minaccioso
(*It.*) In a menacing, threatening manner.

Mineur (*Fr.*) Minor.

Minikin. A small sort of gut string formerly used on the lute, viol, and other stringed instruments.

Minim. *Halbnote* (*Ger.*), *Bianca* (*It.*), *Blanche* (*Fr.*) A time character of the value of two crotchets. In modern music it is second in value to the semibreve now held to be the time standard, but in ancient music it was, as its name implies, of the shortest duration. Morley, in his "Introduction to Practical Musicke," 1601, ascribes the first use, if not the invention, of the minim to Philippo de Vitriaco, a musician of the 14th century, who is also credited with the invention of the crotchet. [Nomenclature.] [Notation.]

Minnesingers. The German name for poets of the Troubadour character, who devoted their talents to the production of love songs (Minnelieder). They enjoyed a certain amount of popularity in the higher grades of society for more than two hundred years (1090-1290), when they fell out of popular estimation, and were succeeded by the Meistersingers, *q.v.* [Troubadours.]

Minnim (*Heb.*) This word, which occurs in Ps. xlv. 8, and Ps. cl. 4, is probably a poetical allusion to stringed instruments generally. It is so rendered in the Bible and Prayer-book versions of the latter psalm. "Praise him with *stringed instruments* and organs." (*Bible.*) "Praise him upon the *strings* and pipe" (*Prayer-book*). In order to bring out its meaning in Ps. xlv. 8, it has been proposed to alter the text thus: "out of the ivory palaces, *stringed instruments* have made thee glad."

Minor (*Lat.*) Less, smaller. (1) Intervals are said to be minor when they contain one semitone less than major. (2) A scale is said to be in the minor mode when its third and sixth are minor. Formerly minor music was described as "with the lesser third." (3) Flute minor, Kleinflöte, a small flute stop on the organ, of 4ft. or 2ft. pitch.

Minstrel. [Troubadour.]

Minuet (*Fr.*) The name of a graceful

Digitized by Google

dance said to have been invented in Poitou about the middle of the 17th century. The tunes for the first minuets are said to have been composed by Jean Baptiste Lully. A minuet was danced by Louis XIV. before his courtiers at Versailles. The Marquis de Flamarens introduced it into England in the reign of Charles II., who made the dance fashionable by having it frequently danced at Court. It continued to be popular as a court dance until the reign of George III.

The minuet form was used by many composers after Lully as a movement in sonatas, overtures, and the like, but composers made a few slight variations in the time and treatment in order to obtain as much originality as possible. The minuet with more or less success continued to be employed as a movement in instrumental music to the time of Beethoven, who abandoned it in favour of the Scherzo.

Independently of its use for dancing purposes the Minuet, almost immediately after its introduction, was adopted as part of a Suite, sometimes in the simple form as required for the dance and sometimes with a second minuet of a slightly different character and key appended. This second minuet was so arranged as to demand a return to the original key. The title of Trio for the second minuet is said to have been first suggested from the circumstance that only three parts or instruments were employed in the performance of that section; it is also stated that J. S. Bach is entitled to the credit of being the first so to call the second minuet. But the subjoined Round O (Rondo) minuet by Mr. Jerry Clark, c. 1700, shows an earlier division of the composition into three parts or strains, and the form itself suggests that the third part which takes the place of the trio in more modern minuets, is more likely to be called trio because of its third place than because it was written in three parts or for three instruments.

MINUET.

LULLY.

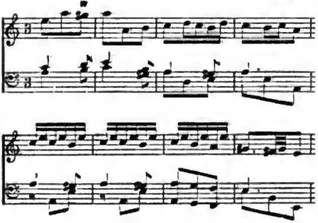

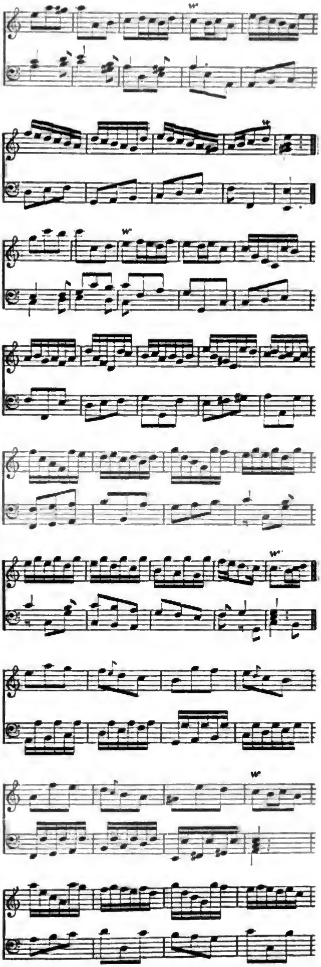

Digitized by Google

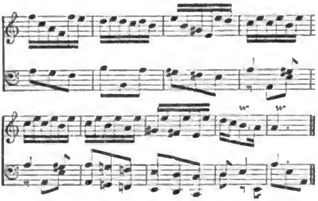

MINUET.

H. PURCELL. Choice Lessons.

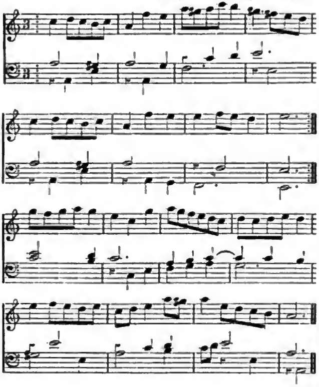

ROUND O MINUET.

1st Straine. By Mr. JERRY CLARK.

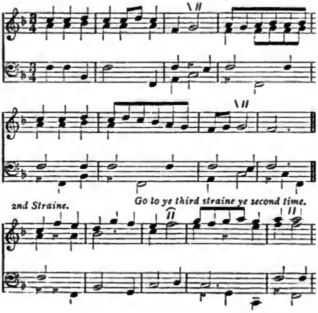

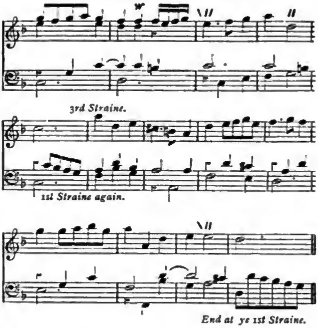

Minuetto (*It.*) [Minuet.]

Miscella. A mixture stop of an organ.

Miserere. The 51st Psalm sung in the *Tenebræ* service in the Roman Catholic Church.

Mishrokitha, *or* **Masrakitha** (*Chaldaic*). This word occurs four times in the book of Daniel, and is probably rightly interpreted in the Septuagint by *Syrinx.*

Missa (*Lat.*) A mass; as, *missa pro defunctis*, a requiem mass; *missa solemnis*, a solemn mass, &c. [Mass.]

Misshällig (*Ger.*) Dissonant, discordant.

Missklang (*Ger.*) Discord, cacophony.

Misteriosamente, Misterioso (*It.*) Mysteriously.

Misura (*It.*) A measure, a bar.

Misurato (*It.*) Measured, in strict time.

Mit (*Ger.*) With; as, *mit Begleitung*, with accompaniment; *mit Bewegung*, con moto; *mit Empfindung*, with feeling.

Mitos (*Gk.*) μίτος. A thread; a term sometimes used for the string of a lyre.

Mittelcadenz (*Ger.*) An imperfect cadence, or half-close. [Cadence.]

Mittelstimmen (*Ger.*) Inner parts.

Mixed Cadence. An old name for a cadence, consisting of a subdominant followed by a dominant and tonic chord; so called because the characteristic chords of the plagal and authentic cadences both appear.

Mixed Voices. [Equal Voices.]

Mixolydian. [Greek music.]

Mixture. An organ stop, consisting of several ranks of pipes to each note. It is only used in combination with the foundation and compound stops, as it consists of high harmonics of the ground tone. [Organ.]

Mode. (1) A scale [Greek music.] (2) A species of scale; as, *major mode*, *minor mode*. (3) A church scale. [Plain Song.]

Moderatamente, Moderato, Moderazione, con (*It.*) Moderately.

Moderatissimo (*It.*) Very moderate.

Modes (secular). A mediæval attempt at creating time-modes (various successions of mensurable notes) analogous to Church scale-modes. Franco enumerates five.

Modinha. A Portuguese love song.

Modulation. (1.) Movement or graduation of sound; (2.) A change of key.

One of the chief characteristics of modern music is the constant change of key. In many of the recognized "Forms" it is absolutely necessary to introduce "second" subjects in a different, although a related key to the chief theme. Hence, in old Treatises on Music the subject is barely touched upon, whereas modern authors enlarge upon it. Modulation is of three kinds, Diatonic, Chromatic, and Enharmonic. The first of these is sometimes called natural; the last two, artificial.

A Diatonic modulation is the passing from one key to another by using chords from *relative keys.*

Ex. 1. BEETHOVEN, Symphony No. 1.

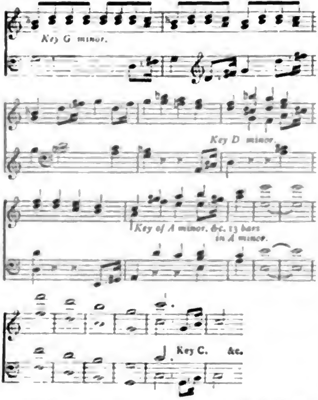

In the above (Ex. 1) a modulation is made from G minor to C, thus: G minor is a relative key of D minor, which is a relative key of A minor, which is the relative minor of C.

When a remote key is reached by relative keys, the modulation is by some said to be *extraneous.*

A chromatic modulation is the passing from one key to any other, by means of unrelated keys.

Ex. 2. BEETHOVEN, Symphony No. 3.

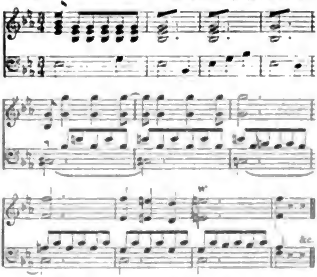

The above (Ex. 2) shows a chromatic modulation from the key of E♭ to that of F.

An Enharmonic Modulation is the passing from one key to another by the alteration of the *notation* of some of the over-lapping sounds, *e.g.:*

SCHUMANN, Song "Widmung."

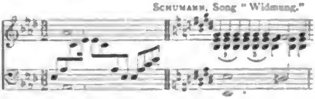

It is not easy to distinguish between the use of the terms Modulation and Transition. It is generally understood that the former denotes an entry into a new key, with the intention of remaining for some time in it; the latter, the passing through a key rapidly, as for instance is often done in sequences. But by some, the word Transition is used for a *rapid modulation.* The Tonic Sol-faists call modulation in general, transition. Music-masters, who delight in technicalities, have, in addition to the above, divided modulation into (1) appropriate and digressive; (2) simple and compound; (3) partial and complete. (1) Appropriate modulation is when the sense of key is not disturbed by the movement of chords; digressive, when a complete change of key is made. (2) A simple modulation is a change of key into the next remove, that is, into a key having one more sharp or flat in the signature; a compound is the passing through relative keys into a remote key. (3) Partial modulation is a temporary change of key; complete, the establishment of a new key.

Moduliren (*Ger.*) To modulate.

Modus (*Lat.*) Mode. (1) A scale, as *Dorian mode*, &c. [Greek music.] (2) One of the three divisions of mensurable music [see Tempus and Prolatio]. Modus major was the

Digitized by Google

division of a Maxim (*notula maxima*) into Longs. Modus minor the division of a Long into Breves. The "Modus major" was perfect when the Maxim contained three Longs, imperfect when it contained two. The "Modus minor" was perfect when the Long contained three Breves, imperfect when it contained two. The following (from Franchinus) exhibits the different kinds of Modus.

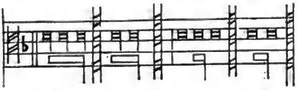

The division of Modus into "major" and "minor" was later in date than the division of mensurable music, into Modus, Tempus, Prolatio. Originally Modus (like Tempus and Prolatio) was only divided into Perfect and Imperfect. The former having the sign [· · ·], the latter [· ·| This seems the more consistent, as then *Modus* is the division of Longs into Breves; *Tempus* the division of Breves into Semibreves; *Prolatio* the division of Semibreves into Minims. Each of these is *Perfect* when the division under it is trinary, *Imperfect* when binary.

(3) A Gregorian tone. This use of the word is not strictly correct, and has always been condemned.

Moll (*Ger.*) Minor.

Molle (*Lat.*) Soft. A term applied in mediæval music to B *flat*, as opposed to B natural, which was called B *durum*. Hence, the term came to signify major and minor mode, as in the German, *e.g.*, A dur, the key of A major; A moll, the key of A minor. Hence too, the French formed the word *bémol*, a *flat*.

Mollemente (*It.*) Softly, sweetly.

Molltonart (*Ger.*) The minor mode.

Molto (*It.*) Much, very; as, *molto adagio*, very slow; *molto allegro*, very quick; *molto sostenuto*, much sustained.

Monaulos (*Gk.*) μόναυλος. A Greek single-pipe made of a reed. A flûte à bec.

Monferina (*It.*) An Italian peasant dance.

Monochord. A single string stretched across a board or soundboard, under which a moveable bridge can be moved at pleasure. By placing under the string a diagram of the proportionate lengths of string required for the production of just intervals, the ear can be trained and experiments can be made. It was anciently called, or rather, the results obtained from experiments with it, the harmonic canon. [Acoustics.] [Temperament.]

Monocordo (*It.*) *Monocorde, à* (*Fr.*) On one string.

Monodia (*It.*)
Monodie (*Fr.*)
Monody.
A song for a single voice, generally of a plaintive character. The term was originally applied to vocal solos in the church service.

Monodrama. A dramatic piece for one performer only.

Monotone, to. To recite words on a single note without inflections.

Montant (*Fr.*) Ascending.

Monter in ut, re, &c. (*Fr.*) To sing a scale of c, d, &c.

Montré (*Fr.*) Mounted diapason. An organ stop whose pipes form part of the case or are placed away from the soundboard. One of the foundation stops is generally used for this purpose.

Morceau (*Fr.*) (1) A piece; a small composition of an unpretending character. (2) An excerpt.

Mordente (*It.*), *Beisser* (*Ger.*) A beat, or turn, or passing shake.

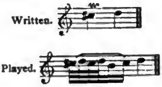

Morendo (*It.*) Dying away. A direction that the sounds of voices or instruments are to be gradually softened, and the pace slackened.

Moresca (*It.*) The Morris dance.

Morisco (*It.*) In the Moorish style.

Moriscoe, to. To dance the Morris dance.

Mormorando
Mormorevole
Mormorosa
(*It.*) In a gentle, murmuring, whispering manner.

Morris dance. *Moresca* (*It.*) A rustic dance, performed in spring and summer time. There are many records extant to prove the universal popularity of this dance, both in the parish accounts of several dates and in the writings of poets of various periods. Douce in his illustrations to Shakespeare, supposes "that the Morris-dance derives its name from the Moors, among whom it originated, and that it is the same that gave rise to the Fandango. It was probably brought to England in the time of Edward III., when John of Gaunt returned from Spain. Few vestiges can be traced of it beyond the time of Henry VII., about which time, and later, the churchwardens' accounts show the dance to have been very popular at parochial festivals." Laneham, 1590, in his description of a Bride Ale, mentions "a lively Morisdauns according to the auncient manner; six dauncers, Mawd-Marion and the fool."

The pipe and tabor were the ancient and are the present accompaniments to the dance, which is still occasionally performed at rural festivals. Tunes of various kinds were associated with it.

Morte (*Fr.*) The death note of any hunted animal sounded upon a bugle. "And whan the hare is take, and your houndes have ronne well to hym, ye shul the morte blowe oftirward, and ye shul yif to your houndes the halow." Twety in Rel. Ant., 1, 153, quoted in Furnivalle's reprint of the Percy MS.

> "And then to sigh, as't were
> The mort of the deer."
>
> SHAKESPEARE.

Mosso (*It.*) Moved, as *più mosso*, more moved or faster; *meno mosso*, less fast.

Mostra (*It.*) A direct ✓. A sign, suggested by Avison, for pointing out to a performer the entry of a particular point or subject.

Motet. A vocal composition in harmony, set to words generally selected from the Scriptures, or to paraphrases of the sacred writings. The motet was, at one time, a varied treatment of a given theme similar to the poem called in Spanish a "moto," referred to in the article Madrigal. Like the madrigal, the motet was at first set to words of a profane character, and there are ecclesiastical decrees extant forbidding its use in church. We read in Durandus "De modo generalis concillii celebrandi," cap. xix. "Videtur valde honestum esse quod cantus indevoti et inordinati motetorum et similium non fierent in ecclesia." And that the character of the motet was at one period the reverse of sacred, the following quotations from "Le Roman de la Rose," conjectured by Warton to belong to the 13th century, will show:

> "Qu'il faist rimes jolivettes
> Motes, fabliaux, et chansonettes
> Qu'il veuille à sa mie envoier."

And further:

> "Chantant en pardurableté
> Motes, gaudias, et chansonettes."

In the Constitut: Carmelit: Lib. III.:

> "Neque motetos, neque uppaturam vel aliquam cantum magis ad lasciviam quam devotionem provocantem, aliquis decantare habeat, sub pœna gravioris culpæ."

And Du Cange quotes (verbo Motulus) a decree of Odo, Archbishop of Rouen, in which it is said:—"In festo S. Johannis et Innocentium nimia jocositate et scurrilibus cantibus utebantur utpote farsis, conductis, motulis; præcepimus quod honestius et cum majori devotione alias se haberent."

The word motet was synonymous with *pulpitre* in the 15th century. In the account of the solemn entry of John of Burgundy, Bishop of Cambray in 1442, we read, "en vidant de l'église, les enfants d'autels canterent le motet ou pulpitre, tournez le visage vers l'autel."

Morley, in his "Introduction to Musicke," imperfectly describes a motet as "a song made for the church, either upon some hymn or anthem, or such like; and that name I take to have been given to that kinde of musicke in opposition to the other, which they called canto fermo, and we do commonlie call plainsong, for as nothing is more opposit to standing and firmness than motion, so did they give the mòtet that name of moving, because it is in a manner quight contrarie to the other, which after some sort, and in respect of the other, standeth still."

Probably for the reason above quoted, Du Cange suggests that the motet was originally of a gay and lively nature, and the similarity of style between the motet and madrigal of the 16th century would point to the conclusion that the titles were interchangeable until the character of the words of the madrigal fixed the title for this sort of writing, and the word motet remained to describe movements more especially intended for the church service.

Metrical psalms and hymns in which the several verses are sung to a varied setting are called motets in the Roman Catholic Church. Many sacred cantatas of unconnected movements are also described as motets. The early anthems took the place of the motet in the Anglican church, and many of the church compositions of the Elizabethan musicians are of the motet form. [Anthem.] [Madrigal.]

Motetto (*It.*) [Motet.]

Motion. (1) The movement of a single part with reference to intervals taken by it. *Conjunct motion* takes place when the sounds move by single degrees of the scale, *e.g.*, C, D, E, F; *disjunct motion* is when they move by skips, *e.g.*, C, F, D, G. (2) The movement of two or more parts with relation to each other. *Similar* or *direct motion* is when parts move in the same direction either by single degrees or by skips; *contrary motion* is when parts move in opposite directions; *oblique motion* is when one part remains stationary while another moves.

Motivo (*It.*) *Motive.* (1) The sort of movement indicated by the opening notes of a sentence. (2) A subject proposed for development.

Moto (*It.*) (1) Motion, movement, as *con moto*, with spirited movement, keeping up the interest of the music. Hence *con moto* has become a time-sign, signifying *rather fast.* (2) *Moto continuo*, continuous motion, the constant repetition of a particular musical figure or group of figures. (3) *Moto con-*

Digitized by Google

trario, contrary motion; *moto obbliquo*, oblique motion; *moto retto*, direct or similar motion [Motion]. (4) *Moto precedente*, at the preceding pace; *moto primo*, at the first pace.

Motteggiando (*It.*) In a bantering, jesting, jocose manner.

Motus (*Lat.*) Motion; *contrarius*, contrary; *obliquus*, oblique; *rectus*, direct or similar; *conjunctivus*, conjunct; *disjunctivus*, disjunct. [Motion.]

Mounted cornet. [Cornet, § 3.]

Mouth. The speaking part of an organ pipe, as opposed to the *foot*, through which the wind enters.

Mouth organ. [Pan's-pipes.]

Mouthpiece, *Embouchure* (*Fr.*); *Imboccatura* (*It.*); *Mundstück* (*Ger.*) That part of a wind instrument which is put into the mouth of the performer. In the case of brass instruments the end of the instrument is placed on the exterior of the lips, and in the case of reed instruments the reed itself is inserted in the mouth. *Kessel* is the German name for the hollow or *cup* of a brass instrument by means of which the lips set the column of air in vibration, and *Schnabel*, or *beak*, is the pointed end of oboes, clarinets, &c.

Movement. (1) Motion of melody, or of parts [Motion]. (2) A division, or definite portion of a work, as *first movement*, *slow movement*, &c., of a sonata or symphony, or other extended composition. (3) A portion of a musical piece separated from the rest by a complete change of time or key.

Mund (*Ger.*) Mouth of a pipe or musical instrument; *Mundstück*, mouthpiece.

Munter (*Ger.*) Lively, *allegro*.

Murky. A piece of harpsichord music, having a bass consisting of broken octaves, thus:

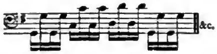

Musars. Ballad singers of the troubadour period.

Musette (*Fr.*) (1) A small bagpipe (*corna musa*) formerly much used by the various people of Europe. (2) The name of a melody, of a soft and sweet character, written in imitation of the bagpipe tunes. (3) Dance tunes and dances in the measure of those melodies. (4) A reed stop on the organ.

Musica (*It.*) Music. *Musica di camera*, chamber music; *musica di chiesa*, church music; *musica da teatro*, operatic music.

Musica ficta. (1) Music which singers al[illegible]d by the introduction of accidentals [illegible]ng to traditional rules. These rules [illegible] into force as early as 1300, if not earlier, and mark the gradual separation by the people of their scale (our scale) from the church modes. For example:

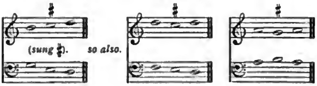

but such accidentals were only sung where we should naturally expect a leading note. The same melodic progression would not be altered in such cases as these:

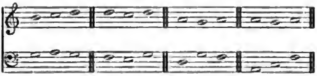

In descending, a flat was often introduced on the sixth degree; thus:

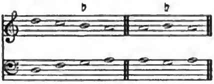

But here again attention has to be given to the harmony suggested. Up to the latter part of the sixteenth century, or even later, composers were often very undecided as to the use of accidentals. (2) The term was sometimes applied to music in which accidentals were marked, that is to say, in keys remote from church-modes (see "Erotemata musices practicæ," 1563, Cap. V.).

Musical Box. A portable instrument, the sounds of which are produced by a steel comb having teeth of graduated length. Projecting pegs or stops, in a metal barrel turned by clockwork, set the teeth in vibration.

Musical Glasses. [Armonica.] [Glass.]

Musici. A name given to the followers of the Aristoxenian system of music. (See Pythagoreans.)

Musico (*It.*) A name given to a male soprano.

Muta (*It.*) A direction to a player on a horn, trumpet, &c., or on drums, to change the *key* of his instruments; as, *muta in A, B*.

Mutation Stops. In an organ, stops giving sounds not in unison or octaves with the foundation stops, *e.g.*, the twelfth, tierce, &c.

Mute. A small instrument of brass, wood, or ivory, so made that it can be readily fixed upon the bridge of a violin or violoncello, to damp or deaden the sound. The direction for its use is written *con sordini* or *muta*, its discontinuance by *senza sordini*. A leather pad of a pear shape is employed as a mute for brass instruments, which, inserted in the bell, produces the effect of sound at a distance.

M. V. Abbreviation of *mezza voce*.

Digitized by Google

N

Nabla *νάβλα* (*Gk.*) [Nebel.]

Nablium (*Lat.*) [Nebel.]

Nacaire (*Fr.*) A large drum.

Naccare or **Gnaccare** (*It.*) [Castanets.]

Nacchera (*It.*) A military drum. [Nakeres.]

Naccherone (*It.*) A large military drum.

Nachahmung (*Ger.*) [Imitation.]

Nachdruck (*Ger.*) Emphasis, accent.

Nachspiel (*Ger.*) A postlude.

Nächstverwandte Töne (*Ger.*) The nearest Relative Keys, *q. v.*

Nachthorn (*Ger.*) Literally "Night-horn," an organ stop consisting of stopped wood pipes of a moderately large scale, the tone of which is somewhat like that of a horn.

Nænia (*Lat.*) A funeral song of the Romans, mentioned in the laws of the twelve tables. "Honoratorum virorum laudes in concione memorantor; easque næniæ ad Tibicinem prosequuntor.

Nafiri. An Indian trumpet.

Nagaret or **Nagareet.** An Abyssinian drum; a kind of kettledrum. [Nakeres.]

Naguar. An Indian drum with one head only.

Naïf (*Fr.*) Simple, *naïvement*, artlessly, unaffectedly.

Nail Violin. [Eisenvioline.]

Naked fifth. The interval of a fifth without a third.

Naked fourth. The interval of a fourth without the addition of any other interval.

Naker. A drum. [Nakeres.]

Nakeres (*Old Eng.*) The explanations of this word given by various authors are somewhat conflicting. Albert Way, in his edition of the "Promptorium Parvulorum," considers the word identical with the nagârah, or drum of the Arabs and Moors. Joinville speaks of the minstrels of the Soudan "qui avoient cors Sarrazinois, et tabours, et nacaires." Most probably nakeres were small metal drums, used in pairs. In the poem, "Sir Gawayn and the Grene Knyzt," (Early English Text Society,) the word occurs:

> "Trumpez & nakerys
> Much pypyng per repayres."

A genitive case "nakeryn" is found in the early English alliterative poems published by the same society. "Belshazzar's Feast."—(No. xiii.):

> "& ay ðe nakeryn noyse, notes of pipes,
> Tymbres & taborns, tulket among,
> Symbales & sonetez swared ðe noyse."

Chaucer, too, makes mention of them in his knight's tale:

> "Pipes, trompes, nakerers, and clariounes,
> That in the bataille blowen blody sounes."

In Strutt's "Sports and Pastimes," a payment to Janino le Nakerer of sixty shillings as one of the minstrels of the King (Edward II.), is recorded.

Nanga. A negro harp. [Harp.]

Narrante (*It.*) As if narrating. A direction to a singer, implying that the music is to be subordinate, as it were, to the recitation of the words.

Nasard, Nazard or **Nassat.** An organ stop sounding a twelfth above the foundation stops. It will therefore be 2⅔ feet in length.

Nason Flute. A stop of 4ft. tone, sometimes found in old organs. It is of a soft and delicate quality of sound, and consists of stopped pipes.

Nassat. [Nasard.]

National Air, or **National Music.** Popular music, peculiar to, or characteristic of, a particular nation. It often happens, from the fact of a melody possessing the characteristics which render it generally popular, that it is passed rapidly *vivâ voce* from one person to another, with less consideration of the author than of the music he has composed: hence a tendency to lose the names of the composers of national music; but of course, music is none the less national because it has its author's name attached, and recent research has discovered the authorship of many a tune which was supposed to have this small claim to nationality.

It is impossible to describe a piece of music of any kind as national, until it has proved, by its continued existence, that it has those qualities which will enable it to live. Thousands of melodies, some of them unfortunately of the most worthless kind, from time to time take a firm hold of the common peopl[illegible]f this and other countries; but the hold is not lasting, their popularity is but ephemeral, hence they are not truly national airs; and the airs which have become entitled to be called national, numerous though they are, are probably a mere fraction of the vast number which have had a short existence, and then sink into complete oblivion. Bearing these considerations in mind, it will not be difficult to give a definition of a "national air," as follows: an Air which, by its reflection or representation of a sentiment, taste, or habit of a nation, either through the music alone, or words and music combined, has become so commonly known and used by that nation, as to be inseparable from the idea of its special or characteristic music.

When the customs or tastes of nations are very similar, if their musical scale has also great similarity, we shall, of course, find many tunes common to all, and claimed by each. The more general the adhesion to the sentiment, or the greater the area over which the custom extends, or the more general the use of the scale from which it is formed, the more cosmopolitan will a piece of music become. The climate of a country, by acting upon the temperament of the inhabitants, influences indirectly the style of its music. It is incorrect to suppose that the minor mode is a universal characteristic of national music. From a statement drawn up in a tabular form by Carl Engel, in his important work, "An Introduction to the Study of National Music," it appears that in Sweden, Norway, Russia, Finland and Hungary, the minor mode predominates in the national music. In some countries, as for instance, in Moldavia, and Wallachia, and in Russia, the music seems to vacillate between the two modes, the tunes sometimes commencing in the major and ending in the minor. Sometimes also, tunes commence in the minor and end in the major. As might be expected, some melodies become altered in various ways in passing from generation to generation by oral tradition; our English Ballad literature, so ably collected and arranged by W. Chappell, abounds with examples. Some nations add profuse grace notes to a well known melody; this is the case with the Spaniards, but chiefly with the Arabs and Persians. The Welsh performance called *Pennillion* consists of varied accompaniments to the voice parts, and the tunes played on the violin by the Transylvanian gipsies abound in flourishes. As regards the scale in use in different countries, it does not happen, as might be expected, that a simple form of scale is used among nations of a low order of civilisation, and a more complicated form by a more civilised race. The contrary is often the case. The scale used by the Maories is enharmonic, that is, contains intervals smaller than a semitone; while that of the Chinese is pentatonic, that is, consists of five sounds succeeding each other in a series similar to that which would be produced by playing only on the black keys of a pianoforte. The power of the ear to distinguish between sounds differing from each other by only a very small interval, varies considerably in different nations, some savage nations, the Maories for instance, having a perfect power of distinguishing between quarter-tones. The study of national music leads to the inevitable conclusion that all musical scales are purely conventional; the modern diatonic scale having no better claim to be founded on nature, or natural harmonics, than the complicated scales of the Arabs, Egyptians, or Hindus. Closely connected with the history of national music, is the history of national musical instruments, the two subjects throwing much valuable light on each other. The quality and pitch of that particular voice most common in any nation, will also much influence the style of the popular music. "Some Asiatic nations," says Engel, "sing in shrill notes by straining the voice to its highest pitch; others delight in a kind of vibration or *tremolando*. Some sing habitually in an undertone; others in a nasal tone. Others, again, cultivate with predilection the falsetto, and usually introduce it into their vocal performances." It is a fact generally known in this country that Russia abounds in deep Bass voices, and Italy in fine Tenor or Baritone voices. It is said that Hottentots are chiefly Tenors, and that the Chinese sing entirely in falsetto. A proverb current in the 14th and 15th centuries (see Chappell's Pop. Mus., p. IX., Vol. I.) says, "Galli cantant, Angli jubilant, Hispani plangunt, Germani ululant, Itali caprizant." The allusion to the English is undoubtedly meant for a compliment, they "sing cheerfully;" in "caprizant" is probably an allusion to the frequent use of the *tremolo*. It was commonly stated by English authors that England possessed no national music, but the researches of Chappell have brought together a collection of melodies of various kinds, which cannot be surpassed by any European nation. Unfortunately, our national songs are but little known in our large towns and cities; they are, however, handed down carefully in many rural districts where they may be heard in a more or less correct form.

Natural, ♮. A sign which restores a note to its position in the natural scale. It has the effect of sharpening a note previously flattened, or of flattening a note previously sharpened. It is an accidental, that is, it does not occur in the signature of a piece

Digitized by Google

of music, unless for the purpose of cancelling all or part of a previous signature:

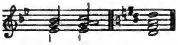

The earliest known use of the sign is found in Bonaffino's "Madrigali Concertati," 1623, a work in which also bars are employed as marking the correct divisions of time.

Natural harmonics. The sounds produced by dividing a string into aliquot parts. [Acoustics.]

Natural key. Key of C.

Natural modulation. Diatonic modulation, as opposed to Chromatic. [Modulation.]

Natural pitch. The pitch of a pipe before it is overblown.

Nay. A flute used in modern Egypt, but almost identical with the long flute anciently employed in that country. The most common *nay* is called the "Dervish flute," because with it those people accompany their songs at their zikrs or religious dances. Lane describes it as "a simple reed, about eighteen inches in length, seven-eighths of an inch in diameter at the upper extremity, and three-quarters of an inch at the lower. It is pierced with six holes in front, and generally with another hole at the back." It seems to have been played like our own flute, and like it to be capable of additional compass when blown hard.

Nazard. [Nasard.]

Neapolitan sixth. A name, apparently without much reason, given to a chord occurring on the subdominant of a minor key, and consisting of a minor third and minor sixth, *e.g.*:

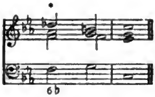

It evidently can have nothing to do with the *key* of D♭ major in the position given above, although it appears at first sight to be the first inversion of the common chord of D♭. Nor can it be derived from the root G, inasmuch as a flattened fifth of a root is not admitted to exist by the best theorists. Some authors have explained the nature of the chord by calling the F A♭, notes derived from G, and the D♭ a note derived from C. This would make the chord a double-root chord (from G C) having the minor ninth of both roots, heard with the minor seventh of the upper root. Another explanation found is that it is derived from the minor scale of the subdominant, in its modern form, *e.g.*:

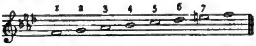

If we combine 1 3 6 of the above scale, we obtain the chord under examination. That it is so derived seems the more probable, inasmuch as it is often *preceded* by a common chord of that key-note, *e.g.*:

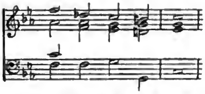

and sometimes *followed* by one, *e.g.*:

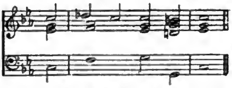

It is sometimes followed at once by a major tonic chord, *e.g.*:

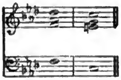

Nebel (*Heb.*) One of the most important of the stringed-instruments of the ancient Hebrews. It was not as ancient as the kinnor, but was probably of more elaborate construction. It is almost always rendered, in our version of the Bible, by the word psaltery, two or three times by *viol*, once (Ps. lxxxi. 2) by the word psalm, "Take a *Psalm*, bring hither the tabret." This use of the word psaltery is most unfortunate. because it has by almost universal consent been the name of the ancient dulcimer, *psanterin*, *psalterion*, *psaltery*, *sautry*, &c. This has led to a double error, many theologians thinking that a psaltery is a *harp*, many musicians thinking that a nebel was a *dulcimer*. In all probability the nebel was a harp. It was portable because Saul met a company of prophets "coming down from the high place with a *nebel*." That it was used on secular occasions is proved by Isaiah's words (v. 12), "The harp, and the *nebel*, the tabret, and pipe are in their feasts." David was proficient on this instrument, and the list of his nebel-players is given in 1 Chron. xxv. 1, and elsewhere.

Nebendominant (*Ger.*) The dominant of the dominant, as D is in the key of C.

Nebengedanken (*Ger.*) Accessory ideas, or subordinate subjects.

Digitized by Google

Nebenstimmen (*Ger.*) (1.) Unessential parts, as for instance those notes which may be added to a triad, or those parts which are in unison with others. (2.) Accessory stops of an organ.

Nechiloth (*Heb.*) The generic name for wind-instruments.

Neck. That part of instruments of the violin and guitar class, which lies between the peg-box and the belly. To its upper surface is attached the finger-board or fret-board.

Nefer. An Egyptian guitar, called also Nofre. [Guitar.]

Neghinoth. [Nechiloth.]

Negligente, negligentemente, negligenza, con (*It.*) In a negligent or careless manner.

Nel battere (*It.*) At the down stroke of the bar; on the beat.

Nel stilo antico (*It.*) In the old manner or style.

Nel tempo (*It.*) In time (after an *ad libitum*), or, in the previous time.

Nero (*It.*) A crotchet. ♩

Nete (*Gk.*) [Greek music.]

Nettamente, netto (*It.*) Neatly, with precision.

Neumes. The musical notations employed from the eighth or ninth to the twelfth century. Their origin is doubtful; Kiesewetter considers them to be the ancient *nota Romana*, others believe them to have been of Asiatic origin. There can, however, be little doubt that the earliest system of musical notation was merely a series of directions as to the intonation (accentus) of the voice; the *acute* accent, directing the raising of the voice; the *grave*, the flattening of it; the *circumflex*, a movement up and down. These afterwards were modified into conventional cadences of various kinds, new signs being added from time to time to those already recognized, and the force and meaning of the old signs being amplified.

The progress of these signs from the acute, grave, and circumflex into *neumes* of different shapes can be traced with tolerable precision. The acute accent grew into the *virgula*, the grave accent into the *punctus*, the circumflex into the *clivus* and *podatus*. Hence, Coussemaker has, in his valuable "Historie de l'Harmonie au Moyen Age," divided neumes into *generators* and *derivatives*, separating the former into two kinds, "simple" (virgula and punctus) and "compound" (clivus and podatus); the latter class also into two kinds, "bound" and "conjoint." The following shews the accentual neumes as used in the eleventh century:—

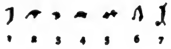

The names above are as follows:—1 Accentus acutus; 2 accentus gravis; 3 percussionalis brevis; 4 percussionalis longa; 5 inflatilis; 6 circumflexa; 7 muta.

The next illustration of neumes is part of an interesting list given in Gerbertus "De Cantu et Musica Sacra," Vol. II.:

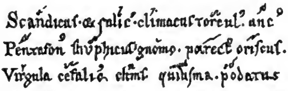
Scandicus · et salic' · climacus torcul' · anc'
Pentafon strophicus gnomo · porrect' oriscus ·
Virgula cefalic' clivis quilisma · podatus

It will be seen that the above shews the form of the scandicus, salicus, climacus, torculus, ancus, pentafon (spelt also penta phone), strophicus, gnomo, porrectus, oriscus, virgula, cefalicus (cephalicus), clivus, quilisma, podatus. These are *dis*-arranged for the purpose of getting them into verse. But those who are curious on the subject will find two admirable explanatory tables in the work above alluded to, pp. 184, 185.

Neumes were originally written above the words to be sung to them. After the ninth century they began to assume graduated height and position, this was succeeded by the addition of lines and clefs. Neumes then grew into ligatures and into notes representing sounds of different lengths. The foundation of the modern system of musical notation was thus laid. [Notation.] The word neuma is probably connected with the Greek pneuma (πνεῦμα), a breath, and signified originally a group of sounds to be sung to a syllable; but later on this particular meaning seems to have merged into the more general one of *notes;* and *neumes* and *notes* are used almost as synonyms. When applied to the system of notation the word is spelt without the letter p (neuma); when applied to a series of notes to be sung to one syllable, the word seems generally to have retained its p (pneuma). The practice of singing pneumas has been at times carried to an almost ridiculous extent. They may be divided into two classes: first, the expansion of a melody in the middle of a word, as in modern runs or divisions; *e.g.*:

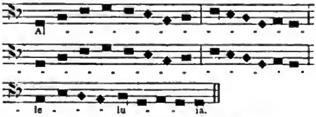

(*From the Compline Office on Holy Sunday.*)

This is not opposed to modern notions of musical form.

The second kind may be described as a sort of *coda*, or tail-piece to the final word of a sentence, and such pneumas were generally sung to the last syllable of the last word; *e.g.*:

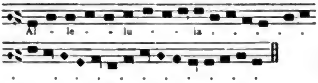

Neuvième (*Fr.*) The interval of the ninth.

Nexus (*Lat.*) One of three branches of Melopœia—*Ductus*, *nexus*, *petteia*. Ductus was the movement up or down or concurrent by single degrees. Nexus was movement by an interval, and was of three kinds; *rectus*, as *fa-gb-ac*; *revertens*, as *ca-bg-af*; *circumstans*, as *af-gb-af-ge*.

1. Direct.

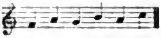

2. Reversed

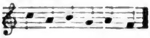

3. Circumcurrent.

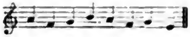

[Melopœia.]

Niederschlag (*Ger.*) The accented part of a bar.

Nineteenth. An organ stop. [Larigot.]

Ninna, nanna (*It.*) A cradle song.

Ninth, interval of a. A compound interval, equal to a second in the superior octave. It may be major, minor, or augmented, *e.g.*:

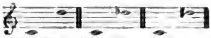

Ninth, chord of the major. A chord formed by a combination of thirds starting with the dominant or fifth of the scale, called by some writers the "added ninth," because it consists of a chord of the dominant seventh, with the addition of the ninth; by others the "dominant ninth," because it occurs on a dominant bass.

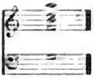

It is composed of five sounds, and, therefore has four inversions. Like all chords of the ninth, in its inversions the root or ground-note is seldom heard. They may be therefore written thus:

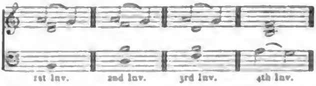

In the above, the ninth is made to resolve first, leaving the seventh to be resolved afterwards. In the following, both discordant notes are resolved in the next chord:

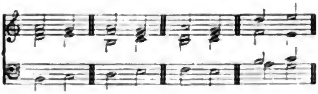

The fourth inversion, having the ninth in the bass, is not so commonly met with as the others. In the resolutions given above it will be seen that the third of the dominant, that is, the leading note, ascends to its tonic; the minor seventh descends; the major or added ninth descends also. The first inversion of this chord is sometimes termed the "chord of the seventh on the leading note," because the leading note of the scale is in the bass, and an interval of a seventh is found in the chord.

For similar reasons the second inversion is sometimes called the chord of $\substack{6\\5}$ on the supertonic, the third inversion that of $\substack{6\\4\\3}$ on the sub-dominant, the fourth inversion that of $\substack{6\\4\\2}$ on the superdominant. This chord and its inversions (excepting the fourth) are often used without preparation.

Ninth, chord of the minor. One of the most important ingredients of modern music. Not only is it exceedingly beautiful to the ear, but from its peculiar form it gives the greatest possible facilities for modulation from key to key, whether closely related or not. It consists of a dominant, its major third, major (perfect) fifth, minor seventh, and minor ninth, *e.g.*:

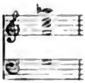

In its inversions the dominant, that is, the root or ground-note, is nearly always omitted, *e.g.*:

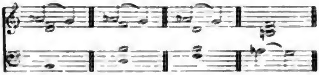

In the above examples the ninth only has been resolved, leaving the seventh unresolved; but in resolving the seventh it will be found that the major or minor third of the succeeding chord may be used, thus:

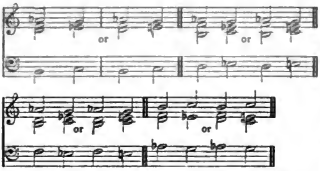

Hence the chord is as often found in music in a major key, as in that in a minor. The beauty of the inversions no doubt arises from the fact that they consist practically of a combination of minor thirds, *e.g.*:

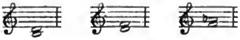

The alteration of the notation of the inversions of this chord gives scope for rapid enharmonic modulation. Take for example the second inversion:

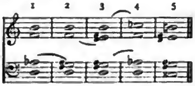

No. 1 is derived from G, dominant of C
No. 2 „ E, „ A
No. 3 „ C♯, „ F♯
{No. 4 „ B♭} „ E♭}
{No. 5 „ A♯} „ D♯}

Nos. 4 and 5 are a complete change from a flat to a sharp key. Without altering the *sound* of the above chord it may be made to lead into any of the above keys or the keys related to them. Such a change of notation as this, without a change of sound, is termed an enharmonic modulation. It has already been shewn that the chord may resolve either into the major or minor mode; therefore the tonic minors of the above notes, and also the keys related to them, may be reached with equal facility. On looking at the five changes given above, the reason why the ground-note or root is omitted from the chords becomes obvious, for not only would the symmetry of form produced by the conjunction of the minor thirds be destroyed, but also a definite resolution would become absolutely necessary, or, in other words, the invaluable property of "doubtfulness of key" belonging to such chords would be removed. One more fact should be noticed. If the list of the notes given above as the ground-notes of the five examples be read separately, thus: G, E, C♯, B♭, or, as Nos. 4 and 5 are convertible, G, E, C♯, A♯, and their tonics, C, A, F♯, E♭, or C, A, F♯, D♯, we get the following:

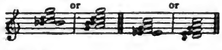

which will be found to be identical with those derived from them when separated into ground-notes, both as to the nature of their construction and their capability for enharmonic modulation. The first inversion of this chord is often called the chord of the *diminished seventh*, because of the interval between the bass note and minor ninth of the root.

Ninth, chord of the suspended. A name given to the chord of the ninth on the tonic, as opposed to that of the ninth of the dominant, owing to the fact that the former is more often used as a *prepared* discord than the latter. The ninth may or may not be accompanied by the seventh (in this case the leading note). In the following examples the seventh is omitted:

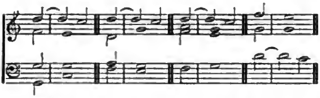

The seventh is included in the following examples:

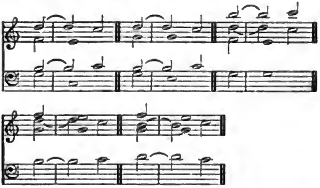

It will be seen that the note on which the suspension resolves is not heard with the suspension, except in the original position of the chord; also, that the third inversion is only to be obtained when the seventh is used. The ninth is not unfrequently resolved upwards, in which case some authors would not call it a "suspension," but a "retardation," *e.g.*:

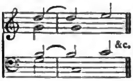

The ninth and seventh are both used with the fourth, or, as it is more properly termed, the eleventh on the tonic; under "Suspension," examples of their combination will be found.

Digitized by Google

The division of discords into discords of suspension, discords by retardation, discords by addition, and fundamental discords, is purely arbitrary, consequently hardly any two authors apply these names in the same way. It is highly desirable that a simple and consistent method of arranging chords should be generally adopted.

Nobile, nobilmente, con nobilità (*It.*) With grandeur, nobly.

Nocturne (*Fr.*) [Notturno.]

Nocturns. Services of the church held during the night, for which certain portions of the Psalms are set aside, each of which is termed in the Breviary a nocturn. [Horæ Canonicæ.]

Nodal line. [Acoustics, § 13.]

Node. [Acoustics, § 7.]

Nodus (*Lat.*) *Lit.* a knot. A Canon. So called because compositions of this class were sometimes given as enigmas, the meaning of which had to be unravelled. [Canon.]

Noël (*Fr.*) "Good news." *Old Eng.* Nowell. A word used as a burden to Carols at Christmas. Hence, Carols are sometimes called noels, or nowells. [Carol.]

Nofre. [Nefer.]

Noire (*Fr.*) A crotchet ♩. The black note.

Noise (*Old Eng.*) Music, or a performance of music. As in Shakespeare's Henry IV., part II., Act 4, "See if thou canst find Sneak's *noise*." So also in Milton's "Ode on the morning of Christ's nativity:"

"When such music sweet
Their hearts and ears did greet,
As never was by mortal finger strook;
Divinely warbled voice
Answering the stringed noise,
As all their souls in blissful rapture took."

The Prayer-Book version of the Psalms has "God is gone up with a merry *noise*." ("Ascendit Deus in jubilo." *Vulgate.*)

Nomenclature. In music the terms applied to the various signs employed to stand as the representatives of time, sounds, pitch, pace, and expression. Mr. Hullah, in a paper recently read before the Musical Association, says: "The signs relating to time are the breve, semibreve, minim, crochet, quaver, semiquaver, and demisemiquaver. Of these names the first three have lost their significance; the fourth is no longer appropriate; the fifth, sixth, and seventh, are arbitrary.

"The Germans call these notes, beginning from our semibreve, the *whole note*, the *half note*, the *quarter note*, and so on. These appellations, so far as they express the proportion of the first note named to those which follow it, are convenient; they form of themselves a time-table. But it is an imperfect one; for they do *not* show, without further calculation, any intermediate proportion. They show at once that eight quavers = one semibreve, but not at once that four quavers = one minim. But I have a much more serious charge to bring against them. They assume what, if not always false, is, as it seems to me, not always true—that the semibreve is, or that any form of note can be absolutely a 'whole note.' What is, or what should be regarded as a whole note? If I were sure that the word "phrase" represented to all of us the same idea as it does to me, I should answer unhesitatingly that a whole note was any note that could be divided into a phrase, or—to be a little more precise—any note divided or undivided, which would fill either an entire measure or require as many beats as would make one. This would give us practically four claimants to the title of *whole note*;—the *breve*, the average whole note of the sixteenth century; the semibreve, the average whole note of our own time; the minim, and even the crotchet. For, that movements innumerable of four beats in a measure, each of which is a quaver, exist, I need not say, nor that the measure even of four semiquavers has been occasionally employed.

"It is certain that a sound lasting four beats may be expressed, and has been expressed by six different forms—the *maxim*, the *long*, the *breve*, the *semibreve*, the *minim*, and the *crotchet*. Perhaps some musician of the future may think proper to express such a note by a quaver.

"Let us now consider the names used by the French—a people possessing in high perfection the power of clear exposition of what they themselves see clearly. As usual, they leave or throw on one side whatever they regard as uncertain, or equivocal, or not commonly accepted, and proceed to deal with the undisputed and indisputable facts or portions of facts before them. And what is there in respect to the forms which express the relative durations of sounds? First, that they are forms, and secondly, that they are different forms;—that one is an oval or circle, that another is a circle with a stem, and another a circular spot also with a stem; and that all other notes are opaque and have not only stems but hooks varying in number. They call these notes or forms as they find them,—*round*, *white*, *black*, *hooked*, *twice-hooked*, and *thrice-hooked*. I certainly prefer the German nomenclature, which, though raised on a false basis is consistent, to our own, which is inconsistent as well as false; but I prefer the French to the German, because, not pretending to do so much, it does what it pretends to do perfectly.

"The nomenclature, not of sounds, but of the relations between them—the nomen-

Digitized by Google

clature of musical *intervals*, is a subject on which English theorists and practitioners are by no means agreed.

"I believe that the *seconds* and *thirds* and their inversions the *sevenths* and *sixths* found in the so-called 'natural' scale, and all scales made like it, are very generally called among us, *major* and *minor;* and that six of the *fourths* and their inversions the *fifths* are as generally called *perfect*. Here, however, agreement ends. For the one exceptional fourth and the one exceptional fifth rejoice each in as many aliases as a swindler finally run down by the police detective. To the exceptional fourth, which, according to the old theorists 'diabolus est,' I have heard and seen applied the name 'tritone,' and the epithets *sharp*, *superfluous*, *redundant* and *augmented;* to the exceptional fifth the epithets *flat*, *false*, *imperfect*, *diminished* and *equivocal*. Others might possibly be added to this list. To the name *tritone* no objection is, I think, open; it expresses the contents of the interval—three tones; but it carries with it the disadvantage of there being no corresponding name for its inversion, the exceptional fifth. *Augmented* and *diminished* are no doubt antonyms; but both are epithets which, as I shall try to show, ought to be reserved exclusively for another class of intervals—the chromatic. *Superfluous* and *redundant* are, I think, clumsy epithets; but if either is to be applied to the exceptional fourth, its antonyms *scanty* or *insufficient* should be applied to the exceptional fifth. If this last interval is to be called *false*, its inversion (the tritone) should be called *true*. Only one of these epithets seems to me quite unobjectionable—*imperfect* as applied to the exceptional fifth. As an antonym to this I have long used the epithet *pluperfect*, which has been very largely adopted.

"I objected just now to the epithets *augmented* and *diminished* as applied to these particular intervals, the exceptional fourth and fifth. I think these should be reserved exclusively to chromatic intervals. I know of course that my objection involves a principle, or rather begs a question. What *is* a chromatic interval? This question, as often happens, throws us back on another. What is a chromatic *scale?* A chromatic scale I should define, with Dr. Crotch, to be a scale containing *more than two semitones*. The so-called 'natural' scale, and all other scales made like it, is not a chromatic scale, neither are any of the ancient scales formed from the arrangement of the same series of sounds in a different order. Of these last the 'natural' *minor* scale is one, and the only one familiar to the modern musician. Only however by means of a most serious alteration has it been reconciled to modern tonality, which above all things demands, as the unequivocal sign, seal, or confirmation of a key, the combination known as the 'discord of the dominant seventh.' Such a combination on the 5th of the natural minor scale is only possible by an alteration or non-naturalization which at once brings it under Dr. Crotch's definition. In the series A, B, C, D, E, F♮, G♯, and A, we find three semitones, and one interval greater than a tone. Moreover, by skips from one note to another of a scale so constituted, we get three other intervals alien to the natural scale, the inversion of the altered second formed by F-G♯, the altered fifth formed by C-G♯, and its inversion. These intervals are, I conceive, augmentations or diminutions of intervals which would have remained unaltered, but for the artificial process needed to reconcile the minor key with modern tonality; they are therefore, I believe, generally called *augmented* and *diminished* accordingly. So all intervals, which the cultivated ear does not reject as cacophonous, formed by notes one or both of which are foreign to the key to which they are introduced, are but augmentations or diminutions of those that are natural to it. *Without change of key* we can augment certain of the unisons, seconds, fifths, and sixths; and diminish certain of the octaves, sevenths, fourths, and thirds. Now, as we have seen, in the unaltered or natural scale, major or minor, we find no examples of any one of these intervals; they are uniformly the result of artificial treatment. But, with the exceptional fourth and fifth the case is altogether different; they are not the results of artificial treatment, we find them ready to our hands; and they are as much constituent parts of the scale in which we find them as is the semitone between the third and fourth sounds. How then can the interval F-B in the scale of C, be augmented or B-F diminished? Of what are they augmentations or diminutions? Of F-E♭, or of F♯-B? Are B♭ or F♯ constituents of the scale—I do not say the key—of C? If they are, our modern tonality must be reconstructed *de fond en comble*, and every scale must be allowed three dominants instead of one. Again, we find that the intervals of the natural scale which bear augmentation are the largest of their kind in it, and those which will bear diminution the smallest. Of the seconds we can augment only the *major;* of the thirds we can diminish only the *minor*. Can we augment the tritone, the largest fourth in the scale, or diminish its inversion, the smallest fifth? Both have reached their utmost limits, and resist and defy all attempts to put them further asunder or bring them nearer together. If it be answered that they are already augmented and diminished, I ask again what was their original condition? The tritone and its inversion are, I repeat, constituents of the diatonic scale; and they are diatonic intervals accordingly. For the latter an epithet, *imperfect*, is already largely accepted; I submit to you, in the absence of a better, the epithet *plu*perfect for its inversion.

"Before quitting this second division of my subject—pitch—I will ask you to give me your attention for a few moments longer. It seems to me that musicians have much cause to complain of the way in which not merely general *littérateurs* but even scientific writers employ words to which, since music has been an art, musical artists have agreed in attaching certain definite significations. Perhaps the most glaring instance of this, and it is the only one which I shall give, is the employment of the word 'tone' to express the thing or sensation which we and they also sometimes call 'sound.' A *tone* with us is not a sound, but the relation or difference between one sound and another. This acceptation of the word would seem to be, if not as old as the musical art itself, at least of great antiquity, as is shewn in the co-existence of two such words as *tetrachord* and *tritone;* the one, observe, representing a passage of four *sounds*, or strings which produce them. The other, an interval which, though it includes four sounds, is named after the three intervals—*tones*—which separate them. We hear now of *over-tones*, or the acute sounds resulting from spontaneous vibration; and of *under-tones*, meaning grave sounds resulting from the combination of others. Some of us have occasionally been at a good deal of pains to explain that a major third consists of or includes *two* tones: if a tone be a sound, a major third must consist of three or even of five tones—or of both three and five. The most recent and extravagant employment of this word, in this sense, is in its application to great composers. Beethoven especially we often hear of as a great 'tone-poet.' I should say that if this terminology is to be accepted at all, it should be graduated or made more precise; so as to express the rank of the poet to whom it is applied. If Beethoven be a *tone* poet, some of our contemporaries should be authorized to call J. S. Bach an *augmented* tone poet, and, *e converso*, Rossini a *semitone* poet. What designation should be applied to the vast crowd of less successful aspirants to musical fame I know not. Perhaps they might be put off with some of those minute intervals the excess or insufficiency of which disturbs the minds of those who still generously devote themselves to the search after that philosopher's stone of our art—Perfect intonation.

"I pass on now to the consideration of *expression*, under which term we may class words or signs indicative of *pace*, *intensity*, and *style*.

"A growing disposition has been observable of late among the different musical peoples of Europe to use their own languages as vehicles for these indications. I think this is to be regretted: (1) as inconvenient to foreigners among whom their music is likely to go. It seems hard on an English, French, Italian, Hungarian, or Bohemian musician, that to understand a piece of music by any eminent modern German master, he must not only be a musician, but a linguist; that he should not merely be able to appreciate the musical sound of the notes in the score before him, but have also a vocabulary, practically unlimited, of German words. He opens, say, Schumann's Overture to 'Genoveva.' He sees by the position of the stave, headed by the C clef and designated '*Bratsche*,' what is the meaning of that word; he need not have much doubt about the stave similarly headed and holding three parts, against which is written *Posaunen;* by the shapes of the passages intended for them, he may construe *Ventilhorn* in Es, *Waldhorn*, and *Pauken;* and by the help of the metronome mark he may come at the meaning of *langsam*. But a little further on he encounters *leidenschaftlich bewegt*, which is harder upon him; and a little further still, *sehr frisch*, which is really too bad. This example has lately found imitators among the Scandinavians, who expound their musical intentions in words which a German of philological tastes and pursuits could doubtless make out, but which to the average German must be as unintelligible as to the average Englishman. Strange to say, the French, who take it for granted that everybody understands their language, or ought to, have not sinned in this way so much as the Germans. It is true that the scores of their operas are covered with phrases like 'avec chaleur,' 'très-simplement,' 'à demi-voix,' 'avec ironie,' but these may be regarded as 'stage directions' addressed to, and inevitably intelligible to those who are to play the parts as well as sing the music to which they refer. Otherwise French composers limit themselves in their scores to a few native words such as *détaché*, *doucement*, and the like. As for ourselves, our modern musical publications would indicate, what certainly is the reverse of true, that we are the greatest linguists on earth. It is needless to present examples of what everybody is familiar with; we have all seen, and see daily, title-pages, for instance, in which two, three, and even four languages are employed. (2) Irrespective of its practical inconvenience, which I do not wish to overrate, this practice takes from music its noblest characteristic—its catholicity. We musicians are able to discourse in a language touching to the hearts, if not always clear to the intelligences, of every people on the face of the globe; and we are furnished with an alphabet in which to write this language, which is not the invention of a single mind, but of a thousand minds—a thing which has marched on to its present perfection *pari passu* beside music itself; an alphabet so clear, to him who knows how to read it, that a musical composition, no matter of what intricacy, composed, let us say at Moscow, can, without any serious violation of the intentions of its author, and without his personal assistance, be performed within a few days, weeks, or months, in London, Paris, New York, or Melbourne—wherever there are artists to interpret it. Let us cherish this precious possession, and do what we can to prevent its acquiring a sectarian, provincial, or even national character, through the introduction of any peculiarities whatever. The directions of which I have spoken were, up to a comparatively recent time, made all the world over in one language — Italian. And even to this hour, the most Teutonic of musical composers are still obliged to resort to that language. In the score of which I have just spoken, Schumann's 'Genoveva,' there are as many Italian words or abbreviations of Italian words as there are measures, very often more. The words *dolce*, *sempre*, *basso*, *soli*, *divisi*, and the like, are of frequent occurrence; and as for the contraction of *piano*, *forte*, *crescendo*, *diminuendo*, *sforzato*, and the like, they may be counted by hundreds. Granted that Italy has not been observant of Andrew Marvel's caution—

> "The same arts that did attain
> A power, must it maintain:"

granted that she has not held her own against such competitors as Germany during the last hundred years has brought into the field, are we to lose all veneration for the people whom the slightest acquaintance with musical history will show to have been the musical teachers of all the world? Are we to kick down the ladder by which we have risen to our present superiority—to turn a cold shoulder on old friends because we have become better off, and it may be wiser than they? Not, however, to throw more sentiment round this matter than it will bear, it does seem unwise to subject ourselves to the inconveniences of which I have spoken, when they can be avoided by the simple process of using only *one* vocabulary, and that not a new or unaccustomed one, but one with which every musical people is at least partially familiar."

Nomos. *νόμος* (*Gk.*) A song. *νόμοι πολεμικοι*, war-songs.

Nona (*It.*) A ninth.

Nonenakkord (*Ger.*) The chord of the ninth.

Nonetto (*It.*) A piece of music for nine voices or instruments.

Nonny, or "Hey nonny." A common burden to old English ballads, as "fa, la, la," was to madrigals.

Nonuplet. A collection of nine notes to be played in the time of eight, or six.

Normal Pitch. [Pitch.]

Normalton (*Ger.*) The standard sound. [A. § 5.]

Normaltonart (*Ger.*) The normal scale, C major and its relative minor A.

Nota (*It.*) A note, as, *nota buona*, an accented note; *nota cambiata*, in counterpoint, the proceeding from a discord to a concord by a skip, *e.g.*:

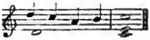

nota caratteristica, a leading or characteristic note, *nota cattiva*, an unaccented note; *nota d'abellimento*, a grace note, note of transition; *nota di passaggio*, a passing note; *nota sensibile*, the leading note or subtonic.

Nota, or Figura (*Lat.*) (1) In general, any musical sign. (2) In particular, the signs placed upon the stave which shewed by their shape and position the duration and pitch of sound. They constitute the essence of *mensurable music* (cantus mensurabilis) as opposed to mere "signs of *intonation*," such as were the neumes. The first division of notes was, as might be expected, into long and short (breve or brief), "mensurabilis musica est cantus longis brevibusque mensuratus." (Johannes de Moravia.) The long was a four-sided note with a tail (No. 2); the breve a four-sided note without a tail (No. 3); the semibreve a diamond note (No. 4).

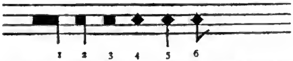

These notes varied in length (1) according to the time-signature, (2) according to the notes which preceded or followed them. (1) In *duple* measure each was equal to *two* of the next order; in *triple* measure each was equal to three. Thus, an imperfect (duple) long was equal to two breves; a perfect (triple) long to three breves, and so on. (2) A long was perfect when placed before another long; imperfect when preceded or followed by a breve, &c. The double long or *maxim* or *large* is shewn in No. 1. It was in form like a long, but had an extended head.

The introduction of the *dot* (punctus) fortunately relieved musicians of the complications above mentioned, by adding half its value to a note regardless of its position or the time-signature. In course of time six musical figures were admitted, as shown above: 1, double long; 2, long; 3, breve; 4, semibreve; 5, minim; 6, semiminim.

In the fourteenth century an extraordinary movement to create novelties in notation seems to have arisen, and treatises of that date abound with rules as to the position of the *tails* of notes; for their absolute length or proportion to each other was made to depend on this. A breve was allowed a tail to the left, a long to the right. Also, notes with tails up and down were invented, called *dragme* or *fuiscée*, and notes with two tails in the same direction These, and other similar complicated systems, never appear to have been generally received. In plain-song of course the length of notes is indeterminate, and consequently their shapes have been differently interpreted.

On the disuse of the double long (No. 1) the note No. 2 seems to have been made into the breve, No 3 into the semibreve, and No. 4 into the minim. The note No. 5 then became a *semiminima* (crotchet), and No. 6 the fusa or *unca* (hooked-note or quaver).

The notes Nos. 2, 3 and 4 were adopted by Marbecke, 1550, as his breve, semibreve, and minim, and are so named in one of the latest and best treatises on Church-song, "Les vrais principes du chant Grégorien" (Malines, 1845). But it should be remarked that in the "Traité théorique et pratique du plain-chant" (Paris, 1750). No. 2 is called the long, and No. 4 the breve. Other authorities might be cited to show that uniformity of nomenclature has never existed with regard to this subject. In the majority of modern works on Plain Song, No. 2 is the breve, No. 3 the semibreve, No. 4 the minim.

Notation, Early Systems of Musical, down to the invention of Notes.—The most ancient system known to us of expressing musical sounds upon paper, or other material employed for that purpose, is that of Alypius for the music of ancient Greece. This treatise supplies a complete method of representing the notes by letters of the Greek alphabet, the letters being sometimes upright, sometimes on their sides, and sometimes upside down, or broken in half. Unluckily, there are but three extant specimens of Greek music, in the genuineness of which any reliance can be placed, to which this system can be applied. The three are Greek hymns, of uncertain date, which were first published by Galilei in 1580, and by successive writers down to Dr. Burney; but there are evidently mistakes, as well as omissions, in the manuscripts from which the printed copies were derived, and the results are therefore unsatisfactory. The system of Alypius does not supply an adequate clue to the scales that are included in the later Greek treatise of Aristides Quintilianus. Scales are a sure test as to whether a clue be right or wrong, especially where the accompanying text gives an explanation of the intervals that should occur, as in that case. The system of Alypius has therefore become, for practical purposes, useless. It may suffice then to say, that it may be referred to in the collection of Greek writers upon music published by Meibomius at Amsterdam in 1652. If any further specimens of Greek music should be discovered, it seems more probable that the later system of musical notation included in the treatise of Aristides Quintilianus will be of avail than that of Alypius. Aristides begins with the double *omega* for *gamma*—the *g* on the lowest line of the bass clef—and carries up the scale to the extent of three octaves and a third, viz. to *b*, including every intervening semitone as well as tone. This notation will be found on the lower half of p. 27 of his treatise, which is also included in Meibomius's collection of the Greek Authors on Music. The notation may be described as a later form of that employed by Alypius. Variations between the manuscripts used by Meibomius and others are noted at pp. 243 to 245 of the same work.

One other Greek system remains to us of uncertain date. This is the ecclesiastical notation of the Greek Church. It is supposed to have originated in Greek accents, and to have been gradually enlarged into a complicated system of signs for chanting. These were written over the words without lines or spaces. Specimens of this kind of writings of various ages, may be seen in Gerbert's *De Cantu et Musica Ecclesiae* (vol. ii. p. 56 *et seq.*), and the signs are explained by Chrysante de Madyte, Archbishop of Durazzo in Illyria, in the third chapter of his Introduction to the theory and practice of ecclesiastical music,[*] published in Paris in 1821.

The Chinese have a good diatonic scale, but have now degenerated into a state of musical barbarism, and do not know how to use it. The Chinese system of musical notation is explained by Morrison in his Dictionary of the Chinese language. Chinese musicians use principally the five tones of the scale without the semitones, but they occasionally use semitones also, as is proved by the song, "Nien lai yun chue koo woo e," and by some others, printed in China about 1790. In these cases, supposing the interval between E and F to be a semitone as in our tonal system, it is to be found in passages both ascending and descending. As the fact of the use of semitones by the Chinese is denied by M. Fétis, it is well to give the title of the work from which this opposite conclusion is drawn. It is "Sin ting kew kung ta ching nan tsze kung poo," and is compiled by Chow Tseang-yŭh and others. The music and the words of the song were kindly transcribed for the writer by Mr. Robert K. Douglas, of the British Museum, who adopted Morrison's interpretation of the musical notes.

---

[*] Εἰσαγωγη εἰς τὸ θεωρητικὸν καὶ πρακτικὸν τῆς εκκλησιαστικῆς μουσικῆς (8vo.) Paris, 1821.

We next come to the Romans, who derived both their system of music and of musical notation by letters of the alphabet, from the Greeks. The Romans adopted only one of the Greek divisions of the scale, viz., the diatonic, consisting of tones and semitones, as in modern music. They had therefore a sufficient number of characters in their alphabet without breaking the letters into parts, and without turning them round about, like the Greeks.

We have no extant specimen of the music of classical Rome, and there remains but one Roman treatise on music that has any claim to completeness, and yet it is far from being complete. This is by the "ultimus Romanorum" Boethius, who was put to death by Theodoric the Goth in 525. It is entirely copied from the Greeks by a philosopher who had read Greek treatises, but who seems to have had no practical knowledge of music, and in many cases to have mistaken the meaning of Greek musical terms, even of those which express the lowest and the highest sounds.

In later ages, Boethius was claimed as a Christian, and as he wrote in Latin, instead of Greek, there was a double reason why his work should be chiefly followed by writers on Church music.*

Nevertheless, the seven letters of the alphabet, A to G, which were used as names for the seven notes of the scale, were referred back by them to Virgil, as the "septem discrimina vocum," alluded to in the Æneid (vi. 645).

In the treatise of Boethius the letters run beyond G, and up to O and P,† for the second octave, but they are intended as demonstrations of the diagrams in the fourth book, rather than as musical notes. This is proved by A being there marked to the note that would be C in music, and by the same system of illustration by letters being adopted throughout his treatise.

The mediæval system of musical notation was sometimes by letters of the alphabet, from A to G for the lowest seven notes, and from H to O for the second seven; but it was far more generally by marks or signs over the words called *neumes*, or *pneumata*, breathings for sounds. These names were taken from the Greek words, *πνεῦμα* and its plural *πνεύματα*. There are cases, but they are rare, in which the two systems of letters and neumes are combined, the letters giving greater certainty to the neumes. An early instance of this combination is found in a collection of Latin hymns, anciently used at St. Augustine's, Canterbury. The writing is of the tenth century, and the hymn is addressed to the Saint Augustin of the English (Austin the Monk). It begins: "Gemma sacerdotum! rutilans lux alma piorum Anglorum."

The compass of this hymn is nine notes, from bass C to tenor D, and the letters go up to L. (Cotton MSS., Brit. Mus. Vesp. D., vi. fol. 77.) A book was printed in facsimile in Brussels, a few years ago, entitled *Antiphonaire de Saint Grégoire*, supposed at first to be from a contemporary manuscript, and afterwards from a copy made about the year 790. The facsimile shows writing not older than the tenth century, and Father Schubiger conclusively disproves the assumed age and character of the manuscript by identifying one of the sequences included in it, "Laus tibi, Christe" (p. 62) as written by Notker, Abbot of the Monastery of Saint Gall. There is little reason for believing that the chants in any antiphonary of the time of Gregory the Great would be intelligible now, or indeed that they ever were legible without living help, for Saint Isidore, Bishop of Seville, who was a junior friend of Gregory's, and one who had long outlived him, says in the third book of his *Origines*, or Etymologies, that "unless sounds are retained in the memory, they perish, because they cannot be written."‡ This Saint Isidore wrote at the beginning of the seventh century. It seems to argue that the letters of the Roman alphabet were not in use for noting down Church music at that early date.

Neumes which had neither lines nor letters added to them were but marks or signs over the words, to serve as rough guides to the eye, whether the voice should rise or fall. They could only serve to remind the singer of chants which he had first learnt by ear. He could not tell exactly how far to ascend or to descend, and whether by tone or semitone, nor the note to begin upon. There was no

---

* This had an unfortunate effect upon Church music, and we propose to show elsewhere that at least one of its defects may be traced back to this cause.

† Our i and j count but as one letter.

‡ "Nisi enim ab homine memoria teneantur, soni pereunt, quia scribi non possunt."

Digitized by Google

precise measure of time in them. The earlier neumes were but an expansion of the system of accents and stops. At a later date, the flourish at the end of a chant, or of an *alleluia* was called the *pneuma*. These had no words to them, and were to be sung in one breath.

The next system was one without neumes, but with a ladder of letters placed one above the other at the beginning of the chants. The syllables of the words were then disconnected in order to place each syllable on a level with the letter which represented the note to which it was to be sung.

This system seems to have been invented in Flanders, by Hucbald, a monk of Saint Amand, who died at an advanced age in 930 or 932. Instead of using seven letters, he employed but four (on the Greek tetrachord system), viz., D, E, F, and G. He turned these four into four different positions, to make four disjoined tetrachords or fourths. The twisting about of the letters was probably borrowed from the Greek notation of Alypius. They were first in their ordinary position, then faced the reverse way, and then turned upside down, and faced to the left and to the right. By this means he obtained a scale of sixteen notes, and, for the seventeenth and eighteenth, to complete his number, he turned the first and second letters jacent on their faces.

It must be noted that Hucbald's scale was not what the ecclesiastical, or "Gregorian" scale is commonly supposed, and said, to have been. He made all his fourths to have the semitone between the second and third notes, as in D, E, F, G. This has been overlooked by all who have written about his notation, and Kiesewetter, in his *History of Music*, has translated Hucbald's examples without marking a single sharp or flat. Yet Hucbald's text is clear enough to any one not prepossessed with the immutability of "Gregorian" music, for he says repeatedly that his tetrachords have the same succession of intervals whether taken up or down.[*]

It is a mistake to suppose that what is called "Gregorian music" is of the age of Saint Gregory. The word means nothing more than the "use of Rome." "Nos Gregoriani et nos Ambrosiani"—"We who follow the use of Rome, and we who follow the use of Milan." As another proof that the music is changed since the tenth century, a second writer of the same age may be cited. Notker says in his *De Octo Tonis*, that *every* chant of the first and second tones ends in B; of the third and fourth in C; of the fifth and sixth in D; and of the seventh and eighth in E.[†] This differs much from the law of later times.

The "Gregorian" tones have been changed by altering the positions of the semitones in the scales. The first and second of later dates end on D; the third and fourth on E; the fifth and sixth on F; and the seventh and eighth on G. The music cannot be the same, because the intervals follow in a different succession. It is not surprising that the music should have been changed since the tenth century, the only wonder would be, if it had not.

Hucbald's tetrachord, with its semitone in the middle, became the parent of the hexachord system, or six-note scale. It consisted of a tetrachord of the same kind as Hucbald's in the middle, and a tone added at each end. It had quite the same object—that of bringing the semitone into the middle (between *mi* and *fa*) so as to make the succession of intervals the same whether the scale were taken up or down.

Hucbald's scale was as follows: Γ (gamma) A, B♭, C,—D, E, F, G,—a, b♮, c, d,—e, f♯, g, *a*,—*b*, *c*♯. His letters were placed in spaces formed between lines, which lines were designed to represent strings. In the following example, our ordinary letters are used, instead of Hucbald's oddly shaped antiquities, because they are more readily intelligible to general readers.

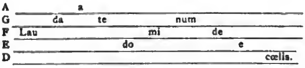

The words are: "Laudate Dominum de cœlis."

One great objection to this system was the dislocation of syllables; and another was, that it rendered necessary the doubling and trebling of the vowels when there were two or three notes to one syllable. Lastly, Hucbald's musical scale was not ruled by musical laws. He had b♮, f♯, and *c*♯ as octaves to B♭, F, and c. Surely then some licence must have been taken by the singer, for no man with ears about him could sing such octaves. These were indeed the dark ages of music. Hucbald's harmony is equally barbarous.

---

[*] See Hucbald's *Musica Enchiriadis*, in Gerbert's *Scriptores de Musica Ecclesiæ*, v. 1, p. 152. In col. 1, lines 17 and 18, "ut semper quatuor et quatuor ejusdem conditionis sese consequantur." In col. 2, lines 2 and 3, "Secundus deuteros, tono distans a proto; tertius tritos, semitonio distans a deutero." At p. 156, lines 1 to 4, "Sed dum forte in sono aliquo dubitatur qualis sit, tum, a semitoniis quibus constat semper deuterum tritumque disjungi; toni in ordine rimentur." P. 152, col. 2, "Sive sursum sive jusum sonos in ordine ducas." If it is to be the same up or down, the semitone must always be in the middle.

[†] "Ex septem bis quatuor sunt, nempe B, C, D, E, in quas omnis cantus desinit: qui primi et secundi toni est, desinit in B; qui tertii et quarti in C; qui quinti et sexti in D; qui septimi et octavi in E." Gerbert's *Scriptores*, I. 96. There were two Notkers at St. Gall, and this treatise is attributed by Gerbert to Notker Labeo (large lipped), but both were of the tenth century. The first was the abbot, and the second a monk of St. Gall.

In all countries which were in communion with the Church of Rome, the system of writing down the music of the chants, either by neumes or by letters, endured for several centuries. The great preponderance was by neumes; indeed, the use of letters was comparatively rare. The neumes were written over the words, without either lines or spaces to fix them to any pitch, or to distinguish tone from semitone. The earliest improvements upon this general practice are to be found in the Prosæ et Sequentiæ which each country produced for itself.

Much attention has of recent years been given to the question how to transcribe neumes into modern notation, but without much success, though with a vast amount of dogmatic assertion as to accuracy. Some, however, assert that they were not a musical notation at all, but only represented directions for modulation of the *reading voice*. Others go to the opposite extreme, and say that they were not only a complete system of notation, but included signs of expression, *p*, *f*, *cres.*, *dim.*, &c. When neumes were arranged in an ascending and descending position for the purpose of assisting the singer, later hands tried to draw lines over the neumes in order to arrive at their relative pitch. The absolute pitch of plain-song could not of course be determined until letters were placed on the lines; at one period *any letter* might be thus used, later only F, C, or a flat (for B♭).

The English continued to use any one of the seven letters at the signature down to the first half of the thirteenth century. At that time notes had been invented, and neumes were being gradually changed into the forms of notes.

Although early English service books are exceedingly scarce, owing to the wholesale destruction of them enforced by the severest penalties for having any in possession in the reign of Henry VIII., yet there are a few still extant. Among these are the Saint Alban's Gradual in the British Museum (MSS. Reg. 2 B. iv.), and an equally beautiful manuscript in the Cotton Collection (Caligula A. 14). These contain some of the same hymns which are included in the Winchester manuscript, and to the same music. An examination of the foreign collections of hymns, enumerated by Daniel, Mone, and G. Morel, does not reveal any of the same as having been in use abroad, neither does the use seem to have extended beyond the southern half of England, all yet discovered belonging to the province of Canterbury.

Early English advances in music are to be traced to the same cause as the early proficiency in other arts and sciences, and in Greek and Latin. It was the fortune of England, about seventy years after the conversion of the southern part of the kingdom by Saint Augustine, to have a very learned Greek as Archbishop of Canterbury, and he came to England, accompanied by an almost equally learned African, who became Abbot of Saint Peter's, afterwards called Saint Augustine's, at Canterbury.

Pope Vitalian had offered the Archbishopric to Hadrian, but Hadrian pleaded his youth and unworthiness, and recommended Theodore in his place. The Pope was doubtful whether Theodore might not introduce some of the usages peculiar to the Eastern Church, and therefore sent Hadrian to accompany him, and to keep an eye over him in that respect. Theodore remained in England till his death, viz., from 669 to 690, and Hadrian survived him till 709. These two taught the arts and sciences as well as the languages of Greece and Rome, and the gain to the nation was soon apparent through the many learned men that England speedily produced.

The venerable Bede says that some of their scholars were living in his own time, who were as well versed in Greek and Latin as in their native tongue.* Aldhelm, who was one of Hadrian's pupils, was also one of the first Englishmen who became celebrated for his skill in music, in poetry, and in other liberal arts. It is clear that the organ was introduced into England about the time of, and probably by, Theodore, from Aldhelm's full description of it in his *Laus Virginitatis*.

It was a great advantage for us to have a Greek master, for the Romans, from Boethius downwards, knew very little about music. Boethius seems only to have known, or taught, the antiquated Pythagorean division of a scale, with all major tones in it, and so false thirds. The celebrity of Greek teachers endured at least till the 11th century, for Guido d'Arezzo, in his letter *De ignoto Cantu*, thus alludes to them: "I have seen many very acute philosophers who, for the study of this art of music, have not only sought Italian, French and German masters, but even the very Greeks themselves.† Neither did the cultivation of music, once introduced, die away in England, for, in the time of St. Dunstan, who was himself an organ-builder, the Winchester organ had 400 pipes. The large number of men necessary to blow this organ (which was to be heard all over Winchester), seems to prove the bellows to have been on the Greek construction. Wolstan (or rather Wulstan) of Winchester, who describes this great organ in his metrical life of Saint Swithun, was the author of a treatise on harmony (*De Tonorum Harmonia*); which continued in use in the 13th century. William of Malmesbury describes it as a most useful book (*valde utile*), nearly a hundred and fifty years after it had been written, while St. Dunstan was the author of a still extant "Kyrie," a creditable specimen for the age, in good Greek Dorian or D minor.

---

* Bede's *Ecclesiastical History*, Book 4, cap. 2.

† "Ipsos que etiam Græcos quæsivere magistros."

Some of the Anglo-Latin hymns, written upon lines and spaces, are of a very florid character. (See, for instance, Cotton MSS., Julius A., vi. fols. 18 and 19.)

The next musical system in order of date was that of the hexachord. This endured for many centuries, and yet it seems impossible to find any thing in it to commend. It is Hucbald's system applied to the later church-scale, and with extra complexity.

The one argument in its favour adduced by old writers is the ultra-perfection of the number 6. This admirable quality is discovered through its containing within itself all its aliquot parts, viz., 3, 2, and 1, and "such perfect numbers are rare."* To all this, it might have been answered that the hexachord does not contain six equal parts; but, in any case, its perfections will weigh lightly in modern estimation against the far more ancient, the one and only true system of the octave. Eventually the hexachord system was combined with that of the octave.

The hexachord required not only the alphabetical name for the note (which sufficed by itself in the octave system) but also to tag on two or three other names. Wherever the interval between two notes was but a semitone, there was placed a *mi* for the one, and a *fa* for the other; and as there was a b flat in use, as well as a b natural in the *acute* and *super-acute* parts of the scale, so there must be a separate hexachord for each of the two, and at the interval of only one tone, the first hexachord beginning on F, and the second on G. The addition of the tone below Hucbald's tetrachord seems, by accident rather than by design, to have changed the hexachord from minor into part of a major scale, by moving the position of the semitone to the interval of one tone higher. If any one had designed such a reason for the change, there would have been some sense in the system, although it was still but an imperfect attempt to return to the octave. As it was, it may have been an accidental stepping-stone to the use of the major scale, by placing C below Hucbald's D, and thus making the intervals in that position, C, D, E, F, G, A, but this order of notes already existed in the scale. The ancient Greek scale was wholly minor, and beginning and ending on the third note of a minor scale must make a "relative" major. The sharp seventh is a comparatively modern addition to minor scales.

The following is the Hexachord system included in the octave scale:

Super-acute.	*e*†							la
	d						la	sol
	c						sol	fa
	b♮							mi
	b♭						fa	
	a					la	mi	re
Acute.	g					sol	re	ut
	f					fa	ut	
	e				la	mi		
	d			la	sol	re		
	c			sol	fa	ut		
	b♮				mi			
	b♭			fa				
	a		la	mi	re			
Grave octave.	G		sol	re	ut			
	F		fa	ut				
	E	la	mi					
	D	sol	re					
	C	fa	ut					
	B	mi						
	A	re						
	Γ	ut						

All this surplusage of names does not suffice to distinguish the notes of one octave from another. There are two E-la-mis, two F-fa-uts, two G-sol-re-uts, two A-la-mi-res, two b-fas, and three b-mis. The only notes distinguished are C and D, so there could hardly be a more complicated and useless system of nomenclature. It was taken from the initial syllables of a hymn to Saint John the Baptist, which is too short to be omitted.

Ut queant laxis   *Sol*ve polluti
*Re*sonare fibris,   *La*bia reati,
*Mi*ra gestorum   Sancte Johannes.
*Fa*muli tuorum:

Here was a *sa* for the seventh note of the scale; but, on account of the perfection of the number 6, it was not employed. In later use, in order to mark another semitone by the vowel i (as in *mi*) *sa* was turned into *si*. *Ut* was also changed to *Do* (France excepted) for the sake of openness of the vowel, although it was already included in *sol*.

Guido Aretino, or d'Arezzo, had for a long time the credit, or discredit, as some may think it, of having invented the hexachord system. He makes no claim to it in any of his works, and as M. Fétis has justly remarked, Guido's fame has rested far more upon what has been attributed to him, than upon what he really did. His reputation as a teacher was well deserved, for he directed his pupils to sing intervals without always referring to the monochord, but, instead of it, to think of similar intervals in some well-known hymns, such as this "Ut queant laxis." It appears from John of Cotton,* who wrote soon after Guido, and who was one of his great admirers, that this had been long the practice with French, Germans, and English, and even upon this particular hymn, but it was unknown to Cotton that the Italians had ever employed it for that purpose before Guido's time.†

---

* Walter Odington, *apud* De Coussemaker's *Scriptorum*, I. 215.

† This "*e*" in the super-acute part of the scale is an addition made in the 14th century.

A second system, which has been attributed to Guido, is that of the red and yellow lines for F and C. In his *Micrologus* Guido says: "In order that sounds may be discerned with certainty, we mark some lines with various colours, so that the eye may immediately distinguish a note, in whatever place it may be. For the third of the scale [C] a bright saffron line. The sixth [F], adjacent to C, is of bright vermilion, and the proximity of others to these colours will be an index to the whole. If there were neither letter nor coloured line to the neumes, it would be like having a well without a rope—the water plentiful, but of no use to those who see it."‡ This exactly describes the state of all music with neumes only, and it sufficiently accounts for all the changes that have occurred in the traditions of several centuries. Of the fact of change there can be no reasonable doubt. Much of the Gregorian music is now just as deficient in tonality as the Chinese; but a skilful organist can, by the use of anti-Gregorian harmonies, bring some of it into shape.

It will be observed that Guido does not claim to have invented the use of the red or yellow line, but to have employed them. There are specimens of both colours, among the fragments printed in Martini's *Storia della Musica*, which have been judged to be decidedly older than Guido's time, and M. Fétis (in his memoir of Guido) states that he has in his own library some more ancient still. The superior antiquity of the red line seems to be beyond question. This system was undoubtedly an improvement upon the oldest—that of having neither line nor letter to the neumes—but it was not one that would lead up to the use of alternate line and space, because the lower C, the third of the scale, was below F, the sixth, and there could be no line between them, unless lines only, and not lines and spaces, were employed. D and E only intervened between C and F, and so a single line with its space on each side would have been too much.§

Guido himself makes claim but to one invention, which he describes as "most useful" and "hitherto unheard."‖ John of Cotton gives him the credit for the same.¶ It is but a modified copy from Hucbald, utilising the spaces without employing the lines. Instead of twisting round the four letters, as Hucbald did, Guido placed the five vowels, a, e, i, o, u at the ends of the spaces; and when more notes than five were wanted, then the a, e, i, o and u were repeated. The system is shown in the following example:

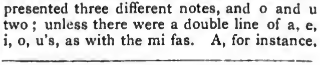

The words are: "Maria, veri solis mater, ubera tuos." This system will hardly be judged as an improvement upon that of Hucbald. The vowels were not even distinguished by capital and small letters, so that a, e, and i, each represented three different notes, and o and u two; unless there were a double line of a, e, i, o, u's, as with the mi fas. A, for instance, might be either gamma, the lowest note of the scale, the E above it, or the c above E.

---

* This author is largely quoted by foreign writers, and always as Johannes Anglicus, but his precise birthplace is not known. There are four places named Cotton in England (besides Cotton Abbots and Cotton Edmunds). They are in Suffolk, Yorkshire, Cheshire, and Shropshire. It may be conjectured that he was one of the monks driven from England in the time of William I., for, if he had remained here, his excellent treatise would surely have been quoted by some English writers, such as Odington, and he could hardly have escaped notice by Bale and others. Cotton dedicates his treatise "Domino et patri suo venerabili antistite Fulgentio"—not "*Anglorum* antistite," as printed by Gerbert. There is an imperfect copy of his treatise, written in the 12th century, in the Cotton Collection (*Vespasian*, A., xi. fol. 131). It is to be inferred that he went from one of the northern counties, in the province of York, as he does not seem to have known Canterbury use.

† Cap. i. *apud* Gerbert's *Scriptores*, 2, 232.

‡ Regulæ Musicæ Rythmicæ, apud Gerbert, ii. 30-31.

"Ut proprietas sonorum discernatur clarius,
Quasdam lineas signamus variis coloribus;
Ut quo loco quis sit tonus mox discernat oculus
Ordine tertiæ vocis splendens crocus radiat,
Sexta ejus, sed affinis, flavo rubet minio;
Et affinitas colorum reliquis indicio.
At si littera vel color neumis non intererit,
Tale erit quasi funem dum non habet puteus,
Cujus aquæ, quamvis multæ, nil prosunt videntibus."

§ Some years ago the writer exhibited several specimens of this musical notation with the red and yellow lines to the Musical Society of London. The parchments had been cut up to line the covers of old books. The chants were plagals, so the C line was always below F.

‖ "utilissimum usui, licet hactenus inauditum," *Micrologus*, Cap. 17.

¶ "Adhuc et aliam modulandi monstramus viam. pulcram sane, sed ante Guidonem inusitatam" Cap xx., Gerbert's *Scriptores*, II. 256.

Guido's invention was dying out of use about 1220, when Walter Odington wrote his treatise.[*] Odington does not in any way notice the red and yellow line system, which, if it was ever adopted by English scribes (and that seems doubtful), must soon have passed away. His own examples are upon the English plan of employing any one of the seven letters as a clef note, and of utilising four lines, and spaces, changing the clef (*clavis*) when the compass extended beyond them.[†]

There are some difficulties in fixing the exact date of Guido Aretino, or d'Arezzo. The one fact which ought to be beyond question is the evidence of Guido's cotemporary, Adam of Bremen, who says that, about 1067, Hermann, Archbishop of Bremen, called Guido, the musician, to Bremen, to correct the music, as well as the monastic discipline of his See.[‡] At the other extreme of date stands the prose version of the *Micrologus*, which is dedicated to Theobald, Bishop of Arezzo, who, according to Ughelli, died in 1037. There seems but little doubt that Cardinal Baronius (writing in the 16th century) was wrong in placing Guido the musician, under the year 1022. The very quotation that he copies supplies a sufficient correction. It is from a manuscript of the *Micrologus*, ending: "Explicit Micrologus Guidonis, suæ ætatis anno *trigesimo* quarto,[§] Johanne *Vigesimo* Romanam gubernante ecclesiam, sub quo ejusdem Guidonis librorum editio facta est." The Cardinal adds of Pope John XX.: "Successit hic Benedicto," and again: "ex cujus numeratione pariter intelligimus Joannem Benedicto successorem, *Vigesimum* ordine nuncupatum ejus nominis Romanum Pontificem." (*Annales Ecclesiæ*, anno 1022.) As Benedict VIII. was Pope in 1022, that date could not be right for either of the Johns. John XIX. succeeded Benedict VIII. in 1024, and John XX. followed Benedict IX. in 1044, and was deposed by the Council of Sutri in 1046, after which Benedict IX. was restored.[‖] The short reign of John XX., his speedy deposition, and the restoration of Benedict, have caused John to be overlooked by some chronologers, and there is a gap in the list of Johns which they supply variously. M. Fétis has adopted John XIX. as Guido's John XX., and it has seemingly been the cause that induced him to question, if not wholly to reject, the most reliable of all the evidence—that of Adam of Bremen—because Guido might be supposed to be too old to be sent for in 1067. But the date of John XX., 1044 to 1046, would restore the probability, and Baronius's error of 1022 may be accounted for by a confusion between Guido Aretino the musician, and another Guido Aretino, who became prior of the convent of Santa Cruce at Avellano in 1029. The latter has often been mistaken for Guido the musician, and M. Fétis has enumerated some of the cases.[¶]

Having touched upon the date of Guido, it is of more importance to our present subject to point out some anachronisms in the copy of the *Micrologus*, printed by the Abbé Gerbert. A manuscript of the earliest date should have been selected for publication, for, in so popular a work, it might have been expected that the notation of the examples would be adapted by the copyist to the use of his own age. Those at pages 10 and 12 of Gerbert's edition have been thus modernised. They are printed on lines and spaces, which Guido did not use together, and a comparison of the whole text with that of one of the manuscripts in the British Museum,[**] proves that large additions have been interpolated to Guido's text in Gerbert's edition, not only in the *Micrologus*, but also more largely in the *De ignoto Cantu*.

The transition from neumes to notes was an easy one. It was called "quadrating" them, *i.e.*, squaring them like the old shape of the natural B. They thus filled the spaces between the lines. But although such changes were easy, they were practically slow, because old habits and prepossessions worked against them.

---

[*] "Sed talis modus componendi jam evanescit." See Coussemaker's *Scriptorum*, I. 217.

[†] See "De Clavibus," p. 214. "Unam dictarum clavium semper una linearum lineatur, aliter spaciatur; linearis pro eo quod lineæ applicatur; spatialis pro eo quod spatium occupat, ut hic." This most useful treatise has been recently printed by M. E de Coussemaker, in his *Scriptorum de Musica medii Ævi nova Series a Gerbertina alterum*. The only writer unquestionably English in Gerbert's Series is John of Cotton. M. de Coussemaker has added John Hothby (called Ottebi by Italians), Walter Odington, John of Garland, John of Tewkesbury, Simon Tunstede, Robert de Handlo, John Hanboys, Theinred of Dover, and the works of some anonymous writers, together with a greatly increased number of good foreign authors. The limited number of copies printed makes the speedy acquisition of this collection desirable to all those who are interested in the subject.

[‡] "Musicum Guidonem Hermannus, Archiepiscopus, Bremam adduxit, cujus industria melodiam et claustralem disciplinam correxit." *Historia Ecclesiarum Hamburgensis et Bremensis*, &c., *ab anno* 788 *ad* 1072, lib. 2, cap. 102, p. 30.

[§] M. Fétis has printed it 24 instead of 34 in his Memoir of Guido. *Biographie Universelle des Musiciens*, v. 4, p. 147, col. 1.

[‖] "Silvestre III., antipape, étoit évêque de Sabine, lorsqu'après avoir chassé Benoit IX., en 1044, les Romains le mirent à sa place. Trois mois après, Benoit fut retabli par sa famille: son rival forcé de céder, vendit sa dignité à un prêtre, qu'il sacra sous le nom de Jean XX., mais resta lui même à Rome, et conserva ses partisans. Le conseil de Sutri, 1046, les déclara tous les deux usurpateurs—Voyez Benoit IX. et Gregoire VI." (*Dictionnaire Général de Biographie et d'Histoire*, par Ch. Dezobry et Th. Bachelet, Paris. 8vo, 1857.)

[¶] *Biographie Universelle*, v. 4, p. 148, col. 2.

[**] Addit. MSS., No. 10,335, eleventh century

The best explanation of neumes is that of M. de Coussemaker, in his *Histoire de l'Harmonie au Moyen Age*, where two tables of neumes with their interpretations will be found at p. 184. Whoever shall desire to modernise neumes will do well to consider first the country and the age in which the manuscript was made. For instance, the *Quilisma* may be either a mark for a tremulous expression without altering the note,* or it may be a sign of graduated ascent of two or three notes.† The first is the earlier use (and just what its form would seem to have been invented to express) and the second is the later use, when neumes were soon to be altered into notes. The reader will find many pages of fac-similes of neumes in M. de Coussemaker's great work, and some in Gerbert's *De Cantu et Musica Ecclesiae*, v. 2, about and after p. 60. A third source for the student is at the end of the so-called *Antiphonaire de Saint Grégoire*, which he should bear in mind to be only a fac-simile of a manuscript of the tenth century. The names of the neumes transformed into notes will be found in Walter Odington's treatise (p. 213), and in some other treatises in M. de Coussemaker's collection.

It was not until notes began to take the place of neumes that it became possible to write *mensurable* music with any probability of the value of the notes being reproduced in performance according to the intention of the composer. The first writer on mensurable music was Franco of Cologne, who is cited by nearly all subsequent theorists as the earliest authority on this branch of the art. His treatise, entitled *Ars Cantus Mensurabilis*, has been printed by both Gerbert and Coussemaker, and was probably written shortly before A.D. 1200. The system of notation which he lays down is, in substance, that which was universally employed up to the latter part of the sixteenth century, and which our Ravenscroft, in his *Briefe Discourse*, tried to perpetuate as late as 1614.

Franco knows only four notes, the double long or large, the long, the breve, and the semibreve. The double long is equal in value to two longs. The long may be either perfect, *i.e.*, equal to three breves; or imperfect, *i.e.*, equal to two breves. The breve may be either "recta," *i.e.*, equal to one-third of a long; or, "altera," *i.e.*, equal to two-thirds of a long. The semibreve may, in the same way, be either minor or major in relation to the breve. A long followed by a long is always perfect; a long followed by a single breve is made imperfect unless there is placed between them "quidam tractulus qui signum perfectionis dicitur, qui etiam alio nomine divisio modi appellatur" (see "Punctus divisionis"), in which case the long remains perfect, and the breve makes imperfect the long which follows it. If a long be followed by two breves and a long, the first long will be perfect, the first breve "recta," the second breve "altera," and the second long perfect, unless a "tractulus divisionis" be placed between the two breves, in which case both the longs will be imperfect and both the breves "rectæ."

If a long be followed by three breves, the long will be perfect, and each of the three breves "recta," unless the "tractulus divisionis" be placed after the first breve, in which case the long will be imperfect, the first two breves "rectæ," and the third "altera," and so on.

These rules as to the value of longs and breves apply equally to breves followed by breves or by one or more semibreves, and when notes of lesser value were introduced in the fourteenth century, they were also applied to semibreves and minims, though, curiously enough, a minim was never allowed to contain more than two semiminims. This system, by which notes of the same form could be either single or double in value ("recta" or "altera"), is based upon the triple measure which so dominated early music that it was considered impossible for two notes of the same form occurring together to be equal in value; they had therefore to be wrested into triple measure by doubling the value of one or other of them, and it was the second of the two that was always selected for this doubling or "alteratio." Neither in Franco's treatise nor in that of Walter of Oddington (written about 1220) is duple measure admitted for any note of less value than a long, and duple measure seems to have been successfully excluded from mensurable music for at least a century after Franco.

Franco then describes the nature and forms of "plicæ" (see "Plica"), a form of notation which was freely used till the beginning of the fifteenth century, when it seems suddenly to have dropped out of use.

---

* [Quilisma est] "unisonum quia non habet arsim et thesim" [*i.e.*, neither rise nor fall] "nec, per consequens, intervallum vel distanciam, sed est *vox tremula:* sicut est sonus flatus tubæ vel cornu, et designatur per neuman quæ vocatur Quilisma" (B. Engelbertus, per Gerbert, *De Cantu*, v. 2, p. 60).

† "Hæ Antiphonæ, licet a finali incipiant, tamen, quia per Quilismata, quæ nos *gradatas neumas* dicimus, magis gutturis." (Bernonis Augiensis *Tonarius*, Gerb. *Scrip.*, v. 2, p. 80.)

Digitized by Google

He then deals with ligatures, which he defines as "conjunctiones figurarum simplicium per tractus debitos ordinatæ"; they consist of square notes—the lozenge-shaped semibreve is never admitted in ligature—with or without tails, connected together either at their angles or in a transverse bar, thus:

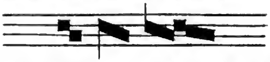

the bar, however long, representing two notes only, viz., that on which it begins and that on which it ends. The scholasticism of the age shows itself in the "distinctions" or "differences" which were said to belong to ligatures. They are three in number:

1. A ligature is either "ascendens" or "descendens," according as the second note is above or below the first.
2. A ligature may be either "cum proprietate," "sine proprietate," or "cum opposita proprietate." A descending ligature is "cum proprietate" when the first note has a tail downwards on the left, as in plain-chant; it is "sine proprietate" when the first note has no tail. An ascending ligature, on the other hand, is "cum proprietate" when the first note has no tail, and is "sine proprietate" when the first note has a descending tail on the left or on the right ("quod magis proprium est"). Both ascending and descending ligatures are "cum opposita proprietate" when the first note has a tail ascending on the left.
3. A ligature is "cum perfectione" when the last note is square and is placed immediately above the penultimate, thus: 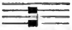; or, when the last note is square and is placed below the penultimate "averso capite," thus: 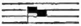; if the last note is "oblique," *i.e.*, forms part of a transverse bar, or if it is placed above the penultimate "averso capite," thus: 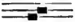 the ligature is "sine perfectione."

The rules for determining the value of notes in ligature are as follows:—

The first note of every ligature "cum proprietate" is a breve, of every ligature "sine proprietate" a long. The last note of every ligature "cum perfectione" is a long, of every ligature "sine perfectione" a breve.

The first *two* notes of every ligature "cum opposita proprietate" are semibreves, and, subject to this last rule, every note in a ligature other than the first and last is a breve.

With two slight exceptions, these rules for ligatures given by Franco continued in use till the sixteenth century, as long, in fact, as his system of notation lasted. The two changes were (1) that the first note of an ascending ligature, if meant for a long, always had the tail descending on the *right*. (2) That, owing to difficulty in writing the ligature 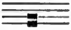 when the two notes were contiguous, the form 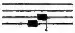 became more general, though the former was still, and is now, used in plain-song. The rules therefore for determining the value of ligatures may be stated as follows:—

1. Every note in a ligature having a tail downwards on the right is a long.
2. If a note in ligature have a tail upwards on the left, that *and the succeeding note* are semibreves.
3. Subject to the above rules, every note in ligature, other than the first or last, is a breve.
4. The first note, being without a tail, is a breve if the next note ascends, but a long if the next note descends.
5. The first note, having a tail downwards on the left, is always a breve.
6. The last note, being square and without a tail, is a breve if the preceding note is below it, but a long if the preceding note is above it.
7. The last note, being oblique and without a tail, is always a breve.

The rests employed were—

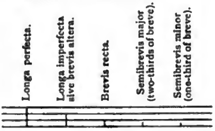
Longa perfecta.
Longa imperfecta sive brevis altera.
Brevis recta.
Semibrevis major (two-thirds of breve).
Semibrevis minor (one-third of breve).

When the minim was introduced in the fourteenth century, the rests used were—

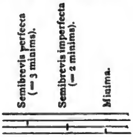
Semibrevis perfecta (= 3 minims).
Semibrevis imperfecta (= 2 minims).
Minima.

Early in the fifteenth century the second of these semibreve rests fell into disuse, and from that date the rests are the same as those now employed, except that they are drawn vertically instead of horizontally.

The accidentals employed were two only, viz., the ♭ rotundum, which was used not only for flattening the notes B, E, and A, but to prevent the sharpening of other notes, more particularly F, C, and G, where we should now use a natural. The other accidental, originally the ♮ quadratum, assumed various forms, such as ♮, ⊔, ♯, and was used both as a sharp, with F, C, and G, and as a natural, with B, E, and A.

The repeat mark in use consisted of one or more vertical lines placed between pairs of dots, thus: 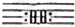; the number of lines indicated the number of times the preceding section was to be sung or played; in the above example, for instance, the piece would be sung through four times, *i.e.*, *repeated* three times only.

Notazione musicale (*It.*) The system of musical notation.

Note. A sign of a sound made of various shapes to denote relative duration. Hence, the term is used generally for the sounds of which notes are signs, as when we say of a singer that his high *notes* are good, or that a player plays wrong *notes*. [Nota.]

Note (*Fr.*) A note, as, *note dièsée*, a sharpened note; *note d'agrément*, a grace note, note of transition; *note de passage*, a passing-note; *note sensible*, the leading note or subtonic; *notes de goût*, notes of embellishment.

Notturno (*It.*) Originally, a kind of serenade; now a piece of music of a gentle and quiet character.

Notula (*Lat.*) The notes used in writing ligatures. [Ligature.]

Noursingh. An Indian horn or trumpet, formed of a straight metallic tube. It is supported in a horizontal position by means of a long stick.

Novemole. [Nonuplet.]

Nuances (*Fr.*) Shades of musical expression.

Nuovo, di (*It.*) Again.

Nut. The fixed bridge formed by a slight prominence or ridge at the upper end of the strings of instruments of the violin and guitar family.

Digitized by Google

O.

O. A circle, formerly the sign of Tempus perfectum, as the incomplete circle C was of Tempus imperfectum. [See Tempus, Modus and Prolatio.]

O (*It.*) Or; as, *flauto o violino*, flute or violin.

Oaten-pipe. The simplest form of a reed pipe, a straw with a strip cut to form the reed, at the end closed by the knot

Ob. Abb: of *Oboe*, also of *obbligato*.

Obbligato (*It.*) An instrumental part or accompaniment of such importance that it cannot be dispensed with.

Ober (*Ger.*) Over or upper, as, *Obermanual*, or *Oberwerk*, the upper manual; *Oberstimme*, and *Obertheil*, upper part.

Obligé (*Fr.*) [Obbligato.]

Oblique motion. When one part moves and the other remains stationary. [Motion.]

Oboe. One of the most ancient, as also one of the most charming instruments of music. In some of the Egyptian pipes in the British Museum were found pieces of thin wood or straw inserted into the tubes in such a manner as to suggest at once the similarity between them and the oboe. In all probability the Greeks used instruments of this class, although they called them by the general name of "flute." That the reeds used were very small, and of more perishable material than the tubes into which they were inserted, is quite sufficient to account for the fact that we have not discovered a pipe with a reed fixed in it ready for use, among Greek antiquities. Moreover, it must be remembered that musicians do and always did carry their reeds and mouthpieces about, separated from the instrument. The word γλωσσόκομον proves that Greek players were not exceptions to this rule. Almost all European nations have some form or other of the oboe in use, though the different kinds vary much in their merits as musical instruments, some being in the simplest form of a reed-pipe played by mountaineers, as the *chalumeau* (from *calamus*, a reed), which is still played by the peasantry in the Tyrol, and the *piffero pastorale* of the Italians, a similar instrument to the *chalumeau*, called *Schalmey* by the Germans, and formerly *shalm* or *shawm* in England. The immediate forerunner of the modern oboe was the bombardino, or little bombardo, called by the Italians *bombardo piccolo*, a kind of *chalumeau*. The date of the introduction of the oboe is about 1720, but from the mention in Bach's Passion-music of two kinds of oboe, the oboe d'amore and oboe da caccia, it is evident that well-known varieties were in use in his time. The oboe d'amore, which was also called *oboe luonga*, produced a delicate and sweet tone, while the *oboe da caccia* corresponded to the tenoroon oboe, or corno inglese. The latter, though not in common use, is occasionally introduced into the scores of modern operas, as in Halévy's *Jewess*, Meyerbeer's *Huguenots*, &c.

The oboe now in use owes its large compass, like many other wind instruments, to the overblowing of the player, which brings into existence a set of *overtones*. The overtones of the oboe are similar to those of an open pipe, that is, they are represented by the numbers 1, 2, 3, 4, &c., while on the other hand, the overtones of a clarinet correspond to those of a stopped pipe as represented by 1, 3, 5, 7, &c. The notes lying just between the sounds produced by the natural length of the pipe (as shortened also by finger-holes and keys), and those produced by an altered pressure of wind from the player's mouth, always present difficulties of execution when required to be played in rapid succession. The clarinet has a single reed; the oboe, a double one. The extreme compass of the oboe now in use is two octaves and one fifth, namely:

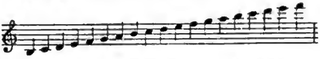

with the intermediate semitones, subject to certain limitations in their use. Some instruments have the low B♭

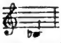

but it is not yet commonly met with. The compass given above should be very much

Digitized by Google

curtailed for general use, the four or five lower notes being weak and thin, the two or three upper notes harsh and shrill. The latter, however, are not objectionable in *fortissimo* passages in *tutti* parts. The following shakes, lying between intervals played with difficulty, are given by Berlioz (Instrumentation) as difficult:

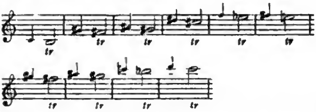

The following are impossible (and all tonal or semitonal shakes above this):

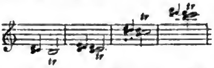

The oboe is more effective, because more easily played, in simple keys, than in keys with many sharps or flats; and in simple passages, than such as are complicated or rapid. As an instrument of a pastoral character it can be cheerful or plaintive; it possesses also in its peculiar quality of tone, the wail of grief or agitation, but it has also the power of soothing and calming the mind after disturbing influences.

Oboe. An organ stop consisting of reed pipes, slightly conical and surmounted by a bell and cap, of 8ft. pitch. The tone is thin and soft. An *orchestral* oboe is an organ stop intended as a more perfect imitation of the orchestral instrument than that ordinarily made.

Oboe da caccia. The "hunting" oboe, a large kind of oboe. [Oboe.]

Oboe d'amore. The oboe "of love." A small oboe. [Oboe.]

Oboist. A player on the oboe.

Oboista (*It.*) An oboist.

Ocarine (*It.*) A series of seven musical instruments made of terra cotta pierced with small holes, invented by a company of performers calling themselves the Mountaineers of the Apennines. With these instruments, which are of a soft and sweet, yet "travelling" quality of tone, operatic melodies with simply harmonised accompaniments were given.

Octave. (1) The interval of an eighth. It may be major, minor, or augmented, *e.g.*:

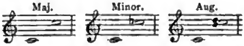

It was the *diapason* of the Greek system. (2) The first note of the harmonic scale. (3) An organ stop of 4ft. pitch on the manuals, or 8ft. on the pedals. (4) The eight days following a great festival of the Church.

Octiphonium (*Lat.*) [Ottetto.]

Octochord. An instrument with eight strings.

Octuor (*Fr.*) [Ottetto.]

Octuplet. A group of eight notes which are to be played in the time of six.

Odeon. ᾠδεῖον (*Gk.*), **Odeum** (*Lat.*) A building in which odes or other works could be performed for public approval or private rivalry.

Œuvre (*Fr.*) Opus or work; as, *œuvre première*, Opus 1; *chef d'œuvre*, a masterpiece.

Offen (*Ger.*) Open; as, *Offenflöte*, open flute, an organ flute-stop. [Flute.]

Offertoire (*Fr.*), **Offertorio** (*It.*), **Offertorium** (*Lat.*) [Offertory.]

Offertory. (1) The collection of alms during the celebration of Holy Communion. (2) A piece of music performed during the Offertory.

Oficleida (*It.*) [Ophicleide.]

Ohne (*Ger.*) Without; as, *ohne Ped.*, without Pedals; *ohne Begleitungen* without accompaniments.

Oioueae. The vowels of "World without end, Amen," an imitation of the *Evovae*, the vowels of "seculorum amen," used to designate the ending of a mode. [Evovae.]

Oliphant. The name of an obsolete species of Horn, so-called because it was made of ivory (olifaunt, olivant, olyfaunce, being old forms of the word elephant). Three specimens of this instrument are in Kensington Museum, two of the eleventh century, one of the fifteenth.

Ombi. A harp used by negroes in Western Africa, the strings of which are made of fibrous root or creepers. It bears a strong resemblance to the Oriental harp.

Omnes, omnia (*Lat.*) Chorus or Tutti.

Omquäd. The name of the refrain of some old Danish ballads, called the "Kämpe Viser."

Onágon. A Chippawa drum.

Ondeggiamento, ondeggiante (*It.*) With an undulating, or quivering sound; making a tremolo.

Onduliren (*Ger.*) To make a tremolo, or produce an undulating tone.

Ondulé (*Fr.*) Undulating. [Ondeggiamento.]

'Ood or **Oud.** An Egyptian instrument of a similar character to the guitar, *q. v.* It is flat on the upper surface and convex at the back. The neck is like that of the guitar but without frets, and there are seven pairs of strings, each pair tuned in unison. The player uses a plectrum formed of a strip of vulture's quill. [Lute.]

Open Diapason. The name in England of the chief open foundation stop of an

Digitized by Google

organ, called by Germans "Principal" or "Prestant," and by French "Principal," or if in front, "Montre." On the manuals it is of metal, on the pedal organ of metal or wood. The metal pipes of this stop are cylindrical, and of the best material. A great number of foreign organs, and a few in England, contain open diapasons made of pure tin (Organ, § 13). It is, however, more commonly made of spotted metal, or of an equal mixture of tin and lead. If two or more open diapasons are allotted to one row of keys, they are made of different scales, in order that they may not destroy each other. One is then sometimes called large, another small. This stop is of 8ft. length on the manuals, and 16ft. on the pedals, unless stated to the contrary on the register.

Open Harmony. Chords formed by as equidistant a disposition of the parts as possible, *e.g.*:

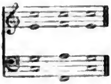

as opposed to

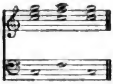

Open Notes. Of stringed instruments,—the notes of the open strings (*q. v.*) Of wind instruments, such as the horn, trumpet, &c., the series of natural harmonies which can be produced by the *lip* of the performer without the assistance of a slide, key, or piston.

Open Pipe. A pipe open at the top, as opposed to one closed at the top. The pitch of a closed pipe is approximately one octave lower than that of an open pipe of the same length. [Organ, § 14.]

Open Score. When each part has a separate line assigned to it, music is said to be *in open score*. When more than one part is written in each line, in *close* or *short* score.

Open Strings. Strings producing the sounds assigned to them according to the "accordatura," or system of tuning belonging to the particular instrument. Strings are said to be *stopped* when their pitch is altered by the pressure of the finger.

Opera. A dramatic entertainment, of which music forms an essential and not merely an accessory part. Acting and music have been associated in some way or other from time immemorial, and if our Opera is looked upon as nothing more than a rather superior kind of mixed acting and music, it may be called one of the oldest entertainments in the world. But the modern Opera has so many distinctive features which it has certainly not inherited from its ancestry, that it is unwise to definitely ascribe its origin to the Greek play, or the mediæval mystery, or even to those rustic plays with music which enlivened market days and fairs, although these last are in close contact with our present conception of an Opera. The birth of modern Opera took place just about the year 1600, and its birthplace was Florence, where some enthusiastic devotees of what might then have well been called the Renaissance *cult* were hoping to recover, or at least find a musical substitute for the lost art of declamation as used in the Greek tragedy. Vincenzo Galilei (father of the astronomer), Caccini, Peri, Strozzi, Renuccini, and others used to meet at the Palace of Giovanni Bardi, Count of Vernio. The advent of this style necessarily involved the abandonment of the devices, character, and very essence of the vocal polyphonic school, and yet some historians roundly abuse the founders of the operatic school, the monodists, for casting aside their allegiance to the past. But the new vocal style, called *Lo stilo rappresentativo*, or *Musica parlante*, also compelled composers to throw fresh responsibilities on musical instruments; from the former grew the recitative, and to the latter we owe that development of the orchestra which now gives it such an all-important share in dramatic representation. A dramatic piece, "Dafne," by Peri, was privately performed in the Palazzo Corsi in 1597, but the first public performance of an Opera took place at Florence in December, 1600, when Peri's "Euridice" was produced on the occasion of the marriage of Henry IV. of France with Maria de Medici. Cavaliere and Caccini deserve to share with Peri the honour of being the earliest writers of Opera. In 1607 Monteverde produced his "Arianne" to a libretto by Rinuccini, Gagliano at the same time, setting Rinuccini's "Dafne." But Monteverde's success eclipsed that of all his contemporaries, and his bold genius was destined to guide the art into fields hitherto unknown, and provide it with resources which have only been fully developed in our own day.

The following extract, from the fragmentary remains of "Arianne" which have survived, is the Lament which is said to have moved the audience to tears. It certainly is capable of the tenderest expression, and having been written only ten years after this style was first attempted, it proves how rapidly and fully Monteverde grasped its power. It should be noticed that this specimen is moulded in a form which was afterwards to become of permanent utility; the first portion consists of six bars, the second of eight, and the third is a repetition of the first. The treble and bass are from the original, the slight harmonies have been added to assist the reader in forming some idea of the actual manner of its performance when the bass and treble were filled up.

IL LAMENTO D'ARIANNA.

MONTEVERDE, 1607.

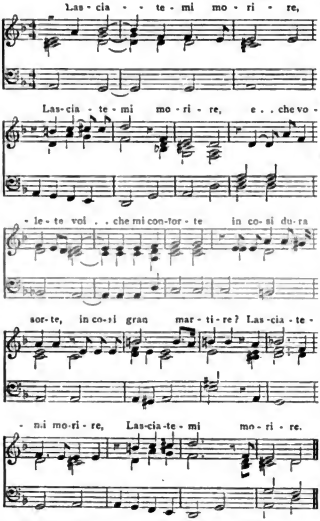
Las - cia - - te - mi mo - ri - re,
Las-cia - te - mi mo - ri - re, e . . che vo -
- le - te voi . . che mi con-for - te in co- si du-ra
sor- te, in co-si gran mar - ti- re? Las-cia - te -
- mi mo-ri - re, Las-cia-te - mi mo - ri - re.

Monteverde, in 1608, wrote "Orfeo," the opening symphony of which, a *toccata* followed by a *ritornello*, has often been quoted as the earliest specimen of an operatic overture. The chief successors of Monteverde were his pupil, Cavalli (who, between 1639 and 1665, wrote no less than thirty-four Operas), Cesti (a pupil of Carissimi), Legrenzi (who, between 1664 and 1684, wrote seventeen Operas). But a large supply must have been required by the eleven Opera-houses said then to exist in Venice.

Rome opened its first Opera-house in 1671, and very rapidly the love of Opera spread to Naples, Bologna, Padua, Vienna, Dresden, and Paris. The name which stands prominently forward in this period of art was that of Alessandro Scarlatti. He showed that the old declamation, or *Recitativo secco*, could be modified into the accompanied recitative, *Recitativo stromentato*, and could thereby become a more powerful expression of emotion. To Scarlatti also was due the finished form of the *Aria*—that is, a first portion closing in the key and generally with a symphony, followed by a second portion in a closely related key (relative minor, or relative major, or dominant, sometimes actually in same key as first portion, but containing some central modulation), at the close of which (often a half-close) *Da capo al fine* is marked, and the repetition of the first portion brings the *Aria* to a close. Previous writers of Operas had, perhaps unconsciously, made essays in the direction of this form, but had not given it a definite shape. Although it may be found in the earliest popular melodies of many nations, its permanent function in the Opera gave it an influence which eventually introduced it into pure instrumental music, and for "slow movements" and separate pieces it has been constantly used by composers from Scarlatti's time to the present day. The first of Scarlatti's Operas was performed in Rome in 1680; between that date and his death, in 1725, he composed 108 more! His most distinguished contemporaries were Stradella, Rossi, Caldara, Lotti. Soon after this period there began a decadence of the Opera in Italy. In the list of cities in which Opera took root we just now named Paris, and it is necessary to go back to the year 1646, when a boy, aged thirteen years, by name Jean Baptiste Lulli, was brought to Paris by the Chevalier de Guise, and was employed in the kitchen by Mademoiselle de Montpensier, niece of Louis XIV. This is not the place to tell the strange story of his life, but when dismissed from his post in the kitchen, his musical ability secured him a place in the king's band and eventually gave him entire control over a small band, "les petits violons." Commencing as a composer of ballets, which were then much in vogue at Court, he afterwards gradually developed a distinct school of Opera. No doubt he was wise enough to study the Italian school, then in its vigour; he had also before him the efforts of many previous Frenchmen—the Abbé Mailly, who wrote the Opera "Akebar, Roi de Mogul," and Cambert, who has the honour of being the composer of the first French Opera ever performed in Paris, "Pomone," played in 1671 at the Hôtel de Nevers. In 1672 Lulli, whose influence at Court was always paramount, obtained a Royal Patent giving him entire monopoly of the Académie de Musique, which had been three years previously (1669) founded for the public presentation of "operas and dramas with music, and in French verse," and had been granted to Perrin, Cambert, and the Marquis de Sourdéac. This was a sad blow to Cambert, fresh from his success as composer of "Pomone." He left France for

London, was kindly received by Charles II., but soon after died of disappointment and grief, aged 49. The removal of Lulli's only dangerous rival left him in absolute control over the French Opera. Amongst the improvements he introduced may be named first the creation of the Overture-Form which bears his name. He discarded the previous short prelude, and in its place introduced a slow Introduction, a Fugato movement, and (often) a Minuet or other dance tune. He culled and used all that was best in the accompanied recitative recently introduced by Cavalli and others; he caught the spirit of Italian melody and rhythm, and showed the utmost skill in the manner he dealt with the text and prosody of his poetry. If to this be added his complete mastery over stage requirements, scenic effects, and the use of the chorus, it will be understood that Lulli's equipment for founding a School could not have been surpassed.

Till the time of his death, in 1687, Lulli not only retained his hold on the Académie de Musique, but he produced, with Quinault as librettist, twenty grand operas, not counting smaller works. Anyone who takes the trouble to study Lulli's music must be struck with the richness of his invention and the sincerity of his effort to make music the interpreter of his words. His successors were not men of genius, although prolific writers; but it must be allowed that some of them enlarged the scope of the orchestral element. Among them were Charpentier (1634—1702), Campra (1660—1744), Monteclair (1666—1737), and others. The last-named introduced the three-string double bass into the orchestra. The next man of note and influence was Rameau, who, though about thirty-eight years of age when he arrived in Paris in 1721, raised the standard of Opera to a very high level. He wrote numerous Grand Operas in which the use of the orchestra was still further expanded. During ten years after his death no work of first importance was produced; then, in 1774, a new era was in store for French Opera, Gluck produced his "Iphigénie en Aulide," and the arrival soon after of Piccini brought about that struggle between realism and conventionalism which was to end in the victory of the former. We next find Gossac, Spontini, Kreutzer, and the renowned Cherubini, ultimately followed by Rossini, Auber, Halévy, Meyerbeer, Hérold, and, in our own days, Gounod, Thomas, Massenet, and Saint-Saëns.

But England had not been altogether idle during this long period, Purcell, born about 1658, composed, when a mere lad, the opera "Dido and Æneas," and during his short, but brilliant career, produced a long series of Operas, Plays with Incidental Music, Overtures, &c. But, unfortunately, the English Opera had neither the ancestry nor entourage of our French neighbours; it was a child of the popular entertainment, the masque, and seems never to have attempted to utilise all the experience and improvements going on around. Even the genius of Purcell, his beautiful melody, natural dramatic force, and boldness in harmonic progression, were unable to found a permanent English Opera; and after his time even Handel found it a difficult task to establish Italian opera in London. How different the influence of Handel on English music might have been had he carried on English Opera! And after Handel's time the so-called English Opera never rose above a drama with incidental songs, choruses, &c. The first Operas heard in Germany were Rinuccini's "Dafne" in 1627, "Ferrari's "L'Inganno d'Amore" in 1653, all sung to Italian words. Later on Italian Operas were given in Munich and in Vienna. But German Opera had its birthplace in Hamburg, where "Adam und Eva," a Songspiel by Johann Theile, was publicly performed in 1678. The successors of Theile were Strunck, Franck, Förtsch, Conradi, Mattheson, and others. But the most distinguished among them was Keiser (1673—1739), whose distinctively German thought and method found their finest exponents in Mozart (1756—1791) and Weber (1786—1826). As their works are still part of every Opera repertory, it is unnecessary to speak of them here. We are now reaching the period when Wagner (1813—1883) appears on the scene, bringing new principles, new conceptions, new methods, which have placed Germany in the front rank of Opera-producing nations. His talents expanded with his experience. It is too early to say what his influence may be; but Italy, after falling to a low ebb, as exemplified in the works of Bellini, Donizetti, and a host of minor lights, has, in Verdi, Boïto, and others, found men worthy of accepting the work of the Operatic Reformation, and it may even be said that some of the later French composers have not despised the lessons taught by Wagner.

But a result has followed upon the modern method of writing Opera which deserves notice. In proportion as the music is made absolutely suitable to the perpetually changing character of the dialogue, soliloquy, or libretto generally, so do "set-pieces" become of less common occurrence; and therefore the performance of the music becomes more confined and limited to the opera-house itself. When an Opera was, as formerly, a

Digitized by Google

series of separate movements separated by very conventional recitative, these movements, songs or otherwise, found their way, as "arrangements" of various kinds, into every musical home, into the repertory of every itinerant band, and on to the barrels of innumerable barrel-organs and mechanical pianofortes. With the exception of a few Songs and Marches there is but little of Wagner which is suitable for home consumption or street-desecration.

The Overture of an opera, whether founded on themes afterwards occurring in the opera or not, is almost without exception thrown into modern symphonic form, and is therefore capable of separate performance without injury to its effect. Interludes or *entr'actes* are often performed separately. The composer Mascagni can hardly be gratified when hearing the *entr'acte* from his "Cavalleria Rusticana" set as a song to the words "Ave Maria!" Though Bishop, Balfe, Wallace, and other English composers have left operas of a very melodious type which are yet very popular, we have not as yet found a native composer capable of taking a high rank in this branch of music. If the much-loved Handel had started an English instead of an Italian Opera in London, it might have had good results, but in this, as in many other respects, his influence on English music was far from beneficial. His misdirected influence is the more to be regretted because England made a good start in this direction. Some of Purcell's highest flights of genius are to be found in his Operas; and so fresh is his masterly touch that when, in 1896, his "Dido and Eneas" was revived by students of the Royal College of Music, a critical audience was greatly impressed by its beauty.

Operist (*Ger.*) An opera singer.

Ophicleide. A keyed serpent. (ὄφις and κλεὶς). A brass instrument of large compass and great power, but having so peculiar a quality of tone as to necessitate careful treatment by a composer. There are two sorts of ophicleides, alto and bass. The bass instrument has a compass of three octaves and one note, from

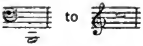

with all the intermediate semitones. Bass ophicleides in B♭ produce, of course, sounds one whole tone below those written, and those in A♭, sounds a major third below those written. The ophicleide is capable of a great deal of execution, if not too chromatic and too low in its compass. The alto ophicleide has a compass similar in extent to that of the bass instrument, but starting from a note one octave higher, that is from

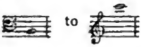

But as the alto ophicleide is in F or E♭, the real sounds produced will be a major fifth or major sixth lower than the notes written. Three or four of the lowest notes and two or three of the highest notes should be avoided. This instrument is not so satisfactory as its bass fellow, and is therefore but rarely heard. The quality of the tone is not unpleasant, but it does not combine well with that of other members of the orchestra. This unpleasant prominence of character tends to limit the use of the instrument.

Opus (*Lat.*) A work. Composers number their works for purposes of distinction and reference.

Opusculum (*Lat.*) A short work.

Oratorio (*Ital.*) A composition for voices and instruments illustrating some subject taken directly from scripture or paraphrased upon some theme in sacred history. The music consists of symphonies or overtures, airs, recitatives, duets, trios, choruses, &c., with accompaniments for orchestra or organ.

The origin and growth of oratorio is almost coeval with that of opera, both being developments of the early forms of drama, or, to speak more clearly, oratorio may be said to be the successor of the mediæval Mystery play, as was the opera of the Morality. In both cases the poetical description of the subjects chosen is accompanied with music, but with this difference, that whereas the opera requires also scenery and dramatic action in its representation, oratorio is now performed without either.

The soldiers and pilgrims of the first crusade, in their attempts to delineate dramatically the Passion of our Lord, and other incidents of sacred history, together with the legends of the deeds and endurance of saints and martyrs, imitated the practice observed in the performance of secular plays by the introduction of music at these representations. It is probable that more than one of the hymns written by St. Bernard of Morlaix and others living about the time were contributed for this purpose, and that the audience joined in the singing. As many of the first oratorios were mystical expositions of doctrine such as that described in many of the early hymns, some colour is given to the supposition that musical exposition and teaching in the form known as oratorio may be as old as the time of the Crusades. It was not until five centuries later that it was made a recognised and dis-

Digitized by Google

tinct medium of instruction and pleasure. St. Philip Neri, in the endeavour to establish firmly the institution he had founded in Rome, and also moved by a desire to win the people to the observance of religious duties, relied upon the universal love for music to gain his object. With this intention he engaged Giovanni Animuccia, a Florentine, as his Maestro di Cappella, who during his connection with the oratory produced several musical pieces, consisting of motetts, psalms, and songs of praise set to Latin and Italian words, known by the general title of "Laudi Spirituali." Animuccia published his first collection at Rome in 1563. The character of these compositions is similar to that known as the madrigal style, and these "laudi" have been taken as the origin of that class of composition. Animuccia was afterwards appointed choirmaster to the Vatican, a post he held until his death in 1571, when he was succeeded by Palestrina.

The experiment succeeded so well that not only was a new form of composition originated, but what was probably more to the purpose of the founder, the congregation of the Fathers of the Oratory was placed upon so firm a basis that it exists to the present day. These musical performances were divided into two parts, a short form of prayer preceding the first, a sermon the second, the whole being concluded with religious exercises. By this plan those who came for the main purpose of the music were constrained to take part in the devotional observances, and doubtless, obtained much spiritual advantage. If in the performance of oratorio at the annual festivals in some of our English cathedrals the same or a similar arrangement were adopted, much of the scandal now accruing might be unquestionably avoided if not totally suppressed.

By degrees, the psalms and spiritual songs gave place to sacred stories or events in scripture written in verse, sometimes in dialogue, and set to music. The subjects most popular in early times were—*The Good Samaritan*, *The Prodigal Son*, *Tobit with the Angel*, *The Sacrifice of Abraham*, and others, and the name of the place where these were first heard was given to the class of music performed. The first work in any way corresponding to the more modern form of oratorio, with solos, recitatives, choruses, and orchestral accompaniments was called *La Rappresentazione di Anima e di Corpo*, composed by Emilio del Cavaliere, performed in the oratory of the Church of Santa Maria della Vallicella at Rome in 1600. Ten years previously Cavaliere had produced two pastoral dramas at Florence, and he has, therefore, claim to the honour of being considered as one of the first writers, if not the inventor of opera. Cavaliere's oratorio was represented in action on a stage erected in the church, with scenery and costumes; dances were also introduced. The recitatives in both his operas or oratorios were furnished by Jacopo Peri, and it was supposed that in the introduction of recitative the ancient use of the Greeks and Romans was recovered. In the preface to Cavaliere's oratorio each singer is required "to have a fine voice, perfectly in tune, and free from all defects in the delivery of it, with a pathetic expression, the power of swelling and diminishing the tones," and is enjoined to be "particularly attentive to the articulation and expression of the words, and to have an equal respect for the composer as for the poet.". The instruments, consisting of a double lyre, double guitar, a harpsichord, and two flutes, accompanied this oratorio behind the scenes. The choruses served as the music for the dancers, and madrigals commenced and concluded the performance. The violin was only sparingly used in the early oratorios, one of the first writers who wrote freely for that instrument being Domenico Mazzochi, and for this reason his name deserves mention among the early composers; beyond this fact there is nothing remarkable in any of the works produced until the time of Stradella, 1670, for the only noticeable points in most of these compositions are the curious mixtures of piety and profanity. Stradella's best oratorio was on the subject of *St. John the Baptist*, and, judging from the specimens printed in Martini's "Exemplare di contrappunto fugato," is remarkable for beauty and sublimity; the scoring also shows some progress in musical thought, but the work is little known, as it still remains in manuscript. The next writer of importance was Carissimi; his compositions are marked by sweetness of melody, clever modulations, and skilful harmony; his most noted oratorios are the *Judgment of Solomon*, and *Jephtha*. One of the choruses in the latter work, "Plorate filiæ Israel," was appropriated by Handel, and introduced into *Samson* to the words "Hear Jacob's God." The chorus "Exululantes filii Ammon" also forms the conclusion of "With thunder armed" in *Samson*. "Et clangebant tubæ" is imitated in the symphony of "We come in bright array" in *Judas Maccabæus*, and the recitative in *Alexander's Feast*, "He chose a mournful muse" is note for note identical with "Heu mihi, filia mea," from Carissimi's *Jephtha*. Many other instances might be cited of Handel's obligations to this musician, the most striking and important being that all his recitatives were formed upon the improved model furnished by Carissimi, who, in addition to other advancements in music, gave great importance to the use of stringed instruments in his accompaniments, made the basses of his compositions move more freely than former masters permitted, and first recommended the introduction of the cantata upon the stage. He died in 1675 at a great age. Dean Aldrich, who was an enthusiastic admirer of his works, made a large collection of them, which he placed in the library of Christ Church, Oxford, where it still remains. After Carissimi, Scarlatti deserves the next place of honour, not only as a prolific, but as a thoughtful composer; his works are rarely heard now, except when some of the choruses with Handel's name attached are performed. The device of accompanying recitatives with a string quartett was first tried by Scarlatti, and this plan was followed by Leonardo Leo, whose contrapuntal writing, similar in character to Handel's in sublimity and massiveness, forms the chief beauty of his oratorios and sacred pieces. As a master of counterpoint Caldara is worthy of mention, while as melodists Hasse and Pergolesi are best known; the beauty and grace of melody in these two composers is also to be found in a certain degree in the writings of Jomelli, who was the last of the oratorio writers of the Italian school.

Like the opera, oratorio arose and was nourished in Italy, to be developed to its highest form by German writers, and, although it was permitted to be used in the offices of the Roman church, and was to some extent abused by Italian writers, yet the value of oratorio was recognised by the German reformers of the sixteenth century, who encouraged its introduction into their churches with a result that ultimately led to its most complete forms. It was formerly the custom in the German Lutheran church to have oratorios performed with instrumental accompaniments on solemn occasions; the congregation was encouraged to join at intervals in the chorales or Psalm tunes which were introduced for this purpose. *The Passion, according to St. Matthew*, by John Sebastian Bach, is an excellent example of this kind of work. [Passion music.] Bach, one of the most original, versatile, and prolific writers that the world has yet seen, was born at Eisenach in 1685, and died at Leipsic in 1750. The study of his music has influenced the thoughts and writings of most composers of importance since his time, his organ compositions have created a new school of players, and the comparatively recent knowledge of his oratorios and vocal music has shown the extraordinary power and value of his genius, which, unknown or not understood in his own time, is now thoroughly appreciated. It is quite reasonable to assume that, had his vocal works been earlier known, a greater advance in oratorio music would have been made than has been made. The *Passion Music* alluded to above is written for two choirs and two orchestras. The choruses and chorales are exquisitely harmonised, the airs are remarkable for tender pathos and truth of expression, and the recitatives are notable for the admirable force with which the meaning and intention of the text is conveyed. Bach's oratorios were written for the service of the church, and were never intended for any other purpose, the introduction of the oratorio into secular places being a later idea. The subjects are partly in narrative and partly deductions from the incidents, after the manner of the old Greek chorus. This is the true form of oratorio as adapted to worship; but as the oratorio was disused as an aid to devotion, and employed in the theatre and concert room, the narrative form was the one most favoured for the purpose, and it is in this form that the majority of Handel's oratorios are written.

Handel, who is allowed to be the representative composer of oratorio, produced his first work, *La Resurrezione*, while he resided in Italy, before he was twenty years of age. Fifteen years later, while he was organist to the Duke of Chandos, *Esther*, the first oratorio written by him in England, was brought forward privately; it was, however, not until 1732 that it was produced in public, having been laid aside for twelve years. In the advertisement announcing this performance it was considered necessary to give the following explanation of the plan, as oratorios were then a novelty in England: "By His Majesty's command, at the King's Theatre in the Haymarket, on Tuesday, the 2nd May, will be performed the sacred story of *Esther*, an oratorio in English, composed by Mr. Handel, and to be performed by a great number of voices and instruments. N.B. There will be no acting on the stage, but the house will be fitted up in a decent manner for the audience." This explanation was probably made in deference to the objection popularly entertained at the time against any performance savouring of Popery and profanity, and was needful because of the prevalent prejudice against the dramatic performance of sacred subjects. After the production of *Esther* Handel was engaged in the production of operas and other secular works, and in the establishment of his short-lived academy of music—this entailed a loss of health and of much money. After the advanced age of 53, and in defiance of prejudice and failure, he gave to the world those immortal productions with which his name is now familiarly associated. The majority of these works were performed at the "theatre in the Haymarket," *Deborah* being first given in 1733, *Athaliah* in 1734, *Israel in Egypt* in 1738, *The Messiah* in 1741, *Samson* in 1742, *Judas Maccabæus* in 1746, *Joshua* in 1747, *Solomon* in 1749, and *Jephtha* in 1751. The greatest works in this list are *Israel in Egypt* and *The Messiah.* In all these, with the exception of the last-named, Handel availed himself freely of the license of appropriating other men's works and incorporating them in his own. It has been said that he has always improved and invigorated all pieces so borrowed, but it is hard to believe that the mere adaptation of words different to those originally set can be considered an improvement. Notwithstanding this pillage there is enough originality in Handel to constitute him a great composer, especially when it is considered that the work by which he is most popularly and extensively known, *The Messiah,* is for the most part his own. It is in broad choral writing that Handel's great strength exists, for he only sparingly employed instrumental effects, for the orchestra was only imperfectly developed in his time, and he generally employed his instruments more for the sake of supporting the voices than for any peculiar effects of colouring to be obtained from their use. There are some exceptions to this plan, but they are very few; the scantiness of Handel's effects, and the more extensive means employed for the performance of his works, gave rise to the additional parts which have been made for instruments either only partly known in his time, or of subsequent invention. The first of these additional accompaniments was furnished by Mozart to *The Messiah;* Mendelssohn, Macfarren, Perry, Costa, and Sullivan have supplied parts to the other of the more frequently performed works with more or less success.

The example set by Handel was followed by such writers as Stanley, Dr. Arne, Dr. Worgan, Dr. Arnold, but little encouragement was offered to their efforts, so that no new oratorio of importance was heard in London until Haydn's *Creation* was performed.

Haydn was the author of but few oratorios, *The Return of Tobias, The Seasons, The Seven Last Words,* and *The Creation,* being all his compositions that can be in any way classed under this head. *The Seven Last Words* is more in accordance with the character of the Lutheran oratorios, being intended to follow as many short sermons on the last words of our Lord. *The Seasons* is somewhat secular in character; and the same sweetness of melody prevalent in this is found in Haydn's best known work, *The Creation,* produced in 1798. The special qualities of this oratorio, besides those already alluded to, are found in the brightness of the choruses and the interest of the instrumentation. In general design *The Creation* is inadequate to the subject treated of, and while the ear is pleased by the sparkling and varied treatment, the heart is never moved by emotions which the grandeur of the theme should have inspired.

Beethoven's *The Mount of Olives* is a drama rather than an oratorio, full of sublime and noble thoughts, but ineffective without the aid of scenery and accessories.

Spohr's oratorios, *The Crucifixion, The Last Judgment,* and the *Fall of Babylon,* contain many grand and surprising thoughts, much beauty of melody and clever harmonies; the restless modulations employed by him cease to excite wonder after a time, and so, to a certain extent, their end is defeated. In the peculiarity of enharmonic changes on a dominant harmony, Spohr so far exhausted the possible combinations, that composers in imitating his style simply reproduce his thoughts. The most successful composer in this manner of modern date was Mendelssohn, and his *Elijah* and *St. Paul* served to revive the drooping taste for oratorio. *Elijah* was produced at Birmingham in 1846, and *St. Paul* ten years previously at Düsseldorf, and both works have since held high rank as well in the estimation as in the affections of musicians. Of the two, *St. Paul* more completely fulfils the conditions of oratorio proper, in the happy arrangement of its narrative and didactic portions. The production of this form is due to the assiduous study of the works of its great perfecter, John Sebastian Bach, and for this reason, and for its intrinsic merit, *St. Paul* is held to be Mendelssohn's best oratorio. *Elijah* is more dramatic in constitution, and if the prejudice should ever be sufficiently overcome to permit of its performance on the stage, with proper scenery and action, it will attain even greater importance than that it at present enjoys.

Of the more recent contributions to oratorio music, there are few worthy to be particularised, as they are for most part feeble filterings of an almost exhausted stream.

Oratorio has ever been more patronised in England than on the Continent, but until the establishment of the Sacred Harmonic Society in 1832 the opportunity for hearing it was of rare occurrence, being confined almost to the annual meetings of the three choirs of Gloucester, Hereford, and Worcester in the provinces, and to the Lenten performances in London. The example set by this Society has been imitated with success, not only in London, but in the country, and oratorio per-

Digitized by Google

formances are now frequent, and upon a scale of grandeur, magnificence, precision, and perfection hitherto unattained and completely unknown elsewhere.

Orchesis (*Gk.* ὄρχησις). **Orchestik** (*Ger.*) The art of rhythmical dancing as it existed in the Greek Theatre.

Orchester (*Ger.*) Orchestra.

Orchestra (*Gk.*) (1) The ὀρχήστρα or "dancing-place" in the Greek theatre was that circular space, immediately below the lowest seats of the semicircular part or κοῖλον, which was given up to the chorus. The level of the orchestra was twelve feet below that of the lowest seat, and twelve feet below that of the proscenium (προσκήνιος), but between the orchestra and the proscenium was a platform six feet higher than the orchestra, called the thymele (θυμέλη). If, therefore, the part-circle formed by the front of the lowest row of seats were supposed to be completed, it would be found that the orchestra would contain an area about one-third of the circle in that part farthest from the stage. (2) The place where the band, or band and chorus, are placed in modern concert-rooms, theatres, &c. (3) The collection of instruments of varied compass and quality of tone which constitutes a *full band*. There are no orchestral scores earlier than the latter part of the sixteenth century, so all statements as to concerted instrumental music before that time are wholly conjectural.

Many of the early poets speak of harmony in a manner that shows that their construction of the term was the same as our own. It is well known that harmony was not a sudden discovery, but the result of slow growth, and it is reasonable to suppose that one who could sing part-music according to a recognised system, and also play upon an instrument, would be led to select one as much in accordance with the compass of his own voice as possible, and so would *play* the part he was accustomed to *sing*. Accidental circumstances might suggest combinations which were afterwards made permanent, and from small beginnings the adjustment of tone considered necessary in a modern orchestra might be gradually arrived at.

One of the first authentic attempts to gain special effects may be traced in the disposition of the instruments accompanying the first operas. The score of Monteverde's *Orfeo* performed in 1608, contains indications of a purpose in the employment of what appears in the present as a somewhat strange arrangement of accompanying instruments. Orpheus sings to the accompaniment of the basses, Apollo to the organ, Pluto to the trombones, Eurydice to the violas, and Charon to the lutes or guitars. The use of particular instruments to accompany the music sung by special characters was revived by Hoffman in his opera *Undine*, and introduced by Wagner as a new invention; but the idea belongs to Monteverde, as the directions in the score quoted above prove. Information regarding this branch of musical history is to be obtained only by reference to scores of various dates, by which it will be seen that composers, for a long time subsequent to Monteverde, did not follow out the hints he had given. They made some omissions in the number and arrangement of the instruments, and in most cases simply contented themselves with violins, and sometimes added an organ for the accompaniment of everything but recitative, for which the harpsichord or lute was employed. The lute was used at the opera-house so late as 1790, the harpsichord much longer; the musician who sat at the latter instrument was called the conductor, and it was his business to give the time of the several pieces to the leader of the band. It is only within the present century that an independent time-beater, one not playing an instrument, has been fully recognised, and his position in the orchestra established. The story of Dr. Philip Hayes standing up to beat time for a band which Cramer was leading, and the refusal of the latter to play until "de fat man was sit down," shows that the practice of "conducting" was not general at the latter part of the last century.

In addition to violins, wind instruments were used, sparingly at first, but afterwards with great liberality. In Handel's time the proportion of treble reed wind instruments to the violins were as two to three, that is to say, for every six fiddles four oboes were employed. This is a matter which should not be lost sight of in calculating Handel's orchestral effects, for there is every reason to believe, however strange the combination might be to modern ears, that a restoration of the proportions of the orchestra for which Handel wrote would place his music in a far different light from that in which it is viewed at present, not that it would lose, but perhaps would rather gain in grandeur. It can scarcely be said that Handel did much towards developing the resources of the orchestra; he seemed to be mainly content with treating it as an accompanying body, for even in his concertos, he was contented with the performance of a phrase by the full body, which had been given out by a single instrument to a part. In a few cases he seems to have made a happy choice of an accompanying instrument, but even then it is considered so doubtful as to whether he meant what is called an "orchestral effect" or not, that many composers, following the

Digitized by Google

example of Mozart, consider themselves justified in adding instrumental parts to Handel's scores, not always with the laudable desire of exhibiting an improved knowledge of instrumental quality, not always with correct taste or judgment. Mozart himself declared to Attwood that his sole reason for adding parts to the scores of Handel's works was to make his countrymen think as highly of the compositions as he himself did; so that all the stories about his undertaking the task at the request of a noble patron, against his own inclination, are pure inventions. Mozart's additions to the *Messiah* are specimens of noble orchestral colouring, and give so much beauty to the score that a performance of the work without them would seem tame and uninteresting to ears that have grown accustomed to them.

The composer in whose works the first consistent and concentrated efforts in the direction of improvement of the orchestra are to be traced is Gluck. All his biographers agree in declaring that his sole aim in orchestral writing was "expression." He disregarded the usages of fashion and custom, and thought for himself, and like every independent thinker founded a school of his own, to become in time the conventionality of a future age.

In the introduction of new instruments by which fresh combinations and effects can be obtained, a difficulty will be always present, in that skilful players are not readily found, and musicians do not care to study an instrument that is rarely employed, and therefore of little profit to them professionally. This is the one reason why trumpet playing is comparatively a lost art. There is no remedy against this drawback, and therefore many instruments of unquestionable value in an orchestra are neglected and disused.

Meyerbeer did much towards obtaining a recognition for certain disregarded instruments, and although he was fortunate enough to find players with sufficient enthusiasm or interest in the matter to take them up and study them, he could not secure a succession, and, consequently, in the performance of his operas, such passages as are written in the score for obsolete instruments have to be performed by those in general use.

The introduction of new effects is always a work of time, and the development of some degree of daring. Lully wrote parts for drums to some of his choruses to strengthen the deep tones; Beethoven wrote solos for the same instrument. The prominence given by a composer to a certain instrument is to be accounted for in more than one way. He may have had a special training in that particular instrument, a partiality for its tone, or he may have in his orchestra players of exceptional ability. Thus Weber writes so wonderfully for the horns, because he was a player; Spohr for the violin, an instrument on which he was a master; Spontini loved the tone of the trombone, and wrote often for it; Handel had Valentine Snow to play his trumpet parts, and Snow had a special gift for trumpet playing; and many other instances may be quoted if it were needful.

Haydn has been considered the father of modern orchestral writing; and in the utilization of accepted materials his genius as a composer is best shown. In charm of orchestral shading Mozart's are perfect studies; while for nobility of effect Beethoven has never been surpassed.

The tendency of the present order of writers for the orchestra is to seek to obtain a fictitious grandeur by the liberal use of "Janitscharen-musik," cymbals, bells, big drum, &c. This is all very well when it is desirable to conceal the defective playing of the strings and wind, but when an orchestra comprises performers of excellent skill, having familiar knowledge of the capabilities of their several instruments, such noise is not only needless, but offers a great temptation to carelessness in playing on the more delicate instruments.

Ordinario (*It.*) Ordinary, as *a tempo ordinario*, in the usual time, or, at a moderate pace.

Orecchia musicale (*It.*) An ear for music, or musical ear.

Oreille musicale (*Fr.*) [Orecchia.]

Organ (from the *Gk.* ὄργανον). This word originally signified an *implement*, *instrument*, or *piece of mechanism* for any purpose. Hence, it came to be applied to any *musical* instrument, and later still to the compound wind-instrument now known as the organ, which, no doubt, grew out of the *ugab* of the Hebrews and the *syrinx* of the Greeks (*Pan's-pipes*), when combined with the wind-chest, as naturally suggested by the leathern reservoir of the *bagpipe*. Only, it must be remembered, that the pipes of Pan had, each separately, to be made into a sort of *flûte à bec*, having a foot to convey the air to the block and lips; also, that under each pipe a little slide had to be placed in order to prevent all the pipes from speaking at once. This early slide, which permitted particular pipes to speak at the will of the performer, has now developed into the *pallet* and the *key-action;* and we now apply the word *slide* or *slider* only to that strip of wood which, passing under a row of pipes from right to left, admits the air to a particular *row of pipes* or stops. Neither of these two arrangements—the *key*-slide and *stop*-slide—required any special mechanical skill, and it may safely be said that they were used at a much earlier date than is commonly supposed.

Digitized by Google

The real difficulty of ancient organ-builders was *inequality* of the pressure of wind. It is needless to observe that the leathern reservoir of a bagpipe could never be so equally pressed all round as to produce an uniform current of air. Nor could the simple forms of bellows as used by blacksmiths to this day produce an equal pressure, because, when they are being filled, the air is under too great compression, and as the top sinks the pressure becomes less and less. There can be but little doubt that the so-called *hydraulic* organ owed its utility and consequent fame to the fact, that in it water was used in such a manner as to counterbalance the hitherto variable pressure. The ignorant, no doubt, thought that the water went near to, or even entered, the pipes, but this was palpably not the case. Those who are interested in the supposed details of the hydraulic organ will find information in the "Bible Educator" (Parts 8 and 10, Vol. II.), in Rimbault and Hopkins on the Organ, and in Chappell's "History of Music" (p. 325). The ordinary organ in which no water was used came to be called the *pneumatic* or *wind-organ*. In order to keep up a due supply of wind, while bellows were being refilled, organ builders multiplied their number. The modern inventions of the double-action feeder, &c., fully explained further on, made this system of blowing unnecessary. If the account given of the Magrepha (*q.v.*) be true, it must have really been an organ. It will be unnecessary to trouble the reader with long untrustworthy quotations from old writers in explanation of, or in praise of, the wonders of organs. It will suffice to state where and when some of the oldest known instruments were built, and note the progress in their construction. Julian, the Apostate (who died A.D. 363), mentions an organ in a Greek epigram, but in such a manner as to make it doubtful whether it had *keys*. Aldhelm, (who died A.D. 709), mentions an organ which had *gilt pipes*. An organ having leaden pipes was placed in the Church of S. Corneille, at Compiegne, in the middle of the eighth century. The story of St. Dunstan's organ with its brass pipes is well known. One of the most interesting of the mediæval organs was that built for Winchester Cathedral at the close of the tenth century. According to the account given by the Monk Wulston it had *thirteen pairs of bellows* and *four hundred pipes* distributed so that forty pipes were under the control of each key. By pressing down the key belonging to any one set of pipes, they all sounded together after the manner of the mixture; the separation of the wind-chest by means of slides was apparently not attempted in this case.

The large pipes of every key of the oldest organs stood in front; the whole instrument sounded and shrieked in a harsh and loud manner. The key-board had 11, 12, even 13 keys in diatonic succession without semitones. It was impossible to get anything else than a choral melody for one voice only, on such an organ.

By degrees the keys and pipes in organs increased in number, but no alteration was made in the register of the pipe-work; the chromatic intervals, too, were added. The pedals were devised in 1470 by Bernhard the German, a skilful musician at Venice, and so quickly came into fashion that in Germany few organs were built without them. Originally they had only eight keys—♮B, C, D, E, F, G, A, ♮B, which were attached to the valves of the pipes with cord. As time went on the key-boards of the manuals became increased to three. This was the case in the large organ (built in 1361) at Halberstadt, described by Prætorius after its renovation in 1495, in "Syntagma Musicum" (1619). It had three key-boards and pedals, the latter, however, being added at the time the organ was restored. The Præstant, or row of pipes standing in front (the Principal or open diapason), was separated from the other portion of the pipe-work, which was almost detached, and was called Hintersatz. A difference of tone could thus at least be obtained by using the principal alone, or by coupling it with the whole hintersatz and pedals. The highest row of keys, called "Discant," had a compass from B♮ C, C♯ chromatic to G, A, and served for the full organ principal and hintersatz coupled. The second key-board of the same compass, and also called Discant, served for the principal alone without the hintersatz. The third row of keys was a bass key-board, contained a large bass principal in side-towers, with a compass from C to C, and was used with the left hand as principal in the second row of keys. The pedals extended from C to B♭.

It must not be thought that the manner of playing these old organs was anything like that now-a-days: the breadth of a key-board containing nine keys extended to three-quarters the length of a yard, that of the single key amounted to three inches (Bedos de Celles speaks of organ keys even from five to six inches). The keys were of a peculiar shape, *e.g.*:

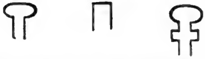

The valves of the keys and the whole mechanism being clumsy, playing with the finger was not to be thought of, but the keys were obliged to be struck with the clenched fist, and the organist was often called "pulsator organorum." It was consequently impossible to play chords with one hand, but a kind of duet or trio could be produced with the assistance of the pedals. The early organs were often a tone or a tone and a half above the choral pitch [v. Chorton], and lower than the kammerton by about a tone.

But when the organ had once attained a certain degree of perfection, it proceeded quickly towards further improvement. The keys were made gradually smaller, so that the Fifth had a width of a span, something similar to our octave, as in the old organ at Braunschweig. In the keys of the organs in the monastery at Bamberg (enlarged 1493), in the church of the Barefooted Friars at Nuremberg (1475), in the cathedral at Erfurt (1483), in that of St. Blasius at Braunschweig (1499), the width of the span of the octave amounted only to about one key more than at the present time, but the action was still clumsy. The compass of the manuals, mostly beginning at B flat, had extended to EE and FF, but the pedals were still limited in compass from A or B flat to B flat or B natural.

But the most considerable stride in improvement was the division of the Hintersatz into separate registers, which could now be used singly through the invention of the spring-box. Prætorius, about 1619, describes the slide-box as being quite common, and further states that it had already been in use for two hundred years, so that it must have been invented at the end of the fourteenth or the beginning of the fifteenth century, but was a long time gaining recognition. In the beginning of 1500 the pipe-work itself had also begun to be perfected in different ways. Hitherto it had the scale of the principal throughout. Pipes began to be covered probably about 1530. In later times reed-pipes and pointed flutes came into vogue, and by degrees bellows of very thin board were put in the place of the old bellows, which were constructed in folds, and the portable chamber-organ [Regals], since grown out of use, introduced. About the year 1677, Christian Förmer (born 1610) invented the anemometer.

The mistaken zeal of the Puritans caused a wanton destruction of organs and the discouragement of organ-building, so that when a revival of interest in the instrument took place, it was found that England had lost her ancient supremacy, and there was no longer a sufficient number of builders in the country. Foreign organ-builders were invited to England, and the most celebrated of these—Bernhard Schmidt (generally called Father Smith) with his nephew and Renatus Harris, established themselves in this country. Their immediate successors were Christopher Schroeder, Snetzler, and Byfield. These were followed by Avery and Green, and from them sprung the present generation of organ-builders, whose skill and ingenuity have done much to restore the ancient supremacy of organ-building in England.

Organ Construction. (§ 1.) A complete organ may be said to consist of five parts: choir organ, great organ, swell organ, solo organ, and pedal organ. These are all, however, in reality constructed on the same principle, that is to say, the manner of connection between the keys and the pallets, which admit a current of air to the pipes, is similar in principle in all. A large organ therefore consists of a number of small organs differing in quality of tone, and so arranged as to be under the control of one performer.

(§ 2.) Before explaining the connection between a key and a pipe, it will be necessary to explain the construction of the bellows, a receptacle which encloses atmospheric air and forces it through passages called "wind-trunks" to the sound-boards over which the pipes are arranged. The construction of modern bellows, termed horizontal to distinguish them from the old diagonal bellows (*q.v.*), and for the invention of which organists are probably indebted to an Englishman, Samuel Green, who is known to have furnished two organs with them in 1787 and 1788, may be best explained by the following figure:

Fig. 1.

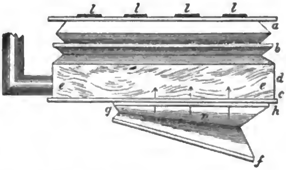

The top board *a* is substantially and strongly made of pine, ledgered at each end, or in large bellows panelled; some of the panels being moveable allow easy access to the interior for repairs. The middle frame *b* is merely a surrounding frame of wood to which is attached the lower edge of the upper ribs and the upper edge of the lower ribs. The middle board *c* is slightly larger than the top-board, and on it rests a strong ridge *d* called the trunk-band or lining, to which the wind trunks can be at any point joined, as at *e*. The bottom board *f* is the bottom of the feeder *f g h*, and is sometimes nearly as large as the top board. In large bellows provision is made for the due support of the weight of the top board by placing posts inside, or a ridge of wood extending the whole length of the middle board, of such a height that when *a* and *b* are at the position rest, *a* is touching the top of the posts or ridge. On the bottom of the middle board are a number of holes for the admission of air from the feeder, covered by leather valves or suckers, which fall and prevent the exit of the air by the way it came. The valves being hinged at one end only, rise for the entry of the air from the feeder, and then fall by their own weight. The bottom of the feeder *f* is supplied with a similar set of suckers, or valves, so that the operation of "blowing" is as follows: To the end of *f* are attached, in any way most convenient, the levers connected with the bellows handle, so that by its stroke *f* is raised towards *c*. The atmospheric contents of the feeder are then forced through the interior valves † † † and the top board rises. When *f* has reached its nearest point to *c*, and begins to return to its position in the figure, the leather valves fall over the openings † † †, and the air is forced by the heavy top board and its incumbent weights *l l l l* to depart by the only way open to it, namely, by the wind trunk *e*. It will be noticed that one set of ribs is turned inwards, the other outwards. This ingenious though simple arrangement, which was first suggested by a clock-maker named Cummins, and immediately adopted by Flight, the organ-builder, entirely removes the inequality of pressure spoken of in the article on "diagonal bellows." But as the feeder *f* is descending, no air can of course be supplied to the bellows. The simplest means of obtaining a constant supply of wind is to have two feeders, or, as it is termed, a double-feeder, which may, according to circumstances, be placed either under the side or end of the bellows.

Fig. 2.

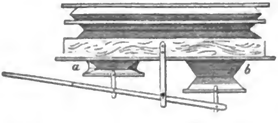

As one feeder *a* is descending, the other *b* is ascending. A like result is produced by the use of the *cuckoo feeder*, so-called because it is constructed on the same principle as the little bellows of the cuckoo toy.

Fig. 3.

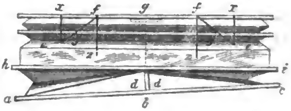

The bottom board *a c* is hinged at *b*, so that when *a* ascends, *c* descends, both *a* and *c* being, of course, provided with the ordinary valves. The under lining *d d* is sometimes exactly reversed in position, being attached to the middle board instead of the bottom board. Before leaving the subject of bellows, the *counter-balances* must be explained. Their object is to make both sets of ribs expand or contract equally, and prevent the swinging to and fro of the top of the bellows. They consist of three flat pieces of iron, and are fixed to the upper board at *x* (Fig. 3), the middle board at *y*, and the trunk lining at *z*; and being moveable at every joint but *x* and *z*, as *x* rises, *e f* impart a proportionate rise to *y*. In order to prevent an undue rising of the bellows when more wind is supplied than used, a *waste-pallet* is placed in every bellows. It consists of a valve (*g* Fig. 3) pressed under an opening in the top board by a spring, having attached to its under-side a piece of cord (dotted line) which is fixed to the bottom board *h i*. When the top board rises higher than the length of the cord, the valve is of course pulled downwards, and the superfluous air escapes. If any part of the middle board is not over a feeder, the pallet may be placed inside the middle board, and the cord fixed to the top board, in which case the pallet will be pulled up when the top board, by its rising, pulls the cord.

(§ 3.) But notwithstanding all these improvements in feeders and bellows, it was found that an unsteadiness of wind was caused, if many of the larger pipes were suddenly made to speak, or as suddenly, to cease speaking. In the former case so large a supply of atmosphere was taken from the wind chest that the air remaining expanded by its own elasticity; in the latter, the supply of air did not cease with the demand. Both these defects are remedied by the clever invention by Mr. Bishop, the organ-builder, of the concussion-bellows, which are a reservoir hinged at one end and therefore wedge-shaped, which is attached to the side of the wind-trunk, with a spring tending to force the moveable board towards the trunk.

Fig. 4.

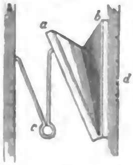

When there is no air being forced through the trunk *d*, *a* reaches its nearest point to *b*. When the trunk is filled with compressed air the spring *c* just counterbalances it. When a sudden demand is made upon the contents of the trunk *d* the spring forces *a* nearer to *b*; when the air in the trunk is more than usually compressed, the spring allows *a* to recede from *b*.

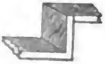

(§ 4.) The wind-trunk is made of an oblong form, and is shaped so as to carry the air by the most convenient route, to the wind-chest, which is immediately under the sound-board. It will be perhaps better before entering into details to give at once a general idea of the wind-chest and sound-board.

Fig. 5.

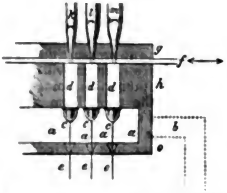

The above figure (Fig. 5) represents a section of a sound-board and wind-chest viewed from front. *a* is the wind-chest into which compressed atmospheric air has been introduced either through the side or bottom, from the end of the wind-trunk *b*. The pallets *c c c* are held tightly against the openings *d d d*, leading from the wind-chest to the mouth of the pipes by springs underneath them, thus:

F.g. 6.

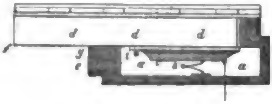

The spring *s* keeps the pallet *c* against the opening into *d*. The wires called *pull-downs* (*e e e* Fig. 5), which pass through small holes in the bottom of the wind-chest and are in connection with the key-board, are attached to a loop of wire called the *pallet-eye*, fastened to the moveable end of the pallet. A piece of wire is placed on each side of every pallet to steady it and keep it in the perpendicular during its ascent and descent, and every pallet is covered at top with soft leather, to make it fit closely and work quietly. When *e* is pulled down (Fig. 5) the pallet *c* descends, and air from the wind-chest *a a a* rushes through *d* into the pipe over it. But the slider *f* is a narrow strip of wood, so placed between the woodwork *g* and *h* that it may be moved backwards and forwards from right to left, and is pierced with holes corresponding throughout to those just under the pipes. If the apertures in the slider are under the pipes (as represented by dotted lines in Fig. 5), the opening of a pallet will make a pipe speak; if, however, the slider has been moved so that the apertures do not correspond (as shewn in Fig. 7 by dotted lines), even if the pallet be opened and the chest full of air from the trunks, no sound will be produced.

Fig. 7.

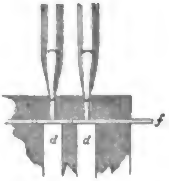

When the apertures in the slider are under those below the pipe, the "stop," the handle of which controls the position of the slider, is said to be *out*, or *drawn*. When the apertures do not correspond, the stop is said to be *in*. Thus it is that when no stops are drawn no sound is produced, even although the wind-chest be full of air from the wind-trunk. Looking at Fig. 5 again, it will be observed that if another row of pipes (or a "stop") were placed immediately behind *k*, *l*, *m*, with another slider under them, the same openings *d* could supply both sets of pipes, subject to the control of the sliders. The only limit to the number of rows of pipes will therefore be the depth of the air passage *d* over which the pipes stand. But, inasmuch as *one* pallet under the control of *one* key will admit air to

Digitized by Google

all the pipes placed over any one air-passage *d*, it is evident that only those pipes must be placed over *d* which belong to one particular note on the key-board. The air passage *d* is technically termed a *groove*, the wood which separates one groove from another is technically termed a sound-board *bar*. Every row of keys on an organ will therefore have in connection with it a sound-board, having as many grooves and pallets as there are notes in its compass; and as many sliders as it has stops. Having made a slight digression in order to give a general idea of the construction of this important part of the instrument, it is now time to enter more into detail, and first, it will naturally follow that large pipes will require a larger groove and pallet than those that are smaller, hence the grooves vary in size throughout the whole sound-board.

(§ 5.) The usual arrangement of pipes on a sound-board is to have the largest on the outside, say on the right and left sides, and then to proceed by degrees to the smallest which thus come nearest to the middle of the sound-board. But the arrangement of pipes must be spoken of again hereafter. If a greater supply of wind than usual is required for the due speaking of any large pipe, two grooves may be allotted to that note. These *double-grooves*, as they are termed, are said to have been first introduced by England, the organ-builder. By reference to Fig. 6, it will be seen that the groove *d* runs much farther back than the wind-chest *a*, *a*, *a*. This is, of course, necessary where a large number of stops has to be placed over it. The width and depth of the grooves varies according to the number of the stops, and the size of the pipes. A sound-board is made by taking a strong piece of wood of the required size, turning it over and then glueing the sound-board bars to it, placing between the ends of the bars pieces of wood called *fillings in*, the width of the groove; then the ends of the bars and the fillings in which form the front and back (called the *cheeks*) are planed down and covered firmly with a strong piece of mahogany or other hard wood. Then the wind-chest, pallets, &c., cover over part of the grooves, and the remainder (*f*, *g*. Fig. 6), the openings of which have hitherto been exposed to view, are covered by parchment or leather. This, of course, becomes the underside of the sound-board when it is turned over into its proper position. The wood to which the grooves are glued now becomes the *table*, and is bored to admit air to the pipes. On the table are arranged the sliders, over the sliders are the upper-boards, into the holes of which the ends (*noses*) of the pipe enter, above the upper-boards are the pipe-racks, simple frameworks of wood which keep the pipes in a perpendicular position.

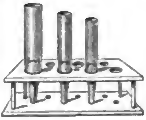

As the amount of room which an organ occupies is generally a matter of some importance, the pipes in the sound-board are rarely arranged in a straight line thus

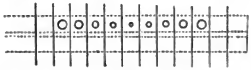

but are placed slightly zig-zag, thus:

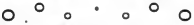

but as the grooves run in the direction indicated by the plain lines, and the sliders as indicated by the dotted lines, it is evident that such an arrangement only affects the piercing of the table, slider, and upper boards. The sliders might, if their edges touched one another, disturb each other's position when pulled in and out; to prevent this, between each slider and that next to it is placed a *bearer* or immoveable strip of wood, which receives the weight of the upper-boards, and to which the upper-boards are screwed. There should of course be no pressure on the sliders, otherwise their free movement will be hindered.

(§ 6.) Having thus given some account of the progress of the wind from the time it enters the bellows until the drawing down of the pallet, and the position of the slider has permitted it to enter into the pipes and make them speak, it remains to show, first, by what means the key is enabled to pull down the pallet; next, by what method the stop-handle or register acts upon the slider. But before the former can be explained it will be necessary to say a few words as to the arrangement of the pipes on a sound-board, for on this will depend to a great extent the kind of connection between key and pallet. The simplest arrangement will of course be to place the smallest pipe on the right-hand side of the sound-board and to proceed thence by gradation to the largest pipe on the left-hand side. This would be naturally suggested by the order in which the keys are placed, the highest note, produced by the smallest pipe, being on the right; the lowest note, produced by the largest pipe, on the left. This arrangement is known as Vogler's system, because that learned musician and organist

Digitized by Google

brought it prominently before the public. But it has some great disadvantages, the chief of which are as follows: first, as the lowest notes are produced by the largest pipes, all the largest pipes, and in consequence the greater part of the *weight* of an organ, rests on one side of the sound-board; secondly, the largest pipes occupy by far the greatest space, hence, the left-hand side of a sound-board must be proportionately roomy in length and depth; thirdly, one end only of the wind-chest is taxed with the supply of all the big pipes; lastly, if it is found necessary to *convey* wind from the grooves to ornamental pipes standing in the front, the distance is extremely great. In order to explain this last objection it must be here stated that if a pipe does not stand in its proper place on the sound-board, either because there is not room for it, or because it is placed at a distance in the front of the organ, a roughly made pipe of light metal is glued at one end into the proper hole and carried by any course most convenient, forming sometimes the most fantastic shapes, to the base of the new position of the pipe, the other end is then firmly glued into a small hollow immediately under the pipe. Therefore, if *conveyances* (as these supplementary tubes are called) are very long, the pipe does not speak very readily, the compressed air when the pallet is opened, having to travel from the place where the pipe ought to be, to the spot where it may happen to be. The third objection above stated may be surmounted by making larger wind-trunks and wider grooves. But even if there be no case to an organ, the Vogler system has its disadvantages, for the outline formed by an unbroken succession of pipes from 16ft. to a few inches in length, standing side by side, in the short space of a sound-board, can hardly be rendered pleasing to the eye. The second system of arranging pipes has already been alluded to, namely, that of placing the largest pipes alternately on the outside, proceeding by degrees until the smallest are in the centre, and the outline formed by the tops of the pipes nearly assumes the shape of a slightly flattened V. The third system is the reverse of the above, the large pipes being placed in the middle and the small ones at the sides, the outline of the pipes will be therefore not unlike Λ.

The fourth system is a compromise between the first and second, most of the pipes being arranged semitonally, but some of the larger pipes being placed on either side. The outline of the pipes will, therefore, be of this kind. These four are the chief arrangements of pipes in vogue, but of course exceptional positions for an organ necessitate exceptional dispositions of the pipes.

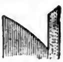

(§ 7.) Before explaining the means which connect keys with pallets, it must be pointed out that there are two kinds of keyboards or *claviers*, one of which is played by the hands, called the manual-clavier or *manual;* the other played by the feet, called the pedal-clavier or *pedals*. The long keys of a manual are generally made of lime-tree wood covered with ivory, and the short keys of ebony, or some other wood stained black. Both long and short keys are balanced on a ledge of wood called the *pin-rail*, because metal pins, fixed into it and entering into punctures in the keys, steady them perpendicularly whilst allowing them to move on them, as on pivots. A similar rail passes immediately under the front of the keys, with another set of pins, to produce steadiness of movement, only these punctures do not go completely through the key as in the previous case. The pieces of wood on each side of the manual, to which the pin-rails are firmly fixed, are called *key-checks*. To prevent an undue rising of the key when the finger is suddenly removed, a piece of board covered on the underside with baize, and loaded with lead, called the *thumping-board*, lies on the keys transversely immediately behind the part which is used by the fingers. The pedal keys are made of oak, mahogany, or any hard wood, and work on centres at the heel end, being steadied at the toe end. A spring under each key holds it in position subject to the pressure of the foot.

(§ 8.) On pressing down a manual key, the back of one key rises, and raises a *sticker*, a thin piece of wood, round or square, and of length varying from a few inches to a couple of feet, provided with a pin of wire firmly fixed in the upper end, which passes into the end of the *back-fall*, a horizontal lever, to which is attached a tracker, a strip of light wood provided with a tap wire and leather button at the end, passing through the back-fall, and with a hook of copper at the end connected with the pull-down, thus—

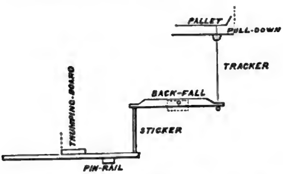

If, then, the arrangement of the pipes be according to the first plan given (largest on the left hand side, smallest on the right), the back-falls are arranged thus, on the back-fall frame, so that the end of the back-fall may come under the pull-down. This kind of movement is, from its appearance, called the *fan-frame* movement.

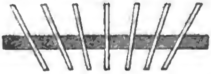

But if the pipes are arranged according to the second plan given, a piece of mechanism called the roller-board comes between the end of the back-fall and the puller down. It is a series of rollers so arranged on a flat board standing upright, that whilst one end of the roller is under the pull-down, to which it is attached by a second tracker, the other end is over the corresponding back-fall, the back-falls, in this case, lying parallel to each other. Thus, by means of the roller board, the finger is enabled to communicate with pipes on either side of the sound-board. The system employed must, however, depend greatly on the position of the sound-board with relation to the keys.

Where corners intercept action, or where it is necessary to carry action at right angles, either horizontally or perpendicularly, *squares* are largely used. The square is of wood, and of this shape, and generally in two pieces, dove-tailed, so that no strain may cause one arm to snap off. The arms are generally about 3 inches in length. The block in which the squares work is termed the *square-frame*. Of course the length of stickers, trackers, and back-falls must entirely depend on the space to be occupied by an instrument, or the position in which it is placed. Long stickers or trackers are steadied by being made to pass through holes in a flat piece of wood called a register.

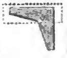

The connection between pedal-keys and pedal-pipes does not differ in principle from that between manual-keys and manual-pipes, but from the usual position of the pedal organ squares have frequently to be used.

(§ 9.) The draw-stop handle acts upon the slider by means of a series of levers, the position of which varies greatly in different instruments. The most common system is called the wooden-trundle draw-stop action.

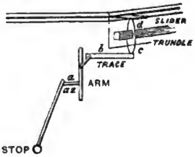

When the stop is pulled out, the arms *a a* draw the trace *b* from right to left, the end of the trundle *c* being attached to the trace is moved in a similar direction, whilst the other end of the trundle *d* moves in an opposite direction, and draws out the slider. When metal trundles are used, this action is simplified. Sliders are sometimes made to move by means of small bellows placed under them, which, upon being filled with compressed air, suddenly expand, and as they expand draw the slider. A large organ, it will be now understood, is practically several distinct instruments, not differing in the principles of their construction so much as in quality of tone. But if this were all, the interior of an organ would appear very simple—there are still to be considered those contrivances which enable a performer to play upon more than one set of keys simultaneously.

(§ 10.) The draw stops producing combinations of manuals or of manuals to pedals are termed couplers. Suppose an organ to consist of four manuals and pedals, *viz.:* solo organ, swell organ, great organ, choir organ, and pedal organ, the couplers assigned to it would probably be "Solo to Great," "Swell to Great," "Swell to Choir," "Choir to Great," "Solo to Pedals," "Swell to Pedals," "Great to Pedals," "Choir to Pedals." To these are sometimes added couplers combining the super-octave or sub-octave of the swell with the great, or similar intervals of the swell organ on itself.

The construction of unison manual-couplers is generally of the following kind. A rod passes between the back of two sets of manual keys, which is pierced with a series of holes in which are placed little stickers of this shape. The length of the sticker is the distance between the two sets of keys to be combined, so that when the lower one is pressed down by the finger the sticker is pushed up by the back of the key and raises the back of the key lying above it. When not required for use the coupler-stop leaves the stickers standing in a hollow purposely made in the keys, thus :—

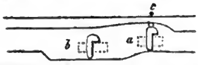

a representing the coupler when ready for use, or "drawn," *b* when not required, and *c* the regulating button. The *tumbler* coupler is now almost obsolete. It consisted of a rod containing stickers, which were turned at an angle when not required for use thus : but it will be seen that if a coupler of this kind is "drawn" while the fingers are on the lower row of

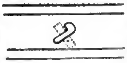

keys, the fingers will have to withstand a sudden pressure. This evil is avoided in the modern coupler. Octave couplers, coupling the octave either above or below, are of the backfall description. As the pedal-keys occupy more space than the manual-keys a roller-board is used, made on the same principle and practically for the same purpose as that already described. One end of the roller is over the pedal-key and the other under the manual-key desired to be affected. A sticker rests on the end of the backfall, when the pedal-coupler is "out," but is thrown out of the perpendicular when the coupler-stop is "in." A coupler attaching the great to the swell organ, termed a sforzando coupler, capable of being used for a single chord if required, is formed by a double backfall thus,

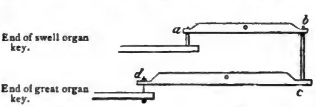

so that when *a* rises *b* is depressed and forces down *c* which in turn raises *d*.

(§ 11.) To enable a performer to make rapid changes of stops, short iron pedals are placed over the pedal-clavier, called composition-pedals or combination-pedals, whose office it is to produce certain definite arrangements of the registers when pressed down with the foot. A single-action composition pedal is one that draws out a certain number of stops, but does not alter the arrangement of the other stops on the same manual during the operation. A double-action composition pedal is one that has the power of drawing *in* any stops not belonging to its prescribed combination which may happen to be *out;* for example, supposing the trumpet and sesquialtera on the great organ are drawn out, and the foot is then placed on the composition pedal which draws out the diapasons, if the action of the pedal be *single* the trumpet and sesquialtera will remain out as the diapasons are drawn; if the action be *double*, as the diapasons are being drawn out, the trumpet and sesquialtera will be *drawn in.*

Composition pedals act by leverage either on the back of the draw stop handles or on the ends of the sliders, but it would exceed the due limits of this article to explain even a few of the many systems of connection between composition pedals and sliders which are in use among organ builders. Pneumatic bellows are not unfrequently employed for purposes of combining registers, and are often set in motion by small pegs or nobs projecting between the manuals, instead of iron pedals. In cases where several sound-boards are attached to one manual, containing each of them part of the stops, changes of combination are sometimes obtained by pedals or pegs, called *ventils*, which admit the wind to, or cut the wind off from, the separate sound-boards.

(§ 12.) The pneumatic action is a contrivance for lightening the touch of manual or pedal keys by placing near each key a small bellows, having two openings on the under side, one of which is a pallet in a chest of compressed air, the other a waste-valve. The following diagrams will explain this:—

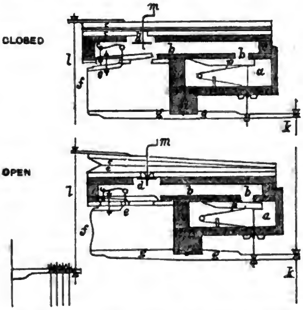

When either the finger or foot is pressed upon a key connected with *k*, the outer end of the back-fall *gg* is pulled down, which opens the pallet *p*. The compressed air in *a* then rushes through the groove *bb* into the bellows *cc*, which rises and lifts with it all the action attached to it by *l*. As the top of the bellows *cc* rises, it lifts up the throttle-valve *d* (regulated by the wire *m*) which prevents the ingress of any more compressed air by *bb*. But the action of the key on *gg*, which opened the pallet *p* also allowed the double-acting waste-valve *e* to close, and the tape *f* hangs loose. The compressed air, therefore, as it is admitted through *bb* cannot escape, but on the other hand when the key releases the outer end of *g*, and lets it rise up again, the tape *f* becomes tightened and opens the waste-valve, the bellows *cc* then drops into its closed position.

The principle of this invaluable invention was introduced by Mr. Hamilton of Edinburgh, or Mr. Barker. The latter gentleman succeeded in making its importance understood by some leading French organ-builders, by whom it was at once adopted, and whence it found its way back to this country. English organ-builders have, however, altered very much the original form of the pneumatic lever. The diagrams given above represent the lever as improved by the eminent organ-builder, Henry Willis, to whose talent organists are also indebted for the compressed-air action instead of trackers for long distances.

Tubular Pneumatic action, as this is termed, is thus formed. Ordinary tubing is taken, from the point where the impetus is given, to the point where the action is to be set in motion. When compressed air is admitted at the former end of the tube, it immediately forces up a disc placed at the other end, to which the required action is attached.

(§ 13.) Organ pipes are divided, with reference to their material, into metal or wood; with reference to their manner of producing tone, into reed and flue. Of the metals used in making organ pipes tin is the best, lead the worst. A mixture of tin and lead in equal proportions is technically termed "metal." An excess in the relative quantity of lead renders "metal" of proportionately less value. Zinc is used sometimes for large pipes, where cheapness is an object. Antimony may be used, but not in too great quantities, its brittleness being a serious drawback to its durability. Spotted metal is so called because of the mottled appearance of its surface. The spots are considered a guarantee of a fairly good material, as they do not rise unless there is one third of tin in the metal. Metal is prepared for use by being, when melted, poured into a wooden trough with a slit in the back, and passed rapidly along a bench faced with tick. The thickness of the sheet of metal depends, of course, on the size of the slit in the trough. It is afterwards planed down carefully to an uniform surface.

Deal, pine, and cedar, are the woods most commonly used for pipes; more rarely, oak, mahogany, maple, &c.; the harder woods being generally chosen for pipes to be used in hot countries.

The tone of a flue-pipe is produced by the resonance of a column of air, which is set in vibration by a current of compressed air forced through a slit, over which is placed a sharp edge. The pitch is therefore dependent on the length of the pipe. The tone of a reed-pipe is produced by a thin strip of metal which is placed over the entrance of a tube, in such a manner that the compressed air, in trying to force its way into the tube sets the strip of metal into vibration. The pitch of the reed-pipe is, therefore, dependent on the number of vibrations of the reed (strip of metal), *not necessarily on the length of the tube.*

(§ 14.) Flue pipes are of two kinds, open or closed. The tone of a closed pipe is, in its pitch, an octave below that of an open pipe of the same length, the cause of which is, that in an open pipe the vibrating column of air is divided by a node (or point of rest) at its centre, whereas in a closed pipe the column of air is undivided. An open metal flue pipe consists of *foot*, and *language*, and *body*. The foot is that on which it stands, and into which the compressed air enters.

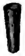

FOOT.

LANGUAGE.

The language is a flat piece of metal fastened by its edge to the top of the foot, and which by its shape, only permits the air to leave the foot in one direction. The whole of the pipe above the foot and language is termed the body. The lower edge of the body is indented in front thus, the indentation is called the *leaf*. The lower edge of the leaf forms the *upper lip* and the corresponding edge of the foot which is also indented is termed the *lower lip*. The space between the lips is termed the *mouth*. A wooden flue-pipe consists of a *foot*, *cap*, *block*, and *body*. The foot is a hollow wooden tube on which the pipe stands, and through which the air enters into it.

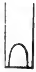

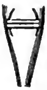

The air on entering is arrested in its upward course by the *block*.

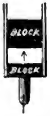

It, therefore, comes out towards the front through the *throat*, as the opening in the block is termed, and passes along the cap by which it is forced through the small opening between the top of the cap (the *under-lip*) and the top of the block, called the *wind-way*. The four boards used in making the pipe form the *body*. The *upper-lip* of a wooden pipe is generally made by bevelling off that part of the body of the pipe just over the upper edge of the cap,

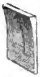

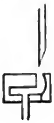

but the mouths are sometimes inverted thus:

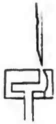

Occasionally also the block is superseded by a wooden "language," similar in its position and office to that in a metal pipe, thus:

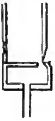

The block is sometimes bevelled off on its underside, while the cap remains straight, thus:

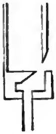

The following section of a flue-pipe will probably make its construction clear:—*t* is the foot through which air enters; *c* the space between the block and cap; *d e* the slit through which the air is directed to strike against the lip, *a b*; and *r* is inside of the pipe, which acts as a resonance-box.

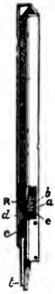

(§ 15.) A reed-pipe consists of a *boot*, *block*, *reed*, *tongue*, *wedge*, *tuning wire*, and *tube*.

The *boot* is a metal case corresponding somewhat in its office to the *foot* of a flue-pipe, inasmuch as the weight of the pipe rests upon it, and through it the air passes upwards. The *block* is a round piece of metal fitting into the boot, pierced in two places, one to contain the reed, the other to admit the tuning wire. It would appear thus in section. Into the larger orifice the reed is forced, through the smaller the tuning wire is passed. The rim prevents it from being forced into the boot by the superincumbent weight of tube.

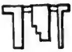

The *reed* is a brass tube rather wider at the bottom than the top, and closed at the bottom, having a narrow orifice over which lies the *tongue*, a thin elastic piece of brass large enough to cover the orifice and its edges.

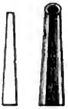

It is kept in its position by a small wedge of wood which is forced into the orifice in the block with the reed. The lower end of the tongue is therefore perfectly free. The *tuning-wire* is a piece of common wire bent in such a manner that a flat part of it lies across the *tongue*. By moving the wire from the outside, more or less of the tongue is allowed to vibrate, and thus the pitch is regulated. The upper end stands just in front of the *tube* of wood or metal, into which the air passes when leaving the reed, and which adds so materially to the tone of a reed-pipe. In small pipes the tube is generally of metal, but in large pipes often of wood; it is conical in shape if a full tone is required; if not, circular or cylindrical with a bell-top; and is either soldered to the top of the block, as is the case in small pipes, or is dropped into a small socket on the top of the block. Though, as before stated, the tube is not essential for the regulation of the pitch, its length and shape greatly influence the quality of sound produced. The *closed reed* is a reed which admits a much smaller quantity of wind than the open or common reed.

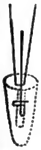

(§ 16.) The free or common harmonium reed is a reed the tongue of which does not lie over the edge of the aperture, but is able to pass freely backwards and forwards. Hence its name *free*, in opposition to the kind of reeds above described which are termed "percussion reeds." They are not commonly used in organs, but are more often met with in continental than in English organs.

(§ 17.) The swell organ is the successor of the old echo organ, a portion of a few stops enclosed in a wooden box in order to produce a subdued tone, or the effect of distance. In 1712, Abraham Jordan invented the "Nag's-head swell," as it was afterwards termed. It consisted of an echo organ, having, instead of a fixed front, a moveable shutter working up and down in a window sash. A pedal on the right hand side of the performer raised and closed the shutter at will, thus producing a *crescendo* and *diminuendo*. But the mechanism of this old "swelling organ" was most unsatisfactory. A very heavy pressure on the pedal was required before the shutter would stir, but once in motion it would often run up to its extreme limit against the wish of the performer, coming down to its closed position again when the pedal was released with a loud "bang." After half a century, England, and Avery, and afterwards Green, made use of the Venetian Swell, which consists of a series of shutters from six to ten inches in breadth, and from a half to two and a half inches in thickness, each hung on an axis one-third from the top of its breadth. The edges of each shutter are bevelled away at an angle of about forty-five degrees, so that each two contiguous boards overlap each other, and also are covered with felt in order to fit close, shut quietly, and shut in the sound thoroughly. The inside of the whole box may be lined with thick brown paper, glued on to the wood-work; this prevents the escape of sound, and does not injure the reverberation when the shutters are open. The Venetian Swell is so named because of the similarity of the appearance of the shutters to that of a common Venetian blind. The shutters sometimes are placed perpendicularly instead of horizontally, and occasionally open at the sides or back, as well as in the front.

(§ 18.) The quality of tone, and to some extent the pitch of pipes, whether they are wood or metal, reed or flue, depends upon their shape or outline, their scale, and the pressure of the wind to which they are subject. Metal pipes are in shape either cylindrical, conical, or inverted-conical, or surmounted by a bell whether cylindrical or conical. Wood pipes are four-sided, the opposite sides being equal in width, or pyramidal, or inverted-pyramidal.

(§ 19.) By Scale is meant the relation between the width and length of a pipe. If a pipe is wide in proportion to its length it is said to be of a large scale; if narrow in proportion to its length of a small scale. Inasmuch as scale affects to some degree the pitch of a pipe, it necessitates a proportionate alteration in its length, thus a pipe of a small scale must be made somewhat longer than a pipe of a large scale sounding the same note.

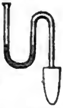

(§ 20.) The amount of the pressure or weight of wind is regulated by the weights placed on the bellows. It is found, by using a wind-gauge or anemometer, a bent glass tube, which is fixed into a socket, and into which a little water is poured. On placing the socket on one of the holes in the upper board of a sound-board, and admitting the air from the bellows, the water is forced up the tube, and the difference between the level of the two surfaces of water is measured in inches. The usual weight of wind on the swell, great, and choir organs of a church organ, is from two and a half to three inches, the pedal organ (where a separate bellows makes it feasible) receiving rather a higher pressure. The fourth manual or solo organ is nearly always voiced to a very high pressure, as very fine effects are in this way obtained both from reed and flue pipes. Organs in concert rooms and public buildings are often voiced throughout to a higher pressure than those in churches. The average amount of surface weight on the bellows required to produce a three-inch wind is nine pounds and three quarters per superficial foot.

(§ 21.) As regards the disposition of the various qualities of tone between the manuals,—the choir organ, being as its name implies, intended for the accompaniment of voices, generally receives stops of a delicate and subdued quality of tone, such as the Spitzflöte, Gemshorn, Viol di Gamba, Dulciana, Keraulophon, &c. The great organ being required for grand effects, has allotted to it the large scale Diapasons, together with a complete set of octave, twelfth, super-octave, mixture stops, and a proportionate number of large scale reed-stops. The swell organ has a larger proportion of reeds than any other manual, chiefly because of the fine effects they produce during the opening and closing of the shutters of the swell box. The reed-stops are generally placed in the front of the swell sound-board, just inside the box, not only because in that position the effect of a *crescendo* is increased, but also because they are more easily within the reach of a tuner. The Pedal Organ, as an independent organ, cannot be said to have been known in England for more than the last half century, although apparently always a constituent part of a good German instrument. The deepest tones (32 feet) are consigned to the Pedal Organ, as also a large number of 16 feet stops, to serve as a bass to the manuals, which have rarely more than one or two of that pitch, on any separate manual.

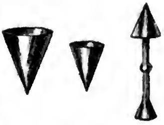

(§ 22.) An organ is tuned by means of hollow cones and reed-knives. The former are used for tuning metal flue-pipes, and are used in this manner. If a pipe is too flat, the apex of the cone is inserted into the top of the pipe, and the cone being forced in enlarges the orifice and raises the pitch. If a pipe is too sharp, a cone whose base is so broad that it will admit the top of the pipe,

Digitized by Google

is placed on the pipe and squeezed down until the orifice is reduced in size, and the pitch is flattened enough. Cones are made of metal or wood, more often the latter, because of its comparative lightness. Wooden flue-pipes are tuned, if stopped, by moving the stopper up or down; if open, by a piece of lead placed on the back of the top of the pipe for the purpose.

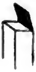

Large open wood pipes are tuned by cutting off part of the top if too flat, and by raising a piece of wood screwed for the purpose near the top if too sharp.

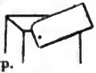

The largest metal-pipes have generally an opening at the back of the top, which can be opened or closed as it is necessary to make them flatter or sharper.

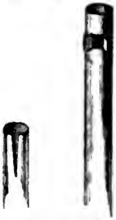

Metal pipes, with soldered covers, with or without chimneys, are tuned by altering the position of the ears, which are made specially long.

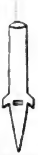

(§ 23.) The reed-knife is a long piece of common iron or steel in this shape, hence its name.

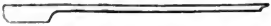

The tuning wire is provided at its upper extremity with a hook or projection by which the edge of the knife is able, by gentle taps, to gradually raise it. The wire is pushed in by being struck on the top. By referring to the account of the tuning wire given in §15, it will be understood that raising the wire flattens the note, depressing it sharpens it.

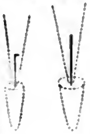

Organetto (*It.*) A little organ.

Organic Music. A mediæval name for instrumental music.

Organo (*It.*) Organ. *Organo pleno*, the full organ.

Organochordium. An instrument having pipes and strings, invented by Vogler.

Organo di legno. A Claque-bois. [Ligneum Psalterium.]

Organophonic. The name assumed by a band of Polish performers, who imitated the sounds made by various instruments with their mouths.

Organo portabile (*It.*) A *portative* organ. A small organ capable of being easily carried about.

Organ-point. [Fugue.]

Organum. (1) An early form of harmony, called also *diaphony*. [Descant.] (2) An organ.

Organum hydraulicum (*Lat.*) Hydraulic organ.

Organum pneumaticum (*Lat.*) The common organ. [Pneumatic Organ.]

Orgel (*Ger.*) An organ.

Orgelgehäuse (*Ger.*) An organ case.

Orgelpunkt (*Ger.*) A pedal point. [Fugue.]

Orgue (*Fr.*) Organ or harmonium.

Orgue expressif (*Fr.*) Harmonium.

Orgue portatif (*Fr.*) A portable organ.

Orificium (*Lat.*) The mouth of an organ-pipe.

Original position. A chord is said to be in its original position when the ground-note is in the bass; in other words, before it has undergone inversion: or, when its upper notes are in the order 3, 5, 8.

Ornamenti (*It.*) Grace-notes and other embellishments.

Ornatamente, Ornato (*It.*) With embellishments.

Orpharion. A kind of lute, having *wire* strings. [Lute.]

Osservanza, con (*It.*) With exactness.

Ossia (*It.*) Or else, as *ossia più facile*, or else in this more easy way.

Ostinato (*It.*) *Lit.* obstinate, used in the sense of "frequently repeated," as *basso ostinato*, a ground-bass, *q.v.*

Oton. An Indian wind instrument producing only a single sound, employed during the dancing of the Bayadères. The player holds the oton in his left hand while with his right he beats a tambourine fastened to his belt. [Tabor.]

Ottava (*It.*) An octave, as *ottava alta* (to be played), an octave higher, *ottava bassa*, an octave lower.

Ottavina (*It.*) The little or higher octave.

Ottavino flauto (*It.*) Piccolo Flute. [Flute, § 3.]

Ottemole. [Octuplet.]

Ottetto (*It.*) A composition in eight parts.

Ouïe (*Fr.*) The hearing.

Ouvert (*Fr.*) Open, as *à livre ouvert*, at sight.

Ouverture (*Fr.*) [Overture.]

Overblow. A pipe is said to be overblown when the pressure of air forces it to speak an upper partial instead of its fundamental note.

Overtura di ballo (*It.*) An overture in the style of dance-music.

Overture. Overtura (*It.*) **Ouverture** (*Fr.*) An introductory symphony for instruments. It is usual to say that there are four kinds of overtures, two in the ancient, and two in the modern style. This may be true with regard to the limit placed upon the ancient overtures, but it is not entirely so with respect to the modern ones, for they are as various as the fancies of composers, and defy all attempts to reduce them to classification.

Old overtures may be arranged into two classes, the French and the Italian, or to describe them by the names of their reputed inventors into the Lully and Scarlatti overtures. The Lully overture usually opened with a slow and stately movement, by way of introduction. The succeeding theme was treated in imitation in the old Canzona style; if divided into sections, the first section ended in the key of the dominant if it began in a major key, and in the relative major if it began in the minor. All kinds of free imitation were allowed in the next section. After this a "grave" was repeated, and a minuet, or some lively movement concluded the overture. This form is employed by Handel, Bach, Keiser, Telemann, Hasse, Graun, and other composers of the early part of the 18th century, with more or less alterations according to fancy. When Sonata Form began to be understood, overtures in the form described above were more rarely written, but the style was not wholly disused until the commencement of the present century.

The Scarlatti form of overture consisted of three movements or sections; the first and third were quick or lively, and the second was slow. In the slow movement, a change of pace, and also of accent or rhythm was used, and fewer instruments were employed, so that the scoring might afford a strong contrast to the other portions. A near approach to the arrangement of the Sonata and Symphony was thus made, and the Italian overture or "Sinfonia" may be called the immediate precursor of the symphony as it now stands.

The modern overture dates from the latter part of the 18th century, and may be described in a few words as a composition in the form of the first movement of an ordinary sonata or symphony, with the exception that the sections are not always marked for repetition. It has occasionally a short introduction in a slow *tempo*, of a marked and stately character. Its principal themes are often taken from the work it precedes. Sometimes the "free fugue" style is chosen, and occasionally the sonata and fugue form are happily united, as in Mozart's overture "Die Zauberflöte." (Magic flute)

Many modern overtures consist almost entirely of a series of clever contrasts independent of form. These should be described as *medley* overtures, though the medley overture proper includes portions of the chief melodies in an opera. One of the most effective works of this class is the overture to "Guy Mannering" by Sir H. R. Bishop. The invention of the medley overture is attributed to Richard Charke, the son-in-law of Colley Cibber.

Overtures not connected with, or part of a larger work, are known as "Concert Overtures!" but it does not follow that overtures originally written as preludes to a special work, may not be treated as "Concert" pieces, even if the work to which they were attached has ceased to be known.

Oxybaphon (*Gk.*) ὀξυβάφον. *Lit.* a vinegar jar. Hence (like the acetabulum) an earthenware vessel used for experiments as to sound.

Oxypycni (Med. *Lat.* from *Gk.* ὀξὺς, acute, and πυκνὸς, close.) Church Modes having a pycnon or semitone high in the tetrachord *e.g.*:

Lydian.

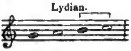

Hypolydian.

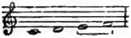

Digitized by Google

P.

P. Abbreviation of *piano*, soft.

Paduana, Padouana. [Pavan.]

Pæan, παιάν. The ancient choral song addressed to Apollo, named after its burden, ἰὼ παιάν. It was sung sometimes before battle, and sometimes after a victory. Donaldson thinks it probable that it was at first accompanied on the phorminx, which, however, was afterwards superseded by the flute. From the ancient Pæan sprang the gymnopœdic, pyrrhic, and hyporchematic dances.

Pæon. A foot consisting of one long syllable and three short. The position of the long syllable can be varied in four ways, hence the Pæon is said to be primus, secundus, tertius, or quartus. [Metre.]

Pair of Organs. The word pair simply means a *set*, and it is used in this sense in the expression a "pair of stairs," or a "pair of scissors." It was formerly in more general use than now, and "pair of cards," and "pair of beads," are met with in the old poets. A "pair of organs" means, then, an organ having a complete set of pipes.

Palalaika. [Balalaika.]

Palco (*It.*) The stage of a theatre, &c.

Palilia. The rustic dance at the Festival of Pales, in which the shepherds leapt round and over burning straw. It may have been to some extent the origin of the May-pole dance.

Pallet. [Organ, § 4.]

Palmula (*Lat.*) Manual of an organ.

Pambe. A small Indian drum.

Pandean-pipes. [Pan's-pipes.]

Pandoura. [Bandore.]

Pan's-pipes, or mouth-organ. The earliest form of a compound wind instrument, undoubtedly the precursor of the organ. It was the σύριγξ of the Greeks, *fistula* of the Romans, and probably the *ugab* of the Hebrews. It was formed of seven, eight, or nine short hollow reeds, fixed together by wax, and cut in graduated lengths, so as to produce a musical scale. The lower ends of the reeds were closed and the upper open and on a level, so that the mouth could easily pass from one pipe to another.

Pantaleone. An instrument invented by Hebenstreit in 1705. It was a sort of dulcimer, but strung with catgut and metal strings in two series.

Pantalon (*Fr.*) One of the movements of a quadrille.

Parrallelbewegung (*Ger.*) [Parallel Motion.]

Parallelen (*Ger.*) Consecutives.

Parallel motion. The movement of two or more parts at fixed intervals, as thirds, sixths. Parallel fifths are under certain limitations forbidden. [Consecutives.]

Paramese (*Gk.*) παραμέση. [Greek music.]

Paranete (*Gk.*) παρανήτη. [Greek music.]

Parhypate (*Gk.*) παρυπάτη. [Greek music.]

Parlando, Parlante (*It.*) In a declamatory manner, as if speaking.

Parlante (*It.*) *Lit.* speaking. (1) In the style of a recitative; (2) played in the style of vocal recitation.

Paroles (*Fr.*) Words; as, *Sans paroles*, without words.

Part. (1) So much of a piece of music, or work, as is performed by any one voice or instrument; (2) Division of a work.

Parte (*It.*) Part. *Parte cantante*, vocal part. *Partie* (*Fr.*)

Partial tones. Those simple sounds which in combination form an ordinary sound and cause its special quality of tone. [Acoustics.]

Partimenti (*It.*) Figured basses, used as exercises.

Partition (*Fr.*) **Partitur** (*Ger.*) [Score.]

Partitura, Partizione (*It.*) [Score.]

Part-Music. Music in two or more parts, performed by more than one person.

Part-Song. A vocal composition, having a striking melody harmonized by other parts more or less freely, but from which counterpoint is for the most part excluded. The Part-song owes its origin to the habit prevalent among the Germans of adding simple harmonies to their "Volksgesänge." On the model of these harmonized airs, fresh compositions have from time to time been produced, sometimes with a little more elaboration in their construction than their models, in such numbers, that there is now an extensive literature of this kind of music. Almost every town in Germany has its "Singverein," "Liedertafel," or "Männergesangverein," all of which largely encourage this special class of composition. From Germany Part-songs have been imported into England, and our native composers have produced a large number, many of which are of great merit. The Part-song holds now the position which

Digitized by Google

the glee held in this country from the middle of the last century to the beginning of this, and which the madrigal held from the middle of the sixteenth to the middle of the seventeenth century. The simplest form of Part-song is that in which the same music (*tutti*) is repeated for each verse of the words; the most elaborate that in which soli parts occur, or a separate solo accompanied by the other parts, either *pianissimo* or *con bocca chiusa*. Part-songs may be for soprani, contralti, tenori, and bassi, or for men's voices only. In the latter case, the upper part is in Germany sung by high tenors, but in England, where the alto voice is still cultivated by alti, the score includes an alto, two tenor parts, and a bass part, instead of the ordinary two tenor and two bass parts of those by German authors. A Part-song differs from a madrigal in its exclusion of contrapuntal devices; from a glee, in its being sung by many voices instead of by one only to each part. The treatment of the chords and the harmonical progressions sufficiently stamp the madrigalian style, while the peculiar melodiousness of the inner parts of a glee, coupled with its special adaptability for chamber use, mark its character distinctly; and on the other hand, the Part-song by its strong outlines and modern harmonization can easily be distinguished from a glee.

Pas (*Fr.*) A dance, as *pas seul*, *pas de deux*, a dance by one or two performers. A step, as *pas redoublé*, a quick step, or march.

Paspy. The English name for the dance *Passe-pied*, called also *Passa-mezzo* by the Italian and *Paspie* by the Spanish writers. It was a precursor of the minuet, some of the tunes called by the title Paspy resembling the minuet in rhythm and measure. Hawkins says it "is said to have been invented in Bretagne, and it is in effect a quick minuet." The old English writers call it passa-measure, passy-measure, passing-measure, or simply measure. It was a favourite dance in the time of Queen Elizabeth, and from the fact that examples exist by writers as late as Purcell and Croft, it could not have been out of fashion in their time.

PASPY. H. PURCELL.

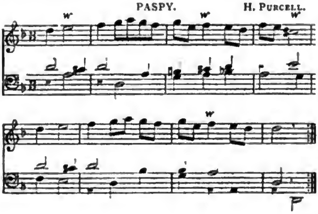

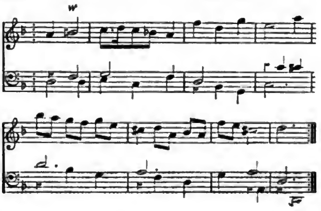

THE ENGLISH PASPY.

CROFT, c. 1700.

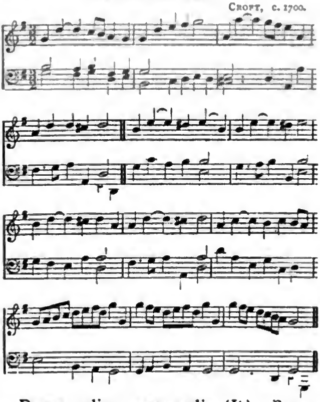

Passacaglio, passacaglia (*It.*); *Passacaille, passe-rue* (*Fr.*); *Passacalle* (*Sp.*); *Gassenhauer* (*Ger.*) A dance in triple measure, constructed on a ground bass. Bach has used this form for one of his grandest compositions for the organ. [Gassenhauer.] [Chaconne.]

Passage. (1) A phrase of music. (2) A figure. (3) A run.

Passage (*Fr.*) [Passage.]

Passage-boards. Boards placed in different parts of an organ, on which the tuner can walk, and from whence he can reach pipes or mechanism.

Passaggio (*It.*) A passage or phrase, as *passaggio cromatico*, a chromatic passage. Also a run, or rapid succession of notes.

Passa-mezzo (*It.*) [Paspy.]

Passe-pied. (*Fr.*) [Paspy.]

Passing discord. [Passing note.]

Passing note. A note not essential to harmony, forming an unprepared discord, which is not objectionable because it is a fragment of a scale. It is a necessary characteristic of a passing-note, that it should have a degree of the scale on each side of it. Passing-notes having degrees of a diatonic scale on each side are said to be *diatonic;* those having degrees of a chromatic scale on each side, are said to be *chromatic*.

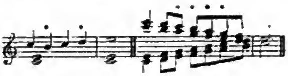

Diatonic passing notes.

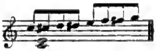

Chromatic passing notes.

Passing notes generally occur on unaccented parts of a bar. When unessential notes occur on accented parts of a bar, they are generally called *auxiliary notes*, *e.g.*:

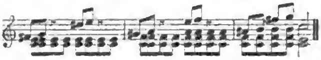

Passionatamente, passionate, passionato, con passione (*It.*) In an impassioned manner.

Passione (*It.*) A cantata founded on the incidents of the Passion, or on the seven last words.

Passion Music. Music set to the narrative of our Lord's Passion in the Gospels. Dramatic representations of the subject date from a very early period, there being still extant a play ascribed, though somewhat doubtfully, to Gregory Nazianzen, Bishop of Antioch. It is in 4th century Greek, and is entitled "*χριστός πασχῶν*." This, like many of the religious plays of a subsequent date, was designed only to be spoken as in ordinary dramatic representations, and is independent of musical aid. For, although music was occasionally employed to add to the attraction of the performance of these Mysteries or Miracle plays in early times, it was only in a secondary manner, and not so connected with or inseparable from the action as in still later times, or in the form now known. Passion plays are still occasionally performed in remote continental places, with music as an accessory.

The dramatic performance of the Passion arose in imitation of the ancient custom, still observed in the Roman Catholic Church of dividing the recital of the Gospel of the Passion in Holy Week between two, three, or more readers, assigning those parts which reproduce the words of the people (turbæ) to the congregation or choir. To one priest was assigned the part of Jesus, to others those of Pilate, Judas, the High Priest, &c., and the connecting portions of the narrative by a narrator. All these parts were recited according to the rules of the *accentus ecclesiasticus*, while the turbæ or people's part was delivered in monotone.

When the Passion was played at other times than during the course of the service, it is supposed that the words were spoken without the aid of music. So popular did these plays become that as early as 1264 a company of monks was formed for the express purpose of performing the "Sufferings of Christ." Bayle, Burney, and others give a series of dates of several notable mediæval performances of the Passion, for example, at Padua in 1243, at Friuli in 1298, at Civita Vecchia in the same year, at Paris before Philip le Bel, and our own Edward II. in 1313, at St. Maur near Paris in 1404, at Constance during the Council in 1417, by the English fathers, at Paris, on the entrance of the Kings of England and France into that capital in 1420, in the same city in 1424, in 1431 on the occasion of the entry of Henry VI. of England, as King of France, and at Poitiers with great magnificence in 1486. In the Imperial library at Paris is preserved a copy of "Le Mystère de la Passion Jesus Christ," printed on vellum by Antoine Verard, 1490, in which a MS. note describes the performance of this play at Metz and at Veximiel, with some curious details respecting the representation at the latter place. In Rymer's "View of Tragedy," there is a quotation from the Register of the Parliament at Paris in which the Procureur General du Roi lays a complaint against four laymen for having undertaken to perform *Christ's Passion* and the *Acts of the Apostles*, "with a company of illiterate and incompetent men; that they had moreover introduced apocryphal matters, and interlarded their plays with drolls and farces, so as to spin out the play and make it last six or seven months; that they had played for lucre, and made their performance so attractive that nobody went to church, charity grew cold, and immoral excesses were occasioned." Their proceedings were stopped during the King's pleasure, and they were condemned to pay a fine of 800 livres out of their profits.

The reformation witnessed, among the many important changes to which it gave rise, the modification of the representation of the Passion. Luther encouraged the continuance of the practice, and a development of the ancient custom by which the several incidents told in the gospels were recited by different individuals, gave opportunities to the composers of the period to set music of

Digitized by Google

their own to the story, in the place of the traditional church tones, and by this means the story was told, not only at the times commanded by the ritual of the church, but at other periods, in other places, and on other occasions. One of the first versions of the text of these improvements was printed at Wittenberg in 1573, with the music for the recitative and the choruses for the turbæ in four-part harmony. The author of this interesting specimen of composition is unknown, though some writers attribute it to Antonio Scandell. Fifteen years later, namely, in 1588, Bartholomæus Güse, or Gese, produced a more elaborate composition, in which the expressions uttered by the priests, people, or disciples, were called by the general name "Turbæ." These were set in harmony or plain-song and sung in chorus. The words of our Lord, of Peter, Pilate, Judas, the High Priest, &c., were usually sung by separate solo voices; but in the work of Bartholomæus Gese, a different practice is resorted to. The words of the maid-servant are set for *two* voices, those of Peter and Pilate for *three*, those of the Saviour for *four*. In this work, the words of Holy Scripture alone are used as the text.

There was little attempt to vary the old plain-song in the early settings of the Passion, until Heinrich Schütz, or as he sometimes styled himself in accordance with the fanciful custom of his time, Henricus Sagittarius, introduced a new feature into the Passion music; that of the reflective chorus, as supposed to be sung by the "Christian community,"—and while the greater part of his harmonies are based upon the old church modes, there is a noteworthy attempt in his recitatives to free himself from the *accentus ecclesiasticus*, and to make them more independent. The design thus suggested was carried out by Johann Sebastiani, who was organist and Kapellmeister at Brandenburgh, in 1672, to whom also the introduction of orchestral, in the place of an organ accompaniment, is due. A still further deviation from the ancient plan was made by Reinhard Keiser at the beginning of the 18th century. Taking, instead of the gospel narratives of the Passion, original poems on the subject as his libretto, he made his "Passion Music" of like character with his many operas. In one of these works, "Der Blutige und Sterbende Jesus," the libretto by Menantes, he employs the term "soliloquia" to describe the reflective or interjaculatory passages set in "arioso recitative," so frequently employed later by John Sebastian Bach.

The words of Reinhard Keiser's "Passion" were written by Brockes. a Hamburg poet, who added to the Bible narrative, soliloquia, or reflections for the "Daughter of Sion," and "The Believing Soul." Telemann, Mattheson, and Handel set Brockes' poem to music, and traces of the influence of the poem are to be found in Bach's Matthew "Passion." A statement is made by Rietz that in the MS. score in Bach's handwriting certain solos, and solos with chorus, are marked as though they were to be sung by the Daughter of Zion, or Zion, or the Faithful. As there are several pieces in both the existing works on the Passion by Bach (the St. Matthew and the St. John), similar, if not identical with Brockes' soliloquia, there is a reasonable ground for believing the statement of Rietz. The credit of the introduction of the chorale has been assigned to Bach, but although he employed it freely, Telemann, Mattheson, and Handel, also used chorales in their works on the Passion. The earlier writers, Gese, Schütz, Keiser, and Sebastiani did not use the chorale as Bach probably did, with the design of enlisting the voices as well as the sympathies of the congregation. Bach selected from among a wealthy store of well known hymn-tunes those which had the greatest influence over the minds of German worshippers, tunes set to various words inseparably associated with the sentiments, affections, life, and citizenship of the people.

Passy-measure (*Old Eng.*) A corruption of the word *passa-mezzo*. [Paspy.]

Pastete (*Ger.*) [Pasticcio.]

Pasticcio (*It.*) *Lit.* a pie, or pasty. *Pastiche* (*Fr.*), *Pastete* (*Ger.*) An opera, cantata, or other work, the separate numbers of which are gleaned from the compositions of various authors, or from several disconnected works of one author.

Pastiche (*Fr.*) [Pasticcio.]

Pastoral. (1) A simple melody in six-eight time in a rustic style. (2) A cantata, the words of which are founded on pastoral incidents. (3) A complete symphony, wherein a series of pastoral scenes is depicted by sound-painting, without the aid of words.

Pastorale (*It.*) [Pastoral.]

Pastorella (*It.*) A short pastoral.

Pastorelle (*Fr.*) [Pastorella.]

Pastorita. A shepherd's pipe. *Fistula pastoricia* (*Lat.*) An organ stop, the Nachthorn.

Pastourelle (*Fr.*) One of the movements of a quadrille.

Pateticamente, Patetico (*It.*) In a pathetic manner.

Pathétique, Pathétiquement (*Fr.*) In a pathetic manner.

Patimento (*It.*) Suffering, as *con espressione di patimento*, with an expression of suffering.

Digitized by Google

Pauken (*Ger.*) [Kettle-drums.]

Pausa (*Lat.*) A sign for silence. A rest. Ancient authors allowed five kinds of rests: the long rest being a perpendicular stroke drawn from a line to the second line below; the breve rest, a similar stroke from line to line; the semibreve rest, a stroke commencing from any line, and descending half way between the lines; the minim rest, the reverse of the semibreve, that is, a stroke ascending half way between the two lines; the semiminima rest (crochet), a line equal to half the distance between the lines, but bent to the left. These are, of course, practically the same as modern rests, but the way in which they were written was far from preventing confusion. There were also varieties of form or position to indicate rests in triple or common time.

Pausa (*It.*) A rest or pause. *Pausa generale*, a rest or pause for all the performers.

Pause (*Fr.*), **Pause** (*Ger.*) (1) A rest or pause. (2) A bar's rest.

Pavan. A dance tune of a stately character, deriving its title from Padua, where it is said to have been invented. Others find the origin of the word in "Pavo," a peacock, because it was danced with "such circumstance of dignity and stateliness."

Morley's "Introduction to Practicall Musicke," speaks of it in these terms :—

"The next in grauety and goodnes vnto this is called a pauane, a kind of staide musicke, ordained for graue dauncing, and most commonlie made of three straines, whereof euerie straine is plaid or song twice, a straine they make to containe 8, 12, or 16, semibreues, as they list, yet fewer then eight I haue not seene in any pauan."

"After euery pauan vsually set a galliard."

"The Italians make their galliardes (which they tearm *saltarelli*) plaine, and frame ditties to them, which in their *mascaradoes* they sing and daunce."

By this it would appear that the galliard and the pavan, when played upon the virginals, harpsichord, or spinet, formed a primitive sort of suite or sonata.

That the pavan was not used entirely for the dance may be gathered from the title of a work published by John Dowland, "Lachrymæ or Seaven Teares, figured in Seaven passionate Pavans, with divers other Pavans, Galliards, and Almands."

The pavan, when in duple time, is not unfrequently spoken of as a "quadron or quodran pavan." The "saltarello," of which Morley has made mention, was in triple measure. When the pavan was played faster it was called a passamezzo.

PAVAN, TRANSLATED FROM TABLATURE IN WOLF HECKEL'S "LAUTTEN BUCH," 1562.

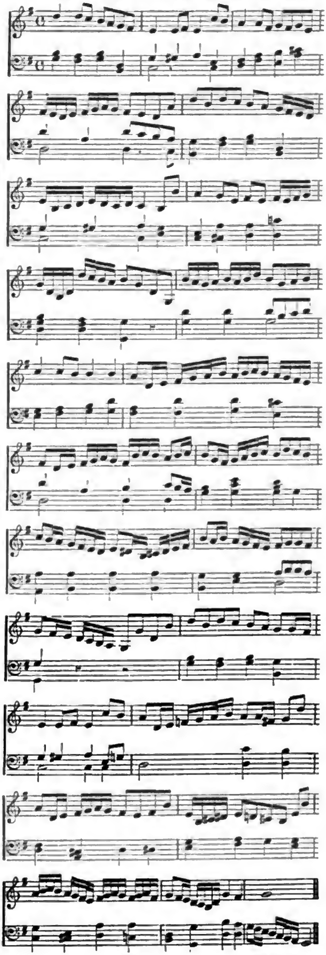

Pavana (*It.*) [Pavan.]

Pavane (*Fr.*) [Pavan.]

Paventato, Paventoso (*It.*) With an expression of fear.

Pavillon (*Fr.*) The bell of a horn, or other instrument of a like kind. *Flûte à Pavillon* (*Fr.*) an organ stop, the pipes of which are surmounted by a bell. *Pavillon chinois*, an instrument consisting of little bells attached to a staff.

Pectis or Pektis. A Greek instrument of the lute or dulcimer kind, variously described.

Ped: Abb. (1) of pedal. In pianoforte music a direction to press down the pedal which raises the dampers. (2) Of pedals. In organ music a direction that the lowest part is to be played by the feet.

Pedal. Any projecting piece of metal or wood which is to be pressed down by the foot. (1) On the pianoforte there are usually two pedals, one of which enables the performer to play only on one string (*una corda*); the other, to remove the dampers (*sordini*). (2) On the organ there are combination pedals, which alter the arrangement of the registers, and a pedal clavier or keyboard, on which the feet play. (3) On the harp there are pedals, each of which has the power of flattening, sharpening, or making natural, one note throughout the whole compass of the instrument.

Pedalclaviatur (*Ger.*) The pedal keyboard of an organ.

Pedal coupler. An accessory stop of an organ, by means of which the pedal keys are enabled to draw down the keys of a manual.

Pedale (*It.*) (1) A pedal key of an organ. (2) A pedal note, or pedal point. [Fugue.] (3) The pedal of a pianoforte, which raises the dampers.

Pedalflügel (*Ger.*) A pianoforte, to which a set of pedal keys is attached.

Pedaliera (*It.*) The pedal keys of an organ.

Pedalion. A set of pedals acting upon strings, producing notes of a deep pitch, so constructed as to be capable of being used with a pianoforte.

Pedal-pipes. The one stop which was often, in former times, in this country the sole representative of an independent pedal organ. It consisted of open wooden pipes of a large scale.

Pedal Point. [Fugue.]

Pedal Sound-board. The sound-board containing the pipes belonging to the pedal organ.

Pedicula, Scabella, or Scabilla (*Lat.*) [κρούπεζα.] [Croupeza.]

Pedarii, Podarii, and Pedicularii. The beaters of time in the performance of Greek music. It was done with the foot, which was furnished with wooden or iron sandals for the purpose. [Croupeza.]

Peg of an Instrument. A small round piece of wood or metal, placed in a hole, or two holes, so as to be capable of being turned round, and pierced to receive that end of a string or wire which is not fixed. It is prevented from slipping round under the influence of the tension of the string or wire, either by being made slightly tapering in form, and then forced into its hole as in a violin; or by being made with serrated edges, as in a pianoforte. The outer end of the peg expands into a broad flat surface which can be easily turned round between the thumb and forefinger, or has a four-sided head on to which a key can be fixed.

Penillion singing. A Welsh custom of singing improvised verses on a given theme to a melody either well-known or then and there learnt from the harper who accompanies the Penillion. There is a similar custom among the Hungarian gipsies called the "Verbunges." [Dutch concert.]

Penorcon. An obsolete instrument of the guitar family, somewhat broader and shorter than the Pandora, with a very broad neck over which passed nine brass strings, which were played upon by the fingers.

Pentatone. An interval of five whole tones, an augmented sixth.

Pentatonic scale. The name given by Carl Engel to the ancient musical scale which is easiest described, as that formed by the black keys of the pianoforte. It consists of the 1st, 2nd, 3rd, 5th, and 6th degrees of a modern diatonic scale, *e.g.*:

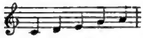

Per (*Lat.*) By; as, *per diminutionem*, by diminution; *per inversionem*, by inversion.

Per (*It.*) By or for; as, *per il violino*, for the violin.

Percussion. An ingenious contrivance whereby a hammer strikes the tongue of a reed and sets it in motion simultaneously with the admission of air from the windchest, thus securing the rapid speech of the reed. Were it not for the percussion, the reed would be only gradually set in motion by the admission of the current of air, and the sound would not instantly follow the striking of the key. It is commonly used in harmoniums, but has also been applied to the largest reeds of an organ.

Percussion of a discord. The striking of a discord, which takes place after its preparation, and which is followed by its resolution.

Percussione (*It.*) The striking of a chord.

Perd. Perden. *Abbreviations of*

Perdendo, Perdendosi (*It.*) Losing time and power, implying both *dim.* and *rall.*

Digitized by Google

Perfect. Complete. (1) Perfect cadence, an authentic or plagal cadence. [Cadence.] (2) Perfect concord, a common chord in its original position. (3) Perfect consonance, the consonance produced by the intervals fourth, fifth, or octave. (4) Perfect interval, one of the divisions of intervals. [Interval.] (5) Perfect time. An old name for triple time.

Performing Rights. [See Copyright.]

Perijourdine. A song-dance, a ballad, named after the district in which it was used.

CHANSON À DANSER DE SARLAT EN PERIGORD.

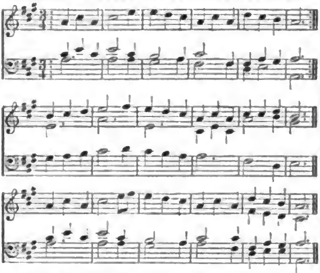

Period, *Periode* (*Fr.*), *Periodo* (*It.*) A complete musical sentence. [Form.]

Periodenbau (*Ger.*) The construction of musical periods.

Periodic Vibration. [Acoustics, § 9.]

Period of Vibration. [Acoustics, § 3.]

Perlé (*Fr.*) *Lit.*, pearled. *Cadence perlée*, a brilliant cadence. A *pearly touch* is an expression of our own for a delicate and bright touch in playing the pianoforte.

Perpetual canon. A canon so constructed that it may be repeated constantly without a break in the time or rhythm. [Canon.]

Per recte et retro. A species of imitation in which the antecedent is repeated at the unison, reading the notes backwards. The following chant by Dr. Crotch is a simple example. The numbers over the bars will show the retrogressive movement.

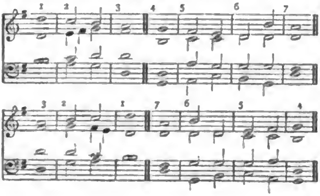

Pesante (*It.*) With weight, or importance. Impressively.

Petite flûte (*Fr.*) **Kleine Flöte** (*Ger.*) Piccolo. [Piccolo flute.]

Petite mesure à deux tems (*Fr.*) Two-four time.

Petteia (*Gk.*) A term of Greek "Melopœia." Perhaps the same as *extensio*, the holding out of a sound, or its frequent repetition to different words or syllables. [Melopœïa.]

Petto (*It.*) The chest. *Voce di petto*, the chest voice.

Peu à peu (*Fr.*) Little by little, gradually.

Pezzi (*It.*) Pieces, as *pezzi concertanti*, concert pieces or *concerti; pezzi di bravura*, pieces for the display of rapid execution and manual dexterity.

P.f., abb. of (1) Pianoforte. (2) *Piano, forte*, soft then loud. (3) *Più forte*, louder.

Pfeife (*Ger.*) Pipe, fife, little flute.

Pfeifenwerk (*Ger.*) The pipe work (of an organ), including both reed and flue pipes.

Phantasie (*Ger.*) [Fantasia.]

Phantasy, pl. **Phantasies** (*Old Eng.*) [Fancies.]

Phonascus (*Lat.*, from *Gk.* φωνασκός). A master who trained the voices of those who were preparing themselves for orators, singers, &c.

Phonometer. An instrument for ascertaining the number of vibrations of a given sound.

Phorminx (*Gk.*), φόρμιγξ. The ancient lyre or cithara, which had probably not more than seven strings. [Lyre.]

Phrasing. The proper rendering of music with reference to its melodic form. A performer who brings into due prominence the grouping of sounds into figures, sentences, &c., is said to *phrase* well. Formerly the phrasing of music was left to be discovered by the intelligence and taste of the performer. Modern writers use various signs as indications of their wishes, such as, the *slur*, *sf*, &c. [Expression.]

Piacere, a (*It.*) At pleasure (the time be modified at the performer's pleasure).

Piacevole, Piacevolezza con, Piacevolmente (*It.*) In a playful or light style. Pleasantly.

Piacimento (*It.*) [Piacere.]

Pianette (*Fr.*) A small pianoforte.

Piangendo, Piangevole, Piangevolmente (*It.*) As if weeping. In a plaintive manner.

Pianino (*It.*) [Pianette.]

Pianissimo (*It.*) As softly as possible, *abbd. pp.*

Piano (*It.*) Softly, *abbd. p.*

Piano à queue (*Fr.*) A grand pianoforte.

Piano carré (*Fr.*) A square pianoforte.

Piano droit (*Fr.*) An upright pianoforte.

Digitized by Google

Pianoforte. (1) The pianoforte is a development of the harpsichord and clavichord, but differs from either in the production of the tone. In the harpsichord and spinet this was effected by quill points or studs of leather acting as *plectra*. In the clavichord by the impact of a brass striker, rather pusher, or *tangent*. In the pianoforte the strings are struck by hammers mechanically, like a dulcimer is struck by hammers the player holds. The dulcimer is the prototype of the pianoforte.

(2) The clavichord may have been the earliest member of the interesting group of piano precursors, from the simplicity of its construction and the certainty that it owed its origin to the mediæval monochord, which had, while retaining the name, become a polychord. The invention, as also that of the harpsichord and spinet, is nowhere recorded. The few facts and scattered notices that have come to us would, in detail, make this article too long. They will be found with the authorities quoted in A. J. Hipkins's "Description and History of the Pianoforte, and of the older keyboard instruments" (Novello & Co., London, 1896). The earliest mention yet found of the clavichord, together with the clavicymbal or harpsichord, occurs in some rules of the German Minnesingers, dated A.D. 1404. The monochord was a vibrating string stretched over a sound-box, and stopped by little bridges moved by hand to obtain the notes required. Increasing the number of strings, of the same length and pitch as in an Æolian harp, and ultimately adding a keyboard to stop the notes with little tangents fixed in the farther ends of the keys, the clavichord was invented. The tangents also determined the vibrating length of the strings at one end, the other end passing over bridges glued to the sound-board, finally one long wooden bridge. A strip of cloth interwoven among the strings behind the tangents instantly damped the sounds when the fingers released the keys and allowed them to rise. The strings were scaled in pairs, like lutes and similar instruments, and by curving the key levers two, three, and even four tangents were made to act upon one pair of strings, such clavichords being in German called *gebunden*, or fretted. It was not until the early years of the eighteenth century that the clavichord had its full number of strings. It was then capable of being tuned in Equal Temperament, and was the favourite *clavier* of Bach, for which he composed his Forty-eight Preludes and Fugues, and, from internal evidence, the Fantasia Cromatica e Fuga. It is the only keyboard instrument upon which the *vibrato*, called in German *Bebung*, is possible.

CLAVICHORD TANGENT.

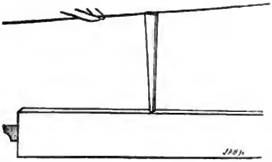

(3) The spinet, or virginal, to use the English name current from the time of Henry VII. to Charles II., was in vogue during the same period as the clavichord, and was a *plectrum* instrument like the harpsichord. The plectrum was usually of quill inserted in a tongue centred in a wooden upright called a *jack*. The key being put down, the jack rose up, forcing the quill point past the string, and setting it in vibration. The jack returning as the key rose, a little cloth damper descended and silenced the note. There can only be one string for a jack; therefore in the spinet or virginal there was only one string for each note.

SPINET OR HARPSICHORD JACK.

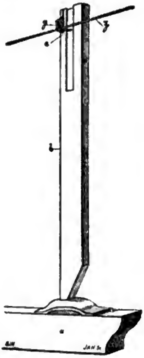

The normal keyboard compass was forty-five keys, apparently from E below the bass clef to C above the treble; but this compass was extended downwards by the *short octave* expedient common in the 16th and 17th centuries to all keyboard instruments, including the organ. By it the lowest E was tuned down to C, F sharp to D, and G sharp to E. The lowest chromatic notes were

Digitized by Google

therefore rejected. Not unfrequently the apparent compass was carried down in the bass to B, which was then tuned G, and C sharp, A. D sharp was usually retained to serve for E flat, a useful note in ecclesiastical music. The oldest spinet known is dated A.D. 1480. Like the harpsichord and clavichord, the spinet remained in use until the opening years of the present century. It was oblong, pentagonal, or hexagonal in shape.

(4) The clavicytherium or upright spinet was raised upon a stand. There is a remarkable and very early specimen in the Donaldson Museum, at the Royal College of Music. It appears to have had brass plectra, which probably preceded the use of quill or leather.

(5) The harpsichord is a long wing-shaped instrument like a grand pianoforte, whence its German name—*Flügel*, a wing. In construction it is a double or treble, rarely a quadruple spinet, the jacks being racked in movable slides, acting with hand stops, analogous to the registers of an organ. The best 17th and 18th century harpsichords had three strings to each note—two tuned in unison and one an octave higher. When there was a fourth string it was an octave lower than the unisons. The two *unisons* are controlled by the stops to the right hand; the *octave*, the *lute*, produced by sounding the first unison nearer the bridge, and the *harp*, a muting or *pizzicato* contrivance, are to the left hand. With the later English harpsichords there are two pedals, the one for the left foot coming into service when the *machine* stop at the left of the keyboards is pushed back; the pedal then removes the first unison and octave registers, leaving on the lower keyboard the second unison, and on the upper the lute only. The right foot pedal is for a *swell*, produced at first by raising a portion of the top or cover, but later by Burkat Shudi's *Venetian swell*, patented by him in 1770, and very soon transferred to the organ. Shudi (Tschudi) was the founder of the house of Broadwood. The earliest known harpsichord is in South Kensington Museum; it is Roman and is dated A.D. 1521. The compass of the harpsichord was gradually extended from the normal four octaves, C—C, the range of the human voice, to five octaves, F—F, the octave string giving yet another octave higher. The famous harpsichord makers of the 17th century were the family of Ruckers of Antwerp; of the 18th, Kirkman, and Shudi and Broadwood, of London.

A RUCKERS HARPSICHORD.

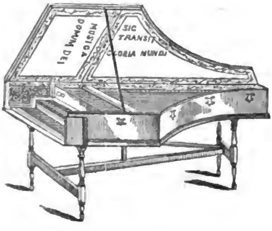

(6) The pianoforte was invented about A.D. 1709, by Bartolommeo Cristofori, a harpsichord maker in Florence. Grand pianos made by him are still existing, dated respectively 1720 and 1726. The former is in the Metropolitan Museum, New York, the latter in the Kraus Museum, Florence. Cristofori named his invention "Gravecembalo col Piano e Forte" (harpsichord with soft and loud), whence the descriptive name, Pianoforte. He solved several important problems—the first was to counteract the strain of thicker strings necessary to withstand the impact of hammers; the second to make up for the weakness caused by the necessary opening between the tuning-pin block, technically *wrestplank*, and the soundboard, to let the hammers rise. The third was the mechanical control of the rebound of the hammer from the strings, technically *escapement*, so that the hammer should not

CRISTOFORI'S ACTION.

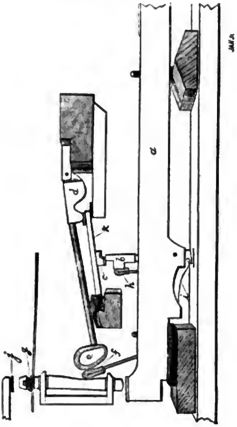

block against the strings and prevent vibration; and, lastly, he invented the *check*, a rest for the hammer tail, to regulate the position of the hammer for another blow, the first step towards *repetition*.

In the diagram *a* is the key; *b* the hopper; *c* the notch for the hopper to raise an escapement lever, lettered *k*, which raises the hammer-butt *d*. The hammer-head is *e*. The spring *i* maintains the position of the hopper in the notch, assisted by the small hopper-check *h*. The hammer-check is *f*; the damper *g*; the damper-stop *j*. It will be seen that Cristofori tried to retain the shallow measure of an Italian single harpsichord. The grand pianos by Silbermann, acquired by Frederick the Great about 1746, and still at Potsdam, have Cristofori's framing, scale, and action (*see* A. J. Hipkins's "History of the Pianoforte," already referred to). The invention of the pianoforte has, in Germany, been claimed for a German organist, Schroeter, who, but not until 1763, published diagrams of actions he contrived between 1717 and 1721. Neither is of much mechanical value, and it does not appear that he ever made a pianoforte, or had one constructed according to his design.

In France a claim has been set up for Marius, a harpsichord maker, who, in 1716, sent models of actions for *clavecins à maillets* (harpsichords with hammers) to the Académie Royale des Sciences, not one of which has those essential features which distinguish Cristofori's invention, or would even compare with Schroeter's simple contrivance.

(7) There have been several epoch-making inventions in pianoforte making. It is not known who invented an early German escapement obtained by raising the hammer until an extension of the hammer-butt in contact with a rail at the back of the action jerked the hammer to the string. Stein, of Augsburg, about 1777, added a hopper escapement which obviated any blocking or jamming of the hammer against the strings, and won thereby the gratitude of Mozart. Stein's action, continued by his successor, Streicher, has not yet gone out of use. Cristofori's invention may have come early to England in a piano made by an English monk at Rome, Father Wood, and such instruments were being made by the second-rate harpsichord makers when J. C. Bach arrived in London in 1759. Burney mentions Americus Backers as one who made them. The first important step in England with grand pianos was when Backers produced his direct lever action, known as the English action, which he brought out in an "Original Forte-Piano," at the Thatched House, in St. James's Street, in 1771. A grand piano of his make has been found at Pistoia, in Italy. It is dated A.D. 1773. From 1777 Robert Stodart made rapid way with Backers' model, and very soon after John Broadwood, by adding the loud and soft pedals and making a separate bridge for the bass strings, completed the wooden grand piano, used by Clementi and the young J. B. Cramer; by Haydn, at that time in London, and Dussek.

In Paris, Erard's first grand piano was made in 1796. He followed at first the German system, but soon began those experiments in *repetition* which culminated in the action patented by his nephew in 1821, the use of which, with some modification, has become general.

GRAND PIANOFORTE ACTION, S. AND P. ÉRARD, PARIS.

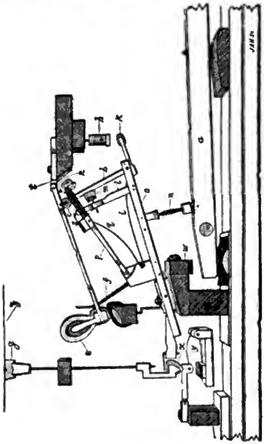

In this diagram the key, *a*, raises by a pilot, *n*, a carriage, *o*, which carries the hopper, *b*, that escapes forward after delivering its thrust to the nut or roller, *c*, beneath the hammer fork, *t*; a prolongation of the carriage forms the escapement lever, *k*, which controls the escapement by the button, *h*. The lower wire spring, *i*, is the hopper spring. With the hammer-head, *e*, and the check, *f*, the single escapement action is complete. The rise of the repetition lever is governed by a small screw, *r*, which acts upon the nose of the lever to cause the second escapement so that the hopper can repeat a blow with the key risen but a little way. The repetition lever, *p*, rises as the upper wire spring, *q*, is bent from the depression of the key. Through this spring it is hardly possible to put the key down without producing sound. The damper is *g*, the damper lifter *x*, and the spring to send it up *y*.

(8) Pianofortes remained to the year 1820 wooden structures, but in that year an invention of metal plates and tubing by William Allen, more for compensation than resistance, taken up by the firm of Stodart, led to a great revolution; the movable tubing suggesting the fixed metal bars of Erard, patented in Paris, 1822, and the movable plates, the fixed plate introduced by Broadwood in 1821, and patented by him in combination with bars, 1827. Thus the built-up wood and iron grand piano came to an advanced stage of efficiency.

(9) As the harpsichord had led up to the grand pianoforte, the clavichord suggested the square (oblong) one. Indeed, many clavichords were, by the addition of a further bridge and a simple jack, made into square pianos, the depth of case allowing for this unsatisfactory change. It is said Friederici of Gera made the first square piano about the year 1758; a specimen of his instrument is not however forthcoming. Zumpe, in London, two or three years later, brought out the small English square piano which was extensively made by him, and his successors and rivals, until the last years of the century, when Geib's hopper action and Broadwood's new framing, with the wrestplank transferred from the old clavichord position to the back of the case, entirely superseded Zumpe's model. Fétis began his study of the piano on a Zumpe square one, dated 1762. Erard in 1777 began in Paris with this model. With time and the requirements of the players, by whom power is always sought, the square piano became enlarged; the sound-board was extended, the check action introduced. But it gave way in the sixties before the smaller grand and upright piano, and is no longer made.

(10) The use of iron was initiated in American pianos by Alphaeus Babcock, of Boston, who, in 1825, invented a frame in a single casting for square pianos. This invention was further developed by Meyer and by Chickering, and ultimately Steinway of New York combined an over or cross-strung scale with it, and thus in 1859 was established another epoch in pianoforte construction. The success of this principle in the Paris Exhibition of 1867 led to the general adoption of it by German pianoforte makers, with such modifications as may be found in the pianos of Bechstein, Blüthner, Bösendorfer, and others. It now prevails in England, and the house of Broadwood has, since 1888, gone yet farther by the invention of a steel frame in which barring is entirely done away with, thus returning to a desirable simplicity of construction with concurrent advantages.

(11) The upright piano at first followed the upright spinet and harpsichord in being merely the horizontal instrument turned up upon its broad end and placed upon a stand. The Italian expert, Signor Ponsicchi, has met with an upright grand piano made in Italy by Domenico Del Mela da Gagliano as early as 1739. It has a good action, adapted from Cristofori. There is one at Brussels, made in 1745, by Friederici of Gera, but with an original, simple action. But a disposition of the keyboard, so that the strings could descend below it, rendering a stand unnecessary, was the invention of Hawkins, an Englishman, in 1800. He was at that time living in America, but took out a patent in England for this and other new ideas in construction that have since been extensively drawn upon.

UPRIGHT PIANOFORTE ACTION.

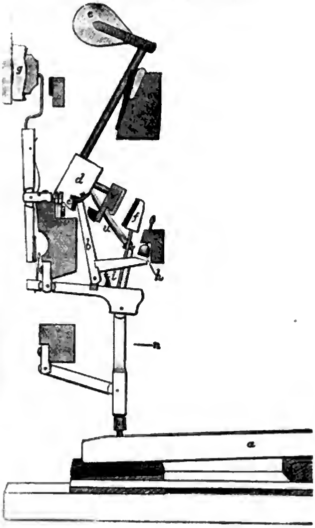

Southwell planned the cabinet, a high upright piano, in 1807. Wornum followed with low upright instruments in 1811-13, patenting, in its main features, the now universal upright action in 1826. The diagram on the previous page shows Broadwood's model of that mechanism. It is hardly necessary to say that in making these actions there are variations, due to the intention or experience of each maker who is not a mere copyist.

a is the key, *b* is the hopper, *c* the notch, *d* the hammer-butt, in which is inserted the shank of *e*, the hammer-head; *f* is the check, *g* the damper, *h* is the set-off button held by a rail, *i* is the hopper spring—in this diagram the spiral spring—invented by the late A. Bord. In the actions of the higher upright instruments there is necessarily a pilot, *n*, to raise the hopper and hammer to the height required for their service. The special merit of Wornum's invention is the tape *u*, which is tightened by the rise of the hammer when the key is struck, so as to pull the butt back, and, of course, the hammer with it, when the key returns, to assist the repetition; the button *v*, projecting from the hammer-butt, falling into check, as it is called—in other words, preventing, by the contact of check and button, the hammer falling too far.

(12) The original compass of the piano keyboard, five octaves, has been increased to seven octaves and more. The high tension of modern pianofortes is permitted by the remarkable improvement in tenacity of cast steel music wire, the tension of a modern concert grand being little less in the aggregate than thirty tons. The pedals, sympathetic aids to vibration, remain much as at first introduced; but a third pedal is occasionally found that confines prolonged vibration to a note required, say as a *point d'orgue*, although useful, it is not much in request. The two pedals indispensable to the complete piano are known as the *piano* and *forte* pedals. The left foot presses upon the *piano* or soft pedal, which, in grand pianos always, in upright pianos occasionally, shifts the keyboard action bodily, thereby removing the attack of the hammers from three to two strings, and leaving one untouched, but not silent, for it vibrates through the influence of the sounding strings it is in tune with. In many upright pianos a soft pedal is contrived by mechanically dropping a strip of cloth, or thin felt, between the hammers and the strings, making, as it were, a thicker clothing for the hammers, to diminish the amount of tone. This plan goes by the name *celeste*. There is another kind of soft pedal now much used, obtained by shortening the radius of the hammer blow, which can only be said to replace the due modifications for *piano* and *pianissimo* of the player's touch. The right foot or *forte* pedal does away with the use of the dampers throughout. It should not, however, be named the *forte* pedal, as it is not less used in *piano* and *pianissimo* playing. It should be invariably described as the *damper* pedal. While it is held down by the foot it makes the instrument an undamped one, and its employment allows any note the prerogative of calling forth by sympathetic vibration all higher notes that are in unison with its own partial tones, and the partial tones of all lower notes that happen to be in unison with its own note or prime. The player controls all this wealth of sympathetic vibration with the damper pedal, bringing down the dampers upon the strings whenever there is any confusion or any discord displeasing to the ear. A study of the indications given by the best modern composers and editors will show the student the refinement in controlling dissonance the mastership of the pedal brings. The high tension and more equally balanced strain of modern pianos increases the damper pedal effect, we may say, by the square; it originates a new manner of playing, and gives the frequent use of the soft pedal a wider justification.

Pianoforte score. A score of a vocal or instrumental composition, under which is written in two lines a condensed form of the harmonies for the use of a pianoforte.

Pianograph. An ingenious machine, which, on being attached to a pianoforte, inscribed what was played. Invented by M. Guérin.

Piatti (*It.*) [Cymbals.]

Pibroch. The wild, irregular martial music of the Scottish Highlanders played on the bagpipe. The term is also figuratively employed for the instrument itself.

> "Cast your plaids, draw your blades
> Forward each man set;
> Pibroch of Donuil Dhu
> Knell for the onset."—*Scott.*

The form of a pibroch is that of an air with variations, the air being twice played: at the beginning and at the end. The word is derived from the Gaelic "Piob," a pipe; a foreign word to that language showing a late introduction. [Bagpipe.]

Picciolo (*It.*) Small; as, *violino picciolo*, a small-sized fiddle.

Piccolo. An organ stop of 2 ft. length—the pipes are of wood, the tone bright and piercing.

Piccolo flute, or piccolo. A small flute, having the same compass as the ordinary orchestral flute, but whose sounds are one

Digitized by Google

octave higher than the notes as they are written. Its compass may be said to be from

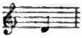

to

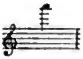

or even one note higher, but the lowest octave is so weak as to be far more effective when played by the higher octave of the common flute, and the upper notes are so shrill as to be only available in *fortissimo tutti* passages, or for peculiar and striking effects. Both Gluck and Beethoven use it in representing a storm, the former in his "Iphigenia in Tauris," the latter in his "Pastoral Symphony." Spontini combines the short piercing scream of the piccolo with the crash of cymbals in his "Danaïdes." The tendency of modern orchestral writers is to use the piccolo too often, and in inappropriate passages. It should be reserved for rare and special effects.

Picco pipe. A small pipe having two ventages above and one below. It is blown by means of a mouthpiece like a *flûte à bec* or whistle; and in playing, the little finger is used for varying the pitch by being inserted in the end. The player, Picco, after whom it was named, produced a compass of three octaves from this primitive instrument.

Piccorn. *Lit.*, horn-pipe. [Horn-pipe.]

Pièce (*Fr.*) A piece, as *suites de pièces*, sets of pieces. [Suite.]

Pieds (*Fr.*) The feet, as, *avec les pieds*, with the feet, *con pedale*.

Pieno, Piena (*It.*) Full; as, *a piena orchestra*, for a full band.

Pietosamente, Pietoso (*It.*) With pity, tenderly.

Piffaro, *or* **Piffero** (*It.*) The old form of the oboe, still in use in some districts of Italy and the Tyrol, formerly called the schalmy. [Oboe.]

Piffarari, *or* **Pifferari** (*It.*) Players on the piffaro.

Pifferino (*It.*) A little piffero.

Piffero (*It.*) [Piffaro.]

Pileata (*Lat.*) Capped (of an organ pipe). Stopped. *Pileata major*, 16 feet stopped. *Pileata minor*, 4 feet stopped. *Pileata diapente*, stopped quint.

Pincé (*Fr.*) Pinched. (1) A direction in French music equivalent to pizzicato, *q.v.* (2) The name of an ornament in harpsichord playing, indicated by the signs here shown:

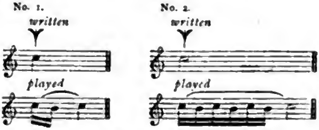

The first was called *pincé simple*, the second *pincé double*. The note of ornament of the pincé was below the note written, while the added note in the *tremblement* was above.

Pincé bemolisé (*Fr.*) A pincé or trill with a flattened note. An ornament in harpsichord playing.

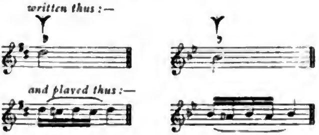

Pincé dièsé (*Fr.*) A pincé or trill with a sharpened note

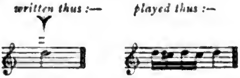

Pincement (*Fr.*) [Pincé.]

Piqué (*Fr.*) On string-instruments a kind of bowing, by which a *staccato* is produced by allowing the bow to lie perfectly still on the string between each separate note. It is written with dots and a slur, thus:

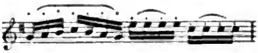

It is also called *spiccato*.

Piquiren (*Ger.*) To play *piqué*.

Pitch. (1) The position of a sound with reference to the number of vibrations which cause it. The relative height of a sound. An early clue to a definite pitch is to be found in "Musica getutscht und auzgezogen," by Arnold Schlick, Basle, 1511. He gives a linear indication by which the late A. J. Ellis ("History of Musical Pitch," 1880, p. 305) was enabled to determine a low pitch of A. 374 and a high pitch of A. 506, according to the organ pipe being used for F or C. which effected a transposition of a fourth to help singers of the Church Modes (see

Digitized by Google

Plagal). A. 506 is deduced from Praetorius for the pitch of the old 14th century organ that was in his day at Halberstadt. Our oldest Cathedral music suits this very high pitch and was probably composed for it. When the necessity for these extremes no longer existed, they converged until they merged in the compromise pitch of the 18th century, A. 422·5, shown by Handel's tuning fork. This pitch was approximately the Chor Ton of Praetorius ("Syntagma Musicum," Wolffenbüttel, 1618), but the "Kammer Ton" he gives for voices and instruments is a whole tone higher; it may best be taken as a tone above French pitch = A. 489. There is an organ at Hamburg still at this pitch; it had a stop a minor third lower and, according to Praetorius, this low pitch was also used in England. The introduction of the violin seems to have banished the high pitches, and the A. 422·5 lasted approximately until the foundation of the London Philharmonic Society in 1813. From 1816 a gradual rise in pitch was accompanied and perhaps caused by the invention and remarkable development of the brass instruments, so that the pitch of Costa's Philharmonic Orchestra (1846-54) was A. 452·5. About this time the pitch in Paris was A. 448, which, being found too high, the Diapason Normal was introduced and made law in France. It is officially stated as A. 435 at 15° Cent.: (59° Fahrenheit). Concert rooms and opera houses being usually nearer 68° at time of performance, the rise in pitch of the wind instruments owing to the increased temperature compels a correction of this vibration number, and the Philharmonic Society, in adopting the Paris pitch in 1896, accepted the present writer's suggestion of a fork tuned to A. 439 (C. 522), through which the band with diapason normal instruments, including the organ, plays in good tune. The various pitches accepted by the Philharmonic Society of which we have forks have been:—

(1813)	A. 423·3
Sir George Smart's (date uncertain)	A. 433·2
Sir Michael Costa's (1846-54)	A. 452·5
Maximum (1874)	A. 454·7
(1897)	A. 439·0

It will be observed that A. 439 is exactly intermediate between 423·3 and 454·7. This was not intentional. A medium pitch, A. 446, introduced by Messrs. Broadwood in 1849, has no longer any vogue. The Society of Arts has abandoned a medium pitch in favour of the French. The reader is referred to A. J. Ellis's translation of Helmholtz (2nd Edition), which contains the gist of his Lecture, now out of print, in an appendix.

(2) As to the measurement or exact record of vibration numbers, a true determination has only been made in the present century, and practically since the French "Diapason Normal" was settled. Before that date (1859) there was no certainty, as the methods adopted were upon trial and unproven. Mersenne (1648) with a stretched string and monochord measures lead the way. Euler and also Bernouilli in the 18th century endeavoured to work out the problem mathematically, but, from the inherent difficulties connected with it, were unable to give absolutely accurate results. Delezenne (1854) returned to the string with nearer measures, nearer than Griesbach (1860), who had a method of his own with the string, but failed to ensure accuracy even within rather wide limits. His forks prepared for the Society of Arts in that year and intended to be A 444 and C 528 have proved by our present reliable methods to be A 445·7 and C 534·5. Hullah's fork, issued as C 512, was measured in the same way as 521·6, but is really 524·8.

The Siren, invented by Cagniard de la Tour (1819) was a more dependable pitch measurer, the sounds being produced from a rotating disk pierced with holes and blown by a pipe. The air passing through the hole when the disk is turned rapidly causes a sound, but is silent when the substance of the plate intervenes. A series of puffs is thus generated, the number of which is equal to the number of holes multiplied by the number of times the disk is turned—say in a second of time. But the least inexactitude in the turning causes error. To obviate this imperfection M. Cavaillé-Col, the well-known organ builder, contrived an improvement on the counting apparatus which he called "Soufflerie de Précision," and with this improved siren M. Lissajous determined the official standard fork for the "Diapason Normal" (1859), now preserved in the Museum of the Paris Conservatoire. It was intended to be A 435 double vibrations in a second (French counting by single vibrations, 870), at a temperature of 15° Cent.: (59° Fahrenheit). But it has been proved by M. Rudolf Koenig, of Paris, to be 435·45. He arrived at this result by a most ingenious apparatus in which a large tuning-fork acts upon a clock, instead of a pendulum. A chronometer is attached to check the accuracy of the clock, and after comparison and rectifying any error, it is easy to calculate the number of vibrations with an accuracy beyond question.

Other methods of recent invention are the Optical (McLeod and Clarke) and the Electrographic (Mayer), but they are

Digitized by Google

difficult of manipulation and are of use rather in corroborating the results of simpler pitch measurers such as the Tuning-fork Tonometer of Scheibler (1834). This tonometer is composed of a number of forks, beating within a countable number, say four in a second, through the extent of an octave. Each beat is equivalent to a double or complete vibration in a second, and the sum of the tonometer beats within the octave is identical with the vibration number of the lower note of the octave—the ratio of that interval being 1—2. Scheibler's original tonometer of 52 forks seems to be hopelessly lost, but the writer had the good fortune to meet with another of 56 forks, belonging to the Scheibler family, but lent to Herr Amels, a musical amateur of Crefeld, where Scheibler lived, and he very kindly allowed the late Dr. A. J. Ellis to use it for the basis of his History of Musical Pitch ("Journal of the Society of Arts," March 5 and April 2, 1880, and January 7, 1881). In this exhaustive work, p. 300, the exact vibrations of this tonometer are given. A. J. H.

(3) The method, now generally adopted, of showing the particular *octave* in which a sound occurs, is as follows :—

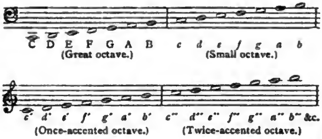

Sounds in the octave next below the *great octave* C, D, and so on. Sounds above the *twice-accented octave c''' d''* and so on.

The pitch of organ pipes is, however, still named after the old system, as follows :—

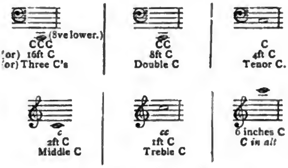

The octave below CCC being CCCC, and that above *c in alt*, *c in altissimo*.

Pitch-pipe. A wood or metal pipe having a moveable stopper or *tampion*, with marks, showing where it should be placed for the production of different notes of the scale. This form of pipe has been almost entirely superseded by an ingenious invention known as "Eardsley's Patent." It consists of a free reed capable of being lengthened or shortened by a graduated curve, on which are registered the notes of the scale.

Più (*It.*) More, as *più allegro*, faster; *più forte*, louder; *più lento*, slower; *più piano*, softer; *più presto*, more rapid; *più stretto*, more urged or closer; *più tosto allegro*, rather quicker.

Piva (*It.*) A bagpipe.

Pizz. Abbreviation of *pizzicato*.

Pizzicato (*It.*), **Pincé** (*Fr.*), **Gekneipt** (*Ger.*) *Lit.* pinched. A direction to violinists to produce the tone by plucking the string with the finger instead of using the bow. The place at which the use of the bow is to be resumed, is pointed out by the words *coll' arco*, or simply *arco*.

Placidamente, Placido (*It.*) Quietly, peacefully, placidly.

Plagal. The name given to those Church modes which were formed from the four older or authentic modes by taking the fourth below as the new key-note, and proceeding thence to the fifth above, *e.g.*:

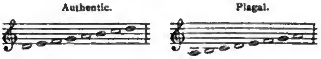

The plagal modes were distinguished by the addition of ὑπὸ, *e.g.*: Doric, an authentic mode; Hypodoric, a plagal mode formed from the Doric. [Plain Song.]

Plagal cadence. The cadence formed when a subdominant chord immediately precedes the final tonic chord, *e.g.*:

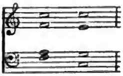

The plagal cadence of the old church writers corresponded more to the modern half-close, which is formed when the final chord is dominant, and is preceded by a tonic chord, *e.g.*:

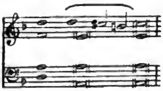

Plagiaulos (*Gk.*) πλαγίαυλος. (*Lat.*) *Tibia* *obliqua*. The cross-flute, or *flauto traverso*. [Flute.]

Plain-Chant. [Plain Song.]

Plain Song. The kind of music used from time immemorial in the worship of the ancient Jewish or Israelitish Church, as well as among all the churches of Christendom throughout the world.

The first principles of all music being found in nature, and its various forms having been gradually evolved by scientific and artistic industry and research from those principles, there must necessarily be found much of coincidence, and similarity, if not an actual identity, in the practical results of such inquiries and experiments. Hence it is not strange to find some Greek Music traced to the Hebrew, nor that of the early Christian Church to both. In fact, so far as in any country the natural development of the musical art has grown up healthily, according to its inner principle of life, there can be little doubt that, in all essential particulars (allowing, of course, a certain amount of characteristic individuality) there will be so much of resemblance to that of other countries as to lead historians (rather too fondly) to trace a parentage and derivation, where, be the fact as it may, neither are necessary to account for the likeness. The Plain Song of the Christian Church is a well known, and universally practised form of musical accent, recitation, and simple melody, written records of which are extant dating from the eighth century; and these are themselves records of Christian Plain Song of still earlier times, of which some, it is not unreasonable to believe, are the very melodies sung by Our Lord Jesus Christ and His Apostles, and of Templar origin.

The music of the Christian Church, like the gospel itself, was doubtless propagated orally, and by tradition, in the first years of its history; and as in the New Testament traces are discoverable of a primitive liturgy and hymnody, so we also find in early writers indications of certain efforts to fix and rule the vocal performance of these, by an appointed choral arrangement, with somewhat of a scientific connection.

One point of primary importance in this respect would naturally be the pitch most convenient for the singing of the various parts of Divine Psalms, Hymns, and Spiritual Songs; the relation in which the several parts of the service, musically considered, were to stand to each other; as also the interchanges to be made as to singing, either by course in alternate choirs, or by the clergy and people, in solo or in chorus.

The writings of the New Testament and of the early Fathers are not without various interesting indications, both of the important functions discharged by vocal music in Christian worship, and also of various ritual *uses;* in such wise, moreover, as excited *then*, as they do *now*, the approval and admiration of some, the indignation and blame of others.

Thus, "Sabellius and Marcellus took occasion to incense the Church of Neocesarea (A.D. 363), against their Bishop S. Basil, as being an author of new devices in the service of God," when he introduced the antiphonal chanting after the custom of Antioch, where it had been taught by S. Ignatius, a contemporary and disciple of the Apostles, specially of S. John. S. Ambrose, Bishop of Milan, was a great admirer of S. Basil, and introduced the same kind of chanting into his church as had originally prevailed in the East, as we learn from S. Basil's refutation of the heretics' charge of novelty and singularity, alleging the example, not only of the Church at Antioch, but also "of the Churches of Egypt, Lybia, Thebes, Palestine, the Arabians, Phœnicians, Syrians, Mesopotamians, and of *all that reverenced the custom of singing Psalms together.*"

Here, it would seem, we have ample grounds for the very general belief that in setting the music at first traditionally sung to various Psalms and Hymns, and the portions specially vocal in their Liturgies, the characters then used in noting music would be taken advantage of, to fix and perpetuate the arrangements of the early Christian music to the words. But the notation of the most artistic music then known was the Greek, which, like other refinements, was also at first used by the Romans, and subsequently improved by the use of the Roman letters, instead of the more complicated Greek characters. But with the Greek notation would, to a great extent, be adopted Greek modes of arrangement, and Greek notions as to the construction of the scales in which any chant, antiphon, or hymn tune might most fitly be said to stand. If its general character, *e.g.*, were major or minor, a similar mode or scale of the Greek music would be selected, and to this it would be assigned, just as in the present day a modern musician's arrangement of any ancient melody would, in the absence of any other system to which he might be able to assign it, necessarily be in the modern major or minor scale. But in our inquiries as to the ancient Plain Song, we must always remember that both it and the Greek modes were founded simply on *melodic*, and not on *harmonic* considerations, in our present meaning of those terms. With their usual acumen, the Greeks, in their *diatonic*, *chromatic*, and *enharmonic* genera, analyzed all the tone modulations of the human voice, with a discrimination most admirable, and effects most wonderful, as all are ready to admit who have studied their theoretic treatises, and the accounts recorded of the practical results of their performances; results, it may be suggested, such as poetic, oratorical, and musical recitation can even now occasionally produce in the case of singers of rare eloquence and sympathetic voices, on an appreciating and impressible audience. Sentiment and feeling, the soul and spirit of all true music, are not confined to any set of outward appliances, however vast, as in the present orchestra; but may be felt in the impassioned delivery of a Simple Plain Song, or the thrilling notes of an unaccompanied solo instrument. It is probable that the first kind of musical notes used by the Christian Church, called neumes (from the Greek *πνεῦμα*, *breath*) were derived from the Greek characters, after the fashion of the Hebrew accents for reading. The diatonic genus of the Greek music most nearly resembled (if it was not, in some modes, identical with it) the tonality of the Christian Plain Song and the ancient Hebrew music, of which it was, in part undoubtedly, the traditional record and sacred depository after the dispersion of the Jews and the consequent disuse by them of their Temple music, which, in their state of humiliation, they deemed it unlawful any more to sing. What music they now practice in their synagogues is still, however, eminently of the same character as the ancient Plain Song of the Christians. The fact above stated forestalls any objection that might be made from the modern Jews not now using them, to the belief that the well-known Ecclesiastical Psalm Tones have descended, through the "ages all along," the thirty centuries from the days of King David, the sweet singer of Israel, in an unbroken tradition, to our own day.

The sources of information as to the history of Plain Song, in addition to the various references scattered throughout the early writers and Fathers of the Church, may be found in the voluminous pages of *Burney* and *Hawkins* in English, and of *Gerbertus* in Latin; as also in the original authors quoted by them, specially the *Septem Scriptores Antiqui* of Meibomius; in the French *Histoire de L'Harmonie au Moyen Age*, by E. De Coussemaker; in the XXIXth Vol. of the *Nouvelle Encyclopédie Théologique*, edited by M. L'Abbé Migne, being a *Dictionnaire de Plain Chant, et de Musique Religieuse;* in the pages of Alfieri, La Feillé, Clement, Janssen, Lambillotte, Alix, and in English in the various editions of Latin service and instruction books by Lambert, Benz, Chas. Child Spencer, and the prefaces to the Rev. T. Helmore's Accompanying Harmonies to the Psalter Noted and Hymnal Noted; more satisfactorily still in the MS. and printed books of Plain Song in various libraries at home and abroad. The subject is very wide, and cannot easily be studied without a large acquaintance with the numerous books of Ritual music, both of the present time, and of all the centuries backwards, till the very first in which any written remains of Plain Song can be discovered. A sketch of what may be gathered as the result of such an investigation is all that can be here attempted.

PLAIN SONG, then, it may be stated, is a definite, well known system of Church Song, which all Catholic musicians are obliged by their duties to study, and from which all the great composers, down to our own times, have more or less tinctured their highest and best compositions.

It has already been stated that the ancient church modes of the Plain Song or Gregorian music had their prototypes, if not their very birth, from the Diatonic genus of the Greeks. Perhaps it would be more true, and in accordance with the notions before expressed, to say that the Greek Diatonic genus was identical with the ancient Hebrew music, and with that of the Christian Church which perpetuated it. Here, then, is a digest of what can be gathered from ancient writers as to the Greek Diatonic scale.

It embraces fifteen sounds, which practically might be produced by the strings of a harp, fifteen in number; or by the voice of any man of ordinary medium vocal power.

It is divided into tetrachords, or divisions of four sounds each,—beginning, not on the first of the system, which (as its name implies) was originally an extra note added below, but on the second, or B. In the lower nine sounds the second and third tetrachords are reckoned to begin, each upon the fourth sound of the foregoing; and are therefore styled *conjoined tetrachords.* The first pair of tetrachords is also succeeded by a tetrachord, an octave above the first, beginning on the sound next to the fourth of the second tetrachord, and consequently denominated the *disjoined* tetrachord. Then a fifth tetrachord occupied the highest part of the system: all which will be better seen and understood by Diagram I., in which the Greek names for the various sounds (written in English letters) are given, with a translation of their characteristic nomenclature.

Digitized by Google

DIAGRAM I

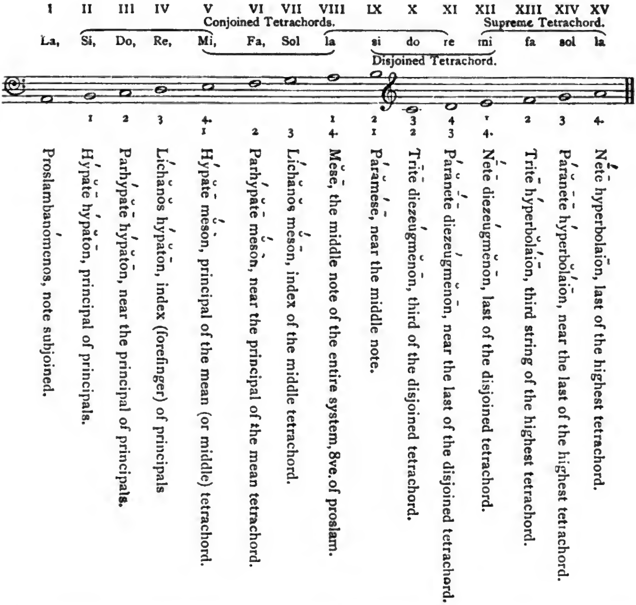
I II III IV V VI VII VIII IX X XI XII XIII XIV XV
Conjoined Tetrachords.
Supreme Tetrachord.
La, Si, Do, Re, Mi, Fa, Sol la si do re mi fa sol la
Disjoined Tetrachord.
1 2 3 4. 1 2 3 4. 2 3 4 1 2 3 4.
1 2 3 4 1 2 3 4.
Proslambanomenos, note subjoined.
Hypate hypaton, principal of principals.
Parhypate hypaton, near the principal of principals.
Lichanos hypaton, index (forefinger) of principals
Hypate meson, principal of the mean (or middle) tetrachord.
Parhypate meson, near the principal of the mean tetrachord.
Lichanos meson, index of the middle tetrachord.
Mese, the middle note of the entire system, 8ve. of proslam.
Paramese, near the middle note.
Trite diezeugmenon, third of the disjoined tetrachord.
Paranete diezeugmenon, near the last of the disjoined tetrachord.
Nete diezeugmenon, last of the disjoined tetrachord.
Trite hyperbolaion, third string of the highest tetrachord.
Paranete hyperbolaion, near the last of the highest tetrachord.
Nete hyperbolaion, last of the highest tetrachord.

N.B.—The accents and prosodaic marks are here used simply for the English reader. They are not Greek.

It will soon be seen that these same fifteen natural notes are those from which the Church modes are all constructed. The musical theorists who, about the eighth century, disputed as to the number of modes or scales to be recognized in Church music, seem to have laid down the principle that a scale or mode must consist of the combination of a perfect fifth (diapente) with a perfect fourth (diatessaron). The division of a scale into these two component parts might be made so as to have either a diapente or a diatessaron in the lower position. The division which had the fifth in the lower and the fourth in the higher place was called *Harmonic*, as being that which music naturally suggests; and the other division of the eight notes of a scale where the fourth was in the lower, and the fifth in the upper portion was called *arithmetical*, because, in the science of numbers, the greater numbers take precedence of the less, and are counted their superiors. This fanciful nomenclature is mild in comparison with many other conceits and far-fetched analogies of ancient theorists.

The Church modes, then, are formed of the seven natural notes of the scale and their octaves, beginning upon any one of the seven, and extending to its octave, subject to the rule just stated that the fifths and fourths of which any scale was composed must be perfect.

Now the Church melodies were generally contained within the compass of an octave or a little more. Some had a much less range from the highest to the lowest notes. When the melody ascended to the fifth above the *final* and the fourth above the fifth, and did not descend below the final, it was said to be *authentic;* when it extended to a fifth above the final and a fourth *below it*, it was called plagal. The sounds of which these two sorts of melodies consisted being set down in orderly succession from the lowest to the highest, naturally arranged themselves into two corresponding classes; and the ecclesiastical modes are therefore matched, or (as mediæval writers have termed it) wedded to one another in pairs, authentic and plagal, in which male and female characteristics are respectively developed.

Now, taking the notes before given as the Greek Diapason, and supposing any Antiphon or other piece of Plain Song to descend to the lowest note A, and, after ascending by various intervals to the A an octave above it, to settle down and close on D, we should, accordingly, by arranging these sounds in an orderly sequence from the lowest to the highest, have a scale whose final or closing note is D, and its range of notes from A to A, and this, as before explained, would be the lower, collateral, and plagal scale or mode of another beginning on D, the final, and rising to its octave, which would be its own proper authentic, or ruling and authoritative mode. Such, in fact, are the first two modes of the Gregorian Plain Song, and the pair, as husband and wife, have reciprocal functions, and though two, are yet one by participation of one common bond of union. viz., the diapente D, E, F, G, A.

In Diagram II. we see, then, the 11 notes of the Dorian Modes, Authentic and Plagal, called the 1st and 2nd.

The 1st is authentic, and ranges from the 4th to the 11th note, having a perfect fifth from D to A, which same fifth is the upper part of the plagal mode. From A the 8th to D the 11th is a perfect fourth, and thus fulfils the requirements before mentioned as essentials of a Church mode; while the same law is obeyed in the case of the 2nd or plagal mode, as from A the 1st to D the 4th is the required perfect diatessaron.

The terms *hyper* (*above*), and *hypo* (*below*), are used to distinguish the two, though the *hyper* is often omitted as unnecessary; and besides, it seems to have been used in another way by some of the Greek writers.

We are told that in Charlemagne's reign (about A.D. 768) a great dispute arose as to the allowable number of the modes, and that the Emperor at first decided in favour of the first eight alone; but that on the remonstrance of the malcontents, the 9th, 10th, 11th, and 12th, were also allowed. The ritual books of the church have, however, been generally arranged for the eight tones only, and those portions of the Plain Song which formerly were assigned to the 9th, 10th, 11th, and 12th modes were so re-arranged by a transposition, as to *appear* to have been composed in some one or other of the first eight; though a careful inspection of these transpositions will serve to show that the means at command were not (while no other change of sound was allowable than the B flat) sufficient to effect a *perfect* transposition. This will appear in the subsequent examples.

DIAGRAM II.

1st AND 2nd MODES.

Dorian or Hyper-Dorian, and Hypo-Dorian.

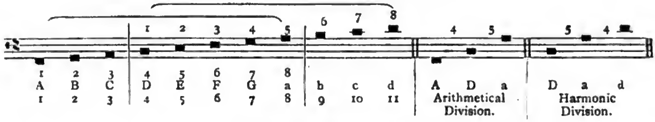

3rd AND 4th MODES.

The next pair is the Phrygian, thus:—Phrygian or Hyper-Phrygian, and Hypo-Phrygian. Final E.

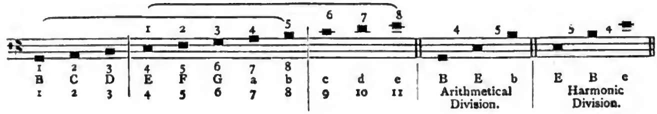

5th AND 6th MODES.

These are the Lydian or Hyper-Lydian, and Hypo-Lydian. Final F.

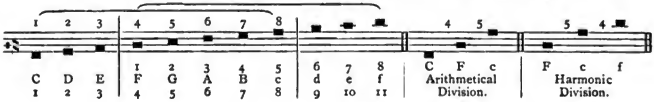
1 2 3 4 5 6 7 8
4 5
5 4
1 2 3 4 5 6 7 8
C D E F G A B c d e f
1 2 3 4 5 6 7 8 9 10 11
C F c
Arithmetical Division.
F c f
Harmonic Division.

7th AND 8th MODES.

The Mixo-Lydian or Hyper-Mixo-Lydian, and the Hypo-Mixo-Lydian. Final G.

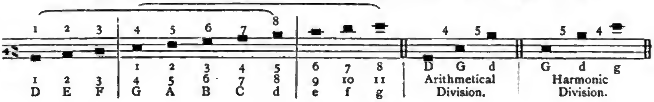
1 2 3 4 5 6 7 8
4 5
5 4
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8 9 10 11
D E F G A B C d e f g
D G d
Arithmetical Division.
G d g
Harmonic Division.

9th AND 10th MODES.

The Æolian or Hyper-Æolian, and Hypo-Æolian. Final A.

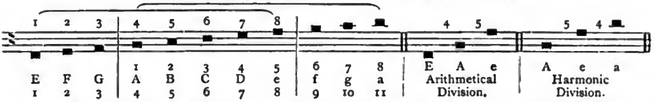
1 2 3 4 5 6 7 8
4 5
5 4
1 2 3 4 5 6 7 8
E F G A B C D e f g a
1 2 3 4 5 6 7 8 9 10 11
E A e
Arithmetical Division.
A e a
Harmonic Division.

B cannot become the final of a mode as the 4th below it is F, and the interval a pluperfect 4th, and the 5th above it is also F, and the interval an imperfect 5th.

11th AND 12th MODES.

The Ionian (or Iastian) or Hyper-Ionian (or Hyper-Iastian), and Hypo-Ionian (or Hypo-Iastian). Final C.

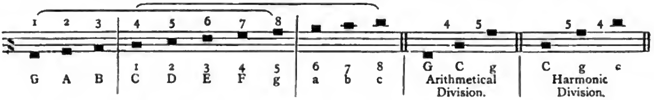
1 2 3 4 5 6 7 8
4 5
5 4
1 2 3 4 5 6 7 8
G A B C D E F g a b c
G C g
Arithmetical Division.
C g c
Harmonic Division.

Thus it will be seen that the ancient system of scales took every one of the six natural notes on which the union of the perfect diatessaron and perfect diapente can be effected as the final of a pair of modes, thus placing the diatonic intervals of tone and semitone in every possible position with respect to the two tetrachords of each, and of the final of each pair.

There was yet another method of securing to each mode its own characteristic features, by which it is clearly to be discerned from all the rest. In each there is a note higher than the final which assumes the next importance to it, from being another point of sound about which the melodies principally, as it were, revolved, and on which the voice more frequently and (in the case of the recitation of any psalm, antiphon, or canticle) most continuously dwelt. From thus predominating over the other sounds of each scale it is called the DOMINANT. But it must not be confounded with what in general is now understood by that term, nor with the harmonic associations which it suggests to the mind of a modern musician.

In every authentic mode the dominant is the 5th above the final, except in the third mode, where B, the fifth, being a changeable note, *i.e.*, sometimes natural and sometimes flat, and also standing, unlike any other note in the scales, at the forbidden interval of a tritone, or pluperfect fourth from the F below it, and of an imperfect fifth from the F above it, is rejected, and C, the nearest sound to it, is the dominant, being a minor 6th above the final E

In the six plagal modes the dominant is the third note in the scale below the dominant of each relative authentic mode, except in the 8th, where B, the 3rd below D, is also on the same grounds discarded for C.

Digitized by Google

DIAGRAM III.

Finals.

1. Authentic, Hyper-Dorian .... $d\ e\frown f\ g\ \overset{\text{Dom.}}{a}\ b\frown\bar{c}\ \bar{d}$
2. Plagal, Hypo-Dorian .... $A\ B\frown c\ d\ e\frown\overset{\text{Dom.}}{f}\ g\ a$ — — —
3. Authentic, Hyper-Phrygian .... $e\frown f\ g\ a\ b\frown\overset{\text{Dom.}}{\bar{c}}\ \bar{d}\ \bar{e}$
4. Plagal, Hypo-Phrygian .... $B\ c\ d\ e\frown f\ g\ \overset{\text{Dom.}}{a}\ b$ — — —
5. Authentic, Hyper-Lydian .... $f\ g\ a\ b\frown\overset{\text{Dom.}}{\bar{c}}\ \bar{d}\ \bar{e}\frown\bar{f}$
6. Plagal, Hypo-Lydian .... $c\ d\ e\frown f\ g\ \overset{\text{Dom.}}{a}\ b\frown\bar{c}$ — — —
7. Authentic. Hyper-Mixo-Lydian .... $g\ a\ b\frown\bar{c}\ \overset{\text{Dom.}}{\bar{d}}\ \bar{e}\frown\bar{f}\ \bar{g}$
8. Plagal, Hypo-Mixo-Lydian .... $d\ e\ f\ g\ a\ b\frown\overset{\text{Dom.}}{\bar{c}}\ \bar{d}$ — — —
9. Authentic, Hyper-Æolian .... $a\ b\frown\bar{c}\ \bar{d}\ \overset{\text{Dom.}}{\bar{e}}\ \bar{f}\ \bar{g}\ \bar{a}$
10. Plagal, Hypo-Æolian .... $e\ f\ g\ a\ b\frown\overset{\text{Dom.}}{\bar{c}}\ \bar{d}\ \bar{e}$ — — —
11. Authentic, Hyper-Ionian or Iastian .... $\bar{c}\ \bar{d}\ \bar{e}\frown\bar{f}\ \overset{\text{Dom.}}{\bar{g}}\ \bar{a}\ \bar{b}\ \bar{\bar{c}}$
12. Plagal, Hypo-Ionian or Iastian .... $g\ a\ b\ \bar{c}\ \bar{d}\ \overset{\text{Dom.}}{\bar{e}}\frown\bar{f}\ \bar{g}$ — — —

Thus A is the Final of the 9th and 10th, and Dominant of the 1st, 4th, and 6th.
„ C „ 11th „ 12th „ „ 3rd, 5th, 8th, and 10th.
„ D „ 1st „ 2nd „ „ 7th.
„ E „ 3rd „ 4th „ „ 9th and 12th.
„ F „ 5th „ 6th „ „ 2nd.
„ G „ 7th „ 8th „ „ 11th.

The following formula committed to memory will be of immense use in grasping the otherwise troublesome varieties of the Church modes.

Finals ...... D	D,	E	E,	F	F,	G	G,	A	A,	C,	C.
Dominants *A*	*f*	*c*	*a*	*c*	*a*	*d*	*c*	*e*	*c*	*g*	*e*
Mode No. 1	2	3	4	5	6	7	8	9	10	11	12

[N.B.—The dominants spell the mnemonic word Afcăcăd[é]cĕcgĕ.]

Of this ancient system the German Professor of Music at the University of Berlin, Dr. Adolph Bernhard Marx, justly remarks that its "profound ideality is undeniable," and while thoroughly imbued with the notions and feelings of a modern, confesses that in many points it displays nicer distinctions and more striking characteristics than our own. Whatever view, however, may be true with regard to the discoveries and improvements of the musical art in our day, it is certain that for sacred music generally, and for Church music in particular, the study of Plain Song as it has come down to us, with all the hallowed associations of a definite history of twelve centuries and with an indefinite period of great extent before them, together constituting a sequence of near upon three thousand years' use in the highest acts of human concern, cannot be neglected without much injury to the sentiment and feeling of composers, and a secularising of Church music, both in its composition and in the style of its execution, highly to be deprecated by every true lover of Sacred Art, and of the Christian Church throughout the world.

It remains to say a few words on the time, accent, rhythm, and metre of Plain Song, all of which have unfortunately suffered much in the opinion of mankind by being misunderstood in theory, and dreadfully distorted in practice.

It is the opinion of those who, like Coussemaker, have most deeply studied the subject, that there was by no means that monotonous equality in the length of the notes, which we have found practised and thought right since the Plain Song of the Church has been so generally neglected for more strictly mensurable modern compositions.

A manifest tendency has long existed to lengthen the short notes of ancient MSS., in order probably that the accompanying organists may give (as they intend) a greater dignity to the solemn march of their harmonies, and get rid of the charming varieties of what may be styled the *bars*, speaking after the fashion of modern music. Two striking instances of this will explain what is meant better than further description.

There is in the British Museum, among the Lansdown MSS. (No. 463), a fine well-written copy of the ancient *Sarum Antiphonary;* and in the Bodleian Library, Oxford, there is a corresponding MS. most splendidly written, which formerly belonged to Archbishop Laud; in both of these the writer of this article has found very many notes short (*i.e.*, the ◆ ◆ diamond shaped semibreve) where in the printed folio (c 35 l., Paris, 1519. 20 B. M.), they are changed into breves thus ■ ■. The difference thus made in the music may be compared to an alteration of all the quavers in "God save the Queen" into crotchets!

Another instance of the same perverse treatment may be seen in the version of the melody "Veni, veni, Emmanuel," set to the words "O come, O come, Emmanuel," in *Hymns Ancient and Modern* (No. 36, first tune), where all the short notes, originally printed (in the *Hymnal Noted*) correctly from the MS. copied by the late J. M. Neale from a French Missal in the National Library, Lisbon, are deliberately, and of set purpose, turned into long ones, seemingly to avoid the interspersion of a triple bar, in several places, among the duple or quadruple measures. See *Hymnal Noted*, "Veni, veni, Emmanuel," 65 Ac. Har., p. 213, "Draw nigh, draw nigh, Emmanuel."

The time or rhythm of the Plain Song, where best understood and properly rendered, is a most entire and definite flow of melody in accordance with the accents of the words and the construction of the musical phrase combined. The fullest development of this is perhaps in the Ancient Metrical Hymns, in which the varied prosodaic metres are always complied with by the Plain Song, the accents of which correspond to those of the words; and the notes fall into what we should now call *bars*, sometimes regular in their formation, and of one time, but often in duple and triple forms mixed.

There is, to the lovers of Plain Song who have caught the spirit and intention of its various phrases, a charming variety and freedom from stiffness in these changes of measure, somewhat akin to the pleasure we derive in verse from similar departures from one stereotyped succession of accents or quantities. For example, after the regular iambics of this couplet—

"Thro' life's long day and death's dark night,
O gentle Jesus, be our light—"

how pleasing is the introduction of the *trochee* in the beginning of the next verse:

"Labour is sweet for thou hast toiled."

Something of the same kind seems to occur, in such a phrase as this, put into measured time, as it should be sung from the usual notes of the Gregorian Plain Song.

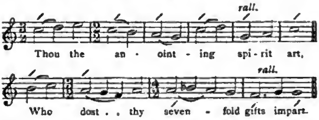
rall.
Thou the an - oint - ing spi - rit art,
rall.
Who dost . . thy seven - fold gifts impart.

In conclusion, it may be remarked that the more highly people are educated in all that concerns the employment of art in the Service of the Church, and the more deeply they enter into the *rationale* of the use of music as the handmaid of religion, the better will they appreciate the value of Plain Song, the better will they execute it, and the better too will they know and feel what other music is fittest, and really the best, to stand side by side with it in the Lord's House, as the offering of the first-fruits of man's highest art and most exquisite skill, to Him from whom all skill and all talent are derived, and who will be worshipped by each according to that he hath, be it little or much. The Plain Song of the Church gives all who can sing the best means of discharging this sacred duty.

Plainte (*Fr.*) An elegy, or lament.

Plaisanterie, or **Pleasantrie.** A kind of concerto for a solo instrument, in which various cheerful tunes or dance melodies were introduced.

Planxty. A harp tune of a sportive and animated character, moving in triplets. It is not intended for or often adaptable to words, and is slower in pace than the jig.

Plaqué (*Fr.*) Struck at once. A direction that a chord is not to be played *arpeggio*.

Play-house Tunes. Musical compositions performed between the acts of the tragedies or comedies in the old theatres of London.

After the Restoration, music became more general as a relief to the performers and the performances in theatres; compositions, called Playhouse or Act Tunes, were written and played in concert, and not in unison as formerly, and the theatre music was of a superior character. "Insomuch," says Sir John Hawkins, "that to say of a performer on any instrument that he was a playhouse musician, or of a song that it was a playhouse song, or a playhouse tune, was to speak of each respectively in terms of the highest commendation." The most popular were the "Genius of England," Madame Subligny's "Minuet," the "Louvre" (or "Loure,") and many others. The principal composers of this kind of music were Mr. John Reading, John Bannister, Godfrey Finger, Matthew Lock, Henry and John Eccles, Raphael Courteville, &c. A coranto written by Bulstrode Whitelock, one of the Lords Commissioners of Charles II., in conjunction with Henry Lawes, was played at the theatres for more than thirty years regularly "by the Blackfriars musicians, esteemed the best in London."

Plectrum (*Lat.*) πλῆκτρον. A little staff made of ivory, horn, quill, or metal, with which (having it in his right hand) the player on a lyra or cithara, set the strings in vibration. It cannot yet be said to be obsolete, inasmuch as one is used by the performer on a modern mandoline and zither. From the method of using a plectrum to the cithara, the more perfect instrument, the dulcimer (Hackebrett), probably took its origin; and from the dulcimer came the spinet, harpsichord, and eventually the pianoforte. But it should be remembered that while the quill of a harpsichord plucks, as it were, a string, just as a plectrum would, the hammer of a pianoforte *strikes* the string. The plectrum was not invariably used by the ancients, for some of the paintings discovered at Herculaneum depict citharists using the ends of their fingers.

Plein jeu (*Fr.*) With the full power.

Plica (*Lat.*) A term used by the old writers on mensurable music the exact meaning of which it is difficult to discover. Franco of Cologne (Gerbertus, Script. iii. 6) says, "Plica est nota divisionis ejusdem toni in gravem et acutam" (Plica is a sign of the division of the same note into a higher and lower sound). This would seem to describe it as an ornament or grace. Joannes de Muris says, "Plica dicitur a plicando, et continet notas duas, unam superiorem et aliam inferiorem." This would make *plica* a sort of appoggiatura. The following were the only *plicae* used: Longa plica descendens Longa plica ascendens Brevis plica descendens Brevis plica ascendens They were disused early in the fifteenth century.

Ploke (*Gk.*), πλοκή. [Melopœïa.]

Plures ex una (*Lat.*) Many parts from one. The old name of a canon because originally written only on one line.

Plus (*Fr.*) More, as, *plus animé*, more animated, with greater spirit.

Pneuma. [Neuma.]

Pneumatic bellows. [Organ, § 12.]

Pneumatic Organ. Organum pneumaticum. The ordinary organ as opposed to the ancient hydraulic organ, or organum hydraulicum, in which water was used, not as was popularly supposed at the time, for the purpose of entering the sound-boards or pipes, but to act as a regulator of the pressure of air. [Organ.]

Pocetta (*It.*) [Poche.]

Poche, Pochette (*Fr.*) A little fiddle used by dancing-masters. [Kit.]

Pochettino, Pochetto (*It.*) A little, as, *ritard un pochettino*, making a slight *rallentando*.

Poco (*It.*) A little, as *poco a poco*, little by little; *poco animato*, rather animated; *poco lento*, rather slow; *mosso poco meno*, rather less quick; *poco piano*, rather soft; *poco più allegro*, rather faster; *poco presto*, somewhat rapid.

Poggiato (*It.*) Leant upon, dwelt upon.

Poi (*It.*) Then, as *piano poi forte*, soft then loud. *Poi segue*, then follows.

Point (*Fr.*) A dot.

Point d'orgue (*Fr.*) A pedal-point [Fugue.]

Point de repos (*Fr.*) A pause.

Pointé (*Fr.*) Dotted.

Polacca (*It.*) Polish. A title applied to melodies written in imitation of Polish dance tunes. It is synonymous with the French word polonoise, or polonaise. It is a term of modern introduction, as neither Brossard, Grassineau, nor Rousseau, have the word in their dictionaries. Kollmann on "Style and National Music," describes it as "a particular characteristic piece in $\frac{3}{4}$ time, and its movement like a majestic but fluent andante or andantino. It deviates from the general rule respecting simple measure in making every rhythmical cæsure, not on the first, but on the last time or crotchet in the bar."

In No. 3 of Handel's twelve grand concertos is a polonaise or polacca, of a slightly different rhythm, and there are other polaccas which differ still more from the model described above.

Polka. A dance of world-wide popularity, the music to which is in 2/4 time, with the third quaver accented. There are three steps in each bar, the fourth beat is always a rest, the three steps are performed on the three first beats of every bar.

It is stated that Anna Slezak, a farm servant at Elbeteinitz, near Prague, invented the polka about 1830. The room in which she was accustomed to dance being of small dimensions, the movements of her feet were short, and so the dance was called the "Pulka" dance, that is the "half" dance. The title was changed when it was introduced into France, as were also the steps and the character of the dance. In Bohemia, where it is said to have originated, it is danced with a peculiar alternative of hurrying and halting steps to which the music has to be accommodated.

It was brought into England about the year 1843, and was received with an extraordinary enthusiasm by all classes. It effected a revolution in the style of dancing which had prevailed up to that period. Country-dances and quadrilles were even performed with livelier steps than hitherto, for the polka was so popular that it absorbed every other dance for a time. Articles of food, of clothing, and of ornament, were named after it. Comic songs and farces were written on the *polka-mania*, as it was called. To such a height did this absurd fancy reach, that public-houses were called the "Polka Arms," and newly-built streets "Polka" terraces or crescents. When the mania subsided these names were altered to titles less suggestive of a passing folly.

Polonaise *or* **Polonoise** (*Fr.*) [Polacca.]

Polychord. An instrument invented by Fried. Hillmer of Leipzig, in 1799. It was strung with ten strings. In shape it was not unlike a double bass without a neck. It was never generally used.

Polycephalus. One of the neumes. [Neumes.]

Polymorphous. Of many shapes. A term applied to compositions the parts of which are capable of inversion, as in double counterpoint; or of augmentation, diminution, *per thesin et arsin*, and other devices, as in Canon.

Pommer. [Bombardo.]

Pomposamente, Pomposo (*It.*) Pompously.

Ponderoso (*It.*) With weight, impressively.

Ponticello (*It.*) *Lit.*, a little bridge. (1) The bridge of a violin or other instrument. (2) The break in the voice. [Break.]

Pont-neuf (*Fr.*) A common ballad, such as was formerly sold on the Pont-neuf, Paris.

Poogye. The nose-flute of the Hindoos. Probably blown by the nose instead of the mouth, in order to avoid possible defilement of caste.

Portamento (*It.*) A lifting of the voice, or gliding from one note to another.

Portando la voce (*It.*) Sustaining the voice, or gliding from one note to another.

Portative organ. A little organ which could be carried about, as opposed to a *positive* organ, which was fixed. Many consider the portative organ to have been identical with regals; they were certainly of the same class. A portative organ was often carried in processions, on a man's shoulder, who, when it was required, placed it on a stool and worked the bellows.

Portato (*It.*) Lengthened, sustained.

Port de voix (*Fr.*) A kind of appoggiatura combined with the Pincé, *q.v.*, written in old harpsichord music thus:

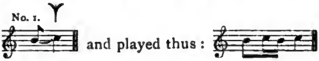

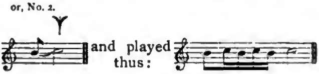

No. 1 is the *Port de voix simple*, and No. 2 the *Port de voix double.*

Portée (*Fr.*) Stave.

Porter la voix (*Fr.*) [Portando la voce.]

Portunal-flute. An organ stop, the pipes of which are of wood, and are open, and larger at the top than at the mouth.

Posato (*It.*) Quietly.

Posaune (*Ger.*) [Trombone.]

Posaune. A reed-stop on the organ, of a rich and powerful tone. Its pipes are of a very large scale. It is of 8 ft. on the manuals, and of 16 ft. or 32 ft. (*contra-posaune*) on the pedals. The tubes of the manual stop are generally of metal, sometimes of tin; those of the pedal stop, sometimes of metal, often of zinc or wood.

Posément (*Fr.*) Gravely, sedately.

Position. (1) A chord is said to be in its *original position* when the groundnote is in the bass, in other *positions* when the relative arrangement of the component notes is changed. (2) The position of a chord is the same as the *disposition* of its parts. A close position is *close harmony;* an open position *open harmony.* (3) A *position*, on a violin or other string instrument, is to use the fingers otherwise than in their normal place. [Shift.]

Position (*Fr.*) [Position 3.]

Positive Organ. An old name for the

choir organ. Originally a positive organ was a fixed organ. [Portative organ.]

Positif (*Fr.*), **Positiv** (*Ger.*) [Positive.]

Position. (1) The arrangement of the constituent notes of a chord. (2) The position of the hand on string instruments. [Shift.]

Possibile (*It.*) Possible, as, *il più forte possibile*, as loud as possible.

Post-Horn. (1) A wind instrument consisting of a simple metal tube, without valves or pistons, blown by postmen. It can hardly be termed a musical instrument. (2) A piece of music suitable to, or in imitation of the notes or passages of a post-horn. [Bugle.]

Postludium (*Lat.*), **Nachspiel** (*Ger.*) Postlude; a concluding voluntary; a piece played at the end of the service.

Potpourri (*Fr.*) A medley; a collection of various tunes linked together; a capriccio or fantasia on popular melodies.

Poule, la (*Fr.*) One of the movements of a quadrille.

Poussé (*Fr.*) The up stroke of a bow. [Bowing.]

PP. Abb. of *pianissimo*.

Pralltriller (*Ger.*) A transient shake, signified by *w*.

Preambulum, Preludium. *Vorspiel* (*Ger.*) An introductory voluntary.

Precentor, Præcentor, Cantor, Caput Scholæ (*Lat.*), **Capiscol, Cabischol, Armarius, Primicerio** (*Sp.*), **Préchantre** (*Fr.*), **Grand Chantre** (*Fr.*), **Vorsänger** (*Ger.*) An officer in a cathedral, who in dignity ranks next to the dean. His stall is on the opposite side of the choir, and that side is called cantoris side, the side of the *cantor*, as the other is called decani, the side of the *dean*. The office of the Precentor was an important one in olden times, and the duties are variously defined. The Rev. Mackenzie Walcott, Precentor of Chichester, in his valuable book "Cathedralia," thus speaks of the Precentor and his duties: "The primicerius, the first named on the waxen tablets or roll of singers, or precentor, was the bishop's vicar, who governed the inferior clergy, presided at the canonical hours, directed the lectures by the clerks, and controlled the order and mode of singing by the choir. From a mere office the precentorship in cathedrals became a dignity. He only sang on the greater feasts. In some places he had authority over all the city and diocesan schools, and was known as *caput scholæ*, *magister scholæ*, being also *scholasticus*. No one could open a song-school without his leave. He saw that all were attentive in the choir and observant of the statutes, he corrected faults and chastised offenders. In England the office of precentor was instituted at an early period—at Exeter, c. 1080; at Salisbury in 1091; at Chichester, 1115; at Wells, c. 1135; at St. David's, 1224; at Hereford, c. 1195; at Lichfield, 1130; Lincoln, 1097; York, 1090; and at St. Paul's in the reign of King John. The duties of the Precentor were generally alike, but differed only in a few local or diocesan peculiarities. A summary of his duties as laid down in the statutes of the several places may be briefly told. He had to direct the divine offices as regards the chant; to select candidates to supply vacancies; to provide and keep the choir-books in good condition; to table the weekly rota, or list of officers, with their duties; to choose and present a succentor when necessary; to instal the dean and canons, and to assign stalls to the prebends; to superintend the training of the choristers, not only as regards food and clothing, but also their morals and choral instruction. On great feasts he 'ruled the choir,' when two canons, robed in the red soutane (cassock) were the chanters, whom he followed up and down the choir, regulating the offices, and overseeing the service in rochet and cope, holding a staff in his hand, which was gloved and ornamented with a ring. He further observed that all the choir sang with proper modulation, and that various and proper chants, according to the day and festival, were used. He gave the note to the canon celebrant at the altar, distributed the copes in the choir, and having presented offenders to the dean and chapter, left correction to them."

In most cathedrals, the revenues of the office have been taken away and its dignity and status lowered, but nevertheless, during the last few years there has been a general attempt to revive the *musical* responsibilities of the holder.

Precipitatamente, Precipitato, Precipitazione, con; Precipitoso (*It.*) With precipitation, impetuously, hurriedly.

Precipité (*Fr.*) [Precipitatamente.]

Precisione, con (*It.*) With exactitude or precision.

Prefectus Chori. Director of the choir. Precentor.

Prelude. A movement played before, or an introduction to, a musical work or performance.

Preludio (*It.*) A prelude, introduction, or introductory voluntary.

Preludium (*Lat.*) [Preludio.]

Première (*Fr.*) First, as *première fois*, the first time, *prima volta* (*It.*); *première dessus*, first treble, *canto primo*.

Preparation. The causing a discord to be heard as a concord immediately before its percussion. It must take place in the same part as that which has the discord.

Preparazione (*It.*) [Preparation.]

Presa (*It.*) *Lit.* taken, caught. A cha

Digitized by Google

racter or mark used generally in continuous fugues or canons to mark the point of entry for the voices or instruments. A lead.

Pressante (*It.*) Pressing on, hurrying the time.

Pressiren (*Ger.*) To hurry the time.

Prestamente, Prestezza, con. (*It.*) Hurriedly, with rapidity.

Prestant. The open diapason of an organ, sometimes of 16 feet, sometimes of 8 feet in length.

Prestissimamente (*It.*) As rapidly as possible.

Prestissimo (*It.*) Very fast indeed.

Presto (*It.*) Fast. *Presto assai*, very fast.

Prick-Song. Written music, as opposed to extempore descant.

Prima, *fem.* (*It.*) First; as, *prima buffa*, chief comic actress or singer; *prima donna*, chief female singer in the opera; *prima viola*, first viola; *prima vista*, at first sight; *prima volta*, the first time, *i.e.*, before repeating.

Prime. (1) Tonic or generator. (2) The lowest note of any two notes forming an interval. (3) The first partial-tone.

Primicerius (*Lat.*) Prior scholæ cantorum. In cathedrals the precentor, succentor or cantor.

Primo, *masc.* (*It.*) First; as, *tempo primo*, at the original pace or time; *violino primo*, first fiddle; *primo buffo*, chief comic actor or singer; *primo musico* and *primo uomo*, principal male singer in the opera.

Principal (*Old Eng.*) The subject of a fugue, the answer being termed the Reply.

Principal (*Ger. and Fr.*) An organ stop. [Open diapason.]

Principal. On English organs the chief open metal stop, one octave higher in pitch than the open diapason. On the manual 4 feet, on the pedal 8 feet in length.

Principal. Chief, as *principal violin*, the leader; *principal vocalists*, those to whom the solos are allotted.

Principal. The name given by Handel to the third trumpet in the "Dettingen Te Deum."

Principale (*It.*) [Principal 4.]

Principalis mediarum (*Lat.*) Hypate meson. [Greek Music.]

Principalis principalium (*Lat.*) Hypate hypaton. [Greek Music.]

Principalis extenta (*Lat.*) Lichanos hypaton. [Greek Music.]

Principal Subject or Theme. One of the chief subjects of a movement in sonata form, as opposed to a subordinate theme.

Printing of Music. Music is printed in various ways: from plates, from lithographic stones, or from moveable type.

The plates for music are generally of pewter, and the notes and characters are stamped with punches of the shape and character required. Corrections are made by beating up the back of the plate, so that the surface may be as equal as possible for making the alterations. The music is either printed direct from the plate, or an impression is transferred to a lithographic stone, and so printed. This latter plan is the most usual when large impressions are needed, or when it is desirable to preserve the plates.

In the earliest engraved music copper-plates were used, and every note was made with the *graver*. Both words and notes were formed in imitation of the writing of the time. "Parthenia,"—"the first musicke that was ever printed for the Virginals,"—in which there are compositions by Byrd, Bull, and Gibbons, was "engraven on copper," in 1611, and the character employed is very like writing. This style prevailed in France and Italy, as well as in England, until the commencement of the eighteenth century, when stamping musical notes on plates became more general for this class of work. The process of stamping, which has been more or less improved according to the skill of the stamper, has always been called "engraving," though, strictly speaking, it ought to be called "punching."

The invention of moveable music type is claimed for Ottaviano Petrucci, who established a printing press at Venice, and was working as early as 1503, the year in which he published some Masses by De La Rue. Petrucci, removing to Fossombrone in 1513, obtained a patent from Leo X. for his invention of moveable types, for the sole printing of *Cantus Figuratus* and *Organorum Intabulaturæ*, for a term of twenty years. There is, therefore, good reason for believing his claim to be supported by trustworthy contemporary evidence. The very first known example of printed music is in the Milan edition of Franchinus. The notes there printed are not, however, music types, but engraved wood blocks, like the musical characters in Higden's "Polychronicon," translated by Trevisa, and printed by Wynkin de Worde in Westminster in 1495.

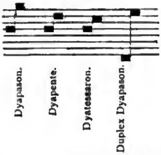

These characters were improved in a book published by Grafton in 1550, entitled "The

Digitized by Google

Booke of Common Praier Noted," by John Marbecke, Organist of Windsor.

The printer thought it necessary to prefix an explanation of the types he employed, which he does in the following words: "In this booke is conteyned so much of the order of Common Praier as is to be song in Churches, wherein are used only these iiij. sortes of notes:—

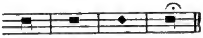

The first note is a strene note and is a breue, the second note is a square note and is a semybreue, the iij. a prycke and is a mynymne. And when there is a prycke by the square note, that prycke is half as muche as the note that goeth before it. The iiij. is a close, and is vsed at only yᵉ end of a verse."

John Day, who in 1560 printed the Church Service in three and four parts, and in 1562 the "Whole Booke of Psalmes," by Sternhold and Hopkins, used music types, as did also many of the printers who immediately succeeded him.

1575. Thomas Vautrollier, in Blackfriars, printed, among other works, Tallis and Byrd's "Cantiones," under a patent to the authors. The patent allowed them the monopoly of "ruling" paper, and as music was printed on ruled lines they managed to include it in their patent.

It would form an interesting episode in the history of printing if special attention were to be given to the progress of the art of printing music from types in Europe, so that detailed accounts of the chief works produced by the several "Fathers of the Chapel" from time to time might be given. Such a list would be entirely beyond the limits of the present article. It has therefore been thought advisable only to include a few of the chief and, to antiquarians, well-known works of the early printers before Playford.

After Vautrollier in chronological order comes Thomas Este, the assigne of William Byrd, living and working by Paul's Wharf. The chief of his productions were the following:—

1588. Byrd's Psalmes, Sonets and Songs; Younge's Musica Transalpina.

1589. Byrd's Songs of sundrie natures. Black Horse, Aldersgate Street.

1590. Watson's Italian Madrigalls Englished.

1591. Farmer's Divers and Sundry ways of two parts in one, to the number of 40 upon one playn song.

1593. Morley's Canzonets, 3 voices.

1594. Mundy's Songs and Psalmes. Morley's Madrigals, 4 voices.

1595. Morley's Ballets, 5 voices; Morley's Canzonets, 2 voices.

1597. Weelkes' Madrigals, 3, 4, 5, and 6 voices; Yonge's Musica Transalpina, 2nd Book; Morley's Canzonets, 4 voices; Kirbye's Madrigals.

1598. Wilbye's Madrigals, 1st Set. Morley's Madrigals, 5 voices; Weelkes' Ballets and Madrigals, 5 voices.

1600. Weelkes' Madrigals of 5 and 6 parts, as assigne of Thomas Morley; Weelkes' Madrigals of 6 parts; Dowland's Songs or Ayres of 4 parts, Book II.

1601. Triumphs of Oriana; Jones's First Book of Ayres.

1603. Robinson's Scoole of Musicke. [In tablature.]

1604. M. Este's Madrigals; Bateson's 1st Set of Madrigals.

1605. Pilkington's Songs or Ayres of 4 parts.

1606. Danyel's Songs.

1608. Youll's Canzonets; Croce's Musica Sacra, Englished.

1609. Wilbye's 2nd Set of Madrigals;[*] Rosseter's Lessons for Consort;[*] Ferrabosco's Ayres.

1610. M. Este's 3rd Set of Madrigals; Campian's Two Books of Ayres.

1611. Byrd's Psalmes, Songs, and Sonnets; Morley's Consort Lessons; Maynard's XII. Wonders of the World.

1612. Gibbons's Madrigals; Campian's 3rd and 4th Books of Ayres; Corkine's 2nd Book of Ayres; Dowland's Pilgrim's Solace.

1613. Ward's Madrigals; Pilkington's Madrigals.

1614. Lichfild's Madrigals.

1615. Robert Tailour's Sacred Hymnes.

1618. Bateson's 2nd Set of Madrigals; M. Este's 4th and 5th Set of Madrigals; Ravenscroft's Pammelia; Mason and Earsden's Ayres in the King's Entertainment.

1619. Vautor's first Set of Ayres.

1620. Martin Pearson's Private Musicke.

1621. Adson's Courtly Masquing Ayres.

1622. Thomas Tomkins's Songs of 3 and more parts; Attey's Ayres.

1624. M. Este's 6th Book (Anthems); Pilkington's 2nd Set of Madrigals.

Contemporary with him was Peter Short, at "The Starre, on Bread Street Hill." He produced—

1597. Morley's Introduction to Practical Musicke; Dowland's Songes or Ayres of 4 parts, Book I.; Morley's Canzonets, 5 and 6 voices; Holborne's Cittharn Schoole.

1598. Farnaby's Canzonets, 4 voices.

1599. Cavendish's Ayres for 4 voices.

---

[*] On the title-pages he calls himself Thomas Este, *alias* Snodham. He subsequently used the latter name only.

Digitized by Google

1601. Rosseter's Ayres; Jones's Second Book of Ayres.

1603. Dowland's Third Book of Songs or Ayres.

And William Barley, the assigne of Thomas Morley, in Little St. Helen's, and sold at his shop in Gratious Street.

1596. Booke of Tabliture.

1599. Farmer's Madrigals; Bennet's Madrigals.

1608. Weelkes' Ayeres or Phantasticke Spirites.

1609. Ravenscroft's Pammelia.

In later years Edward Allde printed—

1614. Ravenscroft's Briefe Discourse.

1615. Amner's Sacred Hymnes.

Following next on the list was Humphrey Lownes, who produced in 1608 Morley's Introduction; and in 1627 Hilton's Ayres or Fa las.

John Leggatt, London, was the printer in 1637 of "Sandys's Paraphrase on the Psalmes of David, with tunes by Henry Lawes."

Edward Griffin in 1641 printed Barnard's "Selected Church Music," a book of which no complete copy is known.

John Haviland, London, printed in 1636 Charles Butler's "Principles of Musik," a book remarkable for its quaint diction; and William Turner, Oxford, in 1634, the same writer's "Feminine Monarchie."

James Young, London—

1648. Henry and William Lawes's Choice Psalmes.

William Godbid—

1656. Matthew Locke's Little Consort.

1657. Walter Porter's Mottets. (No place of printing mentioned.) Gamble's Ayres and Dialogues.

1659. Gamble's Ayres and Dialogues, 2nd Book; Select Ayres and Dialogues.

1662. Dering's Cantica Sacra (Inner Temple); Courtly Masquing Ayres, various authors.

1668. Tompkins's Musica Deo Sacra (Little Britain).

1669. Henry Lawes's Ayres, Third Book.

1674. Dering and others, Cantica Sacra. (No place.)

William Hall, Oxford—

1661. Lowe's Directions for performance of Cathedral Service.

1664. Ditto, 2nd edition.

1668. William King's Songs and Ayres.

J. Heptinstall, London—

1690. Purcell's Songs in Dryden's Amphitryon.

1694. Purcell, &c. Don Quixote. Parts I. and II.

1696. Blow's Ode on the death of Dryden.

1697. Purcell's Ten Sonatas in 4 parts; Purcell's Ayres for the Theatre; Purcell's Te Deum and Jubilate for S Cecilia's Day, 1694.

Some writers declare that the Germans invented moveable music types, but this is not the case. It is certain that they were employed in Germany for books at a very early period. One of the first, if not actually the first book printed in Germany with music was entitled:—"Musicæ activæ micrologus, libris quatuor digestus, omnibus musicæ studiosis non tam utilis quam necessarius." Leipsic, 1517. Andreas Ornithoparcus. It was translated into English by Dowland in 1609.

When type was employed all musical notes were printed separately; quavers and semiquavers, no matter how many there were in succession, were all distinct. In or about the year 1660 John Playford invented what he called the "new tyed note," that is to say, he used a common stroke or strokes to connect a series of quavers or semiquavers. This improvement, which made music neater to the eye and easier to read, was soon adopted by other nations, the Dutch first, the French next, and after them the Germans. The Italians did not adopt it until many years later, for Marcello's Psalms, printed in 1734, had the old disjointed notes. To continue the record of improvements made in type music-printing, mention must next be made of William Pearson, who with John Heptinstall mentioned above, was employed by Henry Playford, the son of old John. In 1699, Playford published "Twelve New Songs, with a Thorow Bass to each Song, Figur'd for the Organ, Harpsichord or Theorbo, chiefly to encourage William Pearson's New London Character." This improvement of William Pearson's consisted in *matrices* for music type, so cut that the several pieces cast in them could be better "ranged" than formerly.

The commencement of the eighteenth century brought with it a change concerning music printing. Music types were considered ugly, and the pages they formed inelegant, and so a strong tide in favour of engraved music set in. In order to make the present sketch to a certain extent complete, and to show the state of music printing in its several branches, it will not be considered superfluous if we turn back to speak of engraved music, and of its progress up to the period at which we have traced type music.

Frescobaldi's "Toccate d'Intavolatura di Cimbalo et Organo, partite di diverse Arie e Corrente, Balletti, Ciacone, Passaghali," Rome, 1637, was one of the earliest books printed in Italy from copper plates; a facsimile specimen of the character of the engraving may be seen in the article "Leger-lines." Hawkins implies that the English adopted the practice from the Italians, in company with the Germans and the French.

Digitized by Google

"The English," says he, in speaking of Frescobaldi's publications, "also gave into it, as appears by a collection of lessons by Dr. Bull, Byrd, and others, entitled 'Parthenia,' already alluded to;" but this work was produced in 1611, twenty-six years before Frescobaldi's books, and it is not easy to see how a previous work can be indebted to a later one.

Type, however, was most generally used for printing music in England as well as abroad. In fact, there was only one engraver of music in London at the beginning of the eighteenth century, one Thomas Cross. The great labour of engraving music, the cost of the plates, and other incidental expenses, made music type preferable to the pocket of the producer, although it was inferior in appearance and elegance. Some Dutch printers are said to have discovered a means of softening copper so as to make it "susceptible of an impression from the stroke of a hammer or a punch, the point whereof had the form of a musical note." There were many works produced by this process, and for more than a quarter of a century, from the year 1700, the Dutch printers had the monopoly of work. Sir John Hawkins further states that "the difficulty of getting music from abroad, and the high duty on the importation of it, were motives to an attempt of a somewhat similar kind in England." The attempt included the use of pewter for copper, as the former material was more workable with punches than the latter. The enterprising publishers who carried this plan to a successful end were John Walsh, of the "Harp and Hoboy," in Catherine Street in the Strand, and John Hare, of the "Viol" in St. Paul's Churchyard. The first book printed from their plates was issued in 1710. The method of stamping which they introduced has been continued with scarcely any improvement in England to the present time. The Germans are more accomplished in this branch of the art of music printing than other European nations. Their punches are of more elegant shapes, and they can produce their work at a less cost than other folk. It is, therefore, not a matter for wonder when it is found that a great many English works are engraved and even printed in Germany in the present day.

Many of the books of Walsh and Hare were clumsily and inelegantly sent forth, but still these men kept the trade in their own hands, and outlived all attempts to beat them off the ground. Richard Mears attempted to rival them, but ruined himself. Walsh's son produced better works than his father, and with almost the same materials published much neater music. Better punches were cut by a Welshman named Phillips, who, with his wife and son, "improved the practice of stamping to so great a degree that music is scarce anywhere so well printed as in England" in his time. A native of Lapland, named Fougt, obtained about the year 1764, in England, a patent for printing with moveable types of his own cutting and founding, but he was undersold by the pewterplate music engravers, and compelled to leave the country.

The French engravers exhibited more taste and elegance than any others. A reference to any work printed in Paris about the years 1710—20 will show this. Their work was expensive, as music was cultivated only as a luxury in France. When it began to be popular it became necessary to produce cheaper work, but the artistic excellence which the engravers had attained in former times did not desert them, and their cheap work was less clumsy in appearance than that issued in England. When the French printed from music types they employed English founts, for they were the best in Europe, as they had been for a century and a half before, and are still to the present day. Any one who has leisure, will, and opportunity, may see by comparison that books printed in Venice, in Holland, and even in Spain, were set up from the same founts of type with which English books were printed at the beginning and nearly all through the seventeenth and eighteenth centuries.

A vast improvement in music type took place at the commencement of the present century. A series of oratorios in folio size, published by Jones and Lackington, at the "Temple of the Muses," Finsbury Square, were printed in beautiful, legible, and sharp outlined characters. It is not necessary to trace the art of music printing beyond that point, or to say more than that it has arrived at a pitch of elegance and usefulness probably undreamed of by the early and enthusiastic printers.

Proasma. An introduction, or a short symphony.

Probe (*Ger.*) A rehearsal.

Proceleusmaticus. A foot consisting of four short syllables, or of two Pyrrhics. [Metre.]

Program, *or* Programme. A list of the names of the pieces of music to be performed at a concert or other musical entertainment.

Progression. There are two kinds of progression, melodic and harmonic. Speaking in general terms, the former is a "succession of sounds forming a tune or melody," but the term is also applied to an "imitative succession of melodic phrases," that is to a melodic sequence. Harmonic progression is "the movement of one chord to another," and is diatonic or chromatic. The term is also sometimes used as synonymous with sequence

Progressionsschweller (*Ger.*) A contrivance of Vogler for gradually drawing out, and then in, the stops of an organ in their harmonic order, thus producing a crescendo, forte, and diminuendo.

Prolatio (*Lat.*) Prolation, one of the three divisions of mensurable music is, according to Franchinus, "essentialis quantitas semibrevibus ascripta." Prolatio is perfect or imperfect; perfect when the semibreve is divided into three minims, imperfect when divided into two. The perfect, or ternary division of the semibreve is called by some "prolatio major;" the imperfect, or binary division, "prolatio minor." The presence of a dot (punctus) in the time-signature (in temporali signo) shews that the prolation is perfect; the absence of the dot shows that it is imperfect. Perfect and imperfect prolation may each occur in perfect or imperfect time, that is, in time having three semibreves equal to the breve, or two. If the prolation be perfect, the measure will be compound.

Prolatio is the subdivision of a semibreve into minims, as Tempus is of a breve into semibreves, and as Modus is of a long into breves, or of a maxim into longs.

Promptement (*Fr.*) Quickly.

Prontamente, pronto (*It.*) Readily, quickly.

Prope media (*Lat.*) [Paramese.]

Proper-chant. An old name for the key of C major which had its *mi in B*, that is which had B for its leading note.

Proportio *(Lat.)* "Duorum numerorum ad invicem habitudo." The ratio of two numbers to each other. The discussion of the ratios of intervals formed one of the chief parts of mediæval treatises on music. Proportio is of three kinds: (1) Multiplex. (2) Superparticularis. (3) Superpartiens. *Proportio multiplex*, is when the larger number contains the smaller so many times without a remainder, as 2:1 (dupla), 3:1 (tripla), 4:1 (quadrupla). *Proportio superparticularis* is when the larger number exceeds the smaller by one only, as 3:2 (sesquialtera), 4:3 (sesquitertia), 5:4 (sesquiquarta). *Proportio superpartiens* is when the larger number exceeds the smaller by more than one, as, 5:3 (superbipartienstertias), 7:4 (supertripartiensquartas), 9:5 (superquadripartiensquintas). The following tables from Boethius, explain fully the kinds of proportio:

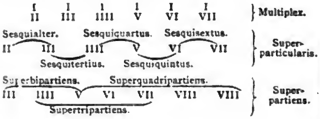

Thus, it will be understood, that instead of giving simply the ratio between two numbers, early writers on arithmetic and geometry, as well as music, coined a single word to express that ratio; for example, 17:5 was said to be *Triplasuperbipartiensquintas, i.e.*, that the larger number contained the smaller number three times (tripla) with two remainder (bipartiens). Again, *Triplasupertripartiensquartas* proportio, signified that the larger contained the smaller three times and three over, as 15:4, 27:8, &c., the last part of the compound word always pointing out the smaller of the numbers compared, or an exact multiple of it. Lastly, the addition of *sub* showed that the smaller number was compared to the larger, *e.g.*, 4:15 would be called *Subtriplasupertripartiensquartas* proportio. This system of proportion was used not only with reference to intervals but also to the comparative length of notes.

Proposta (*It.*) The subject of the fugue.

Prosa (*Lat.*); *Prose* (*Fr.*); *a prose.* A hymn sung after the gradual and before the gospel in the Roman Church. Some authors believe them to be of high antiquity and to have had their origin in popular hymns, sometimes metrical and written in the vulgar tongue. S. Cyprian uses the word in this sense in his life of S. Cæsar of Arles. "Adjecit et compulit ut laicorum popularitas psalmos et hymnos pararet, altaque et modulata voce instar clericorum alii Græce, alii Latine, *prosas* antiphonasque cantarent."

Other authors say that the prose was the result of setting words to the long neumes of alleluias, a specimen of which will be seen under the head Neume. From this cause a prose came to be called *Suite de l'alleluia*, or an *alleluiatic sequence*, or in short, a *sequence* (*sequentia*). Notker, a monk of S. Gall in the 9th or 10th century, is said to have introduced this custom of setting words to the "notæ pneumaticæ."

Their use spread with great rapidity, and about the twelfth century they began to assume the form of rhythmical and rhyming poetry set to a well defined melody. They then threw off the former rule that the prose should be if possible in the *mode* of the alleluia, and the result is that the music of the proses (a large number of which are in triple time) are some of the most beautiful melodies handed down to us.

Their use was limited by Pius V. to Easter-day ("Victimæ paschali laudes"), Whitsunday ("Veni Creator Spiritus"), Holy Sacrament ("Lauda Sion"), and Commemoration of the Dead ("Dies Iræ"), and the Seven Dolours B.V.M ("Stabat Mater"). These were written respectively by Notker,

Digitized by Google

St. Thomas Aquinas, Robert King of France, Thomas of Celano, and Jacopon.

Prosæ Sequentiæ (*Lat.*) [Prosa.]

Prosarium (*Lat.*) A book of proses.

Proscenium. (1) The quadrangular space behind the *logeum* or stage. (2) The stage front; all that part of the stage between the footlights and the curtain.

Proslambanomenos. (*Gk.*) *προσλαμβανόμενος.* [Greek Music.]

Prosody. [Metre.]

Prova (*It.*) **Probe** (*Ger.*) A rehearsal.

Provençales. Troubadours of Provence, in which country the rhymers and minstrels of mediæval times seem to have had their origin. So little of their music has been handed down that it is difficult to form a just opinion on its merits. It is certain, however, that they paid little regard to the laws of music as laid down by their contemporaries, but constructed rhythmical melodies, in whatever scale was pleasing to the popular ear. The date of the rise of provençal poetry is by some stated to have been as early as the tenth century, but was more probably the twelfth.

Psallettes. *Maîtrises* (*Fr.*) Schools in which, at the time of Franco of Cologne, descant was taught.

Psalm-Melodicon. An instrument invented in 1828 by Schuhmacher Weinrich. It was a wind instrument with keys and ventages, imitating the tone of several orchestral instruments. In 1832 it was improved by Leo Schmidt, and named the Apollo-Lyra.

Psalmody. [Hymn tunes.]

Psalmistæ. An order of clergy instituted about the time of the introduction of the Cantus Ambrosianus; for the special duty of singing from the Ambo, such music as would have been marred by the singing of the congregation. At the council of Laodicea, held between the years 360 and 370, it was ordered that no one but the canonical singing-men should presume to sing in the church; and by a canon of the fourth council of Carthage the following form of words was prescribed for their ordination. "Vide ut quod ore cantas corde credis, et quod corde credis operibus comprobes."

Psalterium (*Lat.*) (1) A *Psalter*, generally with musical notation above the words. (2) A Psaltery. [Nebel.]

Psaltery. [Nebel.]

Psaltriæ (*Lat.*) Female musicians who sang and played during a banquet

Pulpitum (*Lat.*) (1) The λογεῖον, or stage of the Greek Theatre. A wide but shallow space in the shape of a parallelogram, in the middle of which the chief actors usually spoke. Behind it was another quadrangular space termed the *proscenium*, not so wide as the *logeum*. The remainder of the *logeum*, right and left of the *proscenium*, had at the back a wall rising as high as the uppermost tier of spectators; at the front, a way down to the orchestra, or space set aside for the chorus. (2) A motet. *Pulpitre* (*Old Fr.*) [Motet.]

Pulsatile. A term applied to instruments of percussion, such as the drum, gong, cymbals, &c.

Pulsator organorum. An organ-player, at the time when the keys were very large and had to be struck sharply. [Organ.]

Punctus (*Lat.*) A point or dot. A punctus was of various kinds. The *punctus additionis*, or *augmentationis*, added to a note one half of its value, as does the modern dot, making an *imperfect* note (or binary), into a *perfect* (or ternary). The *punctus divisionis*, sometimes also formed thus ✓, separated ternary measures. The *punctus perfectionis*, similar in shape to the former, was in effect the same as a *punctus additionis*. The *punctus prolationis* was the dot inserted in the circle or half circle ⊙ (·, which distinguished *prolatio perfecta* from *prolatio imperfecta*.

Punkt (*Ger.*) Dot. *Punktirte Noten*, dotted notes

Punta (*It.*) The point, as *colla punta dell'arco*, with the point or tip of the bow.

Puntato (*It.*) Pointed, detached, made staccato.

Pupitre (*Fr.*) [Pulpitum § 2.]

Pyknon (*πυκνόν*). The close note. (1) A name given to those half or quarter tones which came together in the chromatic and enharmonic genera of the Greeks. (2) In mediæval music, a semi-tone.

Pyramidon. An organ stop of 16 ft. or 32 ft. tone, the pipes of which are closed at the top, and pyramidical in shape, the top being more than four times the width of the mouth. From a pipe only 2 ft. 9 in. in length, 2 ft. 3 in. square at the top, and 8 inches at the block, the note C C C is produced.

Pyrophone. An instrument invented by Kastner, the sounds of which are produced by jets of gas burning under tubes.

Pyrrhic. (1) A dance among the Greeks, danced by boys in armour, accompanied on the lyre or flute. In it was much warlike gesticulation and rapid movement. (2) A metrical foot consisting of two short syllables. [Metre.]

Pythagoreans. The followers of the system of Pythagoras, in which the consonance or dissonance of an interval was judged by the ratio of the vibrations without appeal to the ear. The Aristoxenians, on the other hand, held that the ear should be the sole judge of right or wrong in music. The former were called Canonici, because they appealed to the monochord or harmonic canon for their laws, the latter Musici, because they made the ear and practice their guide.

Digitized by Google

Q.

Quadrate (or B quadratum). The sign b, used originally to raise B rotundum ♭, one semitone. Hence arose its general use for the raising of all flattened notes, as exemplified in its modern form of a natural, ♮

Quadratum (*Lat.*) A breve, 𝅜

Quadrible. [Quatrible].

Quadricinium *or* **Quatricinium.** A composition in four parts.

Quadrille. A well-known dance, consisting of five movements; Le Pantalon, La Poule, L'Eté, La Trenise (or la Pastourelle), La Finale. [Dancing.]

Quadruple Counterpoint. Counterpoint of four parts, so constructed that all the parts may be transposed among themselves without transgressing the laws of progression. A perfect piece of this kind of counterpoint will be capable of twenty-four different dispositions of the parts. It is only feasible at the interval of the octave. [Counterpoint.]

Quadruple Croche (*Fr.*) A hemi-demi-semiquaver, 𝅘𝅥𝅱

Quarte (*Fr.*) The interval of a fourth.

Quarta (*It.*) The interval of a fourth.

Quarta (*Lat.*) An interval of a fourth, as *major, minor, abundans* (superflua), a major, minor, or augmented fourth. *Quarta modi*, *Quarta toni*, the fourth of the scale, the modern subdominant.

Quart de mesure (*Fr.*) A crotchet rest.

Quart de soupir (*Fr.*) A semiquaver rest.

Quarte de nazard (*Fr.*) An organ stop of 2 ft. length, so called because it is a fourth above the nazard or twelfth.

Quarte du ton (*Fr.*) The subdominant.

Quarter note. A crotchet, ♩

Quarter tones. A general name of intervals less than a semitone, introduced into enharmonic instruments.

Quartet. (1) A composition in four parts, or for four performers. (2) Part of a movement sung by four voices *soli* as opposed to *coro*. (3) A composition for four instruments, in complete symphonic form, consisting of an introduction (occasionally), (1) allegro, (2) andante or adagio, (3) minuet and trio, or scherzo, (4) finale. Each of these movements has its form. [Sonata form.] By far the largest number of instrumental quartets are composed for two violins, a viola, and a violoncello, not only on account of the smoothness and evenness of their individual and collective tone, but also because of the compass of each instrument allowing an ever-varying disposition of the harmony. The complete quartet, quintet, or sestet, is to chamber music what the symphony is to concert music. Both are, in their way, the highest production of which instrumental music, apart from vocal, is capable. Attempts at a separation of instrumental and vocal music were made in this country as early as the beginning of the seventeenth century, by John Jenkins (b. 1592), one of whose three-part *Fancies* has been reprinted by Hullah, in his "Transition Period of Musical History," p. 194. Allegri, who died in 1652, composed a quartet for two violins, a viola, and basso di viola. But the string-quartet did not reach its prime until the time of Haydn and Mozart, simultaneously with the settlement of the form of the symphony.

Quartetto (*It.*) [Quartet.]

Quartfagott (*Ger.*) A kind of bassoon, a fourth lower than that commonly in use.

Quartflöte (*Ger.*) A small flute, a fourth higher in pitch than the common flute.

Quartgeige (*Ger.*) A small sized fiddle. *Violino piccolo.*

Quarto d'aspetto (*It.*) A semiquaver rest, 𝄿

Quasi (*It.*) As if, or in the style of. Used to qualify certain terms, as, *quasi allegretto*, somewhat allegretto; *quasi sonata*, a composition very similar to a sonata, but in which sonata form is not strictly adhered to; *quasi fantasia*, a piece in which form is displaced by the style of a fantasia.

Quatre mains, à (*Fr.*) For four hands.

Digitized by Google

Quatrible (*Old Eng.*) To descant by singing fourths on a Plain Song (see Quinible).

Quatricinium. [Quadricinium.]

Quatricroma (*It.*) [Quadruple croche (*Fr.*)]

Quattro mani, a (*It.*) For four hands.

Quatuor (*Fr.*) [Quartet.]

Quaver. The eighth part of a semibreve, ♪ *Achtelnote* (*Ger.*) *Croche* (*Fr.*)

Querflöte (*Ger.*) **Flauto traverso** (*It.*) The flute played sideways, as opposed to the flute which was blown at one end, and held straight in front of the performer. [Flute.]

Querpfeife (*Ger.*), **Fiffaro, Fifre.** The Swiss fife; a small kind of flute, with six holes, but no keys. It has an incomplete compass of two octaves.

Querstand (*Ger.*) [False Relation.]

Queue (*Fr.*) (1) The tailpiece of a violin or other instrument. (2) The tail of a note.

Quick Step. A quick march.

Quinible (*Old Eng.*) To descant by singing fifths on a Plain Song (see Quatrible). In Chaucer's "Miller's Tale," it is said of Absolon:

> "In twenty manere coud he trip and dance,
> (After the scole of Oxenforde tho)
> And with his legges casten to and fro;
> And playen songes on a smal ribible;
> Thereto he song sometime a loud quinible."

Quindecima. [Quinta decima.]

Quint. (1) The interval of a fifth (2) An organ stop, sounding a fifth above the foundation stops, of $5\frac{1}{3}$ ft. length on the manuals, $10\frac{2}{3}$ ft. on the pedal. It should not be used without a double diapason, to which it forms the second natural harmonic, or twelfth. It is sometimes used on the pedal organ without a double diapason (32 ft.), but with questionable effect.

Quintabsatz (*Ger.*) A half close. The imperfect cadence, the penultimate chord of which is a tonic triad; the final chord, a dominant triad. [Cadence.]

Quinta decima (*Lat.*) (1) The interval of a double octave. (2) An organ stop, sounding the double octave of the foundation stops, fifteenth.

Quinta modi (*Lat.*) The fifth of the scale. The modern dominant.

Quintatön (*Ger*). An organ stop consisting of closed metal pipes, of a small scale, so voiced that the twelfth is heard with the ground-tone.

Quinta toni (*Lat.*) [Quinta modi.]

Quinte, Quintsaite (*Ger.*) *Chanterelle* (*Fr.*) The E string of a violin. The lowest string of violoncello and viola being C, A is their fourth string, hence the higher string of the violin came to be called the Quinte, or Quintsaite.

Quinterna, *or* **Chiterna.** A species of guitar not unlike a violin in shape, having three, or four, or five pairs of catgut strings, and sometimes two single strings covered with wire in addition, played with the fingers, not with a plectrum. About two centuries ago it was commonly used in Italy by the lower orders of musicians and comedians.

Quintet. (1) A composition in five parts, or for five performers. (2) Part of a movement sung by five voices *soli*, opposed to *coro*. (3) A composition for two violins, two tenors, and a violoncello; or two violins, a tenor, and two violoncellos; or two violins, a tenor, a violoncello, and double bass, having the same form as a sonata.

Quintfagott (*Ger.*) [Basson quinte.]

Quintfuge (*Ger.*) A fugue, the subject of which is answered at the interval of a fifth.

Quintole. A group of five notes to be played in the time of four—

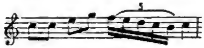

Quinton (*Fr.*) The five-stringed viol.

Quintoyer (*Old Fr.*) to descant at the fifth; to quinible.

Quintuor (*Fr.*) [Quintet.]

Quintviola. An organ stop of the gamba species, but of the pitch of a quint, or of a twelfth.

Quire (*Old Eng.*) The collective title of the body of trained and authorized singers in a church.

Quirister. A member of a Quire, whether man or boy. [Chorister.]

Quodlibet. (1) A sort of Fantasia; (2) a pot-pourri. (3) A Dutch concert.

At the annual re-unions of the members of the Bach family singing or improvising *quodlibets* was one of the amusements indulged in.

Digitized by Google

R

R, abb. for right, as R.H., with the right hand.

Rabanna, *or* **Rabani.** A kind of Indian drum, of a small size, beaten by the hand.

Rabbia, con (*It.*) With fury.

Rackett, Rankett. (1) An obsolete wind-instrument of the double bassoon kind, it had ventages but not keys. It was not of an extended compass, being incapable of producing harmonics. It was a double-reed instrument, the reed being at the end of a tube through which the player blew. The tone was nasal and produced with difficulty. The rackett was improved by Denner at the beginning of the last century, but was not able to hold its own against the then much superior *bassoon*. (2) An organ stop of 16 ft. or 8 ft. pitch now obsolete.

Raddolcendo, Raddolcente (*It.*) With gradual softness and sweetness.

Raddoppiamento, Raddoppiato (*It.*) The doubling of an interval or part.

Radical bass. The fundamental bass, ground note, or root of a chord.

Ragoke. A small Russian horn.

Rags, Raginees. Certain Hindoo melodies founded on fixed scales. They were of three kinds, *sumpoornu*, or those comprising seven notes in a determined succession; *khadoo*, or such as comprised six notes; *oodoo*, or those comprising only five notes.

Rall., abb. for rallentando.

Rallentamento (*It.*) At a slower pace.

Rallentando, Rallentato (*It.*) Getting gradually slower.

Rank of pipes. A row of pipes (of an organ) belonging to one stop. A stop is said to be of two, three, four, or five *ranks*, according to the number of rows of pipes under the control of its one register.

Rant. An old dance; a sort of country dance. This name is often attached to tunes to which country dances were performed. It is perhaps a corruption of the word *coranto*.

Ranz des vaches (*Fr.*), *Kuhreihen* (*Ger.*) The tunes or flourishes blown by Swiss shepherds on their cow-horns or Alpine-horns (long tubes of fir-wood), as signals to the animals under their charge, such as the following:—

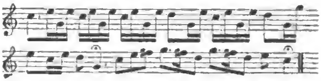

The notes marked F♯ are not properly thus represented, they being the natural harmonic lying between E and G, consequently a sound between F and F♯.

Rapidamente, Rapidità, con, Rapido (*It.*) With rapidity.

Rasgado (*Sp.*) To *sweep* the strings of a guitar with the thumb, for the purpose of producing a full chord, *arpeggio*.

Rastral. [Rastrum.]

Rastrum (*Lat.*) (*Ger.* Harke.) A rake. The name given to the five-pointed instrument for ruling the stave.

Räthselcanon (*Ger.*) A riddle-canon. *Canon ænigmaticus*. One part and the number of parts being given, the student to write it out in full.

Ratio (*Lat.*) Relation or proportion. [Acoustics.] [Proportio.]

Rattenendo, Rattenuto (*It.*) Restraining or holding back the time.

Rauscher (*Ger.*) A passage of repeated notes.

Rauschpfeif, Rauschquint, Rauschwerk, Rauschflöte, Ruszpipe. A stop in old organs of two ranks of pipes, consisting of a twelfth and fifteenth, or a fifteenth and octave twelfth.

Ravanastron. A stringed instrument played with a bow; in use among the Buddhists. [Violin.]

Ravvivando. (*It.*) Reviving. Becoming again animated. *Ravvivando il tempo*, quickening the time.

Re. The name of the second note of the scale. [See Notation, early systems of.]

Digitized by Google

Real Fugue. A strict fugue. The term is now used in opposition to a *tonal* fugue. The answer in a real fugue being a fifth higher or a fourth lower than the subject, note for note; that in a tonal fugue being so far altered that dominant answers tonic and *vice versâ*. So that in a tonal fugue, a subject occupying a compass of five notes, namely from a tonic to its dominant, has to be answered in a compass of four notes, namely from the dominant to the tonic lying above it. [Fugue, § 3.]

Rebab. [Rebec.]

Rebec, *or* **Rebeck.** The English name of a three-stringed instrument played with a bow. It was of Arabian or Turkish origin and was introduced into Spain by the Moors, under the name *rabel* or *rebel*. It found its way thence, in the ninth or tenth century, into Italy under the name of *rebica*, and into England under the name *rebec*. In other European countries it was variously called rebeb, or *rebebe*, *reberbe*, *rebesbe*, *rubebe*, or *erbeb;* in Egypt and Asia, *rebab*. In its earliest form it probably had a long neck and small round body, made of cocoa-nut shell, or some such material, over which parchment was stretched to form the sound-board. After its introduction into Europe, the third string was added, for although the Persians have now a three-stringed *rebab*, the older form was probably only two-stringed. After its introduction into England, the rebec gradually assumed the form of a viol, of which instrument it was the precursor. [Violin.]

Re bémol (*Fr.*) The note D♭. *Re bémol majeur*, the key of D♭ major.

Rebibe, Rebible (*Old Eng.*) A small rebec or three-stringed viol. [Rebec.]

Recheat. A hunting signal, which recalls the hounds.

Recht (*Ger.*) Right, as *rechte Hand*, the right hand.

Recitando, Recitante (*It.*) In the manner of a recitative. As if reciting.

Recitatif (*Fr.*) [Recitative.]

Recitative. Musical declamation. An art lost to Europe in the destruction of Greek music, and not revived till the early part of the seventeenth century. It grew out of the *aria-parlante* or *monody*, which was an attempt on the part of certain Florentine *dilettanti* to restore the ancient recitation of poetry. The names associated in this work, which exercised such a wonderful influence over the art of music, were Vecchi, Galilei, Caccini, Peri, Cavaliere, and Monteverde. It is impossible to point to any one of these as the actual inventor of *recitative*, because the style cannot be said to have been perfected until it had received the free handling of Carissimi and Scarlatti, but their names deserve to be held in reverence. Recitatives were for a considerable period accompanied, at performances both of oratorios and operas, on a harpsichord with a double bass supporting the *basso continuo*. On the removal of harpsichords from our theatres and concert-rooms, which took place at the close of the last century, *an arpeggio chord* on the violoncello was substituted for the harpsichord-part, a double-bass (as before) sustaining the lowest note of each chord. If a band did not possess a very excellent violoncellist these arpeggio chords were often cruelly out of tune. Hence modern composers had no choice but to use soft chords in four-parts played by the whole string-band. This has been gradually followed by the use of the full band in recitatives, and as a natural result, pure *declamation* is to some extent merging into a semi-strict *arioso*. It is quite possible that this is what the authors of *aria parlante* were aiming at in their early efforts. [Opera.]

Recitativo (*It.*) [Recitative.] *Alla recit.*, in the style of a recitative.

Recit. accomp. Recitative with accompaniment.

Recit. secco. A recitative supported only by a chord from a violoncello or double bass, or by a cembalo. [Recitative.]

Recit. stromentato. Recitative accompanied by a band.

Record, to, Recorders (*Old Eng.*) The verb *to record* is used with reference to the singing of birds, as in Shakspeare (Pericles, Act IV.)

> . . . . . "or when to the lute
> She sung, and made the night-bird mute
> That still *records* with moan;" . . . .

The instrument "Recorder" was originally a *flageolet* or *tibia minor*, but the name was afterwards used as synonymous with *flute*.

Recte et retro. [Per recte et retro.]

Re dièse (*Fr.*) The note D♯.

Redita (*It.*) A return. A repeat.

Redondillas (*Sp.*) Roundelays.

Redoublement (*Fr.*) *Raddoppiamento* (*It.*) A doubling of an interval or part.

Redowa, Redowak, Redowazka. A Bohemian dance, originally in $\frac{2}{4}$ and $\frac{3}{4}$ time alternately. The time was afterwards altered, and the dance was made into a sort of polka.

Reductio modi (*Lat.*) (1) The bringing back a transposed mode to its original pitch. (2) The conversion of an old mode into its corresponding modern scale.

Reed. A thin strip of metal or cane set in vibration by a current of air; the vibrations so caused, at the same time, dividing the current of air into rapid discontinuous puffs which produce a musical sound. The reed itself does not produce the sound, but is only

Digitized by Google

a means of obtaining the sound from the current of air directed against it. "It is constructive, not generative" (Tyndall on Sound, p. 192). Reeds are of two kinds, *striking* and *free*. A *striking* reed is rather larger than the aperture, and is placed on that side of it against which the air is directed, being slightly bent upwards at its unattached end. As the current of air attempts to pass by it, the reed is forced against the sides of the aperture, and the progress of the air is suddenly checked, but the elasticity of the reed causes it immediately to recover its former position, when the current of air again rushes through, and so on. A *free reed* is of such size that it will freely pass in and out the aperture. The current of air forces it upwards, and its own elasticity restores it to its place again, and so on.

The *striking* reed is that commonly used in an organ, the *free reed* in a harmonium.

The tone of reeds is greatly intensified by the addition of a pipe or tube, care being taken that the tube should contain a column of air whose vibrations synchronise with the note produced by the reed, or with one of its overtones. In instruments of the horn or trumpet class, the lips perform the function of the reed, the notes produced being the fundamental note of the tube and its harmonics or overtones. The human voice is a reed instrument. [Larynx.]

Reed instruments of an orchestra. Oboes, clarinets and bassoons, with others of their class.

Reed of harmonium. [Reed.]

Reed of organ. [Reed] and [Organ, § 15.]

Reed-stops [Organ, § 15.]

Reel (*Old Eng.*) *Kreol.* A lively rustic dance, popularly supposed to be Scotch, but probably of Scandinavian origin. The Danish *kreols* are very similar to the reel. [Country dance.]

Refrain [Burden.] [Chorus.] [Ballad.]

Regals, Rigals, Rigoles. These terms seem to be synonymous with "Portative Organs," although distinguished by some authors. The word regal is supposed to have come from *rigabello*, mentioned in the following passage: "In Æde Sancti Raphaelis Venetiis, instrumenti musici cujusdam forma extat, ei nomen *rigabello;* cujus in ecclessiis usus fuerit ante organa illa pneumatica quæ hodie usurpantur. *Rigabello* successit aliud quod *turcello* dictum est, cujus Venetias usum induxit homo Germanus." Regals had generally only one row of pipes, and were probably used to support the treble voices. In an inventory of the musical instruments of Henry VIII., taken after his death (Sir Henry Ellis's original letters. Second series. Vol. I., p. 272), we read of "thirteen pair of single regalls," and "five pair of double regalls." Double regals had two rows of pipes.

Bernard Gates in 1767 received a salary as "tuner of regalls" in the Royal Chapel, but in 1770 he is called "tuner of organs.' Snetzler the great organ builder could remember the regals in use in Germany.

Regales de bois (*Fr.*) Claque-bois. [Ligneum Psalterium.]

Regel der Octav (*Ger.*) [Rule of the octave.]

Register. An organ stop; (1) in a limited sense, "the handle on which is written the name of the stop"; (2) in a general sense, a stop or "the pipes belonging to, and acted upon by one slider."

Register (of an organ.) A frame through which long trackers pass. [Organ, § 8.]

Register (of a voice.) (1) Class; as, soprano, &c.; (2) Portion; as, head, chest, &c.; (3) Compass; as, high, low, &c.

Règle de l'octave (*Fr.*) [Rule of the octave.]

Regular form. A work is said to be "not in regular form" if its subjects and their disposition depart from the plan or form conventionally considered most suitable to a composition of its kind.

Regular fugue. A *strict*, as opposed to a *free* fugue, or one in which the laws are not strictly obeyed. [Fugue.]

Regular motion. Similar motion.

Regulation (of a keyed instrument). The adjustment of the touch; in the pianoforte, by means of the regulating-pin; in the organ by means of leather buttons on a tap-wire, which when turned round shorten or lengthen parts of the action.

Rehearsal. *Probe* (*Ger.*), *Prova* (*It.*) A general practice before a performance. *Full Rehearsal*, a rehearsal at which soloists, band and chorus are present. *Public rehearsal*, a rehearsal to which the public are admitted.

Relatio non harmonica (*Lat.*) [False relation.]

Relative chord. A common chord made up of notes taken from the scale, *e.g.*:

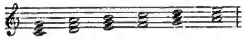

The chords of D minor, E minor, F major, G major and A minor are therefore relative to the chord or scale of C, these being the only common chords which can be made from the scale of C.

Relative key. A key whose tonic chord is a relative chord; that is to say, a key whose first, third, and fifth degrees form a common chord made up of notes of the key to which it is related. Thus D minor, E minor, F major, G major, and A minor are relative keys of C. The first, third, and fifth of each of these scales forming one of the relative chords of C.

Religiosamente, Religioso (*It.*) In a religious or devotional manner.

Remplissage (*Fr.*) A filling up. Intermediate part.

Renversement (*Fr.*) Inversion.

Renvoi (*Fr.*) A repeat.

Repeat. *Wiederholungszeichen* (*Ger.*) A sign that a movement or part of a movement is to be twice performed. That which is to be repeated is generally included within the sign of two or four dots in the spaces, thus,

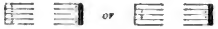

When the performer does not, on repeating, go so far as the last dot-sign, but finishes at a previous cadence, it is usual to write over the repeat, Da Capo, placing a pause and *fine* over the chord at which the performer is to stop. If the signs of the repeat do not coincide with a well-defined portion of a movement the sign 𝄋 is sometimes added thus,

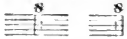

A few bars are sometimes marked *bis*,

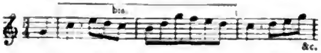

but this sign is only used over a very short phrase.

Répétition (*Fr.*) A rehearsal.

Repetizione (*It.*) Repetition, as *senza repetizione*, without repeating.

Replica (*It.*) Repetition, as *senza replica*, without repeating; *con replica*, with repetition.

Replicato (*It.*) Repeated.

Reply. The answer in a fugue, the subject being called *principal*.

Réponse (*Fr.*) The answer in a fugue.

Repos (*Fr.*) A pause.

Reprise (*Fr.*) (1) Burden of a song. (2) *Reprise d'un opéra*, the reproduction of an opera. (3) A repeat.

Requiem (*Lat.*) A name given to the "Missa pro defunctis," because the words "requiem eternam dona eis," occur in it.

Research. An extemporaneous performance on the organ or pianoforte in which the leading themes or subjects in the piece to which it serves as prelude are suggested and employed.

Resin, *or* **Rosin.** A gum, the viscid exudation of certain trees, chiefly of the fir tribe, which is obtained in large quantities by cutting away part of the bark, a vessel being placed below to catch the gum as it exudes. When purified and prepared it is used to rub over the hair of a bow, the surface of which it renders rough and so enables it to "grip" the strings. [Colophony.]

Resolution. (1) The moving of a discordant note to another which produces a satisfactory effect. This is done sometimes by taking the discord downwards one degree, as,

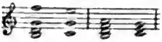

sometimes by taking it upwards, as

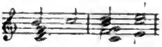

Resonance. [Acoustics, § 21.]

Resonance-body, Resonance-box, *Resonanzboden* (*Ger.*) The hollow part of a stringed instrument which reinforces the sound of the vibrating strings. Its shape is of the utmost importance, and, in the case of the violin has only been definitely fixed after great practical and scientific research. The resonance-box has certain openings to admit of the escape of the reinforcing vibrations.

Respiro (*It.*) A breath. A semiquaver rest.

Response. The answer to a versicle in the Church Service. The following are the signs for these words, ℣, ℟.

Responsorium (*Lat.*) A response-book; a choir-book containing the music of the versicles and responses.

Rests. Signs enjoining the silence of a performer for a given length of time. Each note has its corresponding rest, *e.g.*:

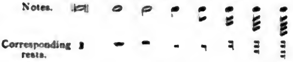

There is, unfortunately, no distinctive sign for a bar's rest. If the bar contain less than a semibreve, as in $\frac{3}{4}$, $\frac{3}{8}$ times, &c., a semibreve rest denotes a bar's rest; but, of course, rest for part of a bar is denoted by its proper signs, thus:—

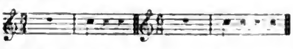

If the bar contain more than a semibreve, a bar's rest is usually denoted by the sign of a *breve rest*, but this is not universally adhered to, as some authors use a semibreve rest for a bar's rest in $\frac{3}{2}$ time.

Dots may be affixed to rests and have the same effects upon them as upon notes, *e.g.*: 𝄽 · is equal to a three-quavers' rest; 𝄽 · ·, equal to a seven-semiquavers' rest. For an account of earlier forms of rest see *Pausa*.

Resultant Tones. [Acoustics, § 19.]

Retardation. (1) A gradual slackening of pace in the performance of a passage. (2) The holding on of a concordant note into the succeeding chord, in such a manner that it becomes a discord, which is resolved *upwards*. A discord of retardation is thus opposed to a discord of suspension, the latter being resolved downwards, *e.g.*:

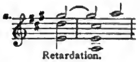
a. Retardation.

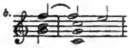
b. Suspension.

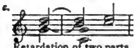
c. Retardation of two parts. (or a double Retardation.)

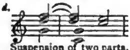
d. Suspension of two parts. (or a double Suspension.)

Three or more parts may be retarded or suspended, and retardations and suspensions may occur in the same chord.

Retro. [Per recte et retro.]

Retrogrado (*It.*) [Retrogradus.]

Retrogradus (*Lat.*). *Motus retrogradus*, reading music backwards. *Imitatio retrograda*, imitation *per recte et retro, q.v. Contrapunctus retrogradus*, counterpoint *per recte et retro.*

Retto (*It.*) Direct, as *moto retto*, direct or similar motion.

Reveil, Revelly (*Old Eng.*) Music which wakens from sleep. A signal given by drum to soldiers at dawn (from *Lat.* revigilare).

Reveillé, Réveil (*Fr.*) [Reveil.]

Reverse motion *or* **movement** Movement by inversion of intervals.

Rf., *abb. of* Rinforzando.

Rhapsodes (ῥαψῳδοί). Wandering minstrels in ancient Greece, of the Ionian race, who formerly recited epics in public places, holding in their hands a staff (ῥάβδος) as a sign of their calling. It is doubtful whether the rhapsode had always a musical accompaniment to his recitation, as one of his hands would be occupied by his staff. Rhapsodical recitation must be regarded as the forerunner of stage-acting, and as forming when conjoined with the Bacchic chorus, the complete Greek drama.

Rhapsodie (*Ger.*)
Rapsodie (*Fr.*)
Rhapsody.
} A composition of irregular form, and in the style of an improvisation.

Rhythm. [Metre.]

Rhythmopœïa (*Gk.*) The due arrangement of *arsis* and *thesis* in metre. [Metre.]

Ribattuta, Ribattitura (*It.*) A beat (w) or passing note.

Ricercari (*It.*) (1) Difficult passages or flourishes. (2) Exercises.

Ricercata (*It.*) (1) A sort of fantasia or toccata. (2) *Fuga ricercata*, a fugue containing nothing but various treatments of the subject; a fugue without episodes; a master-fugue, highly elaborated and containing all possible scientific feats. [Fugue.]

Ridotto (*It.*) Reduced from a full score.

Rifacimento (*It.*) A reconstruction or restoration of a work.

Rifiorimenti (*It.*) Extemporaneous embellishments.

Rigabellum (*Lat.*); *Rigabello* (*It.*) [Regals.]

Rigadoon, Rigaudon (*Fr.*) An old lively dance performed by a man and a woman, as the jig is danced in some places. It is said to have been invented in Provence by one Rigaud, and from him to have taken its title, but as there are more examples of melodies by English composers than by French, it is not unlikely that the word is English, coming from the same root as "Rig," which means wanton, lively. The character of the rigadoon would justify this derivation.

The two dances without authors' names are from a work entitled "The newest Minuets, Rigadoons, and French Dances, perform'd at Court and Publick Entertainments. London, 1716."

Rigadoon. Mr. H. Purcell. Choice Lesson, 1705.

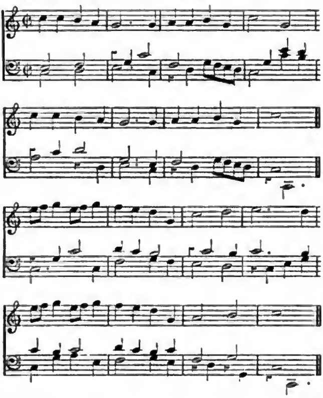

Rigadoon, 1716.

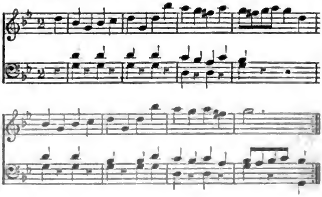

Digitized by Google

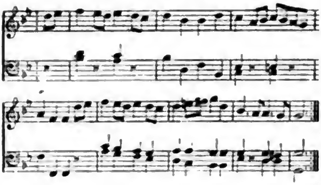

Rigadoon, 1716.

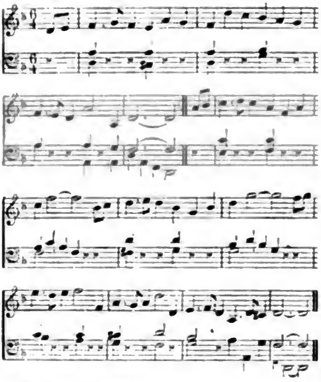

Rigols. [Regals.]

Rigore (*It.*) Strictness, as, *al rigore di tempo*, in strict time; *con rigore*, with exactness.

Rigoroso (*It.*) [Con rigore.]

Rikk. A small tambourine of modern Egypt.

Rilasciando (*It.*) Relaxing the time.

Rilch, Rilka. A Russian lute.

Rinf., *abb.* of Rinforzando.

Rinforzando, Rinforzare, Rinforzato (*It.*) Reinforcing, or strengthening the power and emphasis of a musical sentence.

Ringelpauke (*Ger.*) A sistrum having rings on bars, which rattled when the instrument was shaken. [Sistrum.]

Ripienist. A performer who only assists in the *ripieno* parts.

Ripieno (*It.*) (1) An additional or *filling-up* part. Any part which is only occasionally required for the purpose of adding to the force of a *tutti*, is said to be *ripieno*. (2) A mixture stop on Italian organs; as, *ripieno di due, tre, quattro, cinque*, &c., a mixture stop of two, three, four, five ranks, &c.

Ripresa (*It.*) (1) A reprise or burden. (2) A repeat.

Risentito (*It.*) With expressive energy.

Risolutamente, Risoluto, Risoluzione, con (*It.*) With resolution. *Risolutissimo*, very resolutely.

Risonanza (*It.*) [Resonance.]

Risposta (*It.*) A reply or answer to a fugue-subject.

Ristretto (*It.*) Stretto. [Fugue.]

Risvegliare (*It.*) To rouse up, awaken, reanimate; *risvegliato*, in an animated manner.

Rit., abb. of *Ritardando*.

Ritardando, Ritardato (*It*). With gradually increasing slowness of pace.

Ritardo (*It.*) [Retardation (1).]

Riten., abb. for Ritenendo, &c.

Ritenendo, Ritenente, Ritenuto (*It.*) Holding back the pace.

Ritmo di tre battute (*It.*) In the rhythm of three beats, or triple measure. An expression used when the group of accents is formed by three bars in rapid time, *e.g.*:

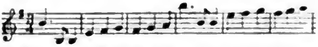

The above passage is marked by Beethoven (Choral Symphony) *ritmo di tre battute*, but,

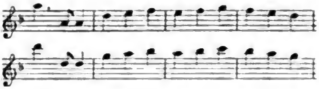

is in *ritmo di quattro battute*. The former is practically in compound triple time, the latter in compound common time.

Ritornello. [Interlude.]

Riverso. [Rovescio.]

Rivolgimento, Rivoltato, Rivolto (*It.*) The inversion or transposition of the parts in double counterpoint.

Robusto (*It.*) Robust, strong, powerful. *Tenore robusto*, a Tenor voice of a powerful or massive quality of tone.

Roccoco, Rococo (*It.*) Old fashioned, queer.

Rock harmonicon. An instrument, the sounds of which are produced by striking graduated lengths of rock-crystal with a hammer.

Rohr (*Ger.*) A reed.

Rohrflote (*Ger.*) Reed-flute. An organ stop consisting of closed pipes, the tone of which is slightly reedy in quality, but very sweet.

Rohrwerk (*Ger.*) Reedwork, the collective name of the stops consisting of reed-pipes, as opposed to flue-work or stops consisting of flue-pipes. [Reed.] [Flute.] [Organ.]

Rôle (*Fr.*) The extract from a drama which an actor has to commit to memory. From *Lat. rotula.* The part assigned to an actor.

Roll. The regular and rapid beating of a drum by two sticks so as to make the sound as far as possible *continuous.* It is commonly expressed thus:

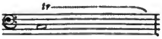

In the case of a tambourine, the *roll* is produced by a rapid succession of blows from the knuckles, as the hand is swung backwards and forwards.

Rolle (*Ger.*) A *run*, a group or series of groups of short notes. In vocal music they are mostly sung to one syllable of a word, *e.g.*:

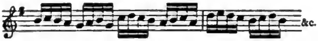

Rollo, Rollando (*It.*) Roll of a drum or tambourine. [Roll.]

Romance (*Eng., Fr., Sp.*), **Romanza** (*It.*), **Romaunt** (*Old Eng.*) The dialect spoken in the south of France, in parts of Spain and of Italy, and elsewhere in the south of Europe during the middle ages; so called, because it was founded on the *Roman* or Latin language. In the Romance dialect the greater portion of Troubadour poetry was recited, hence the term "Romance" came to be applied to any touching love-story, and in music is now frequently given to any simple rhythmical melody which is suggestive of such a story. [Troubadour.]

Romanesca (*It.*) An Italian dance, a galliard.

Romanesque (*Fr.*) A galliard.

Roman-strings. Fiddle-strings made of the intestines of lambs, although commonly called "cat-gut." Italy still supplies the finest quality of strings, known as *Roman.*

Romantique (*Fr.*) In the style of a romance, imaginatively.

Romera. A Turkish dance.

Römischer Gesang (*Ger.*) Catholic Plain Song. Not correctly called *Roman*, because common to the whole church. England had its own old-established *uses*, before any Roman Plain Song had entered the country. [Plain Song.]

Romanzesco (*It.*) [Romantique.]

Ronde (*Fr.*) The round note, *i.e., a semibreve.* [Nomenclature.]

Rondeau (*Fr.*) [Rondo.]

Rondiletta, Rondinetta, Rondino, Rondoletto (*It.*) A short rondo.

Rondo. [Form.]

Root, called also fundamental note, generator, and ground-note. (1) A note which, besides its own sound, gives over-tones or harmonics. (2) That note from amongst whose over-tones any chord may be selected, *e.g.*:

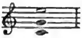

is produced from the vibration of the lowest note C, therefore C is said to be the *root* of this chord. An attempt to reduce chords to their *roots* forms the chief part of many treatises on harmony, but almost insuperable difficulties are met with in consequence of certain over-tones being omitted in our scale and other sounds being introduced which can only be obtained by a minute sub-division of the monochord. The flat seventh and the eleventh of nature are unused, and various notes are arbitrarily inserted in the modern scale in order to obtain more or less of temperament (*q.v.*) Some authors derive all their chords, or rather all those called *fundamental* (which constitute but a very small number of the chords actually in use), from *three* roots—the tonic, sub-dominant, and dominant. Others, again, insist on only *two* roots, the tonic and dominant. Not a few modern musicians use the word *root* without reference to any mathematical laws, and only as describing a note on which, when either expressed or implied, a chord is built up.

Rosalia (*It.*) The repetition of a phrase or passage, raising the pitch one note at each repetition. Ex. 1 is from Beethoven's *Symphonia Eroica*, Ex. 2 is from a Litany by Mozart. In the first the passage mounts by semitones, in the second by tones.

Ex. 1.

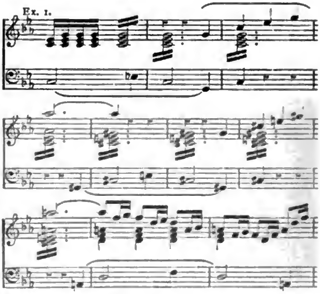

Ex. 2.

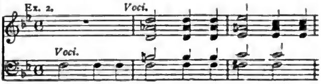

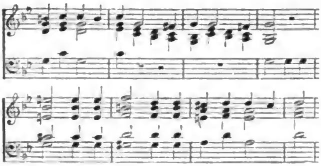

Rosin. [Resin.]

Rota (*Lat.*) A Round, but the word is sometimes applied to anything with frequent repeats, as for instance a Hymn tune.

Rote (*Old Eng.*) [Hurdy Gurdy.]

Rotondo (*It.*) Round or full, with reference to quality of tone.

Rotruenges. Roundelays of the minstrels.

Rotulæ (*Lat.*) *lit.* little rounds. A term applied to Christmas Songs or Carols.

Roulade (*Fr.*) An embellishment, a flourish, ornamental passage of runs.

Round. A composition in which several voices starting at stated distances of time from each other, sing each the same music, the combination of all the parts producing correct harmony. It differs from a Canon therefore in that it *can only* be sung at the unison or octave. It differs from a Catch, which is like it in construction, only in the character of the words. The catch should be amusing, the round may be even sacred. A round may be written out in the form of a canon, if it is of an elaborate construction, or has an independent accompaniment. When sung at the unison, a Round is said to be for *equal voices.*

TRAVERS.

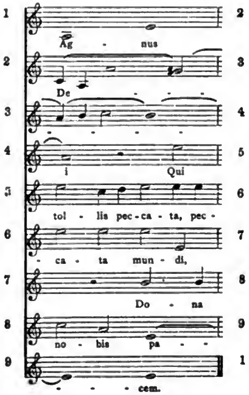

Roundel. A dance in which all joined hands in a ring. It was sometimes called a Round and a Roundelay. Minshew explains the latter word to mean "Shepheards' daunce."

> "When that Arcite had romed all his fill
> And songen all the roundel lustily,
> Into a studie he fell sodenly."
> CHAUCER, *Knight's Tale*

> "And arm in arm
> Tread we softly in a Round."
> BEAUMONT & FLETCHER, *The faithful Shepherdess.*

Roundelay. (1) A poem, certain lines of which are repeated at intervals. (2) The tune to which a Roundelay was sung.

Round O. A Rondo. [Minuet.]

Rovesciamento. [Rovescio.]

Rovescio, *al rovescio, alla riverso* (*It.*), *motus contrarius* (*Lat.*) By inversion. The contrary motion between two parts, caused when one ascends the exact diatonic intervals which the other descends, *e.g.:*

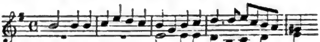

The above, the commencement of the chorus "Egypt was glad at their departing," (*Israel in Egypt,*) is said to be a subject answered *al rovescio.* Imitation *al rovescio* is when the converse of the intervals is not rigidly adhered to. Fugue subjects when treated by inversion are still often amenable to the laws which regulate the use of the plagal and authentic portions of the scale. A Canon *by inversion* is formed when the answer is in *contrary motion* to the subject *throughout the movement.* Perhaps the most ingenious specimen of this difficult construction is by Purcell, in the Gloria to his "Deus Misereatur" in B♭.

Ruana. A Hindoo instrument of the violin class.

Rubato (*It.*) *lit. stolen* or *robbed. Tempo rubato* represents the alteration made in the time, when some notes are held for more, and others for less than their strict duration.

Rückfall (*Ger.*) A back-fall, a kind of grace note, *e.g.*:

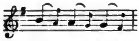

Rückpositiv (*Ger.*) A back choir organ, *i.e.*, a choir organ which is behind the player, the connecting mechanism of which passes under his feet.

Rückung (*Ger.*) [Syncopation.]

Ruhepunkt, Ruhezeichen (*Ger.*) A point of rest, *i.e.*, a pause.

Ruhig (*Ger.*) Quiet, tranquil, calm.

Rule (*Old Eng.*) A line of the stave, *e.g.*: "The F fa ut cliefe standing on the fourth *rule* from beneath." (Morley's Introduction.) "And to distinguish the *cliffs* and *notes* as they be in *rule* or *space;* for knowing the *notes'* places, their names are easily known." (Playford, Introduction to the skill of musick.)

Rule of the Octave. A name given to a system of adding harmonies to the diatonic scale, using it as the lowest part. From the nature and relation of the chords added, many laws as to progression and modulation were deduced; in fact it was formerly taught as a formula for the assistance of students, who committed to memory the harmony or harmonies which each degree was capable of bearing.

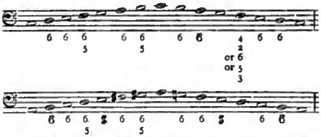

The above shows the simple form of these harmonies, to which great importance was once attached.

Rullante (*It.*) Rolling, as, *tamburo rullante*, the small military drum, the side drum.

Run. A rapid succession of notes. In vocal music usually sung to one syllable, hence called in the German "Silbendehnung." Except for the purpose of training the voice, *runs* may be said to be out of fashion in this half-century. In writing the words under a run in vocal music, it is usual now to place under the first note as much of the word as will show what the whole consists of. This was not formerly the custom, and in one of the songs in Blow's "Amphion Anglicus," the words under a lengthy run are thus divided, "*th - - - - - - - - - - - - us fl - - - - - - - - - - - y*," an absurd division, as the combination of letters to which the run is supposed to be sung closes the mouth effectually. The word *run* being English, is avoided in fashionable critiques and *roulade* is generally adopted.

Running. The improper sounding of an organ pipe or pipes from a defect in the sound-board, or other causes. A running is not heard until keys are held down. The most common way of testing the workmanship of a sound-board is to put in *all* the registers, and hold down full chords. If there is any fault in its construction, a *running* will then be immediately heard.

Ruollo (*It.*) *Lit.* a roll; according to some an Italian dance, a waltz, from which the expression used by children "rowly powly" is derived. (*Fr. rouler*, to roll, and *poulie*, anything which rolls round, a *pulley*.)

Russian horn-band. A band of players whose instruments are so designed that each produces a single note only. For the performance of a simple theme at least twenty horns are required, but the complete band numbers thirty-seven, comprising a compass of three octaves, the tones and semitones, in their relative degrees, being gained by the use of instruments of various lengths, the longest being more than twelve feet, and the smallest nine inches. The effect of the music played is said by those who have heard it to be extremely fine and of pure tone. As each player can give no more than his one note, his attention is fully engaged during the performance. The effort of mind needful being purely mechanical, similar to that required for hand-bell ringing, the players are not necessarily men of a high intellectual order. The invention of this horn music is assigned to J. A. Maresch, a Bohemian in the service of Prince Narischkin in 1751, who being permitted for the purposes of his experiments to employ serfs, did not scruple to use severity in order to make them perform with accuracy, rapidity, and distinctness. Having drilled his forces for four years, Maresch conducted the first performance in the presence of the Imperial Court in 1755, and the tradition he then established has been, or was until recently, strictly maintained. A Russian horn-band visited England in the year 1834.

Rusticano, Rustico (*It.*) In a simple, rustic manner.

Rutscher (*Ger.*) *Lit.* the slider, the galopade.

Rymour (*Old Eng.*) A bard or minstrel.

S.

S. Abb. of *sinistra*, left; *subito*, suddenly; *segno*, sign; *solo*, &c.; as M.S., *manu sinistra*, with the left hand; V.S., *volti subito*, turn over quickly; D.S., *dal segno*, from the sign; V.S., *voce solo*, voice alone, &c.; also Abb. for (1) *scriptus*, written, as MS., manuscript, hand-written; (2) *senza*, as *S. Ped.*, without the pedal or pedals, *S. Sordini*, without the dampers, or without mutes.

𝄋 A sign used to point the extent of a repeat, as *al segno* (𝄋) go back *to the sign* 𝄋, *dal segno* (𝄋), repeat *from the sign* 𝄋. The word *fine* (*It.*) is generally placed over the last chord of a movement repeated by the above directions.

Sabeca. One of the musical instruments mentioned in Dan. iii. 5, 7, 10, 15. It is generally supposed to have been identical with *sambuka*, a large species of harp, perhaps the large Egyptian harp. In the authorised version it is unfortunately rendered *sackbut*, an utterly unwarranted translation.

Saccade (*Fr.*) Strong pressure of a violin bow against the strings, which by forcing them to a level enables the player to produce three or four notes simultaneously.

Sackbut. (1) One of the Babylonian musical instruments mentioned by Daniel in chap. iii. v. 5, 7, 10, 15. It is the translation in the English version of the Bible of the word *sabeca*. Some authors identify it with the sambuka (σαμβύκη) of the Greeks and Romans, a kind of harp. [Sambuka.] (2) The old English sackbut or sagbut was a bass trumpet, with a slide like the trombone. "As he that plaies upon a sagbut by pulling it up and down alters his tones and tunes."—*Burton's "Anatomy of Melancholy."*

Sackpfeife (*Ger.*) [Bagpipe.]

Sacring bell. [Sanctus bell.]

Sailours. A word used in Chaucer which has given some trouble to commentators,

> "There was many a timbestere,
> And *sailours*, that I dare well swere
> Yeothe ther craft full parfitly."
> *Romaunt of the Rose.*

It probably signifies *dancers*, and is derived from the *Fr. saillir, Lat. salio*.

Saite (*Ger.*) [String.]

Saitenhalter (*Ger.*) [Tail-piece.]

Salamanie. An oriental flute.

Salicional, Salicet, Salcional, Solcionell. Terms derived from the Latin *salix* (a willow). An organ stop of soft and delicate quality, supposed to be similar in character with the *salicis fistula*, or *withy-pipe*. It is generally placed in the choir organ, but sometimes in the swell, in either case taking the place of the dulciana, to which it bears a strong resemblance.

Salii. Priests of Mars Gradivus, twelve in number, who had the care of the twelve Ancilia, and who, during the feast of that God, were accustomed to go through the city carrying the Ancilia, singing and *dancing*.

Saltando (*It.*) Jumping. In violin music, "with a dancing bow."

Saltarello (*It.*) (1) A dance in which leaping steps are introduced, similar to the Siciliano and Forlana of Italy, and the Jig of England. It is triple in time, with a triplet always at the commencement of each phrase. Saltarelli are frequently found as movements or separate pieces in harpsichord and pianoforte music. (2) A harpsichord jack, so-called because it jumps when the note is struck. (3) Counterpoint is said by old authors to be "in Saltarello," when six quavers of the accompaniment are given to each minim of the Canto fermo, *e.g.*:

Gerlach.

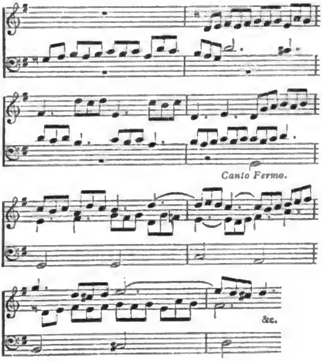

Salterio } (*It.*) The Psalter, or book of
Saltero } Psalms.

Salto (*It.*) (1) A dance in which there is much leaping and skipping. (2) A leap, or skip from one note to another beyond the octave.

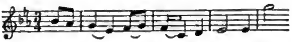

(3) Counterpoint is said to be *di salto* when the part added moves in skips.

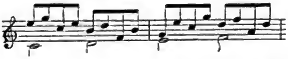

Sambucistria (*Lat.*) A player on the sambuka.

Sambuka (*Gk.*) *σαμβύκη*. This word, though applied sometimes to several musical instruments of different kinds, such as a lyre, a dulcimer, a triangular harp or trigon, and a large Asiatic harp, seems to have been chiefly used as a term for the last-named instrument. By some authors it has been identified with the large Egyptian harp, illustrations of which are so familiar to all of us as to render one unnecessary here. It is generally thought that the *sabeca* mentioned in the book of Daniel, iii. 5, 7, 10, 15, and which is improperly rendered *sackbut* in our version, was a *sambuka* or large harp. It has been suggested that "sambuka" was used as a general term for instruments made of elder-wood (*sambucus*); but the best authorities will not allow any relationship between the two words.

Sampogna, Zampogna. [Bagpipe.]

Sancho. A negro instrument of the guitar species, made of hollowed wood and furnished with a long neck. It is strung with the tough fibres of a creeping plant. It is tuned by means of sliding rings.

Sanctus (*Lat.*) *Ter Sanctus*, or, *Trisagion* (*Gk.*) *Holy*, *Holy*, *Holy*, *&c.* A part of the Communion Service in the Church of England, and a part of the Mass in the Church of Rome. In many cathedrals where it is not usual to celebrate the Holy Eucharist chorally, the Sanctus is used as an Introit, a custom which cannot be too strongly condemned. [Cathedral Music.]

Sanctus bell, Saints' bell. A small bell which is rung in order to mark the progress of the office of the mass. In some churches bells of this kind are placed outside the church, so that those unable to be present inside may be reminded of important parts of the service.

Sanft (*Ger.*) Soft. *Mit sanften Stimmen*, with soft stops.

Sans (*Fr.*) Without; as, *sans pédales*, without the pedals.

Santoral (*Sp.*) Church choir book.

Saquebute (*Fr.*) [Sackbut.]

Saraband (*Old Eng.*), **Sarabanda** (*It.*), **Zarabanda** (*Sp.*), **Sarabande** (*Fr.*) A Spanish dance of Moorish origin, for a single performer, who accompanies himself with the castanets. The tune is in $\frac{3}{4}$ time, but slow and stately, and with a strong accent on the second beat in the bar. This peculiarity would identify it with the Moresca. In olden times it was accompanied with singing the *coplas* or *canciones*, poetry of a trivial sort. Jenkins, Purcell, Blow, Bach, Handel, and Scarlatti, in their *suites* for the harpsichord or clavichord, frequently employed the saraband as a movement.

There is a Spanish proverb with reference to weak verses, that "they are not worth as much as the couplets of the saraband." "No vale las coplas de la sarabanda."

Saraband. (From E. Roger's Virginal book.)

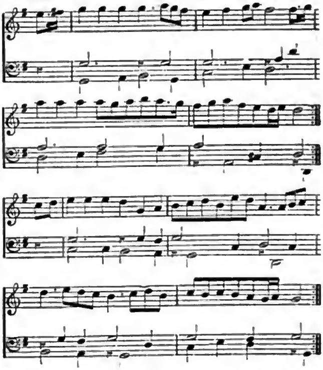

Saraband. H. Purcell.

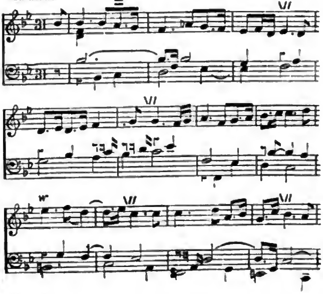

Digitized by Google

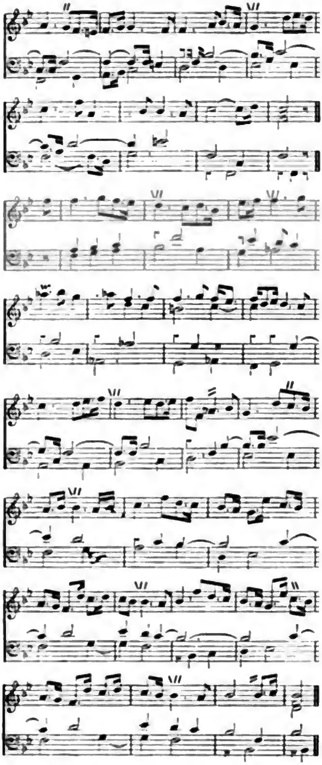

Saroh. Indian instruments played with a bow.

Sattel (*Ger.*) [Nut.]

Satz (*Ger.*) A theme, a subject, a composition, a piece.

Sautereau (*Fr.*) The jack of a spinet.

Sauterie (*Old Eng.*) A psaltery; dulcimer.

Sautillé (*Fr.*) Springing bow; in violin music indicated by dots over the notes.

Sax-horns. Cylinder horns invented by Antoine (usually called Adolph) Sax. They have 3, 4 or 5 cylinders, so that each horn is capable of playing all the notes of its scale without difficulty. A sax-horn band comprises seven instruments, a small high horn, a soprano, an alto, a tenor, a baritone, a bass, and a double-bass. [Metal wind instruments.]

Saxophones. Brass wind instruments, the invention of M. Sax. They are played with a single reed and a clarinet mouth-piece. The quality of tone is soft, yet telling and expressive. They are six in number, the high, soprano, alto, tenor, baritone and bass.

Saxotrombas, Saxtrombas. Brass cylindrical wind instruments with wide mouthpieces, of a shrill and piercing tone, a combination of the trumpet and bugle quality. The complete set is six, divided as the saxophones.

Saxtuba. Deep-toned brass bass instruments of similar character to the saxtrombas.

Saynetes (*Sp.*) Interludes or *entremeses*, introduced between the prologue and the principal comedy in the Spanish drama, in which music and dancing form prominent features. Saynetes are generally of a burlesque or humorous character.

Sbalzo (*It.*) A leap, a skip.

Sbarra (*It.*) A bar. *Sbarra doppia.* double bar.

Scala (*It.*) A scale or gamut.

Scald. [Skald.]

Scale. The graduated series of sounds into which the octave is divided.

To give a history of the scale would be to give a history of music itself; it must suffice, therefore, to say a few words on the growth of the scale to its present shape. Nothing is known with certainty of the nature of the scales of any of the most ancient nations. If it be admitted that the Greeks obtained their notions of music from the Egyptians, it may be hazarded, merely as a supposition, that the Egyptian scale was tetrachordal, that is, consisting of groups of four notes.

In the article Greek Music the reader will find a full explanation of the manner in which the *octave system* became practically a part of the ancient tetrachordal system, which it was destined afterwards to supersede entirely. Although our modern scale was unquestionably a development of the diatonic scale of the Greeks, yet, for several centuries, a hexachordal system was in use, a full account of which will be found under the head "Notation." The Church modes were probably the connecting link between the ancient Greek music and the modern diatonic scale. [Plain Song.] The division of the octave into twelve parts, called semitones, each of which can be used as a key-note, became only feasible when keyed instruments were tuned on the system known as *equal temperament*. [Temperament.] This gives to the chromatic notes of our scale a far greater value than the chromatic or enharmonic notes of the an-

Digitized by Google

cients, as it is probable they were never used but as passing or auxiliary notes. The whole system of music hangs upon the relationship of the sounds used to a *tonic*, which, in modern music, is always the first note of whatever octave system (key) is chosen, but in Greek music and early Church-song was a note at or near the *middle* of the scale.

The old Church mode corresponding to the modern scale was the Ionic or Iastian, but when this was finally adopted as the normal scale, a still older form was retained for use with it, founded on the Dorian and Hypo-Dorian modes, to which, now slightly modified, we give the name *minor mode*, and by starting from any one note in the semitonal scale, we can have twelve minor modes. As a minor mode largely consists of the notes of the major scale beginning on its third degree, it is said to be *relative* to that scale. The form of the minor mode has varied from time to time, and even now cannot be said to be definitely settled.

Ex. 1.

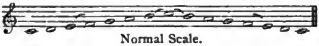

Normal Scale.

Ex. 2.

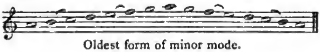

Oldest form of minor mode.

Ex. 3.

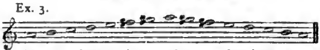

Form of minor mode sometimes used melodically.

Ex. 4.

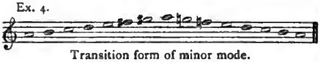

Transition form of minor mode.

Ex. 5.

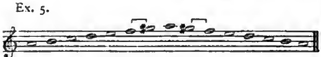

Modern minor mode.

The scale shewn in Ex. 2 is by no means obsolete. The revival of old Church-song has caused many composers to study its forms, and their works are, perhaps insensibly, tinctured with its influence. The somewhat odd *descent* of the scale in Ex. 3 is to be met with in Handel, Bach, and other writers. When this *descending* scale is harmonized, it is generally founded on the following bass :—

A	E	B	E	G♯	A	E	A
5	5	♯5	5	6	5	5	5
3	♯	♯	♯	5	3	♯	3

It will thus be seen that the descending F♯ of the scale is made part of a chord of B major, a sort of new dominant to E, the dominant of A minor.

The form of Ex. 4 is familiar to all, and has not yet been entirely superseded by that of Ex. 5, which always presents both harmonic and melodic difficulty in its augmented second from F♮ to G♯. There seems to be no reason why one particular form of the minor scale should be used to the exclusion of another; all are at times useful and beautiful. Perhaps the great importance of the form given in Ex. 5 arises from the fact that it is the veritable source of the chord of the diminished seventh—

F
D
B
G♯

The musical scales of extra-European countries are so varied in character that it is impossible to draw any reliable conclusions from their form. The Arabs, Indians, and many uncultured tribes in all quarters of the globe have more than twelve divisions in the octave, that is, use *enharmonic* scales. The Chinese have the old five-note scale, called by Engel, Pentatonic.

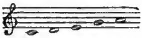

This scale is associated also with Scotch and other Celtic melodies.

In some nations the natural harmonic, known as the sharp eleventh, which we discard, is in use, probably because it is produced upon their simple tube instruments.

The degrees of the ascending scale are distinguished in harmony by the following names.

First	Tonic.
Second... ...	Supertonic.
Third	Mediant.
Fourth	Subdominant.
Fifth	Dominant.
Sixth	Superdominant.
Seventh ...	Subtonic or leading note.

Scagnello (*It.*) The bridge of a stringed instrument.

Scemando (*It.*) An equivalent to *diminuendo*, gradually decreasing in power.

Scena (*It.*) (1) A scene. (2) A solo for a single voice, in which various dramatic emotions are displayed.

Scenario (*It.*) The plot or main incidents of an opera or drama.

Scenici (*Lat.*) Games instituted, according to Livy, B.C. 364. They were the germ of the Roman play.

Schäferlied, Schäferspiel (*Ger.*) A pastoral.

Schallbecken (*Ger.*) Cymbals. *Lit. Sound-cups.* [Cymbals.]

Digitized by Google

Schallhorn } (*Ger.*) Bell of a horn, cornet,
Schallstück } trumpet, &c.

Schalmey. [Chalumeau.]

Scharf (*Ger.*) *Lit.* Sharp. A mixture stop of an organ, formed of a combination of acute harmonics.

Scherz (*Ger.*) Drollery, fun.

Scherzando, Scherzandissimo, Scherzante, Scherzevole Scherzevolmente, Scherzoso (*It.*) (1) Playful, lively, jokingly, merry. (2) A movement of a lively and droll character.

Scherzhaft (*Ger.*) Droll, funny.

Scherzo (*It.*) A term, meaning literally a *jest*, applied to a movement in a Sonata or Symphony of a sportive, playful character. Beethoven introduced it in the place of the minuet, and sometimes in addition to it, and it has been supposed that the credit of its invention is due to him. But Haydn had previously done a similar thing in his quartetts, and J. S. Bach had also called one of the pieces in a suite, by the name. Other writers had directed certain movements to be played "Scherzando," but only in connection with some other word descriptive of the general character of the movement. Thus Schobert, 1724-1768, marked the second movement of his Sixth Sonata, "Badinage Scherzando." In 1734 George Philip Telemann published at Hamburg "Scherzi melodichi, per divertimento di coloro che prendono l'acque minerali in Pirmonte, con ariette semplici e facili, a violono, viola e fondamento," and nearly 130 years before, Monteverde issued at Venice "Scherzi musicale a tre voci." These were similar in style to the "Cantici" or humorous songs in parts, the predecessors of the Catch. Though they cannot fairly be considered as belonging to the class of composition included under the general term *Scherzo*, yet they serve to show that Monteverde believed music to have a merry as well as a serious side. The Scherzo of Bach, alluded to above, is here added:

Scherzo. J. S. BACH.

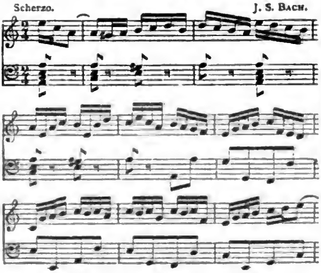

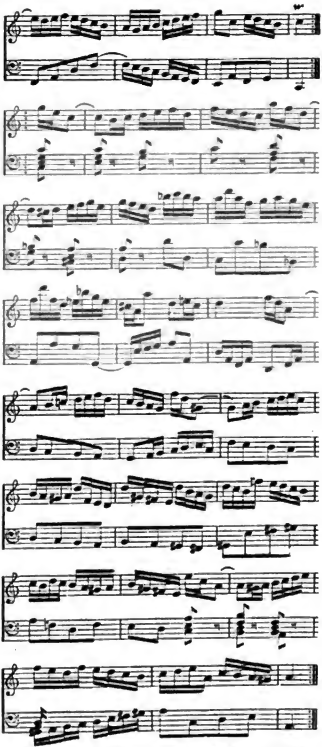

Schiettamente (*It.*) Pure, simple, neat.

Schietto (*It.*) Pure, neat.

Schisma (*Gk.*) σχίσμα. An approximate half of a Pythagorean comma, that is, half of the difference between twelve fifths and seven octaves.

Schlag (*Ger.*) A stroke or beat, (1) of time, (2) of a vibrating reed, (3) of an instrument of percussion.

Schlagfeder (*Ger.*) [Plectrum.]

Schleifbogen, } (*Ger.*) *Lit. Slide-bow.*
Schleifezeichen } A slur. The sign ⌒.

Schleifen (*Ger.*) To slide, to glide.

Schleifer (*Ger.*) (1) Slurred note. (2) A grace. [Harpsichord graces.]

Digitized by Google

Schluss (*Ger.*) The conclusion, finale. *Schlusschor*, final chorus.

Schlüssel (*Ger.*) A clef.

Schlussfall (*Ger.*) Cadence.

Schlussreim (*Ger.*) The burden or refrain of a song.

Schlussstück (*Ger.*) Finale.

Schmelzend (*Ger.*) *Lit.* Melting away. Dying away, diminishing.

Schmerz (*Ger.*) Grief. *Schmerzhaft*, sorrowful, dolorous.

Schnabel (*Ger.*) The mouth-piece of the clarinet, flûte-à-bec, and instruments blown in like manner.

Schnarrpfeifen, Schnarrwerk (*Ger.*) Reed pipes or stops in an organ.

Schnell (*Ger.*) Quick. *Nach und nach schneller*, quicker and quicker. *Schnell und beweglich*, quick and with emotion.

Schneller *or* **Schnelzer** (*Ger.*) A trill.

School. (1) A method or system of teaching; as, Rink's *organ school*, Crivelli's *vocal school*, Spohr's *violin school*, &c. (2) Style; as, the *madrigal school*, *ecclesiastical school*, *operatic school*, &c. (3) The manner or characteristics of a composer, or performer; as, Handel's *school*, Rossini's *school*. (4) A group of composers whose works mark an epoch in the history of music. Such a school is generally named after a place where such composers resided; as, the *Venetian school*, the *Neapolitan school*, &c.

Schophar *or* **Shophar.** A Hebrew trumpet.

Schottische (*Ger.*), *Ecossaise* (*Fr.*) *Lit.* The Scotch dance. A slow dance of modern introduction, written in $\frac{2}{4}$ time.

Schreibart (*Ger.*) Style.

Schrittmässig (*Ger.*) Slowly. *Andante.*

Schultergeige (*Ger.*) The shoulder-fiddle, or common violin, as opposed to the *Kniegeige*, or *viola da gamba*, the viol played between the legs.

Schusterfleck (*Ger.*) A facetious name for a rosalia, *q.v.*

Schwach (*Ger.*) Weak, feeble, soft; *piano.*

Schwärmer (*Ger.*) *Bombo.* An old name given to four or more notes repeated rapidly on the same degree of the scale.

Schweigezeichen (*Ger.*) A rest.

Schweinskopf (*Ger.*) *Lit.* Pig's head. A name given to grand pianofortes because of their outline, as viewed from the side.

Schweizerpfeife, Schweizerflöte, Feldpfeife. The name given in Germany to the old *flauto traverso* or *Querpfeife*, which, when introduced into this country, was known as the *German flute.*

Schweller (*Ger.*) The swell organ

Schwermüthig (*Ger.*) Sad, pensive.

Schwingungen (*Ger.*) Vibrations. [Acoustics.]

Scioltamente, Scioltezza, con, Sciolto (*It.*) Freely, with agility, easily. *Fuga Sciolta*, a free fugue.

Scolia. Short songs of the Greeks. They were distinct from the religious hymns, as they were frequently sung during feasts or banquets. Though at first they were like the Pæans sung by the whole company assembled, they were afterwards performed by each guest alone, holding in his hand a myrtle branch, which, at the conclusion, he passed to his neighbour, and so on in turn. Sometimes they were sung to an accompaniment.

Scolia were of three kinds: (1) Moral songs, like some of those preserved by Athenæus. (2) Mythological hymns and historical songs. (3) Miscellaneous songs, chiefly on love and wine.

Scordato (*It.*) Out of tune.

Scordatura (*It.*) The mis-tuning of an instrument. When a violinist alters the *accordatura* of his instrument for a special purpose, he is sometimes said to make a *scordatura*. [Accordatura.]

Score. *Partitur* (*Ger.*), *Partition* (*Fr.*), *Partizione*, *Partitura*, *Sparto*, *a partito* (*It.*) A copy of a musical work in which all the component parts are shewn, either fully, or in a compressed form. (1) A *short* or compressed score is when all the parts are arranged or transcribed so that they shall appear in two staves. Such scores are commonly used for hymn tunes, and also (especially in Germany) for part-songs. When counterpoint is introduced into vocal music, or when the parts frequently cross, short-scores become too complicated for general use. But in simple vocal music, more or less note against note, they effect a great saving of space. In transcribing four-part music into short score, the two upper parts are arranged in the treble stave; the two lowest in the bass, the tails to the notes of the first and third parts being invariably turned *up*, those of the second and fourth parts *down.* (2) A pianoforte or organ score is one in which the voice-parts are written out in full on separate lines, and the instrumental accompaniment is arranged in two lines, treble and bass, for performance on a pianoforte or organ. (3) A vocal score is (or was formerly understood to be) one in which the voice-parts are written out in full, and the accompaniment (if any) is indicated by a figured bass. (4) A full score is one in which each part is written on a separate line one over the other, subject, however, to the modification that the parts to be played by two wind instruments of the same name and compass may be included on one line. For example, the parts of the two oboes, two flutes, two clarinets, &c., and, sometimes, of the alto and tenor trombones are written on one line;

Digitized by Google

the tails of the notes to be played by the first instrument being uniformly turned upwards; those to be played by the second, downwards. The first and second violin parts are never written on the same line unless they are to play in unison; but the violoncello and double-bass parts are usually written on one line. The parts of similar instruments are not *necessarily* written on the same line, unless it is desirable to economise space. The order in which the instrumental parts are arranged in a score has from time to time varied considerably, the only point of uniformity being that voice parts are nearly always placed immediately above the line or lines set aside for the violoncello and double-bass part.

The following are a few examples of the manner of grouping instruments and voices:

Ex. 1.
Trombe
Timpani
Corni
Flauti
Oboi
Fagotti
Violino 1mo.
Violino 2do.
Viola
1
2 Voci
3
4
Violoncello
e Basso
Organo

Ex. 2.
Violino 1mo.
Violino 2do.
Viola
Fagotti
Oboi
1
2 Voci
3
4
Organo e
tutti Bassi

Ex. 3.
Violino 1mo.
Violino 2do.
(in unison and on one line.)
Oboe
Viola
Basso

Ex. 4.
Timpani
Trombe (or) Clarini
Corni
Clarinetti
Flauti
Oboi
Fagotti
Violino 1mo.
Violino 2do.
Viola
Violoncelli (e)
Bassi

Ex. 5.
Pauken
Trompeten
Hörner
Flöten
Hoboen
Clarinetten
Fagotte
Alt und
Posaunen Tenor
Bass
Violine, 1ste
Violine, 2te
Bratsche
Violoncell
Contrabass

Ex. 6.
Flutes
Oboes
Clarinets
Bassoons
Horns
Trumpets
Trombones
Drums
Violins, 1
Violins, 2
Violas
1
2 Voices
3
4
Violoncellos and
Basses
Organ

Exs. 1, 2, and 3 illustrate a disposition often followed by Handel; Ex. 4 the arrangement often found in the works of Mozart, Haydn, and others. Ex. 5 shows the arrangement sometimes followed by Schumann, which is peculiar as separating the trombones (Posaunen) from the trumpets. The above examples (1 to 6) show only a few of the many arrangements sometimes found. The disposition of the parts now most generally followed is shown in Ex. 6.

Where more than the usual number of instruments are employed, this last disposition may be indefinitely extended, if only the instruments are grouped into classes, as follows:

1. Wood wind instruments.
2. Soft-toned brass instruments.
3. Loud brass instruments.
4. Instruments of percussion.
5. String instruments (excepting the bass part).
6. Voices.
7. Bass string instruments.
8. Organ or pianoforte.

Modern scores owe much of their apparent intricacy to the individual existence which is now given to almost every separate instrument in an orchestra, for the purpose of obtaining constant change of "colour," as it has been called. In the scores prior to the commencement of this century, it is no uncommon thing for the first oboe to be directed to play throughout a whole movement with the first violins, the second oboe with the second violins, the bassoons with the violoncellos. It is still more remarkable that the *viola* should often have been made to play the *bass* part in unison, or (when below its compass) at the octave; for the power and beauty of the "string-quartet" was thus wantonly destroyed.

Printers of music use the word score to denote any lines (two or more) united by one *brace*.

Scorrendo (*It.*) Gliding from one sound to another. [Glissando.]

Scorrevole (*It.*) Running, flowing, gliding.

Scotch Scale. [Pentatonic Scale.]

Scotch Snap. A peculiarity of the comparatively modern Scotch melodies in which a short note precedes a long one:

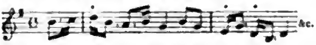

It is the characteristic of Strathspey tunes. Reels and jigs are distinguished from the Strathspey by the absence of the snap. The best informed among the Scotch writers on music declare it to be the mark of imitations of Scotch melodies, such as, "Within a Mile of Edinboro' Town" and the *Scherzo* in

Digitized by Google

Mendelssohn's "Scotch" Symphony, as none of the ancient tunes contain any examples of its use. It has been conjectured that its introduction into Scotland is due to the Hungarian gipsies, as the snap is characteristic of their tunes.

Scozzese (*It.*) Scotch, as, *alla scozzese*, in the Scotch style.

Scriva (*It.*) Written. *Si scriva*, as written.

Sdegno, con; Sdegnosamente, Sdegnoso (*It.*) Scornfully, disdainfully.

Sdrucciolando (*It.*) Sliding.

Sdrucciolare (*It.*) (1) To slide, by pressing down the keys of a pianoforte in rapid succession and lightly, with the finger nails. Scales upon the harpsichord were frequently played thus. (2) *Glissando.*

Se (*It.*) As, if; *se bisogno*, if required.

Sec (*Fr.*), **Secco** (*It.*) Dry, unadorned, plain, as, *recitativo secco*, plain recitative, that is, without band accompaniments. [A table sec.]

Seccarara (*It.*) A Neapolitan dance.

Sechsachteltakt (*Ger.*) Six-eight time, $\frac{6}{8}$. *Sechsvierteltakt*, $\frac{6}{4}$ time.

Sechssaitig (*Ger.*) Six stringed.

Sechstheilig (*Ger.*) In six parts.

Sechszehnfüssig (*Ger.*) Of sixteen feet. The unison pitch of the pedal organ.

Sechszehntheilnote (*Ger.*) A semiquaver. The sixteenth part of a semibreve.

Second, Chord of the. An old abbreviation of the chord $\begin{smallmatrix}6\\4\\2\end{smallmatrix}$ [Figured Bass.]

Second, Interval of a. [Interval.]

Second (*Fr.*) Second; as, *second dessus*, the second treble; *seconde fois*, the second time.

Secondo (*It.*) Second; as, *violino secondo*, second violin; *seconda parte*, the second part; *seconda volta*, the second time.

Secund (*Ger.*) The interval of a second. *Die kleine Secund*, the minor second; *übermässige Secund*, an augmented second.

Secundakkord (*Ger.*) [Second, chord of the.]

Sedecima (*Lat.*) A sixteenth. A name formerly given improperly to the *fifteenth stop* of an organ.

Seg. Abb. of *segue*; also, of *segno*.

Segno (*It.*) The sign 𝄋 [Al segno.]

Segue (*It.*) Follows, succeeds, comes after; as, *segue il coro*, the chorus follows; *segue l'aria*, the aria follows.

Seguendo, Seguente (*It.*) Following, as, *attacca subito il seguente*, begin the following at once.

Seguenza (*It.*) A sequence.

Seguidilla (*Sp.*) A lively Spanish dance, similar to the country dance; the tune is in $\frac{3}{4}$ or $\frac{3}{8}$ time.

SEGUIDILLA.

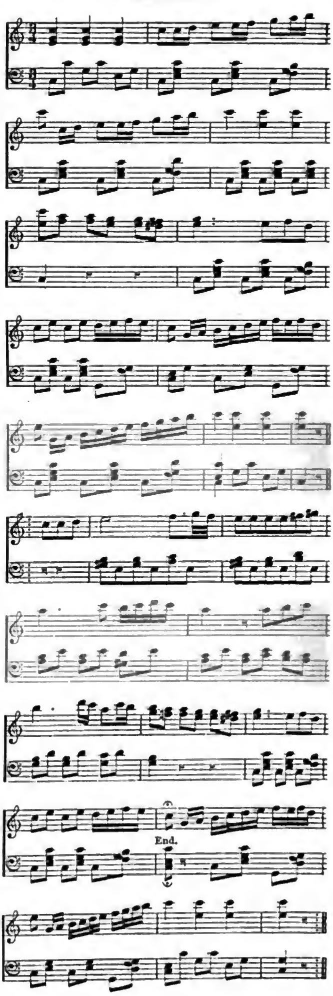

Digitized by Google

Seguite (*It.*) [Segue.]

Sehnsucht (*Ger.*) Ardour, desire, fervour, longing.

Sehr (*Ger.*) Very, much, extremely; as, *sehr lebhaft*, very lively; *sehr langsam*, very slow.

Sei (*It.*) Six; as, *sei stromenti*, six instruments.

Seitenbewegung (*Ger.*) Oblique motion. [Motion.]

Seizième de soupir (*Fr.*) A semiquaver rest.

Semeia (*Gk.*) *σημεῖα*. The musical characters.

Semibreve. [Nota.]

Semichorus. A direction that the passage so marked is to be sung by half the chorus, or a selected portion of it.

Semicroma. A semiquaver 𝅘𝅥𝅯

Semidemisemiquaver. A half demisemiquaver, the 64th part of a semibreve.

Semi-diapason (*Lat.*) An imperfect octave. *Octava deficiens.*

Semi-diapente (*Lat.*) An imperfect or diminished fifth. *Quinta deficiens.*

Semi-diatessaron (*Lat.*) An imperfect or diminished fourth. *Quarta deficiens.*

Semi-ditonus (*Lat.*) A minor third. *Semi-ditonus cum diapente.* A minor seventh.

Semifusa (*Lat.*) A semiquaver.

Semiminima (*Lat.*) A crotchet.

Semiquaver. The 16th part of a semibreve.

Semiquaver rest. The sign 𝄿 representing silence for the length of a semiquaver.

Semisuspirium (*Lat.*) A quaver rest.

Semitone. A half a tone, or an approximate half of a tone. [Temperament.]

Semitonium (*Lat.*) A semitone. *Semitonium modi*, the leading note, sensible, master note, or major seventh, called also *subsemitonium modi.*

Semituono (*It.*) A semitone.

Semplice (*It.*) Pure, plain, simple, unadorned.

Semplicemente (*It.*) Purely, plainly, simply, without ornament.

Semplicità, con (*It.*) With simplicity, unaffectedly.

Sempre (*It.*) Always, ever, continually; as, *sempre forte*, loud throughout; *sempre legato*, smooth throughout; *sempre piano*, always soft; *sempre più affrettando il tempo*, continually hastening the time; *sempre più forte*, continually increasing in loudness; *sempre ritardando*, continually slackening the time; *sempre staccato*, staccato throughout.

Sennet, Synnet, Signet, Synet. (1) A word chiefly occurring in the stage directions of the old plays, indicating the sounding of a note seven times:

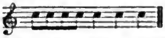

(2) A *flourish* consisting of a phrase made of the open notes of a trumpet or other tube-instrument.

Sensibile (*It.*) Expressive.

Sensibilità (*It.*) Sensibility, feeling.

Sensible (*Fr.*) The leading note of a scale, the major seventh, *note sensible.*

Sentie (*Fr.*) Expressed; *mélodie bien sentie*, the melody well expressed or marked.

Sentimental. A term applied to songs having words that are not positively humorous, comic, or bacchanalian, whether they express a special sentiment or not.

Senza (*It.*) Without; as, *senza accompagnamento*, without accompaniment; *senza bassi*, without the basses; *senza fiori*, without ornaments, plainly, simply; *senza interruzione*, without interruption, go on without stopping; *senza oboe*, without the oboe; *senza organo*, without organ; *senza ornamenti*, without embellishments or ornaments; *senza pedale*, without the pedal; *senza repetizione*, *senza replica*, without repetition; *senza rigore*, not in strict time; *senza sordini*, without the dampers in pianoforte playing; *senza sordino*, without the mute of a violin; *senza stromenti*, without instruments; *senza tempo*, without time, in no definite or exactly marked time.

Separation. An old name for a grace or passing note "not reckoned in the measure or time, put between two real notes rising a third, and only designed to give a variety to the melody."

Se piace (*It.*) At will, as it pleases the performer.

Septet, Septetto (*It.*), **Septuor** (*Fr.*) (1) A composition for seven voices or instruments. (2) A piece in seven parts.

Septième (*Fr.*), *Septime* (*Ger.*) The interval of a seventh.

Septimenakkord (*Ger.*) The chord of the seventh.

Septimole. A group of seven notes to be played in the time of four or six.

Sequence. (1) A Prose. [Prosa.] (2) The recurrence of a harmonic progression or melodic figure at a different pitch or in a different key to that in which it was first given. Ex. 1 shows the repetition (five times) of the progression from a minor common chord to the major common chord of the third below:

Beethoven.

Ex. 1.

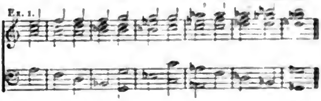

In Ex. 2 will be found the recurrence of a short phrase at the interval of one note higher:

Ex. 2.

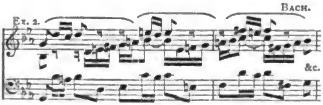

In modern music a striking effect is often obtained by the repetition of a phrase or theme at the interval of one semitone, (Ex. 3):

Ex. 3.

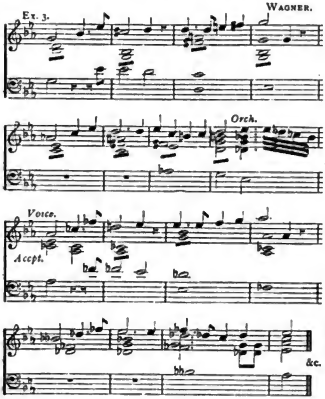

Sometimes a striking melodic figure and a harmonic progression are combined in the formation of a sequence:

Ex 4.

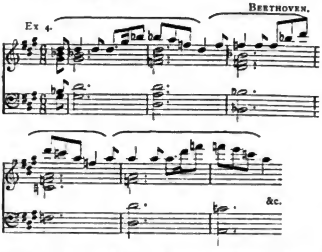

Some authors divide sequences into *tonal* and *real;* others make a like division but term them *diatonic* and *chromatic.* A tonal or diatonic sequence is when no modulation takes place (see also Exs. 1 to 4 of Suspensions). A chromatic or real sequence takes place when the recurrence of a phrase at an exact interval causes a change of key, as Exs. 1, 2, 3 above.

Seraphine. An instrument introduced in the early part of this century, the sounds of which were produced by free-reeds, but being very coarse and unpleasant in tone, it rapidly disappeared on the introduction of the harmonium.

Serenade. Originally a vocal or instrumental composition for use in the open air at night, generally of a quiet, soothing character. The term in its Italian form, *serenata,* came to be applied afterwards to a cantata having a pastoral subject, and in our own days has been applied to a work of large proportions in the form, to some extent, of a symphony. Serenades were sometimes called Ständchen (*Ger.*)

Serena (*It.*) An evening song; an *Abendlied.*

Sereno (*It.*) Calm, serene, tranquil.

Seria (*It.*) Serious, grave, tragic, as, *Opera seria,* a tragic opera.

Serinette (*Fr.*) A bird-organ.

Serioso (*It.*) In a serious, thoughtful manner.

Serpent. Serpente (*It.*) A bass instrument of a powerful character. It is of wood, twisted into a curved form, and covered with leather, with a mouth-piece like a horn or trombone, and keys for producing the notes. It was invented by a French priest at Auxerre in 1590, and is frequently used in the orchestra to strengthen the bass part; but it requires to be very skilfully blown, as it is capable of producing a quarter tone above or below the note intended, and there are three notes in its compass

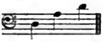

of greater power than the rest of its scale, which ranges from

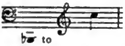

with every intermediate tone and semitone. The serpent is a transposing instrument, being in B flat, and the part it is to take is sometimes written a note higher than its real sound. [Ophicleide.]

Serpentono (*It.*) [Serpent.]

Serrata (*It.*) A concluding performance.

Service. As a term used in church music, this word signifies a musical setting of those portions of the offices which are sung by the choir, such as the Canticles, Sanctus, Gloria in excelsis, &c. A "Burial Service" is a setting of those portions of the Office for the

Burial of the Dead which may be sung by a choir.

Servi Symphoniaci, *or* **Pueri Symphoniaci.** The band of musicians kept by persons of rank among the Romans.

Sesquialtera. (1) Numbers in the proportion 3 : 2. [Proportio.] (2) An organ stop consisting of several ranks of pipes, sounding high harmonics for the purpose of strengthening the ground tone. [Organ.]

Sesta, Sesto (*It.*) A sixth.

Sestet, Sestetto (*It.*) A composition for six voices or instruments.

Settima, Settimo (*It.*) The interval of a seventh.

Setzkunst (*Ger.*) The art of musical composition.

Seventh. [Interval.]

Seventh, Chords of the. Chords are named on two principles: (1) by the largest interval contained in their component notes; (2) by the largest interval they contain when referred to their supposed fundamental-bass or root.

On the former of these principles sevenths have been divided into *primary* and *secondary;* primary being those which have the tonic or dominant for their bass note; secondary, those which have the second, third, fourth, sixth, and seventh of the scale for their bass, *e.g.*:

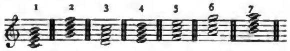

Primary chords of the seventh. Secondary chords of the seventh.

(1) The inversions of No. 1 are

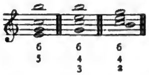

and its most common resolutions, as follows:

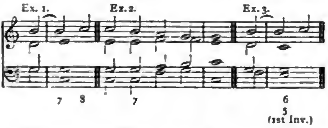
Ex. 1.
Ex. 2.
Ex. 3.
7 8
7
6
5
(1st Inv.)

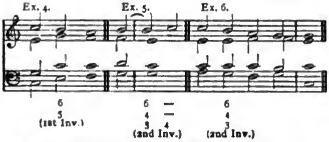
Ex. 4.
Ex. 5.
Ex. 6.
6
5
(1st Inv.)
6 —
4 —
3 4
(2nd Inv.)
6
4
3
(2nd Inv.)

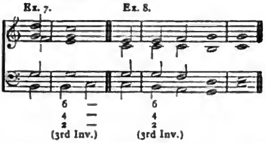
Ex. 7.
Ex. 8.
6 —
4 —
2
(3rd Inv.)
6
4
2
(3rd Inv.)

From the above it will be seen that the seventh from the tonic sometimes ascends, sometimes descends. In the former case (Exs. 1, 3, 5, 7), it is called the suspended leading-note or sub-tonic; in the latter, the chord of the tonic seventh.

(2) The next primary seventh

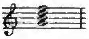

is called the chord of the dominant seventh, or minor seventh, and, by some, the fundamental seventh. It has three inversions, namely:

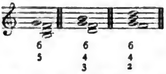

It is only possible to give a few of its many possible resolutions:

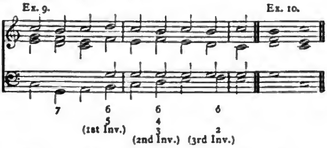
Ex. 9.
Ex. 10.
7
6
5
(1st Inv.)
6
4
3
(2nd Inv.)
6
2
(3rd Inv.)

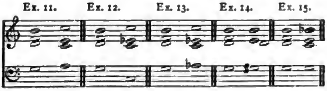
Ex. 11.
Ex. 12.
Ex. 13.
Ex. 14.
Ex. 15.

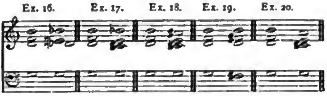
Ex. 16.
Ex. 17.
Ex. 18.
Ex. 19.
Ex. 20.

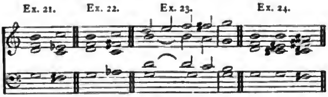
Ex. 21.
Ex. 22.
Ex. 23.
Ex. 24.

These resolutions have been divided into diatonic, chromatic, and enharmonic; also, into simple and compound; and into related and extraneous, &c., &c.

(3) The secondary sevenths are generally

named after the degree of scale on which they occur; thus, chord No. 3 (Ex. 25) is called the "seventh on the supertonic;" No. 4 (Ex. 26) "seventh on the mediant;" No. 5 (Ex. 27) "seventh on the subdominant;" No. 6 (Ex. 28) "seventh on the submediant;" No. 7 (Ex. 29) "seventh on the leading-note."

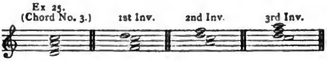
Ex 25.
(Chord No. 3.)
1st Inv.
2nd Inv.
3rd Inv.

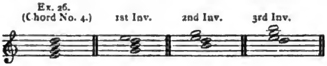
Ex. 26.
(Chord No. 4.)
1st Inv.
2nd Inv.
3rd Inv.

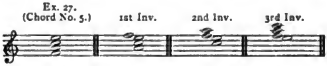
Ex. 27.
(Chord No. 5.)
1st Inv.
2nd Inv.
3rd Inv.

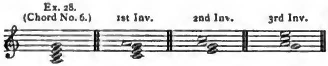
Ex. 28.
(Chord No. 6.)
1st Inv.
2nd Inv.
3rd Inv.

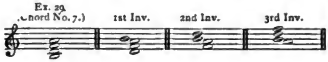
Ex. 29.
(Chord No. 7.)
1st Inv.
2nd Inv.
3rd Inv.

The following are the most common resolutions of chord No. 3, the seventh on the supertonic:

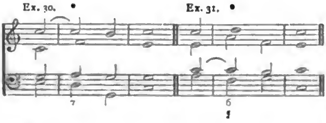
Ex. 30.
Ex. 31.
7
6

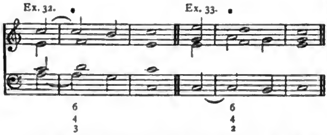
Ex. 32.
Ex. 33.
6
4
3
6
4
2

The first inversion of the supertonic seventh (Ex. 31) is also known as the chord of the added sixth, *q.v.* By some authors, the chord No. 3 (Ex. 25) is said to be derived from G, it becomes therefore the chord of the 11, 9, 7, 5, of that fundamental bass. On this principle, the chord * (Ex. 30) is the second inversion of the chord of the eleventh on G; the chord * (Ex. 31) the third inversion of the same chord, and so on.

(4) The resolutions of chord No. 4 (Ex. 26) are commonly as follow:

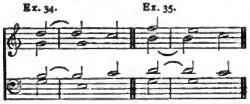

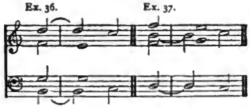

Most authors treat this as part of the chord of the ninth on C; namely, C, E, G, B, D, with the C omitted. [Ninth, chord of.]

(5) The resolutions of chord No. 5 (Ex. 27) are generally as follow:

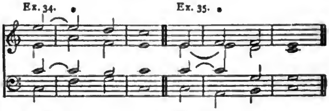

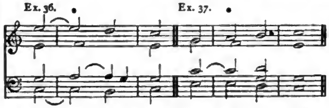

Many authors call the above the third (Ex. 34), fourth (Ex. 35), fifth (Ex. 36), and sixth inversions of the chord of the dominant thirteenth of G, of which fundamental bass they form the intervals of the 7th, 9th, 11th, and 13th.

(6) The following are common resolutions of chord No. 6 (Ex. 28):

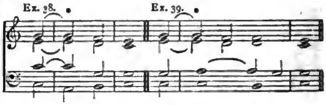

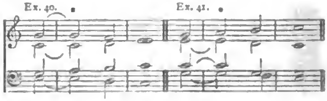

This chord is by some authors considered as a derivative of the subdominant F. From this point of view the chords * (Exs. 38, 39, 40, 41) become the first, second, third, and fourth inversions respectively of a fundamental F. Others look upon this chord as a thirteenth from G, and as analogous to that described in No. 5.

(7) The following are the common resolutions of chord No. 7 (Ex. 29), known as the "seventh on the leading-note."

Ex. 42. Ex. 43. Ex. 44. Ex. 45.

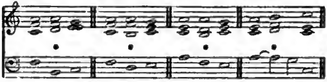

This chord is attributed by almost universal consent to the fundamental note G, and is called the chord of the "dominant ninth," or "added ninth," or "fundamental ninth," or "major ninth."

Chords of the seventh are also called fourfold chords, because in their full form they are made up of four sounds.

The chord of the diminished seventh will be found described under the chord of the minor ninth.

Severamente (*It.*) Severely, strictly, exactly.

Severità (*It.*) Severity, strictness, exactness.

Sexquialtera (*Lat.*) [Sesquialtera.]

Sext (*Ger.*) (1) A sixth. (2) The name of an organ stop of two ranks, having the interval of a sixth between them, namely, a twelfth and tierce.

Sextetto. [Sestetto.]

Sextolet, *or* **Sextuplet**. A double triplet, six notes to be performed in the time of four.

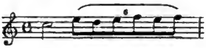

Sextuor (*Fr.*) A composition in six parts.

Sf. *or* **Sfz.** Abb. of *Sforzando* or *Sforzato*.

Sforzando (*It.*) [Sforzato.]

Sforzato (*It.*) Forced. A term signifying that the note or notes pointed out by the sign *sf.* are to be emphasised more strongly than they would otherwise be in the course of the rhythm.

Sfz. p. Abb. for *Sforzato piano*, a sudden *forte* followed by a *diminuendo* or *piano*. Also indicated by the sign >.

Sfuggito (*It.*) Avoided, shunned, as *Cadenza sfuggita*, an interrupted cadence.

Shading of pipes. The placing of anything so near the top of an organ pipe as to affect the vibrating column of air which it contains.

Shake. *Trillo* (*It.*), *Pralltriller* (*Ger.*) An ornament produced by the rapid alternations of two notes, either a tone or semitone apart, as the case may be. The sign of a shake is *tr.* (the first two letters of the word *trillo*) placed over the chief note:

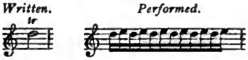

an indefinite number of times according to the fancy of the performer or the duration of the note. A shake, preceded by an appoggiatura is generally finished with a turn:

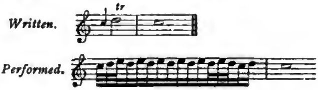

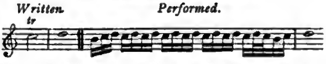

A succession of shakes is called a chain, *Catena di trilli*. A shake which commences with a *turn* is called a *prepared shake*. In harpsichord music a shake was written thus:

and played thus: 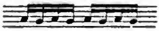

a plain note and shake, written thus:

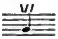 played 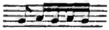

a turned shake written thus:

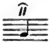 played 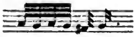

Sharp. (1) The sign which raises a note one semitone above the normal or *natural* pitch. A note so affected is restored to its normal pitch by the use of a natural. In old music sharps were often used to raise notes which had been previously flattened, for which purpose a natural is always now used. (2) An augmented interval is said by some to be *sharp*. In old writers a major third is called a *sharp* third. (3) Out of tune, by being higher in pitch than is just. (4) Shrill or acute, as *sharp mixture*, an organ stop.

Shawm. [Chalumeau.]

Shift. To change the position of the hand when playing on string instruments. On the violin, the first position is when the first finger is nearest the extreme (scroll) end of the finger-board, *i.e.*, on the note F, covering the notes to B with the four fingers; the second position is when the first finger is on G, covering to C, which is called also the *half shift*; the third position is when the first finger is on the note A, called also the first *whole shift*; the fourth position is when the first finger is on B, called also the *double* *shift*. But the term has been quite superseded by Position.

Short octaves. In old organs, in order to avoid the expense of large pipes which were not frequently used, only the most important notes between *c c* and *g g g* were employed. The following was a common system of arranging the four lowest sounds in English organs :—

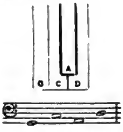

8ve lower.

The keys as above *appeared*, of course, to be B, C, C♯, D.

Si. The name of the seventh degree of the scale of Do. It was first suggested as a solfeggio syllable by Ericius Puteanus, of Dodrecht, in 1580, and again by Lemaire, of Paris, about the year 1690. In the scale as divided into hexachords by Guido, the seventh note as the first was called Ut; but the use of solmisation rendered a seventh name necessary. Za and Sa were both suggested at different times, the latter because it was a portion of the first syllable of the word *sancte*, one of the concluding words of the verse of the hymn which gave the names to the other notes. [Aretinian syllables.]

Andreas Lorente, in a part of his book, "El porque de la musica," 1673, suggested the syllable Bi for the leading note of the scale, but the syllable adopted by Puteanus and Lemaire took firmer hold of the musical mind of the public, from a supposed notion that the sibilant sound indicated to some extent the peculiarity of the tone.

Si bémol (*Fr.*) The note B♭.

Siciliana, Siciliano (*It.*) A graceful dance of the Sicilian peasantry, set to a melody in $\frac{6}{8}$ or $\frac{12}{8}$ time, of a simple pastoral character. Compositions or movements of like character are so named.

Side-drum. A small military drum frequently used in the orchestra. It is suspended at the side of the player and beaten with two wooden sticks on the upper head or surface, the lower having catgut strings called snares, stretched across to check the reverberation. Weber was the first to employ the side-drum as an orchestral instrument in his "Kampf und Sieg," 1815. Rossini used it two years later. [Drum.]

Siebenklang (*Ger.*) (1) A chord of the seventh. (2) A heptachord or scale of seven notes.

Siebenpfeife des Pan. [Pan's pipes.]

Siegeslied (*Ger.*) A song of triumph.

Siegue. [Segue.]

Sifflöte, Sufflöte (*Ger.*) [Flute.]

Signa. (1) An old name for large Church bells. (2) Characters and signs in mediæval music.

Signalist (*Ger.*) A military trumpet player.

Signatur (*Ger.*) [Signature.]

Signature. The signs placed at the commencement of a piece of music. There are two kinds of signature, the time-signature and the key-signature, the latter requiring a clef to show the pitch. (1) The key signatures, including the clefs, are usually written on every stave; the time-signature only at the commencement of the first line and where changes occur. It would be more proper to call the time-signature the *measure-sign*, as it shows the contents of a bar but not the *pace* at which the music should be performed.

The signatures of minor keys are the same as those of their relative majors. This leads to some inconvenience, as it is often necessary to look into a piece before determining whether it is in the major or minor. Various remedies have been proposed for this, the most important of which is to mark the raised leading-note and minor sixth in the signature, *e.g.*:

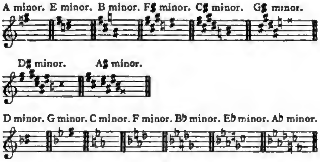

It is remarkable that writers before the middle of the eighteenth century frequently placed one sharp less in the signature than was absolutely necessary, and consequently were obliged to use an accidental at every recurrence of the leading-note. It seems difficult to account for this custom, unless such writers thought it wrong to acknowledge that the force of the leading-note overthrew the old church modes. Some writers consider that it was done in order not to disturb the proper position of the *mi* and *fa* in sol-faing. [Sol-faing.]

(2) Time-signatures are expressed by fractional parts of a semibreve. But the signs C and ₵ are, it is to be regretted, still much in vogue in common time. The former is a corruption of the semicircle C, which was used to denote duple or *imperfect* measure.

Digitized by Google

the whole circle O being used to denote *perfect* or triple measure. The stroke through the sign 𝄵 has two distinct meanings, namely, (1) a halving of the *contents* of the bar; (2) a halving of the *pace* of the music. Thus, 𝄴 formerly denoted *four minims* in a bar; and 𝄵 *two minims* in a bar, that is, half-measure. But in modern music 𝄵 generally indicates four minims in a bar, *allegro*, that is half the pace. There is not the smallest necessity for the use of either of these signs, as will be seen by the following excellent tables, drawn up by the Rev. J. Troutbeck, M.A. (Troutbeck and Dale, *Music Primer*.)

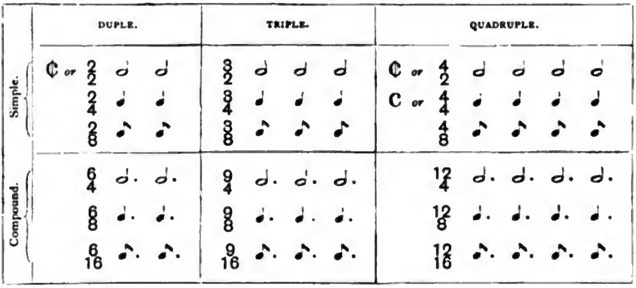
DUPLE.
TRIPLE.
QUADRUPLE.
Simple.
Compound.
𝄵 or 2/2
2/4
2/8
3/2
3/4
3/8
𝄵 or 4/2
𝄴 or 4/4
4/8
6/4
6/8
6/16
9/4
9/8
9/16
12/4
12/8
12/16

An excellent suggestion is thrown out by the above writer, the adoption of which would point out where the *accent* falls in compound times. He proposes to say $\frac{2}{2\cdot}$ (two dotted minims) instead of $\frac{6}{4}$, and $\frac{2}{4\cdot}$ (two dotted crotchets) instead of $\frac{6}{8}$, and so on. Not only are the number of beats, and consequently, the position of accents in this way shewn, but a simple and uniform table can be drawn up thus (*Music Primer*, p. 41):

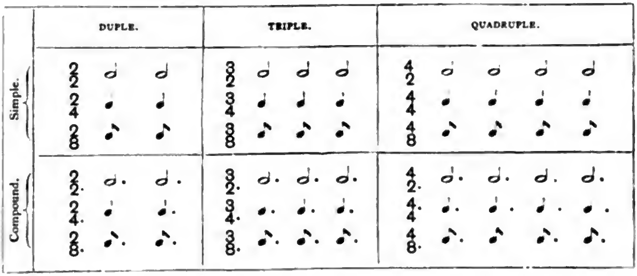
DUPLE.
TRIPLE.
QUADRUPLE.
Simple.
Compound.
2/2
2/4
2/8
3/2
3/4
3/8
4/2
4/4
4/8
2/2·
2/4·
2/8·
3/2·
3/4·
3/8·
4/2·
4/4·
4/8·

In ancient music, when the semibreve was really a short note, a duple measure of 2 (𝅝 𝅝); a triple measure of 3 (𝅝 𝅝 𝅝); a quadruple measure of 4 (𝅝 𝅝 𝅝 𝅝); with compound measures $\frac{6}{2}$ (𝅝· 𝅝·) and $\frac{9}{2}$ (𝅝· 𝅝· 𝅝·); are to be found.

Sign. A note or character employed in music.

Signe (*Fr.*) The sign 𝄋. [Segno.]

Siguidilla. [Seguidilla.]

Silence (*Fr.*) **Silenzio** (*It.*) A rest.

Si leva il sordino (*It.*) A direction that the mute (sordino) is to be taken off.

Silver strings. The covered strings used on violins, tenors, violoncellos, guitars, &c [String.]

Sim. Abb. of *simile*.

Simicion *or* **Simekion** (*Gk.*), *σιμίκιον.* A harp with thirty-five strings, known to, and occasionally used by the Greeks.

Simile (*It.*) Like; in the same manner. A direction that a method of performance previously ordered is to be adhered to in all similar passages.

Similar motion. [Motion.]

Simpla (*low Lat.*) *Semiminima.* A crotchet.

Simple. (1) Not florid; as, *simple counterpoint.* (2) Not developed; as, *simple imitation.* (3) Not exceeding an octave; as, *simple interval.* (4) Containing only one group of notes; as, *simple measure, simple time.* (5) Without valves or pistons; as, a *simple tube.* (6) That which cannot be resolved into constituents; as, *a simple tone.*

Sin' al fine (*It.*) To the end.

Sinfonia (*It.*) **Symphonie** (*Fr.*) [Symphony.]

Singakademie, Singschule (*Ger.*) An academy or school for singing.

Singend (*Ger.*) [Cantabile.]

Singetänze (*Ger.*) Song-dances. Ballads.

Singhiozzando (*It.*) In a sobbing style.

Single action. [Pianoforte.] [Harp.]

Single chant. [Chant.]

Single fugue. A composition in which only one subject is employed. [Fugue.]

Single relish. An old ornament in harpsichord music, violin playing, or singing. It was also called a *cadent.*

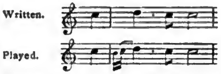

Singschule (*Ger.*) A school for teaching vocal music. A song-school.

Singspiel (*Ger.*) [Opera.]

Singstimme (*Ger.*) A vocal part.

Sinistra (*It.*) The left hand.

Sinkapace. [Paspy.]

Si piace (*It.*) At pleasure, *ad libitum.*

Si replica (*It.*) To be repeated.

Si scriva (*It.*) As written, without impromptu embellishment or alteration.

Si segue (*It.*) As follows, go on.

Sister (*Ger.*) An old German guitar having seven gut strings, the three lowest covered. It was tuned to G c f g c′ e′ g′.

Sistro (*It.*) A triangle.

Sistrum (*Gk.*), *σεῖστρον* (from *σείω* to shake). A rattle used by the ancient Egyptians, the Greeks, and Romans. It is not improbable sistrums were known also to the Hebrews, if the word *menaaneim* is correctly traced to a root signifying to "rattle." Its common form was that of a handle surmounted by a loop of metal having cross-bars on which rings were sometimes placed. (Fig. 1.)

Fig. 1.

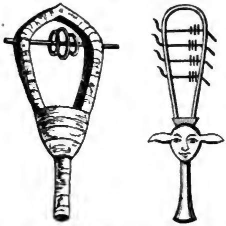

Others had bars of unequal length without rings, a fact which has led some writers to suppose that they were struck with a piece of metal held in the other hand. (Fig. 2.)

Fig. 2.

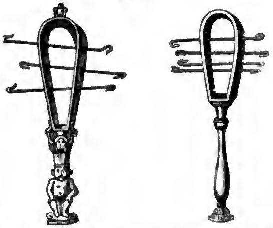

Fig. 3 shows Egyptian ladies rattling sistrums at a religious ceremony.

Fig. 3.

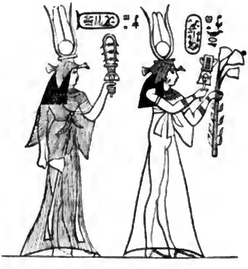

Sitole. [Citole.]

Si tace (*It.*) Be silent.

Si volta (*It.*) Turn over.

Sixième (*Fr.*) The interval of a sixth.

Six pour quatre (*Fr.*) A sextuplet, *q.v.*

Sixte (*Fr.*) The interval of a sixth.

Sixteen feet. The length of the open pipe which gives the unison of the pedal organ and the double of the manuals.

Sixteenth note. A semiquaver, the sixteenth part of a semibreve.

Sixth, added. [Added sixth.]

Sixth, chord of the. The first inversion of the common chord; it consists of a note with its third and sixth.

Sixth, chord of the French, German, Italian. [Extreme Sixth.]

Sixth, Neapolitan. [Neapolitan Sixth.]

Skald *or* **Scald.** A Gothic poet, priest, or bard. According to Percy the word denotes "smoothers and polishers of language," but others derive it from *gala* to sing, whence *galld*, enchantment, and *sgalld*, *skalld*, an enchanter. The root "gala" appears in the termination *gall* in the word nightingale. The name was especially applied to those who, in addition to natural gifts, possessed some degree of education, that is to say, a knowledge of versification, mythical imagery, and the traditions of their country. Skaldic poetry had for its object the celebration of the deeds of living warriors or their ancestors. The Skalds were attached to the courts of the Scandinavian princes, it being accounted honourable to be possessed of the most skilful of these poets. There are few complete poems of the Skalds extant, but a large number of fragments are preserved, partly by the younger Edda, partly in the Sagas and the Heimskringla. The Eddas are songs as old as the 6th century, the Sagas are historical and legendary tales. The Heimskringla, or Mythic ring of the world, records the history of the kings of Norway from the earliest times to the year 1177, the year in which the historian Snorri Sturleson was born.

Skip. A movement from any one note to another which is at a greater interval than one degree. [Disjunct motion.]

Skizzen (*Ger.*) Sketches. Short pieces, sometimes suggestive of some particular subject, not in any prescribed form.

Slancio, con (*It.*) With eagerness, impetuosity; from *slanciare*, to rush upon.

Slargando, Slargandosi (*It.*) Widening, opening, extending. Used as an equivalent for *rallentando*.

Slentando (*It.*) Slackening the time, becoming slower by degrees.

Slide. (1) An arrangement in the trumpet and trombone, by means of which the tube can be lengthened so as to generate a new series of harmonics. (2) To slide is to pass from one note to another without any cessation of sound, or distinction between the intervals. (3) A slider of an organ.

Slider. [Organ.]

Sliding relish, *Coulé* (*Fr.*) A grace in old harpsichord music,

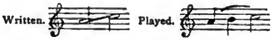

Slur. A curved line placed over notes directing that they are to be played or sung legato. [Bind.] A slur is often used in modern music to show the phrasing. In violin music a slur directs that the notes under it are to be played with one bow. [Bowing.]

Small Octave. The name given by the Germans to the notes,

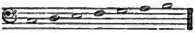

and their intermediate semitones. It is also called the lesser octave, and is described by the small letters c, d, e, f, g, a, b. [Pitch.] [Tablature.]

Smaniante, Smaniare, Smanioso (*It*) Furious, frantic, with rage.

Smanicare (*It.*) To shift. [Shift.]

Sminuendo, Sminuito, Smorendo (*It.*) Diminishing, decreasing gradually, lessening the time and tone.

Smorfioso (*It.*) Affected, coquettish.

Smorzando, Smorzato (*It.*) Gradually fading away.

Snap. [Scotch Snap.]

Snare Drum. [Side Drum.]

Soave, Soavemente (*It.*) Agreeably, delicately, gently, softly, sweetly.

Sobb. An old word for *damping* in lute playing. "Cause them (the strings) to sobb, by slacking your stopping hand so soon as they are struck; yet not to unstop them, but only so much as may dead the sound on a sudden. This gives great pleasure in such cases." Mace, 1676.

Soggetto (*It.*) Subject, theme, motive, proposition of a fugue.

Sol. The note G. [Sol-faing.]

Sol-bémol (*Fr.*) The note G flat. *Sol-bémol majeur*, the key of G flat major. *Sol-bémol mineur*, the key of G flat minor. *Sol-dièse*, the note G sharp.

Solennemente (*It.*) Solemnly.

Solennità (*It.*) Solemnity, pomp.

Sol-fa (*It.*) A general name for the notes in music. [Sol-faing, Solfeggi, Tonic Sol-fa.]

Sol-faing. A system of singing; a composition in which the names of the notes are employed instead of the words to which it may be set. Formerly only four of the seven names of the notes—Ut, Re, Mi, Fa, Sol, La, Si, were used, namely, Mi, Fa, Sol, La. These were applied to every note in the scale.

Digitized by Google

on the principle that it is naturally divided into two halves of similar proportions.

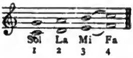

Thus, from 1 to 2 is a tone, from 2 to 3 is also a tone, and from 3 to 4 a semitone, whether the upper or lower series be the first reckoned. So, as in the scale of C, from Fa to Sol and from Sol to La are each a tone apart, and from Mi to Fa only a semitone; *all* tones in the scale were distinguished by these names for the purpose of Sol-faing.

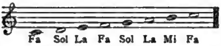

Mi was always used for the leading, or master note. This series, repeated to any extent, was supposed to express all the different orders of tones and semitones in the diatonic scale. "Above Mi will stand Fa, Sol, La, and below it the same inverted La, Sol, Fa, and one Mi is always distant from another an octave, which cannot be said of any other of them, because after Mi ascending come always Fa, Sol, La, Fa, which are repeated invertedly, descending."

The old rules for remembering Mi in all keys were thus set forth, the positions of the notes being called by the seven letters of the alphabet in use then as now;

If that no flat is set in B,
Then in that place standeth yͤ Mi:

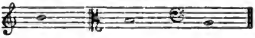

But if your B alone is flat,
Then E is Mi. be sure of that:

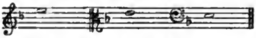

If both be flat, your B and E,
Then A is Mi, here you may see:

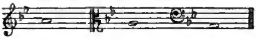

If these be flat, E, A and B,
Then Mi alone doth stand in D:

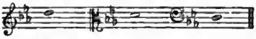

If all be flat, E, A, B, D,
Then surely Mi will stand in G:

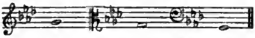

Learn this, and learn it well by rote,
That Mi is aye the last sharped note
For if a sharp on F be set,
To call that Mi, do not forget:

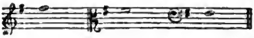

And if another on C be found,
Remember there your Mi to sound:

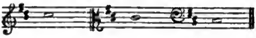

And if one more be set on G,
Then in that place will stand your Mi:

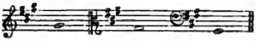

If all be sharp, F, C, G, D,
Then Mi alone will stand in D:

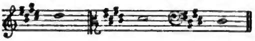

In the modern method of Sol-faing no distinction is made between tones and semitones by the use of a fixed nomenclature for the proportions of the scale, but the notes in any key are called by the names they bear in the key of Do or C.

Thus, a piece in E or E flat would be Sol-faed by calling the scale notes—Mi, Fa, Sol, La, Si, Do, Re, Mi, whether they were sharp, flat, or natural. The Tonic Sol-fa method calls all scales starting from any one of the twelve semitones by the same names as it gives to the notes of the scale of C, namely, Doh, Ray, Me, Fah, Soh, Lah, Te, Doh.

Solfège (*Fr.*) A vocal exercise in which the notes are called by the several names Do, Re, Mi, Fa, Sol, La, Si. [Sol-faing.]

Solfeggiamenti (*It.*) Solfeggi.

Solfeggiare (*It.*) To practise solfeggi.

Solfeggio (*It.*) [Solfège.]

Solist. [Soloist.]

Solito (*It.*) Usual, used, accustomed. In the ordinary manner.

Sollecito (*It.*) Careful; a word directing a careful and attentive manner of performance.

Solmisare (*It.*), **Solmizare** (*It.*), **Solmisiren** (*Ger.*) To Sol-fa. [Sol-faing. Solfeggio.]

Solmisation. [Sol-faing.]

Solo (*It.*) Alone. *Soprano solo*, the soprano alone; *voci soli*, voices alone, &c.

Soloist. One who sings or performs alone, with or without the aid of accompaniment.

Solo pitch. The tuning of an instrument a little higher than the ordinary pitch in order to obtain brilliancy of tone with a certain amount of ease to the player. [Accordatura.]

Solospieler (*Ger.*) A solo player.

Solostimme (*Ger.*) A solo part.

Son (*Fr.*) Sound, tone.

Sonabile (*It.*) Sounding, resonant.

Sonare (*It.*) To sound, to play upon. *Sonare alla mente*, to play extempore. *Sonare il violino*, to play upon the violin.

Sonata. The word Sonata is supposed by some to be derived from the Italian word *sonare*, to sound, to play, but by others from *sonetto*, a sonnet.

The term Sonata or Suonata, as applied to a musical composition, was first used about the beginning of the 17th century. Those of that time so called, had but one movement; they were in fact, simply *airs* arranged in parts, for an instrument or instruments. But the title of *airs* was given to some Sonatas as late as 1770, for in Hoyle's "Dictionarium musicæ" of that date, *voce* "Suonata," we read "of Corelli's Musick, the first and third operas are Church Sonatas, and the second and fourth, Chamber Sonatas; though the common distinction among us is made by the name of *airs*."

Other terms were occasionally used to describe Sonatas, such as Consorts, Ayres, Lessons, Fantasies or Fancies, "so made as they must be plaid and not sung," thus showing the antithesis between a sonata and a cantata.

Sir John Hawkins, in his "History of Music," says these titles were disused about the middle of the 17th century, when new forms of concerted pieces of a more elegant character came into vogue; these were called "the Sonata di Chiesa and the Sonata di Camera." The first of these, as being adapted to Church Service, was grave and solemn, consisting of slow movements, intermixed with fugues; the other admitted of a variety of airs to regular measures, such as the Allemande, the Courant, the Sarabande, &c.

The slow movements "intermixed with fugues," arose from the introduction of the Canzona, and Sir John Hawkins in mentioning this as a characteristic of the Sonata, unconsciously proves that the influence of the Canzona style had not weakened or faded at the time he wrote.

The connection of the Canzona with the Sonata is to a certain extent indicated in the pieces by Frescobaldi, published at Venice (1634). "Canzone da Sonare a una, due, tre, et quattro, con il basso continuo," which are exactly similar in style to the compositions recorded in early times as Sonatas. When in subsequent compositions of this kind two or more movements were employed, those most favoured were a Canzona or something in Canzona style, and a dance tune, such as a Pavan, an Allemande, or a Coranto. The Canzona was probably selected by the composer as a vehicle for the display of art and skill, and the dance tune was offered as a conciliation to the taste of his hearers. Long after the form was fixed, compositions bearing the name of Sonatas, constructed on the principles and according to the models of older times, were published, some as late as the end of the last century, as intimated above.

Frescobaldi's Canzone consist of only one movement, with various changes of time. A short analysis of the whole piece from which the illustration to the word Canzona is taken, will give an idea of the character of these early Sonatas. The first phrase of 19 bars is in fugal style, then follow 12 bars in $\frac{6}{4}$, after which are 27 bars in duple time in which free imitation is employed, then 9½ bars in $\frac{6}{4}$ again, the music of a different character to that in the former movement, of the same pulse measure, then 2 bars by way of a coda or conclusion, in common time. No portion is marked for repetition, but the whole thing goes straight on from the first note to the last.

The earliest compositions to which the title of Sonata or Suonata was attached, were written by Bonifacio Graziani (1609-1672), Marc Antonio Cesti (1624-1675), and Paolo Colonna (1630-1690), among others. It is presumed that some of the works of these musicians were those brought to England by John Jenkins (1592-1680), who afterwards, upon the models so suggested, published in 1660 in London "Twelve Sonatas for two violins and a bass, with a thorough bass for the organ." These were the first compositions of the kind by an Englishman. Jenkins was already well known as an agreeable writer of "fancies for viols," and his Sonatas show a certain amount of artistic progress in the arrangement of contrasted movements. Most writers on musical history declare that Francis Henry Biber (1648-1698), was the first who published a work with the title of a Sonata, but his compositions did not appear until 1681, more than twenty years after those by Jenkins with the same title; and there were also the still earlier Italian writers named above, from whom Jenkins confessedly obtained the idea. Considering the variety of the German tongue, and the unwillingness of the German people to use a foreign term when a native equivalent can be found, it is scarcely likely that the term Sonata would be first attached to a German composition by a German composer; therefore it is reasonable to assume that the word would have been adopted by the musicians of the country to which it belongs. Graziani, Cesti, and Colonna, who died before the time Biber published his Sonatas, used the term to describe certain of their compositions; but supposing the honour of first using it belonged to a German, then Johann Rosenmüller, who published at Venice "XII. sonate a camera, a 5 stromenti," in 1667, has a prior claim to Biber.

Henry Purcell (1658-1695), who was one of the earliest English writers of Sonatas, has almost as strong a claim to an early use of the word as Biber. Purcell composed Sonatas which were published in 1683, with the title of "Twelve sonatas of three parts, two violins and a base, to the organ or harpsychord." These were issued in separate parts, and in the sixth Sonata of this set is the melody whose character has given rise to the erroneous statement that Purcell composed the air upon which "God save the King" is founded. Furthermore, some writers declare that there is evidence in the construction of this set of Sonatas, that Purcell was indebted to Corelli for his ideas. If there is a sufficient similarity of style to warrant the assertion, there still remains the doubt whether Purcell could have seen the work of his Italian contemporary before his own was published, as both sets were issued in the same year, and communication, especially with a foreign country, was not so rapid then as now.

In the preface to this book of Sonatas, Purcell states that "he has faithfully endeavoured a just imitation of the most famed Italian masters, principally to bring the seriousness and gravity of that sort of musick into vogue and reputation among our countrymen, whose humour 'tis time now should begin to loath the levity and balladry of our neighbours." He further states, "He is not ashamed to own his unskilfulness in the Italian language, but that is the unhappiness of his education, which cannot justly be counted his fault; however, he thinks he may warrantably affirm he is not mistaken in the power of the Italian notes, or elegancy of their compositions." If Purcell benefited by the writings of any Italian musician, it could hardly have been by those of Corelli.

Purcell also composed a second set about the same time, but they were not printed until after his death. Among this second set was one well-known by the name of the "Golden Sonata," and as it is a very fair example of Purcell's compositions of this kind, and indeed of most others of the period, a short description may not be out of place here. It is written in three parts, for two violins and a figured bass. The key is F, and the first movement, only twenty-six bars in length, is a *largo* in common time of four crotchets. The subject, proposed by the bass and imitated by the trebles, is repeated in the key of the dominant at the 8th bar, but by the bass only, the trebles having a second subject. At the 12th bar the subject is inverted in the bass and answered according to the inversion by the other parts in fugal form, modulating into the key of the relative minor, in which key the subject re-appears, having for counterpoint an inversion of the first and second subjects; and with an imitation of the first episode as coda the movement ends.

There are altogether five movements, a *largo* of great beauty of melody, the second an *adagio* in $\frac{3}{2}$ time, with all the notes not in the key of F major written as accidentals, the third called *canzona allegro* is in F major (duple time), the fourth a *grave* in D minor (duple time), and the fifth is an *allegro* in F major $\frac{3}{8}$. No portion of any one movement is marked for repetition. The movements are each short, the last being the longest, and there is but little indication of that which is known as Sonata form in any, though each is written in a form regular enough in itself. Every device of imitation, inversion, and augmentation of the subjects proposed is employed freely, though not to a great extent. Lully [1634-1687], Pachelbel [1653-1706], Buononcini [1658-1702], the great Arcangelo Corelli, and Kuhnau [1667-1722], may be mentioned as Sonata writers contemporary with Purcell. Kuhnau being, perhaps, next to Corelli, the most remarkable. His early Sonatas were in three movements, an *allegro*, *andante*, and *allegro*, and for this reason his claim to merit as the suggestor of the modern form has been made out. One of these Sonatas or suites (No. 3) in E minor, may be taken as an example not only of the rest of his works, but of the form the Sonata had assumed in his day. Commencing with a Præludium of 22 bars in which sentences given out by the right hand are imitated in the left, there follows an Allemande with the half close in B minor, a Courante and a Sarabande with the first sections ending in B major; in none are the subjects developed, but the Gigue with which the suite concludes contains the nearest approach to modern form, in that the second section commences with the subject first proposed, though by inversion, and oddly enough the second half of the Gigue contains two bars less than the first, a very unusual circumstance—as in nearly all other instances the latter portion of a Gigue is the longer. There is not any attempt in either movement to introduce a subject of sufficiently marked character to justify its being called a second or distinctive theme.

The yearning after a fixed form, and the suggestion of the treatment which became expanded into the Sonata form may certainly be traced in Kuhnau's "Frische Clavierfrüchte," seven Sonatas, published in 1703 (not in 1696, as some say). In melodic treatment and expansion of subject these show a decided advance. That which in Frescobaldi appeared to be a mere capricious change of *tempo*, in Kuhnau became developed into movements of respectable length, but still without any remarkable development of chosen themes. Kuhnau often

Digitized by Google

makes the first subject of some of his movements heard in the dominant (whether the suite is in a major or a minor key) in the middle of a movement, but does not call, as it were, special attention to the fact by means of a double bar or a repeat. Johann Mattheson (1681-1722) frequently imitates his first subject in the second movement by inversion, and occasionally shows that he contributed something towards the settlement of form by the manner in which he treats his "Gigues," as may be seen in the example quoted *sub voce* "jig." His Sonata, published in 1713, and dedicated "to the person who shall best perform it" is in one movement only, and is capriccio-like in treatment. It may be here mentioned as an interesting fact that in the majority of the suites a Gigue is chosen as the concluding movement, and in its lively character as well as in the style of its construction is more in accordance with modern Sonata form than any other portion; this was the method often employed by Bach, Handel, and others to a later time. The thirty Sonatas of Alessandro Scarlatti (1659-1725), the next writer of importance after Mattheson, have each two movements, in which may be discerned a still further attempt to fix the style, and to impart some degree of unity.

Contemporary with Scarlatti was Tomaso Albinoni [1674-1745], whose Sonatas were at one time so popular in England, that the common fiddlers were able to play movements from them; but there is little indication of Sonata form in any one of his compositions. Domenico Alberti [1705-1739], whose name is associated with the arpeggio bass, said to have been invented by him, was one among the first, if not the very first who employed the nearest approach to modern Sonata form that had as yet been made. In his "VIII. Sonate per Cembalo" (1737), each one has two movements, and nearly each movement two well-defined subjects, properly introduced. With the exception of an occasional opening or closing chord, the whole of these movements are written in two-part harmony, and they are pleasingly effective. C, F, G major and minor, and A are the keys selected in obedience to the usual custom of the time of avoiding extreme keys, but Alberti employs accidentals freely for his bold and advanced modulations.

In the works of these earlier writers an advance towards the Sonata form may be traced, for in more than one there is a close upon the dominant in the first section, and the original subject is announced in the key of the dominant at the commencement of the second section; but there is little indication, if any at all, of what might be called a second or subsidiary theme.

In five out of eight of the Sonatas of Domenico Alberti, the Sonata form is nearly observed in the first movement, and often in the second, and there are few compositions of better construction, either for voices or instruments, of earlier date than his "Sonata in stilo nuovo" (1737).

Alberti's Sonatas are in two movements only, like those of many of those of his contemporaries and immediate successors, particularly those by Dr. Croft (1679-1727), Nicolo Porpora (1685-1767), Dr. Boyce (1710-1779), Fedeli (1715-1762), Carlo Tessarini (1715-1765), Frederic Theodor Schumann (1729-1760), Valentin Roeser (1740-1787), &c. Francesco Durante (1684-1755), who, as a teacher, enjoyed a high and honourable reputation, as a writer of Sonatas displays a refined and correct taste, but very little originality of conception, many of his thoughts being based upon the ideas of Scarlatti his master. The Sonatas and clavecin compositions of Domenico Scarlatti (1683-1757), were, according to Burney, "the wonder and delight of every hearer who had a spark of enthusiasm about him, and could feel new and bold effects, intrepidly produced by the breach of almost all the old and established rules of composition," but they contribute little towards a settlement of form.

It is probable that Handel noted the growth of the form, and occasionally employed it, some of the movements in his "Suites" having the orthodox first and second subject while others have only a single subject in each section. Some of the songs in his Italian operas are almost in Sonata form, wanting only the observance of certain details.

Some of the Sonatas of Christopher Wagenseil (1688-1779) are in correct form, though the second subject is timidly introduced and employed. The lingering fondness for the same tonality, which is the characteristic of the Suites, and the earliest Sonatas, is to be found in Wagenseil; one of his Sonatas in the key of F, a fair specimen of his work in this style, has an *allegro assai* as the first movement, an *andante grazioso* as the second, and a *minuet* as the third. The first and last movements are in F major, and the *andante* in F minor. The collection in which this Sonata is to be found was published between 1740 and 1750.

The changes of *tempo* in the earliest compositions of this kind, probably suggested a division into separate movements. Inherent musical feeling would doubtless prompt the composer to make the several movements offer as much contrast as possible, and the variety thus introduced would be accepted as the first canon of this class of composition.

Digitized by Google

The arrangement of the earliest "Suites des pièces" may be studied with advantage in reference to this point. In the "Pièces de Clavecin" by François Couperin, 1713, there is a set in C minor, consisting of an Allemande, followed by a first and second Courante, then a Sarabande, a Gavotte, and finally a Minuet, each movement varying in *tempo* as well as in character, though not in key. The number of movements in a Suite alternated between five and seven for many years, but by degrees they became lessened in number. Although the names of dance tunes ceased to be attached to the several movements, it was easy to see that much of their character was retained; for as composers began to feel that the measure of the dance tune had a tendency to cramp their musical thoughts, their models were altered or abandoned, and they expanded their movements at pleasure, without reference to the needs of the dance. Later, when three movements were adopted, the dance tune was restored, a preference being given to the Minuet or something in its kind, and this, with a slow and a quick movement, for a long time made up the recognised constitution of a Sonata. Beethoven added a fourth movement, the Scherzo, which he used sometimes instead of, sometimes in addition to, the Minuet; but he was not, as some say, the inventor of that movement, as Haydn in his quartets, and Bach in his Suites had previously employed a movement called *Scherzo*.

The "Clavier-Uebung," of J. S. Bach, published singly between 1726 and 1730, marked in Hoffmeister's edition as Œuvre I., No. 1, containing six Suites in various keys, offers a very remarkable series of studies, inasmuch as they show the master mind yielding to the custom of his time in the order and style of the arrangements of his suites or Sonatas, but as free from conventionality as possible, and pointing in the direction so worthily followed by later writers. In the first number there are six movements, all in the key of B flat, namely, a *prelude*, an *allemande*, a *courante*, a *sarabande*, a *minuet* and *gigue*. Each, with the exception of the Prelude, is, as near as possible, in Sonata form, though the second subject is not always treated in the tonic in the second section. As usual, there is a second Minuet in the place now occupied by the trio, of which more presently. The second Suite is in C minor, and commences with a *symphony* or overture of two contrasted movements, but free in form; the next movement is an *allemande*, the first section closing on the dominant (G major); the third movement a *courante*, similarly written; the fourth a *sarabande*, with the first close in the relative major; the fifth is a *rondeau*, in the form proper to such pieces; the last is a *caprice*, the second section having the first subject inverted. The third Suite, in A minor, opens with a *fantasia*, followed by an *allemande*, which is succeeded by a *courante*, more or less free in treatment. The next piece is a *sarabande* in good form, then a *burlesca*, next a *scherzo* in duple time (given as an illustration to the article "Scherzo"), lastly a *gigue* in fantasia style. The fourth Suite, in D major, begins with an *overture* of some length; then comes a *courante* not in form, then an *allemande*, then an *aria*, then a *sarabande*, the three last named obeying the rules of form; then there is a short *minuet* of two movements, and the last is a *gigue*. The fifth Suite, in G major, opens with what he calls a *preamble*, after which an *allemande*, a *courante*, a *sarabande*, a *tempo di minuetto*, a *passepied* and a *gigue*, the saraband only of all these movements answering to any extent the requirements of form. The sixth Suite, in E minor, the most remarkable and difficult of the whole set, begins with a long and beautiful *toccata*, having next an *allemande* with the close of the first section on the dominant B major, next a *courante*, treated in like fashion, then an air in rondo form, then a *gavotte* in Sonata form, a *sarabande*, written in fantasia style, and a *gigue* in the somewhat unusual *tempo* "alla breve," the first bar of which is quoted under the head "alla breve." In the third book of the second set, "Zweiter Theil der Clavier-übung," there is a minuet in B minor which, instead of a second minuet in the relative major, according to custom, has a "trio" in B minor; this is noteworthy, as it is probably the earliest instance of this use of the word. The construction and arrangement of this second set is so similar to the first, that with the exception just now pointed out, there is little or no point of difference to call for special remark.

The "Sechs leichte Clavier-Sonaten," of C. Philipp Emanuel Bach (1711-1788), published at Leipsic, 1766, contain each the orthodox three movements, a moderate, a slow and a quick; and the first, and sometimes the last movement of each is written in Sonata form, with the exception of No. 6, which begins with a movement in rondo form. Not one has the name of a dance tune attached, though there are several which would answer all needs, were they so named. Many of the Sonatas of his brother Wilhelm Friedemann Bach (1710-1784) though, cleverly and ably written, show a less regard for the symmetrical form so characteristic of the works of Philipp Emanuel. More than one of Friedemann's Sonatas, like those of the older writer, consist of a continuous movement with a frequent change of time, rhythm and tonality. Some portions fulfil

Digitized by Google

all the requirements of form, but they are not marked for repetition. There are no separate and distinct movements, and the form chosen might be called, if not a caprice, at all events an original idea of the rondo, for after many and varied contrasts of time and subject, the Sonata often concludes with the phrase originally given out.

Schobert's Sonatas, published in 1741, are especially remarkable, because the second subjects are introduced in proper keys, after due preparation, although they are not so fully developed as those by later writers. There is an evident design in expanding the subjects to a greater extent than that which had been the custom before his time. So many of the Sonatas by E. Bach and Schobert are constructed according to Sonata form as now accepted, that the honour of having fixed the form may fairly be divided between these two composers.

Schobert's works were at one time the most popular of any performed in England, Dr. Burney having introduced them here in 1766. This writer makes a mistake when he says that Schobert published nothing until 1764, for he was already well known as a composer when he was invited to become "Musician to the Prince de Conti" in 1760, and copies of his Sonatas were in the hands of every harpsichord player before that time. Emanuel Bach knew Schobert's value as a musician, for Dr. Burney states that "his (Bach's) party allowed Schobert to be a man of genius, but spoiled by his affectation of a new and extraordinary style. They further accused him of frequently copying himself." His writing is perfectly individual, is fresh and novel, and more like an anticipation of Haydn and Mozart than Bach, from whom all the musicians of the time were wont to copy.

It is scarcely necessary to pursue the question further, or to enter into any elaborate argument in support of the claim of any particular musician to the invention of that to which probably all have contributed more or less. For at the period of time to which our inquiries have now led, the Sonatas of Haydn and Mozart became the recognised form of this species of composition, and they are happily so well known that a particular description is unnecessary. But, notwithstanding the existence of such noble models, not every composer of a Sonata cared to use the recognised or Sonata form; even Cherubini, with his love for form, sometimes wrote Sonatas not in true form.

It is unnecessary, perhaps, to say that the Sonata was brought to its present perfection by Beethoven. The works of Clementi, the father of pianoforte-playing, Dussek, Hummel, Weber, Field, Onslow, Moscheles, and Schubert show no further advance in progress in development.

A modern Sonata is generally constructed upon the following plan:—

The *first movement* is an *allegro*, sometimes with an introduction, but more frequently without one; the *second movement*, ordinarily called "the slow movement," is set in any time, between *adagio* and *andante;* and the *final movement* is an *allegro*, written either in Sonata or in rondo form. If there is a *fourth movement*, it is usually placed after the slow movement, and is either a scherzo or minuet and trio. This is the broad outline; the details require a little more particular description.

The first movement should have two themes, unlike each other in character; for example, one vigorous and spirited, the other tender and expressive; and each should be capable of varied treatment. After the first subject has been well announced, a modulation into the key of the dominant, if the subject starts in the major, and into the relative major if the subject commences in the minor, should be made. In either case the second subject must be heard in the changed key before the half close or perfect cadence concludes the first part of the *allegro*. In the second portion of the *allegro*, a greater amount of license is permitted, this section often partaking of the nature of a free fantasia. The two principal themes are subjected to varied development, and are introduced in various keys at the will of the composer. Having exhausted all chosen devices, the first subject must now enter in its entirety, then the second subject should be heard, this time in the tonic or key in which the movement is written, if that mode be major; or it may appear in the key of the tonic major, if the mode be minor; but the episode and cadence which are to usher in the conclusion of this section, must be in the tonic key. Sometimes a coda, formed of a portion of the first subject, is added, and the movement ends in the key first proposed. This method of treatment is called the "Sonata" or "Binary" form.

The second movement may be written in several different ways, that is to say, treated like the first movement with two principal themes, or with only one principal theme.

Whatever be the precise form selected, the style of the music should afford a great contrast to that of the *allegro*.

The style of the last movement should afford a still greater contrast by its character, whatever be the form selected; if it is a Rondo, it should be lighter than the first or second movements. Many instances could be quoted in which the final movement is made the vehicle for the expression of quaint musical humour. [Form.]

Sonata da chiesa (*It.*) A church sonata, as opposed to a *sonata da camera.*

Sonate (*Fr.*) [Sonata.]

Sonatina (*It.*), **Sonatine** (*Fr.*) A short sonata. One in which the subjects are not developed at length.

Sonatore (*It.*) An instrumental performer.

Sonevole (*It.*) Sonorous, sounding, resonant.

Song. (1) A short poem intended for music. (2) A musical setting of a short poem or portion of prose. The word is generally applied to solos, but sometimes also to compositions for two or more voices. (3) The second subject of a sonata is sometimes called the "song-theme." [Air.] [Ballad.] [Lied.]

Sonometer. An instrument for measuring the vibrations of sounds.

Sonoramente (*It.*) Sonorously.

Sonore (*Fr.*), **Sonoro** (*It.*) Sonorous, resonant, harmonious.

Sonorità (*It.*) Harmony, resonance, sonorousness.

Sonorophone. A metal wind instrument of the Bombardon class.

Sons (*Fr.*) The name given by the Provençal poets to their lyrical productions.

Sons étouffés (*Fr.*) Stifled, veiled, or muffled tones. Sounds produced by a muted instrument.

Sons harmoniques (*Fr.*) Harmonic tones or sounds.

Sons pleins (*Fr.*) Full tones, applied to the production of full round tones by a voice or instrument.

Sonus (*Lat.*) Sound.

Sopra (*It.*) Above, before, over, upon, upper, as *Di sopra*, as above; *Come sopra*, as above or before; *Nella parte di sopra*, in the upper or higher part; *Contrappunto sopra il soggetto*, counterpoint over the subject.

Sopran (*Ger.*), *Soprano* (*It.*) The highest kind of female or boy's voice. Also the singer possessing that voice.

Soprana corda (*It.*) The highest string on the fiddle, *chanterelle.*

Soprano clef. The C clef upon the first line of the stave. [Clef.]

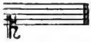

Sorda (*It.*) Muffled, veiled, muted, damped.

Sordamente (*It.*) Softly, gently, silently.

Sordini (*It.* pl.) (1) Mutes. Small instruments of metal, bone, or wood made to fix upon the bridge of a violin to damp or deaden the sound, by intercepting the vibrations. Mutes of wood covered with leather, of a pear shape, are sometimes used to check the sound of horns, trumpets, cornets, clarinets and oboes, the mute being inserted in the bell. (2) When mutes are required the direction *con sordini* is placed above the part so to be played, the contrary direction is *senza sordini.* (3) Dampers of a pianoforte.

Sordino (*It.*) A mute. A small pocket fiddle, a *pochette* or *kit*, formerly used for the purpose of giving the pitch, &c., at music parties.

Sordo, sorda (*It.*) Damped with a mute; as, *clarinetto sordo*, *tromba sorda*, &c. [Sordini.]

Sordun, Sordono (*It.*) (1) An old form of wood wind instrument, having a double reed, with twelve ventages and two keys. (2) A sort of mute for a trumpet. (3) An organ reed stop of 16 ft. pitch.

Sortisatio (*Lat.*) Counterpoint *alla mente.* [Alla mente.] [Chant sur le livre.]

Sortita (*It.*) (1) A word applied to the first piece sung by any one character in an opera. Thus, "Come per me" is the *aria sortita*, *aria d'entrata*, or entrance air for Amina in *La Sonnambula.* (2) A concluding voluntary, played as the congregation leaves the church.

Sospensivamente (*It.*) Doubtfully, irresolutely, waveringly.

Sospirando, Sospirante, Sospirevole, Sospiroso (*It.*) Sighing, subdued, wretched, doleful.

Sospiro (*It.*) A crotchet rest. In old music a minim rest.

Sost. Abb. of *Sostenuto.*

Sostenendo, Sostenuto (*It.*) Sustaining. Maintaining the tone for the full duration of the notes written.

Sotto (*It.*) Below, under; as, *sotto voce*, in an undertone; *sotto il soggetto*, below the subject.

Soubasse (*Fr.*) Sub-bass. A stop in the organ, of 32 ft. pitch.

Soubrette (*Fr.*) A serving maid. A female singer in a minor part of a comic opera.

Souchantre (*Fr.*) Succentor.

Souffárah (*Persian.*) The general name among the Persians and Arabs for wind instruments without reeds.

Soufflérie (*Fr.*) The apparatus connected with the bellows of an organ. [Organ.]

Souffleur (*Fr.*) (1) A prompter in a theatre. (2) An organ blower.

Soum. A Burmese harp.

Sound. [Acoustics.]

Sound-waves. [Acoustics § 3.]

Sound-board. (1) A piece of fir or other resonant wood placed behind the strings of a pianoforte for the purpose of increasing the power of the sounds. [Pianoforte.] (2) In an organ, the sound-board is the chamber of air into which the feet of the pipes are placed. [Organ.] (3) A wood screen placed behind a pulpit for the purpose of "reflecting" the preacher's voice; or over it, to prevent the

Digitized by Google

sound from ascending into a lantern-tower, or a dome.

Sound-body. Sound-box. [Body.] [Resonance Box.]

Soupir (*Fr.*) A crotchet rest 𝄽.

Soupir de croche (*Fr.*) A quaver rest 𝄾.

Soupir de double croche (*Fr.*) A semiquaver rest 𝄿.

Soupir de triple croche (*Fr.*) A demisemiquaver rest 𝅀.

Sourdeline (*Fr.*) A small kind of bagpipe, or musette. [Bagpipe.]

Sourdine (*Fr.*) (1) A mute (Sordino.) (2) A stop on the harmonium, which by limiting the supply of wind to the lower half of the instrument, enables the performer to play full chords *piano*.

Sous (*Fr.*) Under; as, *Sous-dominante*, the sub-dominant or fourth of the scale. *Sous-médiante*, the sub-mediant, or sixth of the scale. *Sous-tonique*, the sub-tonic or seventh of the scale; the leading or *note sensible*.

Spaces. The intervals between the lines of the stave. The stave consists of five lines and includes four spaces, but notes in spaces between leger lines above and below the stave, are employed. [Stave.]

Spagnoletta (*It.*) A dance in the Spanish style.

Spanisches Kreuz (*Ger.*) The Spanish cross, the sign of a double sharp, ×.

Spanish guitar. [Guitar.]

Spart. [Sparto.]

Sparto (*It.*) Scattered, distributed; hence, *a Score*, as the parts are arranged on several lines. [Score.]

Spartito (*It.*) Scored.

Spassapensiere (*It.*) Jew's harp.

Spatium (*Lat.*) } (1) A space on the stave,
Spazio (*It.*) } (2) an interval.

Spianato (*It.*) Smooth, level, even.

Spiccatamente (*It.*) Brightly, brilliantly.

Spiccato (*It.*) Distinct, detached, pointed. The direction for this is by *dots* over notes. [Bowing.]

Spielart (*Ger.*) Manner of playing, method of performance.

Spielen (*Ger.*) To play upon an instrument.

Spieler (*Ger.*) A player.

Spielmanieren (*Ger.*) Ornaments, graces, *broderies*.

Spinæ (*Lat.*) *Lit.*, thorns. A name given to the quills and jacks of a spinet, and sometimes to the instrument itself.

Spinet. Couched harp. *Spinett* (*Ger.*) *Spinetta* (*It.*) *Épinette* (*Fr.*) An ancient keyed instrument similar in construction to, but smaller in size than, the Harpsichord. The strings, which were placed at an angle with the keys, were sounded by means of leather or quill plectra (Spinæ.) [Clarichord.] [Harpsichord.] [Pianoforte.] [Virginals.]

Spirito, con, } (*It.*) In a spirited,
Spiritosamente } lively, animated, brisk
Spiritoso } manner.

Spirituale (*It.*) Sacred, spiritual.

Spirituel (*Fr.*) Ideal, pure, ethereal.

Spissa (*Lat.*) Close. (1) Intervals in the enharmonic and chromatic genus were said to be *spissa*, (πυκνά.) (2) In mediæval music a *pycnon* (πυκνόν) or *spissum* was a *semitone*.

Spitzflöte, Spindelflöte (*Ger.*) *Cuspida, flauto cuspido* (*It.*) An organ stop of 8 ft. or 4 ft. pitch, consisting of open flue-pipes, of a conical shape. Its tone is thin and reedy. *Spitzquint*, a quint-stop or twelfth of conical pipes.

Spondalium, *or* **Spondaulium** (*Lat.*) A Hymn sung during a sacrifice, accompanied by a flute.

Spondee (*Lat.*) A metrical foot consisting of two long syllables. [Metre.]

Spring. An old word for a tune

Square-piano. [Pianoforte.]

Squillante (*It.*) Ringing, sounding, bell-like in tone, from *squilla*, a little bell.

Sroutis. The name of the twenty-two parts into which the Hindu scale is divided.

Sta (*It.*) As it stands; to be performed as written.

Stabat Mater. A well-known Latin Hymn on the Crucifixion, sung during Passion week in the Roman Catholic Church. Jacopone, a Franciscan who lived in the thirteenth century, is supposed to have been the author of the words. In addition to the ancient setting, probably contemporary with the words, many composers have written music to the Stabat Mater, but the compositions which are best known are those by Palestrina, Pergolesi (the last effort of his life), and Rossini. The first of these three is a noble work, the second is full of pathos and expression, and the last is a quaint unison of operatic effects and florid vocal writing.

Stabile (*It.*) Firm, steady.

Stacc. Abb. of Staccato.

Staccare (*It.*) To make *staccato*.

Staccatissimo (*It.*) As *staccato* as possible.

Staccato (*It.*) Detached, taken off, separated. In music the word signifies a detached, abrupt method of singing or playing certain notes, by making them of less duration than they otherwise would be. A small dash over a note signifies that it is to be played *staccato*.

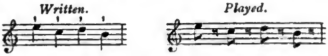

Sometimes a dot over a note is called a

Digitized by Google

staccato mark, but it is more properly the sign of a *spiccato*. Staccato notes are played shorter than those marked *spiccato*.

Stadtmusikanten, Stadtpfeiffer, Stadtzinkenisten, Kunstpfeiffer, Hausleute (*Ger.*) Town musicians. [Waits.]

Staff. [Stave.]

Stammakkord (*Ger.*) A key-chord or fundamental chord.

Stampita (*It.*) A sonata, an air, a song.

Standard pitch. [Pitch.]

Ständchen (*Ger.*) A serenade.

Standhaft (*Ger.*) Firm, steady, steadfast.

Stanghetta (*It.*) A bar-line. The vertical line placed on the stave to mark the division of bars.

Stanza (*It.*) *Lit.*, A station or resting place. (1) A series of metrical lines forming a verse or subdivision of a poem. (2) A strophe.

Stark (*Ger.*) Strong, loud; as, *mit starken Stimmen*, with loud stops.

Stave. Liniensystem (*Ger.*) Portée (*Fr.*) (1) A term applied to the five horizontal and parallel lines in music, upon which the notes or rests are placed

(2) Kircher describes a Greek manuscript which he found in the Jesuits' library of S. Salvator at Messina, the age of which he declared to be more than seven hundred years, in which hymns are set to music, written on staves of eight lines, no use being made of the intermediate spaces. Guido d'Arezzo, to whom the invention of the stave is usually assigned, may only have reduced the number of the lines, by making use of the spaces as steps in representing diatonic degrees. In the thirteenth century the number of lines was further reduced to four, a number still found sufficient for the purposes of the Plain Song.

Staves of various numbers, from three to six lines, are found in mediæval music, and in fact it was not until the invention of printing that the number of lines was settled at five. It is by no means an unusual thing to find six lines ruled in manuscript music books of the seventeenth and early part of the eighteenth centuries, and in some of the early printed books a variable number of lines is occasionally found. Frescobaldi in his "Tocate, canzone versi d'Hinni Magnificat, Gagliardi, Correnti, et Altri Partite di Cimbalo et organo," which was printed from copper plates in 1637, employs a stave of eight lines.

The stave of eleven lines, called also the "grand stave," has been in use for many generations; the following, from "L'Armonica pratico el Cimbalo di Francesco Gasparini," Venice, 1729, shows the form with the clefs then in use in Italy.

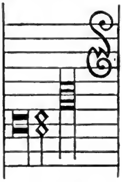

Sir John Hawkins, in his "History of the Science and Practice of Music," quotes a canzona from Frescobaldi's collection, but he has modernised the stave. Although five is now the general number of lines employed as the stave for all voices and instruments, yet it is by no means an unfrequent thing to find a stave of one line only for pulsatile instruments with a single unvarying tone, as cymbals, gongs, triangle, great drum, &c.

(3) In modern music the position of the clef upon the stave gives a special title to the stave; the G clef upon the second line, the C clef upon the first, second, third or fourth lines, and the F clef upon the fourth line, cause the stave to be known as the treble, soprano, mezzo-soprano, alto, tenor or bass staves respectively. A double stave, with the G and F clefs connected by a brace, is used for pianoforte and organ music, &c.

(4) A stanza, a portion of a song. A verse.

Steg (*Ger.*) The bridge of a violin, &c. *Ponticello* (*It.*), *Chevalet* (*Fr.*)

Stem. *Cauda* (*Lat.*), *Virgula* (*Lat.*) The line attached to the head of a note. All notes but the semibreve, or whole-note, have stems; quavers and their subdivisions have stems and hooks. In writing a "single part" for a voice or instrument, it is usual to turn the stems of notes lying below the middle line of the stave *upwards*, of notes lying above the middle line *downwards*. Notes on the middle line have their stems up or down as seems best, *e.g.*:

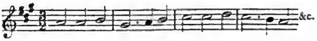

When two parts are written on one line, the tails of the upper part are *always* turned upwards, those of the lower part downwards. When both parts play the same note the stem is drawn both up and down, and if it is a semibreve, two notes intersect each other, *e.g.*:

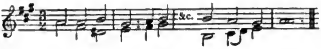

Digitized by Google

In *short* score, or a compression of four parts into two lines, the tails of notes in the treble part are always turned up, those of the alto down, those of the tenor up, those of the bass down; the treble and alto parts being confined to the upper stave, the tenor and bass to the lower, *e.g.*:

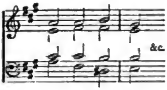

In groups of notes falling between the upper and lower stave the stems are sometimes partly up and partly down for convenience, *e.g.*:

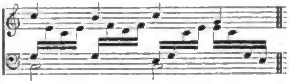

Up to the middle of the seventeenth century each note having a stem and hook was printed separately, instead of being bound together as in the above groups of semiquavers. [Printing of Music.] In mediæval music, for a short period, the position of the stem, up or down, affected the length of the note.

Stentando (*It.*) Delaying, retarding.

Stentato (*It.*) Forced, emphasised.

Steso (*It.*) Extended, spread, diffused. *Steso moto*, slow movement.

Stesso (*It.*) The same. *L'istesso tempo*, the same time.

Sthenochire. A hand strengthener. A machine for imparting strength and flexibility to the fingers for pianoforte-playing.

Sticcado *or* **Sticcato.** An instrument composed of pieces of wood of graduated lengths, flat at the bottom and rounded at the top, resting on the edges of an open box, and tuned to a diatonic scale. The tone is produced by striking the pieces of wood with small hard balls at the end of flexible sticks. A similar instrument made of glass or metal is called a Harmonica. [Ligneum Psalterium.]

Stiefel (*Ger.*) Boot of a reed-pipe in an organ. [Organ.]

Stift (*Ger.*) The jack of a spinet or harpsichord. [Jack.]

Stil (*Ger.*) Style.

Stile (*It.*) Style. [Stilus.]

Stillgedackt (*Ger.*) An organ stop of soft tone.

Stilus (*Lat.*) Style, as, *stilus choraicus* (*stilo coraico, It.*) the dance style; *stilus ecclesiasticus* (*stile ecclesiastico, It.*) the church style; *stilus familiaris* (*stile familiare, It.*) the simple style, note against note; *stilus hyporchematicus*, the theatrical style; *stilus madrigalescus*, the madrigal style; *stilus melismaticus*, the florid style; *stilus motecticus*, the motet style; *stilus phantasticus* (*stilo fantastico, It.*) the fantasia style; *stilus recitativus* (*stile rappresentativo* or *drammatico, It.*) the dramatic style; *stilus syllabicus*, the syllabic style, in which not more than one note is given to each syllable; *stilus symphoniacus* (*stilo sinfoniaco, It.*), the instrumental style.

Stimme (*Ger.*) (1) The voice. (2) Sound. (3) The sound-post of a violin or violoncello. (4) A part in vocal or instrumental music. (5) An organ stop, or rank of pipes.

Stimmgabel (*Ger.*) Tuning-fork.

Stimmhölzchen (*Ger.*) Sound-post of a fiddle, &c.

Stimmhorn (*Ger.*) Tuning cone. [Organ.]

Stimmpfeife (*Ger.*) A pitch-pipe.

Stimmschlüssel, Stimmhammer (*Ger.*) A tuning key or hammer-shaped instrument.

Stimmstock (*Ger.*) The sound-post of a violin or violoncello.

Stinguendo (*It.*) Fading away, dying away.

Stiracchiato, Stirato (*It.*) Widening, enlarging, retarding the time.

Stockfagott. [Rackett.]

Stonante (*It.*) Discordant, untuneful.

Stop. (1) The pressure by the fingers of the strings upon the fingerboard of a stringed instrument. (2) A fret upon a guitar or similar instrument. (3) A collection, register, or row of pipes in an organ.

Stopped diapason. [Organ.]

Stopper. The plug inserted in the top of an organ pipe, in order to "close" it.

Stopples. Plugs inserted in some of the ventages of the flute in order to accommodate its scale to some particular mode.

Storto, Storta (*It.*) A name formerly given to the horn, serpent, &c., because of their twisted form.

Strain. A musical subject forming part of, and having relation to, a general whole. At one time every subordinate portion of a composition either marking rhythmical pauses or completed sentences was distinguished by a double bar, and therefore the double bar was held to mark the strain, a practice which is still observed in writing chants and hymn tunes, where the double bar marks a strain, but not a completed phrase or subject.

Strascicando (*It.*) Dragging or drawling.

Strascicato (*It.*) Dragged, drawled.

Strascinando (*It.*) [Strascicando.]

Strascinando l'arco (*It.*) Drawing or dragging the bow over the strings so as to bind the notes together.

Strascino (*It.*) A drag or slur; a term applied to a slurring movement from sound down to sound, the pace at the same time being slightly slackened.

Strathspey. A Scotch dance in duple time, invented about the beginning of the 18th century, and first danced in Strathspey, from whence it derives its name. The "Scotch Snap" is one of the peculiarities of the tunes for this dance.

Stravagante (*It.*) Extravagant, capricious, fantastical.

Stravaganza (*It.*) Extravagance, fancy.

Street Music. [Licensing.]

Streichinstrument (*Ger.*) A stringed instrument played by the *stroke* of a bow.

Streichquartett (*Ger.*) String quartet.

Streichzither (*Ger.*) A zither played with a bow.

Strene. The name given by Marbecke to a breve.

Two explanations have been given of the meaning of the word *strene;* the first is, that it is a note which may be *stretched* or *strained* for the purpose of recitation; the second is that is bounded or *constrained* by two lines. Marbecke's use of the note certainly does not justify the first of these meanings.

Streng (*Ger.*) Strict, severe, rigid. *Streng gebunden*, strictly tied or legato, exceedingly smooth. *Streng in tempo*, strictly in time. *Strenge Fuge*, a strict fugue.

Strepito (*It.*) Noise.

Strepitosamente (*It.*) Noisily.

Strepitoso (*It.*) Noisy, impetuous.

Stretta (*It.*) A coda, a final passage taken in quicker time than the preceding movements. The conclusion of the chorus in Haydn's *Creation*, "The heavens are telling," is a *stretta*.

Stretto (*It.*) Contracted, close. A stretto in a fugue is the bringing closely together the subject and its answer. [Fugue.]

Striking-reed. A percussion reed. [Reed.]

String. Abb. of *stringendo*.

String. Prepared wire or catgut, plain or covered, used for musical instruments. Strings of steel or brass wire are used for all instruments which are struck with hammers or plectra, as dulcimers, zithers, mandolines, and pianofortes, and strings of catgut for instruments played with the unprotected fingers, or with a bow, as guitars, harps, violins, violas, violoncellos, and double-basses. Violin strings are made of catgut, each string being of a different thickness according to the tone and tension required, the fourth string being covered with a fine wire either of silver or white metal; hence it is called the silver string. Violas and violoncellos have each two silver strings, the object in using covered strings being to ensure a sufficient gravity of tone without having too clumsy a material. The covered strings on the guitar are upon a basis of silk instead of catgut, and the double bass strings are of thick gut uncovered. A large quantity of catgut strings for musical instruments is made in England, but the best are imported from Italy, which has, from time immemorial, been famous in this branch of industry. Rome, Venice, Pistoja, Lyons, were mentioned by Thomas Mace, 1676, as the most famous places from whence strings were brought in his day.

Silk has been sometimes used as the material of first violin strings, but with questionable success.

Among uncivilised people strings are made of the hair of animals, the fibres of creeping plants, of fibrous roots of trees, of cane, and of thongs of leather.

Stringendo (*It.*) Pressing, hastening on the time.

String-gauge. A small instrument for measuring the thickness of strings for violins, guitars, &c., consisting of a disc or an oblong piece of metal, with a graduated slit and engraved table.

String Quartet. (1) A quartet for two violins, viola, and violoncello. The most important and highest form of chamber music. Nearly all the great composers have left specimens of this class of composition. The classical form of the quartet may be said to have been established by Haydn and Mozart, and it has not been materially altered, though considerably developed, in the later works of Beethoven, Spohr, Onslow, and Mendelssohn. The mighty genius of Beethoven seemed to find in the string quartet a vehicle of his profoundest inspirations. (2) The group of string instruments in a band; sometimes also called the string quintet, because it includes the double-bass. But as the double-bass is chiefly used for the reduplication of the violoncello, the term "string quartet" is quite legitimate. The treatment of the double-bass is in many respects analogous to that of pedal-pipes in organ music: generally reinforcing, but often separated with good effect.

Strisciando (*It.*) Creeping, gliding, slurring smoothly from one note to another.

Strofa (*It.*) A strophe. [Stanza.]

Strohfiedel (*Ger.*) [Ligneum Psalterium.]

Stroke of the bow. [Bowing.]

Strombettare (*It.*) To sound a trumpet.

Strombettiere (*It.*) A trumpet-player.

Stromentato (*It.*) Instrumented, scored for an orchestra.

Stromento (*It.*) An instrument. *Stromento di fiato* or *di vento*, a wind instrument. *Stromento di corda*, a stringed instrument.

Strophe (*Gk.*) στροφή. *Lit.*, A turning. The turning of the Greek Chorus towards a particular part of the orchestra in dancing; as *antistrophe* (ἀντιστροφή) was their *returning*. Hence, the term came to be applied to the portions of the poem sung during these movements.

Stubenorgel (*Ger.*) A chamber organ.

Stück (*Ger.*) A piece, air, tune, composition.

Study. A term applied to an exercise for the pianoforte or other instrument.

Stufe (*Ger.*) A step, a degree. *Stufe der Tonleiter*, a degree of the scale.

Stürmisch (*Ger.*) Boisterously, furiously, impetuously.

Style. Character, form, or temperament of music with reference to, (1) the result of individual influence; as, *Handel's style*, *Spohr's style*; (2) the conformity of music to the purpose for which it was written, as the *Church style*, the *Glee style*; (3) the conventional or national method of performance; as, the *Italian style*, the *Scotch style*; (4) its construction; as, the *Chromatic style*, the *Fugue style*.

Suabe flute. [Flute, § 6.]

Suave (*It.*) Sweet, agreeable, pleasant.

Suavemente, Suavità, con (*It.*) Sweetly, with delicacy.

Sub-bass, Sub-bourdon. A pedal register in the organ, of 32 ft. tone. [Organ.]

Sub-chanter. Succentor.

Subdiapente, Subdominant. The fifth below or the fourth above any key note.

Subduple proportion. [Proportio.]

Subitamente, Subito (*It.*) Suddenly, without pause. *Volti subito*, turn quickly.

Subject. The theme or principal phrase of any movement, from which all the subordinate ideas spring or are developed. In sonata form there should be two chief subjects, called first and second; in rondo form, one is sufficient. In a fugue the subject is called also the exposition, dux, proposition. [Fugue.] [Sonata.]

Submediant. The sixth of the scale.

Suboctave. A coupler in the organ which pulls down keys one octave below those which are struck.

Subprincipal. An organ stop consisting of open pipes, of 32 ft. pitch on the pedals, and of 16 ft. pitch on the manuals.

Subsemifusa (*Lat.*) A demisemiquaver

Subsemitone. [Leading note.]

Subsemitonium modi (*Lat.*) The leading note.

Subtonic. The leading note. *Note sensible* (*Fr.*, Master-note, the semitone below the tonic.

Succentor, Sub-chanter, Souchantre (*Fr.*) An officer of the choir, the deputy of the Præcentor, appointed by the Dean and Chapter of a cathedral. In "Cathedralia" we find that:—"The succentor originally led the antiphonal chant on the side opposite the Præcentor, as S. Augustine includes under *cantores* the præcentor 'qui vocem præmittit in cantu;' and the succentor 'qui subsequenter canendo respondet.' The succentor major of canons was first instituted at Wells A.D. 1130-74. At Hereford the succentor or præcentor represented absent Vicars. The sub-chanter at York directed the minor canons, and the succentor, as in other cathedrals, was the præcentor's vicar, with regard to vicars-choral and songmen, and delated offenders to the Saturday chapter. At Chichester and Exeter he tabled the duties of the vicars in the week. At Salisbury and York he ranked next to the archdeacons. He supplied the præcentor's place during absence, and ruled the song-school by his officer. At Wells and Lichfield he wrote out all chants not in the table, arranged the method and order of the processions, enjoined the lections on greater doubles and in masses, and after Benedicite on Saturday in chapter arranged the table of services for the ensuing week. At St. Paul's he acts in the præcentor's absence, as regards the regulation of the service, and in olden times had to be obeyed by major canons, minor canons, and all other ministers. At Chichester his duty was to give stripes, seven or fourteen in his discretion, to the boys, if they behaved badly in the choir, and at Hereford his duty was to bear the 'burden' of the psalmody and chant, to distribute copes on the greater festivals, to order processions, to punish clerks of the first form who were not of the family of a canon, and suspend others offending. He delated offenders to the chapter, he took care that the singing was reverently conducted, and appointed five boys of clerks of the first form, removeable at his will, to sing the antiphons, and to carry tapers and thuribles. An honest robe, shoes, and stockings were provided for these boys out of the allowances of the succentor."

Succession. (1) The order in which the notes of a melody proceed. There are two sorts of succession, regular or conjoint, and disjunct. A regular or conjoint succession is that in which the notes succeed each other in the order of the scale to which they belong, either ascending or descending. In a disjunct succession the melody is formed of intervals greater than a second.

(2) A sequence is sometimes spoken of as a succession, and passages of similar chords or progressions are described as a succession of thirds, fourths, fifths, sixths, sevenths or octaves, as the case may be.

Sudden modulation. [Modulation.]

Sufflöte, Sifflote (*Ger.*) [Flute.]

Suite (*Fr.*) A set, series, or succession of movements in music. The term was

Digitized by Google

applied at an early period to collections of dance tunes of contrasted character but similar tonality. Galliards and Pavans chiefly formed the Suites composed until the middle of the seventeenth century. After which, almains, or allemands, corantos, grounds (passecailles), sarabands, jigs, minuets, passepieds, gavots, &c., were more favoured; and in the middle of the eighteenth century the titles of all the dances were dropped with the exception of the minuet, which was retained. Then the Suite became a Sonata, and the word Sonata, instead of being loosely applied to "Suites of Lessons for the Harpsichord," became employed to describe a composition of definite form and arrangement.

Suivez (*Fr.*) Follow. A direction to an accompanist to watch the singer or soloist, and accommodate the accompaniment to his singing or playing.

Sujet (*Fr.*) A subject, melody, phrase or theme.

Sul, Sull', Sulla (*It.*) On, upon, by. *Sulla soprano corda*, upon the first string; *sul ponticello*, by or near the bridge, in violin playing; *sulla tastiera*, upon the key-board.

Sultana. A violin with strings of wire in pairs, like the cither or cittern. It was similar to the *Streichzither.*

Summation tones. [Acoustics, § 19.]

Suo loco (*It.*) In its own place, in the register as written.

Suonata (*It.*) A sonata. *Suonata di camera*, a chamber sonata, a secular piece. *Suonata di chiesa*, sonata for the organ, piece for church use.

Super (*Lat.*) Above, over. *Superdominant*, the note next above the dominant, the sixth of the scale. *Supertonic*, the second of the scale.

Superfluous intervals. Those intervals greater by a semitone than major or perfect. [Interval.]

Superius (*Lat.*) A name given to the upper part in a composition by the writers of the sixteenth century.

Superoctave. (1) An organ stop tuned two octaves above the diapasons. (2) A coupler pulling down keys one octave above those struck.

Supersus. A term formerly applied to high treble parts.

Supertonique (*Fr.*) Supertonic. [Scale.]

Supplichevole (*It.*), **Supplichevolmente** (*It.*) In an imploring, supplicating manner.

Supposed bass. A term applied to any bass note forming one of the inversions of a chord, in contradistinction to the real bass or generator.

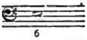

E is the supposed, C the real bass.

Sur (*Fr.*) Upon, on, over. *Sur la quatrième corde*, upon the fourth string; *sur une corde*, on one string.

Suspended cadence. An interrupted cadence. [Cadence.]

Suspension. The holding or prolongation of a note in any chord into the chord which follows, thereby often producing a discord. The first appearance of the note to be suspended is called its preparation; its presence as a discord, its percussion; its removal to a note of *rest in key*, or some legitimate sound of a sequence, its resolution. Suspensions are named after the interval of the note forming the discord. Two suspended notes form a *double* suspension; three a *triple* suspension, and so on. The intervals most commonly suspended are the fourth, sixth, seventh, and ninth.

The percussion of a discord of suspension is generally on the strong accent of a bar.

Ex. 1.

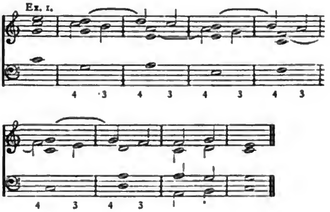

The above (Ex. 1) shows the suspension 4 3 on every note of the scale. It cannot legitimately appear on the subtonic, owing to the imperfection of the fifth.

The next (Ex. 2) shows suspended sixths on every degree of the scale but the seventh, on which it is not considered a genuine suspension.

Ex. 2.

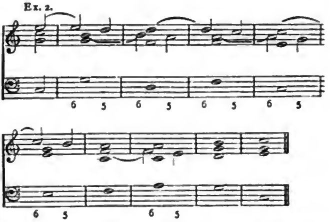

The next (Ex. 3) shows that suspended

Digitized by Google

sevenths can occur on every degree of the scale.

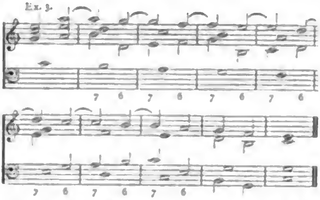

In the following (Ex. 4) a ninth is suspended on every degree of the scale. The chord marked * is rarely met with, and probably would not be tolerated unless as part of sequence such as that in which it occurs.

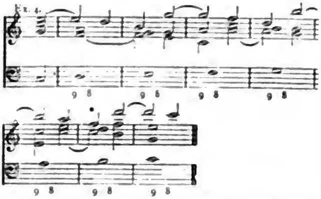

In Ex. 5, at *a* will be found the double suspension $\begin{smallmatrix}9&8\\4&3\end{smallmatrix}$; at *b* the double suspension $\begin{smallmatrix}9&8\\7&6\end{smallmatrix}$; at *c* the triple suspension $\begin{smallmatrix}9&8\\6&5\\4&3\end{smallmatrix}$; at *d* the triple suspension $\begin{smallmatrix}9&8\\7&6\\4&3\end{smallmatrix}$; at *e* the quadruple suspension $\begin{smallmatrix}9&8\\7&6\\6&3\\4&\end{smallmatrix}$.

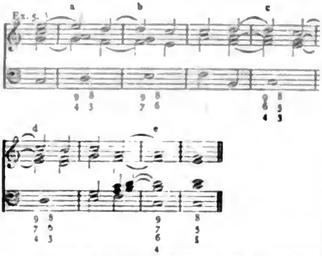

An exhaustive list of suspensions has never yet been attempted. The constant use of chromatic progressions in modern music has led to the formation of a vast number of *chromatic* suspensions which early writers on harmony never could have anticipated, and which modern theorists make no attempt to explain.

Süss (*Ger.*) Sweet.

Sussurando, Sussurante (*It.*) Whispering, murmuring.

Sustained note. A name given to prolonged notes which partake of the character of a pedal-point by their immunity from ordinary harmonic rules, but which cannot with propriety be called pedal-points owing to their occurrence in the middle or upper part, *e.g.*:

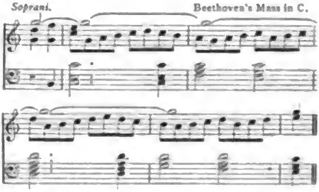

Svegliato (*It.*) Awakened, brisk, lively.

Svelto (*It.*) Swift, light, quick, free, easy.

Swell. [Organ.]

Symphonion. A combination of the pianoforte and harmonium, invented by Kauffman of Dresden.

Symphony, Symphonie (*Fr.*), **Sinfonia** (*It.*) (1) A composition for an orchestra, similar in construction to the sonata, which is usually for a single instrument. A symphony has several varied movements, generally four, never less than three. The first, an allegro; the second, a largo, adagio, or andante; the third, a scherzo, or minuet and trio; and the fourth, an allegro. The form of the first and last movement is usually that of the sonata. The scherzo, or the minuet, in some symphonies, is placed before instead of after the slow movement.

The arrangement of the symphony after this order is due to Haydn. About the same period Gossec composed a symphony, the twenty-first out of twenty-nine of which he was the author, which was a decided advance in musical form, not only upon his own previous compositions, but also upon those of others. Gossec's symphonies are of varied character, and according to the short description given by Fétis, belong chiefly to the class of concerto rather than to that of symphony, as they are for the most part for certain obbligati instruments. The "Symphonie en ré," written for the "Concert des Amateurs," gave the first impulse towards the perfection of instrumentation in France. It was written for two violins, viola, violoncello, contrabass, two oboes, two clarinets, a flute, two bassoons, two horns, two trumpets, and drums. "L'effet en fut très remarquable." It is, however, to Haydn that the perfecting of the symphony is due. For he possessed an extensive acquaintance with the character and resources of every known orchestral instrument. "If any doubts arose during composition, his situation at Eisenstadt gave him the power of resolving them into certainties immediately. He rang the bell for a rehearsal, the performers appeared, he caused them to execute the doubtful passages two or three various ways, and making his choice of one of them, dismissed the orchestra, and returned to his labours." [Sonata.]

(2) Formerly overtures were called symphonies; in the Dublin MS. of the "Messiah," the property of Sir Frederick Ouseley, Handel has called the overture "Sinfonica," and it was a common practice in his time to name any long instrumental piece after this manner.

(3) The introductory, intermediate, and concluding instrumental parts of a song or other vocal piece are also called symphonies.

(4) In the seventeenth century the virginal was sometimes spoken of as a symphony.

(5) A bagpipe has also been called a symphony, perhaps a corruption of the word *sampogna*.

Symposiac. A term applied to cheerful and convivial compositions for voices, as glees, catches, rounds, &c.

Synaphe (*Gk.*) The conjunction of two tetrachords. [Greek Music.]

Synaulia (*Gk.*) A concert of flute players performing alternately.

Syncopation, Syncopatio (*Lat.*), **Syncope** (*Fr.*) Suspension or alteration of rhythm by driving the accent to that part of a bar not usually accented. Syncopation may be completed in a bar:

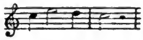

Or it may be carried by sequence through several bars:

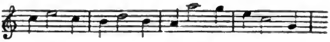

Or it may be so that more than one bar is involved in the syncopation:

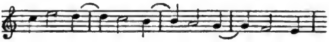

Syncopated counterpoint is the fourth species of counterpoint.

Syncopiren (*Ger.*) To syncopate. [Syncopation.]

Synemmenon. [Greek Music.]

Syren. [Acoustics.]

Syrinx. [Pan's Pipes.]

Szopelka (*Russ.*) A kind of oboe, about fifteen inches in length, made of elder wood, having a brass mouth-piece and eight large and seven small finger holes. It is a popular instrument in Southern Russia, and is the Western representative of the Eastern *Zourna*.

Digitized by Google

T.

T. Abb. of *tasto; t.s., tasto solo*; also of *tenor* and *tutti*.

Ta, te, tee, to. (τα, τε, τη, τω) (*Gk.*) Syllables used by the Greeks for purposes of solmization.

Taballo (*It.*) Kettle-drum.

Tabarde, *or* **Tabarte** (*Old Eng.*) Tabor.

Tabl. An Egyptian drum formed from a hollowed block of wood, or made of earthenware, with a skin stretched over one end.

Tablature (*Fr.*) **Intavolatura** (*It.*) **Tablatur** (*Ger.*) (1) A general name for all the signs and characters used in music. Those who were well acquainted with these signs were said to sing by the Tablature. (2) A peculiar system of notation employed for instruments of the lute class, for viols, and certain wind instruments. The earliest systems of notation, like the music of Asiatic nations to this day, were different sorts of Tablature of the character shown in the following example from Fétis:—

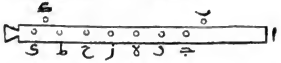

NOTES OF A PIPE IN TABLATURE.

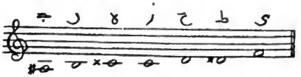

TABLATURE NOTES TRANSLATED BY FÉTIS.

That which may be called the modern Tablature was invented not earlier than the sixteenth century. The general character of the tablature employed was nearly the same in England, France, Spain, and Italy, though each country made variations in it according to fancy, and sometimes the several professors who taught in those countries had little peculiarities of their own in which they differed from their brother professors. These differences doubtless contributed towards making the general system as perfect as it could be in the end. Sir John Hawkins, quoting Mersennus, says that several skilful men had laboured to improve the Tablature, but they affected to make a mystery of it, and this gave rise to a diversity of notation between them. It is therefore probable that systems of writing in Tablature were in existence long before the publication of treatises on the subject. Adrien le Roy, a bookseller in Paris, published about 1570 a book with the title of "Briefve et facile Instruction pour apprendre la Tablature, à bien accorder, conduire, et disposer la main sur la Guiterne;" the popularity of this work was so great that it was translated into sundry languages as soon as possible after publication. An English translation, the basis of all subsequent systems published in England, was made by one signing himself with the initials F. K., and published by John Kingston in 1574.

The tuning of the lute according to Adrien le Roy was as follows:—

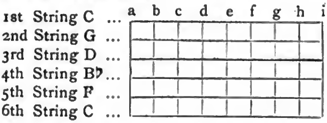

so that, with the use of the frets, usually eight, and lettered from *a*, the open string, to *i*, each string was capable of giving a minor sixth from its open tone. The letters refer to the frets which were on the finger-board, each fret representing a semitone, the letter *a* always standing for the open string. The horizontal lines stand for the strings, and the letters, the points at which they are to be stopped, so forming the notes of the composition. The signs for time were

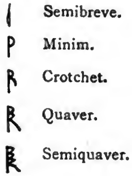

placed over the stave, and were intended to correspond with the semibreve, minim, crotchet, &c., respectively. It was to be understood that the sign after the bar represented the value of all the notes in that bar unless otherwise contradicted.

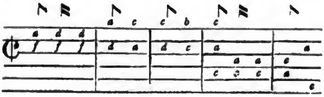

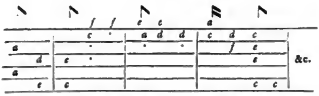

which being translated into ordinary notation would read :—

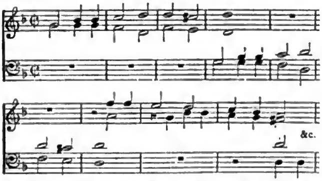

It is a remarkable fact that while in music the ordinary notation bars were either entirely omitted, or used as guides to the eye rather than as exact divisions of the notes into measurable bars, in Tablature the bars were more often properly divided according to the value of the time signature than not. An earlier publication than that of Adrien le Roy is the *Lautten Buch* of Wolf. Heckel, a writer concerning whom nothing is known, beyond the fact of the production of this book, copies of which are extremely rare; one is preserved among the treasures of the Sacred Harmonic Society's library. It bears date Strasbourg, 1562, and its title runs thus, "Lautten Buch, von mancherley schönen und lieblichen stucken, mit zweyten lautten zusammen zuschlagen, und auch sonst das mehrer theyl allein für sich selbst. Gute Teutsche, Lateinsche, Frantzösische, Italienische Stuck oder lieder. Auch vilfaltige Newe Tentz, sampt mancherley Fantaseyen, Recercari, Pavani, Saltarelli, und Gassenhawer, &c." The book is an oblong quarto, printed in a bold type, and the Tablature is entirely different to that employed by the English, French, or Italian writers for the lute, as the following example will show, when compared with the other specimens.

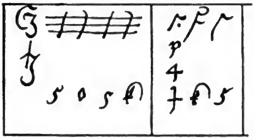

The above cut shows the first two bars of the Tablature of the dance tune given under "Gassenhauer."

The Italian and Spanish writers employed figures instead of letters to indicate the particular fret to be used, as in "Il Fronimo, Dialogo sopra l'arte del ben intavolare e rettamente suonare la Musica," Venezia 1583, by Vincentio Galilei—the father of the famous mathematician. In the "Nobilta di Dame," Venice, 1605, the tunes are set in "l'Intavolatura del Liuto" as well as in the ordinary notation. The Italian Tablature was used in England for MS. lute music, as shown in the subjoined specimen from a book written about the end of the sixteenth or beginning of the seventeenth century.

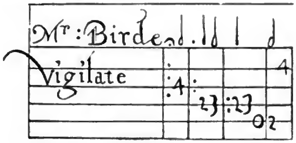

Guitar music was written in "an Alphabet.'

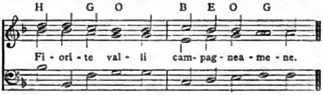

The above specimen is taken from Kapsperger's "Libro Primo di Villanelle à 1, 2 et 3 voci accommodate per qual si vogliæ strumento con l'intavolatura del Chitarone et *alfabeto* per la Chitarra Spagnuola," Rome, 1610. The Tablature for the arch-lute (Chitarone) is after the ordinary Italian manner, while the alphabet for the guitar is placed over the usual notation, implying, however, the employment of a single note at a time only, and answering the purpose of figures for fingering, as in elementary pianoforte music of the present day, unless the letters refer to a system of chords. The precise value of this "Alfabeto" is not now clear.

In England Tablature was employed for all stringed instruments, and as those most in use at the end of the sixteenth century were those plucked by the fingers, the greater number of the early specimens existing refer to this class of instrument. The following portion of an "Almaine for the Lute".

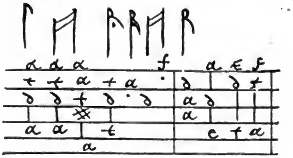

Digitized by Google

is to be found in "A new Booke of Tabliture, containing sundry easie and familiar Instructions, shewing howe to attaine to the knowledge to guide and dispose thy hand to play on sundrie Instruments, as the Lute, Orpharion, and Bandora: Together with divers new Lessons to each of these Instruments. Whereunto is added an Introduction to Pricke-Song, and certain familliar rules of Descant, with other necessary Tables plainely shewing the true vse of the Scale or Gamut, and also how to set any Lesson higher or lower at your pleasure. Collected together out of the best Authors professing the practise of these Instruments." London, 1596.

A like form of Tablature was employed for viols, as in John Dowland's second book of Songs or Ayres, printed in 1600, in which is a lesson in this peculiar notation for the bass viol, entitled "Dowland's Adeu for Master Oliver Cromwell;" and the "Schoole of Musicke," by Thomas Robinson, lutenist, has a song for the viol in Tablature.

The number of lines employed in Tablature was regulated by the number of strings the instrument possessed. The following portion of a "Pavin" is for the cittern, which had four strings:—

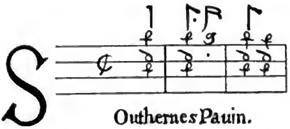

From Philip Rossetor's "Lessons for Consort," London, 1609.

To show a little variation in the Tablature in use in these and similar books, the opening notes of a "Coranto," from "Musick's Recreation on the Lyra Viol. Being a choice collection of New and Excellent Lessons for the Lyra Viol, both easie and delightfull for yong Practitioners," c. 1656; are here added:—

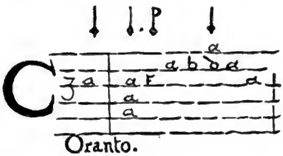

So late as the year 1682 Playford published his "Apollo's Banquet, containing instructions and variety of new tunes Ayres and Jiggs, for the treble violin (fretted)," in which Tablature is employed, with the compass of the four strings thus set forth—the upper line indicating the frets to be used:—

The First or Treble.

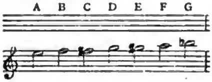

The Second or Small Mean.

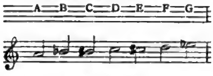

The Third or Great Mean.

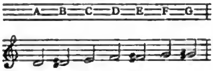

The Fourth String or Base.

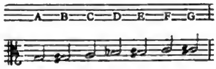

TUNING OF THE LUTE.

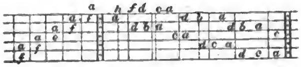

Music for viols was written sometimes in the ordinary notation, or *gamut-way;* but lute music was always written in Tablature or *Lyra-way*.

Tablature for wind instruments was expressed by dots on a stave of six, seven, or eight lines, according to the number of holes in the instrument, the number of dots signifying the number of holes to be stopped by the fingers. This method, employed for the German flute, hautboy and flageolet, is shown in the following example, taken from "The Pleasant Companion; or New Lessons and Instructions for the Flagelet," by Thomas Greeting, gent. London, 1680.

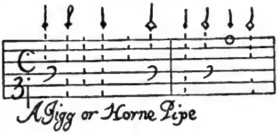

Digitized by Google

To use the author's own words:—"All tunes or lessons for the Flagelet are prick'd upon six lines, answering to the six holes in the instrument, by certain characters called dots. These dots direct what holes are to be stopt, there being so many, and the same holes to be stopt on the Flagelet each breath, as there are dots placed perpendicularly on the six lines." The time was marked in a manner similar to that in lute Tablature.

Another system of Tablature will be best explained by the following diagram (altered from that given by Koch in his Lexicon) which shows the sounds, proceeding by semitones, by the side of their letter signs. It will be seen that the strings are called by the old names fifth-string or quint, small fifth-string, great fifth-string, small prummer,* middle prummer, great prummer. Capital letters are reserved for the deepest string, but the alphabetical system is carried out by the small letters beginning from the a of the mittelprummer working upwards to the letter e of the quint; then from the mittelprummer f up to k, and so on.

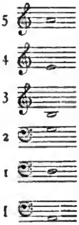

5	Quintsait	e (b♭)	k (b♮)	p (c)	v (c♯)	9 (d)	ee (d♯)	kk (e)	ll (f)
4	Clainsancksait	d (f)	i (f♯)	o (g)	t (g♯)	& (a)	dd (b♭)	ii (b♮)	
3	Grossancksait	c (c)	h (c♯)	n (d)	s (d♯)	z (e)	cc (f)	hh (f♯)	
2	Clainprummer	b (g♯)	g (a)	m (b♭)	r (b♮)	y (c)	bb (c♯)	gg (d)	
1	Mittelprummer	a (d♯)	f (e)	l (f)	q (f♯)	x (g)	aa (g♯)	ff (a)	
I	Grossprummer	A (b♭)	F (b♮)	L (c)	Q (c♯)	X (d)	Aa (d♯)	Ff (e)	

Organ Tablature was a system of writing the notes without the stave by means of letters. Thus the several octaves were called great, little, one and two-line octaves, according to the style of letter employed to indicate them.

Fig. 1.

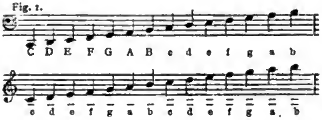

This system is liable to be confused with another of a similar character, which was adopted at the commencement of the sixteenth century, for organs which only went down to F.

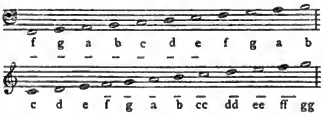

In these systems sharpened notes were indicated by a downward curved line at the side, a‿, c‿ (a♯, c♯) and flattened notes by an upward curved line a′, B′, c′ (a♭, B♭, c♭).

The duration of notes and rests was expressed by the following signs—

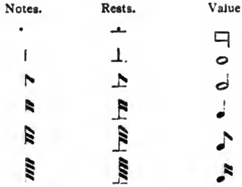

When several notes were employed together, the tails were united thus: or when rapidly written as they appear in the fac-simile from Wolf. Heckel's *Lautten Buch*, given above.

Figured bass has also been called Tablature.

The latest publication in which this peculiar form of notation was employed was in "The New Flute Master," printed in 1704. The dot notation for wind instruments survived the Tablature for the lute from which it was derived by a few years only. With the disuse of the lute, arising from the trouble and expense of keeping it in order, the employment of Tablature faded away, and as it was found that the common notation was equally available for all kinds of voices and instruments, the employment of a special notation for individual instruments was completely dispensed with.

---

* The word *Prummer* or *Brummer* means literally a *growler* or *grumbler*, and its use with reference to deep-toned strings is still kept up in this country by rustics, who call the violoncello a *grumbo*, as the Germans call it *brumm*

Tabl el musahhir. Also called the Baz, a drum used by Egyptian criers during the Ramadan or annual fast, to accompany the religious and congratulatory sentences, uttered before the houses of the wealthy. The Baz or Tabl is also employed by the Dervishes in their religious dances called zikrs.

Tabl Shamee. An Egyptian drum, suspended from the neck and beaten with two small sticks.

Table d'harmonie (*Fr.*) (1) A table or diagram of chords, &c. (2) A sounding-board.

Table d'instrument (*Fr.*) The belly of an instrument of the violin or guitar classes. The upper plate.

Table music. (1) Compositions intended to be sung by several persons sitting at a table. Many of the early printed music books of madrigals, psalm tunes, &c., had the parts so arranged on one page that two or more persons sitting opposite each other at a table could sing from the same book. The following duet, by Dr. Rogers, is given that readers may try for themselves how far this is a convenient system:

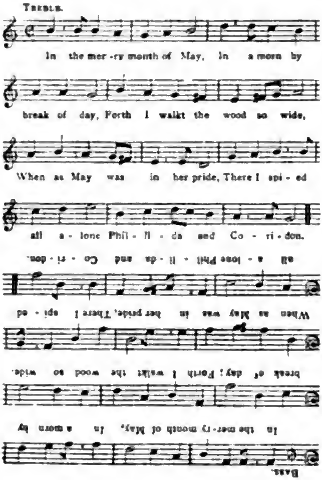
TREBLE.
In the mer-ry month of May, In a morn by
break of day, Forth I walkt the wood so wide,
When as May was in her pride, There I spi-ed
all a-lone Phil-li-da and Co-ri-don.
Bass.

The ingenuity of composers was sometimes exercised to produce pieces apparently in one part but really in two, like the subjoined, which may be performed by two players sitting opposite to each other, each beginning at the top of the page relatively to him:

Allegro Moderato. Duet for two violins, composed by Nicolo Mestrino, 1720-1790.

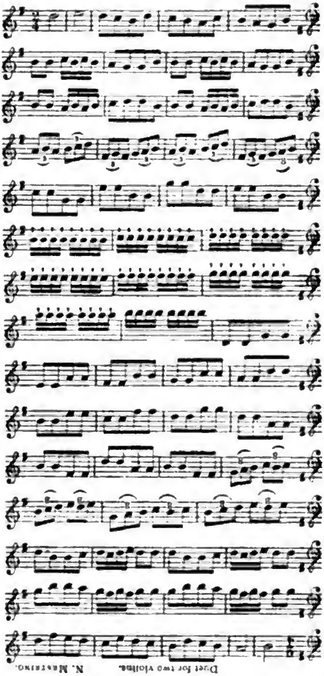

(2) German part-songs (from the word *Liedertafel*).

Tabor. This instrument, under the name *toph* (*Arabian*, *aduf*), is several times mentioned in the Bible. It is probable that it only differed from the tambourine by being without jingles in the hoop. It is often associated with a word which some translators give as *pipe*, but which in the authorized version of the Bible is rendered *dancing* or *dances*.

The old English tabor was hung round the neck and beaten with a stick held in the right hand, while the left hand was occupied in fingering a pipe. The pipe and tabor were the ordinary accompaniment of the morris-dance.

Taboret, Tabouret, Tabourin (*Fr.*) Tabor.

Tabret. [Tabor.] [Tambourine.]

Tacet (*Lat.*), **Tace** (*It.*), **Taciasi** (*It.*) Be silent. *C.B. tacet*, let the contrabasso be silent, a direction that the violoncelli only are to play the bass part.

Tact (*Ger.*) [Takt.]

Tactus (*Lat.*) The stroke of the hand or *bâton* in beating time. In mediæval music the time stroke was called *tactus major* when the time consisted of a breve in a bar, and *tactus minor* when it was a semibreve.

Tafelmusik (*Ger.*) [Table music.]

Tail. [Stem.]

Taille (*Fr.*) (1) The tenor voice or tenor part. (2) The tenor violin, the viola.

Tail-piece. That part of an instrument of the violin kind to which the strings are fastened at the lower end. The tail piece is usually of ebony.

Takt (*Ger.*) Time, measure, bar; as, *Taktart*, the sort of time, whether duple or triple; *Taktfest*, steady in keeping time: *Taktführer*, conductor; *Takthalten*, to keep time; *Taktmässig*, according to the time; *Taktmesser*, a metronome; *Taktnote*, a semibreve; *Taktpause*, a bar-rest; *Taktschlagen*, to beat time; *Taktstock*, a *bâton*, or stick for beating time; *Taktstrich*, a bar-line; *Takttheil*, a division of time; *Taktzeichen*, a time-sign.

Takigoti, *or* **Takigoto.** A species of dulcimer in use among the Japanese, provided with moveable bridges, to alter and regulate the pitch. It is played with the fingers and with plectra.

Talabalacco. A Moorish drum.

Talon (*Fr.*) The heel of a bow. [Bow.]

Tambour (*Fr.*) Drum, the great drum.

Tamboura. An instrument of the guitar species, with strings of wire struck with a plectrum. The neck is long, and the body, of gourd-shape, is often beautifully ornamented. The Tamboura is found in Persia, Turkey, Egypt, and Hindustan, and it was known to the Assyrians and Egyptians under various names. The Egyptians called it *nofre* or *nefer*, a term said to be synonymous with *nebel*, the Hebrew word for a stringed instrument. [Guitar.]

Tambour de Basque (*Fr.*) A tabor with jingles, a tambourine. [Tambourine.]

Tambourin (*Fr.*) (1) A stage dance formerly popular in France. It was of a lively measure, and accompanied with a pedal bass in imitation of the drone caused by rubbing the thumb over the skin of a tambourine. (2) A movement in a suite, of which the following is a specimen:

TAMBOURIN D'ALCIMADURE.

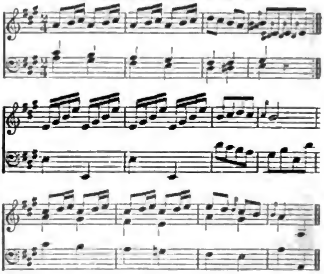

From "Premier Recueil de Menuets, Allemandes, &c., Entremelés D'airs agréables à Chanter, avec leurs accompagnements, arrangés exprès pour Le Cythre ou Guitthare Allem^{de}. Par Mr. L'Abbé Carpentier." Paris, c. 1760.

Tambourine, Tambour de Basque, Timbrel. An ancient pulsatile instrument of the drum class, popular among all European people, but particularly those of the South. The Biscayan and Italian peasantry employed it on every festal occasion. It is formed of a hoop of wood, sometimes of metal, over which is stretched a piece of parchment or skin; the sides of the hoop are pierced with holes, in which are inserted pieces of metal in pairs, called jingles. Small bells are sometimes fastened on to the outer edge of the hoop. It is sounded by being struck with the knuckles, or by drawing the fingers or thumb over the skin, which produces what is called "the roll," a peculiar drone mingled with the jingle of the bells or pieces of metal. In a book of instruction for the Tambourine the names of the several effects and the manner of writing for them, and of producing them, are thus set forth:—

"'Flamps' are made with the knuckles near the centre of the *skin* of the instrument; they are thus indicated:

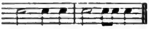

'Semi-flamps' are struck nearer the *rim;* you make them where you see the music written thus:

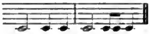

To make the 'Travale'

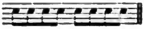

Digitized by Google

draw your wetted thumb in a circular direction over the skin. The 'double-travale'

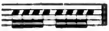

is twice as quick. Use the 'jingles' where the music is marked thus:

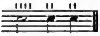

and the 'Roll' when the tails of the notes are waved:

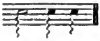

The 'Roll' is performed by shaking the instrument.

"N.B.—There are no sharps or flats in tambourine playing."

Tambourineur (*Fr.*) Drummer, tambourine player.

Tambour major (*Fr.*) Drum major.

Tamburino (*It.*) A drummer.

Tamburo (*It.*) A drum.

Tamburone (*It.*) The great drum.

Tam-tam (*Fr.*) [Gong.]

Tändelnd (*Ger.*) In a playful style.

Tangente (*Ger.*) The striking pin of a clarichord.

Tanto (*It.*) So much, as much; *Allegro non tanto*, not so fast; *a tanto possibile*, as much as possible.

Tanz (*Ger.*) A dance. *Tanzkunst*, the art of dancing.

Tarantella (*It.*) A rapid Neapolitan dance in triplets, so called because it was popularly thought to be a remedy against the supposed poisonous bite of the tarantula spider. Others say that the bite of the spider causes this delirious dance, which ends fatally.

TARANTELLA, 1654.

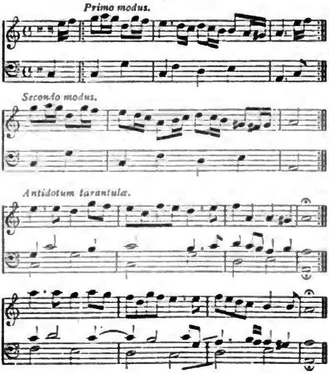

Tardamente (*It.*) Slowly.

Tardando (*It.*) Retarding the time.

Tardo (*It.*) Slow, dragging.

Tartini's tones. [Acoustics, § 19.]

Tastame (*It.*), **Tastatur, Tastenbrett** (*Ger.*), **Tastatura** (*It.*), **Tastiera** (*It.*), **Tastenleiter** (*Ger.*) The key-board of a pianoforte or organ. A hand-guide, *guidemain*. *Tastenschwanz* (*Ger.*), the extremity of the keys. *Tastenwerk* (*Ger.*), a keyed instrument.

Tasto (*It.*) (1) A key of a pianoforte. (2) The touch of a pianoforte or organ.

Tasto solo (*It.*) A direction to play the part without accompanying chords; in unison, or octaves.

Tatto (*It.*) The touch.

Tattoo, *or* **Taptoo.** The beat of the drum at night to call soldiers to their quarters; the morning beat is called Reveillé or Revelly.

Tche *or* **Tsang.** A Chinese instrument strung with wire, tuned by means of pegs and moveable bridges, and played with the fingers like a guitar.

Teatro (*It.*) A theatre.

Technik (*Ger.*) A general name for the systems, devices, and resources of musical art.

Tedesca, alla (*It.*) In the German style.

Tell-tale. A moveable piece of metal or bone attached by a cord to the bellows of an organ, which gives notice to the blower or the performer of the quantity of wind they contain. [Organ.]

Tema (*It.*) A theme, or subject; a melody.

Temperament. The division of the octave. After discussing this, an appendix on the calculation of intervals and beats shall be given.

Nature of Problem.

Notation required.

The problem is, to divide the octave into a number of intervals such, that the notes which separate them shall be suitable in number and arrangement for the purposes of practical harmony. We must provide ourselves with a notation and means of expression suited to the subject before we can discuss the treatment it has received.

Intervals expressed in eq. temp. semitones.

The simplest form of temperament is that in common use, which divides the octave into twelve equal semitones. It is most convenient to express all intervals in terms of these semitones. The perfect fifth contains 7.019550008654 semitones; the perfect third 3.863137138649 semitones. Five places are enough for all practical purposes.

Series of fifths.

In all harmonious music the fifth to any given note may be required at any time. Hence all systems provide series of fifths of a more or less complete character. We shall found on this remark a notation suitable for general discussion.

If a series of perfect fifths be constructed starting from *c*, octaves being disregarded, it

Digitized by Google

will not return exactly to *c* again. Taking the fifth to be 7.01955 semitones, each note in the series lies further than the last from the equal temperament note of the same name; and this *departure* increases by .01955 for each step in the series, the equal temperament fifth being seven semitones. Thus the *c* to which we return after twelve fifths is higher than that from which we started by 12 × .01955 = .23460 of a semitone. This interval is called the comma of Pythagoras.

Notation for series of fifths.

We employ the following notation to represent this *departure*. We take the series of fifths,
f♯—*c*♯—*a*♭—*e*♭—*b*♭—*f*—*c*—*g*—*d*—*a*—*e*—*b*, for a standard series. On passing to the next *f*♯ we denote that it is higher than the first one by prefixing to it a mark of elevation (/*f*♯); thus *b*—/*f*♯ is a perfect fifth. We proceed to form another series like the first, to all of which the mark (/) is prefixed; so that /*c* is the note to which we return on completing the circle of twelve fifths from *c*. We may extend this indefinitely. Thus we have a succession of notes, as *c*—/*c*—//*c*—///*c* . . . , each additional (/) representing the change of pitch caused by rising twelve fifths in the series, *i.e.* the Pythagorean comma. Similarly we may extend the series in the other direction: thus the fifth below *f*♯ is *b*—*f*♯, where (\) is a mark of depression, and we have such notes as *c*—\c . . . . , each (\) representing the depression of a Pythagorean comma caused by traversing a circle of twelve-fifths downwards in the series. In these series such distinctions as that between *a*♯ and *b*♭ are not observed. The place in the series is entirely determined by the prefix. The use of this notation may be extended to fifths which are not perfect.

Thirds formed by eight fifths down.

Theorem. If from any note (*c*) eight perfect fifths be tuned downwards, a note (*e*) is determined which forms an approximately perfect third to (*c*). This theorem is the foundation of Helmholtz's system. For if we tune *c*—*f*—*b*♭ *e*♭—*a*♭—*d*♭—*g*♭—*b*—*e*, we depress the pitch 8 × .01955 = .15640 below equal temperament. But the perfect third is 3.86314 semitones, *i.e.*, .13686 below the third of equal temperament, which differs from the preceding value of (*e*) by about $\frac{1}{50}$ of a semitone.

In the practical application of this, it is generally attempted to distribute the error over the eight fifths; but for all ordinary approximate purposes it does not matter where the error lies. We note here that the notes *b*, *e*, *a*, *d* have their major thirds in the same series, as *d*—*f*♯; all other notes in the series below, as *c*—*e*.

The following definitions will now be received without difficulty:—

Definitions.

Regular System.

Regular systems consist of notes which form a continuous series of fifths, *e.g.* the system of perfect fifths.

Regular Cyclical System.

Regular cyclical systems consist of notes which form a continuous series of fifths, and divide the octave into a certain number of equal intervals.

Order of Systems. Primary regular cyclical systems are those in which the departure of twelve of the approximate fifths of the system from the starting point is equal to one unit of the system. Secondary systems are those in which the departure of twelve fifths from the starting-point is two units, and so on.

Positive and Negative systems.

Positive systems have fifths sharper than equal temperament fifths; *negative* systems have fifths flatter than equal temperament fifths.

The division of the octave into 53 equal intervals furnishes an important primary positive system; into 118, a secondary positive system; into 31, a primary negative system; into 50, a secondary negative system.

HISTORICAL.

Pythagorean system.

The earliest systematic division of the octave on record is known as the Pythagorean system; it consists of a series of perfect fifths. The third employed was that formed by four fifths up, which still bears the name of the Pythagorean or dissonant third. Thus where *c*—*g*—*d*—*a*—*e* are perfect fifths, *c*—*e* is the Pythagorean or dissonant third. The true third is said to have been discovered by Archytas.

Monochord.

Vibration of strings.

Vibration numbers.

Vibration ratios.

The Greeks were acquainted with the principle of the monochord, and with the numerical ratios obtainable from that instrument; a short account of these is necessary to our subject. We shall include vibration numbers and ratios.

The number of vibrations per second given by a string of varying length, tension and weight of string per unit of length being the same, varies inversely as the length. Thus half the string vibrates twice as fast as the whole string.

If two vibrating strings have the same tension and weight per unit of length, but lengths differing in a given ratio, they will always sound notes which include the same interval; and hence:—

If two sounds have different vibration numbers, the interval between them is always the same when the ratio of the vibration numbers is the same. Thus if the lengths are as 1:2, the vibration numbers are as 2:1, and the interval is an octave; if the lengths are as 1:3, the vibration numbers are as 3:1, and the in-

Digitized by Google

Harmonics.

terval is a twelfth, and so on. The notes thus produced by fractions of the length of a string are called harmonics. The order of a harmonic is the denominator of the corresponding fraction. The following table exhibits the vibration, ratios, and intervals of the first sixteen harmonics, the intervals being taken to three places of decimals :—

Table of Harmonics :—distance from fundamental = *n* equal temperament semitones.
order of Harmonic = denominator of string fraction ;—
= number of vibrations relatively to number of fundamental ;—
= *x*.

First Octave.		Second Octave.		Third Octave.		Fourth Octave.	
x	*n*	*x*	*n*	*x*	*n*	*x*	*n*
1 (Fundamental)	0·000	2 (Octave)	12·000	4 (Fifteenth)	24·000	8	36·000
						9	38·039
				5 (Tierce)	27·863	10 (Octave Tierce)	39·863
						11	41·513
		3 (Twelfth)	19·020	6 (8ve. Twelfth)	31·020	12	43·020
						13	44·405
				7 (Seventh Harmonic)	33·688	14	45·688
						15	46·883
2 (Octave)	12·000	4 (Fifteenth)	24·000	8	36·000	16	48·000

Mersenne.

Intervals.

Mersenne's work on Harmony was published in 1636. Amongst other things he treats particularly of intervals. The following table contains a list of the principal intervals discussed by theorists. The last two columns show the number of units of the division of the octave into 53 equal parts which represents each interval most closely, with the equivalent of this representation in equal temperament semitones. This arrangement forms a classification of the various intervals in question, which will much assist the mind in the comprehension of their relations. The system of 53 is the most important primary positive system. One or two of the intervals are from Kircher, who wrote soon after Mersenne.

Mersenne and Kircher.	Difference of.	Ratio.	Equal Temperament Semitones.	$\frac{1}{53}$rds of Octave.	$\frac{1}{53}$rds in Eq. Temp. Semitones.
Octave		2 : 1	12·00000	53	12·00000
Fifth		3 : 2	7·01955	31	7·01902
Fourth	Fifth and Octave	4 : 3	4·98045	22	4·98098
Major Third		5 : 4	3·86314	17	3·84914
Minor Third	Major Third and Fifth	6 : 5	3·15641	14	3·16988
Major Tone	Fourth and Fifth	9 : 8	2·03910	9	2·03778
Minor Tone	Major Tone and Major Third	10 : 9	1·82404	8	1·81136
				7	1·58494
Semitone Maximus	Major Tone and Minor Semitone	27 : 25	1·33237	6	1·35852
Major Semitone	Fourth and Major Third	16 : 15	1·11731	5	1·13210
Apotomè Pythagorica	Seven Fifths up, Four Octaves down	2187 : 2048	1·13685		
Semitone Medius, or Chromatic Semitone	Major Tone and Major Semitone	135 : 128	·92179	4	·90568
Pythagorean Semitone	Fourth and 2 Major Tones or 5 Fifths down and 3 Octaves up	256 : 243	·90225		
Minor Semitone Semitone Minimus Chromatic Diesis	Major and Minor Third or Minor Tone and Major Semitone	25 : 24	·70673	3	·67926
Enharmonic Diesis	Major Semitone and Minor Semitone (Chromatic Diesis)	128 : 125	·41058	2	·45284
Comma (Major)	Major and Minor Tone or 2 Major Tones and Perfect Thirds	81 : 80	·21506	1	·22642
Comma of Pythagoras	Twelve Fifths either way and 7 Octaves the reverse		·23460		
Diaschisma	Half Enharmonic Diesis		·20529		
Comma (Minor)	Enharmonic Diesis and Major Comma	10240 : 10125	·19552		
Schisma	Fraction of Comma				

Mersenne's Systema Perfectum in F.

Mersenne gives numerous systems of scales which admit of the construction of perfect concords. We will give one as an example. It is a scale of the key of F, with 18 intervals in the octave; the division is irregular, but can be represented as follows by means of a broken series of fifths:—

\\b—\f♯—\c♯—\g♯; \g—\d—\a—\e—\b;
e♭—b♭—f—c—g; /g♭—/d♭—/a♭—/e♭.

The resources of this system are very limited. We have:—

Major chords of *c—f—b♭—e♭*
——— *\e—\a—\d—\g* thirds to the above:
——— */a♭—/d♭—/g♭* to which the first given three notes are thirds.
Minor chords of—*c—f—b♭—e♭*;
——— *\e—\a—\d—\g*, thirds to the above:
——— *\\b—\f♯—\c♯*, thirds of the chords of *\g—\d—\a.*

The minor chord of *\g* gives the perfect chord of the sixth on the subdominant in the key of F, *i.e.*, (*b♭*, *\d*, *\g*), thus requiring two different keys (*g*, *\g*) to represent the second of the key of F. A key-board for the system is delineated in Mersenne, in which the double G key appears. This double second has always been a characteristic of the more intelligent attempts at systems of pure intonation. Several other key-boards for more complex systems are drawn in Mersenne.

Huyghens. **Smith.**

In a tract, "Cyclus Harmonicus," (1698) Huyghens first described correctly the properties of the division of the octave into 31 equal intervals, which was previously known to be of interest. It is the most important primary negative system. Smith's Harmonics, 1759. Three systems are principally discussed, one called equal harmony, which has very flat fifths and flat thirds; it is negative, and resembles the system of 50, as Smith points out: the mean tone or old unequal temperament, which resembles the system of 31, and has flat fifths and perfect thirds: and a system in which the thirds are just as sharp as the fifths are flat.

Woolhouse.

An important tract is Woolhouse's Essay on Musical Intervals (1835). He performs a part of his reckoning with the notation of equal temperament semitones.

In the solution of the problem of temperament, Woolhouse adopts as basis the division of the octave into 50 equal intervals; this is a secondary negative system. But instead of treating it as a regular system, he selects from it notes sufficient to form a certain limited number of scales. This treatment we regard as imperfect. He distinguishes between such notes as *c♯* and *d♭*. We shall see later that this distinction is true and sufficient in negative systems only.

Woolhouse also gives scales based on the system of 31. The same remarks apply.

De Morgan.

A paper by De Morgan (Cam. Phil. Trans. X. 129) "On the beats of imperfect consonances," contains some details on the calculation of intervals. The treatment of the problem of beats is fundamentally erroneous, as are all accounts of this subject before that of Helmholtz. De Morgan employs equal temperament semitones as the measures of intervals, and gives rules for calculation nearly identical with those used independently by the writer. See Appendix.

Herschel.

A paper by Herschel (Quarterly Journal of Science, V. 338) contains proposals for systems of temperament. The character of these will be sufficiently illustrated by one example. In the table at top of p. 348, we have substantially a Pythagorean system, with twelve notes in the octave; the break is made at *d—\a*; this fifth is a Pythagorean comma out of tune. Such a fifth cannot be used in music; and this system would exclude from use the keys of G, D, and A, both major and minor. The Pythagorean system is also defended by name.

Thompson. **Poole.**

Two important practical attempts are those of Gen. T. P. Thompson and Mr. H. W. Poole (Silliman's American Journal, Vol. xliv.) The limits of this article forbid more than a concise reference to the former. Gen. T. P. Thompson in his Enharmonic organ, arranged three key-boards, each starting from a key of the ordinary board in perfect tune, and admitting of performance in related keys by means of auxiliary notes; there were 40 notes in the octave. Looking at the material of notes from our point of view, it may be regarded as constituting a series of approximately perfect fifths extending from *\\c♯* to */c*, with the omission of two single notes *\\d* and *b*. These omissions, and the distribution between different key-boards, greatly diminished the resources of the instrument, which were, however, far greater than those of any instrument previously constructed. We cannot omit to notice Gen. Thompson's method of using the monochord. He varied the weight by which the string was stretched, as well as the length. We judge that his form of the instrument is probably the most perfect that has been constructed.

Helmholtz.

The subject of temperament owes much to Helmholtz. The principal conclusions which we shall require to borrow from him, are:—The approximate formation of a perfect third by eight fifths tuned downwards; and the theories of harmonics, and of difference tones (subharmonics or Tartini tones) so far as we require them for the calculation of beats. We must also note his definition of dissonance, now universally received by physicists; viz., the intermittent excitation of the ear by a sound. Thus consonant and dissonant properties of chords depend on the beats they furnish. Such combinations as fourths and harmonic sevenths, which give no beats, may be distinguished as *unsatisfied combinations*.

Harmonics. Beats.

The theory of harmonics;—Musical notes consist of combinations of harmonics with their fundamentals (*i.e.* every note contains 8ve, 12th, &c., in very considerable strength). The ear analyses this complex note by receiving it on a resonant instrument, which may be roughly compared to a harp. The tones of different parts of the scale are thus so far separated that no direct interference can take place between tones more than a minor third apart.

The beats of imperfect consonances arise: (1) From the interference of pairs of harmonics nearly coinciding in pitch (*e.g.* imperfect fifth, by interference of 12th of C with octave of G).

(2) From the interference of difference tones (Tartini tones) with each other or with harmonics, in pairs nearly coinciding in pitch. For Ex. see Appendix.

Difference Tones. (Tartini Tones.)

Difference tones (Tartini tones) are such that their vibration number is the difference of the vibration numbers of their primaries. Ex., if c''—d'' be a major tone (8 : 9) [nine] its difference tone is (1) [one] *i.e.* C, the lowest note of the organ key-board. If c''—d' be sounded on a clarabella or harmonic flute, the difference tone will be distinctly heard, in the equal temperament of course not exactly in tune. For the somewhat abstruse theory of difference tones we must refer to Helmholtz. The old theory, in which their origin was ascribed to beats, is no longer received by physicists.

Rule for Beats.

The number of beats per second made by two notes nearly coinciding in pitch, is the difference of the vibration numbers of the two notes.

The above principles are required for numerical calculation, see Appendix.

Distinction between (*c*♯ *d*♭) when admissible.

The remark already made, that in positive systems the distinction between say (*c*♯ *d*♭) fails, while in negative systems it is true and essential, is the generalization of an observation of Helmholtz. As this is very important we will shortly explain the reason.

Not in positive systems.

The perfect third is below the equal temperament third in pitch. Hence in positive systems, which have larger fifths than the equal temperament, a third formed by downward fifths will approximate to the true third. That is to say the thirds of these systems will be formed by eight fifths down (see Th. p. 422); according to this, the major third to *a* would be the note commonly called *d*♭. But it must be written *c*♯; so in order to be clear, we cease to recognise any distinction between these two expressions, and denote the position in the series of fifths by the notation before described.

But in negative.

But in negative systems, where the fifths are less than equal temperament fifths, fifths up depress the pitch; so that the third formed by four fifths up is the representative of the perfect third, and the distinction between *c*♯ *d*♭ becomes true and essential.

Ellis.

A paper by Mr. A. J. Ellis, F.R.S., (Proceedings of Royal Society, 1864,) contains much information and copious references on the subject of temperament.

GENERAL THEORY OF REGULAR CYCLICAL SYSTEMS.

Importance of Regular Systems.

The importance of Regular Systems arises from the symmetry subsisting between the various scales to which they give rise. No idea in the present day is more connected with progress in music than that of the similarity of all keys.

Regular Cyclical Systems.

The importance of regular *cyclical* systems arises from the infinite freedom of modulation which is possible in properly arranged systems of this class. In non-cyclical systems, modulation is apt to bring us to the end of our material.

We shall consider the theory of regular cyclical systems in its simplest form.

Seven-fifths Semitone.

Definitions. The interval formed by tuning seven fifths *up* (disregarding octaves), is called a seven-fifths semitone, and it lies above the starting point.

Five-fifths Semitone.

The interval formed by tuning five fifths *down* is called a five-fifths semitone, and lies above the starting point.

The seven-fifths semitone is the Apotomè Pythagorica, (Table of Intervals) when the fifths are perfect. The five-fifths semitone is the Pythagorean semitone when the fifths are perfect.

Theorem on Semitones in Octave.

Theorem. In any regular system, five seven-fifths semitones and seven five-fifths semitones make up an exact octave.

For the departures from equal temperament of five seven-fifths semitones are due to thirty-five fifths taken upwards, and the departures of seven five-fifths semitones are due to thirty-five fifths taken downwards; so that the departures destroy one another, and leave the twelve semitones of the equal temperament, which give an exact octave.

Theorem on difference of the semitones.

Theorem. In positive systems of the first order (primary), the seven-fifths semitone is greater than the five-fifths semitone by one unit; in positive systems of the second order (secondary), by two units; and so on. In negative systems less.

For, in regular systems, the seven-fifths semitone has the departure from equal temperament due to seven-fifths up. (*Ex. gr.*, *c—g—d—a—e—b—ˌf♯—ˌc♯.*) The five-fifths semitone, taken downwards, has the departure due to five more fifths taken up. (*Ex. gr.*, *ˌc♯—ˌa♭—ˌe♭—ˌb♭—ˌf—ˌc.*) Hence, in positive systems, the seven-fifths semitone exceeds the five-fifths semitone by the departure of twelve fifths from the starting point (*c—ˌc*). But this is one unit in primary systems, two in secondary, and so on. (Definition.) And in negative systems (in which the fifths are less than equal temperament fifths, and twelve fifths fall short of the octave), the seven-fifths semitone is less than the five-fifths semitone by the departure of the twelve fifths, which is one unit in primary systems, two in secondary, and so on.

These two theorems permit us to construct positive or negative cyclical systems of any order. Those few alone which present some points of interest are entered in the following scheme:—

Scheme of Regular Cyclical Systems.

Seven-fifths Semitone, = s units.	Five-fifths Semitone, = f units.	Octave (Th. p. 425), = $5s + 7f$.
	PRIMARY POSITIVE.	
2	1	17
3	2	29
4	3	41
5	4	53
6	5	65
	SECONDARY POSITIVE.	
11	9	118
	PRIMARY NEGATIVE.	
1	2	19
2	3	31
3	4	43
	SECONDARY NEGATIVE.	
3	5	50

PROPERTIES OF INTERVALS OF THE ABOVE SYSTEMS.

Departure, means displacement from equal temperament. *Error*, from the perfect interval.

Units in Octave, = n.	Value of Unit in Semitones, $= \frac{12}{n}$.	Departure of 12-fifths.	Error of Single Fifth.	Departure of Thirds.	Error of Thirds.
17	·70588	·70588	·03927 Sharp.	·47059	·33373 Flat.
29	·41379	·41379	·01493 Sharp.	·27586	·13900 Flat.
41	·29268	·29268	·00484 Sharp.	·19512	·05826 Flat.
53	·22642	·22642	·00068 Flat.	·15640	·01954 Flat.
65	·18462	·18462	·00416 Flat.	·12308	·01378 Sharp.
118	·10169	·203388	·00260 Flat.	·13559	·00127 Sharp.
19	·63158	·63158	·07218 Flat.	·21053	·07367 Flat.
31	·38710	·38710	·05181 Flat.	·12903	·00783 Sharp.
43	·29707	·297077	·04432 Flat.	·09902	·03784 Sharp.
50	·24000	·48000	·05955 Flat.	·16000	·02314 Flat.

Selection of systems.

On inspecting the columns of errors, we at once see that the system of 118 affords the greatest combined perfection of fifths and thirds; and next to it comes the system of 53, which we prefer in practice as more manageable. The system of 31 has fifths $\frac{1}{20}$ of a semitone flat; and this is enough to be disagreeable with sharp qualities of tone, but its thirds are very good. This system forms very nearly a cyclical form of the mesotonic system, which, in an imperfect condition, was the old unequal temperament. The rule of that system was simply that all the fifths of the continuous series were made so flat as to bring down the third made by four fifths up to a perfect third.

We cannot here enter on the subject of sevenths, except to remark that the systems of 53 and 31 both afford good approximations to the harmonic seventh.

PRACTICAL EMPLOYMENT OF THE SYSTEM OF 53.

System of 53.

We will now point out shortly the arrangement which has been adopted for the practical treatment of the system of 53; it is applicable to all regular systems.

Position relations of general key-board.

A key-board is arranged according to position in a series of fifths. There are twelve vertical divisions in the octave; in the *c* division there are notes such as *‶c*—*‵c*—*c*—*′c*—*″c*; these are placed in ascending and receding order.

The vertical displacement *c*—*′c* is divided equally amongst the 12 intermediate fifths. Thus the whole tones of two-fifths each form diagonal lines, and six whole tones lead from *c′* to *′c″*. The following scheme shows the relative positions of a portion of this arrangement, with the characteristic numbers of the system of 53.

[SCHEME.]

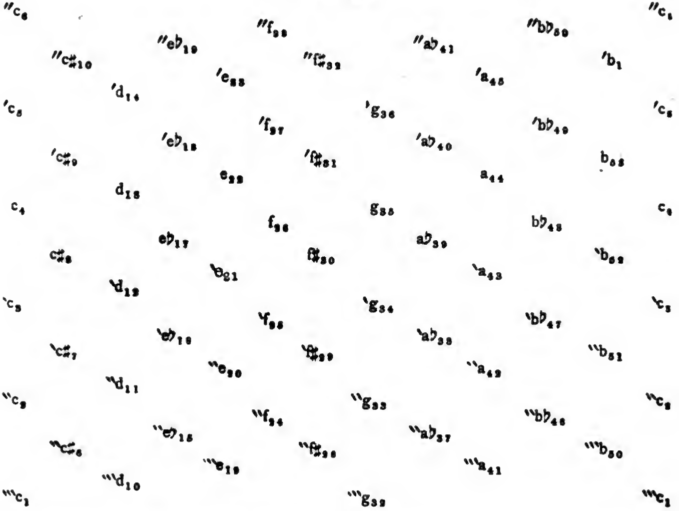
″c_{6}
″f_{28}
″$b\flat_{50}$
″c_{6}
″$e\flat_{19}$
″$a\flat_{41}$
″$c\sharp_{10}$
″$f\sharp_{32}$
′b_{1}
′e_{23}
′a_{45}
′d_{14}
′c_{5}
′g_{36}
′c_{5}
′f_{27}
′$b\flat_{49}$
′$e\flat_{18}$
′$a\flat_{40}$
′$c\sharp_{9}$
′$f\sharp_{31}$
b_{53}
e_{22}
a_{44}
d_{13}
c_{4}
g_{35}
c_{4}
f_{26}
$b\flat_{48}$
$e\flat_{17}$
$a\flat_{39}$
$c\sharp_{8}$
$f\sharp_{30}$
‵b_{52}
‵e_{21}
‵a_{43}
‵d_{12}
‵c_{3}
‵g_{34}
‵c_{3}
‵f_{25}
‵$b\flat_{47}$
‵$e\flat_{16}$
‵$a\flat_{38}$
‵$c\sharp_{7}$
‵$f\sharp_{29}$
‶b_{51}
‶e_{20}
‶a_{42}
‶d_{11}
‶c_{2}
‶g_{33}
‶c_{2}
‶f_{24}
‶$b\flat_{46}$
‶$e\flat_{15}$
‶$a\flat_{37}$
‶$c\sharp_{6}$
‶$f\sharp_{28}$
‷b_{50}
‷e_{19}
‷a_{41}
‷d_{10}
‷c_{1}
‷g_{32}
‷c_{1}

It is to be noted that, by the symmetry of the principle of arrangement, the scales on a keyboard constructed on this principle are the same in all keys, as far as form and fingering go.

According to the formation of the system of 53, we see that the seven-fifths semitones exhibit differences of 5 units, the five-fifths semitones of 4—*Ex. gr.* c_4—′$c\sharp_9$; and c_4—$c\sharp_8$.

Enharmonic harmonium.

A harmonium has been constructed with a key-board of 84 keys to the octave, the position relations of a portion of which are shown by the preceding scheme. Some notes at the top are identical with adjoining notes at the bottom, on the right. Thus infinite freedom of movement is secured, and any one of the 53 notes can be used as key-note in exactly the same manner, and with the same facility. The rule for identifications is as follows: "A note near the top of one division is identical with one near the bottom of the adjoining division on the right, when the lower division is white (♮), if the sum of the marks is 4, and when the lower division is black (♯ or ♭), if the sum of the marks is 5."

E.g.: $//c_6 = \backslash\backslash c\sharp_6.$

$/c\sharp_{10} = \backslash\backslash\backslash d_{10}.$

We have altogether four position marks in the first pair, five in the second.

A few examples are subjoined, to illustrate the practical employment of the notation.

(1)

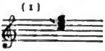

Common major chord of C.

(2)

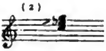

Common minor chord of C.

(3)

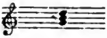

Common chord of dominant, with first form of second.

(4)

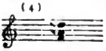

Chord of sixth on sub-dominant, with second form of second.

(5)

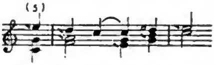

Example of the successive use of the two forms of the second; the (o) is used to nullify the preceding mark of depression (\).

(6)

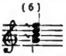

Approximate harmonic seventh.

Hence the theorem: if the harmonic seventh be used on the dominant, it must not be suspended to form a fourth with the tonic. For if we suspend the \f in the above, we obtain the fourth *c*—\f, a comma flat, which is unbearable.

(7)

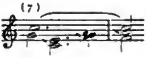

The passing note illustrates the combination formed by the division of the octave into three major thirds. The major third being $\frac{17}{53}$ of an octave, three major thirds fall $\frac{2}{53}$ short. We may distribute this error in a variety of ways. The example shows the case in which two of the thirds, *c*—\e—\g♯, are made perfect. The remaining third, \g♯—*c* is $\frac{19}{53}$, instead of $\frac{17}{53}$. Noticing that $\frac{18}{53}$ is the ordinary dissonant third, we may call \g♯—*c*, a super-dissonant third. If we employ *g*♯ as the passing note, we should have two ordinary dissonant thirds. In the equal temperament, the error of nearly half a semitone is divided equally among the three intervals. The most favourable distribution is judged by the writer to be that given in the example. Another common instance of the same combination is the following:

(8)

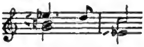

Here \b— /*e*♭ is a super-dissonant third.

The application of this arrangement to the systems of 118 and 31 presents considerable interest, but our limits preclude the discussion.

Non-cyclical regular systems.

If the above scheme be considered without the characteristic numbers of the system of 53, we have the form of application to any non-cyclical regular system. The only difference then is that, there being no identifications, continuous modulation, upwards or downwards, on the scheme would soon bring us to the end of the material of notes provided. This displacement takes place, in positive systems, most rapidly by modulation between major and minor keys; *e.g.*:

(9)

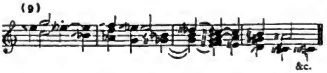

The *a*♭ at the beginning of the second bar, and /*d* in the last, are derived from inversions of chords into which harmonic sevenths enter.

Defects of equal temperament.

The defects of the equal temperament may be estimated physically by means of Helmholtz's definition of dissonance. Beats begin to be disagreeable when their number exceeds 3 or 4 in the second, attain the maximum of dissonance when about 33 in the second, and remain audible as beats up to considerably beyond 100 per second, when produced by interference of tones of sufficient strength. The application of this remark to the results calculated in the Appendix serves for the estimation of the dissonances of the equal temperament. An ear accustomed to pure chords recognizes these dissonances immediately, even in the organ and in the full-toned modern pianoforte; but especially in the harmonium.

In cases where difference tones are strongly formed, as with wide-scaled organ-pipes, or with two treble voices singing duets, errors in small intervals displace the difference tone by many times their own amount; and to educated ears this displacement, though not always involving beats, is very offensive. Helmholtz, who has written much on this subject, regards this as the most offensive characteristic of the equal temperament. *E.g.*: The difference tone of *e′—g′* should be C (8 ft.). In the equal temperament it bears to that note the ratio $\frac{21}{22}$ very nearly, which is roughly about $\frac{3}{4}$ of a semitone.

Melodic sequences.

The question of melodic progressions, as affecting the excellence of temperaments, is too extensive for our limits. We believe that it is a matter entirely of custom and education, and that the ear can accustom itself to any melodic sequences whatever.

The writer has accustomed his ear to such sequences as the following, which are new in music, whatever may be the opinion as to their effect:—

(10)

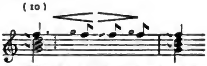

(The o nullifies the elevation and depression marks.)

The following example illustrates an instance in which a direct depression of pitch by a comma (\\) has proved to have a good effect.

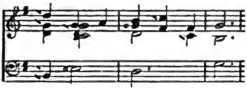

The second chord contains a depression of the tonic *g*. It may be regarded as derived from a passing dominant harmony on *a*, containing \g as harmonic seventh.

APPENDIX.

On the Calculation of Intervals, and of Beats.

(1) To find the equivalent of a given vibration ratio in equal temperament semitones.

Rule.—Take the common logarithm of the ratio; subtract $\frac{1}{300}$ of itself, and call this the first improved value. From the original logarithm subtract $\frac{1}{300}$ of the first improved value, and $\frac{1}{10,000}$ of the first improved value. Multiply the remainder by 40. The result is the required equivalent. If we take logs. to seven places, five will generally be correct in the result.

Example.—To find the equivalent of a perfect fifth, whose ratio is $\frac{3}{2}$.

$$\begin{array}{rlr}
\log.\ 3 = & \cdot4771213 & \log.\ \tfrac{3}{2} = \cdot1760913 \\
\log.\ 2 = & \cdot3010300 & \cdot0005850 \\
\hline
\log.\ \tfrac{3}{2} = & \cdot1760913 & \cdot1755063 \\
\tfrac{1}{300} = & \cdot0005870 & 1755 \\
\hline
\text{First improved value.}\ \} & \cdot1755043 & \cdot17548875 \\
 & & 40 \\
\hline
\tfrac{1}{300} & \cdot0005850 & 7\cdot01955|00
\end{array}$$

In this case the whole seven places are correct, but this is accidental.

(2) To find the vibrations ratio of an interval given in equal temperament semitones.

To the given number add $\frac{1}{300}$ of itself and $\frac{1}{10,000}$ of itself; divide by 40. The result is the logarithm of the ratio required. We must not take more than six places; then as many as we take will be correct in the answer.

Example.—To find the vibrations ratio of the equal temperament third, consisting of 4 semitones.

$$\begin{array}{rl}
 & 4\cdot000000 \\
\tfrac{1}{300} = & \cdot013333 \\
\tfrac{1}{10,000} = & \cdot000400 \\
\hline
40\ ) & 4\cdot013733 \\
\hline
 & \cdot1003433 = \log.\ 1\cdot259921
\end{array}$$

The ratio of the perfect third is 1·250000, so that the ratio of the equal temperament third to the perfect third is very nearly $\frac{126}{125}$.

The calculation of beats is essential to the accurate construction of systems of tuning. The principles to be applied have been already enunciated.

Ex. 1.—To determine the number of beats per second in the equal temperament fifth *c′—g′* (*c′* = 256).

The twelfth of *c′* is *g″*=768. This interferes with the octave of *g′*; and this is ·0195500 flat, since a perfect fifth contains 7·0195500 semitones.

To find the vibration number of the note ·0195500 below *g″* = 768.

Proceeding by Rule 2 we have—

$$\begin{array}{rl}
 & \cdot0195500 \\
 & \cdot0000652 = \tfrac{1}{300} \\
 & 19 = \tfrac{1}{10,000} \\
\hline
40\ ) & \cdot0196171 \\
 & \cdot0004904 = \text{log. ratio of note to } g''
\end{array}$$

$$\begin{array}{rl}
\text{Again—log. } 768 = & 2\cdot8853613 \\
\text{log. ratio} = & \cdot0004904 \\
\hline
\text{log. tempered } g'' = & 2\cdot8848709 \\
 = & \text{log. } 767\cdot133
\end{array}$$

The number of beats per second is the difference of the vibration numbers.

$$\begin{array}{r} 768{\cdot}000 \\ 767{\cdot}133 \\ \hline {\cdot}867 \end{array} = \text{no. of beats per second.}$$

$$\begin{array}{r} {\cdot}867 \\ 60 \\ \hline 52{\cdot}02 \end{array} = \text{no. of beats per minute.}$$

Ex. 2.—To determine the number of beats per second in the equal temperament third c'—e', caused by the interference of the tierce e' with the double octave of c'.

$$\begin{array}{lr} \text{From p. 25—log. ratio } \frac{e'}{c'} & = {\cdot}1003433 \\ \text{log. } c''' = \text{log. } 1024 = & 3{\cdot}0103000 \\ \hline \text{log. tempered } e''' = & 3{\cdot}1106433 \\ & = \text{log. } 1290{\cdot}16 \\ \text{and } e''' = 5 \times 256 = & 1280{\cdot}00 \\ \hline & 10{\cdot}16 \end{array}$$

or 10 per second very nearly.

These beats are, then, rapid enough to be offensive in the higher parts of the scale, where the notes used contain the harmonics of the 4th and 5th order in sufficient strength.

Beats of difference tones (Tartini Tones).

Ex.—To find the number of beats per second produced by the difference tones of c'—e', e'—g', in the equal temperament triad.

From the above examples we have—

$$c' = 256 \quad g', = \frac{767{\cdot}133}{2} = 383{\cdot}566$$

$$e' = \frac{1290{\cdot}16}{4} = 322{\cdot}54$$

$$\text{whence} \quad \begin{array}{r} 322{\cdot}54 \\ 256{\cdot} \\ \hline \end{array} \quad \begin{array}{r} 383{\cdot}56 \\ 322{\cdot}54 \\ \hline \end{array}$$

$$\text{diff. of } e'—c' = 66{\cdot}54 \quad 61{\cdot}02 = \text{diff. of } g'—e'$$

These are the vibration numbers of the two difference tones; and we have—

$$\begin{array}{r} 66{\cdot}54 \\ 61{\cdot}02 \\ \hline 5{\cdot}52 \end{array}$$

or 5½ beats per second nearly.

Application to the construction of systems.

For this purpose it is necessary to know the absolute pitch, as all the numbers in question vary with it. The following method can be carried out with a few organ pipes or harmonium reeds :—

RULE.—To the note (c) to be determined tune a perfect fifth (c—g), and to this (g) a perfect fourth down (g—d). Then (c—d) is a major tone.

Interpolate between these two notes so many others that the beats between each pair are just slow enough to admit of counting. Count the number of beats given by each of the pairs in two minutes: add, and divide the sum by 15. The result is the number of vibrations of (c) in one second.

Ex.—A harmonium reed sounded a note near tenor c; d was made a major tone to it as above: four notes (I., II., III., IV.) were arranged so as to give countable beats. The numbers observed in a quarter of a minute were:—

$$\begin{array}{r} c\text{— I.} = 58 \\ \text{I.— II.} = 44{\cdot}5 \\ \text{II.—III.} = 70 \\ \text{III.—IV.} = 49 \\ \text{IV.— } d = 40 \\ \hline 261{\cdot}5 \end{array}$$

Multiply by 8 we have the number for two minutes = 2092; and dividing by 15 we get 139·5 for the vibration number of the given note.

When we have the vibration number of the starting-point, we can proceed to obtain numbers of beats for guidance in tuning.

E.g.—To tune the equal temperament by fifths.

The following are the beats per minute when $c_{\prime} = 264$:—

c' —g'	53·4
$c'\sharp$—$g'\sharp$	57·
d' —a'	60·3
$e'\flat$—$b'\flat$	64·2
e' —b'	67·5
f' —c''	71·6
$f'\sharp$—$c''\sharp$	75·6
g' —d''	80.4
$a'\flat$—$e''\flat$	85·2
a' —e''	90·
$b'\flat$—f''	95·4
b' —$f''\sharp$	101·4

The methods indicated in the foregoing brief sketch are sufficient for the solution of problems in tuning. Our limits preclude further discussion on the subject.

Tempestosamente (*It.*) Impetuously, furiously.

Tempestoso (*It.*) Tempestuous, moved, agitated.

Tempête (*Fr.*) A dance invented in Paris about twenty-five years ago. The dancers are arranged as in a quadrille, in parties of four couples. Two couples stand side by side facing their respective *vis-à-vis*, so that as there are no side couples as many sets can be arranged as the ball-room will accommodate. " The step is the same as the quadrille, varied

Digitized by Google

sometimes by the introduction of the galop step, when the couples cross to each others' places, or advance to the lines of the next set." La Tempête is danced to quick music in $\frac{2}{4}$ time.

Tempo (*It.*) Time or measure. *Tempo comodo*, convenient, easy, moderate time; *Tempo di Ballo*, dance time; *Tempo di Cappella*, in the time of Church music [A Cappella]; *Tempo di Gavotta*, in the time of a gavot [Gavot]; *Tempo di Marcia*, in marching time; *Tempo di Minuetto*, in the time of a minuet [Minuet]; *Tempo di Polacca*, in the time of a polacca [Polacca]; *Tempo di prima parte*, in the time of the first part, or original movement; *Tempo di Valse*, in waltz time [Waltz]; *Tempo frettevole* or *frettoloso*, in quick, hastened, hurried time; *Tempo giusto*. [A tempo giusto]; *Tempo ordinario*, in ordinary or usual time; an ordinary walking pace, an *andante*. *Tempo perduto*, lost, interrupted, broken, and irregular time. *Tempo primo*, first or original time; a direction to resume the pace with which the movement started after an alteration. *Tempo reggiato*, regulated time. A direction to accommodate the pace to the solo performer. *Tempo rubato*, robbed or stolen time. Time occasionally slackened or hastened for the purposes of expression.

Tempo wie vorher (*Ger.*) The time as before.

Temps, *or* **Tems** (*Fr.*) (1) Time; (2) the parts or divisions of a bar. *Temps faible*, (1) weak time; (2) the unaccented parts of a bar. *Temps fort*, (1) strong time; (2) the accented parts of a bar. *Temps frappé*, the down beats in a bar [Down beat]. *Temps levé*, the up beats in a bar [Up beat].

Tempus (*Lat.*) Time, one of the three early divisions of mensurable music, which were—(1) Mode; (2) Time; (3) Prolation. *Modus* was the division of a maxim into longs, or of a long into breves; *Tempus*, the division of a breve into semibreves; *Prolatio*, the division of a semibreve into minims. Tempus was of two kinds, "perfectum" and "imperfectum;" in the former, the breve was divided into three semibreves; in the latter, into two semibreves. The sign of the former is a complete circle, O; that of the latter, an incomplete circle, C. It is from this latter sign, in all probability, we derive our mark for common time now usually written as a C.

Ten., abb. of *tenuto*.

Tendrement (*Fr.*) Tenderly, delicately.

Tenebræ (*Lat.*) An office celebrated on the afternoons of Maundy Thursday and Good Friday, and on other special days in the Roman Catholic Church, to commemorate the darkness over the earth at the time of the Crucifixion.

Tenendo il canto (*It.*) Sustaining the melody.

Teneramente (*It.*) Delicately, tenderly.

Tenerezza, con (*It.*) With tenderness, softness, delicacy.

Tenero (*It.*) Tender, soft, delicate, sensible, careful.

Tenete (*It.*) Keep, hold, sustain.

Tenor, Tenore (*It.*), **Taille** (*Fr.*) (1) The third of the four kinds of voices arranged with regard to their compass. It is the highest of male chest voices, and its extent lies between tenor C and treble A.

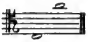

(2) The tenor voice is sometimes called by way of distinction "the human voice," from an idea that it is the quality and compass of voice most common to man. The Plain Song of the Church was formerly given as a tenor part, the harmonies being constructed above and below it; the supposed derivation of the word tenor from *teneo*, I hold, is supported by the fact that the *cantus firmus* was known and sung by the congregation or greater body of singers. Many of the hymn tunes employed up to the first quarter of the eighteenth century, as well as many of the arrangements of the responses used in the Church Service, were written with the melody in the tenor part.

(3) In old music the tenor voice was divided into three classes—*altus*, *medius*, and *bassus*—high, mean, and low tenor.

(4) The larger violin of low pitch is called the tenor, viola, bratsche, and sometimes alto violin. [Viola.]

(5) The largest bell in a peal, or set, is distinguished by the name of tenor bell.

Tenor C. (1) The lowest C in the tenor voice.

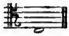

(2) The lowest string of the tenor violin.

Tenor clef. The C clef placed upon the fourth line of the stave.

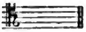

It is used for the tenor voice, tenor trombone, the higher register of the bassoon and violoncello, &c. The treble clef is sometimes employed for the tenor voice, but the notes are then expressed an octave above their true sound. The late Thomas Oliphant suggested that two treble clefs conjoined should be the sign of the tenor G clef.

Tenore (*It.*) (1) Tenor voice. (2) A tenor singer; *Tenore buffo*, a tenor singer to whom

Digitized by Google

is assigned a comic part in an opera; *Tenore leggiero*, a tenor singer with a voice of light, small quality; *Tenore robusto*, a tenor singer with a full, strong, sonorous voice.

Tenorist. One who sings the tenor part, or plays the tenor violin.

Tenoroon. (1) The name of an old tenor oboe with a compass extending downwards to tenor C. (2) A word affixed to an organ stop to denote that it does not proceed below tenor C, as, *tenoroon hautboy*. A *tenoroon diapason* is a double diapason which does not extend below tenor C.

Tenorposaune (*Ger.*) Tenor trombone. [Trombone.]

Tenorschlüssel (*Ger.*) Tenor clef.

Tenor trombone. A trombone with a compass of two octaves and a fifth 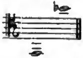 for the C trombone and 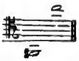 for the B flat trombone. [Trombone.]

Tenor violin. The viola. [Viola.]

Tenorzeichen (*Ger.*) The tenor clef. [Clef.]

Tenth. (1) A compound interval comprising an octave and a third, nine conjoint degrees, or ten sounds. The tenth is the octave of the third, and may be major or minor, diminished or augmented. [Intervals.] (2) An organ stop, tuned a tenth above the diapasons, called also *double tierce* or *decima*. [Organ.]

Tenuto (*It.*), **Tenu** (*Fr.*) Held on, sustained, kept down for the full time.

Teorbo (*It*) [Theorbo.]

Teoria (*It.*) Theory. *Teoria del Canto*, theory of singing; *Teoria d'armonia*, theory of harmony.

Tepidamente (*It.*) In a lukewarm manner, with indifference.

Tepidità (*It.*) Coldness, indifference.

Ter (*Lat.*) Thrice.

Tercet (*Fr.*) A triplet. [Triplet.]

Ternary form. Rondo form. [Form.]

Ternary measure. Triple time. Perfect time. [Signature § 2.]

Terpodion. An instrument invented by David Buschmann of Hamburg, in 1816, resembling in appearance the pianoforte, but the tone was produced from blocks of wood struck with hammers. It contained also a contrivance by which the sound might be increased or diminished at pleasure.

Tertia (*Lat.*) [Tierce.]

Tertian. An organ stop composed of two ranks of pipes, sounding a major third and fifth of the foundation pipes, in the third octave above; a *Tierce* and *Larigot* on one slider.

Ter unca (*Lat.*) Thrice hooked. A demisemiquaver.

Terz (*Ger.*) [Tierce.]

Terza (*It.*) (1) The third. *Opera terza*, the third work; *Violino terzo*, the third violin. (2) [Tierce.]

Terzdecime (*Ger.*) A thirteenth.

Terzdecimole (*Ger.*) A group of thirteen notes to be performed in the time of eight, or of twelve.

Terzettino (*It.*) A short composition for three performers.

Terzetto (*It.*) A composition for three performers.

Terzflöte (*Ger.*) (1) A flute sounding a third above the notes written. (2) An organ stop. [Tierce.]

Terzina (*It.*) A triplet.

Terzo mano (*It.*) *Lit.* a third hand. An octave coupler on Italian organs.

Terzquartakkord (*Ger.*) The $\begin{smallmatrix}4\\3\end{smallmatrix}$ or $\begin{smallmatrix}6\\4\\3\end{smallmatrix}$ chord. The second inversion of the chord of the dominant seventh.

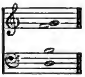

Terzquartsextakkord (*Ger.*) [Terzquartakkord.]

Terzquintsextakkord (*Ger.*) The $\begin{smallmatrix}6\\5\\3\end{smallmatrix}$ chord. The first inversion of the chord of the dominant seventh, figured $\begin{smallmatrix}6\\5\end{smallmatrix}$.

Testo (*It.*) (1) The text, theme, subject of a composition. (2) The libretto of an opera, or the words of a song.

Testudo (*Lat.*) A name for the lyre, because the sounding part or hollow was made of the shell of the sea tortoise or turtle. [Lyre.]

Tetrachord. A scale-series of *four* notes. The word in its modern sense signifies a half of the octave scale, *e.g.:*

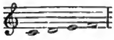

First Tetrachord.

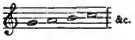

Second Tetrachord.

It will be seen that the position of the tones and semitones is similar in both tetrachords. A third tetrachord placed above these two would lead into the key of G., *e.g.:*

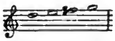

and another into the key of D,

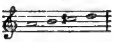

and so on through all the sharp keys. Simi-

Digitized by Google

larly, tetrachords placed below the first lead into flat keys, *e.g.*:

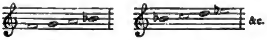

Tetrachords which overlap are said to be *conjunct;* those having a degree between them, *disjunct.*

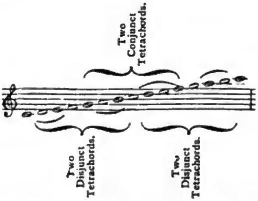
Two Conjunct Tetrachords.
Two Disjunct Tetrachords.
Two Disjunct Tetrachords.

Similar disjunct tetrachords necessarily pass through the whole key-series, and a combination of conjunct and disjunct tetrachords is required to form a diatonic scale of more than one octave in compass. [Greek music.]

Tetrachordal system. The early form of the system now known as *Tonic Sol-fa.* [Tonic Sol-fa.]

Tetrachordon, Bogenclavier (*Ger.*) An instrument similar in appearance to a cottage pianoforte, and like it played by a finger-board, but the tone, instead of being produced by striking, is obtained by means of a cylinder of india-rubber charged with rosin, kept in motion by a pedal, variety of tone being gained by the depth of pressure on the keys by the fingers. It is called the tetrachordon from an idea that its sounds are similar to those produced by a string quartet. The instrument is constructed also with self-acting machinery. [Xänorphica.] [Hurdy Gurdy.]

Theil (*Ger.*) (1) A part or division of a bar. (2) A phrase, strain, or part of a piece. (3) A piece, composition, work.

Thema (*Ger.*) [Theme.]

Thême (*Fr.*) [Theme.]

Theme. (1) One of the divisions of a subject, in the development of sonata-form. [Form.] (2) The *cantus firmus* on which counterpoint is built. (3) The subject of a fugue. (4) A simple tune on which variations are made.

Théorbe (*Fr.*) [Theorbo.]

Theorbo, Tiorba (*It.*); **Théorbe** (*Fr.*); **Basslaute** (*Ger.*) (1) An old stringed instrument resembling the lute in form or tone. It had two necks, to the longest of which the bass strings were attached. It was employed for accompanying voices, and was in great favour during the seventeenth century. Mace describes it as "no other than that which we call'd the old English lute." It differed from the lute in the possession of its two necks, from whence it is sometimes called cithara bijuga.

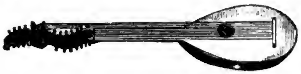

The strings were usually single in the Theorbo, and when double, or tuned in octaves or unison with the bass or treble notes, the instrument was called the arch-lute, or chittarone.

(2) Its invention has been variously assigned to an Italian, Signor Tiorba, from whom it is said to have derived its name; to one called Bardella about the year 1600; and to Hotteman, a German, living in France in 1650.

Sir John Hawkins attributes the invention to a nameless Neapolitan, who called it Tiorba from its resemblance to an instrument used for pounding perfumes so called. Johannes Kapsberger, a German of noble birth, who died about 1630, was a skilful performer on the instrument, and wrote a large quantity of music in tablature for it.

(3) The Theorbo was used as an instrument in the orchestra as late as the year 1708 by Francesco Conti. It was also employed in the performance of sonatas in the place of the cembalo, from its power of rendering a figured or thorough bass. Corelli's third set of sonatas, published in Bologna in 1690, contains a part for the Theorbo or violone.

Theoretiker (*Ger.*) A theorist.

Théoricien (*Fr.*) A theorist.

Theorist. One who studies the nature of sound or the principles of musical art.

Theory of music. The science of music. The speculations arising from a knowledge of the principles of sound. The rules for composition and arrangement of music for voices and instruments in rhythm, melody, harmony, counterpoint, and instrumentation.

Thesis (*Gk.*) The downward wave of the hand to denote the absence of accent. [Accent, § 5.] [Arsis.] [Metre.]

Thin. (1) Meagre and scanty harmony. (2) A poor quality of tone in a voice or on an instrument.

Third. [Interval.]

Third flute. [Terzflöte.]

Third stave. A name given to the stave upon which pedal music is written for the organ.

Thirteenth, chord of the. A chord called by some a suspension; by others, a secondary seventh (*see* Seventh, Exs. 34—41). It consists generally of the 3rd, 7th, and 13th of the dominant, and is used both in the major and minor modes. The following are its most common forms.

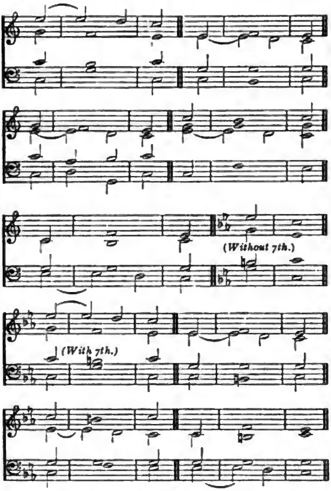

Thirty-second note. A demisemiquaver.

Thorough bass, Basso continuo (*It.*), **Basse contrainte** (*Fr.*) A species of musical shorthand, reduced to a system by Ludovico Viadani, about the year 1605, which has remained substantially unimproved since his day. It consists of a bass part with the accompanying harmonies indicated by figures. Henri Dumont was the earliest musician in France who made use of thorough bass, about 1640, and the first treatise on thorough bass, published in England, was the work of Matthew Lock, issued under the title of "Melothesia; or, Certain General Rules for Playing upon a Continued Bass, with a choice collection of Lessons for the Harpsichord and Organ of all sorts," London, 1673.

Francis de la Fond in 1725, suggested a new method of figuring a bass by calling the notes of the chromatic scale by the numbers one up to thirteen. [Figured Bass.]

Threnody. An elegy, or funeral song, from the Greek *θρηνῳδία*.

Thrice-marked-octave (*dreigestrichene Octav*). The name given in Germany to all the notes of the octave above C in alt. [Pitch.]

Thrum. (1) To play without skill upon a stringed instrument. (2) The sound so made.

Thürmer (*Ger.*) A town musician. [Waits.]

Tibia (*Lat.*) [Flute.]

Tibiæ pares (*Lat.*) Two flutes of the same pitch, which were played together by the same performer.

Tibiæ utriculariæ (*Lat.*) [Bagpipe.]

Tibia major. An organ flute-stop of 16-ft. pitch. [Flute.]

Tibicen (*Lat.*) (From *tibia* and *cano.*) A flute-player; *tibicina*, a female flautist; *tibicinium*, a piping; *tibicino*, to pipe.

Tie. (1) A curved line placed over two or more notes in the same position on the stave:

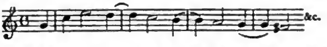

The tie is also called a *bind*, and the curved line, when used over notes representing different sounds, is called a *slur*. [Bind.] [Slur.] (2) When two or more quavers, semiquavers, &c., are united, instead of being written with separate tails, they are said to be *tied*. [Stem.] [Printing of Music.]

Tief (*Ger.*) Deep, low.

Tierce (*Fr.*) (1) A third. (2) An organ stop tuned a seventeenth above the diapason. [Organ.] (3) The service which took place at the third hour of the day, the undersang of the Anglo Saxons. [Horæ Canonicæ.] (4) The natural harmonic produced by ⅕ of a vibrating string. [Acoustics.]

Tierce de Picardie (*Fr.*) The sharpened third in the concluding chord of a composition in the minor mode.

Tierce coulée (*Fr.*) A slurred third. In old harpsichord music the *tierce coulée* was expressed by a dash through the notes of the third, in a diagonal direction, thus:

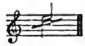

when it was called the rising slurred third, *tierce coulée en montant*, and was played:

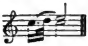

A stroke in the opposite direction

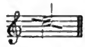

was called the falling slurred third, *tierce coulée en descendant*, and was played:

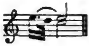

Timb. Abb. for Timbales. [Kettle-drums.]

Timbale (*Fr.*) [Kettle-drum.]

Digitized by Google

Timballo (*It.*) [Kettle-drum.]

Timbre (*Fr.*) Quality of tone or sound. *Klang* or *Klangfarbe* (*Ger.*) [Acoustics, § 16.]

Timbrel. [Tabor.] [Tambourine.]

Time. (1) The division of musical phrases into certain regulated portions measured with regard to the value of the notes with respect to the semibreve, which, in modern music, is held to be the standard of time. There are two sorts of time, duple with two, four, or eight beats in a bar, and triple with three beats in a bar. There is also compound time, or time formed of the union of triple with duple, and triple with triple, each having a distinctive time signature. [Signature, § 2.]

(2) The pace at which a movement is performed is called its *time* or *tempo*.

Timorosamente (*It.*) Timidly, hesitatingly, with fear.

Timoroso (*It.*) Timorous, with hesitation.

Timp. Abb. of Timpani.

Timpani (*It.*) Kettle-drums.

Timpanista (*It.*) A drummer.

Tintement (*Fr.*) The tinkling of a bell.

Tintinno (*It.*) [Tintement.]

Tintinnabulum (*Lat.*) A rattle (*Gk.* πλαταγή) formed either of small bells or little plates of metal.

Tiorba (*It.*) [Theorbo.]

Tipping. [Double tongueing.]

Tirade. The filling up of an interval between two notes with a run, in vocal or instrumental music, *e.g.*:

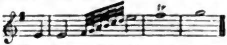

Tirante (*Sp.*) The brace of a drum.

Tirasse (*Fr.*) The pedals of an organ which act on the keys or manuals.

Tirato (*It.*) (1) A down bow. [Bowing.] (2) A scale passage in notes of equal length.

Tira tutto (*It.*) A pedal commanding the full power of the organ. (*Fr.*) *Grand jeu*.

Tiré (*Fr.*) (1) Drawn, pulled. (2) A down bow. [Bowing.] (3) The drawing out of an accordion.

Titty, tziti, toutari. An Indian bagpipe. [Bagpipe.]

Toccata (*It.*) (1) A prelude or overture. The overture to Rinuccini's opera "Orfeo," 1609, is called a *toccata*, and is directed to be performed three times, "Avanti il lever de la tela," "before the rising of the curtain." (2) Compositions written as exercises. (3) A fantasia. (4) A suite.

Toccatina (*It.*) A short toccata.

Todtenmarsch (*Ger.*) A funeral march.

Tombestere *or* **Tymbestere** (*Old Eng.*) A female dancer, who accompanied herself upon a tambourine, occasionally throwing her instrument in the air and catching it.

> "There was many a tymbestere,
> Couthe her crafte full parfytly.—*Chaucer.*

Tom-tom. (1) An Indian drum. (2) A Chinese gong.

Ton. (*Fr.* and *Ger.*) (1) Tone, sound. (2) The interval of a second.

Tonabstand (*Ger.*) An interval.

Tonada (*Sp.*) A tune, air, or melody.

Tonadilla (*Sp.*) A short tune, an interlude, ritornello, symphony to a song.

Tonal Fugue. [Fugue.]

Tonarion *or* **Tonarium.** A pitch-pipe (according to Quintilian) used by the Latin orators for the purpose of regulating the pitch of their speaking voice, called also *fistula eburneola*. Dionysius limits the compass of the oratorical voice to *five notes*. When this compass was exceeded, an attendant blew the *tonarion* to enable the orator to recover his proper pitch.

Tonart (*Ger.*) Mode, tune, key, scale-system, *tonality*.

Tonausweichung (*Ger.*) Modulation.

Ton bas (*Fr.*) A deep, low tone.

Tondichter (*Ger.*) A composer. A poet *in sounds*, as a painter may be described as a poet *in colours*. This word has been badly rendered "tone-poet."

Tondichtung (*Ger.*) A musical composition. A sound-poem.

Tone. (1) Sound. (2) Quality of tone. [Acoustics, § 16.] (3) The interval of a second. [Interval.] (4) A Gregorian chant. [Plain Song.]

Tonfall (*Ger.*) A cadence.

Tonfolge (*Ger.*) A succession of sounds. A melody.

Tonführung (*Ger.*) (1) A melodic succession. (2) [Modulation.]

Tonfuss (*Ger.*) A foot. [Metre.]

Tongang (*Ger.*) [Tonführung.]

Ton-générateur (*Fr.*) [Root.]

Tongeschlecht (*Ger.*) The character of the modes. There are two *Tongeschlechter*—major and minor. [Scale.] [Mode.]

Ton haut (*Fr.*) An acute sound.

Tonic, Tonica (*It.*), **Tonique** (*Fr.*) (1) The key-note of any scale. The ground-tone or basis of a scale or key. (2) The key-chord in which a piece is written, and with which it concludes.

Tonic Sol-fa. A letter-system of notation. Many attempts have from time to time been made to produce a simpler notation than the stave, clefs, signature, &c., of the so-called "Old Notation."

As early as the year 1672, Thomas Salmon wrote a book entitled, "An essay to the advancement of music, by casting away the perplexities of different cliffs, and uniting all sorts of music, lute, viol, violins, organ, harpsichord, voice, &c., in one universal character." In this he proposed to write all music on a stave of four lines, which should give the notes the names of the first seven letters of the alphabet, octaves above or below being marked with the name of the octave to which they belong. The proposition was violently opposed by Matthew Lock, but other musicians treated it with contempt, and Salmon's book and proposition now exist only among the literary curiosities of music. Jean Jacques Rousseau suggested a notation whereby the notes of the scale were indicated by the numbers 1 to 7. This, or rather an improved form of it, is still largely used in France. Miss Sarah A. Glover, of Norwich, about thirty years ago projected and taught successfully a system which she called the tetrachordal system, which was the Tonic Sol-fa notation in its original form; but it has since received such important modifications and additions from the hands of the Rev. J. Curwen, that it is now justly associated with his name.

Its chief *raison d'être* is that the ordinary notation, just in the degree that it accommodates itself to the keyboard and the theory founded more or less thereon, is not fitted as a notation for the greatest of all instruments, the human voice. Tonic Sol-faists nevertheless maintain that their notation may be used in all branches of the art, and pupils are taught to play instruments, to study harmony, musical form, and composition, entirely from the syllables. The leading features of the notation are as follows: Of the two relationships of musical sounds, those of pitch and key, the latter is of transcendant importance. It is argued, therefore, that it is of the first consideration that this supremely important fact should be prominently shown. The keynote of a piece is, therefore, always called *doh*, the second of the scale *ray*, and so on, *me*, *fah*, *soh*, *lah*, *te*. The reason for this departure from the ordinary spelling is, that the above is considered easier for English people to pronounce. In printing music, the initial letter indicates the scale note. *Si* and *soh* having the same initial, the former is altered to *te*. Higher or lower octaves are shown by figures placed by the side of the notes, d^1, d^2, m^2, and s_1, m_2, d_2. The first part of the National Anthem is written in tune thus, *d d r* t_1 *d r m m f m r d r d* t_1 *d*. The particular pitch of the key-sound is shown by the statement at the beginning of the piece, key G, key E♭, key A, &c. The minor mode is regarded as derived from the relative major, the tonic being called *lah*. It is maintained that to call the tonic of the minor scale *doh*, would lead to extraordinary practical difficulties, besides being false in theory. Changes of key in the course of a piece are met by what are called *bridge tones*. The note of the key quitted is placed side by side with the note of the same pitch in the key approached, and the pupil is taught to think and sing the sound of the first note, and to call it by the name of the second. Thus, *d r m f* ${}^{s}d$ t_1 *d* would show a transition, say from key C to key G. By this means changes of the most complex nature are simply represented by Tonic Sol-faists, and they assert that the music has yet to be quoted which cannot be expressed in their notation. The chromatic scale is named by adding the vowel *e* to the initial of sharpened notes, and *a* (pronounced *aw*) to notes to be flattened. Thus *de*, *re*, *fe*, *se*, are respectively *d*, *r*, *f*, *s*, sharp, and *ma*, *la*, *ta*, are *m*, *l*, *t*, flat. The sharp sixth of the minor scale is called *bah* to distinguish it from *fe* the sharp fourth of the major. Time and accent are indicated by measurement across the page thus:

| : | : | : |

the space between one sign and the next, representing the pulse or beat; the line showing the stronger beat or accent, and the colon the weaker. For short divisions a dot | . : in the centre of the pulse divides it into halves, and commas | , . , : are used to divide into quarters, and other divisions are similarly shown. A stroke — through a pulse means that a previous note is to be continued. Sol-faists believe that their fixed standard of a pulse or beat gives them considerable advantage over the ever-shifting standard of the ordinary notation. "Rule Britannia" is thus written:

.s_1 |d :d |d,r.m,f :s .d |r :r .m ,f |m .etc.

No account of the Tonic Sol-fa Notation would be complete without reference to its indispensable adjunct the Tonic Sol-fa Method, *i.e.*, the distinctive plan of teaching the musical facts indicated by the notation. This method is the outcome of years of laborious enquiry by Mr. Curwen, and of the collated experience of all the best teachers of the system. Great stress is given to the doctrine of *mental effect*, by which is meant the various impressions or colours of the notes of the scale when sung slowly. Thus *doh*, is considered firm; *te*, sharp and piercing; *lah*, sorrowful; *fah*, gloomy; *soh*, bright and clear, &c., &c. Teaching by pattern is also insisted on; the scale is taught in the following order, first the notes of the tonic chord *d m s* and their replicates, next the dominant chord *s t r*, then the sub-dominant chord *f l d*. In developing the scale, large use is made of what has been justly called the backbone of the system, the MODULATOR, without a proper use of which, it is not too much to say, the method cannot be fairly taught or learned.

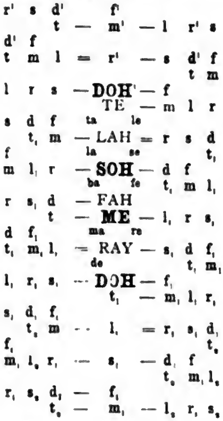

This Modulator is a sort of map of musical sounds. It represents pictorially in an upright position the relative places of the notes of a scale, its minor mode, its chromatics, and its more closely related scales. By frequent and systematic handling of this sheet, the upward or downward motion of the notes printed all on one level, is adequately realised by the pupil. The Modulator is manipulated with great effect by the best teachers of the system. In teaching Time an adaptation of the time-names used in the French Cheve system has lately been more or less adopted by Sol-faists. Syllables are used to show the length of notes just as they are to show the relation of sounds. Thus *taa* is the name of a pulse, *taa-tai* of half pulses, and *tafatefe* of quarter pulses. Continuations of any kind are met by dropping the consonant. It is thought that the finer accents of divided pulses, as well as the broader accents of the measure, can be better taught by the use of this language of duration than on any other plan.

Tonkunst (*Ger.*) The art and science of music.

Tonkünstler (*Ger.*) A musician. A musical artist.

Tonleiter (*Ger.*) A scale. [Scale.]

Ton majeur (*Fr.*) Major key or mode.

Tonmalerei (*Ger.*) Composition, invention, sound-painting.

Tonmessung (*Ger.*) Tone-measuring, metre, rhythm.

Ton mineur (*Fr.*) Minor key or mode.

Tonos (*Gk.*) τόνος. [Tone.]

Tonsatz (*Ger.*) A musical composition.

Tonschluss (*Ger.*) A cadence.

Tonschlüssel (*Ger.*) Claves signatæ (*Lat.*) Key clef. [Key, § 5.]

Tonschrift (*Ger.*) (1) Written music. (2) Musical notes or characters.

Tons de la trompette (*Fr.*) Crooks used to alter the pitch of a trumpet. [Crook.]

Tons de l'église (*Fr.*) Church modes or tones. [Plain Song.]

Tonsetzer (*Ger.*) (1) A composer. (2) (In a sarcastic sense) a music-maker.

Tonsetzkunst (*Ger.*) The art of musical composition.

Tonsetzung (*Ger.*) A musical composition.

Tonspieler (*Ger.*) A performer.

Tonsprache (*Ger.*) The art of expressing thoughts and feelings in music.

Tonstück (*Ger.*) A piece of music, a composition.

Tonstufe (*Ger.*) A step or degree of a scale.

Tonsystem (*Ger.*) (1) The systematic arrangement of sounds according to the rules of melody, harmony, and rhythm. (2) A scale.

Tonverhalt (*Ger.*) Rhythm.

Tonzeichen (*Ger.*) (1) A note or other character used in music. (2) Accent.

Toomourah. An Indian tambourine.

Tooroorec. A trumpet used by the Brahmins in their religious processions.

Toph (*Heb.*) A small hand-drum; a timbrel or tabret. [Tabor.] [Tambourine.]

Torcelli. A name anciently given to organs in Italy.

Tosto (*It.*) Quick, swift, rapid. *Più tosto*, more rapid; *Tostamente*, quickly, rapidly; *Tostissamamente*, *Tostissimo*, very quickly, with great rapidity.

Touch. (1) The resistance made to the fingers by the keys of a pianoforte or organ. Thus the touch of the keyboard may be hard or light accordingly as the resistance is great or little. (2) The peculiar manner in which a player presses the keyboard, whether light, pearly, heavy, clumsy, firm, &c.

Touches (*Fr.*) The keys of a pianoforte, organ, harmonium, and concertina.

Touquet (*Fr.*) [Toccata.] [Tucket.]

Tours de force (*Fr.*) (1) Roulades, runs, or divisions for the voice. (2) Passages of rapid execution upon an instrument.

Toys. An old English name for dance tunes and light and trifling pieces of music. "Pavens, Galiards, Almaines, Toies, Jiggs, Thumpes, and such like." Ford's "Musicke of sundre kindes," 1607.

Trachea. [Larynx.]

Trackers. [Organ.]

Tractur (*Ger.*) [Trackers.]

Tractus (*Lat.*) A melody sung in the Roman Catholic Church during Lent, instead of the Alleluia.

Tradotto (*It.*) Transposed, arranged, translated. [Arrangement.]

Trainé (*Fr.*) Slurred, bound, dragged.

Trait (*Fr.*) (1) A run, or division; *trait de chant*, a melodious vocal phrase. (2) A special passage or phrase for a body of instruments of the same class. Like the *trait des violons* in Cherubini's overture to "Anacreon," or the passage for strings in Beethoven's No. 3, "Leonora." (3) *Trait d'harmonie*, a sequence in harmony. (4) *Trait d'octave*, rule of the octave.

Traité (*Fr.*) A treatise on the theory or practice of music.

Tranquillamente (*It.*) Tranquilly, calmly, quietly.

Tranquillità, con (*It.*) With tranquillity, calmness.

Transcription. The arrangement or modification of a composition for some instrument or voice other than that for which it was originally written.

Transient Modulation. The temporary introduction of chords or progressions from an unrelated key. [Modulation.]

Transitio (*Lat.*) Change of key.

Transitus (*Lat.*) Progression by passing notes; *transitus regularis*, diatonic progression, the passing notes on the unaccented portions of the bar; *transitus irregularis*, progression in which some of the notes of the scale are omitted; passing notes on the accented part of the bar.

Transition. (1) A modulation. [Modulation.] (2) A passing note. [Passing Note.]

Transponiren (*Ger.*) Alteration of the original key. [Transposition.]

Transponirende Instrumente (*Ger.*) [Transposing Instruments.]

Transpose. To alter the key in which a piece is set, by changing it into a higher or lower scale.

Transposing Instruments. A general name for all instruments which do not produce the exact sounds written on paper for them.

Thus, a B♭ clarinet is so called because the *written* note C, when sounded, is B♭; its part, therefore, is written one note higher than the actual sounds required. Similarly, the A clarinet is so called because the *written* note C, when sounded, is A; its part, therefore, is written a minor third higher than the sounds actually required. The C clarinet is so called because it plays *as written*. The Cor Anglais and Corno di bassetto produce sounds a major fifth below those written.

A C horn produces sounds one octave below those actually written; a D horn, a minor seventh below; an E horn, a minor sixth below, and so on. A B♭ *alto* horn produces sounds a whole tone below that written; a B♭ *basso* horn, a ninth below.

Similarly, a trumpet in B♭ produces sounds one note below those actually written, and a D trumpet sounds *one note above*. Drum parts are usually written as if always in the key of C, directions being given as to the tuning, at the commencement of each movement.

The only transposing string-instrument is the double bass, which produces sounds an octave below those written, as far as its compass will permit. Flutes, oboes, bassoons, and trombones, are not transposing instruments; but the *piccolo* flute produces sounds one octave higher than those written, and the *double bassoon* sounds one octave lower.

Transposing Piano. A pianoforte so constructed that its key-board may be moved to admit of its giving sounds, other than those which the scale used would seem to imply. The key-frame is made in duplicate, and on it is screwed, by means of ordinary thumb-screws, the action, making it perfectly rigid. At each extreme end of the keys the block of wood, called the "key-block," is also attached to the key-frame, and thus rendered moveable. These blocks are pierced with holes exactly a semitone apart, in which small pegs of wood are inserted. When it is required to alter the piano to a higher or lower pitch the pegs are taken out, and the key-board can then be moved up or down the scale.

Transposition. (1) A change of key. [Transpose.] (2) An inversion of parts in counterpoint.

Trascinando (*It.*) Dragging, delaying the time.

Trascritto (*It.*) Transcribed, copied.

Trauermarsch (*Ger.*) A Funeral March.

Traversière (*Fr.*) Across. *Flûte traversière*, the flute held crossways, as is now usual; the *flûte-à-bec* being blown with a mouthpiece like the oboe. [Flute.]

Traverso (*It.*) [Traversière.]

Tre (*It.*) Three. *A tre voci*, for three voices; or, in three parts.

Treble. (1) The highest vocal or instrumental part, sung by women or boys, or played by violins, flutes, oboes, clarinets, or other instruments of acute tone. (2) The treble or soprano voice is the most flexible of all vocal registers, its ordinary compass is from middle C upwards to the extent of a twelfth, its exceptional range a fifteenth, or even beyond this. [Triplex.]

Treble Clef. The G clef on the second line of the stave, used for treble voices and instruments of high and medium pitch, such as flutes, oboes, clarinets, horns, violins, and trumpets. [Stave.] [Clef.]

Trem. Abb. of *tremando* and *tremolando*.

Tremolando (*It.*) Trembling, wavering. (1) A chord or note played or bowed with great rapidity so as to produce a quivering effect. (2) Vibration of the voice in singing, arising from nervousness, or a bad production; or used for the purpose of producing a special effect. [Vibrato.]

Tremolant, *or* **Tremulant.** An organ and harmonium stop which causes the air as it proceeds to the pipes or reeds to pass through a valve having a moveable top, to which a spring and weight are attached. The up and down movement of the top of the valve gives a vibratory movement to the air which similarly affects the sound produced. On American organs, a fan-wheel by rotating in front of the wind chest causes a *tremolando*.

Tremore (*It.*) [Tremolando.]

Tremoroso (*It.*) [Tremolando.]

Trenchmore. An old English country dance, or Hey-de-guy. It was of a lively character. Dr. Burton in his "Anatomy of Melancholy," 1621, says: "There is no remedy; we must dance Trenchmore over tables, chairs, and stools;" and Selden in his "Table-Talk" speaks of the dance as an "Omnium Gatherum, tolly polly, hoite cum toite." It is mentioned by several of the dramatic writers of the time of Queen Elizabeth.

TRENCHMORE.

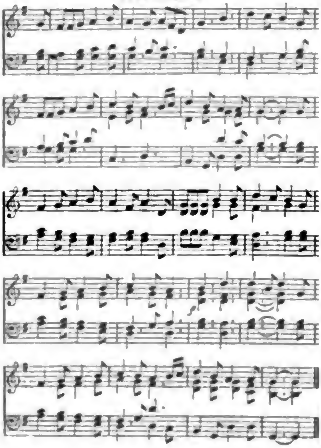

Trenise (*Fr.*) The fourth figure in a quadrille.

Très (*Fr.*) Very. *Très animé*, very animated; *très vif*, very lively.

Triad. (1) A chord of three notes. (2) A common chord. Triads are said to be major, *e.g.* 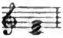; minor, 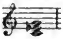; augmented, 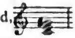; diminished, 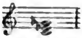

Triangle. An instrument of steel bent into a three-sided form. It is usually held by a string in the left hand and struck with a small bar of iron or steel with the right. It is employed with effect occasionally as an orchestral instrument.

Trias deficiens (*Lat.*) Imperfect triad. [Triad.]

Trias harmonica (*Lat.*) Perfect or major triad. [Triad.]

Tribrach. A foot consisting of three short syllables. [Metre.]

Tricinium (*Lat.*) A composition in three parts.

Trill. A shake.

Trill. Abb. of *trillando*.

Trillando (*It.*) Shaking. A lengthened vocal or instrumental shake.

Trille (*Fr.*), **Triller** (*Ger.*), **Trillo** (*It.*) A shake. In a tract entitled "A brief discourse of the Italian manner of singing, wherein is set down the use of those graces in singing, as the trill and gruppo, used in Italy, and now in England; written some years since by an English gentleman who had lived long in Italy, and, being returned, taught the same here," published by Playford about 1683; the trill is described as a shake upon one note only; it would therefore be similar to the effect called now the *vibrato*, while the gruppo was the *shake* as now practised.

Trillerkette (*Ger.*) A chain or succession of shakes. *Catena di trilli.*

Trillo caprino (*It.*) A goat-like shake.

Trinklied (*Ger.*) A song in praise of drinking.

Trio. (1) A composition for three voices or instruments. (2) A part of a minuet, march, &c. [Form.] [Minuet.]

Triole, Triolet (*Fr.*) A triplet. Three notes played in the time of two of the same value.

Triomphal, Triomphant (*Fr.*), **Trionfale, Trionfante** (*It.*) Triumphal, triumphant.

Tripedisono. A *capo tasto*.

Tripeltakt (*Ger.*) Triple time.

Triphonisch (*Ger.*) Having three sounds.

Triple croche (*Fr.*) A demisemiquaver.

Triple counterpoint. A counterpoint in three parts, so contrived that each part will serve for bass, middle, or upper part as required. [Counterpoint.]

Digitized by Google

Triplet. A group of three notes performed in the time of two. The triplet is always indicated by a slur and the figure 3:

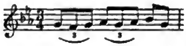

Triple time. Time of three beats, or three times three beats in a bar. [Signature § 2.]

Triplex, Triplum (*Lat.*) The name originally given to a third part when added to two other parts, one of which was a *canto fermo*, the other a *counterpoint*. This additional part was generally the *upper part*, hence the word *treble* or *triplex* came to be applied to the *canto primo*. (2) A motet or other composition in three parts.

Trisagion (*Gk.*) Thrice holy. *Ter Sanctus* (*Lat.*) The opening words of the Sanctus.

Trite. [Greek music.]

Tritone, Triton (*Fr.*), **Tritono** (*It.*), **Tritonus** (*Lat.*) An augmented fourth, containing three whole tones:

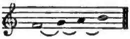

The use of the tritone was anciently forbidden in harmony or counterpoint, as it was regarded in the light of what is called a *false relation*. It was not permitted to be employed in the upper note of one chord and the lower note of the following, as in the subjoined examples:

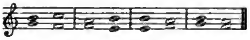

In each case it was called *mi contra fa*, the leading or sensitive note being known as *mi*, and according to the old rules, *mi contra fa diabolus est*.

Tritt (*Ger.*) Tread, treadle, step.

Trittschuh (*Ger.*) A place for the foot on the bellows of old organs.

Trochee. A foot consisting of one long and one short syllable – ⏑. [Metre.]

Troll. To take part in a catch or round, the voices succeeding each other at regulated intervals with the same melody.

Tromb. Abb. for *tromba* (trumpet) and trombone.

Tromba (*It.*) (1) A trumpet. (2) An organ reed stop of 8 ft. pitch.

Tromba bassa (*It.*) A bass trumpet.

Tromba cromatica (*It.*) A keyed trumpet capable of producing intermediate semitones. [Trumpet.]

Tromba marina (*It.*) [Trumpet marine.]

Tromba sorda (*It.*) A trumpet whose sound is stifled by the insertion of a mute in the bell.

Tromba spezzata (*It.*) An obsolete name for the bass trumpet.

Trombetta (*It.*) A small trumpet.

Trombone. (1) A large, deep and loud toned instrument of the trumpet species, the name being an augmentative of *tromba*. It consists of two tubes, so constructed that one may slide in and out of the other, and thus form one tube that can be lengthened at will and made of varying pitch. There are three kinds of trombones, called after their compass the alto, tenor, and bass trombones. Soprano trombones have also been made, but they are rarely used. The general compass of the trombone is a little more than two octaves, the pitch of the instrument varying with the length of the tube. Thus an alto trombone, the part for which is written sometimes in the alto clef, sometimes in the tenor clef, can play all notes between

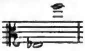

a tenor trombone all between

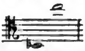

and a bass trombone all between

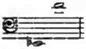

including every intermediate semitone. Each instrument can also sound the note an octave below the first note shown as the commencement of the compass, besides other notes outside the notes indicated above, but as they are difficult and uncertain they are very rarely written. The trombone when judiciously used has a very fine effect, but the modern custom of forcing its tone, and playing in short sharp barks, is much to be regretted. The value of the instrument was well understood by Gluck, (who was probably the first who employed it in the orchestra), as well as by Mozart, Beethoven, Spontini and Weber, as a reference to the scores of Alceste, Die Zauberflöte, Fidelio, La Vestale, and Der Freischütz will prove. (2) A powerful reed stop in the organ of 8 ft. or 16 ft. scale on the manuals and 16 ft. or 32 ft. on the pedals.

Trommel (*Ger.*) Military drum.

Trommelklöpfel, *or* **Trommelschlägel** (*Ger.*) Drumstick.

Trommelschläger (*Ger.*) A drummer.

Trompe (*Fr.*) (1) A trumpet. (2) A hunting horn.

Trompe de Béarn (*Fr.*) [Jew's-harp.]

Trompete (*Ger.*) A trumpet.

Trompetenzug (*Ger.*) Trumpet stop or register in an organ.

Trompette (*Fr.*) (1) A trumpet. (2) A trumpeter. (3) A reed stop in an organ.

Trompette à clefs (*Fr.*) A keyed trumpet.

Trompette à pistons (*Fr.*) A valve trumpet

Digitized by Google

Troop. (1) A march in quick time.

> "When the drums and fifes sounding a troop
> Off they briskly set."—DEFOE.

(2) The second beat of the drum as the signal for marching.

Troparia (*Gk.*) τροπάρια. Hymns which probably had their origin in the ancient custom of inserting ejaculations in the Psalms, especially when used as introits. "Tanto opere enim christiani primis sæculis medii ævi psalmorum cantui, patrum institutis quasi consecrato, favebant, ut nova cantica integra comprobare non auderent, sed *in brevibus strophis psalmorum versibus intexendis vel adfigendis acquiescerent*." But in time the Christian poets *did* dare to write complete new poems which grew out of the ancient *troparium* just as the later prose or sequence grew out of the early attempts to set words to *neumes*. [Prosa.] It is an interesting fact that in the early Greek Church a troparium for a great festival was called an *ἀκολουθία*, which, of course, is actually synonymous with *sequentia*.

Troparion. An office-book of the Greek Church containing the sequences or chants sung after the lessons. An ancient troparion with the musical notes is preserved in the Bodleian library at Oxford.

Troppo (*It.*) Too much. *Troppo caricata*, too much loaded, overloaded with ornaments or accompaniment. *Non troppo allegro*, not too quick.

Troubadour. A musical poet, who, unlike the jongleur, did not wander about the country singing for hire. In the days when all classes of the community were equally unrefined, there was no such distinction, every verse-maker was called a troubadour, a word derived probably from the Provençal *trobar*, to invent or find; and ultimately the term came to be confined to kings, princes, and nobles, who practised poetry for pleasure, or out of chivalrous gallantry. The minstrels or jongleurs only recited or chanted poems, but did not write or invent them; or perhaps accompanied on some instrument the troubadour who sang his own compositions. It was not an unusual thing for a troubadour to have several minstrels or jongleurs in his service; the word minstrel meant probably a minister or servant.

The minstrels in later times formed a separate guild, uniting for the purposes of mutual protection and support; but the troubadours were always independent, and this independence gives a character to their individual productions, unfettered as they are by the trammels of any particular school of thought. The same free spirit gives a colour to the rude outline of their history, so that the practice of their art is to be traced, not so much by its expansion and development in classes as in individuals. Troubadours frequently attached themselves to the courts of kings and nobles, whom they praised or censured in their songs; but it was a rule that some lady was selected as the "dame de du cœur," and to her, under some general or fanciful title, love songs, complaints, and other poems were addressed. The "love service" of the troubadours was often nothing more than a mere artificial gallantry, but there are instances on record where it became something more earnest.

Contests, competitions, or verse battles were sometimes entered into, generally on questions of gallantry suggested by the ladies who presided as judges over a tribunal called the Court of Love, and awarded prizes to the victors. The poems of the troubadours were not always confined to subjects of gallantry, sometimes they treated of the conditions of society, the evils of the times, the degeneracy of the clergy and other subjects.

There is reason for supposing that the art of the troubadours, generally called the *gay science*, was derived from the East, coming into Europe through the Spaniards, and the troubadours of Provence learning from their neighbours of Spain. Troubadour poetry was cultivated in Provence, Toulouse, Dauphiné, and other parts of France south of the Loire, as well as in Catalonia, Arragon, and Valencia in Spain, and in the north of Italy. Its duration was about 200 years (1090-1290), the period when Eastern customs were giving a tone to those of Western Europe, and while chivalry began to redeem men from barbarism. Warton (History of English Poetry) rightly estimates the value of the poems of the troubadours, when he says:—"The Provençal writers established a common dialect, and their examples convinced other nations that the modern languages were no less adapted to composition than those of antiquity. They introduced a love of reading, and diffused a general and popular taste for poetry by writing in a language intelligible to ladies and the people. Their verses, being conveyed in a familiar tongue, became the chief amusement of princes and feudal lords, whose courts had now began to assume a greater brilliancy; a circumstance which necessarily gave great encouragement to their profession, and by rendering those arts of ingenious entertainments universally fashionable, imperceptibly laid the foundation of polite literature. From these beginnings it were easy to trace the progress of poetry to its perfection through John de Meum in France, Dante in Italy, and Chaucer in England."

Trouveur, Trouvère, Trouverre, *or* **Rymour.** According to Ritson, one who composed *Romants*, *Contes*, *Fabliaux*,

Digitized by Google

Chansons, and *Lais*, whilst those who devoted themselves to the composition of *Contes* and *Fabliaux* were called *Contours*, *Conteurs*, or *Fabliers*. [Romance.] [Troubadour.]

Trugschluss *or* **Trugcadenz** (*Ger.*) An interrupted or deceptive cadence.

Trumbscheit (*Ger.*) [Trumpet marine.]

Trummel (*Ger.*) [Trommel.]

Trump. A poetical name for the trumpet.

Trumpet, Tromba (*It.*), **Trompette** (*Fr.*), **Trompete** (*Ger.*) A metal wind instrument of bright and penetrating tone, formed of a single tube curved into a convenient shape, with a mouthpiece at one end, the other having a bell. Its part is usually written in the key of C with the treble clef, though by means of crooks or lengthening pieces the sounds produced may be in various keys. The trumpet required for a piece is indicated at the commencement; as, trumpet in C, D, E flat, E, F, or G.

The scale of the instrument is formed of the harmonics of an ordinary open pipe.

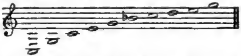

By means of the slide, the B flat shown above, which is a little flat, may be sounded in tune, and certain intermediate semitones obtained.

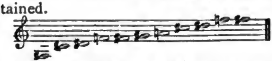

In addition to the notes shown above, the trumpet is capable of producing,

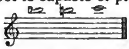

and

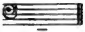

but the higher tones are very difficult to obtain. Handel and other writers of the last century frequently wrote them, but trumpeters of the present day omit them.

In old scores the trumpet was often coupled with the drums, the notes of the one corresponding to the beats on the other.

Gluck and Handel were among the earliest writers to discover the effect of long-holding notes, and the power of the softer tones of the trumpet.

Trumpets with pistons or valves capable of producing every chromatic sound within their compass are sometimes used, but the tone is by no means to be compared with the true trumpet tone.

Trumpet marine. An instrument formed of a triangular chest, over one side of which is stretched a thick gut string, passing over a bridge slightly uneven on its feet, one side being fastened and the other free. When the string is set in vibration by means of a bow, the rapid impact of the loose foot of the bridge on the belly slightly checks the vibration and causes the sound to resemble that of the trumpet. The fingers of the left hand of the player, being passed lightly over the strings, the ordinary harmonics of an open string are produced.

Tuba (*Lat.*) (1) A trumpet. (2) A powerful reed stop in the organ.

Tuba major, *or* **Tuba mirabilis.** An 8-ft. high-pressure reed.

Tubular Pneumatic Action. [Organ Construction, § 12.]

Tucket, Touquet (*Fr.*) A flourish on a trumpet. Mr. Staunton, in his notes on Shakspeare, conjectures it to be derived from the Italian *toccata*, or the Spanish *tocar; tocar trompeta*, to sound a trumpet.

Tune. (1) A melody or air. [Air.] (2) Just intonation.

Tuning Fork. An instrument of steel with two prongs, which when set in vibration gives out a musical sound varying in pitch according to the thickness of the metal or the length or width apart of the prongs. It was invented by John Shore, sergeant trumpeter to George I. There is a considerable variety in the pitch of tuning forks, arising from the absence of any recognised standard of tonality. [Pitch.] The ordinary fork gives out a single note only, but one has been introduced from Germany which has a slider on each prong which can be moved up or down so as to alter the pitch.

Tuning. The adjustment of the sounds naturally produced by any instrument to some standard pitch and to their proper relation to each other. Wind instruments played by the mouth are tuned by adjusting the length of the tube to one particular note; and after that the just intonation to a great extent depends upon the construction of the instrument, subject of course to certain artificial alterations made by varying the pressure of wind, or altering the length of the vibrating column of air by the insertion of the hand, &c.

String instruments of the violin, guitar, and pianoforte class are tuned by altering the tension of the strings at the end where they are carried round a moveable peg. In the first of these three instruments, after the strings or open notes have been tuned, true intonation depends on the correct *ear* and mechanical skill of the performer; in the second class the position of the *frets* determines the pitch throughout. In the pianoforte and the organ, the intonation is entirely out of the control of the player, and owing to the large number of octaves in their compass certain notes are first tuned to the system of temperament adopted, and from these the whole instrument is tuned. These sounds first adjusted are called *bearings*. But the whole subject of tuning these two instruments is so intimately connected with Temperament that the reader is referred to that article.

Digitized by Google

Bells are tuned by paring off a portion of the inner side of the sound-bow, thus altering the relation of diameter to thickness.

Musical glasses can be altered in pitch by pouring a liquid into them. Drums are tuned by tightening or relaxing their parchments.

Reeds in an organ or harmonium can be tuned by altering the length of the part allowed to vibrate, also by varying their thickness or breadth.

Organ flue-pipes are tuned by opening or closing the orifice, a process which has practically the same effect as shortening or lengthening the column of vibrating air.

Tuning hammer. An instrument employed for tuning pianofortes or harps. It is in shape like a common hammer, but has a head of wood instead of iron and a shank of iron instead of wood. In the bottom end of the shank is a square or oblong hole made to fit the tops of the pegs round which the strings of the instrument are twisted; these pegs being turned to the right or left, tighten or loosen the strings as needed. The old name for the tuning hammer was *wrest*.

Tuono (*It.*) (1) Body of tone, sound. (2) A tune.

Tuono ecclesiastico (*It.*) Accentus ecclesiasticus. [Accent, § 4.]

Turbæ (*Lat.*) The chorus part or voice of the multitude in a Passion-Music.

Turca (*It.*) Turkish; *alla turca*, in the Turkish style.

Turkish Music, *or* **Janitscharenmusik.** Noisy music produced solely by instruments of percussion such as cymbals, gongs, &c.

Turn. An ornament in music formed by taking the adjoining notes above or below the principal note, according to the position of that note in the diatonic scale. The turn indicated by the sign ~ must be performed in the time the note it alters would occupy without it. Thus the common turn, which takes a *higher* note first in the change:

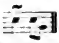 should be performed— 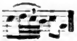

The back-turn taking a lower note first in the change:

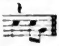 should be performed— 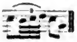

When the turn appears over a note followed by a rising or a falling interval:

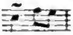 or 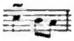

The turn should consist of four notes besides the principal:

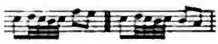

or in the case of the back-turn, written thus:

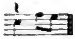

it should be performed:

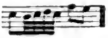

If it is desired to alter the note above or below a turn, that is to say, to change the character as it appears in the scale, the following signs are used. The position of a sharp or flat underneath the sign of the turn shows whether the note above or below is to be sharpened or flattened; the dash through the turn always signifying a sharpened note.

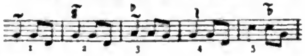

Performed:

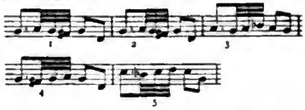

A turn over a note in the unaccented part of the bar is usually performed with the changed note first.

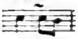
Written.

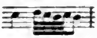
Played.

Turr. A Burmese violin with three strings.

Tutta (*It.*) All, the whole; *tutta forza*, the full power or force; *tutto arco*, the whole length of the bow.

Tutte corde (*It.*) A direction to play on all the strings instead of *una corda*.

Tutti (*It.*) All. Every performer to take part in the execution of the passage or movement.

Tuyau d'orgue (*Fr.*) An organ pipe.

Tuyaux à anche (*Fr.*) Reed pipes.

Tuyaux à bouche (*Fr.*) Open pipes.

Twelfth. (1) An interval of twelve diatonic degrees, the replicate of the fifth. (2) An organ stop tuned twelve notes above the diapasons.

Tympani. [Timpani.]

Tympanon (*Fr.*) A dulcimer.

Tyro. A learner or beginner in music or any other science.

Tyrolienne. (1) A song accompanied with dancing. Rossini's "Toi que l'oiseau," in the third act of *Guillaume Tell*, is the earliest specimen of a Tyrolienne other than the popular *Volkslieder*. (2) Popular songs or melodies in which the jödl, *q.v.*, is freely used.

Tzetze. An Abyssinian instrument of the guitar kind, formed of a long carved neck attached to a gourd. It has frets and one string usually made of the tough fibre of a palm-tree.

U.

U.C. Abb. of *una corda.* [A una corda.]

Uebelklang
Uebellaut (*Ger.*) Discord, cacophony.

Uebereinstimmung (*Ger.*) Consonance, harmony.

Uebergang (*Ger.*) Passage, transition.

Uebermässig (*Ger.*) Augmented.

Uebung (*Ger.*) An exercise.

Ugab, Huggab, Agub (*Heb.*) The first wind instrument mentioned in the Bible, rendered *organ* in the authorized version, "such as handle the harp and the organ" (Gen. iv., 21.) In the Septuagint it is sometimes translated by *cithara,* sometimes by *psalm,* sometimes by *organ.* It is probable that in its earliest form the *ugab* was nothing more than pans-pipes or a syrinx, but that it gradually developed into a more important instrument. In Psalm cl. it may be possibly used in a poetical sense of all wind instruments, as *minnim* is of string instruments. "Praise Him upon the strings (minnim) and pipe (ugab)."

Uguale (*It.*) Equal, like, similar, *Canone a tre voci uguali,* a canon for three equal voices.

Ugualmente (*It.*) Equally, alike, similarly.

Umana,
Umano (*It.*) Human; as, *voce umana,* the human voice.

Umfang (*Ger.*) Compass, extent; as *Umfang der Stimme,* compass of the voice.

Umkehrung (*Ger.*) Inversion.

Un
Una
Uno (*It.*) One; as, *Una corda,* one string; *Una volta,* once, &c.

Unca (*Lat.*) A quaver.

Uncoupled. A direction that the manual or pedal is to be detached from the row to which it was coupled. [Organ.]

Unda maris. *Lit. Wave of the sea.* A name given to the "Vox Angelica" organ-stop, because of its wavy undulating tone. [Vox Angelica.]

Undecima (*Lat.*) An eleventh.

Undecimole. A group of eleven notes to be played in the time of eight of the same name.

Undersong. [Burden.]

Undertones. [Acoustics, § 19.]

Undulazione (*It.*) The tremulous sound produced by violinists by the vibratory pressure of the finger upon the strings.

Unequal Temperament. [Temperament.] [Wolf.]

Unequal Voices. Voices of mixed qualities, those of women and men combined.

Unessentials. Notes not forming a necessary part of the harmony. Passing, auxiliary, or ornamental notes.

Ungerader Takt (*Ger.*) Triple time.

Unharmonischer Querstand (*Ger.*) [False relation.]

Unison. (1) Having the same number of vibrations; homophonous. (2) Music *in octaves* for mixed voices or instruments.

Unisoni (*It.*) Unisons; two or more parts playing in unison with each other, or at the octave, according to the character of the instrument or voice.

Unisono (*It.*)
Unisonous (*Eng.*)
Unisonus (*Lat.*) In unison, or in octaves.

Unitamente (*It.*) Together, unitedly.

Uno a uno (*It.*) One by one, severally.

Un peu plus lent (*Fr.*) A little more slow.

Un pochettino,
Un pochino (*It.*) A little, a very little; as, *un pochino più mosso,* a very little more lively.

Un poco (*It.*) A little; as, *un poco allegro,* rather quick; *un poco più,* a little more; *un poco ritenuto,* held back a little, slightly retarded.

Unterbass (*Ger.*) A double-bass.

Unterdominant (*Ger.*) Subdominant *Unterdominant-Akkord,* the chord of the subdominant.

Unterhalbton (*Ger.*) The leading note.

Unterleitton (*Ger.*) The dominant seventh.

Untersatz (*Ger.*) Sub-bass.

Unterstimme (*Ger.*) An under-part.

Upbeat. The beat of a bar at which the hand is raised. An unaccented beat. [Accent.] [Arsis.] [Metre.]

Up bow. [Bowing.]

Uppatura (*Med. Lat.*) A song of a profane character, forbidden to be sung in church by the Constitution of the Carmelite order. [Motet.]

Urh-heen. The Chinese fiddle. [Violin.]

Ut. [Aretinian syllables.] [Notation.]

Ut bémol (*Fr.*) The note C flat.

Ut dièse (*Fr.*) The note C sharp. *Ut dièse mineur,* the key of C sharp minor.

Utriculariæ (*Lat.*) Tibiæ utriculariæ. [Bagpipe.]

Ut supra (*Lat.*) As above, as before. *Gloria Patri ut supra,* the Gloria as before.

Uvula. [Larynx.]

Digitized by Google

V.

V. Abb. of *verte*, *violino*, *violini*, *voce*, *voci*, *volta*, *volti*, *&c.*

Va. Abb. of *viola.*

Va (*It.*) Go on; as *va crescendo*, go on increasing the power; *va rallentando*, go on dragging the time.

Vacillando (*It.*) Wavering, uncertain as regards the time or tone.

Vago (*It.*) With a vague, indefinite expression.

Valce (*It.*), **Valse** (*Fr.*) Waltz.

Valeur (*Fr.*), **Valore** (*It.*) The value, worth, or length of a note.

Valse à deux temps (*Fr.*) A form of waltz, now most commonly danced, in which two steps are made to each measure of three beats.

Vamp. To improvise an accompaniment.

Variamente (*It.*) In a varied manner.

Variations, Variationen (*Ger.*) **Variazioni** (*It.*) Certain modifications with regard to the time, tune, and harmony of a theme proposed originally in a simple form. At one period it was considered indispensable that the subject chosen should be heard unchanged through all the variations, that no alteration should be made either in the relation, length, or melodic progression of the sounds. This only permitted the employment of the several species of counterpoint as variations. Afterwards a ground bass was selected, or written, in which more freedom and variety was attainable. Subsequently composers altered their melodies by the addition of florid passages, but not to such an extent as to make the subject not easily recognisable. There are many instances in the works of most of the classical writers in which this artifice is successfully employed. The old composers called their variations "Doubles." Modern transcriptions are often nothing more than elaborate variations of a theme.

Variato (*It.*), **Varié** (*Fr.*) Varied, changed, altered, with variations.

Vaudeville (*Fr.*) A term originally applied to a country song of like kind with those written by Oliver Basselin, of the valleys of Vaux de Vire in Normandy, in the fifteenth century. These songs, which were satirical, had for their subjects love, drinking, and passing events. They became very popular, and were spread all over France under the name of *Lais des Vaux de Vire.* The peculiarity of their character lived after their origin was forgotten, and plays, interspersed with songs of this description, came to be called Vaudevilles, and occasionally *Virelais.* The songs in Vaudevilles should form an integral portion of the plot or subject.

The following melody of a Vaudeville is from the "Second livre de Pièces de Viole avec la Basse continue. Composé par Mons. De Caix D'Hervelois." Paris, c. 1719.

VAUDEVILLE GAVOTTE.

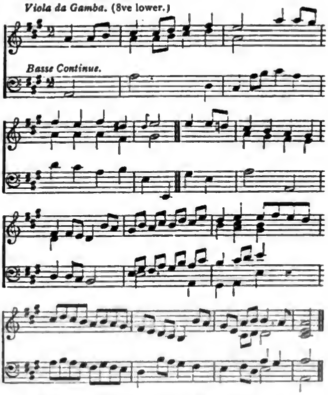

Veemente (*It.*) Vehement, forcible.

Veemenza, con (*It.*) With vehemence, force.

Velato, Velata (*It.*) Veiled, as *voce velata*, a voice the tone of which is not clear.

Vellutata, Vellutato (*It.*) From *velluta*, velvet. In a soft, smooth, velvety manner.

Veloce (*It.*) Rapid, swift.

Velocissimamente (*It.*) Very swiftly

Velocissimente (*It.*) Swiftly.

Velocità, con (*It.*) With swiftness, rapidity.

Digitized by Google

Velocity of sound. [Acoustics § 5.]

Ventil (*Ger.*), **Ventile** (*It.*) (1) A valve, by means of which brass tubes may be made to sound the semitones and tones between the natural open harmonics. [Metal wind instruments.] (2) A mechanical contrivance on an organ for the purpose of cutting off the wind from a particular sound-board.

Venusto (*It.*) Beautiful, graceful, fine.

Veränderungen (*Ger.*) (1) Variations. (2) The mechanical arrangement by which ' stops " on a harpsichord were moved in and out.

Verbindung (*Ger.*) Combination, union, connection, binding.

Vergnügt (*Ger.*) Pleasantly.

Verhallend (*Ger.*) Dying away, *decrescendo.*

Verhältniss (*Ger.*) Proportion; *ratio.*

Verilay (*Fr.*) [Vaudeville.]

Vermindert (*Ger.*) Diminished; as *verminderter Dreiklang*, diminished triad.

Verschiebung (*Ger.*) The soft pedal of a pianoforte; *mit Verschiebung*, *una corda*.

Verse. (1) Those portions of an anthem or service intended to be sung by a single voice to a part. [Anthem.] (2) A separate stanza of a song or a ballad. (3) A verse anthem is one which begins with soli portions as opposed to a full anthem, which commences with a chorus.

Versette (*Ger.*) Short pieces for the organ intended as preludes or voluntaries.

Versetzen (*Ger.*) To transpose.

Versetzungszeichen (*Ger.*) Accidentals.

Verspätung (*Ger.*) Retardation, delay.

Verstimmt (*Ger.*) Out of tune.

Versus Fescennini (*Lat.*) Nuptial songs, so-called because they were first used by the people of Fescennia in Etruria. From this kind of poetry arose the Epithalamium, a needful refinement upon the Fescenninian verses.

Vertatur, Verte (*Lat.*) Turn over.

Verwandt (*Ger.*) Related, as *verwandte Tonarten*, related keys.

Verwechselung (*Ger.*) Changing, altering, as to tone, time, key.

Verweilend (*Ger.*) Delaying, *rallentando.*

Verwerfung (*Ger.*) Transposing.

Verzierungen (*Ger.*) Ornaments, decorations, embellishments.

Verzögerung (*Ger.*) Retardation.

Vezzosamente (*It.*) Tenderly, softly, gracefully.

Vezzoso (*It.*) Tender, sweet, graceful.

Vibrante (*It.*) Vibrating, tremulous.

Vibrato (*It.*) A tremulous quality of tone, as opposed to a pure equal production.

Vibration. [Acoustics § 3.]

Vicar choral. [Lay vicar.]

Vide (*Fr.*), *Vido* (*It.*) Open. [Vuide.]

Viel (*Ger.*) Much; as, *mit vielem Tone*, with much tone.

Vielle (*Fr.*) The hurdy-gurdy.

Vielstimmig (*Ger.*) For many voices.

Vierfach (*Ger.*) Four-fold; having four ranks of pipes. [Organ.]

Viergesang (*Ger.*) Song for four parts.

Vierspiel (*Ger.*) A composition in four parts. Quartet.

Vierstimmig (*Ger.*) For four voices or instruments, or in four parts.

Vierstück (*Ger.*) A piece for four players, a quartet.

Viertelnote (*Ger.*) Quarter note, a crotchet, the fourth part of a semibreve.

Viertheilston (*Ger.*) A quarter tone, half a semitone.

Viervierteltakt (*Ger.*) Duple time of four crotchets in a bar.

Vierzweiteltakt (*Ger.*) Duple time of four minims in a bar.

Vif (*Fr.*) Lively, brisk, quick.

Vigorosamente (*It.*) Vigorously.

Vigoroso (*It.*) Vigorous, bold, forcible.

Vigüela (*Sp.*) A simple kind of guitar.

Villancico (*Sp.*) A species of song of two or more stanzas, each containing seven lines, belonging to the poetry of the 15th century, which, like the madrigal, is of an epigrammatic form—formerly very popular in Spain. The composers of that country employed the tunes of many of these songs as themes for counterpoint in church motets. Those motets which are sung during high mass on Christmas-eve are always called Villancicos.

Villanella (*It.*) A rustic dance accompanied with singing. The melody was usually lively and the rhythm well marked. The words, when any were used, were commonplace epigrams, and were also called madrigals and ballets. The villanella or villotte, and villancico were the precursors of the madrigal. When the title *villanella* began to be used to describe compositions in rustic style, not intended for voices, those compositions were embellished with variations upon the original subject.

Villotte (*It.*) The name given to the first secular pieces in harmony after the rules of counterpoint were fixed. Every so-called " license " in harmony was used in this kind of composition; whereas church music could only be written according to strict rules; therefore, secular pieces were called *vile* or *rustic*, as being more fitted for clownish than for courtiers' ears. Morley, speaking of the villanella says, " many perfect chords of one kind, nay, even disallowances may be taken at pleasure, uniting a clownish music to a clownish matter." [Villanella, Villancico, Madrigal.]

Digitized by Google

Vina. [Bina.]

Viol. A stringed instrument, shaped like the violin but of larger size; it was furnished in England with five or six strings, had a fretted finger-board and was played with a bow. The viol was called in mediæval Latin, vitula, and is found depicted in MSS. as early as the 11th century. In France, Germany, and Italy the number of the strings varied between three and six. It is supposed that they were tuned in fourths and thirds. A chest of viols consisted of six instruments of various sizes, the smaller ones were called in England, *treble*, the next *mean*, and the larger *bass* viols. In Germany the names were Bratsche, Schultergeige, and Bassgeige, and in Italy viola da braccio and viola da gamba were the names given to distinguish the several sizes.

Viola (*It.*) *Bratsche* (*Ger.*) The tenor violin. It has four strings, tuned thus:

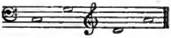

The two lowest are covered strings. Music for this instrument is written in the *alto* clef, whence it is sometimes called *alto viola*.

Viola di Bordone (*It.*) An instrument of the violin kind, strung with six or seven catgut strings tuned in the following manner:

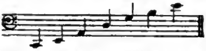

Beneath the gut were metal strings, varying in number from sixteen to as many as forty-four, arranged in a diatonic order from

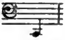

as the lowest tone. The sympathetic strings were occasionally plucked with the left hand in playing. The instrument is now obsolete, but was in use in the time of Haydn, who at the request of Prince Esterhazy composed upwards of sixty pieces for it. It is also called *viola di fagotto*, and *baryton*.

Viola Pomposa (*It.*) A species of *viola da gamba*, invented by John Sebastian Bach. It had five strings, the four lower strings were tuned like the violoncello, in fifths, and the fifth string was tuned to E, by means of which greater facility in the execution of extended passages was possible. Improved skill on the part of violoncello players made the *viola pomposa* unnecessary.

Viola da Gamba (*It.*) One of the larger instruments among a chest of viols. It received its name from being held within the knees of the performer. It had a fretted finger-board and six strings, tuned thus:

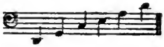

It was particularly adapted to the performance of broken harmony and extended chords.

Viola d'amore (*It.*) An obsolete instrument of the violin family. In addition to catgut strings, metal strings were placed under the finger-board, which, by the production of sympathetic sound, gave a peculiar quality of tone to the instrument.

The tuning varied according to the piece to be played, the "accordatura" being often indicated at the commencement of the copy which was played from, but the most usual method of tuning was the following:

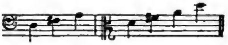

Meyerbeer has written an obbligato part for this instrument in the first act of the "Huguenots."

Violetta. A little viol.

Violetta Marina (*It.*) A stringed instrument similar in tone to the viol d'amour. It was called *violetta piccola*, and by the French *Haute-contre*.

Violin. **Discantgeige** (*Ger.*) **Violon** (*Fr.*) **Fiddle** (*Eng.*) The most familiar of all stringed instruments played with a bow. It is somewhat smaller than the old viol, as its name implies, violin being a diminutive of viol. Like the rest of the family it represents, it consists of a wooden chest of peculiar form, made of two curved surfaces, called the back and the belly, united by sides, and with a hollow half-way in the length. A neck at one end has a finger-board, over which the four strings pass, being fastened at one end of the chest or body to a tail-piece, and kept in tune and position by a series of pegs at the end of the neck. The strings are raised above the belly by the bridge. In the belly are two holes, called the *f* holes from their similarity to the shape of that letter. The sound is produced by drawing a bow of horse-hair charged with rosin across the strings, which are tuned in fifths:—

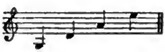

the changes of pitch being gained by "stopping" the strings with the fingers of the left hand against the finger-board, thus shortening the vibrating portion of the string. The harmonics of the violin are very telling in quality, and are produced by touching the strings lightly instead of pressing them upon the finger-board. The *sordino* or *mute*, placed upon the bridge, produces a peculiar modification of tone, and a good effect is gained by plucking the strings, as in playing a guitar. [Pizzicato]

Digitized by Google

The compass of the violin ranges between:—

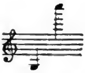

with every intermediate shade of sound. The extreme high notes are rarely used, as they are harsh and shrill. The violin is capable of producing a limited harmony by means of double stops and bowing in "arpeggio," while as to power of expression and execution there is no other instrument which can be compared to it. It has a wide range of sounds, to which any degree of loudness or softness, *staccato* or *legato*, can be given. This variety of tone, added to the capability of being played a great length of time with less fatigue than any other orchestral instrument, renders it invaluable, either for solos or combinations. The violin or string quartet, as it is called, that is to say, 1st and 2nd violins, viola, and violoncello, form an indispensable portion of a score; while the same combination, used for quartets, is productive of the highest pleasure when well and skilfully managed.

Stringed instruments played with a bow are in use among many Eastern nations, the form of the instrument so played varying considerably. The number of strings and the form of the bow also differ. The *Urh-heen*, or Chinese fiddle, has no finger-board. There are usually only two strings, but some specimens have as many as four. The bow is twisted under the strings, and it is said that the Chinese fiddlers have to practise assiduously in order to produce the tone properly from strings so placed. The resonance body of the Urh-heen is covered with snake-skin.

CHINESE URH-HEEN.

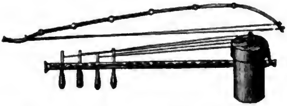

The "Burmese Thro, Theyaou, or Tarau," has three strings of silk, and although the tone is nasal, it is not of unpleasant quality.

BURMESE THRO OR TARAU.

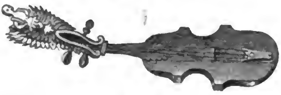

The "Kemangeh," another bowed instrument, is in use among Mohammedan nations. It is usually furnished with three strings of gut, and there are two sound-holes beside the bridge. The *Kemangeh* is made in several sizes, like the European violin family. The larger instruments are supported on an iron peg.

KEMANGEH.

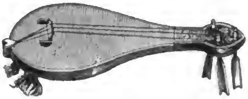

The Hindu "Chikarah" is supposed by some to be the modern representative of the ancient *Ravanastron*, the oldest form of fiddle.

CHIKARAH.

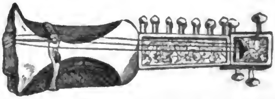

The "Rebab" of Egypt has one string of horse-hair stretched over a resonance-body of skin. In Egypt it is used to accompany the recitations of the storytellers and the motions of the dancers.

REBAB.

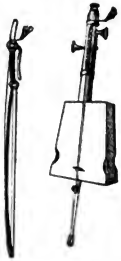

It is sometimes made in shape like the "Kemangeh," and furnished with three or even more strings. Whatever the form may be, it is a favourite with the country-people who use it. It is also called *rabel* or *arrabel*. When it was introduced into Europe it retained its Eastern name, with certain modifications; *rebebe, reberbe, rubebe, rubeck, rebec; but it was altered in shape.* The old English *jig*, *jegg*, or *gig*, as it is variously spelt, was also called *rebec*, the "jocund rebeck" mentioned by Milton and other

Digitized by Google

poets; it had three strings like its Eastern prototype, the rebab.

Modern writers find the origin of the violin in the above-mentioned Indian *Ravanastron*, an instrument still existing, and used by the poor Buddhist begging monks. From the Ravanastron descends the rebab, kemangeh, whence comes the Scandinavian guddok, the ancestor of the Welch and Anglo-Saxon crwth. From the crwth, the fithele, vitula, or viol descended, and from the viol the violin, the whole progress representing a period of nearly 5,000 years in history. The violin has not altered its form since the 16th century, many instruments of that time being still in actual use. One of the earliest makers of violins was Gasparo da Salò, in Brescia, 1542-1609.

In Cremona, during the 17th century, the Amati family, Andrea, his sons Girolamo and Antonio, and Nicola the son of the latter, were famous makers of violins. Antonio Stradiuarius, also of Cremona, pupil of Nicola Amati, surpassed, if possible, the productions of the Amatis, and the reputation of Cremona for violins was maintained by Guarnerius and Ruggeri. The Tyrolese makers, Jacob Stainer and Mathias Klotz (and his sons) became as famous as the Italians for violin-making. Gand, Lupot, and Vuillaume, of Paris, are celebrated modern makers.

Violin clef. The G clef placed upon the first line of the stave.

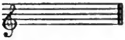

In this position it is known as the *French violin clef*. Now obsolete.

Violino principale (*It.*) Solo violin, or leader. *Violino primo*, first violin. *Violino secondo*, second violin. *Violino ripieno*, a violin part required only to fill in and strengthen the *tutti*.

Violinsteg (*Ger.*) The bridge of a violin.

Violon (*Fr.*) The violin; in Germany the same word is used for the double bass.

Violoncello (*It.*) The little *violone*. When the violoncello supplanted the bass viol or the viola da gamba in the early part of the eighteenth century, many of the other instruments were so altered, by having the number of strings reduced, as to make them playable like the violoncello. Excepting occasionally, when the exigences of an old score demand the use of a viola da gamba, no other small bass stringed instrument is employed in the orchestra, as the variety of tone it is capable of—almost equal in resource to the violin—renders the use of others unnecessary. The violoncello is strung with four gut strings, the lower two covered with silver wire; all are tuned a fifth apart.

The compass usually employed extends between

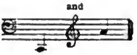

though soloists play an octave higher, with all the intermediate semitones. Music for it is written in the bass clef up to

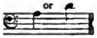

but the C clef, or sometimes the G clef, is employed for the higher notes of the scale.

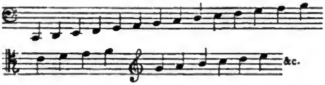

The tone of the violoncello is peculiarly sympathetic, almost tearful in the higher register, and these notes are frequently employed with great effect as well in solos as in orchestral music. In the band the violoncello is coupled with the double-bass, and the union of the tones of the two instruments is wonderfully telling and solid. Mendelssohn has used the double-bass without the violoncello in *Elijah*, and the effect is stormy and uncertain, while the violoncello without the double-bass is light and expressive of unrest. The violoncello is capable of giving rapid passages with more clearness than the double-bass, and so it is frequently found that composers give essential notes to the larger, and intermediate notes to the smaller of the two orchestral basses.

BEETHOVEN (Pastoral Symphony).

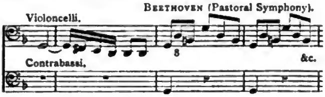

HANDEL (Joshua).

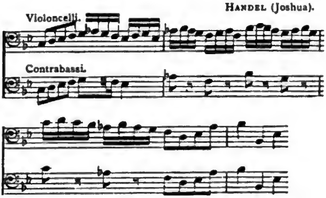

The division of the two parts in this fashion was at one time a matter of necessity, as double-bass players were not sufficiently skilful to be able to perform such passages, but at the present time there is scarcely a double-bass player who is not able to take such divisions with ease.

In addition to the single notes of its scale the violoncello can give double notes, seconds, thirds, fifths, sixths, sevenths, and octaves in nearly every key between the two notes

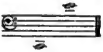

Seconds, thirds, and fourths can be made when the upper note is an open string, as

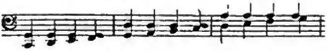

Fifths, sixths, and sevenths when the lower note is stationary:

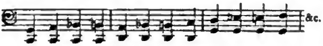

Violone (*It.*) The double bass.

Virelay (*Fr.*) [Vaudeville.]

Virginal. A stringed instrument played by means of a keyboard, like the modern pianoforte. It was in form like a box, or desk of wood without legs or supports, and was usually placed upon a table or stand. The strings were of metal, one for each note, and the sound was made by means of pieces of quill, whalebone, leather, or occasionally elastic metal, attached to slips of wood called "jacks," which were provided with metal springs. The compass was about three octaves. This title has been by some supposed to have originated as a compliment to Queen Elizabeth, who was an expert performer upon the instrument; but as the word was in use before her birth this supposition is not valuable. Others say that the name was applied because of the smallness and delicacy of its tone, as it was and is considered proper for a woman "to speak small." It is more likely to have received its name from the word "virgæ." The virginal, also called spinet, from the quills with which the string was sounded, was the precursor of the harpsichord, now superseded by the pianoforte. [Dulcimer, Spinet, Pianoforte.]

Virgula (*Lat.*) (1) The stem or tail of a note. (2) A neume.

Virtuoso (*It.*), **Virtuose** (*Ger.*) A skilled performer on some particular instrument, specially the violin. The term was formerly applied to an amateur, "one who feels delight in, and possesses a taste for, the musical science."

Vista (*It.*) Sight; as, *a prima vista*, at first sight.

Vistamente, Vitamente (*It.*) Briskly, quickly, rapidly, with life.

Vivace, Vivacemente (*It.*) Livel quickly, sprightly.

Vivacetto (*It.*) Rather lively.

Vivacezza, Vivacità (*It.*) Liveliness vivacity.

Vivacissimo (*It.*) Very lively.

Vivamente (*It.*) Lively, briskly.

Vive (*Fr.*) Brisk, quick, lively.

Vivente (*It.*) Animated, lively.

Vivezza, con (*It.*) With life, liveliness.

Vivido (*It.*) Brisk, lively.

Vivo (*It.*) Alive, brisk, lively, animated.

Vocal. (1) For, or by the voice; musi intended to be sung. (2) Compositions so written as to be easy and effective for the voice. (3) The "singing" quality of tone obtained from an instrument.

Vocale (*It.*) Belonging to the voice.

Vocalezzo (*It.*) An exercise for the voice.

Vocalisation. (1) Control of the voice and vocal sounds. (2) Method of producing and phrasing notes with the voice.

Vocalizzi (*It.*) Vocal exercises, solfeggi.

Vocal Score. [Score.]

Voce (*It.*) The voice; as, *voce di camera*, literally, a voice for the chamber, applied to a voice of small quality, better fitted to the limited area of a private room than for a public concert room. *Voce di gola*, a guttural or throaty voice; *voce di petto*, the chest voice, the lower register of the voice; *voce di testa*, the head voice, the upper range of the voice; sometimes the falsetto is so termed. *Voce sola*, the voice alone, unaccompanied.

Vogelflöte (*Ger.*) A bird-call, flageolet, or whistle. Mozart employs one in the score of the *Zauberflöte*.

Vogelgesang (*Ger.*) A musical instrument composed of a series of small pipes standing in water, through which the wind had to pass. A *merula*.

Vogelpfeife (*Ger.*) [Vogelflöte.]

Voglia (*It.*) Ardour, desire, longing.

Voice. Voices may be arranged in seven orders or classes, according to gravity or acuteness—viz., The bass, baritone, tenor, alto, contralto, mezzo-soprano, and soprano. The first four are the natural voices of men and the last three those of women. The compass or range of notes is different in each voice, but it is not compass alone which determines the class to which any voice may belong, as very frequently a baritone quality of voice is limited to the range of a bass, and a tenor quality to the compass of a baritone. It is almost impossible to describe in words the general character of the several voices, it must be sufficient to give the number of notes they are capable of using.

The part for the bass voice is written in the F clef and its most effective notes are those indicated in the following by

Digitized by Google

crotchets, though for solo purposes the notes shown by the minims are often used.

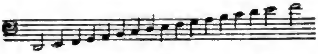

The baritone voice has a compass between

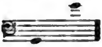

its quality and character best fit it for solo singing. Its part is now written in the bass or F clef, on the fourth line, though there are instances, as, for example, in Dr. Cooke's setting of Collins' "Ode to the Passions," where the baritone clef, that is to say, the F clef on the third line, is used for the voice.

The compass of the tenor voice is between

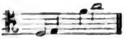

The most serviceable notes being those between the two crotchets. Its part is written in the C clef upon the fourth line. The alto voice is an artificial and not a real voice. It is made by cultivating the falsetto notes instead of the true chest voice. The C clef is also used for this voice, but it is placed upon the third line, and the compass employed is between the notes:

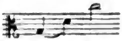

All the tones between the two crotchets are produced from the chest, the D generally dividing the two qualities of chest and falsetto, the point of junction being known as "the break."

The contralto voice is of the lowest tone in female voices, and nearly corresponds in range to the bass voice in men, only, the sounds produced are an octave higher. Music is written for it in both the alto and treble clefs, more frequently the latter. Its range is between

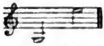

The mezzo-soprano, often confounded with the contralto voice, has a compass of an octave and a fifth or sixth.

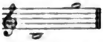

It is more flexible than the contralto, though less brilliant in tone than the soprano voice, which latter has effective and useful notes between

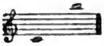

and even beyond these limits, in exceptional cases.

Voicing. The regulation of the tone and power of an organ-pipe. [Organ.]

Voix (*Fr.*) The voice.

Voix céleste (*Fr.*) [Vox Angelica.]

Volante (*It.*) Flying, applied to the execution of a rapid series of notes, either in singing or playing.

Volata (*It.*) A run, or division; a light and rapid series of notes.

Volkslied (*Ger.*) A popular song. [Lied.]

Voll (*Ger.*) Full; as, *Vollgesang*, with the full chorus; *mit vollem Werk*, with the full power.

Vollstimmig (*Ger.*) Full-voiced, full-toned.

Vollstimmigkeit (*Ger.*) Fulness of tone.

Volonté (*Fr.*) Will, pleasure, *à volonté*, at will.

Volta (*It.*) Turn, or time; as, *Una volta*, once; *due volte*, twice; *prima volta*, first time; *seconda volta*, second time.

Volta (*It.* and *Fr.*) An old dance. [Lavolta.]

Volteggiando (*It.*) Crossing the hands in pianoforte-playing.

Volteggiare (*It.*) To cross the hands.

Volti (*It.*) Turn; as, *Volti subito*, or abbreviated, *v.s.*; turn quickly.

Volume. A term applied to the power and quality of the tone of a voice or instrument, or of a combination of sounds.

Voluntary. An organ solo played before, during, or after any office of the Church; hence, called respectively introductory, middle, or concluding. Such solos were formerly, and are often now unpremeditated, or improvisations, as the name *Voluntary* seems to imply. Towards the end of the last century the style of the music performed had greatly degenerated; the ordinary voluntary consisted of running passages played by the right hand on a cornet-stop or flute-stop, while the left hand sustained a few soft chords by way of accompaniment. The protests of clergy and sound musicians against such performances had not such a powerful influence in eradicating them as the general introduction into this country of complete sets of pedal-keys, and the consequent study of John Sebastian Bach's compositions for the organ. Arrangements from instrumental and orchestral works are at the present time largely used as voluntaries, and although it must be said in their favour that they bring the themes of great masters under the notice of many who would

Digitized by Google

not otherwise hear them, yet it is to be feared that they tend to discourage the composition of genuine organ-music, and foster the mischievous notion that the organ is intended to be an imitation of a full band.

The custom of performing voluntaries and interludes in Church was at one time carried to an absurd extent. Not only was the Psalm-tune introduced with a "very pretty flourish," but interludes were played between each line of the tune, whether the sense was complete or not. A specimen of the style of these performances may be seen in the following, from "The Psalmes set full for the Organ or Harpsicord as they are Plaid in Churches and Chappels in the maner given out, as also with their Interludes of great variety," by D. Purcell, c. 1700.

VOLUNTARY.
St. David's Tune given out.

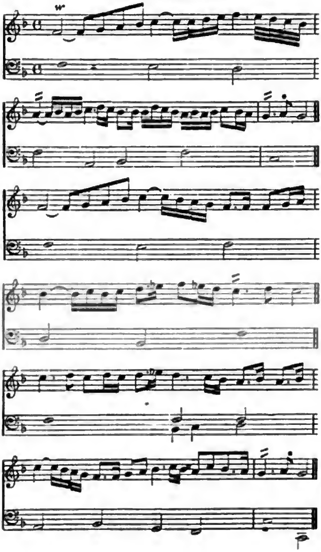

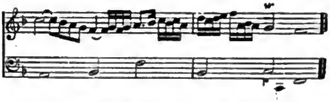

Vorausnahme (*Ger.*) (1) Anticipation. *e.g.*:

Simple form.

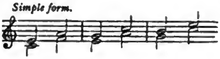

With anticipation.

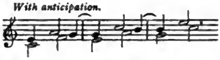

(2) Preparation of a discord.

Vorbereitung (*Ger.*) Preparation of discords.

Vorgeiger (*Ger.*) Leader. First fiddle.

Vorgreifung, Vorgriff (*Ger.*) Anticipation.

Vornalt (*Ger.*) (1) Syncopation. (2) Suspension.

Vorpause (*Ger.*) A preliminary rest.

Vorschlag (*Ger.*) Appoggiatura, beat.

Vorspiel (*Ger.*) Prelude, introductory movement, overture.

Vorspieler (*Ger.*) A leader or principal performer upon any instrument.

Vortänzer (*Ger.*) Leader of a dance.

Vorzeichnung (*Ger.*) Signature, *q.v.*

Vox (*Lat.*) (1) A voice; as *vox humana*, the human voice. (2) A part; as *tres voces*, three parts. (3) A sound; as, *inter aliquam vocem et octavam a se*, between any sound and its octave. (4) A key. [Acutæ claves.] (5) A theme; as, *vox antecedens*, the subject of a fugue.

Vox Angelica (*Lat.*) An organ stop consisting of two ranks of pipes of small scale and delicate quality of tone, one of which is tuned slightly sharp, in order to produce a wavy and tremulous sound. Called also *Voix céleste*, *unda maris*, &c.

Vox humana (*Lat.*) A reed stop in the organ, consisting of a large reed and short tube, intended to imitate the sounds of the human voice; called *voce umana* in Italian, *voix humaine* in French, and also *anthropoglossa*.

V.S. Abb. for *volti subito*, and for *violino secondo*.

Vuide (*Fr.*) Open, as *corde vuide*, open string, on the violin or violoncello.

Digitized by Google

W.

Waits, *or* **Wayghtes.** Originally certain minstrels or musical watchmen attached to the households of kings and other great persons, who paraded an assigned district sounding the hours at night. Until very recently, the Waits of the City of Westminster were regularly sworn before the "Court of Burgesses." In the "Liber niger domus regis" quoted in Rymer's "Fœdera," in an account of the musicians of the household of Edward IV., mention is made of "A Wayte, that nightely from Mychelmas, to Shreve Thorsdaye, pipe the watche within this courte fowere tymes; in the somer nightes iij tymes, and makyth Bon Gayte at every chamber, doare, and offyce, as well for feare of pyckeres and pillers." Many cities and towns, both English and foreign, encouraged and licensed their "waits," Exeter among other places having a regular company as early as the year 1400.

In the "Coxcomb," by Beaumont and Fletcher, we find this allusion to a neglect of duty on the part of the watchmen.

> "Where were the Watch the while? Good sober gentlemen
> They were like careful members of the City,
> Drawing in diligent ale and singing Catches."

The word was sometimes used to describe those who acted as the town musicians but who did not do duty as watchmen. It was also given to any company of performers when employed as serenaders.

The instruments used were a species of hautboys, called also shawms, and from their use "waits."

Dr. Busby in his Musical Dictionary [1800] says "The noun formerly signified hautboys, and (which is remarkable) has no singular number. From the instruments, its signification was, after a time, transferred to the performers themselves."

In *The famous history of Dr. Faustus*, the word is in a list of musical instruments. "Lastly was heard by Faustus all manner of instruments of music—as organs, clarigolds, lutes, viols, citterns, *Waits*, hornpipes, anomes, harps, and all manner of other instruments of music." Butler in his "Principles of Musick in singing and setting, with the two-fold use thereof [Ecclesiastical and Civil]" published in 1636, identifies waits with the hautboy; and Mr. H. Coleridge adds his testimony founded upon a passage in the romance of *Kynge Alysander and Sir Eglamore*, that the waits were wind instruments.

A writer in the "Gentleman's Magazine," in 1756, describing the method of constituting freemen at Alnwick, says:—"They (the proposed freemen) are generally met by women dressed up with ribbons, bells, and garlands, who welcome them with dancing and singing, and are called *timber-waits*, probably a corruption of *timbrel waits*, players on timbrels."

At the present day the waits are detached bodies of impromptu musicians who make night hideous for three weeks before Christmas, with wretched performances of indifferent melodies. The waits or town musicians in Germany were called *Zinkenisten*—players of the *Zink*, a coarse reed instrument.

Waldflute, Waldflöte, Waldpfeife (*Ger.*) Forest flute. An organ stop of 4 ft. pitch consisting of open wood pipes. [Flute.] Waldquinte or Waldflötenquinte is a similar stop one fifth higher in pitch.

Waldhorn (*Ger.*) A hunting horn. A French horn.

Walnika, *or* **Walynka** (*Russ.*) A kind of bagpipe used among the Russian peasantry.

Waltz, Valse (*Fr.*), **Valce** (*It.*), **Walzer** (*Ger.*) A dance said to have originated in Bohemia, now of almost universal adoption. The time is of triple measure in crotchets or quavers, and consists of eight or sixteen bar phrases. Modern waltz writers frequently add to the original dance-form an introduction and coda. The "Vienna" waltz is characterised by a rapid movement and strict unbroken time. Ländler are slower and more dignified than the waltz.

"Classical waltzes" are compositions in waltz-form intended for set pieces, not for dance tunes. In them, greater scope is given to the composer and performer than is compatible with the rhythm of the dance.

Walze (*Ger.*) A roll. A symmetrical run or division, *e.g.*:

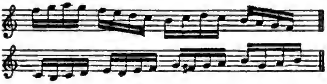

Wasserorgel (*Ger.*) Hydraulic organ.

Waves of sound. [Acoustics § 3.]

Wayghtes. [Waits.]

Wechselgesang (*Ger.*) Responsive or antiphonal song.

Wechselnote (*Ger.*) *Nota cambiata* (*It.*) In counterpoint, proceeding from a discord by a skip, *e.g.*:

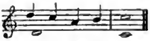

Weich (*Ger.*) (1) Minor. (2) Soft and rich.

Weihnachtslieder (*Ger.*) Carols.

Weissenote (*Ger.*) White note; *nota bianca* (*It.*), a minim.

Weiteharmonie (*Ger.*) Dispersed harmony.

Welsh harp. [Harp.]

Wenig (*Ger.*) Little, as *ein wenig stark*, rather loud.

Wesentlich (*Ger.*) Essential; as *wesentliche Septime*, essential, or dominant seventh.

Wheel. The refrain or burden of a ballad.

> "You must sing, *Down, a-down, an you call him* a-downa.
> O, how the *wheel* becomes it!"—Shakspeare.

Whiffler. A wand-bearer to head a procession. A fifer. "Whifflers originally headed armies or processions as fifers or pipers; in process of time the word *whifflers*, which had always been used in the sense of a *fifer*, came to signify any person who went before in a procession." Douce's "Illustrations of Shakspeare."

Whistle. (1) To make a musical sound with the lips and breath without using the vocal cords; the hollow of the mouth forming a resonance-box. The pitch of whistling is an octave higher than is generally supposed.

Whistle, *tin whistle, penny whistle.* The common wood or tin *flûte-à-bec* having six ventages. Its scale is:

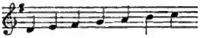

when softly blown, which may of course be considerably extended upwards by increasing the pressure of wind, as in other flutes.

Whole note. A semibreve.

Wiederholung (*Ger.*) Repetition.

Wiegenlied (*Ger.*) A lullaby, a cradle-song.

Wie oben (*Ger.*) As above.

Wind chest. Windlade (*Ger.*) [Organ Construction, § 4.]

Wind gauge. [Organ Construction, § 20.]

Wind trunk. [Organ Construction, § 4.]

Wind instrument. A musical instrument whose sounds are produced by the breath of the player, or by means of a pair of bellows.

Wirbel (*Ger.*) (1) A peg of a violin, &c. (2) The stopper of a closed organ pipe. (3) *Wirbelstock*, a sound-board into which pegs are fixed. (4) *Wirbeltanz*, a whirling or circular dance.

Wolf. (1) The bad effect produced when playing in certain keys on an organ tuned to "unequal temperament." It is well known that tempered thirds are more distressing to the ear when heard from instruments of continuous-tone like the organ and harmonium than from pianofortes, &c. To obviate this difficulty, tuners of organs formerly made certain of their thirds untempered, that is, true to nature, in the ratio 4:5. Thus C♯ was made a true third above A; E♭ a true third below G; F♯ a true third to D; G♯ a true third to E; B♭ a true third below D, *e.g.*:

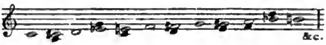

The common chords, therefore, of C, D, E♭, E, F, G, A, were perfectly in tune. But as will be seen from the above, the following would be heard for the chord of some remoter keys:

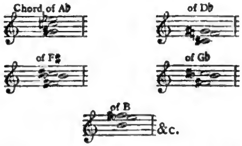

The divergence of these intervals from just pitch was painfully apparent. Modern organs are almost without exception tuned to equal temperament. [Temperament.]

(2) Some particular note often found on a violin, violoncello, or other stringed instrument, the intonation of which is not true.

Wood wind, *or* **Wood wind-band.** The flutes, oboes, clarinets, bassoons, and instruments of their nature, in an orchestra.

Wood stops. Organ stops, the pipes of which are of wood.

Wrest (*Old Eng.*) A tuning-hammer.

Würde (*Ger.*) Dignity. *Mit Einfalt und Würde*, with simplicity and dignity.

Digitized by Google

X——Y——Z.

Xänorphica. A key-violin. An instrument somewhat like the tetrachordon, invented by Röllig (1761-1804), the sounds of which were produced by bows set in motion by a pedal, and acted upon by keys.

Xyloharmonica, Xylosistrum. An harmonicon consisting of graduated blocks of wood, struck with hammers acted upon by keys. Instruments of this class were made by Röllig (1761-1804), and by Uthe (1810).

Xylophone.
Xylorganum. } [Ligneum Psalterium.]

Yang Kin. A Chinese instrument furnished with brass strings, which are struck with two small hammers, like a dulcimer.

Yo. An Indian flute.

Yu. An interval of the Chinese scale. The ancient Chinese divided the octave into twelve equal parts, like the semitones of our chromatic scale, which were called *lu*. Their scale, as commonly used, consisted, however, of only five notes, which were called koung, chang, kio, tché, and yu, and which corresponded to our F, G, A, C, D; koung or F being considered to be the normal key.

Yue Kin. A guitar called by the Chinese, "Moon Guitar." [Guitar.]

Za. Formerly a *solfeggio* name for B♭.

Zählzeit (*Ger*). Time-measure, time-unit.

Zambomba (*Sp.*) A common toy instrument in Spain, formed of an earthen pot tightly covered at its open end with a parchment, into which is inserted the stem of a reed. When the reed is rubbed up and down with rough or moist fingers, the vibrations are transferred to the air enclosed in the vessel, and a hollow rumbling sound is produced.

Zampogna (*It.*) A bagpipe in use among Italian peasants. The name is supposed to be a corruption of *Symphonia.* [Bagpipe.] A rough-toned reed instrument without a bag is also called Zampogna or Zampugna.

Zaner. [Zummarah.]

Zanze. A negro instrument. Known also by the names of mambira, ambira, marimba, ibeka, vissandschi, in different parts of Africa. It consists of a wooden box, on which a number of sonorous slips of wood or tongues of iron are fixed, in such a position as to admit of their being made to vibrate by pressing them down with the thumb or with a stick. The compass of two such instruments is given in Engel's "Music of the most ancient nations," pp. 12, 13.

Zapatadeo (*Sp.*) A dance in which the heel is struck violently on the ground to mark the rhythm.

Zarabanda (*Sp.*) [Saraband.]

Zarge (*Ger.*) The sides of instruments like the violin or guitar.

Zart (*Ger.*) Soft, delicate; as, *mit zarten Stimmen*, with soft stops; *sehr zart*, very soft.

Zartflöte (*Ger.*) A delicate-toned flute.

Zärtlich (*Ger.*) Softly, delicately.

Zarzuelas (*Sp.*) A sort of drama, said to have been first produced at Zarzuela in the time of Philip IV.

Zeichen (*Ger.*) A musical sign, note, or character.

Zeitmass (*Ger.*) Time, measure.

Zèle (*Fr.*), **Zelo** (*It.*) Zeal, energy.

Zelosamente (*It.*) Zealously, ardently, earnestly.

Zeloso (*It.*) Zealous, energetic.

Zergliederung (*Ger.*) *Lit.* dissection. The reduction of a subject to its component figures in order to expand it by their repetition.

Zerstreut (*Ger.*) Dispersed, spread, open, as, *zerstreute Harmonie.* dispersed or open harmony.

Ziemlich (*Ger.*) Moderately, as *ziemlich langsam*, moderately slow.

Ziffern (*Ger.*) To cypher.

Ziganka. A dance popular among the Russian peasantry, similar in its figures to the English country-dance. The tune is lively, and the accompaniment is usually a "burden," or "bag-pipe bass," as shown in the following example:

ZIGANKA.

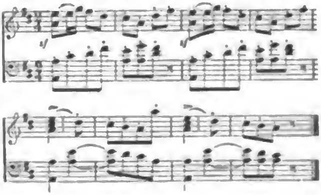

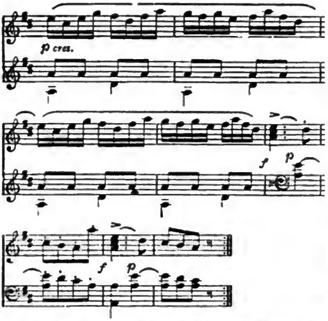

Zikrs. Religious dances of the dervishes in Egypt.

Zingaresca. A gipsy song or dance.

Zinken (*Ger.*) **Cornetto curvo** (*It.*) [Cornet, § 1.]

Zither (*Ger.*) Cither. A flat stringed instrument, placed upon a table or on the knees, having brass strings played with the thumb of the right hand, while the melody is brought more prominently out by the use of a plectrum. A bow is sometimes employed.

Zögernd (*Ger.*) *Ritardando.*

Zoppa (*It.*) [Alla zoppa.]

Zufällige (*Ger.*) Accidentals.

Zufolo (*It.*) A flageolet or whistle.

Zug (*Ger.*) (1) A drawstop in an organ. (2) A pedal of a pianoforte. *Zugwerk*, the mechanical appliances of an instrument.

Zuklang (*Ger.*) Unison, consonance.

Zummarah. An Egyptian wind instrument. [Bassoon.]

Zunehmend (*Ger.*) *Lit.* adding, *crescendo.*

Zunge (*Ger.*) (1) The tongue of a reed-pipe. (2) The metal tongue in the reed of a harmonium.

Zurna. A Turkish wind instrument, similar in character to the oboe.

Zurückhaltung (*Ger.*) Retardation.

Zusammengesetzt (*Ger.*) Compound. *Zusammengesetzte Taktarten*, compound time.

Zusammenklang, Zusammenlaut (*Ger.*) Harmony, consonance.

Zweifach (*Ger.*) (1) Two-fold. (2) Having two ranks of organ pipes. (3) Compound intervals.

Zweifacher (*Ger.*) A dance having alternate triple and duple time. It is also called *Grad und Ungrad.*

ZWEIFACHER.

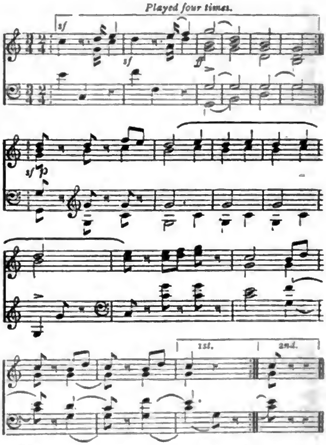

Zweigestrichen (*Ger.*) Having two strokes. [Pitch.] [Tablature.]

Zweiunddreissigtheil-Note (*Ger.*) A demisemiquaver.

Zweivierteltakt (*Ger.*) Two-four time, $\frac{2}{4}$.

Zweizweiteltakt (*Ger.*) Two-two time, $\frac{2}{2}$.

Zwerchflöte (*Ger.*) The German flute. *Flauto traverso* (*It.*)

Zwerchpfeife (*Ger.*) The piccolo flute, or fife.

Zwischengesang (*Ger.*) An interpolated song.

Zwischenräume (*Ger.*) The spaces of the stave.

Zwischensatz (*Ger.*) An episode.

Zwischenspiel (*Ger.*) An interlude played between the verses of a hymn or choral.

Zwölfachteltakt (*Ger.*) Twelve-eight time.

Zymbel (*Ger.*) [Cymbal.]

Digitized by Google

Digitized by Google

Digitized by Google

Digitized by Google

$9.95 in USA

Essential Dictionary of Music Notation

A Practical and Concise Source for Students to Professionals

Take the mystery out of music notation. Whether you're using a pencil to sketch a melody or a computer to score a symphony, this multi-purpose quick reference is for you.

Every aspect of notation is covered in a versatile, easy-to-use format. From beginning music students to seasoned engraving professionals, everyone notating music will love this dictionary!

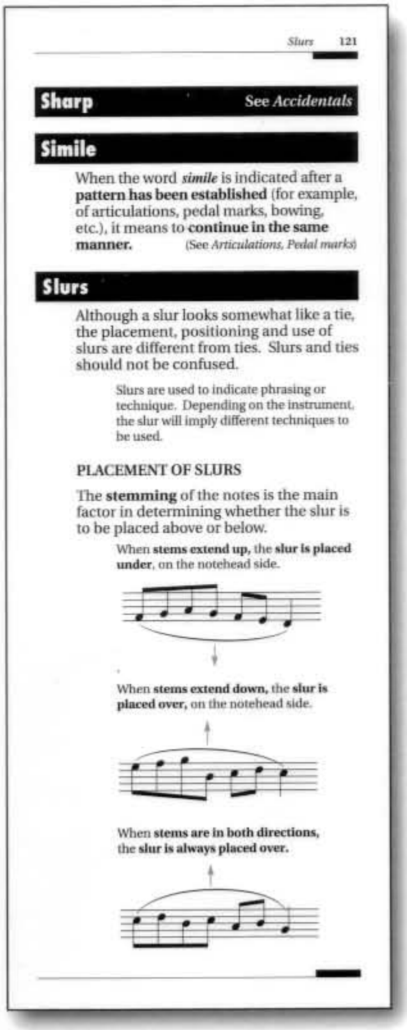

Slurs 121

Sharp See *Accidentals*

Simile

When the word ***simile*** is indicated after a **pattern has been established** (for example, of articulations, pedal marks, bowing, etc.), it means to **continue in the same manner.** (See *Articulations, Pedal marks*)

Slurs

Although a slur looks somewhat like a tie, the placement, positioning and use of slurs are different from ties. Slurs and ties should not be confused.

Slurs are used to indicate phrasing or technique. Depending on the instrument, the slur will imply different techniques to be used.

PLACEMENT OF SLURS

The **stemming** of the notes is the main factor in determining whether the slur is to be placed above or below.

When **stems extend up,** the **slur is placed under**, on the notehead side.

When **stems extend down,** the **slur is placed over,** on the notehead side.

When **stems are in both directions,** the **slur is always placed over.**

- Easy–to–use dictionary format
- Concise text is clearly illustrated
- Key words and concepts are highlighted
- Over 75 topics spanning 158 pages

Alfred Publishing Co
16380 Roscoe Blvd.
P.O. Box 10003
Van Nuys, CA 91410

17241

AN ALFRED HANDY GUIDE

Essential Dictionary of Music Notation

mio celo egressio eius et oc
cursus eius alleluya.
Celi enarrant. primo
est passio
nem su

The most practical and concise source for music notation available for students to professionals.

TOM GEROU • LINDA LUSK

Essential Dictionary of Music Notation

The most practical and concise source for music notation

TOM GEROU
LINDA LUSK

- Easy-to-use format, including cross-referencing
- For any musician, whether using pencil or computer
- Emphasis on correct and current notation practices
- Numerous examples, kept simple and to-the-point

Thanks to Susy Christiansen, Mitchell Newman, Greg Plumblee and Patrick Wilson

Cover design: Carol Kascsak

Cover art, center photo: From an Antiphoner,
or book of music for the choir services.

Copyright © Sotheby's

Copyright © MCMXCVI by Alfred Publishing Co., Inc.
All rights reserved. Printed in USA.

ISBN 0-88284-768-6

APPROACHING MUSIC NOTATION

Music notation is a language and, as such, it must be communicable. The closer the notation adheres to common rules and guidelines, the more successfully the music will be performed. The rules and guidelines presented here will make the notation easier to read and will help to make the musical intent clear. Although these rules are not inflexible, they should not be ignored.

For music notation, the first priority is always clarity. Choices need to be made that are solely dependent on the situation; a rule may not provide the most desirable answer to a problem. (It would be an endless task to list every musical exception to a rule.) Problems are usually solved by deciding which rule is more flexible (least important to clarity).

Although meticulous adjustment can only clarify and beautify the notation, sometimes such attention to detail is not practical. All decisions must depend on a wide variety of situations. An understanding of the rules allows these choices to be made wisely—the notation can only be as good as the knowledge at hand.

ABOUT THIS BOOK

The *Essential Dictionary of Music Notation* is in an easy-to-use format, which includes cross referencing. At the top of each page, running heads show exactly which topic is being treated on that page. Bold treatments of text highlight important points for clarity and quick reference.

Text is accompanied by numerous examples directly related to the topic being discussed. These examples are kept as simple and complete as possible, with an emphasis on presenting correct and current notation practices. Obsolete rules are mentioned only when necessary.

The focus of this book has been narrowed, avoiding the discussion of orchestration or music theory, in order to present as much information as possible that is directly related to the actual *notation* of music. The serious musician is encouraged to seek out other sources that more thoroughly address the subjects of orchestration and music theory.

Whether you are using a pencil or a computer, all principles within this book are applicable. No computer software or specific tool is referenced, in order to make this a useful guide for any musician.

An *Index of Topics* is provided on the last page.

A cappella

Vocal music that is to be sung **without accompaniment** (a cappella) often has a piano part for rehearsal only.

The piano part may be full size or cue size (see *Cue notes*). If cue size, the staff should be cue size also. (Cue-size notes on a full-size staff are not as easy to read.)

The indication ***for rehearsal only*** should be clearly placed above the piano part.

A due (a2)

See *Sharing a staff*

A tempo

Place above the staff (sometimes within the grand staff) to indicate a return to the normal tempo. Used after tempo alterations such as ***ritardando, accelerando, più lento*** or ***ad libitum.*** (See *Tempo marks*)

Accents

See *Articulations*

Accidentals

Temporary alterations to the pitch of a note.

THE 5 SYMBOLS

If appearing in the music, apart from use in the key signature, these symbols are called *accidentals.*

♯ **Sharp**
Raises a note one half step.

𝄪 **Double-sharp**
Raises a note two half steps.

♭ **Flat**
Lowers a note one half step.

𝄫 **Double-flat**
Lowers a note two half steps. Notice that the flats touch, forming one symbol.

♮ **Natural**
Cancels any of the above, whether indicated in the key signature or as an accidental.

ONE ACCIDENTAL PER NOTE

Do not use multiple accidentals on one note. (A double-flat is a single accidental.

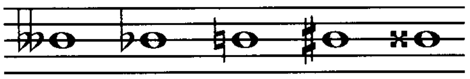

Accidentals shown altering one note from lowest to highest.

The lowest a note can be lowered is a 𝄫.

The highest a note can be raised is a 𝄪.

WITHIN A MEASURE

Any accidental is **cancelled when a new accidental appears** on the like pitch.

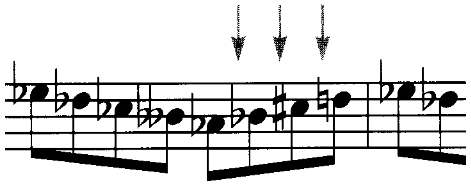

Notice that only the flat (not ♮♭) is needed to change from a double-flat to a flat.

An accidental affects all **subsequent notes of the same pitch** within the measure.

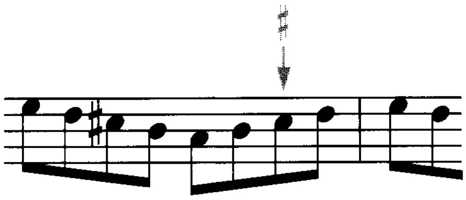

An accidental **must be added again** in the following measure.

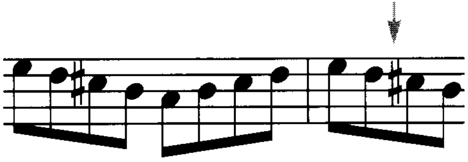

An accidental **carries over to a tied note** extending from one measure to the next.

The accidental will reappear only when the note is re-struck.

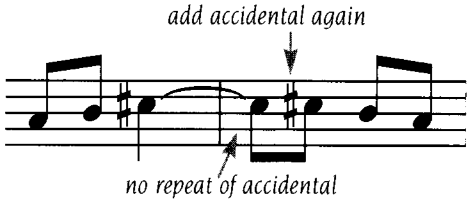

OCTAVE ALTERATIONS

Accidentals **do not apply if an *8va*** sign is added; repeat the accidentals.

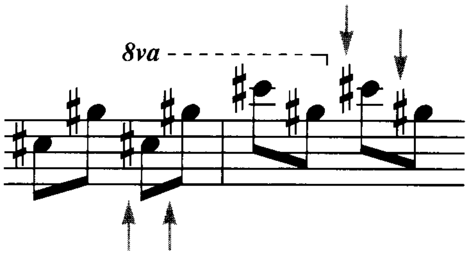

COURTESY ACCIDENTALS

Used as **reminders of the key signature** when an accidental has been used in the previous measure, or to **avoid ambiguity** in certain situations.

Use a courtesy accidental as a **reminder** to return to the original pitch.

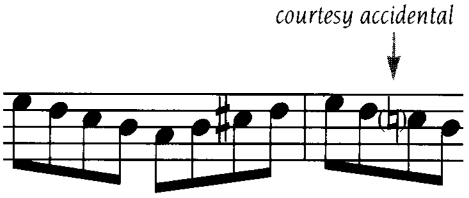

Courtesy accidentals may or may not be enclosed in **parentheses.** The parentheses show that it *is* a courtesy accidental, and help avoid confusion about the key signature. However, if many courtesy accidentals are used, and the piece is also crowded, the parentheses could make it look cluttered and more complicated. For this reason some editors choose not to use parentheses at all.

If a chord has more than one courtesy accidental, **each accidental gets its own pair of parentheses.**

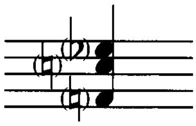

It is best to **add a courtesy accidental for octave shifts** although the notes of the same pitch in different octaves are considered separately.

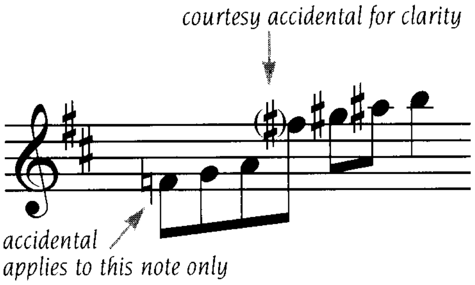
courtesy accidental for clarity
accidental applies to this note only

GRACE NOTES WITH ACCIDENTALS

The accidental placed on a grace note will be smaller (cue size), like the grace note.

The accidental **affects all subsequent notes of the same pitch** within the measure.

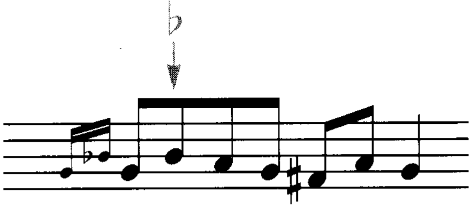

Depending on the clarity of the situation, **courtesy accidentals are often needed** to avoid confusion.

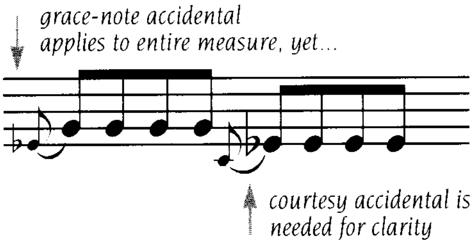
grace-note accidental applies to entire measure, yet...
courtesy accidental is needed for clarity

ALIGNMENT FOR INTERVALS (2 Notes)

For intervals of a **2nd through a 6th,** place the upper accidental closest to the note, the lower accidental to the left.

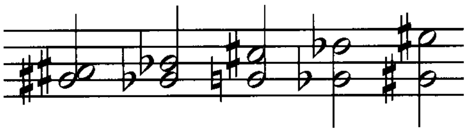

When intervals are **greater than a 6th,** accidentals align vertically.

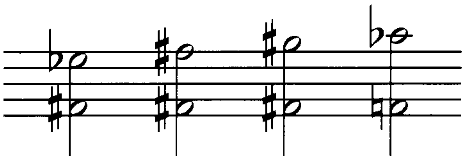

For the **interval of a 6th,** if the two accidentals don't collide they may be aligned vertically.

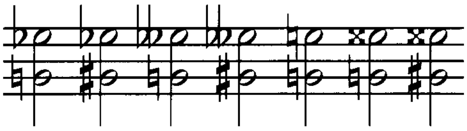

ALIGNMENT FOR CHORDS (3 Notes)

When outer notes are a **6th or less,** the upper accidental is closest to the note, the lower accidental is placed left, and the middle accidental is placed farthest left.

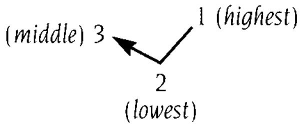
(middle) 3
1 (highest)
2
(lowest)

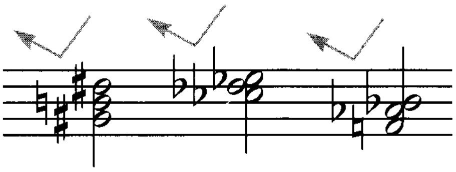

When outer notes are **greater than a 6th,** upper and lower accidentals are aligned closest to the notes—the middle accidental is placed to the left.

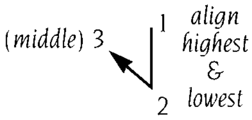
(middle) 3
1
align highest & lowest
2

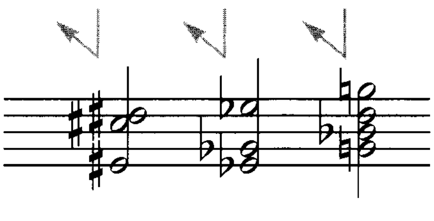

ALIGNMENT FOR CHORDS (More than 3 Notes)

When dealing with accidentals for complex chords, rules are treated more as guidelines or suggestions. Keep the arrangement of the accidentals as compact as possible.

Make the arrangement of accidentals as **easy to read as possible** for the situation.

Align the highest and lowest accidentals whenever possible.

The **center accidentals are usually arranged diagonally** from highest to lowest.

Align accidentals for octaves whenever possible.

Accidentals for the 2nd should usually be **shaped like the 2nd** (if the outer notes are greater than a 6th).

The upper accidental is placed closest to the note, the lower accidental to the left.

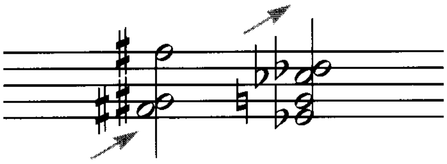

Accidentals are always placed before the entire note structure. (Do not place an accidental between notes that are played together, even if they are stemmed in opposite direction.)

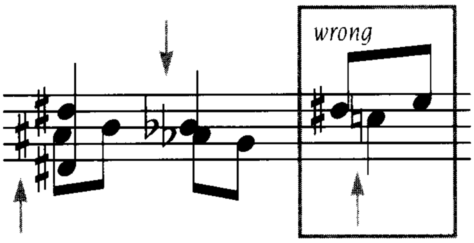

TWO PARTS ON A STAFF

For multiple notes played by **one instrument,** an accidental carries through the measure. Cancellations must be applied.

one instrument—
the accidentals carry through the measure

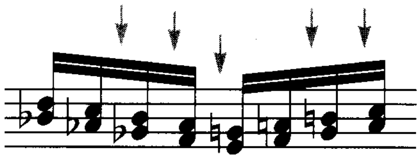

Parts on the same staff that are played by **more than one instrument** (including vocal parts) are treated individually. Repeat accidentals in the same measure if the voices cross.

two instruments sharing a staff—
the accidentals are considered separately

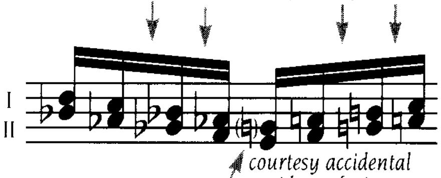

courtesy accidental
avoids confusion

CHROMATIC PASSAGES

In general, use **sharps when ascending, flats when descending.**

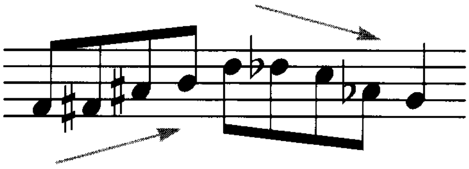

When choosing accidentals, consider the context of the music, such as key signature, harmony and voice leading.

In the chord below, both a sharp and a flat are required (C♯, not D♭; E♭ not D♯).

augmented 6th chord

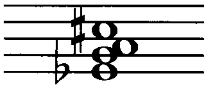

ACCIDENTALS WITH TRILLS (See *Trill*)

ACCIDENTALS IN HARP MUSIC

(See *Pedal marks*)

Alternating instruments

CONCERT ORCHESTRA / BAND

1. To alternate instruments, enough **time must be given** to allow the change.
2. During the measures of rest the indication ***"Muta in [second instrument]"*** is indicated **below** the staff as soon as the first instrument stops.
3. Any change to the **key signature** occurs after the barline following the rest, before the new instrument begins.
4. The **new instrument is indicated** above the entrance of the instrument.

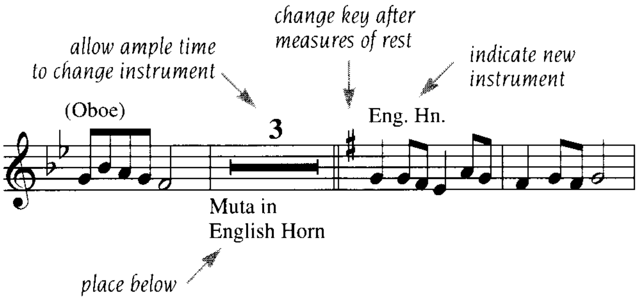
allow ample time to change instrument
change key after measures of rest
indicate new instrument
(Oboe)
3
Eng. Hn.
Muta in English Horn
place below

INCORPORATING CUES

1. Cue notes should be **added after many measures of rest,** to help the performer make a correct entrance.
2. The cued instrument is **indicated above the passage** in a smaller type size.
3. The **key signature changes at the beginning** of the cue.
4. **Full-size rests** are indicated below (or above, depending on stem direction).

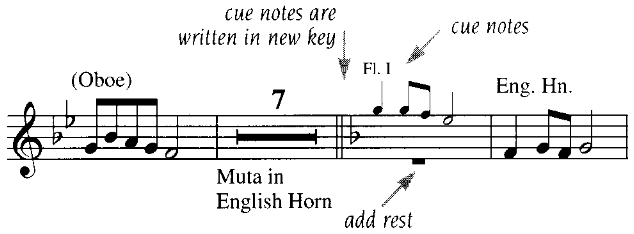
cue notes are written in new key
cue notes
(Oboe)
7
Fl. I
Eng. Hn.
Muta in English Horn
add rest

(See *Cue notes*)

JAZZ ENSEMBLE

All rules above apply except that "muta in [...]" is written **"Change to [new instrument]"** and is placed **above** the staff.

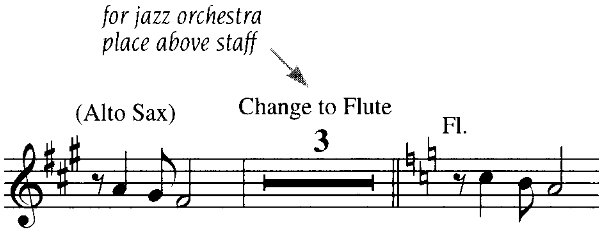
for jazz orchestra place above staff
(Alto Sax)
Change to Flute
3
Fl.

COMMONLY ALTERNATING INSTRUMENTS

The following is a listing of commonly alternating instruments in a standard orchestra:

	Alternate
Flute III	Piccolo
Oboe III	English horn
Clarinet III	Bass clarinet
Bassoon III	Contrabassoon

In smaller orchestras, the second player alternates in all cases above.

Arpeggio signs

A vertical wavy line before a chord or interval indicates an arpeggio (rolled chord).

ARPEGGIO DIRECTION

The chord is to be played quickly from **bottom to top** (no direction arrow).

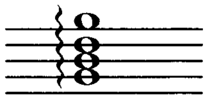

played upward
(no direction arrow)

If the arpeggio is to be played from **top to bottom, an arrowhead is added** to the bottom of the wavy line.

played downward

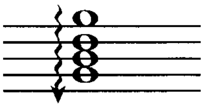

When the arpeggio returns to normal (bottom to top) or alternates with an arpeggio played downward, an arrowhead is added to the top of the wavy line.

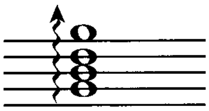

played upward

The arpeggio sign **should not extend much past the chord** affected.

PLACEMENT WITH ACCIDENTALS

Arpeggio signs are always **placed before accidentals**.

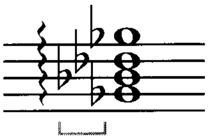

allow space for accidentals

FOR KEYBOARD INSTRUMENTS

The arpeggio sign **can extend across both staves.** The right hand begins after the left hand has completed the chord.

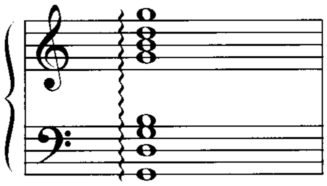

played upward, from bottom to top

If the wavy line is broken between staves, both hands begin the arpeggio simultaneously.

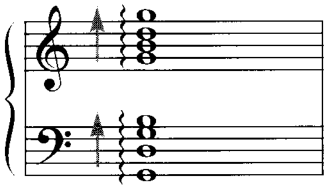

hands start simultaneously

Articulations

Articulations are symbols used to indicate *how* a note or chord is played, alone or in relation to other notes and chords.

There are five main articulations: ***staccatissimo, staccato, tenuto, accent, marcato.***

An articulation's **influence on a note or chord varies** depending on its context. A staccato note in a slower tempo is not played as short as it would be played in a faster tempo. An accent is softer in the dynamic ***p*** than in the dynamic ***f***.

One articulation applies to all notes of an interval or chord (on one stem).

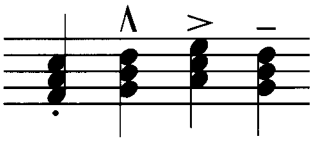

DURATIONAL ARTICULATIONS

There are three articulations that **affect the duration** of a note or chord: *staccatissimo* (▾), *staccato* (.) and *tenuto* (–).

The following illustrates durational articulations compared to an unmarked note, from shortest to longest.

SHORTER → LONGER

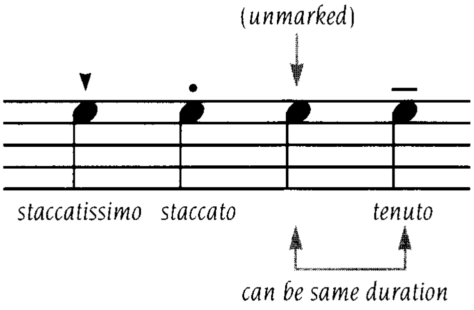

Staccatissimo

The shape varies slightly among engravers, but is always wedge-shaped. It indicates that a note or chord is to be played as **short as possible.**

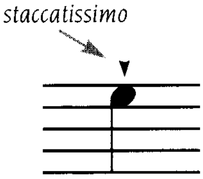

Used only with note values of a quarter note or less.

The *staccatissimo* is **centered on the notehead**. Placement is in the next space from the notehead, whether the note is on a line or space.

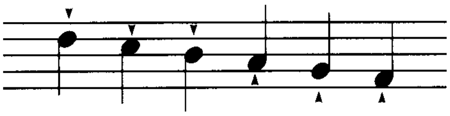

Center on the stem for opposite stem direction.

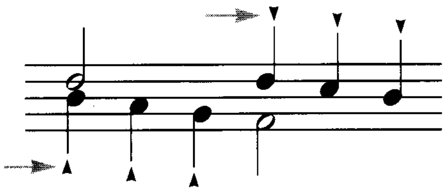

Staccato

A dot above or below a note or chord indicates that the note or chord is to be played **short.**

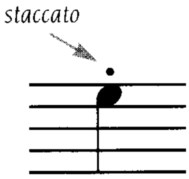

Used with note values of a quarter note or less.

The *staccato* is **centered on the notehead.** Placement is in the next space from the notehead, whether the note is on a line or space.

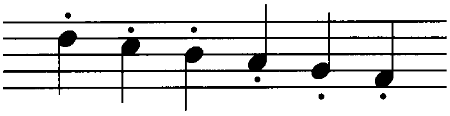

Center on the stem for opposite stem direction.

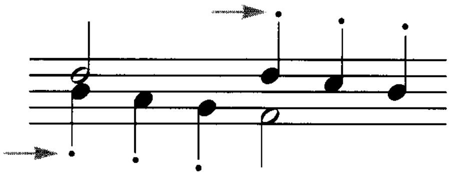

Tenuto

A short, heavy line above or below a note or chord, **thicker than the staff lines.**

A tenuto indicates that a note or chord is to be **held for its full value** (durational articulation), or the intent may be to apply a **slight stress** (articulation of force).

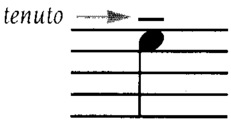

However it is interpreted, the tenuto is meant to ensure that a note or chord will be given a **certain amount of attention.**

This emphasis allows the tenuto to be **flexible in its application,** especially when combined with other articulations.

The tenuto is sometimes referred to as a *sostenuto* or *stress.*

Used with notes of any rhythmic value.

The *tenuto* is **centered on the notehead.** Placement is in the next space from the notehead, whether the note is on a line or space.

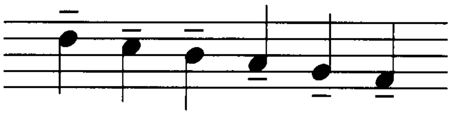

Also **center to the notehead,** not the stem, for **opposite stem direction.**

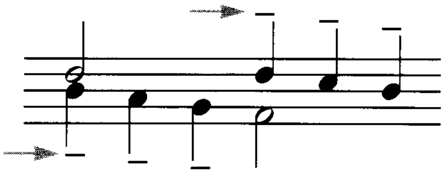

ARTICULATIONS OF FORCE

There are three articulations that affect the **force of attack** of a note or chord: the *tenuto* [–] discussed earlier, *accent* [>] and *marcato* [Λ].*

The following illustrates articulations of force compared to an unmarked note, from lesser to greater force of attack.

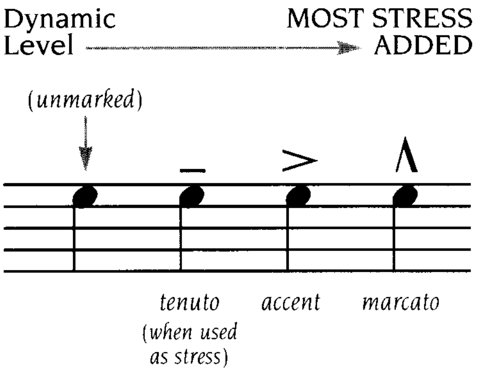
Dynamic Level
MOST STRESS ADDED
(unmarked)
tenuto
(when used as stress)
accent
marcato

*Terminology varies for the symbols [>] and [Λ]. For the sake of this discussion, the words *accent* and *marcato* have been attached to the symbols.

Accent

When an *accent* is placed over or under a note or chord, the note or chord is to be played with **more attack, more marked.**

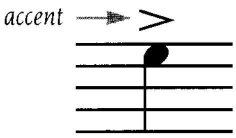

Used with notes of any rhythmic value.

The *accent* is **centered on the notehead.** Preferred placement is **outside the staff.**

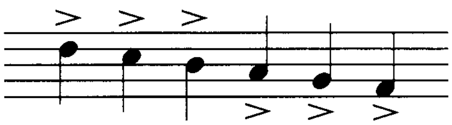

Although it is to be avoided, sometimes placing an accent **within the staff** is acceptable. A decision needs to be made whether it is preferred to have the accent closer to the notehead, or outside the staff.

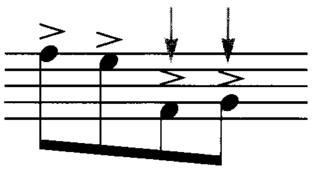

Center to the notehead, not the stem, for **opposite stem direction.**

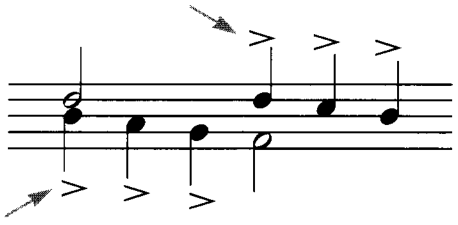

Marcato

When the *marcato* is placed over or under a note or chord, the note or chord is to be played with **even more attack, and more marked,** than an *accent.*

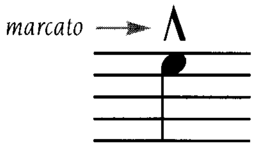

Used with notes of any rhythmic value.

The *marcato* is **centered on the notehead.** Preferred placement is **outside the staff** and ***above* the staff.**

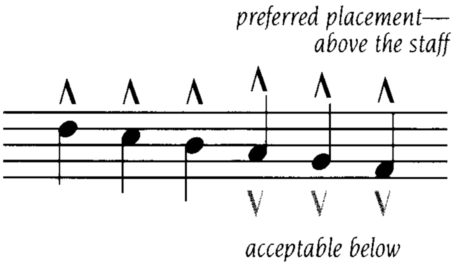
preferred placement—
above the staff
acceptable below

Also **center to the notehead** (not the stem) for **opposite stem direction.**

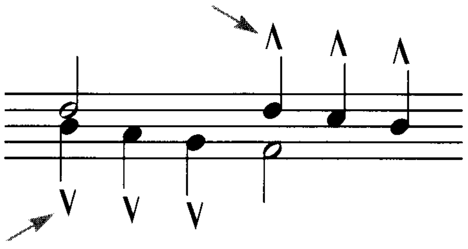

OTHER ARTICULATIONS OF FORCE

Sforzando (*sf*), *forzando/forzato* (*fz*) and *sforzato* (*sfz*) are also included in this articulation group and are interchangeable with accents and marcato accents, depending on the dynamic range.

These three articulations have the benefit of a built-in **indicator of dynamics.**

For instance, a sf should be indicated only in the dynamic ranges from ppp to f. With each increase of dynamic level, the sforzando reflects the higher dynamic level: sff in ff, sfff in fff.

The following chart illustrates the relationships between *accent* [>] and *marcato* [Λ] to *sforzando* [*sf*], *forzando* [*fz*], and *sforzato* [*sfz*] within various dynamic levels.

Dynamic Range

+	*ppp* to *f*	*ff*	*fff*
>	*sf*	*sff*	*sfff*
Λ	*fz*	*ffz*	*fffz*
Λ̲	*sfz*	*sffz*	*sfffz*

A [>] in *f* is comparable to a [*sf*].
A [Λ̲] in *fff* is comparable to a [*sfffz*].

COMBINED ARTICULATIONS

Durational articulations can be combined with articulations of force to create variations of the basic articulations.

The following examples are the most used combinations.

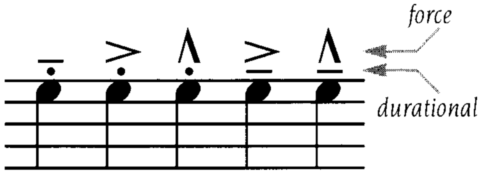
force
durational

Staccatissimo and **staccato** are never combined (they basically have the same function in varying degrees).

Accent and **marcato** are never combined for the same reason.

Staccatissimo and **tenuto** are not commonly combined.

Staccatissimo and **marcato** are not commonly combined.

Combining articulations with tenuto

The **dual function** of the tenuto—the idea of *more importance, more attention,* whether of duration or force (or both)—allows the tenuto to be combined with either articulation group.

When a tenuto is **combined with an articulation of force,** the tenuto has more of a *durational function,* but may also retain its idea of force.

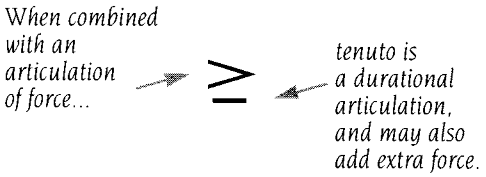
When combined with an articulation of force...
tenuto is a durational articulation, and may also add extra force.

When a tenuto is **combined with a durational articulation,** the tenuto has the function of an *articulation of force,* while retaining its durational function.

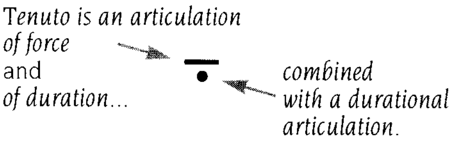
Tenuto is an articulation of force and of duration...
combined with a durational articulation.

Usual interpretation: play detached and with a slight stress.

PLACEMENT WHEN COMBINED

The following can be used as guides for the placement of combined articulations:

tenuto & staccato

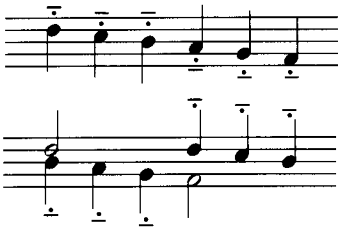

accent & staccato

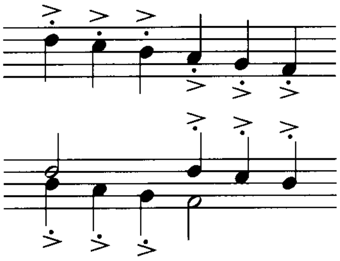

marcato & staccato

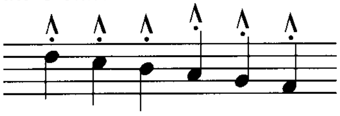

Keep the two combined articulations together (on the same side of the note) whenever possible.

Notice that when a **staccato** is combined with another articulation (and is on the same side of the note), it is **centered on the notehead,** even when it is on the stem side.

SIMILE

When an articulation (or pattern of articulations) is applied in a repetitious and predictable manner, the word ***simile*** may be used.

Simile is indicated **after the pattern has been clearly established** (usually two or three measures).

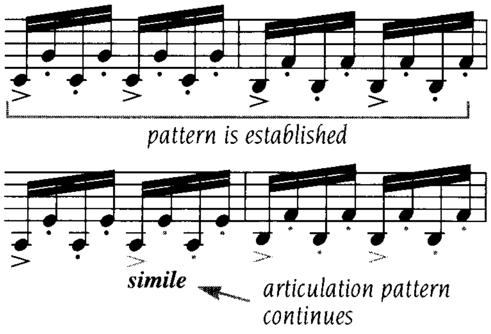

Notice that the ***simile*** was **not used at the beginning of the line,** even though the pattern had already been established.

The ***simile*** is ended when the pattern ends or when different articulations are added.

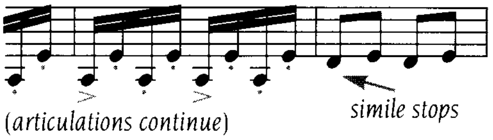

SLURS WITH ARTICULATIONS (See *Slurs*)

Augmentation dot

PLACEMENT

Always place **one dot per space** and **only in a space**—never on a staff line or on the same level as a leger line.

For a line note—place to the right of the notehead, **in the space above.**

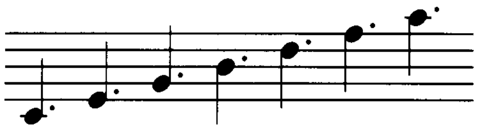

For a space note—place to the right of the notehead **in the same space** as the note.

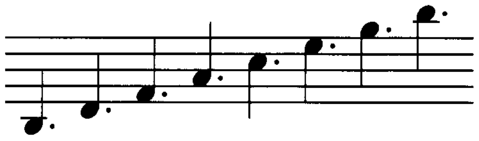

For notes with a flag—the dot is placed further right, altogether **avoiding the flag.** Never place a dot between the notehead and its flag.

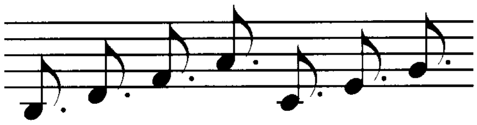

Double and triple dots are placed directly to the right of the first dot. Horizontal spacing is equal to that of the first dot. *Multiple dots should be used only in situations where they will be easily understood.*

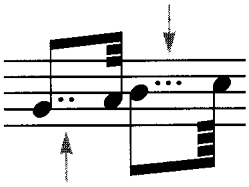

Never separate a dot from its notehead. Stem direction and creative positioning of a dotted note must be considered.

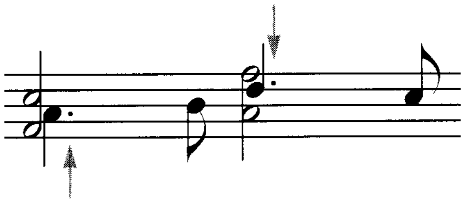

Ties always avoid the dot—place the tie clearly to the right of the dot.

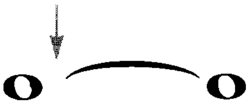

For two or more notes in a chord with the same stem direction, **always align the dots vertically.**

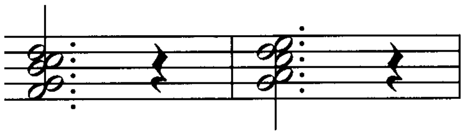

Two dots never share the same space.

FOR INTERVALS OF A SECOND

The dot for the space note is always placed in the space to the right of the notehead.

If the **line note is above,** place the dot in the space above. If the **line note is below,** place the dot in the space below.

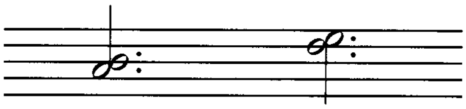

FOR TWO OR MORE PARTS

For stems up, place dots for line notes in the space above. **For stems down,** place in the space below.

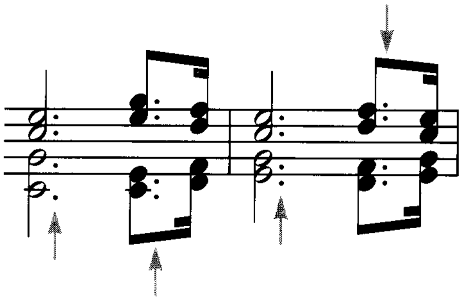

DOTTED RESTS

Rests are dotted in the same way notes are dotted, but with some restrictions on their use.

Place the dots as follows:

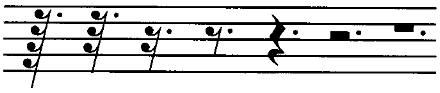

Dotted rests are used mainly to **clarify the subdivision of a measure.** (See *Meter*)

Dotted whole rests are the least used of all the dotted rests. They are never used in simple meter.

Possible use of dotted whole rest:

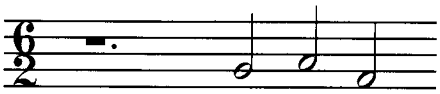

Likewise, **a dotted half rest** should not be used in simple meter.

Possible uses of dotted half rest:

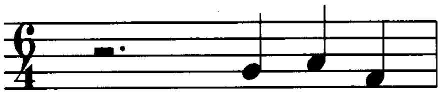

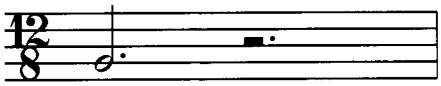

Dotted quarter rests are used in compound meters. Avoid using them in simple meters.

In $\frac{6}{8}$ time, dotted quarter rests can be used on beats 1 and 4 only.

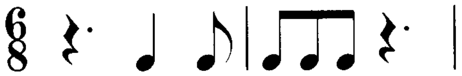

In $\frac{9}{8}$ time, dotted quarter rests can be used on beats 1, 4 and 7 only (the pulses).

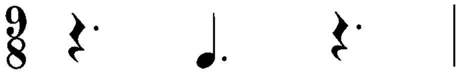

In $\frac{12}{8}$ time, do NOT use a dotted half rest across the middle of the measure. Use dotted quarter rests.

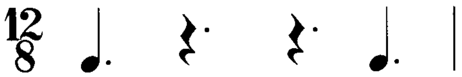

It is acceptable to use the **dotted eighth rest, dotted sixteenth rest, etc.,** for a cleaner look.

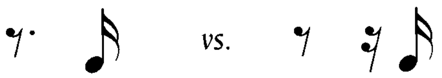

Barlines

Barlines are vertical lines that divide a staff into measures.

SINGLE BARLINE

The **thickness of a barline** is equal to, or greater than, the staff line thickness (staff lines being thicker than stems).

To help the performer follow the music from staff to staff, **avoid the alignment of barlines** from staff to staff (or from system to system).

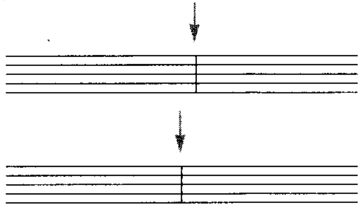

Allow a small amount of space after a barline—depending upon the density of notes within the measure.

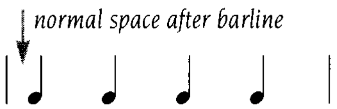

Less space may be allowed **after** the barline if the **density of notes is high.**

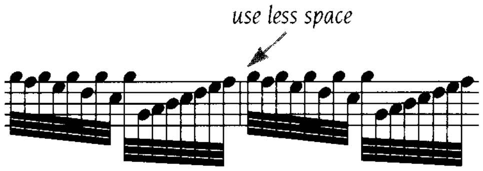

More space may be allowed **after** the barline if the **density of notes is low.**

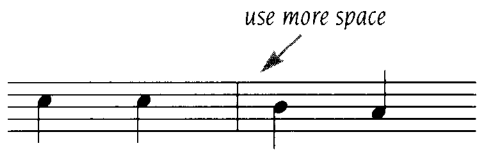

Less space may be allowed **before** the barline if **density is high.** Avoid *adding* space before the barline.

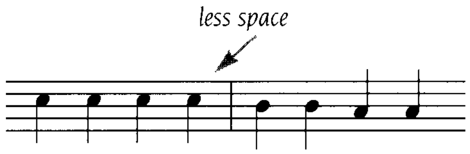

DOUBLE BARLINE
(Thin/thin)

Used to clarify the end of a section of music, such as in a *Da Capo* layout.

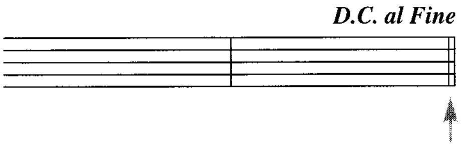

FINAL DOUBLE BARLINE
(Thin/thick)

Marks the **end of a composition, or the end of a movement** within a larger work.

Consists of a thin barline followed by a thick barline.

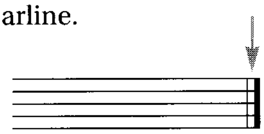

BROKEN BARLINES

Broken (dashed) barlines are sometimes used when a situation calls for the completion of a section **mid-measure,** although the **final double barline** is preferred.

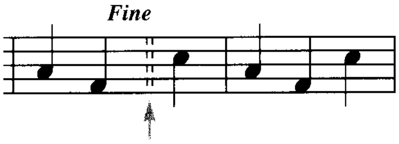

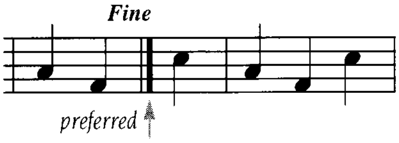
Fine
preferred

For composite meters, a broken barline is sometimes used to **clarify the meter division** within a measure.

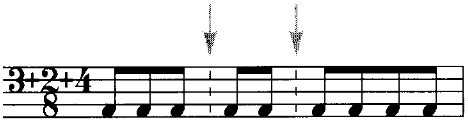

THE LEFT BARLINE

A single staff should ***never* begin** with a barline.

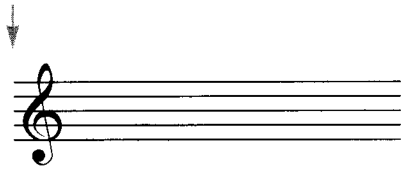

SYSTEMIC BARLINE

A systemic barline connects two or more staves together, as in the grand staff.

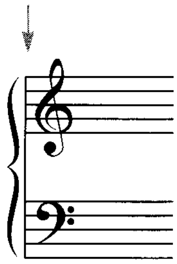

Barlines are **used to group staves of like instruments** in a score layout.

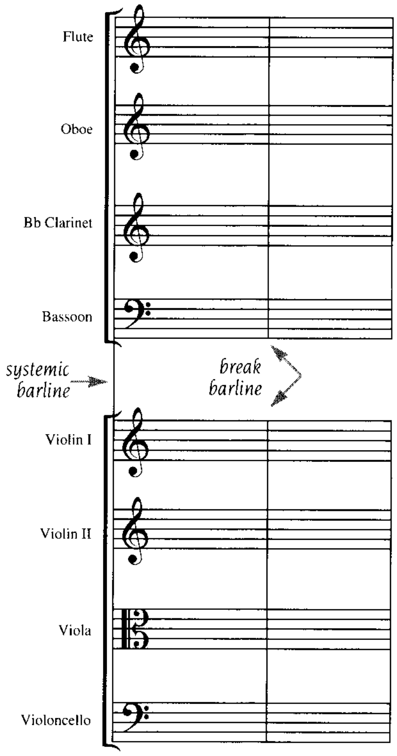
Flute
Oboe
Bb Clarinet
Bassoon
systemic barline
break barline
Violin I
Violin II
Viola
Violoncello

WITH CHANGING CLEFS

The new clef (cue size) is placed **before the barline** if the clef changes mid-staff.

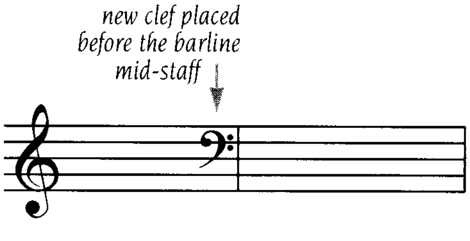
new clef placed
before the barline
mid-staff

A **courtesy clef is placed before the last barline** of the staff if the new clef begins immediately on the next staff.

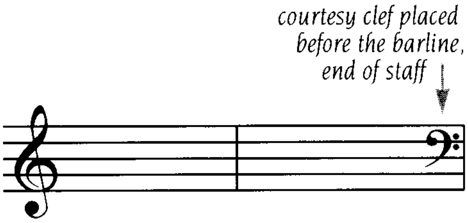
courtesy clef placed
before the barline,
end of staff

(See *Clef signs*)

WITH CHANGING KEY SIGNATURES

The new key signature is **placed after** an added **double barline** if the key signature changes mid-staff.

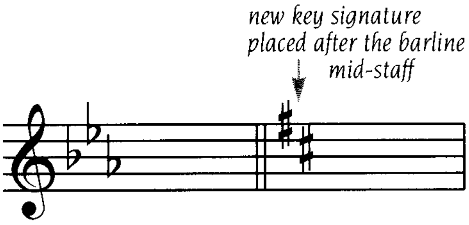
new key signature
placed after the barline
mid-staff

A **courtesy** key signature is **placed after** the last barline (add double bar) if the new key signature begins on the following staff.

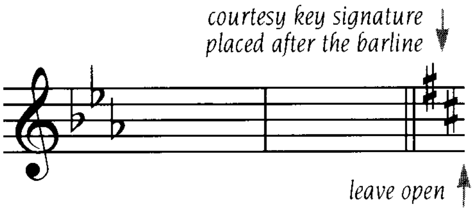
courtesy key signature
placed after the barline
leave open

Note: *A single barline may be used in both the above instances, but a double bar calls more attention to the change.*

WITH CHANGING TIME SIGNATURES

The time signature is **placed after** the barline if the time signature changes mid-staff.

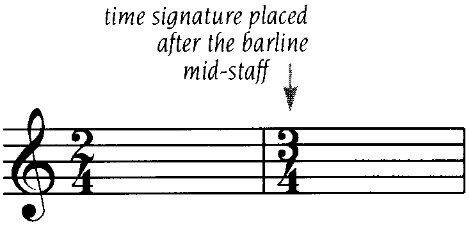
time signature placed
after the barline
mid-staff

A **courtesy** time signature is **placed after** the last barline if the new time signature begins on the following staff.

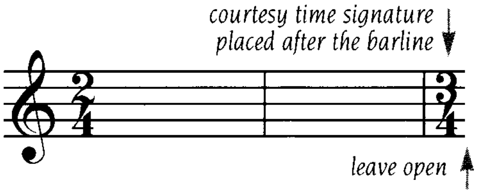
courtesy time signature
placed after the barline
leave open

WITH MULTIPLE COURTESY SIGNS

If a courtesy clef, key signature and time signature all appear together, they follow the rules for each individual courtesy change.

The order is the same as the order presented at the beginning of the piece (clef, key signature, time signature).

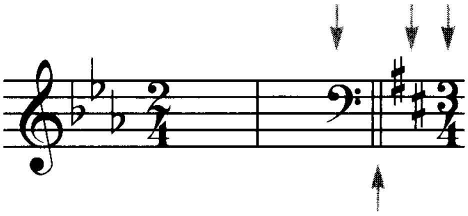

INCOMPLETE MEASURES

If a piece begins with a **pick-up,** the barline is positioned **according to the time value of the beat(s).**

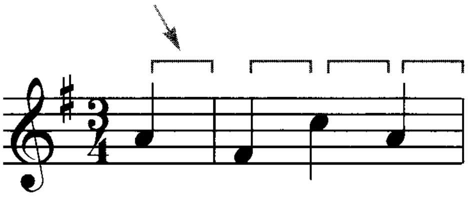

Beams

Although sometimes referred to as "balkans" or "ligatures," **"beam"** is the preferred term for the connection between two or more consecutive notes, replacing the flag.

Beams greatly simplify the reading of music, **substituting for individual flags** in groupings of **notes smaller than a quarter note.**

Because of the **ease in reading of beams,** the **use of flags in vocal music**—in relation to the lyric—has become **obsolete.**

CONNECTION TO THE STEM

Beams are connected to the side of the stem. They must be **flush to the stem,** both vertically and horizontally.

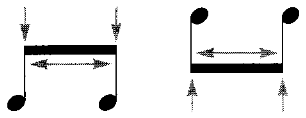

PRIMARY AND SECONDARY BEAMS

The beam that is furthest from the noteheads and connects a group of notes is called a *primary beam.* This beam **remains unbroken** throughout the grouping.

Any beam other than the primary beam is a *secondary beam.* The secondary beam **may be broken,** to divide the grouping into smaller units for easier reading.

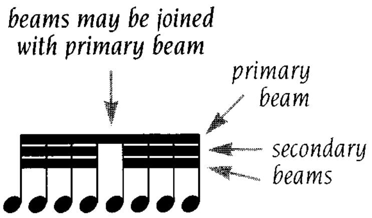

Stems extend through secondary beams and connect to the primary beam.

DOUBLE BEAMS

Normal stem length will accommodate a double beam—the stem need not be adjusted.

TRIPLE AND QUADRUPLE BEAMS

When a third or fourth beam is added below the secondary beam, **the stems must be extended** by approximately one staff space per beam.

The extension of the stems allows the space between the noteheads and the lowest beam to remain normal.

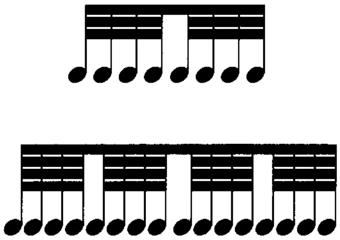

If 32nd- and 64th-note values occupy the majority of the rhythmic values, it may be better to double the rhythmic values of the entire piece and alterthe tempo.

FRACTIONAL BEAMS

A fractional beam is also a **secondary beam.**

Fractional beams are **associated with only one note.**

The **length** of a fractional beam is the same as the width of a notehead.

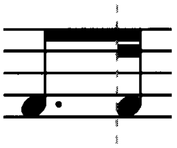

Place exactly the same as a full-length secondary beam would be placed. The beam is always *inside* the grouping, and it usually follows or precedes a beamed dotted note.

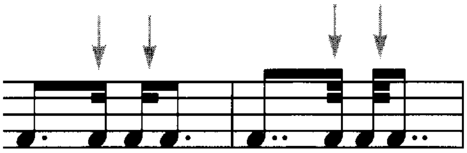

BEAMING AND METER

The basic purpose of beaming is to **connect two or more notes within the same beat (or pulse).** Meter must be considered when grouping notes with a beam. (See *Meter*)

In **any simple meter,** each beat is divisible by two; a beam may connect the two notes. Simple duple, simple triple and simple quadruple meters each have slightly different considerations.

Simple duple meter ($\frac{2}{2}$, $\frac{2}{4}$, $\frac{2}{8}$)

In some cases, the entire measure may be beamed.

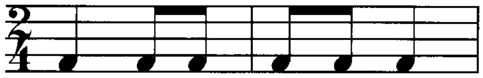

or

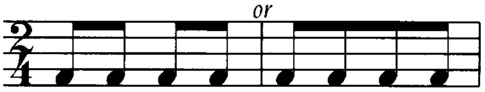

Some combinations of beams, flags and rests in simple duple meter:

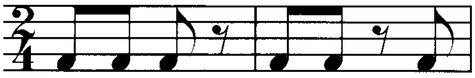

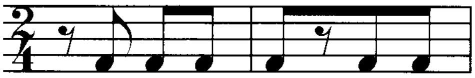

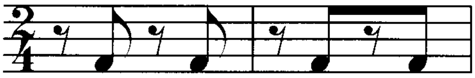

Although some contemporary engravers beam over a rest at the beginning or ending of a beamed grouping 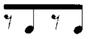 , it is not the preferred notation.

Simple triple meter ($\frac{3}{2}$, $\frac{3}{4}$, $\frac{3}{8}$)

Consideration is given to **visually maintain** the three beats within the measure. In some cases the entire measure may be beamed.

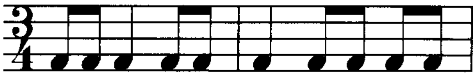

or

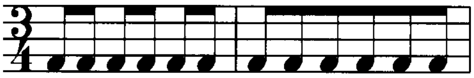

It is better to **avoid arbitrary groupings** of 4+2 or 2+4. However, they are sometimes used to emphasize phrasing.

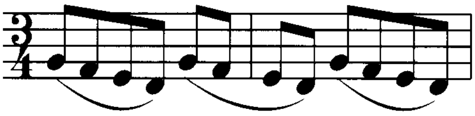

The combined use of **flags and beams** clarify the meter.

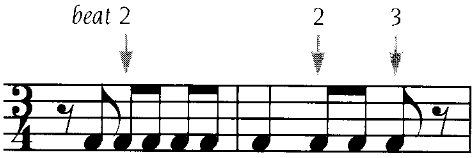

Avoid beaming in a manner that would suggest $\frac{6}{8}$ time. Use flagged notes to differentiate $\frac{3}{4}$ meter from $\frac{6}{8}$ meter.

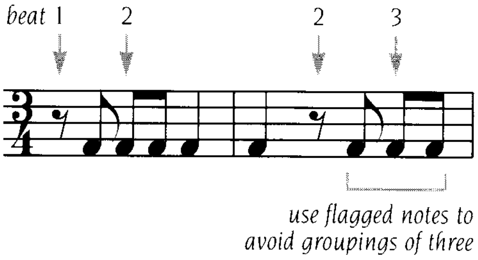

Beaming to include rests inside the grouping is acceptable.

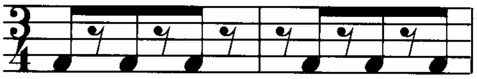

Simple quadruple meter ($\frac{4}{4}$, $\frac{4}{8}$)

In $\frac{4}{4}$ time, beats 1 & 2 and beats 3 & 4 may be beamed together to form a unit.

Never connect the notes of beat 2 with beat 3.

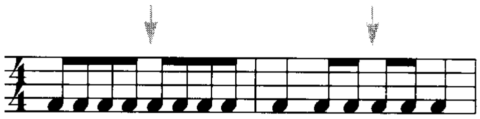

The basic rule in $\frac{4}{4}$ time is that the **two halves** of the measure (beats 1 & 3) **must be immediately recognizable.** The only exception to this would be a very simple syncopation (such as ♩𝅗𝅥 ♩).

Because of this rule, all the notes in the measure would *not* be beamed together, as is possible in the other simple meters.

or

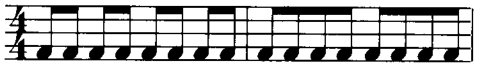

THE SAME RHYTHM THREE WAYS

Each could be the best for a particular situation.

Notice that **each beat** is recognizable in the following example.

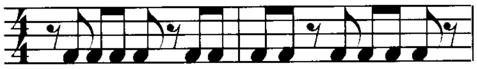

Notated with beams only, the **two halves** of the measure are easily recognizable:

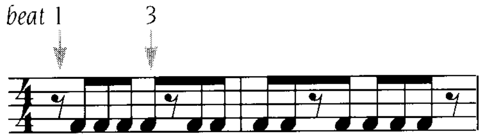

If the above example were in a piece that contained triplets, it could at first glance be **mistaken for triplets.** In that case, a better notation would be:

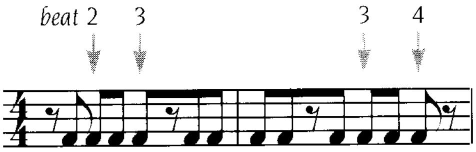

For any simple meter, **when a beat is uneven or contains more than two notes,** the two adjacent beats should not be beamed together as a unit.

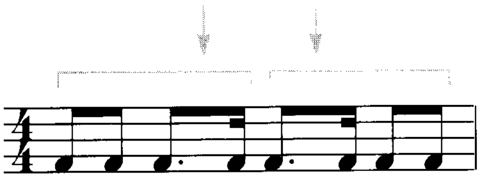

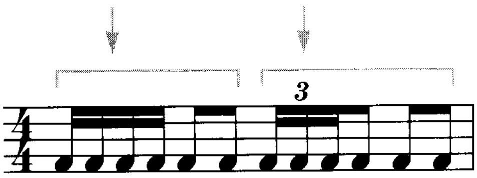

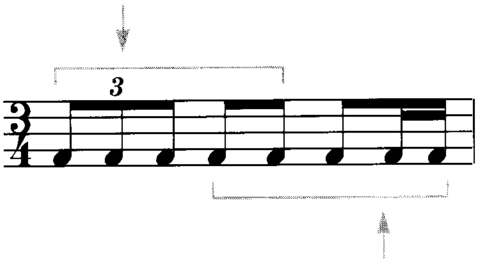

Compound meter

When each **pulse** of the measure is **divided into three beats,** the meter is called "compound." A beam may connect the three beats.

compound duple (2 pulses, 6 beats)

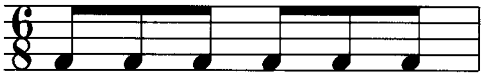

compound triple (3 pulses, 9 beats)

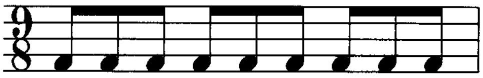

compound quadruple (4 pulses, 12 beats)

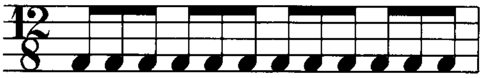

The **units** of three beats **for each pulse** in compound meter **must always be maintained.** Even when rests are included in the measure, as in the example below, beats 1, 4 & 7 must be clearly visible.

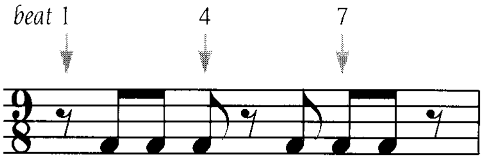

The three units in the above example could be made even more obvious by including the rest on beat 5 in a beamed unit.

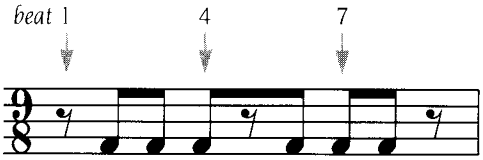

It is **possible to group all notes within a pulse.** All the notes in the *measure,* however, cannot be beamed together—this could cause the measure to appear to be in some other meter.

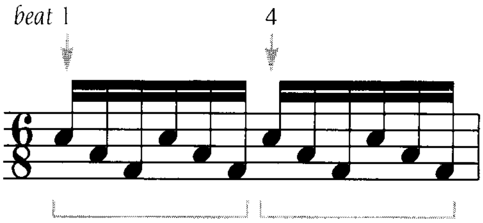

Although different groupings are sometimes used for phrasing or interpretation, the above should be generally adhered to.

TIES AND BEAMS

Avoid using tied notes *within* a beamed grouping. The **unit should be broken** where the tie is placed.

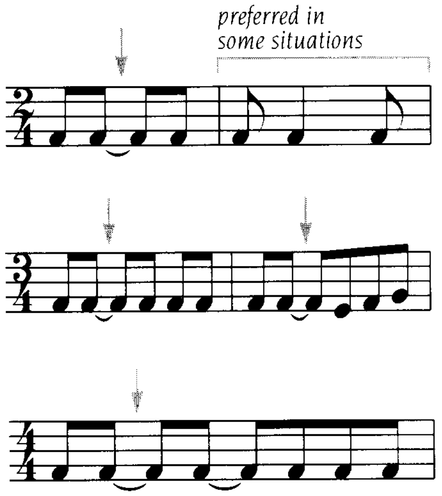
preferred in
some situations

In this case, there would not be a continuous beam, since beats 2 and 3 should not be beamed together in this meter.

STEM DIRECTION OF BEAMED GROUPS

Simple rules for stemming of single notes apply to beamed groups when possible (for example, if all notes in a beamed grouping are on or above the middle line of the staff, stems are down), but some additional guidelines are needed.

(See *Stems*)

For two beamed notes

Stem direction is **determined by the note farther from the middle staff line.**

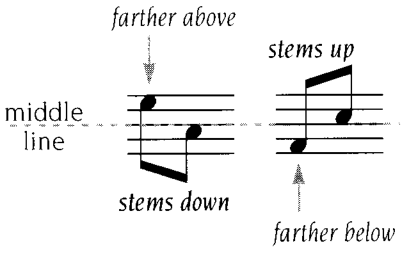
farther above
stems up
middle
line
stems down
farther below

If both notes are **the same distance** from the middle line, the preference is for **stems down.**

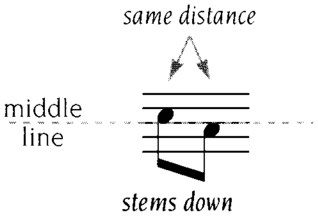
same distance
middle
line
stems down

If all other stems in the surrounding measures are up-stem, that direction is sometimes continued if the two notes are the same distance from the middle line. (This is a very subtle point, and somewhat obsolete.)

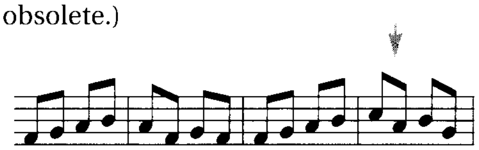

For three or more beamed notes

If the majority of the notes are **on or above** the middle line, the stems go down.

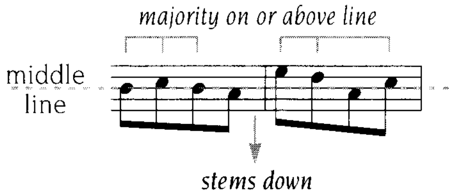
majority on or above line
middle
line
stems down

If the majority of the notes are **below** the middle line, the stems are up.

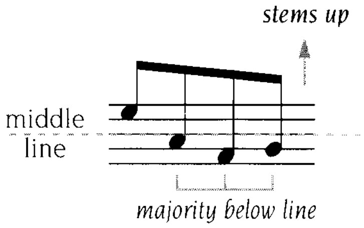
stems up
middle
line
majority below line

If the number of notes **above and below the middle line are equal,** the note farthest from the middle line dictates the stem direction. (If all notes are the same distance, the preference is down.)

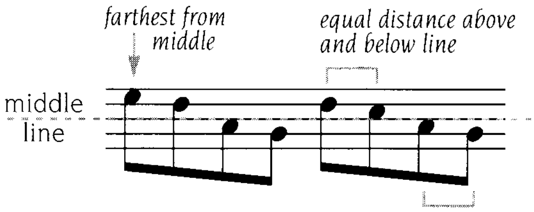
farthest from
middle
equal distance above
and below line
middle
line

BEAM SLANT & PLACEMENT

Concerning the computer

With most computer engraving today, the direction, slant and placement of the beam are automatic.

However, the defaults—or even the settings allowable—may not conform to traditional practices. Even though some of these practices are in the process of change, there are some guidelines to keep in mind for a more professional look.

Beginning placement of the beam

When **stems are up,** the stem length of the **highest note** determines the beam placement.

When **stems are down,** the stem length of the **lowest note** determines the beam placement.

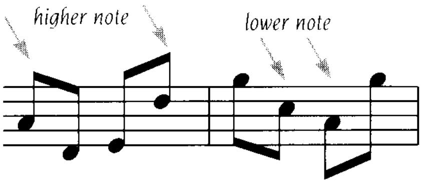

Direction of slant

Beams *usually* **follow the contour of the notes,** ascending or descending.

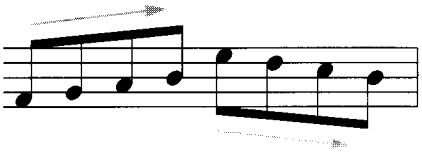

For mixed direction of notes, the beam *usually* **follows the direction between the first and last note** of a grouping.

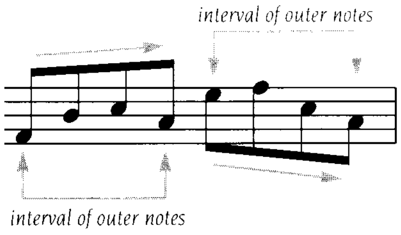

When two or more notes share a stem, the note **closest to the beam** dictates the beam direction.

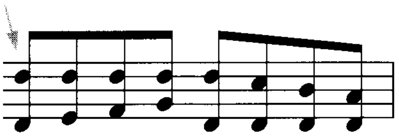

Horizontal beaming

A beam should be horizontal if the first and last note of the beamed group is on the **same pitch level.**

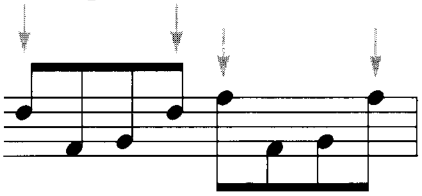

If the grouping consists of **repeated intervals,** the beam should be horizontal.

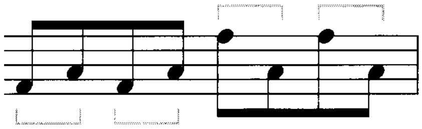

Horizontal beams may be used if inner notes **do not follow the interval direction** of the outer notes.

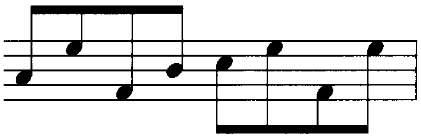

Some other examples of horizontal beaming:

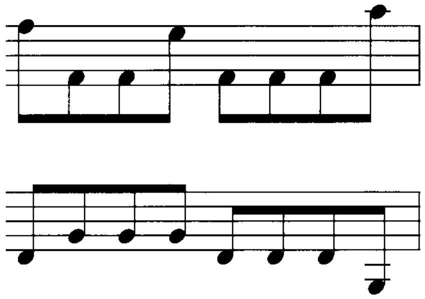

Interval-specific beam slant

Determining the placement and slant of beams in traditional plate engraving was complicated and detailed.

The following are *approximate* guidelines, **modified from the strict traditional** rules for beaming of intervals. There are two main reasons for this modification:

1. The consistency of the computer and the quality of modern printing make it no longer necessary to avoid the small "wedges" of white space that in the past often filled in with ink.
2. Adjusting many beams individually on the computer is very time-consuming.

A pair of **notes of different pitch should not have a horizontal beam.** (Exception: two notes on the same line or space, one or both altered with an accidental)

If a **beam slants, it is usually from 1/2 to 2 staff spaces**—very seldom more.

Intervals of a **2nd** slant 1/2 space.

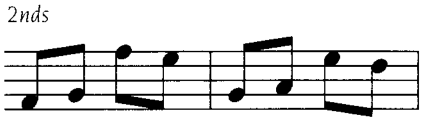

Intervals of a **3rd** slant 1 staff space.

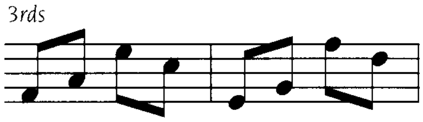

Intervals of a **4th** slant from 1 to 1-1/2 staff spaces.

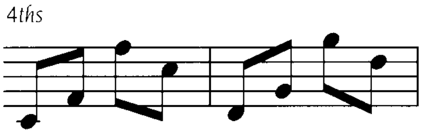

Intervals of a **5th, 6th & 7th** each slant approximately 1 to 1-1/2 staff spaces.

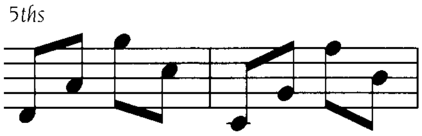

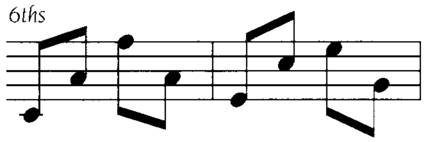

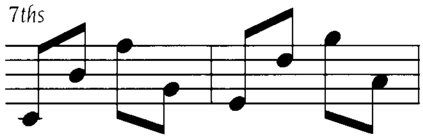

Intervals of an **octave and greater** slant 1-1/2 to 2 staff spaces.

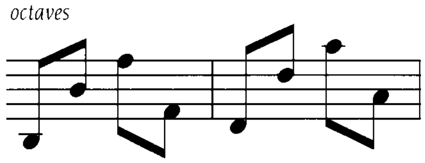

Beamed notes with **leger lines** slant 1/2 staff space. (Remember that stems must extend to the middle line.)

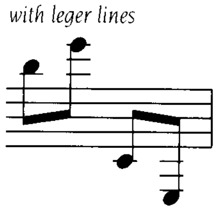

Depending upon the **amount of space** allowed between the beamed notes, the slant of the beam naturally *decreases* with an *increase* in the amount of space allotted, if stem lengths remain the same.

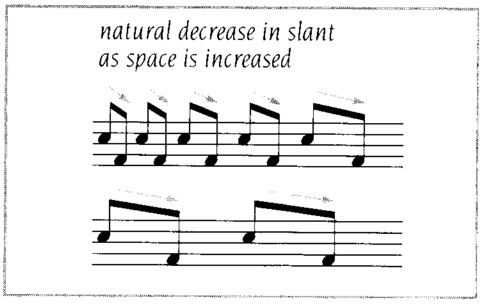

For tightly spaced beamed notes, **adjust the slant** (by lengthening the stem of the farthest note) so that it is not so extreme.

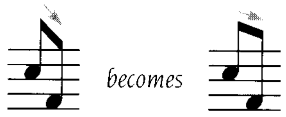

Alternative beaming

For grand staff notation, **cross-staff beaming** is acceptable in some situations. The slant of the beam is determined by the amount of space between the two staves, the length of the stems, and by avoiding exaggerated angles.

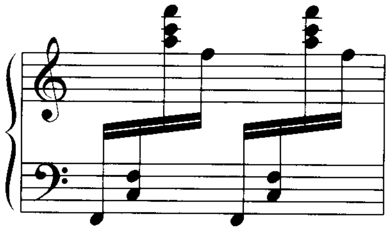

For single-staff music, alternative beaming should be used only in **exceptional circumstances.** Stems must not be too short.

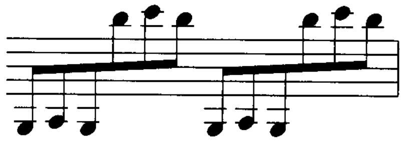

Bowing

Down-bow and **up-bow** markings are centered to the notehead and are placed above the staff, except when two parts share a staff.

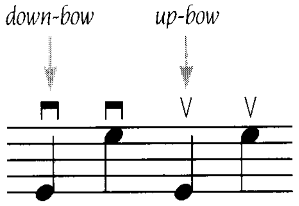

String techniques such as ***spiccato, legato, portato*** and ***jeté*** are notated conventionally, with slurs and/or articulations.

The study of orchestration is necessary to properly specify string indications.

Slurs can indicate either phrasing or bowing, or both.

Brackets to the left of the chord may be used for **multiple stops** if the chord is to be divided a particular way. Usually no bracket is used.

Multiple stops may be *double, triple* or *quadruple* stops.

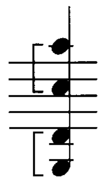

PLACEMENT OF TEXT

All text for bowing indications is **placed above** the staff.

When a particular string is to be played exclusively, ***sul*** [string] is indicated.

sul G

Pizzicato is cancelled with the ***arco*** indication.

pizz. ***arco***

With any of the following indications, ***ord.*** or ***ordinario*** is commonly used to indicate the return to normal playing.

non vib. ***sul pont.*** ***sul tasto*** ***col legno***

Brace

The brace is used to connect the two staves in scores for instruments such as piano, organ, harp and marimba.

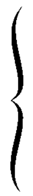

The brace connects two staves to form a **grand staff.**

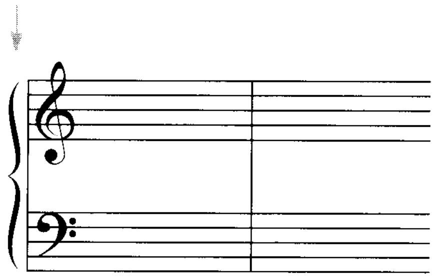

Bracket

The bracket is used **to connect two or more staves** to form a system or to **group certain instruments** within a system.

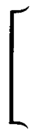

The bracket is different from the brace in that it connects the staves for two or more **separate instruments,** while the brace connects staves for one instrument (such as keyboard).

The grouping may be two or more **unlike** instruments, **identical** instruments, or a **family** of instruments, such as strings.

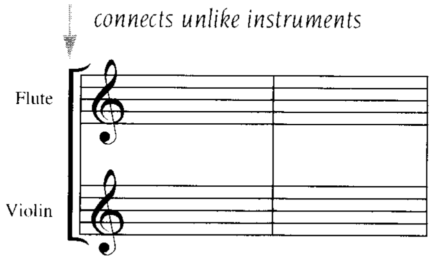

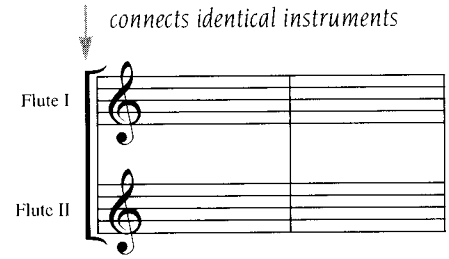

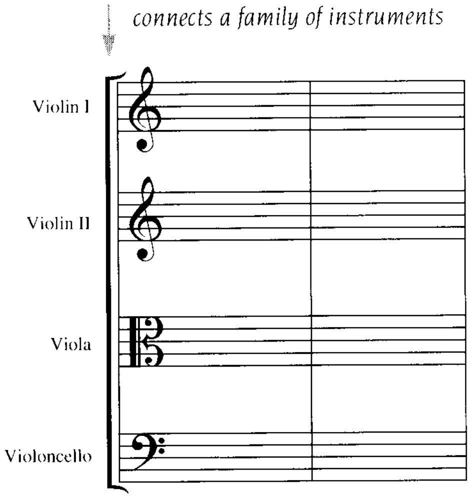

(See *Barlines, Scores*)

Breath mark

The breath mark (❜) has two similar uses:

- **an actual breath indication** for instruments (including voice) requiring the breath to sound.
- **a small pause or break** (as if taking a breath) for instruments *not* requiring breath to sound.

A breath mark may or may not interrupt the tempo. If it does interrupt the tempo, the break or pause is shorter than that of a caesura or fermata.

PLACEMENT

The breath mark is indicated just above the staff, at the point **where a breath or break is desired.**

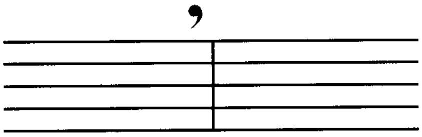

If the breath mark occurs between two notes, the placement is toward the **end of the space for the full value** of the note.

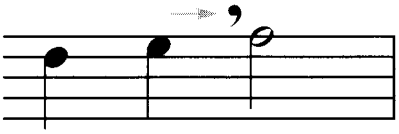

If the breath mark occurs at the end of a measure, the placement is just **before the barline.**

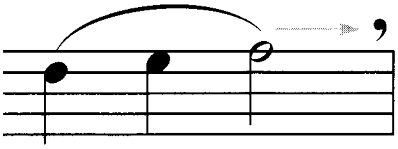

Caesura

The caesura (//) is a small pause or break in the tempo (implying a break in sound).

The pause of the caesura is **slightly longer than a breath mark,** but shorter than a fermata.

PLACEMENT

The caesura is placed on the **fourth line from the bottom** of the staff (extending to the first leger line), at the point **where a break is desired.**

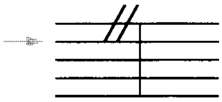

The caesura is placed toward the **end of the full value** of the note.

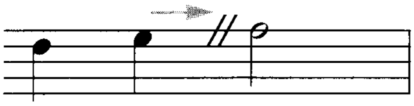

If space is tight, the caesura **can be centered** between two notes.

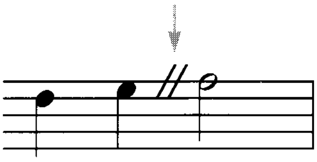

If the caesura occurs at the end of a measure, the placement is just **before the barline.**

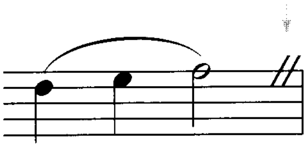

If a longer pause is desired, **a fermata may be placed above** the caesura.

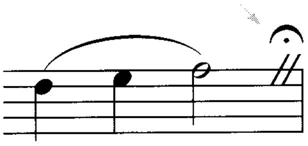

Chord frames & symbols

Chord frames are diagrams that contain all the information necessary to play a particular chord on a fretted instrument.

The photograph below shows the finger numbers corresponding to the fingerings used for guitar.

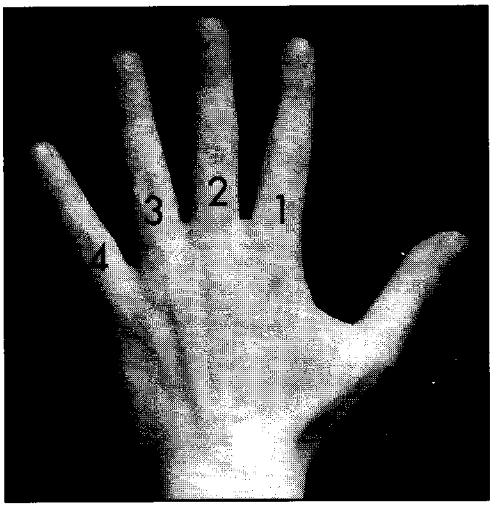

The number of strings and tunings vary between string instruments. The **chord frame is modified** (number of strings, tunings) according to the differences between each instrument.

The following examples explain the various **chord frame elements** for guitar.

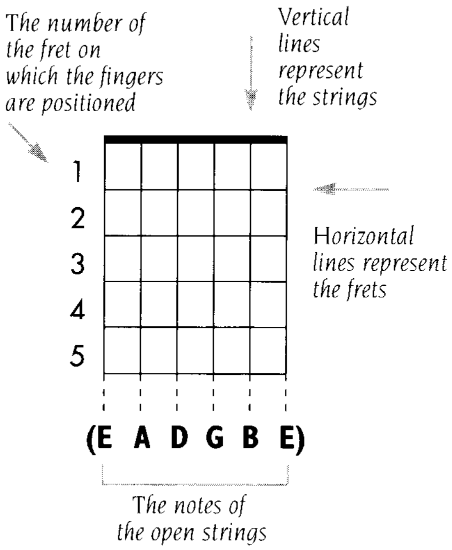
The number of the fret on which the fingers are positioned
Vertical lines represent the strings
Horizontal lines represent the frets
1
2
3
4
5
(E A D G B E)
The notes of the open strings

The **fingerings, note names** and **position** of the chord on the neck are all provided on the chord frame.

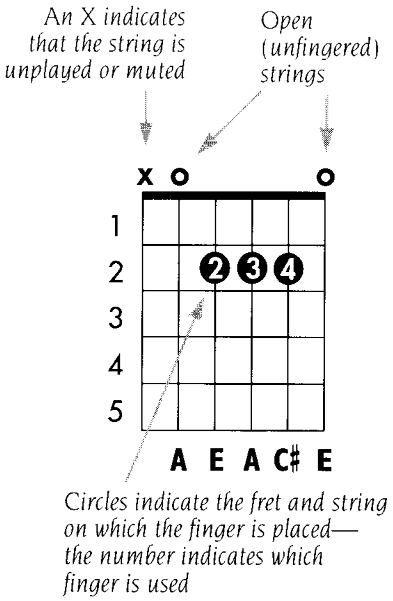
An X indicates that the string is unplayed or muted
Open (unfingered) strings
X O O
1
2
3
4
5
2 3 4
A E A C♯ E
Circles indicate the fret and string on which the finger is placed— the number indicates which finger is used

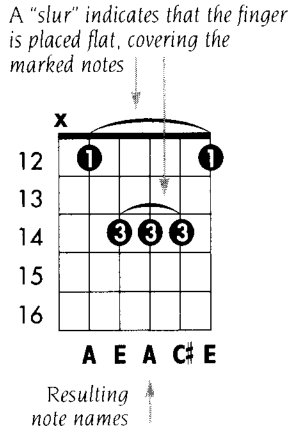
A "slur" indicates that the finger is placed flat, covering the marked notes

x
12
13
14
15
16
1
1
3
3
3
A E A C♯ E

Resulting note names

CHORD SYMBOLS

The **chord symbol** is always placed above the frame to identify the chord.

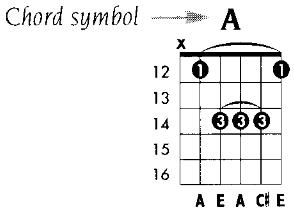
Chord symbol
A
x
12
13
14
15
16
A E A C♯ E

Careful attention must be given to the **consistent identification** of chords, in the **least ambiguous way** possible.

Chord symbols often appear without a frame. Whether or not there is a frame, space must be allowed so that the symbol is **aligned with the proper beat.**

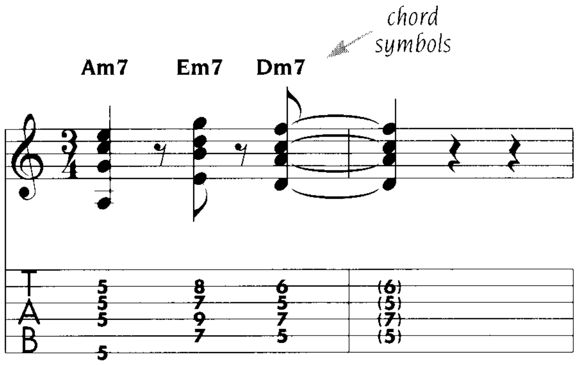
chord symbols
Am7
Em7
Dm7
T
A
B
5 5 5 5
8 7 9 7
6 5 7 5
(6) (5) (7) (5)

(See *Tablature*)

A complete listing of guitar chords, along with chord theory, is provided in Alfred's Guitar Chord Encyclopedia *(Item #4432). It also includes a listing of chord symbols and a listing of ambiguously named symbols.*

Circle of fifths See *Key signatures*

Clef signs

The following are commonly used clefs.

Middle C is shown for comparison in all examples except neutral clefs.

TREBLE AND BASS CLEFS

Treble and bass clefs are most frequently used. Keyboard, harp and organ music is written using only these clefs.

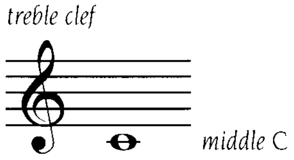
treble clef
middle C

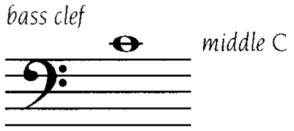
bass clef
middle C

The **octave treble clef** is sometimes used for tenor vocal parts, to indicate that the part is one octave lower than written.

The **double treble clef** serves the same function as the octave treble clef but is much less frequently seen.

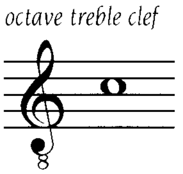
octave treble clef
8

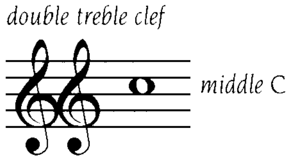
double treble clef
middle C

The regular **treble clef** is also commonly used for tenor vocal parts; the octave is assumed.

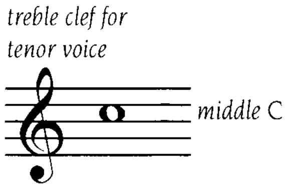
treble clef for tenor voice
middle C

C CLEFS

Alto clef is commonly used for viola, to avoid excessive leger lines. (Viola also uses treble clef.)

alto clef

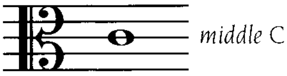

Tenor clef is primarily used for bassoon, cello and trombone when playing in higher registers.

tenor clef

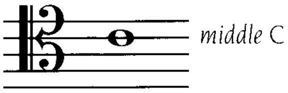

The following C clefs are **obsolete:**

soprano clef

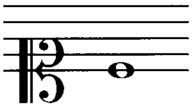

mezzo-soprano clef

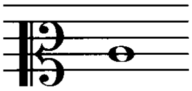

baritone clef

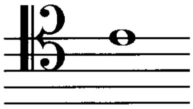

NEUTRAL CLEFS

Neutral clefs are used by percussion instruments of indefinite pitch. The following is most commonly used.

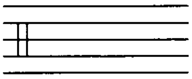

Neutral clef for single-line percussion:

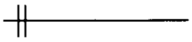

POSITIONING

Clefs are **slightly indented on the staff,** to the right of a systemic barline or the open single staff. Be sure to leave the ***single* staff open—do not put a barline before the clef!**

indent on single staff—no left barline

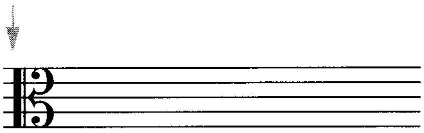

indent on staff with systemic barline

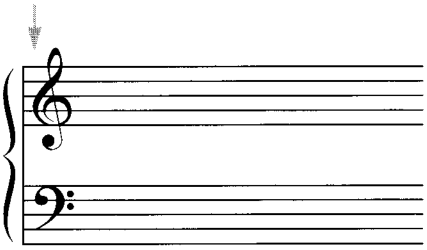

The clef is always **placed before the key signature and time signature.**

Notice the distance from the end of the staff to the clef, from the clef to the key signature, and from the key signature to the time signature.

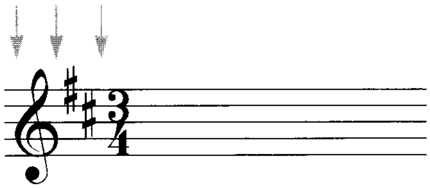

CLEF CHANGES

When the clef changes **within a staff,** a cue-size clef is used (usually 75% of the original clef size). Courtesy clefs are also cue size.

If a clef change affects a complete measure, it is always placed before the barline.

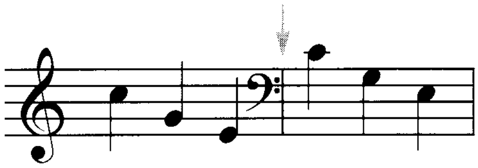

Only a *systemic* barline can precede a clef sign. (See *Systems*)

If a clef change is within a measure, the clef is placed directly before the first note involved.

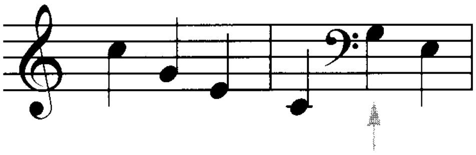

If a clef change **begins after a rest and is on a beat,** the clef is placed directly before the first note affected.

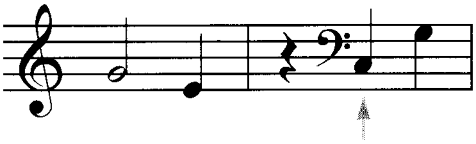

If a clef change **begins after a rest but is not on a beat,** the clef precedes the rest.

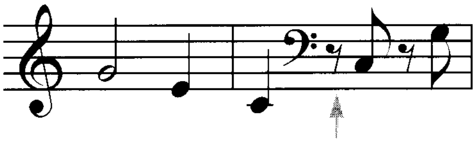

If a clef change **occurs at the beginning of a staff,** a courtesy clef is placed before the barline at the end of the previous staff, followed by the new clef at full size in the new staff.

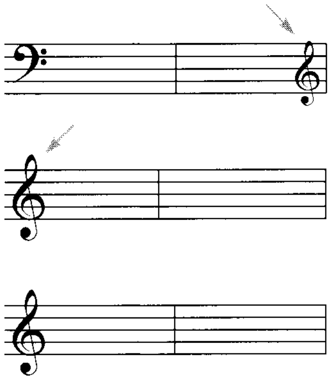

The rules for clef changes are not affected by the **addition of a key change and/or time signature change.** The clef sign precedes the barline; the key signature and time signature follow the barline.

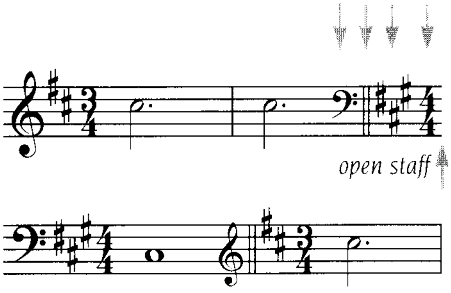

Notice that the **order** of courtesy clef, key signature and time signature is the same as the order at the beginning of a piece.

The **courtesy clef** is cue size; the key signature and time signature are normal size.

Also notice the **open staff** after the *courtesy* key signature and time signature.

Coda sign See *D.C. al Coda, D.S. al Coda*

Courtesy accidentals See *Accidentals*

Credits

Credits may consist of the **composer, lyricist** and **arranger,** or a word such as *traditional, anonymous, folk song,* etc.

1. The **composer** credit is placed flush right with the page margin. It is not horizontally aligned with the tempo marking.
2. The **arranger** credit is placed below the composer credit (also not horizontally aligned with the tempo marking).
3. The **lyric** credit is placed flush left with the page margin, horizontally aligned with the composer credit.

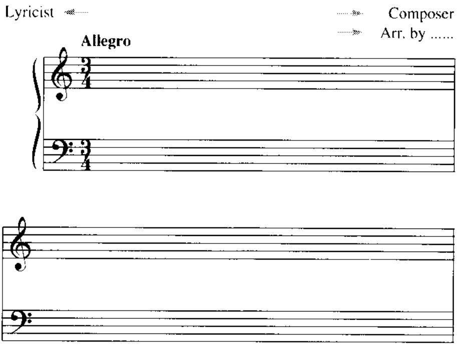

The typeface for all three credits is usually Roman.

For a less formal style, such as pop piano/vocal music, the lyric credit would go under the composer credit, with the arranger under that. **All credits would be flush right** with the page margin.

Some examples:

Words and Music by...................

Words and Music by
....

Music by...................
Words by...................
Arranged by...................

The word "arranged" may be abbreviated (Arr.).

Cue notes

Cue notes may be given as guidance only, **to assist a performer's entrance** after numerous measures of rest.

Cue notes may also be given for **possible performance** if the cue instrument is weak or missing. (See *Alternating instruments*)

The size of cue notes is somewhat **smaller than normal note size,** but still large enough to be legible (65–75% of normal note size).

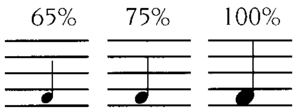

All musical elements associated with the cue notes will also be at **cue size.**

There are **two ways to notate** cue passages.

With **reverse stem direction** and added rests:

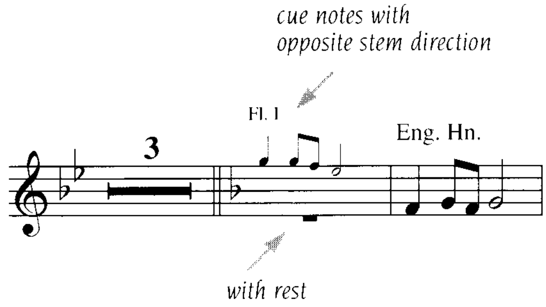
cue notes with
opposite stem direction
Fl. 1
3
Eng. Hn.
with rest

With **normal stem direction** and no added rests:

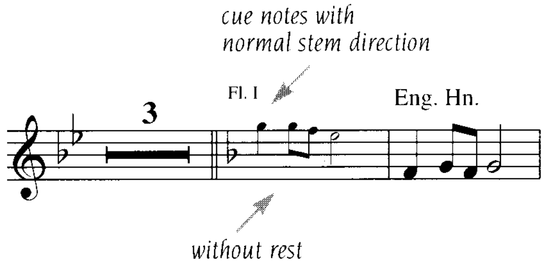
cue notes with
normal stem direction
Fl. 1
3
Eng. Hn.
without rest

Piano accompaniment parts may include a cue-size solo part above the grand staff.

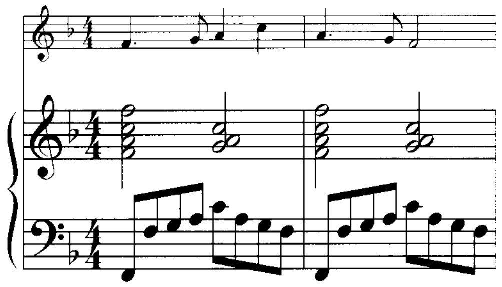

Not all notes that are cue *size* are used as a cue.

Ossia passages and **grace notes** are typically notated at cue size. (See *Ossia, Grace notes*)

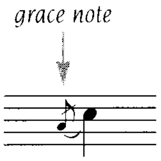

D.C. al Coda (da capo al coda)

Means go back to the beginning, play to the "to coda" sign (𝄌), then jump to the coda to finish the piece.

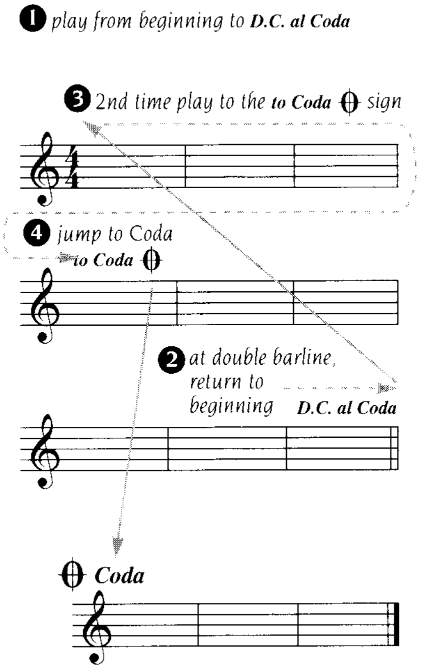
❶ play from beginning to D.C. al Coda
❸ 2nd time play to the to Coda 𝄌 sign
❹ jump to Coda
to Coda 𝄌
❷ at double barline, return to beginning
D.C. al Coda
𝄌 Coda

Align the edge of the "to coda" sign **flush right** with the barline.

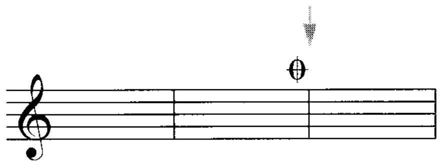

The words ***to Coda*** are often omitted from the "to coda" sign.

Place a **double barline** (thin/thin) at the ***D.C. al Coda.*** **(*D.C. al Coda*** can also be written ***D.C. al Coda* 𝄌)**

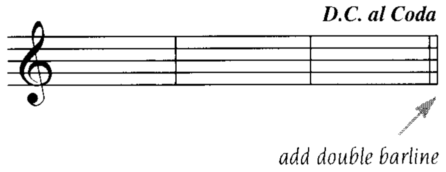
D.C. al Coda
add double barline

Separate the coda from the staff above (optional) and **indent.**

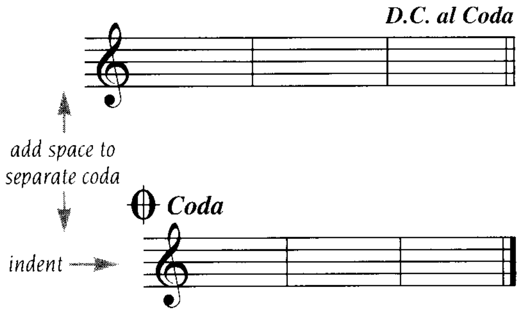
D.C. al Coda
add space to separate coda
Coda
indent

Align the coda sign with the beginning of the staff. It should be somewhat **larger** than the "to coda" sign.

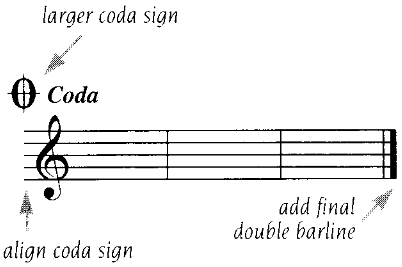
larger coda sign
Coda
add final double barline
align coda sign

The coda will end with a **final double barline** (thin/thick).

It is sometimes necessary to begin the coda on the same staff as the ***D.C. al Coda.*** If so, **break the staff** and **add a clef and key signature.** (Do not add a time signature unless it changes at the coda.)

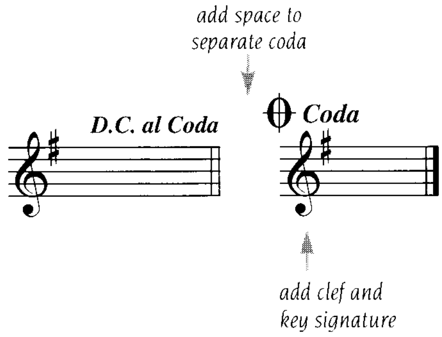
add space to separate coda
D.C. al Coda
Coda
add clef and key signature

D.C. al Fine (da capo al fine)

Means go back to the beginning and play to the ***Fine.***

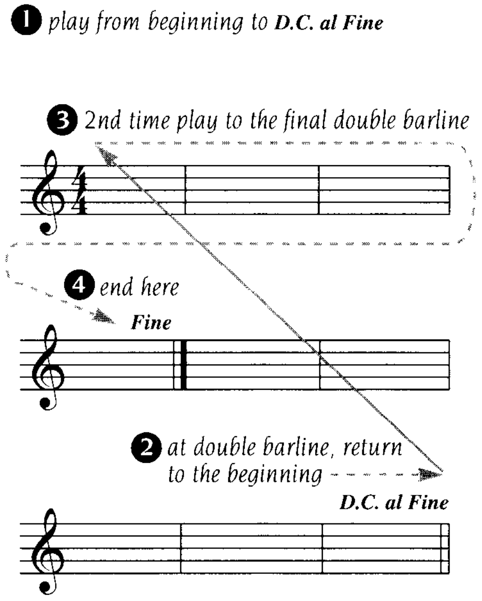
❶ play from beginning to D.C. al Fine
❸ 2nd time play to the final double barline
❹ end here
Fine
❷ at double barline, return to the beginning
D.C. al Fine

Place a **double barline** (thin/thin) at the ***D.C. al Fine.***

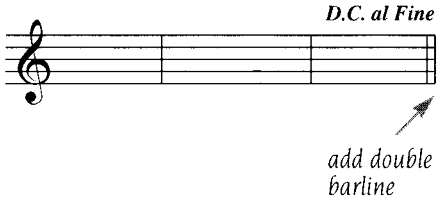
D.C. al Fine
add double barline

A **final double barline** (thin/thick) is added at the ***Fine.***

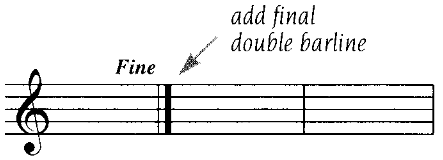
add final double barline
Fine

D.S. al Coda (dal segno al coda)

Means go back to the sign (𝄋), play to the "to coda" sign (𝄌), then jump to the coda to finish the piece.

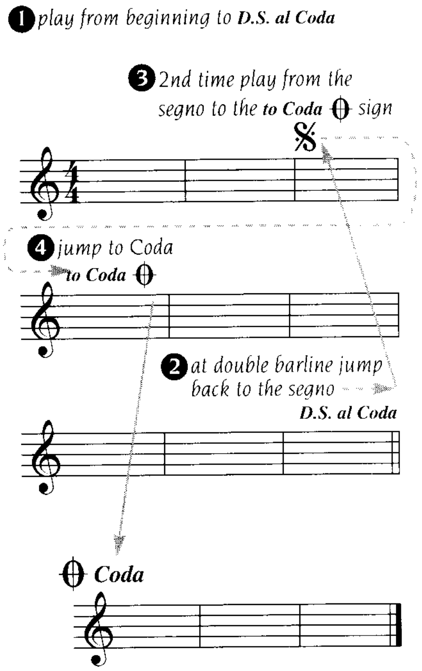
❶ play from beginning to D.S. al Coda
❸ 2nd time play from the segno to the to Coda 𝄌 sign
❹ jump to Coda
to Coda 𝄌
❷ at double barline jump back to the segno
D.S. al Coda
𝄌 Coda

The segno 𝄋 should be **aligned flush left** with the barline.

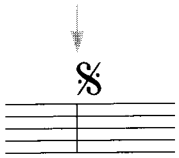

Align the edge of the "to coda" sign **flush right** with the barline.

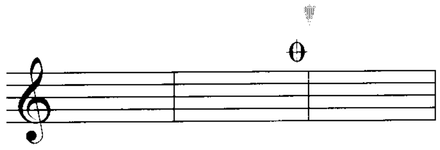

The words *to Coda* are often omitted from the "to Coda" sign.

Place a **double barline** (thin/thin) at the *D.S. al Coda.* (*D.S. al Coda* can also be written *D.S.* 𝄋 *al Coda* 𝄌*.)*

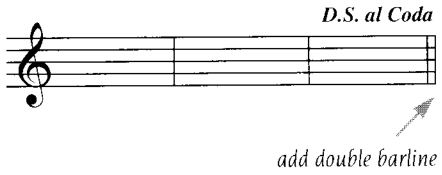
D.S. al Coda
add double barline

Separate the coda from the staff above (optional) and **indent.**

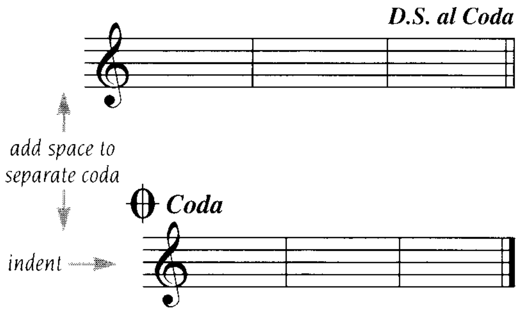
D.S. al Coda
add space to separate coda
𝄌 Coda
indent

Align the coda sign with the beginning of the staff. It should be somewhat **larger** than the "to coda" sign.

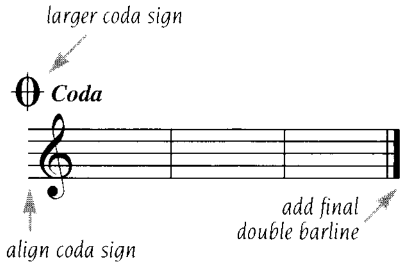
larger coda sign
𝄌 Coda
add final double barline
align coda sign

The coda will end with a **final double barline** (thin/thick).

D.S. al Fine (dal segno al fine)

Means go back to the sign and play to the ***Fine***.

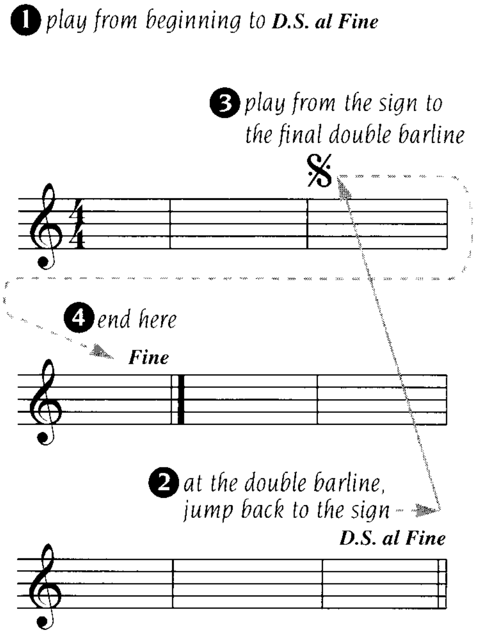
❶ play from beginning to D.S. al Fine
❸ play from the sign to the final double barline
❹ end here
Fine
❷ at the double barline, jump back to the sign
D.S. al Fine

The segno 𝄋 should be **aligned flush left** with the barline.

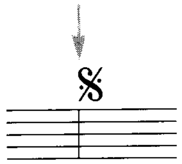

Add a **double barline** (thin/thin) at the ***D.S. al Fine.***

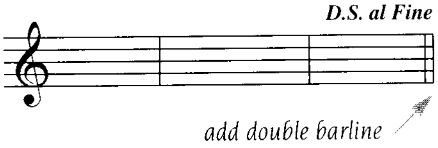

A **final double barline** (thin/thick) is added at the ***Fine.***

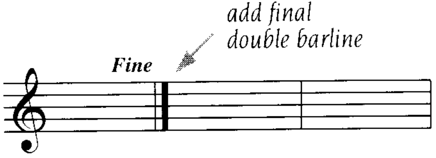

Divisi

Divisi is used for instruments (such as strings) that can produce **more than one note at a time,** indicating that the notes should be divided among the *group* of instruments (i.e., instead of multiple stops). (See *Sharing a staff*)

Divisi dots

See *Tremolo*

Dotted rhythms

See *Augmentation dot*

Dynamics

Dynamics indicate the varying degree of volume or intensity of a note, phrase or section of music.

The following lists the most common dynamics from the softest to the loudest:

softest → *loudest*

ppp pp p mp mf f ff fff

Dynamics **softer than *ppp* or louder than *fff* are impractical,** although theoretically any dynamic is possible.

COMBINED DYNAMICS

Dynamics such as ***mf, f, sf*** can be combined with ***p*** or ***pp***. The resulting dynamics function as a somewhat broader form of articulation.

fp sfp sfpp

fp *would mean to play loudly, then immediately play softly.*

(See *Articulations*)

CRESCENDO & DIMINUENDO

The **crescendo sign** is wedge-shaped and **opens toward the right.**

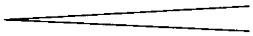

The **diminuendo sign** is wedge-shaped and **opens toward the left.** (The words *diminuendo* and *decrescendo* are interchangeable, but *diminuendo* is preferred.)

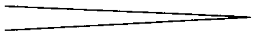

A crescendo or diminuendo sign is used when a **gradual increase or decrease** is wanted from one dynamic to another.

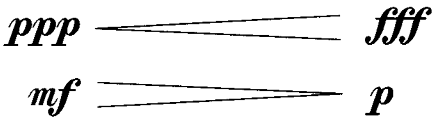

Musical elements should not be placed within the opening of the signs.

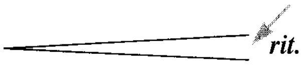

The word ***crescendo*** or ***diminuendo (decrescendo)*** may be written or abbreviated ***cresc., dim. (decresc.).***

The **effect is identical** whether using the sign, word or abbreviation.

If space doesn't allow the use of the sign, or the passage is a long one, it is best to use the word.

A **dashed line** may be used if it helps clarify the length of the passage.

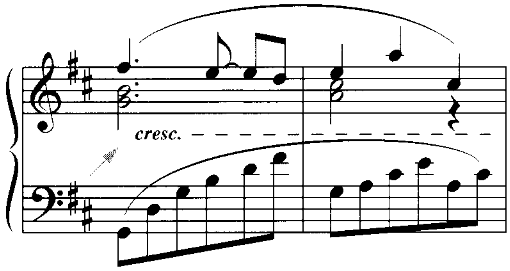

The word is sometimes **hyphenated** to extend throughout a section of music.

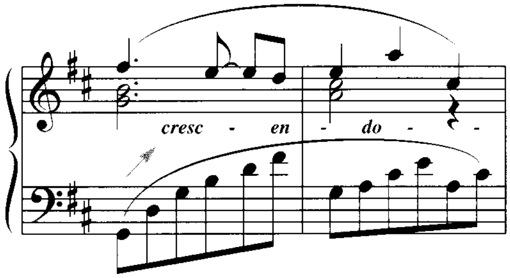

GENERAL PLACEMENT

Place the dynamic sign *slightly* before the notehead whenever possible.

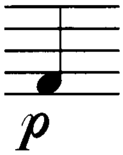

For single-staff instruments

Dynamics are placed **below the staff.**

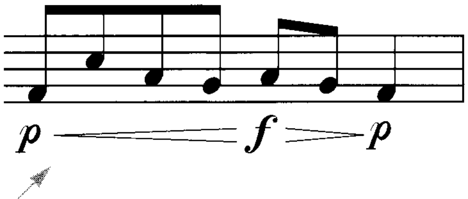

dynamics below for instruments

When two parts share a staff and the dynamics vary between parts, the dynamics will be placed above and below the staff.

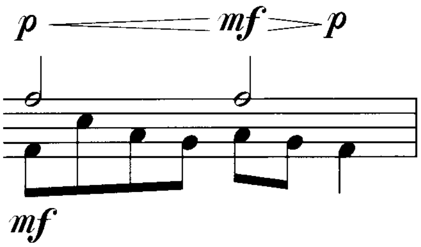

For vocal music

Dynamics are placed **above the staff** to avoid conflicting with the lyric.

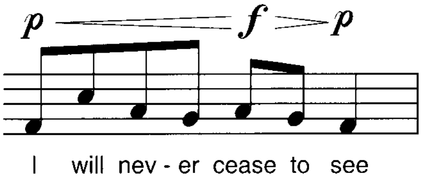

For the grand staff

Dynamics are placed **between the two staves.**

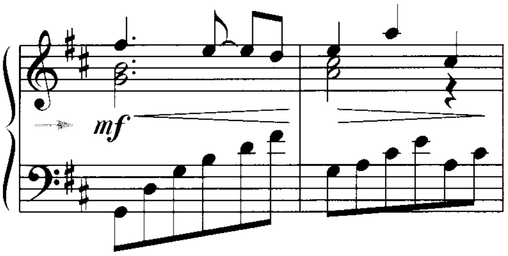

If two different dynamics are needed for each staff, the dynamics may be placed above for the top staff and below for the bottom, or in positions that clarify the situation best.

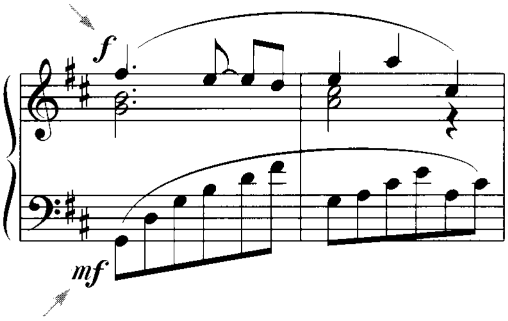

Horizontal placement is preferred for crescendo and diminuendo signs, although an angled placement is acceptable if necessary.

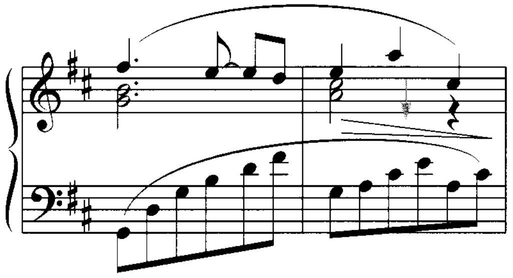

The signs may also be **placed outside the staff** if space is tight within the grand staff.

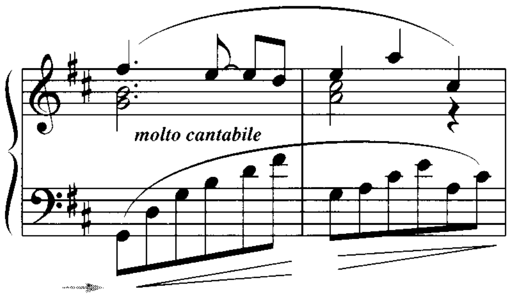

If a dynamic mark must be placed on a barline, the barline should be **broken.**

For crescendo and diminuendo signs, the **barline is left intact.**

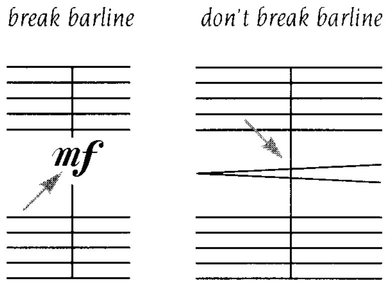

Dynamics with crescendo/ diminuendo signs

The intensity of the increase/decrease in volume is **relative to the dynamic level** of the section of music.

> *If a dynamic is not indicated at the end of a crescendo sign, and the music is consistently gentle, the increase in volume would probably be subtle.*

The **level at the beginning of the sign** is the same as **the dynamic last indicated** in the music.

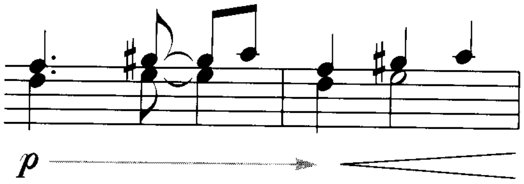

It is advised to **carefully begin and end the sign** so that it makes sense.

While the example below would make perfect sense in a clarinet part, it would make no sense in a piano part.

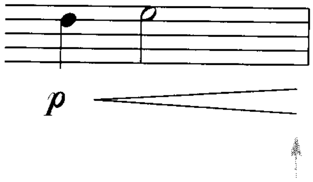

Endings

See *Repeat signs*

Fermata

A hold or pause sign (𝄐). It is used with any rhythmic value—single notes, intervals, chords or rests.

PLACEMENT

Preferred placement is above, regardless of stem direction, except when parts share a staff.

The fermata is **centered on the notehead** (or rest), whether it is on the notehead side or the stem side.

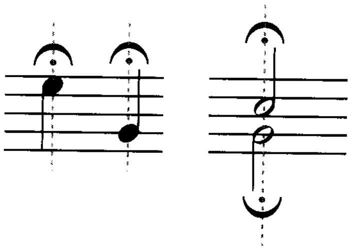

One fermata applies to all notes of an interval or chord.

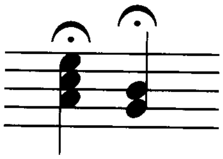

Place outside the staff if at all possible.

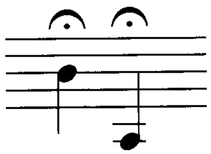

Place outside articulations and slurs.

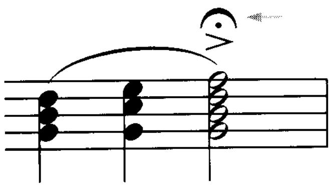

Place the fermata on the exact beat that is to be held, so that there can be no confusion. In the first example below, it might be wondered whether the fermata over the whole notes would apply to the eighth rest in the lower voice.

Confusing:

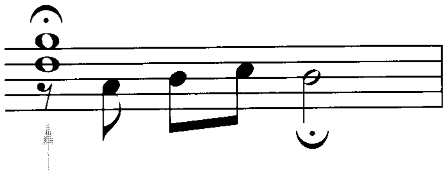

Same example, more clearly written:

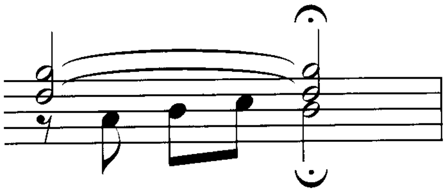

TIME VALUE

A fermata over a note or chord not only indicates that the tempo is interrupted but that the note or chord is **sustained.**

The length of a fermata is **determined by the performer** and is relative to the musical situation.

In a fast tempo, placed on a sixteenth note, the fermata would be a shorter value than if placed on a whole note, in a slow tempo.

USING WITH RESTS

In the following example, the fermata **prolongs silence rather than sound.**

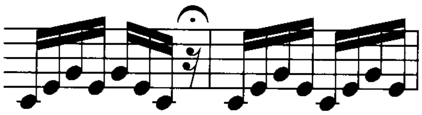

ABOVE A CAESURA

The following example is similar to the last one. The caesura indicates a small break in the sound; the addition of a *fermata* **lengthens the break.**

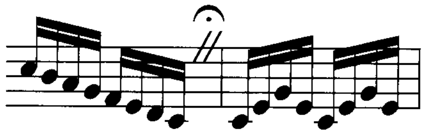

Occasionally a fermata is seen over a final double barline before another movement. However, since a barline has no time value, **avoid using a fermata over a single barline.**

ACCOMPANYING TERMS

Lunga *(long)* or **poco** *(little* or *short)* may be used with a *fermata* for a more specific intent.

These should be **centered above** the *fermata.*

Fingerings (for keyboard)

GENERAL PLACEMENT

Fingerings for the **right hand** should be placed **above** the staff, away from all musical elements.

Fingerings for the **left hand** should be placed **below** the staff, away from all musical elements.

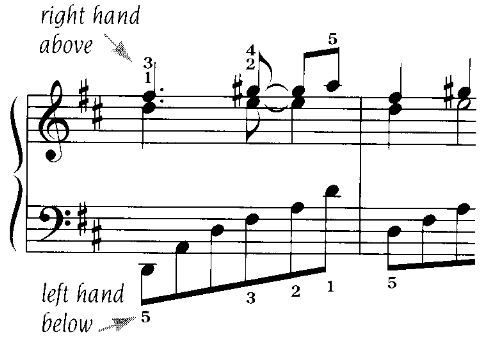

Fingerings are **centered** above or below the notehead, when possible.

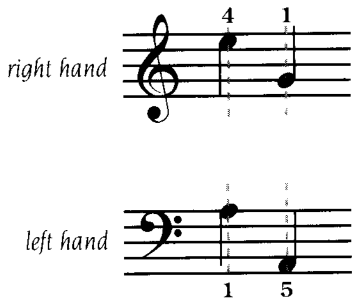
4
1
right hand
left hand
1
5

A finger number is preferably placed **outside a beam.**

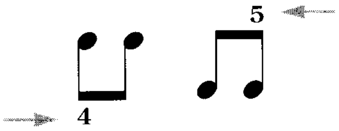

CHORDS

Fingerings are **stacked above or below** a chord; all notes should have a fingering.

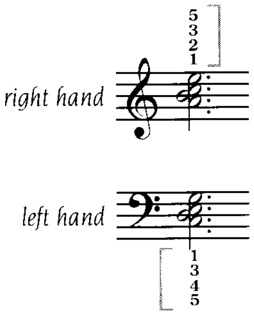
5
3
2
1
right hand
left hand
1
3
4
5

If a chord includes **notes that are tied** and a fingering is indicated for a note that is *not tied,* only that fingering is necessary.

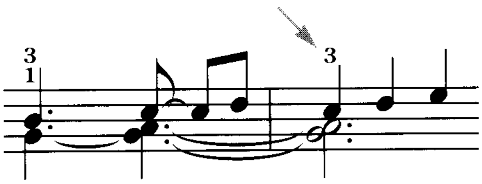

ALTERNATING FINGERINGS

If a fingering changes on a note that is held, a hyphen (-) or an "en-dash" (–) **separates the two fingerings.** The first fingering is centered on the note; the second is to the right.

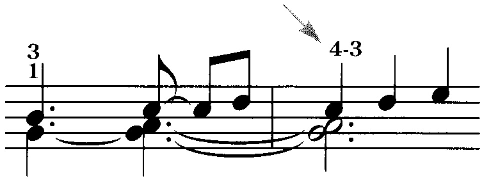

An alternate method uses a small slur instead of the hyphen or en-dash.

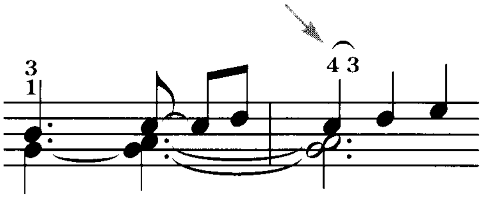

USE WITH TRIPLETS

The triplet ***3*** is sometimes mistaken for a fingering when the fonts are similar.

The **fingering font should be smaller** and of a **different typeface** than the font used for triplets.

Avoid positioning the triplet *3* on the same side of the staff as a fingering.

Instead of moving the fingering, **move the triplet *3*** to the opposite side from the fingering. The triplet ***3*** would normally be on the *beam* side; **if on the notehead side** it is helpful to **add a bracket.** (See *Tuplets*)

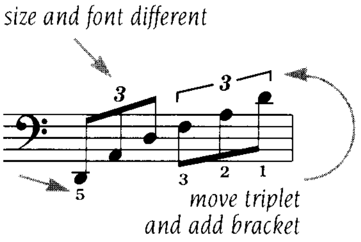
size and font different
3
3
5
3
2
1
move triplet and add bracket

TWO HANDS ON ONE STAFF

If both hands share the same staff, the fingering for the **right hand is above,** the **left hand below.**

RH (right hand) or **LH** (left hand) may be added for **clarification.**

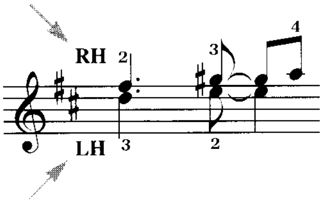

AMOUNT OF FINGERING

Fingering should be **used sparingly** and only to establish a pattern or hand position.

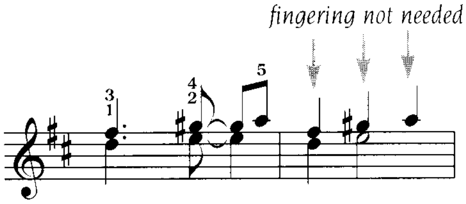

OVERLAPPING ELEMENTS

Fingerings **may overlap a slur.** The preference is to break the slur at the point of intersection, if possible.

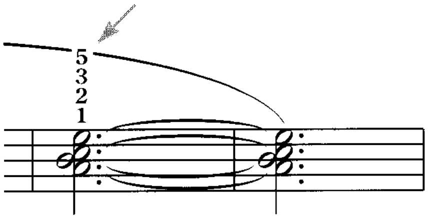

The vertical space sometimes dictates that fingerings be placed within the staff. The **staff lines should be broken** to accommodate the overlapping fingerings.

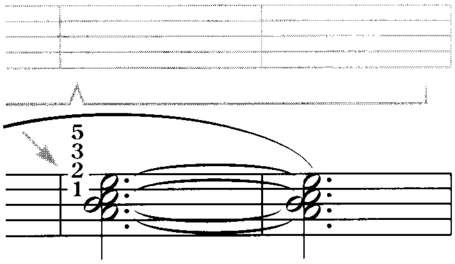

Flat

See *Accidentals, Key signatures*

Glissando sign

A glissando is indicated by a wavy or straight line placed at an angle, ascending or descending.

Indicate the beginning and ending notes of the glissando.

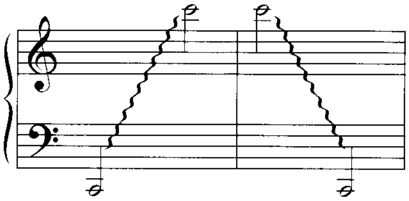

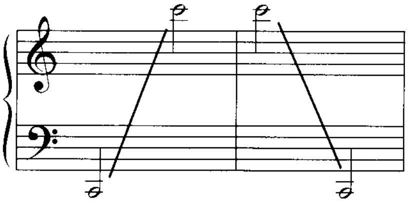

ADD THE ABBREVIATION

It is helpful to **add the abbreviation** (*gliss.*), placing it at the same angle as the symbol.

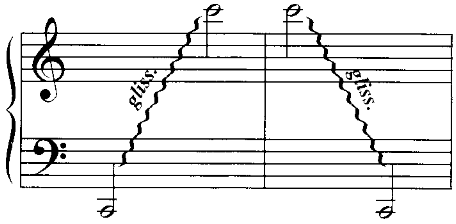

TWO OR MORE NOTES

If the glissando affects more than one note, add a **glissando symbol for each note.** The word (*gliss.*) is used only once regardless of the number of notes.

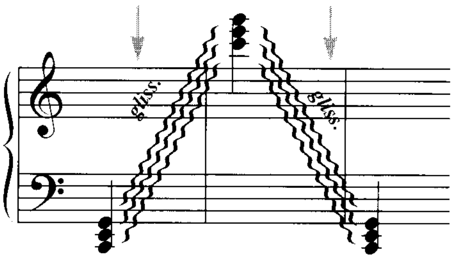

Grace notes

Grace notes are small notes **without a rhythmic value** of their own, taking their value from the previous or following beat.

Grace notes are notated at **cue size** or slightly smaller (65% of normal size works well).

SINGLE GRACE NOTES

Single grace notes are **considered eighth notes.**

Place the **stem up** regardless of the main note's stem direction (except when two parts share a staff).

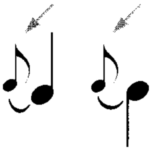

Usually a **small slur connects** the grace note to the main note; the slur usually begins below the grace note.

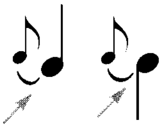

The slur **end does not need to be centered** on the main notehead like slurs normally do. It may end **slightly to the left** of the notehead. (See *Slurs*)

For single grace notes with a flag, a **small line** can intersect the flag, always **slanting upward, from left to right,** regardless of stem direction.

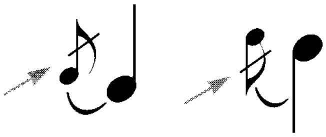

(for a lower voice sharing a staff)

PLACEMENT ON THE STAFF

The grace note is centered in a space or placed directly on a line.

Grace-note **accidentals** are also cue size. (See *Accidentals*)

Stems for grace notes on leger lines do not have to be lengthened to meet the middle line.

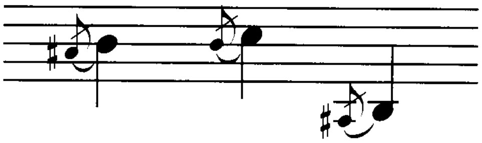

SHARING A STAFF

When two parts share a staff, the grace note stems follow the stem direction of the parts.

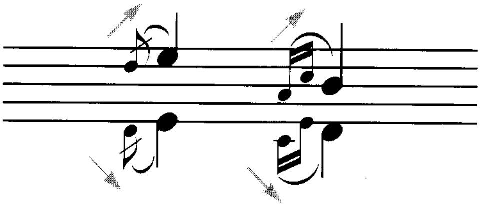

MULTIPLE GRACE NOTES

Multiple grace notes should **never have a slanted line** through the stems or beams. The slanted line is only for single, flagged grace notes.

If a slur is used, the **slur connects** all the grace notes **to the main note.**

Two grace notes are **considered sixteenth notes** and should be beamed accordingly. Three and four can be either sixteenth or 32nd notes.

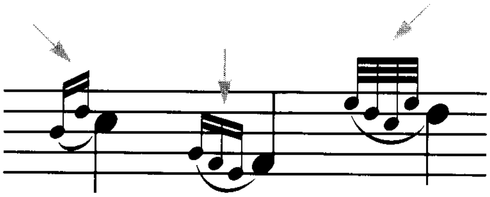

Larger groupings, such as five and six grace notes, are usually beamed as 32nd notes.

DIRECTION OF GRACE-NOTE SLURS

If a grace note precedes an interval or chord, the **slur will follow the direction of resolution** (from the grace note to the appropriate main note).

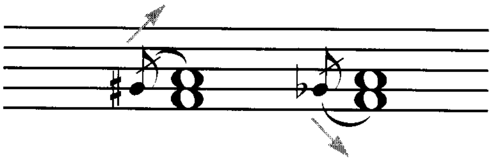

WITH LEGER LINES

If leger lines are used, the slur may be **placed above to avoid the leger lines.**

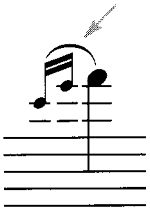

The grace-note leger lines can be the same thickness as staff lines, but shorter.

Grand pause

The grand pause or general pause (G.P.) over a measure of rest indicates that the **entire orchestra is silent,** often unexpectedly.

Grand staff

The grand staff is used to notate scores for instruments such as piano, organ, harp and marimba.

The grand staff consists of **two staves joined by a brace** and a systemic barline. (See *Brace*)

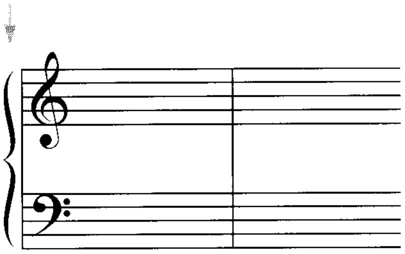

Although the "normal" clefs for the grand staff are treble above and bass below, **either staff may have either clef,** and may change clefs at any time. (See *Clef signs*)

FOR ORGAN

Organ scores use the grand staff connected to a staff used to notate music for the organ pedals.

Except for the systemic barline, **barlines are separated** between the manuals and pedal.

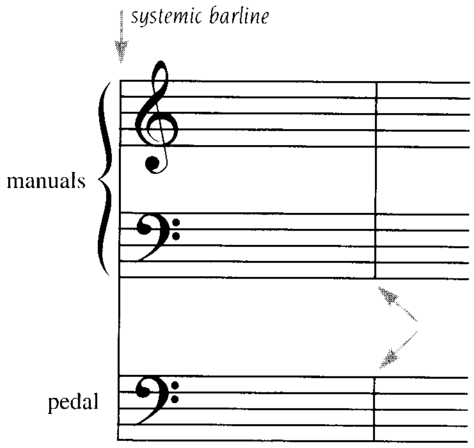

FOR INSTRUMENT & PIANO

Scores for an instrument (including voice) and piano use a **grand staff, with the instrument staff above.**

The staff above the grand staff is sometimes **cue size,** if the instrument has its own full-size, separate part.

Except for the systemic barline, **barlines are broken** between the instrument staff and the grand staff.

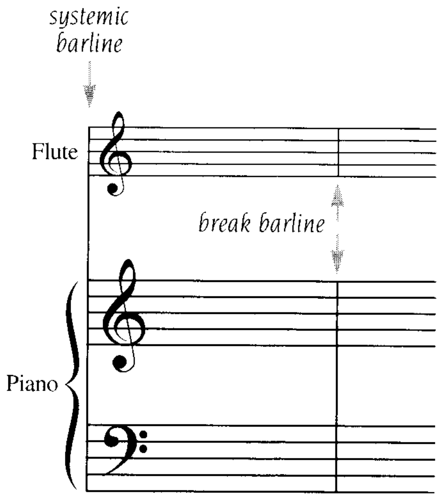

TWO-PIANO MUSIC

Two-piano music requires two grand staves joined together with a **systemic barline.**

The measure **barlines connect staves for each grand staff only.**

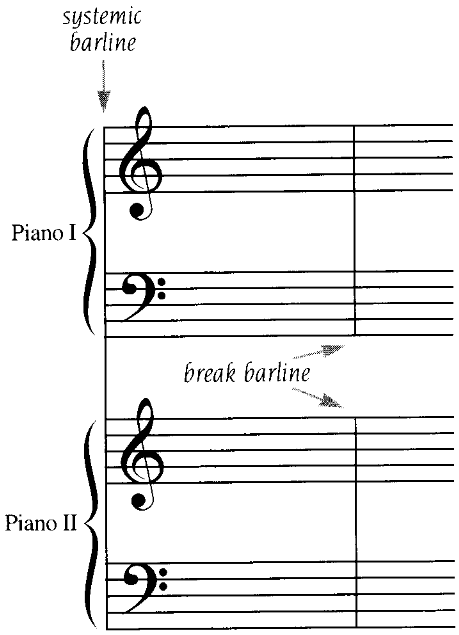
systemic barline
Piano I
break barline
Piano II

ONE PIANO—FOUR HANDS

Facing pages are required.

The **Secundo** part is notated on the left-hand page; usually both staves are in **bass clef.**

The **Primo** part is notated on the right-hand page; usually both staves are in **treble clef.**

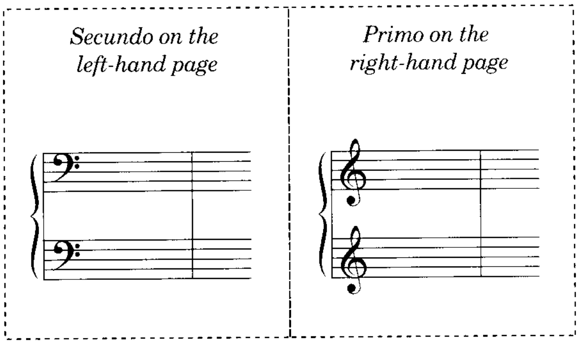
Secundo on the left-hand page
Primo on the right-hand page

The **page layout** for primo and secondo **is preferably identical**, system-for-system; if they are not identical, the facing pages *must* begin and end with the same measures.

Harmonics

NATURAL HARMONICS

Natural harmonics are indicated by a **small circle** centered above the notehead.

The indicated note is the SOUNDING NOTE.

Such notation indicates a **natural harmonic.**

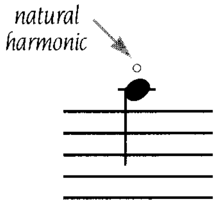
natural harmonic

Careful consideration of an instrument's capabilities is necessary. Only some instruments are capable of producing harmonics.

Another way to notate a natural harmonic (for strings) is to use a **diamond-shaped notehead.**

The position on the staff **indicates where the finger is placed**—NOT the sounding note.

The sounding note may be indicated in cue size and in parentheses.

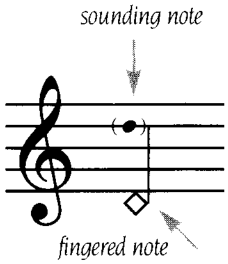
sounding note
fingered note

ARTIFICIAL HARMONICS (STRINGS)

A diamond-shaped notehead with a standard notehead comprises an artificial harmonic. The **standard notehead indicates the placed finger**; the diamond indicates where the second finger lightly touches the node to produce the harmonic.

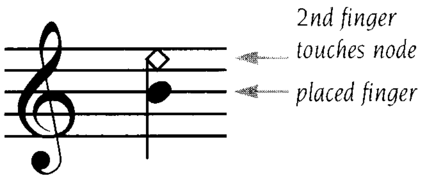
2nd finger touches node
placed finger

Often the sounding note is indicated in cue size and in parentheses.

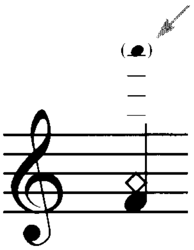

Key signatures

Key signatures appear **after the clef but before the time signature.**

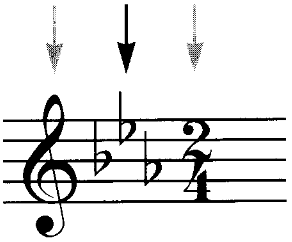

CHANGING KEYS

The preference is to make key changes at the beginning of a staff or system, with a **courtesy key signature** at the end of the previous one.

1. A key change is traditionally preceded by a **double barline.**
2. If the key change is at the beginning of a staff or system, a **courtesy key signature** is placed at the end of the previous one.
3. The **staff is left open** after the courtesy key signature.
4. The **new key signature is indicated** on the next staff.

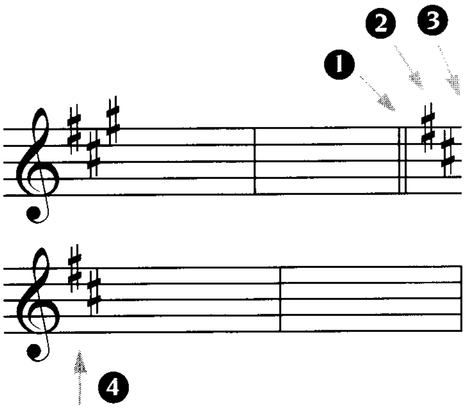

If a key **changes mid-staff:**

1. a double barline is added,
2. followed by the new key signature.
3. The new key signature continues on the next staff.

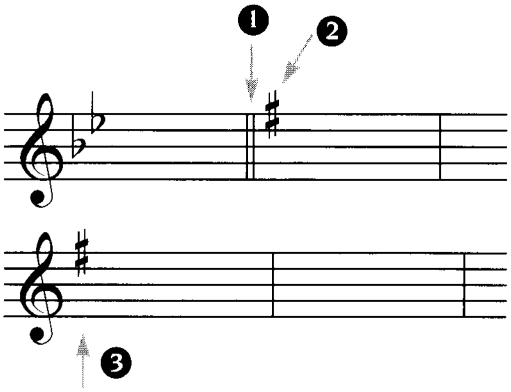

CANCELLATIONS

Cancellations are no longer considered necessary, unless the new key is C major or A minor (no sharps or flats). In that case, cancel the old sharps or flats with natural signs, in the same order as the old key signature .

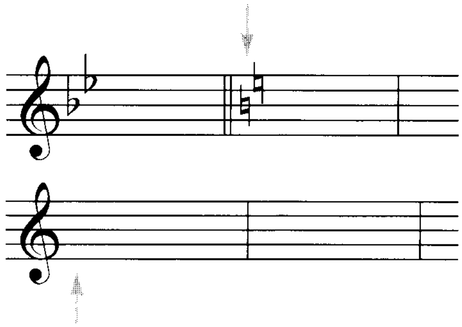

The following illustrates the positioning of the clef, key signature and time signature if all three change simultaneously.

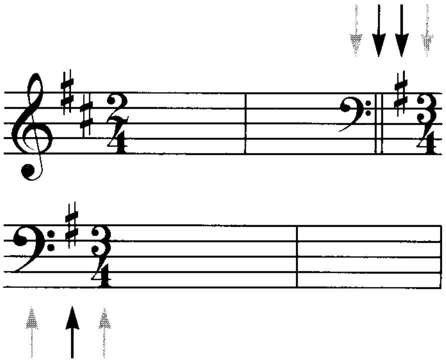

Key signatures in the **TREBLE CLEF**

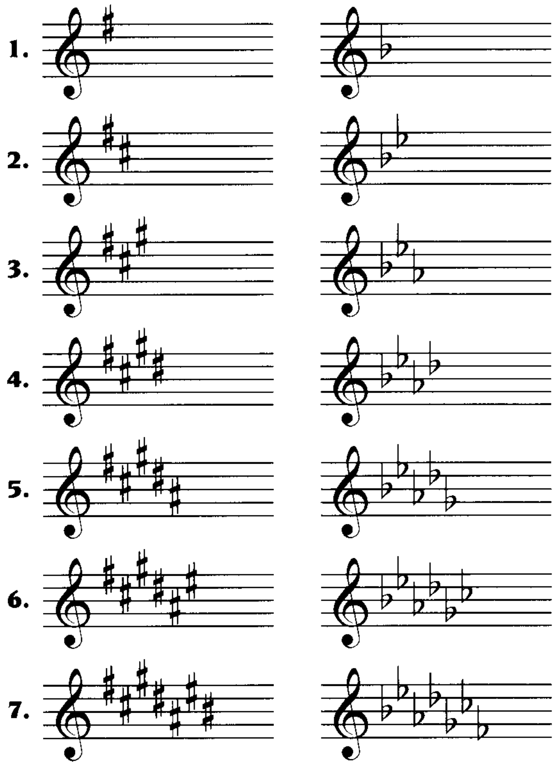

Key signatures in the **BASS CLEF**

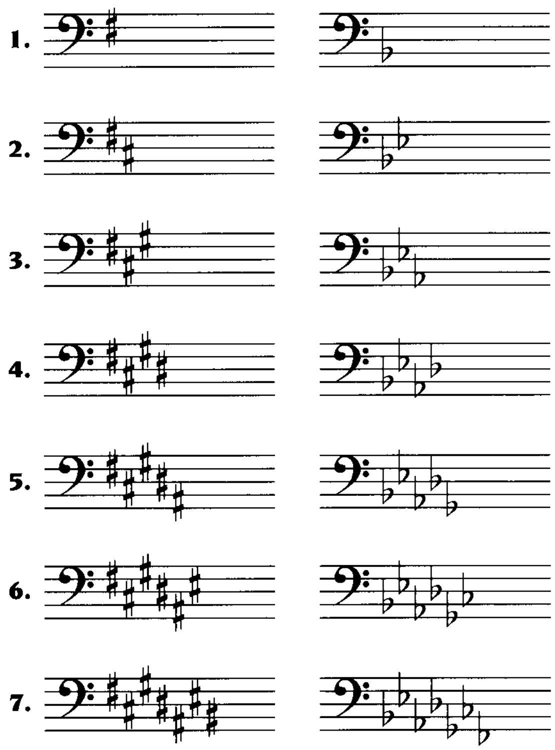

Key signatures in the **TENOR CLEF**

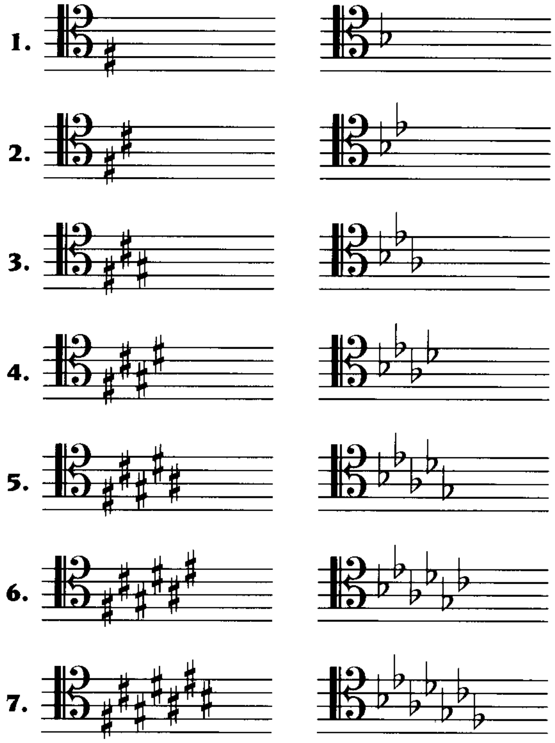

Key signatures in the **ALTO CLEF**

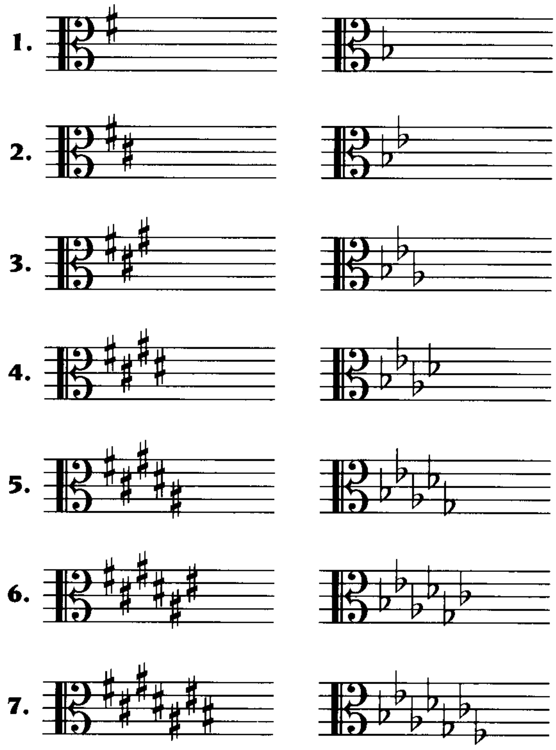

THE CIRCLE OF FIFTHS

The Circle of Fifths serves as a quick reference guide to the relationship of the keys and key signatures, and how they can be figured out in a logical manner.

The Circle of Fifths will also help to clarify which keys are enharmonic equivalents. (*Enharmonic* means notes that sound the same but are written differently.)

Clockwise movement *(up a perfect 5th) provides all of the sharp keys by adding one sharp to the key signature progressively.*

Counter-clockwise *(down a perfect 5th) provides the flat keys by adding one flat similarly.*

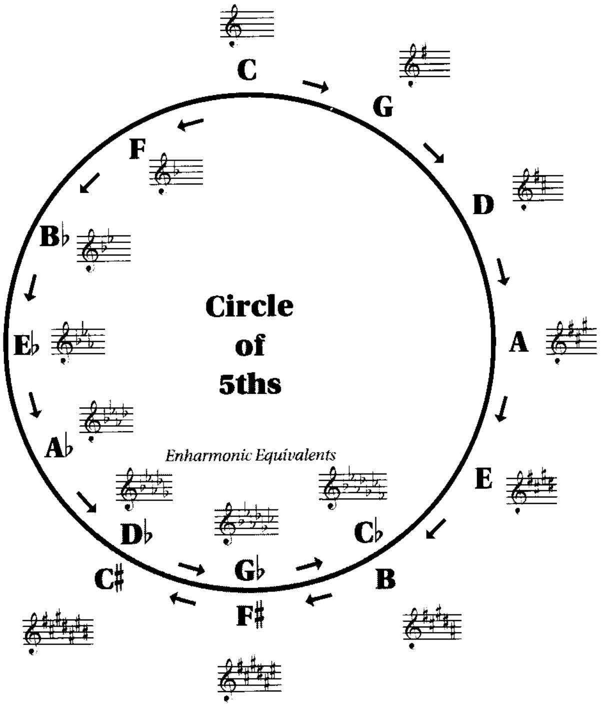

Layout

See *Spacing*

Leger lines

Leger lines are used to extend the range of the five-line staff.

Leger lines are the **same line weight as staff lines,** or slightly heavier.

The **vertical spacing** of leger lines must be identical to the vertical spacing between staff lines.

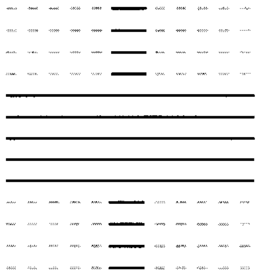

Leger lines **extend slightly past the notehead.** They will need to extend sufficiently enough to be seen.

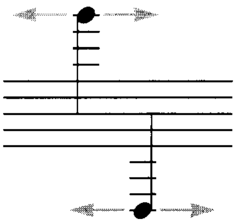

STEM LENGTH

Stems of all notes ABOVE or BELOW the first leger line past the staff **must extend at least to the middle staff line.**

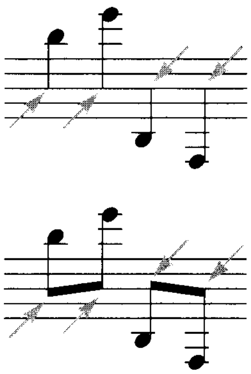

OVERCROWDING

To avoid leger lines that touch, adjustments must be made either in the **horizontal spacing** of the music or by **shortening the leger line length.** The first is preferred, but is not always possible.

In the first example below, the leger lines are overcrowded and appear to be touching.

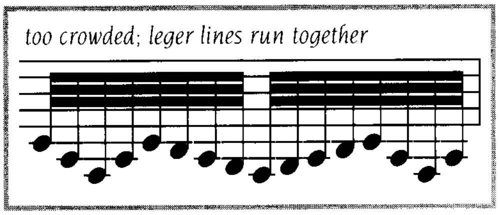

Corrected by shortening the leger lines to create a slight amount of space between the lines.

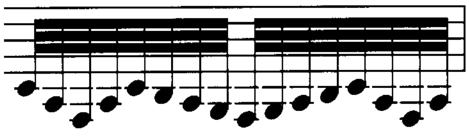

CLEF CONSIDERATIONS

Choose the appropriate clef for the situation. The viola, for instance, may be written in alto clef to avoid excessive leger lines. (See *Clefs*)

The following examples illustrate the same musical example in three different clefs.

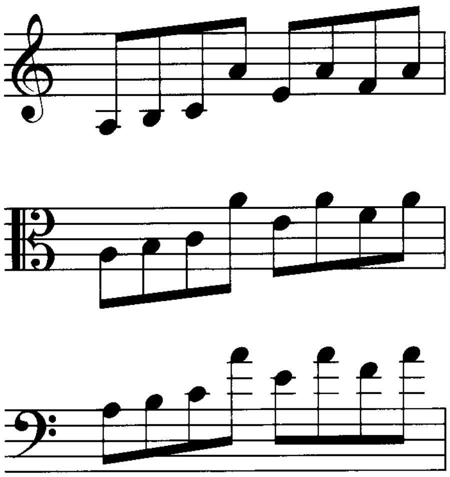

Lyric

Beaming of notes associated with a lyric now follows standard notational practice.

Traditional practice, now obsolete, was to use flags for eighth notes, sixteenth notes, etc. Beams were used only for melismas.

However, in several other ways the lyric makes the notation of a vocal part different from that of other parts.

PLACEMENT OF STAFF INDICATIONS

Dynamics and tempo indications go **above the staff** to avoid conflicting with the lyric.

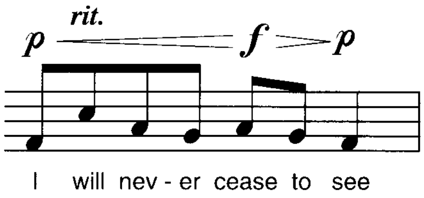

USE OF SLURS

Slurs are **seldom used in a vocal part for anything other than a *melisma*** (a term used here simply to mean a syllable or word that is sung to more than one note).

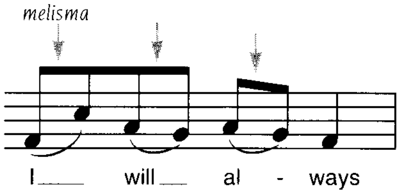

ALIGNMENT OF LYRIC FOR MELISMA

Another help in reading a melisma is the proper **alignment of the lyric.** The word or syllable, instead of being centered under the note, should be aligned **flush left with the left edge of the notehead** (see above).

EXTENDER LINES

In addition to the slur, **extender lines at the baseline of the lyric** help to **show the length of a melisma.** However, extender lines are only used for **one-syllable words, or for the last syllable of a word.** A hyphen is used in the middle of a word.

The **thickness of extender lines** should be less than that of the staff lines.

The **end of the extender line** should be **aligned flush right** with the last note of the melisma. *The most common mistake in the use of lyric extender lines is to extend the line for the full value of the note.*

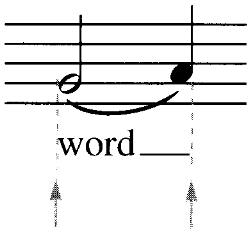

When an extender line is called for, there should always be one, even if the words are crowded and it has to be very short. (In this case, the line can extend slightly past the note.)

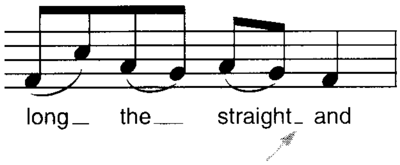

When an extender line wraps to the next staff, begin the line at the end of the key signature.

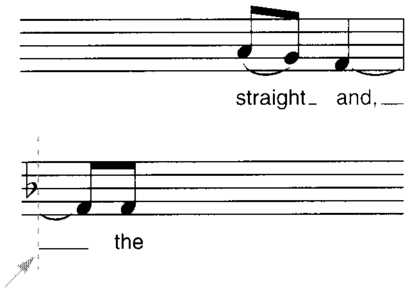

Notice the comma in the previous example. Any **punctuation always precedes an extender line.**

TIES

Ties are treated exactly like melismas, including the use of extender lines and alignment of lyric.

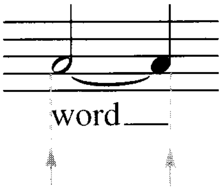

WORD PLACEMENT (NORMAL)

Words or syllables for notes other than melismas or tied notes are **centered to the notehead.** If the lack of space does not allow a word or syllable to be exactly centered, at least some part of the word or syllable should be under the note.

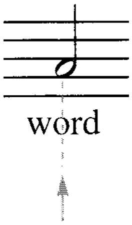

HYPHENATED WORDS

Center the *hyphen* (not a dash or extender line) between syllables.

If the melisma (or tie) is extremely long, **more than one hyphen** may be used.

If the hyphenated word **wraps to the next line, another hyphen *may* be used** at the beginning of the line.

Hyphenate words **according to the dictionary.** Do not hyphenate according to how the word *sounds* like it should divide. If there is any question at all, check the dictionary.

PUNCTUATION

Use **normal punctuation and capitalization.** (Sometimes capitalization is done as if the words were poetry, with each "line" capped, but this is becoming less common.) *A common mistake is to use little or no punctuation.*

MULTIPLE VERSES & REPEATS

Broken slurs

A **broken slur** is used when one verse has a melisma, especially if it is not the first verse.

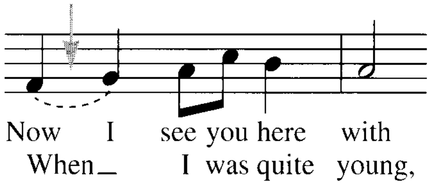

Treat the lyric for each verse separately. Notice in the above example that the melisma is aligned *as a melisma,* while the previous verse is aligned normally.

Omission lines

An omission line is sometimes used to clarify that there is **no word or syllable sung.**

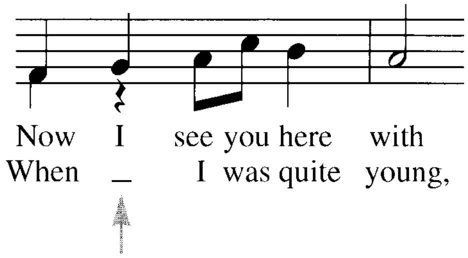

In many cases, however, there is a better way to treat the situation. In the previous example, the omission line could be mistaken for an extender line, so it would be better without it, letting the cue-size rest serve the same purpose.

Cue-size notes, rests

Use cue-size notes and rests **when a rhythm is different from the first verse** (see previous example). Stems and beams should also be cue size.

If there are **more than two verses** with different rhythms, the notation can become quite complicated. It may be better to write out the music.

Numbering

Multiple verses are numbered. There are several ways to treat these when there are pick-ups, repeats, first and second endings, etc. Be clear and consistent.

→ 1. Verse, verse, verse
→ 2. Verse, verse, verse
→ 3. Verse, verse, verse

Using a brace

When multiple lines converge into one line, such as a *chorus* or *refrain*, a **brace** is used.

→ 1. Verse, verse, verse
→ 2. Verse, verse, verse } chorus or refrain →
→ 3. Verse, verse, verse

Meter

The understanding of meter is essential to the placement of ties, rests, syncopated rhythms and proper beaming.

Time signatures identify the meter.

There are two overall classifications of meter: **perfect** and **imperfect** (odd) meter.

PERFECT METER

In perfect meter the measure can be divided into **equal halves or thirds**.

There are two classifications of perfect meter: **simple meter** and **compound meter.**

In **simple** meters each **BEAT** can be subdivided by 2.

In **compound** meters each **PULSE** can be subdivided by 3.

Metric subdivisions should not be confused with tuplets.

SIMPLE METER

The classification of **duple, triple** or **quadruple** reflects the number of **BEATS** in the measure.

Simple duple meter

Simple duple time signatures have a 2 as the upper number of the time signature.

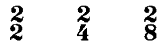

Each measure contains **2 BEATS.**

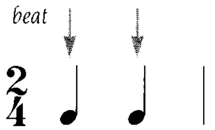

Each BEAT can be **subdivided by 2**.

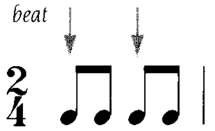

Only **rests of equal value to the beat** or less may be used (except in a complete **measure of rest,** which is indicated by a whole rest).

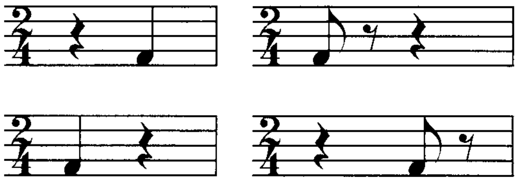

Rests of equal value to the beat are used ONLY on the beat.

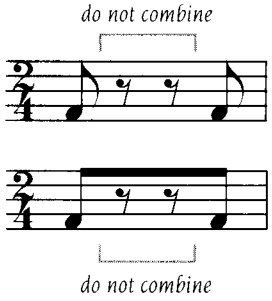

The following are all acceptable examples in simple duple meter.

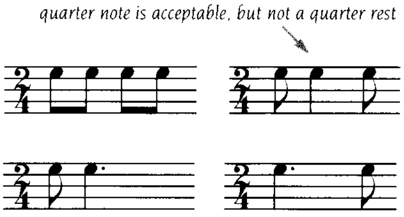

Simple triple meter

Simple triple time signatures have a 3 as the upper number of the time signature.

$\genfrac{}{}{0pt}{}{3}{2}$ $\genfrac{}{}{0pt}{}{3}{4}$ $\genfrac{}{}{0pt}{}{3}{8}$

Each measure contains **3 BEATS.**

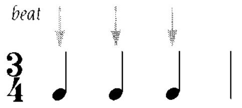

Each beat can be **subdivided by 2.**

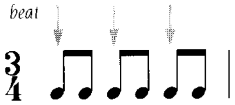

Only **rests of equal value to the beat** or less may be used (except in a complete **measure of rest,** which is indicated by a whole rest).

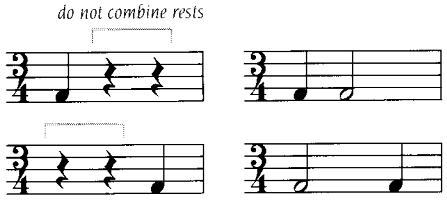

Beaming should reflect the simple triple meter, NOT compound duple meter.

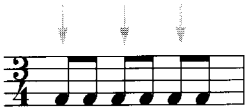

NOT

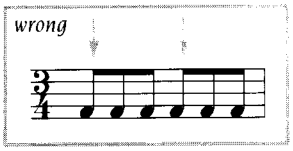

Simple quadruple meter

Simple quadruple time signatures have a 4 as the upper number of the time signature.

$\genfrac{}{}{0pt}{}{4}{2}$ $\genfrac{}{}{0pt}{}{4}{4}$ $\genfrac{}{}{0pt}{}{4}{8}$

Each measure contains **4 BEATS.**
Each beat can be **subdivided by 2.**

To successfully notate *simple quadruple* meter, it is best to consider the meter as a combination of **2 units of duple meter.**

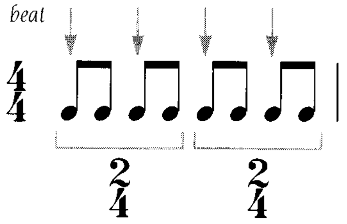

*In simple quadruple meter, these **2 units** must always be clearly distinguished. The only exception to this would be a very simple syncopation (such as ♩𝅗𝅥 ♩).*

The following examples illustrate the **subdivision of the 2 units.** The notation of each unit, separately, is identical to that of simple duple meter.

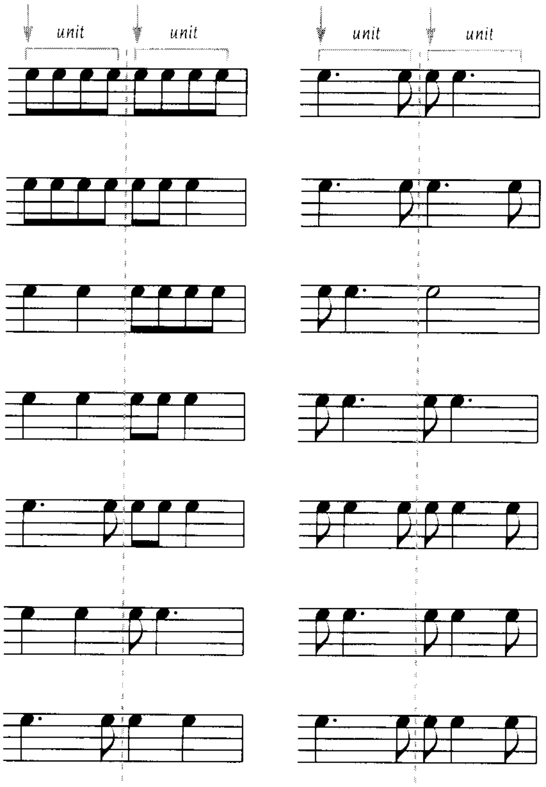

Whole rests are used to indicate a complete measure of rest.

Rests twice the value of the beat are used ONLY on beats 1 & 3 (the first beat of each half of the measure).

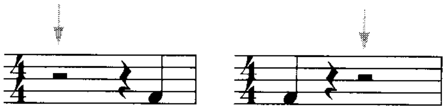

Rests of equal value to the beat can ONLY appear on the beat.

Rests less than the value of the beat are used freely as long as the two units are apparent.

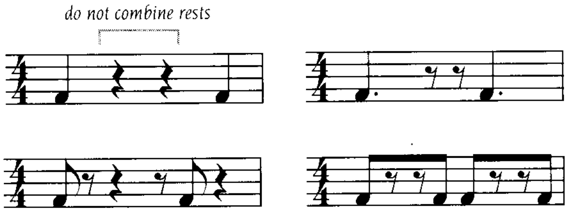

COMPOUND METER

The classification of **duple, triple** or **quadruple** reflects the number of **PULSES** in the measure.

Compound duple meter

Compound duple time signatures have a 6 as the upper number of the time signature.

$\begin{smallmatrix}6\\4\end{smallmatrix}$ $\begin{smallmatrix}6\\8\end{smallmatrix}$

Each measure contains **2 PULSES.**

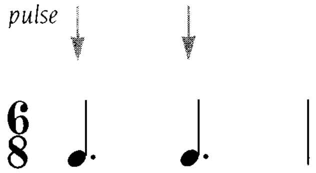

Each PULSE can be **divided into 3 BEATS.**

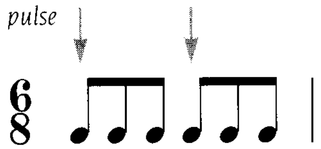

Only **rests of equal value to the pulse** or less may be used (except in a complete **measure of rest** which is indicated by a whole rest).

Rests of equal value to the pulse can ONLY appear on the pulse.

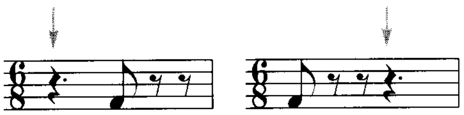

Rests of $2/3$ the value of a pulse may be used ONLY on the pulse.

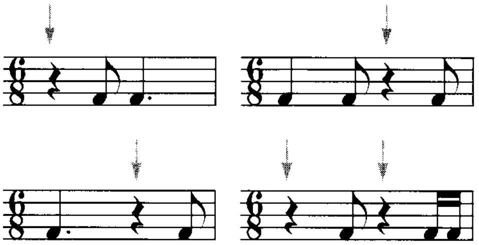

Rests of equal value to the beat, **not appearing on the pulse,** CANNOT be combined.

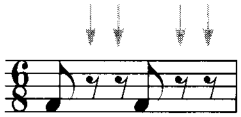

Beaming should reflect the compound duple meter, NOT simple triple meter.

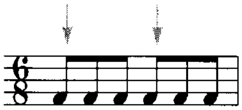

NOT

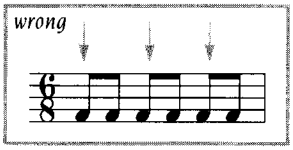

Compound triple meter

Compound triple time signatures have a 9 as the upper number of the time signature. An example is $\begin{smallmatrix}9\\8\end{smallmatrix}$.

Each measure contains **3 PULSES.**

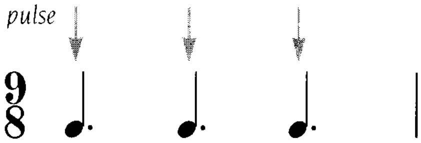

Each PULSE can be **divided into 3 BEATS.**

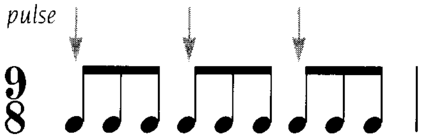

Rules governing **rests and beaming** are identical to compound duple meter.

Compound quadruple meter

Compound quadruple time signatures have a 12 as the upper number of the time signature. An example is $\frac{12}{8}$.

Each measure contains **4 PULSES.**

Each PULSE can be **divided into 3 BEATS.**

To successfully notate *compound quadruple* meter it is best to consider the meter as a combination of **2 units of compound duple meter.**

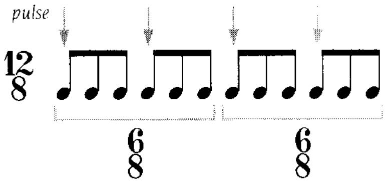
pulse
12/8
6/8
6/8

Rules governing **rests and beaming** are identical to compound duple meter.

IMPERFECT METER (ODD METER)

In imperfect meter (odd meter) the measure **CANNOT be divided into equal halves or thirds.**

Imperfect meter can be considered as any combination of **duple or quadruple meter** AND **triple meter.**

$\frac{7}{8}$ time can be considered a combination of $\frac{2}{4} + \frac{3}{8}$ or $\frac{3}{8} + \frac{2}{4}$.

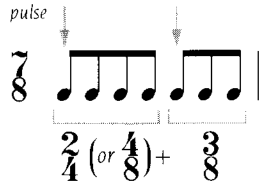
pulse
7/8
2/4 (or 4/8)+ 3/8

Notice that 7 is not divisible by 2 or 3.

MODERN APPROACHES TO METER

A **broken barline** is sometimes added to clarify imperfect meters.

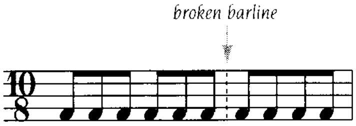
broken barline
10/8

A **composite meter** is another option to clarify the subdivision of imperfect meters.

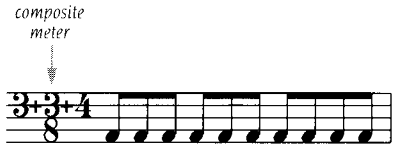
composite meter
3+3+4/8

When two or more measures alternate meters in a manner that repeats predictably, a time signature reflecting the **alternating meter** may be used.

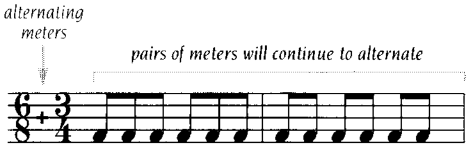
alternating meters
pairs of meters will continue to alternate
6/8 + 3/4

Metronome marks See *Tempo marks*

Natural sign See *Accidentals*

Notes

Traditionally there are five main shapes representing noteheads:

Pitched	Non-pitched
○ *whole note*	◇ *whole note, half note*
○ *half note*	
● *quarter, eighth, sixteenth, 32nd, etc.*	X *quarter, eighth, sixteenth, 32nd, etc.*

PLACEMENT ON THE STAFF

On a **five-line staff,** notes are either in a space or on a line.

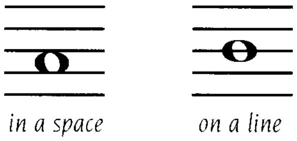
in a space
on a line

On a **single-line staff,** the notehead usually intersects the line, with the stem always up.

When two instruments share a staff, one part sits on the line; the other hangs from the line.

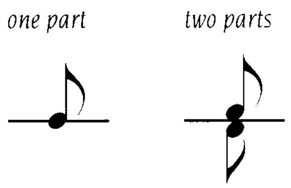

STEMS AND NOTEHEADS

Double whole notes (or *breves*) and **whole notes** are without stems.

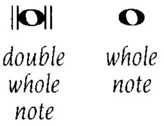

Half notes and **quarter notes** are always stemmed.

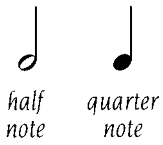

Single **eighth, sixteenth, 32nd and 64th notes** are always stemmed and with a flag(s).

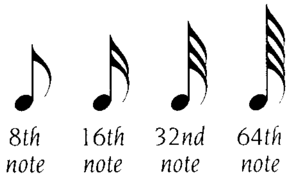

Two or more eighth, sixteenth, 32nd and 64th notes are usually **beamed into groups.**

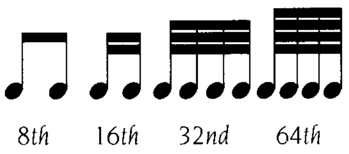

An **X** notehead is used primarily in **non-pitched percussion** music. It is also seen in vocal music to indicate approximate-pitched singing *(Sprechstimme),* rhythmically spoken parts, clapping, etc.

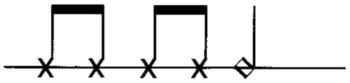

The diamond-shaped notehead serves as a half note (with stem) or whole note (without stem).

When used as a **harmonic,** the diamond-shaped **notehead is always open,** whatever the rhythmic value.

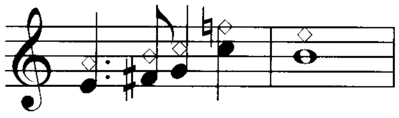

In some software programs the characters for the diamond-shaped notehead and the harmonic are slightly different.

Octave signs

The use of octave signs is primarily to **avoid multiple leger lines.** While frequently used in piano music, octave signs are only selectively used in instrumental music.

8va or *all'ottava* means "at the octave."

8va (or sometimes *8*) is indicated above a **treble clef staff only.** Do not use above any other clef.

Usually accompanied by an extender line—**a broken line with a downstroke** at the end.

The alignment of the *8va* is centered above the notehead.

The end of the *8va* bracket **extends slightly past the last note** affected.

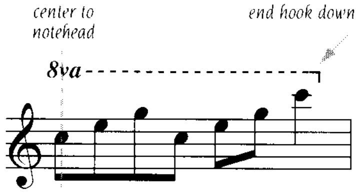

8va bassa or *ottava bassa* means "at the octave below."

*8va (or 8) below the staff, especially in keyboard music, is always understood to mean **8va bassa.***

8va bassa is indicated below a **bass clef staff only.** Do not use for treble, alto or tenor clefs.

Usually accompanied by an extender line—**a broken line with an upstroke** at the end.

The beginning of the ***8va bassa*** is placed below the notehead.

The end of the ***8va bassa*** line **extends slightly past the last note** affected.

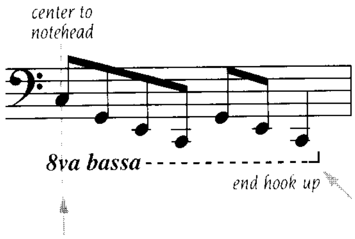

The abbreviation ***8vb*** is *only* a copyist's shorthand and **should not be used in engraved music.**

PLACEMENT CONSIDERATIONS

The ***8va*** or ***8va bassa*** with its extender line should be **clearly placed,** to avoid conflict with as many musical elements as possible and yet be as close to the notes affected as possible.

In the example below, the 8va is placed entirely above the slur.

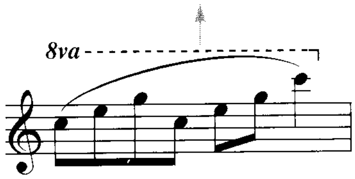

In this example, the 8va intersects the slur due to a possible limitation of space.

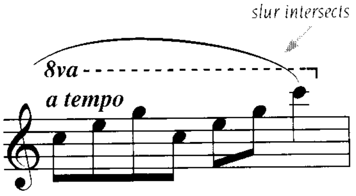

WITH ACCIDENTALS

Accidentals must be repeated for the notes included within ***8va*** or ***8va bassa*** octave alterations.

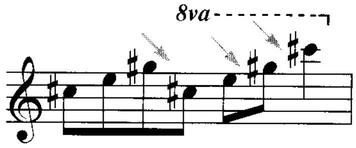

WRAPPING FROM STAFF TO STAFF

The ***8va*** or ***8va bassa*** **breaks at the end of the first staff,** aligned with the right barline.

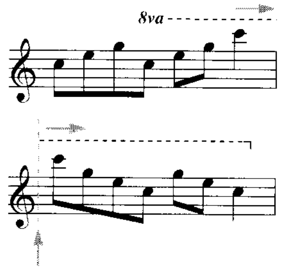

The extender line **begins again just past the clef** or key signature.

A **courtesy *8va*** or ***8va bassa*** may also be indicated at the beginning of the second staff when wrapping from staff to staff.

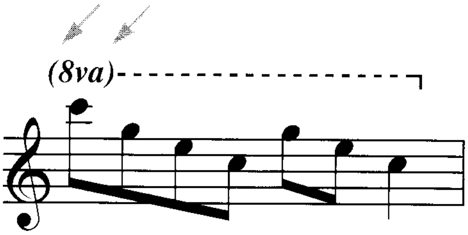

LOCO

Although the ending of the extender line should suffice, the word ***loco*** (meaning "at place") may be added as a reminder that the ***8va*** or ***8va bassa*** no longer applies.

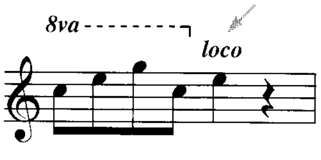

ON AN ENDING NOTE

The *8va* or *8va bassa* **is centered to the notehead** and, since it is the last note of the piece, it may or may not be accompanied with a downstroke (or upstroke).

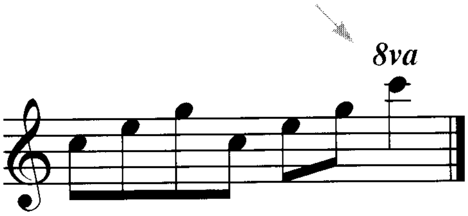

If a single note is altered but is **not the last note** of the piece, an **upstroke (or downstroke) *must* be added.**

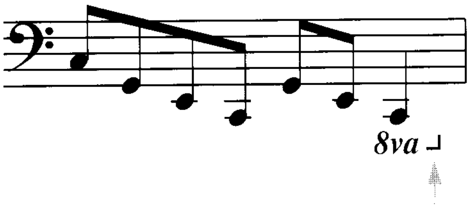

SPECIAL SITUATIONS

The following examples are incorrectly notated. The *8va* or *8va bassa* is **never placed between the two staves** of a grand staff.

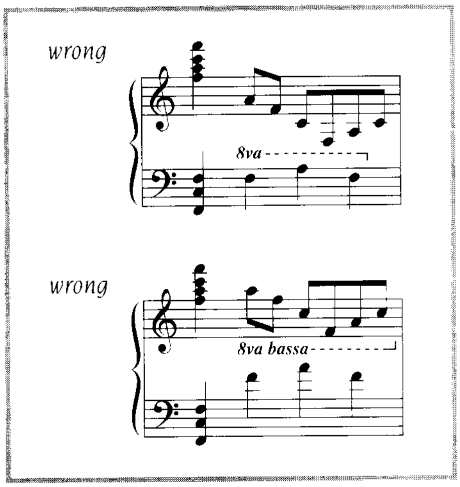

These two incorrect examples would need to be rewritten—perhaps in one of the following ways:

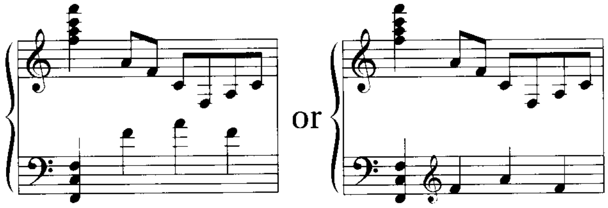

WITH RESTS

If the *8va* or *8va bassa* is interrupted with short rests, the **extender line may continue over** them.

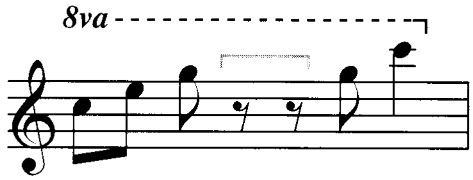

If the rests are of a **larger value, the line should be broken** and reintroduced.

When **rests precede** the *8va* or *8va bassa,* begin on the first note after the rests.

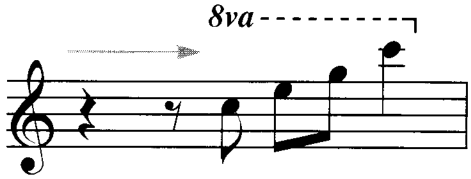

When **rests follow** the *8va* or *8va bassa,* the octave alteration ends with the last note in the grouping—the rests should not be included in the grouping.

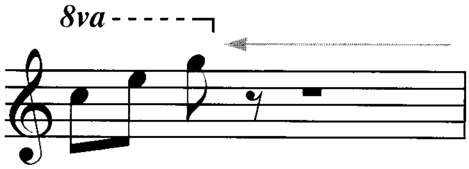

CUE SIZE

For any passage at **cue size** the *8va* will likewise be at cue size.

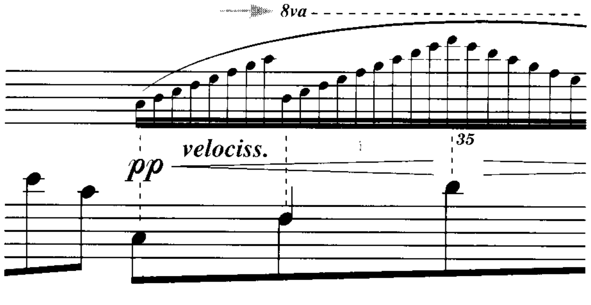

15MA

15ma or *quindicesima* means "at the fifteenth" (two octaves). The situations to use ***15ma bassa*** are very rare.

All rules for ***8va*** apply to ***15ma.***

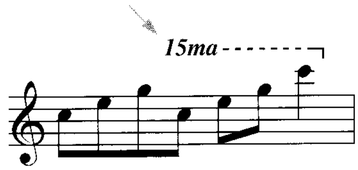

COLL' 8VA

Coll' 8va (or ***coll' 8***) is an abbreviation for *coll' ottava* meaning **"with the octave."** It is a shorthand system of writing octaves, **discouraged in traditional engraving.**

The function of ***coll' 8va*** differs from ***8va*** in that the written notes are **doubled at the octave.**

For the extender line, a longer dash length or a solid line is recommended, to make the ***coll' 8va*** distinguishable from ***8va.***

The ***coll' 8va*** doubles at the **octave above** when notated in the **treble clef.**

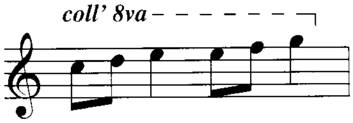

is played:

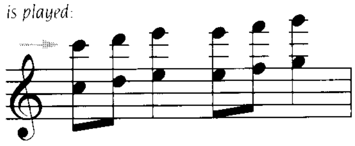

The ***coll' 8va*** doubles at the **octave below** when notated in the **bass clef.**

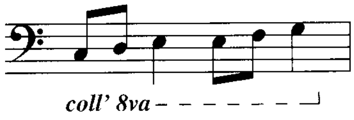

is played:

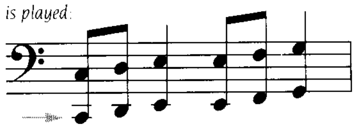

INSTRUMENT CONSIDERATIONS

ORCHESTRA / BAND SCORES

Often a score will have **limited space,** causing leger lines to become too crowded and illegible. Octave signs help to **eliminate vertical overcrowding.**

WOODWINDS

Wind players prefer to read leger lines.

STRINGS

Octave signs should be used very sparingly for viola and cello. Violins are accustomed to reading either leger lines or ***8va*** (but not ***15ma***).

KEYBOARD INSTRUMENTS

Celesta uses ***8va*** but not ***8va bassa*** (the range does not extend low enough).

Harpsichord rarely uses ***8va*** or ***8va bassa.*** Leger lines are more commonly used.

Piano uses ***8va*** and ***15ma*** in the treble clef and ***8va bassa*** in the bass clef.

Organ uses ***8va*** but not ***8va bassa*** (the range does not extend low enough).

Ornaments

Ornaments are primarily used in Baroque, Classical or Romantic period music. The interpretation of an ornament **varies according to the musical period** or composer. It is advised to consult detailed references on ornament usage and interpretation.

Ornament symbols are **placed above the staff** regardless of stem direction.

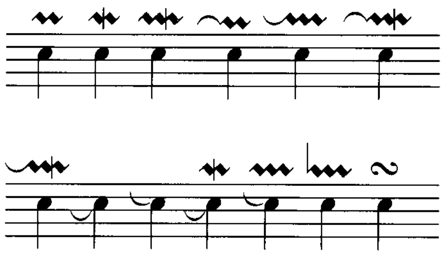

When two parts share a staff, an ornament symbol applying to the lower part is placed below the staff.

Ornaments are usually **centered** to the notehead. **Depending upon the position** of the ornament, the realization of the ornament will vary.

In the example below, the turn symbol is presented two ways, centered above the note and positioned after the note.

Notice the different realizations of the ornament.

this turn CENTERED *above the notehead indicates:*

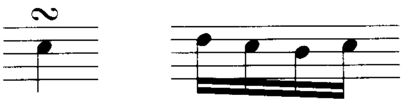

this turn placed AFTER *the note indicates:*

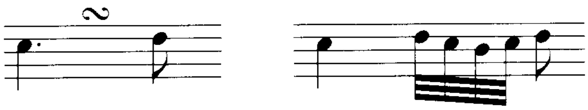

Ossia

Ossia is an indication for an **alternate version,** which may be easier or more difficult. It is treated the same as a realized ornament.

The ossia passage is **placed above** the full-size staff. The passage is cue size (65–75%).

Only the **passage concerned is notated**—the staff begins directly before and ends directly after.

Align all beats.

ossia (more difficult)

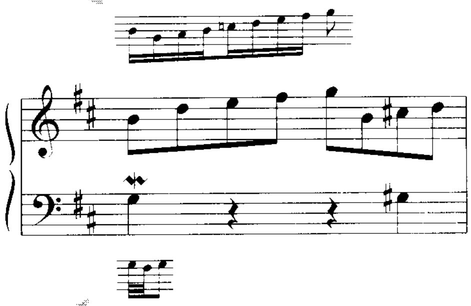

realized ornament

Pedal marks

FOR THE PIANO

There are three pedals on a grand piano: the **damper pedal** (right), the **sostenuto pedal** (center) and the **una corda pedal** (left). Some upright pianos have only two pedals—the damper and una corda.

Damper pedal

Markings for the damper pedal (sustain pedal) are **always placed below the grand staff** and are usually placed below all other musical elements.

The most commonly used pedal marks consist of three elements: *pedal down, pedal change and pedal up.*

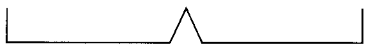

pedal down *pedal change* *pedal up*

The **pedal-down marking** begins vertically aligned with the left edge of the notehead.*

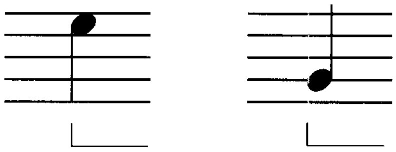

begin left of the notehead

The *point* of the **pedal-change marking** is vertically aligned with the center of the notehead.

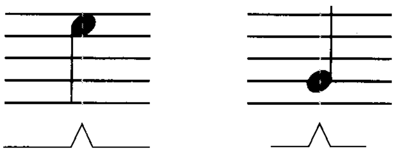

centered on notehead

*Some publishers center or align flush right. In any case, a style decision should be made.

The **pedal-up marking** is vertically aligned to the left of the notehead.

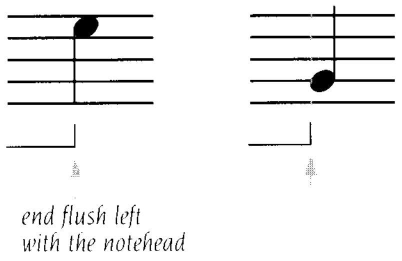
end flush left
with the notehead

The traditional method of pedal marking consisted of **pedal down** and **pedal up.**

This method is somewhat obsolete.

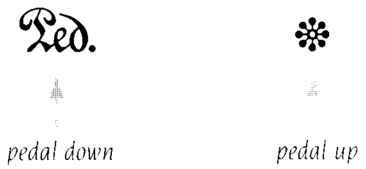
Ped.
pedal down
pedal up

Drawbacks of this method of pedal indication include the potential for inaccuracy and a cluttered look.

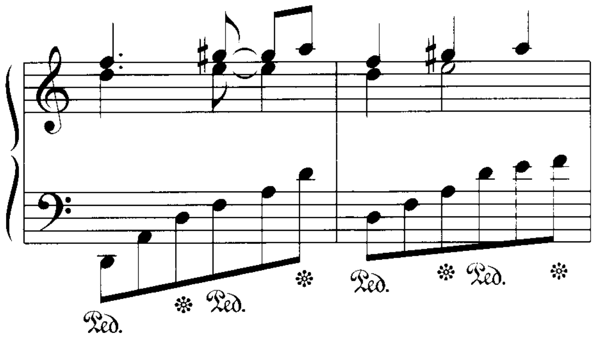
Ped.
Ped.
Ped.
Ped.

Flutter Pedal

To notate the rapid fluctuating of the pedal, a **zigzag pattern** interrupts the usual pedal marking.

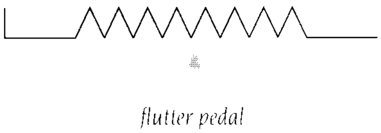
flutter pedal

Wrapping from staff to staff

The pedal mark is interrupted at the end of the staff. It continues on the next staff, **beginning immediately after the key signature.**

If the pedal mark continues to the end of the piece, it is vertically aligned with the thin line of the final double barline.

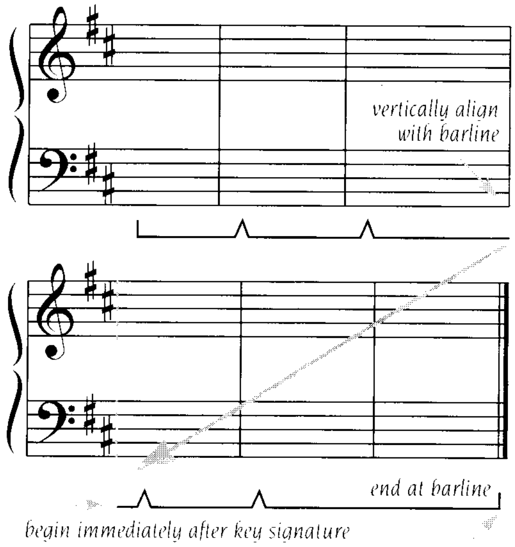
vertically align
with barline
end at barline
begin immediately after key signature

Simile

When the pattern **has been clearly established,** ***simile*** may be used instead of continuing the pedal marks. (See *Simile*)

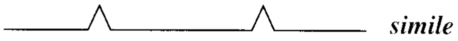

If the pattern changes, simply begin the pedal marks again.

Half pedal

The clearest indication is to simply **add "1/2" to the pedal marking,** whether at the down pedal or at the pedal change.

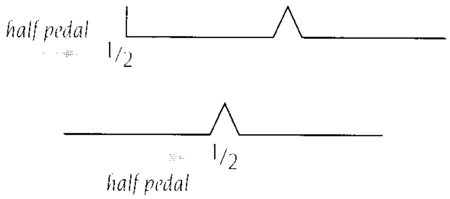
half pedal
1/2
1/2
half pedal

Una corda pedal

The indication ***una corda*** (one string) or ***u.c.*** is placed where the pedal is pressed; ***tre corde*** (three strings) or ***t.c.*** indicates the return to normal.

Sostenuto pedal

This is indicated by the abbreviation ***sos.*** (sometimes in combination with a bracket).

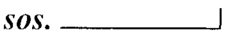

FOR THE ORGAN

Two symbols are used, one for the *heel,* one for the *toe.*

U Λ

heel *toe*

The symbols are **NOT inverted** when moved from one side of the staff to the other.

The pedal marks for the right foot are always **placed above the staff and centered on the notehead.**

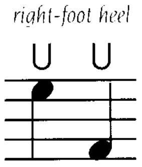
right-foot heel
U U

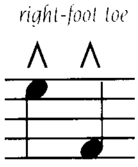
right-foot toe
Λ Λ

The pedal marks for the left foot are always **placed below the staff and centered on the notehead.**

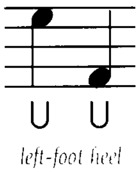
U U
left-foot heel

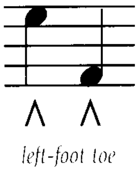
Λ Λ
left-foot toe

HARP PEDALS

Pedal indications for harp are **settings for tuning.**

The left and right foot have **separate sets** of pedals.

Notation for the harp is closely linked to the position of each of these pedals. Therefore, **enharmonic spellings** are frequent in harp music.

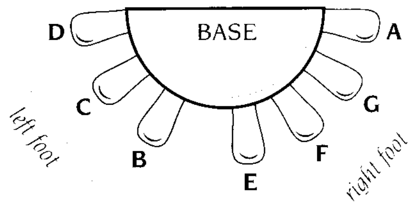
BASE
D
C
B
E
F
G
A
left foot
right foot

The tuning of the harp is based on the three **notch positions** of the pedals.

When a pedal is in its resting position, **center,** the string is tuned to the note's letter name (D, C, B, E, F, G & A).

When a pedal is **raised,** the note is lowered a half step (loosens the string).

When a pedal is **lowered,** the note is raised a half step (tightens the string).

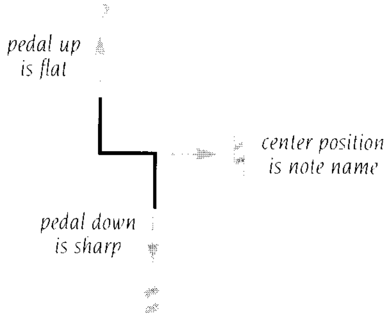
pedal up is flat
center position is note name
pedal down is sharp

The following three examples illustrate possible ways to indicate harp pedal settings in a score.

They may be listed **horizontally, stacked or using the Salzedo diagram.**

D♭ C♯ B♭ E F♭ G A♯

or

D♭ C♯ B♭
E F♭ G A♯

or

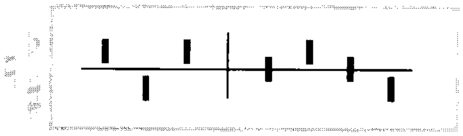

the Salzedo diagram

Picking

Picking is notated with *down-stroke* and *up-stroke* symbols. The same symbols may be used as those that are used for down-bow and up-bow.

Down-stroke and **up-stroke** indications are centered to the notehead and are always placed above the staff.

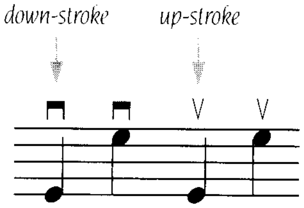
down-stroke
up-stroke

Repeat signs

Indications for repeats occur on three levels: **beat repetitions, measure repetitions and section repetitions.** The first two are used almost exclusively in manuscript or in popular music.

BEAT REPETITIONS

Angled slashes following a single beat indicate that whatever is written on the **first beat is to be repeated.** These are commonly **used in manuscript.** Do not use in engraved music.

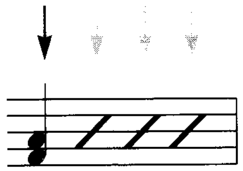

Another use of beat repeats is in **guitar or rhythm parts,** where they help pinpoint where chord changes occur. These are used in engraved music. Stems, beams and ties may be used when necessary.

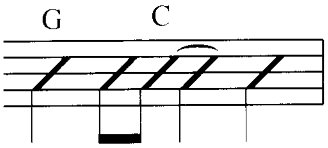

ONE-MEASURE REPEAT

This sign (𝄎), meaning **repeat the previous measure,** consists of a slash with two dots and is centered in an otherwise empty measure. In engraved music, its use should be limited to rhythm parts.

When several consecutive measure repeats are used, numbering is helpful. **Place a small number** above the staff and **center it on the measure repeat sign.**

The **notated measure is considered the first measure,** with the actual numbering beginning on the first repeated measure.

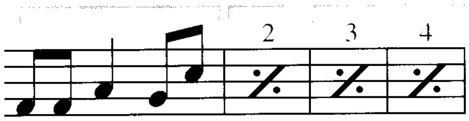

Write out the music **at key points** such as rehearsal numbers, page turns, etc.

TWO-MEASURE REPEAT

When **two measures are to be repeated,** a sign consisting of *two* slashes with dots is placed **on the center bar line** of two otherwise empty measures, **with the number 2 above it.** The number is in the same typeface and size as a time signature number.

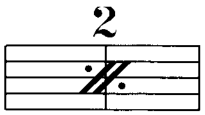

This sign should *not* be used when one measure is to be repeated twice. It should only be used for **two consecutive measures that are *different*,** and are to be repeated. In engraved music, it should only be used in rhythm parts.

If more than two measures are to be repeated, the **repeat sign** should be used.

SECTION REPETITIONS

The **repeat sign** is used to indicate the repeat of one or more measures, or an entire section.

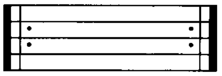

The repeat sign looks like a final double barline, with two dots positioned in the spaces above and below the center line of the staff.

The right-facing repeat sign is **placed at the beginning** of the first measure of the section to be repeated. The left-facing sign is **placed at the end** of the last measure of the same section.

Thus, the section that is to be repeated is **framed** by the inward-facing repeat signs.

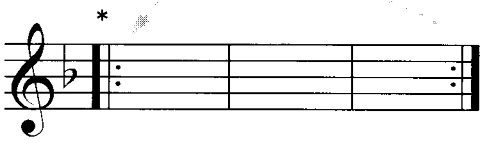

*If the repeat begins **at the beginning of a piece, the right-facing sign is not needed.**

For **two adjoining repeated sections,** repeat signs are placed, as before, framing each section. The **adjoining repeat signs may share the thick bar.**

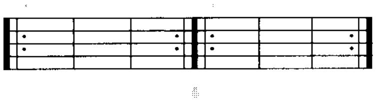

With key changes

When a key signature change occurs at the beginning of a repeated section, a double bar is added and the **key signature is placed before the first repeat sign**.

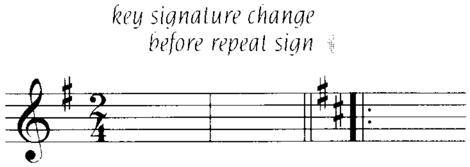
key signature change
before repeat sign

With multiple endings

A section that is repeated may have a different ending each time it is played.

The repeat sign is placed at the beginning of the section and at the last bar line of the 1st ending.

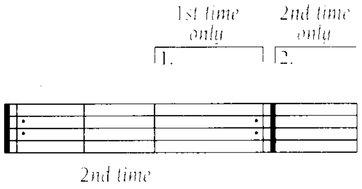
1st time only
2nd time only
1.
2.
2nd time

The **brackets do not intersect or touch the staff** but are placed above the staff, avoiding all musical elements. They are aligned vertically with the barlines.

1st ending brackets are **always closed.**

If the **2nd ending** is **not the end of the piece,** the bracket will be **open-ended.**

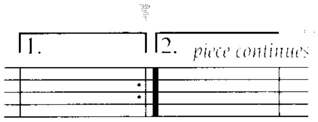
1.
2. piece continues

If it is the **end of the piece** (or movement), the bracket will be **closed.**

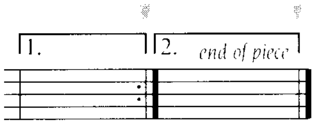
1.
2. end of piece

If a section is played three or more times but there are only two endings, number the brackets accordingly. This occurs often when there are lyrics.

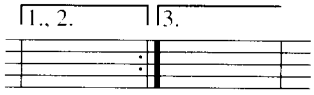

Rests

This section deals mainly with the positioning of rests. For more on the use of rests, see *Meter* and *Augmentation dot.*

TABLE OF RESTS

In the listing of rests below, each rest is half the value of the previous rest, starting with the double whole rest.

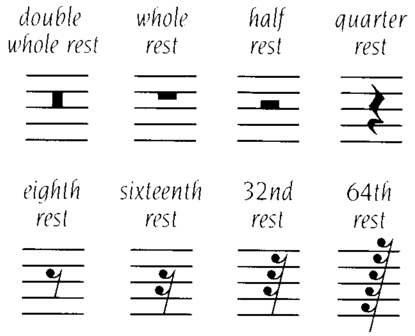
double whole rest
whole rest
half rest
quarter rest
eighth rest
sixteenth rest
32nd rest
64th rest

DOUBLE WHOLE REST (BREVE)

The double whole rest is placed between the third and fourth staff lines, touching both lines.

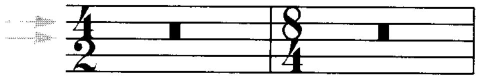

WHOLE REST

With the exception of $\frac{4}{2}$ and $\frac{8}{4}$, the whole rest is used to signify a **complete measure of rest** in any time signature.

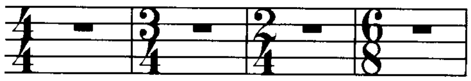

It **hangs** from the fourth staff line.

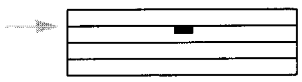

Whole rests are **centered in the measure.**

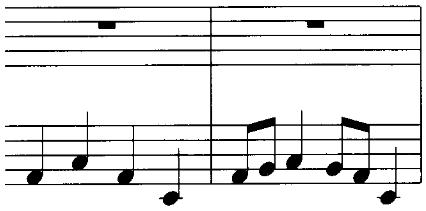

When two parts share a staff, whole rests are positioned on the **fifth line for the top** part and the **first line for the bottom.**

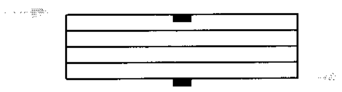

If either part conflicts with the placement of the rest, the rest may be **positioned outside the staff.**

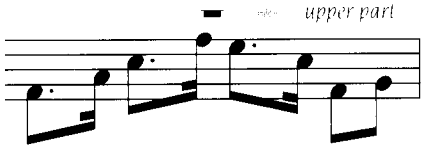

HALF REST

The half rest is positioned to **sit on the third staff line.**

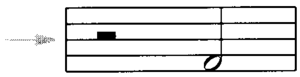

When two parts share a staff, the half rest will sit on **either the fifth line or the first line.**

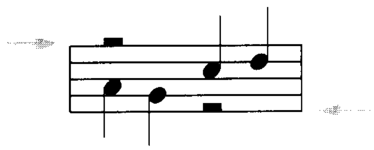

If either part conflicts with this placement, the rest may be positioned **outside the staff.**

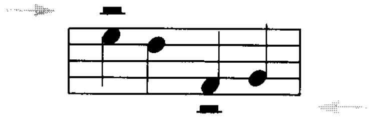

Half rests are **positioned at the beginning of beat 1 or 3**—never centered between two beats.

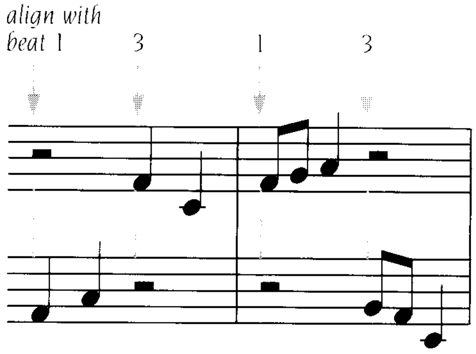

In the above example the use of half rests instead of two quarter rests is correct. However, meter must be considered—do not automatically substitute a half rest for two quarter rests.

(See *Meter)*

QUARTER REST

The quarter rest is positioned as follows. Notice the position of the bottom hook.

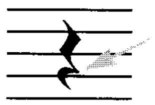

When two parts share a staff, the quarter rest is moved **up or down** from its normal position. The **relationship** of the rest **to lines and spaces** must be the same as in normal position.

The rest may be completely outside the staff when necessary.

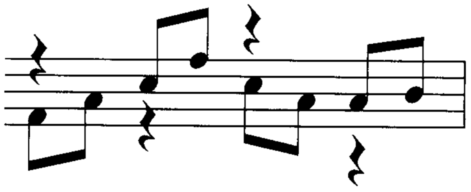

In some situations the above rests could be left out entirely. In other situations, if *both parts* were resting, one rest in normal position could serve for both.

EIGHTH, SIXTEENTH & 32ND RESTS

The guidelines for positioning quarter rests in the example above also apply for eighth, sixteenth and 32nd rests.

The ***hook* of these rests is always placed in a space** (or between imaginary leger lines if placed outside the staff).

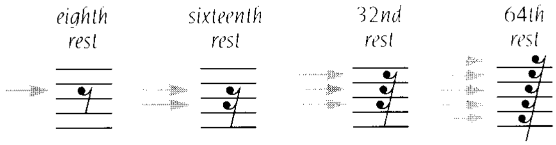

If one of the above rests **interferes with a beam, the rest moves up or down** so that stem lengths may be more normal. Other rests in the measure remain in normal position.

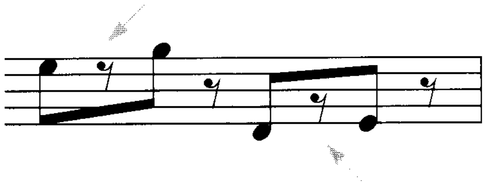

MULTIPLE-MEASURE RESTS

For a rest of more than one measure, a **horizontal line** (the thickness of a beam) is placed on the middle staff line, with a thinner vertical line at each end extending from the second to the fourth staff line.

A number the size and font of a time-signature number is placed above the staff and centered in the measure.

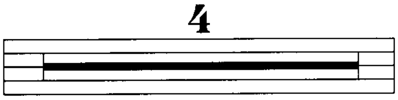

Break the multiple-measure rest for rehearsal numbers, tempo changes, etc.

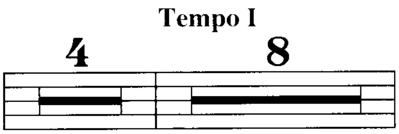

RESTS ON A SINGLE-LINE STAFF

When there is only one part on a staff, position the rests as follows:

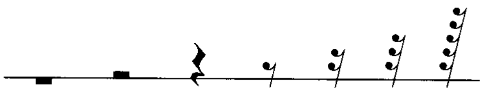

When **two parts share a single-line staff,** place the rests above and below the line, not touching the line. Use a leger line for whole and half rests.

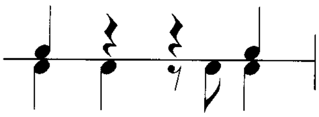

Scores

On a complete score all parts are included on a system and are arranged in a certain order.

All information contained on the parts and score must be **identical.**

On the first page of the score the system (or first system) is **indented.** Brackets connect instrument *choirs.* (See *Bracket, Systems*)

The **full name of each instrument** is given to the left of the first system.

Abbreviations are substituted for the full instrument names on subsequent systems.

Rehearsal numbers (indicated at the top of the score) serve as key reference points.

They may be **letters** or **measure numbers** enclosed in a circle or square.

All **extracted parts** will have **identical rehearsal** numbers.

SCORE ORDER

Instruments should be arranged in what is known as *score order.*

Woodwinds
Brasses
Percussion
Other instruments
Strings

Each **family** of instruments is predominantly **grouped together** from highest-pitched instruments to lowest-pitched instruments, with some traditional exceptions.

The French horns, although not the highest-pitched, are traditionally placed above the other brass instruments.

Sharing a staff

In band or orchestra music some parts of instrument families share the same staff. Combine parts 1 & 2 or 2 & 3; avoid combining three instruments on one staff.

BEGINNING A PASSAGE

If parts are divisi, no indication is necessary unless one or both instruments (such as strings) can play more than one note at a time. (See *Divisi*)

Indicate ***a2*** above staff if both parts play in unison. Use only one stem per note.

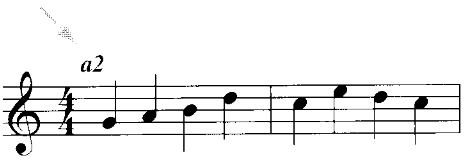

If the unison continues after a page turn, repeat ***a2*** in parentheses.

If like instruments are sharing a staff and only one part plays, indicate which part: I. or II.

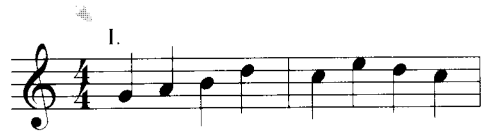

If unlike instruments are sharing the staff and one part plays, indicate which instrument is playing.

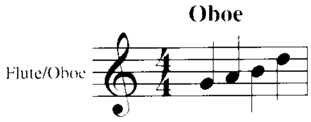

STEMMING OF TWO PARTS

If two parts are playing **different pitches but the same rhythm,** they share **one stem.**

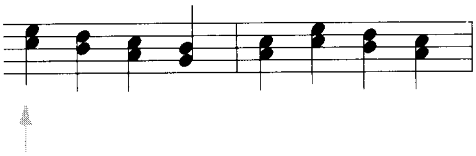

If the parts are **alternating between unison and divisi, and share the same rhythm,** opposite stem direction makes the unisons clear.

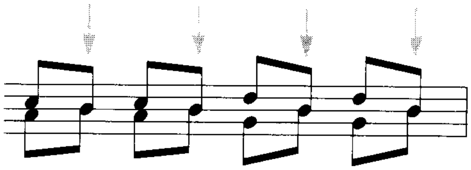

If both parts have **different rhythms,** opposite stem direction is necessary.

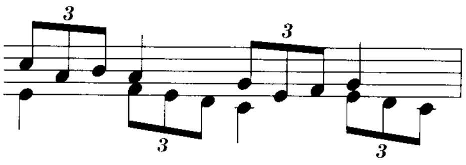

(See *Accidentals, Rests*)

FROM TWO PARTS TO UNISON

Opposite stem direction is used for clarity on **unison notes in short passages.**

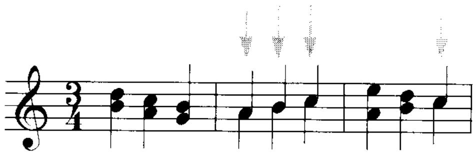

Indicate ***a2*** above staff and use one stem for longer passages.

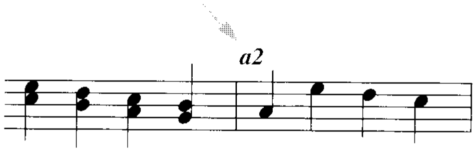

FROM TWO PARTS TO ONE:

Use rests in shorter passages.

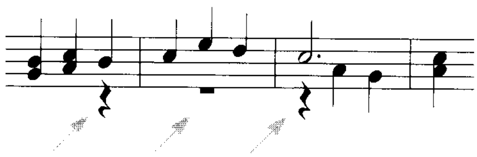

Place I. or II. above the staff and use one stem (for two *like* instruments in longer passages or after long rests).

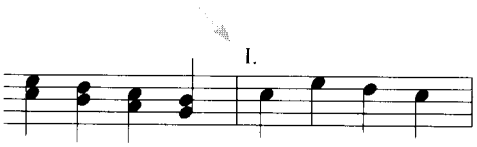

Mark passage with instrument name (for two *unlike* instruments in longer passages or after long rests).

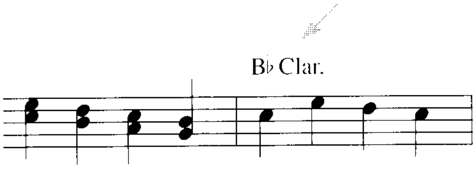

STAFF INDICATIONS

Staff indications (dynamic, slurs, tempi) are **placed below** the staff.

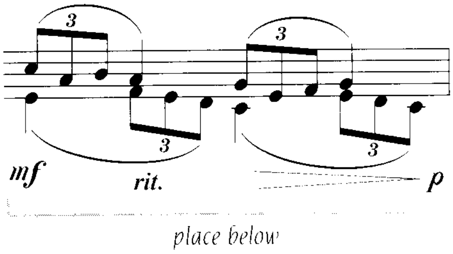

If the two parts move **independently,** a **separate set of directions** is needed for each part, above and below the staff.

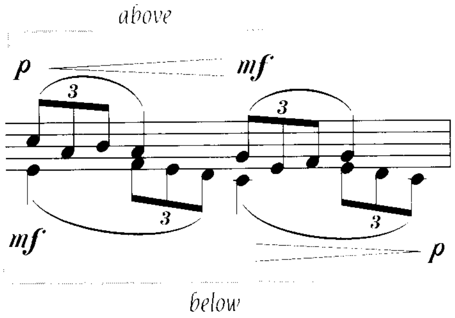

WHEN NOT TO SHARE A STAFF

1. If the range of one part tends to cross the range of the other, the parts should have **separate staves.**
2. If the parts require **different clefs** or **clef changes** the parts should be separated.
3. If the two parts are **too rhythmically diverse,** they should be separated.

In all cases **a bracket should be used,** not a brace.

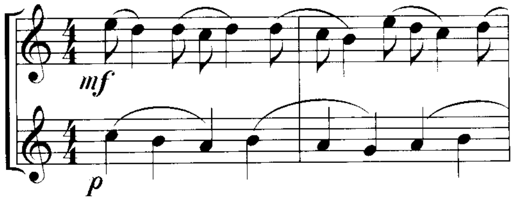

Sharp — See *Accidentals*

Simile

When the word ***simile*** is indicated after a **pattern has been established** (for example, of articulations, pedal marks, bowing, etc.), it means to **continue in the same manner.** (See *Articulations, Pedal marks*)

Slurs

Although a slur looks somewhat like a tie, the placement, positioning and use of slurs are different from ties. Slurs and ties should not be confused.

Slurs are used to indicate phrasing or technique. Depending on the instrument, the slur will imply different techniques to be used.

PLACEMENT OF SLURS

The **stemming** of the notes is the main factor in determining whether the slur is to be placed above or below.

When **stems extend up,** the **slur is placed under**, on the notehead side.

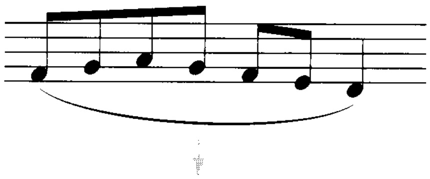

When **stems extend down,** the **slur is placed over,** on the notehead side.

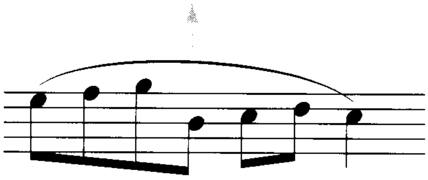

When **stems are in both directions,** the **slur is always placed over.**

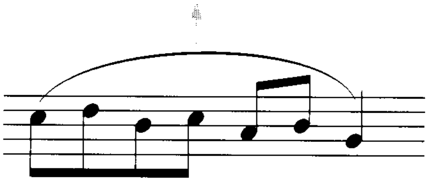

If the notes are whole notes, *imagine* the stems!

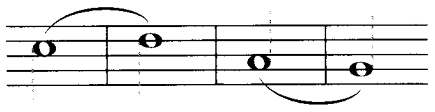

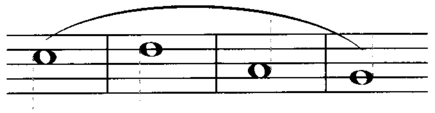

If the **phrase is substantially long,** the slur may be placed over, regardless of stem direction.

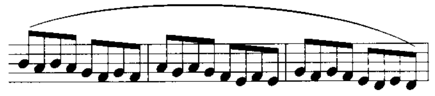

PLACEMENT ON THE NOTEHEAD

Slurs have a **definite beginning and ending;** they should not be "hanging in space" so that the beginning and ending notes must be guessed at.

If the slur is on the notehead side, the beginning and ending of the slur should be **centered on the notehead.** It should not touch the notehead.

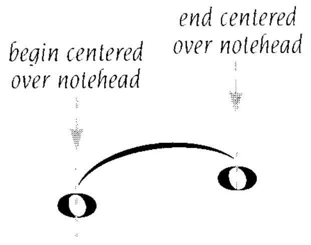
end centered
over notehead
begin centered
over notehead

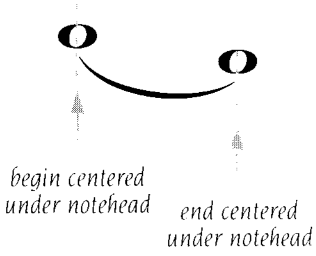
begin centered
under notehead
end centered
under notehead

BEGINNING FROM THE STEM

Place the beginning of the slur so it **starts mid-stem** (to the **right** of the stem—never begin to the left, crossing the stem).

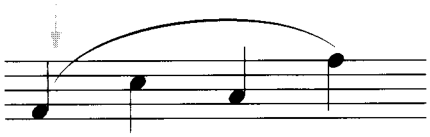

In some cases the slur may begin **at the end of the stem.**

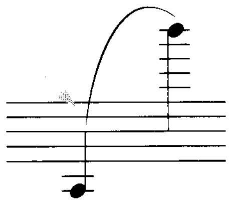

Begin at the end of the stem **if beaming is involved.** Never cross the beam with a slur.

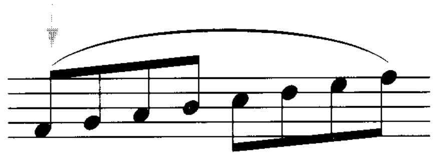

ENDING ON THE STEM

Place the end of the slur so it **slopes towards the notehead,** away from the end of the stem. **Center the slur** on the notehead.

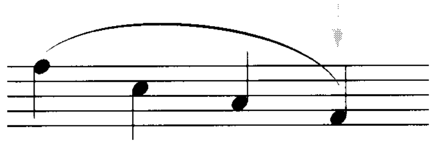

The slur may also end **at the end of the stem** if it would make a more pleasing shape.

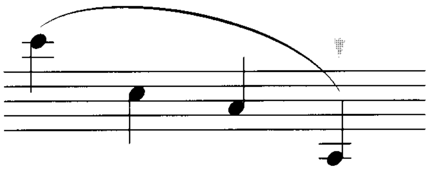

End at the end of the stem **if beaming is involved.** Never cross the beam with a slur.

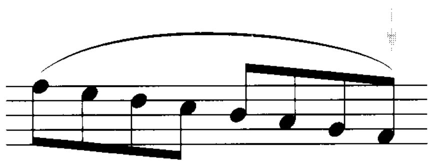

SLUR DIRECTION

If the direction of the **phrase ascends,** the slur will **slope upward.** If the direction of the **phrase descends,** the slur will **slope downward.**

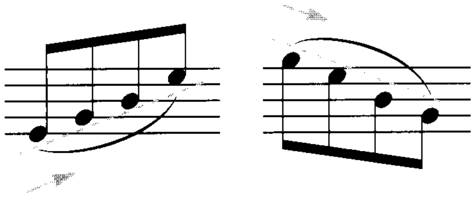

When a slur **begins and ends on the same note**, the slur can remain on the same horizontal level.

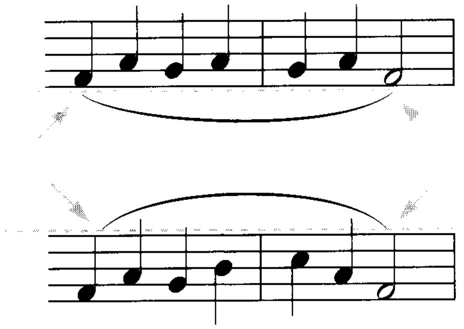

SLURS WITH CHORDS

Only one slur is needed when there is only one stem per chord.

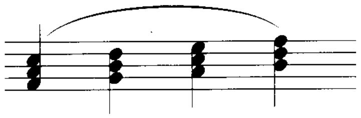

SHARING THE SAME STAFF

For two parts, place slurs **above for the upper part, below for the lower part.**

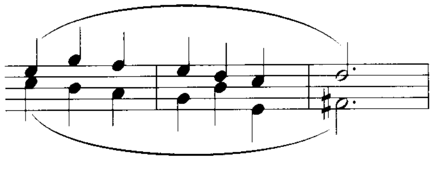

For two *sets* of parts, place slurs **above for the upper parts, below for the lower parts.**

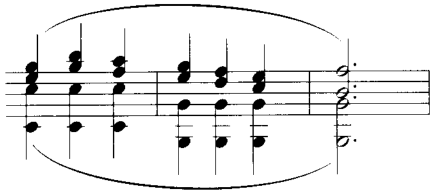

WRAPPING FROM STAFF TO STAFF

Align the end of the slur with the end of the staff. The slur ends at an **angle** (ascending or descending) at the end of the staff.

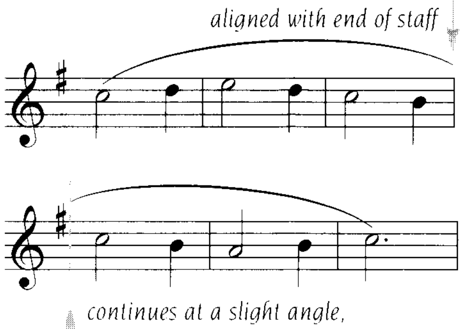

ends at a slight angle, aligned with end of staff

continues at a slight angle, after key signature or clef

Continue the slur in the lower staff immediately after the key signature (or clef), also at an **angle.** Be sure the first note is far enough to the right so that it is very clear that the slur does not begin on the note.

If the slur is over the staff it **remains over** when continuing on the next staff. If it is under, it **remains under** on the next staff.

The **direction of the slur** (ascending or descending) continues logically in the following staff.

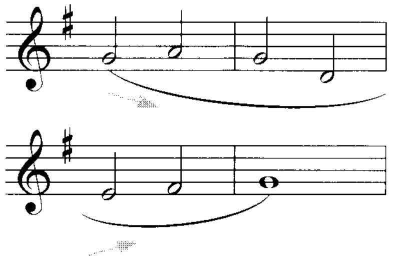

slur ascends to follow direction of notes

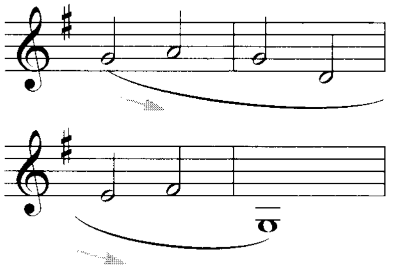

slur descends to follow direction of notes

DIFFERENTIATING SLURS FROM TIES

When slurs and ties are both involved in a chord progression such as the following, **position the slur opposite to the tie** and at a noticeable **angle.**

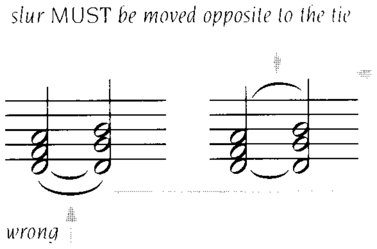
slur MUST be moved opposite to the tie
wrong

An even better way to write the above:

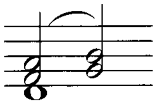

If a slur **begins and ends on the same pitch** and could possibly be mistaken for a tie, position the slur on the stem side.

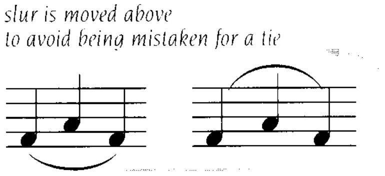
slur is moved above
to avoid being mistaken for a tie

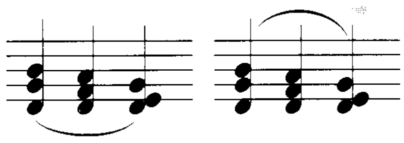

When a slur connects chords with **tied notes within**, combine the tied notes into a single note with opposite stem direction.

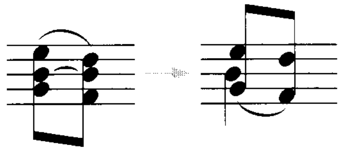

BEGINNING OR ENDING WITH A TIE

A slur **must include both notes of the tie,** whether the tie is at the beginning or at the ending of the slur. (For an exception, see "Elisions," on the next page.)

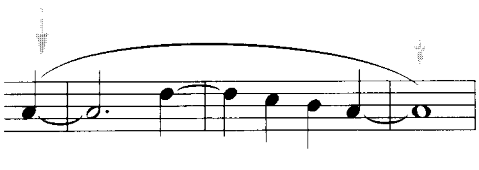

ELISIONS

In keyboard music, one slur might end at the same point where another slur begins.

The two slurs **share the center of the notehead** (elide) without touching each other.

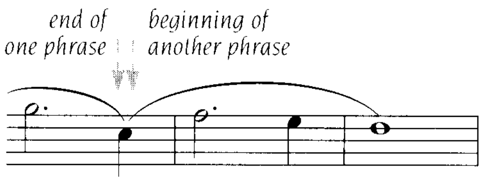
end of
one phrase
beginning of
another phrase

When the **elision occurs on tied notes**, the first phrase may end on the first tied note. The second phrase can then begin on the second tied note.

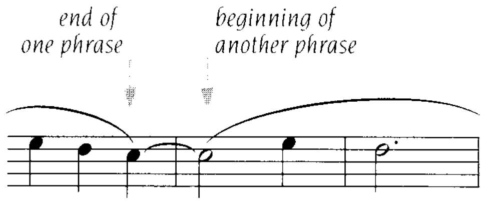
end of
one phrase
beginning of
another phrase

GRACE-NOTE SLURS

Grace notes **may or may not have slurs.** The slurs are usually somewhat smaller to match the grace notes.

Since grace notes are always stemmed up (unless there are two voices), the slur is usually **placed below.**

Grace-note slurs may be placed more to the inside edges of the noteheads.

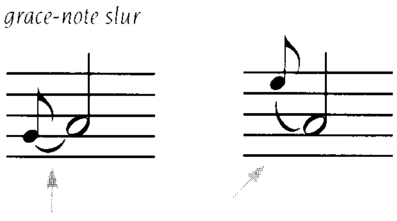
grace-note slur

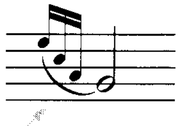

slur for beamed
grace notes

Grace-note slurs elide with slurs.

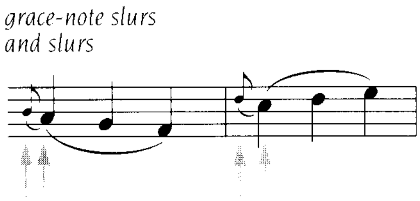
grace-note slurs
and slurs

SLURS AND ARTICULATIONS

Beginning or ending on notes with an articulation:

The beginning or ending of a slur is placed **outside staccato and tenuto** marks.

Center over or under the articulation.

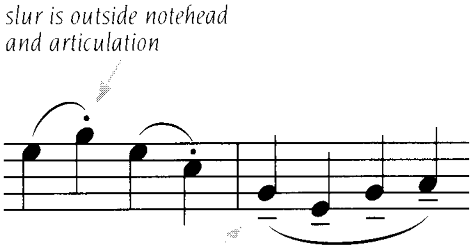
slur is outside notehead
and articulation

The beginning or ending of a slur is usually placed **between accents** and the notehead.

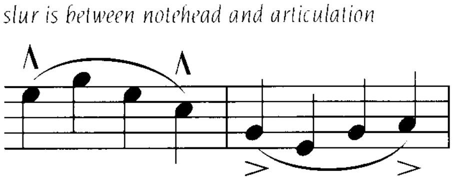
slur is between notehead and articulation

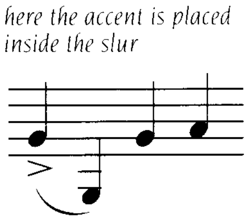
here the accent is placed
inside the slur

Articulations between the beginning and ending notes remain inside the slur.

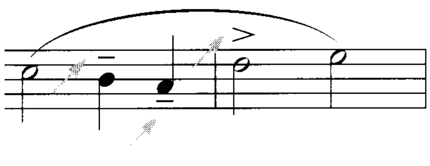

Fermatas are placed **outside** the end of a slur.

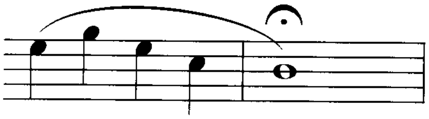

"LET RING" SLURS

The use of a short slur to indicate continued vibration should be placed on notes of longer duration only. Angle the slur slightly.

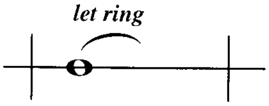

INSTRUMENT CONSIDERATIONS

Woodwinds and brass

Legato technique is indicated by the use of a slur. The slur might also indicate breathing and phrasing.

Slurs in combination with staccato or tenuto marks show varying degrees of legato.

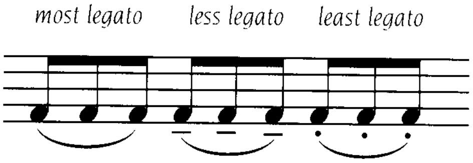
most legato
less legato
least legato

For keyboards

If the space between the two staves is limited, slurs may be placed outside the staff.

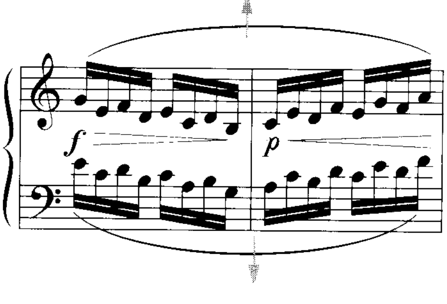

An S-shaped slur is sometimes used in music notated on the grand staff.

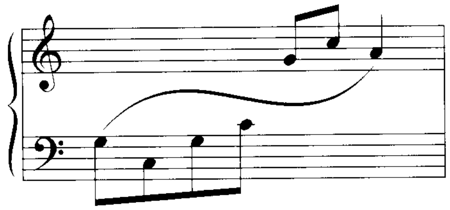

For strings

Slurs can only partially indicate phrasing. Slurs are also used to indicate various bowing techniques. (See *Bowing*)

For vocal music

A **slur is used for a *melisma*** (two or more notes that are to be sung on a single syllable).

The word or syllable should be placed flush left with the notehead.

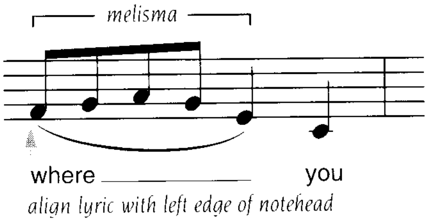

Slurs to indicate phrasing in vocal music are seldom used. Since breathing is a major factor in phrasing, rests and breath marks can be used as a partial indicator of phrases.

Slurs can be drawn using a **broken line** instead of a solid one. The broken slur is most often **used to accommodate lyrics for multiple verses.**

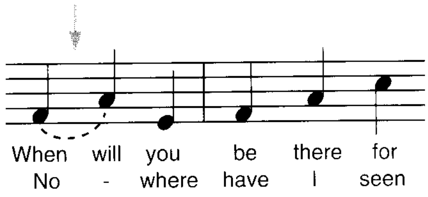

Similarly, **broken slurs** may also be used for syllable differences of translated text.

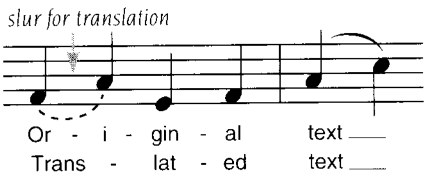

Spacing

Spacing is of primary importance to the **visual quality** of any musical notation. While the attention to *musical* detail may be excellent, poor judgment in spacing may make the score look ill-conceived.

An awareness of spacing occurs on many levels, from the **smallest element** (the beat) **to the largest** (the score or book). The desired effect is that the notation *and* layout will immediately convey the musical intent. Ultimately, any score should look **accessible** regardless of the level of difficulty.

PROPORTIONAL SPACING

The concept of proportional spacing takes into consideration the **mathematical spacing** of a beat, measure and system, in relation to the **density of activity** (the number of notes per beat, measure and system).

With proportional spacing, a half note, for example, must get *more* space than a quarter note, but not *twice* as much.

SPACING WITHIN THE MEASURE

Notes are spaced within the measure according to a **compromise between** two extremes:

1) **Space given according to note-value** (for example, a half note getting twice as much space as a quarter note).

2) **Equal spacing given to each note.**

Compare the normal (proportional) spacing with the two extreme examples.

Normal spacing—*proportional*

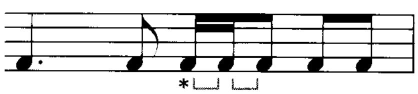

Spaced according to note-value—mathematical

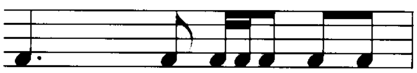

Wrong: *each note spaced equally*

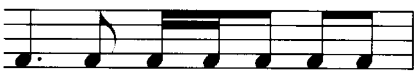

*Make sure that consecutive notes of equal value have equal space *following* each note.

MEASURES WITHIN THE STAFF

The size of the measure within the staff is determined according to a **compromise between** two extremes:

1) **Equal space** for each measure.

2) Measure size **based on note density only.**

Normal spacing of measures

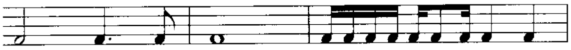

Wrong: *spaced according to equal measure size*

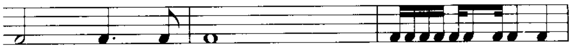

Wrong: *spaced according to measure density only*

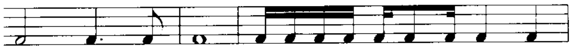

VERTICAL SPACING

Between staves

Musical elements (notes on leger lines, articulations, slurs, pedal marks) **projecting beyond a staff should never conflict** with the elements of another staff or system. There should be enough space allowed for ease of reading.

Inside systems

On the other hand, the **staves of a system should not be too far apart.** This is especially important for a system that is to be read by one performer—a keyboard player, for example. The eye should not have to jump too far from one staff to the other, in order to read notes for the right and left hand simultaneously.

> *The most common misjudgment is to place staves (or systems) exactly the same distance from one another without adjusting to compensate for the projecting musical elements.*

Between systems

More space should be be provided between systems than between the *staves* of the systems.

The **spacing between systems should be balanced,** with the page margins and the number of systems on the page in mind. If the **page is dense,** the systems will be closer to each other and to the minimum vertical margins, while allowing enough distance between systems to distinguish one from another. If the **page is sparse,** the upper and lower page margins could be larger.

GENERAL PAGE LAYOUT

Page numbers

The position of the page number in relation to the page edges should *never* change, regardless of page spacing. Printers can use the page number as a guide for positioning the image area on the page.

If page numbers are placed to the left and right, as is usual, **even numbers** are always to the **left, odd numbers** always to the **right.** Page numbers may be centered, but this is not the usual practice. Centered page numbers interfere more with the music and text.

Page numbers should be positioned **no closer than 1/4 inch from** the top (or bottom) of the **page edge.**

Title page

The title, credits, copyright and any accompanying text can be **vertically adjusted to accommodate** a loosely spaced page or a tightly spaced page. Such type should always be well placed, considering the space allowed.

Facing pages

Facing pages **should look good together.** As much as possible, they should have a **balanced look** (notes on a system, number of systems, white space, etc.). The vertical page margins should be the same if at all possible.

Page turns

It is very important to lay out the measures and systems in a manner that will allow a good page turn. Having **rests in the last measure** before the page turn is the most desirable. Second to rests would be perhaps a whole note (that could be held with the pedal) or a sparse measure—anything that would allow a hand to be free to turn the page.

Try to **minimize the number of page turns** by adjusting the amount of music on the page or by changing the sequence of pieces.

Book layout

Calculating the number of pages needed or what will work in the allowed number of pages is crucial to an appealing layout.

If a book contains many short pieces and a particular order is not important (or can be adjusted slightly), there are several things that can be done to make a good layout.

Often a piece can be **condensed or expanded** to offer choice page turns and good facing pages.

Consider reordering the pieces to **minimize the number of page turns. Group pieces** with an odd number of pages together so that the ones with an even number of pages can have facing pages.

A poorly placed piece at the beginning of a book can throw off the facing pages for the rest of the book.

Backing up

Expect to go back and adjust **various elements** (page turns, systems, staves, measures or notes within a measure) when planning a layout for a book. A change in the order of pieces and the number of pages might create density problems on other pages.

The part must be considered with the whole and the whole considered with the part.

White space

With experience comes the **appreciation of white space** on a page or within a book. There is nothing more daunting than overwhelming density. The eye should be guided from note to note, measure to measure, staff to staff and page to page with a minimum of surprises. If a performer must stop playing in order to understand the notation, or loses his place because of poor placement, the notation and layout are *not* successful.

Staccato — See *Articulations*

Staff

The staff commonly consists of five lines.

The **staff line weight** should be thick enough to be clearly legible but thin enough for the notes of the staff to be easily read.

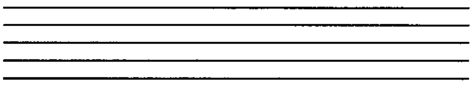

A staff line will be **thicker than stem lines** and **equal to** (or thinner than) **barlines.**

If a note is on the line, the line will always run through the **center of the notehead,** regardless of the note value.

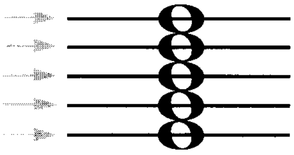

A notehead in a space is always clearly positioned between the two staff lines.

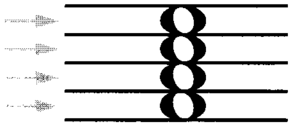

A **one-line staff** is used for non-pitched percussion.

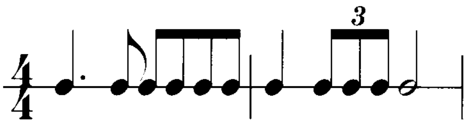

or

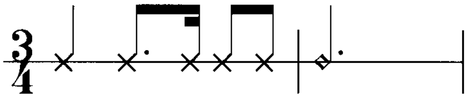

Leger lines extend the range of the staff.

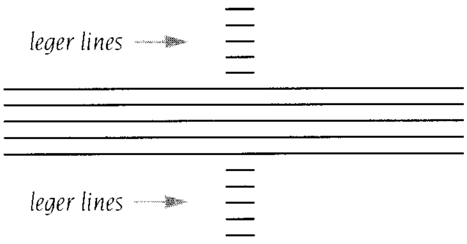
leger lines
leger lines

Barlines divide the staff into measures.

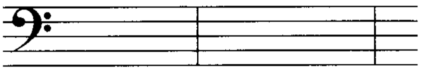

A **clef** placed on a five-line staff determines the pitches for each line and space.

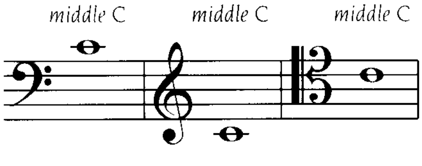
middle C
middle C
middle C

A **key signature** determines the key of the music notated on the staff.

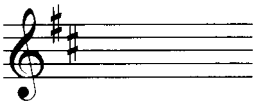

Two or more staves grouped together (with a systemic barline) form a **system.**

A **bracket** groups staves into systems or parts of systems. (See *Bracket*)

A **brace** groups staves into a **grand staff.** (See *Brace*)

Some examples of systems:

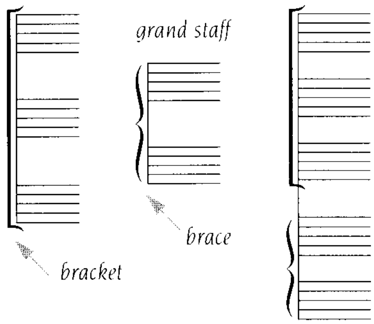
grand staff
brace
bracket

The **first system** (or staff) of music is usually **indented** about 1/2 inch.

Stems

Stems are **thinner than staff lines and barlines.**

STEM LENGTH

The **normal stem length** is 3 1/2 spaces (one octave).

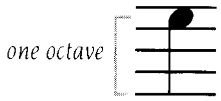

With leger lines

When a note extends **beyond one leger line,** the stem **must touch the middle staff line.**

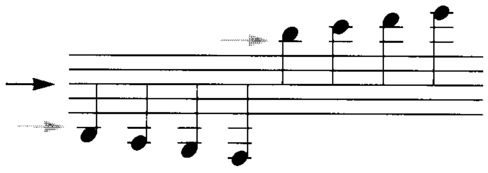

When parts share a staff

Up-stem notes **above the middle line are shorter** than normal. Traditionally, their stem length is relative to surrounding notes, getting progressively shorter as the notes go higher. The shortest stem length is 2 1/2 spaces (interval of a 6th).

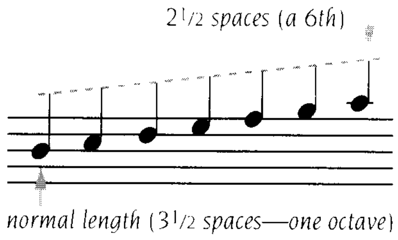
2 1/2 spaces (a 6th)
normal length (3 1/2 spaces—one octave)

For the computer, a setting should be chosen that works well for most situations.

Down-stem notes **below the middle line** follow the same guidelines as the above.

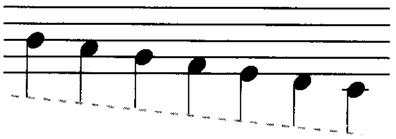

With the interval of a 2nd

The stem is always placed **between the two notes** of an interval of a 2nd, with the **upper note always to the right,** the **lower note always to the left.**

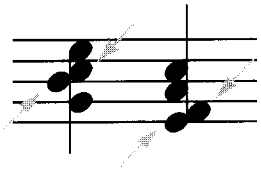

STEM DIRECTION

For notes on the **middle line and above,** the stem is down.

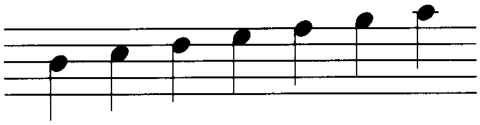

For notes **below the middle line,** the stem is up.

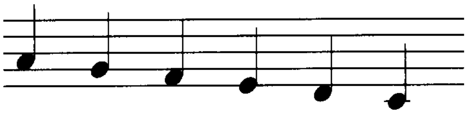

When sharing a staff

Notes have **opposite stem direction** when sharing a staff. (Stems in the "wrong direction" will have shorter stems, as previously mentioned.)

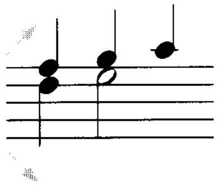

When notes share a stem

If the **note above is farther** from the middle line than the note below, the **stem goes down.**

If the note **below is farther** from the middle line, the **stem goes up.**

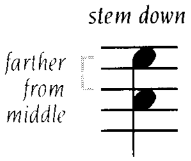

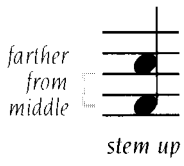

When the two notes are an **equal distance from the middle line,** the preferred direction is down.

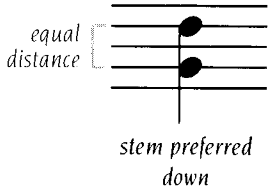

If a **pattern of stem direction** is established, stem direction is *sometimes* maintained when the notes are an equal distance from the middle. (This practice is all but obsolete.)

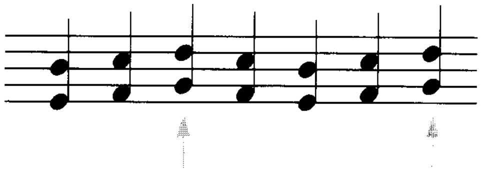

The same rules apply to *more than two notes* sharing a stem. The **distance of the OUTER NOTES from the middle line** determines stem direction.

When the **outer notes** are an **equal distance from the middle line** and the majority of notes are above the middle line, the stem goes down.

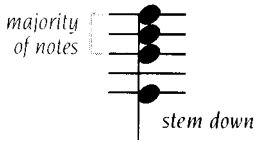

When the **outer notes** are an **equal distance from the middle line** and the majority of notes are below the middle line, the stem goes up.

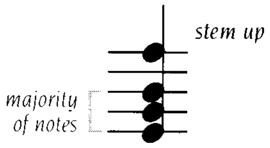

For beamed groups

Simple rules for stemming of single notes and chords apply to beamed groups when possible (for example, if all notes in a beamed grouping are on or above the middle line of the staff, stems are down).
(See *Beams*)

Systems

Two or more staves grouped together (with a systemic barline) form a system.

A **bracket** groups staves into systems or parts of systems. (See *Bracket*)

A **brace** groups staves into a **grand staff.** (See *Brace* and *Grand staff*)

Some examples of systems:

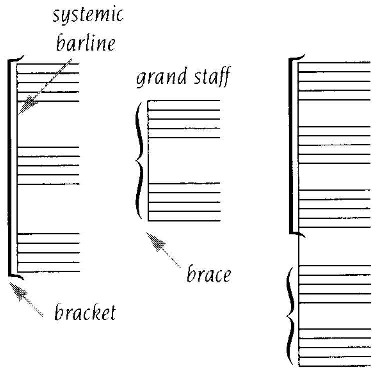
systemic barline
grand staff
brace
bracket

(See *Barlines*)

The **first system** of music is usually **indented** about 1/2 inch.

Tablature

Tablature is a system of notation that graphically represents the strings and frets of a stringed instrument (such as guitar).

The letters **"TAB"** are placed where a clef would normally be. The only function of these letters is to identify the staff as being tablature.

Each note is indicated by **placing the fret number on the appropriate string.**

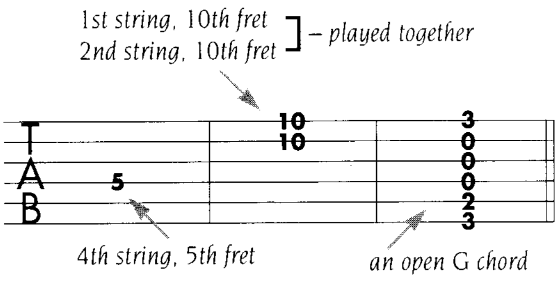
1st string, 10th fret
2nd string, 10th fret
– played together
T
A
B
5
10
10
3
0
0
0
2
3
4th string, 5th fret
an open G chord

Tablature does NOT indicate rhythmic values; therefore, a five-line staff using **standard notation** is placed above the tablature staff.

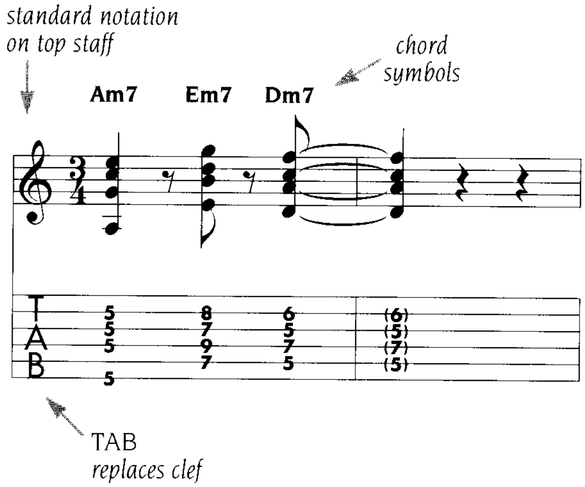
standard notation on top staff
chord symbols
Am7
Em7
Dm7
T
A
B
5
5
5
5
8
7
9
7
6
5
7
5
(6)
(5)
(7)
(5)
TAB replaces clef

A **systemic barline connects** the two staves together. Measure **barlines** are broken.

Chord symbols (when appropriate) are often added above the top staff.

Picking or fingering techniques notated on the top staff are repeated on the tablature staff.

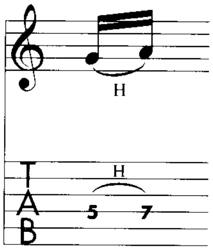

Every note on the top staff must have a **corresponding finger number** on the tablature staff—**tied notes** are in parentheses.

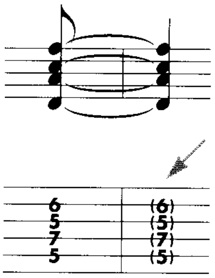

Tablature may consist of **four lines** (for electric bass and mandolin) or **five lines** (five-string banjo).

The **notation** for guitar tablature outlined on the previous pages applies to all tablature, regardless of the number of strings.

Electric bass

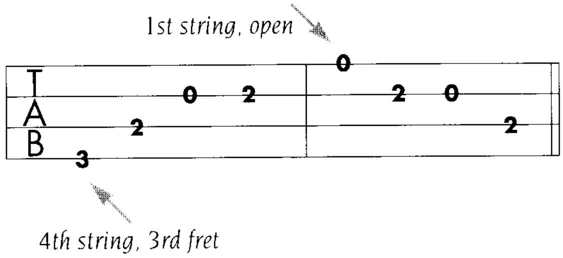

Five-string banjo

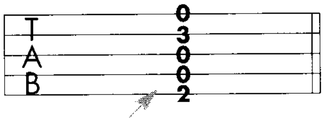

a G7 chord

Tempo marks

Tempo marks indicating the rate of speed may be followed by an expression mark and/or metronome mark.

The left edge of the tempo mark is **vertically aligned to the left edge of the time signature.**

The type is bold Roman.

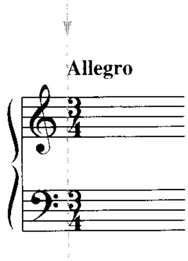

WITH EXPRESSION MARKS

Tempo marks are often accompanied by an indication of **expression;** however, expression marks should **never** ***replace*** tempo marks.

Presto con brio

Slowly, with much emotion

WITH METRONOME MARKS

The tempo mark may be more precisely indicated with the addition of a metronome mark.

The metronome mark may be ***specific,*** a tempo ***range*** or ***general.*** The type is slightly smaller than that of the tempo and is usually enclosed in parentheses. The note is cue size or smaller.

Allegro con moto **(♩ = 144)**

Allegro con moto **(♩ = 132–144)**

Allegro con moto **(♩ = ca. 144)**

TEMPO VARIATIONS

If the tempo is varied by the use of markings such as ***ritardando, rallentando*** or ***accelerando,*** the term ***a tempo*** is placed at the point where the tempo is reinstated. *The type will be bold italic.*

If the tempo changes, and then returns to the original tempo later in the piece, the term **Tempo I** is indicated **above the staff.** *The type will be bold Roman.*

Tenuto

See *Articulations*

Ties

A tie connects **two consecutive notes of the same pitch,** extending the duration of the first note to include the second. (The second note does not have its own attack.)

The shape of a tie is somewhat similar to a slur, but the **placement, positioning and use of ties and slurs are different;** the two should not be confused.

BEGINNING AND ENDING

Begin and end on the same horizontal level.

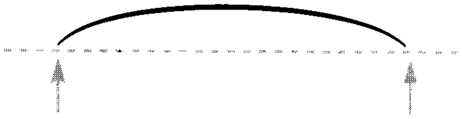

Begin to the right of the notehead.
End to the left of the next notehead.
The tie **does not touch the noteheads.**

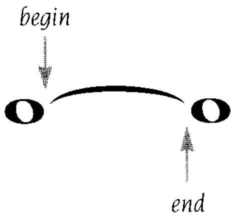

AVOID AUGMENTATION DOTS

Ties should never collide with augmentation dots. **Begin the tie to the right of an augmentation dot.**

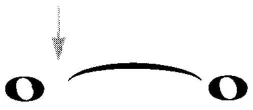

PLACEMENT ON THE STAFF

Avoid touching staff lines.
Place the **center** of the curve in a "space."

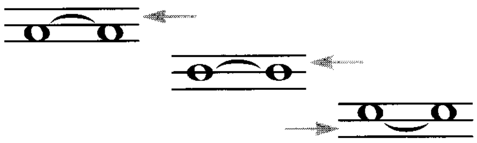

Ends of ties may cross staff lines—it is better to cross even slightly, than to just touch a staff line.

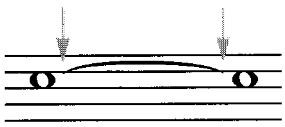

Adjust curve for longer ties.
Raise or lower the center of the curve to the next space.

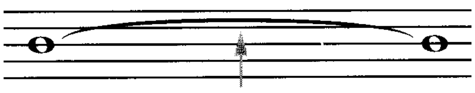

DIRECTION ON THE STAFF

Curve ties opposite of stem direction.
Stem down—curve above notehead.
Stem up—curve below notehead.

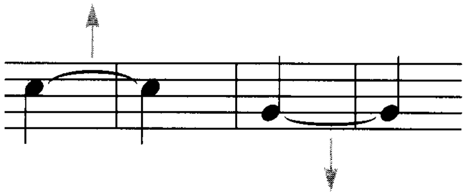

Ties on whole notes are treated as if the whole notes had stems.

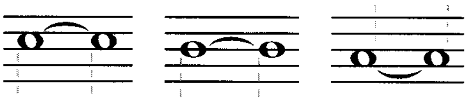

For mixed stem direction, the tie is always placed **above.**

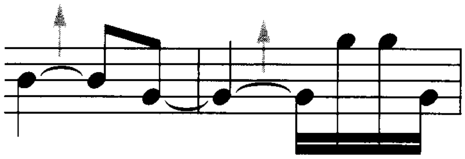

SHARING THE SAME STAFF

For two parts, place ties **above for the upper part, below for the lower part.**

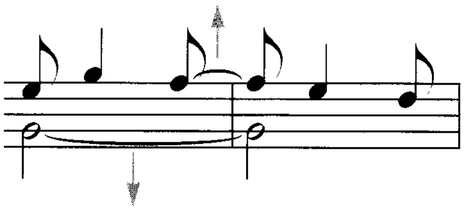

For two *sets* of parts, place ties **above for the upper two parts, below for the lower two parts.**

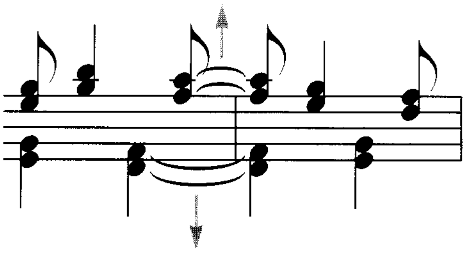

MULTIPLE TIED NOTES

When more than two notes are tied, ties must be added from note to note.

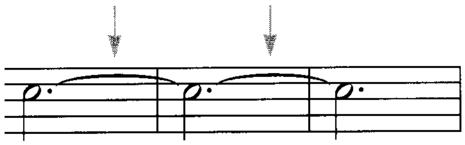

WRAPPING FROM STAFF TO STAFF

Align end of tie with end of staff.

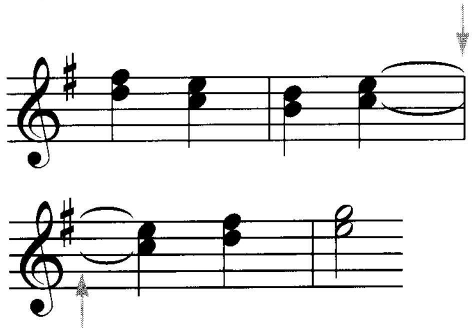

Continue tie in the lower staff immediately after the key signature (or clef).

Beginning and ending points always remain on the **same horizontal level.**

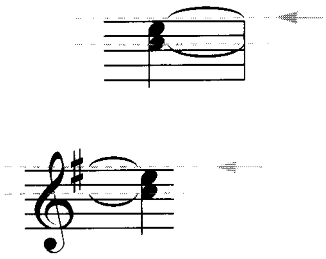

DIRECTION OF TIES FOR CHORDS

Highest note—tie is **above.**

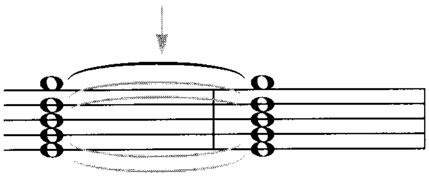

Lowest note—tie is **below.**

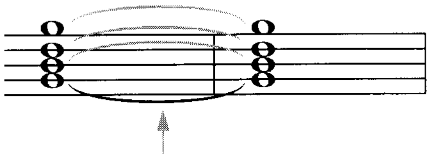

Notes between highest and lowest—on the **middle line and above,** ties are **above.**

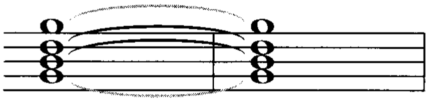

Below the **middle line,** ties are **below.**

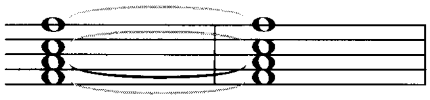

For **intervals of a 2nd**, ties will be in **opposite directions** whenever possible.

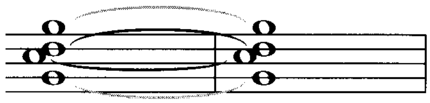

If some notes are not tied within the chords, the uppermost tied note has the tie above; the lowermost tied note has the tie below.

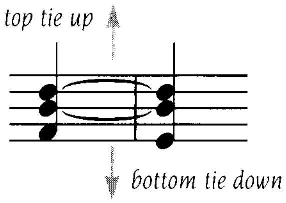

ALIGNMENT OF TIE ENDS

Align ends of ties within a chord.

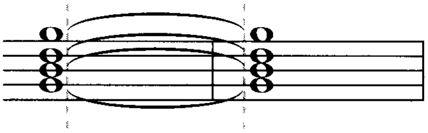

Tie ends for intervals of a 2nd should be adjusted if possible.

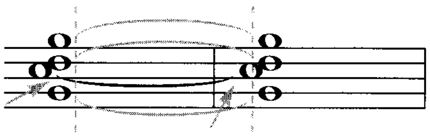

ADJUSTMENT OF TIES

Whenever possible, **maintain the curve** of a tie in uncommon situations. It should look like a tie, not a slur.

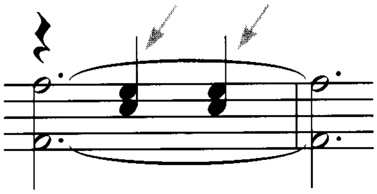

In this instance it is better to have the top tie curved up even though it is a part of the lower voice.

If a time signature interrupts a tie, break the tie to avoid colliding with the time signature.

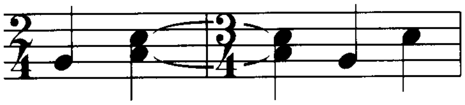

BROKEN (DASHED) TIES

Ties using a broken line instead of a solid one are often **used to accommodate lyrics for multiple verses.**

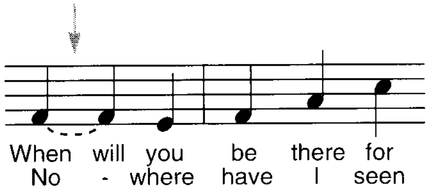

TIES IN MEASURED ARPEGGIOS

Although traditionally considered incorrect, ties are sometimes **extended to the unbroken chord** to avoid a cluttered look.

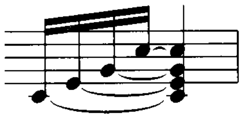

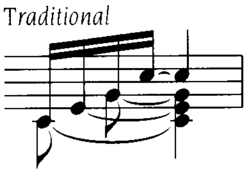

Time signatures

Time signatures indicate the meter for a measure(s) or an entire piece.

Since meter is so closely connected with the proper use of ties, rests, beams and syncopated rhythms, the **correct time signature** must be chosen.

The **upper number** indicates the number of beats per measure.

The **lower number** indicates the note that gets one beat (2 = half note, 4 = quarter note, 8 = eighth note, 16 = sixteenth note, etc.).

All measures after a time signature will be in the same meter until a new time signature is indicated.

PLACEMENT

The **upper number** is always placed between the top line and the middle line of the staff.

The **lower number** is always placed between the middle line and the bottom line of the staff.

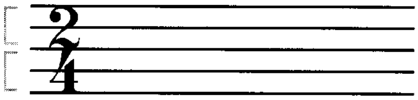

For a **single-line staff,** the line intersects the two numbers.

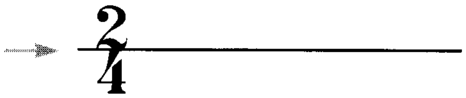

The time signature is **indicated after the clef and key signature.**

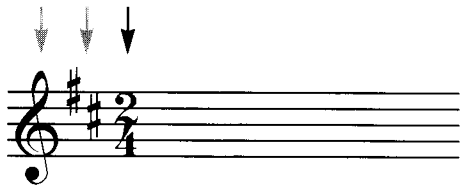

CHANGING TIME SIGNATURES

Time signatures may change from measure to measure, applying to all measures until the next time signature change. **A change in time signature always appears AFTER a barline.**

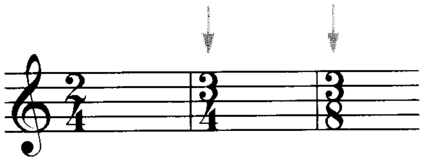

If the time signature changes at the beginning of a line, a **courtesy time signature** is added at the end of the previous staff, *after* the last barline. The new time signature appears after the clef (and key signature, if any) on the next staff.

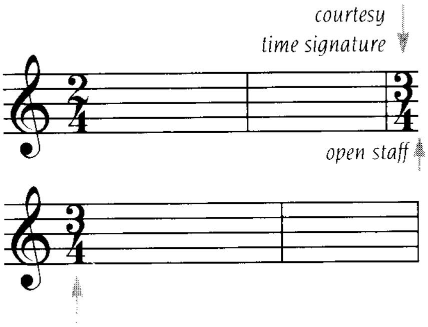

The following illustrates the positioning of the clef, key signature and time signature if all three change simultaneously.

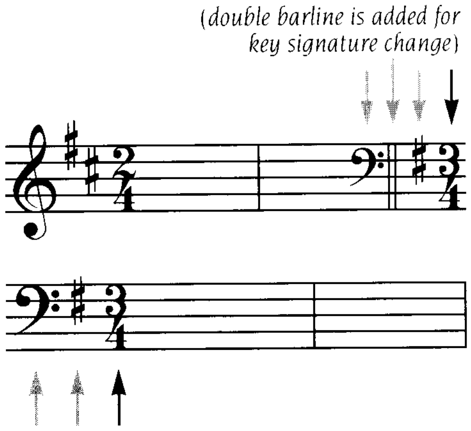

COMPOSITE METER (See *Meter*)

ALTERNATING METER (See *Meter*)

Tremolo

A tremolo may be **on one note,** or **between two or more notes.** It may be **measured** (a clear subdivision) or **unmeasured** (as fast as possible).

ON ONE NOTE

Measured tremolo

A **short slanted line,** ascending from left to right, through the stem of a note indicates that a note is to be subdivided.

Each line placed on the stem indicates the **rhythmic value** of the tremolo.

For instance, one line will indicate eighth notes,
two lines—sixteenth notes.

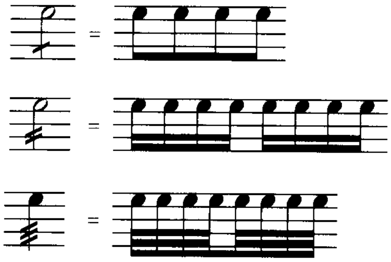

In string music, a measured tremolo is sometimes indicated by placing **divisi dots** above the note, in addition to the slashes through the stem.

The divisi dots precisely indicate the divisions per note. The slashes through the stem indicate the rhythmic value.

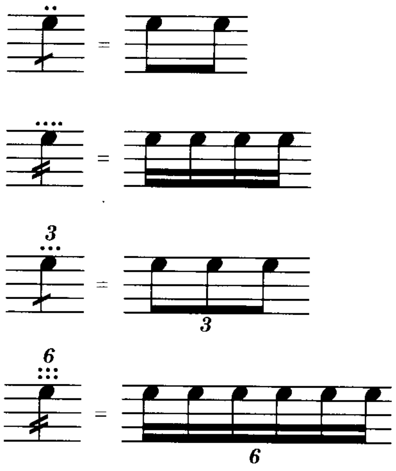

Unmeasured tremolo

An **unmeasured tremolo** (on one note) is indicated by placing three lines through the stem and adding the abbreviation ***trem.*** above the note. (If placed on a beamed note, the beam counts as one of the lines.) Most often seen in notation for strings, percussion and fretted instruments.

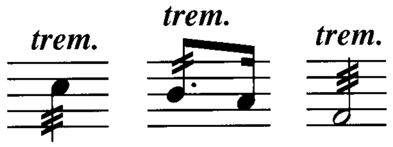

Tremolo lines on whole notes are traditionally **indicated above or below the note** (positioned as if there were a stem, but centered on the notehead).

traditional style

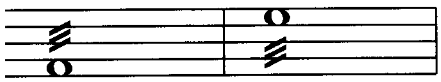

Modern style breaks the whole note into two half notes and places the slanted lines through the stems as usual.

modern style

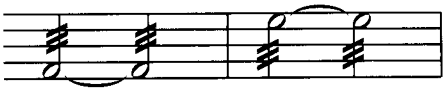

ON MORE THAN ONE NOTE

When a tremolo is indicated between two (or more) different pitches, the notes are connected with a beam, with incomplete beams placed between to indicate the rhythmic value.

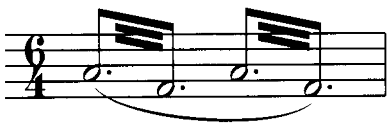

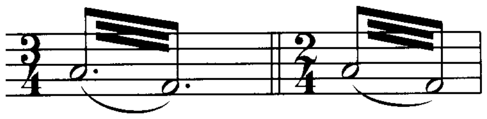

Stems are lengthened to accommodate added beams.

With such a subdivision of the tremolo, **each note equals the total rhythmic value of the tremolo.**

unmeasured whole-note tremolo (traditional style)

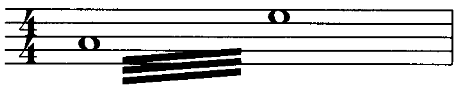

modern style

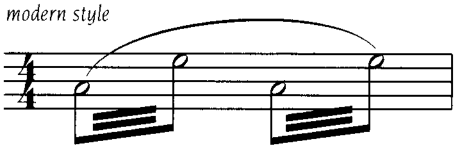

In the previous example, a complete measure of tremolo in $\frac{4}{4}$ time consists of two sets of half notes beamed together.

INTERVAL / CHORD TREMOLO

For some instruments, a tremolo is possible between two, three or more notes.

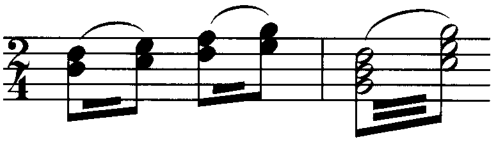

Trill

The abbreviation ***tr*** indicates a trill.

The ***tr*** is always **placed above the note,** regardless of stem direction (unless two parts share a staff).

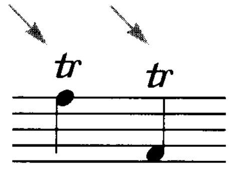

Each trilled note must have its own trill symbol.

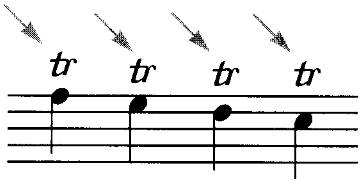

A **wavy line** may be indicated after the ***tr***, especially if the trill is extended.

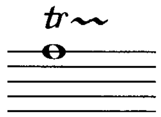

If the trill **extends over two or more notes,** the wavy line continues to the end of the notehead of the last note affected by the trill.

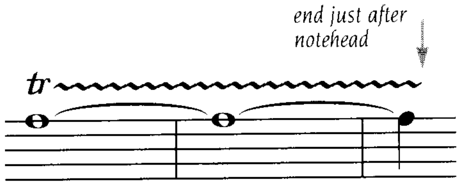

If the **wavy line extends over consecutive notes of the same pitch** but without ties, the notes are to be clearly articulated.

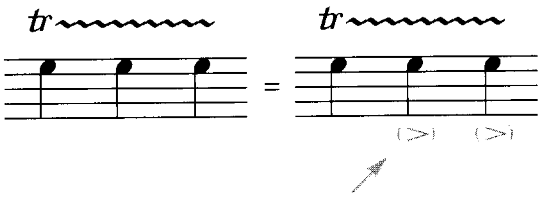

When two parts share a staff, the trill for the lower part is placed below the note.

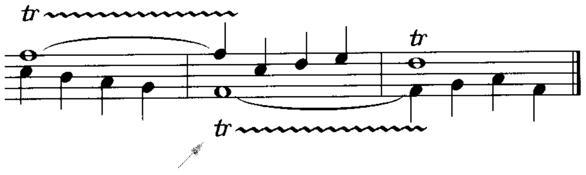

When the upper note of a trill is **altered from the diatonic scale,** a small **accidental** is used with the trill sign. The accidental is placed after the trill sign but before the wavy line (if used.)

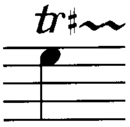

The accidental **may also appear above** the trill sign.

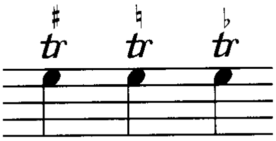

A **cue-size notehead** is sometimes used.

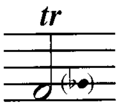

Terminations can be indicated by cue-size or full-size notes.

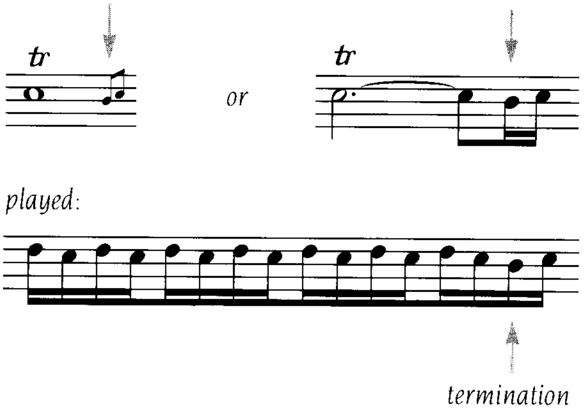

Triplet

See *Tuplets*

Tuplets

Rhythmic groupings of notes that are NOT metric groupings are known as **tuplets.**

Duplets, triplets, quadruplets, quintuplets, sextuplets and septuplets are all examples of **tuplets.**

TUPLET NUMERALS

The numeral is larger than a finger number and is **distinctly italic, preferably bold italic.** (See *Fingerings [for keyboard]*)

For beamed tuplets, the numeral is preferred **on the beam side, centered.**

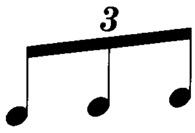

Do not use a tuplet slur with the numeral. This is an obsolete practice and causes confusion between normal slurs and tuplet slurs.

Place the tuplet numeral **outside the staff** whenever possible.

If placed within the staff, the **numeral must be clearly positioned,** avoiding staff lines as much as possible.

The numeral **remains placed with the beam,** whatever the stem direction.

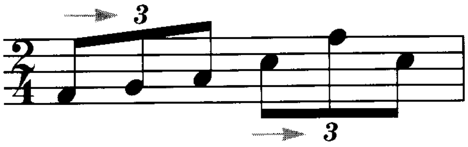

When parts share a staff, the numeral is placed normally (beam/stem side).

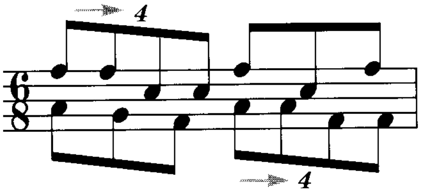

After a **pattern of tuplets** has been established, the numeral may be omitted. ***Simile*** may also be used for clarity.

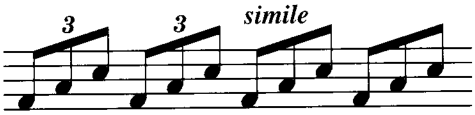

TUPLET BRACKETS

If the tuplet is not beamed, a **bracket is added** and the **numeral is centered** within the bracket.

The bracket always **begins flush left** with the notehead and **ends flush right** with the notehead, whether on the notehead or stem side.

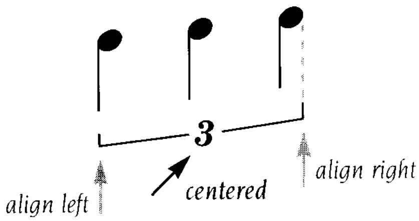

The **bracket is broken** to accommodate the number; the **ends** are always vertical.

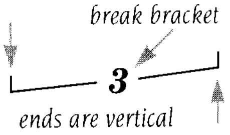

If the tuplet numeral is **moved to the notehead side** of a beamed tuplet (for example, to avoid fingerings), add a bracket for clarity.

The **bracket is placed** at the same angle as the beam.

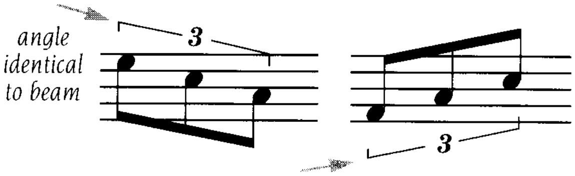

When placing a **bracket with unbeamed notes,** the bracket angle can vary.

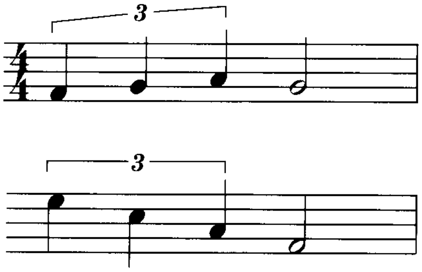

BRACKETS AND UNEQUAL VALUES

For **unequal rhythmic values** within the tuplet group, the bracket must **include all notes** of the group.

The **numeral remains centered** within the bracket.

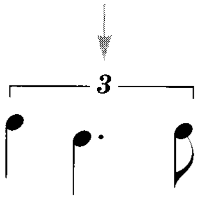

The tuplet unit is clearly seen if the bracket extends to **include the entire allotted space** for the tuplet (as in the first tuplet below).

The bracket does not extend to include the flag (as shown in the second tuplet).

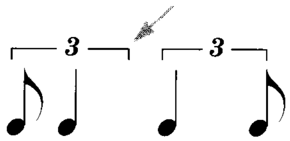

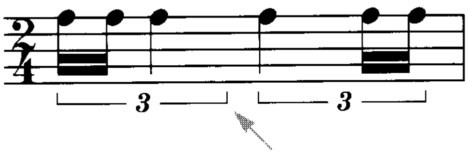

WITH RESTS

The **bracket extends** to include rests.

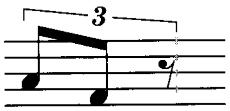

Although the bracket is preferred, an option is to **extend the beam** to include the rest within the tuplet.

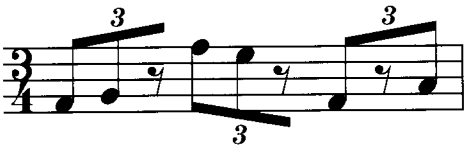

WITH FINGERINGS

It is advised to **move the tuplet figure to avoid confusion with finger numbers.**

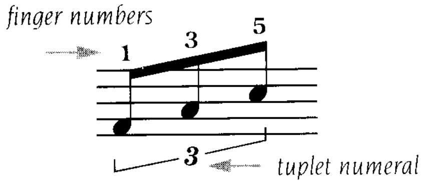
finger numbers
1
3
5
3
tuplet numeral

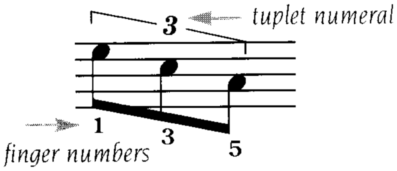
3
tuplet numeral
1
3
5
finger numbers

WITH ARTICULATIONS

If the bracket must be placed on the same side as articulations, the bracket is **placed outside articulations.**

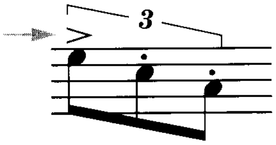

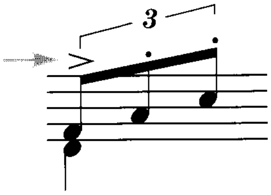

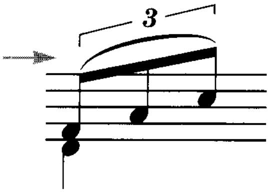

Short slurs may be considered articulations.

DETERMINING NOTE VALUE

A tuplet may occupy a beat, a pulse, a *portion* of a beat or pulse, or an entire measure.

To determine the rhythmic value of a tuplet that occupies **one beat,** subdivide the beat metrically as many times as you can without exceeding the number of notes in the tuplet. The note value of the tuplet will be equal to the note value of that metric subdivision.

In the example below, if each beat were subdivided further, it would have more notes than the tuplet; therefore the tuplet notes are sixteenth notes.

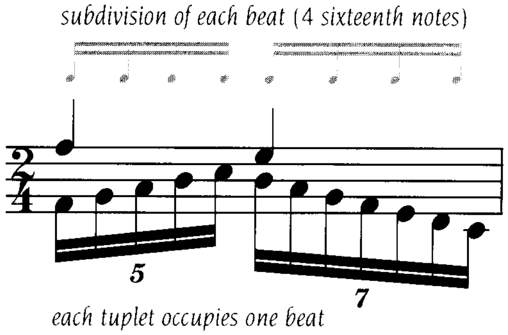
subdivision of each beat (4 sixteenth notes)
5
7
each tuplet occupies one beat

Use the same principle to determine the note value of tuplets that occupy an **entire measure** (or another unit of a measure).

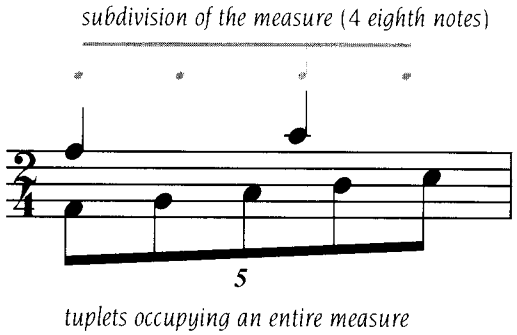
subdivision of the measure (4 eighth notes)
5
tuplets occupying an entire measure

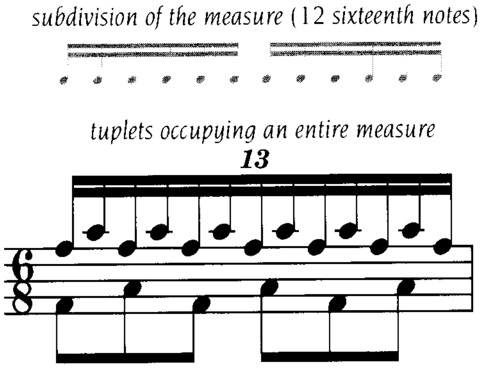
subdivision of the measure (12 sixteenth notes)
tuplets occupying an entire measure
13

Duplets are the only exception to the previous rule deciding tuplet rhythmic value.

Duplets are **equal in rhythmic value to the beat** (not to be confused with pulse).

Duplets only occur in **compound meter.**

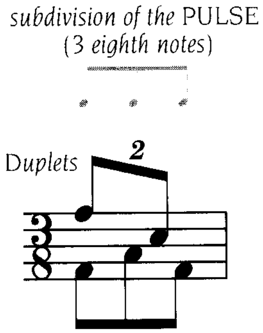
subdivision of the PULSE
(3 eighth notes)
Duplets
2

HORIZONTAL SPACING

A tuplet is spaced mathematically, as proper spacing dictates—not aligned with the metric rhythm.

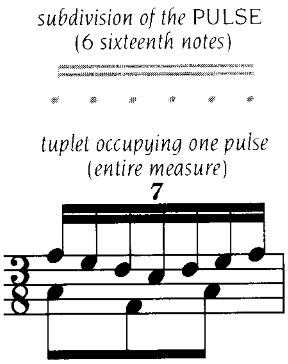
subdivision of the PULSE
(6 sixteenth notes)
tuplet occupying one pulse
(entire measure)
7

Words

See *Lyric*

X notehead

See *Notes*

INDEX OF TOPICS

A cappella 3
A due (a2) 3
A tempo 3
Accents 3
Accidentals 3
Alternating instruments 10
Arpeggio signs 11
Articulations 13
Augmentation dot 21
Barlines 25
Beams 29
Bowing 42
Brace 43
Bracket 43
Breath mark 44
Caesura 45
Chord frames & symbols 46
Circle of fifths 49
Clef signs 49
Coda sign 53
Courtesy accidentals 53
Credits 53
Cue notes 54
D.C. al Coda 55
D.C. al Fine 57
D.S. al Coda 58
D.S. al Fine 59
Divisi 60
Divisi dots 60
Dotted rhythms 60
Dynamics 60
Endings 65
Fermata 65
Fingerings (keyboard) 67
Flat 71
Glissando sign 71
Grace notes 72
Grand pause 74
Grand staff 74
Harmonics 76
Key signatures 78
Layout 82
Leger lines 83
Lyric 84
Meter 88
Metronome marks 95
Natural sign 95
Notes 95
Octave signs 97
Ornaments 103
Ossia 104
Pedal marks 105
Picking 109
Repeat signs 110
Rests 113
Scores 117
Sharing a staff 118
Sharp 121
Simile 121
Slurs 121
Spacing 131
Staccato 134
Staff 135
Stems 137
Systems 140
Tablature 140
Tempo marks 142
Tenuto 143
Ties 143
Time signatures 149
Tremolo 151
Trill 153
Triplet 155
Tuplets 155
Words 160
X notehead 160

SMITHSONIAN

NATURE GUIDE

CHALCANTHITE

ARAGONITE

ORPIMENT

GRANITE

OLIVINE

ROCKS AND MINERALS

THE WORLD IN YOUR HANDS

NATUREGUIDE

ROCKS
AND MINERALS

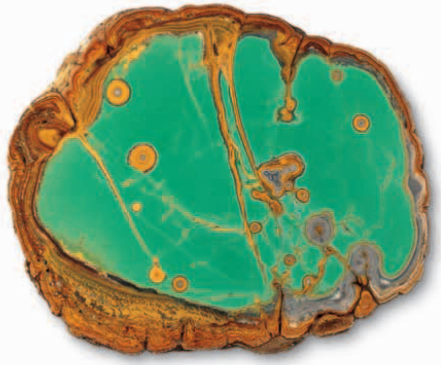

DK SMITHSONIAN

NATURE GUIDE

ROCKS AND MINERALS

Ronald Louis Bonewitz

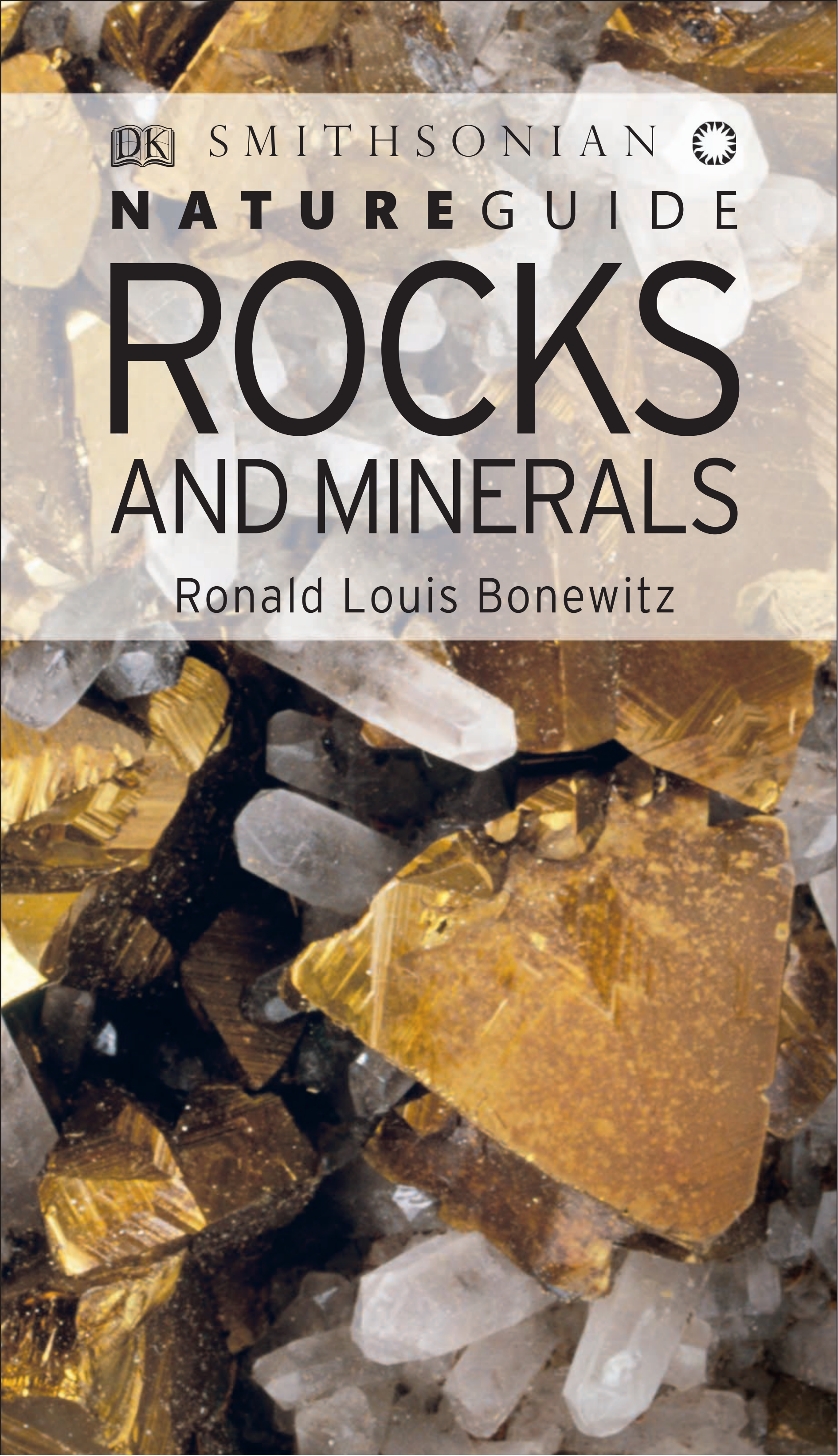

**LONDON, NEW YORK, MELBOURNE,
MUNICH, AND DELHI**

DORLING KINDERSLEY

Senior Editor
Peter Frances

US Editors
Jill Hamilton
Rebecca Warren

Jacket Editor
Manisha Majithia

Jacket Designer
Laura Brim

Managing Editor
Camilla Hallinan

Associate Publishing Director
Liz Wheeler

Publishing Director
Jonathan Metcalf

Senior Art Editor
Spencer Holbrook

Production Editor
Rebekah Parsons-King

Production Controller
Erika Pepe

DK Picture Library
Rose Horridge

Picture Researchers
Jo Walton, Julia Harris-Voss

Managing Art Editor
Michelle Baxter

Publisher
Sarah Larter

Art Director
Philip Ormerod

DK INDIA

Managing Editor
Rohan Sinha

Deputy Managing Editor
Alka Thakur Hazarika

Senior Editor
Soma B. Chowdhury

Editors
Pragati Nagpal, Neha Pande, Priyaneet Singh

DTP Designers
Sourabh Challariya, Arvind Kumar,
Arjinder Singh, Jagtar Singh, Rajesh Singh,
Bimlesh Tiwary, Tanveer Zaidi

Production Manager
Pankaj Sharma

Deputy Managing Art Editor
Mitun Banerjee

Senior Designer
Ivy Roy

Designers
Arijit Ganguly, Mahua Mandal, Tanveer Zaidi

Assistant Designer
Sanjay Chauhan

Consultant Art Director
Shefali Upadhyay

Picture Researcher
Sakshi Saluja

DTP Manager
Balwant Singh

CONSULTANT

Dr. Jeffrey E. Post, Geologist, Curator-in-Charge, National Gem and Mineral Collection,
National Museum of Natural History, Smithsonian Institution

First American Edition, 2012
Published in the United States by DK Publishing
375 Hudson Street, New York, New York, 10014
12 13 14 10 9 8 7 6 5 4 3 2 1
001 – 181829 – Jul/2012

Copyright © 2012 Dorling Kindersley Limited
All rights reserved

Without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted, in any form, or by any means (electronic, mechanical, photocopying, recording, or otherwise), without the prior written permission of both the copyright owner and the above publisher of this book.

Published in Great Britain by Dorling Kindersley Limited

A catalog record for this book is available from the Library of Congress.

ISBN 978-0-7566-9042-7

DK books are available at special discounts when purchased in bulk for sales promotions, premiums, fund-raising, or educational use. For details, contact: DK Publishing Special Markets, 375 Hudson Street, New York, New York, 10014 or SpecialSales@dk.com

Reproduced by Bright Arts, China, and MDP, UK
Printed and bound in China by Leo Paper Products

Discover more at
www.dk.com

CONTENTS

Introduction

What is a Mineral? 8
Mineral Groups and Associations 10
Classifying Minerals 12
Identifying Minerals 14
What are Crystals? 18
Crystal Habits 20
Crystal Systems 22
Gems 24
What is a Rock? 26
Collecting Rocks and Minerals 28
Equipment 30
Organization, Storage and Cleaning 32

Minerals

Native Elements 36
Sulfides 48
Sulfosalts 65
Oxides 74
Hydroxides 100
Halides 106
Carbonates 113
Borates 126
Sulfates, Molybdates, Chromates and Tungstates 131
Phosphates, Vanadates, and Arsenates 147
Silicates 166
Organics 246

Rocks

Igneous Rocks 256
Metamorphic Rocks 286
Sedimentary Rocks 304
Meteorites 334

Glossary 338
Index 343
Acknowledgments 351

HOW THE ROCK AND MINERAL PROFILES WORK

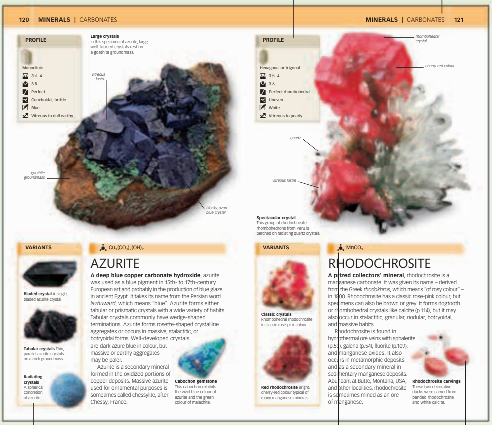

profile information (including illustration of crystal system in mineral entries)

name of mineral or rock group

variants panel containing named varieties and additional specimens

chemical formula of mineral

example of rock or mineral application

KEY

- Hardness
- Specific gravity
- Cleavage
- Fracture
- Streak
- Luster
- Type
- Origin
- Temperature of formation
- Pressure of formation
- Structure
- Grain size
- Major minerals
- Minor minerals
- Color
- Precursor rock
- Fossils

INTRODUCTION

WHAT IS A MINERAL?

A mineral is a naturally occurring solid with a specific chemical composition and a distinctive internal crystal structure. Most minerals are formed inorganically but some, such as those found in bone, are formed organically (by living organisms).

WHAT MINERALS ARE MADE OF

Most minerals are chemical compounds composed of two or more chemical elements. However, copper, sulfur, gold, silver, and a few others occur as single "native" elements. A mineral is defined by its chemical formula and by the arrangement of atoms within its crystals. For example, iron sulfide has the chemical formula FeS_2 (where Fe is iron and S is sulfur). Iron sulfide can crystallize in two different ways. When it crystallizes in the cubic system (pp.22–23), it is called pyrite; when it crystallizes in the orthorhombic system, it becomes the mineral marcasite. Minerals are classified by their chemical content: for example, those containing oxygen ions are called oxides and those having carbon and oxygen ions are called carbonates.

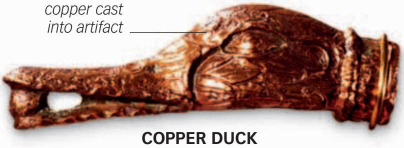

COPPER DUCK

Native elements
Native copper was probably the first metal used by humans. This duck's head was made in North Africa about 1,900 years ago.

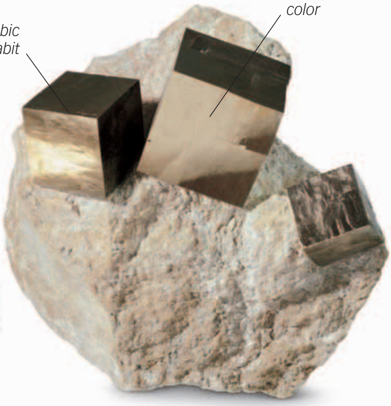

PYRITE

Same composition but different structure
Though pyrite and marcasite have the same chemical composition and are both iron sulfide, their differing crystal structures make them different minerals.

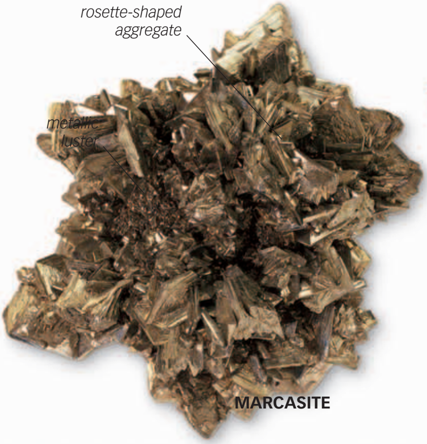

MARCASITE

Native sulfur
Sulfur is mined at Kawah Ijen, Java. Volcanic gases escaping from small openings in the ground (fumaroles) carry sulfur vapors to the surface, where it is deposited as a yellow crust.

Volcanic rhyolite
The Rhyolite Hills in Iceland are formed of rhyolite, a silica-rich rock produced as a result of volcanic activity. Rhyolite is made up of crystals of high-silica minerals.

ELECTRICAL CHARGE AND COMPOUNDS

A mineral compound is based on an electrical balance between a positively charged metal and a negatively charged part. In many minerals, negative charge is carried by a "radical": a combination of atoms acting as a single unit. For example, carbon and oxygen combine in a 1:3 ratio to give the CO_3 radical, which acts as a single, negatively charged unit.

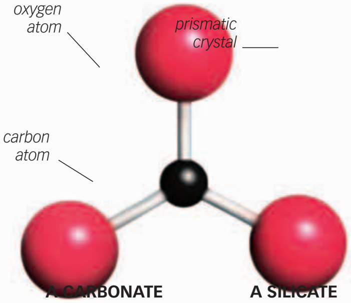
oxygen atom
prismatic crystal
carbon atom

A CARBONATE RADICAL

A SILICATE MINERAL

Simple and complex compounds
In carbonates, a simple carbon and oxygen group known as a radical combines with one or more metals.

COMMON MINERALS

There are more than 500 known minerals, but only about 100 of these are common. Silicon and oxygen make up about three-quarters of the crust by weight, and silicate minerals such as quartz, feldspar, and olivine are by far the most common minerals in rocks, making up 90 percent of the rocks at Earth's surface. The carbonates calcite and dolomite form sedimentary rocks, such as limestone.

Silicates
Silica tetrahedra link to form quartz. They can act as a radical to combine with one or more metals or semimetals to form other silicate minerals.

crystal face

QUARTZ CRYSTAL

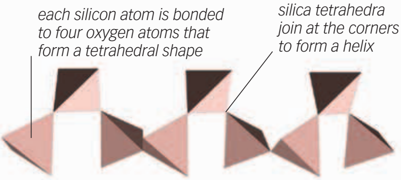
each silicon atom is bonded to four oxygen atoms that form a tetrahedral shape
silica tetrahedra join at the corners to form a helix

STRUCTURE OF QUARTZ

MINERAL GROUPS AND ASSOCIATIONS

Some minerals belong to chemical groups or series called solid solutions. In some circumstances, minerals are found together in groupings known as associations or assemblages. These patterns of occurrence can provide clues as to the minerals' origin.

SOLID SOLUTIONS

Some minerals do not have specific chemical compositions. Instead, they are homogenous mixtures of two minerals. These homogenous mixtures are known as solid solutions. For example, the olivine group of silicates includes forsterite and fayalite. Forsterite is a magnesium silicate, while fayalite is an iron silicate. Most olivine specimens are homogenous mixtures of the two, with the relative content of magnesium and iron varying in specimens. These minerals are described as part of a solid-solution series, in which forsterite and fayalite are the end-members.

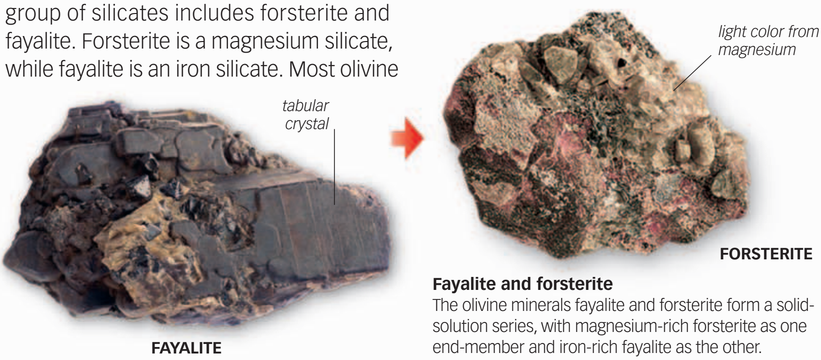

Fayalite and forsterite
The olivine minerals fayalite and forsterite form a solid-solution series, with magnesium-rich forsterite as one end-member and iron-rich fayalite as the other.

PRIMARY AND SECONDARY MINERALS

Primary minerals crystallize directly from magma and remain unaltered. They include essential minerals used to assign a classification name to a rock and accessory minerals that are present in lesser abundance and do not affect the classification of a rock. Secondary minerals are produced by the alteration of a primary mineral after its formation. For example, when copper-bearing primary minerals come into contact with carbonated water, they alter into secondary azurite or malachite.

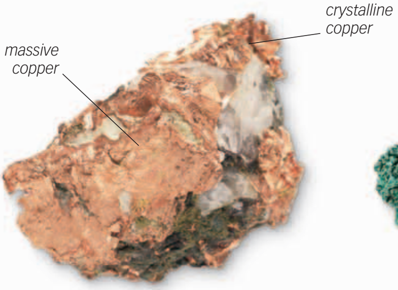

Primary copper mineral
Primary minerals, such as native copper, form directly in igneous rocks and remain unaltered. Their eventual alteration products are secondary minerals.

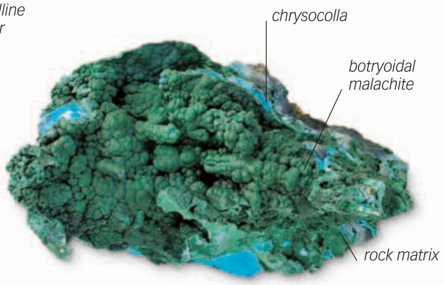

Secondary copper mineral
Chrysocolla and malachite are secondary minerals derived from the chemical weathering of primary copper minerals, such as native copper and bornite.

MINERAL ASSOCIATIONS

Some minerals are consistently found together over large areas because they are found in the same rock type. Other associations occur in encrustations, veins, cavities, or thin layers. The fact that certain minerals are likely to be found together can help in the discovery and identification of minerals. Lead and zinc ore minerals are often associated with calcite and barite, while gold is frequently found in association with quartz.

Associated minerals that form almost simultaneously and are usually present in a specific rock type make up an assemblage. Orthoclase, albite, biotite, and quartz form an assemblage for granite, and plagioclase, augite, magnetite, and olivine for gabbro. Assemblages are key indicators of the environments in which minerals form.

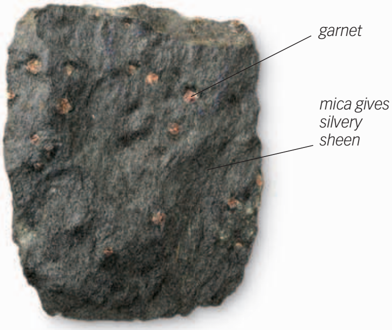

Metamorphic mix
The assemblage of garnet, quartz, and mica in this specimen indicates that this rock formed at moderate pressure and low temperatures (up to 400°F/200°C).

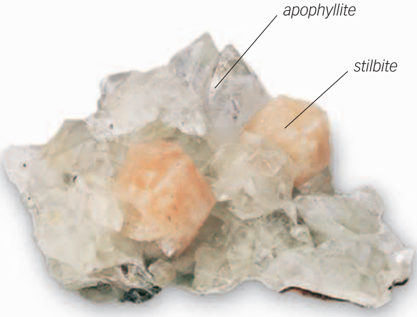

Zeolite association
Minerals belonging to the zeolite group of silicates, such as these crystals of apophyllite and stilbite, are often found in association with one another.

Layered rocks in the San Juan River
Erosion at this canyon in Utah, USA, has exposed layers of shale. Differences in the assemblage of minerals in various shale layers can reveal much about the geological history of the region.

CLASSIFYING MINERALS

Classification of minerals is an ongoing study among mineralogists—geologists who specifically study minerals. The ability to delve deep into the structure and chemistry of minerals has increased dramatically with advances in instruments and techniques.

MINERAL OR NOT?

The term "mineral" is commonly applied to certain organic substances, such as coal, oil, and natural gas, when referring to a nation's wealth in resources. However, these materials are more accurately referred to as hydrocarbons. Gases and liquids are not, in the strict sense, minerals. Although ice—the solid state of water—is a mineral, liquid water is not; nor is liquid mercury, which can be found in mercury ore deposits. Synthetic equivalents of minerals, for example emeralds and diamonds produced in the laboratory, are not minerals because they do not occur naturally. The "minerals" referred to in foods are also not strictly minerals—they refer to elements, such as iron, calcium, or zinc.

Synthetic ruby boule
Rubies and other gems grown synthetically are not classified as minerals. Some gems, such as yttrium-aluminum garnet, do not even occur in nature.

CHEMICAL FORMULAE

A chemical formula identifies the atoms present in a mineral and their proportions. In some minerals, the atoms and their proportions are fixed. Pyrite, for example, is always FeS_2, denoting iron (Fe) and sulfur (S) in a 1:2 ratio. In solid solutions, the components may be variable. For olivine, where complete substitution is possible between iron and magnesium (Mg), the formula is $(Fe,Mg)_2SiO_4$, indicating that iron and magnesium are found in varying amounts.

CHEMICAL ELEMENTS

Symbol	Name	Symbol	Name	Symbol	Name	Symbol	Name
Ac	Actinium	**Er**	Erbium	**Mo**	Molybdenum	**Sb**	Antimony
Ag	Silver	**Es**	Einsteinium	**N**	Nitrogen	**Sc**	Scandium
Al	Aluminum	**F**	Fluorine	**Na**	Sodium	**Se**	Selenium
Am	Americium	**Fe**	Iron	**Nb**	Niobium	**Si**	Silicon
Ar	Argon	**Fm**	Fermium	**Nd**	Neodymium	**Sm**	Samarium
As	Arsenic	**Fr**	Francium	**Ne**	Neon	**Sn**	Tin
At	Astatine	**Ga**	Gallium	**Ni**	Nickle	**Sr**	Strontium
Au	Gold	**Gd**	Gadolinium	**No**	Nobelium	**Ta**	Tantalum
B	Boron	**Ge**	Germanium	**Np**	Neptunium	**Tb**	Terbium
Ba	Barium	**H**	Hydrogen	**O**	Oxygen	**Tc**	Technetium
Be	Beryllium	**He**	Helium	**Os**	Osmium	**Te**	Tellurium
Bi	Bismuth	**Hf**	Hafnium	**P**	Phosphorus	**Th**	Thorium
Bk	Berkelium	**Hg**	Mercury	**Pa**	Protactinium	**Ti**	Titanium
Br	Bromine	**Ho**	Holmium	**Pb**	Lead	**Tl**	Thallium
C	Carbon	**I**	Iodine	**Pd**	Palladium	**Tm**	Thulium
Ca	Calcium	**In**	Indium	**Pm**	Promethium	**U**	Uranium
Cd	Cadmium	**Ir**	Iridium	**Po**	Polonium	**V**	Vanadium
Ce	Cerium	**K**	Potassium	**Pt**	Platinum	**W**	Tungsten
Cf	Californium	**Kr**	Krypton	**Pr**	Praseodymium	**Xe**	Xenon
Cl	Chlorine	**La**	Lanthanum	**Pu**	Plutonium	**Y**	Yttrium
Cm	Curium	**Li**	Lithium	**Ra**	Radium	**Yb**	Ytterbium
Co	Cobalt	**Lu**	Lutetium	**Rb**	Rubidium	**Zn**	Zinc
Cr	Chromium	**Lw**	Lawrencium	**Re**	Rhenium	**Zr**	Zirconium
Cs	Cesium	**Md**	Mendelevium	**Rh**	Rhodium		
Cu	Copper	**Mg**	Magnesium	**Rn**	Radon		
Dy	Dysprosium	**Mn**	Manganese	**S**	Sulfur		

CLASSIFYING MINERALS

Minerals are primarily classified according to their chemical composition. Shown below are the major chemical groups, with an example of each. Minerals are further classified into subgroups, with each subgroup taking its name from its most typical mineral. A radical is a group of atoms that acts as a single unit.

GRAPHITE

Native elements
Minerals formed of a single chemical element—metals such as gold and copper and nonmetals such as sulfur and carbon—are called native elements.

BRUCITE

Hydroxides
Hydroxide minerals contain a hydroxyl (hydrogen and oxygen) radical combined with a metallic element. In brucite, the metallic element is magnesium.

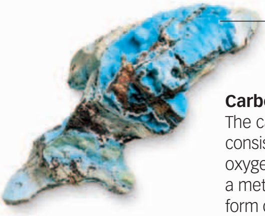

SMITHSONITE

Carbonates
The carbonate radical, consisting of carbon and oxygen, combines with a metal or semimetal to form carbonate minerals. In smithsonite, the metal is zinc.

COLEMANITE

Borates and nitrates
Borates contain radicals of boron and oxygen, and nitrates, radicals of nitrogen and oxygen. In colemanite boron and oxygen combine with calcium and water.

AMETHYST

Silicates
In this group, silicon and oxygen form a silica radical that combines with metals or semimetals. Silica occurs alone as quartz, as in this amethyst specimen.

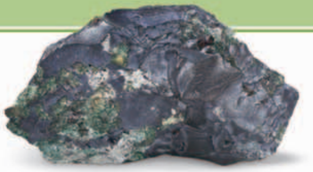

CHALCOCITE

Sulfides
The sulfides are formed when a metal or semimetal combines with sulfur. In chalcocite, the metallic element is copper.

RUBY

Oxides
When oxygen alone combines with a metal or semimetal, an oxide is formed. Corundum is aluminum oxide, with a red variety called ruby.

Halides
A halogen element (chlorine, bromine, iodine, or fluorine) combined with a metal or semimetal makes a halide. Sylvite is a compound of chlorine and potassium.

SYLVITE

APATITE

Arsenates, phosphates, and vanadates
In these minerals, a radical of oxygen and either arsenic, phosphorus, or vanadium combines with a semimetal or metal. Apatite is a phosphate.

Sulfates, chromates, tungstates, and molybdates
Sulfur, molybdenum, chromium, or tungsten form a radical with oxygen that combines with a metal or semimetal. Celestine is a sulfate.

CELESTINE

AMBER

Organic minerals
This group includes some naturally occurring substances, such as shell and coral, that are generated by organic means. Amber is a fossil resin.

IDENTIFYING MINERALS

There are certain physical properties determined by the crystalline structure and chemical composition of a mineral. These can commonly help to identify minerals without the use of expensive equipment. Even a beginner can readily use these pointers.

COLOR

Some minerals have characteristic colors—the bright blue of azurite, the yellow of sulfur, and the green of malachite allow for easy identification. This is not true of all minerals—fluorite occurs in virtually all colors, so it is best identified by other properties.

In minerals, color is caused by the absorption or refraction of light of particular wavelengths. This can happen for several reasons. One is the presence of trace elements—"foreign" atoms that are not part of the basic chemical makeup of the mineral in the crystal structure. As few as three atoms per million can absorb enough of certain parts of the visible-light spectrum to give color to some minerals. Color can also result from the absence of an atom or ionic radical from a place that it would normally occupy in a crystal. The structure of the mineral itself, without any defect or foreign element, may also cause color: opal is composed of minute spheres of silica that diffract light; and the thin interlayering of two feldspars in moonstone gives it color and sheen.

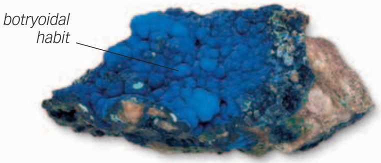

Azurite
Some minerals can be identified by their characteristic color. The copper carbonate azurite is always azure blue.

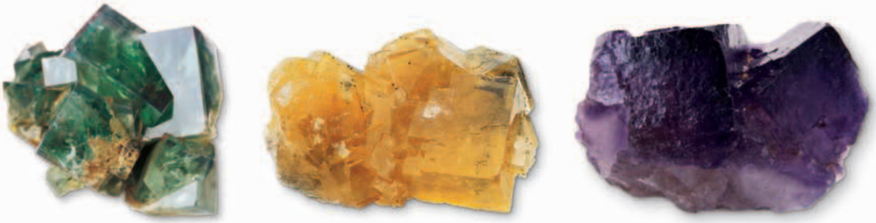

GREEN FLUORITE

YELLOW FLUORITE

PURPLE FLUORITE

Color range
These specimens show only a few of the many colors that can occur in fluorite. Different coloration depends on a number of factors.

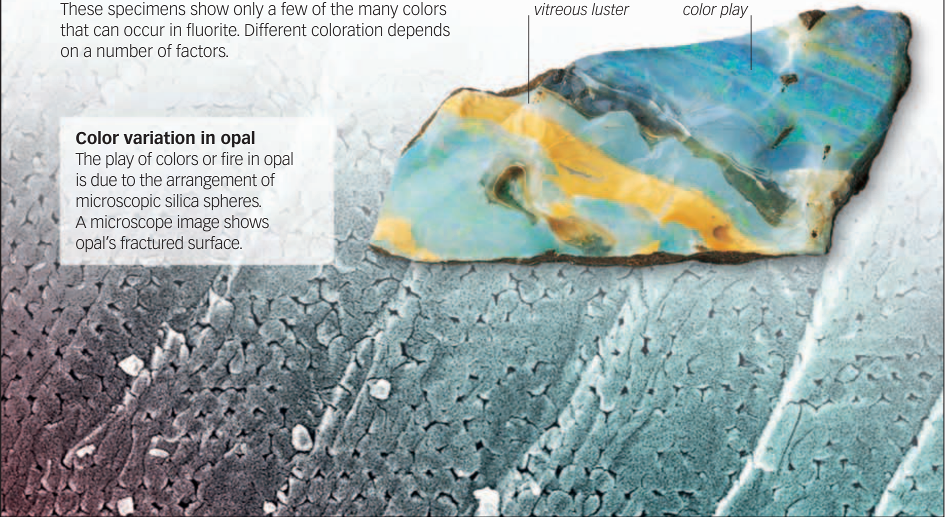

Color variation in opal
The play of colors or fire in opal is due to the arrangement of microscopic silica spheres. A microscope image shows opal's fractured surface.

LUSTER

A mineral's luster is the appearance of its surface in reflected light. There are two broad types of luster: metallic and nonmetallic. Metallic luster is that of an untarnished metal surface, such as gold, silver, or copper. These minerals tend to be opaque. Minerals with nonmetallic luster commonly show transparency or translucency. Vitreous describes the luster of a piece of broken glass; adamantine, the brilliant luster of diamond; resinous, the luster of a piece of resin; and pearly, the luster of mother-of-pearl or pearl. Greasy luster refers to the appearance of being covered with a thin layer of oil, and silky, the appearance of the surface of silk or satin. Dull luster implies little or no reflection, and earthy luster the nonlustrous look of raw earth.

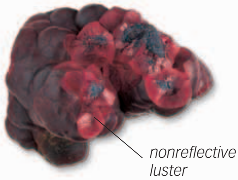

Dull
A dull luster is seen in this specimen of hematite. It is nonreflective but not as granular in appearance as earthy luster.

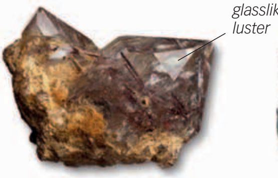

Vitreous
Many silicate minerals, such as this quartz crystal, have a vitreous luster. This luster appears similar to the surface of glass.

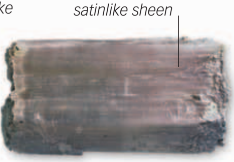

Silky
The borate ulexite exhibits a silky luster, with the surface sheen resembling a bolt of satin or silk.

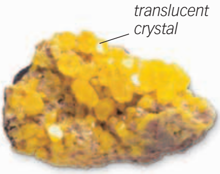

Resinous
Native sulfur crystals are transparent or translucent, with a resinous luster that resembles the surface of tree resin.

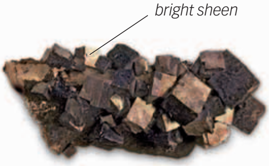

Metallic
The sulfide galena has a metallic luster and a distinctive cleavage. Metallic luster looks like the reflection from new metal.

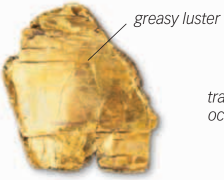

Greasy
Orpiment can appear greasy—resembling an oily surface—or resinous. The difference between the two lusters is subjective.

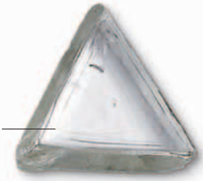

Adamantine
Adamantine is the brightest of lusters, with an appearance similar to the surface of this diamond. It is brighter than vitreous luster.

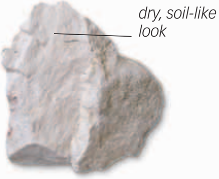

Earthy
Minerals with an earthy luster, such as this fine-grained calcite, have the look of freshly broken, dry soil.

STREAK

The color of the powder produced when a specimen is drawn across a surface such as unglazed porcelain is known as streak. A mineral's streak is consistent and is a more useful diagnostic indicator than its color, which can vary. Streak can help distinguish between minerals that are easy to confuse. For example, the iron oxide hematite has a red streak, while magnetite, another iron oxide, gives a black streak.

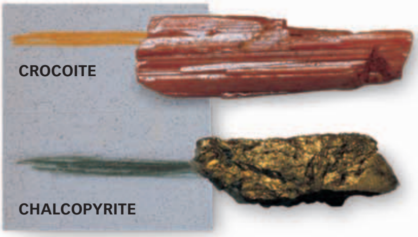

Consistent streak
The streak of a mineral is consistent from specimen to specimen, as long as an unweathered surface is tested. It is the same as the color of the powdered mineral.

CLEAVAGE

The ability of a mineral to break along flat, planar surfaces is called cleavage. It occurs in the crystal structure where the forces that bond atoms are the weakest. Cleavage surfaces are generally smooth and reflect light evenly. Cleavage is described by its direction relative to the orientation of the crystal and by the ease with which it is produced. If cleavage easily produces smooth, lustrous surfaces, it is called perfect. Distinct, imperfect, and difficult indicate less easy kinds of cleavage. Minerals may have different quality cleavages in different directions. Some have no cleavage at all.

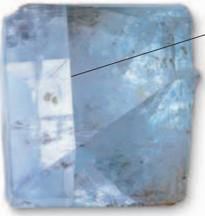

Perfect cleavage
This topaz crystal exhibits perfect cleavage. It breaks cleanly parallel to its base, and is thus said to have perfect basal cleavage.

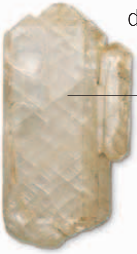

Clear breaks
The cleavage planes of this baryte crystal are clearly visible. Baryte has perfect cleavage in different directions, as seen in this specimen.

FRACTURE

Some minerals can break in directions other than along cleavage planes. These breaks, known as fractures, help in identifying minerals. For example, hackly fractures (with jagged edges), are often found in metals, while shell-like conchoidal fractures are typical of quartz. Other terms for fractures include even (rough but more or less flat), uneven (rough and completely irregular), and splintery (with partially separated fibers).

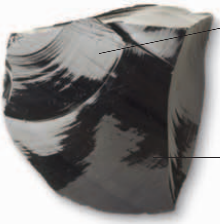

Conchoidal
This obsidian nodule shows conchoidal fracture, with fractures shaped like a bivalve seashell. It is commonly seen in silicates.

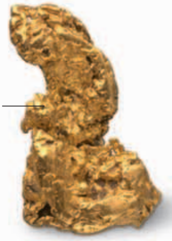

Hackly
This gold nugget shows hackly fracture, with sharp edges and jagged points. It is characteristic of most metals.

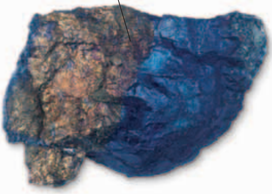

Uneven
This specimen of chalcopyrite shows uneven fracture. Its broken surface is rough and irregular, with no pattern evident.

TENACITY

The term tenacity describes the physical properties of a mineral based on the cohesive force between atoms in the structure. Gold, silver, and copper are malleable and can be flattened without crumbling. Sectile minerals can be cut smoothly with a knife; flexible minerals bend easily and stay bent after pressure is removed; ductile minerals can be drawn into a wire; brittle minerals are prone to breakage; and elastic minerals return to the original form after they are bent.

Ductile copper
Like many other native metals, copper is ductile. This means that it can be drawn into a wire without breaking.

Malleable gold
The malleability of gold allows it to be wrought into elaborate shapes. It can also be hammered into sheets thinner than paper.

HARDNESS

The hardness of a mineral is the relative ease or difficulty with which it can be scratched. A harder mineral will scratch a softer one, but not vice versa. Minerals are assigned a number between 1 to 10 on the Mohs scale, which measures hardness relative to ten minerals of increasing hardness. Hardness differs from toughness or strength; very hard minerals can be quite brittle. Most hydrous minerals—those that contain water molecules—are soft, as are phosphates, carbonates, sulfates, halides, and most sulfides. Anhydrous oxides—those without water molecules—and silicates are relatively hard.

Fingernail test
The fingernail is about 2½ on the Mohs scale and can scratch talc and gypsum. The hardness of other common items is also noted on the scale.

THE MOHS SCALE OF HARDNESS

Hardness	Mineral	Other materials for hardness testing
1	Talc	Very easily scratched by a fingernail
2	Gypsum	Can be scratched by a fingernail
3	Calcite	Just scratched with a copper coin
4	Flourite	Very easily scratched with a knife but not as easily as calcite
5	Apatite	Scratched with a knife with difficulty
6	Orthoclase	Cannot be scratched with a knife but scratches glass with difficulty
7	Quartz	Scratches glass easily
8	Topaz	Scratches glass very easily
9	Corundum	Cuts glass
10	Diamond	Cuts glass

REFRACTIVE INDEX

Light changes velocity and direction as it passes through a transparent or translucent mineral. The extent of this change is measured by the refractive index: the ratio of light's velocity in air to its velocity in the crystal. A high index causes dispersion of light into its component colors. Refractive indices can be found using specialized liquids or inexpensive equipment.

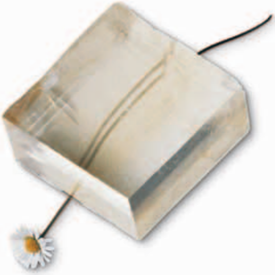

Double refraction
A calcite rhomb is said to be double refractive. It refracts light at two different angles, thus creating a double image.

FLUORESCENCE

Some minerals exhibit fluorescence—that is, they emit visible light of various colors when subjected to ultraviolet radiation. Ultraviolet lights for testing fluorescence can be obtained from dealers selling collectors' equipment. Fluorescence is an imperfect indicator of a mineral's identity because not all specimens of a mineral show fluorescence, even if they look identical and come from the same location.

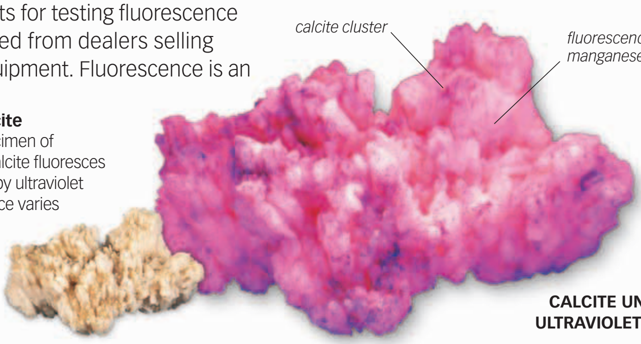
calcite cluster
fluorescence from manganese
CALCITE UNDER NATURAL LIGHT
CALCITE UNDER ULTRAVIOLET LIGHT

Manganoan calcite
This yellowish specimen of manganese-rich calcite fluoresces rose pink when lit by ultraviolet light. Its fluorescence varies with manganese concentration.

WHAT ARE CRYSTALS?

Virtually all minerals are crystalline—solids in which the component atoms are arranged in a particular, repeating, three-dimensional pattern. All crystals of a mineral are built with the same pattern. Some are 10 feet long; others can only be seen with a microscope.

ATOMIC STRUCTURE

A crystal is built up of individual, identical, structural units of atoms or molecules called unit cells. A crystal can consist of only a few unit cells or billions of them. The unit cell is repeatedly repeated in three dimensions, forming the larger internal structure of the crystal. The shape of the unit cell and the symmetry of the structure determine the positions and shapes of the crystal's faces.

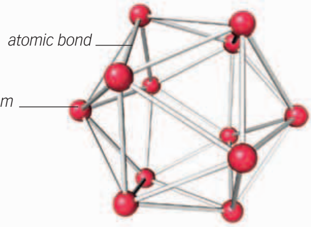

Crystal structure
Stick-and-ball diagrams, such as this one, show how each atom in the structure of a crystal is bonded to others.

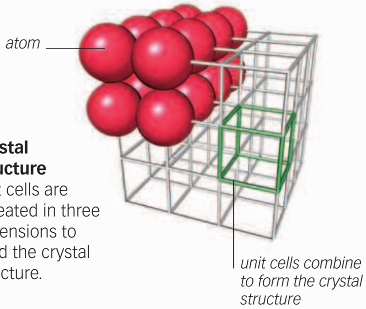

Crystal structure
Unit cells are repeated in three dimensions to build the crystal structure.

Crystals of many different minerals have unit cells that are similar in shape but are made of different chemical elements. The final development of the faces of a crystal is determined by the symmetry of the atomic structure and by the geological conditions at the time of its formation. Certain faces may be emphasized, while others disappear altogether. The final form taken by a crystal is known as its habit (pp.20–21).

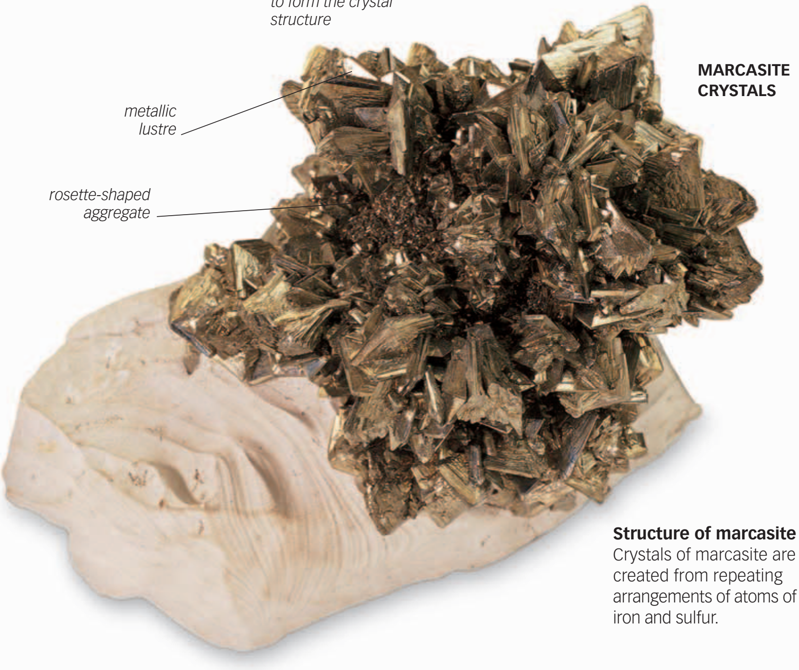

Structure of marcasite
Crystals of marcasite are created from repeating arrangements of atoms of iron and sulfur.

CRYSTAL SYMMETRY

All crystals exhibit symmetry because each crystal is built up of repeating geometric patterns. These patterns of crystal symmetry are divided into six main groups, or crystal systems (pp.22–23). The first of these symmetrical patterns is the cubic system, in which all crystals exhibit cubic symmetry. The characteristics of cubic symmetry may be explained as follows: if opposite face centers of a cube-shaped cubic crystal, such as halite, are held between the thumb and forefinger and the crystal is rotated through 360 degrees, the pattern of faces will appear identical four times as the different faces and edges come into view.

All cubic crystals have three axes of fourfold symmetry. They have other axes of symmetry, but these differ among classes within the cubic system. For example, cube-shaped crystals of halite have three axes of fourfold symmetry, in addition to its four axes of threefold symmetry.

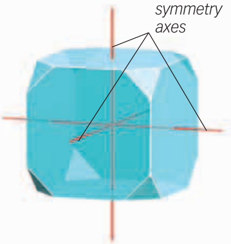

Cubic symmetry
All cubic crystals, such as those of halite (right), have three axes of fourfold symmetry.

cubic crystal
rock matrix

HALITE

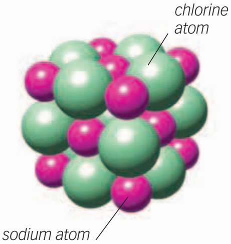

Halite atomic structure
This diagram shows the cubic arrangement of sodium and chlorine atoms in the halite structure.

TWIN CRYSTALS

When two or more crystals of the same species (a group of minerals that are chemically similar), such as gypsum or fluorite, form a symmetrical intergrowth, they are referred to as twinned crystals. Twins can be described as interpenetrating or contact. Penetration twinning may occur with individual crystals at an angle to one another—for example, forming a cross. It can also occur with individual crystals parallel to one another, as in Carlsbad twinning. If a twin involves three or more individual crystals, it is referred to as a multiple twin or a repeated twin. Albite often forms multiple twins. Many other minerals form twins, but they are particularly characteristic of some, such as the "fishtail" contact twins of gypsum or the penetration twins of fluorite.

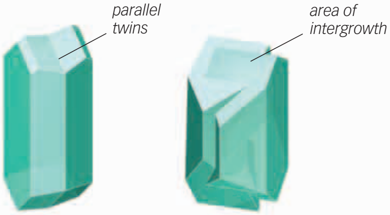

CONTACT TWIN

CARLSBAD PENETRATION TWIN

Contact and penetration twins
Parallel twinning is a kind of contact twinning in which two or more crystals share a common face or faces. Penetration twinning results from crystals growing into each other.

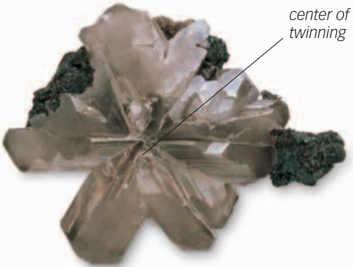

Cyclic twin
Cyclic twins occur when more than two crystals are twinned at a common center. This specimen of cerussite shows the cyclic twinning of three crystals all at 60° angles to each other.

CRYSTAL HABITS

Habit refers to the external shape of a crystal or an assemblage of intergrown crystals. It includes names of crystal's faces, such as prismatic and pyramidal, names of forms, such as cubic and octahedral, and descriptive terms, such as bladed and dendritic.

CRYSTAL FACES

The three types of crystal face—prism, pyramid, and pinacoid—are determined by a relationship to a crystallographic axis (p.22). Prism faces are parallel to the axis; pyramid faces cut through the axis at an angle; and pinacoid faces are at right angles to the axis. A crystal may have numerous sets of pyramid faces, each at a different angle to the c axis. Crystals may also have major and minor prism faces with edges parallel to each other. In most crystals, some faces are more developed than others.

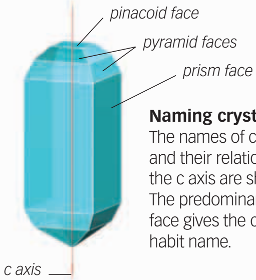

Naming crystal faces
The names of crystal faces and their relationship to the c axis are shown here. The predominant crystal face gives the crystal its habit name.

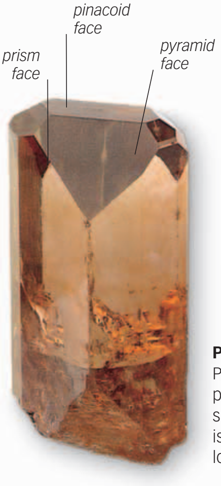

Prismatic topaz
Although this topaz crystal exhibits prismatic, pryamidal, and pinacoidal faces, the bulk of the crystal is defined by its prism faces and it is therefore called prismatic.

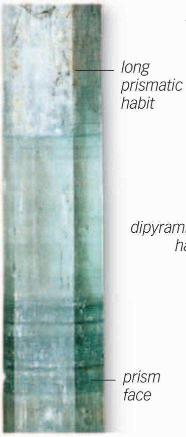

Prismatic
Prism faces clearly predominate in this long specimen of beryl. Its habit is therefore described as long prismatic.

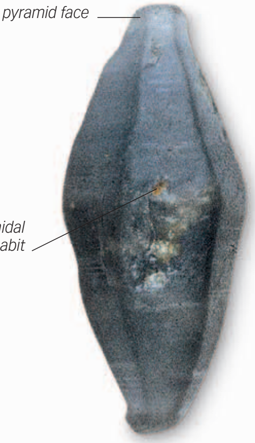

Pyramidal
If pyramid faces dominate in one direction, the habit is pyramidal. If pyramid faces dominate in both directions, as in this specimen of sapphire, the habit is dipyramidal.

CRYSTAL FORMS

Habits can be named after crystal forms: "cubic" implies crystallizing in the form of cubes; "dodecahedral," in the form of dodecahedrons; and "rhombohedral," in the form of rhombohedrons. When crystals of one system crystallize in forms that appear to be the crystals of another system, the habit name is preceded by the word "pseudo." When terminations take different forms in the same crystal, the habit is known as hemimorphic.

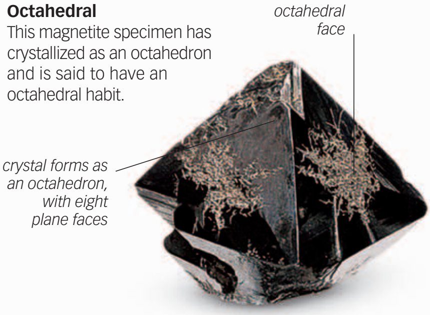

Octahedral
This magnetite specimen has crystallized as an octahedron and is said to have an octahedral habit.

AGGREGATES

Aggregates are groups of intimately associated crystals. In general, aggregates are intergrowths of imperfectly developed crystals. In some aggregates, the crystals may be microscopic. The type of aggregation is often typical of a particular mineral species. Terms used to describe aggregates include granular, fibrous, radiating, botryoidal, stalactitic, geodic, and massive.

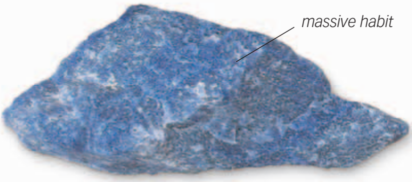

Massive
The massive habit occurs when there is a mass of crystals that cannot be seen individually, as in this specimen of dumortierite.

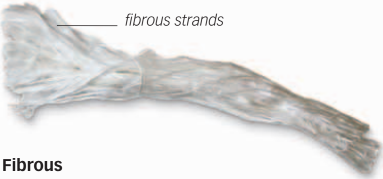

Fibrous
The fibrous habit is an aggregate, consisting of slender, parallel, or radiating fibers. This tremolite specimen is a good example.

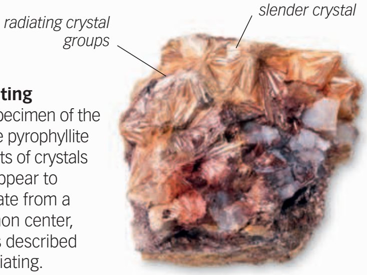

Radiating
This specimen of the silicate pyrophyllite consists of crystals that appear to originate from a common center, so it is described as radiating.

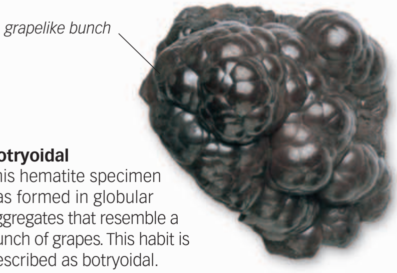

Botryoidal
This hematite specimen has formed in globular aggregates that resemble a bunch of grapes. This habit is described as botryoidal.

CRYSTAL APPEARANCE

Some habits are descriptions of the general appearance of a crystal. The term "tabular" describes a crystal with large, flat, parallel faces; "bladed" describes elongated crystals that are flattened like a knife blade; "stalactitic" describes crystal aggregates shaped like stalactites; and "blocky" or "equant" describes crystals with faces that are roughly the same size in all directions.

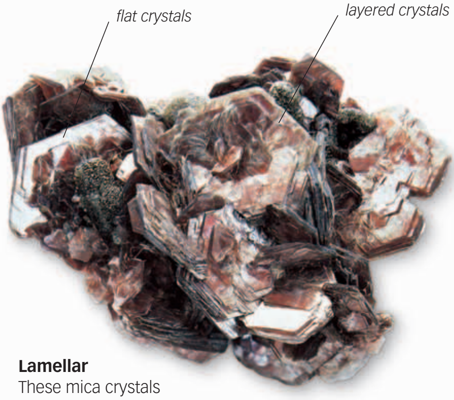

Lamellar
These mica crystals are described as having a lamellar habit. They are flat, platelike, individual crystals arranged in layers.

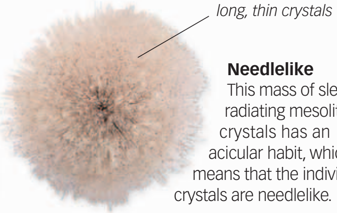

Needlelike
This mass of slender, radiating mesolite crystals has an acicular habit, which means that the individual crystals are needlelike.

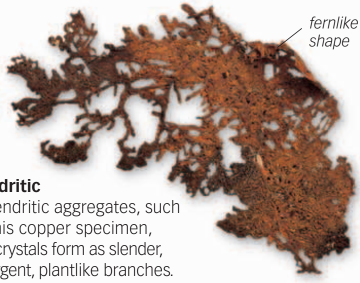

Dendritic
In dendritic aggregates, such as this copper specimen, the crystals form as slender, divergent, plantlike branches.

CRYSTAL SYSTEMS

Crystals are classified into six different systems according to the maximum symmetry of their faces. Each crystal system is defined by the relative lengths and orientation of its three crystallographic axes—imaginary lines that pass through the centre of an ideal crystal.

CUBIC

Cubic crystals have three crystallographic axes (a_1, a_2, and a_3) at right angles and of equal length, and four threefold axes of symmetry. The main forms within this system are cube, octahedron, and rhombic dodecahedron. Halite, copper, gold, silver, platinum, iron, fluorite, and magnetite crystallize in the cubic system, which is also known as the isometric system.

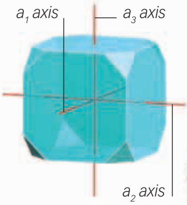

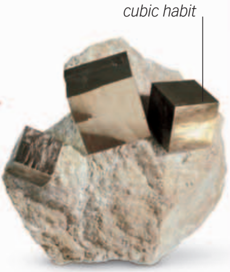

Cubic crystal system
Pyrite crystals commonly form as cubes, but they can also occur as pentagonal dodecahedra and octahedra, or combinations of all three forms.

TETRAGONAL

Tetragonal crystals have three crystallographic axes at right angles—two equal in length (a_1 and a_2), and the third (c) longer or shorter. These crystals have one principal, fourfold axis of symmetry. Crystals look like square or octahedral prisms in shape. Rutile, zircon, cassiterite, and calomel are minerals that crystallize in the tetragonal system.

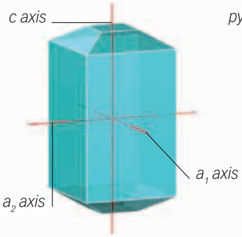

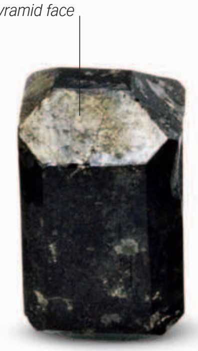

Vesuvianite
This vesuvianite crystal—with prismatic, pyramidal, and pinnacoid faces—shows a classic tetragonal form.

HEXAGONAL AND TRIGONAL

Some crystallographers consider hexagonal and trigonal crystals to comprise a single system, whereas others regard them as forming separate systems. Both crystalline forms have three crystallographic axes (a_1, a_2, and a_3) of equal length. These are at 120 degrees to one another and to a fourth axis (c), which is perpendicular to the plane of the other three axes. Trigonal crystals have only threefold symmetry, whereas hexagonal crystals have sixfold symmetry. Minerals that crystallize in the hexagonal system include beryl (emerald and aquamarine) and apatite. Some of the minerals that crystallize in the trigonal system are calcite, quartz, and tourmaline.

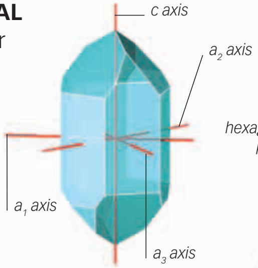

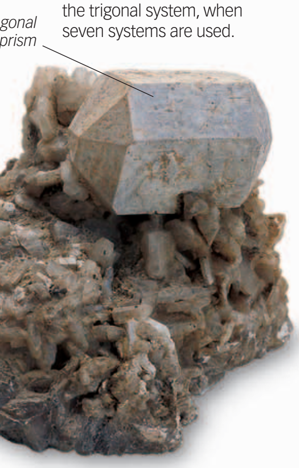

Apatite
Apatite forms hexagonal-shaped prisms, and its overall symmetry is hexagonal. Many "hexagonal" minerals fall into the trigonal system, when seven systems are used.

MONOCLINIC

The term “monoclinic” means “one incline.” Monoclinic crystals have three crystallographic axes of unequal length. One (c) is at right angles to the other two (a and b). These two axes are not perpendicular to each other, although they are in the same plane. The crystals have one twofold axis of symmetry. More minerals crystallize in the monoclinic system than in any other crystal system. Examples are gypsum, orthoclase, malachite, and jadeite.

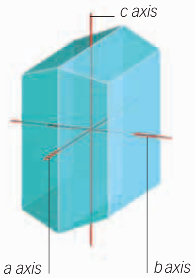

Gypsum
The parallelogram shape of this crystal of gypsum demonstrates the two unequal crystallographic axes and the third axis at right angles of the monoclinic system. Orthoclase, which belongs to the same system, often forms twinned crystals.

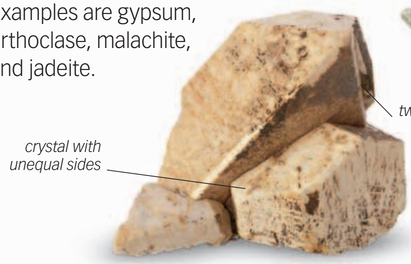

ORTHOCLASE

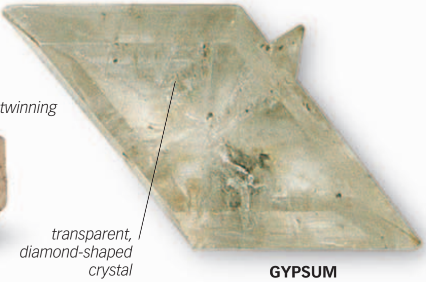

GYPSUM

ORTHORHOMBIC

Orthorhombic means “perpendicular parallelogram”. Crystals in this system have three crystallographic axes (a, b, and c) at right angles, all of which are unequal in length. They have three twofold axes of symmetry. Minerals that crystallize in this system include olivine, aragonite, topaz, marcasite, and barite.

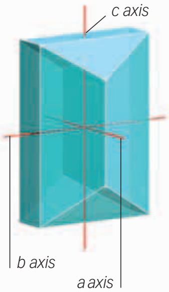

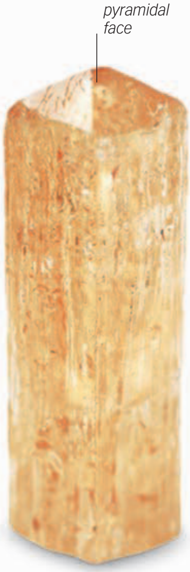

Topaz
The mineral topaz often forms beautiful, orthorhombic prismatic crystals that are usually terminated by pyramids or other prisms. The mineral barite also forms orthorhombic prisms.

TRICLINIC

Triclinic crystals have the least symmetrical shape, with three crystallographic axes of unequal length (a, b, and c) inclined at angles other than 90 degrees to each other. The orientation of a triclinic crystal is arbitrary. Minerals that crystallize in this system include albite, anorthite, kaolin, and kyanite.

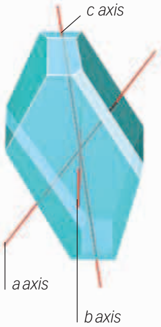

triclinic axinite crystal

Axinite
The silicate axinite is a classic triclinic mineral. Several feldspars, including albite and microcline, are also triclinic.

GEMS

A gem is any mineral that is highly prized for its beauty, durability, and rarity. It is enhanced in some manner by altering its shape, usually by cutting and polishing. Most gems begin as crystals of minerals or as aggregates of crystals.

HISTORY OF GEMS

The use of gemstones in human history goes back to the Upper Paleolithic Period (25,000–12,000 BCE). People were initially drawn by the bright colors and beautiful patterns of gems. When the shaping of stones for adornment first began, opaque and soft specimens were used. As shaping techniques improved, harder stones began to be cut into gems. Beads of the quartz varieties hard carnelian and rock crystal were fashioned in Mesopotamia (now Iraq) in the 7th millennium BCE. Records of the time suggest that people thought that stones had a mystic value—a belief that persists to the present.

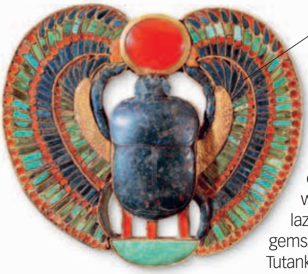

Ancient masterpiece
This ancient Egyptian chest ornament is inlaid with gold, finely cut lapis lazuli, carnelian, and other gems. It is from the tomb of Tutankhamun (*c.*1361–1352 BCE).

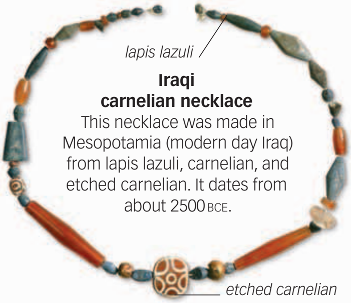

Iraqi carnelian necklace
This necklace was made in Mesopotamia (modern day Iraq) from lapis lazuli, carnelian, and etched carnelian. It dates from about 2500 BCE.

GEM MINING

Gemstone deposits form in different geological environments. Perhaps the best known are the "pipes" of kimberlite, from which most diamonds are recovered by the hard-rock methods of drilling and blasting. Other gems also recovered from the rock in which they form are quartz varieties, opal, tourmaline, topaz, emerald, aquamarine, some sapphires and rubies, turquoise, lapis lazuli, and chrysoberyl. Hard and dense gemstones that are impervious to chemical weathering are carried by water to placer deposits such as river beds, beaches, and the ocean floor. Placer mining techniques mimic the creation of the placer by separating denser minerals in running water. The simplest methods are panning and sieving, or passing gravel through a trough of flowing water with baffles at the bottom. The lighter material washes away but denser gemstones remain.

Gem panning
Many gemstone minerals, such as sapphire and ruby, are heavier than normal stream gravels. These can be recovered using the slow but thorough panning method.

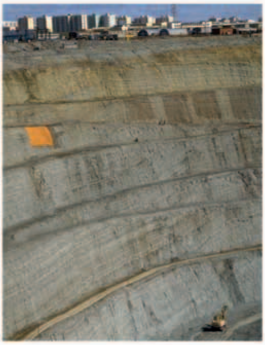

Diamond mine in Siberia
Russia has become a major supplier of diamonds. In this mine, diamonds are being recovered from a diamond pipe.

FACETING

Gemstones can be shaped in several ways. Opaque or translucent semiprecious stones, such as agate and jasper, are tumble-polished, carved, engraved, or cut with a rounded upper surface and a flat underside. Grinding and polishing of flat faces on the stone is called faceting. Facets are placed in specific geometric positions at specific angles according to the bending of light within a particular stone. Transparent stones, such as amethyst, diamond, and sapphire, are faceted to maximize their brilliance and "fire" or enhance color. Although much material is ground away while cutting, the final value is much enhanced.

Cutting a brilliant
While faceting gemstones, care must be taken to preserve the maximum material and produce the best brilliance and color.

Rough choice
The faceter selects his rough based on color, clarity, and shape, which determine the cut for the final gem.

Sawn in two
The rough is sawn to roughly the final shape of the gem. Accurate sawing saves time in the grinding process.

Faceting begins
The major facets are first ground onto the gem. The accuracy of these determines the final brilliance.

Further facets
Smaller facets are cut after the major facets. Based on the cut, there may be only a few or dozens of these.

Finished off
After the first side of the stone is cut, it is reversed and facets are placed on the second side in the same order.

GEM CUTS

There are three basic types of facet cut: step (with rectangular facets), brilliant (with triangular facets), and mixed (a combination of the two). The first faceting probably involved diamond cutting in Italy prior to the 15th century. First, only the natural faces of octahedral diamond crystals were polished. The rose cut was developed in the 17th century. By about 1700, the brilliant cut (today's favorite for diamonds and other colorless gems) was created. The emerald cut was soon developed to save valuable material, as its rectangular cut conforms to the shape of emerald crystals. Today there are hundreds of possible gem cuts.

Gemstone shapes
A principal criterion for the cutter in choosing a gemstone shape is the shape of the rough gemstone. This ensures that a minimum of valuable material is lost.

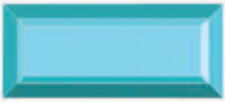

BAGUETTE

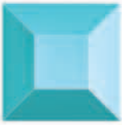

TABLE

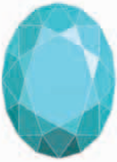

OVAL

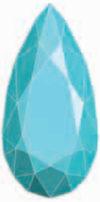

PENDELOQUE

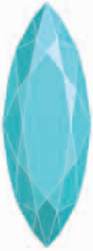

MARQUISE

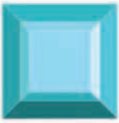

SQUARE

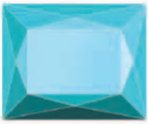

SCISSOR

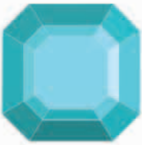

OCTAGONAL

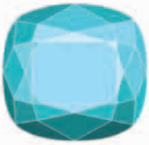

CUSHION

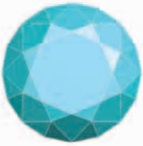

ROUND

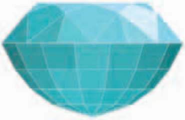

MIXED

WHAT IS A ROCK?

A rock is a naturally occurring and coherent aggregate of one or more minerals. There are three major classes of rock—igneous, sedimentary, and metamorphic. Each of these three classes is further subdivided into groups and types.

TYPES OF ROCK

Igneous rocks form from melted rock called magma. When magmas solidify underground, intrusive rocks such as granite are created. Intrusive rocks are also known as plutonic rocks. If the magma flows onto the surface of the land or ocean bed, extrusive rocks such as basalt, are formed.

Sedimentary rocks are usually made of deposits laid down on Earth's surface by water, wind, or ice. They almost always occur in layers or strata. Stratification survives compaction and cementation and is a distinguishing feature of sedimentary rocks. Some sedimentary rocks are of chemical origin, having been deposited in solid form from a solution. Others are of biochemical origin and are composed predominantly of the compound calcium carbonate.

When existing rocks are subjected to extreme temperatures or pressures, or both, their composition, texture, and internal structure may be altered to form metamorphic rocks. The original rocks may be igneous, sedimentary, or metamorphic.

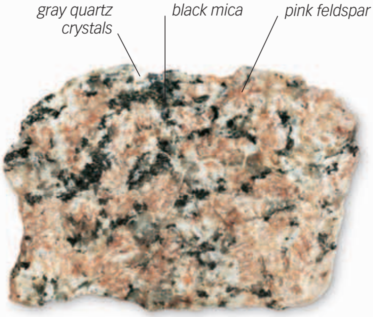

Pink granite
In this specimen of igneous pink granite, the three essential components of all granites can be seen: quartz, alkali feldspar, and mica.

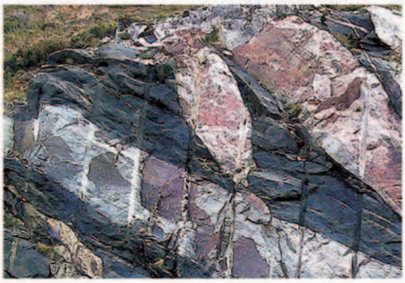

Pegmatite dike
Light-colored bands of igneous hydrothermal pegmatite, composed principally of quartz, can be seen here cutting across darker bands of metamorphic gneiss.

Volcanic growth
The eruption of extrusive magmas can create volumes of igneous rock measured in cubic miles on the surface of Earth.

Marble quarry
This dazzling white marble being quarried at Carrara, Italy, is metamorphosed from a very pure limestone.

Meteorites are not considered igneous, sedimentary, or metamorphic but are a group of their own. Many are remnants of asteroids, which are themselves remnants of the formation of the Solar System. Some meteorites are remains of the nickel-iron cores of asteroids; some contain nickel-iron and minerals such as olivine from the mantles of asteroids; and others are made up principally of silicate minerals.

Sedimentary layers
The Colorado River cuts through layers of sedimentary rock in the Grand Canyon, USA. The highest layers are the youngest, while the deepest are the oldest.

THE ROCK CYCLE

The series of processes by which rocks are created, broken down, and reconstituted as new rocks is known as the rock cycle. These processes depend on pressure, temperature, time, and changes in environmental conditions in Earth's crust and surface. At various stages in the rock cycle, old rocks are broken down, new minerals form, and new rocks originate from the components of the old. Thus a rock that began at the surface as an igneous rock may be reworked into a sedimentary rock, metamorphic rock, or new igneous rock to continue the cycle again.

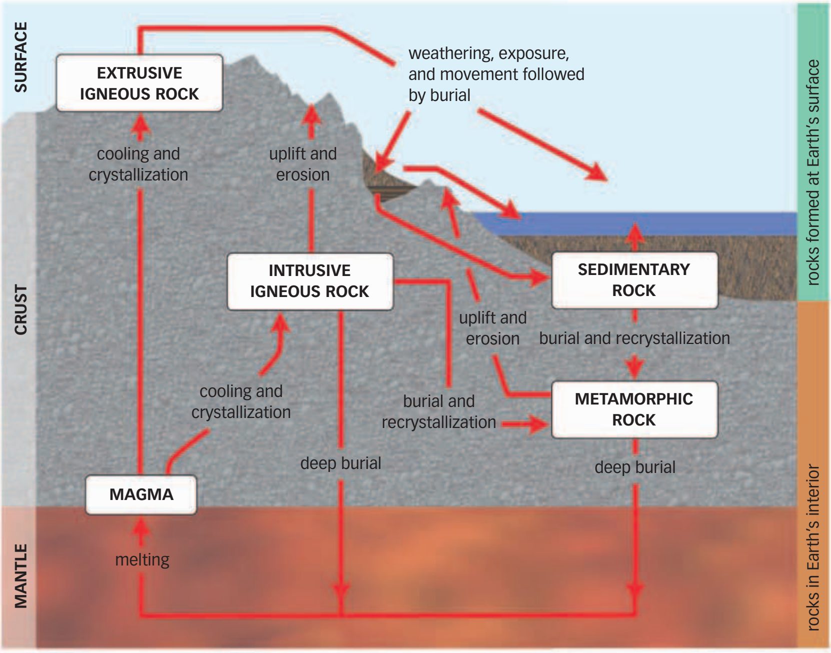
SURFACE
CRUST
MANTLE
EXTRUSIVE IGNEOUS ROCK
weathering, exposure, and movement followed by burial
cooling and crystallization
uplift and erosion
INTRUSIVE IGNEOUS ROCK
SEDIMENTARY ROCK
uplift and erosion
burial and recrystallization
cooling and crystallization
burial and recrystallization
METAMORPHIC ROCK
deep burial
deep burial
MAGMA
melting
rocks formed at Earth's surface
rocks in Earth's interior

The rock cycle
This diagram summarizes the various elements of the rock cycle, from the creation of fresh igneous rock through erosion, deposition, and its reconstitution into new rock.

COLLECTING ROCKS AND MINERALS

The world of rocks, minerals, gems, and fossils offers endless possibilities for the hobbyist. Only a small amount of specialized knowledge is required to open a whole world of enjoyment of some of nature's finest creations.

WHERE TO LOOK

Most collectors begin by just accumulating rocks, minerals, and fossils. As their collection grows, they start being more selective, keeping only specimens with better color and crystallization and more interesting crystal forms. A wide range of specimens can be purchased from dealers, but it is often more enjoyable to find your own. In many countries, there are guidebooks that give precise directions to collecting localities for rocks, minerals, and fossils.

Sample collection is not without its constraints: working mines and quarries have legal restrictions on people permitted on their premises; old mines are dangerous; old mine dumps have been gone over for decades by other collectors; and public access to land is often restricted. However, traditional collecting sites, such as road cuttings and eroded cliffs on shorelines, continue to provide excellent opportunities for collectors.

Field experience
Rock collecting can be a hobby for a lifetime, as the collector develops knowledge and skills to enhance the activity.

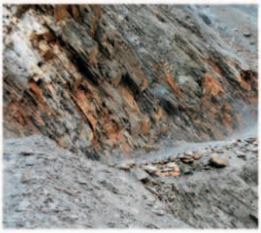

Road cutting
The bank of a road cut through this pegmatite rock reveals giant feldspar crystals. Many fine rock and mineral specimens are derived from road cuttings.

Looking for gold
The gold pan is an essential piece of kit for a collector. Many gemstones, such as garnet and sapphire, can be found by panning.

There is also an increasing number of collecting localities that are open to the public on the payment of a fee. Some clubs for collecting enthusiasts have their own collecting sites, and they also arrange trips to sites that are otherwise inaccessible to the public. Collectors should bear in mind that permission must always be sought to collect samples on private property.

Old working
While mine dumps are good sources of specimens, hidden workings and old machinery can pose a hazard to unwary collectors.

Tempting tunnels
Old mine shafts can be tempting, but are often highly dangerous places. In most cases better specimens are usually found in the mine dumps outside.

SAFETY AND THE COLLECTING CODE

While mineral collecting is generally a safe hobby, there are a few definite hazards that a collector needs to be aware of. The most dangerous collecting localities are around old mines and workings. Tunnels should never be entered—shoring timbers rot quickly, and cave-ins and rock falls are almost guaranteed to happen. Collectors must also pay attention to what is underfoot—old shafts are sometimes covered over. In any case, there is often remarkably poor collecting inside old mines because most material of value has usually already been removed by miners. Mine dumps, by contrast, can be a good source of specimens. However, caution should be exercized because mine dumps are often loosely piled and can be unstable.

When collecting in beach cliffs, road cuttings, and rock falls, pay attention not only to loose material underfoot but also to anything that may fall or roll from above. It is best to avoid a collecting locality if you are not sure that is safe.

TAKING NOTES

When they start out, new collectors often ignore the need to write down information about their finds. But experience soon shows that investing in a notebook and devoting the minimal amount of time it takes to keep at least basic notes is essential. It is especially important to make notes about exactly where specimens were found. A considerable time may go by before you revisit the locality, and by then, in the absence of notes, you will probably be unable to find the spot again. It is useful to make a sketch of important landmarks or outcrops, because these can help relocate a specific spot.

Map and compass
Tools such as a compass and a map or a GPS receiver are essential for identifying localities and relocating them at a later date.

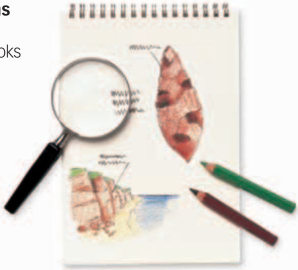

Drawing locations
It is useful to make drawings in notebooks of locations and the specimens they have yielded.

Correcting fluid
Number each specimen with a note about their find-spot. A dab of correction fluid makes a good label and can be removed if necessary.

EQUIPMENT

Mineral collecting is a safe hobby, but some simple pieces of equipment increase the safety factor dramatically. Just a few basics, such as the right hammer and chisel, a hard hat, goggles, gloves, and things you already have, will get you started.

FIELD EQUIPMENT

In addition to the basic collecting tools described here, safety equipment should be considered essential. Access to some collecting localities requires safety clothing such as a hard hat and fluorescent vest. Carry a cell phone with a fully charged battery with you even if you are only going a short distance from the car. A fall into a ravine or another low

HARD HAT
LEATHER GLOVES
SAFETY GOGGLES

Head and hand protection
Flying rock splinters and falling rocks cause injuries to collectors every year. Hands, eyes, and heads are particularly vulnerable areas. Goggles are recommended when breaking or splitting stone.

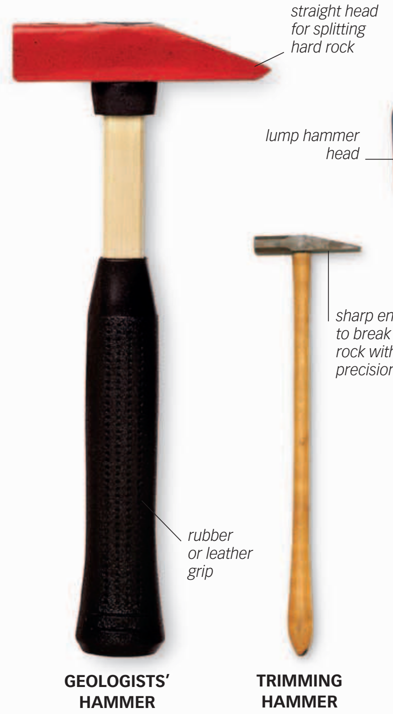
straight head for splitting hard rock
lump hammer head
sharp end to break rock with precision
rubber or leather grip
GEOLOGISTS' HAMMER
TRIMMING HAMMER

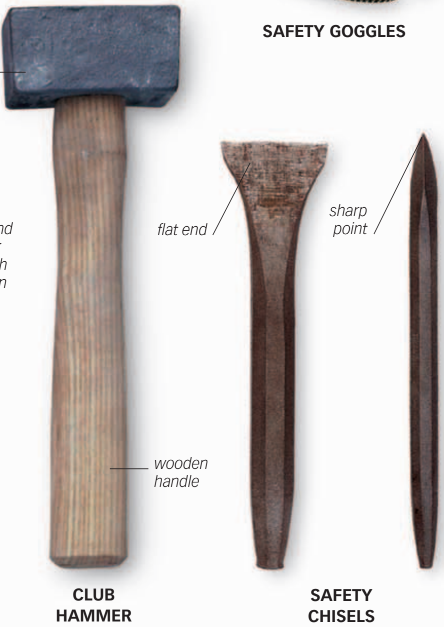
flat end
sharp point
wooden handle
CLUB HAMMER
SAFETY CHISELS

Hammers
Every year rock collectors are injured—sometimes blinded—by using the wrong hammers. Geologists' hammers are made of special steels. Their striking ends are beveled to prevent steel splinters from flying off.

Chisels
Like rock hammers, the chisels used by geologists are made from special steels that resist splintering. Not all are essential but having two or three of different sizes will make cutting rock safer.

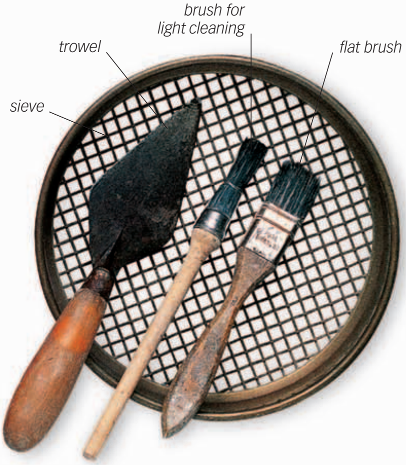

Extra tools
The experienced collector has a range of equipment for all collecting possibilities, from sieves and pans to various brushes and trowels. Most of these can be bought a few at a time as new collecting localities are visited.

spot may take you out of sight of potential help and add hours to the time it takes to find you. In desert country, an adequate supply of water is essential, and if you are in snake country take an appropriate snake-bite kit. Clothing suitable to the weather and terrain is, of course, vital. Leave your low-cut shoes and sneakers at home. Leather boots offer better protection from snake bites, cactus spines, sharp stones, jagged metal, and rolling stones, and ensure much better traction.

MAGNIFICATION

There is an entire area of mineral collecting devoted to tiny crystals known as micromounts. Small crystals often develop superb forms and groupings that are obscured as the process of crystallization progresses. Micromount collectors need effective microscopes, or at the least large magnifiers, to examine and enjoy these minute specimens. For collectors not wishing to incur the expense of a microscope, a simple hand lens will reveal much of the beauty of the tiny micromounts.

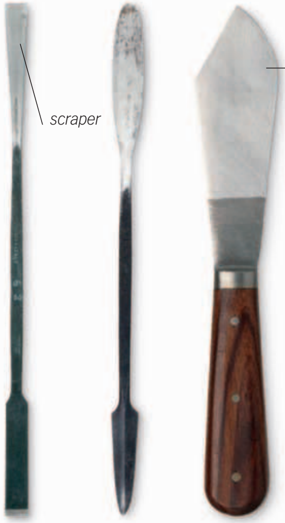

KNIVES AND SPATULAS

PEN KNIFE

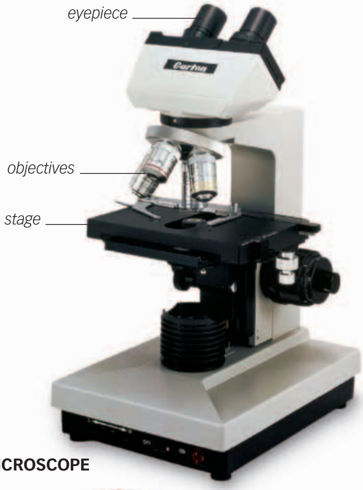

MICROSCOPE

magnifying glass

HAND LENS

Cleaning tools
There are two types of cleaning tool: those for field use and those for cleaning specimens at home. Tools for field use are more robust and are used for separating specimens from adhering rock.

A closer look
Most collectors of small crystals have a microscope to examine their specimens. The field equipment of every geologist and collector should include a hand lens with a magnification of about 10 times.

ORGANIZATION, STORAGE, AND CLEANING

Finding mineral specimens is only the first stage of collecting. The number of specimens damaged in the course of the journey home or while cleaning can be large. Care must therefore be taken from the moment a specimen is collected.

TRANSPORTING SPECIMENS

Wrapping of some sort is essential when transporting newly collected specimens, whether they are being carried in a backpack or a car. Delicate specimens should be wrapped first in tissue and then in newspaper. If your wrapping material is used up, try leaves, grass, or pine needles as a natural alternative. Unwrap wet specimens and let them dry as soon as you get home. Cotton balls and cellulose wadding should be kept entirely away from specimens, because the fibers are almost impossible to remove.

In the bag
Rock samples can be carried in a cloth specimen bag. More sensitive specimens require elaborate wrapping so that they can be transported safely.

CLEANING SPECIMENS

As a general rule, clean specimens as little as possible, starting with the gentlest methods first. Begin by using a soft brush to remove loose soil and debris. Hard rock specimens, such as gneiss or granite, are unlikely to be damaged by vigorous cleaning. With delicate minerals, such as calcite crystals, it is essential to use a fine, soft brush. Never use hot water to wash a specimen, because the heat may cause some minerals to crack or shatter. Toothbrushes that use a pulsing water jet are useful cleaning tools. Soaps should be avoided, but if you must use them, choose liquid dishwashing soaps over hand or toilet soaps, which have additives that can penetrate specimens. The use of ultrasonic cleaners is not recommended—they can shatter delicate specimens even at low intensities. Certain

Muddy rocks
Many specimens will be muddy or dirty when collected. Most dirt is more easily removed when it is dry and can be lightly brushed off.

DISTILLED WATER
HYDROCHLORIC ACID

Cleaning liquids
Distilled (or deionized) water is good as a final wash for minerals. Weak hydrochloric acid is good for cleaning silicates, but always be aware of the risks involved.

FINE POINTED SCRAPER

acids are suitable for cleaning specific minerals. Silicates are not harmed by weak acids, but carbonates and phosphates can be damaged by them. If you do use acids, seek specific information on their use from specialized books or other collectors.

Cleaning up
Removing rock with fine specialist tools is often necessary when collecting fossils. The mineral collector, by contrast, is more likely to brush or wash off dirt from specimens.

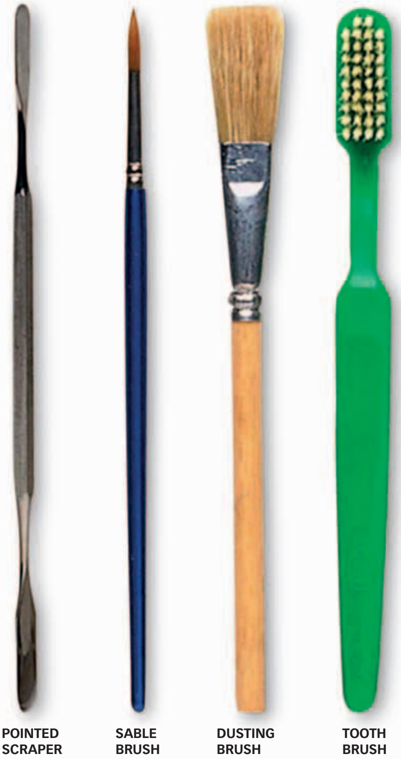

Cleaning tools
A variety of tools is useful for cleaning specimens. Each specimen will present a different cleaning problem, so a selection of tools is necessary.

STORAGE AND DISPLAY

Once specimens have been collected and cleaned, they need to be stored or, in the case of the most attractive pieces, displayed. Many collectors like to store specimens in card trays inside shallow drawers. Once collected, some minerals are liable to experience physical and chemical effects that may change or sometimes even destroy them. Fortunately, these problems are well known and preventative measures can be taken in advance.

Every specimen collected should be accompanied by a label with as much information about it as is feasible. For display, use a sturdy, preferably glass-fronted cabinet or shelf. Many guests will wish to handle specimens, but they may not be aware that handling can damage delicate examples.

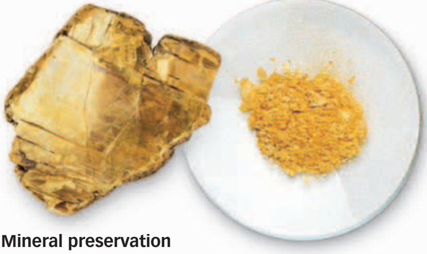

Mineral preservation
Minerals such as orpiment and realgar are sensitive to light and need special storage methods. Other minerals may require either dry or humid conditions.

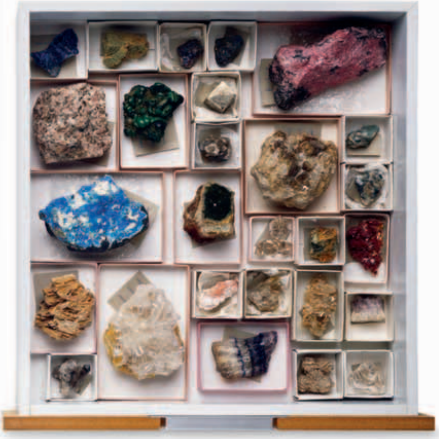

Informative display
People will admire your best specimens and also value information about them. Some collectors choose to provide museum-style information about specimens.

MINERALS

NATIVE ELEMENTS

There are 88 chemical elements known to occur in nature. Of these, less than two dozen are found uncombined with other elements. This group is called the native elements. Only eight of these native elements are found in significant quantities.

COMPOSITION

The native elements are classified into three groups: metals like copper and gold; semimetals like arsenic; and nonmetals like sulfur and carbon. The metals rarely form well-defined crystals; the semimetals typically occur as nodular masses; and the nonmetals form distinct crystals.

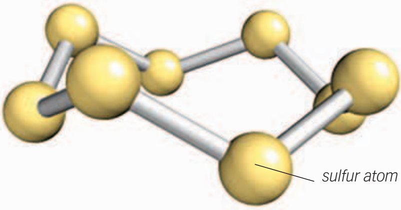

Sulfur crystal structure
In the orthorhombic crystal structure of sulfur, strongly bonded rings of eight sulfur atoms are weakly bonded to neighboring rings.

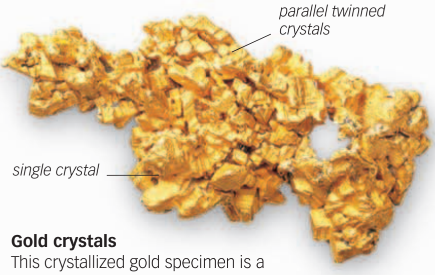

Gold crystals
This crystallized gold specimen is a rarity, because native metals rarely form well-defined crystals. Most occur in wirelike and branching forms or as nuggets.

OCCURRENCE AND USES

Native elements are known to form under a wide range of geologic conditions and in a variety of rock types. A native element can occur in several different environments. Some are found in sufficient concentrations to form economically important deposits.

Native gold and silver have been media of exchange for three millennia, and native copper and meteoric iron were among the first metals to be used by humans.

Industrial tools
This tool-maker is producing a diamond-edged industrial cutting tool. Although partly replaced by synthetic diamond, natural diamond continues to be used as an industrial abrasive.

Sulfur crust
Native sulfur builds up around fumaroles, where sulfur-rich gas is vented around volcanoes. These fumaroles often produce magnificently crystallized specimens.

PROFILE

Cubic
2½–3
8.9
None
Hackly, ductile
Rose
Metallic

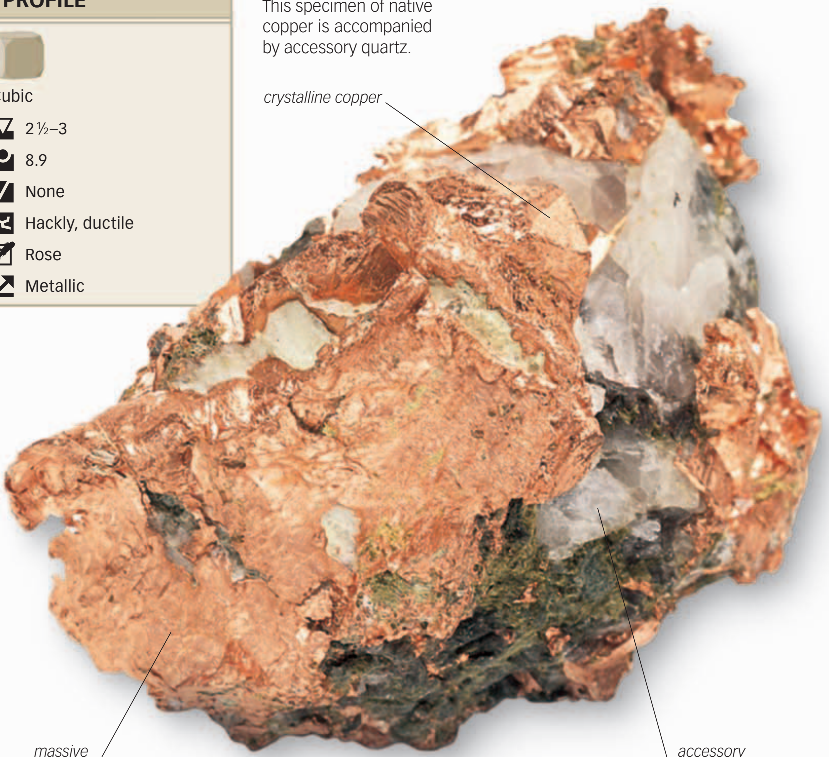

Native copper This specimen of native copper is accompanied by accessory quartz.

VARIANT

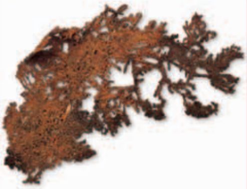

Dendritic copper A specimen of crystalline copper in the branching form

Cu

COPPER

In its free-occurring metallic state, copper was probably the first metal to be used by humans. Neolithic people are believed to have used copper as a substitute for stone by 8000 BCE. Around 4000 BCE, Egyptians cast copper in molds. By 3500 BCE, copper began to be alloyed with tin to produce bronze.

Copper is opaque, bright, and metallic salmon pink on freshly broken surfaces but soon turns dull brown. Copper crystals are uncommon, but when formed are either cubic or dodecahedral, often arranged in branching aggregates. Most copper is found as irregular, flattened, or branching masses. It is one of the few metals that occur in the "native" form without being bonded to other elements. Native copper seems to be a secondary mineral, a result of interaction between copper-bearing solutions and iron-bearing minerals.

Plumbing joint Because it is easy to shape and roll the metal, copper is widely used to make household pipes.

PROFILE

Cubic
4–4½
14.0–19.0
None
Hackly
Whitish steel-gray
Metallic

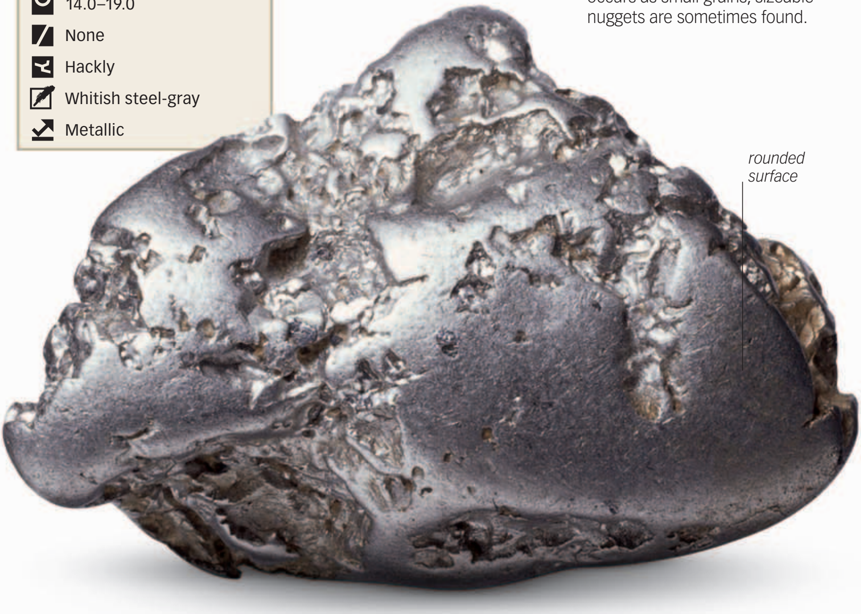

Platinum nugget Although most of the platinum mined from placer deposits occurs as small grains, sizeable nuggets are sometimes found.

VARIANTS

Granular habit Most platinum is recovered as small grains

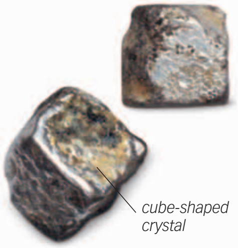

Platinum crystals Isolated cubic crystals of platinum

Pt

PLATINUM

The first documented discovery of platinum was by the Spaniards in the 1500s, in the alluvial gold mines of the Río Pinto, Colombia. They called it *platina del Pinto*, from *platina,* which means "little silver," thinking that it was an impure ore of silver. It was not recognized as a distinct metal until 1735. It is opaque, silvery gray, and markedly dense.

Platinum usually occurs as disseminated grains in iron- and magnesium-rich igneous rocks and in quartz veins associated with hematite (p.91), chlorite, and pyrolusite (p.80). When rocks weather, the heavy platinum accumulates as grains and nuggets in the resulting placer deposits. Crystals are rare, but when found they are cubic. Most platinum for commercial use is recovered from primary deposits. Native platinum typically contains iron and metals such as palladium, iridium, and rhodium.

Platinum ring A 2.5-carat, brilliant-cut diamond has been set in a platinum mounting in this ring.

Iron meteorite
Most native iron is in Earth's core, but iron from meteorites, such as this one, was used from about 3000 BCE. Native iron is usually alloyed with nickel.

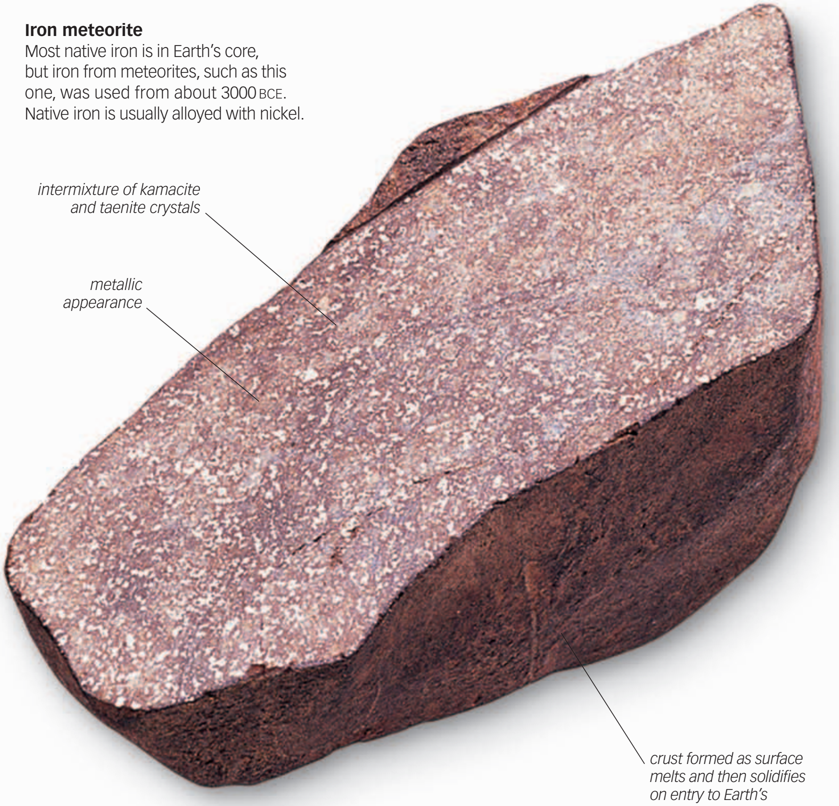
intermixture of kamacite and taenite crystals
metallic appearance
crust formed as surface melts and then solidifies on entry to Earth's atmosphere

PROFILE

Cubic
- 4
- 7.3–7.9
- Basal
- Hackly
- Steel-gray
- Metallic

Fe,Ni

IRON

Five percent of Earth's crust is made up of iron. Native iron is rare in the crust and is invariably alloyed with nickel. Low-nickel iron (up to 7.5 percent nickel) is called kamacite, and high-nickel iron (up to 50 percent nickel) is called taenite. Both crystallize in the cubic system. A third form of iron-nickel, mainly found in meteorites and crystallizing in the tetragonal system, is called tetrataenite. All three forms are generally found either as disseminated grains or as rounded masses.

Kamacite is the major component of most iron meteorites (p.335). It is found in most chondritic meteorites (p.337), and occurs as microscopic grains in some lunar rocks. Taenite and tetrataenite are mainly found in meteorites, often intergrown with kamacite. Iron is also plentiful in the Sun and other stars.

Viking axe head
This iron Viking axe head from Frykat, Denmark, has a shape commonly used in weapons.

PROFILE

Trigonal or hexagonal

- 2–2½
- 9.7–9.8
- Perfect basal
- Uneven
- Silver-white
- Metallic

Crystalline bismuth This group of intergrown bismuth crystals shows typical metallic luster and iridescence.

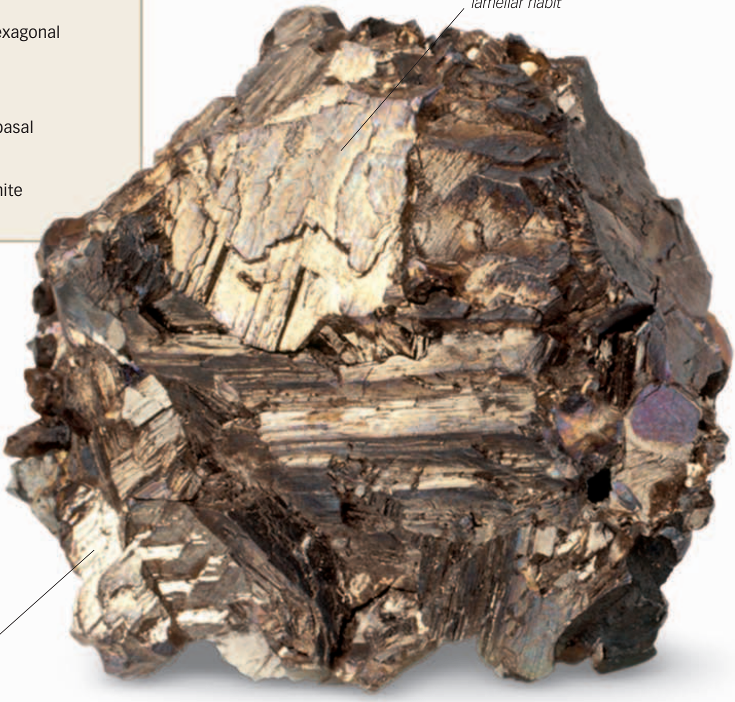

VARIANT

Native bismuth Partly crystalline bismuth on rock

Bi

BISMUTH

As a native metal, bismuth has been known since the Middle Ages. A German monk named Basil Valentine first described it in 1450. Bismuth is often found uncombined with other elements, forming indistinct crystals, often in parallel groupings. It is hard, brittle, and lustrous. It is also found in grains and as foliated masses. Silver-white, it usually has a reddish tinge that distinguishes it. Specimens may have an iridescent tarnish.

Bismuth is found in hydrothermal veins and in pegmatites (p.260) and is often associated with ores of tin, lead, or copper (p.37), from which it is separated as a by-product. Bismuth expands slightly when it solidifies, making its alloys useful in the manufacture of metal castings with sharp detailing. Bismuth salts are often used as soothing agents for digestive disorders.

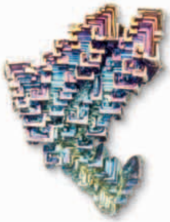

Hopper-shaped crystals Laboratory-grown bismuth crystals with cavernous faces like these exhibit an array of colors.

PROFILE

Hexagonal
- 3–3½
- 6.6–6.7
- Perfect basal
- Uneven
- Gray
- Metallic

Massive antimony
This specimen of massive antimony has a pale silvery gray color and the occasional small crystal.

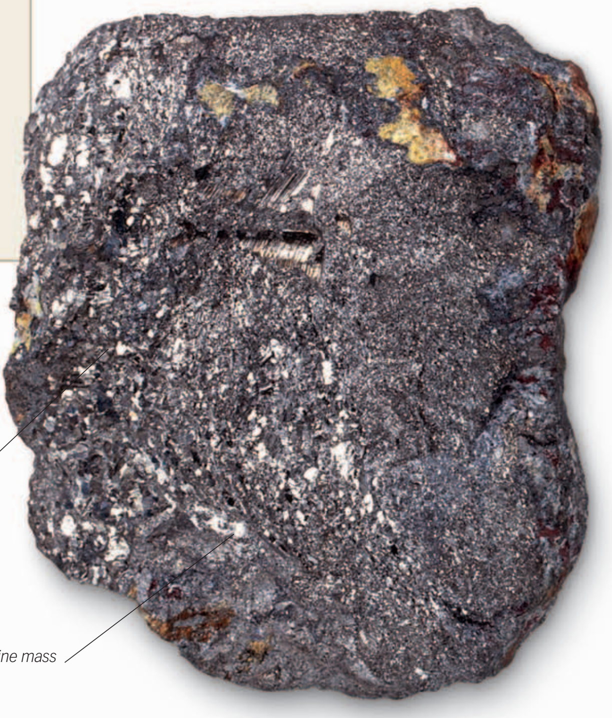

VARIANT

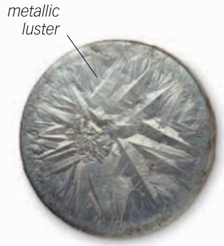

Antimony star A starry pattern of antimony formed when molten antimony is cooled

Sb

ANTIMONY

Although recognized as a metal since the 8th century or earlier, antimony was only identified as an element in 1748. Crystals are rare but when found are either psuedocubic or thick and tabular. Antimony usually occurs in massive, foliated, or granular form. It is lustrous, silvery, bluish white in color, and has a flaky texture that makes it brittle. It almost always contains some arsenic and is found in veins with silver (p.43), arsenic (p.45), and other antimony minerals.

Antimony is extremely important in alloys. Even in minor quantities, it imparts strength and hardness to other metals, particularly lead, whose alloys are used in the plates of automobile storage batteries, in bullets, and in coverings for cables. Combined with tin and lead, antimony forms antifriction alloys called babbitt metals, which are used as components of machine bearings. Like bismuth (p.40), antimony expands slightly on solidifying, making it a useful alloying metal for detailed castings.

PROFILE

Cubic

2½–3

19.3

None

Hackly

Golden yellow

Metallic

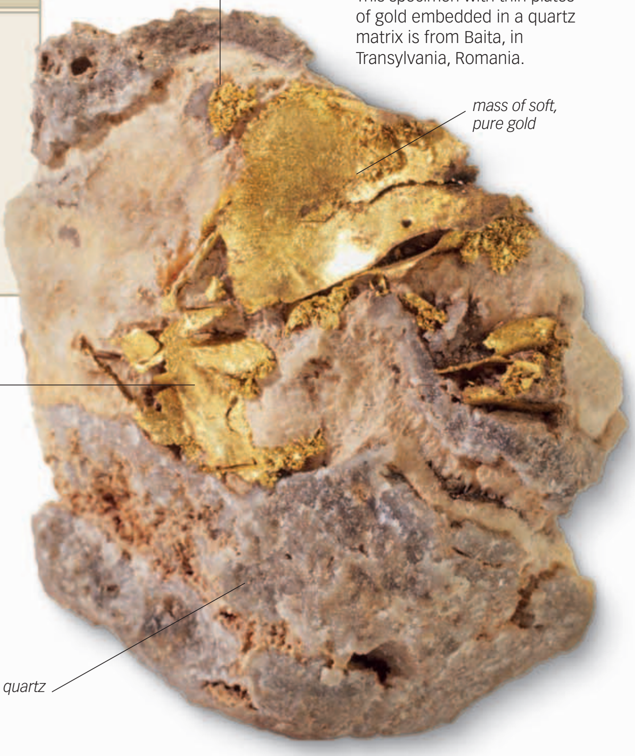

Scales of gold This specimen with thin plates of gold embedded in a quartz matrix is from Baita, in Transylvania, Romania.

VARIANTS

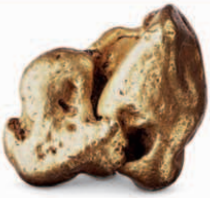

Gold nugget An irregularly shaped gold nugget

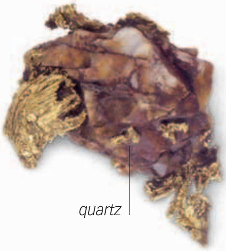

Gold crystals Crystalline gold in a dull quartz matrix

Au

GOLD

Throughout human history, gold has been the most prized metal. It is opaque, has a highly attractive metallic golden yellow color, is extremely malleable, and is usually found in a relatively pure form. It is remarkably inert, so it resists tarnish. These qualities have made it exceptionally valuable. Gold usually occurs as treelike growths, grains, and scaly masses. It rarely occurs as well-formed crystals, but when found these are octahedral or dodecahedral.

Gold is mostly found in hydrothermal veins with quartz (p.168) and sulfides. Virtually all granitic igneous rocks—in which it occurs as invisible, disseminated grains—contain low concentrations of gold. Almost all of the gold recovered since antiquity has come from placer deposits—weathered gold particles concentrated in river and stream gravel.

Garnet in gold This gold ring has an unusual demantoid (yellow-green) garnet set in it.

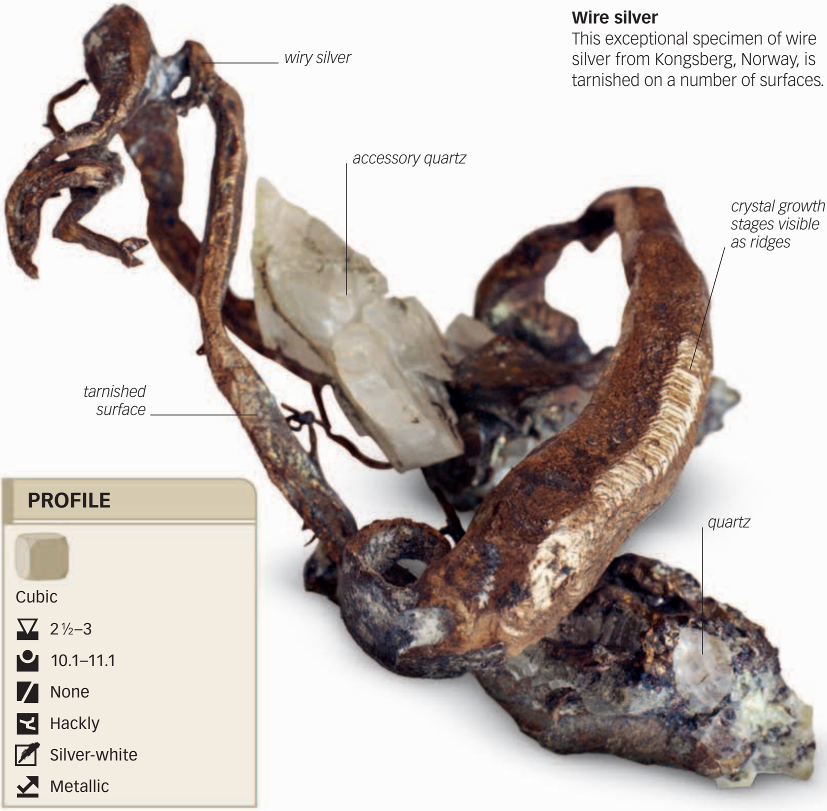

Wire silver This exceptional specimen of wire silver from Kongsberg, Norway, is tarnished on a number of surfaces.

PROFILE

Cubic
2½–3
10.1–11.1
None
Hackly
Silver-white
Metallic

VARIANTS

Tarnished silver A tarnished specimen of wiry silver

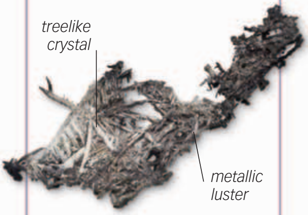

Dendritic silver Superbly crystalline, dendritic silver

Ag

SILVER

The earliest silver ornaments and decorations were found in tombs that date as far back as 4000 BCE. Silver coinage began to appear around 550 BCE. Opaque and bright silvery white with a slightly pink tint, silver readily tarnishes to either gray or black. Natural crystals of silver are uncommon, but when found they are cubic, octahedral, or dodecahedral. Silver is usually found in granular habit and as wiry, branching, lamellar, or scaly masses.

Widely distributed in nature, silver is a primary hydrothermal mineral. It also forms by alteration of other silver-bearing minerals. Much of the world's silver production is a by-product of refining lead, copper (p.37), and zinc. Silver is the second most malleable and ductile metal, and it is important in the photographic and electronic industries.

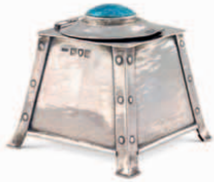

Silver inkwell This Guild of Handicraft textured silver inkwell of square, tapering form has a blue enamel cabochon.

PROFILE

Orthorhombic

1½–2½

2.1

Indistinct

Conchoidal to uneven, brittle

White

Resinous to greasy

Sulfur crystals Yellow orthorhombic crystals of sulfur are set in a rock matrix in this specimen from Conil, Andalucía, Spain.

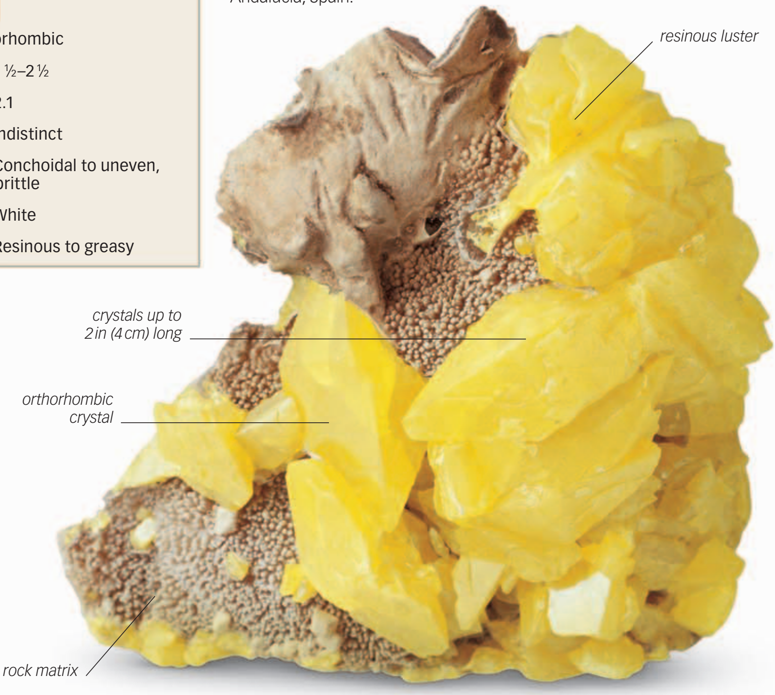

VARIANTS

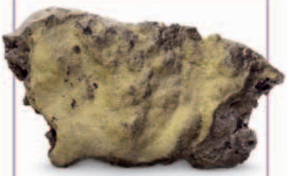

Fumarole crystals A crust of very small sulfur crystals from a fumarole in Java, Indonesia

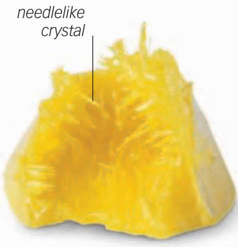

Acicular sulfur Elongated sulfur crystals on rock

S

SULFUR

The ninth most abundant element in the Universe, after oxygen and silicon, sulfur is the most abundant constituent of minerals. It occurs in the form of sulfides (pp.49–64), sulfates (pp.132–41), and elemental sulfur. The bright yellow or orangish color of sulfur makes the mineral easy to identify. Sulfur forms pyramidal or tabular crystals, encrustations, powdery coatings, and granular or massive aggregates. Crystalline sulfur may exhibit as many as 56 different habits.

Most sulfur forms in volcanic fumaroles, but it can also result from the breakdown of sulfide ore deposits. Massive sulfur is found in thick beds in sedimentary rocks, particularly those associated with salt domes. Sulfur is a poor conductor of heat, which means that specimens are warm to the touch.

Powdered sulfur Sulfur is used in a number of industrial and medicinal applications, including in the production of sulfuric acid.

PROFILE

Hexagonal
- 3½
- 5.7
- Perfect, fair
- Uneven, brittle
- Gray
- Metallic or dull earthy

Botryoidal arsenic In this specimen, native arsenic has a metallic luster and exhibits a typical botryoidal habit.

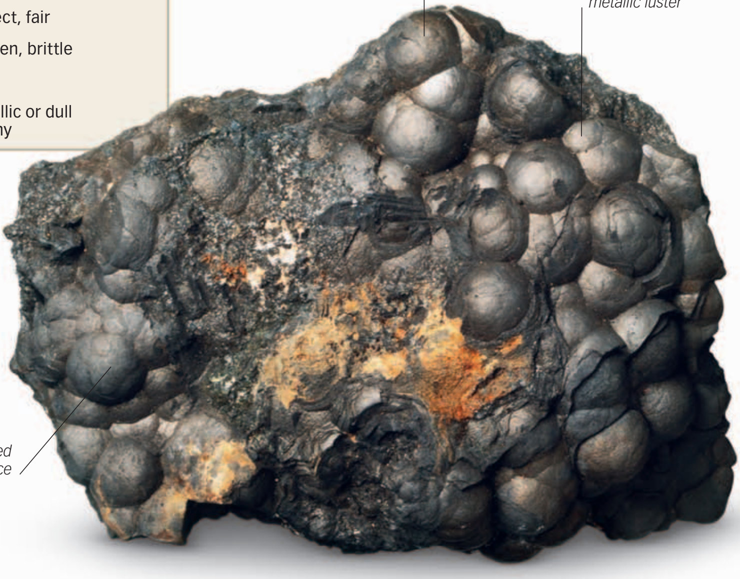

VARIANT

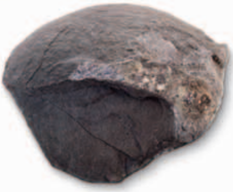

Massive arsenic A darkly tarnished, massive specimen of native arsenic

As

ARSENIC

Known since antiquity, arsenic is widely distributed in nature, although it is unusual in native form. It is classified as a semimetal, because it possesses some properties of metals and some of nonmetals. Crystals are rare, but when found they are rhombohedral. Arsenic usually occurs in massive, botryoidal to reniform, or stalactitic habits, often with concentric layers. On fresh surfaces, arsenic is tin-white, but it quickly tarnishes to dark gray.

Native arsenic is found in hydrothermal veins, often associated with antimony (p.41), silver (p.43), cobalt, and nickel-bearing minerals. It is highly poisonous, although it is used in some medicines to treat infections. Arsenic-based compounds can be used in alloys to increase high-temperature strength and as a herbicide and pesticide.

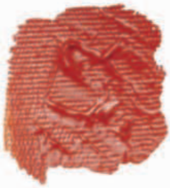

Arsenic paint Ancient Egyptian artists used orange-red colors made from powdered arsenic sulfide.

PROFILE

Hexagonal

1–2

2.2

Perfect basal

Uneven

Black to steel-gray, shiny

Metallic or dull earthy

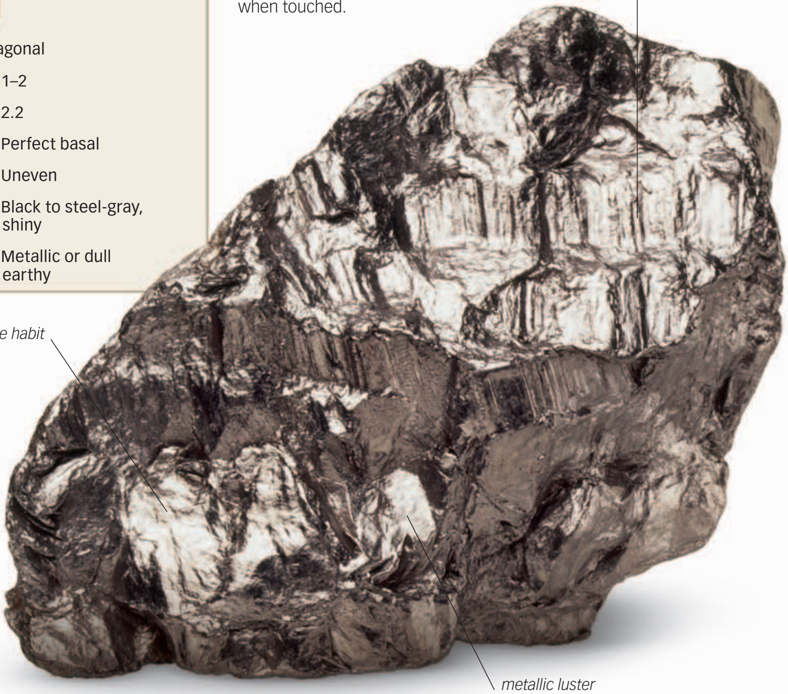

Massive graphite As seen in this massive specimen, graphite has a soapy or greasy feeling when touched.

VARIANTS

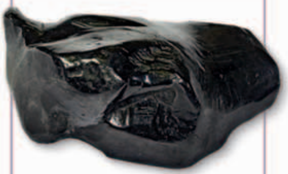

Black graphite A lump of compact, black graphite

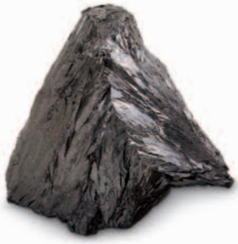

Crystalline graphite A graphite crystal exhibiting metallic luster

C

GRAPHITE

Like diamond, graphite is a form of native carbon. It takes its name from the Greek term *graphein*, which means “to write”—a reference to the black mark it leaves on paper. Graphite is opaque and dark gray to black. It occurs as hexagonal crystals, flexible sheets, scales, or large masses. It may be earthy, granular, or compact.

Graphite forms from the metamorphism of carbonaceous sediments and the reaction of carbon compounds with hydrothermal solutions. Graphite looks dramatically different from diamond and is at the other end of the hardness scale. Graphite’s softness is due to the way carbon atoms are bonded to each other—rings of six carbon atoms are arranged in widely spaced horizontal sheets. The atoms are strongly bonded within the rings but very weakly bonded between the sheets.

Graphite pencil The familiar pencil “lead” contains graphite. The first use of graphite pencils was described in 1575.

PROFILE

Cubic

10

3.4–3.5

Perfect octahedral

Conchoidal

Will scratch streak plate

Adamantine

Diamond in a matrix
An octahedral diamond crystal rests in the kimberlite matrix in which it was found.

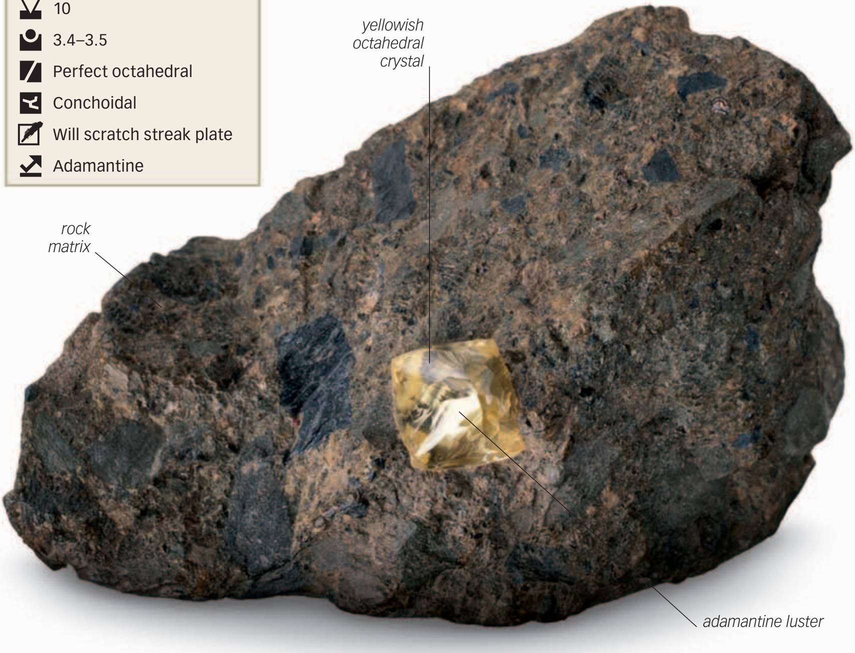

VARIANTS

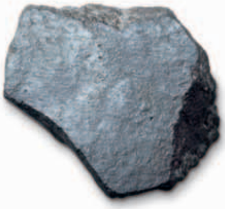

Carbonado A form of black industrial diamond

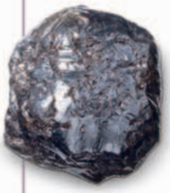

Bort diamond A crystal of black bort diamond

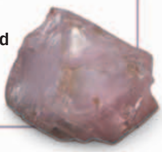

Pink diamond A rare pink diamond crystal

C

DIAMOND

The hardest known mineral, diamond is pure carbon. Its crystals typically occur as octahedrons and cubes with rounded edges and slightly convex faces. Crystals may be transparent, translucent, or opaque. They range from colorless to black, with brown and yellow being the most common colors. Other forms include bort or boart (irregular or granular black diamond) and carbonado (microcrystalline masses). Colorless gemstones are most often used in jewelry.

Most diamonds come from two rare volcanic rocks—lamproite and kimberlite (p.269). The diamonds crystallize in Earth's mantle, generally more than 95 miles (150 km) deep, and are formed up to Earth's surface through volcanism. Diamonds are also found in sediment deposited by rivers or melting glaciers.

Hope Diamond
Blue in color, the 45.5-carat Hope diamond is probably the world's most famous diamond.

SULFIDES

Sulfides are minerals in which sulfur (a nonmetal) is combined either with a metal or a semimetal. Some sulfides are brilliantly colored, and most of them have low hardness and high specific gravity. Sulfides are common and are found widely in nature.

Red River deposit
Sulfide deposits in the Rio Tinto area of southwestern Spain have been mined for silver, zinc, and copper since 800 BCE.

COMPOSITION

Most sulfides have simple atomic structures, in which sulfur atoms are stacked alternately with metal or semimetal atoms and arranged as cubes, octahedra, or tetrahedra. This yields highly symmetrical crystal forms. Except in a few sulfides, such as orpiment and realgar, the symmetrical form also gives rise to many of the properties also found in metals, including metallic luster and electrical conductivity.

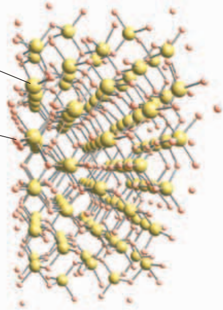

Crystal structure of bornite
Bornite is a compound of copper, iron, and sulfur, whose atoms link together to form tetrahedra.

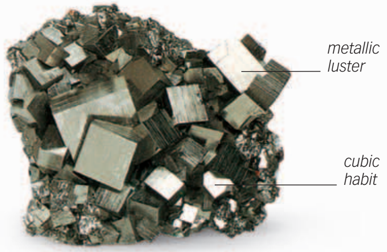

Crystalline pyrite
The iron sulfide pyrite, also called fool's gold, is one of the most common sulfides. The cubic crystals of pyrite reflect its simple atomic structure.

OCCURRENCE AND USES

Sulfides tend to form primarily in hydrothermal veins, from fluids circulating within fractures in Earth's crust. Sulfides such as pyrite and marcasite can form in sedimentary environments; others may form in magmas. It is common to find several sulfide minerals together.

Sulfides are the major ore minerals of many metals, including lead, zinc, iron, antimony, bismuth, molybdenum, nickel, silver, and copper—all of which have industrial uses. Gold is commonly found in sulfide deposits.

Die production
Sphalerite—a zinc sulfide—is the principal ore of zinc, which is used for die-cast components to galvanize, or coat, iron and steel.

PROFILE

Monoclinic
2–2½
7.2–7.4
Indistinct
Subconchoidal, sectile
Black
Metallic

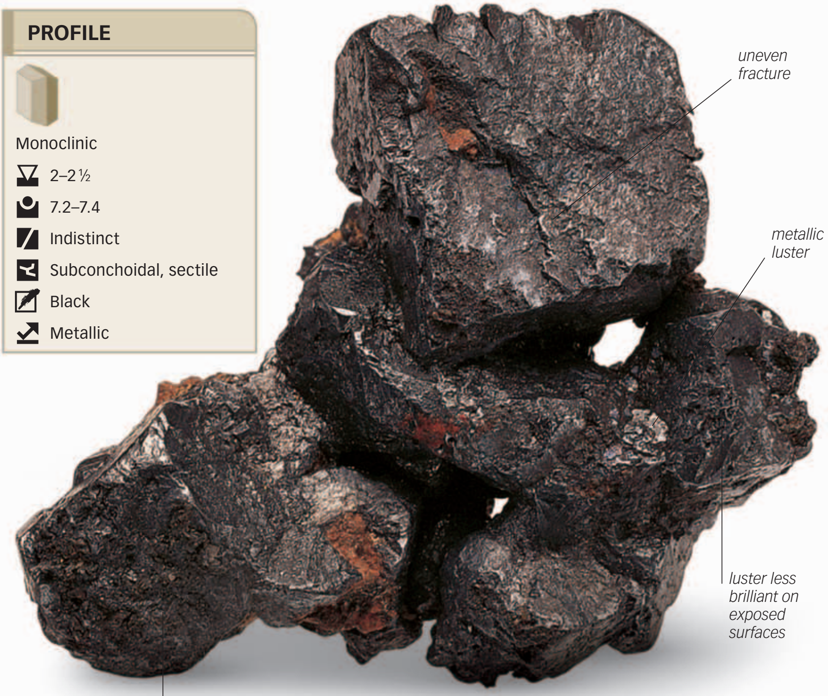
uneven fracture
metallic luster
luster less brilliant on exposed surfaces
darkened, weathered surface

Blocky, prismatic crystals In this pseudomorph specimen, argentite has replaced acanthite, while retaining the outward form of acanthite's cubic symmetry.

VARIANT

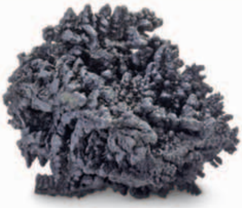

Thorny acanthite Dark, spiky acanthite crystals

Ag_2S

ACANTHITE

A silver sulfide, acanthite is the most important ore of silver. It takes its name from the Greek *akantha*, which means "thorn" and refers to the spiky appearance of some of its crystals. It also occurs in massive form and has an opaque, grayish black color. Above 350°F (177°C), silver sulfide crystallizes in the cubic system, and it used to be assumed that cubic silver sulfide—known as argentite—was a separate mineral from acanthite. It is now known that they are the same mineral, with acanthite crystallizing in the monoclinic system at temperatures below 350°F (177°C).

Acanthite forms in hydrothermal veins with other minerals, such as silver (p.43), galena (p.54), pyrargyrite (p.70), and proustite (p.72). It also forms as a secondary alteration product of primary silver sulfides. When heated, acanthite fuses readily and releases sulfurous fumes. The most famous locality of acanthite, the Comstock Lode in Nevada, USA, was so rich in silver that a branch of the US mint was established at nearby Carson City to coin its output.

PROFILE

Orthorhombic

- 3
- 5.1
- Poor
- Uneven to conchoidal, brittle
- Pale grayish black
- Metallic

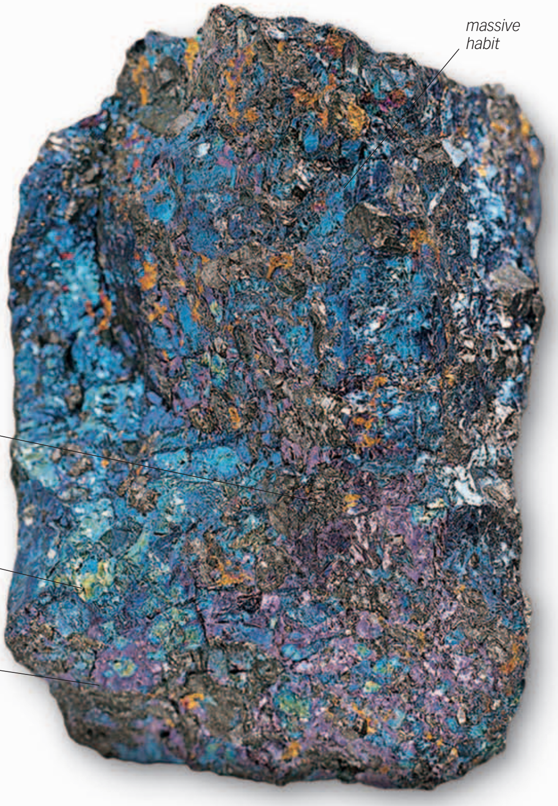

Massive bornite This specimen of tarnished bornite shows the oxidation colors that give it the names "purple copper ore" and "peacock ore."

VARIANT

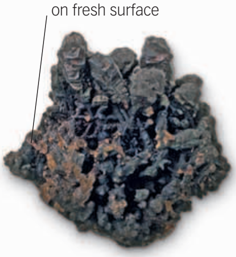

Bornite crystals Well-developed bornite crystals with curved faces

Cu_5FeS_4

BORNITE

One of nature's most colorful minerals, bornite is a copper iron sulfide named after the Austrian mineralogist Ignaz von Born (1742–91). A major ore of copper, its natural color can be coppery red, coppery brown, or bronze. It can also show iridescent purple, blue, and red splashes of color on broken, tarnished faces, which explains its common name, "peacock ore." Bornite is also known as "purple copper ore" and "variegated copper ore."

Bornite crystals are uncommon. Although they exhibit orthorhombic symmetry, crystals, when found, are cubic, octahedral, or dodecahedral, often with curved or rough faces. Bornite is frequently compact, granular, or massive and alters readily to chalcocite (p.51) and other copper minerals upon weathering. It forms mainly in hydrothermal copper ore deposits with minerals such as chalcopyrite (p.57), pyrite (p.62), marcasite (p.63), and quartz (p.168). It also forms in some silica-poor, intrusive igneous rocks and in pegmatite veins and contact metamorphic zones.

PROFILE

Monoclinic
2½–3
5.5–5.8
Indistinct
Conchoidal
Blackish lead gray
Metallic

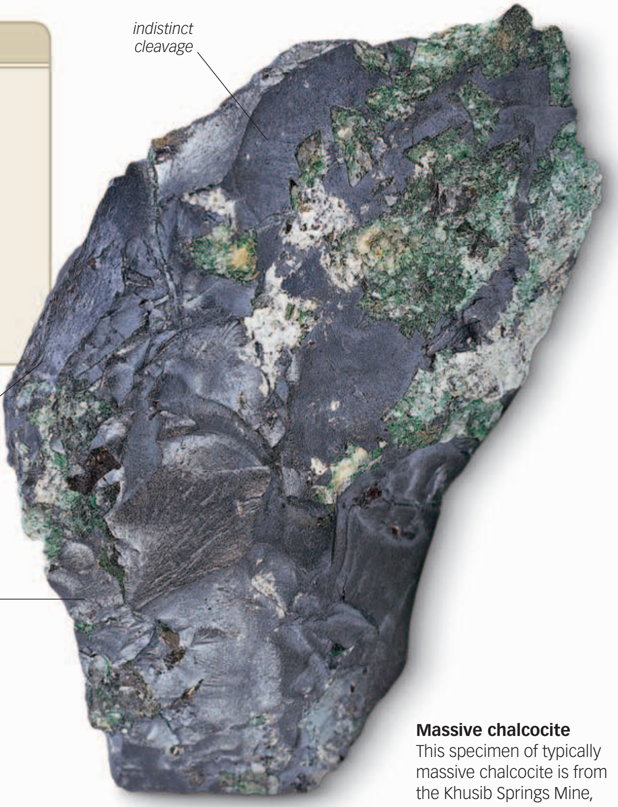

Massive chalcocite This specimen of typically massive chalcocite is from the Khusib Springs Mine, Otavi, Namibia.

VARIANT

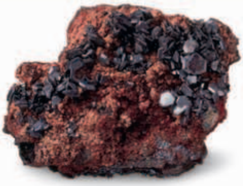

Prismatic crystals Short, prismatic chalcocite crystals on dolomite

Cu_2S

CHALCOCITE

The name chalcocite is derived from the Greek word for copper, *chalcos*. Chalcocite is one of the most important ores of copper. It is usually massive but, on rare occasions, occurs in short, striated prismatic or tabular crystals or as pseudohexagonal prisms formed by twinning. It is opaque, dark metallic gray, and becomes dull on exposure to light. Chalcocite was formerly known as chalcosine, copper glace, and redruthite, but these names are now obsolete.

Chalcocite forms at relatively low temperatures (up to 400°F/200C°), often as alteration products of other copper minerals such as bornite (p.50). It is found in hydrothermal veins and porphyry copper deposits with other minerals—bornite, covellite (p.52), sphalerite (p.53), galena (p.54), chalcopyrite (p.57), calcite (p.114), and quartz (p.168). Deposits in Cornwall, England, have been worked since the Bronze Age. Concentrated in secondary alteration zones, chalcocite can yield more copper than the element's primary deposits.

PROFILE

Hexagonal
- 1½–2
- 4.6–4.7
- Perfect basal
- Uneven
- Lead gray to black, shiny
- Submetallic to resinous

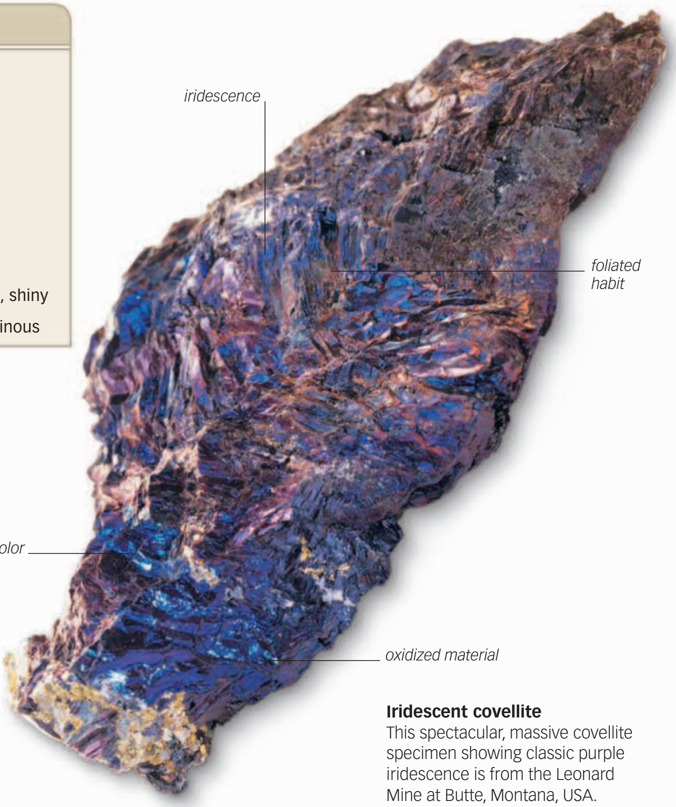

Iridescent covellite This spectacular, massive covellite specimen showing classic purple iridescence is from the Leonard Mine at Butte, Montana, USA.

VARIANT

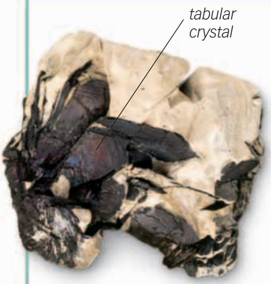

Tabular covellite Rare covellite crystals in their tabular habit

CuS

COVELLITE

Named in 1832 after the Italian minerologist Niccolo Covelli, who first described it, covellite is a copper sulfide. A minor ore of copper (p.37), covellite is opaque, with a bright metallic blue or indigo color. It is easy to recognize because of its brassy yellow, deep red, or purple iridescence. Covellite is generally massive and foliated in habit, although sometimes spheroidal. In crystalline form, it occurs as thin, tabular, and hexagonal plates, which are flexible when thin enough. Plates formed from its perfect basal cleavage are likewise flexible. It fuses very easily when heated, emitting a blue flame.

Covellite is a primary mineral in some places, but it typically occurs as an alteration product of other copper sulfide minerals such as bornite (p.50), chalcocite (p.51), and chalcopyrite (p.57). It sometimes forms as a coating on other copper sulfides. It rarely occurs as a volcanic sublimate, as on Mt. Vesuvius, where Niccolo Covelli first collected it. Covellite is abundant in the massive copper mines in Arizona, USA.

PROFILE

Cubic

3½–4

3.9–4.1

Perfect in six directions

Conchoidal

Brownish to light yellow

Resinous to adamantine, metallic

Sphalerite crystals These superbly formed sphalerite crystals occur with well-crystallized pyrite and quartz. They are from Casapalca, Lima, Peru.

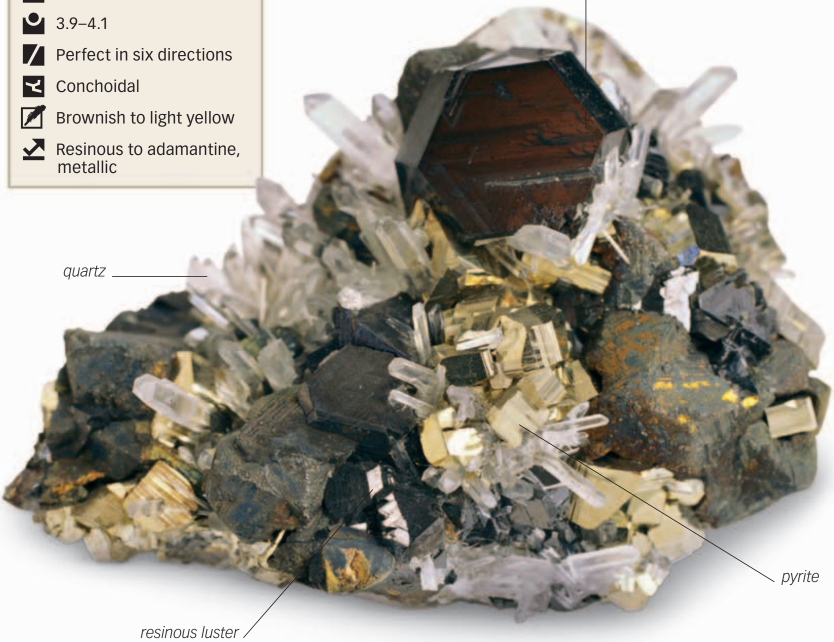

VARIANT

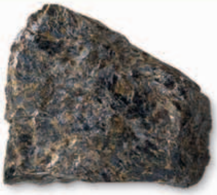

Massive sphalerite The most common habit of sphalerite

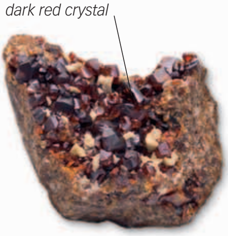

Ruby blende Brilliant red crystals of ruby blende sphalerite

ZnS

SPHALERITE

Sphalerite is the principal ore of zinc. Pure sphalerite is colorless and rare. Normally, iron is present, causing the color to vary from pale greenish yellow to brown and black with increasing iron content. Its complex crystals combine tetrahedral or dodecahedral forms with other faces. Sphalerite gets its name from the Greek *sphaleros*, meaning "deceitful," because its lustrous dark crystals can be mistaken for other minerals. It is often coarsely crystalline or massive, or forms banded, botryoidal, or stalactitic aggregates.

Sphalerite is found associated with galena (p.54) in lead-zinc deposits. It occurs in hydrothermal vein deposits, contact metamorphic zones, and replacement deposits formed at high temperature (1,065°F/575°C or above). It is also found in meteorites and lunar rocks.

Oval cut This oval cut shows off the golden brown color of sphalerite. Such stones are cut for collectors.

PROFILE

Cubic

2½

7.6

Perfect

Subconchoidal

Lead-gray

Metallic

Galena crystals Galena is usually found in cube-shaped crystals, but the crystal shape can also incorporate the faces of octahedra, as here.

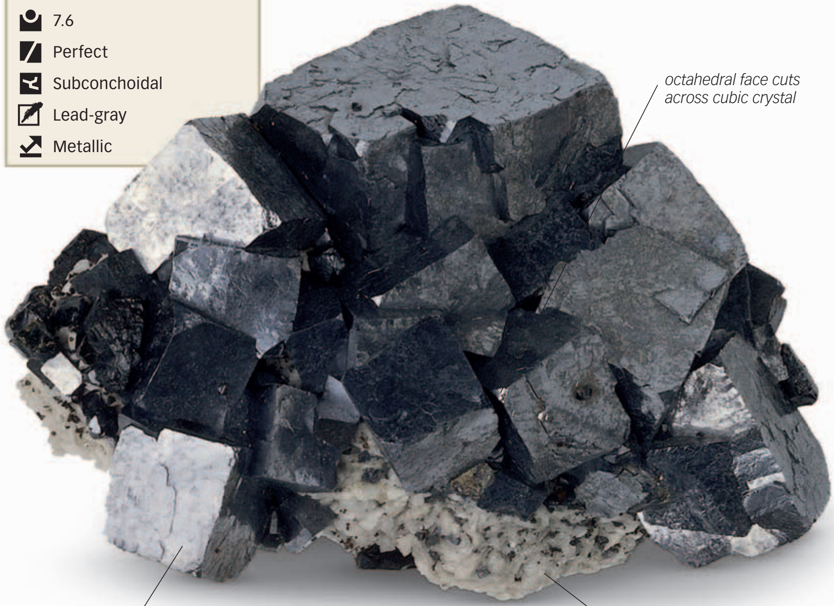

octahedral face cuts across cubic crystal

metallic luster

accessory dolomite

VARIANT

Perfect cleavage Cubic galena crystals with perfect cleavage in three directions

PbS

GALENA

There are more than 60 known minerals that contain lead, but by far the most important lead ore is galena, or lead sulfide. It is possible that galena was the first ore to be smelted to release its metal—lead beads found in Turkey have been dated to around 6500 BCE. Galena is opaque and bright metallic gray when fresh, but it dulls on exposure to the atmosphere. Its crystals are cubic, octahedral, dodecahedral, or combinations of these forms. Irregular, coarse, or fine crystalline masses are common.

Galena is common in hydrothermal lead, zinc, and copper (p.37) ore deposits worldwide and is often associated with sphalerite (p.53), chalcopyrite (p.57), and pyrite (p.62). It is also found in contact metamorphic rocks. Galena weathers easily to form secondary lead minerals, such as cerussite (p.119), anglesite (p.132), and pyromorphite (p.151). Galena is both the principal ore of lead and the main source of silver (p.43)—it often contains a considerable amount of silver in the form of acanthite as an impurity. It can also be a source of other metals.

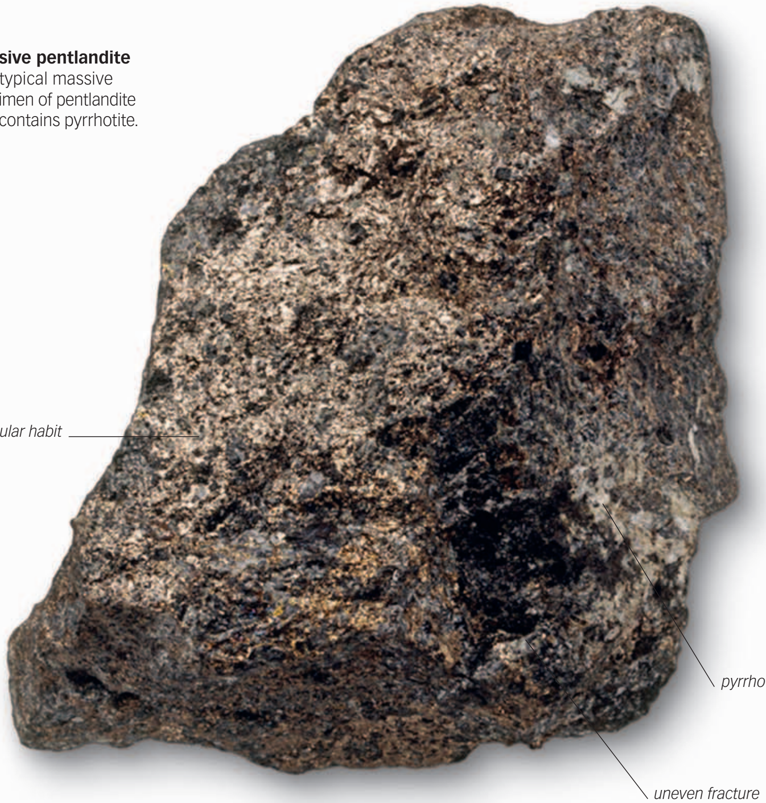

Massive pentlandite
This typical massive specimen of pentlandite also contains pyrrhotite.

PROFILE

Cubic
3½–4
4.6–5.0
None
Conchoidal
Bronze-brown
Metallic

$(Fe,Ni)_9S_8$

PENTLANDITE

Named in 1856 after the Irish scientist Joseph Pentland, its discoverer, pentlandite is a nickel and iron sulfide. Nickel is usually a smaller component than iron, but both may be present in equal parts. Pentlandite mainly has a massive or granular habit, and its crystals cannot be seen by the naked eye. It is opaque, metallic yellow in color, and has a bronzelike tarnish.

Pentlandite occurs in silica-poor, intrusive igneous rocks. It is almost always accompanied by pyrrhotite, with other sulfides such as chalcopyrite (p.57) and pyrite (p.62), and with some arsenides. The chief ore of nickel, pentlandite is relatively widespread, but commercial deposits are scarce. In Ontario, Canada, nickel from an ancient meteorite is thought to have enriched the ore. Pentlandite is also found as an accessory mineral in some meteorites. Silver (p.43) can be present in the pentlandite structure, yielding the mineral argentopentlandite; when cobalt replaces the iron and nickel, the mineral becomes cobaltpentlandite.

PROFILE

Hexagonal
2–2½
8.0
Perfect
Subconchoidal to uneven
Scarlet
Adamantine to dull

Crystalline cinnabar This massive specimen from Monte Amiata, Tuscany, Italy, also contains cinnabar crystals.

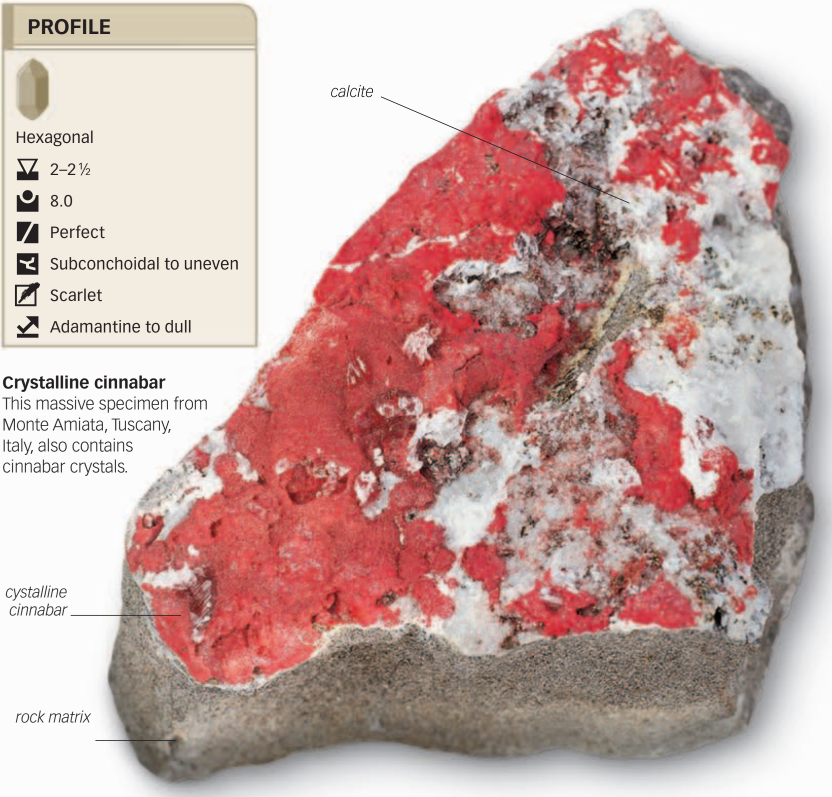

VARIANT

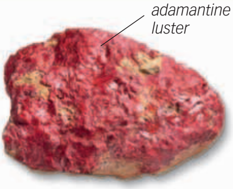

Massive cinnabar A specimen of massive cinnabar with a nonmetallic, adamantine luster

HgS

CINNABAR

A mercury sulfide, cinnabar takes its name from the Persian *zinjirfrah* and Arabic *zinjafr*, which mean "dragon's blood." It is bright scarlet to deep grayish red in color. It is the major source of mercury. Crystals are uncommon but when found they are rhombohedral, tabular, or prismatic. It usually occurs as massive or granular aggregates, but sometimes powdery coatings.

Cinnabar is often found with other minerals—such as stibnite (p.61), pyrite (p.62), and marcasite (p.63)—in veins near recent volcanic rocks. It is also found around hot springs. Cinnabar is believed to have been mined and used in Egypt in the early 2nd millennium BCE. It has also been mined for at least 2,000 years at Almadén, Spain. This site still yields excellent crystals.

Powdered cinnabar Since ancient times, artists have used bright red powdered cinnabar for the pigment vermillion.

PROFILE

Tetragonal

3½–4

4.2

Distinct

Uneven, brittle

Green-black

Metallic

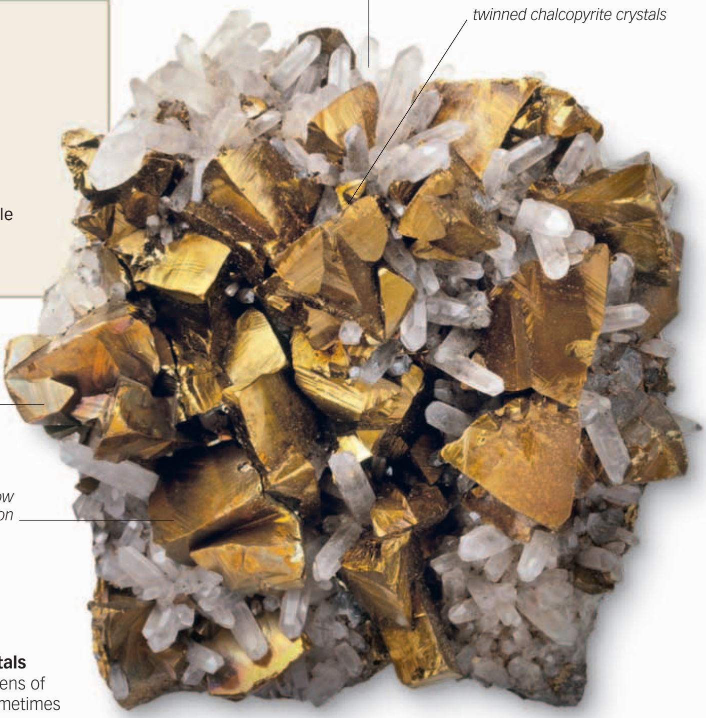

Chalcopyrite crystals Crystallized specimens of chalcopyrite can sometimes contain both twinned crystals and quartz crystals.

$CuFeS_2$

CHALCOPYRITE

One of the minerals worked at Rio Tinto, Spain, since Roman times, chalcopyrite is a copper and iron sulfide. It is opaque and brassy yellow when freshly mined, but it commonly develops an iridescent tarnish on exposure to the atmosphere. This tetragonal mineral forms tetrahedral crystals, which can be up to 4 in (10 cm) long on a face. It commonly occurs as massive aggregates and less frequently as botryoidal masses or as scattered grains in igneous rocks.

Chalcopyrite forms under a variety of conditions. It is mostly found in hydrothermal sulfide veins as a primary mineral deposited at medium and high temperatures (400°F/200°C or above), and as replacements, often with large concentrations of pyrite (p.62). It is also found as grains in igneous rocks and is an important ore mineral in porphyry copper deposits. Rarely, it occurs in metamorphic rocks. Chalcopyrite is an important ore of copper owing to its widespread occurrence. In some cases, selenium can replace a portion of the sulfur.

VARIANTS

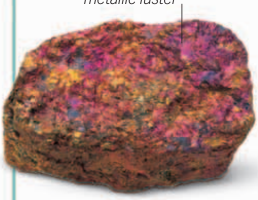

Massive chalcopyrite A specimen with an iridescent tarnish to it

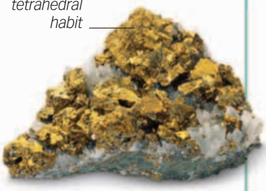

Brassy yellow chalcopyrite A specimen with an uneven fracture and tetrahedral habit

Realgar crystals
These bright red, prismatic realgar crystals are in a rock matrix and accompanied by gray quartz.

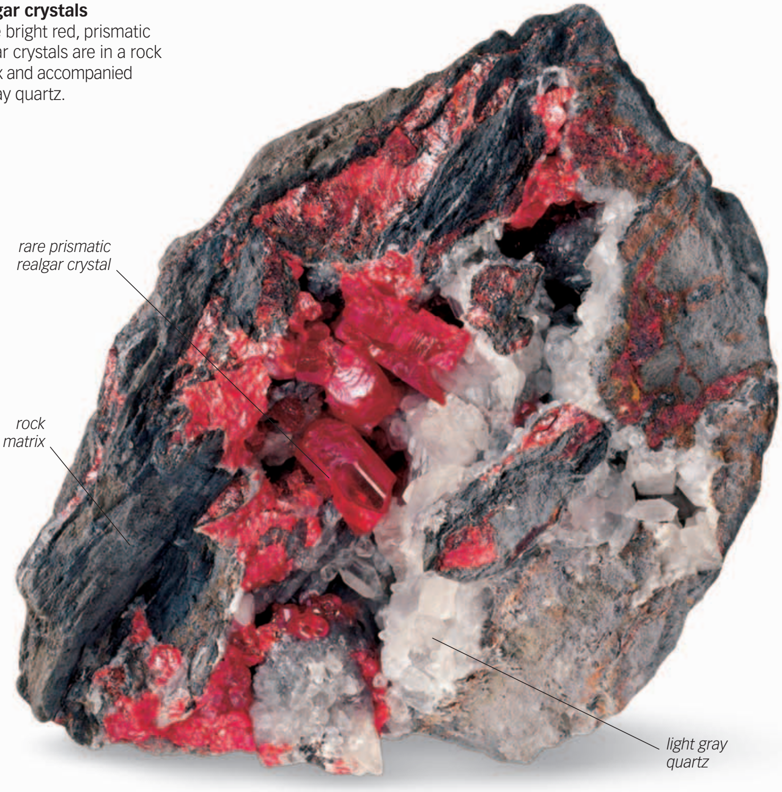

PROFILE

Monoclinic
1½–2
3.6
Good
Conchoidal
Scarlet to orange-yellow
Resinous to greasy

AsS

REALGAR

An important ore of arsenic, realgar is bright red or orange in color. Crystals are not often found, but when they occur they are short, prismatic, and striated. Realgar mostly occurs as coarse to fine granular masses and as encrustations. Realgar disintegrates on prolonged exposure to light, forming an opaque yellow powder, which is principally pararealgar. Therefore, specimens are kept in darkened containers.

Realgar is typically found in hydrothermal deposits at low temperature (up to 400°F/200°C) often with orpiment (p.59) and other arsenic minerals. It also forms as a sublimate around volcanoes, hot springs, and geyser deposits and as a weathering product of other arsenic-bearing minerals. Realgar is often found with stibnite (p.61) and calcite (p.114).

Powdered realgar
Scarlet to orange-yellow in color, powdered realgar was once used as a pigment and in fireworks.

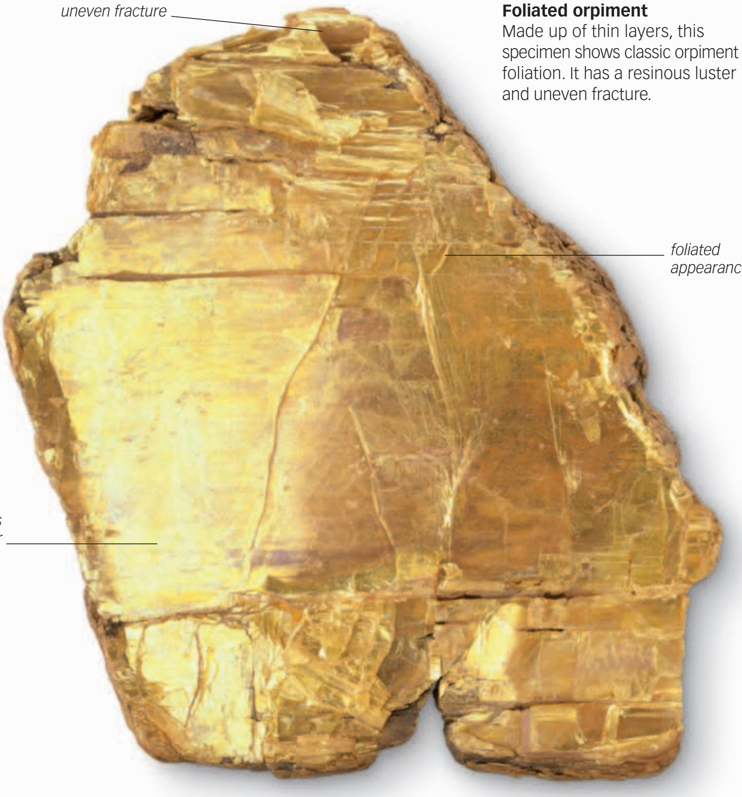

Foliated orpiment Made up of thin layers, this specimen shows classic orpiment foliation. It has a resinous luster and uneven fracture.

PROFILE

Monoclinic
- 1½–2
- 3.5
- Perfect
- Uneven, sectile
- Pale yellow
- Resinous

VARIANT

Crystalline orpiment Rare, stubby, prismatic crystals

As_2S_3

ORPIMENT

An arsenic sulfide, orpiment is a soft yellow or orange mineral. Widely distributed, it is typically powdery or massive, but it is also found as cleavable, columnar, or foliated masses. Distinct crystals are uncommon, but when found they are short prisms. Orpiment occurs in hydrothermal veins at low temperature (up to 400°F/ 200°C), hot spring deposits, and volcanic fumaroles, and it may occur with stibnite (p.61) and realgar (p.58). It also results from the alteration of other arsenic-bearing minerals.

When heated, orpiment gives off the garlic odor typical of arsenic minerals. The luster is resinous on freshly broken surfaces but pearly on cleavage surfaces. It was used as a pigment, mainly in ancient times in the Middle East. It was also used later in the West but soon replaced due to its toxicity.

Yellow pigment Powdered orpiment was used as a yellow pigment, especially to make gold-colored paint.

PROFILE

Hexagonal

- ⧖ 3–3½
- ● 5.5
- ◪ Perfect
- ▨ Uneven, brittle
- ✎ Greenish black
- ✓ Metallic

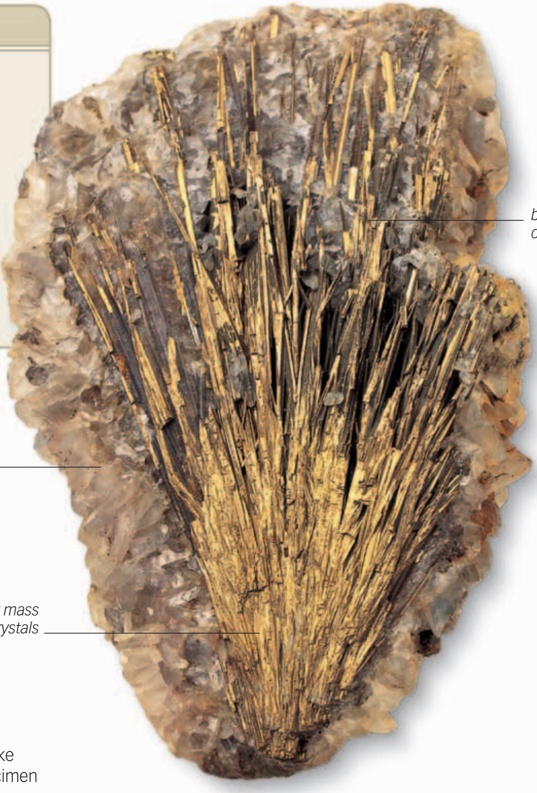

Millerite needles
The metallic yellow, needlelike crystals of this millerite specimen form a radiating spray.

VARIANT

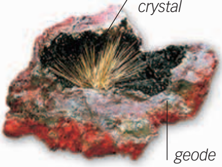

Millerite geode Thin, radiating crystals of millerite have formed in this hollow space

NiS

MILLERITE

A nickel sulfide, millerite commonly occurs as delicate, needlelike, opaque golden crystals. It can form free-standing, single crystals or occur as tufts, matted groups, or radiating sprays. It is also massive and frequently found with an iridescent tarnish. Nickel is more abundant in Earth's crust than copper (p.37), but it is generally more dispersed. Millerite is an ore of the element nickel, which is used in corrosion-resistant metal alloys, especially in the copper-nickel coinage that has replaced silver. It was named in 1845 after the English mineralogist W.H. Miller, who was the first person to study it.

Millerite normally forms at low temperatures (up to 400°F/200°C). It is often found in cavities in limestone (p.319) or dolomite (p.320), in carbonate veins and other associated rocks, within coal (p.253) deposits, and in serpentinite (p.298). It can occur as a later-formed mineral in nickel sulfide deposits and as an alteration product of other nickel minerals. Millerite is also found in meteorites and as a sublimate on Mount Vesuvius, Italy.

PROFILE

Orthorhombic
2
4.6
Perfect
Subconchoidal
Lead-gray to steel-gray
Metallic

Stibnite crystals This group of long, prismatic, striated stibnite crystals is on a quartz and barite matrix.

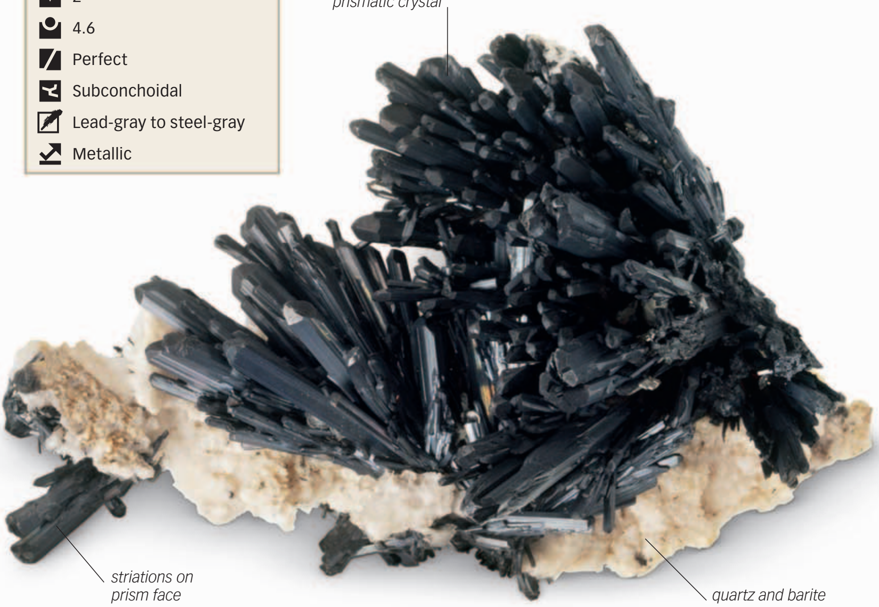

VARIANTS

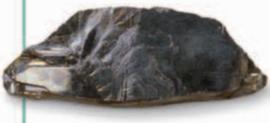
Stibnite sheets Thin layers of stibnite with sheetlike cleavage

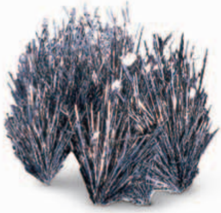
Acicular stibnite A mass of radiating, needlelike crystals

Sb_2S_3

STIBNITE

The principal ore of antimony, stibnite is antimony sulfide. Its name comes from the Latin *stibium*. Lead-gray to silvery gray in color, it often develops a black, iridescent tarnish on exposure to light. It normally occurs as elongated, prismatic crystals that may be bent or twisted. These crystals are often marked by striations parallel to the prism faces. Stibnite typically forms coarse, irregular masses or radiating sprays of needlelike crystals, but it can also be granular or massive.

A widespread mineral, stibnite occurs in hydrothermal veins, hot-spring deposits, and replacement deposits that form at low temperatures (up to 400°F/200°C). It is often associated with galena (p.54), cinnabar (p.56), realgar (p.58), orpiment (p.59), pyrite (p.62), and quartz (p.168). It is found in massive aggregates in granite (pp.258–59) and gneiss (p.288) rocks. Stibnite is used to manufacture matches, fireworks, and percussion caps for firearms. Powdered stibnite was used in the ancient world as a cosmetic for eyes to make them look larger.

PROFILE

Cubic

6–6½

5.0

None

Conchoidal

Greenish black to brownish black

Metallic

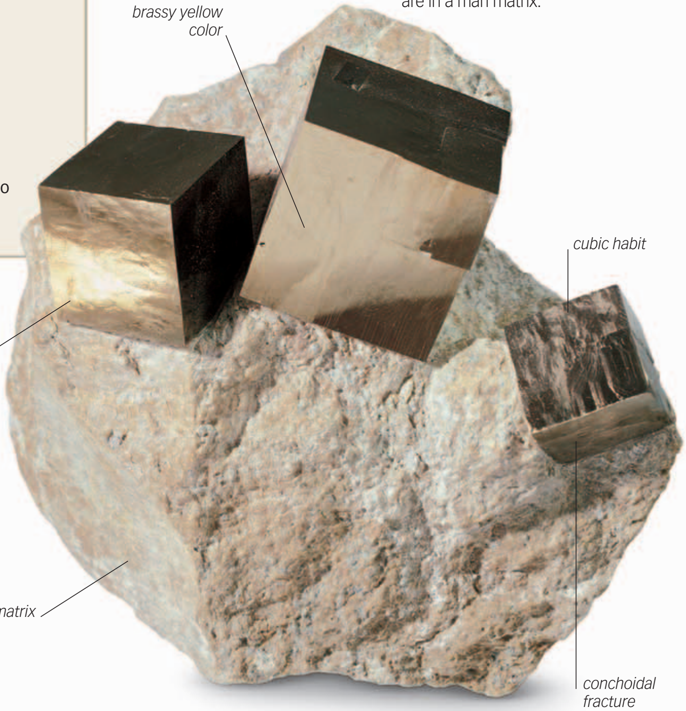

Cubic pyrite These three perfectly formed pyrite crystals—up to 1½in (3.5cm) wide—from Navajún, La Rioja, Spain, are in a marl matrix.

VARIANTS

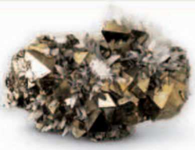

Octahedral pyrite A group of octahedral crystals with quartz

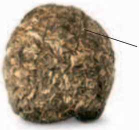

Pyrite nodule A ball-shaped, nodular group of pyrite crystals

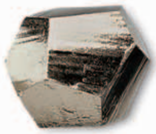

Pyritohedral pyrite A classic pyrite pyritohedral crystal

FeS_2

PYRITE

Known since antiquity, pyrite is commonly referred to as "fool's gold." Although much lighter than gold, its brassy color and relatively high density misled many novice prospectors. Its name is derived from the Greek word *pyr*, meaning "fire," because it emits sparks when struck by iron. It is opaque and pale silvery yellow when fresh, turning darker and tarnishing with exposure to oxygen. Pyrite crystals may be cubic, octahedral, or twelve-sided "pyritohedra," and are often striated. Pyrite can also be massive or granular, or form either flattened disks or nodules of radiating, elongate crystals.

Pyrite occurs in hydrothermal veins, by segregation from magmas, in contact metamorphic rocks, and in sedimentary rocks, such as shale (p.313) and coal (p.253), where it can either fill or replace fossils.

Pyrite beads With care, brittle pyrite can be ground into beads, such as those strung together in this necklace.

PROFILE

Orthorhombic

6–6½

4.9

Distinct

Uneven or irregular

Gray to black

Metallic

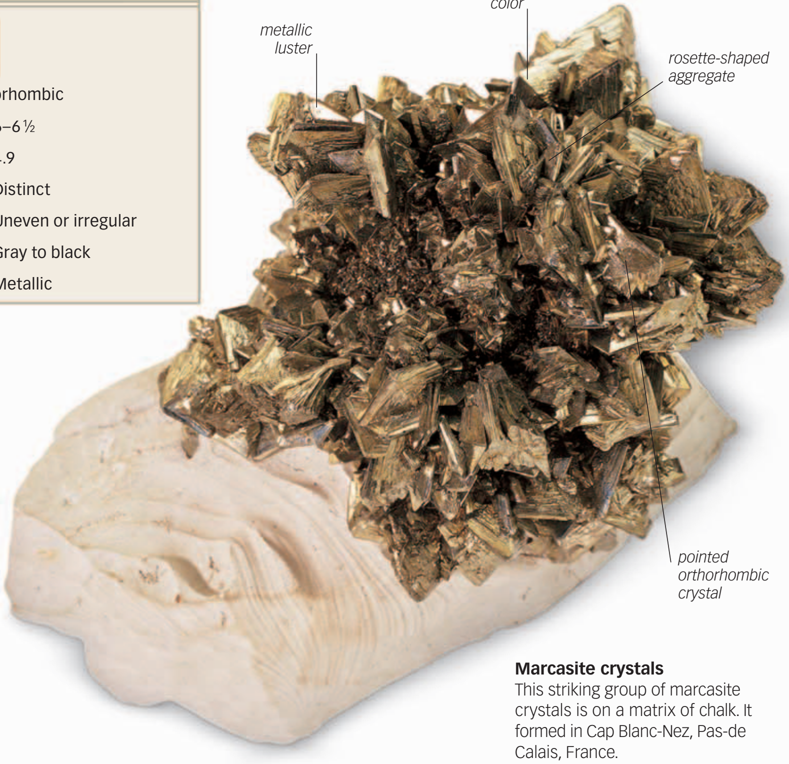
silvery yellow color
metallic luster
rosette-shaped aggregate
pointed orthorhombic crystal

Marcasite crystals
This striking group of marcasite crystals is on a matrix of chalk. It formed in Cap Blanc-Nez, Pas-de Calais, France.

VARIANT

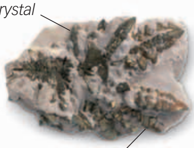
spear-shaped crystal
limestone matrix

Spear-shaped marcasite
Several groups of spear-shaped, twinned crystals

FeS_2

MARCASITE

An iron sulfide, marcasite is chemically identical to pyrite (p.62), but unlike pyrite it has an orthorhombic crystal structure. Marcasite is opaque and pale silvery yellow when fresh but darkens and tarnishes on exposure. It has a predominantly pyramidal or tabular crystal form. It is also found in characteristic twinned, curved, sheaflike shapes that resemble a cockscomb. Nodules with radially arranged fibers are common. Marcasite can also be massive, stalactitic, or reniform.

Marcasite is found near Earth's surface. It forms from acidic solutions percolating downward through beds of shale (p.313), clay, limestone (p.319), or chalk (p.321), where it often fills or replaces fossils. Marcasite also occurs as nodules in coal (p.253).

Art Deco jewelry
Marcasite was a popular choice for Victorian and Art Deco jewelry, although most of the material used was actually pyrite.

Layered masses
The crystallized molybdenite masses in this specimen show a typical layered structure.

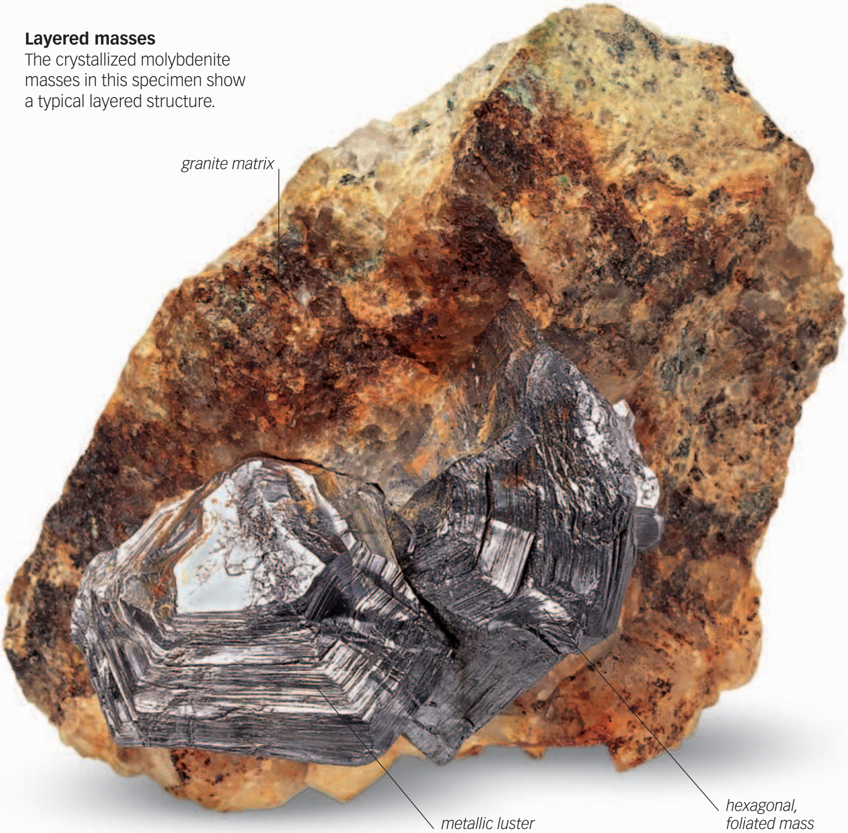

PROFILE

Hexagonal or trigonal

1–1½

4.7

Perfect basal

Uneven

Greenish or bluish gray

Metallic

MoS_2

MOLYBDENITE

A molybdenum sulfide, molybdenite is the most important source of molybdenum, which is an important element in high-strength steels. Molybdenite was originally thought to be lead, and its name is derived from the Greek word for lead, *molybdos*. It was recognized as a distinct mineral by the Swedish chemist Carl Scheele in 1778.

Molybdenite is soft, opaque, and bluish gray. It forms tabular hexagonal crystals, foliated masses, scales, and disseminated grains. It can also be massive or scaly. The platy, flexible, greasy-feeling hexagonal crystals of molybdenite can be confused with graphite (p.46), although molybdenite has a much higher specific gravity, a more metallic luster, and a slightly bluer tinge. Molybdenite occurs in granite (pp.258–59), pegmatite (p.260), and hydrothermal veins at high temperature (1,065°F/575°C or above) with other minerals—fluorite (p.109), ferberite (p.145), scheelite (p.146), and topaz (p.234). It is also found in porphyry ores and in contact metamorphic deposits.

SULFOSALTS

Sulfosalts are a group of mostly rare minerals that contain two or more metals in combination with sulfur (a nonmetal) and semimetals such as arsenic and antimony. Sulfosalt minerals have a high density, a metallic luster, and are usually brittle.

COMPOSITION

Sulfosalts have complex crystal structures. The structures of many sulfosalts appear to be based on fragments of simpler sulfur compounds. Metals commonly found in sulfosalts are lead, silver, thallium, copper, tin, bismuth, and germanium.

OCCURRENCE

Sulfosalts occur in small amounts in hydrothermal veins formed at low temperatures (up to 400°F/200°C). They are generally associated with the more common sulfides. A single Swiss deposit is known to have yielded up to 30 different sulfosalt minerals.

USES

Sulfosalts are typically found in small amounts but in a few deposits are economically important. Sometimes, they can constitute minor ores of silver, mercury, and antimony.

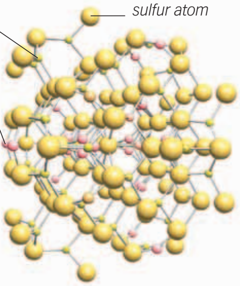

Crystal structure of tetrahedrite
The sulfosalt tetrahedrite is a complex compound of sulfur, copper, and antimony, and may contain silver, iron, zinc, and lead.

Silver coins
Pyragyrite, a sulfosalt mineral, yields the metal silver, which was used to produce the ancient coins seen here.

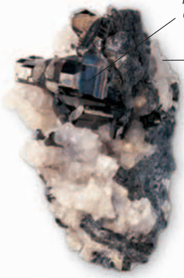

Bournonite
This specimen shows crystals typical of the sulfosalt bournonite. Twinned crystals growing parallel to each other give the mineral its informal name, cogwheel ore.

Sulfosalt deposit
Many sulfosalt minerals are found in Cornwall, England. These stone buildings once housed steam engines to pump water from mine shafts.

PROFILE

Cubic

- 4
- 4.6–4.7
- None
- Subconchoidal to uneven, brittle
- Black
- Metallic

Crystalline tennantite This mass of tetrahedral tennantite crystals is set in a rock matrix and has an iridescent tarnish.

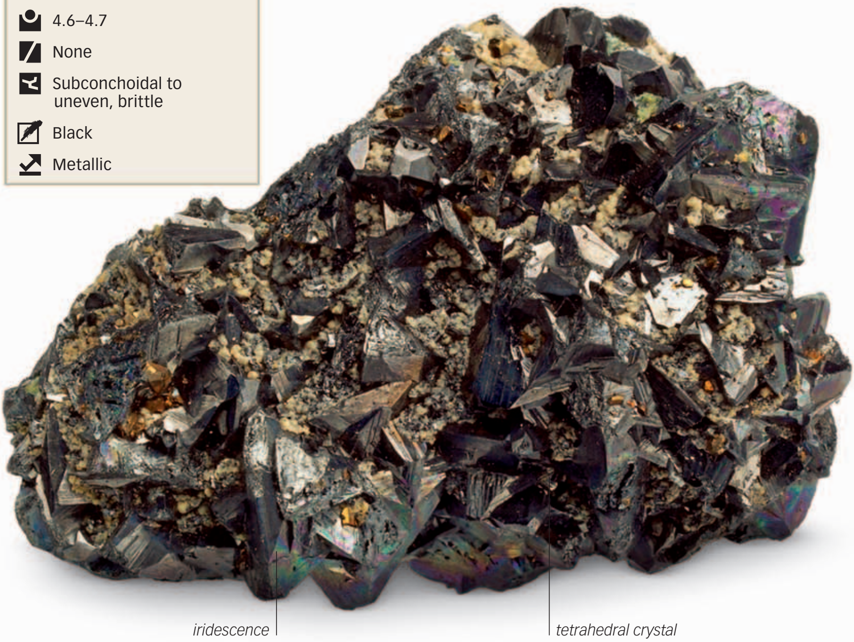

VARIANT

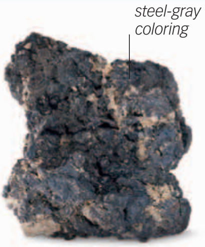

Massive tennantite A specimen of massive tennantite from Cornwall, England

$(Cu,Fe)_{12}As_4S_3$

TENNANTITE

Named in 1819 after the English chemist Smithson Tennant, tennantite is a copper iron arsenic sulfide. Iron, zinc, mercury, bismuth, and silver may substitute for up to 15 percent of the copper in tennantite. Tennantite is gray-black, steel-gray, iron-gray, or black in color. It forms cubic and tetrahedral crystals. It may also occur in massive, granular, and compact forms. Tennantite is an end member of a solid-solution series with the similar mineral tetrahedrite (p.73). The two have very similar properties, making it difficult to distinguish between them. Their crystal habits are similar and both exhibit contact and penetration twinning.

Tennantite is found in hydrothermal and contact metamorphic deposits, often associated with sphalerite (p.53), galena (p.54), chalcopyrite (p.57), fluorite (p.109), barite (p.134), and quartz (p.168). Deposits are found in Freiberg, Saxony, Germany; Lengenbach, Switzerland; and Butte, Montana, and Aspen and Central City, Colorado, USA.

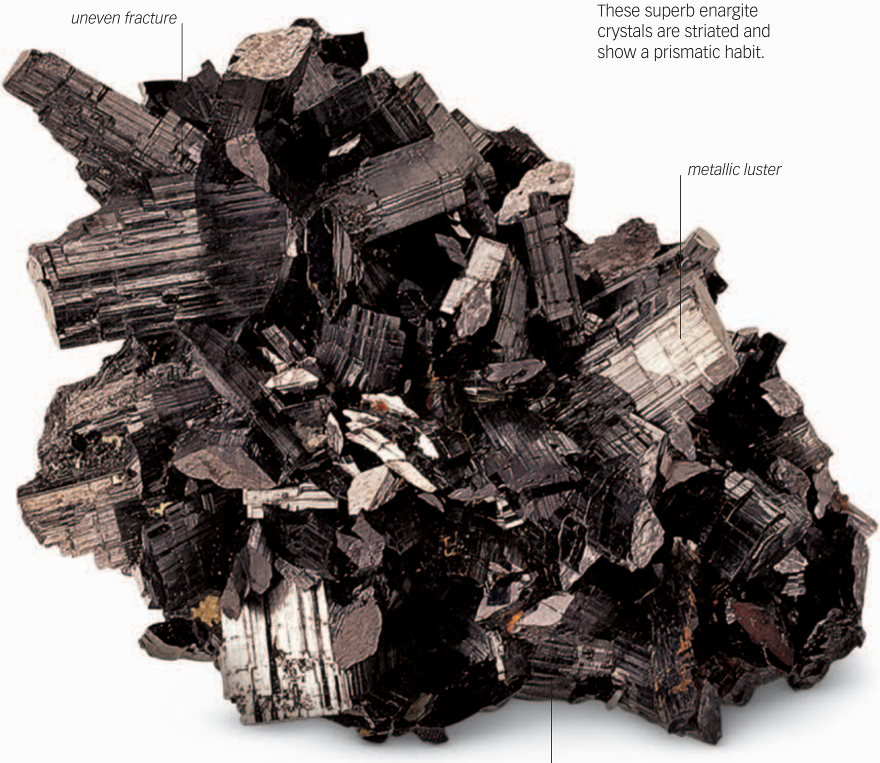

Enargite crystals
These superb enargite crystals are striated and show a prismatic habit.

PROFILE

Orthorhombic
3
4.4–4.5
Perfect
Uneven, brittle
Black
Metallic

Cu_3AsS_4

ENARGITE

A copper arsenic sulfide, enargite takes its name from the Greek word *enarge*, which means "distinct"—a reference to its perfect cleavage. An important ore of copper, it has a bright metallic luster, is opaque, and has a gray-black to iron-black to violet-black color when fresh. It turns dull black on exposure to light and pollutants. Enargite may occur in massive or granular habits. Crystals are usually small, either tabular or prismatic, sometimes pseudohexagonal or hemimorphic (with different terminations at each end), and have striations along the prism faces. Enargite crystals occasionally form star-shaped multiple twins.

Enargite forms in hydrothermal vein deposits at low to medium temperature (up to 1,065°F/575°C) and in replacement deposits, where it is associated with bornite (p.50), covellite (p.52), sphalerite (p.53), galena (p.54), chalcopyrite (p.57), pyrite (p.62), and other copper sulfides. It also occurs in the cap rocks of salt domes, with minerals such as anhydrite (p.133).

Fibrous habit
This jamesonite specimen, set in a rock matrix, has the fibrous habit typical of the mineral.

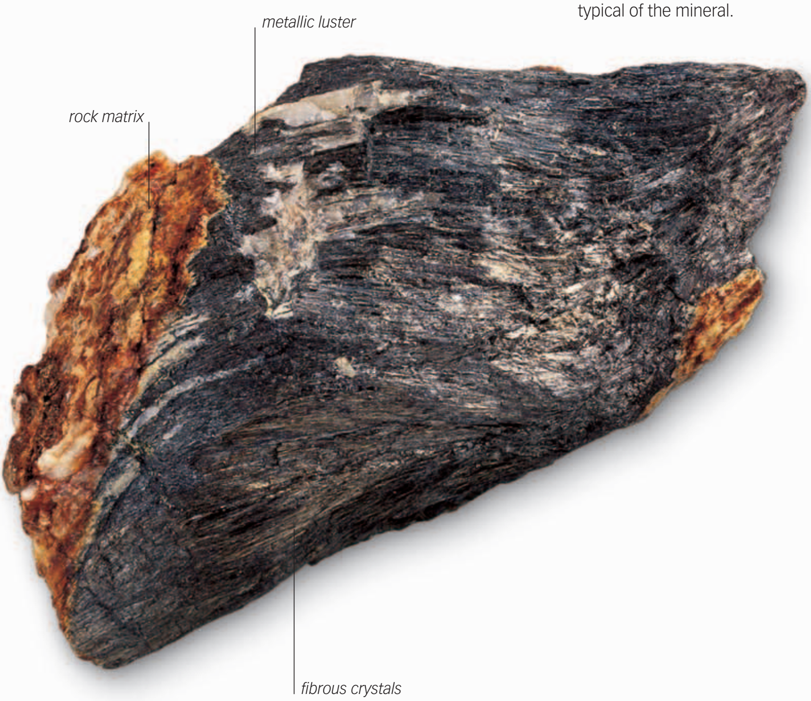

PROFILE

Monoclinic
2–3
5.5–6.0
Good
Uneven to conchoidal
Grayish black
Metallic

$Pb_4FeSb_6S_{14}$

JAMESONITE

Named in 1825 after the Scottish mineralogist Robert Jameson, jamesonite is a lead iron antimony sulfide. It is opaque lead-gray, but can often develop an iridescent tarnish. Jamesonite is normally found as needlelike or fibrous crystals combined together into columnar, radiating, plumose (featherlike), or feltlike masses.

Jamesonite occurs in hydrothermal veins at low or medium temperature (up to 1,065°F/575°C), where hot, chemical-rich fluids have permeated joints and fault lines, depositing minerals during cooling. In hydrothermal veins, it often occurs with other lead and antimony sulfides and sulfosalt minerals. Jamesonite also occurs in quartz associated with carbonate minerals, such as calcite (p.114), dolomite (p.117), and rhodochrosite (p.121). Jamesonite is a minor ore of antimony, which is used as a strengthening agent in alloys. It is widespread in small amounts, with good specimens coming from Freiburg, Saxony, Germany; Yakutia, Russia; Trepca, Serbia; Dachang, China; Cornwall, England; and Oruro, Bolivia.

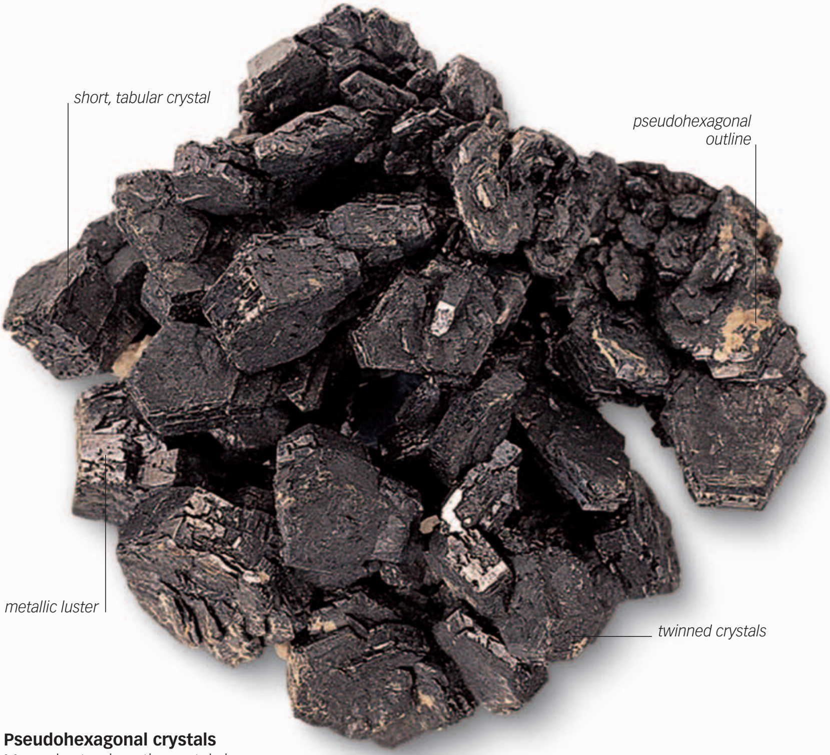

Pseudohexagonal crystals
Many short, prismatic crystals in this stephanite specimen show pseudohexagonal twinning.

PROFILE

Orthorhombic
2–2½
6.2–6.5
Imperfect
Subconchoidal to uneven, brittle
Iron-black
Metallic

Ag_5SbS_4

STEPHANITE

A silver antimony sulfide, stephanite was named in honor of Archduke Victor Stephan, the mining director of Austria, in 1845. It is sometimes called brittle or black silver ore. It is opaque, iron-black to black in color, and has a metallic luster on fresh faces. Stephanite crystals range from short prismatic to tabular and are repeatedly twinned to form pseudohexagonal groups. Stephanite may also occur in massive and granular habits.

Stephanite is generally found in small amounts in late-stage hydrothermal silver veins associated with native silver (p.43), sulfides, and other sulfosalts, such as acanthite (p.49) and tetrahedrite (p.73). It was found in sufficient quantity to be an ore of silver in Comstock Lode, Nevada, USA.

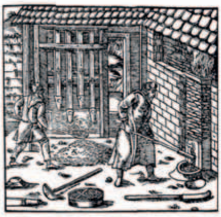

Historic silver processing
This 1550 woodcut from Georgius Agricola's treatise *De Re Metallica* shows silver ore being processed.

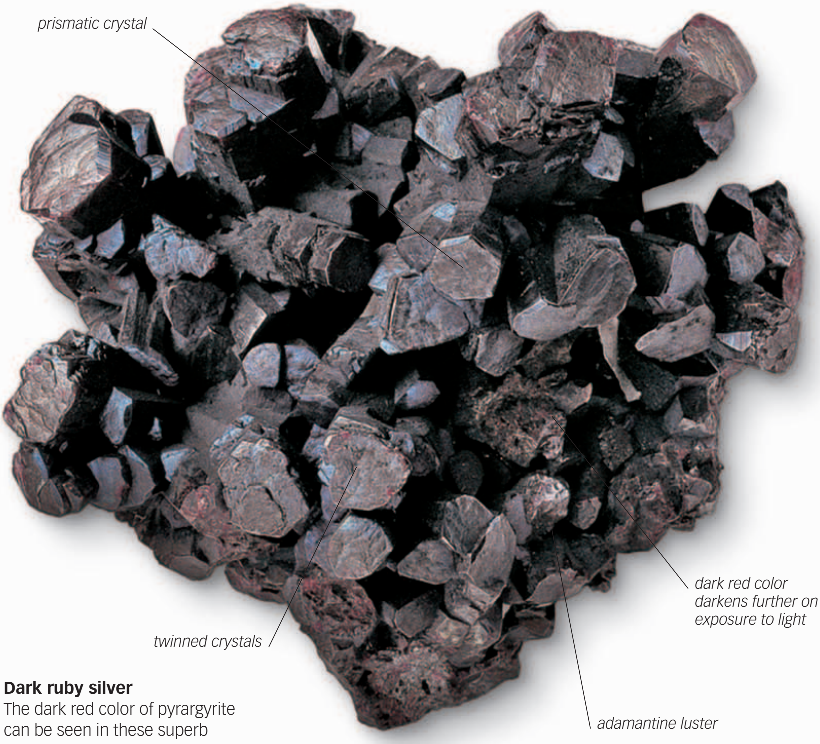

Dark ruby silver
The dark red color of pyrargyrite can be seen in these superb twinned, prismatic crystals.

PROFILE

Hexagonal

2½

5.8

Distinct

Conchoidal to uneven, brittle

Purplish red

Adamantine

Ag_3SbS_3

PYRARGYRITE

An important ore of silver, pyrargyrite takes its name from the Greek words *pyros*, which means “fire,” and *argent*, which means “silver”—an allusion to its silver content and its translucent, dark red color. Also known as dark ruby silver, pyrargyrite turns opaque dull gray when exposed to light. Therefore, prized specimens are stored in the dark. Pyrargyrite is typically massive or granular. It can also occur as well-formed prismatic crystals with rhombohedral, scalenohedral, or flat terminations, different at each end and frequently twinned.

Pyrargyrite forms in hydrothermal veins at relatively low temperature (up to 400°F/200°C) with the minerals sphalerite (p.53), galena (p.54), tetrahedrite (p.73), proustite (p.72), and calcite (p.114). It also forms by the alteration of other minerals.

Roman silver
This Roman *denarius* (silver coin) of the first century BCE shows gladiators fighting.

PROFILE

Orthorhombic

2½–3

5.8

Indistinct

Subconchoidal to uneven

Steel-gray

Metallic

Cogwheel ore This specimen of bournonite, accompanied by white baryte, shows the twinned habit that gives the mineral its informal name.

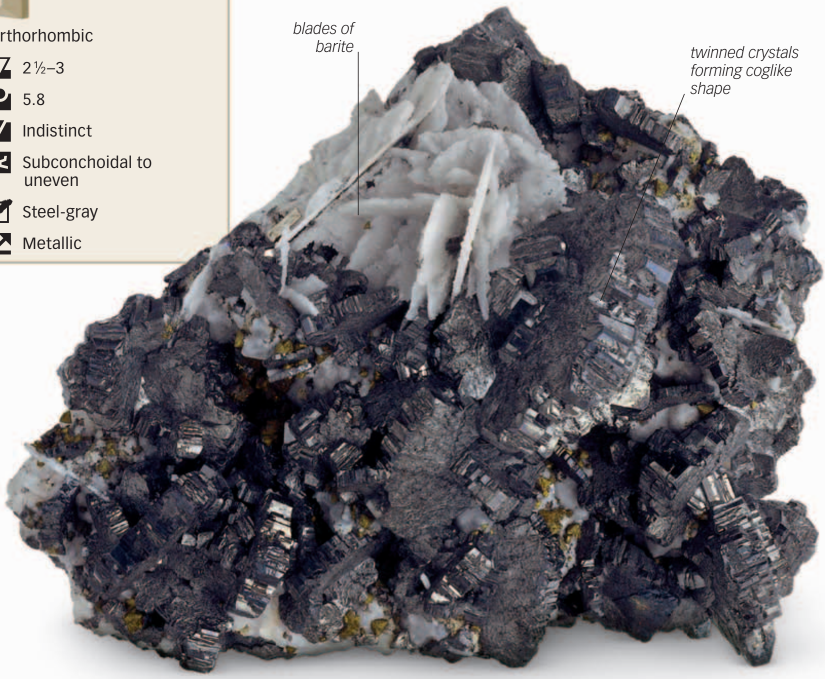

VARIANT

prismatic bournonite crystals
quartz matrix

Prismatic bournonite A group of twinned prismatic crystals of bournonite

$PbCuSbS_3$

BOURNONITE

A lead copper antimony sulfide, bournonite occurs as heavy, dark crystal aggregates and masses, as well as interpenetrating cruciform (crosslike) twins. When repeatedly twinned, bournonite has the appearance of a toothed wheel, giving rise to the informal name cogwheel ore. Untwinned crystals of this opaque mineral are tabular or short prismatic and usually have smooth and bright faces. Bournonite was first mentioned as a mineral in 1797 but was named only in 1805 after the French mineralogist Count J.L. de Bournon.

A widely distributed mineral, bournonite is found in hydrothermal veins at medium temperatures (400–1,065°F/200–575°C) and associated with sphalerite (p.53), galena (p.54), chalcopyrite (p.57), pyrite (p.62), tetrahedrite (p.73), and other sulfide minerals. Particularly prized specimens of bournonite come from the Harz Mountains of Germany, where a few crystals exceed ¾ in (2.2 cm) in diameter. This mineral has been used as a minor ore of antimony (p.41).

PROFILE

Hexagonal

2–2½

5.8

Distinct rhombohedral

Conchoidal to uneven, brittle

Vermilion

Adamantine to submetallic

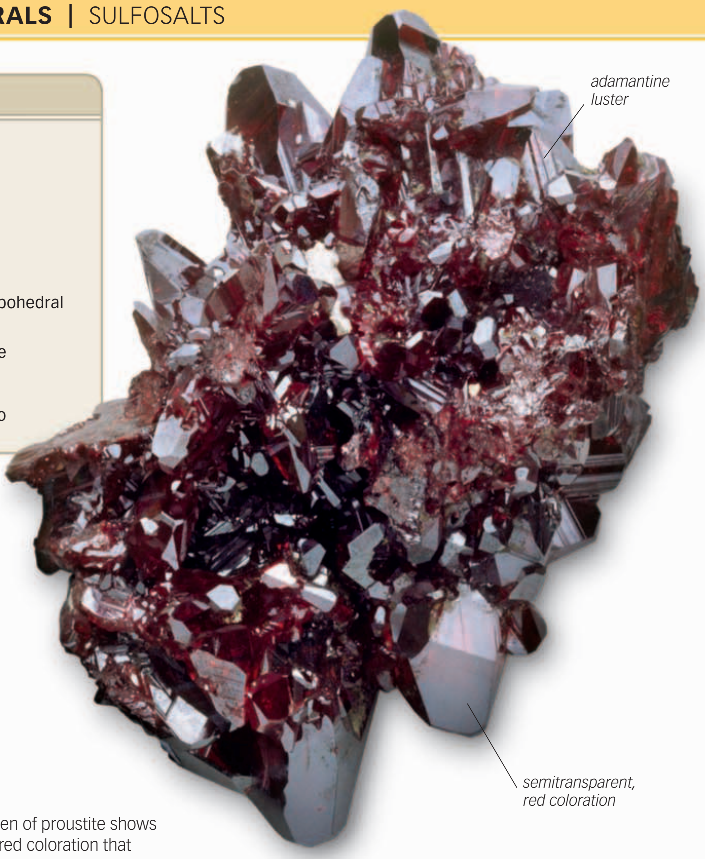

Ruby silver This stunning specimen of proustite shows the semitransparent red coloration that gives it the name common ruby silver.

Ag_3AsS_3

PROUSTITE

As its original name ruby silver ore suggests, proustite is translucent and red and is an important source of silver (p.43). It has also been called light red silver ore. The name proustite comes from the French chemist Joseph Proust, who distinguished it from the related mineral pyrargyrite (p.70) by chemical analysis in 1832. Its striated, often brilliant crystals are typically prismatic with rhombohedral or scalenohedral terminations, often resembling the dogtooth spar form of calcite (p.114) in habit. Proustite also occurs as massive or granular aggregates. The mineral turns from transparent scarlet to dull opaque gray in strong light, so specimens are stored in the dark.

Proustite forms in hydrothermal veins at low temperature (up to 400°F/200°C) with other silver minerals, such as acanthite (p.49), stephanite (p.69), and tetrahedrite (p.73), and with native arsenic (p.45), galena (p.54), and calcite (p.114). It also forms in the secondary zone of silver deposits. Large crystals come from Chañarcillo, Chile, and Freiburg, Saxony, Germany.

VARIANTS

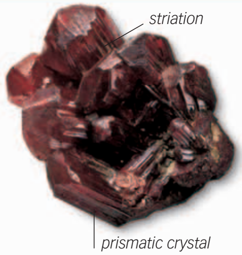

Striated proustite A prismatic, semitransparent crystal

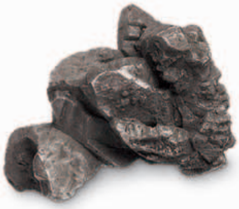

Dull proustite A dull, opaque specimen after exposure to light

Tetrahedral crystals
This group of relatively rare tetrahedrite crystals shows twinning and coats a rock matrix.

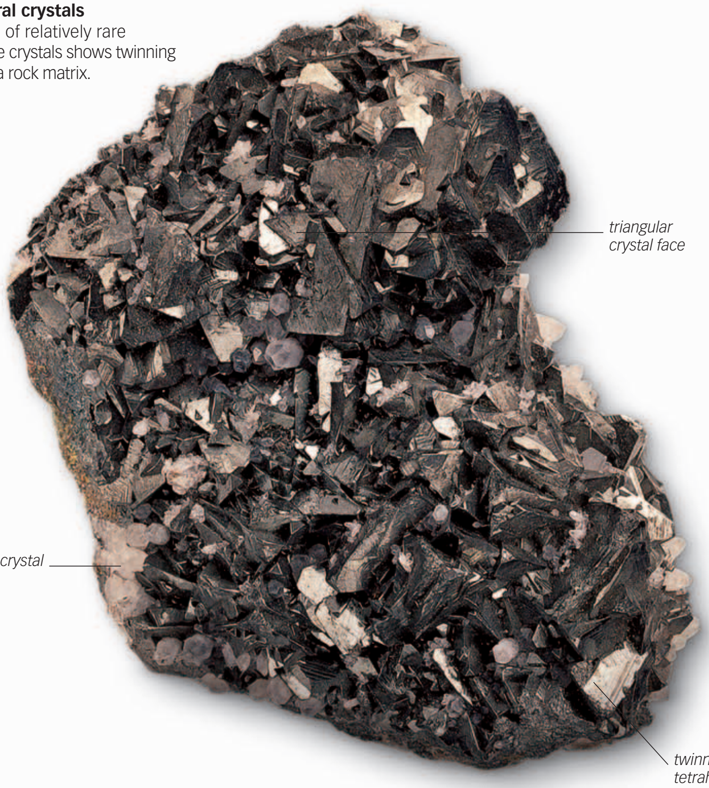

PROFILE

- Cubic
- 3–4
- 4.6–5.1
- None
- Subconchoidal to uneven
- Brown to black to cherry-red
- Metallic

$(Cu,Fe)_{12}Sb_4S_{13}$

TETRAHEDRITE

The name tetrahedrite comes from this mineral's characteristic tetrahedral crystals, although it also occurs as massive, compact, or granular aggregates. Tetrahedrite is opaque, metallic gray, or nearly black, and it sometimes coats or is coated with brassy yellow chalcopyrite (p.57). It forms a continuous solid-solution series with the similar mineral tennantite (p.66), in which arsenic replaces antimony in the crystal structure. Bismuth also substitutes for antimony and forms bismuthian tetrahedrite or annivite.

Tetrahedrite is an important ore of copper and sometimes silver. It forms in hydrothermal veins at low to medium temperatures (up to 1,065°F/ 575°C), often with bornite (p.50), galena (p.54), chalcopyrite, pyrite (p.62), barite (p.134), and quartz (p.168). It is also found in contact metamorphic deposits.

Copper ore
This 9th-century brass Arabic astrolabe is believed to have been made of copper extracted from tetrahedrite.

OXIDES

The minerals in this group have crystal structures in which metals or semimetals occupy spaces between oxygen atoms. The properties of oxides vary: the metallic ores and gemstone varieties tend to be hard and have a high specific gravity.

Queensland
The Queensland region of eastern Australia is a treasure trove of minerals. Several deposits of alluvial sapphire (aluminum oxide) are found here.

COMPOSITION

Oxides can be either simple or multiple. Simple oxides, such as cuprite (Cu_2O), contain only one metal or semimetal and oxygen. Multiple oxides have two different metal sites, both of which may be occupied by several different metals or semimetals. The minerals in the spinel ($MgAl_2O_4$) group are examples of multiple oxides.

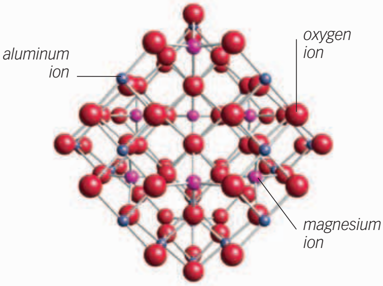

Spinel crystal structure
In spinel, magnesium and aluminum combine with oxygen. Other metals can replace magnesium and aluminum to form the spinel series of minerals.

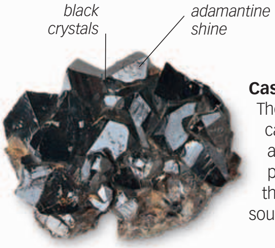

Cassiterite crystals
The simple tin oxide cassiterite, seen here as a group of twinned prismatic crystals, is the world's primary source of tin.

OCCURRENCE AND USES

Oxides occur as accessory minerals in many igneous rocks, especially as early crystallizing minerals in ultrabasic rocks, in pegmatites, and as decomposition products of sulfide minerals. Many resist weathering and are found concentrated in placers.

Many oxide minerals are important ores of chromium, uranium, tantalum, zinc, tin, cerium, tungsten, manganese, copper, and titanium. Other oxides, such as quartz and corundum, are important gemstone minerals.

Chrome bumper
The large, chrome-plated front bumper of this classic American 1956 Chevrolet is a dramatic example of chromium derived from the oxide chromite.

PROFILE

Hexagonal

Varies

1.0

Perfect, difficult

Conchoidal, brittle

White

Vitreous

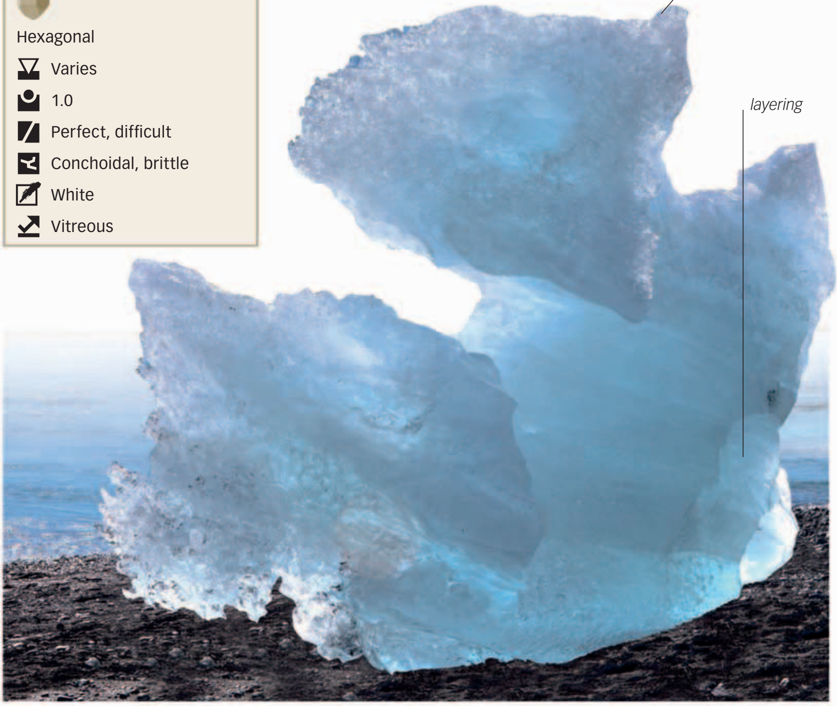

Iceberg This small, beached iceberg still shows some of its original depositional layering.

VARIANTS

Frost Crystalline ice in a frostlike form

Hailstone A huge (2 in x 3½ in / 5 cm x 9 cm) hailstone

H_2O

ICE

Although largely absent at lower latitudes, ice is probably the most abundant mineral exposed on Earth's surface. Liquid water is not classified as a mineral because it has no crystalline form. As snow, ice forms crystals that seldom exceed ¼ in (7 mm) in length, although as massive aggregates in glaciers, individual crystals may be up to 17½ in (45 cm) long. Other forms of ice include branching, treelike frost, skeletal, hopper-shaped, prismlike frost, and hailstones and icicles made up of many randomly oriented crystals.

Ice crystals are generally colorless, but the common white color of ice is due to gaseous inclusions of air that reflect light. There are at least nine polymorphs—different crystalline forms—of ice, each forming under different pressure and temperature conditions, but only one form exists at Earth's surface. The hardness of ice varies with its crystal structure, purity, and temperature. At temperatures found in the Arctic and high-alpine zones, ice is so hard it can erode stone when windblown.

PROFILE

Tetragonal

5½–6

3.9

Perfect

Subconchoidal

White to pale yellow

Adamantine to metallic

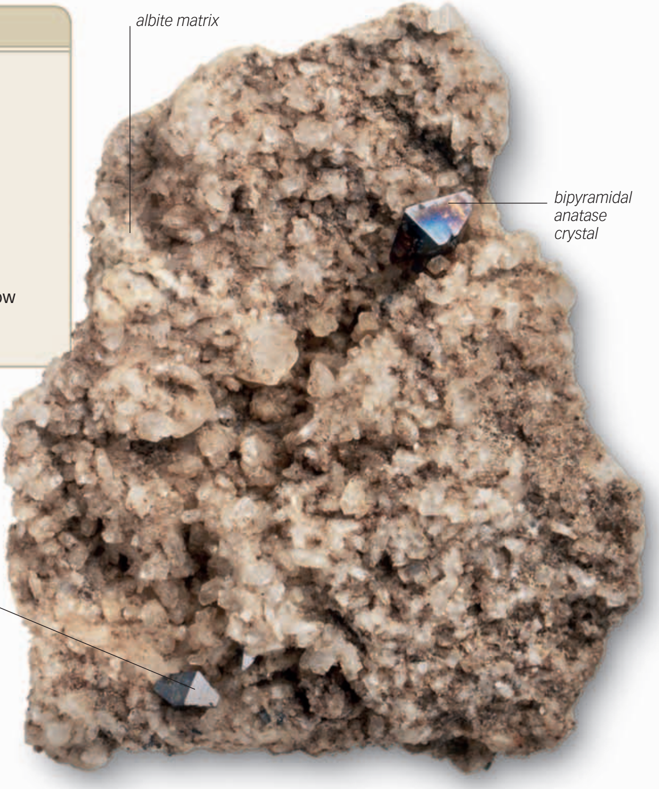

Bipyramidal anatase This specimen shows two bipyramidal anatase crystals perched on a matrix of albite crystals.

VARIANTS

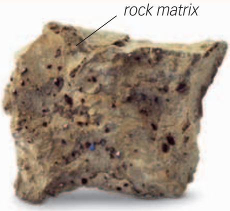

Black anatase Schist speckled with tiny black anatase crystals

Octahedral crystal A perfectly formed, modified bipyramidal anatase crystal

TiO_2

ANATASE

Formerly known as octahedrite, anatase is a polymorph of titanium dioxide. Its name comes from the Greek word *anatasis,* which means "extension"—a reference to the elongate octahedral crystals that are the most common habit of anatase. Anatase crystals can also be tabular and, rarely, prismatic. Hard and brilliant, the crystals can be brown, yellow, indigo-blue, green, gray, lilac, or black in color.

Anatase forms in veins and crevices in metamorphic rocks, such as schists (pp.291–92) and gneisses (p.288), and is derived from the leaching of surrounding rocks by hydrothermal solutions. Anatase also forms in pegmatites (p.260), often in association with the minerals brookite (p.77), ilmenite (p.90), fluorite (p.109), and aegirine (p.209). It is found in sediments and is sometimes concentrated in placer deposits. Much anatase is formed by the weathering of titanite (p.234). Weathered anatase becomes rutile (p.78). Although rutile replaces anatase, it retains the anatase crystal shape.

PROFILE

Orthorhombic
5½–6
4.1
Indistinct
Subconchoidal to uneven
White, grayish, yellowish
Metallic to adamantine

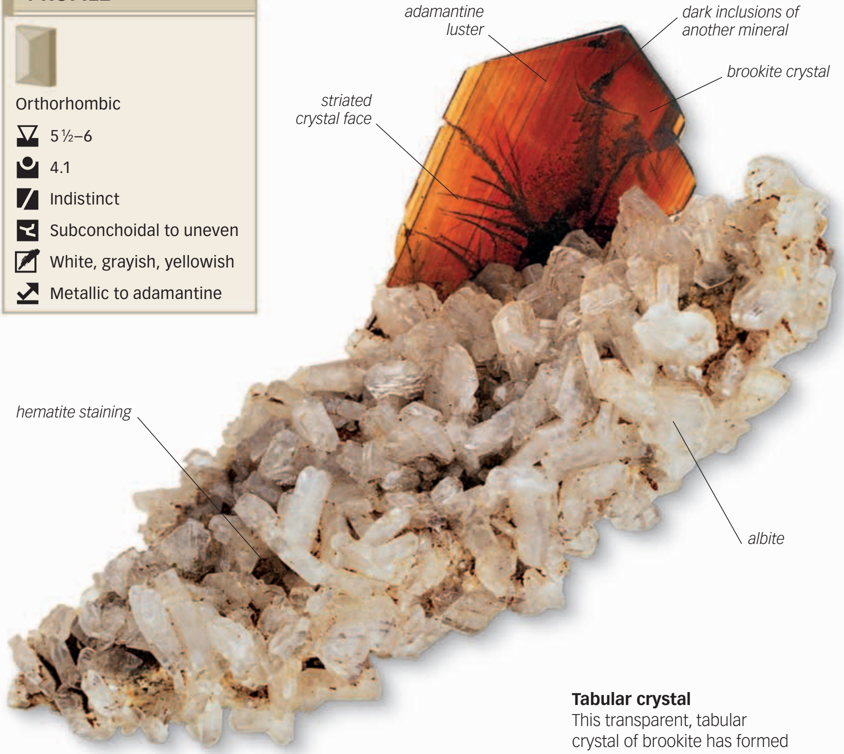

Tabular crystal
This transparent, tabular crystal of brookite has formed on a mass of albite crystals.

VARIANT

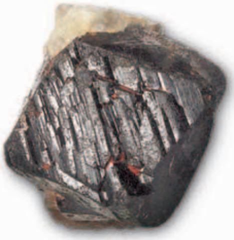

Dipyramidal crystal A black brookite specimen with metallic luster

TiO_2

BROOKITE

Named in 1825 after British crystallographer H.J. Brooke, brookite, like anatase (p.76) and rutile (p.78), is composed of titanium dioxide. However, unlike anatase and rutile, brookite exhibits orthorhombic symmetry. Usually brown and metallic, brookite may also be red, yellow-brown, or black. Crystals can be tabular or, less commonly, pyramidal or pseudohexagonal. They may be thin or thick and up to 2 in (5 cm) long. Iron is almost always present in this mineral's structure to a small degree, and brookite containing niobium is also known.

Brookite occurs in hydrothermal veins, in some contact metamorphic rocks, and as a detrital mineral in sedimentary deposits. Being relatively dense, it is common in areas with natural concentrations of heavy minerals, such as the diamond placer deposits of Brazil. It generally occurs with other minerals, including rutile, anatase, and albite (p.177). Brookite is widespread in mineral veins in the Alps. In the Fronolen locality in northern Wales, UK, it forms crystals on crevice walls in diabase rock.

PROFILE

Tetragonal

6–6½

4.2

Good

Conchoidal to uneven

Pale brown to yellowish

Adamantine to submetallic

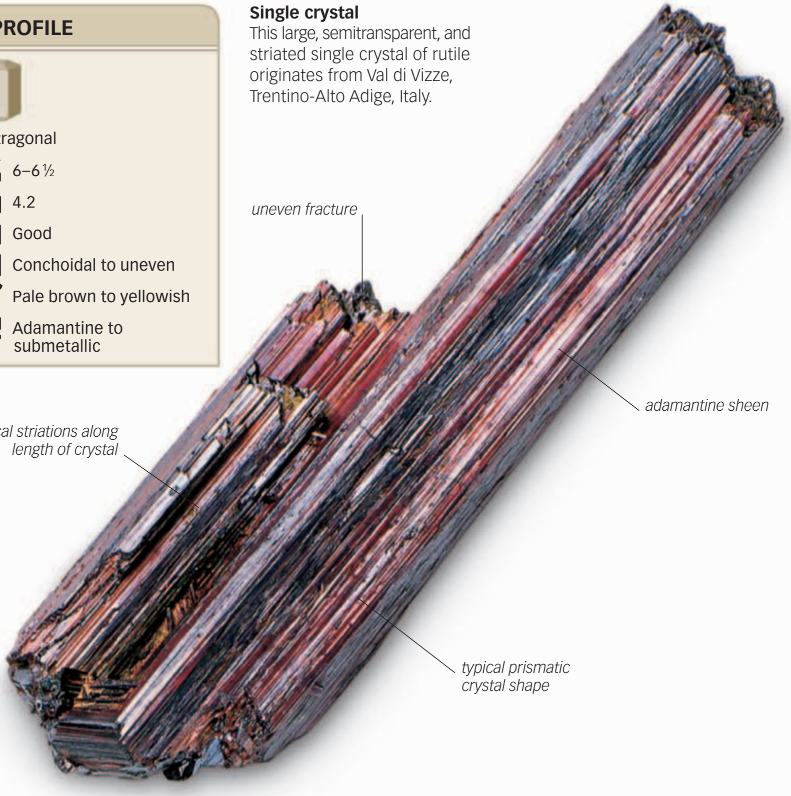
uneven fracture
adamantine sheen
vertical striations along length of crystal
typical prismatic crystal shape

Single crystal This large, semitransparent, and striated single crystal of rutile originates from Val di Vizze, Trentino-Alto Adige, Italy.

VARIANTS

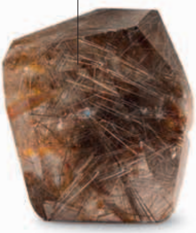
rutile needle

Rutilated quartz Pale-golden rutile crystals in polished quartz

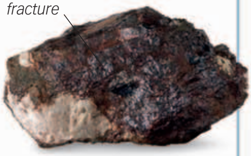
uneven fracture

Massive rutile Dark-hued crystals in rock matrix

TiO_2

RUTILE

A form of titanium oxide, rutile takes its name from the Latin *rutilis*, which means “red” or “glowing.” It often appears as pale golden, needlelike crystals inside quartz (p.168). When not enclosed in quartz, it is usually yellowish or reddish brown, dark brown, or black. Crystals are generally prismatic but can also be slender and needlelike. Multiple twinning is common and is either knee-shaped, net- or latticelike, or radiating, forming wheel-like twins. Rutile may also radiate in starlike sprays from hematite crystals.

Rutile often occurs as a minor constituent of granites (pp.258–59), gneisses (p.288), and schists (p.291), and also in hydrothermal veins and in some clastic sediments. It commonly forms microscopic, oriented inclusions in other minerals, producing an asterism effect.

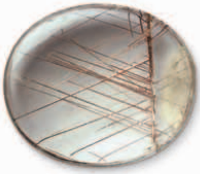

Quartz rutile cabochon Slender rutile crystals are clearly visible inside this polished, convex-cut, colorless quartz.

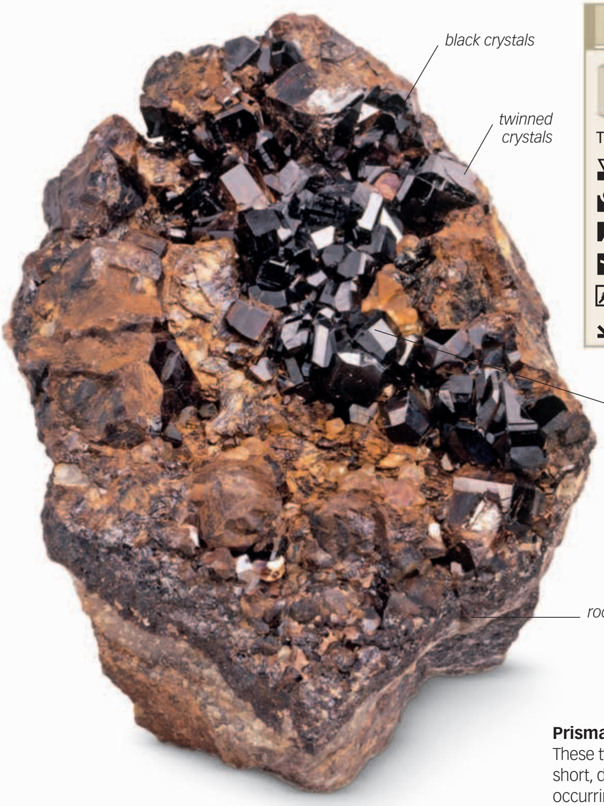

Prismatic crystals These twinned cassiterite crystals are short, dark-colored, and prismatic, occurring on a rocky matrix.

PROFILE

Tetragonal

6–7

7.0

Indistinct

Subconchoidal to uneven

White, grayish, brownish

Adamantine to metallic

VARIANT

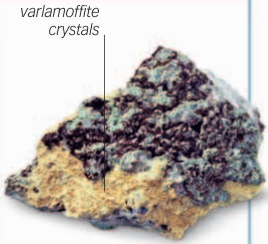

Varlamoffite cassiterite A specimen displaying the yellow variety of tin oxide

SnO_2

CASSITERITE

The tin oxide cassiterite takes its name from the Greek word for tin, *kassiteros*. Also called tinstone, it is the only important ore of tin. Colorless when pure, it commonly appears brown or black due to iron impurities. Rarely, it is gray or white. Its crystals are usually heavily striated prisms and pyramids. Twinned crystals are quite common. It can also be massive, occurring as a botryoidal, fibrous variety (wood tin) or as water-worn pebbles (stream tin).

Cassiterite forms in association with igneous rocks in hydrothermal veins at high temperature (1,065°F/575°C or above), with tungsten minerals such as ferberite (p.145), and with topaz (p.234), molybdenite (p.64), and tourmaline (p.224). Durable and relatively dense, it becomes concentrated in placer deposits after erosion from its primary rocks.

Brilliant gemstone This faceted, golden orange cassiterite gem is transparent with a resinous luster.

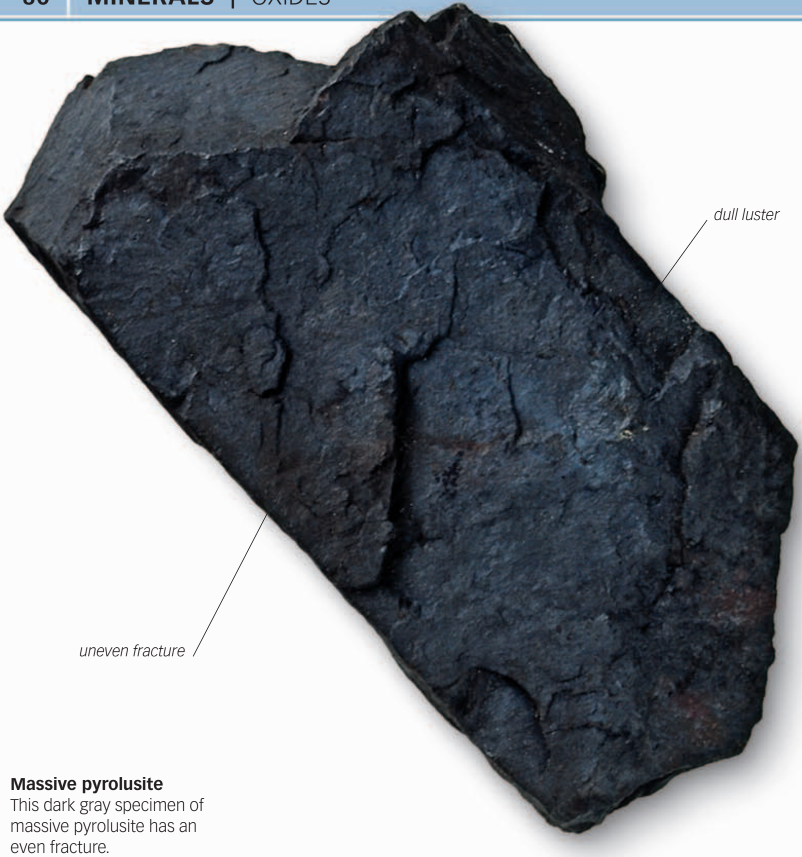

Massive pyrolusite
This dark gray specimen of massive pyrolusite has an even fracture.

PROFILE

Tetragonal
6–6½
4.4–5.1
Perfect
Uneven, brittle, splintery
Black or bluish black
Metallic to earthy

MnO_2

PYROLUSITE

Pyrolusite is the primary ore of the element manganese. Specimens are typically light gray to black in color. Pyrolusite usually occurs as massive aggregates. It also forms metallic coatings, crusts, fibers, nodules, botryoidal masses, concretions, and coatings that may be powdery or branching. Crystals are rare; when found, they are opaque and prismatic.

Pyrolusite forms under highly oxidizing conditions as an alteration product of manganese minerals, such as rhodochrosite (p.121). It has been found in bogs, lakes, and shallow marine environments and as a deposit laid down by circulating waters. Excellent crystals are found at Horni Blatna, Czech Republic, and at Bathurst, New Brunswick, Canada. The mineral is mined extensively in Russia, India, Georgia, and Ghana. Pyrolusite is used as a decolorizing agent in glass, as a coloring agent in bricks, and in dry cell batteries. It is also used in the manufacture of steel and saltwater-resistant manganese-bronze, which is used to make ships' propellers.

PROFILE

Orthorhombic

8½

3.7

Distinct

Uneven to conchoidal

Colorless

Vitreous

Cyclic twin The cyclic twinning of chrysoberyl exhibited by this specimen is common in the mineral.

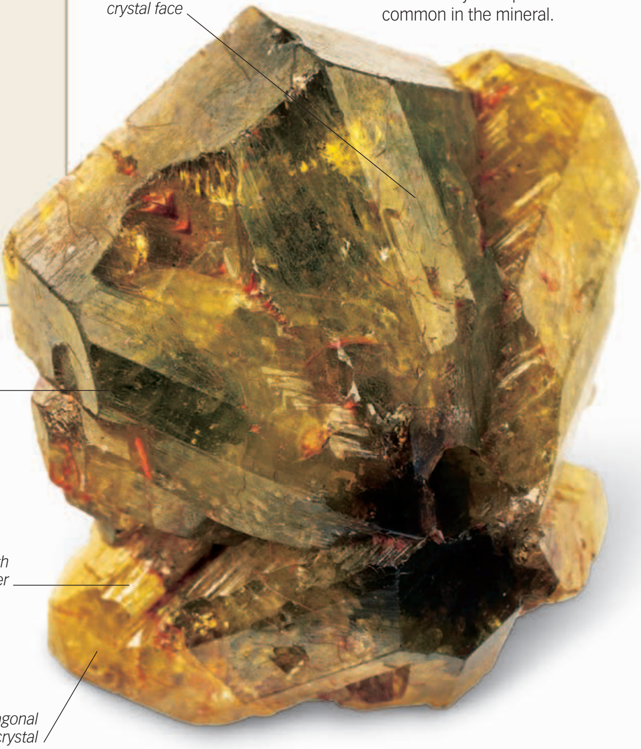
striation on crystal face
greenish yellow twinned crystal
transparent with vitreous luster
pseudohexagonal twinned crystal

VARIANTS

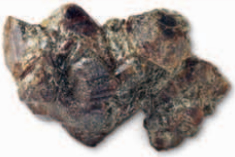

Siberian alexandrite A group of twinned alexandrite crystals with mica from Russia

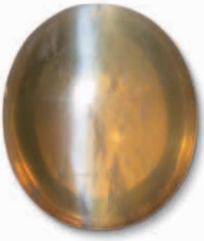

Yellow gemstone Cat's eye chrysoberyl in the most desirable honey-yellow color

$BeAl_2O_4$

CHRYSOBERYL

A beryllium aluminum oxide, chrysoberyl is hard and durable. It is inferior in hardness only to corundum (p.95) and diamond (p.47). Chrysoberyl is typically yellow, green, or brown in color. It forms tabular or short prismatic crystals and heart-shaped or pseudohexagonal twinned crystals. Alexandrite, one of its gemstone varieties, is one of the rarest and most expensive gems. Another variety, cat's eye, is also prized as a gemstone. It contains parallel fibrous crystals of other minerals that reflect light across the surface of a polished gemstone—an effect known as chatoyancy.

Chrysoberyl occurs in some granite pegmatites (p.260), gneisses (p.288), mica schists, and marbles (p.301). Crystals that weather out of the parent rock are often found in streams and gravel beds.

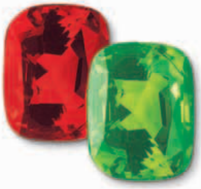

Color change Alexandrite exhibits color change—from brilliant green in daylight to cherry-red under tungsten light.

PROFILE

Orthorhombic
6–6½
5.2–8.0
Distinct
Subconchoidal or uneven
Red, brown, or black
Submetallic to resinous

Ferrocolumbite This opaque, tabular crystal of ferrocolumbite exhibits a submetallic to resinous luster.

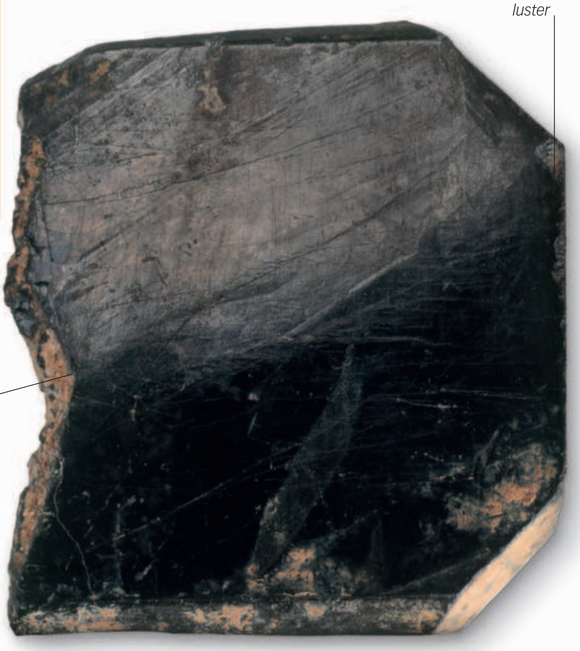

VARIANT

Yttrotantalite Dark crystals of the coltan series mineral yttrotantalite (yttrium-rich tantalite) in a light matrix

$(Fe,Mn)(Nb,Ta)_2O_6$–$(Fe,Mn)(Ta,Nb)_2O_6$

COLUMBITE–TANTALITE

Columbite forms the coltan series—a nearly complete solid-solution series—with the mineral tantalite. Minerals at the columbite end of this series are niobium-rich, and those at the tantalite end are tantalum-rich. Tantalite and columbite have similar crystal structures, but tantalite is denser, and tantalum atoms replace niobium atoms in the columbite crystal structure. The name of the mineral is prefixed with "ferro-" or "mangano-" depending on the content of iron or manganese. Ferrocolumbite is the most common mineral of the coltan group. Scandium and tungsten may also be present as minor constituents.

Coltan minerals are brown or black in color and are often iridescent. They are either massive or form tabular or short, prismatic crystals. They are the most abundant and widespread of the niobates and tantalates, and are the most important ores of niobium and tantalum. Coltan minerals mainly occur in granite pegmatite rocks (p.260) and in detrital deposits.

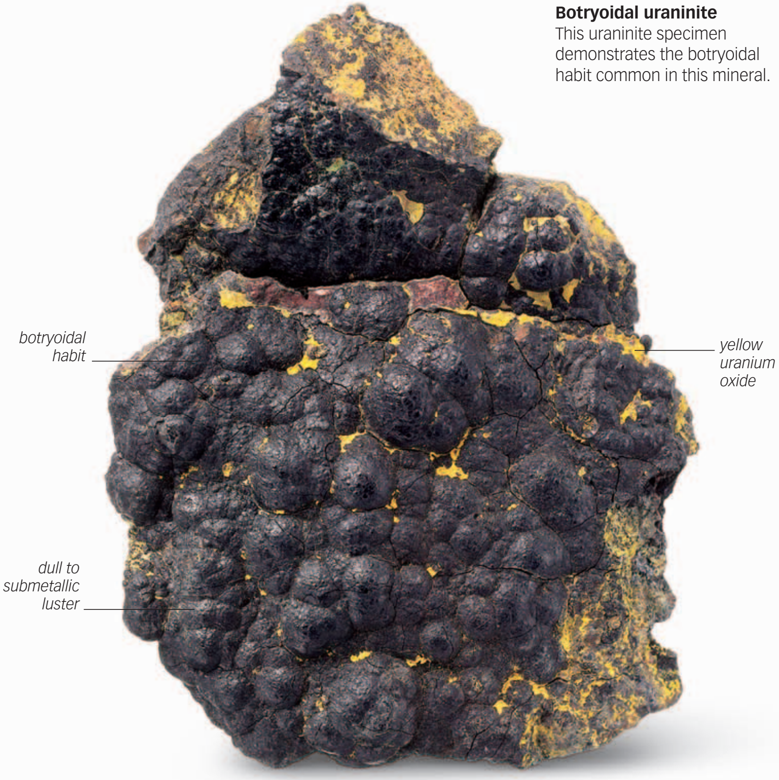

Botryoidal uraninite
This uraninite specimen demonstrates the botryoidal habit common in this mineral.

PROFILE

Cubic
5–6
6.5–11.0
None
Uneven to subconchoidal
Brownish black
Submetallic, pitchy, dull

UO_2

URANINITE

Discovered by the German chemist M.H. Klaproth in 1789, uraninite is a major ore of uranium. The pioneering work on radioactivity by Pierre and Marie Curie was based on uranium extracted from uraninite ores. It is black to brownish black, dark gray, or greenish. It commonly occurs in massive or botryoidal forms, or in banded or granular habits, and less commonly as opaque octahedral or cubic crystals.

Uraninite crystals occur in granitic pegmatites (p.260). Uraninite forms with cassiterite (p.79) and arsenopyrite in hydrothermal sulfide veins at high temperatures (1,065°F/575°C or above). It also forms at medium temperatures (400–1,065°F/200–575°C) as pitchblende. It also occurs as small grains in sandstones and conglomerates, where it may have weathered into secondary uranium minerals.

Uranium pellets
These ceramic pellets of enriched uranium are ready for use in nuclear reactors.

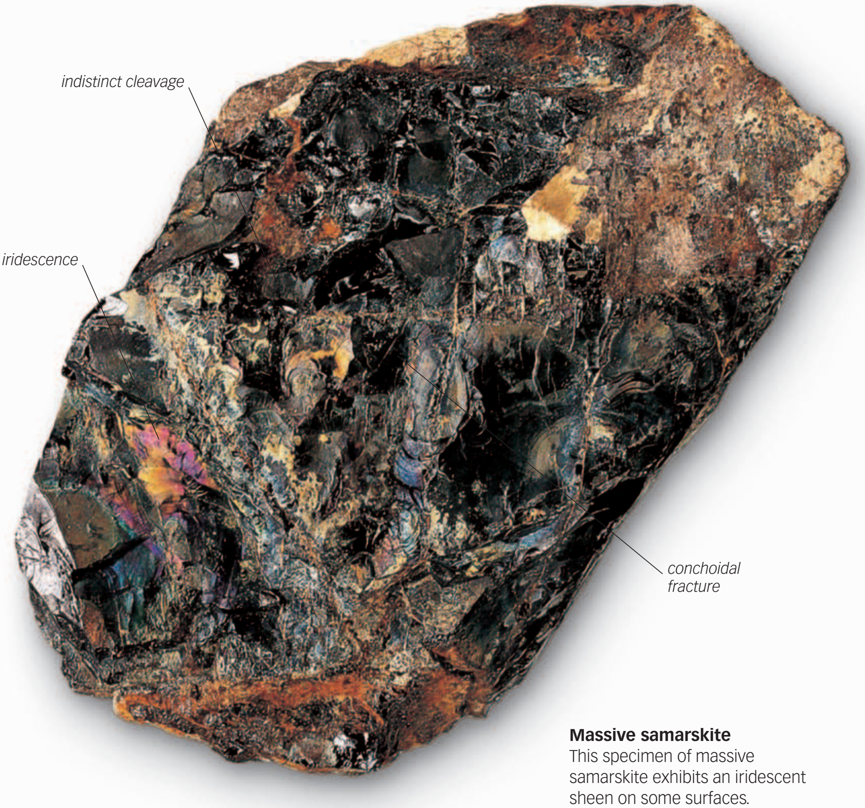

Massive samarskite
This specimen of massive samarskite exhibits an iridescent sheen on some surfaces.

PROFILE

Orthorhombic
5–6
5.7
Indistinct
Conchoidal, brittle
Dark reddish brown to black
Vitreous to resinous

$(Y,Fe,U)\ (Nb,Ta)O_4$

SAMARSKITE

Named in 1847 after Vasili Yefrafovich von Samarski-Bykhovets of Russia, samarskite is a complex oxide of yttrium, iron, tantalum, niobium, and uranium. Two types of samarskite are recognized—samarskite-(Y) or yttrium samarskite; and samarskite-(Yb) or ytterbium samarskite. The mineral is usually black and opaque but translucent in thin fragments. Crystals are stubby, opaque, and prismatic with a rectangular cross section—although samarskite is commonly found in the massive form. It is often brown or yellowish brown due to surface alteration. Specimens with high uranium content have a yellow-brown, earthy rind. Samarskite samples are usually radioactive.

Samarskite is usually found in rare, earth-bearing granitic pegmatites (p.260). It forms in similar conditions as columbite (p.82), so the minerals are closely associated. Samarskite is also associated with monazite (p.150), garnet, and other minerals. Yttrium from samarskite has been used in cathode-ray televisions, optical glass, and special ceramics.

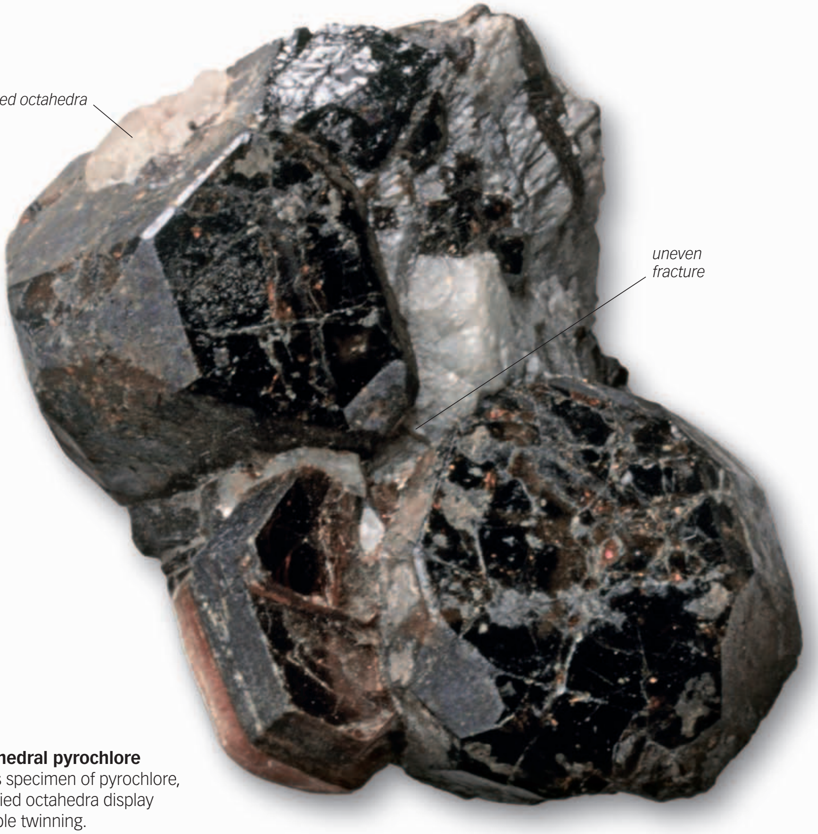

Octahedral pyrochlore
In this specimen of pyrochlore, modified octahedra display multiple twinning.

PROFILE

Cubic
5–5½
4.5
Distinct
Subconchoidal to uneven
Light brown, yellowish brown
Vitreous to resinous

$(Na,Ca)_2Nb_2(O,OH,F)_7$

PYROCHLORE

A major source of the element niobium, pyrochlore is a complex niobium sodium calcium oxide. Its name comes from the Greek *pyr* and *chloros*, which mean “fire” and “green” respectively—a reference to some specimens that turn green after heating. Pyrochlore is orange, brownish red, brown, or black in color. Crystals are typically well-formed octahedra with modified faces. They are frequently twinned or occur as either granular or massive aggregates. Pyrochlore often contains traces of uranium and thorium, and it may be radioactive. In such cases, its internal structure may be disrupted.

Pyrochlore forms in pegmatite rocks (p.260) and in igneous rocks dominated by carbonate minerals. It is an accessory mineral in silica-poor rocks, often occurring with magnetite (p.92), apatite (p.148), and zircon (p.233). It also accumulates in some detrital deposits. Niobium is a major alloying element in nickel-based superalloys. It has been used either alone or together with zirconium in claddings for nuclear-reactor cores.

Microlite crystals
This specimen of crystalline microlite contains tantalum in place of the niobium typically found in pyrochlore.

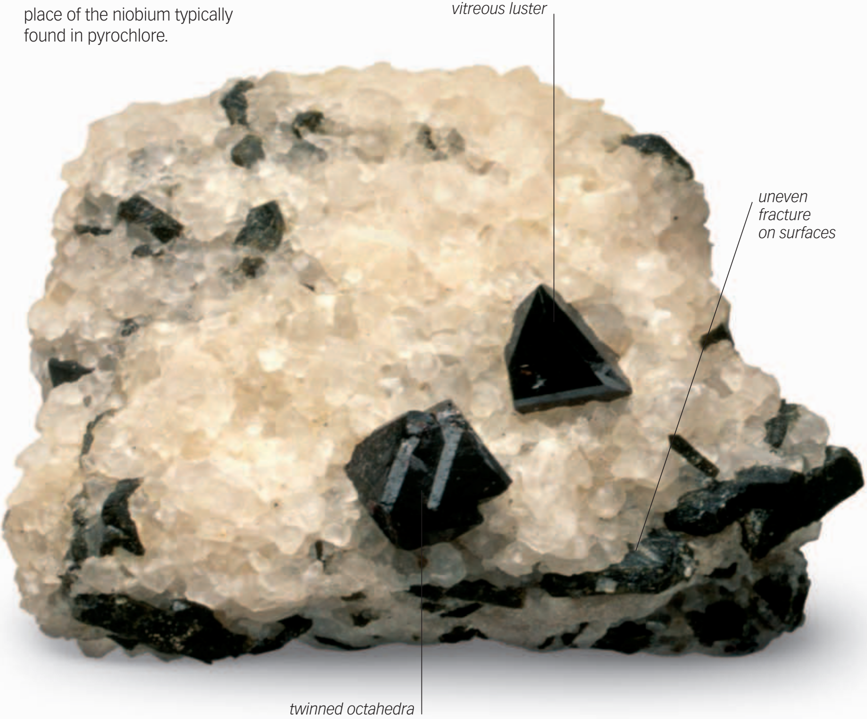

PROFILE

Cubic
5–5½
6.4
Distinct to difficult
Subconchoidal to uneven
Yellowish to brownish
Resinous to vitreous

$(Na,Ca)_2Ta_2O_6(O,OH,F)$

MICROLITE

Named in 1835, microlite takes it name from the Greek word *micros,* which means "small"—a reference to the small size of the crystals found in the locality where the mineral was first discovered. Microlite can, in fact, form excellent octahedral crystals, which can be up to ⅜ in (1 cm) on an edge. It also occurs as irregular grains. Specimens can be yellow, brown, black, green, or reddish. Microlite is related to pyrochlore (p.85), and both minerals are dominated by rare-earth elements: microlite by tantalum and pyrochlore by niobium.

Microlite is found in pegmatites (p.260), especially those rich in lepidolite (p.198) or other lithium-bearing minerals, and in albite (p.177). It is a major ore of tantalum, which is especially useful in high-capacitance electronic devices, particularly those used in miniaturized circuitry. Microlite is also used in corrosion-resistant chemical equipment. Excellent crystals are found at Dixon, New Mexico, USA; Shingus, Gilgit, Pakistan; Mattawa, Ontario, Canada; and at numerous localities in Brazil.

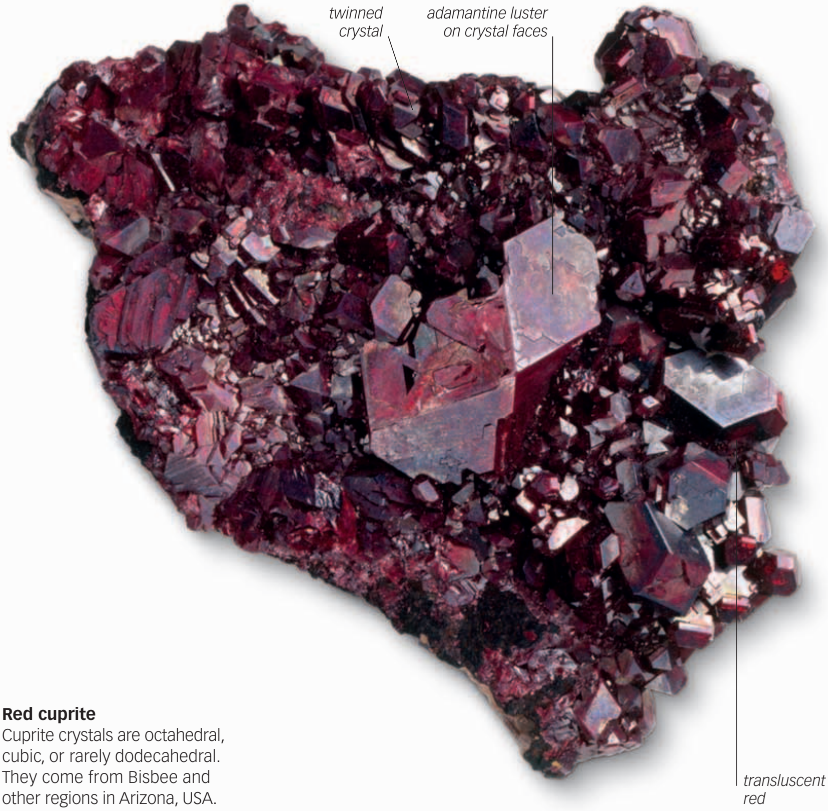

Red cuprite
Cuprite crystals are octahedral, cubic, or rarely dodecahedral. They come from Bisbee and other regions in Arizona, USA.

PROFILE

Cubic

- 3½–4
- 6.1
- Distinct
- Conchoidal, brittle
- Brownish red, shining
- Adamantine, submetallic

VARIANT

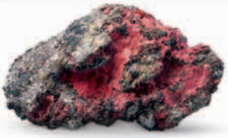

Chalcotrichite Bright red, hairlike crystals of the chalcotrichite variety

Cu_2O

CUPRITE

A relatively soft, heavy copper oxide, cuprite is an important ore of copper. Its crystals are either cubic or octahedral in shape and commonly striated. Massive or granular aggregates with the appearance of sugar are common. Cuprite is translucent and bright red when freshly broken but turns to a dull metallic gray color on exposure to light and pollutants. Cuprite is sometimes known as ruby copper due to its distinctive red color.

In the variety called chalcotrichite or plush copper ore, the crystals are a rich carmine color, fibrous, capillary, and are silky in appearance. They are found in loosely matted aggregates. Cuprite of the tile ore variety is soft, earthy, brick-red to brownish red in color, and often contains intermixed hematite (p.91) or goethite (p.102).

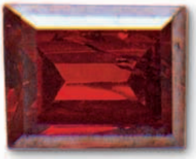

Step cut
Rare transparent cuprite is sometimes cut for collectors, as in this rectangular step cut.

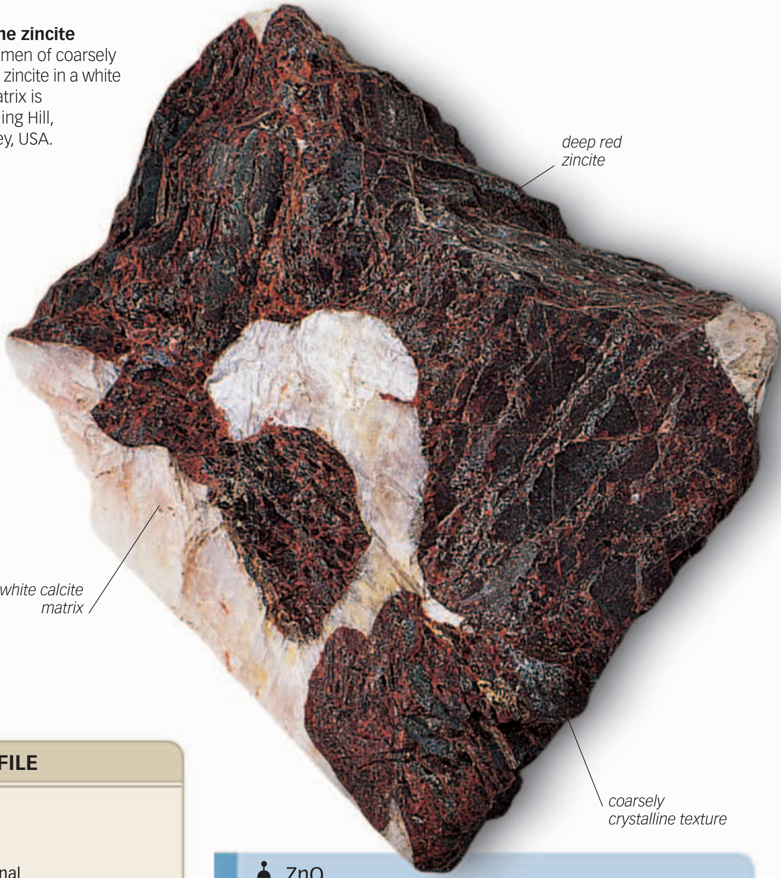

Crystalline zincite This specimen of coarsely crystalline zincite in a white calcite matrix is from Sterling Hill, New Jersey, USA.

PROFILE

Hexagonal
- 4–5
- 5.7
- Perfect
- Conchoidal
- Orange-yellow
- Resinous, submetallic

VARIANT

Granular habit Granular zincite with black franklinite

ZnO

ZINCITE

Red oxide of zinc is another name for zincite, which is a minor ore of zinc. Zincite occurs mostly as cleavable or granular masses. Natural crystals are rare, but when they occur they are pyramidal, pointed at one end and flat at the other. These crystals can be orange, red, yellow, or green.

Zincite is found mainly as an accessory mineral in zinc-ore deposits and is commonly associated with black franklinite and white calcite. It may also be a rare constituent of volcanic ash. Crystals are found only in secondary veins or fractures, where zincite forms by the chemical alteration or metamorphism of zinc deposits. Some so-called natural zincite crystals in the collectors' market are, in fact, large crystals that have formed in the chimneys of smelters. Natural crystals are rarely fluorescent; artificial crystals may range from fluorescent green to fluorescent yellow. The classic locality for fine zincite crystals is Franklin, New Jersey, USA. It is also found at Varmland and Nordmark, Sweden.

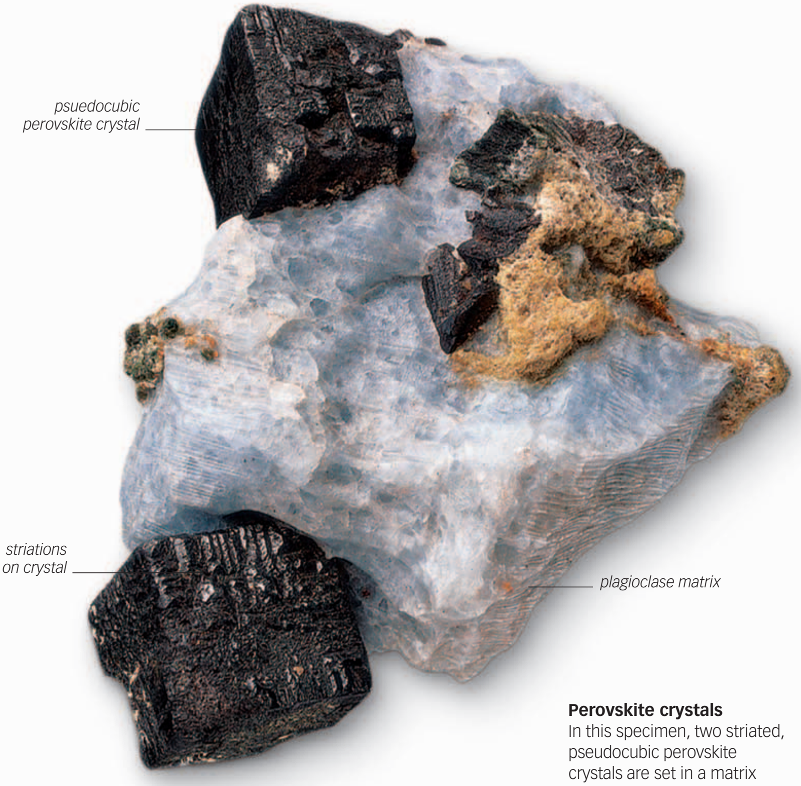

Perovskite crystals
In this specimen, two striated, pseudocubic perovskite crystals are set in a matrix of plagioclase feldspar.

PROFILE

Orthorhombic

5½

4.0

Imperfect

Subconchoidal to uneven

Gray to colorless

Adamantine or metallic

$CaTiO_3$

PEROVSKITE

A calcium titanium oxide, perovskite was named after the Russian mineralogist Count Lev Alekseevich Perovski in 1839. The composition of perovskite varies considerably: niobium can substitute for up to 44.9 percent titanium by weight, and cerium and sodium can substitute for calcium. When specimens are black, they have a metallic luster; when brown or yellow, they appear adamantine. Although perovskite is an orthorhombic mineral, its crystals are usually pseudocubic. Perovskite crystals can be pseudooctahedral in varieties where niobium or cerium has replaced a large amount of titanium. The crystals tend to be deeply striated and are frequently twinned.

Perovskite occurs in igneous rocks that are rich in iron and magnesium. It also occurs in contact metamorphic rocks associated with magnesium- and iron-rich intrusive igneous rocks and in some chlorite and talc schists. It is also found in carbonaceous chondrite meteorites (p.337).

PROFILE

Hexagonal

5–6

4.7

None

Conchoidal

Black to reddish brown

Metallic to submetallic

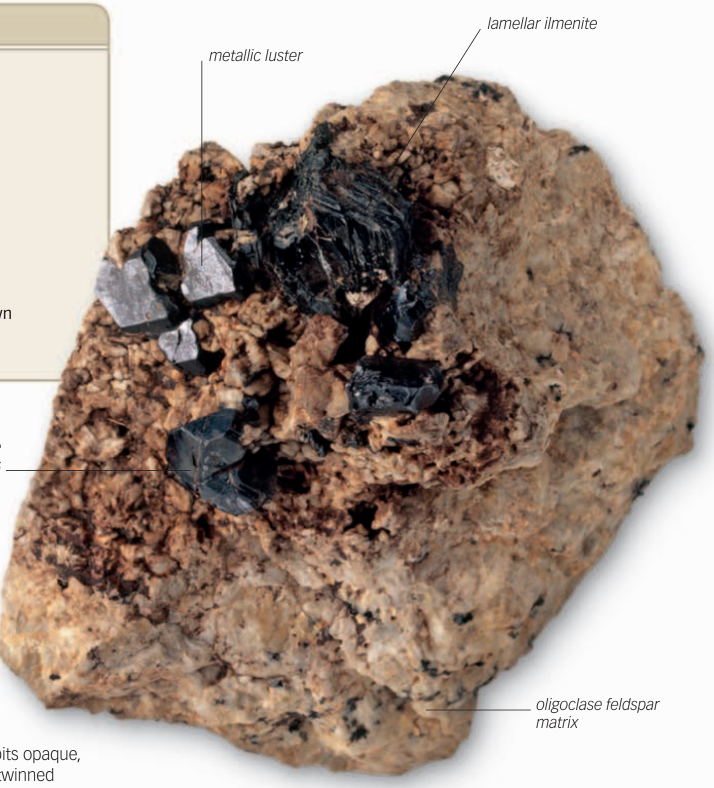

Ilmenite crystals This specimen exhibits opaque, black, lamellar, and twinned crystals of ilmenite.

VARIANT

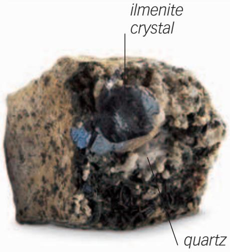

Tabular crystals Thin, gray, tabular ilmenite crystals with actinolite and quartz

$FeTiO_3$

ILMENITE

Named after the Il'menski Mountains near Miass, Russia, where it was discovered, ilmenite is a major source of titanium. Usually thick and tabular, its crystals sometimes occur as thin lamellae (fine plates) or rhombohedra. Ilmenite can also be massive, or occur as scattered grains. Intergrowths with hematite (p.91) or magnetite (p.92) are common, and ilmenite can be mistaken for these minerals because of its opaque, metallic, gray-black color. Unlike magnetite, however, ilmenite is nonmagnetic or very weakly magnetic; and it can be distinguished from hematite by its black streak. It may weather to a dull brown color.

Ilmenite is widely distributed as an accessory mineral in igneous rocks, such as diorite (p.264) and gabbro (p.265). It is a frequent accessory in kimberlite rocks (p.269), associated with diamond (p.47). It is also found in veins, pegmatite rocks (p.260), and black beach sands associated with magnetite, rutile (p.78), zircon (p.233), and other heavy minerals.

Rhombohedral hematite These superb hematite crystals from Elba, Italy, demonstrate hexagonal or rhombohedral form and metallic luster.

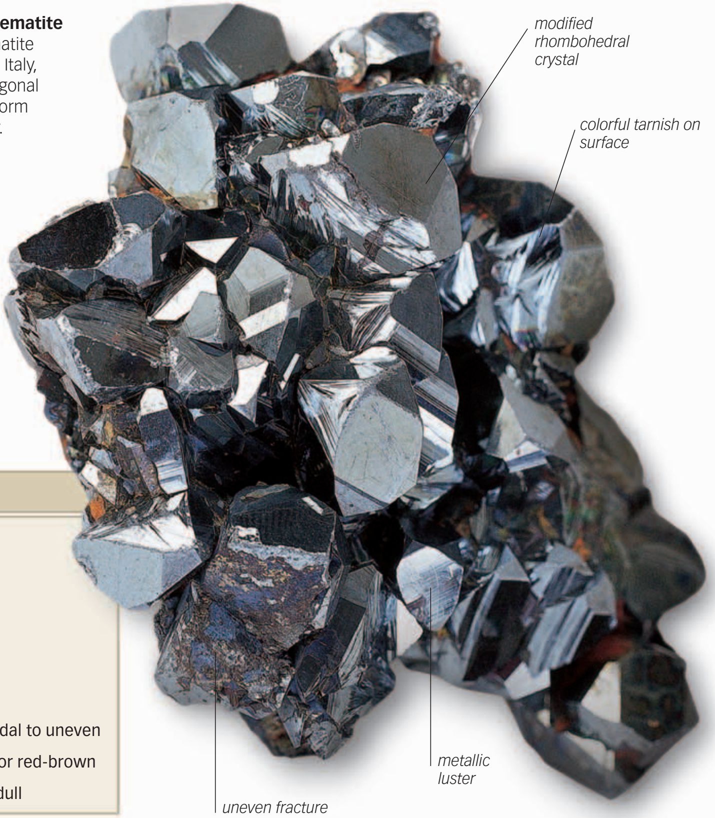

PROFILE

Hexagonal

5–6

5.3

None

Subconchoidal to uneven

Cherry-red or red-brown

Metallic to dull

VARIANTS

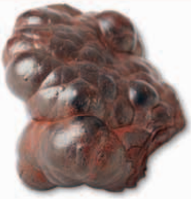

Kidney ore A perfect example of hematite's botryoidal habit

Specular hematite Brilliant platy crystals of specular hematite

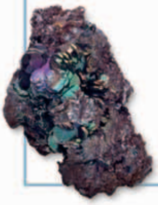

Iridescent hematite An iridescent crystal on rock

Fe_2O_3

HEMATITE

Dense and hard, hematite is the most important ore of iron (p.39) because of its high iron content and its abundance. The mineral occurs in various habits: steel-gray crystals and coarse-grained varieties with a brilliant metallic luster are known as specular hematite; thin, scaly forms make up micaceous hematite; and crystals in petal-like arrangements are called iron roses. Hematite also occurs as short, black, rhombohedral crystals and may have an iridescent tarnish. The soft, fine-grained, and earthy form of hematite is used as a pigment.

Important hematite deposits occur in sedimentary beds or in metamorphosed sediments. A compact variety known as kidney ore has a kidney-shaped surface. A form of ground hematite called rouge is used to polish plate glass and jewelry.

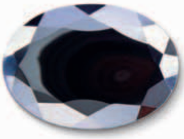

Oval cabochon This oval cabochon of black hematite is faceted on top. Hematite cabochons have been sold as "marcasites."

PROFILE

Cubic

5½–6

5.2

None

Conchoidal to uneven

Black

Metallic to semimetallic

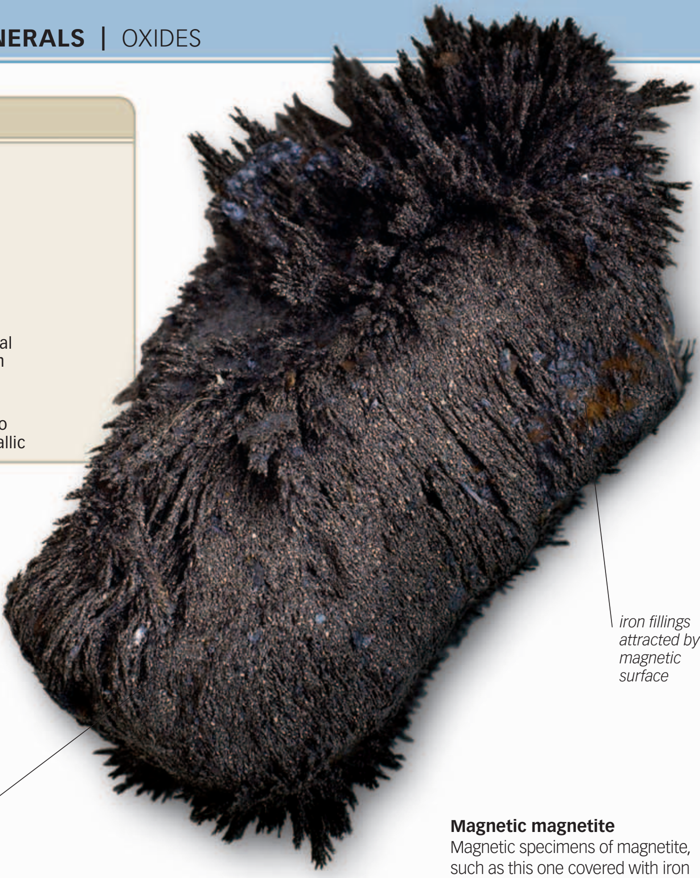

Magnetic magnetite Magnetic specimens of magnetite, such as this one covered with iron filings, are known as lodestones.

VARIANTS

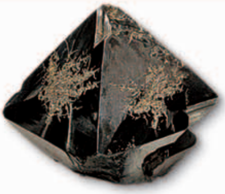

Octahedral crystal A magnetite crystal showing classic octahedral form

Magnetite crystals A cluster of black magnetite crystals

Fe_2O_4

MAGNETITE

An iron oxide, magnetite is named after the Greek shepherd boy Magnes, who noticed that the iron ferrule of his staff and the nails of his shoes clung to a magnetite-bearing rock. All magnetite can be picked up with a magnet, but some magnetite is itself naturally magnetic and attracts iron filings and deflects compass needles. Magnetite usually forms octahedral crystals, although it sometimes occurs as highly modified dodecahedrons. Specimens can also be massive or granular, occurring as disseminated grains and as concentrations in black sand. Magnetite is similar in appearance to hematite (p.91), but hematite is nonmagnetic and has a red streak.

Magnetite occurs in a range of geological environments. It forms at high temperatures (1,065°F/575°C or above) as an accessory mineral in metamorphic and igneous rocks and in sulfide veins. A major ore of iron (p.39), magnetite forms large ore bodies. Economically important deposits occur in silica-poor intrusions of igneous rocks and in banded ironstones (p.329).

Swedish fergusonite
In this specimen from Ytterby, Sweden, fergusonite crystals rest on a matrix of feldspar.

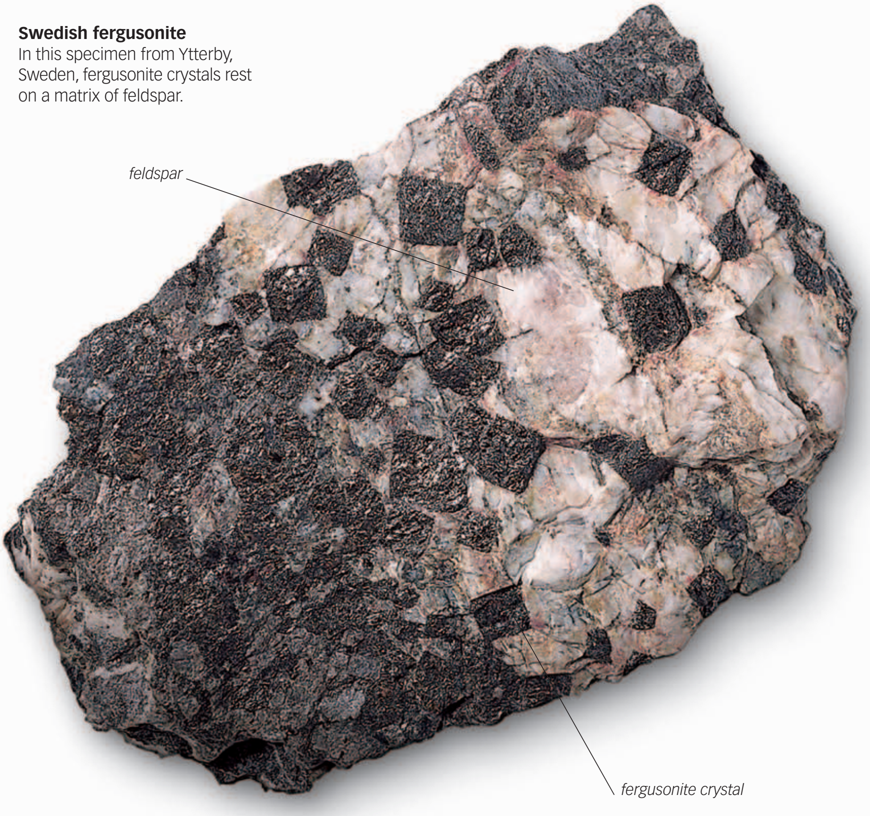

PROFILE

Tetragonal

5½–6½

4.2–5.7

Poor

Subconchoidal, brittle

Brown, yellow-brown, greenish gray

Vitreous to submetallic

$(Ce,Y,La,Nd)NbO_4$

FERGUSONITE

Named after the Scottish mineralogist Robert Ferguson (1767–1840), the fergusonite group contains several minerals. All fergusonites may be considered as sources of the rare metals they contain. The most common is fergusonite-(Y), which is rich in yttrium. Its crystals are prismatic to pyramidal in shape and black to brownish black. Fergusonite-(Ce) is cerium-rich, dark red to black in color, and forms prismatic dipyramidal crystals—although these are rare. Fergusonite-(Nd), a neodymium-bearing fergusonite, is usually granular. Another member of the fergusonites, formanite-(Y), is found as tabular crystals and anhedral pebbles. Yet other fergusonites, most of which appear in minor quantities, bear the prefix “beta.”

Fergusonites can also have varying amounts of erbium, lanthanum, niobium, dysprosium, uranium, thorium, zirconium, and tungsten. They can be found in granitic pegmatites (p.260) associated with other rare-earth minerals and in placer deposits.

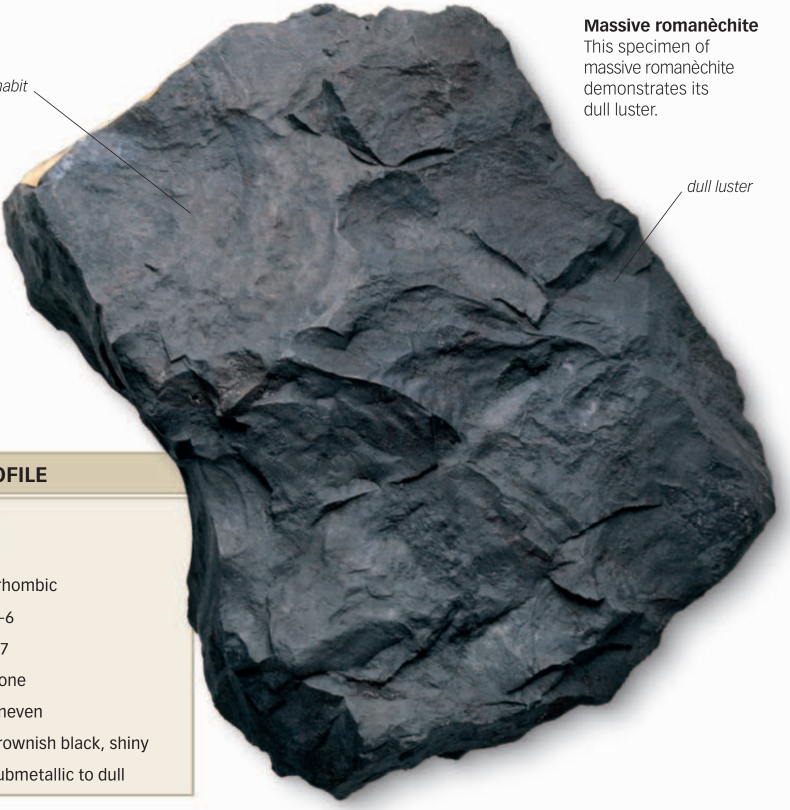

Massive romanèchite This specimen of massive romanèchite demonstrates its dull luster.

PROFILE

Orthorhombic
5–6
4.7
None
Uneven
Brownish black, shiny
Submetallic to dull

VARIANT

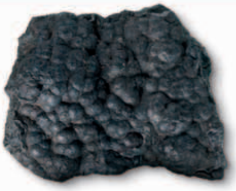

Botryoidal romanèchite Dense, submetallic, botryoidal romanèchite

$(Ba, H_2O)\ (Mn_4+Mn_3+)_5O_{10}$

ROMANÈCHITE

A hard, black, barium manganese oxide, romanèchite is named for its occurrence at Romanèche-Thorins, France. It is one of the manganese oxides that were formerly grouped together under the name psilomelane, which has been applied to several distinct minerals. Although the name psilomelane is no longer used to refer to a particular mineral, it continues to be used as a term of convenience for a group of barium-bearing manganese oxides. Romanèchite specimens are usually fine-grained or fibrous. Crystals are rare; when found, they are prismatic.

Romanèchite forms as an alteration product of other manganese minerals and is an ore of manganese. The mineral also forms in bogs, lakes, and shallow seas. Although romanèchite is named after a French locality, it was first identified at Schneeberg, Saxony, Germany. Other important deposits of romanèchite are at Tekrasni, India; Pilbara, Australia; Cornwall, England; and Hidalgo County, New Mexico, USA.

PROFILE

Hexagonal

9

4–4.1

None

Conchoidal to uneven

Colorless

Adamantine to vitreous

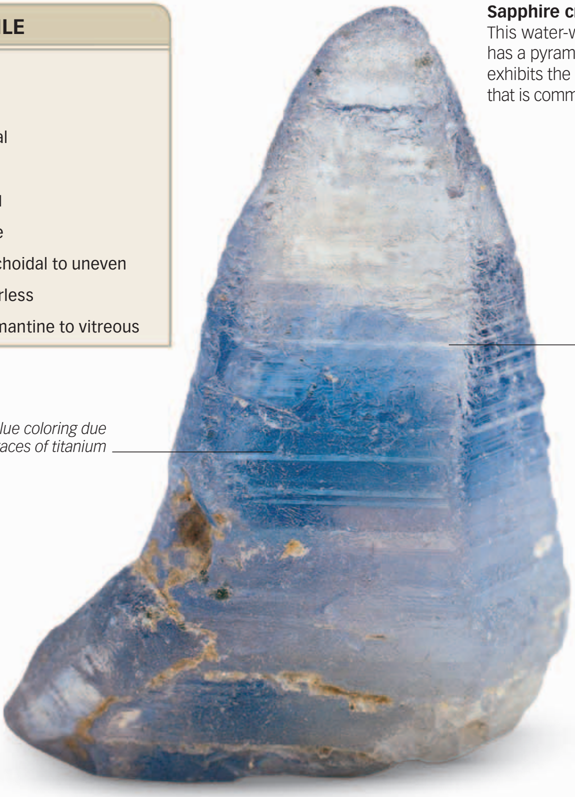

Sapphire crystal
This water-worn crystal has a pyramidal form and exhibits the color zoning that is common in sapphire.

color zoning

blue coloring due to traces of titanium

VARIANTS

vitreous luster

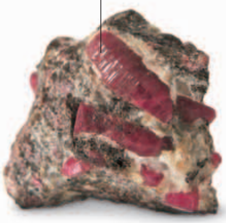

Ruby in matrix
Prismatic Kashmir rubies embedded in a rock matrix

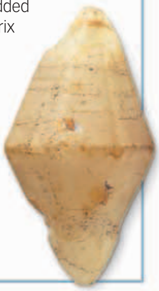

Common corundum
An opaque, dipyramidal crystal of common corundum

Al_2O_3

CORUNDUM

After diamond, corundum is the hardest mineral on Earth. The name corundum comes from the Sanskrit *kuruvinda*, meaning “ruby”—the name given to red corundum. Ruby and sapphire are gem varieties of corundum. An aluminum oxide, corundum is commonly white, gray, or brown, but gem colors include red ruby and blue, green, yellow, orange, violet, and pink sapphire. Colorless forms also occur. Ruby forms a continuous color succession with pink sapphire; only stones of the darker hues are considered to be ruby.

Corundum crystals are generally hexagonal, either tabular, tapering barrel-shaped, or dipyramidal. Corundum can also be massive or granular. It forms in syenites (p.262), certain pegmatites (p.260), and in high-grade metamorphic rocks. It is concentrated in placer deposits.

Antique ruby ring
In this ring, a square-cut ruby has been set at right angles to its square setting.

PROFILE

Cubic
7½–8
3.6–4.1
None
Conchoidal to uneven
White
Vitreous

Spinel octahedrons In this specimen, octahedral crystals of pleonaste, or black spinel, are set in a quartz matrix.

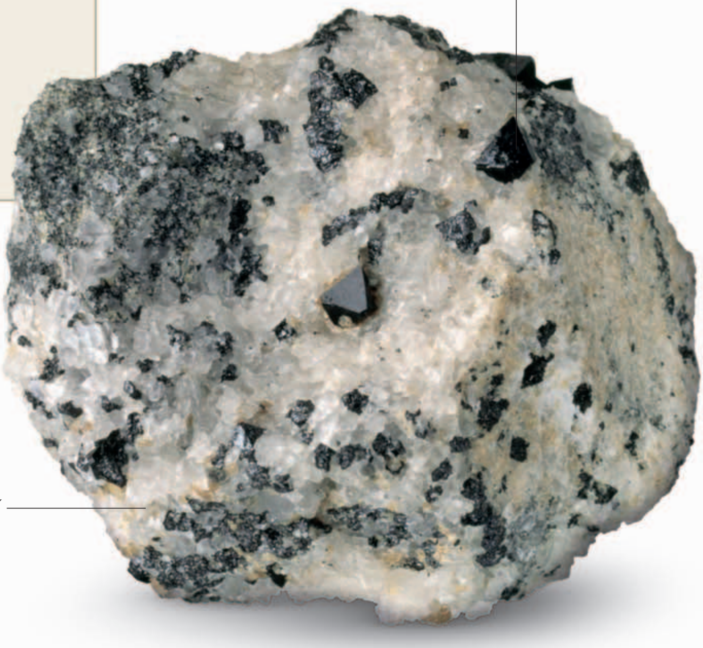

VARIANTS

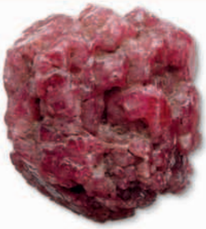

Spinel aggregate Numerous ruby spinel crystals

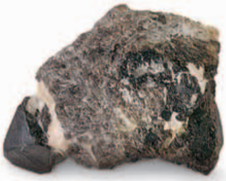

Black spinel A modified octahedron of black spinel on a rock matrix

$MgAl_2O_4$

SPINEL

Spinel is the name of both an individual mineral and of a group of metal-oxide minerals that share the same crystal structure. Minerals in this group include gahnite (p.97), franklinite, and chromite (p.99). Spinel is found as glassy, hard octahedra, or as grains or masses. Although familiar as a blue, purple, red, or pink gemstone, spinel also occurs in other colors. Red spinel is called ruby spinel; its blood-red color is due to the presence of chromium.

A minor constituent of peridotites (p.266), kimberlites (p.269), basalts (p.273), and other igneous rocks, spinel also forms in aluminum-rich schists (pp.291–92) and metamorphosed limestones. Water-worn crystals come from stream deposits. The earliest known spinel dates back to 100 BCE and was discovered near Kabul, Afghanistan.

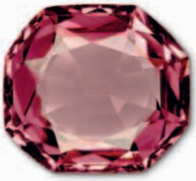

Spinel gemstone This superb faceted spinel shows excellent red-lavender color and good clarity.

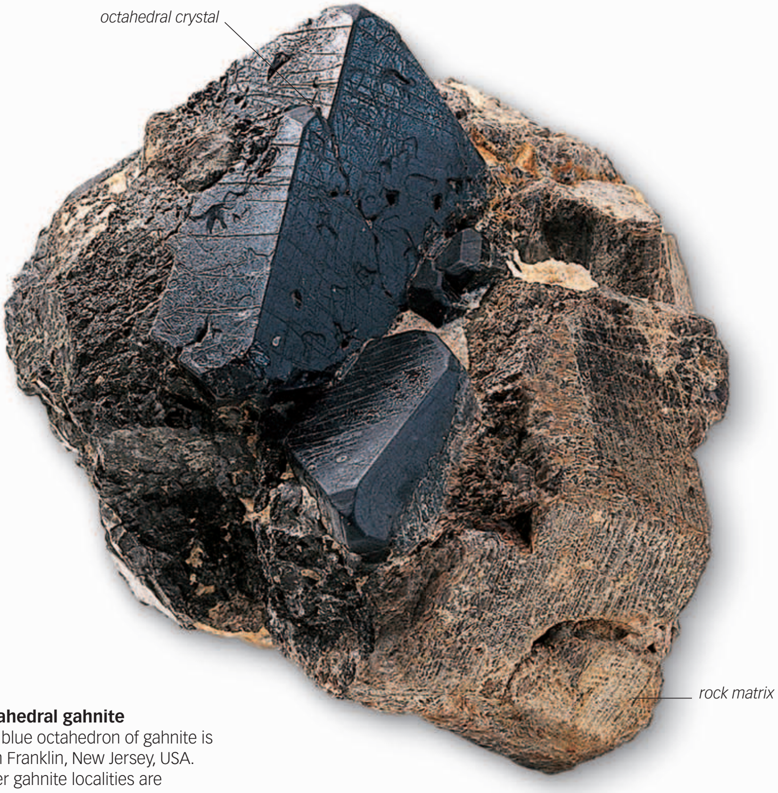

Octahedral gahnite
This blue octahedron of gahnite is from Franklin, New Jersey, USA. Other gahnite localities are Colorado and Maine, USA.

PROFILE

Cubic
7½–8
4.6
Indistinct
Conchoidal, irregular
Grayish
Vitreous

$ZnAl_2O_4$

GAHNITE

A zinc aluminum oxide, gahnite is a member of the spinel group and frequently forms the simple octahedral crystals typical of the group. Crystals usually show good external form. They may be striated on faces and cleavage surfaces. Usually dark green or blue to black in color, they can reach up to 4½ in (12 cm) on an edge. Crystals can sometimes be gray, yellow, or brown in color. Gahnite also occurs as irregular grains and masses, and in some lithium pegmatites (p.260) as gem-clear nodules.

Gahnite was named in 1807 after the Swedish chemist and mineralogist John Gottlieb Gahn. It is found in crystalline schists (pp.291–92) and gneisses (p.288), in granites (pp.258–59) and granitic pegmatites, and in contact metamorphosed limestones. It sometimes forms from the low-grade metamorphism of bauxite (p.101) and is also found in placer deposits. Superb crystals occur at Salida and Cotopaxi, Colorado, USA; at Falun, Sweden; and at Minas Gerais, Brazil.

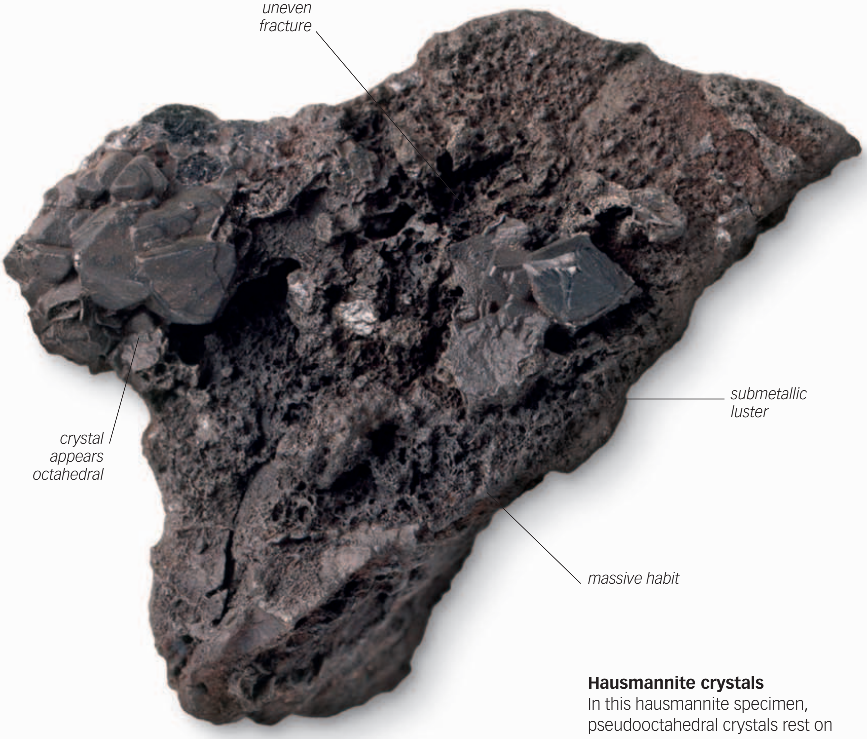

Hausmannite crystals
In this hausmannite specimen, pseudooctahedral crystals rest on a base of massive hausmannite.

PROFILE

Tetragonal
5½
4.8
Perfect
Uneven
Reddish brown
Submetallic

$Mn^{2+}Mn_2^{3+}O_4$

HAUSMANNITE

Named in 1827 after Johann Friedrich Ludwig Hausmann, a German professor of mineralogy, hausmannite is dark brown or black and is usually granular or massive. Well-formed crystals are uncommon yet distinctive. They are pseudooctahedral in shape but often have additional faces. Small amounts of iron and zinc may substitute for manganese in the hausmannite structure. Hausmannite forms in hydrothermal veins, and it also occurs where manganese-rich rocks have been metamorphosed. It is often found associated with other manganese oxides, such as pyrolusite (p.80), romanéchite (p.94), and the manganese–iron mineral bixbyite. Superb crystals, up to 1½in (4 cm) long, come from Brazil, South Africa, and Germany.

Hausmannite is an ore of manganese, which is added to aluminum and magnesium alloys to improve corrosion resistance. Manganese oxides are important in the manufacture of steel, where they absorb the sulfur in iron ores and impart strength.

PROFILE

Cubic
5 ½
4.7
None
Uneven
Brown
Metallic

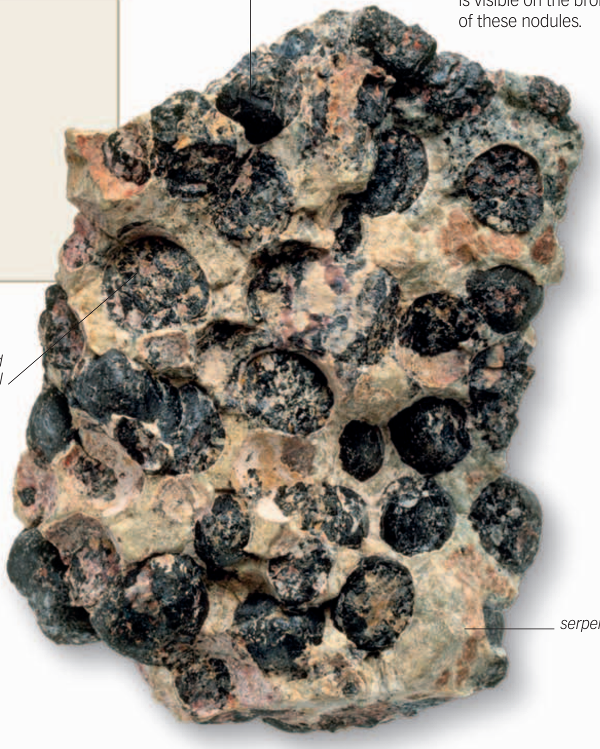

Nodular chromite The metallic luster of chromite is visible on the broken surfaces of these nodules.

VARIANT

Massive chromite A glossy black specimen of chromite

$FeCr_2O_4$

CHROMITE

A member of the spinel mineral group, chromite is an iron chromium oxide and the most important ore of chromium. Crystals are uncommon, but when found they are octahedral. Chromite is usually massive or in the form of lenses and tabular bodies, or it may be disseminated as granules. It is sometimes found as a crystalline inclusion in diamond. Chromite is dark brown to black in color and can contain some magnesium and aluminum.

Chromite is most commonly found as an accessory mineral in iron- and magnesium-rich igneous rocks or concentrated in sediments derived from them. It occurs as layers in a few igneous rocks that are especially rich in iron and magnesium. Almost pure chromite is found in similar layers in sedimentary rocks. The layers are preserved when the sedimentary rocks metamorphose to form serpentinite (p.298). Referred to as chromitites, these rocks are the most important ores of chromium. The weathering of chromite ore bodies can also lead to its concentration in placer deposits.

HYDROXIDES

Hydroxides form when metallic elements combine with a hydroxyl radical. They are found predominantly as weathering products of other minerals. Hydroxide minerals are usually less dense and softer than oxide minerals. Many hydroxides are important ore minerals.

COMPOSITION

Nearly all hydroxides form at low temperatures (up to 400°F/200°C), when water reacts with an oxide. They contain the hydroxyl radical, which is a single chemical unit made up of one atom of hydrogen and one atom of oxygen.

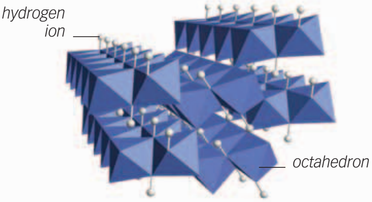

Crystal structure of diaspore
In the aluminum hydroxide diaspore, aluminum ions are in octahedral coordination with hydroxyl groups, forming strips of octahedra.

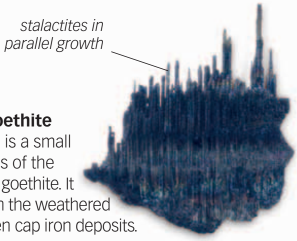

Stalactitic goethite
This specimen is a small stalactitic mass of the iron hydroxide goethite. It mainly forms in the weathered zones that often cap iron deposits.

OCCURRENCE AND USES

Hydroxide minerals are found in most places where water has altered primary oxides. Some hydroxides are also precipitated directly. They are often important ore minerals. The aluminum hydroxides diaspore, bohemite, and gibbsite constitute bauxite, the ore of aluminum. Goethite, an iron hydroxide, is an ore of iron.

UMBER
OCHER

Artists' pigments
Although now mostly replaced by synthetics, goethite (formerly called limonite) has provided the pigment in umbers and ochers for millennia.

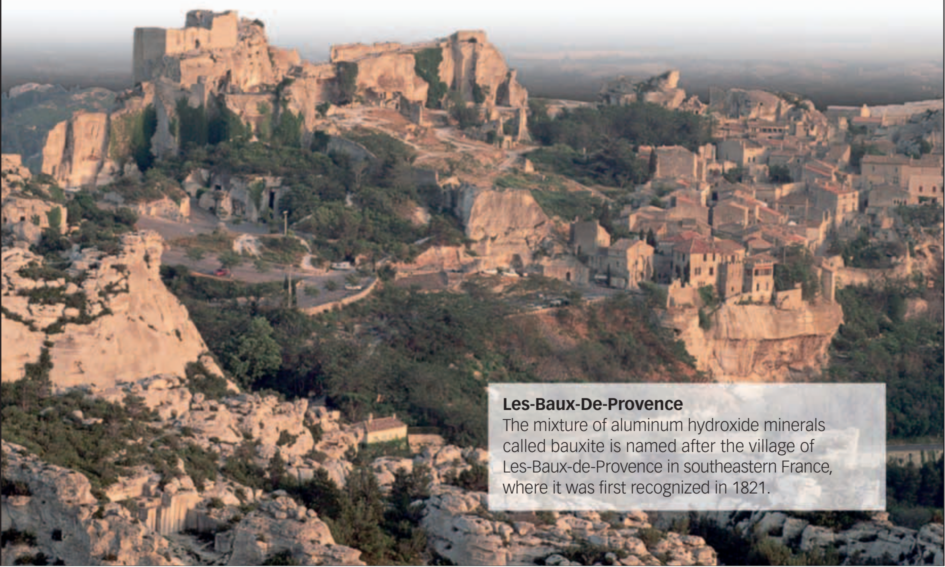

Les-Baux-De-Provence
The mixture of aluminum hydroxide minerals called bauxite is named after the village of Les-Baux-de-Provence in southeastern France, where it was first recognized in 1821.

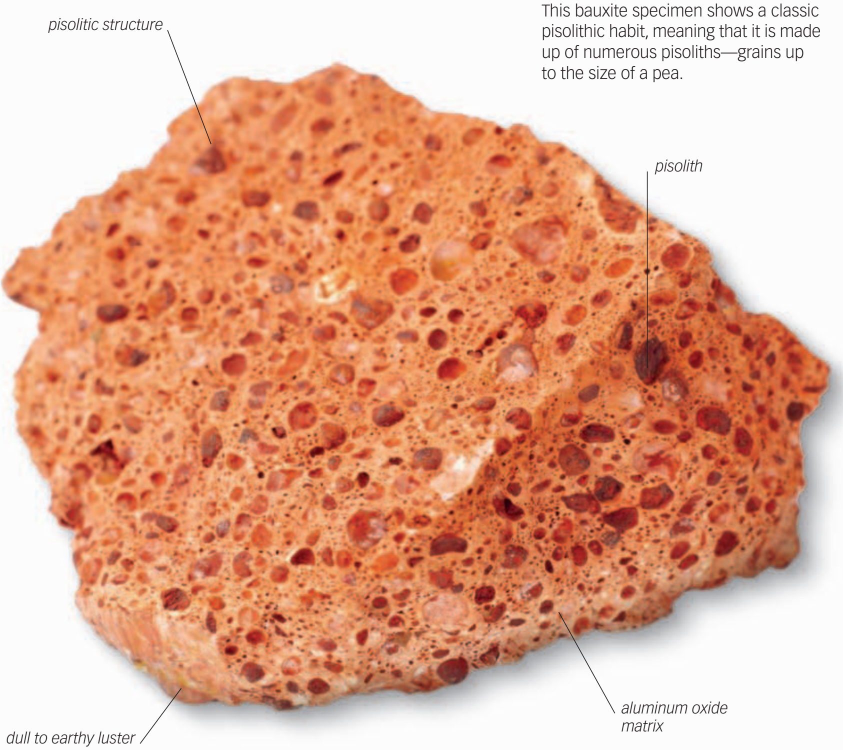

Pisolitic bauxite
This bauxite specimen shows a classic pisolithic habit, meaning that it is made up of numerous pisoliths—grains up to the size of a pea.

Mixture of hydrous aluminum oxides

BAUXITE

Although bauxite is not a mineral, it is one of the most important ores because it is the sole source of aluminum. The product of weathering of aluminum-rich rocks, it contains several constituent minerals. Bauxite is variably creamy yellow, orange, pink, or red because of the presence of quartz (p.168), clays, and hematite (p.91) and other iron oxides in addition to several hydrated aluminum oxides.

Bauxite forms as extensive, shallow deposits in humid tropical environments. It may be nodular, pisolitic, or earthy. Deposits are soft, easily crushed, and textureless or hard, dense, and pealike. Bauxite may also be porous but strong and stratified, or it may retain the form of its parent rock.

Versatile aluminum
A key metal of the modern age, aluminum is used to make products ranging from takeout trays to spacecraft.

PROFILE

Crystal system None

1–3

2.3–2.7

None

Uneven

Usually white

Earthy

VARIANT

Bauxite as an ore Primary ore of the metal aluminum

PROFILE

Orthorhombic

5–5½

4.3

Perfect

Uneven

Brownish yellow to ocher-red

Adamantine to metallic

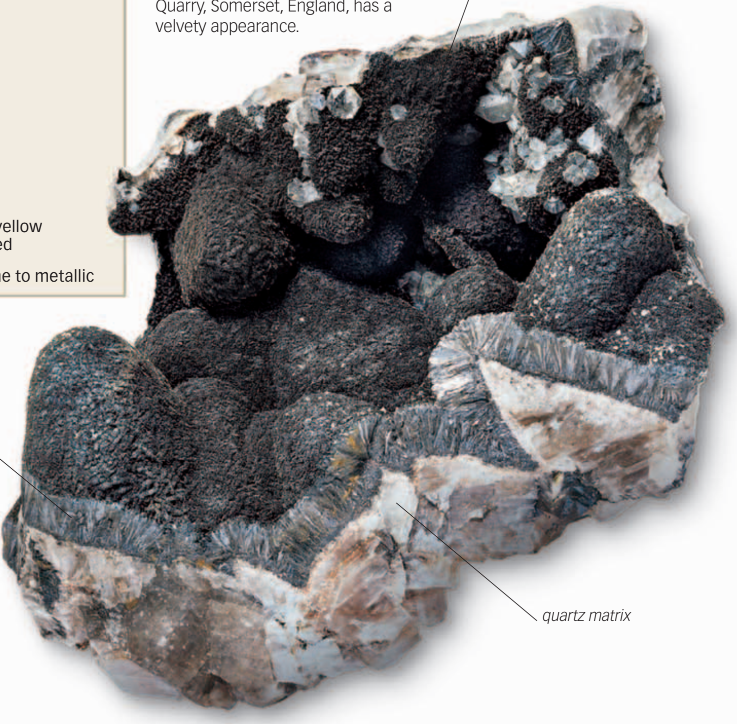

Velvety goethite This botryoidal mass of radiating goethite crystals from Merehead Quarry, Somerset, England, has a velvety appearance.

VARIANTS

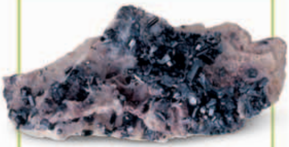

Prismatic crystals Striated crystals in quartz matrix

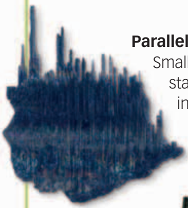

Parallel growth Small goethite stalactites in parallel growth

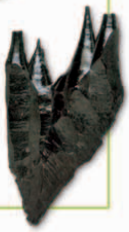

Stalactitic goethite Hard, glossy goethite in classic stalactitic habit

$FeO(OH)$

GOETHITE

Named after the German mineralogist Johann Wolfgang von Goethe in 1806, goethite is a common mineral. It can be brownish yellow, reddish brown, or dark brown in color, depending on the size of the crystal in the specimen—small crystals appear lighter, and larger ones darker. It can occur as opaque black, prismatic and vertically striated crystals; velvety, radiating fibrous aggregates; flattened tablets or scales; and reniform or botryoidal masses. Goethite can also occur in stalactitic or massive forms and in tufts and drusy coatings.

Goethite is an iron oxide hydroxide, although manganese can substitute for up to 5 percent of the iron. It forms as a weathering product in the oxidation zones of veins of iron minerals, such as pyrite (p.62), magnetite (p.92), and siderite (p.123). Goethite may occur with these minerals in the gossan, or iron hat, which is the weathered capping of an iron ore deposit. It also occurs in a form called bog iron ore, which can be produced by living organisms.

PROFILE

Monoclinic

- 4
- 4.3
- Perfect, good
- Uneven
- Reddish brown to black
- Submetallic

Prismatic manganite This specimen is a mass of pseudoorthorhombic prisms showing typical deep striations on the crystal faces.

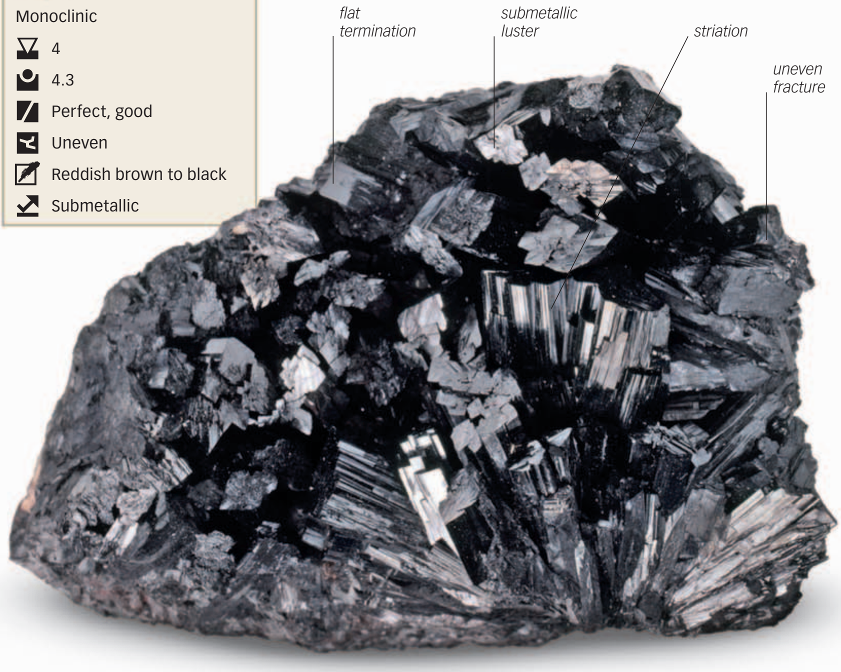

VARIANT

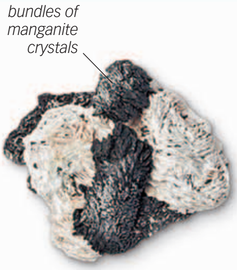

Crystal bundles Manganite crystals grouped in bundles

$MnO(OH)$

MANGANITE

A widespread and important ore of manganese, manganite is hydrated manganese oxide. The mineral had been described by a number of different names since 1772, but was finally given its current name, which it owes to its manganese component, in 1827. Opaque and metallic dark gray or black, crystals of manganite are mostly pseudoorthorhombic prisms, typically with flat or blunt terminations, and are often grouped in bundles and striated lengthwise. Multiple twinning is common. Manganite can also be massive or granular; it is then hard to distinguish by eye from other manganese oxides, such as pyrolusite (p.80).

An important ore of manganese, manganite occurs in hydrothermal deposits formed at low temperature (up to 400°F/200°C) with calcite (p.114), siderite (p.123), and barite (p.134), and in replacement deposits with goethite (p.102). Manganite also occurs in hot-spring manganese deposits. It alters to pyrolusite and may form by the alteration of other manganese minerals.

Dark red diaspore
In this specimen, a mass of dark red, thin, platy diaspore crystals rests in a matrix of corundum.

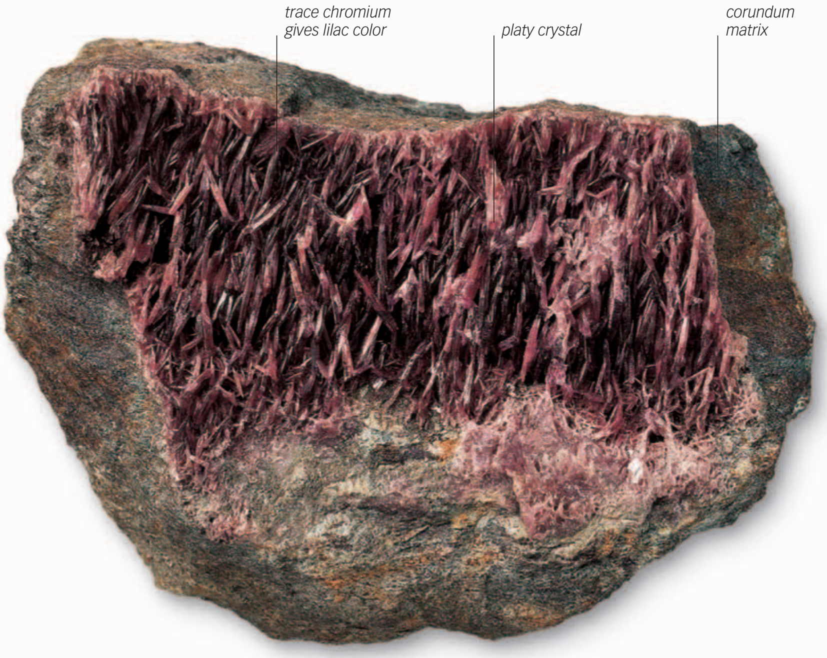

PROFILE

Orthorhombic

6½–7

3.4

Perfect, imperfect

Conchoidal, brittle

White

Vitreous

$AlO(OH)$

DIASPORE

Diaspore takes its name from the Greek word *diaspora*, which means "scattering"—a reference to the way diaspore crackles and depreciates under high heat. Its crystals are thin and platy, elongated, tabular, prismatic, or needlelike and are often twinned. Diaspore can be massive or can occur as disseminated grains. It may be colorless, white, grayish white, greenish gray, light brown, yellowish, lilac, or pink in color. The same specimen can appear to have different colors when viewed from different directions.

Diaspore forms in metamorphic rocks, such as schists (p.292) and marbles (p.301), where it is often associated with corundum (p.95), spinel (p.96), and manganite (p.103). It is widespread in bauxite (p.101), laterite (p.326), and aluminous clays.

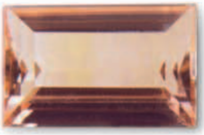

Faceted gem
Zultanite, which is a rare, transparent type of diaspore crystal from Turkey, is a collector's gem.

PROFILE

Hexagonal

- 2½
- 2.4
- Perfect
- Uneven, sectile
- White
- Waxy to vitreous/pearly

Fibrous brucite
This fibrous mass of brucite with a vitreous luster is from Timmins, Ontario, Canada.

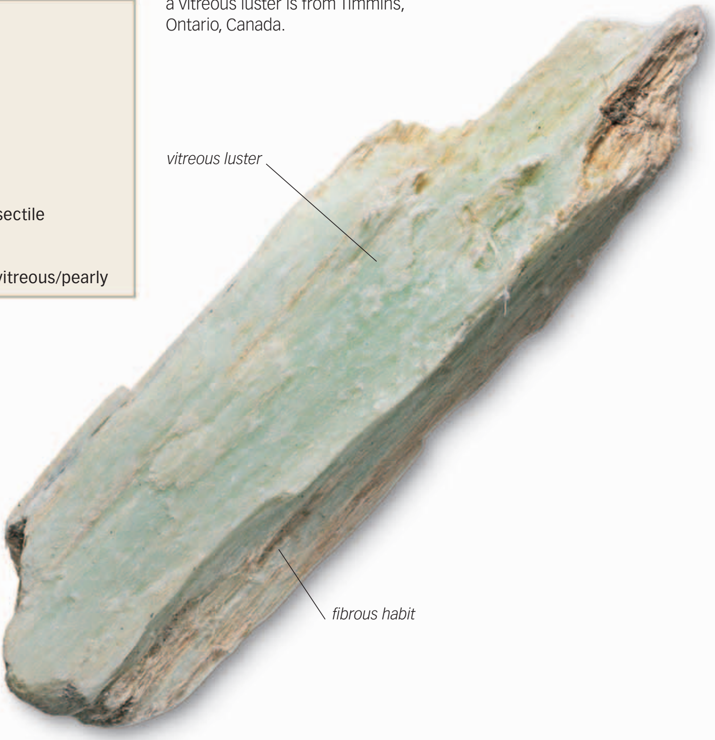

VARIANTS

brucite crystal

Tabular crystals Tabular brucite in a rock matrix

Nemalite A fibrous variety of brucite

$Mg(OH)_2$

BRUCITE

Named after the American mineralogist Archibald Bruce in 1824, brucite is magnesium hydroxide. Usually white, it can be pale green, gray, or blue. Manganese may substitute to some degree for magnesium, producing yellow to red coloration. Its crystals can be tabular or form aggregates of plates. They tend to be soft, and range from waxy to glassy in appearance. Fine large crystals have been collected from nemalite, a variety of brucite that occurs in fibres and laths. Brucite may also occur in massive, foliated, fibrous, or, more rarely, granular habits.

Brucite is found in metamorphic rocks, such as schist (p.292), and in low-temperature hydrothermal veins (up to 400°F/200°C) in marbles (p.301) and chlorite schists. It is used as a primary source of medical magnesia and as a fire retardant.

Kiln lining
Because of its high melting point, brucite is used to line kilns, such as the potter's kiln being used here.

HALIDES

Minerals in this group consist of metals combined with one of the four common halogen elements: fluorine, chlorine, iodine, or bromine. Halides tend to be soft and many crystallize in the cubic system.

COMPOSITION

Compositionally and structurally, there are three broad categories of halide mineral: simple halides, halide complexes, and oxyhydroxy-halides. Simple halides form when a metal combines with a halogen. Halite and fluorite are examples of simple halides. In halide complexes, the halide is usually bound to aluminum, creating a molecule that behaves as a single unit, which is in turn bound to a metal. For example, in cryolite, fluorine and aluminum are bound to sodium. Oxyhydroxy-halides are very rare. Atacamite is an example of these halides.

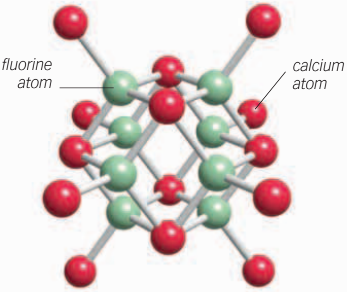

Fluorite crystal structure
In the crystal structure of fluorite, every calcium atom is coordinated with eight fluorine atoms at the corners of a cube. This yields cubic crystals.

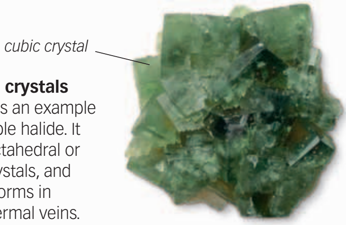

Fluorite crystals
Fluorite is an example of a simple halide. It forms octahedral or cubic crystals, and usually forms in hydrothermal veins.

OCCURRENCE

Many halides occur in evaporite deposits. Others occur in hydrothermal veins or form when halide-bearing waters act upon the oxidation products of other minerals.

USES

Halides are important industrial minerals. Halite, or table salt, is the classic example. Other halides are used as fertilizers, in glass making and metal refining, and as minor ores.

Fertilizing crops
The halides sylvite and carnallite are important sources of potash for fertilizers. Potash reduces many diseases, rot, and mildew of food plants.

Salt Lake
Thick, white crusts of the halide mineral halite encrust rocks along the edge of the Great Salt Lake, Utah, USA.

Granular carnallite
This granular mass of carnallite has a red color due to inclusions of hematite.

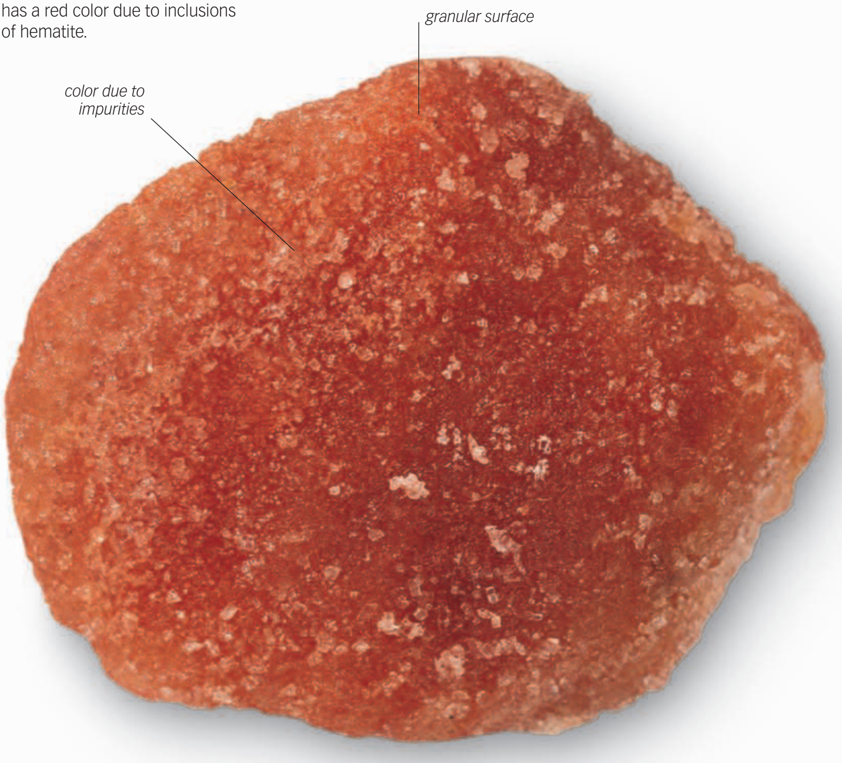

PROFILE

Orthorhombic

2½

1.6

None

Conchoidal

White

Greasy

$KMgCl_3 \cdot 6H_2O$

CARNALLITE

First discovered in Germany, carnallite was named after Rudolph von Carnall, a Prussian mining engineer, in 1856. It is usually white or colorless but may appear reddish or yellowish depending on the presence of hematite (p.91) or goethite (p.102) impurities. Hydrated potassium and magnesium chloride, carnallite is generally massive to granular in habit. Crystals are rare because they absorb water from the air and dissolve. When found, crystals are thick and tabular, pseudohexagonal, or pyramidal.

Carnallite forms in the upper layers of marine evaporite salt deposits, where it occurs with other potassium and magnesium evaporite minerals. The mineral is Russia's most important source of magnesium. Caustic potash, a potassium hydroxide, is produced from carnallite.

Potash fertilizer
Carnallite is an important source of potash, which is used in fertilizers, and caustic potash.

Cryolite crystals
This mass of translucent cryolite crystals on rock has patches of siderite on it.

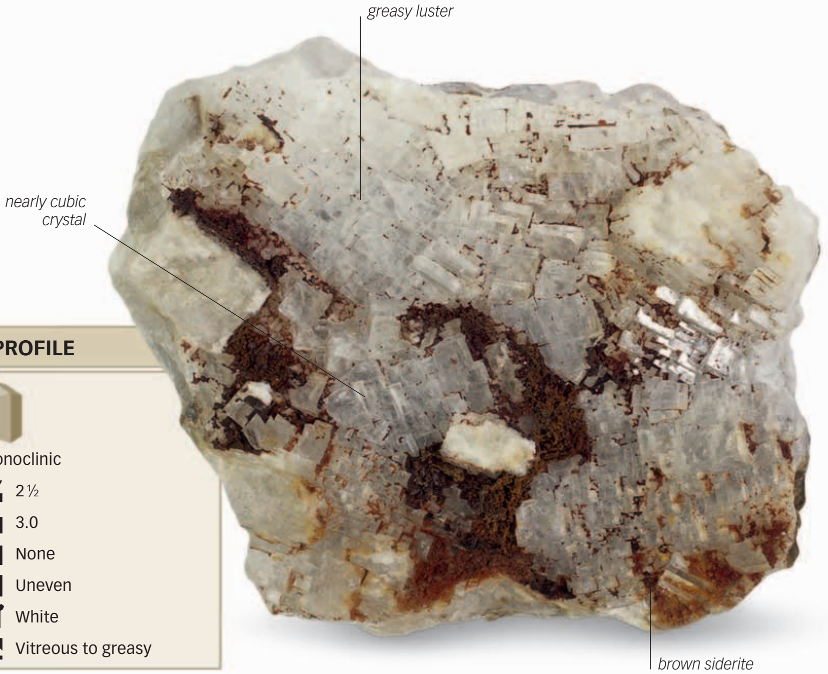

PROFILE

Monoclinic
2½
3.0
None
Uneven
White
Vitreous to greasy

VARIANT

pseudocubic outline
greasy luster

Massive cryolite specimen
A close-up of a massive cryolite fragment

Na_3AlF_6

CRYOLITE

Few people have heard of cryolite, but it is one of the most important minerals of our age. Aircraft could not fly without it, and modern engineering of all kinds would be stunted in its absence. Synthetic cryolite is an essential ingredient in aluminum production. The mineral takes its name from the Greek terms *kryos* and *lithos*, which mean "ice" and "stone"—an allusion to its translucent, icelike appearance. Cryolite is usually colorless or white. Rarely it can be brown, yellow, reddish brown, or black. It occurs commonly as coarse, granular, or massive aggregates and rarely, as pseudocubic crystals.

Cryolite forms mainly in certain granites (pp.258–59) and granitic pegmatites (p.260). The largest deposit of cryolite, at Ivigtut, Greenland, is now exhausted. Lesser amounts are found in Spain, Russia, and the USA.

Cryolite in aviation
Synthetic cryolite is used to separate aluminum—an indispensable metal in aviation—from its ores.

PROFILE

Cubic
4
3.0–3.3
Perfect octahedral
Flat conchoidal
White
Vitreous

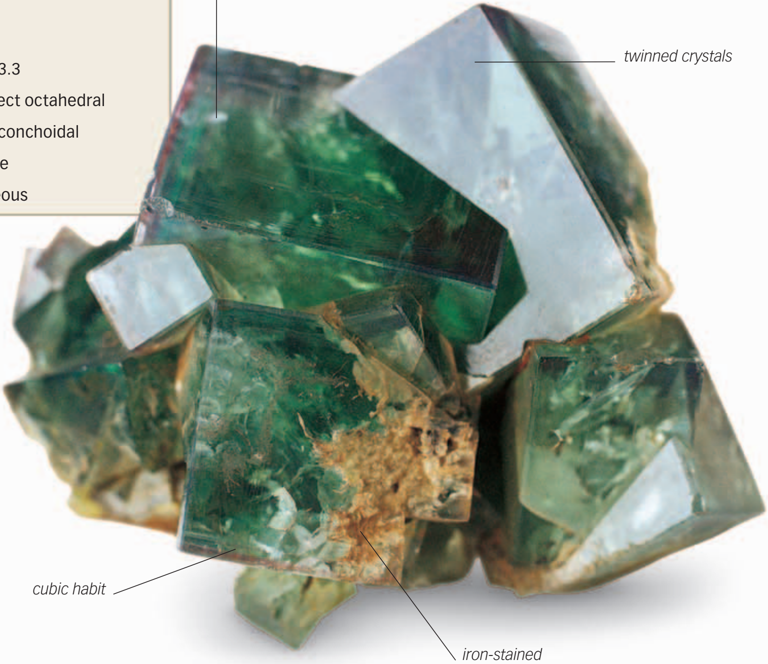

Twinned crystals This group of fine, green cubic fluorite crystals exhibits the classic penetration twins, which are typical of this mineral.

VARIANTS

Pink octahedron An octahedral crystal of rare pink fluorite

Yellow fluorite A group of bright yellow fluorite cubes

Near-white fluorite A group of unusually colorless cubes

CaF_2

FLUORITE

An important industrial mineral, fluorite used to be known as fluorspar. The name fluorite comes from the Latin word *fluere*, which means "to flow"—a reference to its use in iron smelting to improve the fluidity of slags and the refining of metals. Fluorite commonly occurs as vibrant, well-formed crystals. A single crystal may have zones of different colors that follow the contour of the crystal faces. Fluorite crystals are widely found in cubes, while fluorite octahedra—which are often twinned—are much less common. The mineral can also be massive, granular, or compact.

Fluorite occurs in hydrothermal deposits and as an accessory mineral in intermediate intrusive and silica-rich rocks. It is used in the manufacture of high-octane fuels and steel and in the production of hydrofluoric acid.

Blue John Veins of banded purple, white, and yellow fluorite, known as Blue John, are visible in this vessel.

PROFILE

Cubic

2½

2.1–2.6

Perfect cubic

Conchoidal

White

Vitreous

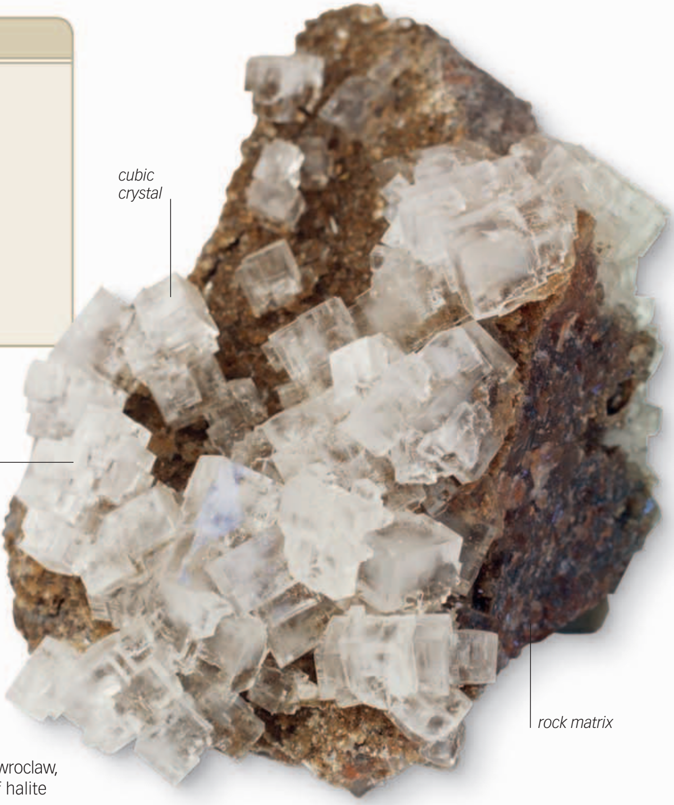

Halite crystals In this specimen from Inowroclaw, Poland, cubic crystals of halite cover a rock matrix.

VARIANTS

Cubic halite Twinned, cubic crystals on rock matrix

Massive halite A specimen of massive pink halite

Blue halite Unusual blue cubic halite on rock

NaCl

HALITE

Culinary rock salt is actually halite. Its name is derived from the Greek word *hals*, which means “salt.” Most halite is colorless, white, gray, orange, or brown, but it can also be bright blue or purple. The orange color comes from inclusions of hematite (p.91), while the blue and purple colors indicate defects in the crystal structure. Halite is commonly found in massive and bedded aggregates as rock salt. It also occurs in coarse, crystalline masses or in granular and compact forms.

Halite crystals are usually cubic. Sometimes, halite may form “hopper” crystals—in which the outer edges of the cube faces have grown more rapidly than their centers, leaving cavernous faces. It is widespread in saline evaporite deposits.

Table salt Mined since ancient times and also used as a currency, common table salt is the mineral halite.

PROFILE

Cubic

- 2½
- 2.0
- Perfect cubic
- Uneven
- White
- Vitreous

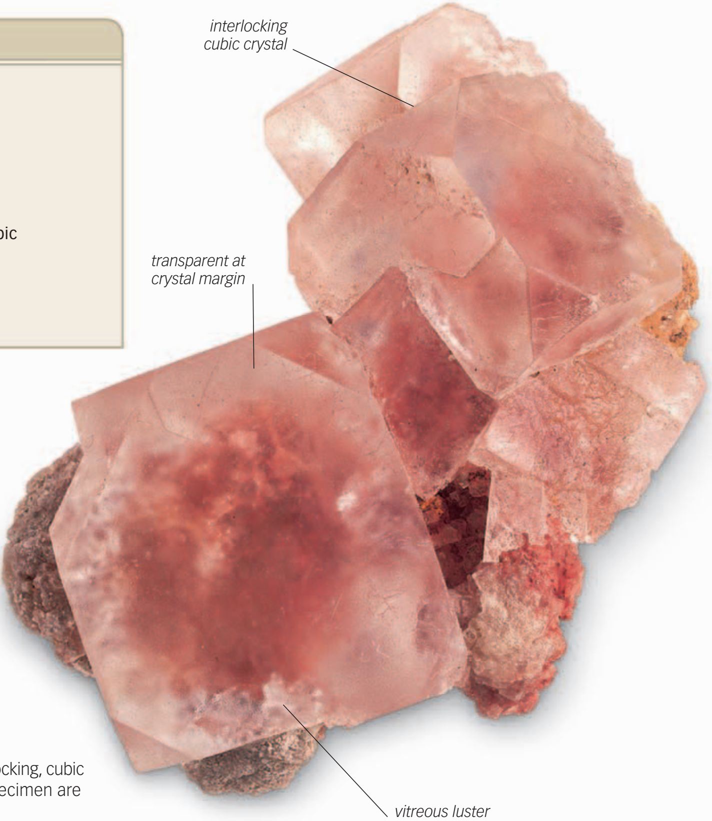

Sylvite crystals The pinkish, interlocking, cubic crystals in this specimen are typical of sylvite.

VARIANT

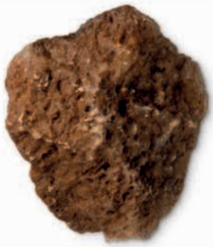

Sylvite in potash A specimen of massive potash containing the mineral sylvite

KCl

SYLVITE

Millions of tons of sylvite are mined annually for the manufacture of potassium compounds, such as potash fertilizers. Sylvite is also used to manufacture metallic potassium. The mineral was first discovered in 1823 on Mount Vesuvius, Italy, where it occurs as encrustations on lava. The name sylvite comes from its Latin medicinal name, *sal digestivus Sylvii*, which means "digestive salt," and it is also known as sylvine. Usually colorless to white or grayish, sylvite can be tinged blue, yellow, purple, or red. Sylvite crystals are cubic, octahedral, or both. It commonly occurs as crusts and as columnar, granular, or massive aggregates.

Sylvite is found in thick beds either mixed or interbedded with halite (p.110), gypsum (p.136), and other evaporite minerals, although it is rarer than halite.

Sylvite fertilizer Crushed potash, as seen here, is used as a fertilizer and comes from the mineral sylvite.

PROFILE

Tetragonal

1–2

6.5

Distinct

Conchoidal

Pale yellow-white

Adamantine

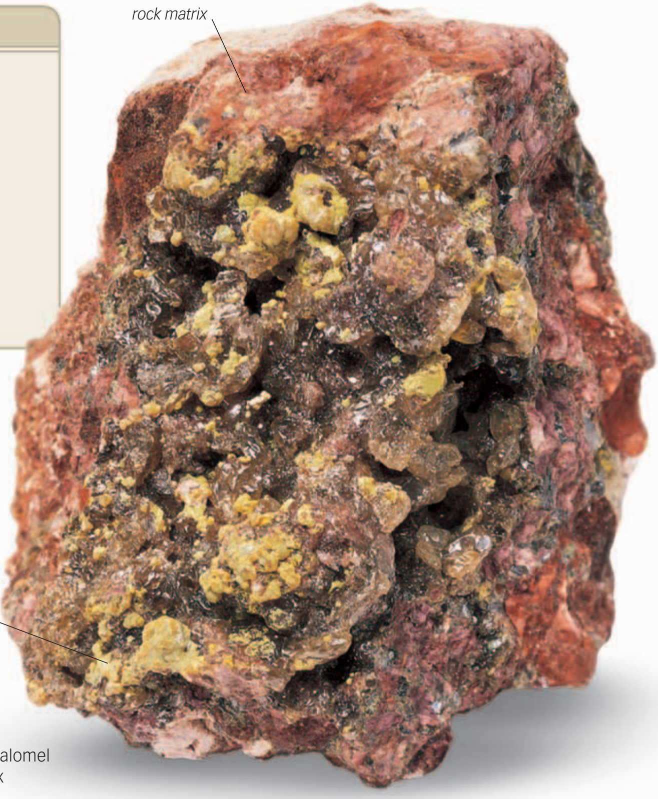

Calomel encrustations A thin crust of yellowish calomel crystals coats a rock matrix in this specimen.

VARIANT

black calomel crystal

Calomel in matrix Crystals of black calomel in a rock matrix

HgCl

CALOMEL

A mercury chloride, calomel takes its name from two Greek words: *ómorfi*, which means “beautiful,” and *méli*, which means “honey”—an allusion to its sweet taste, although it is, in fact, toxic. Calomel is also referred to as horn quicksilver and horn mercury. Specimens are soft, heavy, and plasticlike, with crystals that are pyramidal, tabular, or prismatic, often with complex twinning. Calomel is also found as crusts and can be massive and earthy. It fluoresces brick red.

Calomel occurs as a secondary mineral in the oxidized zones of mercury-bearing deposits, together with native mercury, cinnabar (p.56), goethite (p.102), and calcite (p.114). It was used as a laxative and a disinfectant as well as in the treatment of syphilis from the 16th century until the early 20th century, when the toxic effect of its mercury component was discovered. Calomel’s use as a teething powder in Britain was suspended only in 1954, following widespread poisoning. It is still used as an ore of mercury and in insecticides and fungicides.

CARBONATES

There are approximately 80 known carbonate minerals. Most of them are rare, but the common carbonates calcite and dolomite are major rock-forming minerals. Carbonates form rhombohedral crystals and are soft, soluble in hydrochloric acid, and often vividly colored.

COMPOSITION

All carbonates contain the carbonate group CO$_3$ as the basic compositional and structural unit. This group has a carbon atom in the center of an equilateral triangle of oxygen atoms, giving rise to the trigonal symmetry of many carbonate minerals. This basic unit is joined by one or more metals or semimetals such as calcium, sodium, aluminum, manganese, barium, zinc, and copper.

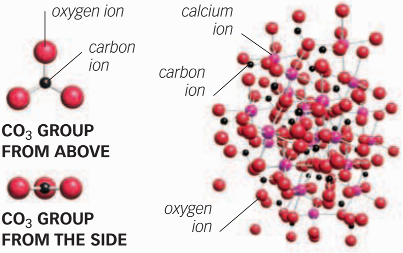

Calcite crystal sructure
In calcite, three oxygen ions surround each carbon ion in a CO$_3$ group. Each calcium ion combines with six oxygen ions to form an octahedron.

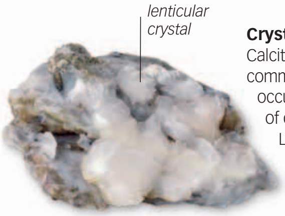

Crystalline calcite
Calcite is the most common carbonate and occurs in a wide range of crystalline forms. Lenticular and scalenohedral crystals are seen in this specimen.

OCCURRENCE

Calcite and dolomite are found in sediments such as chalk and limestone. They also occur in seashells and coral reefs, in evaporate deposits, and in metamorphic rocks, such as marble. Other carbonates, such as rhodochrosite, azurite, and malachite, are principally secondary minerals.

USES

The carbonate minerals calcite and dolomite are important in the manufacture of cement and building stone. Other carbonates find uses as ores of metals: witherite of barium; strontianite of strontium; siderite of iron; rhodochrosite of manganese; smithsonite of zinc; and cerussite of lead.

intricate growth pattern

Malachite jewel box
The copper carbonate malachite has been a favorite carving stone for three millennia.

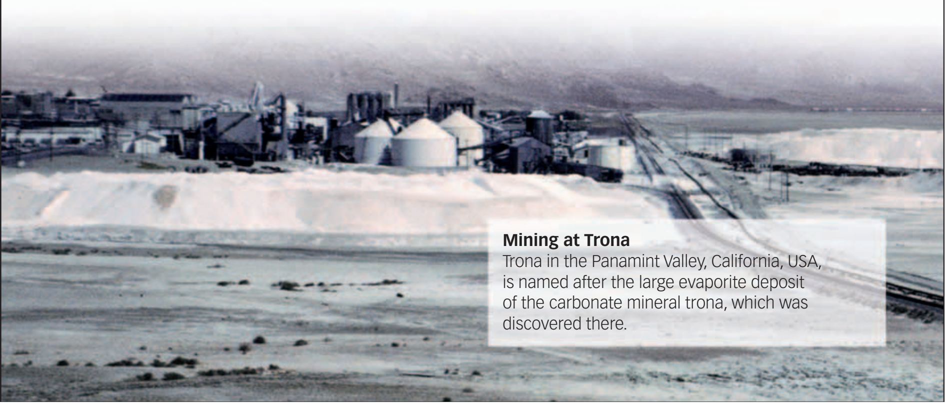

Mining at Trona
Trona in the Panamint Valley, California, USA, is named after the large evaporite deposit of the carbonate mineral trona, which was discovered there.

PROFILE

Hexagonal
3
2.7
Perfect rhombohedral
Subconchoidal, brittle
White
Vitreous

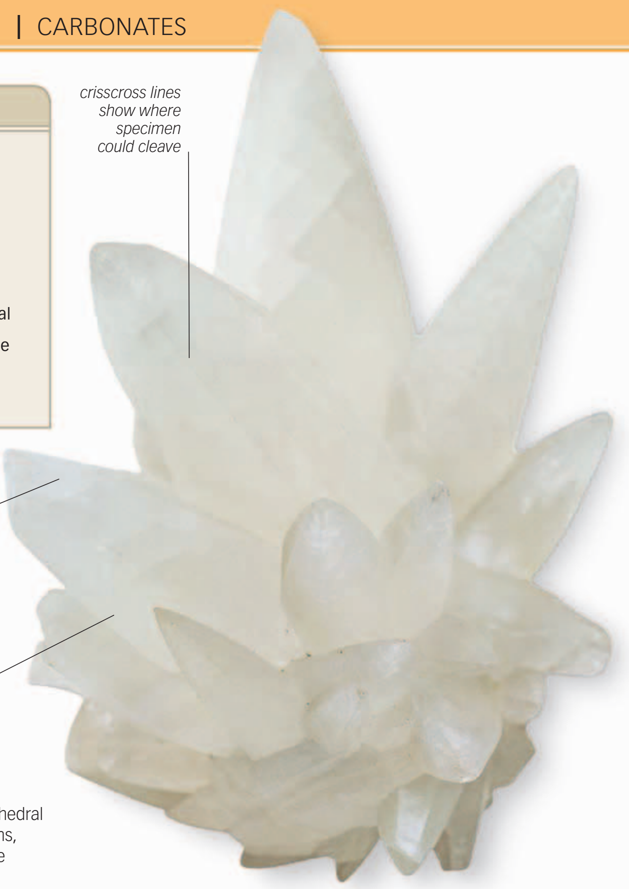

crisscross lines show where specimen could cleave
vitreous luster
group of dogtooth crystals

Dogtooth spar
Crystals with steep, rhombohedral or scalenohedral terminations, such as in this specimen, are known as dogtooth spar.

VARIANTS

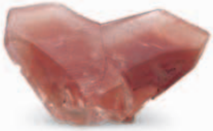

Butterfly twin A twinned, pink crystal of calcite

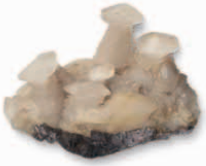

Nailhead spar A rhombohedral calcite crystal on galena

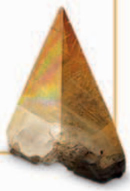

Scalenohedron A single scalenohedral calcite crystal

$CaCo_3$

CALCITE

The most common form of calcium carbonate, calcite is known for the variety and beautiful development of its crystals. These occur most often as scalenohedra and are commonly twinned, sometimes forming heart-shaped, butterfly twins. Crystals with rhombohedral terminations are also common; those with shallow rhombohedral terminations are called nailhead spar. Highly transparent calcite is called optical spar.

Although calcite can form spectacular crystals, it is usually massive, occurring either as marble (p.301) or as limestone (p.319). It is also found as fibers, nodules, stalactites, and earthy aggregates. Calcite specimens can occur in metamorphic deposits, igneous rocks, and hydrothermal veins.

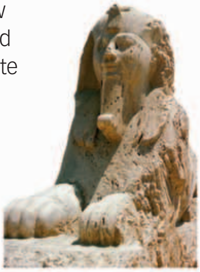

Alabaster sphinx
Virtually all ancient Egyptian “alabaster,” such as that used to make this small sphinx, was actually calcite.

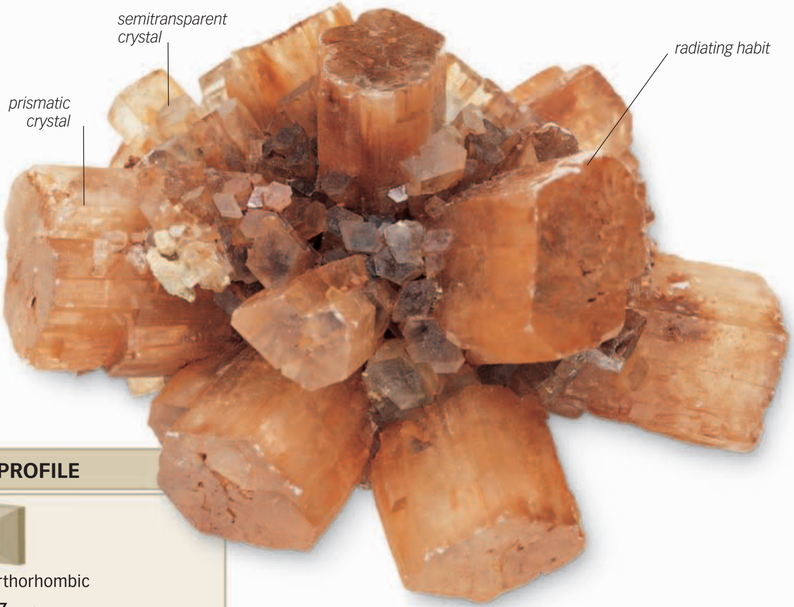

PROFILE

Orthorhombic

3½–4

2.9

Distinct

Subconchoidal, brittle

White

Vitreous inclining to resinous

Pseudohexagonal crystals This specimen consists of a radiating group of prismatic, semitransparent, pseudohexagonal, twinned aragonite crystals.

VARIANTS

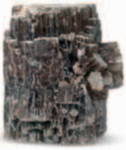

Intergrown crystals A mass of pseudohexagonal crystals of aragonite

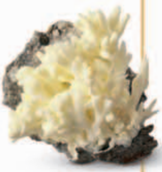

Flos ferri Coral-like aragonite crystals on rock matrix

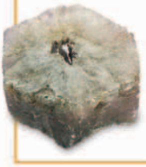

Cyclic twin A classic aragonite cyclic twin from Spain

$CaCo_3$

ARAGONITE

Although aragonite has the same chemical composition as calcite (p.114), its crystals are different. They are tabular, prismatic, or needlelike, often with steep pyramidal or chisel-shaped ends, and can form columnar or radiating aggregates. Multiple twinned crystals are common, appearing hexagonal in shape. Although aragonite sometimes looks similar to calcite, it is easily distinguished by the absence of rhombohedral cleavage. Specimens can be white, colorless, gray, yellowish, green, blue, reddish, violet, or brown.

Aragonite is found in the oxidized zones of ore deposits and in evaporites, hot spring deposits, and caves. It is also found in some metamorphic and igneous rocks. Banded stalactitic aragonite can be polished as an ornamental stone.

Mother of pearl Aragonite is also produced by some living animals. It is seen here forming the inner layer of a marine mollusk shell.

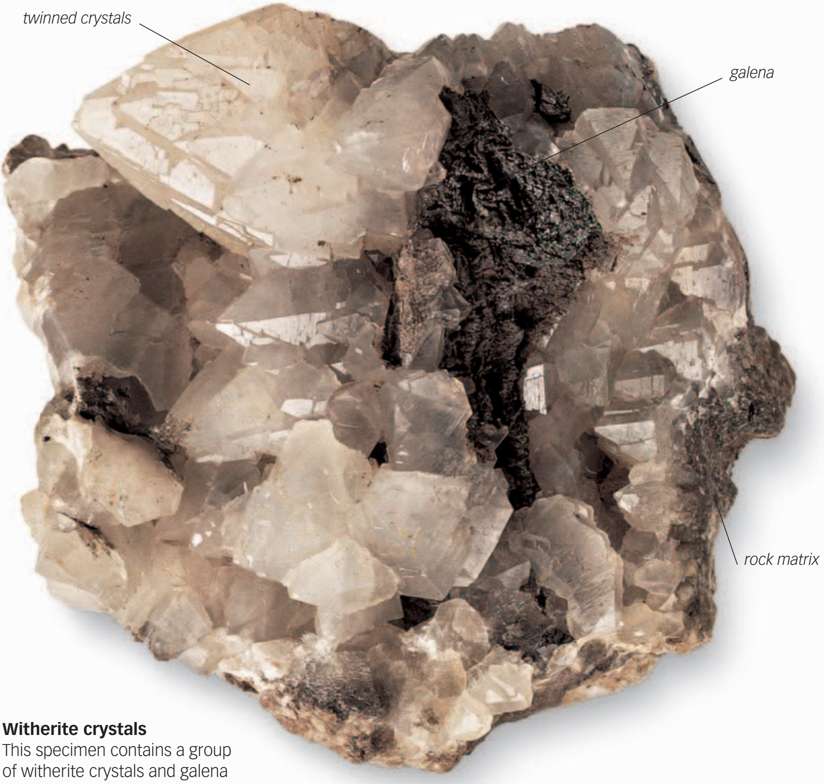

Witherite crystals
This specimen contains a group of witherite crystals and galena on a rock matrix.

PROFILE

Orthorhombic
3–3½
4.3
Distinct, imperfect
Uneven, brittle
White
Vitreous

$BaCo_3$

WITHERITE

This barium carbonate was named in 1790 after the English mineralogist William Withering. Witherite is white, colorless, or tinged yellow, brown, or green. Its crystals are always twinned, either as prisms which appear hexagonal in shape, or as pyramids, which are frequently paired. They can also be short to long prismatic or tabular and may have striations running across the prism faces. Witherite can also be fibrous, botryoidal, spherular, columnar, granular, or massive.

Most witherite comes from hydrothermal veins formed at low temperatures (up to 400°F/200°C), usually resulting from the alteration of baryte (p.134). Specimens feel relatively heavy for their size due to the presence of the high-density element barium. Witherite is preferred over the commonly found barium mineral barite for the preparation of other barium compounds because it is more soluble in acids. These compounds are used in case-hardening steel, in copper refining, in sugar refining, in vacuum tubes, and in many other applications.

PROFILE

Hexagonal

3½–4

2.8–2.9

Perfect rhombohedral

Subconchoidal

White

Vitreous

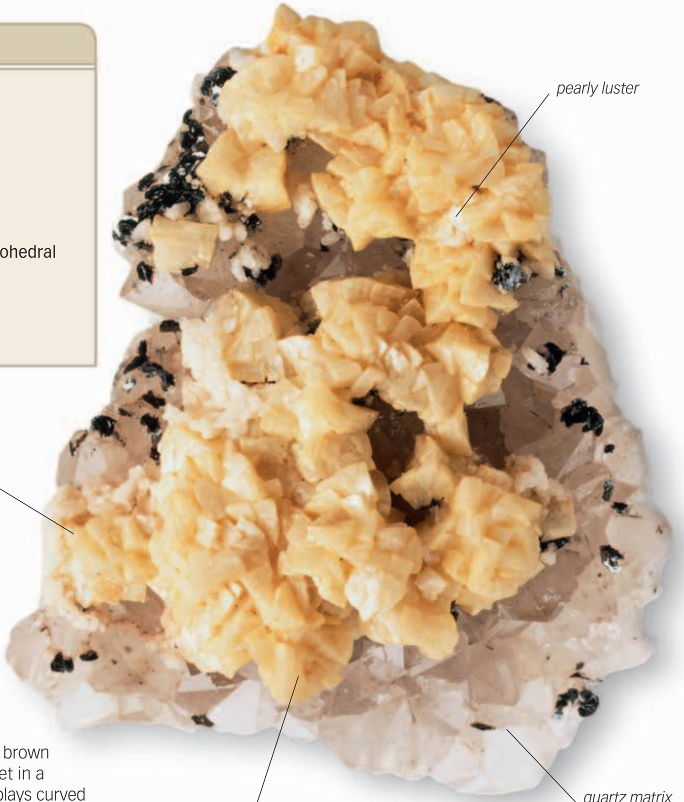

Pink dolomite This specimen of pale brown dolomite crystals is set in a quartz matrix and displays curved crystal faces.

VARIANT

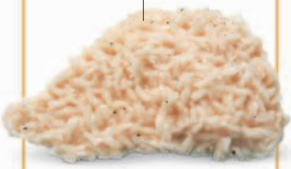

Tabular crystals Pink, saddle-shaped, tabular crystals of dolomite in a crustlike form

$CaMg(CO_3)_2$

DOLOMITE

An important rock-forming mineral, dolomite is named after the French mineralogist Déodat Gratet de Dolomieu. It is a colorless to white, pale brown, grayish, reddish, or pink mineral. Its crystals are commonly rhombohedral or tabular, often have curved faces, and sometimes cluster in saddle-shaped aggregates. Dolomite may be striated horizontally and twinned. Some crystals may be up to 2 in (5 cm) long. It can also be coarse to fine granular, massive, and, rarely, fibrous.

Dolomite is the main constituent in dolomite rocks and dolomitic marbles. It occurs as a replacement deposit in limestone (p.319) affected by magnesium-bearing solutions, in talc schists, and in other magnesium-rich metamorphic rocks. Dolomite is found in hydrothermal veins associated with lead, zinc, and copper ores. It is also found in altered, silica-poor igneous rocks, in some carbonatites (p.272), and in serpentinites (p.298). Crystals of dolomite frequently form in cavities in limestone and marble (p.301).

PROFILE

Hexagonal

4

3.0

Perfect rhombohedral

Conchoidal, brittle

White

Vitreous

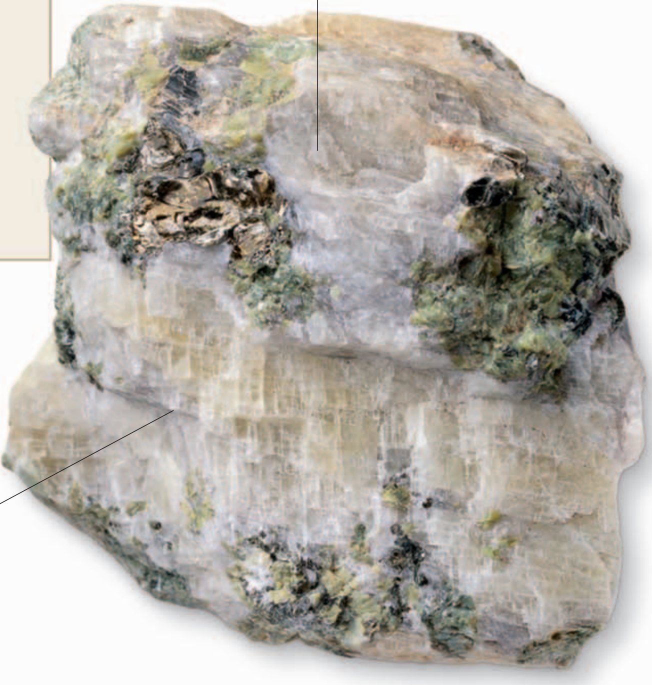
coarse crystalline habit

perfect rhombohedral cleavage

Magnesite crystals This crystalline mass of magnesite is on a rock matrix.

VARIANT

Rhombohedral magnesite Rare, rhombohedral crystals of the colorless form of magnesite

$MgCO_3$

MAGNESITE

This carbonate of magnesium takes its name from its magnesium component. It is generally massive, lamellar, fibrous, chalky, or granular. Distinct crystals are rare, but when found they are either rhombohedral or prismatic. Most commonly white or light gray, magnesite can be yellow or brownish when iron substitutes for some of the magnesium.

Magnesite forms mainly as an alteration product in magnesium-rich rocks, such as peridotites (p.266). It can occur as a primary mineral in limestones (p.319) and talc or chlorite schists, in cavities in volcanic rocks, and in oceanic salt deposits. It is also found in some meteorites (pp.335–37). An important source of magnesium, magnesite is used as a refractory material, as a catalyst and filler in the production of synthetic rubber, and in the manufacture of chemicals and fertilizers. Magnesium derived from magnesite is alloyed with aluminum, zinc, or manganese for use in aircraft, spacecraft, road vehicles, and household appliances.

Tabular crystals In this specimen, a mass of tabular cerussite crystals covers a rock matrix.

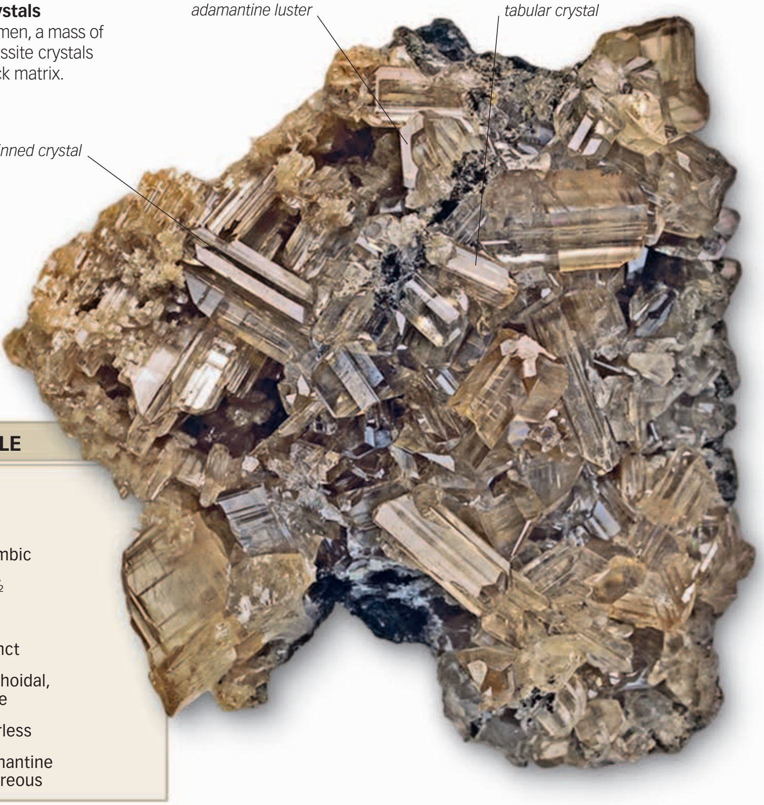

PROFILE

Orthorhombic

3–3½

6.5

Distinct

Conchoidal, brittle

Colorless

Adamantine to vitreous

$PbCO_3$

CERUSSITE

Known since antiquity, cerussite is named after the Latin word *cerussa*, which describes a white lead pigment. After galena (p.54), it is the most common ore of lead. Cerussite is generally colorless or white to gray, but may be blue to green due to copper impurities. Its crystal habits are highly varied. Cerussite forms tabular or pyramidal crystals or, sometimes, twins that may be star-shaped or reticulated (netlike) masses. Fragile aggregates of randomly grown prismatic crystals known as jack-straw cerussite are also common. The adamantine luster of cerussite crystals is particularly bright.

A widespread secondary mineral that occurs in the oxidation zones of lead veins, cerussite is formed by the action of carbonated water on other lead minerals, particularly galena and anglesite (p.132).

Collector's gem Faceted cerussite stones, such as this rare gem, are brilliant but too soft to be worn.

VARIANTS

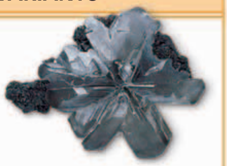

Cyclic twin A star-shaped, crystalline specimen of cerussite from Zambia

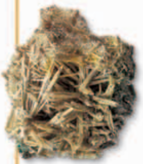

Jack-straw cerussite Delicate, needlelike crystals of jack-straw cerussite

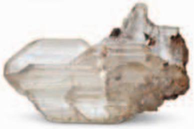

Prismatic crystal A striated, colorless, prismatic crystal of cerussite

PROFILE

Monoclinic

- 3½–4
- 3.8
- Perfect
- Conchoidal, brittle
- Blue
- Vitreous to dull earthy

Large crystals In this specimen of azurite, large, well-formed crystals rest on a goethite matrix.

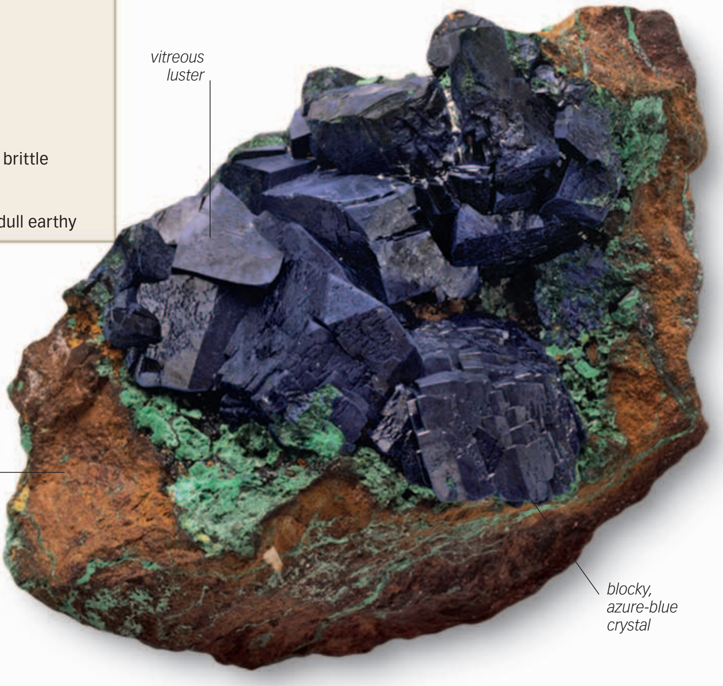

VARIANTS

Bladed crystal A single, bladed azurite crystal

Tabular crystals Thin, parallel azurite crystals on a rock matrix

Radiating crystals A spherical concretion of azurite

$Cu_3(CO_3)_2(OH)_2$

AZURITE

A deep blue copper carbonate hydroxide, azurite was used as a blue pigment in 15th- to 17th-century European art and probably in the production of blue glaze in ancient Egypt. It takes its name from the Persian word *lazhuward*, which means "blue." Azurite forms either tabular or prismatic crystals with a wide variety of habits. Tabular crystals commonly have wedge-shaped terminations. Azurite forms rosette-shaped crystalline aggregates or occurs in massive, stalactitic, or botryoidal forms. Well-developed crystals are dark azure blue in color, but massive or earthy aggregates may be paler.

Azurite is a secondary mineral formed in the oxidized portions of copper deposits. Massive azurite used for ornamental purposes is sometimes called chessylite, after Chessy, France.

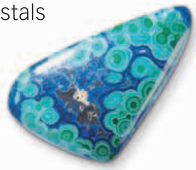

Cabochon gemstone This cabochon exhibits the vivid blue color of azurite and the green color of malachite.

PROFILE

Hexagonal
3½–4
3.6
Perfect rhombohedral
Uneven
White
Vitreous to pearly

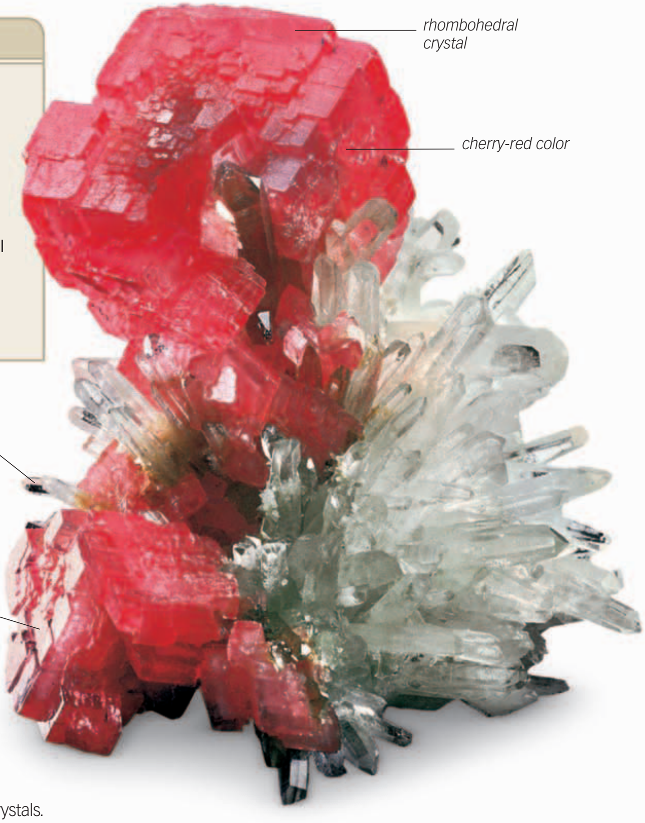

Spectacular crystal This group of rhodochrosite rhombohedrons from Peru is perched on radiating quartz crystals.

VARIANTS

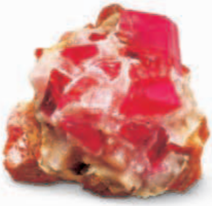

Classic crystals Rhombohedral rhodochrosite in classic rose-pink color

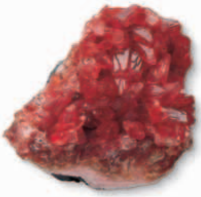

Red rhodochrosite Bright, cherry-red color typical of many manganese minerals

$MnCO_3$

RHODOCHROSITE

A prized collectors' mineral, rhodochrosite is a manganese carbonate. It was given its name—derived from the Greek *rhodokhros*, which means "of rosy color"—in 1800. Rhodochrosite has a classic rose-pink color, but specimens can also be brown or gray. It forms dogtooth or rhombohedral crystals like calcite (p.114), but it may also occur in stalactitic, granular, nodular, botryoidal, and massive habits.

Rhodochrosite is found in hydrothermal ore veins with sphalerite (p.53), galena (p.54), fluorite (p.109), and manganese oxides. It also occurs in metamorphic deposits and as a secondary mineral in sedimentary manganese deposits. Abundant at Butte, Montana, and other localities, rhodochrosite is sometimes mined as an ore of manganese.

Rhodochrosite carvings These two decorative ducks were carved from banded rhodochrosite and white calcite.

Ankerite rhombohedra
This group of ankerite rhombohedra is set in a rock matrix.

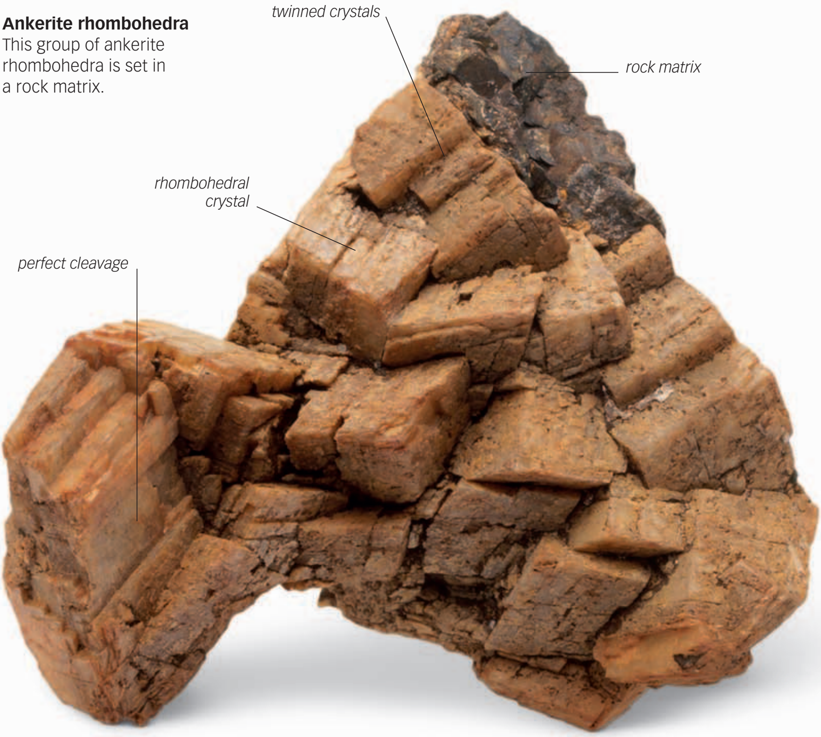

PROFILE

Hexagonal
3½–4
2.9
Perfect
Subconchoidal
White
Vitreous to pearly

Ca(Fe,Mg,Mn,)$(CO_3)_2$

ANKERITE

Considered a rock-forming mineral, ankerite is calcium carbonate with varying amounts of iron, magnesium, and manganese in its structure. It was named in 1825 after the Austrian mineralogist M.J. Anker. Although usually pale buff in color, ankerite can be colorless, white, gray, or brownish. Much ankerite becomes dark on weathering, and many specimens are fluorescent. Ankerite forms rhombohedral crystals similar to those of dolomite (p.117), often with similarly curved faces forming saddle-shaped groups; it can also form prismatic crystals. However, ankerite is more commonly massive or coarsely granular.

Ankerite forms as a secondary mineral from the action of iron- and magnesium-bearing fluids on limestone (p.319) or dolomite rock (p.320). It is a waste mineral in hydrothermal ore deposits and also occurs in carbonatites (p.272), low-grade metamorphosed ironstones, and banded ironstone formations (p.329). Ankerite is also found in iron ore deposits with siderite (p.123).

PROFILE

Hexagonal

3½–4

3.9

Perfect rhombohedral

Uneven or subconchoidal

White

Vitreous to pearly

Rhombohedral crystals This group of well-formed siderite rhombohedra with many twinned crystals rests on a rock matrix.

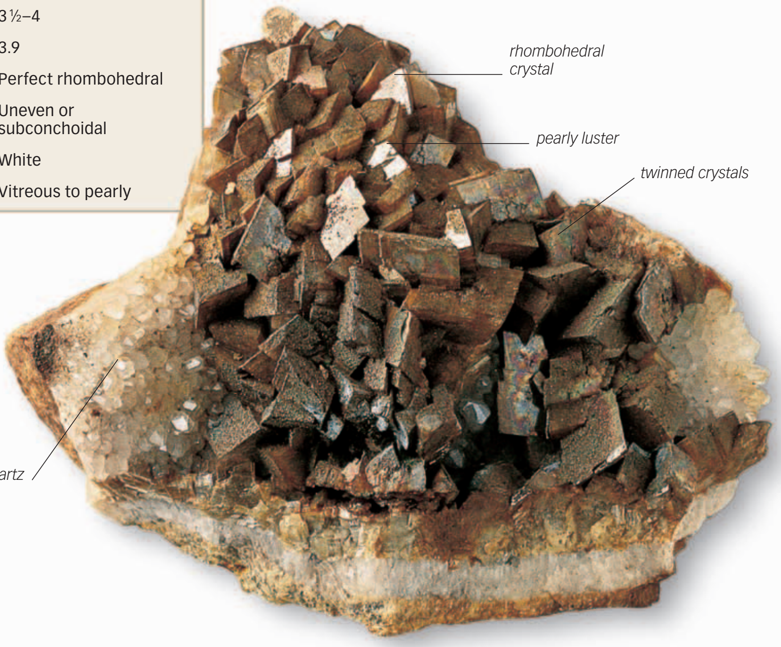

$FeCO_3$

SIDERITE

An ore of iron, **siderite takes its name** from the Greek word *sideros*, which means “iron.” Formerly known as chalybite, siderite can form rhombohedral crystals, often with curved surfaces. The mineral can also form scalenohedral, tabular, or prismatic crystals. Single crystals up to 6 in (15 cm) long are found in Quebec, Canada. However, siderite is more commonly massive or granular and sometimes botryoidal or globular in habit.

A widespread mineral, siderite occurs in igneous, sedimentary, and metamorphic rocks. In sedimentary rocks, siderite occurs in concretions (p.333) and in thin beds with coal (p.253) seams, shale (p.313), and clay. Well-formed crystals are found in hydrothermal metallic veins and in some granitic and syenitic pegmatites (p.260). An outcrop of siderite originally mined for iron by American colonists is still visible at Roxbury, Connecticut. Rare transparent siderite is sometimes cut as gemstones for collectors.

VARIANTS

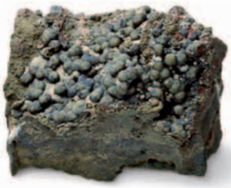

Botryoidal siderite Grapelike siderite bunches on a base of massive siderite

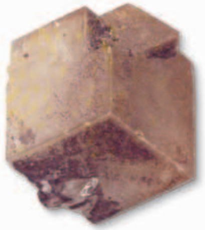

Single crystal A large, rhombohedral, single crystal of siderite

PROFILE

Hexagonal
4–4½
4.4
Perfect rhombohedral
Uneven to conchoidal
White
Vitreous to pearly

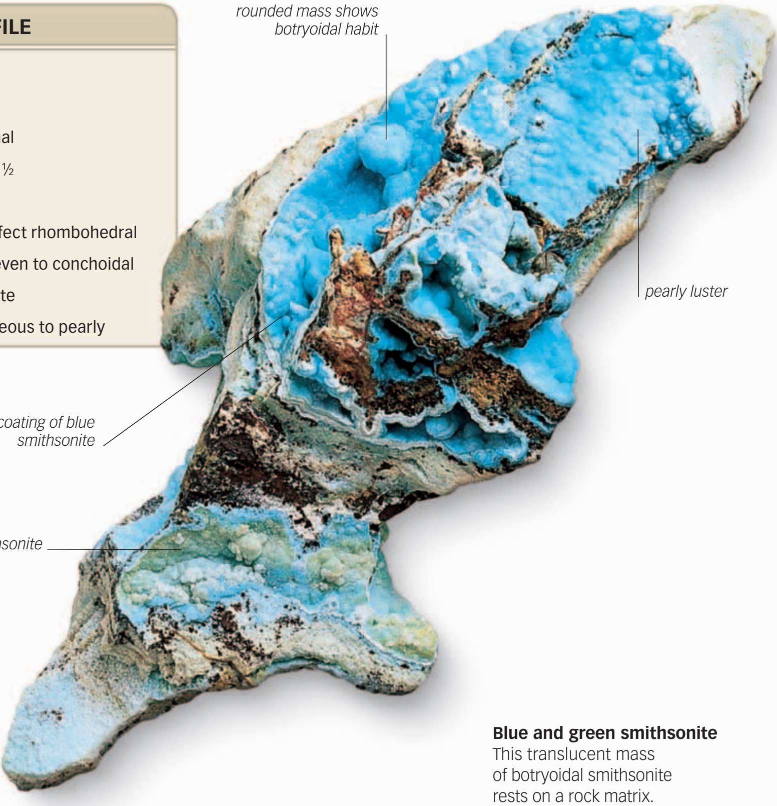
rounded mass shows botryoidal habit
pearly luster
coating of blue smithsonite
green smithsonite

Blue and green smithsonite This translucent mass of botryoidal smithsonite rests on a rock matrix.

VARIANT

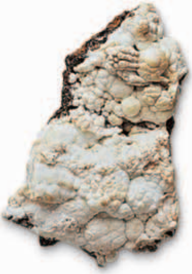

White smithsonite A mass of earthy smithsonite on a rock matrix

$ZnCo_3$

SMITHSONITE

An ore of zinc that continues to be frequently mined, smithsonite may have provided the zinc component of brass in ancient metallurgy. Specimens can be of various colors, such as yellow, orange, brown, pink, lilac, white, gray, green, and blue. Although smithsonite rarely forms crystals, when found, they are prismatic, rhombohedral, or scalenohedral and often have curved faces. A zinc carbonate, smithsonite commonly occurs as massive, botryoidal, spherular, or stalactitic masses, or sometimes, as honeycombed aggregates called dry-bone ore.

Smithsonite is a common mineral, found in the oxidation zones of many zinc ore deposits and in adjacent calcareous rocks. It is often found with malachite (p.125), azurite (p.120), pyromorphite (p.151), cerussite (p.119), and hemimorphite (p.227).

Cabochon Soft smithsonite is occasionally cut into cabochon gemstones for collectors.

Botryoidal malachite
This specimen of malachite on chrysocolla is from Etoile du Congo Mine in Katanga province, Congo.

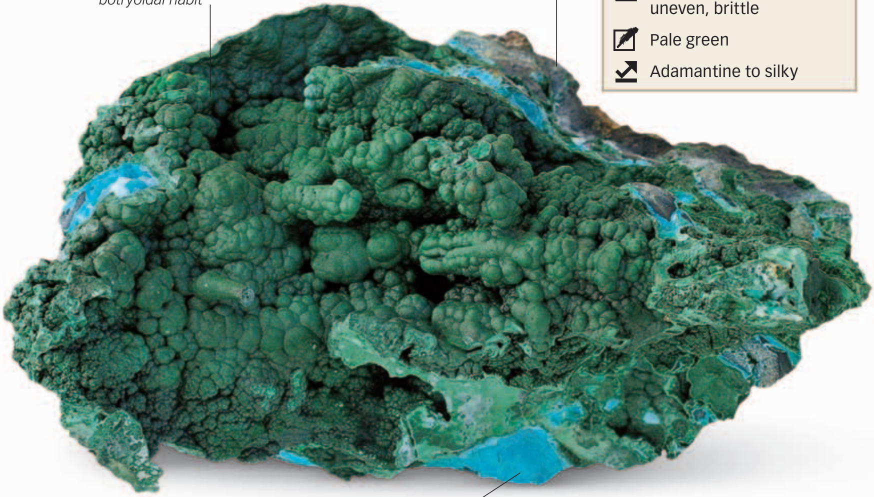

PROFILE

Monoclinic

3½–4

3.9–4.0

Perfect

Subconchoidal to uneven, brittle

Pale green

Adamantine to silky

VARIANTS

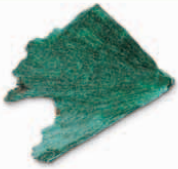

Fibrous malachite
A radiating group of fibrous malachite crystals

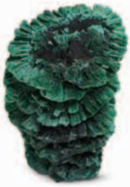

Stalactitic malachite
A group of radiating, fibrous malachite crystals

Malachite section
A section cut through a malachite stalactite

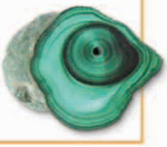

$Cu_2CO_3(OH)_2$

MALACHITE

Possibly the earliest ore of copper, malachite is believed to have been mined in the Sinai and eastern deserts of ancient Egypt from as early as 3000 BCE. Single crystals are uncommon; when found, they are short to long prisms. Malachite is usually found as botryoidal or encrusting masses, often with a radiating fibrous structure and banded in various shades of green. It also occurs as delicate fibrous aggregates and as concentrically banded stalactites.

Malachite occurs in the altered zones of copper deposits, where it is usually accompanied by lesser amounts of azurite (p.120). It is primarily valued as an ornamental material and gemstone. Single masses that weighed up to 51 tons were found in the Ural Mountains of Russia in the 19th century.

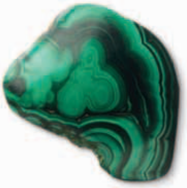

Polished malachite
This specimen of the mineral malachite has been polished to show dark and light color bands.

BORATES

Borate minerals are compounds containing boron and oxygen. Most borate minerals are rare, but a few, such as borax, ulexite, colemanite, and kernite, form large, commercially mined deposits. Borates tend to be soft and either white or colorless.

COMPOSITION

Structurally, boron and oxygen may form a triangle (BO_3) or a tetrahedron (BO_4), each with a boron atom. These structures act as a single chemical unit that bonds to a metal, such as sodium in borax and calcium in colemanite. Borates tend to contain water molecules or a hydroxyl (OH) group, which acts as a chemical unit bonded into their structure. Some borates contain both.

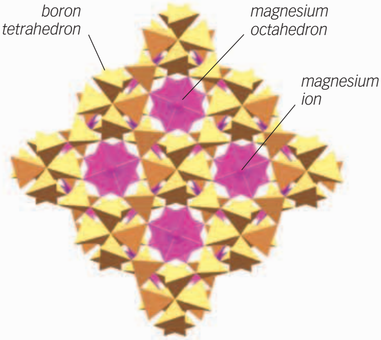

Crystal structure of boracite
In boracite, densely packed boron tetrahedra combine with the metals magnesium and iron (not shown). The borate radical is in the form of tetrahedra.

silky luster
crystals have translucent ends

Ulexite
This is a classic evaporite borate. A hydrous sodium calcium borate, ulexite can form parallel, fibrous crystals that act as fiberoptics when viewed from an end.

OCCURRENCE AND USES

Borates appear in two geologic environments. In the first, borate-bearing solutions that result from volcanic activity flow into a closed basin, where evaporation takes place. Basin deposits usually occur in desert regions, such as the Mojave Desert and Death Valley in California. Borax, ulexite, and colemanite occur in these evaporate deposits. In the second environment, borate minerals are formed as a result of rocks being altered by heat and pressure at relatively high temperatures (1,065°F/575°C or above).

Borates are used as pottery glazes, solvents for metal-oxide slags in metallurgy, welding fluxes, fertilizer additives, soap supplements, and water softeners.

Fireworks
Boron carbide is used to give a green color to fireworks, in place of the toxic barium compounds that were once used.

Borax crust
This crust formed at the edge of Searles Lake in southern California, USA, is principally composed of borax, a borate produced by evaporation.

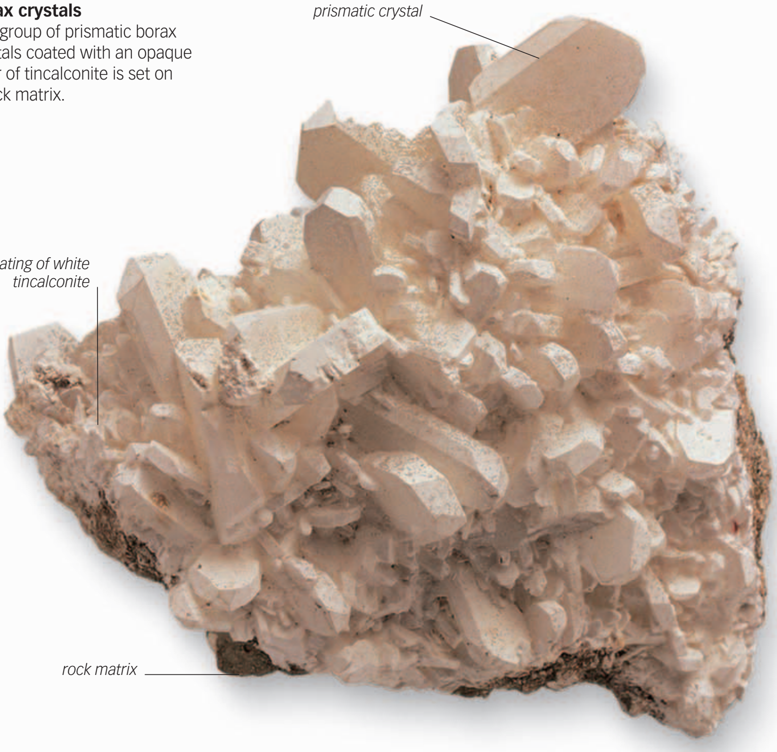

Borax crystals
This group of prismatic borax crystals coated with an opaque layer of tincalconite is set on a rock matrix.

PROFILE

Monoclinic
2–2½
1.7
Perfect, imperfect
Conchoidal
White
Vitreous to earthy

$Na_2B_4O_5(OH)_4 \cdot 8H_2O$

BORAX

An important source of boron, borax has been mined since ancient times. A hydrated sodium borate, borax's colorless crystals dehydrate in air to become the chalky mineral tincalconite. Specimens can also be white, gray, pale green, or pale blue. Borax has short prismatic to tabular crystals, although in commercial deposits it is predominantly massive.

Borax is an evaporite formed in dry lake beds with halite (p.110) and other borates and evaporite sulfates and carbonates. It is used in metal-casting and steel-making. Molten borax beads were historically used to test the composition of other minerals—powdered minerals were fused with the beads, and color change in the beads revealed what the minerals contained.

Boron soap
Compounds derived from borax and, to a lesser extent, ulexite, are key components of many soaps.

PROFILE

Monoclinic

2½

2.0

Perfect

Uneven

White

Vitreous to silky

Ulexite slice This ulexite specimen has a fibrous structure and has been sliced and polished to show its fiberoptic effect.

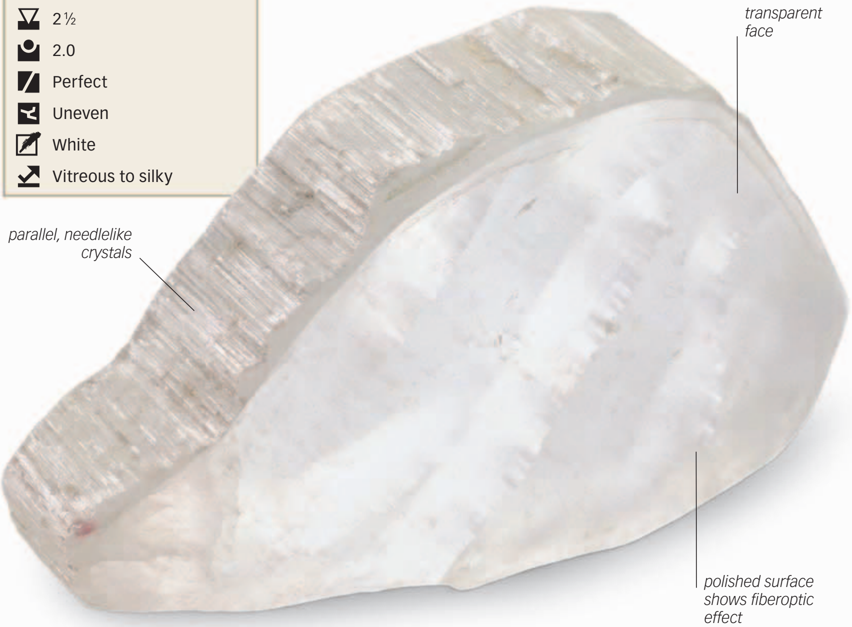
transparent face
parallel, needlelike crystals
polished surface shows fiberoptic effect

VARIANT

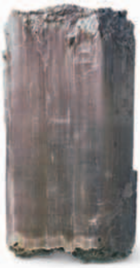

Fibrous crystals Parallel, fibrous ulexite crystals with a silky luster

$NaCaB_5O_6(OH)_6 \cdot H_2O$

ULEXITE

An important economic borate mineral, ulexite is named after the German chemist George Ludwig Ulex, who determined its composition in 1850. It is either colorless or white and has a number of habits. It is commonly found in nodular, rounded, or lenslike crystal aggregates, which often resemble balls of cotton. Less commonly, ulexite is found in dense veins of parallel fibers known as television stone because the fibers transmit light from one end of the crystal to the other. Ulexite also occurs in radiating or compact aggregates of crystals.

Ulexite is found in playa lakes and other evaporite basins in deserts, where it is derived from hot, boron-rich fluids. The mineral commonly occurs with colemanite (p.130), anhydrite (p.133), and glauberite (p.141).

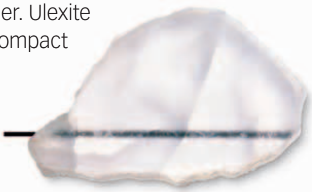

Television stone An unusual property of the form of ulexite shown above is its ability to "transmit" images.

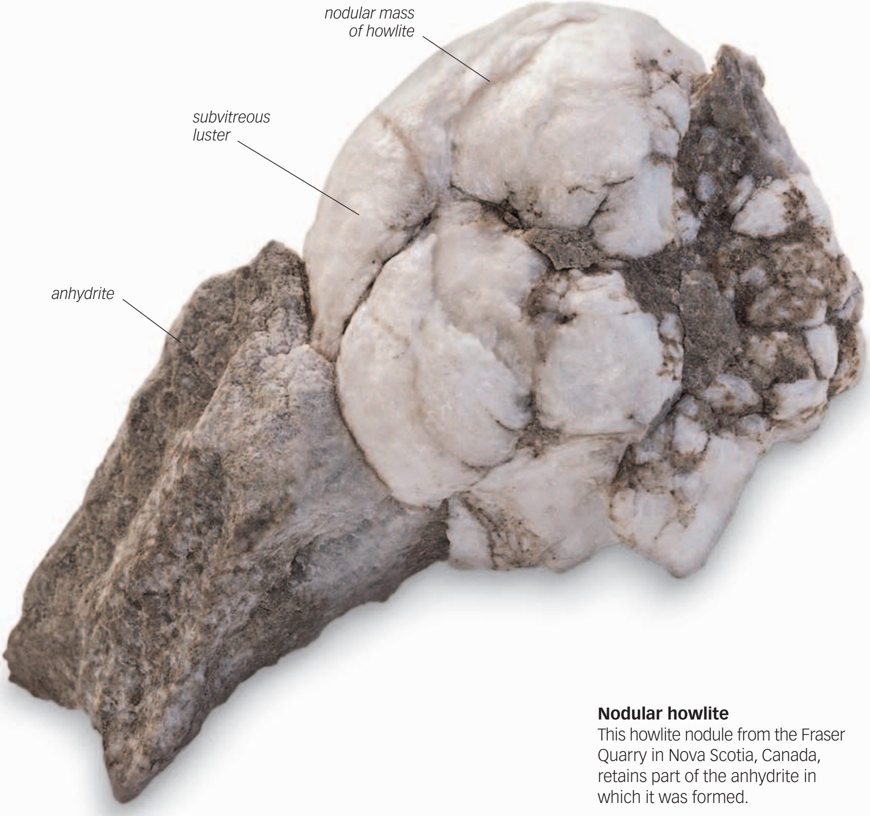

Nodular howlite
This howlite nodule from the Fraser Quarry in Nova Scotia, Canada, retains part of the anhydrite in which it was formed.

PROFILE

Monoclinic
3½
2.6
None
Conchoidal to uneven
White
Subvitreous

$Ca_2B_5SiO_9(OH)_5$

HOWLITE

Named in 1868 after the Canadian chemist, geologist, and mineralogist Henry How, howlite is a calcium borosilicate hydroxide. It generally forms cauliflowerlike nodular masses. The nodules are white, with fine gray or black veins of other minerals running across in an erratic, often weblike pattern. Crystals are rare, but when found they are tabular and seldom exceed $\frac{3}{8}$ in (1 cm) in length. When dyed, howlite specimens resemble and are sometimes sold as turquoise (p.154), although they are easily distinguished by their inferior hardness and lighter color.

Howlite usually occurs associated with other boron minerals, such as kernite and borax (p.127). It is easily fused and is used to make carvings, jewelry components, and other decorative items.

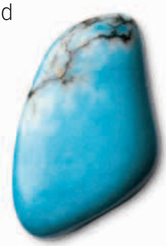

Stained howlite
This tumble-polished and dyed or stained piece of howlite looks similar to turquoise.

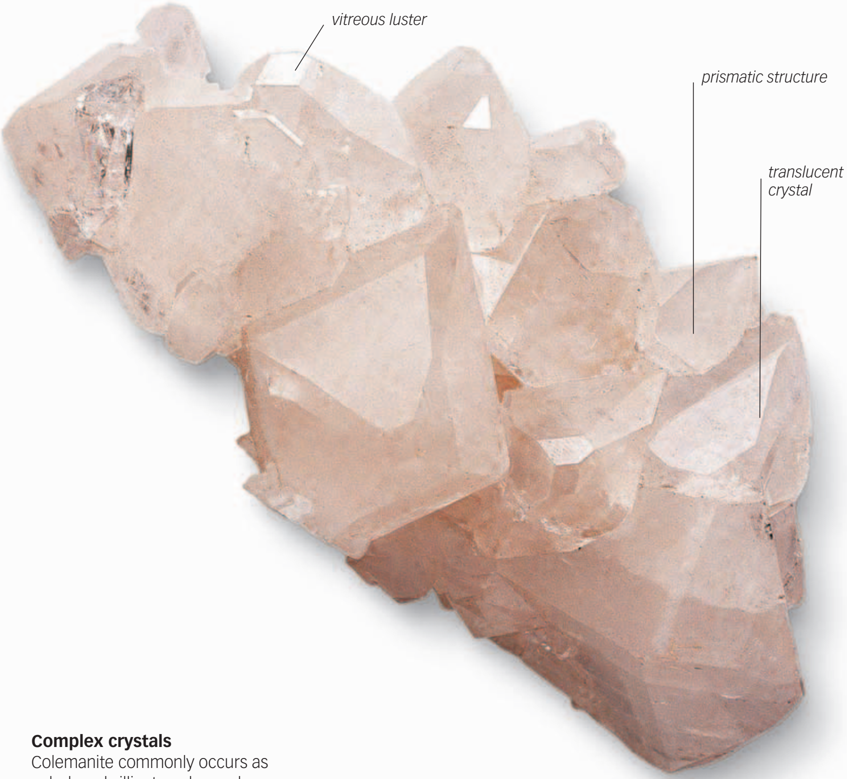

Complex crystals
Colemanite commonly occurs as colorless, brilliant, and complex crystals, as in this specimen.

PROFILE

Monoclinic
4–4½
2.4
Perfect, distinct
Uneven to subconchoidal
White
Vitreous to adamantine

$CaB_3O_4(OH)_3 \cdot H_2O$

COLEMANITE

An important source of boron, colemanite was named in 1884 after William Coleman, the owner of the mine in California where it was discovered. It is colorless, white, yellowish white, or gray. Colemanite occurs as short prismatic or equant crystals in nodules or as granular or coarse, massive aggregates. It is usually massive in commercial deposits, but individual crystals up to 8 in (20 cm) long have also been found.

Colemanite is found in playas and other evaporite deposits, where it replaces other borate minerals, such as borax (p.127) and ulexite (p.128), which were originally deposited in huge inland lakes. Borosilicates derived from colemanite and other minerals are used to make glass that is resistant to chemicals, electricity, and heat.

Heat-resistant glass
Borosilicate glass is used in car headlights, laboratory glassware, ovenware, and industrial equipment.

SULFATES, MOLYBDATES, CHROMATES, AND TUNGSTATES

Sulphates, molybdates, chromates, and tungstates share similar structures and chemical behavior. Sulfates are soft and lightweight, chromates are rare and brightly colored, and tungstates and molybdates are dense, hard, brittle, and vividly colored.

COMPOSITION

Sulfate minerals have a tetrahedral crystal structure, with four oxygen atoms at each corner and a sulfur atom in the center. The sulfate tetrahedron behaves chemically as a single, negatively charged radical or unit. All sulfates contain an SO_4 group.

The basic structural unit of the chromates, molybdates, and tungstates is also a tetrahedron formed from four oxygen atoms, with a central chromium (Cr), molybdenum (Mo), or tungsten (W) atom, respectively. The chromate minerals all contain a CrO_4 group, the molybdates an MoO_4 group, and the tungstates a WO_4 group.

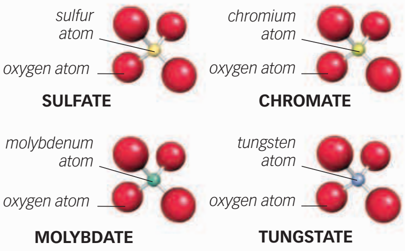

Crystal structure
Tetrahedra are the structural basis of the sulfates, chromates, tungstates, and molybdates. The central metal atom gives each group its name.

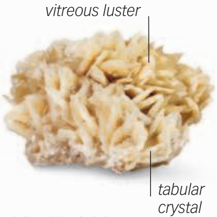

Barite crystals
This large group of tabular barite crystals is from the Wet Grooves mine, Yorkshire, England. Barite has important industrial and medicinal uses.

OCCURRENCE

Sulfates, such as gypsum, occur in evaporite deposits; others, such as barite, mainly occur in hydrothermal veins. Many tungstates are found in hydrothermal veins and pegmatites. Chromates and molybdates are often found as secondary minerals.

USES

The sulfates gypsum and barite are major industrial minerals. Chromates, tungstates, and molybdates are rare but when found concentrated are important ores of the metals they contain.

Plaster cast
About 75 percent of the calcium sulfate gypsum that is mined is used to make plaster of Paris. Most is used for wallboards, but some finds medical uses, such as making plaster casts.

Mineral-rich Madagascar
The island of Madagascar is rich in minerals. It is a prime locality for rich, blue crystals of the sulfate celestine, which is mostly mined for collectors.

PROFILE

Orthorhombic
2½–3
6.4
Good, distinct
Conchoidal, brittle
Colorless
Adamantine to resinous, vitreous

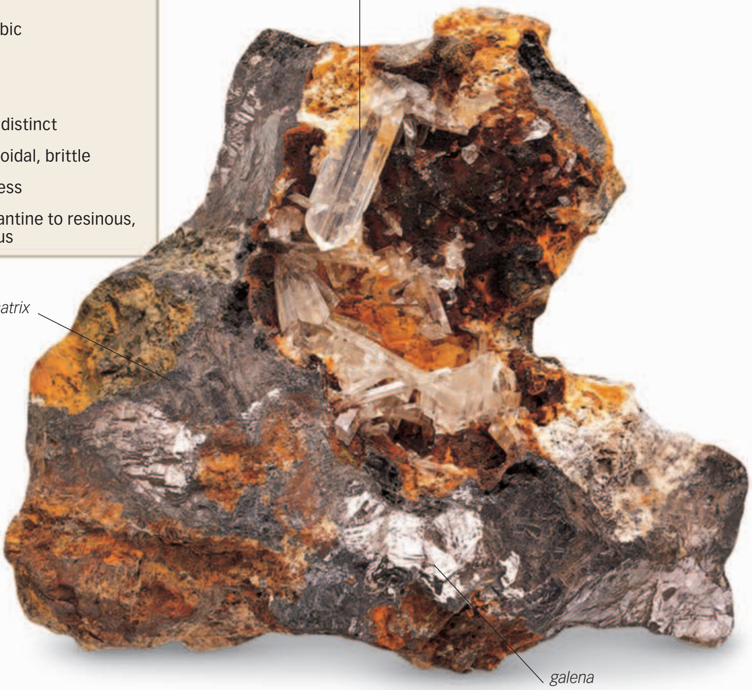

Anglesite crystals These striated prismatic crystals of anglesite are on a rock matrix with galena.

VARIANTS

Pyramidal crystal A pointed crystal of anglesite with galena

adamantine luster

Single crystal A crystal of anglesite that has an adamantine luster

$PbSO_4$

ANGLESITE

Named in 1832 after the large deposit of this mineral found on the island of Anglesey in Wales, anglesite is colorless to white, grayish, yellow, green, or blue and often fluoresces yellow under ultraviolet light. It commonly occurs in massive, granular, or compact forms. It has a number of crystal habits: thin to thick tabular, prismatic, pseudorhombohedral, and pyramidal with striations along the length. Exceptionally large crystals—up to 31 in (80 cm) long—have been found.

Used since ancient times as an ore of lead, anglesite forms in the oxidation zones of lead deposits. It is an alteration product of galena (p.54), formed when galena comes into contact with sulfate solutions. Anglesite is sometimes found in concentric layers with a core of unaltered galena.

Oval-cut anglesite Anglesite is soft and easily cleaved. It is one of the stones used to test the skills of master gem cutters.

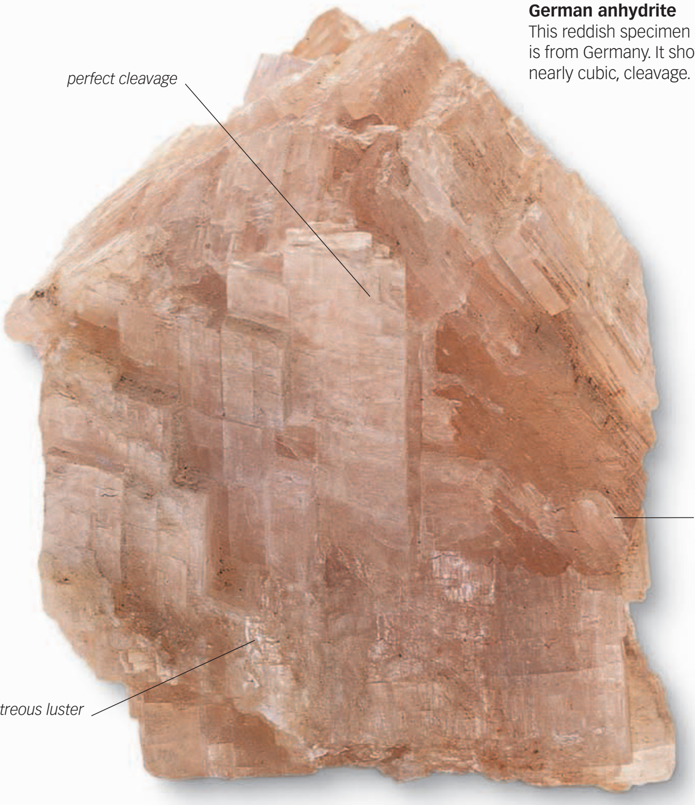

German anhydrite
This reddish specimen of anhydrite is from Germany. It shows perfect, nearly cubic, cleavage.

PROFILE

Orthorhombic

3½

3.0

Perfect, good

Uneven to splintery

White

Vitreous to pearly

$CaSO_4$

ANHYDRITE

An important rock-forming mineral, anhydrite is a calcium sulfate. It takes its name from the Greek word *anhydrous*, which means "without water." Anhydrite is usually colorless to white. Specimens can also be brownish, reddish, or grayish or pale shades of pink, blue, or violet. Individual crystals are uncommon, but when found they are blocky or thick tabular. Crystals up to 4 in (10 cm) long come from Swiss deposits. Anhydrite is usually massive, granular, or coarsely crystalline.

Anhydrite is one of the major minerals in evaporite deposits and commonly occurs in salt deposits associated with halite (p.110) and gypsum (p.136). It alters to gypsum in humid conditions. Anhydrite is often a constituent of cap rocks above salt domes that act as reservoirs for natural oil. It also occurs in volcanic fumaroles and in seafloor hydrothermal "chimneys." Anhydrite is used in fertilizers and as a drying agent in plasters, cement, paints, and varnishes.

PROFILE

Orthorhombic

- 3–3½
- 4.5
- Perfect
- Uneven
- White
- Vitreous, pearly, resinous

Barite crystals
This large group of tabular barite crystals is from the Wet Grooves Mine in Yorkshire, England.

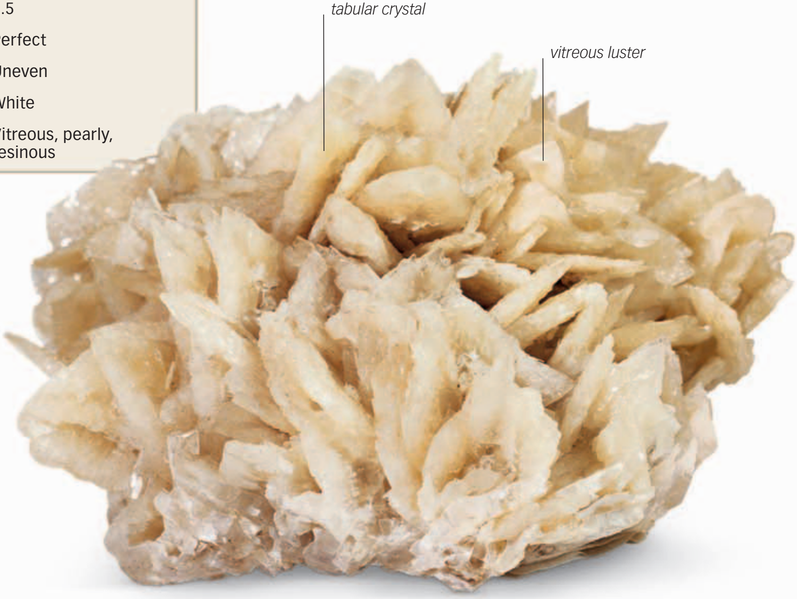

VARIANTS

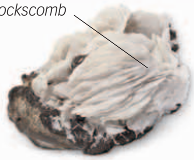

Cockscomb White cockscomb barite resting on sphalerite

Prismatic crystals
A group of yellow prismatic barite crystals

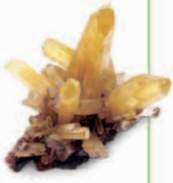

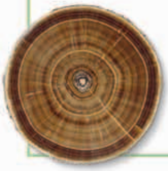

Stalagmite section
Barite in a stalagmitic form

$BaSO_4$

BARITE

The barium sulfate barite takes its name from the Greek word *barys*, which means "heavy"—a reference to its high specific gravity. It has also been called heavy spar. Barite crystals are sometimes tinged yellow, blue, or brown. Golden barite comes from South Dakota. Crystals are well formed, usually either prismatic or tabular. Cockscomb (crested aggregates) and desert roses (rosette aggregates) of crystals are common. Transparent, blue barite crystals may resemble aquamarine but are distinguished by their softness, heaviness, and crystal shape. Barite can also be stalactitic, stalagmitic, fibrous, concretionary, or massive.

Barite is a common accessory mineral in lead and zinc veins. It is also found in sedimentary rocks, clay deposits, marine deposits, and cavities in igneous rocks.

Barite gemstone
Although transparent barite is soft and difficult to cut, it is sometimes faceted for collectors.

PROFILE

Orthorhombic

3–3½

4.0

Perfect

Uneven

White

Vitreous, pearly on cleavage

Celestine crystals
This superbly crystallized specimen of blue celestine crystals is from Madagascar. The largest crystal is more than 1½ in (3.5 cm) long.

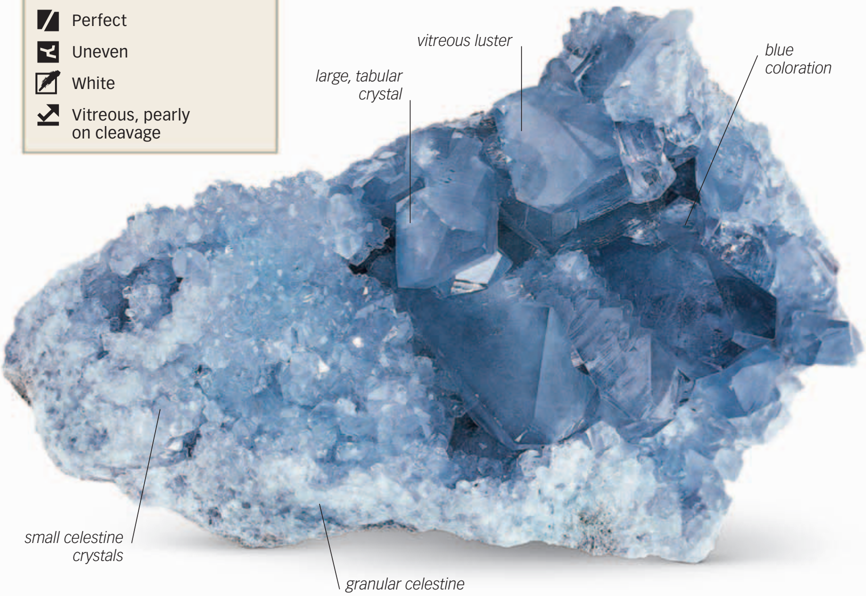

VARIANTS

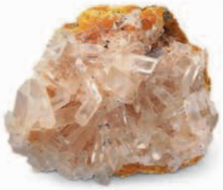

Colorless celestine
Prismatic, colorless crystals on a sulfur matrix

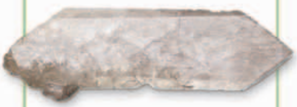

Single crystal A light blue prismatic crystal of celestine

$SrSO_4$

CELESTINE

Often light blue in color, celestine takes its name from the Latin word *coelestis*, which means "heavenly"—an allusion to the color of the sky. Specimens can also be colorless, white, light red, green, medium to dark blue, or brown. Celestine crystals are commonly more than 4 in (10 cm) long. Well-formed, transparent, light- to medium-blue, tabular crystals are common, and some have been known to reach more than 30 in (75 cm) in length. Crystals can also be blocky, bladed, or form elongate pyramids. Celestine may also be massive, fibrous, granular, or nodular in habit.

Celestine forms in cavities in sedimentary rocks (pp.306–33). It commonly occurs in evaporite deposits and can also be precipitated directly from seawater. It can occasionally form in hydrothermal deposits. Celestine is an ore of strontium.

Collector's gem
Celestine is too soft to wear. Faceted celestine demonstrates the skills of master cutters.

PROFILE

Monoclinic

2

2.3

Perfect

Splintery

White

Subvitreous to pearly

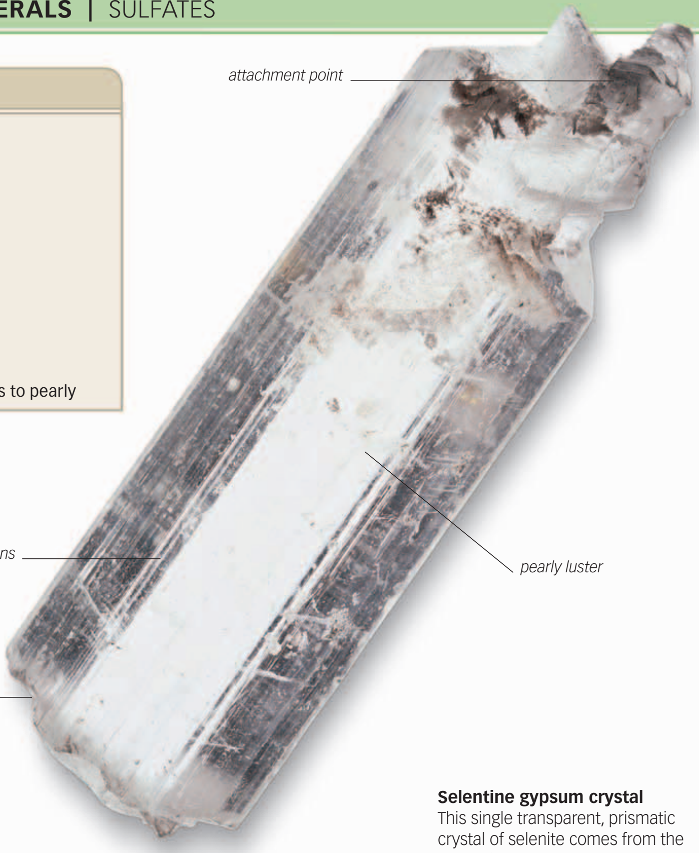

Selentine gypsum crystal This single transparent, prismatic crystal of selenite comes from the Cave of Swords in Mexico.

$CaSO_4 \cdot 2H_2 0$

GYPSUM

A widespread calcium sulfate hydrate, gypsum is found in a number of forms and is of great economic importance. It is colorless or white but can be tinted light brown, gray, yellow, green, or orange due to the presence of impurities. Single, well-developed crystals can be blocky with a slanted parallelogram outline, tabular, or bladed. Twinned crystals are common and frequently form characteristic “fishtails.” Numerous transparent, swordlike selenite gypsum crystals 6½ ft (2 m) or more long can be found at the Cave of Swords, Chihuahua, Mexico, one of the world’s most spectacular mineral deposits.

Gypsum occurs in extensive beds formed by the evaporation of ocean brine. It also occurs as an alteration product of sulfides in ore deposits and as volcanic deposits.

Cat’s eye sheen Satin spar, a fibrous variety of gypsum, can be cut into a cabochon gem with a cat’s eye sheen.

VARIANTS

silky sheen

Gypsum satin spar Fibrous gypsum crystals

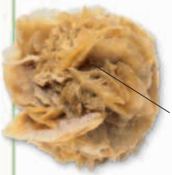

Desert rose Spherical clusters of bladed selenite

Fishtail twin Colorless, translucent selenite gypsum with fishtail twinning

Massive melanterite
This nodule of melanterite shows typical massive form and some minor crystallization.

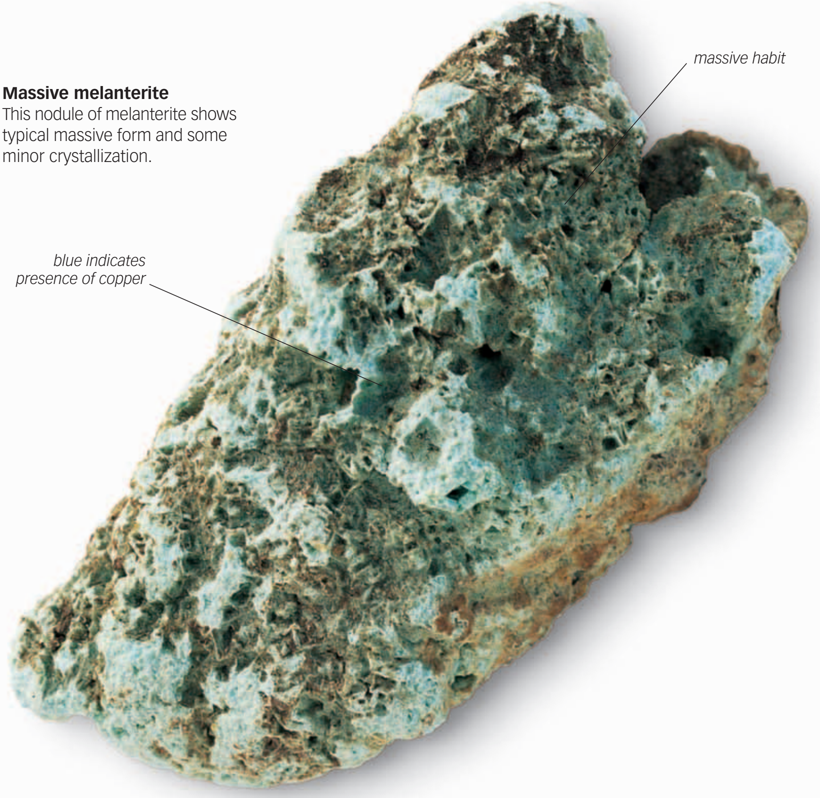

PROFILE

Monoclinic
2
1.9
Perfect
Conchoidal, brittle
White
Vitreous

$FeSO_4 \cdot 7H_2O$

MELANTERITE

A hydrous iron sulfate, melanterite takes its name from the Greek word *melas*, which means "sulfate of iron." Melanterite is the iron analog of the copper sulfate chalcanthite, the two minerals having similar molecular structures. Most specimens of melanterite are colorless to white but can become green to blue as copper increasingly substitutes for iron. Melanterite is generally found in stalactitic or concretionary masses and rarely forms crystals. When crystals occur, they are short prisms or pseudo-octahedrons.

Melanterite is a secondary mineral formed by the oxidation of pyrite (p.62), marcasite (p.63), and other iron sulfides. It is frequently deposited on the timbers of old mine workings. Melanterite also occurs in the altered zones of pyrite-bearing rocks, especially in arid climates and in coal (p.253) deposits, where it is an alteration product of marcasite. Iron sulfate is used in water purification as a coagulant and also as a fertilizer.

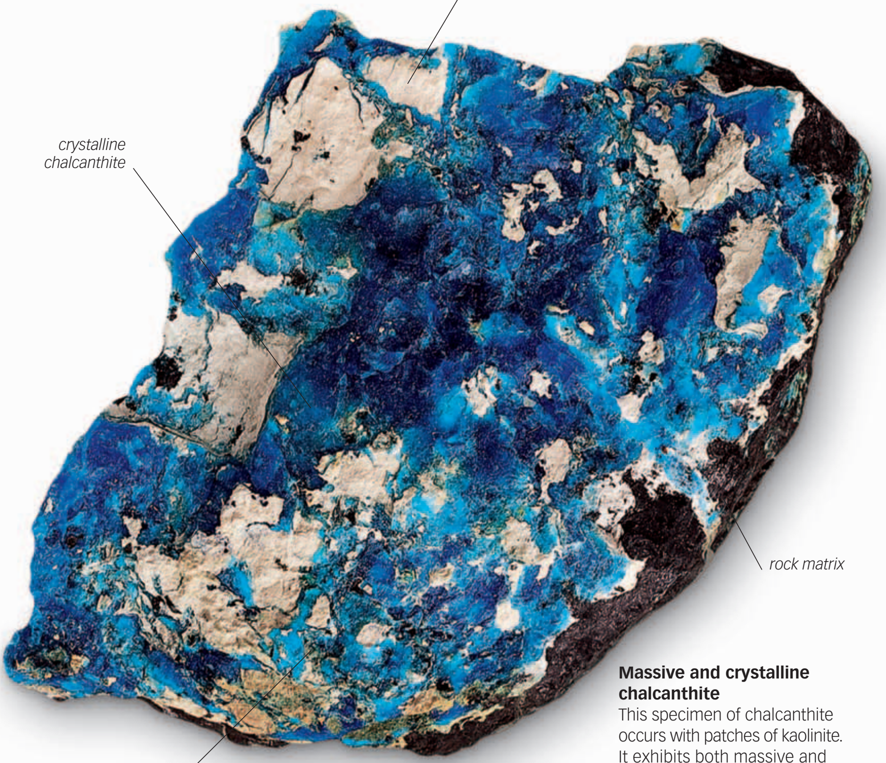

Massive and crystalline chalcanthite This specimen of chalcanthite occurs with patches of kaolinite. It exhibits both massive and crystalline forms of the mineral.

PROFILE

Triclinic

2½

2.3

Not distinct

Conchoidal

Colorless

Vitreous

VARIANT

Stalactite Chalcanthite in the form of a stalactitic aggregate

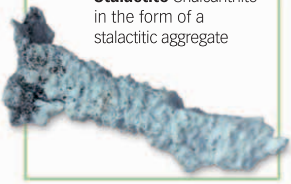

$CuSO_4 \cdot 5H_2O$

CHALCANTHITE

A hydrated copper sulfate, chalcanthite takes its name from the Greek words *khalkos*, which means “copper,” and *anthos*, which means “flower.” It used to be known as blue vitriol. It is commonly peacock blue, although some specimens are greenish. Natural crystals are relatively rare. Chalcanthite usually occurs in veinlets and as massive and stalactitic aggregates.

This widespread, naturally occurring mineral forms through the oxidation of chalcopyrite (p.57) and other copper sulfates that occur in the oxidized zones of copper deposits. Being a water-soluble mineral, it is often found forming crusts and stalactites on the walls and timbers of mine workings, where it crystallizes from mine waters. In arid areas, such as Chile, chalcanthite concentrates in sufficient quantities without being dissolved away to constitute an important ore of copper. Although chalcanthite is a sought-after collectors’ mineral, its crystal structure disintegrates over time because it readily absorbs water.

PROFILE

Monoclinic

3½–4

4.0

Perfect

Uneven to subconchoidal

Pale green

Vitreous

Acicular brochantite
This brochantite specimen from Chile has needlelike crystals on a matrix of iron oxides.

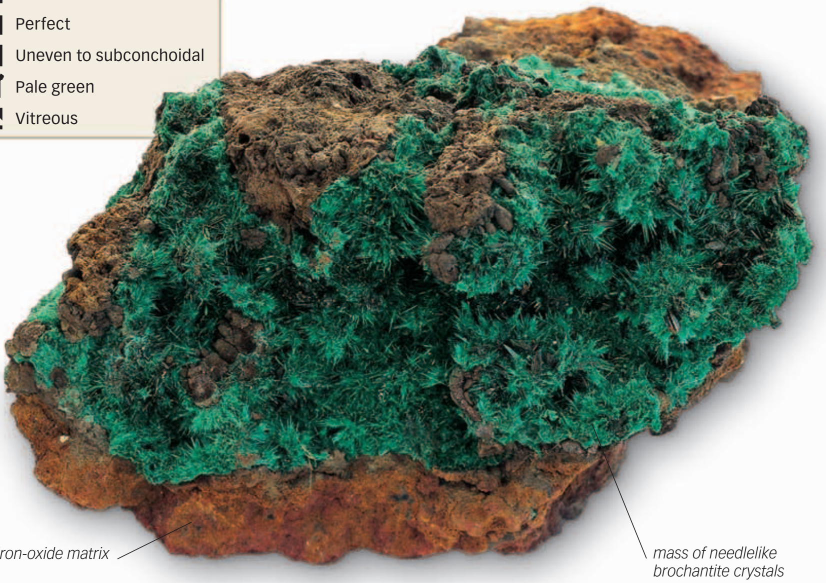

VARIANT

blue azurite

Brochantite on azurite
Green brochantite with blue azurite

$Cu_4SO_4(OH)_6$

BROCHANTITE

A hydrous copper sulfate, brochantite is emerald green, blue-green, or blackish green in color. It was named in 1824 after the French geologist and mineralogist A.J.M. Brochant de Villiers, who was the first pupil admitted to the École des Mines, Paris, and who later became its Professor of Geology and Mines. Brochantite usually forms prismatic or needlelike crystals, which rarely exceed a fraction of an inch in length. Twinning is common in crystals. Brochantite is also found in tufts and druse crusts and as fine-grained masses.

Brochantite forms in the oxidation zones of copper deposits, especially those that occur in the arid regions of the world. In these regions, brochantite is usually associated with azurite (p.120), malachite (p.125), and other copper minerals. In Arizona, and Chile, the mineral is abundant enough to be an ore of copper. Splendid specimens of brochantite come from Namibia, and Bisbee, Arizona, where prismatic crystals may exceed ⅜in (1 cm) in length.

PROFILE

Orthorhombic

2–2½

1.7

Perfect

Conchoidal

White

Vitreous to silky

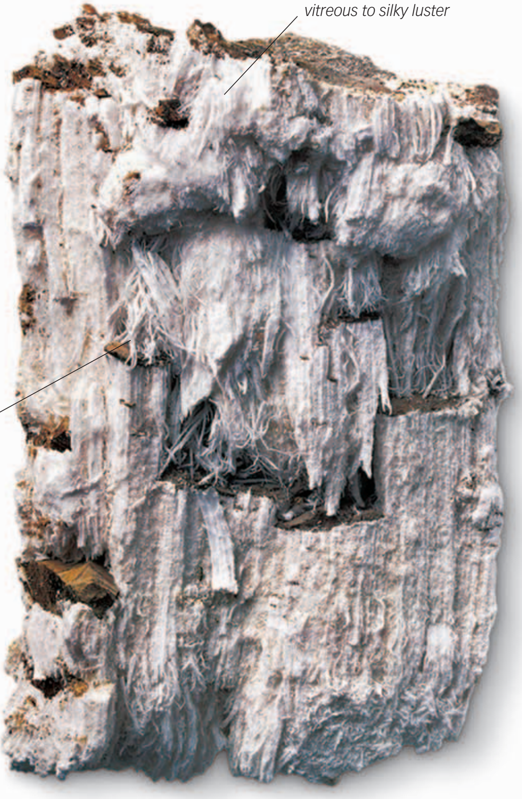

Fibrous epsomite
This epsomite specimen occurs in a fibrous habit and shows a vitreous to silky luster.

VARIANT

Powdery mass Epsomite coating on a rocky matrix

$MgSO_4{\cdot}7H_2O$

EPSOMITE

Epsom salts is the common name for this hydrated magnesium sulfate mineral. It was first found around springs near the town of Epsom in Surrey, England, and was named after that locality in 1805. It is colorless, white, pale pink, or green. Epsomite crystals are rare; when found, they are either prismatic or fibrous. Epsomite usually occurs as crusts, powdery or woolly coatings, or sometimes as botryoidal or reniform masses.

Magnesium sulfate occurs in solution in seawater, saline lake water, and spring water. When the water evaporates, epsomite precipitates, forming deposits. It is also found with coal (p.253), in weathered magnesium-rich rocks, sulfide ore deposits, and dolomite (p.320) and limestone caves.

Refined epsom salt
This widely used medication is derived from epsomite. One common use is as a natural laxative.

PROFILE

Monoclinic

2½–3

2.8

Perfect, indistinct

Conchoidal

White

Vitreous to waxy

Glauberite crystals This group of dipyramidal glauberite crystals is from Ciempozuelos, Madrid, Spain.

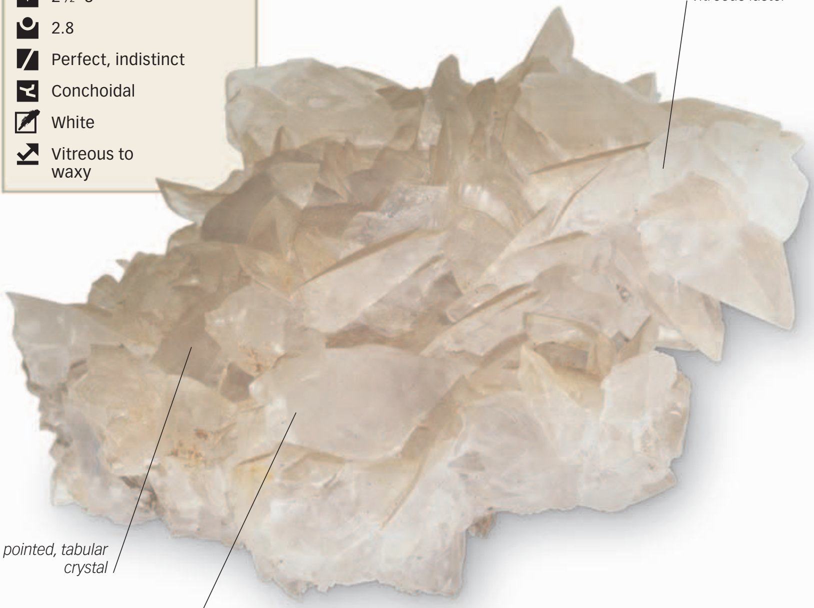

VARIANTS

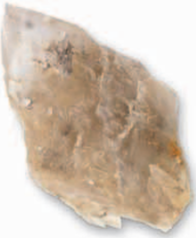

Single crystal A single pyramidal crystal of glauberite

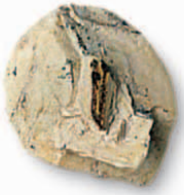

Pseudomorph A specimen with glauberite replaced by calcite

$Na_2Ca(SO_4)_2$

GLAUBERITE

This mineral was named in 1808 for its similarity to another chemical, Glauber's salt, which in turn was named after the German alchemist Johann Glauber. Glauberite is a sodium calcium sulfate. It can be colorless, pale yellow, reddish, or gray, and its surface may alter to white, powdery sodium sulfate. Crystals can be prismatic, tabular, and dipyramidal, all with combinations of forms and all of which may have rounded edges. Glauberite crystal pseudomorphs form when other minerals, such as calcite (p.114) and gypsum (p.136), replace it. Glauberite has a slightly saline taste, turns white in water, and fuses to a white enamel.

This mineral forms under a variety of conditions. It is primarily an evaporite, forming in both marine and salt-lake environments. It is also found in cavities in basaltic igneous rocks and in volcanic fumaroles. Molds and casts of quartz (p.168) and prehnite (p.205) formed from glauberite are frequently found in basalt cavities in Patterson, New Jersey, USA.

PROFILE

Monoclinic

2½–3

6.0

Distinct in one direction

Conchoidal to uneven, brittle

Orange-yellow

Vitreous

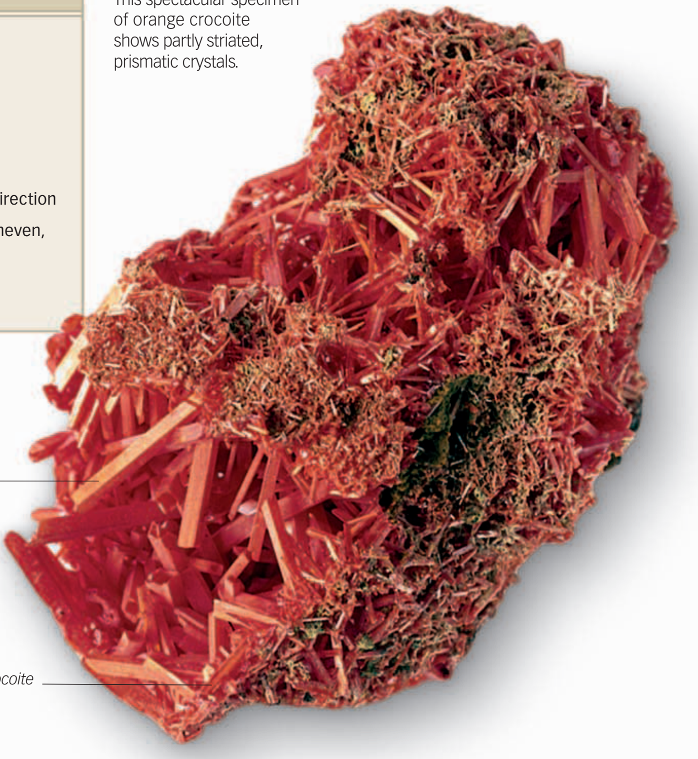

Prismatic crystals This spectacular specimen of orange crocoite shows partly striated, prismatic crystals.

VARIANTS

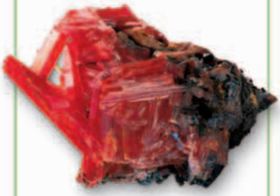

Red crocoite Red, prismatic crystals of crocoite in a rock matrix

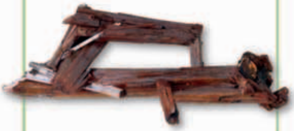

Atypical crocoite An almost reticulated growth of crocoite

$PbCrO_4$

CROCOITE

One of the most eye-catching of minerals, crocoite is bright orange to red in color. It takes its name from the Greek word *krókos*, which means “saffron.” Crocoite crystals are prismatic, commonly square-sectioned, slender and elongated, and sometimes cavernous or hollow. They may be striated along their length and may rarely show distinct terminations. Crystals usually occur in radiating or randomly intergrown clusters. Crocoite can also occur in granular or massive forms. On exposure to light, much of the translucence and brilliance of the mineral is lost.

Crocoite is a rare mineral, as specific conditions—an oxidation zone of lead ore and the presence of low-silica igneous rocks that serve as the source of chromium—are required for its formation. It is the official mineral emblem of Tasmania, where exceptional crystals, 3–4 in (7.5–10 cm) long and having a brilliant luster and color, are found. Crocoite is identical in composition to the pigment chrome yellow.

PROFILE

Tetragonal
2½–3
6.5–7.0
Distinct
Subconchoidal to uneven
White
Subadamantine to greasy

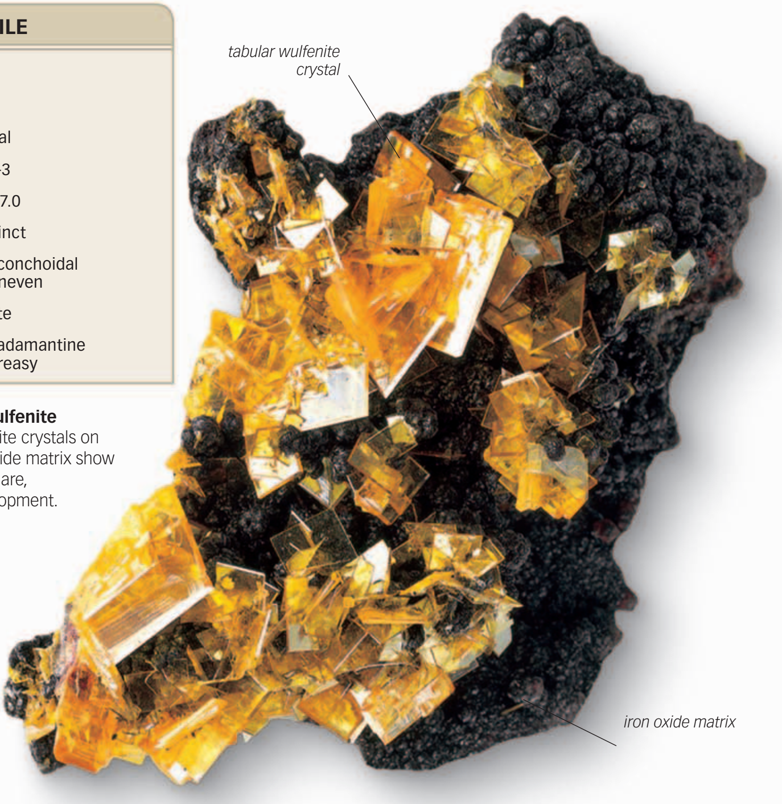

Yellow wulfenite The wulfenite crystals on this iron oxide matrix show classic, square, platy development.

VARIANTS

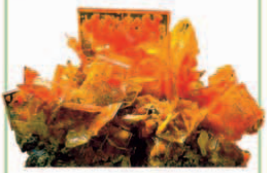

Square crystals Typical thin, tabular crystals of wulfenite

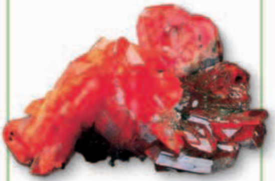

Red cloud wulfenite Crystals found in Red Cloud Mine in Arizona

$PbMoO_4$

WULFENITE

The second most common molybdenum mineral after molybdenite, wulfenite was named after F.X. Wülfen, an Austro-Hungarian mineralogist, in 1841. The color of specimens varies and can be yellow, orange, red, gray, or brown. Wulfenite usually forms as thin, square plates or square, beveled, tabular crystals. Crystals sometimes show different terminations on each end, probably due to twinning. Bright, colorful, and sharply formed crystals are popular with collectors. Wulfenite also occurs in massive, earthy, and granular forms.

Wulfenite is a minor source of molybdenum. Tungsten substitutes for the molybdenum, although in most specimens it is present only in trace amounts. Wulfenite is a secondary mineral formed in the oxidized zones of lead and molybdenum deposits, and it occurs with other minerals, including cerussite (p.119), pyromorphite (p.151), and vanadinite (p.155). It is relatively widespread and is often found in superb crystals, occasionally up to 4 in (10 cm) on an edge.

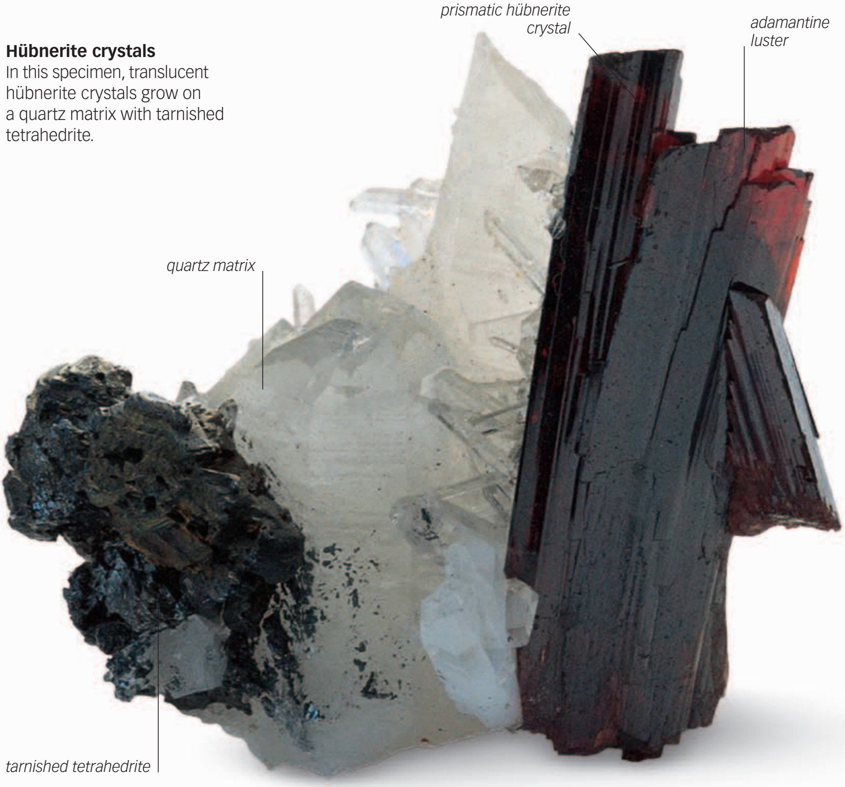

Hübnerite crystals
In this specimen, translucent hübnerite crystals grow on a quartz matrix with tarnished tetrahedrite.

PROFILE

Monoclinic
4–4½
7.3
Perfect
Uneven
Yellow to brown
Submetallic/adamantine to resinous

$MnWO_4$

HÜBNERITE

Named after the German mineralogist Adolf Hübner, who first described it in 1865, hübnerite is an important ore of tungsten. It is found as prismatic, long prismatic, tabular, or flattened crystals with striations and is commonly twinned. It can also form groups of parallel or subparallel crystals or radiating groups. Hübnerite is generally reddish brown. In transparent crystals, it can change color when viewed from different directions and show strong internal reflection.

Hübnerite is the manganese end member of a manganese–iron solid-solution series. It occurs in granitic pegmatites (p.260) and in thermal veins at high temperatures (1,065°F/575°C or above). The mineral is also recovered from alluvial gravels, in which it can concentrate.

Bulb filament
Hübnerite is an ore of tungsten, which is mainly used in light bulb filaments.
]

Ferberite crystal
This ferberite crystal is from Cínovec, Czech Republic. It shows the prismatic habit of the mineral.

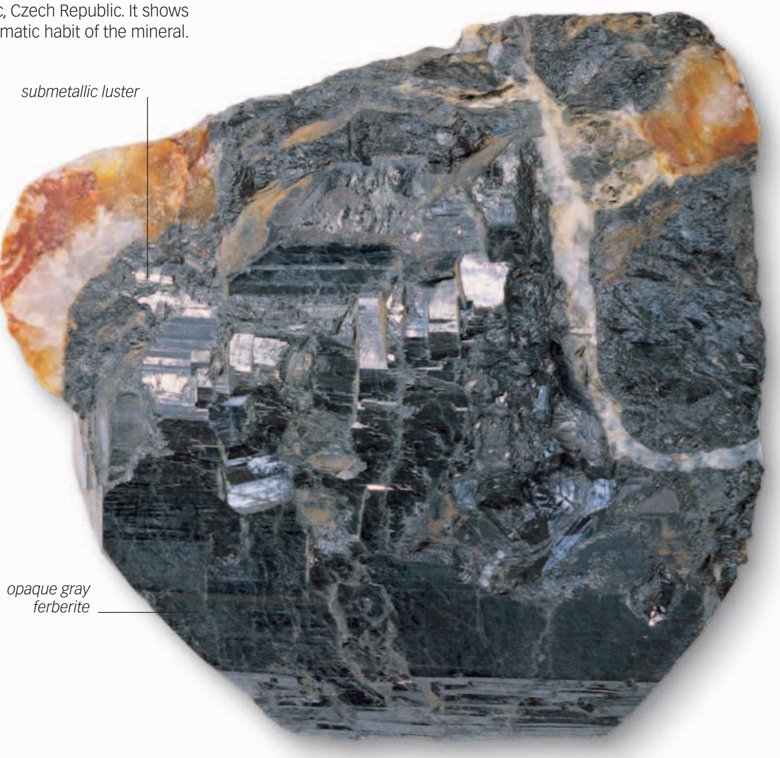

PROFILE

Monoclinic
4–4½
7.5
Perfect in one direction
Uneven, brittle
Black to brown
Submetallic

$FeWO_4$

FERBERITE

The principal ore of tungsten, ferberite is an iron tungstate. It was named in 1863 after Moritz Rudolph Ferber, a German industrialist and mineralogist. Ferberite forms black crystals, which are commonly elongated or flattened with a wedge-shaped appearance. Twinning and striations are common in crystals. Ferberite is also found as granular masses.

Ferberite is the iron end-member of a solid-solution series it forms with hübnerite (p.144), the manganese end-member. Together, they constitute the mineral formerly called wolframite. Ferberite occurs in hydrothermal veins at high temperatures (1,065°F/ 575°C or above) and in granitic pegmatites (p.260) with other minerals.

Tungsten steel
Rocket nozzles, such as those used in Saturn V, are made of heat-resistant, hard, and strong tungsten steel.

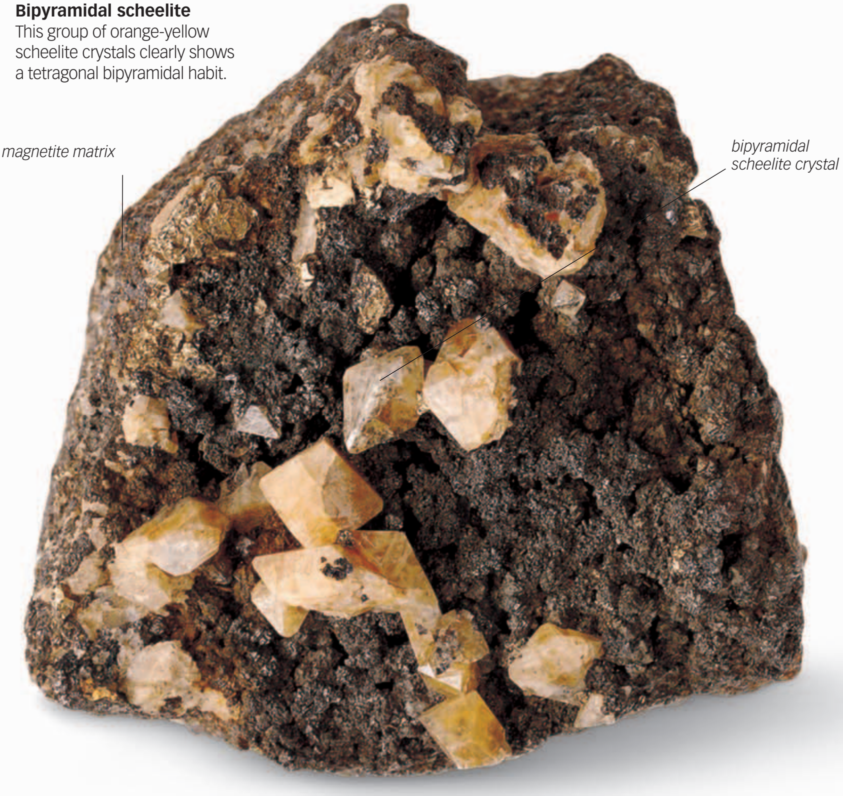

Bipyramidal scheelite This group of orange-yellow scheelite crystals clearly shows a tetragonal bipyramidal habit.

PROFILE

Tetragonal

4½–5

6.1

Distinct

Uneven to subconchoidal

White

Vitreous to greasy

$CaWO_4$

SCHEELITE

Named in 1821 after the Swedish chemist C.W. Scheele, scheelite is calcium tungstate. Its crystals are generally bipyramidal and twinned but also form in granular or massive aggregates. Irregular masses of colorless, gray, orange, or pale brown scheelite can be difficult to spot, but they fluoresce vivid bluish white under a short-wave ultraviolet light. Scheelite is sometimes associated with native gold (p.42), and its fluorescence is used by geologists in their search for gold deposits.

Scheelite commonly occurs in contact with metamorphic deposits, in hydrothermal veins formed at high temperatures (1,065°F/575°C or above), and less commonly in granitic pegmatites (p.260). Opaque crystals weighing up to 15½ lb (7 kg) come from Arizona. Scheelite is a major source of tungsten.

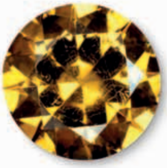

Brilliant cut scheelite Transparent scheelite is relatively rare. Stones faceted from it are only for gem collectors.

PHOSPHATES, VANADATES, AND ARSENATES

The phosphate, arsenate, and vanadate minerals are grouped together because their crystal structures are similar. The phosphates are the most numerous of the three groups, with more than 200 known minerals.

COMPOSITION

Phosphates contain phosphorus and oxygen in a 1:4 ratio, written as PO_4. The combined atoms act as a single unit that in turn combines with other elements to form phosphate minerals. Arsenates have a basic structural unit of arsenic and oxygen, written as AsO_4, which combines with other elements to form arsenate minerals. Most arsenates are rare and many are brilliantly colored. Vanadates mostly contain the same type of structural tetrahedra as the phosphates and arsenates, written as VO_4. The structures of vanadates are complex, and these minerals are relatively rare.

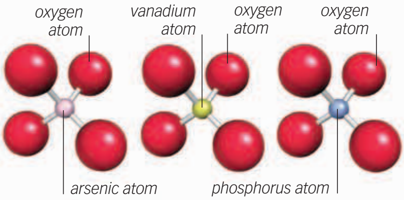

Crystal structure
The arsenate (AsO_4), vanadate (VO_4), and phosphate (PO_4) ions consist of a metal atom bonded to four oxygen atoms. Each ion acts as a single unit.

OCCURRENCE

Primary phosphates usually crystallize from aqueous fluids derived from igneous crystallization; secondary phosphates, when primary phosphates are altered in the presence of water; and rock phosphates, from phosphorus-bearing organic material.

USES

Phosphates are of major economic importance as fertilizers. Vanadates are minor ores of vanadium and have no other economic importance. The only exception is carnotite, an important source of uranium.

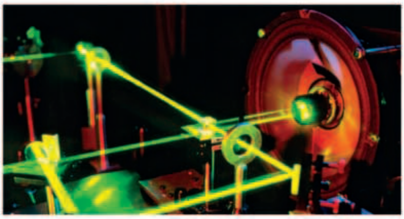

High-speed laser
Garnet-containing yttrium, which is derived from the yttrium phosphate xenotime, is used to make lasers. The direction of the laser beams is changed using mirrors.

Colorado Plateau
Spread across the Colorado Plateau in western Colorado, USA, are extensive deposits of carnotite (a phosphate of uranium, vanadium, and potassium).

PROFILE

Monoclinic

5

3.1–3.2

Indistinct, variable

Conchoidal to uneven

White

Vitreous, waxy

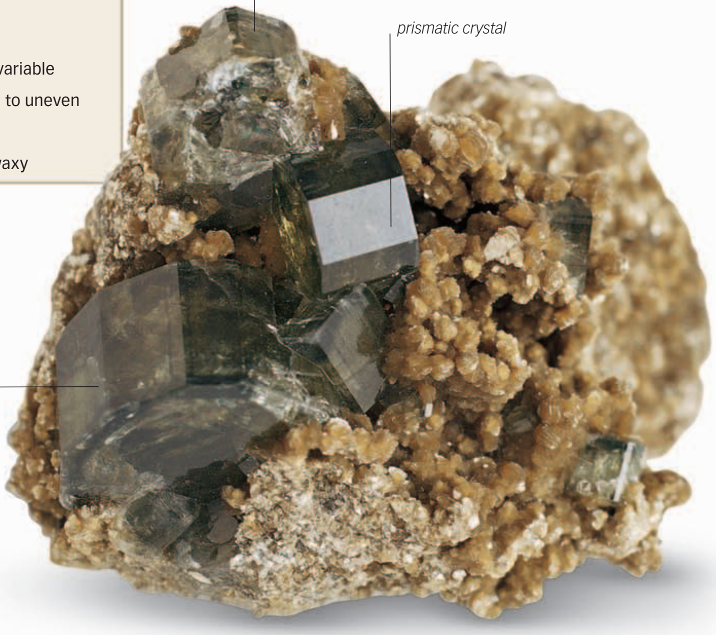

Apatite crystals
These spectacular apatite crystals from Panasqueira Mine, Beira Baixa, Portugal, occur with muscovite and a small amount of arsenopyrite.

VARIANTS

Chlorapatite crystal
Double-terminated chlorapatite

Brilliant apatite
A single yellow crystal of hydroxylapatite

Hydroxylapatite A specimen with a waxy luster

$Ca_5(PO_4)_3(F,OH,Cl)$

APATITE

A series of calcium phosphate minerals that differ in composition are known as apatites. The name apatite is derived from the Greek *apate*, which means "deceit"—a reference to its similarity to crystals of aquamarine, amethyst, and olivine (p.232). Apatites can occur as green, blue, violet-blue, purple, colorless, white, yellow, pink, or rose-red specimens. All the apatites are structurally similar and are commonly found as transparent, well-formed, glassy crystals and in masses or nodules. Crystals are short to long prismatic, thick tabular, or prismatic with complex forms.

Apatites occur in marbles (p.301), skarns (p.302), and other metamorphic rocks. Rich deposits of apatite also occur in sedimentary rocks. As an accessory mineral, it occurs in a wide range of igneous rocks and in hydrothermal veins.

Step-cut gemstone
Owing to the brittleness of apatite, an edge of one facet of this blue gemstone has become chipped.

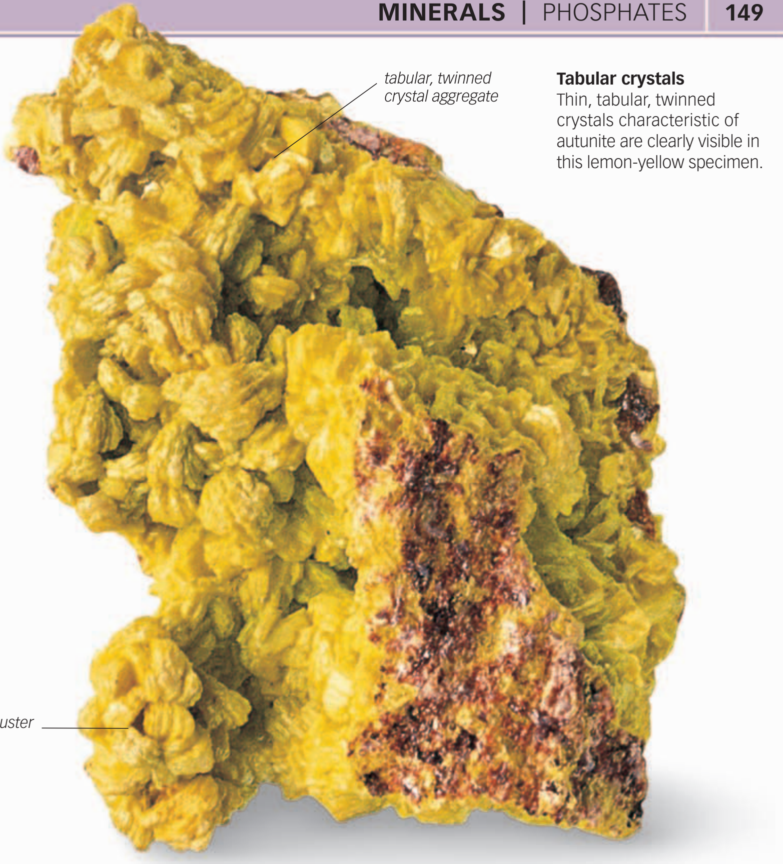

Tabular crystals Thin, tabular, twinned crystals characteristic of autunite are clearly visible in this lemon-yellow specimen.

PROFILE

Tetragonal
2–2½
3.1–3.2
Perfect basal
Uneven
Pale yellow
Vitreous to pearly

$Ca(UO_2)_2(PO_4)_2\cdot10\text{–}12H_20$

AUTUNITE

A popular collector's mineral, autunite is a calcium uranium phosphate. Greenish or lemon yellow in color, autunite specimens fluoresce green under ultraviolet light. Crystals of autunite have a rectangular or octagonal outline. Coarse groups are found, but scaly coatings are more common. Autunite is also found as crusts with crystals standing on edge, giving a serrated appearance.

Autunite is named after Autun, France, where this mineral was discovered. It is formed in the oxidation zones of uranium ore bodies as an alteration product of uraninite (p.83) and other uranium-bearing minerals. It also occurs in hydrothermal veins and in pegmatites (p.260). Since autunite contains uranium and is radioactive, it must be stored carefully and handled as little as possible. When mildly heated, tetragonal autunite dehydrates into orthorhombic meta-autunite. Most museum and collector specimens of autunite have been converted to meta-autunite. A moist atmosphere helps prevent dehydration.

PROFILE

Monoclinic

5

4.6–5.4

Perfect, good, poor

Conchoidal to uneven, brittle

White

Resinous, waxy, or vitreous

Monazite crystal This striated crystal of monazite is from Arendal, Aust-Agder, Norway, which is an important monazite locality.

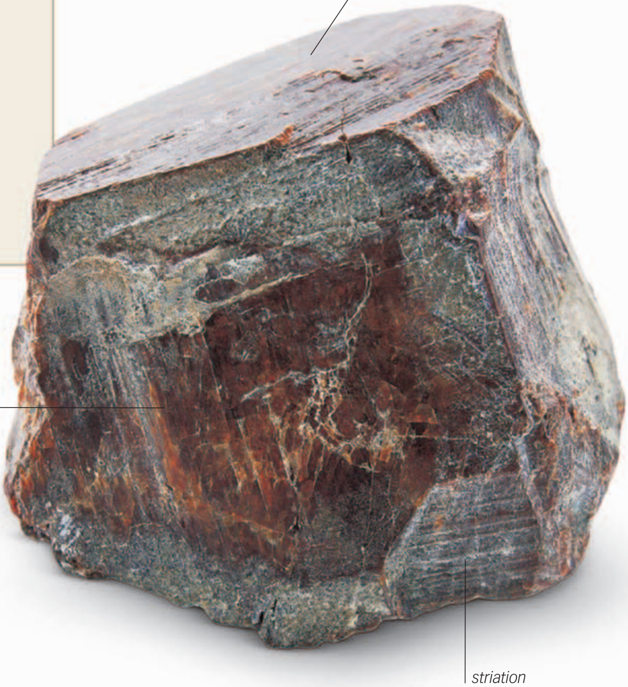

VARIANT

Monazite fragment A brown crystal fragment showing growth lamellae

$(Ce,La,Th,Nd)PO_4$

MONAZITE

The monazite group consists of three different phosphate minerals, all sharing the same crystal structure. The most widespread is monazite-(Ce), cerium phosphate, which is yellowish or reddish brown to brown, greenish, or nearly white. Monazite-(Ce) forms prismatic, flattened, or elongated crystals, which are occasionally large, coarse, and commonly twinned. Two other species of monazite are monazite-(La), which is lanthanum phosphate, and monazite-(Nd), which is neodymium phosphate.

Monazite is a common accessory mineral in granites (pp.258–59) and gneisses (p.288) and in pegmatites (p.260) and fissure veins. Detrital monazite can accumulate as monazite sands. Lanthanum is used in oil refining. Neodymium is used for coloring glass.

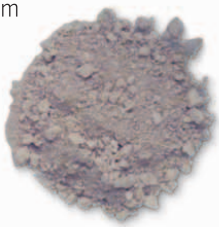

Cerium oxide Monazite-(Ce) is a source of cerium. Cerium oxide is used for polishing glass, stone, and gemstones.

PROFILE

Hexagonal

- 3½–4
- 7.0
- Poor
- Uneven to subconchoidal, brittle
- White
- Resinous

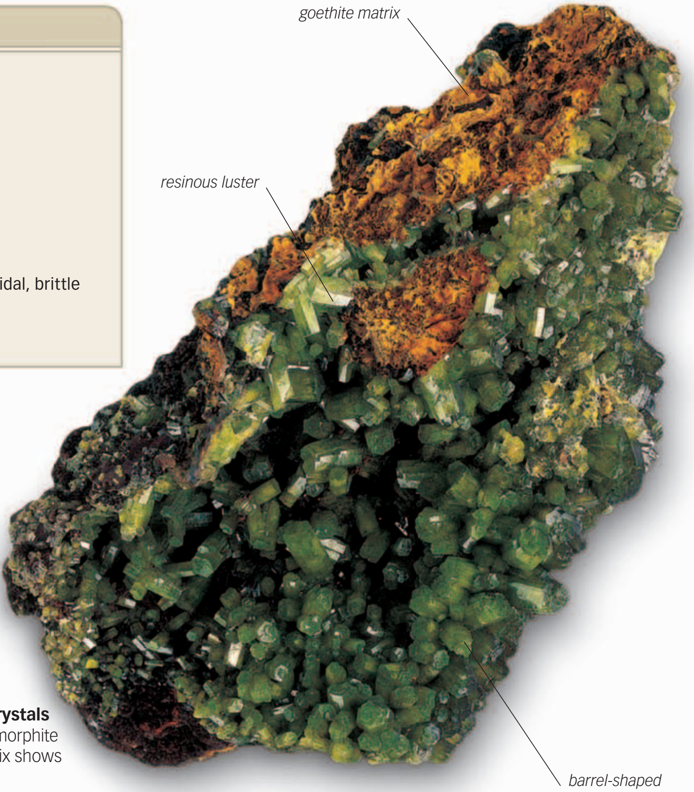

Barrel-shaped crystals This mass of pyromorphite on a limonite matrix shows its typical barrel-shaped crystals.

VARIANTS

Lime-green crystals Crystals showing pyromorphite's intense coloration

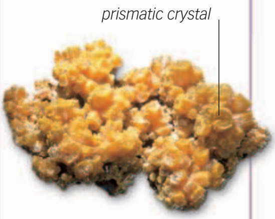

Yellow-green crystals A specimen of pyromorphite with yellow-green prismatic crystals

$Pb_5(PO_4)_3Cl$

PYROMORPHITE

A minor ore of lead but a popular collector's mineral, pyromorphite forms a continuous chemical series with mimetite (p.164) in which phosphorus and arsenic replace each other. Pyromorphite gets its name from the Greek words *pyr*, which means "fire," and *morphe*, which means "form"—an allusion to its property of becoming crystalline on cooling after it has been melted to a globule. It is dark green to yellow-green, shades of brown, a waxy yellow, or yellow-orange. Crystals may be either simple hexagonal prisms or rounded and barrel-shaped, spindle-shaped, or cavernous. Some crystals exhibit different colors when viewed from different directions and some produce electricity on application of mechanical stress. The mineral can also be globular, reniform, or granular in habit.

Pyromorphite occurs as a secondary mineral in the oxidized zones of lead deposits with galena (p.54), goethite (p.102), cerussite (p.119), smithsonite (p.124), and vanadinite (p.155). Pseudomorphs of pyromorphite after galena are common.

Torbernite crystals In this specimen, tabular crystals of torbernite rest on an iron-rich matrix.

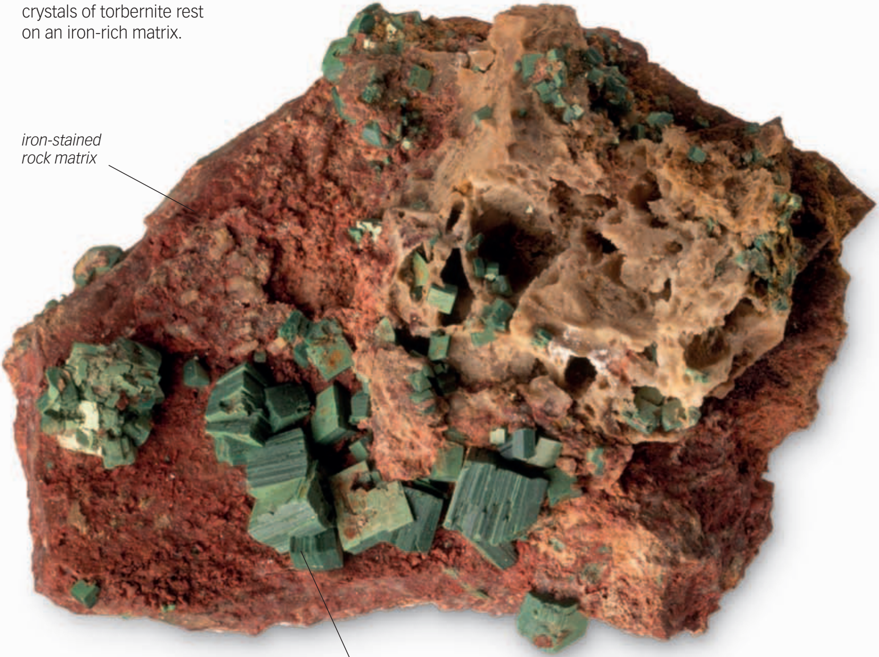

PROFILE

Tetragonal

- 2–2½
- 3.2
- Perfect, basal
- Uneven
- Pale green
- Vitreous to subadamantine

VARIANT

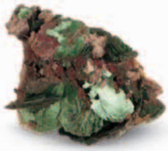

Metatorbernite Green sheaves of metatorbernite crystals in rock matrix

$Cu(UO_2)_2(PO_4)_2 \cdot 8\text{–}12H_2O$

TORBERNITE

Named in 1793 after the Swedish mineralogist Torbern Olaf Bergmann, torbernite is a uranium-bearing mineral and a minor ore of uranium. Torbernite forms thin to thick tabular crystals that are commonly square in outline, foliated micalike masses, sheaflike crystal groups, or scaly coatings. Specimens are bright mid-green, emerald green, leek green, or grass green in color. Torbernite is chemically unstable, and with increased hydration it transforms to metatorbernite. In fact, all specimens are probably metatorbernite. It is also radioactive and needs to be handled with appropriate care.

Torbernite is found as a secondary mineral formed in the oxidation zones of deposits containing uranium and copper and is associated with other phosphate minerals. It forms as an alteration product of uraninite (p.83) or other uranium-bearing minerals. Torbernite is also associated with uraninite, autunite (p.149), and carnotite (p.159). Fine specimens occur in Cornwall, England, in the Flinders Range of Australia, and elsewhere.

Massive aggregate
This specimen of massive triplite is from Megiliggar Rocks in Cornwall, England.

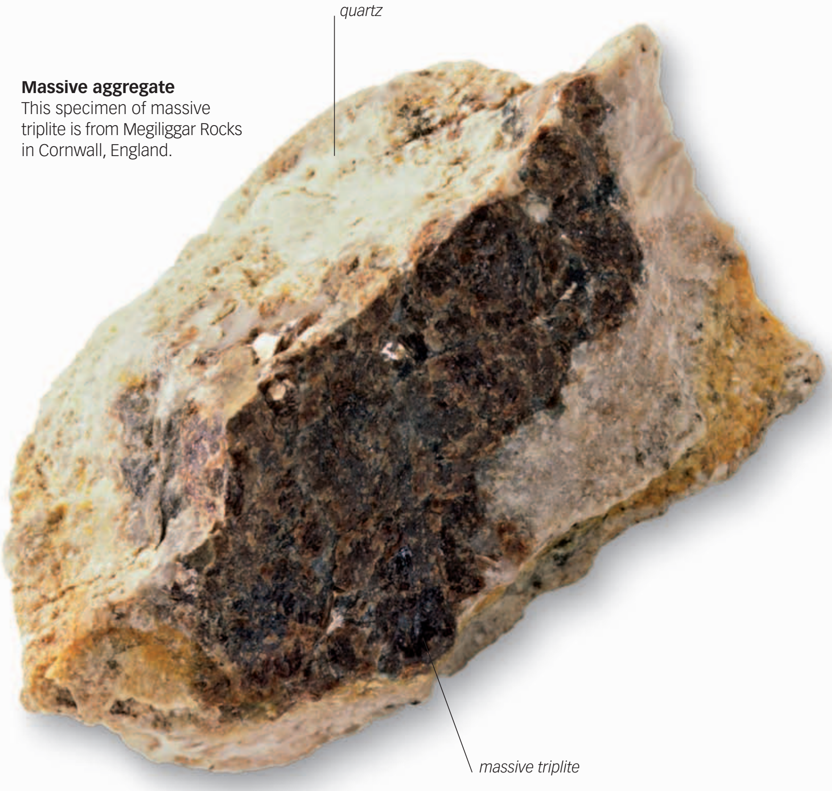

PROFILE

Monoclinic
5–5½
3.5–3.9
Good in three directions
Uneven to subconchoidal
White to brown
Vitreous to resinous

$(Mn,Fe,Mg)_2PO_4(F,OH)$

TRIPLITE

The first occurrence of triplite was described in 1813 in Chanteloube, Limousin, France. Although it is a fluoridated manganese phosphate, in most triplite samples iron partially replaces manganese. Triplite takes its name from the Greek word *triplos,* which means "triple"—a reference to its three cleavages oriented at right angles to each other. Its crystals are typically rough and poorly developed but may have many indistinct forms. Triplite is more commonly nodular or massive. Specimens may be chestnut brown, reddish brown, flesh red, or salmon pink in color. If altered, they may be brownish black to black. Translucent crystals may also exhibit different colors when viewed from different directions (a phenomenon known as pleochroism), going from yellow-brown to reddish brown.

Triplite is a primary mineral in granite pegmatites (p.260) with complex zones and in some hydrothermal tin veins. It may be accompanied by sphalerite (p.53), pyrite (p.62), apatite (p.148), and tourmaline (p.224).

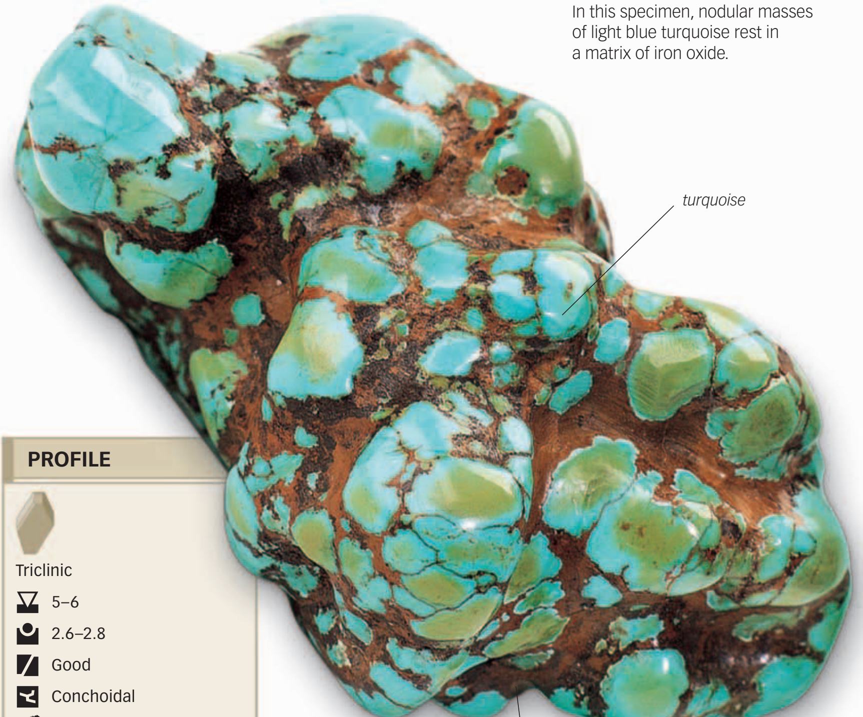

Blue turquoise
In this specimen, nodular masses of light blue turquoise rest in a matrix of iron oxide.

PROFILE

Triclinic
5–6
2.6–2.8
Good
Conchoidal
White to green
Waxy to dull

VARIANTS

Turquoise vein Massive turquoise in a small vein

Green turquoise A hard, green nugget from the USA

Turquoise in rock Blue-green, massive turquoise in rock

$CuAl_6(PO_4)_4(OH)_8{\cdot}4H_2O$

TURQUOISE

One of the first gemstones to be mined, turquoise is a hydrous copper aluminum phosphate. Beads made of turquoise that date back to *c.*5000 BCE have been recovered in Mesopotamia (present-day Iraq). This mineral usually occurs in massive or microcrystalline forms, as encrustations or nodules, or in veins. Crystals are rare; when found, they occur as short, often transparent prisms. Turquoise varies in color from sky-blue to green, depending on the amount of iron and copper it contains.

"Turquoise" is derived from the French word for "Turkey," because it was first transported to Europe through Turkey. Turquoise occurs in arid environments as a secondary mineral probably derived from the decomposition of apatite (p.148) and some copper sulfides.

Carved elephant
Turquoise is a favorite of Chinese stone carvers, who produced this turquoise elephant.

Red vanadinite
Vanadinite crystals are often brilliantly colored in shades of red and yellow. This specimen has smooth-faced, prismatic crystals.

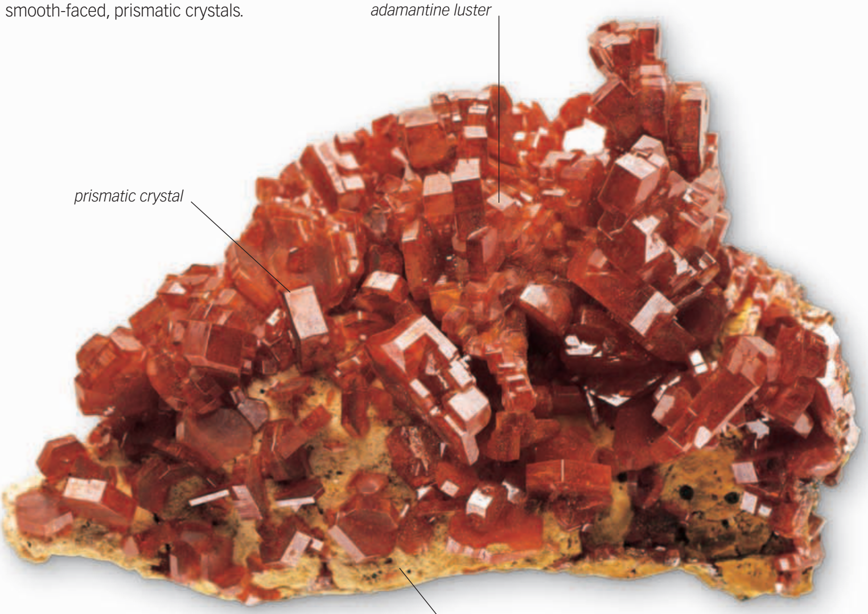

PROFILE

Hexagonal
3
6.9
None
Uneven, brittle
Whitish yellow
Adamantine

$Pb_5(VO_4)_3Cl$

VANADINITE

A relatively rare mineral, vanadinite is a lead chloro-vanadate. The bright red or orange-red colors of vanadinite make it popular among mineral collectors, although it is sometimes brown, red-brown, gray, yellow, or colorless. Crystals are usually in the form of short, hexagonal prisms but can also be found as hexagonal pyramids or as hollow prisms. They can also be needlelike. Small amounts of calcium, zinc, and copper may substitute for lead, and arsenic can completely substitute for vanadium in the crystal structure to form the mineral mimetite (p.164). The mineral is also found as rounded masses or crusts.

Vanadinite forms as a secondary mineral in oxidized ore deposits containing lead, often associated with galena (p.54), goethite (p.102), barite (p.134), and wulfenite (p.143). Vanadium from vanadinite is used to make strong vanadium steels.

Steel spanner
Vanadium imparts strength and hardness to steel that is used to make high-stress tools, such as this spanner.

Concretionary variscite
Variscite is often found in nodules and concretions like the sliced specimen shown here. It can be sometimes mistaken for turquoise.

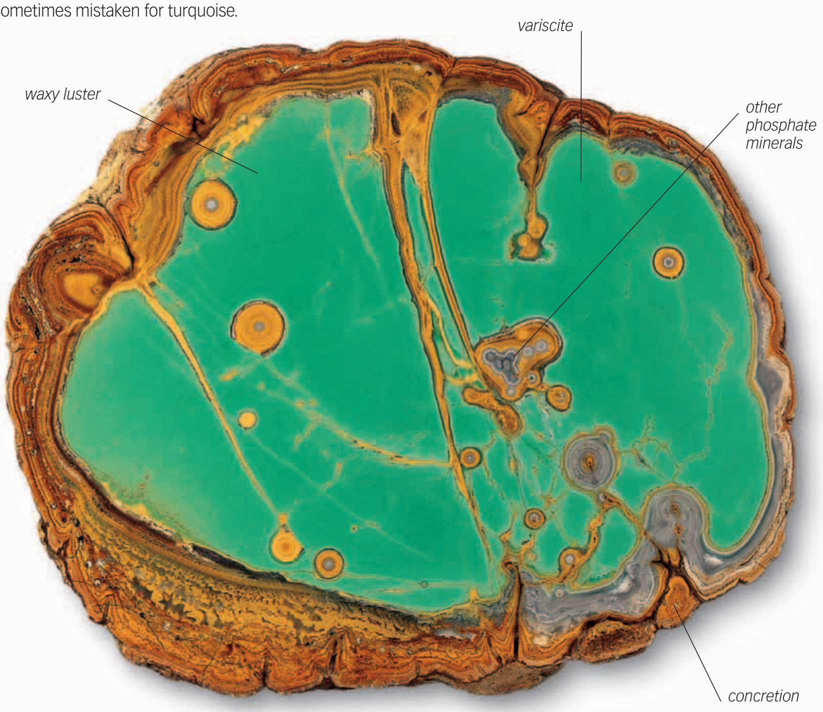

PROFILE

Orthorhombic
4½
2.6
Good but rarely visible
Splintery in massive
White
Vitreous to waxy

$AlPO_4 \cdot 2H_2O$

VARISCITE

This mineral was named after Variscia, the old name for the German district of Voightland, where it was first discovered, in 1837. Variscite is pale to apple green in color. It is predominantly found as cryptocrystalline or fine-grained masses and in veins, crusts, or nodules. It rarely forms crystals.

Variscite forms in cavities produced by the action of phosphate-rich waters on aluminous rocks. It commonly occurs in association with apatite (p.148) and wavellite (p.158). It is valued as a semiprecious gemstone and is used for carvings and as an ornamental material. Variscite is porous, and, when worn next to the skin, tends to absorb body oils, which discolor it. A mineral that appears to be turquoise (p.154) but is actually variscite is sometimes marketed by the name variquoise.

Cabochon
Variscite can be polished into inexpensive gems, but their softness makes them vulnerable to wear.

PROFILE

Monoclinic
1½–2
2.7
Perfect
Uneven
Bluish white
Vitreous to earthy

Needlelike vivianite This specimen of radiating vivianite crystals is set on a rock matrix.

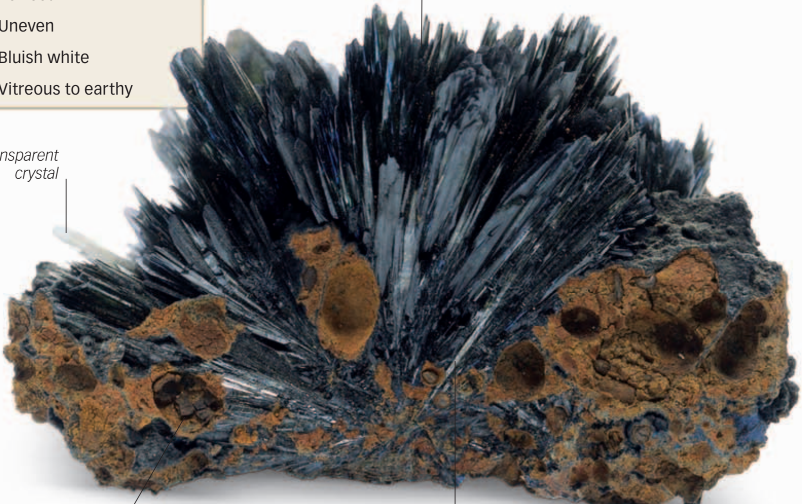
radiating, bladed crystal
thin, transparent crystal
crystals in cavity left by fossil shell
vitreous luster

VARIANT

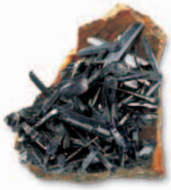

Prismatic crystal Blue-black, elongated prismatic crystals of vivianite

$Fe_3(PO_4)_2 \cdot 8H_2O$

VIVIANITE

Named in 1817 after the English mineralogist John Henry Vivian, vivianite occurs as elongated, prismatic, or bladed tabular crystals. Specimens may be rounded, corroded, concretionary, earthy, or powdery in form. Vivianite can also form starlike groups or encrustations or occur in massive or fibrous forms. Sometimes colorless when freshly exposed, vivianite becomes either pale blue to greenish blue or indigo blue on oxidation. Before the development of modern synthetic chemicals, vivianite was the source of the sought-after blue paint pigment blue ocher.

Vivianite is a widespread secondary mineral, forming in the weathered zones of iron ore and phosphate deposits and in complex granite pegmatites (p.260).

Blue ocher Powdered vivianite was used to make blue ocher, a rare and expensive pigment.

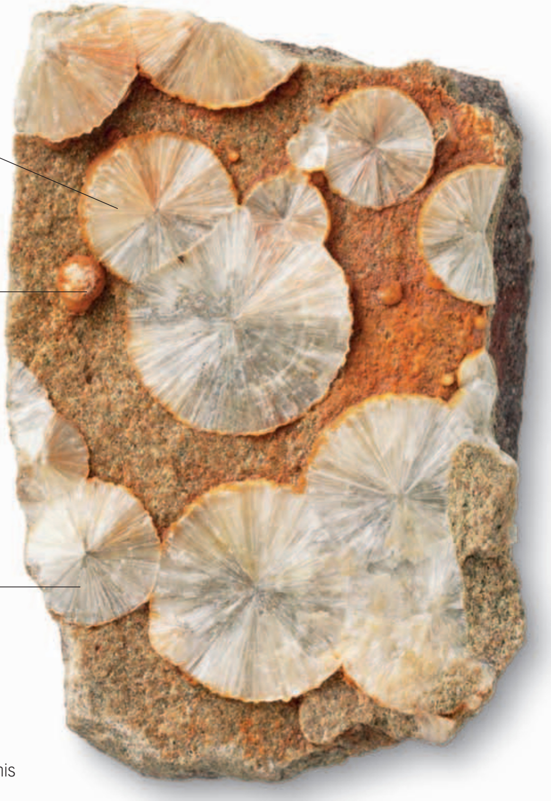

Radiating crystals
Needlelike crystals of wavellite form radiating aggregates on this rock matrix.

PROFILE

Orthorhombic
3½–4
2.4
Good
Subconchoidal to uneven
White
Vitreous to resinous

$Al_3(PO_4)_2(OH,F)_3\cdot 5H_2O$

WAVELLITE

Named in 1805 after the amateur English mineralogist William Wavell, wavellite is a hydrated aluminum phosphate. Specimens are usually green but can also be white, greenish white, green-yellow, yellowish brown, turquoise-blue, brown, or black. They may exhibit color zoning. Crystals are uncommon but when found are short to long prismatic, elongated, and striated parallel to the prism faces. Wavellite is commonly found as translucent, greenish, globular aggregates of radiating crystals up to 1¼in (3cm) in diameter, as crusts, or as stalactitic deposits.

Wavellite is a secondary mineral that forms in low-grade, aluminous, metamorphic rocks; goethite (p.102) and phosphate rock deposits; and, rarely, in hydrothermal veins. It is also found in areas where phosphate minerals have been weathered in granites (p.258) and granitic pegmatites (p.260).

Match production
Phosphorous sulfate derived from wavellite and other phosphates is a major component of matches.

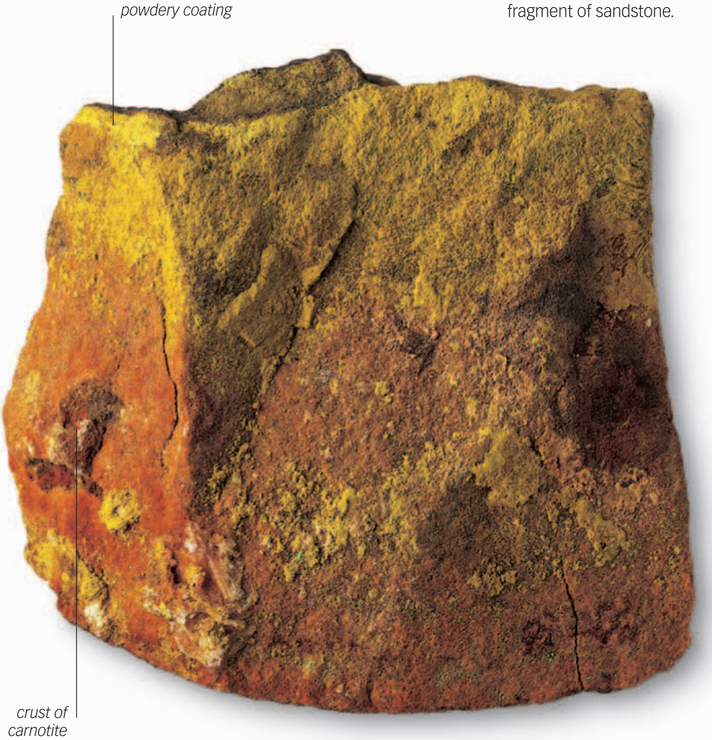

Carnotite crust
A crust of vivid yellow, powdery radioactive carnotite coats this fragment of sandstone.

PROFILE

Monoclinic
2
4.7
Perfect
Uneven
Yellow
Pearly to dull

$K_2(UO_2)_2(VO_4)_2 \cdot 3H_2O$

CARNOTITE

A radioactive mineral, carnotite was named in 1899 after the French chemist and mining engineer Marie-Adolphe Carno. It is bright to lemon yellow or greenish yellow. Carnotite is generally found as powdery or microcrystalline masses; tiny, disseminated grains; or crusts. Crystals are platy, rhombohedral, or lathlike.

Carnotite is a secondary mineral formed by the alteration of primary uranium-vanadium minerals. It occurs chiefly in sandstone (p.308), either disseminated or in concentrations around fossil wood or other fossilized vegetable matter. Pure carnotite contains about 53 percent uranium by weight, which is used to generate nuclear energy and in atomic weapons. It has also been mined for vanadium and radium, from World War II onward.

Radium dial
Sourced from carnotite, radium has been used to create illuminated watch hands and dials.

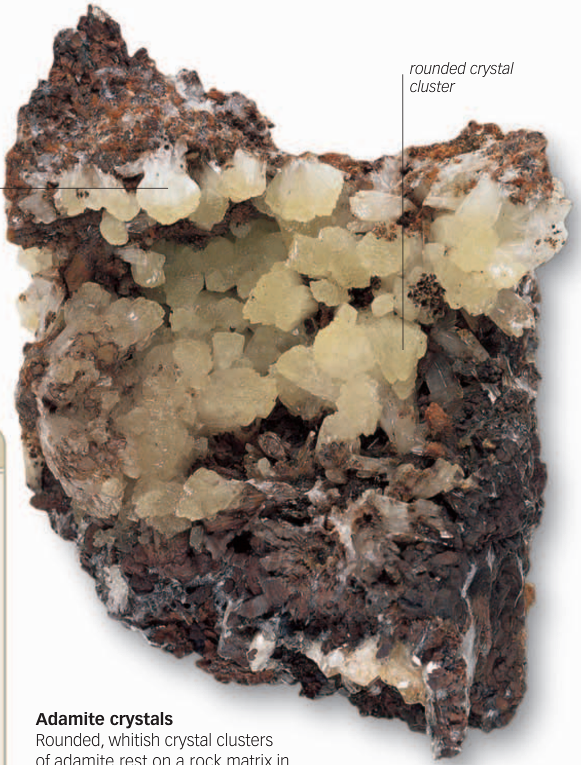

Adamite crystals Rounded, whitish crystal clusters of adamite rest on a rock matrix in this specimen.

PROFILE

Orthorhombic

3½

4.4

Good

Subconchoidal to uneven, brittle

White

Vitreous

VARIANTS

Reddish adamite Adamite crystals on a reddish orange iron-oxide matrix

adamite crystal

Spheroidal adamite A cluster of yellow, spheroidal adamite crystals on a limonite matrix

$Zn_2AsO_4(OH)$

ADAMITE

Named in 1866 after the French mineralogist G.J. Adam, who discovered adamite in Chile, this mineral is a zinc arsenate hydroxide. It is rarely colorless or white, and many specimens fluoresce green under ultraviolet light. Adamite is often brightly colored due to traces of other elements: copper commonly substitutes for zinc to yield yellow or green crystals depending on its concentration; manganese may substitute for zinc to yield crystals that are pink or violet. Adamite crystals are elongated, tabular, or blocky. This mineral also occurs as rosettes and spherical masses of radiating crystals.

Adamite forms as a secondary mineral in the oxidized zones of zinc and arsenic deposits, often associated with goethite (p.102), azurite (p.120), smithsonite (p.124), mimetite (p.164), scorodite (p.165), hemimorphite (p.227), and olivenite. Although adamite has no commercial uses, its bright and lustrous crystals are highly sought after by mineral collectors.

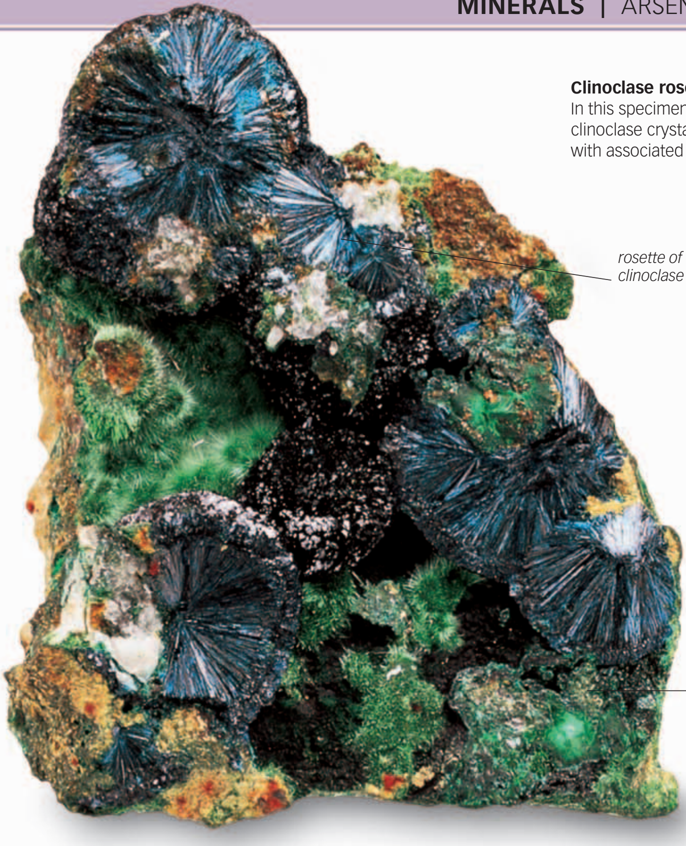

Clinoclase rosettes
In this specimen, rosettes of clinoclase crystals are seen with associated olivenite.

PROFILE

Monoclinic
2½–3
4.3
Perfect
Uneven, brittle
Bluish green
Vitreous to pearly

$Cu_3(AsO_4)(OH)_3$

CLINOCLASE

Discovered in 1830 in the Wheal Gorland mine in Cornwall, England, clinoclase was named in 1868. It takes its name from the Greek *me klísi*, which means "to incline," and *gia na spásei*, which means "to break"—a reference to its oblique basal cleavage. The vitreous crystals of clinoclase are translucent dark blue to dark greenish blue. They can be elongated or tabular or occur as single, isolated crystals that appear rhombohedral. Specimens can also form rosettelike aggregates or occur as crusts or coatings with a fibrous structure.

Clinoclase forms as a secondary mineral in the oxidized zones of deposits containing copper sulfides. Associated minerals include goethite (p.102), azurite (p.120), malachite (p.125), brochantite (p.139), adamite (p.160), quartz (p.168), and olivenite. Specimens come from Broken Hill, New South Wales, Australia; Tintic district, Utah, and Majuba Hill, Nevada, USA; the Vosges, France; and the Tsumeb Mine, Namibia.

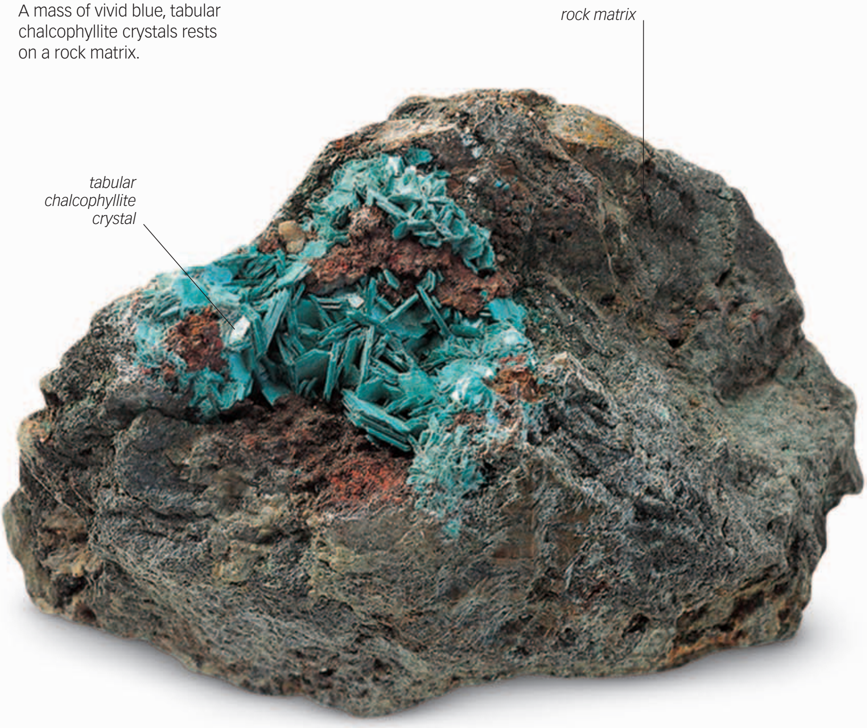

Tabular chalcophyllite
A mass of vivid blue, tabular chalcophyllite crystals rests on a rock matrix.

PROFILE

Hexagonal
2
2.7
Perfect basal
Uneven to subconchoidal
Pale green
Pearly to vitreous

$Cu_{18}Al_2(AsO_4)_2(SO_4)_3(OH)_{27}\cdot 33H_2O$

CHALCOPHYLLITE

A vivid blue-green in color, chalcophyllite takes its name from the Greek words *chalco*, which means "copper," and *phyllon*, which means "leaf"—an allusion to its copper content and its common foliated habit. Chalcophyllite was first described after material collected in Germany and named in 1847. Translucent crystals exhibit a blue-green color when viewed from one direction and appear almost colorless from another direction. Crystals are platy, six-sided, and flattened and may have triangular striations. Chalcophyllite may also be rosettelike, tabular, drusy, or massive.

This mineral occurs in hydrothermal copper deposits, often accompanied by cuprite (p.87), azurite (p.120), malachite (p.125), brochantite (p.139), and clinoclase (p.161).

Statue of Lamma
Chalcophyllite was used as an ore of copper when this copper statue was made in the period 1800–1600 BCE.

PROFILE

Monoclinic

1½–2½

3.1

Perfect

Uneven, sectile

Pale red

Adamantine to vitreous, pearly

Acicular crystals These brightly colored, needlelike crystals of erythrite are from Bou Azzer, Morocco.

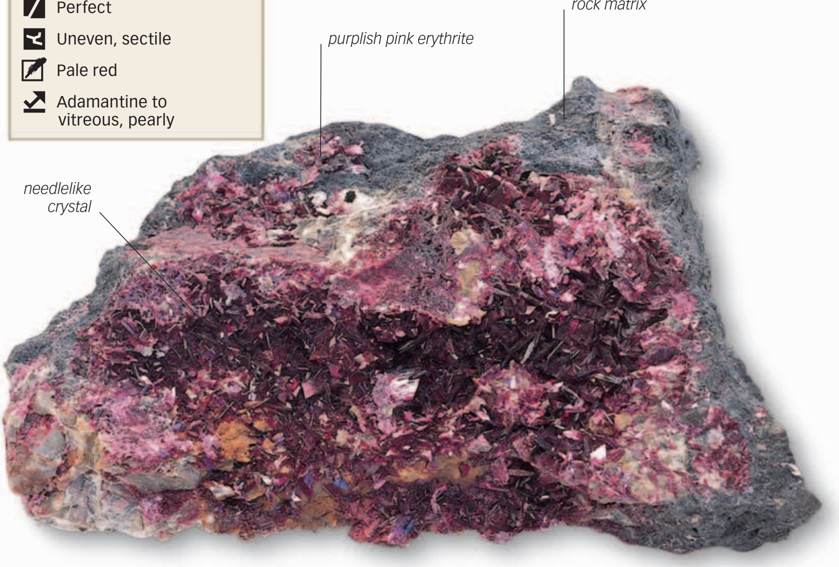

VARIANT

Erythrite crust A thin crust of erythrite on a brown rock base

$CO_3(AsO_4)_2\cdot 8H_2O$

ERYTHRITE

Although of little commercial value, erythrite is an important tool for prospectors looking for cobalt and related silver deposits. The bright purplish pink color of erythrite in a rock indicates the presence of cobalt. This explains why miners call erythrite “cobalt bloom.” Erythrite is a cobalt arsenate hydrate. It forms a chemical replacement series with annabergite, in which nickel replaces cobalt in the erythrite structure. Its color may vary from crimson red to peach red, with the lighter colors indicating a higher nickel content. The coloration may also occur in bands. Well-formed crystals are rare, but when found they occur as deeply striated, prismatic to needlelike, commonly radiating, globular tufts of crystals, or as powdery coatings.

Erythrite is a secondary mineral found in the oxidized zones of cobalt-nickel-arsenic deposits. Fine specimens come from Canada and Morocco. Erythrite is also found in Mexico, France, southwestern USA, the Czech Republic, Germany, Australia, and elsewhere.

PROFILE

Hexagonal

3½–4

7.3

Poor

Conchoidal to uneven, brittle

White

Resinous

Mimetite on manganese oxide This specimen from England contains "campylite"—a rounded variety of mimetite—and barite on nodules of manganese oxide.

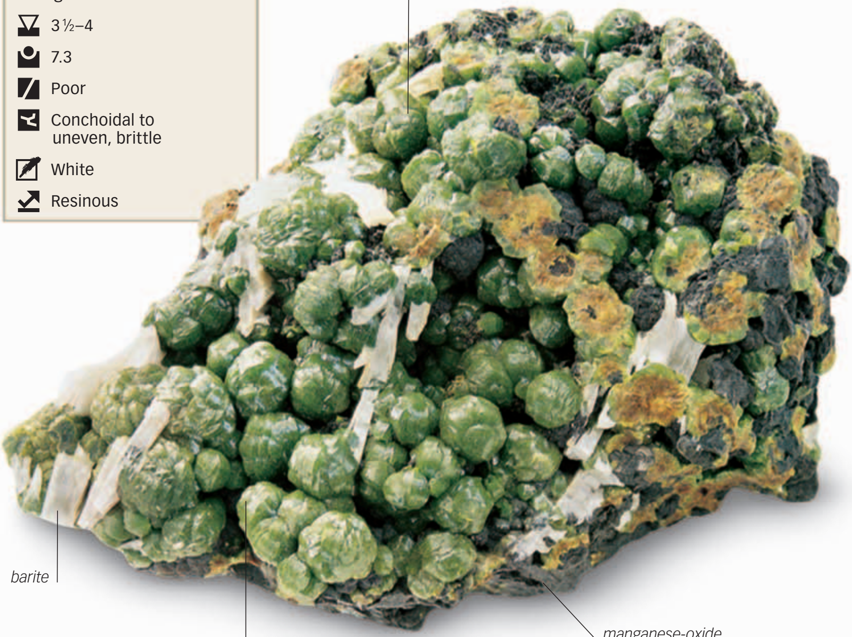

VARIANTS

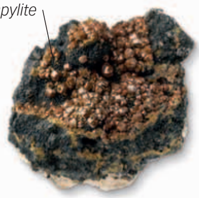

Campylite Massive and crystalline varieties of campylite set in a rock matrix

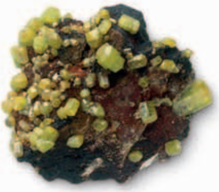

Prismatic crystals Mimetite in the form of prismatic, barrel-shaped crystals

$Pb_5(AsO_4)_3Cl$

MIMETITE

An arsenate mineral, mimetite is the end-member of a solid-solution series with pyromorphite (p.151). It is named after the Greek word *mimetes*, which means "imitator"—a reference to its resemblance to pyromorphite. Although similar in physical characterisitics and crystal form to pyromorphite, mimetite is a less common mineral. It forms heavy, barrel-shaped, hexagonal crystals or rounded masses, but is also found as botryoidal, granular, tabular, and needlelike aggregates. Mimetite specimens may be colorless or occur in shades of yellow, orange, brown, and green.

Mimetite is a secondary mineral, which forms in the oxidized zone of lead deposits and in other localities where the elements lead and arsenic occur together. Excellent specimens come from Chihuahua, Mexico; Saxony, Germany; Attica, Greece; Broken Hill, Australia; and Bisbee and Tombstone, Arizona, USA. A single crystal mined from Tsumeb in Namibia measured 2½ in (6.4 cm) in length.

PROFILE

Orthorhombic

3½–4

3.1–3.3

Imperfect

Subconchoidal

White

Vitreous to resinous or waxy

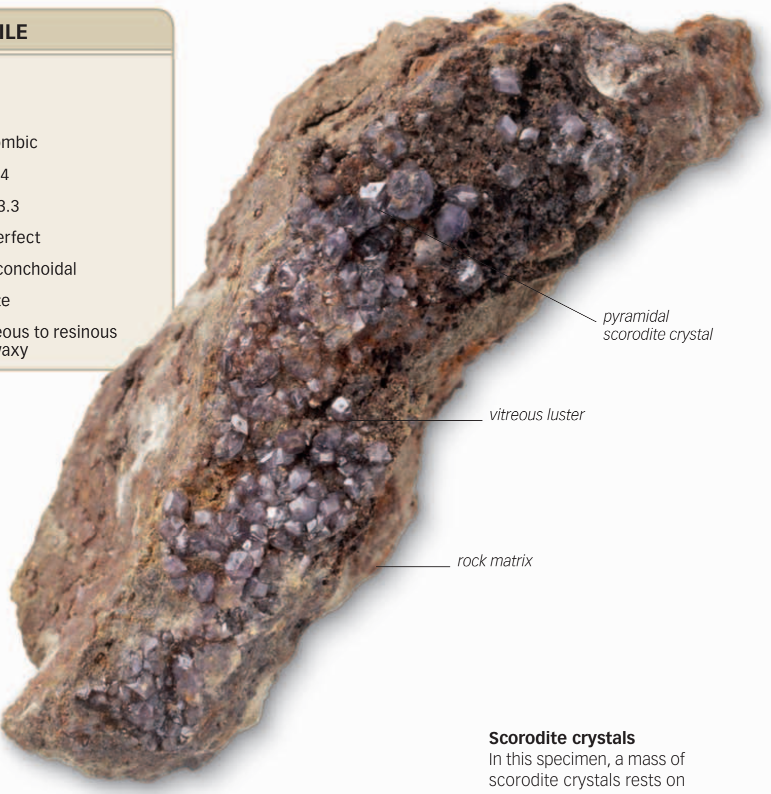

Scorodite crystals In this specimen, a mass of scorodite crystals rests on a rock matrix.

VARIANT

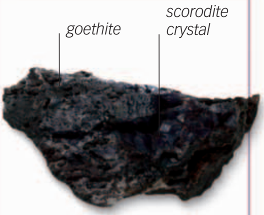

Crystal aggregate A specimen of scorodite with goethite

$FeAsO_4 \cdot 2H_2O$

SCORODITE

A hydrated iron arsenate mineral, scorodite takes its name from the Greek word *scorodion*, which means "garliclike"—an allusion to the odor emitted by the arsenic when specimens are heated. Scorodite can vary considerably in color depending on the light under which it is seen: pale leek green, grayish green, liver brown, pale blue, violet, yellow, pale grayish, or colorless. It may be blue-green in daylight but bluish purple to grayish blue in incandescent light; in transmitted light it may appear colorless to pale shades of green or brown. Crystals are usually dipyramidal, appearing octahedral, and may have a number of modifying faces. They may also be tabular or short prisms. Drusy coatings are common. Scorodite can also be porous and earthy or massive.

Scorodite is found in hydrothermal veins, hot spring deposits, and oxidized zones of arsenic-rich ore bodies. Associated minerals may be pharmacosiderite, vivianite (p.157), adamite (p.160), and various iron oxides.

SILICATES

The silicates constitute around 25 percent of all known minerals and nearly half of the most common ones. All silicates are built around a basic structure of silicon and oxygen. They are a major component of Earth and occur in lunar samples and meteorites.

COMPOSITION

The silicates make up about 95 percent of the crust and upper mantle of Earth. All silicates contain silicon and oxygen. Silicon is a lightweight, shiny metal; oxygen is a colorless, odorless gas.

In silicates, silicon and oxygen combine to form structural tetrahedra, each with a silicon atom in the center and oxygen atoms at the corners. Silicate tetrahedra may exist as discrete, independent units and connect only with other silicate tetrahedra (as in quartz), or they may link with other elements such as iron, magnesium, and aluminum. Tetrahedra may also share their oxygen atoms at corners, edges, or, more rarely, faces, creating various structures. The different linkages also create voids of different sizes, which are occupied by ions of various metals. Substitutions can occur where atoms are of a relatively similar size. Silicates are divided into six main groups (see panel, below) according to the structural configurations that result from the different ways in which tetrahedra and other elements are linked. Within these main groups are further subdivisions based on chemistry—that is, the type and location of other atoms in the structure. Many groups are solid-solution series, such as the feldspars (see panel, opposite) and the garnets, in which the ions of various metals and semimetals substitute for each other within the silicate structure.

rich purple coloring

Amethyst crystal
The tectosilicate mineral quartz occurs in several differently colored varieties. These include amethyst (above), rock crystal, smoky quartz, and citrine.

SILICATE GROUPS

The six silicate groups are based on the different ways in which the basic silica tetrahedra are linked. These differing linkages create voids of different sizes and configurations that allow positively charged atoms of different sizes to fit into the structure.

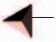
no shared corners

NESOSILICATES
Isolated groups of silica tetrahedra, with an SiO$_4$ unit in their formulae, are called nesosilicates.

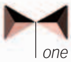
one shared corner

SOROSILICATES
These are double-tetrahedral groups with an oxygen atom shared by two tetrahedra.

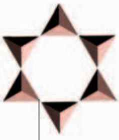
two shared corners to form rings

CYCLOSILICATES
These are closed, ringlike circles of tetrahedra that share corners.

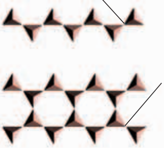
two shared corners form chains

two or three corners shared to form double chains

INOSILICATES (SINGLE-CHAIN AND DOUBLE-CHAIN)
Chains of silica tetrahedra with Si$_2$O$_6$ groups are called single-chain inosilicates.

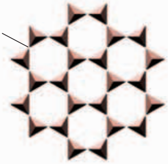
three shared corners

PHYLLOSILICATES
Flat sheets of silicate tetrahedra, with an Si$_2$O$_5$ unit in their chemical formulae, are called phyllosilicates.

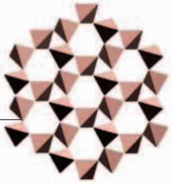
all four corners shared

TECTOSILICATES
A three-dimensional network of silica tetrahedra, linked at each corner, makes up a tectosilicate.

THE FELDSPARS

The feldspars are a group of aluminosilicate minerals that contain calcium, sodium, or potassium. As shown here, there are three major solid-solution series within the group. Feldspars are the most abundant mineral in Earth's crust.

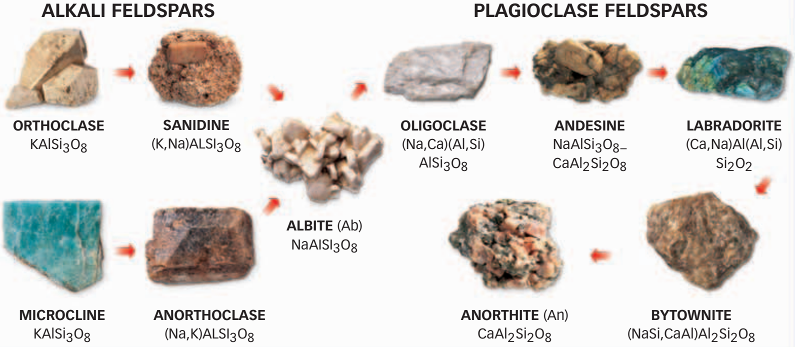
ALKALI FELDSPARS
PLAGIOCLASE FELDSPARS
ORTHOCLASE $KAlSi_3O_8$
SANIDINE $(K,Na)ALSI_3O_8$
ALBITE (Ab) $NaAlSI_3O_8$
OLIGOCLASE $(Na,Ca)(Al,Si)AlSi_3O_8$
ANDESINE $NaAlSi_3O_8–CaAl_2Si_2O_8$
LABRADORITE $(Ca,Na)Al(Al,Si)Si_2O_2$
MICROCLINE $KAlSi_3O_8$
ANORTHOCLASE $(Na,K)ALSI_3O_8$
ANORTHITE (An) $CaAl_2Si_2O_8$
BYTOWNITE $(NaSi,CaAl)Al_2Si_2O_8$

OCCURRENCE AND USES

The ultimate source of all silicates is igneous rock in which tectosilicate feldspar minerals are the major component. Silicates are found not only on Earth but also on the Moon and in meteorites. After feldspar, quartz is the most abundant mineral in the crust and upper mantle. It occurs in nearly all high-silica igneous, metamorphic, and sedimentary rocks. In silica-poor rocks where quartz does not form, other silicate minerals develop. Since many silicates, especially quartz and its varieties, are resistant to weathering, they form the major component of most detrital sediments. There are numerous uses of silicates. Quartz and its varieties find use as gemstones, in electronic and optical applications, and as abrasives. The feldspars are used in glass and ceramics, as gemstones, and as abrasives. Other silicates are ores, and yet others are important gem, ornamental, and industrial minerals. The tough rocks formed from silicate minerals are used as major building and industrial materials.

Jadeite mask
The inosilicate jadeite was one of the first minerals used by humans. Although difficult to shape, it has been used for tools and ornaments.

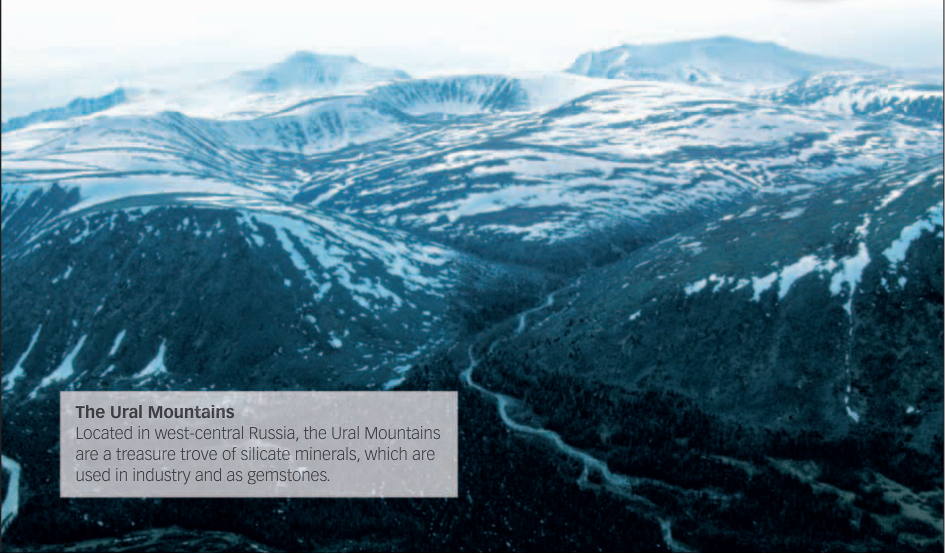

The Ural Mountains
Located in west-central Russia, the Ural Mountains are a treasure trove of silicate minerals, which are used in industry and as gemstones.

PROFILE

Hexagonal

7

2.7

None

Conchoidal

White

Vitreous

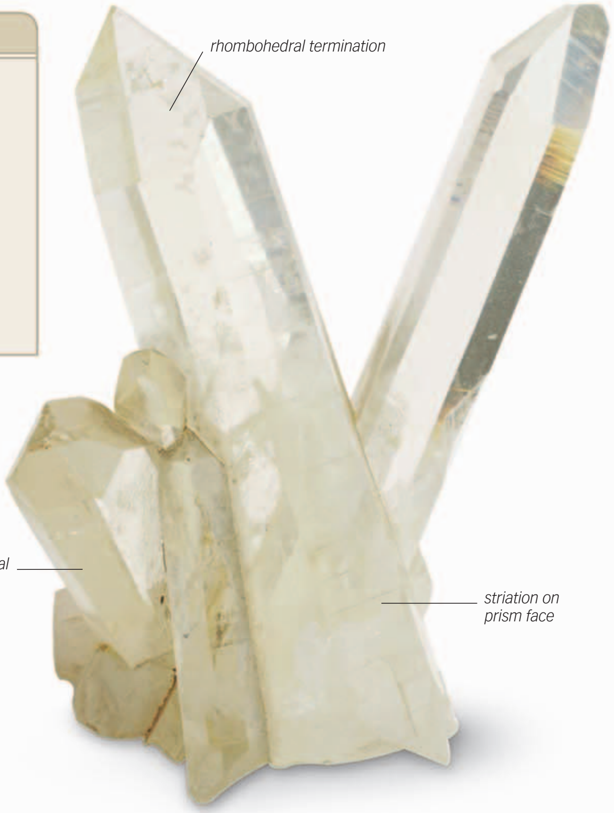

Prismatic quartz This group of long, prismatic quartz crystals is from the Dauphiné province of France.

VARIANTS

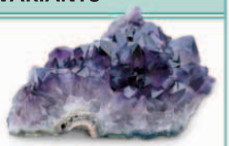

Pyramidal amethyst An amethyst specimen with pyramidal terminations

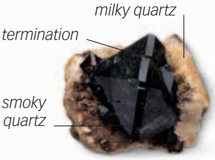

Smoky quartz Double-terminated smoky quartz in milky quartz

termination

Milky quartz A white, terminated quartz prism

SiO_2

QUARTZ

One of the most common minerals in Earth's crust, quartz has two forms: macrocrystalline (with crystals that can be seen by eye) and cryptocrystalline (formed of microscopic crystals). Macrocrystalline quartz is usually colorless and transparent, as in rock crystal, or white and translucent, as in milky quartz. Colored varieties include: pink and translucent rose quartz; transparent to translucent lavender or purple amethyst; transparent to translucent black or brown smoky quartz; and transparent to translucent yellow or yellow-brown citrine. All crystalline varieties form hexagonal prisms and pyramids.

Cryptocrystalline varieties of quartz include chalcedony (p.169), agate (p.170), and jasper (p.171). Quartz occurs in nearly all silica-rich sedimentary, igneous, and metamorphic rocks.

Oval citrine This large, oval-cut citrine is set in a silver brooch. It is encircled by silver leaves and faceted amethysts.

PROFILE

Hexagonal
7
2.7
None
Uneven
White
Waxy to dull

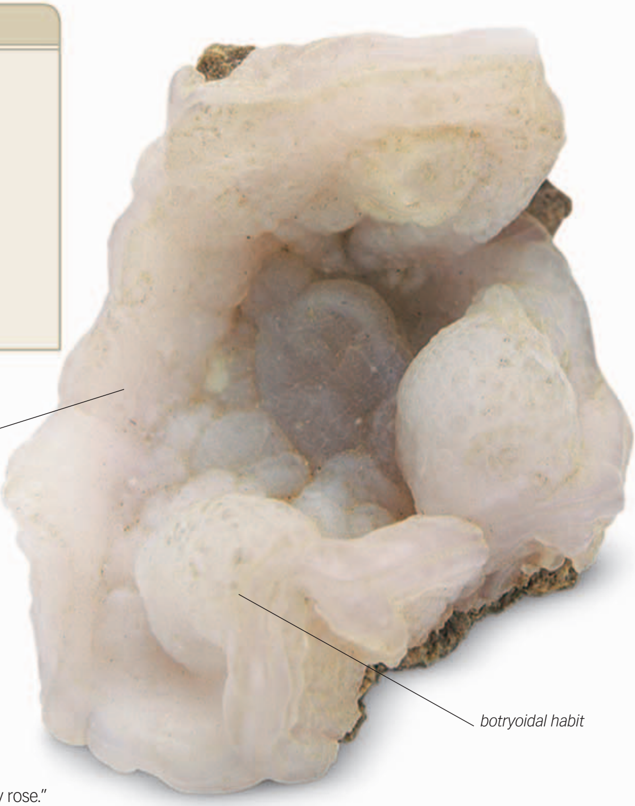

Pink chalcedony
This form of botryoidal pink chalcedony is sometimes referred to as a "chalcedony rose."

VARIANTS

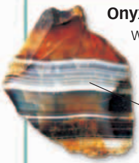

Onyx Chalcedony with straight banding

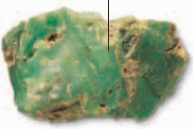

Chrysoprase Chalcedony colored green by nickel

Carnelian A piece of red-orange chalcedony

SiO_2

QUARTZ: CHALCEDONY

A compact variety of microcrystalline quartz (p.168), chalcedony may have been named after the ancient port of Khalkedon in Asia Minor (now Turkey), where there were extensive deposits of this mineral. Chalcedony is white when pure, but it may contain microscopic inclusions of other minerals, which give it a range of colors. This mineral is composed of microscopic fibers and can be mamillary, botryoidal, or stalactitic.

Many chalcedonies are semiprecious gems and have their own names. Chalcedony with distinct banding is called agate (p.170). All varieties of chalcedony occur worldwide. It is found in veins, concretions, and geodes. It forms in cavities, cracks, and when silica-rich waters at low temperatures (up to 400°F/200°C) percolate through existing rocks.

Chalcedony blade
This Aztec sacrificial knife has a finely chipped chalcedony blade and a mosaic handle.

PROFILE

Hexagonal

7

2.7

None

Conchoidal

White

Vitreous to waxy

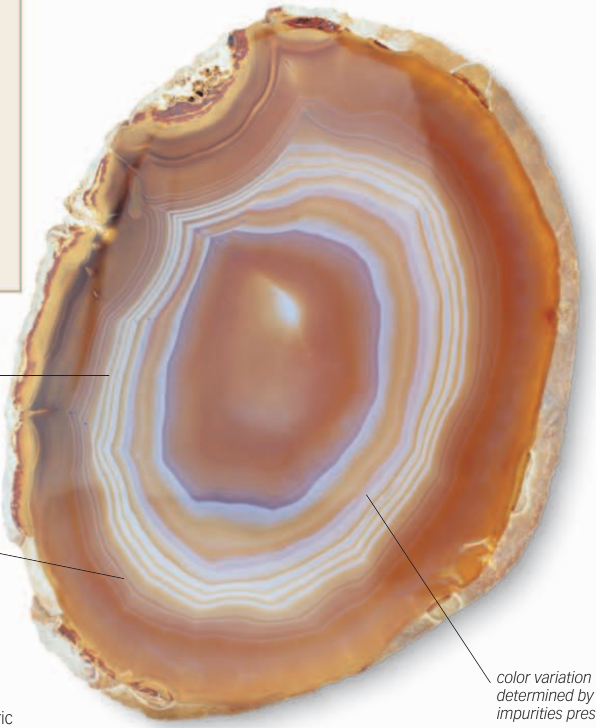
concentric bands of agate
shapes of bands follow outline of cavity
color variation determined by impurities present

Brazilian agate
This cross section of a Brazilian agate nodule shows the concentric layering typical of agate.

VARIANT

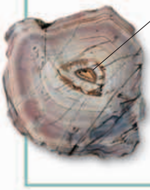
concentric bands

Fortification agate
Banded agate with angular bends

SiO_2

CHALCEDONY: AGATE

A common, semiprecious chalcedony, agate has been worked since prehistoric times. It is a compact, microcrystalline variety of quartz (p.168), and it has the same physical properties as quartz. Agate is characterized by concentric color bands in shades of white, yellow, gray, pale blue, brown, pink, red, or black.

Other names often precede the word agate to indicate the mineral's visual characteristics or place of origin. One of these is fire agate, which has inclusions of reddish to brown hematite that give an internal iridescence to polished stones. Another is fortification agate, which has concentric bands of color resembling an aerial view of an ancient fortress. Most agates are found in cavities in ancient lavas or other extrusive igneous rocks.

Snuff bottle
The 19th-century Chinese snuff bottle seen here has been carved from agate. It has a jade stopper.

PROFILE

Hexagonal

7

2.7

None

Conchoidal

White

Vitreous

Color variation
Hematite colors this example of jasper brownish red. Threads of white quartz veins make a crisscross pattern on this specimen.

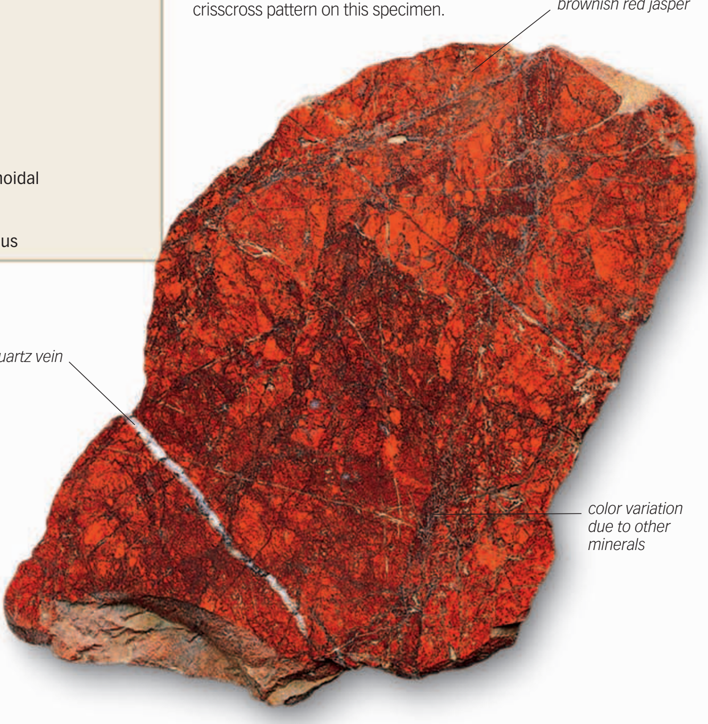

VARIANTS

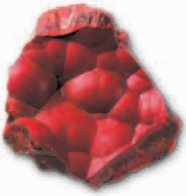

Mammillary jasper Red jasper in mammillary form

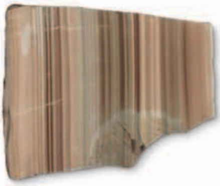

Ribboned jasper
A specimen of jasper with parallel, reddish bands

SiO_2

QUARTZ: JASPER

An impure variety of cryptocrystalline quartz (p.168), jasper takes its name from the Greek word *iaspis*, which is probably of Semitic origin. It is fine-grained or dense, and it contains various amounts of other materials, which give it opacity and color. Hematite (p.91) gives jasper a brick-red to brownish red color; clay a yellowish white or gray color; and goethite (p.102) a brown or yellow hue.

Jasper forms when silica-rich waters at low temperatures (up to 400°F/200°C) percolate through cracks and fissures in other rocks, incorporating a variety of materials and leaving behind deposits. It is found worldwide wherever cryptocrystalline quartz occurs. The classification and naming of jasper varies greatly and often incorporates place names or colors. Only some of these are formally recognized as varieties of jasper, leaving great latitude in defining which jasper is which. Color names such as "red" or "green" can apply to a range of shades, while locality names, such as "Bruneau jasper" after a canyon in Idaho, tend to be more specific.

PROFILE

Crystal system Amorphous

5–6

1.9–2.3

None

Conchoidal

White

Vitreous

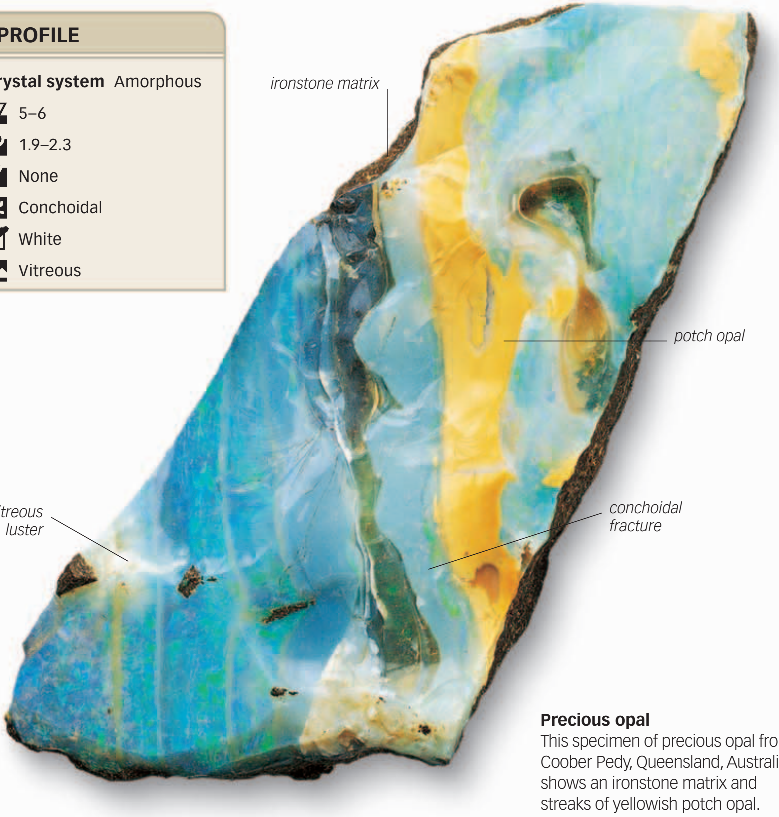

Precious opal
This specimen of precious opal from Coober Pedy, Queensland, Australia, shows an ironstone matrix and streaks of yellowish potch opal.

VARIANTS

Boulder opal Blue mass of opal in an iron-oxide nodule

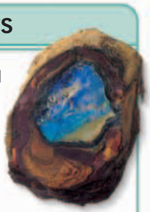

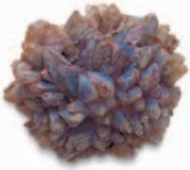

Opal pseudomorph Crystals of glauberite replaced by opal

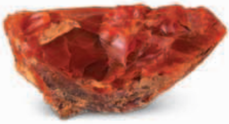

Fire opal Nonidirescent, transparent opal

$SiO_2 \cdot nH_2O$

OPAL

Known since antiquity, opal derives its name from the Roman word *opalus,* which means "precious stone." Although it is colorless when pure, the vast majority of common opal occurs in opaque, dull yellows and reds. It varies from essentially amorphous to partially crystalline. Precious opal is the least crystalline form of the mineral, consisting of a regular arrangement of tiny, transparent, silica spheres. Regularly arranged spheres of a particular size create a diffraction effect called color play.

Opal is widespread and is deposited at low temperatures (up to 400°F/200°C) from silica-bearing, circulating waters. It is found as nodules, stalactitic masses, veinlets, and encrustations in most kinds of rocks. Opal constitutes important parts of many sedimentary accumulations, such as diatomaceous earth.

Victorian ring
Some cut opal dries and cracks with age and needs to be kept moist. The opal in this ring is well preserved.

PROFILE

Monoclinic

6–6½

2.5–2.6

Perfect

Subconchoidal to uneven, brittle

White

Vitreous

Orthoclase prisms In this specimen, white, blocky prisms of orthoclase are associated with cleavelandite albite and set in pegmatite.

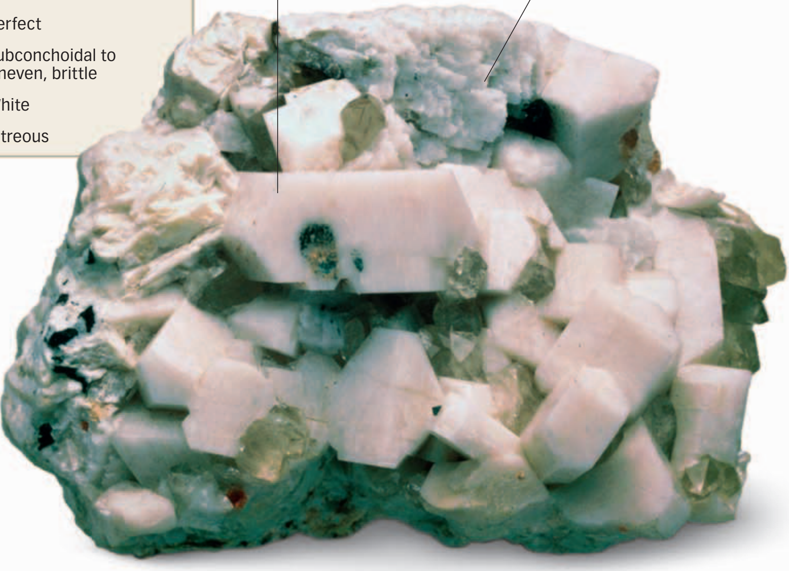

VARIANTS

Yellow orthoclase A crystal of yellow orthoclase

Moonstone rough An opalescent variety of orthoclase

twinned crystal
prismatic crystal

Orthoclase crystals Twinned orthoclase with smaller prism

$KAlSi_3O_8$

ORTHOCLASE

An important rock-forming mineral, orthoclase is the potassium-bearing end member of the potassium—sodium feldspar solid-solution series. It is a major component of granite (pp.258–59)—its pink crystals give granite its typical color. Crystalline orthoclase can also be white, colorless, cream, pale yellow, or brownish red. Orthoclase appears as well-formed, short, prismatic crystals, which are frequently twinned. It may also occur in massive form. Moonstone is a variety of orthoclase that exhibits a schiller effect.

Pure orthoclase is rare—some sodium is usually present in the structure. Specimens are abundant in igneous rocks rich in potassium or silica, in pegmatites (p.260), and in gneisses (p.288). This mineral is important in ceramics, to make the item itself and as a glaze.

Moonstone-set brooch Orthoclase exhibits the schiller effect which creates the shimmer seen on the moonstones in this brooch.

Prismatic sanidine
This single, well-formed prismatic crystal of sanidine rests in a matrix of the volcanic rock trachyte.

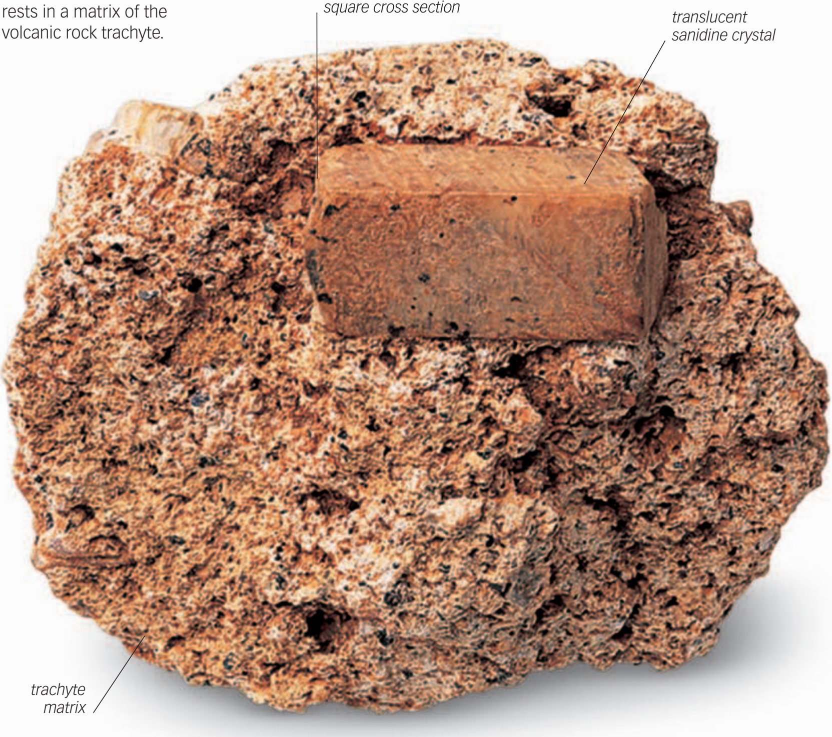

PROFILE

Monoclinic
6–6½
2.6
Perfect, good
Conchoidal to uneven
White
Vitreous

$(K,Na)AlSi_3O_8$

SANIDINE

A member of the solid-solution series of potassium and sodium feldspars, sanidine is the high-temperature form of potassium feldspar, forming at 1,065°F (575°C) or above. Crystals are usually colorless or white, glassy, and transparent, but they may also be gray, cream, or occur in other pale tints. They are generally short prismatic or tabular, with a square cross section. Twinning is common. Crystals have been known to reach 20in (50cm) in length. Sanidine is also found as granular or cleavable masses.

A widespread mineral, sanidine occurs in feldspar- and quartz-rich volcanic rocks, such as rhyolite (p.278), phonolite, and trachyte (p.279). It is also found in eclogites (p.299), contact metamorphic rocks, and metamorphic rocks formed at low pressure and high temperature. Sanidine forms spherular masses of needlelike crystals in obsidian (p.280), giving rise to what is called snowflake obsidian. Significant occurrences of sanidine are at the Alban Hills near Rome, Italy; Mont St.-Hilaire, Canada; and Eifel, Germany.

PROFILE

Triclinic

6–6½

2.6

Perfect, good

Conchoidal to uneven, brittle

White

Vitreous, dull

Prismatic microcline Numerous prismatic crystals of light-colored microcline sit atop a pegmatite matrix, along with smoky quartz.

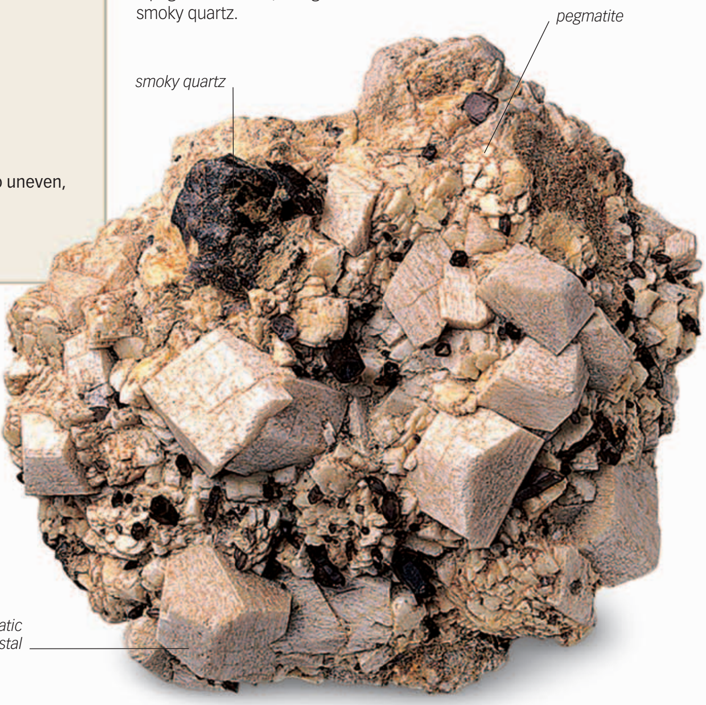

VARIANT

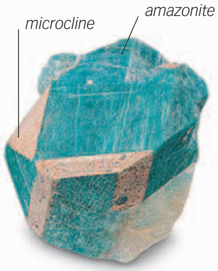

Amazonite A single crystal of blue-green amazonite, a variety of microcline

$KAlSi_3O_8$

MICROCLINE

Used in ceramics and as a mild abrasive, microcline is one of the most common feldspar minerals. It can be colorless, white, cream to pale yellow, salmon pink to red, or bright green to blue-green. Microcline forms short prismatic or tabular crystals that are often of considerable size: single crystals can weigh several tons and reach yards in length. Crystals are often multiply twinned, with two sets of fine lines at right angles to each other. This gives a "plaid" effect that is unique to microcline among the feldspars. Microcline can also be massive.

The mineral occurs in feldspar-rich rocks, such as granite (pp.258–59), syenite (p.262), and granodiorite (p.263). It is found in granite pegmatites (p.260) and in metamorphic rocks, such as gneisses (p.288) and schists (p.291–92).

Amazonite cabochon This Arts and Craft ring exhibits an asymmetrically set cabochon of amazonite in a rose-and-foliage design.

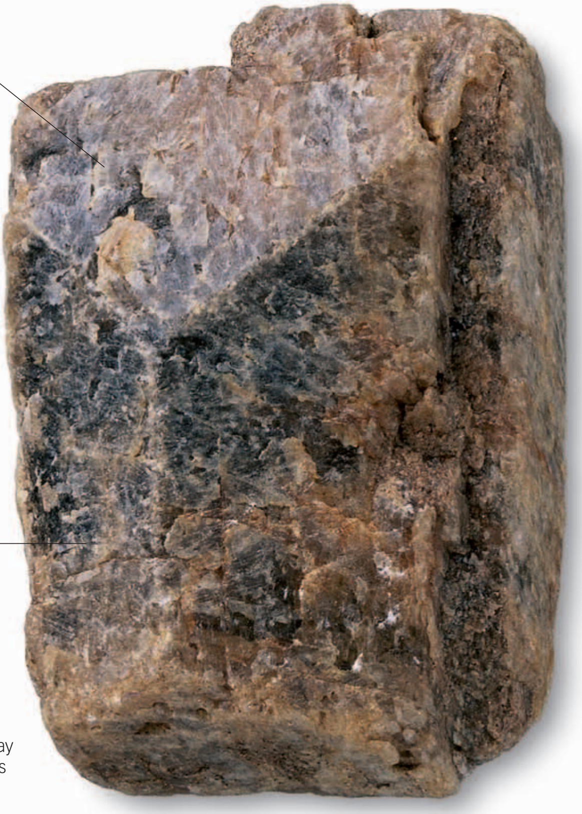

Prismatic anorthoclase
This specimen of pink and gray prismatic anorthoclase shows well-developed crystal faces.

PROFILE

Triclinic
6–6½
2.6
Perfect, good
Conchoidal to uneven, brittle
White
Vitreous

$(Na,K)AlSi_3O_8$

ANORTHOCLASE

This member of the sodium- and potassium-rich feldspar group takes its name from the Greek word *anorthos*, which means “not straight”—a reference to its oblique cleavage. Anorthoclase is colorless, white, cream, pink, pale yellow, gray, or green. Its crystals are prismatic or tabular and are often multiply twinned. Anorthoclase crystals can show two sets of fine lines at right angles to each other like microcline (p.175), but the lines are much finer. Specimens can also be massive or granular.

Anorthoclase forms in sodium-rich igneous zones. It commonly occurs with ilmenite (p.90), apatite (p.148), and augite (p.211). Much anorthoclase exhibits a gold, bluish, or greenish schiller effect, making it one of several feldspars known as moonstone when cut *en cabochon*. A type of the igneous rock syenite (p.262) called larvikite has large schillerized crystals of anorthoclase and is highly prized as an ornamental stone. Anorthoclase is widespread, but fine examples come from Cripple Creek, Colorado, USA; Larvik, Norway; and Fife, Scotland.

PROFILE

Triclinic

6–6½

2.6

Perfect, good

Conchoidal to uneven, brittle

White

Vitreous to pearly

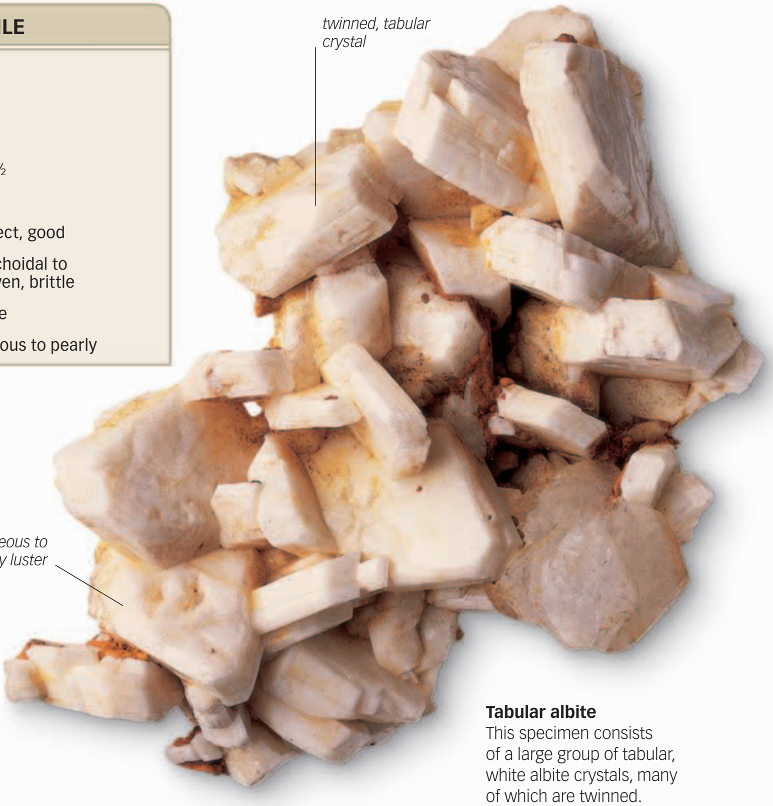

Tabular albite This specimen consists of a large group of tabular, white albite crystals, many of which are twinned.

VARIANT

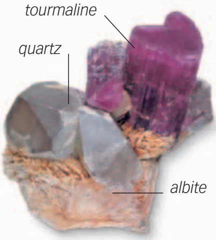

Albite base Tourmaline and quartz crystals on albite

$NaAlSi_3O_8$

ALBITE

A rock-forming mineral, albite takes its name from the Latin word *albus*, which means “white”—a reference to its usual color. Specimens can also be colorless, yellowish, pink, or green. Albite occurs as tabular or platy crystals that are often twinned, glassy, and brittle. It can also be massive or granular. Albite was named in 1707.

This mineral is the solid-solution end member of both the plagioclase and the sodium- and potassium-rich feldspars. It occurs in pegmatites (p.260) and in some feldspar- and quartz-rich igneous rocks. Albite also forms through chemical processes in certain sedimentary environments and occurs in low-grade metamorphic rocks. The cleavelandite variety occurs in complex pegmatites as thin plates or scales.

Facet-grade albite Faceted albite, although fragile, is sometimes used in jewelry, along with albite’s moonstone variety.

PROFILE

Triclinic

6

2.6

Perfect

Conchoidal to uneven, brittle

White

Vitreous

Massive oligoclase This typical massive specimen of oligoclase is from Penland, Mitchell County, North Carolina.

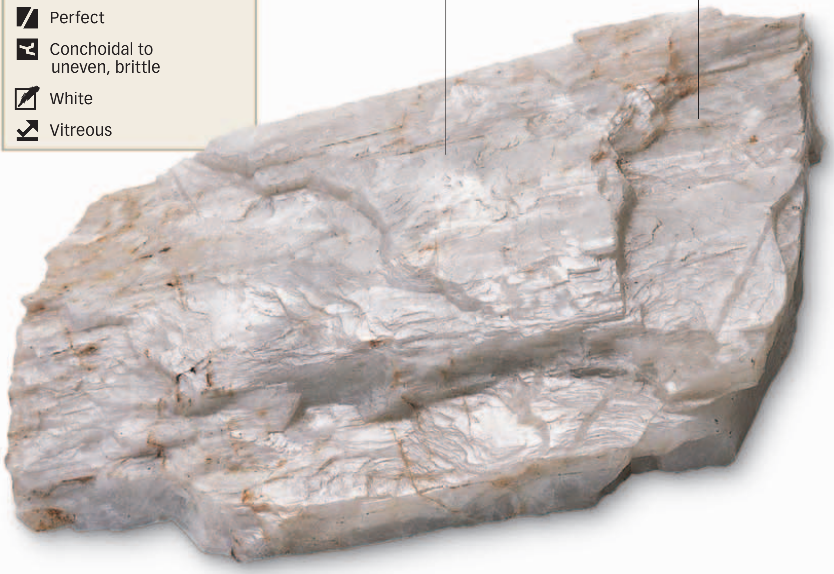

VARIANTS

Sunstone rough An uncut specimen of sunstone oligoclase

smoky quartz

Oligoclase crystal A pink crystal accompanied by smoky quartz

$(Na,Ca)Al_2Si_2O_8$

OLIGOCLASE

In 1826, the German mineralogist August Breithaupt named this mineral after two Greek words: *oligos*, which means “little,” and *clasein*, which means “to break”—because it was thought to have a less perfect cleavage than albite (p.177). Oligoclase can be gray, white, red, greenish, yellowish, brown, or colorless. Its usual habit is massive or granular, although it can form tabular crystals that are commonly twinned.

Oligoclase is the most common of the plagioclase feldspars. It occurs in granite (pp.258–59), granitic pegmatites (p.260), diorite (p.264), rhyolite (p.278), and other feldspar- and quartz-rich igneous rocks. It also occurs in high-grade, metamorphosed gneisses (p.288) and schists (pp.291–92).

Semiprecious oligoclase Sunstone oligoclase, such as the oval example seen here, has hematite or goethite inclusions.

PROFILE

Triclinic

6–6½

2.7

Perfect

Conchoidal to uneven, brittle

White

Vitreous

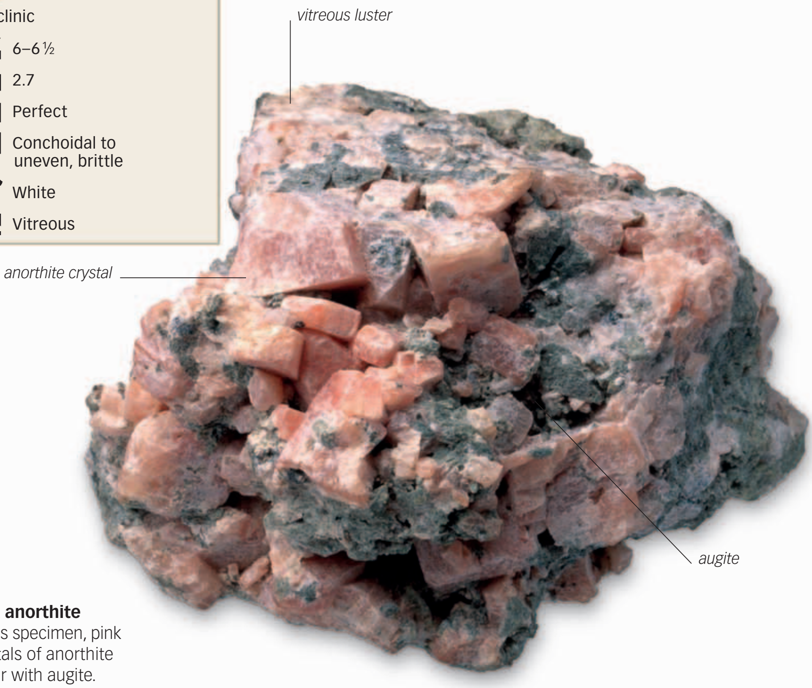

Pink anorthite In this specimen, pink crystals of anorthite occur with augite.

VARIANT

Anorthite aggregate A mass of blue-gray anorthite

$CaAl_2Si_2O_8$

ANORTHITE

The calcium-rich end-member of the plagioclase-feldspar solid-solution series, anorthite takes its name from the Greek word *anorthos*, which means “not straight”—a reference to its triclinic form. Its brittle, short, glassy crystals are well-formed prisms that can be colored white or shades of gray, pink, or red. Specimens are also massive or granular. Anorthite is a calcium aluminosilicate and can contain up to 10 percent albite (p.177).

Anorthite is a major rock-forming mineral present in many magnesium- and iron-rich igneous rocks, contact metamorphic rocks, and chondroditic meteorites (p.337). Pure anorthite is uncommon; it weathers readily and is rare in rocks exposed at the surface for long periods. Anorthosite (p.261), a rock composed mainly of anorthite, makes up much of the lunar highlands. The so-called Genesis Rock, brought back by Apollo 15, is made of anorthosite and dates back to the formation of the Moon, which occurred about 4.1 billion years ago. Anorthite was also discovered in the comet Wild 2.

PROFILE

Triclinic

6–6½

2.7

Perfect

Uneven to conchoidal

White

Vitreous

Blue labradorite
This specimen shows polysynthetic twinning typical of plagioclase feldspars. This is evident as a series of parallel lines on the broken faces.

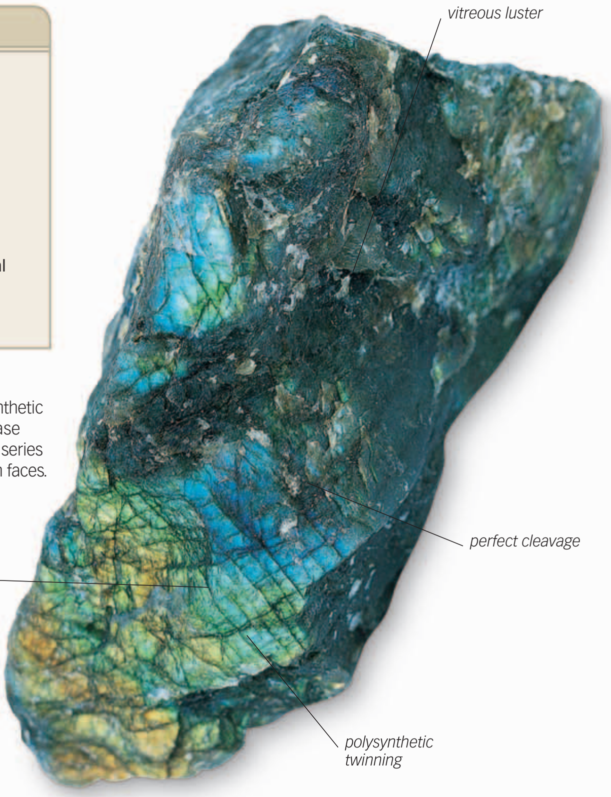

VARIANTS

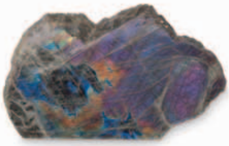
Schiller effect Orange, purple, and blue flashes visible in a specimen of labradorite

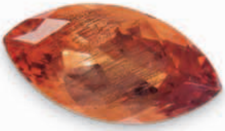
Orange sunstone Labradorite "sunstone" from Oregon

$NaAlSi_3O_8$–$CaAl_2Si_2O_8$

LABRADORITE

The calcium-rich, middle-range member of the plagioclase feldspars, labradorite is characterized by its schiller effect—a rich play of iridescent colors, mainly blue, on cleavage surfaces. Specimens are generally blue or dark gray but can also be colorless or white. When transparent, labradorite is yellow, red, orange, or green. This mineral seldom forms crystals, but when crystals do occur, they are tabular. It most often forms masses with crystals that can be microscopic or up to 3 ft (1 m) or more wide.

Labradorite is a major constituent of certain medium-silica and silica-poor igneous and metamorphic rocks, including diorite (p.264), gabbro (p.265), basalt (p.273), andesite (p.275), and amphibolite (p.296). Gem-quality labradorite from Finland is known as spectrolite.

Semiprecious gemstone The polished oval of labradorite in this choker beautifully displays the stone's rainbow iridescence.

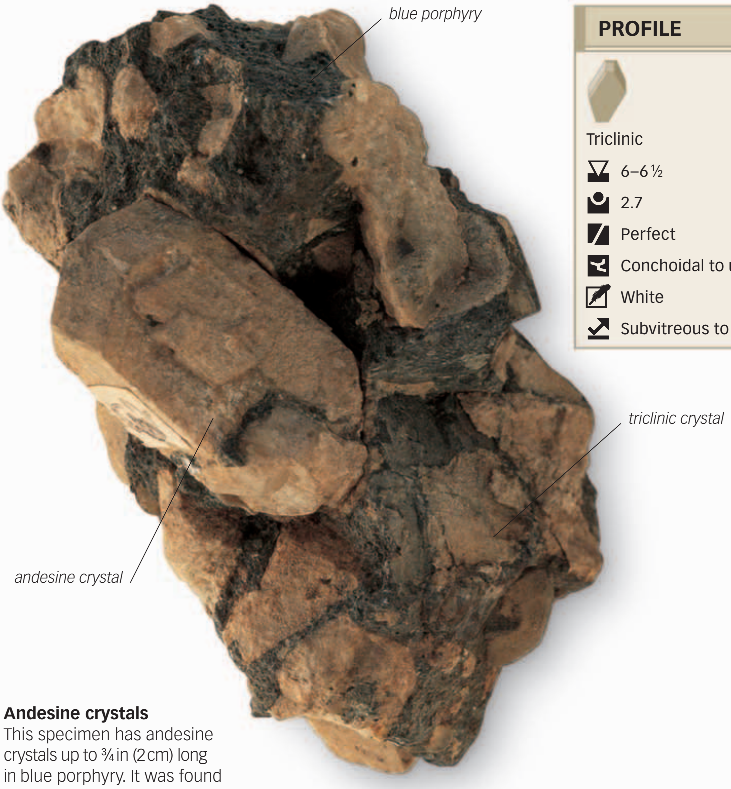

Andesine crystals This specimen has andesine crystals up to ¾ in (2 cm) long in blue porphyry. It was found in Estérel, Var, France.

PROFILE

Triclinic
6–6 ½
2.7
Perfect
Conchoidal to uneven
White
Subvitreous to pearly

VARIANT

Andesite porphyry Andesite is a major constituent of this porphyry rock

$NaAlSi_3O_8–CaAl_2Si_2O_8$

ANDESINE

The plagioclase feldspar andesine is named after the Andes Mountains in South America, where it is abundant in andesite lavas. A white, gray, or colorless mineral, andesine often forms well-defined crystals that usually exhibit multiple twinning. It can also be massive or occur as rock-bound grains.

A sodium calcium aluminosilicate, andesine is an intermediate member of the plagioclase solid-solution series. It occurs widely in igneous rocks of medium silica content, especially in andesite (p.275). Andesine is also found in other intermediate igneous rocks, such as syenite (p.262) and diorite (p.264). Specimens are commonly associated with magnetite (p.92), quartz (p.168), biotite (p.197), and hornblende (p.218). Andesine typically occurs in metamorphic rocks formed under high pressure and temperatures (1,065°F/575°C or above). It is also found as detrital grains in sedimentary rocks. The accurate identification of individual specimens involves detailed study and analysis.

Massive nepheline
This specimen of nepheline from Arkansas shows the mineral's most typical massive habit.

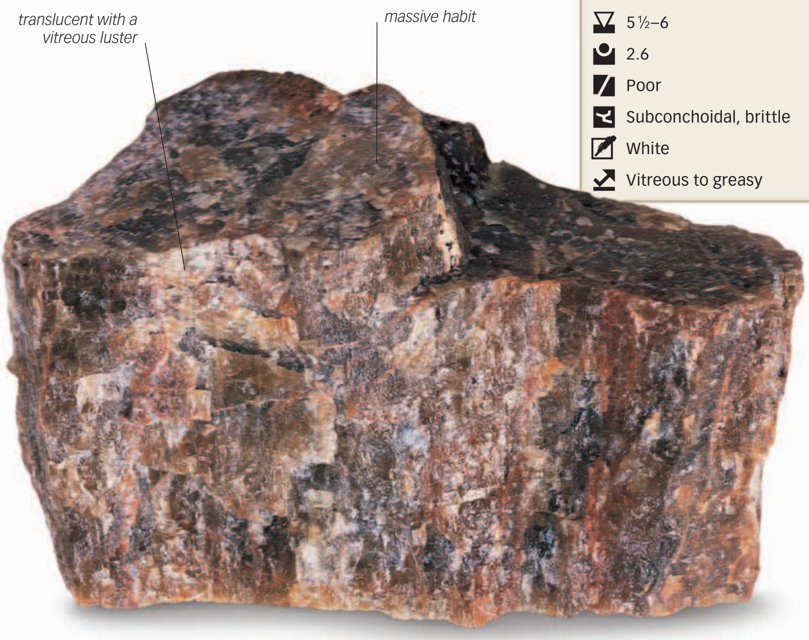

PROFILE

Hexagonal
5½–6
2.6
Poor
Subconchoidal, brittle
White
Vitreous to greasy

VARIANT

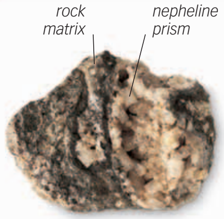

Nepheline crystals Well-developed crystals within cavities in a rock matrix

$(Na,K)AlSiO_4$

NEPHELINE

The most common feldspathoid mineral, nepheline takes its name from the Greek word *nephele*, which means "cloud"—a reference to the fact that the mineral becomes cloudy or milky in strong acids. Specimens are usually white in color, often with a yellowish or grayish tint. They can also be colorless, gray, yellow, or red-brown. Nepheline is generally massive. Crystals usually occur as hexagonal prisms, although they can exhibit a variety of prism and pyramid shapes. Nepheline also forms large, tabular phenocrysts in igneous rocks.

This rock-forming mineral is found in iron- and magnesium-rich igneous rocks with perovskite (p.89), spinel (p.96), and olivine (p.232). It also occurs in intermediate igneous rocks with aegirine (p.209) and augite (p.211) and in some volcanic and metamorphic rocks.

Ceramic bowl
Nepheline is sometimes used as a substitute for feldspars in ceramics, such as this porcelain bowl.

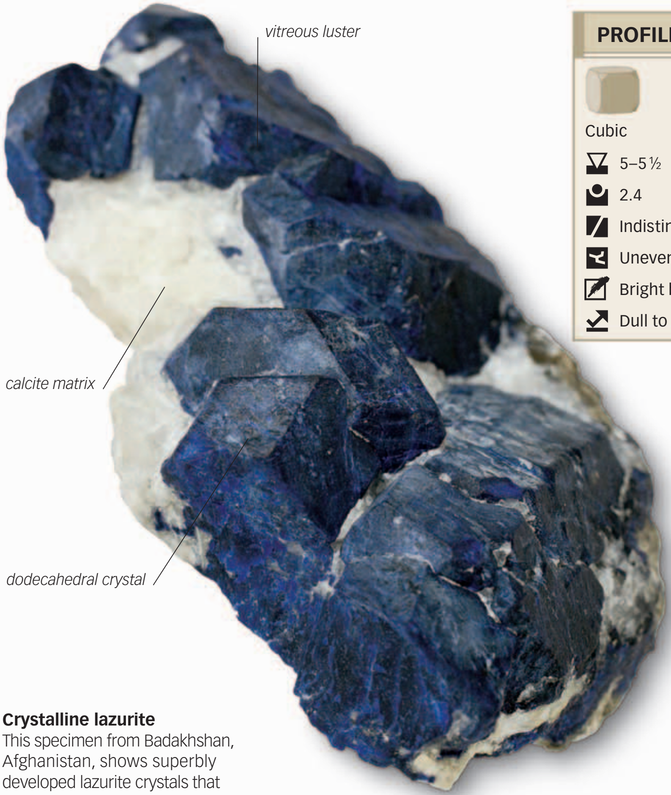

Crystalline lazurite This specimen from Badakhshan, Afghanistan, shows superbly developed lazurite crystals that are up to ¾ in (1.9 cm) long.

PROFILE

Cubic
5–5½
2.4
Indistinct
Uneven, brittle
Bright blue
Dull to vitreous

VARIANTS

Polished slab A slice of lazurite showing intense color

Lapis lazuli An uncut piece of lapis lazuli streaked with calcite

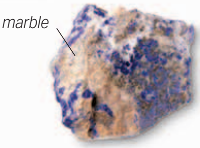

Lazurite in marble Lazurite dispersed in marble

$Na_3Ca(Al_3Si_3O_{12})S$

LAZURITE

A sodium calcium aluminosilicate, lazurite is the main component of the gemstone lapis lazuli and accounts for the stone's intense blue color, although lapis lazuli also typically contains pyrite (p.62), calcite (p.114), sodalite (p.184), and haüyne. Lazurite specimens are always deep or vibrant blue. Distinct crystals were thought to be rare until large numbers were brought out of mines in Badakhshan, Afghanistan, in the 1990s. These are usually dodecahedral and are much sought after. Most lazurite is either massive or occurs in disseminated grains.

Lapis lazuli is relatively rare. It forms in crystalline limestones (p.319) as a product of contact metamorphism. The best quality lapis lazuli is dark blue with minor patches of calcite and pyrite. In addition to its use as a gemstone, lapis lazuli was used as one of the first eye shadows.

Expensive pigment Powdered lapis lazuli was once used to make ultramarine, one of the most expensive pigments.

PROFILE

Cubic

5½–6

2.1–2.3

Poor to distinct

Uneven to conchoidal

White to light blue

Vitreous to greasy

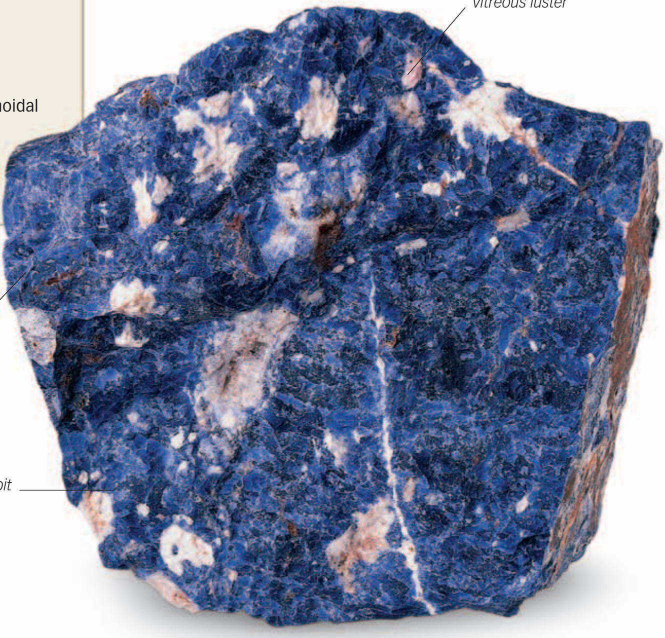

Massive sodalite This sodalite specimen shows intense blue color, which can sometimes lead to the mineral being mistaken for lapis lazuli.

VARIANTS

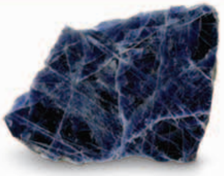

Polished sodalite A specimen that has been polished to bring out its color

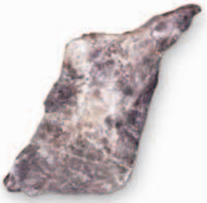

Indian sodalite A specimen of light blue sodalite found in India

$Na_4Al_3Si_3O_{12}Cl$

SODALITE

Named in 1811 after its high sodium content, sodalite is sodium aluminum silicate chloride. Specimens can be blue, gray, pink, colorless, or other pale shades. They sometimes fluoresce bright orange under ultraviolet light. Sodalite nearly always forms massive aggregates or disseminated grains. Crystals are relatively rare; when found, they are dodecahedral or octahedral.

Sodalite occurs in igneous rocks and associated pegmatites (p.260). It is sometimes found in contact metamorphosed limestones (p.319) and dolomites (p.320) and in rocks ejected from volcanoes. Rare crystals are found on the Mount Vesuvius volcano in Italy. Uncommon transparent specimens from Mont St.-Hilaire, Canada, are faceted for collectors.

Sodalite beads This unusual modern Egyptian necklace has beads made of blue sodalite and red carnelian.

PROFILE

Cubic
Tetragonal

5½–6
2.5
Poor
Conchoidal, brittle
White
Vitreous

Leucite crystals
In this specimen formed at a high temperature, fine psuedotrapezohedral crystals of leucite rest in cavities in a rock matrix.

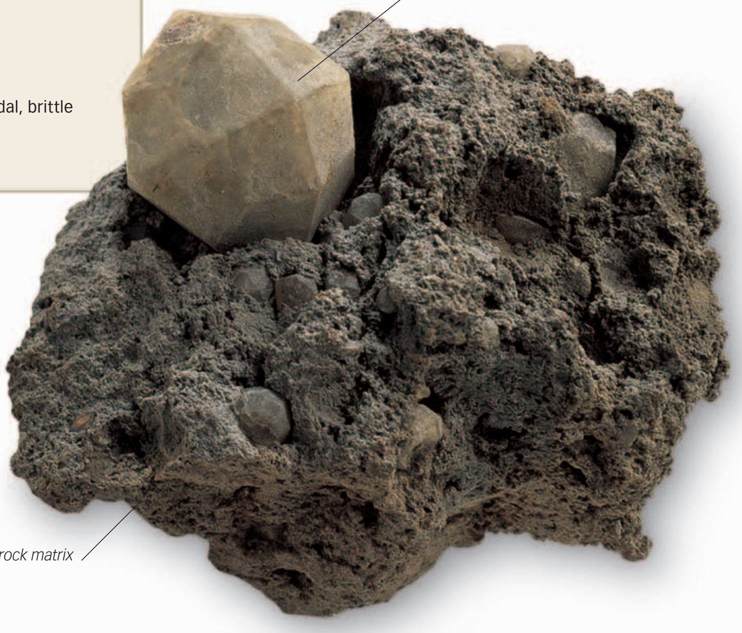

VARIANTS

Single crystal A single yellowish crystal of leucite

Italian leucite A crystal in pseudotrapezohedral form from Casserta, Italy

$KAlSi_2O_6$

LEUCITE

The name leucite comes from the Greek word *leukos*, which means "matt white"—a reference to the mineral's most common color. Specimens can also be colorless or gray. Crystals are common and can be up to 3½ in (9 cm) wide. More often, leucite occurs as massive or granular aggregates or as disseminated grains. It is tetragonal at temperatures below 1,155°F (625°C) and cubic with trapezohedral crystals at higher temperatures. The trapezohedral form is preserved as the mineral cools and develops tetragonal symmetry.

Leucite is found in potassium-rich and silica-poor igneous rocks. It is found with nepheline (p.182), sodalite (p.184), natrolite (p.186), analcime (p.190), and sodium- and potassium-rich feldspars, and occurs worldwide.

Potassium fertilizer
Because of leucite's high potassium content, the mineral is used as a fertilizer in some countries.

PROFILE

Orthorhombic
5–5½
2.3
Perfect
Uneven, brittle
White
Vitreous to pearly

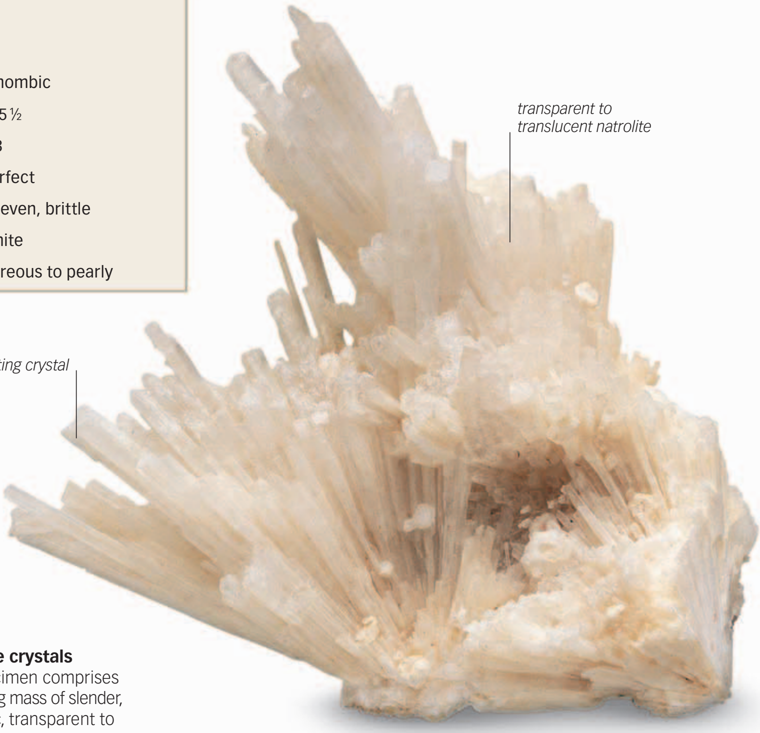

Natrolite crystals
This specimen comprises a radiating mass of slender, prismatic, transparent to translucent natrolite crystals.

VARIANT

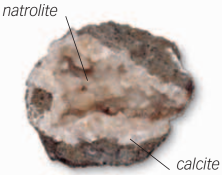

Natrolite and calcite
A specimen including white calcite and light orange natrolite

$Na_2Al_2Si_3O_{10}\cdot 2H_2O$

NATROLITE

A hydrated sodium aluminosilicate, natrolite takes its name from the Greek word *natrium*, which means “soda”—a reference to the sodium content of this mineral. Natrolite can be pale pink, colorless, white, red, gray, yellow, or green. Some specimens fluoresce orange to yellow under ultraviolet light. Natrolite crystals are generally long and slender, with vertical striations and a square cross section. They may appear tetragonal and can grow up to 3 ft (1 m) in length. Natrolite is also found as radiating masses of needlelike crystals and as granular or compact masses. This mineral produces an electric charge in response to both pressure and temperature changes.

Natrolite is found in cavities or fissures in basaltic rocks (p.273), volcanic ash deposits, and veins in granite (pp.258–59), gneiss (p.288), and other rocks. It also occurs in altered syenites (p.262), aplites, and dolerites (p.268). Specimens are often associated with quartz (p.168), heulandite (p.187), apophyllite (p.204), and other zeolites.

PROFILE

Monoclinic

3½–4

2.2

Perfect

Uneven, brittle

Colorless

Vitreous to pearly

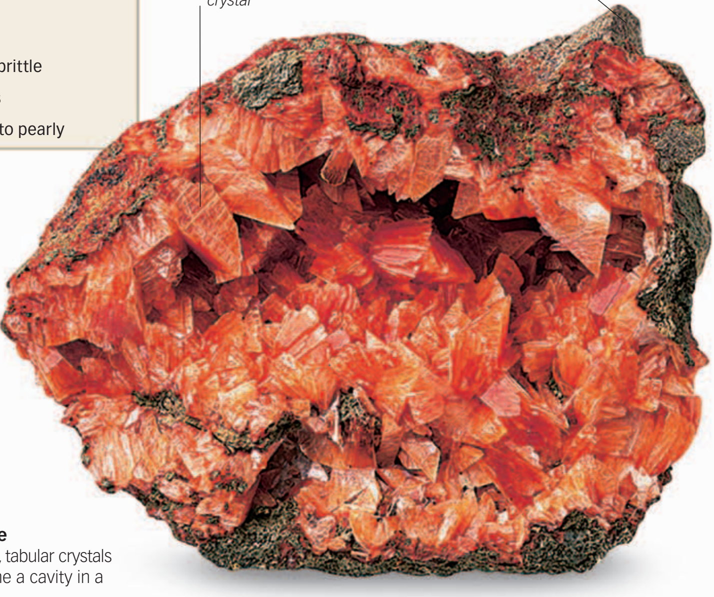

Red heulandite
In this specimen, tabular crystals of heulandite line a cavity in a basalt matrix.

VARIANT

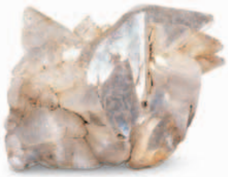

Colorless crystals Typical, colorless, coffin-shaped heulandite crystals

$CaAl_2Si_7O_{18}\cdot 6H_2O$

HEULANDITE

The name heulandite is used to refer to a series of five zeolite minerals, all of which look the same but vary in composition. The group was named in 1822 after the British collector and mineral dealer J.H. Heuland. Heulandite is usually colorless or white but can also be red, gray, yellow, pink, green, or brown. When found, crystals are elongated, tabular, and widest at the center, creating a coffin shape. Occasionally, trapezohedral crystals are found. Heulandite specimens can also be granular or massive.

Heulandite forms at low temperatures (up to 400°F/200°C) in a wide range of environments: with other zeolites filling cavities in granites (pp.258–59), pegmatites (p.260), and basalts (p.273); in metamorphic rocks; and in weathered andesites (p.275) and diabases.

Oil refining
Heulandite and other zeolites are used to filter out unwanted molecules during oil refining.

PROFILE

Monoclinic

5

2.3

Perfect

Uneven, brittle

White

Vitreous to silky

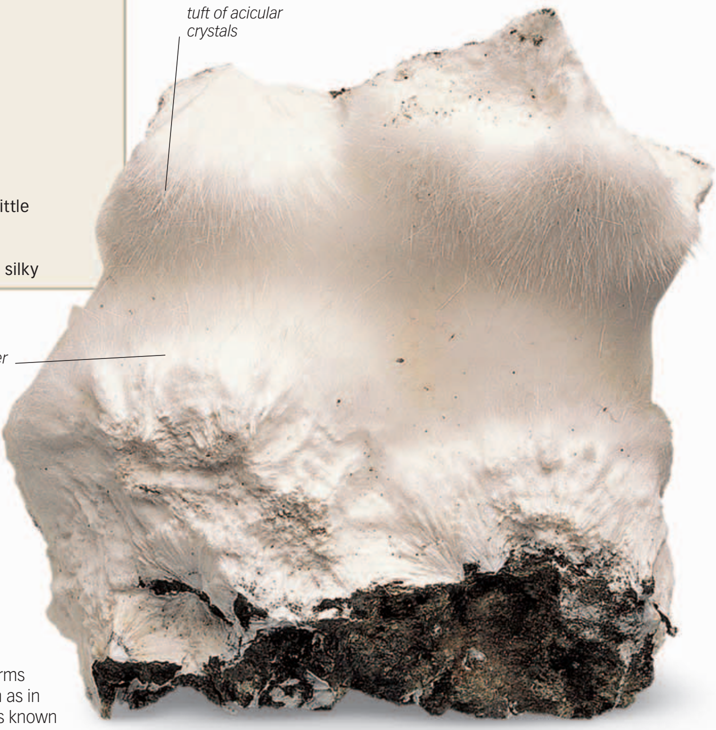

Cotton stone When mesolite forms hairlike tufts, such as in this specimen, it is known as cotton stone.

VARIANTS

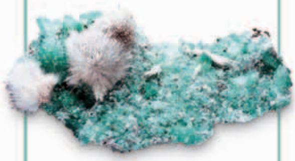

White mesolite Needles of mesolite on green apophyllite

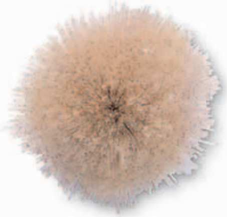

Acicular mesolite A radiating mass of needlelike crystals

$Na_2Ca_2(Al_6Si_9)O_{30}\cdot 8H_2O$

MESOLITE

First described in 1816, mesolite takes its name from two Greek words: *mesos*, which means “middle,” and *lithos*, which means “stone”—a reference to the fact that this mineral is chemically intermediate in composition between scolecite and natrolite (p.186). Mesolite is structurally identical and similar in appearance to scolecite and natrolite, which makes it difficult to identify in hand specimens. Specimens can be white, pink, red, yellowish, green, or pale colored. It occurs as long, slender needles, radiating masses, prisms, and, less commonly, compact masses or fibrous stalactites.

Mesolite is found in cavities in basalts (p.273) and andesites (p.275), where delicate, glassy prisms can occur with stilbite, heulandite (p.187), and green apophyllite (p.204). It is also found in hydrothermal veins. Exceptional specimens occur in Ahmadnagar and Poona, India; Neubauerberg, the Czech Republic; Naalsoy in the Faroe Islands; Victoria Land, Antarctica; and in the states of Washington, Oregon, and Colorado.

Pseudocubic chabazite This group of pseudocubic chabazite crystals is from the Bay of Fundy in Nova Scotia, Canada.

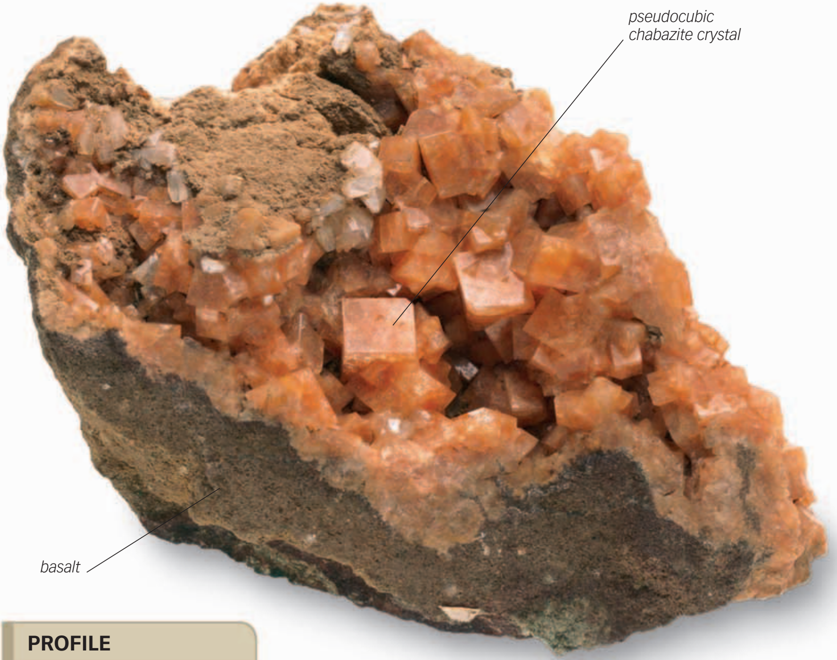

$(Na,Ca_{0.5},K)_4(Al_4Si_8O_{24})\cdot 12H_2O$

CHABAZITE

Common zeolites, the chabazite group consists of three distinct minerals that look alike: chabazite-Ca, chabazite-K, and chabazite-Na. The name is derived from the Greek *chabazios* or *chalazios*, both of which mean "hailstone." Specimens are colorless, white, cream, pink, red, orange, yellow, or brown. Chabazite crystals occur as distorted cubes or pseudorhombohedrons consisting of multiple twins. They may also be prismatic. Twinning is common in all forms of chabazite.

Chabazite is found in cavities in pegmatites (p.260), basalt (p.273), andesite (p.275), volcanic ash deposits, and granitic (pp.258–59) and metamorphic (pp.288–303) rocks. It is widespread, with crystals that are 1–2 in (2.5–5 cm) long occurring in several locations. Chabazite and some other zeolites have an open crystal structure that is sievelike and permits small molecules to pass through, while preventing the passage of larger molecules. This structure, for example, helps filter methane from gases emitted by decaying organic waste matter.

PROFILE

Triclinic (Pseudohexagonal)

4–5

2.0–2.2

Indistinct

Uneven, brittle

White

Vitreous

VARIANT

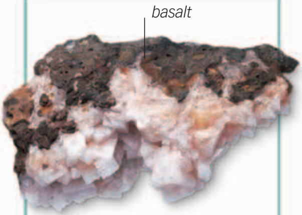

Chabazite in basalt A group of white chabazite crystals in a hollow in basalt

PROFILE

Cubic

5–5½

2.3

None

Subconchoidal, brittle

White

Vitreous

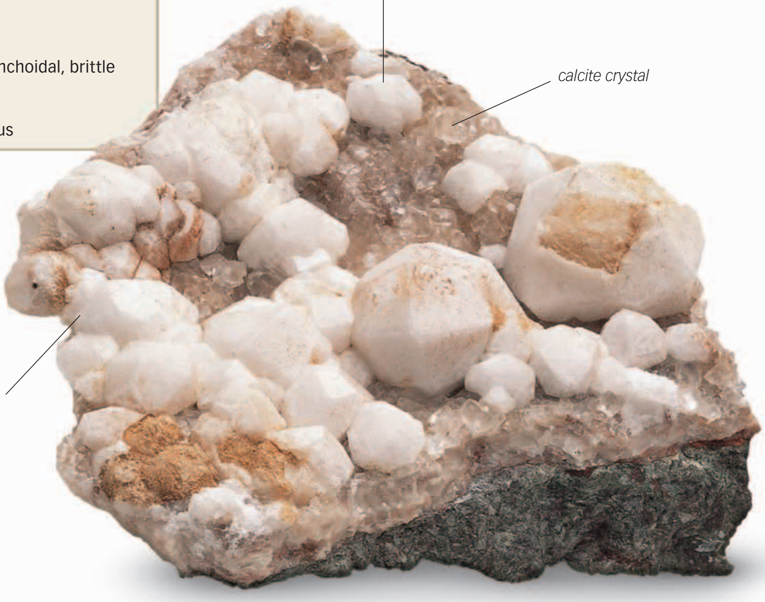

Analcime trapezohedrons This group of superbly crystallized analcime trapezohedrons from the Dean Quarry, Cornwall, England, rests on a bed of calcite crystals.

VARIANT

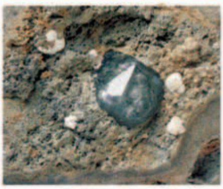

Colorless analcime A single crystal of colorless analcime on a rock matrix

$Na(AlSi_2)O_6\cdot H_2O$

ANALCIME

Formerly grouped with the feldspathoids, analcime is now classified as a zeolite. A sodium aluminum silicate, analcime is named after the Greek word *analkimos*, which means "weak"—a reference to the weak electrical charge that this mineral produces when it is heated or rubbed. Specimens are usually colorless or white but can also be yellow, brown, pink, red, or orange. Most analcime crystals are trapezohedral. Variations in the ratio and order of the sodium–aluminum portion in analcime can lead to structural variations and variation in crystal system.

Analcime occurs in seams and cavities in granite (pp.258–59), basalt (p.273), gneiss (p.288), and diabase, associated with calcite (p.114), prehnite (p.205), and other zeolites. It also occurs in extensive beds formed by precipitation from alkaline lakes.

Silica dessicator Made from analcime, silica gel, such as in this dessicator, rapidly absorbs moisture and has many drying uses.

PROFILE

Monoclinic or triclinic

3½–5½

2.5–2.6

Perfect but not visible

Conchoidal to splintery

White

Subvitreous to greasy, resinous, earthy, dull

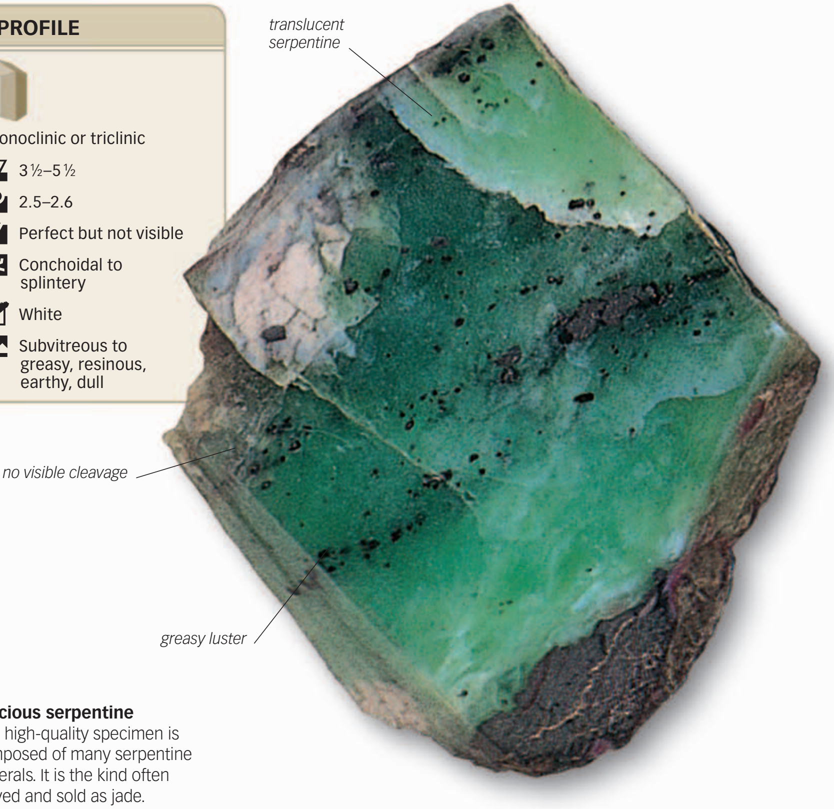

Precious serpentine This high-quality specimen is composed of many serpentine minerals. It is the kind often carved and sold as jade.

VARIANTS

Lizardite A specimen of this fine-grained serpentine mineral from Cornwall, UK

Antigorite A specimen of this serpentine mineral with characteristic, corrugated plates

$(Mg,Fe,Ni)_3Si_2O_5(OH)_4$

SERPENTINE

Resembling snakeskin in appearance, serpentine is a group of at least 16 white, yellowish, green, or gray-green magnesium silicate minerals. Although they usually form mixtures, individual members of the group can sometimes be distinguished. Four common serpentine minerals include chrysotile (p.192), antigorite, lizardite, and amesite, which occur in platy or pseudohexagonal, columnar crystals. Although their chemistry is complex, these minerals look similar.

Serpentines are secondary minerals derived from the chemical alteration of olivine (p.232), the pyroxenes, and the amphiboles. It is found in areas where highly altered, deep-seated, silica-poor rocks are exposed, such as along the crests and axes of great folds, in island arcs, and in Alpine mountain chains.

Williamsite cabochon A variety of serpentine, williamsite is an ornamental stone that is sometimes cut as an inexpensive gem.

PROFILE

Monoclinic or orthorhombic

2–3

2.6

Perfect

None

White

Subresinous to greasy

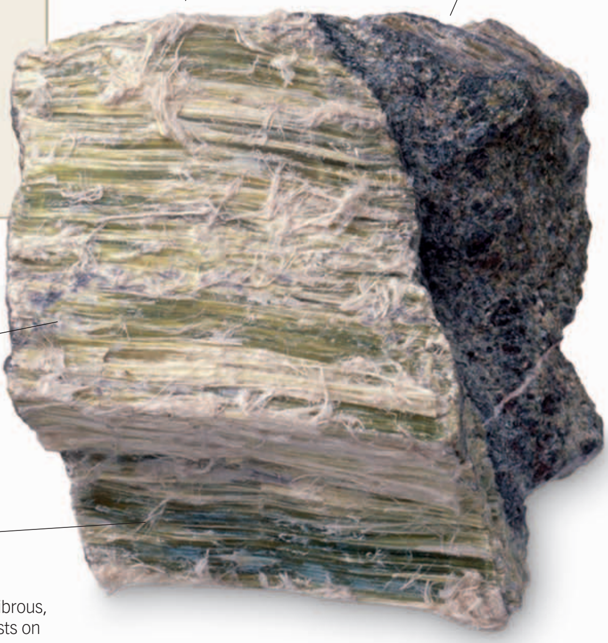

Chrysotile fibers
In this specimen, a mass of fibrous, flexible chrysotile crystals rests on a rock matrix.

VARIANT

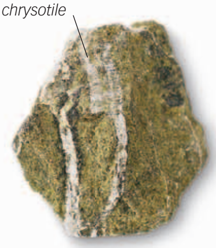

Asbestos mineral A specimen of fibrous chrysotile

$Mg_3Si_2O_5(OH)_4$

CHRYSOTILE

The fibrous serpentine mineral chrysotile is the most important asbestos mineral. Also known as white asbestos, it accounts for about 95 percent of all asbestos in commercial use. Chrysotile fibers are tubes in which the structural layers of the mineral are rolled in the form of a spiral. Individual chrysotile fibers are white and silky, while aggregate fibers in veins are green or yellowish. The fibers are generally oriented across the vein and less than ½ in (1.3 cm) in length, but they can be longer. The mineral sometimes appears golden, and its name is derived from the Greek for "hair of gold."

Chrysotile can take three different forms: clinochrysotile, orthochrysotile, and parachrysotile. These are chemically identical, but orthochrysotile and parachrysotile have orthorhombic rather than monoclinic crystals. These forms are indistinguishable in hand specimens, and clinochrysotile and orthochrysotile may occur within the same fiber. Specimens occur as veins in altered peridotite (p.266) with other serpentine minerals.

PROFILE

Triclinic or monoclinic

1

2.8

Perfect

Uneven to subconchoidal

White

Pearly to greasy

Foliated talc
This specimen of green, foliated talc exhibits its pearly luster and micalike cleavage.

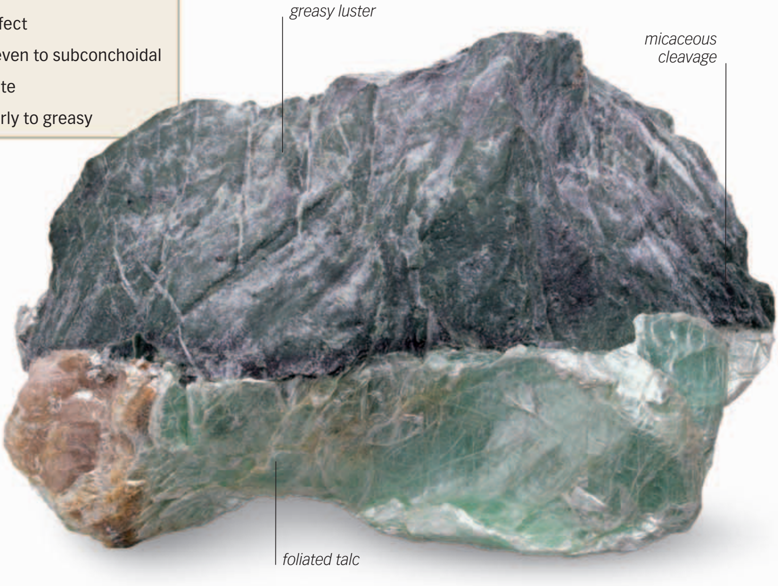

VARIANT

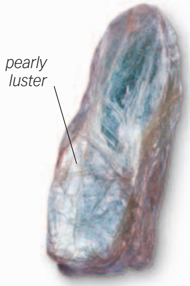

Pearly talc A pearly, toothlike piece of talc

$Mg_3Si_4O_{10}(OH)_2$

TALC

Easily distinguishable by its extreme softness, talc is white, colorless, pale to dark green, or yellowish to brown. Crystals are rare; talc is most commonly found in foliated, fibrous, or massive aggregates. It is often found mixed with other minerals, such as serpentine (p.191) and calcite (p.114). Dense, high-purity talc is called steatite.

Talc is a metamorphic mineral found in veins and magnesium-rich rocks. It is often associated with serpentine, tremolite (p.219), and forsterite (p.232) and occurs as an alteration product of silica-poor igneous rocks. Talc is widespread and is found in most areas of the world where low-grade metamorphism occurs. The name soapstone is given to compact masses of talc and other minerals due to their soapy or greasy feel.

WITH THE
Beatles
TALC

Talcum powder
Talc is the principal mineral used to make talcum powder. It acts as an astringent on the skin.

PROFILE

Triclinic or monoclinic

1–2

2.7–2.9

Perfect

Uneven

White

Pearly to dull

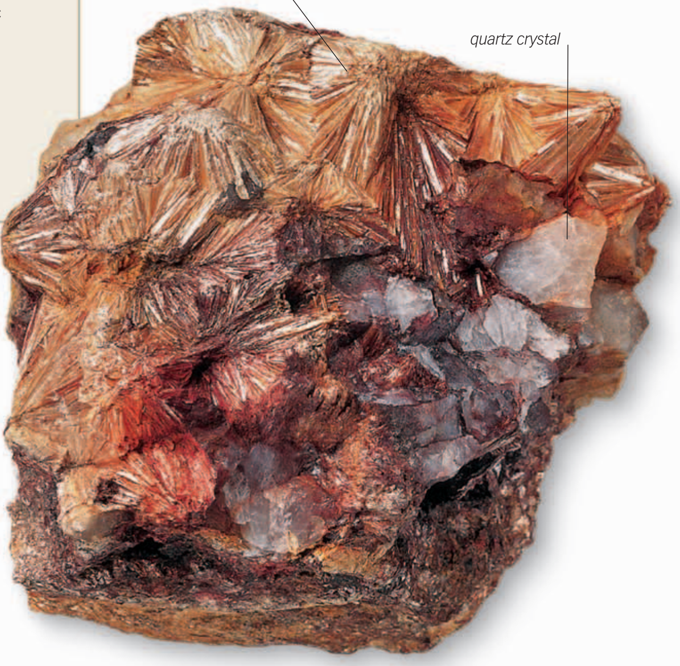

Pyrophyllite stars This aggregate of pyrophyllite displays radiating, starlike groups of laths with associated quartz.

VARIANT

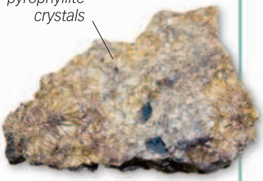

Pyrophyllite on rock Groups of pyrophyllite crystals on a rock matrix

$Al_2Si_4O_{10}(OH)_2$

PYROPHYLLITE

An aluminum silicate hydroxide, pyrophyllite takes its name from the Greek words *pyr* and *phyllon*, which respectively mean “fire” and “leaf”—a reference to the mineral’s tendency to exfoliate when heated. Pyrophyllite can be colorless, white, cream, brownish green, pale blue, or gray. It is usually found in granular masses of flattened lamellae. Pyrophyllite rarely forms distinct crystals, although it is sometimes found in coarse laths and radiating aggregates. Specimens are often so fine-grained that the mineral appears textureless.

Pyrophyllite forms by the metamorphism of aluminum-rich sedimentary rocks, such as bauxite (p.101). It is a good insulator and is used in heat-resistant applications, such as in making fire bricks.

High gloss Bright, reflective flakes of powdered pyrophyllite are added to lipstick to give it a high sheen.

Colorful fuchsite
Fuchsite is a minor variety of white muscovite. Specimens such as this one are colored by traces of chromium.

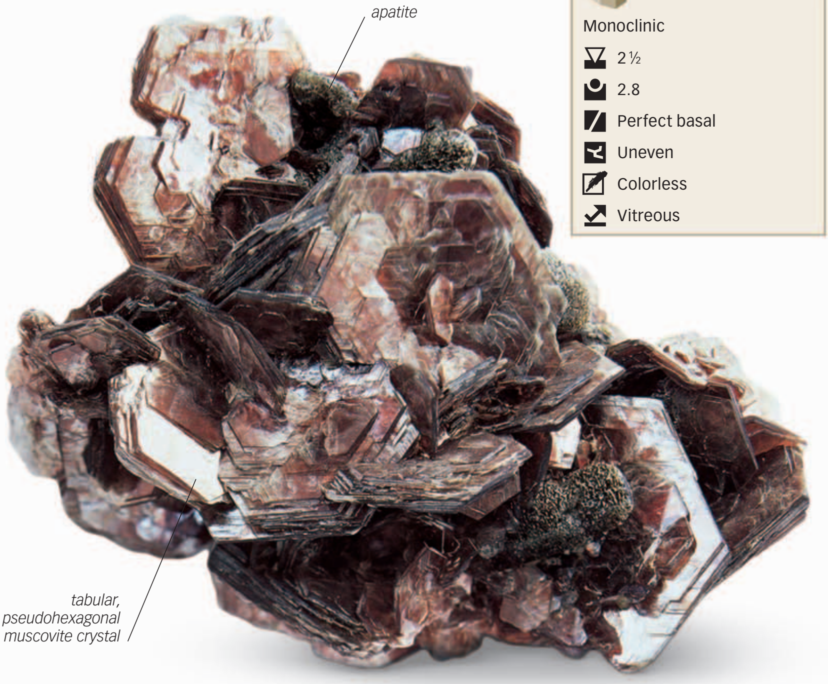

PROFILE

Monoclinic
2½
2.8
Perfect basal
Uneven
Colorless
Vitreous

VARIANTS

Tabular crystals Silver-brown crystals of tabular muscovite

Platy muscovite Crystals of muscovite in a rock matrix

Green fuchsite Bright green fuchsite in a rock matrix

$KAl_2(Si_3Al)O_{10}(OH,F)_2$

MUSCOVITE

Also called common mica, potash mica, or isinglass, muscovite is the most common member of the mica group. Specimens are usually colorless or silvery white but can also be brown, light gray, pale green, or rose red. Muscovite typically occurs as tabular crystals with a hexagonal or pseudohexagonal outline. Crystals can be up to 9¾ ft (3 m) in diameter. Muscovite can also form thin, flat sheets and fine-grained aggregates. Fine-grained muscovite is called sericite or white mica, while bright green specimens rich in chromium are called fuchsite.

A common rock-forming mineral, muscovite occurs in metamorphic rocks, such as gneisses (p.288) and schists (pp.291–92), and in granites (pp.258–59), veins, and pegmatites (p.260). It is also found in some fine-grained sediments. Muscovite has considerable commercial importance. Its low iron content makes it a good electrical and thermal insulator. In Russia, thin, transparent sheets of muscovite, called muscovy glass, were used as window panes.

PROFILE

Monoclinic

2

2.4–2.9

Perfect basal

Uneven

Green

Dull to earthy

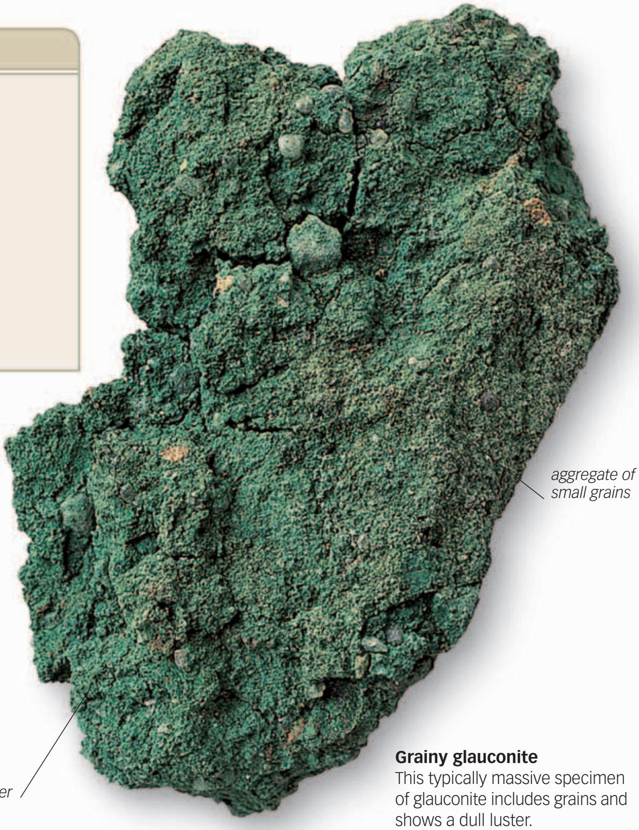

Grainy glauconite
This typically massive specimen of glauconite includes grains and shows a dull luster.

VARIANTS

Glauconite sandstone
Sandstone with a high percentage of glauconite

Nodular glauconite Light green nodules of glauconite

$(K,Na)\ (Mg,Al,Fe)_2(Si,Al)_4O_{10}(OH)_2$

GLAUCONITE

A member of the mica group, glauconite was named in 1828 after the Greek word *glaukos*, which means “blue-green”—a reference to the mineral’s usual color. Specimens can also be olive green to black-green. The mineral usually occurs as rounded aggregates or pellets of fine-grained, scaly particles. It weathers readily and easily crumbles to a fine powder.

A widespread silicate, glauconite forms in shallow marine environments, where it is used as a diagnostic mineral to identify continental-shelf deposits with slow rates of accumulation. The sedimentary rock greensand (p.309) is so called because of the green color imparted by glauconite pellets, which in turn, may incorporate other minerals. Glauconite can also be found in impure limestone (p.319), chalk (p.321), and sand and clay formations. The mineral has long been used as a pigment in artists’ oil paint, especially in the paintings of Russian icons. It has also been used in wall paintings dating back to Roman Gaul.

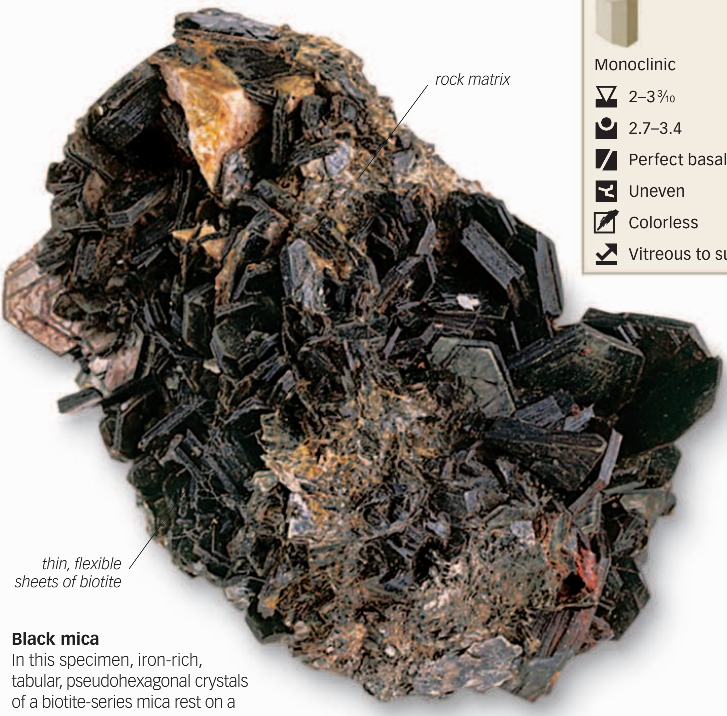

Black mica
In this specimen, iron-rich, tabular, pseudohexagonal crystals of a biotite-series mica rest on a rock matrix.

PROFILE

Monoclinic

$2–3\frac{3}{10}$

2.7–3.4

Perfect basal

Uneven

Colorless

Vitreous to submetallic

VARIANT

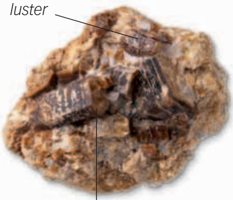

Phlogopite Crystals of the magnesium end-member of the biotite solid-solution series

$KFe_3,Mg_3(AlSi_3O_{10})(OH,F)_2$

BIOTITE

Once considered a mineral in its own right, biotite is now recognized as a solid-solution series with the mineral annite as the iron end-member and phlogopite as the magnesium end-member. It was named in honor of the French physicist Jean-Baptiste Biot in 1847. Micas of the biotite series usually form large, tabular to short, prismatic crystals that are often pseudohexagonal in cross section. They also occur in thin layers or as scaly aggregates or disseminated grains. Specimens are black when iron-rich, and brown, pale yellow to tan, or bronze with increasing magnesium content. They readily cleave into thin, flexible sheets.

Biotite-series micas are widespread. They are a key constituent of many igneous and metamorphic rocks, including granites (pp.258–59), nepheline syenites (p.262), gneisses (p.288), and schists (pp.291–92). They are also found in potassium-rich hydrothermal deposits and some clastic sedimentary rocks. Biotite is used extensively to date rocks.

PROFILE

Monoclinic
2½–3½
3.0
Perfect basal
Uneven
Colorless
Vitreous to pearly

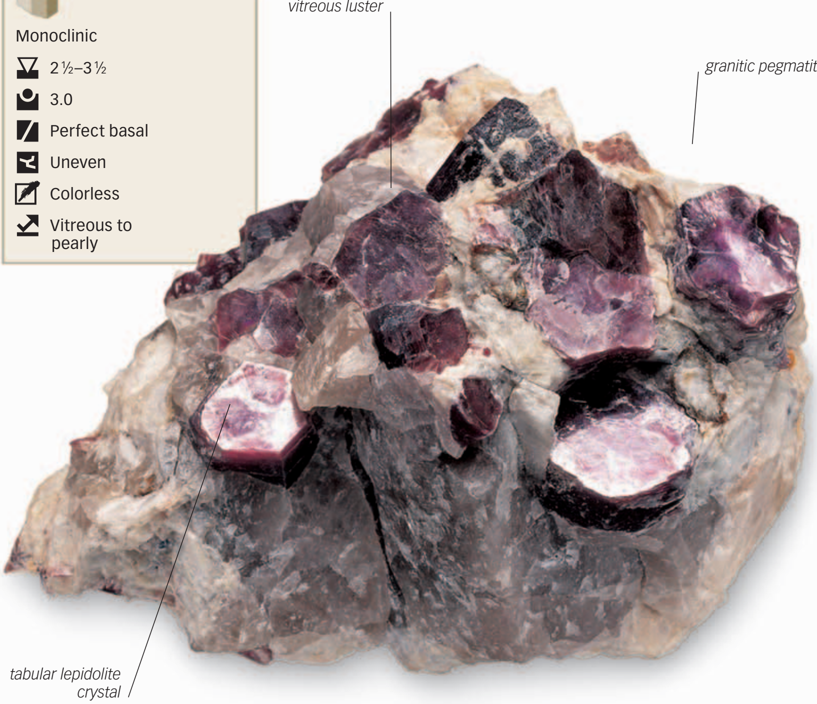

Distinctive color
Numerous violet pseudohexagonal lepidolite crystals protrude from this pegmatite specimen.

VARIANT

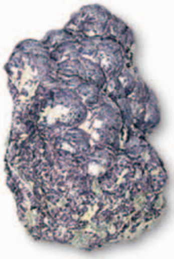

Botryoidal lepidolite
Lepidolite in botryoidal habit

$K(Li,Al_3)(AlSi_3)O_{10}(OH,F)_2$

LEPIDOLITE

A light mica, lepidolite is Earth's most common lithium-bearing mineral. Its name is derived from two Greek words: *lepidos*, which means "scale," and *lithos*, which means "stone." Although typically pale lilac, specimens can also be colorless, violet, pale yellow, or gray. Lepidolite crystals may appear pseudohexagonal. The mineral is also found as botryoidal or kidneylike masses and fine- to coarse-grained, interlocking plates. Its perfect cleavage yields thin, flexible sheets.

Lepidolite occurs in granitic pegmatites (p.260), where it is associated with other lithium minerals, such as beryl (p.225) and topaz (p.234). The mineral is economically important as a major source of lithium, which is used to make glass and enamels. It is also a major source of the rare alkali metals rubidium and cesium.

Lithium battery
Extracted from lepidolite, the metal lithium has many industrial uses, such as in lithium batteries.

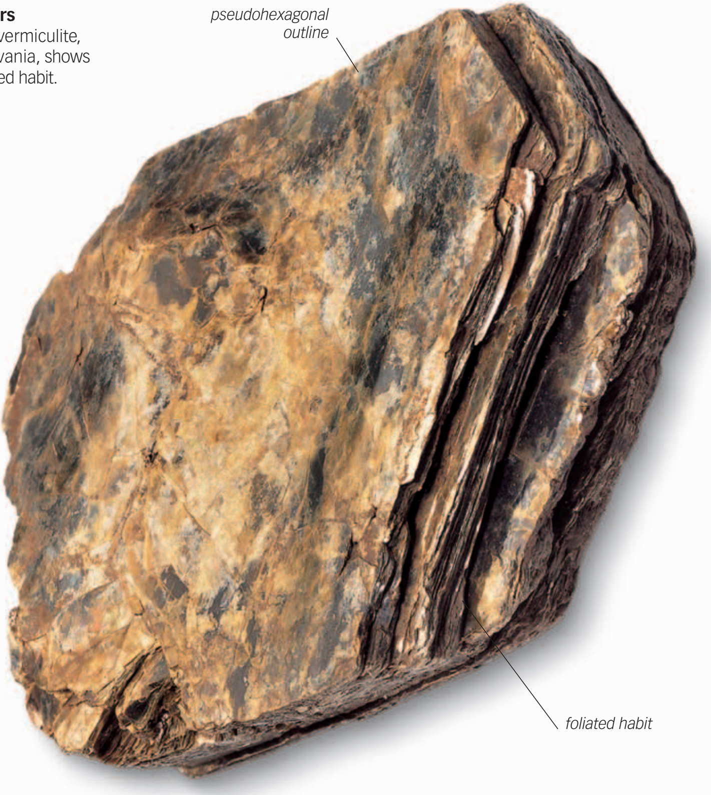

Vermiculite layers
This specimen of vermiculite, mined in Pennsylvania, shows the mineral's foliated habit.

PROFILE

Monoclinic
1–2
2.6
Perfect
Uneven
White
Oily to earthy

$(Mg,Fe,Al)_3(Al,Si)_4O_{10}(OH)\cdot 4H_2O$

VERMICULITE

The name vermiculite is applied to a group of mica minerals in which various chemical substitutions occur in the molecular structure. Vermiculite may be completely interlayered with other micas and claylike minerals. Specimens are green, golden yellow, or brown in color. Vermiculite usually forms tabular, pseudohexagonal crystals or platy aggregates.

Vermiculite occurs as large pseudomorphs replacing biotite (p.197), as small particles in soils and ancient sediments, and at the interface between feldspar-rich and iron- and magnesium-rich igneous rocks. It also forms by hydrothermal alteration of iron-bearing micas. When heated to nearly 572°F (300°C), vermiculite can expand quickly and strongly to 20 times its original thickness.

Potting soil
Expanded vermiculite is a good growing medium for new plants; it retains water and offers good aeration.

PROFILE

Monoclinic
1–2
2.1
Perfect
Uneven
White to buff
Earthy

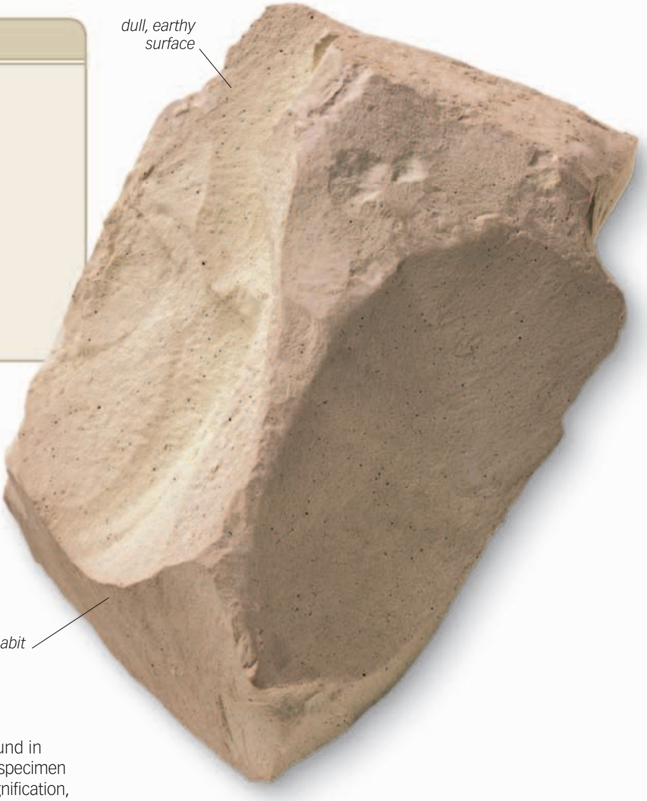

Massive bentonite Bentonite is always found in massive form, as in the specimen shown here. Under magnification, tiny, hexagonal plates are visible.

VARIANTS

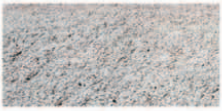

Desert cracking Loss of water from bentonite clays causes shrinkage and cracking

Bentonite sediments Layers of bentonite-rich clay

$(Na,Ca_{0.5})_{0.33}(Al,Mg)_2Si_4O_{10}(OH)_2 \cdot nH_2O$

BENTONITE

This group of minerals are all kinds of clay that expand as they absorb water and shrink as they dry. In regions underlain by bentonites, this property causes immense problems with building foundations. There are three bentonite minerals, each named after the respective dominant element: potassium bentonite, sodium bentonite, and calcium bentonite. The minerals are generally yellow, white, or gray in color. They occur as microscopic crystals and are earthy and frequently stained.

Although the term bentonite has been used for clay beds of uncertain origin, this mineral group generally forms from volcanic ash that has weathered in the presence of water. Important industrial minerals, bentonites are used as sealants and in oil drilling.

Potting clay The potting clay used to make this bowl contains bentonite, which is also used in bricks and ceramics.

PROFILE

Triclinic

2–2½

2.6

Perfect

Unobservable

White

Earthy

Powdery kaolinite In this specimen, powdery kaolinite coats a piece of granite.

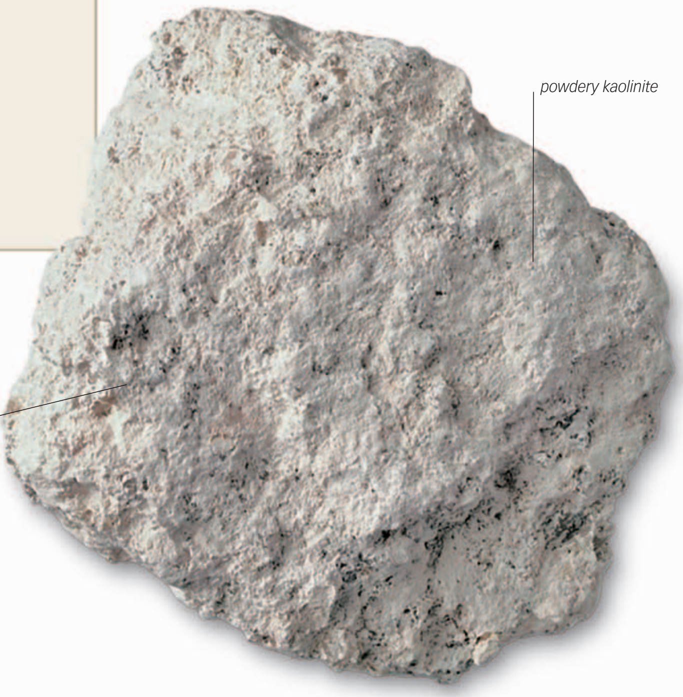

VARIANTS

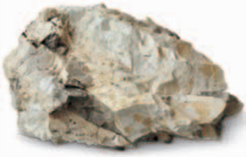

Blocky kaolinite Blocky, typically white kaolinite

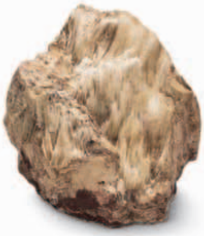

Iron staining A specimen of kaolinite mixed with iron oxides, which give it an orange color

$Al_2Si_2O_5(OH)_4$

KAOLINITE

Clay minerals are far removed in their outward appearance from more attractive and glamorous minerals, such as gold and diamond. Yet, by providing the raw material for brick, pottery, and tiles, they have played a vital part in the progress of human civilization. Important among these minerals is kaolinite. Kaolinite forms white, microscopic, pseudohexagonal plates in compact or granular masses and in micalike piles. Three other minerals—dickite, nacrite, and halloysite—are chemically identical to kaolinite but crystallize in the monoclinic system. All four have been found together and are often visually indistinguishable.

Kaolinite is a natural product of the alteration of mica, plagioclase, and sodium–potassium feldspars under the influence of water, dissolved carbon dioxide, and organic acids. It is used in agriculture; as a filler in food, such as chocolate; mixed with pectin as an antidiarrheal; as a paint extender; as a strengthener in rubber; and as a dusting agent in foundry operations.

PROFILE

Monoclinic
1–2
2.6–2.9
Perfect
None
White
Dull

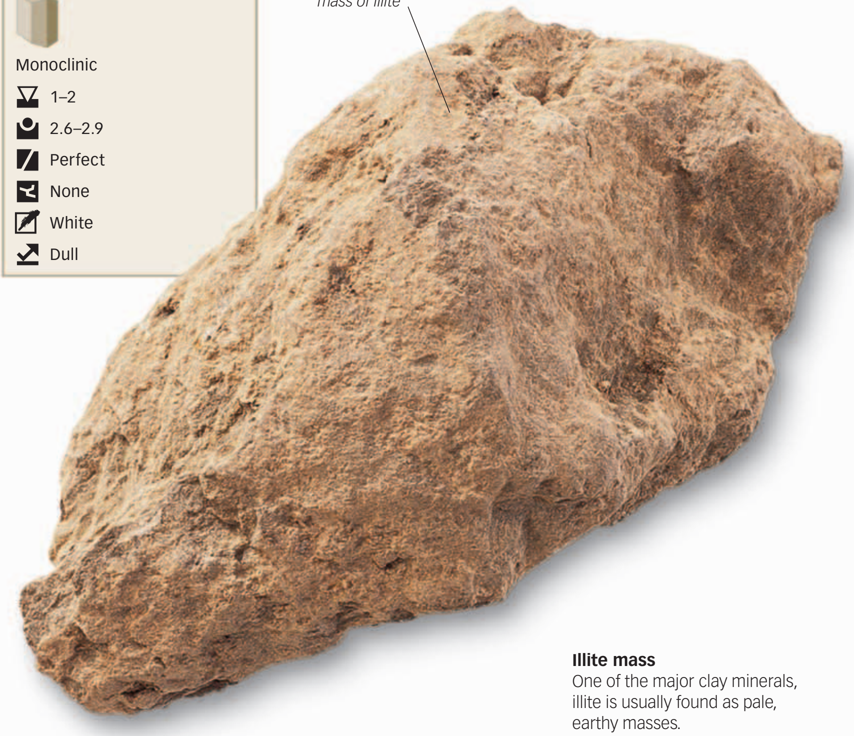

Illite mass
One of the major clay minerals, illite is usually found as pale, earthy masses.

VARIANT

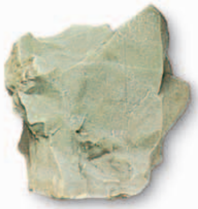

Solid illite Solid masses of illite are occasionally found

$K_{0.65}Al_2Al_{0.65}Si_{3.35}O_{10}(OH)_2$

ILLITE

Once regarded as a clay mineral, illite is now classified as a group of mica minerals that bear many structural similarities to the white mica muscovite (p.195). Illite takes its name from its type location in Illinois. It is white, but impurities may tint it gray and other pale colors. It occurs as fine-grained aggregates. Individual hexagonal crystals can only be seen using an electron microscope. Because of its minute crystals, illite can only be positively identified by x-ray diffraction. The degree of crystallization of illite has been used as an indicator of metamorphic grade in clay-bearing metamorphic rocks.

Illite is found in sedimentary rocks and soils. It is the most abundant clay mineral in shales (pp.313–14) and clays. It appears to be derived from the weathering of muscovite and feldspar (pp.173–81).

Mud bricks
Ancient buildings, such as the Funerary Temple in Egypt, were often made from clays bearing illite.

Chrysocolla with azurite
In this specimen, chrysocolla can be seen with the carbonate mineral azurite in a rock matrix.

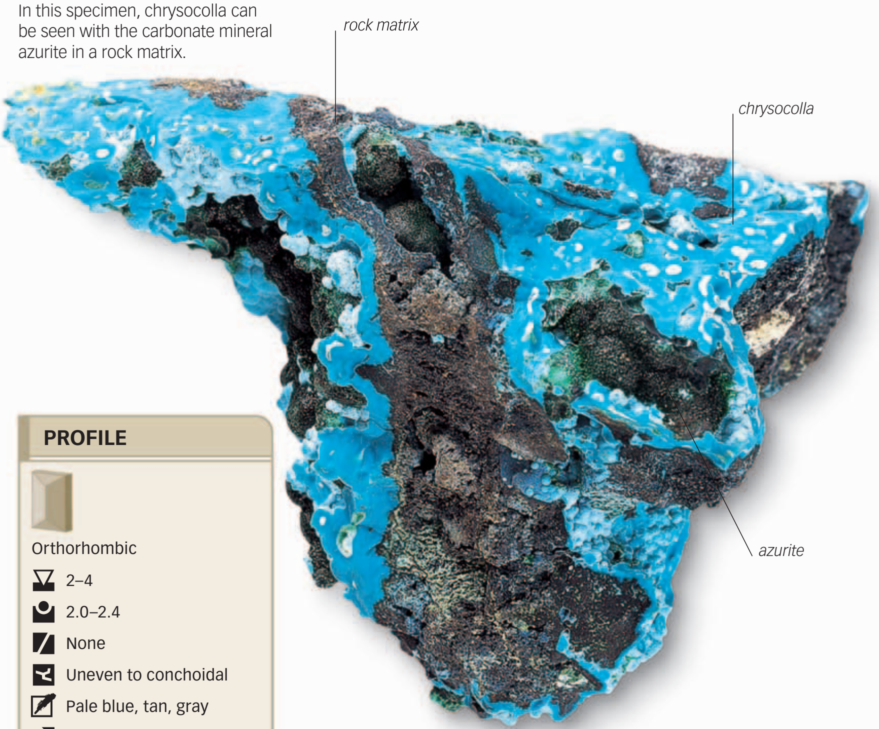

PROFILE

Orthorhombic
2–4
2.0–2.4
None
Uneven to conchoidal
Pale blue, tan, gray
Vitreous to earthy

VARIANTS

Rough chrysocolla
A specimen intergrown with turquoise and malachite

chrysocolla

Cabochon Green chrysocolla within reddish iron oxide

$Cu_2H_2(Si_2O_5)(OH)_4 \cdot nH_2O$

CHRYSOCOLLA

The term chrysocolla was first used by the Greek philosopher Theophrastus in 315 BCE to refer to various materials used in soldering gold. The name is derived from two Greek words: *chrysos*, which means “gold,” and *kolla*, which means “glue.” A copper aluminum silicate, chrysocolla is generally blue-green in color. Specimens are commonly fine grained and massive. Crystals are very rare but when found occur as botryoidal, radiating aggregates.

An occasional ore of copper, chrysocolla is a decomposition product of copper minerals, especially in arid regions. It is frequently intergrown with other minerals, such as quartz (p.168), chalcedony (p.169), and opal (p.172), to yield a gemstone variety. Gemstones can weigh more than 5 lb (2.3 kg).

Chrysocolla bracelet
Rich blue-green chrysocolla, such as the cabochon in this antique bracelet, is highly prized as a gemstone.

Green apophyllite In this specimen, green apophyllite occurs in a basalt matrix with a white zeolite mineral.

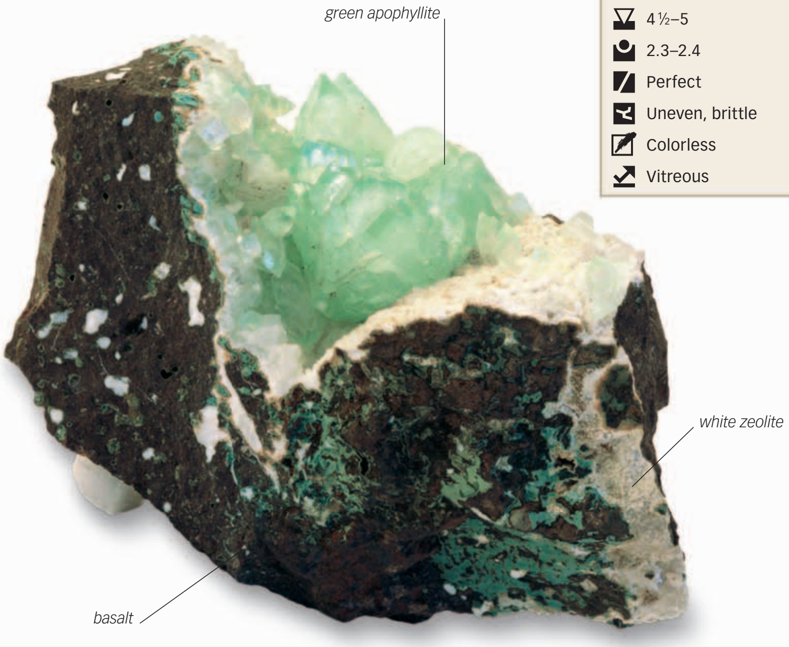

PROFILE

Tetragonal
4½–5
2.3–2.4
Perfect
Uneven, brittle
Colorless
Vitreous

VARIANTS

Pink crystal An unusual single crystal of pink apophyllite

Blocky crystals Green crystals on stilbite

apophyllite
stilbite

Reddish pink crystals Apophyllite with white stilbite

$KCa_4Si_8O_{20}(F,OH)\cdot 8H_2O$ (fluorapophyllite)

APOPHYLLITE

The name apophyllite comes from the Greek words *apo* and *phyllazein*, which mean “to get” and “leaf” respectively—a reference to the way in which the mineral separates into flakes or layers when it is heated. Once considered to be a single mineral, apophyllite is now divided into two distinct species—fluorapophyllite and hydroxyapophyllite. These species form a solid-solution series in which fluorine can predominate over oxygen and hydrogen, and vice versa. Apophyllite specimens are green, pink, colorless, or white. Crystals are transparent or translucent and up to 8 in (20 cm) in length. They occur as square-sided, striated prisms with flat ends and may appear cubic. Apophyllite crystals may also show steep pyramidal terminations.

The mineral frequently occurs with zeolite minerals in basalt (p.273) and less commonly in cavities in granites (pp.258–59). It is also found in metamorphic rocks and in hydrothermal deposits. Colorless and green specimens from India are faceted as collectors’ gems.

PROFILE

Orthorhombic
6–6½
2.9
Distinct basal
Uneven, brittle
White
Vitreous

Botryoidal prehnite A group of radiating crystal masses of prehnite resting on a rock matrix gives a botryoidal form to this specimen.

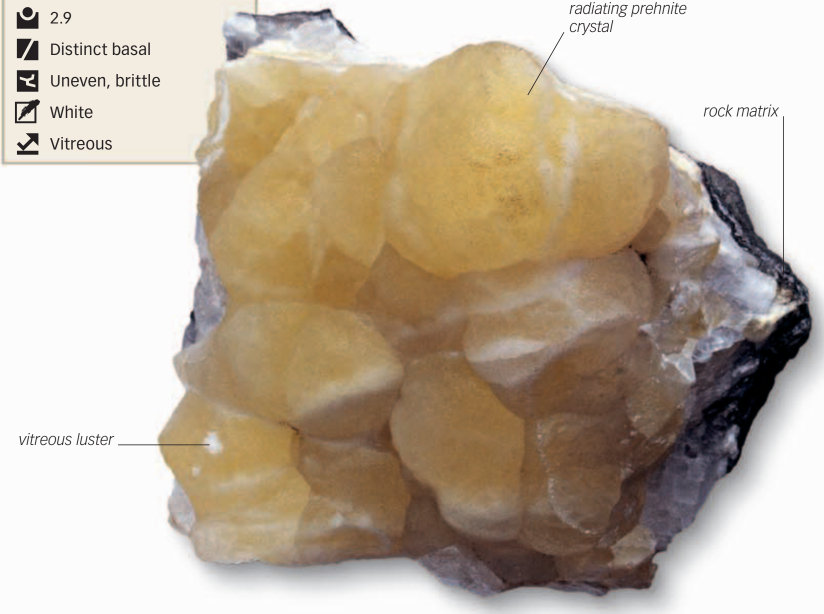

VARIANT

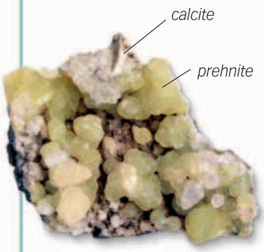

Green prehnite Spherical masses of prehnite occuring with white calcite on a rock matrix

$Ca_2Al_2Si_3O_{10}(OH)_2$

PREHNITE

A calcium aluminum silicate, prehnite was named in 1789 after its discoverer Hendrik von Prehn, a Dutch military officer. Specimens can be pale to mid-green, tan, pale yellow, gray, light blue, or white. Prehnite commonly occurs as globular, spherical, or stalactitic aggregates of fine to coarse crystals. Rare individual crystals are short prismatic to tabular with square cross sections. Many of these have curved faces.

Prehnite is often found lining cavities in volcanic rocks, associated with calcite (p.114) and zeolites (pp.185–90), and in mineral veins in granite (pp.258–59). Crystals up to several inches long come from Canada. Transparent specimens from Australia and Scotland are faceted for gem collectors.

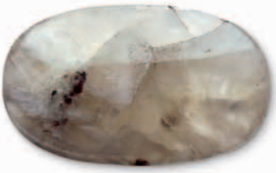

White cabochon Prehnite gems, such as this creamy white cabochon with dark inclusions, are almost too soft to wear.

PROFILE

Orthorhombic

2–2½

2.1–2.3

Good but rarely seen

Uneven

White

Dull to earthy

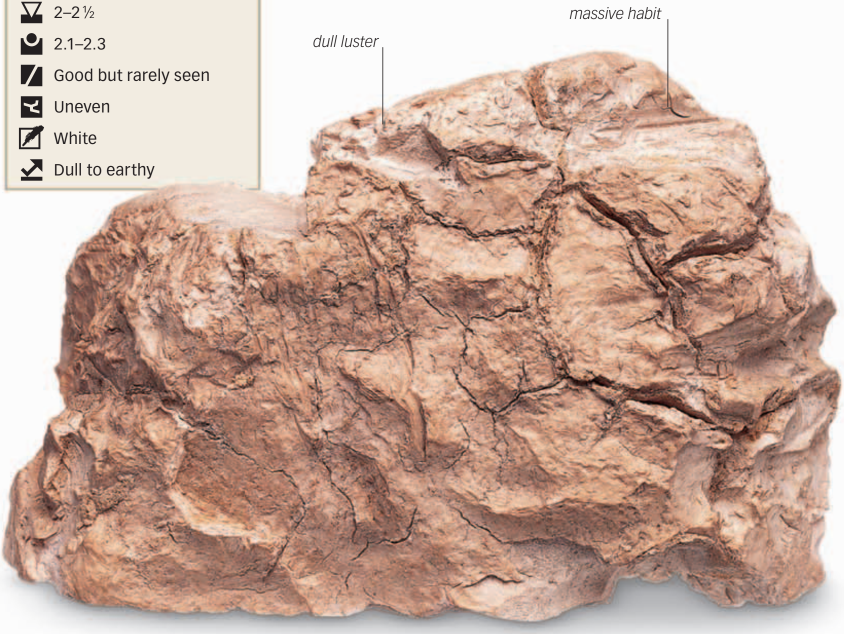

Massive sepiolite This specimen of massive sepiolite shows a characteristic dull, earthy luster.

VARIANT

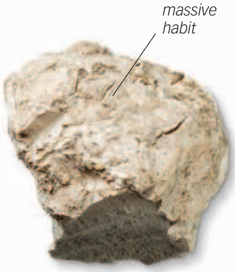

Meerschaum A specimen of massive white sepiolite

$Mg_4Si_6O_{15}(OH)_2.6H_2O$

SEPIOLITE

A compact, claylike, often porous mineral, sepiolite is best known by its popular name *meerschaum*, which is the German word for “sea-foam.” The name sepiolite comes from the mineral’s resemblance to the light and porous bone of cuttlefish from the genus *Sepia*. Sepiolite is usually white or gray and may be tinted yellow, brown, or green. It is usually found in nodular masses of interlocking fibers, which give it a toughness contrary to its mineralogical softness. Sepiolite also occurs in porous aggregates.

Sepiolite is an alteration product of minerals such as magnesite (p.118) and rocks, such as serpentinite (p.298). It is found as irregular nodules in Turkey and elsewhere, and in large sedimentary deposits.

Meerschaum cigar holder Sepiolite is used in carved tobacco pipes and cigar holders, which develop a brown patina when smoked.

Purple-brown pigeonite
This specimen of pigeonite comes from the Kovdor Pit, Kola Peninsula, Russia.

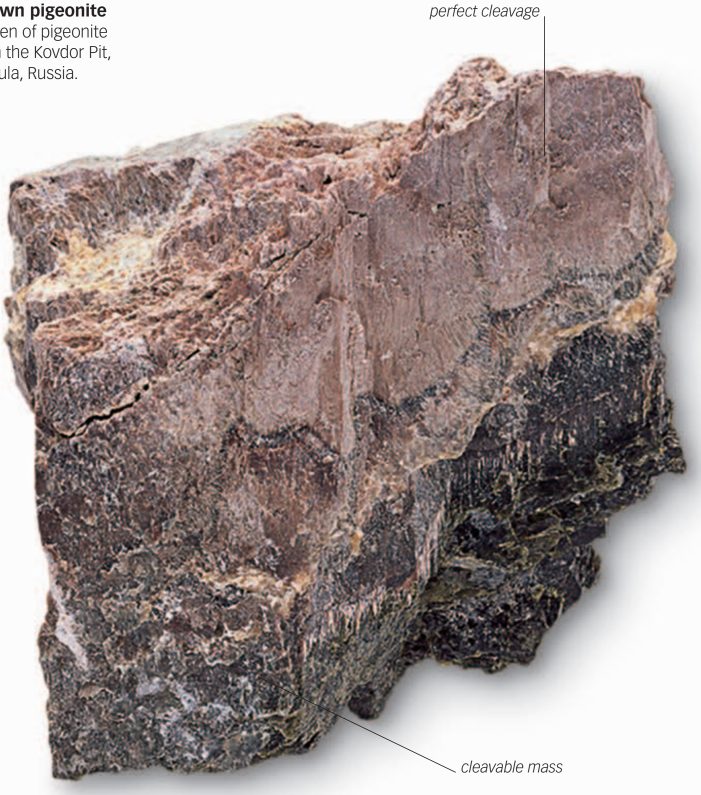

PROFILE

Monoclinic
6
3.2–3.5
Good
Uneven to conchoidal, brittle
White to pale brown
Vitreous

$(Mg,Fe,Ca)_2(Si_2O_6)$

PIGEONITE

A member of the pyroxene group of minerals, pigeonite is named after Pigeon Point, Minnesota, USA—the locality where it was first identified. Specimens are brown, purplish brown, or greenish brown to black in color. Pigeonite is generally found as disseminated grains. Well-formed crystals are relatively rare. An iron-rich variety of pigeonite is sometimes called ferropigeonite.

Pigeonite is an important mineral in lunar rocks and also occurs in meteorites (pp.335–37). It is found in lavas and smaller intrusive rock bodies as the dominant pyroxene and as an important component of dolerites (p.268) and andesites (p.275). The temperature limit of pigeonite formation indicates the crystallization temperature of the magma from which it has originated. Mare—the large, dark, flat areas of the Moon once believed to be seas—are in fact basalts (p.273) containing pigeonite. Notable localities on Earth include Skaergaard, Greenland; Mull, Scotland; Labrador, Canada; Mount Wellington, Tasmania; and Goose Creek, Virginia, and New Jersey, USA.

PROFILE

Orthorhombic

5–6

3.1–3.9

Good to perfect

Uneven

Gray to white

Vitreous

Prismatic crystals This mass of small, prismatic enstatite crystals rests on a rock matrix.

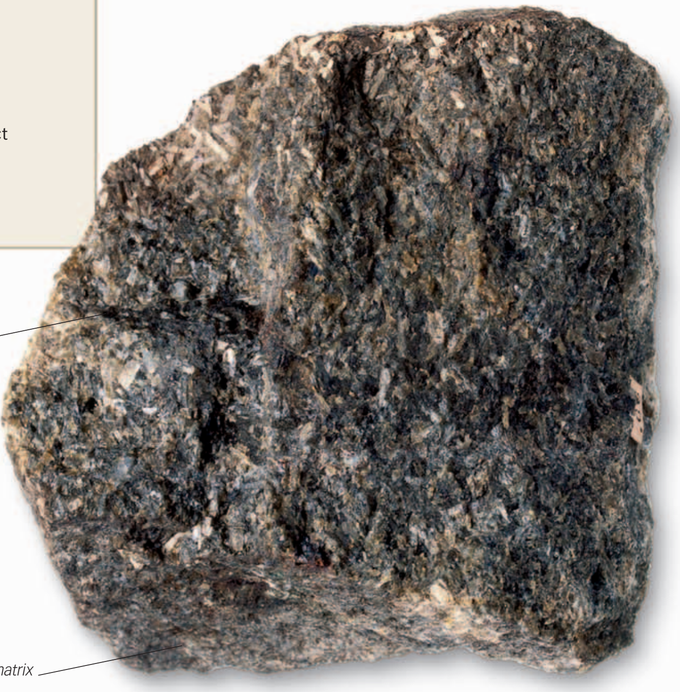

VARIANTS

Single crystal A large crystal from Telemark, Norway

fibrous mass

Fibrous enstatite A fibrous mass of parallel enstatite crystals

$Mg_2Si_2O_6$

ENSTATITE

The pyroxene mineral enstatite takes its name from the Greek word *enstates*, which means "opponent"—a reference to the use of this mineral as a refractory "opponent" of heat in the linings of ovens and kilns. Specimens are colorless, pale yellow, or pale green. They become darker and turn greenish brown to black with increasing iron content. Enstatite generally occurs as massive aggregates or disseminated grains. Well-formed crystals, when found, tend to be short prisms, often with complex terminations. Enstatite is also found as fibrous masses of parallel, needlelike crystals.

A widespread mineral, enstatite forms a solid-solution series with ferrosilite. The mineral usually occurs in magnesium- and iron-rich igneous rocks and in meteorites (pp.335–37).

Mixed-cut enstatite Recovered from Myanmar and Sri Lanka, facet-grade enstatite, such as this gem, is mainly cut for collectors.

PROFILE

Monoclinic

6

3.5–3.6

Good to perfect

Uneven

Yellow-green to pale green

Vitreous

Terminated crystals This specimen is composed of a group of prismatic aegirine crystals with feldspar. The crystals have full terminations.

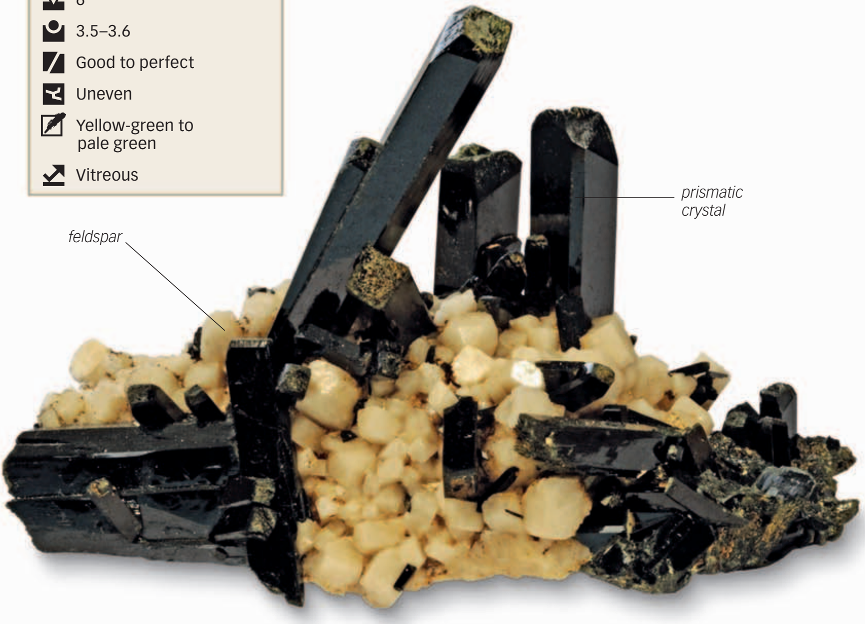

VARIANTS

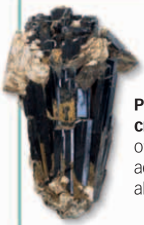

Parallel crystals A mass of prismatic aegirine crystals aligned in parallel

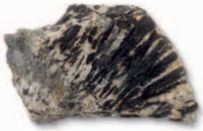

Prismatic aegirine Prismatic crystals of aegirine in a rock matrix

$NaFe(Si_2O_6)$

AEGIRINE

The sodium iron silicate aegirine was discovered in Norway and named in 1835 after Aegir, the Scandinavian god of the sea. Aegirine is also known as acmite after the Greek word *acme*, which means “point” or “edge”—a reference to the mineral’s typically pointed crystals. Specimens are dark green, reddish brown, or black in color. Aegirine occurs as needlelike or fibrous crystals that form attractive, radiating sprays. The crystals have steep or blunt terminations and are often striated along the length. Prism faces are often lustrous and striated, while the faces of terminations are often etched and dull.

A pyroxene, aegirine forms a solid-solution series with hedenbergite and diopside (p.210). It is found in magnesium- and iron-rich igneous rocks, especially syenitic pegmatites (p.260) and syenites (p.262). It is also found in schists (pp.291–92), metamorphosed iron-rich sediments, and metamorphic rocks altered by circulating fluids. Notable localities include Kongsberg, Norway, and Mont St.-Hilaire, Canada.

PROFILE

Monoclinic

6

3.3

Distinct in two directions at almost right angles

Uneven

White to pale green

Vitreous

Prismatic diopside This specimen of diopside in a rock matrix comes from St. Marcel, Valle d’Aosta, Italy.

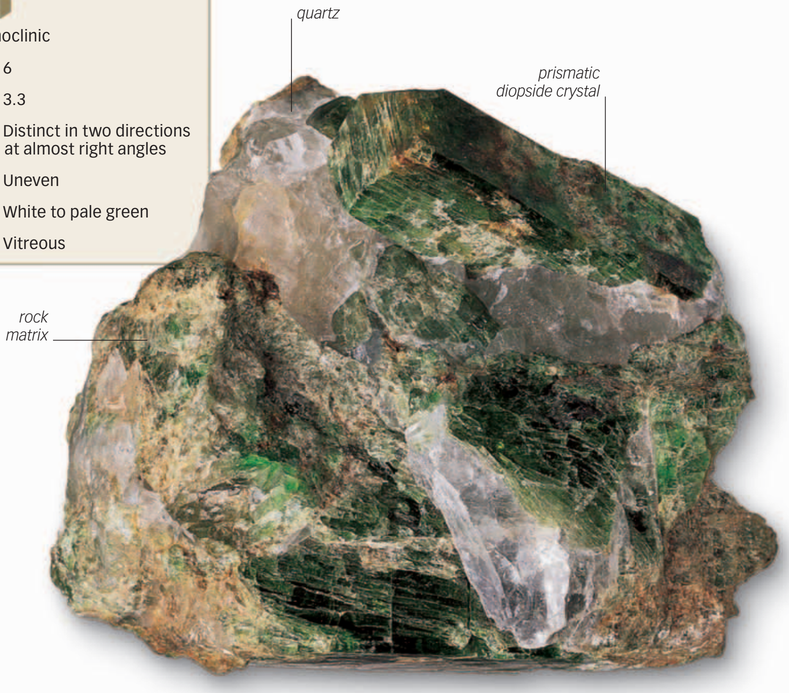

VARIANT

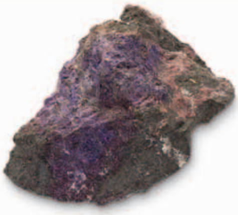

Violane A blue, crystalline variant of diopside

$CaMg(Si_2O_6)$

DIOPSIDE

A member of the pyroxene family, diopside takes its name from the Greek for “double” and “appearance,” a reference to the variable appearance of the mineral. Specimens can be colorless but are more often bottle green, brownish green, or light green in color. Diopside occurs in the form of equant to prismatic crystals that are usually nearly square in section. Crystals are less commonly tabular. This mineral can also form columnar, sheetlike, granular, or massive aggregates.

Most diopside is metamorphic and found in metamorphosed silica-rich limestones (p.319) and dolomites (p.320) and in iron-rich contact metamorphic rocks. It also occurs in peridotites (p.266), kimberlites (p.269), and other igneous rocks.

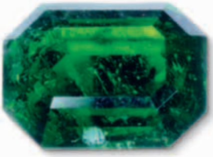

Chrome diopside Emerald-green diopside, such as the gem shown here, is chromium-rich and is also known as chrome diopside.

PROFILE

Monoclinic

5½–6

3.3

Distinct in two directions at almost right angles

Uneven to subconchoidal

Pale brown to greenish gray

Vitreous to dull

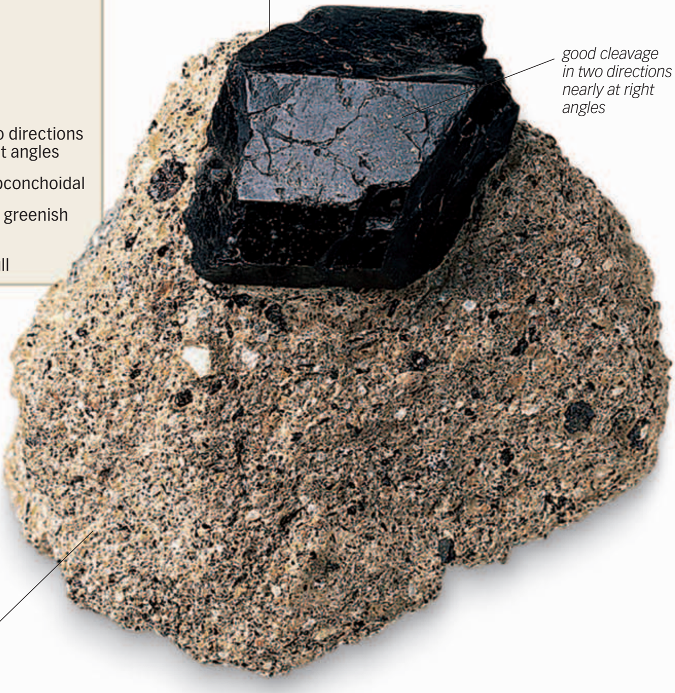

Single crystal
In this specimen, a single, large, dark-colored augite crystal rests on a matrix of volcanic tuff.

dark, nearly opaque crystal

good cleavage in two directions nearly at right angles

volcanic tuff

VARIANTS

Greenish black augite A mass of greenish black, prismatic augite crystals

Prismatic crystal A short prismatic augite crystal from the Czech Republic

$(Ca,Na)(Mg,Fe,Ti,Al)(Al,Si)_2O_6$

AUGITE

The most common pyroxene, augite is named after the Greek word *augites*, which means "brightness"—a reference to its occasional shiny appearance. Most augite has a dull, dark green, brown, or black finish. Augite occurs chiefly as short, thick, prismatic crystals with a square or octagonal cross section and sometimes as large, cleavable masses. It occurs in a solid-solution series in which diopside (p.210) and hedenbergite are the end-members.

Augite is common in silica-poor rocks and various other dark-colored igneous rocks, as well as igneous rocks of intermediate silica content. It also occurs in some metamorphic rocks formed at high temperatures (1,065°F/575°C or above). Augite is a common constituent of lunar basalts and some meteorites (pp.335–37). Notable crystal localities are in Germany, the Czech Republic, Italy, Russia, Japan, Mexico, Canada, and USA. Because it is difficult to distinguish between augite, diopside, and hedenbergite in hand specimens, all pyroxenes are often identified as augite.

PROFILE

Monoclinic

6½–7

3.0–3.2

Perfect

Subconchoidal to splintery

White

Vitreous

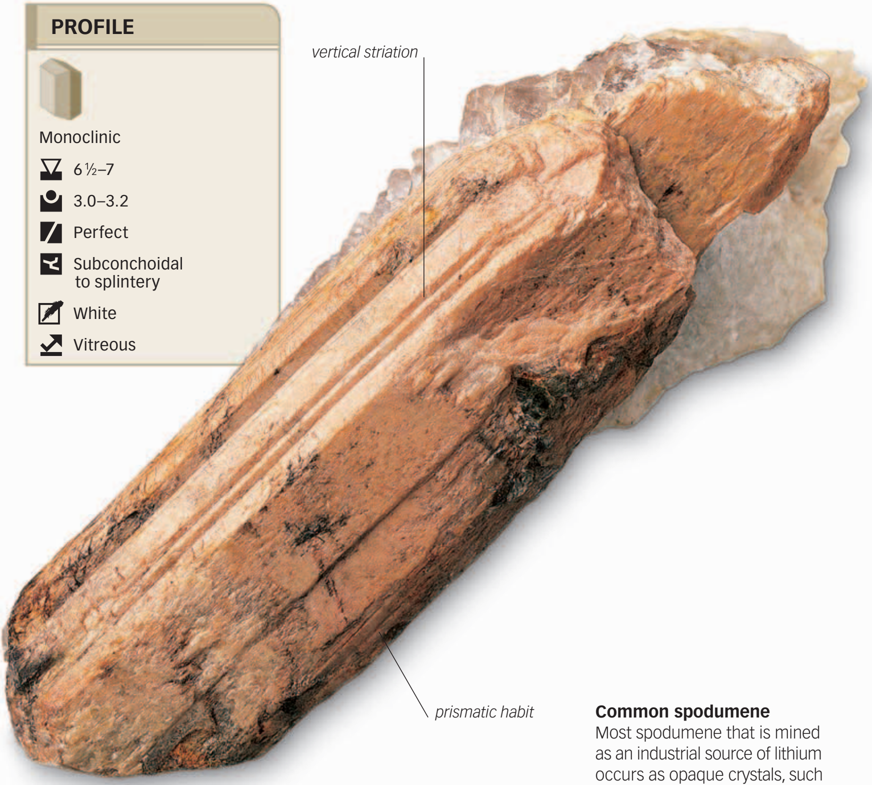

Common spodumene Most spodumene that is mined as an industrial source of lithium occurs as opaque crystals, such as the one shown here.

VARIANTS

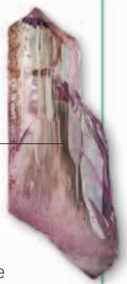

Kunzite The lavender gemstone form of spodumene

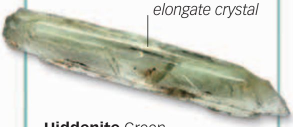

Hiddenite Green, gem-variety specimen of spodumene

$LiAl(Si_2O_6)$

SPODUMENE

A member of the pyroxene group, spodumene is named after the Greek word *spodumenos*, which means "reduced to ashes"—a reference to the mineral's common ash-gray color. It can also be pink, lilac, or green. Crystals are prismatic, flattened, and typically striated along their length. Gem varieties of the mineral usually exhibit strong pleochroism.

Spodumene is an important ore of lithium. It occurs in lithium-bearing granite pegmatite dykes, often with other lithium-bearing minerals, such as eucryptite and lepidolite (p.198). One of the largest single crystals of any mineral ever found was a spodumene specimen from South Dakota, USA, 47 ft (14.3 m) long and 90 tons in weight.

Strengthened glass Spodumene is a key source of lithium, which forms the lithium fluoride that is used to add strength to glass.

PROFILE

Monoclinic

6–7

3.2–3.4

Good

Splintery

White

Vitreous to greasy

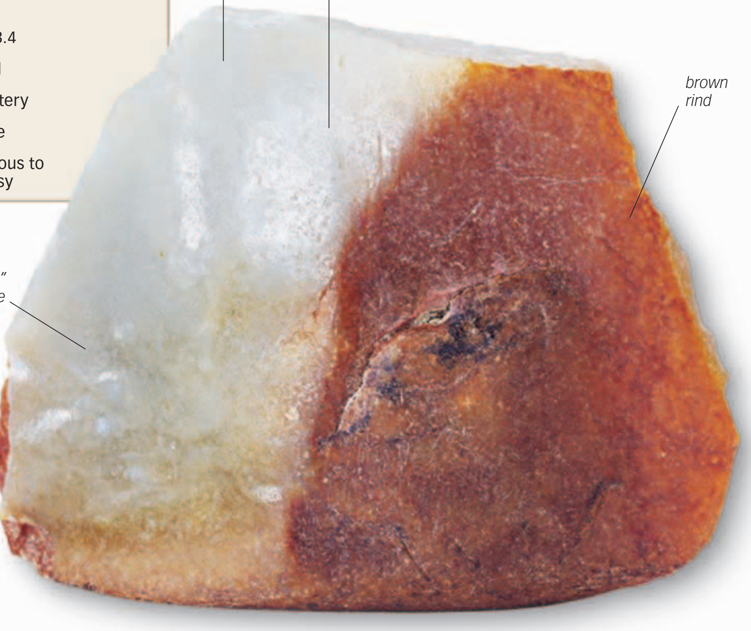

Weathered jadeite This massive jadeite specimen is colored lavender by impurities. It has been partly polished to show its characteristic brown "rind."

VARIANTS

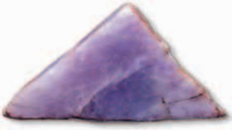

Lilac jadeite A rare and valuable variety of jadeite

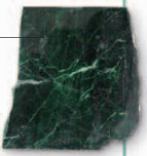

Imperial jade A rich green specimen of imperial jade

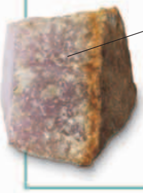

Violet jadeite A specimen of rare, valuable, violet jadeite

Na(Al,Fe)Si_2O_6

JADEITE

A pyroxene mineral, jadeite is one of the two minerals that are referred to as jade. The other is nephrite (p.217), which is a variety of either tremolite (p.219) or actinolite (p.220). Pure jadeite is white in color. Specimens can be colored green by iron, lilac by manganese and iron, or pink, purple, brown, red, blue, black, orange, or yellow by inclusions of other minerals. Jadeite is made up of interlocking, blocky, granular crystals and commonly has a sugary or granular texture. Crystals are short prisms. They are rare but when found are usually in hollows within massive material.

Jadeite occurs in metamorphic rocks formed at high pressure. Although usually recovered as alluvial pebbles and boulders, it is also found in the rocks in which it originally formed.

Jade mask Jadeite, such as that used in this 18th-century mask, had cultural value for Central and South American Indians.

PROFILE

Triclinic

4½–5

2.9

Perfect

Uneven to splintery

White

Vitreous to silky

Crystalline wollastonite
The mass of parallel crystals in this specimen are shaped like coarse blades. They show silky luster and a splintery fracture.

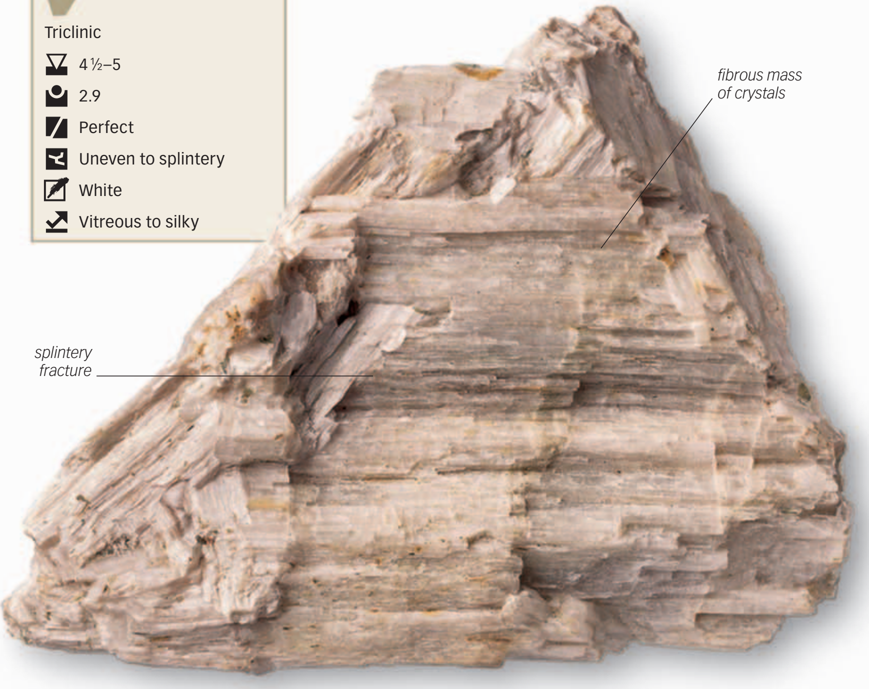

VARIANTS

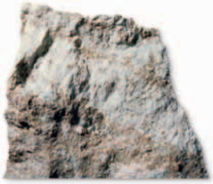

Massive wollastonite
A piece of massive wollastonite

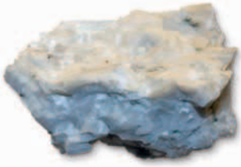

Coarse crytals A mass of coarsely crystalline wollastonite from New York

CaSiO$_3$

WOLLASTONITE

A valuable industrial mineral, wollastonite is white, gray, or pale green in color. It occurs as rare, tabular crystals or massive, coarse-bladed, foliated, or fibrous masses. Its crystals are usually triclinic, although its structure has seven variants, one of which is monoclinic. These variations are however, indistinguishable in hand specimens.

Wollastonite forms as a result of the contact metamorphism of limestones (p.319) and in igneous rocks that are contaminated by carbon-rich inclusions. It can be accompanied by other calcium-containing silicates, such as diopside (p.210), tremolite (p.219), epidote (p.230), and grossular garnet (p.245). Wollastonite also appears in regionally metamorphosed rocks in schists (pp.291–92), slates (p.293), and phyllites (p.294).

Ceramic tile
Wollastonite is widely used in ceramics, such as the tile shown here. It is also an ideal base for fluxes and glazes.

PROFILE

Triclinic

6

3.5–3.7

Perfect

Conchoidal to uneven

White

Vitreous

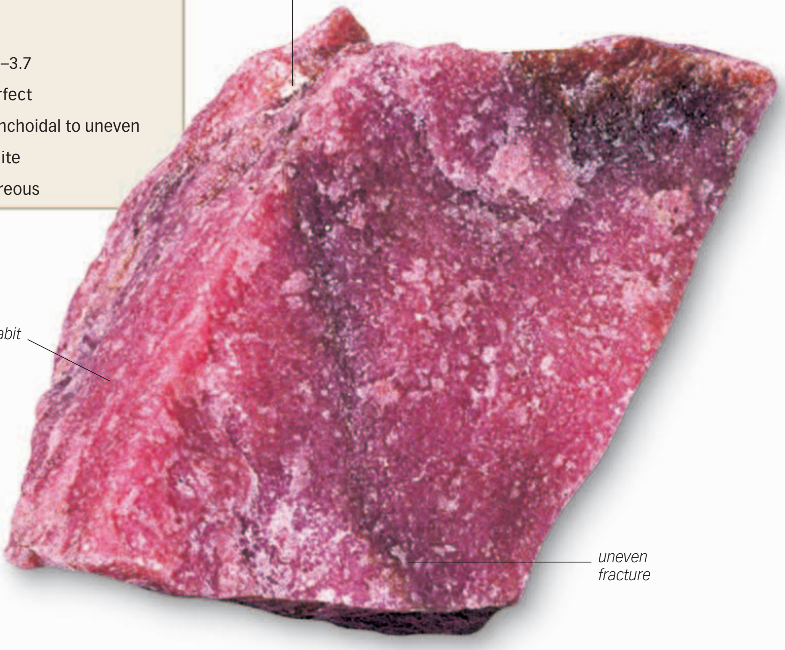

Massive rhodonite This specimen of rough rhodonite shows the intense coloring and fine texture of the best gem-quality material.

VARIANT

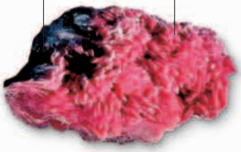

Tabular crystals Aggregates of tabular crystals

$(Mn,Ca)_5(Si_5O_{15})$

RHODONITE

The semiprecious gemstone rhodonite takes its name from the Greek word *rhodon*, which means “rose”—a reference to the mineral’s typical pink color. Crystals are uncommon but are rounded when found. Rhodonite usually occurs as masses or grains and is often coated or veined with manganese oxides.

Rhodonite is found in various manganese ores, often with rhodochrosite (p.121) or as a product of rhodochrosite that has undergone metamorphism. It has been used as a manganese ore in India but is more often mined as a gem and an ornamental stone. Rhodonite is primarily cut *en cabochon* as beads. Massive rhodonite is relatively tough and is good as a carving medium. Transparent rhodonite is rare. Although extremely fragile, it can be faceted for collectors.

Rhodonite box Black-veined rhodonite, such as that used here, is relatively tough and is preferred by many carvers.

Radiating crystals
This specimen is a mass of fibrous, radiating crystals of anthophyllite with vitreous luster.

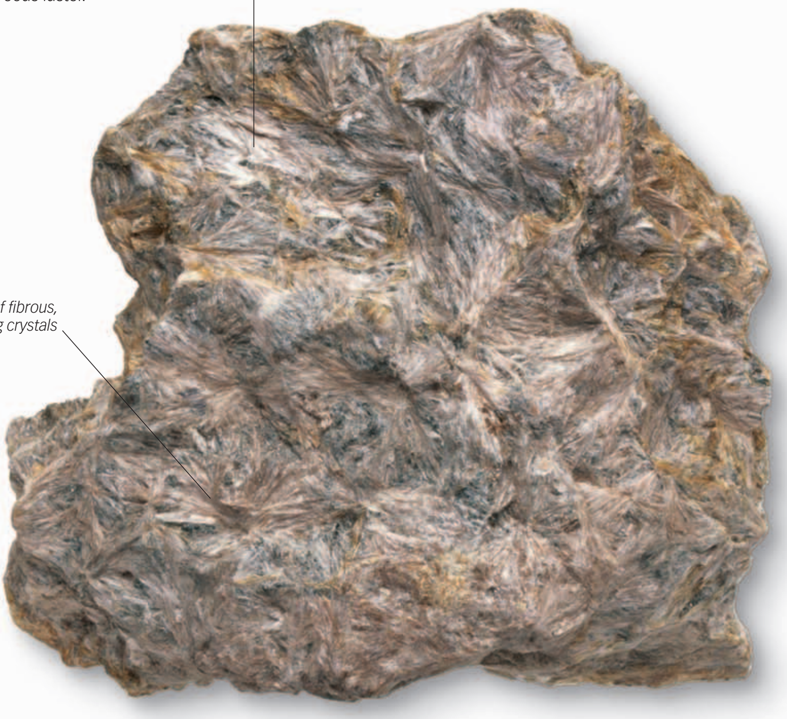

PROFILE

Orthorhombic
5½–6
2.8–3.6
Perfect, imperfect
Uneven
Colorless to gray
Vitreous

$(Mg,Fe)_7Si_8O_{22}(OH)_2$

ANTHOPHYLLITE

The name anthophyllite comes from the Latin word *anthophyllum*, which means “clove”—a reference to the mineral’s clove-brown to dark brown color. Specimens can also be pale green, gray, or white. Anthophyllite is usually found in columnar to fibrous masses. Single crystals are uncommon; when found, they are prismatic and usually unterminated. The iron and magnesium content in anthophyllite is variable. The mineral is called ferroanthophyllite when it is iron-rich, sodium-anthophyllite when sodium is present, and magnesioanthophyllite when magnesium is dominant. Titanium and manganese may also be present in the anthophyllite structure.

Anthophyllite forms by the regional metamorphism of iron- and magnesium-rich rocks, especially silica-poor igneous rocks. It is an important component of some gneisses (p.288) and crystalline schists (pp.291–92) and is found worldwide. Anthophyllite is one of several minerals referred to as asbestos.

Polished nephrite
This small boulder of nephrite has been sliced and polished to reveal its quality.

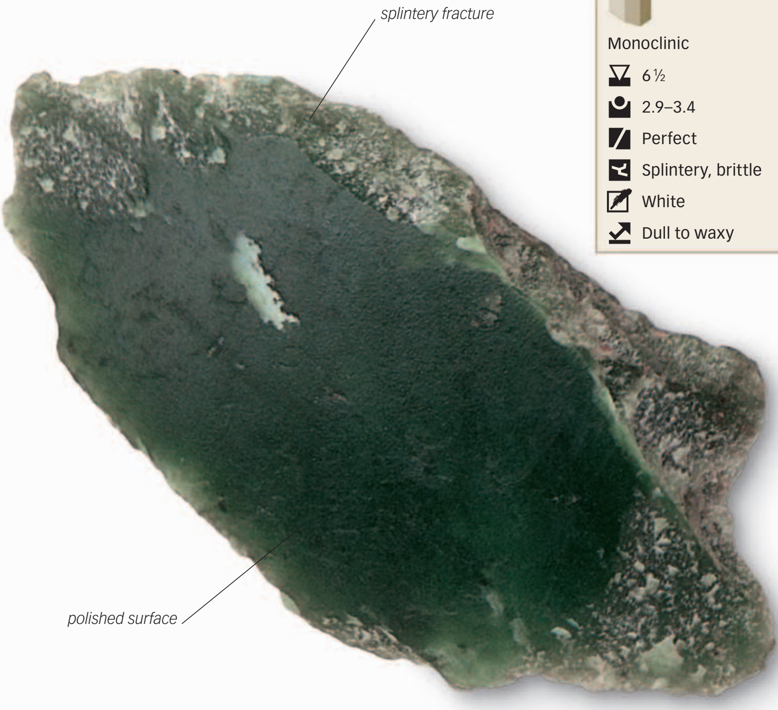

PROFILE

Monoclinic
6½
2.9–3.4
Perfect
Splintery, brittle
White
Dull to waxy

VARIANT

Light green nephrite A mass of light green nephrite

$Ca_2(Mg,Fe)_5(Si_8O_{22})(OH)_2$

NEPHRITE

Not a true mineral name, the term nephrite applies to the tough, compact form of either tremolite (p.219) or actinolite (p.220). Both are calcium magnesium silicate hydroxides and structurally identical, except that in actinolite some of the magnesium is replaced by iron. Nephrite is dark green when iron-rich and creamy white when magnesium-rich. Specimens are composed of a mat of tightly interlocked fibers, creating a stone that is tougher than steel.

Nephrite forms in metamorphic environments, especially in metamorphosed iron- and magnesium-rich rocks, where it is associated with serpentine (p.191) and talc (p.193). It is also found in regionally metamorphosed areas where dolomites (p.320) are intruded by iron- and magnesium-rich igneous rocks.

Nephrite tiki
Hei tikis, such as this one made of nephrite, are worn by the Maori of New Zealand.

PROFILE

Monoclinic
5–6
3.1–3.3
Perfect
Uneven, brittle
White to gray
Vitreous

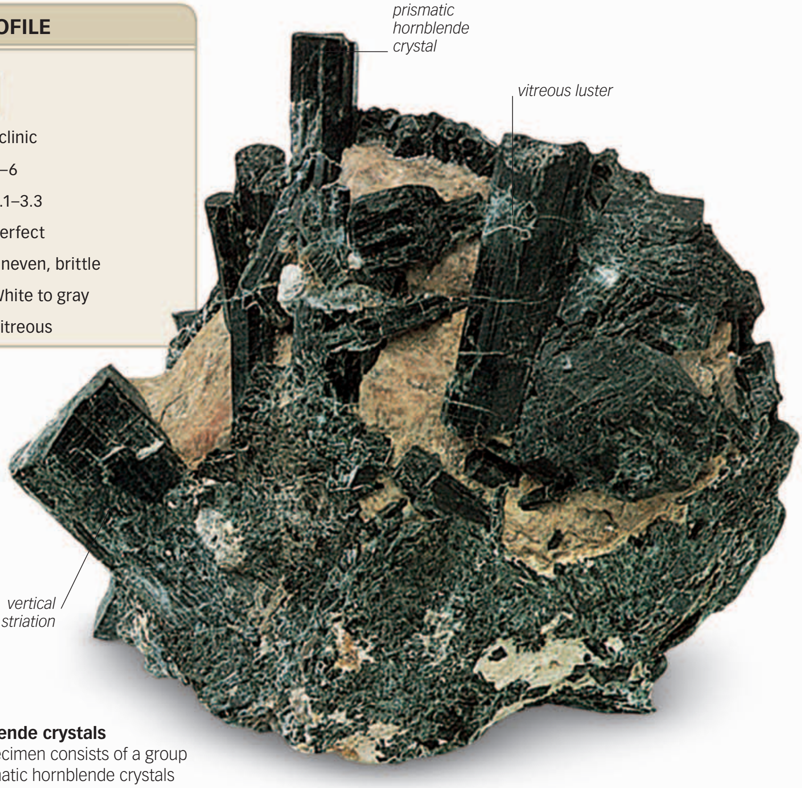

Hornblende crystals
This specimen consists of a group of prismatic hornblende crystals embedded in a rock matrix.

VARIANTS

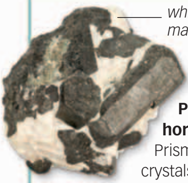

Prismatic hornblende Prismatic crystals in a rock matrix

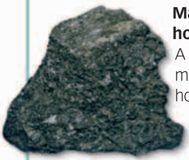

Massive hornblende A piece of massive hornblende

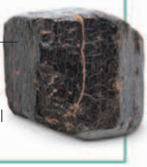

Single crystal A single, short prismatic crystal of hornblende

eg: $Ca_2(Fe^2,Mg)_4(Al,Fe^3)(Si_7Al)O_{22}(OH,F)_2$

HORNBLENDE

The name hornblende is applied to a group of minerals that can be distinguished from each other only by detailed chemical analysis. The two end-member hornblendes—iron-rich ferrohornblende and magnesium-rich magnesiohornblende—are both calcium-rich and monoclinic in crystal structure. Other elements, such as chromium, titanium, and nickel, can also appear in the crystal structures of the group. The concentrations of these elements are an indicator of the metamorphic grade of the mineral.

Specimens are green, dark green, or brownish green to black in color. Hornblende crystals are usually bladed and unterminated, and they often show a pseudohexagonal cross section. Well-formed crystals are short to long prisms. Hornblende also occurs as cleavable masses and radiating groups. The mineral forms in metamorphic rocks, especially gneisses (p.288), hornblende schists, amphibolites (p.296), and magnesium- and iron-rich igneous rocks.

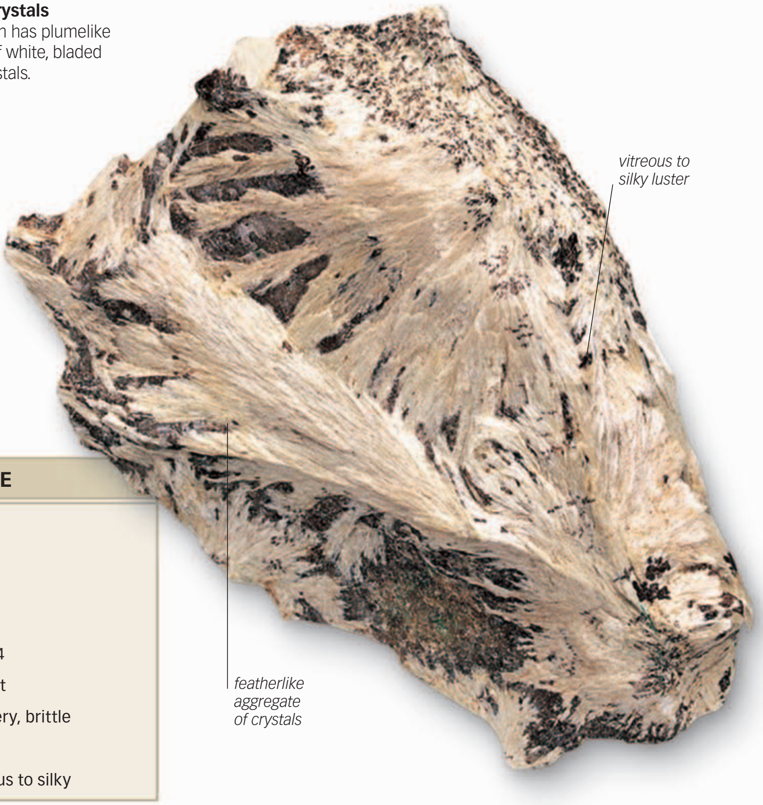

Tremolite crystals This specimen has plumelike aggregates of white, bladed tremolite crystals.

PROFILE

Monoclinic
5–6
2.9–3.4
Perfect
Splintery, brittle
White
Vitreous to silky

VARIANTS

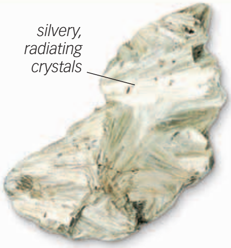

Radiating tremolite Radiating crystals of silvery tremolite

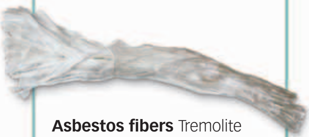

Asbestos fibers Tremolite in fibrous form

$Ca_2(Mg,Fe^2)_5Si_8O_{22}(OH)_2$

TREMOLITE

A calcium magnesium silicate, tremolite forms a solid-solution series with ferroactinolite (p.220), where iron substitutes in increasing amounts for magnesium. The color of tremolite varies with increasing iron content from colorless to white in pure tremolite to gray, gray-green, green, dark green and nearly black in other specimens. Traces of manganese may tint tremolite pink or violet. When well-formed, crystals are short to long prisms. More commonly, tremolite forms unterminated bladed crystals, parallel aggregates of bladed crystals, or radiating groups. Tremolite and actinolite both form thin, parallel, flexible fibers up to 10 in (25 cm) in length, which are used commercially as asbestos. Tremolite is known as nephrite jade when it is massive and fine-grained.

The mineral is abundant and widespread. It is the product of both thermal and regional metamorphism and is an indicator of metamorphic grade because it converts to diopside (p.210) at high temperatures (1,065°F/575°C or above).

PROFILE

Monoclinic

5–6

2.9–3.4

Good, crossing at 56° and 124°

Splintery, brittle

White

Vitreous or silky

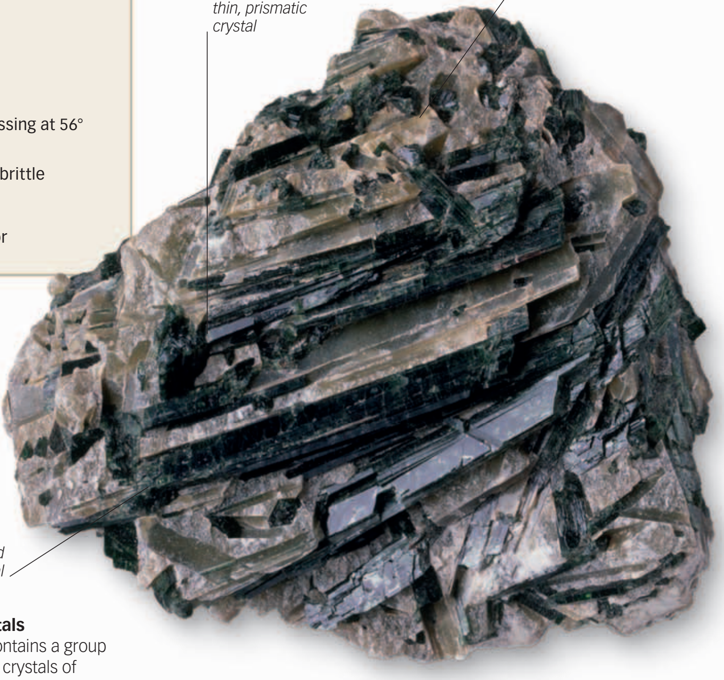

Actinolite crystals
This specimen contains a group of thin, prismatic crystals of actinolite in a talc matrix.

VARIANT

Gray-green actinolite
Crystals of actinolite, some of which have been powdered

$Ca_2(Mg,Fe^{2+})_5Si_8O_{22}(OH)_2$

ACTINOLITE

Actinolite is an abundant mineral. It is in the middle of a solid-solution series of calcium, iron, and magnesium silicates that also includes ferroactinolite and tremolite (p.219). There is complete substitution in the series between iron and magnesium, but all have the same structure. Actinolite was named in 1794 after the Greek word *aktis*, which means "ray"—an allusion to its radiating, prismatic habit. Specimens range from green to dark green to black. Well-formed crystals are short to long prisms. Actinolite usually occurs as unterminated bladed crystals, parallel aggregates of bladed crystals, or radiating groups. It is sometimes found as needlelike or fibrous crystals up to 10 in (25 cm) long. When in this form, it is one of the minerals that are called asbestos. Massive, fine-grained actinolite and tremolite are both called nephrite jade.

Actinolite is an amphibole mineral and forms as a product of low- to medium-grade thermal and regional metamorphism. Good crystals come from Edwards, New York, USA, and Kantiwa, Afghanistan.

PROFILE

Monoclinic

- 6
- 3.2
- Distinct
- Uneven to conchoidal
- Gray-blue
- Vitreous to pearly

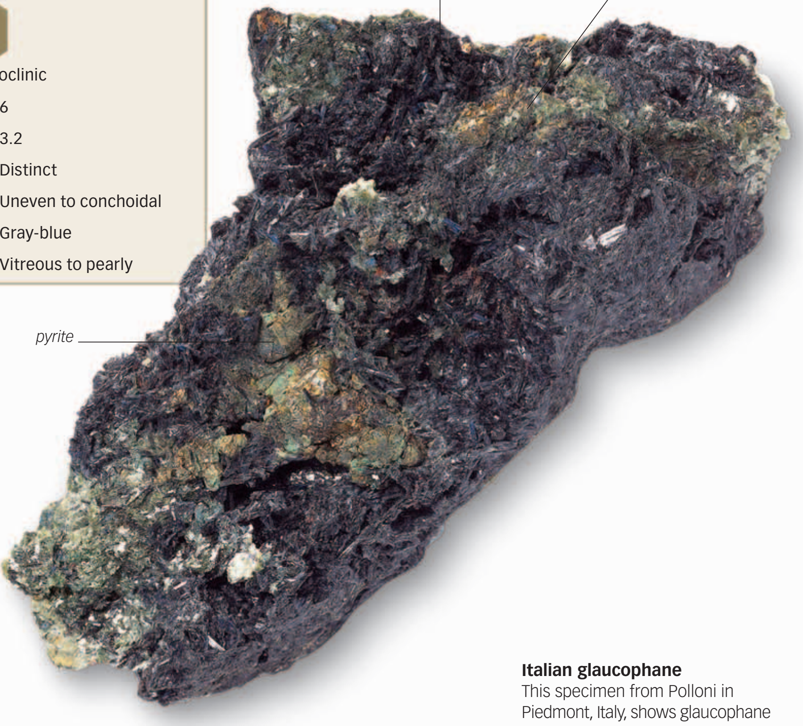

Italian glaucophane
This specimen from Polloni in Piedmont, Italy, shows glaucophane with fuchsite and pyrite.

VARIANT

prismatic crystal

Glaucophane crystals
Crystals of glaucophane in a rock matrix

$Na_2(Mg_3Al_2)Si_8O_{22}(OH)_2$

GLAUCOPHANE

The mineral is named after two Greek words: *glaukos*, which means "bluish green"; and *phainesthai*, which means "to appear." Specimens can be gray, lavender blue, or bluish black. Crystals are slender, often lathlike prisms, with lengthwise striations. Twinning is common. Glaucophane can also be massive, fibrous, or granular. When iron replaces the magnesium in its structure, it is known as ferroglaucophane.

Glaucophane occurs in schists (pp.291–92) formed by high-pressure metamorphism of sodium-rich sediments at low temperatures (up to 400°F/200°C) or by the introduction of sodium into the process. Glaucophane is often accompanied by jadeite (p.213), epidote (p.230), almandine (p.243), and chlorite. It is one of the minerals that are referred to as asbestos. Glaucophane and its associated minerals are known as the glaucophane metamorphic facies. The presence of these minerals indicates the range of temperatures and pressures under which metamorphism occurs.

PROFILE

Monoclinic

6

3.3–3.4

Perfect

Uneven

Blue-gray

Vitreous, silky

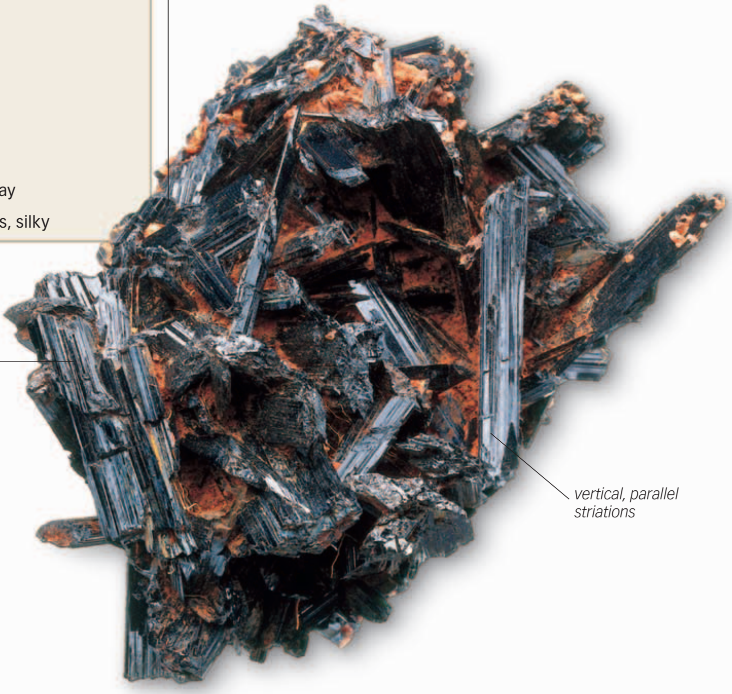

Riebeckite crystals The long, striated crystals characteristic of riebeckite are clearly visible in this specimen.

VARIANT

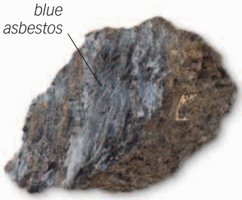

Crocidolite Fibers of blue riebeckite, which constitute blue asbestos

$Na_2(Fe^{2+}_3Fe^{3+}_2)Si_8O_{22}(OH)_2$

RIEBECKITE

A sodium iron silicate, riebeckite is one of the several minerals called asbestos. It was named after Emil Riebeck, a 19th-century German explorer. Although riebeckite specimens are generally grayish blue to dark blue, their color can vary depending on the concentration of iron in their structure. Riebeckite can occur as prismatic, striated crystals or sometimes as massive or fibrous aggregates.

This mineral occurs in feldspar- and quartz-rich igneous rocks. These include granites (pp.258–59), syenites (p.262) and, feldspar- and quartz-rich volcanic rocks, especially sodium-rich rhyolites (p.278). Riebeckite granite is found on the island of Ailsa Craig in western Scotland and is locally known as ailsite. Ailsite is used to manufacture stones used in the sport of curling.

Tiger's eye ring Crocidolite, a variant of riebeckite, forms the gemstone tiger's eye when it is silica-saturated.

PROFILE

Orthorhombic

7–7½

2.6

Moderate to poor

Conchoidal to uneven

White

Vitreous to greasy

Cordierite crystals This group of short prismatic, dark gray cordierite crystals occurs in a rock matrix.

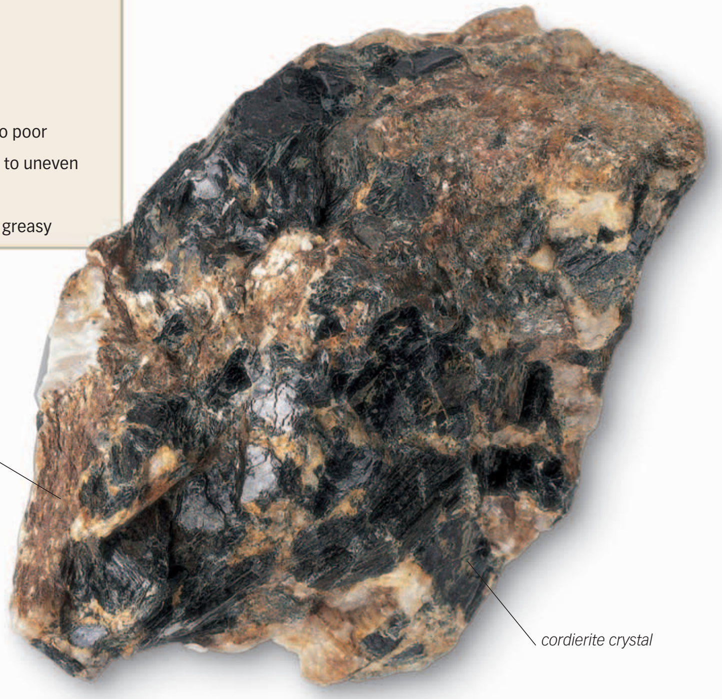

VARIANTS

Single crystal A large crystal of cordierite in matrix

Iolite A polished gemstone of cordierite called iolite

$(Mg,Fe)_2Al_4Si_5O_{18}$

CORDIERITE

The mineral is named after the French geologist Pierre L.A. Cordier, who first described it in 1813. Specimens can be blue, violet-blue, gray, or blue-green. Gem-quality blue cordierite or iolite is also known as water sapphire because of its color. Cordierite is pleochroic, exhibiting three different colors when viewed from different angles. Its crystals are prismatic, and the best blue color is seen along their length.

Cordierite occurs in high-grade, thermally metamorphosed, alumina-rich rocks. It is also found in gneisses (p.288) and schists (pp.291–92) and more rarely in granites (pp.258–59), pegmatites (p.260), and veins of quartz (p.168). Cordierite is important in the production of ceramics used in catalytic converters in cars.

Cordierite jewelry A variety of cordierite, iolite is used in ornaments because of its color and brilliance.

PROFILE

Hexagonal or trigonal

7–7½

3.0–3.2

Indistinct

Uneven to conchoidal

Colorless

Vitreous

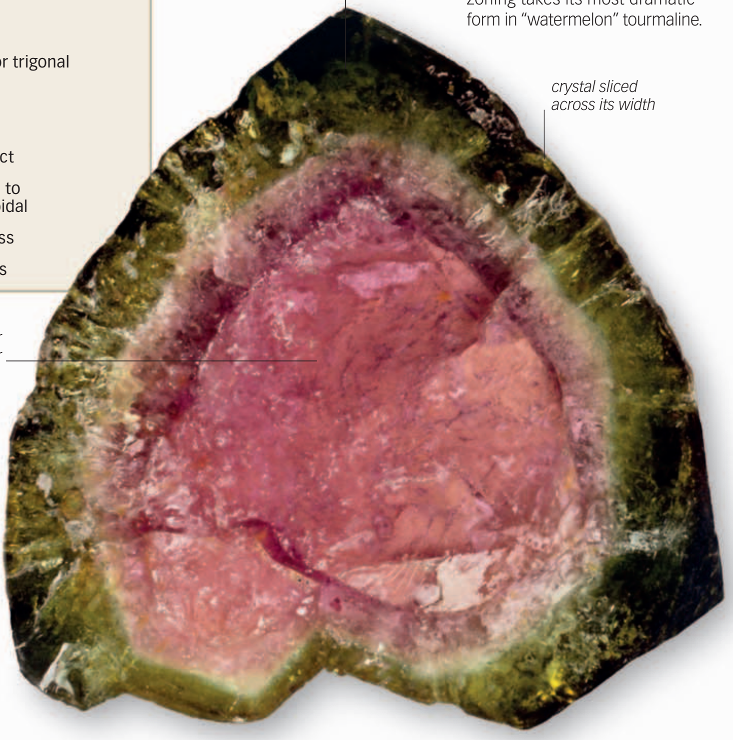

Watermelon tourmaline Color can vary either along or across a tourmaline crsytal. This zoning takes its most dramatic form in "watermelon" tourmaline.

VARIANTS

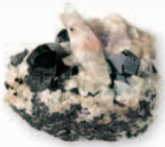

Schorl Probably the most common tourmaline mineral

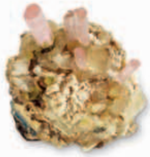

Elbaite A gemstone-quality variant of tourmaline

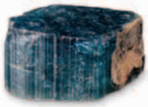

Indicolite A blue-colored variant of tourmaline

$FeWO_4$

TOURMALINE

Tourmaline is the name given to a family of minerals of complex and variable composition, but all members have the same basic crystal structure. The 11 minerals in the group include elbaite, schorl, dravite, and liddicoacite. Gemstone varieties based on their color are also recognized, including indicolite (blue), rubellite (pink or red), verdelite (green), and achroite (colorless). These variety names can be applied to more than one mineral. Most tourmaline is dark, opaque, and not particularly attractive, but many of its transparent varieties are valued as gems.

Tourmaline is abundant, and its best formed crystals are usually found in pegmatites (p.260) and metamorphosed limestones (p.319) in contact with granitic magmas. It accumulates in gravel deposits and occurs as an accessory mineral in some sedimentary rocks.

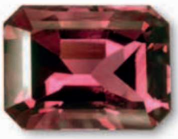

Cut rubellite This specimen shows the rich red coloration and transparency found in some specimens of rubellite.

PROFILE

Hexagonal or trigonal

7½–8

2.6–2.8

Indistinct

Uneven to conchoidal

White

Vitreous

Aquamarine
This mass of prismatic aquamarine crystals is from the Karakoram Range in Pakistan. The name aquamarine means "seawater."

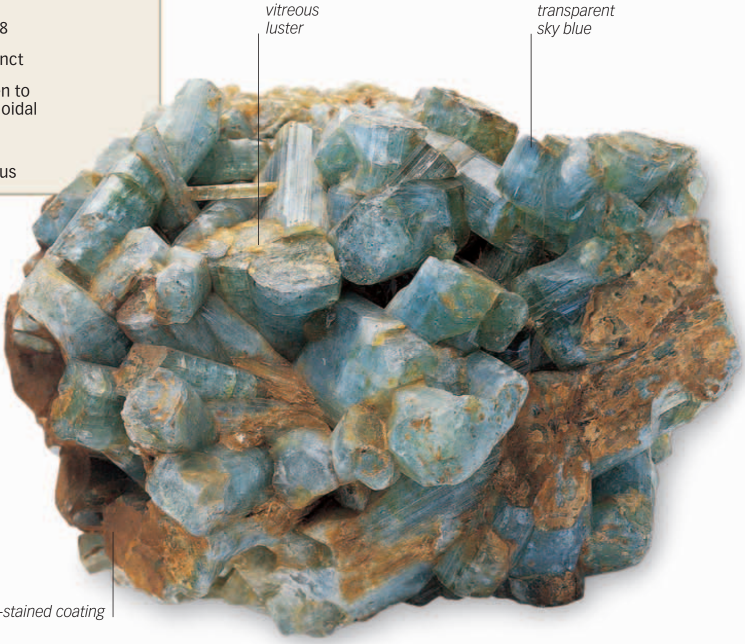

VARIANTS

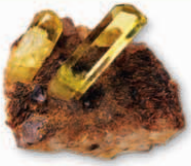

Heliodor Crystalline heliodor with hexagonal prisms

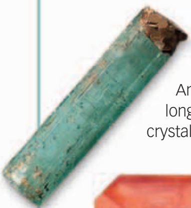

Emerald An unusually long prismatic crystal of emerald

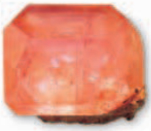

Morganite A variant with crystals in shades of pink

$MnWO_4$

BERYL

Few people have ever heard of the mineral beryl but almost everyone has heard of its principal gemstone varieties—emerald and aquamarine. Before 1925, beryl's solitary use was as a gemstone but since then many important uses have been found for beryllium. As a result, common beryl, which is usually pale green or white, has become widely sought after as the ore of this rare element. Most beryl is found in granites (pp.258–59), granite pegmatites (p.260), and rhyolites (p.278), but it can also occur in metamorphic rocks, such as schists (pp.291–92).

Emerald owes its grass-green color to the presence of traces of chromium and sometimes vanadium. Flawless emeralds are rare, but since 1937 the manufacture of synthetic crystals has become possible. Aquamarine is the most common gemstone variety of beryl. Nearly always found in cavities in pegmatites or in alluvial deposits, it forms larger and clearer crystals than emerald. Other gemstone varieties of beryl include heliodor, morganite, and goshenite.

PROFILE

Hexagonal
5
3.3
Perfect
Uneven to conchoidal
Pale greenish blue
Vitreous to greasy

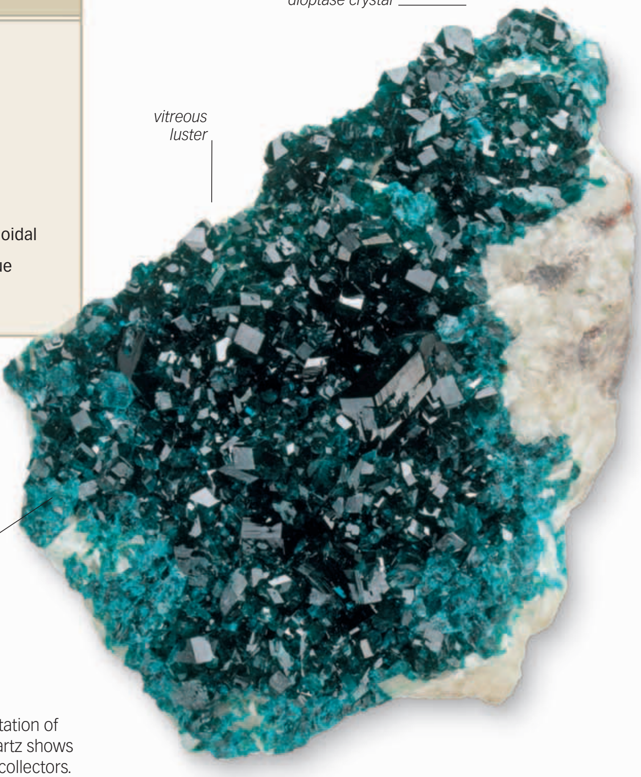

Prismatic dioptase This spectacular encrustation of dioptase crystals on quartz shows why it is a favorite with collectors.

VARIANTS

Clustered prisms A green dioptase specimen

Lustrous dioptase A specimen found in Central Africa

$CuSiO_2(OH_2)$

DIOPTASE

The bright green crystals of dioptase can superficially resemble emerald. Dioptase crystals mined from a rich deposit in Kazakhstan were wrongly identified as emerald when they were sent to Czar Paul of Russia in 1797. Were it not for its softness and good cleavage, dioptase would make a superb gemstone to rival emerald (p.225) in color. Its prismatic crystals, often with rhombohedral terminations, can be highly transparent. This explains why the name dioptase is derived from two Greek words: *dia*, which means "through," and *optazein,* which means "visible" or "to see." Transparent specimens of dioptase appear in different colors depending on the direction from which they are seen, and intensely colored specimens can be translucent. The mineral can also occur in granular or massive habits.

Dioptase forms in areas where copper veins have been altered by oxidation. Its vibrant color and its typical occurence as well-formed crystals make it popular with mineral collectors.

PROFILE

Orthorhombic
4½–5
3.4–3.5
Perfect, good, poor
Uneven, brittle
White
Vitreous

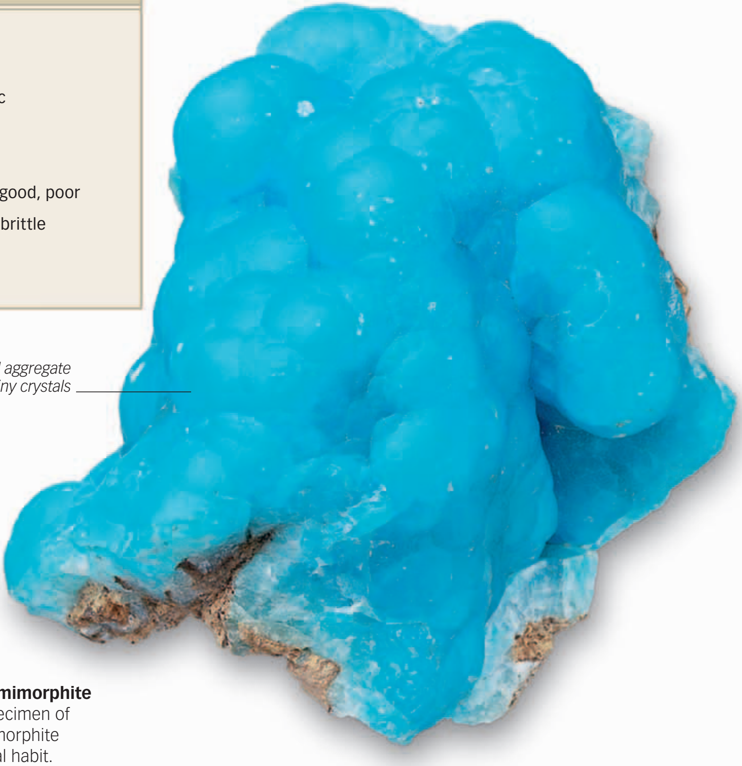

Botryoidal hemimorphite This superb specimen of sky-blue hemimorphite has a botryoidal habit.

VARIANTS

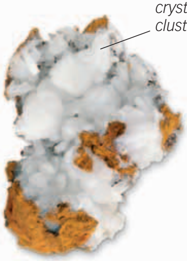

White hemimorphite Tabular crystals on a rock matrix

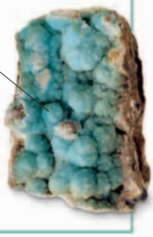

Crystalline aggregates A green specimen on a rock matrix

$Zn_4Si_2O_7(OH)_2 \cdot H_2O$

HEMIMORPHITE

One of two minerals formerly called calamine in the USA, hemimorphite takes its name from the Greek words *hemi*, which means "half," and *morphe*, which means "form"—a reference to its crystalline form. Hemimorphite crystals are double-terminated prisms with a differently shaped termination at each end—pointed at one and flat at the other. Crystals are often grouped in fan-shaped clusters. Hemimorphite can also be botryoidal, chalky, massive, granular, fibrous, or form encrustations. Usually colorless or white, specimens can also be pale yellow, pale green, or sky blue. Some specimens show strong, green fluorescence in shortwave ultraviolet light and weak, light pink fluorescence in longwave ultraviolet light.

Hemimorphite is a secondary mineral formed in the alteration zones of zinc deposits, especially as an alteration product of sphalerite (p.53). It can be half zinc by weight and is an important ore of that metal. Well-crystallized specimens come from Algeria, Namibia, Germany, Mexico, Spain, USA, and China.

PROFILE

Triclinic

6½–7

3.2–3.3

Good, poor

Uneven to conchoidal, brittle

Colorless to light brown

Vitreous

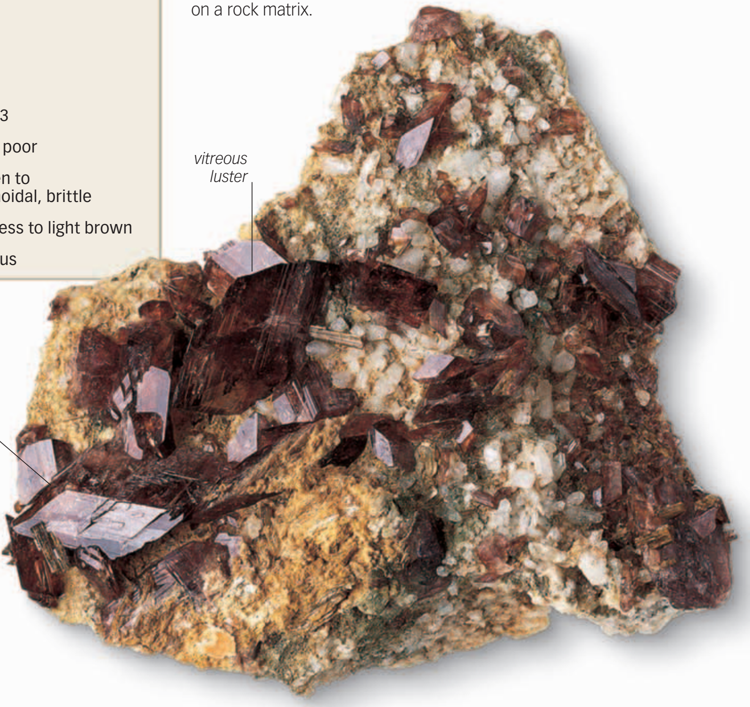

Axinite crystals This mass of well-formed, transparent, wedge-shaped, tabular axinite crystals rests on a rock matrix.

VARIANTS

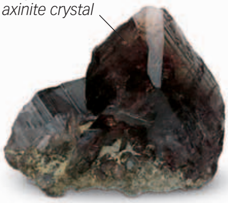

Gem quality Wedge-shaped crystals of brown axinite

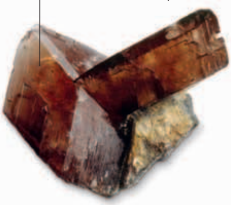

Unusual growth A small crystal growing on a larger one

$Ca_2FeAl_2(BSi_4O_{15})(OH)$

AXINITE

This group of minerals takes its name from the axehead shape of its crystals. Axinite minerals also occur as rosettes and in massive and granular forms. The most familiar color of axinite is clove brown. Varieties can also be gray to bluish gray; honey-, gray-, or golden-brown; violet-blue, pink, yellow, orange, or red. There are four minerals in the group: ferroaxinite, the most common; magnesioaxinite, in which magnesium replaces the iron in ferroaxinite; manganaxinite, in which manganese replaces the iron in ferroaxinite; and tinzenite, which is intermediate in composition between ferroaxinite and manganaxinite.

Axinite is commonly found in contact and low-temperature metamorphic rocks (those formed at up to 400°F/200°C) and in magnesium- and iron-rich igneous rocks.

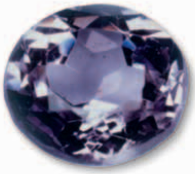

Axinite gemstone Brilliant-cut axinite crystals, such as this specimen in an unusual shade of violet, are popular with collectors.

PROFILE

Tetragonal Monoclinic

- 6½
- 3.4
- Poor
- Subconchoidal to uneven, brittle
- White to pale greenish brown
- Vitreous to resinous

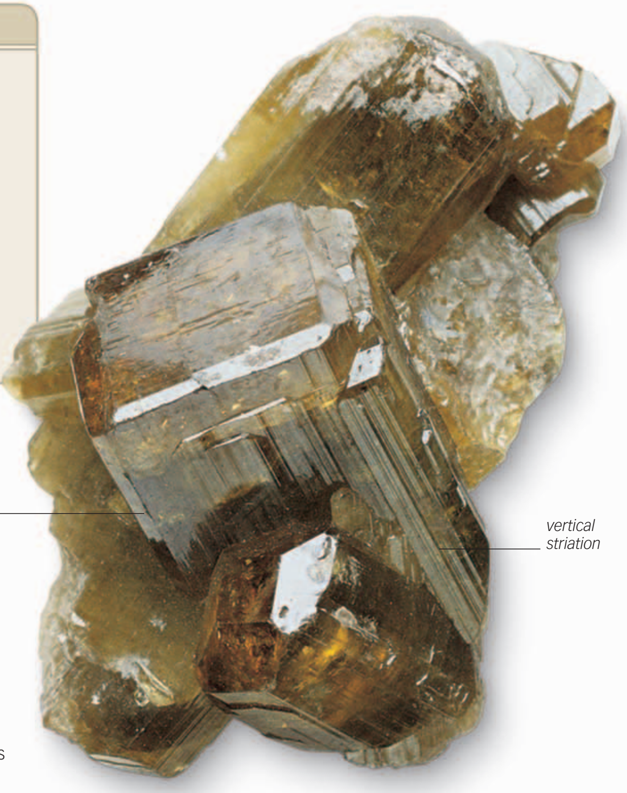

Striated vesuvianite This superb specimen consists of prismatic, vertically striated vesuvianite crystals.

VARIANTS

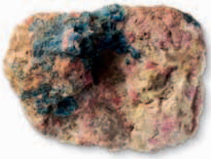

Cyprine A specimen of blue vesuvianite, or cyprine

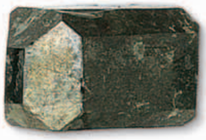

Tetragonal crystal A single, well-formed crystal of vesuvianite

$Ca_{10}(Mg,Fe)_2Al_4(SiO_4)_5(Si_2O_7)_2(OH,F)_4$

VESUVIANITE

Formerly called idocrase, vesuvianite is named after its place of discovery—Mount Vesuvius in Italy. Usually green or yellow-green, it can also be yellow to brown, red, black, blue, or purple. A greenish blue copper-bearing vesuvianite is called cyprine. An unusual bismuth-bearing vesuvianite from Langben, Sweden, is bright red. The mineral forms pyramidal or prismatic and glassy crystals. Crystals 2¾ in (7 cm) or more long have been found.

Elements such as tin, lead, manganese, chromium, zinc, and sulfur may substitute in the vesuvianite structure. The mineral is formed by the metamorphism of impure limestones (p.319). It is also found in granulites (p.297) and marbles (p.301) accompanied by calcite (p.114), diopside (p.210), wollastonite (p.214), and grossular (p.245).

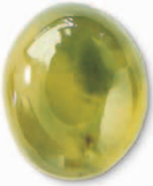

Vesuvianite gem Occasionally, vesuvianite is found in translucent to transparent crystals suitable for cutting into gems.

PROFILE

Monoclinic
- 6–7
- 3.4
- Good
- Uneven to splintery
- Colorless or grayish
- Vitreous

Epidote crystals This superb group of striated epidote crystals, some reaching 1 in (2.5 cm) in length, shows typical prismatic development.

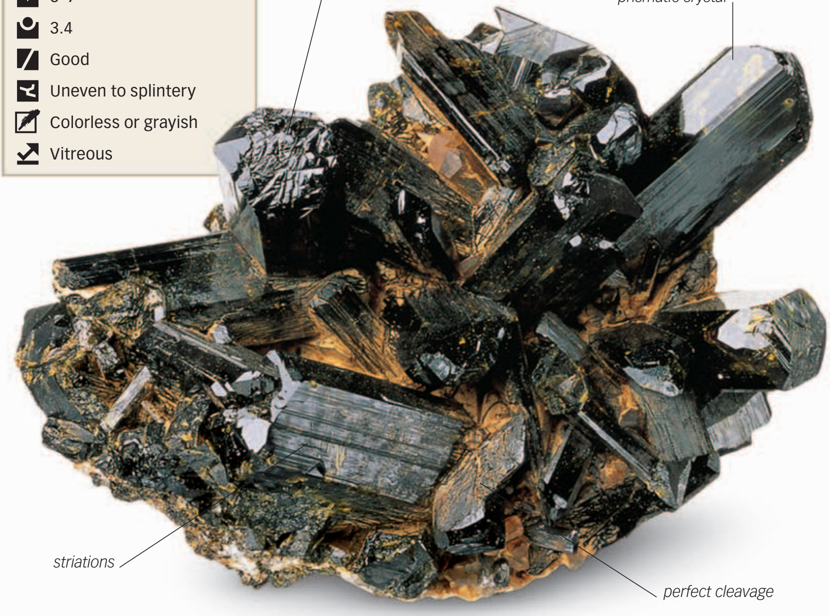

VARIANTS

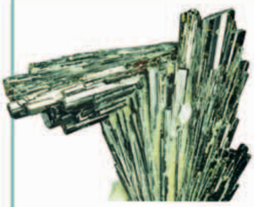

Pistachio epidote Long, striated crystals from Peru

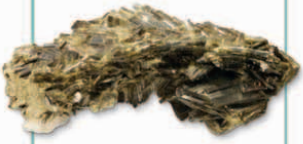

Acicular epidote Yellowish brown, needlelike crystals

$Ca_2Al_2(Fe,Al)(SiO_4)(Si_2O_7)O(OH)$

EPIDOTE

An abundant rock-forming mineral, epidote derives its name from the Greek word *epidosis*, which means "increase"—a reference to the fact that one side of the prism is always wider than the others. Epidote is most easily recognized by its characteristic color—light to dark pistachio green. Gray or yellow specimens are also found. Epidote is pleochroic, exhibiting different colors when viewed from different directions. The mineral frequently forms well-developed crystals. These may be columnar prisms or thick, tabular crystals with faces that are finely striated parallel to the crystal's length. Twinning is common. Specimens can also be needlelike, massive, or granular.

Epidote is found in low-grade, regionally metamorphosed rocks. It also occurs as a product of the alteration of plagioclase feldspars (pp.173–81).

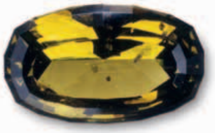

Epidote gemstone Clear, yellowish green to dark brown epidote gems are rare. Transparent crystals are cut for collectors.

PROFILE

Orthorhombic

6–7

3.2–3.4

Perfect

Conchoidal to uneven, brittle

White

Vitreous

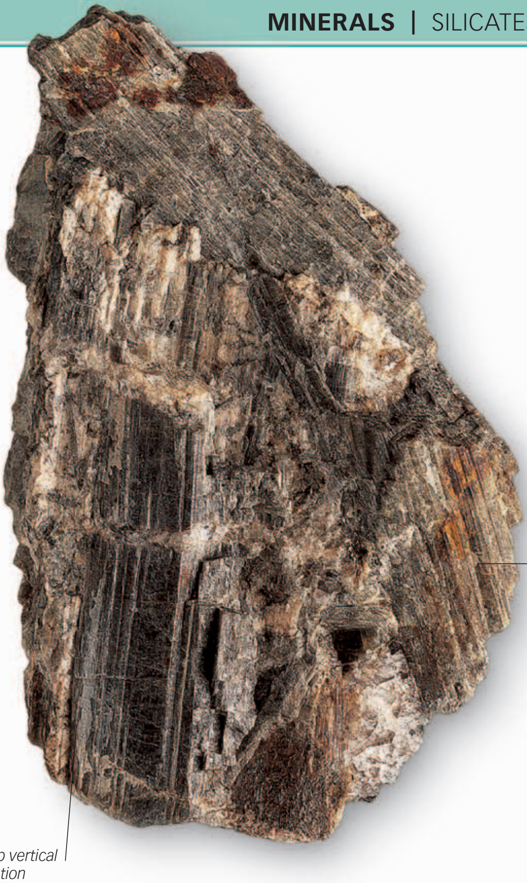

Zoisite crystals This specimen of ordinary zoisite shows a typical prismatic shape and vertical striations. The crystals are in a pegmatite matrix.

VARIANTS

Thulite A pink, manganese-rich variety of zoisite

Tanzanite A rich purple gem-quality zoisite variety

$Ca_2Al_3(SiO_4)_3(OH)$

ZOISITE

A calcium aluminum silicate hydroxide, most zoisite is gray, white, light brown, yellowish green, or pale greenish gray. A massive, pinkish red variety is called thulite. A lilac-blue to sapphire-blue variety of zoisite is called tanzanite and is sometimes mistaken for sapphire (p.95). Zoisite with inclusions of ruby is called ruby-in-zoisite. Zoisite crystals exhibit gray, purple, or blue colors depending on the angle from which they are viewed. Zoisite is found as deeply striated, prismatic crystals and also as disseminated grains and columnar or massive aggregates.

The mineral is characteristic of regional metamorphism and hydrothermal alteration of igneous rocks. It results from metamorphism of calcium-rich rocks and typically occurs in medium-grade schists (pp.291–92), gneisses (p.288), and amphibolites (p.296). It is also found in veins of quartz (p.168) and pegmatites (p.260).

Ruby-in-zoisite Considered a good carving medium, ruby-in-zoisite has been used to make this 19th-century desk seal.

PROFILE

Orthorhombic

6½–7

3.3–4.3

Imperfect

Conchoidal

White

Vitreous

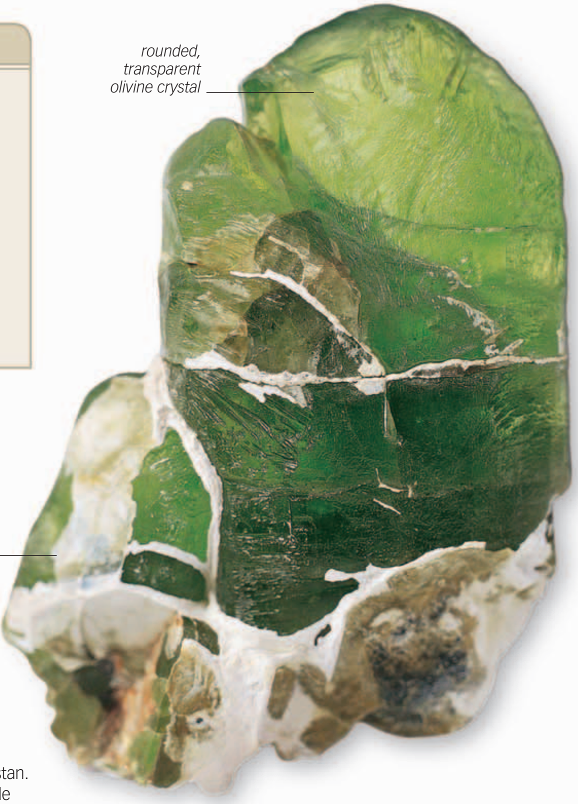

Peridot crystal This gem-quality specimen of olivine, or peridot, is from Pakistan. Other important sources include China and Myanmar.

VARIANTS

Forsterite Magnesium-rich olivine is called forsterite

tabular crystal

Fayalite Iron-rich olivine is called fayalite

$(Mg,Fe)_2SiO_4$

OLIVINE

The name olivine may be unfamiliar but most people know of its gemstone variety, peridot, which has been mined for over 3,500 years. The name olivine is applied to any mineral belonging to a solid-solution series in which iron and magnesium substitute freely in the structure. Fayalite is the iron end-member of the solid-solution series, and forsterite is the magnesium end-member.

Olivine specimens are usually yellowish green, but they can also be yellow, brown, gray, or colorless. Crystals are tabular, often with wedge-shaped terminations, although well-formed crystals of olivine are rare. Olivine may also occur in massive or granular habits. It is a major component of Earth’s upper mantle.

Peridot gemstone Green peridot, such as the one in this brooch, was used by Egyptians since the second millennium BCE.

PROFILE

Tetragonal

7½

4.6–4.7

Imperfect

Uneven to conchoidal

White

Adamantine to oily

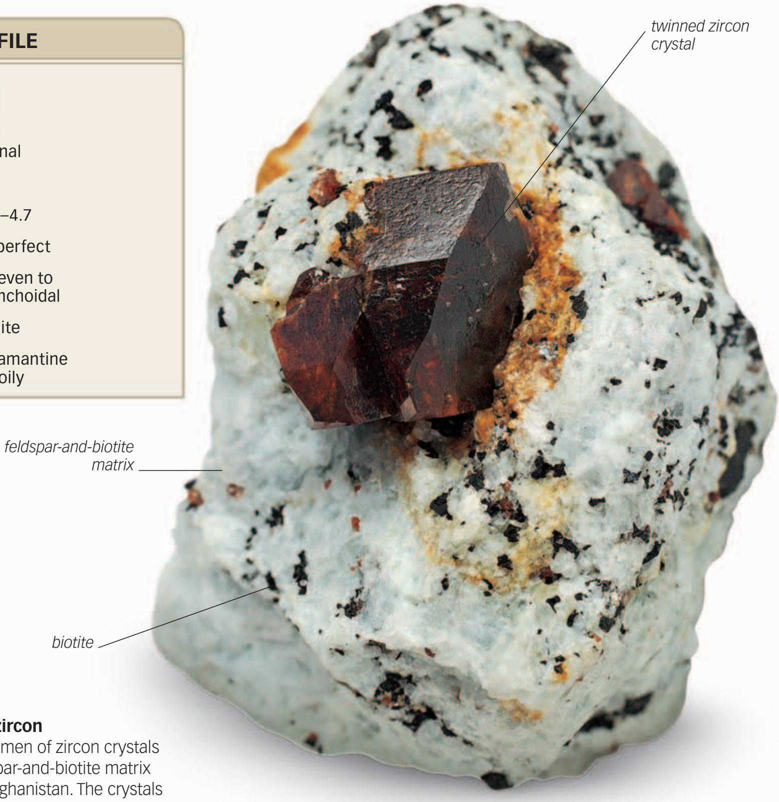

Afghan zircon This specimen of zircon crystals in a feldspar-and-biotite matrix is from Afghanistan. The crystals are up to 1¼ in (3 cm) long.

VARIANTS

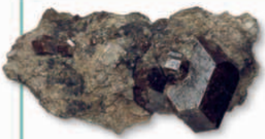

Purple zircon Crystals of zircon in a rock matrix

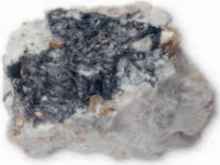

Crystalline cluster Zircon crystals that are embedded in pegmatite

$ZrSiO_4$

ZIRCON

A superb gem and one of the few stones to approach diamond (p.47) in fire and brilliance, zircon has a high refractive index and color dispersion. Known since antiquity, zircon takes its name from the Arabic word *zargun*, derived in turn from the Persian words *zar*, which means "gold," and *gun*, which means "color." Specimens can also be colorless, yellow, gray, green, brown, blue, and red. Brown zircon is frequently heat-treated to turn it blue. The mineral forms prismatic to dipyramidal crystals. Single crystals can reach a considerable size: specimens weighing up to 4½ lb (2 kg) and 8¾ lb (4 kg) have been found in Australia and Russia, respectively.

Zircon is found in metamorphic rocks and silica-rich igneous rocks. It resists weathering and, because of its relatively high specific gravity, concentrates in stream and river gravels and beach deposits.

Zircon bracelet Gem zircons, such as the colorless, faceted zircons in this bracelet, have been mined for over 2,000 years.

PROFILE

Orthorhombic

8

3.4–3.6

Perfect basal

Subconchoidal to uneven

Colorless

Vitreous

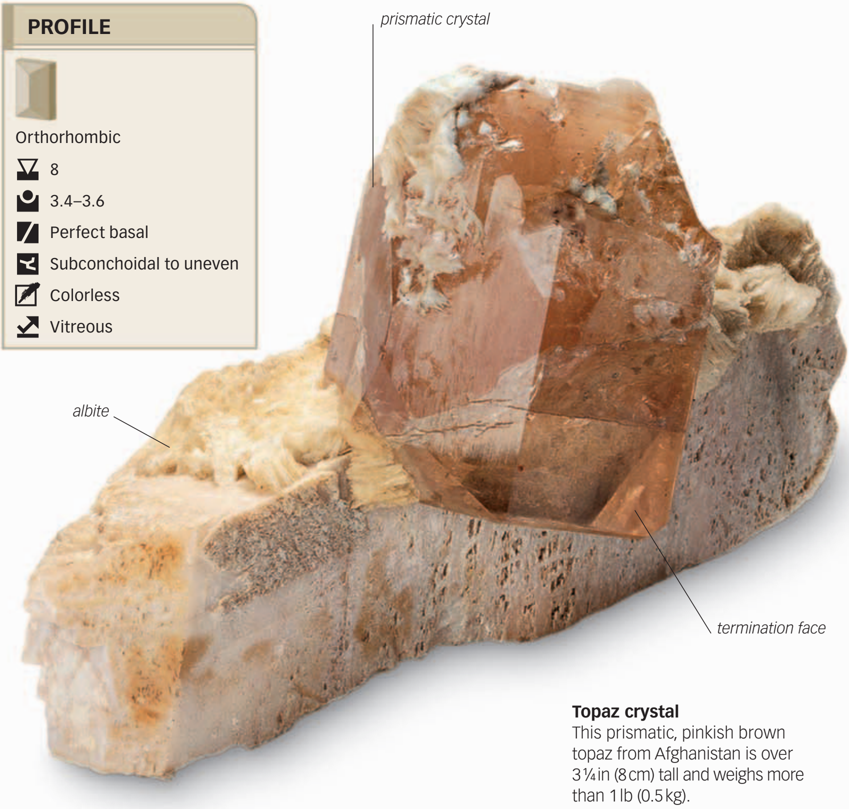

Topaz crystal
This prismatic, pinkish brown topaz from Afghanistan is over 3¼ in (8 cm) tall and weighs more than 1 lb (0.5 kg).

VARIANTS

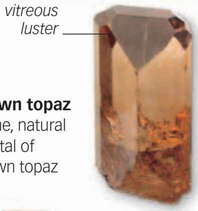

Brown topaz
A fine, natural crystal of brown topaz

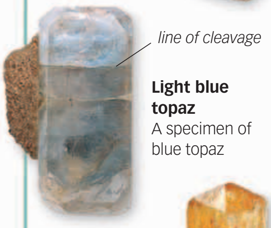

Light blue topaz
A specimen of blue topaz

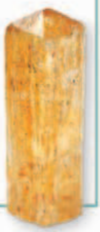

Imperial topaz
A golden imperial topaz from a deposit in Brazil

$Al_2SiO_4(F,OH)_2$

TOPAZ

The name topaz is thought to have been derived from the Sanskrit *tapaz*, which means "fire." Topaz occurs in a wide range of colors, with sherry-yellow and pink stones being particularly valuable. Colorless topaz has such a high refractive index that brilliant-cut specimens are mistaken for diamond (p.47). Well-formed prismatic crystals have a characteristic lozenge-shaped cross section and striations parallel to their length. The mineral is also found in massive, granular, and columnar habits.

Topaz is formed by fluorine-bearing vapors released in the late stages of crystallization of igneous rocks. It occurs in granites (pp.258–59), rhyolites (p.278), pegmatite dykes, and hydrothermal veins. Rounded pebbles are also found in river deposits. The world's largest faceted topaz weighs over 36,000 carats.

Pink topaz
A clear, octagonal step cut, pink topaz is set here in a gold ring. Natural pink topaz is rare.

Titanite crystals
These interpenetrating, wedge-shaped titanite crystals, 1½ in (3.5 cm) long, are from Russia.

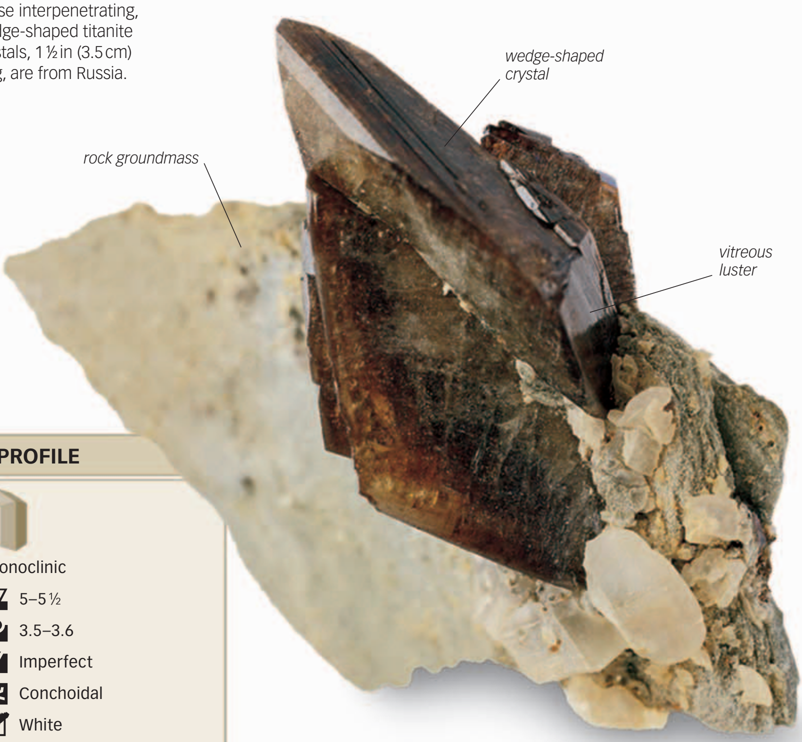

PROFILE

Monoclinic
5–5½
3.5–3.6
Imperfect
Conchoidal
White
Vitreous to greasy

VARIANT

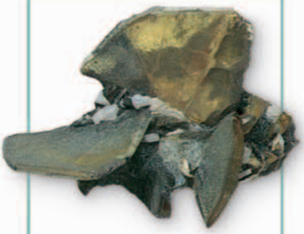

Crystal group Wedge-shaped titanite crystals

$CaTiSiO_5$

TITANITE

Formerly called sphene, titanite is a calcium titanium silicate. The name sphene originates from the Greek word *sphen*, which means "wedge"—a reference to the typical wedge-shaped crystals of the mineral. Crystals can also be prismatic. Gem-quality crystals occur in yellow, green, or brown colors. Specimens can also be black, pink, red, blue, or colorless. Titanite is strongly pleochroic, exhibiting different colors when seen from different directions. The mineral can also be massive, lamellar, or compact. Faceted titanite is one of the few stones with a color dispersion higher than that of diamond (p.47).

Titanite is widely distributed as a minor component of silica-rich igneous rocks and associated pegmatites (p.260). It is also found in the metamorphic rocks gneiss (p.288), schist (pp.291–92), and marble (p.301).

Titanite ring
Faceted titanites, such as the brilliant cut set in this gold ring, have superb fire and intense colors.

PROFILE

Orthorhombic

6½–7½

3.2

Good to perfect, poor

Conchoidal

White

Vitreous

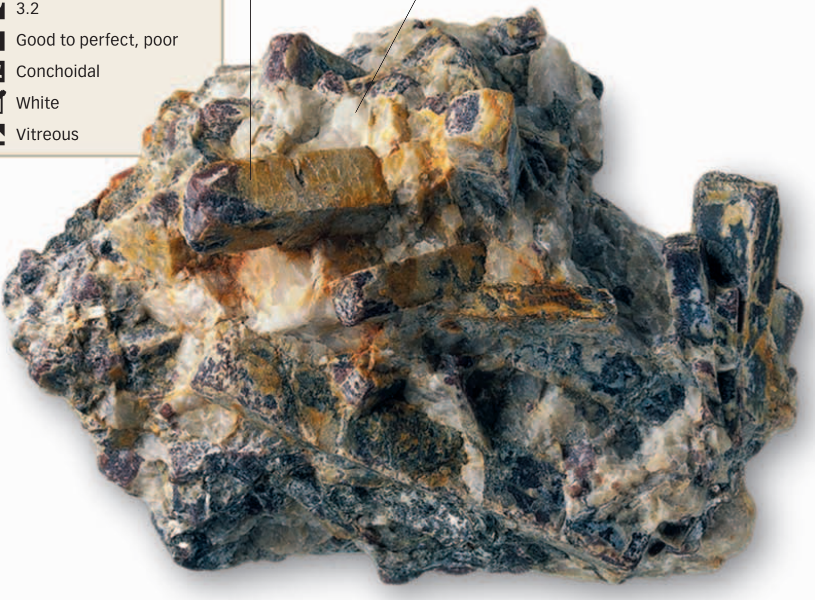

Andalusite crystals This group of prismatic andalusite crystals from the Austrian Tyrol is in a matrix of quartz.

VARIANTS

Brown andalusite Prismatic crystals on a rock matrix

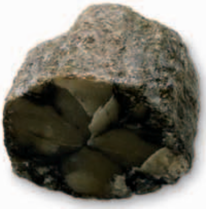

Chiastolite A yellowish brown andalusite crystal with cross-shaped inclusions of carbon

Al_2OSiO_5

ANDALUSITE

Named after the locality in Andalusia, Spain, where it was first described, andalusite is aluminum silicate. It is pink to reddish brown, white, gray, violet, yellow, green, or blue. Gem-quality andalusite exhibits yellow, green, and red colors when viewed from different directions. Andalusite crystals are commonly prismatic with a square cross section. They can also be elongated or tapered. Andalusite can also occur in massive form. A yellowish gray variety called chiastolite occurs as long prisms enclosing symmetrical wedges of carbon-rich material.

Andalusite is found in regional and low-grade metamorphic rocks, where it is associated with corundum (p.95), cordierite (p.223), sillimanite (p.237), and kyanite (p.238). It is rarely found in granites (pp.258–59) and granitic pegmatites (p.260) .

Rectangular step cut Relatively uncommon, transparent andalusite is too brittle to be worn. It is faceted for gem collectors.

PROFILE

Orthorhombic

7

3.2–3.3

Perfect

Uneven

White

Silky

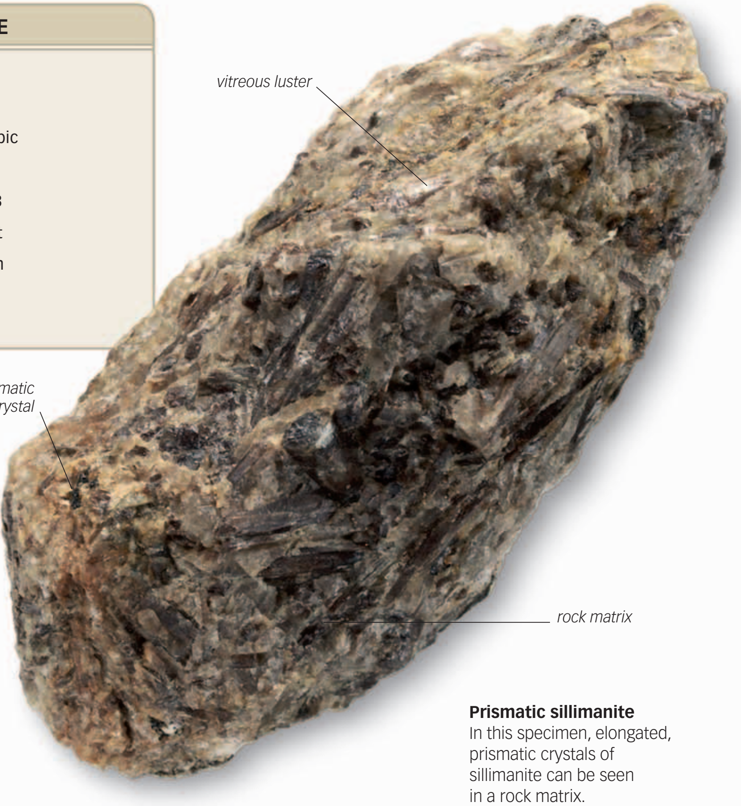

Prismatic sillimanite In this specimen, elongated, prismatic crystals of sillimanite can be seen in a rock matrix.

VARIANTS

Crystals in rock Sillimanite in a rock matrix

silky luster

Fibrous sillimanite A mass of parallel, fibrous crystals

Al_2OSiO_5

SILLIMANITE

Named after the American chemist Benjamin Silliman, sillimanite is one of three polymorphs of aluminum silicate. Commonly colorless to white, sillimanite can also be pale yellow to brown, pale blue, green, or violet. A single specimen may appear yellowish green, dark green, or blue when seen from different angles. The mineral occurs in long, slender, glassy crystals or in blocky, poorly terminated prisms.

Sillimanite is characteristic of clay-rich metamorphic rocks formed at high temperatures (1,065°F/575°C or above). The mineral is often found with corundum (p.95), cordierite (p.223), and kyanite (p.238). Specimens also occur in gneisses (p.288), sillimanite schists, hornfels (p.303), and detrital sediments.

Collectors' gem Facet-grade sillimanite, such as this specimen, occurs in the gem gravels of Sri Lanka and Myanmar, and in Brazil.

Blady kyanite
This specimen of kyanite with quartz from northern Brazil shows the characteristic elongated, bladed habit of kyanite crystals.

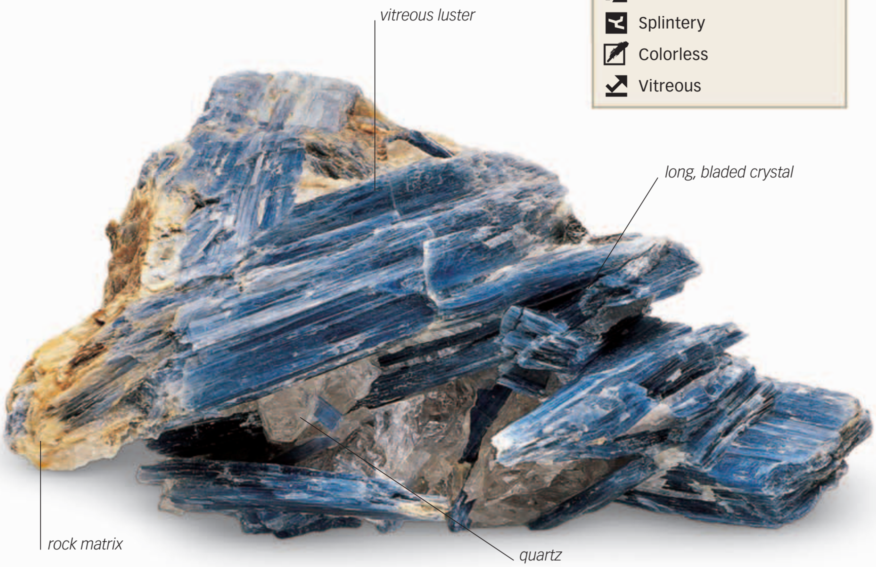

PROFILE

Triclinic

4½; 6

3.6

Perfect

Splintery

Colorless

Vitreous

VARIANT

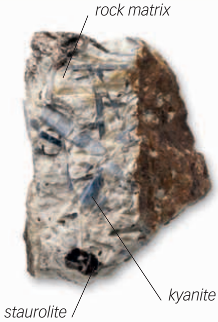

Kyanite in rock Kyanite crystals with staurolite in a rock matrix

Al_2SiO_5

KYANITE

Named after the Greek word *kyanos*, which means "dark blue," kyanite is blue and blue-gray, the colors generally zoned within a single crystal. Kyanite can also be green, orange, or colorless. Specimens have variable hardness: about 4½ when scratched parallel to the long axis but 6 when scratched perpendicular to the long axis. Kyanite occurs mainly as elongated, flattened blades that are often bent and sometimes as radiating, columnar aggregates.

Kyanite is formed during the regional metamorphism of clay-rich sediments. It occurs in mica schists, gneisses (p.288), and associated hydrothermal quartz veins and pegmatites (p.260). It is used to estimate the temperature, depth, and pressure at which a rock has metamorphosed.

Spark plugs
Kyanite is mined for the aluminum silicate mullite, which is used in spark plugs.

PROFILE

Monoclinic

7–7½

3.7

Distinct

Conchoidal

Colorless to gray

Vitreous to resinous

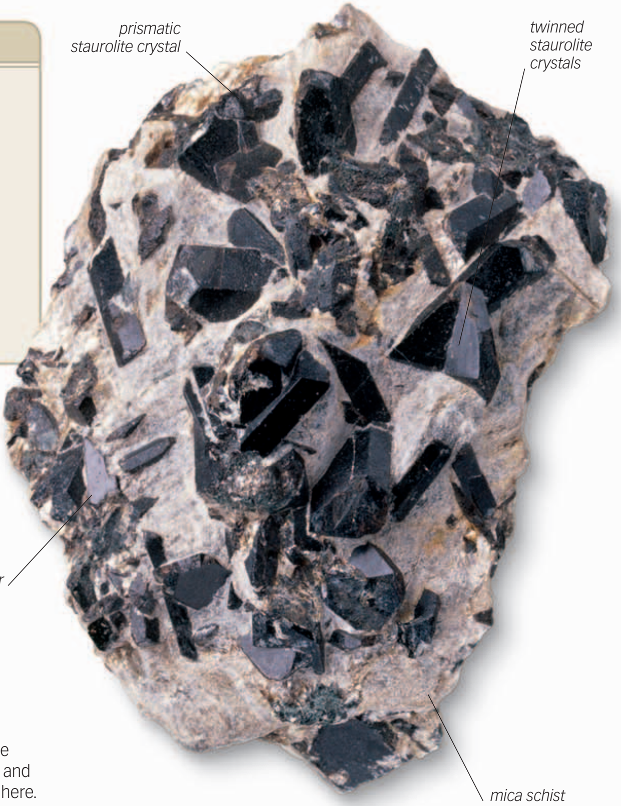

Staurolite crystals This is a specimen of staurolite in a mica schist matrix. Single and twinned crystals can be seen here.

VARIANTS

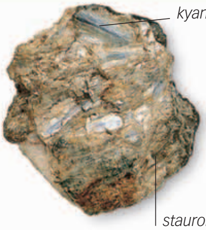

Staurolite in schist Kyanite and staurolite in schist

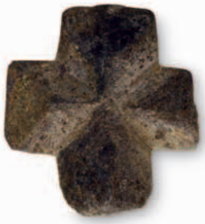

Fairy cross A twinned staurolite, or "fairy cross," crystal

$(Fe,Mg)_4Al_{17}(Si,Al)_8O_{45}(OH)_3$

STAUROLITE

A widespread mineral, staurolite takes its name from two Greek words: *stauros*, which means "cross," and *lithos*, which means "stone"—a reference to its typical crosslike twinned form. Cross-shaped penetration twins of the mineral are common and are in great demand as charms. Staurolite specimens are yellowish brown, reddish brown, or nearly black in color. The mineral normally occurs as prismatic crystals, which are either hexagonal or diamond-shaped in section and often have rough surfaces.

Staurolite occurs in mica schists, gneisses (p.288), and other metamorphosed, aluminum-rich rocks. It forms only under a specific range of pressure and temperature, which helps determine the various conditions under which the metamorphic rock formed.

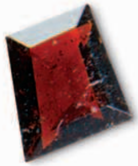

Trapeze-cut staurolite Transparent staurolite, as in this stone, is a rare faceting material because of its dark color and lack of brilliance.

PROFILE

Monoclinic
7½
3.0
Perfect
Conchoidal, brittle
White
Vitreous

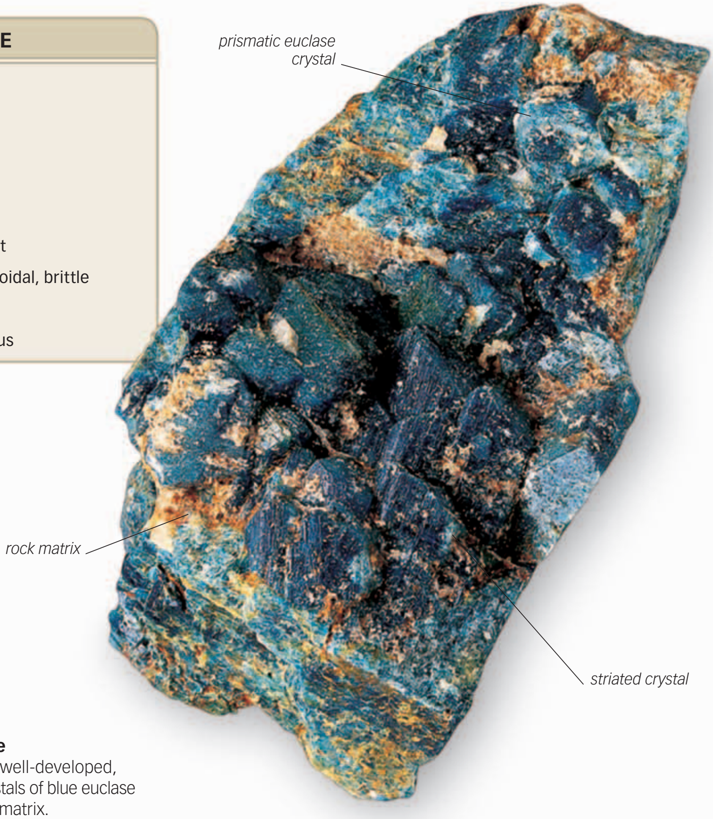

Blue euclase
This mass of well-developed, prismatic crystals of blue euclase is on a rocky matrix.

VARIANT

Transparent euclase
A near-transparent, striated euclase crystal

$BeAlSiO_4(OH)$

EUCLASE

Euclase takes its name from two Greek words: *eu*, which means "good," and *klasis*, which means "fracture"—a reference to its perfect cleavage. Generally white or colorless, euclase can also be pale green or pale to deep blue—a color for which it is particularly noted. It forms striated prisms, often with complex terminations. Massive and fibrous specimens are also found.

Euclase occurs in hydrothermal veins formed at low temperatures (up to 400°F/200°C), granitic pegmatites (p.260), and some metamorphic schists (pp.291–92) and phyllites (p.294). It is also found in stream gravels. Exquisite, colorless, and deep blue color-zoned crystals come from Karoi in Zimbabwe. Cut euclase resembles certain types of beryl (p.225) and topaz (p.234).

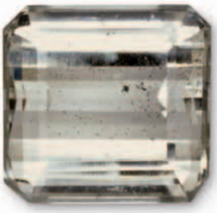

Euclase gemstone
This square-cut euclase gemstone shows small, dark inclusions of another mineral.

PROFILE

Orthorhombic
6–6½
3.2–3.3
Poor
Subconchoidal to uneven
Yellow to orange
Vitreous

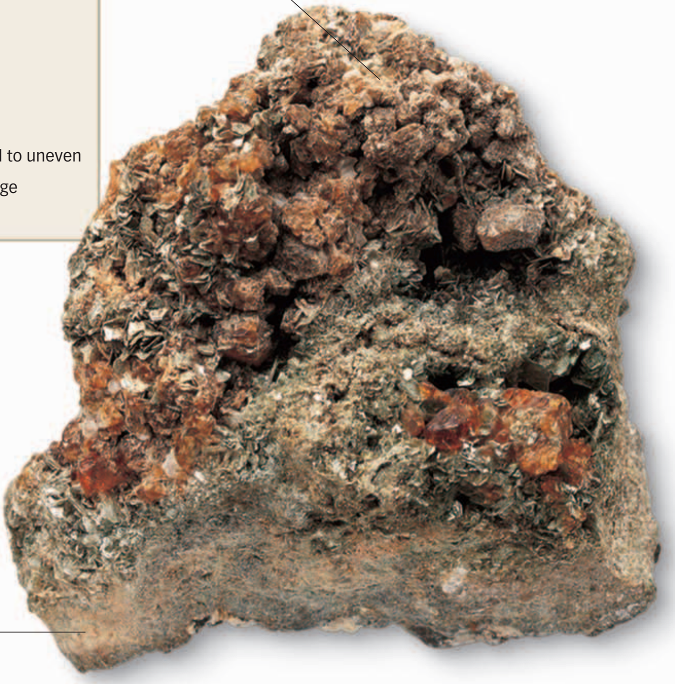

Humite crust In this specimen, a crust of yellowish brown humite crystals covers a rock matrix that also contains accessory mica.

VARIANTS

Yellow humite A specimen of massive, yellow humite

Orange humite Massive, orange humite with a brown weathering crust

$(Mg,Fe)_7(SiO_4)_3(F,OH)_2$

HUMITE

Named in 1813 after the English mineral and art collector Sir Abraham Hume (1749–1838), humite is a silicate of iron and magnesium. Manganese substitutes for iron in the structure to form a complete solid-solution series with manganhumite. Specimens are yellow to dark orange or reddish orange in color, tending toward brown with increasing manganese content. Humite is generally found in granular masses. Well-formed crystals are rare and grow in parallel with one another. Crystals rarely exceed ⅜ in (1 cm) in length and are occasionally twinned.

Humite is found with pyrite (p.62), cassiterite (p.79), hematite (p.91), quartz (p.168), tourmaline (p.224), and mica in contact and regionally metamorphosed limestones (p.319) and dolomites (p.320). Although this mineral occurs worldwide, noteworthy locations include Persberg and elsewhere in Sweden; Isle of Skye, Scotland; Mount Vesuvius, Italy; Valais, Switzerland; and Brewster, New York, USA.

PROFILE

Cubic

7–7½

3.6

None

Conchoidal, brittle

White

Vitreous

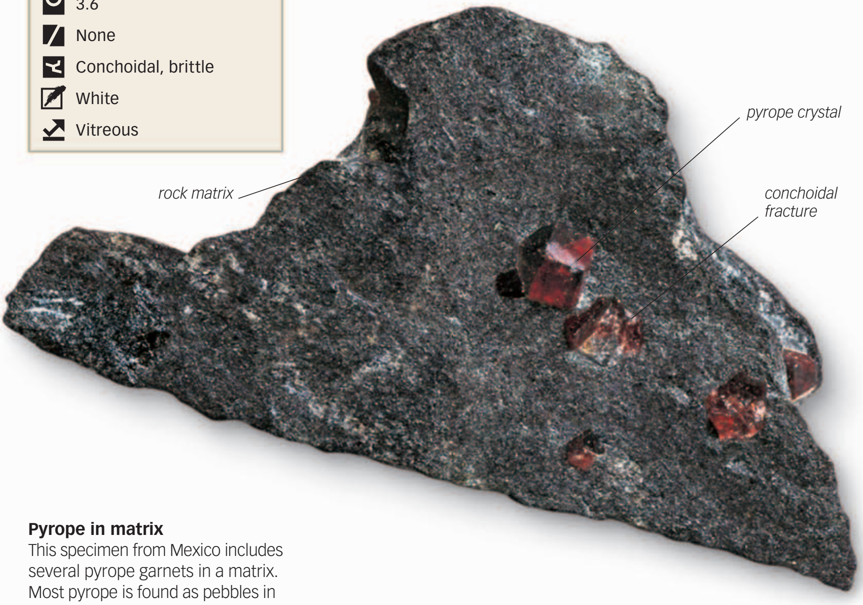

Pyrope in matrix This specimen from Mexico includes several pyrope garnets in a matrix. Most pyrope is found as pebbles in placer deposits with other gems.

VARIANT

Gemstone rough Water-rounded pyrope recovered from a placer deposit

$Mg_3Al_2(SiO_4)_3$

PYROPE

The magnesium aluminum garnet pyrope was named in 1803 after the Greek words *pyr* and *õps*, which mean "fire" and "eye" respectively—a reference to the typical fiery color of specimens. Manganese, chromium, iron, and titanium substitute in the mineral's structure and act as coloring agents to some degree. Specimens can be rich red, dark red, violet-red, rose red, or reddish orange depending on the composition. Crystals are dodecahedral and trapezohedral, with hexoctahedra sometimes present. The mineral is most often found as rounded grains or pebbles.

Pyrope occurs as a high-pressure mineral in metamorphic rocks. It is also found in high-pressure, silica-poor igneous rocks and in detrital deposits derived from them.

Pyrope gemstones Beautiful garnet jewelry comes from Bohemia, Czech Republic, where pyropes as big as hens' eggs are found.

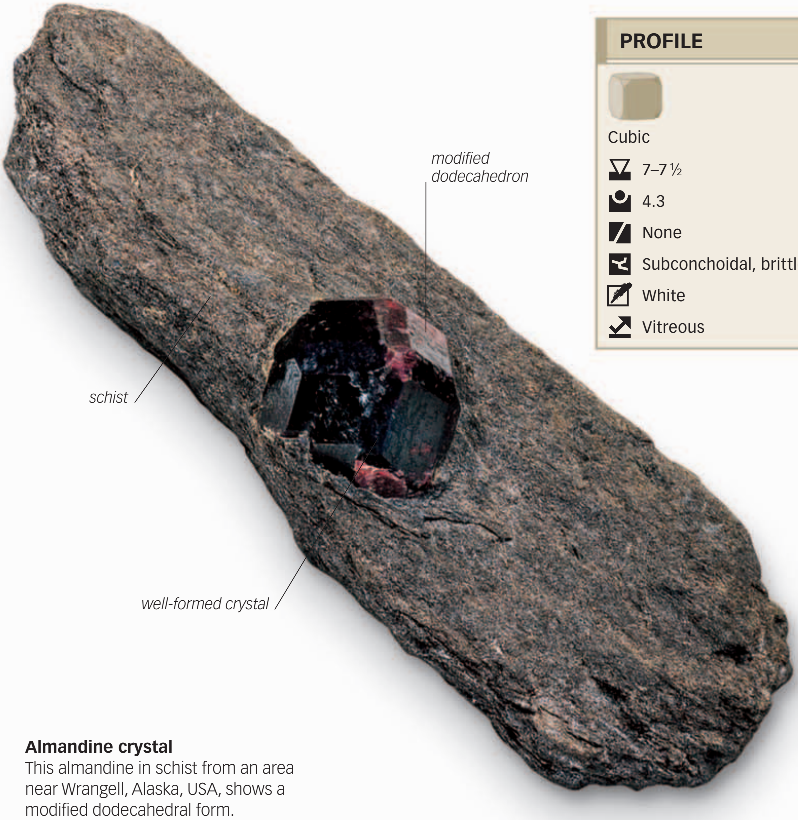

Almandine crystal
This almandine in schist from an area near Wrangell, Alaska, USA, shows a modified dodecahedral form.

PROFILE

Cubic
7–7 ½
4.3
None
Subconchoidal, brittle
White
Vitreous

VARIANTS

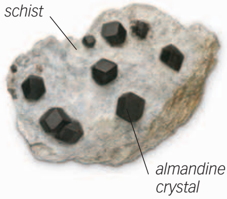

Crystals in schist Numerous almandine dodecahedrons rest on schist

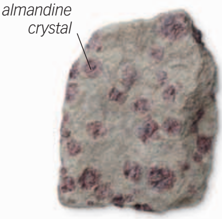

Almandine in granulite
Crystals of almandine in a granulite matrix

$Fe_3Al_2(SiO_4)_3$

ALMANDINE

The most common garnet, almandine is named after Alabanda (now Araphisar) in Turkey, where it has been cut since antiquity. Almandine is always red, often with a pink or violet tinge. Specimens can sometimes be nearly black. This mineral tends to be a pinker red than other garnets. Crystals often have well-developed faces and are dodecahedral or trapezohedral or have other more complex forms. Massive aggregates and rounded grains are also found. Rutile (p.78) needles can show as a four-rayed star when almandine is cut *en cabochon*.

Almandine occurs worldwide. It is found in gneisses (p.288) and mica schists, igneous rocks (pp.256–57), and occasionally as inclusions in diamond (p.47). When it occurs in metamorphic rocks, its presence indicates the grade of metamorphism.

Faceted almandine
Three faceted almandine gems and a seed pearl create a central flower motif in this antique gold brooch.

Spessartine crystals
In this specimen from Norway, well-formed dodecahedral crystals encrust a rock matrix.

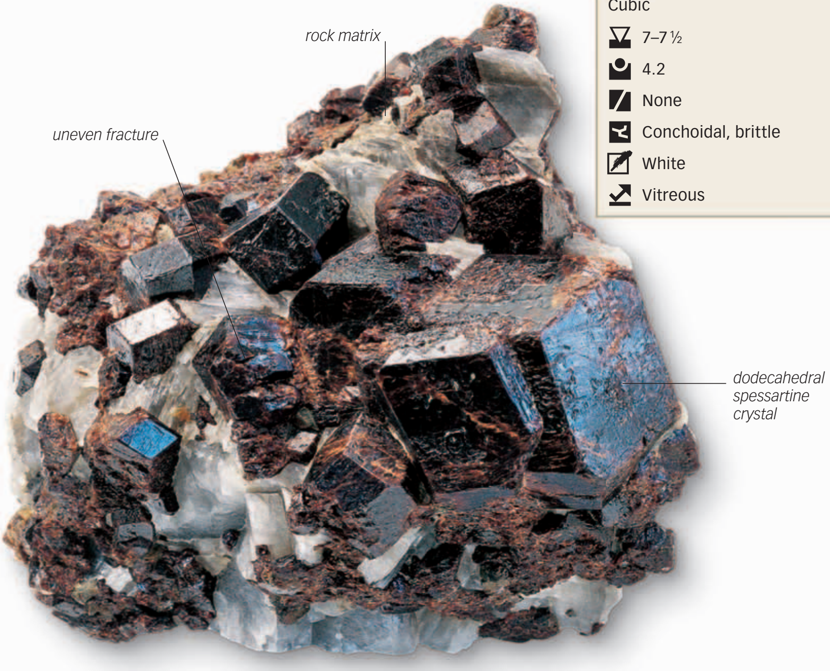

PROFILE

Cubic
7–7 ½
4.2
None
Conchoidal, brittle
White
Vitreous

VARIANT

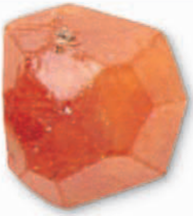
Translucent spessartine
An attractive crystal of translucent spessartine

$Mn_3Al_2(SiO_4)_3$

SPESSARTINE

The manganese aluminum silicate spessartine is named after Spessart—the locality in Germany where it was first described. The mineral is pale yellow when nearly pure and orange to deep red, brown, or black in other specimens. A color change known as the alexandrite effect is occasionally found in spessartine grossular garnet. Crystals are dodecahedral or trapezohedral. Spessartine may also occur as either granular or massive aggregates.

Spessartine almost always contains some amount of iron (p.39). Pure spessartine is relatively rare and is found in manganese-rich metamorphic rocks, granites (pp.258–59), and pegmatite veins (p.260). The heaviest spessartine ever discovered weighs 6,720 carats.

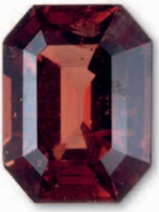
Octagonal step cut
Because of spessartine's rich color, the liquid inclusions under the edge facets in this gem are not very noticeable.

PROFILE

Cubic
- 6½–7
- 3.6
- None
- Conchoidal
- White
- Vitreous

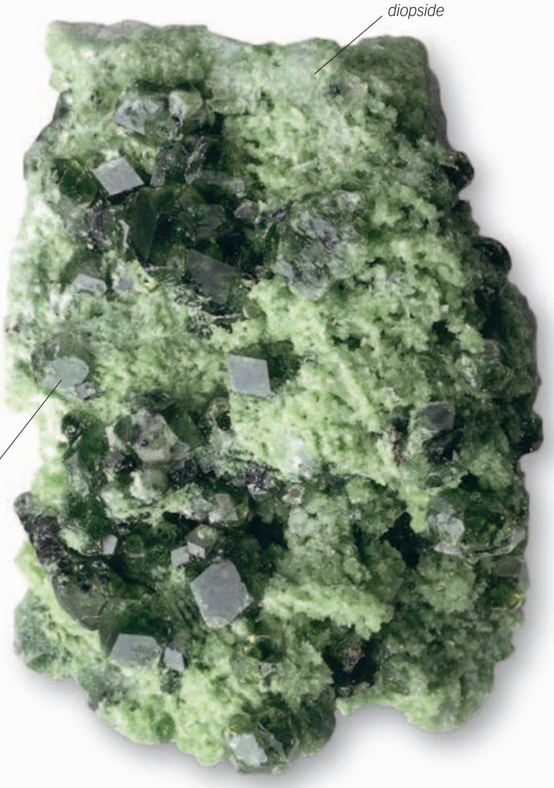

Grossular on diopside These grossular crystals from Piedmont, Italy, are set on a matrix of diopside.

$Ca_3Al_2(SiO_4)_3$

GROSSULAR

The calcium aluminum silicate grossular, a type of garnet, is named after the Latin word *grossularia*, which translates into "gooseberry"—a reference to the mineral's gooseberry-green color. Specimens can also be pale to emerald green, white, colorless, cream, orange, red, honey, brown, or black. Reddish brown or pink grossular is called hessonite or cinnamon stone. Grossular is usually translucent to opaque but can be transparent. It occurs as rounded dodecahedral or trapezohedral crystals that are up to 5 in (13 cm) wide. Specimens can also be granular or massive.

Grossular forms in impure calcareous rocks that have undergone regional or contact metamorphism, in some schists (pp.291–92) and serpentinites (p.298), and occasionally in meteorites (pp.335–37).

Grossular beads This strung group is made up of size-graded, round, luminescent green grossular beads.

VARIANTS

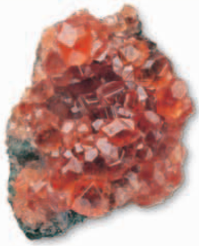

Hessonite Reddish brown dodecahedral grossular crystals

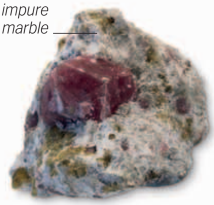

Pink grossular Rounded, pink grossular crystals in marble

ORGANICS

Generated by organic (biological) processes, the organics may or may not be crystalline. In some cases, they contain the same mineral matter—such as calcite or aragonite—as that generated through inorganic processes. Organics are sometimes used as gems.

COMPOSITION

The organics have a highly varied composition. Coral, pearl, and shell contain mineral matter generated by biological processes. Amber is fossilized resin, mainly from extinct coniferous trees, although amberlike substances from even older trees are also known. Copal is a modern equivalent of fossilized amber. Coal is derived from buried organic material, such as peat. Bitumen is a very heavy oil.

Shell
The shells of marine invertebrates capture large amounts of carbon in the form of aragonite. Their remains form extensive beds of carbonate rocks.

OCCURRENCE

The organics are relatively widespread. The shells of freshwater and marine organisms are part of a carbonate cycle that extracts carbon from the environment and returns it as carbonate sediment, either to be reincorporated into other organics or to be lithified. Other organics incorporate carbon in their essential composition.

USES

Coal and bitumen are the organics that find the widest use. Organically formed carbonate rocks are also extensively used as building stone and ballast and in the manufacture of cement. Other organics are used to make ornaments and jewelry.

Coral necklace
This branchlike Native American necklace is made from small, polished branches and tiny beads of red coral.

Coral reef
A coral reef forms in shallow ocean areas. Corals are the most important part of the reef and form its main structural framework. Coral skeletons are made of aragonite.

PROFILE

Trigonal Orthorhombic

Amorphous

- 3½
- 2.6–2.7
- None
- Hackly
- White
- Dull to vitreous

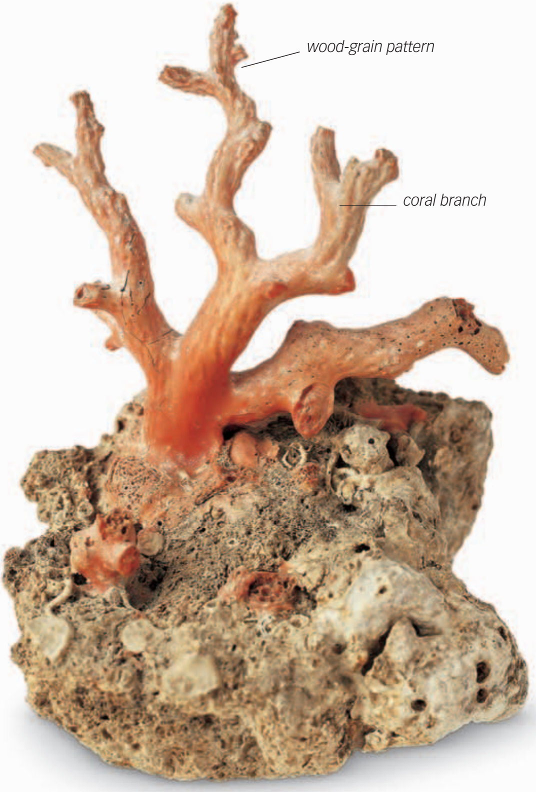

Red coral
The use of red coral dates back to the Iron Age. This specimen from the Mediterranean has a wood-grain pattern on its branches.

VARIANTS

Blue coral
A type of coral used in artifacts and jewelry

Black coral
A variant that is polished to make jewelry

Brain coral
An elaborate confection of organic aragonite

Mostly $CaCO_3$

CORAL

According to Greek legend, coral came from the drops of blood shed when the mythic hero Perseus cut off the head of the monster Medusa. Coral is actually the skeletal material generated by marine animals also known as coral polyps. In most corals, this material is calcium carbonate, but in black and golden corals it is a hornlike substance called conchiolin. Coral has a dull luster when recovered, but it can be polished and brightened. It is sensitive even to mild acids and can become dull with extensive wear.

Coral is variable in color. Red and pink precious coral is found in the warm seas around Japan and Malaysia, in African coastal waters, and in the Mediterranean Sea. Black coral comes from the West Indies, Australia, and from around the Pacific Islands.

Coral necklace
This triple-stranded necklace from Morocco is made of coral, silver, and turquoise.

PROFILE

Orthorhombic
3
2.7
None
Uneven, brittle
White
Pearly

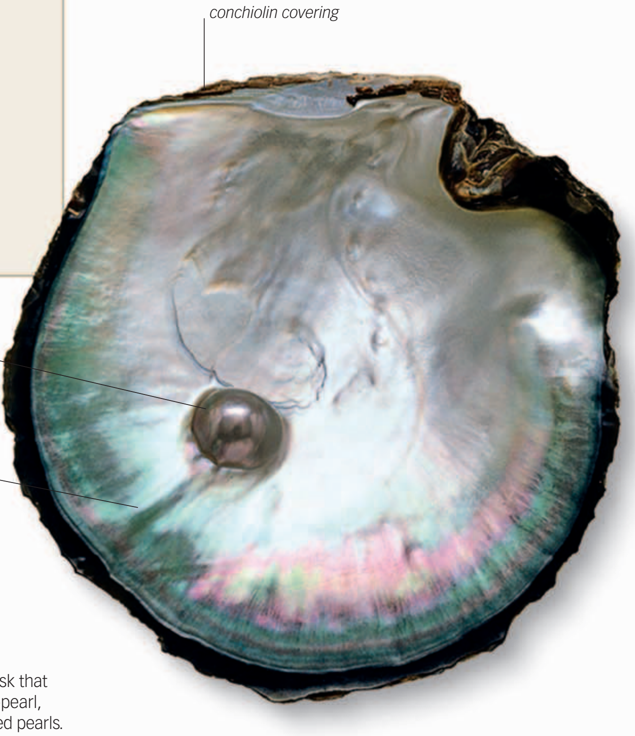

Black mother-of-pearl The black-lip shell, a mollusk that produces black mother-of-pearl, also produces black cultured pearls.

VARIANTS

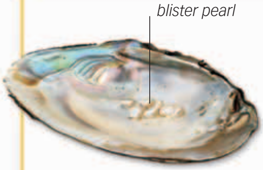

Blister pearls Attached to the shell, these are flat on one side

Freshwater pearls These have the same luster as marine pearls

Mostly $CaCO_3$

PEARL

A concretion formed by a mollusk, a pearl consists mainly of aragonite (p.115), the same material as the animal's shell (p.249). The shell-secreting cells are located in a layer of the mollusk's body tissue called the mantle. When a foreign particle enters the mantle, the cells build up concentric layers of pearl around it. Colors vary with the mollusk and its environment and can be any delicate shade from black to white, cream, gray, blue, yellow, green, lavender, or mauve.

The finest pearls are produced by limited species of saltwater oysters and freshwater clams. A pearl is valued by its translucence, luster, color, and shape. The most valuable pearls are spherical or droplike, with a deep luster and good play of color. Rose-tinted Indian pearls are highly prized.

Pearl bracelet This Cartier Art Deco bracelet has five strands of cultured pearls with a gold and oxidized-silver clasp.

PROFILE

Hexagonal Orthorhombic

Amorphous

2½

About 1.3

None

Conchoidal

White

Dull to vitreous

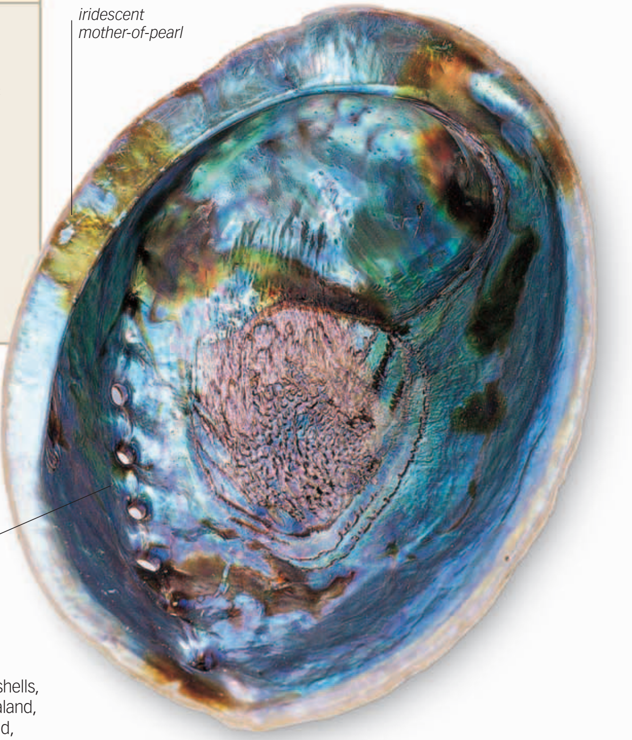

Abalone shell
Found in warm seas, abalone shells, such as this one from New Zealand, are noted for their multicolored, iridescent, mother-of-pearl lining.

VARIANTS

Spider conch A type of shell widely used for ornamental purposes

Tortoise shell The shell of a hawksbill turtle

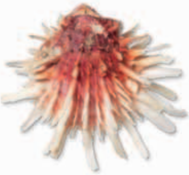
Thorny oyster A shell made up of organic aragonite

$CaCO_3$

SHELL

Like coral (p.247), shell is mineral matter generated by biological processes. The mineral component of shell is either calcite (p.114) or aragonite (p.115), both of which are calcium carbonate. Shell forms as the hard outer covering of many mollusks. It is secreted in calcareous layers by cells in the mantle—a skinlike tissue in the mollusk's body wall. Mollusk shells differ in the composition, number, and arrangement of calcareous layers. These layers form as distinct microstructures with different mechanical properties and, in some shells, different colors.

Both marine and freshwater shells are used for ornamentation. Shells with different colored layers have been carved into cameos since antiquity. Pearly shells were used widely in button making around the 19th century.

Shell perfume bottle
This 19th-century perfume bottle is made of two shells glued together. It has a chain, ring, and pinchbeck stopper.

PROFILE

Crystal system None
2–2½
About 1.1
None
Conchoidal
White
Resinous

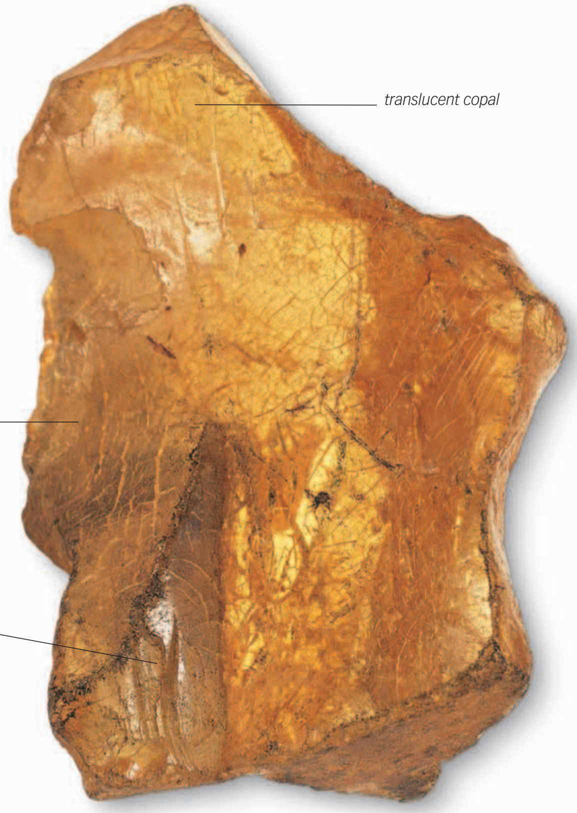

Copal nugget This specimen of copal closely resembles amber. Some copal is used as an amber substitute in jewelry.

VARIANTS

orange-yellow gum

Kauri gum Resin derived from the Kauri conifer

Copal lumps Lumps of resin from the Protium copal or copal tree

Various

COPAL

Named after the Nahuatl word *copalli*, which means "resin," copal is a yellow to red-orange resin obtained from various tropical trees. It can be collected from living trees and from accumulations in the soil beneath the trees. It can also be mined if it is buried. Copals from different sources usually have different chemical properties but can have similar physical properties. Copal has approximately the same hardness as amber (p.251) but unlike amber it is wholly or partially soluble in organic solvents. Buried copal is the nearest to amber in durability and is, in many cases, virtually indistinguishable from it.

Zanzibar in Tanzania is a major source of buried copal. This mineral is also found in China, Brazil, and other South American countries. It is used in varnishes, lacquers, and inks.

Copal beads In this necklace, beads of a tough and compact form of copal alternate with beads carved from seeds.

PROFILE

Crystal system None

2–2½

1.1

None

Conchoidal

White

Resinous

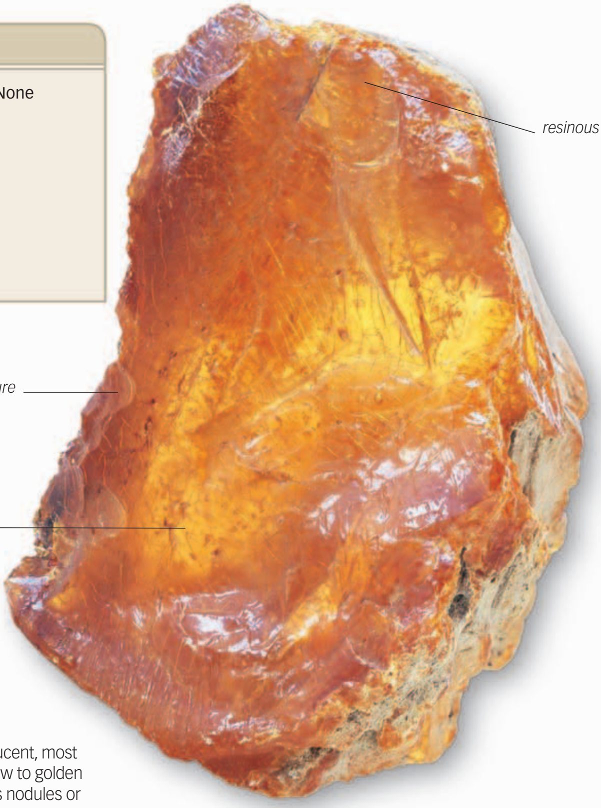

Amber nodule
Transparent to translucent, most amber is golden yellow to golden orange and occurs as nodules or small, irregularly shaped masses.

VARIANTS

Wave-rounded amber
A piece of wave-rounded Baltic amber

Lithuanian amber A group of amber pieces showing color variation

Hydrocarbon (C,H,O)

AMBER

The fossilized resin amber comes mainly from extinct coniferous trees, although amberlike substances from earlier trees are also known. It is generally found in association with lignite coal (p.253). Amber and other partially fossilized resins are sometimes given mineral-like names depending on where they are found, their degree of fossilization, or the presence of other chemical components. For example, resin that resembles copal (p.250) and comes from the London clay region is called copalite. At least 12 other names are applied to minor variants.

For thousands of years, the largest source of amber has been deposits along the Baltic coast, extending intermittently from Gdánsk right around to the coastlines of Denmark and Sweden. Amber has been traded since ancient times.

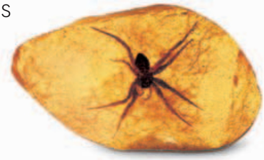

Preserved in amber
As resin dried 40–50 million years ago, insects were sometimes fossilized within the sticky substance.

PROFILE

Crystal system None

2½

About 1.3

None

Conchoidal

Black to dark brown

Velvety, vitreous, or waxy

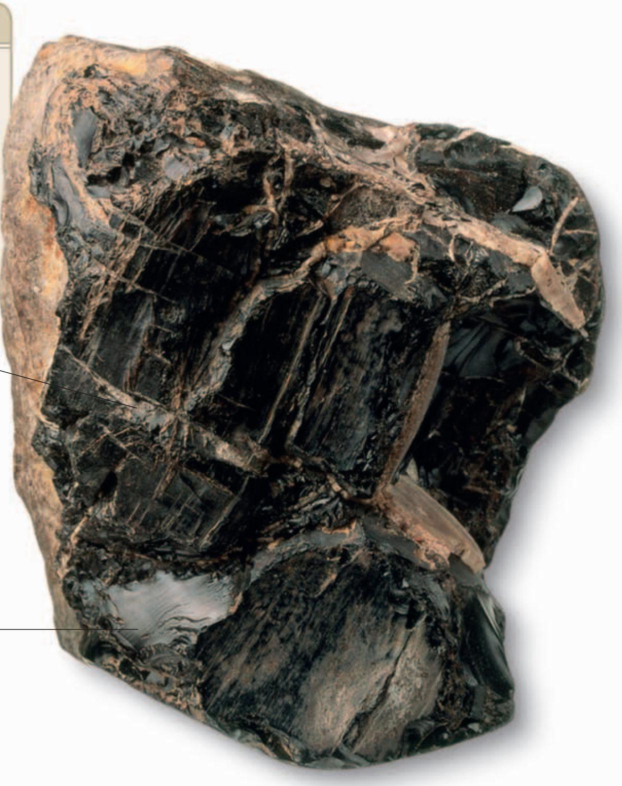

Woody structure This specimen of jet shows the layered, woody structure that is sometimes characteristic of the mineral.

VARIANT

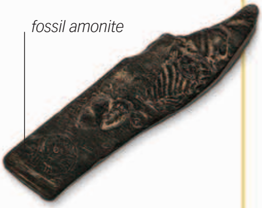

Jet fossil Jet with fossils of marine origin

Various

JET

Generally classified as a type of coal (p.253), jet has a high carbon content and a layered structure. It is black to dark brown in color. Specimens sometimes contain tiny inclusions of pyrite (p.62), which have a metallic luster. Jet is found in rocks of marine origin, perhaps derived from waterlogged driftwood or other plant material. It can occur in distinct beds, such as those at Whitby, England, from which jet has been extracted since the 1st century CE.

Jet has been carved for ornamental purposes since prehistoric times and has been found in prehistoric caves. The Romans carved jet into bangles and beads. In medieval times, powdered jet drunk with water or wine was believed to have medicinal properties.

Jet necklace This Native American necklace is made of high-quality, fine-grained jet, which shows velvety luster.

PROFILE

Crystal system None
2–2½
About 1.1
None
Conchoidal
Black
Nearly metallic

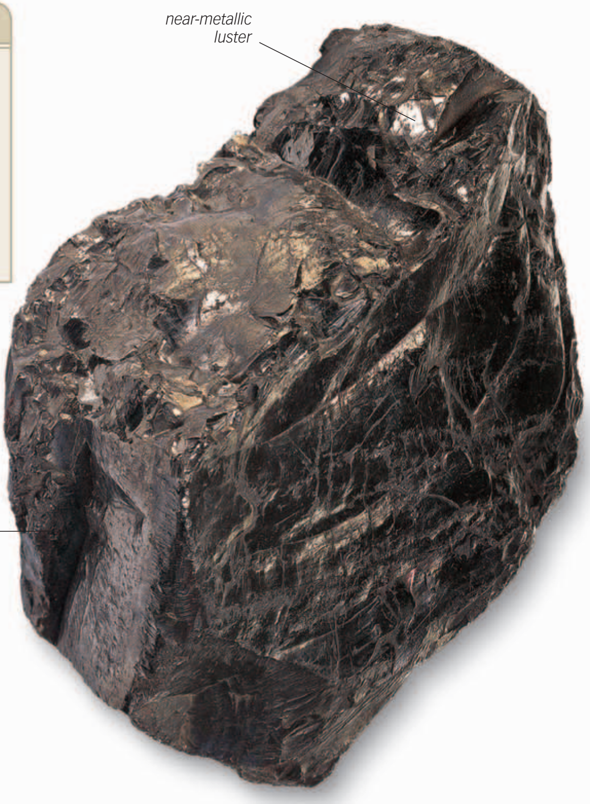

Anthracite
Hard to the touch, anthracite is naturally shiny. It takes a brilliant polish and is used for decorative as well as practical purposes.

VARIANTS

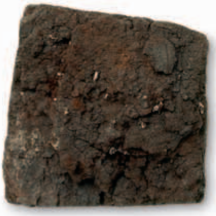

Lignite A variant having a composition between peat and bituminous coal

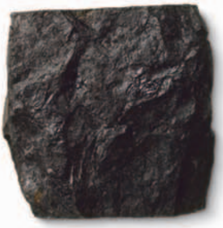

Bituminous coal The most common form of coal

Various

COAL

The fossilized remains of plants, coal usually occurs in layered, sedimentary deposits. It is brown or black and made up of an irregular mixture of chemical compounds called macerals, which are analogous to minerals in inorganic rocks. Unlike minerals, macerals have no fixed chemical composition and no definite crystalline structure.

Different varieties of coal are formed depending on the kinds of plant material, varying degrees of coalification (the process by which plant material is converted to coal), and the presence of impurities. Four varieties are recognized. Lignite is the lowest grade and is the softest and least coalified variety. It forms from peat under moderate pressure. Sub-bituminous coal is dark brown to black. Bituminous coal is the most abundant and is commonly burned for heat generation. Anthracite is the highest grade and the most highly metamorphosed form of coal. It contains the highest percentage of low-emission carbon and would be an ideal fuel if it were not relatively scarce.

ROCKS

IGNEOUS ROCKS

Igneous rocks are formed from magma—molten rock. They are classified as extrusive or intrusive depending on whether or not the magma emerged at Earth's surface before crystallizing. Extrusive rocks form on the surface; intrusive rocks form below it.

INTRUSIVE IGNEOUS ROCKS

Intrusive rocks are categorized as plutonic if formed deep inside the crust and hypabyssal if formed at shallow depths. Plutonic intrusive rocks are characterized by their large crystals and generally form geographically large bodies. For example, a batholith is a large igneous body with a surface exposure of at least 40 square miles (100 square km) and a thickness of about 6–9 miles (10–15 km). Batholiths form the cores of great mountain ranges, such as the Rockies and the Sierra Nevada in North America. Granite, diorite, peridotite, syenite, and gabbro are all plutonic igneous rocks.

Hypabyssal intrusive rocks are formed at shallower depths and are characterized by fine crystallization. They occur in sheetlike bodies called dikes and sills, volcanic plugs, and other relatively small formations. Dikes range from less than an inch to many feet in thickness and can be hundreds of miles in length. Sills are similar to dikes, except that they form parallel to the enclosing rocks and intrude between two strata.

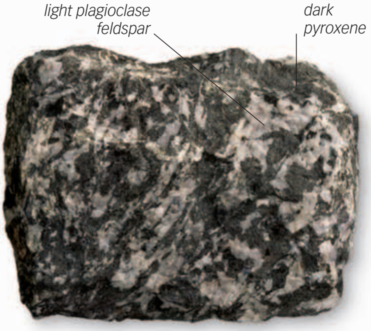
light plagioclase feldspar
dark pyroxene

Gabbro
Most gabbros form as plutonic intrusive rocks. This coarsely crystallized specimen contains light-colored crystals of plagioclase and dark crystals of pyroxene.

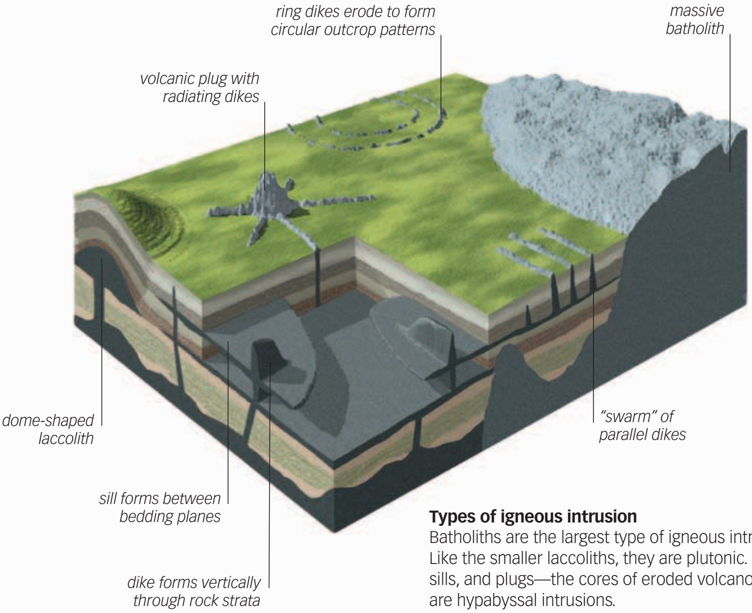
ring dikes erode to form circular outcrop patterns
massive batholith
volcanic plug with radiating dikes
dome-shaped laccolith
"swarm" of parallel dikes
sill forms between bedding planes
dike forms vertically through rock strata

Types of igneous intrusion
Batholiths are the largest type of igneous intrusion. Like the smaller laccoliths, they are plutonic. Dikes, sills, and plugs—the cores of eroded volcanoes—are hypabyssal intrusions.

EXTRUSIVE IGNEOUS ROCKS

Extrusive igneous rocks are also known as volcanic rocks. The principal rock types in this category include basalt, obsidian, rhyolite, trachyte, and andesite. All of these usually form from lava—a magma that has flowed either onto land or underwater. Other extrusive rocks, such as tuff and pumice, form in explosive volcanic eruptions. These "pyroclastic" rocks are porous because of the frothing expansion of volcanic gases during their formation. Basalt is the most common extrusive rock, forming the floor of most oceans and extensive plateaus on land, such as the Deccan Plateau of India and the Columbia River Basalts of Oregon, USA.

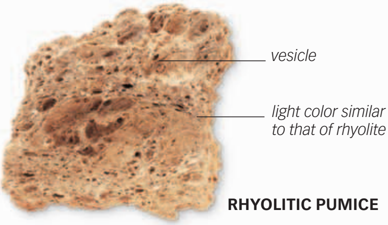

RHYOLITIC PUMICE

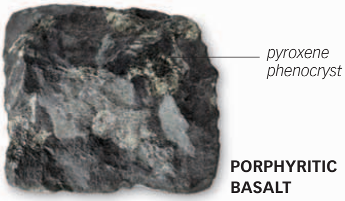

PORPHYRITIC BASALT

Grain size
The origin of igneous rocks is generally indicated by the grain size of their minerals. Small or microscopic grains are found in extrusives; large grains in intrusives.

COMPOSITION OF IGNEOUS ROCKS

Igneous rocks form from magma, which is essentially a silicate melt. Igneous rocks are classified on the basis of silica content. Felsic rocks have over 65 percent silica, intermediate rocks 55–65 percent, mafic rocks 45–55 percent, and ultramafic rocks less than 45 percent silica. The silicate minerals that develop from the melt depend on factors such as silica concentration in the melt, the presence and concentration of other elements such as aluminum, iron, magnesium, calcium, sodium, and potassium within the melt, and the temperature and pressure at which crystallization takes place.

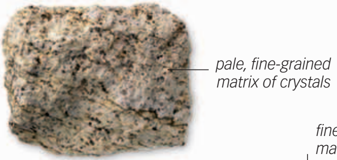

LIGHT COLOR

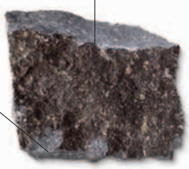

MEDIUM COLOR

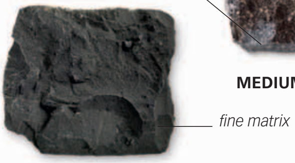

DARK COLOR

Color
Relatively few igneous rocks are identified by specific colors. Instead, they are generally described as light, intermediate, or dark.

Mount St. Helens
In 1980, Mount St. Helens in the USA erupted, sending thousands of tons of pyroclastic debris across northwestern USA. Extensive beds of pumice and tuff form from such eruptions.

PROFILE

- Silica-rich, plutonic
- Crystallization of a silica-rich magma in a major intrusion
- 1/16–3/16 in (2–5 mm), phenocrysts to 4 in (10 cm)
- Potassium feldspar, quartz, mica
- Sodium plagioclase, hornblende
- White, light gray, pink, red

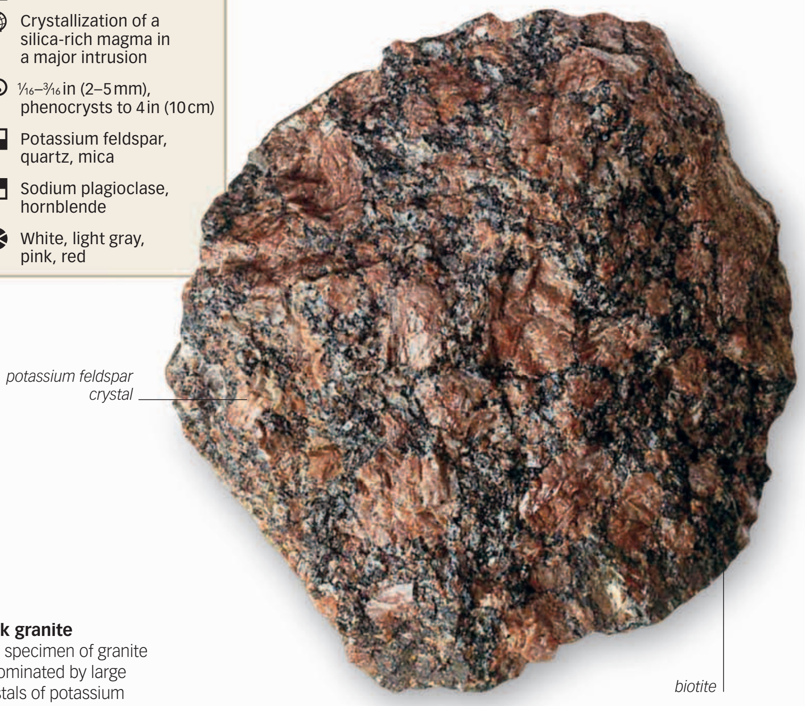

Dark granite This specimen of granite is dominated by large crystals of potassium feldspar, quartz, and biotite.

VARIANTS

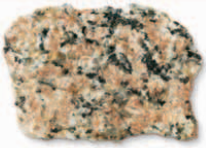

Pink granite Pink feldspars give this granite its color

White granite A specimen dominated by light-colored minerals

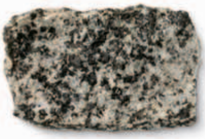

Hornblende granite Granite with black hornblende

GRANITE

The most common intrusive rock in Earth's continental crust, granite is familiar as a mottled pink, white, gray, and black ornamental stone. It is coarse- to medium-grained. Its three main minerals are feldspar, quartz (p.168), and mica, which occur as silvery muscovite (p.195) or dark biotite (p.197) or both. Of these minerals, feldspar predominates, and quartz usually accounts for more than 10 percent. The alkali feldspars are often pink, resulting in the pink granite often used as a decorative stone.

Granite crystallizes from silica-rich magmas that are miles deep in Earth's crust. Many mineral deposits form near crystallizing granite bodies from the hydrothermal solutions that such bodies release.

Granite staircase Granite is used as a building stone. This stairway made of granite ascends to the Bom Jesus Church in Portugal.

PROFILE

- Intrusive
- Intrusive
- Medium to coarse
- Quartz, feldspar, mica
- Hornblende, plagioclase
- Pink, white

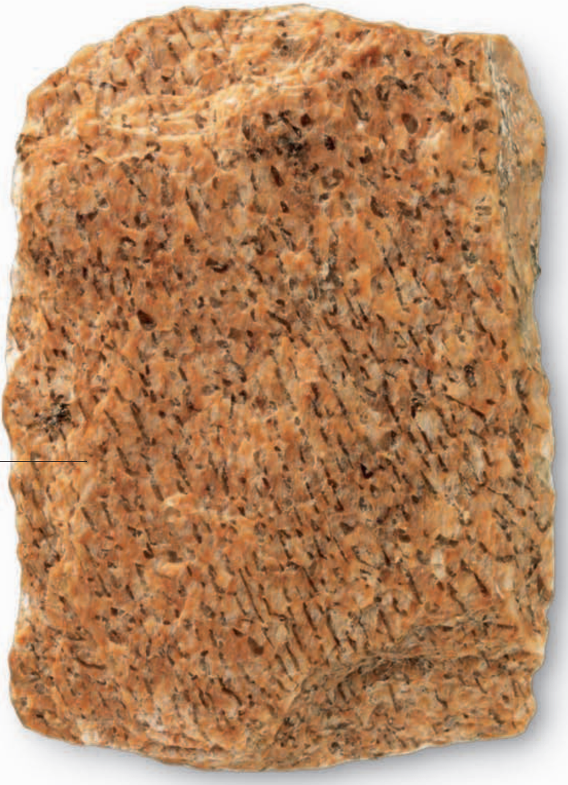

Graphic granite This specimen of granite shows simultaneous growth of quartz and feldspar, which produces a pattern.

VARIANTS

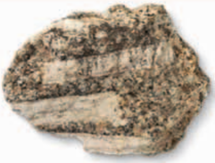

Porphyritic granite A specimen in which large feldspar phenocrysts are set in granite

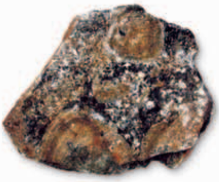

Orbicular granite Granite with spherical phenocrysts of feldspar

TEXTURED GRANITE

Granites with distinct patterns in their crystallization are known as textured granites and in some cases, graphic granites. Graphic granite consists of roughly 30 percent quartz (p.168) and 70 percent feldspar, with a few other minerals. These minerals are intergrown in such a way that straight-sided quartz crystals, which look like hieroglyphic characters, are set in a background of feldspar. The texture forms in pegmatites (p.260) when the main minerals crystallize from the magma at the same time.

Orbicular granite is an unusual but spectacular granite containing spheres (orbicules) of concentric layers of granitic minerals. The orbicules are about 2–4 in (5–10 cm) in diameter and often richer than the granite in darker minerals. They are usually restricted to small areas within the larger granite mass. In porphyritic granite, the feldspar crystals are larger and better formed than the surrounding mineral grains. These granites have been quarried and polished for use as ornamental and building stones.

PROFILE

- Feldspar-rich, plutonic
- Fluid-rich crystallization in the final stages of the formation of a granite
- More than 3⁄16 in (5 mm) to many ft (m)
- Quartz, feldspar, mica
- Tourmaline, topaz
- White, pink

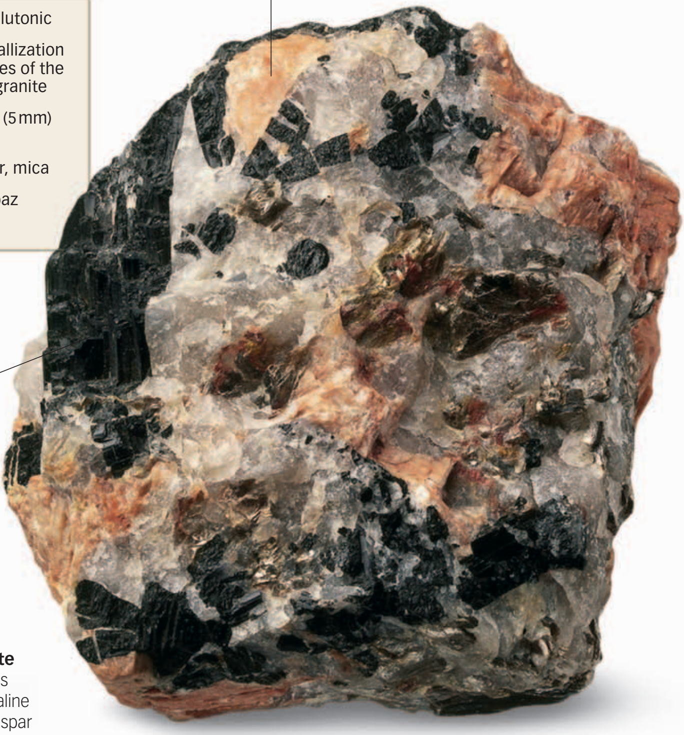

Tourmaline pegmatite
This specimen contains crystals of black tourmaline in a matrix of pink feldspar and white quartz.

PEGMATITE

Important sources of many gemstones, pegmatites are very coarse-grained igneous rocks, mostly of a granitic composition. Although crystals can be huge—over 33 ft (10 m) in some specimens—individual crystals usually average 3¼–4 in (8–10 cm) in length. The large crystals are due to the considerable amount of water in the magma rather than slow cooling. The most perfect crystals are typically found in openings or pockets. Quartz (p.168) and feldspar dominate, but many other minerals can form large, beautiful crystals. Muscovite (p.195) and other micas commonly occur in pegmatites in large, flat sheets known as books.

Pegmatites are light-colored rocks and occur in small igneous bodies, such as veins and dykes, or sometimes as patches in larger masses of granite (pp.258–59). Several gemstones—such as tourmaline (p.224) group minerals, aquamarine (p.225), rock crystal, smoky quartz, rose quartz, and topaz (p.234)—are mined from pegmatites.

VARIANTS

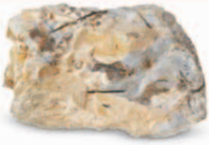

Granitic pegmatite
A specimen with feldspar, quartz, biotite, and needle-shaped tourmaline

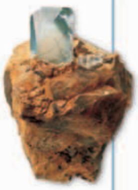

Blue topaz
A topaz crystal in a pegmatite matrix

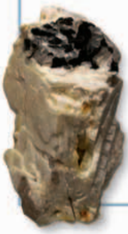

Feldspar pegmatite
A pegmatite dominated by white feldspar, with amphibole

PROFILE

- Ultramafic, plutonic
- Crystallization of a silica-poor magma in a major intrusion
- 1/16–3/16 in (2–5 mm)
- Calcium plagioclase
- Olivine, pyroxene, garnet
- Light gray to white

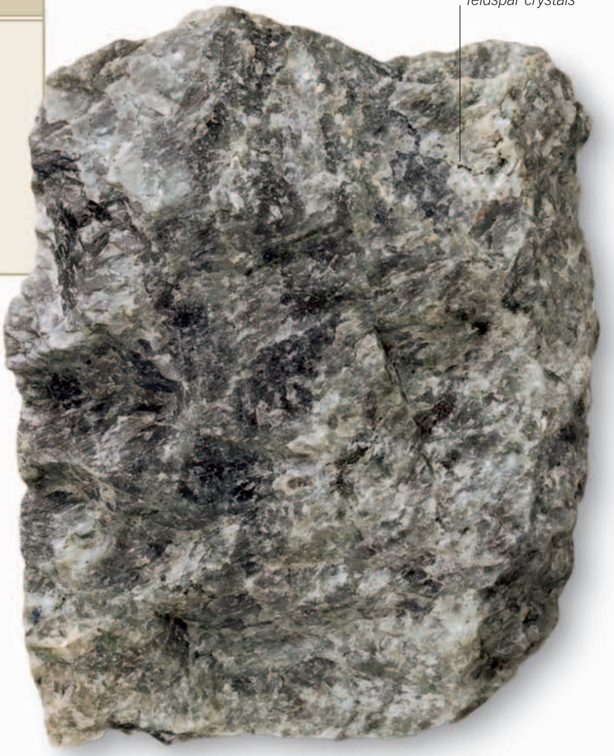

Anorthosite This specimen of anorthosite is dominated by light-colored plagioclase feldspars.

VARIANTS

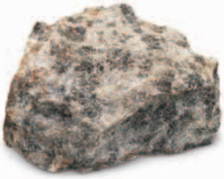

Coarse anorthosite A specimen with pyroxene and orthopyroxene

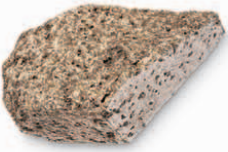

Lunar anorthosite Believed to be the first lunar rock to crystallize

ANORTHOSITE

An intrusive igneous rock, anorthosite is composed of at least 90 percent calcium-rich plagioclase feldspar—principally labradorite (p.180) and bytownite. Olivine (p.232), garnet, pyroxene, and iron oxides make up the remaining 10 percent. Anorthosite is coarse-grained and either white or gray. Specimens can also be green. Many anorthosites have an interesting "cumulate" texture, where well-formed crystals appear to have settled out of the liquid magma, in a similar way as large grains settling in a sediment.

Anorthosite is not a common rock, but where it does occur, it is found as immense masses, or as layers between iron- and magnesium-rich rocks, such as gabbro (p.265) and peridotite (p.266). Anorthosite is, however, common on the surface of the Moon.

Labradorite relief Labrodorite occurs in anorthosite and is used in carvings, such as this 19th-century relief.

PROFILE

- Intermediate silica content, plutonic
- Crystallization of an alkaline intermediate magma in a major intrusion
- $\frac{1}{16}$–$\frac{3}{16}$ in (2–5 mm)
- Potassium feldspar
- Sodium plagioclase, biotite, amphibole, pyroxene, hornblende, feldspathoids
- Gray, pink, or red

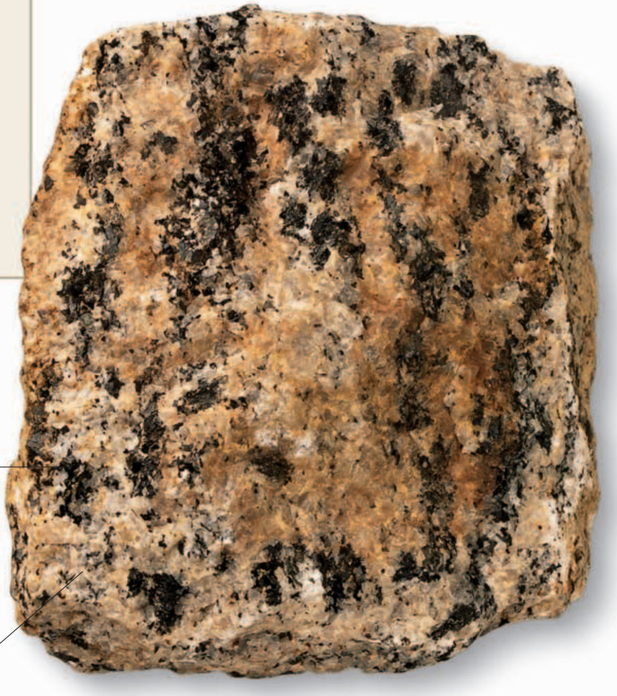

Pink syenite The pink color of this syenite specimen is due to the presence of alkali feldspar, which predominates in syenite.

SYENITE

Visually similar to granite and often confused with it, syenites can be distinguished from granite by the absence or scarcity of quartz. A syenite is any one of a class of rocks essentially composed of: an alkali feldspar or sodium plagioclase (or both); a ferromagnesian mineral, usually biotite (p.197), hornblende (p.218), or pyroxene; and little or no quartz. The alkali feldspars can include orthoclase (p.173), albite (p.177), or less commonly, microcline (p.175). Syenites are attractive, multi-colored rocks—usually gray, pink, or red. Other minerals that occasionally occur in small amounts in syenite include titanite, apatite (p.148), zircon (p.233), magnetite (p.92), and pyrite (p.62). When syenites contain quartz, they are called quartz syenites.

Nepheline (p.182) and alkali feldspar are essential minerals in nepheline syenite, but this rock can contain other minerals, including unusual and attractive ones, such as eudialyte. If the rock includes a pyroxene, it is usually aegirine (p.209); if an amphibole it is usually arfvedsonite, both of which are rich in sodium.

VARIANTS

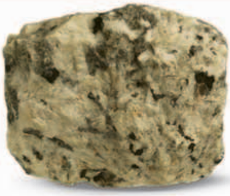

Nepheline syenite Nepheline crystals with black hornblende

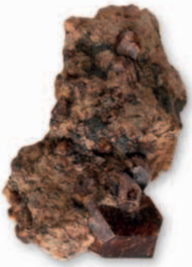

Syenite with zircon Zircon crystals in a syenite matrix

PROFILE

▲ Feldspar-rich, plutonic
🌐 Intrusive
⚲ Coarse
◧ Plagioclase, alkali feldspar, quartz, mica
◰ Hornblende, augite
✱ Gray, white, or pink

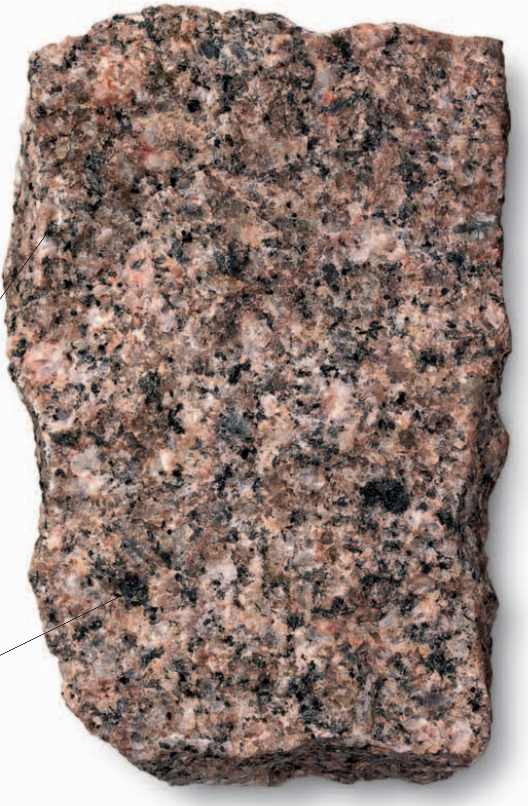

Speckled granodiorite This pink granodiorite has a speckled appearance because of the presence of darker minerals, such as mica and hornblende.

VARIANT

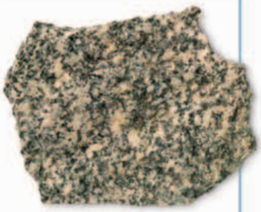

Pink granodiorite A specimen of granodiorite with dark mica and hornblende

GRANODIORITE

Among the most abundant of intrusive igneous rocks, granodiorite is a medium- to coarse-grained rock that is similar to granite (p.258) in texture. Granodiorite can be pink or white, with a grain size and texture similar to that of granite, but abundant plagioclase generally makes it appear darker than granite. The biotite (p.197) and hornblende (p.218) give it a speckled appearance. The mica may occur in well-formed hexagonal crystals, and the hornblende may be present in needlelike crystals. Twinned plagioclase crystals are sometimes wholly encased by orthoclase.

The quartz (p.168) present in granodiorite can be gray to white. With increased amounts of quartz and alkali feldspar, granodiorite grades to granite. With less quartz and alkali feldspar, it becomes diorite (p.264). The volcanic equivalent of granodiorite is dacite (p.274). Two historic stones are granodiorite: the Rosetta Stone was carved from it, and the Plymouth Rock is a glacial erratic boulder of granodiorite.

PROFILE

- ▲ Intermediate silica content, plutonic
- 🌐 Crystallization of a magma with intermediate silica content in a major intrusion
- 🔍 1⁄16–3⁄16 in (2–5 mm)
- ◪ Sodium plagioclase, hornblende
- ◳ Biotite
- ✱ Black or dark green mottled with gray or white

Two-toned diorite
This diorite specimen gets its two-toned appearance from light-colored plagioclase feldspar and black hornblende.

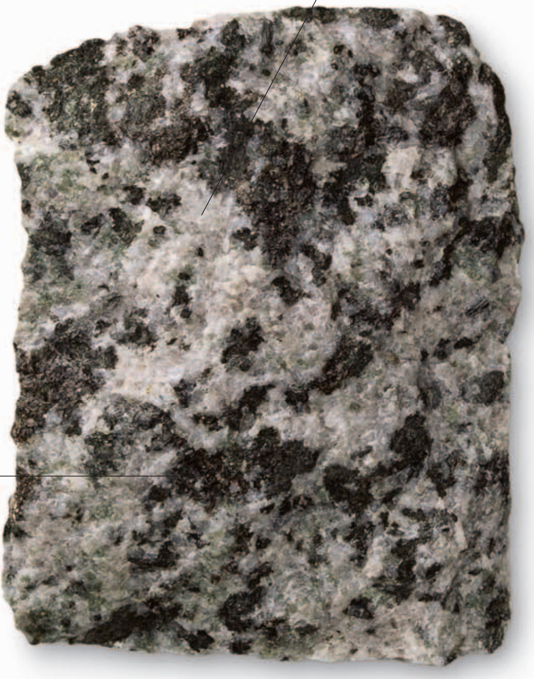

VARIANTS

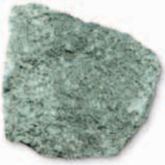

Light-colored diorite
A specimen of diorite with white plagioclase and a minor amount of hornblende

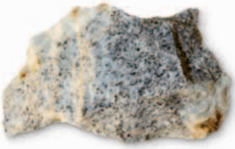

Fine-grained diorite
A specimen of fine-grained diorite with phenocrysts of hornblende

DIORITE

This medium- to coarse-grained intrusive igneous rock is sometimes sold as “black granite.” In general, though, diorite is darker than granite (p.258). It is commonly composed of about two-thirds white plagioclase feldspar and one-third dark-colored minerals, such as biotite (p.197) and hornblende (p.218). The plagioclases in diorite—oligoclase (p.178) or andesine (p.181)—are rich in sodium. Diorite can be of uniform grain size or have large phenocrysts of plagioclase or hornblende.

The rock can occur as large intrusions or as smaller dykes and sills. Most diorite is intruded along the margins of continents. With small amounts of quartz (p.168) and alkali feldspar, it becomes a granodiorite (p.263); with larger amounts, it is classified as granite.

Neolithic axehead
Diorite can be extremely tough and was used to make ancient tools, such as this neolithic axehead.

PROFILE

- Mafic, plutonic
- Crystallization of a silica-poor magma in a major intrusion
- $\frac{1}{16}$–$\frac{3}{16}$ in (2–5 mm)
- Calcium plagioclase feldspar, pyroxene, ilmenite
- Olivine, magnetite
- Dark gray to black

Coarse-grained gabbro This specimen of gabbro has coarse grains, as produced by the formation of large crystals during slow cooling of a magma.

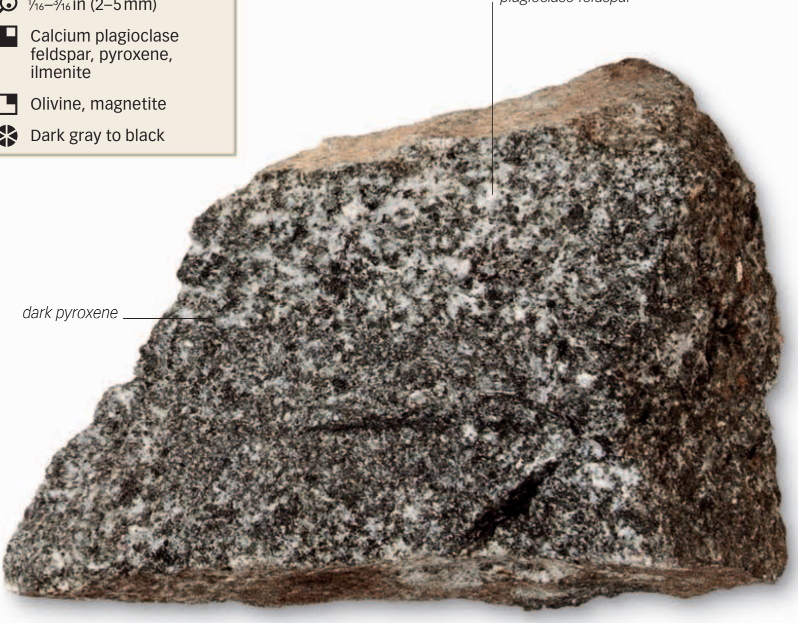

VARIANTS

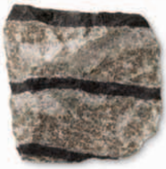

Layered gabbro Bands of light plagioclase and dark ferromagnesian minerals

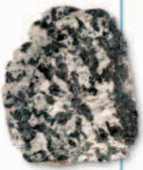

Leucogabbro A gabbro with feldspar-rich crystals

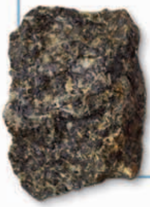

Olivine gabbro A gabbro containing olivine

GABBRO

Medium or coarse-grained rocks, gabbros consist principally of dark green pyroxene (augite and lesser amounts of orthopyroxene) plus white- or green-colored plagioclase and black, millimeter-sized grains of magnetite and/or ilmenite. A gabbro has an intermediate or low silica content and rarely contains quartz (p.168). Gabbro is essentially the intrusive equivalent of basalt (p.273), but unlike basalt gabbro has a highly variable mineral content. It often contains a layering of light and dark minerals (layered gabbro), a significant amount of olivine (olivine gabbro), or a high percentage of coarse crystals of plagioclase feldspar (leucogabbro).

Gabbros are widespread but not common on Earth's surface. They occur as intrusions and as uplifted sections of oceanic crust. Some gabbros are mined for their nickel, chromium, and platinum (p.38). Those containing ilmenite (p.90) and magnetite (p.92) are mined for their iron or titanium.

PROFILE

- Ultramafic, plutonic
- Crystallization of a silica-poor magma in a major intrusion, or Earth's mantle
- 1/16–3/16 in (2–5 mm)
- Olivine, pyroxene
- Garnet, chromite
- Dark green to black

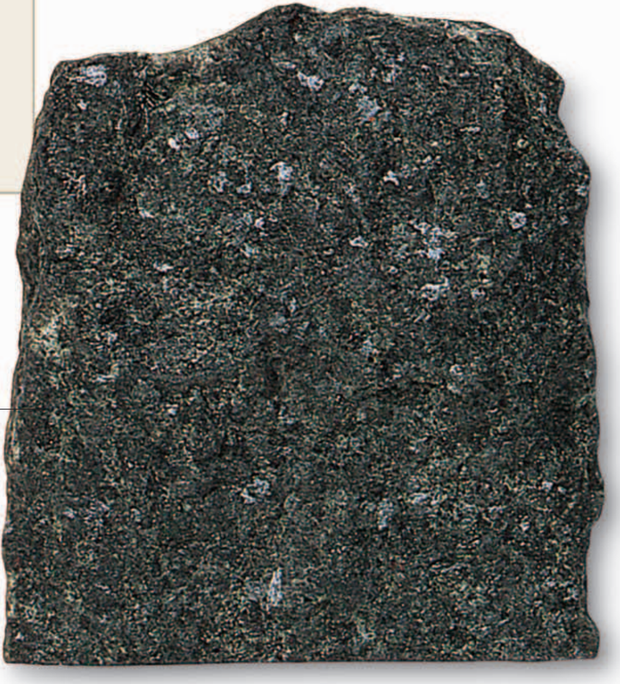

Coarse peridotite This is a specimen of dark, olivine- and pyroxene-rich peridotite from Odenwald, West Germany.

VARIANTS

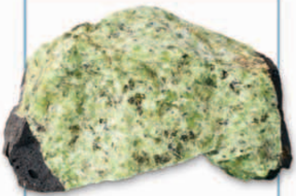

Green peridotite A specimen containing green olivine

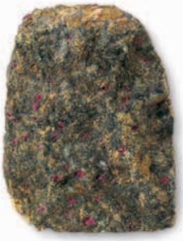

Garnet peridotite Peridotite with red phenocrysts of pyrope garnet

PERIDOTITE

An intrusive igneous rock, peridotite is coarse-grained and dense. It is light to dark green in color. Peridotite contains at least 40 percent olivine (p.232) and some pyroxene. Unlike the olivine grains, the pyroxene grains in peridotite have a visible cleavage when viewed under a hand lens. Peridotite forms much of Earth's mantle and can occur as nodules that are brought up from the mantle by kimberlite (p.269) or basalt (p.273) magmas.

The rock is usually found interlayered with iron- and magnesium-rich rocks in the lower parts of layered igneous rock bodies, where its denser crystals first form through selective crystallization and then settle to the bottom of still-fluid or semi-solid crystallizing mushes. A peridotite specimen that has been altered by weathering becomes serpentinite (p.298). Peridotite and pyroxenite (p.267) form in similar environments, but pyroxenite contains a higher percentage of pyroxene. Peridotites are important sources of chromium and nickel.

Pyroxenite
Pyroxene, the main component of pyroxenite, can be seen in this specimen, along with smaller amounts of plagioclase feldspar and accessory sulfide minerals.

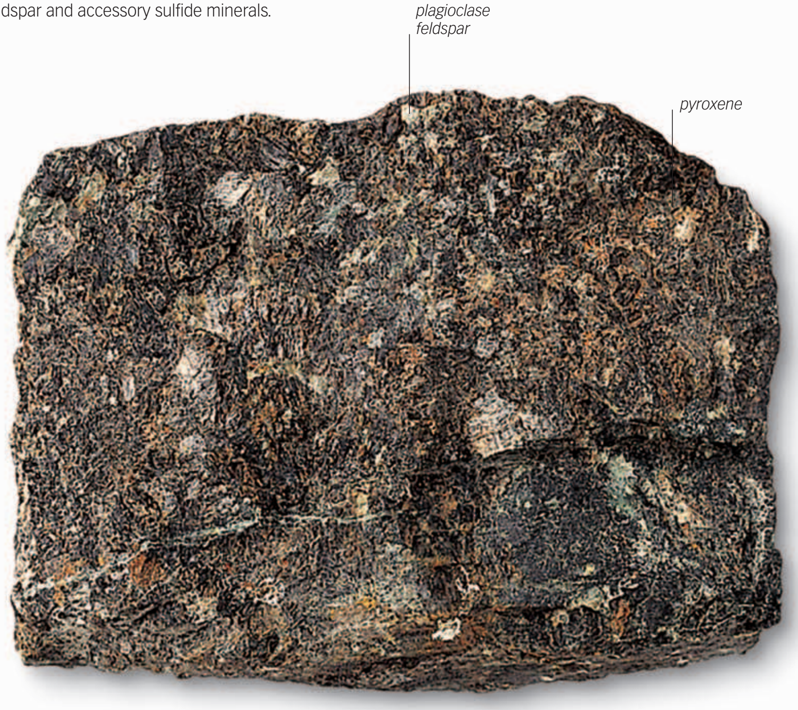

PROFILE

- Intrusive
- Crystallization of a silica-poor magma in a major intrusion
- 1/16–3/16 in (2–5 mm)
- Pyroxene
- Biotite, hornblende, olivine, plagioclase, nepheline
- Light green, dark green, or black

PYROXENITE

This is a coarse-grained, granular igneous rock that contains at least 90 percent pyroxene. Pyroxenite may also contain olivine (p.232) and oxide minerals when it occurs in layered intrusions or nepheline (p.182) when it occurs in silica-poor intrusions. A hard and heavy rock, it is light green, dark green, or black. Its surface often weathers to rusty brown. Individual crystals may be 3 in (7.5 cm) or more in length. Pyroxenites are usually found with gabbros (p.265) and peridotites (p.266). Unlike gabbros, pyroxenite contains almost no feldspars. Also, pyroxenite has less olivine than peridotites.

The principal minerals usually found accompanying pyroxenites, in addition to olivine and feldspar, are chromite (p.99) and other spinels (p.96), garnet, rutile (p.78), and magnetite (p.92). It has been proposed that large volumes of pyroxenite form in the upper mantle. Rare metamorphic pyroxenites are known and are described as pyroxene hornfels.

PROFILE

- Mafic, plutonic
- Crystallization of a silica-poor magma in a minor intrusion
- $\frac{1}{256}$–$\frac{1}{16}$ in (0.1–2 mm)
- Calcium plagioclase, pyroxene
- Quartz, magnetite, olivine
- Dark gray to black, often mottled white

Dark gray dolerite
Dolerite's characteristic medium texture and dark color can be seen in this specimen.

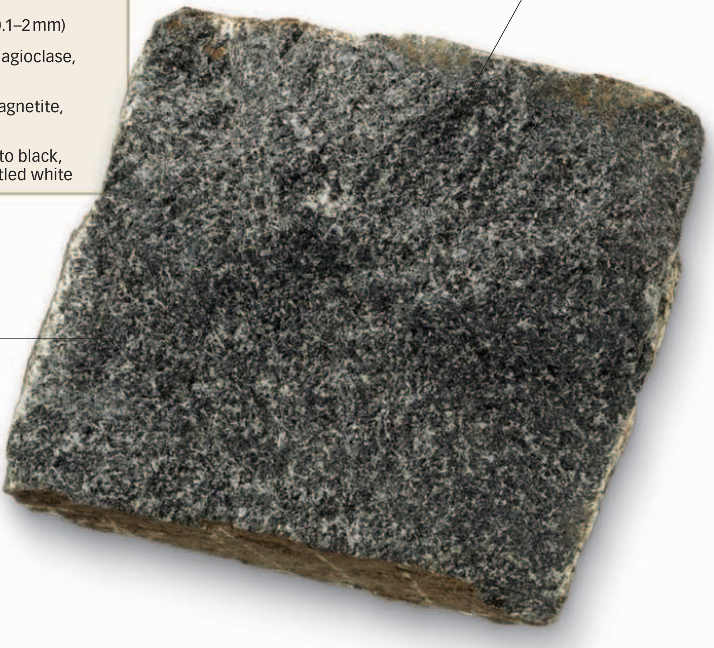

VARIANTS

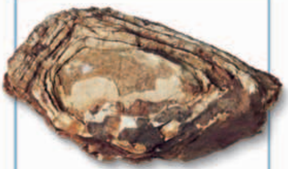

Weathered dolerite A dolerite specimen showing surface flaking from weathering

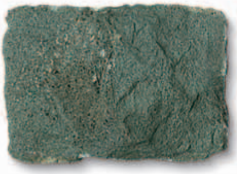

Fine-grained dolerite A specimen of dolerite with fine-grained texture

DOLERITE

A medium-grained rock, dolerite has the same composition as gabbro (p.265): one- to two-thirds is calcium-rich plagioclase feldspar, and the remainder is mainly pyroxene. Specimens can have up to 55 percent silica and up to 10 percent quartz (p.168). Plagioclase crystals commonly occur as tiny rectangular crystals within larger pyroxene grains in dolerite. Olivine (p.232) can occur as a constituent in the form of rounded grains that are often weathered orange-brown.

An extremely hard and tough rock, dolerite occurs in dykes and sills intruded into fissures in other rocks. It is a heavy rock that is often polished for use as a decorative stone. Dolerite is also used in its rough state for paving and is crushed for road stone. It is a stone sold under the name "black granite."

Stonehenge, England The inner circle at Stonehenge in England is made up of about 80 pieces of dolerite "bluestones."

KIMBERLITE

The major source of diamonds (p.47), kimberlite is a variety of peridotite (p.266). It is rich in mica, often in the form of crystals of phlogopite, a type of mica. Other abundant constituent minerals include chrome-diopside (p.210), olivine (p.232), and chromium- and pyrope-rich garnet. Lesser amounts of rutile (p.78), perovskite (p.89), ilmenite (p.90), magnetite (p.92), calcite (p.114), serpentine (p.191), pyroxene, and diamond can also be present.

Kimberlite is typically found in pipes—structures with vertical sides roughly circular in cross section. The rock may have been injected from the mantle into zones of weakness in the crust. Fragments of mantle rock are often brought to the surface in kimberlites, making them a valuable source of information about inner Earth.

PROFILE

- ▲ Silica-poor, volcanic
- 🌐 Extrusion of a fluid part of Earth's mantle
- 🔍 Wide range
- ◩ Olivine, pyroxene, mica, garnet, diopside
- ◲ Ilmenite, diamond, serpentine, calcite, rutile, perovskite, magnetite
- ✱ Dark gray

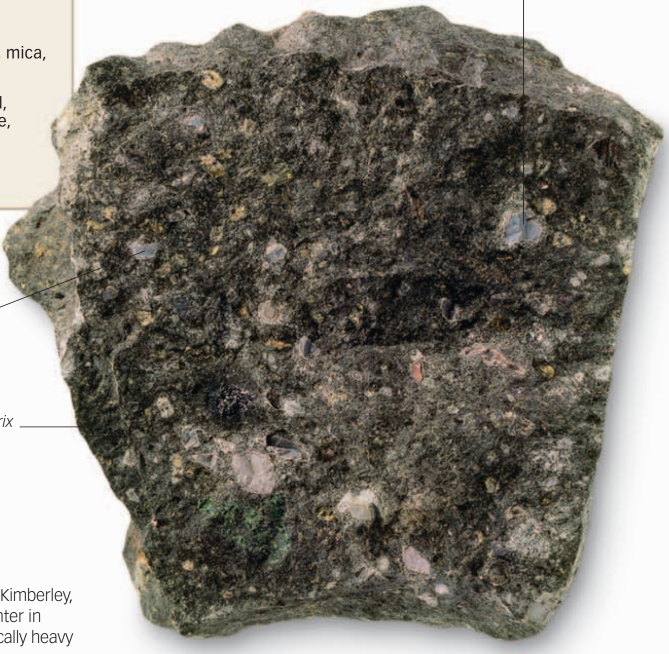

Brecciated kimberlite
Kimberlite is named after Kimberley, the diamond industry center in South Africa. This is a typically heavy and fragmented specimen.

Diamond ring
This ring has an emerald-cut diamond on a gold shank. Kimberlites are the primary source rock for diamonds.

VARIANTS

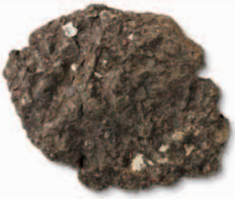

Weathered kimberlite
Heavily weathered kimberlite from Kimberley, South Africa

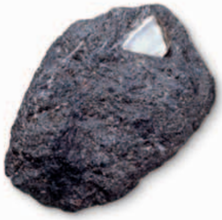

Diamond in kimberlite
An octahedral diamond in a kimberlite matrix

PROFILE

- Extrusive
- Crystallization of an alkaline magma in a minor intrusion
- 1⁄256–1⁄16 in (0.1–2 mm)
- Orthoclase, plagioclase, biotite, hornblende
- Hornblende, magnetite, axinite, amphibole, pyroxene
- Dark brown to black

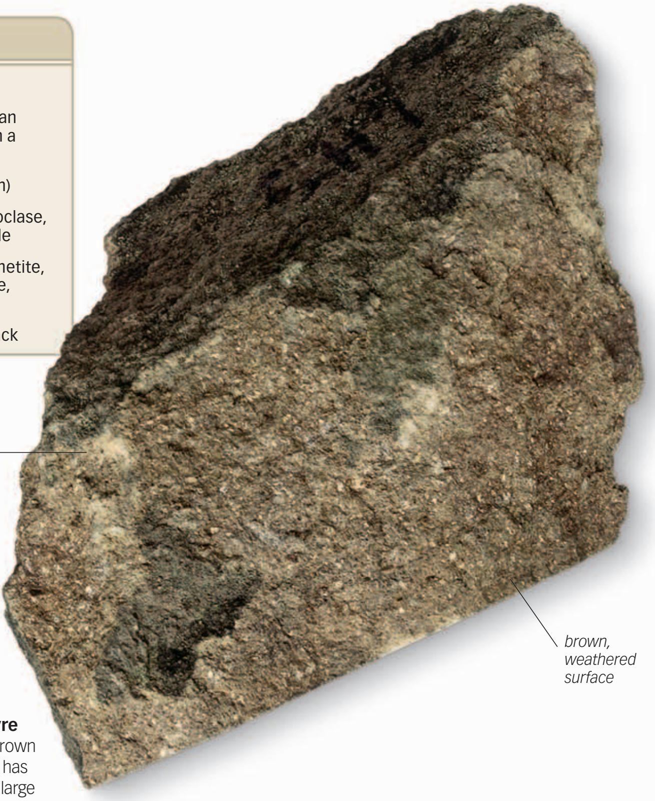

Weathered lamprophyre This weathered grayish brown specimen of lamprophyre has a porphyritic texture, with large crystals set in a fine matrix.

VARIANTS

Dark brown lamprophyre A specimen of lamprophyre with mica flakes

Fine-grained lamprophyre Lamprophyre with fine grains and no phenocrysts

LAMPROPHYRE

The term lamprophyre is used to refer to a group of igneous rocks with high potassium, magnesium, and iron content. Four minerals dominate these rocks: orthoclase (p.173), plagioclase, biotite (p.197), and hornblende (p.218). Amphibole and biotite tend to occur in a matrix of various combinations of plagioclase and other sodium- and potassium-rich feldspars, pyroxene, and feldspathoids (pp.182–84). Because of their relative rarity and varied composition, lamprophyres do not fit into standard geological classifications. In general, they form at great depth and are enriched in sodium, cesium, rubidium, nickel, and chromium, as well as potassium, iron, and magnesium. Some are also source rocks for diamonds.

The exact origin of lamprophyres is still debated. These rocks occur mainly in dykes, sills, and other small igneous intrusions. They form along the margins of some granites (pp.258–59) and are often associated with large bodies of intrusive granodiorite (p.263).

PROFILE

Extrusive

Two-stage crystallization of an igneous rock

Less than $\frac{1}{256}$ in (0.1 mm); phenocrysts up to ¾ in (2 cm)

Various

Various

Red, green, purple

Rhomb porphyry This porphyry has feldspar with rhombic cross sections in a fine-grained matrix.

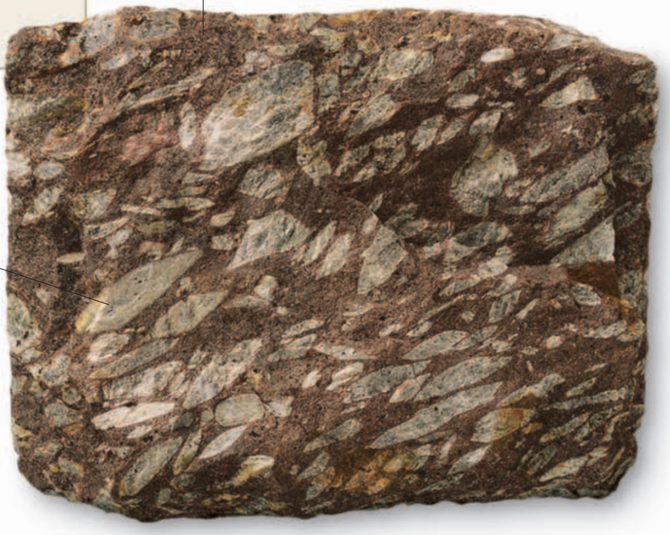

VARIANTS

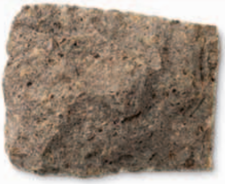

Quartz porphyry Porphyry with phenocrysts of quartz

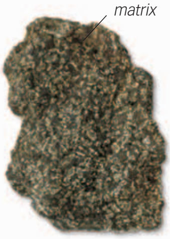

Feldspar porphyry Phenocrysts of feldspar in a reddish brown matrix

PORPHYRY

The name porphyry is a general name and textural term for medium- to fine-grained igneous rocks that contain large crystals (phenocrysts) of other minerals, especially if these minerals are found in the smaller crystals of the matrix. It is most often used for rocks formed in lava flows or minor intrusions. The term porphyry is often prefixed with a reference to the minerals it contains, such as quartz–feldspar porphyry, which contains phenocrysts of the two minerals. Alternatively, the prefix can refer to the composition or texture of the rock. Examples are rhyolite porphyry or rhomb porphyry, respectively.

Porphyries form when crystallization begins deep in Earth's crust and cooling occurs quickly after rapid upward movement of magma. This results in the formation of very small crystals of the matrix. Historically, the name porphyry was used for the purple-red form of the rock, which has been valued since antiquity as an ornamental stone. Many Egyptian, Roman, and Greek sculptures used this type of porphyry.

PROFILE

Intrusive and extrusive

Crystallization of a carbonate magma derived from Earth’s mantle

Wide range

Calcite, dolomite, siderite

Magnetite, apatite, pyrochlore, phlogopite

Cream, yellow, brown

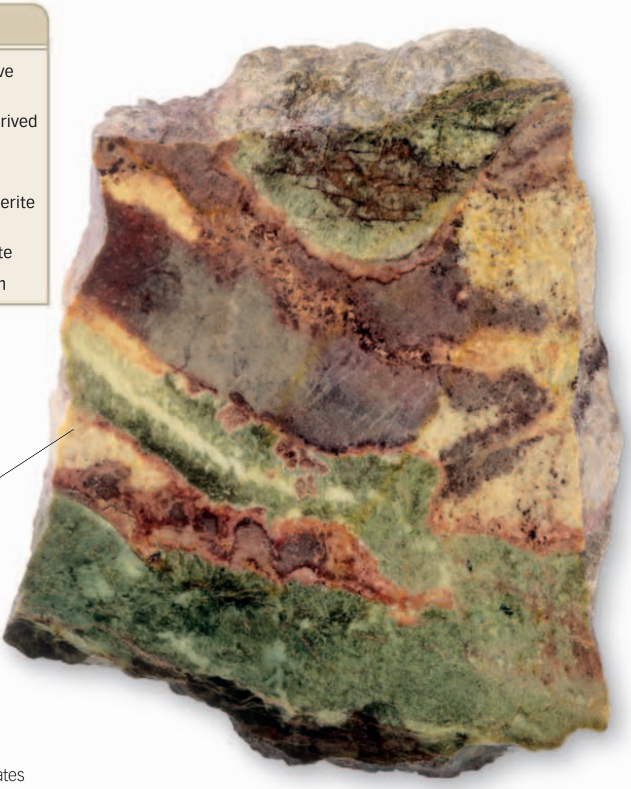

Carbonatite This carbonatite specimen contains a mixture of carbonates and other minerals.

VARIANT

Carbonatite with magnetite A specimen of carbonatite with accessory black magnetite

CARBONATITE

An unusual rock type, carbonatite consists of over 50 percent carbonate minerals—usually calcite (p.114), dolomite (p.117), or siderite (p.123). Most carbonatites also contain some portion of silicate minerals and may contain magnetite (p.92), the brown mica phlogopite, and rare-earth minerals such as pyrochlore (p.85). Specimens are typically cream-colored, yellow, or brown. Carbonatite looks similar to marble (p.301). It may be coarse-grained if intrusive and fine-grained if volcanic.

The process by which carbonatites form is still a matter of conjecture. They are usually found in areas of continental rifting in veins, dykes, and sills around intrusions of sodium- and potassium-rich igneous rocks. The geochemistry of carbonatites is complex and can vary considerably in specimens. Many contain scarce and valuable ores of rare elements, such as niobium, cesium, tantalum, thorium, and hafnium. Some carbonatites also contain significant amounts of platinum (p.38), gold (p.42), silver (p.43), and nickel.

PROFILE

- Mafic, volcanic
- Extrusion of a silica-poor magma
- Less than $\frac{1}{256}$ in (0.1 mm)
- Sodium plagioclase, pyroxene, olivine
- Leucite, nepheline, augite
- Dark gray to black

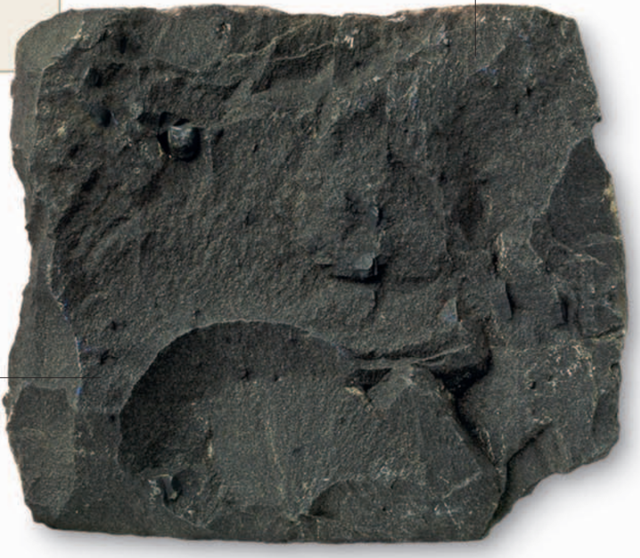

Fine-grained basalt This specimen of basalt shows its characteristic fine-grained texture.

VARIANTS

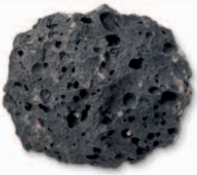

Vesicular basalt A basalt specimen with holes left by gas bubbles during cooling

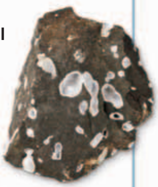

Amygdaloidal basalt A specimen with zeolite crystals in vesicules

Porphyritic basalt A basalt with phenocrysts of pyroxene

BASALT

Basalt is the most common rock on Earth's surface. Specimens are black in color and weather to dark green or brown. Basalt is rich in iron and magnesium and is mainly composed of olivine (p.232), pyroxene, and plagioclase. Most specimens are compact, fine-grained, and glassy. They can also be porphyritic, with phenocrysts of olivine, augite (p.211), or plagioclase. Holes left by gas bubbles can give basalt a coarsely porous texture.

Basalt makes up large parts of the ocean floor. It can form volcanic islands when it is erupted by volcanoes in ocean basins. The rock has also built huge plateaus on land. The dark plains on the Moon, known as maria, and, possibly, the volcanoes on Mars and Venus are made of basalt.

Basalt temple This magnificent, thousand-pillared temple in Andhra Pradesh, India, is made of gray basalt.

PROFILE

- Extrusive
- Extrusion of a magma with intermediate silica content
- Less than 1⁄256 in (0.1 mm)
- Quartz, plagioclase
- Biotite, hornblende, pyroxene
- Gray to pink

Porphyritic dacite This dacite specimen has a porphyritic texture, with prominent phenocrysts of plagioclase and biotite.

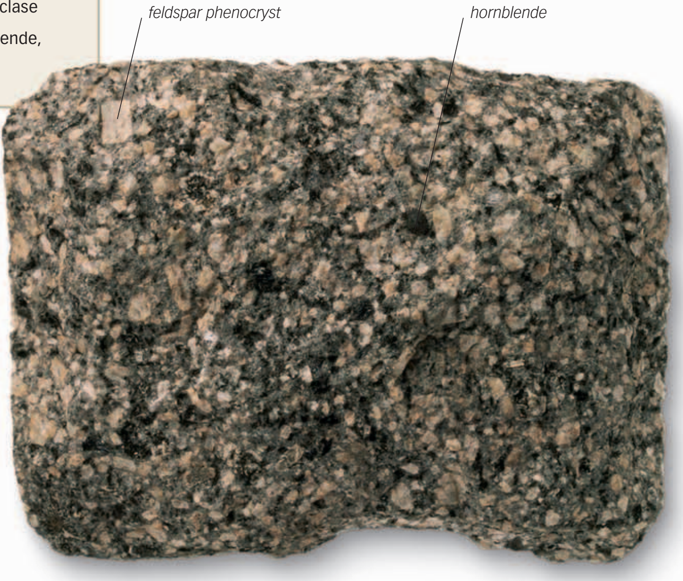

VARIANTS

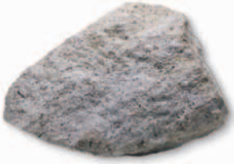

Fine-grained dacite A fine-grained specimen of dacite with small phenocrysts

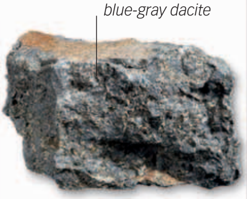

Blue-gray dacite A specimen of dacite with dark phenocrysts

DACITE

An extrusive igneous rock, dacite takes its name from Dacia (modern Romania)—the ancient Roman province where it was first found. The volcanic equivalent of granodiorite (p.263), dacite is usually pink or a shade of gray. It often has flowlike bands. Porphyritic varieties are common, with large crystals usually consisting of blocky plagioclase feldspar or rounded quartz (p.168), or both. Dacite matrix can be cryptocrystalline or glassy.

Dacite occurs with andesite (p.275) on continental margins, and with rhyolite (p.278) in continental volcanic districts. Along continental margins, dacite magmas form in areas where oceanic crust sinks beneath continental crust. Dacite magmas are chemically altered as they reach the mantle. Dacite lavas are quite viscous because of their moderate silica content, and thus can be quite explosive in eruptions. The explosion of Mount Saint Helens volcano in USA in 1980 was a result of dacite domes formed from previous eruptions. The mineral compositions of dacite lavas tell the history of the magma.

PROFILE

- Intermediate silica content, volcanic
- Extrusion of a magma with intermediate silica content
- Less than $\frac{1}{256}$ in (0.1 mm)
- Plagioclase feldspars
- Pyroxene, amphibole, biotite
- Light to dark gray, reddish pink

Porphyritic andesite
This specimen of porphyritic andesite shows light feldspar phenocrysts in a dark andesite matrix.

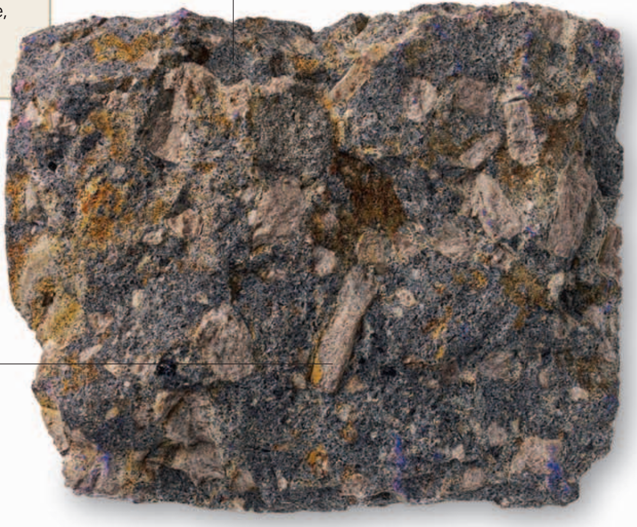

VARIANTS

Fine-grained andesite
A specimen from the Solomon Islands, Pacific Ocean

small phenocrysts

Andesite with plagioclase
A specimen of andesite with phenocrysts of light plagioclase

Amygdaloidal andesite Vesicles of andesite filled with a zeolite

ANDESITE

This volcanic rock is named after the Andes Mountains. Intermediate in silica content, it is usually gray in color and may be fine-grained or porphyritic. Andesite is the volcanic equivalent of diorite (p.264). It consists of the plagioclase feldspar minerals andesine (p.181) and oligoclase (p.178), together with one or more dark, ferromagnesian minerals such as pyroxene and biotite (p.197).

Amygdaloidal andesite occurs when the voids left by gas bubbles in the solidifying magma are later filled in, often with zeolite minerals (pp.185–90). Andesite erupts from volcanoes and is commonly found interbedded with volcanic ash and tuff (p.282). Ancient andesites are used to map ancient subduction zones because andesitic volcanoes form on continental or ocean crust above these zones.

Volcanic andesite
Mount Fujiyama in Yamanashi Prefecture, Honshu, Japan, is the cone of an andesitic volcano.

PROFILE

- Extrusive
- Rubbly top of an extruded, silica-poor magma
- Less than 1⁄256 in (0.1 mm)
- Plagioclase, pyroxene
- Apatite, magnetite, olivine
- Dark brown, black, or red

Vesicular scoria The large vesicles characteristic of scoria can be seen in this specimen.

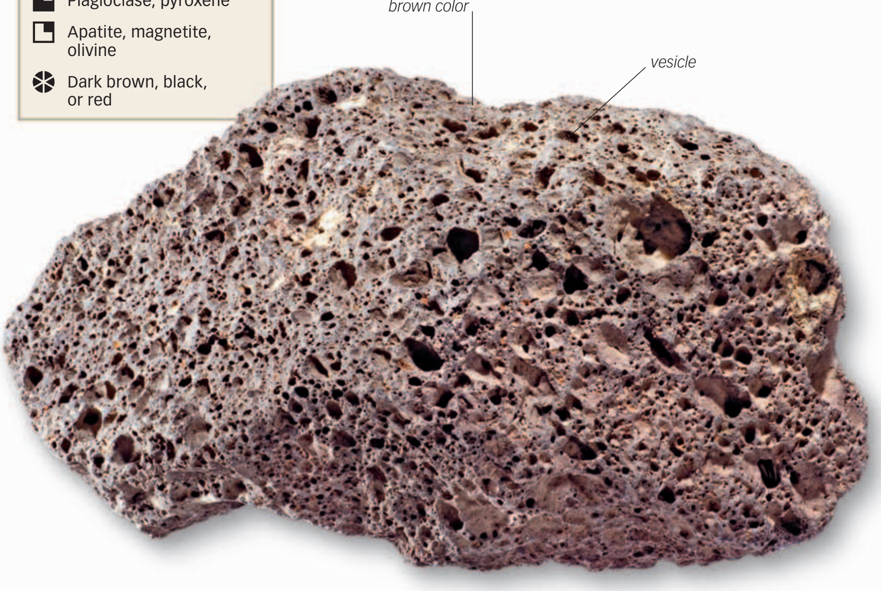

VARIANT

Rubbly scoria A specimen of rubbly scoria from the top of a lava flow

SCORIA

The top of a lava flow is made up of a highly vesicular, rubbly material known as scoria. It has the appearance of vesicular lava. When fresh, scoria is generally dark in color—dark brown, black, or red. Weathered scoria has a medium-brown color and forms piles of loose rubble with small pieces. Most scoria is basaltic or andesitic in composition. This rock forms when gases in the magma expand to form bubbles as lava reaches the surface. The bubbles are then retained as the lava solidifies. Scoria is common in areas of recent volcanism, such as the Canary Islands and the Italian volcanoes.

This rock is of relatively low density due to its vesicles, but it is not as light as pumice (p.277), which floats on water. Scoria also differs from pumice in that it has larger vesicles with thicker walls. Scoria has commercial use as a high-temperature insulating material. It also has applications in landscaping and drainage.

PROFILE

- Volcanic
- Solidification of a silica-rich lava with trapped gas bubbles
- Less than 1⁄256 in (0.1 mm)
- Glass
- Feldspar, augite, hornblende, zircon
- Black, white, yellow, brown

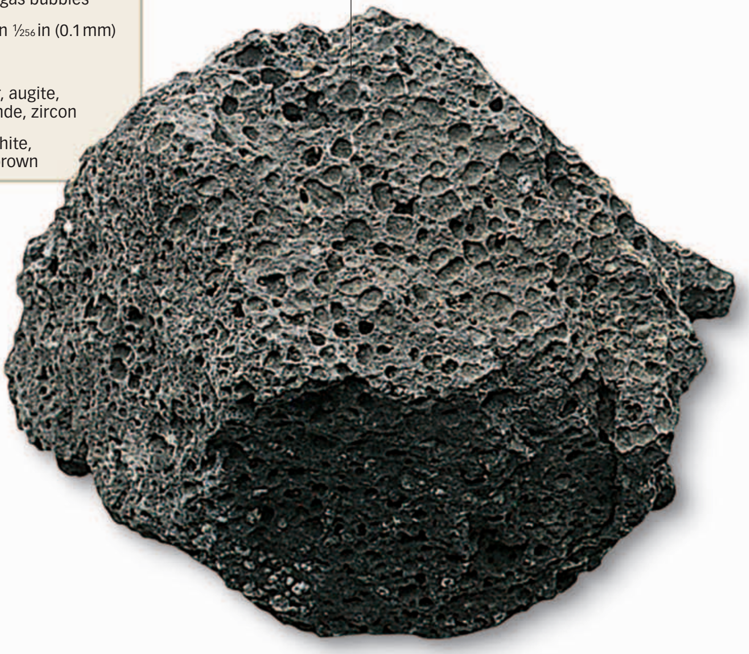

Frothy pumice The hollows (or vesicles) in this pumice clearly show its frothy nature. Vesicles may join together to form hollows or passages.

VARIANTS

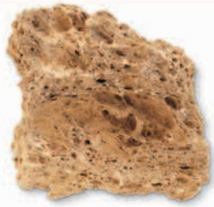

Rhyolitic pumice A light-colored rhyolitic pumice with a frothy structure

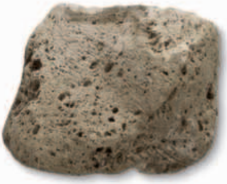

Historic pumice A specimen of pumice from the Krakatoa eruption of 1883

PUMICE

A porous and frothlike volcanic glass, pumice is created when gas-saturated liquid magma erupts like a carbonated drink and cools so rapidly that the resulting foam solidifies into a glass full of gas bubbles. Pumices from silica-rich lavas are white, those from lavas with intermediate silica content are often yellow or brown, and rarer silica-poor pumices are black. The hollows in the froth can be rounded, elongated, or tubular, depending on the flow of the solidifying lava. The glassy material that forms pumice can be in threads, fibers, or thin partitions between the hollows.

Although pumice is mainly composed of glass, small crystals of various minerals occur. Pumice has a low density due to its numerous air-filled pores. For this reason, it can easily float in water.

Pumice stone Pumice is soft and easily shaped. Mildly abrasive, it is often used to remove rough skin.

PROFILE

- Feldspar-rich, volcanic
- Extrusion of a silica-rich magma
- Less than $\frac{1}{256}$ in (0.1 mm)
- Quartz, potassium feldspar
- Glass, biotite, amphibole, plagioclase, pyroxene
- Light to medium gray, light pink

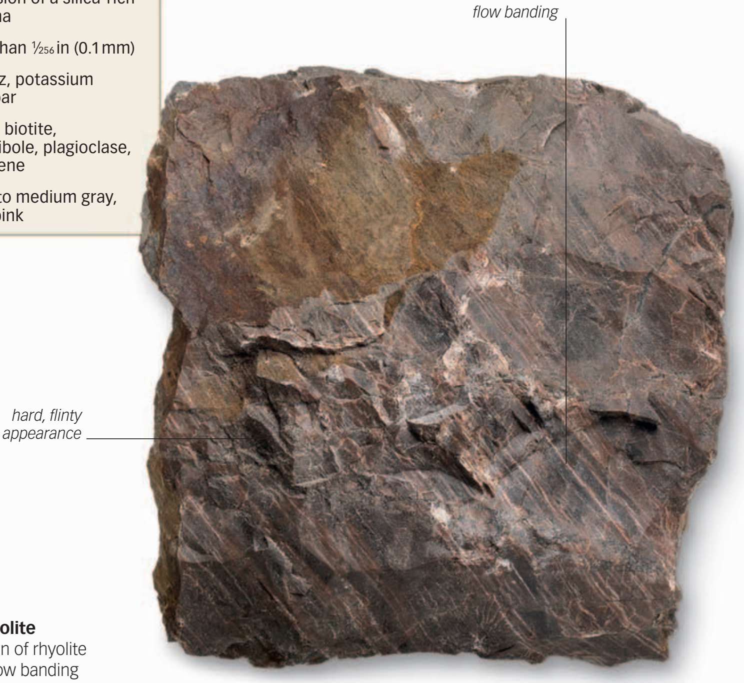

Banded rhyolite
This specimen of rhyolite has visible flow banding across its surface.

VARIANT

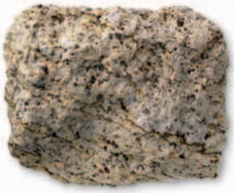

Porphyritic rhyolite
Light-colored rhyolite with phenocrysts of quartz

RHYOLITE

A rare volcanic rock, rhyolite is usually fine-grained. It is often composed largely of volcanic glass (pp.280–81). Individual grains of quartz (p.168), feldspar (pp.173–81), and mica may be present but are too small to be visible. The small size of these grains indicates that crystallization began before the lava flowed to the surface. Rhyolites sometimes have millimeter-scale phenocrysts of quartz, feldspar, or both. Specimens can also include iron- and magnesium-rich minerals, such as biotite (p.197) or pyroxene and amphibole. The granitic magma from which rhyolite crystallizes is very viscous. Therefore, flow banding is often preserved and can be seen on weathered surfaces. Banded rhyolites have few or no phenocrysts. A rhyolite variant with tiny crystals arranged in radiating spheres is called spherulitic rhyolite.

Rhyolite occurs with pumice (p.277), obsidian (p.280), and intermediate volcanic rocks, such as andesite (p.275). Rhyolites are almost exclusively confined to the interiors and margins of continents.

PROFILE

- Intermediate silica content, volcanic
- Extrusion of an alkaline intermediate magma
- Less than 1⁄256 in (0.1 mm)
- Sanidine, oligoclase
- Feldspathoids, quartz, hornblende, pyroxene, biotite
- Off-white, gray, pale yellow, pink

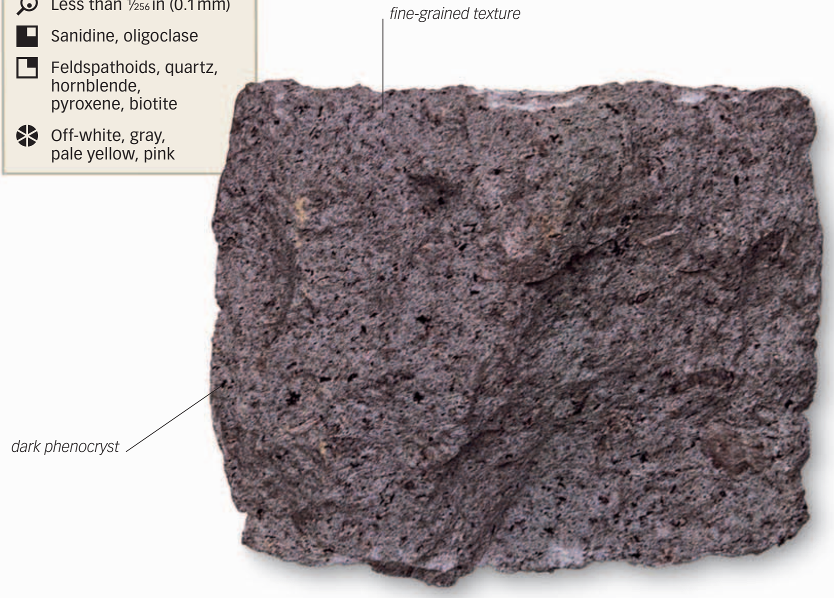

Porhyritic trachyte This specimen of fine-grained trachyte has phenocrysts of a dark mineral.

VARIANTS

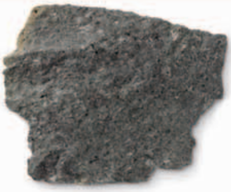

Gray trachyte A fine- to medium-grained specimen of trachyte

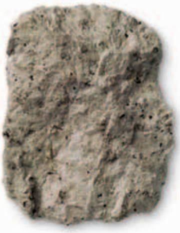

Porphyritic trachyte Light-colored trachyte with dark mineral phenocrysts

TRACHYTE

The name trachyte comes from the Greek word *trachys*, which means “rough”—a reference to the rock’s typical rough texture. Trachyte’s composition is dominated by alkali feldspar—a major component of the fine matrix and of the abundant phenocrysts that are common in the rock. Dark, iron- and magnesium-rich minerals, such as biotite (p.197), pyroxene, and amphibole, can be present in small quantities. Trachyte is similar to rhyolite (p.278) in color and occurence, but it contains little or no quartz (p.168).

Trachyte occurs on continents and oceanic islands with other volcanic rocks that are rich in alkali feldspars, have intermediate to high silica content, and are iron- and magnesium-rich.

Trachyte stonework The cathedral in Morelia, in Mexico’s Michoacán state, has ornate stonework of pink trachyte.

PROFILE

- Feldspar-rich, volcanic
- Extrusion and rapid cooling of a silica-rich magma
- Not granular
- Glass
- Hematite, feldspar
- Black, red, brown

Icelandic obsidian
This classic specimen of obsidian from Iceland perfectly demonstrates the rock's conchoidal fracture and vitreous luster.

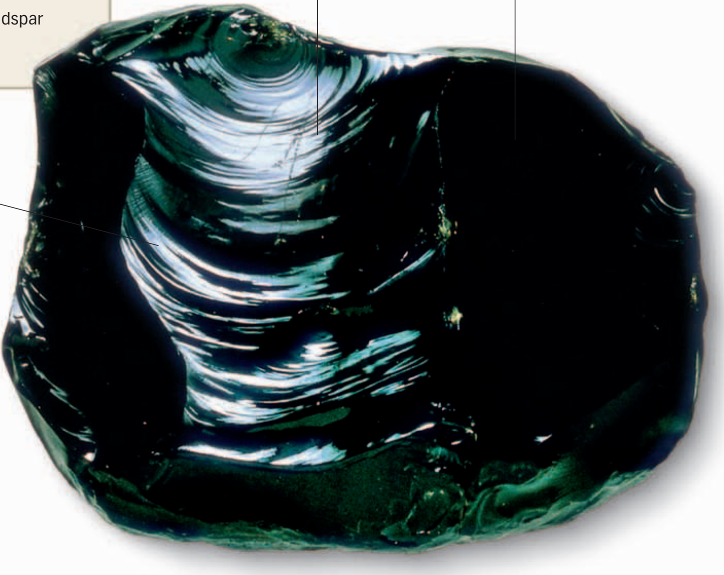

VARIANTS

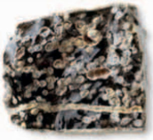

Vesicular obsidian
A specimen with white, silicate-filled vesicles

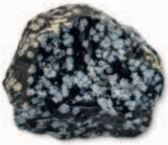

Snowflake obsidian
A specimen with needlelike crystals in spherical aggregates

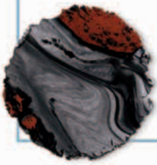

Banded obsidian
Obsidian with red hematite

OBSIDIAN

The natural volcanic glass obsidian forms when lava solidifies so quickly that crystals do not have time to form. Specimens are typically jet black, although the presence of hematite (p.90) can produce red and brown variants. The inclusion of tiny gas bubbles can sometimes create a golden sheen. Tiny crystals of feldspar (pp.173–81) and phenocrysts of quartz (p.168) can also be present. Obsidian can show flow banding.

Although obsidian can have any chemical composition, most specimens have a composition similar to rhyolite (p.278) and are found on the outer edges of rhyolite domes and flows. Like rhyolite, obsidian is also found along the rapidly cooled edges of sills and dykes. Most obsidian is relatively young, as the glass gradually crystallizes into minerals over a period of time.

Obsidian tear drop
These polished obsidian nodules are called Apache Tears after the tears of felled Apache warriors.

PROFILE

- Extrusive
- Rapid cooling of basalt in a lava fountain
- Up to 1⁄32 in (1 mm)
- Basaltic glass
- None
- Brown, black

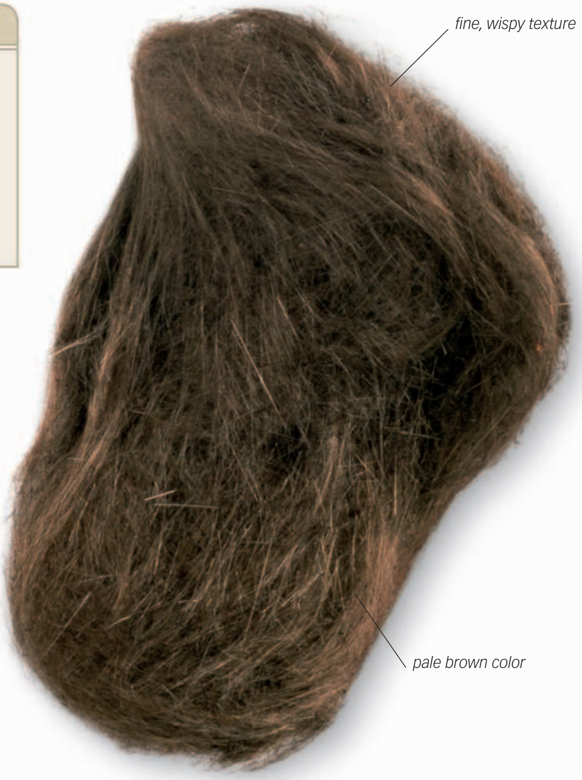

Pele's hair This unusual rock produced in volcanic eruptions is made up of hairlike fibers of volcanic glass.

VARIANTS

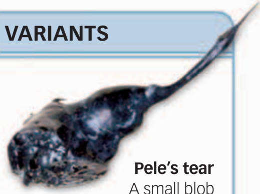

Pele's tear A small blob of volcanic glass elongated by airflow

Golden hair A mass of fine, golden volcanic glass

Long strands Particularly long strands of Pele's hair

PELE'S HAIR AND TEARS

Two of the more unusual kinds of extrusive volcanic rock are named after Pele, the Hawaiian goddess of volcanoes. The first of these is Pele's hair, which refers to threads or fibers of volcanic glass (pp.280–81) formed when small droplets of molten basaltic material are blown into the air and spun out by the wind into long, hairlike strands. Specimens are usually emitted from lava fountains, lava cascades, vents, and vigorous lava flows. They are generally deep yellow or golden in color. A single strand of Pele's hair with a diameter of less than 1⁄16 in (0.5 mm) can be as long as 6½ ft (2 m). Strands can be blown tens of miles away from the vent or fountain where they originated.

The second variety of volcanic extrusive named after Pele is Pele's tears, which are small blobs of volcanic glass formed in much the same way as Pele's hair. Specimens occur as spheres or tear drops that are jet black in color. They are frequently found on one end of a strand of Pele's hair.

PROFILE

- Volcanic
- Pyroclastic accumulation of fine material
- 1/256–1/16 in (0.1–2 mm)
- Glassy fragments
- Crystalline fragments
- Light to dark brown

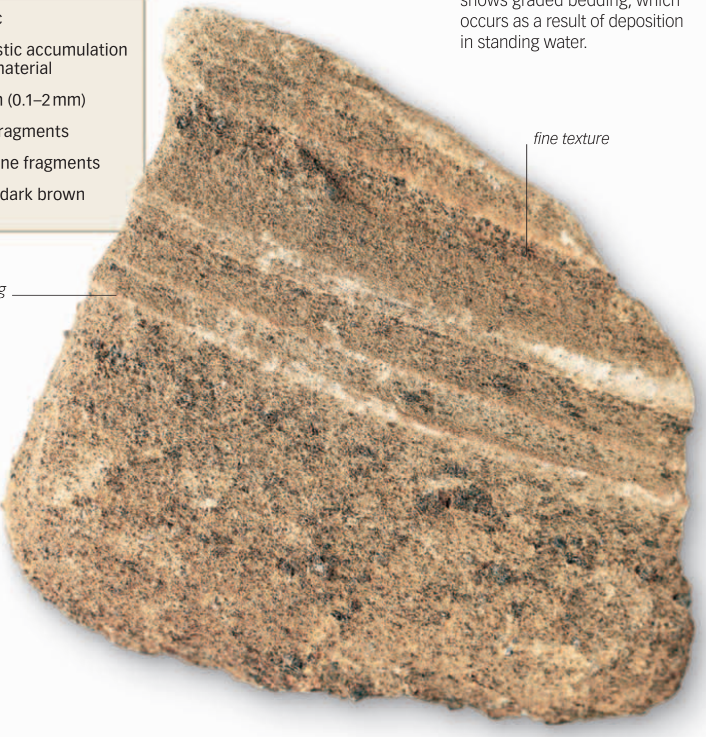

Tuff with graded bedding This tuff specimen from Ireland shows graded bedding, which occurs as a result of deposition in standing water.

TUFF

Any relatively soft, porous rock made of ash and other sediments ejected from volcanic vents that has solidified into rock is known as tuff. Most tuff formations include a range of fragment sizes and varieties. These range from fine-grained dust and ash (ash tuffs) to medium-sized fragments called lapilli (lapilli tuffs) to large volcanic blocks and bombs (bomb tuffs). Tuffs originate when foaming magma wells to the surface as a mixture of hot gases and incandescent particles and is ejected from a volcano.

The conditions under which the ejected ash solidifies determine the final nature of the tuff. Tuffs can vary both in texture and in chemical and mineralogical composition because of variations in the conditions of their formation and the composition of the ejected material. If the pyroclastic material is hot enough to fuse, a welded tuff (called ignimbrite) forms at once. Other tuffs lithify slowly through compaction and cementation, and can stratify when they accumulate under water.

VARIANTS

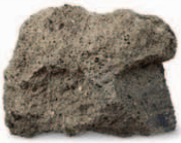

Lithic tuff A specimen of tuff containing a high percentage of small rock fragments

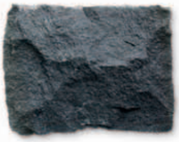

Crystal tuff A specimen of tuff containing a predominance of crystal fragments

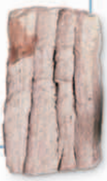

Bedded tuff A specimen of tuff that has fallen in distinct layers

PROFILE

- Extrusive
- Mixing of liquid and solid material during crystallization of a basic magma
- Less than 1/256 in (0.1 mm), clasts 3/16–8 in (0.5–20 cm)
- Various
- Various
- Red, brown, black

Rhyolite breccia This specimen of volcanic breccia incorporates angular fragments of reddish rhyolite.

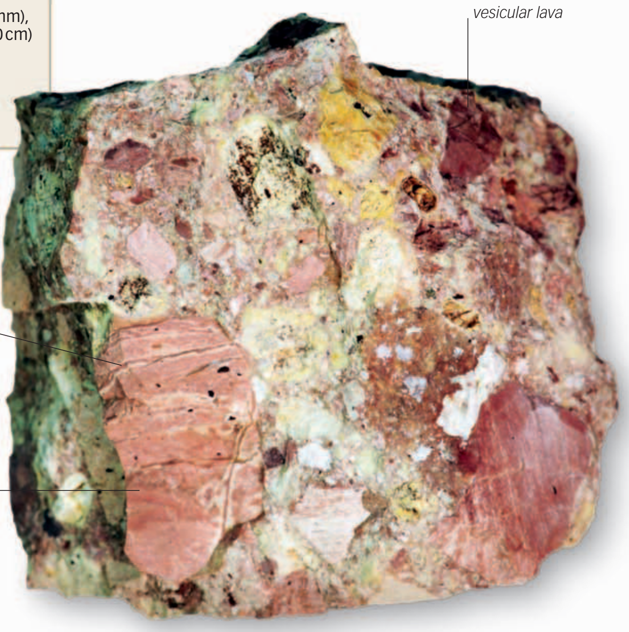
vesicular lava
flow banding from original rhyolite deposit
reddish rhyolite fragment

VARIANT

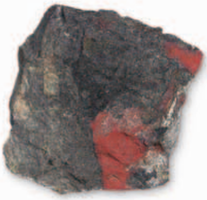

Volcanic breccia A specimen of volcanic breccia that contains large clasts of other volcanics

VOLCANIC BRECCIA

These igneous rocks are formed either by the interaction of lava and scoria (p.276) or by the mixing of cooled lava and flowing lava. Volcanic breccia takes the form of inch-scale angular clasts, which may be rocks broken off the side of a magma conduit or rocks picked up off the surface during a pyroclastic or lava flow. In certain types of lavas, especially dacite (p.274) and rhyolite (p.278) lavas, thick and nearly solidified lava is broken into blocks and then reincorporated into the flow of liquid lava. Flowing lavas can also pick up surface rocks and incorporate them into a solidified breccia. In explosive volcanoes, solidified lava may be reshattered numerous times to be reconstituted as breccias.

Flowtop breccia commonly forms at the top of a lava flow where the moving lava picks up loose debris from previous eruptions and flows. It is especially common between basaltic lava flows, which may occur some time apart. Breccias are different from agglomerates (p.284), in which the clasts are rounded.

PROFILE

- Extrusive
- Pyroclastic accumulation of coarse material
- Less than $\frac{1}{256}$ in (0.1 mm), clasts $\frac{3}{16}$–8 in (0.5–20 cm)
- Igneous rock fragments
- None
- Various

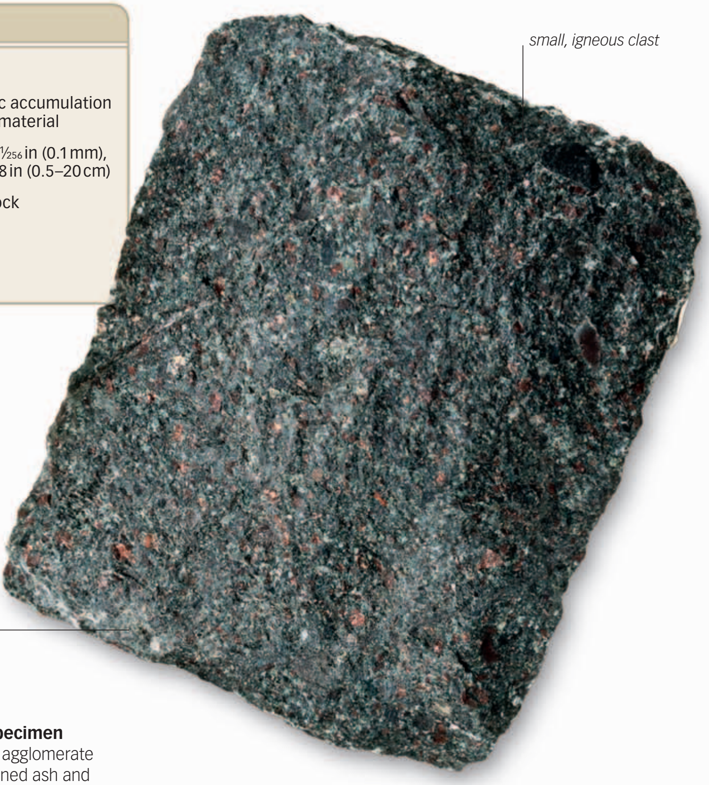

Agglomerate specimen
This specimen of agglomerate contains fine-grained ash and small clasts of other igneous rocks.

VARIANT

red dolomite

Carbonatite agglomerate
A specimen of carbonatite with clasts of red dolomite

AGGLOMERATE

An agglomerate is a pyroclastic rock in which coarse, rounded clasts up to several inches long are set in a matrix of lava or ash. The clasts are fragments that may be derived from lava, pyroclastic rock, or country rock (the rock that surrounds or lies beneath a volcano). The rounding of the clasts may have occurred either in the magma during eruption or by later sedimentary reworking. The rounded nature of these clasts is the key to designating the rock as an agglomerate rather than as a volcanic breccia (p.283). In a volcanic breccia, most of the clasts are angular.

One type of agglomerate, vent agglomerate, is the rock that plugs either the main vent or a satellite vent of a volcano. The outcrop of this rock is of limited extent and appears circular on a geological map. Like other agglomerates, vent agglomerate contains a variety of clasts of different sizes, shapes, and compositions from the lava, other volcanic rocks, or country rocks. These clasts lie in a matrix of fine-grained igneous rock.

PROFILE

- Extrusive
- Mid-air cooling of masses of silica-poor magma
- ¾ in–3¼ ft (2 cm–1 m)
- Basalt
- None
- Dark shades of red, brown, or green

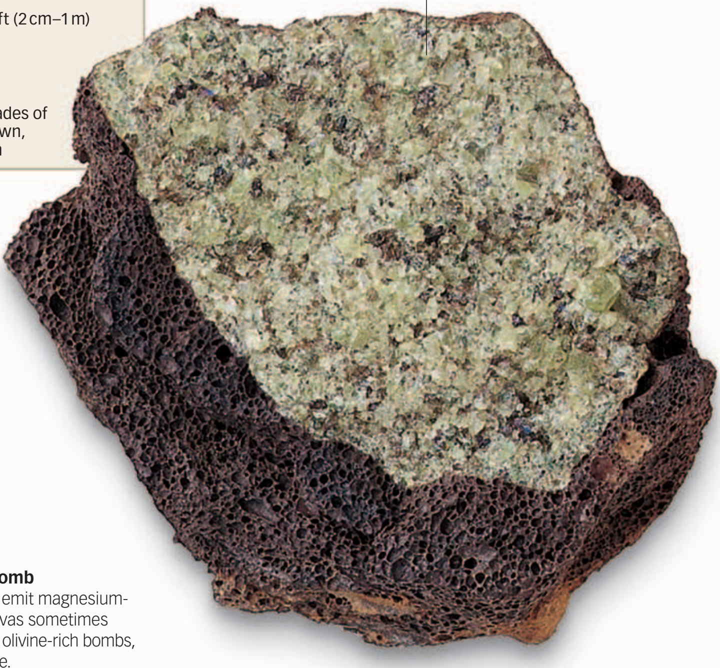

Olivine-rich bomb Volcanoes that emit magnesium- and iron-rich lavas sometimes explosively hurl olivine-rich bombs, such as this one.

VARIANTS

Spindle bomb Lava twisted by aerodynamic forces

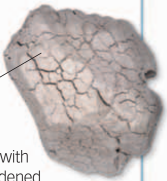

Breadcrust bomb Bomb with crust that hardened before landing

elongated shape

Spiky spindle A spindle bomb violently spun after ejection

VOLCANIC BOMB

Formed by the cooling of a mass of lava while it flies through the air after eruption, a volcanic bomb is a pyroclastic rock. To be called a bomb, a specimen must be larger than 2½ in (6.5 cm) in diameter; smaller specimens are known as lapilli. Specimens up to 20 ft (6 m) in diameter are known. Volcanic bombs are usually brown or red, weathering to a yellow-brown color. Specimens can become rounded as they fly through the air, although they may also be twisted or pointed. They may have a cracked, fine-grained, or glassy surface.

There are several types of volcanic bomb, which are named according to their outward appearance and structure. Spherical bombs are spheres of fluid magma pulled into shape by surface tension. Spindle bombs are formed by the same process as spherical bombs, except that their rotation in flight leaves them elongated. A breadcrust bomb forms if the outside of the lava bomb solidifies during flight and develops a cracked outer surface while the interior continues to expand.

METAMORPHIC ROCKS

Metamorphism occurs when an existing rock is subjected to pressures or temperatures very different from those under which it formed. This causes its atoms and molecules to rearrange themselves into new minerals in the solid state, without melting.

DYNAMIC METAMORPHISM

There are three different ways in which metamorphic rocks are formed. The first of these is dynamic metamorphism. This occurs as a result of large-scale movements in Earth's crust, especially along fault planes and at continental margins where tectonic plates collide. The resulting mechanical deformation produces angular fragments to fine-grained, granulated, or powdered rocks. These rocks are characterized by a foliated appearance, in which mineral grains align as parallel plates.

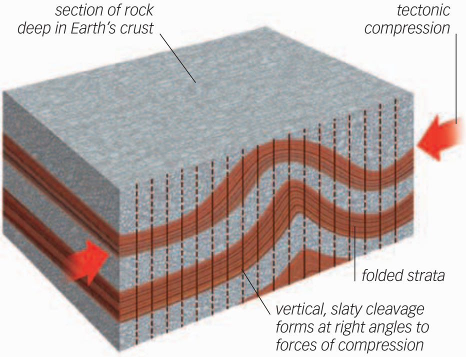

section of rock deep in Earth's crust
tectonic compression
folded strata
vertical, slaty cleavage forms at right angles to forces of compression

Dynamic metamorphism
Tectonic forces transform sedimentary rocks by dynamic metamorphism. Rock strata fold and cleavages develop as minerals align themselves due to the pressure.

REGIONAL METAMORPHISM

The second type of metamorphism is the formation of regional metamorphic rocks. These are associated with mountain building through the collision of tectonic plates. This process increases temperature and pressure over an area of thousands of square miles, producing widespread metamorphism. Important regional metamorphic rocks include slates, schists, and gneisses. Which rock forms depends on the existing rock, the temperatures and pressures to which it is subjected, and the time spent under those conditions. At the extremes, low temperature and pressure produces slates; high temperature and pressure produces gneisses.

Ancient metamorphic rocks
The Elk Mountains in Colorado, USA, have extensive areas of metamorphic rock, including schists and gneisses thought to have metamorphosed 1.7–1.9 billion years ago.

CONTACT METAMORPHISM

The third type of metamorphism is contact metamorphism or thermal metamorphism. This type occurs mainly as a result of increases in temperature, not in pressure. It is common in rocks near an igneous intrusion. Heat from the intrusion alters rocks to produce an "aureole" of metamorphic rock. The rocks nearest the intrusion are subjected to higher temperatures than those farther away, resulting in concentric zones of distinctive metamorphic rocks. The minerals of each zone depend on the original composition of the host rocks.

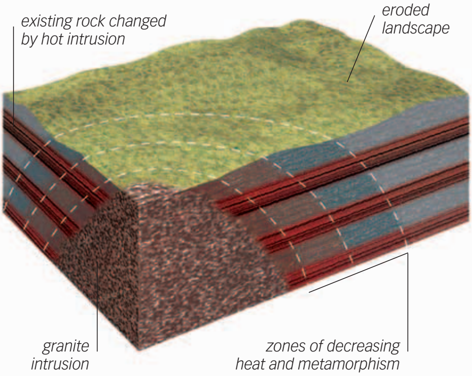
existing rock changed by hot intrusion
eroded landscape
granite intrusion
zones of decreasing heat and metamorphism

Contact metamorphism
A large intrusion of an igneous rock, such as granite, releases heat into the surrounding rocks, altering their mineral content.

CHANGING CHARACTERISTICS

Metamorphism is said to be low grade if it occurs at relatively low temperature and pressure and high grade at the intense end of the temperature and pressure range. The assemblages of minerals in rocks are affected differently depending on the grade of metamorphism and the relative importance of pressure and temperature in the reaction. In some low-grade reactions, the components of existing mineral assemblages are simply redistributed. In other reactions at higher temperatures and pressures, components combine with others present in the rock to form an entirely new set of minerals.

Shale to gneiss
This sequence shows how shale, a sedimentary rock, can be metamorphosed into various other rocks by the application of increasingly high degrees of heat and pressure (from left to right).

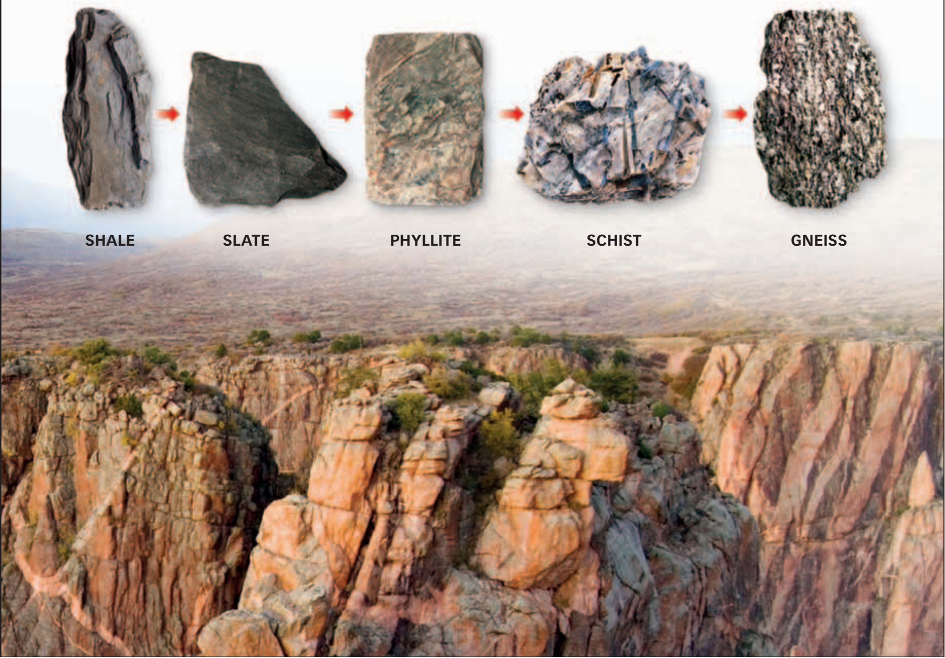

PROFILE

Regional metamorphic

High-grade metamorphism of rocks containing quartz and feldspar

High

High

Foliated, crystalline

1⁄16–3⁄16 in (2–5 mm)

Quartz, feldspar

Biotite, hornblende, garnet, staurolite

Gray, pink, multicolored

Granodiorite, mudstone granite, shale, siltstone, or felsic volcanics

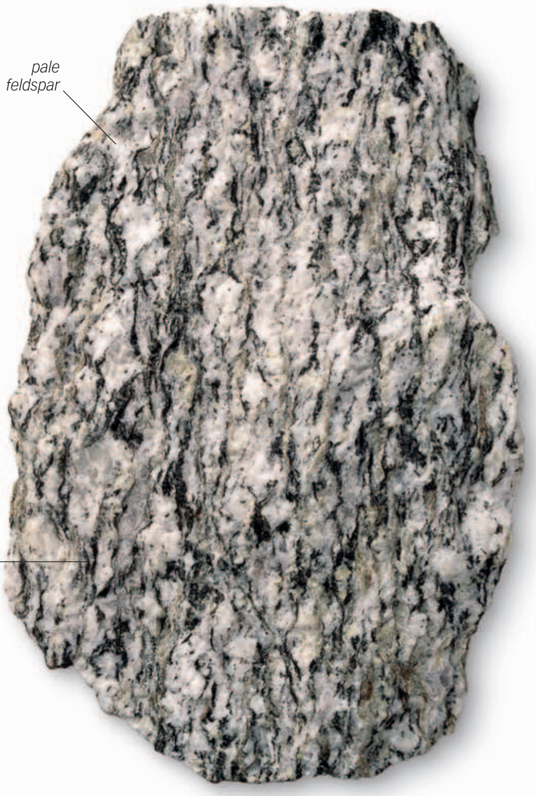

Banded gneiss
This specimen of classic gneiss shows foliated banding of light and dark minerals.

VARIANTS

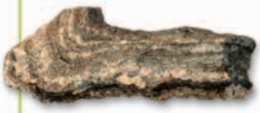

Folded gneiss A specimen with alternating mineral bands and typical folding

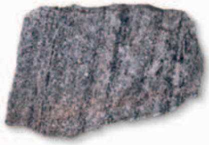

Orthogneiss Gneiss from metamorphosed igneous rocks

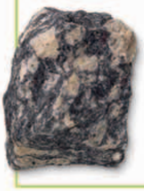

Augen gneiss A specimen with large “eyes” of light-colored feldspar

GNEISS

Distinct bands of minerals of different colors and grain sizes characterize this metamorphic rock. In most gneisses, these bands are folded, although the folds may be too large to see in hand specimens. Gneiss is a medium- to coarse-grained rock. Unlike schist (pp.291–92), its foliation is well developed, but it has little or no tendency to split along planes. Most gneisses contain quartz (p.168) and feldspar (pp.173–81), but neither mineral is necessary for a rock to be called gneiss. Larger crystals of metamorphic minerals, such as garnet, can also be present.

Gneiss makes up the cores of many mountain ranges. It forms from sedimentary or granitic (pp.258–59) rocks at very high pressures and temperatures (1,065°F/575°C or above). A variety called pencil gneiss has rod-shaped individual minerals or mineral aggregates. In augen gneiss, the augens or “eyes” are single-mineral, eye-shaped grains that are larger than other grains in the rock. Orthogneiss is gneiss derived from igneous rock, and paragneiss is gneiss derived from sedimentary rock.

PROFILE

- Dynamic thrust zones
- Stretching of a rock in a large fault
- Low
- Shearing stress
- Streaked out
- Less than 1⁄16 in (2 mm)
- As surrounding rock
- As surrounding rock
- As surrounding rock
- Surrounding rock

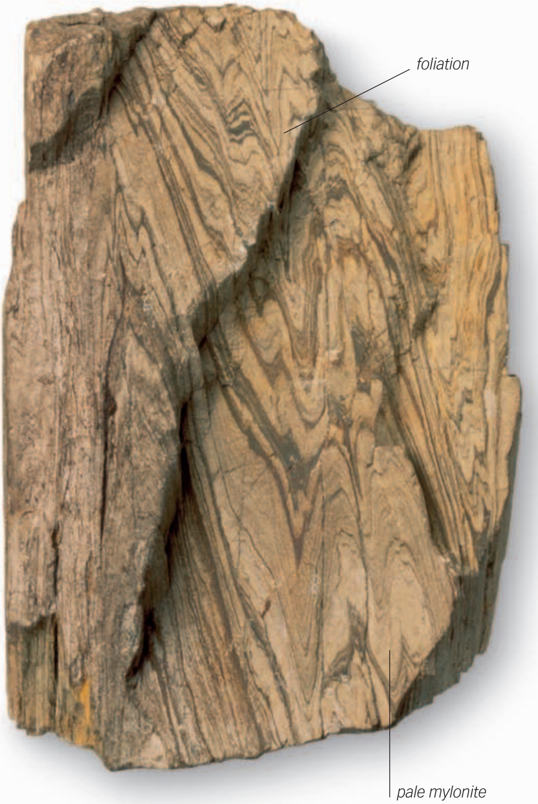

Deformed mylonite
The folded bands in this mylonite specimen indicate that it has been subjected to extreme deformation.

VARIANT

fine grains

Granular mylonite
A specimen with stretched mineral grains

MYLONITE

The term mylonite refers to fine-grained rocks with streaks or rodlike structures produced by the ductile deformation, or stretching, of mineral grains. This classification is based only on the texture of the rock, and specimens can have different mineral compositions. Mylonite with a large percentage of phyllosilicate minerals, such as chlorite or mica, is known as phyllonite. When mylonite is hard, dark, and so fine that it has the appearance of streaky flint, it is known as ultramylonite. Although generally fine grained, a few mylonites are coarse grained and often sugary in appearance. These are referred to as blastomylonites.

There are many different views on the formation of mylonite. It is typically produced in a zone of thrusts or low-angle faults. Fine-grained mylonites may have been produced by recrystallization under pressure. The fact that mylonite grains are stretched rather than sheared makes it evident that the rock has softened in the metamorphic process.

PROFILE

- Regional metamorphic
- Partial melting of rocks containing quartz and feldspar
- High
- High
- Foliated, crystalline
- 1/16–3/16 in (2–5 mm)
- Quartz, feldspar, mica
- Various
- Banded light and dark gray, pink, white
- Various, including granite

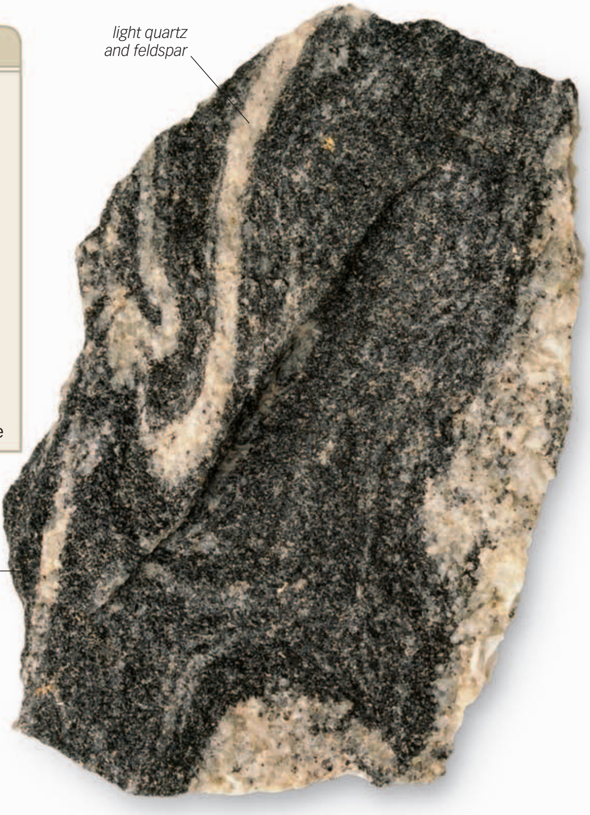

Mixed migmatite The mixing of light igneous and dark metamorphic mineral elements is evident in this specimen of migmatite.

VARIANTS

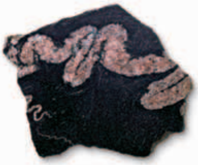

Partial melting Migmatite with a snakelike vein of granite, which indicates partial melting

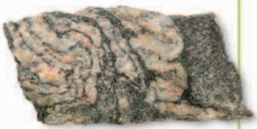

Migmatite folding A specimen of migmatite with distortions produced by extreme temperatures

MIGMATITE

The term migmatite means "mixed rock" and refers to rocks that consist of gneiss (p.288) or schist (pp.291–92) interlayered, streaked, or veined with granite (pp.258–59). The granitic parts consist of granular patches of quartz (p.168) and feldspar (pp.173–81), and the gneissic parts consist of quartz, feldspar, and dark-colored minerals. The granite streaks are a result of the partial melting of the parent rock at temperatures below the melting point of the schist or gneiss. The layering may be tightly folded as a result of softening during heating. Migmatites occur at the borderline between igneous and metamorphic rocks.

The rock forms near large intrusions of granite when some of the magma has intruded into the surrounding metamorphic rocks. Commonly, migmatite occurs within extremely deformed rocks that once formed the bases of eroded mountain chains. It forms deep in the crust at high temperatures (1,065°F/575°C or above) and pressures.

PROFILE

- Regional metamorphic
- Regional metamorphism of fine-grained sediments
- Low to moderate
- Low to moderate
- Foliated
- 1/256–1/16 in (0.1–2 mm)
- Quartz, feldspar, mica
- Garnet, hornblende, actinolite, graphite, kyanite
- Silvery, green, blue
- Mudstone, siltstone, shale, or felsic volcanics

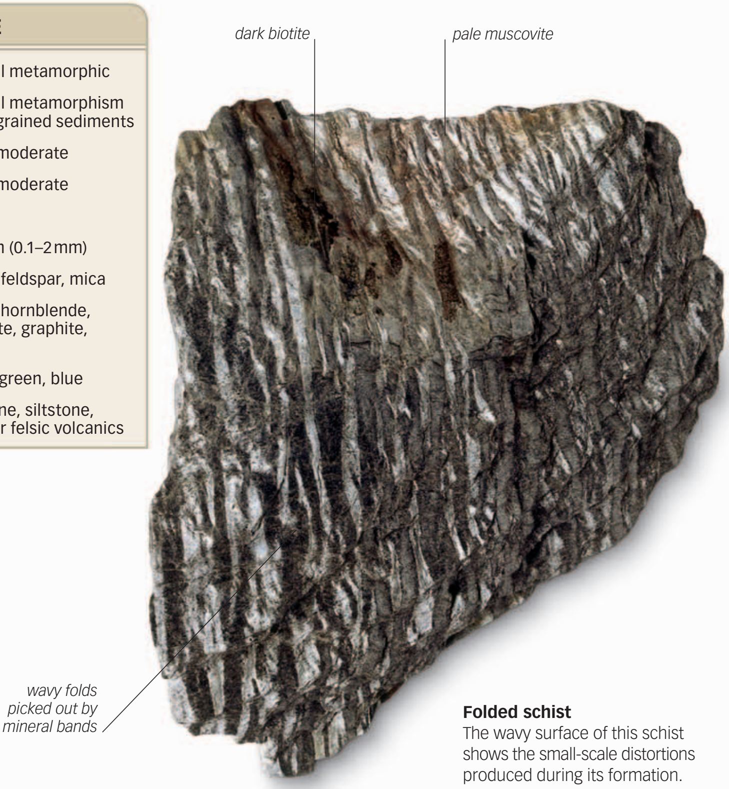

Folded schist The wavy surface of this schist shows the small-scale distortions produced during its formation. It has split along its mica bands.

VARIANTS

Muscovite schist A specimen of schist dominated by white muscovite mica

Blue schist Schist colored blue by glaucophane

Kyanite schist Small, blue blades of kyanite in schist

SCHIST

This metamorphic rock has a flaky and foliated texture. Specimens have wrinkled, wavy, or irregular sheets as a result of the parallel orientation of the component minerals. Schist shows distinct layering of light- and dark-colored minerals. The mineral assemblage varies, but mica is usually present. Most schists are composed of platy minerals, such as chlorite, graphite (p.46), talc (p.193), muscovite (p.195), and biotite (p.197). The mineral composition of a schist depends on its protolith or original rock and its metamorphic environment. The mineral assemblage can thus be used to determine the metamorphic history of the rock.

Indian schist carving Although its texture and composition are often uneven, schist is sometimes used as a carving material.

PROFILE

- Regional metamorphic
- Medium-grade metamorphism of silica-rich rocks
- Low to moderate
- Moderate
- Foliated
- 1⁄16–3⁄16 in (2–5 mm)
- Muscovite, biotite, garnet
- Feldspar, staurolite, sillimanite, kyanite, cordierite
- Light to dark
- Silica-rich rocks

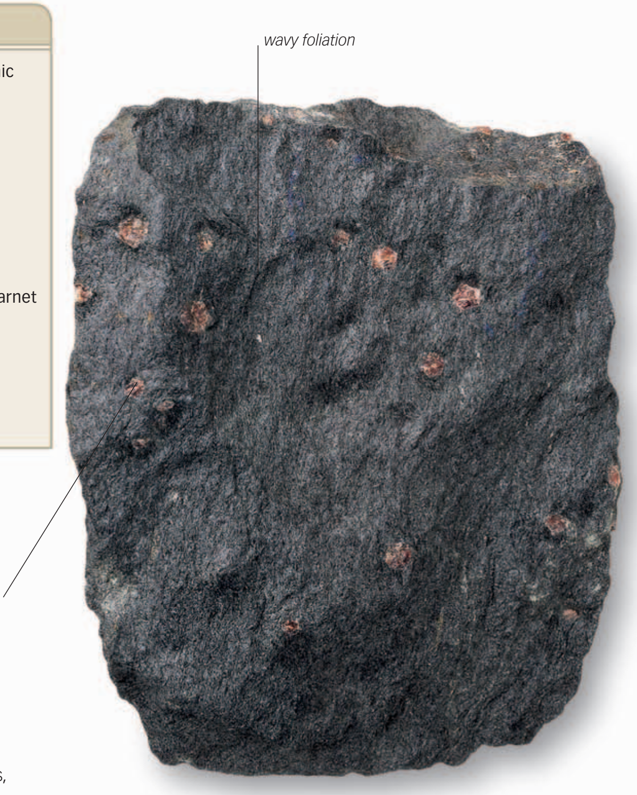

Garnet porphyroblasts Garnet is commonly present in schists, forming large crystals called porphyroblasts, as seen here.

VARIANTS

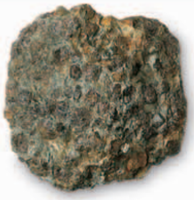

Garnet-chlorite schist Garnet porphyroblasts in chlorite schist

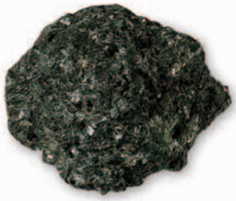

Garnet-muscovite-chlorite schist Garnet porphyroblasts in a specimen of muscovite-chlorite schist

GARNET SCHIST

Like other schists, garnet schist is a metamorphic rock with a characteristic texture: wrinkled, irregular, or wavy as a result of the parallel orientation of its component minerals, such as chlorite, graphite (p.46), talc (p.193), muscovite (p.195), and biotite (p.197). In garnet schist, garnet occurs as porphyroblasts, which are large crystals set in a metamorphic matrix with other smaller crystals. The resulting texture is called porphyroblastic. The equivalent texture in igneous rocks is called porphyritic.

Garnet schist is widespread and usually forms from the metamorphism of fine-grained sediments, especially during the formation of mountains. The mineral assemblage in garnet schist, like in other schists, helps determine both the environment in which the original rock formed and its metamorphic history. Other porphyroblastic schists—such as staurolite schist, corundum schist, kyanite schist, muscovite schist, biotite schist, and schists of other metamorphic minerals—are indicative of other metamorphic histories.

PROFILE

- Regional metamorphic
- Low-grade regional metamorphism of fine-grained sediments
- Low
- Low
- Foliated
- Less than 1/256 in (0.1 mm)
- Quartz, mica, feldspar
- Pyrite, graphite
- Various
- Mudstone, siltstone, shale, or felsic volcanics

Welsh slate This dark gray specimen is from Wales, Britain's principal source of slate. In the USA, slate is quarried in Pennsylvania and Vermont.

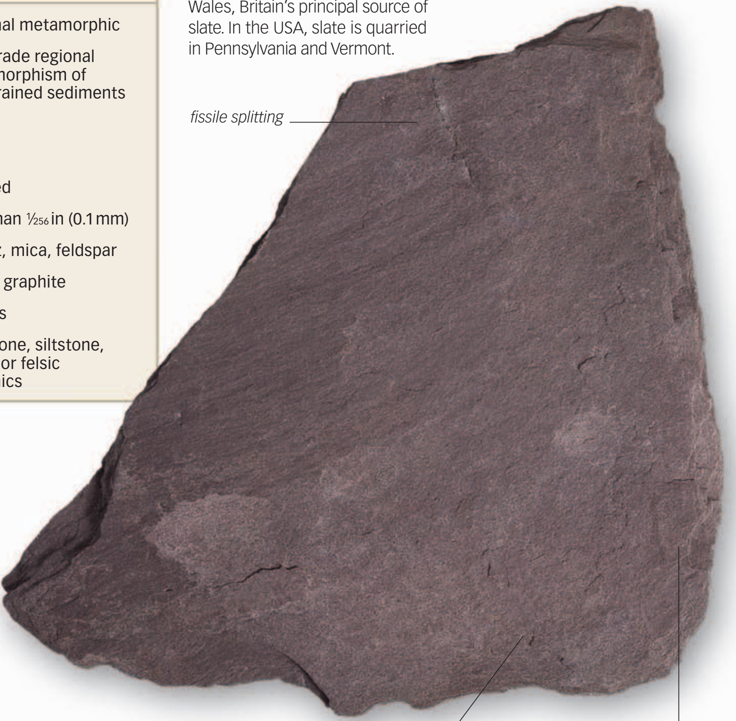

VARIANTS

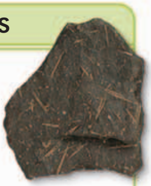

Chiastolite slate High-temperature slate with chiastolite crystals

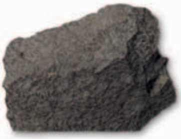

Spotted slate Slate with small aggregates of carbon

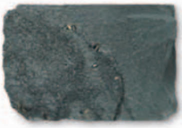

Pyrite Slate with pyrite grains and porphyroblasts

SLATE

A fine-grained metamorphic rock, slate occurs in a number of colors that depend on the minerals in the original sedimentary rock and the oxidation conditions under which that rock formed. Slate has a characteristic cleavage that allows it to be split into relatively thin, flat sheets. This is a result of microscopic mica crystals that have grown oriented in the same plane. True slates split along the foliation planes formed during metamorphism, rather than along the original sedimentary layers.

Slate is common in regionally metamorphosed terrains. It forms when shale (p.313), mudstone (p.316), or volcanic rocks rich in silica are buried as well as subjected to low pressures and temperatures (up to 400°F/200°C). The ability of slate to split into thin sheets makes it ideal as a durable roofing material.

School slate This child's school slate is from Victorian times. In the past, all school "blackboards" were made from slate.

PROFILE

Regional metamorphic

Regional metamorphism of fine-grained sediments

Low to moderate

Low

Foliated

Less than 1⁄256 in (0.1 mm)

Quartz, feldspar, chlorite, muscovite mica, graphite

Tourmaline, andalusite, cordierite, biotite, staurolite

Silvery to greenish gray

Mudstone, siltstone, shale, or felsic volcanics

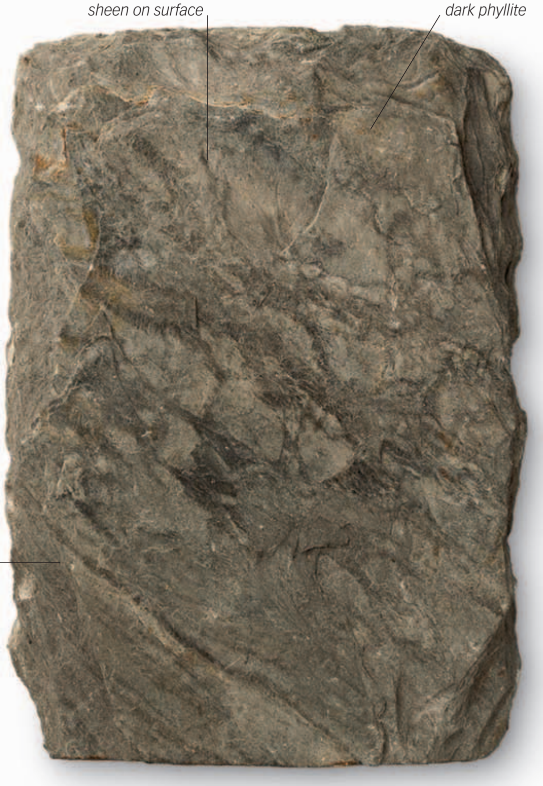

Irregular surface This example of phyllite shows bands of minerals and a cleaved surface that is more irregular than that found in slate.

VARIANTS

Coarse foliation An example of coarsely foliated phyllite

Alternating minerals Phyllite with bands of minerals

Garnet porphyroblasts Small, dark garnets on foliated surface

PHYLLITE

Like slate (p.293), phyllite is a fine-grained metamorphic rock that is usually gray or dark green in color. It has a shinier sheen than slate because of its larger mica crystals. The rock has a tendency to split in the same manner as slate because of a parallel alignment of mica minerals. However, the split surfaces are more irregular than in slate, and phyllite splits into thick slabs rather than thin sheets. Many phyllite specimens have a scattering of large crystals called porphyroblasts, which grow during metamorphism. The rock is often deformed into folds a couple of inches wide and is veined with quartz (p.168). Biotite (p.197), cordierite (p.223), tourmaline (p.224), andalusite (p.236), and staurolite (p.239) are commonly found in phyllite.

This rock occurs in both young and old eroded mountain belts in regionally metamorphosed terrains. It forms when fine-grained sedimentary rocks, such as shales (p.313) or mudstones (p.316), are buried and subjected to relatively low pressures and temperatures (up to 400°F/200°C) for a long period of time.

PROFILE

- Regional metamorphic
- Regional metamorphism of orthoquartzite
- High
- Low to high
- Crystalline
- 1⁄16–3⁄16 in (2–5 mm)
- Quartz
- Mica, kyanite, sillimanite
- Almost any
- Sandstone

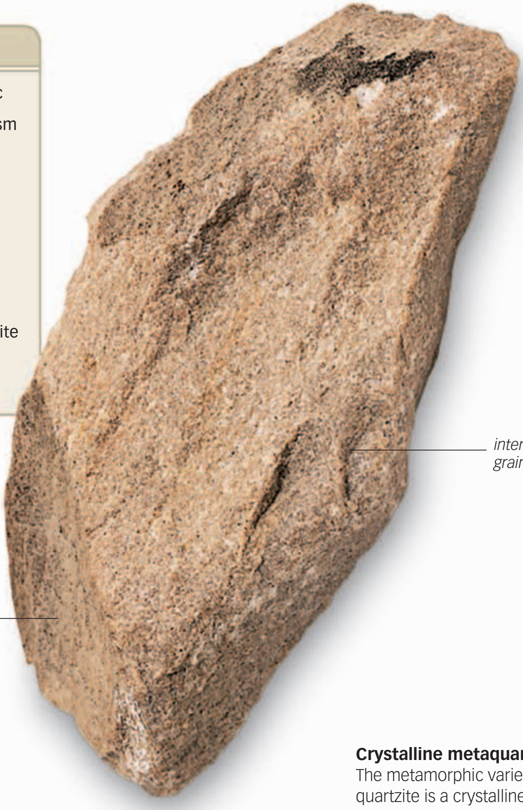

Crystalline metaquartzite
The metamorphic variety of quartzite is a crystalline rock with over 90 percent quartz.

VARIANTS

Pale metaquartzite
A specimen of metaquartzite with a very high percentage of quartz

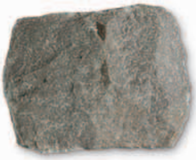

Gray metaquartzite
A specimen that reflects the color of the original sandstone

QUARTZITE

This quartz-rich metamorphic rock is usually white to gray when pure. Specimens can be various shades of pink, red, yellow, and orange when mineral impurities are present. Quartzite is very hard and brittle and shows conchoidal fracture. It usually contains at least 90 percent quartz (p.168).

Metamorphic quartzite, or metaquartzite, forms when sandstone (p.308) is buried, heated, and squeezed into a solid quartz rock. It is found with other regional metamorphic rocks formed during the shifting of Earth's tectonic plates. Quartzite can also refer to sedimentary sandstone converted to a much denser form through the precipitation of silica cement in pore spaces. The two types of quartzite can be distinguished by examining the grains under a microscope: metamorphic quartzite consists of interlocking crystals of quartz, whereas sedimentary quartzite, or orthoquartzite, contains rounded quartz grains. Crushed quartzite is often used as railroad ballast and in the construction of roads.

PROFILE

Regional metamorphic

High-grade metamorphism of silica-poor igneous rocks

Low to moderate temperatures (up to 1,065°F/575°C)

Low to moderate

Foliated, crystalline

1⁄16–3⁄16 in (2–5 mm)

Hornblende, tremolite, actinolite

Feldspar, calcite, garnet, pyroxene

Gray, black, greenish

Basalt, graywacke, dolomite

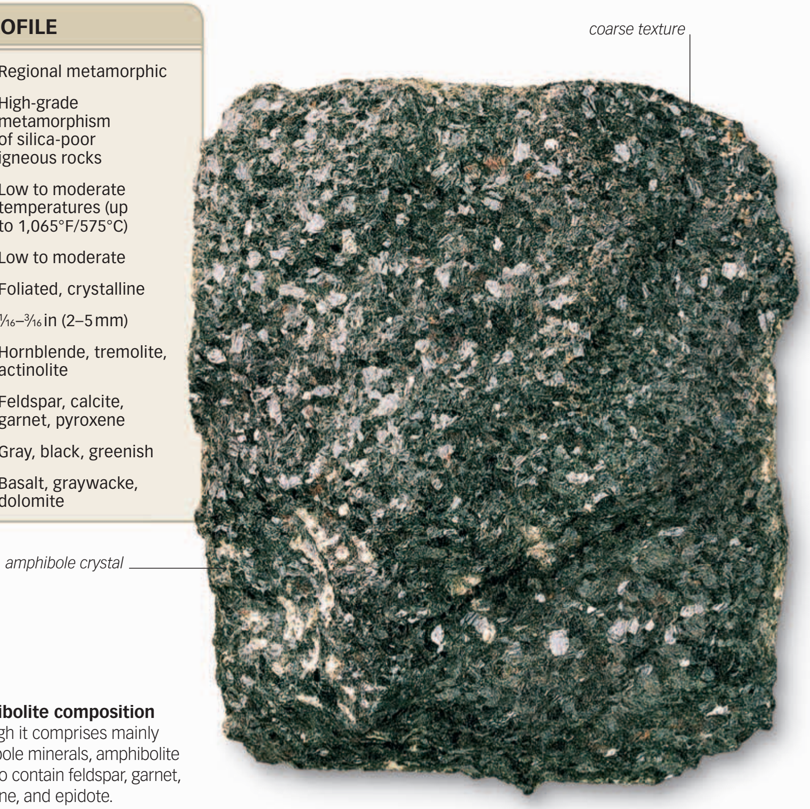

Amphibolite composition Although it comprises mainly amphibole minerals, amphibolite can also contain feldspar, garnet, pyroxene, and epidote.

VARIANTS

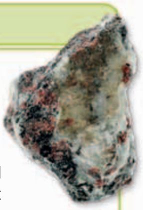

Light-colored amphibolite An unusual amphibolite with light-colored minerals present

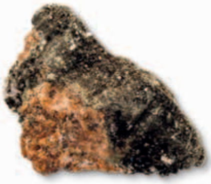

Hornblende-rich amphibolite A hornblende-rich specimen of amphibolite

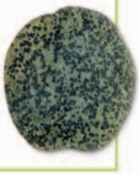

Green amphibolite Amphibolite with many garnet porphyroblasts

AMPHIBOLITE

As the name suggests, amphibolites are dark-colored, coarse-grained rocks that are dominated by amphiboles: the black or dark green hornblende (p.218) and the green tremolite (p.219) or actinolite (p.220). Specimens may contain grains of calcite (p.114), feldspar (pp.173–81), and pyroxene and large crystals of minerals such as garnet. Except for garnet, the mineral grains in amphibolite are usually aligned. The rock can also show banding.

Amphibolites form from the metamorphism of iron- and magnesium-rich igneous rocks, such as gabbros (p.265), and from sedimentary rocks, such as graywacke (p.317). They comprise one of the major divisions of metamorphic rocks as classified by their mineral assemblages. These rocks form under conditions of low to moderate pressures and temperatures (up to 1,065°F/575°C). Amphibolites are used in building roads and in other aggregates where high degrees of strength and durability are required.

PROFILE

- Regional metamorphic
- High-grade metamorphism of silica-poor igneous rocks
- High temperatures (1,065°F/575°C or above)
- High
- Crystalline
- 1/16–3/16 in (2–5 mm)
- Feldspar, quartz, garnet, pyroxene
- Spinel, corundum
- Gray, pinkish, brownish, mottled
- Felsic igneous and sedimentary rocks

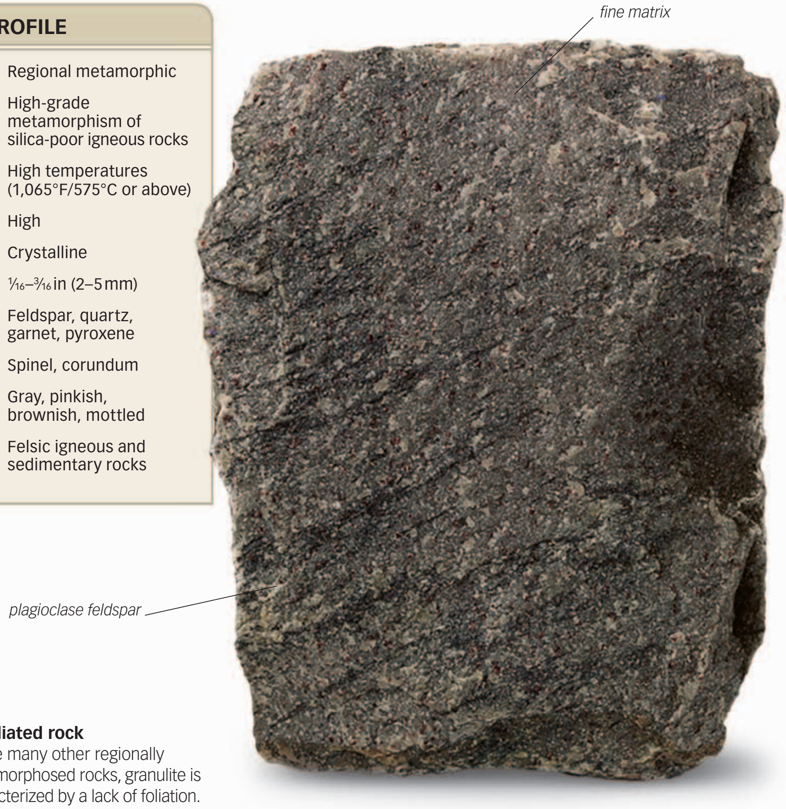

Unfoliated rock Unlike many other regionally metamorphosed rocks, granulite is characterized by a lack of foliation.

VARIANT

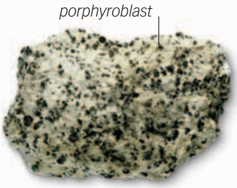

Light-colored granulite This specimen of granulite has numerous porphyroblasts

GRANULITE

This metamorphic rock is named for its even-grained, granular texture. Specimens are typically tough and massive. Granulite has a high concentration of pyroxene, with diopside (p.210) or hypersthene, garnet, calcium plagioclase, and quartz (p.168) or olivine (p.232). It has nearly the same minerals as gneiss (p.288) but is finer-grained and less perfectly foliated, and has more garnet.

Formed at high pressures and temperatures (1,065°F/575°C or above) deep in Earth's crust, granulites are characteristic of the highest grade of metamorphism. Rocks formed under these conditions belong to a category of metamorphic rocks known as the granulite facies. Mineral groups such as micas and amphiboles cannot survive at the high metamorphic grade under which granulites form and are converted into pyroxenes and garnets. Most granulites date from the Precambrian Age, which ended over 500 million years ago. They are of particular interest to geologists because many of them represent samples of the deep continental crust.

PROFILE

Regional

Water-rich, low-grade metamorphism of olivine-rich rocks

Low temperatures (up to 400°F/200°C)

Low

Nonfoliated

Less than 1/256 in (0.1 mm)

Serpentine

Chromite, magnetite, talc

Medium to dark

Olivine-rich igneous rocks

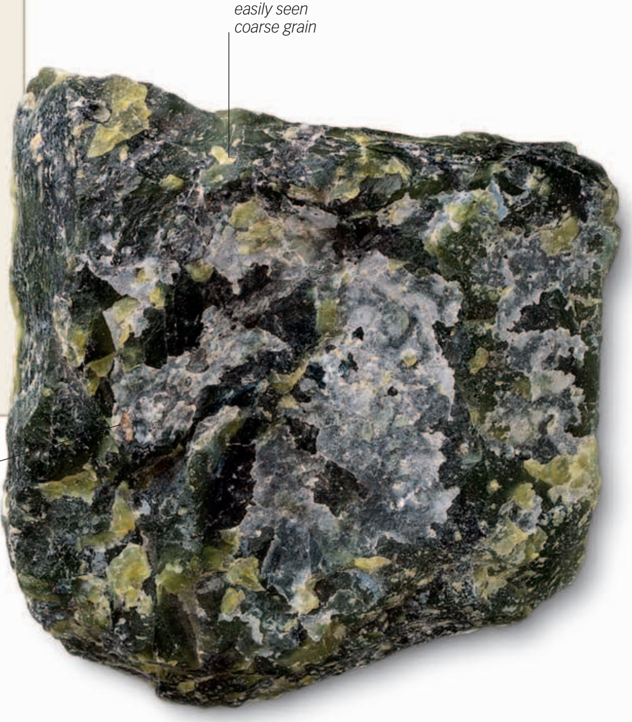

Grainy serpentinite This example of serpentinite clearly shows its fine grains.

VARIANT

Alpine serpentine Red and green serpentinite streaked with calcite from the Alps

SERPENTINITE

An attractive rock, serpentinite is composed of serpentine (p.191) and other serpentine-group minerals. It commonly has flowing bands of various colors, especially green and yellow. Serpentine minerals form by a metamorphic process called serpentinization that alters olivine and pyroxene-rich, silica-poor igneous rocks. This process occurs at low temperatures (up to 400°F/200°C) and in the presence of water. The original minerals are oxidized to produce serpentine, magnetite (p.92), and brucite (p.105). The degree to which a rock undergoes serpentinization depends on the composition of the parent rock and the mineral composition of its components, especially its olivine (p.232). For example, fayalite-rich olivines serpentinize differently than forsterite-rich olivines.

Serpentinite is used as a decorative stone since it can be easily cut and polished. It is also mixed into concrete aggregate and used as a dry filler in the steel shielding jackets of nuclear reactors.

PROFILE

- Regional
- High-pressure metamorphism of silica-poor igneous rocks
- High temperatures (1,065°F/575°C or above)
- High
- Foliated, crystalline
- $\frac{1}{16}$–$\frac{3}{16}$ in (2–5 mm)
- Pyroxene, garnet
- Kyanite, quartz, olivine, diopside
- Various colors
- Silica-poor igneous rocks

Distributed eclogite grains Eclogites, such as this one with red garnet and green pyroxene, are formed under great pressure, and are thought to originate in the mantle.

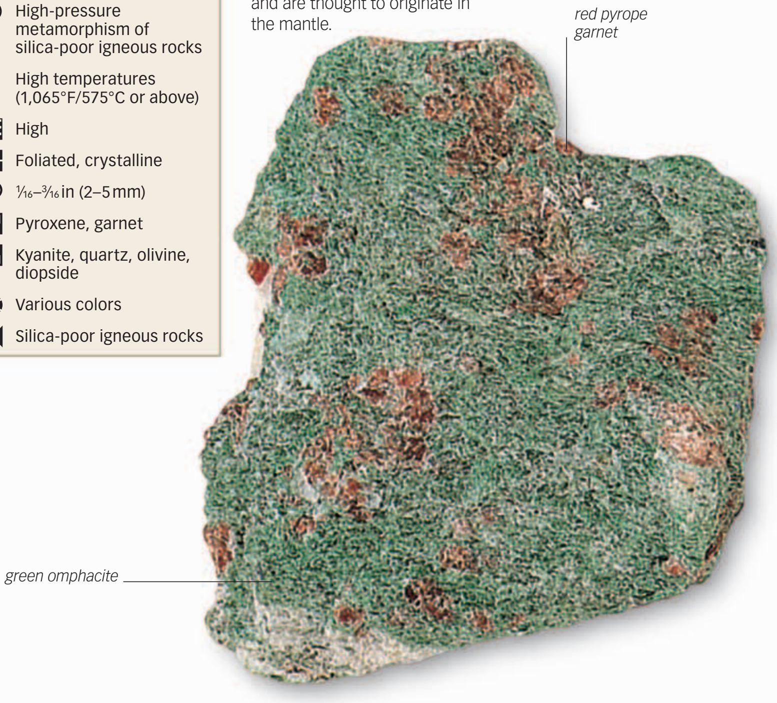

VARIANTS

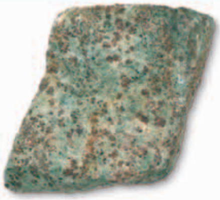

Mantle rock Eclogite with garnet porphyroblasts

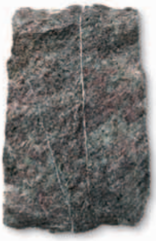

Fine-grained eclogite Garnet- and pyroxene-rich specimen of eclogite

ECLOGITE

A rare but important rock, eclogite is formed only by conditions typically found in the mantle or the lowermost and thickest part of the continental crust. Its overall chemical composition is similar to that of igneous basalt (p.273). It is a beautiful, coarse-grained, dense rock of bright red garnet and contrasting bright green omphacite—a pyroxene characteristic of high-temperature metamorphism. Diopside (p.210) and olivine (p.232) are commonly present as well. Eclogite grains may be evenly distributed or banded.

Eclogite forms at very high temperatures (1,065°F/575°C or above) and pressures from silica-poor igneous rocks. Many investigators believe that eclogite is characteristic of a considerable portion of the upper mantle. When brought upward into the crust, eclogites are mainly found as xenoliths (foreign inclusions) in igneous rocks and as isolated blocks up to 330 ft (100 m) wide in other metamorphic rocks. Although rare, diamonds (p.47) may occur in eclogites.

PROFILE

- ▲ Regional
- 🌐 Metamorphism of serpentine in presence of water
- 🌡 Low temperatures (up to 400°F/200°C)
- ☰ Low
- ▦ Foliated
- 🔍 Less than $\frac{1}{256}$ in (0.1 mm)
- ◼ Talc
- ◻ Chlorite, magnesite
- ✱ White, green, brown, black
- ⏪ Serpentine

Soapstone specimen This specimen of soapstone shows the foliation of the original talc.

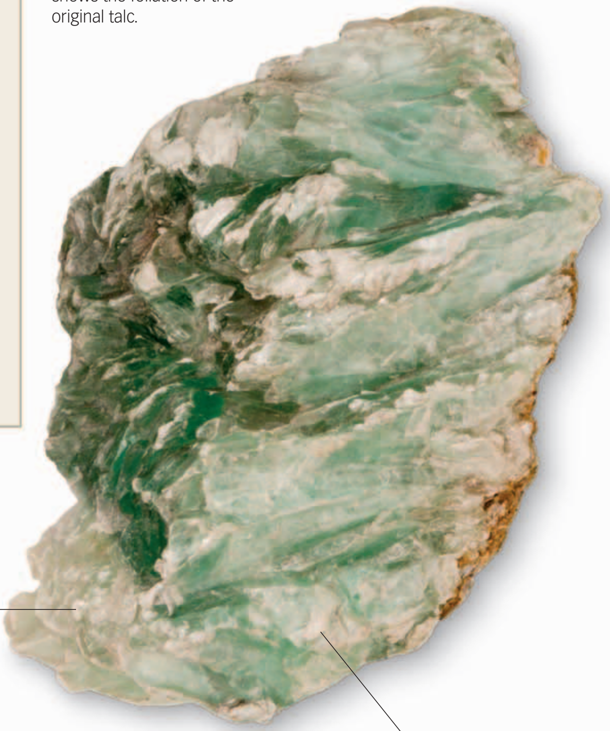

VARIANTS

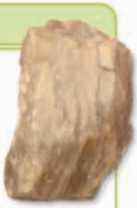

Foliated soapstone Soapstone with visible foliation

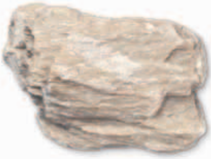

Flaky soapstone Soft soapstone with a flaky surface

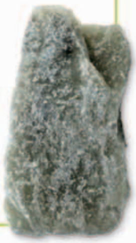

Green soapstone A specimen of soapstone consisting primarily of green talc

SOAPSTONE

Also known as steatite, soapstone is a fine-grained, massive rock. Talc (p.193), one of the softest minerals, is its principal component. Soapstone specimens are easily recognized by their softness: they have a greasy feel and can be scratched with a fingernail. Soapstone may also contain varying amounts of amphiboles, such as anthophyllite (p.216) and tremolite (p.219), and chlorites. Specimens can be green, brown, or black when polished, but they become white when scratched.

Soapstone is found associated with other metamorphosed silica-poor igneous rocks, such as serpentinite (p.298). It is used in fireplace surrounds because it absorbs and distributes heat evenly. It is also carved into molds for soft-metal casting.

Lion dog seal Soapstone, such as that used in this Chinese seal, has been used to carve ornaments for millennia.

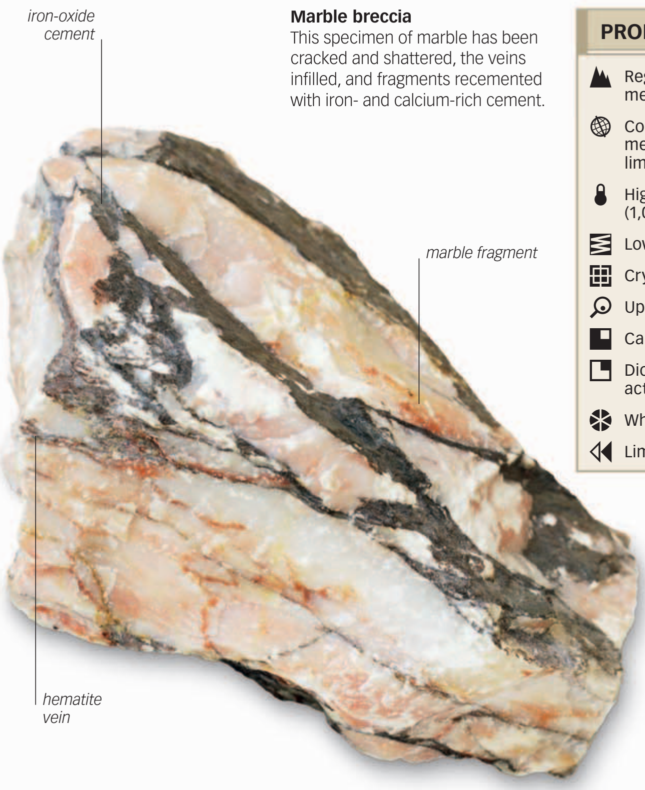

Marble breccia This specimen of marble has been cracked and shattered, the veins infilled, and fragments recemented with iron- and calcium-rich cement.

PROFILE

- ▲ Regional or contact metamorphic
- 🌐 Contact or regional metamorphism of limestone or dolomite
- 🌡 High temperatures (1,065°F/575°C or above)
- ≋ Low to high
- ⊞ Crystalline
- 🔍 Up to ¾ in (2 cm)
- ■ Calcite
- ◩ Diopside, tremolite, actinolite, dolomite
- ✱ White, pink
- ◀ Limestone, dolomite

VARIANTS

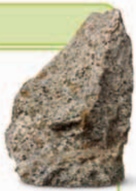

Olivine marble Marble with small crystals of olivine

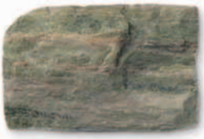

Green marble A specimen of marble colored by green silicate minerals

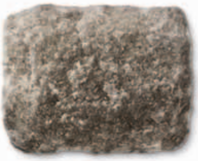

Gray marble Marble from relatively pure limestone

MARBLE

A granular metamorphic rock, marble is derived from limestone (p.319) or dolomite (p.320). It consists of a mass of interlocking grains of calcite (p.114) or the mineral dolomite (p.117).

Marbles form when limestone buried deep in the older layers of Earth's crust is subjected to heat and pressure from thick layers of overlying sediments. It may also form as a result of contact metamorphism near igneous intrusions. Impurities in the limestone can recrystallize during metamorphism, resulting in mineral impurities in the marble, most commonly graphite (p.46), pyrite (p.62), quartz (p.168), mica, and iron oxides. In sufficient amounts, these can affect the texture and color of the marble.

Taj Mahal, India The Taj Mahal is built of Makrana—a white marble that changes hue with the angle of the light.

PROFILE

- Contact
- Contact metamorphism of limestone
- High temperatures (1,065°F/575°C or above)
- Low
- Crystalline
- 1/256–3/4 in (0.1–20 mm)
- Calcite
- Garnet, serpentine, forsterite, vesuvianite
- Variegated
- Limestone, dolomite

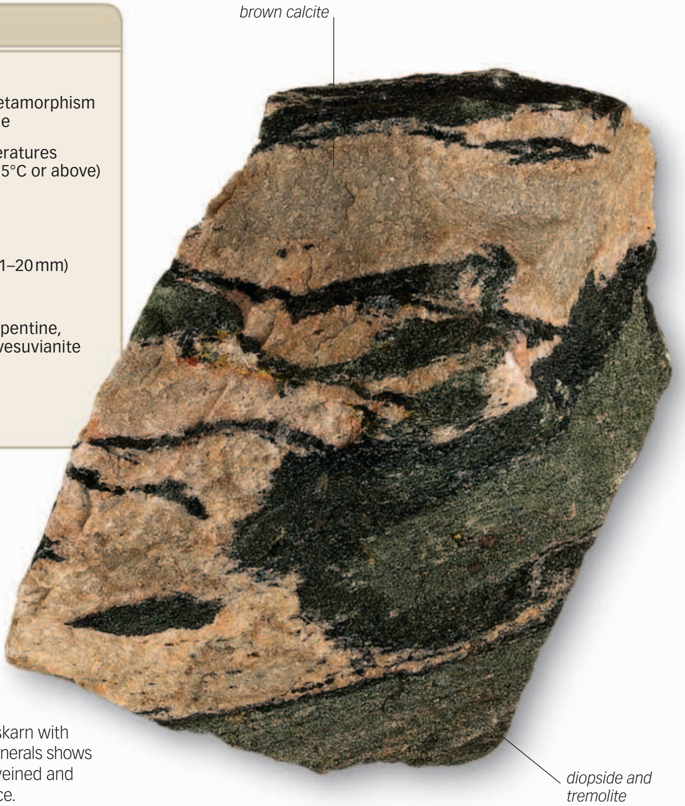

Skarn specimen
This specimen of skarn with calcite and dark minerals shows the rock's typical veined and banded appearance.

VARIANT

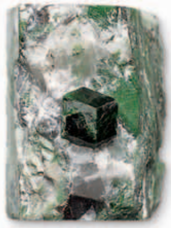

Garnet in skarn A uvarovite garnet porphyroblast in a skarn matrix

SKARN

This rock is a product of the contact metamorphism of limestone (p.319) or dolomite (p.320) by an igneous intrusion, often granite (pp.258–59), with an intermediate or high silica content. Hot waters derived from the granitic magma are rich in silica, iron, aluminum, sulfur, and magnesium. When limestone or dolomite is invaded by this high-temperature hydrothermal solution, the carbonate minerals calcite (p.114) and dolomite react strongly with the slightly acid solution. The elements carried in the solutions combine with the calcium and magnesium in the parent rock to form silicate minerals, such as diopside (p.210), tremolite (p.219), and andradite. The resulting rock is usually a highly complex combination of calcium-, magnesium-, and carbonate-rich minerals.

Skarn minerals can be fine- to medium-grained. They also occur as coarse, radiating crystals or bands. Some skarns are rich in metallic ores and form valuable deposits of metals, including gold (p.42), copper, iron, tin, lead, molybdenum, and zinc.

PROFILE

- Contact metamorphic
- Contact metamorphism of fine-grained sediment
- Moderate to high temperatures (400°F/ 200°C or above)
- Low to high
- Crystalline
- Less than 1/256 in (0.1 mm)
- Hornblende, plagioclase, andalusite, cordierite, and others
- Magnetite, apatite, titanite
- Dark gray, brown, greenish, reddish
- Almost any rock

dark, pyroxene crystals

Pyroxene hornfels In this specimen of hornfels, dark porphyroblasts of pyroxene can be seen.

VARIANTS

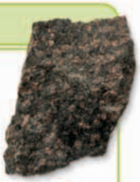
Garnet hornfels A specimen colored red by garnet crystals

Cordierite hornfels Hornfels with small grains of cordierite, quartz, and mica

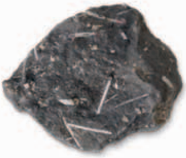
Chiastolite hornfels Hornfels with elongated porphyroblasts of chiastolite (andalusite)

HORNFELS

Formed by contact metamorphism close to igneous intrusions at temperatures as high as 1,300°–1,450°F (700°–800°C), hornfels can form from almost any parent rock and is notoriously difficult to identify. Its composition depends on the parent rock and the exact temperatures and fluids to which the rock is exposed. Specimens are usually dense, hard, and hard to break. They are fine-grained and relatively homogeneous, with a conchoidal, flintlike fracture. The rock may sometimes appear glassy. The color is usually even throughout, although specimens may also be banded.

Hornfels is often categorized by the mix of minerals present in the specimen. Garnet hornfels, for example, is characterized by large crystals of garnet set into a rock matrix. Cordierite hornfels contains large crystals of cordierite (p.223) that can be up to several inches in diameter. An outcrop of hornfels rarely extends more than a few yards from the contact and may pass outward into spotted slate.

SEDIMENTARY ROCKS

Sedimentary rocks are formed at or near Earth's surface either by accumulation of grains or by precipitation of dissolved material. These rocks make up the majority of the rock exposed at Earth's surface, but are only about 8 percent of the volume of the entire crust.

LITHIFICATION

The transformation of loose grains of sediment into clastic sedimentary rock is known as lithification. The grains are often bound together by a cementing agent, which is generally precipitated from solutions that filter through the sediment. In some cases, the cementing agent is created at least in part by the breakdown of some rock particles of the sediment itself. The most common cement is silica (usually quartz), but calcite and other carbonates as well as iron oxides, barite, anhydrite, zeolites, and clay minerals also form cements. The cementing agent becomes an integral and important part of the sedimentary rock once it is formed. In some cases, clastic sedimentary rocks (see opposite) can also be formed by simple compaction—a process in which the grains bind together under extreme pressure.

Lithification can sometimes take place almost immediately after the grains have been deposited as sediment. In other cases, hundreds or even millions of years may pass by before lithification occurs. At any given time, there are large amounts of rock fragments that have been produced by weathering but have not been lithified and simply exist as a form of sediment.

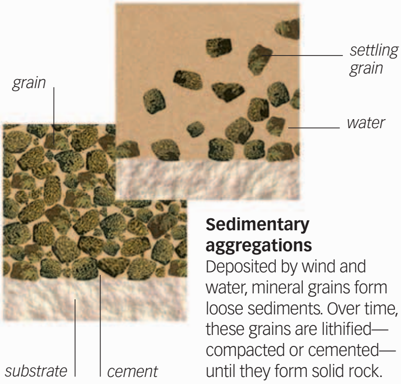
settling grain
grain
water
substrate
cement

Sedimentary aggregations
Deposited by wind and water, mineral grains form loose sediments. Over time, these grains are lithified—compacted or cemented—until they form solid rock.

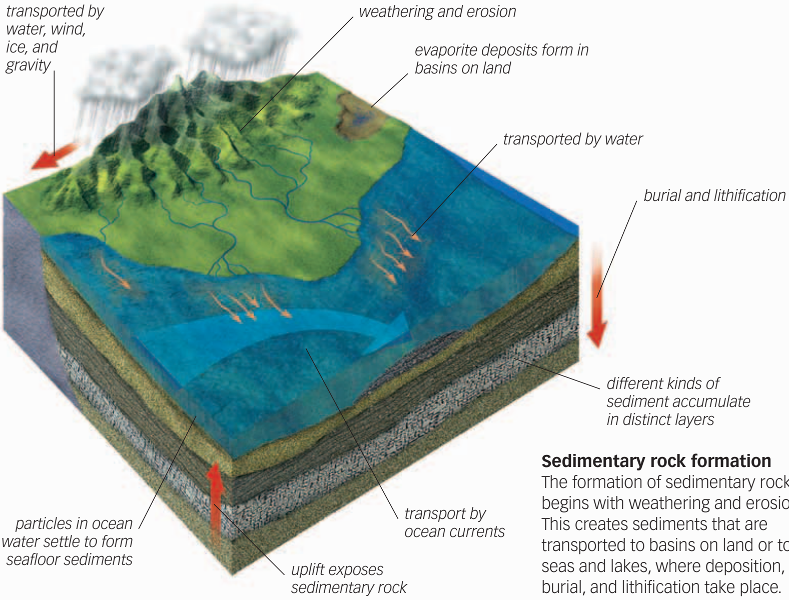
transported by water, wind, ice, and gravity
weathering and erosion
evaporite deposits form in basins on land
transported by water
burial and lithification
different kinds of sediment accumulate in distinct layers
particles in ocean water settle to form seafloor sediments
transport by ocean currents
uplift exposes sedimentary rock

Sedimentary rock formation
The formation of sedimentary rocks begins with weathering and erosion. This creates sediments that are transported to basins on land or to seas and lakes, where deposition, burial, and lithification take place.

CLASTIC ROCKS

Clasts are rock fragments ranging in size from boulders to microscopic particles. Clastic sedimentary rocks are grouped according to the size of the clasts from which they form. Larger clasts such as pebbles, cobbles, and boulder-sized gravels form conglomerate and breccia; sand becomes sandstone; and finer silt and clay particles form siltstone, mudstone, and shale. The mineral composition of clastic rocks can be subject to considerable change over time.

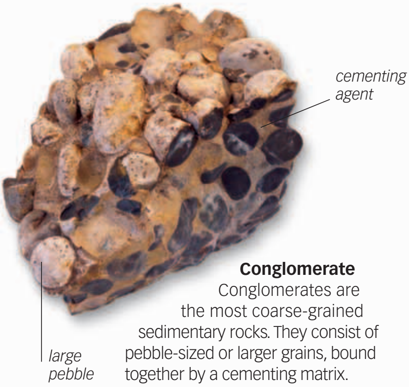

Conglomerate
Conglomerates are the most coarse-grained sedimentary rocks. They consist of pebble-sized or larger grains, bound together by a cementing matrix.

CHEMICAL ROCKS

Chemical sedimentary rocks are formed by the precipitation of the transported, dissolved products of chemical weathering. In some cases, the dissolved constituents are directly precipitated as solid rock. Examples include banded iron formations, some limestones, and bedded evaporite deposits—rocks and mineral deposits of soluble salts resulting from the evaporation of water. In other sedimentary rocks, such as limestone and chert, solid material first precipitates into particles or becomes the shells of organisms, which are then deposited and lithified.

White Cliffs of Dover
Located near the town of Dover in Kent, England, the celebrated White Cliffs of Dover are spectacular deposits of chalk many feet thick.

FOSSILS

A fossil is a remnant, impression, or trace of an organism that lived in a past geologic age. Fossils are preserved almost exclusively in sedimentary rocks. The most common fossils are of aquatic plants and animals. After an organism dies, the soft parts decompose, leaving behind only the hard parts—the shell, teeth, bones, or wood. Buried in layers of sediment, the hard parts gradually turn to stone. In a process known as permineralization, water seeps through the rock, depositing mineral salts in the pores of the shell or bone, thereby fossilizing the remains. In some cases, the organic matter is completely replaced by minerals as it decays. In other cases, circulating acid solutions dissolve the original shell or bone, leaving a cavity of identical shape, which is filled in as new material is deposited in the cavity, creating a cast.

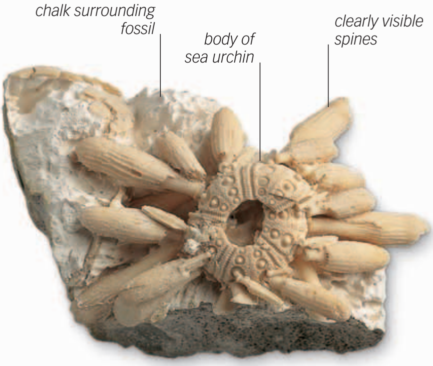

Fossil in chalk
Large sea urchin fossils, such as this well-preserved example, are usually found in chalk, which is itself made up of the fossils of tiny marine organisms.

PROFILE

- Detrital, from pebbles
- Marine, freshwater, glacial
- ¹⁄₁₆ in (2 mm) to several in (cm) in finer matrix
- Any hard mineral can be present
- Any mineral can be present
- Varies
- Very rare

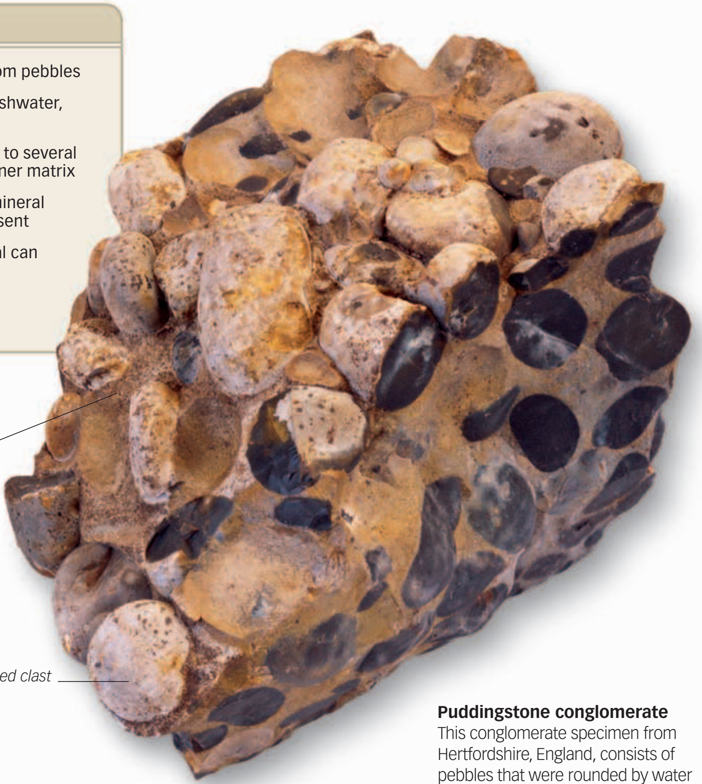

Puddingstone conglomerate This conglomerate specimen from Hertfordshire, England, consists of pebbles that were rounded by water and then cemented together.

VARIANTS

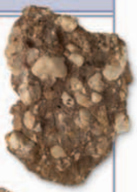

Quartz conglomerate A specimen with large clasts of quartz

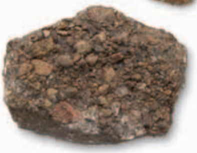

Polygenetic conglomerate A specimen with clasts of various rock types

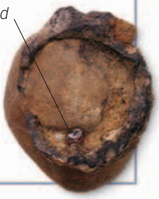

Alluvial diamond A diamond cemented into a conglomerate

CONGLOMERATE

Rocks formed by the lithification of rounded rock fragments that are over ¹⁄₁₆ in (2 mm) in diameter are known as conglomerates. They can be further classified by the average size of their constituent materials—pebble-conglomerate (fine), cobble-conglomerate (medium), and boulder-conglomerate (coarse). Conglomerate can also be known by the rock or mineral fragments in its composition; for example, a quartz pebble conglomerate.

Depending on the environment in which these fragments are deposited, these rocks may be of two types. Well-sorted conglomerates result from water flow over a long period. These have well-sorted pebbles (with a small size variation) generally of only one rock or mineral type, and a few small particles between the pebbles. Poorly sorted conglomerates form from rapid water flow and deposition. They have poorly sorted pebbles (of varying sizes) of mixed rock and mineral types with a number of small particles between the pebbles.

PROFILE

- Detrital, from coarse sediment
- Marine, freshwater, glacial
- 1⁄16 in (2 mm) to 1 in (several cm) in finer matrix
- Any hard mineral can be present
- Any mineral can be present
- Varies
- Very rare

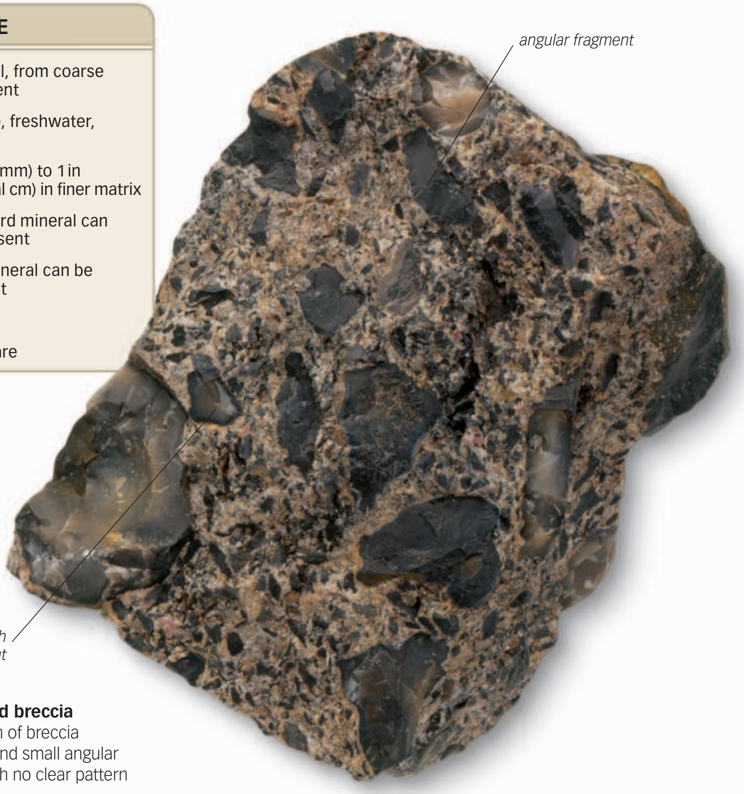

Poorly sorted breccia This specimen of breccia shows large and small angular fragments with no clear pattern of orientation.

VARIANTS

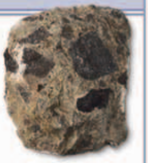

Limestone breccia Fragments of limestone in breccia

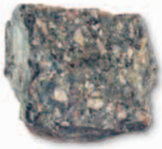

Polygenetic breccia A specimen with clasts of different types of rock

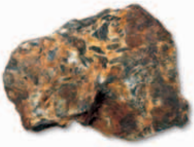

Fault breccia Cemented clasts shattered by faulting

BRECCIA

Lithified sediments with rock fragments that are more than 1⁄16 in (2 mm) in diameter but angular or only slightly rounded are called breccias. The lack of rounding indicates that little or no transportation took place before the fragments became incorporated in the rock.

Breccias can form in several ways. Rocks can shatter—for example, due to frost action or earth movement—and the fragments then become cemented in the new position. Shattered fragments may also move before being cemented—for example, they may accumulate at the base of a cliff or be carried by a flash flood. Breccias can also form in areas of active faulting. In areas where faulting occurs underwater, newly shattered material can also move in underwater landslides and become cemented to form breccia.

Breccia vase This ancient Egyptian carved vase made of breccia or mottle stone was used to store liquids.

PROFILE

Detrital, from sand
Terrestrial
$\frac{1}{256}$–$\frac{1}{16}$ in (0.1–2 mm)
Quartz, feldspar
Silica, calcium carbonate
Cream to red
Vertebrates, invertebrates, plants

Structure in sandstone This sandstone from Yorkshire, England, is derived from deposits of wind-blown sand. The bedding planes, which slumped after deposition, are well preserved.

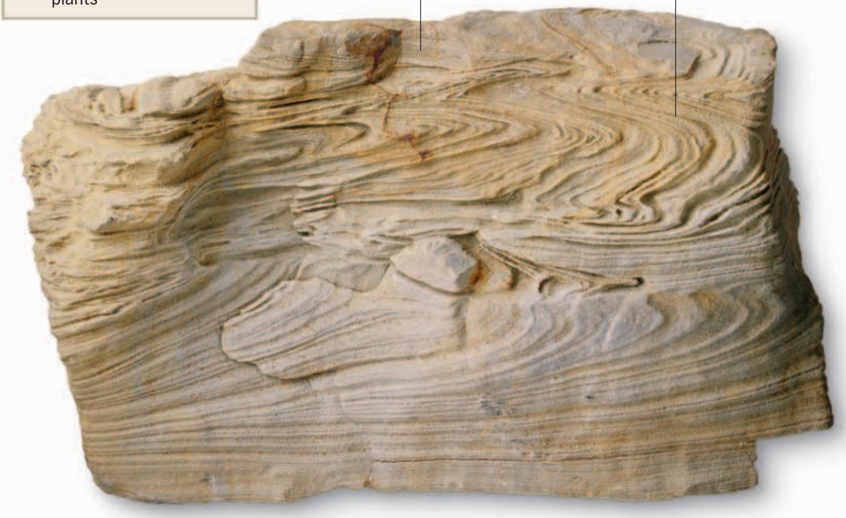

VARIANTS

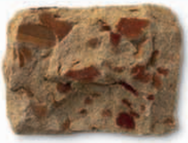

Micaceous sandstone A specimen with flakes of mica and patches of iron oxide

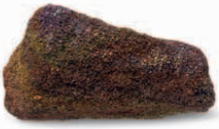

Goethite sandstone Sand grains with red goethite

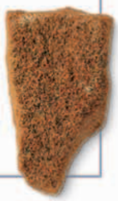

Red sandstone A specimen of quartz sandstone colored red due to the presence of iron oxides

SANDSTONE

The second most abundant sedimentary rock after shale (p.313), sandstone makes up about 10 to 20 percent of the sedimentary rocks in Earth's crust. Sandstones are classified according to texture and mineralogical properties into micaceous sandstone, orthoquartzite (p.310), and graywacke (p.317). They are usually dominated by quartz (p.168) and have visible sandy grains and other minerals present in varying amounts. Well-rounded grains are typical of desert sandstone, while river sands are usually angular, and beach sands somewhere in between.

Bedding is often visible in sandstones as a series of layers representing successive deposits of grains. Bedding surfaces may show either ripples or the cross bedding typical of dunes. Sandstone is an important indicator of deposition and erosion processes.

Sandstone platform This platform in the center of the courtyard pool at Fatehpur Sikri, India, is made from red sandstone.

Marine greensand
This specimen is from the Atlantic coastal plain of North America, where greensand is a common type of rock.

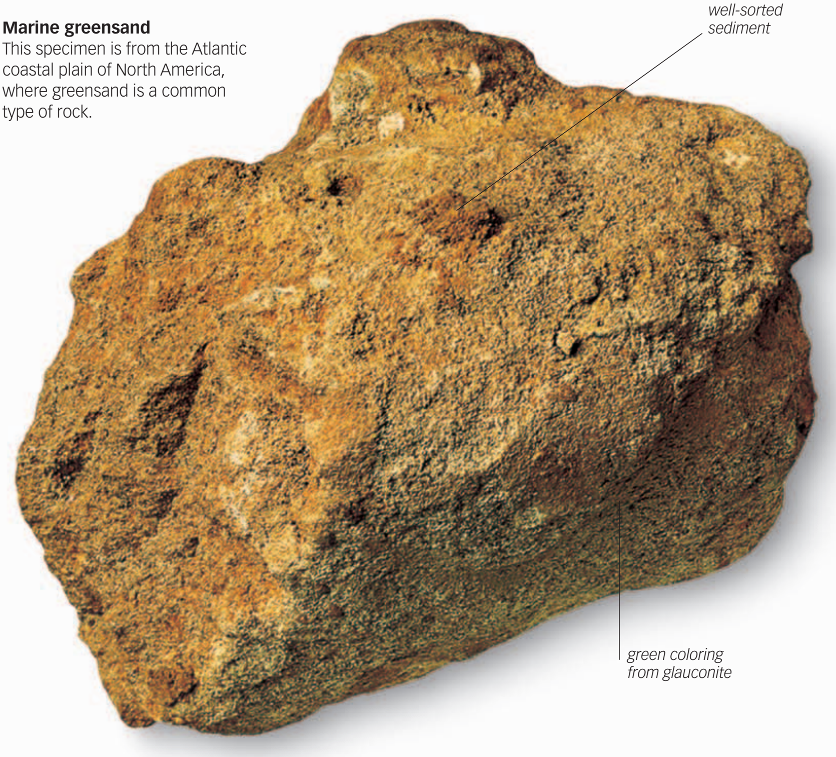

PROFILE

- ▲ Detrital, from glauconite-rich marine sand
- 🌐 Marine
- 🔍 1/256–1/16 in (0.1–2 mm)
- ◼ Quartz, glauconite
- ◻ Feldspar, mica
- ✱ Green
- ✲ Vertebrates, invertebrates, plants

VARIANT

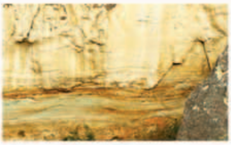

Greensand banding
A band of greensand within a rock outcrop

GREENSAND

A quartz sandstone with a high percentage of the green mica mineral glauconite (p.196) is known as greensand or glauconitic sandstone. The term glauconite is also loosely applied to any glauconitic sediment. Glauconite is believed to form in shallow, oxygen-poor marine environments that are rich in organic detritus. It forms as a result of a slow accumulation of sediments, by the replacement of calcite (p.114), or by primary deposition. It may contain shell fragments and larger fossils. Some glauconite pellets are biogenic, originating as fecal pellets.

Greensand tends to be weak and friable, although relatively hard greensand has been used as building stone. The soil derived from greensand varies, ranging from fertile to sterile. Glauconite is favored in organic cultivation as a natural source of potassium and phosphorus. Its potassium content is useful in radiometric age-dating. Due to its chemical-exchange properties, the glauconite in greensand is used as a water softener and in water-treatment systems.

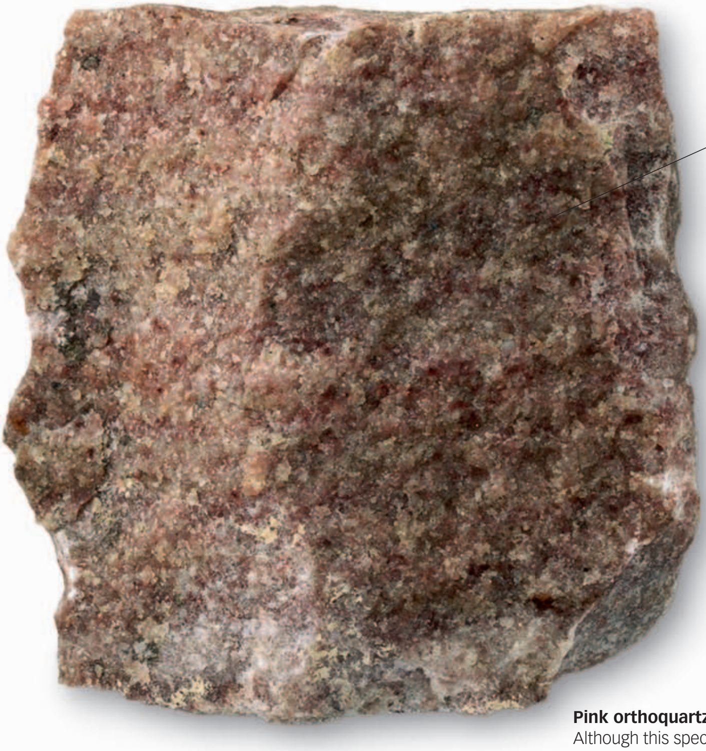

Pink orthoquartzite
Although this specimen of pink orthoquartzite is composed almost entirely of quartz, it is colored by small amounts of iron oxide.

PROFILE

- Detrital
- Terrestrial
- 1⁄256–1⁄16 in (0.1–2 mm)
- Quartz
- Heavy minerals, such as zircon and rutile
- Light to medium
- Rare, vertebrates

VARIANT

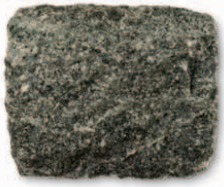

Gray orthoquartzite
A specimen colored gray by quartz grains

ORTHOQUARTZITE

A pure quartz sandstone, orthoquartzite is usually composed of well-rounded quartz (p.168) grains cemented by silica. The high degree of rounding of the grains indicates that they have traveled some distance, which also accounts for the high degree of sorting. Some orthoquartzites are up to 99 percent quartz, with only minor amounts of iron oxide and traces of other erosion-resistant minerals, such as rutile (p.78), magnetite (p.92), and zircon (p.233). Lesser amounts of other minerals can color the generally white or pinkish specimens either gray or red. Orthoquartzites can be differentiated in hand specimens from other sandstones by their lighter color and absence of other minerals.

Orthoquartzites rarely preserve fossils, although the sedimentary structures are usually preserved. The presence of silica cement makes orthoquartzites durable. They tend to resist weathering and form prominent outcrops. Orthoquartzite is distinct from metamorphic quartzite, also known as metaquartzite (p.295).

Gray arkose
Darker variants of arkose, such as this specimen, tend to be older than pink arkose.

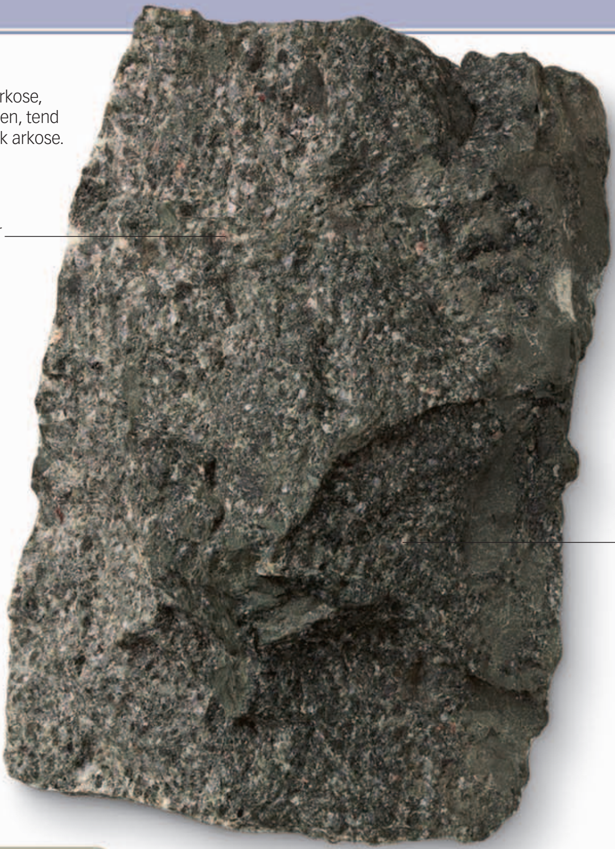

PROFILE

- ▲ Detrital, from feldspar-rich sand
- 🌐 Terrestrial, marine, or freshwater
- 🔍 $\frac{1}{256}$–$\frac{1}{16}$ in (0.1–2 mm)
- ◩ Quartz, feldspar
- ◳ Mica
- ✻ Pinkish, pale gray
- ❂ Rare

VARIANT

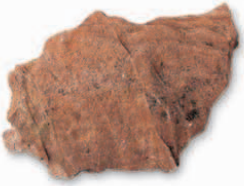

Pink arkose A variant of arkose formed more recently than gray arkose

ARKOSE

A pink sandstone, arkose is colored by an abundance of feldspars (pp.173–81), especially pink alkali feldspars. Its high feldspar content (more than 25 percent of the sand grains) sets it apart from other sandstones. Arkose specimens are relatively coarse and consist primarily of quartz (p.168) and feldspar grains, with small amounts of mica. The grains tend to be moderately well sorted and angular or slightly rounded. They are usually cemented with calcite (p.114) or sometimes with iron oxides or silica. The flat cleavage faces and angular grain shape of the feldspars reflect light under a hand lens. Sandstones with a feldspar content of 5–25 percent are called subarkoses.

Arkose forms from the quick deposition of sand weathered from granites (pp.258–59) and gneisses (p.288). The development of arkoses is thought to indicate either a climatic extreme or a rapid uplift and high relief of the source area. Arkoses are common along the front ranges of the Rocky Mountains.

PROFILE

- Detrital, from grit
- Freshwater
- 1/32–3/16 in (1–4 mm)
- Quartz, feldspar
- Garnet, rutile
- White, pink, gray
- Invertebrates, vertebrates, plants

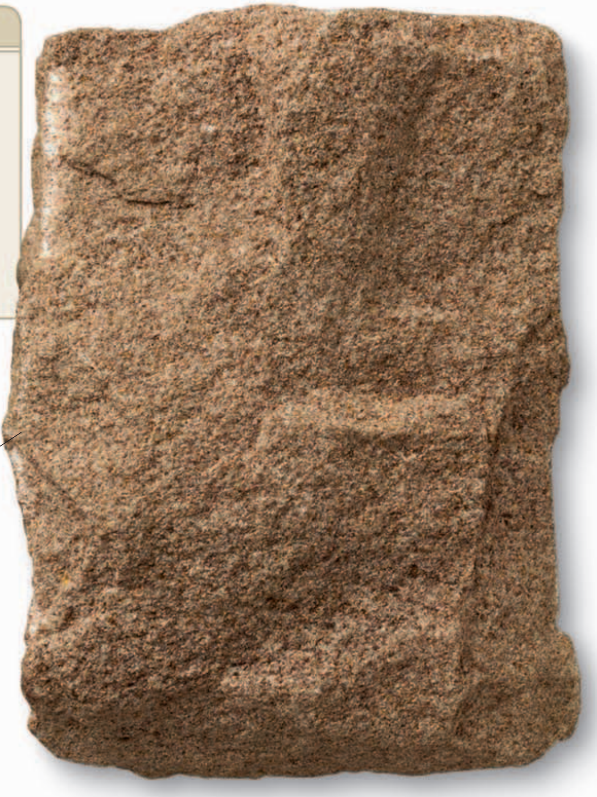
reddish color from iron oxides

Quartz gritstone
Specimens of quartz gritstone, such as this, are composed of more than 75 percent quartz.

VARIANT

Feldspathic gritstone
A variant containing up to 25 percent feldspar

GRITSTONE

Providing material for grindstones and millstones, gritstone has, in the past, been a commercially important sedimentary rock. It is a porous rock composed of cemented, coarse, often angular sand grains, with occasional small pebbles. Quartz (p.168) is always the greatest component, but specimens often contain iron oxides that give them a yellow, brown, or red color. In some, the grains can be easily rubbed out; in others, strong cement makes the rock suitable for use as grinding stones.

Gritstones originate in river deposits and frequently show signs of cross-bedding or current bedding. The Millstone Grit, a gritstone deposit in northern England, was mined for millstones used in flour mills and for grindstones, which were used to make paper pulp from wood and sharpen tools. Gritstone is still quarried for use as building material worldwide. Large exposures of gritstone are favored by rock climbers because the rough surface provides outstanding friction, enabling them to grip the smallest features in the rock.

Gray shale
This specimen of shale from Runswick Bay, Yorkshire, England contains fossils of bivalves and ammonites.

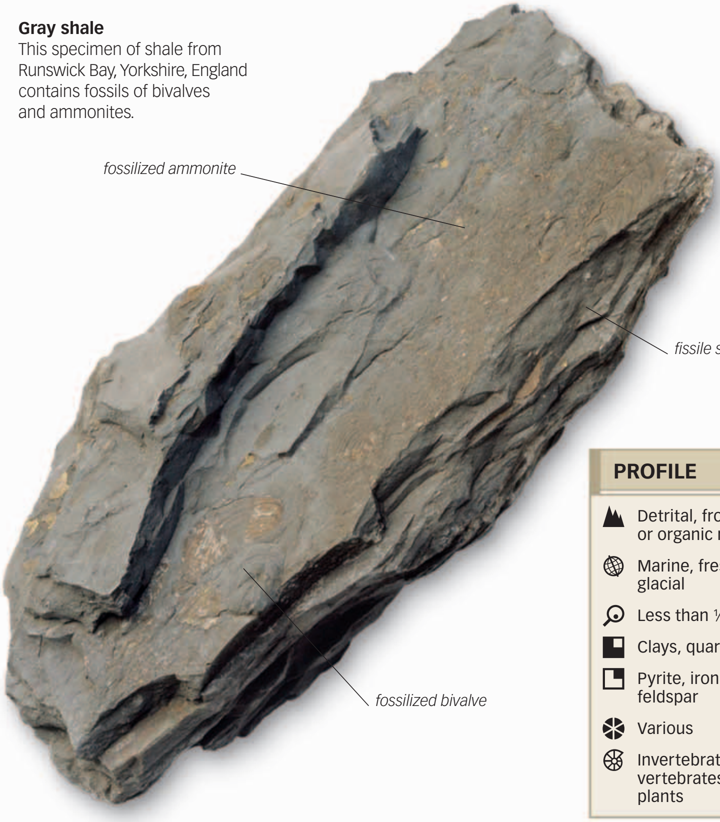

PROFILE

- Detrital, from mud, clay, or organic material
- Marine, freshwater, glacial
- Less than 1/256 in (0.1 mm)
- Clays, quartz, calcite
- Pyrite, iron oxides, feldspar
- Various
- Invertebrates, vertebrates, plants

SHALE

The most abundant sedimentary rock, shale makes up about 70 percent of all sedimentary rocks in Earth's crust. It consists of a high percentage of clay minerals, substantial amounts of quartz (p.168), and smaller amounts of carbonates (pp.114–25), feldspars (pp.173–81), iron oxides, fossils, and organic matter. Shales are colored reddish and purple by hematite (p.91) and goethite (p.102); blue, green, and black by ferrous iron; and gray or yellowish by calcite (p.114). They split easily into thin layers.

Shales consists of silt- and clay-sized particles deposited by gentle currents on deep ocean floors, shallow sea basins, and river floodplains. They occur thinly interbedded with layers of sandstone (p.308) or limestone (p.319) and in sheets up to several yards thick.

Fossil trilobite
Preserved in gray shale, this fossil trilobite is more than 400 million years old.

VARIANTS

Fossiliferous shale Shale with numerous fossils of brachiopods

alum-rich mineral

Light-colored shale
A specimen with alum-rich, light-colored areas

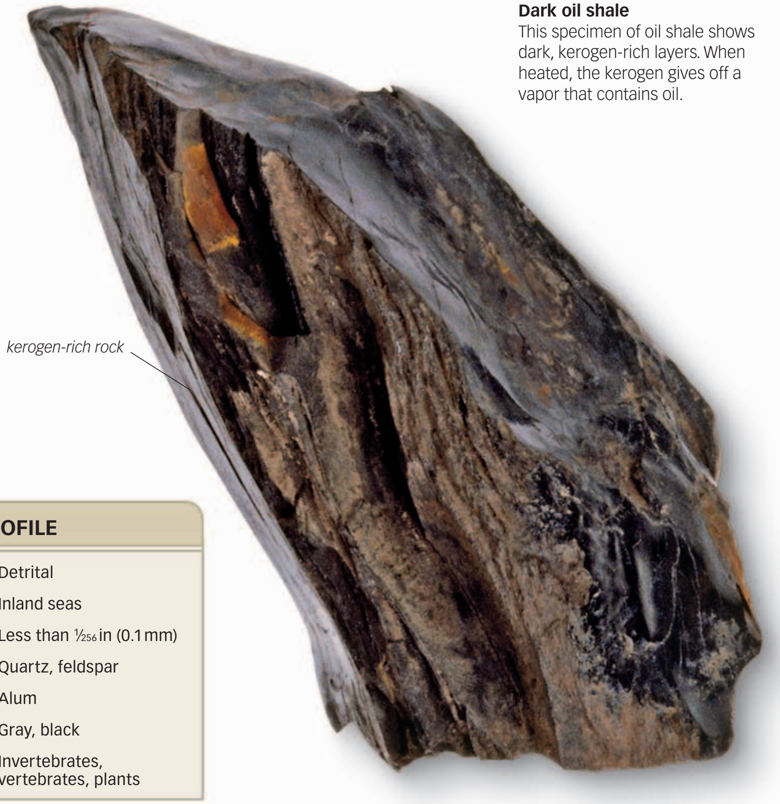

Dark oil shale
This specimen of oil shale shows dark, kerogen-rich layers. When heated, the kerogen gives off a vapor that contains oil.

PROFILE

- Detrital
- Inland seas
- Less than 1/256 in (0.1 mm)
- Quartz, feldspar
- Alum
- Gray, black
- Invertebrates, vertebrates, plants

VARIANT

Oily surface A specimen with condensed droplets of oil visible on the surface

OIL SHALE

The name oil shale is a general term for organic-rich sedimentary rocks that contain kerogen—a chemically complex mixture of solid hydrocarbons derived from plant and animal matter. When subjected to intense heat, these shales yield oil. Oil shales range from brown to black in color. They are flammable and burn with a sooty flame. Some oil shales are true shales in which clay minerals are predominant. Others are actually limestones (p.319) and dolomites (p.320). Much of the original organic material in oil shales is unrecognizable, but it is believed to be derived from plankton, algae, and microorganisms that live in fresh sediment.

In previous centuries, small amounts of oil have been successfully recovered from oil shales. During the past century, oil shales have been mined with rock types varying from shale (p.313) to marl (p.322) and other carbonate rocks. Various pilot plants have been built to extract oil from shales, but the commercial results have been modest so far.

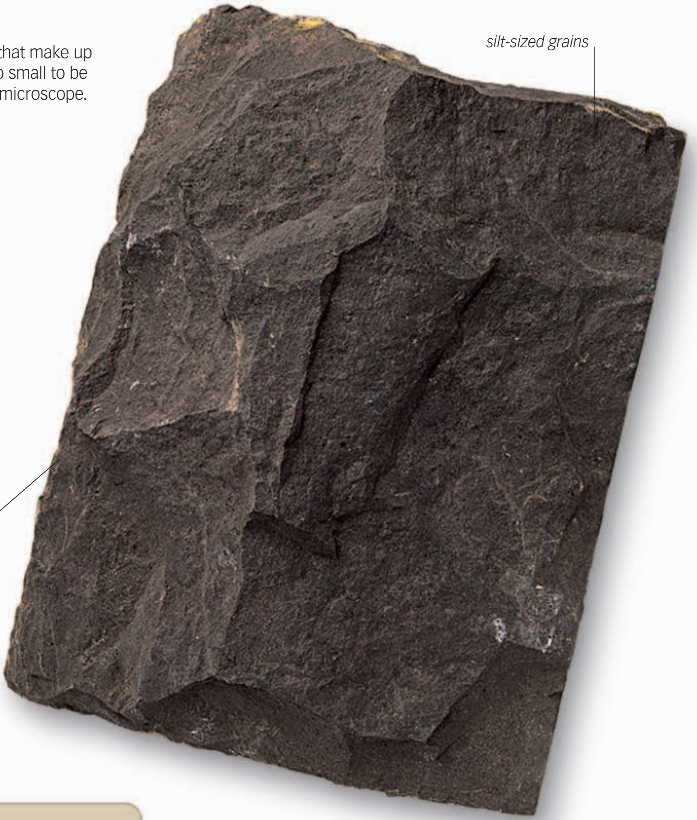

Gray siltstone
The tiny grains that make up siltstone are too small to be seen without a microscope.

PROFILE

- Detrital, from silt
- Marine, freshwater, glacial
- Up to 1⁄256 in (0.1 mm)
- Quartz, feldspar
- Mica, chlorite, mica-rich clay minerals
- Gray to beige
- Invertebrates, vertebrates, plants

VARIANT

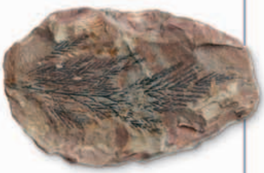

Fossiliferous siltstone
A fern fossil enclosed in siltstone

SILTSTONE

Formed from grains whose sizes vary between that of sandstone (p.308) and mudstone (p.316), siltstone is a sedimentary rock. Like sandstone, it can form in different environments and have different colors and textures. Siltstones are typically red and gray with flat bedding planes. Plant fossils and other carbon-rich matter are common in darker-colored siltstones. Examples tend to be hard and durable and do not easily split into thin layers. However, the presence of mica may produce a siltstone that splits into thicker, flagstonelike sheets. In addition to mica, siltstone may contain abundant chlorite and other mica-rich clay minerals.

Although many shales (p.313) contain more than 50 percent silt, siltstones are usually chemically cemented and show cross-bedding, ripple marks, and internal layering. This indistinct layering tends to weather at oblique angles unrelated to bedding. Siltstone is less common than shale or sandstone and rarely forms thick deposits.

PROFILE

- Detrital, from mud
- Marine, freshwater, glacial
- Less than 1⁄256 in (0.1 mm)
- Clays, quartz
- Calcite
- Gray, brown, black
- Invertebrates, vertebrates, plants

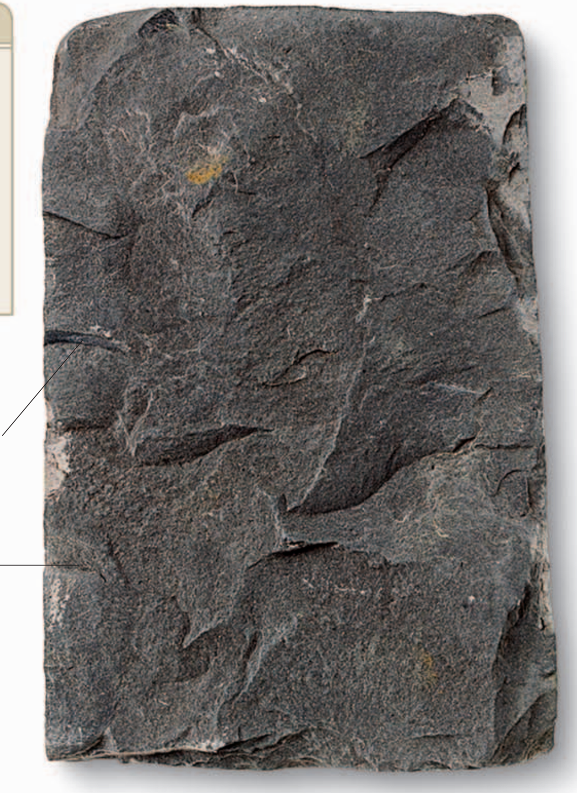

Mudstone specimen
Mudstone is somewhat similar in appearance to siltstone, but its grains are smaller and it has a broken surface with a much finer texture.

VARIANTS

Fossiliferous mudstone
A specimen of mudstone with numerous invertebrate fossils

Calcareous mudstone
A mudstone variant with a substantial amount of calcite

MUDSTONE

A gray or black rock formed from mud, mudstone contains carbon-rich matter, clay minerals, and detrital minerals such as quartz (p.168) and feldspar (pp.173–81). Mudstones look like hardened clay and can show the cracks seen in sun-baked clay deposits.

As in shale (p.313), mudstone's individual grains are clay- and silt-sized particles that can only be seen under a hand lens. Mudstone generally has the same color range as shale, with similar associations between color and content. Unlike shale, mudstone is not laminated during lithification and is not easily split into thin layers. The lack of layering is either due to the original texture or the disruption of layering. Mudstone deposits may be up to 10 ft (several meters) in thickness.

Nautiloid fossil
This fossil nautilus is preserved in mudstone. The iridescence of a part of its shell is still visible.

PROFILE

- Detrital, from muddy sand
- Marine
- 1/256–1/16 in (0.1–2 mm)
- Quartz, feldspar, mafic minerals
- Chlorite, biotite, clay, calcite
- Mostly gray or greenish gray; also brown, yellow, or black
- Rare

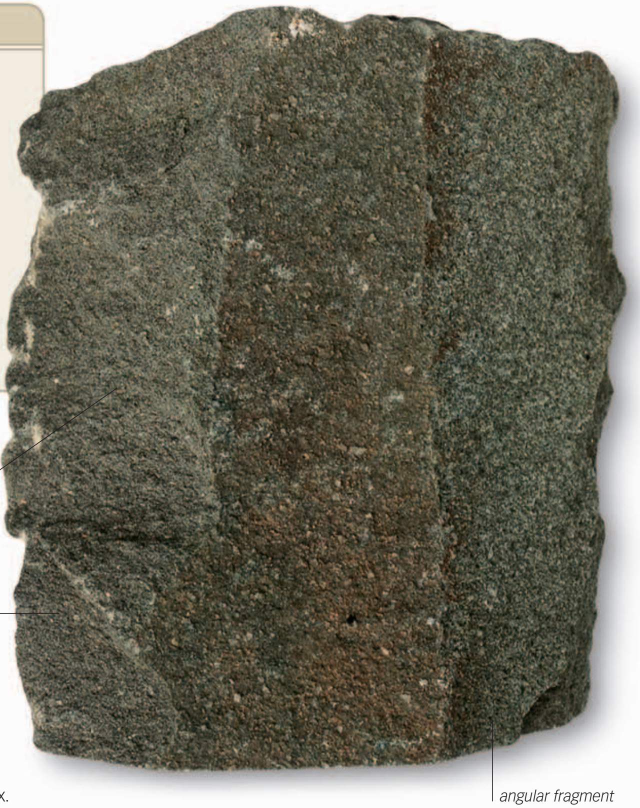

Fine-grained graywacke This specimen shows the poor sorting characteristic of graywacke. The angular clasts are set in a fine-grained matrix.

VARIANTS

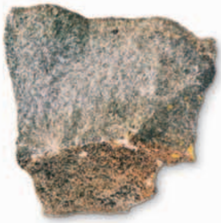

Dirty sandstone A specimen from Canada with typical dirty appearance and poor sorting

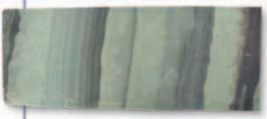

Turbidite A graywacke formed by undersea avalanches

GRAYWACKE

Easily confused with igneous basalt (p.273), graywacke is a turbidite—a rock resulting from rapid deposition in a turbulent marine environment. It gets its name from the German word *grauwacke*, which signifies a gray, earthy rock. Also called dirty sandstone, graywacke is hard, and mostly gray, brown, yellow, or black.

Graywacke is composed of poorly sorted, coarse- to fine-grained quartz (p.168), feldspar (pp.173–81), and dark-colored minerals such as amphibole and pyroxene, set in a fine-grained matrix of clay, calcite (p.114), or quartz. Graywackes may occur in thick or thin beds along with slates (p.293) and limestones (p.319). In ancient Egypt, this stone was used to make sarcophagi, vessels, and statues.

Graywacke carving This ancient Egyptian graywacke carving dates from 1370 BCE. It is an offering to the god of writing.

PROFILE

- Detrital
- Glacial
- From less than 1/256 in (0.1 mm) to many ft (m)
- Rock fragments
- Rock fragments
- Various
- None

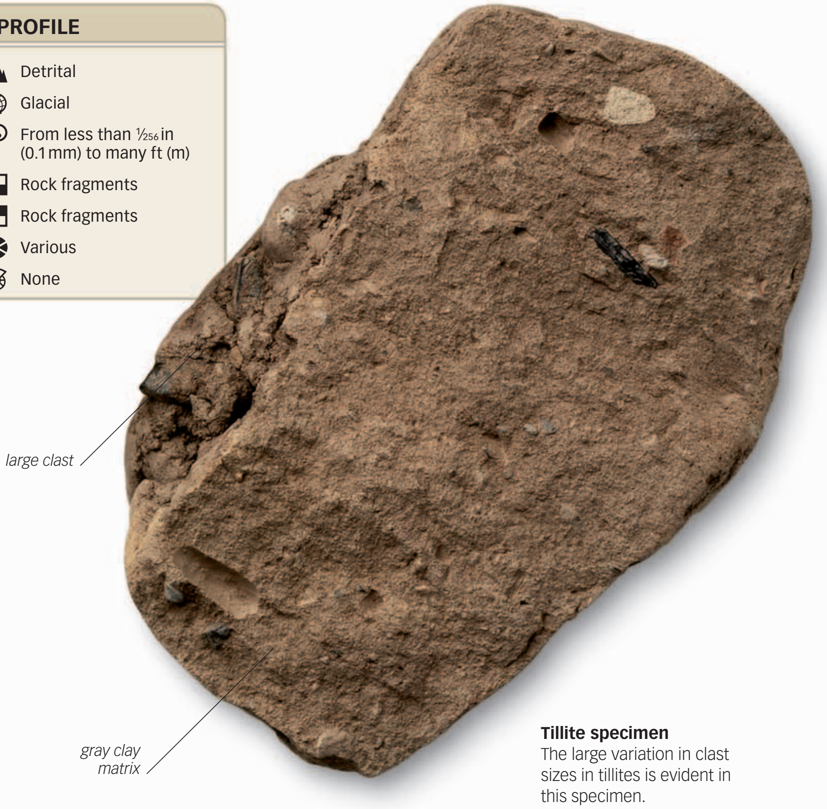

Tillite specimen The large variation in clast sizes in tillites is evident in this specimen.

VARIANTS

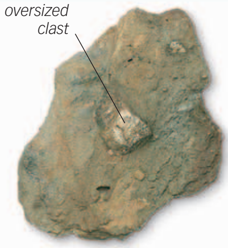

Large clast Tillite with a large clast in a finer matrix

Differing clast sizes A specimen of tillite with rounded clasts of differing sizes

TILLITE

Unsorted and unstratified rock material deposited by glacial ice that has later been lithified is called tillite. As glaciers move down valleys, they erode and transport rocks. On melting, they deposit this material, which is referred to as till. Glacial tills consist of preexisting rock fragments pushed forward or sideways by the glacier, newly eroded material ground or broken up by the glacier, or mixtures of the two.

Tillites are typically made of grains with a wide range of sizes. They include fragments ranging from clay-sized particles to large blocks. The largest pieces give rise to the rock's alternative name, "boulder clay." The matrix, which frequently comprises a large percentage of the rock, is often made from rock flour, an unweathered and finely ground rock powder. Matching beds of ancient tillites on opposite sides of the South Atlantic Ocean provided early evidence for continental drift—the idea that continents move relative to one another—which was an important precursor to the theory of plate tectonics.

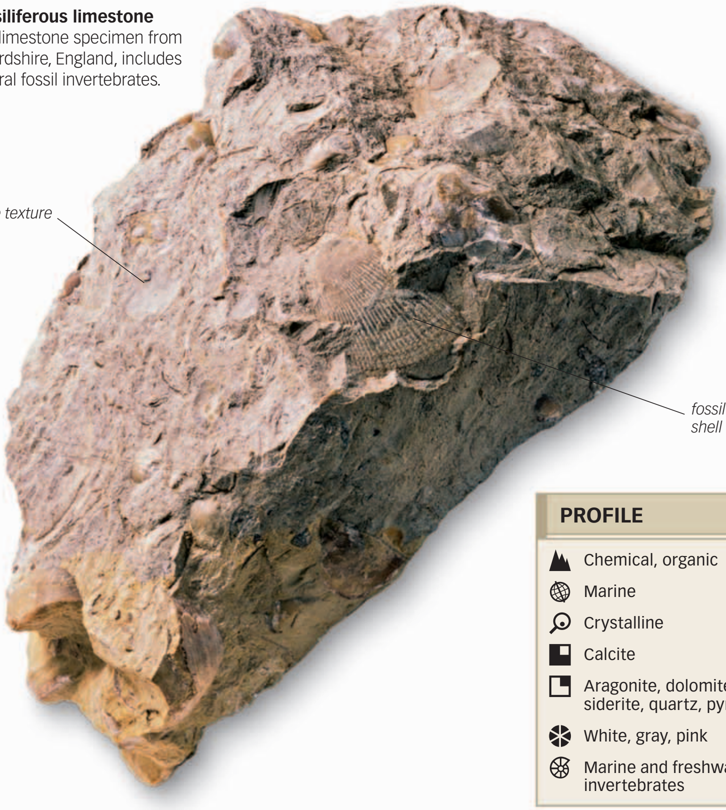

Fossiliferous limestone This limestone specimen from Oxfordshire, England, includes several fossil invertebrates.

PROFILE

- Chemical, organic
- Marine
- Crystalline
- Calcite
- Aragonite, dolomite, siderite, quartz, pyrite
- White, gray, pink
- Marine and freshwater invertebrates

LIMESTONE

Composed mainly of calcite (p.114), this abundant rock forms multiple layers that are thick and extensive. It can be yellow, white, or gray. Specimens can be identified by the rapid release of carbon dioxide and a fizzing sound when they react with dilute hydrochloric acid. Limestones can be compact, grainy, or friable. Many have cross-bedding or ripple marks. The texture of limestone ranges from coarse and fossil-rich to fine and microcrystalline.

Limestone generally forms in warm, shallow seas either from calcium carbonate precipitated from seawater or from the shells and skeletons of calcareous marine organisms. It is used in construction, as a raw material in the manufacture of glass, as a flux in metallurgical processes, and in agriculture.

Limestone mask This face mask dating from the Neolithic Period has been carved out of mottled limestone.

VARIANTS

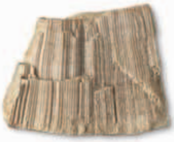

Coral limestone Limestone formed from fossil coral

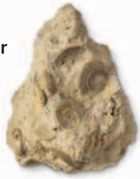

Freshwater limestone Uncommon variant of limestone

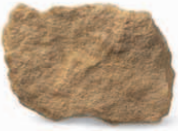

Oolitic limestone Round ooliths set in calcite cement

PROFILE

- Chemical
- Marine
- Crystalline
- Dolomite
- Calcite
- Gray to yellowish gray
- Invertebrates

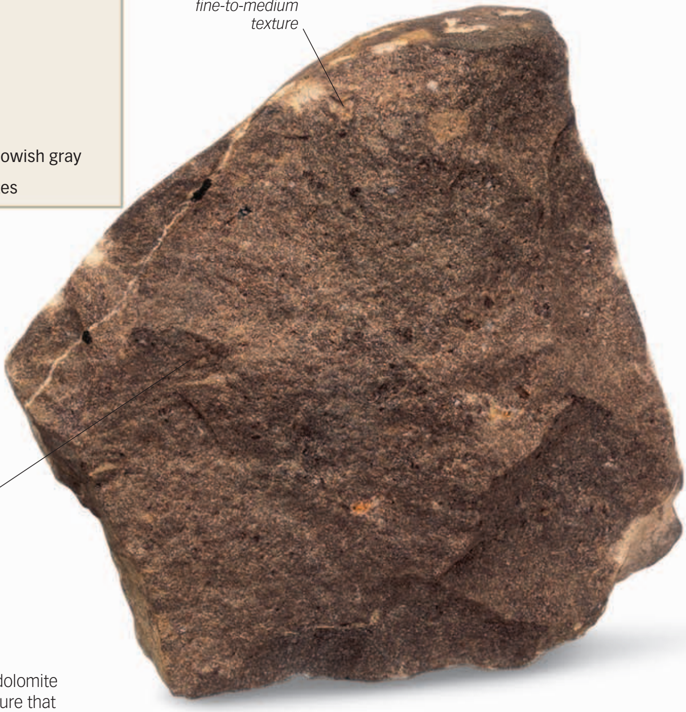

Dolomite rock This specimen of dolomite has a uniform texture that is typical of the rock.

VARIANT

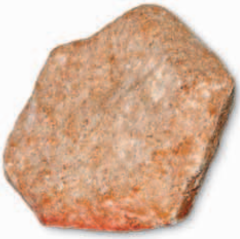

Red dolomite A specimen of dolomite colored with iron oxides and hydroxides

DOLOMITE

This rock, formed exclusively from the mineral dolomite (p.117), is also called dolostone. Most dolomite rocks are believed to be limestones (p.319) in which dolomite has replaced the calcite (p.114) in contact with magnesium-bearing solutions. This process is called dolomitization. Fresh dolomite looks similar to limestone, while weathered dolomite is yellowish gray. Dolomites have fewer fossils than limestones because fossils and other features are destroyed by the dolomitization process. Dolomite fizzes less violently than limestone when in contact with hydrochloric acid.

Dolomite typically occurs as massive layers. It is also present as thin layers or pods within limestone. Dolomite is used as a flux to make steel. It can also be used as a lightweight building aggregate, in breeze blocks, and in poured concrete.

Environment-friendly flux In contrast to other steel-making fluxes, dolomite produces slag that can be reused.

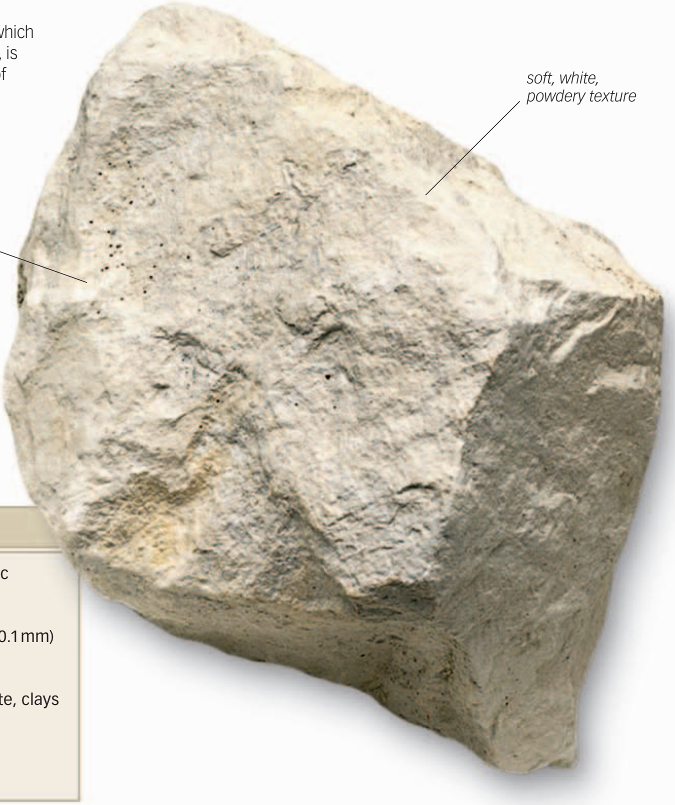

Microscopic fossils This specimen of chalk, which is almost entirely calcite, is made up of tiny fossils of marine organisms.

PROFILE

- ▲ Chemical, organic
- 🌐 Marine
- 🔍 Less than $1/256$ in (0.1 mm)
- ◩ Calcite
- ◳ Quartz, glauconite, clays
- ✱ White, gray, buff
- ⊛ Invertebrates, vertebrates

VARIANTS

Red chalk A specimen of chalk that takes its color from incorporated hematite

Marine chalk A specimen of chalk pitted by the boring of marine animals

CHALK

A soft, fine-grained, easily pulverized, white to grayish variety of limestone is known as chalk. It is composed of calcite shells of minute marine organisms. Small amounts of other minerals, such as apatite (p.148), glauconite (p.196), and clay minerals, are usually present. Silica from sponge spines, diatom and radiolarian skeletons, and nodules of chert (p.332) and flint can also be present.

Extensive chalk deposits were formed during the Cretaceous Period (142 to 65 million years ago), the name being derived from the Latin word *creta*, which means "chalk." Chalk is used to make lime and cement and as a fertilizer. It is also used as a filler, extender, or pigment in a wide variety of materials, including ceramics and cosmetics.

Blackboard chalk One of chalk's longstanding uses has been compression into sticks for writing on blackboards.

Green marl
Marls differ in color depending on their mineral content. Green marls, such as this one, are colored by either glauconite or chlorite.

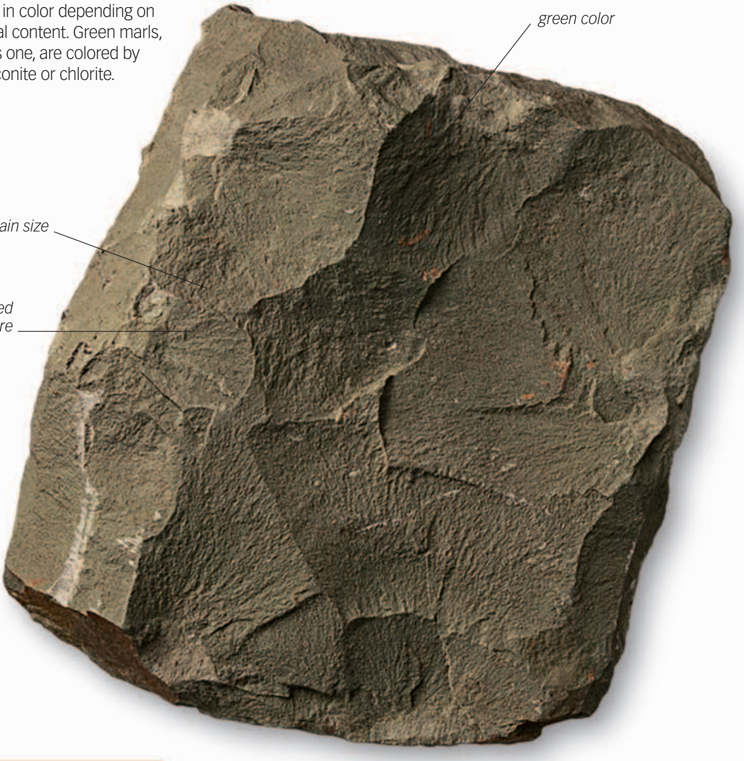

PROFILE

- Detrital, from lime mud
- Marine, freshwater
- Less than 1⁄256 in (0.1 mm)
- Clays, calcite
- Glauconite, hematite
- Various
- Vertebrates, invertebrates, plants

VARIANT

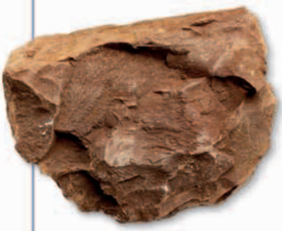

Red marl A specimen colored red by iron oxides

MARL

Calcareous mudstone or marl is a term applied to a variety of rocks that have a range of compositions but are all earthy mixtures of fine-grained minerals. In some countries, marl is referred to by the German terms *mergel* and *seekreide*, both of which mean "lake chalk." Marls usually consist of clay minerals and calcium carbonate. They form in shallow freshwater or seawater. The calcium carbonate content is frequently made up of shell fragments of marine or freshwater organisms or calcium carbonate precipitated by algae. The high carbonate content of marls makes them react with dilute acid.

Marls are whitish gray or brownish in color but can also be gray, green, red, or variegated. Greensand marls contain the green mineral glauconite (p.196), and red marls, iron oxides. Marl is much less easily split than shale (p.313) and tends to break in blocks. Specimens are often nodular, and the nodules are usually better cemented than the surrounding rock.

PROFILE

- Evaporite
- Marine
- Crystalline
- Gypsum
- Anhydrite
- White, pinkish, yellowish, gray
- None

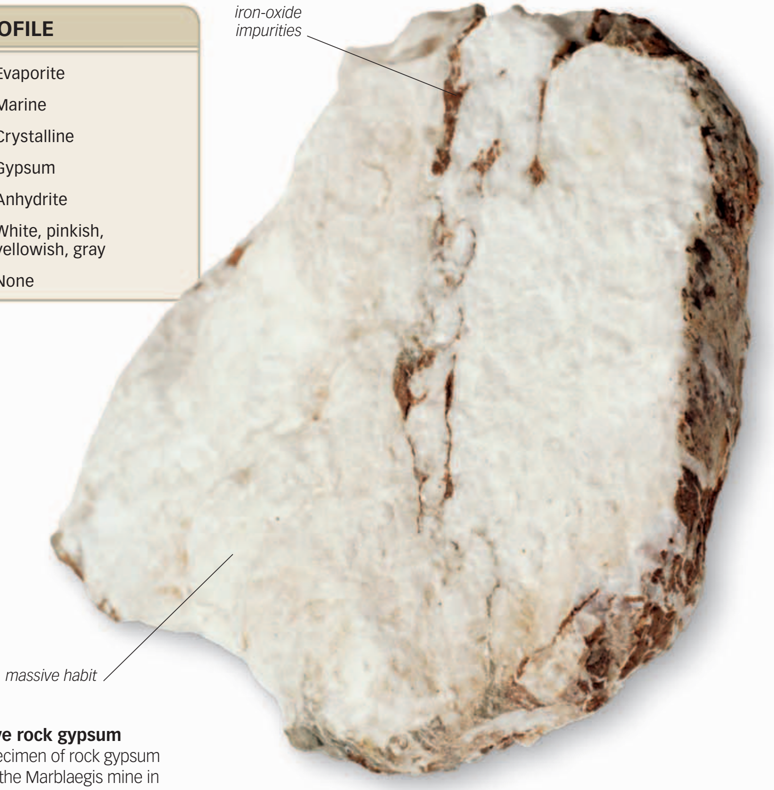

Massive rock gypsum This specimen of rock gypsum is from the Marblaegis mine in Nottinghamshire, England.

VARIANT

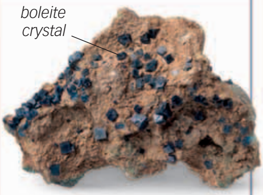

Gypsum with boleite Blue crystals of boleite seen in a matrix of gypsum

ROCK GYPSUM

Also called gyprock, rock gypsum is the sedimentary rock formed mainly from the mineral gypsum (p.136). Although it is commonly granular, it can also occur as fibrous bands. Rock gypsum occurs in extensive beds formed by the evaporation of ocean water, in saline lakes, and in salt pans. It also occurs in some shales (p.313), limestones (p.319), and dolomitic limestones. Rock gypsum is commonly interlayered with other evaporites, such as rock anhydrite (p.133) and salt (p.324).

Most of the gypsum that is extracted—about three-quarters of the total production—is calcined (heated to drive off some of its water) for use as plaster of Paris. Unaltered rock gypsum is used as a fluxing agent, fertilizer, filler in paper and textiles, and retardant in Portland cement.

Mesopotamian seal This ancient cylinder seal, with engravings of stags and scorpions, is made of alabaster gypsum.

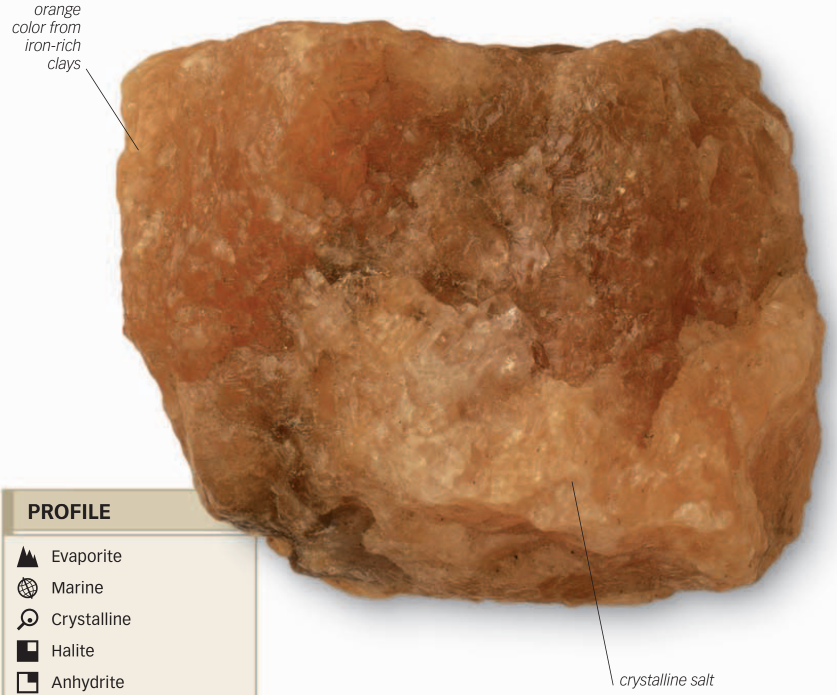

Crystalline rock salt This specimen of crystalline rock salt is colored orange by iron-rich clays incorporated during deposition.

PROFILE

- ▲ Evaporite
- 🌐 Marine
- 🔍 Crystalline
- ■ Halite
- □ Anhydrite
- ✱ White, pinkish, gray, yellowish
- ⊛ None

VARIANTS

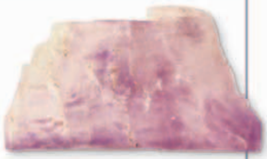

Pink rock salt A specimen colored pink by traces of iron oxides

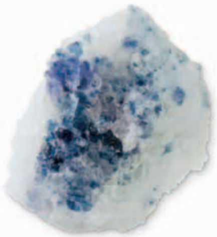

Massive rock salt Rock salt with coarse, white and blue crystals

ROCK SALT

Familiar as common table salt, rock salt is the massive rock form of the mineral halite (p.110). It occurs in beds that range in thickness from 3 ft (1 m) or so to more than 990 ft (300 m). Rock salt forms as a result of the evaporation of saline water in partially enclosed basins. It is commonly interlayered with beds of shale (p.313), limestone (p.319), dolomite (p.320), and other evaporites, such as anhydrite (p.133) and gypsum (p.136).

Rock salt often occurs in salt domes, which consist of a core of salt surrounded by strata of other rock. The domes can be 1 mile (2 km) or more thick and up to 6 miles (10 km) in diameter. They can occur with petroleum deposits: oil from oil-rich shales migrates up and gets caught on the underside of a salt dome.

Natural salt Many people prefer natural salt for their cooking. Much of it comes from evaporating brines.

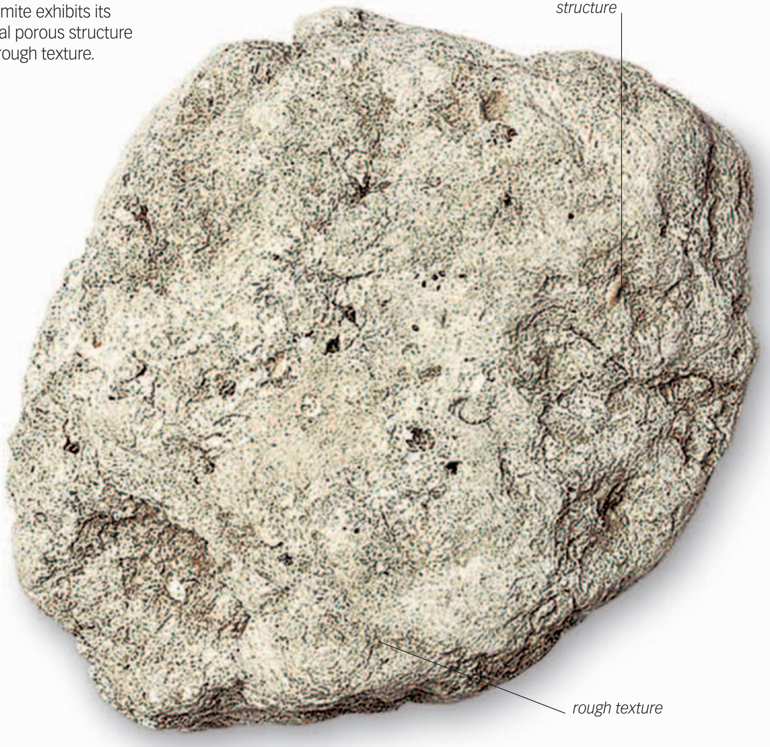

Porous diatomite
This specimen of diatomite exhibits its typical porous structure and rough texture.

PROFILE

- Detrital
- Accumulation of silica-rich organisms
- Less than 1⁄256 in (0.1 mm)
- Opal
- None
- White
- Diatoms

DIATOMITE

Also called diatomaceous earth, diatomite consists of about 90 percent silica; the remainder is made up of compounds such as aluminum oxides and iron oxides. Diatomite is easily crumbled into a fine, white to off-white powder. It is made of the fossilized remains of organisms called diatoms. A form of algae, diatoms have hard shells made of amorphous silica (opal) and containing many fine pores. Most float on the surface of the sea. Occasionally, large quantities are deposited in ocean sediments, eventually forming diatomite. A few significant deposits of freshwater diatomite are also known. Diatoms are so small that diatomite specimens may contain millions of diatom shells per cubic inch.

An important industrial rock, diatomite is easily mined. It is used for the filtration of beverages, liquid chemicals, industrial oils, cooking oils, fuels, and drinking water. Its low-abrasive qualities find use in toothpastes, polishes, and nonabrasive cleaners. It is also used as a filler and extender in paint and paper.

PROFILE

- Chemical
- Postdepositional
- Not granular
- Hematite, goethite, calcite, chalcedony
- Sand, clay
- Red-brown or yellow
- None

Red-brown laterite
This specimen shows the intermixture of iron and aluminum oxides with sand or other rock detritus characteristic of laterites.

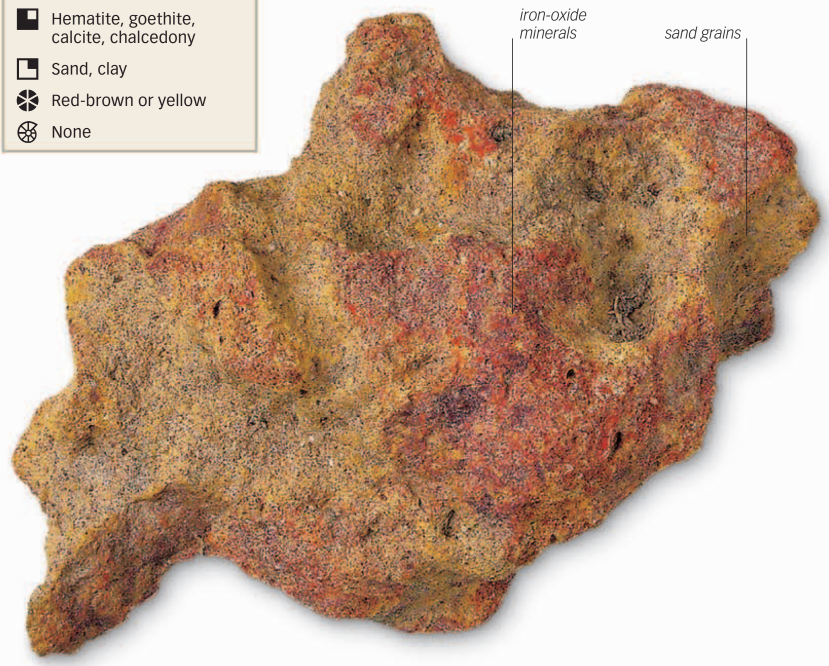

VARIANT

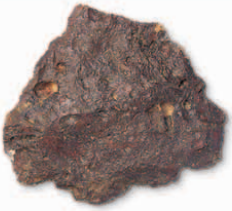

Desert-varnished laterite
A variant of laterite with a polished surface

LATERITE

Laterite is nodular soil rich in hardened iron and aluminum oxides and resembles bauxite (p.101) in composition. Nodules of laterite are red-brown or yellow and contain grains of sand or hardened clay. The rock forms in hot, wet tropical climates, where it develops by intensive and prolonged chemical weathering of the underlying rock. Evaporation and leaching of minerals from rock, loose sediment, and soil leaves behind insoluble salts. This results in a variety of laterites, which differ in their thickness, grade, chemistry, and ore mineralogy.

Laterite is a source of bauxite, which is an ore of aluminum and exists largely in clay minerals and various hydroxides. Laterite ores have also been a source of iron and nickel. Many laterites are solid enough to be used as building blocks. An example is the laterite seen at the famous temple in Angkor Wat, Cambodia. Crushed laterite has been widely used to make roads. Laterites are also being increasingly used in water treatment.

PROFILE

- Chemical
- Swamps, bogs
- Less than 1⁄256 in (0.1 mm)
- Hematite, goethite, calcite, chalcedony
- Sand, clay
- Light to dark
- Plants

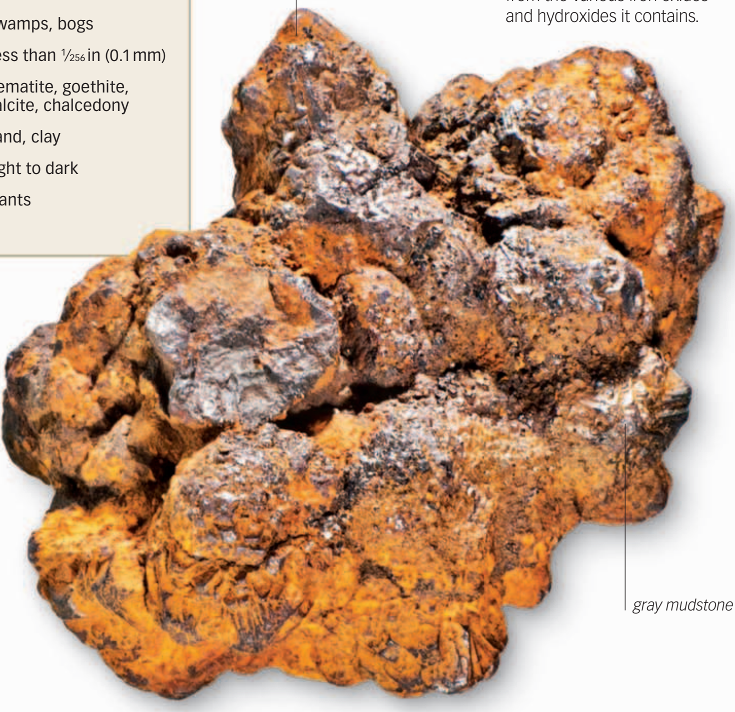

Bog iron nodule This nodule of bog iron shows varied coloration, which results from the various iron oxides and hydroxides it contains.

VARIANT

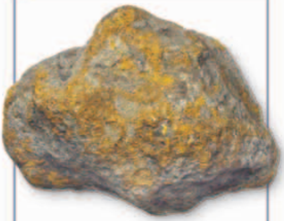

Bog iron ore Gray mudstone with oxides and hydroxides of iron

BOG IRON

Impure iron deposits that develop in bogs or swamps are known as bog iron. Bog iron is typically a brown-yellow mudstone with yellow, red, brown, or black concretions of iron oxides and hydroxides. In general, bog ores consist of iron hydroxides, primarily goethite (p.102). Bog iron can contain up to 70 percent iron oxide. It often contains carbon-rich plant material, which is sometimes preserved by iron minerals. Bog iron typically forms in areas where iron-bearing groundwater emerges as springs. As the iron encounters the oxygen-rich surface water, it oxidizes, and the iron oxides precipitate out.

Bog iron was formerly used as an ore and was widely sought in the preindustrial age. The Romans and the Vikings made extensive use of bog iron as a source of iron. Bog iron was also the principal source of iron in colonial USA. Although it is still forming today, the iron requirements of modern industry demand much larger sources of ore.

PROFILE

- Chemical
- Marine, terrestrial
- 1⁄256–1⁄16 in (0.1–2 mm)
- Hematite, goethite, chamosite, magnetite, siderite, limonite, jasper
- Pyrite, pyrrhotite
- Red, black, brown, yellow
- Invertebrates or plants

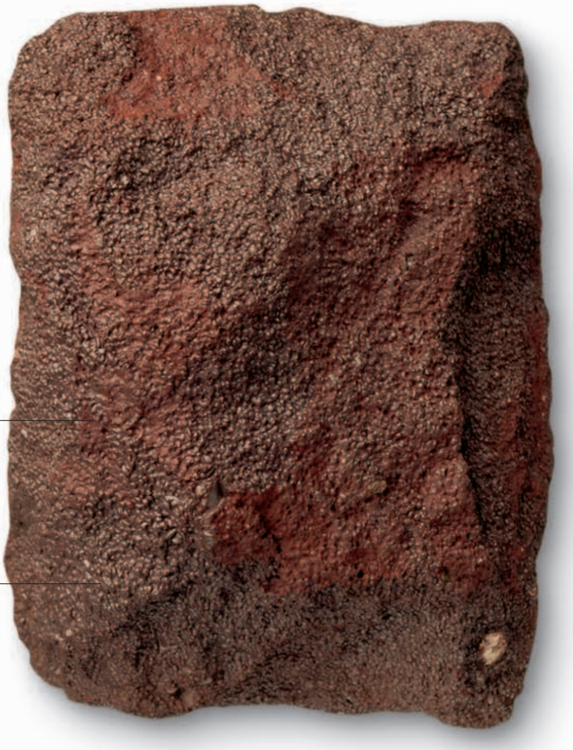

Oolitic ironstone
This specimen of ironstone is made up of small, rounded oolites, which are cemented by hematite.

VARIANT

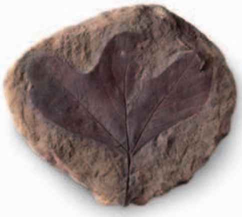

Sandy ironstone A specimen of a leaf fossil beautifully preserved in ironstone

IRONSTONE

The term ironstone is applied to sandstones and limestones that contain more than 15 percent iron. Ironstones are rich in iron-bearing minerals such as hematite (p.91), goethite (p.102), siderite (p.123), and chamosite. These minerals give ironstones a dark red, brown, or yellow color.

Ironstone no longer appears to be forming; partly because of this, the process by which it forms is something of a mystery. Some ironstones seem to have formed early in Earth's history when oxygen was not as abundant in the atmosphere as it is now. Precambrian ironstones, which are more than 500 million years old, as well as slightly later ones, formed prior to 240 million years ago, are common. Many of the later ironstones consist of oolites (small spheres) of hematite and contain fossils. Although historically used as an iron ore, ironstone is too limited in quantity to be an economic modern source of iron.

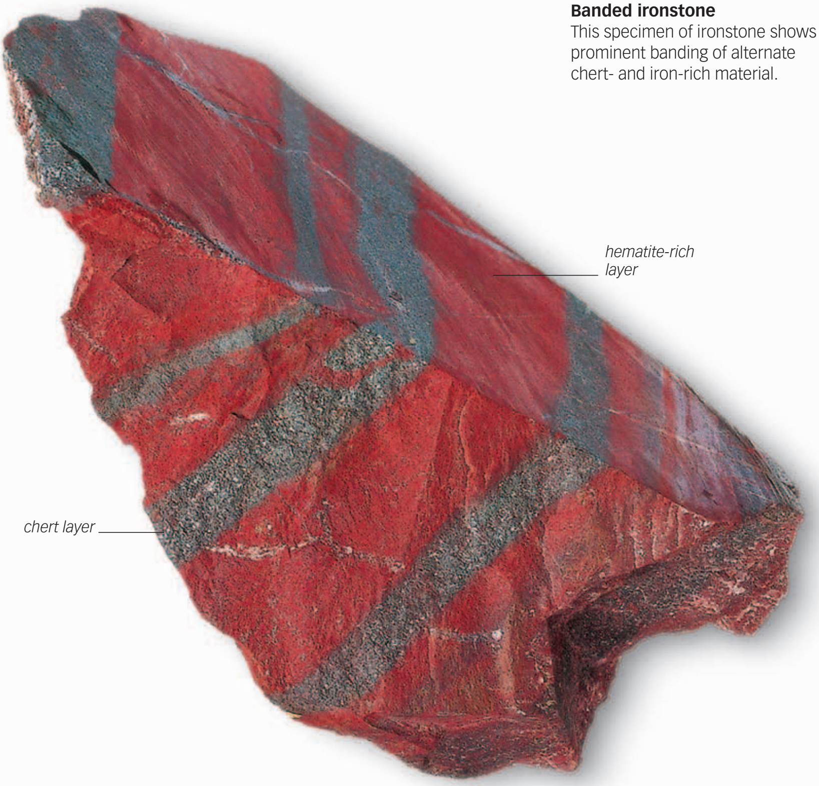

Banded ironstone
This specimen of ironstone shows prominent banding of alternate chert- and iron-rich material.

PROFILE

- Chemical
- Marine
- Neither granular nor very fine
- Siderite, hematite, chert
- Magnetite, pyrite
- Red, black, gray, striped
- None

BANDED IRONSTONE

Also known as banded iron formations, banded ironstone is made up of thin layers of alternating red, brown, or black iron oxides, which may be hematite (p.91) or magnetite (p.92), and gray or off-white shale (p.313) or chert (p.332). It is a very fine-grained rock that breaks into smooth, splintery pieces. Particularly abundant in Precambrian rocks, which are more than 500 million years old, banded ironstone can form very thick sequences, such as in the Hammersley Range of Australia, where it is an economically important iron ore.

Banded iron layers are found in some of the oldest known rock formations, dating from more than 3,700 million years ago. These layers are believed to have formed in the seas as a result of dissolved iron combining with oxygen released by the green algae that flourished in the oceans at that time. This was then precipitated into oxygen-poor seafloor sediments that were forming shale and chert. The banding of ironstone is assumed to result from cyclic variations in the oxygen available.

PROFILE

Chemical
Terrestrial
Less than $\frac{1}{32}$ in (1 mm)
Calcite or silica
Aragonite
White
Rare

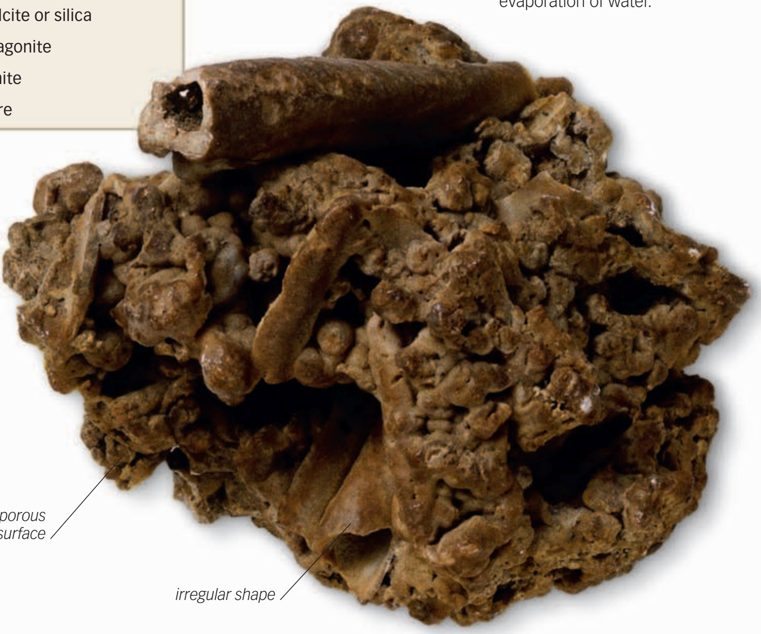

Calcareous tufa
This specimen of calcareous tufa is full of holes and irregular shapes. It was formed by the evaporation of water.

VARIANTS

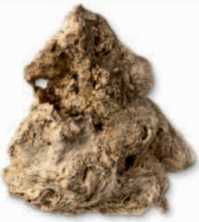

Yellow tufa Irregularly shaped tufa that lacks bedding

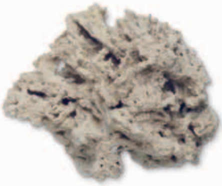

Fossilized tufa Tufa formed around plant remains

TUFA

Two different sedimentary rocks that precipitate from water are known as tufa. Calcareous tufa, or calc-tufa, is a soft, porous form of limestone deposit composed principally of calcium carbonate (calcite) that precipitates from hot springs, lake water, and groundwater. Calc-tufa is often stained red by the presence of iron oxides.

Siliceous tufa, which is also called siliceous sinter, is a deposit of opaline or amorphous silica that forms through the rapid precipitation of fine-grained silica as an encrustation around hot springs and geysers. It is believed to have been partly formed by the action of algae in the heated water. The term “sinter” means that it has several hollow tubes and cavities in its structure, which are often a result of organic matter that has decomposed later.

Stone Wedding
This pink tufa formation is in the Rhodope Mountains, Bulgaria. It is locally called the Stone Wedding.

PROFILE

- Chemical
- Post-depositional
- Fine
- Pyrite, marcasite
- Silica
- Various
- Occasional plants and animals

Flint nodule This nodule shows a variation in color due to differences in the extraneous material enclosed while the silica making the rock solidified.

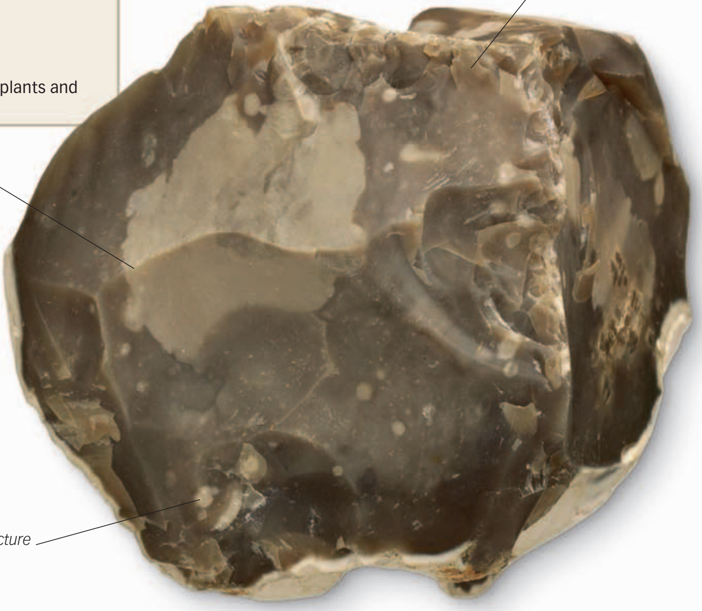

NODULES

A rounded mineral accretion that differs in composition from its surrounding rock is known as a nodule. Nodules are commonly elongated with a knobby irregular surface and are usually oriented parallel to the bedding of their enclosing sediment. Most nodules are formed by the accumulation of silica in sediments and its subsequent solidification. Nodules containing manganese, phosphorous, titanium, chromium, and other valuable metals develop on the seafloor but are uneconomical to mine. Other nodules form around plant and animal remains as part of the fossilization process.

Pyrite (p.62) is commonly found as nodules. It occurs as spheres, rounded cylinders of radiating crystals, and flat, radiating disks, or "suns." Clay ironstone, a mixture of siderite (p.123) and clay, sometimes occurs as layers of dark gray-to-brown nodules overlying coal seams. Chert (p.332) and flint often occur as nodules of nearly pure cryptocrystalline quartz (p.168) within beds of limestone (p.319) or chalk (p.321).

VARIANTS

Melanterite A nodule of hydrous iron sulfate

Pyrite nodule A nodule of radiating pyrite crystals

Marine nodule A nodule of manganese oxides

PROFILE

- Chemical
- Postdepositional
- Not granular
- Chalcedony
- None
- Gray, brown
- Invertebrates, plants

Grainy chert
This specimen of fine-grained chert shows the way the rock breaks along flat to rounded surfaces.

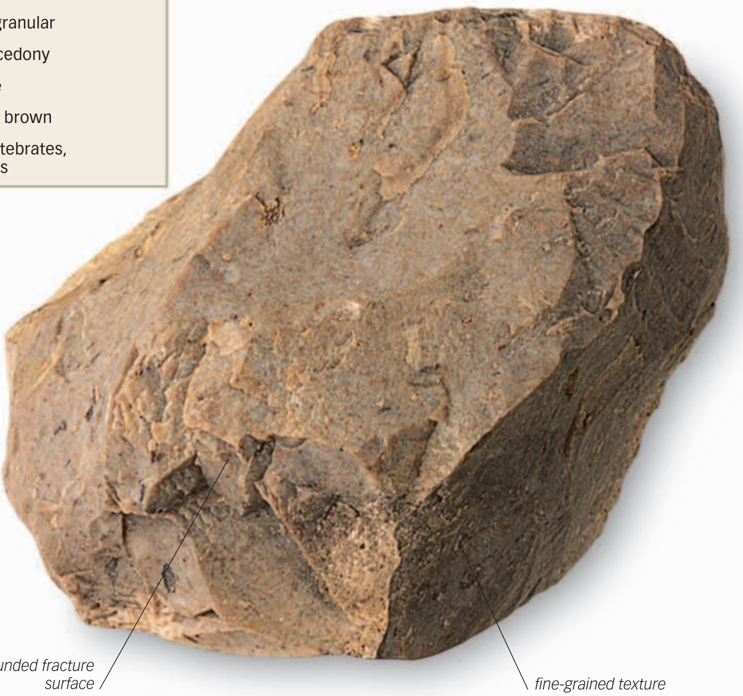

VARIANT

Fossiliferous chert
A specimen containing fossilized primitive plants

CHERT

A rock composed of microcrystalline silica is known as chert. It is commonly gray, white, brown, or black and may contain small fossils. Traces of iron (p.39) may give chert a light green to rusty-red color. Chert occurs as beds or nodules. It breaks along flat to rounded, smooth surfaces and has a glassy appearance. The rock forms by precipitation from silica-rich fluids and colloids.

Flint is chert that occurs in marly limestones (p.319) or chalk (p.321). It is generally gray in color. Flint is found as nodules, often in bands parallel to the bedding planes, and breaks with a conchoidal fracture. The nodules are irregular but rounded. Flint resists weathering, so it can form thick pebble beds on beaches and accumulate in soils that are derived from chalk.

Flint ax
A prized material for tool making, flint was one of the earliest rocks to be mined and extracted.

PROFILE

- Chemical
- Postdepositional
- Crystalline
- Calcite, silica
- Celestine
- Gray, brown
- Invertebrates, plants

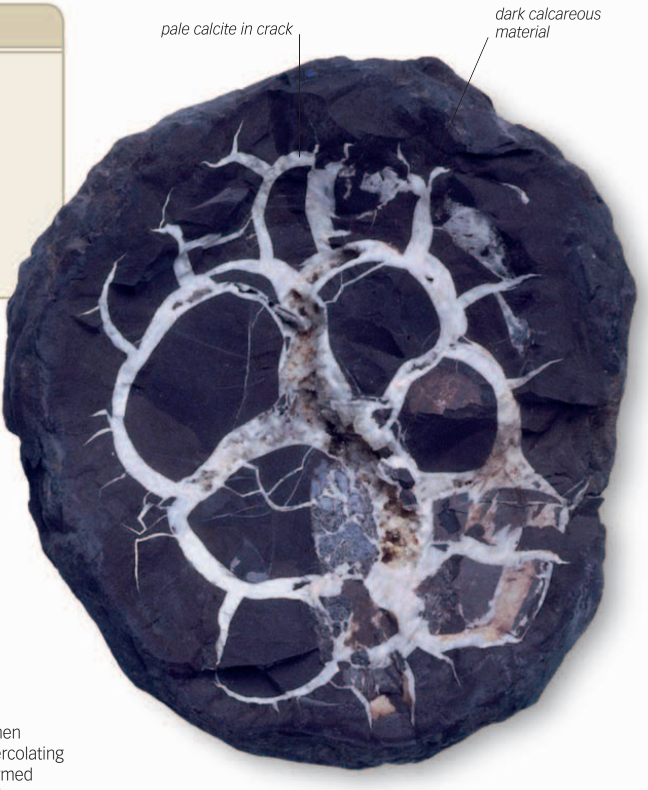

Septarian concretion
This concretion formed when crystals deposited from percolating brines filled cracks that formed inside a nodule as it shrank.

VARIANT

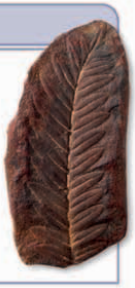

Ironstone concretion
Pteridosperm (seed fern) fossilized in an ironstone concretion

CONCRETIONS

Nodules and concretions are distinctly different geologically, although they have some visual similarities. Unlike nodules, which are of different compositions than their hosts, concretions are made of the same material as their host sediment and cemented by other minerals, mainly calcite (p.114), iron oxides, and silica. Concretions are often much harder and more resistant to erosion than their surrounding rock and can be concentrated by weathering. They usually form early in the burial history of the sediment before the rest of the sediment is hardened into rock.

Concretions vary in shape, hardness, and size. They can be so small that they need to be seen under a magnifying lens. Specimens can also be huge bodies 9¾ ft (3 m) or more in diameter and weighing several tons. In some localities, fossil collectors seek out concretions that have formed around plant and animal remains and perfectly fossilized them. The largest fossil from concretions was an almost complete hadrosaur—a type of dinosaur.

METEORITES

Meteorites are rocks that formed elsewhere in our Solar System and orbited the Sun before colliding with Earth and falling to its surface. The fragments most likely to survive are either very large (weighing ⅓oz/10g or more) or very small (1mg or less).

STRUCTURE

There are three basic types of meteorite: irons, composed mostly of metallic iron and varying amounts of nickel; stony-irons, composed of a mixture of iron and silicate minerals; and stony meteorites, which are further classified into chondrites and achondrites, depending on the presence or absence of small igneous, silicate spheres called chondrules. The most common type of stony-irons are pallasites, which are iron meteorites with centimetre-sized, translucent, olivine or pyroxene crystals scattered throughout.

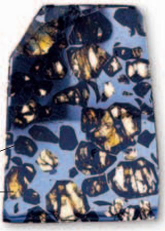

Pallasite section
The stony-iron meteorites known as pallasites are usually one part olivine crystals to two parts metal. In this specimen, olivine is more abundant.

small to medium grains

Achondrite
This stony meteorite found in Haryana, India, lacks chondrules, the small spheres of igneous minerals that would make it a chondrite.

OCCURRENCE

Meteorites are found on every continent. They are often found in environments where they are highly visible—ice caps and deserts being examples. They range in size from microscopic specimens to masses that weigh many tons. Rocks that have been melted by meteorite impacts called tektites, are also widespread.

THE ORIGIN OF THE SOLAR SYSTEM

Meteorites are of particular interest in the study of the origin of the Solar System. Nearly all meteorites are thought to be fragments of asteroids—rocky bodies that formed in the solar nebula at about the same time as Earth. Radiometric dating puts the age of most meteorites at about 4.5 billion years, the same age as Earth.

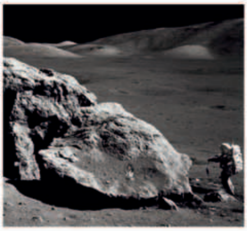

Lunar discoveries
Over 100 meteorites are now believed to be material dislodged from the Moon, providing a more representative sample of the lunar surface than those brought back by astronauts.

Barringer Crater
This well-preserved impact crater is in the Arizona Desert, USA, and is ¾ mile (1.2km) in diameter. The meteorite that created it fell about 50,000 years ago and was about 165ft (50m) across.

PROFILE

- Small
- Iron, nickel
- Olivine, pyroxene, graphite
- Steel gray on fresh surfaces, black on the outside

Iron meteorite The kamacite and taenite crystals in this iron meteorite occur as fine, crosshatched lines, forming what is known as a Widmanstätten pattern.

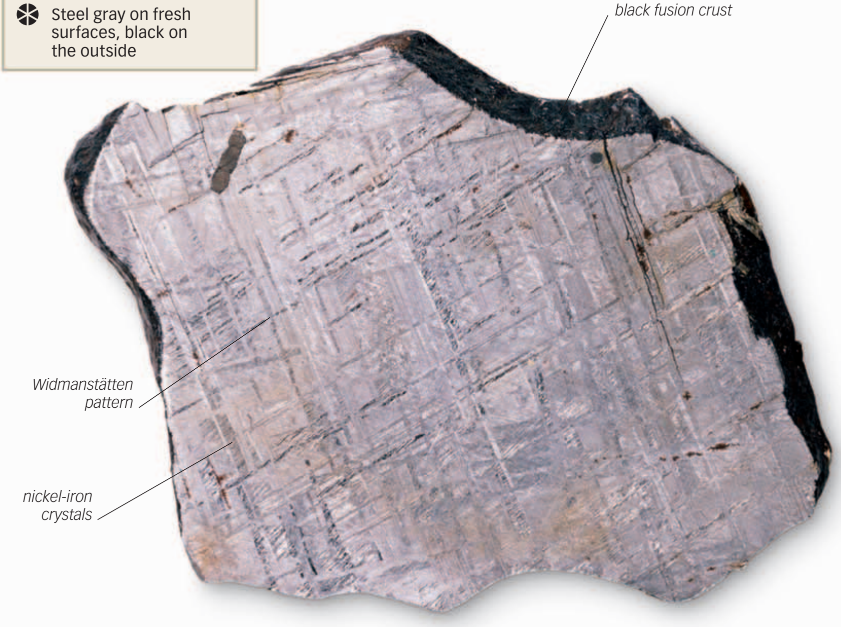

VARIANT

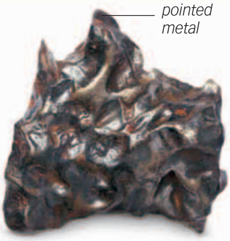

Fusion surface An iron meteorite with a partially vaporized surface

IRON METEORITE

These meteorites are believed to be the shattered fragments of the formerly molten cores of large, ancient asteroids. As the name suggests, iron meteorites are composed mostly of metallic iron (p.39) and up to 25 percent nickel, to produce two minerals, kamacite and taenite, which are otherwise rare on Earth. Other minerals present in minor amounts in iron meteorites include troilite, graphite (p.46), phosphides, and some silicates—most commonly olivine (p.232) and pyroxene. Iron meteorites are divided into a number of subgroups based on their chemical makeup.

Since iron meteorites are so different from most terrestrial rocks, they are recognized more often than other kinds of meteorites and tend to be over represented in meteorite collections. It is estimated that only about 6 percent of meteorites are iron meteorites. Chemical and isotope analysis of a number of iron meteorites indicates that at least 50 distinct parent asteroids were involved in their creation.

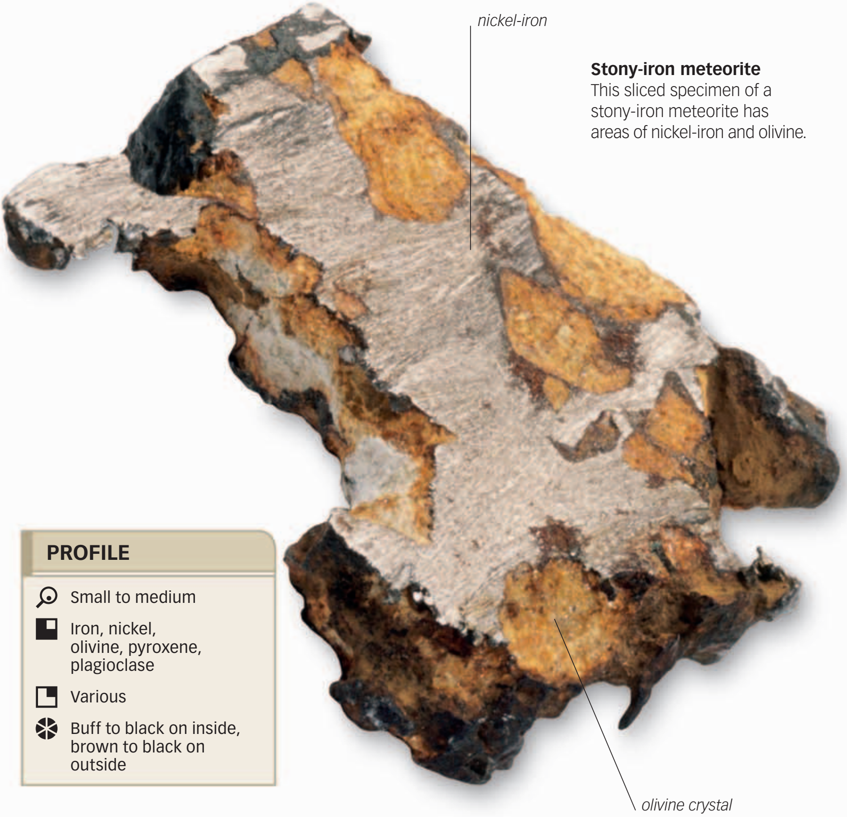

Stony-iron meteorite This sliced specimen of a stony-iron meteorite has areas of nickel-iron and olivine.

PROFILE

- ⌕ Small to medium
- ◩ Iron, nickel, olivine, pyroxene, plagioclase
- ◱ Various
- ✱ Buff to black on inside, brown to black on outside

VARIANTS

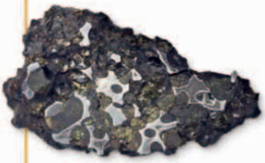

Meteorite from Antarctica Stony-iron meteorite from the Thiel Mountains, Antarctica

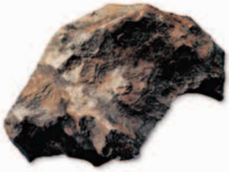

"Thumbprint" surface Meteorite from Chile with a melted surface resembling a thumbprint

STONY-IRON METEORITE

Composed of a mixture of metallic nickel-iron and silicate minerals, stony-iron meteorites make up about 1 percent of recovered meteorites. Pallasites, the most common type of stony-iron meteorites, are iron meteorites with inch-sized, translucent crystals of olivine (p.232), or sometimes pyroxene, scattered throughout. Pallasites were believed to have originated along the boundaries of the iron cores and silicate mantles of large asteroids. But some experts now believe they are impact-generated mixtures of core and mantle materials.

The other main group of stony-iron meteorites are known as the mesosiderites. Relatively rare, these consist of about equal parts of metallic nickel-iron and silicate minerals, namely pyroxene, olivine, and calcium-rich feldspar (pp.173–81). They have an irregular, often brecciated texture, with the silicates and metal occurring as lumps, pebbles, or fine-grained intergrowths.

PROFILE

- 🔍 Small to medium
- ◩ Olivine, pyroxene, plagioclase
- ◳ Various
- ✱ Buff to black on inside, brown to black on outside

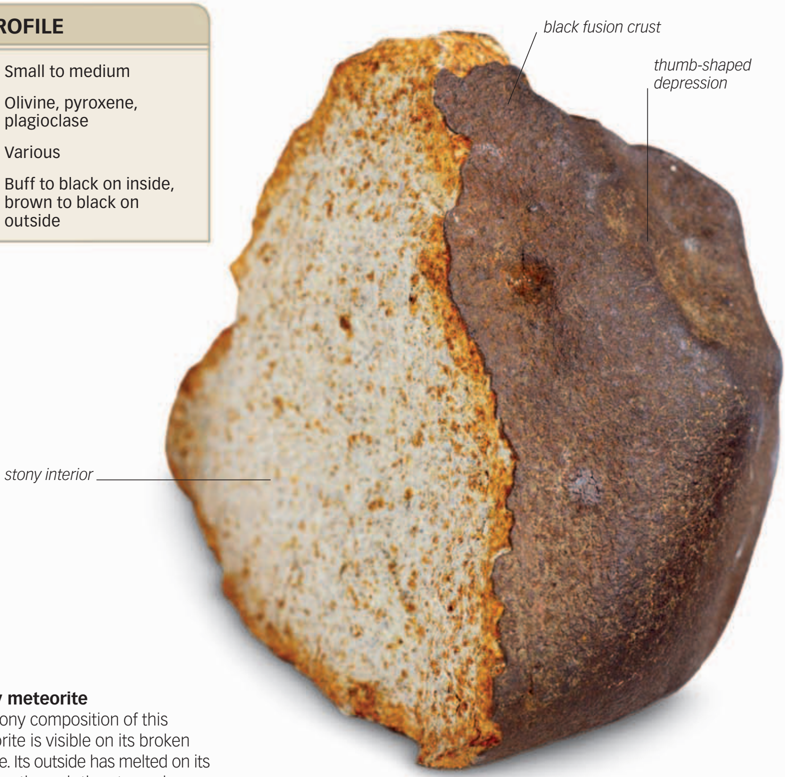

Stony meteorite
The stony composition of this meteorite is visible on its broken surface. Its outside has melted on its passage through the atmosphere.

VARIANTS

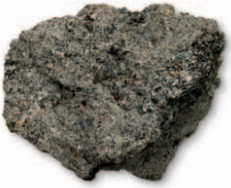

Achondrite A specimen with more varied composition than chondrites

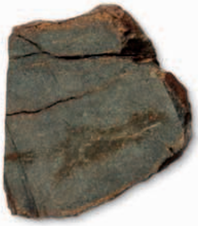

Chondrite Variant of stony meteorites usually composed of olivine and pyroxene

STONY METEORITE

These meteorites can be divided into chondrites and achondrites, which are named after chondrules—small, igneous, silicate spheres that dominate most chondritic meteorites. Chondrules are largely made up of pyroxene and olivine (p.232), with lesser amounts of glass, and may contain calcium-rich feldspar (pp.173–81). They can be up to ⅜ in (1 cm) wide. Chondritic asteroids are some of the oldest and most primitive astronomical bodies in the Solar System. They may have formed by rapid heating in the cloud of dust from which the Solar System originated. More than 80 percent of the meteorites that fall to Earth are chondrites.

Achondritic meteorites, in which chondrules are absent, are believed to represent the crust or mantle of asteroids that developed separate mineral shells around an iron-nickel core. They are similar to igneous rocks (pp.258–85) and impact breccias on Earth. Although most achondrites appear to come from asteroids, about 20 of those recovered are believed to have originated on Mars and another 20 on the Moon.

GLOSSARY

ACCESSORY MINERAL
A mineral that occurs in a rock in such small amounts that it is disregarded in the definition of the rock.

ACICULAR HABIT
A needlelike crystal habit of some minerals. See also *habit.*

ADAMANTINE LUSTER
A type of bright mineral luster similar to that of diamond. See also *luster.*

AGGREGATE
An accumulation of mineral crystals or rock fragments.

ALKALINE ROCK
A class of igneous rocks abundant in potassium- and sodium-rich minerals.

ALTERATION
The chemical, thermal, or pressure process or processes by which one rock or mineral is changed into another.

ALTERATION PRODUCT
A new rock or mineral formed by the alteration of a previous one. See also *alteration.*

ALUM
Any of a group of hydrated double salts, usually consisting of aluminum sulfate, water of hydration, and the sulfate of another element.

AMYGDALE
A secondary in-filling of a void in an igneous rock. Minerals that occur as amygdales include quartz, calcite, and the zeolites.

ASSOCIATED MINERALS
Minerals found growing together but not necessarily intergrown. See also *intergrowth.*

AUREOLE
The area around an igneous intrusion where contact metamorphism has occurred.

BASAL CLEAVAGE
Cleavage that occurs parallel to the basal crystal plane of a mineral. See also *cleavage.*

BATHOLITH
A huge, irregularly shaped mass of igneous rock formed from the intrusion of magma at depth. See also *magma*.

BLADED HABIT
A crystal habit in which wide, flat crystals appear similar to knife blades. See also *habit*.

BOTRYOIDAL HABIT
A mineral habit in which crystals form globular aggregates similar to bunches. See also *aggregate*, *habit*.

BRECCIA
A sedimentary rock made up of angular fragments. See also *igneous breccia*.

CABOCHON
A gemstone cut with a domed upper surface and a flat or domed under surface; gemstones cut in this way are said to be cut *en cabochon*. See also *cut*, *gem*, *gemstone*.

CARAT
A unit of gemstone weight, equivalent to 0.007 oz (0.2 g). Carat (also spelled "Karat") is also a measure of gold purity, the number of parts of gold in 24 parts of a gold alloy: 24 kt is pure gold; 18 kt is three quarters gold. See also *gem, gemstone*.

CHATOYANCY
The cat's-eye effect shown by some stones cut *en cabochon*. See also *cabochon*.

CLAST
A fragment of rock, especially when incorporated into a sedimentary rock.

CLASTIC ROCK
A sedimentary rock composed of cemented clasts. See also *clast*.

CLAY
Mineral particles smaller than about 0.00008 in (0.002 mm).

CLEAVAGE
The way certain minerals break along planes dictated by their atomic structure.

COLLOID
Any substance that consists of particles substantially larger than atoms or molecules but too small to be visible to the unaided eye.

COLOR DISPERSION
The separation of white light into its constituent colors.

CONCHOIDAL FRACTURE
A curved or shell-like fracture in many minerals and some rocks. See also *fracture*.

CONCRETION
A rounded, nodular mass of rock formed from its enclosing rock and commonly found in beds of sandstone, shale, or clay.

CONTACT TWINNING
The phenomenon of two or more crystals growing in parallel contact with each other and sharing a common face. See also *twinned crystals*.

CRYPTOCRYSTALLINE HABIT
A mineral habit that is crystalline but very fine-grained. Individual crystallized components can be seen only under a microscope. See also *habit*.

CUT
The final shape of a ground and polished gem, as in emerald cut. See also *gem*, *gemstone*, *gem cutting*.

DENDRITIC HABIT
A type of habit in which crystals form branching, treelike shapes. See also *habit*.

DETRITAL ROCKS
Sedimentary rocks formed essentially of fragments and grains derived from existing rocks. See also *sedimentary rock.*

DIATOMACEOUS EARTH
A sedimentary rock composed of the siliceous shells of microscopic aquatic plants known as diatoms.

DIFFRACTION
The splitting of light into its component colors. See also *x-ray diffraction*.

DIKE
A sheet-shaped igneous intrusion that cuts across existing rock structures.

DISCOVERY/TYPE LOCALITY
The site where a mineral was first recognized as a new mineral.

DOUBLE REFRACTION
The splitting of light into two separate rays as it enters a stone.

DRUSY COATING
A sheet of numerous, often well-formed, small crystals covering a mineral.

DULL LUSTER
A type of luster in which little or no light is reflected. See also *luster*.

EARTHY LUSTER
A nonreflective mineral luster. See also *luster*.

EQUANT
A mineral habit that refers to a crystal that is of equal size in all directions; a rock composed of grains of equal size.

EUHEDRAL
A term describing crystals with well-formed faces.

EVAPORITE
A mineral or rock formed by the evaporation of saline water.

EXTRUSIVE ROCK
A rock formed from lava that either flowed onto Earth's surface or was ejected as pyroclastic material. See also *intrusive rock*, *lava*.

FACES
The external flat surfaces that make up a crystal's shape.

FELDSPATHOIDS
Minerals similar in chemistry and structure to the feldspars but with less silica.

FELSIC ROCK
An igneous rock with more than 65 percent silica and more than 20 percent quartz. It is also known as acidic rock.

FISSILE TEXTURE
A rock texture that allows the rock to be split into sheets. See also *texture*.

FLOW BANDING
Layering in a rock that originated when the rock was in a fluid, molten state.

FOLIATION
The laminated, parallel orientation or segregation of minerals.

FOSSIL
Any record of past life preserved in the crustal rocks. Apart from bones and shells, fossils can include footprints, excrement, and borings.

FRACTURE
Mineral breakage that occurs at locations other than along cleavage planes. See also *cleavage.*

FUMAROLE
In volcanic regions, an opening in the ground through which hot gases are emitted.

GARNET
A member of a group of silicates with the general formula A3B2(SiO4)3 in which A can be Ca, Fe^{2+}, Mg, or Mn^{2+}; and B can be Al, Cr, Fe^{3+}, Mn^{3+}, Si, Ti, V, or Zr.

GEODE
A hollow, generally rounded nodule lined with crystals. See also *nodule*.

GEM, GEMSTONE
A cut stone worn in jewelry, valued for its color, rarity, texture, or clarity. It may even be an unset stone cut for use as jewelry. See also *cut*, *rough gemstone*.

GEM CUTTING
The process of shaping a gemstone by grinding and polishing. See also *cut*, *gem*, *gemstone*.

GLASS
A solid substance showing no crystalline structure—in effect, a very thick liquid. See also *glassy texture*.

GLASSY TEXTURE
The smooth consistency of an igneous rock in which glass formed due to rapid solidification. See also *glass, texture*.

GRANULAR TEXTURE
A rock or mineral texture that either includes grains or is in the form of grains. See also *texture*.

GRAPHIC TEXTURE
The surface appearance of some igneous rocks in which quartz and feldspar have intergrown to produce an effect resembling a written script. See also *texture*.

HABIT
The mode of growth and appearance of a mineral. The habit of a mineral results from its molecular structure.

HACKLY FRACTURE
A mineral fracture that has a rough surface with small protuberances, as on a piece of broken cast iron. See also *fracture*.

HEMIMORPHIC FORM
A crystalline form with a different facial development at each end.

HOPPER CRYSTAL
A crystal whose incomplete growth has created a hopper-shaped concavity on one or more faces.

HYDROTHERMAL DEPOSIT
A mineral deposit that is formed by hot water ejected from deep within Earth's crust.

HYDROTHERMAL MINERAL
A mineral that is derived from hydrothermal deposition. See also *hydrothermal deposit*.

HYDROTHERMAL VEIN
A rock fracture in which minerals have been deposited by fluids from deep within Earth's crust. See also *hydrothermal deposit*, *pegmatite*, *vein*.

HYPABYSSAL
A term describing minor igneous intrusions at relatively shallow depths within Earth's crust.

IGNEOUS BRECCIA
An igneous rock made up of angular fragments. See also *breccia*.

IGNEOUS ROCK
A rock that is formed through the solidification of molten rock.

INCLUSION
A crystal or fragment of another substance within a crystal or rock.

INTERGROWTH
Two or more minerals that are growing together and interpenetrating each other. See also *associated minerals*.

INTERMEDIATE ROCK
An igneous rock that is intermediate in composition between silica-rich and silica-poor rocks.

INTRUSIVE ROCK
A body of igneous rock that invades older rock. See also *extrusive rock*.

IRIDESCENCE
The reflection of light from the internal elements of a stone, yielding a rainbowlike play of colors.

LAMELLAR HABIT
A type of crystal habit in which plates or flakes occur in thin layers or scales. See also *habit*.

LAVA
Molten rock that is extruded onto Earth's surface. See also *magma*.

LITHIFICATION
The process by which unconsolidated sediment turns to stone. See also *recrystallization*.

LUSTER
The shine of a mineral caused by reflected light.

MAFIC ROCK
An igneous rock with 45–55 percent silica. Such rocks have less than 10 percent quartz and are rich in iron-magnesium minerals.They are also known as basic rocks. See also *ultramafic rock*.

MAGMA
Molten rock that may crystallize beneath Earth's surface or be erupted as lava. See also *lava*.

MAMILLARY HABIT
A mineral habit in which crystals form rounded aggregates. See also *aggregate, habit*.

MASSIVE
A mineral form having no definite shape.

MATRIX
A fine-grained rock into which or on top of which larger crystals appear to be set. It is also known as a groundmass.

METAL
A substance that is characterized by high electrical and thermal conductivity as well as by malleability, ductility, and high reflectivity of light.

METALLIC LUSTER
A shine similar to the typical shine of polished metal. See also *luster*.

METAMORPHIC ROCK
A rock that has been transformed by heat or pressure (or both) into another rock.

METEOR
A rock from space that completely vaporizes while passing through Earth's atmosphere.

METEORITE
A rock from space that reaches Earth's surface.

MICA
Any of a group of hydrous potassium or aluminum silicate minerals. These minerals exhibit a two-dimensional sheet- or layerlike structure.

MICROCRYSTALLINE HABIT
A mineral habit in which crystals are so minuscule that they can be detected only with the aid of a microscope. See also *habit*.

MINERAL GROUP
Two or more minerals that share common structural and/or chemical properties.

MOONSTONE
A gem-quality feldspar mineral that exhibits a silvery or bluish iridescence. Several feldspars, especially some plagioclases, are called moonstone.

NATIVE ELEMENT
A chemical element that is found in nature uncombined with other elements.

NODULE
A generally rounded accretion of sedimentary material that differs from its enclosing sedimentary rock.

NONMETAL
An element, such as sulfur, which lacks some or all of the properties of metals. See also *metal*.

OOLITHS
Individual spherical sedimentary grains from which

oolitic rocks are formed. Most ooliths comprise concentric layers of calcite.

ORE
A rock or mineral from which a metal can be profitably extracted.

OXIDATION
The process of combining with oxygen. In minerals, the oxygen can come from the air or water.

PEGMATITE
A hydrothermal vein composed of large crystals. See also *hydrothermal vein*.

PENETRATION TWINNING
The phenomenon of two or more crystals forming from a common center and appearing to penetrate each other. See also *twinned crystals*.

PHENOCRYST
A large crystal set in an igneous rock matrix, creating a porphyritic texture. See also *porphyritic texture*.

PISOLITIC HABIT
A mineral habit characterized by pea-sized grains with a concentric inner structure. See also *habit*.

PLACER, PLACER DEPOSIT
A deposit of minerals derived by weathering and concentrated in streams or beaches because of the mineral's high specific gravity.

PLASTIC ROCK
A type of rock that is easily folded when subjected to high temperature and pressure.

PLATY HABIT
The growth habit shown by flat, thin crystals. See also *habit*.

PLAYA DEPOSIT
A mineral deposit formed in a desert basin that is intermittently filled with a lake.

PLEOCHROIC, PLEOCHROISM
The phenomenon of a mineral or gem presenting different colors to the eye when viewed from different directions.

PLUTON
A mass of igneous (plutonic) rock that has formed beneath Earth's surface by the solidification of magma.

POLYMORPH
A substance that can exist in two or more crystalline forms; one crystalline form of such a substance. See also *pseudomorph*.

PORPHYRITIC TEXTURE
An igneous rock texture in which large crystals are set in a finer matrix. See also *phenocryst*, *texture*.

PORPHYROBLAST
A relatively large crystal set in a fine-grained matrix in a metamorphic rock.

PORPHYROBLASTIC TEXTURE
A texture characterized by relatively large crystals in a fine-grained matrix. See also *texture*.

PRECIPITATION
The condensation of a solid from a liquid or gas.

PRIMARY MINERAL
A mineral that has crystallized directly from an igneous magma and is unaltered by rain, groundwater, or other agents. See also *secondary mineral*.

PRISMATIC HABIT
A mineral habit in which parallel rectangular crystal faces form prisms. See also *habit*.

PROTOLITH
A rock that existed prior to undergoing metamorphic transformation into a different rock type.

PSEUDOMORPH
A crystal with the outward form of another species of mineral. See also *polymorph*.

PYRAMIDAL HABIT
A crystal habit in which the principal faces join at a point. When two such pyramids are placed base to base, the crystal is said to be di- or bi-pyramidal. See also *habit*.

PYROCLASTIC ROCK
A rock consisting of airborne material ejected from a volcanic vent.

PYROXENE
A member of a group of 21 rock-forming silicate minerals that typically form elongate crystals.

RADIOMETRIC DATING
The determination of absolute ages of minerals and rocks by measuring certain radioactive and radiogenic atoms in them.

RARE-EARTH MINERAL
A mineral containing a significant portion of one or more of the 17 rare-earth elements, principally ytterbium, gadolinium, neodymium, praseodymium, cerium, lanthanum, yttrium, and scandium.

RECRYSTALLIZATION
The redistribution of components to form new minerals or mineral crystals; in some cases new rocks form. It occurs during lithification and metamorphism. See also *lithification*.

REFRACTIVE INDEX
A measure of the slowing down and bending of light as it enters a stone. It is used to identify cut gemstones and some minerals. See also *cut*, *gem*, *gemstone*.

RENIFORM HABIT
A mineral habit with a kidneylike appearance. See also *habit*.

REPLACEMENT DEPOSIT
A deposit formed from minerals that have been altered. See also *alteration product*.

RESINOUS LUSTER
A shine having the reflectivity of resin. See also *luster*.

RETICULATED
Having a network or a netlike mode of crystallization.

ROCK FLOUR
Very fine-grained rock dust, often the product of glacial action.

ROUGH GEMSTONE
An uncut gemstone. See also *cut, gem*, *gemstone.*

SALT DOME
A large, intrusive mass of salt, sometimes with petroleum trapped beneath.

SCHILLER EFFECT
The brilliant play of bright colors in a crystal, which is often due to minute, rodlike inclusions.

SCHISTOSITY
A foliation that occurs in coarse-grained metamorphic rocks. It is the result of platy mineral grains. See also *foliation*.

SCORIACEOUS
A term for lava or other volcanic material that is heavily pitted with hollows and cavities. See also *lava*.

SECONDARY MINERAL
A mineral that replaces another mineral as a result of weathering or alteration. See also *alteration*, *primary mineral*.

SECTILE
The property of a mineral that allows it to be cut smoothly with a knife. See also *fracture*.

SEDIMENTARY ROCK
A rock that either originates on Earth's surface as an accumulation of sediments or precipitates from water.

SEMIMETAL
A metal, such as arsenic or bismuth, that is not malleable. See also *metal*.

SILICA-POOR ROCKS
Rocks containing less than 50 percent silica. See also *silica-rich rocks*.

SILICA-RICH ROCKS
Rocks containing more than 50 percent silica. See also *silica-poor rocks*.

SLATY CLEAVAGE
The tendency of a rock, such as slate, to break along flat planes into thin, flat sheets. See also *cleavage*.

SOLID-SOLUTION SERIES
A series of minerals in which certain chemical components are variable between two end-members with fixed composition.

SPECIFIC GRAVITY
The ratio of the mass of a mineral to the mass of an equal volume of water. Specific gravity is numerically equivalent to density (mass divided by volume) in grams per cubic centimeter.

SPHEROIDAL HABIT
A crystal habit in which numerous crystals radiate outwards to form a spherical mass. See also *habit*.

STALACTITIC HABIT
A mineral habit in which the crystalline components are arranged in radiating groups of diminishing size, giving the appearance of icicles. See also *habit*.

STRIATION
A parallel groove or line appearing on a crystal.

SUBLIMATION, SUBLIMATE
The process by which a substance moves directly from a gaseous state to a solid state. A sublimate is the solid product of sublimation.

SUNSTONE
A gemstone variety of feldspar with minute, platelike inclusions of iron oxide oriented parallel to one another throughout. See also *gem*, *gemstone*.

TABULAR HABIT
A crystal habit in which the crystals have long, flat, parallel faces. See also *habit*.

TERMINATION
Faces that make up the ends of a crystal.

TEXTURE
The size, shape, and relationships between rock grains or crystals.

TWINNED CRYSTALS
Crystals that grow together as mirror images with a common face (contact twins) or grow at angles up to 90 degrees to each other and appear to penetrate each other (penetration twins).

ULTRAMAFIC ROCK
An igneous rock with less than 45 percent silica. It is also known as an ultrabasic rock. See also *mafic rock*.

VEIN
A thin, sheetlike mass of rock that fills fractures in other rocks.

VESICLE
A small, spherical or oval cavity produced by a bubble of gas or vapor in lava, left after the lava has solidified. See also *lava*.

VITREOUS LUSTER
A shine resembling that of glass. See also *luster*.

VOLCANIC PIPE
A fissure through which lava flows. See also *lava*.

WELL-SORTED ROCK
A sediment or sedimentary rock with grains or clasts that are roughly of the same size.

X-RAY DIFFRACTION
The passing of x-rays through a crystal to determine its internal structure by the way in which the x-rays are scattered. See also *diffraction*.

ZEOLITE
A group of hydrous aluminum silicates characterized by their easy and reversible loss of water.

INDEX

Page numbers in **bold** indicate main entries.

A

a axis, crystals 20, 22–23
abalone 249
acanthite **49**, 54, 69, 72
accessory minerals 10
achondrites 334, 337
achroite 224
acicular crystals 21
acicular epidote 230
acids, for cleaning 32–33
acmite 209
actinolite 90, 213, 217, 219, **220**, 291, 296
Adam, G.J. 160
adamite **160**, 161, 165
Aegir 209
aegirine 76, 182, **209**, 262
agate 168, 169, **170**
age-dating, radiometric 309, 334
agglomerate 283, **284**
aggregates, crystal 21
ailsite 222
alabaster 114
albite 76–77, 86, **177**, 262
  in granite 11
  solid-solution series 167
  twinning 19
alexandrite 81
alexandrite effect 244
almandine 221, **243**
aluminum 98–99, 101, 108, 326
  and halides 106
  ores 100
aluminum oxide 74, 95, 101
aluminum phosphate 158
aluminum silicate 194, 236–37
amazonite 175
amber 13, 246, 250, **251**
amesite 191
amethyst 13, 148, 168, 170
amphibole 191, 270, 278–79, 317
amphibolite 180, 218, 231, **296**
analcime 185, **190**
anatase **76**, 77
  and other minerals 124–25, 139, 160–62, 203
andalusite **236**, 294
Andes Mountains 181
andesine 167, **181**, 264, 275
andesite 257, **275**
  and minerals 180–81, 188–89, 207
  and other rocks 274, 278
andradite 302
anglesite 54, 119, **132**
anhydrite 67, 128–29, **133**, 323–24
Anker, M.J. 122
ankerite **122**
annabergite 163
annivite 73
anorthite 167, **179**
anorthoclase 167, **176**
anorthosite 179, **261**
anthophyllite **216**, 300
anthracite 253
antigorite 191
antimony **41**, 45, 61, 65, 68, 71, 73
antimony sulfide 61
Apache Tears 280
apatite 13, 17, **148**
  and other minerals 85, 153–54, 156, 176
  and rocks 262, 321
  crystal system 22
aplite 186
apophyllite 11, 186, 188, **204**
appearance, crystal 21
aquamarine 134, 148, 225, 260
aragonite **115**, 246–49
arfvedsomite 262
argentite 49
argentopentlandite 55
arkose **311**
arsenates 13, **147**, 160–165
arsenic 41, **45**, 58, 59, 72, 73, 151
  in arsenates 147
  native element 36
arsenic sulfide 59
arsenopyrite 83, 148
asbestos 192
  actinolite 220
  anthophyllite 216
  blue 222
  glaucophane 221
  riebeckite 222
  tremolite 219
ash tuff 282
asterism effect 78
asteroids 27, 334–35, 337
atacamite 106
atomic structure 18
augen gneiss 288
augite 176, 179, 182, **211**, 265
  in gabbro 11
aureole 287
autunite **149**, 152
axinite 23, **228**
axis, crystallographic 20, 22–23
azurite 10, 14, 113, **120**

B

b axis, crystals 23
babbitt metals 41
bags, for collecting 32
baguette cut 25
banded iron formation 329
banded ironstone **329**
barite 11, 61, 66, 73, 103, 116, **134**, 155, 164
  cleavage 16
  crystal structure 131
barium 94, 113, 116
barium carbonate 116
barium sulfate 134
basalt 257, **273**
  and other minerals 96, 180, 204, 207, 211
  and other rocks 265–66, 299, 317
  zeolites 186–90
batholith 256
battery, lithium 198
bauxite 97, 100, **101**, 104, 194, 326
beaches, collecting on 28–29
bentonite **200**
Bergmann, Torbern Olaf 152
beryl 20, 198, **225**, 240
beryllium 225
beryllium aluminum oxide 81
Biot, Jean-Baptiste 197
biotite 11, **197**
  and extrusive igneous rocks 274–75, 278–79
  and intrusive igneous rocks 258, 262–64
  and metamorphic rocks 288, 291–92, 294
  and other minerals 181, 199, 233
  and subvolcanic igneous rocks 270
biotite schist 292
bismuth **40**, 41, 66, 73
bitumen 246
bituminous coal 253
bixbyite 98
black coral 247
"black granite" 264, 268
black-lip shell 248
black mica 197
black silver ore 69
bladed crystals 21
blastomylonite 289
blister pearl 248
blocky crystals 21
blue asbestos 222
blue coral 247
Blue John 109
blue labradorite 180
blue ocher 157
blue porphyry 181
blue vitriol 138
bluestone 268
boart diamond 47
bog iron 102, **327**
bohernite 100
boleite 323
bomb, volcanic 282, **285**
bomb tuff 282
boracite, crystal structure 126
borates 13, **126**, 127–130
borax 126, **127**, 129–30
Born, Ignaz von 50
bornite 10, **50**, 51–52, 67, 73
  crystal structure 48
boron 126–27, 130
bort diamond 47
botryoidal habit 21
boulder clay 318
boulder-conglomerate 306
boulder opal 172
Bournon, Count J.L. de 71
bournonite 65, **71**
bradawl, for cleaning 32
brain coral 247
breadcrust bomb 285
breccia **307**
  marble 301
  volcanic **283**, 284
Breithaupt, August 178
brilliant cut, diamond 25
brittle minerals 16
brittle silver ore 69
Brochant de Villiers, A.J.M. 139
brochantite **139**, 161–62
bromine 106
bronze 37

Brooke, H.J. 77
brookite 76, **77**
brown topaz 234
brown zircon 233
Bruce, Archibald 105
brucite 13, **105**, 298
"Bruneau jasper" 171
brushes, for cleaning 31–33
butterfly spar 114
bytownite 167, 261

C

c axis, crystals 20, 22–23
calamine 227
calc-tufa 330
calcareous tufa 330
calcite 9, 11, 17, **114**
  and other minerals 88, 103, 112, 115, 121, 141, 249
  and other rocks 269, 272, 296, 301–02
  and sedimentary rocks 309–10, 313, 316–17, 319–21, 330, 333
  and silicate minerals 183, 186, 190, 193, 205, 229
  and sulfur minerals 51, 58, 68, 70, 72
  as a carbonate 113
  as a cementing agent 304
  as an organic mineral 246
  crystal structure 113
  fluorescence 17
  luster 15
  refraction 17
calcium 89
calcium aluminum silicate 205, 231, 245
calcium aluminosilicate 179
calcium bentonite 200
calcium borosilicate hydroxide 129
calcium carbonate 114, 122, 247, 249
calcium magnesium silicate 217, 219
calcium phosphate 148
calcium sulfate hydrate 136
calcium sulfite 133
calcium titanium oxide 89
calcium titanium silicate 235
calcium tungstate 146
calcium uranium phosphate 149
calomel **112**
campylite 164
carbon 36, 46–47
carbonado 47
carbonate cycle 246
carbonates 8, 9, 13, **113**, 114–125, 313
carbonatite 117, 122, **272**, 284
Carnall, Rudolf von 107
carnallite 106, **107**
carnelian 24, 169
Carno, Marie-Adolphe 159
carnotite 147, 152, **159**
cassiterite 74, **79**, 83, 241
cat's eye chrysoberyl 81
catalytic converters 223
caustic potash 107
celestine 13, **135**
cementing in sedimentary rocks 304
cerium 74, 89, 93
cerium oxide 150
cerium phosphate 150
cerussite 54, 113, **119**, 124, 143, 151
  twinning 19
cesium 198, 270, 272
chabazite **189**
chalcanthite 137, **138**
chalcedony 168, **169**, 170, 203
chalcocite 50, **51**, 52
chalcophyllite **162**
chalcopyrite **57**, 138
  and sulfide minerals 50–52, 54–55
  and sulfosalt minerals 66–67, 71, 73
  fracture 16
  streak 15
chalcosine 13, 51
chalcotrichite 87
chalk 63, 196, **321**, 331–32
chalybite 123
chamosite 328
chatoyancy 81
chemical elements, table of 12
chemical sedimentary rocks 305
chert 321, 329, 331, **332**
  as sedimentary rock 305
chessylite 120
chiastolite 236, 303
chiastolite slate 293
"chimney," seafloor hydrothermal 133
chisels, for collecting 30
chlorapatite 148
chlorine 106
chlorite,
  and other minerals 38, 89, 105, 221
  and rocks 289, 291–92, 294, 300, 315, 322
chlorite schist 118
chocolate 201
chondrites 334, 337
chondrules 334, 337
chromates 13, **131**, 142
chrome diopside 210, 269
chrome yellow 142
chromite 74, 96, **99**
chromium 74
  and minerals 96, 99, 104, 142
  and rocks 265–66, 270, 331
  in chromates 131
  in emerald 225
chrysoberyl **81**
chrysocolla 10, 125, **203**, 226
chrysoprase 169
chrysotile 191, **192**
cinnabar **56**, 61, 112
cinnamon stone 245
citrine 168
clams 248
clastic sedimentary rocks 304–05
clay 63, 101, 123, 134, 171, 202, 313
  boulder 318
clay-based minerals 200–206
clay ironstone 331
cleaning specimens 32–33
cleavage 16
cleavelandite 173, 177
clinochrysotile 192
clinoclase **161**, 162
clothing, for collecting 31
coal **253**
  and minerals 60, 62–63, 123, 137, 140, 246, 251–52
  as a hydrocarbon 12
cobalt 45, 55
cobalt arsenate hydrate 163
cobalt bloom 163
cobaltpentlandite 55
cobble-congolmerate 306
cockscomb barite 134
cogwheel ore 65, 71
Coleman, William 130
colemanite 126, 128, **130**
collecting rocks and minerals **28–29**
  code 29
  equipment **30–31**
color of minerals 14
coltan series 82
columbite 84
columbite-tantalite **82**
comet Wild 2 179
common mica 195
compass, magnetic 29
conch 249
conchiolin 247
conchoidal fracture 16
concretion **333**
conglomerate 83, **306**
contact metamorphism 287
continental drift 318
copal 246, **250**, 251
copalite 251
copper **37**
  and other minerals 40, 43, 48, 74, 117, 120, 125
  and rocks 302
  and sulfur minerals 50–52, 54, 57, 60, 67, 73, 137–139
  ductile 16
  native element 8, 10, 13, 36
copper aluminum phosphate 154
copper aluminum silicate 203
copper arsenic sulfide 67
copper carbonate 120
copper glace 51
copper iron arsenic sulfide 66
copper iron sulfide 50, 57
copper oxide 87
copper sulfate hydrate 138–39
copper sulfide 52, 161
coral 246, **247**, 319
Cordier, Pierre L.A. 223
cordierite **223**, 236–37, 294, 303
correction fluid, for labeling 29
corundum 17, 74, 81, **95**, 104, 236–37
corundum schist 292
cotton stones 188
Covelli, Niccolo 52
covellite 51, **52**, 67
Cretaceous Period 321
crocidolite 222
crocoite 15, **142**
cryolite 106, **108**
cryptocrystalline quartz 331
crystal axes 20, 22–23
crystals **18–23**
  aggregates 21
  appearance 21
  axes 20, 22–23
  cubic 19, 20, 22
  faces 20
  forms 20
  habits 18, 20–21
  hexagonal 22
  monoclinic 23
  orthorhomic 23

crystals *continued*
  - structure 18
  - symmetry 19
  - systems 19, 22–23
  - tetragonal 22
  - triclinic 23
  - trigonal 22
  - twinning 19

cubic crystals 20, 22
cuprite 74, **87**, 162
Curie, Pierre and Marie 83
curling 222
cushion cut 25
cuts, gems 25
cuttlefish 206
cycle, rock 27
cyclic twinning 19
cyclosilicates 166, 223–26
cyprine 229

D

dacite 263, **274**, 283
dark ruby silver 70
de Bournon, Count J.L. 71
de Dolomieu, Déodat Gratet 117
de Villiers, A.J.M. Brochant 139
denarius coins 70
dendritic crystals 21
deposition 304
desert rose barite 134
desert rose gypsum 136
diabase 190
diamond 17, **47**
  - and other minerals 36, 46, 81, 90, 95, 99, 201, 233–35, 243
  - and rocks 269–70, 299, 306
  - faceting 25
  - mining 24
  - synthetic 12

diaspore, crystal structure 100
diatomaceous earth **325**
diatomite **325**
diatoms 325
dickite 201
dike 256
dinosaurs 333
diopside **210**
  - and other minerals 209, 211, 214, 219, 229, 245
  - and rocks 297, 299, 302

dioptase **226**
diorite 90, 178, 180–81, 263, **264**, 275
dirty sandstone 317
diaspore **104**
displaying specimens 33
dodecahedral crystal form 20
dogtooth spar 114
dolerite 186, 207, **268**
Dolomieu, Déodat Gratet de 117
dolomite,
  - and other minerals 54, 60, 68, 122, 140, 184, 210, 217, 241
  - and rocks 272, 284, 301–02, 314, 324
  - as a carbonate 113
  - as a mineral **117**
  - as a rock **320**
  - occurrence 9

dolomitization 320
dolostone 320
double refraction 17
"dragon's blood" 56
dravite 224
ductile minerals 16
dynamic metamorphism 286
dysprosium 93

E

earth, diatomaceous **325**
eclogite **299**
elastic minerals 16
elbaite 224
electrical charge 9
elements,
  - native **36**
  - table of 12
  - trace 14

emerald 12, 225, 226
enargite **67**
enstatite **208**
epidote 214, 221, **230**, 296
epsom salts 140
epsomite **140**
equant crystals 21
equipment, for collecting **30–31**
erbium 93
erythrite **163**
essential minerals 10
euclase **240**
eucryptite 212
eudialyte 262
extrusive igneous rocks 26–27, 257, 273–79
eye shadow 183

F

faces, crystal 20, 22–23
faceting, gems 25
facies, grabulite 297
fayalite 232
  - solid solution 10

feldspar porphyry 271
feldspars 9, 93, 173–181
  - and igneous rocks 258–60, 262, 275, 278, 280
  - and metamorphic rocks 288, 290–91, 293–94, 296
  - and minerals 202, 233
  - and sedimentary rocks 310, 313, 316–17, 336–37
  - plagioclase 261, 264
  - series 167
  - twinning 19

feldspathoids 182–184, 270
felsic rocks 257
Ferber, Moritz Rudolph 145
ferberite 64, 79, **145**
Ferguson, Robert 93
fergusonite **93**
ferroactinolite 219, 220
ferroanthophyllite 216
ferroaxinite 228
ferrocolumbite 82
ferroglaucophane 221
ferrohornblende 218
ferrosilite 208
fertilizers 147, 185
fibrous habit 21
field equipment, collecting **30–31**
field techniques, collecting 28–29
fingernail test 17
fire agate 170
fire opal 172
fireworks 126
"fishtail" gypsum 19, 136
flexible minerals 16
flint 321, 331, 332
flos ferri 115
flour, rock 318
flow banding 278, 280
fluorapophyllite 204
fluorescence 17
fluoridated manganese phosphate 153
fluorine 106
fluorite 17, 64, 66, 76, **109**, 121
  - color range 14
  - crystal structure 106

fluorspar 109
fool's gold 62
formanite-(Y) 93
forsterite 10, 193, 232
fortification agate 170
fossils 63, 310, 313, 315–16, 319, 321, 325, 328, 330, 332–33
  - sedimentary rock 305

fracture 16
franklinite 88, 96
frost 75
fuchsite 195, 221
fumaroles 8, 36, 44, 133, 141
Funerary Temple 202

G

gabbro 90, 180, 261, **265**, 267–68, 296
Gahn, John Gottlieb 97
gahnite 96, **97**
galena 11, **54**
  - and carbonates 116, 119, 121
  - and phosphates 151, 155
  - and sulfur minerals 49, 51, 53, 61, 65, 67, 70–73, 132
  - luster 15

garlic odor 59
garnet 11, 242–45
  - and minerals 42, 84, 147, 214
  - and rocks 261, 267, 269, 288, 291, 296–97, 299, 302
  - pyrope 266
  - solid-solution series 166

garnet-chlorite schist 292
garnet hornfels 303
garnet schist **292**
gas, natural 12
gelane 66
gems **24–25**
Genesis Rock 179
gibbsite 100
glacial till 318
glaciers 75
glass 337
  - muscovy 195
  - volcanic 277–78, 280–81

Glauber, Johann 141
Glauber's salt 141
glauberite 128, **141**, 172
glauconite **196**, 309, 321–22
glaucophane **221**, 291
gloves, for collecting 30
gneiss **288**
  - and feldspars 173, 175, 178
  - and nesosilicates 235, 238–39, 243
  - and other minerals 61, 76, 78, 81, 97, 150
  - and other rocks 290, 297, 310

gneiss *continued*
  and other silicates 186, 190, 195, 197, 216, 218, 223, 231
  as metamorphic rock 286–87
Goethe, Johann Wolfgang von 102
goethite 100, **102**
  and halides 107, 112
  and other minerals 87, 103, 120, 151, 155, 158, 161, 165, 171
  and rocks 313, 327–28
goggles, for collecting 30
gold **42**, 48, 146, 201, 272, 302
  crystals 36
  fracture 16
  malleable 16
  native element 8, 13, 36
goshenite 225
gossan 102
granite **258**
  and cyclosilicates 223, 225
  and feldspar minerals 173, 175, 178
  and nesosilicates 234, 236, 244
  and other minerals 61, 64, 78, 97, 108, 150, 158, 222
  and other rocks 260, 262, 264, 270, 288, 290, 302, 310
  and phyllosilicates 195, 197, 201, 204–05
  and zeolites 186–87, 189–90
  "black" 264, 268
  graphic 259
  hornblende 258
  metamorphism 287
  orbicular 259
  porphyritic 259
  textured **259**
granodiorite 175, **263**, 264, 270, 274, 288
granulite 229, 243, **297**
graphic granite 259
graphite 13, **46**, 64
  and rocks 291–94, 301, 335
green apophyllite 204
greensand 196, **309**
greensand marl 322
graywacke 296, 308, **317**
grinding, gems 25
gritstone **312**
grossular 229, **245**
grossular garnet 214
gyprock **323**
gypsum 17
  and other minerals 111, 133, 141
  and other rocks 324
  mineral 131, **136**
  rock **323**

H

habit, crystal 18, 20–21
hackly fracture 16
hafnium 272
hailstones 75
hair, Pele's **281**
halides 13, **106**, 107–12
halite 106, **110**, 111, 127, 133, 324
  cubic crystal 19
halloysite 201
halogens 106
hammers, for collecting 30
hardness 17
hats, for collecting 30
Hausmann, Johann Friedrich Ludwig 98
hausmannite **98**
haüyne 183
heavy spar 134
hedenbergite 209, 211
heliodor 225
hematite **91**
  and other minerals 38, 78, 87, 90, 92, 101, 107, 110, 171, 241
  and rocks 280, 313, 321, 328–29
  luster 15
  streak 15
hemimorphic crystal form 20
hemimorphite 124, 160, **227**
hessonite 245
Heuland, J.H. 187
heulandite 186, **187**, 188
hexagonal crystal system 22
hiddenite 212
history, gems 24
Hope diamond 47
horn mercury 112
horn quicksilver 112
hornblende **218**
  and igneous rocks 262–64, 270
  and metamorphic rocks 288, 291, 296
  and other minerals 181
hornblende granite 258
hornfels 237, **303**
  pyroxene 267
How, Henry 129
howlite **129**
Hübner, Adolf 144
hübnerite **144**, 145
Hume, Sir Abraham 241
humite **241**
hydrocarbons 12
hydroxides 13, **100**, 101–05
hydroxyapophyllite 204
hydroxyl radical 100, 148
hypabyssal intrusive rocks 256
hypersthene 297

I

ice 12, **75**
idiocrase 229
igneous rocks 26–27, **256–57**
  extrusive 26–27, 257, 273–79
  intrusive 26–27, 256, 258–67
  sub-volcanic 268–72
  volcanic 257, 280–85
ignimbrite 282
illite **202**
ilmenite 76, **90**, 176, 265, 269
imperial jade 213
imperial topaz 234
Indian pearl 248
indicolite 224
inosilicates 166, 207–22
intermediate rocks 257
intrusive igneous rocks 26–27, 256, 258–67
iodine 106
iolite 223
iridium 38
iron 36, **39**
  and meteorites 334, **335**, 336–37
  and other minerals 55, 62, 101, 109, 122, 123, 137
  and oxide minerals 77, 91, 92, 98–99
  and rocks 265, 270, 302
  bog **327**
iron chromium oxide 99
iron hat 102
iron magnesium silicate 241
iron mica 197
iron oxide 92, 102, 326
iron roses 91
iron sulfate hydrate 137
iron sulfide 8, 63
iron tungstate 145
ironstone 92, 122, 172, **328**
  banded **329**
  clay 331
  concretion 333
isinglass 195
isometric crystal system 22

J

jack-straw cerussite 119
jade **213**
  nephrite 219–20
jadeite 167, **213**, 221
Jameson, Robert 68
jamesonite **68**
jasper 168, **171**
jet **252**

K

kamacite 39, 335
kaolinite 138, **201**
kauri gum 250
kernite 126, 129
kerogen 314
Khalkedon 169
kidney ore 91
Kimberley 269
kimberlite 24, 47, 90, 96, 210, 266, **269**
Klaproth, M.H. 83
knives, for collecting 31
Krakatoa 277
kunzite 212
kyanite 236, **238**, 291, 295
kyanite schist 291–92

L

labeling specimens 33
labradorite 167, **180**, 261
laccoliths 256
lamellar crystals 21
Lamma, Statue of 162
lamproite 47
lamprophyre **270**
lanthanum 93, 150
lanthanum phosphate 150
lapilli 285
lapilli tuff 282
lapis lazuli 183, 184
larvikite 176
laterite 104, **326**
lattice, crystal 18
lava 257
layered gabbro 265
lazurite **183**
lead,
  and native elements 40, 41, 43
  and other minerals 113, 117, 119, 132, 134, 142, 151
  and rocks 302
  and sulfide minerals 53–54, 64
"lead," pencil 46
lead chloro vanadate 155
lead copper antimony sulfide 71
lead iron antimony sulfide 68

lead sulfide 54
lenses, for collecting 31
lepidolite 86, **198**, 212
Les-Beaux-De-Provence (bauxite) 100
leucite **185**
leucogabbro 265
liddicoacite 224
light red silver ore 72
lignite 251, 253
lilac jadeite 213
limestone 305, **319**
  and carbonate minerals 114, 117–18, 122
  and metamorphic rocks 301, 302
  and other minerals 60, 63, 96–97
  and sedimentary rocks 313–14, 317, 320, 323–24, 331–32
  and silicate minerals 184, 196, 210, 214, 224, 229, 241
limestone breccia 307
limonite 100, 151, 160
lipstick 194
lithification 304
lithium 86, 97, 198, 212
lizardite 191
lodestone 92
lunar rocks *see* Moon
luster 15

M

maceral 253
mafic rocks 257
magma 26–27, 48, 256–57
magnahumite 241
magnesia, medical 105
magnesioanthophyllite 216
magnesioaxinite 228
magnesiohornblende 218
magnesite **118**, 206
magnesium 98–99, 107, 118, 122, 270
magnesium carbonate 118
magnesium hydroxide 105
magnesium silicate 191
magnesium sulfate 140
magnetite **92**
  and igneous rocks 262, 265, 267, 269, 272
  and other minerals 85, 90, 102, 181
  and other rocks 298, 310, 329
  in gabbro 11
  streak 15
magnifiers, for collecting 31
malachite 10, **125**
  and other minerals 113, 124, 139, 161, 162, 203
malleable minerals 16
mamillary jasper 171
manganaxinite 228
manganese,
  and hydroxide minerals 102–03, 105
  and other minerals 80, 94, 98, 113, 122, 144–45
  and oxide minerals 74, 80, 94, 98
  and rocks 331
  and silicate minerals 215–16
manganese aluminum silicate 244
manganese bronze 80
manganese carbonate 121
manganese oxide 94, 103
manganite **103**, 104
maps, for collecting 29
marble 272, **301**
  and minerals 81, 104–05, 114, 117, 148, 183, 229, 235
  quarrying 26
marcasite 8, 18, **63**
  and other minerals 48, 50, 56, 91, 137
maria (lunar "seas") 207, 273
marl 62, 314, **322**
marquise cut 25
Mars 273, 337
massive habit 21
Medusa 247
meerschaum 206
melanterite **137**, 331
mercury 56, 65–66, 112
  horn 112
mercury chloride 112
mercury sulfide 56
mesolite **188**
mesosiderite 336
meta-autunite 149
metamorphic rocks 26–27, **286–87**, 288–303
metaquartzite 310, 295
metatorbernite 152
meteorites 27, 55, **334**
  and minerals 53, 60, 89, 118, 179, 207, 208, 211, 245
  and native elements 36, 39
  and silicates 167
  iron **335**
  stony **337**
  stony-iron **336**
mica schist 81, 239
micaceous hematite 91
micas 11, 195–99, 201, 202
  and igneous rocks 258, 263, 269, 278
  and metamorphic rocks 289, 291, 293, 295, 301
  and sedimentary rocks 308, 315
microcline 167, **175**, 176, 262
microlite **86**
micromounts 31
microscope, for collecting 31
migmatite **290**
milky quartz 168
Miller, W.H. 60
millerite **60**
Millstone Grit 312
mimetite 151, 155, 160, **164**
minerals **8–17**, 34–253
  accessory 10
  arsenate 13, **147**, 160–65
  associations **10–11**
  borate 13, **126**, 127–30
  carbonate 8–9, 13, **113**, 114–125, 313
  chromate 13, **131**, 142
  classification **12–13**
  cleavage 16
  color 12
  composition 8–9
  essential 10
  fluorescence 17
  fracture 15
  groups 10–11
  halide 13, **106**, 107–12
  hardness 16
  hydroxide 13, **100**, 101–05
  identification **14–15**, **16–17**
  luster 15
  molybdate 13, **131**, 143
  native element 8, 13, **36**, 37–47
  organic 13, **246**, 247–53
  oxide 8, 9, 13, **74**, 75–99
  phosphate 13, **147**, 148–58
  primary 10
  refractive index 16
  secondary 10
  silicate 9, 13, **166–67**, 168–245
  streak 15
  sulfate 131–41
  sulfide 13, **48**, 49–64, 74
  sulfosalt **65**, 66–73
  tenacity 16
  tungstate 13, **131**, 144–46
  vanadate 13, **147**, 159
mines 24, 28–29
mixed cut 25
Mohs scale 17
molluscs 248–49
molybdates 13, **131**, 143
molybdenite **64**, 79, 143
molybdenum 64, 131, 143, 302
molybdenum sulfide 64
monazite 84, **150**
monoclinic crystal system 23
Mont St.-Hilaire 184
Moon,
  anorthosite 261
  basalt 273
  meteorites 334, 337
  minerals 39, 53, 179, 207, 211
  rocks 39, 53, 207
  silicates 167
moonstone 173, 176
morganite 225
mother of pearl 115, 248–49
Mount Fujiyama 275
Mount Saint Helens 257, 274
Mount Vesuvius 184, 229
mudstone 288, 291, 293–94, 315, **316**, 327
  calcareous 322
mullite 238
muscovite 148, **195**, 202, 258, 260, 292
muscovite mica 294
muscovite schist 291–92
muscovy glass 195
mylonite **289**

N

nacrite 201
nailhead spar 114
native elements 8, 13, **36**, 37–47
natrolite 185, **186**, 188
natural gas 12
nautilus 316
nemalite 105
neodymium 93, 150
neodymium phosphate 150
nesosilicates 166, 232–45
nepheline **182**, 185, 267
nepheline syenite 262
nephrite 213, **217**
nephrite jade 219–20
nickel,
  and igneous rocks 265–66, 270, 272
  and meteorites 334–37
  and minerals 39, 45, 55, 60, 85
  and sedimentary rocks 326
nickel iron sulfide 55
nickel sulfide 60

niobates 82
niobium 77, 82, 84–85, 89, 93, 272
niobium sodium calcium oxide 85
nitrates 13
nodule **331**, 333
note-taking, for collections 29

O

obsidian 174, 257, 278, **280**
octagonal cut 25
octahedrite 76
oil (petroleum) 12, 314, 324
oil shale **314**
oligoclase 167, **178**, 264, 275
oligoclase feldspar 90
olivenite 160–61
olivine **232**
  and igneous rocks 261, 266–69, 273, 285
  and metamorphic rocks 297–99, 301
  and meteorites 334–37
  and other minerals 148, 182, 191
  chemical formula 12
  gabbro 11
  occurrence 9
olivine gabbro 265
omphacite 299
oolitic ironstone 328
oolitic limestone 319
opal 14, **172**, 203, 325
opaline 330
optical spar 114
orbicular granite 259
organic minerals 13, **246**, 247–53
organizing collections 32–33
orpiment 15, 33, 48, 58, **59**, 61
orthochrysotile 192
orthoclase 11, 17, 167, **173**, 262, 270
orthogneiss 288
orthopyroxene 265
orthoquartzite 295, 308, **310**
orthorhombic crystal system 23
oval cut 25
oxides 8, 9, 13, **74**, 75–99
oxygen 74
oxyhydroxy-halides 106
oysters 248–49

P

palladium 38
pallasites 334, 336
panning 24, 28
parachrysotile 192
paragneiss 288
pararealgar 58
Paris, plaster of 131, 323
peacock ore 50
pearl 243, 246, **248**
pearly talc 193
peat 253
pebble-conglomerate 306
pegmatite 259, **260**
  and cyclosilicate minerals 223–25
  and feldspar minerals 173, 175, 177–78
  and feldspathoid minerals 184
  and inosilicate minerals 209
  and nesosilicate minerals 234–36, 238, 240, 244
  and other minerals 40, 64, 108, 123
  and oxide minerals 81, 82, 84–86, 90, 93, 95, 97
  and phosphate minerals 149, 150, 153, 157–58
  and phyllosilicate minerals 195, 198
  and sorosilicate minerals 231
  and tungstate minerals 144–46
  and zeolite minerals 187, 189
  dikes 26
Pele's hair **281**
Pele's tears **281**
pencil gneiss 288
pencil "lead" 46
pendeloque cut 25
penetration, twinning 19
Pentland, Joseph 55
pentlandite **55**
peridot 232
peridotite **266**
  and minerals 96, 118, 192, 210
  and other rocks 261, 267, 269
permineralization 305
Perovski, Count Lev Alekseevich 89
perovskite **89**, 182, 269
Perseus 247
petroleum 12, 314, 324
pharmacosiderite 165
phlogopite 269, 272
phonolite 174
phosphates 13, **147**, 148–58
phosphorus 147, 151, 309, 331
phyllite 214, 287, **294**
phyllonite 289
phyllosilicates 166, 191–206
Pigeon Point 207
pigeonite **207**
pigments 196
pinacoid face 20
pink diamond 47
pink topaz 234
pisoliths 101
pisolitic bauxite 101
pistachio epidote 230
pitchblende 83
placer deposits 24
plagioclase 265, 270, 273, 275, 297
  in gabbro 11
  sodium 262
plagioclase feldspar 89, 261, 264, 267–68
plaster of Paris 131, 323
plate tectonics 318
platinum **38**, 265, 272
pleochroism 153
pleonaste 96
plug, volcanic 256
plush copper ore 87
plutonic intrusive rocks 26, 256
Plymouth Rock 263
polishing, gems 25
polyps 247
porphyritic granite 259
porphyroblasts 292, 294
porphyry 181, **271**
porphyry copper 57
potash 106–07, 111
potash mica 195
potassium 107, 111, 309
potassium bentonite 200
potassium magnesium chloride 107
powder, talcum 193
precipitation 305
Prehn, Hendrik von 205
prehnite 141, 190, **205**
primary minerals 10
prism face 20
Prontium copal 250
protective equipment, for collecting 30
Proust, Joseph 72
proustite 49, 70, **72**
Prussian blue 157
psilomelane 94
pteridosperm 333
puddingstone 306
pumice 257, 276, **277**, 278
purple copper ore 50
purple zircon 233
pyrargyrite 49, 65, **70**, 72
pyramid face 20
pyrite 12, **62**
  and other minerals 102, 137, 153, 252
  and rocks 262, 293, 301, 331
  and silicate minerals 183, 221, 241
  and sulfide minerals 50, 53–57, 61, 63
  and sulfosalt minerals 67, 71, 73
  chemical formula 12
pyritohedral crystals 62
pyrochlore **85**, 86, 272
"pyroclastic" rocks 257
pyrolusite 38, **80**, 98, 103
pyromorphite 54, 124, 143, **151**, 164
pyrope **242**
pyrophyllite 21, **194**
pyroxene,
  and extrusive igneous rocks 273, 275, 278–79
  and intrusive igneous rocks 261, 262, 265–66
  and metamorphic rocks 296–99
  and meteorites 335–37
  and minerals 191
  and sedimentary rocks 317
  and subvolcanic igneous rocks 268–70
  in pyroxenite 267
pyroxene hornfels 267
pyroxenite 266, **267**
pyrrhotite 55

Q

quarries 28–29
quartz 9, 17, **168**, 258–59
  and carbonate minerals 117, 121
  and igneous rocks 260, 262–65, 268, 274, 278–80
  and metamorphic rocks 288, 290–91, 293–95, 297, 301
  and native elements 37, 42–43
  and other minerals 101, 141, 144, 161
  and oxide minerals 78, 90
  and sedimentary rocks 306, 310, 312–13, 316–17
  and silicate minerals 166–67, 169, 170–71, 186, 194, 203, 223, 226, 231, 236, 241
  and sulfide minerals 50, 51, 53, 57–58, 61
  and sulfosalt minerals 66, 73
  cementing agent 304
  cryptocrystalline 331

quartz *continued*
  in granite 11
  luster 15
  milky 168
  rose 168, 260
  smoky 168, 175, 178, 260
quartz porphyry 271
quartzite **295**
quicksilver, horn 112

R

radicals, chemical 9, 12, 14
radiometric age-dating 309, 334
radium 159
realgar 48, **58**, 59, 61
record-keeping, collecting 29
Red Cloud wulfenite 143
red oxide of zinc 88
redruthite 51
refractive index 17
refractory materials 208
regional metamorphism 286
resin 250
rhodium 38
rhodochrosite 68, 80, 113, **121**, 215
rhodonite **215**
rhomb porphyry 271
rhombohedral crystal form 20
rhyolite **278**
  and minerals 174, 178, 222, 225, 234
  and rocks 274, 279–80, 283
  as an extrusive rock 257
rhyolite porphyry 271
ribboned jasper 171
Riebeck, Emil 222
riebeckite **222**
rind, jadeite 213
road cuttings, collecting from 28–29
rock cycle 27
rock flour 318
rock gypsum **323**
rock salt 110, 323, **324**
rocks **26–27**
  igneous **256–57**, 258–85
  metamorphic **286–87**, 288–303
  sedimentary **304–05**, 306–33
romanèchite **94**, 98
rose quartz 168, 260
Rosetta Stone 263
rouge 91
round cut gems 25
rubber 201
rubbly scoria 276
rubellite 224
rubidium 198, 270
ruby 12–13, 95
ruby blende sphalerite 53
ruby copper 87
ruby-in-zoisite 231
ruby silver 70, 72
ruby spinel 96
rutile **78**
  and other minerals 76–77, 90, 243
  and rocks 267, 269, 310

S

safety, when collecting 29–31
St.-Hilaire, Mont 184
salt,
  rock 110, 323, **324**
  table 106, 110, 324
salt domes 44, 67, 133
Samarski-Bykhovets, Vasili Yefrafovich von 84
samarskite **84**
sandstone **308**
  and minerals 83, 159, 196
  and other rocks 295, 309–10, 313, 315
  dirty 317
  layering 27
sanidine 167, **174**
sapphire 20, 74, 95, 231
  water 223
satin spar 136
scandium 82
Scheele, C.W. 146
scheelite 64, **146**
schiller effect 180
schist 286–87, **291**
  and cyclosilicate minerals 223, 225
  and feldspar minerals 175, 178
  and inosilicate minerals 209, 214, 216, 218, 221
  and nesosilicate minerals 235, 237–38, 240, 243, 245
  and other minerals 76, 78, 96–97, 104–05
  and other rocks 288, 290
  and phyllosilicate minerals 195, 197
  and sorosilicate minerals 231
  garnet **292**
schorl 224
scissor cut 25
scolesite 188
scoria **276**, 283
scorodite 160, **165**
scrapers, for cleaning 32–33
sea urchin fossils 305
secondary minerals 10
sectile minerals 16
sedimentary rocks 26–27, **304–05**, 306–33
selenite 23
selenite gypsum 23, 136
selenium 57
sepiolite **206**
sericite 195
serpentine 99, **191**, 193, 217, 269, 298
serpentinite **298**
  and minerals 60, 99, 117, 206, 245
  and other rocks 266, 300
serpentinization 298
shale 11, **313**
  and metamorphic rocks 288, 291, 293–94
  and minerals 62–63, 123, 202
  and sedimentary rocks 308, 314–16, 322–24, 329
  metamorphism of 287
  oil **314**
shell 246, **249**
shores, collecting on 28
siderite **123**
  and other minerals 102, 103, 108, 113, 122
  and rocks 272, 328, 331
sieve, for collecting 24, 31
silica 325, 331–33
silica gel 190
silicates 9, 13, **166–67**, 168–245
siliceous tufa 330
silicon 9, 166
sill 256
Silliman, Benjamin 237
sillimanite 236, **237**, 295
siltstone 288, 291, 293–94, **315**
silver 8, 13, 36, **43**, 272
  and other native elements 41, 45
  and sulfide minerals 48–49, 54–55, 60
  and sulfosalt minerals 66, 69–70, 72
  coinage 65
silver antimony sulfide 69
silver sulfide 49
sinter 330
skarn 148, **302**
sketching, collecting locations 29
slate 214, **293**, 303, 317
  as metamorphic rock 286–87
smithsonite 13, 113, **124**, 151, 160
smoky quartz 168, 175, 178, 260
snow 75
snowflake obsidian 174
soapstone 193, **300**
sodalite 183, **184**, 185
sodium 89
sodium aluminum silicate 190
sodium aluminum silicate chloride 184
sodium aluminosilicate 186
sodium-anthophyllite 216
sodium bentonite 200
sodium borate 127
sodium calcium aluminosilicate 181, 183
sodium calcium sulfate 141
sodium chloride 110
sodium iron silicate 209, 222
Solar System 334
solid-solution series 10, 166
sorosilicates 166, 227–31
spatulas, for collecting 31, 33
specimens, cleaning and display of 32–33
spectrolite 180
specular hematite 91
spessartine **244**
sphalerite 11, **53**
  and other minerals 121, 153, 227
  and sulfur minerals 48, 51, 54, 66–67, 70–71, 134
sphene 235
spider conch 249
spindle bomb 285
spinel 74, **96**, 97, 104, 182, 267
spodumene **212**
square cut 25
stalactites 114
stalactitic crystals 21
staurolite 238, **239**, 288, 294
staurolite schist 292
steatite 193, 300
steel 64
Stephan, Archduke Victor 69
stephanite **69**, 72
stibnite 56, 58–59, **61**, 65
stilbite 11, 204
Stonehenge 268
stony-iron meteorite 334, **336**
stony meteorites 334, **337**
storage, for collections **32–33**
strata, rock 26
streak 15
stream tin 79
strontianite 113
strontium 113, 135

sub-bituminous coal 253
subarkose 310
subvolcanic igneous rocks 268–72
sulfates 13, **131**, 132–41
sulfides 13, **48**, 49–64, 74
sulfosalts **65**, 66–73
sulfur 8, 13, 36, **44**, 57
  crystal structure 36
  in sulfates 131
  in sulfides 48
  luster 15
sunstone 180
sunstone oligoclase 178
syenite 95, **262**
  and silicate minerals 175–76, 181, 186, 197, 209, 222
sylvine 111
sylvite 13, 106, **111**
symmetry, crystal 19
syphilis 112
systems, crystal 19, 22–23

T
table cut 25
table salt 106, 110, 324
tabular crystals 21
taenite 39, 335
Taj Mahal 301
talc 17, **193**, 217, 220, 291–92, 300
talc schist 89, 117–18
talcum powder 193
tantalates 82
tantalite **82**
tantalum 74, 82, 84, 86, 272
tanzanite 231
tears, Pele's **281**
tectonic plates 286
tectosilicates 166–72
tektites 334
television stone 128
tenacity 16
Tennant, Smithson 66
tennantite **66**, 73
tetragonal crystal system 22
tetrahedrite 65, 66, 69, 70–72, **73**, 144
tetrataenite 39
Theophrastus 203
thorium 85, 93, 272
thorny oyster 249
thulite 231
tiki 217
till, glacial 318
tillite **318**
tin 40–41, 79, 302
  cassiterite ore 74
tin oxide 79
tincalconite 127
tinstone 79
tinzenite 228
titanite 76, **235**, 262
titanium 74, 89–90, 216, 265, 331
titanium dioxide 76–77
titanium oxide 78
topaz 17, 64, 79, 198, **234**, 240, 260
  cleavage 16
  habit 20
  sherry 23
torbernite **152**
tortoise shell 249
tourmaline 9, 79, 153, 177, **224**, 241, 260, 294
trace elements 14
trachyte 174, 257, **279**
transportation of specimens 32
tremolite **219**
  and other minerals 193, 213–14, 217, 220
  and rocks 296, 300, 302
triclinic crystal system 23
trigonal crystal system 22
trilobite 313
triplite **153**
troilite 335
trona 113
trowels, for collecting 31
tufa **330**
tuff 257, 275, **282**
tungstates 13, **131**, 144–146
tungsten 74, 79, 82, 93, 143–46
  in tungstates 131
turbidite 317
turquoise 129, **154**, 156, 203
turtle 249
twinning, crystals 19

U
Ulex, George Ludwig 128
ulexite 126–27, **128**, 130
  luster 15
ultramafic rocks 257
ultramarine 183
ultramylonite 289
ultrasonic cleaning 32
ultraviolet light 17
unit cells, crystal 18
uraninite **83**, 149, 152
uranium 74, 83–85, 93, 147, 149, 152, 159
uvarovite 302

V
Valentine, Basil 40
vanadates 13, **147**, 159
vanadinite 143, 151, **155**
vanadium 147, 155, 159, 225
variegated copper ore 50
variquoise 156
variscite **156**
varlamoffite 79
vent agglomerate 284
Venus 273
verdelite 224
vermiculite **199**
vermillion 56
vesuvianite 22, **229**
Vesuvius, Mount 184, 229
Villiers, A.J.M. Brochant de 139
violane 210
violet jadeite 213
Vivian, John Henry 157
vivianite **157**, 165
volcanic bombs 282, **285**
volcanic breccia **283**, 284
volcanic glass 277–78, 280–81
volcanic igneous rocks 257, 280–85
von Born, Ignaz 50
von Carnall, Rudolf 107
von Goethe, Johann Wolfgang 102
von Prehn, Hendrik 205
von Samarski-Bykhovets, Vasili Yefrafovich 84

W
water 75
water sapphire 223
watermelon tourmaline 224
Wavell, William 158
wavellite 156, **158**
Wheal Gorland mine 161
white asbestos 192
white mica 195
Widmanstätten pattern 335
Wild 2, comet 179
williamsite 191
Withering, William 116
witherite 113, **116**
wolframite 145
wollastonite **214**, 229
wood tin 79
Wülfen, F.X. 143
wulfenite **143**, 155

X
xenolith 299
xenotime 147

Y
yellow orthoclase 173
ytterbium 84
yttrium 82, 84, 93, 147
yttrotantalite 82

Z
zeolites 11, 185–90
zinc,
  and other minerals 43, 48, 66, 113, 117, 124, 134
  and oxide minerals 74, 88, 98
  and sulfide minerals 53–54
  in rocks 302
zinc aluminum oxide 97
zinc arsenate hydroxide 160
zinc carbonate 124
zinc sulfide 48
zincite **88**
zircon 85, 90, **233**, 262, 310
zirconium 85, 93
zoisite **231**
zultanite 104

ACKNOWLEDGMENTS

Dorling Kindersley would like to thank the following people from Smithsonian Enterprises at the Smithsonian Institution in Washington D.C:

Carol LeBlanc, Vice President
Brigid Ferraro, Director of Licensing
Ellen Nanney, Licensing Manager
Kealy Wilson, Product Development Coordinator

The publisher would like to thank the following people: Frances Green for helping to plan the profile sections; Janet Mohun, Miezan van Zyl, Lizzie Munsey, and Martyn Page for editorial work; Steve Setford for proofreading; Jane Parker for the index; David Roberts for database work; and Sophia Tampakopoulos, Jacket Design Development Manager.

DK India would like to thank Neha Chaudhary, Jubbi Francis, and Suefa Lee for editorial assistance; Vaibhav Rastogi for design assistance; and Mahima Barrow, Nidhilekha Mathur, Swati Mittal, and Neha Samuel for picture selection assistance.

The publisher would like to thank the following for their kind permission to reproduce their photographs:

(Key: a-above; b-below/bottom; c-center; f-far; l-left; r-right; t-top)

6-7 Corbis: William James Warren/ Science Faction. **8 Corbis:** Roger Wood (tr). **8-9 Corbis:** Ludovic Maisant (b). **9 Corbis:** Kevin Schafer (t). **14 Science Photo Library:** Steve Gschmeissner (b). **24 The Bridgeman Art Library:** Ashmolean Museum, University of Oxford, UK (cr); Egyptian National Museum, Cairo, Egypt/Giraudon (cl). **Corbis:** Jack Fields (bl); Bernard Bisson/Sygma (br). **26 Corbis:** WildCountry (br). **Dorling Kindersley:** Neil Fletcher (cl). **27 Corbis:** Momatiuk – Eastcott (t). **28 Dorling Kindersley:** Neil Fletcher (b). **Getty Images:** Keith Douglas (t). 29 Getty Images: John Elk III (t). **32 Corbis:** Jim Sugar (cr). **Getty Images:** American Images Inc (cl). **33 Corbis:** Philippe Eranian (t). **34-35 Getty Images:** National Geographic. **36 Corbis:** Ricki Rosen/SABA (cl). **FLPA:** Silvestris Fotoservice (b). **39 Dorling Kindersley:** Peter Anderson © Dorling Kindersley, Courtesy of the Danish National Museum (br). **40 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (br). **41 Dorling Kindersley:** Clive Streeter, Courtesy of The Science Museum, London (cl). **42 Dorling Kindersley:** Harry Taylor, Courtesy of the Natural History Museum, London (cl). 43 Dorling Kindersley: Judith Miller/Woolley and Wallis (br). **44 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (cl). **45 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (br). **47 Smithsonian Institution, Washington, DC:** (br). **48 Corbis:** Eric and David Hosking (t). **Science Photo Library:** Ben Johnson (b). **60 Smithsonian Institution, Washington, DC:** (cl). 61 Smithsonian Institution, Washington, DC, USA: (bl). **65 Dorling Kindersley:** Alan Hills and Barbara Winter © The British Museum (cr). **Tony Waltham Geophotos:** (b). 69 Corbis: Heritage Images (br). **72 Dorling Kindersley:** Neil Fletcher (bl). **73 akg-images:** Rabatti – Domingie (br). **74 Corbis:** Steve Parish/Steve Parish Publishing (t). **Dorling Kindersley:** Courtesy of Dream Cars (br). **75 Corbis:** Klaus Lang/ All Canada Photos (t); WIN-Images (cl); Jim Reed (bl). **77 Dorling Kindersley:** Neil Fletcher (bl). **81 Smithsonian Institution, Washington, DC:** (br). **83 Corbis:** Bettmann (br). **88 Dorling Kindersley:** Neil Fletcher (bl). **100 Corbis:** Vince Streano (b). **105 Corbis:** Michael S. Yamashita (br). **106 Corbis:** Stevens Fremont/ Sygma (cr). **Tony Waltham Geophotos:** (b). **109 Dorling Kindersley:** Courtesy of the Oxford University Museum of Natural History (br). **113 Dorling Kindersley:** Judith Miller/333 Auctions LLC (cr). **Tony Waltham Geophotos:** (b). **114 Corbis:** (br). **118 Getty Images:** (bl). **119 Dorling Kindersley:** Neil Fletcher (clb). **121 Smithsonian Institution, Washington, DC:** (t). **126 Tony Waltham Geophotos:** (b). **131 Corbis:** Gallo Images (b); Larry Mulvehill (cr). **134 Smithsonian Institution, Washington, DC:** (clb). **140 Dorling Kindersley:** Neil Fletcher (clb). **141 Dorling Kindersley:** Neil Fletcher (bl). **145 NASA:** Marshall Space Flight Center (br). **147 Corbis:** David Muench (b). **Science Photo Library:** Hank Morgan/University of Massachusetts at Amherst (cr). **151 Smithsonian Institution, Washington, DC:** (clb, bl). 159 Dorling Kindersley: Judith Miller/ Mark Laino (br). 162 Dorling Kindersley: British Museum Images (br). **167 Alamy Images:** Pavel Filatov (b). **Dorling Kindersley:** Michel Zabe © CONACULTA-INAH-MEX (cr). **168 Dorling Kindersley:** Judith Miller/Private Collection (br). **169 Dorling Kindersley:** British Museum Images (br). **170 Dorling Kindersley:** Judith Miller/Wallis and Wallis (br). **173 Dorling Kindersley:** Judith Miller/Sylvie Spectrum (br). **175 Dorling Kindersley:** Judith Miller/Dreweatt Neate (br). **178 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (br). **179 Alamy Images:** Piotr & Irena Kolasa (b). **181 Science Photo Library:** Joel Arem (b). **182 Dorling Kindersley:** British Museum Images – DK Images (br). **184 Dorling Kindersley:** Judith Miller/N. Bloom & Son Ltd. (br). **185 Alamy Images:** RF Company (cl). **Science Photo Library:** Joel Arem (bl). **190 Dorling Kindersley:** Neil Fletcher (cl); Clive Streeter, Courtesy of The Science Museum, London (br). **191 Dorling Kindersley:** Neil Fletcher (bl). **193 Dorling Kindersley:** Judith Miller/ Cooper Owen (br). **194 Corbis:** Scientifica/Visuals Unlimited (bl). **196 U.S. Geological Survey:** G.R. Mansfield (bl); E. D. McKee (cl). **198 Dorling Kindersley:** Neil Fletcher (cl). **200 Dorling Kindersley:** Judith Miller/David Rago Auctions (br). **Dreamstime.com:** Gelyngfjell (bl, cl). **206 Dorling Kindersley:** Judith Miller/Dreweatt Neate (br). **209 Alamy Images:** Andre Joubert (t); RF Company (bl). **214 Getty Images:** DEA/C. Bevilacqua (cl). **Science Photo Library:** Scientifica, Visuals Unlimited (bl). **215 Smithsonian Institution, Washington, DC:** (clb). **217 Dorling Kindersley:** Judith Miller/

Blanchet et Associes (br). **Getty Images:** DEA/Photo 1 (cl). **218 Getty Images:** Mark Schneider (clb); Scientifica (cla). **222 Dorling Kindersley:** John Chase, Courtesy of The Museum of London (br). **223 Dorling Kindersley:** Judith Miller/ Lesley Craze Gallery (br). **226 Alamy Images:** Alan Curtis (bl); Greg C Grace (cl). **230 Getty Images:** Mark Schneider (cl); Visuals Unlimited/Scientifica (bl). **231 Dorling Kindersley:** Judith Miller/Woolley and Wallis (br). **232 Dorling Kindersley:** Judith Miller/ HY Duke and Son (br). **241 shutterstone.com:** (cl, bl). **243 Dorling Kindersley:** Judith Miller/ Joseph H Bonnar (br). **246 Dorling Kindersley:** Martin Strmiska (c) Alamy (b). **248 Dorling Kindersley:** Judith Miller/ Beaussant Lefevre (br). **249 Dorling Kindersley:** Judith Miller/ Wallis and Wallis (br). **252 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (cl). **254-255 Corbis:** Chris Harris/All Canada Photos. **257 Corbis:** Gary Braasch (b). **259 Dorling Kindersley:** Neil Fletcher (bl). **261 Dorling Kindersley:** Judith Miller/N. Bloom & Son Ltd. (br). **Smithsonian Institution, Washington, DC:** (bl). **264 Corbis:** Visuals Unlimited (bl). **Getty Images:** Mark Schneider/Visuals Unlimited (cl). **269 Dorling Kindersley:** Judith Miller/HY Duke and Son (br). **270 Dorling Kindersley:** Neil Fletcher (cl). **Getty Images:** DEA/A. Dagli Orti (bl). **272 Alamy Images:** RF Company (t). **274 Alamy Images:** Susan E. Degginger (bl). **Getty Images:** DEA/R. Appiani (cl). **275 Dorling Kindersley:** Image Plan (c) Corbis (br). **276 Alamy Images:** Phil Degginger (t); Susan E. Degginger (b). **281 Alamy Images:** Photo resource Hawaii (clb). **Corbis:** (bl). **283 Dorling Kindersley:** Neil Fletcher (cl). **shutterstone.com:** (t). **286-287 Alamy Images:** Daniel Dempster Photography (b). **291 Dorling Kindersley:** Judith Miller/ Ormonde Gallery (br). **293 Dorling Kindersley:** Courtesy of the Blists Hill Museum, Ironbridge, Shropshire (br). **294 Alamy Images:** Susan E. Degginger (clb); RF Company (cla); geoz (bl). **296 Alamy Images:** geoz (cl); RF Company (bl); Fabrizio Troiani (clb). **297 Alamy Images:** The Art Archive (b). **300 Alamy Images:** Leslie Garland Picture Library (cl). **Corbis:** Scientifica/Visuals Unlimited (t, bl). **Dorling Kindersley:** Neil Fletcher (clb). **305 Corbis:** Ric Ergenbright (cr). **307 Dorling Kindersley:** Peter Hayman/The Trustees of the British Museum (br). **309 fotoLibra :** Brian G Burgess (b). **316 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (cl). **317 The Bridgeman Art Library:** Detroit Institute of Arts, USA/Gift of Lillian Henkel Haass and Constance Haass (br). **319 Dorling Kindersley:** The Trustees of the British Museum (br). **320 shutterstone.com:** (bl). **323 Dorling Kindersley:** The Trustees of the British Museum (br). **326 Dorling Kindersley:** Neil Fletcher (bl, t). **327 fotoLibra :** John Cleare (t). **332 Dorling Kindersley:** Courtesy of the Pitt Rivers Museum, University of Oxford (br). **333 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (cl). **334 Corbis:** (cr). **Smithsonian Institution, Washington, DC:** (c). **335 Dorling Kindersley:** Neil Fletcher (b). **336 Dorling Kindersley:** Colin Keates, Courtesy of the Natural History Museum, London (cl); Courtesy of the Oxford University Museum of Natural History (t). **337 Dorling Kindersley:** Neil Fletcher (bl)

All other images © Dorling Kindersley
For further information see:
www.dkimages.com

THE COMPLETE ENCYCLOPEDIA OF MINERALS

Description of over 600 Minerals
from around the world

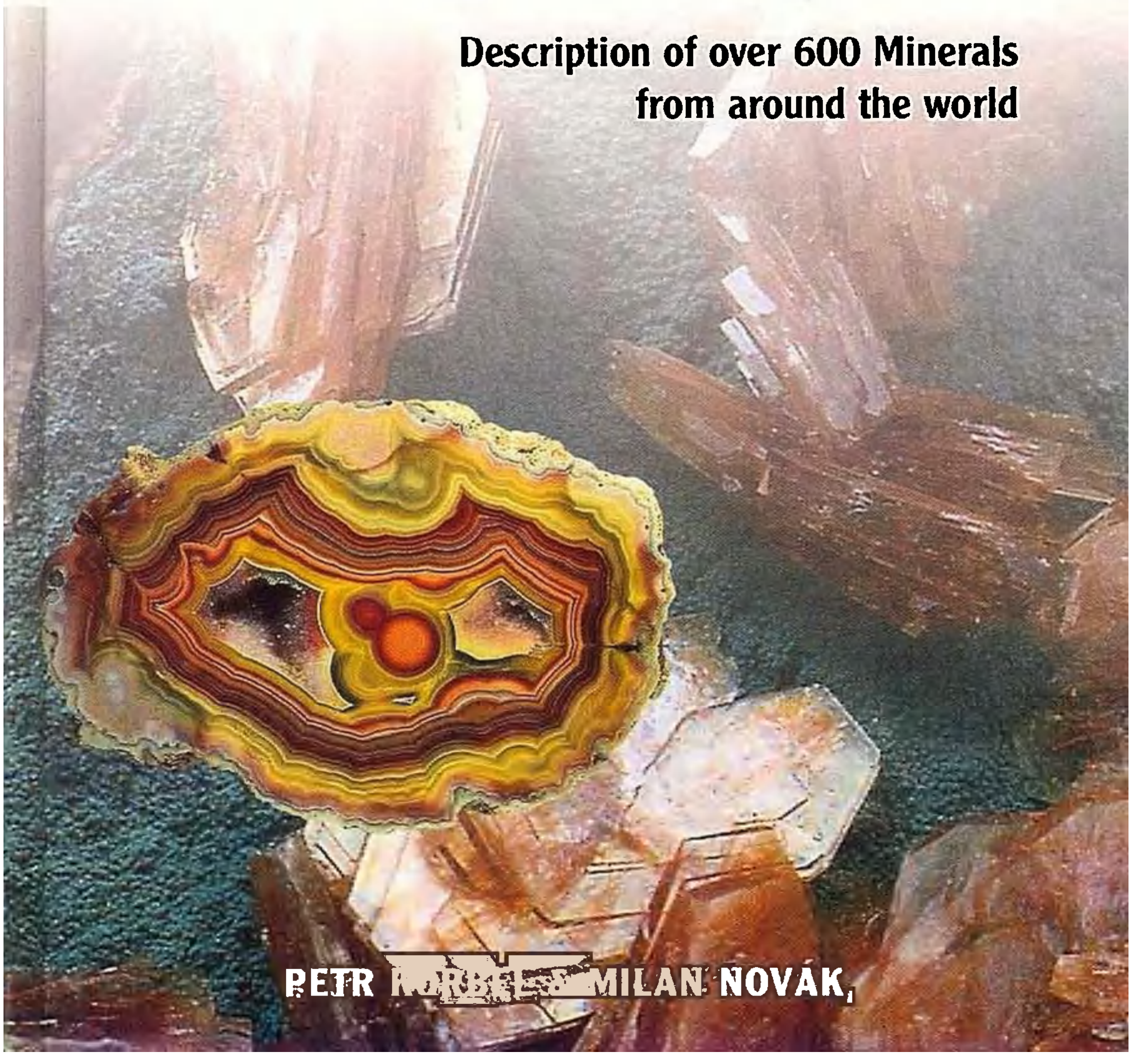

PETR KORBEL & MILAN NOVÁK,

THE COMPLETE ENCYCLOPEDIA OF

MINERALS

THE COMPLETE ENCYCLOPEDIA OF
MINERALS

THE COMPLETE ENCYCLOPEDIA OF

MINERALS

THE COMPLETE ENCYCLOPEDIA OF

MINERALS

Description of over 600 Minerals from around the world

PETR KORBEL & MILAN NOVÁK

Grange
BOOKS

Explanation of the abbreviations used in the book

C – color
S – streak
L – luster
D – diaphaneity
DE – density
H – hardness
CL – cleavage
F – fracture
M – morphology
LU – luminescence
R – remark
X – crystal
XX – crystals

In the figure captions, the bigger dimension is always mentioned

Scale of the frequency of the occurrence

● - very rare
●● - rare
●●● - uncommon
●●●● - abundant
●●●●● - common

© 1999 Rebo International b.v., Lisse, the Netherlands
© 2001 Grange Books PLC, United Kingdom

Text: Petr Korbel, Milan Novak
Fotografie: Fotobanka Granit (100), Petr Korbel (220),
Geza Kulcsar (16), Jeffry A. Scovil (331), Sandor Szakall (34)

Production and layout: Granit Ltd., Praha

ISBN 1 84013 404 6

All rights reserved.

Copying of the entire book as well as of its parts in any way is forbidden by law without prior written consent of the publisher.

Contents

Introduction 7

1. Elements 9
2. Sulphides 19
3. Halides 67
4. Oxides 75
5. Carbonates 113
6. Borates 131
7. Sulfates 137
8. Phosphates 155
9. Silicates 193
10. Organic compounds 281
11. Rocks, meteorites and tektites 285

Recommended literature 291

Index 293

Introduction

The ever increasing number of publications about minerals reflects a growing interest in nature. Most of those publications though only deal with a few dozen of the most common minerals or gemstones. This book fills the gap by also featuring less common and rare minerals. The authors describe over 600 mineral species and varieties, illustrated with about 750 color photographs. In choosing illustrations of particular minerals, aesthetic criteria such as size of crystal and color played a role in addition to their importance and distribution in nature. This book includes some rare minerals, known only from one locality, because they form very attractive crystals or aggregates. There are minerals known to humankind since prehistoric times such as quartz and gold, but also minerals first described quite recently like rossmanite. The photographs show well-formed and colorful crystals but many aggregates, which are more common in nature, are also included. The minerals in the book are listed according to the mineralogical system of Hugo Strunz, in his book *Mineralogische Tabellen* in 1978. The chemical formulae of individual minerals follow the form of Glossary of *Mineral Species* 1995 by M. Fleischer and J.A. Mandarino. The information is complemented in both cases with the latest knowledge from scientific literature, such as new nomenclature of amphiboles, micas and zeolites.

The mineral descriptions cover the basic physical and chemical data, including chemical formula and crystal system. The data provided correspond mainly to the end-members. The less common valence of the chemical elements is marked in the chemical formula (Fe^{3+}, Mn^{3+}, As^{3+}, Mn^{4+}, Pb^{4+}). Where an element features in both valences in the mineral, they are both marked (e.g. ilvaite, braunite).

The origin of individual minerals is described in detail. We chose for a relatively simplified scheme because the normal complexity cannot be described here in detail. Minerals can be distinguished as either primary (resulting directly from a solidifying of magma, crystallizing of an aqueous solution or metamorphism – re-crystallization in a solid state) and secondary (resulting from alteration of the original mineral, e.g. during its oxidation or reduction under low temperature and pressure close to the surface of the Earth). Primary minerals are divided into following groups: 1. magmatic, when a mineral crystallizes directly from a melt (it includes magmatic and effusive igneous rocks, including granitic and alkaline syenite pegmatites and meteorites); 2. sedimentary, when a mineral crystallizes during a process of diagenesis or from hydrous solutions under normal temperature (clastic, organic and chemical sedimentary rocks); 3. metamorphic, when a mineral crystallizes during metamorphic processes in a solid state at a wide range of temperature and pressure (it includes regionally and contact metamorphosed rocks and skarns); 4. hydrothermal, when minerals crystallize from aqueous solutions and fluids under high to low temperatures (it includes ore and the Alpine-type veins, cavities in volcanic rocks, minerals and rocks, hydrothermally altered under high temperature, e.g. greisens).

Secondary minerals are divided into following groups: 1. oxidation, when minerals result from the oxidation (weathering) of the primary minerals in the oxidation zone of ore deposits and other rocks (it includes the origin of malachite and azurite during the chalcopyrite oxidation, also the origin of secondary phosphates in granitic pegmatites during the oxidation of primary phosphates); 2. cementation, when minerals result from the reduction of the primary minerals (the origin of native copper and native silver under the reduction conditions in the cementation zone of ore deposits). This classification is very much simplified of course, because in many cases we cannot readily determine a specific origin of a particular mineral. This relates to minerals that crystallize under conditions which approximately represent a transition between separate phases of the origin, such as the magmatic or hydrothermal origins of elbaite in the pegmatite cavities; the metamorphic or hydrothermal origins of grossular in skarns; the magmatic or metamorphic origins of cordierite in migmatites; and the hydrothermal or secondary origins of some phosphates in granitic pegmatites etc.

With the localities for individual minerals we have tried to list the most important worldwide localities regardless of their recent production but we have also included recent discoveries since these may produce important mineral specimens. Where a mineral has an important use this is listed at the end of mineral description. We would like to acknowledge all who contributed in any way to the production of this book, particularly those private collectors and institutions which loaned minerals for photography. We hope those fascinated in the world of minerals and of nature will find this book a fascinating source of information.

This book is dedicated to the memory of Dr. Jaroslav “venek, who was of extraordinary influence to several generations of Czech and Slovak mineralogists and mineral collectors with his enthusiasm for mineralogy and attitude to life.

Marcasite, 100 mm, Misburg, Germany

1. Elements

Copper

Cu

CUBIC ● ● ● ● ●

Properties: C – light pink to copper-red, it darkens and covers green to black in air; S – red; L – metallic; D – opaque; DE – 8.9; H – 2.5 – 3; CL – none; F – hackly; M – cubic crystals and its combinations, dendritic aggregates, sheets, slabs, massive.

Origin and occurrence: Primary hydrothermal copper is mainly related to basic igneous rocks; it is also common as a product of supergene cementation. It is associated with cuprite, malachite, azurite, silver, chalkocite, bornite and other minerals. The largest accumulations of primary copper are in the Keweenaw Peninsular, Lake Superior, USA, the largest being 15 x 7 x 3m (approx. 50 x 23 x 10ft)and weighing 420 tons. Fine crystals up to 50 mm(approx. 2in) also occur there, as do calcite crystalwith copper inclusions. Superb supergene coppercrystals come from many localities like Tsumeb, Namibia and Chessy, France. Crystals up to 140mm (5½ in) long occurred in the Ray mine and in Bisbee, Arizona, USA. Very fine spinel-law twins up to 5 cm in size and dendritic aggregates come from Mednorudnyansk, Ural mountains, Russia; crystals up to 30 mm (13/16 in) were found in Dzhezkazgan, Kazakhstan. Fine specimens of copper, associated with cuprite, azurite and malachite occurred in Rudabanya, Hungary. *Application:* electronics, electrical engineering, ingredient in gold alloys.

Copper, 42 mm, Cornwall, UK

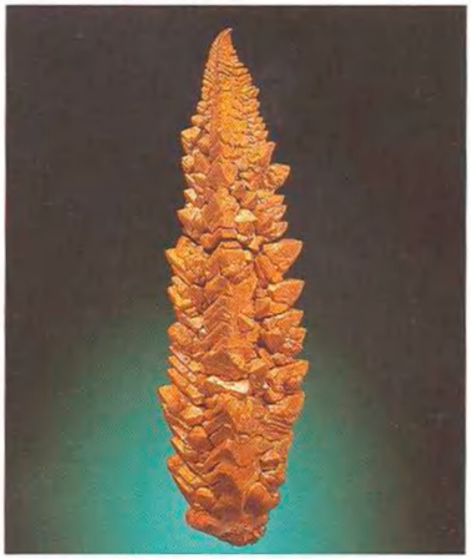

Silver, 52 mm, Freiberg, Germany

Copper, 120 mm, Keweenaw Peninsula, U.S.A.

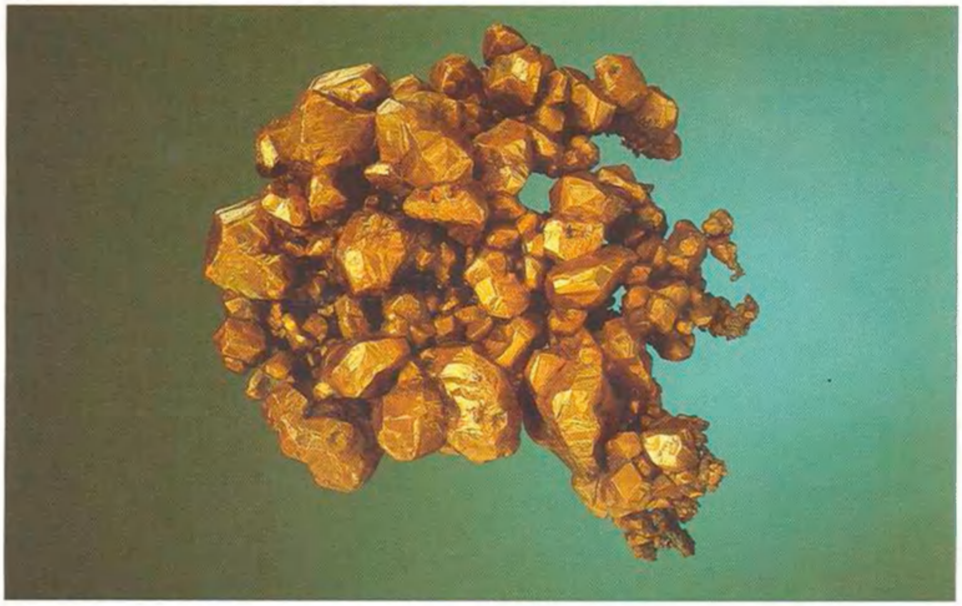

Silver

Ag

CUBIC ● ● ●

Properties: C – silver-white, tarnishes gray to black; S – silver-white; L – metallic; D – opaque; DE – 10.5; H – 2.5 – 3; CL – none; F – hackly; M – cubic crystals, dendritic aggregates, wires, leaves, massive.

Origin and occurrence: Hydrothermal in ore veins and also of secondary cementation origin in association with acanthite, stephanite, proustite, pyrargyrite, copper and many other minerals. The best specimens of crystallized and wire silver come from Kongsberg, Norway, where wires up to 400mm (16 in) long and crystals up to 40 mm ($1\frac{9}{16}$ in) in size have been found. Beautiful specimens of wire silver with wires over 100 mm (4 in) long are known from Freiberg, Schneeberg and St. Andreasberg, Germany. Wires several cm long were also found in Pribram and Jachymov, Czech Republic. Dendritic aggregates from Batopilas, Chihuahua, Mexico, reached up to 150 mm (6 in). Crystals and aggregates of silver, grown together with copper are genetically unique in the basalt cavities in the Keweenaw Peninsula near Lake Superior, Michigan, USA. Wires, up to 100 mm (4 in) long, come also from the San Genaro Mine in Huancavelica and Uchucchaqua, Peru. New finds of silver wires, up to 150 mm (6 in) long, have been made in Dzhezkazgan, Kazakhstan.

Application: photographic industry, jewelry, electronics.

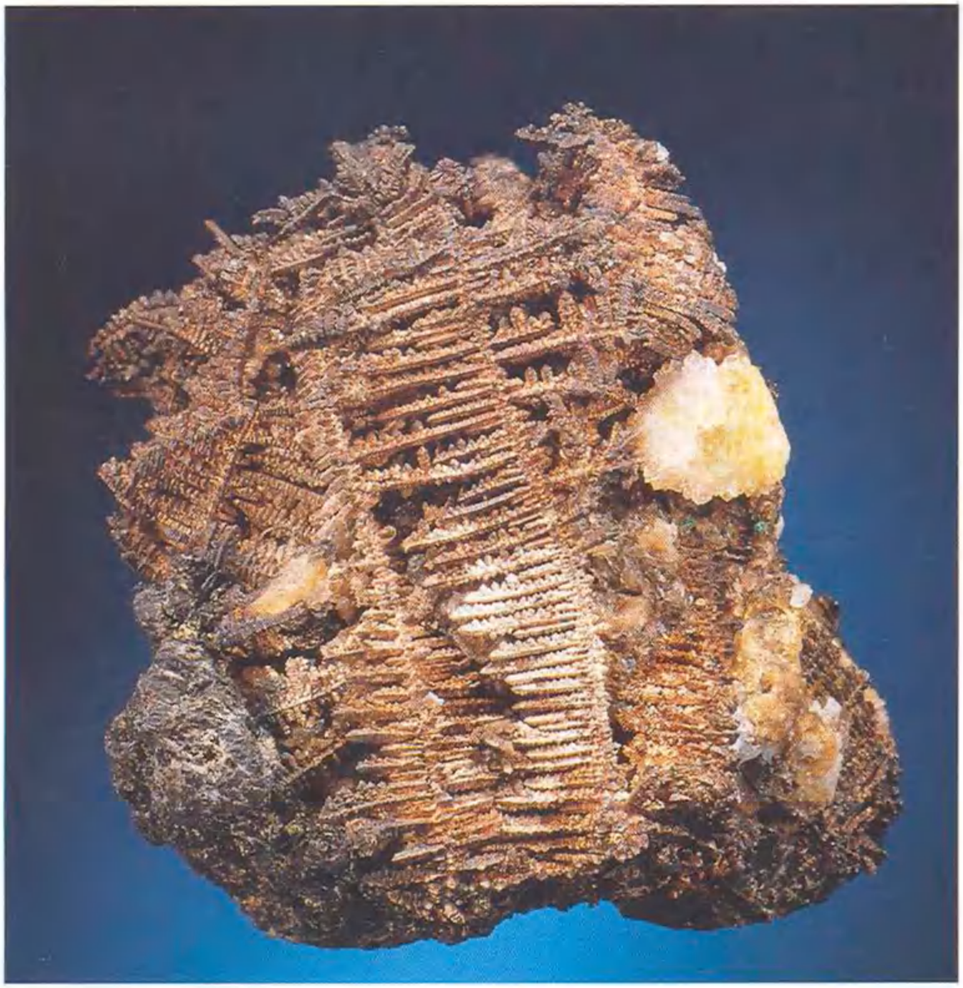

Silver, 55 mm, Schwarzwald, Germany

Gold, 48 mm, El Dorado, California, U.S.A.

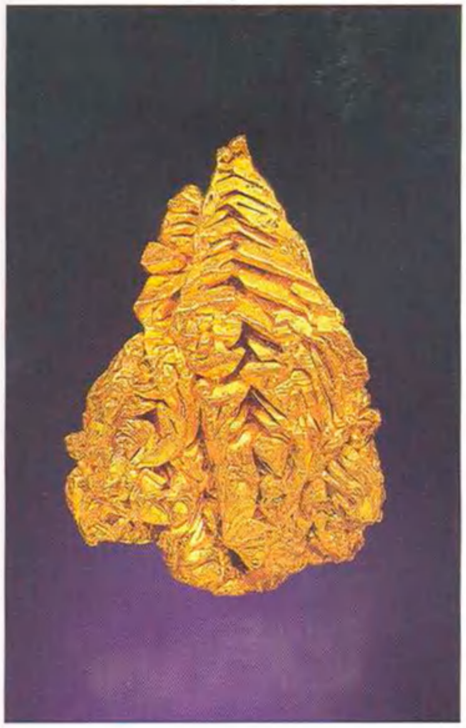

Gold, 68 mm, Eagle's Nest Mine, California, U.S.A.

Gold

Au

CUBIC ● ●

Properties: C – gold-yellow; S – yellow; L – metallic; D – opaque; DE – 19.3; H – 2.5-3; CL – none; F – hackly; M – octahedral and cubic crystals, skeletal and dendritic aggregates, leaves, nuggets.

Origin and occurrence: Primary hydrothermal in ore veins, also in contact metamorphic deposits and pegmatites. Placer deposits are secondary. It occurs with pyrite, arsenopyrite, quartz, sylvanite, calaverite, krennerite and other minerals. Beautiful leaves and crystals of gold found in many localities in California, USA (Colorado Quartz mine, Nigger Hill and others). Fine leaf gold comes from Rosia Montana, Romania. The best crystals, skeletal octahedra, up to 50 mm (2 in) have been found in alluvial sediments near Gran Sabana, Roraima Shield, Venezuela. Gold wires up to 110 mm ($4^{5}/_{16}$ in) long were very rare in Ground Hog mine, Gilman, Colorado, USA. The largest known sheets of crystallized gold occurred in the Jamestown mine, California, USA, where a cavity, which yielded 49 kg (108 lb) of golden leaves, was discovered on 26.12.1992. The largest measures about 300 mm ($11^{13}/_{16}$ in) and has about 25.79 kg (56 lb 13 oz) of gold on it. Typical aggregates of fine gold wires come from Farncomb Hill near Breckenridge, Colorado, USA. Fine crystals were also found in Berezovsk, Ural mountains and in the Lena River basin, Siberia, Russia. Fine scales and larger nuggets from placer deposits were found in Klondike, Alaska; Tuolumne County, California, USA and in Ballarat, Victoria, Australia. Fine dendritic aggregates occurred in the Hope's Nose, Devon, UK. A unique find of leaves up to 100 mm (4 in) was made in Krepice near Vodnany, Czech Republic.

Application: practically the only source of gold as a metal; used in jewelry, electronics and medicine.

Mercury

Hg

TRIGONAL ● ●

Properties: C – tin white; L – metallic to adamantine; D – opaque; DE – 13.6; M – liquid at temperatures above -39°C (-38.2°F); R – very poisonous fumes.

Origin and occurrence: Hydrothermal in low-temperature ore deposits, also connected with hot springs. It is associated with cinnabar, calomel and other Hg minerals. It occurred in Almaden, Spain;

Mercury, 30 mm, Socrates Mine, California, U.S.A.

Idria and Avala, Serbia; New Almaden and New Idria, California, Terlingua, Texas, USA; Dedova hora, Czech Republic and Rudnany, Slovakia as droplets and liquid cavity fillings.

Application: chemical industry, measuring instruments, metallurgy.

Moschellandsbergite

Ag_2Hg_3

CUBIC ● ●

Properties: C – silver-white; S – silver-white; L – metallic; D – opaque; DE – 13.5; H – 3.5; CL – good; F – conchoidal; M – dodecahedral crystals and their combinations, granular, massive.

Origin and occurrence: Hydrothermal in low-temperature deposits, associated with cinnabar, tetrahedrite, pyrite and other minerals. Crystals, several mm long, were found in Moschellandsberg, Germany. They are also known from Sala, Sweden; Les Chalanches, France and Brezina, Czech Republic.

Moschellandsbergite, 6 mm grain, Moschellandsberg

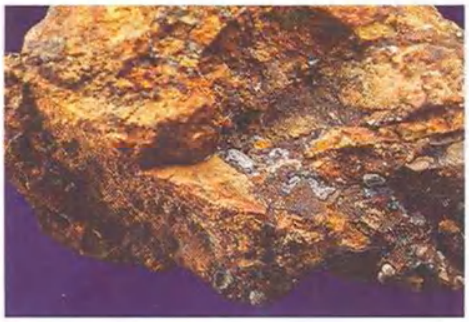

Lead

Pb

CUBIC ● ●

Properties: C – gray-white, tarnishes to lead-gray and gets dull; S – lead-gray; L – metallic; D – opaque; DE – 11.3; H – 1.5; CL – none; M – octahedral and cubic crystals, massive.

Origin and occurrence: Hydrothermal, also sedimentary (authigenic), associated with willemite and other minerals. The best specimens with crystals up to 40 mm (1 9/16 in) in size, come from LDngban, crystallized also from Pajsberg, Sweden. Octahedra, up to 10 mm (3/8in), are described from El Dorado, Gran Sabana, Venezuela. It also occurs in Franklin, New Jersey, USA and Jalpa, Zacatecas, Mexico.

Iron

Fe

CUBIC ● ●

Properties: C – steel-gray to black; S – gray; L – metallic; D – opaque; DE – 7.9; H – 4; CL – perfect; F – hackly; M – crystals, granular, massive.

Origin and occurrence: Terrestric iron occurs mainly in basic rocks, but it is also known from carbonate sediments and the petrified wood. The most famous locality is Blaafjeld near Uivfaq on Disko Island, Greenland, where masses up to 20 tons were found.

Lead, 53 mm, Langban, Sweden

Chunks, weighing over 10 kg (22 lb) come from Bühl near Kassel, Germany. Impregnations of iron in dolerite occur in the Khuntukun massif and masses up to 80 kg (176 lb) are known from Ozernaya Mt., Siberia, Russia.

Platinum

Pt

CUBIC ● ●

Properties: C – steel-gray to dark gray; S – steel-gray to silver-white; L – metallic; D – opaque; DE – 21.5; H – 4-4.5; CL – none; F – hackly; M – cubic crystals, nuggets, grains and scales.

Iron, 80 mm, Bühl, Germany

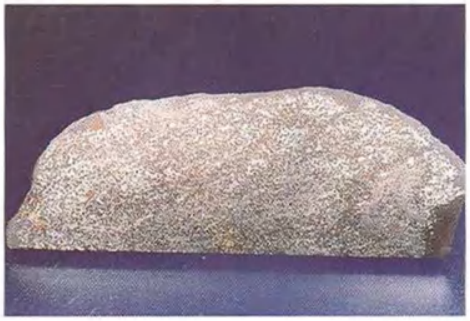

Origin and occurrence: Platinum occurs in magmatic segregations, together with chromite, olivine and magnetite in ultrabasic rocks; secondary in placers. The best crystals up to 15 mm ($^{19}/_{32}$ in) come from Konder in Khabarovsk Region, small crystals, but mainly nuggets, weighing up to 11.5 kg (25 lb 5 oz) found in the Tura River basin near Turiinsk, Ural Mountains. Primary platinum is known from deposits in the vicinity of Nizhniy Tagil, Ural Mountains, Russia; from the Onverwacht mine, Bushveld, South Africa and Sudbury, Ontario, Canada. Fine smaller nuggets, weighing up to 75 g (165 lb) were found in the Trinity River sediments in California, USA and the Choco River sediments in Columbia.
Application: chemical industry, catalytic convertors, rocket industry.

Platinum, 8 mm, Konder, Russia

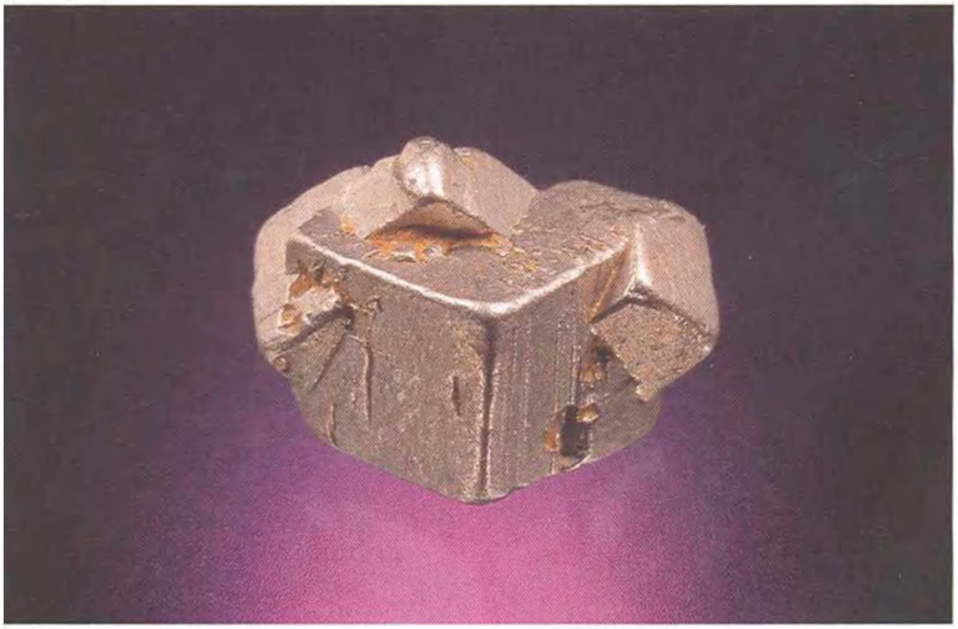

←+

Arsenic, 45 mm, Alden Island, Canada

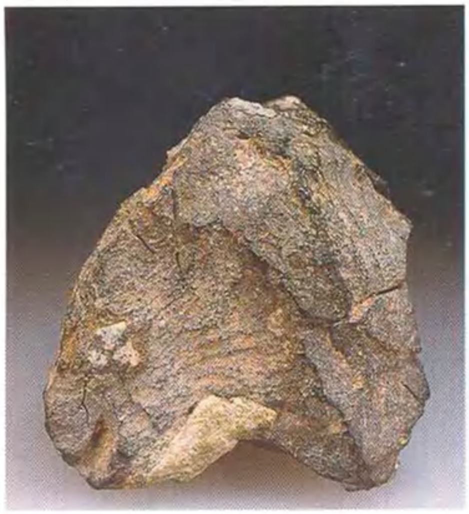

Arsenic

As

TRIGONAL ● ● ●

Properties: C – tin-white, tarnishes quickly to black; SB tin-white; L – metallic; D – opaque; DE – 5.8; H – 3.5; CL – perfect; F – uneven; M – rhombohedral crystals, botryoidal aggregates, granular, massive.
Origin and occurrence: Mainly hydrothermal, together with other As minerals. It forms massive veins, up to 200 mm (7⅞ in) thick, with botryoidal surface in Jachymov, Czech Republic and in Freiberg, Germany. In Akatani, Japan, spherical aggregates consisting of small crystals were found. It is also known from Sacarimb, Romania. Botryoidal aggregates of arsenic with leaf gold were found in the Royal Oak mine, Coromandel, New Zealand. Crystals of metamorphic origin come from Sterling Hill, New Jersey, USA.

Stibarsen, 58 mm, Atlin, Canada

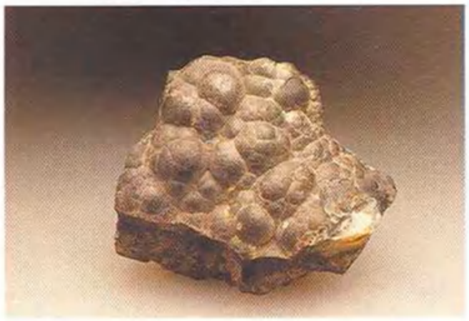

Antimony, 37 mm, New Brunswick, Canada

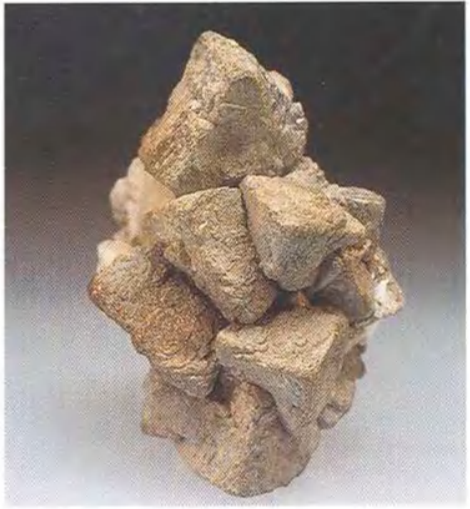

Stibarsen

SbAs

TRIGONAL ● ● ●

Properties: C – tin-white to gray, tarnishes black; S – gray; L – metallic, sometimes dull; D – opaque; DE – 6.3; H – 3-4; CL – perfect; M – indistinct crystals, botryoidal aggregates.
Origin and occurrence: In pegmatites with antimony, stibiotantalite and microlite; hydrothermal in ore veins withpyrargyrite, proustite, pyrostilpnite and dyscrasite. Beautiful botryoidal aggregates of stibarsen (previously labeled as allemontite), up to 100 mm (approx. 4 in) in size, come from Pribram and Trebsko, Czech Republic. Botryoidal aggregates up to 80 mm (3⅛ in) and imperfect crystals were found in quartz veins in Atlin, British Columbia, Canada. Fine specimens occurred in a Li-bearing pegmatite near Varutrask, Sweden.

Antimony

Sb

TRIGONAL ● ● ●

Properties: C – tin-white; S – gray; L – metallic; D – opaque; DE – 6.7; H – 3-3.5; CL – perfect; F – uneven; M – rhombohedral crystals, botryoidal aggregates, massive.
Origin and occurrence: Hydrothermal in ore veins with silver, stibnite, stibarsen, sphalerite and other minerals; also in pegmatites. As veinlets in a pegmatite in Varutrask, Sweden. Cleavable plates, up to 50 mm (2 in) known from Tornäva, Finland.

Massive aggregates up to 200 mm ($7^7/_8$ in) come from Pribram, Czech Republic. Rhombohedral crystals up to 10 mm ($^3/_8$ in) across and accumulations up to 300 mm ($11^{13}/_{16}$ in) in size described from Lake George, New Brunswick, Canada.

Bismuth

Bi

TRIGONAL ● ● ●

Properties: C – silver-white, tarnishes pink; S – silver-white; L – metallic; D – opaque; DE – 9.8; H – 2-2.5; CL – perfect; M – rhombohedral crystals, granular, massive.
Origin and occurrence: It is found in pegmatites, greisens and hydrothermal in ore veins together with chalkopyrite, arsenopyrite, lollingite, nickeline, breithauptite and many other minerals. Common in pegmatites in Anjanabonoina, Madagascar. Very fine crystals, up to 20 mm ($^{25}/_{32}$ in) known from Schlema and Hartenstein, Germany. Skeletal aggregates, overgrown with other arsenides, occurred in Jachymov, Czech Republic. Masses weighing several kg found in Bolivia (Tasna, Velaque) and Australia (Kingsgate, New South Wales). Cleavable masses up to 12 cm ($4^{11}/_{16}$in) described from Cobalt and Gowganda, Ontario, Canada.
Application: Bi ore.

Arsenolamprite

As

ORTHORHOMBIC ●

Properties: C – gray-white, it covers with a black coating; S – black, L – metallic to adamantine; D – opaque; DE – 5.6; H – 2; CL – perfect; M – acicular crystals, tabular and fan-shaped aggregates, massive.
Origin and occurrence: Hydrothermal in ore veins associated with arsenic, bismuth, silver and other minerals. Its crystals and veinlets were found with Cu arsenides in Cerny dul, Czech Republic. Occurs also in Jachymoc, Czech Republic and Marienberg Germany. Found recently in Cavnic, Romania.

Bismuth, 20 mm, Cinovec, Czech Republic

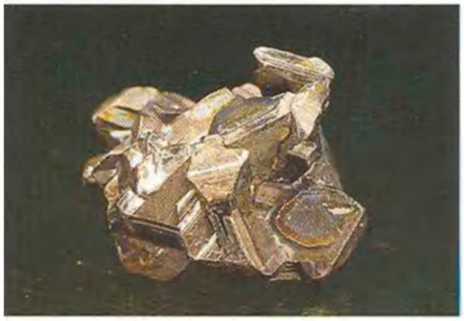

Arsenolamprite, 70 mm, Jachymov, Czech Republic

Graphite

C

HEXAGONAL ● ● ● ● ●

Properties: C – black to steel-gray; S – black to steel-gray; L – metallic, dull, earthy; D – opaque; DE – 2.3; H – 1-2; CL – perfect; M – hexagonal tabular crystals, massive.
Origin and occurrence: Metamorphic, from metamorphism of a sedimentarymaterial with C contents: also primary magmatic.Associated with many materials, stabile underconditions of the graphite origin. Crystals several cm in size known from Nordre Stromfjord. Greenland. Crystals were also found in Sterling Hill, New Jersey and Crestmore, California, USA Foliated aggregates are found in Sri Lanka (Radegara, Galle region). Accumulations in Buckingham and Grenville, Quebec, Canada are industrially important. Also common in Shunga deposit in Karelia, Russia; in Cesky Krumlov, Netolice and Blizna, Czech Republic.
Application: metallurgy, nuclear industry, production of lubricants.

Graphite, 130 mm, Krichim, Bulgaria

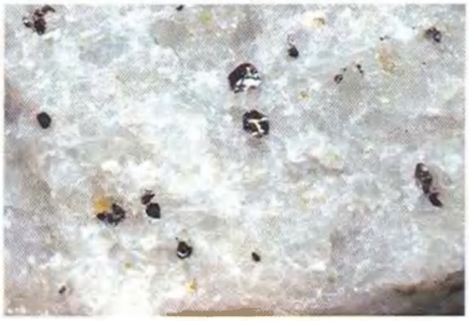

Diamond

C

CUBIC ● ●

Varieties: bort (opaque technical diamonds), balas (dark colored, spherical radial aggregates), carbonado (brown-black to black massive aggregates, up to egg sized)
Properties: C – colorless, yellow, brown, white, pink, black, red, blue, green; S – white; L – adamantine; D – transparent to translucent; DE – 3.5; H – 10; CL – perfect; F – conchoidal; M – octahedral and cubic crystals; LU – sometimes fluorescent, sometimes phosphorescent.
Origin and occurrence: Primary magmatic occurrences are limited to kimberlite pipes, secondary occurrences to placers. Large primary deposits are known from South Africa (Premier mine, Kimberley) and Yakutia, Russia (pipes in the vicinity of Mirnyi). Primary and secondary occurrences of diamonds are located in lamproites and placers near Argyle, Western Australia, Australia. Most historical diamonds from India (Golconda). Brazil (Diamantina, Minas Gerais), Congo, Angola and Namibia were found in placer deposits. Diamonds of industrial grade always prevail over the gem quality stones. The largest gem-grade diamond ever found, the Cullinan, weighing 3106 carats, comes from the Premier mine in Kimberley, South Africa. It yielded gem rough for 104 faceted stones, the heaviest of which weighs 531 carats. The largest faceted diamond known, called Golden Jubilee, was found in the same place in 1986. It weighed 755 carats before cutting and as a finished stone it weighs 545.65 carats. The dark blue Hope (44 carats) and the green Dresden (76 carats) probably came from India. Absolutely unique red diamond, weighing 5 carats, which is at the Smithsonian Institution, Washington, DC, USA, is of unknown origin.
Application: the most popular gemstone, bort and carbonado varieties are used as abrasives.

Sulfur

S

ORTHORHOMBIC ● ● ● ●

Properties: C – sulfur-yellow, yellow-brown, greenish, reddish to yellowish-gray; S – white; L – resinous to greasy; D – transparent to translucent; DE – 2.1; H – 1.5-2.5; CL – imperfect; F – conchoidal to uneven; M – dipyramidal, disphenoidic and thick tabular crystals, botryoidal and stalactitic aggregates.
Origin and occurrence: Hydrothermal product of fumaroles, product of an activity of microorganisms, disintegration of sulfides and acidic chemical reactions; associated with gypsum, anhydrite, aragonite, calcite,

Diamond, 15 mm x, Mirnyi, Russia

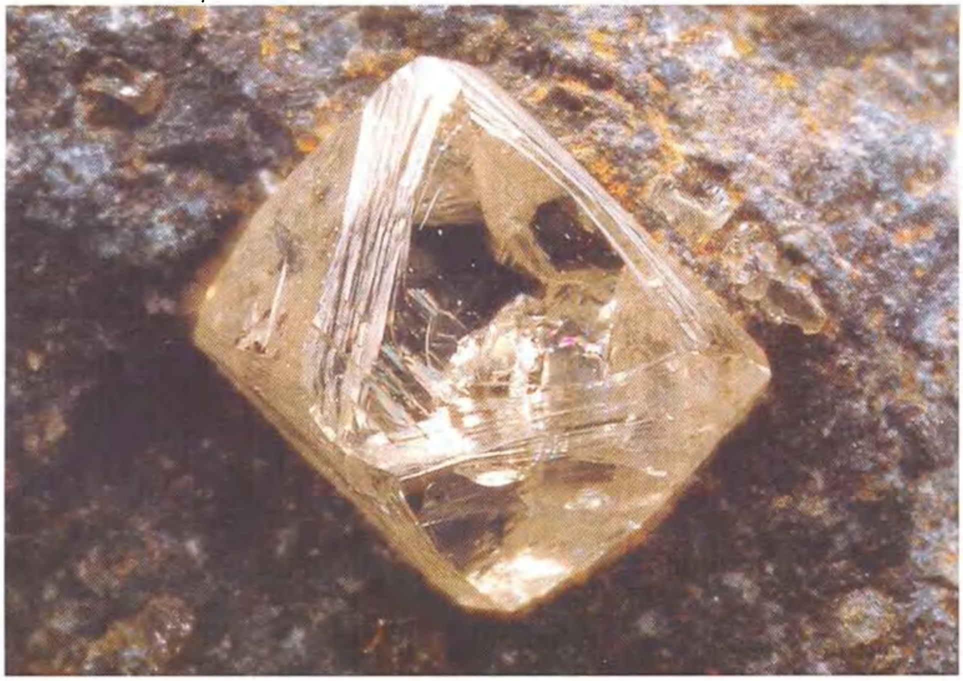

Sulphur, 66 mm, Sicily, Italy

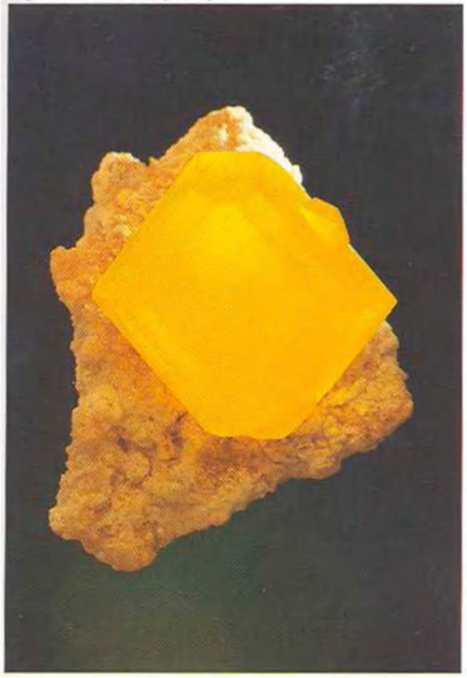

celestite and halite. The world's best crystals come from many localities near Girgenti, Sicily, Italy (Caltanisseta, Cianciana) where they reached up to 12 cm ($4\frac{11}{16}$ in) in size. Fine crystals are also known from Tarnobrzeg, Poland and Yavokskoye near Lvov, Ukraine. As a product of solfataras it occurs in many volcanically active places, like Solfatara near Pozzuoli, Italy or in sulfur lava near Shiretoku, Japan. Sulfur layers, up to 30 m (100 ft) thick, associated with salt diapirs, are located near Charles Lake, Louisiana, USA. It originates during intensive oxidation/reduction reactions of pyrite in Rio Tinto, Spain and in Kostajnik, Serbia.
Application: chemical, paper-making, rubber and leather-making industries, agriculture.

Selenium, 110 mm, Kladno, Czech Republic

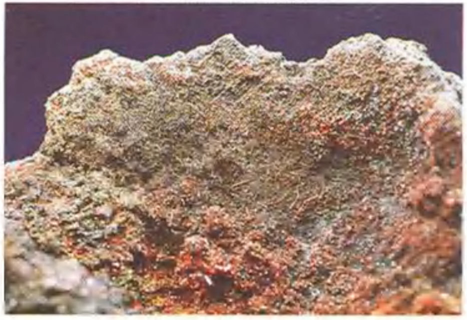

Selenium

Se

TRIGONAL ● ●

Properties: C – gray to red-gray; S – red; L – metallic, D – opaque to translucent; DE – 4.8; H – 2; CL – good; M – acicular crystals, droplets of vitreous surface, felt-like aggregates.
Origin and occurrence: Secondary, resulting from alteration, fumaroles and from burning coal dumps, with sulfur, sal ammoniac and other sulfates. Also from oxidation of organic compounds in U- and V-bearing deposits of the Colorado Plateau type, associated with pyrite, zippeite and other minerals. Red needles up to 20 mm ($\frac{25}{32}$ in) long come from the United Verde mine, Jerome, Arizona, USA. Black selenium needles found in burning coal dumps in Kladno and Radvanice, Czech Republic. Occurs with ores of U and V along sandstone fissures in the Peanut Mine, Bull Canyon, Colorado, USA. Occurred through volcanic activity in Vulcano, Lipari Islands, Italy.

Tellurium

Te

TRIGONAL ● ●

Properties: C – tin-white; S – gray; L – metallic; D – opaque; DE – 6.2; H – 2-2.5; CL – perfect; M – prismatic and acicular crystals, granular, massive.
Origin and occurrence: Primary hydrothermal in low-temperature ore deposits; it originates also as secondary through the oxidation/reduction reactions of tellurides. It is associated with gold, sylvanite, altaite, pyrite and other minerals. Crystals up to 30 mm ($1\frac{3}{16}$ in) long are known from Balya, Turkey. Crystals up to 20 mm ($\frac{25}{32}$ in) long occurred in the Au deposits Cripple Creek and Colorado City, USA. Crystals up to 10 mm ($\frac{3}{8}$ in) across come from Kawazu and Suzuki, Japan. Rich cleavage masses and crystals up to 70 mm (2¾ in) long were found in Uzbekistan. Crystallized tellurium is also known from Fata Baii and Baia de Aries, Romania.

Tellurium, 1 mm xx, Zlatna, Romania

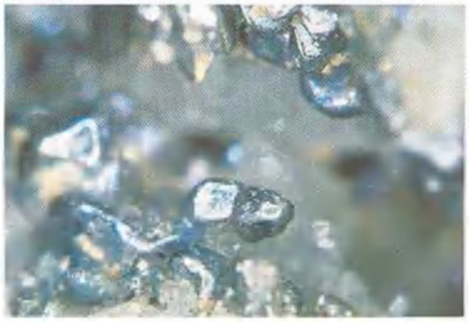

2. Sulfides

Algodonit

Cu_6As

HEXAGONAL ● ●

Properties: C – steel-gray to silver-white, it quickly covers with a brown coating on air; S – gray; L – metallic; D – opaque; DE – 8,7; H – 4; CL – none; F – conchoidal; M – crystals, granular, massive.
Origin and occurrence: Hydrothermal, mainly intimately inter-grown with other Cu arsenides. Its largest accumulations are known from Cu deposits in melaphyres in the Keweenaw Peninsula, Lake Superior, Michigan, USA. It is also known from Chile (Algodones mine near Coquimbo, Atacama). Other localities are Talmessi, Iran and Långban, Sweden.

Domeykite

Cu_3As

CUBIC ● ● ●

Properties: C – tin-white to steel-gray, it tarnishes yellow and covers with a brown coating; S – gray; L – metallic; D – opaque; DE – 7,9; H – 3-3,5; CL – none; F – uneven; M – botryoidal aggregates, massive.
Origin and occurrence: Hydrothermal with copper, cuprite, algodonite and silver. Common in masses, weighing several kg, together with algodonite in the Keweenaw Peninsula, Michigan, USA. The largest accumulations are known from Talmessi and Anarak, Iran. It occurs near Copiap and Chanarcillo, Chile. Massive aggregates in cuprite up to 50 mm (2 in) come from Biloves, Czech Republic.

Sphalerite, 56 mm, Picos de Europa, Spain
Algodonite, 60 mm, Keweenaw Peninsula, U.S.A.

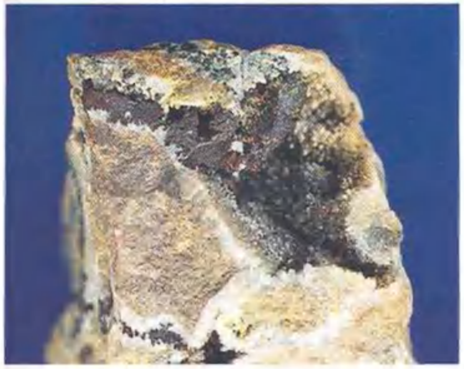

Domeykite, 3 mm x, Rudabanya, Hungary

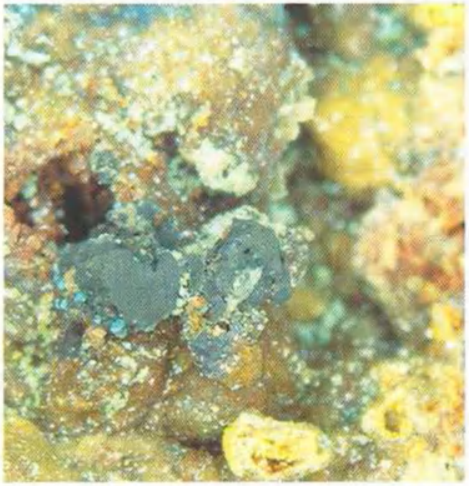

Allargentum

$Ag_{1-x}Sb_x$

HEXAGONAL ●

Properties: C – silver-white; S – gray; L – metallic; D – opaque; DE – 10,1; H – not determined; CL – none; M – small crystals, granular.
Origin and occurrence: Hydrothermal in the silver-bearing ore veins, associated with silver, breithauptite and dyscrasite. Crystals up to 1 mm ($^{1}/_{32}$ in) known from Hartenstein, Germany. Its intergrowths with silver ores were found in Cobalt, Ontario, Canada. It is also known from Broken Hill, New South Wales, Australia and microscopic in Rejska vein in Kutna Hora, Czech Republic.

Allargentum, 50 mm, Schlema, Germany

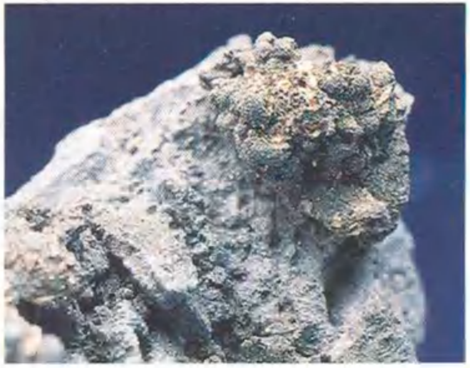

Dyscrasite

Ag_3Sb

ORTHORHOMBIC ● ●

Properties: C – silver-white, it tarnishes yellow to black; S – silver-white; L – metallic; D – opaque; DE – 9,7; H – 3,5-4; CL – good; F – uneven; M – pyramidal and prismatic crystals, granular, massive.
Origin and occurrence: Hydrothermal in the ore veins, associated with silver, stibarsen, pyrargyrite, calcite and other minerals. The best specimens come from the silver-bearing veins, cross-cutting the U deposit Hajc near Pøíbram, Czech Republic, where prismatic crystals up to 50 mm (approx. 2 in) long and striated tabular twins were found. They are mostly embedded in stibarsen; all the specimens, appearing in the mineral shows, are etched out of matrix. Deformed crystals of completely different habit are known from St. Andreasberg, Germany. Crystals occurred also in the Consols mine in Broken Hill, New South Wales, Australia.

Chalcocite

Cu_2S

MONOCLINIC ● ● ● ●

Properties: C – lead-gray to black; S – lead-gray to black; L – metallic; D – opaque: DE – 5,8; H – 2,5-3; CL – imperfect: F – conchoidal; M – prismatic to tabular crystals, granular, massive.
Origin and occurrence: Hydrothermal, also sedimentary and metamorphic, mostly secondary, in the oxidation and cementation zones of ore deposits. It occurs together with pyrite, chalcopyrite, covellite, bornite and other minerals. Crystals up to 25 cm ($9\,^{7}/_{8}$ in) found in the M'Sesa mine, Zaire. Beautiful crystals several cm across come from Redruth and St Just, Cornwall, UK. Crystals over 20 mm ($^{25}/_{32}$ in) across occurred in Bristol, Connecticut and in Butte, Montana. Crystals up to 50 mm (2 in) in size known from the Flambeau mine near Ladysmith, Wisconsin, USA. Shiny cyclic twins of crystals up to 20 mm ($^{25}/_{32}$ in) found in Dzhezkazgan, Kazakhstan. Massive aggregates are important Cu ore in Rio Tinto, Spain; Bor. Serbia; Bisbee, Arizona, USA and Tsumeb, Namibia.
Application: important Cu ore.

Djurleite

$Cu_{31}S_{16}$

MONOCLINIC ● ● ●

Properties: C – lead-gray to black; S – lead-gray; L – metallic: D – opaque; DE – 5,8; H – 2,5-3; CL – none; F – conchoidal: M – short prismatic to tabular crystals, granular, massive.
Origin and occurrence: Secondary as a product of the cementation zone in ore deposits. Crystals up to 10 mm ($^{3}/_{8}$ in) across known from the Botallack mine

Dyscrasite, 75 mm, P?ibram, Czech Republic

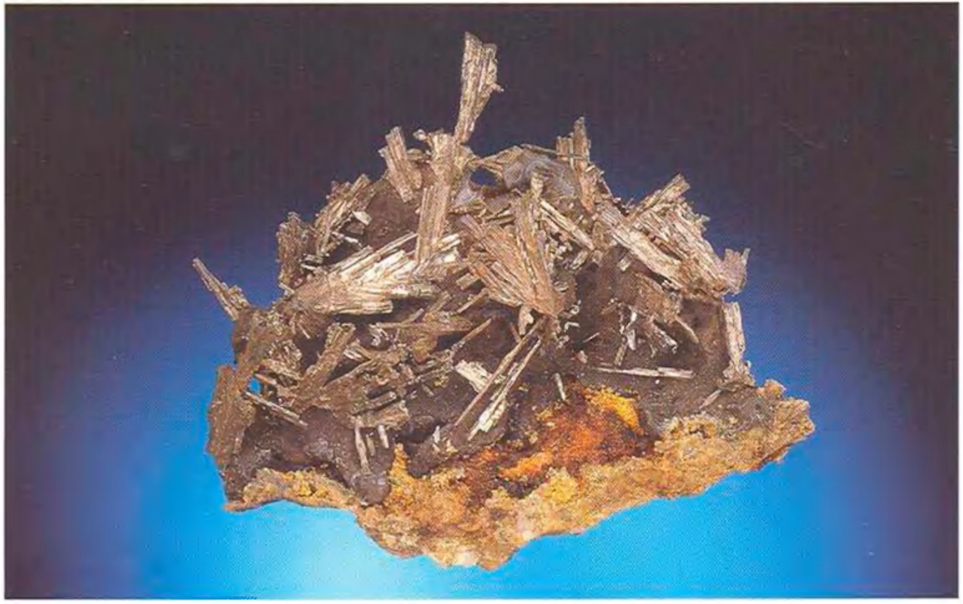

√ +

Chalcocite, 17 mm x, Cornwall, UK

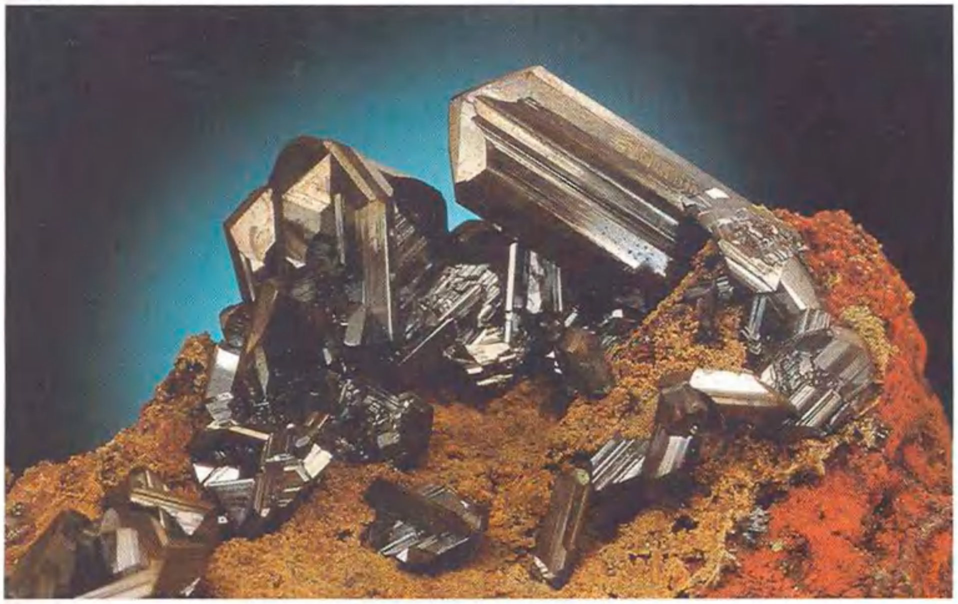

near St. Just, Cornwall, UK. Aggregates of thick acicular crystals up to 30 mm (1 3/16 in) long found in Dzhezkazgan, Kazakhstan.
It occurs in massive form in many porphyry copper deposits (Butte, Montana; Bisbee, Arizona, USA), also in Tsumeb, Namibia, together with chalcopyrite, pyrite and other minerals.

Djurleite, 3 mm xx, Dzhezkazgan, Kazakhstan

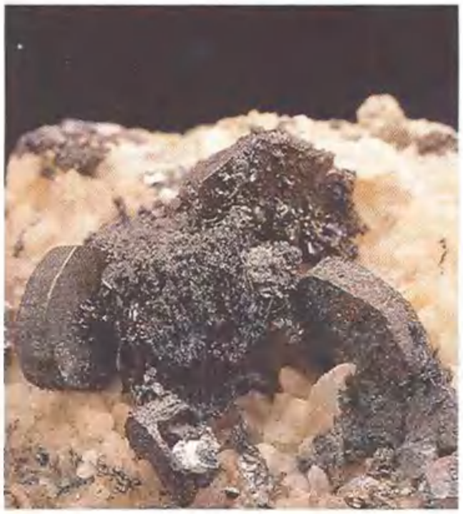

Berzelianite

Cu_2Se

CUBIC ● ● ●

Properties: C – silver-white, tarnishes black; S – silver shiny; L – metallic; D – opaque; DE – 7,3; H – 2,5; CL – none; F – uneven; M – granular, massive.
Origin and occurrence: Hydrothermal, together with other selenides in U, Fe and Au deposits. It is the main mineral in the selenide mineralization in Tilkerode, Germany. Grains, up to several tens of cm across, greenish tarnished, occurred together with other selenides in Bukov, Habøi, Petrovice and Pr edborice , Czech Republic. Similar occurrence is known from near Pinky Fault near Athabasca Lake, Saskatchewan, Canada.

Berzelianite, 60 mm, Bukov, Czech Republic

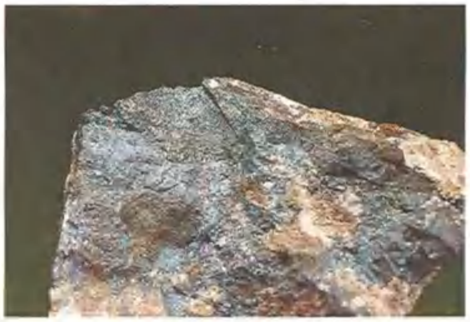

Bornite, 15 mm xx, Dzhezkazgan, Kazakhstan

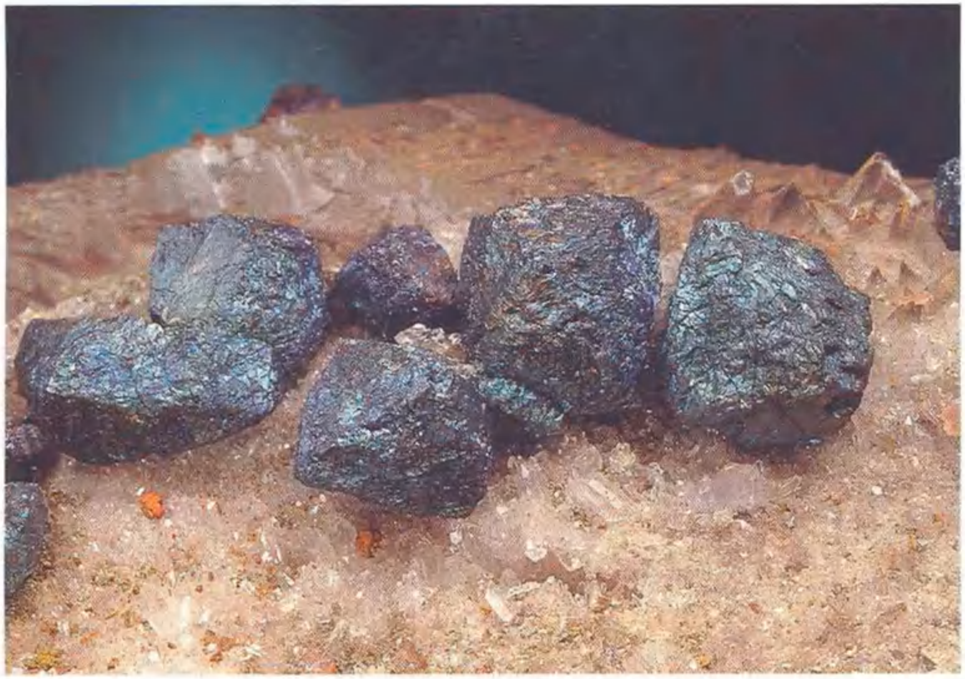

Bornite

Cu_5FeS_4

ORTHORHOMBIC ● ● ● ●

Properties: C – copper-red, tarnishes iridescent; S – gray-black; L – metallic; D – opaque; DE – 5,1; H – 3-3,5; CL – imperfect; F – uneven to conchoidal; M – pseudo-cubic octahedral crystals, massive.
Origin and occurrence: Magmatic, hydrothermal, sedimentary, in skarns and pegmatites together with chalcocite, chalcopyrite, pyrite, quartz and other minerals. Fine crystals up to 10 mm (3/8 in) across are known from Carn Brea, Cornwall, England, UK. Crystals up to 30 mm (13/16 in) come from Likasi, Shaba, Zaire. Beautiful crystals up to 40 mm (1 9/16 in) across were found recently together with chalcocite in Dzhezkazgan, Kazakhstan. Massive aggregates are common and used as Cu ore in Kipushi, Shaba, Zaire. Fine-grained, sedimentary bornite occurs in Cu-bearing shales in Mansfeld, Germany, where it forms the main ore layer. Crystals, up to 20 mm (25/32 in) in size, occurred in the Cole shaft and masses, weighing several thousands of tons, were mined in the Campbell shaft, Bisbee, Arizona, USA. *Application:* Cu ore.

Bornite, 3 mm xx, Dzhezkazgan, Kazakhstan

Umangite

Cu_3Se_2

TETRAGONAL ● ● ●

Properties: C – blue-black with reddish tint, tarnishes purple; S – black; L – metallic; D – opaque; DE – 6,6; H – 3; CL – imperfect; F – uneven to conchoidal; M – granular, massive.
Origin and occurrence: Hydrothermal in ore veins together with other selenides (clausthalite, berzelianite). It is common, associated with berzelianite in Tilkerode, Germany; Sierra de Umango, Argentina and Slavkovice, Czech Republic. Larger accumulations occur in the Martin Lake mine near Athabasca Lake, Canada.

Acanthite

Ag_2S

MONOCLINIC ● ● ●

Properties: C – black; S – black; L – metallic; D – opaque; DE – 7,2; H – 2-2,5; CL – none; F – uneven; M – pseudo-cubic crystals, massive. It mainly occurs as paramorphs after argentite (high-temperature phase of the same composition).

Origin and occurrence: Hydrothermal in ore veins. Beautiful crystals over 50 mm (2 in) long occurred in the Himmelsfürst mine in Freiberg, in Annaberg and Schneeberg, Germany. Acicular crystals are known from Jachymov, Czech Republic. It is common in association with silver, proustite, pyrargyrite, polybasite, stephanite, galena and other minerals in Mexico. Probably the best paramorphs after argentite up to 70 mm (2¾ in) across come from the Rayas mine, Guanajuato. Fine crystals occur in the Las Chispas mine, Arizpe, Sonora and many localities in Zacatecas, Chihuahua.

Application: important Ag ore.

Argentite

Ag_2S

CUBIC ● ● ●

Properties: C – black-gray, tarnishes black; S – black; L – metallic; D – opaque; DE – 7,1; H – 2-2,5; CL – imperfect; F – uneven to conchoidal; M – octahedral and cubic crystals, dendritic aggregates, massive. Stabile at temperatures over 179EC (354.2°F), below this temperature there are paramorphs of acanthite after argentite.

Umangite, 40 mm, Beaverlodge Lake, Canada

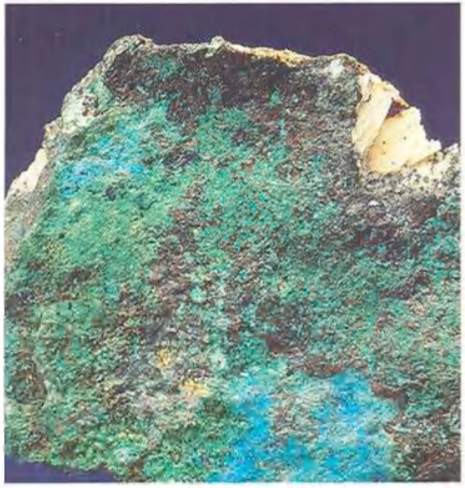

Acanthite, 23 mm, Arizpe, Mexico

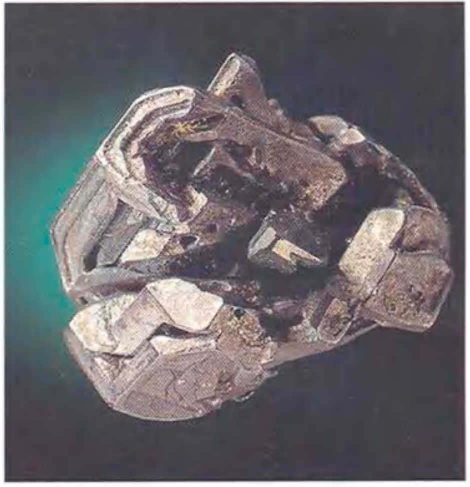

Origin and occurrence: Hydrothermal in low-temperature ore deposits, associated with silver, galena and Ag sufosalts. Occurs between the oxidation and cementation zone with stromeyerite, silver, jalpaite, iodargyrite and other minerals. Fine crystals up to 40 mm ($1^{9}/_{16}$ in) across, are known from Freiberg and Schneeberg, Germany. Similar crystals found in Jachymov and Midinec, Czech Republic. Crystals up to 30 mm ($1^{3}/_{16}$ in) occur in Sarrabus, Sardinia, Italy. Maybe the best argentite crystals occurred in Mexico (Arizpe, Sonora; Zacatecas; Guanajuato), where crystals reached up to 40 mm ($1^{9}/_{16}$ in). Fine crystals up to 20 mm ($^{25}/_{32}$ in) across reported from Chañarcillo in Chile.

Application: important Ag ore.

Argentite, 29 mm, Zacatecas, Mexico

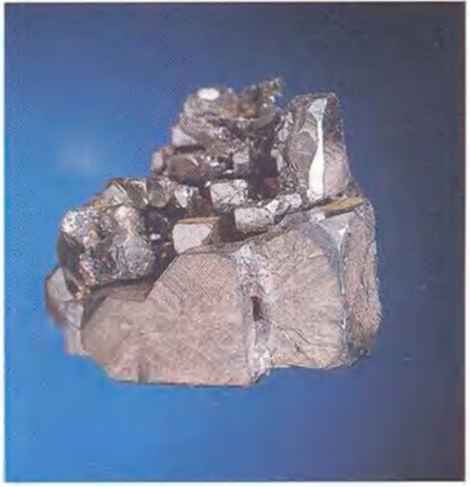

Aguilarite, 4 mm xx, Guanajuato, Mexico

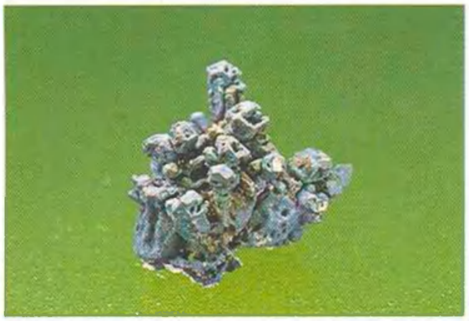

Hessite

Ag_2Te

MONOCLINIC ● ● ●

Properties: C – lead to steel-gray, tarnishes black; S – light gray; L – metallic; D – opaque; DE – 8,4; H – 2-3; CL – imperfect; F – even; M – pseudo-cubic crystal combinations, granular, massive.
Origin and occurrence: Hydrothermal in medium- and low-temperature ore veins together with calaverite, sylvanite, altaite, gold, tellurium and other sulfides. The best specimens with crystals up to 20 mm ($^{25}/_{32}$ in) across ($^{25}/_{32}$ in) and aggregates up to 100 mm (4 in) come from the Bote mine, Romania. Small crystals occur in the Jamestown mine, California, USA. Massive aggregates were found in Gold Hill, Colorado, USA and Moctezuma, Mexico. Aggregates, up to tens of cm^2 (several sq in) in size, were known in the Zavodinskii mine, Altai, Kazakhstan.

Aguilarite

Ag_4SeS

ORTHORHOMBIC ●

Properties: C – lead-gray, tarnishes black; S – gray-black; L – metallic; D – opaque; DE – 7,7; H – 2,5; CL – none; F – hackly; M – skeletal crystals, massive.
Origin and occurrence: Hydrothermal in ore veins, together with silver, stephanite, proustite, pearceite, calcite and quartz. The best crystallized specimens with crystals up to 30 mm ($1^{3}/_{16}$ in) across come from the San Carlos mine, Guanajuato and Chontalpan, Taxco, Guerrero, Mexico. It is also known in inter-

Hessite, 31 mm, Botes, Romania

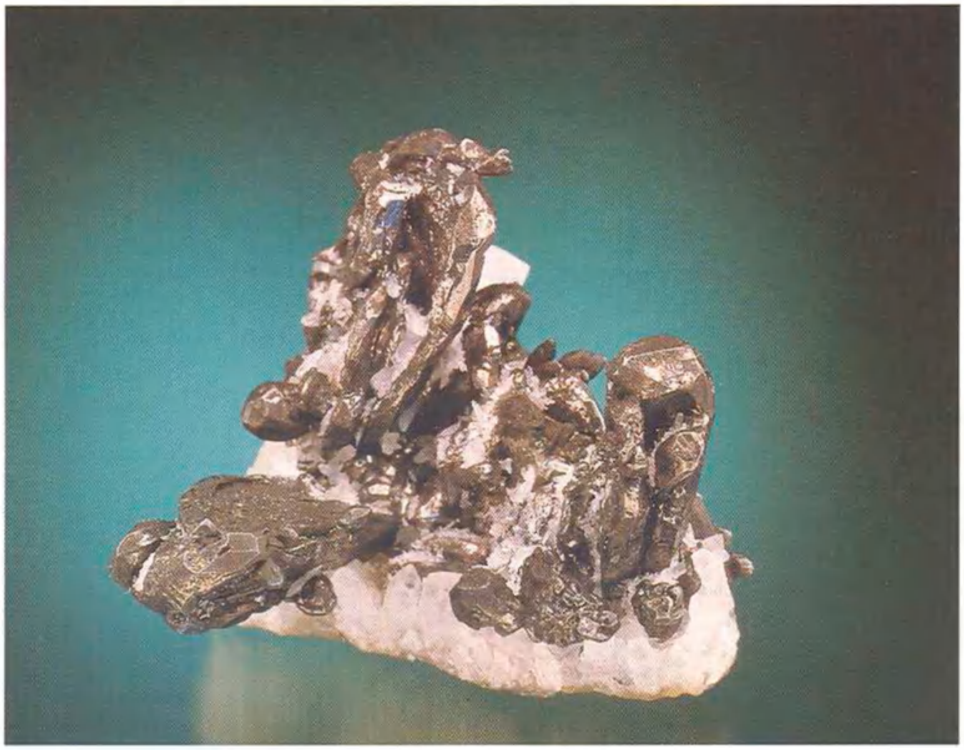

growths with acanthite and naumannite from the Comstock Lode, Virginia City, Nevada, USA.

Argyrodite

Ag_8GeS_6

ORTHORHOMBIC ●

Properties: C – steel-gray, tarnishes black; S – gray-black; L – metallic; D – opaque; DE – 6,3; H – 2,5-3; CL – none; F – uneven to conchoidal; M – combinations of cubic crystals, botryoidal aggregates, massive.
Origin and occurrence: Hydrothermal in low-temperature base metal deposits, associated with Ag sulfosalts. It occurred in crystals in the Himmelsfhrst mine, Freiberg, Germany. Crystals were also found in several localities in Bolivia (Atoche, Colquechaca, Potosí). Crystal measuring 60 mm (24 in) across is reported from Porco, Bolivia.

Stromeyerite

AgCuS

ORTHORHOMBIC ● ● ●

Properties: C – dark steel gray, tarnishes blue; S – steel-gray; L – metallic; D – opaque; DE – 6,3; H – 2,5-3; CL – none; F – conchoidal; M – pseudo-hexagonal tabular crystals, massive.
Origin and occurrence: Mostly secondary in the cementation zone of ore veins, associated with freibergite, bornite, chalcocite, galena and other minerals. Fine tabular crystals up to 10 mm (⅜ in) across found in Dzhezkazgan, Kazakhstan, where pseudo-morphs of stromeyerite after silver wires also occur. Skeletal prismatic pseudo-morphs after chalcocite crystals come from Vraneice, Czech Republic. Massive aggregates are common in many deposits in Colorado, USA (Aspen; Red Mountain), Chile (Copiapo), Bolivia (Potosi) and Canada (Cobalt, Ontario).

Stromeyerite, 6 mm xx, Dzhezkazgan, Kazakhstan

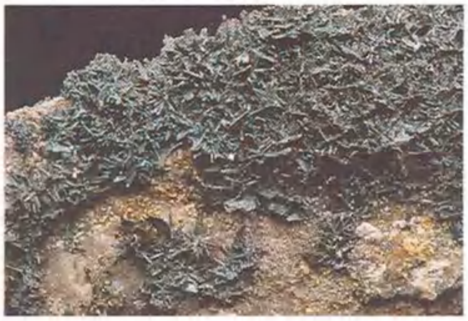

Argyrodite, 34 mm, Mexico

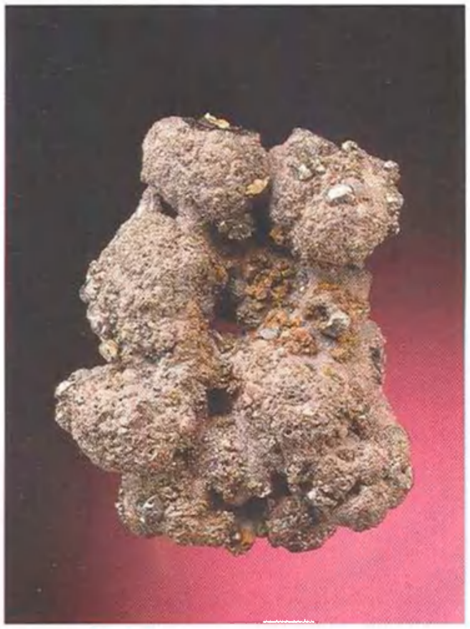

Jalpaite

Ag_3CuS_2

TETRAGONAL ● ●

Properties: C – light gray, tarnishes dark gray to iridescent; S – black; L – metallic; D – opaque; DE – 6,8; H – 2-2,5; CL – good; F – conchoidal; M – crystals, granular.
Origin and occurrence: Hydrothermal in low-temperature ore veins. Crystals up to 25 mm (1 in) across known from Jalpa, Queretaro, Mexico. Crystals up to 30 mm (13/16 in) across come from the Caribou mine, Colorado, USA. Massive aggregates, associated with galena, sphalerite, pyrite, stromeyerite, polybasite and other minerals, occurred in Pøibram, Czech Republic.

Jalpaite, 30 mm veinlet, P?íbram, Czech Republic

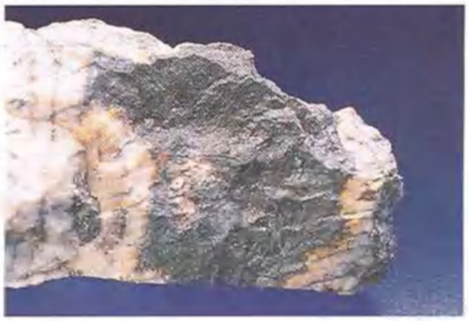

Pentlandite, 90 mm, Sudbury, Canada

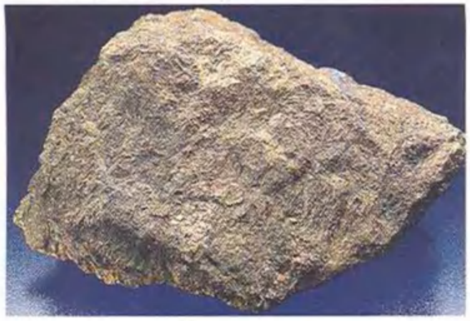

Pentlandite

$(Fe,Ni)_9S_8$

CUBIC ● ● ● ●

Properties: C – light bronze to red-brown; S – light bronze; L – metallic; D – opaque; DE – 5,0; H – 3,5-4; CL – none; F – conchoidal; M – crystals, massive.
Origin and occurrence: Typical magmatic liquid mineral, associated with pyrrhotite and chalcopyrite. It is common in Sudbury, Ontario, Canada, as mostly microscopic inclusions in chalcopyrite, but also as imperfect crystals. Large accumulations are known from Talnakh near Norilsk, Siberia, Russia, where it occurs together with Cu, Pt and Pd sulfides. It is also important ore in the deposit near Rustenburg, South Africa.
Application: the most important Ni ore.

Sphalerite, 70 mm, Olkusz, Poland

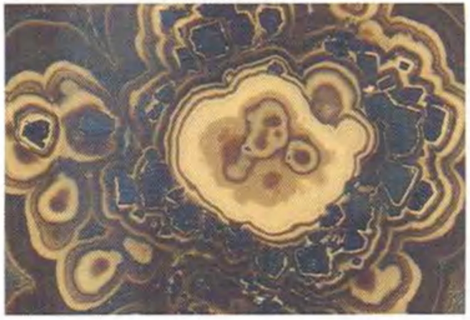

Sphalerite

ZnS

CUBIC ● ● ● ●

Varieties: cleiophane (green, yellow, orange), marmatite (black)

Properties: C – colorless, yellow, orange, green, brown, black; S – brownish, light yellow, white; L – resinous to adamantine; D – transparent, translucent, opaque; DE – 4,1; H – 3,5-4; CL – perfect; F – conchoidal; M – tetrahedral and dodecahedral crystals, botryoidal, fibrous and stalactitic aggregates, massive; LU – sometimes orange.
Origin and occurrence: Magmatic (liquid, in pegmatites); hydrothermal in low- to high-temperature deposits, skarns, hydrothermal sedimentary deposits;

Sphalerite, 31 mm, Elmwood, U.S.A.

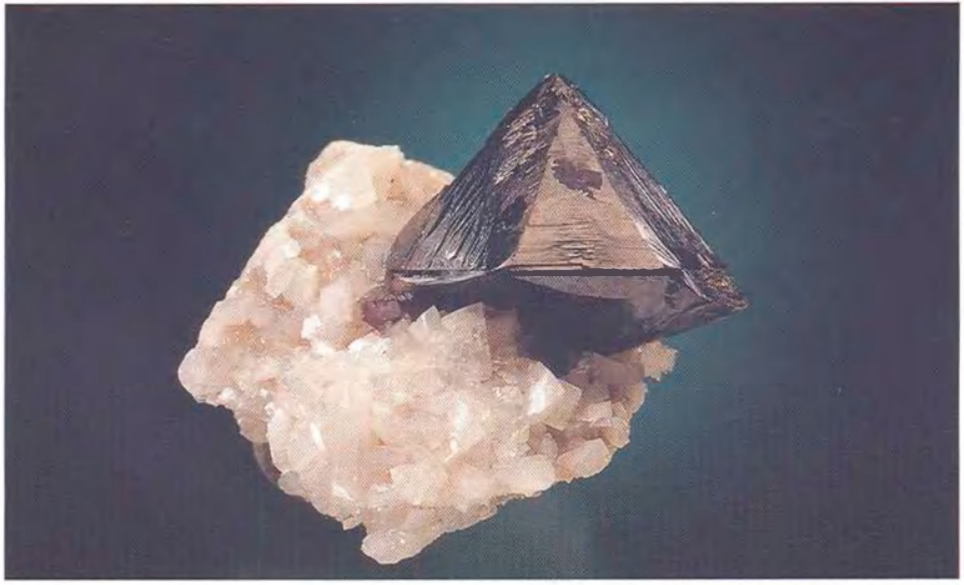

rare sedimentary and metamorphic. It occurs together with galena, pyrite, chalcopyrite, marcasite, fluorite, barite, quartz and other minerals. Beautiful crystals up to 100 mm (4 in) across come from Trepea, Serbia. Green and red crystals up to 100 mm (4 in) known from Cananea, Sonora, Mexico. Fine yellow crystals up to 30 mm ($1^{3}/_{16}$ in) were common in Banska “tiavnica, Slovakia; similar crystals occur in Madan, Bulgaria. The most beautiful yellow, orange and red crystals up to 150 mm (6 in) across found in Picos de Europa, Santander, Spain. They are sometimes faceted. Brown crystals up to 50 mm (2 in) come from Joplin, Missouri; stalactitic aggregates up to 150 mm (6 in) long come from Galena, Illinois, USA. Perfect black, shiny crystals and twins up to 50 mm (2 in) are famous from Dalnegorsk, Russia; yellow crystals, up to 30 mm ($1^{3}/_{16}$ in), occurred in Dzhezkazgan, Kazakhstan. Transparent crystals up to 30 mm ($1^{3}/_{16}$ in) also found in Franklin, New Jersey, USA. Green crystals up to 100 mm (approx. 4 in) across occurred in the Big Four mine, Colorado, USA. Crystals up to 50 mm (2 in) known from the Oppu mine, Aomori, Japan.
Application: principal Zn ore.

Coloradoite

HgTe

CUBIC ● ●

Properties: C – black-gray; S – black-gray; L – metallic; D – opaque; DE – 8,1; H – 2,5; CL – none; F – uneven to conchoidal; M – granular, massive.
Origin and occurrence: Hydrothermal, associated with altaite, calaverite, krennerite, gold, pyrite and other minerals in Au-bearing veins. It was common in Cripple Creek and in the Smuggler mine, Colorado; in the Norwegian mine, California, USA. Grains, reaching up to several mm, come from Jilove, Czech Republic.

Coloradoite, 70 mm, Kalgoorlie, Australia

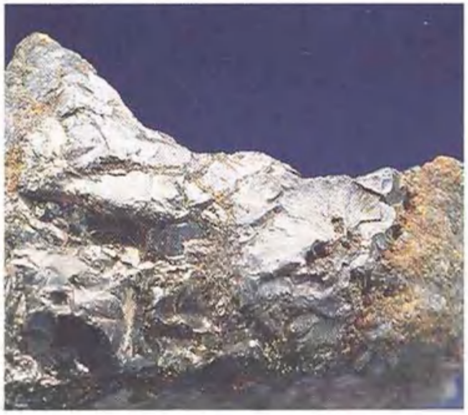

Chalcopyrite, 10 mm xx, Cavnic, Romania

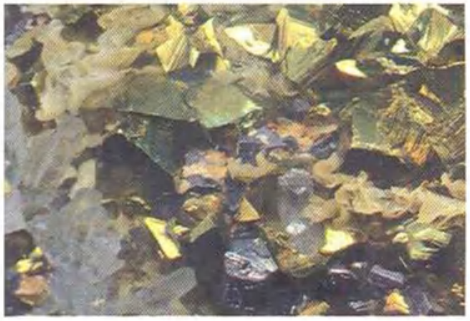

Chalcopyrite

$CuFeS_2$

TETRAGONAL ● ● ● ● ●

Properties: C – brass-yellow, tarnishes iridescent; S – green-black; L – metallic; D – opaque; DE – 4,3; H – 3,5-4; CL – imperfect; F – uneven; M – tetrahedral crystals, botryoidal aggregates, massive.
Origin and occurrence: Magmatic, hydrothermal and sedimentary, in association with sphalerite, galena, tetrahedrite, pyrite and many other sulfides. Fine crystals up to 30 mm ($1^{3}/_{16}$ in) across are known from Banska “tiavnica, Slovakia and from Cavnic, Romania.
Crystals up to 120 mm ($4^{11}/_{16}$in) across, associated with other sulfides, come from the Nikolai mine in Dalnegorsk, Russia. Beautiful crystals up to 120 mm ($4^{11}/_{16}$in) found in Japan (Arawaka, Osarizawa). Fine crystals reaching up to several cm occur in Peru (Huanzala, Huaron). Massive aggregates are important Cu ore in Sudbury, Ontario, Canada; Bingham, Utah; Bisbee, Arizona, USA and Rio Tinto, Spain.
Application: important Cu ore.

Chalcopyrite, 15 mm x, Ground Hog Mine, U.S.A.

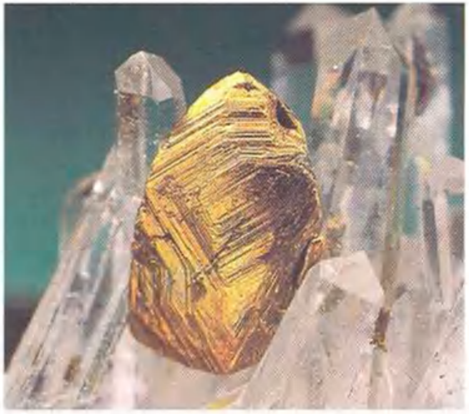

Luzonite, 60 mm, Recsk, Hungary

Luzonite

Cu_3AsS_4

TETRAGONAL ● ●

Properties: C – dark pink-brown; S – black; L – metallic; D – opaque; DE – 4,5; H – 3,5; CL – good; F – uneven to conchoidal; M – crystals, granular, massive.
Origin and occurrence: Hydrothermal in low- to medium-temperature veins, associated with enargite, tetrahedrite, sphalerite, bismuthinite, Ag sulfosalts and other minerals. Crystals up to 40 mm ($1\frac{9}{16}$ in) across come from Quiruvilca; it is common in Cerro de Pasco, Peru. It is also common in the Teine mine, Hokkaido, Japan; crystals also occur in Kinkwaseli, Taiwan. It is known from Bor, Serbia and Recsk, Hungary.

Stannite

Cu_2FeSnS_4

TETRAGONAL ● ● ●

Properties: : C – steel-gray to black, tarnishes blue; S – black; L – metallic; D – opaque; DE – 4,5; H – 4; CL – imperfect; F – uneven; M – pseudo-octahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in high-temperature Sn deposits. The best specimens come from Bolivia; crystals up to 50 mm (2 in) across known from Llallagua; crystals up to 30 mm ($1\frac{3}{16}$ in) from Chocaya and cross-like inter-growths from the San Jose and Itos mines near Oruro. It occurs as massive vein fillings in Carn Brea, Cornwall, UK and Cínovec, Czech Republic. It was also found in amblygonite pegmatites in Caceres, Spain and in quartz-amblygonite veins near Verneøov, Czech Republic.
Application: Sn ore.

Stannite, 10 mm xx, Potosí, Bolivia

Germanite, 40 mm, Tsumeb, Namibia

Germanite

$Cu_{26}Fe_4Ge_4S_{32}$

CUBIC ● ●

Properties: C– pink to purple-grey, S – black, L - metallic, D - opaque, DE – 4,5, H - 4, CL – none, F – uneven, M – tetrahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal, often inter-grown with tennantite, bornite and other minerals. It was important only in one particular place in Tsumeb, Namibia, where larger accumulations were found. Smaller occurrences are known from the Shikanai mine, Japan; small cubic crystals come from the Humboldt mine, Colorado, USA.

Tennantite

$Cu_{12}As_4S_{13}$

CUBIC ● ● ● ●

Properties: C – steel-gray; S – black, brown to dark

red; L – metallic; D – opaque; DE – 4.6; H – 3-4,5; CL – none; F – conchoidal to uneven; M – tetrahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in ore veins and greisens with pyrite, calcite, dolomite, quartz and other sulfides and Cu-Pb-Zn-Ag sulfosalts. Fine crystals are known from Cornwall, UK (Wheal Jewel, Gwennap; Carn Brea). Crystals up to 150 mm (6 in) across, come from Tsumeb, Namibia. Crystals of binnite in Lengenbach, Binntal, Switzerland, up to 30 mm ($1^{3}/_{16}$ in). Crystals up to 20 mm ($^{25}/_{32}$ in) across were found recently in Dzhezkazgan, Kazakhstan. Large masses occurred in Kipushi, Zaire. Crystals up to 25 mm (1 in) known from El Cobre, Zacatecas, Mexico. *Application:* Cu ore.

Tennantite, 26 mm, El Cobre, Mexico

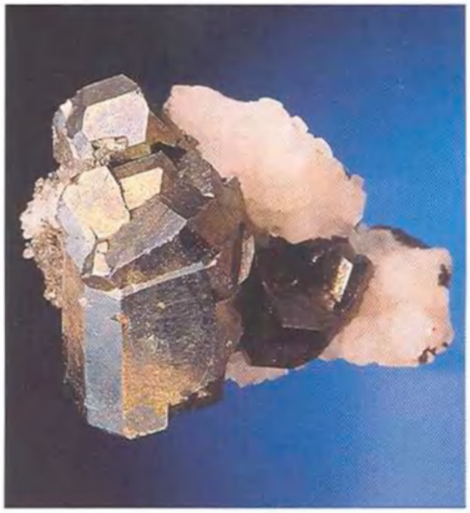

Tetrahedrite

$Cu_{12}Sb_4S_{13}$

CUBIC ● ● ● ●

Properties: C – steel-gray to black; S – black, brown; L – metallic; D – opaque; DE – 5.0; H – 3-4,5; CL – none; F – conchoidal; M – tetrahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in low- to medium temperature veins; in contact metamorphic deposits together with chalcopyrite, galena, sphalerite, pyrite, bornite, calcite, quartz and other minerals. The largest known crystals up to 25 cm ($9^{7}/_{8}$ in) across come from Anzen and Irazein in Pyrennees, France. Common crystals several cm in size occur in Cavnic, Romania. Fine specimens with crystals up to 70 mm (2¾ in) across found in the Mercedes mine in Huallanca, Peru. Other Peruvian localities like Casapalca and Morococha yielded fine crystallized specimens. Fine crystals up to 20 mm ($^{25}/_{32}$ in) in size known from Pøibram, Czech Republic.
Application: ruda Cu.

Tetrahedrite, 10 mm xx, Peru

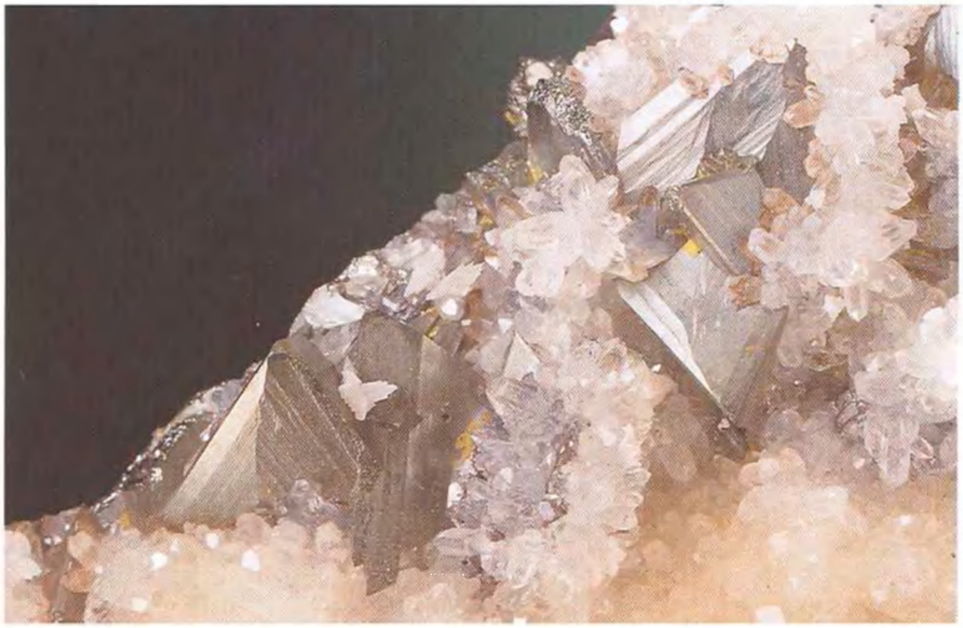

Freibergite, 15 mm xx, Obecnice, Czech Republic

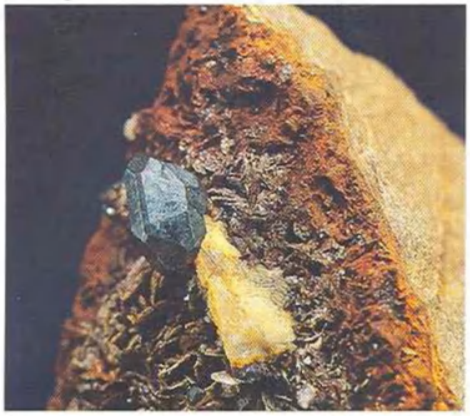

Freibergite

$(Ag,Cu,Fe)_{12}(Sb,As)_4S_{13}$

CUBIC ● ● ●

Properties: C – gray to black; S – black, brown to dark red; L – metallic; D – opaque; DE – 5,4; H – 3-4,5; CL – none; F – uneven to conchoidal; M – tetrahedral crystals, massive.
Origin and occurrence: Hydrothermal in ore deposits, associated with many sulfides and sulfosalts. It is mainly massive in the Himmelsfurst mine in Freiberg, Germany; Cobalt, Ontario, Canada and in Mount Isa, Queensland, Australia. It was a principal Ag ore in Kutná Hora, Czech Republic.
Application: Ag ore.

Wurtzite

ZnS

HEXAGONAL ● ● ●

Properties: C – dark red-brown, dark brown to brown-black; S – brown; L – resinous to submetallic; D – translucent to opaque; DE – 4,1; H – 3,5-4; CL – good; M – pyramidal, prismatic to thick tabular, striated crystals, concentric banded and radial aggregates.
Origin and occurrence: Hydrothermal in ore veins with sphalerite, marcasite, pyrite and other minerals. Also of low-temperature origin along the cracks of clay concretions. The best wurtzite crystals come from Bolivia; crystals up to 40 mm ($1\frac{9}{16}$ in) across from Animas and crystals up to 20 mm ($\frac{25}{32}$ in) across from Llallagua and Potosi. Fine crystals up to 30 mm ($1\frac{3}{16}$ in) across found in Talnakh near Norilsk, Siberia, Russia. Interesting radial aggregates, up to several cm in diameter, occurred in Pøibram, Czech Republic.

Wurtzite, 30 mm xx, Animas, Bolivia

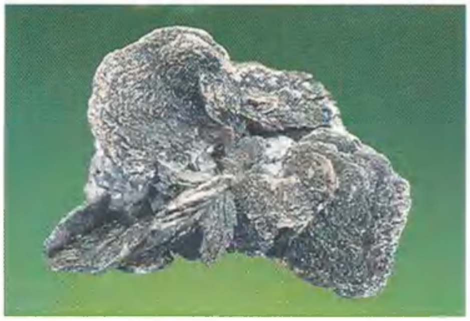

Greenockite

CdS

HEXAGONAL ● ● ●

Properties: C – yellow to orange, dark red; S – orange-yellow to brick-red; L – adamantine to resinous; D – opaque to translucent, DE – 4,8; H – 3-3,5; CL – good; F – conchoidal; M – trillings, pulverulent coatings.
Origin and occurrence: Mainly secondary as pulverulent coatings on sphalerites. It also occurs in cavities of volcanic rocks together with prehnite, zeolites and calcite. Crystals are known from ore veins in Llallagua, Bolivia. Pulverulent coatings are described from Pøibram, Czech Republic; Bleiberg, Austria and from the deposits in the Tri State region, Missouri, USA. Crystals up to 10 mm ($\frac{3}{8}$ in) occurred in the cavities of volcanic rocks near Renfrew, Scotland, UK.

Enargite

Cu_3AsS_4

ORTHORHOMBIC ● ● ● ●

Properties: C – gray-black to black; S – gray-black;

Grenockite, 20 mm, Ocna de Fier, Romania

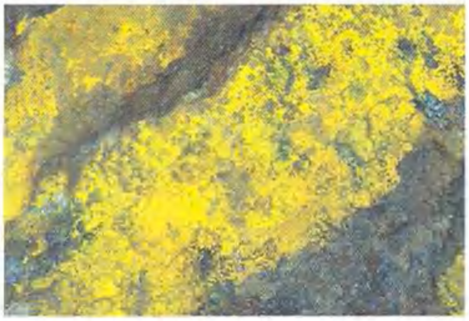

L – metallic; D – opaque; DE – 4,4; H – 3; CL – perfect; F – uneven; M – tabular and prismatic, striated crystals, massive.
Origin and occurrence: Hydrothermal in medium-temperature, sometimes in low-temperature deposits, associated with quartz, pyrite, sphalerite, galena, bornite and other minerals. Beautiful crystals, up to 100 mm (4 in) across, come from the Luz Angelica mine in Quiruvilca; crystals up to 150 mm (6 in) across from Morococha and Cerro de Pasco, Peru. Fine crystals were found in Butte, Montana, USA and in Mancayano, Luzon, Philippines. It occurs as a principal Cu ore in several deposits (Bor, Serbia; Huaron, Peru).
Application: Cu ore.

Enargite, 41 mm, Butte, U.S.A.

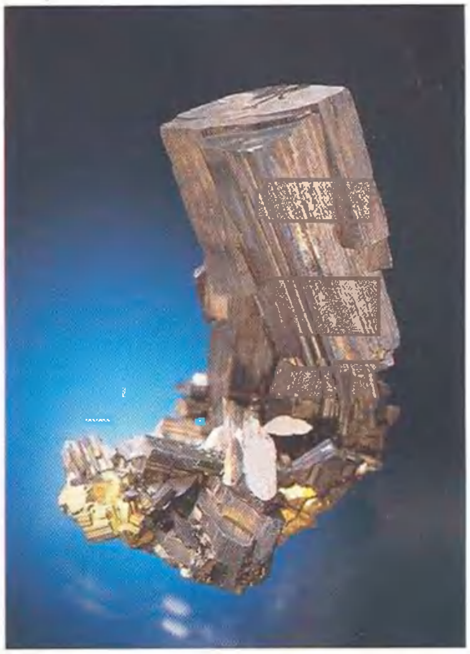

Cubanite

Cu_2FeS_3

ORTHORHOMBIC ● ● ●

Properties: C – brass-yellow to bronze; S – black; L – metallic; D – opaque; DE – 4,1; H – 3,5; CL – none; F – conchoidal; M – tabular crystals, massive.
Origin and occurrence: Magmatic in liquid deposits as inclusions in chalcopyrite, as hydrothermal in high-temperature deposits, associated with chalcopyrite, pyrite, pyrrhotite and sphalerite. The best crystals, twins, up to 40 mm (1$^{9}/_{16}$ in) across, come from the Henderson No.2 mine, Chibougamau, Quebec, Canada. It was also described in crystals from Sudbury, Ontario, Canada and Morro Velho, Brazil. Its large accumulations are important Cu ore in Sudbury, Ontario, Canada; in Mooihoek, South Africa and in Prince William Sound, Alaska, USA.
Application: important Cu ore.

Cubanite, 33 mm x, Chibougamau, Canada

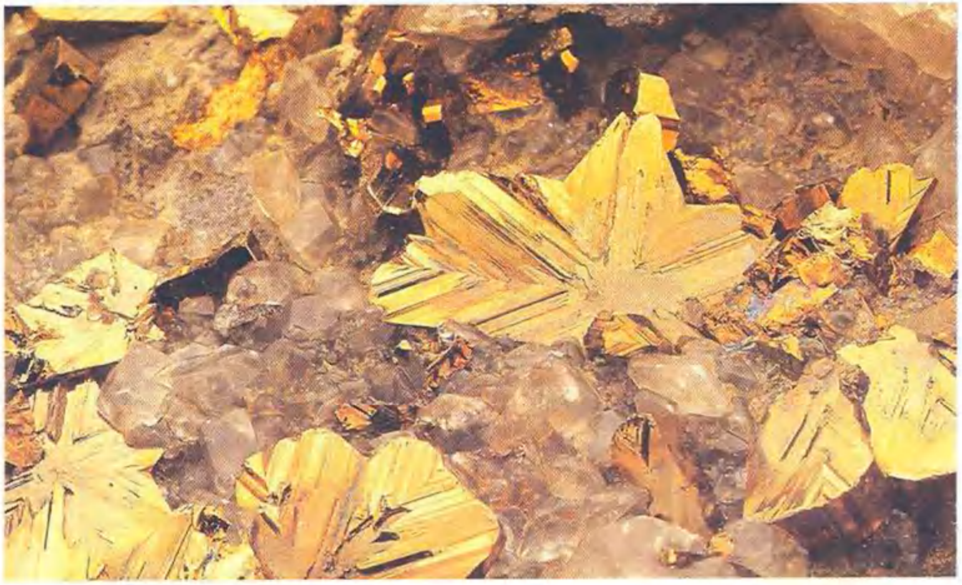

Sternbergite, 20 mm, St.Andreasberg, Germany

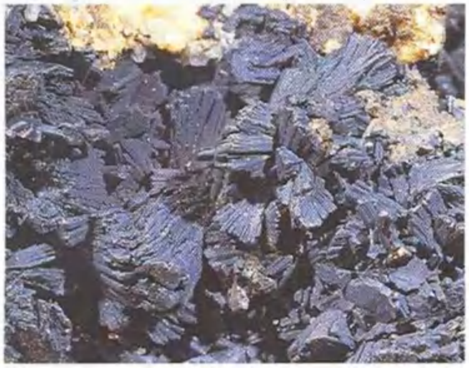

Sternbergite

Ag_2FeS_3

ORTHORHOMBIC ● ●

Properties: C – golden-brown; S – black; L – metallic to adamantine; D – opaque; DE – 4,3; H – 1-1,5; CL – perfect; M – thin tabular pseudo-hexagonal crystals, often in rosettes and fan-shaped aggregates. *Origin and occurrence:* Hydrothermal in Ag-bearing veins, associated with stephanite, acanthite, proustite, argentopyrite and other minerals. Tabular crystals up to several mm across known from Jachymov and Midinec, Czech Republic; from St. Andreasberg, Johanngeorgenstadt, Schneeberg and Freiberg, Germany.

Pyrrhotite, 68 mm, Dalnegorsk, Russia

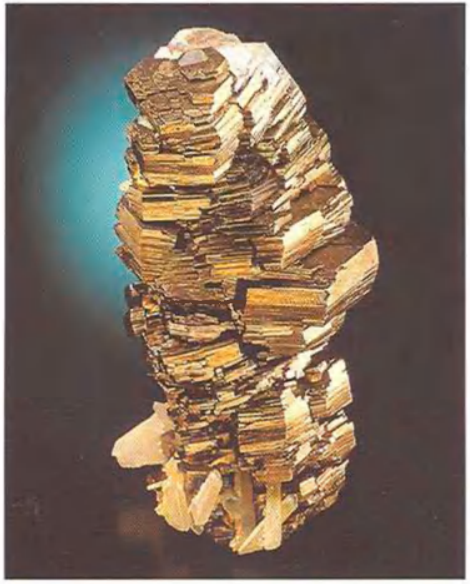

Argentopyrite, 60 mm, M?d?nec, Czech Republic

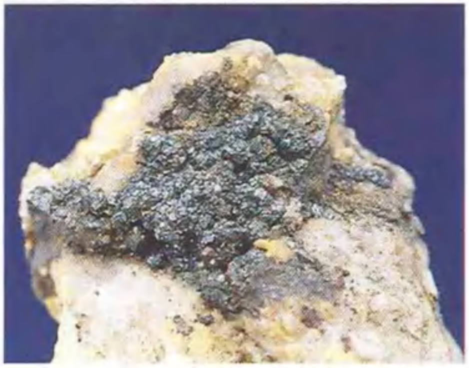

Argentopyrite

Ag_2FeS_3

ORTHORHOMBIC ● ●

Properties: C – gray-white, tarnishes iridescent; S – gray; L – metallic; D – opaque; DE – 4,3; H – 3,5 – 4; CL – none; F – uneven; M – thick tabular pseudo-hexagonal crystals, granular. *Origin and occurrence:* Hydrothermal in ore veins, associated with proustite, pyrargyrite, stephanite, sternbergite, dolomite, quartz and other minerals. Crystals up to 5 mm ($\frac{3}{16}$ in) across come from Jachymov and Midinec, Czech Republic; it is also known from Schlema, Freiberg and Schneeberg, Germany and from Colquechaca, Bolivia.

Pyrrhotite

$Fe_{1-x}S$ ($x = 0\text{-}0{,}17$)

MONOCLINIC ● ● ● ●

Properties: C – bronze-yellow to brown, tarnishes quickly; S – dark gray-black; L – metallic; D – opaque; DE – 4,7; H – 3,5-4,5; CL – none; F – uneven to conchoidal; M – tabular, pyramidal and prismatic crystals, massive.
Origin and occurrence: Magmatic liquid in basic rocks, together with pyrite and pentlandite; in pegmatites; hydrothermal in high-temperature and metasomatic deposits; sedimentary and metamorphic. Tabular crystals up to 300 mm ($11\frac{13}{16}$ in) across come from Trepea, Serbia and Dalnegorsk, Russia; prismatic crystals up to 150 mm (6 in) long found in Santa Eulalia, Chihuahua, Mexico and Chiuzbaia, Romania. Tabular crystals up to 110 mm ($4\frac{5}{16}$ in) occurred in Cavnic, Romania. Large imperfect crystals, coated with wavellite, are known from Llallagua, Bolivia. Huge masses of industrial importance occur in Sudbury, Ontario, Canada; Talnakh near Norilsk, Siberia, Russia and elsewhere. *Application:* sometimes as Fe ore.

Nickeline, 138 mm, Pohla, Germany

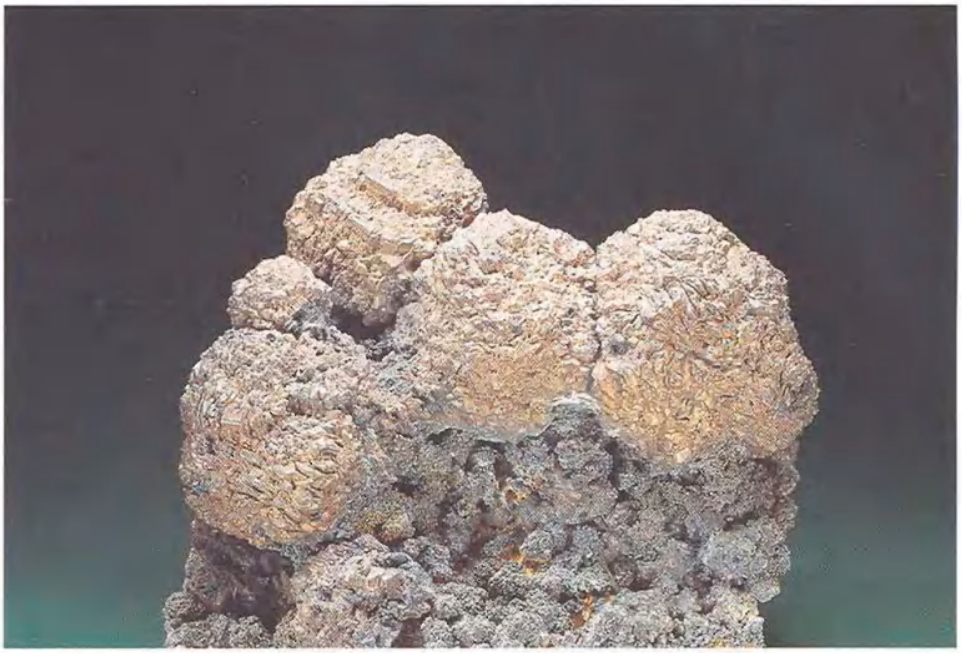

Nickeline

NiAs

HEXAGONAL ● ● ● ●

Properties: C – light copper-red, tarnishes gray to black; S – light brown-black; L – metallic; D – opaque; DE – 7,8; H – 5-5,5; CL – none; F – conchoidal; M – striated crystals, botryoidal and dendritic aggregates, granular, massive.
Origin and occurrence: Hydrothermal in high-temperature ore veins with skutterudite, nickel-skutterudite, safflorite, rammelsbergite and other minerals. Also magmatic in norites and peridotites. Crystals up to 15 mm ($^{19}/_{32}$ in) across come from Pohla; small crystals were found in Richelsdorf, Germany. Fine massive aggregates and small crystals occurred in Bou Azzer, Morocco. Huge accumulations are known in Cobalt and Gowganda, Ontario, Canada.
Application: Ni ore.

Breithauptite

NiSb

HEXAGONAL ● ●

Properties: C – light copper-red, purplish; S – red-brown; L – metallic; D – opaque; DE – 8,6; H – 5,5; CL – none; F – conchoidal to uneven; M – thin tabular crystals, dendritic aggregates, massive.
Origin and occurrence: Magmatic, hydrothermal and metamorphic, associated with silver, nickeline, cobaltite and other sulfides. It occurs inter-grown with pyrrhotite and pentlandite in magmatic liquid deposit Vlakfontein, South Africa. It is common in hydrothermal veins in Cobalt, Ontario, Canada; also known from Sarrabus, Sardinia, Italy and St. Andreasberg, Germany.

Breithauptite, 80 mm, Cobalt, Canada

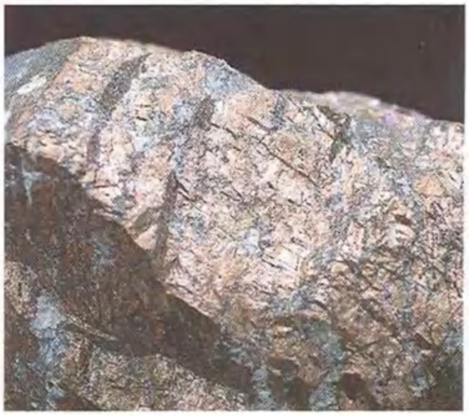

Millerite, 26 mm, Antwerp, U.S.A.

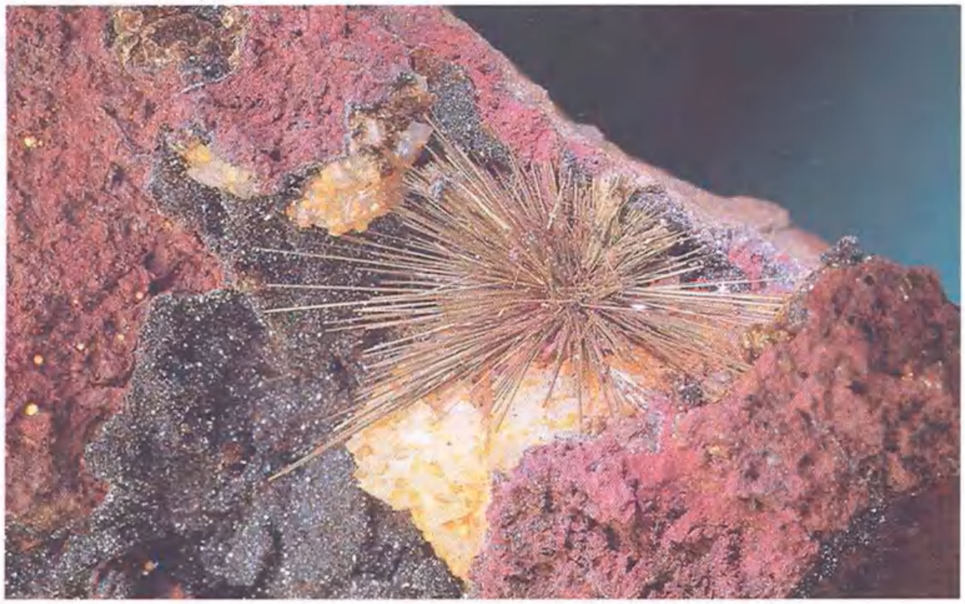

Millerite

NiS

TRIGONAL ● ● ●

Properties: C – light brass-yellow to bronze, tarnishes iridescent; S – greenish-black; L – metallic; D – opaque; DE – 5,4; H – 3-3,5; CL – perfect; F – uneven; M – acicular crystals, cleavable masses, aggregates of parallel inter-grown crystals with velvety surface.
Origin and occurrence: Hydrothermal in low-temperature ore deposits, also as a product of decomposition of Ni sulfides. It is associated with pyrrhotite, ankerite, whewellite, barite and other minerals. Cleavable masses occurred near Temagami near Sudbury, Ontario, Canada and in Kambalda, Western Australia, Australia. Acicular crystals come from ore veins in Jachymov and Príbram, Czech Republic; needles in cavities of siderite concretions in the vicinity of Kladno, Czech Republic, reached up to 70 mm (2¾ in). Acicular crystals up to 80 mm (3⅛ in) long known from the limestone cavities near Dortmund, Germany. Beautiful specimens with crystals up to 50 mm (2 in) long in the cavities in hematite occurred in the Sterling mine, Antwerp, New York, USA. Velvety fibrous aggregates were found in the Thompson mine, Manitoba, Canada. *Application:* Ni ore.

Alabandite, 40 mm, Sacarimb, Romania

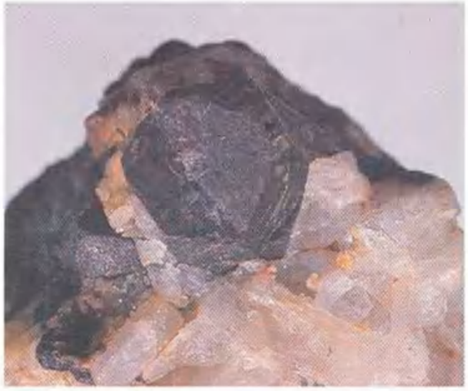

Alabandine

MnS

CUBIC ● ●

Properties: C – black; S – green; L – submetallic; D – opaque; DE – 4,1; H – 3.5-4; CL – perfect; F – uneven; M – cubic and octahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with rhodochrosite, calcite, galena, sphalerite, pyrite and other minerals. Crystals up to 20 mm ($^{25}/_{32}$ in) across together with granular aggregates are relatively common in Romania (Sacarîmb, Baia de Arieş, Roşia Montana). Crystals are also known from the Queen of the West mine, Colorado and from the Lucky Cuss mine, Tombstone, Arizona, USA.

Galena, 56 mm, Neudorf, Germany

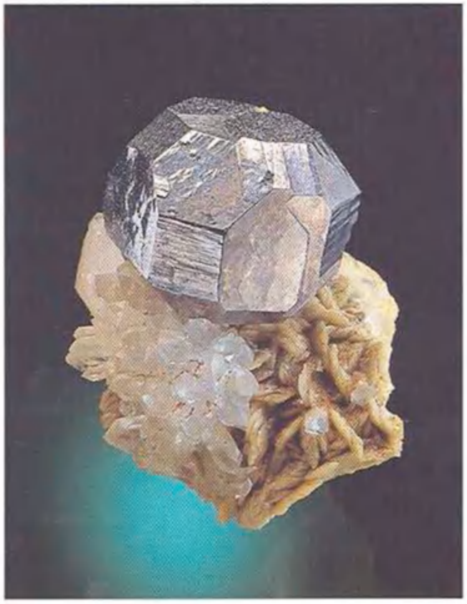

Galena, 66 mm, Picher, U.S.A.

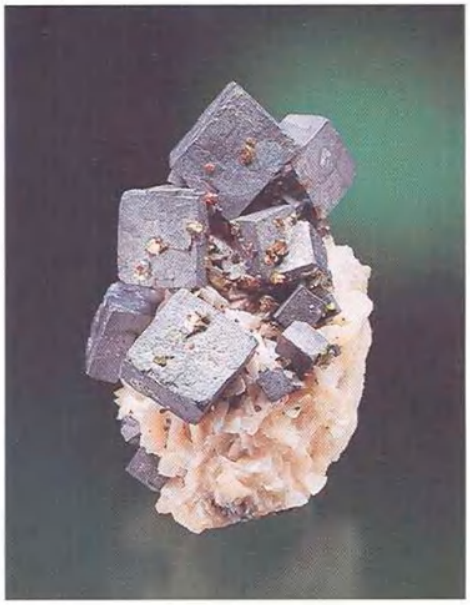

Galena

PbS

CUBIC ● ● ● ● ●

Varieties: steinmannite

Properties: C – lead-gray; S – lead-gray; L – metallic; D – opaque; DE – 7,6; H – 2.5; CL – perfect; F – conchoidal; M – cubic crystals and their complex combinations, tabular crystals, skeletal aggregates, massive.
Origin and occurrence: Magmatic, hydrothermal, metamorphic, very rare sedimentary, associated with sphalerite, chalcopyrite, pyrite, quartz and other minerals. Large crystals up to several tens of cm in size come from many localities in the USA. (Joplin, Missouri; Galena, Kansas; Picher, Oklahoma; Sweetwater mine, Missouri). Beautiful, often skeletal crystals or spinel-law twins up to 200 mm (7⅞ in) across known from the Nikolai mine in Dalnegorsk, Russia. Fine crystals occur also in Naica, Chihuahua, Mexico. Famous complicated combinations of crystals were found in Neudorf, Germany. Octahedral crystals up to 10 mm (⅜ in) (steinmannite variety) were common in Pribram, Czech Republic. Beautiful specimens with cubes up to several cm across come from Madan, Bulgaria; spinel-law twins occurred in Herja, Romania.
Application: the most important Pb ore.

Clausthalite, 20 mm veinlet, Tilkerode, Germany

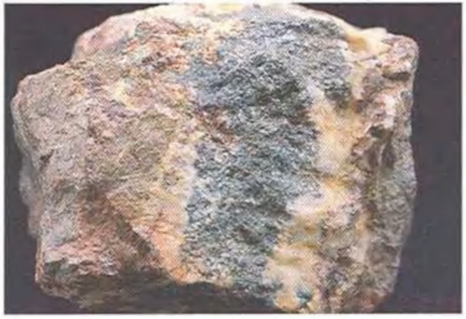

Clausthalite

PbSe

CUBIC ● ● ●

Properties: C – lead-gray, bluish; S – gray-black; L – metallic; D – opaque; DE – 8,3; H – 2,5 – 3; CL – good; F – granular; M – granular, massive..
Origin and occurrence: Hydrothermal in ore veins with a low S content, together with berzelianite, umangite, uraninite and other minerals. It occurs as massive aggregates in calcite veins with other selenides in Clausthal and Tilkerode, Germany. It is similar in Skrikerum, Sweden; common in the U deposits in Predborice, Bukov and Zlatkov, Czech Republic.

Altaite, 65 mm, Stania, Romania

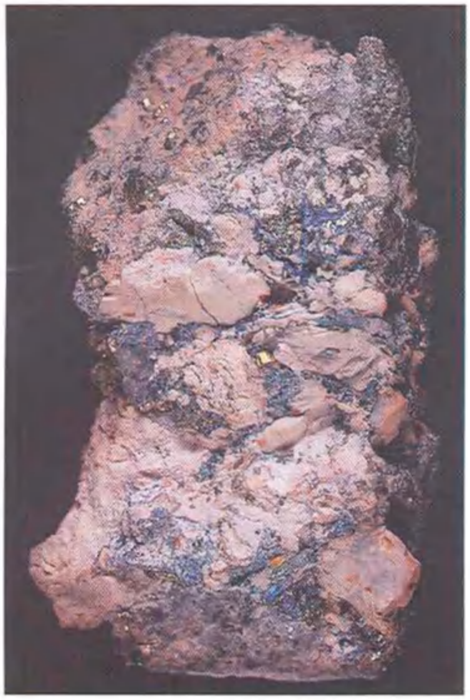

Altaite

PbTe

CUBIC ● ● ●

Properties: C – tin-white to yellowish, tarnishing to bronze; S – black; L – metallic; D – opaque; DE – 8,3; H – 3; CL – perfect; F – conchoidal; M – cubic and octahedral crystals, massive.

Origin and occurrence: Hydrothermal in vein Au deposits, associated with other tellurides, galena and other minerals. Crystals up to 20 mm (25/32 in) across are known from the Revenge mine, Colorado, USA. It is massive with aguilarite in Kalgoorlie, Western Australia, Australia. It is relatively common with other tellurides in Sacarîmb , Romania; in the Zavodinskii mine, Altai, Russia and in Zod near Sevan Lake, Armenia.

Miargyrite

$AgSbS_2$

MONOCLINIC ● ● ●

Properties: : C – black to steel-gray; S – cherry-red; L – metallic, adamantine; D – opaque; DE – 5,3; H – 2,5; CL – imperfect; F – conchoidal; M – thick tabular, striated crystals, massive.

Origin and occurrence: Hydrothermal in low-temperature ore veins, together with proustite, pyrargyrite, polybasite, silver, quartz and other minerals. Fine crystals, up to 10 mm (3/8 in) across, occurred in Príbram and Kutna Hora, Czech Republic; in Braunsdorf and Freiberg, Germany. It is also known from many localities in Bolivia (Tatasi, Oruro – 10 mm (3/8 in) crystals, Potosí) and Mexico (Sombrerete, Catorce). It was also found in Hiendelaencina, Guadalajara, Spain.

Application: Ag ore.

Franckeite

$(Pb,Sn)_6FeSn_2Sb_2S_{14}$

TRICLINIC ● ●

Properties: C – gray-black; S – gray-black; L – metallic; D – opaque; DE – 5,9; H – 2,5-3; CL –

Miargyrite, 20 mm, Kutna Hora, Czech Republic

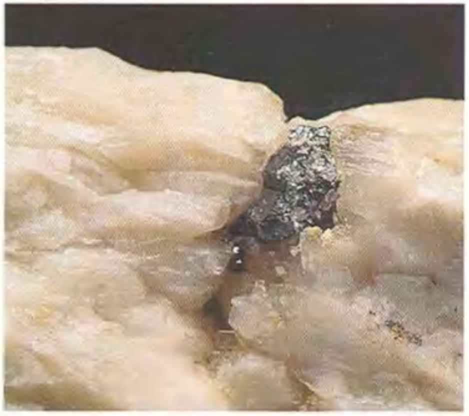

Franckeite, 30 mm, Oruro, Bolivia

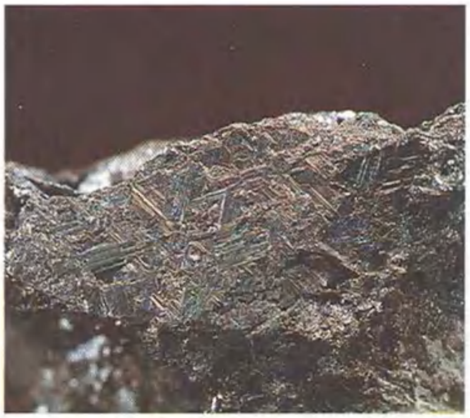

Cylindrite, 50 mm, Poopo, Bolivia

perfect; M – thin tabular, curved crystals, often rosette-like aggregates of crystals.
Origin and occurrence: Hydrothermal in Ag-Sn deposits, associated with cylindrite, zinkenite, cassiterite, wurtzite and other minerals; in contact metamorphic limestones. The best crystals up to 60 mm (24 in) across come from various localities in Bolivia (Mollecagua near Poopo; Huanuni; Chocoya near Potosi; Colquechaca and Oruro). It is known in contact metamorphic limestone from the Kalkar quarry, California, USA.
Application: Sn ore.

Cylindrite

$Pb_3Sn_4FeSb_2S_{14}$

TRICLINIC ● ●

Properties: C – lead-gray to black; S – black; L – metallic; D – opaque; DE – 5,4; H – 2,5; CL – perfect; M – cylindrical, conical and spherical crystal aggregates, massive.
Origin and occurrence: Hydrothermal in Sn-bearing veins with franckeite, stannite, cassiterite, galena and other minerals. World famous specimens with crystals up to 50 mm (2 in) long come from the Trinacria and Santa Cruz mines near Poopo, Bolivia. It was also reported from the Smirnovsk deposit, Transbaikalia, Russia.

Cinnabar

HgS

TRIGONAL ● ● ● ●

Properties: C – red to brownish-red; S – crimson; L – adamantine to metallic, also dull; D – opaque; DE – 8,2; H – 2-2,5; CL – perfect; F – conchoidal to uneven; M – rhombohedral, thick tabular and prismatic crystals, massive.
Origin and occurrence: Low-temperature hydrothermal mineral, associated with realgar, mercury, pyrite, marcasite and other minerals. The world s best specimens with crystals up to 70 mm (2¾ in) across are known from many localities in China (Hunan and Guizhou provinces). Fine shiny crystals up to 10 mm (3/8 in) across come from Nikitovka, Ukraine and in Khaidarkan, Kyrgyzstan. Crystals up to 30 mm (1 3/16 in) also found in Monte Amiata and Rippa near Seravezza, Italy. Its massive aggregates are an important Hg ore in Allchar, Macedonia; Almaden, Spain and elsewhere.
Application: the most important Hg ore.

Cinnabar, 30 mm x, Hunan, China

Covellite, 65 mm. Butte, U.S.A.

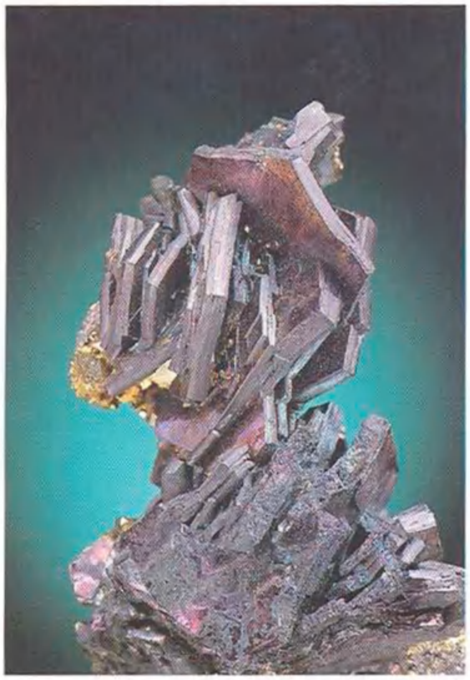

Covellite

CuS

HEXAGONAL ● ● ●

Properties: C – indigo-blue, tarnishing iridescent; S – lead-gray; L – submetallic to resinous; D – opaque; DE – 4,6; H – 1,5-2; CL – perfect; F – uneven; M – hexagonal tabular crystals, massive.
Origin and occurrence: Rare hydrothermal; mainly secondary in the oxidation zone of ore deposits, associated with chalcopyrite, chalcocite, djurleite, bornite and other minerals. Tabular crystals up to 30 mm (1 1/16 in) across are known from Butte, Montana and Summitville, Colorado, USA; also from Sarrabus, Sardinia, Italy. Massive aggregates are common in Bor, Serbia; Bisbee, Arizona, USA and elsewhere.
Application: Cu ore.

Linneite

$Co^{2+}Co^{3+}{}_2S_4$

CUBIC ● ● ●

Properties: C – light gray to steel-gray; S – black-gray; L – metallic; D – opaque; DE – 4,9; H – 4,5-5,5; CL – imperfect; F – uneven to conchoidal; M – octahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in ore veins and metamorphic deposits, together with chalcopyrite, pyrrhotite, millerite, bismuthinite and sphalerite. The best crystals up to 30 mm (1 1/16 in) across come from the Kilembe mine, Uganda. Crystals are also known from Musonoi, Shaba, Zaire; Mhsen, Germany; siderite concretions in Kladno, Czech Republic and from the Bastnas mine near Riddarhyttan, Sweden.

Carrolite

$Cu(Co,Ni)_2S_4$

CUBIC ● ● ●

Properties: C – light to steel-gray, tarnishing red-purple; S – gray; L – metallic; D – opaque; DE – 4,8; H – 4,5-5,5; CL – imperfect; F – conchoidal to uneven; M – octahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with linneite, chalcopyrite and other minerals. It is a principal Co ore in deposits in Zaire. Beautiful crystals up to 20 mm (25/32 in) across are known from the M'Sesa mine near Kambove, Kolwezi and from Kamoto, Shaba.
Application: Co ore.

Linneite, 60 mm, Ruwenzori, Uganda

Carrollite, 8 mm x, Kambove, Zair

Stibnite

Sb_2S_3

ORTHORHOMBIC ● ● ● ●

Properties: C – steel-gray, tarnishing iridescent or black; S – lead-gray; L – metallic; D – opaque; DE – 4,6; H – 2; CL – perfect; F – conchoidal to uneven; M – thick to thin prismatic crystals, thin needles, massive.
Origin and occurrence: Hydrothermal in medium- and low-temperature ore veins with quartz and gold, the other associated minerals are rare (arsenopyrite, berthierite, gudmundite, antimony). The largest known stibnite crystals occurred in the in the Ichinokawa mine, Shikoku, Japan, where they were up to 60 cm (24 in) long. Similar sized crystals were found recently in several localities in China (the Xikuangshan mine, Hunan). Crystals up to 200 mm (7⅞ in) across come from Manhattan, Nevada, USA. Beautiful crystals are known from Romania; long prismatic ones with barite crystals are prevalent in Baia Sprie, clusters of thin needles come from Herja and thick prismatic crystals from Baiuţ. Perfect druses of crystals up to 150 mm (6 in) long, associated with purple fluorite crystals, barite and calcite, occurred in Kadamdzhai, Kyrgyzstan. Beautiful druses of stibnite crystals with quartz come also from Kremnica, Slovakia. Also Kostajnik, Serbia and La Lucette, France yielded fine crystals in the past.
Application: Sb ore.

Stibnite, 187 mm, Ichinokawa, Japan

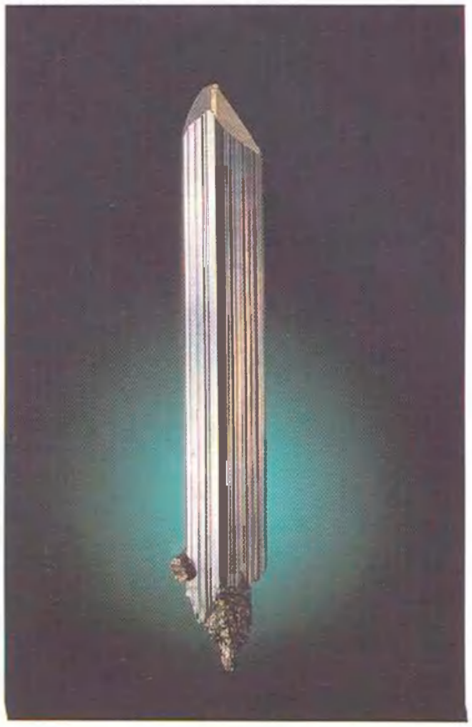

Stibnite, 120 mm, Baia Sprie, Romania

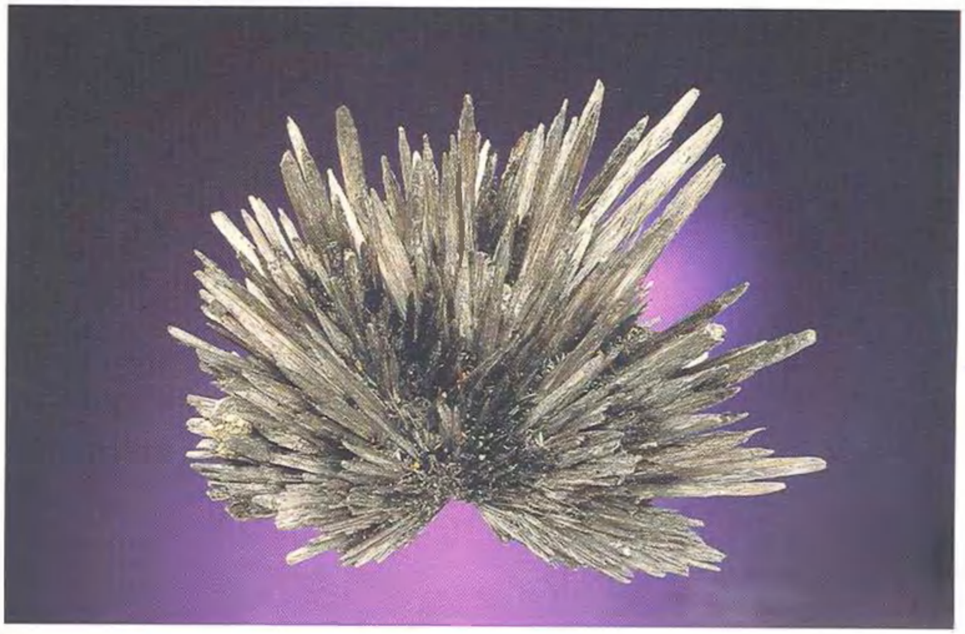

+
←

Bismuthinite, 160 mm, Tasna, Bolivia

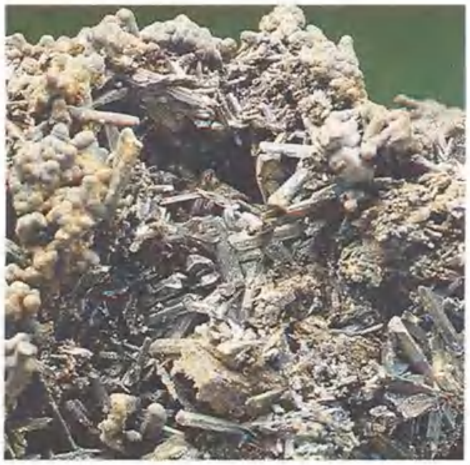

Bismuthinite

Bi_2S_3

ORTHORHOMBIC ● ● ● ●

Properties: C – lead-gray to tin-white, tarnishing yellow and iridescent; S – lead-gray; L – metallic; D – opaque; DE – 6,8; H – 2-2,5; CL – perfect; F – uneven; M – thick prismatic to acicular, striated crystals, fibrous aggregates, massive.
Origin and occurrence: Hydrothermal in ore deposits; also in recent volcanic exhalation deposits, associated with bismuth, arsenopyrite, stannite, galena and other minerals. The world's best crystals over 50 mm (2 in) long, come from Bolivia (Tasna; Huanuni; Llallagua). Fine crystals were also found in Redruth, Cornwall, UK. Rich finds were made in Biggenden, Queensland, Australia. Interesting crystals are known from Spind, Norway.
Application: Bi ore.

Kermesite, 20 mm xx, Freiberg, Germany

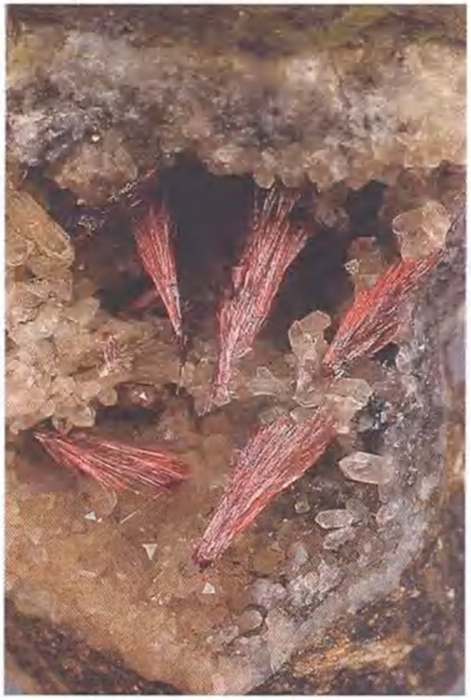

Tetradymite, 80 mm, Ciclova, Romania

Kermesite

Sb_2S_2O

TRICLINIC ● ● ●

Properties: C – cherry-red; S – brown-red; L – adamantine to submetallic; D – translucent to opaque; DE – 4,7; H – 1-1,5; CL – perfect; M – acicular crystals, radial aggregates.
Origin and occurrence: Secondary, as a result of a stibnite oxidation in Sb deposits, associated with stibnite, antimony, senarmontite, valentinite and stibiconite. Famous specimens with needles up to 100 mm (4 in) long, in radial aggregates, come from Pezinok and Pernek, Slovakia. Crystals up to 50 mm (2 in) long found in the Globe and Phoenix mines, Zimbabwe. It is also known from Bolivia (San Francisco mine, Poopo; Oruro) and Braunsdorf, Germany.

Nagyagite, 100 mm, Sacarimb, Romania

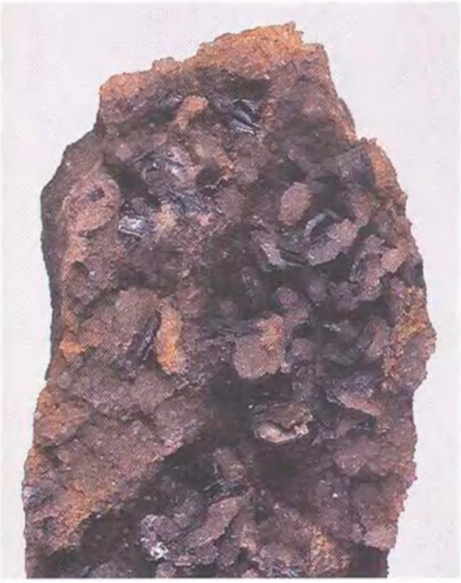

Tetradymite

Bi_2Te_2S

TRIGONAL ● ● ●

Properties: C – light steel-gray, tarnishing dark to iridescent; S – light steel-gray; L – metallic; D – opaque; DE – 7,3; H – 1,5-2; CL – perfect; M – pyramidal crystals, curved bladed aggregates, massive.
Origin and occurrence: Hydrothermal in medium- and high-temperature Au deposits; also in contact metamorphosed deposits, together with gold, hessite, calaverite, pyrite and other minerals. Crystals up to 10 mm (3/8 in) across are known from Župkov, Slovakia. It is common in Colorado (Red Cloud) and in California (Carson Hill), USA. It occurs with gold in skarns in Baiţa Bihorului, Romania.

Nagyagite

$Pb_5Au(Te,Sb)_4S_{5-8}$

TETRAGONAL ● ●

Properties: C – black-gray; S – black-gray; L – metallic; D – opaque; DE – 7,5; H – 1,5; CL – perfect; F – uneven; M – thin tabular to foliated crystals, granular, massive.
Origin and occurrence: Hydrothermal in low-temperature veins, associated with altaite, arsenic, gold, rodochrosite and other minerals. Foliated crystals up to 40 mm (1 9/16 in) across come from Sacarîmb and Baia de Arieş, Romania. Fine specimens were also found in Tavua, Viti Levu, Fiji and in the Sylvia mine, Tararu Creek, New Zealand. It also occurred in Gold Hill and Cripple Creek, Colorado, USA and in Kalgoorlie, Western Australia, Australia.

Sylvanite

$(Au,Ag)_2Te_4$

MONOCLINIC ● ● ●

Properties: C – steel-gray to silver-white, tarnishing to yellow; S – steel-gray to silver- white; L – metallic; D – opaque; DE – 8,2; H – 1,5-2; CL – perfect; F – uneven; M – short prismatic to thick tabular crystals, skeletal aggregates, granular.
Origin and occurrence: Hydrothermal in low-temperature ore veins, also in medium- and high-temperature deposits as one of the latest minerals, associated with gold, calaverite, hessite, krennerite and other minerals. The best specimens come from Sacarîmb and Baia de Arieş, Romania, where it occurs as skeletal crystals and aggregates. Crystals were also found in Cripple Creek, Colorado, USA; crystals up to 10 mm (3/8 in) across come from the Emperor mine, Viti Levu, Fiji.

Sylvanite, 20 mm, Cripple Creek, U.S.A.

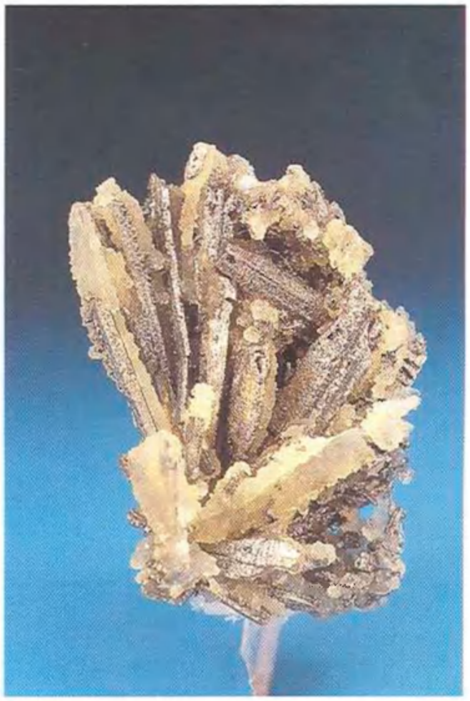

Krennerite, 3 mm xx, Sacarimb, Romania

Krennerite

$AuTe_2$

ORTHORHOMBIC ● ●

Properties: C – silver-white to brass-yellow; S – silver-white; L – metallic; D – opaque; DE – 8,9; H – 2-3; CL – perfect; F – conchoidal to uneven; M – short prismatic, striated crystals, massive.
Origin and occurrence: Hydrothermal in gold-bearing veins, associated with gold, tellurium, pyrite, quartz and other tellurides. Crystals up to 20 mm ($\frac{25}{32}$ in) across found in Tavua, Viti Levu, Fiji. Small crystals occurred in Sacarîmb and Baia de Arieş , Romania and in Cripple Creek, Colorado, USA.
Application: Au ore.

Calaverite

$AuTe_2$

MONOCLINIC ● ● ●

Properties: C – brass-yellow to silver-white; S – greenish; L – metallic; D – opaque; DE – 9,3; H – 2,5-3; CL – none; F – uneven to conchoidal; M – bladed and short prismatic, striated crystals, granular, massive.
Origin and occurrence: Hydrothermal in low-temperature Au-bearing veins, sometimes in medium- and high-temperature deposits, associated with coloradoite, altaite, krennerite and other tellurides. It is common in the Mother Lode in California (Carson Hill). It is important, together with hessite in Cripple Creek, crystals up to 10 mm ($\frac{3}{8}$ in) across come from the Cresson mine, Colorado, USA.. It is also known from several mines near Kirkland Lake, Ontario, Canada.
Application: Au ore.

Pyrite

FeS_2

CUBIC ● ● ● ● ●

Properties: C – light brass-yellow, tarnishing iridescent and darkens; S – green-black to brown-black; L – metallic; D – opaque; DE – 5,0; H – 6-6,5; CL – imperfect; F – conchoidal to uneven; M – combinations of cubic crystals, striated, stalactitic and spherical aggregates, massive.
Origin and occurrence: Magmatic segregations in basic rocks with pyrrhotite and pentlandite, in pegmatites and skarns; hydrothermal in porphyry and vein deposits together with other sulfides; hydrothermal sedimentary, sedimentary and metamorphic. Magmatic segregations are known from Sudbury, Ontario, Canada and Merensky Reef, Transvaal, South Africa. Large crystals up to 200 mm ($7\frac{7}{8}$ in) across are known from Rio Marina, Elba, Italy. Fine octahedra come from Llallagua, Bolivia. Crystals, up to 120 mm ($4\frac{11}{16}$in) across, were found in Bingham and Park City, Utah, USA. Crystals up to 150 mm (6 in) are known from Huanzala and Quiruvilca, Peru. The largest pyrite deposit is Rio Tinto, Spain, where fine-grained pyrite formed accumulations about 1 billion tons. Beautiful cubes up to 80 mm ($3\frac{1}{8}$ in) come from Navajún, Spain. Large deformed crystals

Calaverite, 21 mm, Nagyag, Romania

Pyrite, 74 mm, Navajun, Spain

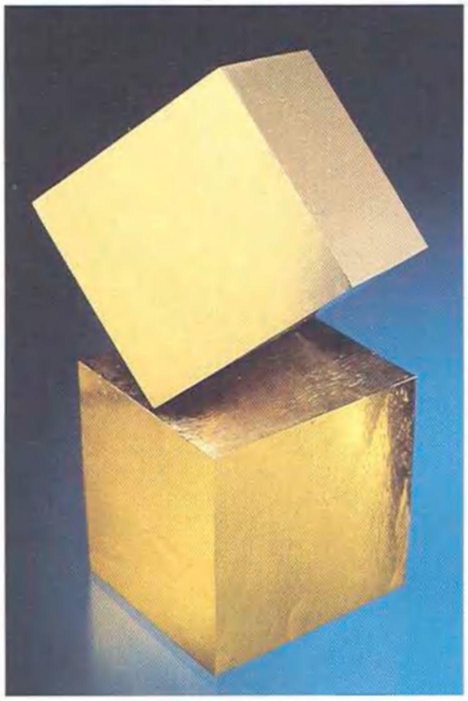

Pyrite, 135 mm, Washington, U.S.A.

up to 200 mm ($7\frac{7}{8}$ in) across occurred in the Samo mine near Hnušt'a , Slovakia. Interesting, complex combinations of crystals found in Nanisivik, on an edge from Climax, Colorado, USA.
Application: production of sulfuric acid.

Hauerite

MnS_2

CUBIC ● ● ●

Properties: C – red-brown to brown-black; S – brown-red; L – metallic to adamantine; D – opaque; DE – 3,44; H – 4; CL – perfect; F – uneven to conchoidal; M – octahedral crystals and their combinations, spherical aggregates..
Origin and occurrence: Low-temperature sedimentary mineral, limited to clays with high S contents. The best crystals up to 50 mm (2 in) across come from the Destricella mine near Raddusa, Sicily, Italy. Crystals up to 25 mm (1 in) and their aggregates are known from Víglašska Huta (former Kalinka) near Zvolen, Slovakia. Crystals up to 15 mm ($\frac{19}{32}$ in) across occur in Tarnobrzeg, Poland. Also found with sulfur, gypsum, realgar and calcite in the salt domes near High Island, Texas, USA.

Hauerite, 39 mm, Raddusa, Italy

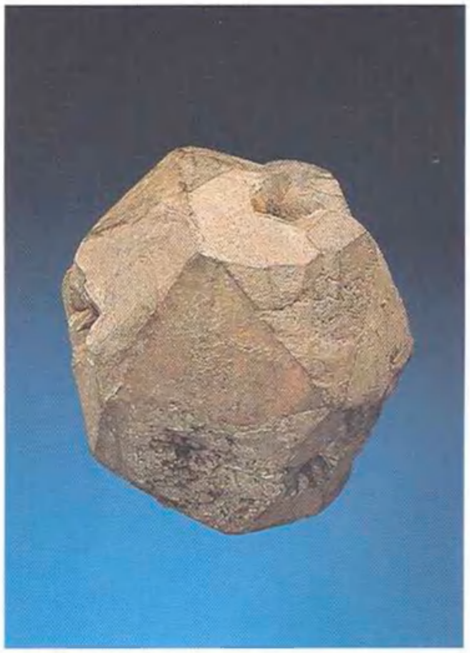

Sperrylite, 55 mm, Talnakh, Russia

Sperrylite

PtAs$_2$

CUBIC ● ●

Properties: C – tin-white; S – black; L – metallic; D – opaque; DE – 10,8; H – 6-7; CL – imperfect; F – conchoidal; M – complex combinations of cubic crystals, massive.
Origin and occurrence: It is mainly of magmatic liquid origin, associated with pyrrhotite, pentlandite, cubanite and other minerals. The best crystals come from Talnakh near Norilsk, Siberia, Russia, where inter-grown crystals reached up to 50 mm (2 in). Crystals up to 40 mm (1 9/16 in) across known from Tweefontein, Potgietersrust, Bushveld, South Africa. It also occurred as disseminated aggregates in Sudbury, Ontario, Canada.

Aurostibite

AuSb$_2$

CUBIC ● ●

Properties: C – gray, tarnishing iridescent; S – gray; L – metallic; D – opaque; DE – 9,9; H – 3; CL Bnone; M – granular.
Origin and occurrence: Hydrothermal in quartz veins, associated with gold and other Sb minerals. Grains up to 5 mm (3/16 in) across come from Krasna Hora, Czech Republic. It also occurs in the Giant Yellowknife mine, Northwest Territories and Hemlo, Thunder Bay, Ontario, Canada and is also reported from Bestyube, Kazakhstan.

Krut'aite

CuSe$_2$

CUBIC ●

Properties: C – gray; S – gray; L – metallic; D – opaque; DE – 6,5; H – 4; M – microscopic crystals, massive.
Origin and occurrence: Hydrothermal, associated with clausthalite, uraninite and other minerals. The richest accumulations of massive aggregates were found in the El Dragon mine, Potosí, Bolivia, where crystals up to 1 mm (1/32 in) occurred as well. It was originally described as microscopic from Petrovice, Czech Republic.

Cobaltite

CoAsS

ORTHORHOMBIC ● ● ●

Properties: C – silver-white; S – gray-black; L – metallic, adamantine to dull; D – opaque;DE – 6,3; H – 5,5; CL – perfect; F – uneven; M – pseudo-cubic crystals, granular, massive.
Origin and occurrence: Hydrothermal in high-temperature ore deposits and metamorphic, together

Aurostibite, 4 mm grain, Krasna Hora, Czech Republic

Krut'aite, 40 mm, El Dragon, Bolivia

Cobaltite, 5 mm x, Hakansboda, Sweden

with magnetite, sphalerite, chalcopyrite, other sulfides and arsenides of Co and Ni. The best dodecahedra, several cm across, come from metamorphic sulfide deposits in Tunaberg, Sweden. Other Swedish localities, like Håkansboda and Ramsberg, yielded crystals up to 60 mm (24 in) across. Fine crystals about 10 mm (3/8 in) occurred in the magnesite deposit Mutnik near Hnust'a , Slovakia. Cubic crystals up to 30 mm (1 3/16 in) across found in Espanola, Ontario, Canada. Beautiful crystals are known from the skarn in Bimbowrie, South Australia, Australia.
Application: Co ore.

Gersdorfite
NiAsS

CUBIC ● ● ●

Properties: C – silver-white to steel-gray, tarnishing gray-black; S – gray-black; L – metallic; D – opaque; DE – 6,0; H – 5,5; CL – perfect; F – uneven; M – octahedral striated crystals and their combinations, massive. *Origin and occurrence:* Hydrothermal in medium-temperature ore deposits in association with nickeline, nickel-skutterudite, ullmanite, siderite and other minerals. Crystals up to 100 mm (4 in) across come from the Snowbird mine, Montana, USA. It is important in several mines near Sudbury and Cobalt, Ontario, Canada. Large cleavable masses are known from Dobsina, Rudnany and Nizna Slana, Slovakia. Crystals from the Aït Ahmane mine near Bou Azzer, Morocco, reached up to 40 mm (1 9/16 in) . *Application:* Ni ore.

Gersdorffite, 10 mm xx, Bou Azzer, Morocco

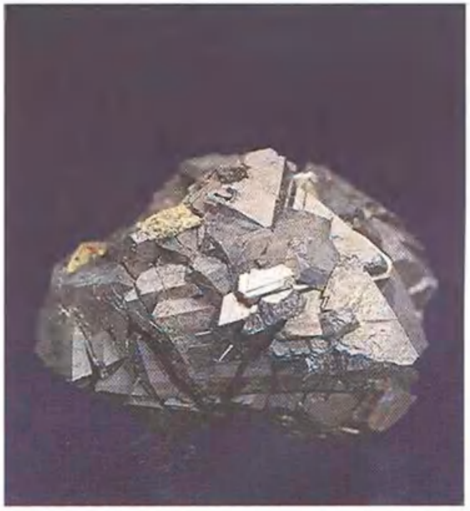

Ullmanite, 60 mm, Bou Cricha, Morocco

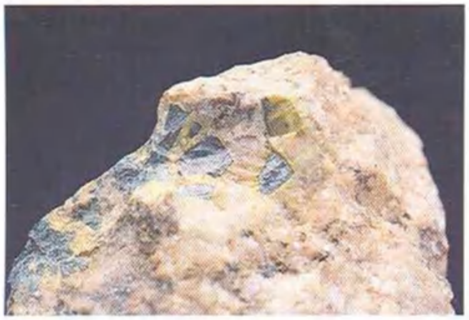

Ullmanite

NiSbS

CUBIC ● ●

Properties: C – steel-gray to silver-white; S – gray-black; L – metallic; D – opaque; DE – 6,8; H – 5-5,5; CL – perfect; F – uneven; M – combinations of cubic crystals, massive.
Origin and occurrence: Hydrothermal in ore veins together with other Ni minerals. Fine crystals, up to 20 mm ($^{25}/_{32}$ in) across, come from Monte Narba near Sarrabus, Sardinia, Italy. Twins are known from Lolling, Austria. Crystals up to 10 mm ($^{3}/_{8}$ in) were found in Ksice near Støibro, Czech Republic. It is common in Broken Hill, New South Wales, Australia and in Cochabamba, Bolivia.
Application: Ni ore.

Marcasite, 42 mm, Vintirov, Czech Republic

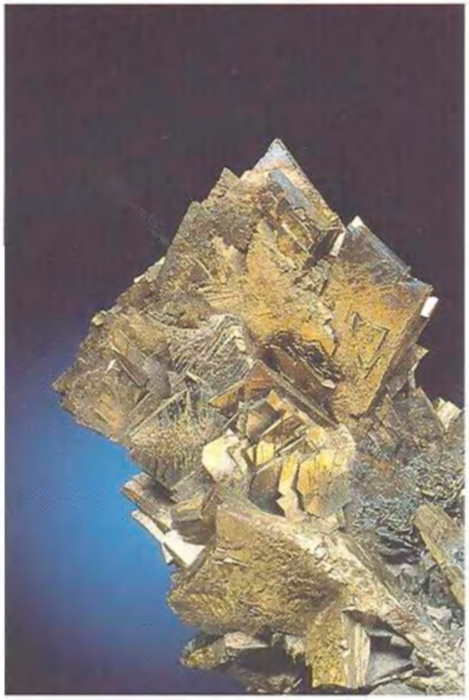

Marcasite

FeS_2

ORTHORHOMBIC ● ● ● ●

Properties: C – tin-white, bronze-yellow, tarnishing iridescent; S – grayish to brownish- black; L – metallic; D – opaque; DE – 4,9; H – 6-6,5; CL – good; F – uneven; M – tabular, pyramidal and prismatic crystals, often twinned into the form of cockscomb-like aggregates, stalactitic, botryoidal and massive.
Origin and occurrence: It originates at low temperatures in very acidic environment, either in sedimentary, or in hydrothermal deposits, associated with pyrite, pyrrhotite, galena, sphalerite, fluorite, dolomite and calcite. Hydrothermal crystals and pseudo-morphs after pyrrhotite are known from Freiberg, Germany; Llallagua, Bolivia and Chiuzbaia, Romania. Crystals from Wiesloch, Germany and Reocin, Santander, Spain are of similar origin. Large crystals occurred in Joplin, Missouri and in Galena, Illinois, USA. The best crystals of sedimentary marcasite come from coal basins. Fine crystals from black coal are known from Essen, Germany. Cockscomb-like aggregates up to 150 mm (6 in) across, come from the brown coal basin in Vintiøov, Czech Republic. Spherical, radial concretions with pyrite are known from Sparta, Illinois, USA and from Champagne, France.
Application: production of sulfuric acid.

Marcasite, 70 mm, Sparta, U.S.A.

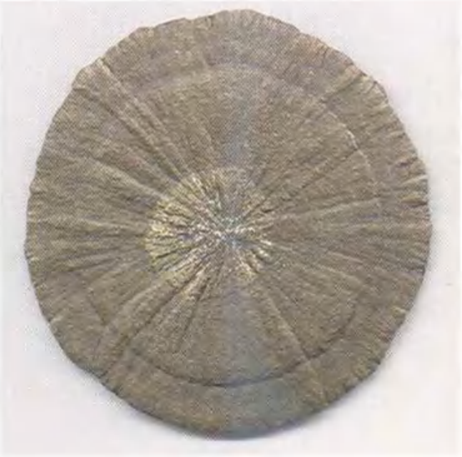

Lollingite, 7 mm xx, Cobalt, Canada

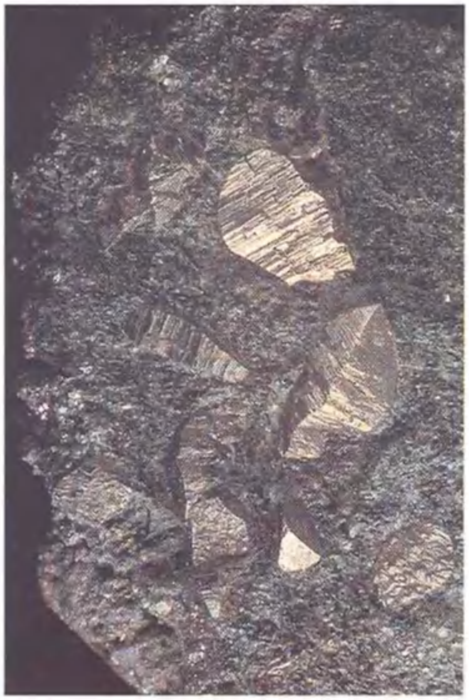

Löllingite

$FeAs_2$

ORTHORHOMBIC ● ● ●

Properties: C – steel-gray to silver-white; S – gray-black; L – metallic; D – opaque; DE – 7,5; H – 5-5,5; CL – sometimes good; F – uneven; M – prismatic crystals, massive.
Origin and occurrence: Magmatic in pegmatites; hydrothermal in greisens and Sn-W veins, rare in the other types of ore veins, together with skutterudite, bismuth, nickeline, siderite, calcite and other minerals. Fine crystals are known from syenite in Langensundsfjord, Norway. Crystals up to 50 cm (20 in) across come from a pegmatite in Kaatiala, Finland. Masses occur in the Kobokobo pegmatite, Kivu, Zaire. Massive aggregates with schorl were found in Dolni Bory, aggregates with cassiterite in Pøebuz, Czech Republic.

Safflorite, 70 mm, Bou Azzer, Morocco

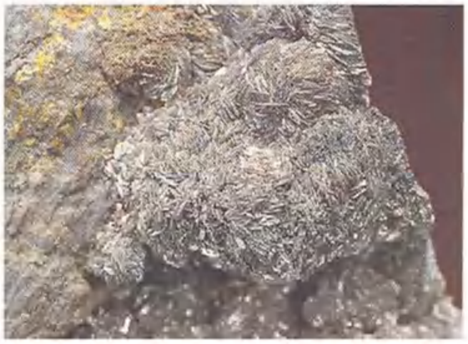

Safflorite

$(Co,Fe)As_2$

ORTHORHOMBIC ● ● ●

Properties: C – tin-white, tarnishing to dark gray; S – gray-black; L – metallic; D – opaque; DE – 7,5; H – 4,5-5; CL – good; F – uneven to conchoidal; M – prismatic crystals, radial aggregates, massive.
Origin and occurrence: Hydrothermal in medium-temperature with skutterudite, rammelsbergite, nickeline, silver, bismuth and lollingite. It is common in Schneeberg, Germany; Cobalt, Ontario, Canada; Batopilas, Chihuahua, Mexico and in Bou Azzer, Morocco.

Rammelsbergite

$NiAs_2$

ORTHORHOMBIC ● ● ●

Properties: C – tin-white, pinkish; S – gray-black; L – metallic; D – opaque; DE – 7,1; H – 5,5-6; CL – good; F – uneven; M – prismatic crystals, radial and fibrous aggregates, massive.
Origin and occurrence: Hydrothermal in medium-temperature veins associated with other Ni and Co minerals. Botryoidal aggregates come from Schneeberg, Germany. It is common in Sarrabus, Sardinia, Italy; in the Eldorado mine near Great Bear Lake; in Cobalt, Ontario, Canada and Bou Azzer, Morocco.

Rammelsbergite, 60 mm, Schneeberg, Germany

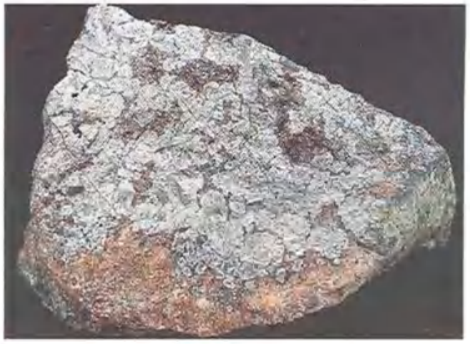

Arsenopyrite, 26 mm x, Portal, Mexico

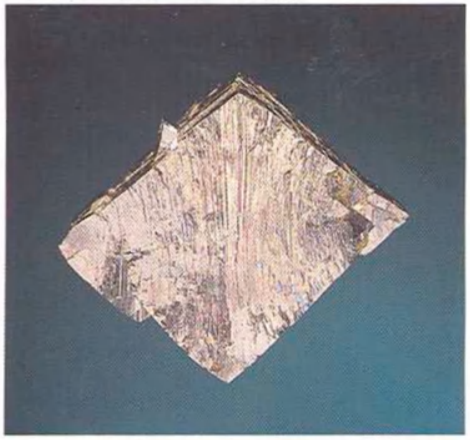

Arsenopyrite

FeAsS

MONOCLINIC ● ● ● ●

Properties: C – silver-white to steel-gray; S – black; L – metallic; D – opaque; DE – 6,2; H – 5,5 – 6; CL – good; F – uneven; M – thick tabular to prismatic striated crystals, granular, massive.
Origin and occurrence: It occurs in pegmatites; hydrothermal in high-temperature vein deposits and greisens; metamorphic in contact metamorphic skarns, gneisses and mica schists. Long prismatic crystals up to 30 cm (12 in) long are known from the Obira mine, Japan. It is very common in greisen Sn and W deposits, fine crystals are known from Horni Slavkov, Czech Republic and Ehrenfriedersdorf, Germany. Beautiful crystals up to 50 mm (2 in) across come from Panasqueira, Portugal, where they occurred associated with fluorapatite, wolframite and siderite. Historically important were large crystals from Tavistock, Devon , UK. Crystals up to 40 mm (1$^{9}/_{16}$ in) found in Llallagua, Bolivia. Crystals up to 50 mm (2 in) across were found recently in the Nikolai mine in Dalnegorsk, Russia. Shiny crystals up to 30 mm (1$^{3}/_{16}$ in) across, come from Hunan province, China. *Application:* As ore.

Gudmundite, 10 mm xx, Polar Urals, Russia

Gudmundite

FeSbS

MONOCLINIC ● ● ●

Properties: C – silver-white to steel-gray; S – black; L – metallic; D – opaque; DE – 7,0; H – 5-6; CL – none; F – uneven; M – prismatic twinned crystals, massive.
Origin and occurrence: Late hydrothermal mineral of ore deposits, also in metamorphic deposits and skarns. It is common in metamorphic sulfide deposits in Sweden (Boliden, Gudmundstorp). Massive aggregates are known from Kutna Hora and Vlastijovice, Czech Republic. It was common in Broken Hill, New South Wales, Australia.

Molybdenite

MoS_2

HEXAGONAL ● ● ● ●

Properties: C – lead-gray; S – blue-gray; L – metallic; D – opaque; DE – 4,0; H – 1-1,5; CL – perfect; M – tabular and prismatic crystals, scaly aggregates.
Origin and occurrence: Magmatic in pegmatites, granites and aplites, hydrothermal in high-temperature veins, also in porphyry ore deposits and in contact metamorphic deposits; associated with chalcopyrite, quartz and other minerals. Large crystals come from pegmatites in Blue Hill Bay, Maine, USA and in Mutue-Fides-Stavoren, Transvaal, South Africa, where they reach several tens of cm in size. Crystals up to 150 x70 mm (6 x 2¾ in) across come from the transitional type between pegmatites and quartz veins near Arendal and Moss, Norway; large crystals also occur in the Temiskaming district, Quebec, Canada; tabular crystals up to 120 mm ($^{24}/_{32}$ in) across found in quartz-molybdenite breccia pipes in veins in Australia (Queensland, New South Wales). Fine crystals are also known from Kladnica near Vito_a, Bulgaria; Horni Slavkov, Czech Republic and Ehrenfriedersdorf, Germany. As a fine grained disseminated ore was mined in Bingham, Utah and Climax, Colorado, USA.
Application: Mo ore.

Skutterudite

$CoAs_{2-3}$

CUBIC ● ● ●

Properties: C – tin-white to silver-gray, tarnishing to

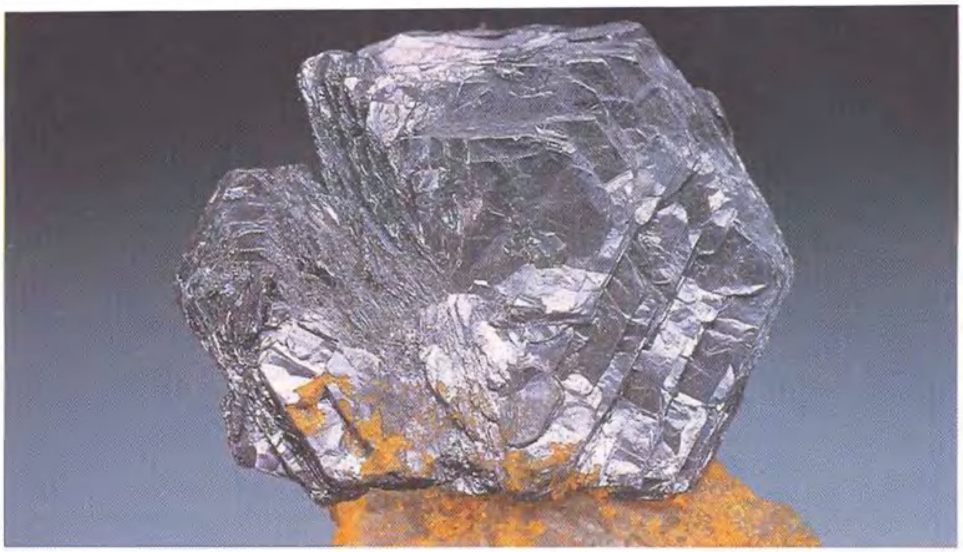

gray and iridescent; S – black; L – metallic; D – opaque; DE – 6,8; H – 5,5 – 6; CL – good; F – conchoidal to uneven; M – combinations of cubic crystals, skeletal aggregates, granular, massive.

Origin and occurrence: Hydrothermal in medium- to high-temperature ore veins, associated with other Ni and Co minerals. Crystals up to 50 mm (2 in) across come from Bou Azzer, Morocco. Crystals up to several cm in size were found in Schneeberg and Annaberg, Germany. Large massive accumulations occur in Cobalt and Gowganda, Ontario, Canada. *Application:* Co ore.

Skutterudite, 25 mm, Schneeberg, Germany

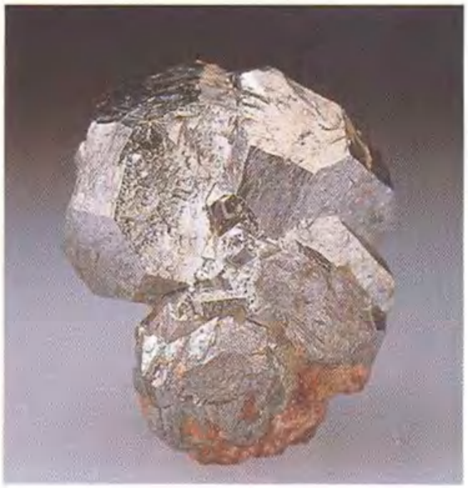

Nickel-skutterudite

$NiAs_3$

CUBIC ● ● ● ●

Properties: C – tin-white to silver, tarnishing to gray and iridescent; S – black; L – metallic; D – opaque; DE – 6,5; H – 5,5 – 6; CL – good; F – conchoidal to uneven; M – combinations of cubic crystals, skeletal aggregates, granular.

Origin and occurrence: Hydrothermal in medium-temperature veins with arsenopyrite, arsenic, bismuth, calcite and siderite. Known in crystals from Chatham, Connecticut and Chester, Massachusetts, USA; also Val d'Anniviers, Wallis, Switzerland. Massive aggregates come from Dobsina, Slovakia; Les Chalanches, France; Mohawk mine, Michigan, USA; Schneeberg, Germany. *Application:* ruda Co a Ni.

Nickel-skutterudite, 36 mm, Saxony, Germany

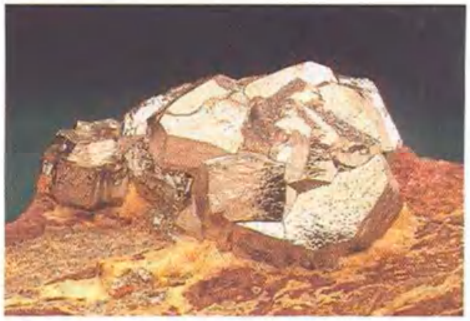

Proustite, 25 mm, Chanarcillo, Chile

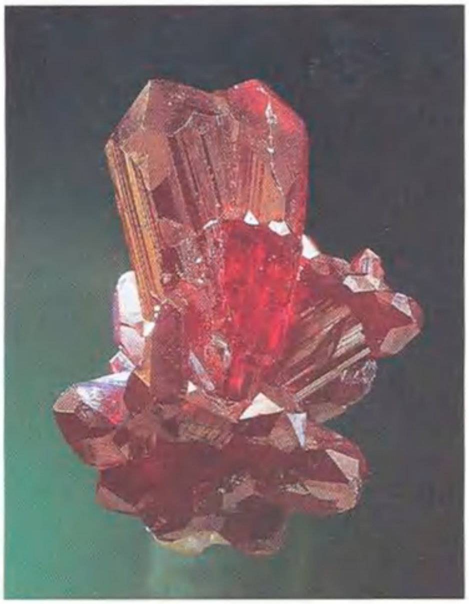

Proustite

Ag_3AsS_3

TRIGONAL ● ●

Properties: C – crimson, darkens upon exposure to light; S – crimson; L – adamantine; D – translucent to opaque; DE – 5,6; H – 2-2,5; CL – good; F conchoidal to uneven; M – prismatic, rhombohedral and scalenohedral crystals, massive.

Origin and occurrence: Low-temperature hydrothermal mineral, also in the oxidation and cementation zone together with stephanite, silver, xanthoconite, acanthite and other minerals. The best specimens with crystals up to 100 mm (4 in) long come from the Dolores mine, Chanarcillo, Chile. Crystals up to 80 mm ($3^{1}/_{8}$ in) long found in the Himmelsfürst mine in Freiberg, Niederschlema and Schneeberg, Germany. Large druses with crystals up to 40 x 20 mm ($1^{9}/_{16}$ x $^{25}/_{32}$ in) across occurred in Jachymov; crystals up to 20 mm ($^{25}/_{32}$ in) across known from Pøibram and Stara Vozice, Czech Republic. Fine crystals come from Batopilas, Chihuahua and Sombrerete, Zacatecas, Mexico. Crystalline masses of proustite, weighing over 250 kg (550 lb), were found in 1865 in the Poorman mine, Silver City, Idaho, USA. *Application:* Ag ore.

Pyrargyrite, 61 mm, Zacatecas, Mexico

+
→

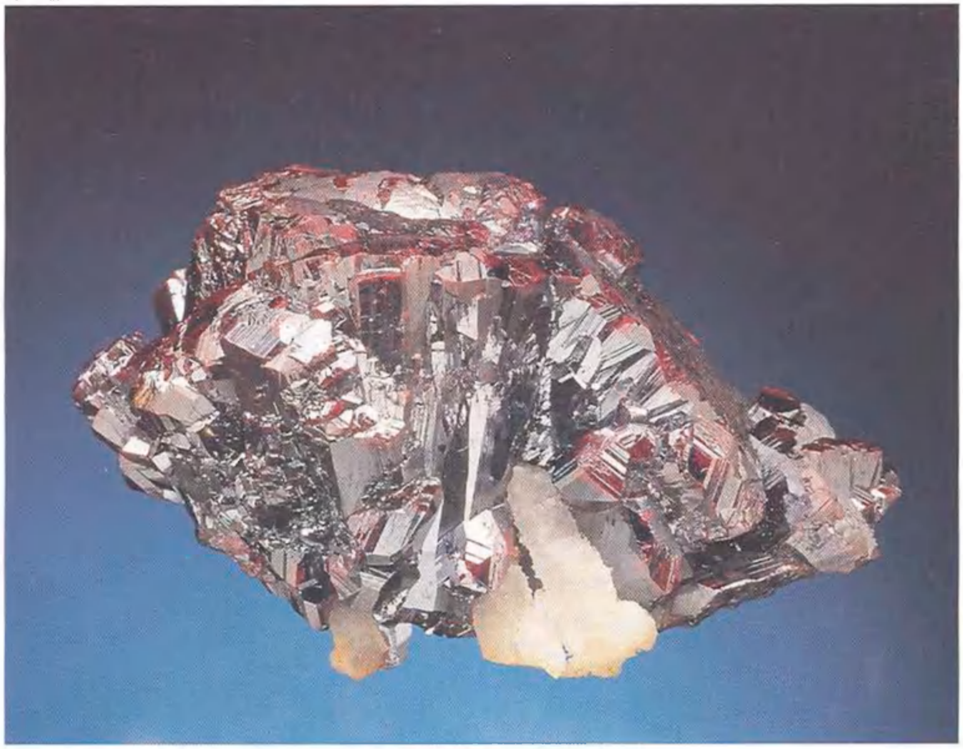

Pyrargyrite

Ag_3SbS_3

TRIGONAL ● ●

Properties: C – dark red, darkens upon exposure to light; S – crimson; L – adamantine; D – translucent to opaque; DE – 5,9; H – 2,5; CL – good; F – conchoidal to uneven; M – prismatic, rhombohedral and scalenohedral crystals, granular, massive.
Origin and occurrence: Low-temperature hydrothermal mineral, also secondary in the oxidation and cementation zone, together with silver, acanthite, other Ag sulfosalts, calcite and quartz. Crystals in Colquechaca, Bolivia and Chañarcillo, Chile reached several cm in size. Crystals up to 70 mm (2¾ in) long occurred in the Santo Nino vein in Fresnillo, Zacatecas, Mexico. San Genaro mine in Huancavelica, Peru yielded crystals up to 50 mm (2 in) across. Crystals up to 40 mm ($1^{9}/_{16}$ in) across found in Freiberg; smaller crystals only are known from St. Andreasberg, Germany. Crystals up to 20 mm ($^{25}/_{32}$ in) across come from Pribram and Stara Vozice, Czech Republic. Crystals up to 50 mm (2 in) across found in the San Carlos mine, Hiendelaencina, Spain.
Application: Ag ore.

Xanthoconite

Ag_3AsS_3

MONOCLINIC ● ●

Properties: C – dark crimson, orange-yellow to yellow-brown; S – orange-yellow; L – adamantine; D – translucent; DE – 5,5; H – 2-3; CL – good; F – conchoidal; M – tabular and lath-like crystals, botryoidal and radial aggregates.

Xanthoconite, 2 mm xx, Marienberg, Germany

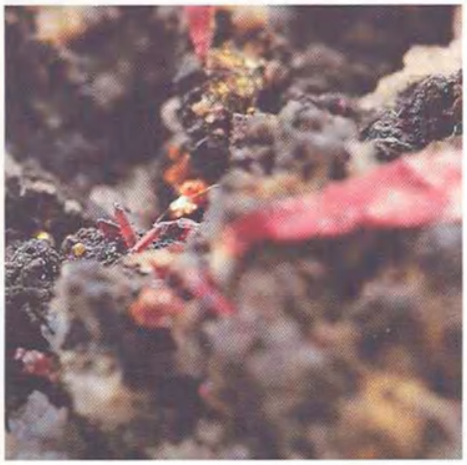

Pyrostilpnite, 60 mm, Potosí, Bolivia

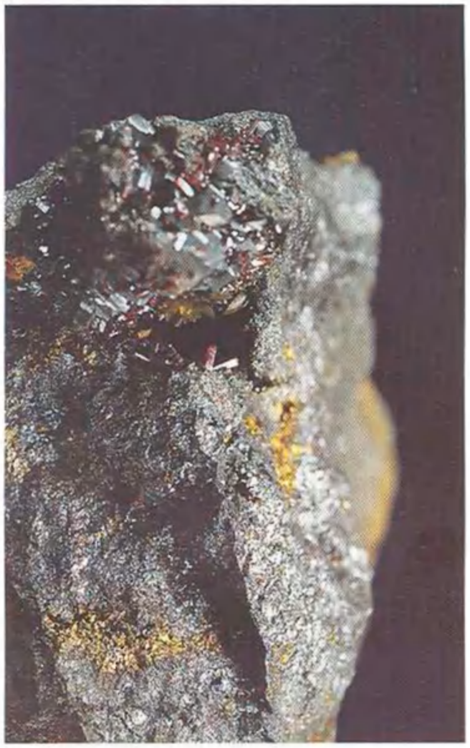

Origin and occurrence: Hydrothermal in ore veins together with proustite, pyrargyrite, acanthite, arsenic and calcite. Botryoidal masses with yellow crystals up to 7mm ($^{9}/_{32}$ in) long come from Freiberg; other important localities are St Andreasberg, Germany; Ste-Marie-aux- Mines, France; Cobalt, Ontario, Canada; Příbram, Trebsko and Jachymov, Czech Republic.

Pyrostilpnite

Ag_3SbS_3

MONOCLINIC ● ●

Properties: C – hyacinth- to orange-red; S – orange-yellow; L – adamantine; D – translucent; DE – 6,0; H – 2; CLB perfect; F – conchoidal; M – tabular to lath-like crystals, radial aggregates.
Origin and occurrence: Hydrothermal in low-temperature veins, associated with pyrargyrite, stephanite, acanthite and other Ag minerals. The best crystals come from St. Andreasberg, Germany. Crystals up to 10 mm ($^{3}/_{8}$ in) long were found in Pribram, Trebsko and Jachymov, Czech Republic. It is also described from Colquechaca, Bolivia and Chañarcillo, Chile.

Samsonite

$Ag_4MnSb_2S_6$

MONOCLINIC ●

Properties: C – steel-gray; S – dark red; L – metallic; D – opaque; DE – 5,5; H – 2,5; CL – none; F – conchoidal; M – prismatic striated crystals.
Origin and occurrence: The only locality, where it occurred in relatively larger amount, was the Samson mine in St. Andreasberg, Germany, where crystals up to 10 mm (3/8 in) across were found.

Samsonite, 12 mm x, St.Andreasberg, Germany

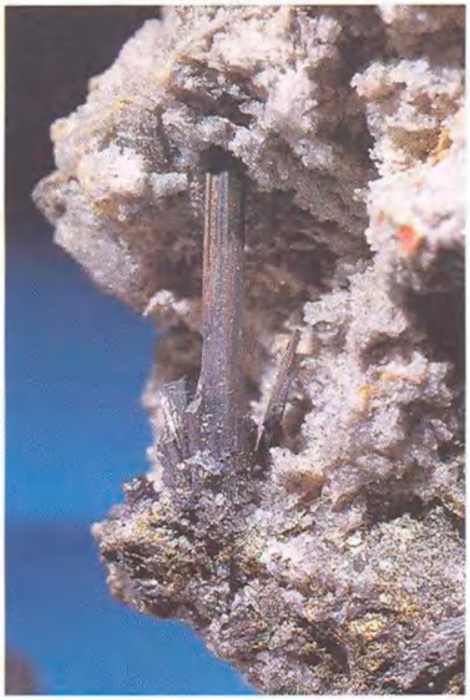

Chalcostibite

$CuSbS_2$

ORTHORHOMBIC ● ●

Properties: C – lead-gray, tarnishing to blue and green; S – lead-gray; L – metallic; D – opaque; DE – 5,0; H – 3-4; CL – perfect; F – conchoidal; M – long prismatic, striated crystals, granular, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with jamesonite, chalcopyrite, tetrahedrite, stibnite, andorite and other minerals. Partly altered crystals up to 100 mm (4 in) long are known from Rar-el-Auz near Casablanca, Morocco. It is often in deposits in Bolivia (Huanchaca, Oruro, Colquechaca), where its crystals reach 10 mm (3/8 in)

Chalcostibite, 29 mm, Saint Pons, France

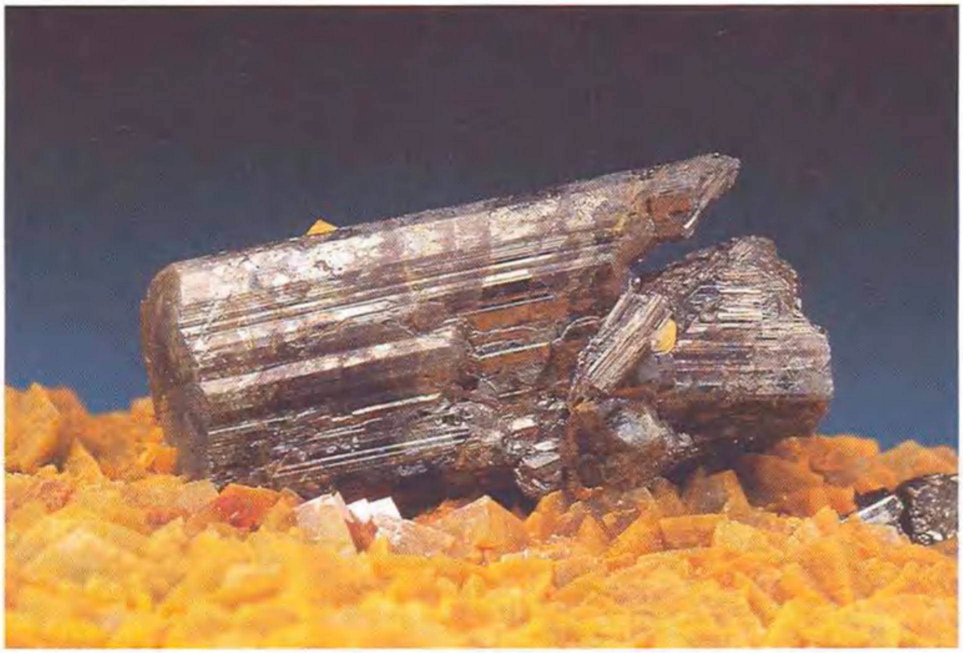

Emplectite, 60 mm, Schwarzenberg, Germany

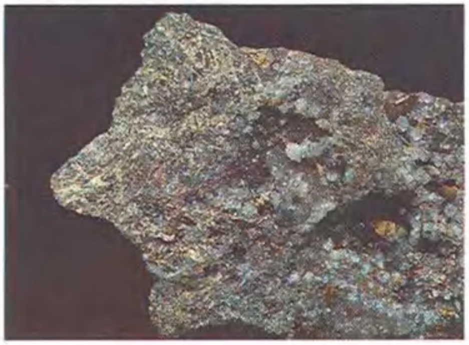

in size. The world's best crystals up to 16 cm ($6^{5}/_{16}$ in) long were discovered recently near St Pons, France.

Emplectite

$CuBiS_2$

ORTHORHOMBIC ● ● ●

Properties: C – gray to tin-white; S – gray; L – metallic; D – opaque; DE – 6,4; H – 2; CL – perfect; F – conchoidal to uneven; M – thin prismatic to acicular striated crystals.
Origin and occurrence: Hydrothermal in high-temperature veins associated with chalcopyrite, molybdenite, quartz, tetrahedrite and other minerals. Fine acicular crystals up to 30 mm ($1^{3}/_{16}$ in) long come from Krupka, Czech Republic. Crystallized specimens were also found in Wittichen and Johanngeorgenstadt, Germany; Colquijirca, Peru and in the Akenobe mine, Japan.

Wittichenite, 50 mm, Wittichen, Germany

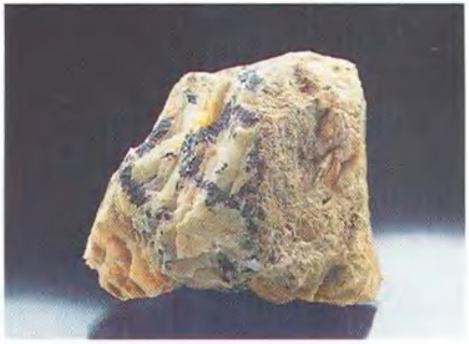

Wittichenite

Cu_3BiS_3

ORTHORHOMBIC ● ●

Properties: C – steel-gray to tin-white, tarnishing yellow to steel-gray; S – black; L – metallic; D – opaque; DE – 6,2; H – 2-3; CL – none; F – conchoidal; M – prismatic crystals, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with other Bi minerals, Cu-Fe sulfides, selenides and secondary U minerals. It occurs in Wittichen, Germany; Baiţa Bihorului, Romania; Tsumeb, Namibia and Cerro de Pasco, Peru.

Berthierite, 98 mm, Herja, Romania

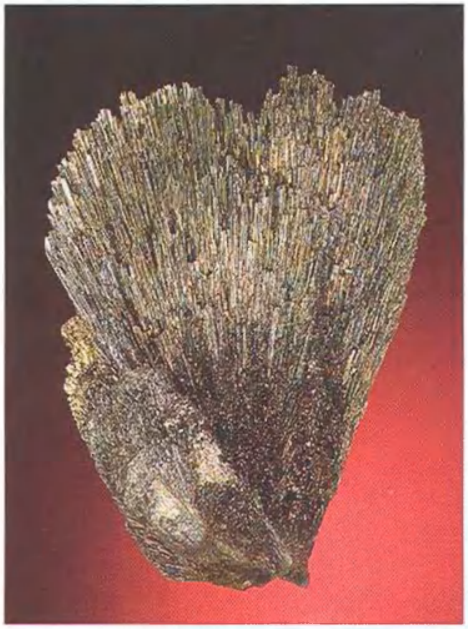

Berthierite

$FeSb_2S_4$

ORTHORHOMBIC ● ● ●

Properties: C – dark steel-gray, tarnishing iridescent to brown; S – dark brown-gray; L – metallic; D – opaque; DE – 4,7; H – 2-3; CL – imperfect; M – long prismatic, striated crystals, fibrous, felt-like and radial aggregates.
Origin and occurrence: Hydrothermal in low-temperature Sb deposits. Acicular crystals up to 10 mm ($^{3}/_{8}$ in) long are known from the St. Antoni de Padua gallery in Kutna Hora, Czech Republic; thick prismatic crystals come from Poproc, Slovakia. Iridescent columnar aggregates up to 200 mm ($7^{7}/_{8}$ in) long occur in Herja, Romania. Fine specimens are known also from Oruro, Bolivia.

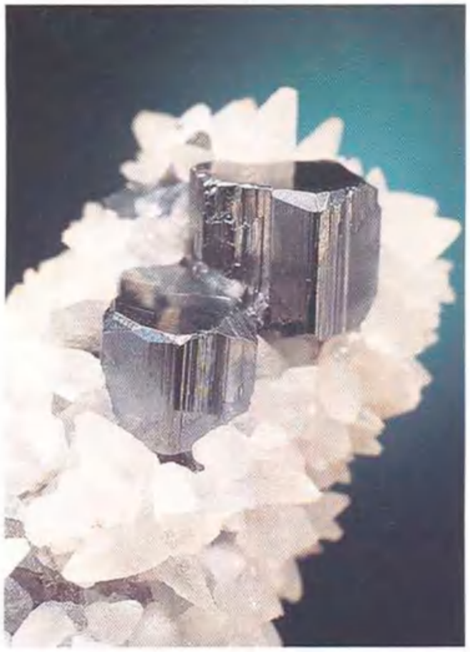

Stephanite, 8 mm xx. Zacatecas, Mexico

Stephanite

Ag_5SbS_4

ORTHORHOMBIC ● ●

Properties: C – black; S – black; L – metallic; D – opaque; DE – 6,3; H – 2-2,5; CL – imperfect; F – conchoidal; M – short prismatic to tabular striated crystals, massive.
Origin and occurrence: Late hydrothermal mineral in Ag deposits, associated with proustite, acanthite, silver, tetrahedrite, galena, sphalerite and pyrite. Crystals up to 40 mm ($1^{9}/_{16}$ in) long come from Pribram and Jachymov, Czech Republic and from St. Andreasberg, Germany. Crystals up to 50 mm (2 in) across found in the Las Chispas mine, Arizpe, Sonora, Mexico and Hiendelaencina, Spain. Smaller crystals occurred in Freiberg, Schneeberg and Annaberg, Germany. *Application:* Ag ore.

Pearceite

$Ag_{16}As_2S_{11}$

MONOCLINIC ● ●

Properties: C – black; S – black; L – metallic; D – opaque; DE – 6,1; H – 3; CL – none; F – conchoidal to irregular; M – short prismatic to tabular crystals and rosette-like aggregates, massive.
Origin and occurrence: Hydrothermal in low- to medium-temperature deposits with acanthite, silver, proustite, quartz, barite and calcite. Crystals several mm across are known from Jachymov, Moldava and Midinec, Czech Republic; from Arqueros, Chile and from the Veta Rica mine, Coahuila, Mexico. Crystals up to 12 mm ($^{15}/_{32}$ in) across were found in the Caribou mine, Colorado, USA; also in Dzhezkazgan, Kazakhstan.
Huge accumulations of almost pure pearceite occurred in the Mollie Gibson mine near Aspen, Colorado, USA.
Application: Ag ore.

Polybasite

$(Ag,Cu)_{16}Sb_2S_{11}$

MONOCLINIC ● ●

Properties: C – black; S – black; L – metallic; D Bopaque; DE – 6,4; H – 2-3; CL – imperfect; F – uneven; M – pseudo-hexagonal tabular crystals, massive.
Origin and occurrence: Hydrothermal in low- to medium-temperature ore veins, associated with

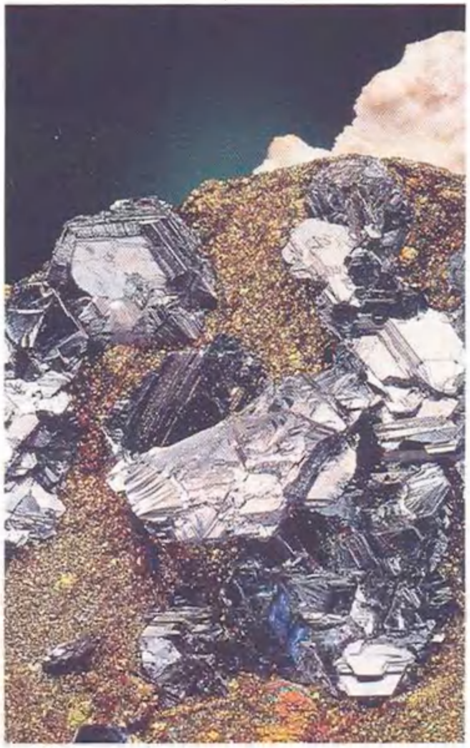

Pearceite, 15 mm xx, Guanajuato, Mexico

Polybasite, 39 mm, Zacatecas, Mexico

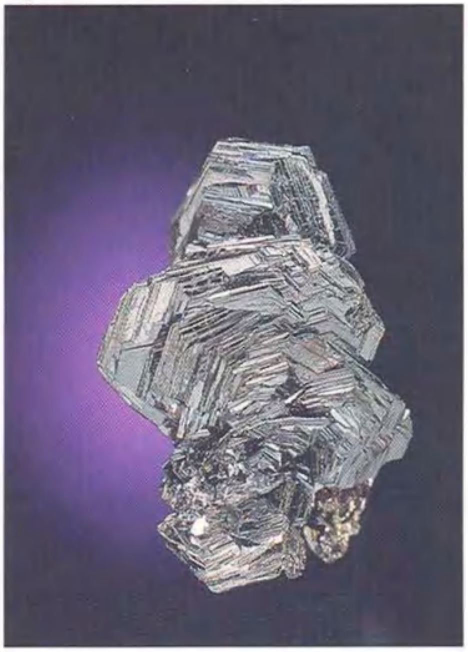

pyrargyrite, tetrahedrite, stephanite, acanthite, quartz and other minerals. Crystals, several cm across, are known from Wolfach, St. Andreasberg, Freiberg and Schneeberg, Germany; also from Guanajuato, Mexico. The best specimens with tabular crystals up to 90 mm ($3^{9}/_{16}$ in) across come from the Las Chispas mine, Arizpe, Sonora, Mexico. *Application:* Ag ore.

Lorandite, 10 mm x, Allchar, Macedonia

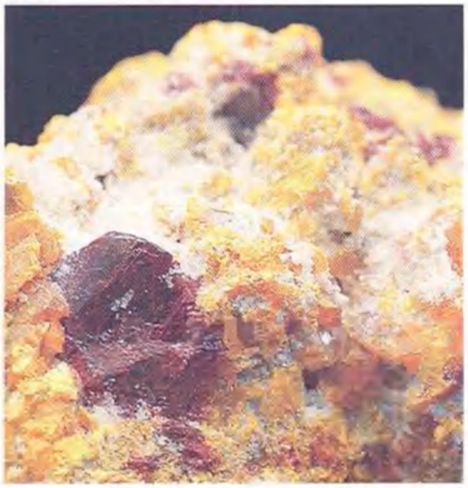

Lorandite

$TlAsS_2$

MONOCLINIC ● ●

Properties: C – crimson, lead-gray, it covers with a yellow coating; S – cherry-red; L – metallic to adamantine; D – translucent to transparent; DE – 5,5; H – 2-2,5; CL – perfect; M – short prismatic to tabular crystals, granular, massive.
Origin and occurrence: Hydrothermal, associated with stibnite, realgar, orpiment, pyrite and other minerals. Crystals up to 50 mm (2 in) across were found in Allchar, Macedonia. It is also known from Djijikrut, Tajikistan and from the cavities in dolomite from Lengenbach, Binntal, Switzerland.

Livingstonite

$HgSb_4S_8$

MONOCLINIC ● ●

Properties: C – black-gray; S – red; L – metallic to adamantine; D – opaque; DE – 5,0; H – 2; CL – perfect; M – acicular crystals, columnar and fibrous aggregates, massive.
Origin and occurrence: Hydrothermal in low-temperature veins, associated with cinnabar, stibnite, getchellite and other minerals. The best specimens with prismatic crystals up to 50 mm (2 in) long are known from Khaidarkan, smaller crystals only found from Kadamdzhai, Kyrgyzstan. It is also described from the La Cruz mine, Huitzuco, Guerrero, Mexico and from the Matsuo mine, Japan.

Livingstonite, 40 mm, Khaidarkan, Kyrgystan

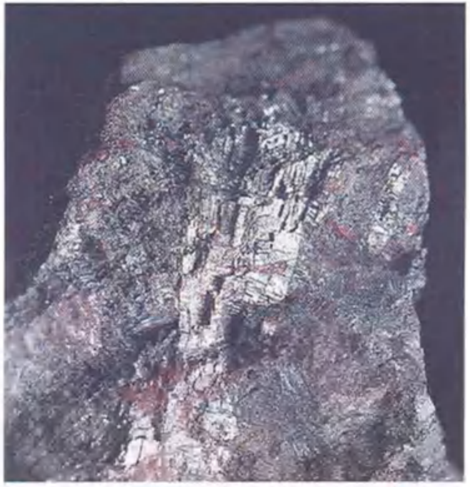

Bournonite, 62 mm, Cornwall, UK

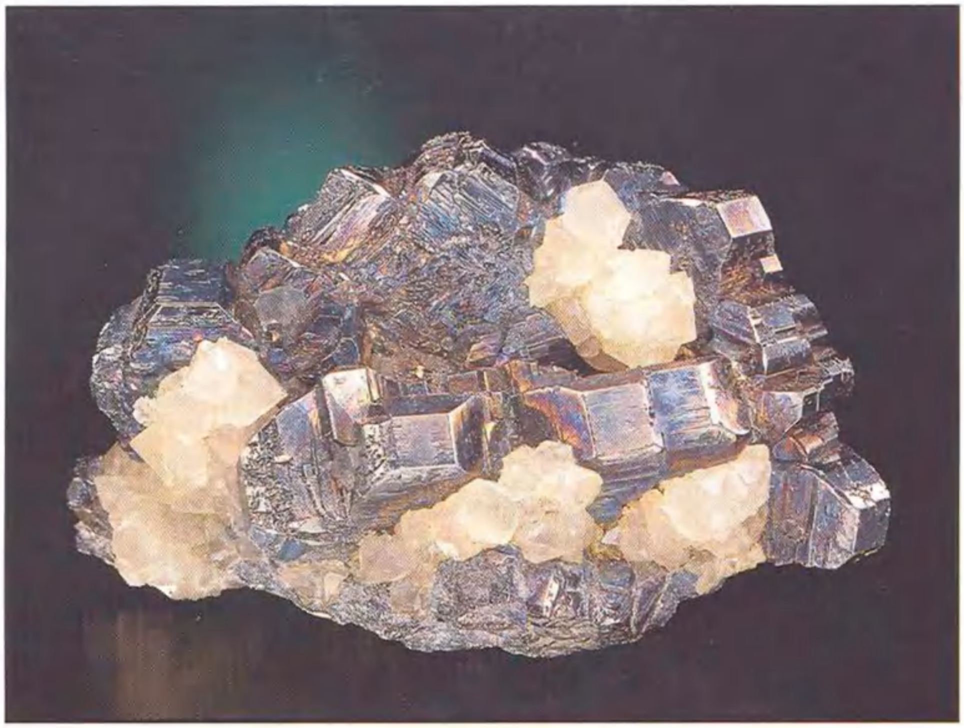

Bournonite

$PbCuSbS_3$

ORTHORHOMBIC ● ● ● ●

Properties: C – steel-gray to black; S – steel-gray to black; L – adamantine to dull; D – opaque; DE – 5,8; H – 2,5-3; CL – imperfect; F – conchoidal to uneven; M – short prismatic to tabular crystals, often striated, granular, massive.
Origin and occurrence: Hydrothermal in medium-temperature ore deposits, together with galena, tetrahedrite, pyrite, siderite and other minerals. The finest crystals, complex twins called cogswheel ore, over 50 mm (2 in) across come from the Herodsfoot mine near Liskeard, Cornwall, UK. Tabular crystals up to 40 mm ($1\frac{9}{16}$ in) across found in siderite cavities in Příbram, Czech Republic; large prismatic and tabular crystals known from Neudorf, Germany. Smaller crystals common in Cavnic and Baia Sprie, Romania. Crystals, up to 100 mm (4 in) across, occurred in Machacamarca, Bolivia. Crystals up to 40 mm ($1\frac{9}{16}$ in) come from Huancavelica, Peru. Crystals up to 100 mm (4 in) across reported from the Les Malines mine, France. Crystals up to 20 mm (25/32 in) across found at Chenzhou, Hunan province, China. *Application:* Pb, Cu and Sb ore.

Aikinite

$PbCuBiS_3$

ORTHORHOMBIC ● ●

Properties: C – black-gray, tarnishing brown; S – gray-black; L – metallic; D – opaque; DE – 7,3; H –

Aikinite, 2 mm xx, Rud?any, Slovakia

Betekhtinite, 10 mm xx, Dzhezkazgan, Kazakhstan

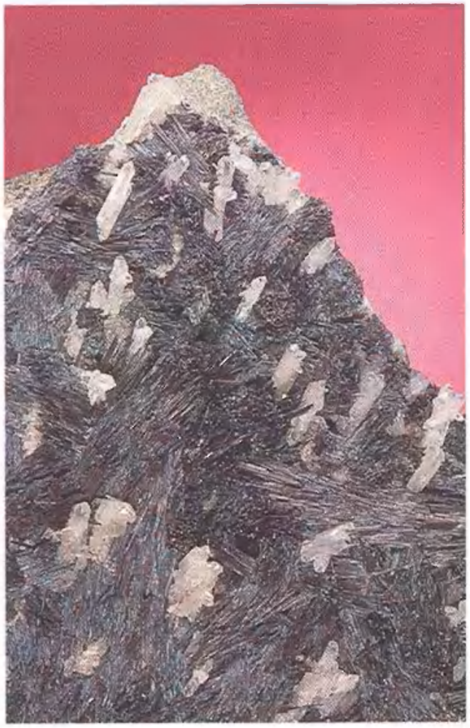

2-2,5; CL – imperfect; F – uneven; M – prismatic to acicular, striated crystals, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with gold, pyrite, galena, tennantite and other minerals. It is common in quartz veins with gold in Berezovsk, Ural Mountains, Russia, where crystals up to 30 mm ($1^{3}/_{16}$ in) long were found; fine crystals come also from La Gardette, Bourg d'Oisans, France. Grains up to 50 mm (2 in) across occurred in the Outlaw mine, Nevada, USA. It is also known in metamorphic veins in Val d'Anniviers, Switzerland.

Betekhtinite

$Cu_{10}(Fe,Pb)S_6$

ORTHORHOMBIC ● ●

Properties: C – black; S – black; L – metallic; D – opaque; DE – 6,1; H – 3-3,5; CL – good; M – acicular crystals, granular.
Origin and occurrence: Hydrothermal in ore deposits. The best specimens come from Dzhezkazgan as clusters of acicular crystals up to 70 mm (2¾ in) long, associated with bornite, chalcocite, djurleite and other minerals. Rich specimens found in Kipushi Kipushi, Shaba, Zaire. Granular aggregates fairly common in calcite veinlets, cross-cutting Cu-bearing shales near Eisleben, Mansfeld, Germany.

Andorite

$PbAgSb_3S_6$

ORTHORHOMBIC ● ●

Properties: C – dark steel-gray, tarnishing to yellow and iridescent; S – black; L – metallic; D – opaque; DE – 5,4; H – 3-3,5; CL – none; F – conchoidal; M – prismatic and tabular striated crystals, massive.
Origin and occurrence: Hydrothermal in ore deposits, together with cassiterite, jamesonite, stannite and other minerals. The world's best specimens come from the Itos and San Jose mines in Oruro and the Tatasi mine in Potosí, Bolivia, where it forms crystals up to 30 mm ($1^{3}/_{16}$ in) across. Thin tabular crystals are known from Baia Sprie, Romania and from the Keyser mine, Nevada, USA. Needles up to 10 mm ($^{3}/_{8}$ in) long occurred in Trebsko, Czech Republic.
Application: ruda Ag.

Andorite, 45 mm, Oruro, Bolivia

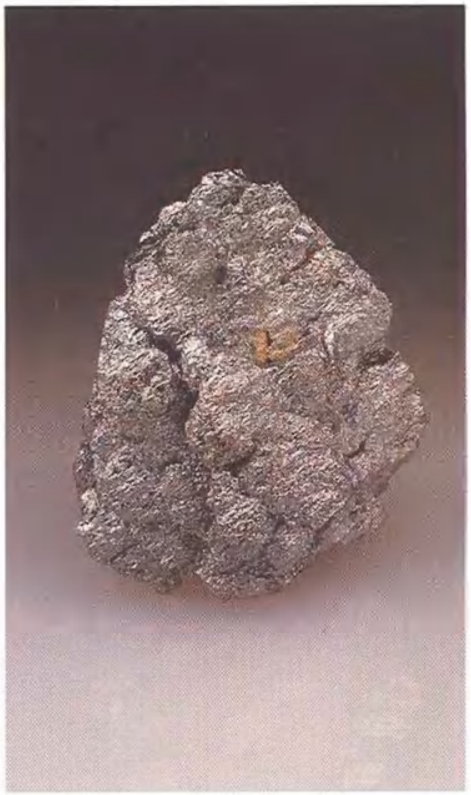

Freieslebenite, 2 mm xx, Baia Sprie, Romania

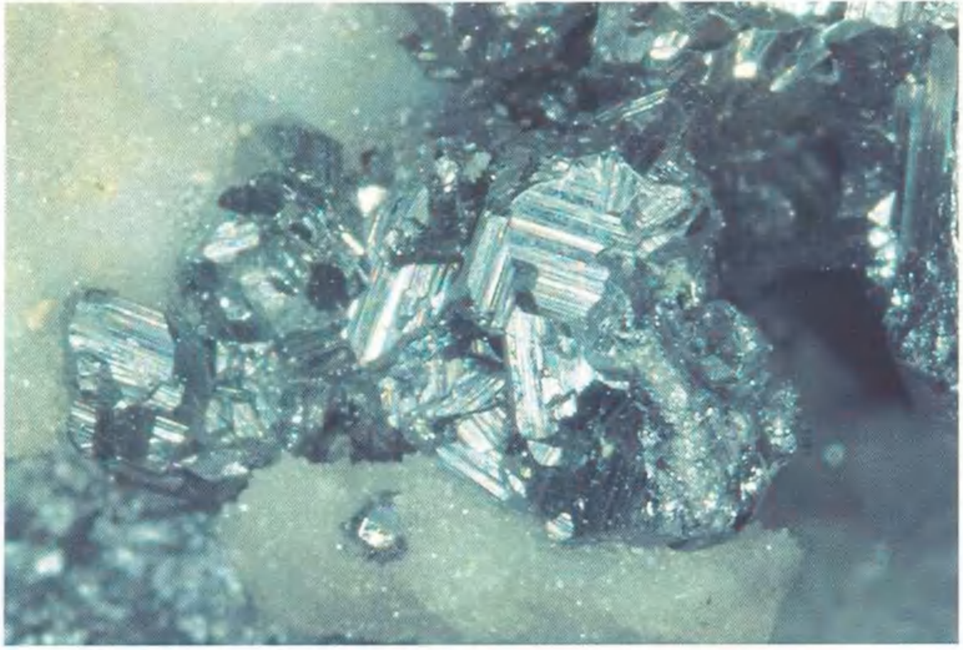

Freieslebenite

$AgPbSbS_3$

MONOCLINIC ● ●

Properties: C – light steel-gray, lead-gray to silver-white; S – light steel-gray, lead-gray to silver-white; L – metallic; D – opaque; DE – 6,2; H – 2-2,5; CL – imperfect; F – conchoidal to uneven; M – prismatic, striated crystals.

Origin and occurrence: Hydrothermal in ore deposits, associated with acanthite, pyrargyrite, silver, galena, siderite and andorite.

Crystals up to 20 mm ($^{25}/_{32}$ in) across come from the Santa Cecilia, Guadalajara, Spain; Freiberg, Germany; Oruro, Bolivia and from the Treasury Lode, Colorado, USA.

Diaphorite

$Pb_2Ag_3Sb_3S_3$

MONOCLINIC ● ●

Properties: C – steel-gray; S – steel-gray; L – metallic; D – opaque; DE – 6,0; H – 2,5-3; CL – none; F – conchoidal to uneven; M – prismatic striated crystals.

Origin and occurrence: Hydrothermal in medium-temperature ore veins, associated with galena,

Diaphorite, 5 mm xx, Pribram, Czech Republic

Sartorite, 4 mm x, Binntal, Switzerland

sphalerite, pyrargyrite, pyrite and other minerals. Beautiful striated crystals up to 10 mm (3/8 in) come from the cavities of quartz veins in Pøíbram; rare small crystals, several mm across found in the cavities of quartz veins in the St. Antoni de Padua gallery in Kutna Hora, Czech Republic. Complicated combinations of crystals were described from Braünsdorf, Germany. Crystals up to 80 mm (3 1/8 in) long occurred in Hiendelaencina, Guadalajara, Spain. It is also reported from Catorce, San Luis Potosí, Mexico.

Sartorite

$PbAs_2S_4$

MONOCLINIC ● ●

Properties: C – dark lead-gray; S – chocolate-brown; L – metallic; D – opaque; DE – 5,1; H – 3; CL – good; F – conchoidal; M – prismatic striated crystals.
Origin and occurrence: Hydrothermal in dolomite, associated with tennantite, dufrenoysite, pyrite and realgar. The best crystals up to 100 mm (4 in) long come from Lengenbach, Binntal, Switzerland. It was also found in the Zuni mine, Colorado, USA.

Baumhauerite

$Pb_3As_4S_9$

TRICLINIC ●

Properties: C – lead- to steel-gray, tarnishing iridescent; S – chocolate-brown; L – metallic; D – opaque; DE – 5,4; H – 3; CL – perfect; F – conchoidal; M – tabular to short prismatic striated crystals, granular.
Origin and occurrence: Hydrothermal, associated with realgar and other sulfosalts. The best crystals up to 25 mm (1 in) across come from Lengenbach, Binntal, Switzerland. Massive aggregates were found in Hemlo, Thunder Bay, Ontario, Canada and Sterling Hill, New Jersey, USA.

Baumhauerite, 50 mm, Binntal, Switzerland

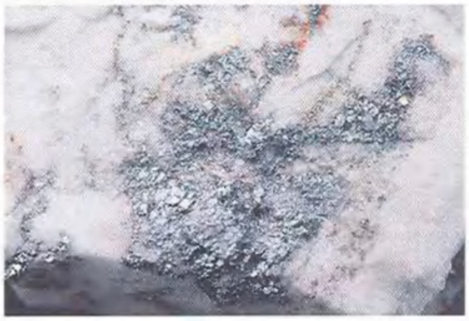

Rathite, 60 mm, Binntal, Switzerland

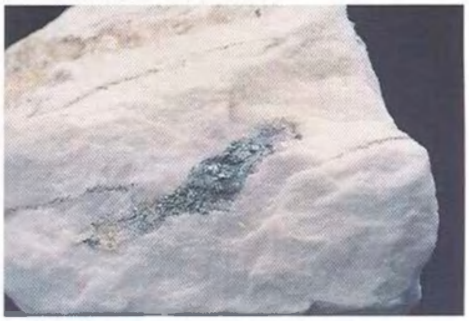

Rathite

$(Pb,Tl)_3As_5S_{10}$

MONOCLINIC ●

Properties: C – lead-gray, tarnishing iridescent; S – chocolate-brown; L – metallic; D – opaque; DE – 5,3; H – 3; CL – perfect; F – conchoidal; M – prismatic striated crystals.
Origin and occurrence: Hydrothermal, associated with other Pb-Tl-As-S minerals. Crystals up to 10 mm (3/8 in) across come from dolomite in Lengenbach, Binntal, Switzerland.

Dufrenoysite

$Pb_2As_2S_5$

MONOCLINIC ● ●

Properties: : C – lead- to steel-gray; S – red-brown to chocolate-brown; L – metallic; D – opaque; DE – 5,6; H – 3; CL – perfect; F – conchoidal; M – elongated striated tabular crystals.
Origin and occurrence: Hydrothermal low-temperature mineral, associated with rathite, sartorite, baumhauerite and realgar. Crystals up to 25 mm (1 in) across come from dolomite in Lengenbach, Binntal, Switzerland. Also found in Batopilas, Chihuahua, Mexico.

Dufrenoysite, 10 mm x, Binntal, Switzerland

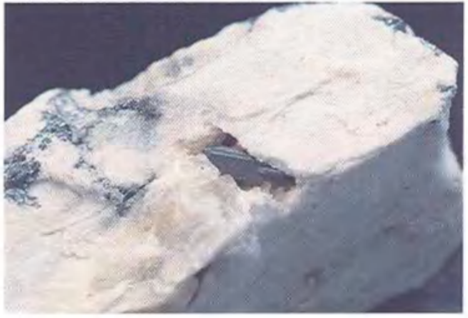

Jordanite, 12 mm x, Binntal, Switzerland

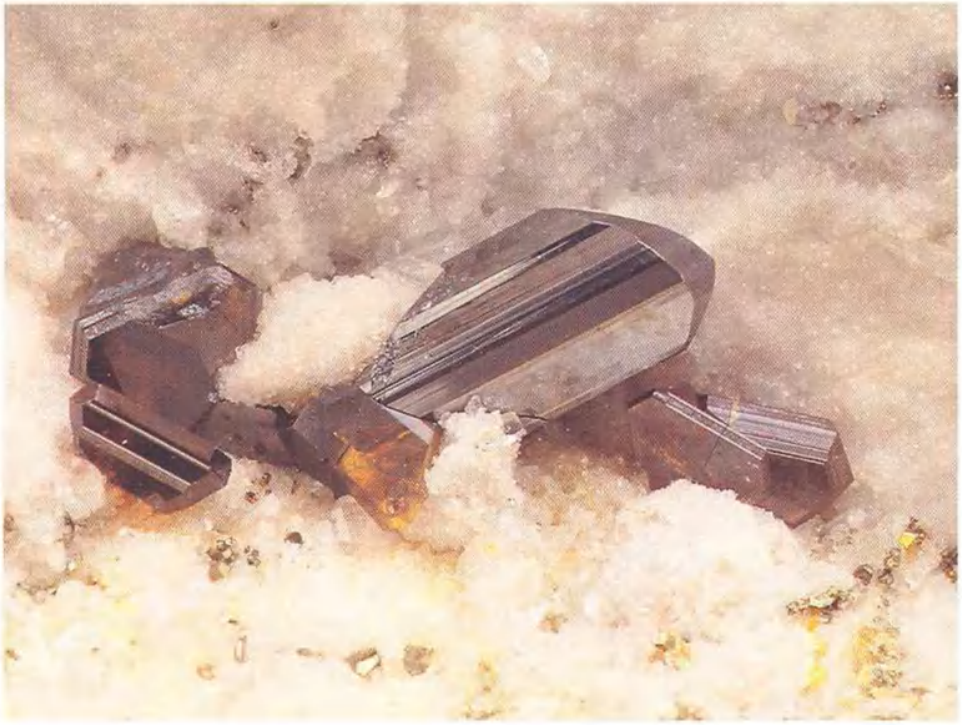

Jordanite

$Pb_{14}(As,Sb)_6S_{23}$

MONOCLINIC ● ●

Properties: C – lead-gray, tarnishing iridescent; S – black; L – metallic; D – opaque; DE – 6,4; H – 3; CL – perfect; F – conchoidal; M – tabular crystals, botryoidal aggregates.

Origin and occurrence: Hydrothermal in low-temperature ore veins, also in metamorphic dolomites, together with tennantite, sphalerite, galena, dolomite and other minerals. The most famous crystals up to 50 mm (2 in) across come from cavities in dolomite in Lengenbach, Binntal, Switzerland. Tabular crystals occurred also in Sacarîmb , Romania. Botryoidal aggregates, growing on barite crystals, were described from the Yunosawa mine, Japan.

Geocronite, 70 mm, Pribram, Czech Republic

Geocronite

$Pb_{14}(Sb,As)_6S_{23}$

MONOCLINIC ● ●

Properties: C – light lead-gray; S – light lead-gray to gray-blue; L – metallic; D – opaque; DE – 6,4; H – 2,5; CL – good; F – uneven; M – tabular crystals, massive. *Origin and occurrence:* Hydrothermal in ore veins, associated with galena, pyrite, tetrahedrite, barite, fluorite and quartz. Crystals, up to 80 mm ($3\frac{1}{8}$ in) across, come from Pietrasanta, Italy. Crystals up to 90 mm ($3\frac{9}{16}$ in) found in Virgem da Lapa, Brazil. Tabular crystals up to 40 mm ($1\frac{9}{16}$ in) across occurred in the Kilbricken mine, Ireland. Massive aggregates are known from Příbram, Czech Republic.

Zinkenite

$Pb_9Sb_{22}S_{42}$

HEXAGONAL ● ● ●

Properties: C – steel-gray, tarnishing iridescent; S – steel-gray; L – metallic; D – opaque; DE – 5,3; H – 3-3,5; CL – imperfect; F – uneven; M – thin prismatic striated crystals, radial to felt-like aggregates, massive.
Origin and occurrence: Hydrothermal in ore veins, associated with stibnite, jamesonite, boulangerite, bournonite, stannite and other minerals. Crystals up to 50 mm (2 in) across are known from the Itos and San Jose mines in Oruro, Bolivia. It was also found in Wolfsberg, Germany; St. Pons, France; Sacarîmb and Baia Sprie, Romania.

Jamesonite

$Pb_4FeSb_6S_{14}$

MONOCLINIC ● ● ● ●

Properties: C – gray-black, tarnishing iridescent; S – gray-black; L – metallic; D – opaque; DE – 5,8; H – 2,5; CL – good; F – uneven; M – acicular crystals, fibrous and felt-like aggregates, massive.

Zinkenite, 39 mm, Oruro, Bolivia

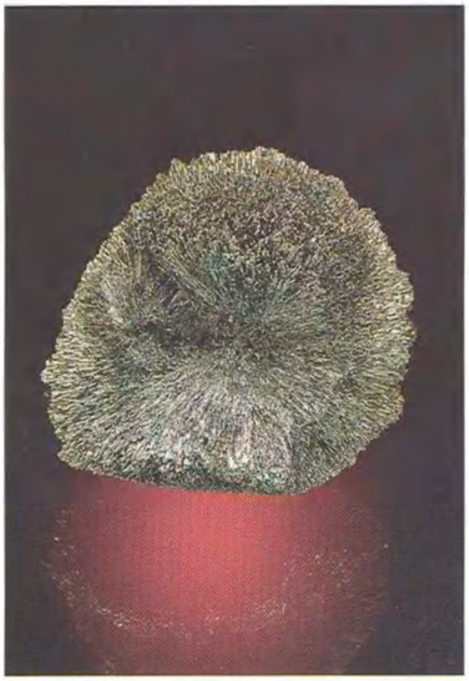

Jamesonite, 30 mm xx, Sombrerete, Mexico

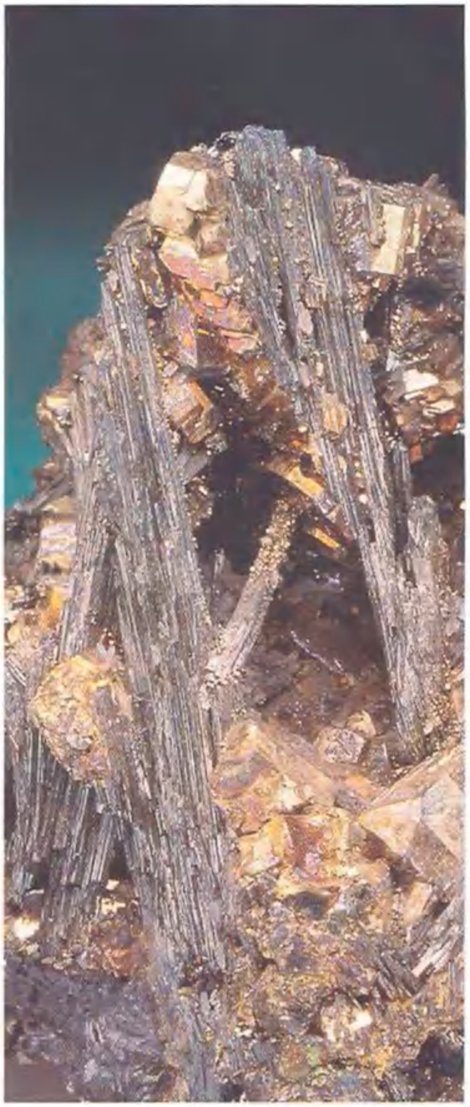

Origin and occurrence: Hydrothermal in medium- and low-temperature base metal ore veins, associated with other Pb-sulfosalts, pyrite, sphalerite, galena, tetrahedrite, quartz and other minerals. Needles up to 80 mm (3⅛ in) long occur in many localities in Bolivia (Tasna; Bolivia mine, Poopo; San Jose and Itos mines, Oruro). It also comes from Wolfsberg and Freiberg, Germany; Nizna Slana . Slovakia and Sombrerete, Zacatecas, Mexico.
Application: Pb and Sb ore.

Semseyite

$Pb_9Sb_8S_{21}$

MONOCLINIC ● ●

Properties: C – gray to black; S – black; L – metallic; D – opaque; DE – 6,1; H – 2,5; CL – perfect; M – tabular and prismatic crystals and their rosette-like aggregates.
Origin and occurrence: Hydrothermal in medium-temperature veins, associated with jamesonite, bournonite, zinkenite, sphalerite and other minerals. Fine rosette-like aggregates of crystals over 10 mm ($^3/_8$ in) across come from Baia Sprie, Herja and Rodna, Romania. It was also found in the San Jose mine, Oruro, Bolivia; Wolfsberg, Germany and Huancavelica, Peru.

Boulangerite

$Pb_5Sb_4S_{11}$

MONOCLINIC ● ● ● ●

Properties: C – lead-gray; S – brownish; L – metallic to silky; D – opaque; DE – 6,2; H – 2,5-3; CL – good; M – acicular striated crystals, fibrous and felt-like aggregates.
Origin and occurrence: Hydrothermal in low- and medium-temperature ore veins, together with other Pb sulfosalts, galena, sphalerite and other minerals. Fine needles over 100 mm (4 in) long come from cavities in quartz in Příbram, Czech Republic. It is common in the Coeur d'Alene district, Idaho, USA. Acicular crystals up to 30 cm (12 in) long were found in Trepca, Serbia and Leadville, Colorado, USA. It is also known from Wolfsberg, Germany and Bolivia (Colquechaca, Huanuni, Isca-Isca).

Cosalite

$Pb_2Bi_2S_5$

ORTHORHOMBIC ● ● ●

Properties: C – lead- to steel-gray, silver-white; S – black; L – metallic; D – opaque; DE – 7,1; H – 2,5-3; CL – none; F – uneven; M – prismatic to acicular crystals, radial and fibrous aggregates.
Origin and occurrence: Magmatic in pegmatites, hydrothermal in medium-temperature deposits; also metamorphic, associated with sphalerite, chalco-

Semseyite, 40 mm, Cavnic, Romania

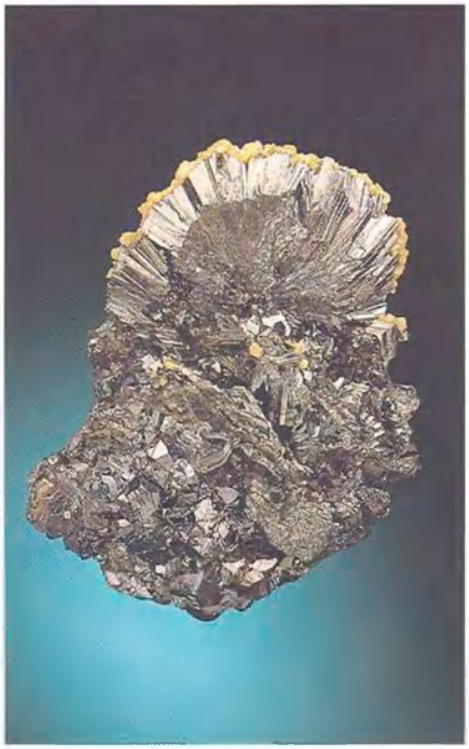

Boulangerite, 40 mm, Zacatecas, Mexico

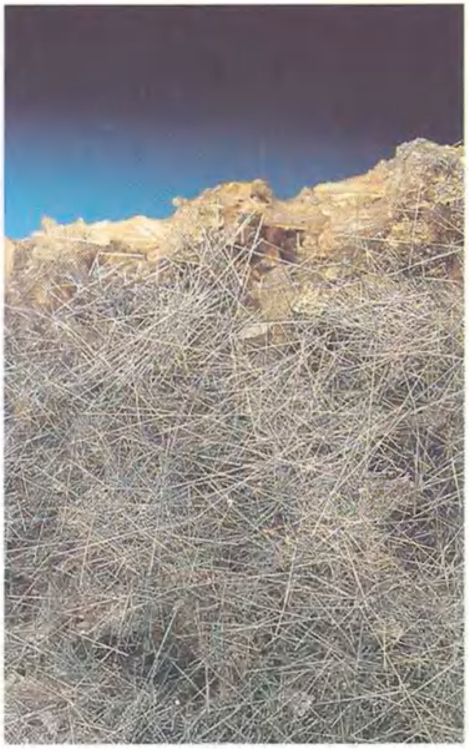

Cosalite, 3 mm xx, Ocna de Fier, Romania

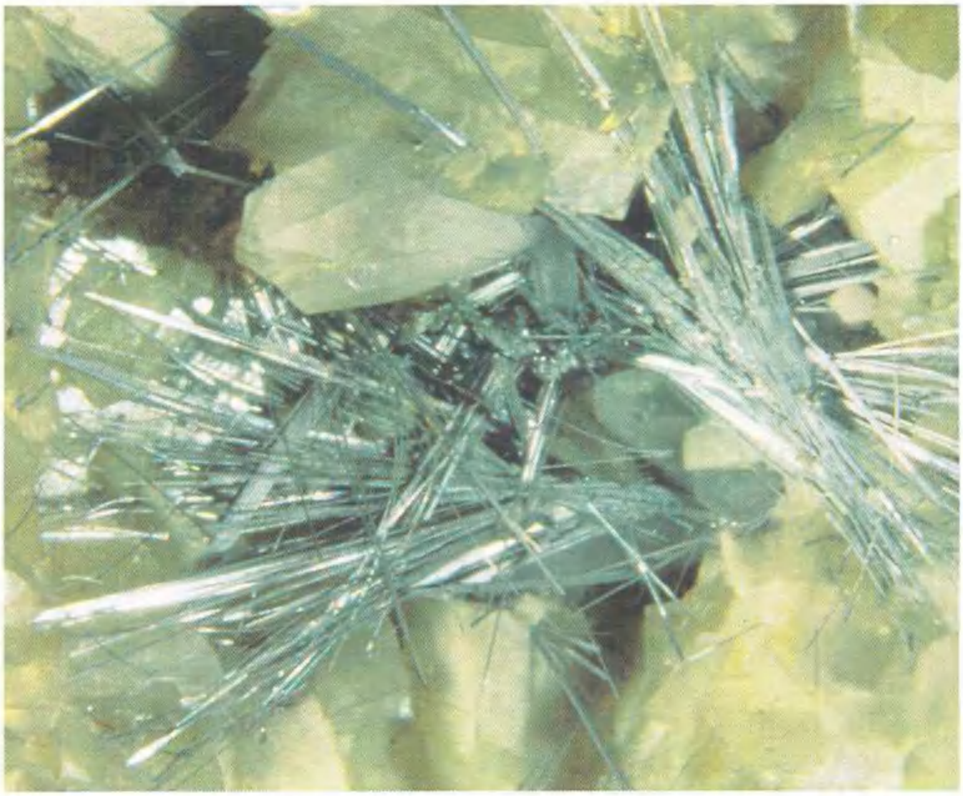

pyrite, pyrite, cobaltite and other minerals. Elongated crystals are known from Crodo, Italy. Fine needles up to 40 mm (1⁹/₁₆ in) long, included in quartz crystals, were found in Kara-Oba, Kazakhstan. It occurs also in Au deposits (Homestake mine, South Dakota, USA) or in skarns (Baiţa Bihorului, Romania).

Kobellite

$Pb_{22}Cu_4(Bi,Sb)_{30}S_{69}$

ORTHORHOMBIC ● ● ●

Properties: C – black-gray to steel-gray; S – black; L – metallic; D – opaque; DE - 6,5; H - 2,5-3; CL – good; M – fibrous aggregates, granular, massive.

Origin and occurrence: Hydrothermal in high-temperature veins and pegmatites, together with cobaltite, arsenopyrite, chalcopyrite and other minerals. It is known from a sulfide rich pegmatite in the Superior Stone quarry, North Carolina, USA. Massive aggregates are common in Jedl'ovec , Slovakia. It was originally described from the Hvena mine near Åskersund, Sweden.

Kobellite, 20 mm xx, Rožňava, Slovakia

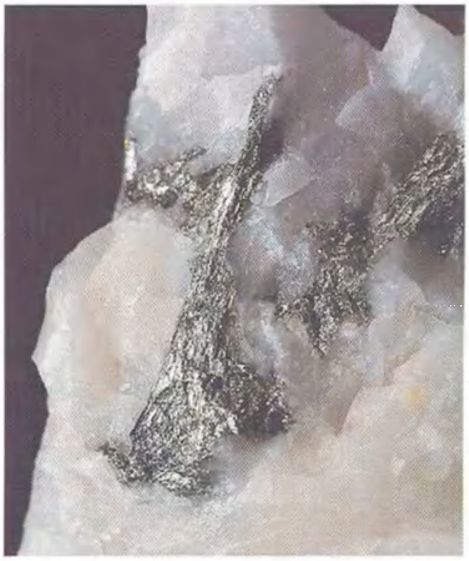

Realgar, 64 mm, Shimen, China

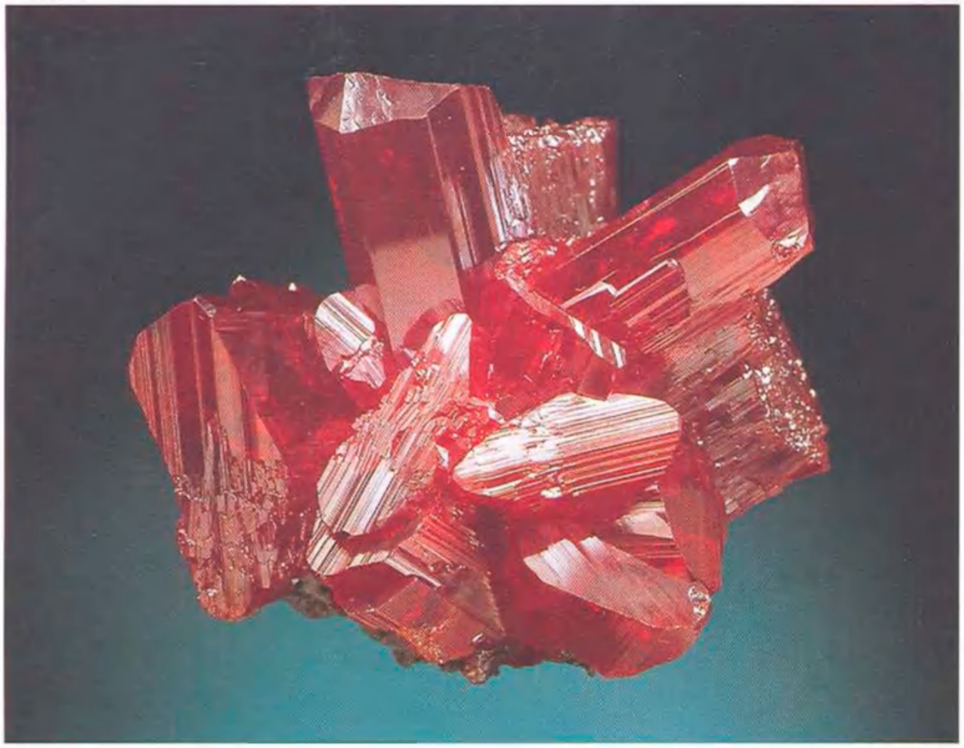

Realgar
AsS

MONOCLINIC ● ● ● ●

Properties: C – red to orange-yellow; S – orange-red to red; L – resinous to greasy; D – transparent to translucent; DE – 3,6; H – 1.5-2; CL – good; F – conchoidal; M – prismatic striated crystals, granular, massive.

Orpiment, 40 mm, Baia Sprie, Romania

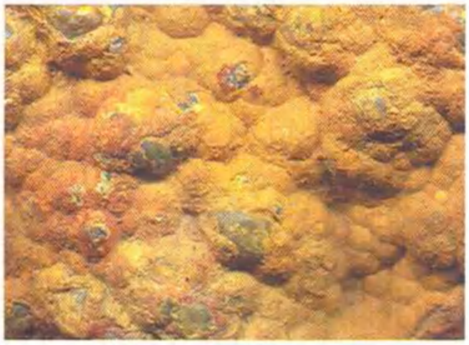

Origin and occurrence: Hydrothermal in low-temperature veins, associated with other As-Sb minerals; also as a sublimation product of volcanic gasses, in hot springs and sediments. The most beautiful crystals over 100 mm (4 in) long come from Shimen, Hunan, China. Crystals up to 70 mm (2¾ in) long occurred in the Getchell mine, Nevada, USA. Crystals up to 80 mm (3⅛ in) long were found in Lengenbach, Binntal, Switzerland. Crystals are also known from Baia Sprie, Romania. Massive aggregates are common in Allchar, Macedonia.

Orpiment
As_2S_3

MONOCLINIC ● ● ● ●

Properties: C – lemon-yellow to bronze-yellow; S – light lemon-yellow; L – resinous to pearly; D – transparent to translucent; DE – 3,5; H – 1,5-2; CL – perfect; M – prismatic crystals, foliated and fibrous aggregates. *Origin and occurrence:* Hydrothermal in low-temperature veins, together with realgar, stibnite, calcite etc., also from hot springs and fumaroles. It is also a common product of realgar oxidation. The best crystals up to 100 mm (4 in) long come from Shimen, Hunan, China. Fine cleavable lamellae occur in Lukhumi, Georgia and Men-Kyule, Yakutia, Russia. Crystals up to 50 mm (2 in) across found in the La Libertad mine, Quiruvilca and Huayllapon, Ancash, Peru. Crystals up to 80 mm (3⅛ in) long described from the Getchell mine, Nevada, USA.. Fine specimens are also known from Allchar, Macedonia and Khaidarkan, Kyrgyzstan. *Application:* As ore, pigment.

Getchellite

$AsSbS_3$

MONOCLINIC ● ●

Properties: C – dark red, tarnishing green and iridescent; S – orange-red; L – pearly to glassy, resinous; D – transparent; DE – 4,0; H – 1,5-2; CL – perfect; F – splintery; M – imperfect curved crystals, massive.
Origin and occurrence: Hydrothermal in low-temperature ore deposits, associated with orpiment, realgar, stibnite, cinnabar and other minerals. It was described from the Getchell mine, Nevada, USA. Beautiful specimens with grains up to several cm across come from Khaidarkan, Kyrgyzstan. It is also known from Zarehshuran, Kurdistan, Iran.

Getchellite, 60 mm, Khaidarkan, Kyrgystan

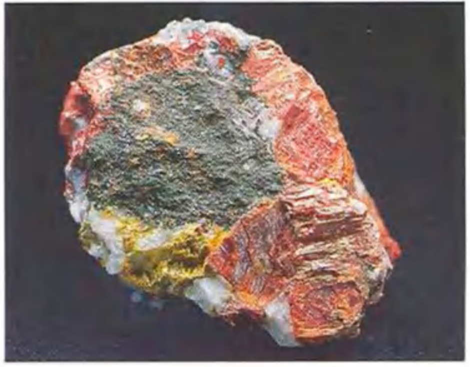

Orpiment, 35 mm, Huayllapon, Peru

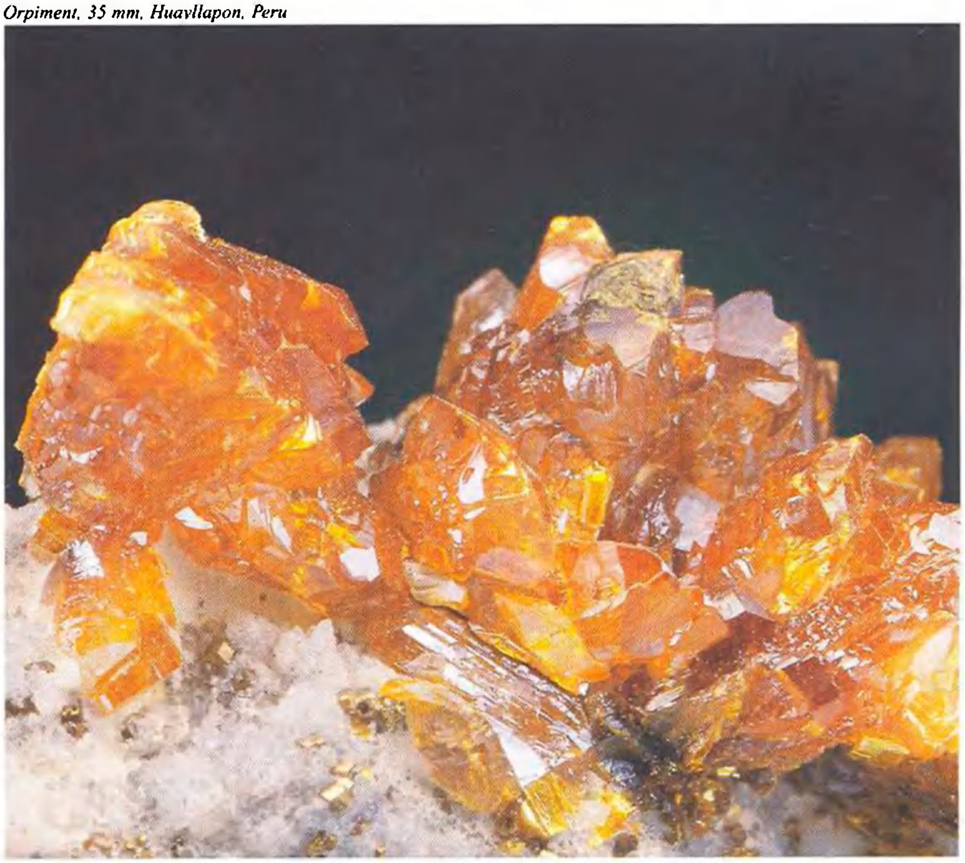

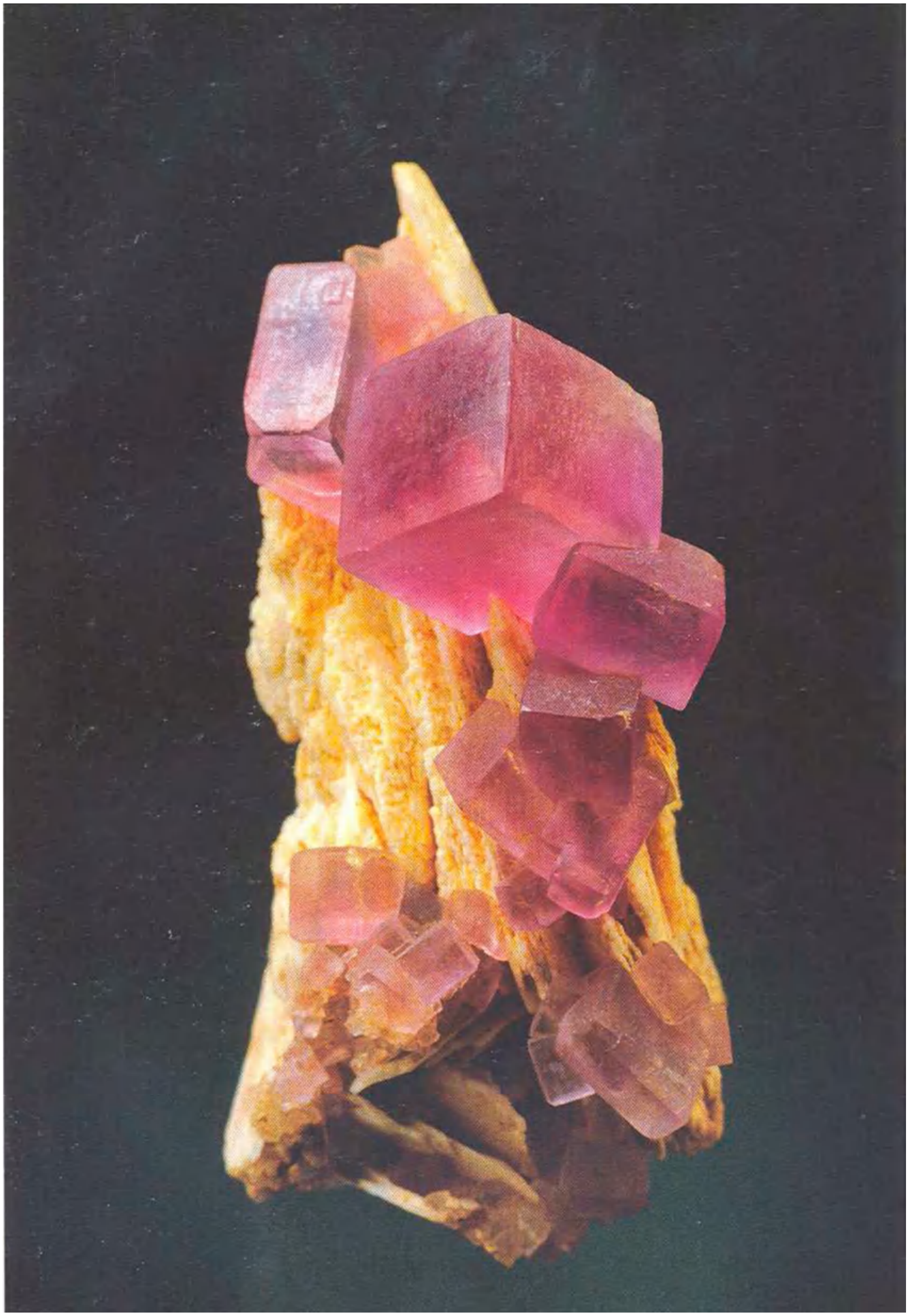

3. Halides

Fluellite

$Al_2(PO_4)F_2(OH) \cdot 7H_2O$

ORTHORHOMBIC ● ●

Properties: C – colorless, white, yellow; S – white; L – vitreous; D – transparent; DE – 2.2; H – 3; CL – imperfect; M – dipyramidal crystals.
Origin and occurrence: Hydrothermal in greisens, also secondary as a result of triplite alteration. Crystal druses over 10 mm (3/8 in) in size come from Horní Slavkov, Czech Republic. Very similar specimens were found in Stenna Gwyn near St. Austell, Cornwall, UK. Also found in pegmatites in Kynzvart, Czech Republic and in Hagendorf, Germany as a product of triplite alteration.

Cryolite

Na_3AlF_6

MONOCLINIC ● ● ●

Properties: C – colorless, white, purple, brownish; S – white; L – greasy to pearly; D – transparent to translucent; DE – 3; H – 2.5; CL – none; F – uneven; M – pseudo-cubic crystals, massive.

Fluellite, 10 mm x, Horni Slavkov, Czech Republic

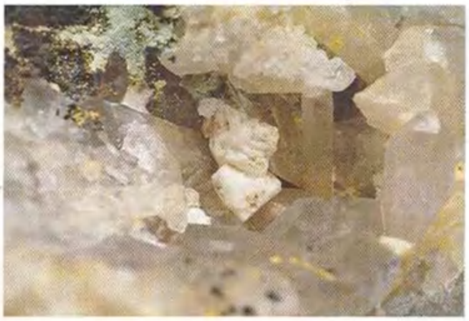

Origin and occurrence: Characteristic mineral of the cryolite pegmatites. Crystals up to 30 mm (1 3/16 in) in size were found in Ivigtut, Greenland, where it was mined as the Al ore for more than 100 years. It was associated with other aluminofluorides, sphalerite, cassiterite, ferrocolumbite and other minerals. It is also known from the Francon quarry in Montreal, Quebec, Canada in crystals up to 10 mm (3/8 in) across. Massive cryolite occurs in Miass, Ural Mountains, Russia and in St. Peter's Dome, Pikes Peak batholith, Colorado, USA.
Application: it was an important Al ore.

Fluorite, 67 mm, Berbes, Spain
Cryolite, 35 mm, Ivigtut, Greenland

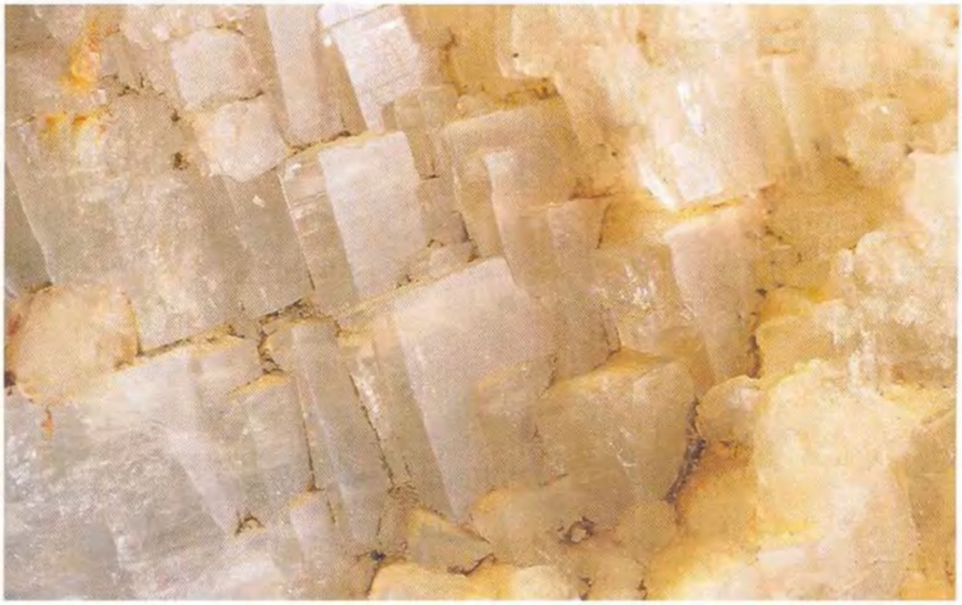

Creedite, 70 mm, Santa Eulalia, Mexico

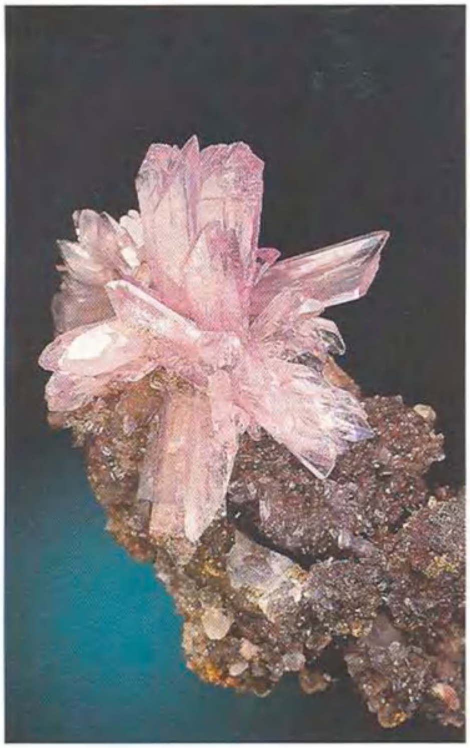

Creedite

$Ca_3Al_2(SO_4)(F,OH)_{10} . 6 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, purple; S – white; L – vitreous; D – transparent; DE – 2.7; H – 4; CL – perfect; F – conchoidal; M – short prismatic to acicular crystals, granular, massive.
Origin and occurrence: Hydrothermal, associated with fluorite and barite. Purple crystals, several cm long, come from Wagon Wheel Gap near Creede, Colorado, USA. Nice druses were found in Santa Eulalia, Chihuahua, Mexico. The best creedite specimens with purple crystals up to 30 mm ($1^{3}/_{16}$ in) long were recently found in Akcha-tau, Kazakhstan.

Carnallite

$KMgCl_3 . 6 H_2O$

ORTHORHOMBIC ● ● ● ●

Properties: C – colorless, white, yellowish, red, blue; S – white; L – vitreous to greasy; DE – 1.6; H – 1-2; CL – none; F – conchoidal; M – pseudo-hexagonal pyramidal and tabular crystals, granular; LU – strong; R – decomposes under wet conditions.
Origin and occurrence: Sedimentary, one of the last products of evaporation of salty solutions; also supergene as a product of a reaction of older salts with solutions, rich in potassium, associated with halite, sylvite and other minerals. Crystals up to 40 mm ($1^{9}/_{16}$ in) across are known from the vicinity of Carlsbad, New Mexico, USA. Nice crystals come also from Stassfurt and Alexanderhall, Germany. Massive aggregates are common in many salt deposits, like in Saskatchewan, Canada; in Kalush, Ukraine and elsewhere.
Application: : the most important potassium salt, used as fertilizer and for production of metal Mg.

Atacamite

$Cu_2Cl(OH)_3$

ORTHORHOMBIC ● ● ●

Properties: C – emerald-green, black-green; S – green; L – vitreous; D – translucent; DE – 3.8; H – 3-3.5; CL – perfect; F – conchoidal; M – prismatic crystals, columnar, radial and lamellar aggregates, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits in the arid climate, associated with other Cu minerals. Crystals up to 10 mm ($^{3}/_{8}$ in) long were described from Burra district, Southern Australia, Australia. Crystals up to 10 mm ($^{3}/_{8}$ in) across come from Tsumeb, Namibia; also from Bisbee, Arizona, USA. Rich aggregates of acicular crystals occur in many localities in Atacama province, Chile (Copiapo, Remolinos).

Carnallite, 80 mm, Merkers, Germany

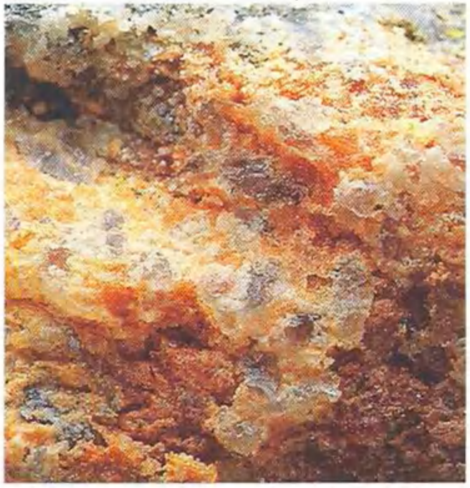

Atacamite, 78 mm, Atacama, Chile

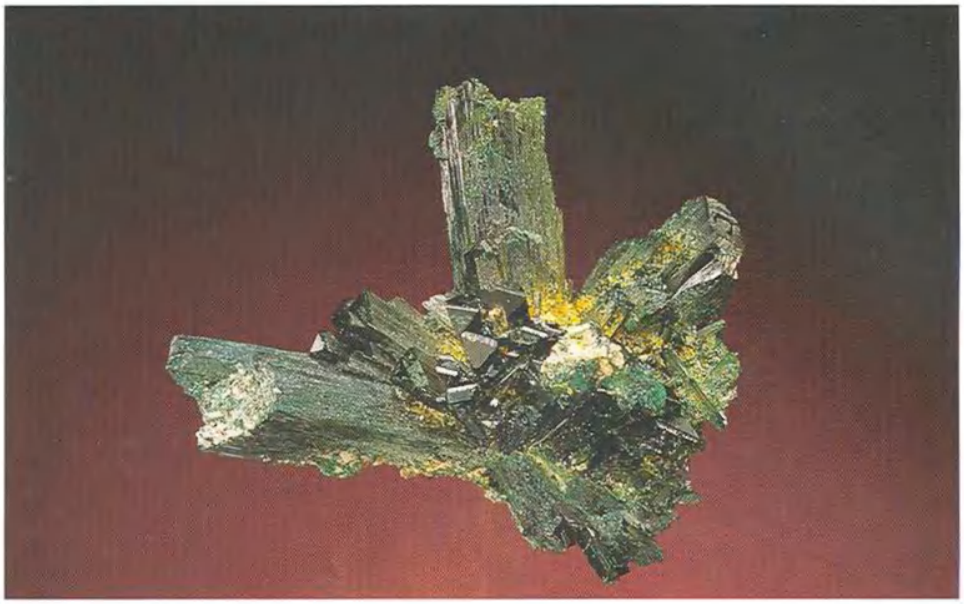

Boleite

$Pb_{26}Ag_9Cu_{24}Cl_{62}(OH)_{48}$

CUBIC ●

Properties: C – blue; S – blue; L – pearly; D – translucent; DE – 5.1; H – 3-3.5; CL – perfect; M – cubic crystals; R – soluble in water.

Origin and occurrence: Secondary, originated in the oxidation zone of Cu deposits in the arid climate. By far the best specimens were found in Boleo, Baja California, Mexico, where cubes up to 25 mm (1 in) in size were found. It is also known from Phillipsburg, Montana, USA and Challacollo, Chile.

Boleite, 65 mm, Santa Rosalia, Mexico

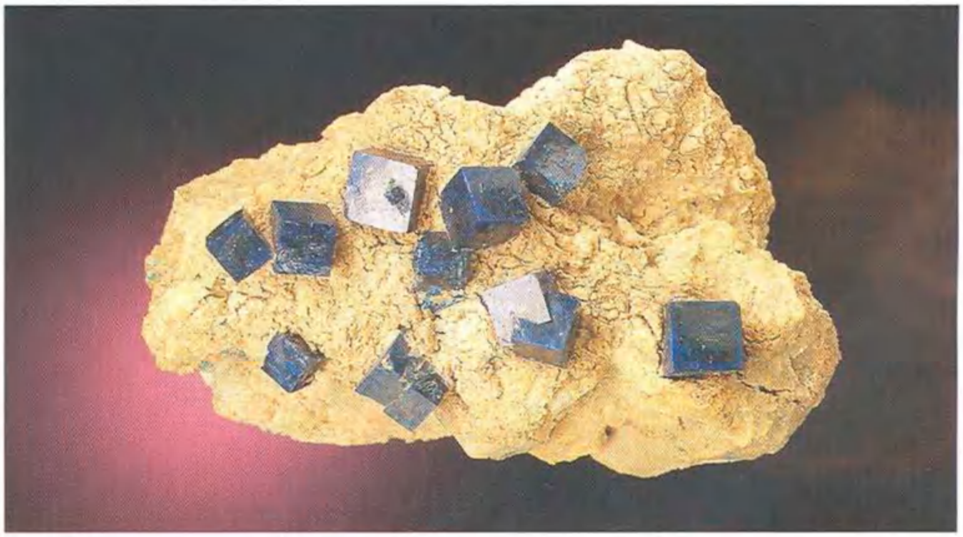

Cumengite, 21 mm, Boleo, Mexico

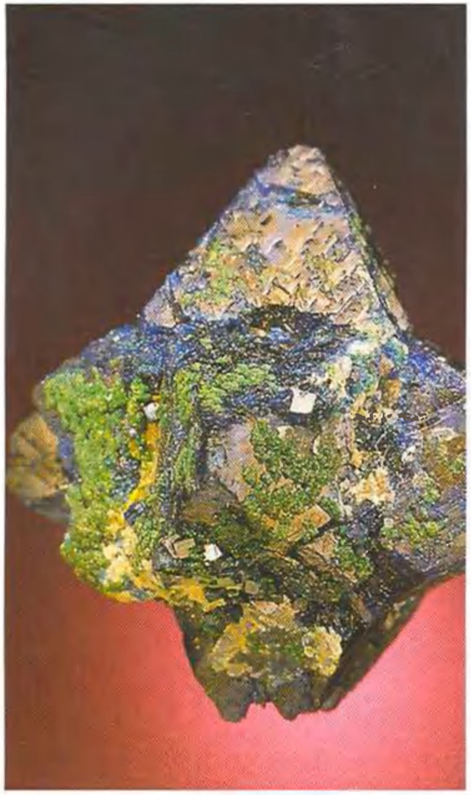

Cumengite

$Pb_{21}Cu_{20}Cl_{42}(OH)_{40}$

TETRAGONAL ●

Properties: C – indigo-blue; S – blue; L – vitreous; D – translucent; DE – 4.7; H – 2.5; CL – good; M – tetragonal pyramidal crystals, also epitaxially overgrown on boleite cubes. *Origin and occurrence:* Secondary in the oxidation zone of Cu deposits in the arid climate, associated with boleite. The largest crystals up to 35 mm (1⅜ in) are known from Boleo, Baja California, Mexico. It is also reported from Newport Beach near Falmouth, Cornwall, UK.

Iodargyrite

β-AgI

HEXAGONAL ● ● ●

Properties: C – colorless, tarnishes to yellow; S – yellow; L – adamantine; D – transparent to translucent; DE – 5.7; H – 1-1.5; CL – perfect; F – conchoidal; M – prismatic to tabular crystals, granular, massive.
Origin and occurrence: Secondary, product by oxidation of Ag ores, with other Ag minerals. Common greenish crystals over 10 mm (⅜ in) in size occur in Broken Hill, New South Wales, Australia. Also found in Vrancice, Czech Republic; Tonopah, Nevada, USA; Chanarcillo and Copiapo, Chile.

Villiaumite

NaF

CUBIC ● ● ●

Properties: C – dark red; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 2-2.5; CL – perfect; M – small crystals, granular, massive; R – soluble in water.
Origin and occurrence: Late mineral in cavities in alkaline igneous rocks (nepheline syenites). Crystals, several cm long, are known from the Rasvumchorr Mountain, Khibiny massif, Kola Peninsula, Russia; only slightly smaller crystals come from Mont St.-Hilaire, Quebec, Canada and Illimaussaq, Greenland.

Iodargyrite, 2 mm xx, Rudabánya, Hungary

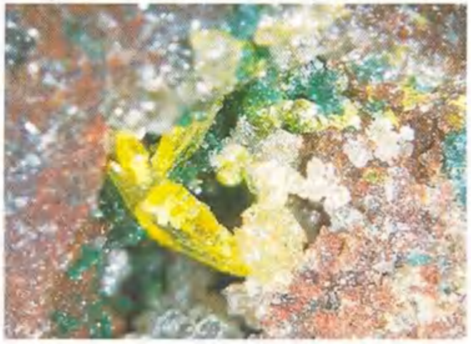

Villiaumite, 10 mm xx, Khibiny Massif, Kola, Russia

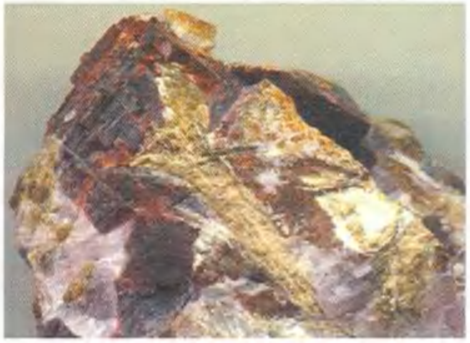

Halite, 55 mm, Trona, U.S.A.

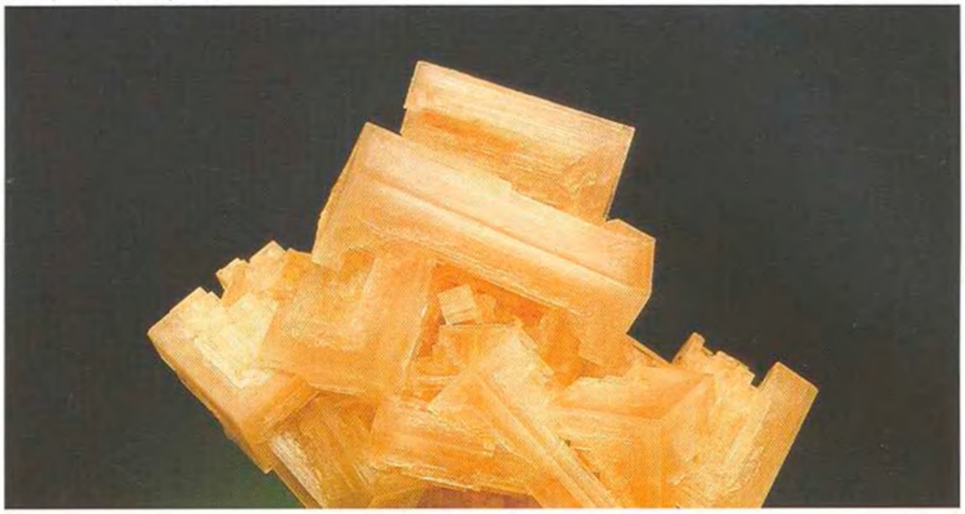

Halite

NaCl

CUBIC ● ● ● ●

Properties: C – colorless, gray, white, red, blue; S – white; L – vitreous; DE – 2.2; H – 2; CL – perfect; F – conchoidal; M – cubic crystals, granular, massive; R – soluble in water.

Origin and occurrence: Product of high-temperature fumaroles (Etna, Mt. Vesuvius; Italy); mainly sedimentary, as a result of evaporation of sea water, associated with sylvite, carnallite and other minerals. Very fine cubes over 10 mm (3/8 in) are known from Weliczka and Bochnia, Poland. Blue cleavable aggregates are found in Bernburg, Germany. Salt deposits in Austria (Hallstatt, Hallein) are also important. Huge halite deposits, associated with potassium salts are mined in the vicinity of Stassfurt, Germany. Fine skeletal crystals are known from many localities in California, USA.

Application: food and chemical industries.

Halite, 18 mm xx, Sonora, Mexico

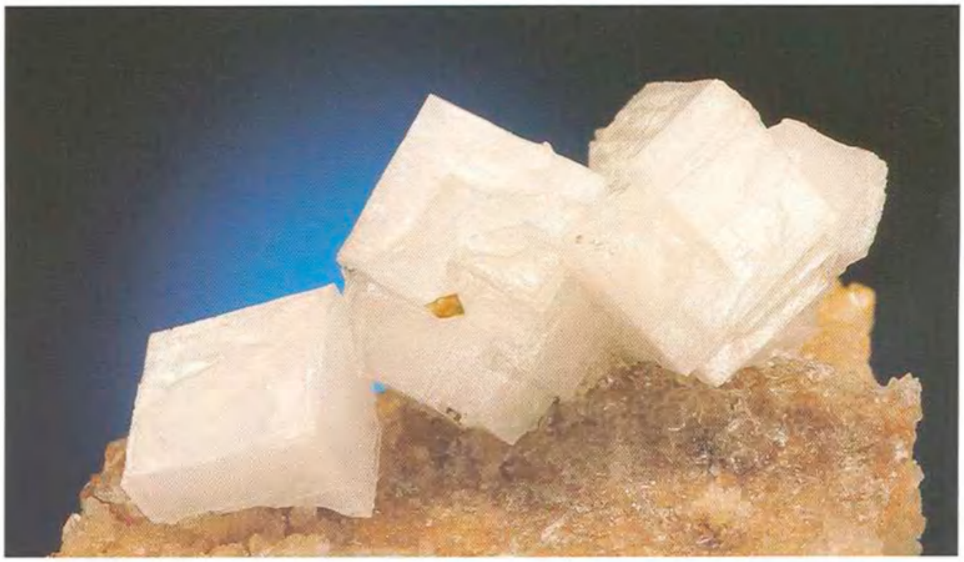

Sylvite, 77 mm, California, U.S.A.

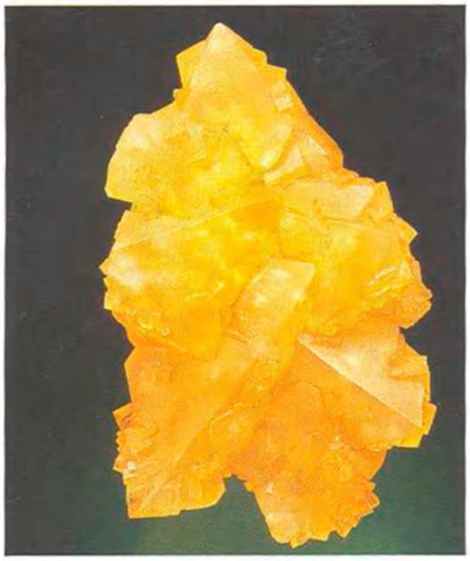

Sylvite

KCl

CUBIC ● ● ●

Properties: C – white, gray, blue, red; S – white; L – vitreous; D – transparent; DE – 2; H – 2; CL – perfect; F – uneven; M – cubic crystals and their combinations; granular, massive; R – soluble in water.
Origin and occurrence: Sedimentary as a result of evaporation of sea water, together with halite, carnallite and other minerals. Nice cubes up to 50 mm (2 in) come from Stassfurt; it forms stalactites in Wathlingen, Germany. Crystals are also known from Kalush, Ukraine and from Salton Sea, California, USA, where it occurs as octahedra on halite crystals.
Application: chemical industry.

Calomel, 30 mm, Terlingua, U.S.A.

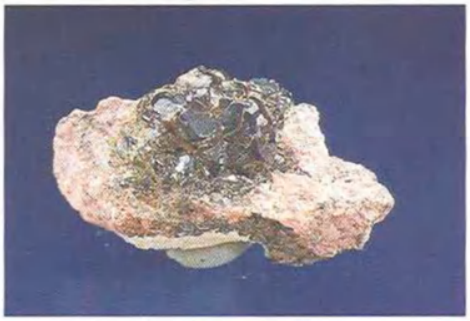

Sal ammoniac, 10 mm x, Batonyterenye, Hungary

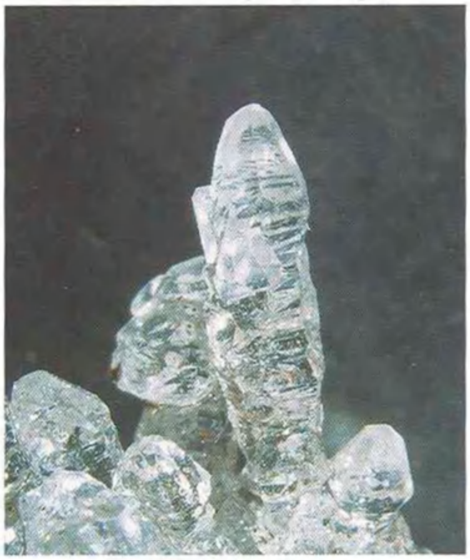

Sal ammoniac

NH_4Cl

CUBIC ● ● ●

Properties: C – colorless, white, gray, yellow, brown; S – white; L – vitreous; D – transparent; DE – 1.5; H – 1-2; CL – imperfect; F – conchoidal; M – combinations of cubic crystals, dendritic and skeletal aggregates; earthy.
Origin and occurrence: Typical mineral for fumaroles and burning coal dumps, associated with sulfur and other minerals. Complicated crystals are known from Mt. Vesuvius, Etna and Vulcano, Italy. Crystals over 10 mm (3/8 in) in size occurred on burning coal dumps near Kladno, Czech Republic; similar from localities in Eastern Pennsylvania, USA and near Ste-Etienne, France.

Calomel

HgCl

TETRAGONAL ● ●

Properties: C – colorless, white, gray, brown, it darkens on air; S – white; L – adamantine; D – transparent to translucent; DE – 7.2; H – 1.5; CL – good; F – conchoidal; M – tabular to pyramidal crystals, coatings, earthy; LU – dark red.
Origin and occurrence: Secondary as a result of alteration of Hg minerals, associated with cinnabar, mercury and other minerals. Crystals were found in Moschellandsberg, Germany; Avala, Serbia;

Khaidarkan, Kyrgyzstan and Terlingua, New Mexico, USA.

Fluorite

CaF_2

CUBIC ● ● ● ● ●

Properties: C – colorless, white, yellow, red, green, blue, purple, brown, black; S – colorless; L – vitreous; D – transparent to translucent, opaque; DE – 3.2; H – 4; CL – perfect; F – conchoidal to splintery; M – combinations of cubic crystals, granular, massive; LU – blue, blue-green, also phosphorescent.

Origin and occurrence: Rare magmatic, mainly hydrothermal and metasomatic. Associations are very diverse, depending on a type of the deposit, in which it occurs. Beautiful crystals are known from many localities all over the world. Pink octahedra, several cm in size, are known from pegmatites in Nagar, Pakistan. Nice crystals were also found in greisens in Cornwall, UK (Wheal Mary mine) and from Horni Slavkov, Czech Republic. Beautiful green cubes up to 20 cm across and colorless cubes up to 10 mm (3/8 in) across from Dalnegorsk, Russia are of hydrothermal origin. Famous green and purple crystals come from Alston Moor and Weardale, England, UK. Nice pink octahedra up to 30 mm (1 3/16 in) occurred in Huanzala, Peru. Beautiful yellow cubes up to 50 mm (2 in) associated with barite, are known from Halsbrücke and Annaberg, Germany. Purple complex combinations of crystals come from La Collada, Spain. Mainly purple cubes up to 10 mm (3/8 in) occurred in Rosiclare, Illinois; in association with honey-yellow calcite crystals are known from the Elmwood mine, Tennessee; similar occurrences are also in several localities in Kentucky, USA. The most valuable fluorite specimens are pink octahedra up to 150 mm (6 in) from Göschenen, Switzerland; Mont Blanc massif, France and other Alpine localities.

Application: metallurgy, chemical industry, special optics.optika.

Fluorite, 50 mm, Argentina

Fluorite, 23 mm x, Chamonix, France

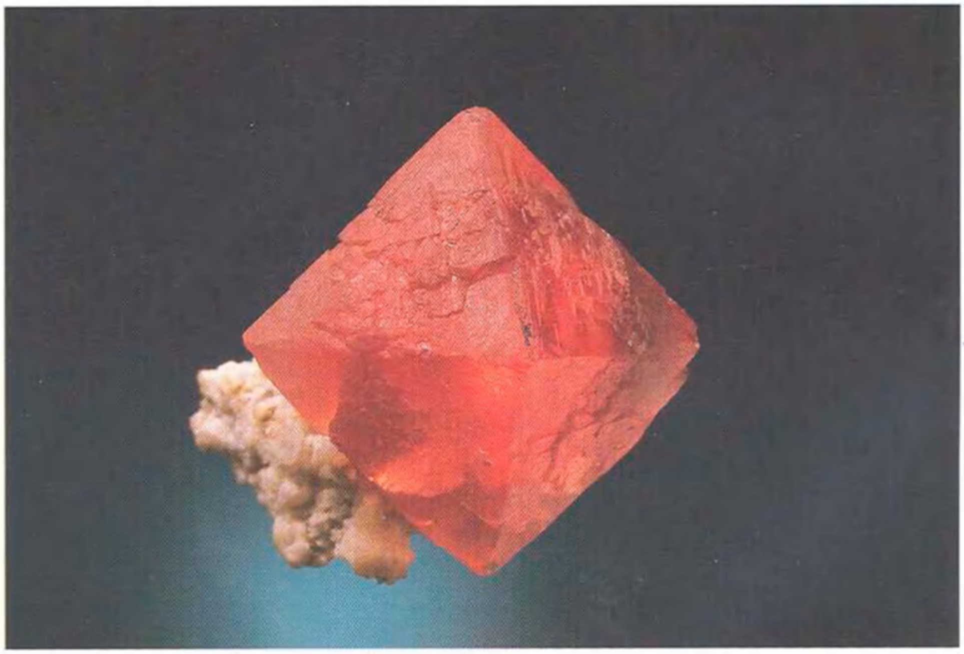

4. Oxides

Cuprite

Cu_2O

CUBIC ● ● ●

Varieties: chalcotrichite (acicular to hair-like crystals)
Properties: C – red; S – red; L – adamantine to submetallic; D – transparent to translucent; DE – 5.8-6.2; H – 3.5-4; CL – imperfect; F – conchoidal to uneven; M – combinations of cubic crystals, hair-like aggregates, granular, massive.
Origin and occurrence: Secondary, as a result of the oxidation of Cu sulfides. Crystals up to 150 mm (6 in) in size, covered with malachite, occur in Onganja, Namibia. Shiny octahedra, up to 40 mm (1$^{9}/_{16}$ in) in size, come from the Mashamba West mine, Zaire. Acicular and fibrous crystals of the chalcotrichite variety were found in Bisbee, Arizona, USA. Combinations of cubic crystals, covered with malachite, are known from Chessy near Lyon, France. Crystals up to 30 mm (1$^{3}/_{16}$ in) across are reported from Tsumeb, Namibia.
Application: important Cu ore.

Zincite

ZnO

HEXAGONAL ● ●

Properties: C – yellow, orange, red; S – orange-yellow; L – adamantine; D – transparent to translucent; DE – 5.7; H – 4.5-5; CL – perfect; F – conchoidal; M – pyramidal crystals, granular, massive.
Origin and occurrence: Metamorphic, associated with willemite and franklinite. It forms very rare crystals up to 40 mm (1$^{9}/_{16}$ in) in size in the metamorphosed Zn deposits in Franklin and Sterling Hill, New Jersey, USA, it is mostly granular and massive. Zincite crystals and aggregates of vitreous luster from Poland, which are offered at the mineral shows, are not of a natural origin, there are smelter products.

Zincite, 11 mm x, Franklin, U.S.A.

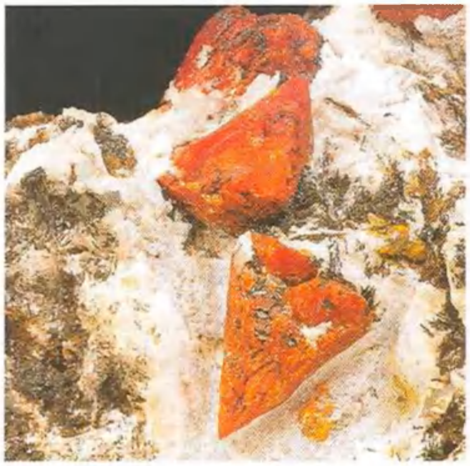

Amethyst, 122 mm, Guerrero, Mexico
Cuprite, 18 mm x, Mashamba West, Zair

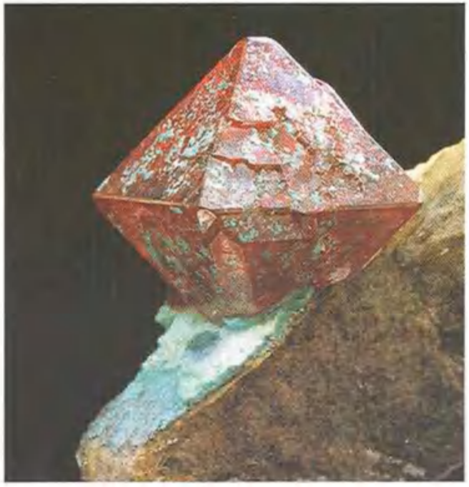

Chalcotrichite, 50 mm, Bisbee, U.S.A.

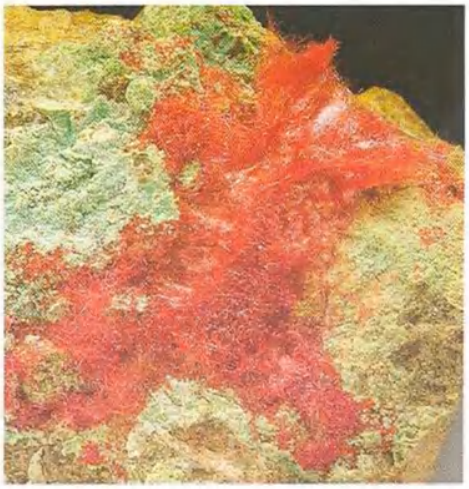

Tenorite

CuO

MONOCLINIC ● ● ●

Properties: C – steel-gray to black; S – gray; L – metallic; D – opaque; DE – 6.5; H – 3.5; CL – imperfect; F – uneven to conchoidal; M – thin tabular to scaly crystals, earthy, massive.

Origin and occurrence: Secondary, in the oxidation zone of Cu deposits, together with other Cu supergene minerals. It was common in Cu deposits in the Keweenaw Peninsula, Michigan, USA. It was mined as Cu ore in Bisbee, Globe and Morenci, Arizona, USA. Thin tabular crystals are known from Tsumeb, Namibia.

Tenorite, 3 mm xx, Bisbee, U.S.A

Spinel

$MgAl_2O_4$

CUBIC ● ● ●

Varieties: pleonast (black)

Properties: C – pink, red, green, blue, brown, black; S – white; L – vitreous to dull; D – transparent to opaque; DE – 3.6; H – 7.5-8; CL – imperfect; F – conchoidal to uneven; M – octahedral crystals, granular, massive.

Origin and occurrence: Magmatic, metamorphic, also in placers, associated with corundum, sillimanite and other minerals. Large pleonast crystals reaching up to 150 mm (6 in) were found in the Aldan massif, Yakutia, Russia. Fine crystals weighing up to 14 kg (30 lb 12 oz) come from Amity, New York. Crystals up to 120 mm (4$^{7}/_{16}$ in) across are known from Sterling Hill, New Jersey. Blue crystals of spinel come from Bolton, Massachusetts, USA and South Burgess, Ontario, Canada. Gemmy pink and red crystals up to 20 mm ($^{25}/_{32}$ in) reported near Ratnapura, Sri Lanka and in Mogok, Burma. Fine pink crystals up to 50 mm (2 in) in size occur in Kukh-i-lal, Tajikistan.

Application: gemstone.

Spinel, 10 mm x, Mogok, Burma

Gahnite

$ZnAl_2O_4$

CUBIC ● ● ●

Properties: C – black-green, black; S – gray; L – vitreous to greasy; D – translucent to opaque; DE –

Pleonast, 15 mm xx, Vietnam

Gahnite, 90 mm, Broken Hill, Australia

4.4-4.6; H – 7.5-8; CL – imperfect; F – conchoidal to uneven; M – octahedral crystals, granular.
Origin and occurrence: Magmatic and metamorphic, associated with wolframite, chalcopyrite and other minerals. Crystals up to 120 mm (4$\frac{7}{16}$ in) in size come from Franklin and Sterling Hill, New Jersey, USA. Crystals from Broken Hill, New South Wales, Australia reached up to 30 mm (1$\frac{3}{16}$ in). Blue-green crystals were found in Rowe, Massachusetts, USA. Cuttable blue crystals occur near Gidan Wayo, Nigeria.

Magnetite
$Fe^{2+}Fe^{3+}_2O_4$

CUBIC ● ● ●

Properties: C – black; S – black; L – metallic; D – opaque; DE – 5.2; H – 5.5-6.5; CL – none; F – uneven to conchoidal; M – octahedral crystals, granular, massive.
Origin and occurrence: Magmatic, hydrothermal and metamorphic, rare sedimentary. Parageneses differ according to the origin. Fine crystals up to 170 mm (6$\frac{11}{16}$ in) in size, come from Traversella, Italy. A crystal 25 cm (9$\frac{13}{16}$ in) in size was found in Vastanfors, Sweden. Fine crystals up to 40 mm (1$\frac{9}{16}$ in) occur in Dashkesan, Azerbaijan, where it is associated with andradite, epidote and apatite. Beautiful shiny octahedra up to 40 mm (1$\frac{9}{16}$ in) are known from Alpa Lercheltini, Binntal, Switzerland. Rare cubes up to 20 mm ($\frac{25}{32}$ in) on edge come from the ZCA No.4 mine, Balmat, New York, USA.. Magnetite crystals reaching up to 10 cm were found in pegmatites in Jaguaraçu, Minas Gerais, Brazil. Crystals up to 20 cm (7$\frac{7}{8}$ in) in size reported from the Gardiner complex, Greenland. Crystals up to 50 mm (2 in) were lately found in Kovdor, Kola Peninsula, Russia.
Application: Fe ore.

Magnetite, 20 mm x, Chester, U.S.A.

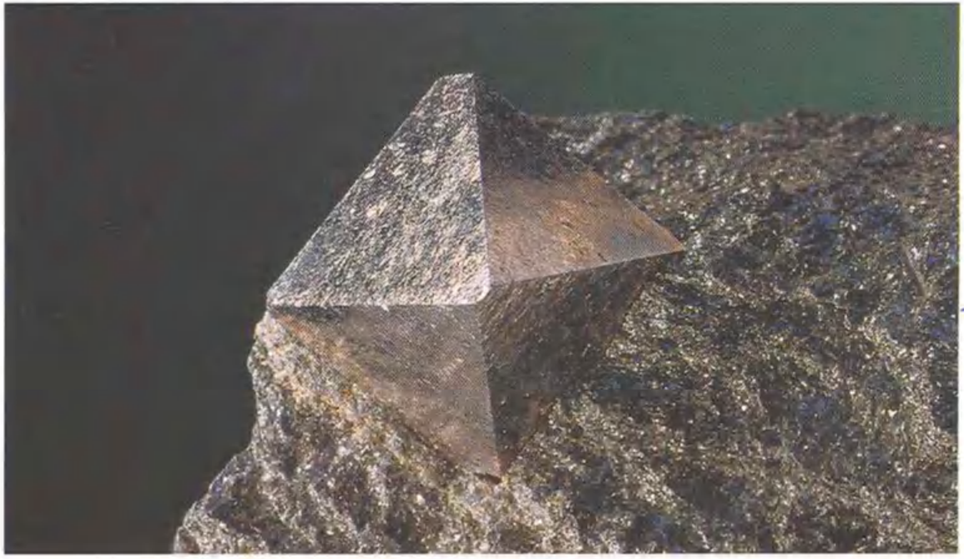

Franklinite, 30 mm, Sterling Hill, U.S.A.

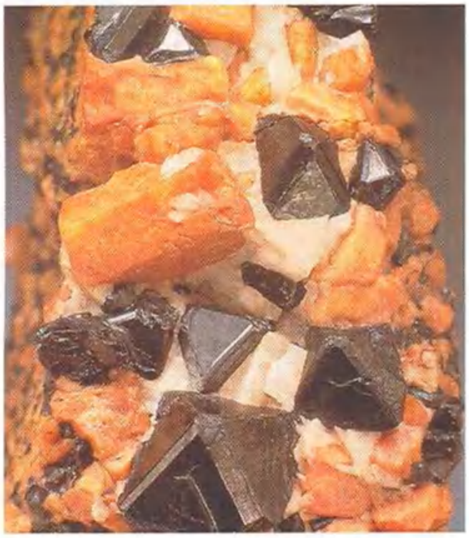

Franklinite

$(Zn,Mn^{2+},Fe^{2+})(Fe^{3+},Mn^{2+})_2O_4$

CUBIC ● ●

Properties: C – black; S – dark brown; L – metallic; D – opaque; DE – 5.1-5.5; H – 6; CL – none; F – uneven to conchoidal; M – octahedral crystals, granular, massive.
Origin and occurrence: Metamorphic, associated with willemite, zincite and other minerals. The only localities, where it is common and occurs in very large accumulations, are Franklin and Sterling Hill, New Jersey, USA. Crystals up to 170 mm ($6^{11}/_{16}$ in) across are known from there. It is rare in Langban, Sweden and Ocna de Fier, Romania.
Application: Zn ore.

Chromite, 87 mm, Finnland

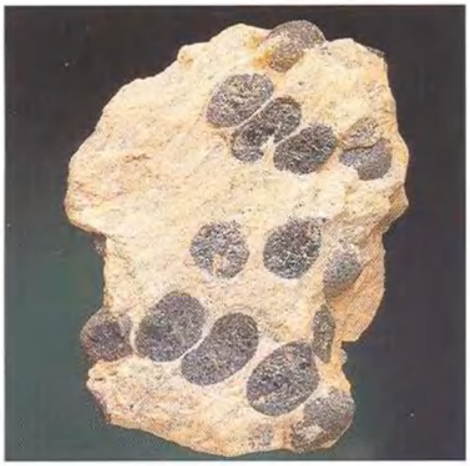

Chromite

$FeCr_2O_4$

CUBIC ● ● ●

Properties: C – black; S – brown; L – metallic; D – opaque; DE – 4.5-4.8; H – 5; CL – none; F – uneven; M – octahedral crystals, granular, massive.
Origin and occurrence: Magmatic, together with magnetite, uvarovite and other minerals. Rare crystals reaching up to 10 mm ($^{3}/_{8}$ in) are known from Uzun Damar, Turkey. It occurs mostly massive, like in deposits in Bushveld, South Africa; in Sarany, Ural Mountains, Russia and in Guleman, Turkey.
Application: Cr ore.

Hausmannite

$Mn^{2+}Mn^{3+}{}_2O_4$

TETRAGONAL ● ● ●

Properties: C – black; S – brown; L – submetallic; D – opaque; DE – 4.8; H – 5.5; CL – perfect; F – uneven; M – pseudo-octahedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in high-temperature Mn deposits, also as a product of the contact metamorphism. The best specimens with crystals up to 30 mm ($1^{3}/_{16}$ in) in size come from the N'Chwaning mine, Kuruman, South Africa. Smaller crystals were found in Ilfeld and Ilmenau, Germany. It also occurred as fine crystals in Langban and Jakobsberg, Sweden.

Minium, 60 mm, Broken Hill, Australia

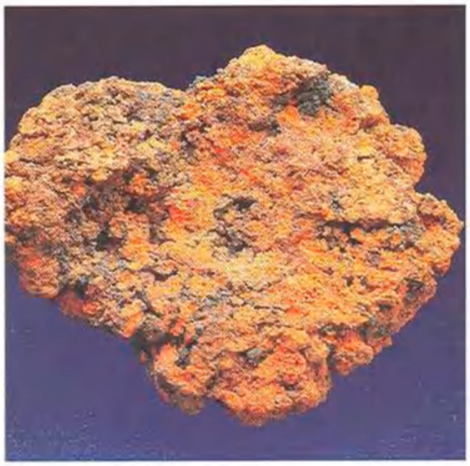

Minium

$Pb^{2+}_2Pb^{4+}O_4$

TETRAGONAL ● ● ●

Properties: C – red, S – orange-yellow; L – dull to greasy; D – opaque; DE – 8.9; H – 2.5; CL – perfect; M – earthy and pulverulent aggregates, massive.

Origin and occurrence: Secondary mineral, as a result of the galena oxidation. It occurs in Langban, Sweden; in Anarak, Iran; in Leadhills, Scotland, and in Broken Hill, New South Wales, Australia.

Hausmannite, 42 mm, Kuruman, South Africa

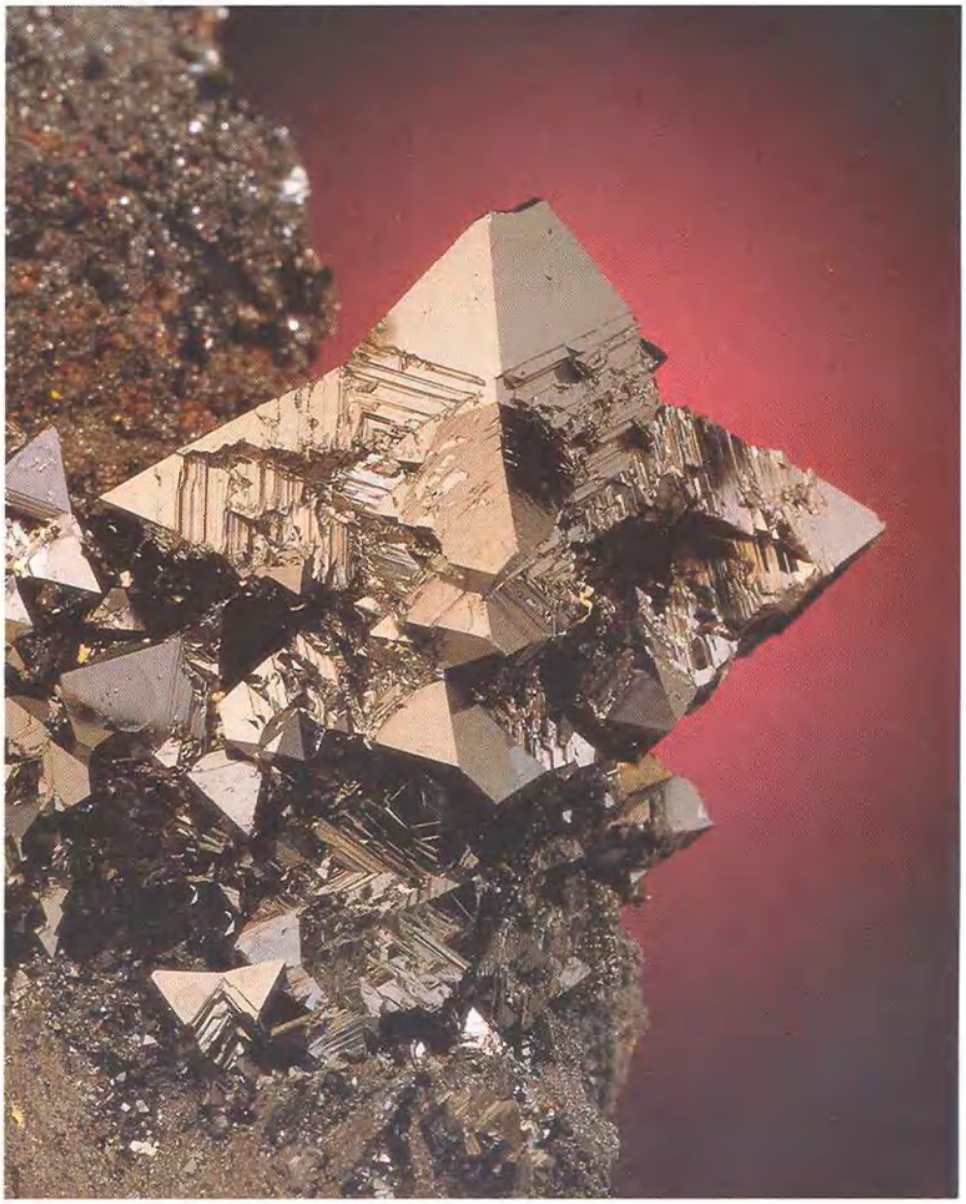

Chrysoberyl, 35 mm, Espírito Santo, Brazil

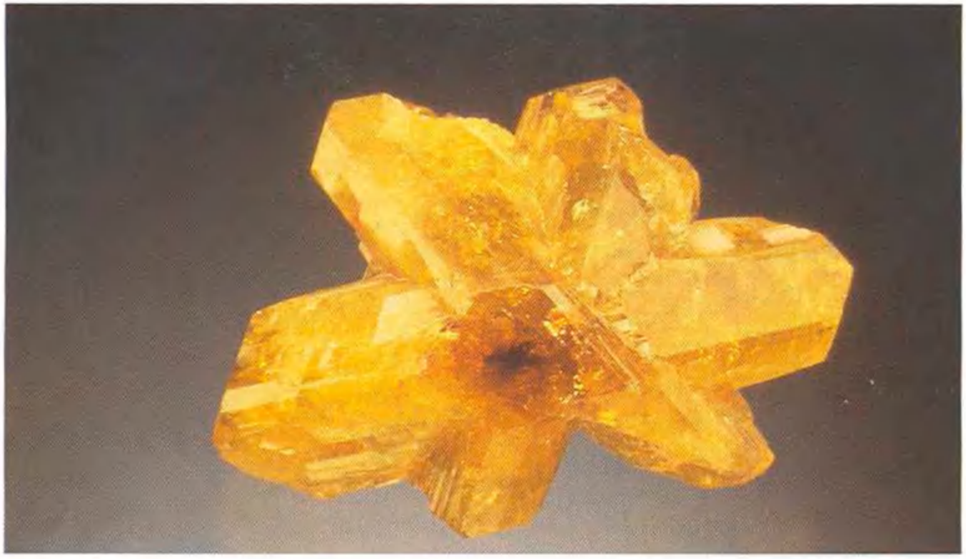

Chryzoberyl

$BeAl_2O_4$

ORTHORHOMBIC ● ● ●

Varieties: alexandrite

Properties: C – yellow-green, yellow, blue-green, alexandrite is green in daylight, purple in artificial light; S – white; L – vitreous; D – transparent to translucent; DE – 3.8; H – 8.5; CL – good; F – conchoidal to uneven; M – thin to thick tabular crystals common cyclic twins.

Origin and occurrence: Magmatic in pegmatites, prevailing as metamorphic, in association with schorl, phenakite and other minerals. Twins up to 22 cm ($8^{11}/_{16}$ in) occurred near Pancas, Espirito Santo, Brazil. Complicated twins up to 100 mm (4 in) in

Alexandrite, 75 mm, Malyshevo, Russia

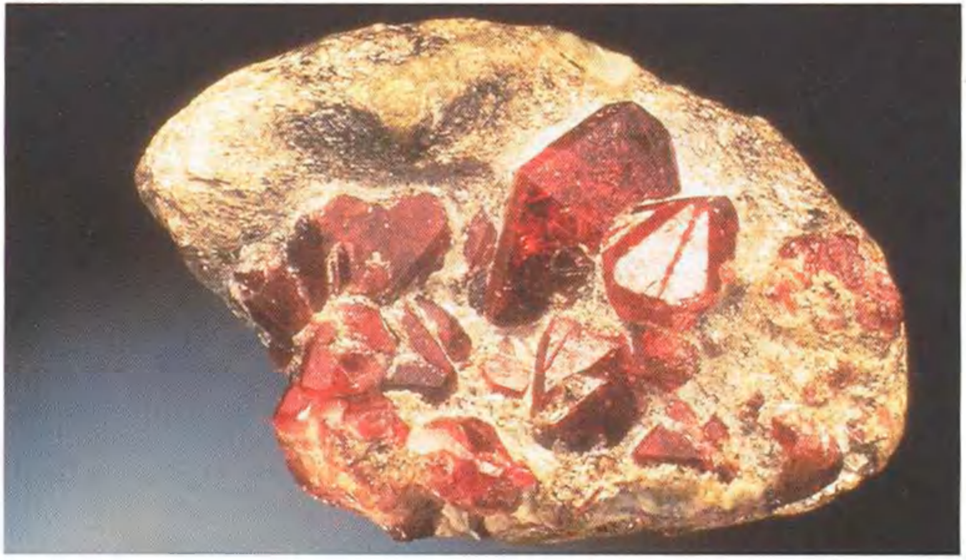

size come from Ambatondrazaka and other localities in Madagascar. Tabular crystals several cm in size, embedded in the sillimanitic rock, were found in Mar_íkov, Czech Republic. Fine alexandrite crystals up to 80 mm ($3^{1}/_{8}$ in) are known from Malyshevo, Ural mountains, Russia, together with emerald and phenakite. Alexandrite crystals reach up to 30 mm ($1^{3}/_{16}$ in) in Nyanda, Zimbabwe. Gemmy chrysoberyls, commonly with a cat's eye effect, come from the vicinity of Ratnapura, Sri Lanka.
Application: gemstone.

Valentinite

Sb_2O_3

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, brownish; S – white; L – adamantine; D – transparent to translucent; DE – 5.7-5.8; H – 2.5-3; CL – perfect; M – prismatic to tabular crystals, radial aggregates, massive.
Origin and occurrence: Secondary mineral, originated in the oxidation of stibnite. The best specimens with crystals up to 30 mm ($1^{3}/_{16}$ in) were found in Pribram, Czech Republic. Fine crystals come also from Braunsdorf, Germany. Crystals up to 20 mm ($^{25}/_{32}$ in) long occur in Oruro, Bolivia. Beautiful radial aggregates up to 40 mm ($1^{9}/_{16}$ in) in diameter, associated with kermesite, are known from Pezinok and Pernek, Slovakia. Pseudo-morphs after stibnite crystals up to 35 cm ($13^{6}/_{8}$ in)long, are reported from the Xikuangshan Mine, Lengshuijiang, China.

Valentinite, 40 mm, Nicolet, Canada

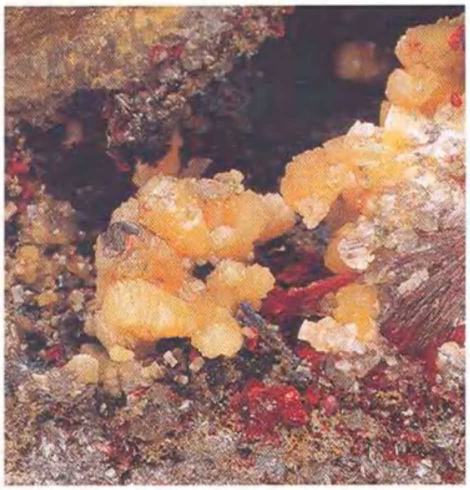

Arsenolite

As_2O_3

CUBIC ● ●

Properties: : C – white; S – white; L – vitreous; D – transparent to translucent; DE – 3.9; H – 1.5; CL – good; F – conchoidal; M – octahedral crystals, crusts, coatings; R – soluble in water.

Origin and occurrence: Secondary mineral, resulting from the oxidation of As ores. Poorly developed crystals several mm long, occur in Jachymov, Czech Republic; in Johanngeorgenstadt and St. Andreasberg, Germany. Crystals up to 20 mm ($^{25}/_{32}$ in) long originated during a mine fire in the White Caps mine, Nevada, USA.

Arsenolite, 3 mm xx, Recsk, Hungary

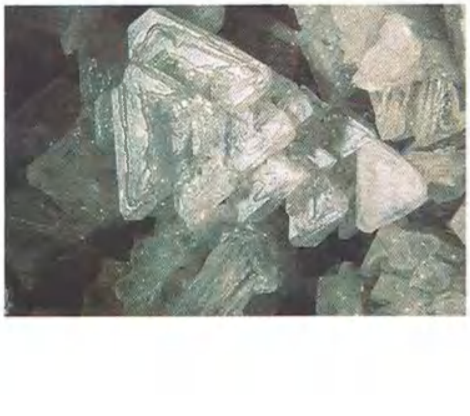

Senarmontite

Sb_2O_3

CUBIC ● ● ●

Properties: C – white, light gray; S – white; L – greasy, vitreous to adamantine; D – transparent to opaque; DE – 5.2-5.8; H – 2-2.5; CL – imperfect; F – uneven; M – octahedral crystals, granular, massive.
Origin and occurrence: Secondary, produced by stibnite oxidation, with valentinite and cerussite. The finest crystals up to 30 mm ($1^{3}/_{16}$ in) in size come from Djebel Hammimate, Algeria. Also occurred in Cetine, Italy and in Dubrava, Slovakia.

Senarmontite, 3 mm xx, Pernek, Slovakia

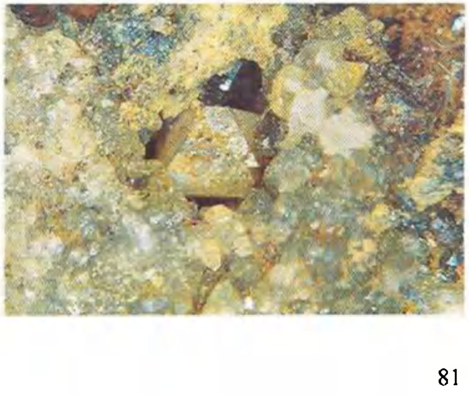

Bixbyite, 11 mm x, Thomas Range, U.S.A.

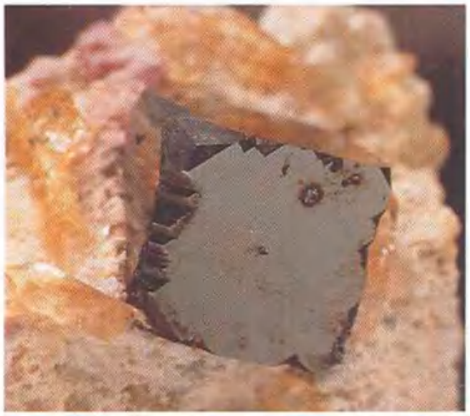

Leucosapphire, 30 mm, Sri Lanka

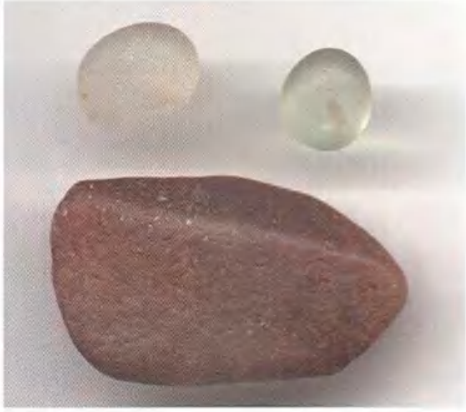

Bixbyite
$(Mn^{3+},Fe^{3+})_2O_3$

CUBIC ● ●

Properties: C – black; S – black; L – metallic to submetallic; D – opaque; DE – 5; H – 6-6.5; CLB imperfect; F – conchoidal to uneven; M – cubic crystals, also twins.

Origin and occurrence: Hydrothermal in rhyolite cavities and metamorphic. Cubes up to 12 mm ($^{15}/_{32}$ in) are found together with topaz in Thomas Range, Utah, USA. Crystals up to 25 mm (1 in) occurred in the Postmasburg mine, South Africa. Crystals up to 80 mm (3$^{1}/_{8}$ in) come from Ultevis, Sweden.

Corundum
Al_2O_3

TRIGONAL ● ● ● ●

Varieties: ruby, sapphire, leucosapphire, emery

Properties: C – colorless (leucosapphire), yellow, pink, red (ruby), blue (sapphire), purple, green,

Corundum, 10 mm xx, Miass, Russia

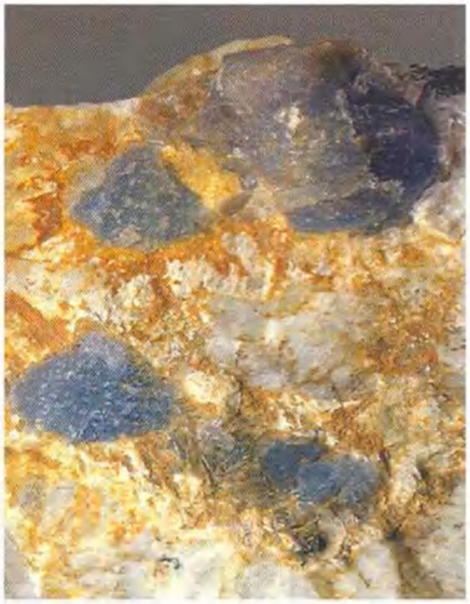

Ruby, 15 mm xx, Kola, Russia

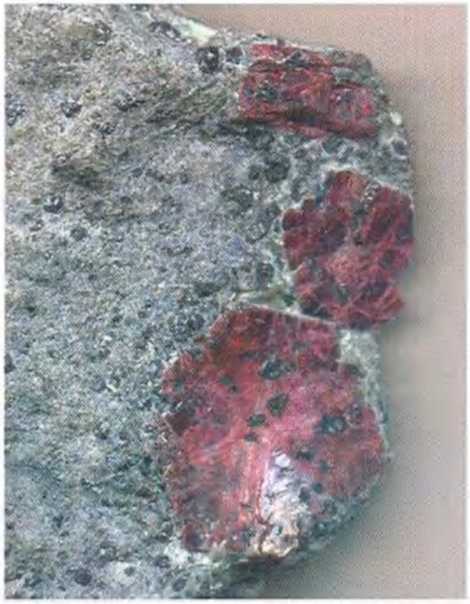

Corundum, 4 mm, Montana, U.S.A.

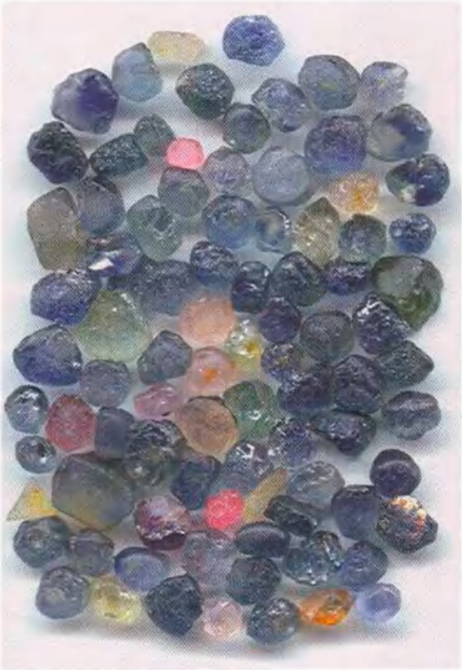

gray; S – white; L – vitreous to adamantine; D – transparent to opaque; DE – 4.0-4.1; H – 9; CL – none; F – conchoidal to uneven; M – long prismatic to barrel-like crystals, pebbles; LU – rare dark red.

Origin and occurrence: Magmatic in andesites, pegmatites and basalts, metamorphic and in placers, in association with andalusite, topaz, spinel and other minerals. Crystals of common corundum weighed up to 30 kg (66 lb) near Bancroft, Ontario, Canada. A crystal, weighing 151 kg (333 lb 3 oz) was also found in the Letaba district, South Africa. Sapphire crystals weighing up to 20 kg (44 lb), come from the vicinity of Ratnapura and Rakwana, Sri Lanka. Fine sapphire crystals are also known from Kashmir, India. Rough gem sapphire is mined from the Yogo Gulch sediments in Montana, USA and from Anikia, Queensland, Australia. Fine blue crystals up to 50 mm (2 in) long, occur near Miass, Ural mountains, Russia. Ruby is even much rarer variety of corundum. Its beautiful crystals up to 50 mm (2 in) long come from Jegdalek, Afghanistan; Mogok, Burma and from Luc Yen, Vietnam. Prismatic crystals of opaque ruby, up to 40 mm (1$^{9}/_{16}$ in) in size were found in the Khit Island near Kola Peninsula, Russia. Ruby crystals up to 30 cm (12 in) in size embedded in green zoisite from the vicinity of Arusha, Tanzania are very decorative.

Application: emery as abrasive material, sapphire and ruby as gemstones.

Ruby, 16 mm x, Jegdalek, Afghanistan

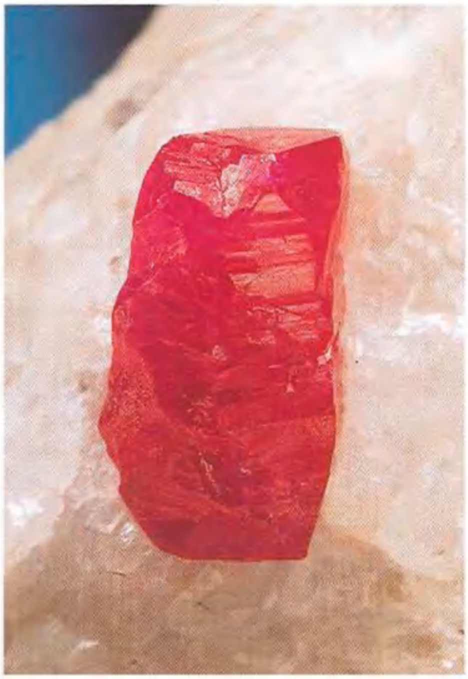

Sapphire, 41 mm, Kashmir, India

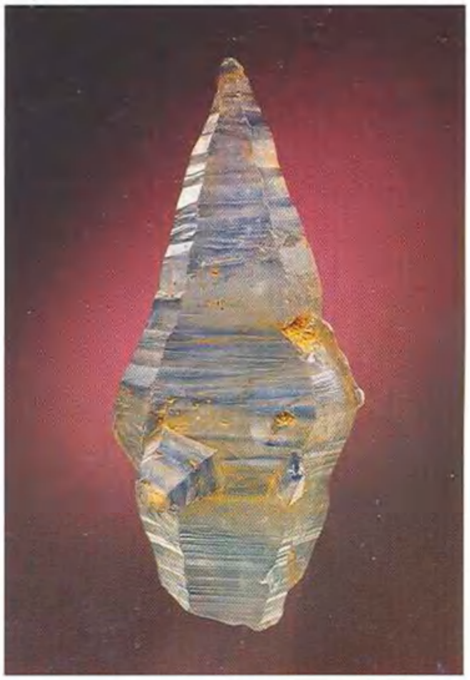

Hematite, 50 mm, La Fibbia, Switzerland

Hematite

Fe_2O_3

TRIGONAL ● ● ● ● ●

Properties: C – red, gray, black; S – red; L – metallic, dull; D – opaque; DE – 5.3; H – 6-6.5, earthy to 1; CL – none; F – uneven to conchoidal; M – thick to thin tabular crystals, massive, earthy.
Origin and occurrence: Magmatic, hydrothermal, sedimentary, also metamorphic, parageneses vary according to the origin. Beautiful crystals up to 100 mm (4 in) in size come from Brumado, BahRa, Brazil. Crystals up to 30 cm (12 in), were found in the Wessels mine, Kuruman, South Africa. So called iron roses reached up to 100 mm (4 in) near St. Gotthard, Switzerland. Fine crystals several cm in size, occurred in Rio Marina, Elba, Italy. Very fine tabular crystals reached up to 70 mm (2¾ in) in Nador, Morocco. New finds of fine crystals up to 40 mm (1 9/16 in) in size were made in the Korshunovskoye deposit, Russia. Fine botryoidal aggregates come from Hradiste and Horni Blatna, Czech Republic and from Botallack, Cornwall, UK. Sedimentary banded iron ores form huge deposits near Krivoy Rog, Ukraine or in the vicinity of Lake Superior (Mesabi Range, Minnesota; Marquette, Michigan, USA).
Application: important Fe ore.

Loparite-(Ce), 10 mm xx, Khibiny Massif, Kola, Russia

Ilmenite

$FeTiO_3$

TRIGONAL ● ● ● ●

Properties: C – black; S – black; L – metallic to dull; D – opaque; DE – 4.5-5; H – 5-6; CL – none; F – conchoidal to uneven; M – thick tabular crystals, granular, massive.
Origin and occurrence: Magmatic, metamorphic and in placers, associated with pyrrhotite, rutile, magnetite and other minerals. Crystals weighing up

Ilmenite, 11 mm x, Binntal, Switzerland

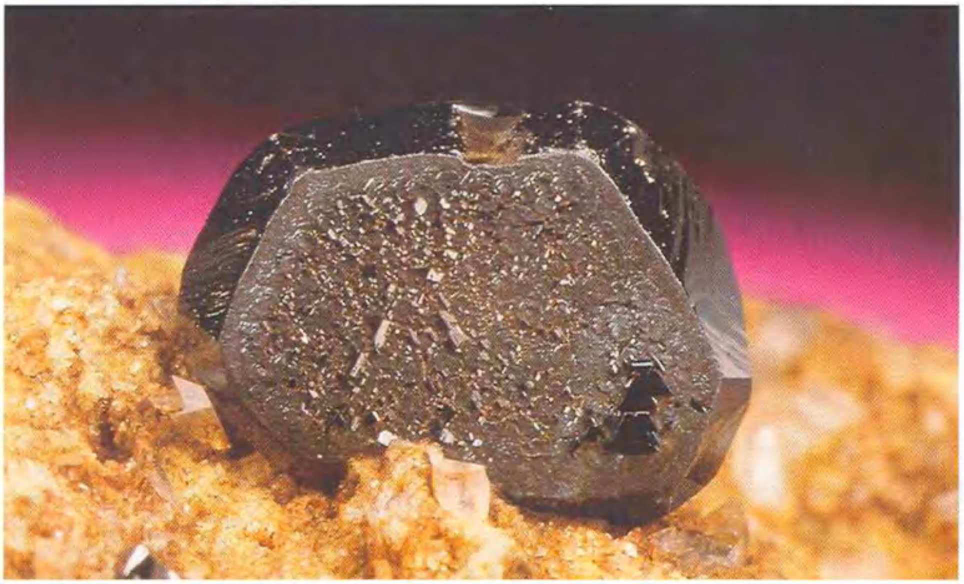

to 30 kg (66 lb) were described from the Faraday mine near Bancroft, Ontario, Canada. Crystals up to 150 mm (6 in) in size occurred near Girardville, Quebec, Canada. Crystals also reached up to 100 mm (4 in) near Miass, Ural mountains, Russia. Crystals up to 120 mm 120 mm (4$^{7}/_{16}$ in), were found in Arendal and Kragero, Norway. Crystal rosettes up to 10 mm ($^{3}/_{8}$ in) in size come from Maderanertal, Switzerland. It is also common in placers (Kamituga, Kivu, Zaire; Sri Lanka; Travancore, India; Madagascar etc.).

Perovskite
$CaTiO_3$

ORTHORHOMBIC ● ● ●

Properties: C – dark brown to black; S – colorless to gray; L – metallic to adamantine; D – opaque; DE – 4.0-4.3; H – 5.5-6; CL – imperfect; F – conchoidal to uneven; M – pseudo-cubic crystals, granular.
Origin and occurrence: Magmatic in basic and ultrabasic rocks, metamorphic, together with magnetite, zircon and other minerals. Fine pseudo-cubic crystals up to 40 mm (1$^{9}/_{16}$ in) in size come from Zlatoust and Akhmatovsk, Ural mountains, Russia. It occurs as crystals up to 80 mm (3$^{1}/_{8}$ in), associated with magnetite crystals in the Gardiner complex, Greenland. Crystals up to 40 mm (1$^{9}/_{16}$ in) were found in Jacupiranga, Sao Paulo, Brazil. Crystals from Val Malenco, Italy, reached up to 20 mm ($^{25}/_{32}$ in). Crystals up to 20 mm ($^{25}/_{32}$ in) were lately found in Afrikanda, Kola Peninsula, Russia.

Loparite-(Ce)
$(Ce,Na,Ca)TiO_3$

ORTHORHOMBIC ●

Properties: C – black, S – dark red-brown, L – metallic; D – opaque; DE – 4.6-4.9; H – 5.5-6; M – pseudo-cubic crystals, granular; R – metamict.
Origin and occurrence: Magmatic in alkaline rocks, with lorenzenite, eudialyte and aegirine. Fine interpenetration twins up to 20 mm ($^{25}/_{32}$ in) in size come from Mount Nyorkpakhk, Kola Peninsula, Russia.

Perovskite, 60 mm, Zlatoust, Russia

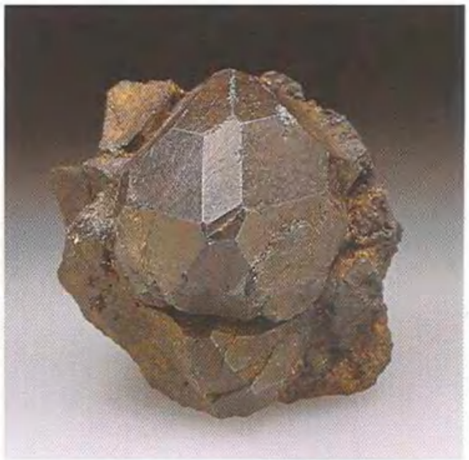

Stibiconite, 230 mm, Catorce, Mexico

Stibiconite

$Sb^{3+}Sb^{5+}{}_2O_6(OH)$

CUBIC ● ● ●

Properties: C – white, creamy, light yellow, brown; S – white; L – vitreous, greasy to dull; D – opaque; DE – 4.1-5.8; H – 3-6; M – pseudo-morphs after stibnite crystals, earthy, massive.
Origin and occurrence: Secondary, as a result of the stibnite oxidation, associated with valentinite and other minerals. Fine pseudo-morphs after stibnite crystals up to 30 cm (12 in) long come from Catorce, San Luis Potosí, Mexico. Similar pseudo-morphs were also found in Kostainik, Serbia; in the Ichinokawa mine, Japan and in Pereta, Italy. Pseudo-morphs after stibnite up to 20 cm ($7^{7}/_{8}$ in) long occur also in Çukuroren, Turkey.

Bindheimite, 4 mm, Rudabánya, Hungary

Bindheimite

$Pb_2Sb_2O_6(O,OH)$

CUBIC ● ● ●

Properties: C – yellow, brown, gray; S – yellow, L – resinous, dull to earthy; D – translucent to opaque; DE – 4.6-5.6; H – 4-4.5; F – conchoidal to earthy; M – botryoidal, nodular and earthy crusts.
Origin and occurrence: Secondary in the oxidation zone of Pb-Sb deposits. Needles up to 10 mm ($^{3}/_{8}$ in) long come from Rudník, Czech Republic. It is common in Broken Hill, New South Wales, Australia; in Bisbee, Arizona, USA and in Sidi-Amor-ben-Salem, Tunisia. Lamellar pseudo-morphs up to several cm in size are known from Tsumeb, Namibia.

Pyrochlore

$(Na,Ca)_2Nb_2O_6(OH,F)$

CUBIC ● ● ●

Properties: C – yellow-brown, brown, black; S – brown; L – vitreous to greasy; D – translucent to opaque, DE – 4.5; H – 5-5.5; CL – locally good; F – conchoidal to uneven; M – octahedral crystals, granular; R – radioactive (admixtures of U, Th).
Origin and occurrence: Magmatic in alkaline rocks, together with zircon, astrophyllite and other minerals. Fine brown shiny crystals up to 20 mm ($^{25}/_{32}$ in) in size come from the vicinity of Vishnevogorsk, Ural mountains, Russia. Crystals reaching 10 mm ($^{3}/_{8}$ in) occur in the Panda Hill deposit, Tanzania. Crystals are also known from Oka, Quebec, Canada. Single octahedra measuring 5 mm ($^{3}/_{16}$ in) were found in Luesha, Kivu, Zaire. *Application:* Nb,U and Th ore.

Microlite, 7 mm x, Gillette Quarry, U.S.A.

Pyrochlore, 58 mm, Vishnevogorsk, Ural Mts., Russia

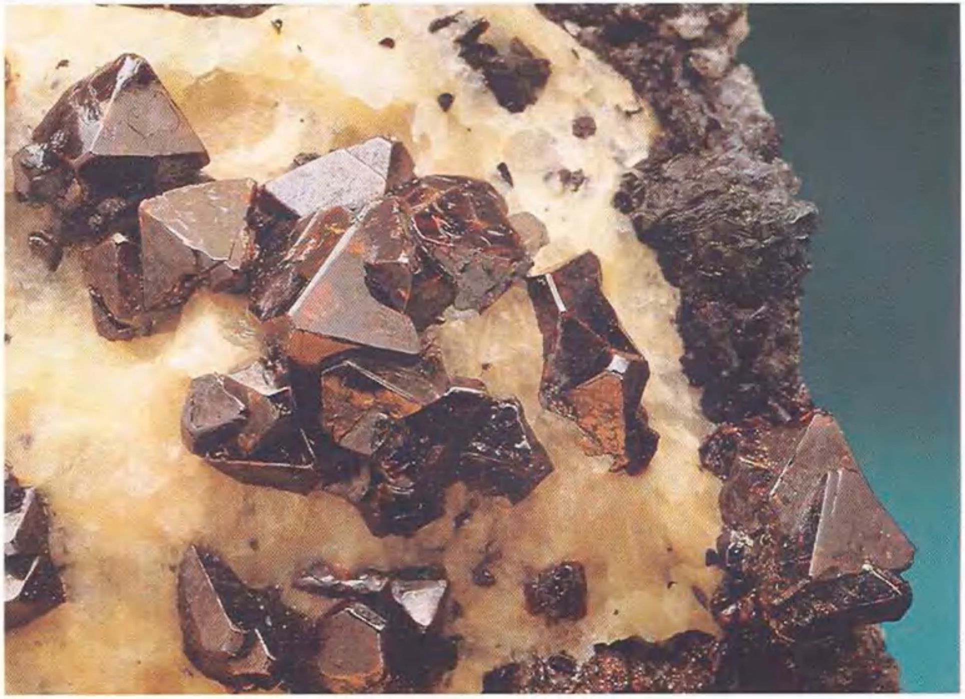

Betafite

$(Ca,Na,U)_2(Ti,Nb,Ta)_2O_6(OH)$

CUBIC ● ●

Properties: C – black, brown, yellow-brown; SB red-brown; L – resinous to greasy; D – translucent to opaque; DE – 4.2; H – 3-5.5; CL – none; F – conchoidal to uneven; M – octahedral crystals; R – radioactive, metamict.

Origin and occurrence: Magmatic in granitic pegmatites, rich in U, Th and rare earth elements, associated with beryl, euxenite-(Y) and other minerals. The world's best specimens come from many localities in Madagascar (Betafo, Ambatofotsikely etc.), where crystals up to 6 kg (13 lb 3 oz) were found. Beautiful specimens with crystals up to 100 mm (4 in) in size occur in the Silver Crater mine near Bancroft, Ontario, Canada. It is also known from Evje, Norway.

Betafite, 20 mm, Silver Crater, Canada

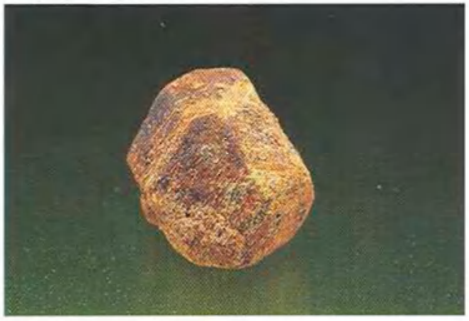

Mikrolite

$(Na,Ca)_2Ta_2O_6(O,OH,F)$

CUBIC ● ● ●

Properties: C – brown, yellow, green, reddish; S – white; L – vitreous to greasy, locally adamantine; D – translucent to opaque; DE – 5-6.4; H – 6-6.5; CL – locally good; F – conchoidal to uneven; M – octahedral crystals, granular, massive.

Origin and occurrence: Magmatic, typical for granitic pegmatites, together with manganocolumbite, manganotantalite and other minerals. Octahedra up to 65 mm (2$^{9}/_{16}$ in) in size occur in Ankola, Uganda. Crystals up to 30 mm (1$^{3}/_{16}$ in) in size come from Virgem da Lapa, Minas Gerais, Brazil. Crystals up to 75 mm (3 in) are reported from the Harding pegmatite, New Mexico, USA. It occurs in important accumulations near Wodgina, Western Australia.

Quartz, 60 mm, Herkimer, U.S.A.

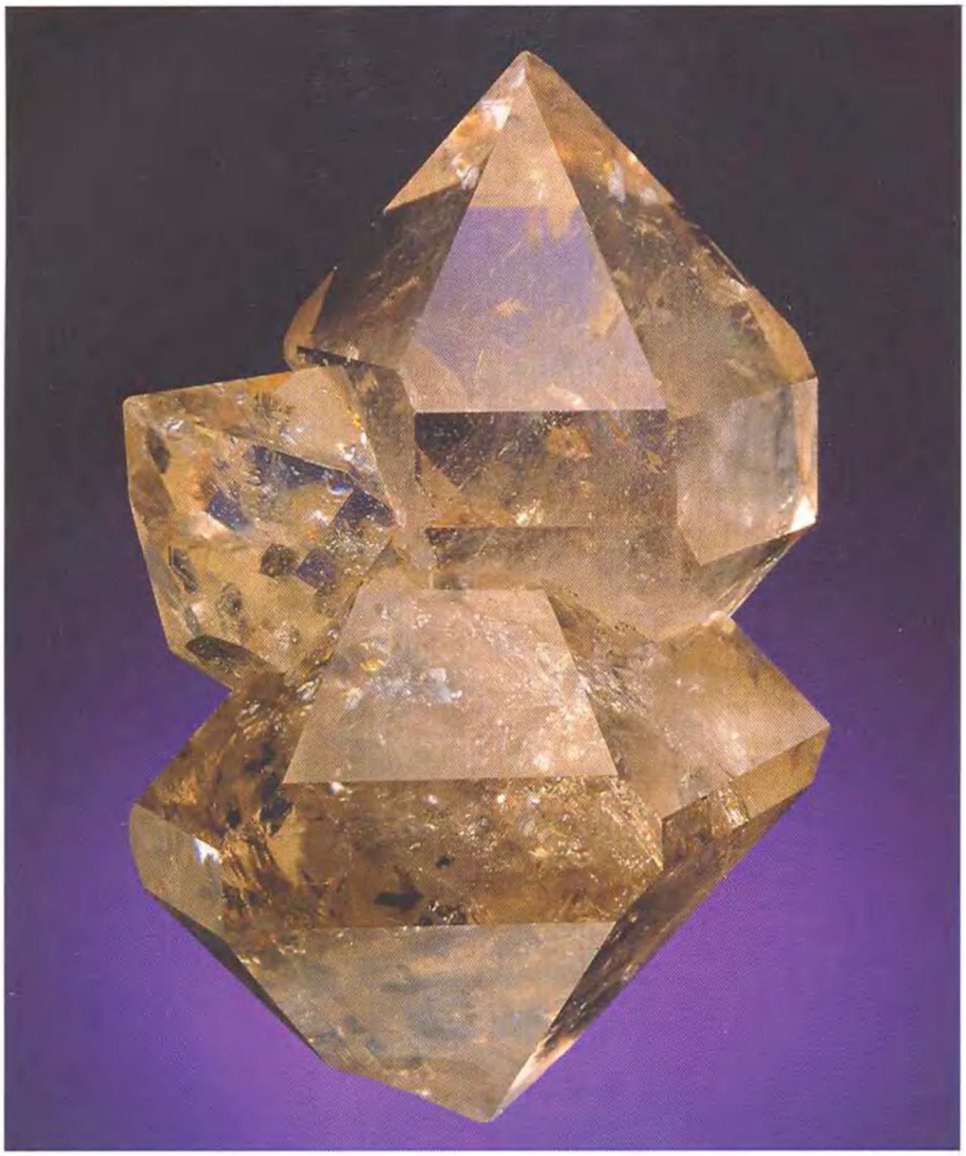

Quartz

SiO_2

TRIGONAL ● ● ● ● ●

Varieties: rock crystal, citrine, smoky citrine, morion, amethyst, rose quartz, chrysoprase, jasper, chalcedony, agate, onyx, sardonyx, aventurine, heliotrope, tiger's eye, falcon s eye.

Properties: C – colorless (rock crystal), white, yellow (citrine), brown (smoky citrine), black (morion), purple (amethyst), pink (rose quartz), green (chrysoprase); D – these varieties are mostly transparent, often translucent; C – other varieties are mainly multicolored, separate colors have different hues and the color is commonly caused by microscopic admixtures of other minerals; varieties: red, green, brown, yellow (jasper), banded with different colors (agate); white

Rock crystal, 95 mm, La Gardette, France

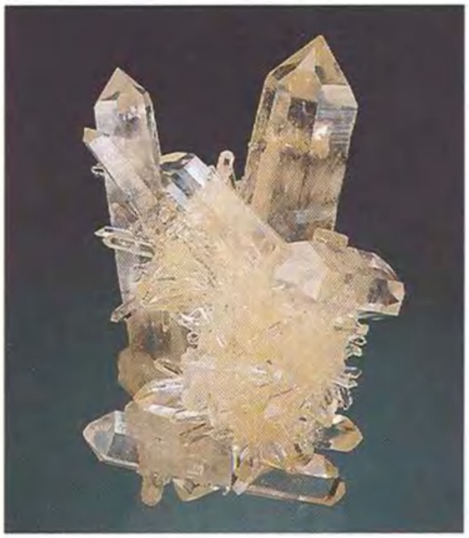

Citrine, 32 mm, Charcas, Mexico

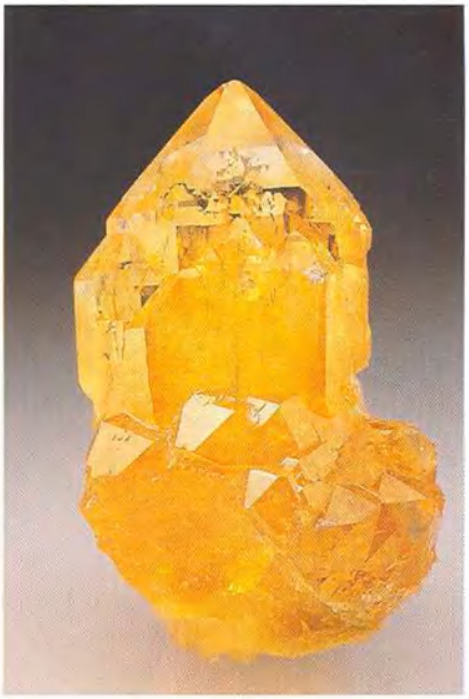

and black bands (onyx), white and red-brown bands (sardonyx), green to red-brown with mica or hematite inclusions (aventurine), dark green with red spots (heliotrope), yellow-brown to black-brown, fibrous with silky luster (tiger's eye), blue-gray to yellow-brown, fibrous with silky luster (falcon's eye); S – white; L – vitreous, silky, dull; D – transparent to translucent, opaque; DE – 2.6; H – 7; CL – none; F –

Smoky quartz, 70 mm, Middle Moat Mt., U.S.A.

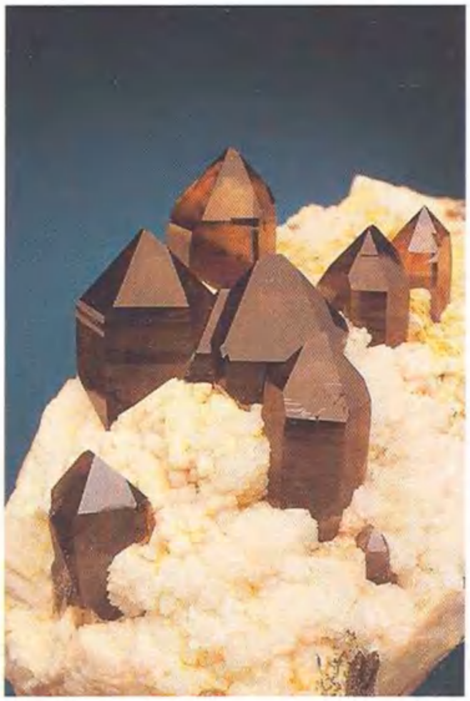

Smoky quartz, 81 mm, Switzerland

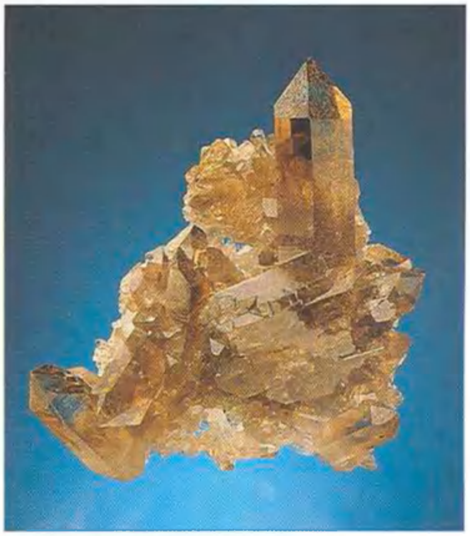

Morion, 100 mm, Agadir, Kazakhstan

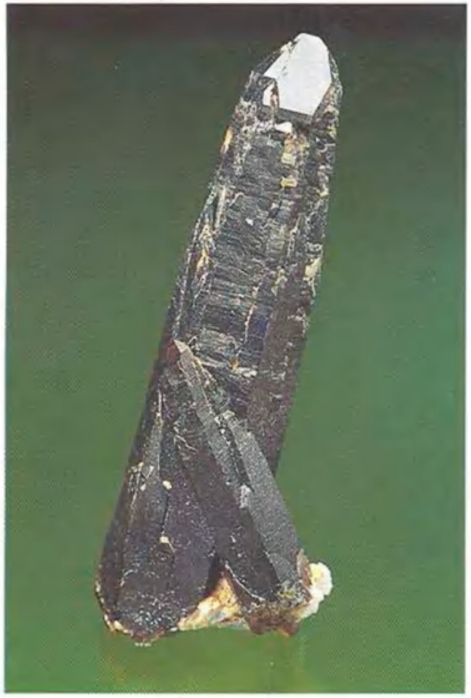

Amethyst, 100 mm, Bochovice, Czech Republic

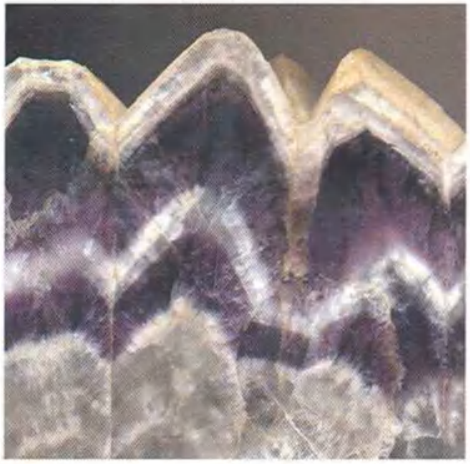

conchoidal; M – long to short prismatic, acicular, dipyramidal to tabular crystals, fibrous, botryoidal and stalactitic aggregates and coatings, concretions, geodes, granular, massive.
Origin and occurrence: Magmatic in different types of rocks, mainly in granites, granitic pegmatites and volcanic rocks; metamorphic in different types of rocks, mainly in quartzites and mica schists; hydrothermal in different types of ore and Alpine-

Rose quartz, 145 mm, Minas Gerais, Brazil

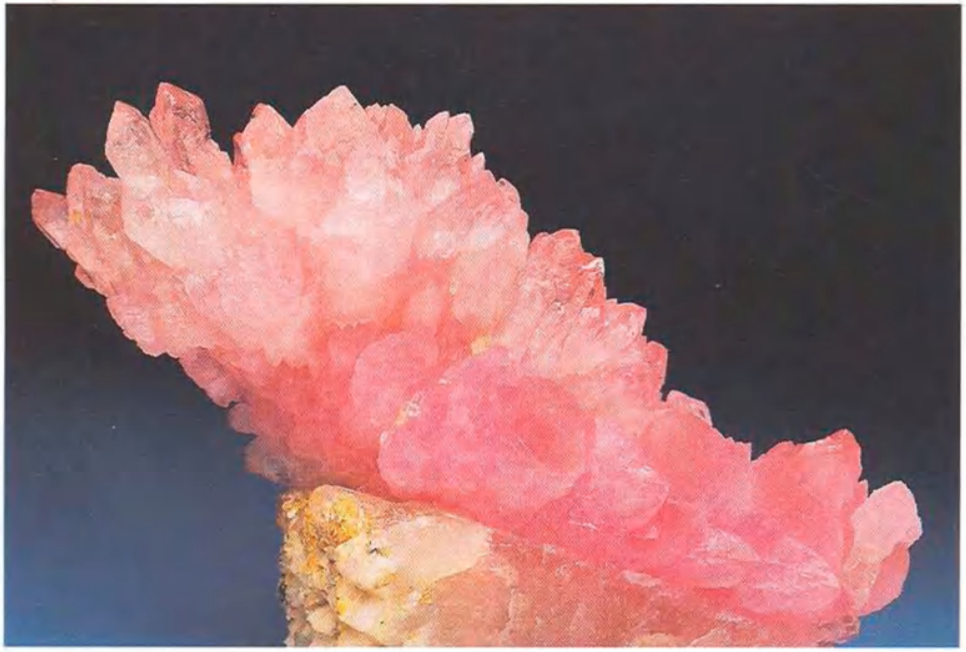

Amethyst, 200 mm, Guerrero, Mexico

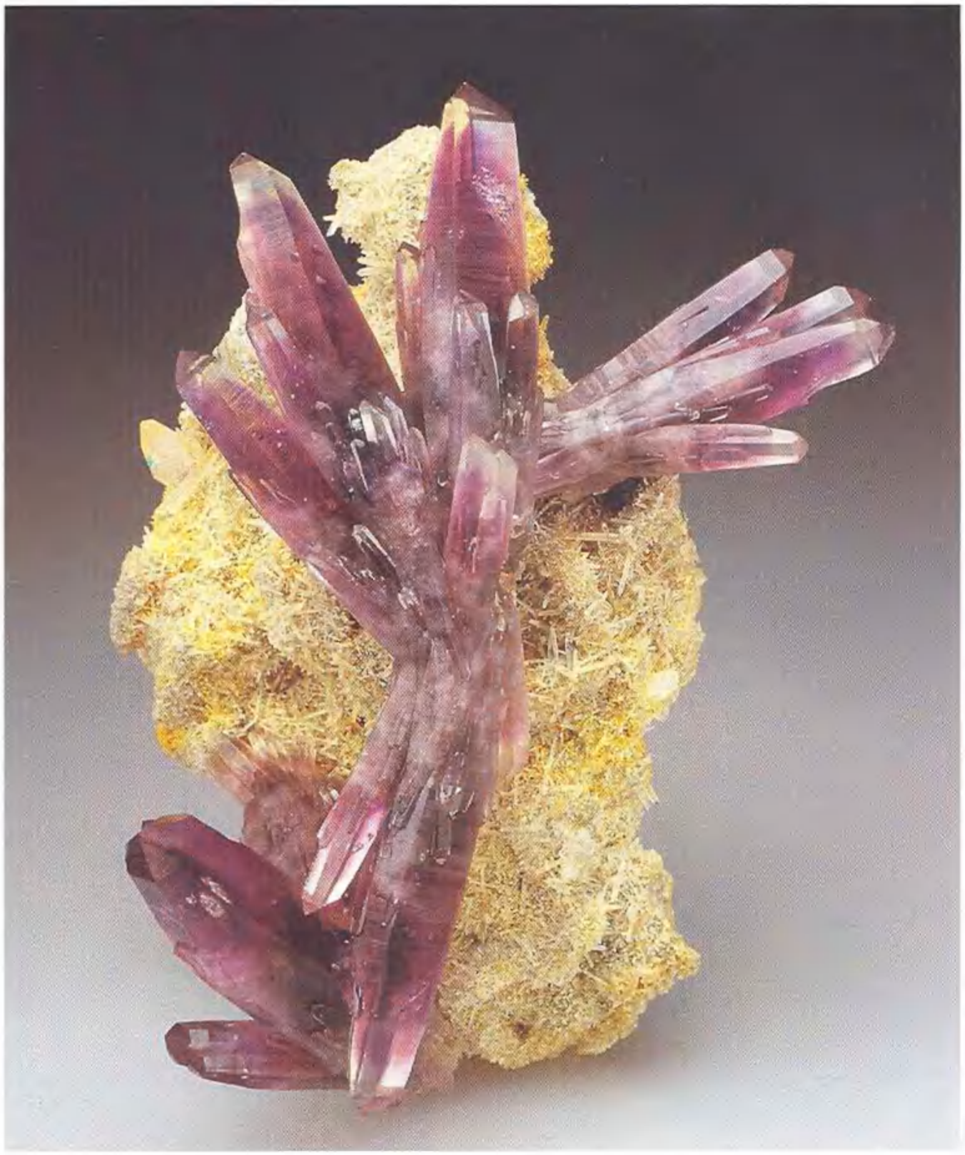

type veins; secondary in the oxidation zone of ore deposits; also in different types of sedimentary rocks and in organic remains, also in placers. Probably the most common mineral in the Earth's crust and the most important rock-forming mineral, as well.

Large crystals of rock quartz up to 7 m (23 ft) long come from pegmatites in the Betafo region in Madagascar and from the Alpine-type veins, like Uri, Grimsel and Furka, Switzerland; perfect crystals are known from the cracks in marbles near Carrara, Italy; it also occurs in the quartz veins in Herkimer, New York and Hot Springs, Arkansas, USA. Citrine occurs mainly in granitic pegmatites and large crystals come from Goias, Brazil; from Suky and Netín, Czech Republic; from Murzinka, Ural mountains, Russia. Smoky quartz originates mostly in granitic pegmatites, it also occurs in the

Chalcedony, 68 mm, High Atlas, Morocco

Carnelian, 40 mm, Nova Paka, Czech Republic

Chrysoprase, 60 mm, Szklary, Poland

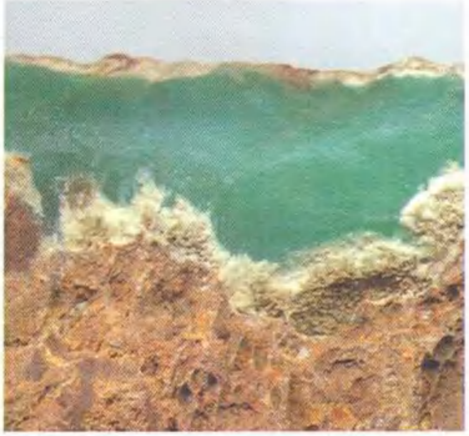

Alpine-type quartz veins and in cavities of volcanic rocks. Perfect crystals up to several meters long, come from many places, the largest crystal, weighing 77 tons , was found in Kazakhstan; perfect crystals occur in pegmatites in many places in Brazil; also in Korostenskiy massif, Ukraine; in the Pikes Peak batholith, Colorado, USA; in the Alpine-type veins in Maderanertal and in Grimsel, Switzerland. Morion crystals, commonly associated with smoky citrine, were found in quartz veins and in pegmatites.
Its crystals are known from St. Gotthard, Switzerland. Amethyst comes from quartz and ore veins, cavities in volcanic rocks, rare in the Alpine-type veins. Famous localities in volcanic rocks are in the states of Rio Grande do Sul and Minas Gerais, Brazil, doubly terminated crystal, weighing 5.5 tons, come from Diamantina. In Serra do Mar, Rio Grande do Sul. a cavity covered with amethyst

Jasper, 50 mm, Ural Mts., Russia

Jasper, 60 mm, Oregon, U.S.A.

Petrified wood, 70 mm, Podkrkonosi, Czech Republic

crystals measuring 10 x 2 x 1 m (33 x 6 x 3 ft 3in) was found; rich druses occur also in the ore veins in Porcura, Romania and Julimes, Mexico. Rose quartz, forming masses up to several meters in granitic pegmatites in the Rose Quartz pit, Quadeville, Ontario, Canada; in Ambositra, Madagascar; crystals up to 10 mm (³/₈ in) long growing on quartz crystals, come from Sapucaia, Minas Gerais, Brazil. Dark green chrysoprase veins up to 50 mm (2 in) thick are known from serpentinites in Szklary,

Iron stained quartz, 95 mm, Horovice, Czech Republic

Agate, 50 mm, Zeleznice, Czech Republic

Onyx, 65 mm, Brazil

Poland. Jasper is known from volcanic rocks and their contacts with sediments, locally as a result of petrification of organic matter, mainly plants, it is also known from quartz veins. Rich aggregates occur in Idar-Oberstein, Germany; in Podkrkonosi region and in Krusne hory mountains, Czech Republic; in the Petrified Forest National Park, Holbrook, Arizona, USA; in Ural mountains, Russia; in Kabamby, Madagascar. Chalcedony is mostly found in quartz veins and geodes in volcanic rocks, also in sediments.

Rich aggregates come from Idar-Oberstein, Germany; Julimes, Mexico; many localities in Uruguay and Brazil; in Huttenberg, Austria. Agates are known from cavities in volcanic rocks, rare in hydrothermal veins and in sediments. The most important localities are located in the southern part of Brazil in the state of Rio Grande do Sul; in Uruguay; also in Yemen; India; Mongolia; in several

Moss agate, 40 mm, Krdjali, Bulgaria

Agate, 140 mm, Horni Halze, Czech Republic

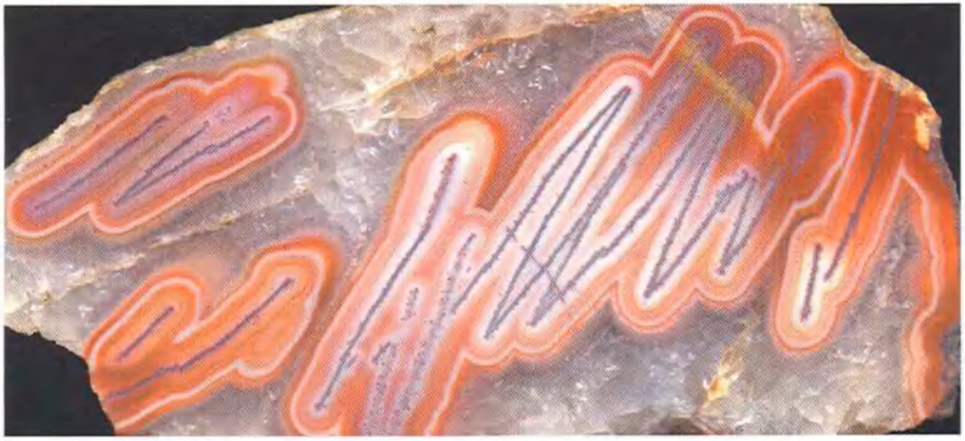

Agate, 60 mm, Brazil

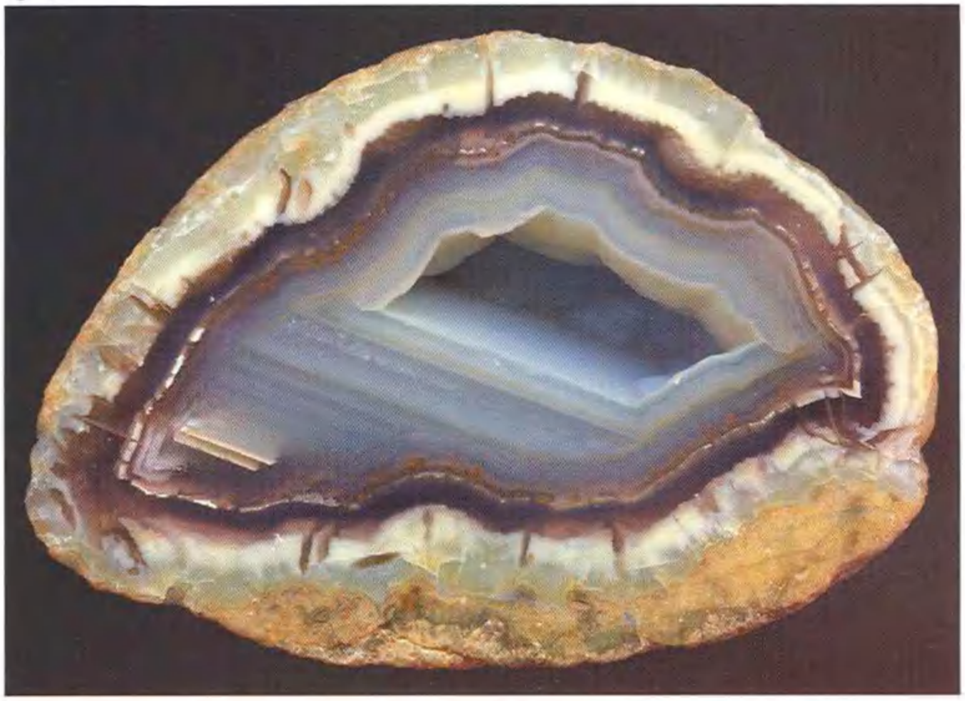

localities in the USA; in Idar-Oberstein, Germany; in Podkrkonosi, Czech Republic. The most famous localities of onyx and sardonyx are in Brazil and Uruguay.

Rich aggregates of aventurine come from Miass, Ural mountains, Russia; Mariazell, Austria; Belany, India. Heliotrope occurs in Idar-Oberstein, Germany; Kozakov, Czech Republic and in Brazil.

Application: important raw material in glass industry, many colored varieties, like amethyst, smoky citrine, citrine, onyx, sardonyx, and heliotrope are cut as gemstones.

Aventurine, 50 mm, India

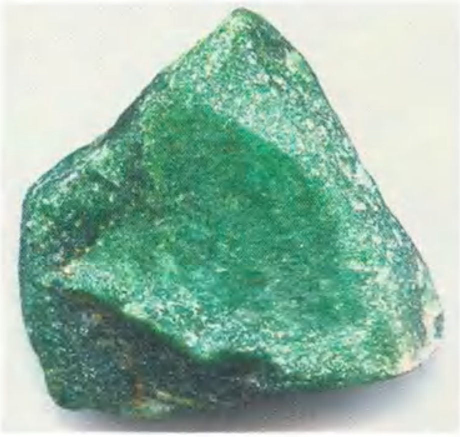

Tiger's eye, 50 mm, Griqualand, South Africa

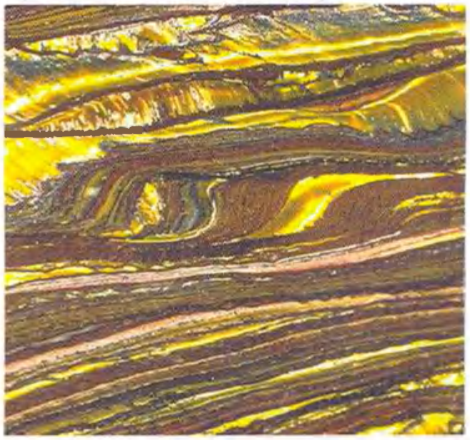

Tridymite, 9 mm, Big Luc Mt., U.S.A.

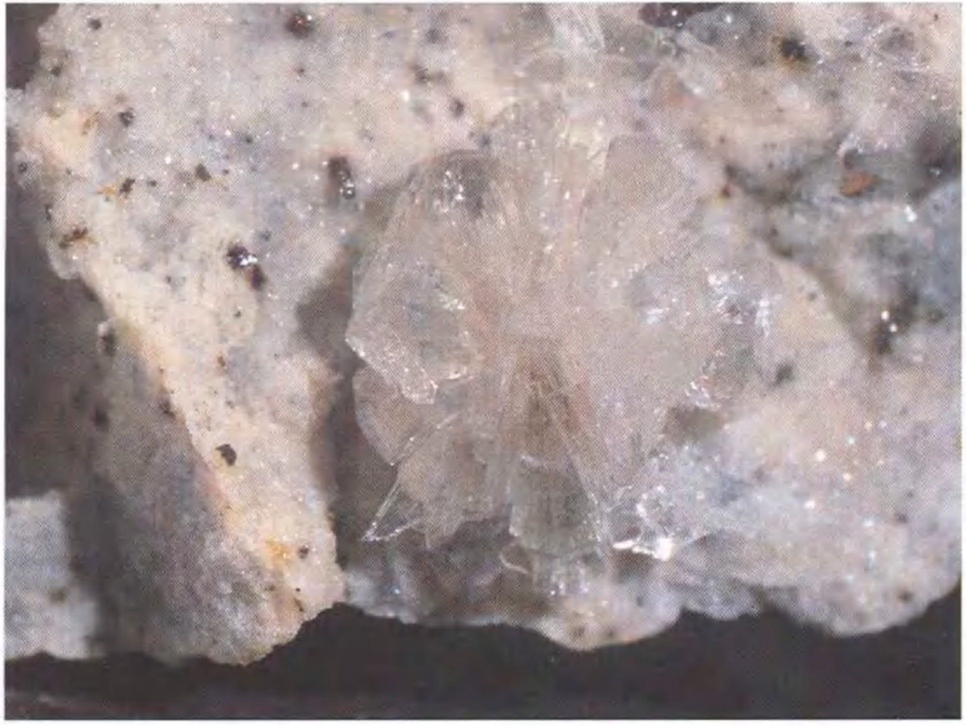

Tridymite

SiO_2

ORTHORHOMBIC ● ● ●

Properties: C – colorless to white; S – white; L – vitreous; D – transparent; DE – 2.3; H – 6.5-7; CL – none; F – conchoidal; M – pseudo-hexagonal tabular crystals.

Cristobalite, 110 mm, Vechec, Slovakia

Origin and occurrence: Magmatic in cavities of young felsic volcanic rocks in association with cristobalite, chalcedony and other minerals. Pseudo-hexagonal tabular crystals up to 10 mm (3/8 in) in size, come from Vechec, Slovakia. Similar crystals found in Ichigayama, Japan. Crystals up to 10 mm (3/8 in) occur with topaz and other minerals in the Thomas Range, Utah, USA.

Cristobalite

SiO_2

TETRAGONAL ● ● ●

Varieties: lussatite (fibrous)

Properties: C – colorless to white; S – white; L – vitreous; D – translucent; DE – 2.3; H – 6.5; CL – none; M – pseudo-octahedral crystals, spherical and botryoidal aggregates.
Origin and occurrence: Magmatic in cavities of young felsic volcanic rocks, associated with tridymite. Crystals up to 4 mm (5/32 in) are known from Cerro San Cristobal, Hidalgo, Mexico. Crystals up to 2 mm (1/16 in) long occur in Vechec, Slovakia. Gray spherical aggregates come from Coso Hot Springs, California, USA.

Common opal, 60 mm, Křemže, Czech Republic

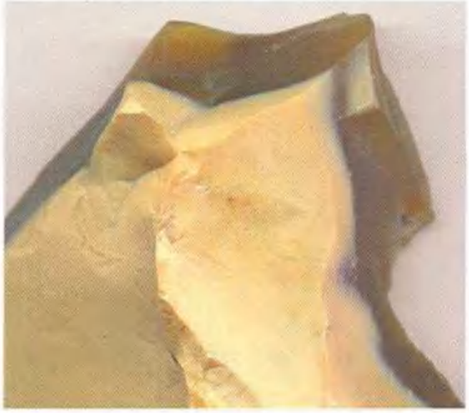

Opal, 50 mm, Herlany, Slovakia

Opal

$SiO_2 \cdot n\ H_2O$

AMORPHOUS ● ● ● ● ●

Varieties: hyalite, milky opal, fire opal, precious opal, wooden opal, geysirite.

Properties: C – colorless (hyalite), white (milky opal), red (fire opal), iridescence (precious opal), brown, red-brown, yellow, green, gray, blue; S – white; L – vitreous, dull, earthy, waxy; D – transparent to translucent, opaque; DE – 2.1; H – 5.5-6.5; CL – none; F – conchoidal; M – botryoidal and stalactitic aggregates, coatings, concretions, geodes, massive; LU – white, yellow-green, green.

Origin and occurrence: Hydrothermal in volcanic rocks and tuffs, also in various types of volcanic rocks and tuffites, in different types of sedimentary rocks, in organic remnants and hot springs, rare in

Hyalite, 45 mm, Valec, Czech Republic

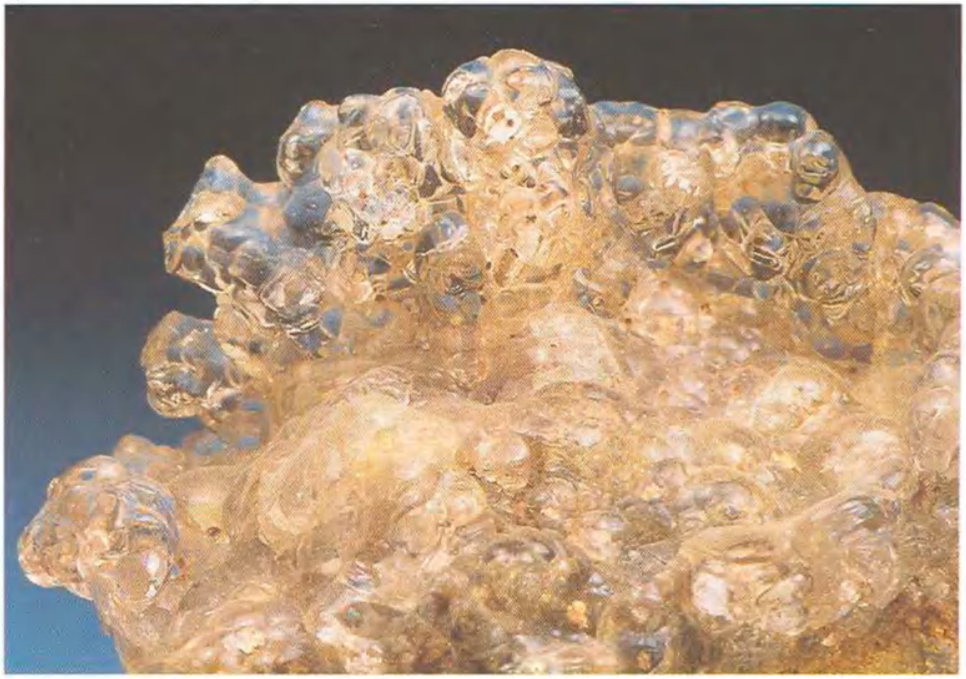

Opal, 47 mm, Opal Butte, U.S.A.

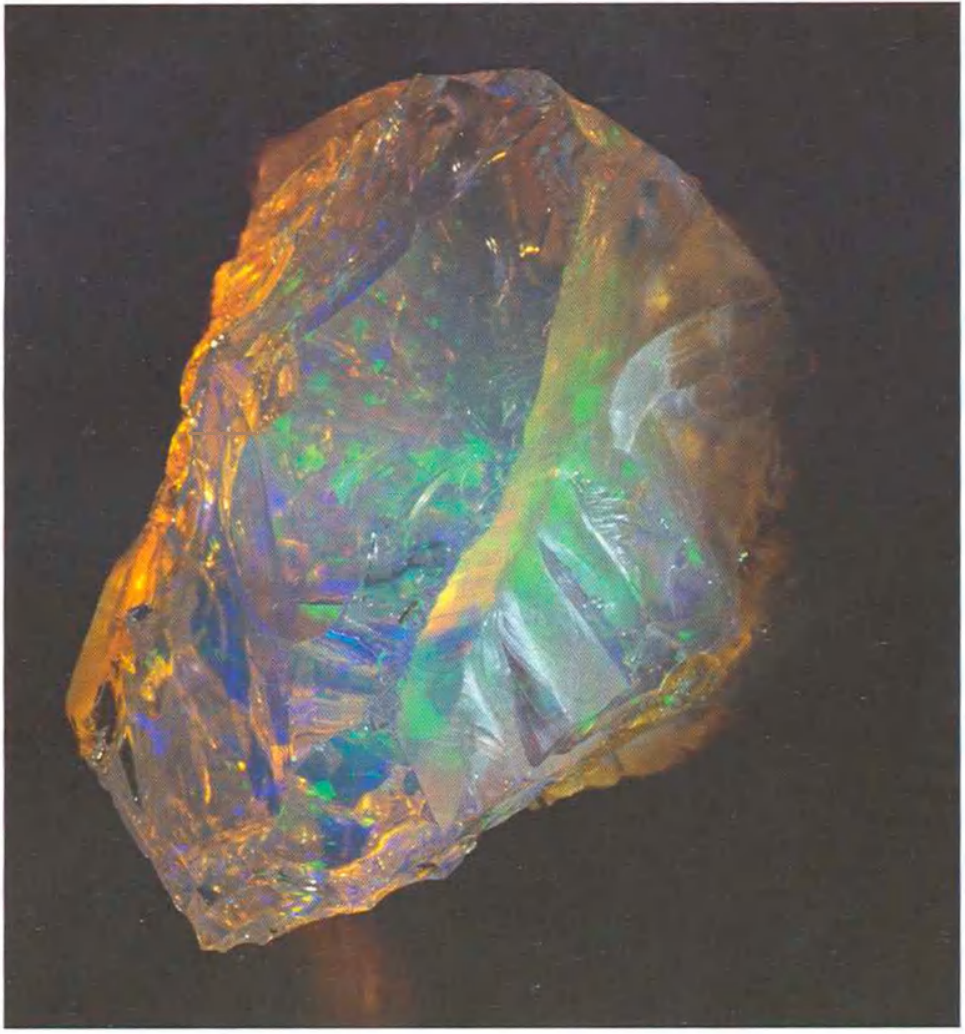

Wooden opal, 40 mm, Lubietová, Slovakia

hydrothermal veins; secondary in the weathering zone of different types of rocks. It is often associated with chalcedony. Coatings and stalactitic aggregates of hyalite up to 50 mm (2 in) thick known from Cerritos, Mexico; Valec, Czech Republic; Klamath Falls, Oregon, USA. Milky opal occurs in Dubník, Slovakia; Smrcek, Czech Republic and many other localities. The most famous locality of fire opal is Zimapan, Hidalgo, Mexico. Precious opal comes from many localities in Australia, e.g. Baracoo River, Queensland; Coober Pedy, Southern Australia and White Cliffs, New South Wales, where it forms rich aggregates and veinlets in sandstones; classic locality is Dubník, Slovakia, where it occurs in

Dendritic opal, 56 mm, Křemže, Czech Republic

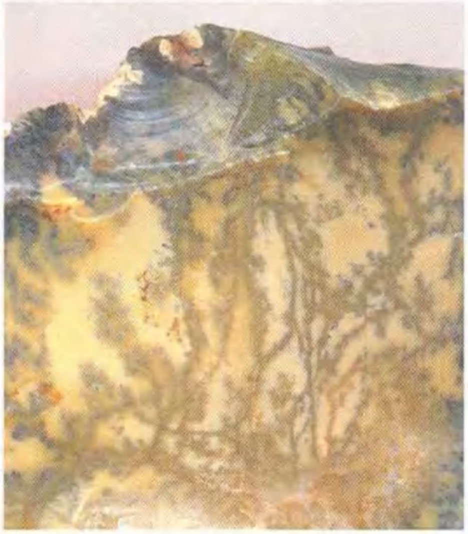

Fire opal, 20 mm, Mexico

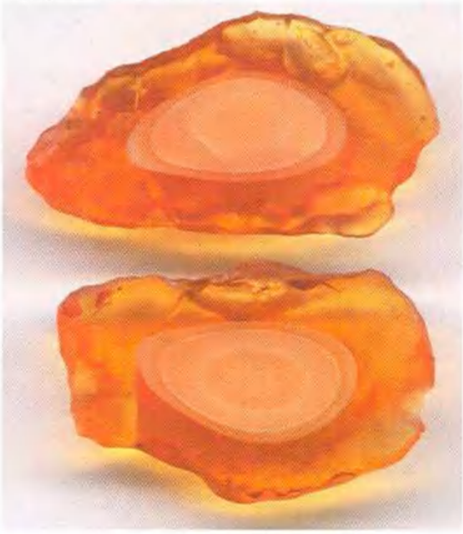

andesites and was probably mined already by ancient Romans. Beautiful precious opals come also from the Virgin Valley, Nevada, USA. Petrified trees, known as wooden opal, reach lengths of several meters in the Petrified Forest National Park, Holbrook, Arizona, USA; in L'ubietova and Povraznik, Slovakia. White geysirite is known mainly from the hot springs in Iceland; Yellowstone National Park, Wyoming, USA; New Zealand.

Application: colored opal varieties, primarily precious opal and fire opal are cut as gemstones, diatomite in chemical industry.

Precious opal, 55 mm, Opal Butte, U.S.A.

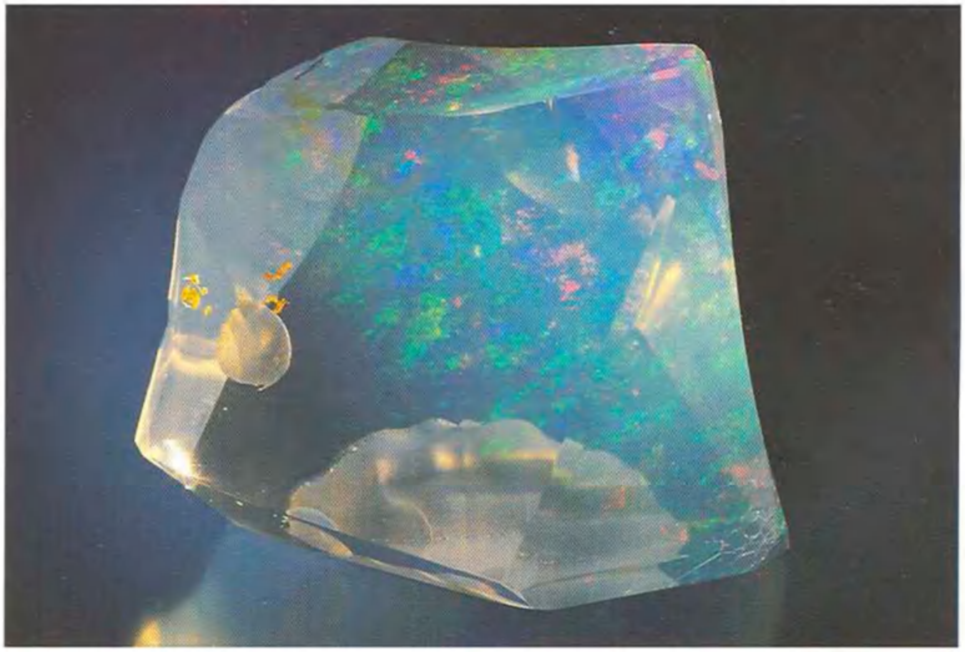

Melanophlogite, 80 mm, Fortullino, Italy

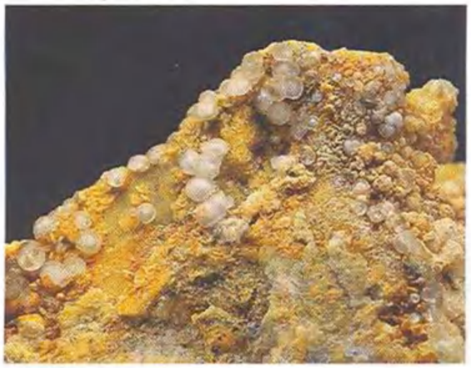

Melanophlogite

SiO_2

CUBIC ●

Properties: C – colorless, white; S – colorless; L – vitreous; D – transparent; DE – 2; H – 6.5; CL – none; M – pseudo-cubic crystals, spherical aggregates.
Origin and occurrence: Hydrothermal, associated with sulfur and other minerals. It was originally described from the sulfur deposit in Racalmuto, Sicily, Italy as crystals up to 4 mm (5/32 in) in size. It is also known from Chvaletice, Czech Republic, where it forms crystals up to 2 mm (1/16 in) in size, in Alpine-type veins.

Rutile

TiO_2

TETRAGONAL ● ● ● ●

Properties: C – red-brown, red, brown, yellowish, black; S – light brown; L – metallic to adamantine; D – transparent to translucent; DE – 4.2; H – 6-6.5; CL – good; F – conchoidal to uneven; M – short prismatic, striated crystals, common twins, acicular crystals, granular, massive.
Origin and occurrence: Magmatic and metamorphic, also in placers, together with monazite-(Ce), topaz, beryl, quartz and other minerals. The largest crystals up to 150 mm (6 in) in size come from the Mount Graves, Georgia, USA. Beautiful epitaxial intergrowths with hematite occur in Cavradischlucht, St. Gotthard, Switzerland and in Ibitiara, Bahía, Brazil. It is common as inclusions in smoky citrine (quartz) crystals from Ibitiara, Bahia and Itabira, Minas Gerais, Brazil. Knee-like crystal twins up to 70 (2¾ in) cm in size were found in the vicinity of Golcuv Jenikov and Sobeslav, Czech Republic.
Application: Ti ore.

Rutile, 21 mm, Ibitiara, Brazil

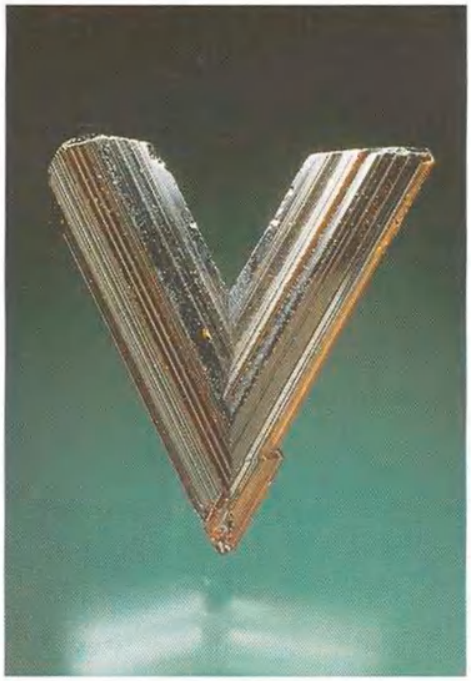

Rutile, 48 mm, Bahía, Brazil

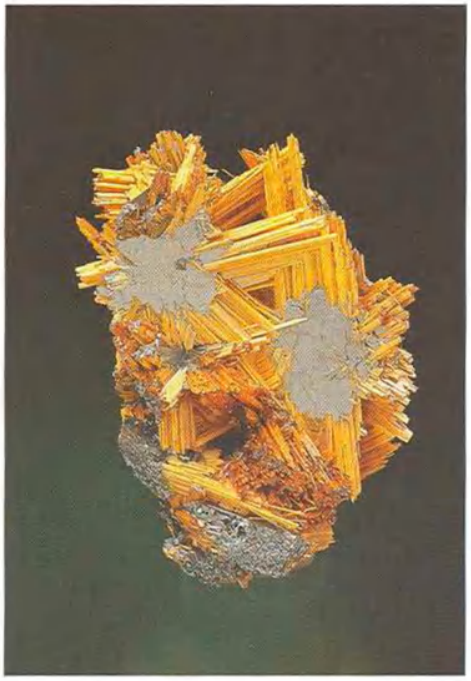

Cassiterite, 110 mm, Yunnan, China

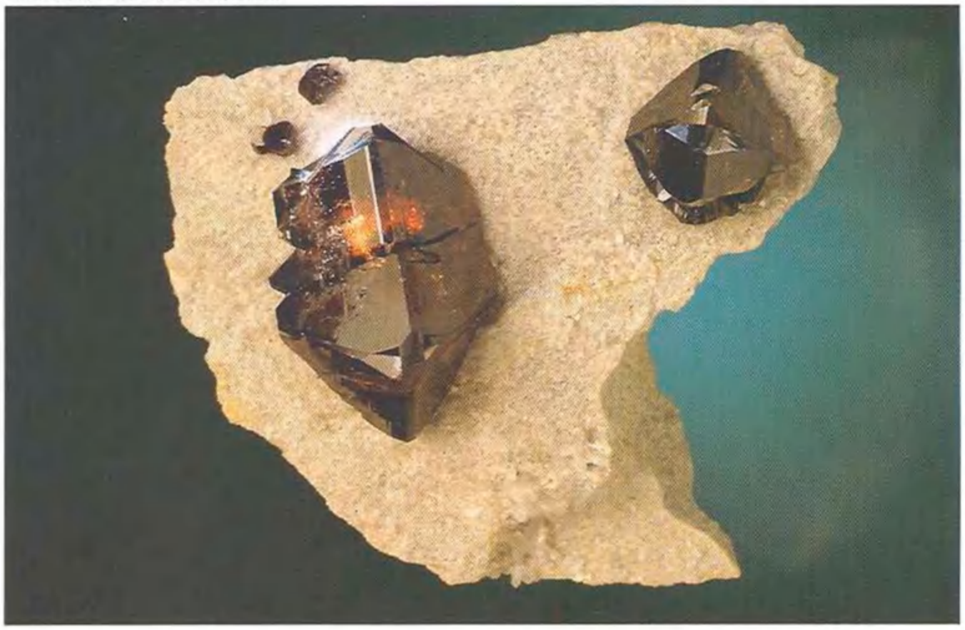

Cassiterite

SnO_2

TETRAGONAL ● ● ● ●

Properties: C – colorless, brown, black; S – white, grayish, brown; L – metallic to adamantine, dull; D – transparent to opaque; DE – 6.3-7.2; H – 6-7; CL – imperfect; F – conchoidal to uneven; M – dipyramidal and short prismatic crystals, multiple twins, granular, massive.
Origin and occurrence: Magmatic in pegmatites, hydrothermal in high-temperature deposits, metamorphic and in placers, together with wolframite, topaz and other minerals. Crystal twins up to 150 mm (6 in) in size come from Horni Slavkov, Czech Republic. Crystals of similar size were also found in Panasqueira, Portugal. Fine twins up to 80 mm ($3\frac{1}{8}$ in) were found in Rossarden, Tasmania, Australia. Crystals up to 70 mm (2¾ in) in size were found in Llallagua, colorless and transparent crystals up to 50 mm (2 in) in size in Viloco, Bolivia. Crystals up to 110 mm ($4\frac{5}{16}$ in) occurred lately in Tenkergin, Chukotka, Russia. Crystals up to 130 mm ($5\frac{2}{16}$ in) in size are known from pegmatites in Minas Gerais, Brazil (Fazenda do Funil). New finds of shiny crystals up to 100 mm (4 in) long were made in Hunan and Yunnan provinces, China.
Application: Sn ore.

Plattnerite

PbO_2

TETRAGONAL ● ●

Properties: C – black; S – brown; L – metallic to adamantine; D – opaque; DE – 9.6; H – 5.5; CL – none; M – acicular crystals, botryoidal aggregates, massive.
Origin and occurrence: Secondary, as a result of the oxidation of other Pb minerals, together with pyromorphite, hemimorphite and other minerals. Fine crystals come from Mina Ojuela, Mapimi, Durango, Mexico and from the Blanchard mine, New Mexico, USA. Botryoidal aggregates, weighing up to 100 kg (220 lb), were found in the Morning mine, Mullan, Idaho, USA.

Plattnerite, 150 mm, Mapimi, Mexico

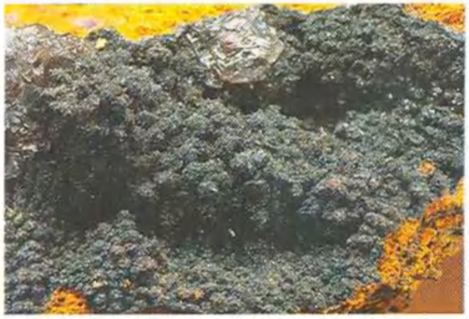

Pyrolusite, 88 mm, Baraga, U.S.A.

Pyrolusite

MnO_2

TETRAGONAL ● ● ● ●

Properties: C – black, steel-gray; S – black; L – metallic to dull; D – opaque; DE – 5.1; H – 6-6.5; CL – perfect; F – uneven; M – prismatic to acicular striated crystals, stalactitic and botryoidal aggregates, granular, massive.
Origin and occurrence: Secondary, as a result of alteration of manganite and other primary Mn minerals, also hydrothermal. Crystals up to 20 mm ($^{25}/_{32}$ in) long come from Horní Blatna, Czech Republic. Radial shiny aggregates were found in Ohrenstock, Germany. It occurred in Ilfeld, Germany, too. Large sedimentary Mn deposits, where pyrolusite is the main constituent, are known near Chiaturi, Georgia or near Nikopol, Ukraine. Crystals are reported also from Tsumeb, Namibia and Hotazel, South Africa.
Application: important Mn ore.

Hollandite

$Ba(Mn^{4+},Mn^{2+})_8O_{16}$

MONOCLINIC ● ● ●

Properties: C – gray-black; S – black; L – submetallic; D – opaque; DE – 5; H – 6; CL – good; M – short prismatic crystals, racemous and columnar aggregates.
Origin and occurrence: Metamorphic in Mn deposits with braunite, scheelite and other minerals, also secondary. Crystals up to 5 mm ($^{3}/_{16}$ in) long come from the Bradshaw mountains, Arizona, USA. It is common in the metamorphosed Mn deposits in Ultevis, Sweden; in Nagpur and Balaghat, India.

Coronadite

$Pb(Mn^{4+},Mn^{2+})_8O_{16}$

TETRAGONAL ● ● ●

Properties: C – black, black-gray; S – brown-black; L – submetallic to dull; D – opaque; DE – 5.4; H – 4.5-5.5; M – botryoidal crusts with fibrous structure, massive.
Origin and occurrence: Secondary in the oxidation zone of Mn deposits. Spherical aggregates up to 100 mm (4 in) in diameter come from the Bou Tazoult mine, Imini, Morocco. Small crystals were found in the Beltana mine, Southern Australia, Australia and in the Silver Bill mine, Arizona, USA.

Todorokite

$(Mn^{2+},Ca,Mg)Mn^{4+}{}_3O_7 . H_2O$

MONOCLINIC ● ● ●

Properties: C – gray, brown-black, black; S – brown;

L – metallic to dull; D – opaque; DE – 3.5-3.8; H – 1; CL – perfect; M – platy crystals, stalactitic and nodular aggregates.
Origin and occurrence: Secondary mineral, as a result of the oxidation of Mn minerals. Crystals are reported from several mines in the Kalahari region, South Africa (Hotazel, Smart). It was originally described from the Todoroki mine, Hokkaido, Japan. It is an important constituent of oceanic Mn concretions.

Ferrotapiolite

$FeTa_2O_6$

TETRAGONAL ● ● ●

Properties: C – black, brown; V – red-brown; L – submetallic, adamantine, resinous; D – opaque; DE – 7-7.8; H – 6-6.5; CL – none; F – uneven to conchoidal; M – dipyramidal and short prismatic crystals, massive.
Origin and occurrence: Magmatic in pegmatites, together with manganotantalite, microlite, cassiterite and other minerals. Crystals up to 40 mm (1$^{9}/_{16}$ in) in size are known from the vicinity of Governador Valadares, Minas Gerais, Brazil. A crystal 120 mm (4$^{7}/_{16}$ in) long has been described from Angarf, Morocco. Short prismatic crystals come from pegmatites near Topsham and Paris, Maine, USA.

Ilmenorutile

(Ti,Nb,Fe)O_2

TETRAGONAL ● ● ●

Properties: C – black; S – brown; L – submetallic; D – opaque; DE – 4.2; H – 6-6.5; CL – good; F – conchoidal to uneven; M – prismatic crystals, granular.
Origin and occurrence: Magmatic in pegmatites, together with schorl, zircon, fluorapatite and other minerals. Prisms several cm long come from Udraz near Pisek, Czech Republic. It is also known from the vicinity of Miass, Ural mountains, Russia and Evje, Norway.

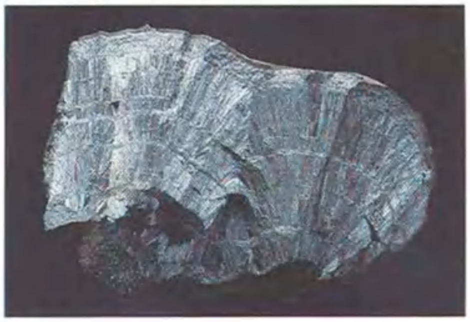

Todorokite, 60 mm, Kamogun, Japan

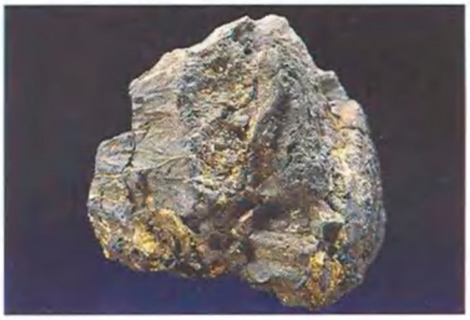

Ferrotapiolite, 60 mm, Maršíkov, Czech Republic

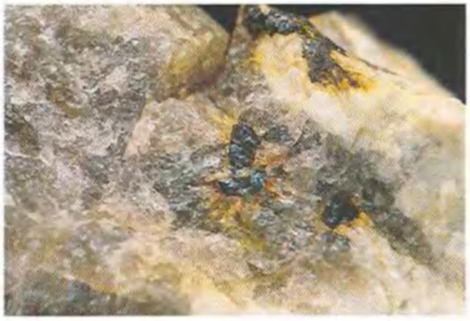

Ilmenorutile, 10mm x, Pisek, Czech Republic

Anatase

TiO_2

TETRAGONAL ● ● ●

Properties: C – black-gray, brown, red-brown, blue, rare colorless; S – white; L – submetallic to adamantine; D – transparent to opaque; DE – 3.8-4; H – 5.5-6; CL – perfect; F – conchoidal; M – dipyramidal and tabular crystals.

Origin and occurrence: Hydrothermal in the Alpine-type veins, associated with brookite and quartz, also sedimentary and metamorphic. Beautiful crystals up to 50 mm (2 in) long were found in Alpa Lercheltini, Binntal, Switzerland. Famous black-blue crystals reaching up to 30 mm ($1\frac{3}{16}$ in) come from Matskorhae, Hardangervidda, Norway. Crystals up to 15 mm ($1\frac{9}{32}$ in) were recently found in Dodo, Polar Ural, Russia. Crystals up to 30 mm ($1\frac{3}{16}$ in) in size occurred in the Old Lot and Vulcan mines, Colorado, USA.

Tellurite

TeO_2

ORTHORHOMBIC ● ●

Properties: C – yellow, yellow-orange; S – yellowish; L – adamantine; D – transparent; DE – 5.8; H – 2; CL – perfect; M – acicular and thin tabular crystals, radial aggregates, pulverulent.

Origin and occurrence: Secondary, resulting from the oxidation of AuBTe ores. Crystals up to 10 mm ($\frac{3}{8}$ in) long occurred in the Kawazu and Susaki mines, Japan. Beautiful specimens were found in Moctezuma, Sonora, Mexico; it also comes from Cripple Creek, Colorado, USA.

Brookite

TiO_2

ORTHORHOMBIC ● ● ●

Properties: C – light to dark brown, yellow-brown, black; S – white to gray; L – submetallic to adamantine; D – transparent to translucent, opaque; DE – 4.1; H – 5.5-6; CL – imperfect; F – conchoidal to uneven; M – tabular, dipyramidal, long and short prismatic crystals.

Origin and occurrence: Hydrothermal along the fissures of the Alpine-type veins and in granitic and alkaline pegmatites; it occurs as pseudo-morphs after titanite and ilmenite; also in sedimentary rocks. Perfect tabular crystals up to 50 mm (2 in) in size found in Rieder Tobel, Switzerland; Magnet Cove, Arkansas, USA; Passo di Viza, Italy. New finds of crystals up to 50 mm in size made in Dodo, Polar Ural, Russia.

Anatase, 16 mm, Hardangervidda, Norway

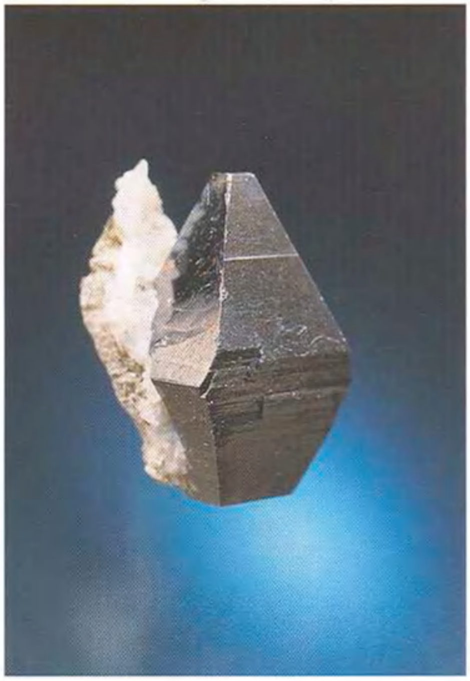

Brookite, 50 mm, Puiva, Polar Urals, Russia

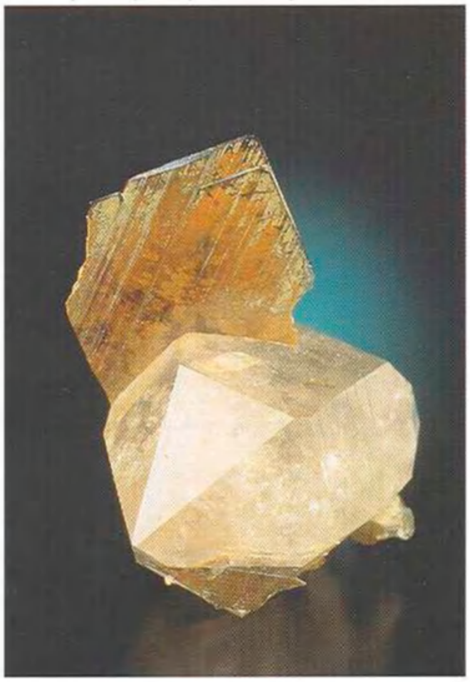

Tellurite, 1 mm x, Fata Baii, Romania

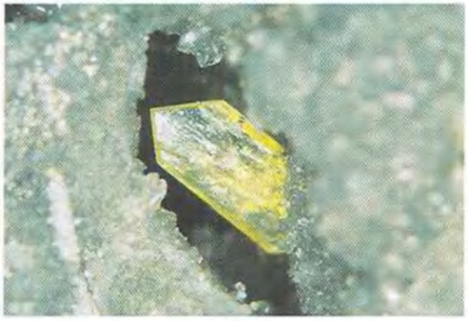

Ferberite

$FeWO_4$

MONOCLINIC ● ● ●

Properties: C – black; S – brown-black to black; L – submetallic; D – opaque; DE – 7.5; H – 4-4.5; CL – perfect; F – uneven; M – short prismatic to tabular crystals, granular, massive.
Origin and occurrence: Hydrothermal in high- to medium-temperature ore veins, in greisens and skarns; rare magmatic in granitic pegmatites and granites; it also occurs in placers. It is usually associated with cassiterite, scheelite, sulfides and quartz. Perfect tabular crystals up to 120 mm ($4^{7}/_{16}$ in) in size found in the Quartz Creek mine, Colorado, USA; also in Cínovec, Czech Republic; Panasqueira, Portugal; Ehrenfriedersdorf, Germany and Potosí, Bolivia.
Application: W ore.

Hübnerite

$MnWO_4$

MONOCLINIC ● ● ●

Properties: C – yellow-brown, red-brown, black; S – yellow-brown to black-gray; L – submetallic; D – translucent to opaque; DE – 7.2; H – 4-4.5; CL – perfect; F – uneven; M – short prismatic to tabular crystals, granular, massive.
Origin and occurrence: Hydrothermal in high- to medium-temperature ore veins, in greisens; rarely magmatic in granitic pegmatites; also in placers. Perfect tabular to short prismatic crystals up to 25 cm ($9^{13}/_{16}$ in) in size, come from the Huayllapon mine, Pasto Bueno, Peru; also from Baia Sprie, Romania; Kara-Oba, Kazakhstan; the Sweet Home mine, Alma and Silverton, Colorado, USA.
Application: W ore.

Ferberite, 68 mm, Mundo Nuovo, Peru

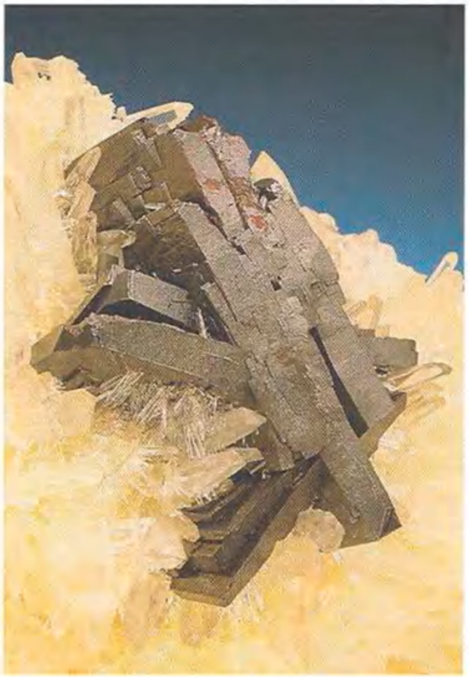

Hubnerite, 39 mm, Silverton, U.S.A.

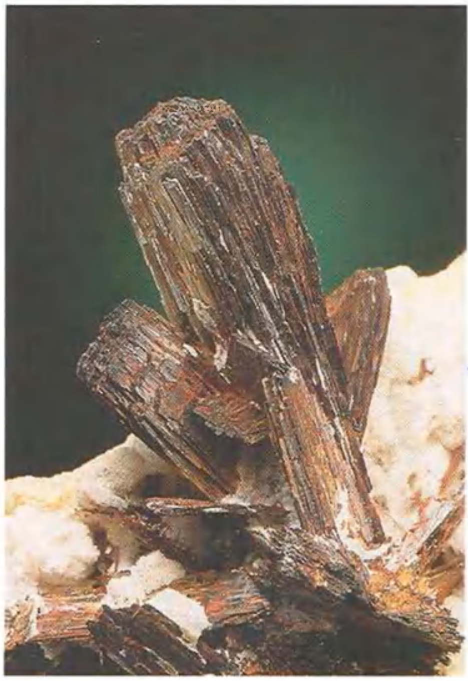

Titanowodginite, 20 mm, Tanco, Canada

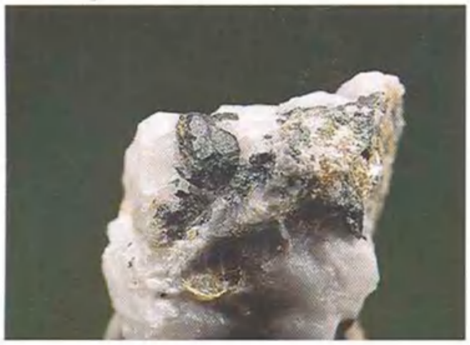

Titanowodginite

MnTiTa$_2$O$_8$

ORTHORHOMBIC ●

Properties: C – dark brown, black; S – dark brown; L – submetallic; D – translucent to opaque; DE – 6.9; H – 5.5; CL – imperfect; F – uneven; M – dipyramidal crystals, granular.
Origin and occurrence: Magmatic in granitic pegmatites. Dipyramidal crystals up to 10 mm ($^{3}/_{8}$ in) long occur in the Tanco mine, Bernic Lake, Manitoba, Canada.
Application: Ta ore.

Ferrocolumbite, 90 mm, Middletown, U.S.A.

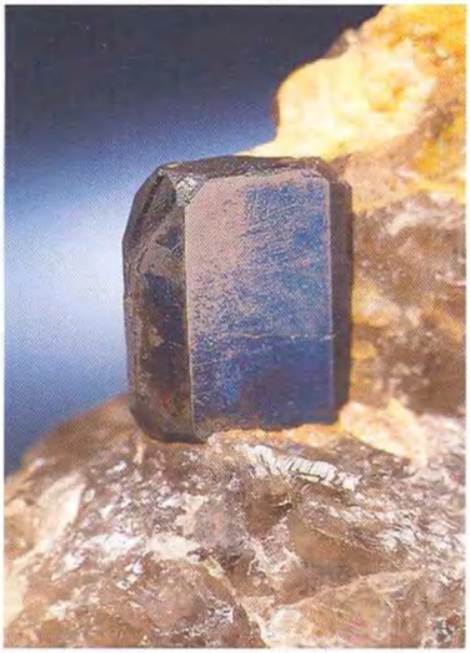

Ferrocolumbite

FeNb$_2$O$_6$

ORTHORHOMBIC ● ● ●

Properties: C – black, red-brown; S – red-brown to black; L – submetallic; D – translucent to opaque; DE – 5.2; H – 6; CL – good; F – uneven to conchoidal; M – long to short prismatic and tabular crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites and granites; rare hydrothermal in high-temperature ore veins and in greisens; also in placers. Tabular crystals up to 1 m ($39^{3}/_{8}$ in) in size occur in granitic pegmatites near Custer and Keystone, South Dakota, USA; in Malakialina, Madagascar; Ichikawa, Japan; masses, weighing up to 270 kg (594 lb) come from the Meyers quarry, Colorado, USA.
Application: Nb ore.

Manganotantalite

MnTa$_2$O$_6$

ORTHORHOMBIC ● ● ●

Properties: C – red, red-brown, black-brown, black; S – dark red to black; L – submetallic; D – translucent to opaque; DE – 8.0; H – 6-6.5; CL – good; F – uneven

Manganotantalite, 19 mm, Nuristan, Afghanistan

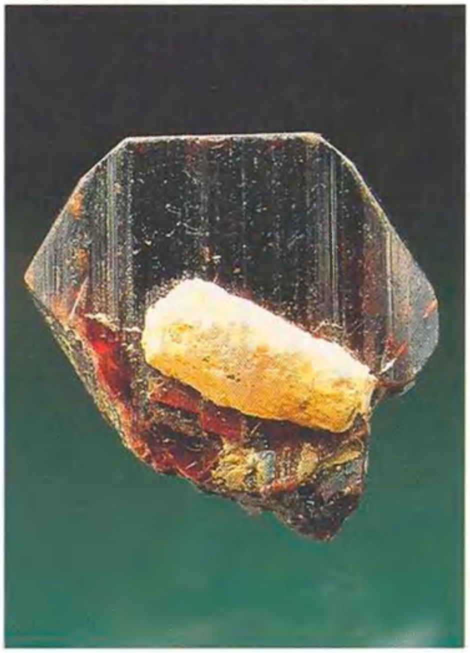

Euxenite-(Y), 40 mm, Ambatofotsy, Madagascar

to conchoidal; M – prismatic to tabular crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites and granites; also in placers. Crystals up to 100 mm (4 in) in size come from Li-bearing pegmatites in the Tanco Mine, Bernic Lake, Manitoba, Canada; also from Sao Jose da Safira, Minas Gerais, Brazil.
Application: Ta ore.

Euxenite-(Y)

$(Y,Ce,U,Th)\ (Nb,Ta,Ti)_2O_6$

ORTHORHOMBIC ● ●

Properties: C – black with brownish and green hues; S – gray, yellowish, brownish; L – submetallic, resinous; D – translucent to opaque; DE – 4.6; H – 6; CL – none; F – conchoidal; M – tabular crystals, granular, massive; R – locally weakly radioactive, commonly metamict.
Origin and occurrence: Magmatic in granitic and alkaline pegmatites; also in placers. Typically associated with monazite-(Ce), zircon, ilmenite and other oxides of rare earth elements. Crystals up to 150 mm (6 in) in size are known from Kragero and Hitterö, Norway; from Ankazobe, Madagascar.

Aeschynite-(Ce)

$(Ce,Ca)\ (Ti,Nb)_2O_6$

ORTHORHOMBIC ● ●

Properties: C – black, red-brown, yellow; S – red-yellow; L – vitreous, resinous, adamantine; D – translucent to opaque; DE – 5.0; H – 5.5; CL – none; F – conchoidal; M – prismatic and tabular crystals, granular, massive; R – locally weakly radioactive, commonly metamict.
Origin and occurrence: Magmatic in alkaline and granitic pegmatites and carbonatites. It is associated with zircon and oxides of rare earth elements. Crystals up to 190 mm (7$^{3}/_{16}$ in) long occur in Quadeville, Ontario, Canada; other localities are Kragero, Norway; Trout Creek Pass, Colorado, USA.

Aeschynite-(Ce), 45 mm, Hittero, Norway

Stibiotantalite

$SbTaO_4$

ORTHORHOMBIC ● ●

Properties: C – yellow, yellow-brown, red-brown, yellow-green; S – yellow-brown; L – submetallic, vitreous, resinous; D – transparent to translucent; DE – 7.5; H – 5-5.5; CL – good; F – uneven to conchoidal; M – prismatic, tabular and dipyramidal crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites; also in placers. Crystals up to 120 mm (4$^{7}/_{16}$ in) in size occur in Muiane, Alto Ligonha, Mozambique; also found in the Little Three mine, Ramona, and the Himalaya mine, Mesa Grande, California, USA. Also known from Greenbushes, Western Australia, Australia.
Application: Ta ore.

Stibiotantalite, 10 mm grain, Dobrá Voda, Czech Republic

Uraninite

UO_2

CUBIC ● ● ● ●

Properties: C – black, black-brown, black-gray; S – black, black-brown to greenish; L – submetallic, greasy, earthy; D – opaque; DE – 7.5-10,6; H – 5-6, earthy aggregates 3; CL – imperfect; F – uneven to conchoidal; M – cubic crystals, botryoidal aggregates, granular, massive; R – strong radioactive.

Origin and occurrence: Mainly hydrothermal in ore veins, skarns; magmatic in granitic pegmatites; in sedimentary rocks; also in placers, usually associated with other U minerals, e.g. coffinite and secondary alteration products, mainly U micas. Perfect crystals up to 100 mm (4 in) in size and weighing up to 2.5 kg (5 lb 8 oz) come from the Fissure mine, Wilberforce, Ontario, Canada, where they occur in simple pegmatites, cross-cutting marbles.

Crystals are also known from Dieresis, Spain and Shinkolobwe, Zair. Rich botryoidal aggregates were found in Jachymov and Slavkovice, Czech Republic; in Bois-Noirs and Margnac, France.

Application: U ore.

Uraninite, 8 mm x, Portland, U.S.A.

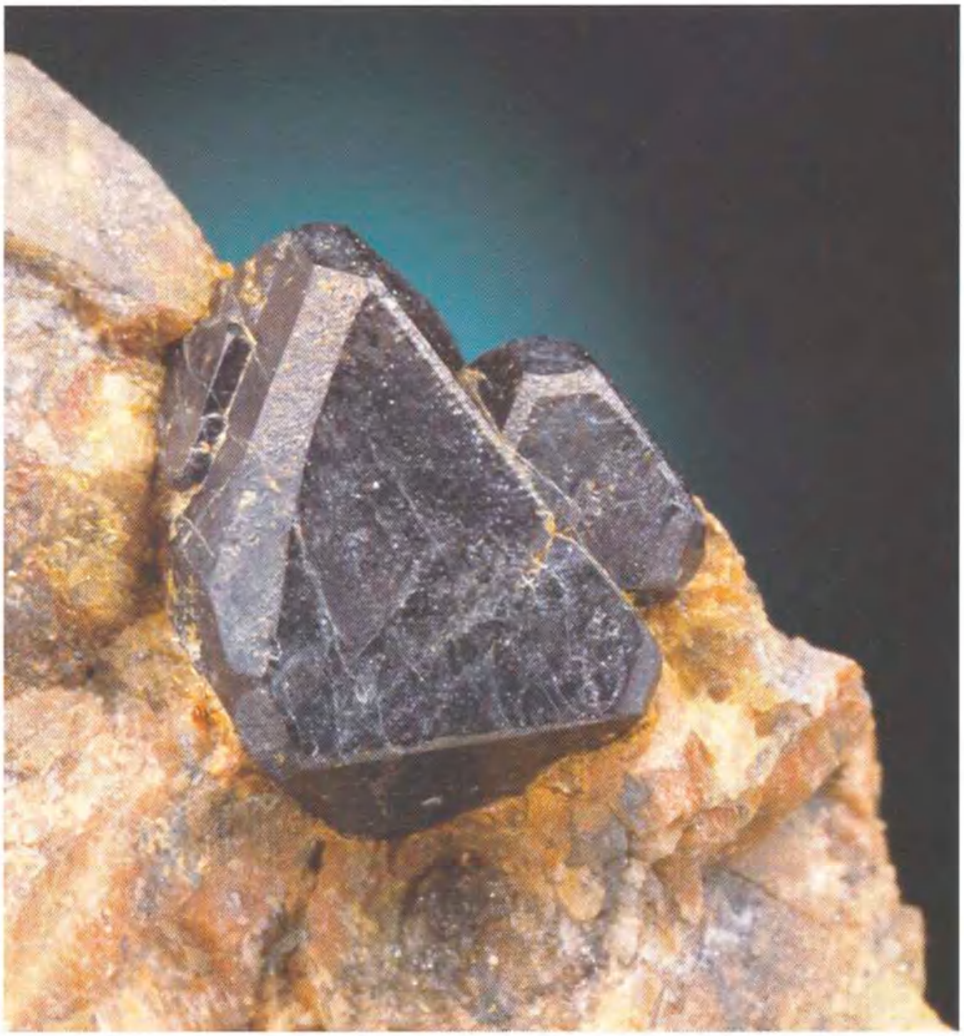

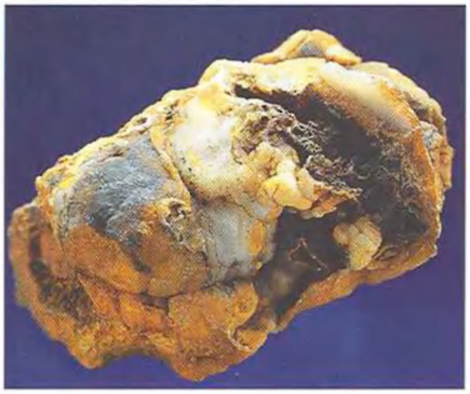

Gibbsite, 80 mm, Gamba, Brazil

Gibbsite

$Al(OH)_3$

MONOCLINIC ● ● ●

Properties: C – colorless, gray, white, greenish; S – white; L – vitreous, pearly; D – transparent to translucent; DE – 2.4; H – 2.5-3.5; CL – perfect; F – uneven; M – tabular crystals, lamellar and earthy aggregates, coatings and stalactitic films, granular, massive.
Origin and occurrence: Hydrothermal as a product of alteration Al-rich rocks; secondary in the oxidation associated with goethite; metamorphic in weakly metamorphosed Al-rich rocks, typically with diaspore; a constituent of bauxites. Tabular crystals up to 100 mm (4 in) in size were found in Zlatoust, Ural mountains, Russia; also in Villa Rica, Minas Gerais, Brazil.

Brucite

$Mg(OH)_2$

TRIGONAL ● ● ●

Properties: C – colorless, gray, white, bluish, blue, yellow, brown; S – white; L – vitreous, pearly; D – transparent to translucent; DE – 2.4; H – 2.5; CL – perfect; F – uneven; M – tabular crystals, foliated, acicular and earthy aggregates, granular, massive.
Origin and occurrence: Hydrothermal in veins in serpentinites or dolomitic marbles, a product of periclase alteration; rare metamorphic in skarns and marbles. Perfect crystals up to 18 cm in size, come from the Low's mine, Pennsylvania and the Tilly Foster mine, New York, USA; also known from Asbestos, Quebec, Canada; Predazzo, the Alps, Italy; blue crystals up to 50 mm (2 in) in size were found in the Bazhenovskoye deposit, Azbest, Ural mountains, Russia.

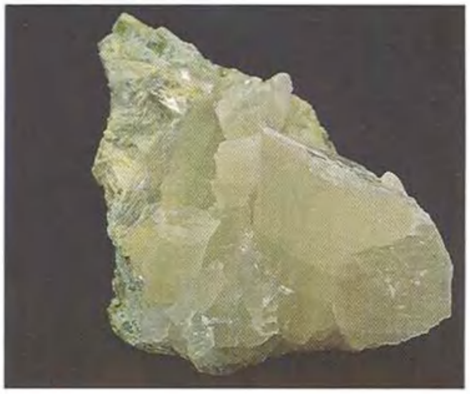

Brucite, 30 mm, Azbest, Russia

Diaspore

$AlO(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, gray, white, greenish, yellowish, pink, purplish; S – white; L – vitreous, pearly; D – transparent to translucent; DE – 3.4; H – 6.5-7; CL – perfect; F – conchoidal; M – tabular crystals, foliated aggregates, stalactitic films, granular, massive.
Origin and occurrence: Hydrothermal as a product of alteration Al-rich minerals, e.g. andalusite, typically with pyrophyllite and corundum; metamorphic in Al-rich rocks; a constituent of bauxites. Tabular crystals up to 120 mm ($4\frac{7}{16}$ in) in size come from Menderess, Turkey; also from Naxos, Greece; Chester, Massachusetts, USA.

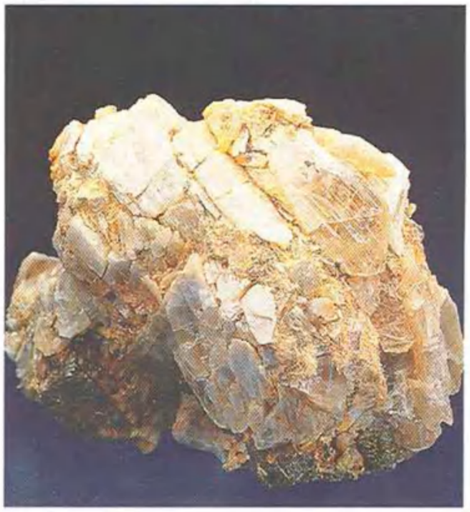

Diaspore, 35 mm, Milas, Turkey

Goethite, 77 mm, Santa Eulalia, Mexico

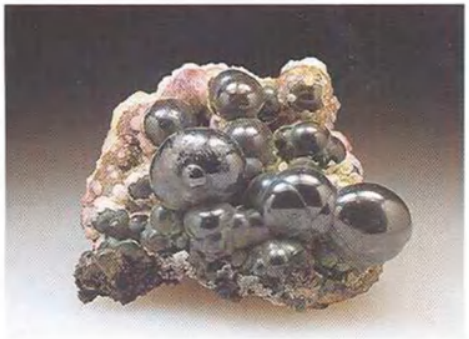

Goethite, 60 mm, Pribram, Czech Republic

Goethite

$Fe^{3+}O(OH)$

ORTHORHOMBIC ● ● ● ● ●

Varieties: velvet ore

Properties: C – black-brown, yellow-brown, brown; S – yellow-brown; L – submetallic, metallic, silky, earthy; D – translucent to opaque; DE – 4.3; H – 5-5.5; CL – perfect; F – uneven to conchoidal; M – acicular to prismatic crystals, botryoidal aggregates, commonly with radial structure, coatings and stalactitic films, earthy, granular, massive.
Origin and occurrence: Secondary as one of the most common minerals of the oxidation zone of ore deposits, it forms a significant part of limonite; hydrothermal in ore veins, in cavities in pegmatites and volcanic rocks. It forms pseudo-morphs after pyrite and other Fe sulfides. Rich botryoidal aggregates of velvet ore with a velvety surface come from Príbram, Czech Republic; acicular crystals up to 50 mm (2 in) long are known from Bottalack and Redruth, Cornwall, UK; it also occurs in Siegen and Horhausen, Germany; in Florissant, Colorado, USA.
Application: Fe ore.

Lepidocrocite, 30 mm, Rudabanya, Hungary

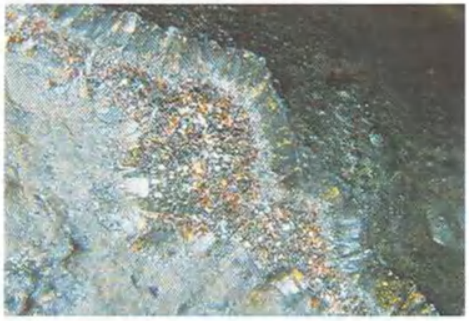

Manganite

$Mn^{3+}O(OH)$

MONOCLINIC ● ● ● ●

Properties: C – black to black-gray; S – red-brown to black; L – submetallic to dull; D – opaque; DE – 4.3; H – 4; CL – perfect; F – uneven to conchoidal; M – long to short prismatic crystals, acicular and earthy aggregates, concretions, granular, massive.
Origin and occurrence: Hydrothermal in low-temperature ore veins, together with quartz; secondary in the oxidation zone of ore deposits; sedimentary and rare metamorphic in Mn-rich rocks. Druses of black crystals up to 40 mm ($1\frac{9}{16}$ in) long come from the classic locality Ilfeld, Germany; it also occurs in Ohrenstock and Ilmenau, Germany; in Nikopol, Ukraine; in Sterling Hill, New Jersey, USA; in the N'Chwaning No. 2 mine, Kuruman, South Africa.
Application: Mn ore.

Lepidocrocite

$Fe^{3+}O(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – dark red to red-brown; S – orange to brick-red; L – submetallic, adamantine to silky; D – transparent to opaque; DE – 4.0; H – 5; CL – perfect; F – uneven to conchoidal; M – tabular to short prismatic crystals, acicular, bladed and earthy aggregates, concretions, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of ore deposits, overgrown on botryoidal goethite. It occurs together with goethite as a constituent of limonite, its tabular crystals and their aggregates are known from Herdorf, Germany and Rancie, France.

Manganite, 85 mm, Ilfeld, Germany

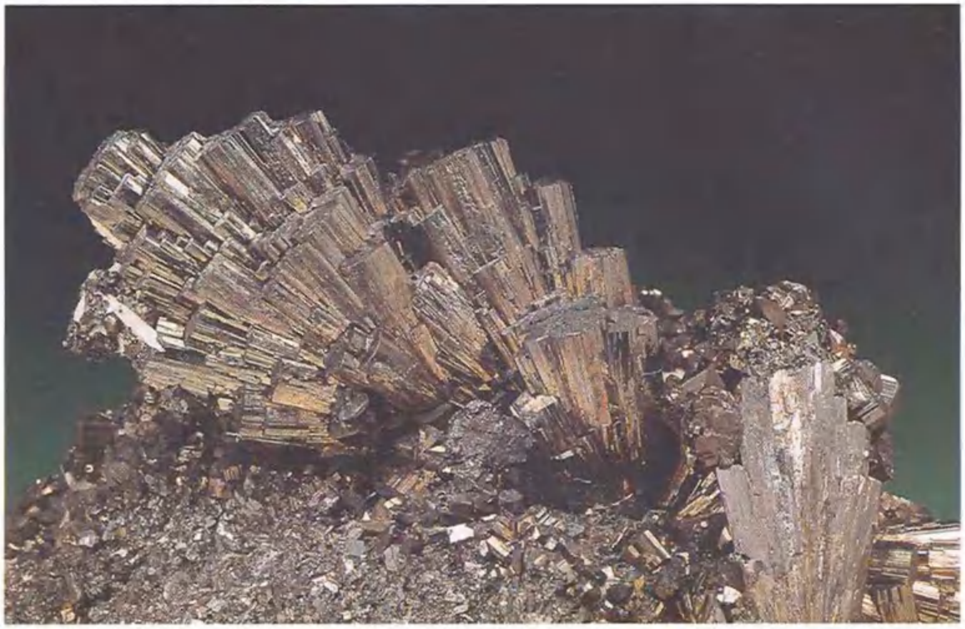

Lithiophorite

$(Al,Li)Mn^{4+}O_2(OH)_2$

MONOCLINIC ● ● ●

Properties: C – black, commonly with bluish tint; S – black-gray to black; L – metallic to dull; D – opaque; DE – 3.3; H – 3; CL – perfect; F – uneven; M – scaly crystals, botryoidal and earthy aggregates, coatings, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of ore deposits and along the cracks in sedimentary rocks. Botryoidal aggregates occur in Schneeberg, Germany; Jivina and Zajecov, Czech Republic; Miyazaki, Japan.

Lithiphorite, 60 mm, Rangersdorf, Germany

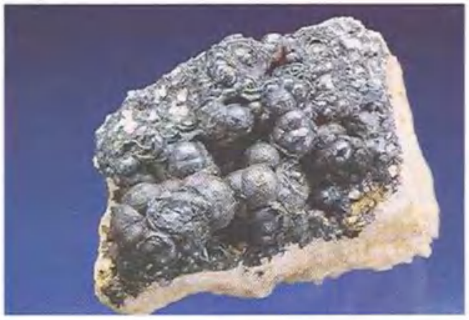

Curite

$Pb_2U_5O_{17} \cdot 4\,H_2O$

ORTHORHOMBIC ● ●

Properties: C – dark orange to red-orange; S – light orange; L – adamantine to earthy; D – transparent to translucent; DE – 7.4; H – 4-5; CL – good; F – uneven; M – acicular crystals, earthy aggregates, coatings, massive; R – strong radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits, associated with other secondary U minerals, e.g. torbernite, kasolite and uranophane. Rich aggregates were found in Shinkolobwe, Zaire; also known from La Crouzille, France; South Alligator, Northern Territory, Australia.

Curite, 30 mm, Shinkolobwe, Zair

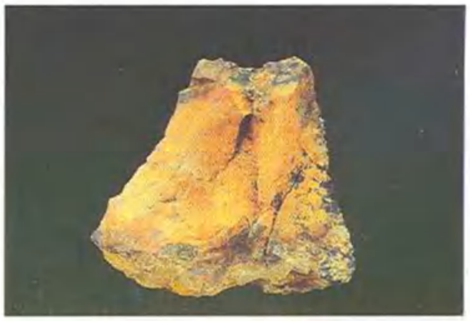

5. Carbonates

Magnesite

$MgCO_3$

TRIGONAL ● ● ● ●

Properties: C – colorless, white, yellowish, brownish, black; S – white; L – vitreous to dull; D – transparent; DE – 3.1; H – 4; CL – perfect; F – conchoidal; M – rhombohedral and prismatic crystals, massive cleavable aggregates, earthy; LU – occasionally blue or green.
Origin and occurrence: Rarely magmatic, mainly hydrothermal metasomatic and metamorphic. The largest crystals are known from Brumado, Bahía, Brazil, reaching up to 1 m ($39^{3}/_{8}$ in) in size, embedded in metamorphosed dolomites. Crystals in cavities in the same locality are up to 50 mm (2 in) in size. Also crystals up to 50 mm (2 in) found in the Eugui quarries, Spain. Crystals up to 10 mm ($^{3}/_{8}$ in) across come also from Val Malenco, Italy. It prevails as massive aggregates, forming huge deposits, like Veitsch, Austria; Liao-Tung, China. Many deposits are located in Slovakia (Jelsavska Dubrava, Hnusta).
Application: heat-resistant material.

Calcite, 40 mm, Houghton Co., U.S.A.
Magnesite, 148 mm, Brumado, Brazil

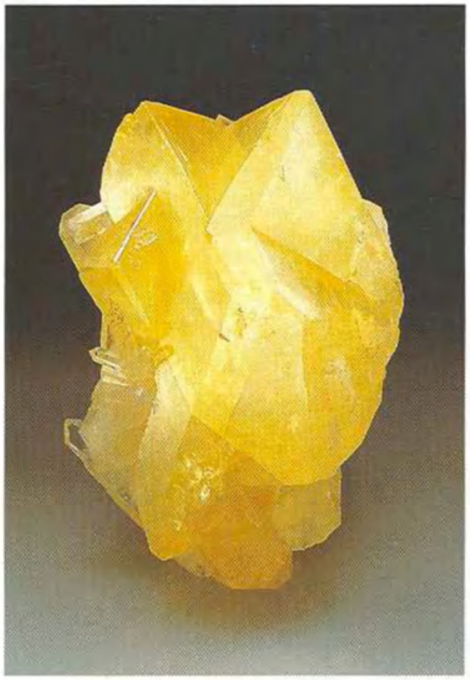

Smithsonite, 32 mm, Tsumeb, Namibia

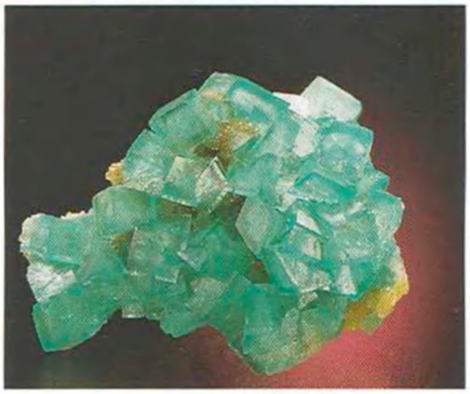

Smithsonite

$ZnCO_3$

TRIGONAL ● ● ●

Properties: C – white, gray, green, pink, blue; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 4.4; H – 4-5; CL – perfect; F – conchoidal to uneven; M – rhombohedral crystals, botryoidal and stalactitic aggregates, massive; LU – sometimes green or blue.
Origin and occurrence: Supergene, as a result of oxidation of the primary Zn ores, associated with other supergene Pb minerals. The largest yellow scalenohedra crystals up to 40 mm ($1^{9}/_{16}$ in) in size come from Broken Hill, New South Wales, Australia. Pink crystals, up to 30 mm ($1^{3}/_{16}$ in) long occurred in Tsumeb, Namibia. World famous blue-green botryoidal crusts up to 100 mm (4 in) thick found in the Kelly Mine, Magdalena, New Mexico, USA. Nice aggregates and banded stalactites described from Monte Poni, Sardinia, Italy.

Smithsonite, 40 mm, New Mexico, U.S.A.

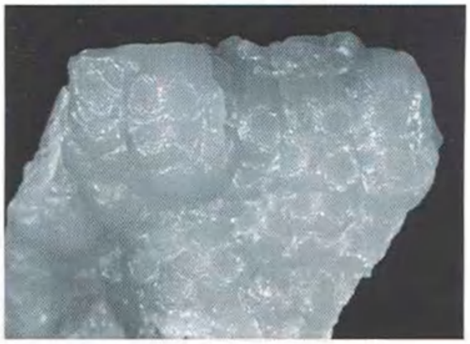

Siderite, 26 mm, Governador Valadares, Brazil

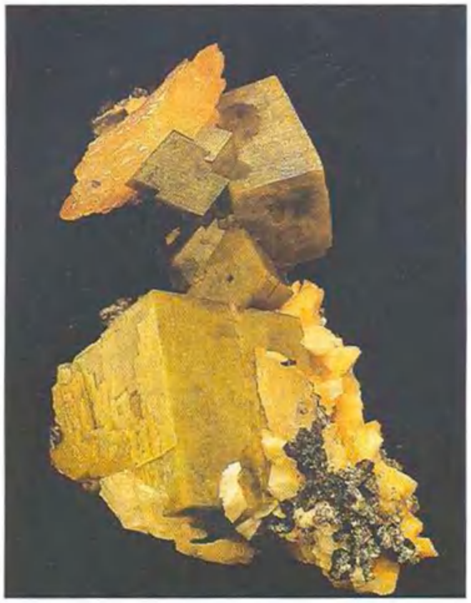

Siderite

$FeCO_3$

TRIGONAL ● ● ● ●

Properties: C – yellow, brown, black; S – yellowish-white; L – vitreous; D – translucent; DE – 4; H-4; CL – perfect; F – uneven to conchoidal; M – rhombohedral crystals, granular, massive.
Origin and occurrence: Hydrothermal in medium- and low-temperature deposits, sedimentary. Crystals up to 40 cm long found in Mont St.-Hilaire, Quebec, Canada. Crystals up to 100 mm (4 in) in size come from Panasqueira, Portugal. Crystals up to 30 mm (1 3/16 in) in size occurred in Neudorf, Germany. Rhombs up to 20 mm (25/32 in) also found in Pribram, Czech Republic. Pseudo-morphs of goethite after siderite up to 70 mm (2¾ in) across described from Pikes Peak, Colorado, USA. Deposits of massive siderite in Erzberg and Huttenberg, Austria yielded crystals up to 50 mm (2 in) long. Fine crystals reported from Tavistock, Devon and Redruth, Cornwall, UK.
Application: important Fe ore.

Sfaerocobaltite

$CoCO_3$

TRIGONAL ● ● ●

Properties: C – pink, gray, brown; S – red; L – vitreous; D – transparent to translucent; DE – 4.1; H – 4; CL – perfect; M – scalenohedral and rhombohedral crystals, radial aggregates, massive.
Origin and occurrence: Secondary, as a product of the oxidation of primary Co minerals. The best specimens, with crystals up to 30 mm (1 3/16 in) long, come from Zaire (Musonoi; Kakanda). Crystals up to 10 mm (4 in) long known from Bou Azzer, Morocco.

Sphaerocobaltite, 16 mm, Bou Azzer, Morocco

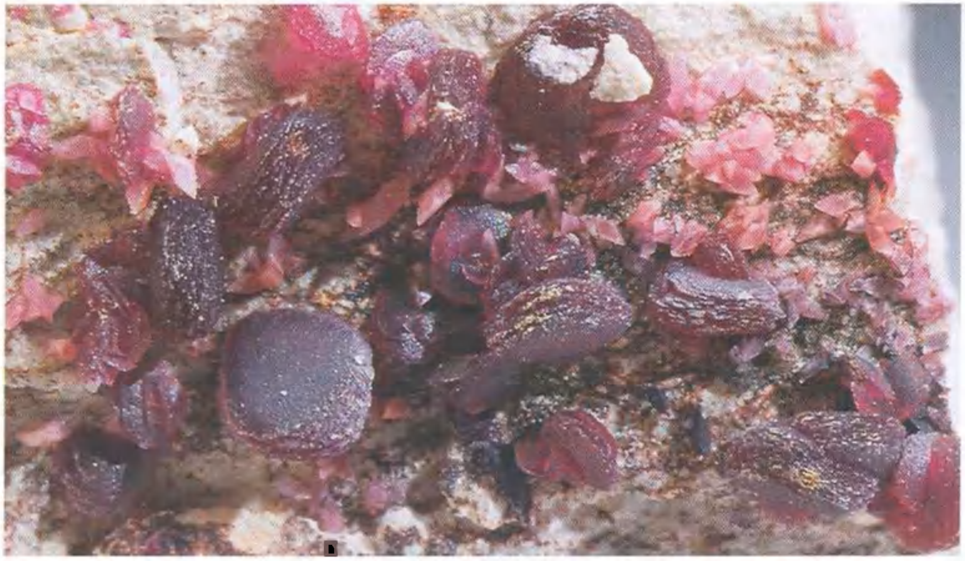

Rhodochrosite, 117 mm, Sweet Home Mine, U.S.A.

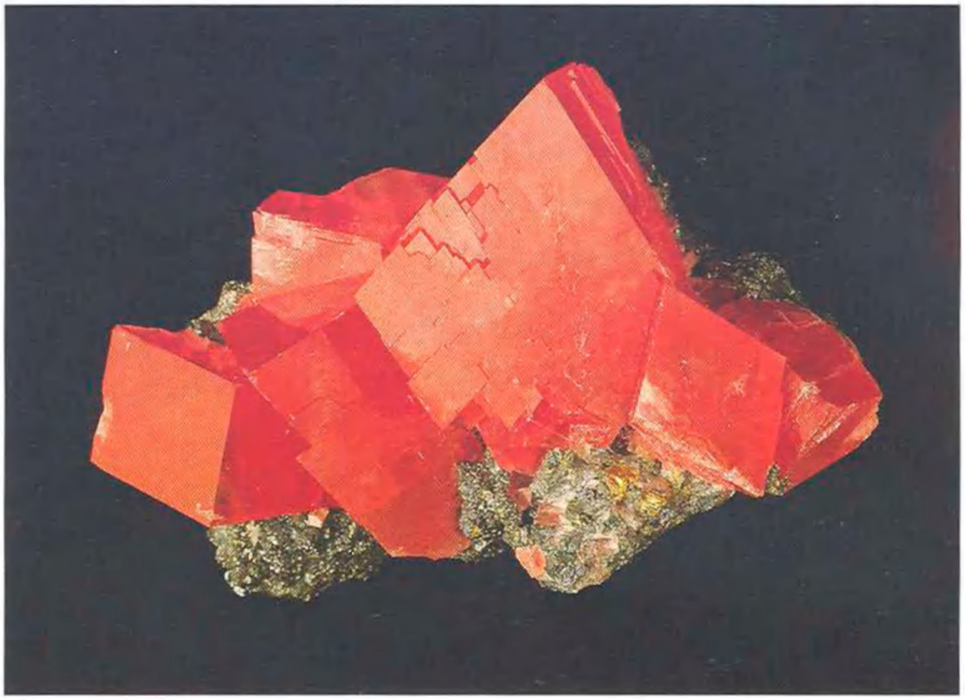

Rhodochrosite, 70 mm, Mina Capillitas, Argentina

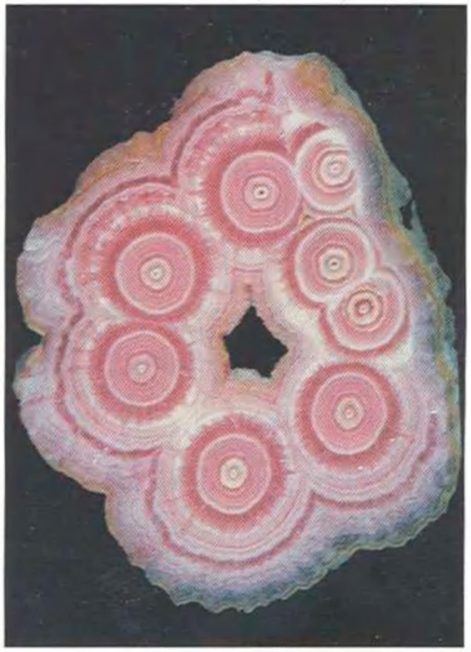

Rhodochrosite

$MnCO_3$

TRIGONAL ● ● ● ●

Properties: C – white, pink, red, brown, locally black coatings on crystals; S – white; L – vitreous; D – transparent to translucent; DE – 3.6; H – 3.5-4; CL – perfect; F – conchoidal to uneven; M – rhombohedral and scalenohedral crystals, hemispherical and botryoidal aggregates, granular, massive.
Origin and occurrence: Only rare in pegmatites, hydrothermal in medium- and low-temperature veins, sedimentary and metamorphic. The most beautiful crystals come from the Sweet Home mine, Alma, Colorado, USA, where rhombs up to 150 mm (6 in) in size occur, associated with purple fluorite, hubnerite, tetrahedrite and other minerals. Beautiful dark red scalehohedra, up to 100 mm (4 in) found in the N'Chwaning No.1 and 2 mines, Kuruman, South Africa. Pink rhombs up to 80 mm ($3\frac{1}{8}$ in) come from Silverton, Colorado, USA. Nice pink hemispheres and botryoidal aggregates are known from Cavnic and Baia Sprie, Romania. Similar specimens occurred in Huaron, Peru. Pink banded crusts and stalactites were found in the Mina Capillitas, Catamarca, Argentina. Pink rhombs, associated with bertrandite, were lately found in Kounrad, Kazakhstan.

Calcite, 150 mm, Elmwood Mine, U.S.A.

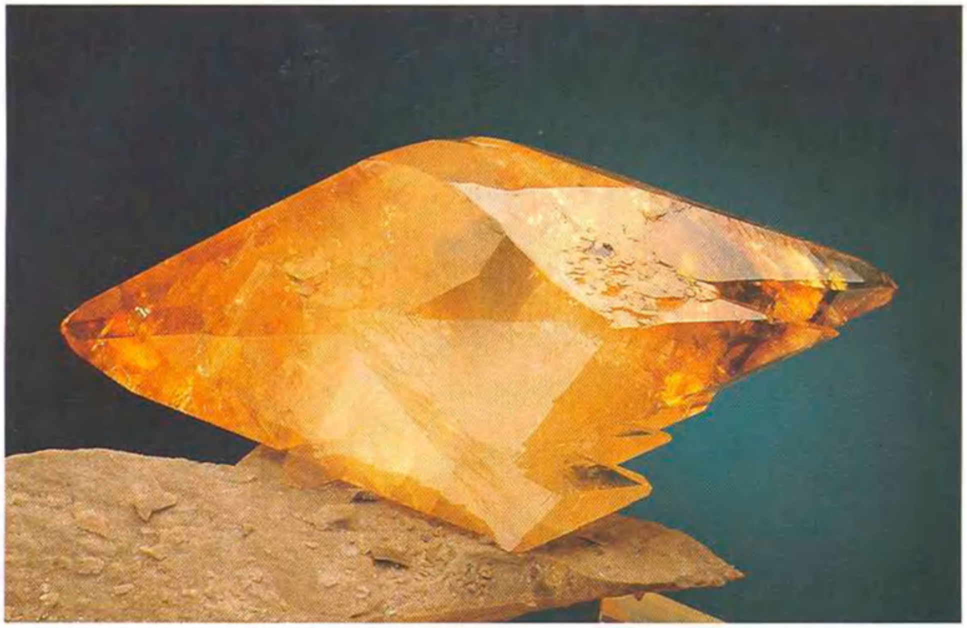

Calcite

$CaCO_3$

TRIGONAL ● ● ● ● ●

Properties: C – colorless, white, gray, yellow, brown, pink, red, blue, green, black; S – white; L – vitreous to pearly; D – transparent to opaque; DE – 2.7; H – 3; CL – perfect; F – conchoidal; M – crystals of various habit, concretions, stalactites, oolitic aggregates, granular, massive; LU – sometimes red to orange.

Calcite, 30 mm, Tunguzka, Russia

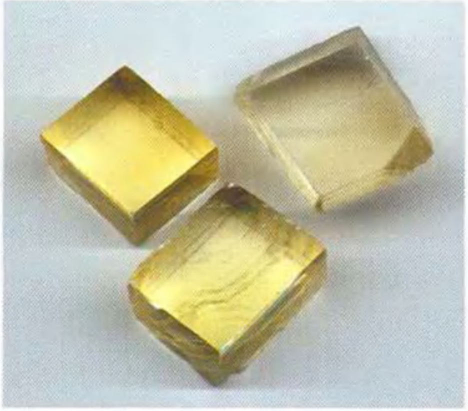

Origin and occurrence: One of the most common minerals, resulting from a wide range of conditions, it is magmatic, hydrothermal, sedimentary, metamorphic and secondary, it occurs in various parageneses. Large crystals found in many localities throughout the world. Pinkish and yellow crystals over 500 mm (20 in) long come from Joplin, Missouri and from the Elmwood mine, Tennessee, USA. Beautiful calcite crystals, crystallographically of very complex habits, found in Dalnegorsk, Russia. Nice calcite specimens occurred also in Mexico (Naica, Chihuahua; Charcas, San Luis Potosí). European localities, like Príbram, Czech Republic; St. Andreasberg, Germany; Kongsberg, Norway are famous by their calcites, too. Wine yellow, complicated crystals are known from the Sarbayskoye deposit in Rudnyi, Kazakhstan. Clear cleavable aggregates of the birefringent calcite (so called Iceland spar) were found in basalt cavities in Helgustadir, Iceland. The largest of them reached up to 6 x 2 m (20 x 6 ft 6 in) in size. Very nice scalenohedra up to 80 mm (3⅛ in) long with copper inclusions occur in Keweenaw Peninsula, Michigan, USA. Perfect scalenohedra and their twins up to 100 mm (4 in) long are known from Egremont and Frizington, UK. Beautiful butterfly twins up to 80 mm (3⅛ in) recently reported from Guiyang, Hunan, China. Crystals of calcite with sand inclusions up to 100 mm (4 in) in size come from the vicinity of Fontainebleau, France.
Application: building industry, optical industry.

Dolomite, 33 mm x, Navarro, Spain

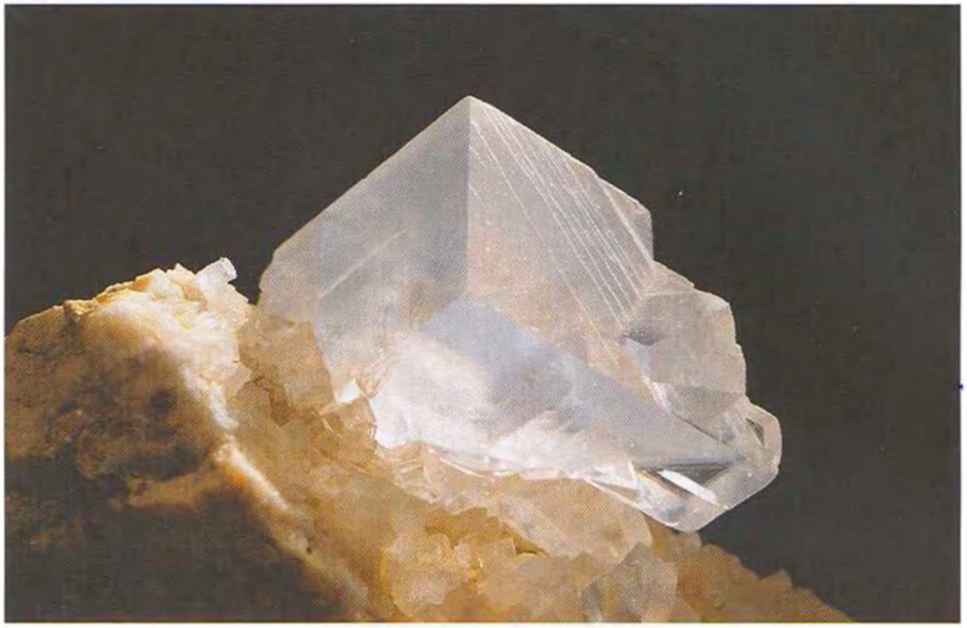

Dolomite

$CaMg(CO_3)_2$

TRIGONAL ● ● ● ● ●

Properties: C – gray-white, pink, red, green, brown, black; S – white; L – vitreous to pearly; D – transparent to translucent, DE – 2.9; H – 3.5; CL – perfect; F – conchoidal; M – rhombohedral crystals, massive.

Origin and occurrence: Magmatic in pegmatites, hydrothermal, metasomatic, sedimentary and metamorphic, together with siderite, magnesite, calcite and other minerals. Crystals up to 100 mm (4 in) long found in Brumado, Bahia, Brazil. Fine crystals also occurred in Banska Stiavnica, Slovakia and in Cavnic, Romania. Crystals up to 200 mm ($7\frac{7}{8}$ in) in size come from Eugui, Spain. Crystals up to 150 mm (6 in) long found in cavities in dolomitic rocks in Lengenbach, Binntal, Switzerland. Crystals several cm long known from Jachymov, Czech Republic. Large accumulations of massive dolomite are common in magnesite deposits.

Application: metallurgy, heat-resistant material.

Ankerite, 20 mm xx, Roudny, Czech Republic

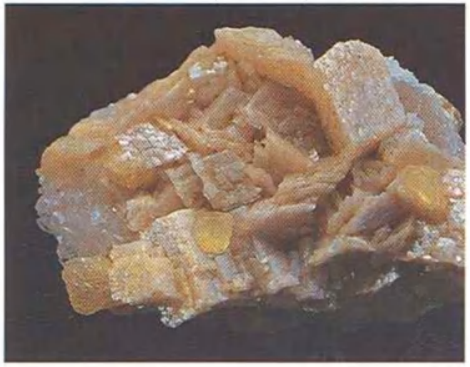

Ankerite

$CaFe(CO_3)_2$

TRIGONAL ● ● ● ●

Properties: C – white, yellowish, brown-yellow; S – white; L – vitreous to pearly; D – translucent; DE – 3; H – 3.5-4; CL – perfect; F – conchoidal; M – rhombohedral crystals, granular.

Origin and occurrence: Hydrothermal in medium- and low-temperature veins, also sedimentary and metamorphic, together with siderite and other minerals. Crystals up to 50 mm (2 in) occurred in the Tui mine, New Zealand. Brown rhombs up to 40 mm ($1\frac{9}{16}$ in) in size known from Gilman, Colorado, USA. Yellowish crystals up to 10 mm ($\frac{3}{8}$ in) come from concretions near Kladno, Czech Republic. Massive aggregates are common in metasomatic deposits of siderite (e.g. Nizna Slana, Slovakia).

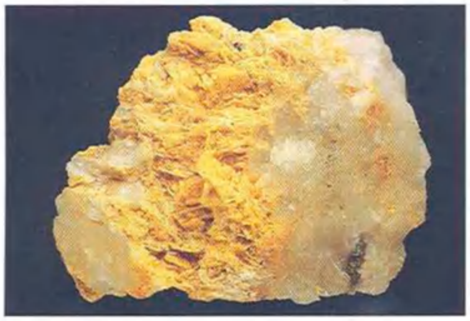
Kutnohorite, 60 mm, Kutna Hora, Czech Republic

Kutnohorite

$CaMn(CO_3)_2$

TRIGONAL ● ● ●

Properties: C – white, gray, pink, yellowish; S – white; L – vitreous; D – translucent; DE – 3.1; H – 3.5-4; CL – perfect; F – conchoidal; M – poorly developed crystals, granular, massive.

Origin and occurrence: Hydrothermal and metamorphic, associated with ankerite, quartz and other minerals. Poorly developed crystals are known from Kutna Hora, Czech Republic. Small crystals several mm in size are described from Mont St.-Hilaire, Quebec, Canada. Large crystals occurred in Moncure, North Carolina, USA. Small gray-white crystals were found in Broken Hill, New South Wales, Australia.

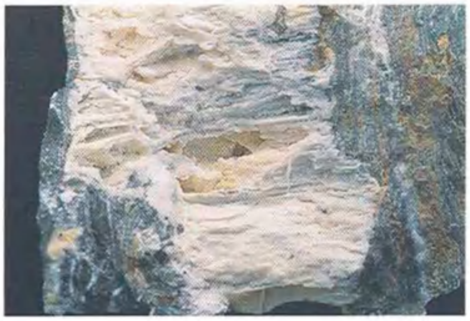
Huntite, 110 mm, Koksin, Czech Republic

Huntite

$CaMg(CO_3)_2$

TRIGONAL ● ●

Properties: C – white; S – white; L – earthy; D – opaque; DE – 2.7; H – 1.5; CL – none; F – conchoidal; M – fibrous aggregates, earthy.

Origin and occurrence: Secondary mineral resulting from the oxidation of dolomite, associated with magnesite and dolomite. Fine fibrous aggregates come

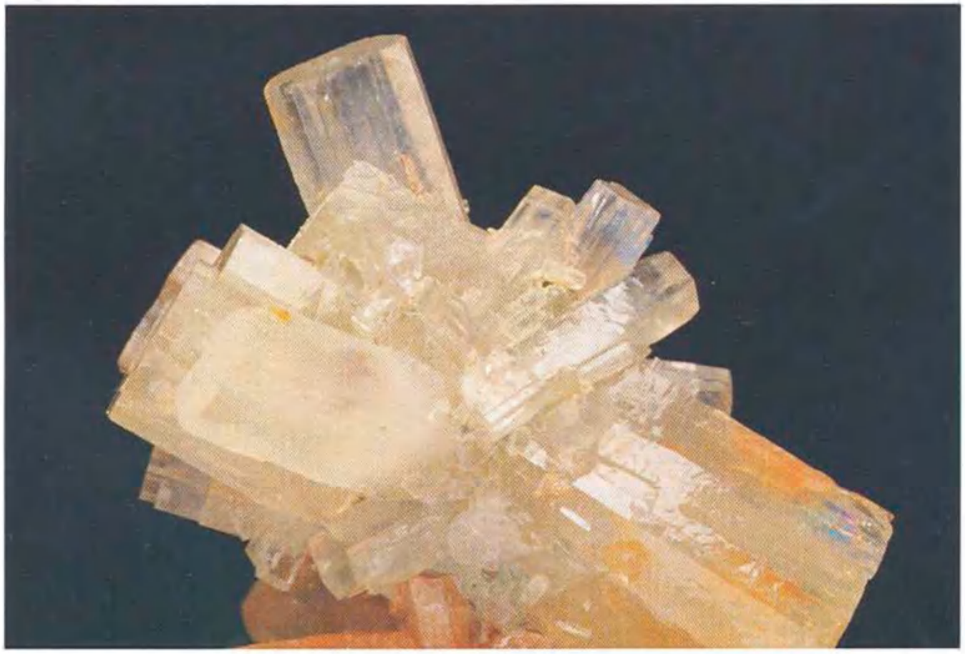
Aragonite, 49 mm, Tazouta, Morocco

Flos ferri, 110 mm, Erzberg, Austria

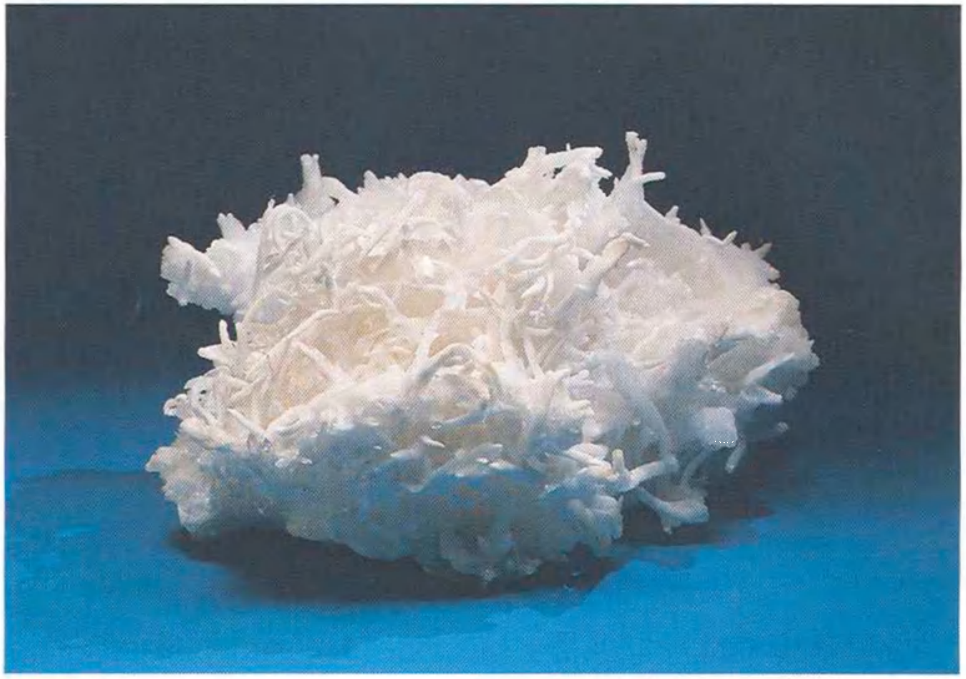

from Koksin, Czech Republic. It also occurs in the Ala-Mar deposit, Nevada, USA. Massive aggregates are known from the Boquira mine, Bahia, Brazil.

Aragonite

$CaCO_3$

ORTHORHOMBIC ● ● ● ●

Varieties: flos ferri, hot-spring tufa, peastone, tarnowitzite.
Properties: C – colorless, white, yellow, reddish, greenish, purplish, bluish, gray; S – white; L – vitreous; D – transparent to opaque; DE – 3; H – 3.5-4.5; CL – imperfect; F – conchoidal; M – prismatic crystals, oolitic, banded, columnar and dendritic aggregates, massive; LU – locally weak cream yellow.

Peastone, 35 mm, Karlovy Vary, Czech Republic

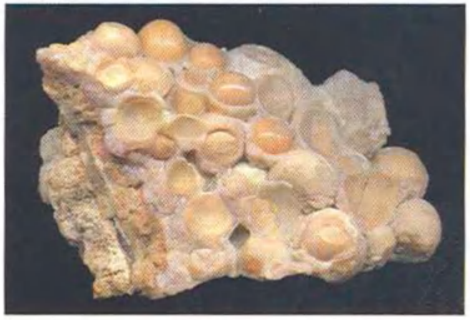

Origin and occurrence: Primary as a late hydrothermal mineral of high-temperature deposits more commonly secondary as a product of the oxidation of siderite and pyrite. It also results from precipitation of thermal springs, it is sedimentary and metamorphic. Fine white prismatic crystals up to 70 mm (2¼ in) long found together with blue celestite crystals in _pania Dolina, Slovakia. Similar crystals come from sulfur deposits in Cianciano, Italy and Tarnobrzeg, Poland. Maybe the best aragonite crystals in the world are known from the gossan of the magnesite deposit in Podrecany, Slovakia, where druses of crystals up to 200 mm (7 7/8 in) long were found. Wine yellow crystals up to 100 mm (3/4 in) long occurred in cavities of volcanic rocks in Horenec near Bilina, Czech Republic. Those crystals were the only gem rough, suitable for facetting, in the world. Very interesting copper pseudo-morphs after aragonite come from Corocoro, Bolivia. Banded and oolitic aggregates found in Karlovy Vary, Czech Republic. Dendritic aggregates from Erzberg, Austria and elsewhere are known as flos ferri.
Application: decorative stone.

Strontianite, 34 mm, Hardin Co., U.S.A.

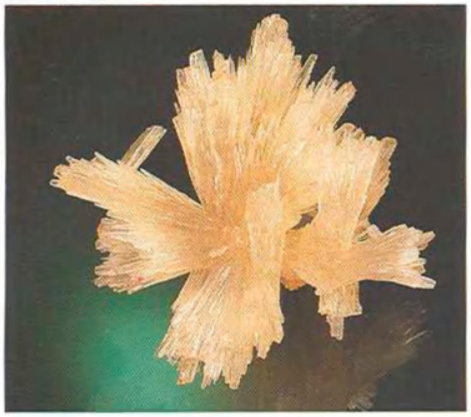

Strontianite

$SrCO_3$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, gray, yellowish, greenish, reddish, brown; S – white; L – vitreous to resinous; D – transparent to translucent; DE – 3.8; H – 3.5; CL – perfect; F – conchoidal to uneven; M – prismatic and acicular crystals, columnar and fibrous aggregates, massive, earthy.
Origin and occurrence: Hydrothermal in low-temperature veins, in cavities of volcanic rocks, mainly sedimentary, together with calcite and zeolites. Crystals up to 80 mm (3⅛ in) long occur near Bleiberg, Austria. Smaller crystals known from Strontian, Scotland. Crystals up to 50 mm long found in marls in Ahlen near Münster, Germany.

Witherite

$BaCO_3$

ORTHORHOMBIC ● ●

Properties: C – white, gray, yellowish; S – white; L – vitreous to greasy. D – translucent; DE – 4.3; H – 3.5; CL – good; F – uneven; M – pseudo-hexagonal dipyramidal crystals, fibrous and botryoidal aggregates, massive.
Origin and occurrence: Hydrothermal in low-temperature deposits with fluorite, barite, calcite and other minerals. Beautiful yellowish crystals up to 120 mm (4$^{11}/_{16}$in) long come from the Minerva No. 1 mine, Cave-in-Rock, Illinois, USA. Crystals, up to 70 mm (2¾ in) in size occurred in Hexham and Alston Moor, England, UK. Its pseudo-hexagonal crystals or botryoidal aggregates were very rare in Pribram, Czech Republic.

Cerussite

$PbCO_3$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, whitish, yellow, black; S – white; L – greasy to adamantine; D – translucent; DE – 6.6; H – 3-3.5; CL – good; F – conchoidal; M –

Witherite, 20 mm x, Alston Moor, UK

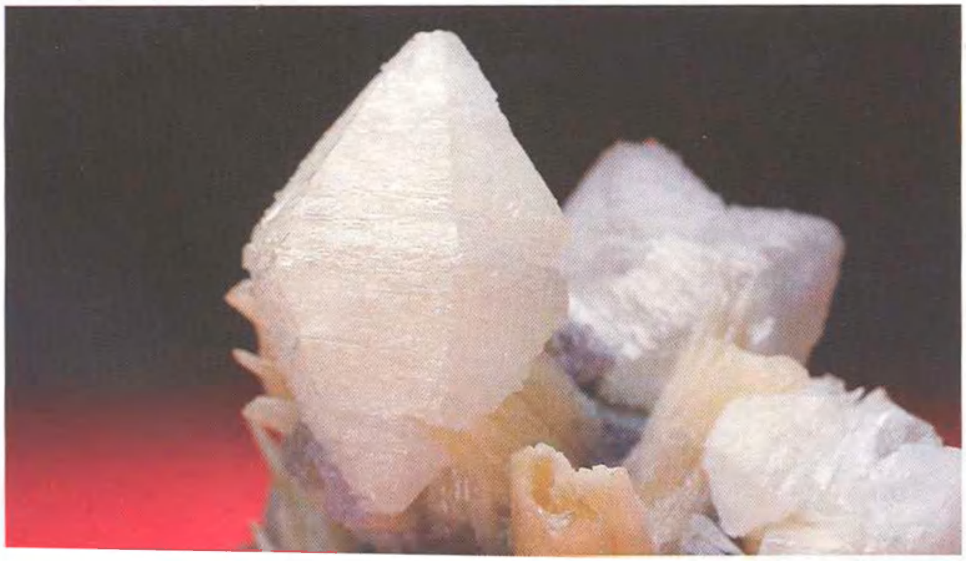

Cerussite, 92 mm, Tsumeb, Namibia

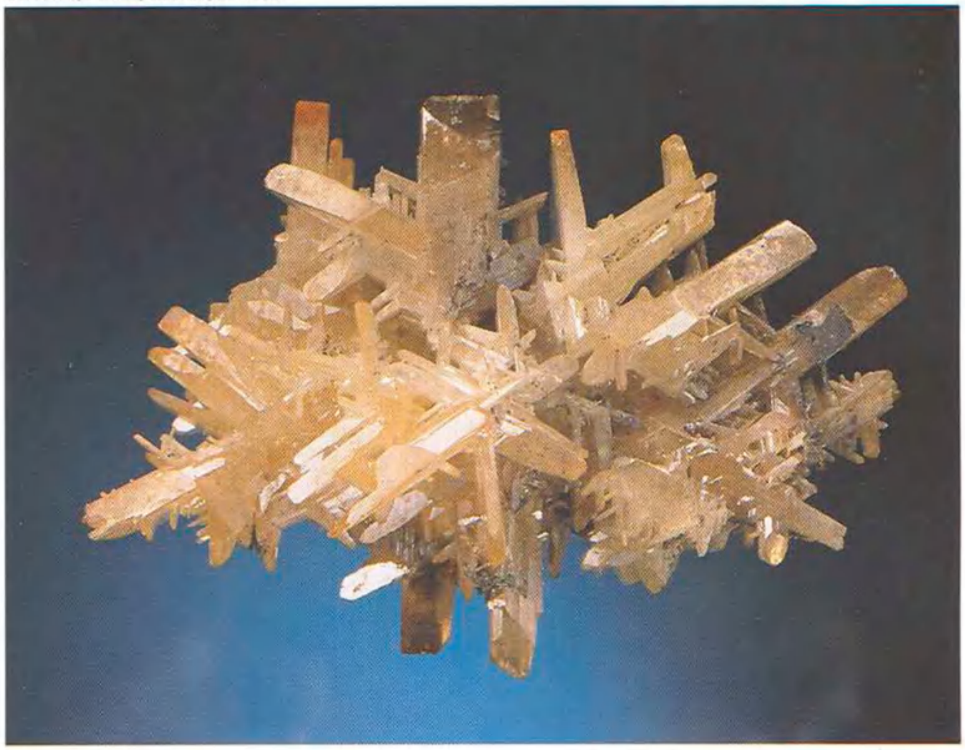

prismatic and pyramidal crystals, common trillings and twins, granular, massive; LU – sometimes yellowish.

Origin and occurrence: Secondary mineral, resulting from the oxidation of galena and other Pb minerals, together with pyromorphite, vanadinite, barite and other minerals. The best specimens come from Tsumeb, Namibia, where trillings up to 200 mm ($7^{7}/_{8}$ in) in diameter occurred. Large twins also known from Broken Hill, New South Wales, Australia. Beautiful crystals up to 50 mm (2 in) in size reported from Mibladen, Morocco. Fine crystals up to 50 mm (2 in) long found in Stribro, Czech Republic. Typical white acicular crystals, up to 60 mm ($2^{3}/_{8}$ in) long come from the Flux mine, Arizona, USA.

Barytocalcite

$BaCa(CO_3)_2$

MONOCLINIC ● ●

Properties: C – colorless, white, gray, yellowish; S – colorless; L – vitreous to resinous; D – transparent to translucent; DE – 3.7; H – 4; CL – perfect; F – conchoidal to uneven; M – prismatic, often striated crystals, massive; LU – light yellow.

Origin and occurrence: Hydrothermal in low-temperature veins, together with calcite, barite and other minerals. Crystals several cm in size and cleavable masses are known from Alston Moor, Cumbria, UK. Imperfect crystals about 10 mm ($^{3}/_{8}$ in) in size occurred in Stříbro, Czech Republic. It is also described from Freiberg, Germany and from Langban, Sweden.

Barytocalcite, 52 mm, Alston Moor, UK

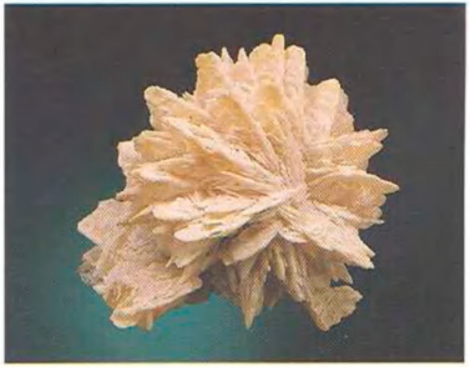

Azurite, 48 mm, Touissit, Morocco

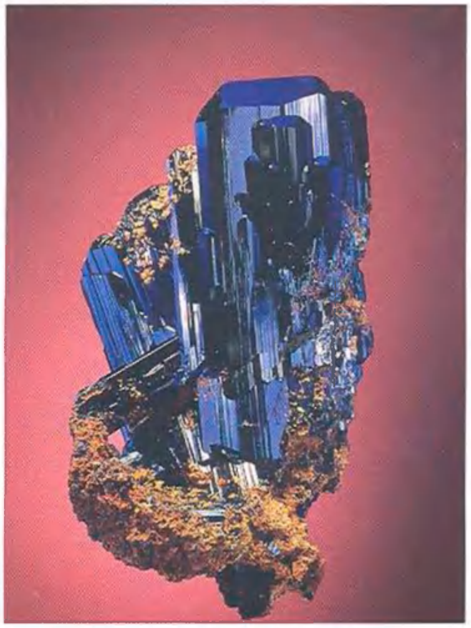

Azurite

$Cu_3(CO_3)_2(OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – blue; S – blue; L – vitreous; D – transparent to opaque; DE – 3.8; H – 3.5; CL – perfect; F – conchoidal; M – tabular and prismatic crystals, pulverulent.
Origin and occurrence: Secondary, resulting from the oxidation of Cu sulfides, mainly associated with malachite. Crystals up to 200 mm ($7^{7}/_{8}$ in) long come from Tsumeb, Namibia. Crystals from Touissit, Morocco reach up to 70 mm (2¾ in). Tabular crystals up to 50 mm (2 in) in size found in Chessy near Lyon, France. Famous crystals up to 50 mm (2 in) across occurred in the Copper Queen mine in Bisbee, Arizona, USA. Fine azurite concretions with crystals on the surface come from La Sal, Utah, USA. Crystal rosettes, reaching up to 130 mm ($5^{3}/_{16}$ in) in size were found in the Yang Chweng Mine, Guang Dong, China.
Application: Cu ore.

Malachite

$Cu_2(CO_3)(OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – green; S – light green; L – vitreous, dull, earthy; D – opaque; DE – 4.1; H – 3.5-4; CL – perfect; F – conchoidal to uneven; M – acicular and prismatic crystals, botryoidal aggregates, stalactites, massive.

Malachite, 63 mm, Mashamba West, Zair

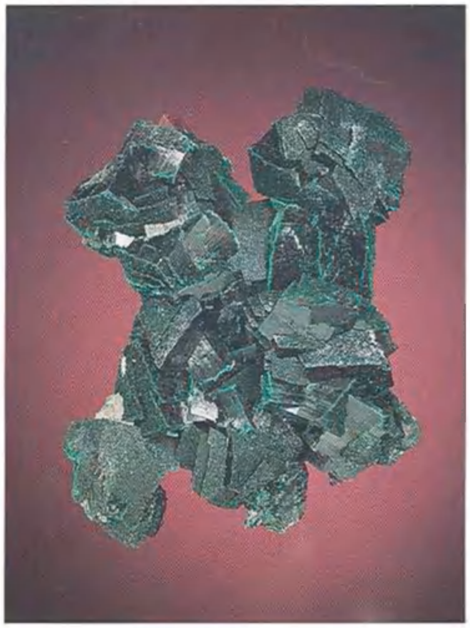

Origin and occurrence: The most common supergene mineral of Cu, associated with azurite, cuprite and other minerals. Crystals up to 30 mm ($1^{3}/_{16}$ in) in size occur in Kambove, Zaire. Crystals up to 20 mm ($^{25}/_{32}$ in) also found in Rudab<nya, Hungary. Pseudomorphs after azurite crystals up to 100 mm (4 in) in size known from Tsumeb; pseudo-morphs after cuprite crystals up to 50 mm in size come from Onganja, Namibia. The blocks of banded malachite weighing up to 250 tons . occurred in Mednorudnjansk, Ural mountains, Russia. Similar material in huge quantities come from many deposits in Shaba

Malachite, 70 mm, Shaba, Zair

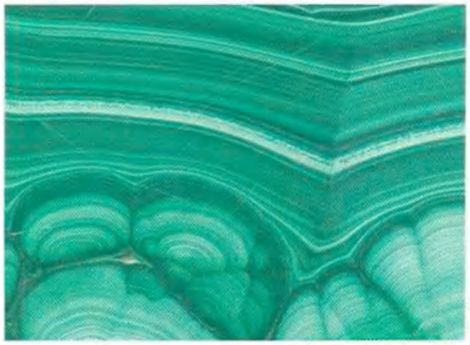

Rosasite, 270 mm. Gleeson, U.S.A.

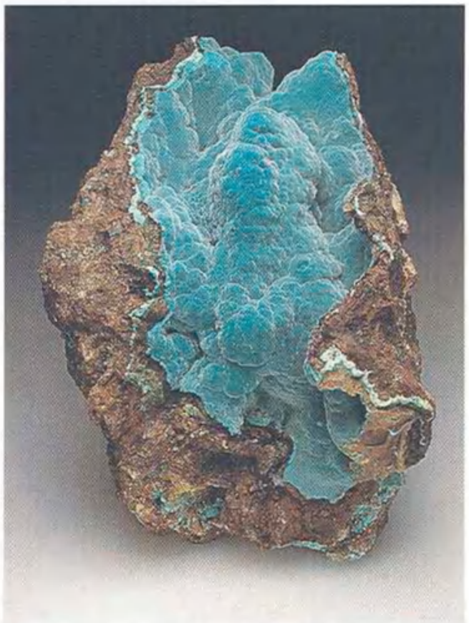

province, Zaire, where stalactites up to 500 mm (20 in) long were also found.
Application: Cu ore, production of decorative objects and jewelry.

Rosasite
$(Cu,Zn)_2(CO_3)(OH)_2$

MONOCLINIC ● ●

Properties: C – green, blue; S – greenish; D – opaque; DE – 4.0-4.2; H – 4.5; CL – good; M – small acicular crystals, fibrous and botryoidal crusts.
Origin and occurrence: Secondary, forming in the oxidation zone of Cu and Zn deposits, together with other secondary minerals of Cu. Spherical aggregates of acicular crystals up to 10 mm ($^3/_8$ in) found in Mina Ojuela, Mapimi, Durango, Mexico. Similar finds were made in Bisbee, Arizona, USA.

Hydrozincite, 65 mm. Tiger, U.S.A.

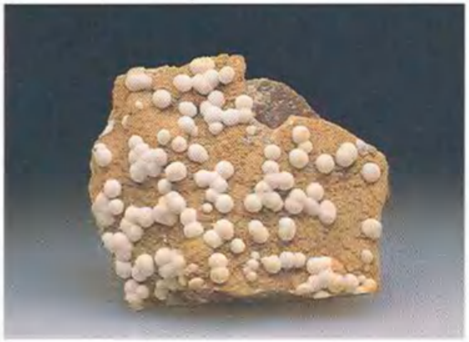

Hydrozincite
$Zn_5(CO_3)_2(OH)_6$

MONOCLINIC ● ● ●

Properties: C – white, yellowish; S – white; L – pearly to dull; D – opaque; DE – 4; H – 2-2.5; CL – perfect; F – conchoidal; M – tabular crystals, massive, earthy; LU – locally blue.
Origin and occurrence: Secondary, resulting from the oxidation of sphalerite, together with cerussite, smithsonite and hemimorphite. Small crystals are known from Mapimi, Durango, Mexico. Spherical aggregates, several mm in diameter, were found in Sterling Hill, New Jersey, USA. Stalactites and thick crusts occurred in Long-Kieng, Burma. Crusts and stalactites were also described from Bleiberg, Austria; from Mezica, Slovenia and from Raibl, Italy.

Aurichalcite
$(Zn,Cu)_5(CO_3)_2(OH)_6$

ORTHORHOMBIC ● ● ●

Properties: C – light green, blue-green, blue; S – blue-green; L – silky to pearly; D – translucent; DE – 4; H – 1-2; CL – perfect; M – acicular crystals, crusts.
Origin and occurrence: Secondary in the oxidation zone of Cu and Zn deposits in the arid climate, associated with linarite and other minerals. Large prismatic crystals come from Mina Ojuela, Mapimi, Durango, Mexico. Nice rosettes of acicular crystals are known from Bisbee and from the 79 mine, Banner district, Arizona, USA. It also occurred in Monteponi, Sardinia, Italy.

Aurichalcite, 80 mm. Arizona, U.S.A.

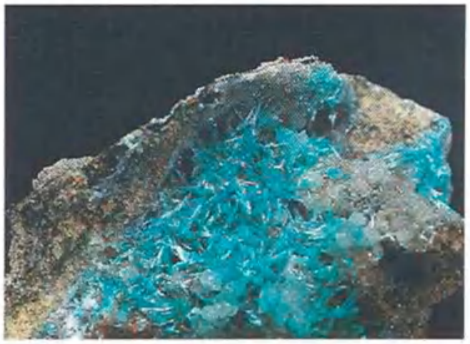

Dawsonite, 2 mm aggregate, Recsk, Hungary

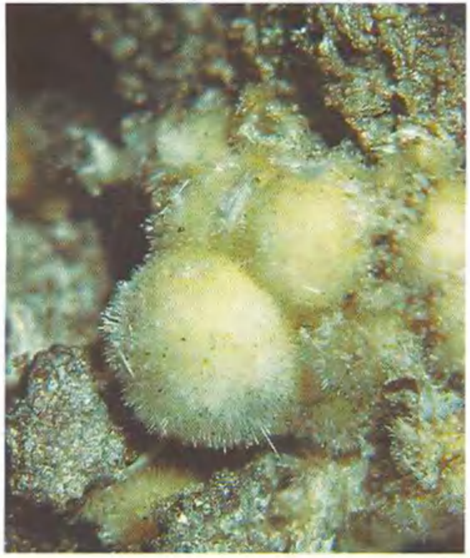

Dawsonite

$NaAl(CO_3)(OH)_2$

ORTHORHOMBIC ● ● ●

Properties: C – colorless to white; S – white; L – vitreous; D – transparent; DE – 2.4; H – 3; CL – perfect; M – acicular to blade-like crystals, radial aggregates.
Origin and occurrence: Hydrothermal in low-temperature deposits, associated with calcite, dolomite and other minerals. Fine acicular crystals up to 35 mm ($1^3/_8$ in) long come from Mont St.-Hilaire, Quebec, Canada. Radial aggregates along the rock cracks were found in Dubník and Zlata Bana, Slovakia.

Bastnäsite-(Ce)

$(Ce,La)(CO_3)F$

HEXAGONAL ● ● ●

Properties: C – yellow to brown; S – yellow-brown; L – vitreous to greasy; D – translucent; DE – 4.8-5.2;

Parisite-(Ce), 71 mm, Boyaca, Colombia

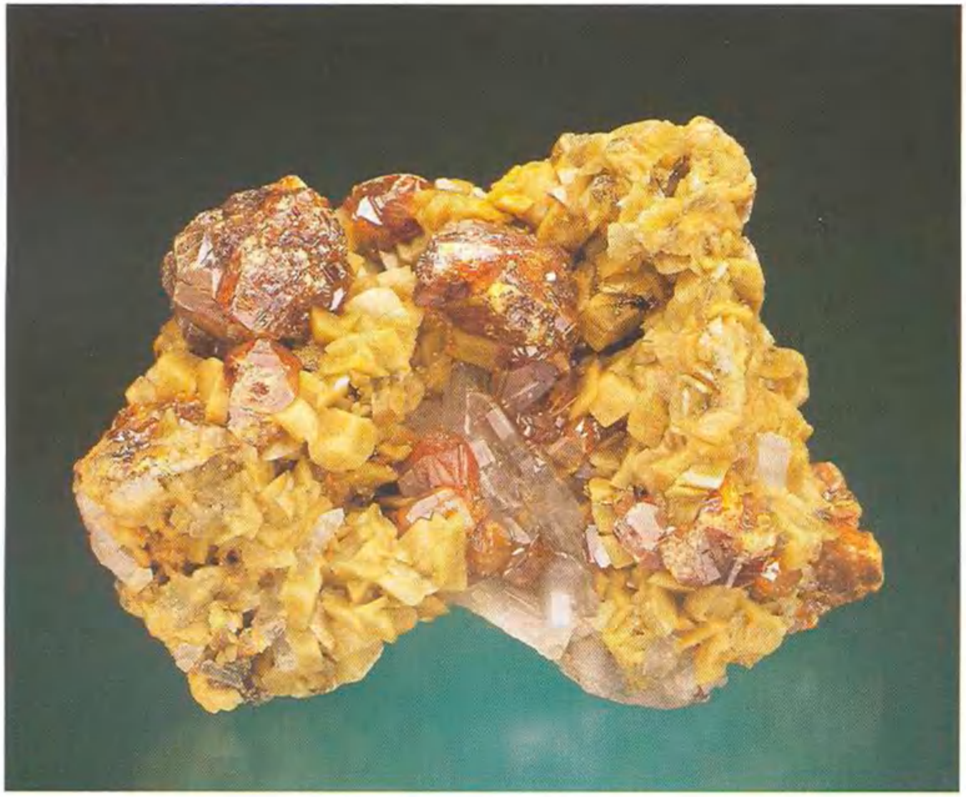

Bastnasite-(Ce). 6 mm x, Ariege, France

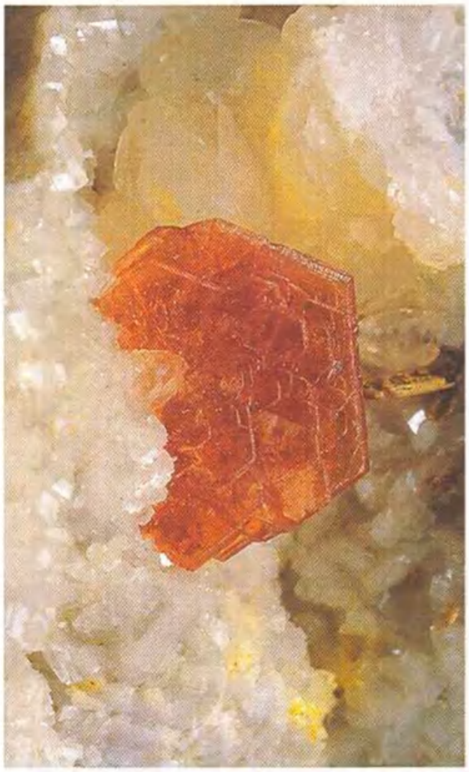

H – 4-4.5; CL – good; F – uneven; M – tabular crystals, granular, massive.
Origin and occurrence: Magmatic in pegmatites, also metamorphic, together with allanite-(Ce) and other rare earth elements minerals. Crystals up to 200 x 150 mm ($7^7/_8$ x 6 in) in size occur in Andekatany, Madagascar. Crystals up to 100 mm (4 in) long are known from Karonge, Burundi. Transparent crystals up to 25 mm (1 in) long were found recently in the Trimouns quarry, France. It forms important local deposits, as those of Mountain Pass, California, USA.*Application:* ore of rare earth elements.

Parisite-(Ce)

$CaCe_2(CO_3)_3F_2$

TRIGONAL ● ● ●

Properties: C – brown, yellow-brown, gray-yellow; S – brownish; L – vitreous to resinous; D – transparent to translucent; DE – 4.4; H – 4.5; CL – perfect; F – conchoidal to splintery; M – dipyramidal striated crystals.
Origin and occurrence: Magmatic in pegmatites, also hydrothermal and rarely metamorphic, together with bastnasite-(Ce) and other minerals of rare earth elements. Crystals several cm long come from Quincy, Massachusetts, USA. Crystals up to 80 mm ($3^1/_8$ in) long occurred in pegmatites near Hundholmen, Norway. Crystals up to 23 cm ($9^1/_{16}$ in) long found in the Snowbird mine, Montana, USA. Transparent crystals up to 15 mm ($1^9/_{32}$ in) were lately reported in the Trimouns quarry, France. Very unusual association have been described from Muzo, Columbia, where crystals up to 50 mm occur together with emeralds.

Phosgenite

$Pb_2(CO_3)Cl_2$

TETRAGONAL ● ●

Properties: C – colorless, white, yellow-white, gray, brown; S – white; L – adamantine; D – transparent to translucent; DE – 6.1; H – 2-3; CL – good; F – conchoidal; M – short to long prismatic and tabular crystals, granular, massive; LU – sometimes yellow.
Origin and occurrence: Secondary, resulting from the oxidation of galena, associated with cerussite and other secondary Pb minerals. Crystals up to 150 x 100 mm (6 x 4 in) across known from Monteponi, Sardinia, Italy. Crystals from Tsumeb, Namibia reached up to 100 mm. Crystals up to 30 mm ($1^3/_{16}$ in) in size found in Matlock, Derby, UK. Crystals up to 35 mm ($1^3/_8$ in) long occurred in the Mammoth mine, Tiger, Arizona, USA.

Phosgenite. 28 mm, Monteponi, Italy

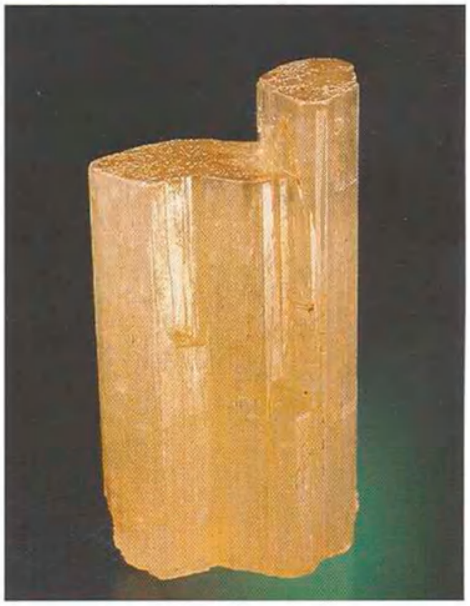

Bismutite, 3 mm xx, Brdarka, Slovakia

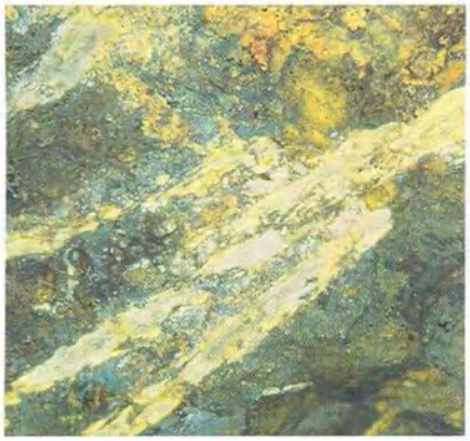

Bismutite

$Bi_2O_2(CO_3)$

TETRAGONAL ● ● ●

Properties: C – yellow, brown, gray, blue, black; S – white; L – vitreous, pearly to dull; D – translucent to opaque; DE – 6.1-7.7; H – 3.5; CL – good; M – spherical, radial aggregates, massive, pulverulent.
Origin and occurrence: Secondary, originated from the oxidation of Bi minerals, associated with bismuthinite and other minerals. It is common from pegmatites in Madagascar (Ampangabe) and in Mozambique. Large pebbles found in Kivu province, Zaire. It also occurred in Tasna, Bolivia.

Weloganite, 30 mm xx, Francon Quarry, Canada

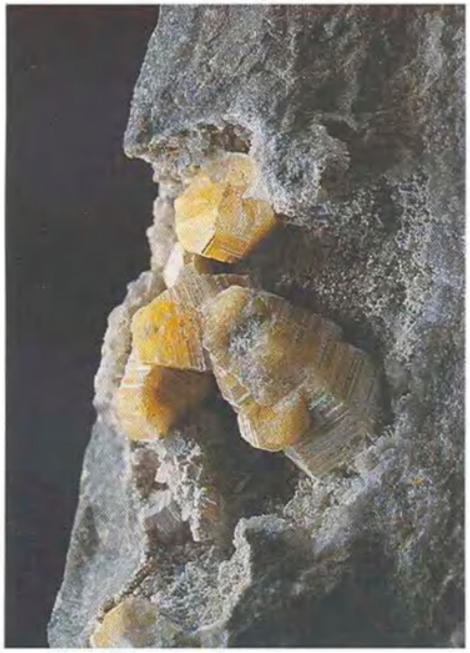

Gaylussite, 30 mm, Natron Lake, Tanzania

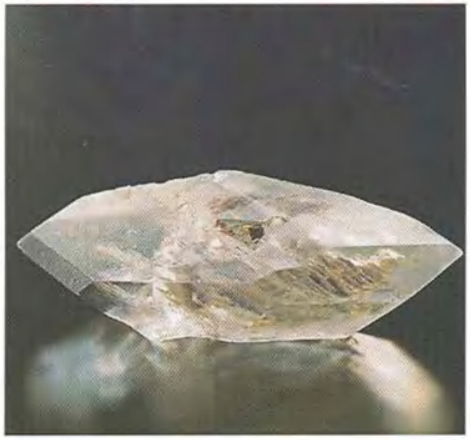

Gaylussite

$Na_2Ca(CO_3)_2 \cdot 5\ H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, gray, yellowish; S – colorless; L – vitreous; D – transparent to translucent; DE – 2; H – 2.5-3; CL – perfect; F – conchoidal; M – lenticular to prismatic crystals.
Origin and occurrence: Sedimentary, typical a constituent of salt sediments. Crystals up to 80 mm ($3\frac{1}{8}$ in) long come from Searles Lake, California, USA. It also occurs in Borax Lake and Mono Lake, California, USA. Large crystals are known from Amboseli Lake, Kenya.

Weloganite

$Na_2Sr_3Zr(CO_3)_6 \cdot 3\ H_2O$

TRICLINIC ●

Properties: C – white, yellow; S – white; L – vitreous; D – transparent to opaque; DE – 3.2; H – 3.5; CL – perfect; F – conchoidal; M – pseudohexagonal striated crystals.
Origin and occurrence: Hydrothermal in the alkaline rocks, together with zircon, dresserite and other minerals. Crystals up to 50 mm (2 in) come from cavities in the Francon quarry, Montreal; crystals up to 30 mm ($1\frac{3}{16}$ in) are known from St. Michel, Quebec, Canada.

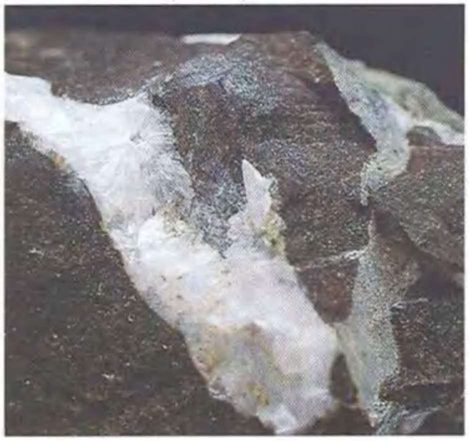

Artinite, 10 mm, Fethiya, Turkey

Artinite

$Mg_2(CO_3)(OH)_2 . 3\ H_2O$

MONOCLINIC ● ● ●

Properties: C – white; S – white; L – silky; D – transparent; DE – 2; H – 2.5; CL – perfect; M – spherical, radial aggregates, veinlets, crusts.
Origin and occurrence: Hydrothermal, originating at low temperatures in serpentinites, associated with magnesite, aragonite and other minerals. Radial aggregates up to 20 mm ($^{25}/_{32}$ in) in size occur in the Gem mine, San Benito County, California, USA. Needles up to 20 mm ($^{25}/_{32}$ in) long come from Val Malenco, Italy. Clusters of acicular crystals were found on Staten Island, New York, USA.

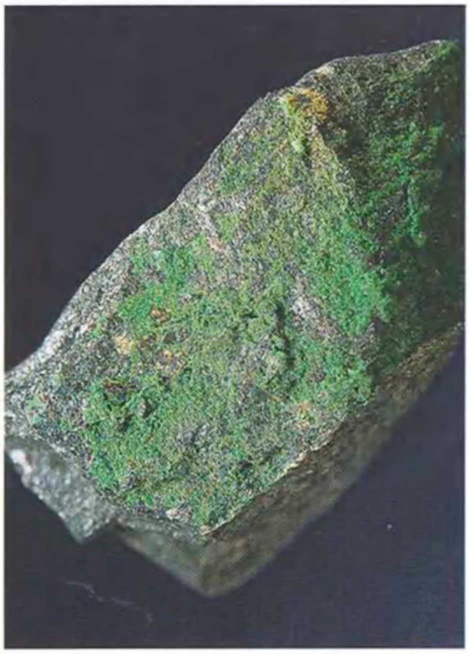

Zaratite, 70 mm, Lawcaster Co., U.S.A.

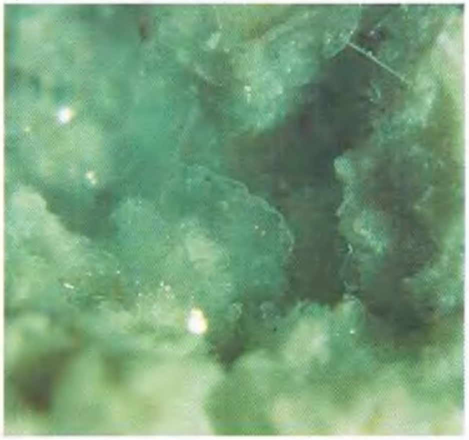

Hydrotalcite, 1 mm xx, Dunabogdány, Hungary

Zaratite

$Ni_3(CO_3)(OH)_4 . 4\ H_2O$

CUBIC ● ● ●

Properties: C – emerald-green; S – light green; L – vitreous to greasy; D – transparent to translucent; DE – 2.6-2.7; H – 3.5; CL – none; F – conchoidal; M – crystalline crusts, stalactites, coatings.
Origin and occurrence: Secondary, originating from the oxidation of Ni minerals, associated with millerite, brucite and other minerals. It occurs as a product of the oxidation of Ni minerals in ultrabasic rocks in Kraubat and Stubachtal, Austria. Large green coatings found in Heazlewood, Tasmania, Australia. It covers millerite needles in the vicinity of Kladno, Czech Republic.

Hydrotalcite

$Mg_6Al_2(CO_3)(OH)_{16} . 4\ H_2O$

TRIGONAL ● ●

Properties: C – white; S – white; L – pearly to waxy; D – transparent; DE – 2.1; H – 2; CL – perfect; M – fibrous and layered aggregates, massive.
Origin and occurrence: : Hydrothermal in ultrabasic and metamorphic rocks. It occurred with serpentine in Nordmark, Norway and also found at Franklin and Sterling Hill, New Jersey, USA.

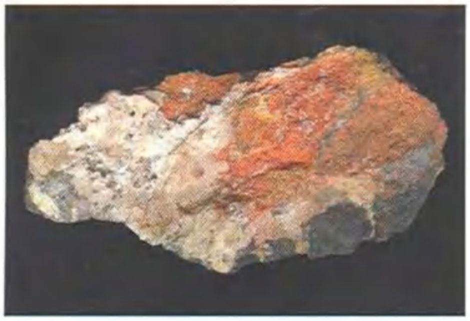

Stichtite, 35 mm, Rouchovany, Czech Republic

Stichtite

$Mg_6Cr_2(CO_3)(OH)_{16} . 4\ H_2O$

TRIGONAL ● ●

Properties: C – pink to purple; S – white to light purple; L – pearly, waxy to greasy; D – translucent; DE – 2.2; H – 1.5-2; CL – perfect; M – lamellar and fibrous aggregates, massive.
Origin and occurrence: It occurs in serpentinites as scaly aggregates and veinlets in Bou Azzer, Morocco; in Dundas, Tasmania, Australia and in Barberton, South Africa.

Alumohydrocalcite

$CaAl_2(CO_3)_2(OH)_4 . 3\ H_2O$

TRICLINIC ● ●

Properties: C – white, gray; S – white; L – earthy; D – opaque; DE – 2.2; H – 2.5; CL – perfect; M – chalky aggregates, consisting of acicular crystals.
Origin and occurrence: Secondary, associated with alophane and other minerals. It was described from the Khakasy deposit, Siberia, Russia, where it originates from the oxidation of alophane. Nice white radial aggregates occurred along the cracks in shales in L'adomírovo, Slovakia.

Ancylite-(Ce)

$Sr_3(Ce,La)_4(CO_3)_7(OH)_4 . 3\ H_2O$

ORTHORHOMBIC ● ●

Properties: C – colorless, yellow, yellow- brown, light purple, brown; S – white; L – vitreous to greasy; D – transparent to opaque; DE – 4; H – 4-4.5; CL – none; F – splintery; M – short to long prismatic and pseudo-octahedral crystals.
Origin and occurrence: Hydrothermal in alkaline rocks, associated with nepheline and other minerals. Crystals up to 6 mm (¼ in) long are known from

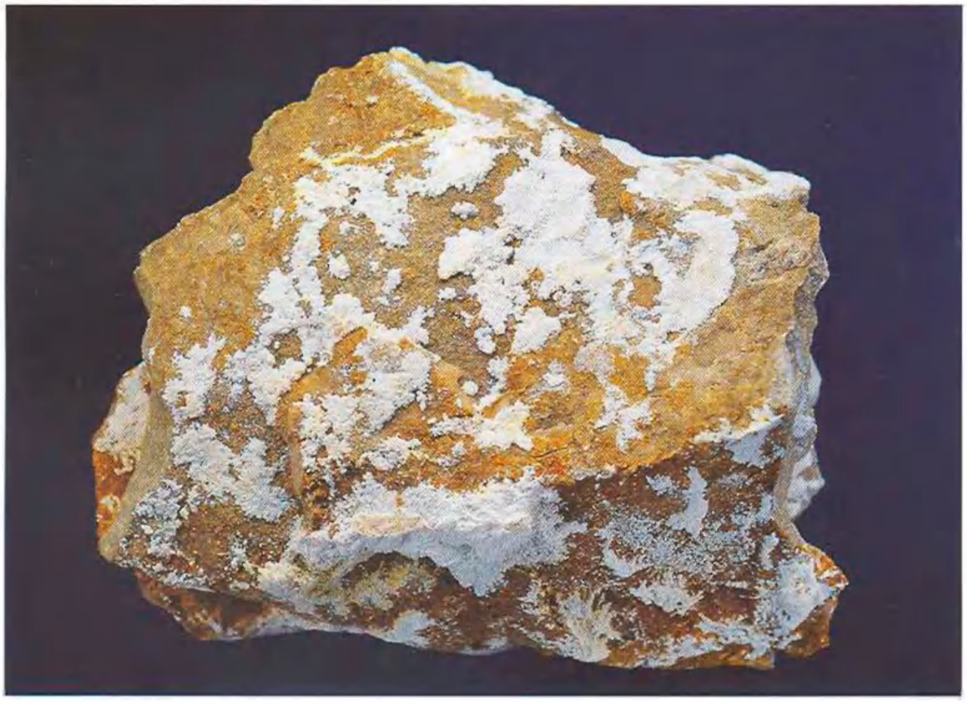

Alumohydrocalcite, 70 mm, Ladomirovo, Slovakia

6. Borates

Ludwigite

$Mg_2Fe^{3+}BO_5$

ORTHORHOMBIC ● ● ●

Properties: C – dark green, black-green, black; S – blue-green; L B silky to dull; D – opaque; DE – 3.9; H – 5; CL – perfect; F B uneven; M – prismatic crystals, fibrous aggregates, granular, massive.
Origin and occurrence: Metamorphic in skarns and dolomitic marbles, locally associated with magnetite and other borates. Rich aggregates are known from Ocna de Fier, Romania; Kamineichi, Japan and the Hol Kol mine, Suan, North Korea.
Application: chemical industry, B ore.

Gaudefroyite

$Ca_4Mn^{3+}_3(BO_3)_3CO_3(O,OH)_3$

HEXAGONAL ●

Properties: C – black; S – black; L – adamantine to dull; D – opaque; DE – 3.4; H – 6; CL – good; F – uneven; M – prismatic crystals, fibrous and acicular aggregates.
Origin and occurrence: Hydrothermal in calcite veins. Prismatic crystals up to 50 mm (2 in) long were found on the mine dumps near Tachgagalt, Morocco and in the N'Chwaning No. 2 mine and the Wessels mine, Kuruman, South Africa.

Gaudefroyite, 40 mm, Kuruman, South Africa
Ludwigite, 120 mm, Ocna de Fier, Romania

Inderite, 230 mm, Boron, U.S.A.

Inderite

$MgB_3O_3(OH)_5 . 5 H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white to pink in aggregates; S – white; L – vitreous; D – transparent to translucent; DE – 1.8; H – 2.5; CL – good; F – conchoidal to uneven; M – prismatic crystals, acicular and fibrous aggregates, nodules, massive.
Origin and occurrence: Sedimentary in boron deposits, commonly associated with colemanite and other borates. Prismatic crystals up to 100 mm (4 in) long occur in Boron, California, USA. It is also known from Inder, Kazakhstan.
Application: chemical industry, B ore.

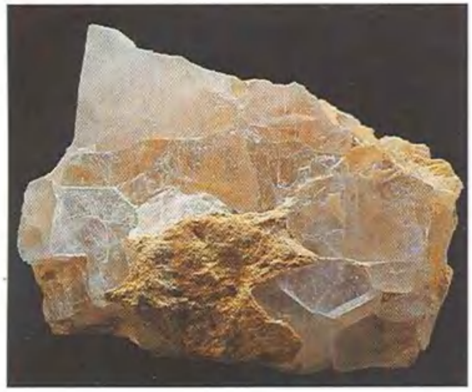
Inyoite, 90 mm, Turkey

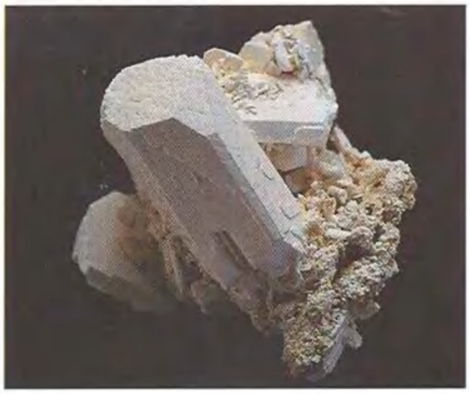
Borax, 60 mm x, Searles Lake, U.S.A.

Inyoite

$CaB_3O_3(OH)_5 . 4 H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white; S – white; L – vitreous; D – transparent to translucent; DE B 1.9; H – 2; CL – good; F – uneven; M – short prismatic to tabular crystals, columnar aggregates, massive.
Origin and occurrence: Sedimentary in boron deposits, commonly associated with colemanite and other borates. Clear tabular crystals up to 100 mm (4 in) long found in Kirka and Emet, Turkey. It also comes from Inder, Kazakhstan and the Corkscrew mine, California, USA.
Application: chemical industry, B ore.

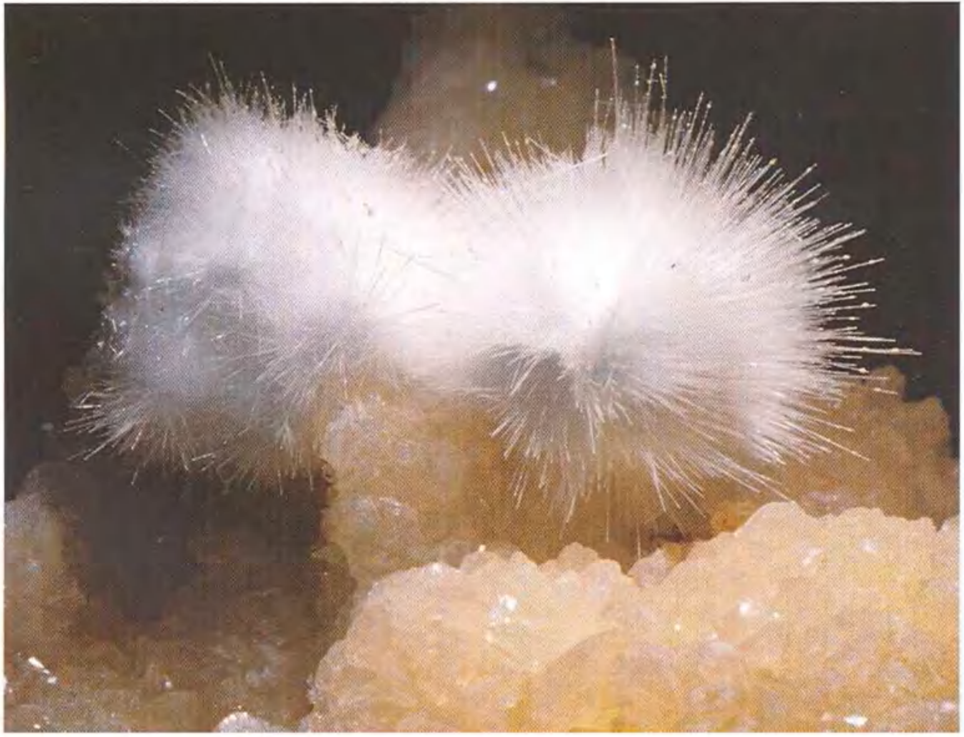
Ulexite, 20 mm aggregate, Death Valley, California, U.S.A.

Borax

$Na_2B_4O_5(OH)_4 . 8 H_2O$

MONOCLINIC ● ● ●

Properties: C – white, colorless, yellowish, gray, greenish; S – white; L B vitreous to dull; D – transparent, translucent to opaque; DE – 1.7; H – 2-2.5; CL – perfect; F – conchoidal; M – short prismatic to tabular crystals, columnar to earthy aggregates, crusts, coatings, granular, massive; R – soluble in water.
Origin and occurrence: Sedimentary in boron deposits, associated with other borates and halite. Prismatic crystals up to 150 mm long are known from Borax Lake, also Searles Lake and Boron, California, USA. It also occurs in Kirka, Turkey.
Application: chemical industry, B ore.

Ulexite

$NaCaB_5O_6(OH)_6 . 5 H_2O$

TRICLINIC ● ● ●

Properties: C – white, colorless, light gray; S – white; L B vitreous to silky; D – transparent to translucent; DE – 2.0; H – 2.5; CL – perfect; F B uneven; M – elongated prismatic to acicular crystals, fibrous aggregates, crusts, nodules, granular, massive.
Origin and occurrence: Sedimentary in boron deposits, associated with other borates. Slabs up to 100 mm (4 in) thick consisting of fibrous aggregates occur in Boron, California; also found in Esmeralda, Nevada, USA; Emet, Turkey and Inder, Kazakhstan.
Application: chemical industry, B ore.

Colemanite

$Ca_2B_6O_{11} . 5 H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, yellowish, light gray; S – white; L – vitreous; D – transparent to translucent; DE – 2.4; H – 4.5; CL – perfect; F B uneven; M – short prismatic and isometric crystals, granular, massive.
Origin and occurrence: Sedimentary in boron deposits, associated with other borates. Perfect crystals up to 200 mm (7⅞ in), come from Emet and Kirka, Turkey. Crystals up to 70 mm (2¾ in) long occur in Gower Gulch, Inyo, California; also known from Esmeralda, Nevada, USA and Inder, Kazakhstan.
Application: chemical industry. B ore.

Kernite, 100 mm, Boron, U.S.A.

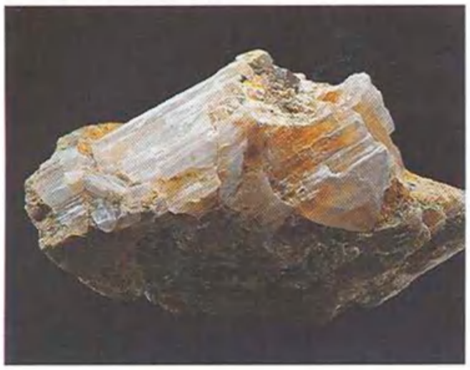

Kernite

$Na_2B_4O_6(OH)_2 . 3 H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, light gray; S – white; L – vitreous to dull, silky; D – transparent, translucent to opaque; DE – 1.9; H – 2.5-3; CL – perfect; F – uneven; M – isometric crystals, fibrous aggregates, granular, massive; R – soluble in water.

Rhodizite, 23 mm, Antsirabe, Madagascar

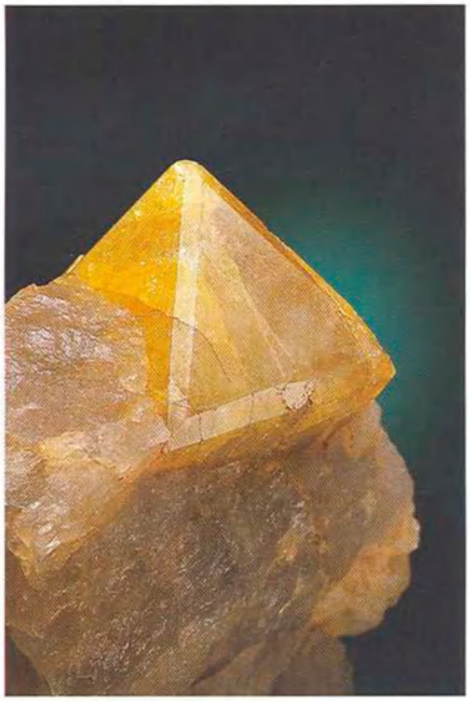

Hambergite, 64 mm, Pamir, Tadzhikistan

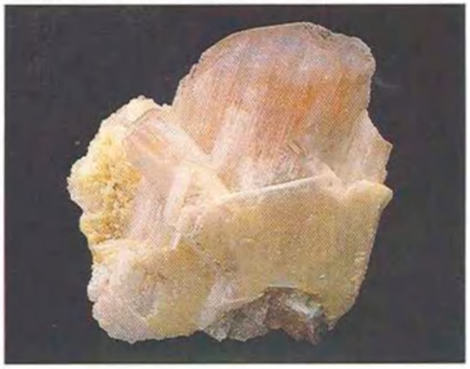

Origin and occurrence: Sedimentary in boron deposits, associated with borax and other borates. Platy aggregates and crystals up to 2.5 x 1 m (8 ft x $39\frac{3}{8}$ in) found in Boron, California, USA; also known from the Tincalayu mine, Salta province, Argentina.

Hambergite

$Be_2BO_3(OH,F)$

ORTHORHOMBIC ● ●

Properties: C – colorless, white, light gray; S – white; L – vitreous; D – transparent to translucent; DE – 2.4; H – 7.5; CL B perfect; F B uneven; M – tabular to prismatic crystals, granular.
Origin and occurrence: Magmatic in granitic and rarely also in alkaline pegmatites; hydrothermal in cavities within the pegmatites, associated with tourmaline, danburite and beryl. Tabular crystals up to 200 mm ($7\frac{7}{8}$ in) from the Little Three mine, Ramona, California, USA. Crystals up to 110 mm ($4\frac{5}{16}$ in) also known from several localities in Madagascar, e.g. Imalo and Anjanabonoina; also found in Hyakule, Nepal and Ctidruzice, Czech Republic .

Rhodizite

$(K,Cs)Al_4Be_4(B,Be)_{12}O_{28}$

CUBIC ● ●

Properties: C – colorless, white, yellow, light gray; S – white; L – vitreous to adamantine; D – transparent to translucent; DE – 3.5; H – 8.5; CL – imperfect; F – conchoidal to uneven; M – isometric crystals, granular.
Origin and occurrence: Magmatic in granitic pegmatites, associated with tourmaline, danburite and beryl. Cubic to tetrahedral crystals, reaching up to 30 mm ($1\frac{3}{16}$ in) occur in several localities in

Madagascar, e.g. Sahatany and Antandrokomby; also known from the Animikie Red Ace pegmatite, Florence, Wisconsin, USA.

Boracite

$Mg_3B_7O_{13}Cl$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish, light gray, light to dark green; S – white; L – vitreous; D – transparent to translucent; DE – 3.0; H – 7-7.5; CL – none; F – conchoidal to uneven; M – isometric crystals, fibrous aggregates, granular, massive.

Origin and occurrence: Sedimentary in evaporitic deposits, together with halite, gypsum and anhydrite; metamorphic in metamorphosed evaporates. Crystals up to 15 mm ($^{19}/_{32}$ in) come from Alto Chapare, Cochabamba, Bolivia; crystals about 5 mm ($^{3}/_{16}$ in) in size, occur in Stassfurt, Hanover and Kahlberg, Germany. Crystals are also known from Choctaw, Louisiana, USA and the Boulby mine, North Yorkshire, England, UK.

Boracite, 12 mm x, Alto Chapare, Bolivia

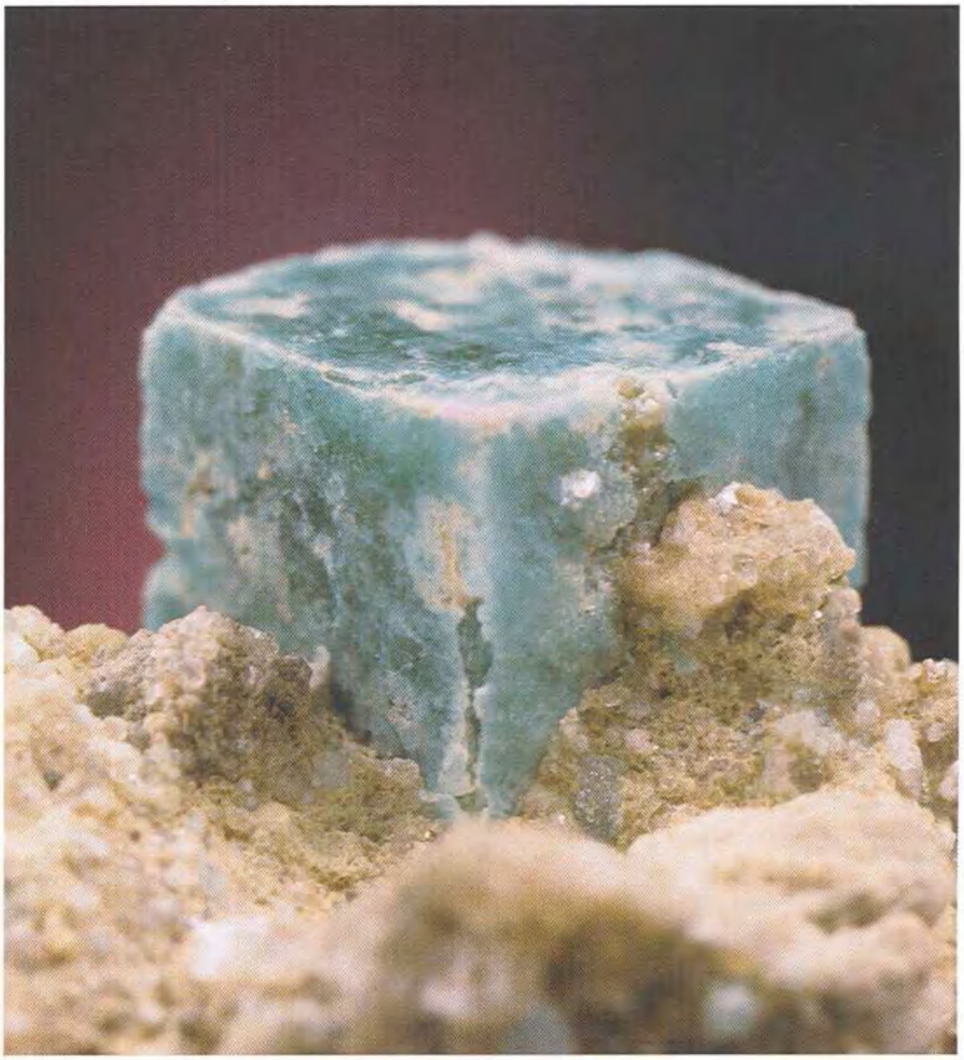

7. Sulfates

Anglesite

$PbSO_4$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish, gray, greenish; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 6.4; H – 2.5 – 3; CL – dobra; F – conchoidal; M – thick tabular crystals, massive, LU – sometimes yellowish.

Origin and occurrence: Secondary, as a result of the galena oxidation, together with cerussite and other minerals. Beautiful yellowish crystals up to 100 mm (4 in) in size come from Touissit, Morocco. Crystals from Tsumeb, Namibia, reached up to 40 mm ($1^{9}/_{16}$ in). Large crystals are also reported from Phoenixville, Pennsylvania, USA. Crystals up to 20 mm ($^{25}/_{32}$ in) occur in cavities in weathered galena in Sardinia, Italy. Crystals, reaching up to 40 mm ($1^{9}/_{16}$ in) also found in Mezica , Slovenia. Prismatic crystals up to 75 mm (3 in) long known from the Bunker Hill mine, Idaho, USA.

Barite, 17 mm xx, Pöhla, Germany

Anhydrite, 55 mm, Simplon Tunnel, Switzerland

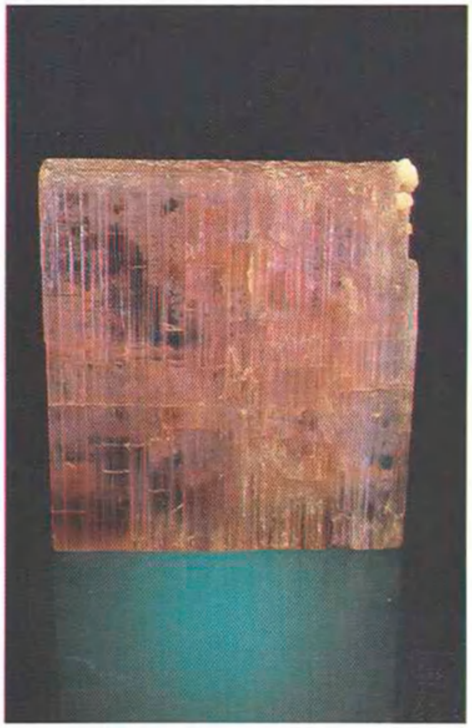

Anglesite, 81 mm, Touissit, Morocco

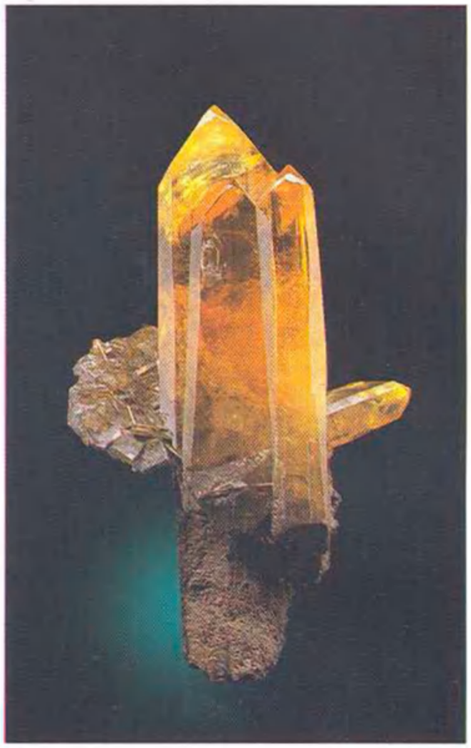

Anhydrite

$CaSO_4$

ORTHORHOMBIC ● ● ● ●

Properties: C – colorless, white, bluish, purplish, red, brown; S – white; L – pearly to vitreous; PS – transparent to translucent; DE – 2.8-3; H – 3-3.5; CL – good; F – splintery to uneven; M – isometric and prismatic crystals, granular, massive.

Origin and occurrence: Mostly sedimentary, as a result of the evaporation of sea water, associated with gypsum, calcite and other minerals; rare in pegmatites and hydrothermal. It is very common in form of massive aggregates in salt deposits in Stassfurt and Wathlingen, Germany, where small crystals also occur. Folded layers in clays known from Wieliczka, Poland. Purplish crystals up to 20 mm ($^{25}/_{32}$ in) found in cracks in metamorphic rocks near Simplon and St. Gotthard, Switzerland. Fine druses of bluish crystals up to 200 mm ($7^{7}/_{8}$ in) long found in Naica, Chihuahua, Mexico.

√+

Celestine, 270 mm, Madagascar

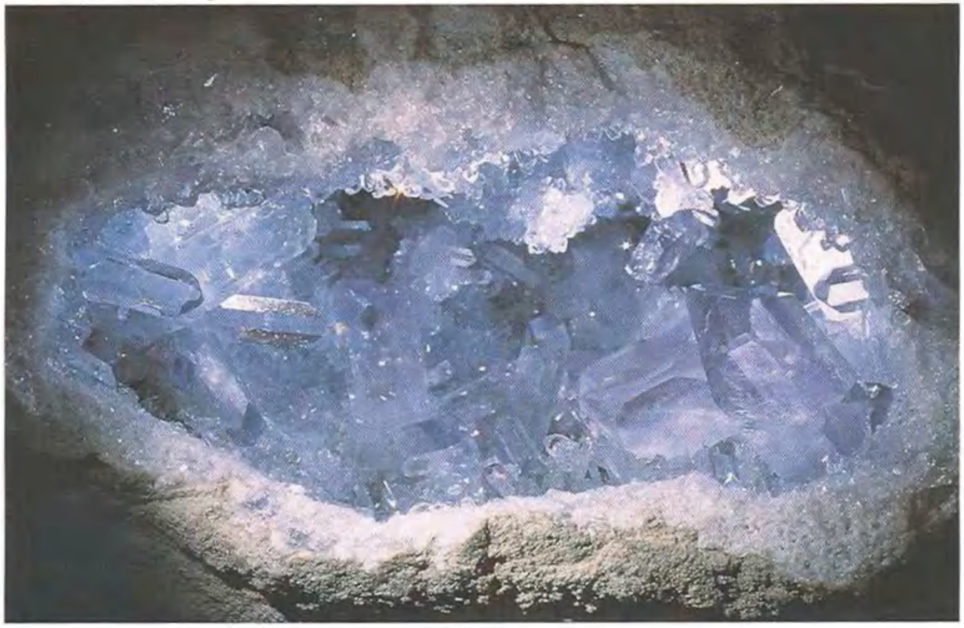

Celestite

$SrSO_4$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish, blue; S – white; L – vitreous; D – transparent to translucent; DE – 4; H – 3-3.5; CL – perfect; F – uneven; M – prismatic and tabular crystals, columnar aggregates, concretions.
Origin and occurrence: Rare hydrothermal, mainly sedimentary, together with halite, anhydrite and gypsum. Beautiful blue crystals several cm long occur in cavities of concretions in the Sakoany mine, Madagascar. Prismatic crystals up to 100 mm (4 in) long in sulfur deposits in Poland (Tarnobrzeg) and Italy (Caltanissetta). Fine blue tabular crystals associated with aragonite crystals found in _pania Dolina, Slovakia. Fine crystals are known from marls in Tunisia and Libya.

Barite

$BaSO_4$

ORTHORHOMBIC ● ● ● ●

Properties: C – white, yellow, blue, red, brown, black; S – white; L – vitreous; D – transparent to translucent; DE – 4.3-4.7; H – 3.5; CL – perfect; F – conchoidal to splintery; M – tabular to prismatic crystals, massive; LU – sometimes blue.
Origin and occurrence: Hydrothermal, it originates under medium and low temperatures, also sedimentary, together with fluorite, calcite, cinnabar and other minerals. Beautiful barite crystals up to 50 mm (2 in) found in hydrothermal veins in Pribram, Czech Republic. Blue tabular crystals up to 100 mm (4 in) from Dedova hora, Czech Republic. Beautiful tabular crystals up to 100 mm (4 in) known from Banska Stiavnica, Slovakia; Cavnic and Baia Spric, Romania. Very fine crystals of up to 200 mm (7⅞ in) occurred in Alston Moor, Frizington and Mowbray, Cumbria, UK. Beautiful druses of honey-brown crystals up to 70 mm (2¾ in) long found in Pohla, Germany. Fine druses of prismatic crystals known from Elk Creek, South Dakota, USA; so called

Antlerite, 60 mm, Chuquicamata, Chile

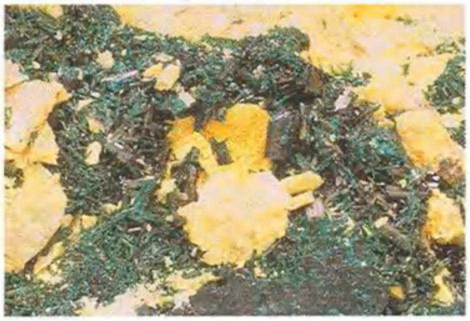

Barite, 100 mm xx, Stoneham Co., U.S.A.

desert roses (crystals with inclusion of sand grains) originating in desert climate occur in the vicinity of Norman, Oklahoma, USA. Shiny crystals come from Freiberg and Halsbrucke, Germany.
Application: an ingredient of drilling fluids, chemical, glass, paper and rubber industries.

Antlerite

$Cu_5(SO_4)_2(OH)_6 \cdot 3\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – green; S – light green; L – vitreous; D – translucent; DE – 3.9; H – 3.5; CL – perfect; F – conchoidal to uneven; M – short prismatic and acicular crystals, crusts and earthy aggregates.
Origin and occurrence: Secondary, resulting from weathering of Cu ores, associated with other secondary Cu minerals. It is a principal mineral of the oxidation zone in Chuquicamata, Chile, where it forms crystals up to 5 mm (3/16 in). Acicular crystals, up to 20 mm (25/32 in) long, occurred in Bisbee, Arizona, USA. Crystalline crusts are known from Spania Dolina and Piesky, Slovakia.

Brochantite, 14 mm aggregate, Bingham, U.S.A.

Brochantite

$Cu_4(SO_4)(OH)_6$

MONOCLINIC ● ● ●

Properties: C – green; S – light green; L – vitreous, in cleavage planes pearly; D – translucent to transparent; DE – 4; H – 3.5-4; CL – good; F – conchoidal to uneven; M – long prismatic to acicular crystals, granular.
Origin and occurrence: Secondary, as a result of the oxidation of Cu ores usually in the arid climate, together with other secondary Cu minerals. Crystals up to 70 mm (2¾ in) long found in Bisbee, together with pseudo-morphs of malachite after brochantite. Crystals are also known from Tsumeb, Namibia. Prismatic crystals up to 50 mm (2 in) long from Cerro Verde, Peru. Emerald-green crusts occur in L'ubietova, Slovakia.

Linarite, 23 mm xx, Graham Co., U.S.A.

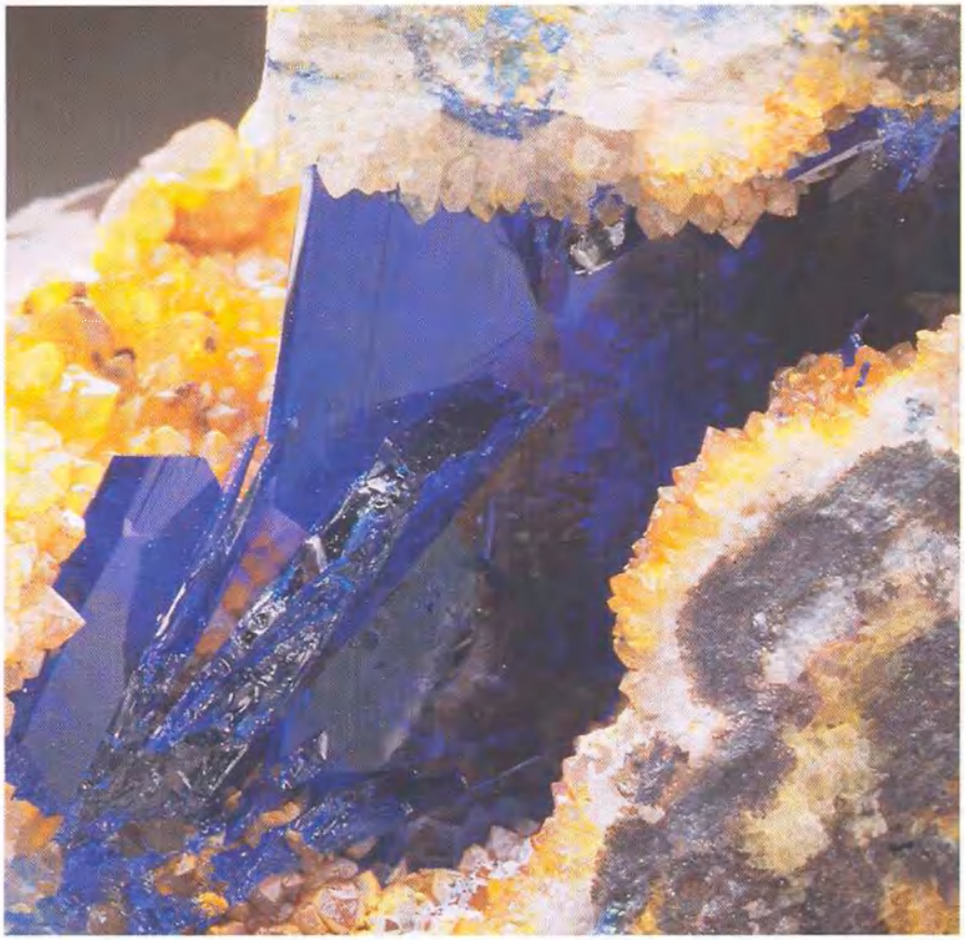

Linarite

$PbCu(SO_4)(OH)_2$

MONOCLINIC ● ●

Properties: C – azure-blue; S – light blue; L – vitreous to adamantine; D – translucent; DE – 5.3-5.5; H – 2.5; CL – good; F – conchoidal; M – tabular and prismatic crystals, coatings.

Origin and occurrence: Secondary, as a result of the oxidation of Cu and Pb sulfides at low pH. The largest crystals, reaching up to 80 mm ($3^{1}/_{8}$ in), come from the Mammoth mine, Tiger, Arizona, USA. Crystals from the Grand Reef mine in Arizona up to 50 mm (2 in) long. Very fine crystals, up to 30 mm ($1^{3}/_{16}$ in) were found in Keswick, Cumbria, England, UK. Large crystals were also described from Tsumeb, Namibia and from Broken Hill, New South Wales, Australia.

Alunite

$KAl_3(SO_4)_2(OH)_6$

TRIGONAL ● ● ●

Properties: C – white, yellowish, gray; S – white; L – vitreous to pearly; D – transparent to opaque; DE – 2.8; H – 3.5-4; CL – perfect; F – conchoidal; M – rhombohedral crystals, porous aggregates, granular.

Origin and occurrence: Secondary, as a result of reactions of sulfuric acid with Al-rich rocks, associated with gypsum. Small crystals are known in Berehovo, Ukraine. Massive aggregates occur together with turquoise in Cu deposits in Arizona, USA. (New Cornelia mine and others). Huge alunite deposits, like Zaglik, Azerbaijan, are mined for Al.

Application: Al ore.

Alunite, 1 mm xx, Nagyegyháza, Hungary

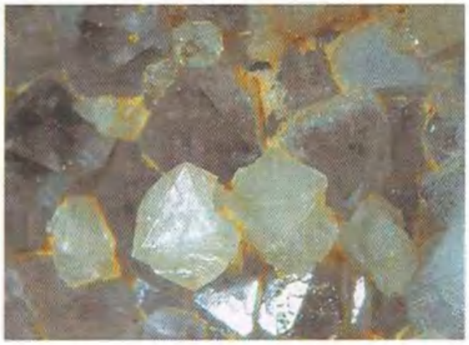

Natrojarosite, 50 mm, Laurion, Greece

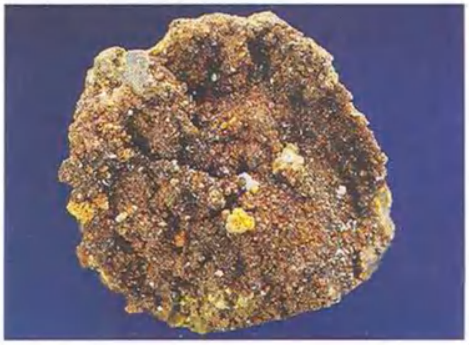

Natrojarosite

$NaFe^{3+}_3(SO_4)_2(OH)_6$

TRIGONAL ● ●

Properties: C – yellow, brown; S – light yellow; L – vitreous; D – transparent to translucent; DE – 3.2; H – 3; CL – good; F – conchoidal; M – tabular to rhombohedral crystals, earthy aggregates..
Origin and occurrence: Secondary, as a result of weathering Fe sulfides, associated with fibroferrite, alunite and other minerals. It occurs in Modum, Norway; Soda Springs Valley, Nevada, USA; Chuquicamata, Chile and elsewhere.

Jarosite

$KFe^{3+}_3(SO_4)_2(OH)_6$

TRIGONAL ● ● ●

Properties: C – yellow; S – light yellow; L – adamantine to dull; D – translucent; DE – 2.9-3.3; H – 2.5-3.5; CL – good; F – conchoidal to uneven; M – rhombohedral to tabular crystals, granular, massive.
Origin and occurrence: : Secondary, as a result of weathering Fe sulfides, usually associated with natrojarosite and other sulfates. The world's best specimens come from Pena Blanca Uranium mine, near Aldama, Chihuahua, Mexico, where crystals, up to 20 mm ($^{25}/_{32}$ in) were found. Crystals, up to 10 mm ($^{3}/_{8}$ in), were found in Tombstone, Arizona, USA. Crystals, together with pseudo-morphs of jarosite after alunite, occurred in Chuquicamata, Chile. Tabular crystals also were described from Horni Slavkov, Czech Republic.

Beudantite

$PbFe^{3+}_3(AsO_4)(SO_4)(OH)_6$

TRIGONAL ● ●

Properties: C – dark green, brown, red-brown; S – greenish, gray-yellow; L – vitreous to resinous; D – transparent to translucent; DE – 4; H – 3.5-4.5; CL – good; M – rhombohedral crystals, crusts.
Origin and occurrence: Secondary, occurring in the oxidation zone of Pb deposits, together with scorodite and other minerals. Crystals, several mm in size, found in Tsumeb, Namibia, are the best in the world. Small crystals are known from Bisbee, Arizona, USA; Ashburton Downs, Western Australia, Australia and Kamarecza near Laurion, Greece.

Jarosite, 2 mm xx, Gyongyösoroszi, Hungary

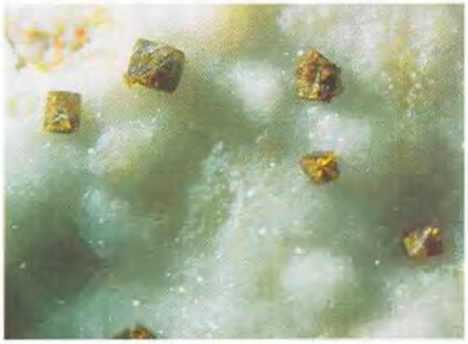

Beudantite, 70 mm, Moldava, Czech Republic

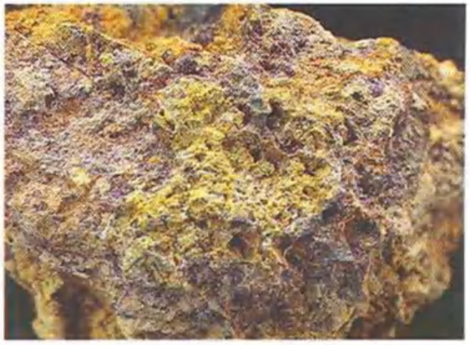

Hanksite

$Na_{22}K(SO_4)_9(CO_3)_2Cl$

HEXAGONAL ● ●

Properties: : C – colorless, yellowish; S – colorless; L – vitreous to dull; D – transparent to translucent; DE – 2.6; H – 3.5; CL – good; F – uneven; M – tabular to short prismatic crystals.
Origin and occurrence: Sedimentary in salt lake sediments, together with halite, borax and other minerals. Crystals up to 200 mm ($7^{7}/_{8}$ in) found in Searles Lake, California, USA. It is also known from borax deposits in the Death Valley, California, USA.

Caledonite

$Cu_2Pb_5(SO_4)_3(CO_3)(OH)_6$

ORTHORHOMBIC ● ●

Properties: C – green to blue-green; S – light green; L – resinous; D – transparent to translucent; DE – 5.6; H – 2.5-3; CL – perfect; F – uneven; M – prismatic crystals, radial aggregates.

Origin and occurrence: Secondary, occurring in the oxidation zone of Cu and Pb deposits, associated with linarite and other minerals. Crystals up to 20 mm ($^{25}/_{32}$ in) come from Leadhills, Scotland, UK. Crystals, up to 15 mm ($^{19}/_{32}$ in), were found in the Mammoth mine, Tiger, Arizona, USA. Rich druses are known from Anarak, Iran. Crystals reaching up to 20 mm ($^{25}/_{32}$ in) occurred in the Blue Bell mine, California, USA.

Leadhillite

$Pb_4(SO_4)(CO_3)_2(OH)_2$

MONOCLINIC ● ●

Properties: C – colorless, yellowish, gray; S – colorless; L – resinous to adamantine; D – transparent to translucent; DE – 6.5; H – 2.5-3; CL – perfect; F – conchoidal; M – pseudo-hexagonal tabular crystals.
Origin and occurrence: Secondary mineral from the oxidation zone of Pb deposits, associated with cerussite, anglesite, linarite and other minerals. The largest crystals, measuring up to 15 cm, come from Tsumeb, Namibia. Crystals up to 25 mm (1 in) found in the Mammoth mine, Tiger, Arizona, USA. Crystals in Leadhills, Scotland, UK were of similar size.

Chalcantite

$CuSO_4 . 5 H_2O$

TRICLINIC ● ● ●

Properties: C – deep blue; S – white; L – vitreous to resinous; D – transparent to translucent; DE – 2.3; H – 2.5; CL – imperfect; F – conchoidal; M – short prismatic to tabular crystals, stalactites, films; R – soluble in water.

Hanksite, 50 mm, Searles Lake, U.S.A.

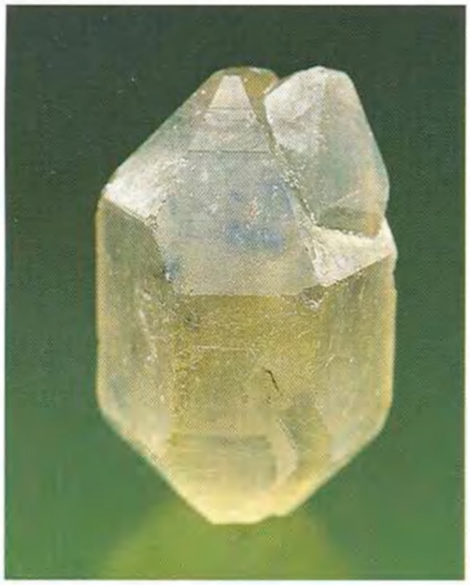

Caledonite, 5 mm x, Graham Co., U.S.A.

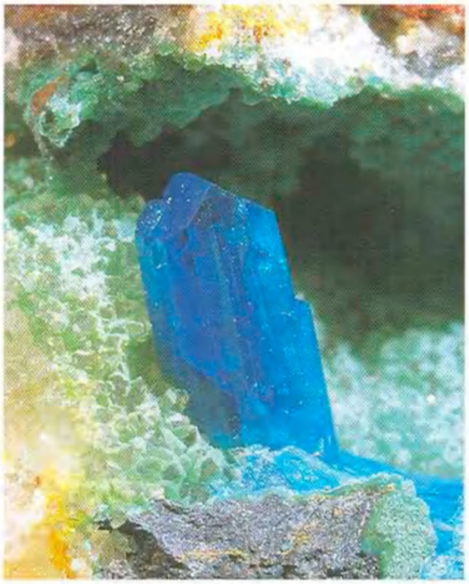

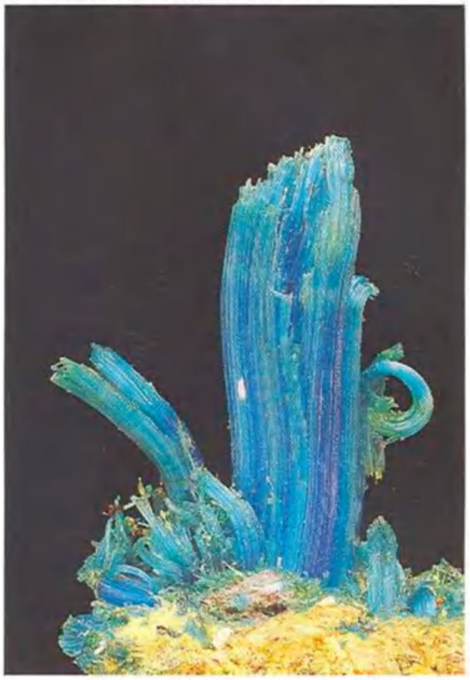

Origin and occurrence: Secondary, as a product of the oxidation of Cu sulfides, stable only in arid climate. Crystals up to 40 mm ($1^{9}/_{16}$ in) in size, and stalactites, up to 1 m ($39^{3}/_{8}$ in) long are known from Bisbee, Arizona, USA. Crystals were also found in the oxidation zone in Chuquicamata, Chile. Beautiful stalactites come from Rio Tinto, Spain.
Application: rarely mined as Cu ore.

Melanterite, 80 mm, Harz Mts., Germany

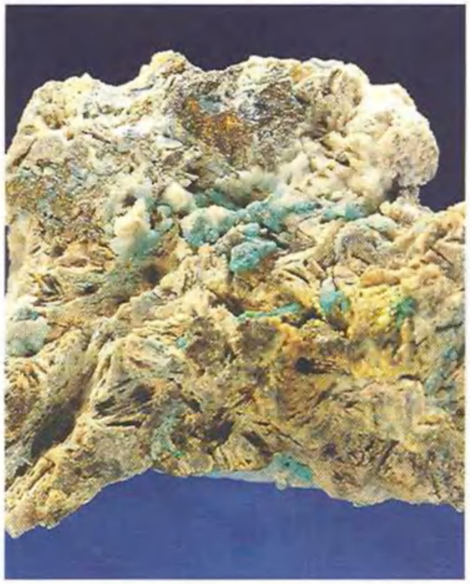

Melanterite

$FeSO_4 . 7 H_2O$

MONOCLINIC ● ● ● ●

Varieties: pisanite (with Cu contents), kirovite (with Mg contents)

Properties: C – light green; S – colorless; L – vitreous; D – translucent; DE – 1.9; H – 2; CL – perfect; F – conchoidal; M – granular crusts, botryoidal and stalactitic aggregates, films; R – soluble in water.
Origin and occurrence: Secondary, as a result of the oxidation of Fe sulfides, together with other sulfates. It is unstable under atmospheric conditions. Crystals up to 20 mm ($^{25}/_{32}$ in) long are known from Bisbee, Arizona and from the Boyd mine, Tennessee, USA. Stalactites, up to 50 cm (20 in) long, occurred in Chvaletice, Czech Republic. It is also common in Rio Tinto, Spain and Banska Stiavnica, Slovakia.

Epsomite
$MgSO_4 . 7 H_2O$

ORTHORHOMBIC ● ● ● ●

Properties: C – colorless to white; S – colorless; L – vitreous to silky; D – transparent to translucent; DE – 1.7; H – 2; CL – perfect; F – conchoidal; M – small crystals, granular, stalactitic aggregates; R – soluble in water, decomposing under atmospheric conditions.
Origin and occurrence: Secondary, as a result of the oxidation of Fe sulfides, also from crystallization of the salt lake water, associated with halite and other minerals. Crystals up to 1 m (39⅜ in) long were found on Mount Kruger, Washington; prismatic crystals up to 50 mm (2 in) long also occur in Bisbee, Arizona and needles of a similar size come from the White Caps mine, Nevada, USA. It is also known as a product of the activity of fumaroles in Mount Vesuvius, Italy.

Morenosite
$NiSO_4 . 7 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – green, green-white; S – greenish; L – vitreous; D – transparent to translucent; DE – 2; H – 2-2.5; CL – good; F – conchoidal; M – stalactitic crusts and efflorescences; R – soluble in water.
Origin and occurrence: Secondary, as a product of the oxidation of Ni minerals, stable mainly in the arid regions. It is very common in Sudbury, Ontario, Canada; also in Richelsdorf, Germany and Potucky, Czech Republic.

Epsomite, 10 mm xx, Stassfurt, Germany

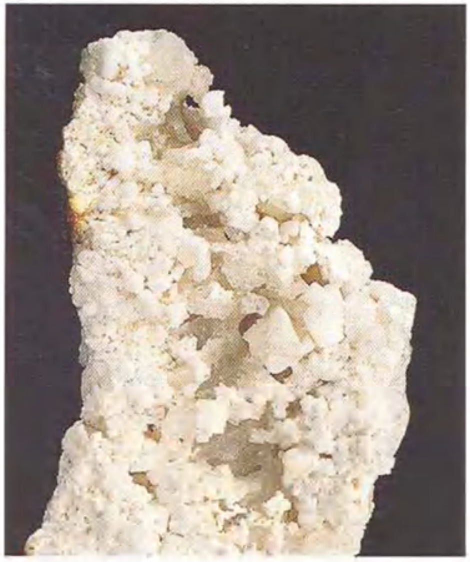

Coquimbite
$Fe^{3+}_2(SO_4)_3 . 9 H_2O$

TRIGONAL ● ● ●

Properties: C – purple, yellow, green, colorless; S – colorless; L – vitreous; D – transparent; DE – 2.1; H – 2; CL – imperfect; F – conchoidal to uneven; M – prismatic to tabular crystals, granular, massive; R – soluble in water.
Origin and occurrence: Secondary, associated with other sulfates. Crystals, several cm long, were found in the Dexter No.7 mine, Calf Mesa, Utah, USA. Small prismatic crystals are known from Zelezník , tabular crystals occurred in Banská Stiavnica, Slovakia. It is very common in Chilean deposits, like Chuquicamata and Tierra Amarilla.

Alunogen
$Al_2(SO_4)_3 . 17 H_2O$

TRICLINIC ● ● ●

Properties: C – colorless, white, gray-yellow; S – colorless; L – silky; PS – transparent; DE – 1.8; H – 1.5; CL – perfect; M – pseudo-hexagonal crystals, granular; R – soluble in water.
Origin and occurrence: Secondary, as a result of the pyrite oxidation, also a product of sublimation on volcanoes and burning dumps, associated with other sulfates. Large crystals are known from the Dexter No.7 mine, Calf Mesa, Utah, crusts, over 1 m (39⅜ in) thick were found on Mount Alum, New Mexico,

Morenosite, 60 mm, Potucky, Czech Republic

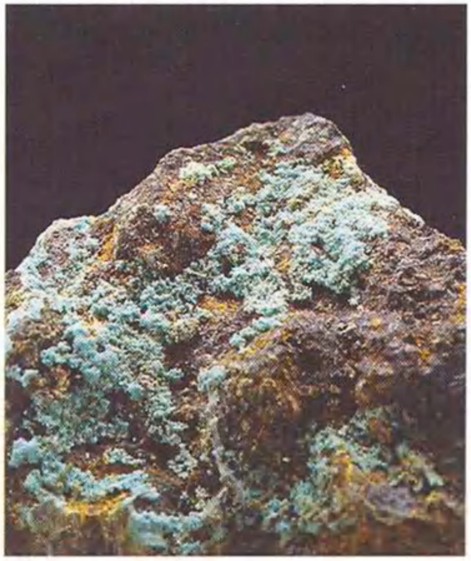

Coquimbite, 110 mm, Cerritos Bajos, Chile

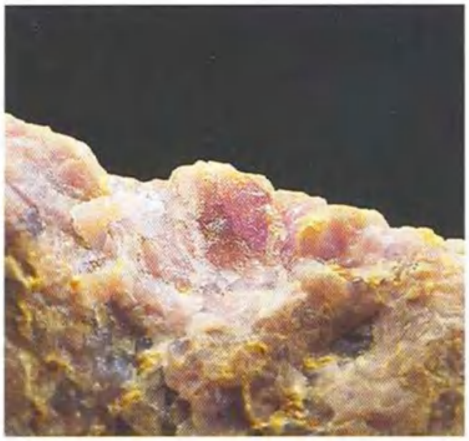

USA. Fine aggregates occur on the walls of underground workings in the old opal mines in Dubník, Slovakia.

Halotrichite

$FeAl_2(SO_4)_4 . 22 H_2O$

MONOCLINIC ● ● ●

Properties: C – white, greenish; S – colorless; L – vitreous; PS – transparent to translucent; DE – 1.9-2.1; H – 1.5-2; CL – imperfect; F – conchoidal; M – acicular crystals, fibrous aggregates, efflorescences; R – soluble in water.

Origin and occurrence: Secondary, as a result of the pyrite oxidation, also a product of the activity of hot springs ans solfataras, associated with other sulfates. Common fibrous crusts occur in Dubnik, Slovakia. It looks similar in Rio Marina, Elba, Italy and in Chuquicamata, Chile. It is known from solfataras in Pozzuoli, Italy. It is a product of hot springs activity in the Lassen Peak National Park, California, USA.

Apjohnite

$MnAl_2(SO_4)_4 . 22 H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, pink, greenish, yellow; S – colorless; L – silky; D – transparent; DE – 1.9; H – 1.5; M – fibrous aggregates, coatings, massive; R – soluble in water.

Origin and occurrence: Secondary, occurring together with other sulfates. Large accumulations are known from Little Pigeon Creek, Alum Cave, Tennessee, USA. It also occurs in Delagona Bay, Mozambique.

Alunogen, 80 mm, Dubnik, Slovakia

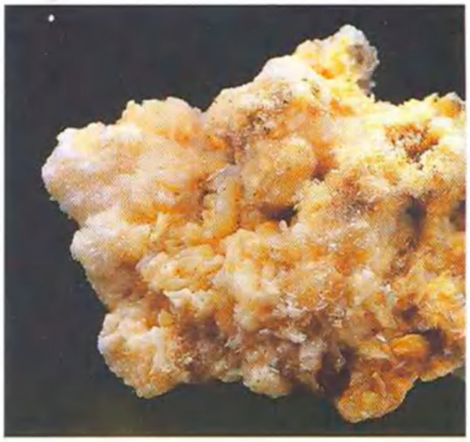

Halotrichite, 90 mm, Recsk, Hungary

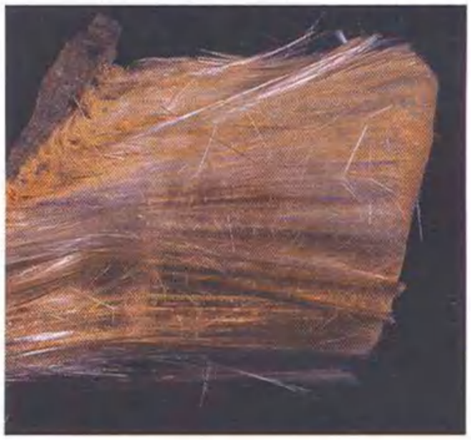

Apjohnite, 60 mm, Smolnik, Slovakia

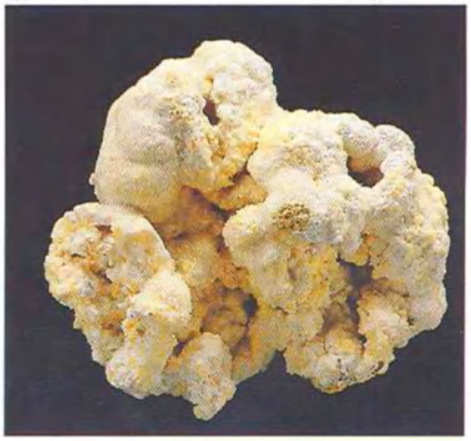

Bilinite, 60 mm, Bilina, Czech Republic

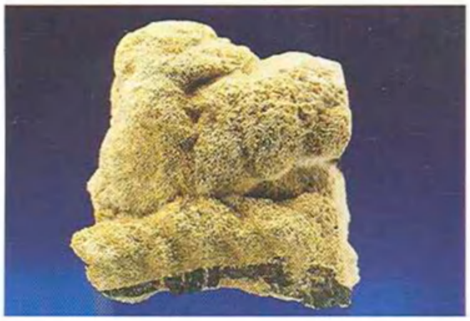

Bilinite

$Fe^{2+}Fe^{3+}{}_2(SO_4)_4 \cdot 22\ H_2O$

MONOCLINIC ● ●

Properties: C – white to wellowish; S – white; L – silky; D – opaque; DE – 1.9; H – 2; M – fibrous aggregates; R – soluble in water.
Origin and occurrence: Secondary, as a result of the pyrite oxidation, associated with other sulfates. Fibrous aggregates are known from Svitec near Bilina, Czech Republic. Also described from Bisbee, Arizona, USA.

Tschermigite

$(NH_4)Al(SO_4)_2 \cdot 12\ H_2O$

CUBIC ● ● ●

Properties: : C – colorless; S – colorless; L – vitreous; D – transparent; DE – 1.6; H – 1.5; F – conchoidal; M – octahedral crystals, fibrous and columnar aggregates; R – soluble in water. *Origin and occurrence:* Secondary in the brown coal, also on burning dumps and as a product of solfataras. Skeletal crystals up to 10 mm (3/8 in) in size are known from the brown coal basin near Most and Bilina, Czech Republic. Original locality Ecrmniky, Czech Republic yielded fibrous crusts, up to 50 mm (2 in) thick. Fine octahedra, up to 10 mm (3/8 in) come from burning dumps near Zastavka near Brno, Czech Republic. Octahedra occur also in Geysers, California, USA. Small crystals on Etna, Sicily, Italy are of volcanic origin.

Tschermigite, 50 mm, Mogyororosbanya, Hungary

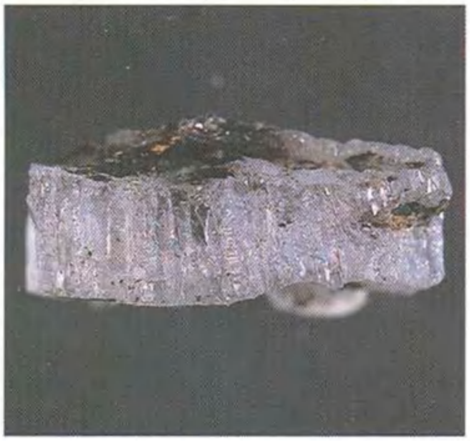

Polyhalite

$KCa_2Mg(SO_4)_4 \cdot 2\ H_2O$

TRICLINIC ● ● ●

Properties: C – colorless, brown, red-brown; S – colorless; L – vitreous; D – transparent to translucent; DE – 2.8; H – 3-3.5; CL – perfect; M – acicular to prismatic crystals, columnar, scaly and fibrous aggregates.
Origin and occurrence: Sedimentary, as a constituent of salt deposits, also as a product of volcanic activity. Tabular and prismatic crystals are found rarely together with massive polyhalite aggregates in Stassfurt, Germany. Fibrous aggregates come from Halstatt, Austria. Large deposits of polyhalite are located near Carlsbad, New Mexico, USA. Coatings occur on Mount Vesuvius, Italy.

Görgeyite

$K_2Ca_5(SO_4)_6 \cdot H_2O$

MONOCLINIC ● ●

Properties: C – colorless, yellowish; S – colorless; L – vitreous; D – translucent; DE – 3; H – 3.5;

Polyhalite, 54 mm, Carlsbad, U.S.A.

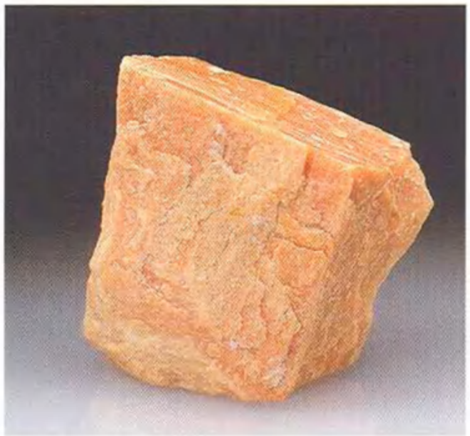

Gypsum, 62 mm, Carneville, Utah, U.S.A.

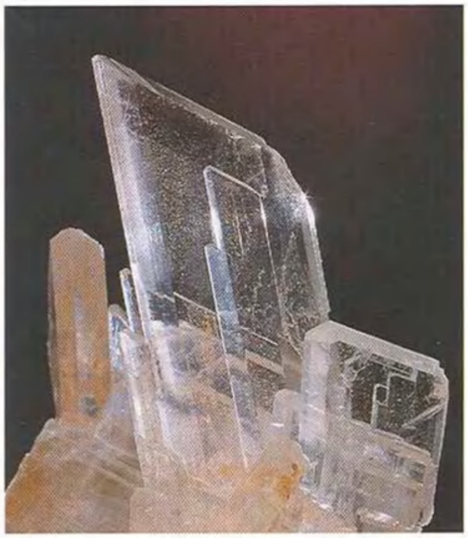

CL – good; F – conchoidal; M – tabular crystals. *Origin and occurrence:* Sedimentary in salt deposits, associated with halite and other minerals. Fine crystals up to 80 mm (3⅛ in) long found near Inder Lake, Kazakhstan; also known from Bad Ischl, Austria.

Gypsum

$CaSO_4 . 2 H_2O$

MONOCLINIC ● ● ● ● ●

Varieties: Maria-glass, alabaster, satin spar, selenite. *Properties:* C – colorless, white, gray, yellowish; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 2.3; H – 1.5-2; CL – perfect; F – conchoidal; M – typical monoclinic crystals, often twins, fibrous and platy aggregates, granular, massive; LU – crystals with inclusions sometimes bluish to yellowish.

Origin and occurrence: Rare primary hydrothermal, mostly sedimentary and as a weathering product, associated with anhydrite, halite and other minerals. Crystals up to 50 mm from Cavnic, Romania are probably hydrothermal. Huge crystals up to 1.5 m (5 ft) long found in karst cavities (Cave of Swords) in gossan in the PbBZn deposit Naica, Chihuahua, Mexico. Crystals up to 9 m (29 ft 6 in) long from Santa Eulalia, Chihuahua, Mexico, where they were found combined with interesting aggregates, called 'ram's horns' from their shape. Crystals up to 4 m (13 ft) long occurred in Tarnobrzeg, Poland. Common crystals are also known from Gorguel, Spain. The 'desert roses' or crystal druses, from Sahara desert in Tunisia and Algeria, with sand grain inclusions are mineralogically interesting. Large slabs of the transparent variety, called Maria-glass, found in Friedrichsrode, Germany. Fine-grained variety alabaster occurs e.g. in Italy. Fibrous variety selenite (sometimes also called satin spar) comes from the Sylva river basin, Perm, Russia.
Application: building, chemical and medical industries.

Gorgevite, 80 mm, Inder Lake, Kazakhstan

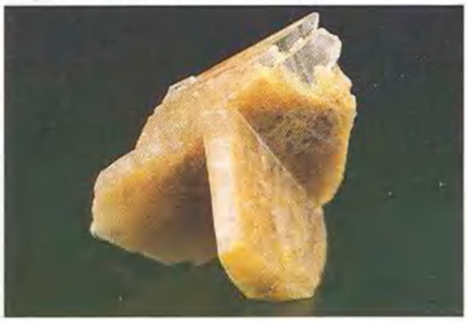

Langite

$Cu_4(SO_4)(OH)_6 . 2 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – blue, blue-green; S – light blue; L – vitreous to silky; D – translucent; DE – 3.3; H – 2.5-3; CL – perfect; M – small isometric crystals, fine-grained crusts, earthy aggregates..
Origin and occurrence: Secondary, as a product of the oxidation of Cu ores, associated with gypsum and other minerals. Crystals known from St. Just, Cornwall, UK. Crystals occurred in Tsumeb, Namibia. Fine specimens come from Spania Dolina and L'ubietova, Slovakia and from Borovec, Czech Republic. It is also common in Ely, Nevada, USA and in El Cobre, Chile.

Langite, 49 mm, Allihies, Cork, Ireland

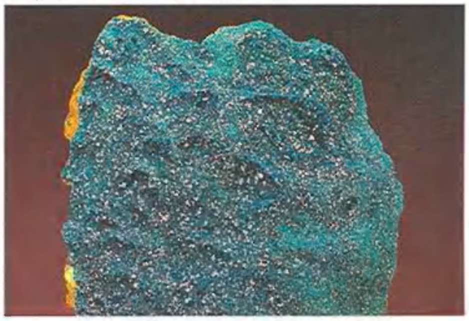

Fibroferrite, 70 mm, Dubník, Slovakia

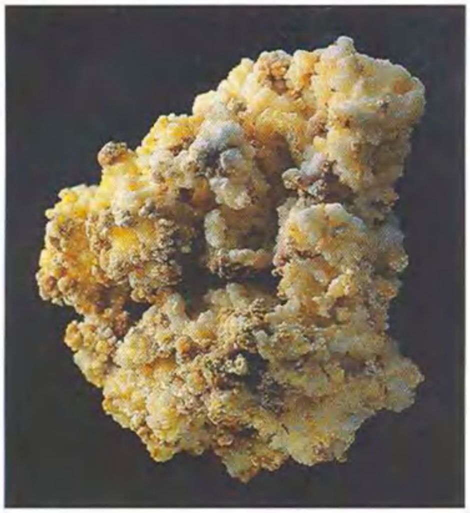

Fibroferrite

$Fe^{3+} (SO_4)(OH) . 5 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – yellowish, greenish, gray; S – white; L – silky; D – opaque; DE – 1.9; H – 2.5; CL – perfect; M – small crystals, fibrous and botryoidal crusts; R – soluble in water.

Origin and occurrence: Secondary as a product of the pyrite oxidation, together with other sulfates. Crystals in cavities in melanterite found in the Dexter No.7 mine, Calf Mesa, Utah, USA. Vein fillings up to 3 m (10 ft) thick come from the Santa Elena mine, La Alcaparrosa, Argentina. Fine fibrous aggregates occur in Dubník, Slovakia. It is common in many localities in Chile (Tierra Amarilla, Chuquicamata).

Aluminite

$Al_2(SO_4)(OH)_4 . 7 H_2O$

MONOCLINIC ● ● ●

Properties: C – white; S – white; L – dull; D – opaque; DE – 1.7-1.8; H – 1-2; CL – none; M – earthy nodules, consisting of microscopic needles.

Origin and occurrence: Secondary, resulting from the reaction of sulfuric acid with Al in Al-rich rocks. Large nodule, 30 cm (12 in) in diameter, known from Newhaven, Sussex, UK. Fine white nodules described from Malá Chuchle, Prague, Czech Republic. It also occurs near Halle, Germany and covers limestone in the vicinity of Joplin, Missouri, USA.

Aluminite, 60 mm, Bicske-Csordakut, Hungary

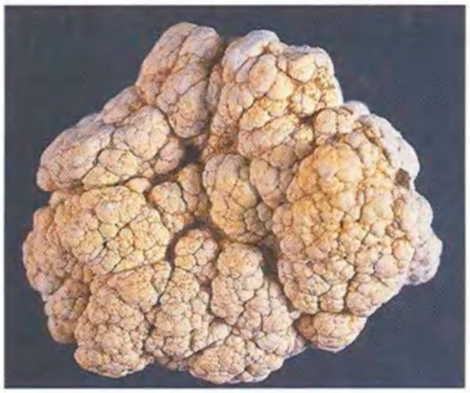

Botryogen

$MgFe^{3+} (SO_4)_2(OH) . 7 H_2O$

MONOCLINIC ● ● ●

Properties: C – orange-red; S – yellow; L – vitreous; D – transparent to translucent; DE – 2.1; H – 2-2.5; CL – perfect; F – conchoidal; M – prismatic striated crystals, racemose and spherical aggregates with radial structure.

Origin and occurrence: Secondary, as a result of the pyrite oxidation in arid regions, together with other sulfates. Crystals up to 35 mm ($1^{5}/_{16}$ in) long from the Libiola mine near Genoa, Italy. It is common in Chuquicamata and Quetena in Chile; also known from Rammelsberg, Germany.

Botryogen, 3 mm xx, Knoxville, U.S.A.

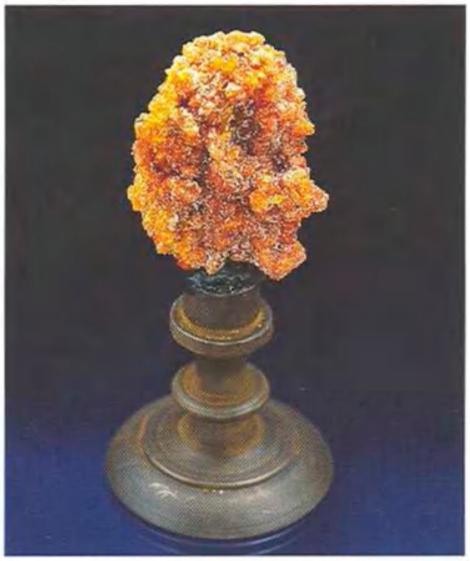

Devilline, 10 mm xx, Spania Dolina, Slovakia

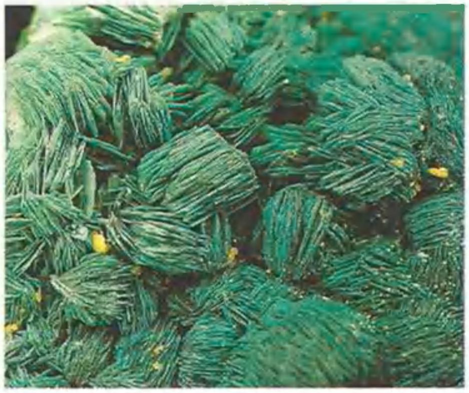

Copiapite

$(Fe^{2+},Mg)Fe^{3+}_4(SO_4)_6(OH)_2 \cdot 20\ H_2O$

TRICLINIC ● ● ●

Properties: C – yellow, green-yellow, orange; S – yellow; L – pearly; D – transparent to translucent; DE – 2.1-2.2; H – 2.5; CL – perfect; M – thin tabular crystals, scaly and pulverulent aggregates, earthy.
Origin and occurrence: Secondary, originated from the pyrite oxidation, together with other sulfates.
Tabular crystals come from the Dexter No.7 mine, Calf Mesa, Utah, USA and also from Zeleznik , Slovakia. Fine crystals were found in many localities in Chile (Tierra Amarilla, Chuquicamata). It also occurs in Rammelsberg, Germany.

Copiapite, 96 mm, Copiapó, Chile

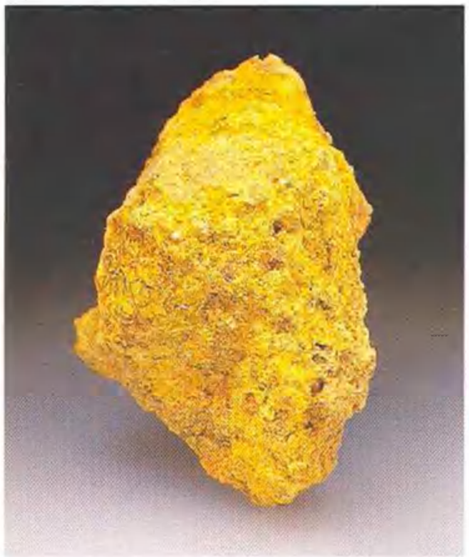

Serpierite, 80 mm, Příbram, Czech Republic

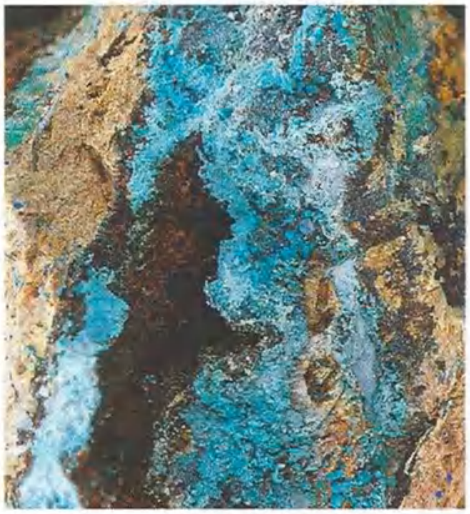

Devilline

$CaCu_4(SO_4)_2(OH)_6 \cdot 3\ H_2O$

MONOCLINIC ● ●

Properties: C – emerald-green; S – white to light green; L – vitreous to pearly; D – transparent to translucent; DE – 3.1; H – 2.5; CL – perfect; M – thin tabular pseudo-hexagonal crystals and coatings.
Origin and occurrence: Secondary, as a result of the oxidation of Cu sulfides, associated with other secondary Cu minerals. The world's best specimens come from Spania Dolina, Slovakia, where crystal rosettes up to 10 mm (⅜ in) in diameter were found in the past century; also known from Botallack, Cornwall, UK and Tsumeb, Namibia.

Serpierite

$Ca(Cu,Zn)_4(SO_4)_2(OH)_6 \cdot 3\ H_2O$

MONOCLINIC ● ●

Properties: C – blue; S – white; L – vitreous; D – transparent to translucent; DE – 3.1; H – 2.5; CL – perfect; M – tabular crystals, coatings.
Origin and occurrence: Secondary, originated in the oxidation zone of Cu-Zn deposits, together with smithsonite and other minerals. Its small crystals and aggregates occurred in Laurion, Greece. Recently confirmed at Příbram, Czech Republic.

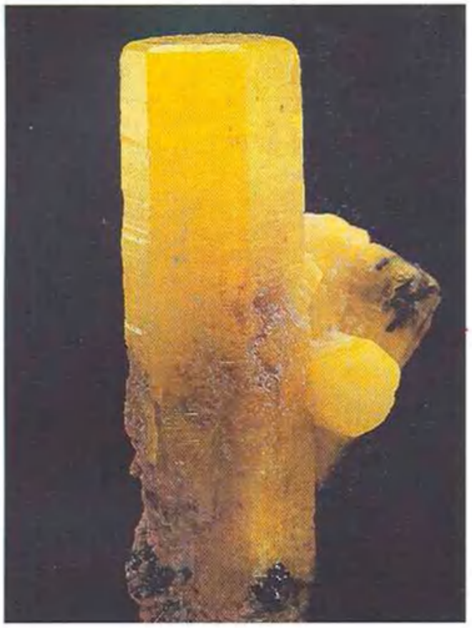

Ettringite, 45 mm, Kuruman, South Africa

Ettringite

$Ca_6Al_2(SO_4)_3(OH)_{12} \cdot 26\ H_2O$

HEXAGONAL ● ●

Properties: C – colorless, yellow; S – white; L – vitreous to silky; D – transparent; DE – 1.8; H – 2.5; CL – perfect; M – prismatic crystals, fibrous aggregates.
Origin and occurrence: Metamorphic, associated with sturmanite. The world's best crystals reaching up to 100 mm (4 in) come from the N'Chwaning mine, Kuruman, South Africa. Crystals up to 4 mm (5/32 in) occurred in Franklin, New Jersey, USA. Also known from the contact metamorphic conditions in Crestmore, California, USA.

Sturmanite

$Ca_6Fe^{3+}{}_2(SO_4)_2[B(OH)_4].(OH)_{12} \cdot 25\ H_2O$

TRIGONAL ●

Properties: C – yellow, yellow-green; S – white; L – vitreous; D – transparent to translucent; DE – 1.9; H – 2.5; CL – perfect; M – flat dipyramidal crystals.
Origin and occurrence: : Probably metamorphic, associated with barite, hematite and ettringite. Crystals up to 140 mm (5½ in) long come from the N'Chwaning mine, Kuruman, South Africa.

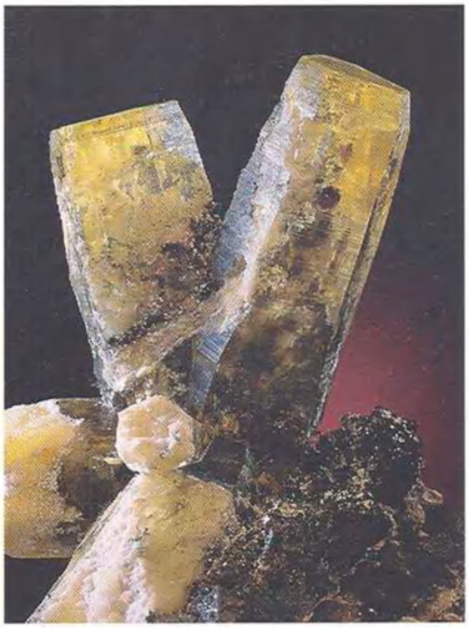

Sturmanite, 35 mm xx, Kuruman, South Africa

Johannite

$Cu(UO_2)_2(SO_4)_2(OH)_2 \cdot 8\ H_2O$

TRICLINIC ● ●

Properties: C – green; S – light green; L – vitreous; D – transparent to translucent; DE – 3.3; H – 2-2.5; CL – good; M – prismatic to thick tabular crystals, scaly aggregates, coatings; R – radioactive.
Origin and occurrence: Secondary, as a result of the uraninite oxidation, together with other secondary U minerals. Crystals are known from Jachymov, Czech Republic and Johanngeorgenstadt, Germany. It was common with zippeite in Central City, Colorado, USA and also reported from Mounana, Gabon.

Zippeite

$K_4(UO_2)_6(SO_4)_3(OH)_{10} \cdot 4\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – orange-yellow; S – yellow; L – dull to earthy; D – opaque; DE – 3.7; H – not determined; CL – perfect; M – acicular and tabular crystals, pulverulent and acicular aggregates, coatings; LU – green; R – radioactive.
Origin and occurrence: Secondary, forming during the uraninite oxidation, associated with other secondary U minerals. Small tabular crystals come from Drmoul, Czech Republic. Coatings and

Johannite, 70 mm, Jachymov, Czech Republic

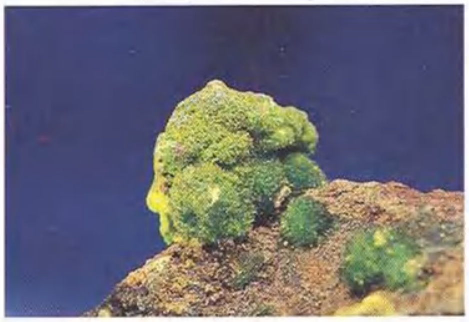

Zippeite, 40 mm, Lodeve, France

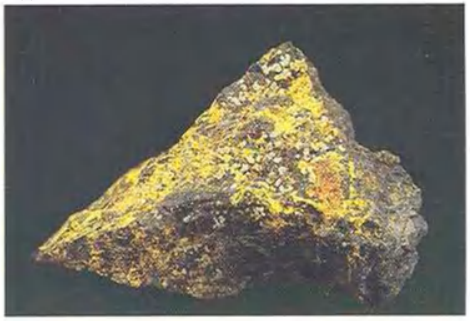

acicular aggregates are known from Jachymov, Czech Republic and Central City, Colorado, USA; also occurs in Shinkolobwe, Zaire.

Scheelite

$CaWO_4$

TETRAGONAL ● ● ● ●

Properties: C – colorless, gray-white, yellow-brown, orange, red, greenish; S – white; L – greasy to adamantine; D – translucent; DE – 6.1; H – 4.5-5; CL – good; F – conchoidal to uneven; M – pseudo-octahedral crystals, granular, massive; LU – blue-white.
Origin and occurrence: Magmatic in pegmatites, hydrothermal in greisens and metamorphic; parageneses vary significantly according to the origin. Beautiful brownish crystals up to 100 mm (4 in) in size come from Taewha and Tongwha in Korea. Similar crystals were recently found in China. Orange crystals up to 40 mm (1$^{9}/_{16}$ in) in size are associated with cassiterite crystals on quartz crystals from Iultin and Tenkergin, Russia. Clear crystals, weighing up to 50 kg (110 lb), occurred in pegmatites near Natas, Namibia. Crystals up to 70 mm (2¾ in) reported from several mines near Traversella, Italy. Beautiful red crystals up to 20 mm ($^{25}/_{32}$ in) were very rare in Pøibram, Czech Republic. Yellow crystals up to 40 mm (1$^{9}/_{16}$ in) are known from quartz veins with pumpellyite in Obøi Dsl, Czech Republic.
Application: W ore.

Scheelite, 77 mm, Sichuan, China

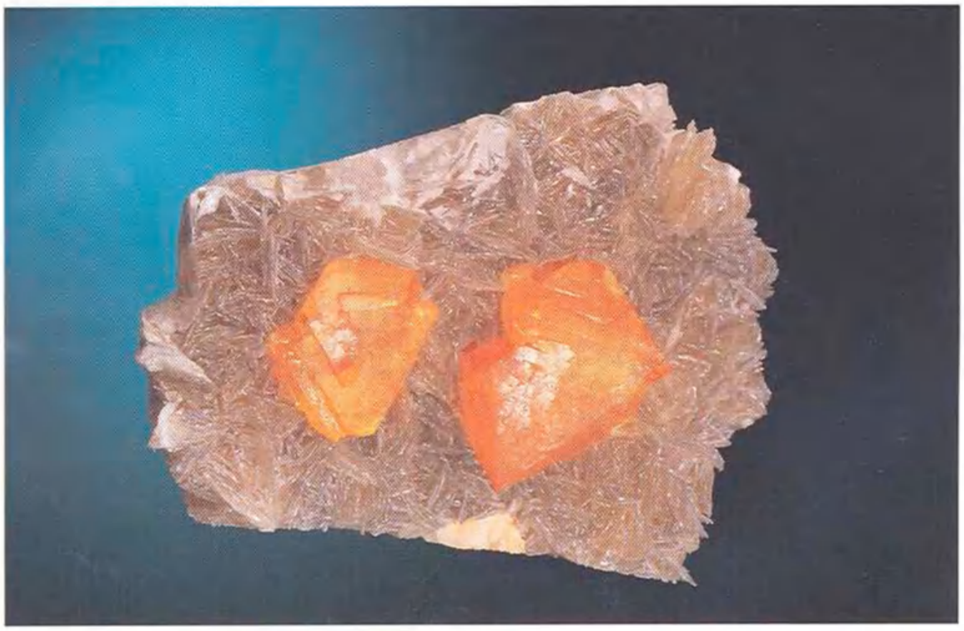

Stolzite, 20 mm, Arizona, U.S.A.

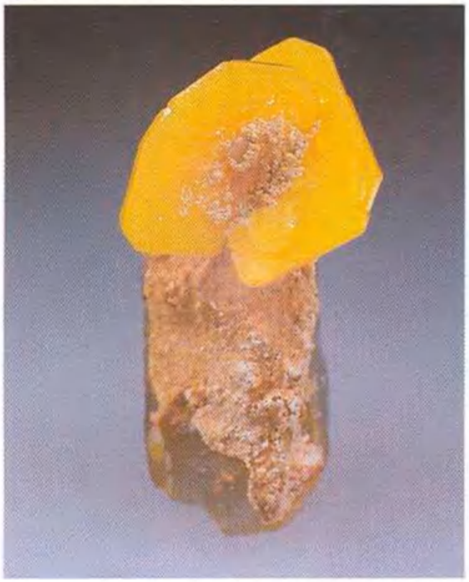

Stolzite

$PbWO_4$

TETRAGONAL ● ●

Properties: C – gray-brown, orange-yellow, red, green; S – colorless; L – adamantine to resinous; D – transparent to translucent; DE – 7.9-8.3; H – 2.5-3; CL – imperfect; F – conchoidal to uneven; M – dipyramidal and thick tabular, striated crystals.

Origin and occurrence: Secondary, as a product of the oxidation of primary W minerals. Crystals up to 60 mm ($2^{3}/_{8}$ in) from St Leger-de-Peyre, France. Prisms and needles up to 25 mm (1 in) long found in Broken Hill, New South Wales, Australia. Crystals up to 25 mm (1 in) also known from Tsumeb, Namibia. Crystals up to 20 mm ($^{25}/_{32}$ in) reported from Cínovec, Czech Republic and from the Black Pine mine, Montana, USA.

Ferrimolybdite, 60 mm, Hurky, Czech Republic

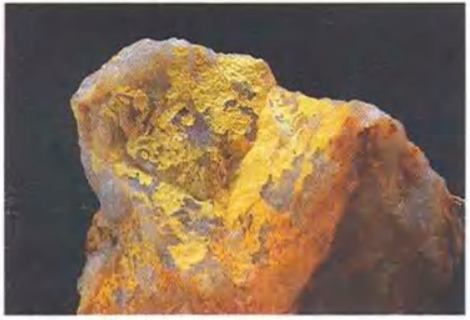

Crocoite, 44 mm, Dundas, Australia

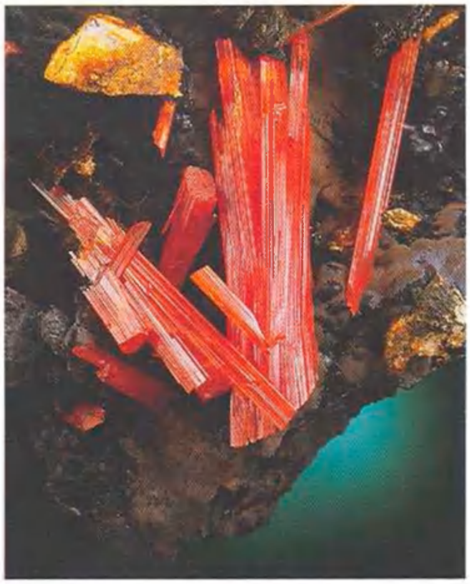

Crocoite

$PbCrO_4$

MONOCLINIC ● ●

Properties: C – orange, red; S – orange-yellow; L – adamantine to greasy; D – translucent; DE – 6; H – 3; CL – good; F – conchoidal to uneven; M – long prismatic to acicular crystals, crusts..

Origin and occurrence: Secondary, as a result of the galena oxidation in basic rocks. The world' best specimens come from the Dundas district, Tasmania, Australia where crystals up to 100 mm (4 in) long were found in several mines. Fine crystals up to 40 mm ($1^{9}/_{16}$ in) long are known from Berezovsk, Ural mountains, Russia. Crystals up to 20 mm ($^{25}/_{32}$ in) were recently found in Callenberg, Germany.

Betpakdalite, 70 mm, Bayan Tsogdo, Mongolia

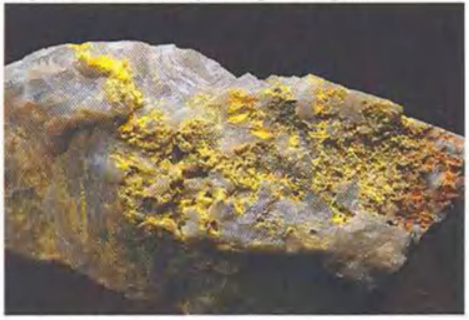

Ferrimolybdite

$Fe_2(MoO_4)_3 . 8 H_2O$

ORTHORHOMBIC ● ●

Properties: C – yellow, whitish; S – light yellow; L – adamantine, silky, earthy; D – opaque; DE – 4.4; H – 1-2; M – small acicular crystals, fibrous and radial aggregates, earthy.
Origin and occurrence: Secondary, as a product of the molybdenite oxidation. Microscopic crystals were found in Glen Innes, New South Wales, Australia. It occurred as yellow coatings in Climax and Telluride, Colorado, in the Getchell mine, Nevada, USA; also in Hsrky near Eista, Czech Republic.

Betpakdalite

$CaFe^{3+}H_8(MoO_4)_5(AsO_4)_2 . 8 H_2O$

MONOCLINIC ●

Properties: C – lemon-yellow; S – yellow; L – vitreous, waxy, dull; D – opaque; DE – 3; H – 3; CL – good; M – short prismatic microscopic crystals, pulverulent.
Origin and occurrence: Secondary mineral. Originally described from Kara-Oba, Kazakhstan. It also occurred in Tsumeb, Namibia and in Krupka, Czech Republic.

Wulfenite

$PbMoO_4$

TETRAGONAL ● ● ●

Properties: C – yellow, orange, brownish, red, greenish; S – white; L – greasy to adamantine; D – transparent to translucent; DE – 6.3-7; H – 3; CL – good; F – uneven to conchoidal; M – thin tabular and dipyramidal crystals, granular, massive.
Origin and occurrence: Secondary, as a result of the galena oxidation, together with cerussite, vanadinite and other minerals. The best specimens are known from the Red Cloud mine near Yuma, Arizona, where red tabular crystals found up to 50 mm (2 in) diameter. Yellow-brown crystals up to 100 mm (4 in) come from the Glove mine, Arizona, USA. Thick tabular orange crystals up to 20 mm ($^{25}/_{32}$ in) in diameter found in the Erupcion mine, Villa Ahumada, Los Lamentos, Chihuahua. Beautiful yellow plates up to 60 mm (2$^{7}/_{8}$ in) with orange mimetite spheres occurred in the San Francisco mine, Magdalena, Sonora, Mexico. Rare tabular crystals up to 70 mm (2¾ in) across come from Tsumeb, Namibia. Fine orange-yellow tabular crystals up to 20 mm ($^{25}/_{32}$ in) across, and pyramidal crystals are known from Bleiberg, Austria and Mežica , Slovenia. Crystals up to 100 mm (4 in) were found recently in Touissit, Morocco.

Wulfenite, 38 mm, Los Lamentos, Mexico

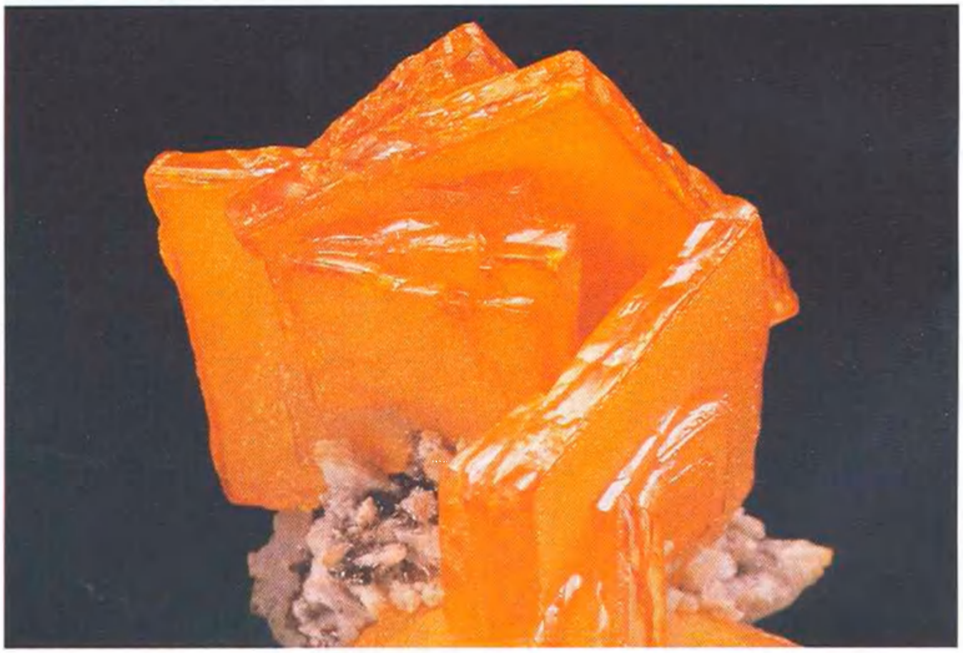

8. Phosphates

Lithiophosphate

Li_3PO_4

ORTHORHOMBIC ●

Properties: C – colorless, white, pinkish; S – white; L – vitreous; D – transparent to translucent; DE – 2.5; H – 4; CL – perfect; F – uneven; M – prismatic crystals, granular.
Origin and occurrence: Hydrothermal in granitic pegmatites, where it forms by replacement of spodumene and montebrasite. Cleavable aggregates up to 100 mm (4in) across occur in the Tanco Mine, Bernic Lake, Manitoba, Canada; crystals up to 25 mm (1 in) across in the Foote mine, Kings Mountain, North Carolina, USA.

Beryllonit

$NaBePO_4$

MONOCLINIC ● ●

Properties: C – colorless, white, yellowish; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 5.5-6; CL – perfect; F – uneven; M – short prismatic to tabular crystals, granular.
Origin and occurrence: Hydrothermal in cavities in granitic pegmatites where it is associated with herderite, albite and tourmaline. Crystals and their twins, up to 150 mm (6 in) across come from Stoneham and Newry, Maine, USA; also from Viitaniemi, Finland and Paprok, Afghanistan.

Adamite, 52 mm, Mapimi, Mexico
Lithiophosphate, 30 mm, Tanco, Canada

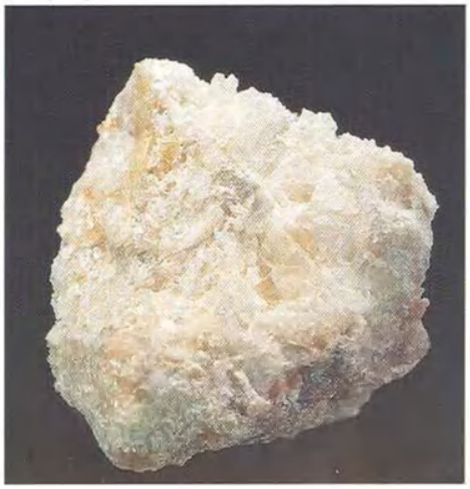

Beryllonite, 31 mm, Paprok, Afghanistan

Triphyllite

$LiFePO_4$

ORTHORHOMBIC ● ● ●

Properties: C – gray-green, gray-blue, gray, brown; S – gray-white; L – greasy to vitreous; D – transparent to translucent; DE – 3.4; H – 4-5; CL – good; F – uneven; M B short prismatic crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites, associated with graftonite, sarcopside and many secondary phosphates. Large triphyllite crystals up to 1.5 m (5 ft) across are known from Hagendorf, Germany; from the Tip Top mine, Custer, South Dakota and Palermo No. 1 mine, North Groton, New Hampshire, USA; also from Hühnerkobel, Germany.

Triphylite, 70 mm, Hagendorf, Germany

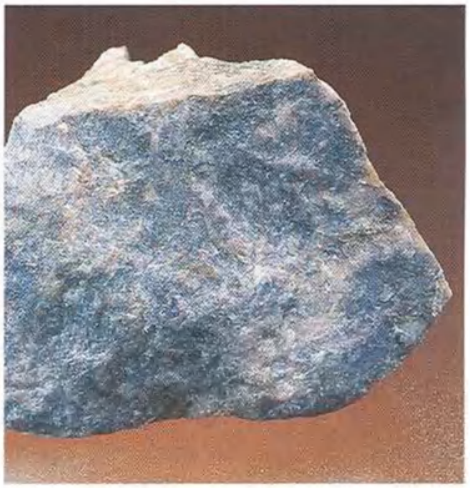

Lithiophilite, 120 mm, Viitaniemi, Finnland

Lithiofilit

$LiMnPO_4$

ORTHORHOMBIC ● ●

Properties: C – pink, red-brown, brown; S – white; L – vitreous; D – transparent to translucent; DE – 3.3; H – 4-5; CL – good; F – uneven; M – short prismatic crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites. It is sometimes associated with triplite and triphyllite, typically replaced by many secondary phosphates. Large masses of lithiophillite, reaching up to 1 m ($39^{3}/_{8}$ in), occur in Karibib, Namibia; Kitumbe, Rwanda. Other localities are Mangualde, Portugal; Tanco mine, Bernic Lake, Manitoba, Canada; Stewart Lithia Mine, Pala, California, USA.

Purpurite, 40 mm, Usakos, Namibia

Berzeliite, 120 mm, Langban, Sweden

Purpurite

$Mn^{3+}PO_4$

ORTHORHOMBIC ● ●

Properties: C – pink, purple, dark brown; S B red-purple; L – dull to velvet; D – translucent to opaque; DE – 3.7; H – 4-4.5; CL – good; F – uneven; M B granular, massive.
Origin and occurrence: Hydrothermal, as a product of lithiophillite replacement in granitic pegmatites. It is usually associated with many secondary phosphates. It is known from Kitumbe, Rwanda; Usakos and Sandamab, Namibia; the Tip Top and Bull Moose mines, Custer, South Dakota and Branchville, Connecticut, USA.

Berzeliite

$(Ca,Na)_3\ (Mg,Mn)_2\ (AsO_4)_3$

CUBIC ● ●

Properties: C – yellow, orange; S – red-purple; L – resinous; D – transparent to translucent; DE – 4.1; H – 4-4.5; CL – none; F – conchoidal to uneven; M – isometric crystals, granular, massive.
Origin and occurrence: Metamorphic, together with haussmanite, rhodonite and tephroite. It occurs as massive in Langban and Nordmark, Sweden.

Whitlockite

$Ca_9(Mg,Fe)H(PO_4)_7$

TRIGONAL ● ●

Properties: C – colorless, white, yellowish, pinkish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 5; CL – none; F – conchoidal to uneven; M – rhombohedral crystals, granular, massive.

Origin and occurrence: Hydrothermal, as a product of replacement of primary phosphates in granitic pegmatites, rare in sedimentary rocks – phosphorites: very rare magmatic in meteorites. Mainly associated with apatite and carbonates. Imperfect crystals around 10 mm (3/8 in) across, occur in the Palermo No. 1 mine, North Groton, New Hampshire; also in the Tip Top mine, Custer, South Dakota, USA.

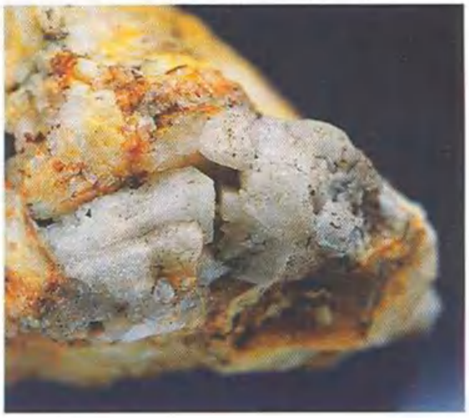
Whitlockite, 10 mm xx, New Hampshire, U.S.A.

Xenotime-(Y)
YPO_4

TETRAGONAL ● ● ●

Properties: C – brown, yellow, gray, greenish; S – white; L – vitreous to resinous; D – transparent, translucent to opaque: DE – 4.5; H – 4.5; CL – good; F – conchoidal to uneven; M – long prismatic to tabular crystals, granular; R – sometimes weakly radioactive and metamict.
Origin and occurrence: Magmatic in granitic and alkaline pegmatites, granites and syenites; hydrothermal in the Alpine-type veins; metamorphic in gneisses; common in placers. It is associated with monazite-(Ce) and zircon. Perfect prismatic crystals up to 100 mm (4in) across occur mainly in pegmatites in Kragero and Hitterö, Norway; Ytterby, Sweden; in several places in Madagascar; in Ichikawa, Japan. Crystals about 20 mm (25/32 in) are known from the Alpine-type veins in Binntal, Switzerland.
Application: ore of rare earth elements.

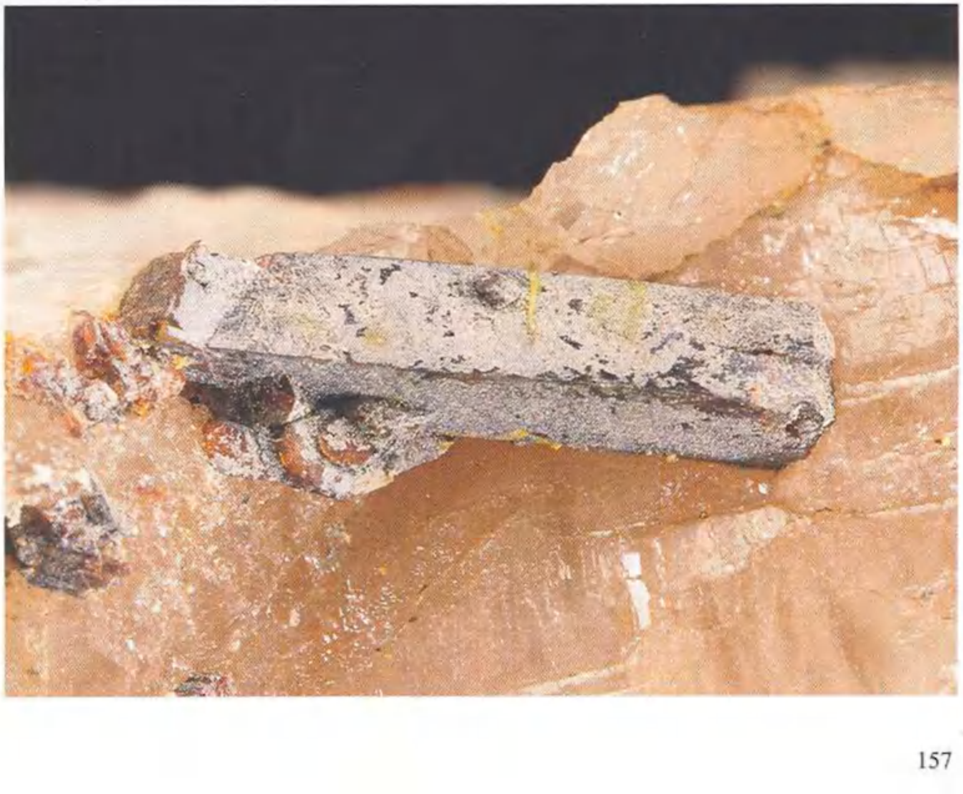
Xenotime-(Y), 14 mm x, Washington Co., U.S.A.

Monazite-(Ce)

$CePO_4$

MONOCLINIC ● ● ●

Properties: C – yellow, brown, red-brown, orange, gray-green; S – white; L – vitreous to resinous; D – transparent, translucent to opaque; DE – 4.6; H – 5-5.5; CL – good; F – conchoidal to uneven; M – long prismatic to tabular crystals, granular; R – sometimes weakly radioactive and metamict.

Origin and occurrence: Magmatic in granitic and alkaline pegmatites, granites, syenites, carbonatites; hydrothermal in the Alpine-type veins and greisens; metamorphic in gneisses; common in placers.

It is associated with apatite, xenotime-(Y) and zircon. Perfect prismatic crystals up to 200 mm (7 7/8 in) come from Mars Hill, North Carolina, Trout Creek Pass, Colorado, USA; also from Arendal, Norway; Ambatofotsikely and Ampangabe, Madagascar, where masses weighing several kg are common; crystals, up to 200 mm (7 7/8 in) across found in Minas Gerais, Jaguaraçu, Brazil.

Application: ore of rare earth elements.

Monazite-(Ce), 31 mm, Felicio, Brazil

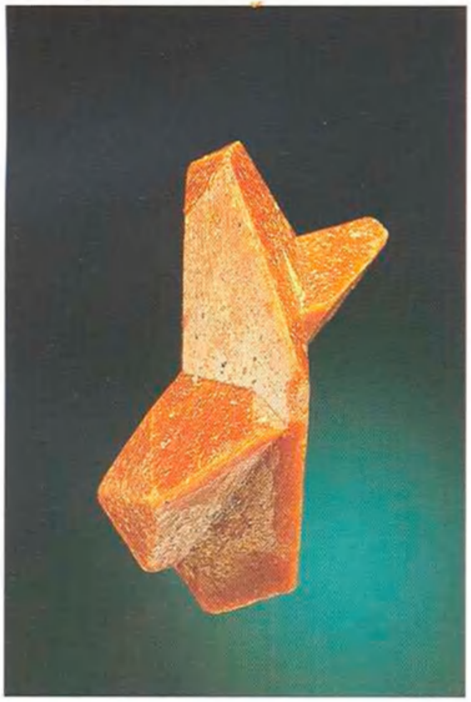

Hydroxylherderite

$CaBe(PO_4)(OH,F)$

MONOCLINIC ● ●

Properties: C – colorless, white, yellowish, greenish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 5-5.5; CL – good; F – conchoidal to uneven; M – prismatic to tabular crystals, radial aggregates, granular.

Origin and occurrence: Hydrothermal in cavities in granitic pegmatites and in greisens. Perfect crystals, up to 120 mm (4 11/16 in) long come from Marilaca and together with colored tourmalines from Virgem da Lapa, Minas Gerais, Brazil; also known from Topsham and Stoneham, Maine, USA.

Amblygonite

$LiAl(PO_4)(F,OH)$

TRICLINIC ● ● ●

Properties: C – colorless, white, yellowish, bluish, gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 5.5-6; CL – good; F – conchoidal to uneven; M – short prismatic crystals, granular, massive.

Hydroxylherderite, 22 mm, Anza, California, U.S.A.

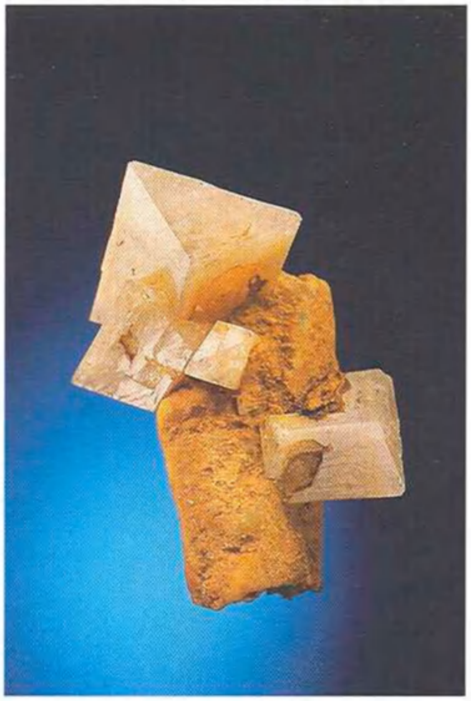

Montebrasite, 80 mm, White Picacho, Arizona, U.S.A.

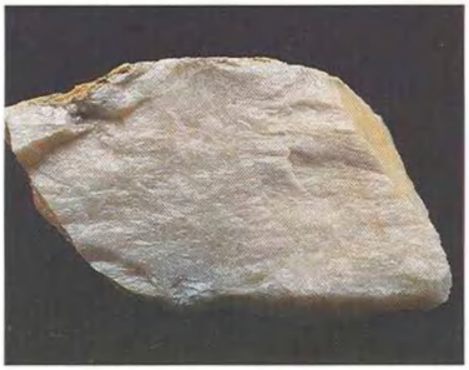

Origin and occurrence: Magmatic in granitic pegmatites and some granites; rare hydrothermal in greisens and in ore veins. Large amblygonite masses, several meters across, come from pegmatites, like the Beecher, Custer and Hugo mines, Keystone, South Dakota, USA, where its blocks weighed up to 200 tons; also known from Viitaniemi, Finland; Uto, Sweden. Typical representative of the quartz-amblygonite veins is Vernerov, Czech Republic.
Application: Li ore and raw material for ceramics.

Montebrasite

$LiAl(PO_4)(OH,F)$

TRICLINIC ● ● ●

Properties: C – colorless, white, yellowish, yellow, bluish, gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 5.5-6; CL – good; F – conchoidal to uneven; M – short prismatic to tabular crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites and granites; hydrothermal in granitic pegmatites. It occurs very frequently with amblygonite, which it replaces. Large masses are known from Montebras, France; also from the Tin Mountain mine, Custer, South Dakota, USA.. Perfect tabular crystals up to 150 mm (6 in) across found in Taquaral, Minas Gerais, Brazil.

Wagnerite

$Mg_2(PO_4)(F,OH)$

MONOCLINIC ● ● ●

Properties: C – light yellow, yellow-green, yellow-brown, green; S – white; L – vitreous to greasy; D – transparent, translucent to opaque; DE – 3.2; H – 5-5.5; CL – imperfect; F – conchoidal to uneven; M – short prismatic crystals, granular, massive.

Wagnerite, 18 mm x, Werfen, Austria

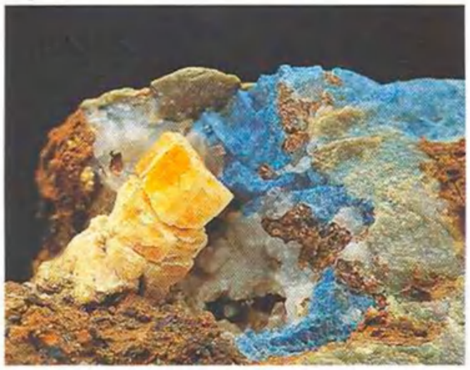

Origin and occurrence: Magmatic in granitic pegmatites; metamorphic in gneisses and eclogites; hydrothermal in quartz veins and in salt deposits. The most famous finds come from the quartz veins in Hollgraben and Radelgraben, the Alps, Austria, where wagnerite occur together with lazulite and forms crystals up to 30 mm ($1^{3}/_{16}$ in) across; also known from Mangualde, Portugal and Bodenmais, Germany.

Amblygonite, 55 mm, Minas Gerais, Brazil

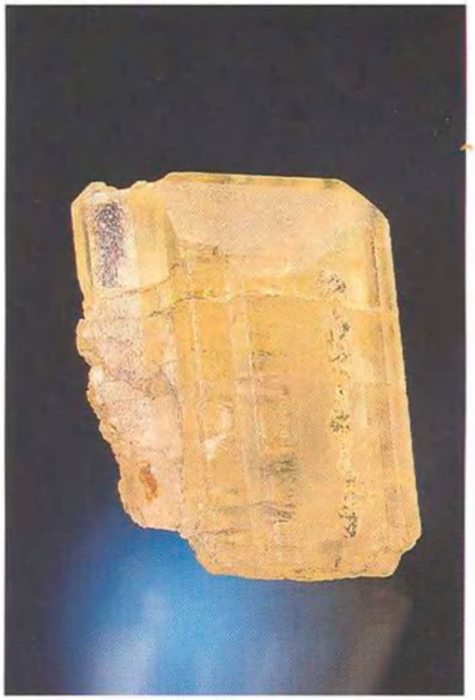

Zwieselite, 30 mm grain, Dolni Bory, Czech Republic

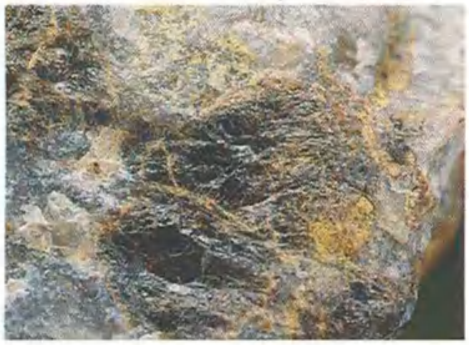

Zwieselite

$Fe_2(PO_4)(F,OH)$

MONOCLINIC ● ● ●

Properties: C – dark brown to black-brown; S – light brown; L – vitreous to greasy; D – translucent to opaque; DE – 4.0; H – 5-5.5; CL – imperfect; F – conchoidal to uneven; M – short prismatic crystals, granular, massive.

Origin and occurrence: Magmatic in granitic pegmatites; rare hydrothermal in greisens. It is associated with apatite, triplite and secondary phosphates. Imperfect crystals and granular aggregates are known from pegmatites near Zwiesel, Germany; Dolni Bory, Czech Republic.

Libethenite, 80 mm, L'ubietova, Slovakia

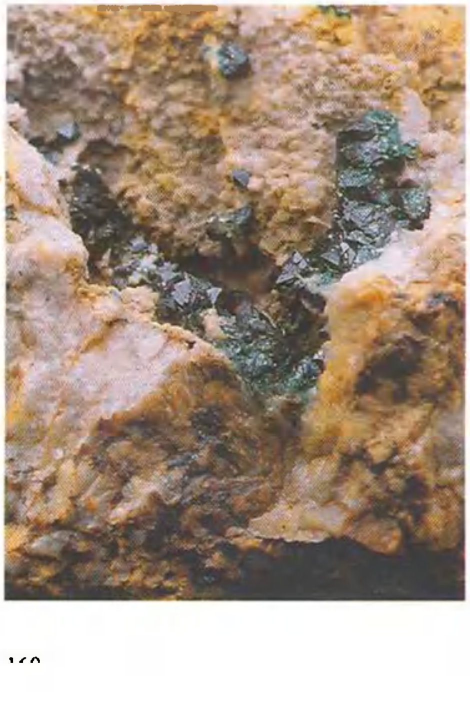

Triplite, 70 mm, Viitaniemi, Finnland

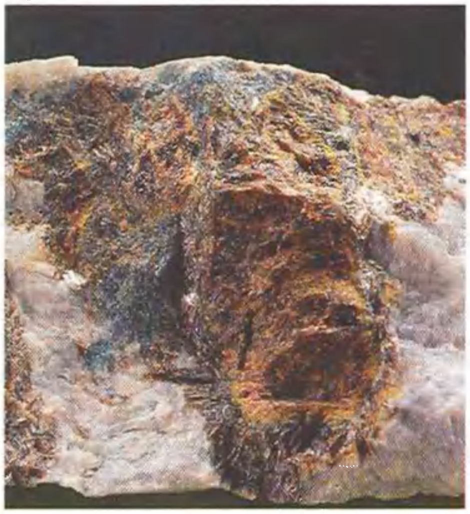

Triplite

$Mn_2(PO_4)(F,OH)$

MONOCLINIC ● ● ●

Properties: C – pink, light brown; S – light brown; L – vitreous to greasy; D – translucent to opaque; DE – 3.8; H – 5-5.5; CL – imperfect; F – conchoidal to uneven; M – short prismatic crystals, granular, massive.

Origin and occurrence: Magmatic in granitic pegmatites and granites; hydrothermal in greisens and in quartz veins. It is usually hydrothermally altered and replaced by secondary phosphates. Huge masses several meters in size come from pegmatites in Karibib, Namibia; it also occurs in Mangualde, Portugal and in Sukula, Finland.

Libethenit

$Cu_2(PO_4)(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – black-green to light green; S – olive-green; L – greasy; D – translucent; DE – 3.9; H – 4; CL – imperfect; F – conchoidal to uneven; M – short prismatic and dipyramidal crystals, botryoidal aggregates, granular, massive.

Origin and occurrence: Secondary in Cu deposits, where it occurs together with malachite, pseudomalachite and brochantite. Perfect crystals, up to 30 mm ($1\frac{3}{16}$ in) across, come from the Rokana mine, Zambia; crystals, up to 10 mm ($\frac{3}{8}$ in), are known from Kambove, Zaire; L'ubietova . Slovakia; Nizhniy Tagil, Ural Mts., Russia.

Olivenite, 80 mm, Tsumeb, Namibia

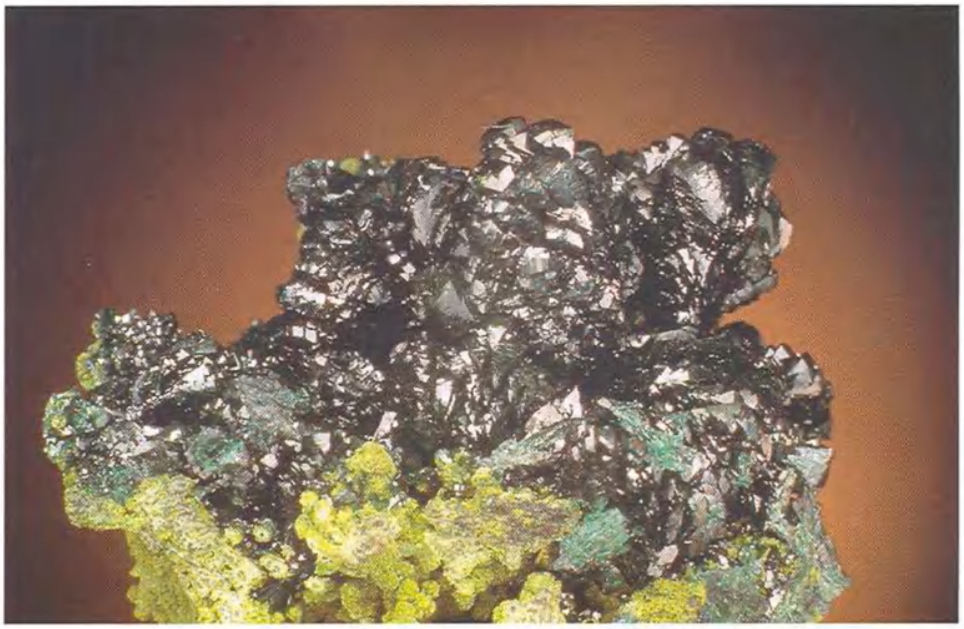

Olivenite

$Cu_2(AsO_4)(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – olive-green, green-brown, gray-green to gray; S – light green; L – greasy; D – translucent to opaque; DE – 4.4; H – 3; CL – imperfect; F – conchoidal to uneven; M – long to short prismatic and dipyramidal crystals, acicular and radial aggregates, massive.
Origin and occurrence: Secondary in Cu deposits. It is associated with other secondary Cu minerals, malachite, azurite and scorodite. Perfect crystals up to 10 mm ($^3/_8$ in) across occurred in Wheal Gorland, Cornwall, and Tavistock, Devon UK; also in Tsumeb, Namibia and Ashburton Downs, Western Australia, Australia.

Adamine

$Zn_2(AsO_4)(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – yellow-green, yellow, green, colorless, purple; S – white; L – vitreous; D – transparent to translucent; DE – 4.4; H – 3.5; CL – imperfect; F – conchoidal to uneven; M – long to short prismatic and dipyramidal crystals, acicular and radial aggregates, massive; LU – yellow-green.

Origin and occurrence: Secondary in Zn deposits, associated with hemimorphite, goethitee and smithsonite. Rich druses of green and rare purple crystals up to 70 mm (2¾ in) across come from Mina Ojuela, Mapimi, Durango, Mexico; also known from Tsumeb, Namibia; Laurion, Greece and Cap Garonne, France.

Adamite, 34 mm, Mapimi, Mexico

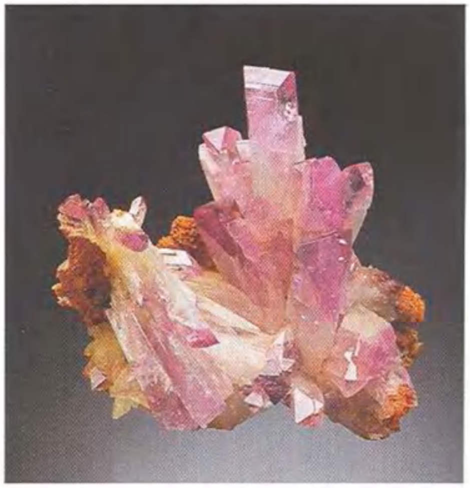

Lazulite, 11 mm x, Lincoln Co., U.S.A.

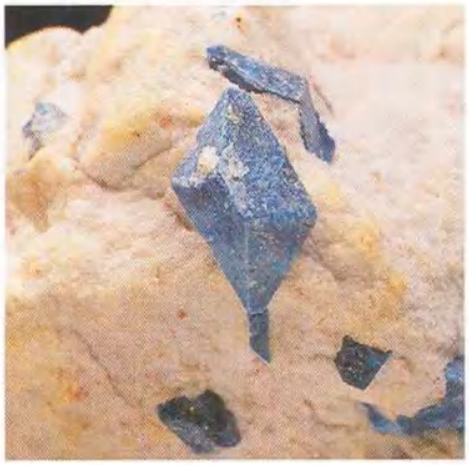

Lazulite

$MgAl_2(PO_4)_2(OH)_2$

MONOCLINIC ● ● ●

Properties: C – dark to light blue, blue-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.1; H – 5.5-6; CL – imperfect; F – conchoidal to uneven; M – short prismatic and dipyramidal crystals, granular, massive.
Origin and occurrence: Hydrothermal in quartz veins and granitic pegmatites, where it is formed by decomposition of primary phosphates; metamorphic in quartzites.
Imperfect crystals up to 100 mm (4 in) across come from Horrsjoberg, Sweden; perfect crystals about 50 mm found in Ashudi, Pakistan; also known from Big Fish River, Yukon, Canada and near Werfen, Austria.

Scorzalite, 80 mm, Palermo No.2 Mine, U.S.A.

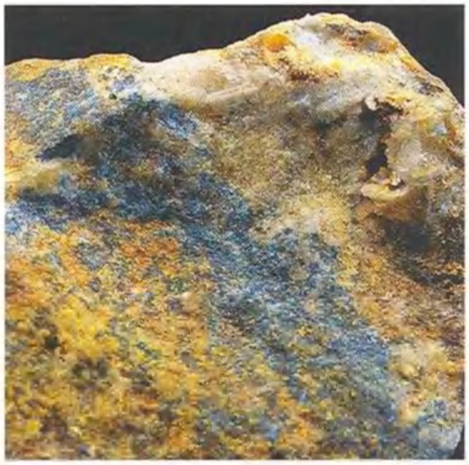

Scorzalite

$FeAl_2(PO_4)_2(OH)_2$

MONOCLINIC ● ● ●

Properties: C – dark blue, blue-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.3; H – 6; CL – imperfect; F – conchoidal to uneven; M – prismatic crystals, granular, massive.
Origin and occurrence: Hydrothermal in granitic pegmatites, as a replacement product of primary phosphates, rare in quartz veins. Dark blue, granular aggregates up to 100 mm (4in) across occur in the Palermo No. 1 and No. 2 mines, North Groton, New Hampshire and the Victory mine, Custer, South Dakota, USA.

Rockbridgeite

$Fe^{2+}Fe^{3+}{}_4(PO_4)_3(OH)_5$

ORTHORHOMBIC ● ● ●

Properties: C – dark and light green, black-green; S – green; L – dull; D – translucent to opaque; DE – 3.4; H – 4.5; CL – good; F – uneven; M – acicular crystals, radial aggregates and crusts, granular, massive.
Origin and occurrence: Secondary in granitic pegmatites and in Fe deposits. It mostly originates from the hydrothermal alteration of primary phosphates, mainly triphyllite, and it is associated with other secondary phosphates. Rich radial aggregates up to 50 mm (2 in) across come from Hagendorf, Germany; the Tip Top mine, Custer, South Dakota and the Fletcher mine, Groton, New Hampshire, USA.

Rockbridgeite, 70 mm, Hagendorf, Germany

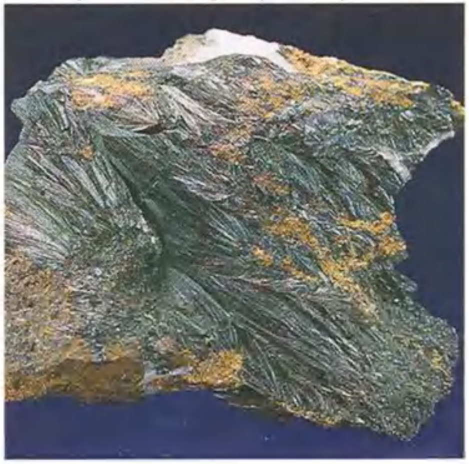

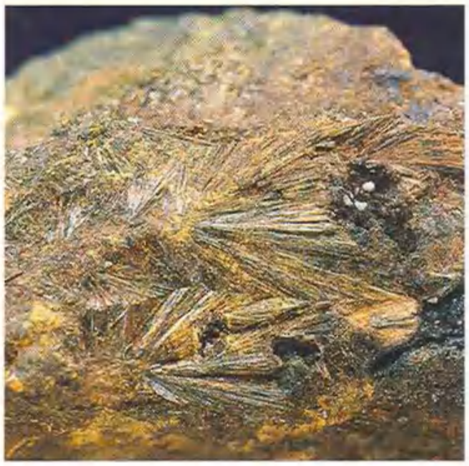

Frondelite, 15 mm xx, Custer, U.S.A.

Frondelite

$MnFe^{3+}_4(PO_4)_3(OH)_5$

ORTHORHOMBIC ● ● ●

Properties: C – light olive-green, brown, black-green; S – green; L – dull; D – translucent to opaque; DE – 3.5; H – 4.5; CL – good; F – uneven; M – acicular crystals, radial aggregates and crusts, granular, massive.

Origin and occurrence: Secondary in granitic pegmatites, where it forms as a result of the hydrothermal alteration of primary phosphates, mainly lithiophillite. Radial aggregates occur in the Fletcher mine, Groton, New Hampshire, USA; also in Sapucaia, Minas Gerais, Brazil.

Dufrenite

$Fe^{2+}Fe^{3+}_4(PO_4)_3(OH)_5 . 2 H_2O$

MONOCLINIC ● ● ●

Properties: C – dark green, black-green, black; S – green; L – vitreous to dull; D – translucent to opaque; DE – 3.4; H – 3.5-4.5; CL – good; F – uneven; M – radial aggregates and crusts, granular, massive.

Origin and occurrence: Secondary in granitic pegmatites and in the oxidation zone of Fe deposits: it forms in pegmatites as a result of the hydrothermal alteration of primary phosphates. Radial aggregates occur in pegmatites in Hagendorf and Hühnerkobel, Germany; also in gossan in the Wheal Phoenix mine, Cornwall, UK.

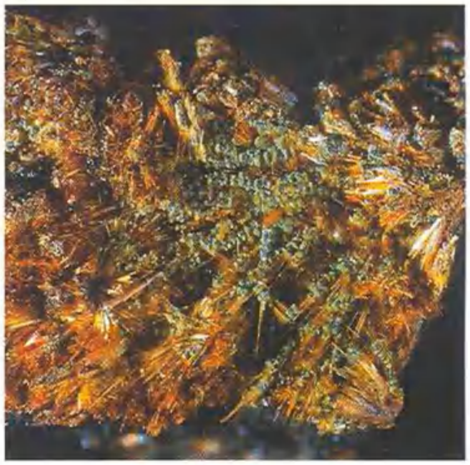

Dufrenite, 9 mm, Lancaster, U.S.A.

Pseudomalachite

$Cu_5(PO_4)_2(OH)_4$

MONOCLINIC ● ● ●

Properties: C – green, black-green; S – green; L – vitreous to dull; D – translucent to opaque; DE – 4.3; H – 4.5-5; CL – good; F – uneven; M – short prismatic crystals, botryoidal aggregates with radial structure, massive.

Origin and occurrence: Secondary in Cu deposits, associated with malachite, chrysocolla, libethenite and goethite. Beautiful botryoidal aggregates and large masses come from Nizhniy Tagil, Ural mountains, Russia; also from Ehl and Virneberg, Germany; from L'ubietova , Slovakia; and also from many localities in Shaba province, Zaire.

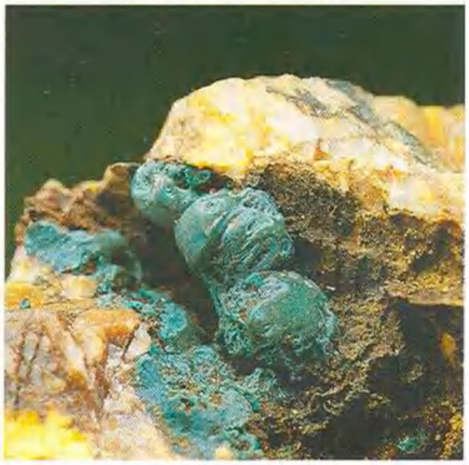

Pseudomalachite, 80 mm, L'ubietova, Slovakia

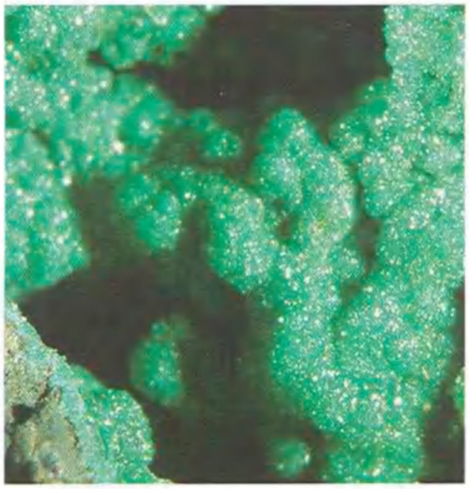

Cornubite, 30 mm, Farbiste, Slovakia

Cornubite

$Cu_5(AsO_4)_2(OH)_4$

TRICLINIC ●

Properties: C – light to dark green; S – light green; L – vitreous to dull; D – translucent; DE – 4.8; H – not determined; CL – not determined; F – uneven; M – botryoidal aggregates, massive.
Origin and occurrence: Secondary in Cu deposits. It was found together with olivenite and clinoclase in the Bedford United quarry, Cornwall, UK; also in Ashburton Downs, Western Australia, Australia and Farbiste, Slovakia.

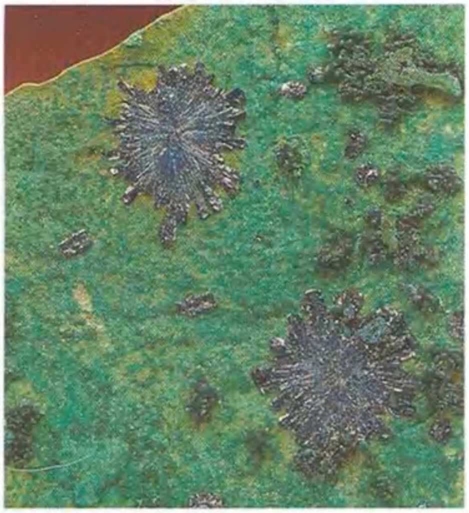

Cornetite, 10 mm aggregate, Lumumbashi, Zair

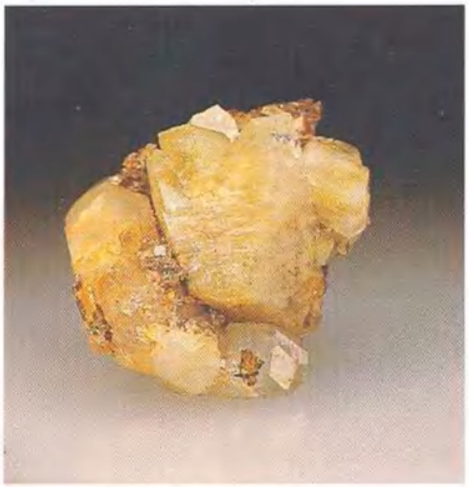

Augelite, 34 mm, Yukon Territory, Canada

Augelite

$Al_2(PO_4)(OH)_3$

MONOCLINIC ● ● ●

Properties: C – colorless, white, yellowish, pinkish; S- white; L – vitreous to dull; D – transparent to translucent; DE – 2.7; H- 4.5-5; CL – good; F – uneven; M – thick tabular to prismatic crystals, granular, massive.
Origin and occurrence: Hydrothermal in granitic pegmatites as a product of primary phosphates replacement; rare metamorphic in quartzites. Massive aggregates several decimeters across occur in pegmatites in Burango, Rwanda; also in the Hugo mine, Custer, South Dakota and Mount White, California, USA. Crystals up to 20 mm ($^{25}/_{32}$ in) across come from Rapid Creek, Yukon, Canada and from the Champion mine, California; rare small crystals found in the Palermo No. 1 mine, North Groton, New Hampshire, USA.

Cornetite

$Cu_3(PO_4)(OH)_3$

ORTHORHOMBIC ●

Properties: C – dark blue to blue-green; S – light blue; L – vitreous; D – transparent to translucent; DE – 4.1; H – 4.5; CL – none; F – conchoidal to uneven; M – short prismatic crystals, coatings.
Origin and occurrence: Secondary in Cu deposits. It is rare in the Etoile mine near Lumumbashi and in Kalagi, Shaba province, Zaire; also occurs in Yerington, Nevada and in Saginaw Hill, Arizona, USA.

Clinoclase, 40 mm, Cornwall, UK

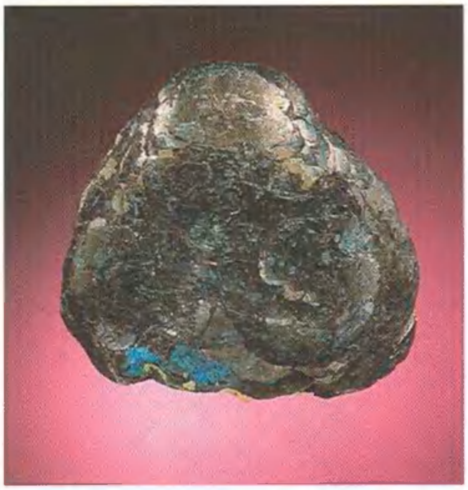

Clinoclase

$Cu_3(AsO_4)(OH)_3$

MONOCLINIC ● ●

Properties: C – dark green-blue to black-green; S – blue-green; L – vitreous; D – transparent to translucent; DE – 4.4; H – 2.5-3; CL – perfect; F – uneven; M – prismatic and tabular crystals, botryoidal aggregates.
Origin and occurrence: Secondary in Cu deposits, often associated with malachite, azurite and other secondary Cu minerals. Spherical aggregates up to 10 mm (/8 in) known from the Majuba Hill mine, Nevada, USA; occurs rarely near Tavistock, Devon, UK; also in Novoveska Huta, Slovakia.

Conichalcite

$CuCa(AsO_4)(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – yellow-green to emerald-green; S – light green; L – vitreous to greasy; D – transparent to translucent; DE – 4.3; H – 4.5; CL – none; F – uneven; M – short prismatic crystals, botryoidal aggregates with radial structure, massive.
Origin and occurrence: Secondary in Cu deposits. It occurs as rich botryoidal aggregates in Otavi, Namibia; Tintic, Utah, in the Higgins mine, Bisbee, Arizona and Yerington, Nevada, USA.

Conichalcite, 70 mm, Tintic, U.S.A.

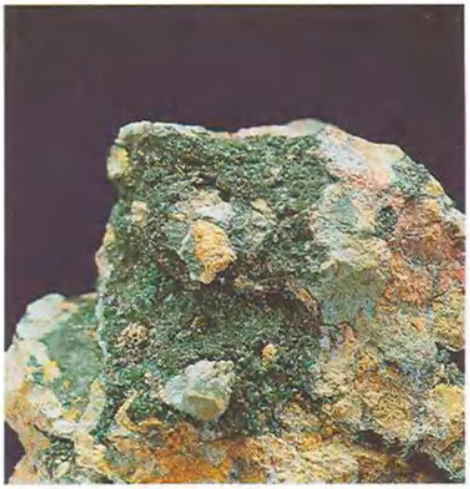

Duftite

$PbCu(AsO_4)(OH)$

ORTHORHOMBIC ● ●

Properties: C – yellow-green, olive-green to gray-green; S – light green; L – vitreous to greasy; D – translucent; DE – 6.5; H – 3; CL – not determined; F – uneven; M – small crystals, botryoidal aggregates and coatings.
Origin and occurrence: Secondary in base metals deposits. It is associated with malachite and azurite in Tsumeb, Namibia; in Mina Ojuela, Mapimi, Durango, Mexico; in Moldava, Czech Republic.

Duftite, 90 mm, Tsumeb, Namibia

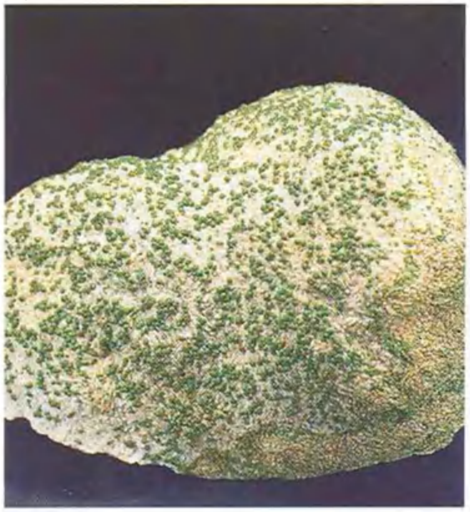

Descloizite, 38 mm, Grootfontein, Namibia

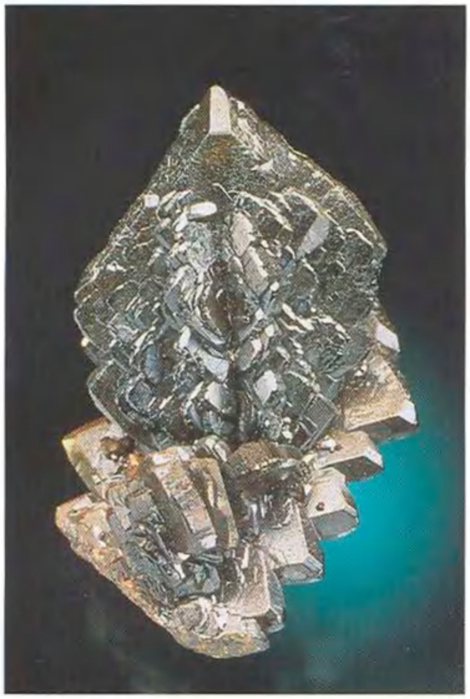

Descloizite

$PbZn(VO_4)(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – red-orange, red-brown to brown-black, gray-green; S – light yellow-brown; L – greasy; D – transparent, translucent to opaque; DE – 6.1; H – 3-3.5; CL – none; F – conchoidal to uneven; M – crystals of different habits, mostly pyramidal or prismatic, botryoidal and skeletal aggregates, massive.
Origin and occurrence: Secondary in base metals deposits. It is mainly associated with pyromorphite, mimetite, vanadinite and other secondary Pb minerals. Occurs as crystals up to 30 mm ($1\frac{3}{16}$ in) long in Tsumeb and Berg Aukas, Namibia; also Broken Hill, Zambia; and the Mammoth mine, Tiger, Arizona, USA.

Arsendescloizite

$PbZn(AsO_4)(OH)$

ORTHORHOMBIC ●

Properties: C – light yellow; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 6.1; H – 4; CL – none; F – conchoidal to uneven; M – tabular crystals, rosette-like aggregates.

Arsendescloizite, 95 mm, Mapimi, Mexico

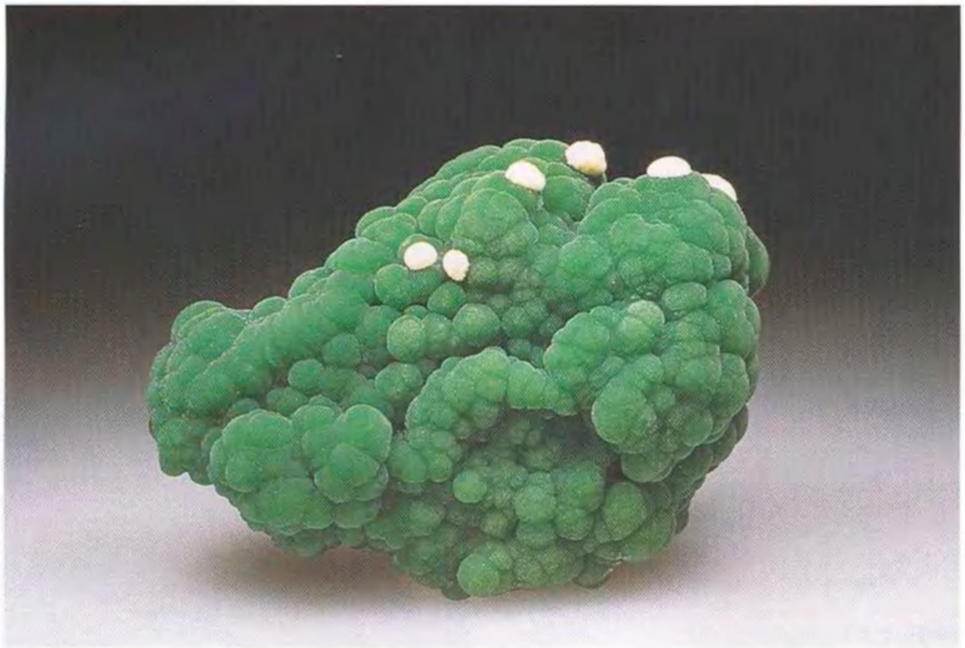

Origin and occurrence: Secondary in base metals deposits, associated with mimetite and goethite. Found rarely as crystals of 1 mm ($^{1}/_{32}$ in) in size in Tsumeb, Namibia; also Mina Ojuela, Mapimi, Durango, Mexico.

Mottramite

$PbCu(AsO_4)(OH)$

ORTHORHOMBIC ● ●

Properties: C – grass-green to black-green; S – light green; L – vitreous to dull; D – transparent to opaque; DE – 5.9; H – 3-3.5; CL – none; F – conchoidal to uneven; M – crystals of different habits, botryoidal and dendritic aggregates, crusts and coatings, massive.
Origin and occurrence: Secondary in base metals deposits, associated with mimetite, descloizite and vanadinite. It occurs in Mottram, Cheshire, UK; in Tsumeb, Namibia; Mammoth mine, Tiger, Arizona, USA.

Brazilianite

$NaAl_3(PO_4)_2(OH)_4$

MONOCLINIC ● ● ●

Properties: C – colorless, white, yellowish, yellow-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.0; H – 5.5; CL – good; F – uneven; M – short prismatic to isometric crystals, radial aggregates, granular.
Origin and occurrence: Hydrothermal in cavities in granitic pegmatites, where it is associated with fluorapatite, albite and tourmaline. It occasionally originates as a product of amblygonite replacement. Perfect yellow-green crystals up to 150 mm (6 in) across found in cavities in pegmatites in Conselheira Pena and Corrego Frio, Linopolis, Minas Gerais and from Pietras Lavradas, Paraiba, Brazil.

Mottramite, 60 mm, Tiger, Arizona, U.S.A.

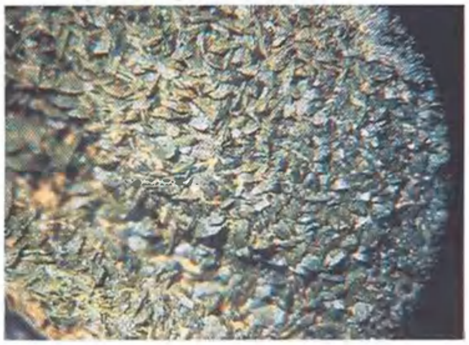

Cafarsite, 9 mm x, Binntal, Switzerland

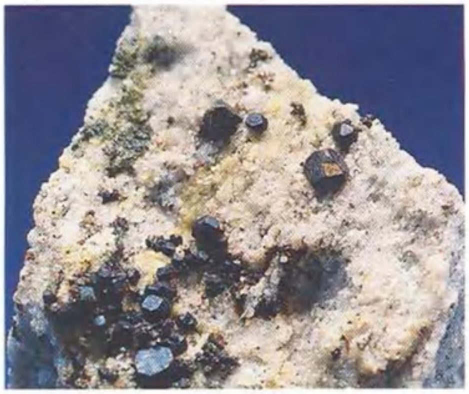

Cafarsite

$Ca_8(Ti,Fe^{2+},Fe^{3+},Mn)(As^{3+}O_3)_{12} \cdot 4\,H_2O$

CUBIC ●

Properties: C – dark brown; S – yellow-brown; L – submetallic; D – translucent; DE – 3.9; H – 5.5-6; CL – none; F – conchoidal; M – isometric crystals.
Origin and occurrence: Hydrothermal along cracks in Alpine-type veins. Cubic crystals up to 30 mm (1$^{3}/_{16}$ in) across come from Binntal, Switzerland and Pizzo Cervandone, Italy.

Brazilianite, 47 mm, Conselheira Pena, Brazil

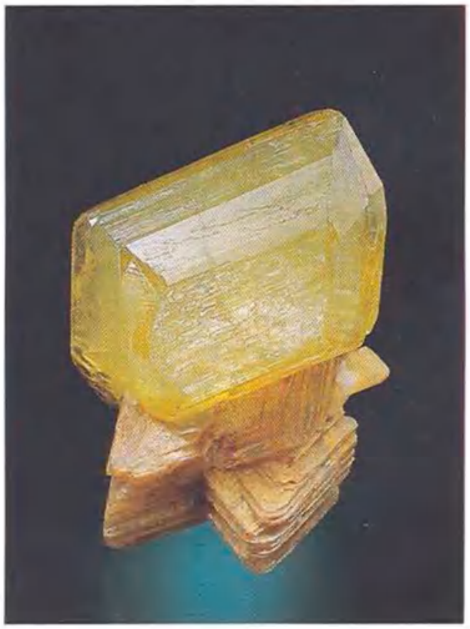

Carminite

$PbFe^{3+}_2(AsO_4)_2(OH)_2$

ORTHORHOMBIC ● ●

Properties: C – crimson-red, red-brown; S – red-yellow; L – adamantine to pearly; D – translucent; DE – 5.5; H – 3.5; CL – good; F – conchoidal to uneven; M – prismatic crystals, acicular, radial, felt-like to porous aggregates, coatings and crusts.
Origin and occurrence: Secondary in base metals deposits, associated with mimetite and scorodite. Crystals up to 10 mm ($^3/_8$ in) across occur in Tsumeb, Namibia; in Mina Ojuela, Mapimi, Durango, Mexico; in Calstock, Cornwall, UK.

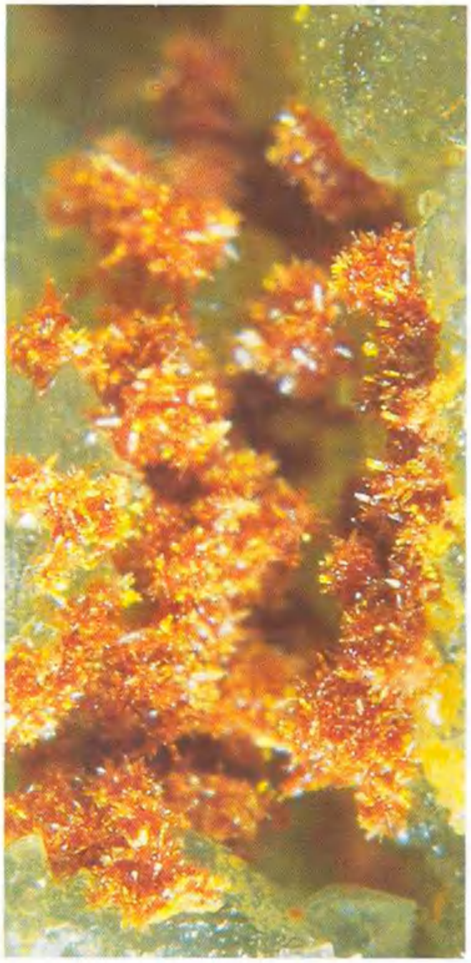

Carminite, 2 mm xx, Nadap, Hungary

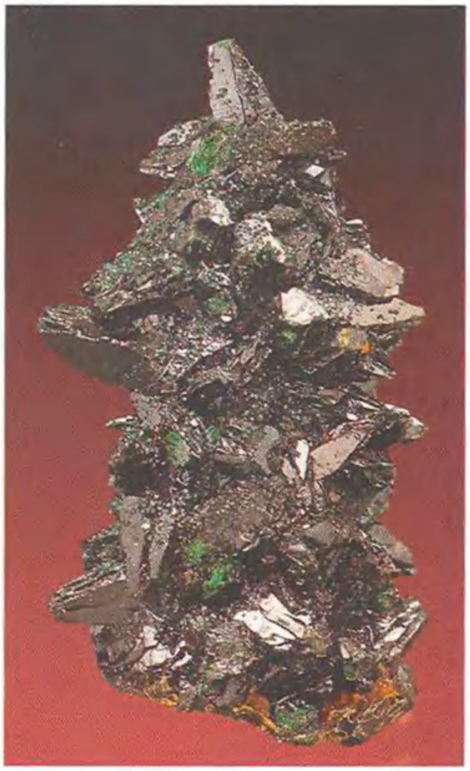

Bayldonite, 34 mm, Tsumeb, Namibia

Bayldonite

$PbCu_3(AsO_4)_2(OH)_2 . H_2O$

MONOCLINIC ● ●

Properties: C – yellow-green, olive-green; S – light green; L – resinous; D – translucent; DE – 5.5; H – 4.5; CL – not determined; F – uneven; M – earthy aggregates, coatings, massive.
Origin and occurrence: Secondary in the oxidation zone in hydrothermal Cu deposits and in greisens. It occurs as common yellow-green coatings and crystals up to 10 mm ($^3/_8$ in) across, associated with large azurite and mimetite crystals in Tsumeb, Namibia; also as coatings in St. Day, Cornwall, UK; in Moldava, Czech Republic.

Vesignieite

$BaCu_3(VO_4)_2(OH)_2$

MONOCLINIC ● ●

Properties: C – yellow-green, olive-green; S – light yellow-green; L – vitreous to dull; D – translucent;

Vesigneite, 80 mm, Vrancice, Czech Republic

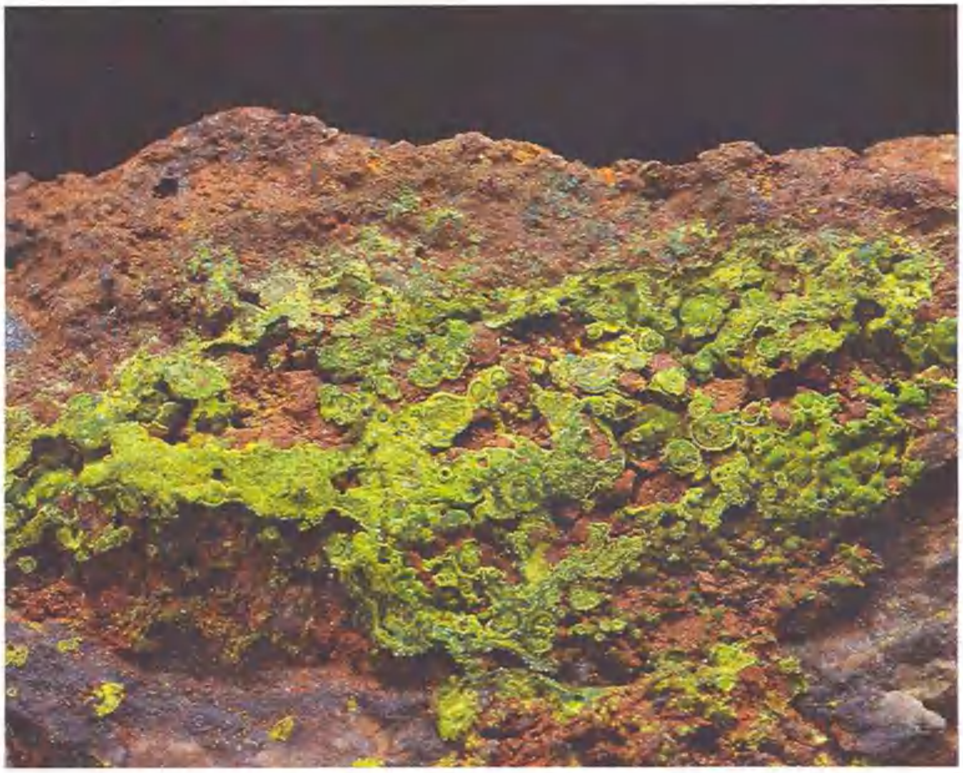

DE – 4.1; H – 3-4; CL – good; F – uneven; M – earthy and pulverulent aggregates and coatings, massive.

Origin and occurrence: Secondary in base metals deposits and in sediments, containing Cu sulfides. It occurs as yellow-green coatings in sediments in Horni Kalna, Czech Republic; also in the Mashamba West mine, Zaire; together with barite and psilomelane in Friedrichsroda, Germany.

Arsentsumebite

$Pb_2Cu(AsO_4)(SO_4)(OH)$

MONOCLINIC ●

Properties: C – bluish-green to light green; S – light green; L – dull; D – translucent; DE – 6.5; H – 3; CL – good; F – uneven; M – earthy aggregates, coatings, massive.

Origin and occurrence: Secondary in the oxidation zone, where it forms as a result of mimetite replacement. Found rarely in Moldava, Czech Republic; in Tsumeb, Namibia, where it forms pseudo-morphs after azurite crystals.

Arsentsumebite, 40 mm, Moldava, Czech Republic

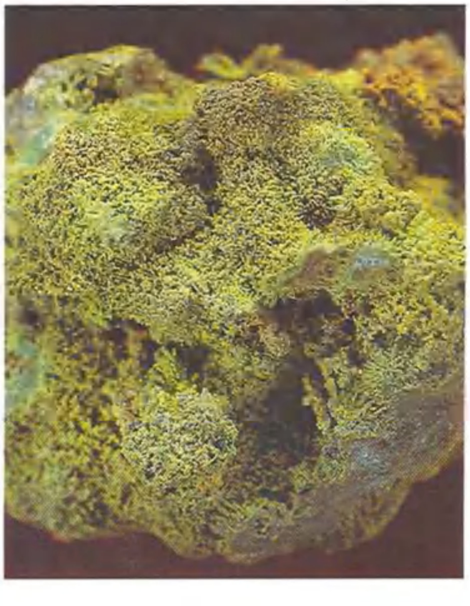

Fluorapatite, 50 mm, Dusso, Pakistan

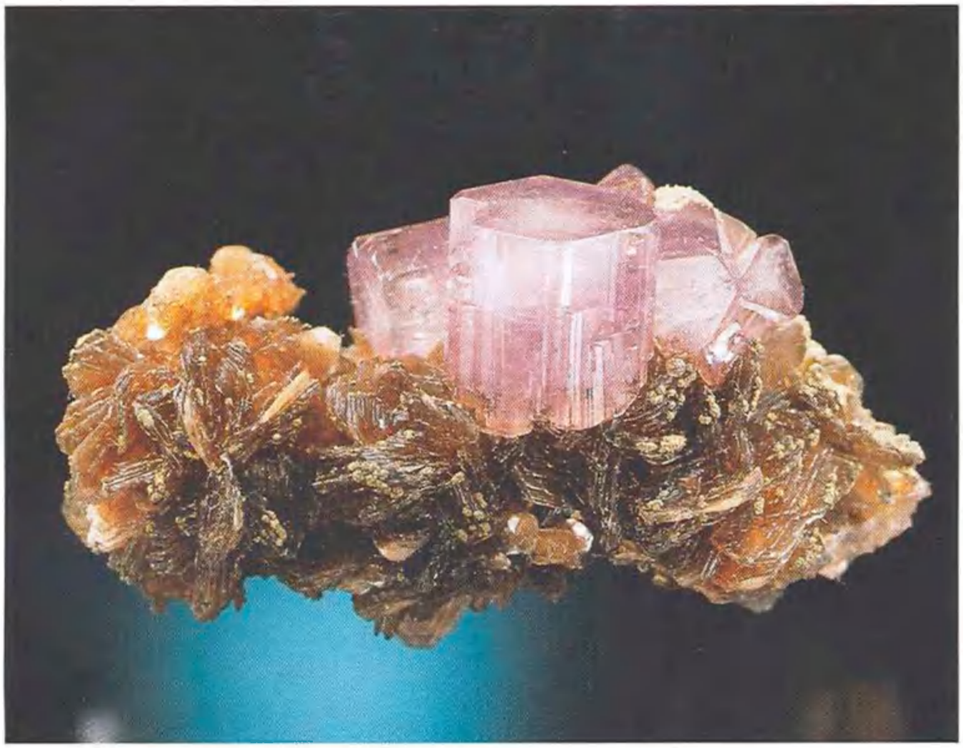

Fluorapatite

$Ca_5(PO_4)_3F$

HEXAGONAL ● ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish, blue, purple, green, brown with various hues; S – white; L – vitreous to dull; D – transparent to translucent, sometimes opaque; DE – 3.2; H – 5; CL – imperfect; F – conchoidal to uneven; M – long prismatic to tabular crystals, botryoidal, earthy and fibrous aggregates, massive; LU – yellow.
Origin and occurrence: Magmatic in granites, syenites, diorites, gabbros and various types of pegmatites, also in volcanic rocks; hydrothermal in quartz veins, ore veins, greisens and Alpine-type veins; metamorphic in different types of gneisses, migmatites, mica schists, skarns and amphibolites; in different types of sedimentary rocks. Perfect, short prismatic, purple transparent crystals up to 40 x 40 mm ($1\frac{9}{16}$ x $1\frac{9}{16}$ in) are renowned from pegmatite in the Pulsifer quarry, Mount Apatite, Auburn, Maine, USA; pink crystals from Dusso, Pakistan. Perfect crystals about 100 mm (4in) across also from Alpine-type veins, e.g. in Fiesch, Switzerland; also known from the quartz veins in greisens, associated with wolframite, in Panasqueira, Portugal; in Horni Slavkov, Czech Republic; Ehrenfriedersdorf, Germany. Also found in skarns at Cerro de Mercado, Durango, Mexico. Large deposits of mas-

Chlorapatite, 20 mm x, Bob's Lake, Ontario, Canada

sive apatite located in Kola Peninsula, Russia; crystals weighing up to 300 kg (660 lb) from the vicinity of Clear Lake, Ontario, Canada.
Application: main source of P, chemical industry, fertilizer.

Chlorapatite

$Ca_5(PO_4)_3Cl$

HEXAGONAL ● ● ●

Properties: C – white, various hues of yellow; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 5; CL – imperfect; F – conchoidal to uneven; M – long prismatic to tabular crystals, granular.
Origin and occurrence: Magmatic in nepheline syenites and their pegmatites, some gabbros and volcanic rocks, also in meteorites; metamorphic in skarns. It is usually associated with scapolite, amphibole, titanite and magnetite. Perfect prismatic crystals up to 35 cm (13$^{12}/_{16}$ in) long, come from pegmatites in Bamle, Norway; also from Kurokura, Japan.

Hydroxylapatite

$Ca_5(PO_4)_3(OH)$

HEXAGONAL ● ● ●

Properties: C – white, yellow, various hues of gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 5; CL – imperfect; F – conchoidal to uneven; M – short prismatic to tabular crystals, acicular aggregates, granular.
Origin and occurrence: Metamorphic in talc schists and serpentinites; hydrothermal in granitic pegmatites; sedimentary in organic remnants. Crystals up to 30 mm (1$^{3}/_{16}$ in) across known from Snarum, Norway; Hospental, Switzerland; Eagle, Colorado, USA.

Carbonate-fluorapatite

$Ca_5(PO_4, CO_3)_3F$

HEXAGONAL ● ● ●

Properties: C – white, gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 5; CL – imperfect; F – conchoidal to uneven; M B spherical and botryoidal aggregates, massive.
Origin and occurrence: Hydrothermal in ore veins and along the cracks in volcanic rocks. Rich botryoidal aggregates are known together with hyalite opal from Valec, Czech Republic; the Wheal Franco mine, Tavistock, Devon, UK.

Hydroxylapatite, 2 mm xx, Tornaszentandrás, Hungary

Carbonate-fluorapatite, 80 mm, Staffel, Germany

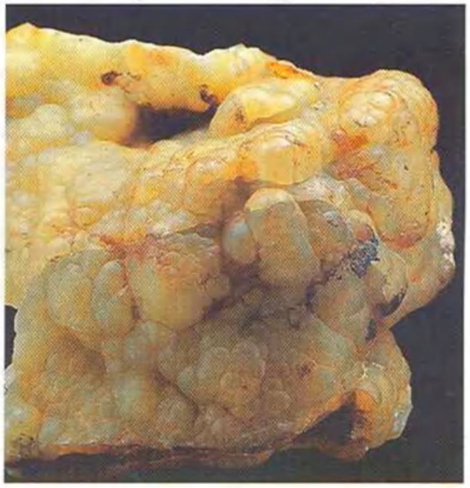

Pyromorphite, 41 mm, Les Farges, France

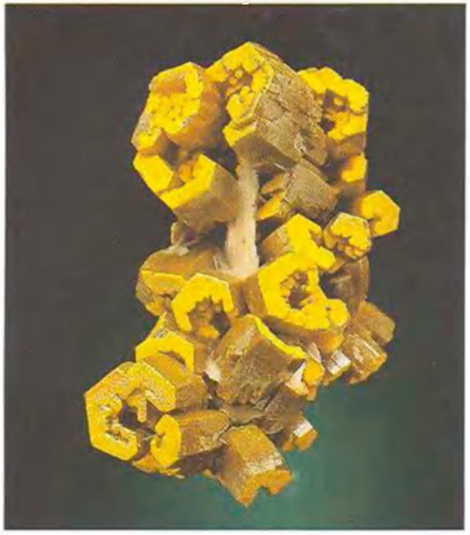

Pyromorfite

$Pb_5(PO_4)_3Cl$

HEXAGONAL ● ● ● ●

Properties: C – green, brown, yellow, orange, white, gray; S -white; L – adamantine to greasy; D – transparent to translucent; DE – 7.1; H – 3.5-4; CL – imperfect; F – conchoidal to uneven; M – long prismatic to tabular and pyramidal crystals, botryoidal aggregates with radial structure, acicular and earthy aggregates, massive.

Origin and occurrence: Secondary in the oxidation zone of Pb deposit, associated with cerussite, goethitee and other secondary minerals. Perfect crystals up to 40 mm (1$^{9}/_{16}$ in) long come from many localities, e.g. green and brown ones from Bad Ems and Zschopau, Germany; yellow-brown ones from Les Farges, France; green, brown and yellow ones from Mina Ojuela, Mapimi, Durango, Mexico. Green, orange and brown crystals, up to 60 mm (2$^{3}/_{8}$ in) long, are known from the Bunker Hill mine, Idaho, USA.

Pyromorphite, 62 mm, Les Farges, France

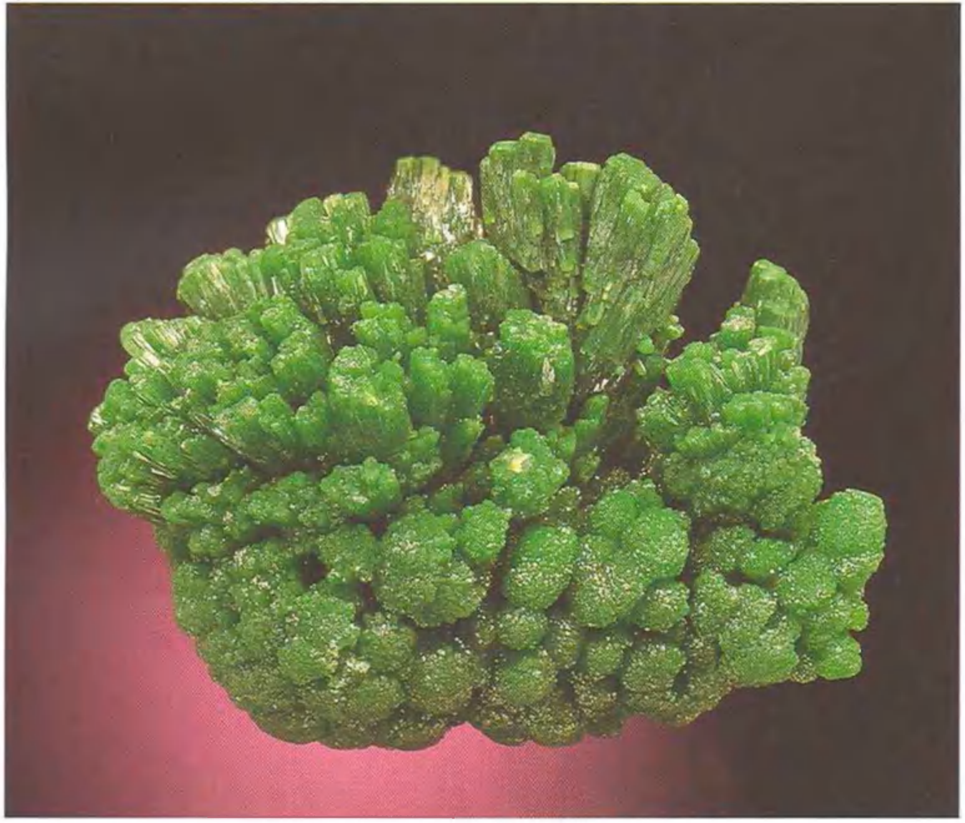

Mimetite, 14 mm x, Johanngeorgenstadt, Germany

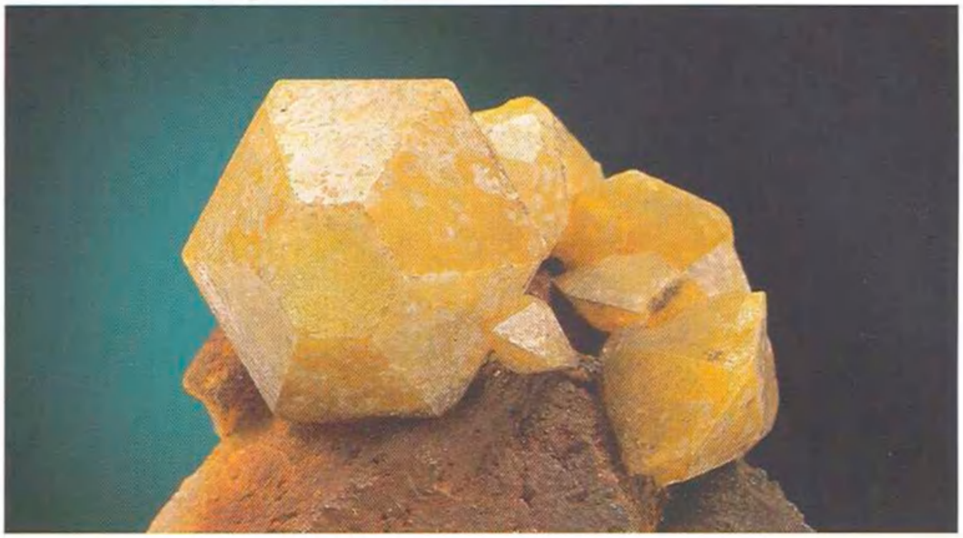

Mimetite

$Pb_5(AsO_4)_3Cl$

HEXAGONAL ● ● ●

Varieties: campylite

Properties: C – yellow, orange, brown, yellow-brown, white, gray; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 7.3; H – 3.5-4; CL – imperfect; F – conchoidal to uneven; M – long to short prismatic and pyramidal crystals, botryoidal aggregates with radial structure, acicular and earthy aggregates, granular.
Origin and occurrence: Secondary in Pb deposits, associated with pyromorphite and goethite. Perfect crystals up to 20 mm ($^{25}/_{32}$ in) long come from Johanngeorgenstadt, Germany; campylite occurs in Drygill, UK; crystals up to 50 mm (2 in) across found in Tsumeb, Namibia; Santa Eulalia and San Pedro, Chihuahua, Mexico. Beautiful yellow crystals up to 30 mm (1$^{3}/_{16}$ in) found recently in Hat Yai province, Thailand.

Vanadinite

$Pb_5(VO_4)_3Cl$

HEXAGONAL ● ● ●

Properties: C – yellow, orange, red, brown, yellow-brown; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 6.9; H – 2.5-3; CL – none; F – conchoidal to uneven; M – long to short prismatic and pyramidal crystals, botryoidal aggregates with radial structure, acicular and earthy aggregates, granular.
Origin and occurrence: Secondary in Pb deposits, associate with pyromorphite, wulfenite and goethite. Perfect crystals up to 130 mm (5$^{3}/_{16}$ in) long come from Djebel Mahseur and Mibladen, Morocco. Prismatic crystals were also found in Tsumeb, Namibia; in the Old Yuma, Red Cloud, Apache and Mammoth mines, Arizona, USA.

Vanadinite, 64 mm, Mibladen, Morocco

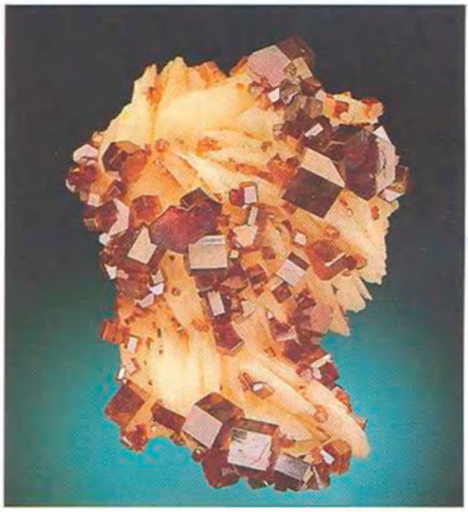

Atelestite, 70 mm, Schneeberg, Germany

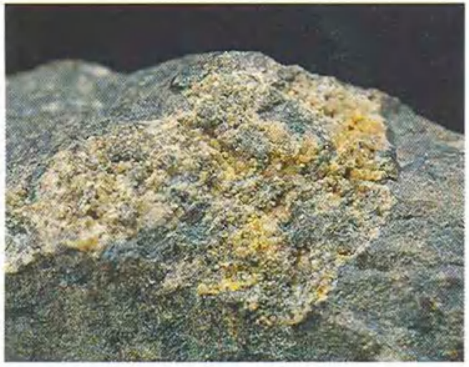

Atelestite

$Bi_2(AsO_4)O(OH)$

MONOCLINIC ●

Properties: C – yellow, yellow-green; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 7.0; H – 4.5-5; CL – imperfect; F – conchoidal to uneven; M – small tabular crystals, spherical aggregates.
Origin and occurrence: Secondary in the oxidation zone of Bi deposits. It was found in Schneeberg, Germany.

Hureaulite

$Mn_5(PO_4)_2[PO_3(OH)]_2 . 4 H_2O$

MONOCLINIC ● ● ●

Properties: C – orange-red, pink, purplish, white, gray; S – white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 3.5; CL – good; F – uneven; M – prismatic to tabular crystals, coatings, granular, massive.

Hureaulite, 53 mm, Shingus, Pakistan

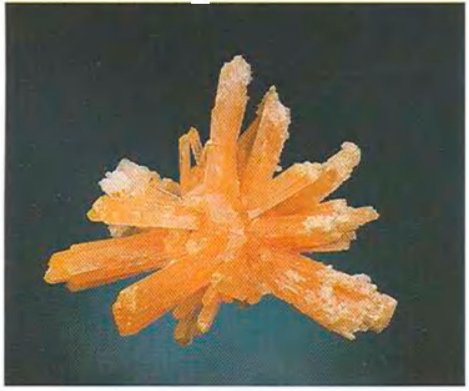

Variscite, 30 mm, Fairfield, U.S.A.

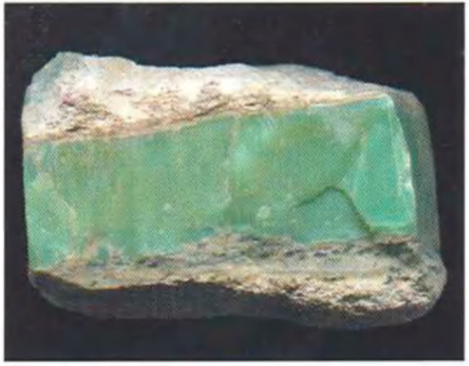

Origin and occurrence: Secondary in granitic pegmatites, as a result of alteration of the primary phosphates, mostly lithiophillite, triphyllite and graftonite. Purplish crystals up to 50 mm (2 in) across come from Shingus, Pakistan. Similar crystals found in Hagendorf, Germany. Also known from Mangualde, Portugal; Sao Jose da Safira, Minas Gerais, Brazil; from the Tip Top mine, Custer, South Dakota, USA.

Variscite

$Al(PO_4) . 2 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, greenish, blue-green; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 3.5-4.5; CL – good; F – uneven; M – isometric crystals, botryoidal aggregates, nodules, coatings, massive.
Origin and occurrence: : Hydrothermal in cracks in sedimentary rocks, rich in Al and P, also in phosphates deposits. It is associated with apatite, wavellite and other phosphates. Renowned greenish nodules up to 30 cm (12 in) in diameter come from Clay Canyon, Fairfield, Utah, USA; also known from Ronneburg, Germany and Jivina near Beroun, Czech Republic.

Strengite

$Fe^{3+}(PO_4) . 2 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, pinkish, red-purple; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 3.5-4.5; CL – good; F – conchoidal; M – isometric, tabular to short prismatic crystals, botryoidal aggregates with radial structure, coatings.
Origin and occurrence: Secondary in granitic

Strengite, 34 mm, Svappavaara, Sweden

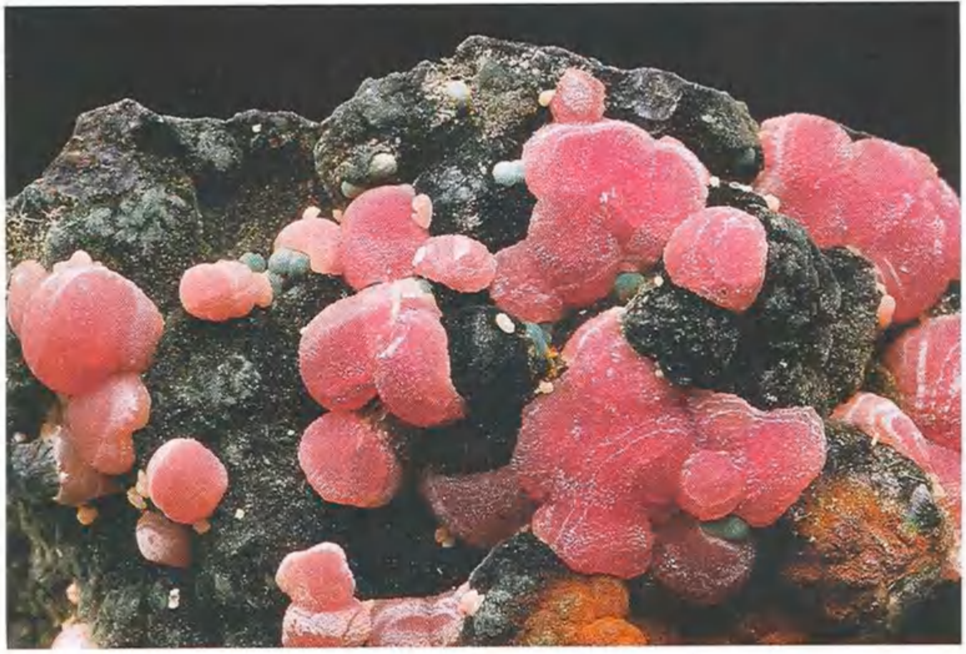

Scorodite, 10 mm x, Zacatecas, Mexico

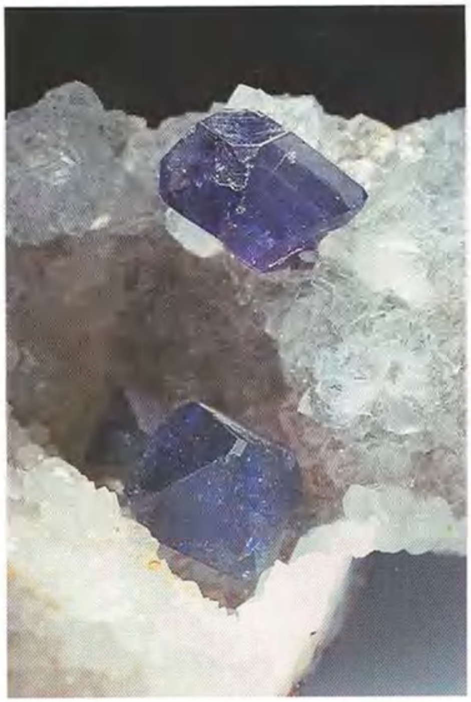

pegmatites where it forms as a result of hydrothermal replacement of primary phosphates; in the oxidation zone of Fe deposit together with goethite. Purple crystals up to 5 mm (3/16 in) across, come from the Bull Moose mine, Custer, South Dakota, USA; also known from Pleystein, Germany and from Teskov, Czech Republic.

Scorodite

$Fe^{3+}(AsO_4) . 2 H_2O$

ORTHORHOMBIC ● ● ●

Properties: B – light green, gray-green, olive-green, colorless, blue, yellow-brown; S – light green; L – vitreous to resinous; D – transparent to translucent; DE – 3.3; H – 3.5-4; CL – imperfect; F – conchoidal; M – dipyramidal to short prismatic crystals, botryoidal and earthy aggregates, coatings, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of ore deposits, associated with arsenopyrite, lollingite and other arsenides; in granitic pegmatites; hydrothermal in the hot springs.

Perfect crystals up to 50 mm (2 in) across come from Tsumeb, Namibia; the Kiura mine, Oita, Japan; Mina Ojuela, Mapimi, Durango, Mexico; green crusts, several cm thick, were found in Djebel Debar, Algeria.

Phosphophyllite, 18 mm x, Potosi, Bolivia

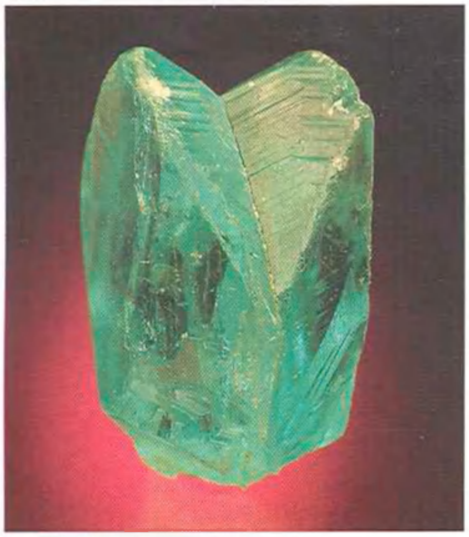

Phosphophyllite

$Zn_2Fe(PO_4)_2 . 4 H_2O$

MONOCLINIC ● ●

Properties: C – colorless, blue-green, blue; S – white; L – vitreous; D – transparent to translucent; DE – 3.1; H – 3-3.5; CL – good; F – uneven; M – long to short prismatic and thick tabular crystals, granular.

Origin and occurrence: Secondary in the oxidation zone of ore deposits; in granitic pegmatites, where it replaces primary phosphates. Perfect crystals, up to 140 mm ($5\frac{1}{2}$ in) across come from sulfide cavities in the Unificada mine, Potosi, Bolivia. Also occurs at Hagendorf, Germany.

Ludlamite

$Fe_3(PO_4)_2 . 4 H_2O$

MONOCLINIC ● ●

Properties: C – light green, green; S – white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 3.5; CL – perfect; F – uneven; M – thin to thick tabular crystals, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of ore deposits; in granitic pegmatites, where it replaces primary phosphates. Perfect crystals up to 90 mm ($3\frac{9}{16}$ in) across known from the San Antonio mine, Santa Eulalia, Chihuahua, Mexico; also from Morococala, Bolivia and the Blackbird district, Idaho, USA.

Ludlamite, 36 mm, Morococala, Bolivia

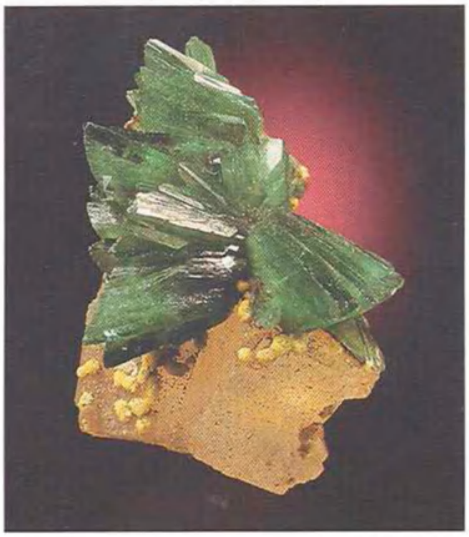

Anapaite

$Ca_2Fe(PO_4)_2 . 4 H_2O$

TRICLINIC ● ●

Properties: C – light to dark green, colorless; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 3.5; CL – perfect; F – uneven; M – thin to thick tabular crystals, rosette-like aggregates, granular.

Origin and occurrence: Secondary in the oxidation zone of Fe deposits and in sedimentary rocks, rich in P. Tabular crystals several mm across and spherical aggregates, up to 30 mm ($1\frac{3}{16}$ in) across, come from the cracks within oolitic ores near Anapa and Kerch, Crimea, Ukraine; also known from Bellaver de Cerdena, Spain.

Vivianite

$Fe_3(PO_4)_2 . 8 H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless when fresh, quickly oxidizes to blue, green, purple, black-blue; S – white to bluish; L – vitreous; D – transparent, translucent to opaque; DE – 2.7; H – 1.5-2; CL – perfect; F – uneven; M – long prismatic to acicular crystals, fibrous, earthy to pulverulent aggregates, coatings, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of Fe deposits; in granitic pegmatites, where it forms by the replacement of primary phosphates and in sedimentary rocks in proximity of organic

Anapaite, 18 mm aggregates, Anapa, Ukraine

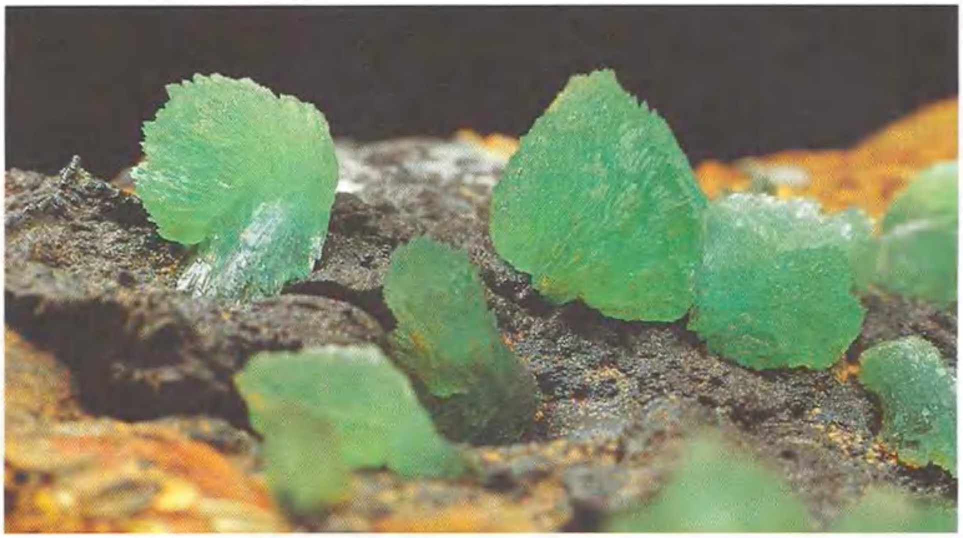

material; hydrothermal in the ore deposits. Crystals up to 1.5 m (5 ft) long found in clay sediments in Anloua, Cameroon; crystals up to 200 mm (7⅞ in) across known from Morococala, Bolivia. Smaller crystals come from Trepca, Serbia; Leadville, Colorado, Bingham Canyon, Utah, USA; also from Kerch and Anapa, Crimea, Ukraine.

Erythrine

$Co_3(AsO_4)_2 \cdot 8\,H_2O$

MONOCLINIC ● ● ●

Properties: C – dark purple, pink, colorless; S – light pink to white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 1.5-2.5; CL – perfect; F – uneven; M – long prismatic, acicular to tabular crystals, earthy aggregates, coatings, granular, massive.

Erythrite, 29 mm, Mexico

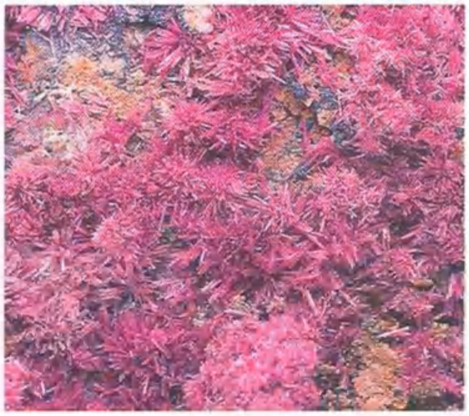

Origin and occurrence: Secondary in the oxidation zone of Co, Ni and U deposits. Tabular crystals up to 60 mm (2⅜ in) long come from Bou Azzer, Morocco. Other important localities are Schneeberg, Germany; Talmessi, Iran; Cobalt, Ontario, Canada; Mount Cobalt, Queensland, Australia.

Vivianite, 48 mm, Morococala, Bolivia

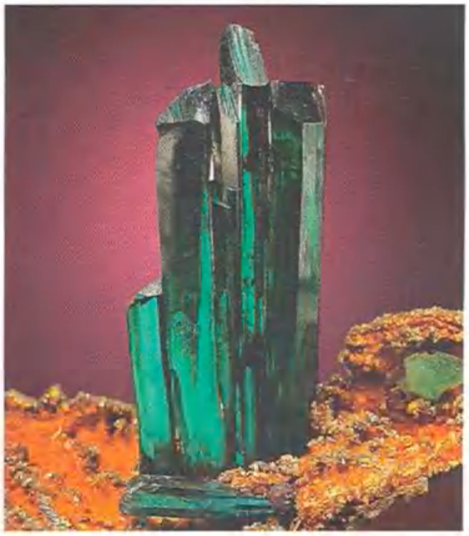

Phosphophyllite, 18 mm x, Potosi, Bolivia

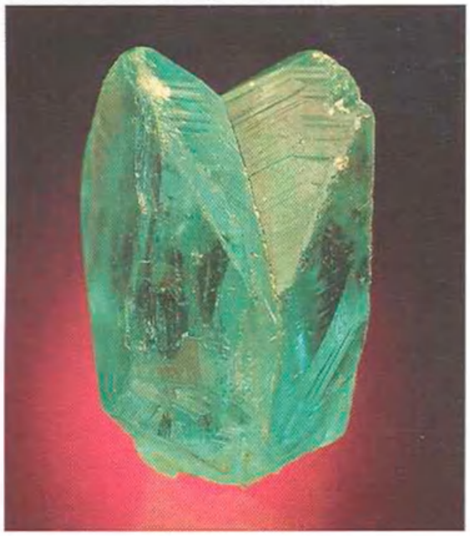

Phosphophyllite

$Zn_2Fe(PO_4)_2 \cdot 4\,H_2O$

MONOCLINIC ● ●

Properties: C – colorless, blue-green, blue; S – white; L – vitreous; D – transparent to translucent; DE – 3.1; H – 3-3.5; CL – good; F – uneven; M – long to short prismatic and thick tabular crystals, granular.

Origin and occurrence: Secondary in the oxidation zone of ore deposits; in granitic pegmatites, where it replaces primary phosphates. Perfect crystals, up to 140 mm ($5\frac{1}{2}$ in) across come from sulfide cavities in the Unificada mine, Potosi, Bolivia. Also occurs at Hagendorf, Germany.

Ludlamite

$Fe_3(PO_4)_2 \cdot 4\,H_2O$

MONOCLINIC ● ●

Properties: C – light green, green; S – white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 3.5; CL – perfect; F – uneven; M – thin to thick tabular crystals, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of ore deposits; in granitic pegmatites, where it replaces primary phosphates. Perfect crystals up to 90 mm ($3\frac{9}{16}$ in) across known from the San Antonio mine, Santa Eulalia, Chihuahua, Mexico; also from Morococala, Bolivia and the Blackbird district, Idaho, USA.

Ludlamite, 36 mm, Morococala, Bolivia

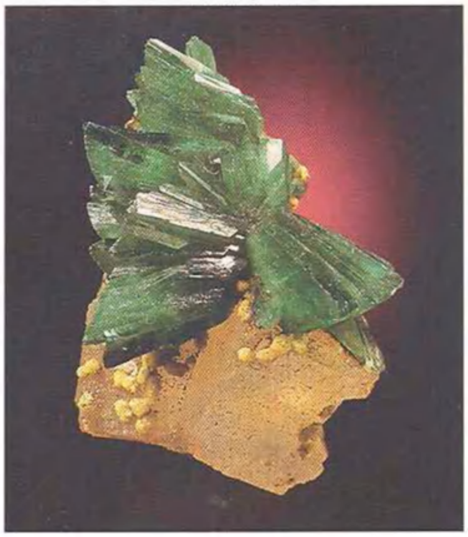

Anapaite

$Ca_2Fe(PO_4)_2 \cdot 4\,H_2O$

TRICLINIC ● ●

Properties: C – light to dark green, colorless; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 3.5; CL – perfect; F – uneven; M – thin to thick tabular crystals, rosette-like aggregates, granular.

Origin and occurrence: Secondary in the oxidation zone of Fe deposits and in sedimentary rocks, rich in P. Tabular crystals several mm across and spherical aggregates, up to 30 mm ($1\frac{3}{16}$ in) across, come from the cracks within oolitic ores near Anapa and Kerch, Crimea, Ukraine; also known from Bellaver de Cerdena, Spain.

Vivianite

$Fe_3(PO_4)_2 \cdot 8\,H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless when fresh, quickly oxidizes to blue, green, purple, black-blue; S – white to bluish; L – vitreous; D – transparent, translucent to opaque; DE – 2.7; H – 1.5-2; CL – perfect; F – uneven; M – long prismatic to acicular crystals, fibrous, earthy to pulverulent aggregates, coatings, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of Fe deposits; in granitic pegmatites, where it forms by the replacement of primary phosphates and in sedimentary rocks in proximity of organic

Anapaite, 18 mm aggregates, Anapa, Ukraine

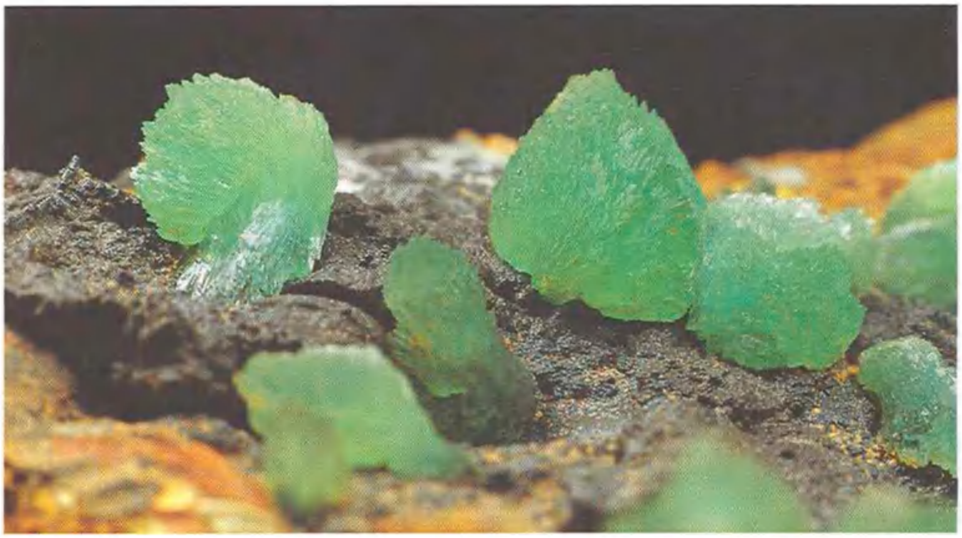

material; hydrothermal in the ore deposits. Crystals up to 1.5 m (5 ft) long found in clay sediments in Anloua, Cameroon; crystals up to 200 mm ($7\frac{7}{8}$ in) across known from Morococala, Bolivia. Smaller crystals come from Trepca, Serbia; Leadville, Colorado, Bingham Canyon, Utah, USA; also from Kerch and Anapa, Crimea, Ukraine.

Erythrine

$Co_3(AsO_4)_2 \cdot 8\ H_2O$

MONOCLINIC ● ● ●

Properties: C – dark purple, pink, colorless; S – light pink to white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 1.5-2.5; CL – perfect; F – uneven; M – long prismatic, acicular to tabular crystals, earthy aggregates, coatings, granular, massive.

Origin and occurrence: Secondary in the oxidation zone of Co, Ni and U deposits. Tabular crystals up to 60 mm ($2\frac{3}{8}$ in) long come from Bou Azzer, Morocco. Other important localities are Schneeberg, Germany; Talmessi, Iran; Cobalt, Ontario, Canada; Mount Cobalt, Queensland, Australia.

Erythrite, 29 mm, Mexico

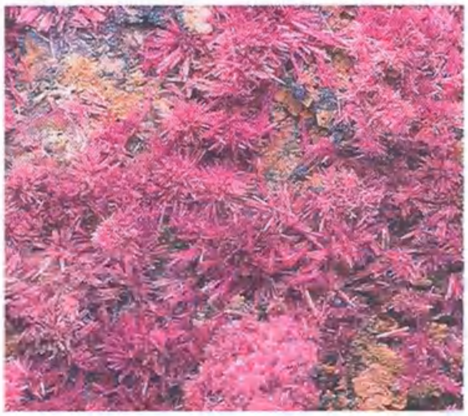

Vivianite, 48 mm, Morococala, Bolivia

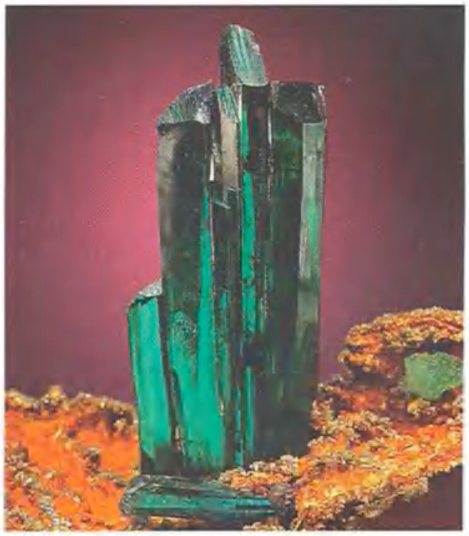

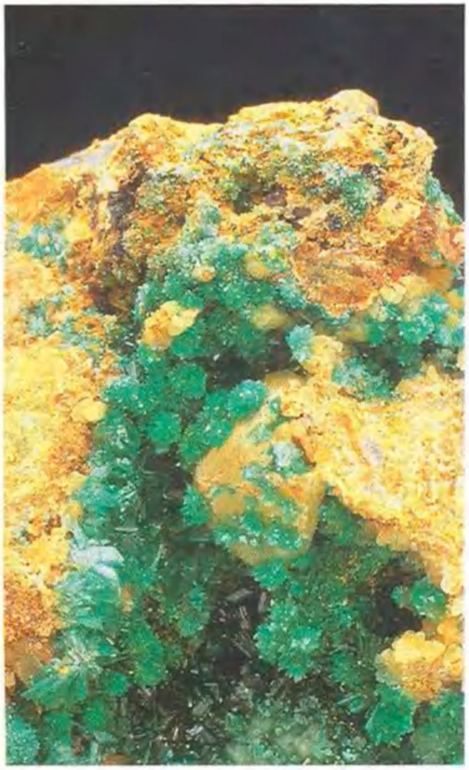

Annabergite, 55 mm, Laurion, Greece

Annabergite

$Ni_3(AsO_4)_2 \cdot 8 H_2O$

MONOCLINIC ● ● ●

Varieties: cabrerite (Mg contents)

Properties: C – light to dark green, white; S – white; L – vitreous; D – transparent to translucent; DE – 3.2; H – 1.5-2.5; CL – perfect; F – uneven; M – long prismatic to acicular and tabular crystals, acicular and earthy aggregates, coatings, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of Ni deposits, associated with erythrite. Crystals up to 5 mm ($^{3}/_{16}$ in) across occur in Gukuroren, Turkey; Sierra Cabrera, Spain; Laurion, Greece. Nodules up to 20 mm ($^{25}/_{32}$ in) come from the Snowbird mine, Montana, USA.

Symplesite

$Fe_3(AsO_4)_2 \cdot 8 H_2O$

TRICLINIC ● ●

Properties: C – bluish, dark blue, light green, black-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.0 H – 2.5; CL – perfect; F – uneven; M – acicular to tabular crystals, spherical aggregates with radial structure, earthy aggregates and coatings, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of ore deposits and in granitic pegmatites with

Symplesite, 12 mm xx, Baia Sprie, Romania

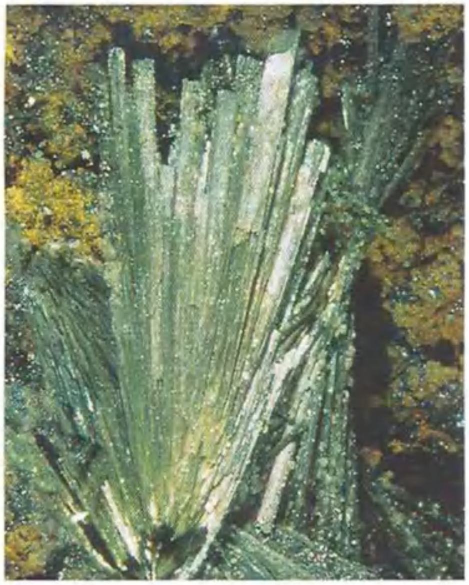

Picropharmacolite, 128 mm, France

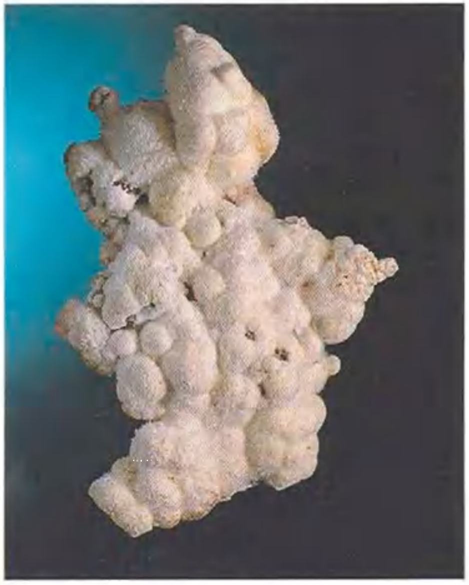

As minerals, mainly arsenopyrite and lollingite. It occurs in Schneeberg, Germany; Baia Sprie, Romania and Trebsko, Czech Republic.

Picropharmacolite

$Ca_4MgH_2(AsO_4)_4 . 12\ H_2O$

TRICLINIC ● ●

Properties: C – colorless, white; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 2.5; CL – perfect; F – uneven; M – acicular crystals, spherical aggregates with radial structure, coatings.
Origin and occurrence: Secondary in the oxidation zone of ore deposits with As minerals, mostly arsenopyrite and lollingite. It is also known from Ste-Marie-aux-Mines and Salsigne, France; Freiberg, Germany and Jachymov, Czech Republic.

Brushite

$CaH(PO_4) . 2\ H_2O$

MONOCLINIC ● ●

Properties: C – colorless, white; S – white; L – vitreous, pearly on cleavage planes; D – transparent to translucent; DE – 2.3; H – 2.5; CL – perfect; F – uneven; M – prismatic, acicular to tabular crystals, earthy aggregates and coatings, massive.
Origin and occurrence: Secondary on bat and bird excrements and bones, it impregnates bones, along the cracks of phosphorites. Tabular crystals up to 20 mm ($^{25}/_{32}$ in) across come from Quercy, France also occurs near Oran, Algeria and Pig Hole, Virginia, USA.

Brushite, 4 mm xx, Domica, Slovakia

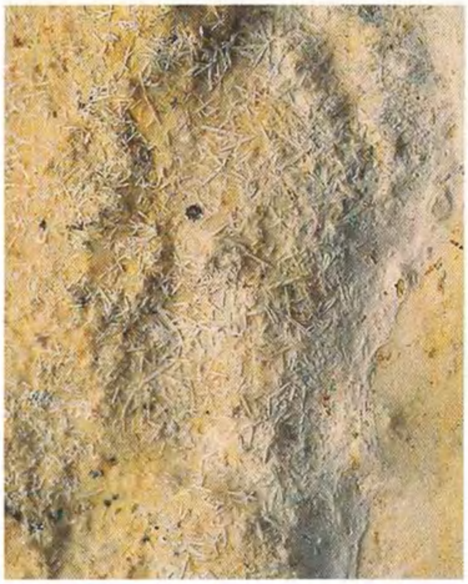

Legrandite, 25 mm xx, Mapimi, Mexico

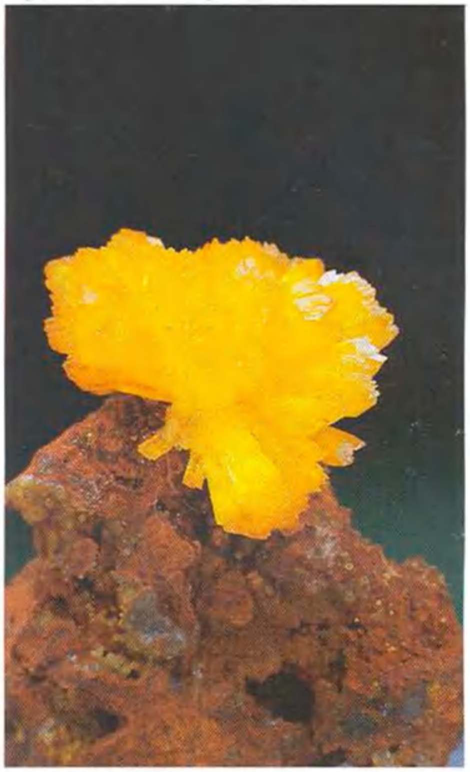

Legrandite

$Zn_2(AsO_4)(OH) . H_2O$

MONOCLINIC ● ●

Properties: C – colorless, yellow, purple; S – white; L – vitreous; D – transparent to translucent; DE – 4.0 H – 4.5; CL – imperfect; F – uneven; M – long prismatic crystals and their inter-growths, radial aggregates.
Origin and occurrence: Secondary in the oxidation zone of Zn deposits and in pegmatites. Prismatic crystals up to 250 mm ($9^{13}/_{16}$ in) long found in Mina Ojuela, Mapimi, Durango, Mexico. It is also known from Galileia, Minas Gerais, Brazil and Tsumeb, Namibia.

Euchroite, 10 mm xx, L'ubietova, Slovakia

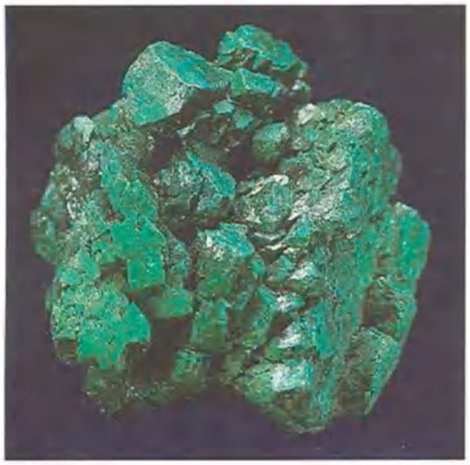

Euchroite

$Cu_2(AsO_4)(OH) . 3 H_2O$

ORTHORHOMBIC ● ●

Properties: C – emerald-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.5; H – 3.5-4; CL – imperfect; F – conchoidal to uneven; M – short prismatic to thick tabular crystals, massive.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, associated with olivenite and malachite. Thick tabular crystals up to 30 mm ($1^{3}/_{16}$ in) come from L'ubietova, Slovakia. It also occurs in Zapacica, Bulgaria and Chessy, France.

Vauxite, 20 mm xx, Llallagua, Bolivia

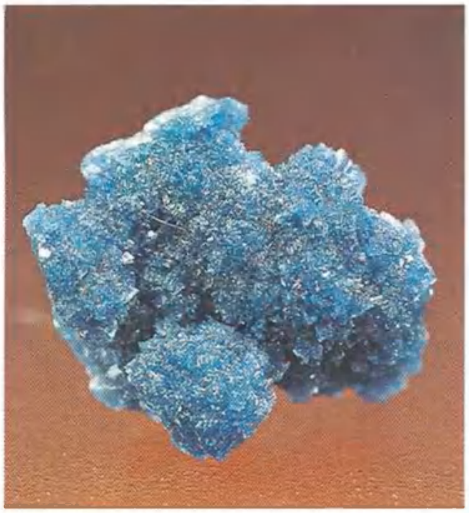

Strunzite, 10 mm xx, Hagendorf, Germany

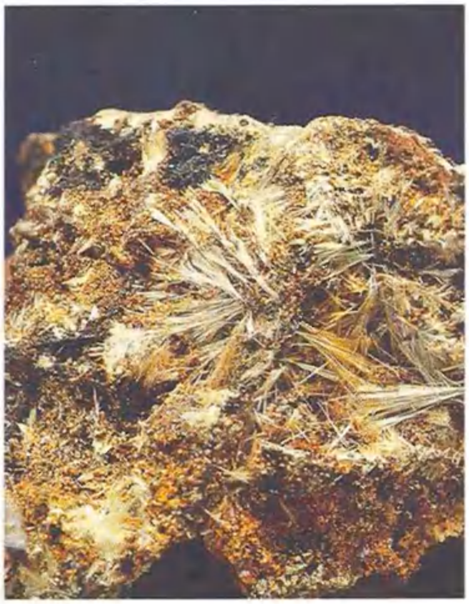

Vauxite

$FeAl_2(PO_4)_2(OH)_2 . 6 H_2O$

TRICLINIC ● ●

Properties: C – light to dark blue; S – white; L – vitreous; D – transparent to translucent; DE – 2.4; H – 3.5; CL – none; F – uneven; M – tabular crystals, radial aggregates, massive.
Origin and occurrence: Secondary in the oxidation zone of Sn deposits, associated with wavellite. Crystals occur in the Siglo XX Mine, Llallagua, Bolivia.

Strunzite

$MnFe^{3+}{}_2(PO_4)_2(OH)_2 . 6 H_2O$

MONOCLINIC ● ● ●

Properties: C – light yellow, yellow-brown; S – yellowish; L – vitreous; D – transparent to translucent; DE – 2.5; H – not determined; CL – not determined; F – uneven; M – acicular crystals, acicular and fibrous aggregates.
Origin and occurrence: Secondary in granitic pegmatites, as a result of weathering of primary phosphates, mostly triphyllite; rarely hydrothermal in the cracks of Fe-rich sedimentary rocks. Acicular crystals up to 20 mm ($^{25}/_{32}$ in) across known from Hagendorf, Germany; the Palermo No.1 and No. 2 mines, North Groton and the Fletcher mine, Groton, New Hampshire, USA.

Cacoxenite, 1 mm xx, Hellertown, Pennsylvania, U.S.A.

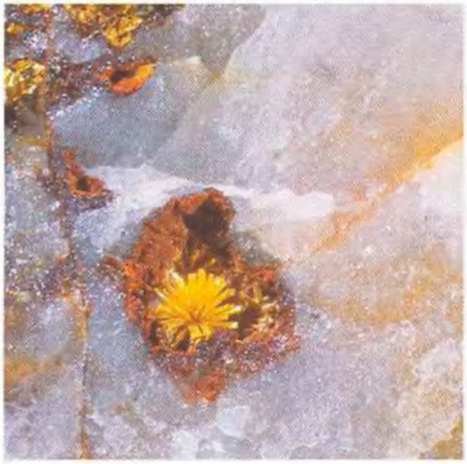

Cacoxenite

$(Fe^{3+},Al)_{25}(PO_4)_{17}O_6(OH)_{12} \cdot 75\ H_2O$

HEXAGONAL ● ● ●

Properties: C – light yellow, yellow-brown, orange; S – yellow; L – silky; D – transparent to translucent; DE – 2.3; H – 3-4; CL – not determined; F – uneven; M – acicular and fibrous aggregates, often with radial structure, botryoidal crusts and coatings.
Origin and occurrence: Hydrothermal on the cracks of sedimentary Fe ores, associated with wavellite; rare as secondary in granitic pegmatites, as a product of weathering of primary phosphates. Crystals up to 10 mm (⅜ in) across come from the Horcajo mine, Ciudad Real, Spain; golden-yellow acicular aggregates occur in Hrbek near Svata Dobrotiva and Trenice, Czech Republic and in Amberg, Germany.

Beraunite

$Fe^{2+}Fe^{3+}_5(PO_4)_4(OH)_5 \cdot 4\ H_2O$

MONOCLINIC ● ●

Properties: C – red-brown, red, gray-green; S – yellow to green-brown; L – vitreous to dull; D – translucent; DE – 3.0; H – 3.5-4; CL – good; F – uneven; M – acicular aggregates, often with radial structure, botryoidal crusts and coatings.
Origin and occurrence: Hydrothermal on the cracks of sedimentary Fe ores, typically together with wavellite; secondary in granitic pegmatites, as a product of weathering of primary phosphates. Acicular aggregates up to 10 mm (⅜ in) across occur in Mount Indian, Alabama, USA; also known from Hrbek near Svata Dobrotiva, Czech Republic and Amberg, Germany.

Beraunite, 3 mm aggregates, Nekeszeny, Hungary

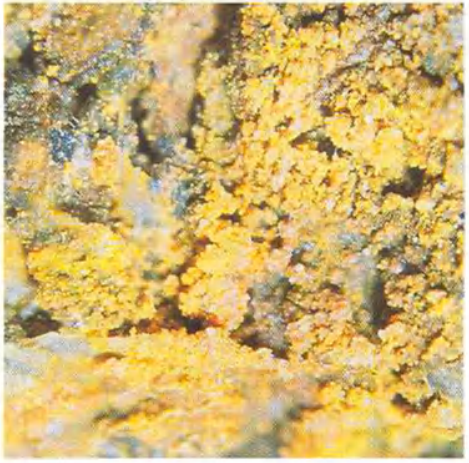

Diadochite

$Fe^{3+}_2(PO_4)(SO_4)(OH) \cdot 5\ H_2O$

TRICLINIC ● ● ●

Properties: C – yellow-brown, brown, red-brown, yellow-green, gray-green; S – yellow to light brown; L – dull, waxy; D – translucent to opaque; DE – 2.0-2.4; H – 3; CL – not determined; F – uneven, conchoidal, earthy; M – nodules, coatings and crusts, massive.
Origin and occurrence: Secondary in the oxidation zone of Fe deposits. Diadochite caves in abandoned mines are known from Saalfeld, Germany; nodules found in New Idria, California and Eureka, Nevada, USA.

Diadochite, 40 mm, Recsk, Hungary

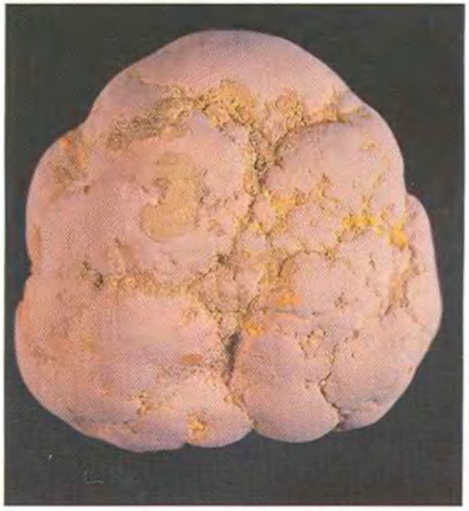

Wavellite, 40 mm, Trenice, Czech Republic

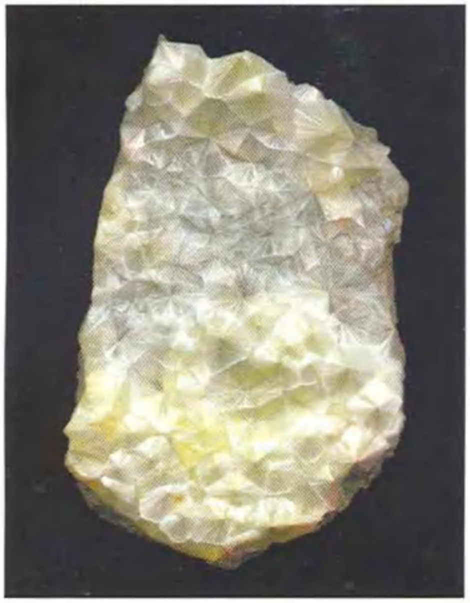

Eosphorite, 4 mm xx, Lavra da Ilha, Brazil

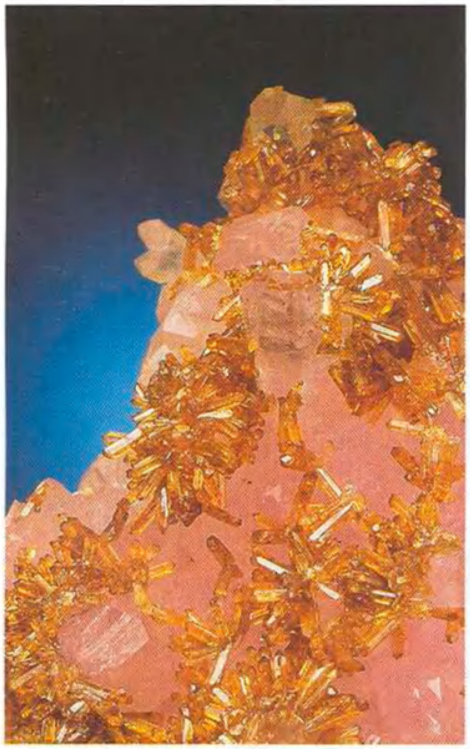

Wavellite

$Al_3(PO_4)_2(OH,F)_3 . 5 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, greenish, light blue-green, green, yellowish; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.4; H – 3.5-4; CL – perfect; F – uneven; M – isometric crystals, hemispherical aggregates with radial structure, botryoidal aggregates, nodules, coatings, massive.
Origin and occurrence: Hydrothermal in the cracks of Al and P-rich sediments, also in phosphate deposits, ore veins and pegmatites. Beautiful hemispherical aggregates up to 40 mm ($1\frac{9}{16}$ in) in diameter occur in Pencil, Garland and Magnet Cove, Arkansas, USA; also known from Trenice and Milina, Czech Republic and Ronneburg, Germany.

Eosphorite

$MnAl(PO_4)(OH)_2 . H_2O$

ORTHORHOMBIC ● ●

Properties: C – pinkish, colorless, white, brownish, red-brown; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 3.1; H – 5; CL – imperfect; F – uneven to conchoidal; M – long to short prismatic crystals, radial aggregates, granular.
Origin and occurrence: Secondary in granitic pegmatites, as a product of hydrothermal replacement of primary phosphates. Crystals up to 100 mm (4in) long found in the Joao Modesto dos Santos mine, Minas Gerais, Brazil. It occurs in Rapid Creek, Yukon, Canada, too.

Turquoise

$CuAl_6(PO_4)_4(OH)_8 . 4 H_2O$

TRICLINIC ● ● ●

Properties: C – blue, blue-green, green; S – white; L – waxy; D – transparent, translucent to opaque; DE – 2.9; H – 5-6; CL – good; F – conchoidal to uneven; M – short prismatic crystals, botryoidal aggregates, coatings, massive.
Origin and occurrence: Secondary in the surface parts of rocks with elevated contents of P and Cu, e.g. in the oxidation zone of some Cu deposits. Small crystals occurred near Lynch Station, Virginia, USA. Massive blue and blue-green concretions come from Mount Ali Mirsai near Maden, Iran. Other localities are Cortez, Nevada, Los Cerillos and Eureka, New Mexico and Bisbee, Arizona, USA.
Application: popular gemstone.

Turquoise, 3 mm aggregates, Humboldt Co., U.S.A.

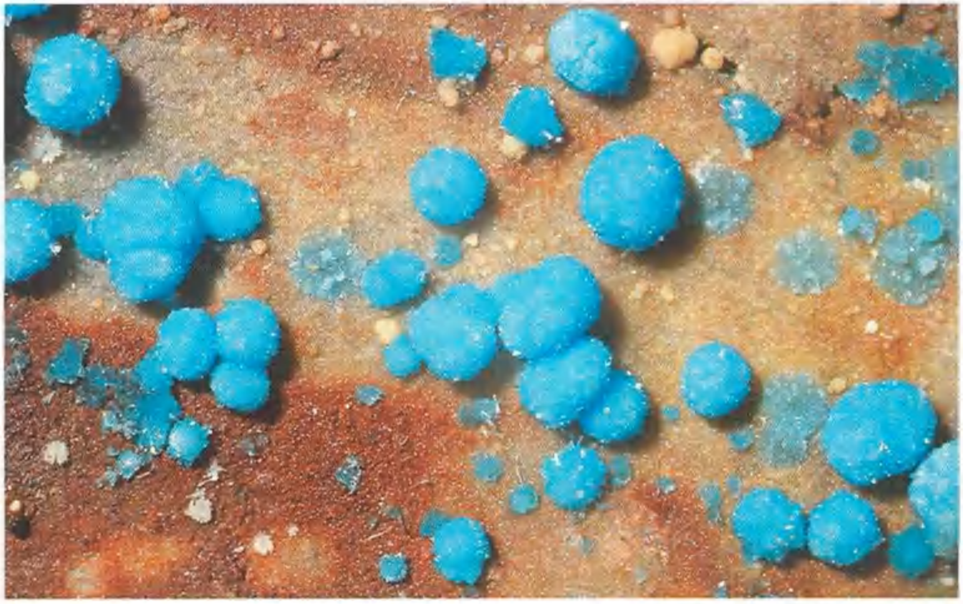

Chalcosiderite

$CuFe^{3+}_6(PO_4)_4(OH)_8 \cdot 4\,H_2O$

TRICLINIC ● ●

Properties: C – dark green; S – white; L – vitreous; D – transparent to translucent; DE – 3.3; H – 4.5; CL – good; F – conchoidal to uneven; M – short prismatic crystals, coatings.

Origin and occurrence: Secondary in the oxidation zone of some Cu deposits, together with goethite, dufrenite and pharmacosiderite. It occurs in Bisbee, Arizona, USA; in the Wheal Phoenix mine, Cornwall, UK; Schneckenstein, Germany; Horni Slavkov, Czech Republic.

Turquoise, 50 mm, Kazakhstan

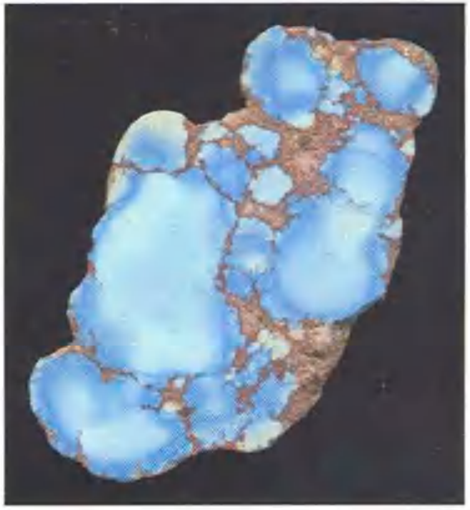

Chalcosiderite, 90 mm, Cornwall, UK

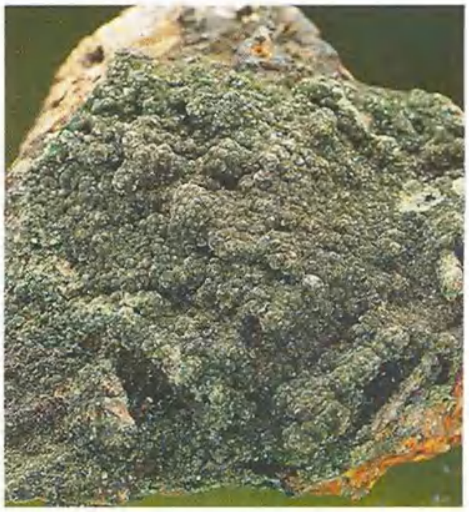

Chenevixite, 100 mm, Chuquicamata, Chile

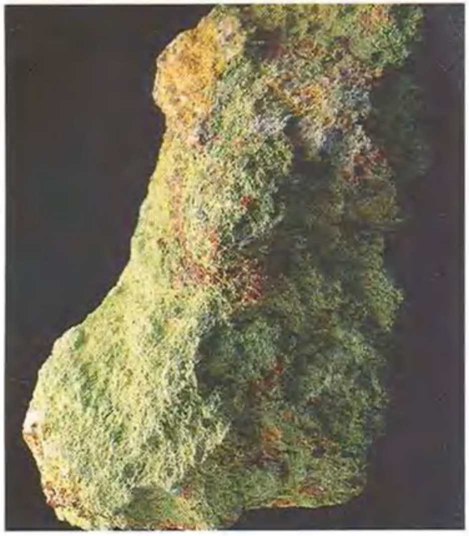

Chenevixite
$Cu_2Fe^{3+}{}_2(AsO_4)_2(OH)_4 \cdot H_2O$

MONOCLINIC ●

Properties: C – dark green, olive-green to yellow-green; S – yellow-green; L – greasy; D – translucent; DE – 3.9; H – 3.5-4.5; CL – not determined; F – conchoidal to uneven; M – earthy aggregates, coatings, massive.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, associated with malachite, tyrolite, azurite and other minerals. Massive aggregates occur in the Mammoth mine, Tintic, Utah, USA; also in Klein Spitzkopje, Namibia.

Tyrolite
$CaCu_5(AsO_4)_2(CO_3)(OH)_4 \cdot 6\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – apple-green, green-blue to blue; S – light green to blue-green; L – vitreous to pearly; D – transparent to translucent; DE – 3.3; H – 2; CL – perfect; F – uneven; M B scaly and fan-shaped aggregates, coatings and crusts.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, frequently associated with chalcophyllite. Rich aggregates occur in the Majuba Hill mine, Nevada, also in Tintic, Utah, USA. It is also known from Brixlegg, Austria; Saalfeld and Schneeberg, Germany; Novoveska Huta, Slovakia.

Tyrolite, 90 mm, Novoveska Huta, Slovakia

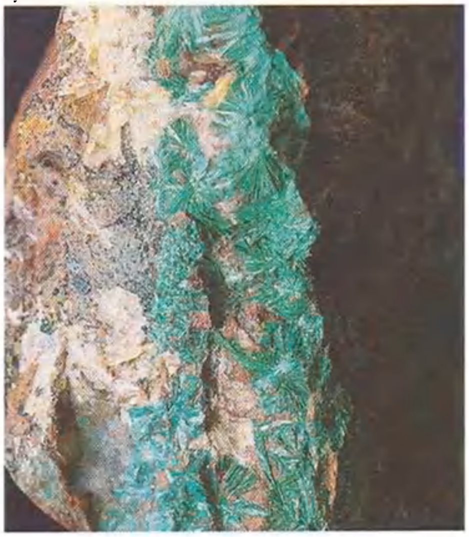

Delvauxite
$CaFe^{3+}{}_4(PO_4,SO_4)(OH)_8 \cdot 4\text{-}6\ H_2O$

AMORPHOUS ● ● ●

Properties: C – yellow-brown, brown, red-brown, black-brown; S – yellow; L – greasy, waxy; D – translucent to opaque; DE – 1.8-2.0; H – 2.5; CL – not determined; F – conchoidal, earthy; M – nodules, stalactites, coatings and crusts, massive.
Origin and occurrence: Secondary in oxidation zone of Fe deposits. Nodules of 50 cm (20 in) across in Czech Republic. Also known in Berneau and Richelle, Belgium; Železník, Slovakia; Kerch, Crimea, Ukraine.

Delvauxite, 80 mm, Nucice, Czech Republic

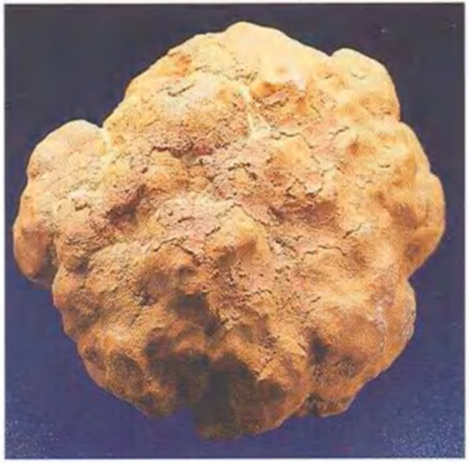

Bukovskyite, 40 mm, Kaňk, Czech Republic

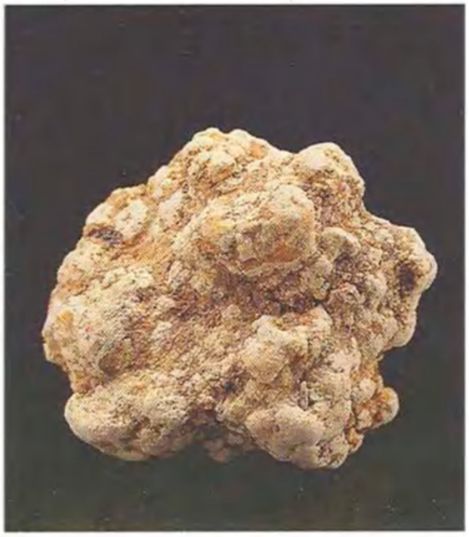

Bukovskyite

$Fe^{3+}{}_2(AsO_4)(SO_4)(OH) . 7 H_2O$

TRICLINIC ●

Properties: C – yellow-green, gray-green; S – yellowish white; L – dull to earthy; D – translucent to opaque; DE – 2.3; H – not determined; CL – not determined; F – earthy; M B botryoidal aggregates and nodules.
Origin and occurrence: Secondary on the old mine dumps where it forms as a product of arsenopyrite weathering. Nodules up to 60 cm (24 in), occur on medieval dumps in Kank near Kutna Hora, Czech Republic.

Chalcophyllite, 36 mm, Chile

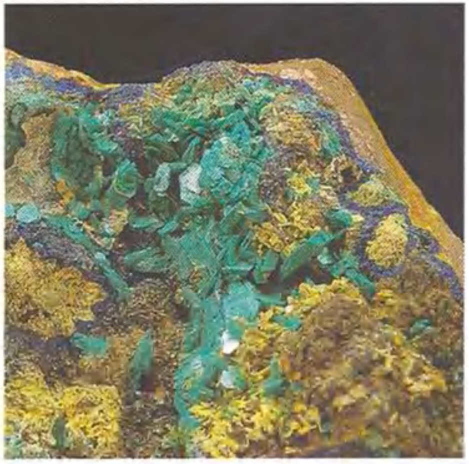

Veszelyite

$(Cu,Zn)_3(PO_4)(OH)_3 . 2 H_2O$

MONOCLINIC ● ●

Properties: C – green, blue-green, dark blue; S – green; L – vitreous; D – translucent; DE – 3.4; H – 3.5-4; CL – good; F – uneven; M – short prismatic to tabular crystals, granular.
Origin and occurrence: Secondary in the oxidation zone of Cu-Zn deposits. Crystals up to 50 mm (2 in) across found in the Black Pine mine, Philipsburg, Montana, USA; also from Moravita, Romania; Arakawa, Japan; Wanlockhead, Scotland, UK.

Chalcophylite

$Cu_{18}Al_2(AsO_4)_3(SO_4)_3(OH)_{27} . 36 H_2O$

TRIGONAL ● ● ●

Properties: C – emerald-green, blue-green; S – light green; L – vitreous, pearly; D – transparent to translucent; DE – 2.6; H – 2; CL – perfect; F – uneven; M – tabular crystals, scaly, fan-shaped aggregates, coatings, massive.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, usually associated with tyrolite. Rich aggregates occur in the Majuba Hill, Nevada and in the Tintic district, Utah, USA. Nice specimens also come from Novoveska Huta and Piesky, Slovakia; Nizhniy Tagil, Ural mountains, Russia; Cap Garrone, France.

Veszelyite, 13 mm, Philipsburg, U.S.A.

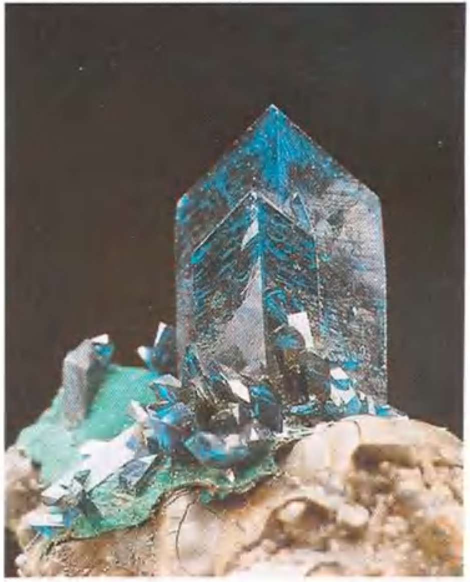

Liroconite, 86 mm, Cornwall, UK

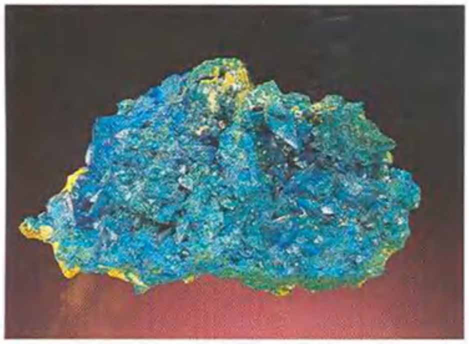

Lirokonit

$Cu_2Al(AsO_4)(OH)_4 . 4 H_2O$

MONOCLINIC ● ●

Properties: C – blue, green; S – light blue; L – vitreous to resinous; D – transparent to translucent; DE – 3.0; H – 2-2.5; CL – imperfect; F – conchoidal to uneven; M – lenticular, dipyramidal crystals, massive.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, together with olivenite, malachite and azurite. Perfect crystals up to 30 mm (1 3/16 in) across come from Redruth and St. Day, Cornwall, UK. It is also known from Cerro Gordo, California, USA.

Evansite, 5 mm aggregates, Sirk – Zelezník, Slovakia

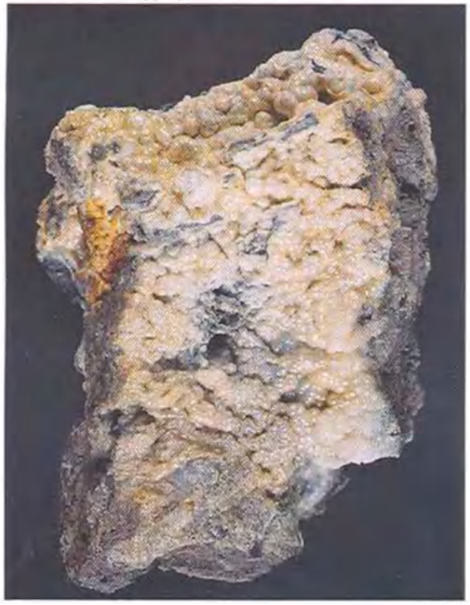

Evansite

$Al_3(PO_4)(OH)_6 . 6 H_2O$

AMORPHOUS 3

Properties: C – colorless, white, greenish, light blue-green, yellowish; S – white; L – vitreous, resinous, waxy; D – transparent to translucent; DE – 1.8-2.2; H – 3-4; CL – none; F – conchoidal; M – botryoidal, stalactitic and hemispherical aggregates and coatings.
Origin and occurrence: Secondary in the oxidation zone of Fe deposits, rich in P, associated with alophane and goethite. Rich stalactitic aggregates come from Zelezník, Slovakia and from Epernay, France.

Whiteite-(CaFeMg)

$CaFeMg_2Al_2(PO_4)_4(OH)_2 . 8 H_2O$

MONOCLINIC ● ●

Properties: C – brown; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 4; CL – good; F – uneven; M – short prismatic crystals.
Origin and occurrence: Hydrothermal in the cracks of P-rich sediments. Perfect crystals up to 20 mm (25/32 in) across come from Big Fish River, Yukon, Canada. Crystals up to 5 mm (3/16 in), are known from Lavra da Ilha de Taquaral, Minas Gerais, Brazil.

Whiteite-(CaFeMg), 65 mm, Yukon Territory, Canada

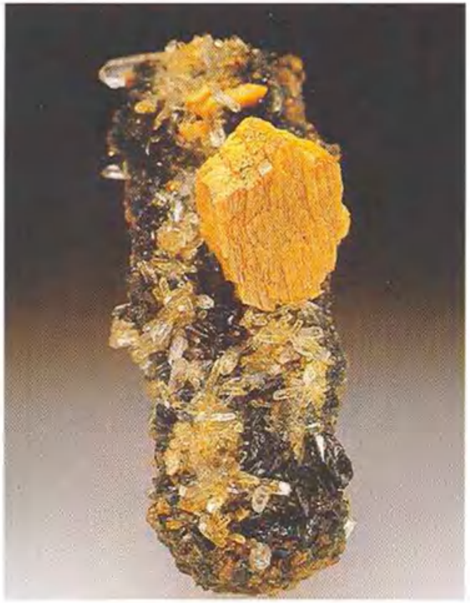

Jahnsite-(CaMnMg), 60 mm, Custer, U.S.A.

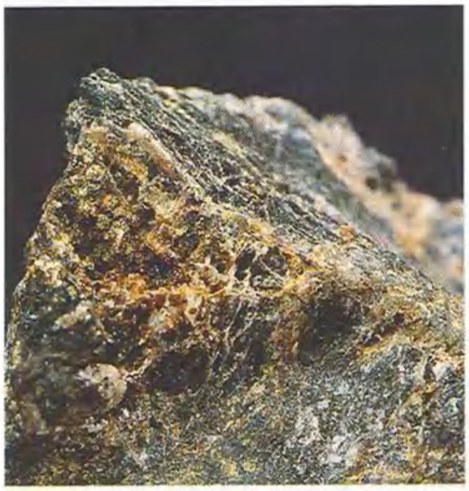

Jahnsite-(CaMnMg)

$CaMnMg_2Fe^{3+}_2(PO_4)_4(OH)_4 \cdot 8\,H_2O$

MONOCLINIC ● ●

Properties: C – yellow, light to dark brown; S – light yellow; L – vitreous; D – transparent to translucent; DE – 2.6; H – not determined; CL – good; F – uneven; M – short to long prismatic crystals, granular.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a result of replacement of primary phosphates. Perfect crystals up to 10 mm (⅜ in) across occur in the Tip Top mine, Custer, South Dakota, USA. It is also known from Hagendorf, Germany.

Wardite

$NaAl_3(PO_4)_2(OH)_4 \cdot 2\,H_2O$

TETRAGONAL ● ● ●

Properties: C – colorless, white, greenish, light blue-green; S -white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 5; CL – perfect; F – uneven; M – dipyramidal crystals, radial and hemispherical aggregates, coatings, crusts, granular.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a product of primary phosphate replacement; hydrothermal in the cracks in P-rich sediments. Perfect crystals up to 30 mm (1⅛ in) across come from Rapid Creek, Yukon, Canada. It is also known from Piedras Lavradas, Paraiba, Brazil.

Wardite, 48 mm, Yukon Territory, Canada

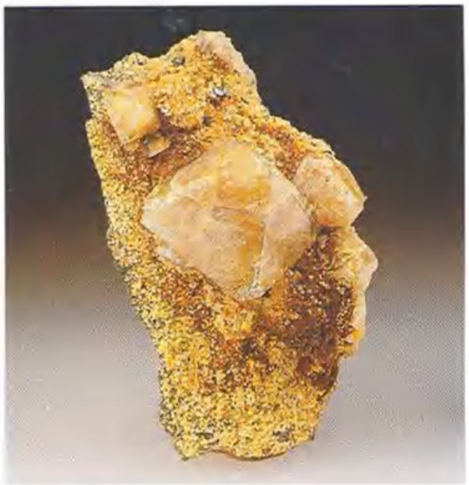

Cyrilovite

$NaFe^{3+}_3(PO_4)_2(OH)_4 \cdot 2\,H_2O$

TETRAGONAL ● ●

Properties: C – yellow, orange, brown-yellow; S – yellow; L – vitreous; D – transparent to translucent; DE – 3.1; H – 4; CL – good; F – conchoidal; M – tabular and dipyramidal crystals, coatings and crusts.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a result of replacement of primary phosphates. Small crystals occur in Cyrilov, Czech Republic; Hagendorf, Germany and Sapucaia, Minas Gerais, Brazil.

Cyrilovite, 60 mm, Iron Monarch, Australia

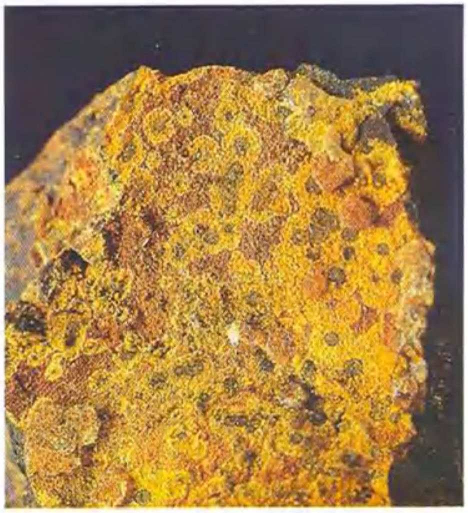

Liroconite, 86 mm, Cornwall, UK

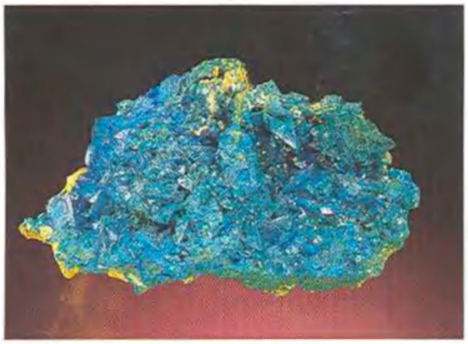

Lirokonit

$Cu_2Al(AsO_4)(OH)_4 . 4 H_2O$

MONOCLINIC ● ●

Properties: C – blue, green; S – light blue; L – vitreous to resinous; D – transparent to translucent; DE – 3.0, H – 2-2.5; CL – imperfect; F – conchoidal to uneven; M – lenticular, dipyramidal crystals, massive.

Origin and occurrence: Secondary in the oxidation zone of Cu deposits, together with olivenite, malachite and azurite. Perfect crystals up to 30 mm (1$^{3}/_{16}$ in) across come from Redruth and St. Day, Cornwall, UK. It is also known from Cerro Gordo, California, USA.

Evansite, 5 mm aggregates, Sirk – Zeleznik, Slovakia

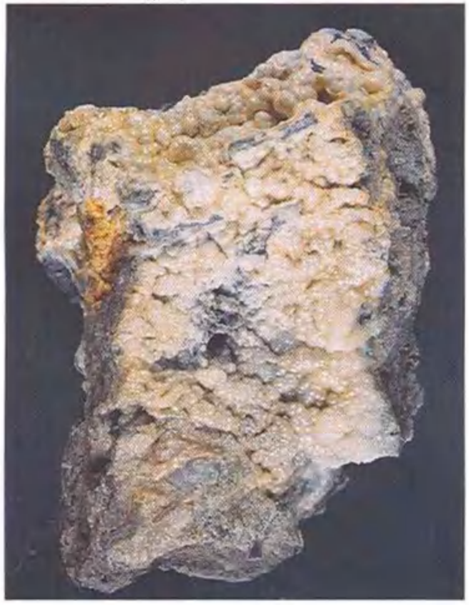

Evansite

$Al_3(PO_4)(OH)_6 . 6 H_2O$

AMORPHOUS 3

Properties: C – colorless, white, greenish, light blue-green, yellowish; S – white; L – vitreous, resinous, waxy; D – transparent to translucent; DE – 1.8-2.2; H – 3-4; CL – none; F – conchoidal; M – botryoidal, stalactitic and hemispherical aggregates and coatings.

Origin and occurrence: Secondary in the oxidation zone of Fe deposits, rich in P, associated with alophane and goethite. Rich stalactitic aggregates come from Zeleznik, Slovakia and from Epernay, France.

Whiteite-(CaFeMg)

$CaFeMg_2Al_2(PO_4)_4(OH)_2 . 8 H_2O$

MONOCLINIC ● ●

Properties: C – brown; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 4; CL – good; F – uneven; M – short prismatic crystals.

Origin and occurrence: Hydrothermal in the cracks of P-rich sediments. Perfect crystals up to 20 mm ($^{25}/_{32}$ in) across come from Big Fish River, Yukon, Canada. Crystals up to 5 mm ($^{3}/_{16}$ in), are known from Lavra da Ilha de Taquaral, Minas Gerais, Brazil.

Whiteite-(CaFeMg), 65 mm, Yukon Territory, Canada

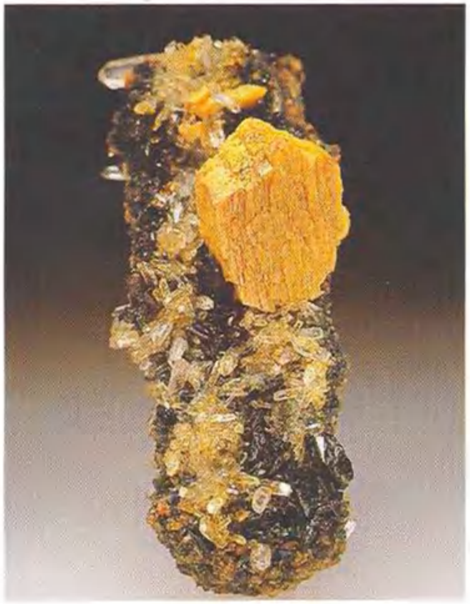

Jahnsite-(CaMnMg), 60 mm, Custer, U.S.A.

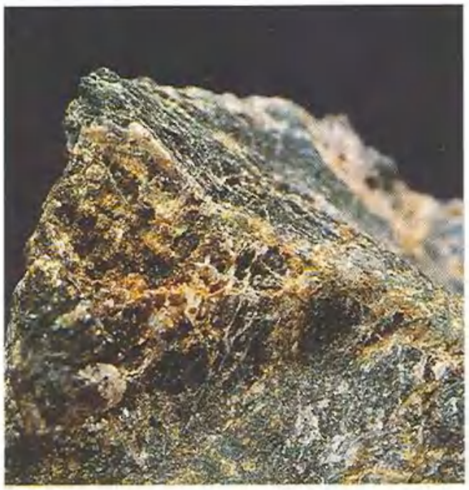

Jahnsite-(CaMnMg)

$CaMnMg_2Fe^{3+}{}_2(PO_4)_4(OH)_4 \cdot 8\,H_2O$

MONOCLINIC ● ●

Properties: C – yellow, light to dark brown; S – light yellow; L – vitreous; D – transparent to translucent; DE – 2.6; H – not determined; CL – good; F – uneven; M – short to long prismatic crystals, granular.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a result of replacement of primary phosphates. Perfect crystals up to 10 mm (⅜ in) across occur in the Tip Top mine, Custer, South Dakota, USA. It is also known from Hagendorf, Germany.

Wardite

$NaAl_3(PO_4)_2(OH)_4 \cdot 2\,H_2O$

TETRAGONAL ● ● ●

Properties: C – colorless, white, greenish, light blue-green; S -white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 5; CL – perfect; F – uneven; M – dipyramidal crystals, radial and hemispherical aggregates, coatings, crusts, granular.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a product of primary phosphate replacement; hydrothermal in the cracks in P-rich sediments. Perfect crystals up to 30 mm (1 3/16 in) across come from Rapid Creek, Yukon, Canada. It is also known from Piedras Lavradas, Paraiba, Brazil.

Wardite, 48 mm, Yukon Territory, Canada

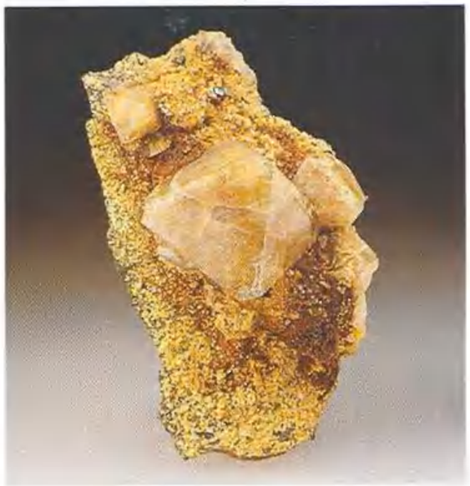

Cyrilovite

$NaFe^{3+}{}_3(PO_4)_2(OH)_4 \cdot 2\,H_2O$

TETRAGONAL ● ●

Properties: C – yellow, orange, brown-yellow; S – yellow; L – vitreous; D – transparent to translucent; DE – 3.1; H – 4; CL – good; F – conchoidal; M – tabular and dipyramidal crystals, coatings and crusts.
Origin and occurrence: Secondary in granitic pegmatites, where it forms as a result of replacement of primary phosphates. Small crystals occur in Cyrilov, Czech Republic; Hagendorf, Germany and Sapucaia, Minas Gerais, Brazil.

Cyrilovite, 60 mm, Iron Monarch, Australia

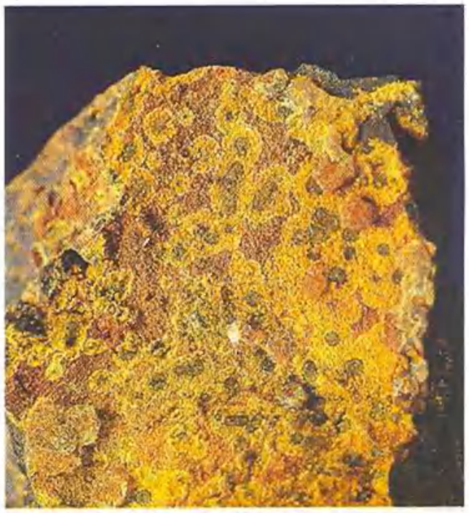

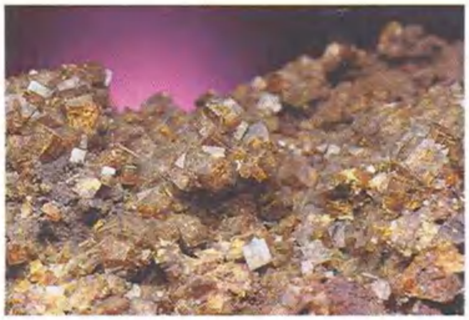
Pharmacosiderite, 35 mm, Cornwall, UK

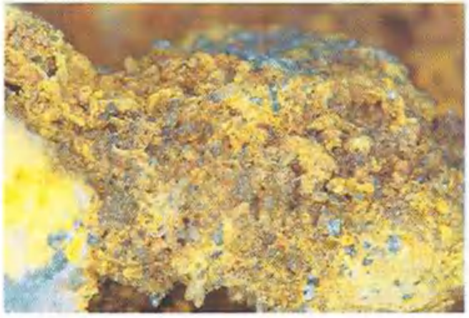
Arseniosiderite, 3 mm, Nagyborzsony, Hungary

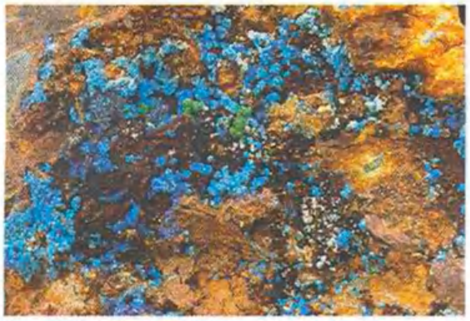
Lavendulan, 36 mm, San Juan, Chile

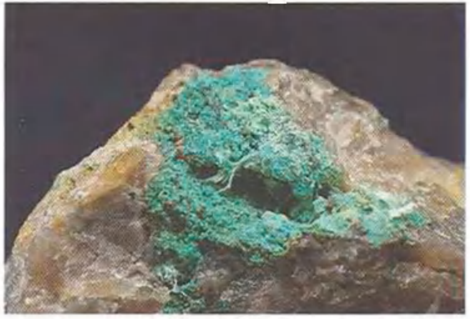
Mixite, 70 mm, Cinovec, Czech Republic

Pharmacosiderite

$KFe^{3+}_4(AsO_4)_3(OH)_4 . 6 H_2O$

CUBIC ● ● ●

Properties: C – green, yellow-brown, brown; S – white; L – adamantine to greasy; D – transparent to translucent; DE – 2.8; H – 2.5; CL – imperfect; F – uneven; M – cubic crystals, coatings, crusts, granular, massive.
Origin and occurrence: Secondary in granitic pegmatites and in the oxidation zone of ore deposits, where it forms as a product of arsenopyrite and lollingite replacement. Cubic crystals up to 10 mm (3/8 in) across occur in St. Day, Liskeard and Redruth, Cornwall, UK. It also comes from the Majuba Hill mine, Nevada, USA; Horhausen, Germany and Cap Garrone, France.

Arseniosiderite

$CaFe^{3+}_3(AsO_4)_3O_2 . 3 H_2O$

MONOCLINIC ● ●

Properties: C – yellow, light to dark brown; S – yellow; L – submetallic to silky; D – opaque; DE – 3.6; H – 1.5; CL – good; F – uneven; M – fibrous and earthy aggregates, coatings, crusts, massive.
Origin and occurrence: Secondary in the oxidation zone of ore deposits and in granitic pegmatites, where it forms as a product of arsenopyrite and lollingite replacement. Rich aggregates occur in Romanche, France; also from the Eureka mine, Tintic, Utah, USA and Wittichen, Germany.

Lavendulane

$NaCaCu_5(AsO_4)_4Cl . 5 H_2O$

ORTHORHOMBIC ● ●

Properties: C – light blue-purple, blue; S – white; L – vitreous to waxy; D – translucent; DE – 3.5; H – 2.5; CL – good; F – uneven; M – acicular crystals, acicular and earthy aggregates, coatings.
Origin and occurrence: Secondary in the oxidation zone of Co, Cu and Ni deposits, as a result of arsenide weathering, associated with erythrite. Rich acicular aggregates come from Talmessi and Anarak, Iran. Small crystals found in the Blanca mine, San Juan, Chile and in Annaberg, Germany. Crystals up to 4 mm (5/32 in) across occur in the Gold Hill mine, Utah, USA.

Kovdorskite, 15 mm x, Kovdor, Kola, Russia

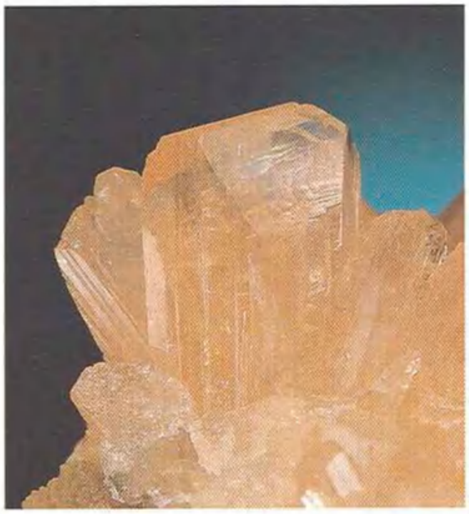

Mixite

$BiCu_6(AsO_4)_3(OH)_6 \cdot 3\ H_2O$

HEXAGONAL ● ●

Properties: C – emerald-green, blue-green, light blue, light green, whitish; S – light green, light blue; L – adamantine to dull; D – translucent; DE – 3.8; H – 3-4; CL – not determined; F – uneven; M – acicular crystals, acicular and earthy aggregates, coatings, massive.
Origin and occurrence: Secondary in the oxidation zone of Bi and Cu deposits, together with bismuthinite. It occurred in Jachymov, Czech Republic; in Tintic, Utah, USA; Schneeberg and Wittichen, Germany.

Kovdorskite

$Mg_5(PO_4)_2(CO_3)(OH)_2 \cdot 4{,}5\ H_2O$

MONOCLINIC ●

Properties: C – light pink-brown, white, blue; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 4; CL – none; F – conchoidal to uneven; M – tabular crystals, massive.
Origin and occurrence: Hydrothermal, associated with magnesite, magnetite and other minerals. Blue and pink-brown crystals up to 25 mm (1 in) across come from the Zhcleznyi mine, Kovdor, Kola Peninsula, Russia.

Torbernite

$Cu(UO_2)_2(PO_4)_2 \cdot 8\text{-}12\ H_2O$

TETRAGONAL ● ● ●

Properties: C – emerald-green to grass-green; S – light green; L – vitreous to dull, pearly on the cleavage planes; D – transparent to translucent; DE – 3.3; H – 2-2.5; CL – perfect; F – uneven; M – tabular to pyramidal crystals, earthy aggregates, coatings, granular, massive; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits, also in pegmatites and sedimentary rocks, resulting from the hydrothermal alteration of uraninite and other U minerals. Emerald-green tabular crystals, several cm across, come from Sabugal, Portugal; Jachymov, Czech Republic; Shinkolobwe, Zaire; Bois-Noirs, France; Moctezuma, Mexico and many localities in the Colorado Plateau, Utah, USA. Beautiful druses of crystals up to 20 mm ($^{25}/_{32}$ in) across, found in the Margabal mine, Aveyron, France.
Application: U ore.

Torbernite, 120 mm, Katanga, Zair

+
→

Autunite, 15 mm xx, Troncasa, Portugal

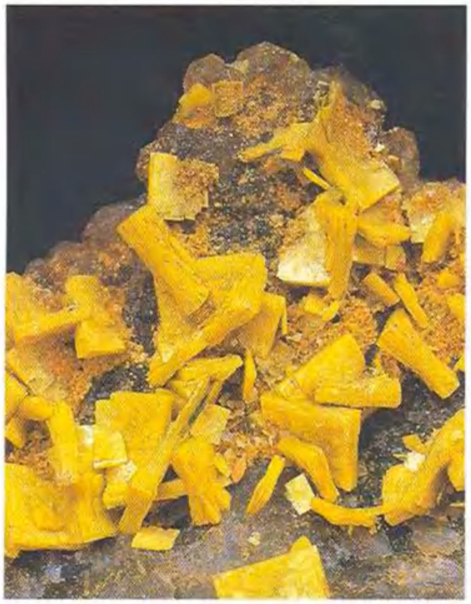

Novacekite, 70 mm, Brumado, Brazil

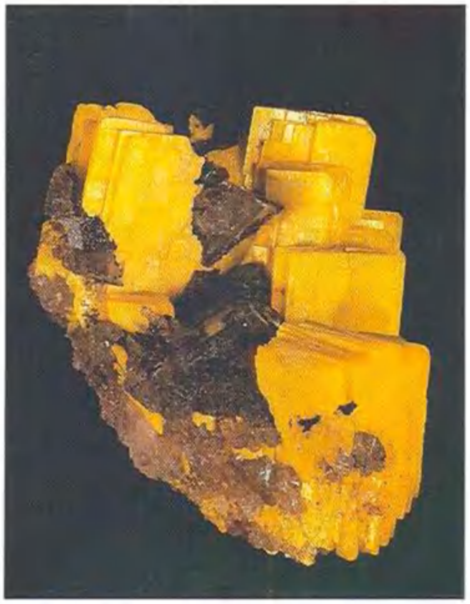

Autunite

$Ca(UO_2)_2(PO_4)_2 \cdot 10\text{-}12\ H_2O$

TETRAGONAL ● ● ●

Properties: C – light to dark yellow, yellow-green to green; S – light yellow; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 2-2.5; CL – perfect; F – uneven; M – tabular crystals, foliated, scaly, earthy and pulverulent aggregates, coatings, massive; LU – yellow-green; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits, in pegmatites and in some U-rich sedimentary rocks, as a result of hydrothermal alteration of uraninite and other U minerals. It is frequently associated with torbernite and other U secondary minerals. Tabular crystals up to 30 mm (1$^{3}/_{16}$ in) across come from Schneeberg and Johanngeorgenstadt, Germany and Autun, France. It is also known from Rum Jungle, Northern Territory, Australia; St. Austel, Cornwall, UK; Mount Spokane, Washington, USA and Jachymov, Czech Republic.
Application: U ore.

Uranocircite

$Ba(UO_2)_2(PO_4)_2 \cdot 10\ H_2O$

TETRAGONAL ● ●

Properties: C – light to dark yellow, light yellow-green; S – light yellow; L – vitreous to dull, pearly on the cleavage planes; D – transparent to translu-

Uranocircite, 78 mm, Minas Gerais, Brazil

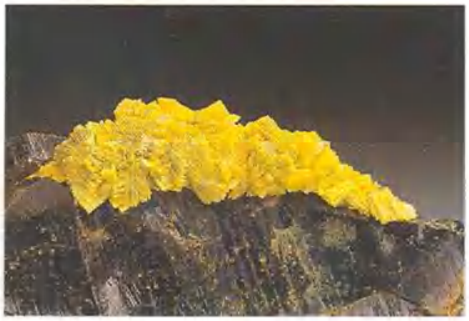

Zeunerite, 5 mm xx, Cinovec, Czech Republic

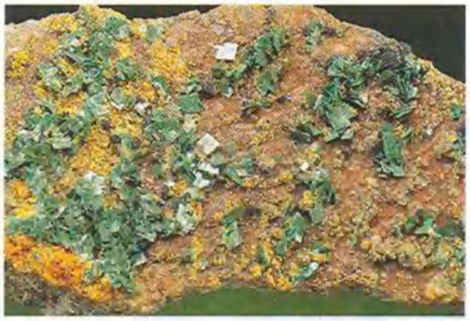

cent; DE – 3.5; H – 2-2.5; CL – perfect; F – uneven; M – tabular crystals, foliated and earthy aggregates, pulverulent coatings, massive; LU – green; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits. Yellow tabular crystals up to 10 mm (3/8 in) across occurred in Dametice, Czech Republic; Bergen and Wolsendorf, Germany and in the Sao Pedro mine, Suaçui, Minas Gerais, Brazil.

Novácekite

$Mg(UO_2)_2(AsO_4)_2 \cdot 10\ H_2O$

TETRAGONAL ● ●

Properties: C – straw-yellow, light yellow; S – light yellow; L – vitreous to dull; D – transparent to translucent; DE – 3.7; H – 2.5; CL – perfect; F – uneven; M – tabular crystals, lamellar, earthy and pulverulent aggregates, massive; LU – dark green; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits. Tabular crystals up to 50 mm (2 in) across come from the Pedra Preta Mine, Brumado, Bahia, Brazil. Lamellar aggregates are known from Zálesi, Czech Republic; Aldama, Chihuahua, Mexico; Wittichen, Germany.

Zeunerite

$Cu(UO_2)_2(AsO_4)_2 \cdot 10\text{-}16\ H_2O$

TETRAGONAL ● ●

Properties: C – emerald-green, yellow-green; S – light green; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 2.5; CL – perfect; F – uneven; M – tabular crystals, foliated aggregates, massive; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of U deposits. Tabular crystals up to 30 mm (1 3/16 in) across, come from the Pedra Preta Mine, Brumado, BahRa, Brazil. It is also known from Zalesí, Czech Republic and Schneeberg, Germany.

Carnotite

$K_2(UO_2)_2(VO_4)_2 \cdot 3\ H_2O$

MONOCLINIC ● ● ●

Properties: C – light to dark yellow, yellow-green; S – light yellow; L – dull; D – transparent to translucent, opaque; DE – 4.9; H – not determined; CL – perfect; F – uneven; M – earthy aggregates, massive; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of sedimentary U deposits, typically associated with tyuyamunite. Platy crystals, up to 2 mm (1/16 in) across found in the Mashamba West mine, Zaire. Earthy and pulverulent aggregates occur in many localities in the Colorado Plateau, e.g. Paradox Valley, Colorado and La Sal, Utah, USA. Also known from Tyuya Muyun, Uzbekistan and Radium Hill, Southern Australia, Australia.
Application: U and V ore..

Carnotite, 55 mm, Utah, U.S.A.

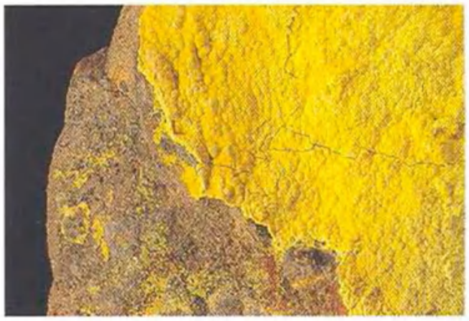

Tyuyamunite

$Ca(UO_2)_2(VO_4)_2 \cdot 5\text{-}8\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – yellow-green, canary yellow; S – light yellow; L – silky to adamantine; D – translucent to opaque; DE – 3.6; H – 2; CL – perfect; F – uneven; M – earthy aggregates, massive; LU – weak yellow-green; R – strongly radioactive.
Origin and occurrence: Secondary in the oxidation zone of sedimentary U deposits, together with carnotite. Common earthy and pulverulent aggregates occur in many localities in the Colorado Plateau, e.g. Paradox Valley, Colorado and Red Creek, Utah, USA. It was described from Tyuya Muyun, Uzbekistan.
Application: U and V ore.

Tyuyamunite, 80 mm, Fergana, Uzbekistan

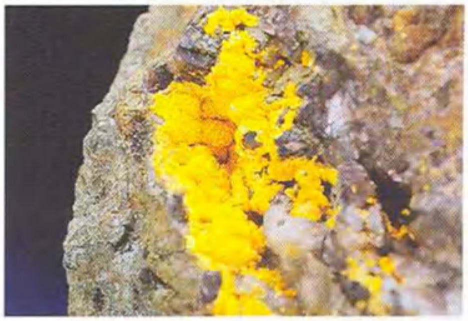

9. Silicates

Phenakite

Be_2SiO_4

TRIGONAL ● ● ●

Properties: C – colorless, white, yellowish; S – white; L – strong vitreous; D – transparent to translucent; DE – 3.0; H – 8; CL – imperfect; F – conchoidal; M – long prismatic to tabular crystals, radial aggregates, granular.
Origin and occurrence: Magmatic in granitic pegmatites; hydrothermal in greisens; metamorphic in mica schists, associated with beryl, chrysoberyl, apatite and quartz. Prismatic crystals up to 250 mm ($9^{13}/_{16}$ in) long occurred in Kragero, Norway. It is also known from Sao Miguel de Piracicaba, Minas Gerais, Brazil in crystals, up to 100 mm (4 in) long. The other localities are Habachtal, Austria; Malyshevo, Russia; Anjanabonoina, Madagascar.
Application: sporadically cut as a gemstone.

Phenakite, 200 mmx, Kragero, Norway

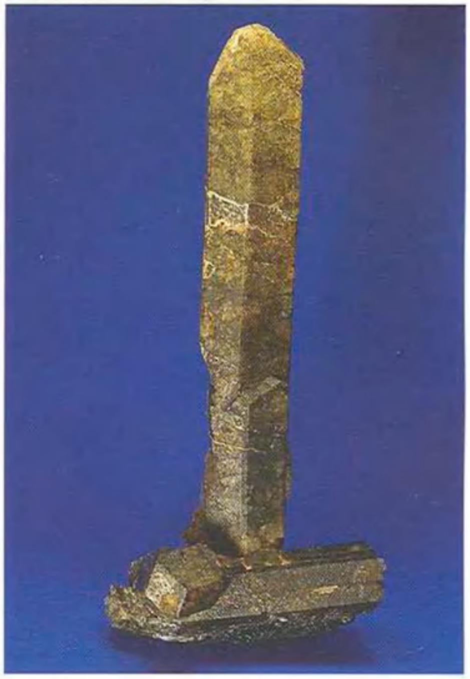

Willemite

Zn_2SiO_4

TRIGONAL ● ● ●

Properties: C – white, yellowish, gray, green; S – white; L – vitreous; D – translucent; DE – 4.0; H – 5.5; CL – good; F B conchoidal to uneven; M – prismatic to tabular crystals, radial aggregates, granular; LU – distinct light green.
Origin and occurrence: Metamorphic in marbles; secondary in the oxidation zone of ore deposits, associated with zincite, franklinite, hemimorphite and smithsonite. Crystals up to 100 mm (4 in) across come from Franklin, New Jersey, USA, Mont St.-Hilaire, Quebec, Canada.
Application: as Zn ore.

Topaz, 60 mm, Thomas Range, U.S.A.
Willemite, 76 mm, Tsumeb, Namibia

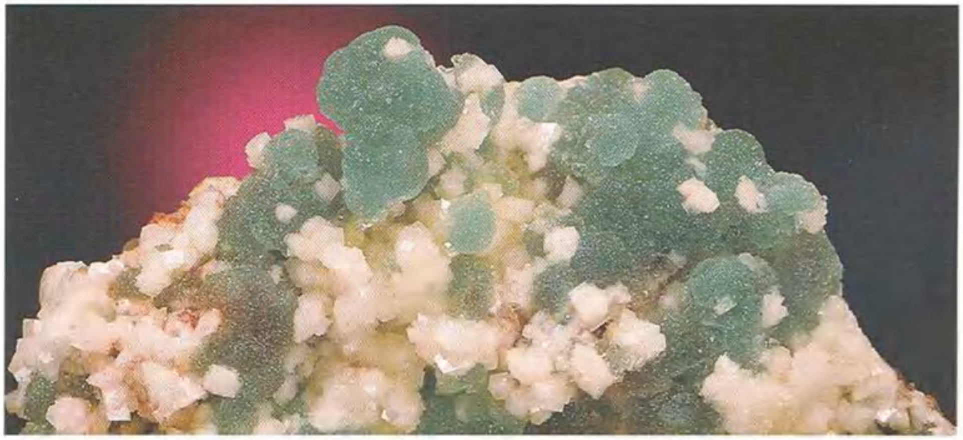

Forsterite, 40 mm, Suppat, Pakistan

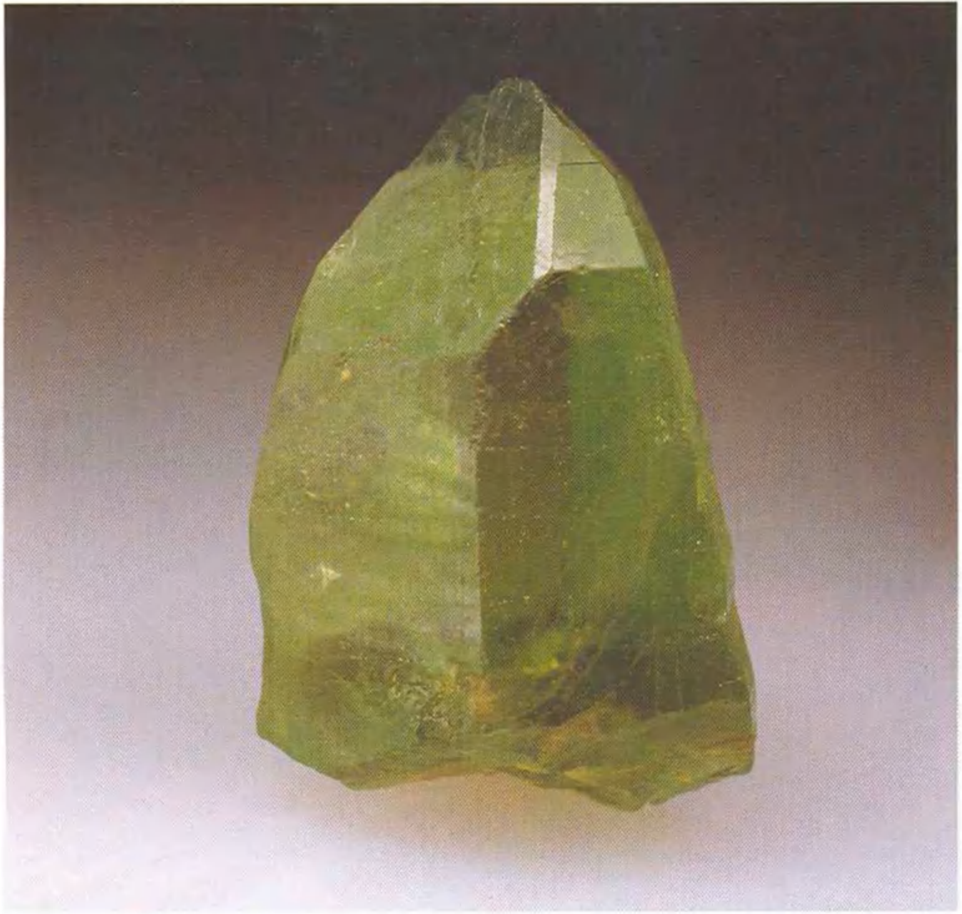

Forsterite

OLIVINE GROUP

Mg_2SiO_4

ORTHORHOMBIC ● ● ● ●

Properties: C – yellowish, greenish, colorless; S – white; L – vitreous; D – transparent to opaque; DE – 3.3; H – 6.5-7; CL – good; F – conchoidal to uneven; M – tabular to prismatic crystals, granular.

Origin and occurrence: Metamorphic in regionally and contact metamorphosed dolomites. Typical rock-forming mineral, associated with enstatite, spinel, phlogopite and chlorite. Green gemmy crystals up to 80 mm (3⅛ in) long come from Suppat, Pakistan. It also occurred in Crestmore, California, USA, Mount Timobly, British Columbia, Canada and Monte Somma, Italy.

Olivine

OLIVINE GROUP

$(Mg,Fe)_2SiO_4$

ORTHORHOMBIC ● ● ● ● ●

Varieties: chrysolite

Properties: C – green, yellow-green (chrysolite), brown-green to black-green; S – white; L – vitreous; D – translucent to opaque; DE – 3.3-3.6; H – 6.5-7; CL – good; F B conchoidal to uneven; M – imperfect crystals, granular.

Origin and occurrence: Magmatic in some ultrabasic rocks, e.g. dunites, lherzolites, peridotites, gabbros and in meteorites. Typical rock-forming mineral, usually associated with diopside, magnetite and pyrope. Classic locality of chrysolite is Zebirget Island in the Red Sea, Egypt, where tabular crystals

Olivine, 60 mm, Smrci, Czech Republic

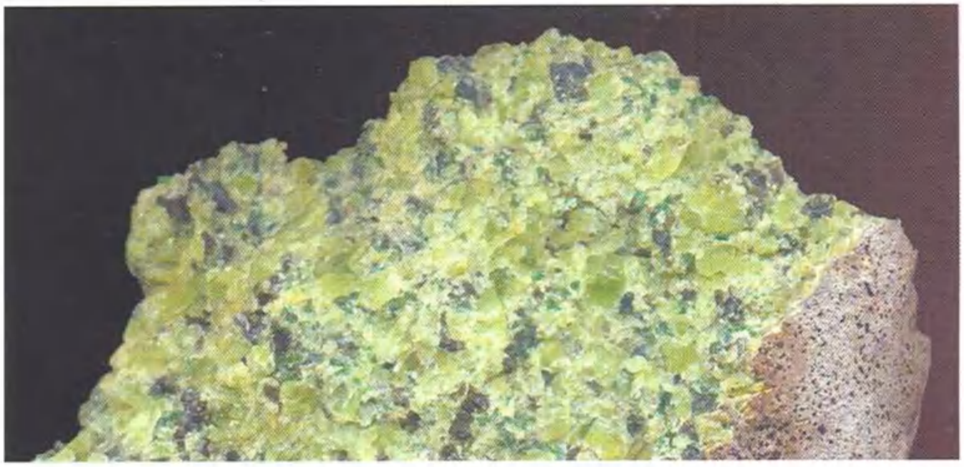

up to 100 mm (4 in) across occurred. It is also known from the San Carlos Indian Reservation, Arizona, USA. Olivine was found in many basaltic rocks in Laacher See, Germany; Rockport, Massachusetts, USA and elsewhere.
Application: chrysolite is cut as a gemstone.

Fayalite

OLIVINE GROUP

Fe_2SiO_4

ORTHORHOMBIC ● ● ●

Properties: C – black-green to black; S – gray; L – dull to vitreous; D – opaque; DE – 4.2; H – 6.5-7; CL – good; F B conchoidal to uneven; M – imperfect prismatic crystals, granular.
Origin and occurrence: Magmatic in granitic pegmatites, granites and syenites, associated with orthoclase, gadolinite-(Y) and epidote; rare metamorphic. Poorly developed crystals up to 150 mm (6 in) long found in pegmatites near Baveno, the Alps, Italy. It is also known from Strzegom, Poland and the Sawtooth Batholith, Idaho, USA.

Fayalite, 80 mm, Rockport, U.S.A.

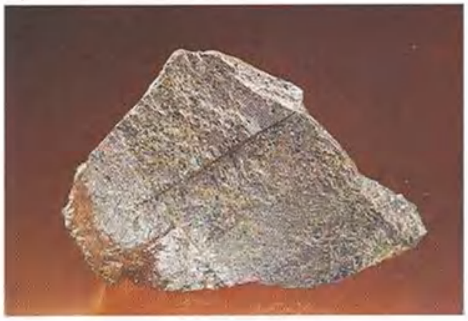

Tephroite

OLIVINE GROUP

Mn_2SiO_4

ORTHORHOMBIC ● ●

Properties: C – gray, olive-green, red-brown; S – white; L – dull to vitreous; D – translucent to transparent; DE – 4.2; H – 6; CL – good; F B conchoidal to uneven; M – prismatic crystals, granular.
Origin and occurrence: Metamorphic in skarns and Mn-rich metamorphosed sediments, together with rhodonite, franklinite and spessartine. Granular aggregates and perfect crystals up to 50 mm (2 in) across known from Franklin, New Jersey, USA; Langban, Sweden. It comes also from Tarnobrzeg, Poland, in crystals, up to 80 mm (3⅛ in) across.

Tephroite, 60 mm, Harstigen, Sweden

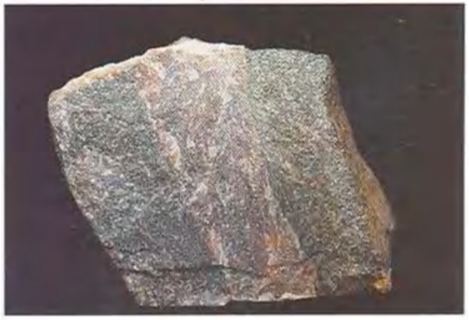

Pyrope, 3 mm grains, Trebivlice, Czech Republic

Pyrope

GARNET GROUP

$Mg_3Al_2Si_3O_{12}$

CUBIC ● ● ● ●

Properties: C – red to purple-red, light purple, black-brown; S – white; L – vitreous; D – transparent to translucent; DE – 3.5; H – 7-7.5; CL – none; F B conchoidal to uneven; M – isometric crystals, granular.

Origin and occurrence: Magmatic in some ultrabasic rocks, e.g. lherzolites, peridotites, kimberlites, eclogites and serpentinites; metamorphic in quartzites; also known from placers. It is associated with diopside, magnetite and diamond. It comes from many localities in ultrabasic rocks, like Trebenice and Mirunice, Czech Republic; Zoblitz, Germany; Madras, India; Kimberley, South Africa. Crystals up to 250 mm ($9^{13}/_{16}$ in) across were found in Dora Maria, the Alps, Italy.

Application: cut as a gemstone.

Almandine, 38 mm x, Ötztal, Austria

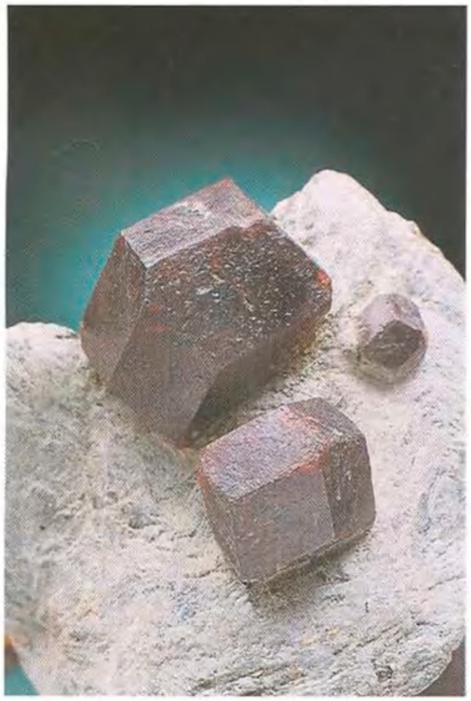

Almandine

GARNET GROUP

$Fe_3Al_2Si_3O_{12}$

CUBIC ● ● ● ●

Properties: C – red to purple-red, black-brown; S – white; L – vitreous; D – transparent to translucent; DE – 4.3; H – 7; CL – none; F B conchoidal to uneven; M B well-formed crystals, granular.

Origin and occurrence: Metamorphic in regionally metamorphosed rocks, as chlorite schists, gneisses,

Spessartine, 96 mm, Gilgit, Pakistan

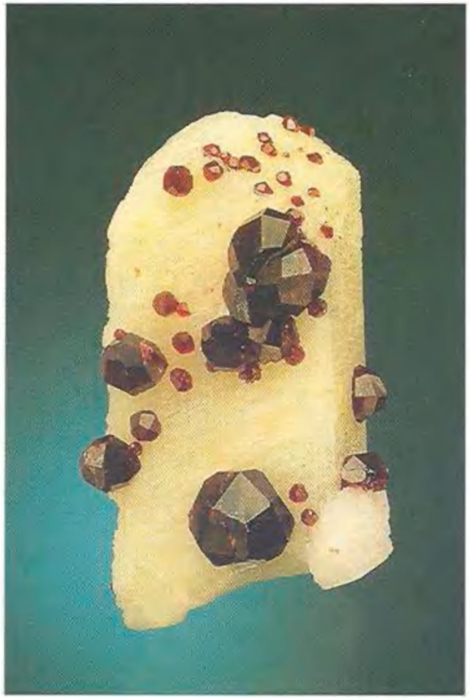

Grossular, 17 mm x, Asbestos, Canada

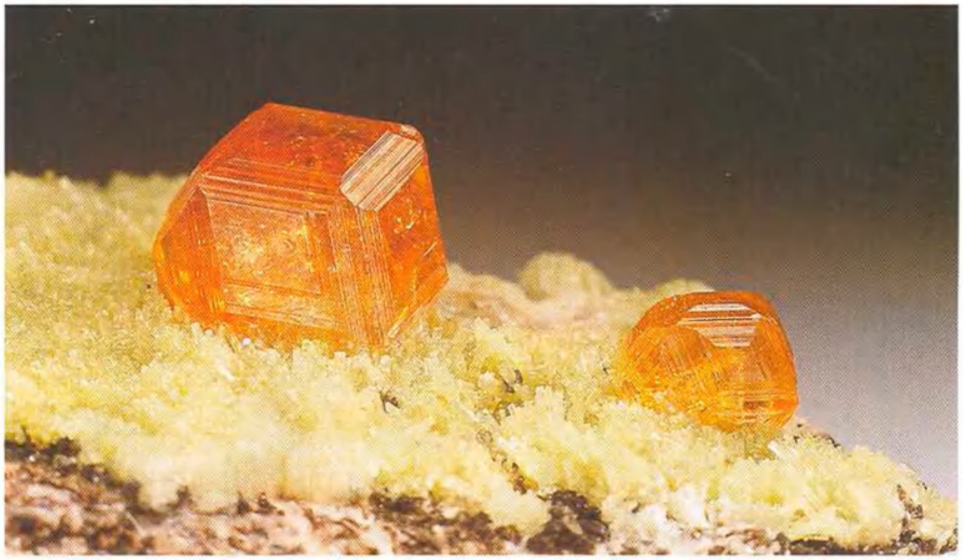

mica schists and migmatites; magmatic in some granites and pegmatites; also in placers. Well-developed crystals up to 150 mm (6 in) across are known from Ishikawa pegmatites, Japan and Shingus, Pakistan. It comes from many mica schists and gneisses as crystals, up to about 50 mm (2 in), like Fort Wrangel, Alaska, USA; Ötztal, Austria; Bodo, Norway. It occurs in placers near Ratnapura, Sri Lanka.
Application: cut as a gemstone, abrasive material.

Spessartine

GARNET GROUP

$Mn_3Al_2Si_3O_{12}$

CUBIC ● ● ● ●

Properties: C – red, orange, light brown to yellowish; S – white; L – vitreous; D – transparent to translucent; DE – 4.3; H – 7-7.5; CL – none; F B conchoidal to uneven; M – perfect crystals, granular.
Origin and occurrence: Magmatic in granitic pegmatites and some granites; hydrothermal in cavities in rhyolites; metamorphic in some skarns and Mn-rich metamorphic rocks. Perfect crystals up to 30 mm ($1^{3}/_{16}$ in) across were found in granitic pegmatites in the Hercules mine, Ramona, California, USA; near Marienfluss river, Namibia and in rhyolite cavities in Nathrop, Colorado, USA.
Gemmy crystal fragments up to 50 mm (2 in) across were recently found in an undisclosed locality in Minas Gerais, Brazil.
Application: cut as a gemstone.

Grossular

GARNET GROUP

$Ca_3Al_2Si_3O_{12}$

CUBIC ● ● ● ●

Varieties: hessonite, tsavorite

Properties: C – red, green (tsavorite), orange, red-brown (hessonite) to colorless; S – white; L – vitreous; D – transparent to tranlucent; DE – 3.4; H – 6.5-7; CL – none; F B conchoidal to uneven; M – perfect crystals, granular.
Origin and occurrence: Metamorphic in Ca-rich, contact metamorphic rocks, skarns, rodingites; hydrothermal along the cracks in these rocks, associated with diopside, vesuvianite, wollastonite, scapolite and epidote. Perfect crystals about 30 mm ($1^{3}/_{16}$ in) across occur in the Jeffrey quarry, Asbestos, Quebec, Canada and Sierra de las Cruces, Coahuila, Mexico.
Tsavorite crystals up to 50 mm (2 in) across come from the Tsavo National Park, Kenya and Merelani Hills, Arusha, Tanzania. The other well-known localities are Ala, Italy and Ciclova, Romania. Grossular crystals up to 100 mm (4 in) in size in Vapenna, Czech Republic; Xalostoc, Mexico and Sandare, Mali.
Application: cut as a gemstone.

Andradite, 30 mm xx, Graham Co., U.S.A.

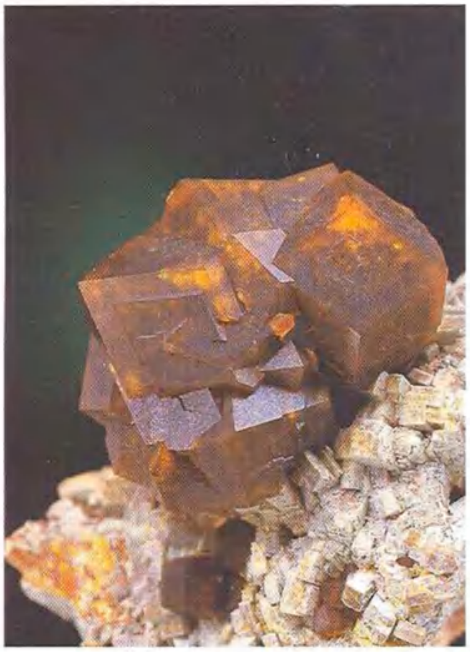

Andradite

GARNET GROUP

$Ca_3Fe^{3+}_2Si_3O_{12}$

CUBIC ● ● ● ●

Varieties: demantoid, melanite

Properties: C – dark red, black-brown, brown, green to yellow-green (demantoid), black-brown to black (melanite); S – white; L – vitreous; D – transparent to translucent; DE – 3.9; H – 6.5-7; CL – none; F B conchoidal to uneven; M B well-developed crystals, granular.
Origin and occurrence: Metamorphic in Ca and Fe rich contact metamorphosed rocks, skarns, rodingites; hydrothermal along the cracks of these rocks; magmatic in some alkaline igneous rocks, usually in the same localities as grossular. Crystals up to 40 mm (1 9/16 in) across come from Sinerechenskoye, Russia. Fine crystals were also found in the Namgar mine, Usakos, Namibia and Ciclova, Romania. Fine melanite crystals occur in Magnet Cove, Arkansas, USA. Demantoid crystals up to 30 mm (1 3/16 in) across known from Val Malenco, Italy and in the Bobrovka river basin, Ural mountains, Russia.
Application: demantoid is cut as a gemstone.

Demantoid, 3 mm xx, Bobrovka, Russia

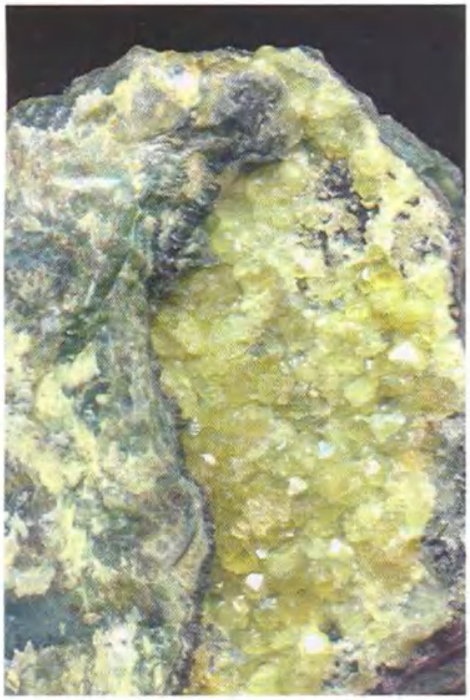

Melanite, 5 mm xx, Rudnyi, Kazakhstan

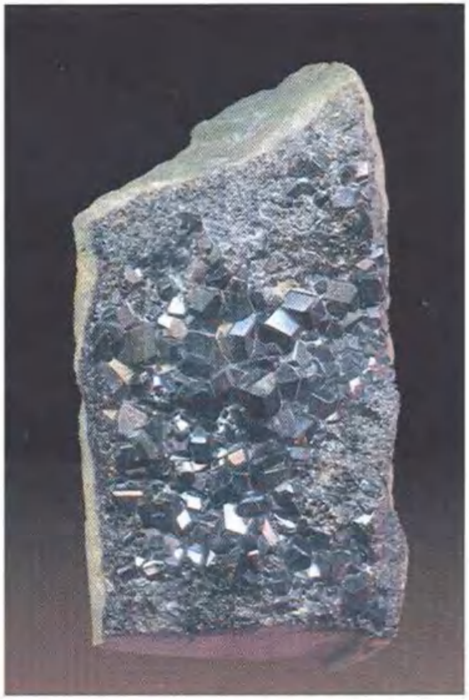

Uvarovite, 2 mm xx, Sarany, Russia

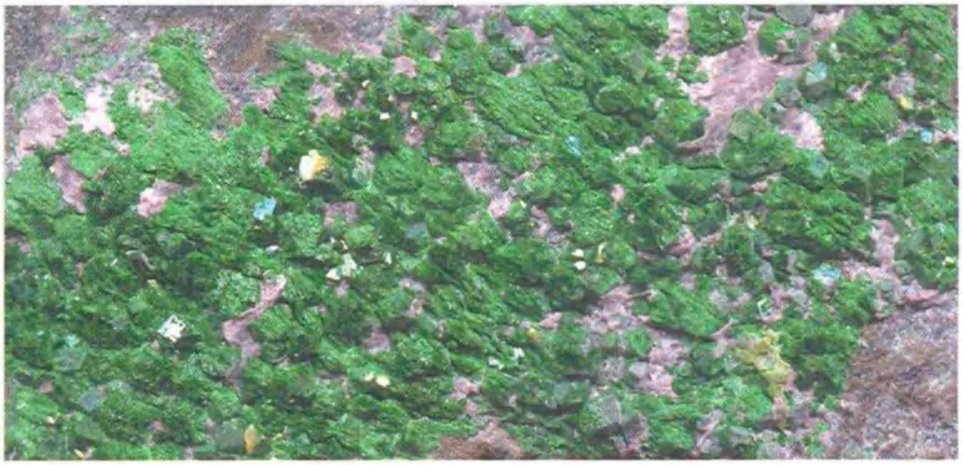

Uvarovite

GARNET GROUP

$Ca_3Cr_2Si_3O_{12}$

CUBIC ● ●

Properties: C – dark emerald-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.9; H – 6.5-7; CL – none; F B conchoidal to uneven; M – perfect crystals, granular.

Origin and occurrence: Metamorphic and also hydrothermal are almost only limited to rocks with increased Cr content, ultrabasic rocks with chromite, serpentinites and skarns. It occurs as crystals up to 8 mm ($^{5}/_{16}$ in) across along cracks in chromite in Sarany, Ural mountains, Russia. Crystals up to 20 mm ($^{25}/_{32}$ in) across, come from Outokumpu, Finland. It is also known from Orford, Quebec, Canada.

Zircon

$ZrSiO_4$

TETRAGONAL ● ● ●

Varieties: jargon, hyacinth.

Properties: C – yellow (jargon), brown, yellow-brown, red-orange (hyacinth), red to colorless; S – white; L – vitreous, greasy to adamantine; D – transparent to translucent; DE – 4.7; H – 7.5; CL – imperfect; F B conchoidal to uneven; M – perfect long and short prismatic crystals, granular; LU – yellow; R B sometimes radioactive and usually metamict.

Origin and occurrence: Magmatic and metamorphic as an accessory mineral in different rock types; rarely hydrothermal in the Alpine-type and quartz veins; also in sediments and placers. Prismatic crystals up to 30 cm (12 in) across, occur in syenite pegmatites in Renfrew and Bancroft, Ontario, Canada. It is also known from Miass, Ural mountains and from Mount Vavnbed, Lovozero massif, Kola Peninsula, Russia. Classic occurrences in granitic pegmatites are Alto Ligonha, Mozambique; Arendal, Norway; Ytterby, Sweden and Jaguaraçu, Minas Gerais, Brazil. Gemmy zircons of different colors up to 80 mm ($3^{1}/_{8}$ in) in size come from placers near Ratnapura, Sri Lanka and elsewhere.

Application: hyacinth and jargon are cut as gemstones, Zr ore.

Zircon, 20 mm x, Vishnevogorsk, Ural Mts., Russia

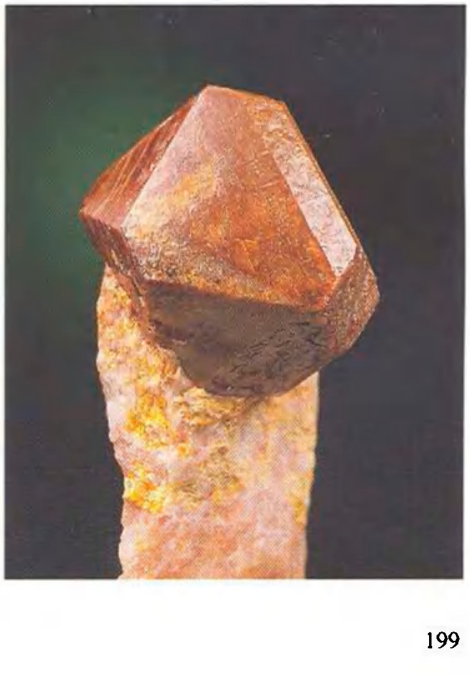

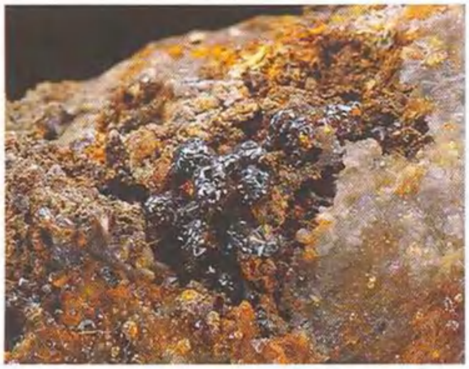
Eulytite, 2 mm xx, Schneeberg, Germany

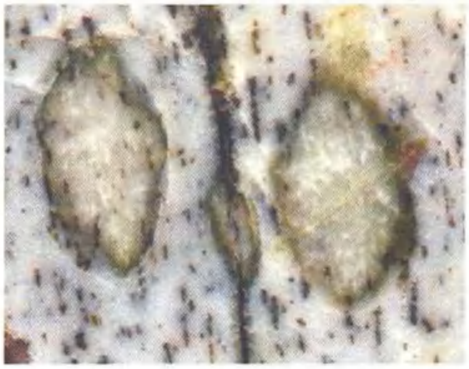
Sillimanite, 10 mm grains, Havlíčkův Brod, Czech Republic

Eulytine

$Bi_4Si_3O_{12}$

CUBIC ● ●

Properties: C – brown, yellow, gray, colorless; S – white; L – greasy; D – translucent; DE – 6.6; H – 4.5; CL – good; F B conchoidal to uneven; M – small dipyramidal crystals, radial aggregates, granular.
Origin and occurrence: Secondary in Bi deposits, typically associated with bismuth.
Its small crystals come from Horní Slavkov, Czech Republic; Schneeberg and Johanngeorgenstadt, Germany.

Euclase

$BeAlSiO_4(OH)$

MONOCLINIC ● ●

Properties: C – colorless, white, greenish, blue; S – white; L – vitreous; D – transparent to translucent; DE – 3.1; H – 7.5; CL – good; F B conchoidal to

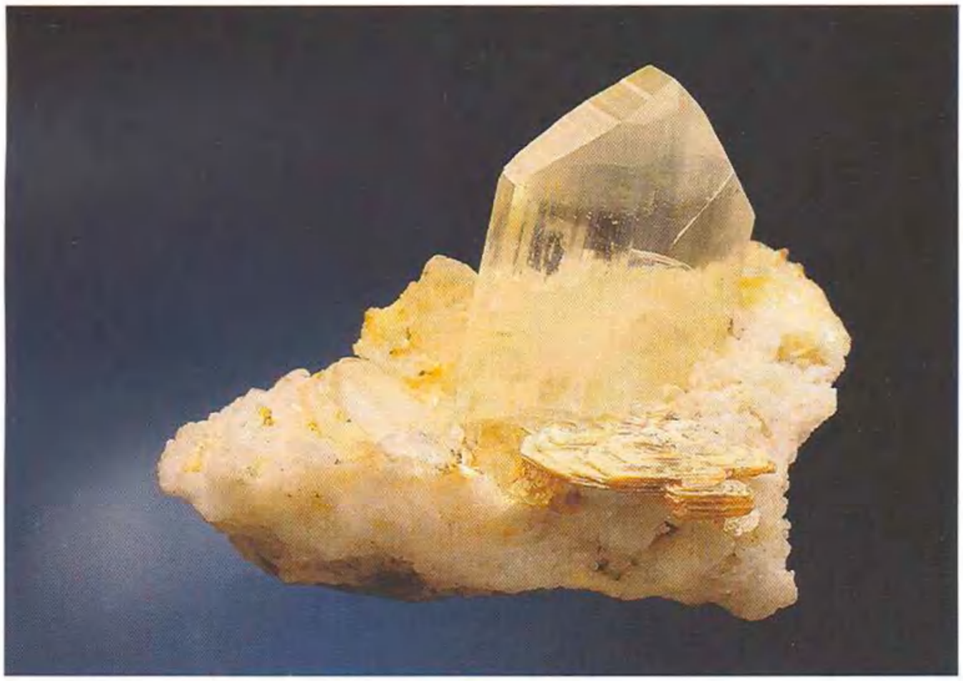
Euclase, 42 mm, Minas Novas, Brazil

uneven; M – long prismatic crystals, radial aggregates, granular.
Origin and occurrence: Hydrothermal occurs in pegmatites, greisens and in quartz and Alpine-type veins; rare in placers. Well-formed crystals up to 80 mm (3$^{1}/_{8}$ in) in size known from Santa do Encoberto, Minas Gerais, Brazil. Blue crystals up to 50 mm (2 in) across occurred in the Last Hope Mine, Karoi, Zimbabwe. It also come from the sediments of Sanarka River, Ural mountains, Russia. Dark blue crystals up to 150 mm (6 in) across were found recently in the Chivor Mine, Colombia.
Application: locally cut as a gemstone.

Sillimanite

Al_2SiO_5

ORTHORHOMBIC ● ● ● ● ●

Properties: C – white, gray, greenish, yellowish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 6.5-7.5; CL – good; F – uneven; M – long prismatic crystals, fibrous aggregates, granular.
Origin and occurrence: Almost exclusively metamorphic in gneisses and migmatites; only rare magmatic in pegmatites and granites; also in placers. Typical rock-forming mineral, very commonly associated with andalusite. Typical localities are Bodenmais, Germany and Mar_íkov, Czech Republic. Gemmy crystals up to 20 mm ($^{25}/_{32}$ in) long are known from Rakwana-Deniyaya, Sri Lanka.

Andalusite

Al_2SiO_5

ORTHORHOMBIC ● ● ● ●

Varieties: chiastolite, viridine

Properties: C – pink, red-brown, red, gray, whitish, green (viridine); S- white; L – vitreous to dull; D – translucent to transparent; DE – 3.2; H – 6.5-7.5; CL – good; F – uneven; M – prismatic crystals, fibrous aggregates, granular, massive.
Origin and occurrence: Metamorphic in regionally and contact metamorphosed rocks; magmatic in pegmatites and granites, hydrothermal in quartzites. Typical rock-forming mineral, commonly associated with sillimanite, corundum and cordierite. Renowned localities are Lisens, the Alps, Austria; Bimbowrie, South Australia, Australia. Green gemmy crystals come from Morro do Chapeu, Bahia, Brazil. Viridine is known from Darmstadt, Germany.

Andalusite, 28 mm x, Aracuay, Minas Gerais, Brazil

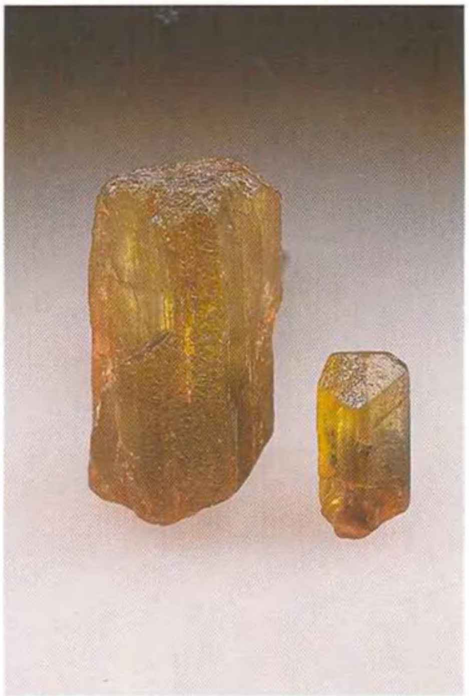

Chiastolith, 20 mm, Bimbowrie, Australia

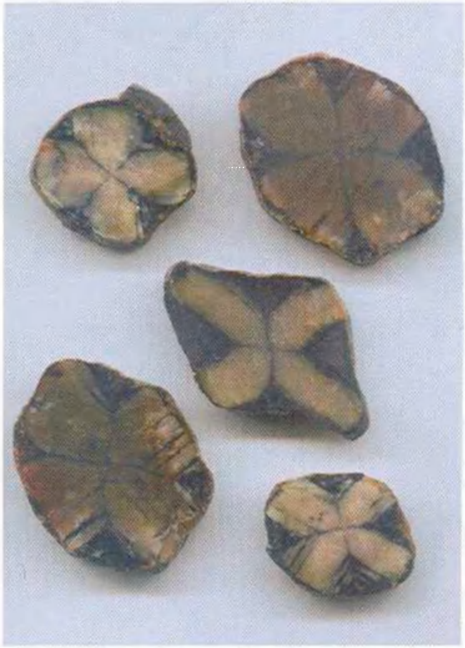

Kyanite

Al_2SiO_5

TRICLINIC ● ● ● ●

Properties: C – blue, gray, white, green, dark gray, colorless; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.6; H – 4.5-7.5; CL – good; F – uneven; M – prismatic to tabular crystals, fibrous aggregates, granular, massive.
Origin and occurrence: Almost only metamorphic in regionally metamorphosed rocks, mica schists, gneisses, granulites and eclogites; less frequently magmatic in pegmatites and granites; rarely hydrothermal in quartz veins. Typical rock-forming mineral, associated with andalusite and sillimanite. Blue columnar aggregates and crystals up to 150 mm (6 in) long occur in Barra do Salinas, Minas Gerais, Brazil. Other renowned localities are Pizzo Forno, Switzerland; Prilep, Macedonia; Keivy, Kola Peninsula, Russia.

Kyanite, 87 mm, Minas Gerais, Brazil

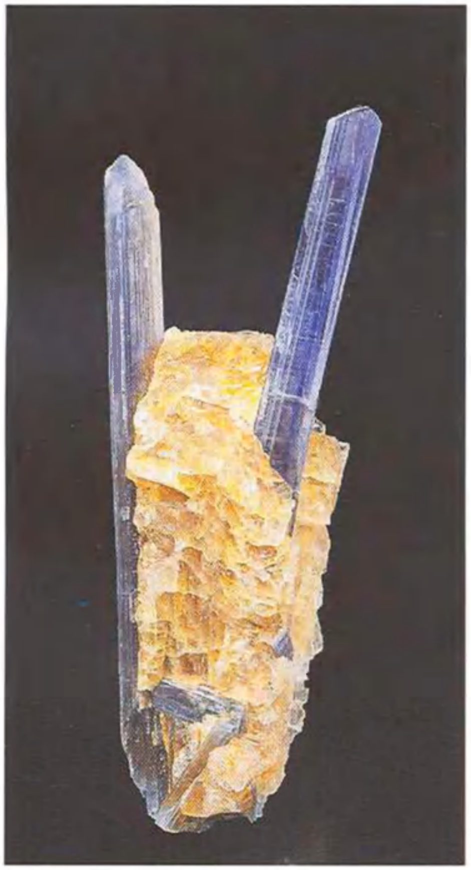

Topaz, 18 mm xx, Ghundao Hill, Pakistan

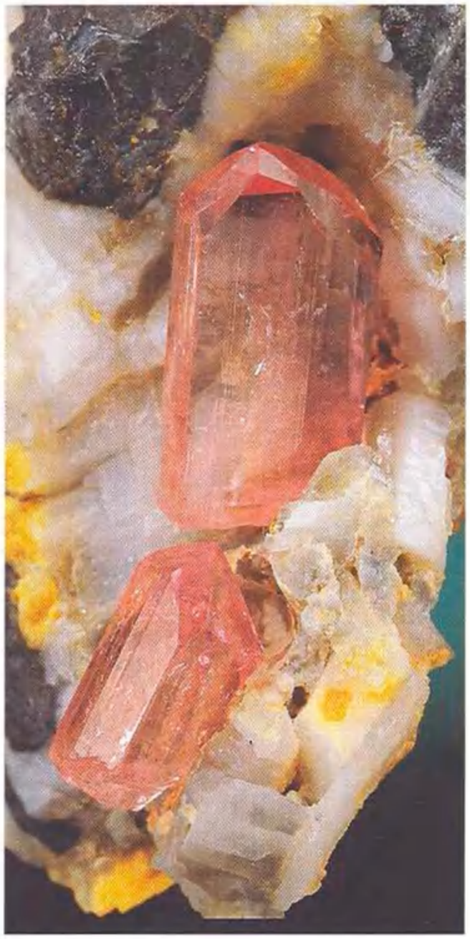

Topaz

$Al_2SiO_4(F,OH)$

ORTHORHOMBIC ● ● ●

Varieties: pycnite

Properties: C – colorless, blue, yellow, gray, white, greenish, pinkish, red; S – white; L – vitreous; D – transparent to translucent; DE – 3.6; H- 8; CL – good; F – uneven; M – perfect prismatic to tabular crystals, radial and columnar aggregates, granular.
Origin and occurrence: Magmatic in pegmatites and granites; hydrothermal in greisens, in rhyolite cavities, in quartz veins, also in placers. Topaz crystals in pegmatites are occasionally very large, like the crystal measuring 80 x 60 x 60 cm (31½ x 24 x 24 in) across from Fazenda do Funil; crystals

Topaz, 49 mm, Ramona, U.S.A.

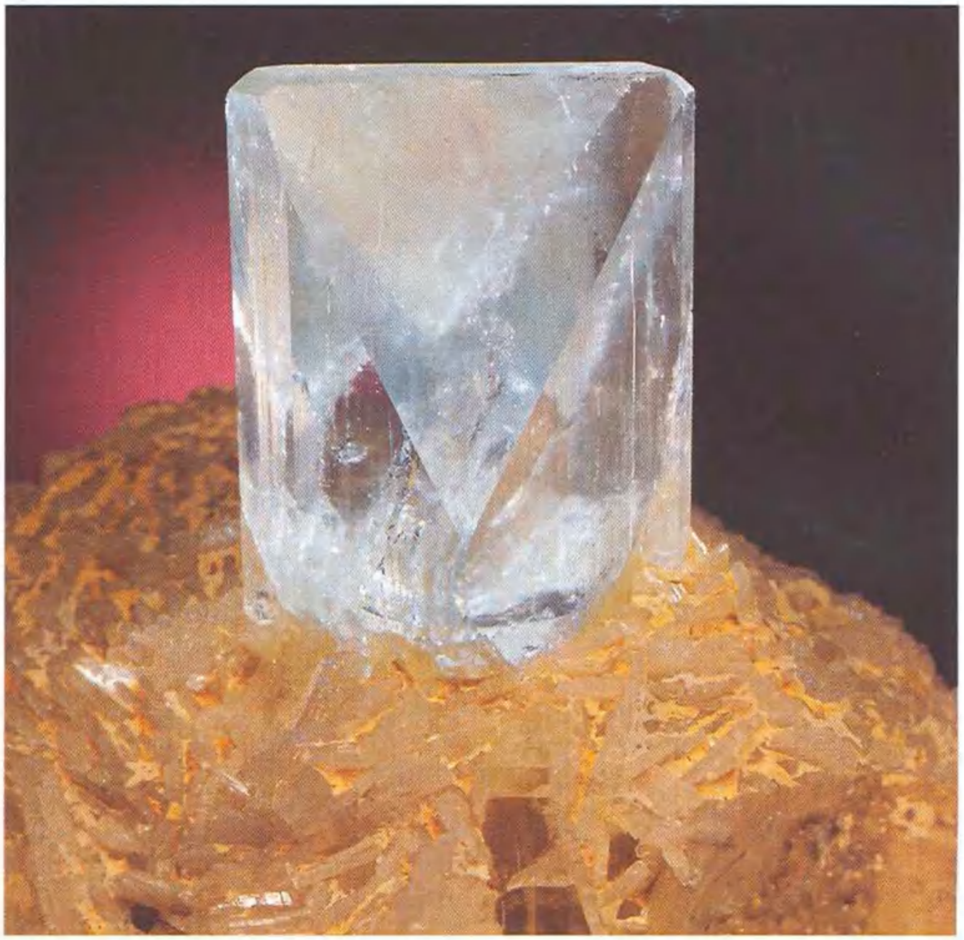

up to 30 cm (12 in) from Virgem da Lapa, both Minas Gerais, Brazil. Blue, brownish and bicolored crystals up to 40 cm (15¾ in) long come from Volodarsk Volynskii, Ukraine. Other topaz localities are Iveland, Norway; Murzinka, Ural mountains, Russia; Little Three mine, Ramona, California; Pikes Peak, Colorado, USA; Gilgit, Pakistan; Spitzkopje, Namibia. Orange to red topaz (imperial topaz) comes from quartz veins near Ouro Preto, Minas Gerais, Brazil. Pink topaz crystals up to 70 mm (2¾ in) found at Mount Ghundao, Mardan, Pakistan. Important topaz specimens were also found in Schneckenstein, Germany; Thomas Range, Utah, USA; Nerchinsk, Siberia, Russia. Columnar aggregates of pycnite come from Cínovec, Czech Republic and Altenberg, Germany. *Application:* cut as a gemstone.

Pycnite, 60 mm, Cínovec, Czech Republic

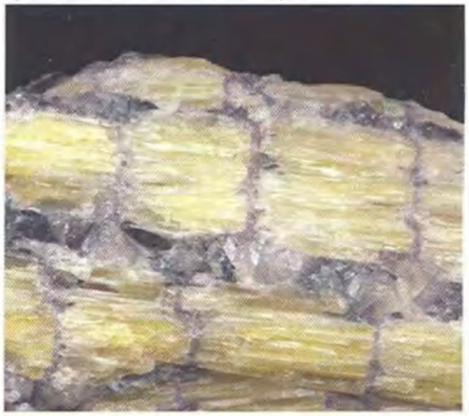

Staurolite, 80 mm, Keivy, Kola, Russia

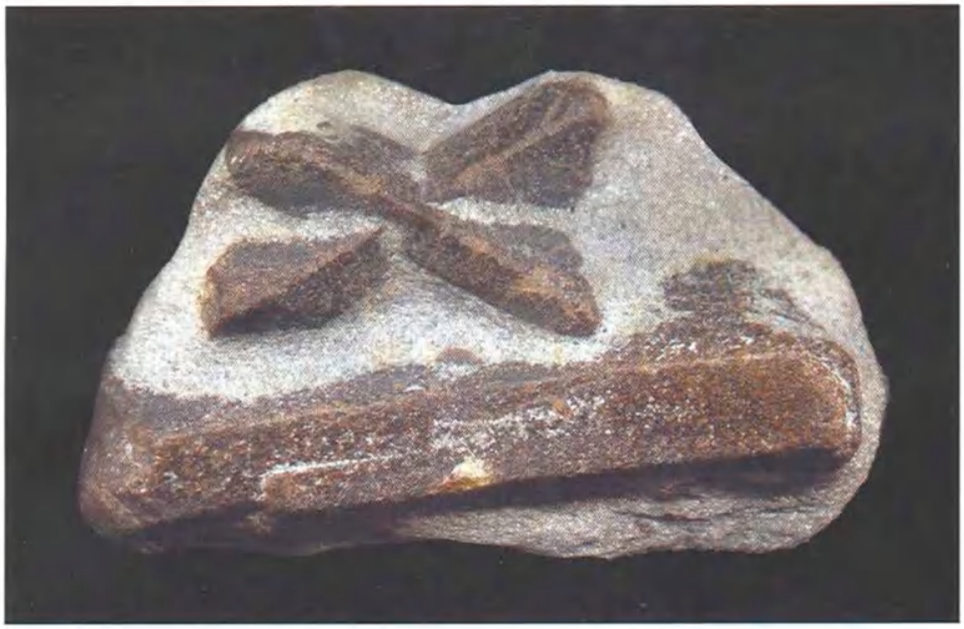

Staurolite

$Fe_2Al_9Si_4O_{22}(OH)_2$

ORTHORHOMBIC ● ● ● ●

Properties: C – dark to light brown, yellow; S – white; L – dull to vitreous; D – transparent to almost opaque; DE – 3.7; H – 7-7.5; CL – good; F – uneven; M – prismatic to tabular crystals and their combinations, granular.

Origin and occurrence: Metamorphic in regionally metamorphosed rocks, gneisses and mica schists; rare magmatic in granites; also in placers, commonly associated with almandine, andalusite and kyanite. Crystals up to 50 mm (2 in) across come from Pizzo Forno, Switzerland. It is also known from Rio Arriba, New Mexico, USA. Its cross-like twins, up to 200 mm ($7^{7}/_{8}$ in) across, occur in Keivy, Kola Peninsula, Russia and Morbihan, France.

Sapphirine, 26 mm, Vohimena, Madagascar

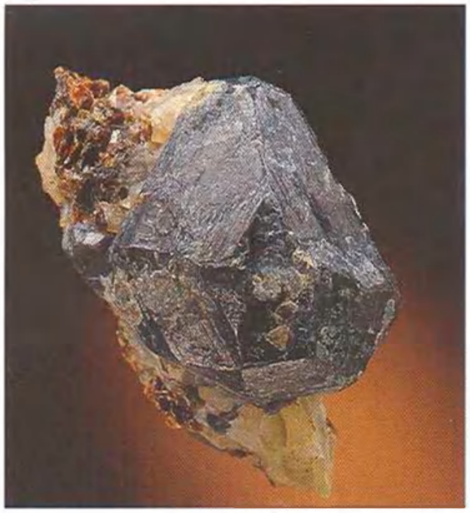

Sapphirine

$Mg_2Al_4SiO_{10}$

MONOCLINIC ● ●

Properties: C – dark to light blue, green; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.5; H – 7.5; CL – imperfect; F – uneven; M – tabular crystals, granular.

Origin and occurrence: Metamorphic in regionally metamorphosed rocks, rich in Al and Mg and poor in Si, associated with spinel and corundum. Crystals up to 40 mm ($1^{9}/_{16}$ in) across come from Fiskenaesset, Greenland; Betroka and Androy, Madagascar. Other localities include Val Codera, Italy and Enderby Land, Antarctica.

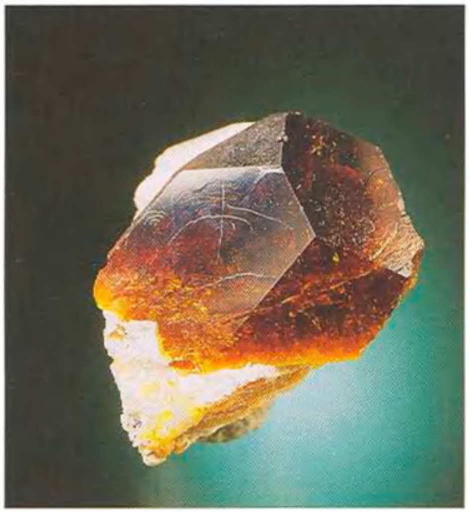
Chondrodite, 15 mm x, Brewster, U.S.A.

Chondrodite
$Mg_5Si_2O_8(OH,F)_2$

MONOCLINIC ● ● ●

Properties: C – yellow, greenish, brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 6-6.5; CL – imperfect; F – uneven; M – crystals of different habits, granular.
Origin and occurrence: Metamorphic in contact and regionally metamorphosed carbonate rocks; rare magmatic in carbonatites, typically associated with spinel, chlorite and phlogopite. Perfect crystals up to 50 mm (2 in) long come from the Tilly Foster mine, Brewster, New York, USA. Other localities are Pargas, Finland; Monte Somma, Italy; Riverside, California, USA.

Clinohumite
$Mg_9Si_4O_{16}(OH,F)_2$

MONOCLINIC ● ● ●

Properties: C – yellow, red, brown, white; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 6; CL – imperfect; F – uneven; M – crystals of different habits, granular.
Origin and occurrence: Metamorphic in contact and regionally metamorphosed carbonate rocks, serpentinites and talc schists, associated with spinel, chlorite, forsterite, serpentine and phlogopite. Gemmy yellow crystals up to 30 mm ($1^{3}/_{16}$ in) across come from Kukh-i-Lal, Pamir, Tajikistan. Other localities are Pargas, Finland; Monte Somma, Italy; Jensen quarry, California, USA.

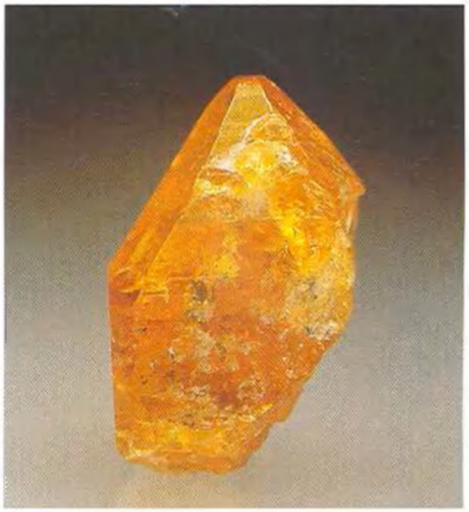
Clinohumite, 19 mm, Kukh-i-lal, Tadzhikistan

Chapmanite
$SbFe^{3+}_2Si_2O_8(OH)$

MONOCLINIC ● ●

Properties: C – olive-green to dark yellow; S – yellowish to yellow; L -dull; D – translucent; DE – 3.7; H – 2.5; CL – imperfect; F – uneven; M – elongated crystals, massive.
Origin and occurrence: Hydrothermal and secondary in the cracks of rocks, sometimes with stibnite. Its massive aggregates occur in Smilkov near Votice, Czech Republic; in the Keely mine, Cobalt, Ontario, Canada.

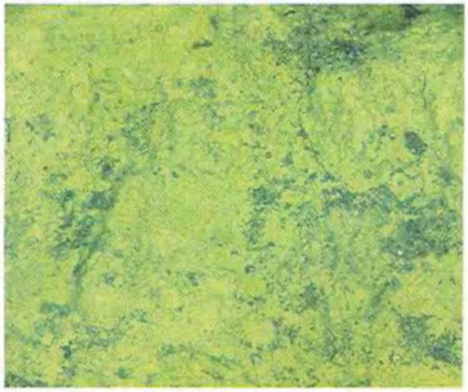
Chapmanite, 4 mm, Pernek, Slovakia

Braunite, 60 mm, Langban, Sweden

Braunite

$Mn^{2+}Mn^{3+}_6SiO_{12}$

TETRAGONAL ● ● ●

Properties: C – black, black-gray, black-brown: S – gray; L – submetallic to metallic: D – opaque; DE – 4.7; H – 6-6.5; CL – perfect; F – uneven to conchoidal; M – dipyramidal crystals, granular.
Origin and occurrence: Metamorphic in regionally and contact metamorphosed rocks; hydrothermal in sedimentary rocks rich in Mn and in hydrothermal veins, associated with hausmanite, pyrolusite and other Mn minerals. Its perfect crystals up to 70 mm (2¾ in) long, come from Kacharhavee and Tirodi, India. It is also known from Langban, Sweden; Ilfeld and Ilmenau, Germany; St. Marcel, Italy; Tizi Bashkun, Morocco.
Application: Mn ore.

Thaumasite

$Ca_6Si_2(CO_3)_2(SO_4)_2(OH)_{12} \cdot 24 H_2O$

HEXAGONAL ● ●

Properties: C – white to colorless; S – white; L – vitreous to dull; D – transparent to translucent; DE – 1.9; H – 3.5; CL – imperfect; F – uneven; M B acicular aggregates, granular, massive; LU – white.
Origin and occurrence: Hydrothermal or metamorphic in contact metamorphosed carbonate rocks, usually associated with other Ca silicates and carbonates, like ettringite and prehnite. Typical localities are Crestmore, California; West Paterson, New Jersey, USA; Langban, Sweden; N'Chwaning mine No. 2, Kuruman, South Africa.

Titanite

$CaTiOSiO_4$

MONOCLINIC ● ● ● ●

Properties: C – colorless, yellow, brown, green, gray to black; S – white; L – vitreous to dull; D – transparent, translucent to opaque; DE – 3.5; H – 5-5.5; CL – good; F – uneven to conchoidal; M – tabular crystals and their combinations, granular, massive.
Origin and occurrence: Metamorphic and magmatic as a common accessory mineral in many igneous and

Thaumasite, 3 mm aggregate, Maglovec, Slovakia

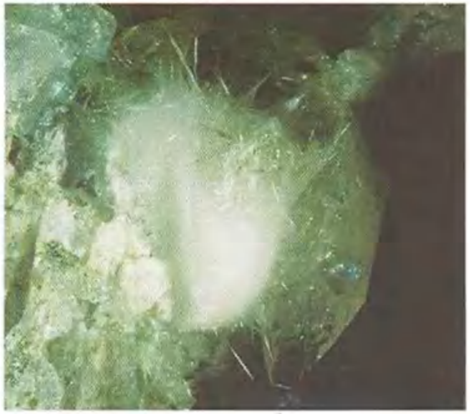

Fersmanite, 10 mm x, Khibiny Massif, Kola, Russia

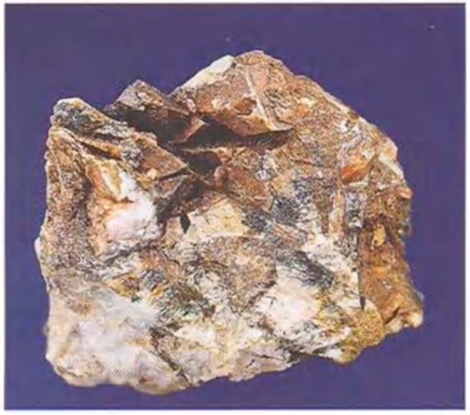

Titanite, 20 mm xx, Stubachtal, Austria

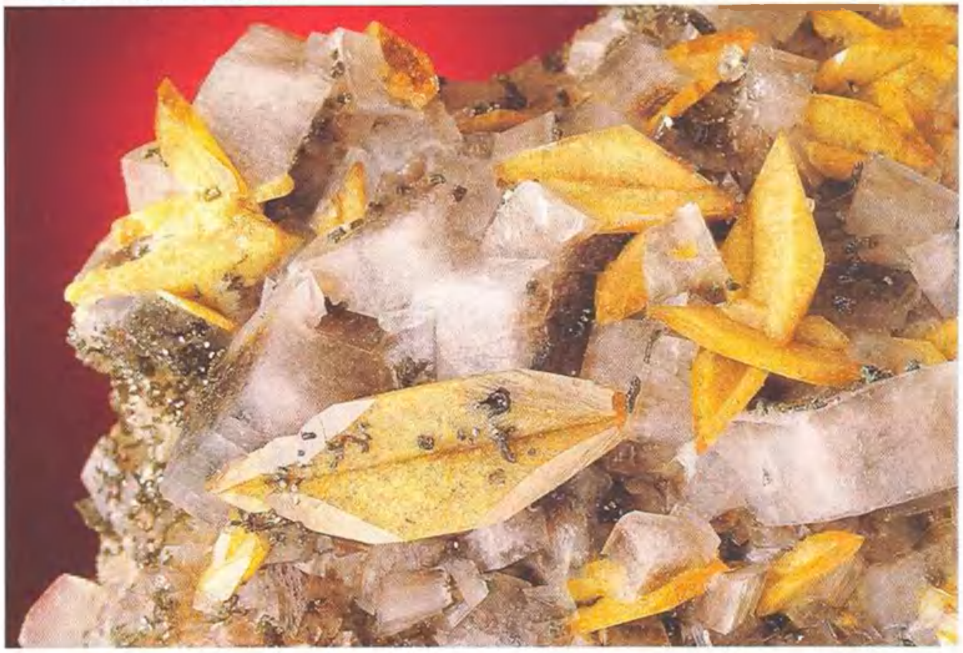

metamorphic rocks and pegmatites; hydrothermal in the Alpine-type veins; also in placers.
The most beautiful crystals up to 180 mm ($7^{1}/_{16}$ in) long occur in Alpine-type veins in Tavetsch and Binntal, Switzerland; Zillertal and Felbertal, Austria; Dodo, Polar Ural, Russia. Large, poorly developed crystals weighing up to 40 kg (88 lb), come from Eganville, Ontario, Canada and Rossie, New York, USA.

Chloritoid, 30 mm, Ile de Croix, France

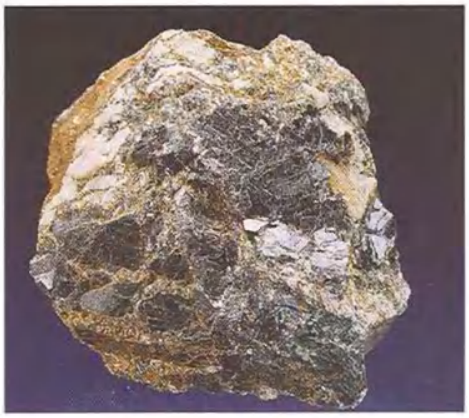

Fersmanite

$(Na,Ca)_4(Ti,Nb)_2Si_2O_{11}(OH,F)_2$

TRICLINIC ● ●

Properties: C – light to dark brown; S -white; L – vitreous to dull; D – translucent; DE – 3.5; H – 5-5.5; CL – none; F – uneven to conchoidal; M – tabular crystals.
Origin and occurrence: Magmatic to hydrothermal in alkaline pegmatites together with pectolite, apatite and sulfides. Crystals up to 30 mm ($1^{3}/_{16}$ in) across known from Mount Eveslogchorr, Khibiny massif, Kola Peninsula, Russia.

Chloritoid

$Fe_2Al_4Si_2O_{10}(OH)_2$

MONOCLINIC, TRICLINIC ● ● ●

Properties: C – dark gray, gray-green to black-green; S – gray; L – vitreous to dull; D – translucent; DE – 3.6; H – 6.5; CL – perfect; F – uneven to conchoidal; M – tabular crystals, foliated and scaly aggregates, granular.
Origin and occurrence: Metamorphic in some mica schists and phyllites; hydrothermal alteration product in lavas. Typical localities are Zermatt, Switzerland; Ottrez, Belgium; Pregraten, Austria.

Datolite, 30 mm xx, Dalnegorsk, Russia

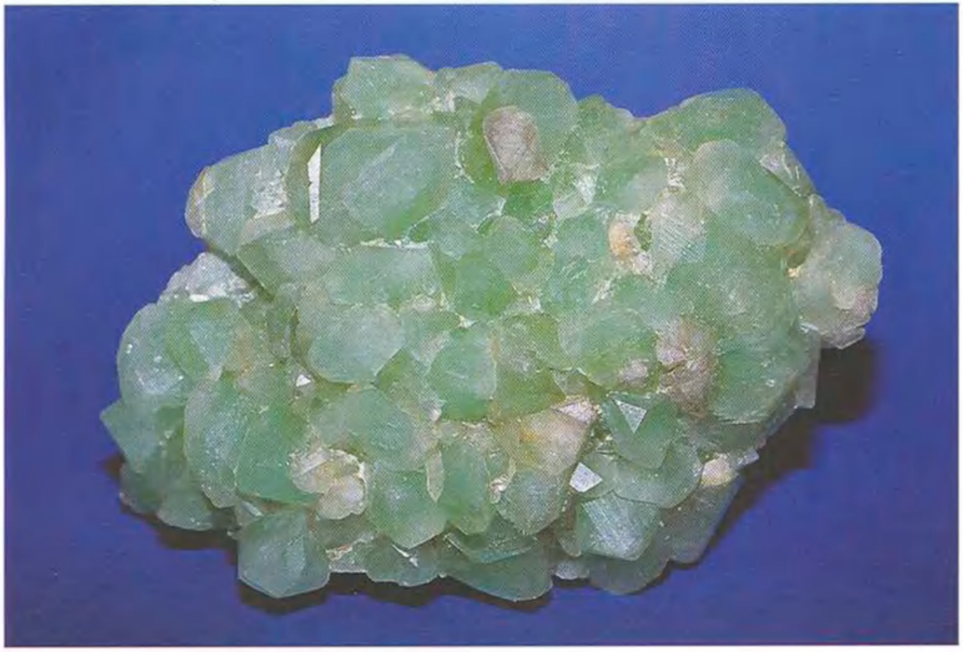

Datolite

$CaBSiO_4(OH)$

MONOCLINIC ● ● ●

Properties: C – white to colorless, yellowish, greenish, gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 5-5.5; CL – none; F – uneven to conchoidal; M – prismatic to tabular crystals, nodules, granular.

Origin and occurrence: Metamorphic and hydrothermal in contact metamorphosed rocks, in cavities in volcanic rocks, in ore veins and pegmatites. It is commonly associated with zeolites, prehnite, calcite and also with tourmaline. Prismatic and tabular crystals up to 100 mm (4 in) long occur in Dalnegorsk, Russia. Other typical localities are West Paterson, New Jersey; Keweenaw Peninsula, Michigan, USA; Haslach, Germany; Charcas, San Luis Potosi, Mexico.

Gadolinite-(Y)

$Y_2FeBe_2Si_2O_{10}$

MONOCLINIC ● ●

Properties: C – black, dark red, brown, greenish; S – gray-green; L – vitreous to greasy, D – transparent to translucent; DE – 4.4; H – 6.5-7; CL – none; F –

Gadolinite-(Y), 60 mm, Hitterö, Norway

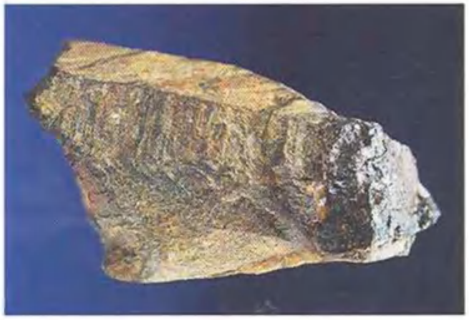

Howlite, 30 mm, California, U.S.A.

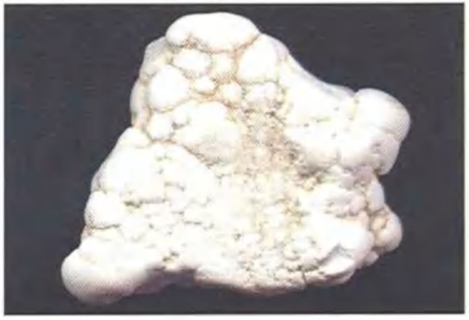

Dumortierite, 60 mm, Dehesa, U.S.A.

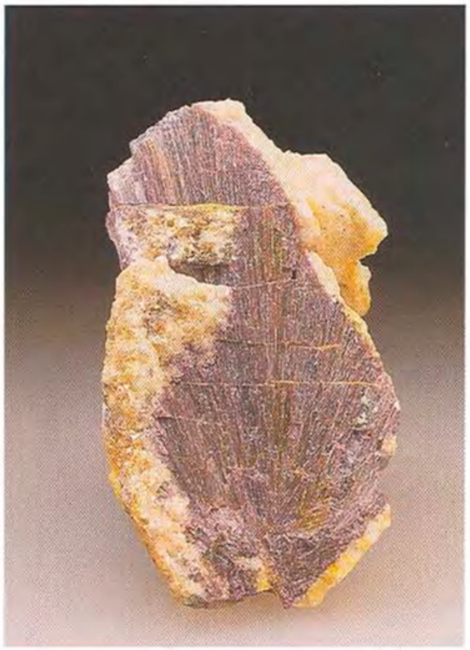

conchoidal; M – prismatic crystals, granular; R – locally metamict.
Origin and occurrence: Magmatic in granites and pegmatites, locally associated with fluorite and allanite; rare hydrothermal in the Alpine-type veins. Aggregates, up to 40 cm (15¾ in) across come from Hittero, Norway. Barringer Hill and Clear Creek, Texas, USA yielded aggregates, weighing up to 90 kg (198 lb). Other famous localities are Blatchford Lake, Northwest Territories, Canada; Baveno, the Alps, Italy; Iveland, Evje, Norway.

Kornerupine

$Mg_4Al_6(Si,Al,B)_5O_{21}(OH)$

ORTHORHOMBIC ● ●

Properties: C – colorless, white, greenish, gray, brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 6.5-7; CL – none; F – conchoidal; M – prismatic crystals, radial aggregates, granular.
Origin and occurrence: Metamorphic in strongly metamorphosed rocks, as granulites. Crystals up to 230 mm ($9\frac{1}{16}$ in) across known from Fiskenaesset, Greensland. It also occurs in Lac Ste-Marie, Quebec, Canada; Waldheim, Germany; Itrongay, Madagascar.

Dumortierite

$(Al,Ti,Mg)Al_6BSi_3O_{16}(O,OH)_2$

ORTHORHOMBIC ● ● ●

Properties: C – purple, pink, blue, brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 8.5; CL – good; F – uneven; M – prismatic crystals, radial and acicular aggregates, massive.
Origin and occurrence: Metamorphic in some Al-rich metamorphic rocks, e.g. in migmatites and gneisses; magmatic in granites and pegmatites; hydrothermal in altered rocks. It comes from Dehesa, California; Rochester, Nevada, USA; Beaunan, France; crystals are known from the vicinity of Kutna Hora, Czech Republic; Soavina, Madagascar.

Howlite

$Ca_2B_5SiO_9(OH)_5$

MONOCLINIC ● ●

Properties: C – white; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.5; H – 3.5; CL – good; F – uneven; M – tabular crystals, nodules, massive.
Origin and occurrence: Hydrothermal in borate deposits, associated with ulexite and colemanite. Crystals, several cm across, come from the vicinity of Bras d'Or Lake, Nova Scotia, Canada. It occurs also in Lang and Daggett, California, USA.

Kornerupine, 90 mm, Lac Ste.-Marie, Canada

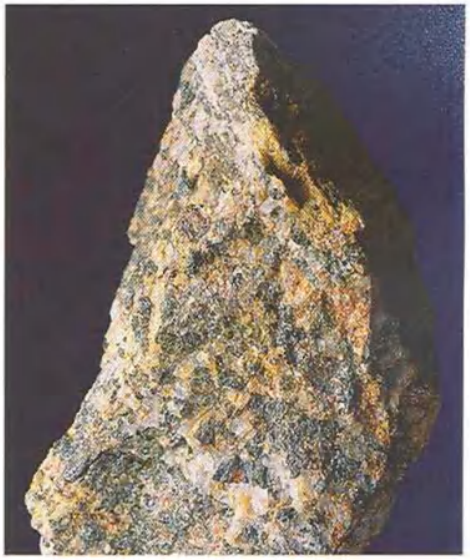

Cuprosklodowskite, 82 mm, Musonoi, Zair

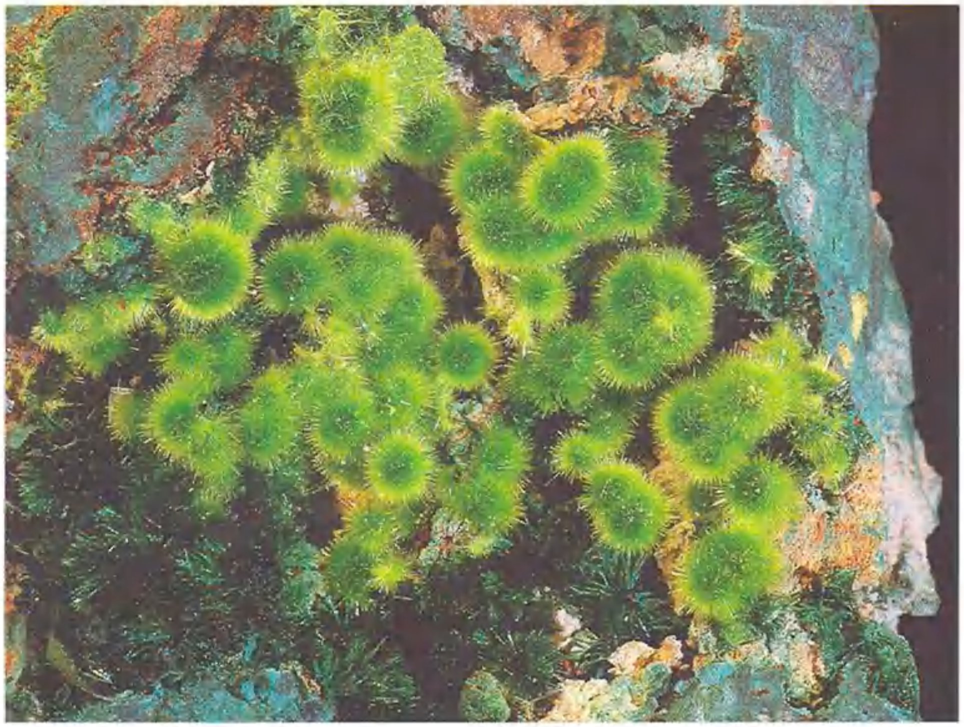

Cuprosklodowskite

$Cu(UO_2)_2Si_2O_7 . 6 H_2O$

MONOCLINIC ● ●

Properties: C – various hues of green; S – greenish; L – vitreous to dull; D – transparent to translucent; DE – 3.8; H – 4; CL – good; F – uneven; M – radial, acicular aggregates, thin coatings, granular, massive; R – strong radioactive.

Origin and occurrence: Secondary in the oxidation zone of U deposits, associated with autunite, torbernite, uranophane and other U secondary minerals. Crystals, up to several cm long, come from

Uranophane, 125 mm, Brewster, Texas, U.S.A.

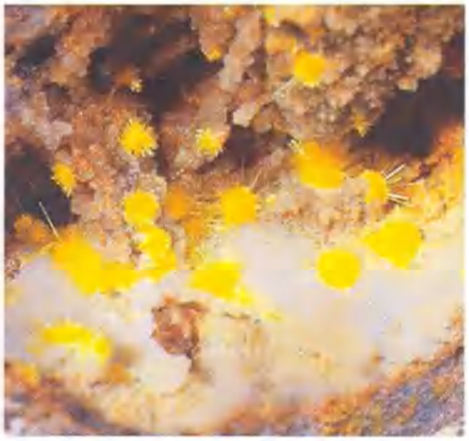

Kasolite, 30 mm, Musonoi, Zair

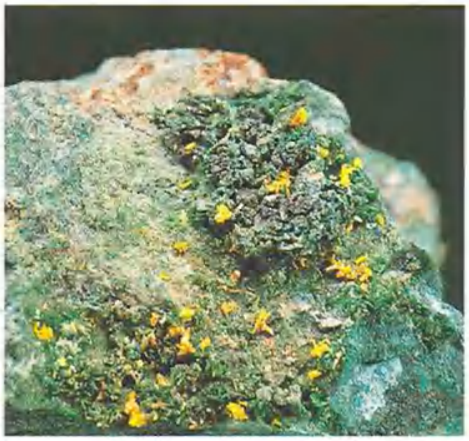

the Mashamba West mine, Musonoi, Zaire. It also occurred in Jachymov, Czech Republic.

Uranophane

$Ca(UO_2)_2Si_2O_7 . 6 H_2O$

MONOCLINIC ● ● ●

Properties: C – various hues of yellow to brown; S – yellowish; L – vitreous to dull; D – transparent to translucent; DE – 3.9; H – 2.5; CL – good; F – uneven; M – radial, acicular aggregates, thin coatings, granular, massive; LU – yellow-green; R – strong radioactive.

Origin and occurrence: Secondary in the oxidation zone of U deposits as a product of uraninite alteration, associated with autunite, torbernite and other secondary U minerals. Needles up to 10 mm ($^3/_8$ in) long, come from Musonoi, Shinkolobwe, Zaire.

It is also known from the Faraday mine, Bancroft, Ontario, Canada. It occurs in Wolsendorf, Germany, as well as in Jachymov, Czech Republic.

Kasolite

$Pb(UO_2)SiO_4 . H_2O$

MONOCLINIC ● ●

Properties: C – various hues of yellow, green to brown; S – yellowish; L – vitreous to dull; D – translucent to opaque; DE – 6.2; H – 4.5; CL – good; F – uneven; M – prismatic crystals, radial and acicular aggregates, thin coatings, granular, massive; R – strong radioactive.

Origin and occurrence: Secondary mineral in the oxidation zone of U deposits. It is a product of uraninite alteration and is associated with uranophane, torbernite and U hydroxides. Crystals up to 10 mm ($^3/_8$ in) long occur in Shinkolobwe, Kasolo, Zaire and Mounana, Gabon.

Akermanite

$Ca_2MgSi_2O_7$

TETRAGONAL ● ● ●

Properties: C – white, gray, green, brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.9; H – 5.5; CL – good; F – uneven to conchoidal; M – short prismatic crystals, granular.

Origin and occurrence: Magmatic in volcanic basic rocks; metamorphic in contact metamorphosed marbles. A typical representative of localities in marbles is Crestmore, California, USA. It occurs also in Ca-rich volcanic rocks in Velardena, Mexico.

Akermanite, 5 mm xx, Mt. Vesuvius, Italy

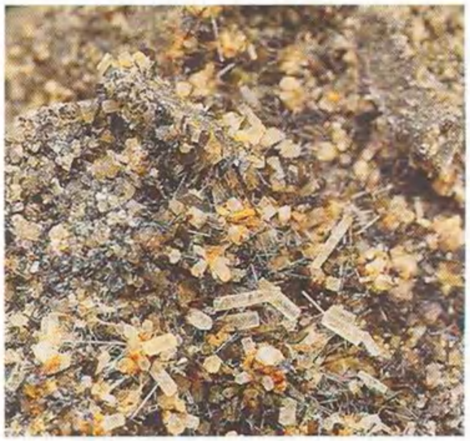

Gehlenite

$Ca_2Al_2SiO_7$

TETRAGONAL ● ● ●

Properties: C – white, gray, yellowish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 5-6; CL – good; F – uneven to conchoidal; M – short prismatic crystals, granular.

Origin and occurrence: Magmatic in volcanic basic rocks; metamorphic in contact metamorphic marbles.

It occurs in Crestmore, California, USA; Monzoni, Italy; Oraviza, Romania and elsewhere.

Gehlenite, 80 mm, Vata de Sus, Romania

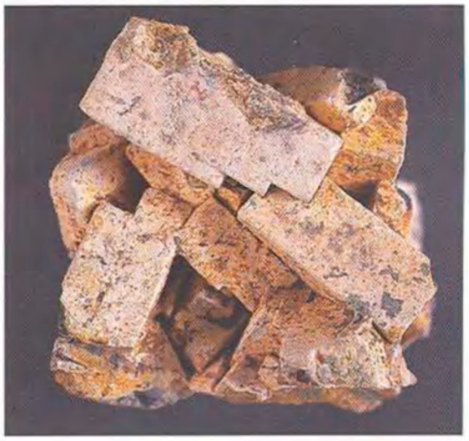

Ilvaite, 20 mm xx, Dalnegorsk, Russia

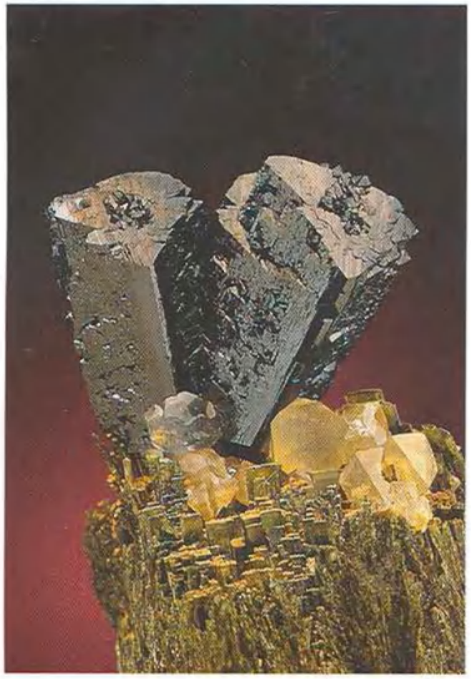

Bertrandite, 36 mm, Mt.Antero, U.S.A.

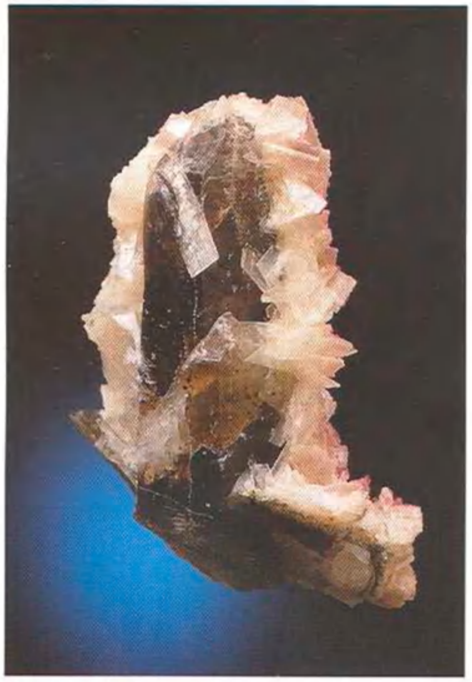

Ilvaite

$CaFe^{2+}_2Fe^{3+}Si_2O_8(OH)$

ORTHORHOMBIC ● ● ●

Properties: C – black to black-gray; S – black; L – submetallic to dull; D – opaque; DE – 4.1; H – 5.5-6; CL – good; F – uneven; M – prismatic crystals, granular.
Origin and occurrence: Metamorphic and hydrothermal in contact metamorphosed Fe, Zn and Cu deposits. Prismatic crystals up to 100 mm (4 in) long come from Rio Marina, Elba, Italy. Crystals up to 30 cm (12 in) long are known from Serifos Island, Greece. Crystals, several cm long occur also in Dalnegorsk, Russia and in the Laxey mine, Idaho, USA.

Bertrandite

$Be_4Si_2O_7(OH)_2$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.6; H – 6-7; CL – perfect; F – uneven; M – tabular crystals and their combinations, radial aggregates, granular, massive.
Origin and occurrence: Hydrothermal in granitic pegmatites, greisens and in hydrothermal veins, together with beryl, also as pseudo-morphs after beryl. Tabular crystals up to 50 mm (2 in) across occur in Conselheira Pena, Minas Gerais, Brazil. It is known from Kounrad and Kara-Oba, Kazakhstan in crystals up to 30 mm (1³/₁₆ in) across. It also comes from Stoneham, Maine, USA; Písek, Czech Republic and Iveland, Norway.
Application: the most important Be ore.

Hemimorphite

$Zn_4Si_2O_7(OH)_2 . H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish, greenish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 4.5-5; CL – perfect; F – uneven; M – tabular crystals and their combinations, botryoidal and radial aggregates, granular, massive.
Origin and occurrence: Secondary in the oxidation zone of Zn deposits, associated with sphalerite, smithsonite and cerussite. Crystals up to 100 mm (4 in) long come from Bisbee, Arizona, USA.. It is also from the El Potosi Mine, Santa Eulalia, Chihuahua, Mexico; Bleiberg, Austria; Cho-Dien, Vietnam.
Application: Zn ore.

Hemimorphite, 30 mm xx, Mapimi, Mexico

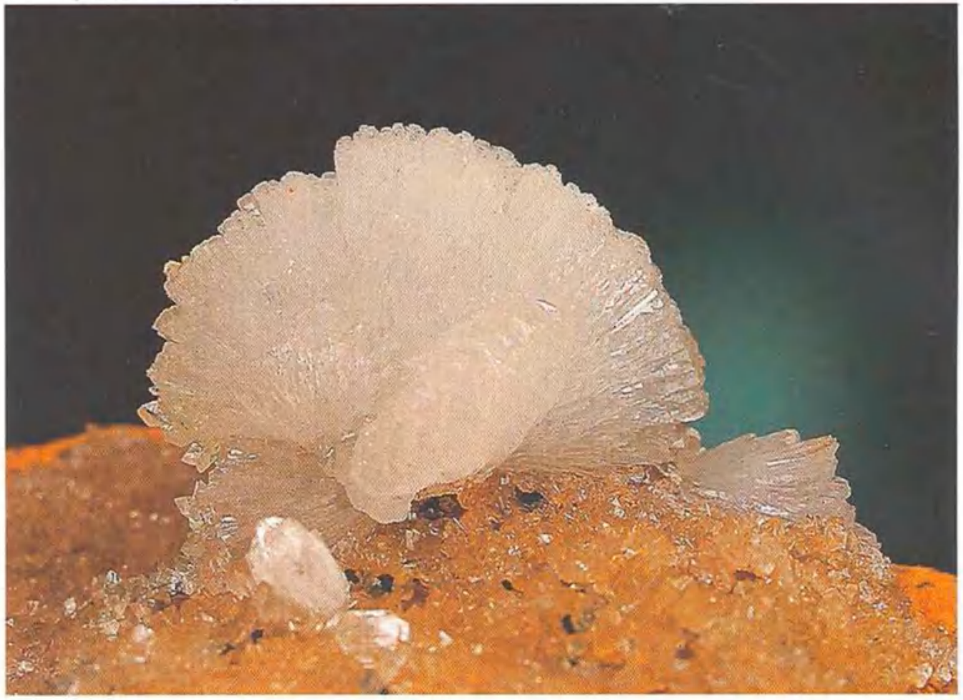

Lamprophylite

$Na_2Sr_2Ti_3Si_4O_{16}(OH,F)_2$

MONOCLINIC ● ●

Properties: C – brown to dark brown; S – white; L – vitreous to submetallic; D – translucent; DE – 3.5; H – 2-3; CL – perfect; F – uneven; M – tabular crystals, radial aggregates.

Origin and occurrence: Magmatic in alkaline syenites and their pegmatites, associated with nepheline, aegirine and eudialyte. The best crystals up to 150 mm (6 in) long come from Mount Flora, Lovozero massif, Kola Peninsula, Russia. It is also known from Langesundsfjord, Norway and Mont St.-Hilaire, Quebec, Canada.

Hemimorphite, 26 mm, Arizona, U.S.A.

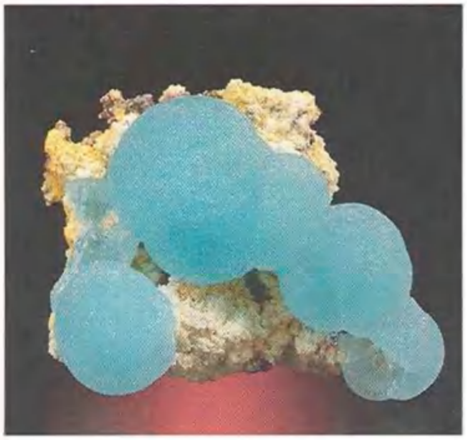

Lamprophyllite, 50 mm xx, Lovozero Massif, Kola, Russia

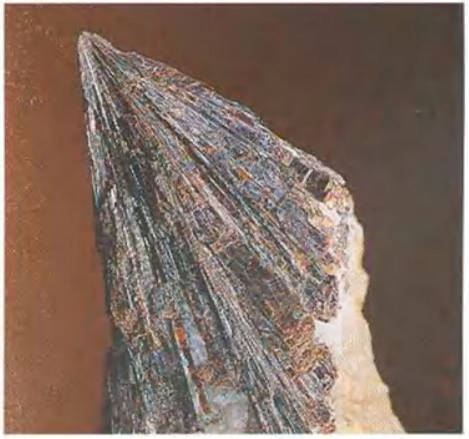

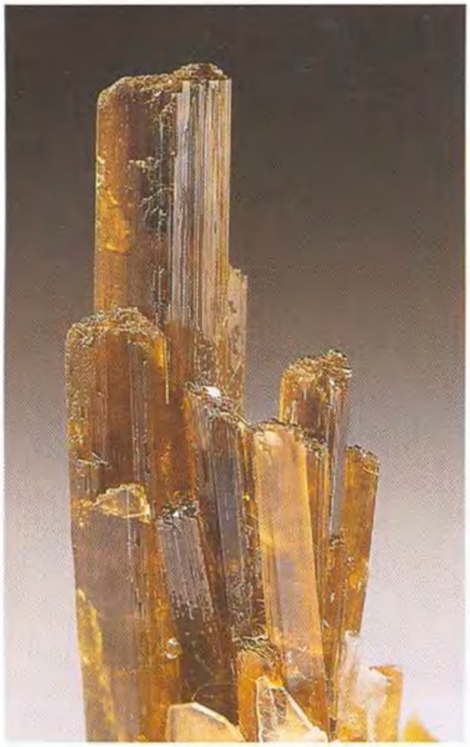
Clinozoisite, 36 mm, Eden Mills, U.S.A.

Clinozoisite

$Ca_2Al_3Si_3O_{12}(OH)$

MONOCLINIC ● ● ● ●

Properties: C – colorless, yellowish, green, pink; S – white; L – vitreous to dull; PS – transparent to translucent; DE – 3.4; H – 6.5; CL – good; F – uneven; M – prismatic crystals, columnar. radial aggregates, granular, massive.
Origin and occurrence: Metamorphic and hydrothermal in contacts of marbles and in the Alpine-type veins. Prismatic crystals up to 60 mm ($2\frac{3}{8}$ in) long come from Radoy Island, Norway. It comes also from Eden Mills, Vermont; Allens Park, Colorado, USA; Pinos Altos, Baja California, Mexico.

Epidote

$Ca_2(Al,Fe^{3+})_3Si_3O_{12}(OH)$

MONOCLINIC ● ● ● ● ●

Properties: C – green, brown, greenish, yellow-green; S – white; L – vitreous to dull; D – transparent to translucent, locally opaque; DE – 3.4; H – 6-7; CL – good; F – uneven; M – prismatic crystals, columnar and radial aggregates, granular, massive.
Origin and occurrence: Metamorphic in marble/granite contacts; hydrothermal in the Alpine-type veins and in hydrothermally altered rocks, typically associated with albite, prehnite and amphibole. Its green crystals up to 140 mm (5½ in) long, were found in Sobotín, Czech Republic. Perfect crystals up to 100 mm (4 in) across come from Knappenwand, Austria. Thick tabular crystals occur in Prince of Wales Island, Alaska, USA. Fine crystals, resembling Austrian crystals, were found recently in Alchuri, Shigar, Pakistan. Fine columnar aggregates of crystals are known from Pampa Blanca, Peru and also reported from Arendal, Norway.

Piemontite

$Ca_2(Al,Mn^{3+})_3Si_3O_{12}(OH)$

MONOCLINIC ● ●

Properties: C – red-brown to black, crimson, red-yellow; S – white; L – vitreous to dull; D – translucent, locally opaque; DE – 3.5; H – 6; CL – good; F – uneven; M – prismatic crystals, radial aggregates, granular.
Origin and occurrence: Metamorphic in shales and Mn-rich metamorphic rocks; rare magmatic in rhyolites and pegmatites. Needles up to 30 mm ($1\frac{3}{16}$

Epidote, Dashkesan, Azerbaidzhan

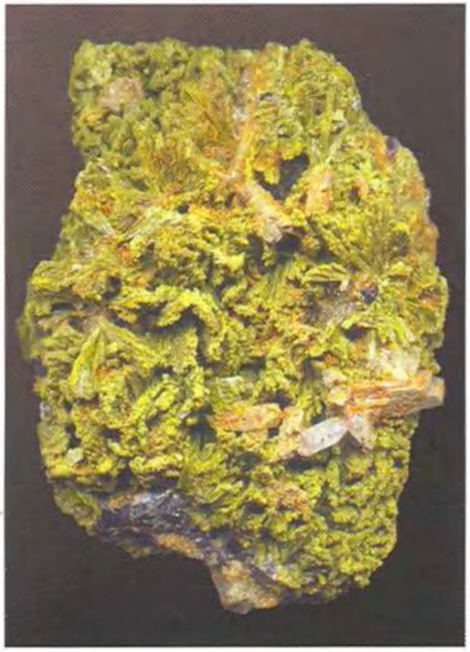

Epidote, 50 mm, Knappenwand, Austria

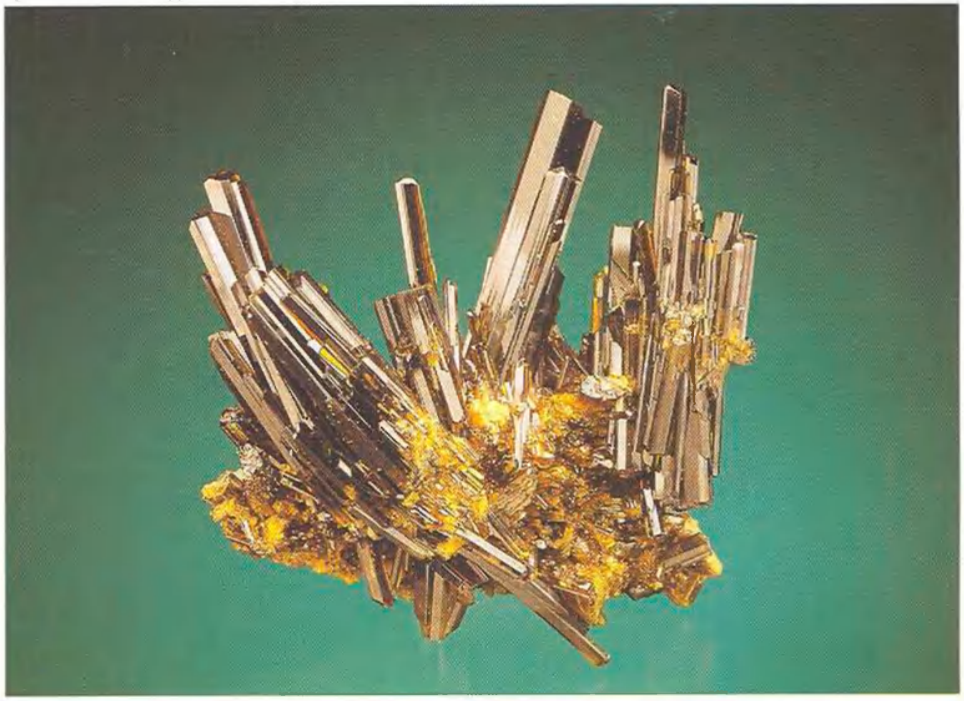

in) long come from St. Marcel, Piedmont, Italy and Otakiyama, Japan.

Allanit-(Ce)

$(Ca,Ce,Y)_2(Al,Fe^{3+})_3Si_3O_{12}(OH)$

MONOCLINIC ● ● ●

Properties: C – black to dark brown; S – light gray; L – greasy to submetallic; D – translucent to opaque; DE – 3.9; H – 5.5-6; CL – none; F – conchoidal to uneven; M – tabular crystals, granular; R – usually metamict.

Origin and occurrence: Magmatic in pegmatites and granites; metamorphic in various types of metamorphic rocks, e.g. migmatites, amphibolites and gneisses. Grains, up to 70 cm (27 9/16 in) across found in pegmatites near Bancroft, Ontario, Canada. Also occurs in Barringer Hill, Colorado; Amelia district, Virginia, USA; Arendal and Hittero, Norway; Ytterby and Riddarhyttan, Sweden; Yates mine, Quebec, Canada.

Piemontite, 45 mm, Todyryact, Morocco

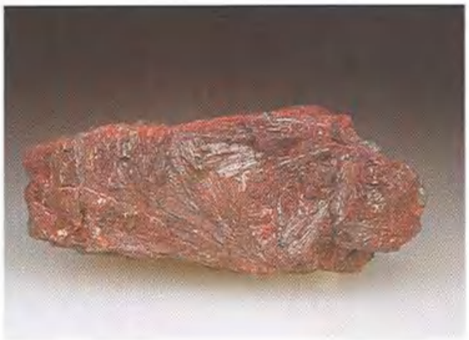

Allanite-(Ce), 40 mm, Vizcaya, Spain

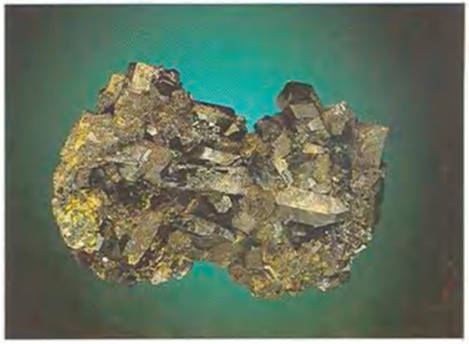

Tanzanite, 37 mm, Arusha, Tanzania

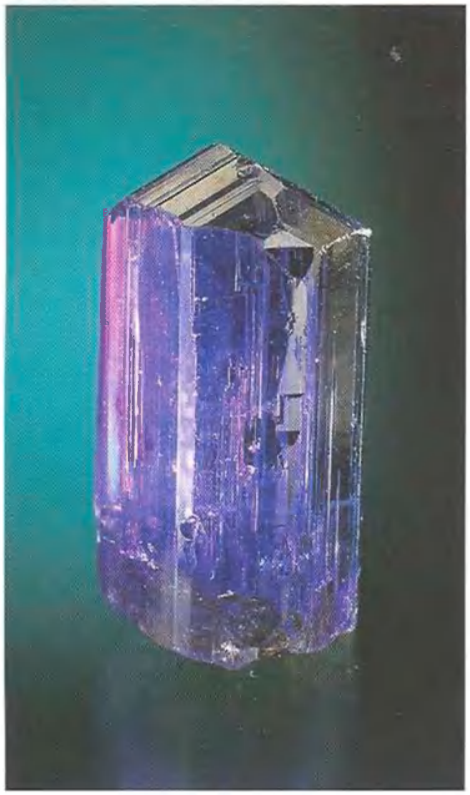

Zoisite, 60 mm, Weisenstein, Germany

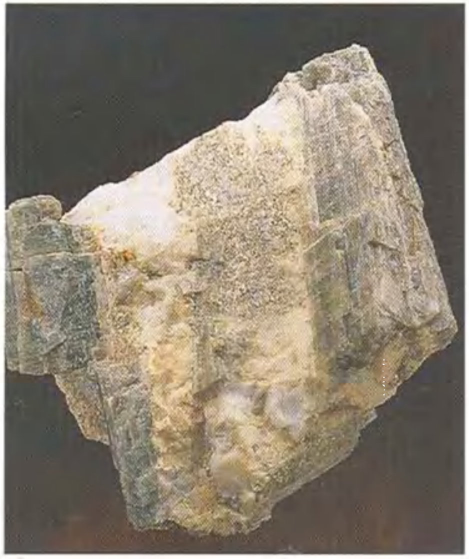

Zoisite

$Ca_2Al_3Si_3O_{12}(OH)$

ORTHORHOMBIC ● ● ●

Varieties: thulite, tanzanite

Properties: C – colorless, yellowish, green, pink, red (thulite), blue (tanzanite); S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 6.5; CL – good; F – uneven; M – prismatic crystals, radial aggregates, granular.
Origin and occurrence: Metamorphic in regionally metamorphosed Ca-rich rocks, mainly in pyroxenic gneisses, amphibolites and in marble contacts. It is known from many localities like Saualpe, Austria; Zermatt, Switzerland; Lexviken, Norway (thulite); Traversella, Italy; Alchuri, Pakistan; Merelani Hills, Arusha, Tanzania (tanzanite), where crystals up to 70 mm (2¾ in) long were found.
Application: tanzanite and thulite are cut as gemstones.

Thulite, 40 mm, Lexviken, Norway

Vesuvianite

$Ca_{19}(Al,Mg,Fe)_{13}Si_{18}O_{68}(OH,F,O)_{10}$

TETRAGONAL ● ● ●

Properties: C – brown, yellowish, green, blue, purple, colorless; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 6-7; CL – imperfect; F – uneven to conchoidal; M – prismatic to tabular crystals, columnar aggregates with radial structure, granular, massive.

Origin and occurrence: Metamorphic and hydrothermal in contact metamorphosed Ca-rich rocks, mainly in skarns, in marble contacts, also in rodingites; rarely magmatic in alkaline rocks. It is usually associated with grossular, wollastonite and diopside. Perfect green crystals up to 180 mm ($7\frac{1}{16}$ in) long and purple crystals up to 70 mm ($2\frac{3}{4}$ in) long come from the Jeffrey quarry, Asbestos, Quebec, Canada. It is also known from Hazlov, Czech Republic; Crestmore, California, Franklin, New Jersey, USA; Monzoni, Italy.

Vesuvianite, 110 mm, Asbestos, Canada

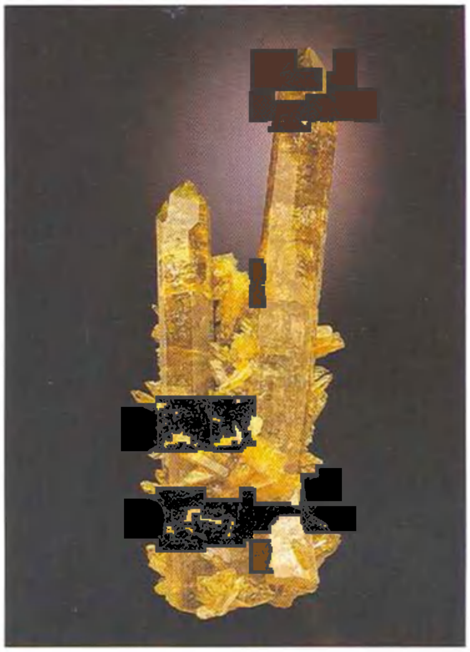

Viluite

$Ca_{19}(Al,Mg)_{13}B_5Si_{18}O_{68}(O,OH)_{10}$

TETRAGONAL ●

Properties: C – dark green, gray-brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 6; CL – imperfect; F – uneven to conchoidal; M – prismatic crystals.

Origin and occurrence: Metamorphic in serpentinized skarn, associated with grossular. Its perfect prismatic crystals up to 50 mm (2 in) long are only known from the Vilui River basin, Yakutia, Russia.

Viluite, 30 mm xx, Vilui River, Russia

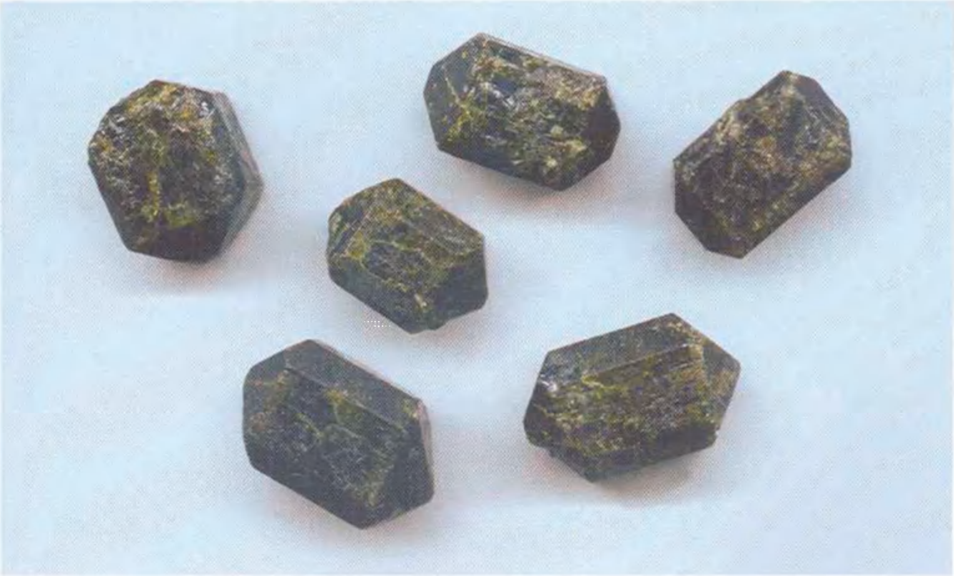

Benitoite

$BaTiSi_3O_9$

TRIGONAL ● ●

Properties: C – blue, pink, white, colorless; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.6; H – 6-6.5; CL – imperfect; F – uneven to conchoidal; M – prismatic to tabular crystals, mainly with trigonal cross-section; LU – bluish.

Origin and occurrence: Hydrothermal in veins, cross-cutting serpentinites, always associated with neptunite and natrolite. Its classic locality is the Benitoite Gem mine, San Benito Co., California, USA, where it forms tabular crystals up to 40 mm ($1^{9}/_{16}$ in) across.

Catapleiite

$Na_2ZrSi_3O_9 . 2 H_2O$

HEXAGONAL ● ●

Properties: C – light yellow, yellow-brown, pink, brown, blue; S – white; L – vitreous to dull; D – transparent to opaque; DE – 2.8; H – 5-6; CL – good; F – uneven; M – thin tabular crystals, lamellar aggregates.

Origin and occurrence: Magmatic in nepheline syenites and their pegmatites, together with aegirine, titanite, nepheline and microcline. It occurs in Mont St.-Hilaire, Quebec, Canada, as tabular crystals up to 150 mm (6 in) across. Crystals up to 30 mm ($1^{3}/_{16}$ in) across come from Mount Yukspor, Khibiny massif, Kola Peninsula, Russia. It is also known from Langesundsfjord, Norway; Magnet Cove, Arkansas, USA.

Eudialyte

$Na_4(Ca,Fe,Ce,Mn)_2ZrSi_6O_{18}(OH,Cl)$

TRIGONAL ● ●

Properties: C – red, pink to brown; S – white; L – vitrous to dull; D – transparent to translucent; DE – 2.8; H – 5-5.5; CL – imperfect; F – uneven; M – prismatic and tabular crystals, granular.

Origin and occurrence: Magmatic in nepheline syenites and their pegmatites, associated with aegirine, nepheline and microcline. Crystals up to 80 mm ($3^{1}/_{8}$ in) across come from Mount Kukisvumchorr, Khibiny massif, Kola Peninsula, Russia. Crystals up to 50 mm (2 in) across are also known from Mont St.-Hilaire, Quebec, Canada. It also occurs in Langesundsfjord, Norway and in Los Island, Guinea.

Ferroaxinite

$Ca_2FeAl_2BSi_4O_{15}(OH)$

TRICLINIC ● ● ●

Properties: C – brown to purple- brown, light purple; S – white; L – vitreous to dull; D – trans-

Benitoite, 28 mm x, San Benito Co., U.S.A.

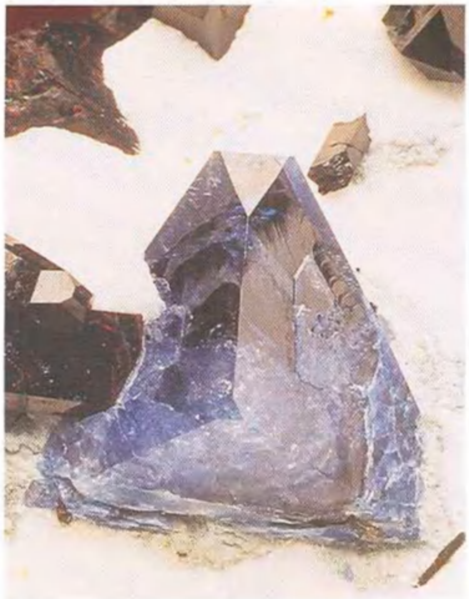

Catapleiite, 37 mm, Mont St.-Hilaire, Canada

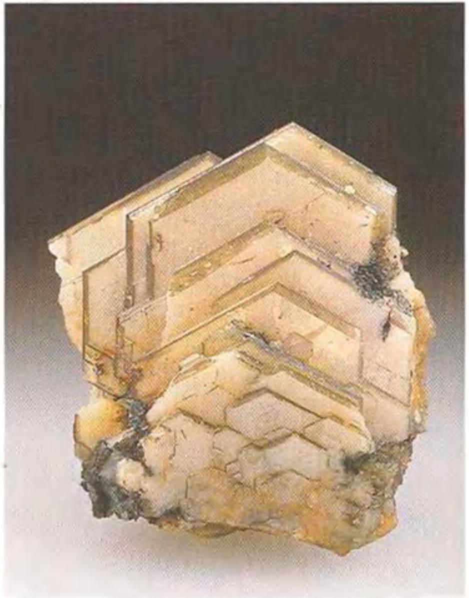

Eudialyte, 20 mm grain, Khibiny Massif, Kola, Russia

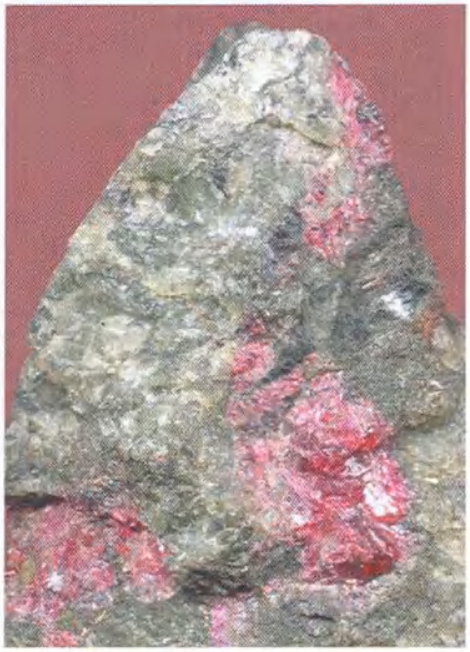

parent to translucent; DE – 3.3; H – 6. 5-7; CL – good; F – uneven to conchoidal; M – tabular crystals, platy aggregates, granular, massive.

Origin and occurrence: Metamorphic and hydrothermal in contacts of marbles and granites, associated with clinozoisite, prehnite, calcite and actinolite, also in the Alpine-type veins and pegmatites. Perfect tabular crystals up to 150 mm (6 in) across come from Puiva, Polar Ural, Russia. Other renowned localities are Obira, Japan; Bourg d'Oisans, France; Monte Scopi, Switzerland.

Tinzenite

$Ca(Mn,Fe)_2Al_2BSi_4O_{15}(OH)$

TRICLINIC ● ●

Properties: C – yellow, orange to red; S – white; L – vitrous to dull; D – transparent to translucent; DE – 3.3; H – 6.5-7; CL – good; F – uneven to conchoidal; M – tabular crystals, platy and fibrous aggregates, granular, massive.

Origin and occurrence: Hydrothermal in the Alpine-type veins, cross-cutting a rock, rich in braunite. It comes from Tinzen, Val d'Err, Switzerland and in the Cassagna mine, Genova, Italy.

Ferroaxinite, 62 mm, Khapalu, Pakistan

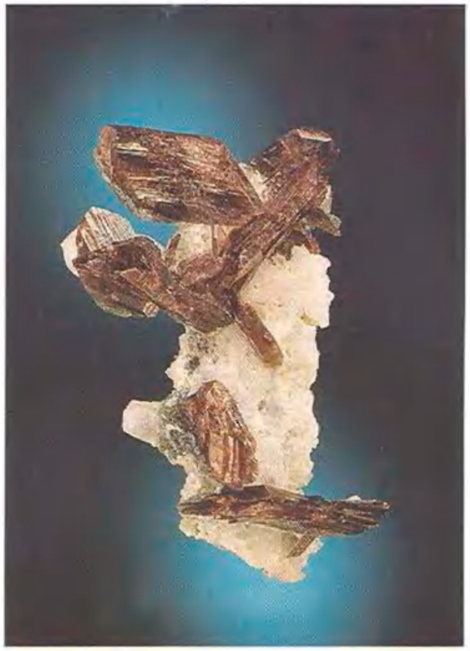

Tinzenite, 10 mm xx, Val Graveglia, Italy

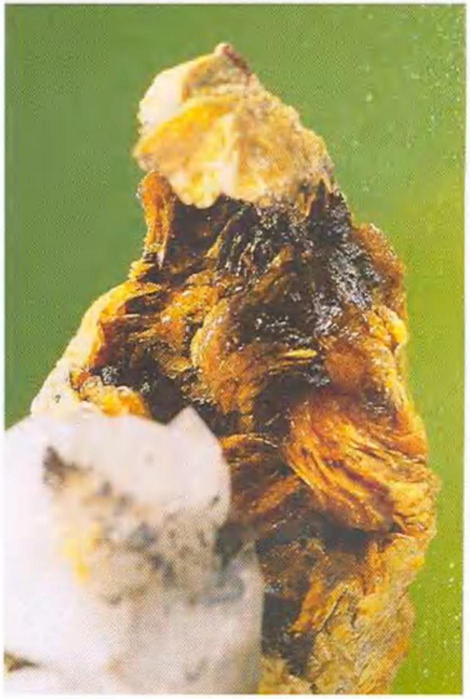

Beryl

$Be_3Al_2Si_6O_{18}$

HEXAGONAL ● ● ● ●

Varieties: emerald, aquamarine, heliodor, morganite, goshenite, red beryl (bixbite).

Properties: C – variable in different varieties; common beryl – mostly yellow, yellow-green, light green to white, rare blue; varieties: emerald – dark emerald-green; aquamarine – light to dark blue-green; morganite – pink; heliodor – light yellow to yellow-green; goshenite – white to colorless; red beryl B red; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.6; H – 7.5-8; CL – imperfect; F – uneven to conchoidal; M – long prismatic to tabular crystals, columnar and radial aggregates, granular, massive.

Origin and occurrence: Magmatic in pegmatites and granites; hydrothermal in greisens, in cavities in rhyolite, in quartz veins; metamorphic in mica schists. Perfect prismatic crystals of common beryl up to 9 m (29 ft 6 in) long found in the Etta mine, Keystone, South Dakota, USA. Crystals weighing up to 177 tons, come from Namivo, Alto Ligonha, Mozambique. Other localities are Pici, Brazil; Iveland, Norway; Antsirabe, Madagascar. Emerald crystals occur in mica-schists, marbles and ultrabasic rocks, associated with other Be minerals, phenakite and

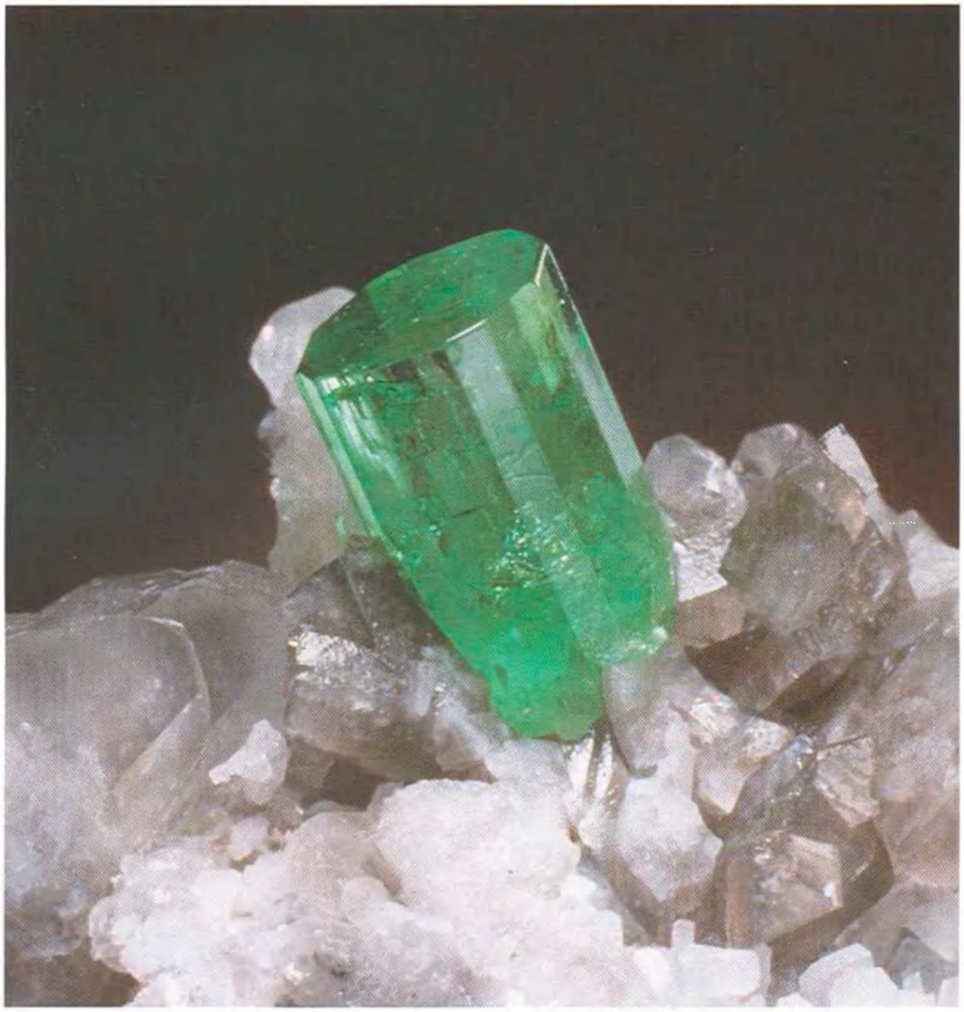

Emerald, 30 mm x, Boyaca, Colombia

Emerald, 30 mm, Ural Mts., Russia

Aquamarine, 97 mm, Koronel Hurta, Brazil

chrysoberyl. Beautiful dark green transparent crystals are known from Malyshevo, Ural Mts., Russia; Muzo and Coscuez, Colombia; Habachtal, Austria, where the largest crystals reach up to 120 mm ($4^{11}/_{16}$ in) in size. Aquamarine is mainly known from pegmatites and hydrothermal veins, commonly

Emerald, 90 mm, Muzo, Colombia

Aquamarine, 158 mm, Shingus, Pakistan

Heliodor, 87 mm, Volodarsk Volynskii, Ukraine

associated with tourmaline, quartz and albite. Intense blue crystals are known from many pegmatites in Minas Gerais, Brazil, where in the Marambaia mine, crystals, up to 70 cm ($27^{9}/_{16}$ in) long and weighing up to 110 kg (242 lb) found. Very beautiful gemmy crystals over 30 cm (12 in) long recently occurred in the Medina Mine. It also come from Murzinka, Ural mountains and Adun Chilon, Siberia, Russia. It also comes from Spitzkopje, Namibia; Gilgit, Pakistan and elsewhere. Morganite is a typical mineral of granitic pegmatites, where it occurs in mainly in cavities, usually associated with color varieties of tourmaline, quartz and albite. Its tabular crystals up to 100 mm (4 in) across found in the White Queen mine, Pala, California, USA. Smaller crystals come from San Piero in Campo, Elba, Italy. Crystals up to 50 cm (20 in) across reported from several localities in Minas Gerais, Brazil. Heliodor also occurs in pegmatite cavities, hydrothermal veins and in gneisses, commonly associated with quartz and albite. Prismatic crystals up to 200 mm ($7^{7}/_{8}$ in) long, come from Volodarsk Volynskii, Ukraine. It is also known from Nerchinsk, Siberia, Russia and several mines in Minas Gerais, Brazil. Goshenite occurs only in pegmatites. Its prismatic crystals come from Goshen, Massachusetts,

Heliodor, 26 mm x, Pamir, Tadzhikistan

Goshenite, 39 mm, Apalygun, Pakistan

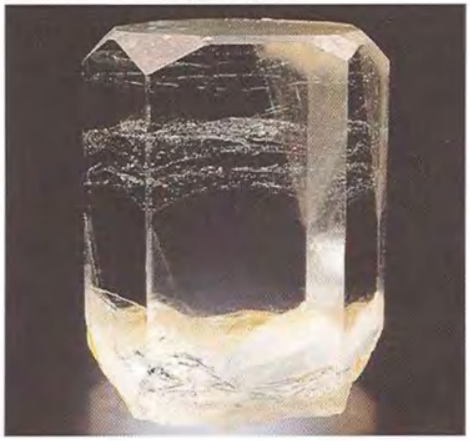

Red beryl, 17 mm x, Wah Wah Mts., U.S.A.

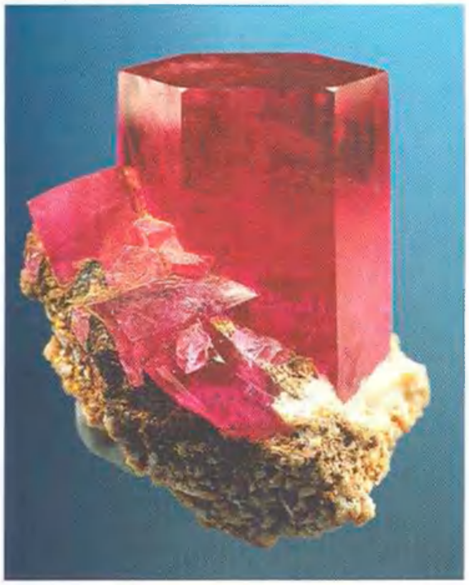

USA; San Piero in Campo, Elba, Italy. Red beryl occurs in Violet Claims, Wah Wah mountains, Utah, USA, where crystals up to 50 mm (2 in) were found. *Application:* Be ore, color varieties are cut as gemstones.

Morganite, 43 mm x, San Diego Co., U.S.A.

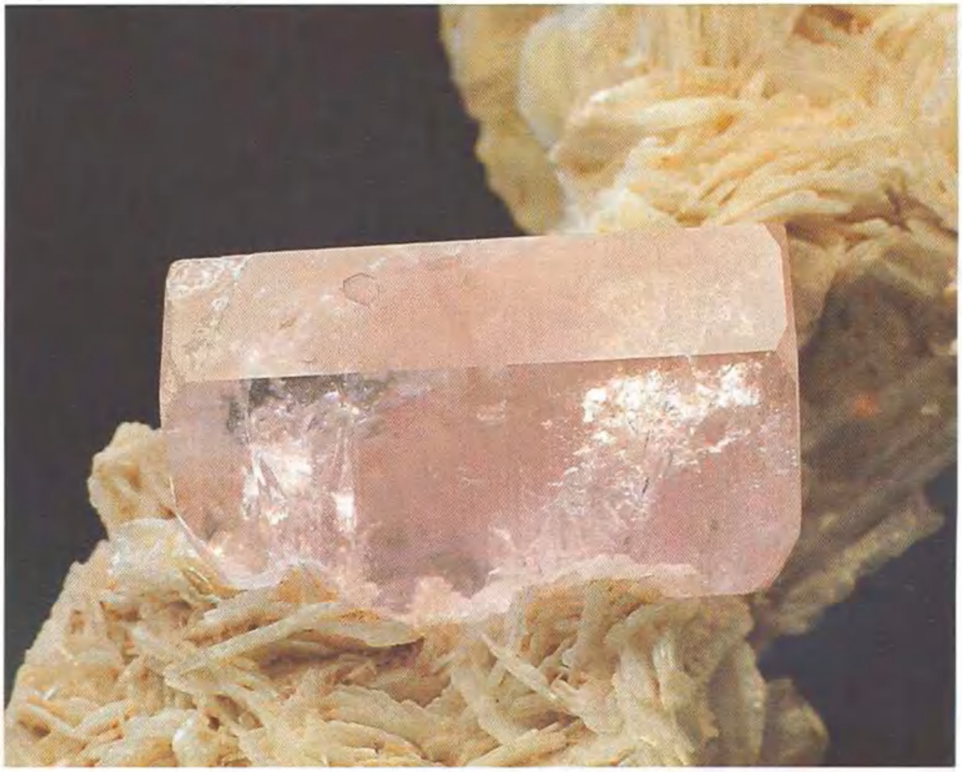

Bazzite, 70 mm, Tordal, Norway

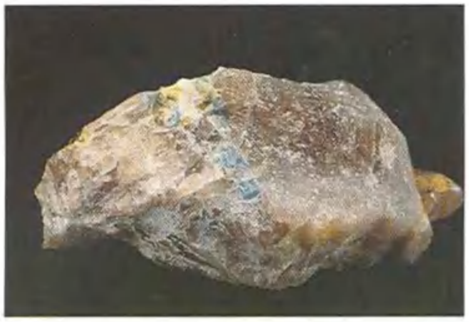

Bazzite

$Be_3(Sc,Al)_2Si_6O_{18}$

HEXAGONAL ●

Properties: C – light to intense blue; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.8; H – 6.5; CL – imperfect; F – uneven to conchoidal; M – long prismatic crystals, columnar and radial aggregates.

Origin and occurrence: Hydrothermal in the Alpine-type veins and pegmatites. Crystals up to 20 mm ($^{25}/_{32}$ in) long come from Tørdal, Norway. It is also known from Lago Maggiore, Italy and St. Gotthard, Switzerland.

Cordierite

$Mg_2Al_4Si_5O_{18}$

ORTHORHOMBIC ● ● ● ●

Varieties: iolite (gemmy blue)

Sekaninaite, 40 mm x, Dolni Bory, Czech Republic

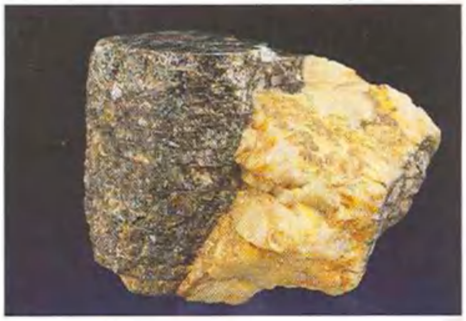

Properties: C – blue, purple, gray, gray-green, gray-brown, trichroic; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.5; H – 7-7.5; CL – imperfect; F – uneven to conchoidal; M – short prismatic crystals, granular.

Origin and occurrence: Metamorphic in migmatites and gneisses and in contact cherts; magmatic in granites and granitic pegmatites, usually associated with andalusite and sillimanite, also known from placers. It is a typical rock-forming mineral. Its prismatic crystals are very rare. Poorly developed transparent crystals up to 200 mm (7⅞ in) long come from Naverberg, Sweden. It also occurs in Orijarvi, Finland; Kragero, Norway; Bodenmais, Germany. Gemmy pebbles are known from the vicinity of Ratnapura, Sri Lanka.

Sekaninaite

$Fe_2Al_4Si_5O_{18}$

ORTHORHOMBIC ● ● ●

Properties: C – blue, purple, strongly pleochroic; S

Cordierite, 50 mm, Fishtail Lake, Canada

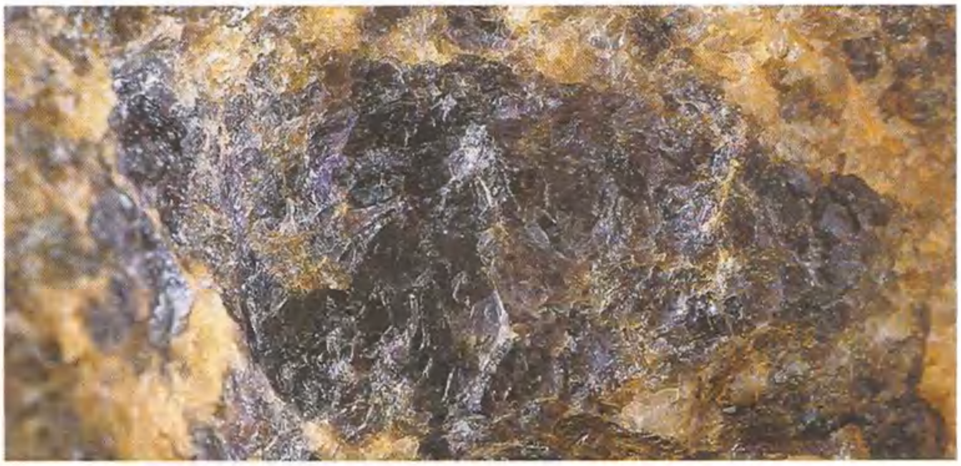

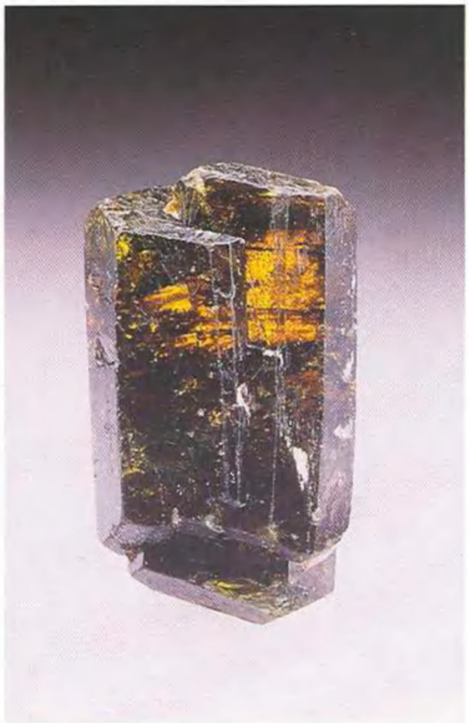

Dravite, 27 mm, Gujarkkot, Nepal

– white; L – vitreous to dull; D – transparent to translucent; DE – 2.8; H – 7-7.5; CL – imperfect; F – uneven to conchoidal; M – short prismatic crystals, granular.

Origin and occurrence: Magmatic in granitic pegmatites and some granites, associated with andalusite and tourmaline; metamorphic in gneisses and migmatites. Conical, imperfect crystals, up to 70 cm ($27\frac{9}{16}$ in) long typically come from Dolni Bory, Czech Republic; also known from San Piero in Campo, Elba, Italy.

Dravite

TOURMALINE GROUP

$NaMg_3Al_6(BO_3)_3Si_6O_{18}(OH)_4$

TRIGONAL ● ● ● ●

Properties: C – light to dark black-brown, blue, colorless, commonly pleochroic; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 7-7.5; CL – none; F – uneven to conchoidal; M – long to short prismatic crystals, columnar to acicular aggregates, granular.

Origin and occurrence: Metamorphic in migmatites, gneisses, mica schists, marbles and in contact metasomatic rocks; magmatic in some granitic pegmatites; hydrothermal in quartz veins and ore veins, also known from placers. Brown to dark brown perfect crystals up to 200 mm ($7\frac{7}{8}$ in) long come from mica schists near Dravograd, Slovenia. It is also known from marbles in Gouverneur, New York, USA. Crystals up to 150 mm (6 in) across found in Yinnietharra, Western Australia, Australia. Crystals up to 50 mm (2 in) long were recently found in Gujarkot, Nepal.

Application: transparent crystals are cut as gemstones.

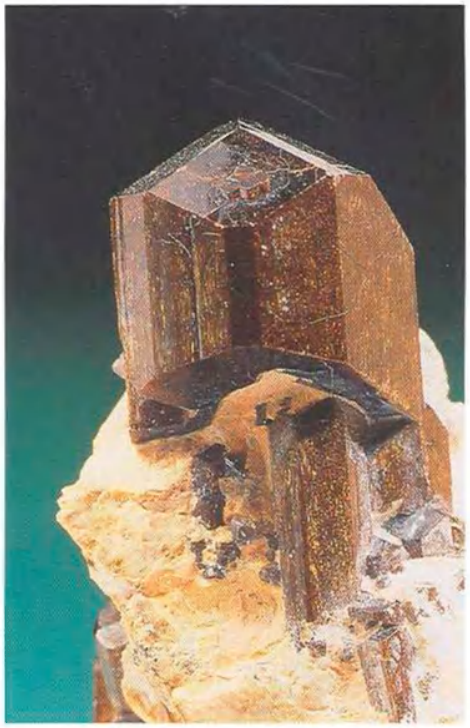

Buergerite, 12 mm x, San Luis Potosí, Mexico

Buergerite

TOURMALINE GROUP

$NaFe^{3+}{}_3Al_6(BO_3)_3Si_6O_{21}F$

TRIGONAL ●

Properties: C – black, strongly pleochroic; S – yellow-brown; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 7; CL – none; F – uneven to conchoidal; M B prismatic crystals, granular.

Origin and occurrence: Hydrothermal in rhyolites. Its black crystals, up to 40 mm (19/16 in) long, are known from Mexquitic, San Luis Potosi, Mexico.

Schorl, 40 mm, Gilgit, Pakistan

Schorl

TOURMALINE GROUP

$NaFe_3Al_6(BO_3)_3Si_6O_{18}(OH)_4$

TRIGONAL ● ● ● ●

Properties: C – black, black-brown, blue-black, strongly pleochroic; S – white; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 7-7.5; CL – none; F – uneven to conchoidal; M – long to short prismatic crystals, columnar to acicular aggregates, granular, massive.

Origin and occurrence: Magmatic in granites and granitic pegmatites; hydrothermal in greisens, in quartz and ore veins; metamorphic in migmatites, gneisses, mica schists and tourmalinites; also known from placers. It is usually associated with muscovite, quartz and albite. Perfect black crystals come from many pegmatite localities.

Its long prismatic crystals, up to 5 m long, come from Arendal, Norway. Very good crystals are also known from Kaatiala, Finland; Dolni Bory, Czech Republic; Conselheira Pena and Galileia, Minas Gerais, Brazil.

Povondraite

TOURMALINE GROUP

$NaFe^{3+}{}_3Mg_2Fe^{3+}{}_4(BO_3)_3Si_6O_{18}(OH)_4$

TRIGONAL ●

Properties: C – black; S – gray; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 7; CL –

Povondraite, 60 mm, Alto Chapare, Bolivia

Elbaite, 115 mm, Tourmaline Queen Mine, U.S.A.

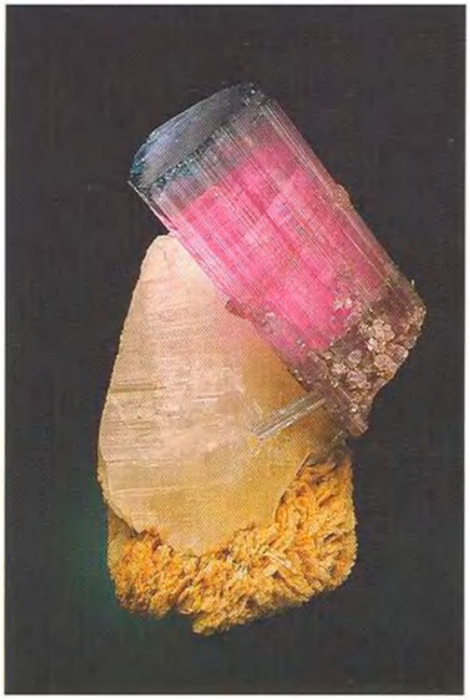

Elbaite, 153 mm, Afghanistan

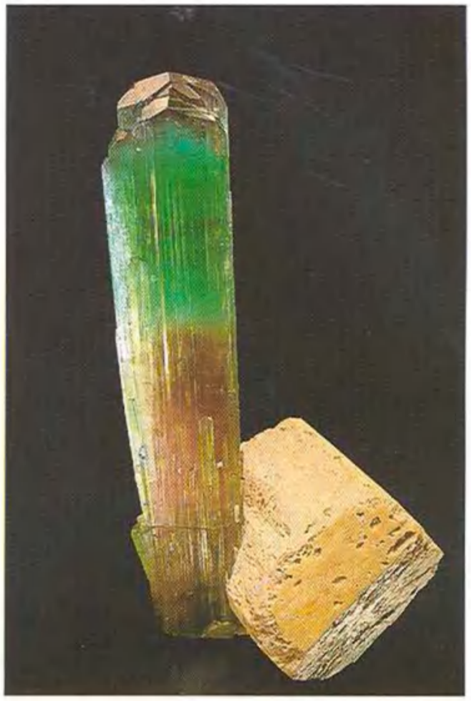

none; F – uneven to conchoidal; M – short prismatic crystals, granular.
Origin and occurrence: Hydrothermal along the cracks in metamorphosed evaporites. Its black crystals up to 10 mm ($^3/_8$ in) long typically come from Alto Chapare, Cochabamba, Bolivia.

Elbaite

TOURMALINE GROUP
$\mathrm{Na(Li_{1,5}Al_{1,5})Al_6(BO_3)_3Si_6O_{18}(OH)_3F}$

TRIGONAL ● ● ●

Varieties: rubellite, verdelite, indicolite, achroite

Properties: C – varies in different varieties, rubellite – pink to red; verdelite – various hues of green; indicolite – blue; achroite – colorless, other colors include yellow, brown and black; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 7; CL – none; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates, granular, massive.
Origin and occurrence: Almost only magmatic and hydrothermal in granitic pegmatites, typically associated with lepidolite and albite; also in placers. Its crystals are mostly known from pegmatite cavi-

Rubellite, 23 mm, Pala, U.S.A.

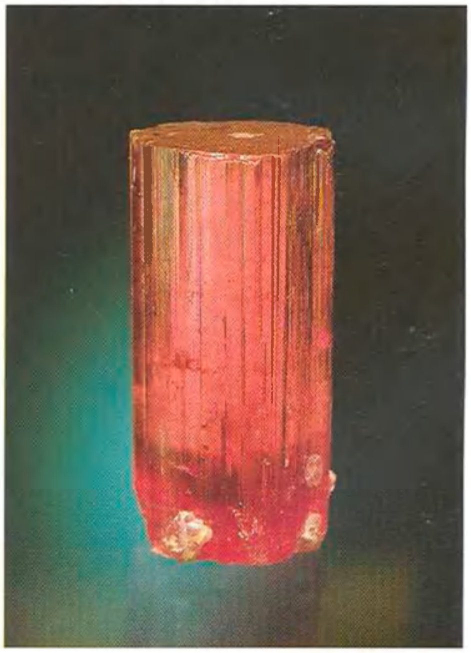

ties in many localities. Rubellite crystal called 'The Rocket' (109 cm/$43\frac{5}{16}$ in long) was found in the Jonas mine, Minas Gerais, Brazil. Rubellite crystals up to 40 cm ($15\frac{3}{4}$ in) long come from Alto Ligonha district, Mozambique. Rubellite crystals up to 250 mm ($9\frac{13}{16}$ in) long occurred in the Stewart Lithia, Tourmaline King and Tourmaline Queen mines, Pala, California, USA. Verdelite crystal 270 mm ($10\frac{5}{8}$ in) long, is known from the Dunton mine, Newry, Maine, USA. Other important localities include San Piero in Campo, Elba, Italy; Malkhan, Transbaikalia, Russia; Rožna , Czech Republic; Uto, Sweden; Gilgit, Pakistan; Paprok, Afghanistan.
Application: commonly used as a gemstone.

Uvite

TOURMALINE GROUP

$CaMg_4Al_5(BO_3)_3Si_6O_{18}(OH,F)_4$

TRIGONAL ● ● ●

Properties: C – gray, black, brown, green, red, pleochroic; S – white; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 7-7.5; CL – none; F – uneven to conchoidal; M – long to short prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Metamorphic in Ca-rich rocks, marbles, skarns; magmatic in some pegmatites; hydrothermal in ore veins.
Perfect green and red crystals, up to 30 mm ($1\frac{3}{16}$ in) across, occur in Brumado, Bahía, Brazil. It is also known from Gouverneur and Pierrepont, New York, USA.

Liddicoatite

TOURMALINE GROUP

$Ca(Li_2Al)Al_6(BO_3)_3Si_6O_{18}(OH)_3F$

TRIGONAL ● ● ●

Varieties: rubellite, verdelite

Properties: C – varies in different varieties, mainly pink, green, green-brown to yellow-brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 7; CL – none; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Magmatic in granitic peg-

Uvite, 41 mm, Brumado, Brazil

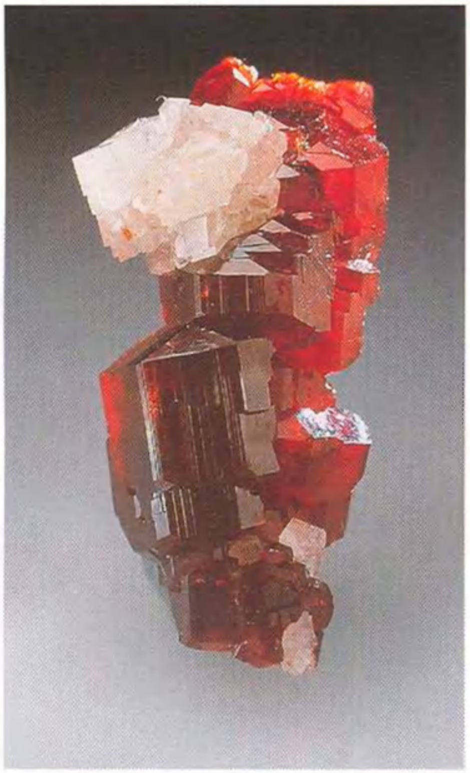

Liddicoatite, 60 mm, Anjanobonoina, Madagascar

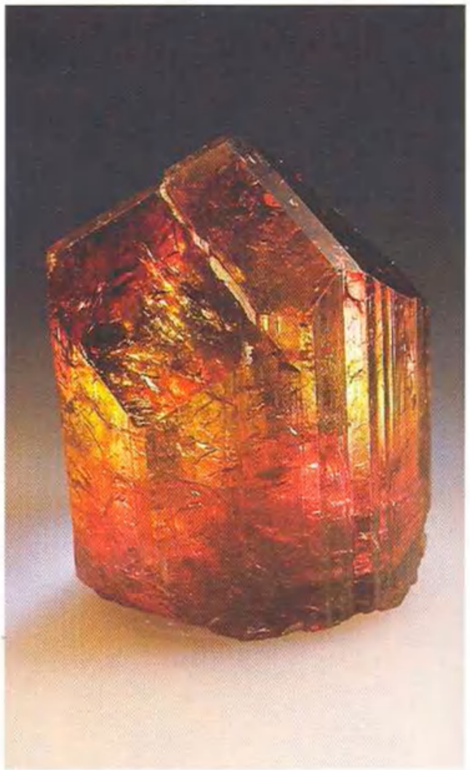

Verdelite, 24 mm, Gillette Quarry, U.S.A.

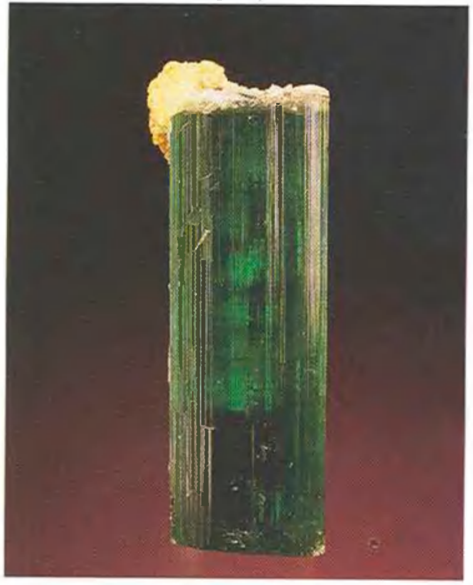

Indigolite, 32 mm, Esmeralda Mine, Mesa Grande, U.S.A.

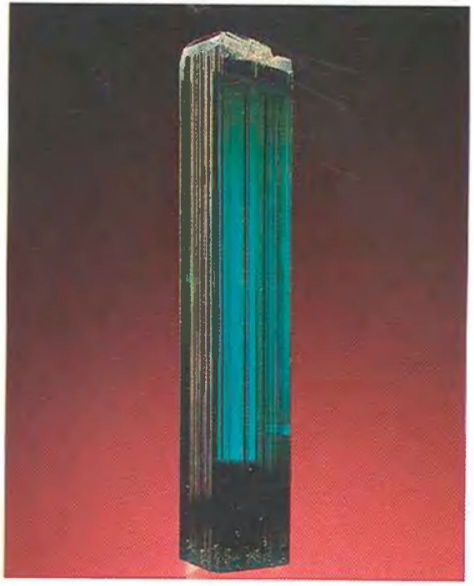

matites in association with lepidolite, spodumene and albite. Its perfect red and green crystals up to 250 mm ($9\frac{13}{16}$ in) long come from pegmatite cavities in many localities in Madagascar, e.g. Sahatany and Anjanabonoina. It is also known from Blizná, Czech Republic.
Application: cut as a gemstone.

Liddicoatite, 41 mm, Madagascar

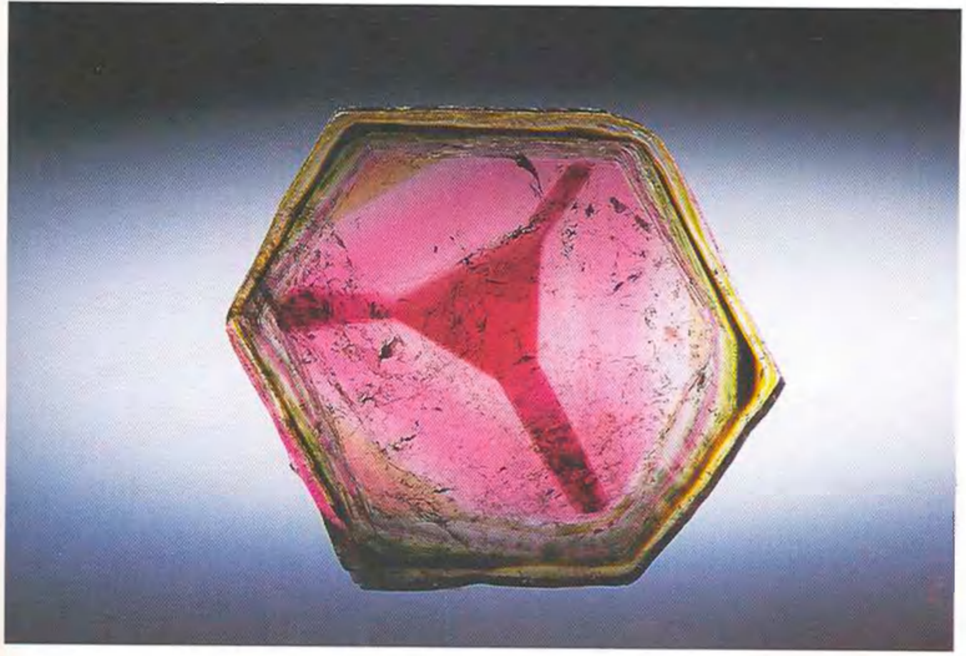

Foitite

TOURMALINE GROUP

☐ $Fe_2Al(Al_6(BO_3)_3Si_6O_{18}(OH)_4$

TRIGONAL ● ● ●

Properties: C – black to black-purple; S – white; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 7; CL – none; F – uneven to conchoidal; M – prismatic crystals, acicular aggregates.
Origin and occurrence: Magmatic in granitic pegmatites and granites; hydrothermal in pegmatite cavities. Foitite typically forms black tips of elbaite crystals (moors heads) in Dobra Voda, Czech Republic; San Piero in Campo, Elba, Italy; White Queen mine, Pala, California, USA. It occurs together with schorl in some granitic pegmatites, like Rozna, Czech Republic.

Rossmanite

TOURMALINE GROUP

☐ $LiAl_2Al_6(BO_3)_3Si_6O_{18}(OH)_4$

TRIGONAL ●

Properties: C – pink; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 7; CL – none; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates.
Origin and occurrence: Magmatic in granitic pegmatites. Prismatic crystals, up to 20 mm (25/32 in) long, were found in massive lepidolite in Rozna, also in Lastovicky, Czech Republic.

Foitite, 20 mm x, Dobra Voda, Czech Republic

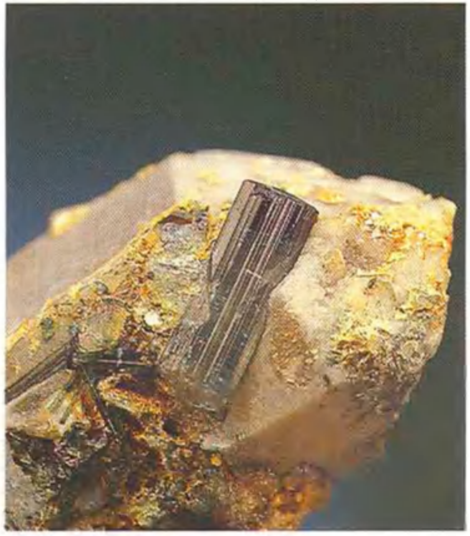

Dioptase

$Cu_6Si_6O_{18} \cdot 6\,H_2O$

TRIGONAL ● ●

Properties: C – emerald-green to blue-green; S – light blue-green; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 5; CL – good; F – uneven to conchoidal; M – long to short prismatic crystals, granular.
Origin and occurrence: Secondary in the oxidation zone of Cu deposits, associated with other secondary Cu minerals. Beautiful crystals up to 50 mm (2 in) across occur in Tsumeb, Namibia; Altyn Tyube, Kazakhstan; Renneville, Congo. It also comes from the Mammoth mine, Tiger, Arizona, USA.

Milarite

$KCa_2AlBe_2Si_{12}O_{30} \cdot H_2O$

HEXAGONAL ● ●

Properties: C – colorless, white, gray, light green; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.5; H – 5-6; CL – none; F – uneven to conchoidal; M – long prismatic to acicular crystals, granular.
Origin and occurrence: Hydrothermal in the Alpine-type veins, in pegmatites and hydrothermal veins, associated with adularia. Crystals up to 40 mm (1 9/16 in) long come from Jaguaraçu, Minas Gerais, Brazil. It was also found in St. Gotthard, Switzerland; Klein Spitzkopje, Namibia; Valencia mine, Guanajuato, Mexico.

Rossmanite, 10 mm xx, Lastovicky, Czech Republic

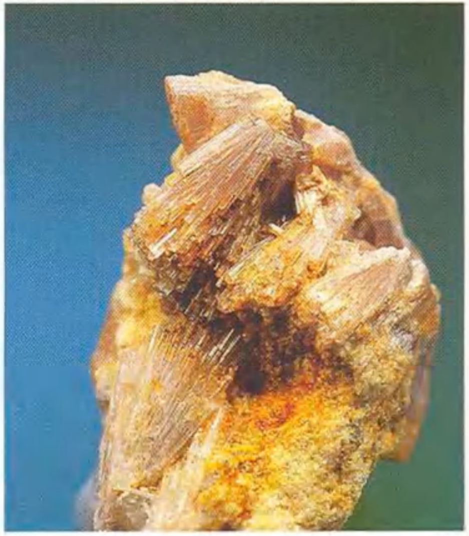

Dioptase, 40 mm, Mindouli, Zair

Sugilite

$KNa_2(Fe,Mn,Al)_2Li_3Si_{12}O_{30}$

HEXAGONAL ●

Properties: C – purple; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.7; H – 6-6.5; CL – none; F – uneven to conchoidal; M – long prismatic to acicular crystals, granular.
Origin and occurrence: Hydrothermal in alkaline syenites, associated with pectolite, albite and aegirine. It occurs in Iwagi Island, Japan and Hotazel, South Africa. *Application:* cut and polished as a gemstone and decorative stone.

Milarite, 22 mm, Jaguaracu, Brazil

Sugilite, 40 mm, Hotazel, South Africa

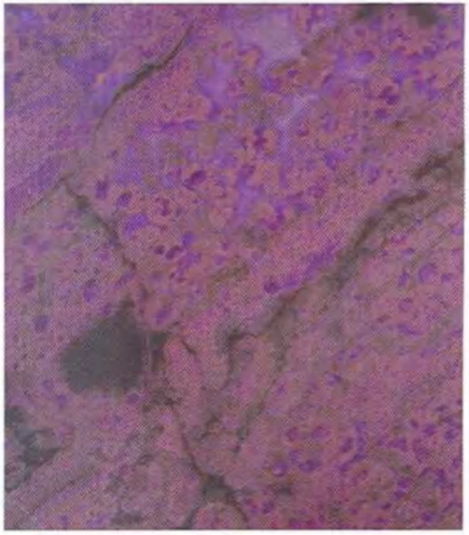

Bronzite, 15 mm, Minas Gerais, Brazil

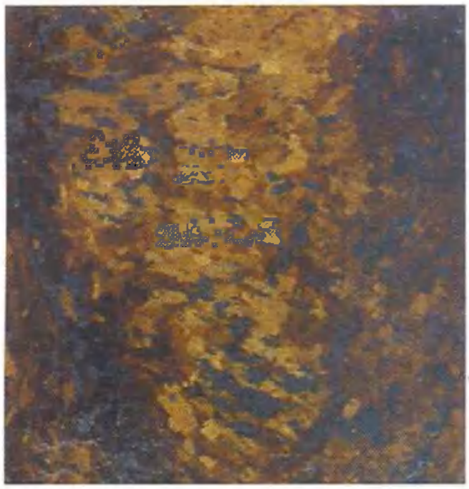

Enstatite

PYROXENE GROUP

$Mg_2Si_2O_6$

ORTHORHOMBIC ● ● ● ●

Varieties: bronzite

Properties: C – colorless, gray, yellowish, greenish, brown (bronzite); S – white; L – vitreous to dull; D – transparent to opaque; DE – 3.2; H – 5-6; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular.
Origin and occurrence: Magmatic in ultrabasic rocks, gabbros, peridotites and in meteorites; metamorphic in marbles, associated mainly with olivine and pyrope. Known from Bamle, Norway in crystals up to 50 cm (20 in) long. It was originally described from Ruda nad Moravou, Czech Republic. Gemmy crystals up to 20 mm ($^{25}/_{32}$ in) long occur in Brumado, Bahia, Brazil.

Diopside

PYROXENE GROUP

$CaMgSi_2O_6$

MONOCLINIC ● ● ● ● ●

Varieties: chrome diopside, fassaite, jeffersonite

Properties: C – light green, dark green (chrome diopside), colorless, white, gray, brown, rare blue; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 5.5-6.5; CL – good; F – uneven to conchoidal; M – long to short prismatic crystals, granular.

Chrome diopside, 5 mm grains, Inagli, Russia

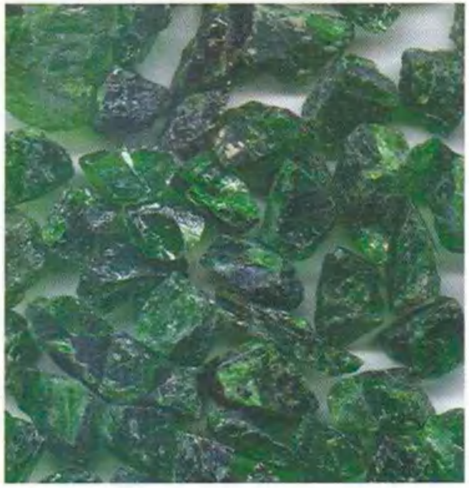

Origin and occurrence: Metamorphic in Ca-rich rocks, skarns, pyroxene gneisses, marbles; magmatic in basic igneous rocks, pegmatites and meteorites; hydrothermal in the Alpine-type veins. Fassaite occurs in skarns, chrome diopside in metamorphic deposits of Cu and Cr; jeffersonite in metamorphic deposits of Mn and Zn. Diopside is a typical rock-forming mineral.

Diopside, 22 mm x, Ala, Italy

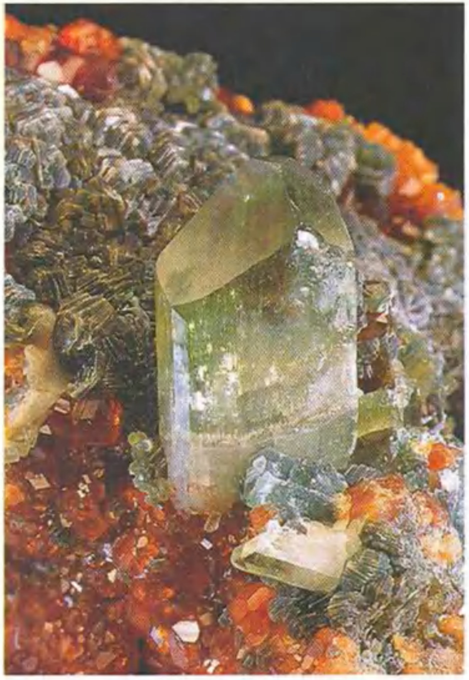

associated with plagioclase, grossular and epidote. Well formed crystals are relatively rare. Jeffersonite forms prismatic crystals up to 250 mm (9$^{15}/_{16}$ in) long in Franklin, New Jersey, USA. Crystals in Corrego Setuba, Minas Gerais, Brazil reach up to 30 cm (12 in). Other typical localities are Zillertal, Austria; Nordmarken, Sweden; Orford mine, Quebec, Canada. Fassaite is known from Val di Fassa, Italy. Chrome diopside crystals reach up to 100 mm (4 in) in Outokumpu, Finland and gem rough recently found in Inagli, Yakutia, Russia.

Hedenbergite

PYROXENE GROUP

$CaFeSi_2O_6$

MONOCLINIC ● ● ● ●

Properties: C – dark green, brown-green, brown to black; S – white to gray; L – vitreous to dull; D – transparent to translucent; DE – 3.6; H – 6; CL – good; F – uneven to conchoidal; M – long to short prismatic crystals, granular.
Origin and occurrence: Metamorphic in Ca-rich rocks, like Fe-skarns and pyroxene gneisses; magmatic in some granites and syenites, associated with magnetite, grossular and epidote. Crystals are relatively rare, reaching up to 50 mm (2 in) from Dalnegorsk, Russia and Franklin, New Jersey, USA. Also occurs in Nordmarken, Sweden. Large lamellar aggregates come from skarns in Rio Marina, Elba, Italy.

Hedenbergite, 40 mm, Dalnegorsk, Russia

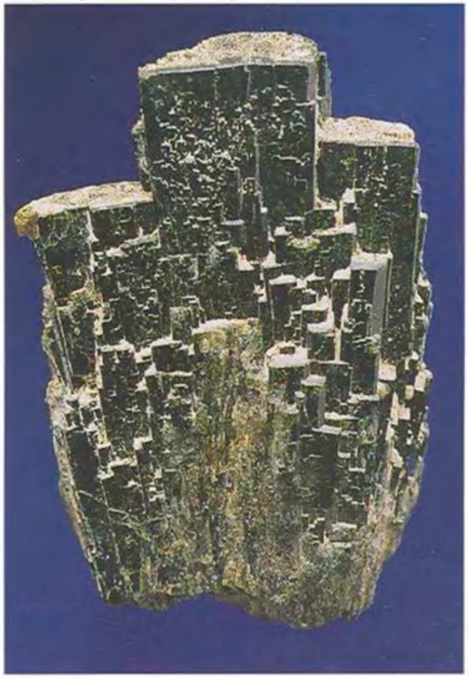

Augite, 17 mm, Firmerich, Germany

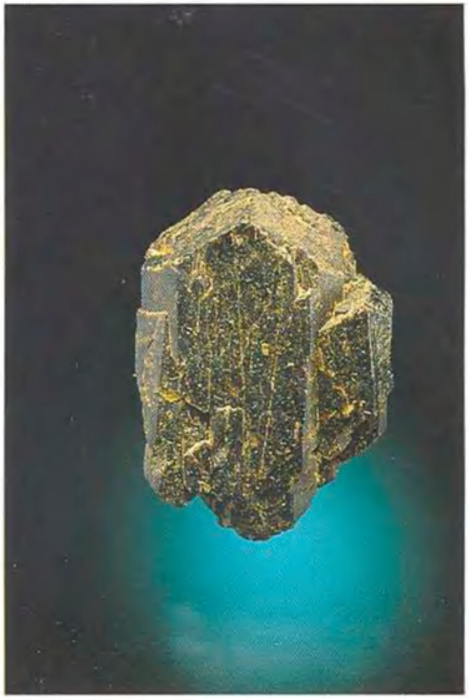

Augite

PYROXENE GROUP

$(Ca,Mg,Fe,Al)(Si,Al)_2O_6$

MONOCLINIC ● ● ● ● ●

Properties: C – dark brown to black; S – gray-green; L – vitreous to dull; D – translucent to opaque; DE – 3.6; H – 6; CL – good; F – uneven to conchoidal; M – short prismatic crystals, granular.
Origin and occurrence: Magmatic in basic rocks (basalts, gabbros, diabases and their tuffs); rare metamorphic in skarns. It is a typical rock-forming mineral, known from many localities. Well formed crystals up to 150 mm (6 in) across occur mainly in volcanic rocks, such as near Lake Clear, Ontario, Canada. Also known from Laacher See, Germany; Lukov and Pa_kapole, Czech Republic, in crystals, up to 50 mm (2 in) in size.

Omphacite, 90 mm, St.Leonhard, Austria

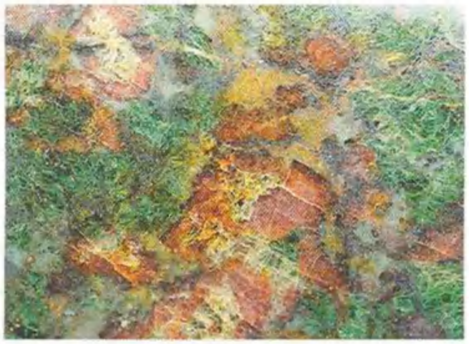

Omphacite

PYROXENE GROUP

$(Ca,Na)(Mg,Fe^{2+},Fe^{3+},Al)(Si,Al)_2O_6$

MONOCLINIC ● ● ● ●

Properties: C – green to dark green; S – gray-white; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 5-6; CL – good; F – uneven to conchoidal; M – short prismatic crystals, acicular aggregates, granular. *Origin and occurrence:* Metamorphic in ultrabasic and basic rocks, originated under high-temperature, usually associated with pyrope, diopside and kyanite. It occurs in eclogites and granulites in Rubinberg, Germany; Headsburg, California, USA and elsewhere.

Jadeite, 85 mm, Guerro Negro, Baja California, Mexico

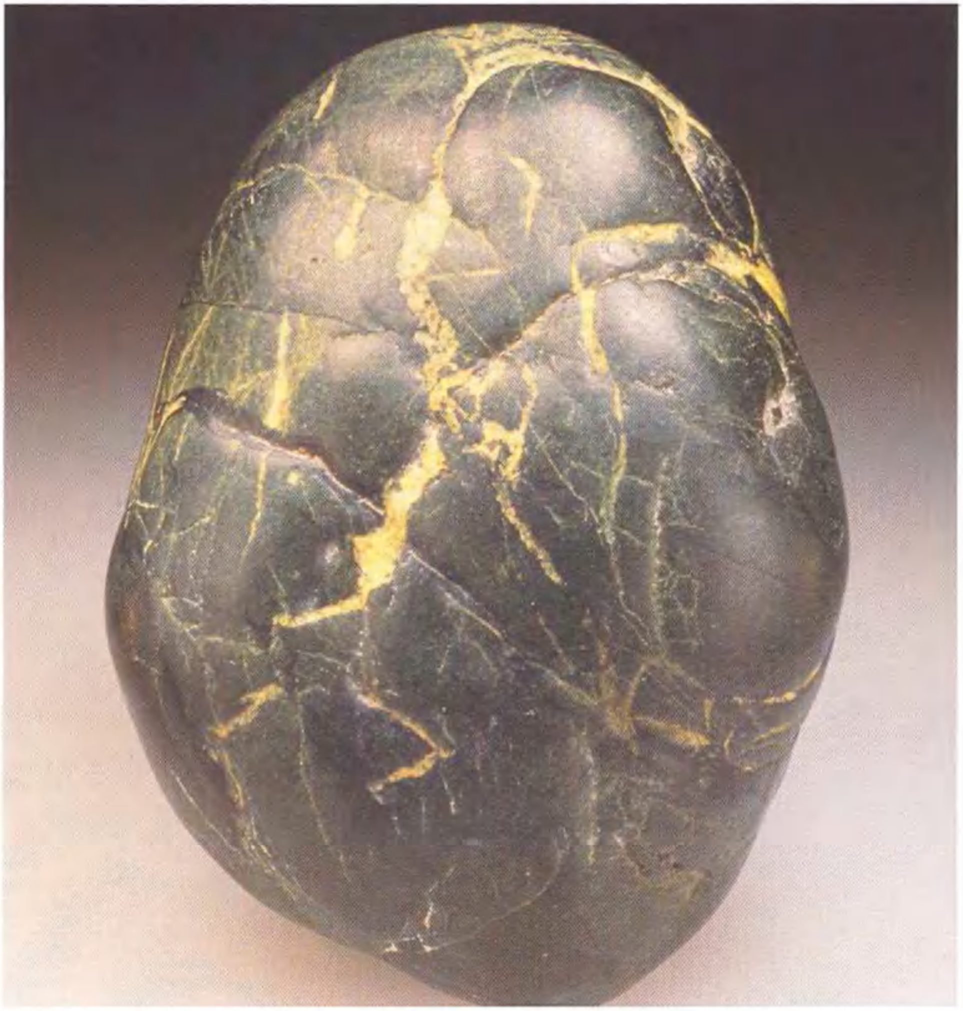

Nephrite, 60 mm, Jordanow, Poland

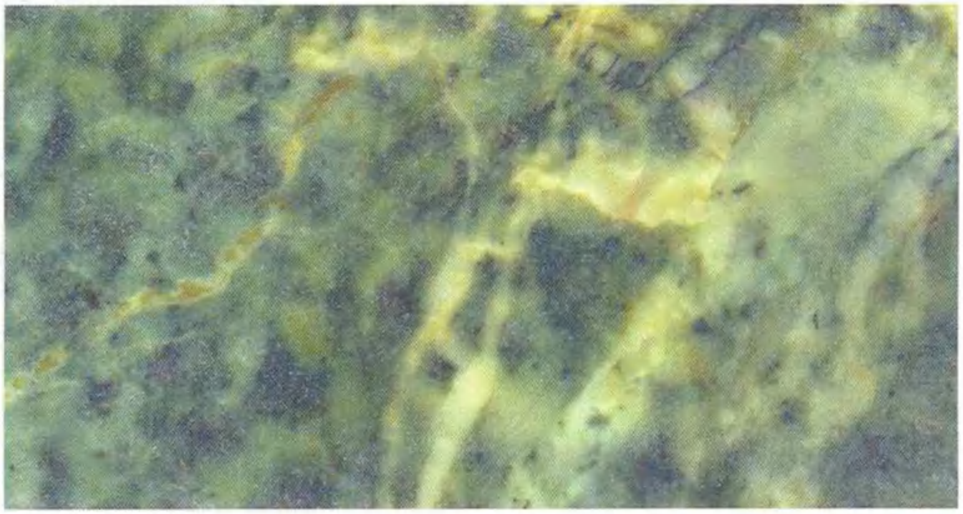

Jadeite

PYROXENE GROUP

$NaAlSi_2O_6$

MONOCLINIC ● ●

Varieties: nephrite

Properties: C – white, lavender to gray, in aggregates also light green (nephrite); S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 6; CL – good; F – uneven to conchoidal; M – prismatic crystals, acicular aggregates, granular.
Origin and occurrence: Exclusively metamorphic in strongly metamorphosed rocks; also in placers. Blocks, weighing several tons, are known from Ben Sur, California, USA. It also come from Tawmaw, Burma; New Zealand; Tibet and elsewhere.
Application: as a material for carvings and decorative purposes.

Aegirine

PYROXENE GROUP

$NaFe^{3+}Si_2O_6$

MONOCLINIC ● ● ●

Properties: C – dark green to black-green; S – light yellow-gray; L – vitreous to dull; D – translucent to opaque; DE – 3.6; H – 6; CL – good; F – uneven to conchoidal; M – long prismatic to acicular crystals, acicular aggregates.
Origin and occurrence: Magmatic, mainly in alkaline magmatic rocks (syenites, carbonatites, alkaline granites and their pegmatites); rare hydrothermal in sediments. Well formed prismatic crystals up to 150 mm (6 in) long come from Mount Malosa, Malawi. Crystals up to 30 cm (12 in) occur in Langensundsfjord, Norway. Other localities are Mount Karnasurt, Lovozero massif, Kola Peninsula, Russia and Mont St.-Hilaire, Quebec, Canada.

Aegirine, 77 mm, Mt.Malosa, Malawi

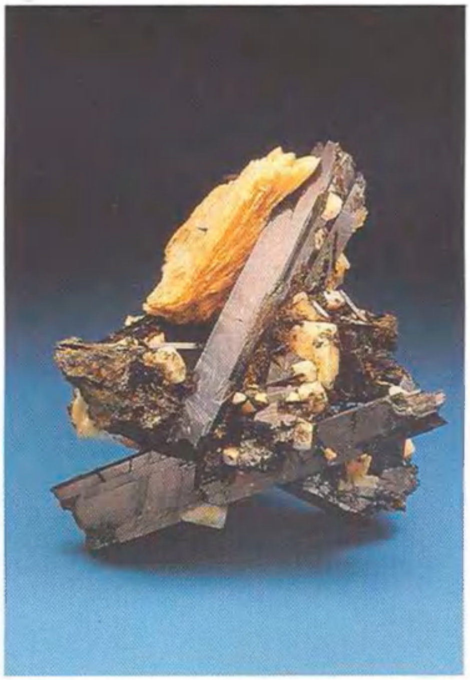

Spodumene, 52 mm x, Kunar, Afghanistan

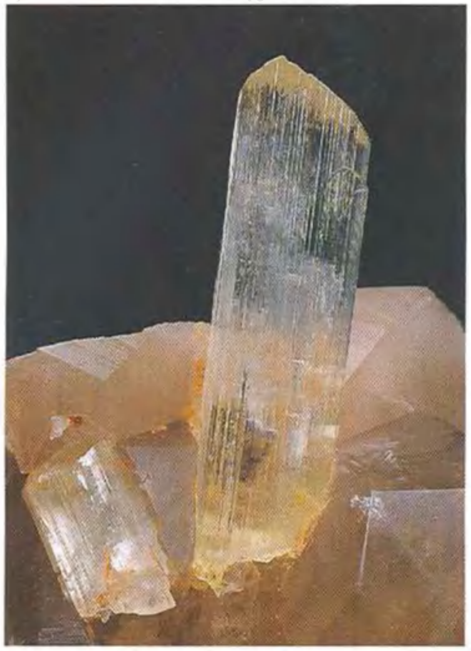

Hiddenite, 35 mm, Adams Property, U.S.A.

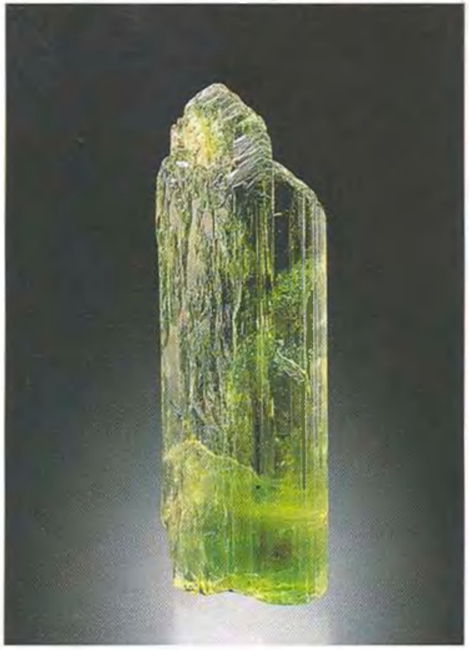

Kunzite, 118 mm, Nuristan, Afghanistan

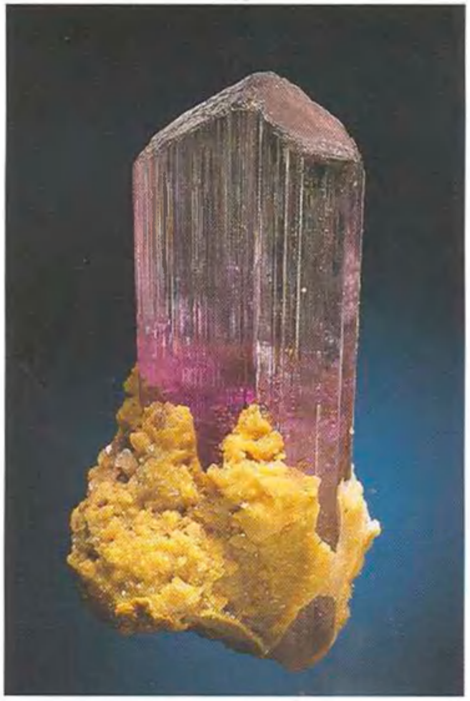

Spodumene

PYROXENE GROUP

$LiAlSi_2O_6$

MONOCLINIC ● ● ●

Varieties: kunzite, hiddenite, triphane

Properties: C – white, gray, yellowish (triphane), green (hiddenite), pink to purple (kunzite); S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 6.5-7.5; CL – good; F – uneven to conchoidal; M – long to short prismatic crystals, cleavable aggregates, granular; LU – orange.
Origin and occurrence: Magmatic in granitic pegmatites; hydrothermal in pegmatite cavities. It is known from many localities. Its poorly developed crystals up to 6 m (20 ft) long occur in the Etta mine, Keystone, South Dakota and in Kings Mountain, North Carolina, USA. Transparent kunzite crystals up to 40 cm (15¾ in) long come from Mawi, Laghman, Afghanistan. Kunzite crystals up to 280 mm (11 in) long found in the Pala Chief mine, Pala, California, USA. Cracked kunzite crystals, up to 1 m (39⅜ in) long known from Araçuai, Minas Gerais, Brazil. Hiddenite occurs in crystals up to 250 mm ($9\tfrac{13}{16}$ in) long in Resplendor, Minas Gerais, Brazil. It also comes from the Adams property, North Carolina, USA.
Application: raw material for ceramics, kunzite and hiddenite are cut as gemstones.

Carpholite, 70 mm, Horni Slavkov, Czech Republic

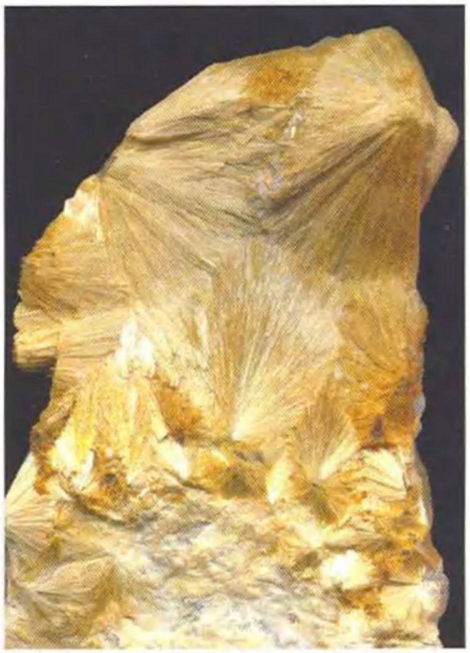

Carpholite

$MnAl_2Si_2O_6(OH)_4$

ORTHORHOMBIC ● ●

Properties: C – various hues of yellow; S – white; L – vitreous to dull; D – translucent; DE – 3.0; H – 5-5.5; CL – good; F – uneven to conchoidal; M – acicular crystals and their aggregates.
Origin and occurrence: Hydrothermal in veins, cross-cutting greisens, associated with fluorite, quartz and cassiterite; also in metamorphic rocks. It was described from Horni Slavkov, Czech Republic, where it forms acicular crystals up to 10 mm ($^3/_8$ in) long and radial aggregates. It is also known from Meuville, Belgium and Wippra, Germany.

Lorenzenite, 20 mm xx, Flora Mt., Kola, Russia

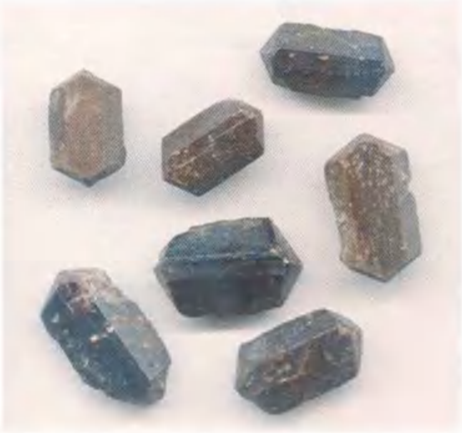

Lorenzenite

$Na_2Ti_2Si_2O_9$

ORTHORHOMBIC ● ●

Properties: C – brown to black; S – yellowish; L – vitreous to dull; D – translucent to opaque; DE – 3.4; H – 6; CL – good; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Magmatic in alkaline syenites and their pegmatites, associated with astrophyllite, nepheline and aegirine. Crystals up to 80 mm (3$^1/_8$ in) long come from Mount Flora, Lovozero massif, Kola Peninsula, Russia. Other localities are Narssarssuk, Greenland and Mont St.-Hilaire, Quebec, Canada.

Ajoite

$(K,Na)Cu_7AlSi_9O_{24}(OH)_6 . 3\ H_2O$

TRICLINIC ●

Properties: C – blue-green; S – light green; L – vitreous to dull; D – translucent to opaque; DE – 3.0; H – not determined; CL – not determined; F – uneven to conchoidal; M – prismatic crystals, massive.
Origin and occurrence: Secondary in Cu deposits, associated with shattuckite. It occurs rarely as small crystals in the New Cornelia mine, Ajo, Arizona, USA. It is also known as inclusions in quartz crystals in the Messina mine, Transvaal, South Africa.

Ajoite, 60 mm, Ajo, U.S.A.

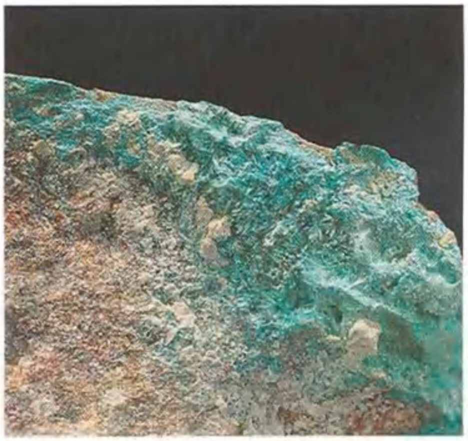

Holmquistite, 60 mm, Lac Malartic, Canada

Antophyllite

AMPHIBOLE GROUP

$Mg_7Si_8O_{22}(OH,F)_2$

ORTHORHOMBIC ● ● ● ●

Properties: C – white, light brown, yellow, light green; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 5.5-6; CL – good; F – uneven to conchoidal; M – long prismatic crystals, columnar and radial aggregates, granular.
Origin and occurrence: Metamorphic in gneisses, serpentinites and at the contact of serpentinites and pegmatites, usually with actinolite or tremolite and cordierite. Crystals up to 150 mm (6 in) long come from the Marbridge No.1 mine, Quebec, Canada. It is also known from Bodenmais, Germany; Snarum, Sweden and Hermanov, Czech Republic.

Anthophyllite, 60 mm, North Carolina, U.S.A.

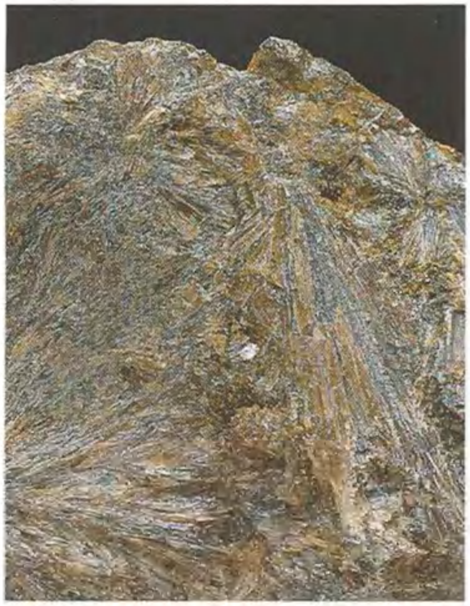

Holmquistite

AMPHIBOLE GROUP

$Li_2Mg_3Al_2Si_8O_{22}(OH,F)_2$

ORTHORHOMBIC ● ● ●

Properties: C – blue, purple, gray, black; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.0; H – 5.5-6; CL – good; F – uneven to conchoidal; M – acicular and prismatic crystals, fibrous aggregates, granular.
Origin and occurrence: Hydrothermal at the contacts between complex Li-bearing pegmatites and amphibolites, sometimes associated with biotite. It occurs in Greenbushes, Western Australia, Australia where its fibers reach up to 180 mm (7 1/16 in) in length. It is also known from Uto, Sweden; Manono, Zaire and Brandbrucken, Austria.

Manganogedrite

AMPHIBOLE GROUP

$Mn_2Fe_5Si_8O_{20}(OH)_2$

MONOCLINIC ● ●

Properties: C – gray, dark green, brown, greenish; S – white; L – vitreous to dull; D – translucent to opaque; DE – 3.5; H – 5-6; CL- good; F – uneven to conchoidal; M – columnar, acicular and radial aggregates, granular.
Origin and occurrence: Metamorphic in contact and regionally metamorphosed Mn rich rocks, associated with chlorite and magnetite. It comes from Dannemora, Sweden and elsewhere.

Manganogedrite, 70 mm, Franz Joseph Iceberg, New Zealand

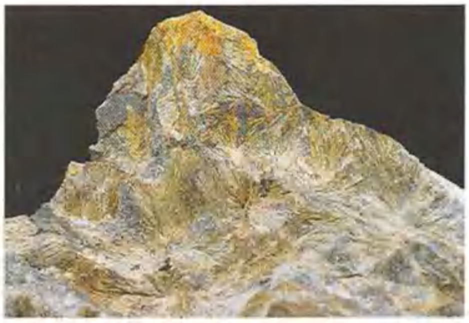

Tremolite, 37 mm, Maricopa Co., Arizona, U.S.A.

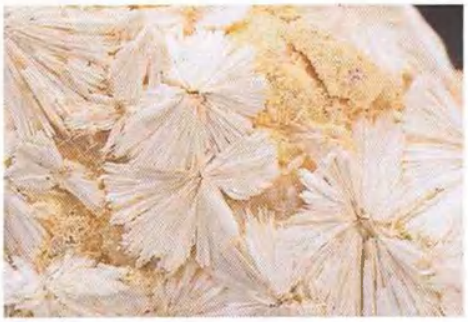

Tremolite

AMPHIBOLE GROUP

$Ca_2Mg_5Si_8O_{20}(OH)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – white, gray, greenish, green, brown, pink; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.9; H – 5-6; CL – good; F – uneven to conchoidal; M – columnar, acicular and radial aggregates, granular.
Origin and occurrence: Metamorphic in contact and regionally metamorphosed rocks, dolomites and ultrabasic rocks; hydrothermal in the Alpine-type veins, typically associated with diopside, talc, dolomite and calcite. It is a typical rock-forming mineral, known from many localities. Prismatic crystals up to 40 cm (15¾ in) long come from Brumado, Bahia, Brazil. It occurs also in Campolungo, Switzerland; Zillertal, Austria; Gouverneur, New York, USA and elsewhere.

Actinolite

AMPHIBOLE GROUP

$Ca_2(Mg,Fe)_5Si_8O_{20}(OH)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – light green to almost black; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 5-6; CL – good; F – uneven to conchoidal; M – columnar, acicular and radial aggregates, granular.
Origin and occurrence: Metamorphic in contact and regionally metamorphosed rocks, also in dolomites and some basic rocks (amphibolites, shales). It is a typical rock-forming mineral, associated with anthophylite, chlorite, talc, dolomite and calcite. It is known from many localities where it forms columnar aggregates up to 250 mm ($9^{13}/_{16}$ in) long, as in Knappenwand and Zillertal, Austria; Val Malenco, Italy; Brumado, BahRa, Brazil and Sobotin, Czech Republic.

Actinolite, 30 mm xx, Austris

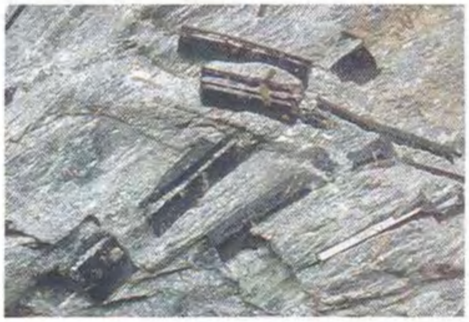

Edenite

AMPHIBOLE GROUP

$NaCa_2(Mg,Fe)_5Si_7AlO_{22}(OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – white, gray to dark green; S – white; L – vitreous to dull; D – transparent, translucent to opaque; DE – 3.1; H – 5-6; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular.
Origin and occurrence: Metamorphic in regionally metamorphosed basic rocks and marbles; magmatic in diorites and gabbros. Typical rock-forming mineral, its crystals up to 40 mm ($1^{9}/_{16}$ in) long come from Edenville, New York, USA and Bancroft, Ontario, Canada.

Edenite, 48 mm, Wilberforce, Canada

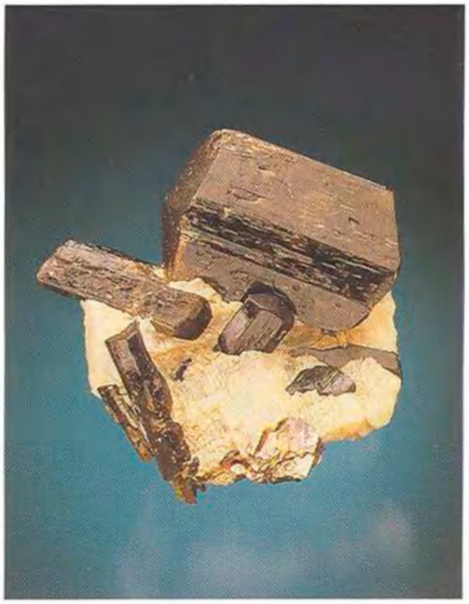

Pargasite

AMPHIBOLE GROUP

$NaCa_2(Mg,Fe)_4AlSi_6Al_2O_{22}(OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – light brown, green-blue, brown to black-brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 5-6; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.
Origin and occurrence: Metamorphic in regionally metamorphosed basic rocks, marbles and skarns; magmatic in diorites and gabbros; typical rock-forming mineral. Crystals up to 80 mm ($3^{1}/_{8}$ in) across come from the Jensen quarry, Riverside, California, USA; Pargas, Finland; Hunza, Pakistan.

Common Amphibole

AMPHIBOLE GROUP

$(Ca,Na)_2(Mg,Fe,Al)_5(Si,Al)_8O_{22}(OH,F)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – dark green, brown to black; S – gray; L – vitreous to dull; D – translucent to opaque; DE – 3.4; H – 5-6; CL – good; F – uneven to conchoidal; M – short prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Metamorphic in basic rocks, as amphibolites and shales; magmatic in some igneous rocks, as syenites, diorites, andesites and basalts. Typical rock-forming mineral, known from many localities. Large crystals up to 1 m ($39^{3}/_{8}$ in) long come from Silver Crater, Ontario, Canada. It is also known from Zillertal, Austria; Lukov, Czech Republic and elsewhere.

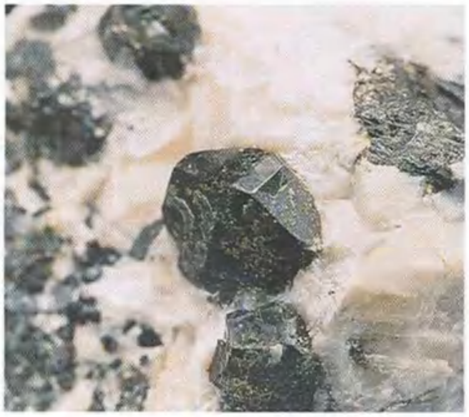

Pargasite, 70 mm, Limberg, Parainen, Finnland

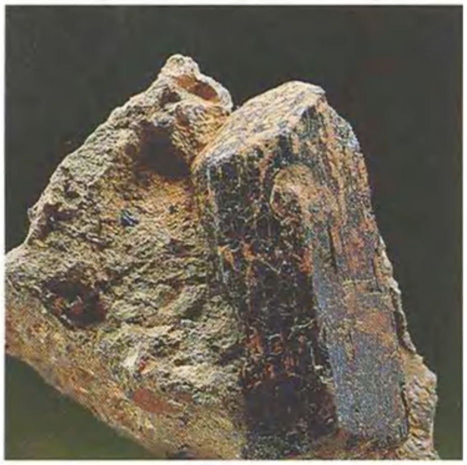

Common Amphibole, 50 mm x, Lukov, Czech Republic

Richterite

AMPHIBOLE GROUP

$Na_2Ca(Mg,Fe,Al)_5(Si,Al)_8O_{22}(OH,F)_2$

MONOCLINIC ● ● ●

Properties: C – brown, yellow, dark red, dark green; S – gray; L – vitreous to dull; D – transparent to translucent; DE – 3.1; H – 5-6; CL – good; F – uneven to conchoidal; M – long prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Magmatic in alkaline rocks, as trachytes and alkaline granites; rare metamorphic in some metamorphosed rocks, as skarns and marbles. Crystals up to 100 mm (4 in) long come from Wilberforce, Ontario, Canada and Langban, Sweden.

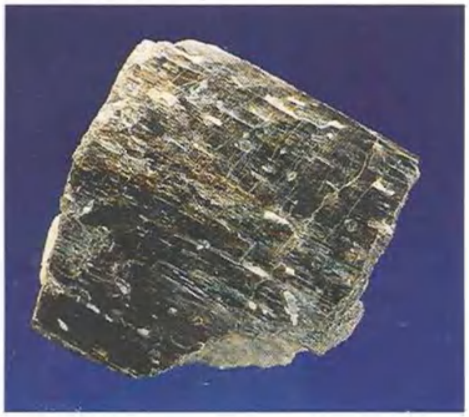

Richterite, 40 mm, Wilberforce, Canada

Riebeckite, 45 mm, Griqualand, South Africa

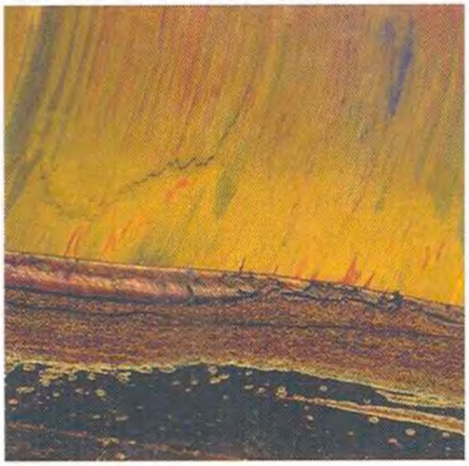

Riebeckite

AMPHIBOLE GROUP

$Na(Fe^{2+},Mg)_3Fe^{3+}{}_2Si_8O_{22}(OH)_2$

MONOCLINIC ● ● ● ●

Varieties: crocidolite (tiger's eye)

Properties: C – dark blue to black, crocydolite – gray-blue; S – white; L – vitreous to dull; D – translucent to opaque; DE – 3.4; H – 5; CL – perfect; F – uneven to conchoidal; M – long prismatic crystals, acicular to fibrous aggregates.
Origin and occurrence: Magmatic in alkaline rocks, as granites, syenites and rhyolites; metamorphic in regionally metamorphosed Fe rich shales. Large crystals up to 150 mm (6 in) long come from Khangay, Mongolia. Crocidolite is famous from Griqualand, South Africa.

Wollastonite, 44 mm, Crestmore, U.S.A.

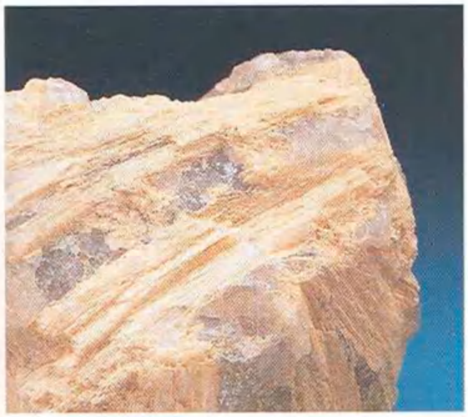

Aenigmatite, 40 mm xx, Eveslogchorr, Kola, Russia

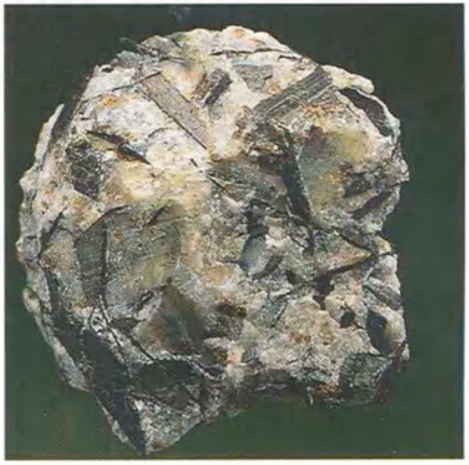

Aenigmatite

$Na_2Fe_5TiSi_6O_{20}$

TRICLINIC ● ●

Properties: C – black; S – red-brown; L – submetallic to dull; D -almost opaque; DE – 3.8; H – 5.5; CL – perfect; F – uneven to conchoidal; M – long prismatic crystals, columnar aggregates, granular.
Origin and occurrence: Magmatic in alkaline syenites and volcanic rocks, associated with aegirine and arfvedsonite. It occurs in Julianehab, Greenland; Lipari Island, Italy and elsewhere.

Wollastonite

$Ca_3Si_3O_9$

TRICLINIC ● ● ● ●

Properties: C - white, gray, light greenish, pink; S – white; L – vitrous to dull; D – transparent to translucent; DE – 3.0; H – 4.5-5; CL – good; F – uneven to conchoidal; M – tabular to short prismatic crystals, columnar and radial aggregates, granular; LU - orange.
Origin and occurrence: Metamorphic mainly in contact metamorphosed rocks, as marbles, skarns, less common in pyroxene gneisses and quartzites, associated with grossular, diopside and vesuvianite. Columnar and acicular aggregates up to 180 mm (7$^{1}/_{16}$ in) long occur in the Strickland quarry, Connecticut, USA. It also come from Crestmore, California, USA; Ciclova, Romania and Zulova , Czech Republic. Short prismatic crystals up to 30 cm (12 in) across are known from the Santa Fe mine, Chiapas, Mexico.

Bustamite, 20 mm xx, Broken Hill, Australia

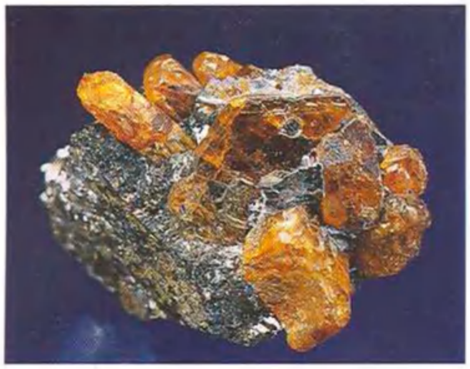

Bustamite

$(Mn,Ca)_3Si_3O_9$

TRICLINIC ● ● ●

Properties: C – pink to red-brown; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.4; H – 5.5-6.5; CL – good; F – uneven to conchoidal; M – tabular crystals, acicular aggregates, massive.
Origin and occurrence: Metamorphic or hydrothermal in Mn rich metamorphic rocks or skarns. It is associated with rhodonite and other Mn silicates in Franklin, New Jersey, USA; Langban, Sweden; Broken Hill, New South Wales, Australia.

Pectolite, 70 mm, Zelechovske udoli, Czech Republic

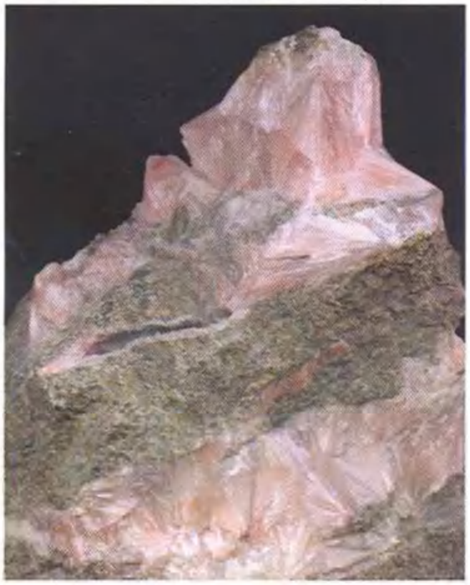

Pectolite

$Ca_2NaSi_3O_8(OH)$

TRICLINIC ● ● ●

Properties: C – white, pinkish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.9; H – 4.5-5; CL – good; F – uneven to conchoidal; M B acicular and radial aggregates; LU – locally yellow to orange.
Origin and occurrence: Hydrothermal in basalt cavities, less frequently along the cracks in marbles, associated with zeolites. Classic localities are West Paterson, New Jersey, USA, where it forms acicular spherical aggregates up to 180 mm ($7\frac{1}{16}$ in) across; also Monte Baldo, Italy; Zelechovske udoli, Czech Republic. Prismatic crystals up to 50 mm (2 in) long come from Mont St.-Hilaire, Quebec, Canada.

Serandite

$Mn_2NaSi_3O_8(OH)$

TRICLINIC ●

Properties: C – pink, red; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.9; H – 4.5-5; CL – good; F – uneven to conchoidal; M – tabular and prismatic crystals, granular.
Origin and occurrence: Hydrothermal in cavities of igneous rocks (alkaline syenite, carbonatite) together with aegirine and analcime. Perfect pink crystals up to 200 mm ($7\frac{7}{8}$ in) long come from Mont St.-

Serandite, 25 mm, Mont St.-Hilaire, Canada

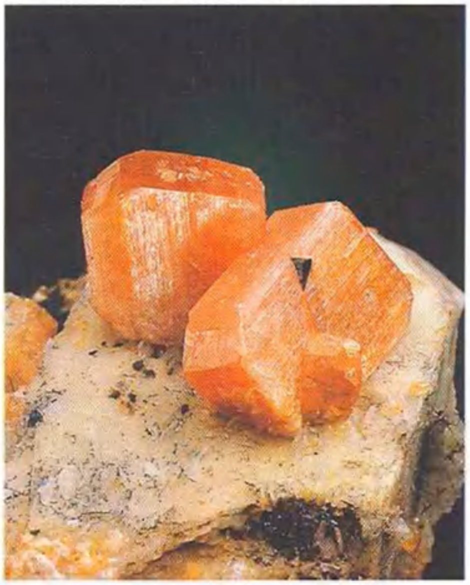

Hilaire, Quebec, Canada; crystals up to 60 mm (2¹/₈ in) long found in the Yubileinaya dike, Mount Karnasurt, Lovozero massif, Kola Peninsula, Russia.

Charoite

$(K,Sr,Ba)(Ca,Na)_2(Si,Al)_4O_{10}(OH,F)$

TRICLINIC ● ●

Properties: C – purple; S – white; L – vitreous to dull; D – translucent to opaque; DE – 2.6; H – 5-6; CL – good; F – uneven; M – fibrous aggregates, massive.
Origin and occurrence: Hydrothermal in alkaline igneous rocks, associated with aegirine and nepheline. Rich aggregates occur in Sirenevyi Kamen, Chara river basin, Murun massif, Yakutia, Russia.
Application: cut and polished as decorative stone.

Charoite, 50 mm, Sirenevyi Kamen, Russia

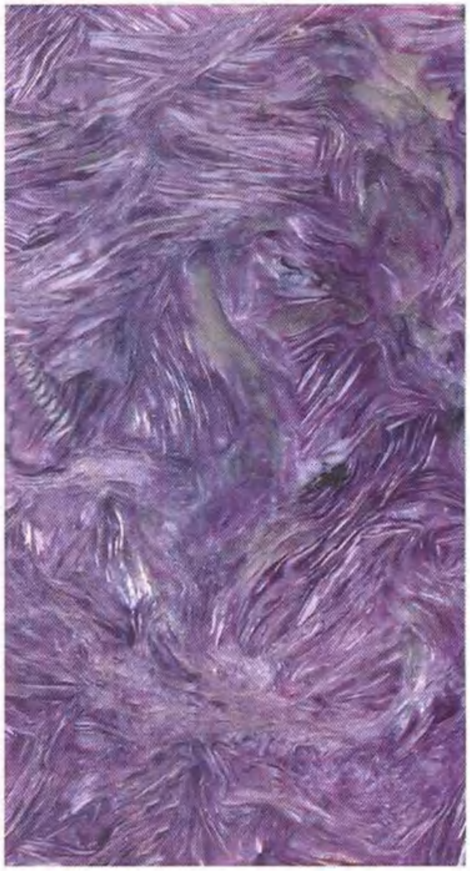

Agrellite

$NaCa_2Si_4O_{10}F$

TRICLINIC ●

Properties: C – gray-white to greenish; S – white; L – vitreous, dull to pearly; D – transparent to translucent; DE – 2.9; H – 5.5; CL – good; F – uneven; M – long prismatic crystals, granular.
Origin and occurrence: Metamorphic in alkaline gneisses. It occurs as prismatic crystals up to 100 mm (4 in) long in the Kipawa river basin, Quebec, Canada.

Okenite

$CaSi_2O_4(OH)_2 . H_2O$

TRICLINIC ● ●

Properties: C – white, yellowish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.3; H – 4.5-5; CL – good; F – uneven to conchoidal; M B blade-shaped crystals, acicular crystals and their spherical aggregates.

Origin and occurrence: Hydrothermal in basalt cavities together with zeolites. Rich acicular spherical aggregates up to 80 mm (3¹/₈ in) in diameter come from basalt cavities in the vicinity of Poona, India; also known from Faeroe Islands.

Agrellite, 120 mm, Kipawa, Canada

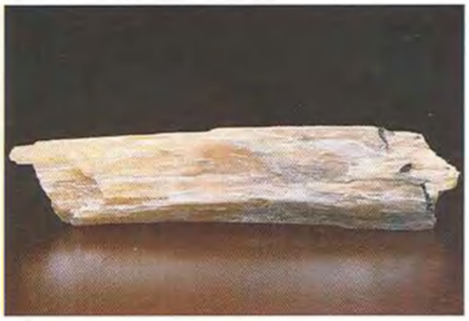

Okenite, 125 mm, Mumbai, India

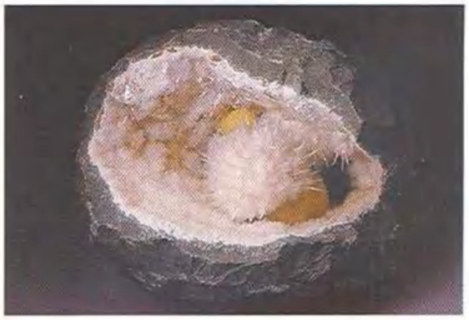

Xonotlite, 20 mm vein, Stare Ransko, Czech Republic

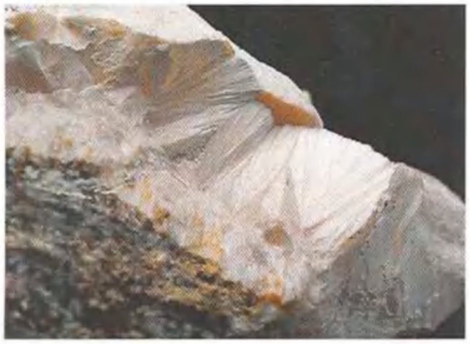

Xonotlite

$Ca_6Si_6O_{17}(OH)_2$

MONOCLINIC ● ●

Properties: C – white, pinkish, gray; S – white; L – vitreous, dull to pearly; D – transparent to translucent; DE – 2.7; H – 6.5; CL – good; F – uneven to conchoidal; M – acicular to fibrous aggregates, massive.
Origin and occurrence: Hydrothermal in cavities of basic and ultrabasic rocks and in the Alpine-type veins, associated with calcite and zeolites; rare metamorphic in skarns and in contact metamorphosed marbles. It comes from Tetela de Xonotla, Puebla, Mexico; Crestmore, California, USA; Mihara mine, Japan. It is was also found in Stare Ransko, Czech Republic.

Elpidite

$Na_2ZrSi_6O_{15} \cdot 3\,H_2O$

ORTHORHOMBIC ● ●

Properties: B – colorless, yellowish, red; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.6; H – 5.5-6.5; CL – good; F – uneven to conchoidal; M – long prismatic crystals, fibrous aggregates.
Origin and occurrence: Hydrothermal in albitized parts of alkaline pegmatites, associated with albite and aegirine. Perfect crystals up to 30 cm (12 in) long come from Tarbagatay, Kazakhstan. Crystals up to 200 mm (7⅞ in) long found at Mont St.-Hilaire, Quebec, Canada. Also known from Khan Bogdo, Gobi desert, Mongolia; Umbozero mine on Mount Alluaiv, Lovozero massif, Kola Peninsula, Russia.

Elpidite, 128 mm, Mont St.-Hilaire, Canada

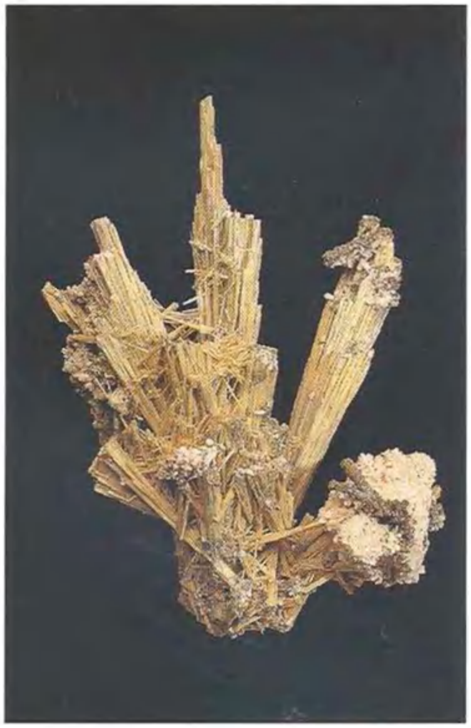

Rhodonite, 35 mm x, Broken Hill, Australia

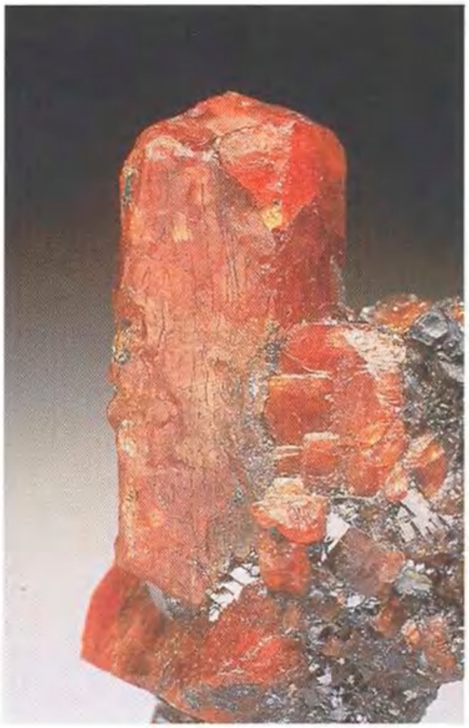

Rhodonite, 70 mm, Maloye Sedelnikovo, Ural Mts., Russia

Rodonite

$CaMn_4Si_3O_9$

TRICLINIC ● ● ●

Properties: C – pink, red, red-brown, gray; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.7; H – 5.5-6.5; CL – good; F – uneven to conchoidal; M – tabular crystals, granular, massive.
Origin and occurrence: Metamorphic in Mn-rich rocks, hydrothermal in ore veins, associated with spessartine and Mn oxides. Perfect crystals up to 200 mm (7⅞ in) long come from Franklin, New Jersey, USA; also from Broken Hill, New South Wales, Australia.
Granular and massive aggregates occur in Langban, Sweden and in Maloye Sidelnikovo, Ural mountains, Russia.
Application: cut and polished as a decorative stone.

Babingtonite

$Ca_2Fe^{2+}Fe^{3+}Si_5O_{14}(OH)$

TRICLINIC ● ●

Properties: C – green-black to black-brown; S – green-gray; L – vitreous to dull; D – translucent to opaque; DE – 3.3; H – 5. 5-6; CL – good; F – uneven to conchoidal; M – short prismatic crystals, granular.
Origin and occurrence: Hydrothermal along the cracks in rocks, associated with prehnite, adularia and epidote.
Well-formed crystals about 20 mm (25/32 in) across are known from Arendal, Norway. Other localities are Westfield, Massachusetts, USA; Baveno, Italy.

Babingtonite, 5 mm x, Mumbai, India

Kinoite

$Ca_2Cu_2Si_3O_8(OH)_4$

MONOCLINIC ●

Properties: C – blue; S – light blue; L – vitreous to dull; D – transparent to translucent; DE – 3.2; H – 5; CL B perfect; F – uneven to conchoidal; M – prismatic crystals, granular.
Origin and occurrence: Hydrothermal along cracks in rocks, associated with apophyllite and copper. Small crystals were found in the Christmas mine, Santa Rita mountains, Arizona, also in the Kearsarge vein, Keweenaw Peninsula, Michigan, USA.

Inesite

$Ca_2Mn_7Si_{10}O_{28}(OH)_2 . 5 H_2O$

TRICLINIC ● ●

Properties: C – pink, brown; S – white; L – vitreous to dull; D – translucent; DE – 3.0; H – 5.5; CL – good; F – uneven to conchoidal; M – short prismatic and tabular crystals, columnar and radial aggregates, granular.
Origin and occurrence: Hydrothermal in ore veins, associated with rhodochrosite, rhodonite and calcite. Rich aggregates and well-formed crystals about 10 mm ($^3/_8$ in) across found in Langban, Sweden; also in Hale Creek, California, USA; Broken Hill, New South Wales, Australia. Its spherical aggregates up to 30 mm ($1^3/_{16}$ in) in diameter are known from the Wessels mine, Kuruman, South Africa.

Neptunite

$KNa_2LiFe_2Ti_2Si_9O_{22}$

MONOCLINIC ● ●

Properties: C – dark brown to black; S – red-brown; L – vitreous to submetallic; D – translucent to opaque; DE – 3. 2; H – 5-6; CL – good; F – uneven to conchoidal; M – long prismatic crystals, granular.
Origin and occurrence: Hydrothermal in veins, associated with natrolite, benitoite and eudialyte. Well-formed prismatic crystals up to 80 mm ($3^1/_8$ in) long come from the Benitoite Gem mine, San Benito Co., California, USA. It is also known from Mont St.-Hilaire, Quebec, Canada and Narssarssuk, Greenland.

Epididymite

$NaBe_2Si_3O_7(OH)$

ORTHORHOMBIC ● ●

Properties: C – white, gray, yellowish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.6; H – 6-7; CL – good; F – uneven to conchoidal;

Kinoite, 36 mm, Christmas Mine, U.S.A.

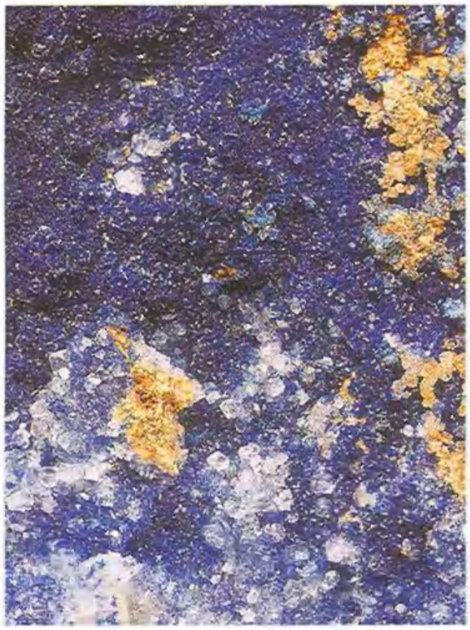

Inesite, 50 mm, Kuruman, South Africa

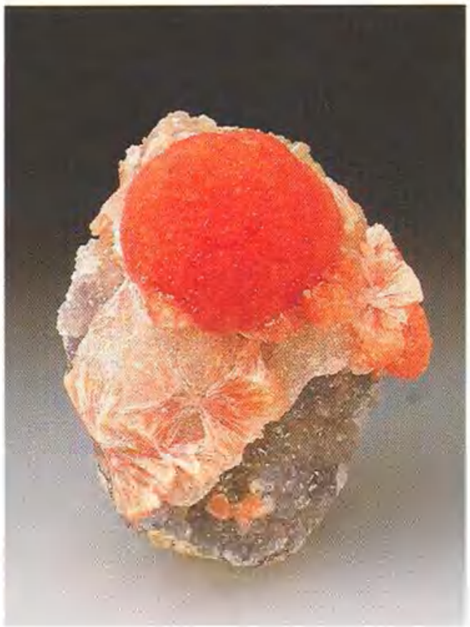

Neptunite, 28 mm, San Benito Co., U.S.A.

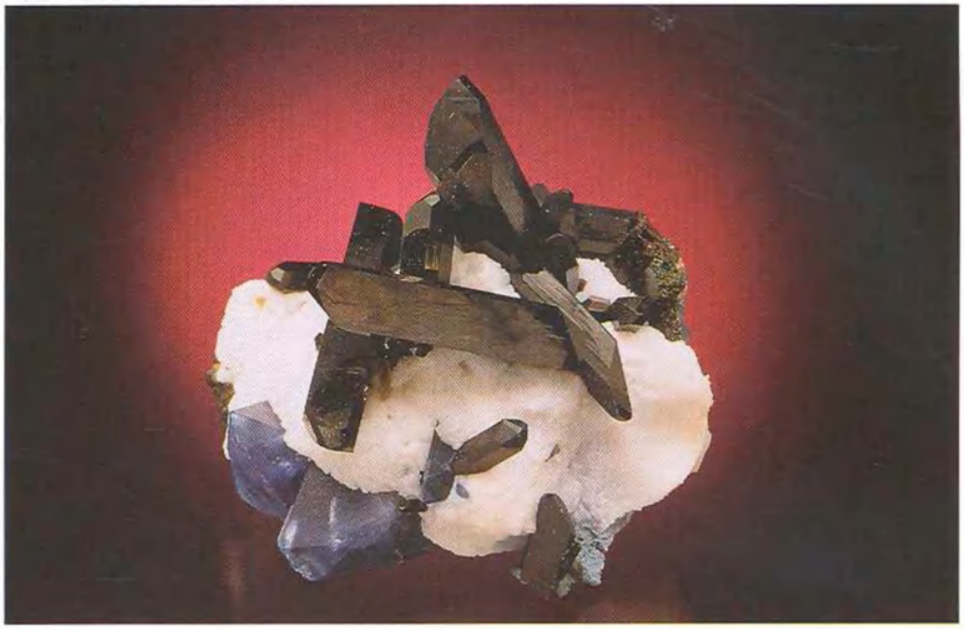

M – tabular crystals, lamellar aggregates, granular. *Origin and occurrence:* Hydrothermal in cavities of alkaline pegmatites, associated with zeolites. Tabular crystals up to 60 mm ($2^{3}/_{8}$ in) across come from Mount Malosa, Malawi. It is also known from Langesundsfjord, Norway; Mont St.-Hilaire, Quebec, Canada; Narssarssuk, Greenland.

Bavenite, 40 mm, Yermolayevskoye, Russia

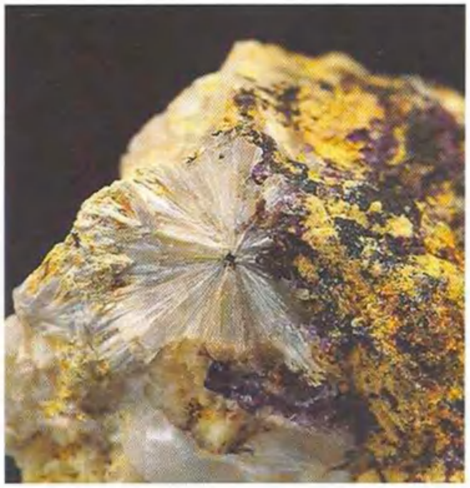

Bavenite

$Ca_4Al_2Be_2Si_9O_{26}(OH)_2$

ORTHORHOMBIC ● ● ●

Properties: C – white, yellowish, pinkish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.8; H – 5.5; CL – good; F – uneven to conchoidal; M – prismatic to tabular crystals, radial and lamellar aggregates. *Origin and occurrence:* Hydrothermal in pegmatic cavities with albite: typically produced by beryl replacement in pegmatites or helvite replacement in skarns. Found in Baveno, Italy with crystals up to 20 mm ($^{25}/_{32}$ in) across. Also known from Strzegom, Poland. Crystals up to 4 mm ($^{5}/_{32}$ in) across found in the Hewitt quarry, USA.

Epididymite, 70 mm, Apatity, Kola, Russia

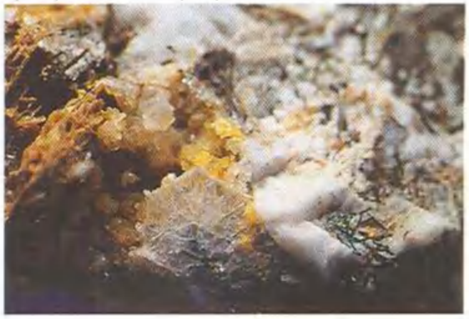

Prehnite, 28 mm, Brandberg, Namibia

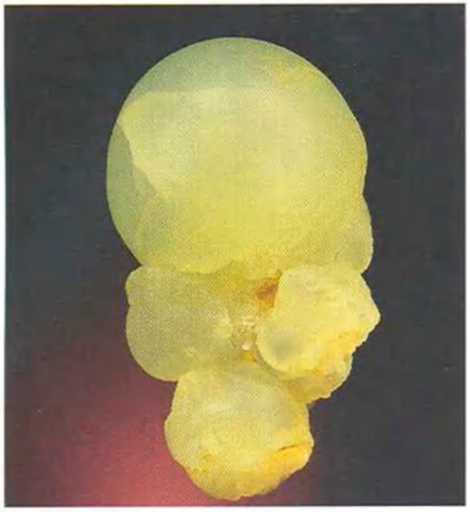

Prehnite

$Ca_2Al_2Si_3O_{10}(OH)_2$

ORTHORHOMBIC ● ● ● ●

Properties: C – white, greenish to green, yellow, gray; S – white; L – vitreous; PS – transparent to translucent; DE – 2.9; H – 6-6.5; CL – good; F – uneven to conchoidal; M – tabular to prismatic crystals, usually forming botryoidal or spherical aggregates, granular.

Origin and occurrence: Hydrothermal in basalt cavities and in the Alpine-type veins, commonly associated with calcite, albite, epidote and zeolites. Rich aggregates are known from West Paterson, New Jersey, USA; Val di Fassa, Italy; Poona, India. Crystals up to 45 mm (1¾ in) across found in Copper Valley, Brandberg, Namibia, Asbestos, Quebec, Canada and Talnakh, Siberia, Russia.

Astrophyllite, 30 mm xx, Khibiny Massif, Kola, Russia

Fluorapophyllite, 47 mm, Poona, India

Astrophyllite

$(K,Na)_3(Fe,Mn)_7Ti_2Si_8O_{24}(O,OH)_7$

TRICLINIC ● ●

Properties: C – various hues of yellow; S – white; L – vitreous to dull; D – transparent to translucent; DE – 3.3; H – 3; CL – good; F – uneven to conchoidal; M – long prismatic crystals, bladed aggregates.

Origin and occurrence: Magmatic in alkaline pegmatites. Crystals and their radial aggregates reaching up to 100 mm (4 in) come from Mount Eveslogchorr, Khibiny massif, Kola Peninsula., Russia. It is also known from Mont St.-Hilaire, Quebec, Canada.

Fluorapophyllite

$KCa_4Si_8O_{20}(F,OH) . 8 H_2O$

TETRAGONAL ● ● ●

Properties: C – white, greenish yellowish, pinkish; S – white; L – greasy to pearly; D – transparent to translucent; DE – 2.4; H – 4.5-5; CL – perfect; F – uneven; M – prismatic and tabular crystals, lamellar aggregates, massive.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, along the cracks in the Alpine-type veins, in ore veins and pegmatites. Well-formed crystals up to 100 mm (4 in) long are known from Jalgaon and Poona, India; Centreville and Fairfax, *Pyrophyllite, 37 mm, California, U.S.A.*

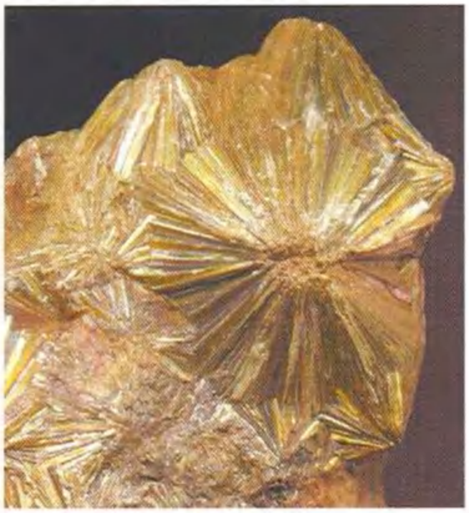

Agalmatolite, 70 mm, India

Virginia, USA. Other localities are St. Andreasberg, Germany; Bento Gonçalves, Rio Grande do Sul, Brazil.

Pyrophyllite

$Al_2Si_4O_{10}(OH)_2$

MONOCLINIC ● ● ● ●

Varieties: agalmatolite

Properties: C – white, greenish, yellow, gray; S – white; L – greasy to pearly; D – translucent to transparent; DE – 2.8; H – 1-2; CL – perfect; F – uneven; M B foliated and lamellar aggregates, massive (agalmatolite), radial to acicular aggregates.
Origin and occurrence: Metamorphic in Al-rich rocks, together with andalusite and kyanite; hydrothermal in veins with quartz and micas. Rich foliated and radial aggregates and poorly-formed crystals up to 10 mm ($^3/_8$ in) across are known from Zermatt, Switzerland; Berezovsk, Ural mountains, Russia and the Champion mine, California, USA.
Application: heat-resistant material, agalmatolite as decorative stone.

Talc

$Mg_3Si_4O_{10}(OH)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – white, greenish, yellow, pinkish, gray; S – white; L – greasy to pearly; D – transparent to translucent; DE – 2.8; H – 1; CL – perfect; F – uneven; M – tabular crystals, foliated aggregates, massive.
Origin and occurrence: Metamorphic in metamorphosed basic rocks and dolomites, together with actinolite, dolomite and chlorite; hydrothermal in veins. Tabular crystals up to 20 mm ($^{25}/_{32}$ in) across come from Brumado, Bahia, Brazil and Chester, Massachusetts, USA. Rich foliated aggregates are known from Zillertal, Austria. Pseudo-morphs after quartz crystals come from St. Gotthard, Switzerland. Large deposits of massive talc are mined in China.
Application: as a filling material in textile, paper and chemical industries, heat-resistant material.

Talc, 61 mm, Rochester, U.S.A.

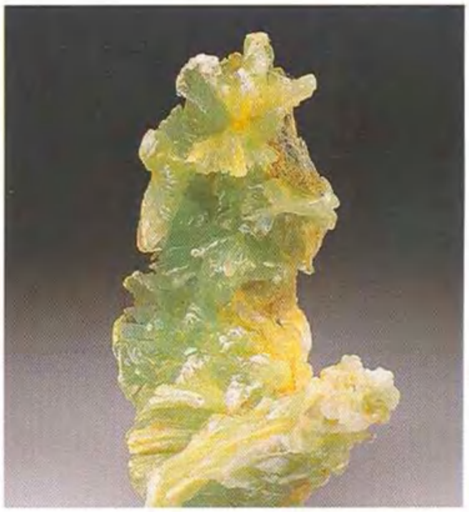

Muscovite

MICA GROUP

$KAl_3Si_3O_{10}(OH,F)_2$

MONOCLINIC ● ● ● ● ●

Varieties: fuchsite
Properties: C – white, greenish, yellowish, pinkish, green (fuchsite), gray; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, lamellar and scaly aggregates, massive.
Origin and occurrence: Metamorphic in various rock types, as mica schists, gneisses; magmatic in granites and pegmatites; hydrothermal in veins or next to ore veins (fuchsite). Important rock-forming mineral, usually associated with quartz, K-feldspar, albite and biotite. Its sheets, reached up to 5 x 3 m in size and weighed up to 85 tons in the Inikurti mine, Nellore, India. Large sheets are also known from Custer, South Dakota, USA. Well-formed crystals up to 100 mm (4 in) across come from Alabashka near Murzinka, Ural mountains, Russia. Other famous localities include Mamsk, Ural mountains, Russia and Cruzeiro mine, Minas Gerais, Brazil, where gemmy crystals were found.
Application: insulation material in electrical applications and construction.

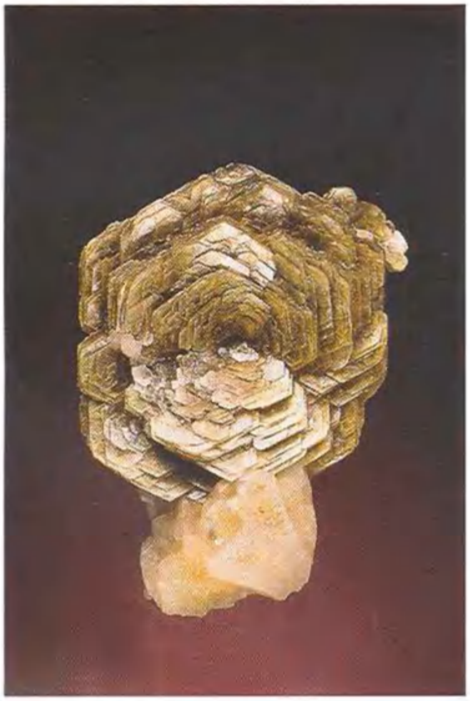

Muscovite, 35 mm, Plumos, California, U.S.A.

Celadonite

MICA GROUP

$KFe^{3+}(Mg,Fe^{2+},Al)Si_4O_{10}(OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – light green to blue-green; S – white; L – dull; D – translucent to almost opaque; DE – 3.0; H – 2; CL – perfect; F – uneven; M – earthy, sometime small scaly aggregates, massive.
Origin and occurrence: Hydrothermal as a product of mafic mineral replacement in volcanic rocks, associated with zeolites, prehnite and calcite. Famous localities are Val di Fassa and Monte Balda, Italy; also known from Bisbee, Arizona, USA.

Boromuscovite

MICA GROUP

$KAl_2BSi_3O_{10}(OH)_2$

MONOCLINIC ●

Properties: C – yellowish; S – white; L – dull; D – transparent to translucent; DE – 2.8; H – 2.5; CL – perfect; F – uneven; M – small scaly aggregates, massive.
Origin and occurrence: Hydrothermal in cavities in

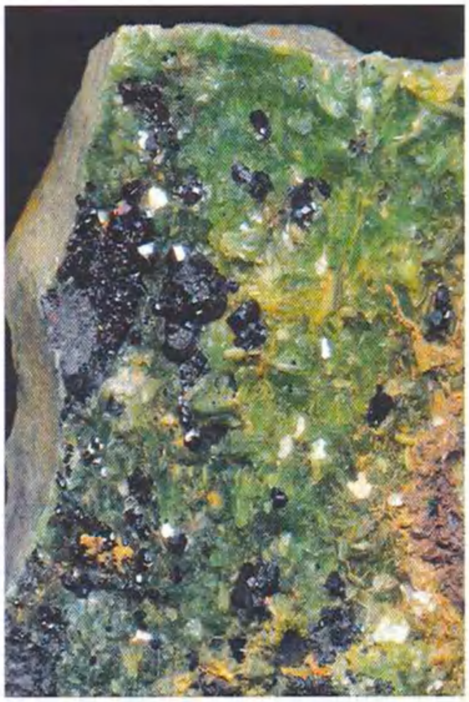

Fuchsite, 4 mm xx, Pamir, Tadzhikistan

Celadonite, 60 mm, Unhošťany, Czech Republic

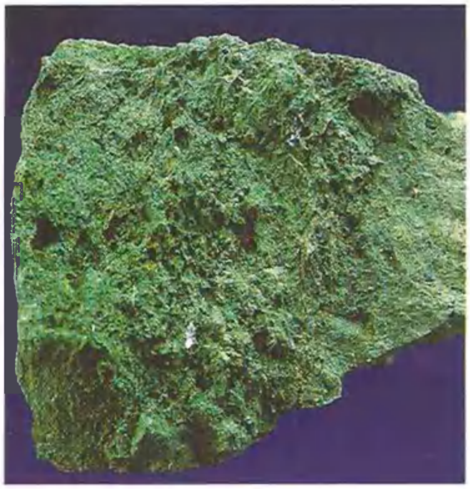

complex pegmatites, associated with albite, elbaite and lepidolite. Fine-grained scaly aggregates were described from the Little Three mine, Ramona, California, USA; also in Řecice, Czech Republic.

Paragonite

MICA GROUP

$NaAl_3Si_3O_{10}(OH,F)_2$

MONOCLINIC ● ● ●

Properties: C – white, greenish, yellowish, pinkish; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.8; H – 2.5; CL B perfect; F – uneven; M – tabular crystals, scaly aggregates, massive.
Origin and occurrence: Metamorphic in various rock types, as mica schists and gneisses. Fine-grained scaly aggregates are known from Pizzo Forno, Switzerland and elsewhere.

Boromuscovite, 40 mm, Ramona, U.S.A.

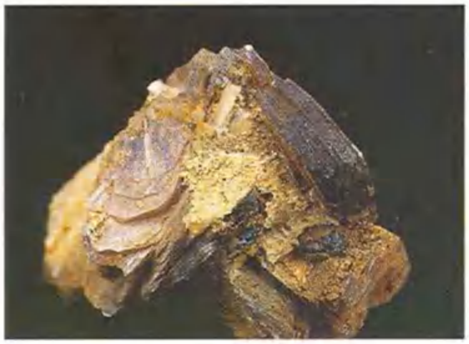

Paragonite, 80 mm, St.Gotthard, Switzerland

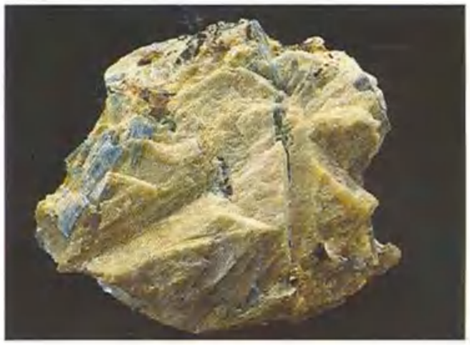

Glauconite

MICA GROUP

$K_{0.8}(Al,Fe^{2+},Fe^{3+},Mg)(Si,Al)_4O_{10}(OH)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – light green, yellow-green to blue-green; S – light green; L – dull; PS – translucent to opaque; DE – 2.9; H – 2; CL B perfect; F – uneven; M B earthy and platy aggregates, massive.
Origin and occurrence: Hydrothermal in sedimentary and volcanic sedimentary rocks, also as a product of replacement of mafic minerals in volcanic rocks. Common in sandstones and limestones, locally associated also with phosphorites. It occurs in many localities in the Karpathians, Poland; in Polabi region, Czech Republic. Massive aggregates are known from the N'Chwaning No. 2 mine, Kuruman, South Africa.

Glauconite, 80 mm, Malomerice, Czech Republic

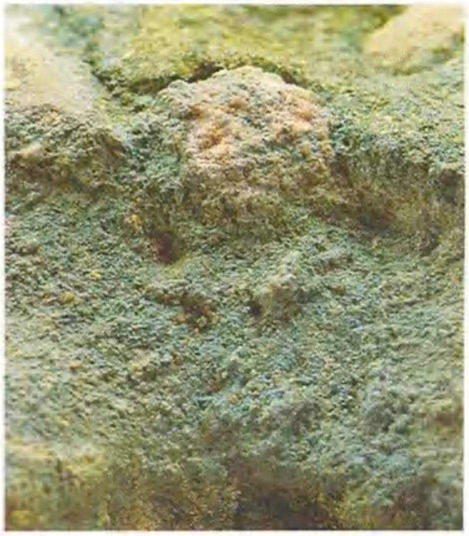

Annite, 120 mm, Budec, Czech Republic

Annite

MICA GROUP

$KFe_3AlSi_3O_{10}(OH,F)_2$

MONOCLINIC ● ● ●

Properties: C – black, locally with reddish or greenish tint; S – colorless; L – vitreous to pearly; D – transparent to translucent, locally opaque; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, lamellar aggregates, massive.
Origin and occurrence: Metamorphic in skarns; magmatic in some pegmatites and granites. Its lamellar aggregates are known from Langban, Sweden and Cape Ann, Massachusetts, USA. Crystals up to 150 mm (6 in) across come from Mont St.-Hilaire, Quebec, Canada.

Phlogopite

MICA GROUP

$KMg_3AlSi_3O_{10}(OH,F)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – light brown, greenish, yellowish, colorless, gray; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, platy and scaly aggregates, massive.
Origin and occurrence: Metamorphic in various rock types, as marbles and some ultrabasic rocks, at the contacts of pegmatites and serpentinites, magmatic in some ultrabasic rocks and pegmatites, associated with dolomite, diopside, anthophyllite. Rich scaly aggregates come from many localities, as Pargas, Finland and elsewhere. Crystals up to 5 m (16 ft) across found in Sludyanka, Siberia, Russia. Crystals up to 50 cm (20 in) across occurred in the Gardiner complex, Greenland. The largest crystals, 10 x 5 m (33 x 16 ft) across, weighing up to 90 tons, come from the Lacy mine, Ontario, Canada.
Application: as insulation material in electrical applications.

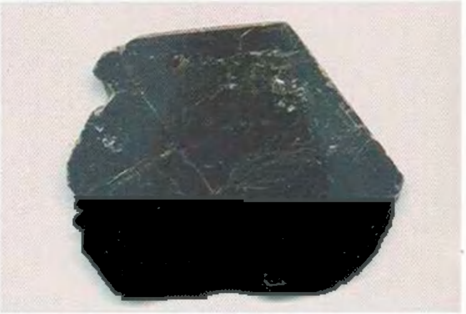

Biotite

MICA GROUP

$K(Fe,Mg)_3AlSi_3O_{10}(OH,F)_2$

MONOCLINIC ● ● ● ● ●

Properties: C – brown to black, commonly with red or green tint; S – colorless; L – vitreous to pearly; D – transparent to translucent, locally opaque; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, scaly aggregates, massive.
Origin and occurrence: Metamorphic in various rock types, as mica schists, gneisses, migmatites and different types of metamorphosed shales; magmatic in pegmatites, granites, syenites and diorites, rare in basalts and ultrabasic rocks. Typical rock-forming mineral, usually associated with quartz, feldspars and muscovite. Rich scaly aggregates are known from a pegmatite in Evje, Norway, where its crystals reach up to several meters across. Other famous localities are Bessnes, France; Uluguru mountains, Tanzania; Silver Crater mine, Ontario, Canada and Laacher See, Germany.

Polylithionite

MICA GROUP

$KLi_2AlSi_4O_{10}(F,OH)_2$

MONOCLINIC ● ● ●

Properties: C – gray, colorless, yellowish, purple; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M B tabular crystals, scaly aggregates, massive.
Origin and occurrence: Magmatic and locally also hydrothermal in alkaline pegmatites, granites and carbonatites, rare in Li-bearing granitic pegmatites. Well-formed crystals up to 40 mm ($1\frac{9}{16}$ in) across are known from cavities in alkaline pegmatites from Mont St.-Hilaire, Quebec, Canada. It comes also from Ilimaussaq, Greenland and from cavities in granitic pegmatites in Recice, Czech Republic and

Biotite, 30 mm x, Ontario, Canada

elsewhere. Large industrial deposits known in Blatchford Lake, Northwest Territories, Canada.
Application: Li and Cs ore.

Trilithionite

MICA GROUP

$KLi_{1,5}Al_{2,5}Si_3O_{10}(F,OH)_2$

MONOCLINIC ● ● ● ●

Properties: C – purple, pink, blue, colorless, green; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, scaly aggregates, massive.
Origin and occurrence: Magmatic in Li-bearing pegmatites and granites, commonly associated with elbaite, spodumene, petalite, quartz and albite; sometimes hydrothermal in quartz veins and pegmatites. Well-formed crystals from pegmatite cavities are known from Virgem da Lapa, Minas Gerais, Brazil. Massive aggregates come from the Stewart Lithia mine, Pala, California and the Brown Derby No. 1 mine, Colorado, USA; Varutrask, Sweden; Rožna , Czech Republic; Meldon Quarry, Devon, UK.
Application: Li and Cs ore.

Polylithionite, 30 mm, Lovozero Massif, Kola, Russia

Zinnwaldite

MICA GROUP

$KLiFeAl_2Si_3O_{10}(F,OH)_2$

MONOCLINIC ● ● ●

Properties: C – gray, colorless, brown; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 3.7; H – 2.5-3; CL – perfect; F – uneven; M – tabular crystals, scaly aggregates, massive.
Origin and occurrence: Hydrothermal in quartz veins and greisens; magmatic in pegmatites and granites, associated with fluorite, cassiterite and wolframite. Rich aggregates with scales up to 100 mm (4 in) across are known from quartz veins in Cínovec, Czech Republic. Crystals up to 150 mm (6 in) across come from pegmatite cavities in Virgem da Lapa, Minas Gerais, Brazil. It also occurs in the Pikes Peak batholith, Colorado, USA; Baveno, Italy.
Application: Li ore.

Trilithionite, 70 mm, Tanco, Canada

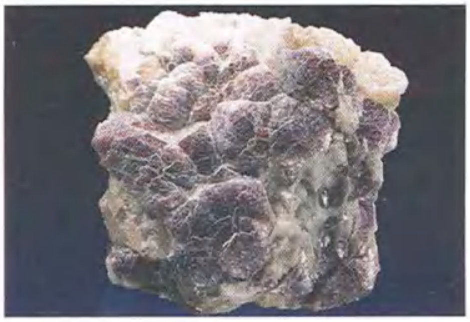

Zinnwaldite, 10 mm xx, Cinovec, Czech Republic

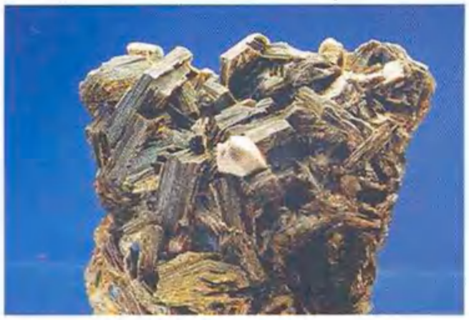

Margarite, 73 mm, Chester, U.S.A.

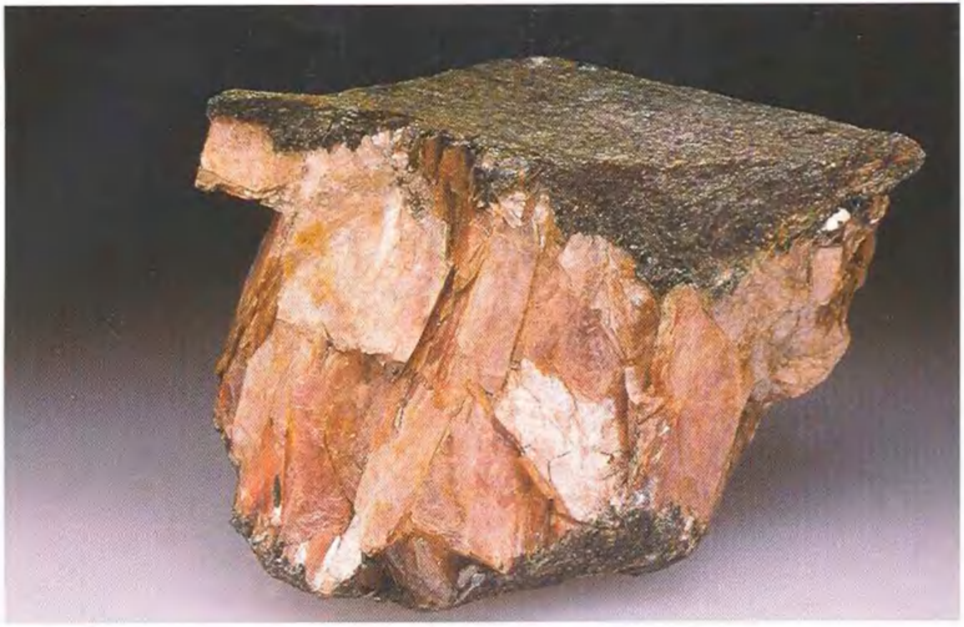

Margarite

$CaAl_4Si_2O_{10}(OH)_2$

MONOCLINIC ● ● ●

Properties: C – pinkish, colorless, yellowish; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 3.1; H – 3.5-4.5; CL B perfect; F – uneven; M – tabular crystals, platy and scaly aggregates, massive.
Origin and occurrence: Metamorphic, associated with corundum, diaspore, tourmaline and staurolite in various types of metamorphosed shales. Coarse platy aggregates come from Chester, Massachusetts and Sterling Hill, New Jersey, USA.

Clintonite, 3 mm xx, Pomaz, Hungary

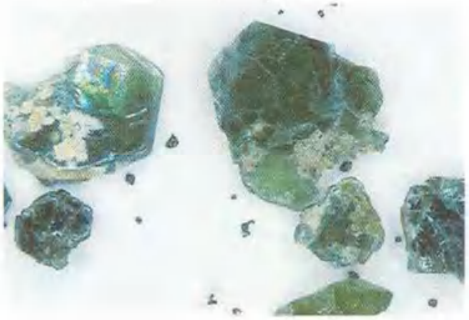

Clintonite

$Ca(Mg,Al)_3(Al_3Si)O_{10}(OH)_2$

MONOCLINIC ● ●

Properties: C – colorless, yellowish, pinkish, greenish; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 3.1; H – 3.5; CL – perfect; F – uneven; M – tabular crystals, lamellar aggregates, massive.
Origin and occurrence: Metamorphic in contact metamorphosed marbles, associated with vesuvianite, grossular, diopside and spinel. Rich aggregates with lamellae up to 20 mm ($^{25}/_{32}$ in) across occur in Green Monster mountain, Alaska and Crestmore, California, USA and Monzoni, Italy.

Stilpnomelane, 100 mm, Horni Udoli, Czech Republic

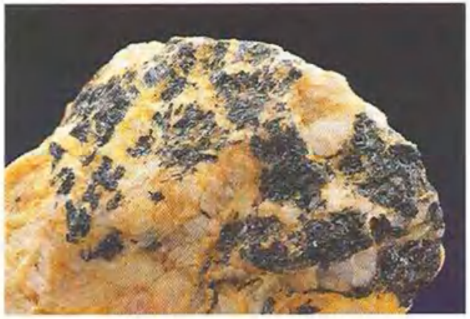

Stilpnomelane

$K(Fe^{2+},Mg,Fe^{3+})_8(Si,Al)_{12}(O,OH)_{27}$

MONOCLINIC ● ● ● ●

Properties: C – black, black-brown, black-green, yellow-brown; S – colorless; L – vitreous to dull; D – translucent to opaque; DE – 2.8; H – 3; CL – perfect; F – uneven; M – tabular crystals, foliated, lath-like and acicular aggregates..
Origin and occurrence: Metamorphic in Fe-rich shales, usually associated with chlorite, magnetite and albite. Crystals up to 20 mm ($^{25}/_{32}$ in) across come from Jim Pond township, Maine, USA. Foliated aggregates are known from Horni Udoli near Zlate Hory, Czech Republic; Mesabi Range, Minnesota, USA.

Montmorillonite

$(Ca,Na)_{0.33}(Al,Mg)_2Si_4O_{10}(OH)_2 \cdot n\ H_2O$

MONOCLINIC ● ● ● ●

Properties: C – white, yellowish, greenish, bluish; S – white; L – greasy; D – translucent to opaque; DE – 2.3; H – 1-2; CL – perfect, wet massive aggregates are plastic; F – uneven; M – earthy aggregates, massive.
Origin and occurrence: Hydrothermal as a product of replacement of other minerals in volcanic rocks, granitic pegmatites and sediments. It is abundant in many localities. Montmorillonite deposits are known from Antrim, Northern Ireland, UK; in Hungary; Slovakia and many places in the USA.
Application: ceramics and chemical industry.

Saponite

$(Ca,Na)_{0.3}(Fe,Mg)_3(Si,Al)_4O_{10}(OH)_2 \cdot 4\ H_2O$

MONOCLINIC ● ● ●

Properties: C – white, yellowish, gray-green, bluish; S – white; L – greasy; D – translucent to opaque; DE – 2.1; H – 1-2; CL – perfect, massive aggregates are plastic under wet conditions; F – uneven; M – earthy, nodular, massive.
Origin and occurrence: Hydrothermal as a product of replacement of other minerals in volcanic rocks and serpentinites. It occurs together with copper in many places in Keweenaw Peninsula, Michigan, USA. Scales up to 10 mm ($^{3}/_{8}$ in) across described from Mont St.-Hilaire, Quebec, Canada.

Montmorillonite, Kisthaas, Syria

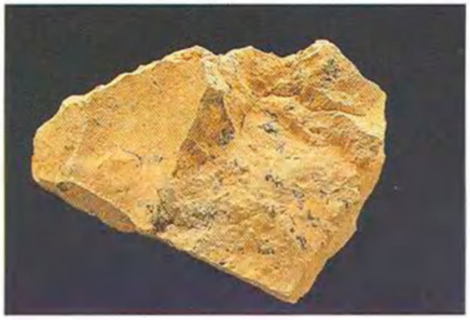

Saponite, 2 mm crust, Erdobenye, Hungary

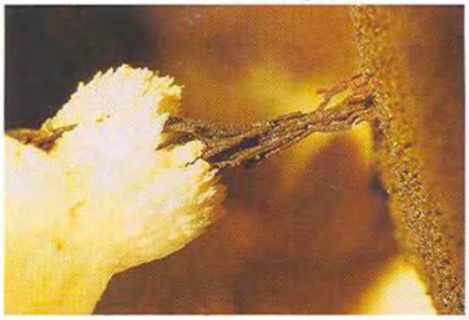

Vermiculite

$(Fe,Mg,Al)_3(Si,Al)_4O_{10}(OH)_2 \cdot 4\ H_2O$

MONOCLINIC ● ● ●

Properties: C – yellow-brown, gray-green, green-brown; S – white; L – greasy; D – translucent to opaque; DE – 2.5; H – 1.5; CL B perfect; F – uneven; M – scaly aggregates, massive; R – it expands when heated.
Origin and occurrence: Hydrothermal as a product of replacement of phlogopite and other mafic micas in various rock types. It occurs in Palbora, South Africa; Milbury, Massachusetts, USA; Kovdor massif, Kola Peninsula, Russia.
Application: electrical applications and paper industry.

Vermiculite, 50 mm, Drahonin, Czech Republic

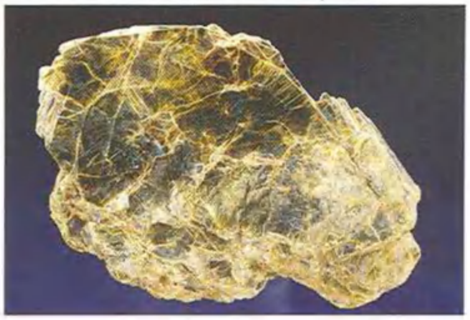

Cookeite, 60 mm, Dobra Voda, Czech Republic

Cookeite

CHLORITE GROUP

$LiAl_4Si_3O_{10}(OH)_8$

MONOCLINIC ● ●

Properties: C – white, yellowish, pinkish, brown; S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 2.6; H – 2.5-3.5; CL – perfect; F – uneven; M B scaly, locally radial aggregates, massive.

Origin and occurrence: Hydrothermal in cavities in Li-bearing pegmatites, associated with elbaite, lepidolite, fluorapatite and albite, it also replaces spodumene; metamorphic in Al-rich shales, together with diaspore and pyrophyllite. Massive aggregates, replacing spodumene, occur in the Tanco mine, Bernic Lake, Manitoba, Canada. It also comes from cavities in pegmatites in the Little Three mine, Ramona and the Himalaya mine, Mesa Grande, California. It is also known from the Pulsifer quarry, Maine, USA and Muiane, Mozambique.

Clinochlore, 5 mm xx, Sarany, Russia

Clinochlore

CHLORITE GROUP

$(Mg,Fe)_5Al_2Si_3O_{10}(OH)_8$

MONOCLINIC ● ● ● ● ●

Varieties: kammererite

Properties: C – greenish, gray, white, yellowish, brown, red-purple (kammererite); S – colorless; L – vitreous to pearly; D – transparent to translucent; DE – 2.7; H – 2-2.5; CL – perfect; F – uneven; M B poorly-developed tabular crystals, foliated and radial aggregates, massive.

Kammererite, 11 mm xx, Kop Daglari, Turkey

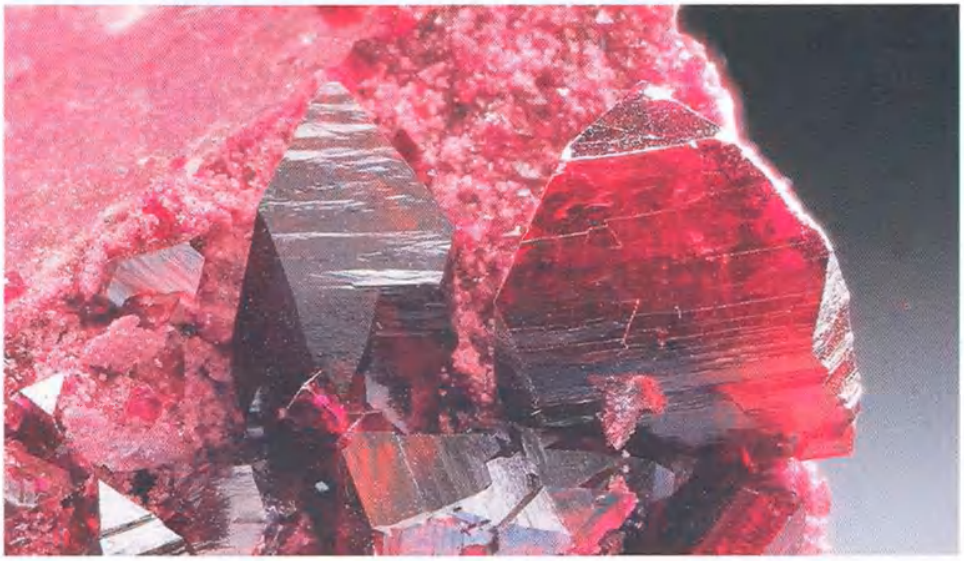

Chamosite, 70 mm, Hajan, Hungary

Origin and occurrence: Metamorphic in various types of shales and marbles; hydrothermal in quartz veins and Alpine-type veins, occasionally replaces certain minerals as biotite. Crystals up to 50 mm (2 in) across, associated with chondrodite and magnetite, are known from the Tilly Foster mine, Brewster, New York and Chester, Pennsylvania, USA. It also occurs in Val d'Ala, Italy and Zillertal, Austria. Scales up to 40 cm (15¾ in) across come from Beramy, Madagascar. Kammererite crystals up to 20 mm (25/32 in) across found in Kop Daolari, Erzerum, Turkey.

Chamosite

CHLORITE GROUP

$(Fe,Mg)_5Al_2Si_3O_{10}(OH)_8$

MONOCLINIC ● ● ● ●

Properties: C – gray, gray-green, brown; S – gray-green; L – vitreous to dull; D – translucent to opaque; DE – 3.2; H – 2-2.5; CL – perfect; F – uneven; M B scaly and oolitic aggregates, massive.
Origin and occurrence: Metamorphic to hydrothermal in various types Fe-rich sediments, typically associated with siderite and magnetite.
It is common in Chamoson, Switzerland; Nučice, Czech Republic. Spherical aggregates up to 15 mm (19/32 in) in diameter found in Mont St.-Hilaire, Quebec, Canada.

Kaolinite

$Al_2Si_2O_5(OH)_4$

TRICLINIC ● ● ● ● ●

Properties: C – white, yellowish, greenish, gray; C – colorless; L – earthy; D – translucent to opaque; DE – 2.6; H – 1; CL – perfect; F – uneven; M – earthy and exceptionally scaly aggregates, massive.
Origin and occurrence: Hydrothermal, as a result of feldspar replacement in various rock types, as granites and arcoses. It forms large deposits in China, France, UK and Czech Republic.
Application: raw material for ceramics.

Dickite, 100 mm, Nowa Ruda, Poland

Dickite

$Al_2Si_2O_5(OH)_4$

MONOCLINIC ● ● ●

Properties: C – white, yellowish; S – colorless; L – earthy to pearly; D – translucent to opaque; DE – 2.6; H – 1; CL – perfect; F – uneven; M – earthy and platy aggregates, massive.
Origin and occurrence: Hydrothermal in cavities in hydrothermal veins, together with quartz, carbonates and sulfides. It occurs in Essen, Germany; Anglesey, Wales , UK; Kladno, Czech Republic. Microscopic crystals come from Mas d'Alary, France.

Kaolinite, 50 mm, St.Austell, UK

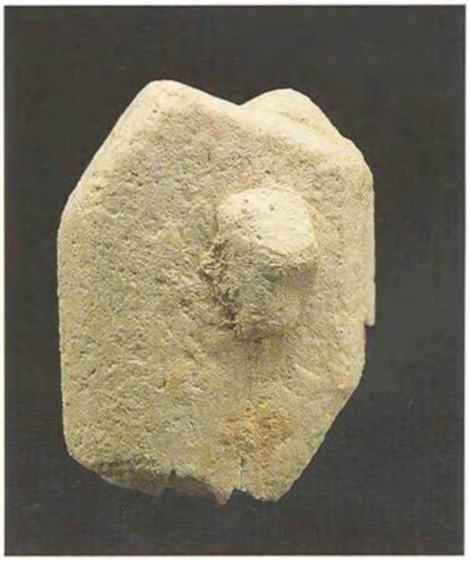

Nakrite, 70 mm, Rochlitz, Germany

Nacrite

$Al_2Si_2O_5(OH)_4$

MONOCLINIC ● ● ●

Properties: C – white, yellowish, gray-white; S – colorless; L – earthy to pearly; D – translucent to opaque; DE – 2.6; H – 1; CL – perfect; F – uneven; M – earthy and scaly aggregates, massive.
Origin and occurrence: Hydrothermal in cavities of hydrothermal veins, associated with quartz, fluorite and topaz. It occurs in Horni Slavkov, Czech Republic; Freiberg, Germany. Pseudo-morphs after topaz come from Ouro Preto, Minas Gerais, Brazil.

Antigorite

$Mg_3Si_2O_5(OH)_4$

MONOCLINIC ● ● ● ● ●

Properties: C – white, yellowish, greenish, gray; S – colorless to gray; L – vitreous, pearly, silky, earthy; D – translucent to opaque; DE – 2.6; H – 2.5-3.5; CL – perfect; F – uneven; M B foliated, lamellar and fibrous aggregates, massive.
Origin and occurrence: Metamorphic in serpentinites, marbles and other Mg-rich rocks; hydrothermal in veins cross-secting these rocks. It is important rock-forming mineral. It occurs in Kraubat, Austria; Hrubsice, Czech Republic; Antigorio, Italy and elsewhere.

Antigorite, 110 mm, Felsocsatar, Hungary

Amesite

$Mg_2Al(SiAl)O_5(OH)_4$

TRICLINIC ● ●

Properties: C – greenish, gray; S – colorless; L – vitreous to pearly; D – translucent; DE – 2.8; H – 2.5-3; CL – perfect; F – uneven; M B foliated to radial aggregates, massive.
Origin and occurrence: Hydrothermal in Mg-rich rocks, associated with diaspore, magnetite and chromite. It comes from Chester, Massachusetts, USA. Crystals up to 40 mm ($1\frac{9}{16}$ in) across found in Sarany, Ural mountains, Russia.

Cronstedtite

$Fe^{2+}_2Fe^{3+}_2SiO_5(OH)_4$

MONOCLINIC ● ●

Properties: C – black, black-brown, black-green; S – dark olive-green; L – vitreous to submetallic; D – translucent to opaque; DE – 3.6; H – 3.5; CL – perfect; F – uneven; M – prismatic crystals, columnar aggregates.
Origin and occurrence: Hydrothermal in ore veins. Fan-shaped aggregates of crystals come from Pribram and Kutna Hora, Czech Republic; Ouro Preto, Minas Gerais, Brazil; Salsigue mine, Auge, France.

Amesite, 40 mm, Sarany, Ural Mts., Russia

Cronstedtite, 42 mm, Hunan, Bolivia

Chrysocolla, 50 mm, Arizona, U.S.A.

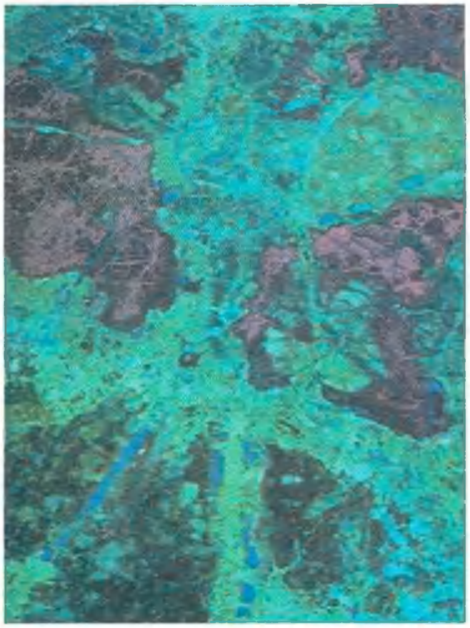

Chrysocolla

$(Cu,Al)_2H_2Si_2O_5(OH)_4 \cdot n\,H_2O$

MONOCLINIC ● ● ● ●

Properties: C – different hues of blue-green; S – white; L – vitreous, greasy to earthy; D – translucent to almost opaque; DE – 2.0-2.4; H – 2-4; CL – none; F – conchoidal to uneven; M – microscopic acicular crystals, botryoidal, stalactitic and earthy aggregates.
Origin and occurrence: Secondary in Cu deposits, associated with malachite and other secondary Cu minerals.
Rich aggregates come from many places in Arizona, USA, like Bisbee and Morenci. It is also known from Mednorudnyansk, Ural mountains, Russia and Broken Hill, New South Wales, Australia.

Allophane

approximately $Al_2SiO_5 \cdot n\,H_2O$

AMORPHOUS ● ● ●

Properties: C – white, gray, bluish, greenish, brown; S -white; L – vitreous, greasy and earthy; D – transparent, translucent to opaque; DE – 1.9; H – 2-3; CL – none; F – conchoidal to uneven; M – botryoidal, stalactitic and earthy aggregates.
Origin and occurrence: Hydrothermal product of alteration along the cracks of sedimentary rocks in coal deposits and in the oxidation zone of ore deposits, associated with other secondary minerals. Rich botryoidal aggregates occur in Dehr, Germany; Moldova Nuova, Romania; New Cornelia mine, Arizona, USA; El Dragon mine, Potosi, Bolivia.

Allophane, 90 mm, Zlate Hory, Czech Republic

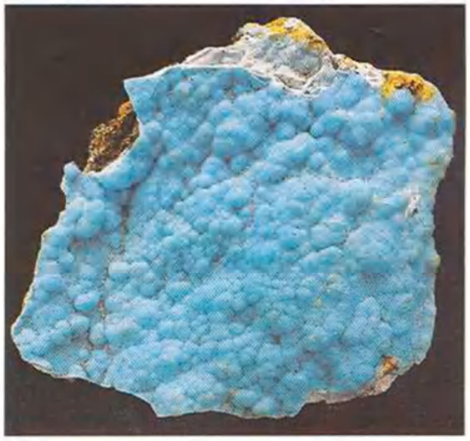

Palygorskite, 120 mm, Burda, Hungary

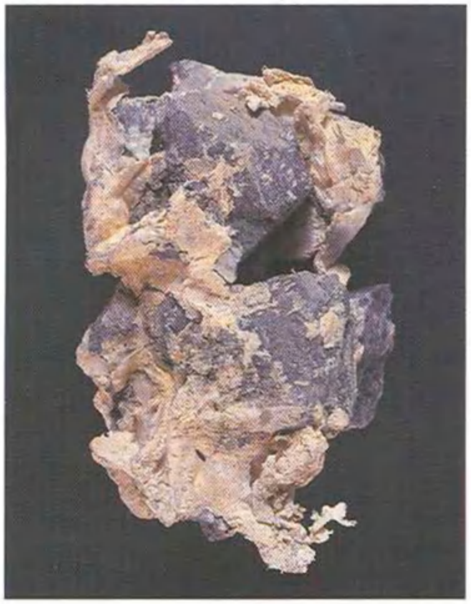

Palygorskite

$(Mg,Al)_2Si_4O_{10}(OH) . 4 H_2O$

MONOCLINIC ● ● ● ●

Properties: C – white, gray; S – white; L – dull; D - translucent to almost opaque; DE – 2.2; H – 1; CL – good; F – uneven, aggregates plastic under wet conditions; M – acicular crystals, fibrous aggregates, massive.
Origin and occurrence: Hydrothermal as a product of alteration of different rock types, such as serpentinites, granites, marbles and graywackes, also in ore veins and the Alpine-type veins. Rich aggregates resembling leather occur along the cracks in marbles in Hejná near Horazd'ovice, Czech Republic; in the Mammoth mine, Tiger, Arizona, USA and in Palygorskava, Russia.

Sepiolite, 70 mm, Eskisehir, Turkey

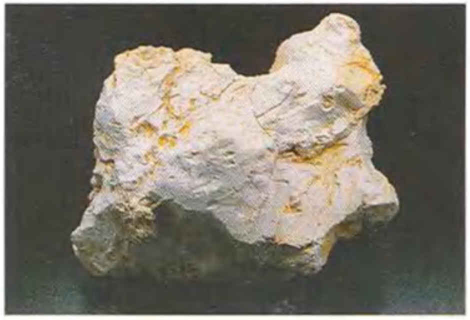

Sepiolite

$Mg_4Si_6O_{15}(OH)_2 . 6 H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – white, yellowish, gray; S – white; L – dull; D – opaque; DE – 2.0; H – 2-2.5; CL – not determined; F – uneven; M – earthy aggregates, massive.
Origin and occurrence: Hydrothermal as a product of serpentinite alteration, typically associated with magnesite. Classic locality is Eskişehir, Turkey; also known from Biskoupky, Czech Republic.

Zeophylite

$Ca_4Si_3O_8(OH,F)_4 . 2 H_2O$

TRICLINIC ● ●

Properties: C – white; S – white; L – pearly; D – transparent to translucent; DE – 2.6; H – 3; CL – perfect; F – uneven; M – platy crystals, commonly with radial structure.
Origin and occurrence: Hydrothermal in cavities of volcanic rocks, associated with zeolites. Its most important locality is Radejeín, Czech Republic, where it forms spherical aggregates up to 10 mm ($^3/_8$ in) in diameter. Also known from Schellkopf, Germany and Monte Somma, Italy.

Cavansite

$CaVOSi_4O_{10} . 4 H_2O$

ORTHORHOMBIC ● ●

Properties: C – green-blue, blue; S – light blue; L – vitreous; D – transparent; DE – 2.2; H – 3-4; CL – good; F – uneven; M – long prismatic to acicular crystals, radial aggregates.
Origin and occurrence: Hydrothermal in cavities of volcanic rocks, associated with calcite, apophyllite and zeolites. Rich radial aggregates up to 30 mm

Zeophyllite, 60 mm, Radejcin, Czech Republic

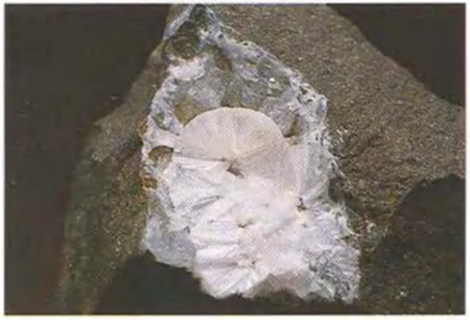

Cavansite, 120 mm, Poona, India

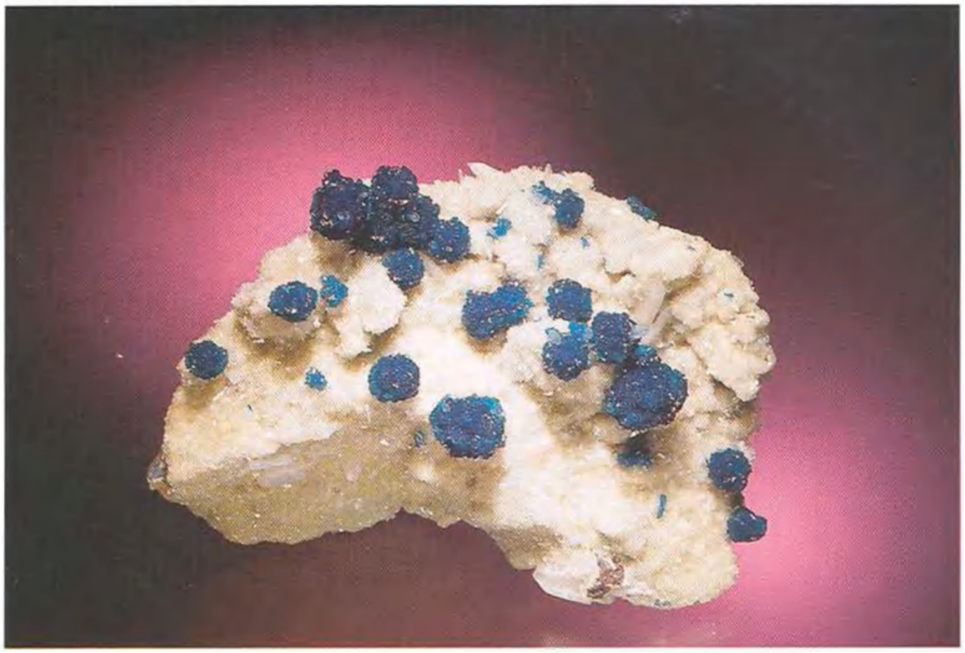

($1^{1}/_{16}$ in) in diameter come from the Wagholi Quarry, Poona, India.

Nepheline
$(K,Na)AlSiO_4$

HEXAGONAL ● ● ● ●

Properties: C – colorless, white, greenish, yellowish, gray, green, brown; S – white; L – vitreous to greasy; D – transparent to translucent; DE – 2.7; H – 5.5-6; CL – none; F – uneven to conchoidal; M – prismatic crystals, granular, massive.
Origin and occurrence: Magmatic in many alkaline rocks, as alkaline syenites and their pegmatites, also in some basalts; rare metamorphic in gneisses. It is a typical rock-forming mineral, associated with leucite, augite and apatite. Well-formed crystals found in Mount Vesuvius, Italy. Rich aggregates are known from many localities in Khibiny massif, Kola Peninsula, Russia. Crystals up to 70 cm ($27^{9}/_{16}$ in) long occur in Davis Hill, Bancroft, Ontario. Perfect crystals up to 35 mm ($1^{3}/_{8}$ in) long, come from Mont St.-Hilaire, Quebec, Canada.
Application: locally as Al ore and in ceramic industry.

Nepheline, 43 mm, Aouli, Morocco

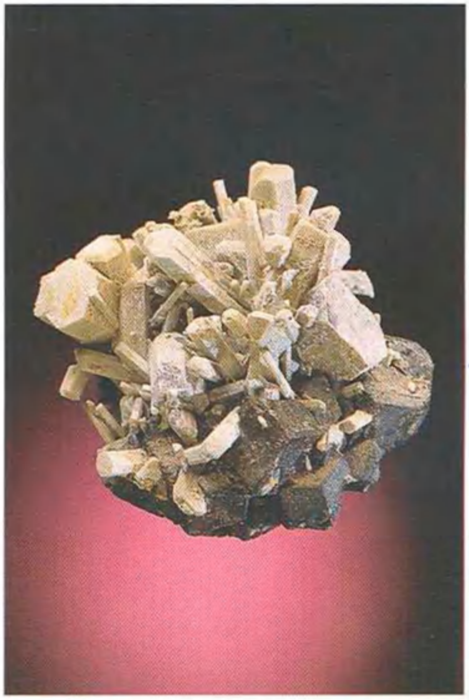

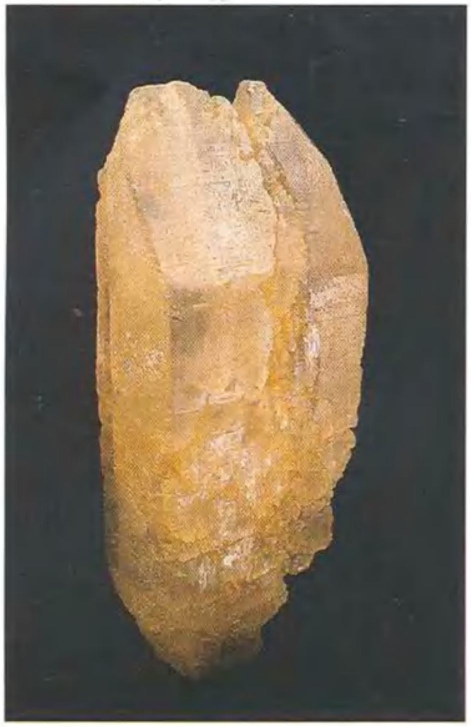
Petalite, 23 mm, Paprok, Afghanistan

Petalite

$LiAlSi_4O_{10}$

MONOCLINIC ● ● ●

Properties: C – colorless, white, greenish, yellowish; S – white; L – vitreous; D – transparent to translucent; DE – 2.5; H – 6-6.5; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular.
Origin and occurrence: Magmatic in Li-bearing pegmatites, associated with lepidolite, elbaite, amblygonite and locally replaced by a mixture of quartz and spodumene. Gigantic crystals several meters long, come from Bikita, Zimbabwe; and Varutrask, Sweden. Clear gemmy crystals up to 230 mm ($9^{1}/_{16}$ in) across found in pegmatite cavities in Paprok, Afghanistanand also in San Piero in Campo, Elba, Italy. Crystals up to 100 mm (4 in) across occur in Araçuai, Minas Gerais, Brazil.
Application: ceramic industry.

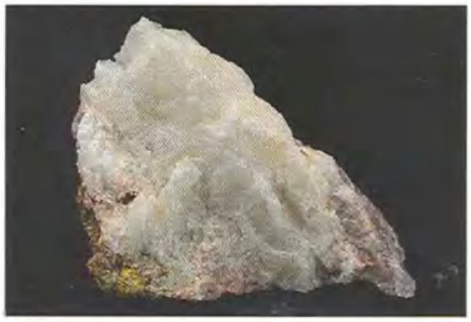
Chkalovite, 70 mm, Umbozero, Kola, Russia

Chkalovite

$Na_2BeSi_2O_6$

ORTHORHOMBIC ● ●

Properties: C – colorless, white; S – white; L – vitreous; D – transparent to translucent; DE – 2.7; H – 6; CL – imperfect; F – conchoidal; M – prismatic crystals, granular, massive.
Origin and occurrence: Hydrothermal in alkaline pegmatites, usually associated with natrolite, sodalite and eudialyte. Crystals up to 200 mm ($7^{7}/_{8}$ in) across come from the Umbozero mine, Mount Alluaiv, Lovozero massif, Kola Peninsula, Russia; also at Julianhab, Greenland and Mont St.-Hilaire, Quebec, Canada.

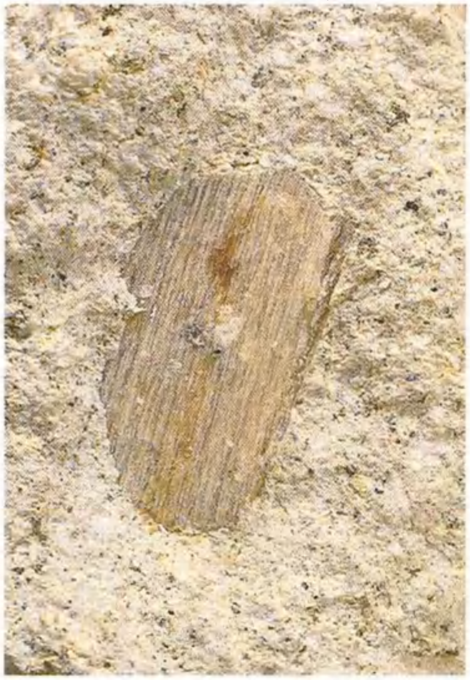
Sanidine, 30 mm x, Drachenfels, Germany

Sanidine

FELDSPAR GROUP

$KAlSi_3O_8$

MONOCLINIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish, gray; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 6-6.5; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.
Origin and occurrence: Magmatic in acid volcanic rocks, as rhyolites and trachytes, a typical rock-forming mineral. Well-formed crystals and their combinations up to 100 mm (4 in) in size occur in rhyolites and trachytes in Drachenfels and Laacher See, Germany. It also comes from Roc de Courlande, France; Beaverdell, British Columbia, Canada; Kyustendil, Bulgaria.

Orthoclase

FELDSPAR GROUP

$KAlSi_3O_8$

MONOCLINIC ● ● ● ● ●

Varieties: adularia, moonstone (gemmy variety with chatoyancy)

Properties: C – colorless, white, yellowish, pinkish, gray, brown, yellow; S – white; L – vitreous to pearly (adularia); D – transparent to translucent; DE – 2.5; H – 6-6.5; CL – good; F – uneven to conchoidal; M – prismatic and tabular crystals and their combinations, granular, massive.
Origin and occurrence: Magmatic in rhyolites, trachytes, granites, syenites and pegmatites; metamorphic in various rock types, as orthogneisses and migmatites; hydrothermal in the Alpine-type veins, ore veins and some sediments, also in placers, a typical rock-forming mineral. Well-formed crystals and twins up to 200 mm (7⅞ in) across come from granites in Twentynine Palms, California, USA; Marina di Campo, Elba, Italy; Loket and Karlovy Vary, Czech Republic; in pegmatite cavities in Strzegom, Poland; San Piero in Campo, Elba, Italy. Yellow gemmy crystals up to 70 mm (2¾ in) across found in Itrongay, Madagascar. Gigantic feldspar crystals up to tens of meters long known from several pegmatite localities in Black Hills, South Dakota, USA; also Hagendorf, Germany. Adularia is known from Alpine-type veins in St. Gotthard, Switzerland; pebbles of moonstone occur in gem-bearing gravels in Ratnapura, Sri Lanka.
Application: ceramic and glass industry, moonstone as a gemstone.

Orthoclase, 32 mm x, Oro Grande, U.S.A.

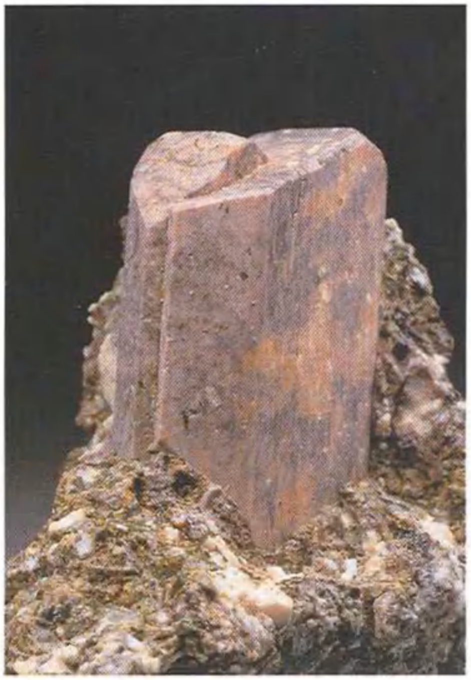

Adularia, 60 mm, Alps, Switzerland

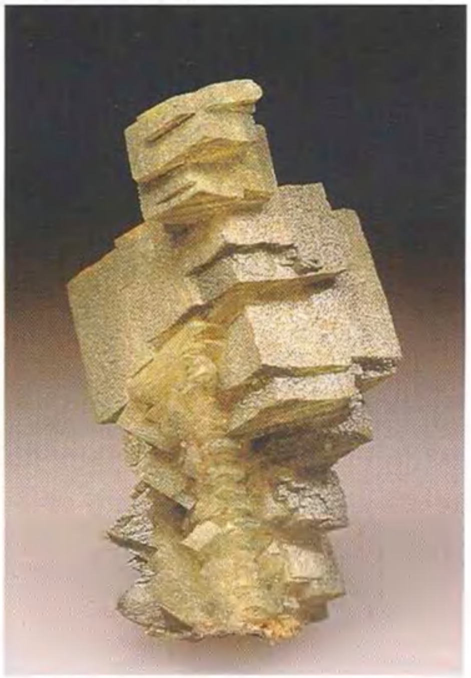

Microcline, 110 mm, Lake George, Teller Co., U.S.A.

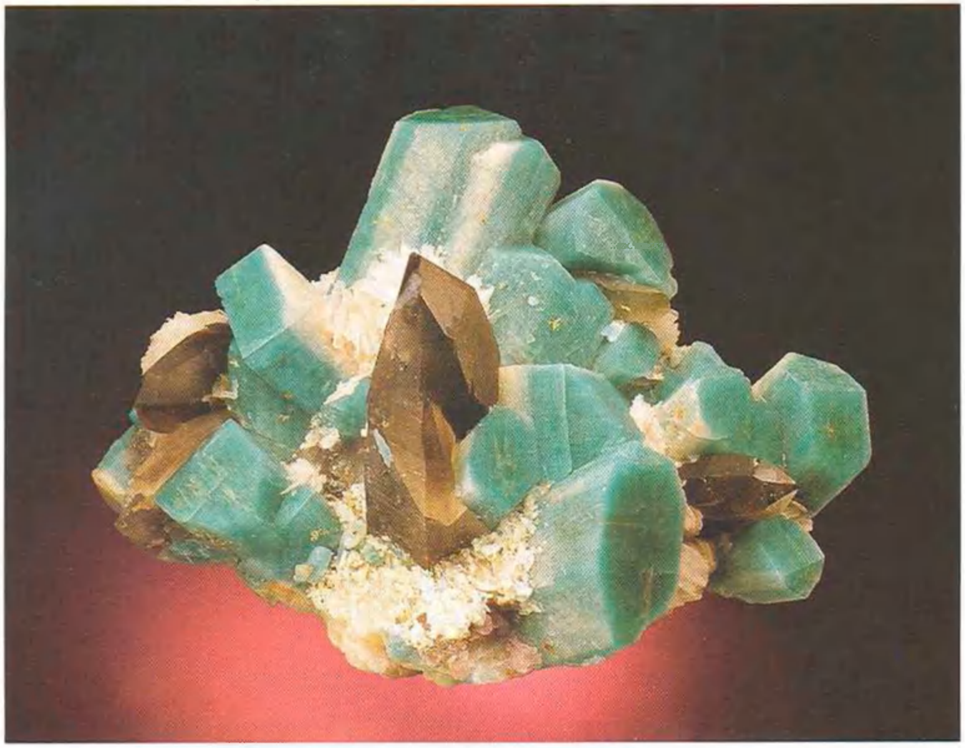

Microcline

FELDSPAR GROUP

$KAlSi_3O_8$

TRICLINIC ● ● ● ● ●

Varieties: amazonite

Properties: C – colorless, white, yellowish, pinkish, gray, light to dark green (amazonite); S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 6-6.5; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.

Origin and occurrence: Magmatic in granites, syenites and pegmatites; metamorphic in various rock types, as orthogneisses and migmatites; hydrothermal in the Alpine-type veins and ore veins; a typical rock-forming mineral. Well-formed amazonite crystals up to 40 cm (15¾ in) across occur in pegmatitic cavities in Crystal Peak, Colorado, USA; Keivy, Kola Peninsula, Russia; Morefield mine, Virginia, USA. Gigantic microcline crystals up to 12 m (39 ft) across come from pegmatites in Black Hills, South Dakota, USA; Kaatiala, Finland.

Application: ceramic industry, amazonite as a decorative stone.

Hyalophane

FELDSPAR GROUP

$(K,Ba)Al(Si,Al)_3O_8$

MONOCLINIC ● ● ●

Properties: C – colorless, white, yellowish; S – white; L – vitreous; D – transparent to translucent; DE – 2.9; H – 6-6.5; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.

Amazonite, 60 mm, Keivy, Kola, Russia

Hyalophane, 50 mm, Busovaca, Bosnia

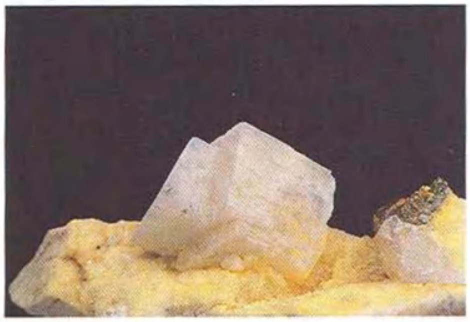

Origin and occurrence: Magmatic in phonolites; metamorphic in various rock types, as marbles, gneisses and Mn-rich rocks; hydrothermal in the Alpine-type and ore veins. Crystals up to 100 mm (4 in) across found in Alpine-type veins near Busovaca, Bosnia-Hercegovina.

Albite

FELDSPAR GROUP

$NaAlSi_3O_8$

TRICLINIC ● ● ● ● ●

Varieties: pericline, cleavelandite

Properties: C – colorless, white, yellowish, pinkish, gray, greenish, bluish; S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 6-6.5; CL – good; F – uneven to conchoidal; M – tabular crystals and their twins, platy aggregates, granular, massive.

Albite, 85 mm, Governador Valadares, Brazil

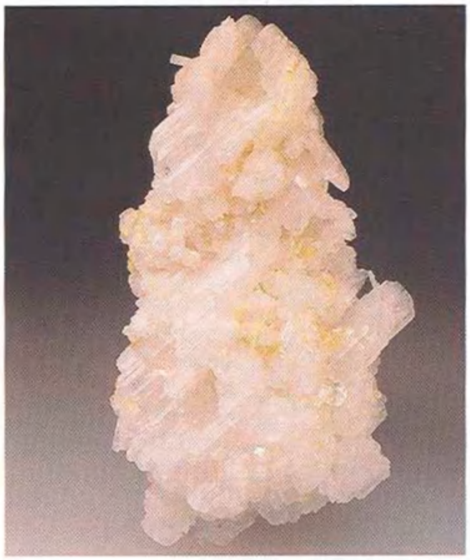

Oligoclase, 40 mm, Ural Mts., Russia

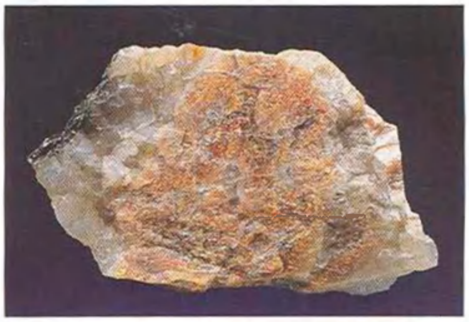

Origin and occurrence: Magmatic in granites, syenites and their pegmatites; metamorphic in various rock types, as orthogneisses, migmatites, phyllites and metamorphosed shales; hydrothermal in the Alpine-type veins and ore veins, a typical rock-forming mineral. Well-formed tabular to platy cleavelandite crystals up to 150 mm (6 in) across occur in pegmatite cavities in the Amelia district, Virginia and the Pala district, California, USA. Albite crystals also come from Strzegom, Poland; San Piero in Campo, Elba, Italy; Murzinka, Ural mountains, Russia; many localities in Minas Gerais, Brazil. Pericline crystals are known from the cracks along Alpine-type veins in Grossgreiner, Austria; St. Gotthard, Switzerland; also from ore veins in Rožnava, Slovakia.

Oligoclase

FELDSPAR GROUP

$(Na_{0,9-0,7}Ca_{0,1-0,3}Al_{1,1-1,3}Si_{2,9-2,7}O_8$

TRICLINIC ● ● ● ● ●

Varieties: peristerite, sunstone

Properties: C – white, yellowish, pinkish, greenish, iridescent (peristerite); S – white; L – vitreous; D – transparent to translucent; DE – 2.6; H – 6-6.5; CL – good; F – uneven to conchoidal; M – tabular crystals, platy aggregates, granular, massive.

Origin and occurrence: Magmatic in granites, syenites, andesites and pegmatites; metamorphic in gneisses and migmatites; hydrothermal in the Alpine-type veins; typical rock-forming mineral. Large crystals, several dm long, are known mainly from pegmatites. Peristerite comes from Arendal, Norway; Miass, Ural mountains, Russia; Quadeville, Ontario, Canada; Tvedestrand, Norway. It also occurs in Vežna , Czech Republic. Sunstone is mined in the Ponderosa mine near Lakeview, Oregon, USA.

Application: peristerite and sunstone are used as gemstones.

Andesine, 50 mm, Bodenmais, Germany

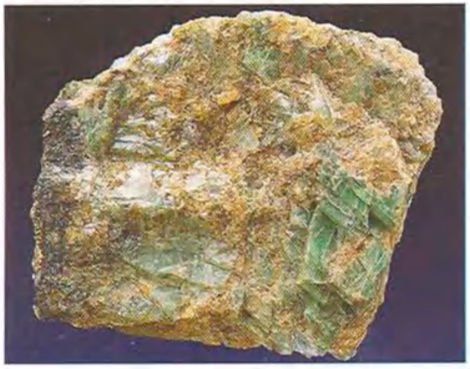

Andesine

FELDSPAR GROUP

$(Na_{0,7-0,5}Ca_{0,3-0,5}Al_{1,3-1,5}Si_{2,7-2,5}O_8$

TRICLINIC ● ● ● ●

Properties: C – yellowish, pinkish, gray, light green; S – white; L – vitreous; D – transparent to translucent; DE – 2.7; H – 6-6.5; CL – good; F – uneven to conchoidal; M – platy aggregates, granular, massive.

Labradorite, 50 mm, Labrador, Canada

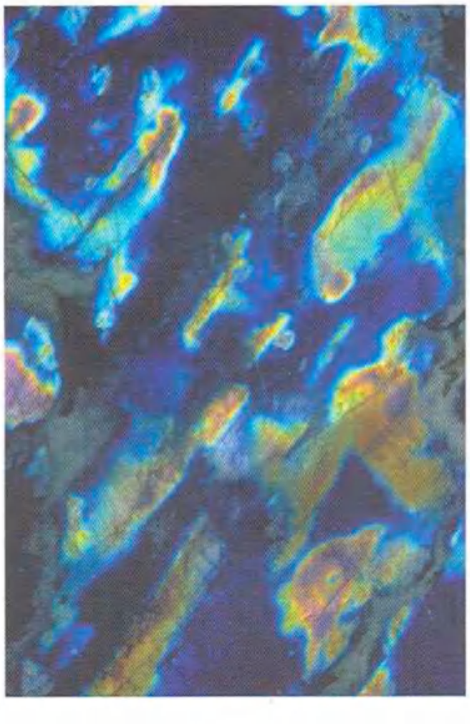

Anorthite, 10 mm xx, Pasmaada, Finnland

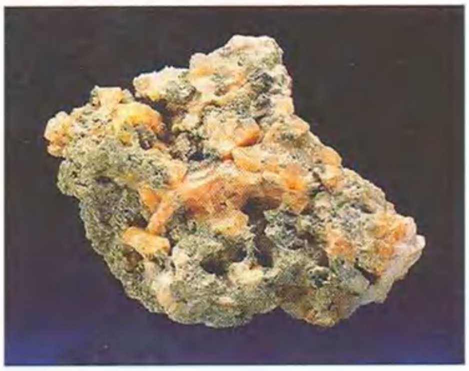

Origin and occurrence: C – white to dark gray, greenish; S – white; L – vitreous, pearly; D – transparent to translucent; DE – 2.7; H – 6-6.5; CL – good; F – uneven to conchoidal; M B cleavable aggregates, granular, massive.

Labradorite

FELDSPAR GROUP

$Ca_{0,7-0,5}Na_{0,3-0,5}Al_{1,5-1,7}Si_{2,5-2,3}O_8$

TRICLINIC ● ● ● ●

Properties: Magmatic in andesites, diorites; metamorphic in gneisses and migmatites; typical rock-forming mineral. Large grains are known from migmatites in Bodenmais, Germany and Adamello, Italy.
Origin and occurrence: Magmatic in gabbros, basalts, anorthosites; metamorphic in amphibolites. Large iridescent aggregates up to 1 m (39⅜ in) across, come from Nain, Labrador, Quebec, Canada and also occur in Korostenskiy massif. Ukraine and Ylamaa, Finland.
Application: as a decorative stone.

Anortite

FELDSPAR GROUP

$CaAl_2Si_2O_8$

TRICLINIC ● ● ●

Properties: C – gray, greenish, pinkish; S – white; L – vitreous, dull; D – transparent to translucent; DE – 2.8; H – 6-6.5; CL – good; F – uneven to conchoidal; M B granular, massive.
Origin and occurrence: Magmatic in gabbros, basalts, anorthosites; metamorphic in contact metamorphosed rocks. Pinkish grains and poorly-developed crystals come from Val di Fassa, Italy. Crystals up to 50 mm (2 in) across are known from Miyake-Jima Island, Japan. It was also found in Mount Erebus, Antarctica and Monte Somma, Italy.

Danburite

$CaB_2Si_2O_8$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, gray, greenish, pinkish, yellow, brown, red-brown; S – white; L – vitreous, dull; D – transparent to translucent; DE – 3.0; H – 7; CL – imperfect; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates, granular, massive.
Origin and occurrence: Hydrothermal in pegmatite cavities, ore veins and the Alpine-type veins; metamorphic in skarns and contact metamorphosed rocks.
Well-formed prismatic crystals up to 250 mm ($9^{13}/_{16}$ in) long come from Russell, New York, USA; Toroku, Japan; Dalnegorsk, Russia; and Charcas, San Luis Potosi, Mexico.

Cancrinite

$Na_6CaAl_6Si_6O_{24}(CO_3) . 2\ H_2O$

HEXAGONAL ● ● ●

Properties: C – colorless, white, yellow, orange, bluish; S – white; L – vitreous; D – transparent to translucent; DE – 2.4; H – 5-6; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.
Origin and occurrence: Magmatic in alkaline syenites; hydrothermal as a product of replacement of volcanic rocks, associated with nepheline and sodalite.
Crystals up to 20 mm ($^{25}/_{32}$ in) across come from Mont St.-Hilaire, Quebec, Canada. It was also found in Litchfield, Maine, USA; Cancrinite Hill, Bancroft, Ontario, Canada and elsewhere.

Cancrinite, 70 mm, Ditrau, Romania

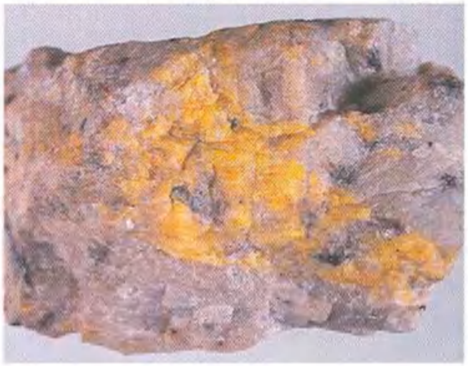

Danburite, 109 mm, Charcas, Mexico

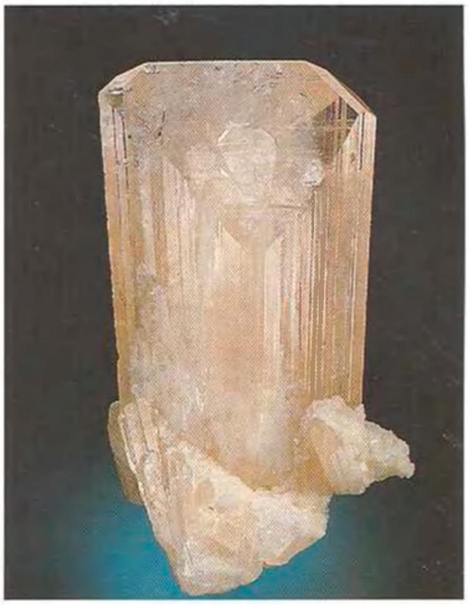

Leifite

$Na_2(Si,Al,Be)_7\ (O,OH,F)_{14}$

TRIGONAL ●

Properties: C – colorless; S – white; L – vitreous; D – translucent; DE – 2.5; H – 6; CL – good; F – uneven; M – acicular crystals.
Origin and occurrence: Hydrothermal in cavities of alkaline pegmatites, associated with microcline and zinnwaldite. Crystals up to 20 mm ($^{25}/_{32}$ in) across occur in Narssarssuk, Greenland. Fine crystals are also known from Mont St.-Hilaire, Quebec, Canada, associated with serandite.

Leifite, 100 mm, Mont St.-Hilaire, Canada

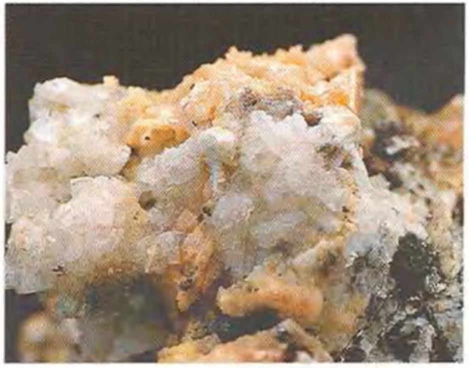

Sodalite, 50 mm, Bancroft, Canada

Sodalite

$Na_8Al_6Si_6O_{24}Cl_2$

CUBIC ● ● ●

Varieties: hackmanite

Properties: C – colorless, white, blue, yellow, pink (hackmanite); S – white; L – vitreous; D – transparent to translucent; DE – 2.3; H – 5.5-6; CL – imperfect; F – uneven to conchoidal; M – isometric crystals, granular, massive; LU – orange-red.
Origin and occurrence: Magmatic in alkaline syenites and phonolites, associated with nepheline, zircon and titanite; hydrothermal in marbles. Granular aggregates are known from the Princess Sodalite mine, Bancroft, Ontario, Canada and Ditrau, Romania. Crystals up to 100 mm (4 in) across come from Kangerdluarssuk, Greenland.
Application: cut as a gemstone.

Noseane

$Na_8Al_6Si_6O_{24}(SO_4)$

CUBIC ● ●

Properties: C – colorless, white, blue, gray, black; S – white; L – vitreous; D – transparent to translucent; DE – 2.3; H – 5-6; CL – good; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Magmatic in alkaline basalts and similar effusive rocks, associated with nepheline and haüyne. Granular aggregates are known from Laacher See, Germany and Monte Somma, Italy.

Lazurite

$(Na,Ca)_8Al_6Si_6O_{24}(S,SO_4)$

CUBIC ● ● ●

Properties: C – dark blue; S -light blue; L – vitreous; D – transparent to translucent; DE – 2.4; H – 5-5.5; CL – imperfect; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Metamorphic in contact metamorphosed marbles, associated with pyrite. Crystals up to 50 mm (2 in) across come from Sar-e-Sang, Badakhshan province, Afghanistan. Granular aggregates are known from many localities, such as Malobystrinskoye deposit near Lake Baikal, Russia or Monte Somma, Italy.
Application: decorative stone and gemstone.

Nosean, 6 mm grains, Laacher See, Germany

Lazurite, 60 mm, Afghanistan

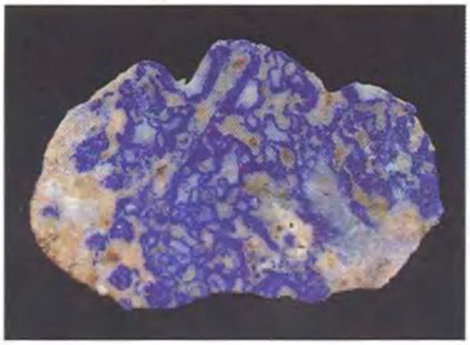

Lazurite, 21 mm x, Nuristan, Afghanistan

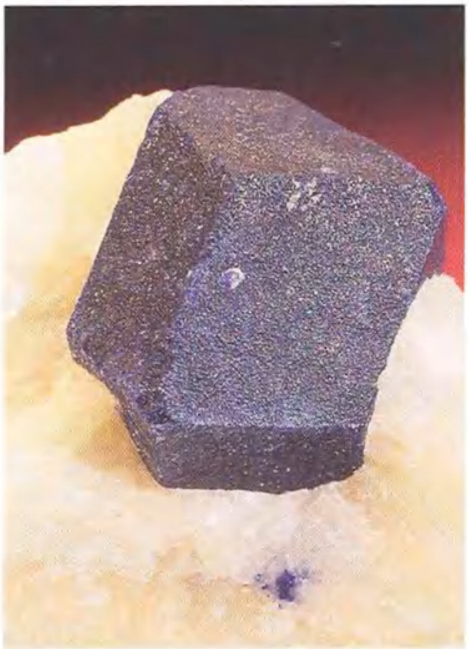

Tugtupite

$Na_8Be_2Al_2Si_8O_{24}Cl_2$

TETRAGONAL ● ●

Properties: C – white, pink, red, bluish; S – white; L – vitreous; D – transparent to translucent; DE – 2.3; H – 5; CL – good; F – uneven to conchoidal; M – small crystals, granular, massive.

Origin and occurrence: Hydrothermal in alkaline syenites and their pegmatites, also as a product of chkalovite alteration. Aggregates up to 60 mm (2$^{3}/_{8}$ in) across come from the Umbozero mine, Mount Alluaiv, Lovozero massif, Kola Peninsula, Russia and IlRmaussaq, Greenland. Small crystals occur in the Poudrette quarry, Quebec, Canada.

Tugtupite, 90 mm, Illimaussaq, Greenland

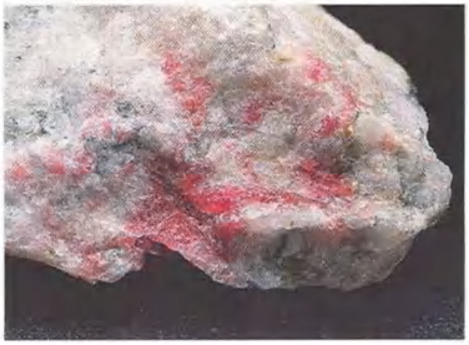

Lazurite, 60 mm, Chile

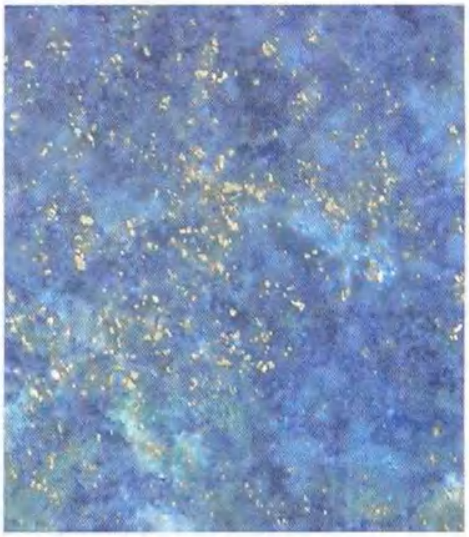

Danalite

$Fe_4Be_3Si_3O_{12}S$

CUBIC ● ●

Properties: C – gray,yellow,pink, red, brown; S – white; L – vitreous; D – transparent to translucent; DE – 3.4; H – 5.5-6; CL – imperfect; F – uneven to conchoidal; M – isometric crystals, granular, massive.

Origin and occurrence: Magmatic in granitic pegmatites; hydrothermal in greisens, skarns and ore veins. It occurs in Cape Ann, Massachusetts, USA; Hortekollen, Norway and Coolgardie, Western Australia, Australia.

Danalite, 70 mm, Yxjoberg, Sweden

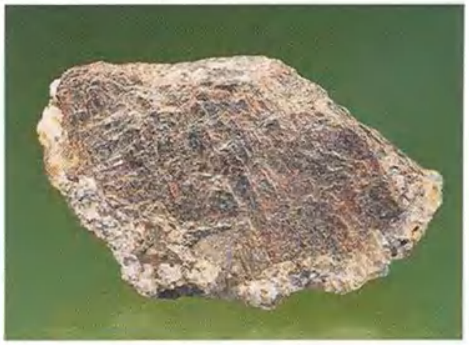

Helvite, 2 mm xx, Cavnic, Romania

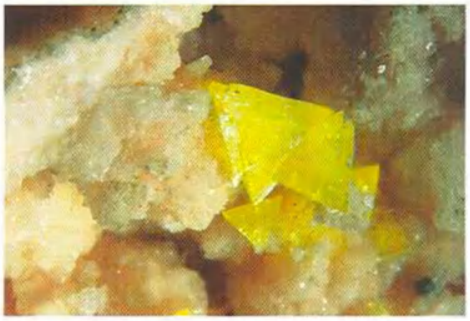

Helvite

$Mn_4Be_3Si_3O_{12}S$

CUBIC ● ● ●

Properties: C – brown, gray, yellow, yellow-green; S – white; L – vitreous; D – transparent to translucent; DE – 3.4; H – 6; CL – good; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Magmatic in granitic pegmatites and alkaline syenites; hydrothermal in greisens, skarns and ore deposits. Cubic crystals up to 25 mm (1 in) across known from the Sawtooth Batholith, Idaho, USA. Crystals also come from Schwarzenberg, Germany; Cavnic, Romania; and Oslofjord, Norway.
Application: Be ore.

Scapolite, 44 mm x, Leslie Lake, Quebec, Canada

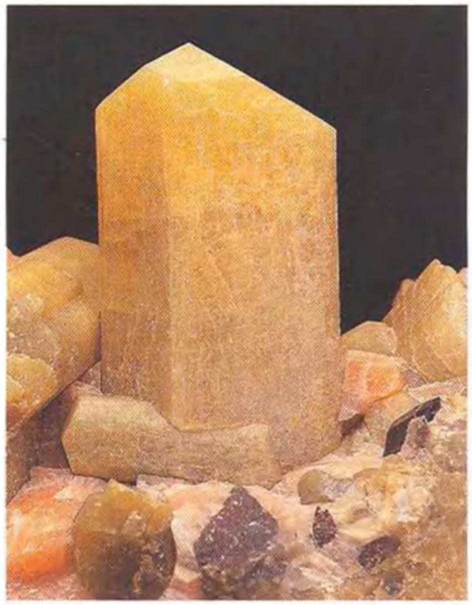

Marialite

SCAPOLITE GROUP

$Na_8(AlSi_3O_8)_6\ (Cl_2,SO_4)$

TETRAGONAL ● ● ●

Properties: : C – colorless, white, gray, purplish, yellow; S – white; L – vitreous, locally pearly; D – transparent to translucent; DE – 2.5; H – 5-6; CL – good; F – uneven to conchoidal; M – prismatic crystals, columnar aggregates, granular, massive; LU – yellow to orange.
Origin and occurrence: Hydrothermal in veins, cross-cutting alkaline metamorphic rocks and in pegmatites, cross-cutting ultrabasic rocks; metamorphic in marbles and metaevaporites. It comes from Ankazobe, Madagascar; Umba, Tanzania and Crestmore, California, USA.

Meionite

SCAPOLITE GROUP

$Ca_8(Al_2Si_2O_8)_6\ (CO_3,SO_4)$

TETRAGONAL ● ● ● ●

Properties: C – colorless, white, gray, purplish, green, blue; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 5-6; CL – good; F – uneven to conchoidal; M – long prismatic crystals, columnar aggregates, granular, massive; LU – yellow to orange.

Leucite, 30 mm, Mt. Vesuvius, Italy

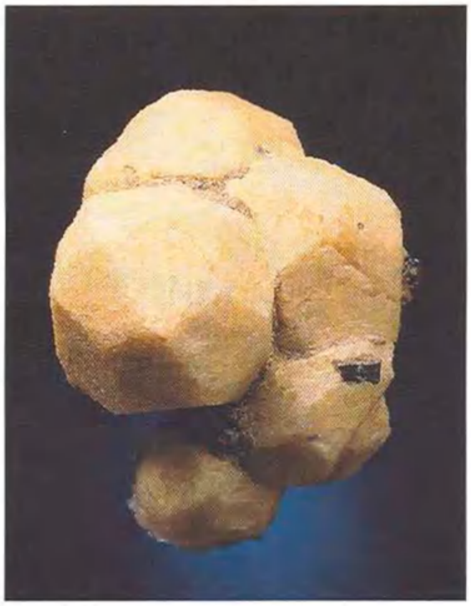

Analcime, 45 mm, Tunguzka, Russia

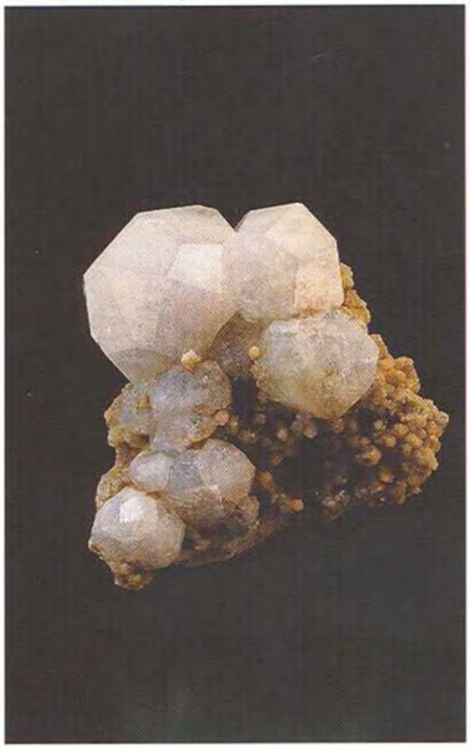

Origin and occurrence: Metamorphic in skarns, marbles, granulites and in contacts of volcanic rocks; hydrothermal in veins, cross-cutting Ca-rich rocks. Well formed crystals up to 40 cm (15¾ in) across come from Lake Clear and Eganville, Ontario, Canada. Crystals are also known from Monte Somma, Italy; Sludyanka, Siberia, Russia and Pargas, Finland.

Leucite

ZEOLITE GROUP

$KAlSi_2O_6$

TETRAGONAL ● ● ● ●

Properties: C – colorless, white, gray; S – white; L – vitreous; D – transparent to translucent; DE – 2.5; H – 5.5-6; CL- imperfect; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Magmatic in effusive K-rich basalts, associated with nepheline and sanidine. Well-formed crystals, several cm in size, come from Mount Vesuvius, Italy and Laacher See, Germany.
Application: ceramic industry.

Analcime

ZEOLITE GROUP

$NaAlSi_2O_6 \cdot H_2O$

CUBIC ● ● ● ●

Properties: C – colorless, white, pinkish, yellowish; S – white; L – vitreous; D – transparent to translucent; DE – 2.3; H – 5-5.5; CL – imperfect; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Mainly hydrothermal as a product of nepheline or sodalite replacement; rare magmatic in effusive rocks; also in sediments, associated with calcite and zeolites. Well-formed crystals up to 30 cm (12 in) across come from Nidym, Siberia, Russia; Lago Maggiore, Italy; Mont St.-Hilaire, Quebec, Canada; West Paterson, New Jersey, USA and elsewhere.

Pollucite

ZEOLITE GROUP

$(Cs,Na)AlSi_2O_6 \cdot H_2O$

CUBIC ● ● ●

Properties: C – colorless, white, gray; S – white; L – vitreous; D – transparent to translucent; DE – 2.9; H – 6.5-7; CL – none; F – uneven to conchoidal; M – isometric crystals, granular, massive.
Origin and occurrence: Magmatic and rarely also hydrothermal in Li-bearing pegmatites, associated with lepidolite, albite, quartz and petalite. Almost monomineral layer of pollucite several meters thick occurs in the Tanco mine, Bernic Lake, Manitoba, Canada. White and colorless crystals from pegmatite cavities up to 60 cm (24 in) across come from Paprok, Afghanistan. Also known from San Piero in Campo, Elba, Italy and Gilgit, Pakistan.
Application: ceramic industry, Cs ore.

Pollucite, 22 mm x, Gilgit, Pakistan

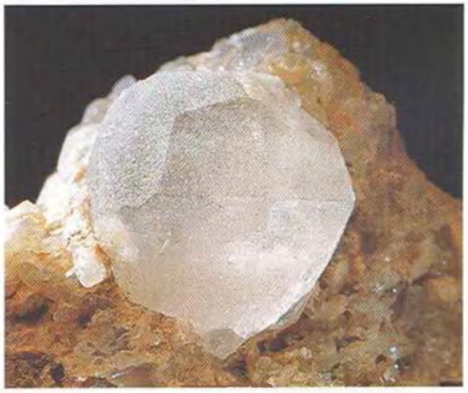

Natrolite, 47 mm, Maricopa Co., Arizona, U.S.A.

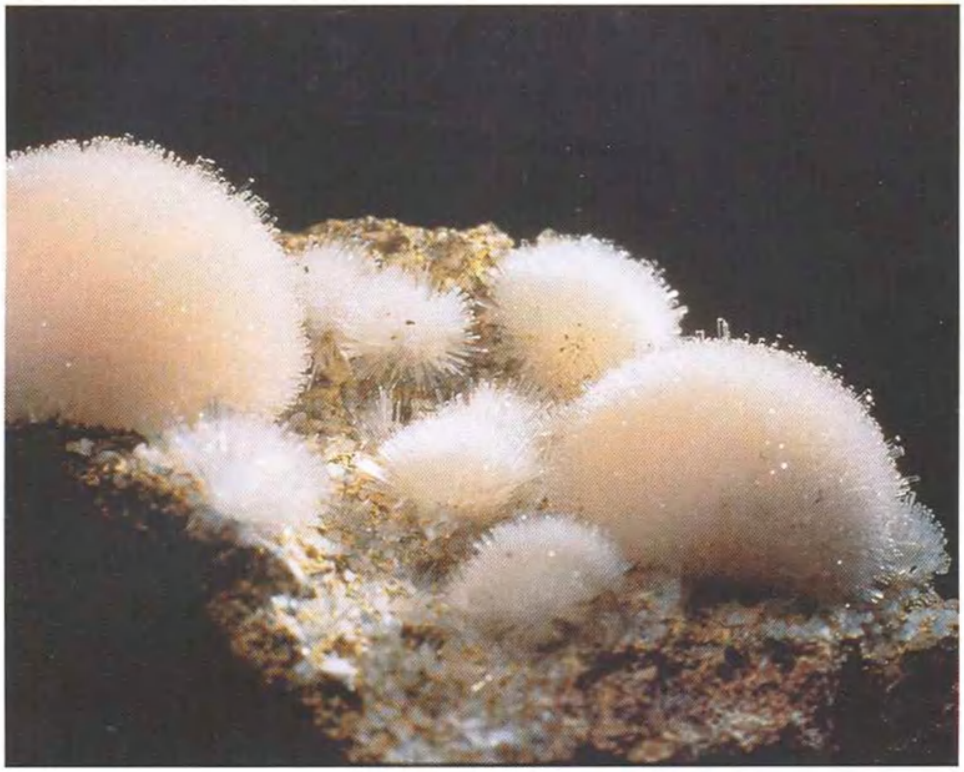

Scolecite, 165 mm, Nasik, India

Natrolite

ZEOLITE GROUP

$Na_2Al_2Si_3O_{10} \cdot 2 H_2O$

ORTHORHOMBIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish; S – white; L – vitreous to silky; D – transparent to translucent; DE – 2.2; H – 5-5.5; CL – perfect; F – uneven to conchoidal; M – long prismatic to acicular crystals, fibrous and radial aggregates, granular, massive.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, alkaline pegmatites, along the Alpine-type fissures, also as a product of plagioclase replacement, commonly associated with calcite and zeolites. Colorless and white acicular crystals up to 30 cm (12 in) long come from cavities of alkaline pegmatites in Mount Putelichorr, Khibiny massif, Kola Peninsula, Russia. Crystals were also found in Narssarssuk, Greensland and Mont St.-Hilaire, Quebec, Canada. Crystals also occur in cavities of basaltic rocks in Teigarhorn, Iceland; Zalezly and Soutisky, Czech Republic and in Faeroe Islands.

Skolecite

ZEOLITE GROUP

$CaAl_2Si_3O_{10} \cdot 3\ H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white; S – white; L – vitreous to silky; D – transparent to translucent; DE – 2.3; H – 5; CL – good; F – uneven; M – long prismatic to acicular crystals, fibrous and radial aggregates.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks and along the Alpine-type fissures, associated with calcite and zeolites. Clear prismatic crystals up to 200 mm ($7\frac{7}{8}$ in) long come from basalt cavities near Nasik, India. Crystals come also from Teigarhorn, Iceland and Suderoy, Faeroe Islands.

Mesolite

ZEOLITE GROUP

$Na_{16}Ca_{16}Al_{48}Si_{72}O_{240} \cdot 64\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white; S – white; L – vitreous to silky; D – transparent to translucent; DE – 2.3; H – 5; CL – good; F – uneven; M - acicular crystals, fibrous and radial aggregates, granular, massive.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, associated with other zeolites. Colorless needles up to 150 mm (6 in) long come from the basalt cavities in Berufjord and Teigarhorn, Iceland. Radial aggregates up to 200 mm ($7\frac{7}{8}$ in) in diameter known from the vicinity of Poona, India. Prismatic crystals up to 100 mm (4 in) long found in Skookumchuck Dam, Washington, USA.

Mesolite, 150 mm, Mahrashitra, India

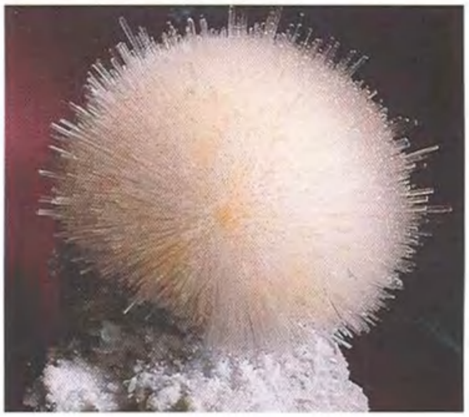

Thomsonite

ZEOLITE GROUP

$Ca_2NaAl_4(Al,Si)_2Si_4O_{20} \cdot 6\ H_2O$

ORTHORHOMBIC ● ● ●

Properties: C – colorless, white, yellowish, brown-red; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.3; H – 5-5.5; CL – good; F – uneven to conchoidal; M – prismatic crystals in clusters with radial structure, botryoidal aggregates, massive.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, also as a product of hydrothermal feldspar replacement. Colorless and white acicular crystals, forming radial aggregates up to 50 mm (2 in) in diameter come from Old Kilpatrick, Scotland, UK and West Paterson, New Jersey, USA. Hemispherical aggregates up to 30 mm ($1\frac{3}{16}$ in) across known from Vinaicka hora near Kladno, Czech Republic.

Thomsonite, 50 mm, Bazsi, Hungary

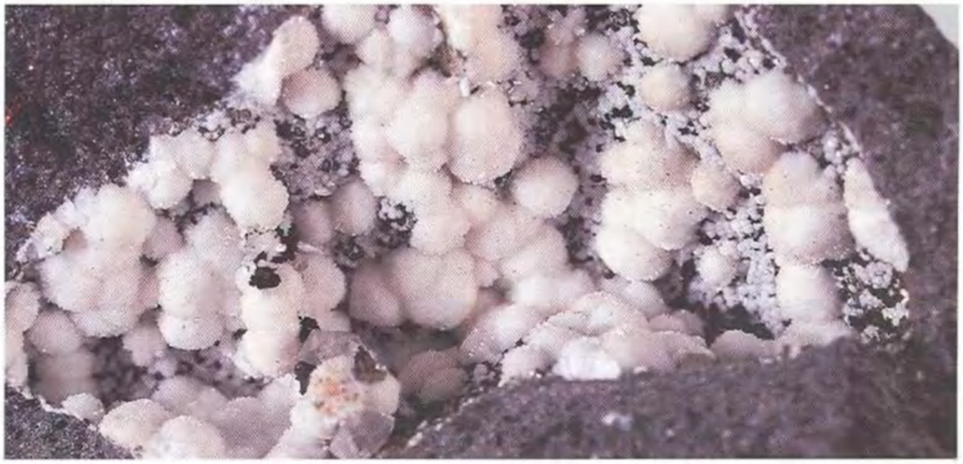

Gonnardite, 10 mm xx, Zalahalap, Hungary

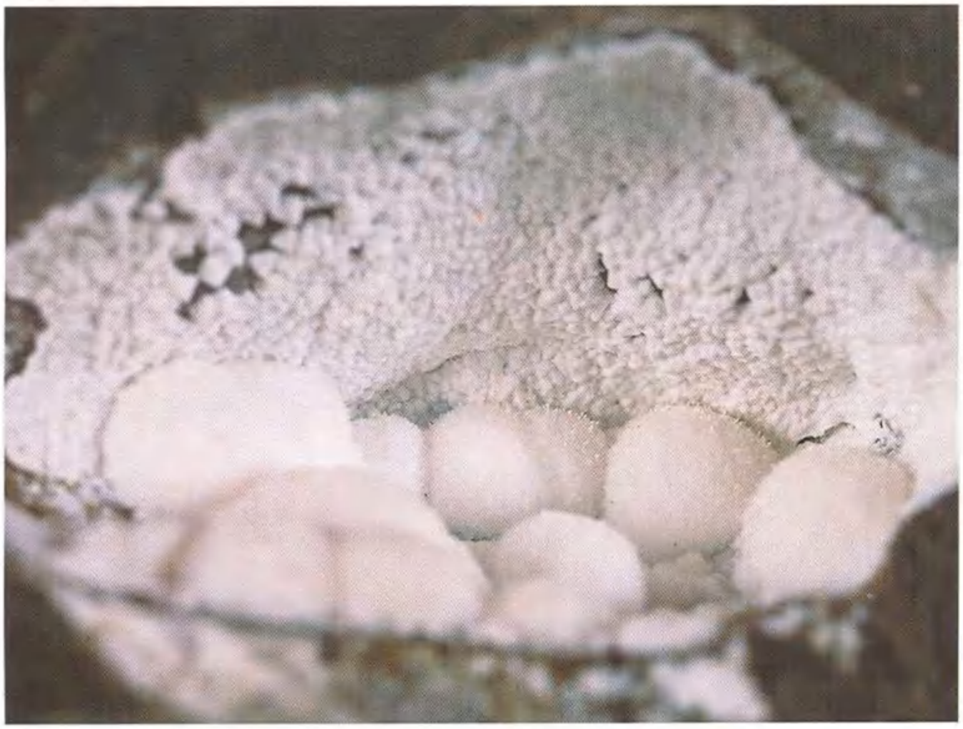

Gonnardite

ZEOLITE GROUP

$(Na,Ca)_{6-8}(Al,Si)_{20}O_{40} \cdot 12\ H_2O$

TETRAGONAL ● ● ●

Properties: C – colorless, white, yellowish; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.3; H – 4.5-5; CL – good; F – uneven to conchoidal; M – prismatic crystals, fibrous and radial aggregates.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, associated with zeolites, also in the contact zone of marbles, associated with wollastonite. Acicular crystals and fibrous aggregates occur in Aci Castello, Sicily, Italy; Weilberg, Germany; Bundoora, Victoria, Australia.

Edingtonite

ZEOLITE GROUP

$BaAl_2Si_3O_{10} \cdot 3\ H_2O$

TETRAGONAL ● ● ●

Properties: C – white, gray, pinkish; S – white; L – vitreous; D – transparent to translucent; DE – 2.8; H – 4; CL – good; F – uneven to conchoidal; M – prismatic crystals, granular, massive.

Origin and occurrence: Hydrothermal in cavities and along the cracks in volcanic rocks, associated with calcite and harmotome. Colorless and white prismatic crystals up to 40 mm ($1\frac{9}{16}$ in) across come from Old Kilpatrick, Scotland, UK; Stare Ransko,

Edingtonite, 100 mm, Stare Ransko, Czech Republic

Czech Republic and Mont St.-Hilaire. Quebec, Canada.

Dachiardite-Ca

ZEOLITE GROUP

$(Ca_{0.5},Na,K)_5Al_5Si_{19}O_{48} \cdot 13\ H_2O$

MONOCLINIC ● ●

Properties: C – colorless, white, yellowish; S – white; L – vitreous to pearly; D – transparent to translucent; DE – 2.1; H – 4-4.5; CL – good; F – uneven to conchoidal; M – prismatic crystals and complex interpenetration twins.
Origin and occurrence: Hydrothermal in cavities in granitic pegmatites, associated with albite, petalite and elbaite, also as pseudo-morphs after petalite. It occurs in San Piero in Campo, Elba, Italy and in the Opal Hill quarry, Riverside, California, USA.

Ferrierite-Ca

ZEOLITE GROUP

$(Ca,Na,K,Mg)_6Al_6Si_{30}O_{72} \cdot 18\ H_2O$

ORTHORHOMBIC ● ●

Properties: C – colorless, white, greenish, pink, brownish; S – white; L – vitreous; D – transparent to translucent; DE – 2.1; H – 3-3.5; CL – good; F – uneven to conchoidal; M – tabular crystals, platy and columnar aggregates, massive.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks, along the cracks of the Alpine-type veins and in volcanic tuffs, associated with calcite and zeolites. Platy aggregates occur in Svojanov, Czech Republic; Albero Bosso and Monastir, Sardinia, Italy; Kamploops Lake, British Columbia, Canada.

Ferrierite-Ca, 2 mm aggregates, Regec, Hungary

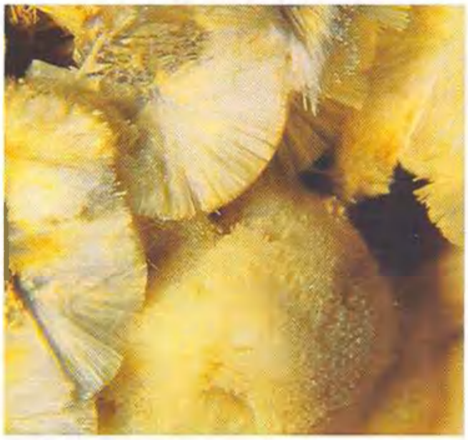

Dachiardite-Ca, 2 mm xx, Regec, Hungary

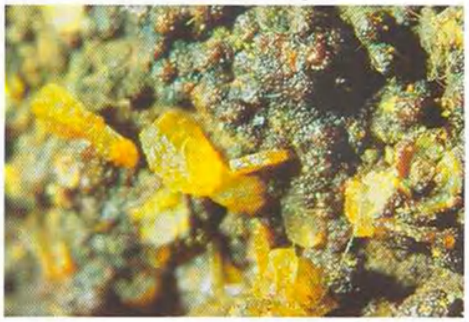

Laumontite

ZEOLITE GROUP

$Ca_4Al_8Si_{16}O_{48} \cdot 18\ H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish; S – white; L – vitreous; D – transparent to translucent, weathered almost opaque; DE – 2.4; H – 3-4; CL – good; F – uneven to conchoidal; M – prismatic crystals, fibrous and radial aggregates.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks, along the cracks in the Alpine-type veins, in ore veins and sediments, associated with calcite and other zeolites. Colorless and white acicular crystals up to 38 cm ($14^{15}/_{16}$ in) long come from the Pandulena Hill Quarries, India. Crystals up to 150 mm (6 in) long known from Pine Creek, Bishop, California, USA. It was also found in Dillenburg, Germany.

Laumontite, 120 mm, Markovice, Czech Republic

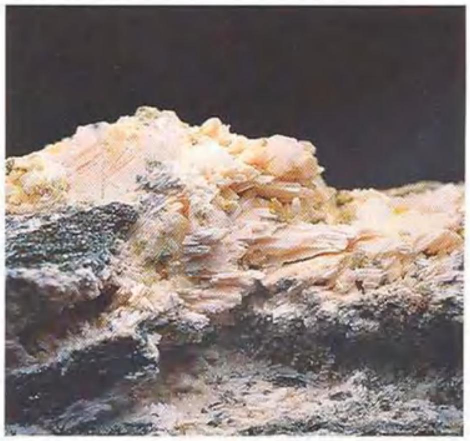

Heulandite-Ca, 8 mm xx, Parana, Brazil

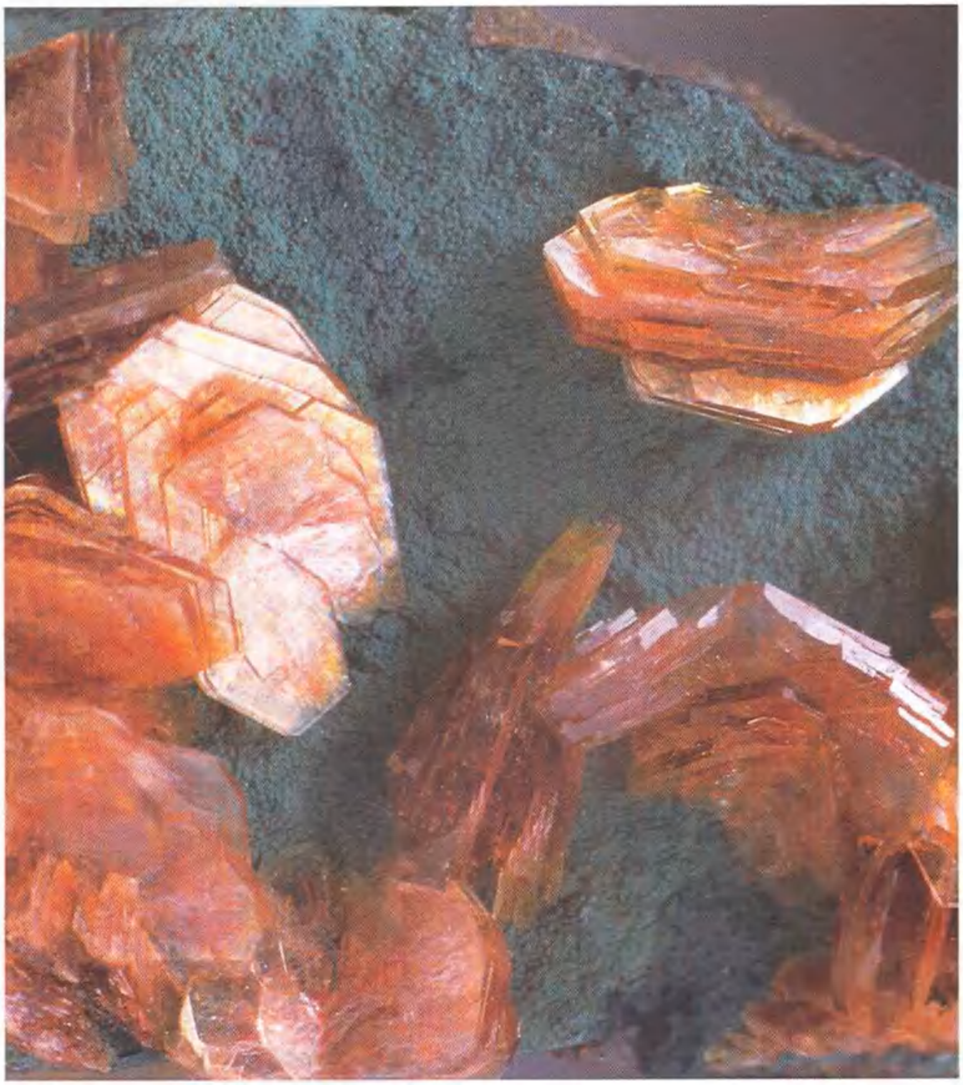

Heulandite-Ca

ZEOLITE GROUP

$(Ca_{0.5},Sr_{0.5},Na,K)_{4.5}Al_9Si_{27}O_{72} \cdot 24\ H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish, red, brown; S – white; L B vitreous to pearly; D – transparent to translucent; DE – 2.1; H – 3.5-4; CL – perfect; F – uneven to conchoidal; M – prismatic to tabular crystals, granular, massive.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks, in the Alpine-type fissures, ore veins and sedimentary rocks, usually associated with calcite and other zeolites. Colorless and white tabular crystals up to 100 mm (4 in) across come from Nasik, India; West Paterson, New Jersey, USA; Teigarhorn, Iceland. Red heulandite found in Val di Fassa, Italy.

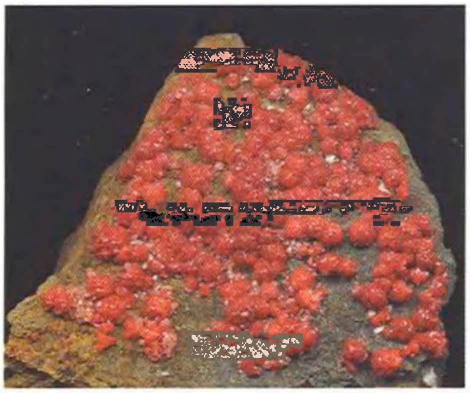

Clinoptilolite-Ca, 2 mm xx, Hon?ova Hurka, Czech Republic

Clinoptilolite-Ca

ZEOLITE GROUP

$(Ca_{0.5},Na,K)_{4.5}Al_9Si_{27}O_{72} \cdot 24\ H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish, red, greenish; S – white; L – vitreous; D – transparent to translucent; DE – 2.2; H – 3.5-4; CL – perfect; F – uneven to conchoidal; M – platy crystals, massive.

Origin and occurrence: Hydrothermal in volcanic-sedimentary rocks. Large industrial deposits occur in New Zealand, Japan and Australia. Crystals come from Agate Beach, Oregon, USA.

Application: construction, chemical industry and agriculture.

Stilbite-Ca

ZEOLITE GROUP

$(Ca_{0.5},Na,K)_9Al_9Si_{27}O_{72} \cdot 28\ H_2O$

MONOCLINIC ● ● ● ●

Properties: C – colorless, white, yellowish, pinkish, brown; S – white; L – vitreous, locally pearly; D – transparent to translucent; DE – 2.2; H – 3.5-4; CL – good; F – uneven to conchoidal; M – prismatic and tabular crystals, commonly complicated interpenetration twins and sheaf-like aggregates, granular, massive.

Origin and occurrence: Hydrothermal in cavities of volcanic rocks, along Alpine-type fissures, in ore veins, also in sedimentary rocks and hot springs, typically associated with calcite and other zeolites. Colorless and white tabular crystals and their combinations up to 200 mm ($7^{7}/_{8}$ in) across come from the Pandulena Hill Quarries, India. Crystals are also known from Teigarhorn, Iceland; Faeroe Islands; West Paterson, New Jersey, USA and elsewhere.

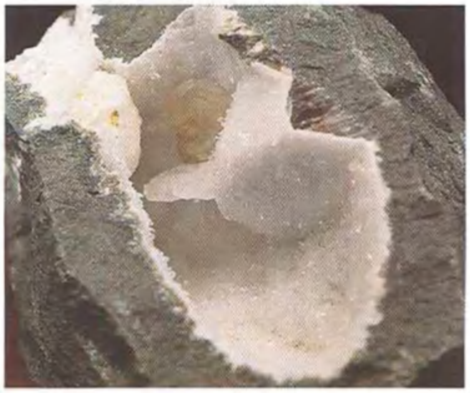

Epistilbite, 15 mm x, Nasik, India

Epistilbite

ZEOLITE GROUP

$(Ca,Na_2)Al_2Si_4O_{12} \cdot 4\ H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, yellowish, pinkish, light brown; S – white; L – vitreous; D – transparent to translucent; DE – 2.2; H – 4; CL – good; F – uneven to conchoidal; M – prismatic crystals, radial aggregates.

Origin and occurrence: Hydrothermal in cavities in volcanic rocks; associated with other zeolites. Clear and white tabular crystals and their twins up to 30 mm ($1^{3}/_{16}$ in) across come from Jalgaon, India; also in Teigarhorn, Iceland and Faeroe Islands.

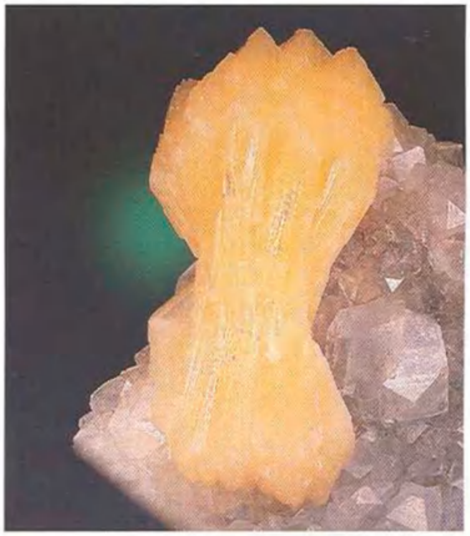

Stilbite-Ca, 83 mm, Nasik, India

Gismondine, 2 mm xx, Tapolca-Diszel, Hungary

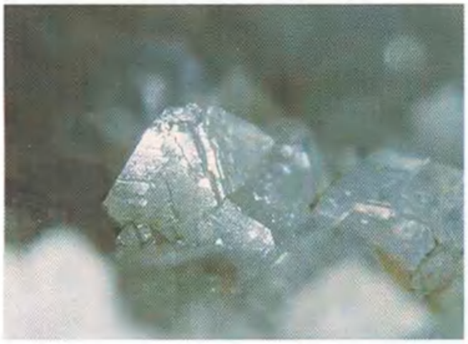

Gismondine

ZEOLITE GROUP

$CaAl_2Si_2O_8 \cdot 4,5\ H_2O$

MONOCLINIC ● ●

Properties: C – colorless, white, bluish, pinkish; S – white; L – vitreous; D – transparent to translucent; DE – 2.3; H – 4.5; CL – good; F – uneven; M – complex twins of crystals, platy aggregates.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks, in hydrothermally altered rocks. It comes from Capo di Bove, Italy; Schiffenberg, Germany and Dobrna, Czech Republic.

Harmotome, 79 mm, Strontian, UK

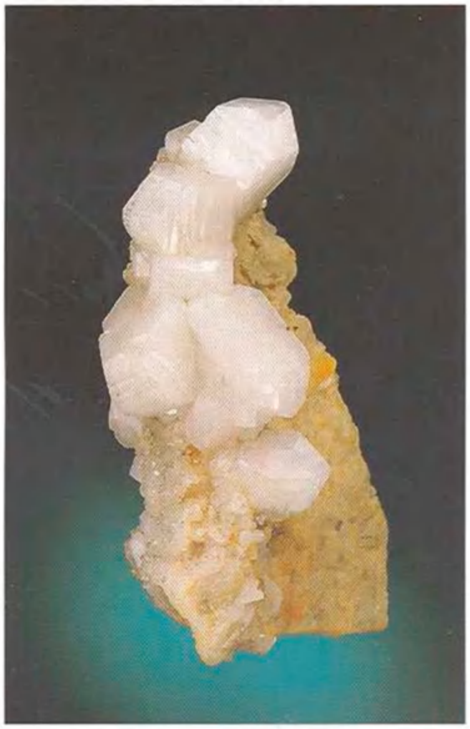

Phillipsite-K, 3 mm xx, Limberg, Germany

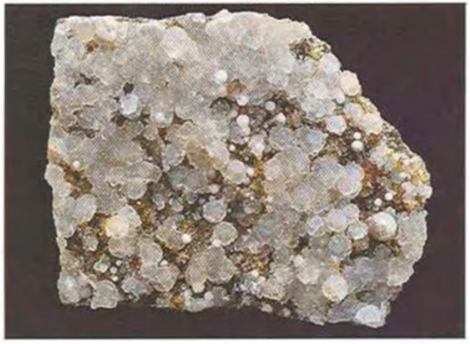

Phillipsite-K

ZEOLITE GROUP

$(K,Na,Ca_{0.5})Al_5Si_{11}O_{32} \cdot 12\ H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, reddish; S – white; L – vitreous; D – transparent to translucent; DE – 2.2; H – 4-4.5; CL – good; F – uneven; M – prismatic crystals and complex interpenetration twins, granular, massive.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks, sedimentary rocks and hot springs; associated with calcite and other zeolites. Colorless and white interpenetration twins up to 20 mm ($^{25}/_{32}$ in) across come from Capo di Bove, Italy; Doughboys, Tasmania, Australia; Sovinec near Litomiøce, Czech Republic and elsewhere.

Harmotome

ZEOLITE GROUP

$(Ba_{0.5},Ca_{0.5},K,Na)_5Al_5Si_{11}O_{32} \cdot 12\ H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, gray, reddish, yellow, brown; S – white; L – vitreous; D – transparent to translucent; DE – 2.4; H – 4.5; CL – good; F – uneven; M – prismatic crystals, commonly complex interpenetration twins, granular massive.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks, in the Alpine-type veins, ore veins and in pegmatites. Prismatic crystals and their twins up to 20 mm ($^{25}/_{32}$ in) across are known from Strontian, Scotland, UK. Crystals also come from St. Andreasberg and Oberstein, Germany; Kozakov, Czech Republic and Kongsberg, Norway.

Goosecreekite

ZEOLITE GROUP

$CaAl_2Si_3O_{10} \cdot 5\ H_2O$

MONOCLINIC ●

Properties: C – white, colorless; S – white; L – vitreous; D – transparent to translucent; DE – 2.2; H – 4.5; CL B perfect; F – uneven to conchoidal; M – prismatic crystals, granular.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks. It is known from Goose Creek, Virginia, USA. Crystals up to 30 mm ($1\frac{3}{16}$ in) across found in the Pandulena Hill quarries, India.

Goosecreekite, 25 mm xx, Maharashtra, India

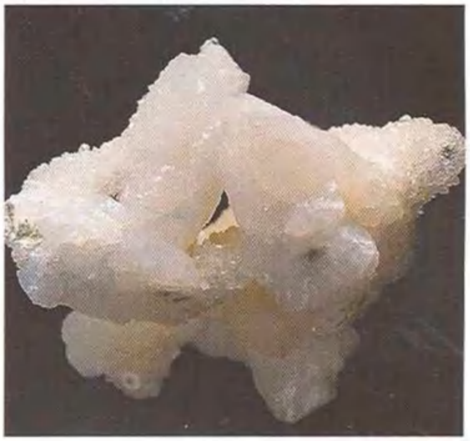

Chabasite-Ca

ZEOLITE GROUP

$(Ca_{0.5},Na,K)Al_4Si_8O_{24} \cdot 12\ H_2O$

TRIGONAL ● ● ●

Properties: C – colorless, white, yellowish, pinkish, greenish; S – white; L – vitreous to dull; D – transparent to translucent; DE – 2.2; H – 4-4.5; CL – imperfect; F – uneven to conchoidal; M – rhombohedral crystals and their twins, granular, massive.
Origin and occurrence: Hydrothermal in cavities in volcanic rocks and pegmatites, along Alpine-type fissures and in hot springs, usually associated with calcite and other zeolites. Colorless and white interpenetration twins up to 60 mm (24 in) across, come from Faeroe Islands. Crystals are also known from Repcice, Czech Republic; Berufjord, Iceland; Panvil, India; Maglovec, Slovakia and elsewhere.

Chabazite-Ca, Passboro, Nova Scotia, Canada

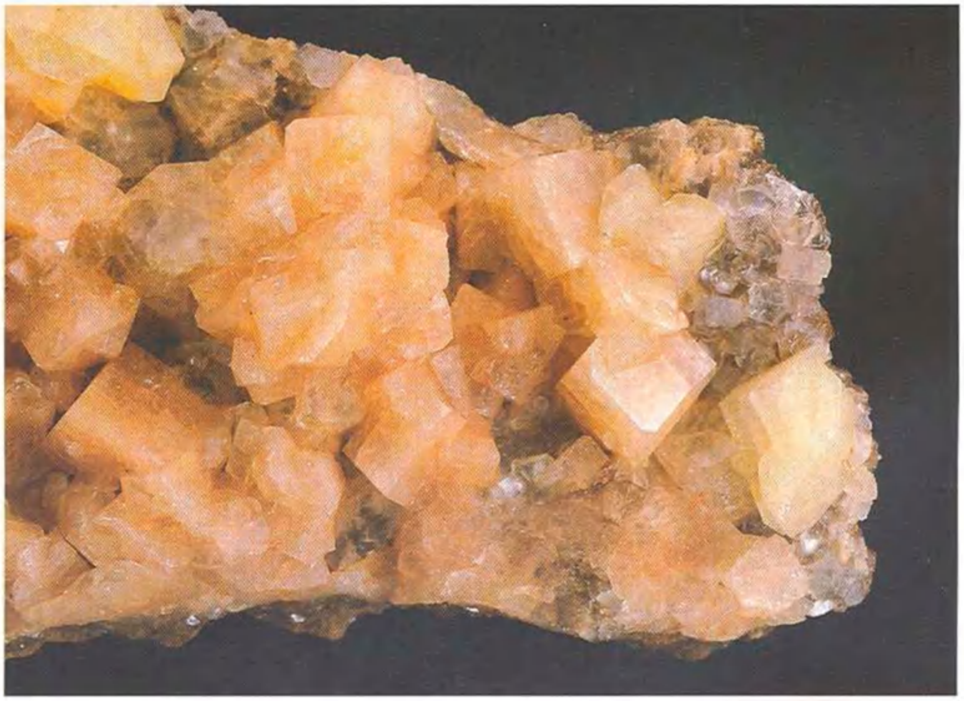

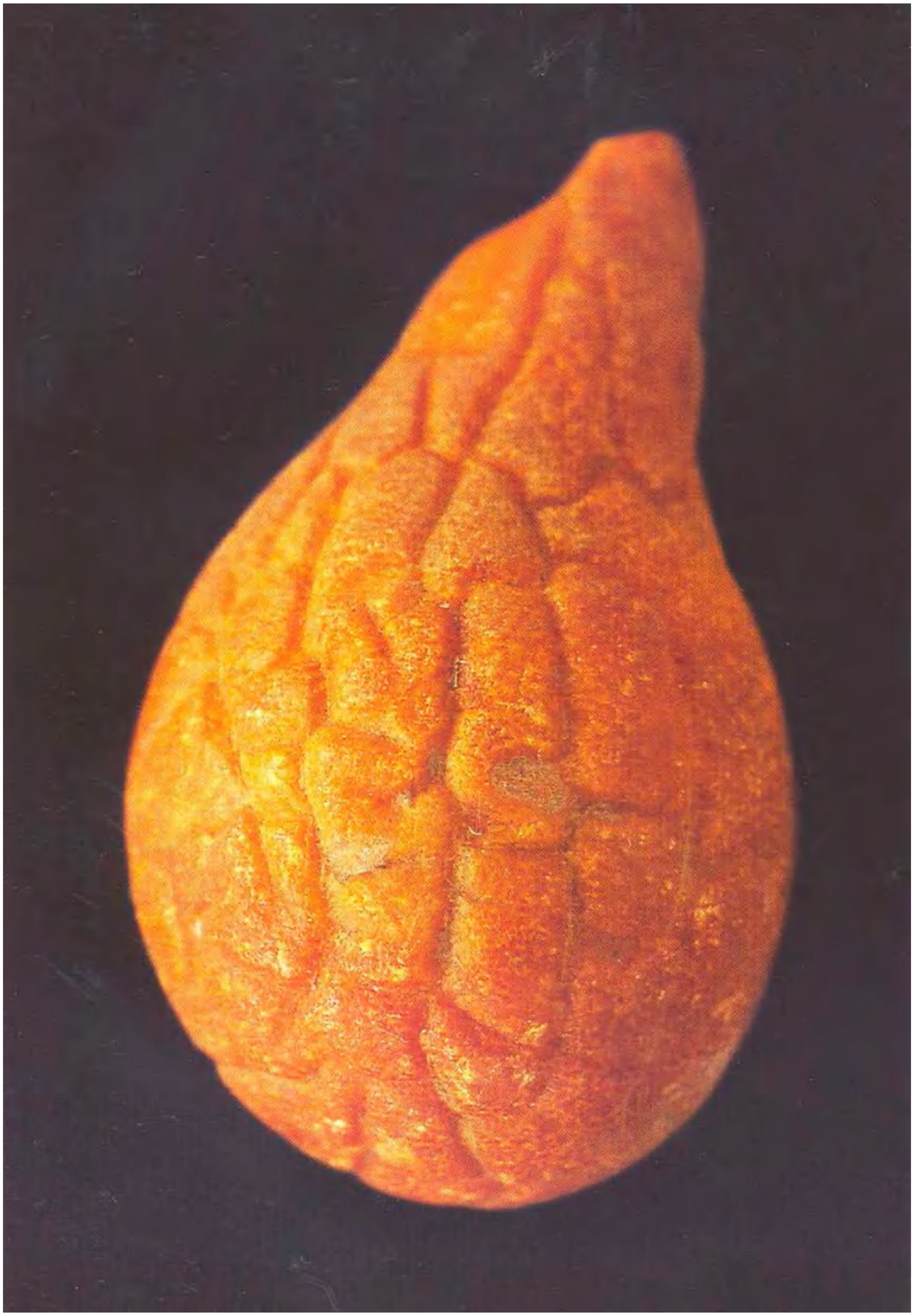

10. Organic compounds

Whewellite

$CaC_2O_4 \cdot H_2O$

MONOCLINIC ● ● ●

Properties: C – colorless, white, grayish; S – colorless; L – vitreous; D – transparent to translucent; DE – 2,2; H – 2,5; CL – good; F – conchoidal; M – prismatic crystals, commonly twinned.

Origin and occurrence: Rare hydrothermal, mainly sedimentary in coal basins, associated with barite, ankerite and other minerals. Hydrothermal crystals up to 70 mm (2¾ in) long, come from U-bearing veins in PrRbram, Czech Republic. Similar crystals, up to 70 mm (2¾ in) long, occurred in Cavnic, Romania. Heart-shaped and butterfly twins, up to 100 mm (4 in) found in concretions near Kladno, Czech Republic. Similar specimens are known from Burgk near Dresden, Germany. Interesting flat radial aggregates found along cracks in clays in the vicinity of Most, Czech Republic.

Amber, 40 mm, Baltic Sea, Latvia
Whewellite, 27 mm x, Burgk, Germany

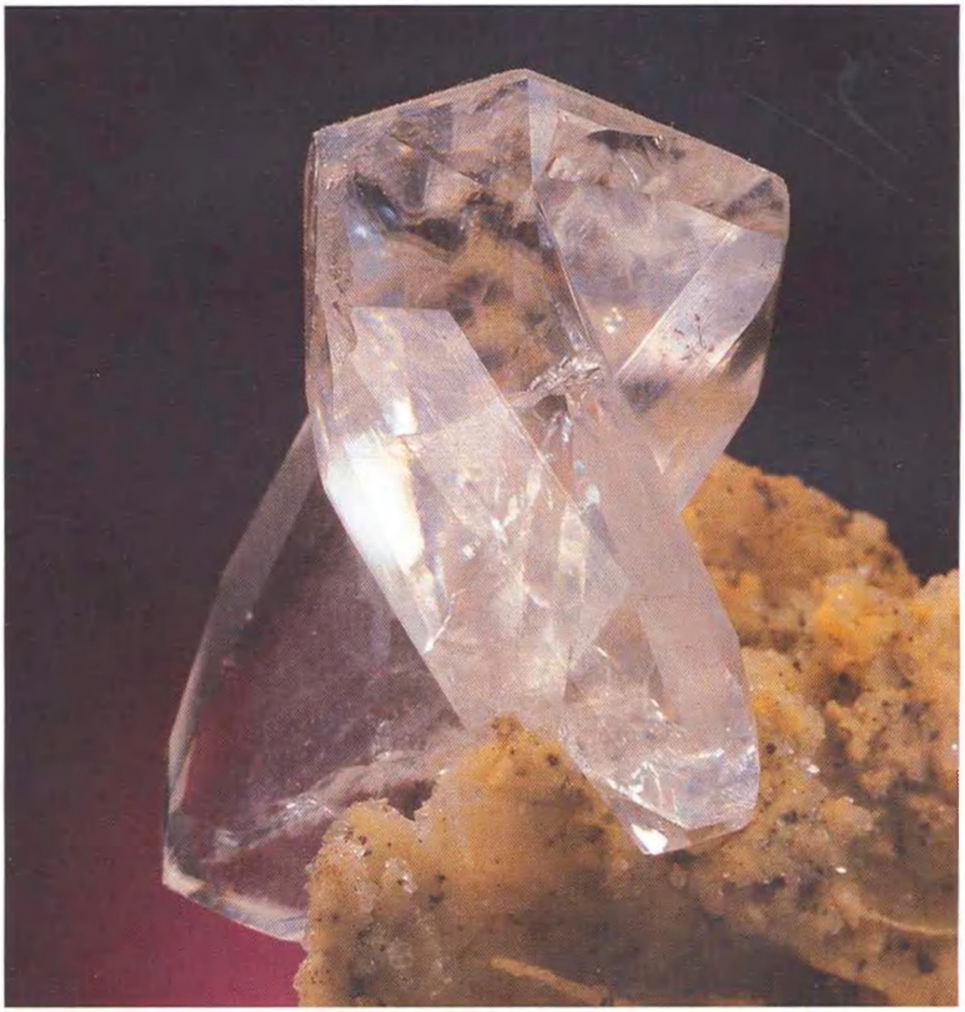

Mellite, 170 mm, Bicske-Csordakút, Hungary

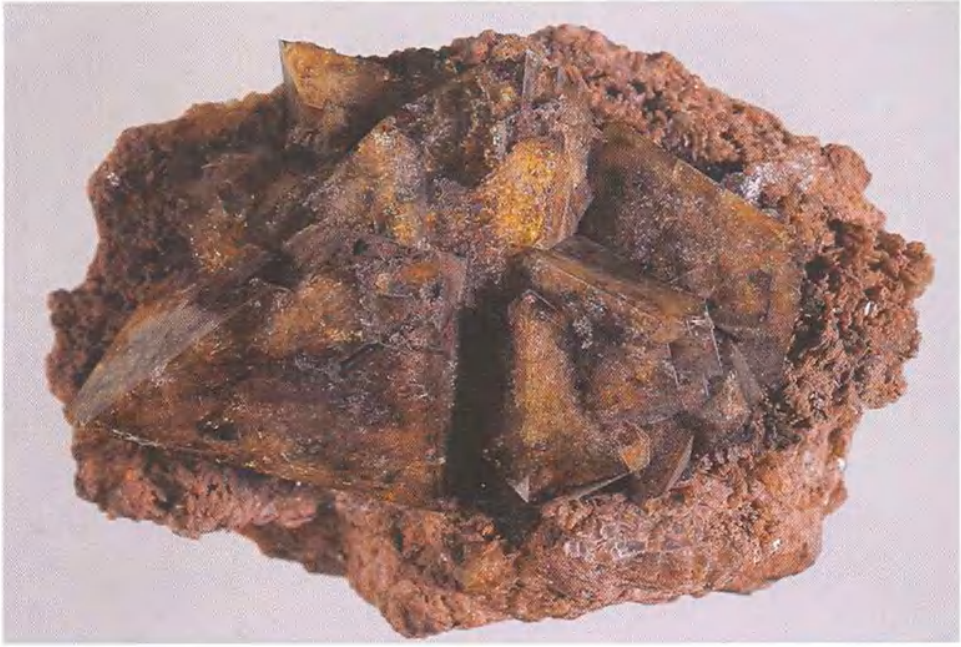

Mellite

$Al_2[C_6(COO)_6] . 18 H_2O$

TETRAGONAL ● ●

Properties: C – honey-yellow; S – white; L – resinous to vitreous; D – transparent to translucent; DE – 1,7; H – 2; CL – imperfect; F – conchoidal; M – dipyramidal crystals, granular; LU – blue.
Origin and occurrence: Secondary in the cracks in brown coal and lignite. The best specimens with crystals up to 40 mm ($1^{9}/_{16}$ in) in size, come from Csordakut near Tatabanya, Hungary. Crystals, up to 10 mm ($^{3}/_{8}$ in) across, were found in Artern, Germany. Granular aggregates are known from Valchov, Czech Republic.

Evenkite

$C_{24}H_{50}$

MONOCLINIC ●

Properties: C – colorless to light yellow; S – colorless; L – waxy; D – translucent; DE – 0,9; H – 1; CL – good; M – pseudo-hexagonal tabular crystals; R – melts at 50EC (122°F).
Origin and occurrence: Probably secondary. It occurs within geodes near Evenki, Siberia, Russia. Also known from the cracks in altered andesite in Dubnik, Slovakia.

Fichtelite

$C_{19}H_{34}$

MONOCLINIC ● ● ●

Properties: C – colorless to yellowish; S – colorless; L – vitreous; PS – transparent to translucent; DE – 1; H – 1; M – thin tabular crystals, scales, crystalline crusts.

Evenkite, 50 mm, Dubnik, Slovakia

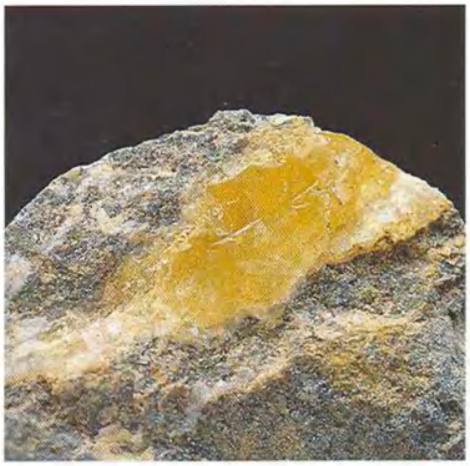

Origin and occurrence: Secondary, typical mineral of peat-bogs. Its crystals are known from Borkovice near Sobeslav, Czech Republic; also known from Marktredwitz, Germany.

Amber

a mixture of hydrocarbons

AMORPHOUS ● ● ●

Properties: C – honey-yellow, yellow-brown, brown, red-brown, blue, green, black; S – white; L – resinous, dull; D – transparent to translucent, rare opaque; DE – 1,0-1,1; H – 2-2,5; CL – none; F – conchoidal; M – massive irregular or drop-shaped aggregates, nodules and fragments; LU – light blue, yellow.

Origin and occurrence: Amber is a petrified resin from Tertiary and Mesozoic conifers, occurring rarely in sediments. Plant or insect remnants are sometimes found trapped in amber. The most famous localities are located along the southern coast of the Baltic Sea in Poland, Germany, Lithuania, Estonia, Latvia and Russia. The largest masses found weighed up to 10 kg (22 lb). Blue amber is known from the Dominican Republic. Other amber localities are found in Syria, Lebanon, Thailand, Vietnam, Canada and the USA.

Application: as a gemstone.

Fichtelite, 80 mm, Mazice, Czech Republic

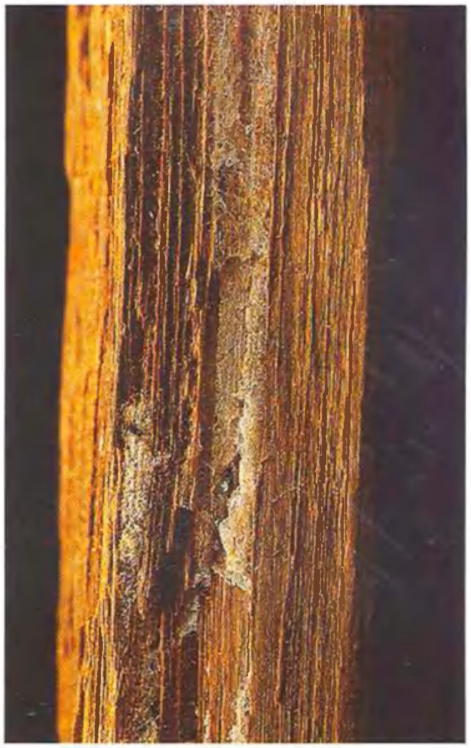

Amber, 95 mm, Madagascar

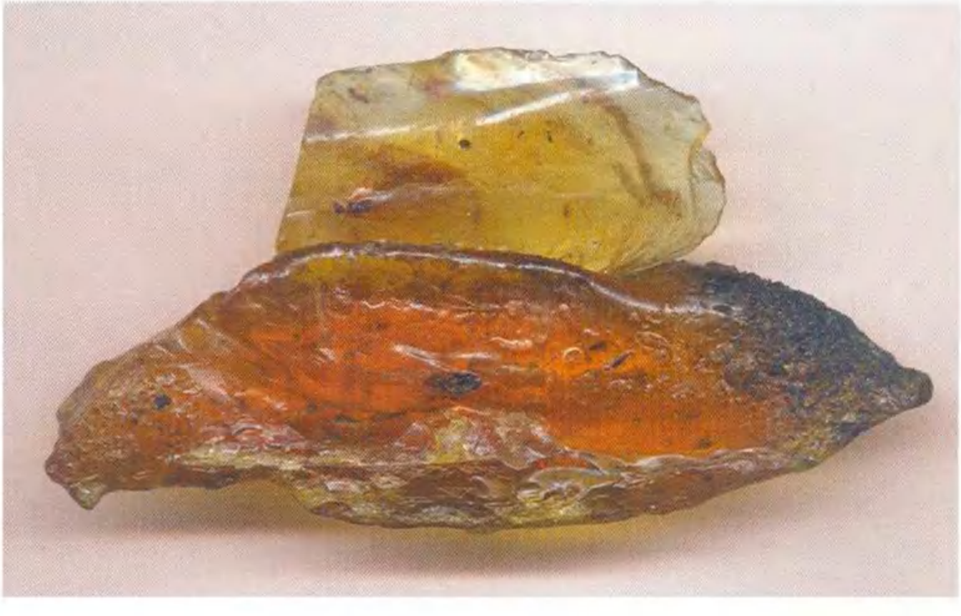

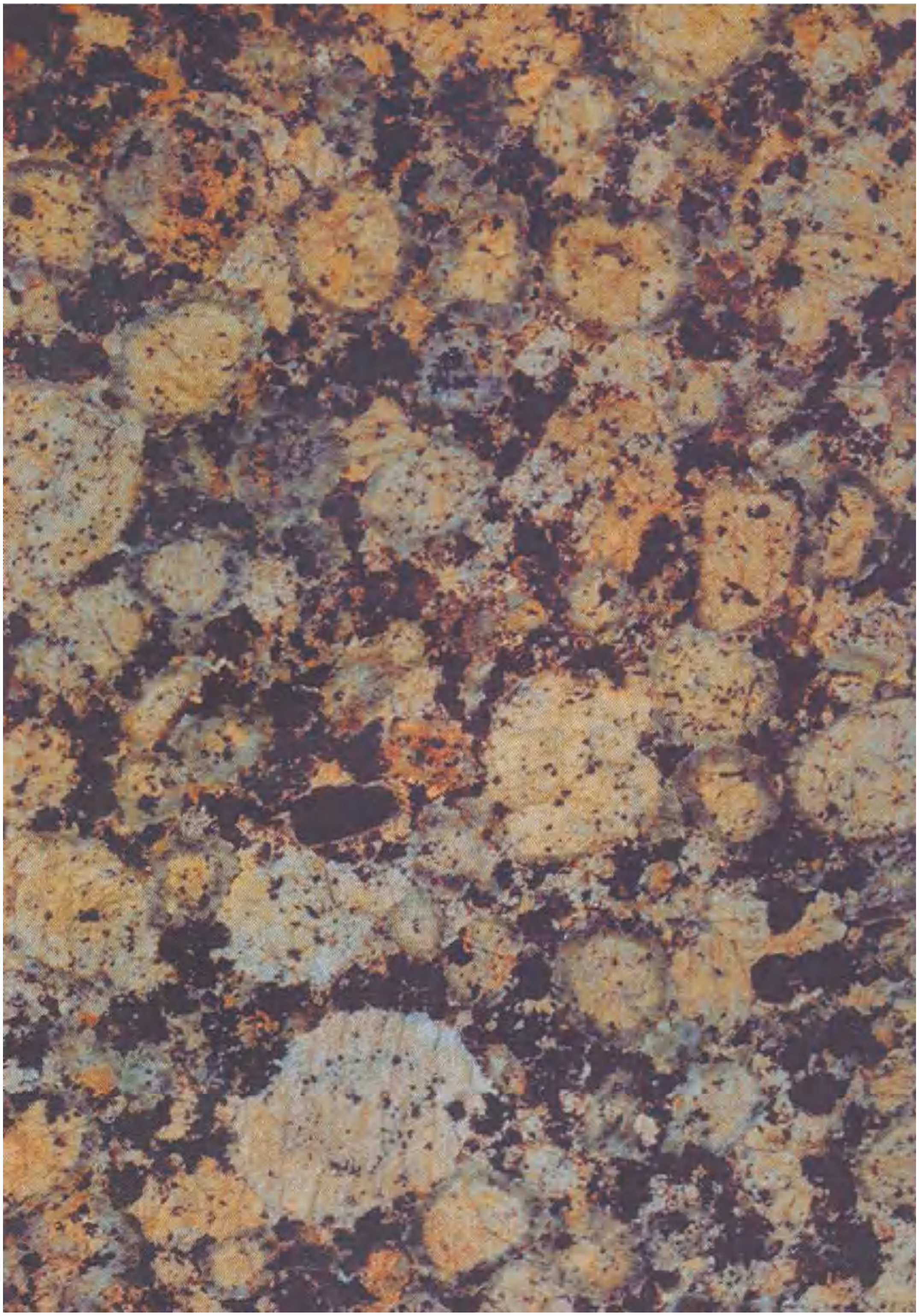

11. Rocks, meteorites and tektites

Rocks consist of minerals and often of fragments of other rocks and organic matter. Since rocks are mixtures, we cannot determine certain physical / chemical data (chemical formula, crystal system, hardness, etc.). Unlike minerals, some rocks form huge bodies which may cover several thousand km^2 so no single localities are shown in this book. Rocks are mainly used for building, in agriculture, and as raw materials for chemicals, ceramics and metals.
Rocks are generally divided into three principal groups according to their origin:
1. **igneous rocks**
2. **sedimentary rocks**
3. **metamorphic rocks**.

1. Igneous rocks

Properties: C – light gray, gray, white, brown, black, gray-green, pinkish, red-brown; D – opaque, rare transparent to translucent; L – dull, rare vitreous; DE – varies from 2.6 up to about 3.8; F – uneven, rare conchoidal; M – coarse to fine grained, commonly massive aggregates, consisting of microscopic or larger grains or crystals of various minerals, up to several decameters across, glass is also present rarely. Minerals in rocks usually have characteristic texture features, like graphic granite and others.
Origin and classification: Igneous rocks form at high temperature and commonly at high pressure in solidification of mainly silicate magma of variable composition. The mineral composition reflects a chemical composition of magma. Typical rock-forming minerals are quartz, orthoclase, microcline, plagioclase, biotite, muscovite, amphiboles, pyroxenes, olivine and nepheline. Rocks may solidify at various depths, according to which they are either intrusive, dyke or effusive. They can be grouped according to chemical composition (SiO2 content):
felsic – granite, syenite, pegmatite, rhyolite, obsidian
intermediate – diorite, andesit **basic** – gabbro, basalt
utrabasic – peridotite, varies from 1.0 up to about 2.8; F – uneven, sometimes conchoidal; M – coarse to fine grained and massive aggregates, consisting of grains or crystals, from microscopic to several dm in size, also of rock fragments and organic matter, banded textures are typical and some rocks are fossiliferous.

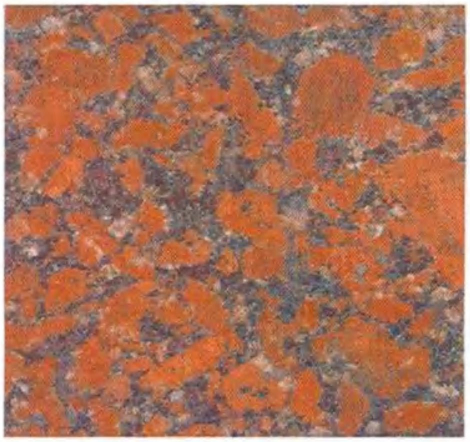
Granite, 65 mm, Norway

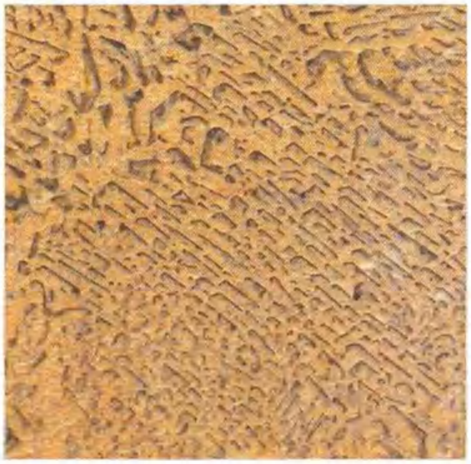
Orbicular granite, 150 mm, Sweden
Pegmatite, 65 mm, Dolni Bory, Czech Republic

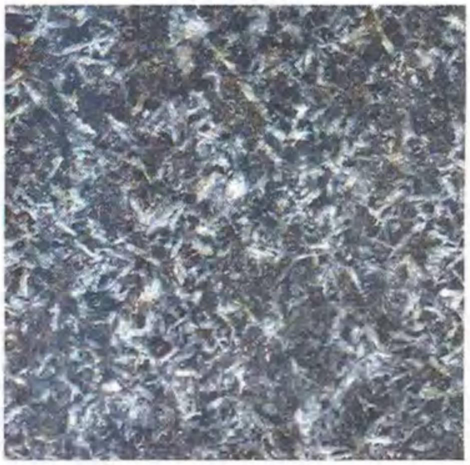
Diabase, 65 mm, Rožany, Czech Republic

Sandstone, 65 mm, Buffalo Gap, South Dakota, U.S.A.

Conglomerate, 65 mm, St.Alban, UK

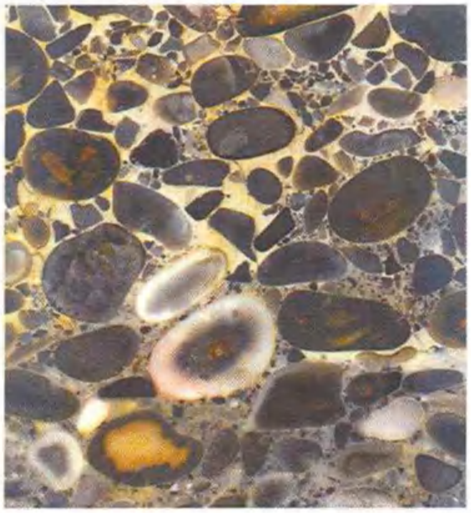

2. Sedimentary rocks

Properties: C – grey, brown, red-brown, black, green-grey, pinkish, white, yellow, often variable even within one rock, banding is typical; D – opaque, rare translucent; L – dull, sometimes vitreous, greasy or earthy; DE -varies from 1.0 up to about 2.8; F - uneven, sometimes conchoidal; M – coarse to fine grained and massive aggregates, consisting of grains or crystals, ranging from microscopic to several dm in size, also of rock fragments and organic matter, banded textures are typical and some rocks are fossiliferous.

Origin and classification: Sedimentary rocks originate under surface temperatures and pressures, as a result of sedimentation of mineral and rock fragments and organic matter of different size through the water and wind activity, or by precipitation from water solutions. Typical rock-forming minerals are quartz, calcite, dolomite, halite, clay minerals and others. According to their origins we can distinguish several groups of sedimentary rocks:

clastic (consisting of rock fragments) – sandstone, conglomerate, quartzite, siltstone.

organic (consisting mainly of organic matter) – limestone, coal.

chemical (originating by precipitation from water solutions) – evaporites, travertine.

Limestone, 65 mm, Italy

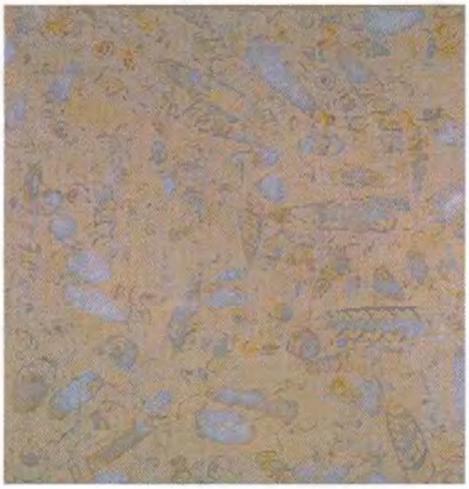

Evaporite, 40 mm, Walkenried, Germany

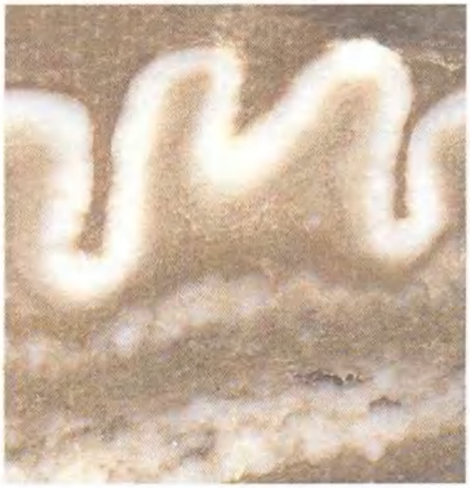

Serpentinite, 65 mm, Hrubšice, Czech Republic

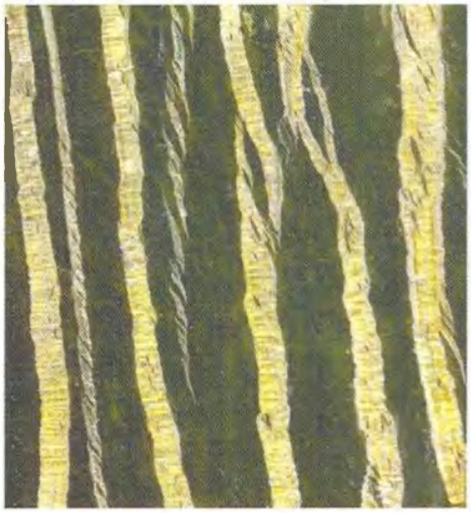

Gneiss, 65 mm, Doubravcany, Czech Republic

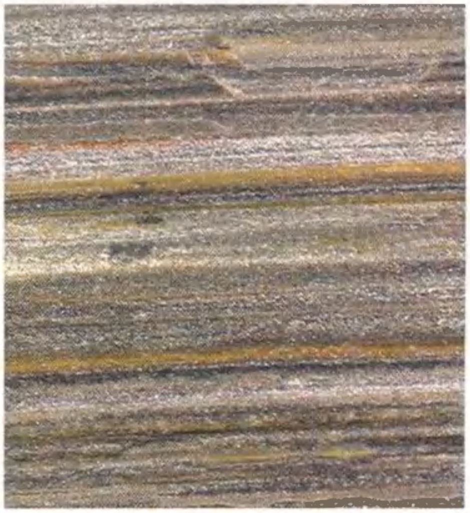

3. Metamorphic rocks

Properties: C – light gray, gray, brown, green, red-brown, black, green-gray, pinkish, white, sometimes variable within one rock, banding is relatively common; D – opaque; L – dull, rare vitreous; DE – varies from 2.5 to 4.8; F – uneven, rare conchoidal; M – coarse to fine-grained, platy, acicular and sometimes massive aggregates, consisting of grains and crystals, ranging from microscopic to several decimeters across. Typical are planar textures and foliation of some minerals.

Origin and classification: Metamorphic rocks originate under higher temperature and pressure during metamorphism of originally igneous or sedimentary rocks. The source of a thermal energy could be magma, then this type is called contact metamorphism, or the thermal source lies in the depth of the earth's crust and effects large areas, then this type is called regional metamorphism. During a process of metamorphism new minerals originate. Typical rock-forming metamorphic minerals are quartz, orthoclase, plagioclases, biotite, muscovite, amphiboles, pyroxenes, calcite, dolomite, sillimanite, kyanite, almandine, staurolite and serpentine.

Regionally metamorphosed rocks: serpentinite, mica schist, gneiss, marble.

Contact metamorphosed rocks: contact chert (porcelanite), skarn.

Marble, 100 mm, Greece

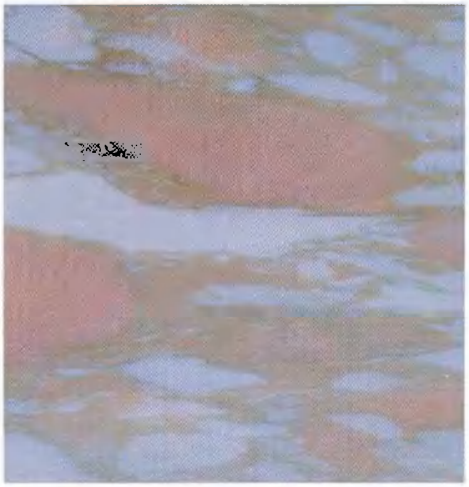

Skarn, 65 mm, Lísna, Czech Republic

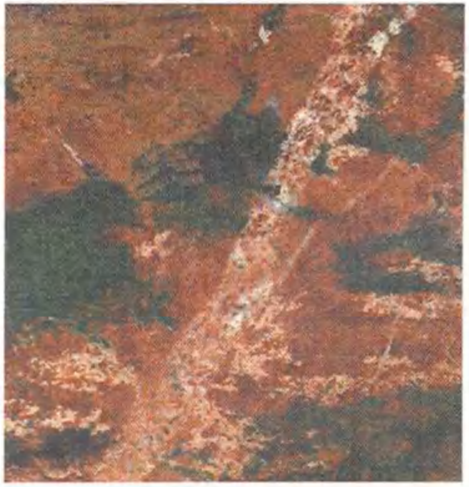

Iron meteorite, 50 mm, Sikhote-Alin, Russia

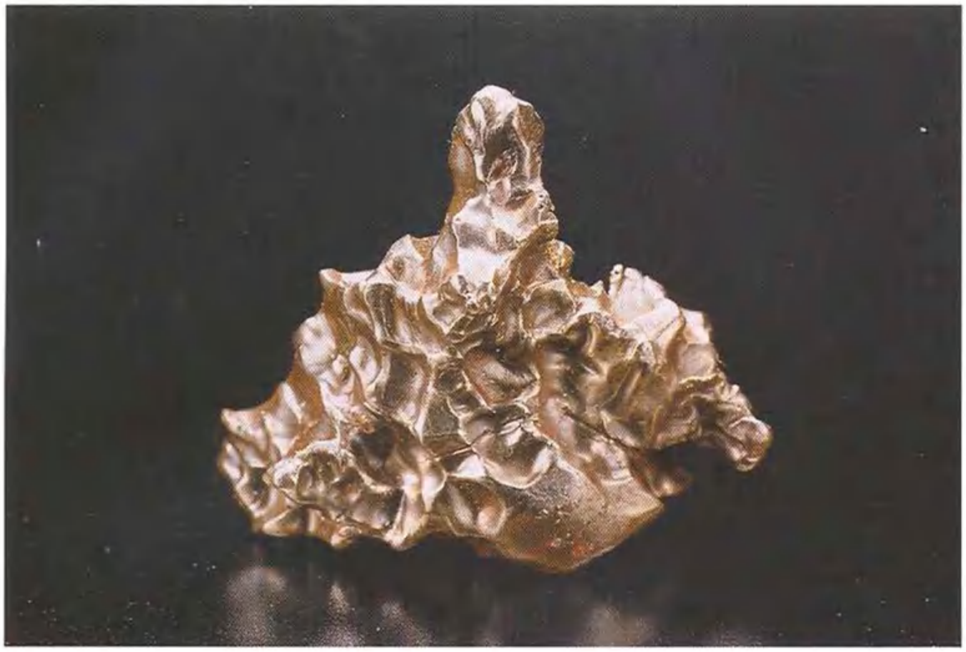

Iron meteorite, 40 mm detail, Gibeon, Namibia

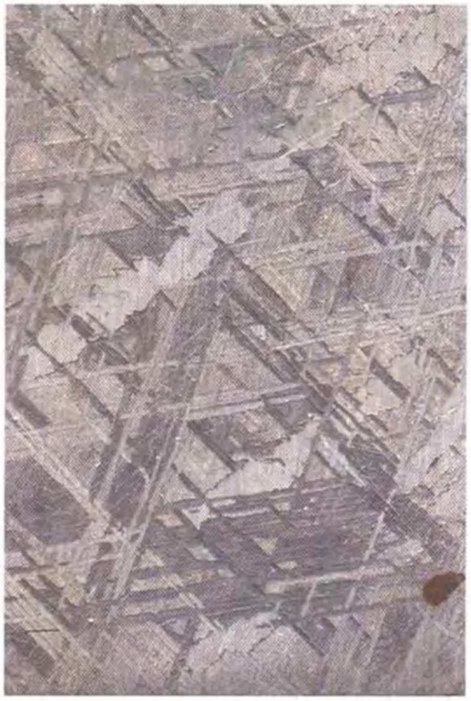

Meteorites

Properties: C - light gray, gray, gray-green, black; D – opaque; L – dull, metallic, rare vitreous; DE – varies from 3.0 to 7.3; F – uneven; M – coarse to fine-grained, sometimes massive aggregates, consisting of irregular grains of different minerals, ranging from microscopic to several cm across.
Origin and classification: Meteorites are igneousrocks formed in space. Most originate in the asteroidbelt between Mars and Jupiter. They consist of various minerals and their chemical composition differs greatly. Typical rock-forming minerals in meteorites are olivine, pyroxenes, plagioclases, Fe and Ni alloys and sulfides, rarely also organic matter. Meteorites fall into four main groups, accordingto metallic iron and silicate component:
iron meteorite;
siderolite;
chondrite;
achondrite;
Occurrence: Meteorite falls are known throughout the world. The largest known iron meteorite, weighing approximately 60 tons , is located near the Hoba farm, Namibia. The largest known chondrite, weighing about 1 ton , fell in 1948 in Norton County, Nebraska, USA. Most meteorites that have been found recently come from large glaciers such as those of Antarctica and from desert in Namibia.

Moldavite, 35 mm, Southern Bohemia, Czech Republic

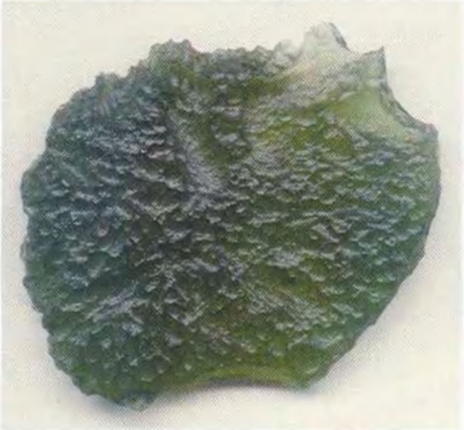

Tectite, 55 mm, Dalat, Vietnam

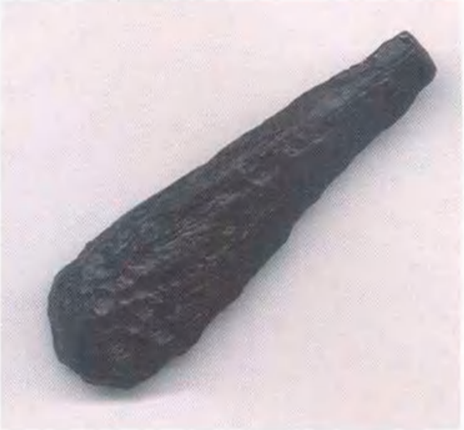

Tektites

Properties: C – light to dark green, yellow-green, brown-green, brown, green-gray, black; D – transparent, translucent to opaque; L – vitreous; DE – varies from 2.3 to 2.6; H – 6-7; F – conchoidal to uneven; M – massive irregular, drop-shaped or disc-shaped aggregates, irregular fragments, sometimes with typical sculptured surface.

Origin, classification and occurrence: Natural glasses, rich in SiO2, which formed as a result of rapid melting of surface rocks during impacts of large meteorites or comets.

They are classified according to their age and occurrence (following sequence from the oldest to the youngest):

Bediasites and georgianites – the USA, Mexico, Barbados, Cuba;

Urengoites – Novyi Urengoi, Russia;

Moldavites – southern Bohemia and western Moravia, Czech Republic;

Ivorites – Ivory Coast;

Irghisites – Zhamanshin, Russia;

Indochinites, phillipinites, javanites, bilitonites – southeastern Asia;

Australites – Australia.

Application: some tektites, mainly moldavites, as gemstones.

Moldavites, 40 mm, Southern Bohemia, Czech Republic

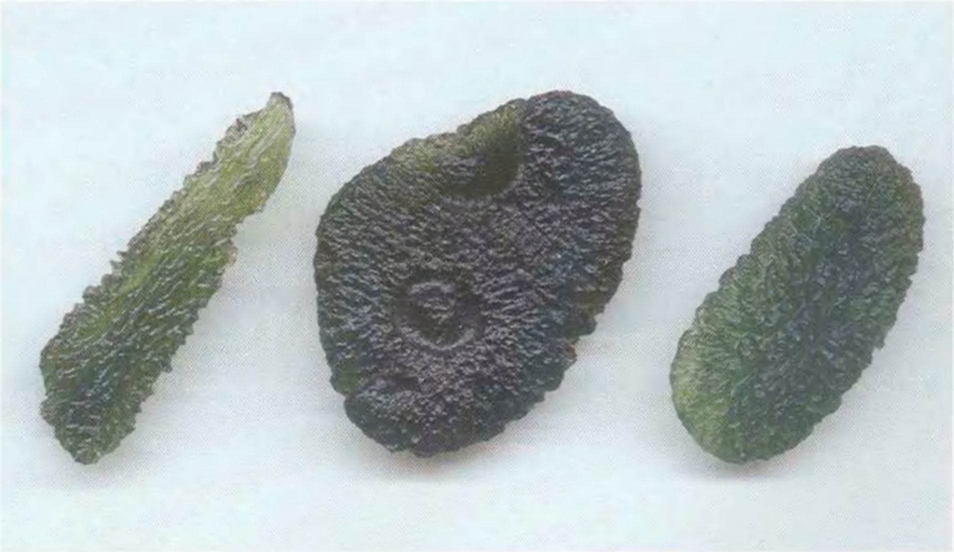

Recommended literature

Anthony J.W. et al. (1990, 1995, 1997):
Handbook of Mineralogy I, II, III,
Mineral Data Publishing, Tucson.

Bernard J.H., Rost R. a kol. (1992):
Encyklopedický prehled mineralů ,
Academia, Prague, in Czech.

Clark A.M. (1993):
Hey's Mineral Index, 3rd print,
Chapman & Hall, London.

Duda R., Rejl L. (1986):
Minerals of the World,
Hamlyn, Twickenham.

Fleischer M., Mandarino J.A. (1995):
Glossary of Mineral Species,
Mineralogical Record, Tucson.

Roberts W.L. et al. (1990):
Encyclopedia of Minerals,
Van Nostrand Reinhold, New York.

Strunz H. (1977):
*Mineralogische Tabellen,*6th print,
Geest & Portig, Leipzig.

Porcelanite, 85 mm, Komňa, Czech Republic
Agate, 50 mm, Zeleznice, Czech Republic

Rejstfiík

A

Acanthite 23
Actinolite 239
Adamite 154, 161
Adularia 263
Aegirine 235
Aenigmatite 241
Aeschynite-(Ce) 107
Agalmatolite 249
Agate 88, 94, 95
Agrellite 243
Aguilarite 24
Achroite 227
Aikinite 56
Ajoite 237
Akermanite 211
Alabandite 34
Alabaster 147
Albite 265
Alexandrite 80
Algodonite 19
Allanite-(Ce) 215
Allargentum 19
Allophane 259
Almandine 196
Altaite 36
Aluminite 148
Alumohydrocalcite 128
Alunite 140
Alunogen 144
Amazonite 264
Amber 280, 283
Amblygonite 158
Amesite 258
Amethyst 74, 88, 91
Analcime 271
Anapaite 176
Anatase 104
Ancylite-(Ce) 128
Andalusite 201
Andersonite 129
Andesine 266
Andorite 57
Andradite 198
Anglesite 137
Anhydrite 137
Ankerite 117
Annabergite 178
Annite 252
Anorthite 266
Anthophyllite 238
Antigorite 258
Antimony 14
Antlerite 139
Apjohnite 145
Aquamarine 220
Aragonite 119
Argentite 23
Argentopyrite 32
Argyrodite 25
Arsendescloizite 166
Arsenic 14
Arseniosiderite 188
Arsenolamprite 15
Arsenolite 81
Arsenopyrite 48
Arsentsumebite 169
Artinite 127
Astrophyllite 248
Atacamite 68
Atelestite 174
Augelite 164
Augite 233
Aurichalcite 123
Aurostibite 44
Autunite 190
Aventurine 88, 95
Azurite 122

B

Babingtonite 245
Balas 16
Barite 136, 138
Barytocalcite 121
Bastnasite-(Ce) 124
Baumhauerite 59
Bavenite 247
Bayldonite 168
Bazzite 224
Benitoite 218
Beraunite 181
Berthierite 53
Bertrandite 212
Beryl 220
Beryllonite 155
Berzelianite 21
Berzeliite 156
Betafite 87
Betekhtinite 57
Betpakdalite 153
Beudantite 141
Bilinite 146
Bindheimite 86
Biotite 252
Bismuth 15
Bismuthinite 40
Bismuthite 126
Bixbite 220, 223
Bixbyite 82
Boleite 69
Boracite 135
Borax 133
Bornite 22
Boromuscovite 250
Bort 16
Botryogen 148
Boulangerite 62
Bournonite 56
Braunite 206
Brazilianite 167
Breithauptite 33
Brochantite 139
Brookite 104
Brucite 109
Brushite 179
Buergerite 225
Bukovskyite 185
Bustamite 242

C

cabrerit 178
cafarsit 167
calaverit 42
caledonit 142
carnallit 68
carnotit 191
carrolit 38
cavansit 260
celestin 138
cerusit 120
cinabarit 37
citrín 88
clausthalit 35
clevelandit 265
clintonit 254
colemanit 133
coloradoit 27
cookeit 256
copiapit 149
coquimbit 144
cordierit 224
cornetit 164
cornubit 164
coronadit 102
cosalit 62
covellín 38
creedit 68
cristobalit 96
cronstedtit 258
cubanit 31
cumengit 70
curit 111
cylindrit 37
cyrilovit 187

â

aaroit 243
akalovit 262

D

dachiardit-Ca 275
danalit 269
danburit 267
datolit 208
dawsonit 124
delvauxit 184
démantoid 198
descloizit 166
devillin 149
diadochit 181
diaforit 58
diamant 16
diaspor 109
dickit 257
diopsid 232
dioptas 230
djurleit 20
dolomit 117
domeykit 19
drah˝ opal 97, 99
dravit 225
dresserit 129
dfievit˝ opal 97, 98
dufrenit 163
dufrénoysit 59
duftit 165
dumortierit 209
dyskrasit 20

E

edenit 239
edingtonit 274
egirin 235
elbait 227
elpidit 244
emplektit 53
enargit 30
enigmatit 241
enstatit 232
eosforit 182
epididymit 246
epidot 214
epistilbit 277
epsomit 144
erythrin 177
ettringit 150
eudialyt 218
euchroit 180
euklas 200
eulytin 200
euxenit-(Y) 107
evansit 186
evenkit 282

F

farmakosiderit 188
fassait 232
fayalit 195
fenakit 193
ferberit 105
ferrierit-Ca 275
ferrimolybdit 153
ferroaxinit 218
ferrocolumbit 106
ferrotapiolit 103
fersmanit 207
fibroferrit 147
fichtelit 282
flogopit 252
fluellit 67
fluorapatit 170
fluorapofylit 248
fluorit 66, 73
foitit 230
forsterit 194
fosfofylit 176
franckeit 36
franklinit 78
freibergit 30
freieslebenit 58
frondelit 163
fuchsit 250

G

gadolinit-(Y) 208
gahnit 76
galenit 35
gaudefroyit 131
gaylussit 126
gehlenit 211
geokronit 60
germanit 28
gersdorffit 45
getchellit 65
geysirit 97
gibbsit 109
gismondin 278
glaukonit 251
goethit 110
gonnardit 274
goosecreekit 279
gorgeyit 146
goshenit 220, 223
grafit 15
greenockit 30
grosular 197
gudmundit 48

H

halit 71
halotrichit 145
hambergit 134
hanksit 142
harmotom 278
hauerit 43
hausmannit 78
hedenbergit 233
heliodor 220, 222
heliotrop 88
helvín 270
hematit 84
hemimorfit 212
hessit 24
hessonit 197
heulandit-Ca 276
hiddenit 236
hollandit 102
holmquistit 238
howlit 209
hrachovec 119
hübnerit 105
huntit 118
hureaulit 174
hyacint 199
hyalit 97
hvalofan 264
hydrotalkit 127
hydroxylapatit 171
hydroxylherderit 158
hvdrozinkit 123

CH

chabasit-Ca 279
chalcedon 88, 92, 93
chalkantit 142
chalkofylit 185
chalkopyrit 27
chalkosiderit 183
chalkosin 20
chalkostibit 52
chalkotrichit 75
chamosit 257
chapmanit 205
chenevixit 184
chiastolit 201
chlorapatit 171
chloritoid 207
chondrodit 205
chromdiopsid 232
chromit 78
chryzoberyl 80
chryzokol 259
chryzopras 88, 92

I

ilmenit 84
ilmenorutil 103
ilvait 212
inderit 131
indigolit 227, 229
inesit 246
inyoit 132

J

jadeit 235
jahnsit-(CaMnMg) 187
jalpait 25
jamesonit 61
jantar 280, 283
jargon 199
jarosit 141
jaspis 88, 92
jeffersonit 232
jodargyrit 70

johannit 150
jordanit 60

K

kakoxen 181
kalcit 112, 116
kalomel 72
kammererit 256
kampylit 173
kankrinit 267
kaolinit 257
karbonado 16
karbonat-fluorapatit 171
karfolit 237
karminit 168
kasiterit 101
kasolit 211
katapleit 218
kermesit 40
kernit 134
kinoit 246
kirovit 143
kleiofan 26
klinohumit 205
klinochlor 256
klinoklas 165
klinoptilolit-Ca 277
klinozoisit 214
kobaltin 44
kobellit 63
konichalcit 165
kornerupin 209
korund 82
kovdorskit 189
krennerit 42
krokoit 152
krokydolit 241
kruĖait 44
kryolit 67
kfiemen 88
kfii‰Èal 88
kunzit 236
kuprit 75
kuprosklodowskit 210
kutnohorit 118
kvanit 202

L

labradorit 266
lamprofylit 213
langit 147
laumontit 275
lavendulan 188
lazulit 162
lazurit 268
leadhillit 142
legrandit 179
leifit 267
lepidokrokit 110
leucit 271
leukosafír 82
libethenit 160
liddicoatit 228
linarit 140
linneit 38
lirokonit 186
lithiofilit 156
lithioforit 111
lithiofosfát 155
livingstonit 55
lollingit 47
loparit-(Ce) 85
lorandit 55
lorenzenit 237
ludlamit 176
ludwigit 131
lussatit 96
luzonit 28

M

magmatické horniny 285
magnesit 113
magnetit 77
malachit 122
manganit 110
manganogedrit 238
manganotantalit 106
margarit 254
marialit 270
marianske sklo 147
markazit 6, 46
marmatit 26
masovˇ opal 97
mastek 249
mûi 9
mejonit 270
melanoflogit 100
melanit 198
melanterit 143
mellit 282
musiani kamen 263
mesolit 273
metamorfované horniny 287
meteority 288
miargyrit 36
mikroklin 264
mikrolit 87
milarit 230
millerit 34
mimetit 173
minium 79
mixit 189
mleanˇ opal 97
molybdenit 48
monazit-(Ce) 158
montebrasit 159
montmorillonit 255
morenosit 144
morganit 220, 223
morion 88, 90
moschellandsbergit 12
mottramit 167
muskovit 250

N

nagyagit 41
nakrit 258
natrojarosit 141
natrolit 272
nefelin 261
nefrit 235
neptunit 246
nikelin 33
niklskutterudit 49
nosean 268
novaaekit 191

O

ohnivˇ opal 97, 99
okenit 243
oligoklas 265
olivenit 161
olivín 194
olovo 12
omfacit 234
onyx 88, 94
opal 97, 98, 99
ortoklas 263

P

palygorskit 260
paragonit 251
pargasit 240
parisit-(Ce) 125
pearceit 54
pektolit 242
pentlandit 26
periklin 265
peristerit 265
perovskit 85
petalit 262
phillipsit-K 278
phosgenit 125
piemontit 214
pikrofarmakolit 179
pisanit 143
platina 13
plattnerit 101
pleonast 76
pollucit 271
polybasit 54
polyhalit 146
polylithionit 252
povondrait 226
prehnit 248
proustit 50
pseudomalachit 163
purpurit 156
pyknit 202
pyrargyrit 51
pyrit 42
pyrofylit 249
pyrochlor 86
pyrolusit 102
pyromorfit 172

pyrop 196
pyrostilpnit 51
pyrrhotin 32

R

rammelsbergit 47
rathit 59
realgár 64
rhodizit 134
riebeckit 241
richterit 240
rockbridgeit 162
rodochrozit 115
rodonit 245
rosasit 123
rossmanit 230
rtuÈ 11
rubelit 227, 228
rubín 82
rutil 100
rÚÏenín 88, 90

S

sádrovec 147
safflorit 47
safír 82
safírin 204
salmiak 72
sametka 110
samsonit 52
sanidin 263
saponit 255
sardonyx 88
sartorit 59
scorzalit 162
sedimentární horniny 286
sekaninait 224
seladonit 250
selen 17
selenit 147
semseyit 62
senarmontit 81
sepiolit 260
serandit 242
serpierit 149
sfalerit 18, 26
sferokobaltit 114
scheelit 151
siderit 114
silimanit 201
síra 16
skolecit 273
skorodit 175
skoryl 226
skutterudit 48
sluneaní kamen 265
smaragd 220
smirek 82
smithsonit 113
sodalit 268
sokolí oko 88
sperrylit 44
spessartin 197
spinel 76
spodumen 236
stannin 28
staurolit 204
stefanit 54
steinmannit 35
sternbergit 32
stibarsen 14
stibiconit 86
stibiotantalit 107
stichtit 128
stilbit-Ca 277
stilpnomelan 255
stolzit 152
strengit 174
stromeyerit 25
stroncianit 120
strunzit 180
stfiíbro 8, 10
sturmanit 150
sugilit 231
sylvanit 41
sylvín 72
symplesit 178

T

tanzanit 216
tarnowitzit 119
tefroit 195
tektity 289
tellur 17
tellurit 104
tennantit 28
tenorit 76
tetradymit 41
tetraedrit 29
thaumasit 206
thomsonit 273
thulit 216
tinzenit 219
tirolit 184
titanit 206
titanowodginit 106
todorokit 102
topaz 192, 202
torbernit 189
tremolit 239
tridymit 96
trifylin 155
trilithionit 253
triplit 160
tsavorit 197
tschermigit 146
tugtupit 269
Èujamunit 191
tygfií oko 88, 95
tyrkys 182
ulexit 133
ullmanit 46
umangit 22
uraninit 108
uranocircit 190
uranofan 211
uvarovit 199
uvit 228

V

valentinit 81
vanadinit 173
variscit 174
vauxit 180
verdelit 227, 228
vermiculit 255
vesignieit 168
vesuvian 217
veszelyit 185
viljujit 217
villiaumit 70
viridin 201
vivianit 176
vfiídlovec 119

W

wagnerit 159
wardit 187
wavelit 182
weloganit 126
whewelit 281
whiteit-(CaFeMg) 186
whitlockit 156
willemit 193
witherit 120
wittichenit 53
wollastonit 241
wulfenit 153
wurtzit 30

X

xanthokon 51
xenotim-(Y) 157
xonotlit 244

Z

záhnůda 88
zaratit 127
zeofylit 260
zeunerit 191
zinkenit 61
zinkit 75
zinnwaldit 253
zippeit 150
zirkon 199
zlato 11
zoisit 216
zwieselit 160

Î

Ïelezn˘ kvût 119
Ïelezo 12

THE COMPLETE ENCYCLOPEDIA OF
MINERALS

THE COMPLETE ENCYCLOPEDIA OF
MINERALS

The Complete Encyclopedia of Minerals is a book for all the people, who are interested in the beautiful world of minerals. They can find here hundreds of photographs of beautiful crystals and crystal aggregates, together with a list of classic and new occurrences for each mineral species. Minerals are arranged into chapters according to the mineral system:

• Elements • Sulfides • Halides • Oxides
• Carbonates • Borates • Sulphates • Phosphates
• Silicates • Organic compounds
• Rocks, meteorites and tektites

ISBN 1-84013-404-6
9 781840 134049

THE PERIODIC TABLE BOOK

A VISUAL ENCYCLOPEDIA OF THE ELEMENTS

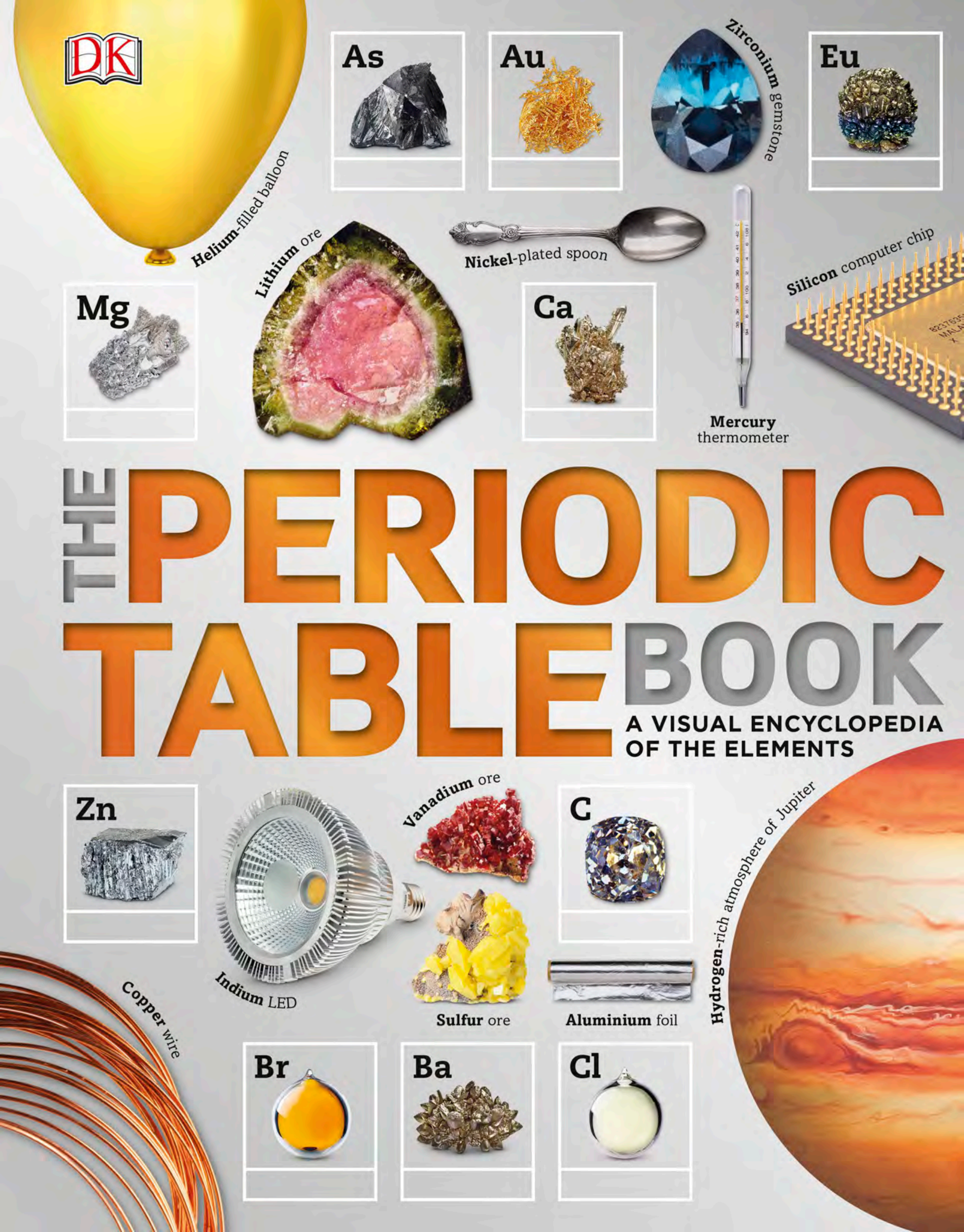
DK
As
Au
Zirconium gemstone
Eu
Helium-filled balloon
Lithium ore
Nickel-plated spoon
Silicon computer chip
Mg
Ca
Mercury thermometer
Zn
Vanadium ore
C
Hydrogen-rich atmosphere of Jupiter
Indium LED
Sulfur ore
Aluminium foil
Copper wire
Br
Ba
Cl

THE PERIODIC TABLE BOOK

A VISUAL ENCYCLOPEDIA
OF THE ELEMENTS

THE PERIODIC TABLE BOOK

A VISUAL ENCYCLOPEDIA OF THE ELEMENTS

WRITTEN BY **TOM JACKSON**
CONSULTANT **JACK CHALLONER**

DK | Penguin Random House

DK India

Senior Editor Bharti Bedi
Project Art Editor Amit Verma
Editorial Team Neha Ruth Samuel, Charvi Arora, Deeksha Saikia
Art Editors Mansi Agrawal, Amisha Gupta, Ravi Indiver
Assistant Art Editors Neetika Malik Jhingan, Nidhi Rastogi
Jacket Designer Suhita Dharamjit
Jackets Editorial Coordinator Priyanka Sharma
Senior DTP Designer Harish Aggarwal
DTP Designers Sachin Gupta, Syed Md Farhan, Nityanand Kumar, Mohammad Rizwan
Picture Researcher Nishwan Rasool
Managing Jackets Editors Saloni Singh, Sreshtha Bhattacharya
Picture Research Manager Taiyaba Khatoon
Pre-production Manager Balwant Singh
Production Manager Pankaj Sharma
Managing Editor Kingshuk Ghoshal
Managing Art Editor Govind Mittal

DK UK

Project Editor Ashwin Khurana
Senior Art Editor Smiljka Surla
Jacket Editor Claire Gell
Senior Jacket Designer Mark Cavanagh
Jacket Design Development Manager Sophia MTT
Managing Editor Dr Lisa Gillespie
Managing Art Editor Owen Peyton Jones
Producers, Pre-production Dragana Puvacic, Catherine Williams
Producer Anna Vallarino
Publisher Andrew Macintyre
Art Director Karen Self
Associate Publishing Director Liz Wheeler
Design Director Phil Ormerod
Publishing Director Jonathan Metcalf

Photographer Ruth Jenkinson
Photography Assistant Julie Stewart

Element samples prepared and supplied by RGB Research Ltd
www.periodictable.co.uk

First published in Great Britain in 2017 by
Dorling Kindersley Limited
80 Strand, London WC2R 0RL

Copyright © 2017 Dorling Kindersley Limited
A Penguin Random House Company
10 9 8 7 6 5 4 3 2 1
001–289022–April/2017

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the copyright owner.

A CIP catalogue record for this book is available from the British Library

ISBN: 978-0-2412-4043-4

Printed in China

A WORLD OF IDEAS:
SEE ALL THERE IS TO KNOW

www.dk.com

CONTENTS

Foreword 6
Elemental building blocks 8
Chemical discoveries 10
Inside an atom 12
Periodic table of elements 14
Reactions and uses 16

Hydrogen 18

Hydrogen 20

Alkali Metals 22

Lithium 24
Sodium 26
Salt flats 28
Potassium 30
Rubidium 32
Caesium, Francium 34

Alkaline Earth Metals 36

Beryllium 38
Magnesium 40
Calcium 42
Fly Geyser 44
Strontium 46
Barium 48
Radium 50

Transition Metals 52

Scandium, Titanium 54
Vanadium, Chromium 56
Manganese 58
Iron 60
Steelmaking 62
Cobalt 64
Nickel 66
Copper 68
Copper wires 70
Zinc 72
Yttrium 74
Zirconium, Niobium 76
Molybdenum, Technetium 78
Ruthenium, Rhodium 80
Palladium 82
Silver 84
Cadmium, Hafnium 86
Tantalum, Tungsten 88
Rhenium, Osmium 90
Iridium 92
Platinum 94
Gold 96
Golden Buddha 98
Mercury 100
Rutherfordium, Dubnium, Seaborgium 102
Bohrium, Hassium, Meitnerium 104
Darmstadtium, Roentgenium, Copernicium 106

Lanthanides 108

Lanthanum, Cerium, Praseodymium 110
Neodymium, Promethium, Samarium, Europium 112
Gadolinium, Terbium, Dysprosium, Holmium 114
Erbium, Thulium, Ytterbium, Lutetium 116

Actinides 118

Actinium, Thorium, Protactinium 120
Uranium, Neptunium, Plutonium, Americium 122
Curium, Berkelium , Californium, Einsteinium 124
Fermium, Mendelevium, Nobelium, Lawrencium 126

The Boron Group 128

Boron 130
Aluminium 132
Jet turbine 134
Gallium, Indium 136
Thallium, Nihonium 138

The Carbon Group 140

Carbon 142
Pink diamond 144
Silicon 146
Germanium, Tin 148
Lead, Flerovium 150

The Nitrogen Group 152

Nitrogen 154
Drag racing 156
Phosphorus 158
Arsenic, Antimony 160
Bismuth, Moscovium 162

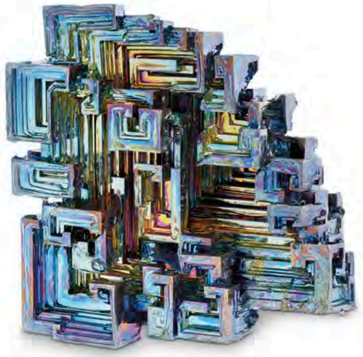

The Oxygen Group 164

Oxygen 166
Sulfur 168
Danakil Depression 170
Selenium, Tellurium 172
Polonium, Livermorium 174

The Halogen Group 176

Fluorine 178
Chlorine 180
Ocean clean up 182
Bromine 184
Iodine, Astatine, Tennessine 186

Noble Gases 188

Helium 190
Nebula 192
Neon, Argon 194
Krypton, Xenon 196
Radon, Oganesson 198

Glossary 200
Index 204
Acknowledgements 208

Chunk of yttrium

Chunk of silver

Zirconium crystal bar

Foreword

Everything in nature, from the mountains and the oceans to the air we breathe and food we eat are made up of simple substances called elements. You may have already heard of several of them, including gold, iron, oxygen, and helium, but these are just four out of a total of 118. Many have unique – and sometimes surprising – chemical and physical properties. Gallium, for example, is a solid but melts in your hand. A compound of sulfur gives off a nasty smell of rotten eggs. Fluorine is a gas that can burn a hole straight through concrete!

The elements are rarely found in their pure form. Mostly, they are combined with each other to make compounds, which make up substances around us. For example, hydrogen and oxygen make water, sodium and chlorine form salt, and carbon is found in millions of compounds, many of which – including proteins and sugars – make our bodies work.

To find out more about the elements, we need to take a good look at the periodic table. This is used by scientists around the world to list and detail the elements. It shows the key information

Nickel balls

Cube of melting gallium

Iodine in a glass sphere

Barium crystals

Chunk of grey selenium

Magnesium crystals

Osmium pellet

for each element, grouping them into similar types. With this information, we can use the elements to make many things we need: a fluorine compound in toothpastes toughens our teeth and silicon crystals engineered into microchips operate our gadgets and phones.

Every element has its own story of where it comes from, what it can do, and how we use it. Let's begin a tour of every element one by one. It's going to be a fascinating journey.

Tom Jackson

Throughout this book you will find boxes with the following symbols. This is what each of them mean.

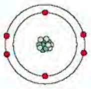

	This shows the structure of an atom of an element, with the nucleus (made of protons and neutrons) at the centre and electrons surrounding it in their shells.
⊖	Electron
⊕	Proton
○	Neutron
State	The state of the element at a temperature of 20°C (28°F). It can be a liquid, solid, or gas.
Discovery	This details the year in which the element was discovered.

Chunk of uranium

Gold crystals

Thulium crystals

Calcium crystals

Elemental building blocks

Elements are everywhere: some you can see, like gold, others are almost invisible, like oxygen gas. An element is a substance that cannot be broken up into simpler ingredients. Each one is made up of tiny building blocks called atoms, which are unique for every element. Most elements are joined with other elements to make compounds, which are made by combining two or more elements. This includes water, which is a compound of hydrogen and oxygen.

Elements in our world

There are 118 elements in the periodic table; 92 of them are found in nature, while the others are made by humans. Every element is unique. Most of the elements are solids, like the metals. At room temperature, 11 elements are gases, while bromine and mercury are the only two liquids.

Bromine liquid with bromine gas

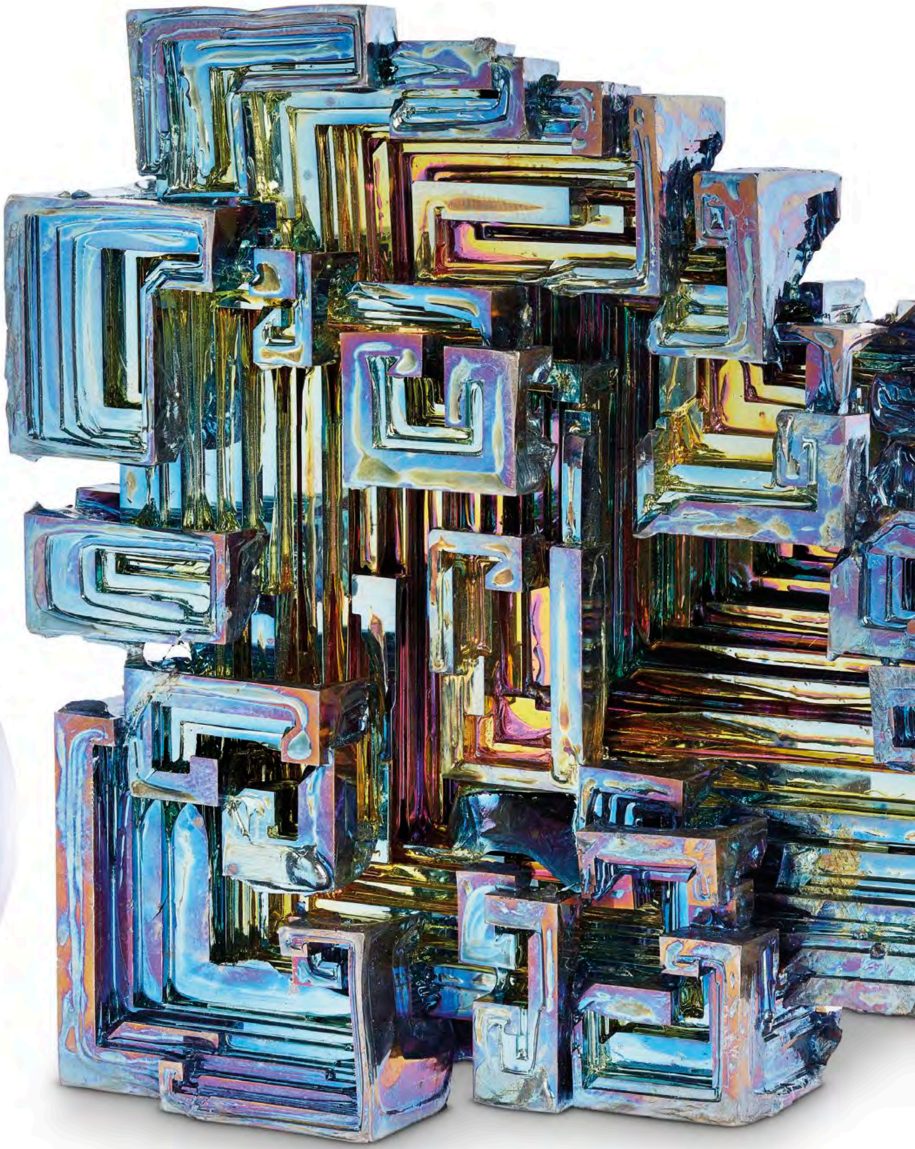

Bismuth crystals

Earth

Water

Air

Fire

Elements in and around us

About 99 per cent of the human body is made from just six elements, though they are combined together to form thousands of different compounds. On the other hand, Earth's atmosphere is a mixture of gases, most of which are pure elements. About 99 per cent of the air is made from nitrogen and oxygen.

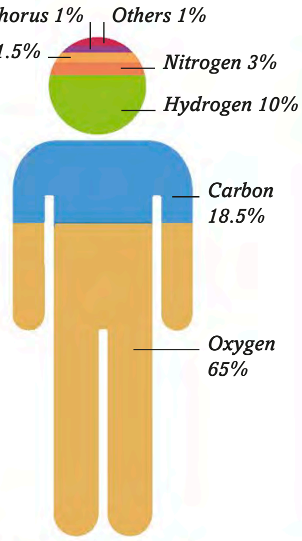

Human body

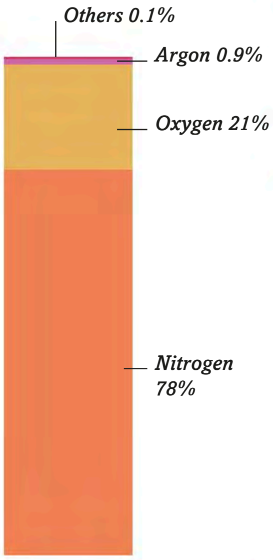

Earth's atmosphere

Gold crystals

Ancient ideas

The idea of elements is very old, dating back about 2,600 years to ancient Greece. However, Greek thinkers believed that the world was made of just four elements: earth, water, fire, and air. Empedocles, an influential scholar, was the first to propose that these elements made up all structures. Only much later did scientists learn that none of these are actually elements. For thousands of years, everybody from ancient Egyptian priests to medieval European alchemists, speculated about the definiton and classification of an element.

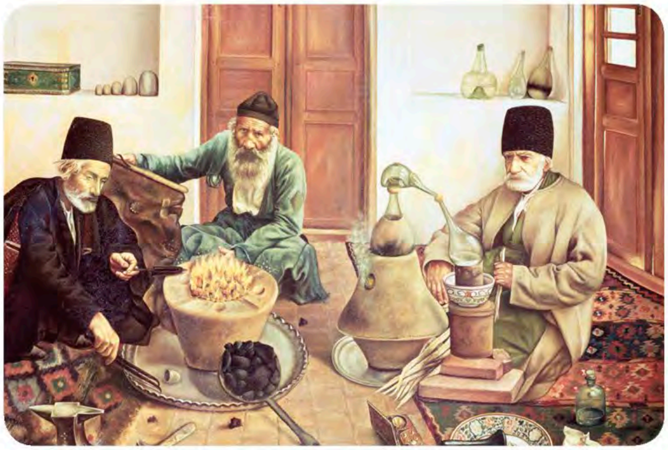

Iranian alchemists in their workshop

Alchemy and mysticism

Chemists are scientists who study elements and compounds. However, before they existed, the alchemists were medieval researchers. Believing in a mixture of science and magic, alchemists tried to change ordinary metals (such as lead) into gold. They failed because elements cannot be changed from one type to another. But, in the process, they discovered many new elements and developed several processes that chemists still use today.

ROBERT BOYLE

The first person to use science to understand the elements was the Irish scientist and inventor Robert Boyle. He pursued science through reason, and in the 1660s he performed the first chemistry experiments to show that much of what the alchemists believed was wrong.

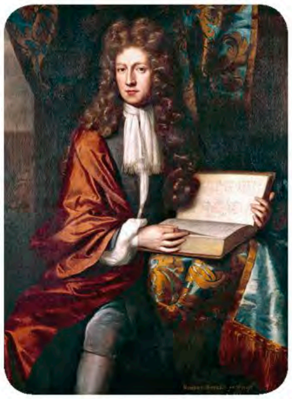

Chemical discoveries

Humphry Davy
In the early 19th century, the English scientist Humphry Davy discovered several new metals. He used a revolutionary process called electrolysis, in which electric currents split chemical compounds into their elements. Davy discovered a total of nine new elements, including magnesium, potassium, and calcium.

The ancient concept of four elements – earth, water, fire, and air – expanded to a belief that every substance on Earth was made from a mixture of these elements. However, many substances including mercury, sulfur, and gold did not fit this idea. Over the last 300 years, chemists have followed a long series of clues to reveal the true nature of elements, their atoms, and what happens to them during chemical reactions.

Pioneering chemists

Many of the first breakthroughs in chemistry came in the 1700s, from investigations into the composition of air. Chemists such as Joseph Black, Henry Cavendish, and Joseph Priestly discovered several different "airs", which we now call gases. They also found that the gases could react with solid substances, which they called "earths". These discoveries began a journey that revealed that there were dozens of elements, not just four. Today, scientists have identified 118 elements, but more may be discovered in time.

Antoine Lavoisier
In 1777, the French scientist Antoine Lavoisier proved that sulfur was an element. This yellow substance was familiar for thousands of years, but Lavoisier performed experiments to show that it was a simple substance that could not be divided up any further. In the same year, he also found out that water was not an element, but a compound of hydrogen and oxygen.

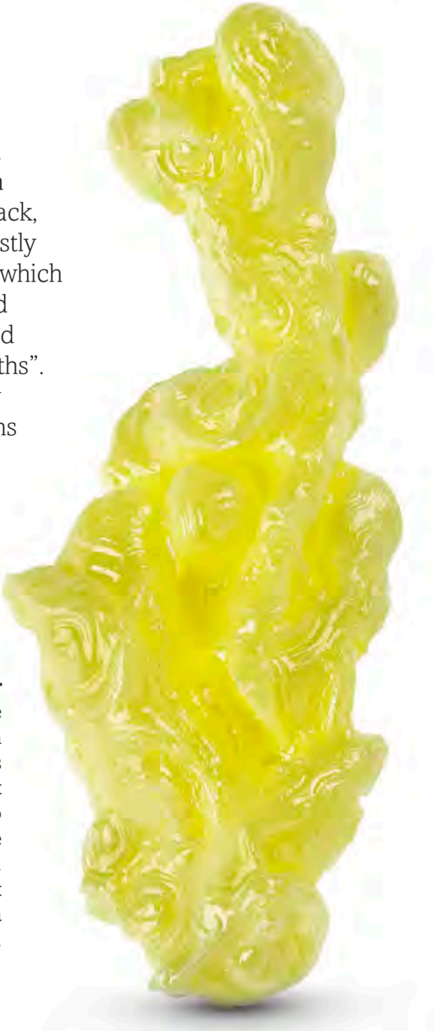

Granule of pure sulfur

Magnesium crystals

JOHN DALTON

Like many scientists of his day, the English scientist John Dalton already believed that matter must be made of tiny particles. In 1803, he began to think about how these particles might join together. He came to realize that there are different particles for every element, and that the particles of one element all have the same mass. He also realized that the particles of different elements combine in simple proportions to make compounds. So, for example, the particles of the elements carbon and oxygen can combine to make carbon monoxide. He suggested that during a chemical reaction, the particles rearrange to make compounds. He formulated the first modern theory of atoms.

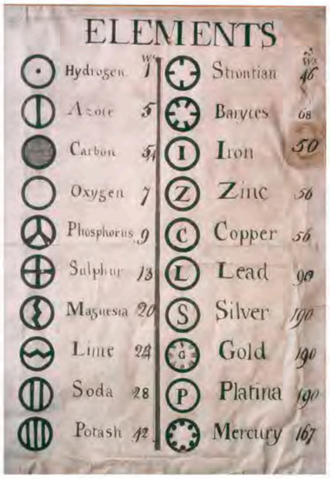

ELEMENTS

	Wt		Wt
Hydrogen	1	Strontian	46
Azote	5	Barytes	68
Carbon	5,4	Iron	50
Oxygen	7	Zinc	56
Phosphorus	9	Copper	56
Sulphur	13	Lead	90
Magnesia	20	Silver	190
Lime	24	Gold	190
Soda	28	Platina	190
Potash	42	Mercury	167

Dalton's table of elements

States of matter

Elements can exist in three states of matter: solid, liquid, and gas. At room temperature, most elements are solids, 11 are gases, and only two are liquids. However, elements can change from one state into another. These changes don't alter the atoms of these elements, but arrange them in different ways.

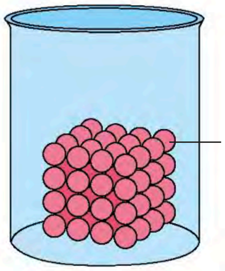

A solid keeps its shape and has a fixed volume.

In a solid, all the atoms are attracted to each other and locked in position.

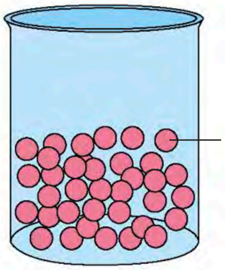

A liquid takes the shape of its container, but its volume remains fixed.

In a liquid, the atoms begin to move around as the attraction between them weakens.

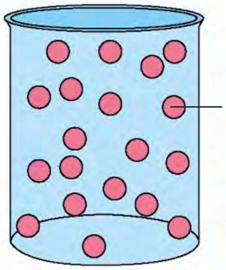

A gas will fill any container, no matter how large or small.

In a gas, the atoms are weakly attracted to each other, so they all move in different directions.

Jacob Berzelius

In the early 1800s, the Swedish doctor Jacob Berzelius investigated chemicals in rocks and minerals. He found two minerals that contained new elements. He named these elements cerium (after Ceres, the dwarf planet) and thorium (after Thor, the Viking god of thunder). Berzelius also invented a system of using symbols and numbers that chemists still use to identify elements and compounds today.

Chunk of pure cerium

Pure caesium inside a sealed container

Robert Bunsen

The German chemist Robert Bunsen is best known for inventing a gas burner that is often used in laboratories. In the 1850s, Bunsen used such a burner – which produced a hot, clean flame – to study the unique flame colours produced by different elements. When an unknown substance made bright blue flames, he named it caesium, meaning "sky blue".

Inside an atom

An atom is the smallest unit of an element. Atoms are too small to see (even with the most powerful microscopes) but they are everywhere. They consist of smaller particles called protons, neutrons, and electrons. Every element has a unique number of protons.

What's the atomic number?

The number of protons in an atom of an element is called the atomic number. The atomic number of an atom identifies the element it belongs to. Every atom also has an equal number of electrons. For elements found naturally on Earth, hydrogen has the smallest atomic number (1), while uranium atoms have the highest atomic number (92).

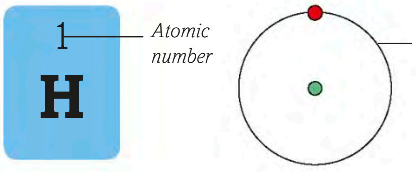

This shell is the space in the hydrogen atom where one electron circles the proton at the centre of the atom.

Hydrogen atom

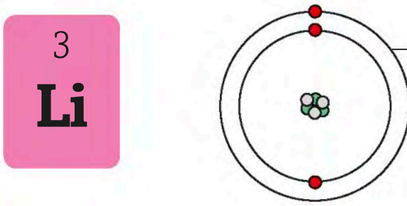

In a lithium atom, two shells house three electrons, which circle the protons and neutrons at the centre.

Lithium atom

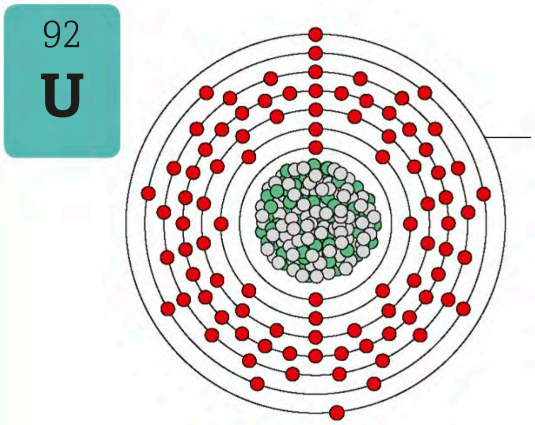

Seven shells house the 92 electrons in a uranium atom.

Uranium atom

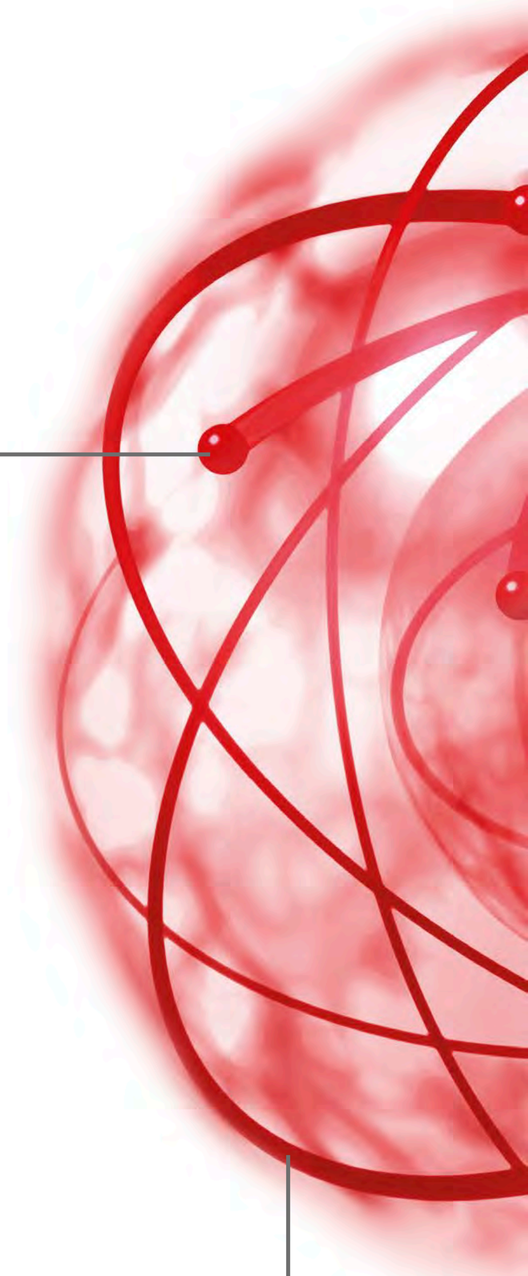

Electron ❯ The tiny, negatively charged particles in an atom are called electrons. They are involved in the way the atoms of an element react and form bonds with the atoms of other elements.

Shell ❯ The electrons in an atom move around the nucleus. They are arranged in layers called shells. When reacting with each other, atoms tend to fill up their outer shells to become more stable.

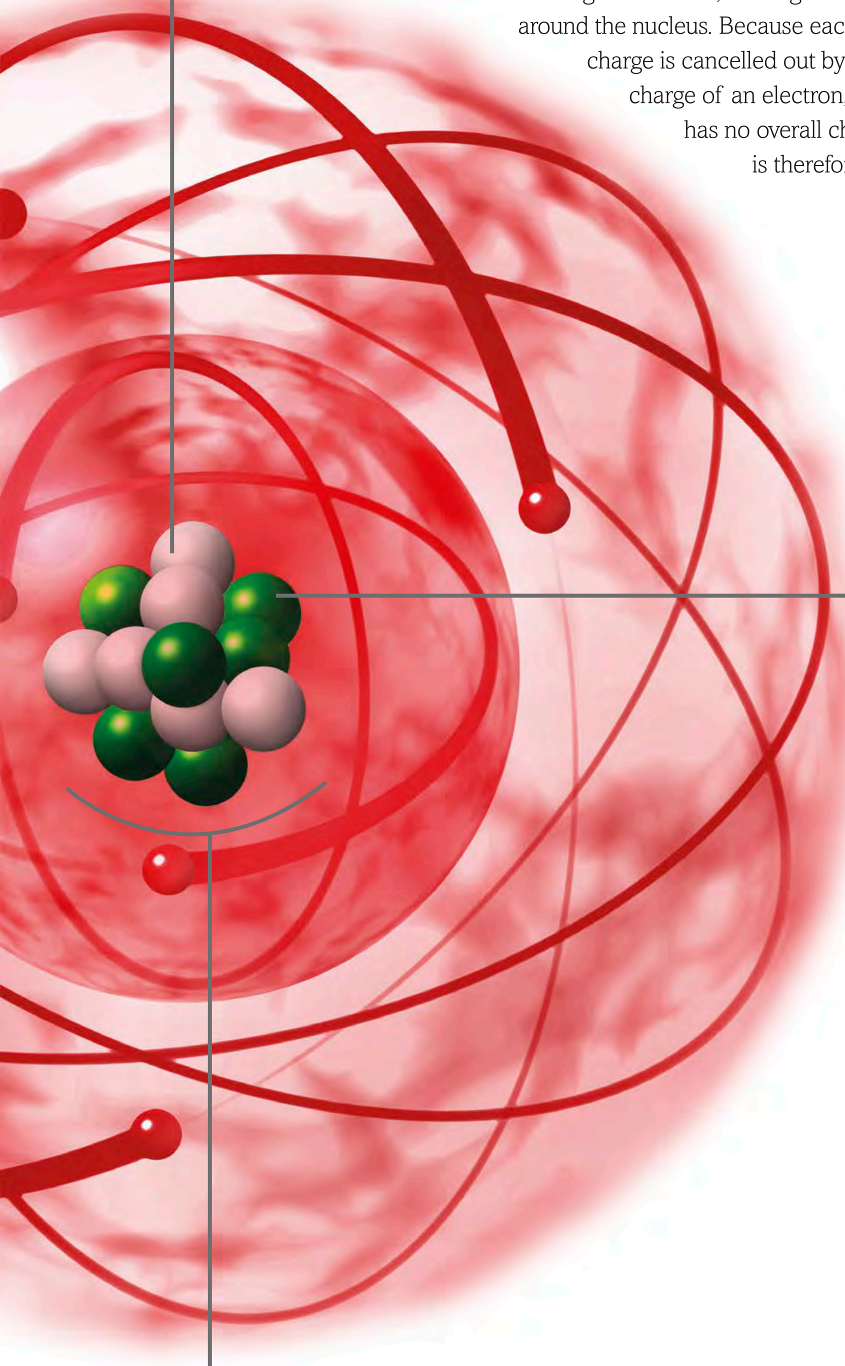

Neutron ❯ As its name suggests, neutrons are neutral particles, which means they do not have an electric charge. A neutron weighs the same as a proton, and much more than a electron.

Proton ❯ Protons have a positive electric charge. This charge attracts the negatively charged electrons, holding them in place around the nucleus. Because each proton's charge is cancelled out by the equal charge of an electron, the atom has no overall charge, and is therefore neutral.

Nucleus ❯ The central core, or nucleus, of an atom is made up of protons and neutrons. Nearly all the mass of the atom is packed into the nucleus, and this gives every element a unique atomic mass.

Atomic facts

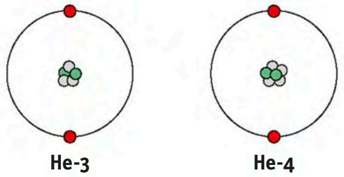

Isotopes

While every element has a unique number of electrons and protons in its atoms, the number of neutrons can vary. These different forms are called isotopes. For example, helium has two isotopes: one contains three neutrons (He-3), the other has four (He-4).

Electromagnet attracts metal pieces

Electromagnetism

Atoms work like tiny magnets. A force called electromagnetism holds them together. It makes particles with opposite charges, such as protons and electrons, attract each other. Those with similar charges repel each other. A magnet is an object in which the magnetic forces of the atoms attract and repel other objects. An electromagnet develops magnetism when an electric current runs through it.

ATOMIC PIONEERS

During his atomic research in the early 20th century, Sir Ernest Rutherford, a New Zealand scientist, expanded our understanding of the structure of atoms. He discovered protons and proved that they were located in an atom's nucleus.

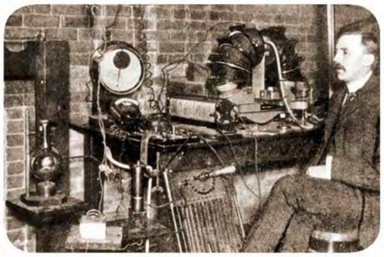

Sir Ernest Rutherford

Periodic table of elements

The periodic table is a useful way of organizing the elements. It arranges the elements in order of their atomic number, which is the number of protons in the nucleus of an atom, and is unique to every element. The table also divides the elements into rows, called “periods”, and columns, called “groups”. Dmitri Mendeleev, the chemist who devised the table, arranged the elements based on the similarity of certain physical and chemical properties.

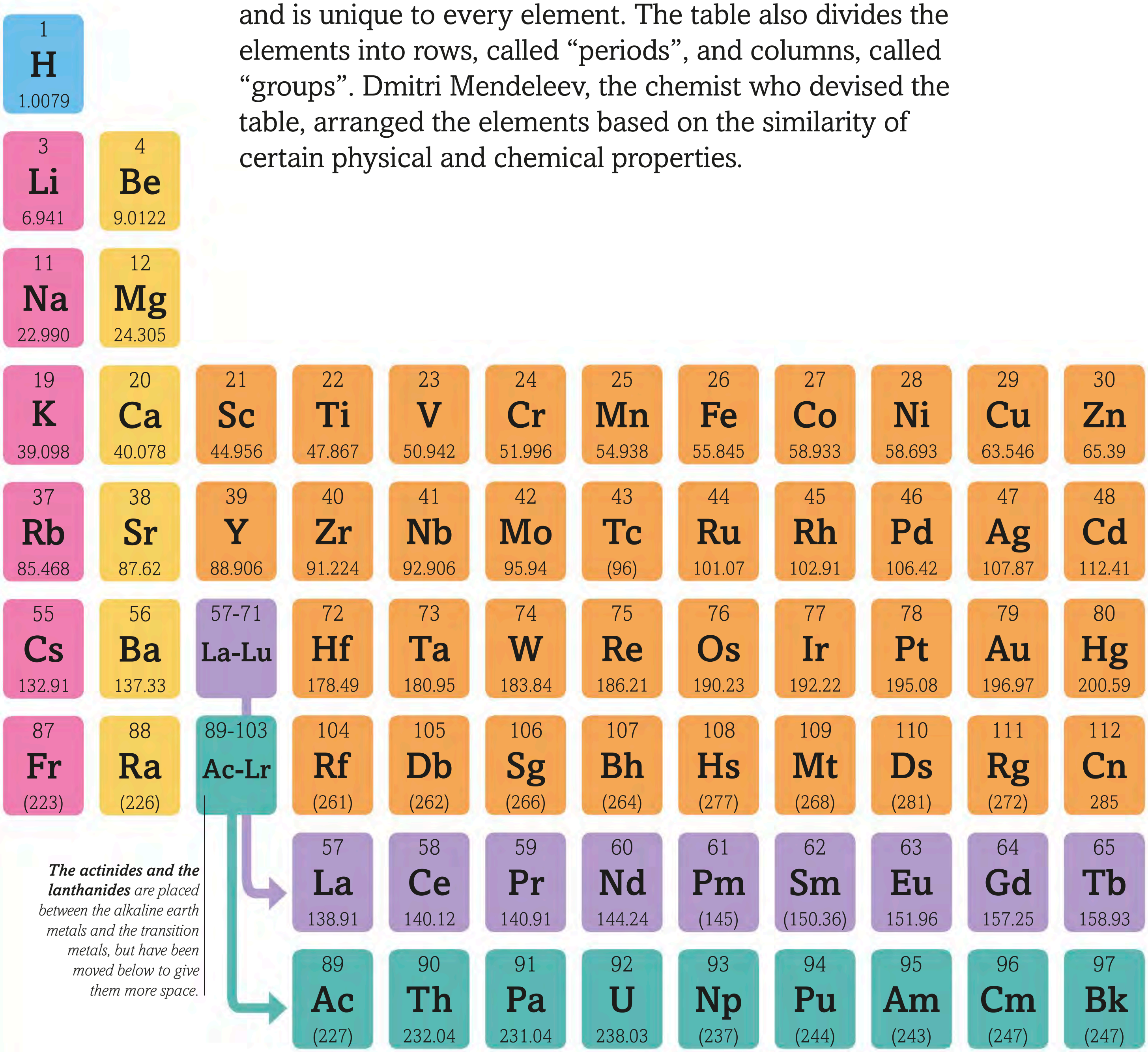

1 **H** 1.0079											
3 **Li** 6.941	4 **Be** 9.0122										
11 **Na** 22.990	12 **Mg** 24.305										
19 **K** 39.098	20 **Ca** 40.078	21 **Sc** 44.956	22 **Ti** 47.867	23 **V** 50.942	24 **Cr** 51.996	25 **Mn** 54.938	26 **Fe** 55.845	27 **Co** 58.933	28 **Ni** 58.693	29 **Cu** 63.546	30 **Zn** 65.39
37 **Rb** 85.468	38 **Sr** 87.62	39 **Y** 88.906	40 **Zr** 91.224	41 **Nb** 92.906	42 **Mo** 95.94	43 **Tc** (96)	44 **Ru** 101.07	45 **Rh** 102.91	46 **Pd** 106.42	47 **Ag** 107.87	48 **Cd** 112.41
55 **Cs** 132.91	56 **Ba** 137.33	57-71 **La-Lu**	72 **Hf** 178.49	73 **Ta** 180.95	74 **W** 183.84	75 **Re** 186.21	76 **Os** 190.23	77 **Ir** 192.22	78 **Pt** 195.08	79 **Au** 196.97	80 **Hg** 200.59
87 **Fr** (223)	88 **Ra** (226)	89-103 **Ac-Lr**	104 **Rf** (261)	105 **Db** (262)	106 **Sg** (266)	107 **Bh** (264)	108 **Hs** (277)	109 **Mt** (268)	110 **Ds** (281)	111 **Rg** (272)	112 **Cn** 285
			57 **La** 138.91	58 **Ce** 140.12	59 **Pr** 140.91	60 **Nd** 144.24	61 **Pm** (145)	62 **Sm** (150.36)	63 **Eu** 151.96	64 **Gd** 157.25	65 **Tb** 158.93
			89 **Ac** (227)	90 **Th** 232.04	91 **Pa** 231.04	92 **U** 238.03	93 **Np** (237)	94 **Pu** (244)	95 **Am** (243)	96 **Cm** (247)	97 **Bk** (247)

The actinides and the lanthanides *are placed between the alkaline earth metals and the transition metals, but have been moved below to give them more space.*

KEY

- ■ Hydrogen
- ■ Alkali Metals
- ■ Alkaline Earth Metals
- ■ Transition Metals
- ■ Lanthanides
- ■ Actinides
- ■ The Boron Group
- ■ The Carbon Group
- ■ The Nitrogen Group
- ■ The Oxygen Group
- ■ The Halogen Group
- ■ Noble Gases

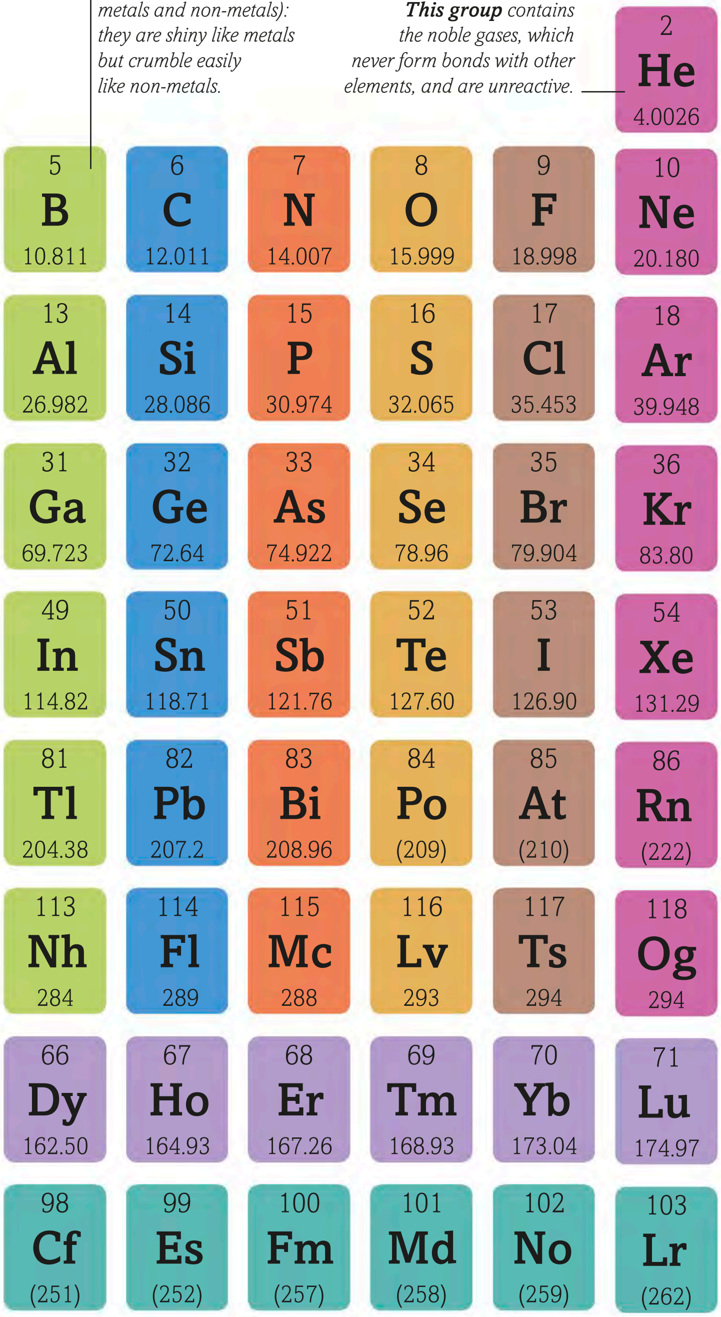

Elements of this group *are semi-metals (elements with the properties of metals and non-metals): they are shiny like metals but crumble easily like non-metals.*

This group *contains the noble gases, which never form bonds with other elements, and are unreactive.*

					2 He 4.0026
5 B 10.811	6 C 12.011	7 N 14.007	8 O 15.999	9 F 18.998	10 Ne 20.180
13 Al 26.982	14 Si 28.086	15 P 30.974	16 S 32.065	17 Cl 35.453	18 Ar 39.948
31 Ga 69.723	32 Ge 72.64	33 As 74.922	34 Se 78.96	35 Br 79.904	36 Kr 83.80
49 In 114.82	50 Sn 118.71	51 Sb 121.76	52 Te 127.60	53 I 126.90	54 Xe 131.29
81 Tl 204.38	82 Pb 207.2	83 Bi 208.96	84 Po (209)	85 At (210)	86 Rn (222)
113 Nh 284	114 Fl 289	115 Mc 288	116 Lv 293	117 Ts 294	118 Og 294
66 Dy 162.50	67 Ho 164.93	68 Er 167.26	69 Tm 168.93	70 Yb 173.04	71 Lu 174.97
98 Cf (251)	99 Es (252)	100 Fm (257)	101 Md (258)	102 No (259)	103 Lr (262)

Reading the table

Element symbol

Every element has a unique symbol of one or two letters. These symbols ensure that scientists who speak different languages do not get confused while describing the same element.

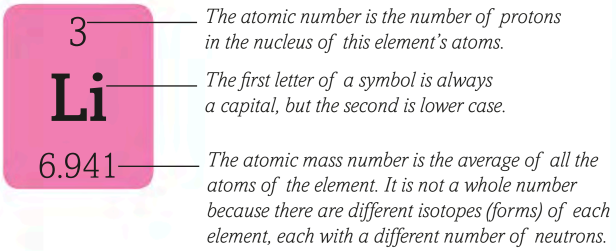

The atomic number is the number of protons in the nucleus of this element's atoms.

The first letter of a symbol is always a capital, but the second is lower case.

The atomic mass number is the average of all the atoms of the element. It is not a whole number because there are different isotopes (forms) of each element, each with a different number of neutrons.

Periods

Elements in the same period, or row, have the same number of electron shells in their atoms. So elements in period one have one electron shell, while those in period six have six electron shells.

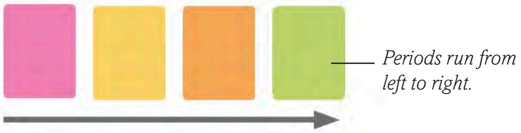

Periods run from left to right.

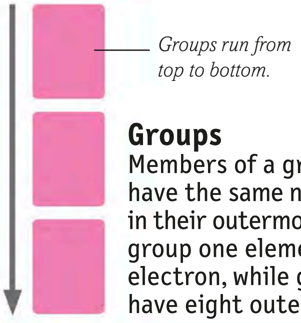

Groups run from top to bottom.

Groups

Members of a group, or column, all have the same number of electrons in their outermost shell. For example, group one elements have one outer electron, while group eight elements have eight outer electrons.

DMITRI MENDELEEV

The periodic table was developed by the Russian chemist Dmitri Mendeleev in 1869. Others had tried before, but his table was periodic, or repeating, because the characteristics of elements follow a pattern. The table was incomplete as some elements had not yet been discovered. However, Mendeleev predicted the positions of the missing elements, and was proved right when they were finally isolated many years later.

Reactions and uses

The elements can combine in different ways to make 10 million compounds, possibly more. As well as learning about the physical and chemical properties of elements, chemists also want to find out how and why certain elements react with each other to form compounds. Chemical reactions are happening all the time. During a reaction, substances change into new substances. The bonds that hold them are broken and then remade in a different combination.

Explosive reaction

In this chemical reaction, pure lithium reacts with air to make the compound lithium oxide. It takes energy to break the links between the lithium atoms and then make bonds with oxygen in the air. Reactions need energy to begin, but they often produce energy as heat and light.

1. This piece of pure lithium is placed on a surface and is exposed to the air.

2. A gas torch is used to heat the lithium, and in just a few seconds it turns red, which is a typical colour for this metal when it becomes hot.

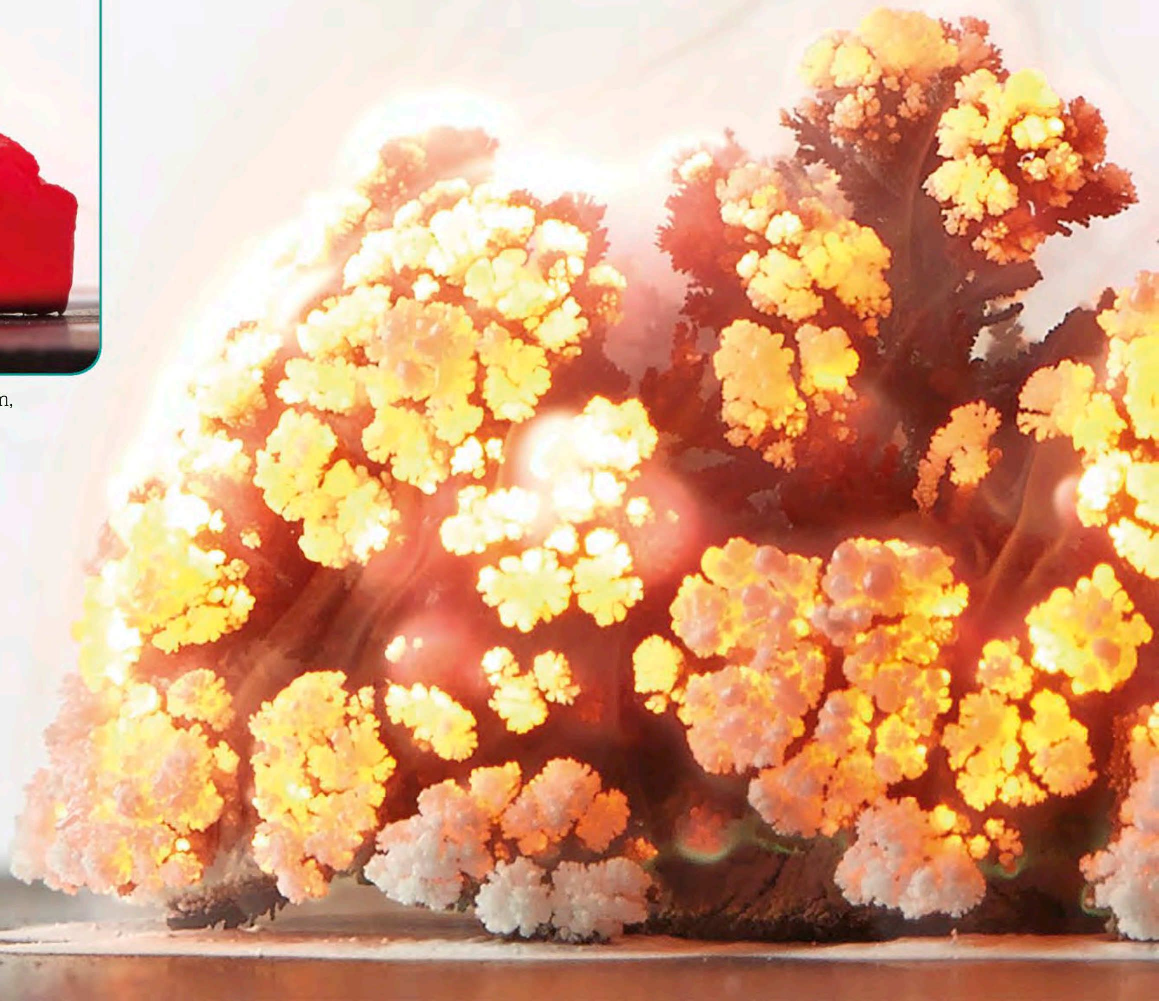

3. Very quickly, the lithium catches fire. The white areas forming here are the compound lithium oxide, which is a combintion of lithium and oxygen.

Mixtures

A mixture is a combination of substances that can be separated by physical means, such as filtering. It is not the same as a compound, where the ingredients are connected by bonds and can only be separated using a chemical reaction. Mixtures can be classified as solutions, colloids, and suspensions.

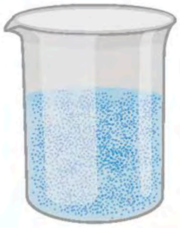

Solution
In this mixture, a substance is completely and evenly mixed, or dissolved, into another substance. Seawater is a solution.

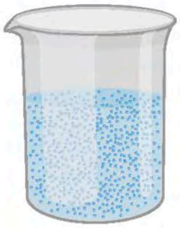

Colloid
This mixture contains unevenly spread particles and clusters that are too small to see. Milk is a colloid.

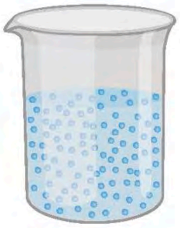

Suspension
This type of mixture consists of large particles of one substance floating in another substance. Muddy water is a suspension.

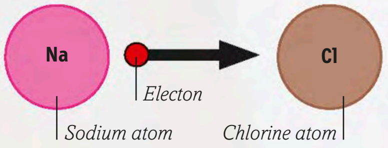

1. A sodium atom donates one electron to a chlorine atom. This gives both atoms full outer electron shells.

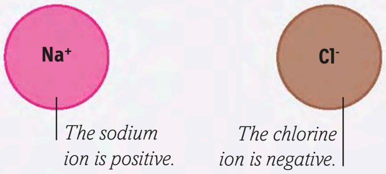

2. These are now charged atoms known as ions. The sodium ion has a positive charge and the chlorine ion has a negative charge.

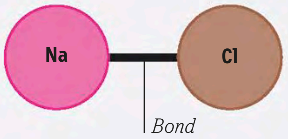

3. Sodium is attracted to – and forms a bond with – chlorine, forming a molecule of the compound sodium chloride.

Forming compounds

There are two kinds of bonds formed between elements during a chemical reaction. In an ionic bond, such as in sodium chloride (above), one atom gives away its electron(s) and another accepts them. This results in each having full outer electron shells. The other type is called covalent bonding. In this, atoms sit together and share their electrons so they both have full outer shells.

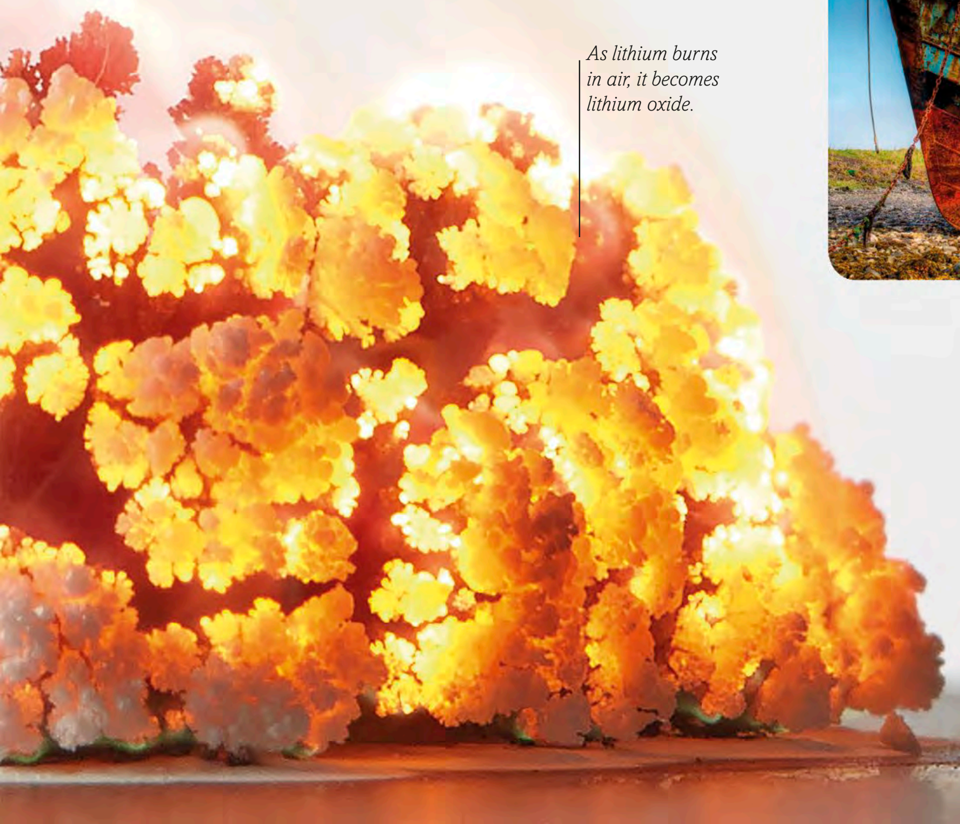

As lithium burns in air, it becomes lithium oxide.

Reactions in the real world

Chemical reactions happen all around us. There are reactions when we cook, take medication, or breathe. The image above shows a rusty iron ship. Over time, the element iron develops this red, flaky layer when it reacts with oxygen present in water or air to form the compound iron oxide – more commonly known as rust.

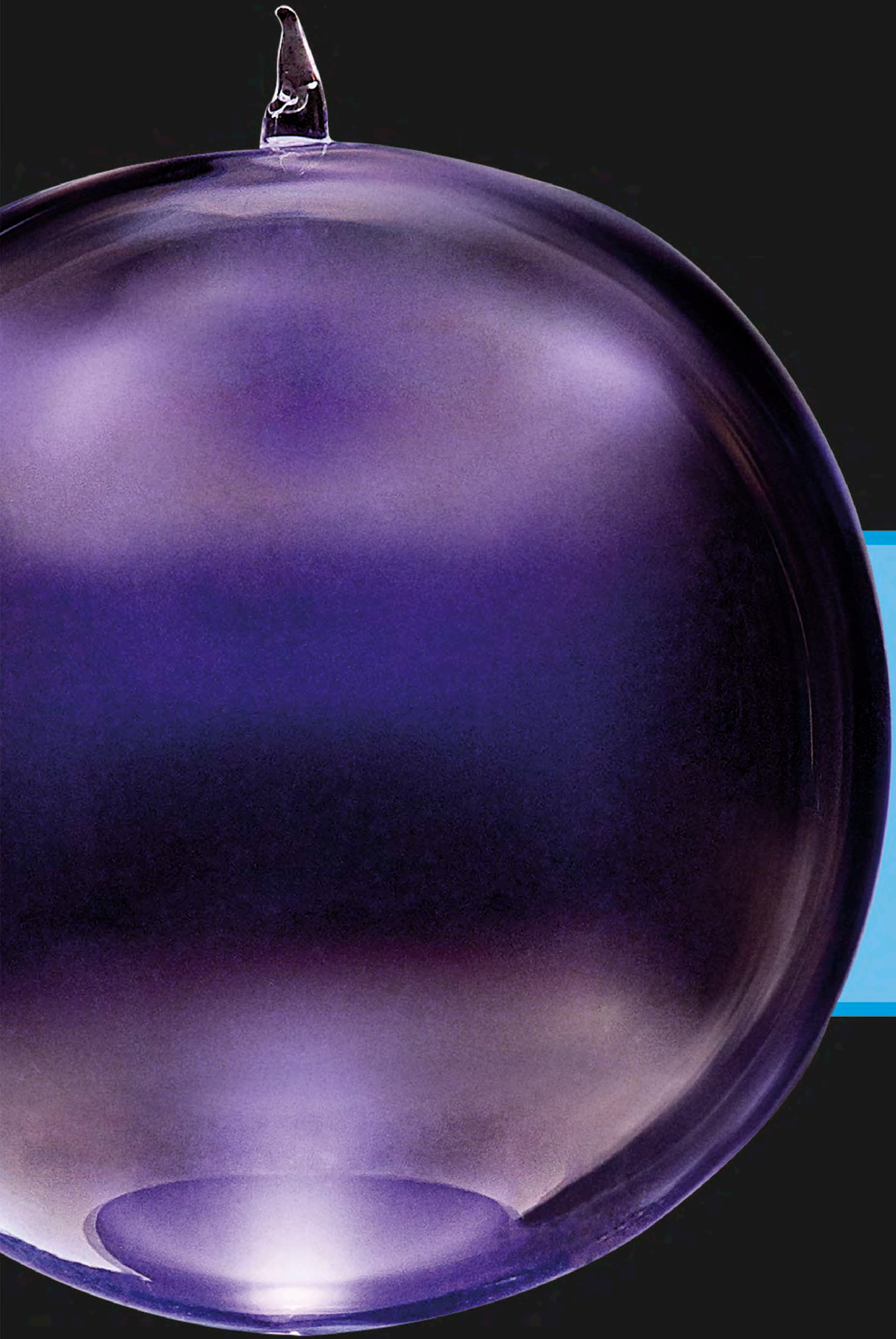

Pure hydrogen (H) fills this glass sphere, and produces a purple glow when electrified.

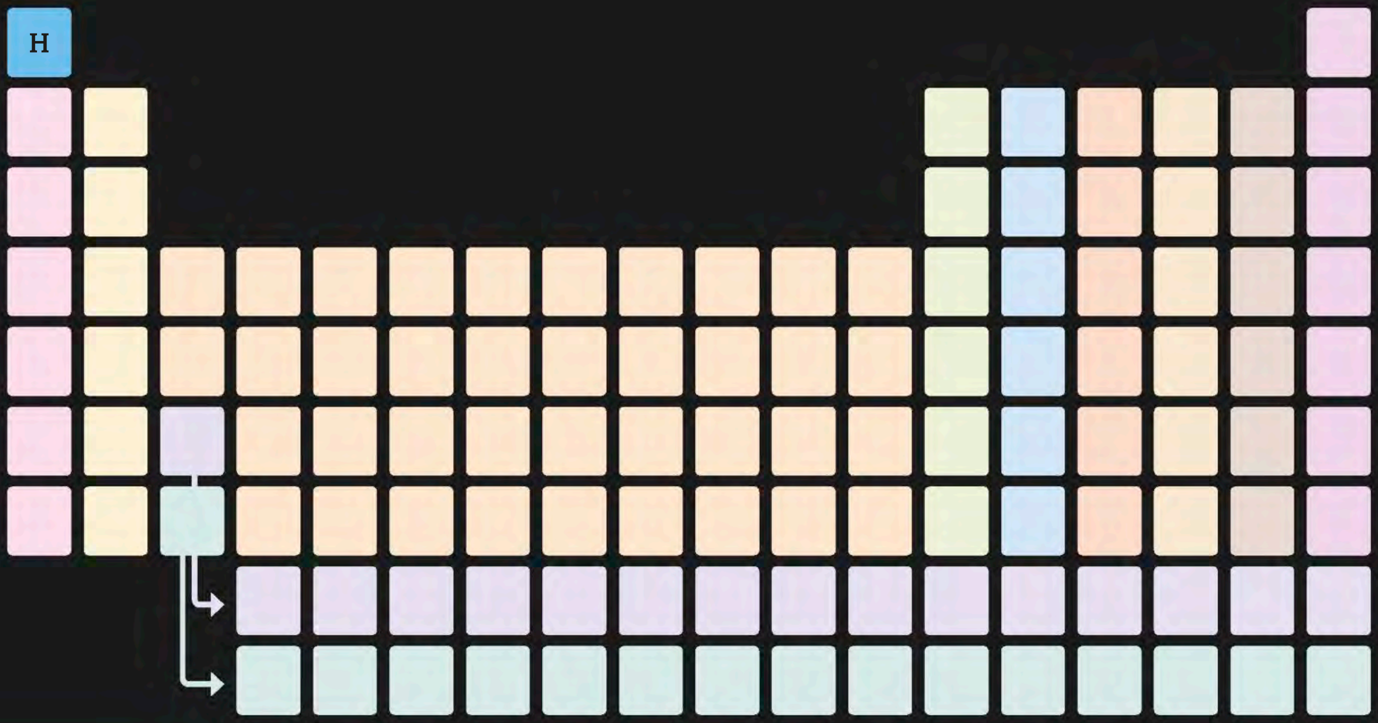

Hydrogen

The first element, hydrogen (H), is located above the alkali metals in the first column of the periodic table. However, because it is so different to the elements below it, hydrogen is not included in their group. This gas has the simplest atoms of any element with one electron and one proton. It is highly reactive and forms compounds with all kinds of other elements.

Atomic structure
A hydrogen (H) atom has one electron moving around a nucleus consisting of a single proton.

Physical properties
Hydrogen gas is the lightest material in the Universe. Pure hydrogen is rare on Earth, as it escapes quickly from the atmosphere into space.

Chemical properties
Hydrogen is highly flammable. It forms compounds with both metals and non-metals.

Compounds
The most common hydrogen compound is water. Acids are compounds that contain hydrogen.

1 H Hydrogen

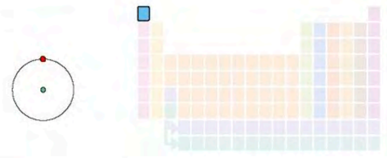

⊖ 1 ⊕ 1 ○ 0

State: Gas
Discovery: 1766

Forms

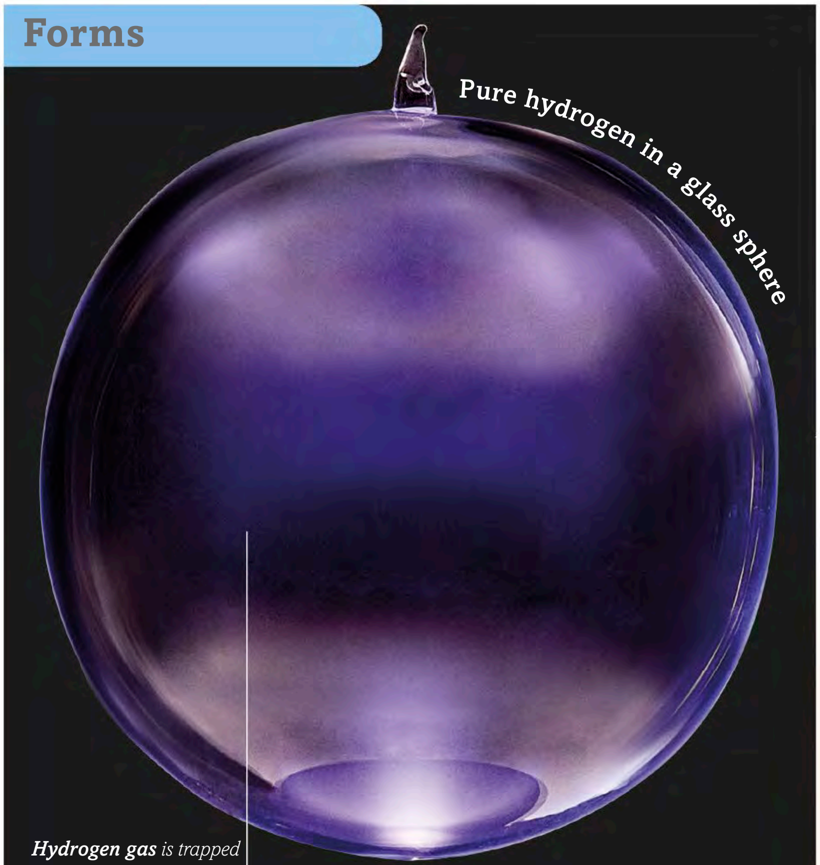

***Hydrogen gas** is trapped inside this glass sphere, and gives off a purple glow when electrified.*

The Orion Nebula

***This gaseous stellar nursery** is giving birth to thousands of stars.*

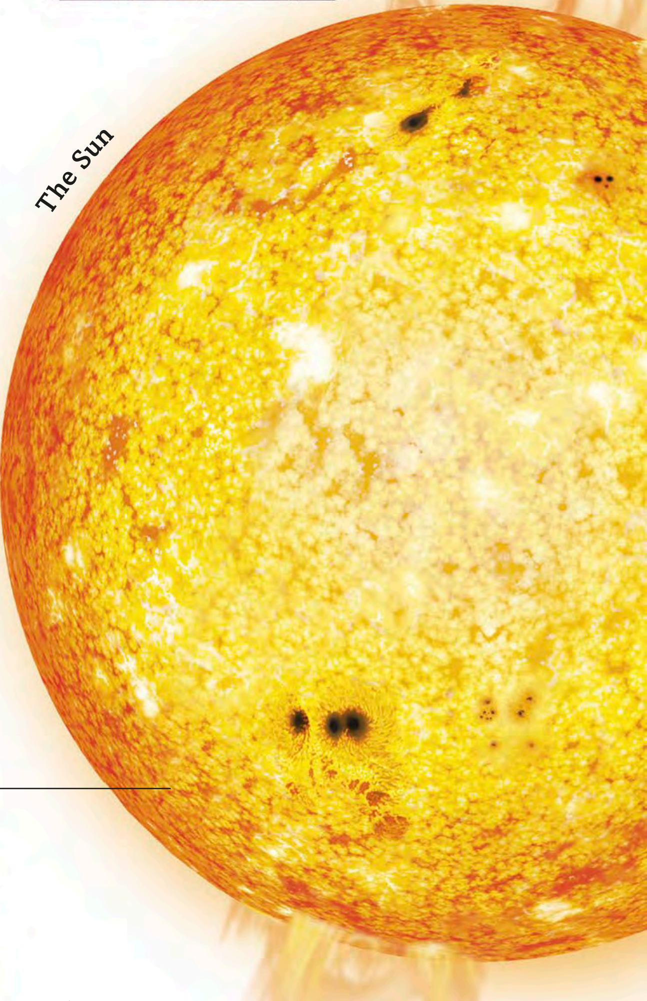

***The Sun** is four-fifths hydrogen.*

***Three quarters of this planet** is made up of layers of gaseous and liquid hydrogen.*

***Each water molecule** has two atoms of hydrogen and one of oxygen.*

Hydrogen is the first member of the periodic table because it has the simplest atoms of all elements: they contain just one proton and one electron. Pure hydrogen is a transparent gas. The biggest planets, such as **Jupiter**, are vast balls of hydrogen mixed with other gases, such as helium and methane. On Earth, hydrogen is commonly found in **water**. Although it is rare in Earth's atmosphere, hydrogen is the most common element in the Universe. Stars, such as the **Sun**, contain large amounts of hydrogen. At the centre of a star, atoms of

Uses

Hydrogen-filled balloon

***This balloon** can rise high into the atmosphere where sensors gather information about atmospheric pressure, temperature, and wind speed.*

HOW ROCKET FUEL WORKS

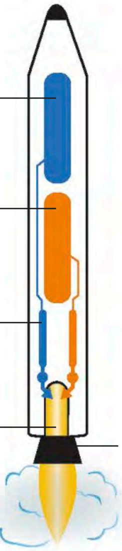

1. This chamber contains a fuel called liquid hydrogen.

2. This chamber contains liquid oxygen, which helps the hydrogen burn.

3. Pumps control the flow of the liquids as they enter the combustion chamber.

4. The combustion chamber is where the liquids mix together, creating an explosion.

5. The nozzle emits hot vapour, pushing the rocket upwards.

Many space rockets use liquid hydrogen as a fuel. The hydrogen reacts with oxygen to form extremely hot steam, which blasts out of the nozzle. This creates thrust, which pushes the rocket upwards.

Margarine

***Margarine** is made of vegetable oils thickened by adding hydrogen.*

Hydrogen peroxide

***This liquid** is used as a cleaner.*

Delta IV rocket

***This powerful rocket** uses 45,460 litres (12,000 gal) of liquid hydrogen as fuel.*

The only **waste product** of hydrogen fuel is **steam**.

Hydrogen bomb explosion

***This powerful explosion was** created by fusing hydrogen atoms.*

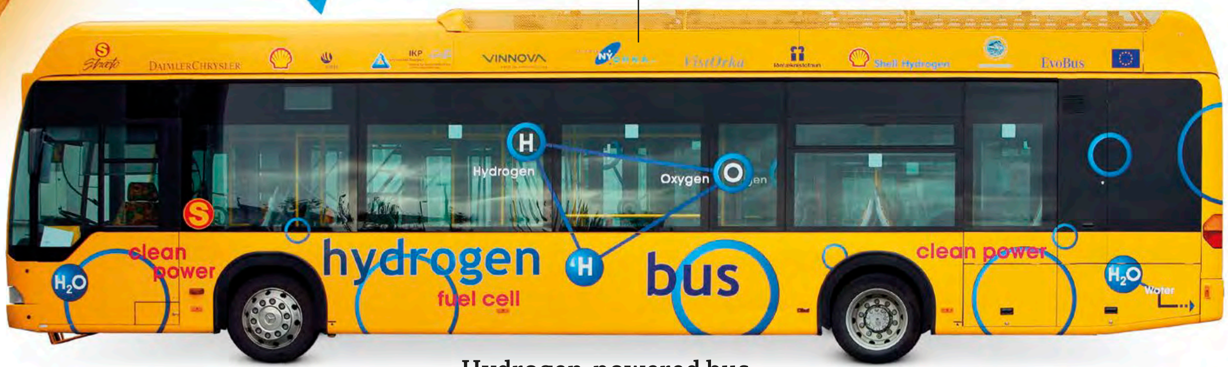

Hydrogen-powered bus

***This energy-efficient bus** runs on a fuel cell fed by hydrogen.*

this element are fused together, releasing heat and light. New stars form inside **nebulae** – such as the **Orion Nebula**. They are clouds of hydrogen gas that slowly collapse in on themselves. Hydrogen gas is the lightest element of all, and much lighter than air. This is why **hydrogen-filled balloons** can fly higher than air-filled ones. Supercold liquid hydrogen is used as **rocket** fuel. Atoms of hydrogen fuse together to produce a lot of energy in **hydrogen bomb** explosions. Pure hydrogen is also a clean energy source used to power some **buses** and cars.

Potassium (K) tarnishes when exposed to air.

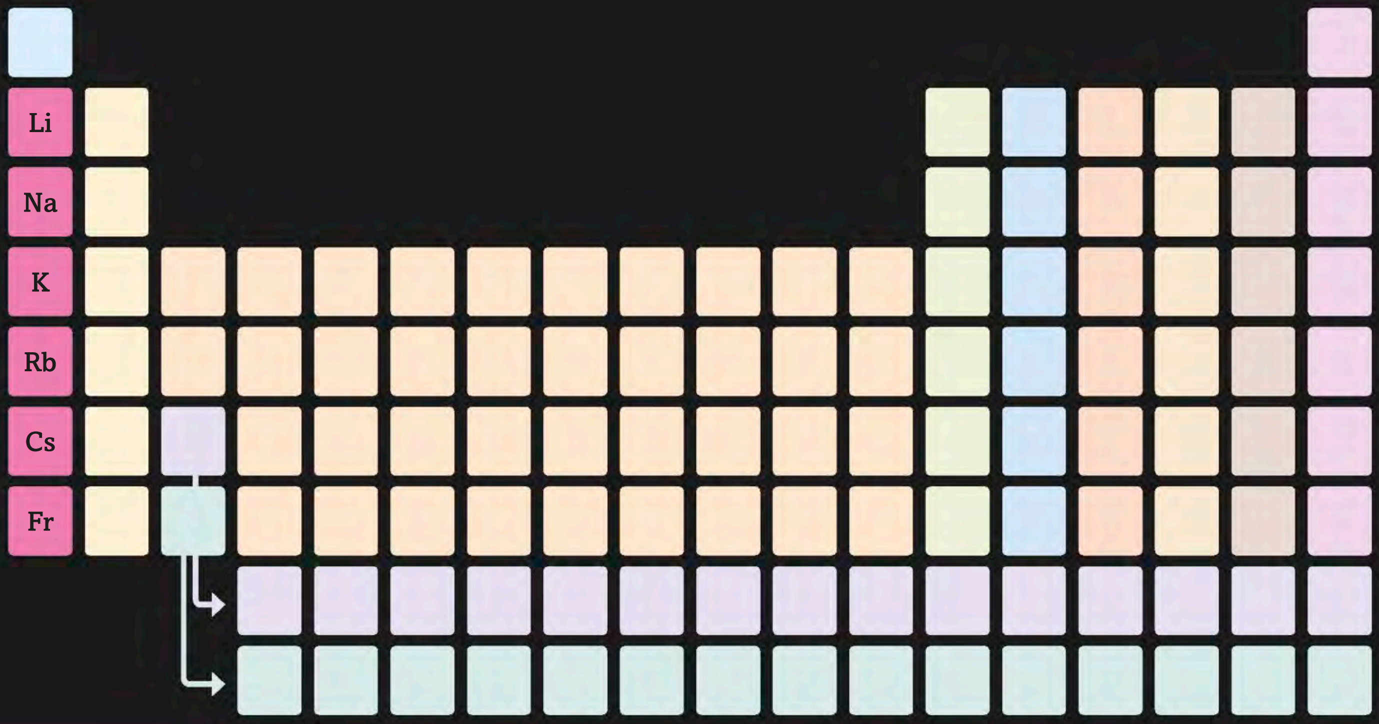

Alkali Metals

After hydrogen (H) – which is in a group of its own – the first column of the periodic table contains the alkali metals. This group gets its name from the way the elements react with water. These vigorous reactions always produce acid-attacking compounds called alkalis. None of the alkali metals are ever found in a pure form in nature. The first three metals are common in many minerals, while the last three are rarer.

Atomic structure
The atoms of all alkali metals have just one electron in their outer shell. Alkali metal atoms are among the biggest of all atoms.

Physical properties
These metals are soft enough to be cut with a knife. They are all silvery and very shiny when clean.

Chemical properties
Alkali metals are highly reactive. They form bonds with other elements, giving away their single outer electron.

Compounds
These metals react with water to form compounds called hydroxides. They react easily with halogens to form salts, such as sodium chloride.

3 Li Lithium

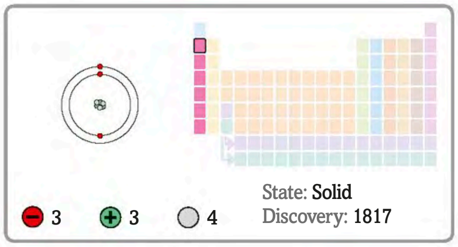

State: Solid
Discovery: 1817

Forms

Drinking water

This water *contains tiny amounts of dissolved lithium minerals.*

Oyster mushrooms

These mushrooms *absorb lithium from the soil.*

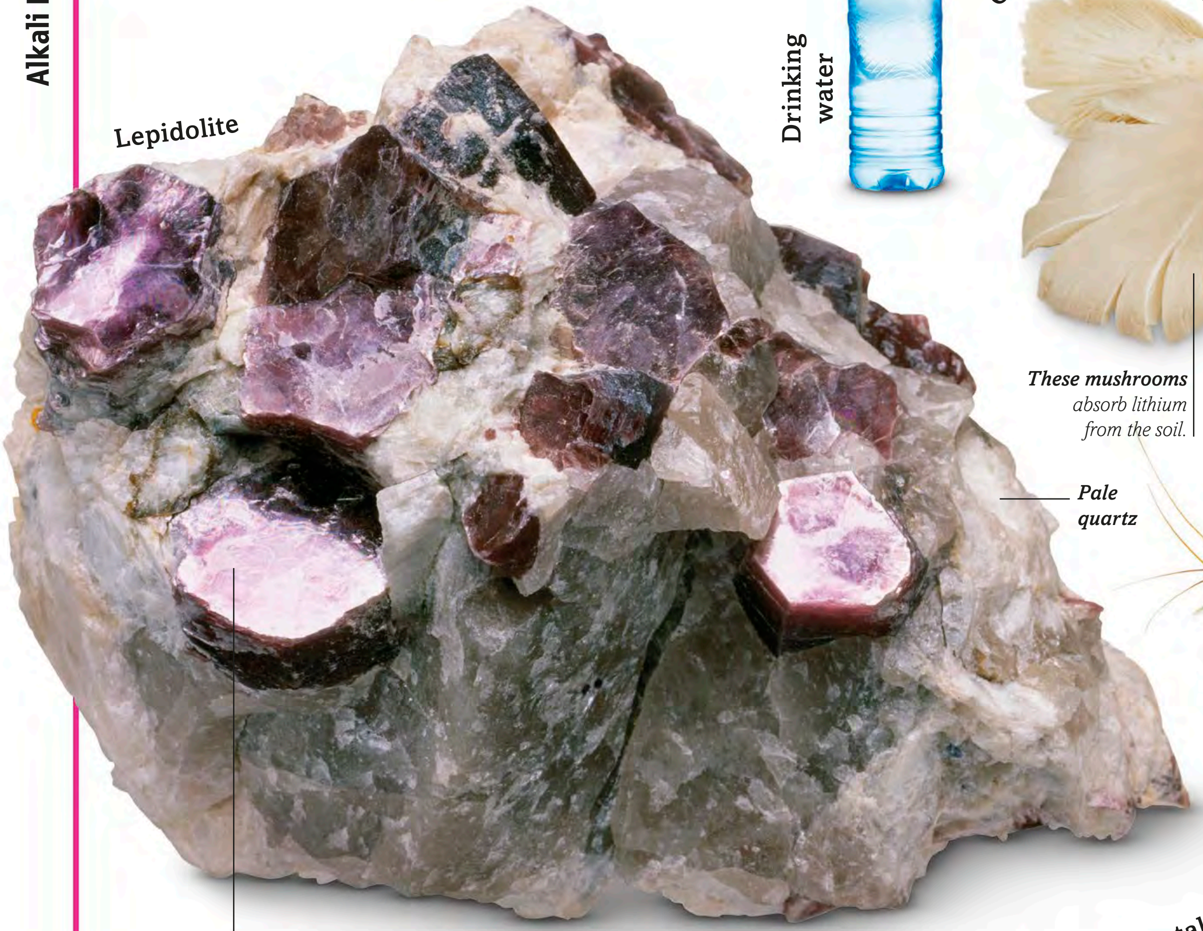

Lepidolite

Pale quartz

Purple crystals *containing lithium*

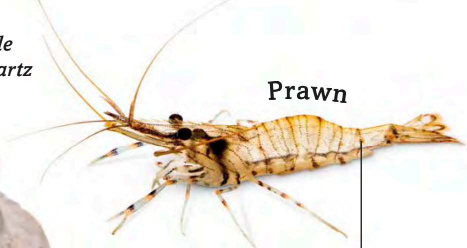

Prawn

Prawns *and other shellfish absorb lithium from seawater.*

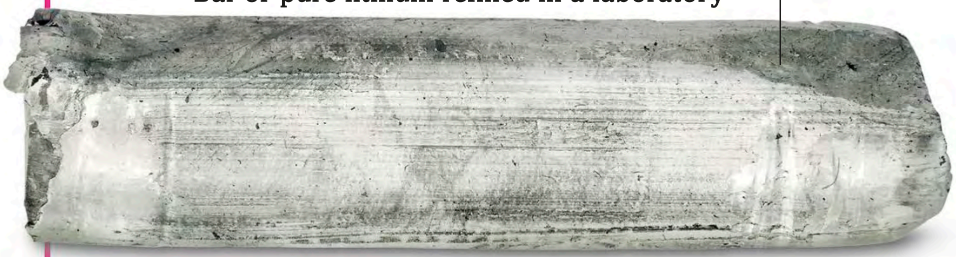

Bar of pure lithium refined in a laboratory

Shiny pure lithium *becomes dull when it is exposed to air.*

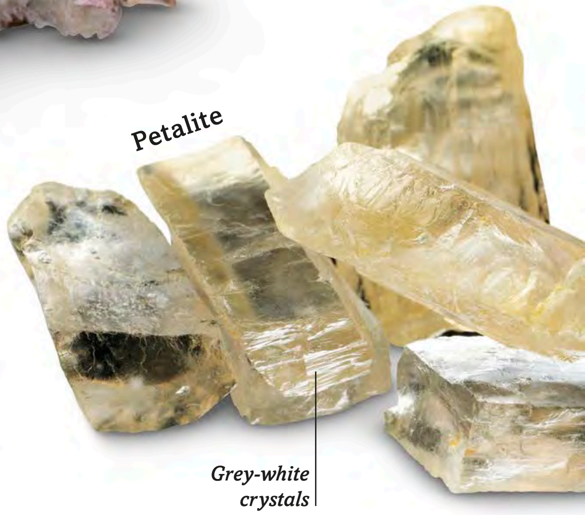

Petalite

Grey-white crystals

Lithium is the the lightest of all metals: in fact, it can easily float on water. **Pure lithium** is very reactive and exists in nature only in minerals, such as **lepidolite** and **petalite**. Many lithium minerals dissolve well in **water**, and the world's seawater contains millions of tonnes of dissolved lithium. Lithium is found in many foods, such as **mushrooms**, **prawns**, nuts, and seeds. It also has many everyday applications. Glass composed of lithium is resistant to heat and is used in scientific equipment, such as **mirrors inside**

Uses

Smartphones run on rechargeable batteries that use lithium to store electricity.

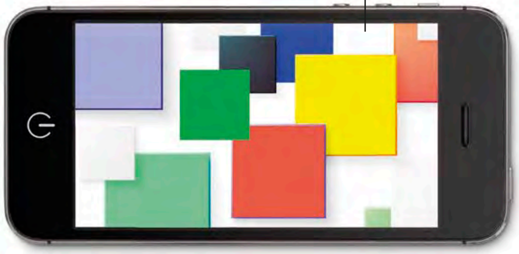

Smartphone

Hale telescope mirror

Lithium added to the glass in this mirror *stops the disc warping at extreme temperatures.*

LITHIUM-ION BATTERY

Lithium-ion batteries are widely used in digital devices. They store electrical energy to power gadgets and are rechargeable. This diagram shows a device's battery in use; when it is charging, this process is reversed.

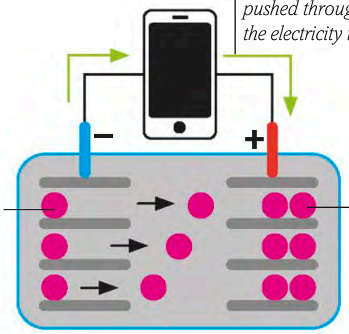

1. Inside the battery, positively charged lithium ions move from the negative electrode (-) to the positive electrode (+).

2. The positive electrode receives lithium ions as the battery loses charge.

3. As ions move inside the battery, negatively charged electrons are pushed through the phone, providing the electricity to make it work.

Syringe

Lithium coating on the inside of some syringes *delays the clotting of the blood sample.*

Lithium-rich grease *is used to keep mechanical parts of engines running smoothly, even when hot.*

TF2
LUBRICANT
Lithium Grease
40gms ℮

Grease

Some artificial teeth *contain lithium disilicate, which makes them strong.*

Artificial teeth

This car runs for at least **64 km** (40 miles) on one charge of its lithium-ion battery.

This charging point *can recharge an electric car in one hour.*

Electric car

This air scrubber *used lithium hydroxide to purify the air inside the Apollo 13 spacecraft.*

Air scrubber

telescopes. The main use for lithium is in rechargeable batteries. Lithium-ion batteries are small but powerful, so they are ideal for **smartphones** and tablet computers. Larger lithium batteries can power **electric cars**, which are less polluting than petrol-powered ones. A soapy compound called lithium stearate is used to make **grease**, which helps automobile engines run smoothly. This element also forms hard ceramics that are used to produce strong **artificial teeth**. Lithium compounds are used in some medicines as well.

11 Na Sodium

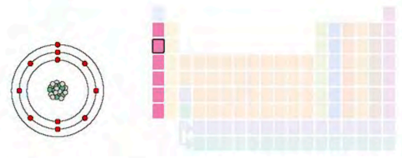

− 11 + 11 ○ 12

State: Solid
Discovery: 1807

Forms

Clinoptilolite

This sodium-rich mineral *is an example of a zeolite, a mineral with tiny holes that can trap gases.*

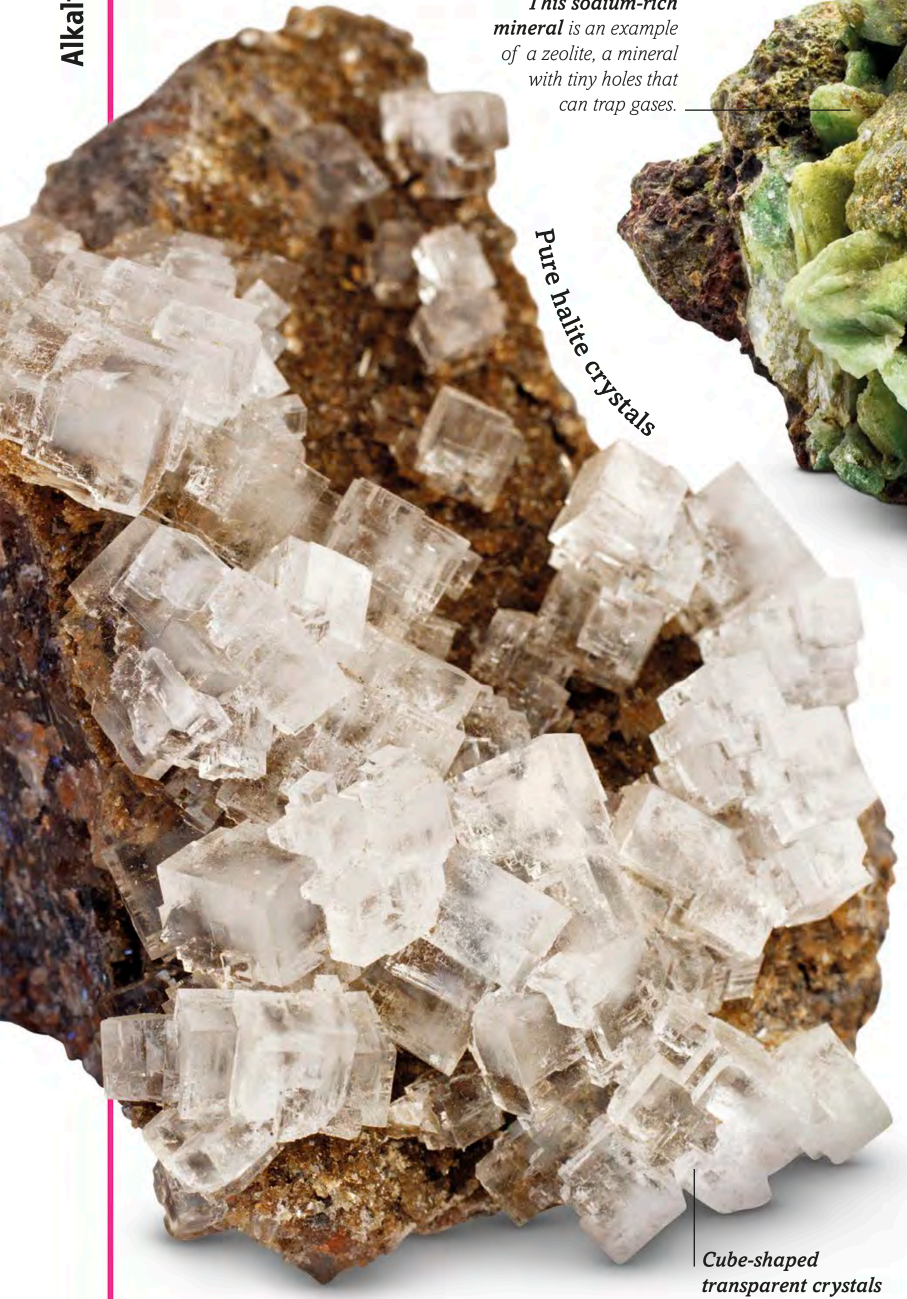

Pure halite crystals

Cube-shaped transparent crystals

Laboratory sample of pure sodium in an airless vial

Soft, shiny metal

This glass case *holding pure sodium has no air in it, to prevent the metal from reacting with oxygen in air.*

Sodalite cabochon

Polished gemstone *made of the mineral sodalite*

The thick, white crust *covering this salt flat contains sodium chloride and other salts.*

Salar de Uyuni, Bolivia

Everyday salt contains lots of sodium. Although abundant on Earth, sodium is never found in its pure form naturally: it forms compounds with other elements. Sodium chloride, which also contains chlorine, is the most common sodium compound. It is also known as the mineral **halite**, and it is what makes seawater salty. Other sodium minerals include **sodalite**, a soft blue stone that can be shaped and polished. **Pure sodium** is soft enough to be cut with a knife. It reacts with oxygen in the air, forming a compound called sodium oxide, and bursts

Uses

Common salt

Edible salt *is made by refining the mineral halite.*

Mummy

This mummified body*, or mummy, was preserved using sodium compounds.*

MUMMIFICATION

Ancient Egyptians believed in life after death and so preserved the bodies of their dead. A dead body was washed and the organs removed, then crystals of sodium compounds were used to dry it out. Finally, the body was wrapped, which completed the process of mummification.

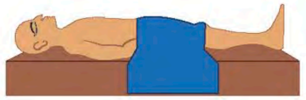

1*. Organs, such as the stomach and lungs, were removed from the dead body.*

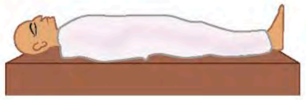

2*. Sodium compounds were spread over the body to dry it.*

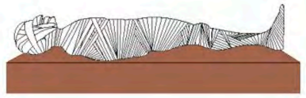

3*. The body was wrapped in cloth to mummify it.*

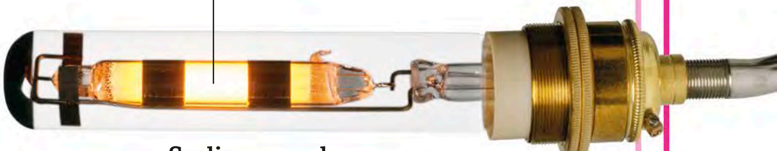

Sodium gas lamp

This tube *glows bright yellow-orange when sodium gas is electrified.*

Sodium fireworks

Bright yellow lights in fireworks *get their colour from burning sodium compounds.*

Cats were sacred in ancient Egypt, so their bodies were mummified.

Bar of soap

Some soaps *contain sodium hydroxide.*

Baking soda

Odourless white powder

Indigo dye powder

Indigo dyes *– often used in blue jeans – contain sodium.*

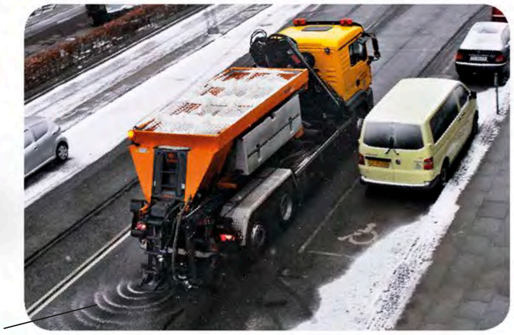

De-icing

Spraying salt *keeps roads free from ice and frost.*

into flames when in contact with water. Sodium compounds in **fireworks** burn with a yellow-orange colour. In ancient Egypt, crystals of sodium compounds were used to preserve dead bodies as **mummies**. Another useful compound is sodium bicarbonate, or **baking soda**, which makes dough rise by releasing bubbles of carbon dioxide. When refined, sodium chloride, or **common salt**, has several uses. It makes ice melt so it is used in salty grit added to slippery, frozen roads. This helps **de-ice** them to make them safer. It is also an important seasoning for meals.

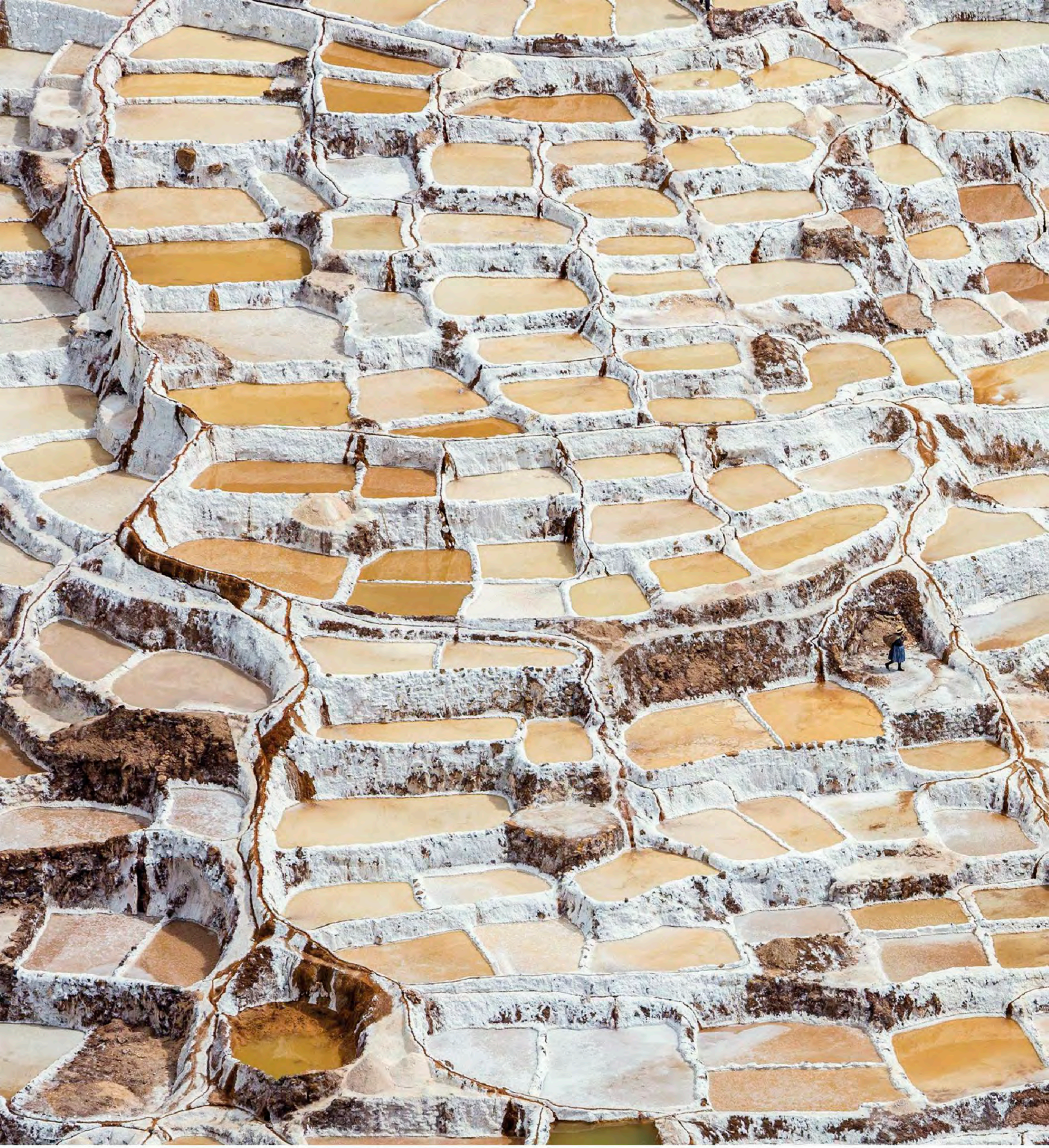

SALT FLATS Hundreds of artificial ponds dot the hillside near the small town of Maras, high in the Andes of Peru. The ponds fill with water from a stream that runs down from the nearby mountains. In the sunshine, the water evaporates, leaving behind a thick salt crust that can be collected. The people of Maras have been gathering salt in this way for at least 500 years.

The salt forms part of rocks deep underground before it is dissolved by the stream and flows into the pools. Evaporation can also be used to collect salt from seawater or other salty water sources (known as brines). Today, however, most of the world's salt comes from underground mines containing thick layers of salt that are a result of ancient seas drying out. Over millions of years, that dry salt has become buried under dense layers of rocks. This so-called "rock salt" is sometimes unearthed using excavators. At other mines, it is washed out by piping in warm water, which dissolves the salt. The brine is then pumped up to the surface for evaporation.

19
K

Potassium

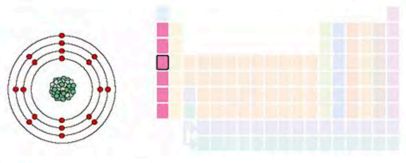

⊖ 19 ⊕ 19 ○ 20

State: **Solid**
Discovery: **1807**

Forms

This mineral *is rich in potassium chloride.*

Potash

Laboratory sample of pure potassium in an airless vial

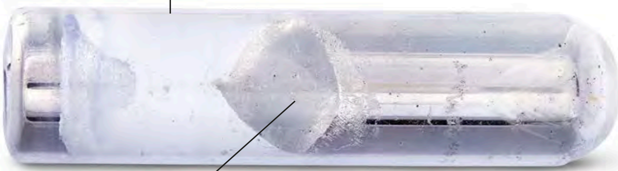

This glass case holding pure potassium *has no air in it, preventing the metal from reacting with oxygen in air.*

Soft and shiny solid

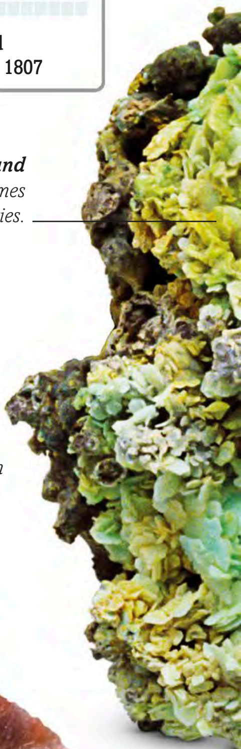

The yellow and green colour *comes from impurities.*

Sylvite

This mineral *contains potassium chloride, which gives it a salty taste.*

Potassium was first found in the dust of burnt plants. It was discovered by Sir Humphry Davy when he experimented with **potash** – a mixture of substances made from the ash of burnt plants soaked in water. The name potassium comes from potash but the element's chemical symbol, K, is taken from *kalium*, a Latin word for "ash". Potassium is never found pure in nature, but is present in minerals such as **aphthitalite** and **sylvite**. Potassium is vital for the human body, helping muscles and nerves work properly. For this, we rely on

Aphthitalite

Potassium-rich food

Banana
Avocado
Sweet potato

Uses

Soda water

Soda water *contains potassium compounds for added flavour.*

Potassium salt

***This salt** contains potassium chloride, which helps lower blood pressure.*

Saline drip

***Potassium solutions** are used to hydrate patients.*

Rebreather

***This cylinder** contains a compound called potassium superoxide.*

Fertilizer

***Potassium-rich fertilizer** is easily absorbed by the soil and boosts plant growth.*

Liquid soap

***This soap** contains potassium hydroxide, which is a cleaning agent.*

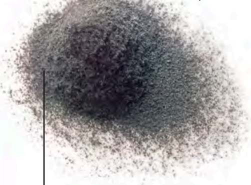

Gunpowder

***This explosive mixture** contains powdered potassium nitrate.*

Toughened glass screen

***This strengthened glass sheet** contains potassium nitrate.*

REBREATHER

A rebreather is a machine used by expert divers so they can stay underwater for long periods.

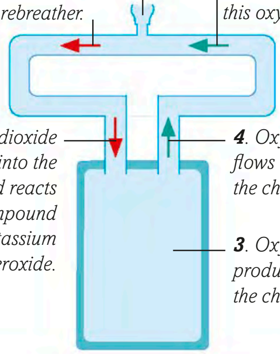

Mouthpiece

***1**. Exhaled air, containing carbon dioxide, enters the rebreather.*

***2**. Carbon dioxide flows into the chamber and reacts with a compound called potassium superoxide.*

***3**. Oxygen is produced in the chamber.*

***4**. Oxygen flows out of the chamber.*

***5**. The diver breathes in this oxygen.*

potassium-rich food, such as bananas, root vegetables, and avocados, which contain potassium chloride. In tiny amounts, this compound can enhance flavours, as it does in **soda water**. It is also a healthy alternative to sodium chloride, or common salt, and an important ingredient in **saline drips** for rehydrating patients who are seriously ill. Potassium nitrate is a compound of potassium, oxygen, and nitrogen, and is found in **gunpowder** and **toughened glass screens** for mobile phones.

37

Rb

Rubidium

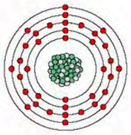

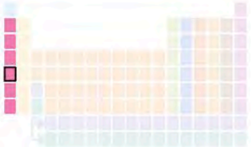

● 37 ⊕ 37 ○ 48

State: Solid
Discovery: 1861

Forms

Laboratory sample of pure rubidium in an airless vial

This glass case contains ***pure rubidium****, preventing it from coming into contact with air and catching fire.*

Leucite

Pale, waxy mineral

Lepidolite

This soft mineral *contains up to 3.5% rubidium.*

Rubidium makes up only about **1 per cent** of this mineral.

Pollucite

This ore *contains caesium and rubidium.*

Rubidium was named after the Latin word *rubidius*, meaning "deepest red". This refers to the red-coloured flame it produces when burned. This highly reactive element ignites on contact with air. On contact with water, it reacts vigorously, producing hydrogen gas and a lot of heat. Rubidium is not often concentrated in particular minerals, but instead is spread in small amounts through a wide range of minerals, such as **leucite** and **pollucite**. The pure metal is sourced mainly from the mineral **lepidolite**. Another mineral called rubicline has even more rubidium in it but is very rare. Rubidium atoms are sensitive to light and can be used in photoelectric cells (devices that convert light energy into electricity) and **night-vision equipment**. This element has radioactive forms, which can be used to measure the age of rocks. When injected into a patient's body, rubidium targets tumours, which show up clearly on **PET (positron emission tomography) scans**. Rubidium is also used by light-sensitive electronics called **photomultipliers**, and in making **insulators** for high-voltage cables and some special types of glass.

Uses

These lenses *contain rubidium, which aids night vision.*

Night-vision goggles

Photomultiplier

This sensitive device *detects light by using a rubidium compound.*

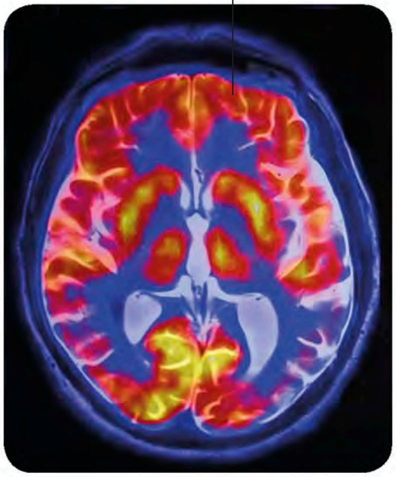

The structure of the brain *can be seen clearly because of the use of radioactive rubidium.*

PET scan

Fireworks

This purple colour *comes from burning a nitrogen rubidium compound.*

Ceramic insulator

Electricity cables *are hung from these rubidium-rich insulators.*

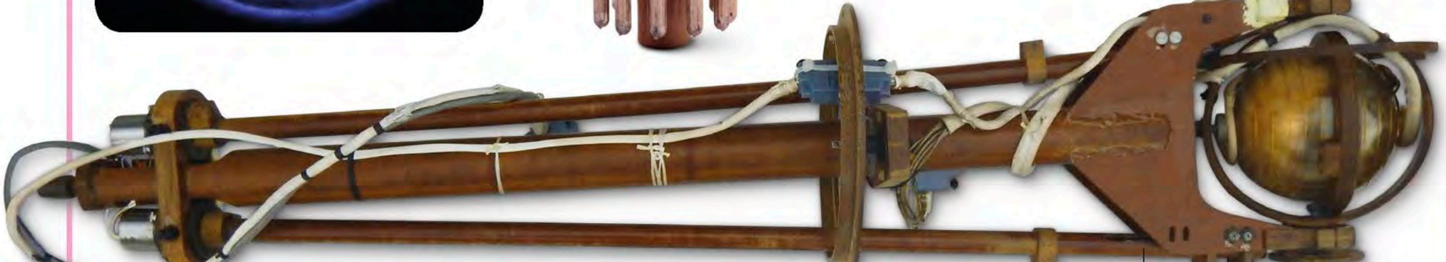

Magnetometer

This device *from the early 20th century used rubidium to measure the strength of magnetic fields.*

RUBIDIUM-STRONTIUM DATING

About a quarter of all rubidium atoms are radioactive. Slowly over time, they break down into strontium atoms. Comparing the amounts of these elements in a rock shows when that mineral was formed. Older rocks have less rubidium and more strontium in them.

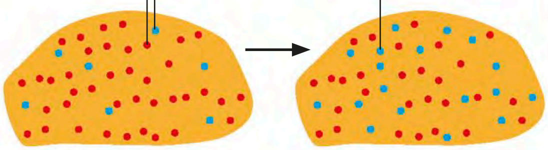

Rubidium-87 atoms (red) decay at a predictable rate.

Only small amounts of strontium-87 (blue) in the rock.

The amount of strontium-87 has increased over time.

Millions of years ago

Present day

55

Cs Caesium

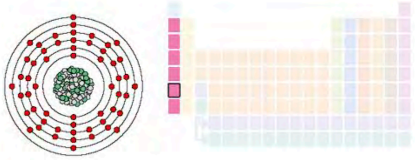

⊖ 55 ⊕ 55 ○ 78

State: Solid
Discovery: 1860

Forms

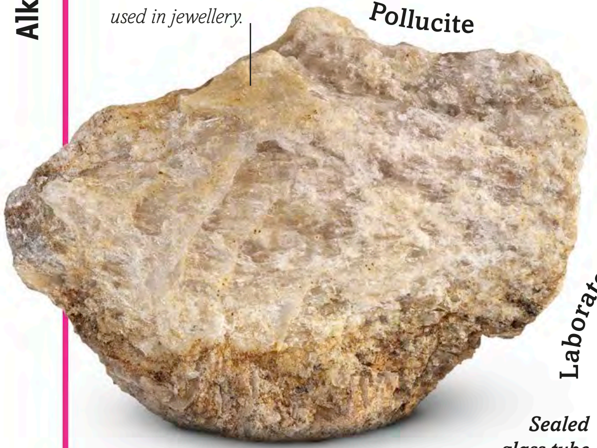

Pollucite

The crystals of this mineral *are used in jewellery.*

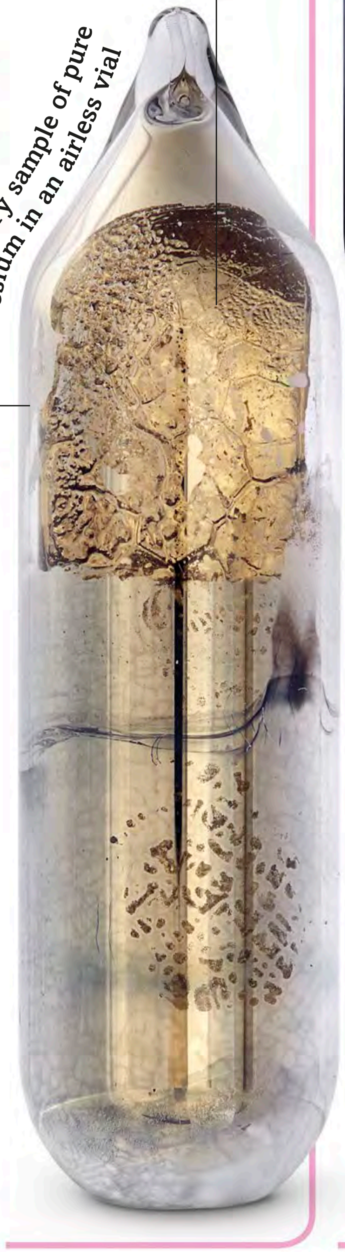

Laboratory sample of pure caesium in an airless vial

Shiny, silver-gold metal

Sealed glass tube

KIRCHHOFF AND BUNSEN

Caesium was discovered in 1860 by German scientists Robert Bunsen and Gustav Kirchhoff. They burned a sample of mineral water on a burner, which split the flame's light into individual colours. One of them was a distinctive light blue, which came from caesium.

Gustav Kirchhoff (left) and Robert Bunsen (right)

Uses

Atomic clock

This highly accurate clock *is also called a caesium clock.*

Drilling fluid

High-density caesium compounds *in this fluid stop toxic gases rising to the surface.*

As the most reactive metal on Earth, caesium explodes into flames if in contact with air or water. Therefore, **pure caesium**, is stored in a sealed glass tube from which all the air has been sucked out. This element is rare, and most of it is extracted from the mineral **pollucite**. Its name means "sky blue" and refers to the colour of caesium's flame when burning. Caesium is used in **atomic clocks**, which measure time down to a billionth of a second. These clocks are so accurate that they would gain or lose no more than one second every 300 years.

87
Fr
Francium

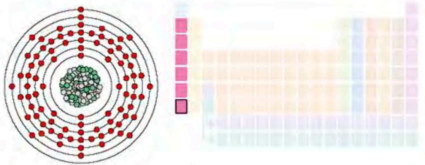

⊖ 87 ⊕ 87 ○ 136

State: **Solid**
Discovery: **1939**

MARGUERITE PEREY

The French chemist Marguerite Perey discovered francium in 1939 while studying the way a pure sample of another radioactive metal – actinium – decayed. She found that actinium broke down to form thorium and a previously unknown element. She named this element francium after her home country.

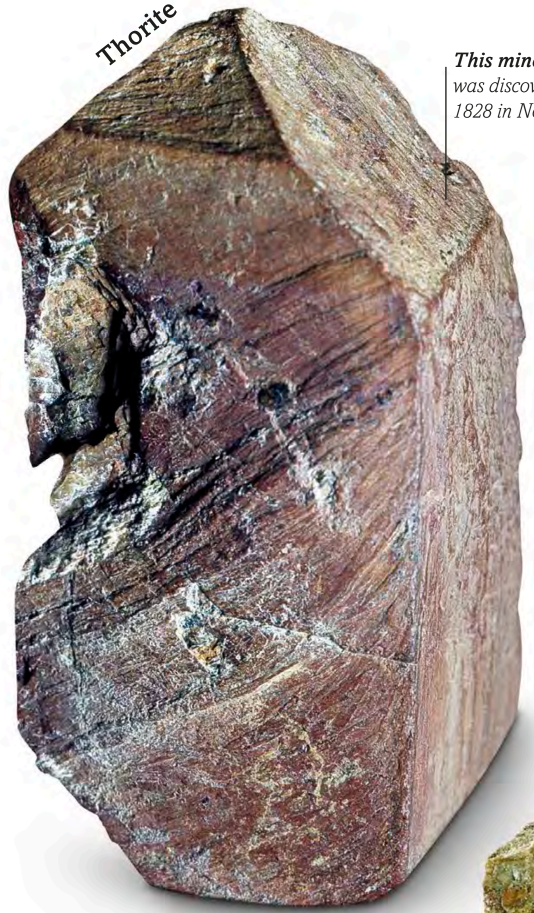

Thorite

This mineral *was discovered in 1828 in Norway.*

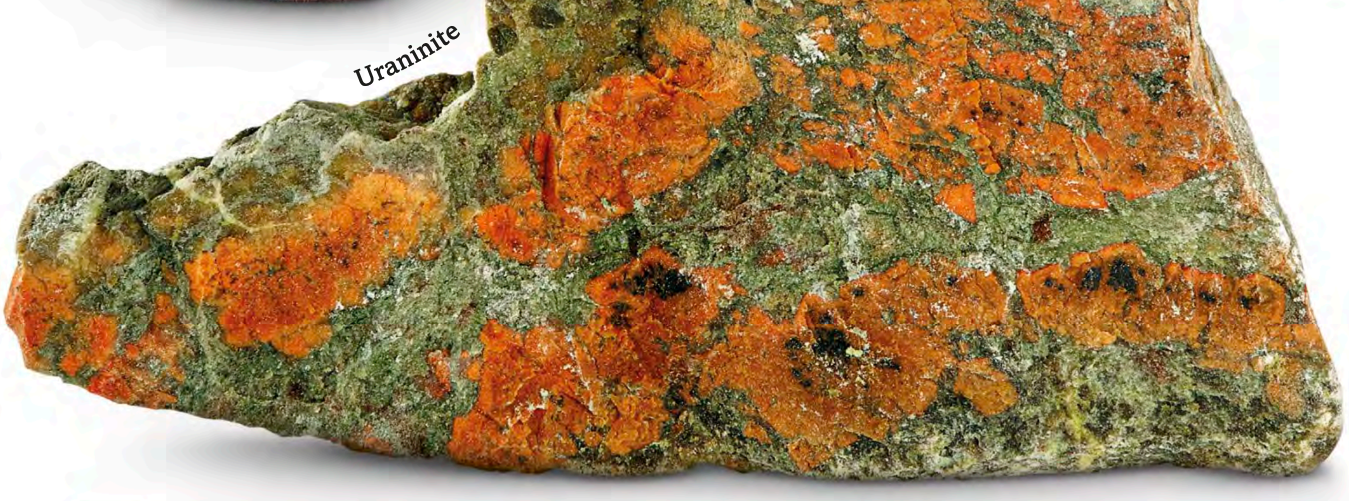

Uraninite

The dark crust *is a uranium mineral that holds tiny amounts of francium.*

Earth's rocks have **one** francium **atom** for every million trillion uranium atoms.

Francium is the rarest natural element on Earth. Scientists think there may be just 30 g (1.1 oz) of francium in Earth's rocks. Francium atoms are created when radioactive elements break down. Francium can be extracted from radioactive ores such as **thorite** and **uraninite**, both of which contain tiny amounts of this element. Even so, to date the largest sample of the metal made contained only 300,000 atoms, and lasted only a few days. Francium has no known uses outside of research.

Barium (Ba) crystals turn black in air.

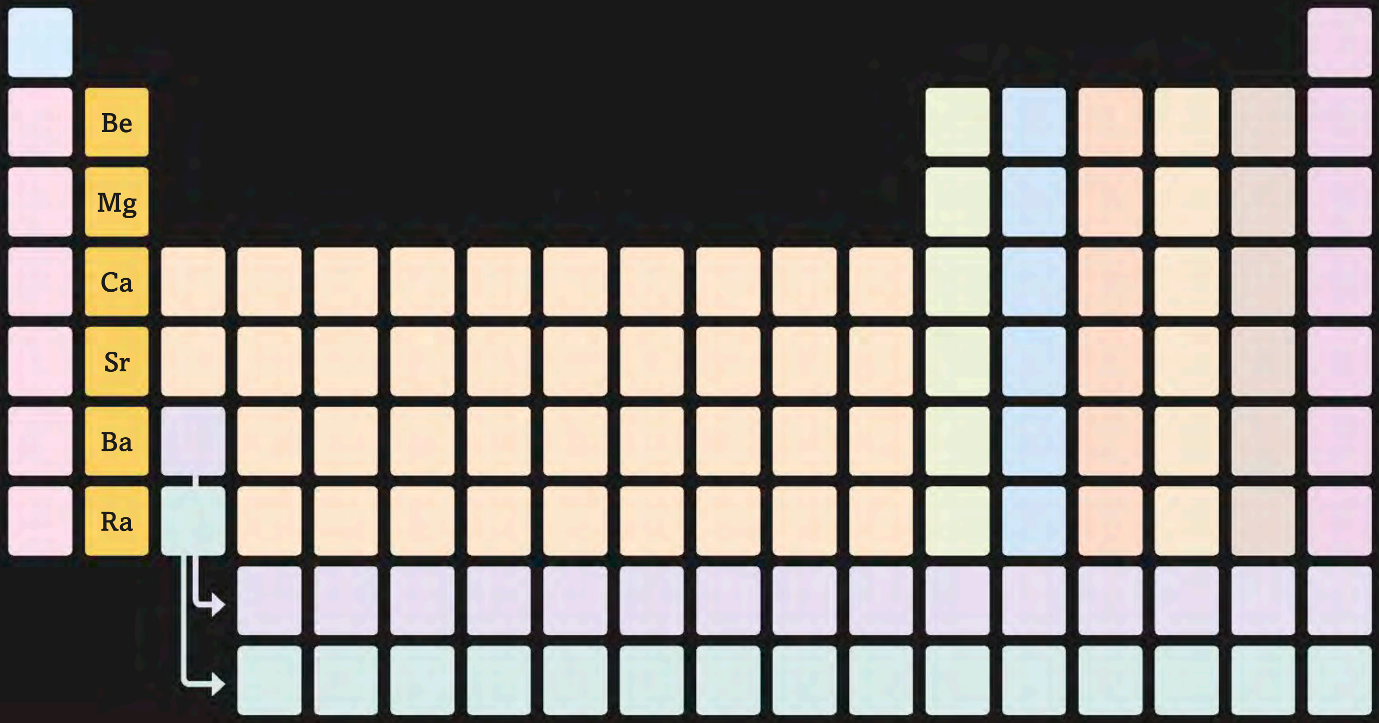

Alkaline Earth Metals

This group is a collection of reactive metals that were discovered as compounds inside common minerals in Earth's crust. Most of these minerals – known in the past as "earths" – are alkaline (alkali-producing), and this is how the group got its name. All alkaline earth metals were first purified in the 19th century.

Atomic structure
The alkaline earth metals have two electrons in their outermost electron shell. Radium (Ra) is the most radioactive member.

Physical properties
All members of this group are soft and shiny when pure. They are solid at room temperature.

Chemical properties
These metals are similar to the alkali metals, but not as reactive. Except for beryllium (Be), all alkaline earth metals react with hot water or steam.

Compounds
These elements form compounds with non-metals by losing their outermost electrons. Several compounds are found in teeth and bones.

4
Be

Beryllium

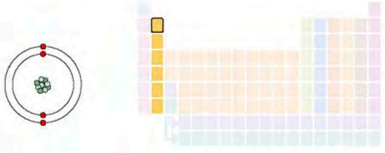

− 4 + 4 ○ 5

State: Solid
Discovery: 1797

Forms

This mineral can also be brown, green, or orange.

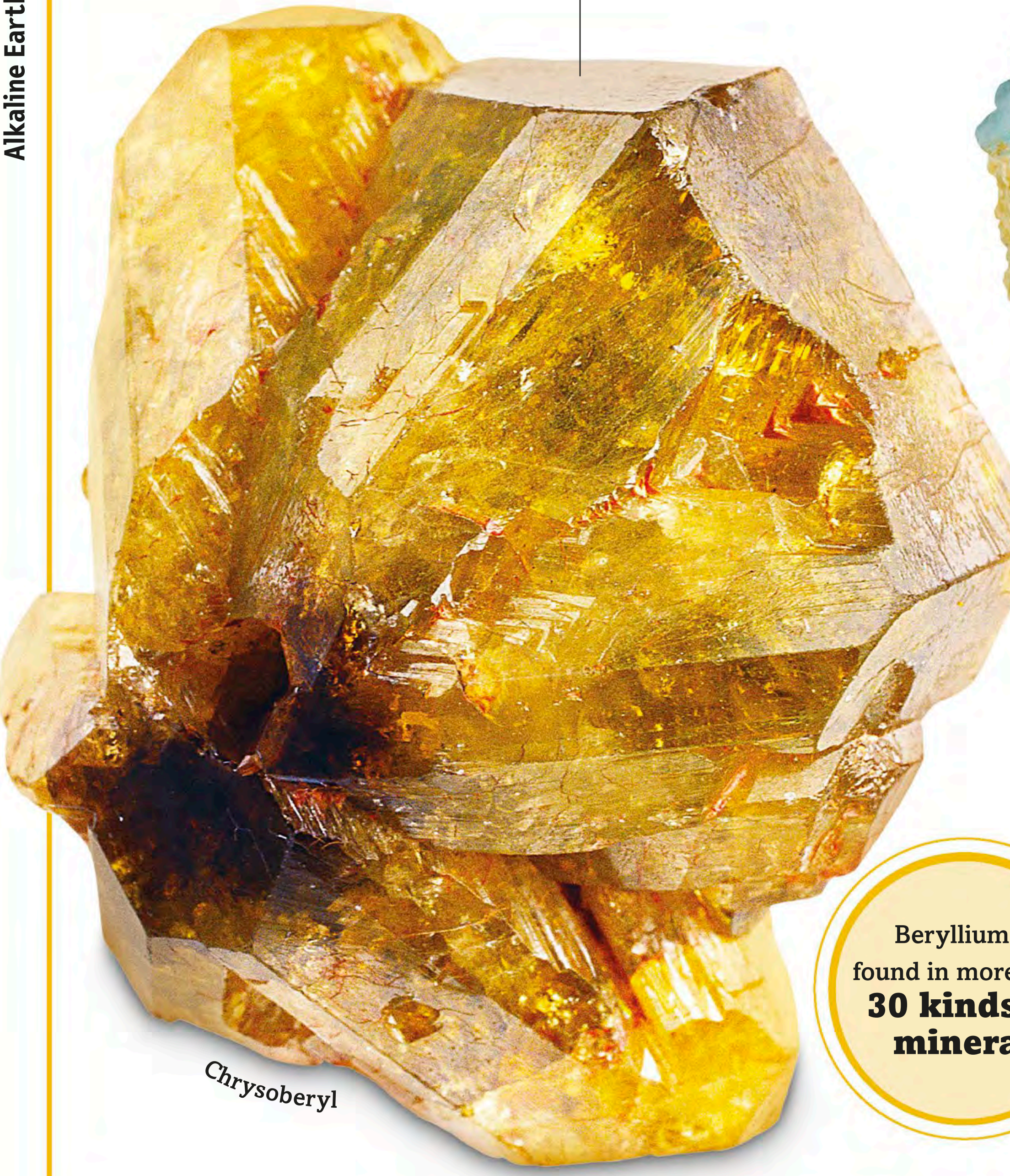

Chrysoberyl

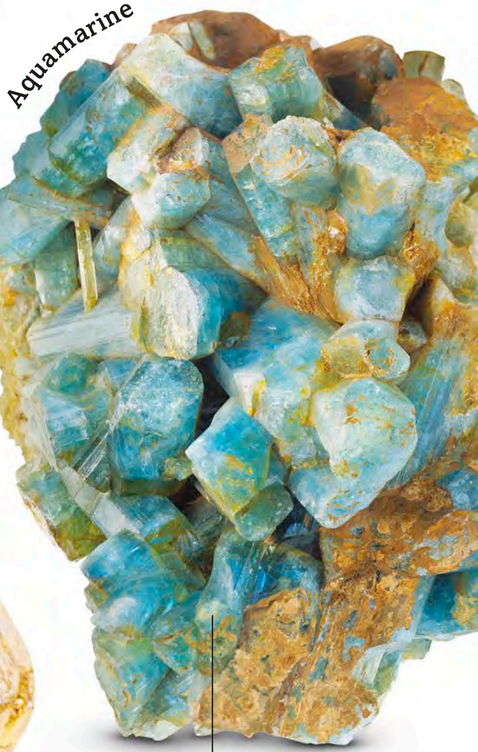

Aquamarine

***These crystals** have this pale blue colour due to iron impurities.*

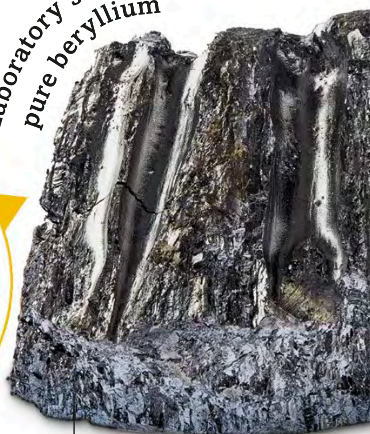

Laboratory sample of pure beryllium

***Lightweight** metal*

Beryllium is found in more than **30 kinds of mineral**.

This widely used element gets its name from the Greek word *beryllos*, after which the mineral beryl is also named. Beryllium is the lightest of the alkaline earth metals, but it does not share many of the group's properties. For example, it does not react with water and is much harder than the other metals in its group. Two common beryllium minerals are **chrysoberyl** and beryl. Beryl has different forms, such as **aquamarine** and emerald. Beryllium is useful in many ways. For example, some military **helicopters** use windows made

Uses

Apache attack helicopter

Beryllium alloy window

LOUIS NICOLAS VAUQUELIN

Beryllium was discovered in 1798 by the French chemist Louis Nicolas Vauquelin. He extracted the pure metal from emerald, which is a valuable green form of the mineral beryl. He had already discovered the element chromium, which is also found in emerald and gives it its green colour.

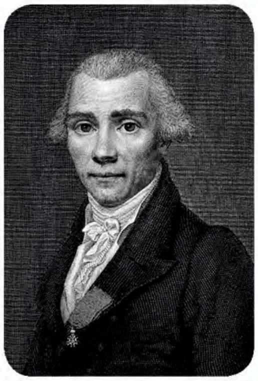

Fire sprinkler

This seal *made of a beryllium-nickel alloy is strong enough to stop high-pressure water supply leaking through.*

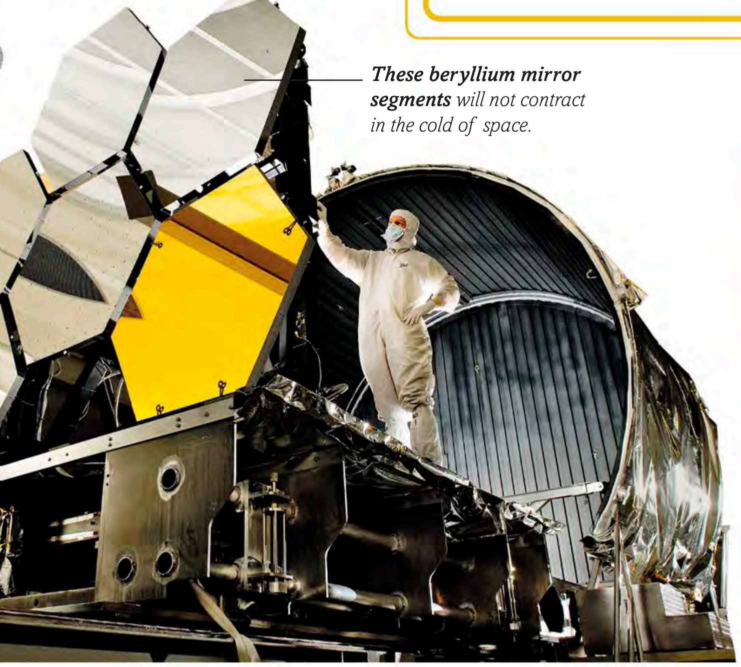

These beryllium mirror segments *will not contract in the cold of space.*

James Webb Space Telescope

This beryllium pipe *delivers beams of protons into this device.*

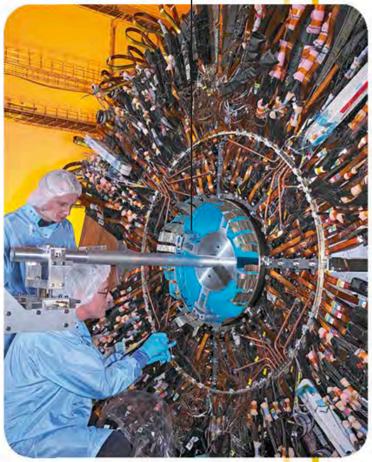

ATLAS, a device at the Large Hadron Collider, CERN, Switzerland

This airbag *is triggered by a sensor that contains beryllium.*

Airbag

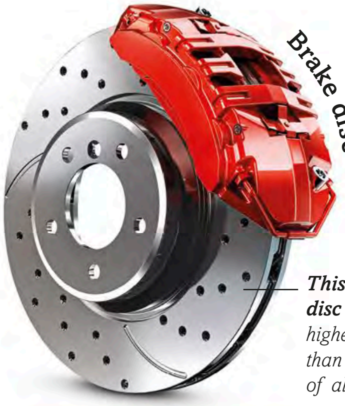

Brake disc

This beryllium disc *can withstand higher temperatures than a disc made of aluminium.*

of beryllium-rich glass to shield optical sensors to aid pilots flying at night or through fog. Objects made of this metal keep their shape well and hardly expand or contract if the temperature changes. This makes beryllium useful in valves for **fire sprinklers** and car sensors that trigger **airbags**. NASA's **James Webb Space Telescope** will use a large beryllium mirror that is light and strong. Beryllium is also used to make **brake discs** for racing cars. Alloys of beryllium and copper are used in springs as well.

12 Mg Magnesium

⊖ 12 ⊕ 12 ○ 12

State: Solid
Discovery: 1755

Forms

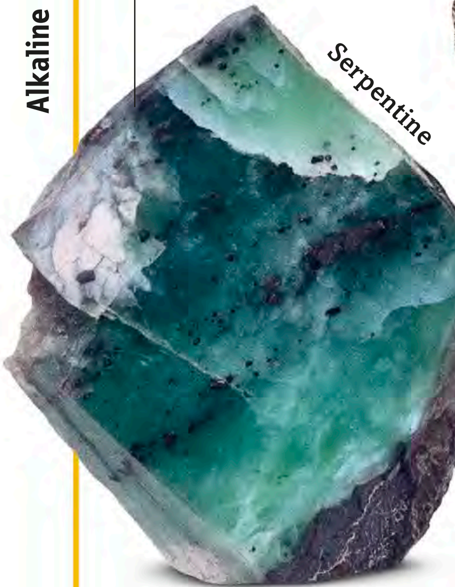

Serpentine

This green, magnesium-rich mineral forms deep underground.

Tremolite

Feather-like appearance

Laboratory sample of pure magnesium

Shiny, grey crystallized form

Magnesium has **22 known isotopes**.

Dolomite

***This ore** is a natural form of magnesium carbonate.*

Magnesium was named after Magnesia in Greece. This element largely exists deep inside Earth's mantle, but it can also be found in seawater and many minerals in our planet's crust, including **serpentine**. Another mineral, **dolomite**, is also a source of **pure magnesium**.

Magnesium has many important applications. Alloys of magnesium are not only strong, but also lightweight, so are used in a range of objects, from **car wheels** to **cameras**. For centuries, many naturally occurring magnesium minerals have been used in traditional medicines.

Uses

Alloy wheel

Magnesium alloy *makes this wheel strong and shiny.*

MAGNESIUM IN CHLOROPHYLL

Chlorophyll is an important molecule in plants and is what makes them green. At its centre sits a magnesium atom, which helps plants convert sunlight into energy in a process called photosynthesis.

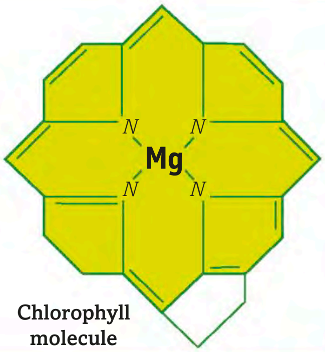

Chlorophyll molecule

Digital camera

The magnesium alloy body *of this camera is lightweight and will not rust.*

Epsom salts

Crystals containing magnesium sulfate *are added to warm water for a soothing bath.*

This powder *makes skin smooth and soft.*

Talcum powder

Milk of magnesia

This indigestion medicine *is a mixture of water and magnesium carbonate.*

White lights *from burning magnesium compounds*

Magnesium fireworks

Portland cement

This widely used cement *contains powdered magnesium oxide.*

Laptop

The magnesium alloy case *of this laptop is strong but lightweight.*

Magnesium carbonate, or **magnesia**, reacts with acid in the stomach to settle indigestion. Heating magnesia produces magnesium oxide, which is one of the ingredients in **cement**. Magnesium compounds are also used in **fireworks**, and they burn hot with a white flame. Salts composed of magnesium, called **Epsom salts**, named after the place in England where they were first mined, work as a muscle relaxant. Magnesium silicate, known as **talc**, is a soft mineral used in body powders.

20 Ca Calcium

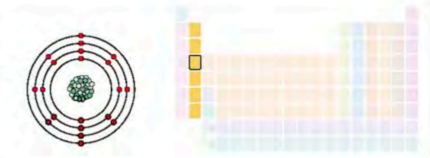

⊖ 20 ⊕ 20 ○ 20

State: Solid
Discovery: 1808

Forms

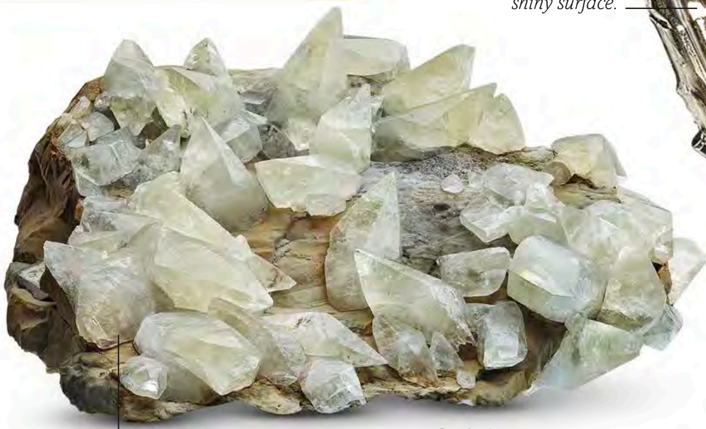

Large, fang-like crystals

Calcite

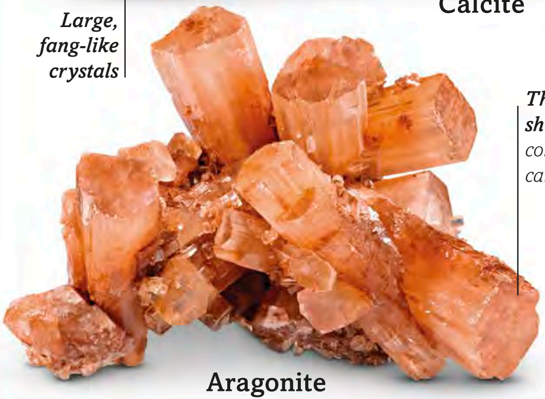

These column-shaped crystals *contain calcium carbonate.*

Aragonite

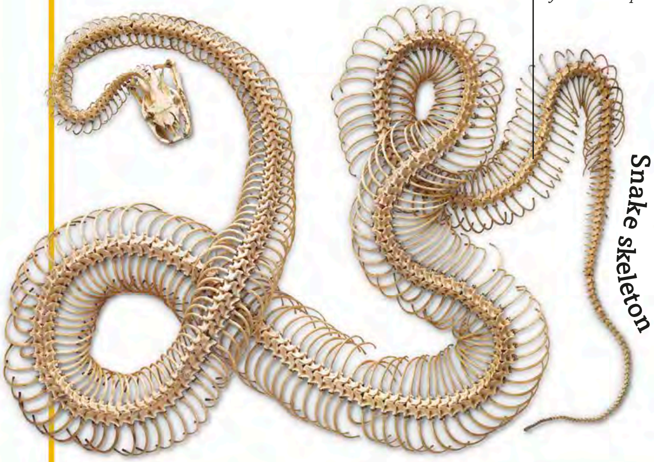

Bones *are hardened by calcium phosphate.*

Snake skeleton

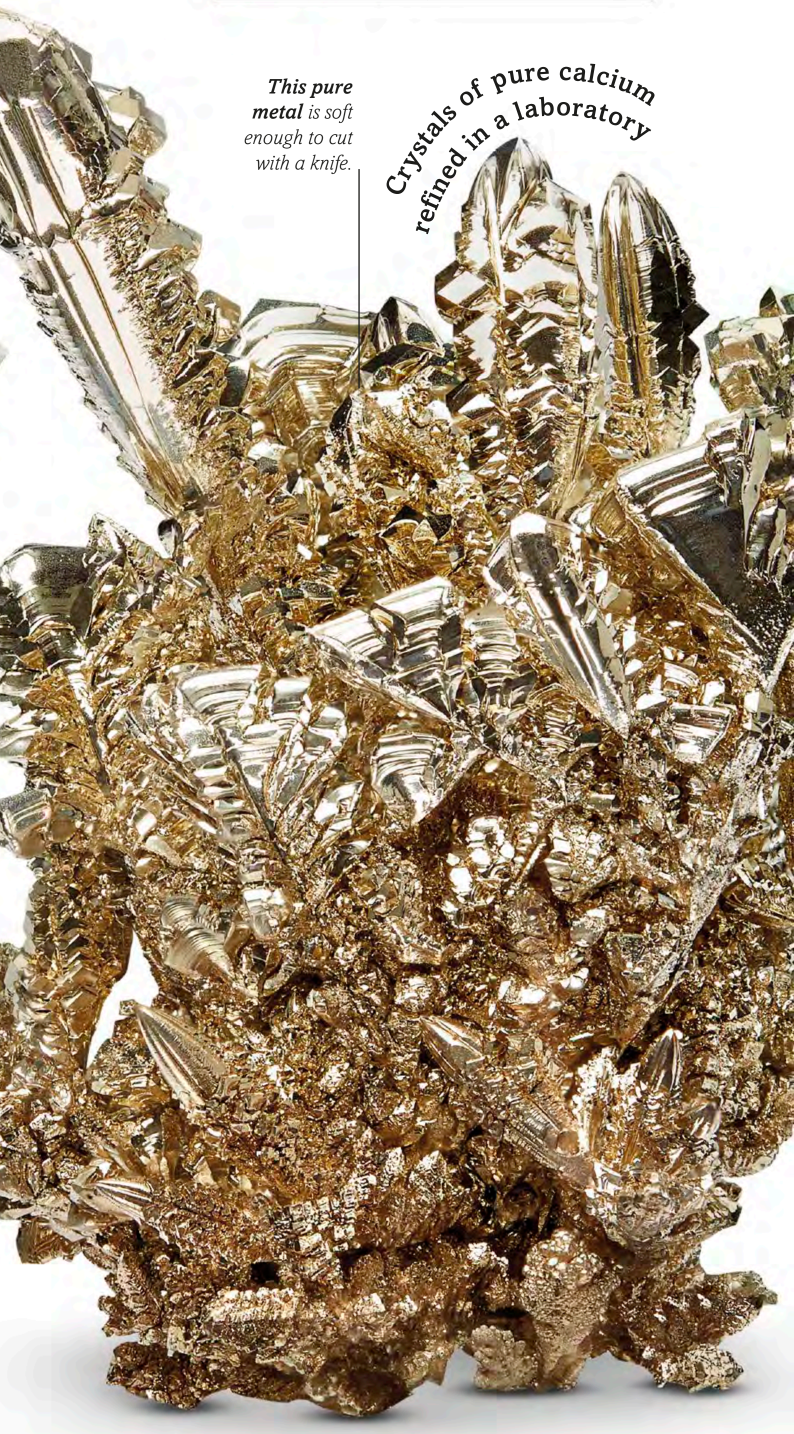

This crystal *has a shiny surface.*

This pure metal *is soft enough to cut with a knife.*

Crystals of pure calcium refined in a laboratory

The most abundant metal in the human body, calcium is also the fifth most common element on Earth. It appears in many minerals: **calcite** and **aragonite** are made of a compound of calcium and carbon called calcium carbonate. Bones in animal **skeletons** contain the compound calcium phosphate. The hard, outer layers of many other animals, such as the **shells** of sea snails, are made of calcium carbonate. Calcium is very important in our diet. We get calcium by eating **calcium-rich food**, including dairy products, green

Sea shell

Shells of sea snails *are hardened by calcium carbonate absorbed from sea water.*

Uses

Antacid tablets

This tablet *contains calcium carbonate, which is an alkali – a substance that balances out acids.*

An adult **human** contains about **1 kg (2 lb) of calcium** in the body.

CALCIUM CAVES

As running water flows into caves, it deposits calcium carbonate. These deposits build up to form structures called stalactites and stalagmites.

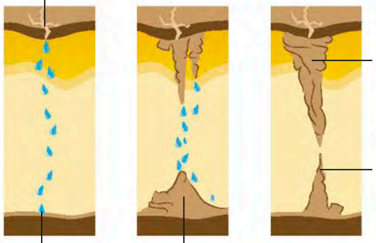
Water with dissolved calcium carbonate flows through a crack and into the cave.
Stalactite hangs from the ceiling.
Stalagmite grows up from the ground.
Water drips onto the ground.
Over time, calcium carbonate starts to build up on the ground and ceiling.

Plaster cast

This plaster of Paris cast *hardens when dry, supporting broken bones.*

Writing chalk

These chalks *contain calcium sulfate.*

Calcium-rich food

Milk
Broccoli
Orange

Marble statue

Marble *forms when limestone comes under high temperature and pressure.*

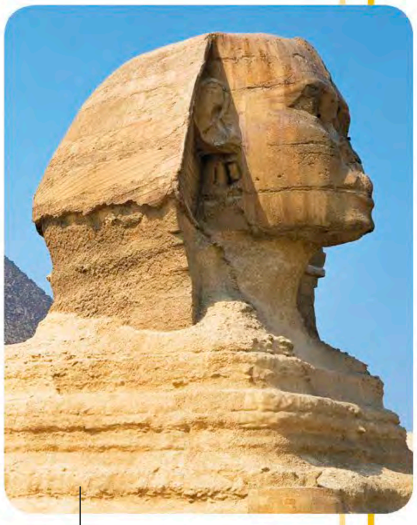

The Sphinx, Egypt

This statue *is made of limestone, a natural rock containing calcium carbonate.*

Oranges are also a good source of calcium, and most orange juices have extra calcium added to them. **Antacid tablets**, used to settle indigestion, contain calcium carbonate. This compound reacts with acid in the stomach. Calcium compounds are also common in construction materials. Plasterboard, which is used to make walls smooth, **writing chalk**, and **Plaster of Paris** are all made from the mineral gypsum. Calcium oxide is an important ingredient in cement and helps turn it into hard concrete.

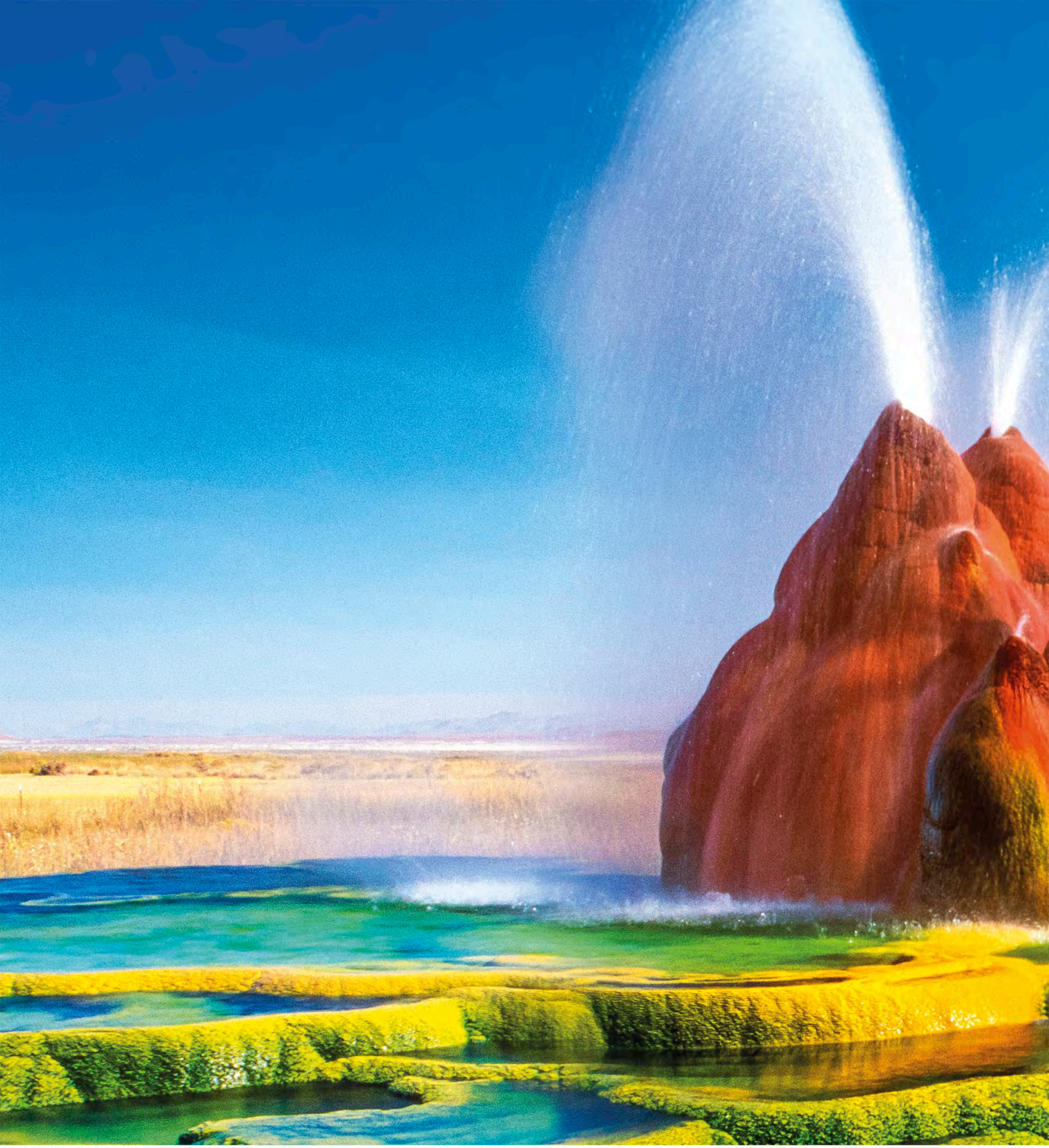

FLY GEYSER The multicoloured Fly Geyser in the Black Rock Desert of Nevada, USA, is made from a mound of calcium carbonate rock. Such mounds and pools are made naturally in many other places where springs gush out warm, calcium-rich waters. The amazing colours of the rocks are caused by algae and bacteria that live in this water.

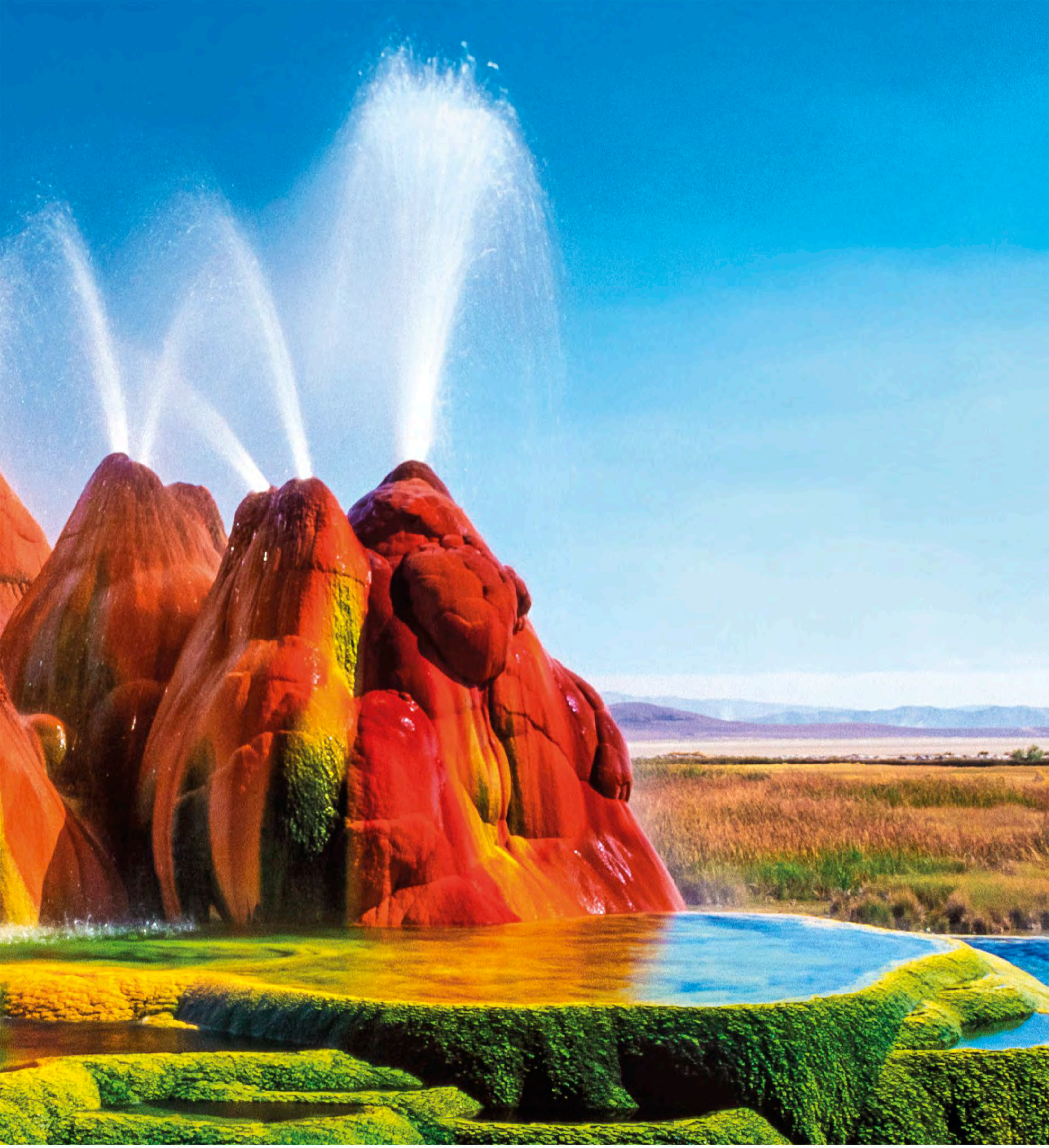

Fly Geyser is not a natural wonder. It was made by accident in 1964 when engineers were drilling a well to find a source of hot water. They did find a small reservoir of water that is heated by volcanic activity deep beneath the surface, but they chose to cap the well and look elsewhere. Eventually, the hot water broke through, creating a natural fountain, or geyser. Over the decades, the calcium deposits have slowly built up. The central mound is now 1.5 m (5 ft) tall and nearly 4 m (13 ft) wide. The scalding water that gushes out can reach a height of 1.5 m (5ft).

38 Sr Strontium

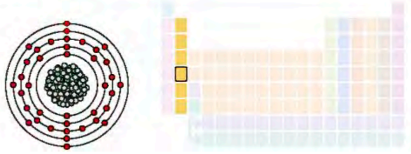

● 38 ⊕ 38 ○ 50

State: Solid
Discovery: 1790

Forms

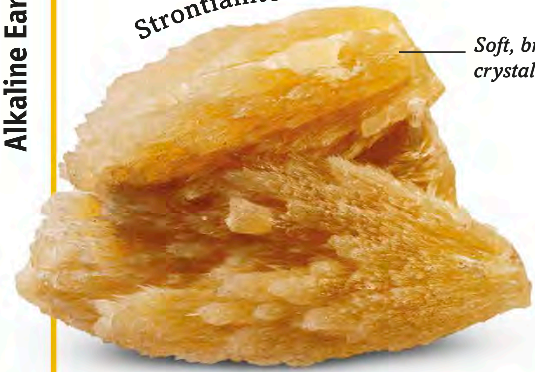
Strontianite
Soft, brittle crystals

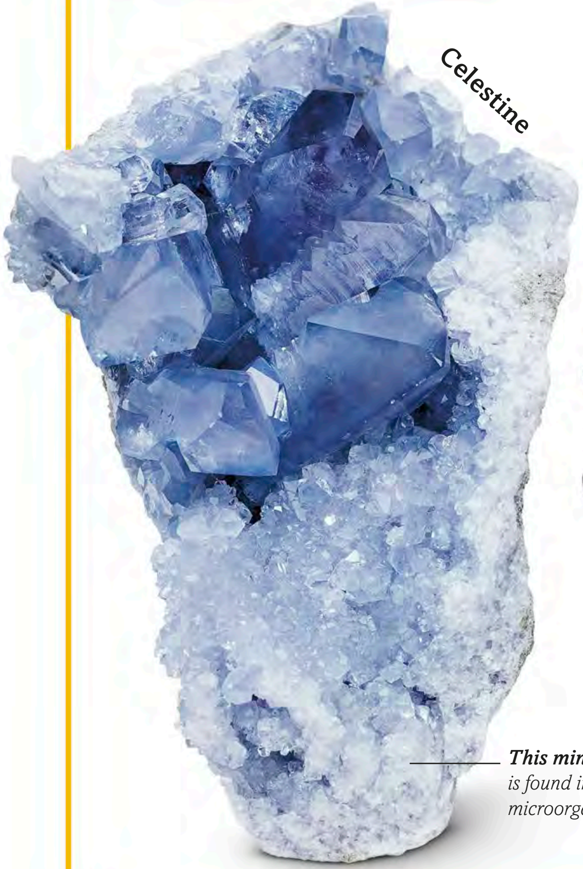
Celestine
This mineral *is found in some microorganisms.*

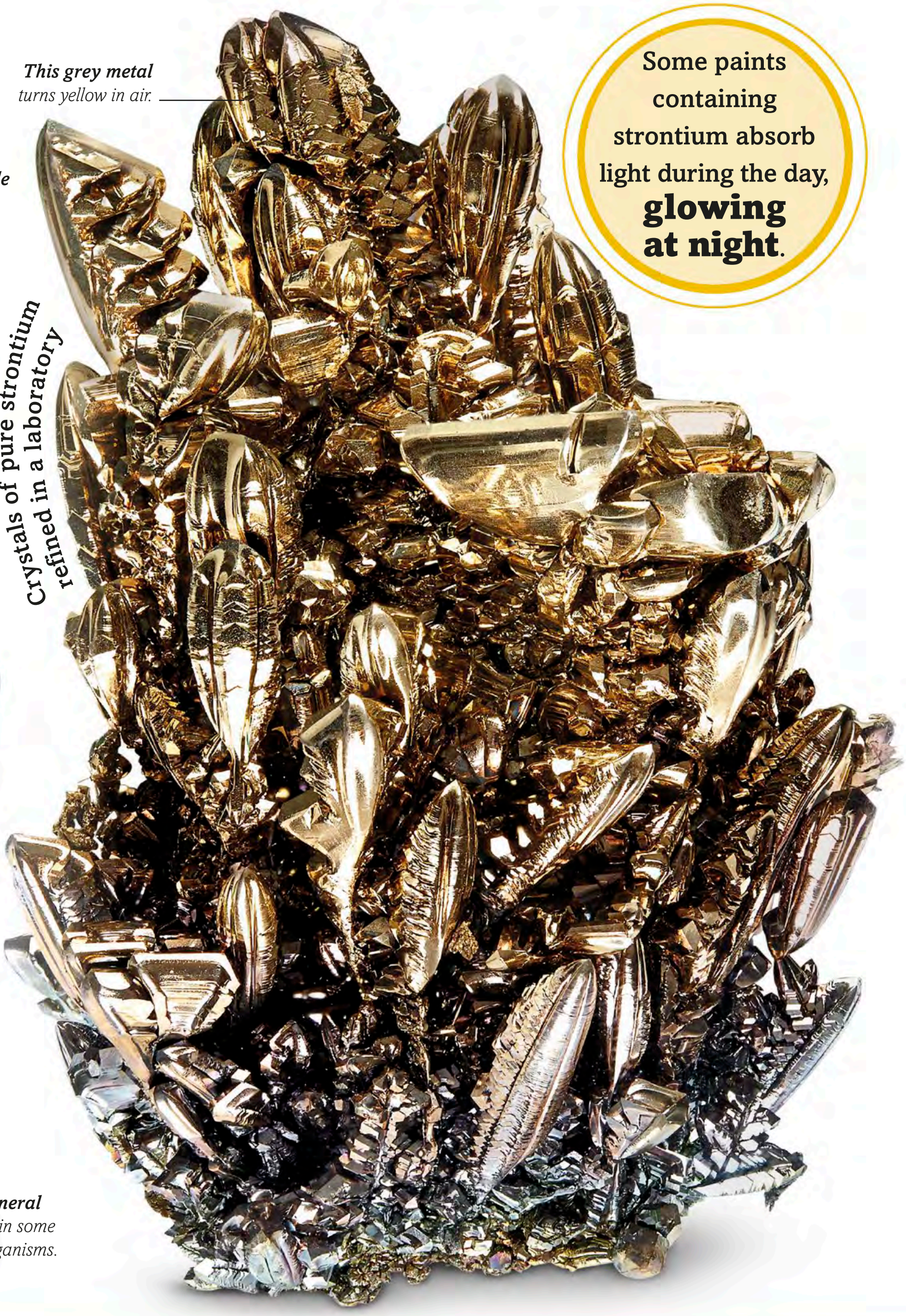
This grey metal *turns yellow in air.*
Crystals of pure strontium refined in a laboratory

Some paints containing strontium absorb light during the day, **glowing at night**.

Strontium was discovered in 1791 in a mineral found near the Scottish village of Strontian. The mineral burned with a bright crimson flame, and Scottish chemist Thomas Charles Hope studied it and found that it contained a new element. This mineral was called **strontianite**, and it is the main ore of strontium. **Pure strontium** was first extracted by British chemist Humphry Davy in 1808, who conducted an experiment using electricity to obtain the element from the mineral. Strontium was once used in television screens, but today

Uses

Glazed ceramic

The bowl's smooth finish *is due to strontium oxide.*

Lights in unmanned buoys *can be powered by radioactive strontium.*

Navigation buoy

Loudspeaker

Magnets inside this loudspeaker *contain strontium.*

Strontium burns *in air with a bright red colour.*

Flare

GENERATING ELECTRICITY

A radioactive form of strontium, called an isotope, can be used to produce electricity. A radioisotopic thermoelectrical generator (RTG) converts heat from the element into electricity for use in spacecraft.

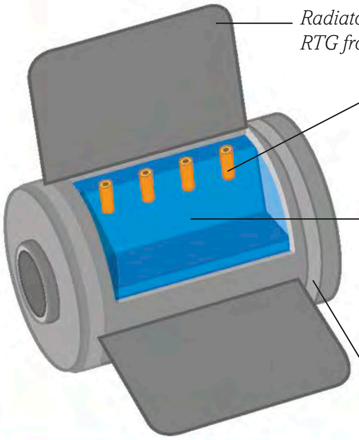

Radiator fin prevents the RTG from overheating.

A thermocouple is a device that converts heat from the radioactive metal into electricity.

Inside the capsule, atoms of radioactive strontium break down into those of lighter elements and produce heat.

RTG is insulated to stop radioactivity from leaking out.

Strontium compounds in some toothpastes *provide relief from pain.*

Toothpaste for sensitive teeth

Unmanned radar stations *run on electricity produced using a form of strontium called strontium-90.*

Weather radar station

there are fewer uses for it. Strontium oxide in pottery and **ceramic** glazes creates distinctive colours, while strontium carbonate produces a red colour in **flares** and fireworks. Magnets that contain iron oxide can be made stronger by adding strontium to them. These strong magnets are used in **loudspeakers** and microwave ovens. Strontium chloride is added to some kinds of **toothpaste**, while radioactive strontium is a source of electricity for **radar stations** in remote places where there are no power lines or fuel supplies.

56
Ba

Barium

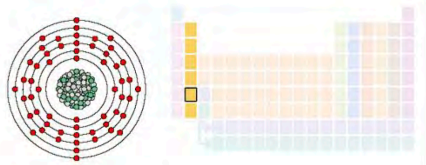

56 56 81

State: Solid
Discovery: 1808

Forms

Farmers used witherite as **rat poison** until the 18th century.

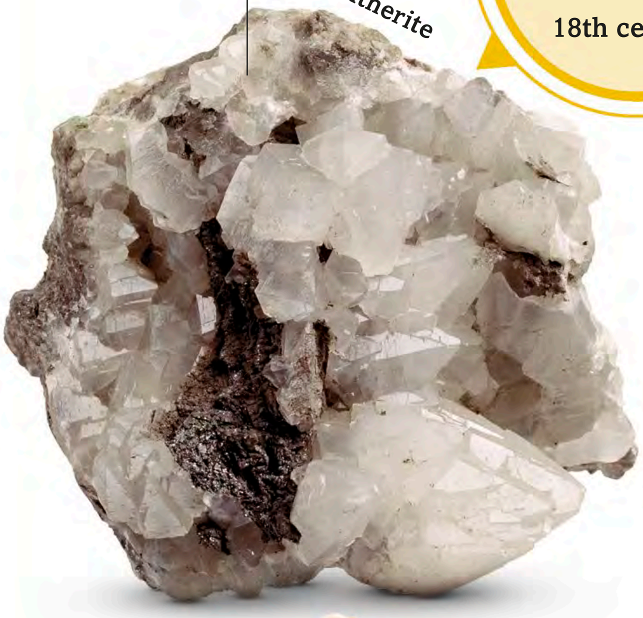

Witherite

This mineral *is used to make ceramic glazes.*

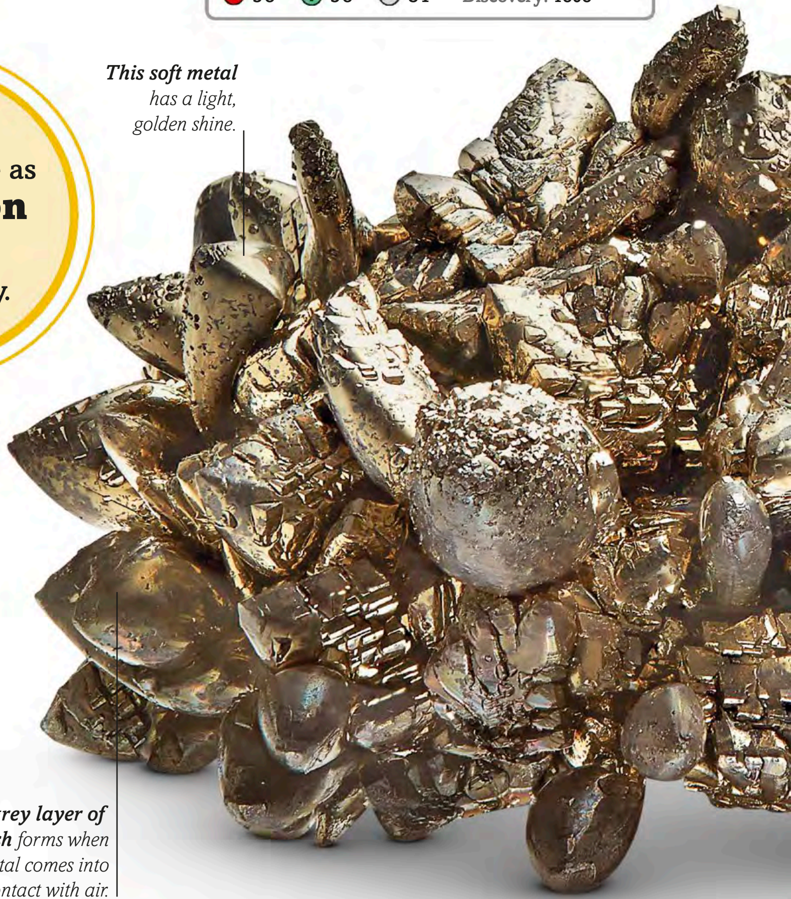

This soft metal *has a light, golden shine.*

A grey layer of tarnish *forms when the metal comes into contact with air.*

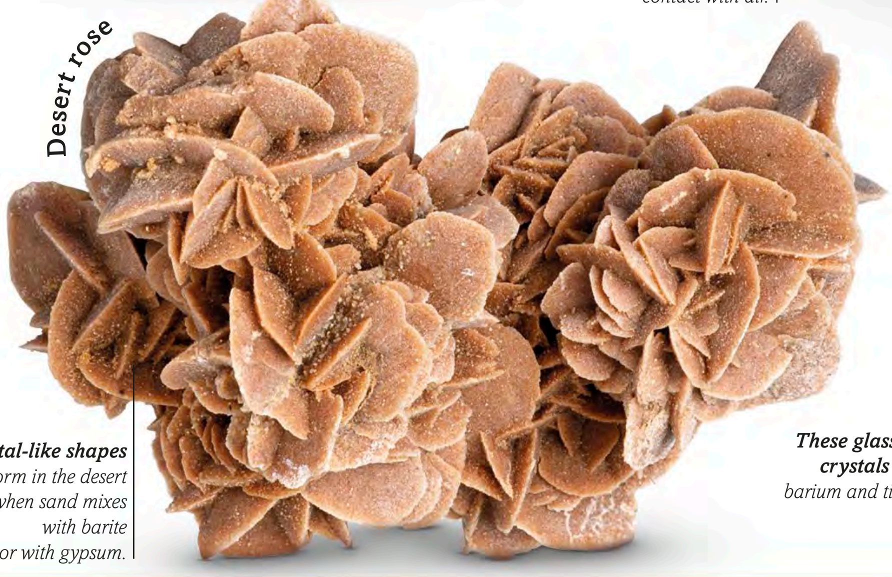

Desert rose

Petal-like shapes *form in the desert when sand mixes with barite or with gypsum.*

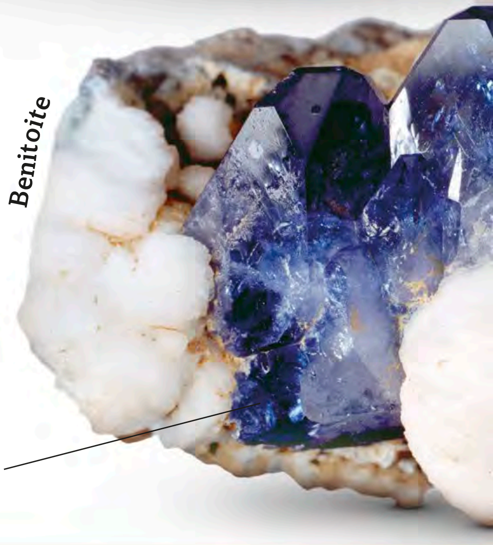

Benitoite

These glassy, blue crystals *contain barium and titanium.*

Barium is named after the Greek word *barys*, which means "heavy", because barium and its minerals are dense. The **pure form** of this element was first discovered in 1808 by the English chemist Humphry Davy, who extracted it from an oxide of barium. This does not exist in nature. Davy obtained it by heating the mineral **witherite**. Today, the main source of barium is barite, a mineral of sulfur that forms in deserts and in rock deposits that come into contact with hot water. A rarer mineral called **benitoite** also contains barium. The

Laboratory sample of pure barium crystals

Uses

Spark plug

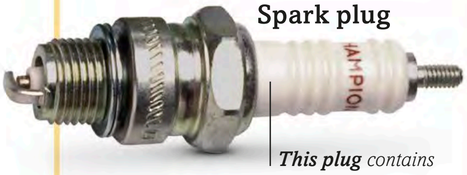

This plug *contains an alloy of barium and nickel.*

Glass-making

This glass *can be made shinier by adding barium oxide and barium carbonate.*

Jasperware pot

This pot *is made from clay that is rich in barium.*

Vacuum tube

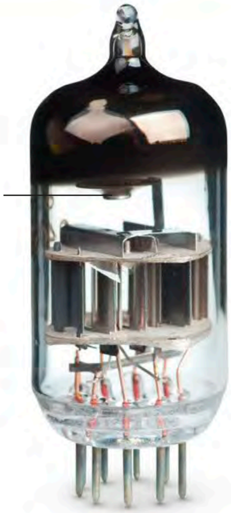

The barium in the metallic strip *absorbs gases in the tube, maintaining a vacuum.*

X-ray scan

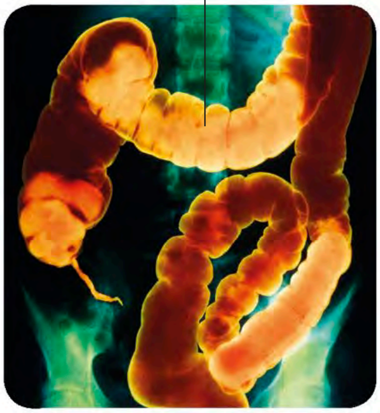

The intestine *is filled with a barium solution.*

BARIUM SOLUTION

Barium is used in a medical test for checking a patient's digestive tract for problems. In this test, a patient swallows a barium liquid solution, which fills the organs in the digestive tract.

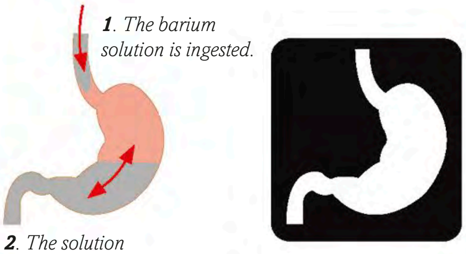

1*. The barium solution is ingested.*

2*. The solution enters the stomach and begins to fill it.*

3*. Under an X-ray scan, the barium-filled stomach shows up clearly.*

element is used in **spark plugs** to make them produce more powerful sparks and is added to **glass** to increase its shine. Barium compounds are added to some types of clay used for making **pots** and vases. In oil wells, barium compounds are added to drilling fluids to increase their density. Doctors make use of barium's density by giving patients a solution of barium compound to swallow, before taking **X-rays** of their digestive system. The barium makes the soft digestive organs denser, allowing them to be seen clearly with an X-ray machine.

88

Ra Radium

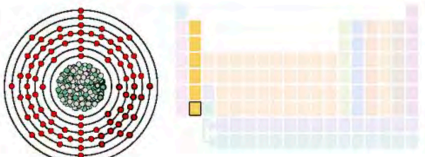

⊖ 88 ⊕ 88 ○ 138

State: Solid
Discovery: 1898

Forms

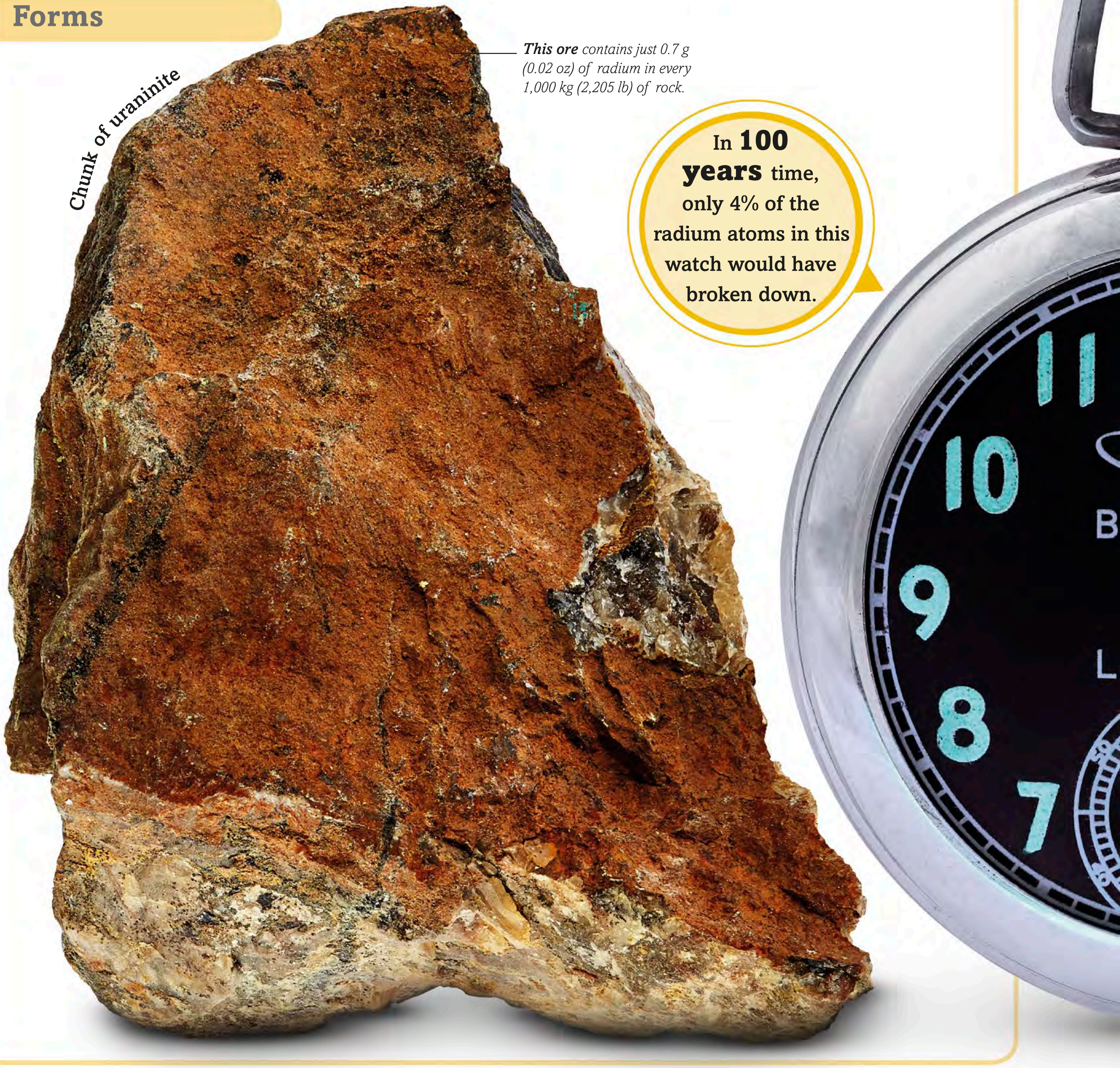

This ore *contains just 0.7 g (0.02 oz) of radium in every 1,000 kg (2,205 lb) of rock.*

In **100 years** time, only 4% of the radium atoms in this watch would have broken down.

Radium is the only radioactive member of the alkaline earth metals. It is also the rarest element in this group, and forms in small amounts when the atoms of more common metals – such as uranium and thorium – break down. Radium atoms do not survive for long, with most of them quickly decaying into radon, a radioactive noble gas. This element is highly dangerous and is rarely used today. However, in the early 20th century, radium compounds were in common use. Luminous paints, like those used to make **watch dials**

Uses

PIERRE AND MARIE CURIE

Radium was discovered in 1898 by Marie and Pierre Curie. They found that uranium ores produced more radioactivity than expected from samples of uranium. They realized another radioactive metal was present and named it radium.

Ingraham
BILTMORE
LUMINOUS

The radium paint *in this clock makes the numbers glow green-blue in the dark.*

Pocket watch with a luminous dial

This vial *contains a liquid called radium chloride.*

Vials for radium treatment

RADIUM-EMANATOR
GUARANTEED
10000 MILLIMICROCURIE
OF RADIUM-EMANATION PER DAY
RADIUM VITA LTD.
LONDON S.W.
Vita

This machine *from the early 20th century mixed radium into water, which was thought to make it healthier to drink.*

Radium emanator

Cosmetics

Nº 1
RADIUM
VITAE

Skin lotions *containing radium were common in the 1920s.*

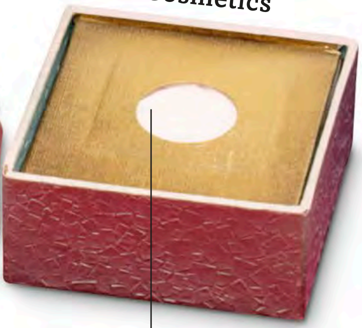

Radium face powder *was once thought to be good for the skin.*

glow in the dark, were created using radium. People working with this paint often became ill, especially with cancer, because the radiation produced by radium damages DNA. Nevertheless, until the 1940s, many people thought radium's radioactivity made them stronger, not weaker. They injected themselves with **vials containing a radium compound**, believing it gave them an energy boost. They also thought that creams and **cosmetics** with radium in them made the skin healthier, even though they did exactly the opposite.

Oddly shaped piece of pure cobalt (Co).

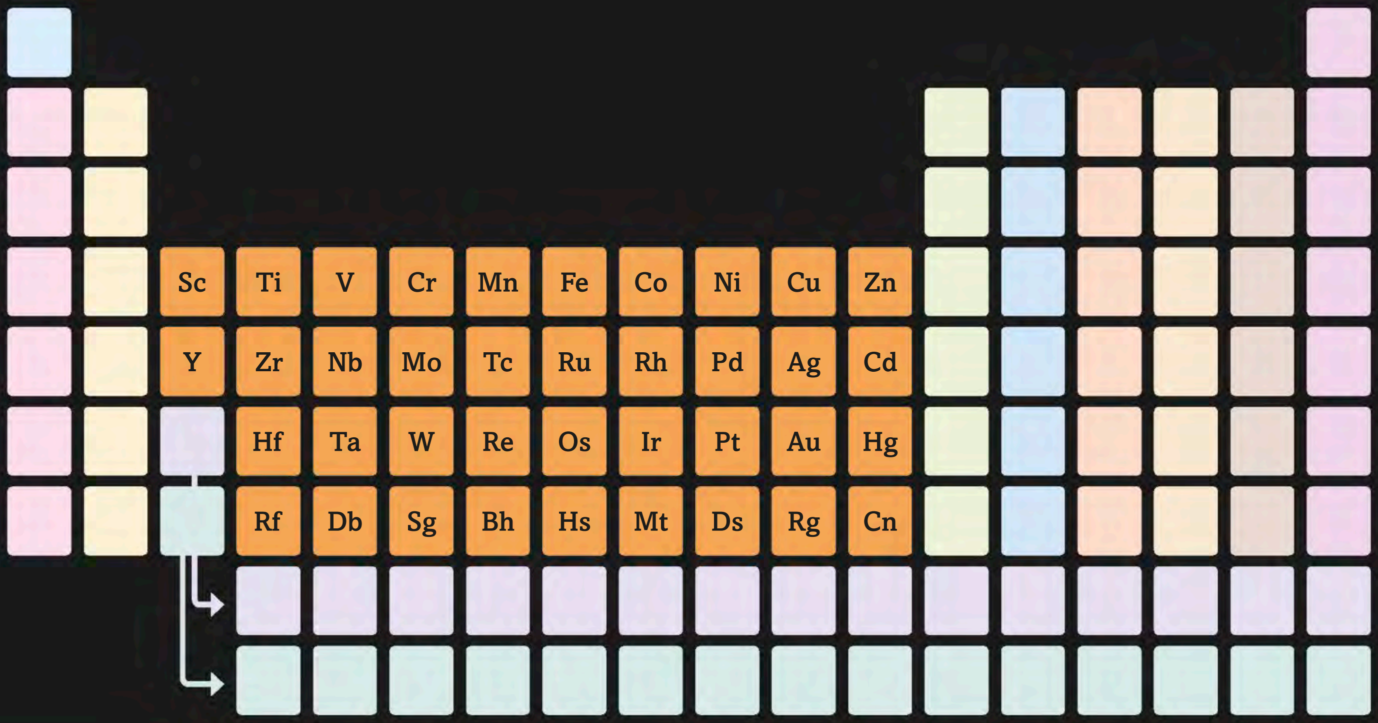
Sc Ti V Cr Mn Fe Co Ni Cu Zn
Y Zr Nb Mo Tc Ru Rh Pd Ag Cd
Hf Ta W Re Os Ir Pt Au Hg
Rf Db Sg Bh Hs Mt Ds Rg Cn

Transition Metals

This is the largest set of elements in the periodic table. This block of metals contains useful elements, such as gold (Au), iron (Fe), and copper (Cu). Many of these metals are easy to shape. The fourth period of the block – from rutherfordium (Rf) to copernicium (Cn) – are artificial and do not occur in nature. They were created by scientists in laboratories.

Atomic structure
Most transition metals have two outer electrons, but a few, such as copper (Cu), have just one.

Physical properties
These elements are generally hard and dense metals. Mercury (Hg), the only element that is liquid at room temperature, also belongs to this group.

Chemical properties
Transition metals are not as reactive as alkali and alkaline earth metals. However, they form many varied and colourful compounds.

Compounds
Many compounds of transition metals are brightly coloured. These metals are often used in alloys, such as brass and steel.

21
Sc

Scandium

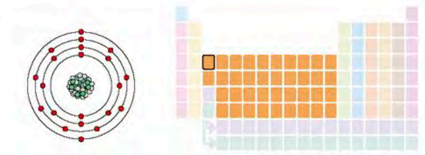

🔴 21 🟢 21 ⚪ 24

State: Solid
Discovery: 1879

Forms

Gadolinite

Greasy texture

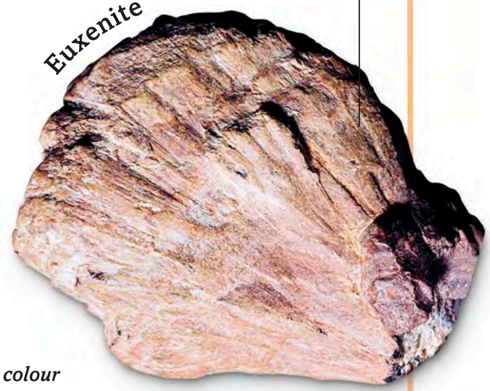

Euxenite

***These crystals** contain only tiny amounts of scandium.*

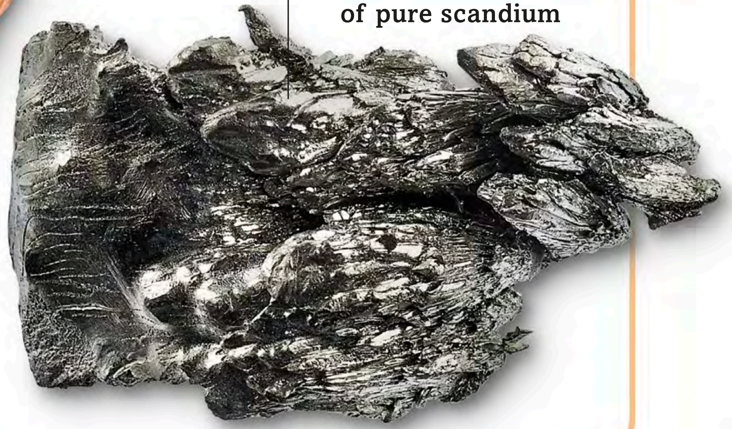

Laboratory sample of pure scandium

***The silvery colour of this metal** turns yellow in the air.*

Uses

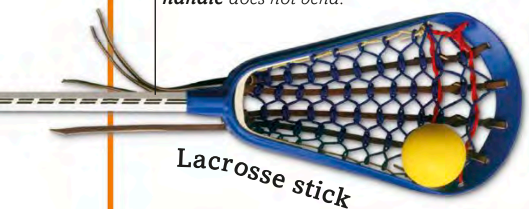

Lacrosse stick

***This lightweight alloy handle** does not bend.*

Metal halide lamp

***The scandium iodide gas** inside the tube glows with a bright bluish light.*

MiG-29 aircraft

***Some high-speed jets** have a fuselage made of scandium alloys.*

A soft and lightweight metal, scandium is similar to aluminium. It is spread so thinly in Earth's rocks that it is very difficult to collect a large amount of this element. Scandium is only used for specialist applications. Its main ores are the minerals **gadolinite** and **euxenite**, which also contain small amounts of many other rare metals, such as cerium and yttrium. Scandium mixed with aluminium makes a strong alloy, which is used in lightweight equipment for sports, such as **lacrosse**, and to make high-speed jets, such as the **MiG-29**.

22
Ti

Titanium

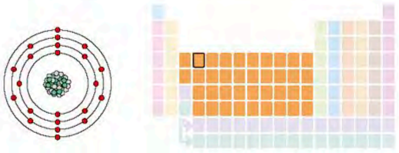

● 22 ⊕ 22 ○ 26

State: Solid
Discovery: 1791

Forms

Perovskite

This grey, cubic crystal *is made of the compound calcium titanium oxide.*

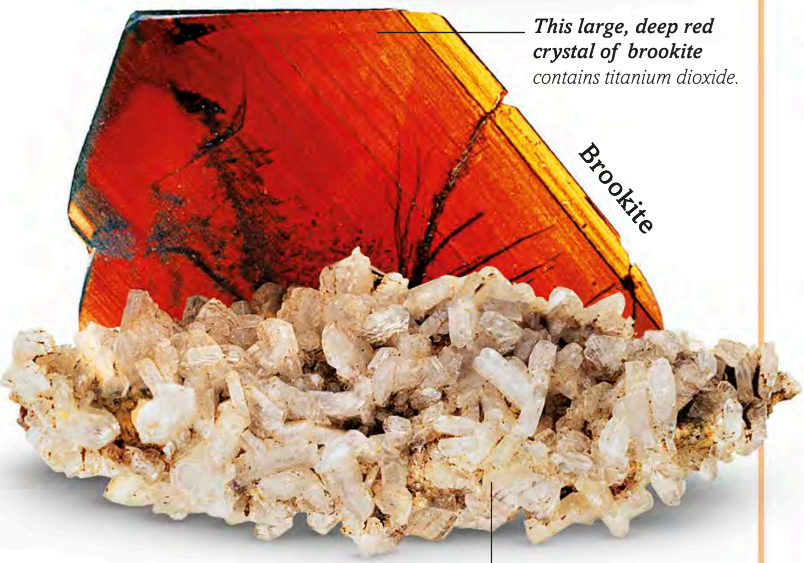

Brookite

This large, deep red crystal of brookite *contains titanium dioxide.*

These are crystals *of the mineral albite.*

This metal's shine fades to grey *when exposed to air.*

Laboratory sample of pure titanium

Uses

The titanium plate *placed inside acts as as shield.*

Body armour

Artificial hip joint

This titanium joint *can replace damaged bone in body.*

Sunscreen

The titanium dioxide *in this sunscreen blocks damaging ultraviolet (UV) radiation in sunlight.*

Drill bit

This drill bit *is hardened with a coating of titanium nitride.*

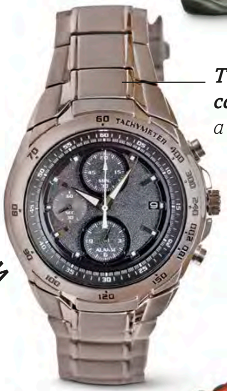

Wristwatch

This watch casing *is made of a titanium alloy.*

Roller blades

This ***titanium frame*** *is lightweight but strong.*

SALOMON
titalium

Named after the Titans, a race of mythic Greek gods, titanium is a silvery metal. It is as strong as steel but much lighter, and it is not corroded by water or chemicals. This strong metal also makes excellent protective shields in **body armour**. Titanium is commonly used to prepare titanium dioxide, a compound of titanium and oxygen that is used in paints and **sunscreen**. Titanium is not toxic so it can be used to make medical implants, such as artificial **hip joints**. **Wristwatches** made with titanium alloys are light and strong.

23 V Vanadium

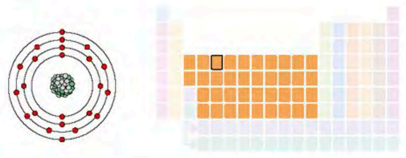

⊖ 23 ⊕ 23 ○ 28

State: Solid
Discovery: 1801

Forms

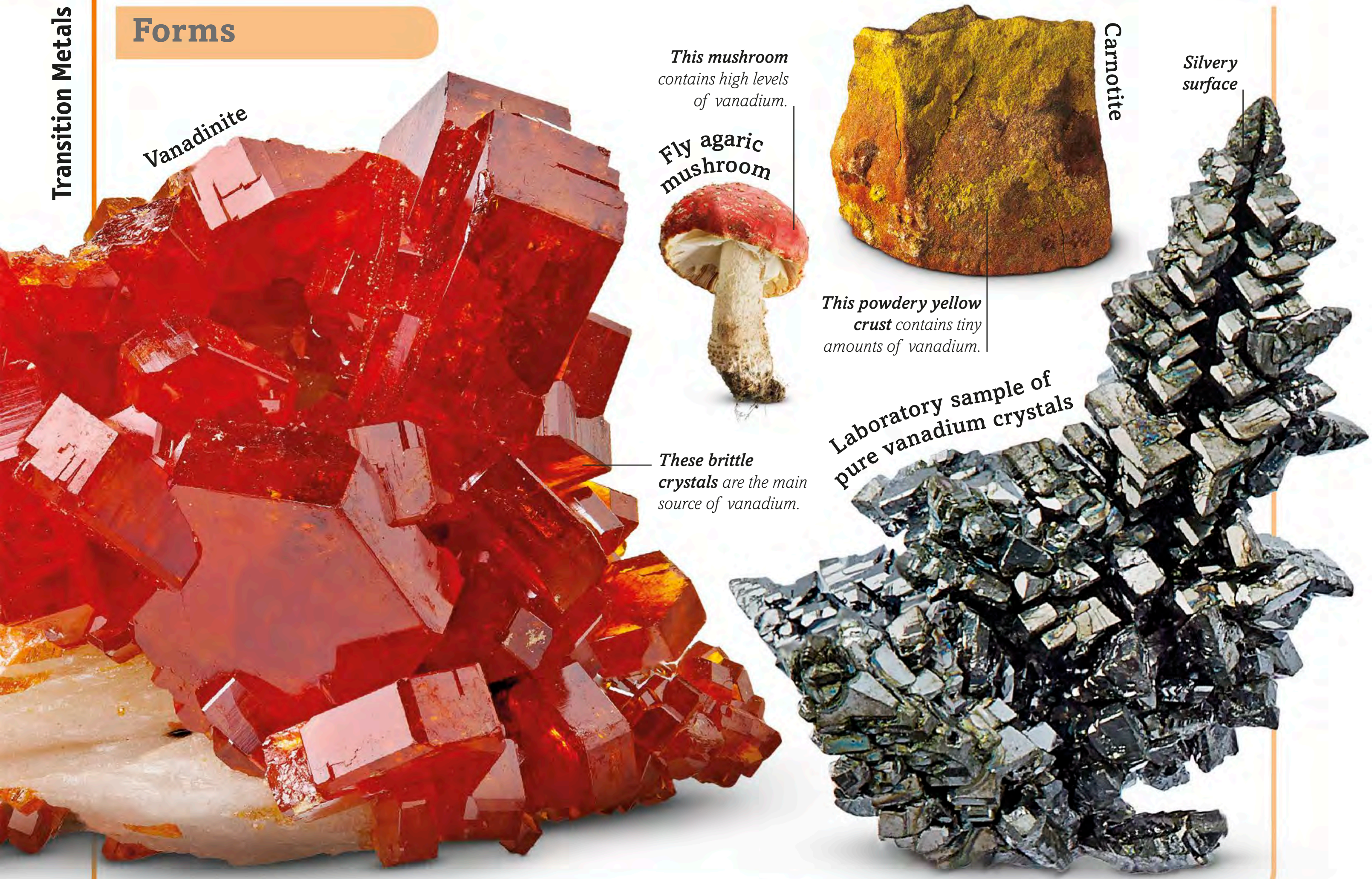
Vanadinite
Fly agaric mushroom
Carnotite
Silvery surface
Laboratory sample of pure vanadium crystals

This mushroom *contains high levels of vanadium.*

This powdery yellow crust *contains tiny amounts of vanadium.*

These brittle crystals *are the main source of vanadium.*

Uses

Tools *made with alloys of vanadium and steel are durable.*

About **85% of all vanadium** is used to toughen steel.

Damascus Steel knife

This knife *has been strengthened by the addition of vanadium.*

Vanadium can be hammered and stretched without breaking. This hard, strong metal is easy to shape. **Vanadium** was first **purified** in 1869 by the British chemist Henry Roscoe. Today, it is commonly extracted from its ore **vanadinite**. Ancient metalworkers used tiny amounts of vanadium compounds to make a very tough substance called **Damascus Steel**. This was named after the capital city of Syria, where ironworkers made the world's sharpest swords. Vanadium is still used to toughen tools, such as **spanners** and knives.

24 Cr Chromium

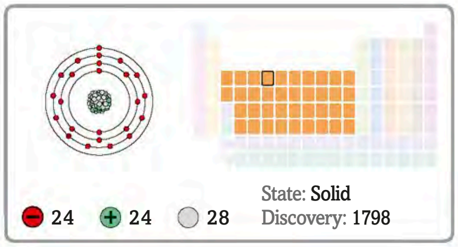
⊖ 24 ⊕ 24 ○ 28

State: Solid
Discovery: 1798

Forms

Chromite

Chromite *is dark grey to brown in colour.*

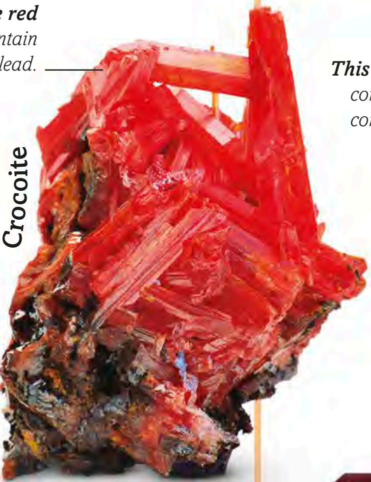
Crocoite

These large red crystals *contain chromium and lead.*

Laboratory sample of pure chromium

This metal stays shiny *even when exposed to air and water.*

Uses

Stainless steel kitchenware

This grater *can resist corrosion because it contains chromium.*

Ruby

This red colour *is due to the presence of tiny amounts of chromium oxide in the crystal.*

Motorcycle

The chromium plating *protects this motorbike from rusting.*

Chromium is named after *chroma*, the Greek word for "colour". Many minerals of chromium, including **chromite** and **crocoite**, are brightly coloured. An artificial form of crocoite, known as "chrome yellow", was once used in paints, but it was banned when scientists discovered it to be poisonous. **Pure chromium** doesn't corrode easily, so it is combined with iron and carbon to produce **stainless steel**. Chromium also gives gemstones, such as **rubies**, their deep-red colour. Some **motorcycles** have chromium-plated bodywork, giving them a shiny finish.

25 Mn Manganese

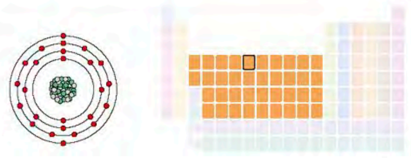

⊖ 25 ⊕ 25 ○ 30

State: Solid
Discovery: 1774

Forms

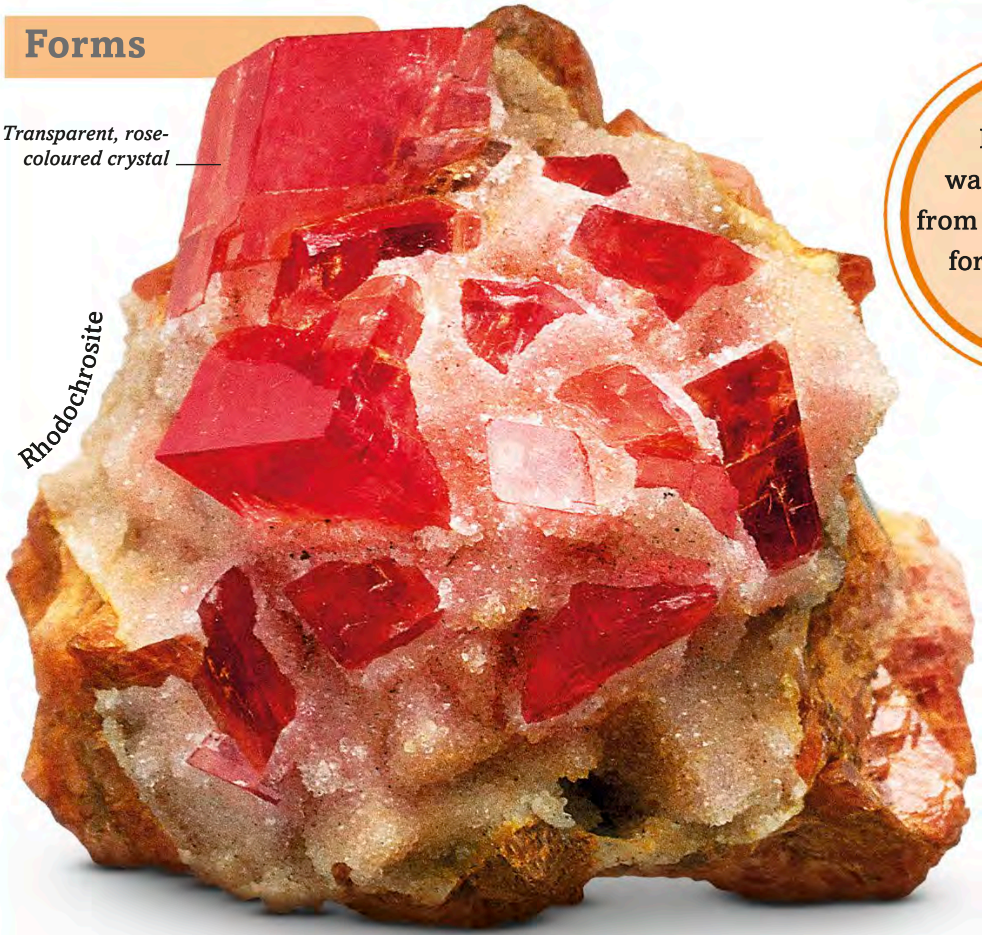
Transparent, rose-coloured crystal
Rhodochrosite

Manganese was **purified** from **pyrolusite** for the first time in 1774.

Shiny, silvery metal

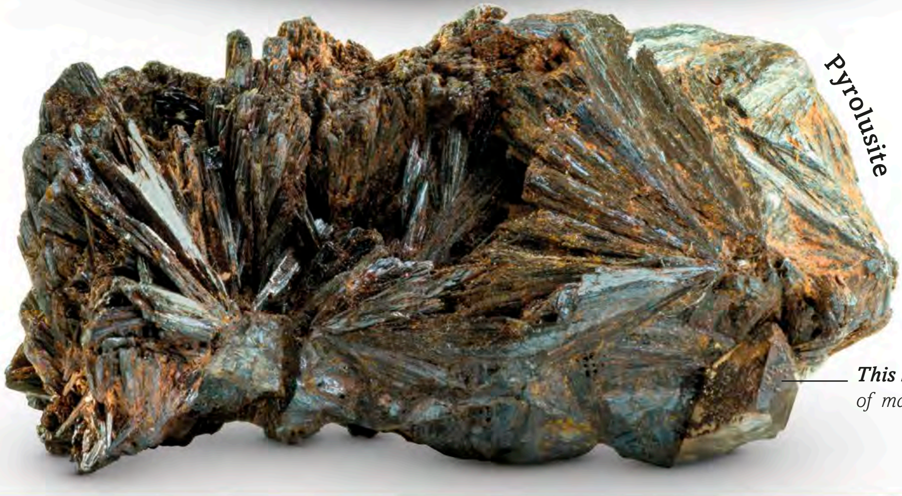
Pyrolusite
***This mineral** is made of manganese dioxide.*

Like magnesium, this element gets its name from the Greek region of Magnesia. There are many manganese minerals, including the colourful mineral **rhodochrosite**. The pure form of the metal is obtained mainly from the ore **pyrolusite**. **Pure manganese** is dense, hard, and brittle. This element is present in seawater as the compounds manganese hydroxide and manganese oxide, which have built up in layers over millions of years to form masses on the sea bed. The human body needs a tiny amount of manganese, which we can get

Pineapple
Oats
Mussel
Hazelnut

Manganese-rich food

Laboratory sample of pure manganese

JOHAN GOTTLIEB GAHN

In 1774, Swedish chemist Johan Gottlieb Gahn discovered manganese by reacting manganese dioxide with charcoal – which contains carbon – under a lot of heat. The carbon took the oxygen away from the compound, leaving behind pure manganese.

Uses

Jefferson nickel

This US coin *from World War II was made using manganese and silver because nickel was in short supply.*

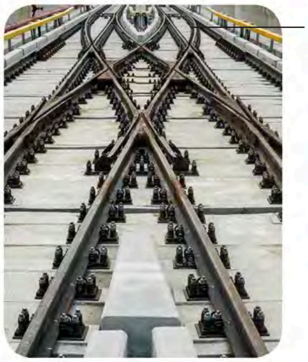

These steel tracks *have manganese added to them to make them stronger.*

Railway tracks

Dry cell battery

This battery *contains manganese dioxide.*

Unleaded petrol

This petrol *contains a manganese compound, which is less toxic than lead.*

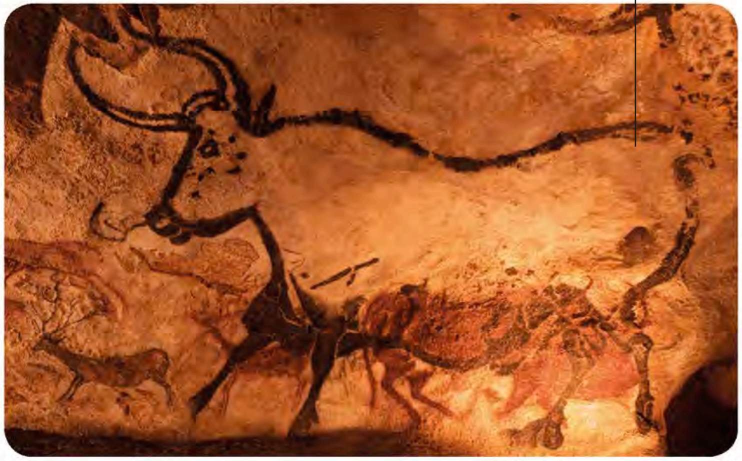

The black colour *comes from manganese dioxide.*

Lascaux cave paintings, France

Purple glass bottle

This glass is coloured *by adding a manganese compound called permanganate.*

from mussels, nuts, oats, and pineapples. The applications of manganese include its use in strengthening steel, which is used in making **railway tracks** and tank armour. Certain **dry cell batteries** carry a mixture containing manganese oxide. Manganese compounds are also added to **petrol** and used to clean impurities from **glass** to make it clear or to give it a purple colour. In prehistoric times, the compound manganese dioxide was crushed to make the dark colours used in **cave paintings**.

26

Fe

Iron

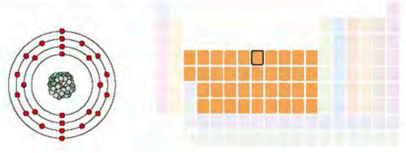

⊖ 26 ⊕ 26 ○ 30

State: Solid
Discovery: c. 3500 BCE

Forms

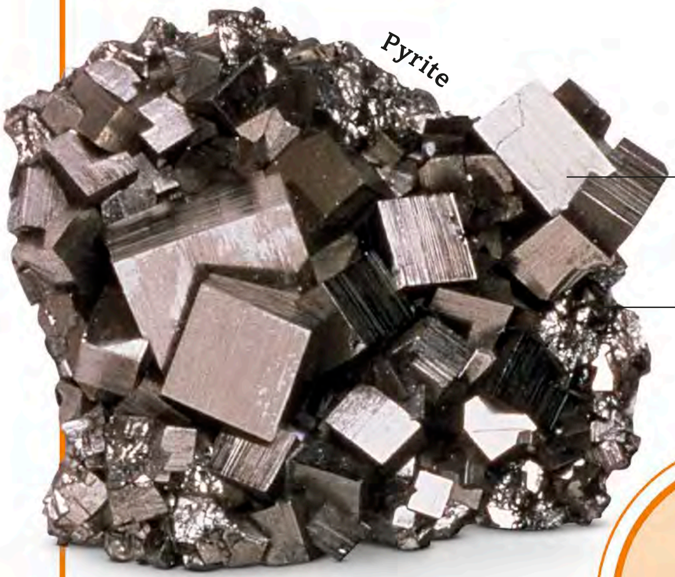

Pyrite

Cube-shaped crystal

***This mineral** is a compound of iron and sulfur.*

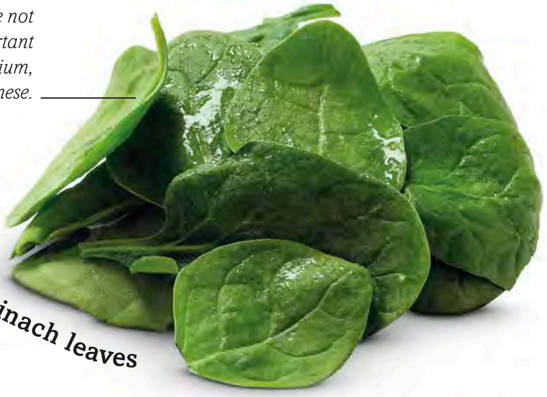

Spinach leaves

***Spinach leaves** provide not only iron but other important elements, such as potassium, calcium, and manganese.*

Iron is the **most common metal** on our planet.

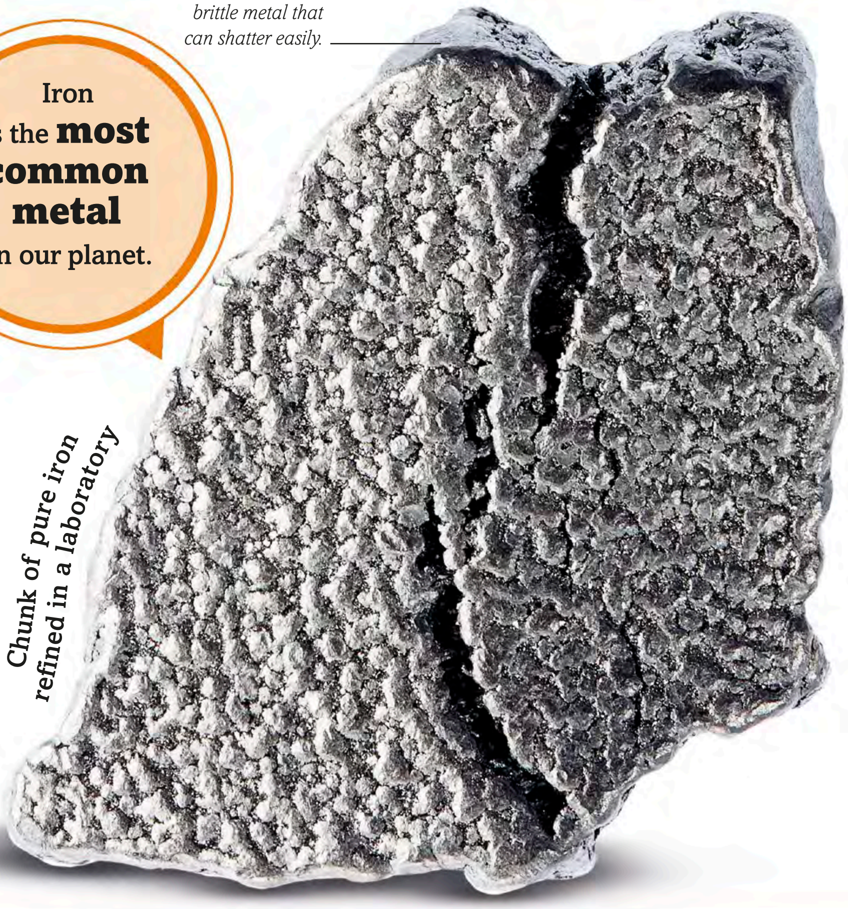

Chunk of pure iron refined in a laboratory

***Pure iron** is a brittle metal that can shatter easily.*

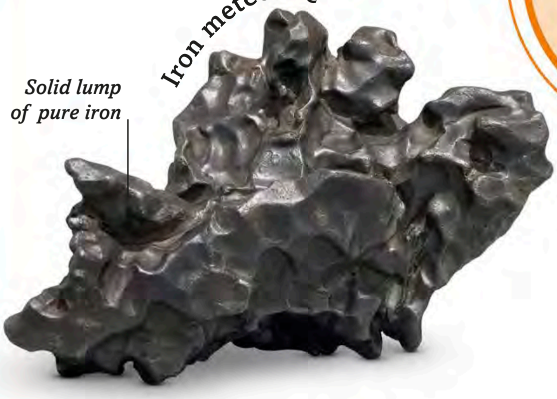

Iron meteorite

Solid lump of pure iron

Blood sample

***Blood** contains almost 70% of the iron in the human body.*

Most of the iron on our planet is locked away in Earth's hot, molten core. This element is widely found in rocks worldwide, and almost 2.5 billion tonnes of iron is purified every year. Mineral ores rich in iron include **pyrite**. Other ores, including haematite, are used to extract **pure iron** in a process called smelting. **Iron-rich meteorites** – chunks of rock from outer space that crash to Earth – are one of very few sources of naturally pure iron. The human body uses iron to make haemoglobin, a substance in blood that carries oxygen around our body (oxygen helps

Uses

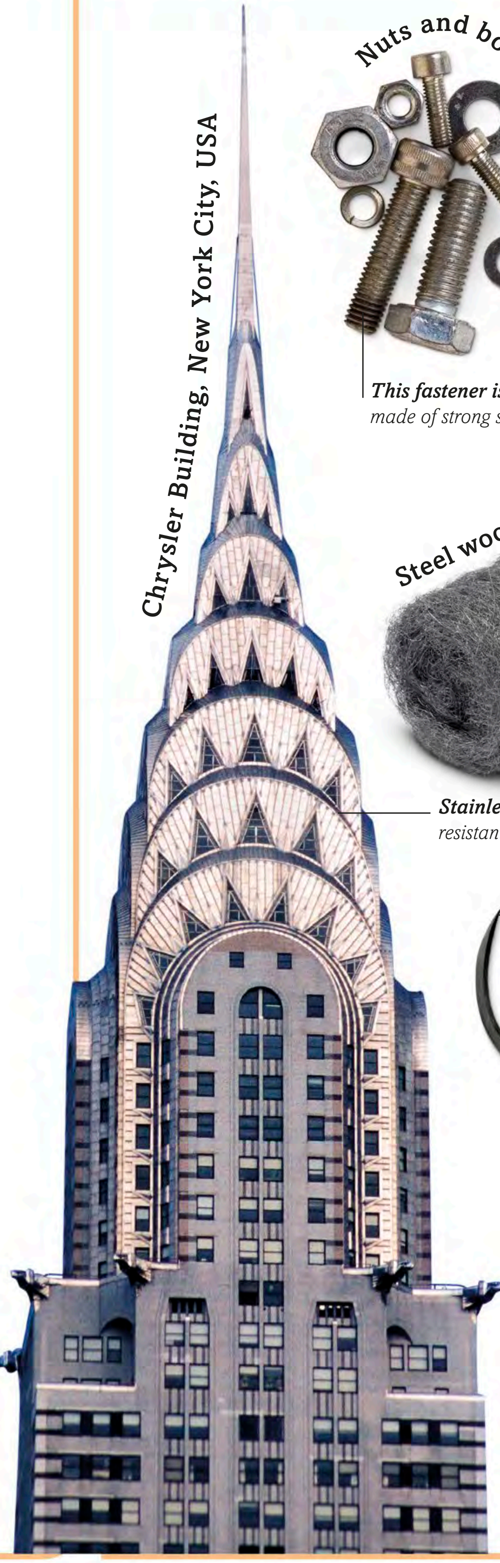
Chrysler Building, New York City, USA

Stainless steel is quite resistant to rain and wind.

Nuts and bolts

This fastener is *made of strong steel.*

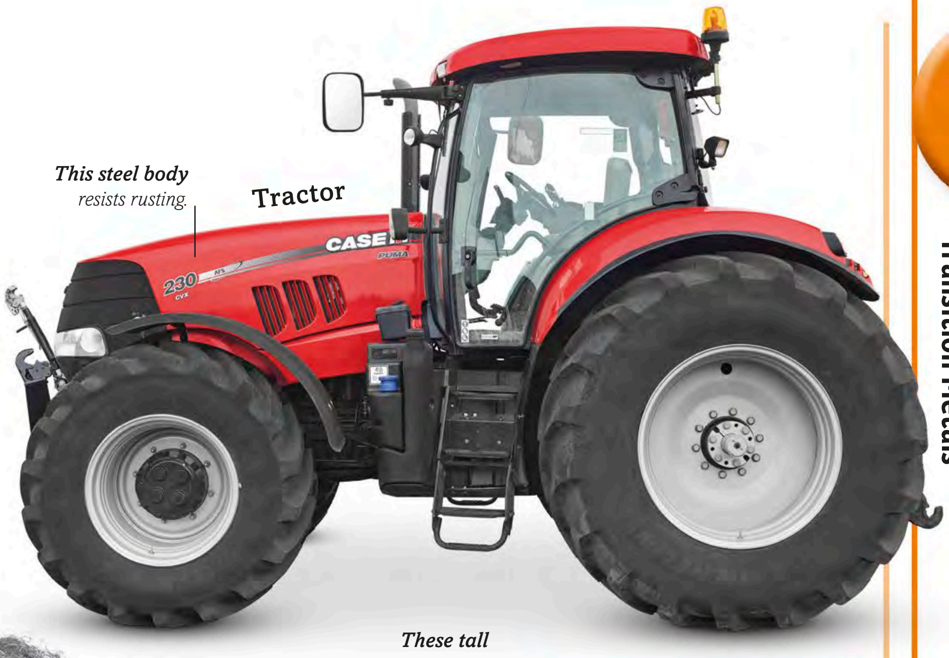
Tractor

This steel body *resists rusting.*

Steel wool

Thin wires *of steel are used to clean hard surfaces.*

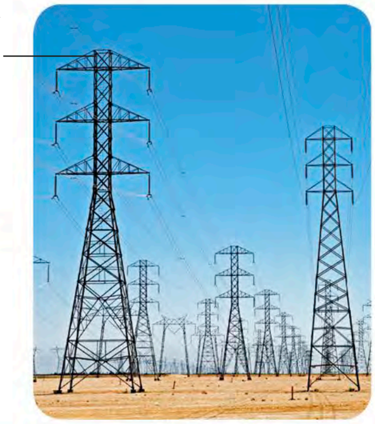
Transmission towers

These tall structures *are made from stiff steel girders.*

Cast iron pot

This iron pot *retains heat well while cooking.*

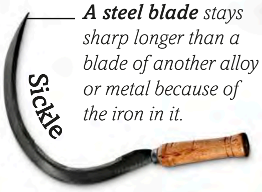
Sickle

A steel blade *stays sharp longer than a blade of another alloy or metal because of the iron in it.*

Iron filings and magnet

These small grains *of pure iron are magnetic and are attracted to the end of a magnet.*

SMELTING

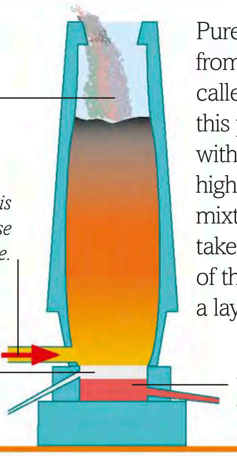
***1**. Iron ore and coal are added to the furnace.*

***2**. Hot air is added here to raise the temperature.*

***3**. Impurities float on the pure metal, then released.*

***4**. Pure iron sinks to the bottom, then removed.*

Pure iron is separated from its ores in a process called smelting. During this process, iron reacts with carbon in coal at a high temperature. As the mixture burns, the carbon takes the impurities out of the ore, leaving behind a layer of pure iron.

our cells produce energy for the body to work). Foods containing iron include meats and green vegetables, such as **spinach**. When pure iron comes into contact with air and water, it develops a flaky, reddish-brown coating called rust, which weakens the metal. In order to make iron tougher, tiny amounts of carbon and other metals, such as nickel and titanium, are added to it. This forms an alloy called steel, which is used to make **bolts** and strong **tractor** bodies, among other applications. Adding the element chromium to steel creates a stronger alloy called stainless steel.

STEELMAKING

A stream of red-hot, liquid metal pours from a furnace at a steelmaking workshop. This is the end of a long process in which iron ore is transformed into steel, a tough alloy that is strong enough to make girders for supporting skyscrapers and bridges. The steel may even be moulded into car bodies, woven into superstrong cables for elevators, or turned into powerful magnets that can levitate maglev trains.

Steel is an alloy of iron that contains about two per cent carbon and some other elements. The carbon locks all the atoms together and prevents the metal from cracking. This makes steel harder than iron: it bends before it breaks and doesn't shatter easily. To make steel, iron ore is smelted in a blast furnace to remove its impurities, such as nitrogen, sulfur, or phosphorus. Other elements can be added to create different varieties of steel. For example, chromium in steel stops it from rusting, while manganese makes it harder. Adding silicon to steel can make the alloy more magnetic, while nickel makes it less brittle at extremely low temperatures.

27
Co

Cobalt

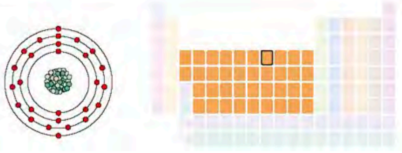

⊖ 27 ⊕ 27 ○ 32

State: Solid
Discovery: 1739

Forms

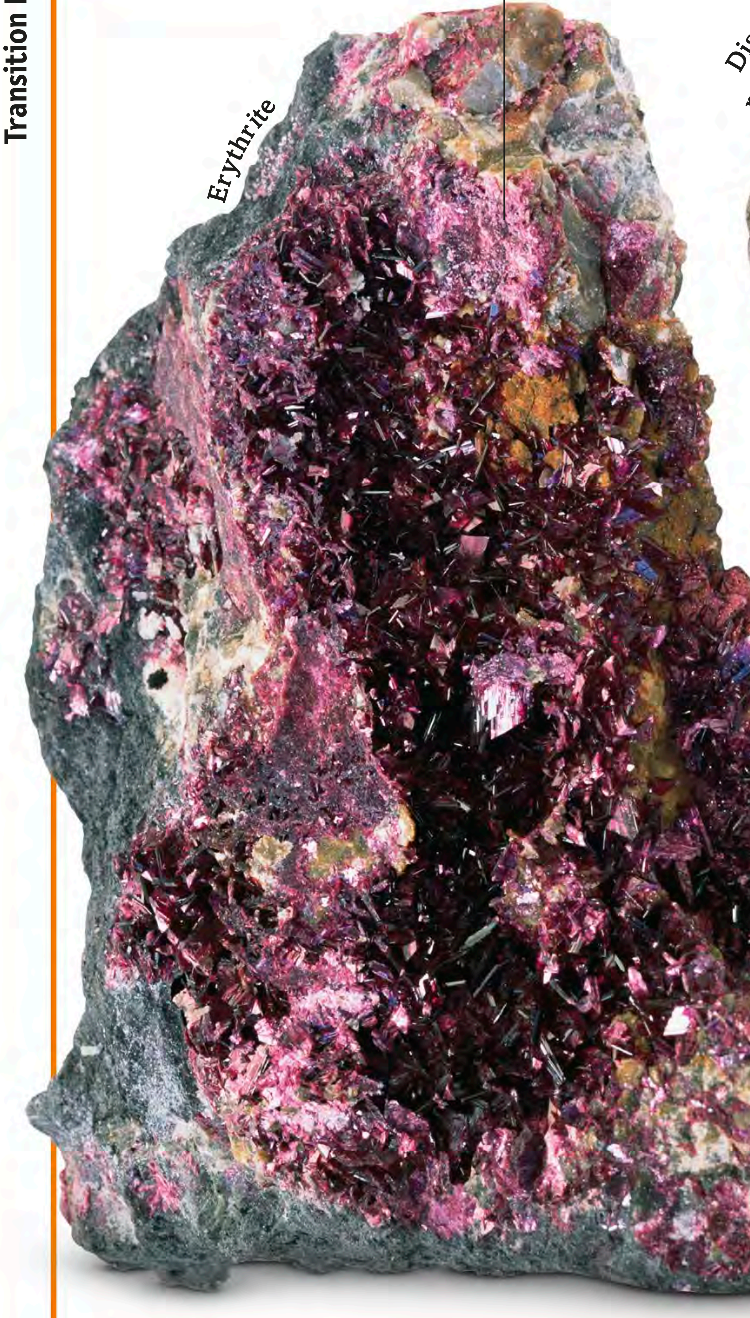

Erythrite

The distinctive purplish colour *gives it the nickname "red cobalt".*

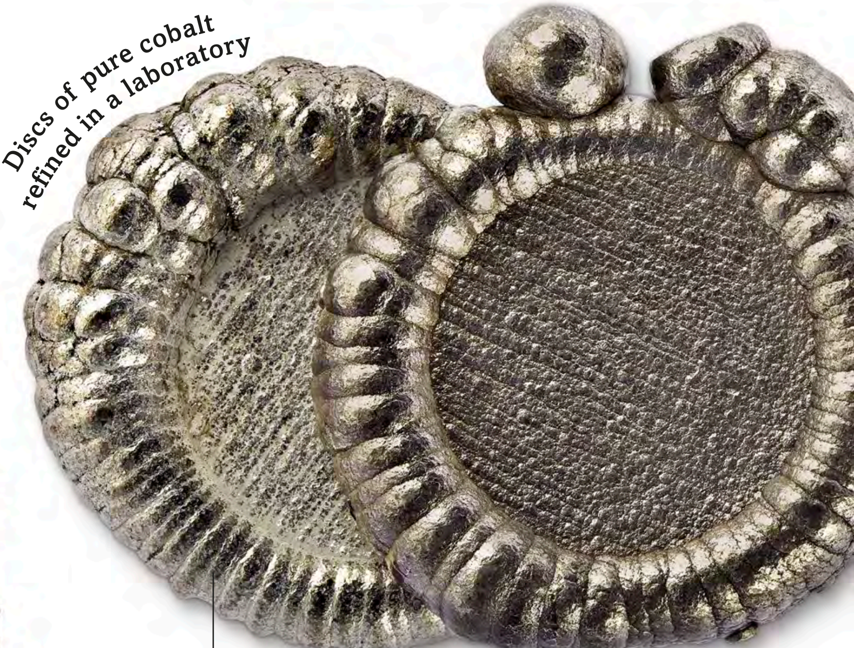

Discs of pure cobalt refined in a laboratory

This shiny metal *is fairly hard.*

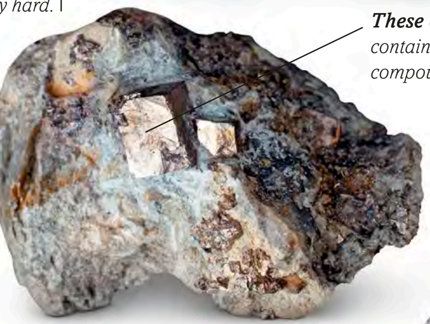

Cobaltite

These cubic crystals *contain a sulfur compound of cobalt.*

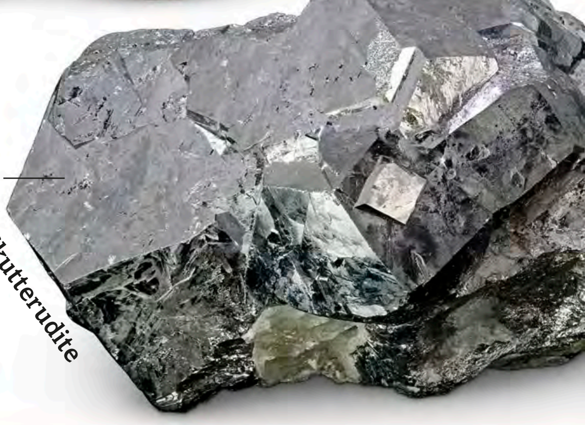

Skutterudite

This silvery mineral *smells like garlic when crushed due to the presence of arsenic.*

Medieval German miners often mistook ores of cobalt for precious metals. When they tried to purify these, the arsenic gas released made them sick. This unwanted side-effect led to the name *kobold*, which is German for "goblin", a mischievous spirit.

Pure cobalt is hard and shiny, and is added to steel and other alloys to make them stronger. Alloys containing cobalt are used in the blades of **jet engines** and in **artificial joints**, such as hip and knee joints. Cobalt is one of the few elements

Uses

Artificial hip joint

Tough and lightweight artificial joints *can be made from an alloy of cobalt and chromium.*

This part of the joint *is screwed into the hip bone.*

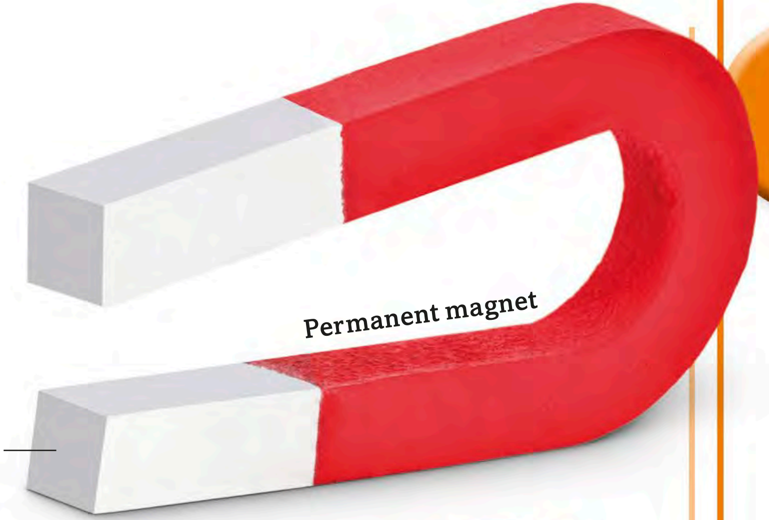

This magnet *can work at temperatures as high as 800°C (1,400°F).*

These blades *made of a cobalt alloy stay hard even when very hot.*

Jet engine turbine

Tiffany lamp

This blue glass *is manufactured using cobalt compounds.*

Cobalt blue paints have been in use **since 3000 BCE**.

Cobalt blue paint

This intense colour *does not fade easily over time or upon exposure to light.*

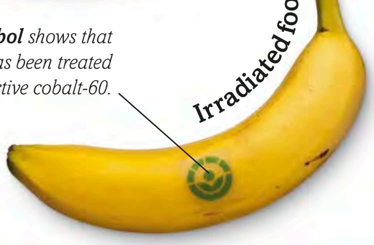

This symbol *shows that this fruit has been treated with radioactive cobalt-60.*

FORMING AN ISOTOPE

Cobalt-60 is an isotope, or form, of this element. It is created artificially in nuclear reactors. Because of its radioactivity, it is useful in some cancer treatments.

A neutron is fired at the cobalt-59 nucleus.

Added neutron

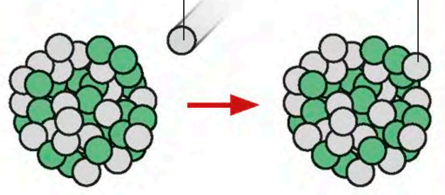

Cobalt-59 is a stable atom with 59 neutrons.

Cobalt-60 is a radioactive atom with 60 neutrons.

that can be used to make a permanent magnet. Large **permanent magnets** are made from a tough alloy of cobalt, nickel, and aluminium, called alnico. A radioactive form of cobalt, called cobalt-60, is produced in nuclear reactors. This form is widely employed to **irradiate food**, a process by which food is exposed to a tiny dose of radiation to kill harmful germs. Cobalt can also produce a deep shade of blue: **cobalt blue paints** and dyes are formed by reacting aluminium with cobalt oxide.

28
Ni

Nickel

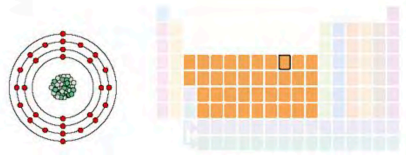

⊖ 28 ⊕ 28 ○ 30

State: Solid
Discovery: 1751

Forms

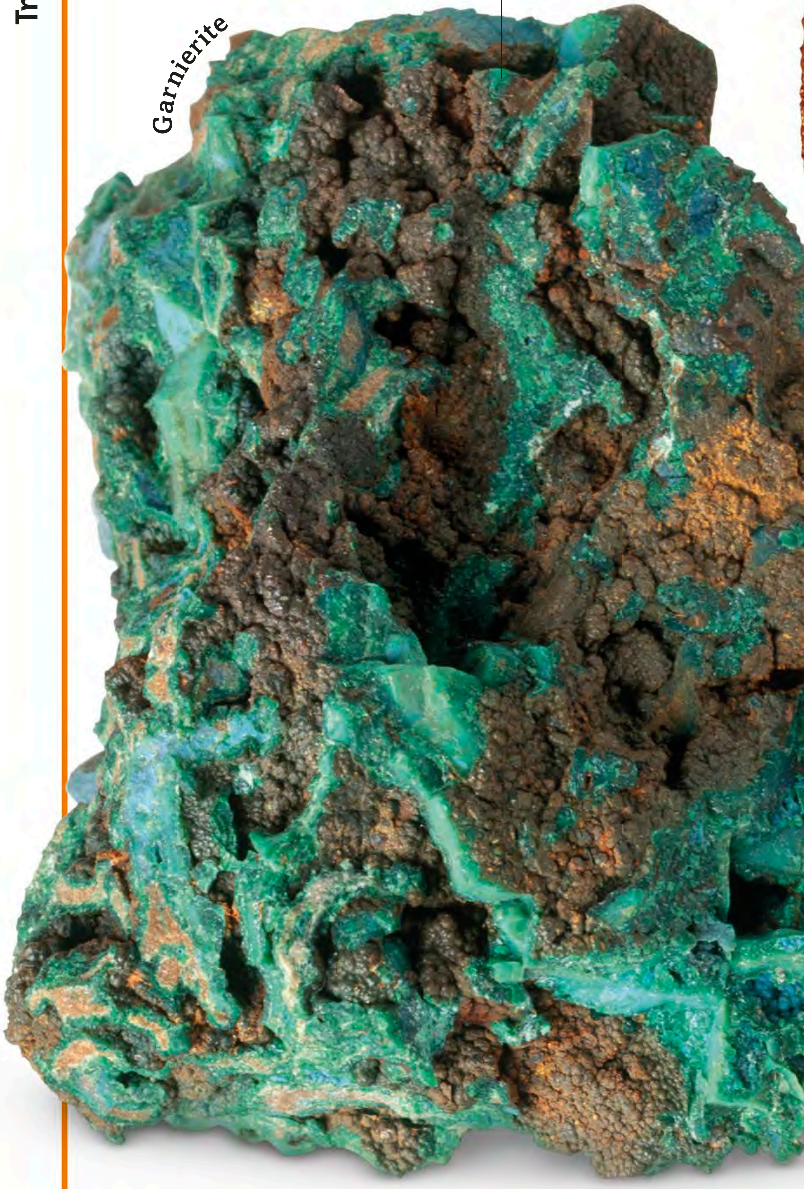

Garnierite

This green colour *comes from the presence of nickel.*

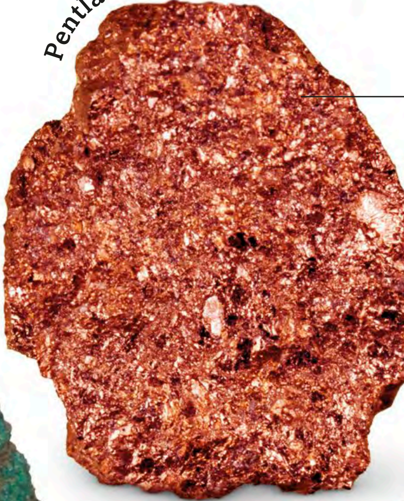

Pentlandite

This reddish mineral *is made of iron and nickel sulfide.*

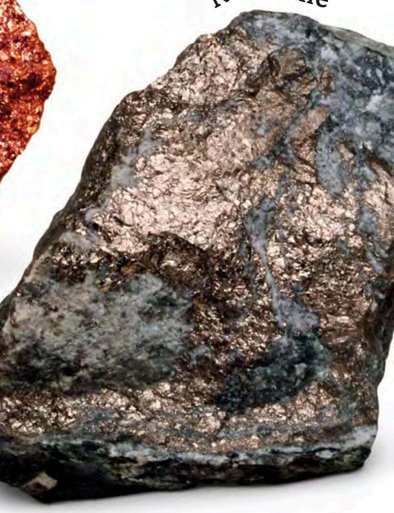

Nickeline

This nickel ore *also contains arsenic.*

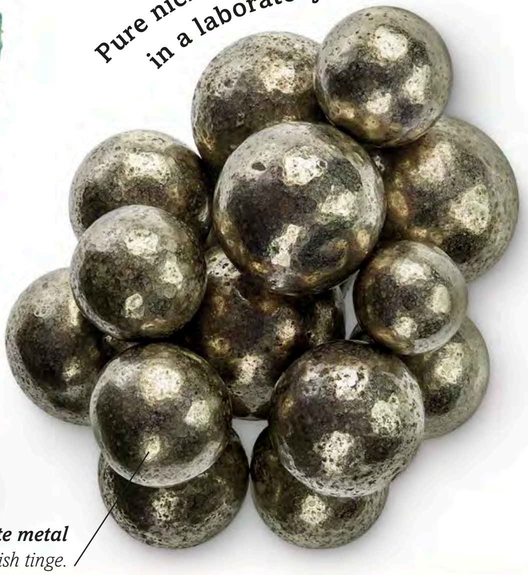

Pure nickel balls refined in a laboratory

These silvery white metal pellets *have a yellowish tinge.*

Nickel is named after Old Nick, a demonic spirit from Christian lore that was believed to live underground. In the 18th century, German miners mistook a poisonous nickel mineral, now known as **nickeline**, for a copper ore. When this mineral failed to yield copper, they named it *Kupfernickel*, meaning "Old Nick's copper". Nickel is also found in other ores, such as **garnierite** and **pentlandite**. This element is one of the most useful metals, with a number of applications. Because **pure nickel** does not rust, it is used to coat objects to make

Uses

Nickel goblet drum

This nickel-plated *instrument has a shiny finish.*

Rust-resistant handle

Nickel-plated dagger

Nickel-based coin

The US five cent coin *is made of an alloy of 75% copper and 25% nickel.*

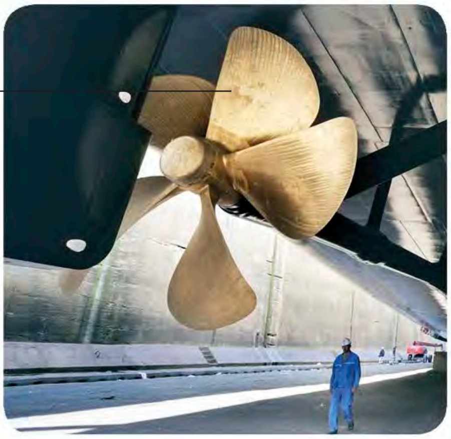

Ship propeller

Nickel-copper alloy plating *makes propellers strong and durable.*

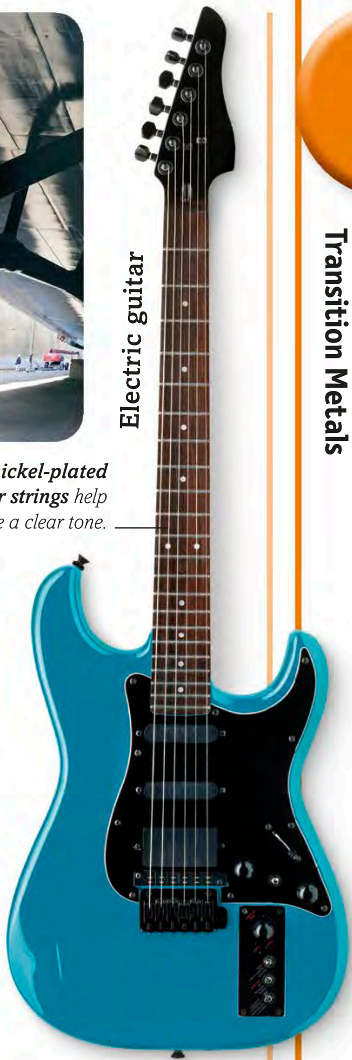

Electric guitar

These nickel-plated guitar strings *help create a clear tone.*

This **coin** from the **USA**, made of a nickel-copper alloy, is called a **nickel**.

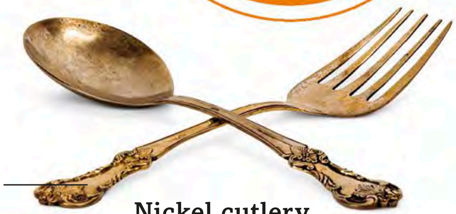

Nickel cutlery

This silver-coated fork *is made of an alloy of nickel, copper, and zinc.*

PERMANENT MAGNETS

A temporary magnet works when it is put in a magnetic field, but a permanent magnet retains its charge even when it is taken out of that field. Nickel is one of only a few elements that can be used to make permanent magnets.

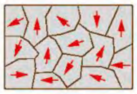

1*. Atoms in a piece of nickel metal are arranged randomly.*

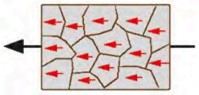

2*. When a magnetic field is applied, the atoms align in one direction.*

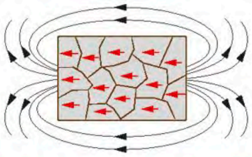

3*. The atoms then stay aligned and produce a magnetic field of their own, even after the external field is removed.*

Toaster

Nickel alloy wires *heat up to make toast.*

them look like silver – a trick still used to make inexpensive ornamental objects. Nickel is also mixed with copper to make an alloy called cupronickel. This is used as plating on **propellers** and other metallic parts of ships, as the alloy does not corrode in seawater. The same alloy is used in most of the world's silver-coloured **coins**. Nickel is used in the strings of **electric guitars**. This element is added to chromium to make an alloy called nichrome. Wires made of this alloy conduct heat very well, so are used in **toasters**.

29 Cu Copper

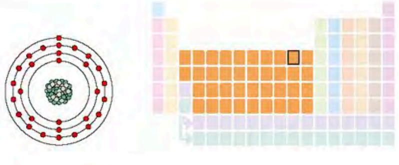

⊖ 29 ⊕ 29 ○ 35

State: Solid
Discovery: Prehistoric

Forms

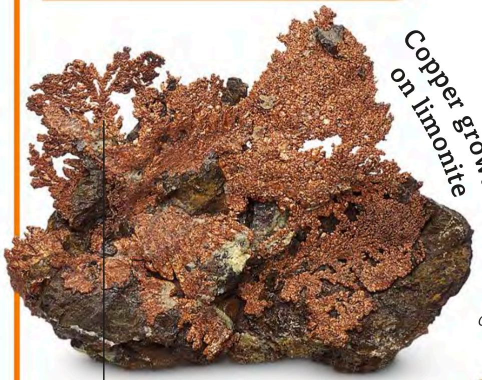

Copper growth on limonite

Branchlike crystals of copper

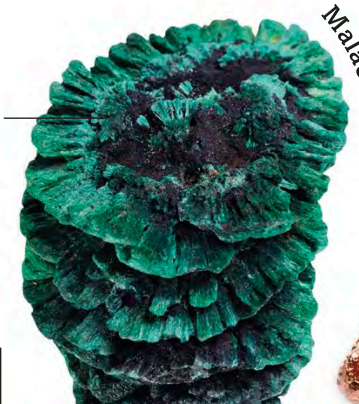

Malachite

In caves, feather-like crystals are often formed.

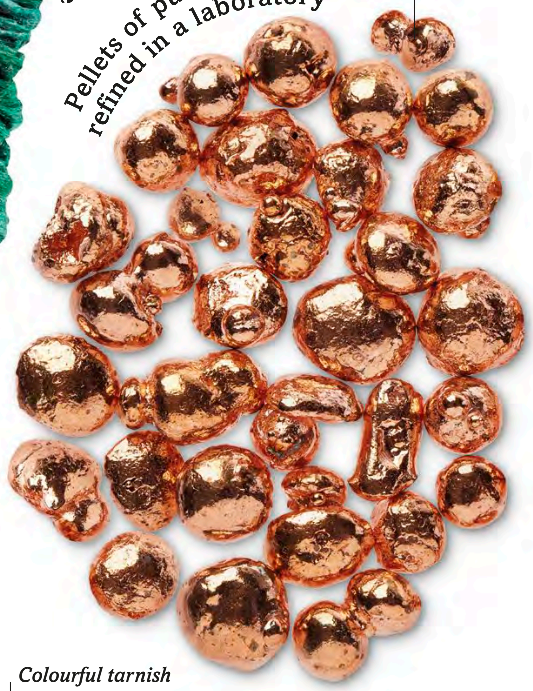

Pellets of pure copper refined in a laboratory

Unique reddish orange colour

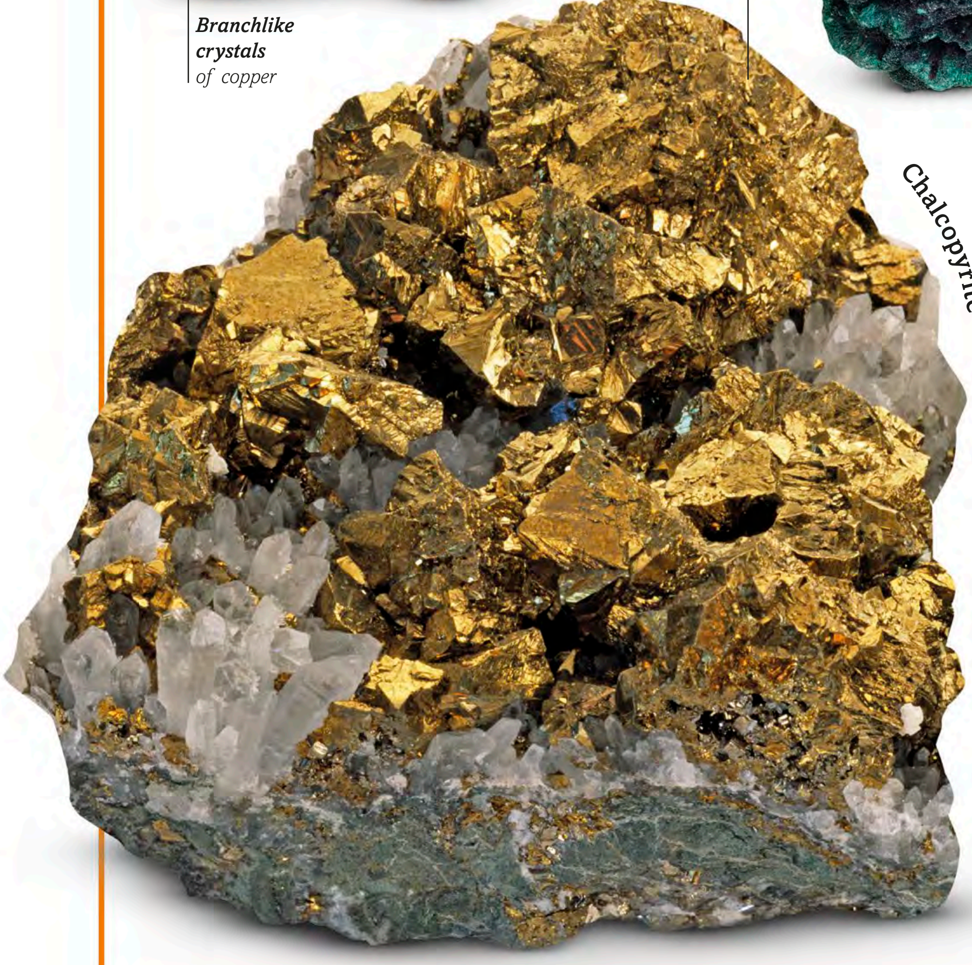

Chalcopyrite

These golden yellow crystals contain copper sulfide.

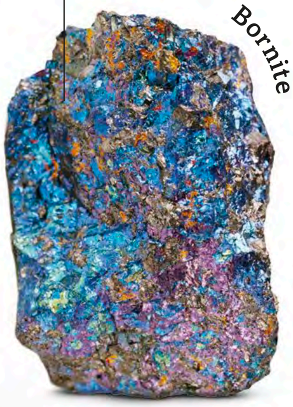

Bornite

Colourful tarnish develops when the mineral reacts with air.

Crustacean blood

Crustacean blood is blue because it contains copper.

Copper is a soft, bendy metal that is an excellent conductor of electricity and heat. Although it is one of the few elements found pure in nature, most of it exists in ores such as **chalcopyrite**. Other copper minerals, such as **malachite** and azurite, are brightly coloured.

Copper is the only metal that has a reddish colour in its pure form. **Pure copper** is mainly used as **wires** in electrical equipment. Copper wire wrapped around an iron core and then electrified helps create an **electromagnet**. Because they can be switched on or off, electromagnets can

Uses

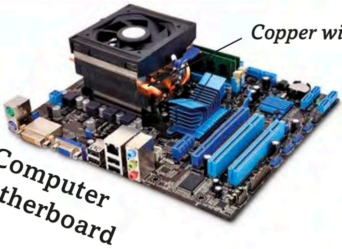

Copper wire

Computer motherboard

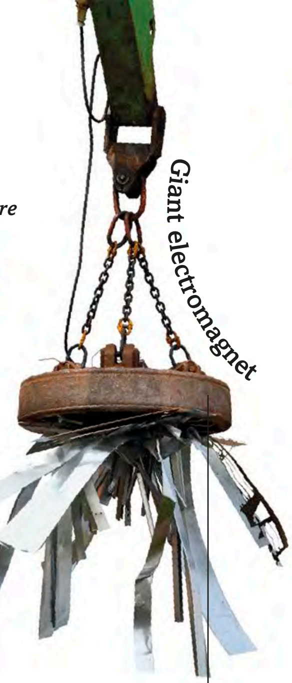

Giant electromagnet

Massive crane-mounted electromagnets *contain huge coils of copper wire.*

Electroplated nails

Copper coating on steel *slows down corrosion.*

ELECTROPLATING

A thin layer of copper can be added to metal, usually iron, to stop it from corroding easily. This process is called electroplating.

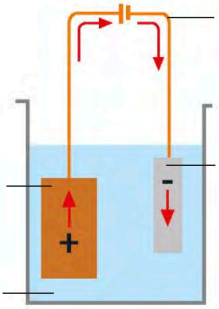

Electrons flow from the positive to the negative electrode.

This positive electrode is made of copper and slowly dissolves in the liquid.

This solution contains dissolved copper ions.

This electrode is made of iron. Copper ions form a protective coating over the metal.

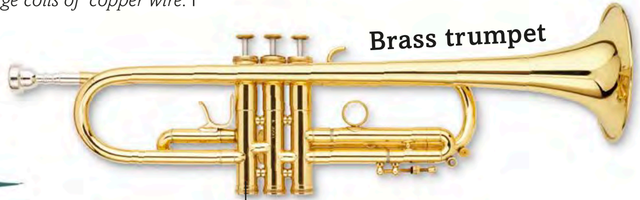

Brass trumpet

The brass tube *contains air that vibrates to produce musical sounds.*

Copper-glazed vase

A copper glaze *on the vase gives it a metallic shine.*

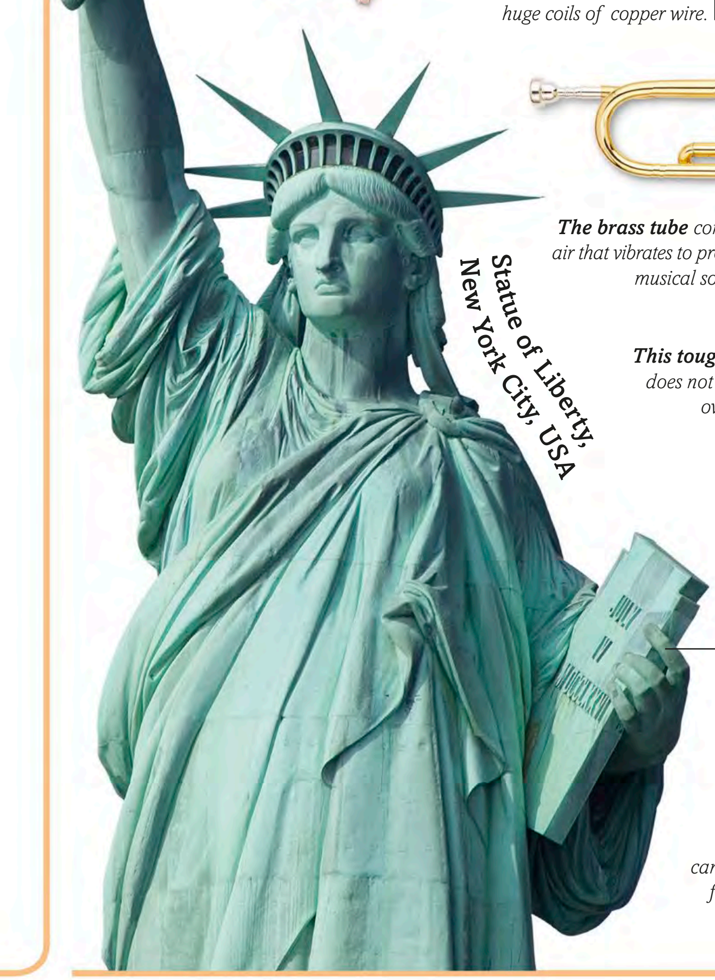

Statue of Liberty, New York City, USA

A verdigris layer *protects pure copper from additional weathering.*

Bronze helmet

This tough alloy *does not weaken over time.*

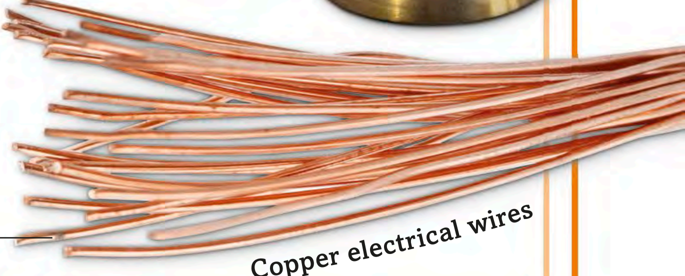

Copper electrical wires

Pure copper *can be stretched to form long wires.*

be magnetic as and when they are needed. They can be much more powerful than normal magnets and can lift heavy objects. Pure copper does not rust, but it reacts with air over time to form a layer of grey-green copper carbonate called verdigris. This can be seen on copper statues, such as the **Statue of Liberty**. Copper is often mixed with other metals to produce tougher alloys. Bronze, a copper-tin alloy, is more durable than pure copper and has been used since ancient times. Brass, a copper-zinc alloy, is used in musical instruments, such as **trumpets**.

COPPER WIRES Not much thicker than a human hair, these copper wires are twisted together and woven into a tight bundle. One of the main uses for these wires is to shield a thicker copper wire that transmits a signal to a television. As the signal carries pictures and sounds in the form of electrical currents, the wires wrapped around it prevent interference from other electrical sources nearby.

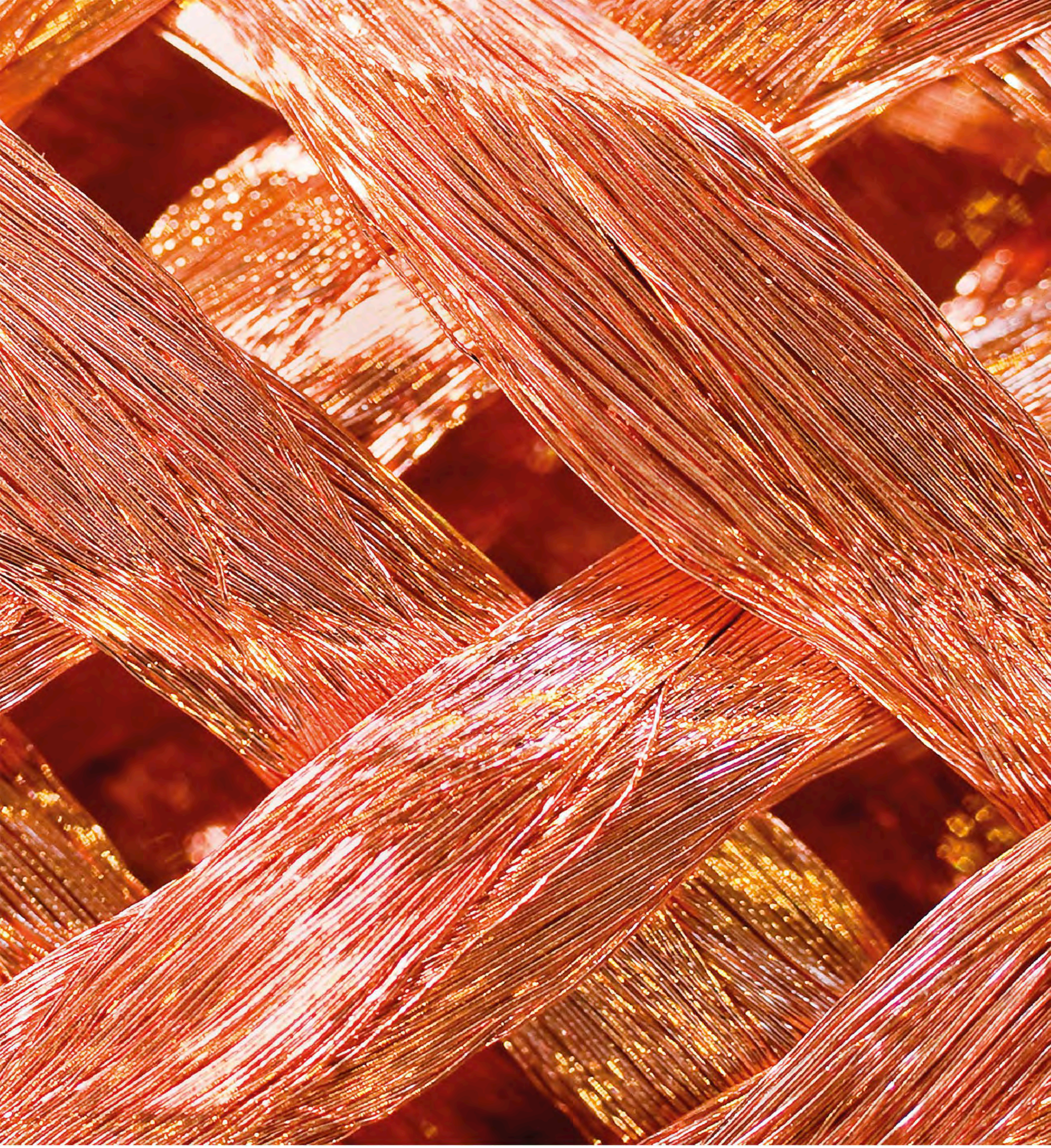

Copper is a very good conductor of electricity, but not the best; silver is better. However, copper is more widely used because it is much cheaper to find and purify. Each year, about 15 million tonnes of pure copper is produced, and more than half of it is used to make electrical components, such as this mesh. Today, more than a billion kilometres of copper wiring is running unseen in power supplies, buildings, and electronics. Copper is now the most common electrical metal, but it has a long history. It was the first element to be refined from ores in large amounts about 7,000 years ago in the region that is now Iraq. Today, Bingham Canyon in Utah, USA, is the world's largest copper mine.

30 Zn Zinc

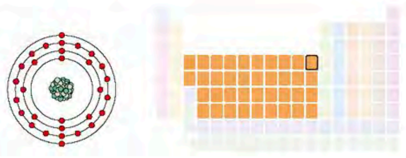

● 30 ⊕ 30 ○ 35

State: Solid
Discovery: 1746

Forms

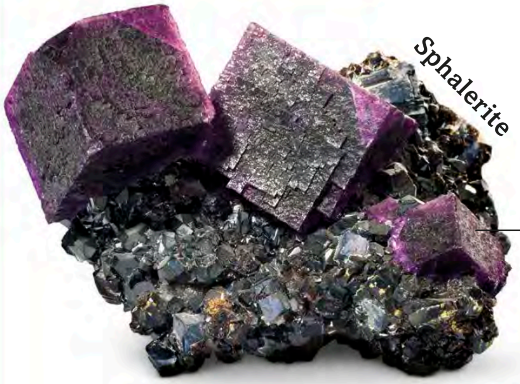

Sphalerite

***This ore** is the main source of zinc.*

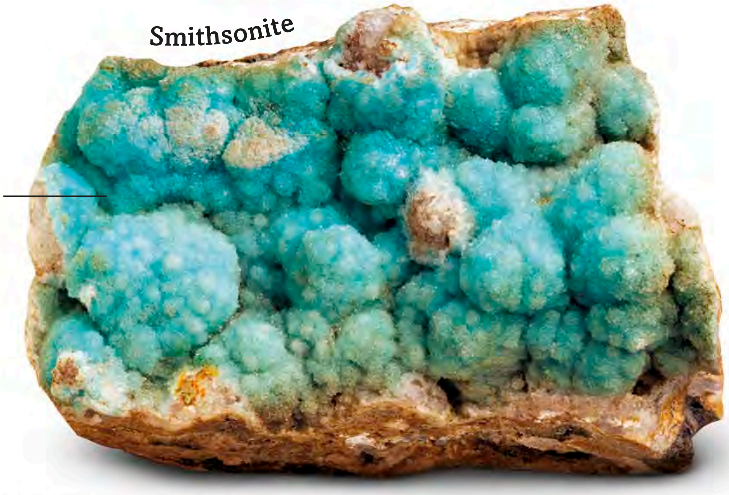

Smithsonite

***This zinc mineral** forms rough nodules inside cracks.*

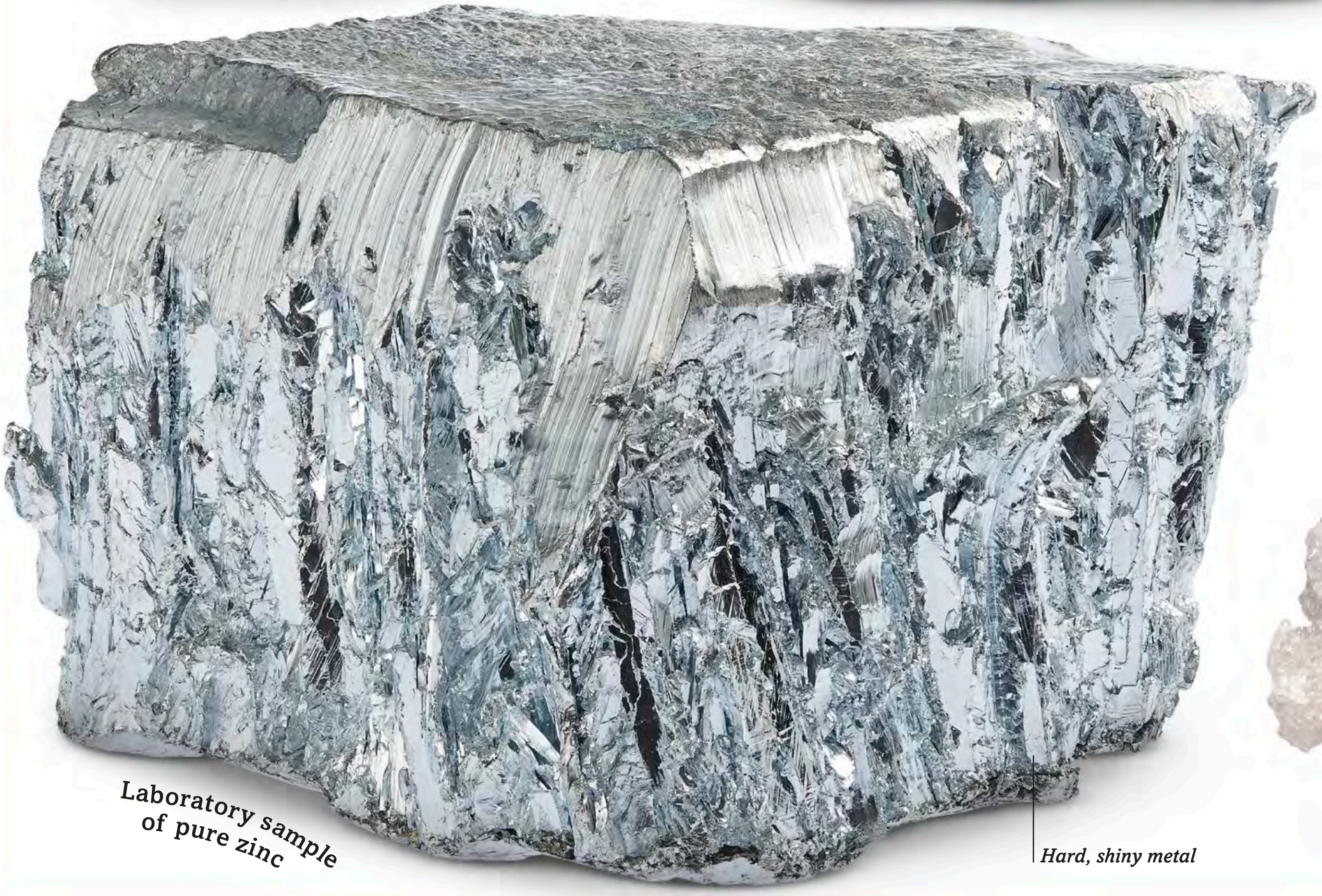

Laboratory sample of pure zinc

Hard, shiny metal

Zinc was used in India and China hundreds of years before the German chemist Andreas Marggraf identified it as a new element in the 18th century. This element is a rare transition metal that is never pure in nature, but is found in many minerals.

The mineral **sphalerite**, containing zinc sulfide, is the major source of **pure zinc**. Another principal mineral, **hemimorphite**, contains zinc and silicon. Zinc is essential in our diet. We consume it from food such as cheese and sunflower seeds. Zinc compounds have a wide

Supernova

Along with many other elements, zinc atoms *are formed inside supernovae (exploding giant stars).*

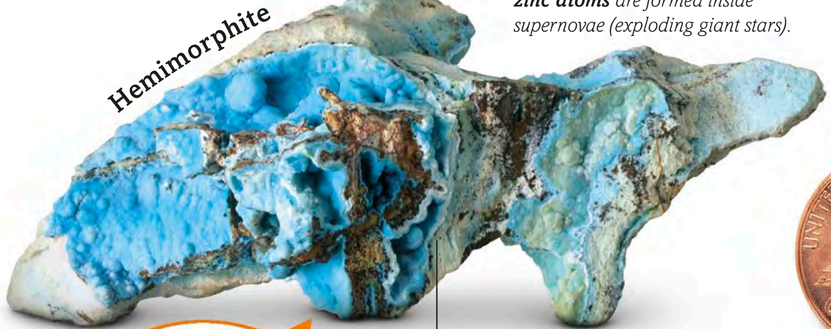

Hemimorphite

This mineral *contains zinc carbonate.*

Hemimorphite was discovered by James Smithson, the founder of the **Smithsonian** Institution.

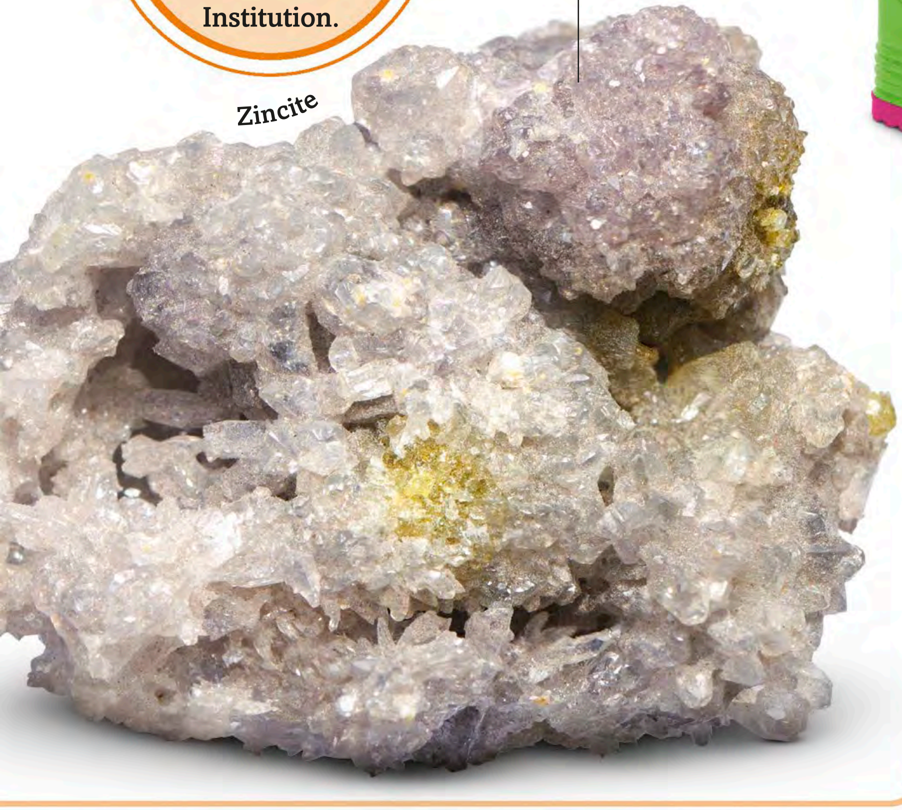

Zincite

Zinc oxide crystals *are generally colourless.*

Uses

First-aid tape

Medical tapes *that contain zinc oxide stop wounds from getting infected with microbes.*

Akashi Kaikyo Bridge, Kobe, Japan

The zinc coating *on this steel bridge protects it from rust.*

American penny

This zinc coin *is coated with copper.*

Rubber boots

This flexible rubber *is made stronger by adding zinc oxide.*

Calamine lotion

This soothing skin lotion *contains a mixture of zinc compounds.*

GALVANIZED STEEL

Steel is protected from corrosion by coating it with zinc. Alloys of iron and zinc form in layers between the steel and pure zinc. This process is called galvanization.

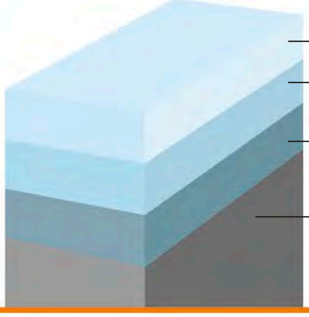

Pure zinc

94% zinc and 6% iron

90% zinc and 10% iron

Steel is an alloy of iron and carbon.

range of applications. For example, a compound of zinc and oxygen called zinc oxide is used in **medical tape** and sunscreen. Zinc oxide can also be used to toughen the rubber used in **boots** and tyres. A compound of zinc and sulfur called zinc sulfide is used to make some paints that glow in the dark. When pure zinc is exposed to air, the metal reacts with oxygen to form a protective layer of an oxide. This coating can prevent objects covered in zinc, such as **bridges**, from corroding easily.

39
Y

Yttrium

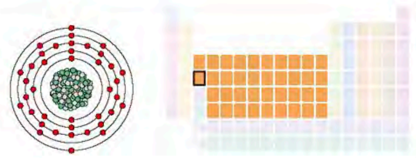

⊖ 39 ⊕ 39 ○ 50

State: Solid
Discovery: 1794

Forms

This mineral *contains trace amounts of uranium, which is radioactive.*

Yttrium is **400 times more common** in Earth's crust than silver.

This rock sample *was brought to Earth by NASA's Apollo 16 mission.*

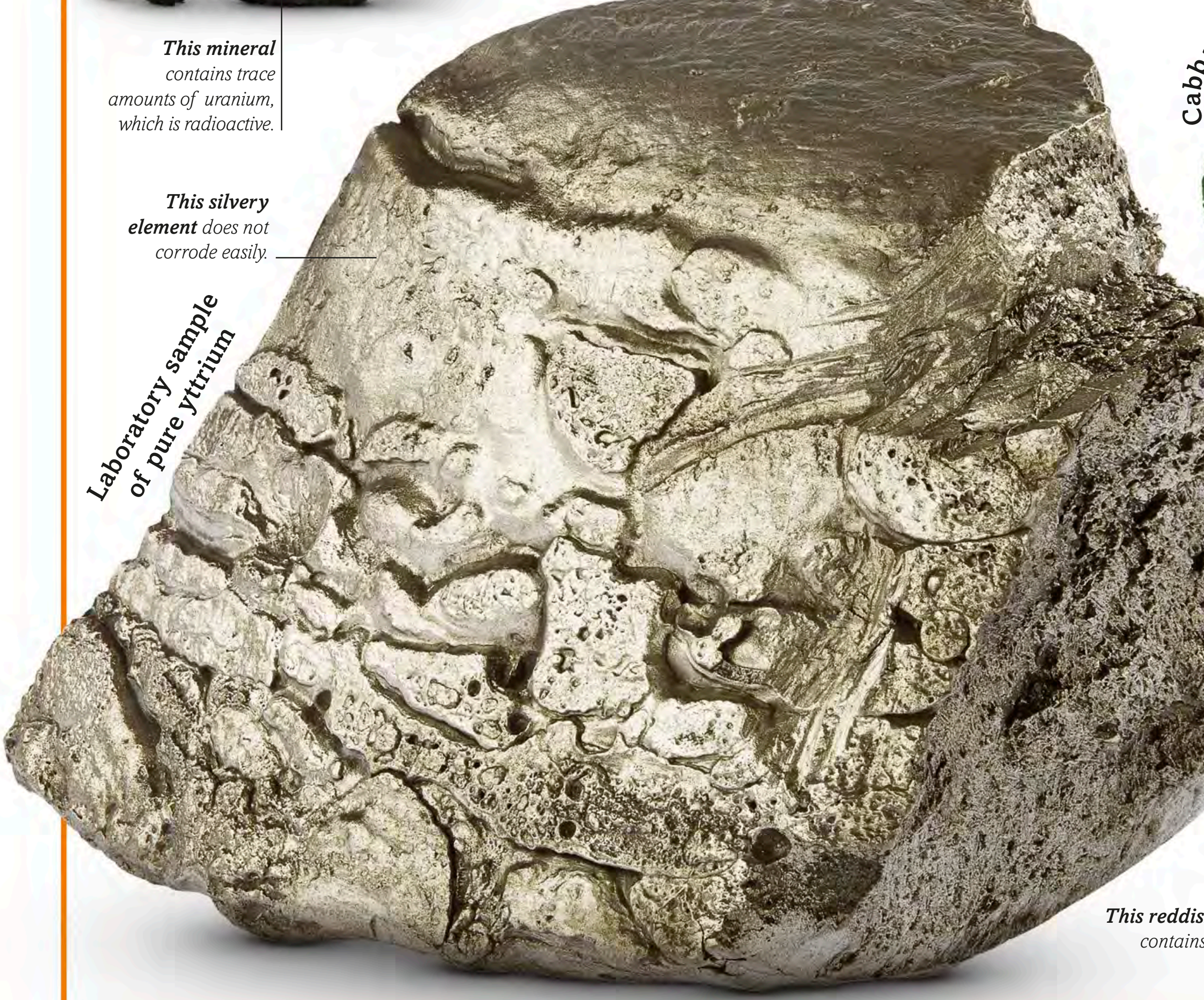

This silvery element *does not corrode easily.*

This vegetable *contains yttrium.*

This reddish-brown mineral *contains about 2% yttrium.*

The samples of rock brought back from the Moon by astronauts in NASA's Apollo missions contained higher levels of yttrium than rocks on Earth. This element is never found in pure form in nature, but small traces of it are present in many minerals, including **xenotime** and **monazite**. Yttrium was discovered in a compound in 1794 by the Finnish chemist Johan Gadolin, but it wasn't isolated until 1828. Other yttrium compounds have since been found in vegetables, including **cabbage**, and in seeds of woody plants. In

Uses

LED lamp

This bulb *contains yttrium compounds that enable it to produce a warmer, more yellow glow.*

Laser

This laser *is powered using crystals composed of yttrium and silicon, and can cut through metal.*

FRIEDRICH WÖHLER

In 1828, the German chemist Friedrich Wöhler became the first person to purify yttrium. He did so by extracting it from the compound yttrium chloride. He was also the first person to extract the metals beryllium and titanium from their ores.

Yttrium-90

This radioactive form of yttrium *is used to treat cancers in the body.*

Yttrium gas mantle

This fabric mantle *holds the hot flame inside.*

Digital camera lens

Shock-proof lenses *are made from yttrium-infused glass to make them tough.*

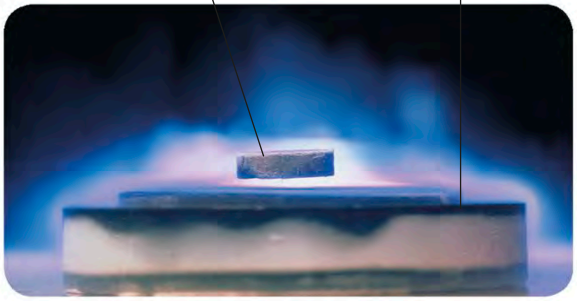

A small magnet *floats over the superconductor.*

This superconductor *produces a magnetic field that repels the magnet above it.*

Yttrium superconductor

NASA spacecraft use yttrium lasers to map the **surface of asteroids** in space.

LED lamps, yttrium converts blue light to other colours. Many **lasers** use an artificial mixture of yttrium and aluminium inside a silicon-rich crystal called garnet. Powerful yttrium lasers are used for treating some skin infections, as well as by dentists during tooth surgery. A radioactive form of this element has medical applications. Yttrium is added to the glass in a **camera lens** to make it tough. Yttrium compounds are also used in **superconductors** – materials that conduct electricity easily when cooled to very low temperatures.

40
Zr

Zirconium

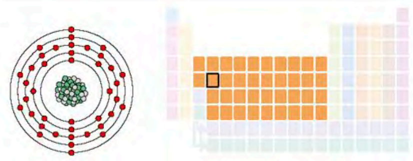

State: **Solid**
Discovery: **1789**

Forms

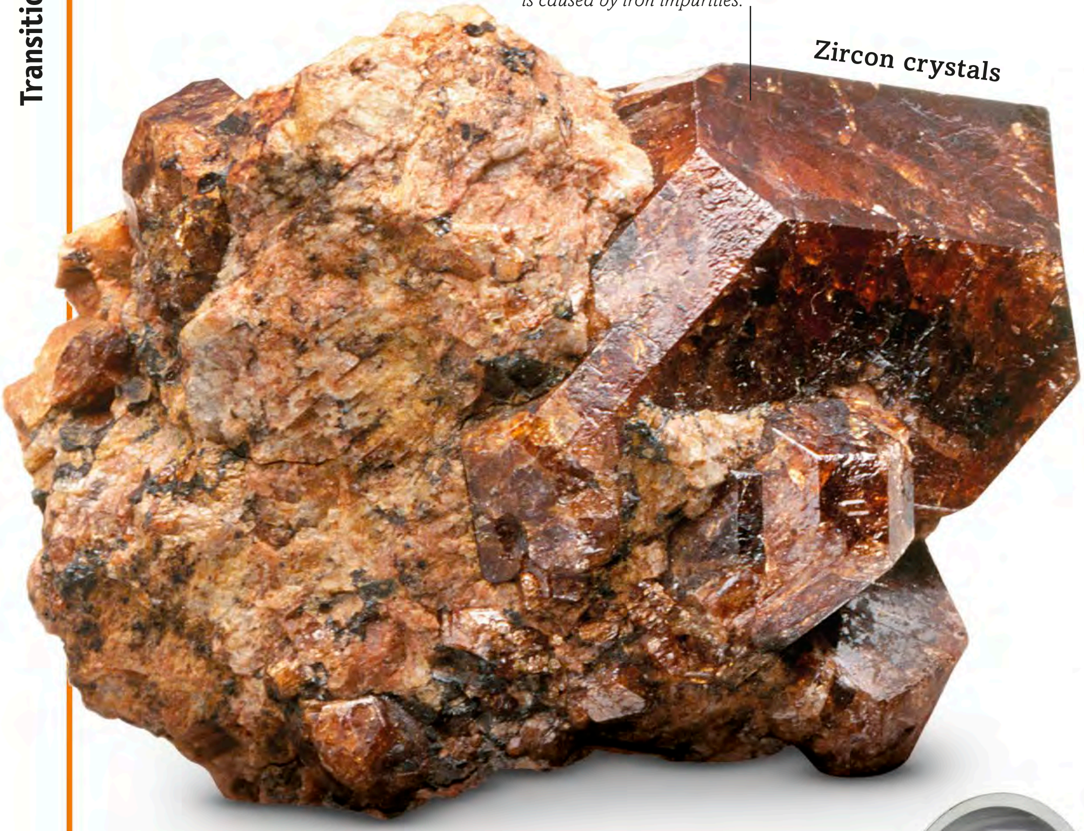

Zircon crystals

This dark brown colour *is caused by iron impurities.*

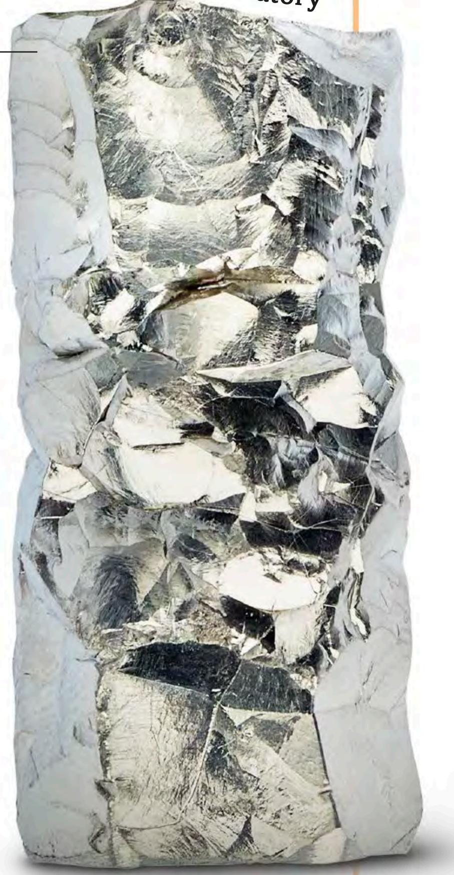

Bar of pure zirconium refined in a laboratory

This greyish-white pure element *is easy to shape.*

Uses

Dental crowns

These tough dental crowns *are made of a zirconium-rich ceramic.*

Ceramic knife

This hard, non-metallic blade *does not require frequent sharpening.*

Camera flash from the 1960s

This zirconium-filled bulb *produces a bright light.*

Zirconia crystal ring

This ring *contains cubic zirconia crystals.*

This element is named after the mineral zircon, which means "golden" in Persian, a reference to the golden-brown colour of its crystals. The Swedish chemist Jacob Berzelius was the first person to isolate **pure zirconium**, in 1824. Today, however, the element is mostly used in the form of the compound zirconium dioxide, or zirconia. Powdered zirconia is heated to produce a hard glass-like ceramic, which is used to create **dental crowns** and sharp **ceramic knives**. Powdered zirconia also forms sparkling **zirconia crystals** that look like diamonds.

41

Nb

Niobium

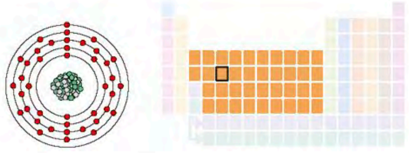

⊖ 41 ⊕ 41 ○ 52

State: Solid
Discovery: 1801

Forms

Columbite

This dark, dense ore *has a light metallic shine.*

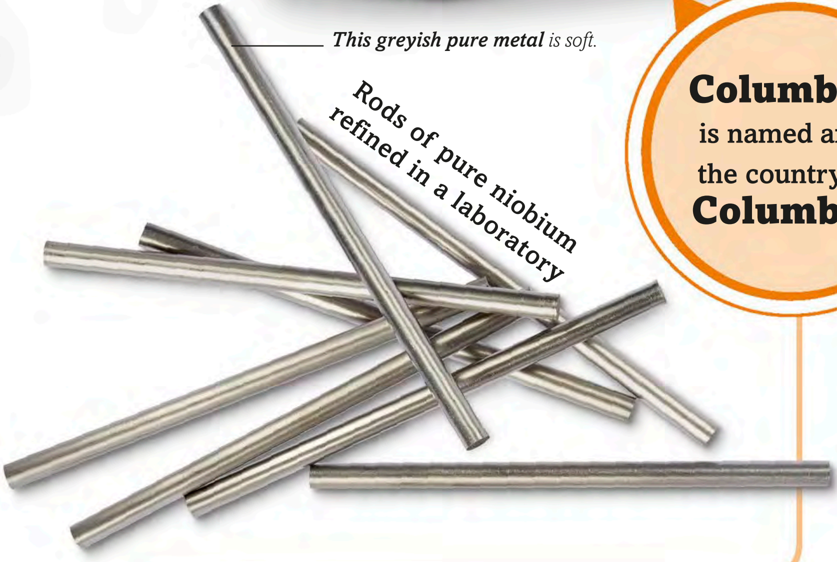

Rods of pure niobium refined in a laboratory

This greyish pure metal *is soft.*

Uses

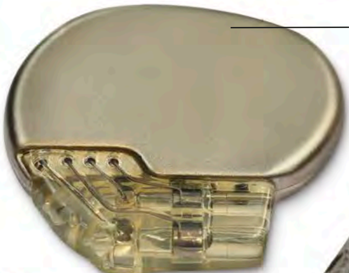

Pacemaker

The niobium case *houses a battery that regulates the beating of a human heart.*

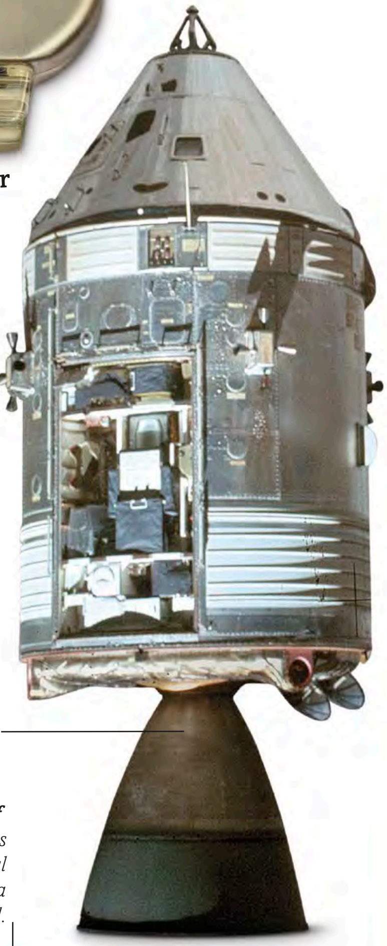

Command Module from the Apollo 15 mission

This nozzle *made of a niobium alloy kept its shape even when very hot.*

Spectacle lenses

This pair of spectacles *has thin, powerful lenses made of a niobium compound.*

Columbite is named after the country of **Columbia**.

Niobium is so similar to the metal tantalum that the two were wrongly thought to be the same element for almost 40 years. The mineral **columbite** is the main source of this shiny metal. Niobium is not found naturally in its **pure form**. When extracted, it has many uses. As the element does not react adversely in the human body, it is used in implants, such as **pacemakers**. Niobium also does not expand when hot, so it is used to make parts of rockets, such as the one on the **Command Module** from NASA's Apollo 15 spacecraft that went to the Moon in 1971.

42
Mo

Molybdenum

Forms

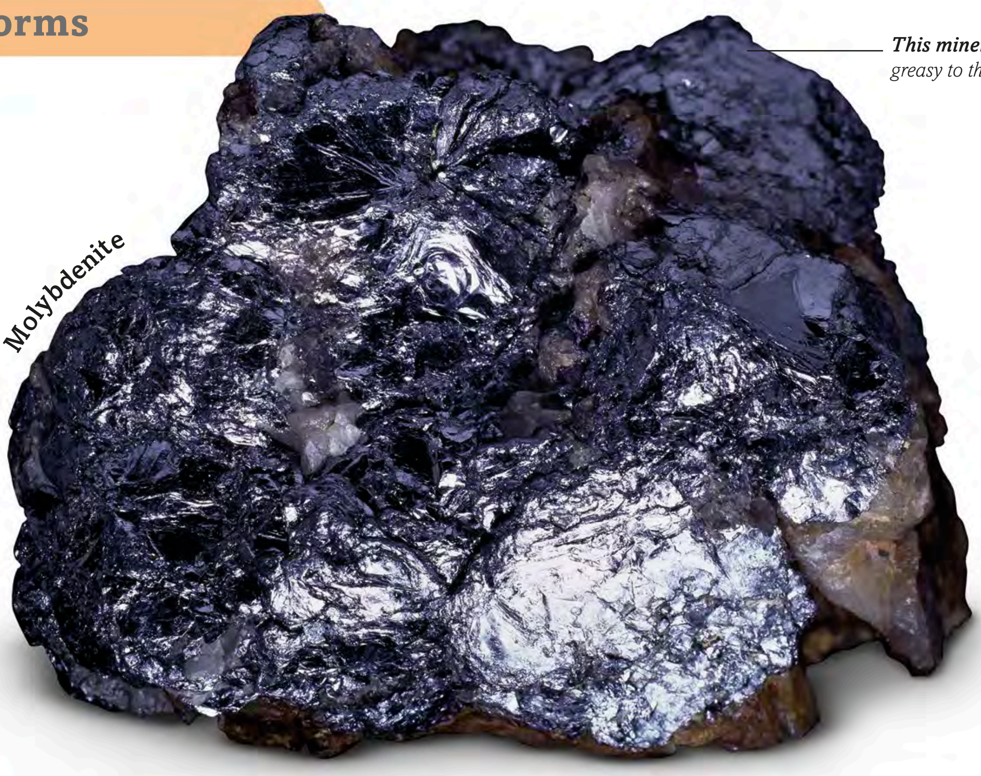
Molybdenite

***This mineral** feels greasy to the touch.*

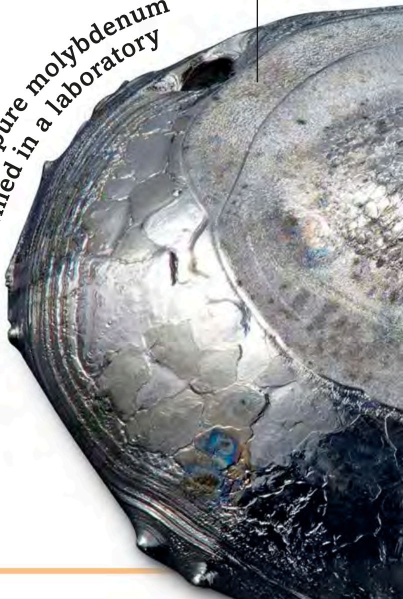
Chunk of pure molybdenum refined in a laboratory

***This metal's pure form** is silver-grey and has a very high melting point: 2,623°C (4,753°F).*

Uses

Lubricant

***This slippery lubricant**, which contains finely powdered molybdenite mixed with oil, protects fast-moving mechanical parts in engines.*

Chrome-Moly steel bike

***This lightweight but stiff frame** is made from a steel containing molybdenum and chromium.*

Ratchet set

***These closely fitting parts** are hard so they do not get damaged easily.*

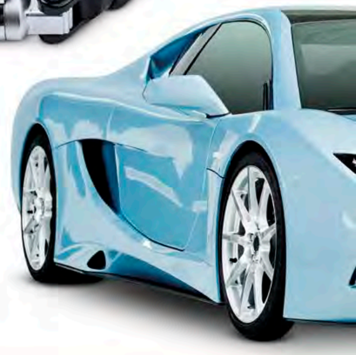

Molybdenum gets its unusual name from the Greek word *molybdos*, which means "lead". Miners once mistook **molybdenite**, a dark mineral containing this metal, for an ore of lead. This element is much harder than lead, so it is easy to distinguish between these two elements when they are pure. Molybdenite is soft and slippery, and it is the main molybdenum ore. **Pure molybdenum** is mainly used to make alloys that are resistant to corrosion. These are lightweight so are ideal for constructing **bike frames**, but

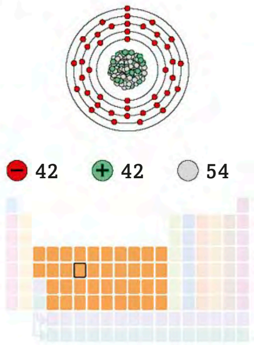

⊖ 42 ⊕ 42 ○ 54

State: Solid
Discovery: 1781

This experimental sports car is built with a rust-resistant alloy that contains molybdenum.

Vencer Sarthe

are hard enough for making sturdy tools, too. Molybdenum alloys are used in the latest designs of supercars, such as the **Vencer Sarthe**.

43 Tc Technetium

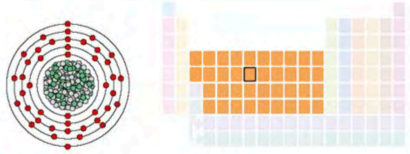

⊖ 43 ⊕ 43 ○ 55

State: Solid
Discovery: 1937

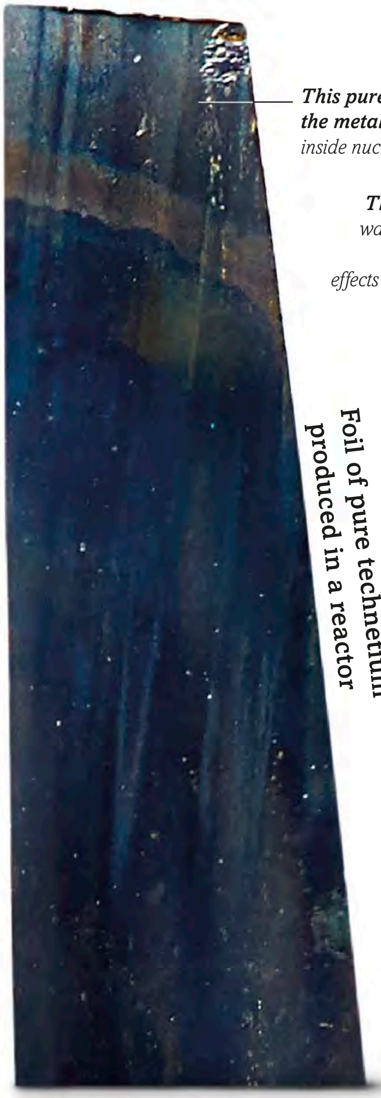

Foil of pure technetium produced in a reactor

This pure form of the metal *is produced inside nuclear reactors.*

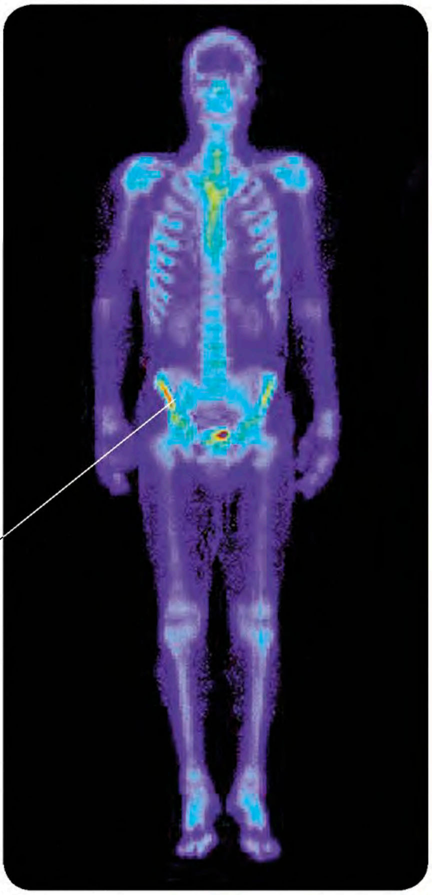

Technetium-based imaging

This body scan *was created using the radioactive effects of technetium.*

Generating technetium

This box *contains radioactive molybdenum, which breaks down into technetium.*

Technetium was the first element to be produced artificially by researchers. It is named after the Greek word for artificial, *tekhnetos*. Technetium does not exist in nature: any of its atoms that once existed on Earth broke down millions of years ago. Tiny amounts of this element were discovered in the waste produced by early nuclear reactors. Technetium is the lightest radioactive element. It is used extensively in **medical imaging**. It is injected into a patient's body, where it emits radiation for a short while. Some machines use this radiation to show bones clearly.

44
Ru
Ruthenium

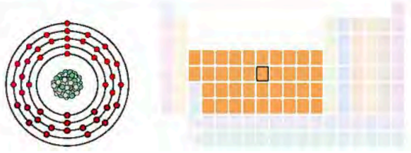

⊖ 44 ⊕ 44 ○ 57

State: Solid
Discovery: 1844

Forms

Pentlandite

***This yellow-brown mineral** is often found deep underground.*

Pure ruthenium refined in a laboratory

***These crystals** have a bright, silver colour.*

Uses

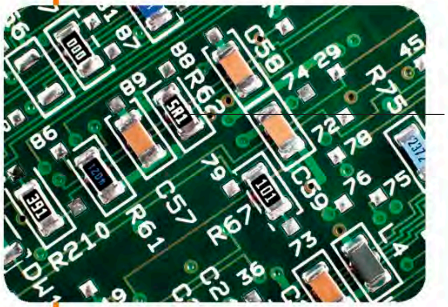

***This component** contains ruthenium dioxide.*

Electronic circuit board

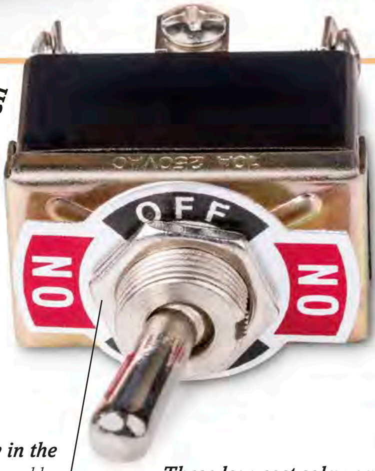

***The metal alloy in the switch** is toughened by adding ruthenium.*

***These low-cost solar panels** are made using ruthenium.*

SwissTech Convention Centre, Switzerland

Ruthenium is named after *Ruthenia*, an old Latin name for Russia. This rare metal is found in the mineral **pentlandite**, and its **pure form** is commonly extracted from this ore. A compound called ruthenium dioxide is used in several components in **electronic circuits**, including resistors and microchips for computers and other digital devices. Adding a small amount of ruthenium makes softer metals, such as platinum and palladium, much tougher. Moving parts in devices such as **switches** benefit from this property.

45
Rh
Rhodium

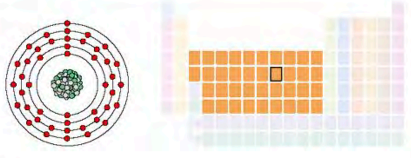

⊖ 45 ⊕ 45 ○ 58

State: Solid
Discovery: 1803

Forms

Pure rhodium pellet refined in a laboratory

***This metal** is shiny and silvery when pure.*

Millerite

***This golden mineral** is defined by its needle-shaped crystals.*

Uses

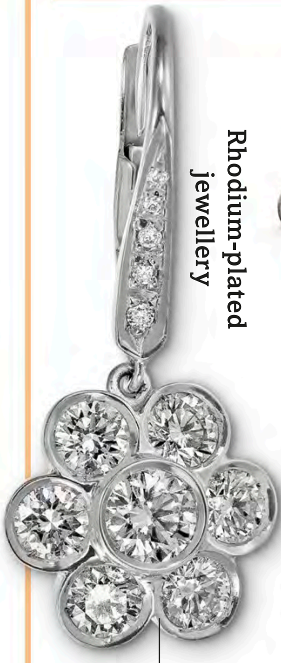

Rhodium-plated jewellery

***The rhodium plating** prevents jewellery from losing its shine.*

Headlight reflector

***This rhodium-alloy reflector** provides a bright light.*

Rhodium-plated microscope

***Parts of this microscope** are coated with rhodium and can resist corrosion.*

Fibreglass production

***These fibreglass strands** are made by passing molten glass through rhodium-enriched trays.*

The rosy red colour of one of its compounds inspired the name rhodium. The Greek word *rhodon* means “rose-coloured”. Rhodium is unreactive and does not form compounds easily. It is a rare metal. Most of the **pure form** is extracted when platinum is mined. Pure rhodium is hard and is used to toughen precious **jewellery**, mirrors, and optical devices, such as **microscopes**. It is mainly used in the production of catalytic convertors for cars. **Fibreglass**, which is often found in protective gear – like helmets – also contains rhodium.

46

Pd

Palladium

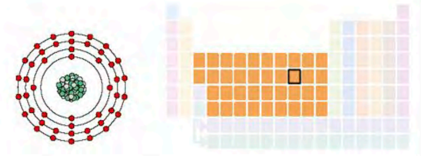

⊖ 46 ⊕ 46 ○ 60

State: Solid
Discovery: 1803

Forms

The mines *in this area have a high concentration of palladium.*

Blue Ridge mines, South Africa

Palladium absorbs **hydrogen**, like a sponge soaking up water.

Pure form *can be produced by separating it from the ores of other metals, such as copper and nickel.*

Pure palladium pellet refined in a laboratory

Uses

The converter *gets hotter as more pollutants enter the exhaust.*

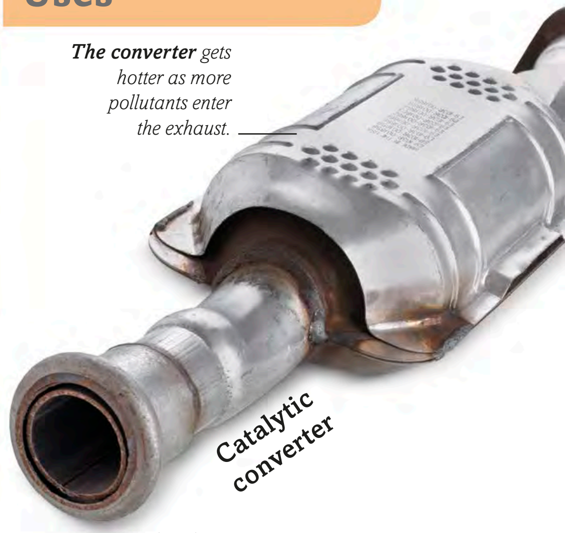

Catalytic converter

This device *uses palladium compounds that change colour when poisonous carbon monoxide is detected, which triggers an alarm.*

CARBON MONOXIDE ALARM
TEST/RESET
Power
Alarm

Carbon monoxide detector

This commemorative coin *is made from the palladium produced by the Stillwater Mining company in Montana, USA.*

STILLWATER PALLADIUM
LEWIS & CLARK CORPS OF DISCOVERY
2004

Palladium coin

Palladium is a rare, precious metal: it is 10 times rarer than silver and twice as rare as gold. Like these metals, palladium has a shiny surface and does not corrode easily. Palladium is found **pure** in nature, but it also has a few rare minerals, such as braggite. Of its many applications, the element's main use is in **catalytic converters**, which are devices used in vehicles to convert poisonous exhaust gases into less harmful ones. A compound called palladium chloride is used in **carbon monoxide detectors**. Because the element is

CATALYTIC CONVERTER

Many cars contain engines that are attached to catalytic convertors. These important devices convert toxic exhaust gases into less harmful pollutants. Palladium plays a key role in the process.

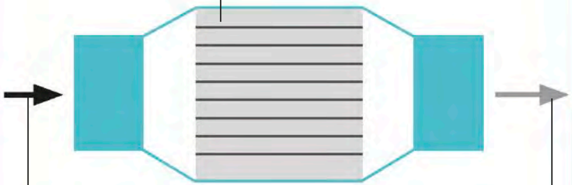

***2**. A chemical reaction occurs in the palladium mesh, which reduces harmful pollutants.*

***1**. Polluting gases enter the converter.*

***3**. Gases with less harmful fumes exit through the exhaust pipe.*

White gold ring

***Palladium is added to gold** to remove its colour.*

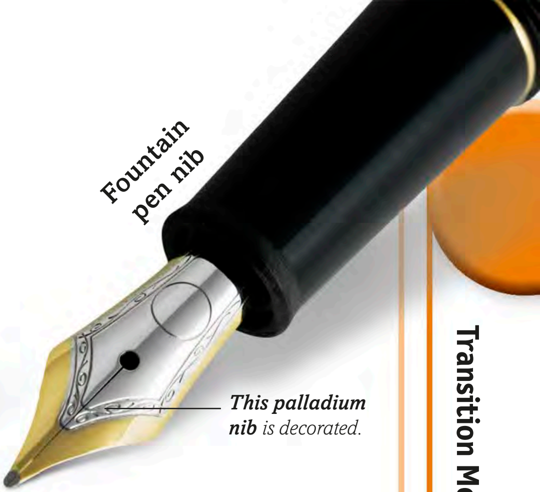

Fountain pen nib

***This palladium nib** is decorated.*

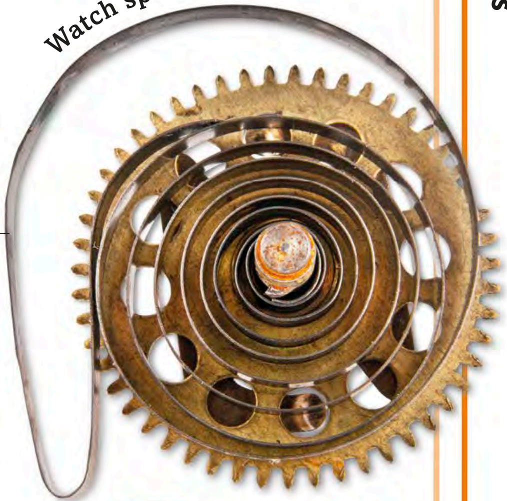

Watch spring

***This tiny spring,** which keeps time in a watch, is made of palladium alloys.*

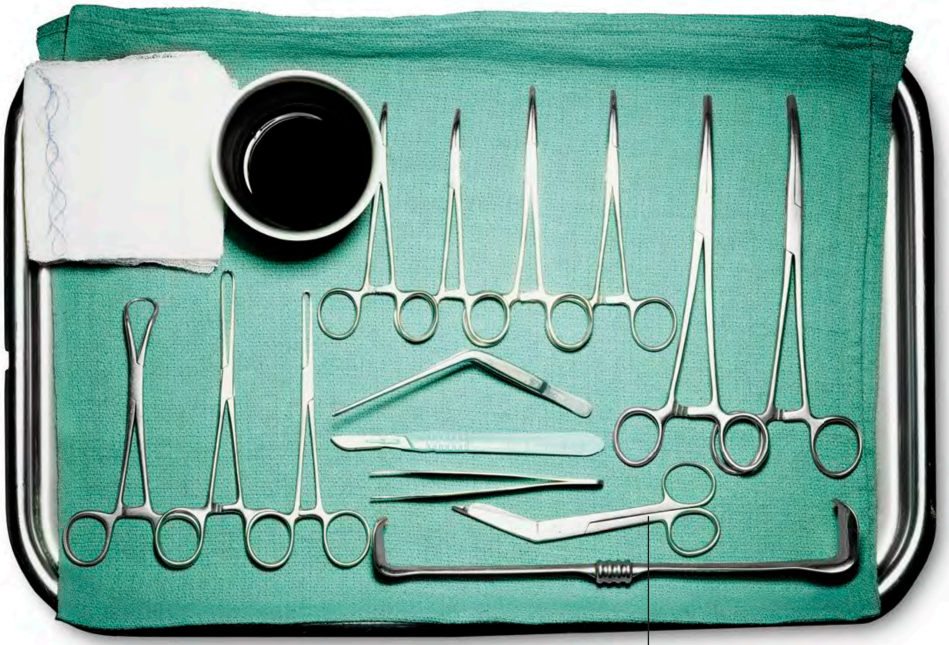

Surgical tools

***These tools** are made of palladium alloys, which keep them sharp for a long time.*

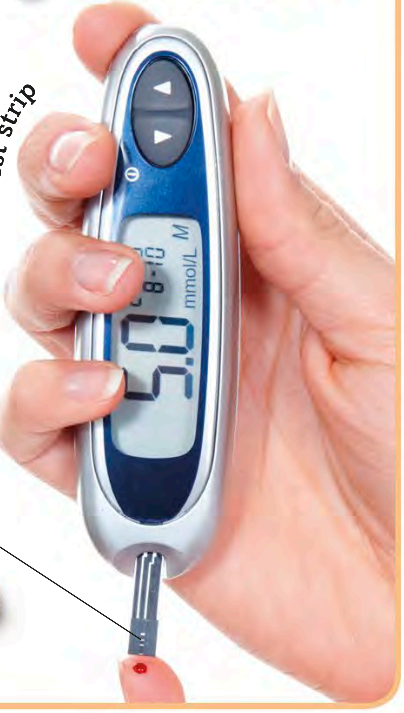

Glucometer test strip

***This test strip** contains palladium, which contributes to a reaction that measures glucose in a drop of blood.*

***This flute** contains palladium that enables it to resist corrosion.*

Orchestral flute

precious, it is used to make **commemorative coins** in some countries. Palladium is alloyed with steel to make it more resistant to corrosion. These alloys are used to make **surgical tools** and expensive musical instruments, such as some **flutes**. Palladium is often mixed with gold to form an alloy called white gold, which is used in jewellery. Some fountain pens have **nibs** decorated with palladium. The element is also used in **glucometer test strips** so that patients can check the level of glucose in their blood.

47 Ag Silver

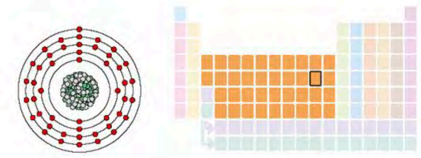

⊖ 47 ⊕ 47 ○ 61

State: Solid
Discovery: c. 3000 BCE

Forms

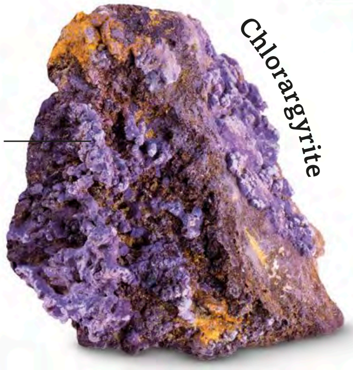

Chlorargyrite

This mineral's colour *changes to purple when it is exposed to bright light.*

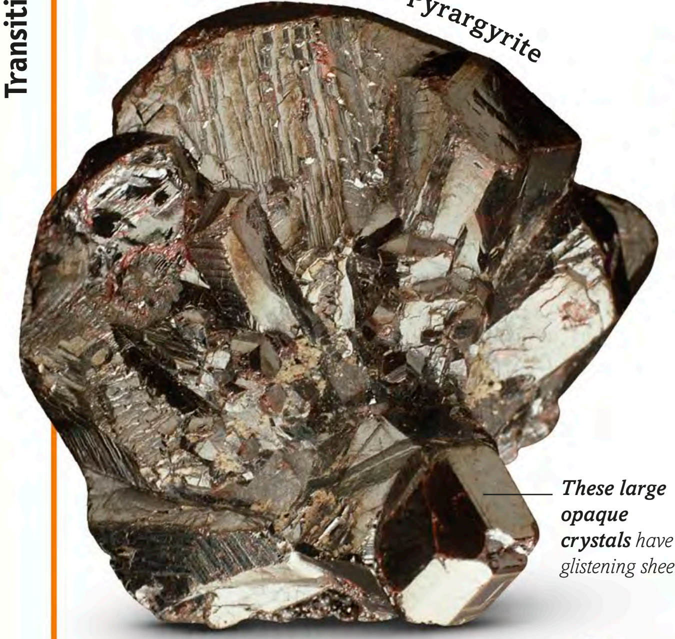

Pyrargyrite

These large opaque crystals *have a glistening sheen.*

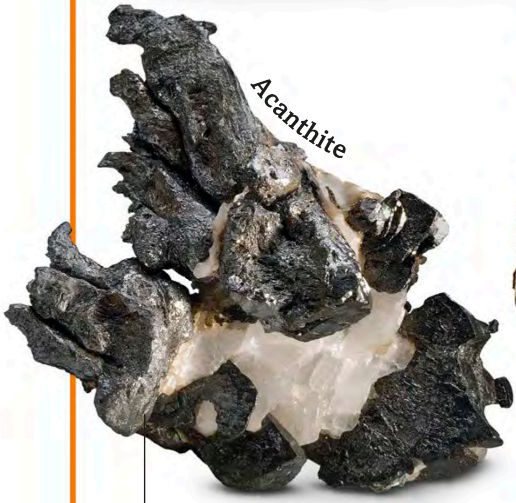

Acanthite

Black silver sulfide *forms twisted crystals.*

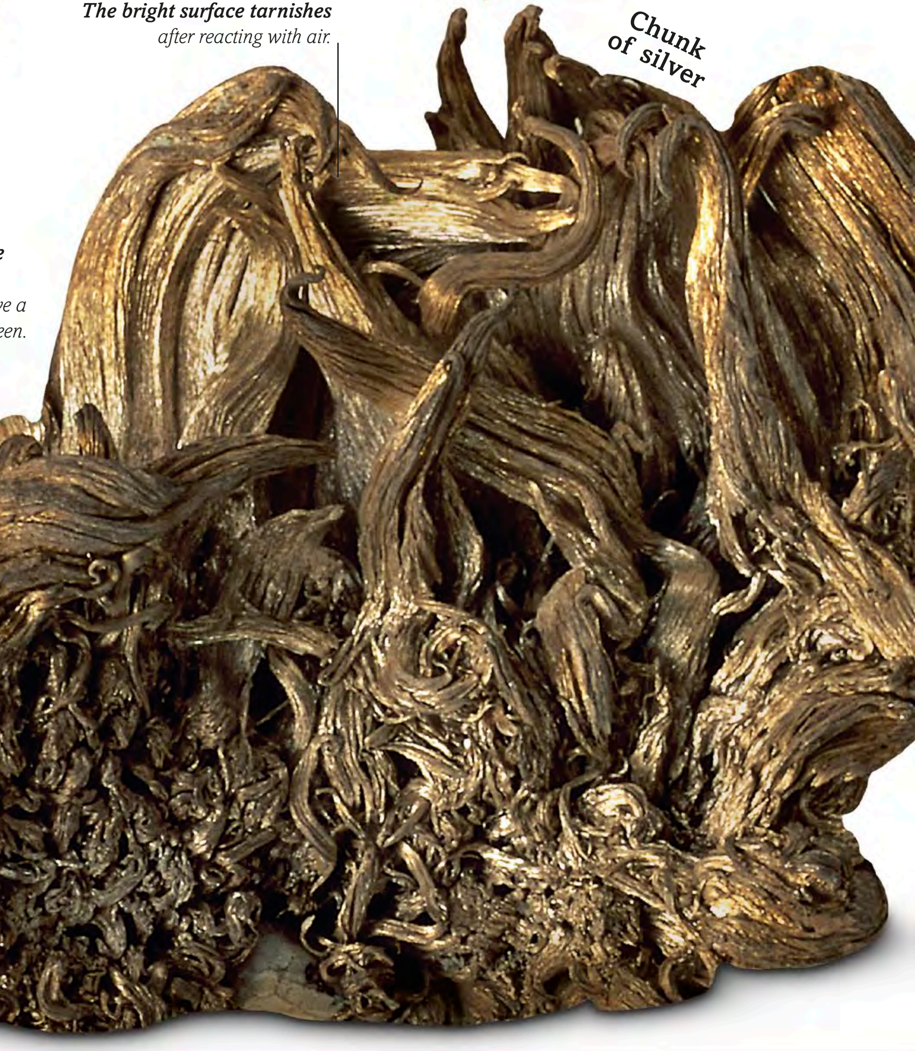

Chunk of silver

The bright surface tarnishes *after reacting with air.*

1 g (0.03 oz) of silver can be drawn into a **2-km- (1.2-mile-) long** wire.

Silver gets its symbol "Ag" from its Latin name *argentum*, which means "shiny white". It is considered a precious metal because its pure form has a grey shine that does not corrode quickly, and it stays untarnished if cleaned regularly. Silver can be found pure in nature, but mostly it is mined from ores, such as **pyrargyrite** and **acanthite**. Because this element is valuable and can be moulded easily, pure silver was used historically to make **coins**. This metal is also ideal for making **bracelets** and settings for gems. Some people even use

Uses

MAKING CLOUDS

Rain is crucial to our Earth, especially for growing healthy crops. Where there are no clouds, scientists can form tiny water droplets that cling to silver iodide powder, forming artificial rain clouds.

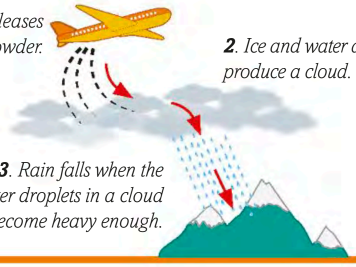

1. Aircraft releases silver iodide powder.

2. Ice and water droplets produce a cloud.

3. Rain falls when the water droplets in a cloud become heavy enough.

__Silver coating__ is used on some circuit board parts.

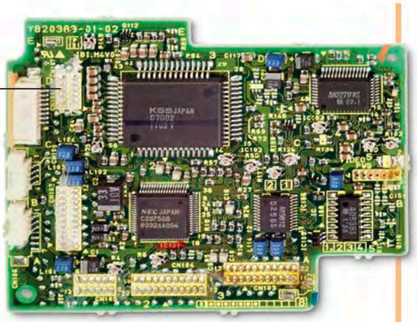

Circuit board

Silver coin

__Soft silver__ is easily pressed into coins.

__The polished surface__ has a pale, metallic shine.

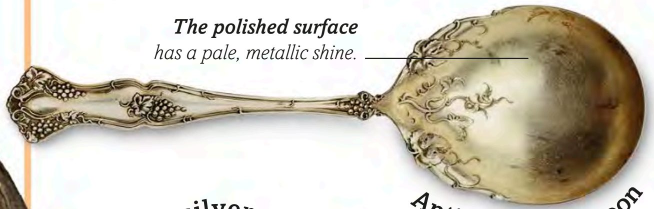

Antique silver spoon

Silver bracelet

__Pure silver__ moulded and cut into varying shapes

Edible silver foil

__These thin sheets of silver__ called "vark" are edible.

SILVER
NITRATE (V)
AgNO₃

__Silver nitrate__ is mixed with water to clean cuts and scrapes.

Silver nitrate

__Glass infused with silver chloride__ turns brownish when exposed to sunlight.

Photochromatic glasses

Photography plate

__An image forms__ when silver bromide darkens quickly on exposure to light.

flattened **silver foil** to decorate food. **Silver spoons** and forks were the only pieces of cutlery that did not create a nasty metallic flavour in the mouth in the days before the invention of stainless steel. Silver conducts electricity better than copper, and is used in some **circuit boards**. **Silver nitrate** (a compound of silver, nitrogen, and oxygen) is a mild disinfectant used in some anti-bacterial soaps. Silver forms light-sensitive compounds with chlorine (used in **sunglasses**) and bromine (used in old **photography plates**).

48
Cd

Cadmium

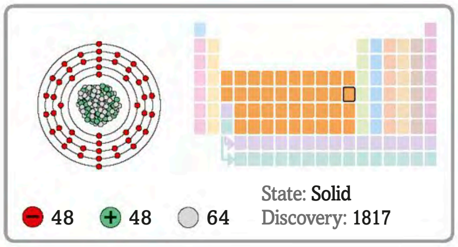

Forms

This mineral *contains a rare form of cadmium sulfide, a compound of cadmium and sulfur.*

Greenockite

This soft metal *has a bluish tinge.*

Pellet of pure cadmium refined in a laboratory

The yellow colour *of this zinc mineral is due to cadmium impurities.*

Smithsonite

Uses

Nickel-cadmium battery

Cadmium and nickel layers *create electricity in this rechargeable battery.*

This deep red pigment *contains powdered cadmium oxide.*

Red paint containing cadmium

This electronic component *used in circuits contains a compound of cadmium and sulfur.*

Light-sensitive resistor

Cadmium-covered screws *do not rust.*

Cadmium-plated screws

This research sample *is being viewed under ultraviolet (UV) light produced by a cadmium laser.*

Fluorescent microscope

Cadmium is a highly toxic metal, and is known to cause cancer. This rare element is found in the ore **greenockite**, but it is mostly obtained as a by-product of zinc extraction. **Cadmium** was discovered in 1817 from a mineral called calamine. Today, this metal is mainly used in conjunction with nickel in **rechargeable batteries**. The compound cadmium oxide was once used in preparing **red paints**, but not anymore because of its toxicity. Cadmium is also used to create lasers for use in powerful **microscopes**.

72

Hf

Hafnium

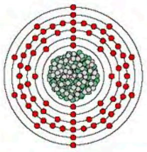

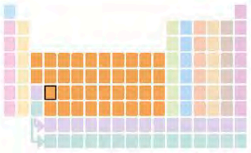

⊖ 72 ⊕ 72 ○ 106

State: Solid
Discovery: 1923

Forms

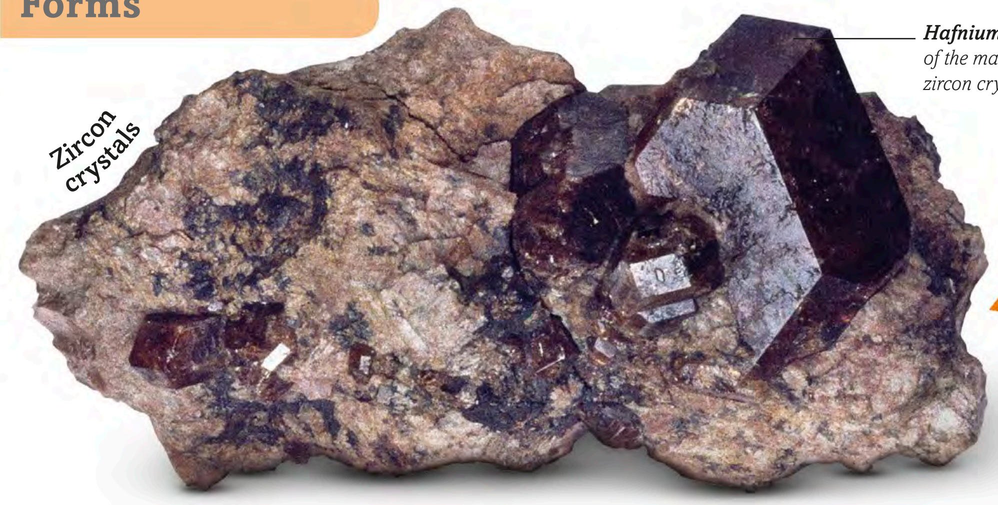

Zircon crystals

***Hafnium** uses up 4% of the mass of this zircon crystal.*

A zircon crystal can be up to **4 billion years** old.

Laboratory sample of pure hafnium

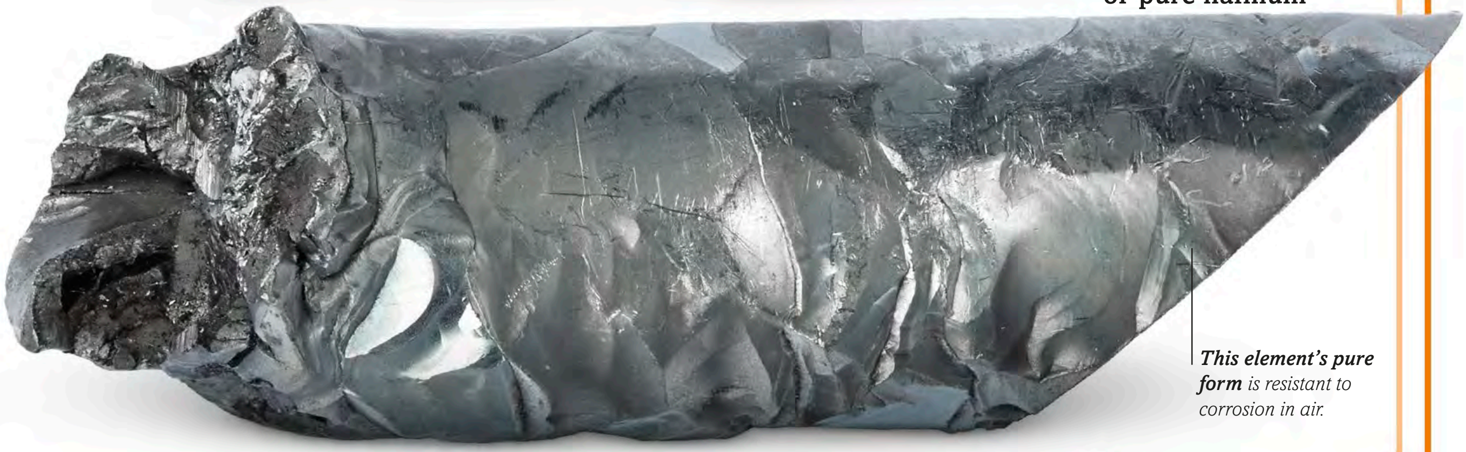

***This element's pure form** is resistant to corrosion in air.*

Uses

***Sections of this cutter** are made of hafnium.*

Metal cutter

***Small electronic components** in this microchip contain hafnium.*

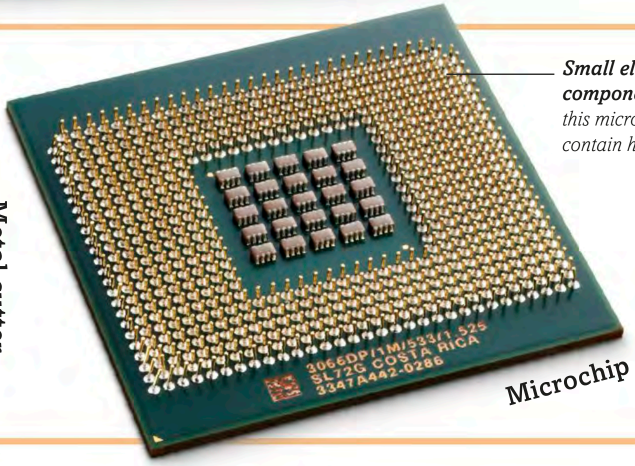

Microchip

Hafnium is named after *Hafnia*, the Latin word for the city of Copenhagen in Denmark. It took a long time to distinguish hafnium from zirconium because the two elements are present together in crystals of the mineral **zircon** and their atoms are similar sizes. Hafnium is used in powerful **cutters** that pierce metallic objects with a hot stream of sparks. It is also used to make ultra-small electronics – only a few millionths of a millimetre wide – in **microchips**.

73
Ta

Tantalum

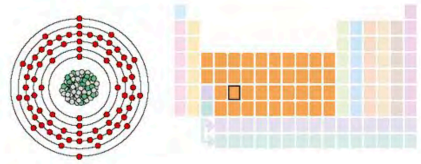

73 73 108

State: Solid
Discovery: 1802

Forms

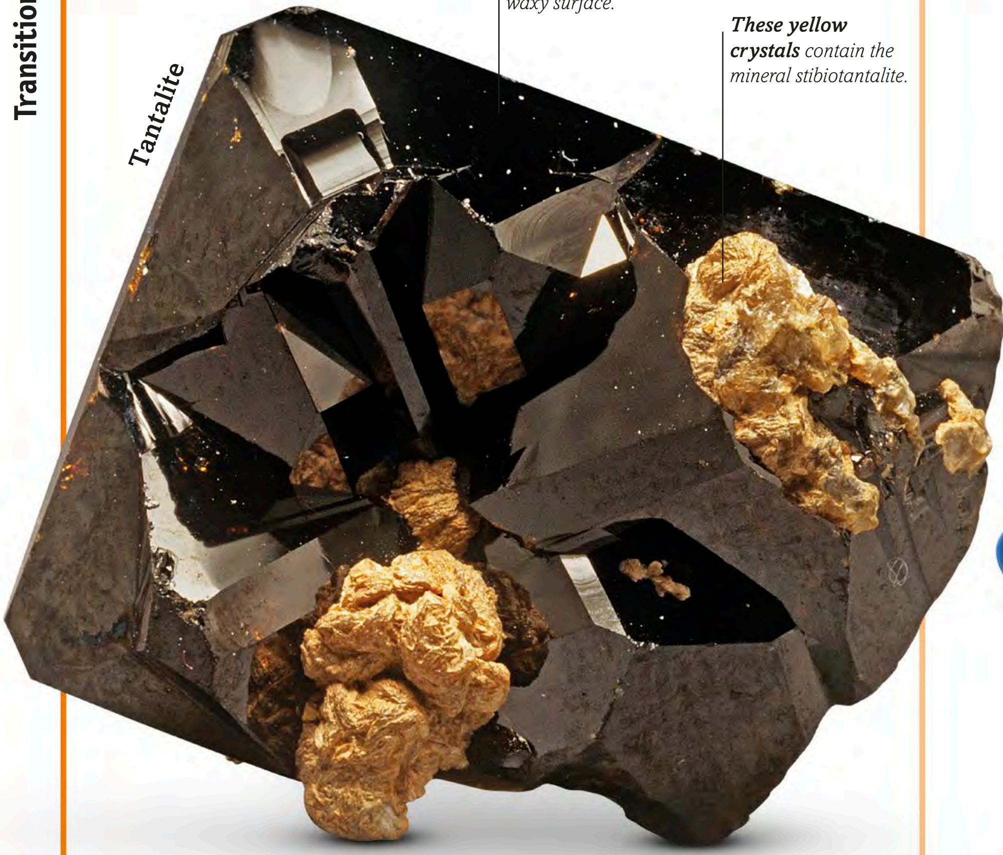

Tantalite

The mineral *has a dark, waxy surface.*

These yellow crystals *contain the mineral stibiotantalite.*

Rods of pure tantalum refined in a laboratory

The pure metal *barely reacts with air, so stays shiny.*

Uses

Artificial joint

The tantalum shell *of this hip implant is lightweight and flexible.*

Tantalum capacitors *are used to store a lot of electric charge in small circuits, such as the ones inside mobile phones.*

Electronic capacitor

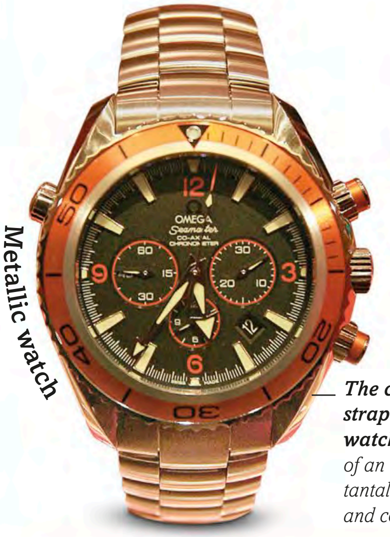

Metallic watch

The case and strap of this watch *are made of an alloy of tantalum, gold, and copper.*

Tantalum is a hard metal named after Tantalus, a man from Greek mythology who was punished by the gods. It is extracted from a rare mineral called **tantalite**. This tough metal is not harmful to the human body, so it is used to make **artificial joints** and other body implants. Tantalum powder is used in **capacitors** – devices used in electronic circuits to store electricity. This strong metal toughens **watches** made of softer, precious metals. Tantalum is also used to create strong turbine blades that do not corrode.

74

W

Tungsten

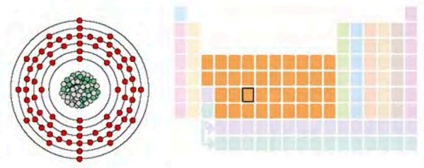

● 74 ⊕ 74 ○ 110

State: Solid
Discovery: 1783

Forms

*This dark, metallic **mineral** contains tungsten and iron.*

Ferberite

*This **mineral** is the main source of pure tungsten.*

Wolframite

*Pure **tungsten** is a hard, grey metal.*

Cylinder of pure tungsten refined in a laboratory

Uses

Drill bit

*This **drill bit** has a coating of tungsten carbide, which protects it from damage.*

Tungsten pigments were in use in Chinese porcelain 350 years ago.

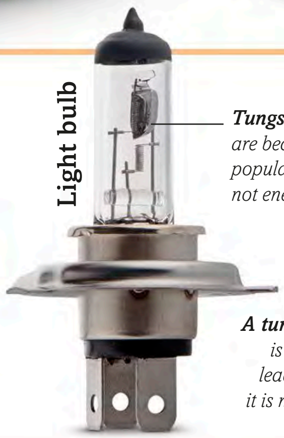

Light bulb

*Tungsten **filaments** are becoming less popular as they are not energy-efficient.*

Fishing sinker

*A **tungsten sinker** is preferred to a lead one because it is not poisonous.*

Tungsten has the highest melting point of any metal: it turns to liquid at a searing 3,414°C (6,177.2°F). It is a very dense metal, and its name comes from the Swedish phrase for "heavy stone". This metal is usually obtained from the mineral **wolframite**. A compound called tungsten carbide is used to harden objects such as **drill bits**. Tungsten's high melting point allows it to be used in the filaments of **light bulbs**. This element is also useful in producing weights, such as **sinkers** used with fishing lures.

75
Re
Rhenium

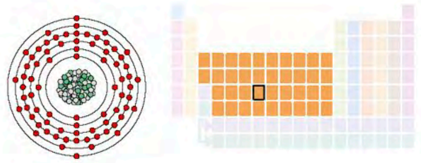

⊖ 75 ⊕ 75 ○ 111

State: **Solid**
Discovery: **1925**

Forms

Molybdenite

This ore *contains molybdenum and small amounts of rhenium.*

Pure rhenium pellet

Pure rhenium *is denser than gold.*

Uses

X-ray tube

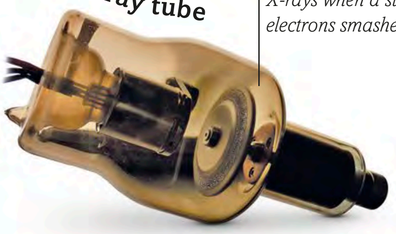

These tubes *contain a rhenium alloy that produces X-rays when a stream of electrons smashes into it.*

Rhenium has the **highest boiling point** of any element.

F-22 Raptor fighter plane

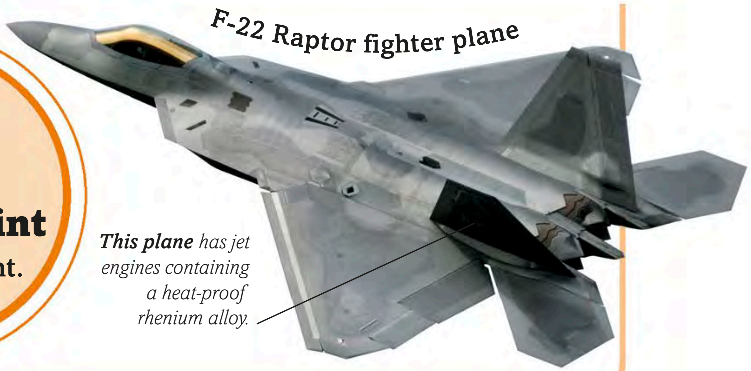

This plane *has jet engines containing a heat-proof rhenium alloy.*

Rhenium is very rare in nature: only one atom out of every billion in Earth's crust is a rhenium atom. Discovered in Germany in 1925, and named after the Rhine river – it was the last stable, non-radioactive element to be found. Rhenium has a very high melting point, and can stay solid at extreme temperatures. This allows alloys made of this element to be used in very hot conditions, such as those inside the tubes of **X-ray machines**, as well those in the exhaust nozzles of rockets and the jet engines of **fighter planes**.

76 Os Osmium

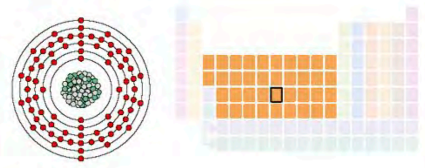

76 76 114

State: Solid
Discovery: 1803

Forms

Osmiridium sand

***This is a natural alloy** of osmium and iridium.*

Pellet of pure osmium refined in a laboratory

***The pure form** is hard but shatters easily.*

Uses

Transmission electron microscope (TEM) image

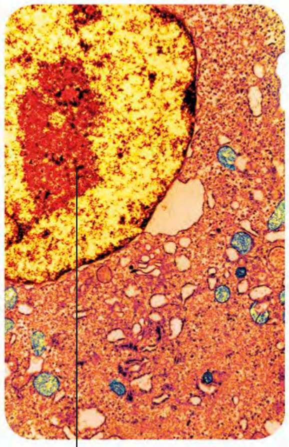

***Osmium oxide** is used to highlight objects inside a cell.*

Fingerprint powder in use

***Black osmium oxide powder** clings to oily fingerprints.*

Record player

***The needle** of this old record player is made of osmium.*

Fountain pen

***The nib of this pen** moves smoothly because of its hardy osmium alloy.*

Osmium is the densest of all naturally occurring elements: 250 ml (8.5 fl oz) of this metal (in its liquid form) weighs 5.5 kg (12 lb). This rare element is found in the ore **osmiridium**. **Pure osmium** reacts with oxygen in the air to form a poisonous oxide, so the metal is used safely by combining it with other elements or alloys. A red osmium oxide stains cells so they can be seen clearly under a powerful **microscope**, while a black oxide powder allows **fingerprints** to be revealed in crime investigations. A hard osmium alloy is used in **fountain pen** nibs.

77 **Ir** Iridium

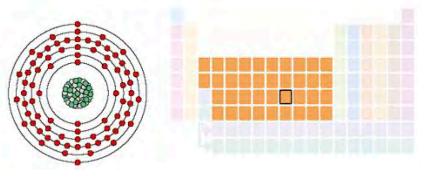

− 77 **+** 77 ○ 115

State: **Solid**

Discovery: **1803**

Forms

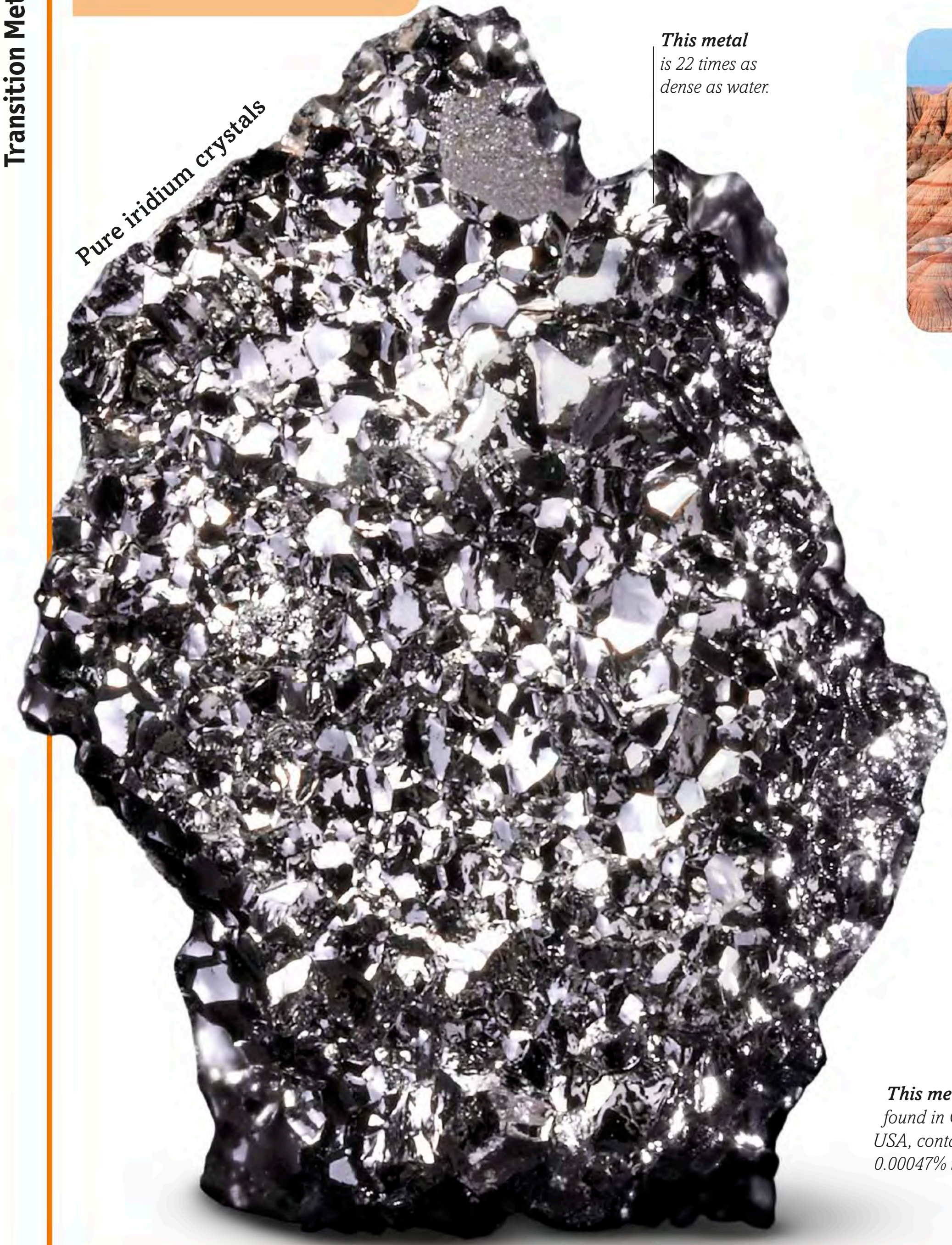

This metal *is 22 times as dense as water.*

These rocks *contain a layer of iridium-rich clay.*

Badlands National Parks, South Dakota, USA

Willamette meteorite

This meteorite*, found in Oregon, USA, contains just 0.00047% iridium.*

Iridium is the rarest natural element on Earth: there is one iridium atom out of every billion atoms in Earth's rocks. This dense metal can be found in its **pure** form in nature as well as in other common ores that contain nickel and copper. Iridium is present in **meteorites** and other space rocks. A layer of iridium-rich clay is found in Earth's crust all over the world, especially in the **Badlands of South Dakota, USA**. Scientists believe this small quantity of iridium in our planet's crust was deposited by the dust from an

Uses

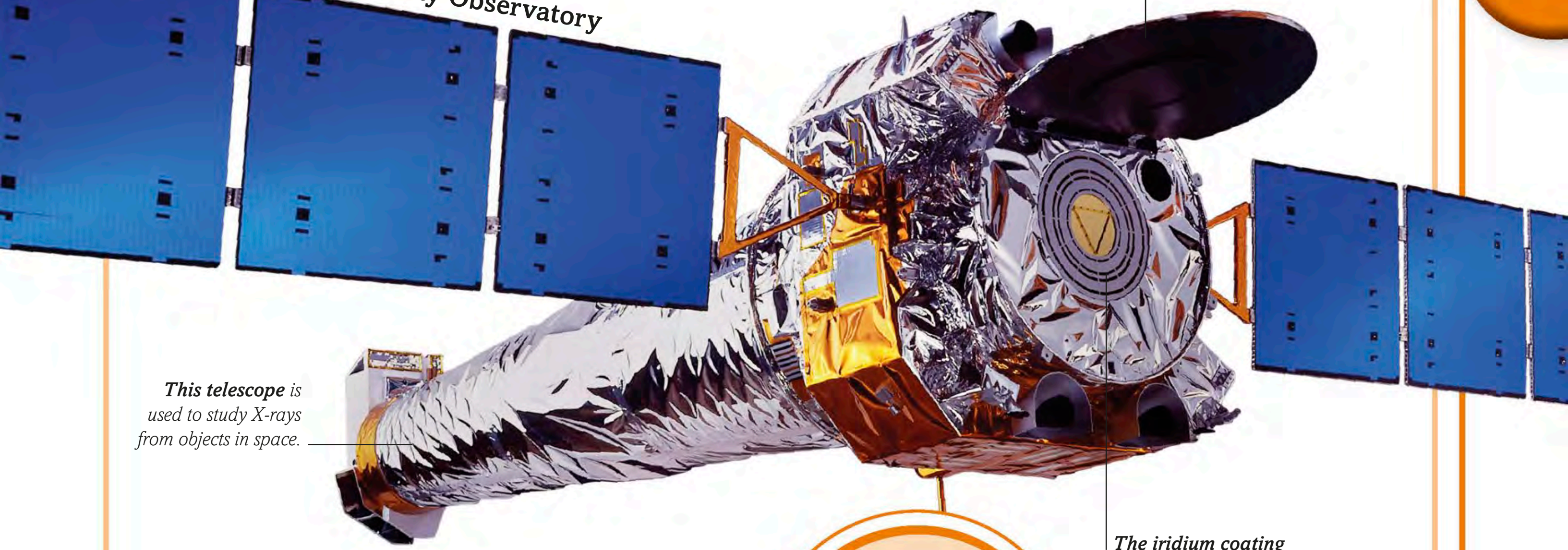

Chandra X-ray Observatory

This door *protects the equipment from the Sun's glare.*

This telescope *is used to study X-rays from objects in space.*

The iridium coating *on the mirror is only a few atoms thick.*

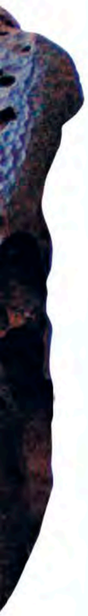

Spark plug

This spark plug *contains a tiny amount of iridium, and can resist the high temperatures created during sparking in vehicles.*

Iridium is the **second densest metal**, after osmium.

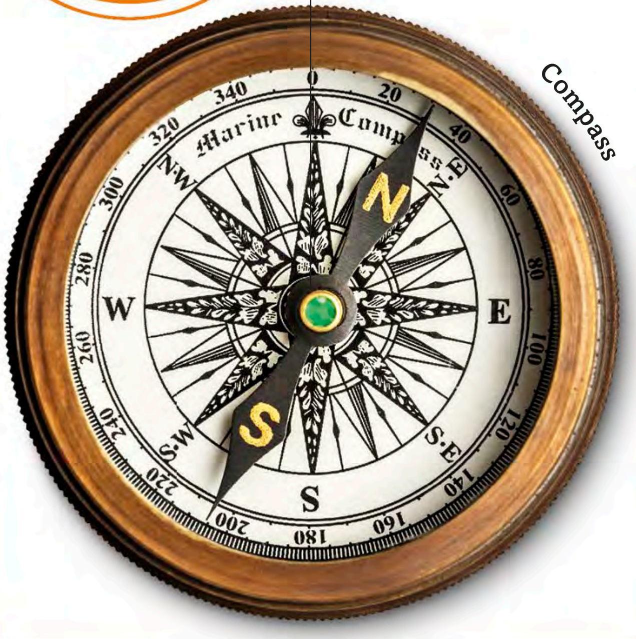

Compass

The moving parts of *some compasses are made of the alloy osmiridium.*

LUIS WALTER ALVAREZ

In 1980, the American physicist Luis Walter Alvarez and his son Walter discovered a layer of iridium-rich clay in rocks all over the world. They suggested that this was the result of a meteorite impact about 66 million years ago, which led to the extinction of the dinosaurs.

explosion 66 million years ago when a large meteorite hit our planet. The applications of this element include coating the mirror of NASA's **Chandra X-ray Observatory**, an Earth-orbiting telescope that studies X-rays from distant stars. Iridium is more durable than platinum and copper, and is therefore preferred over these metals for use in spark plugs. Iridium is also mixed with osmium to make an alloy called osmiridium, which is used in **compasses** and put in nibs for some fountain pens to make them hardy.

78 Pt Platinum

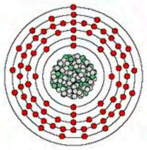

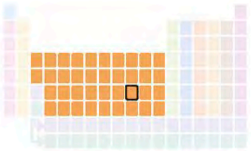

⊖ 78 ⊕ 78 ○ 117

State: Solid
Discovery: Unknown

Forms

Platinum melts at the high temperature of **1,768°C** (3,214°F).

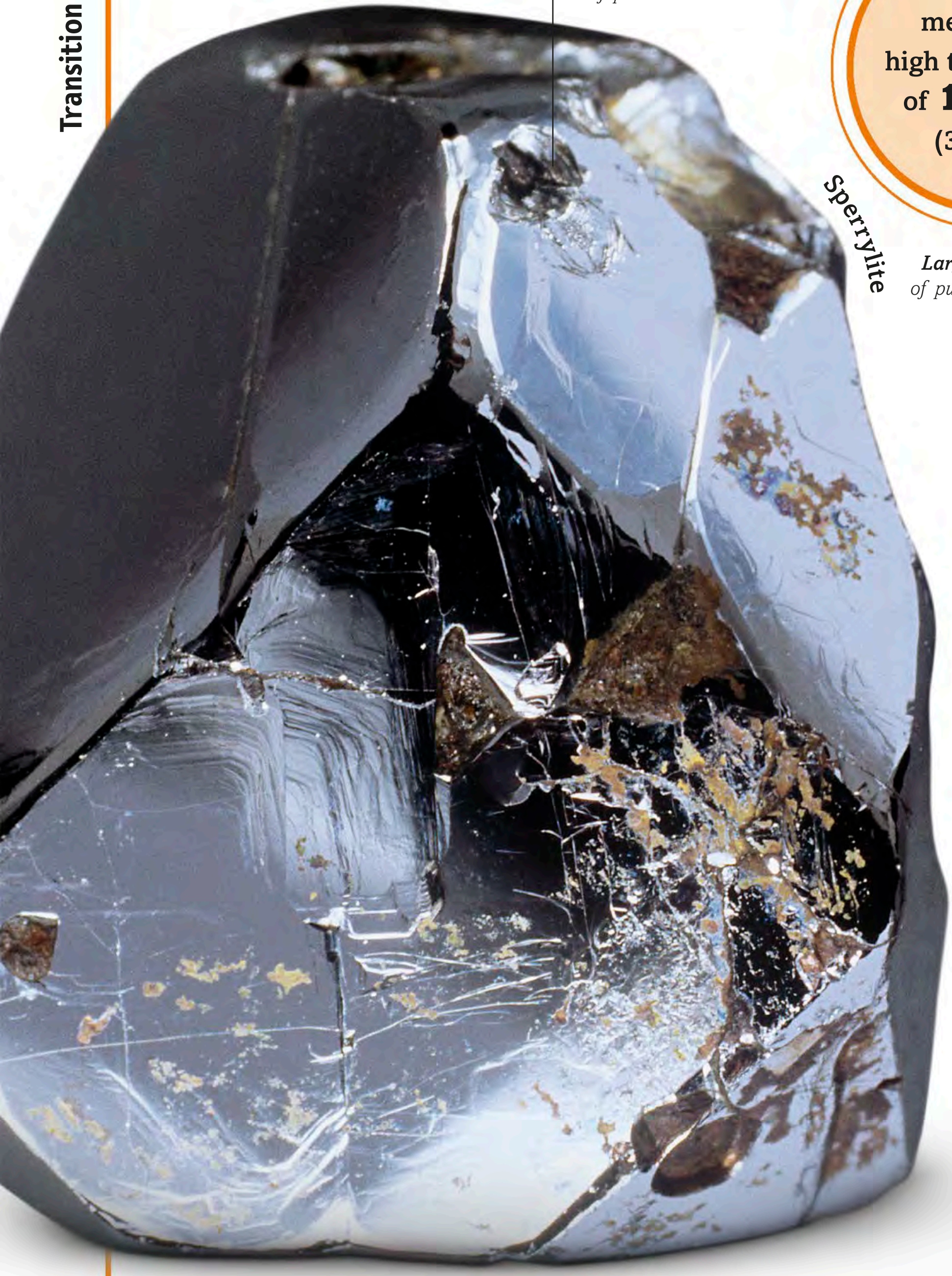

Sperrylite

This dense, shiny mineral *consisting of platinum and arsenic is the most common ore of platinum.*

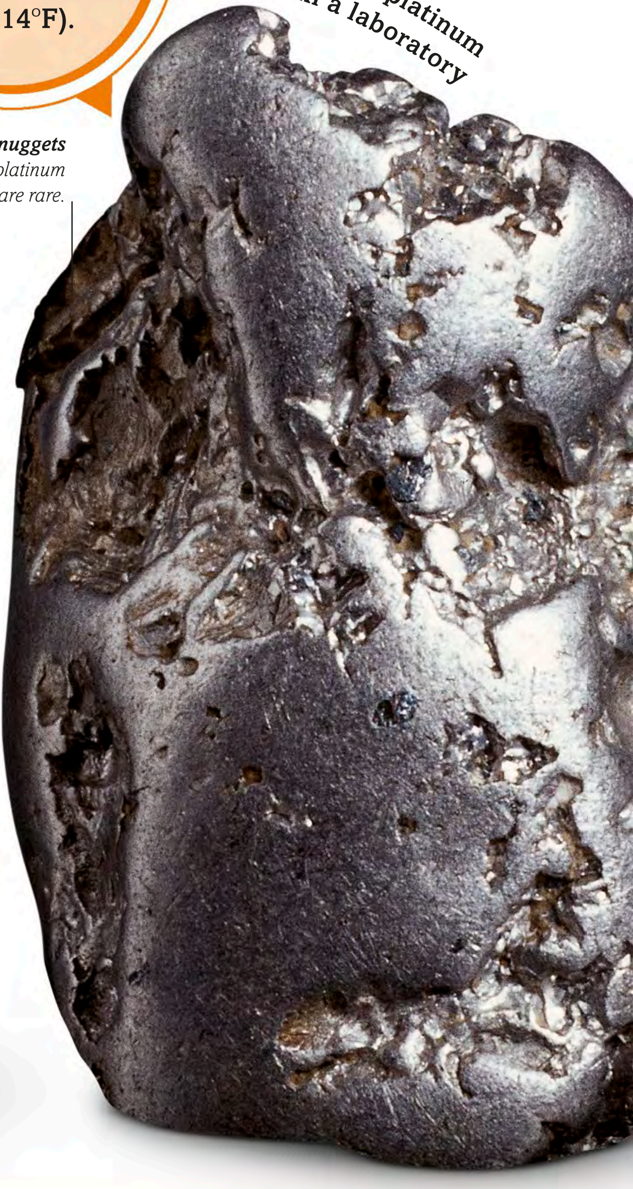

Nugget of pure platinum refined in a laboratory

Large nuggets *of pure platinum are rare.*

Spanish explorers first found platinum in the mines of South America in the 1700s. They obtained a whitish substance that the locals living near there called *platina*, meaning "little silver". This precious metal has a silvery white shine. Platinum rarely reacts with other elements, even at high temperatures. This makes it difficult to extract from its ores, such as **sperrylite**. **Pure platinum** does not corrode or tarnish. It is, however, not easy to shape or mould, so use of platinum was limited to the making of simple jewellery and **watches**. By the 20th century,

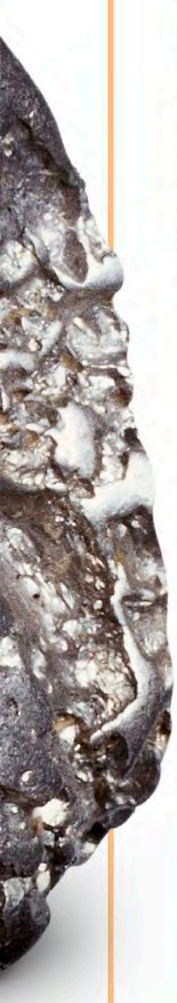

Uses

Platinum resistance thermometer

This thermometer *records temperature by measuring the electric current flowing through a fine platinum wire.*

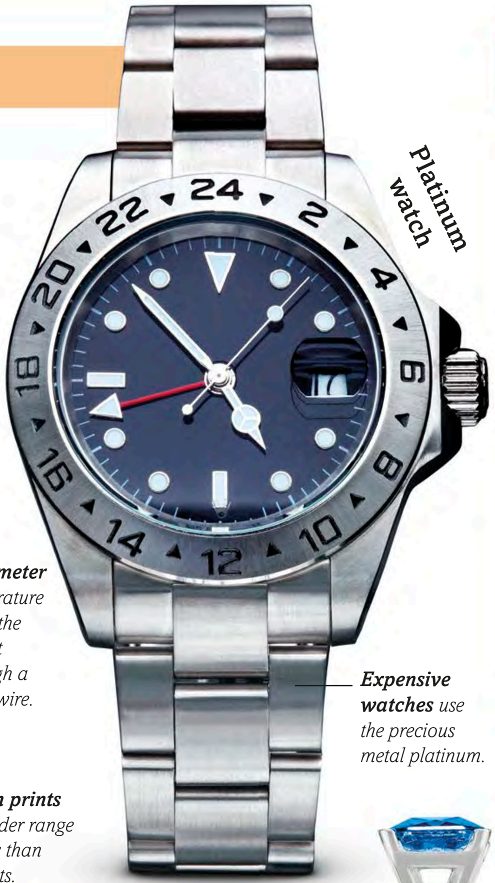

Platinum watch

Expensive watches *use the precious metal platinum.*

Platinum prints *have a wider range of shades than silver prints.*

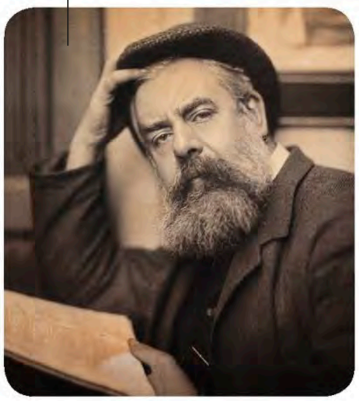

Black and white photographic print

Jewellery *made of platinum does not lose its shine.*

Platinum ring

Dental fillings *once contained platinum and mercury.*

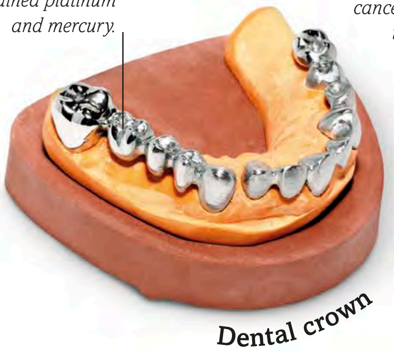

Dental crown

This drug *contains platinum and kills cancer cells in the body.*

Cisplatin
1 mg/ml concentrate for solution for infusion
50 mg
50 ml
Intravenous
For single use only

Cancer drug

This stent made of platinum *is not harmful to the body and anchors a damaged blood vessel as it heals.*

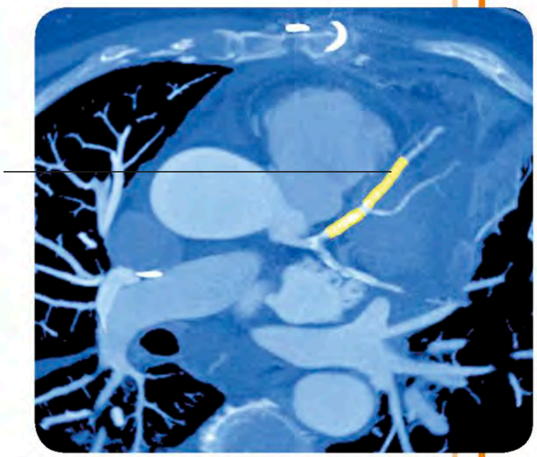

Medical stent

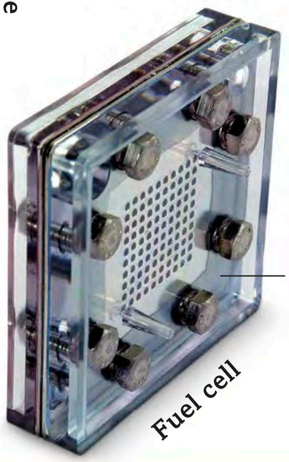

This fuel cell *contains platinum, which speeds up the reaction between hydrogen and oxygen.*

Fuel cell

Platinum was found in an Egyptian casket from the **7th century BCE**.

ANTONIO DE ULLOA

Although platinum had been in use in jewellery on the west coast of South America for more than 2,000 years, it was Spanish naval officer Antonio de Ulloa who made the first major study of it. In 1735, while on a South American expedition, he found grains of the metal in river sands. He brought them back to Spain to examine them.

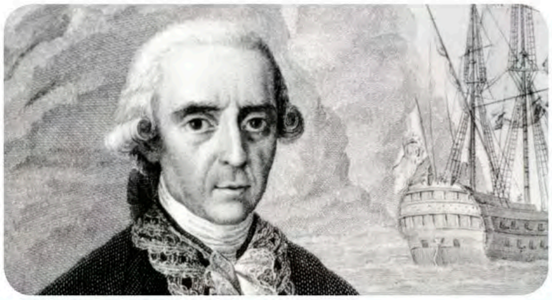

more applications were discovered. Platinum can be used in place of silver to generate **photographic prints**, and in place of gold for making **dental fillings**. Today, platinum plays an important role in various technologies. For example, it is used in **fuel cells** – devices that generate electricity by combining hydrogen and oxygen. These cells do not need to be recharged like other batteries. Powerful **drugs for treating cancer** contain this element, while **stents** made of pure platinum help heal damaged blood vessels.

79 **Au** Gold

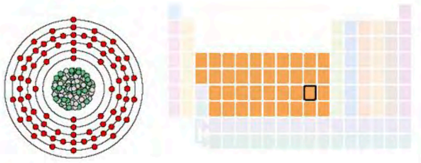

● 79 ⊕ 79 ○ 118

State: Solid
Discovery: c. 3000 BCE

Forms

Gold's chemical symbol **Au**, comes from its **latin name**, *aurum*.

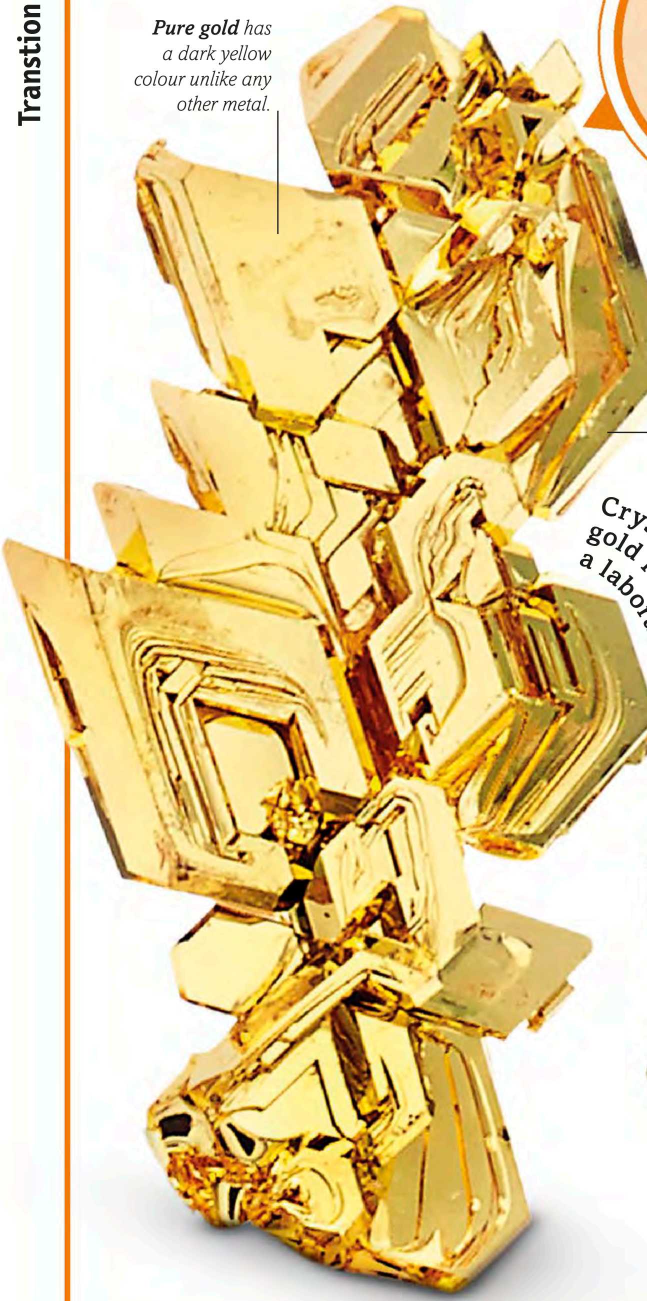

***Pure gold** has a dark yellow colour unlike any other metal.*

***These crystals** may be elongated, rectangular, or cubic.*

Crystals of pure gold refined in a laboratory

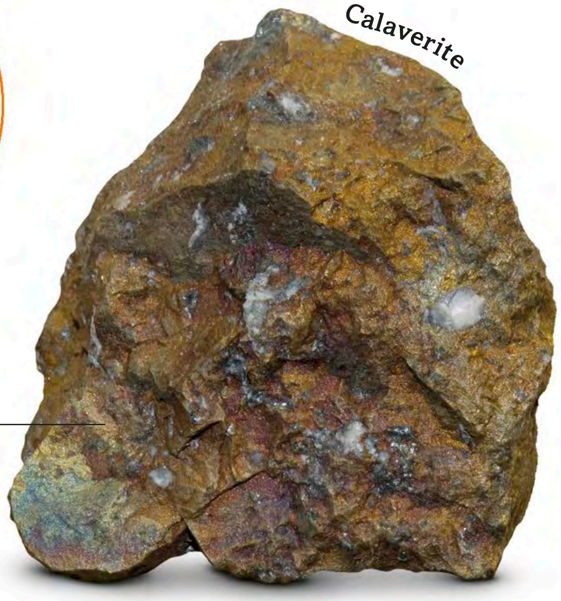

Calaverite

***This mineral ore** contains gold and tellurium.*

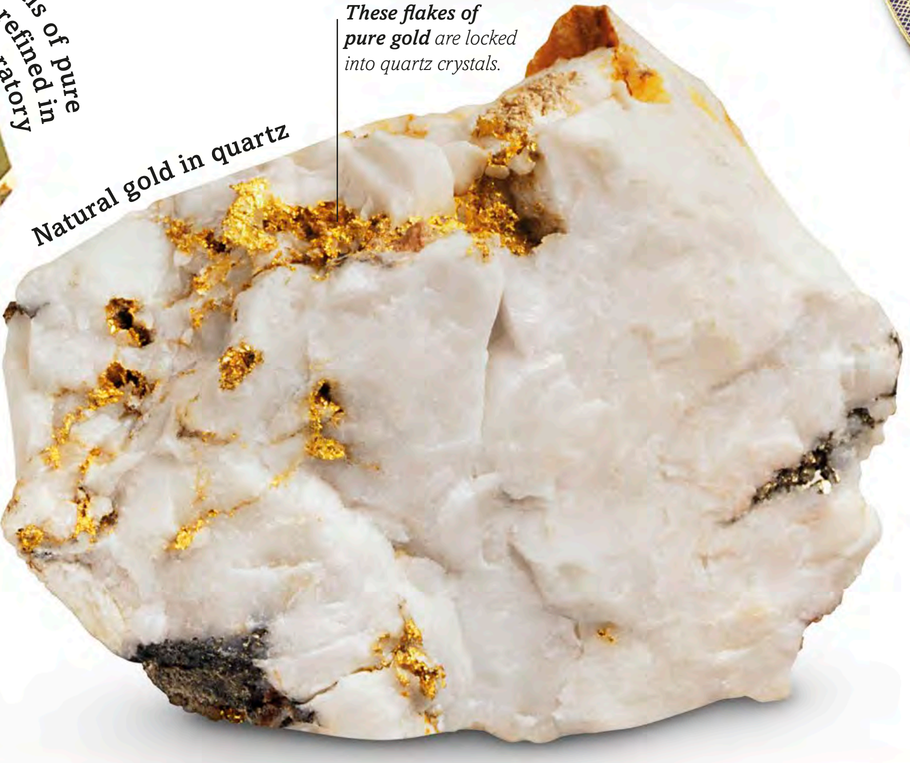

Natural gold in quartz

***These flakes of pure gold** are locked into quartz crystals.*

People were making gold ornaments more than 6,000 years ago. This was many centuries before they learned how to purify copper, iron, and other metals. Gold is believed by many to be the first metal element to be identified. It is a dense, unreactive metal with a distinctive deep yellow colour. Gold is naturally **pure** and seldom makes compounds in nature; the compound in the mineral ore **calaverite** is an exception. **Pure gold** found in nature may form nuggets but mostly is found as tiny specks embedded in rocks. Gold miners crush up these rocks and wash out the

Uses

This very thin layer of gold *protects the astronaut from the Sun's heat.*

Astronaut's visor

This mask *was placed over the pharaoh's mummified face.*

Tutankhamun's death mask

Royal Crown Derby plate

This glass plate *contains specks of gold.*

Gold bars

Gold bars *stored in banks are a sign of wealth.*

FINE GOLD
999.9
NET WT
1000 g

Edible gold flakes

The flakes *decorating this expensive chocolate are edible.*

Gold teeth

These replacement teeth *are made of gold and mercury.*

Wat Phrathat Doi Suthep temple, Thailand

A thin layer of gold *covers this entire temple.*

Gold foil *keeps this car engine at a stable temperature.*

McLaren F1 car engine

Ancient gold jewellery

This neck ornament *is made from cast gold.*

THE HOLTERMANN NUGGET

The largest piece of natural gold was found on 19 October, 1872, near the small town of Hill End in Australia. Named after its discoverer, Bernhardt Holtermann, the piece contained moe than 90 kg (198 lb) of pure gold.

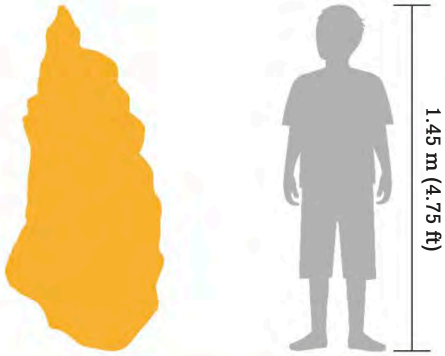

Holtermann Nugget **Child aged 10 years old**

gold dust with water or strong acids. The applications for gold include heat shields in **astronaut's visors**. This metal has always been seen as valuable and many ancient artefacts, such as the **3,300-year-old death mask** of Egyptian pharaoh Tutankhamun, were forged from it. Some of the earliest coins, found in Turkey, were made of it. Gold is used to cover important buildings, such as Thailand's **Wat Phrathat Doi Suthep** temple. This precious metal is most commonly used today in **jewellery** or decorations.

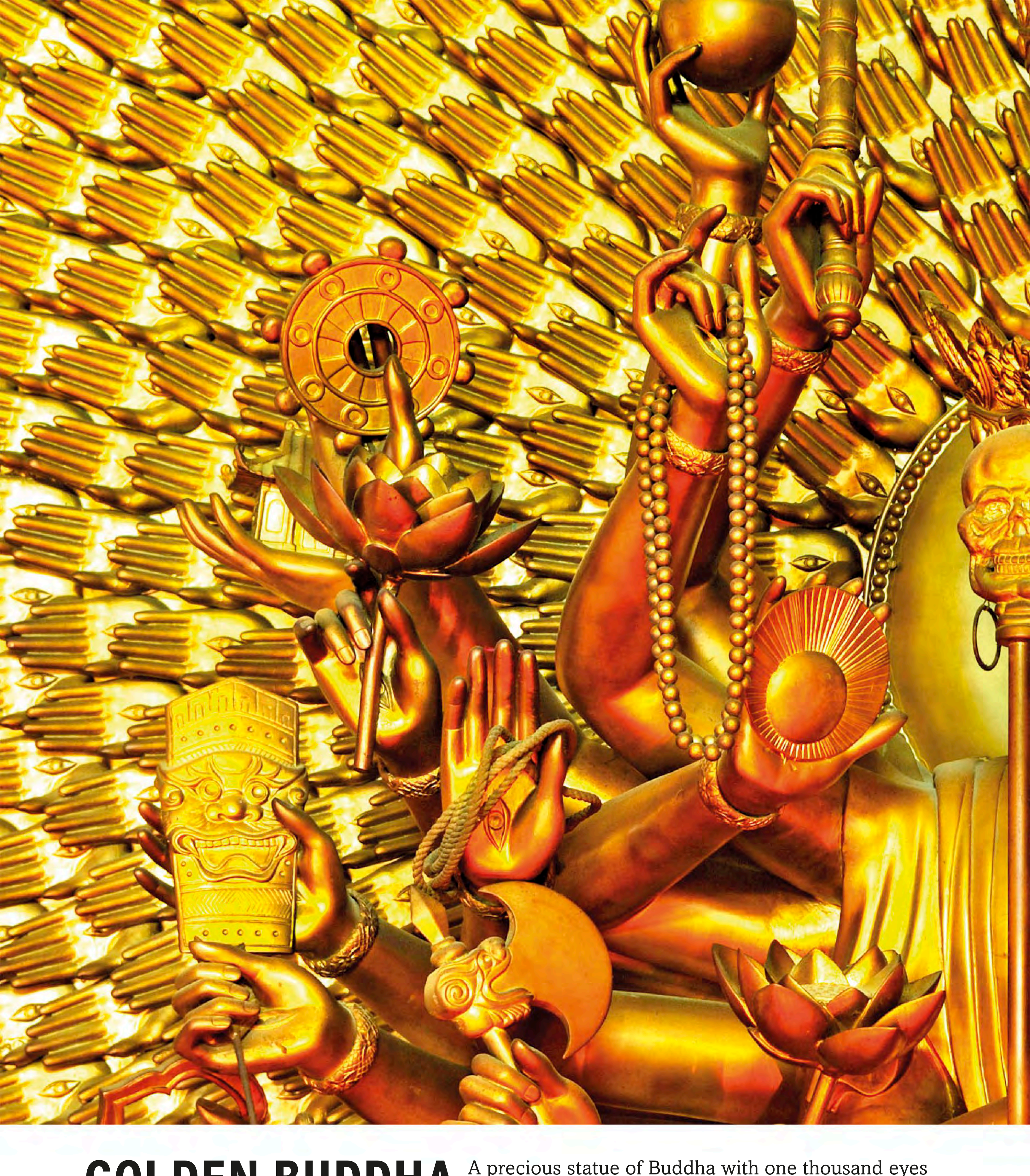

GOLDEN BUDDHA A precious statue of Buddha with one thousand eyes and one thousand hands stands in Long Son Pagoda, a temple in Nha Trang, Vietnam. The Buddha is depicted as holding a range of sacred objects, including scrolls and white lotus flowers. This statue is completely covered in a layer of pure gold, and it draws in hundreds of devotees from across the world.

Although humans have discovered many strong metals and useful elements, gold has remained one of the most valuable. Before people knew what it was, they saw glittering gold dust in river beds or dug large gold nuggets out from rocks. They found that gold has many valuable qualities: it is soft enough to hammer into any shape and can be melted down for moulding into ornaments. Best of all, its gleaming golden colour never fades away. Ancient cultures prized items made of gold: in ancient Egypt, gold was used to make coins as well as to cap the tops of pyramids. Gold is, however, so rare that if all the world's mined gold were forged into a cube, it would fit inside the penalty area of a soccer pitch.

80 Hg Mercury

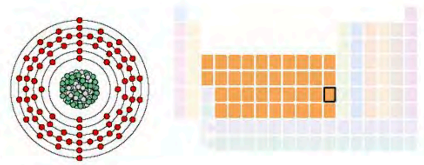

⊖ 80 ⊕ 80 ○ 121

State: Liquid
Discovery: c. 1500 BCE

Forms

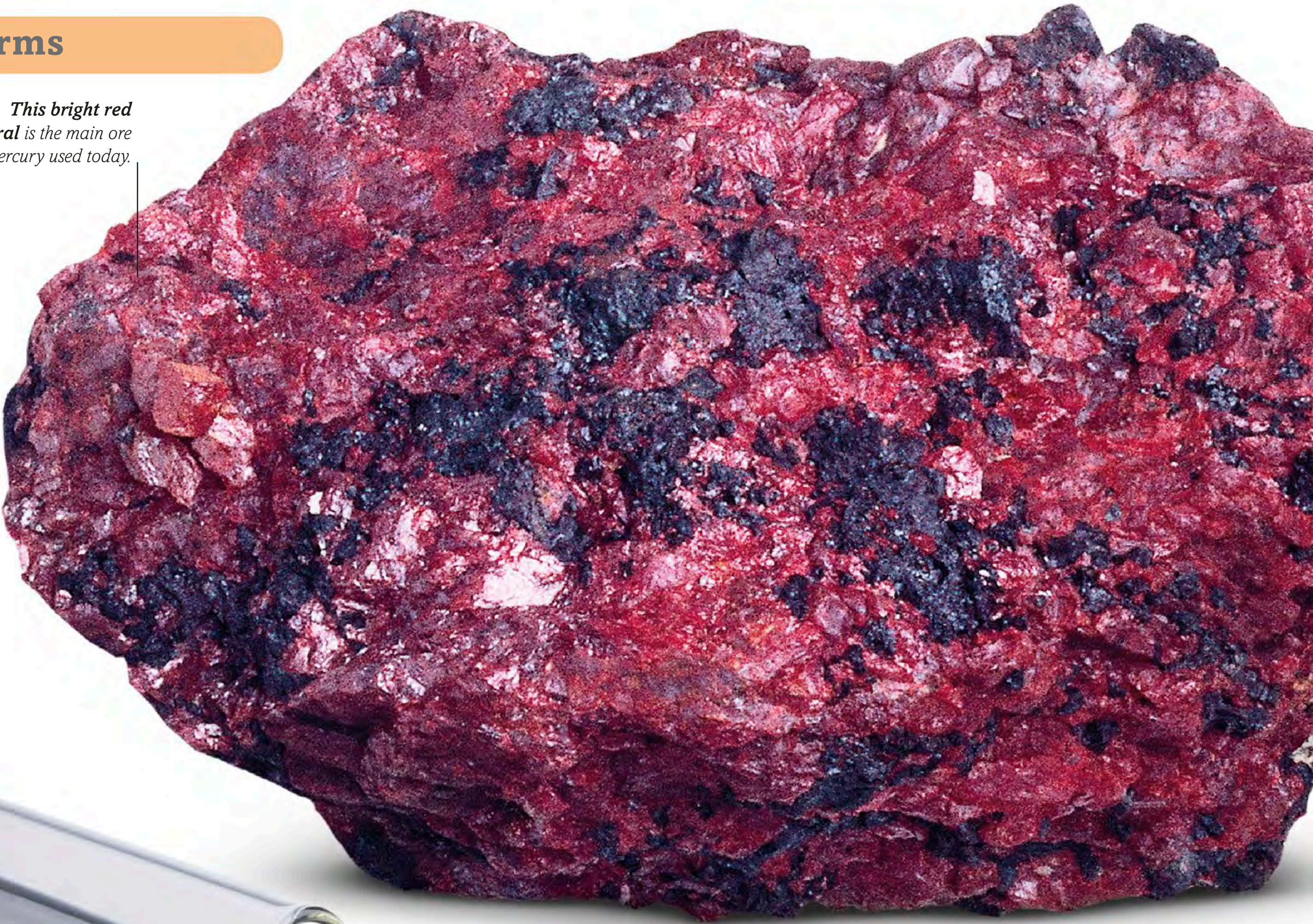

Cinnabar

This bright red mineral *is the main ore of mercury used today.*

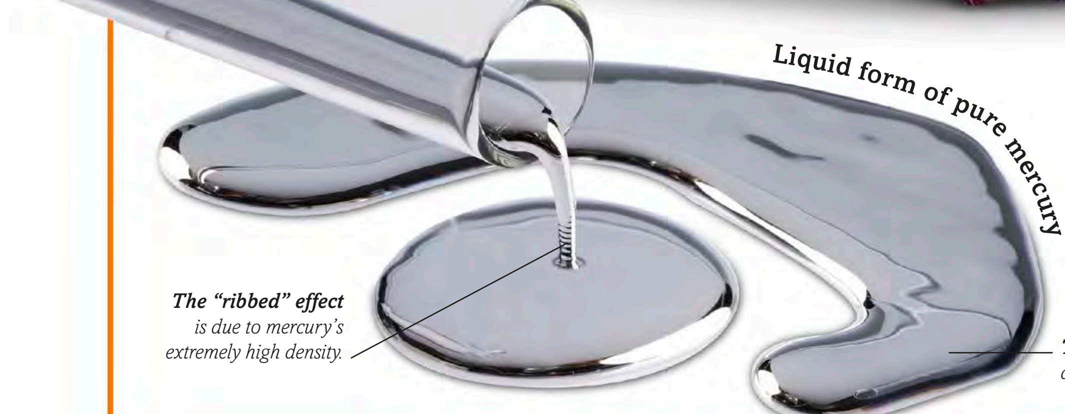

Liquid form of pure mercury

The "ribbed" effect *is due to mercury's extremely high density.*

This metal *melts at −39°C (−38°F).*

Solid mercury is soft enough to be **cut with a knife**.

Mercury is the only metal that is liquid at room temperature. Along with water, it is one of the few liquids found naturally on Earth's surface. **Pure mercury** forms around volcanoes where the heat separates it from its minerals, such as **cinnabar**. This red mineral has been used for many centuries: ancient Romans roasted cinnabar to release a liquid they called *hydrargyrum*, meaning "silver water". This was the element mercury. It was later known as quicksilver because of how fast it flowed as a stream of liquid. This metal is very poisonous:

Uses

94
6
8
100
2
4
6
108 F

This thermometer *contains mercury, which expands as it gets warmer and contracts as it cools.*

Mercury thermometer

Pills containing mercury *were commonly used to treat constipation and toothache.*

Mercury pills

This bulb *glows when the mercury vapour inside it is electrified.*

CFL bulb

Mercury has been in use for more than **4,000 years**.

HOW DOES A BAROMETER WORK?

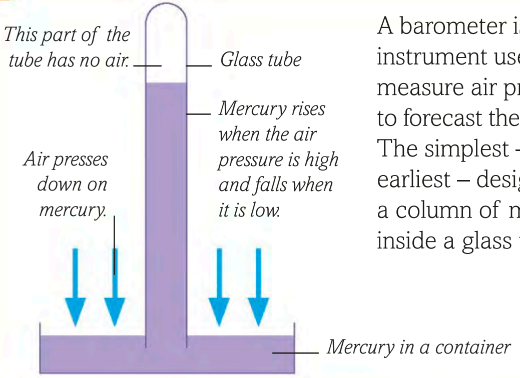
This part of the tube has no air.
Glass tube
Mercury rises when the air pressure is high and falls when it is low.
Air presses down on mercury.
Mercury in a container

A barometer is an instrument used to measure air pressure to forecast the weather. The simplest – and earliest – designs used a column of mercury inside a glass tube.

This bright red paint *is made with powdered cinnabar.*

Red paint

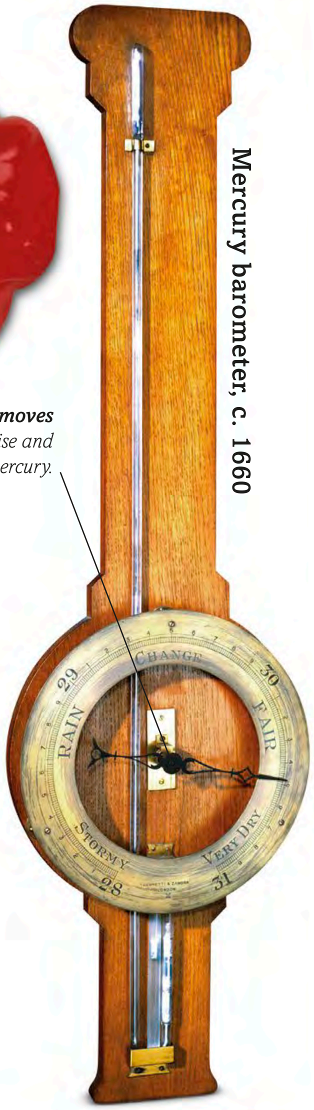
CHANGE
29
30
RAIN
FAIR
STORMY
VERY DRY
28
31

The needle moves *with the rise and fall of mercury.*

Mercury barometer, c. 1660

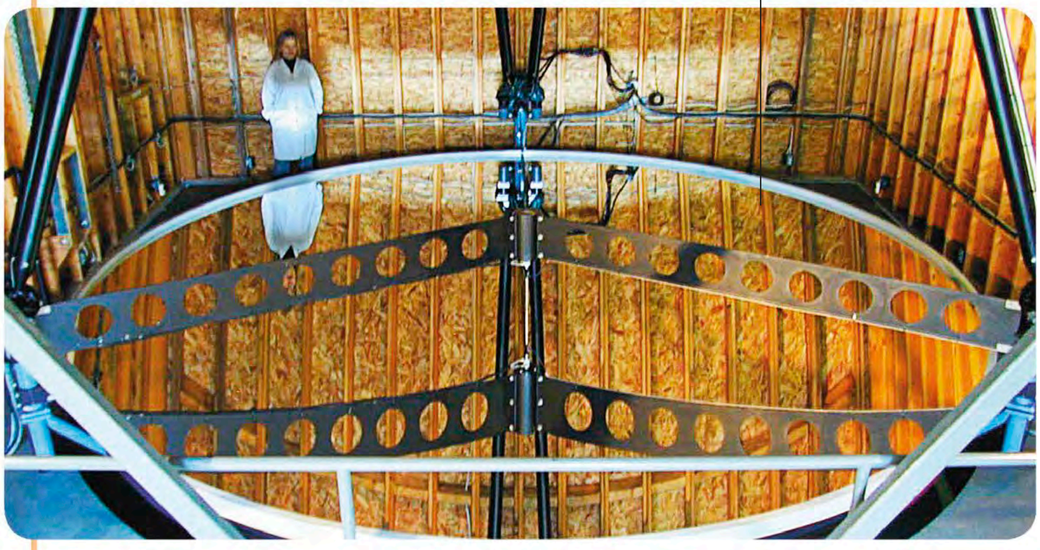

This large, low-cost mirror *made of a pool of pure mercury is used in an astronomical telescope.*

Liquid mirror inside telescope

it can damage organs and nerves if inhaled or swallowed. As a result, the use of this metal is carefully monitored today. Mercury is used in some batteries, some **thermometers**, and in low-energy, **compact fluorescent light (CFL) bulbs**. Its compounds are used to prepare strong, **red paints**. Until the early 18th century, mercury was used in **pills** for treating some common ailments. It gradually fell out of use when it was found to be toxic. The first accurate **barometers** also contained this liquid, but such devices are rarely seen today.

104 Rf Rutherfordium

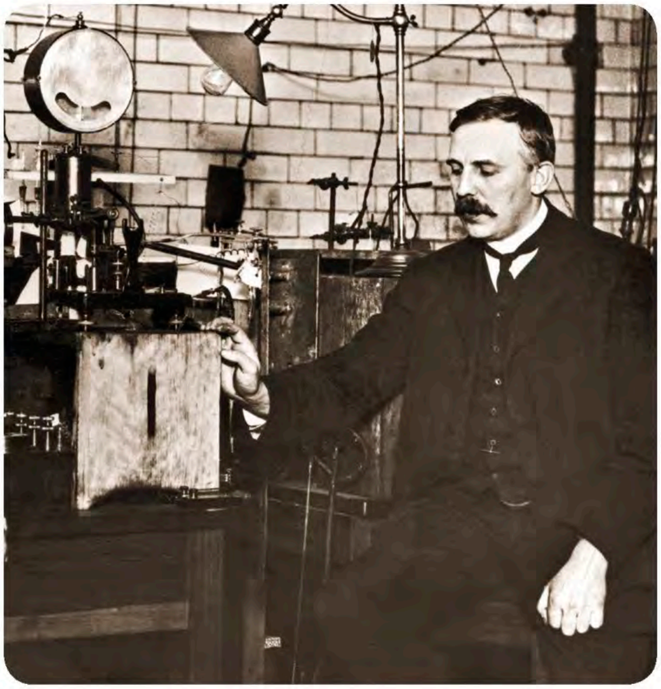
Ernest Rutherford

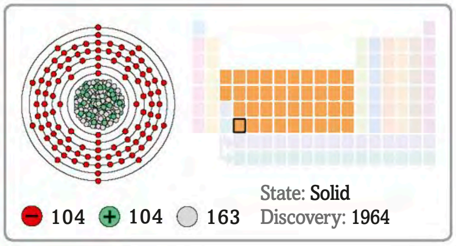

State: Solid
Discovery: 1964

Rutherfordium was the first superheavy element to be discovered. In this type of element, each atom has 104 or more protons in its nucleus. It is named after the New Zealand scientist **Ernest Rutherford**, who, in 1913, suggested that every atom has a nucleus, or core. Pure rutherfordium is synthesized by researchers in a laboratory.

---

105 Db Dubnium

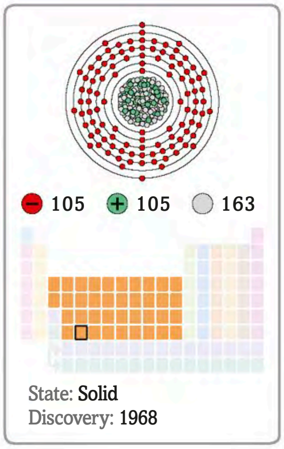

State: Solid
Discovery: 1968

The US scientist Albert Ghiorso discovered **12 elements** in the 20th century.

Albert Ghiorso

It took scientists nearly 30 years to agree on a name for this element. Dubnium was finally named after the Russian city of Dubna, where the first atoms of this artificial, radioactive element were created, in 1968. However, a team of American scientists led by **Albert Ghiorso** also produced samples of the element at the same time. This radioactive element has 12 isotopes, or forms, with different numbers of neutrons.

106 Sg Seaborgium

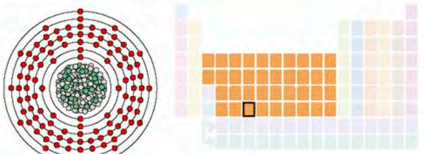

⊖ 106 ⊕ 106 ○ 163

State: Solid
Discovery: 1974

Atoms of seaborgium break apart in about three minutes, so little is known about it. Scientists think it may be a metal. The element was isolated in 1974 in a machine called the **Super Heavy Ion Linear Accelerator** at the Lawrence Berkeley National Laboratory. It was named after the US scientist Glenn T Seaborg.

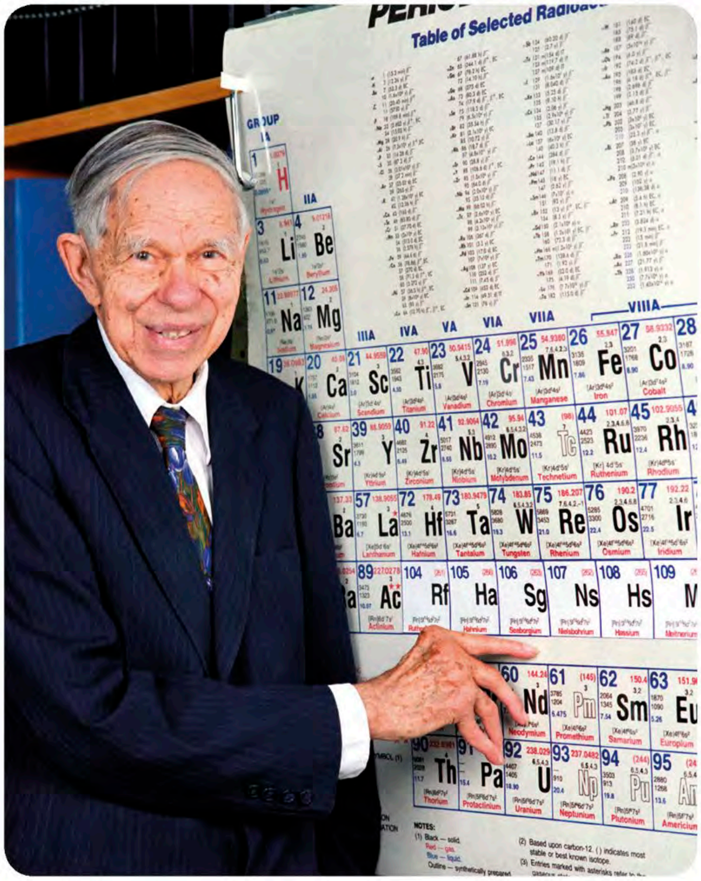

Glenn T Seaborg

This huge machine was used to discover **five new elements**.

***This giant tube** forms part of the Super Heavy Ion Linear Accelerator, which is a type of particle accelerator – a machine in which atoms are smashed together.*

Super Heavy Ion Linear Accelerator, Lawrence Berkeley National Laboratory, California, USA

NOBEL PRIZE IN CHEMISTRY

Glenn T Seaborg and his fellow US researcher Edwin McMillan were awarded the Nobel Prize for Chemistry in 1951 for their work in creating neptunium. This was the first element to be isolated that was heavier than uranium – the heaviest natural element.

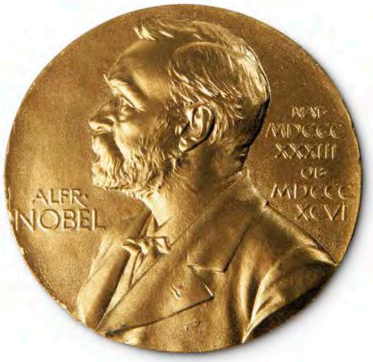

Nobel Prize medal

107

Bh

Bohrium

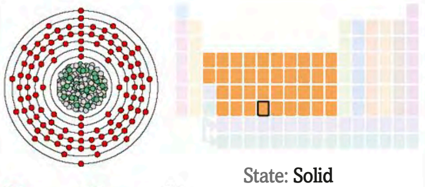

State: Solid

⊖ 107 ⊕ 107 ○ 163 Discovery: 1981

Niels Bohr

Bohrium is an artificial element named after the Danish scientist Niels Bohr. This was to honour his model of the structure of atoms' electron shells. Bohrium was first produced by firing chromium atoms at bismuth atoms in a particle accelerator (a machine in which atoms are smashed together). Atoms of this metal are unstable: half of any sample of bohrium atoms breaks apart in 61 seconds. As a result, it is not very well understood.

108

Hs

Hassium

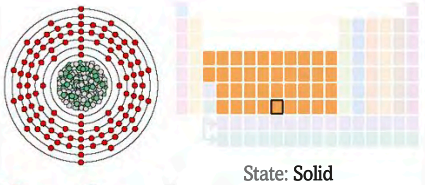

State: Solid

⊖ 108 ⊕ 108 ○ 169 Discovery: 1984

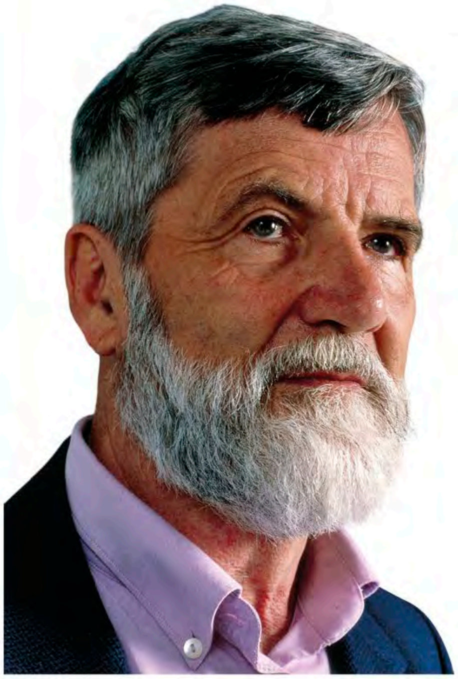

Peter Armbruster

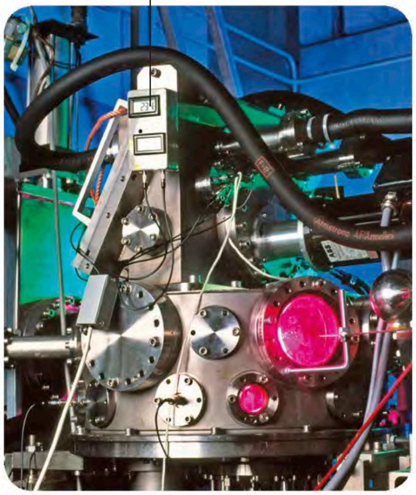

***Hassium** was produced inside this chamber.*

A chamber at Centre for Heavy Ion Research, Darmstadt, Germany

Scientists think hassium is a metal, but they have not been able to produce enough of its atoms to study it in any detail. Hassium is very radioactive, and most of its atoms break apart within a few seconds. This element is named after the German state of Hesse, the location of the **Centre for Heavy Ion Research**, where hassium was first created artificially by a team led by the German physicist **Peter Armbruster**.

109
Mt

Meitnerium

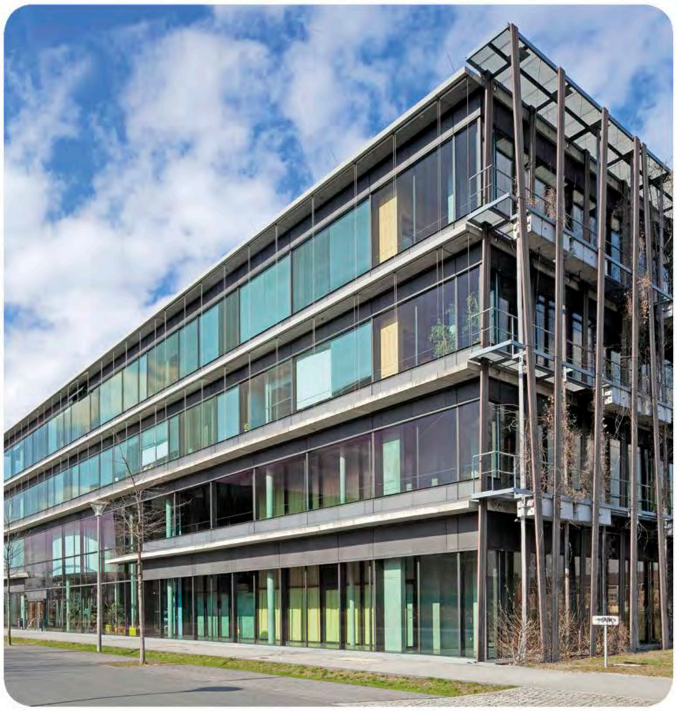

Meitner Haus building, Humboldt University, Germany

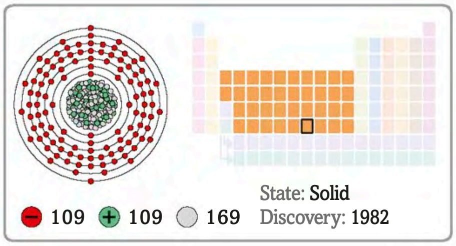

⊖ 109 ⊕ 109 ○ 169

State: Solid
Discovery: 1982

Researchers think meitnerium might be the densest of all elements. It is very unstable, and even the atoms of its most stable isotope, or form, break apart in a matter of seconds. Meitnerium is named after the Austrian physicist **Lise Meitner**, to honour her achievements in physics. Several universities, such as **Humboldt University** in Berlin, Germany, also have buildings in her name.

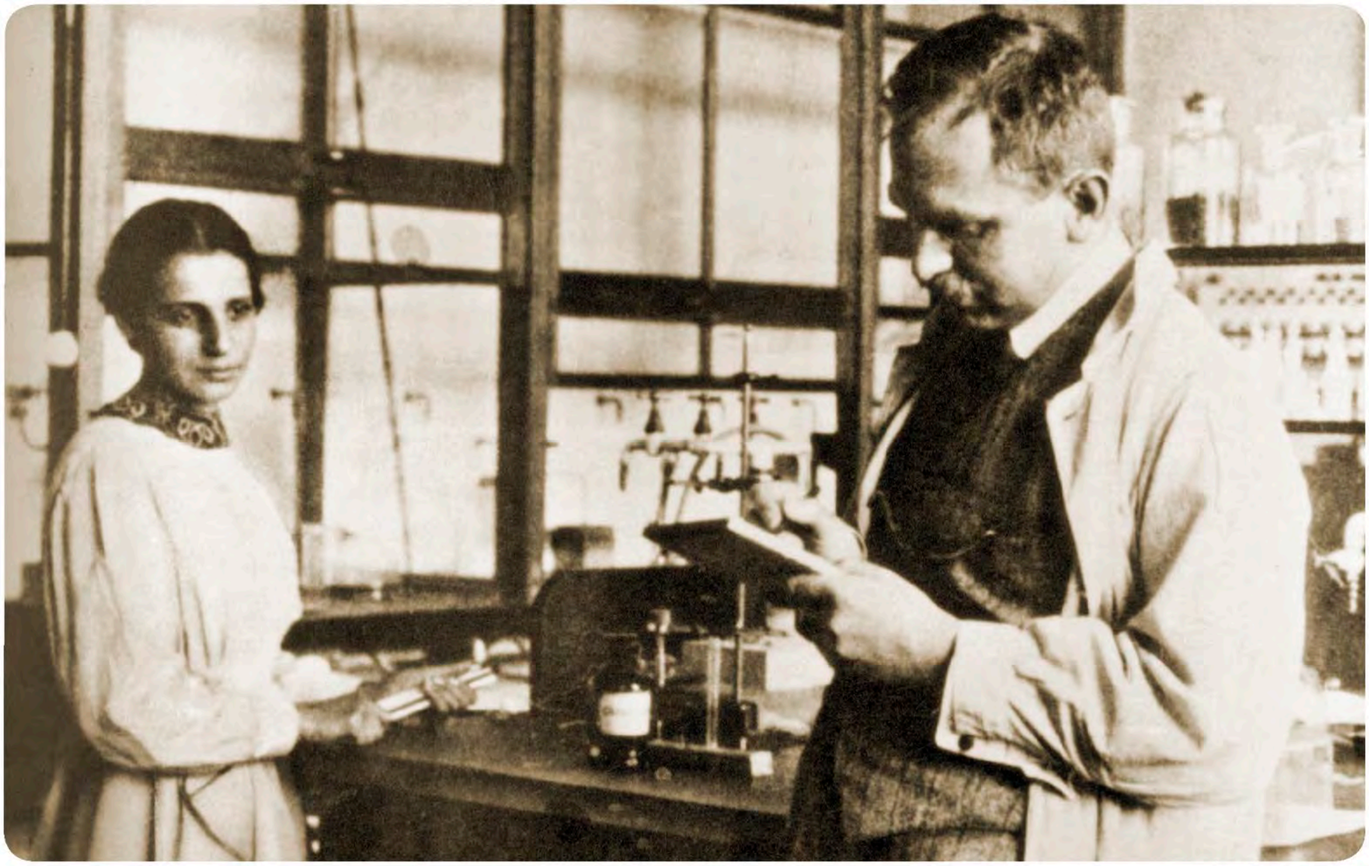

Lise Meitner (left) works with the German chemist Otto Hahn

110

Ds

Darmstadtium

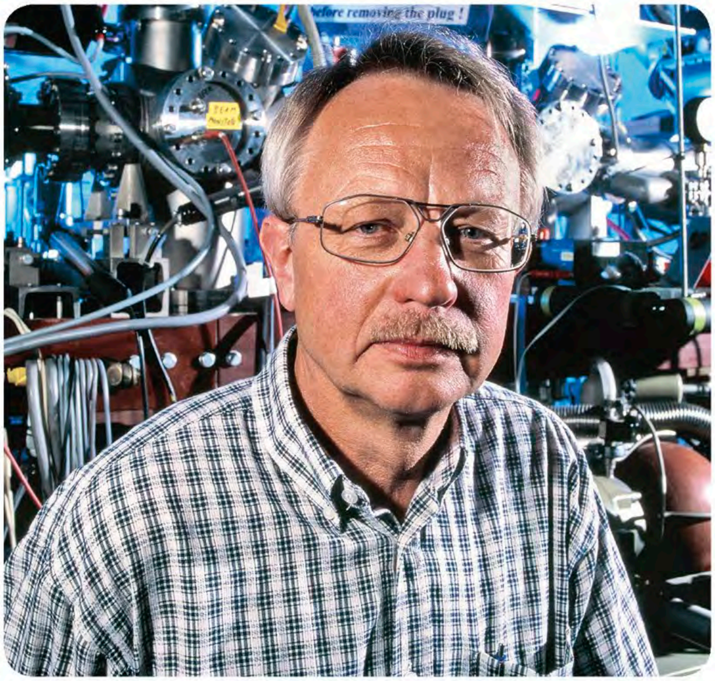

Sigurd Hofmann

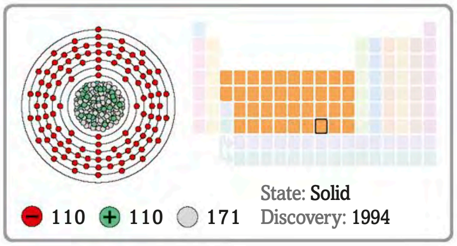

This artificial element is named after the German city of Darmstadt – the home of the Institute for Heavy Ion Research where this element was first produced. A team led by the German physicist **Sigurd Hofmann** created darmstadtium by smashing nickel atoms into lead atoms in a particle accelerator (a machine in which atoms are smashed together).

111

Rg

Roentgenium

Wilhem Röntgen

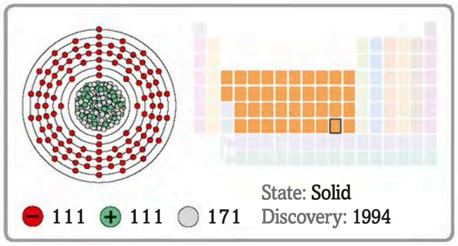

Scientists believe that this metal shares many characteristics with precious metals, such as gold and silver. However, its atoms break apart within seconds, so this has not yet been confirmed. Roentgenium was created in Darmstadt, Germany. It was named after **Wilhelm Röntgen**, the German scientist who discovered X-rays in 1895.

112 Cn Copernicium

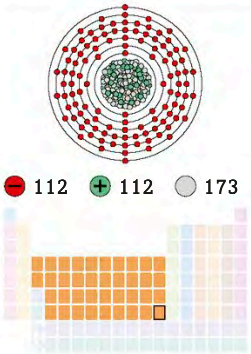

State: Solid
Discovery: 1996

Some scientists think copernicium could be the **only gaseous metal**.

This statue *stands in front of the Polish castle in which Copernicus lived.*
Statue of Nicolaus Copernicus

This German research institute *is where copernicium was discovered.*

Heavy Ion Research Centre, Germany

The atoms of this radioactive element survive only for a few minutes, before breaking down. Copernicium is created in particle accelerators by smashing together atoms of lead and zinc. Only a few atoms of this artificial element have ever been produced. Copernicium is named after **Nicolaus Copernicus**, the Polish astronomer who theorized that our planet orbits the Sun.

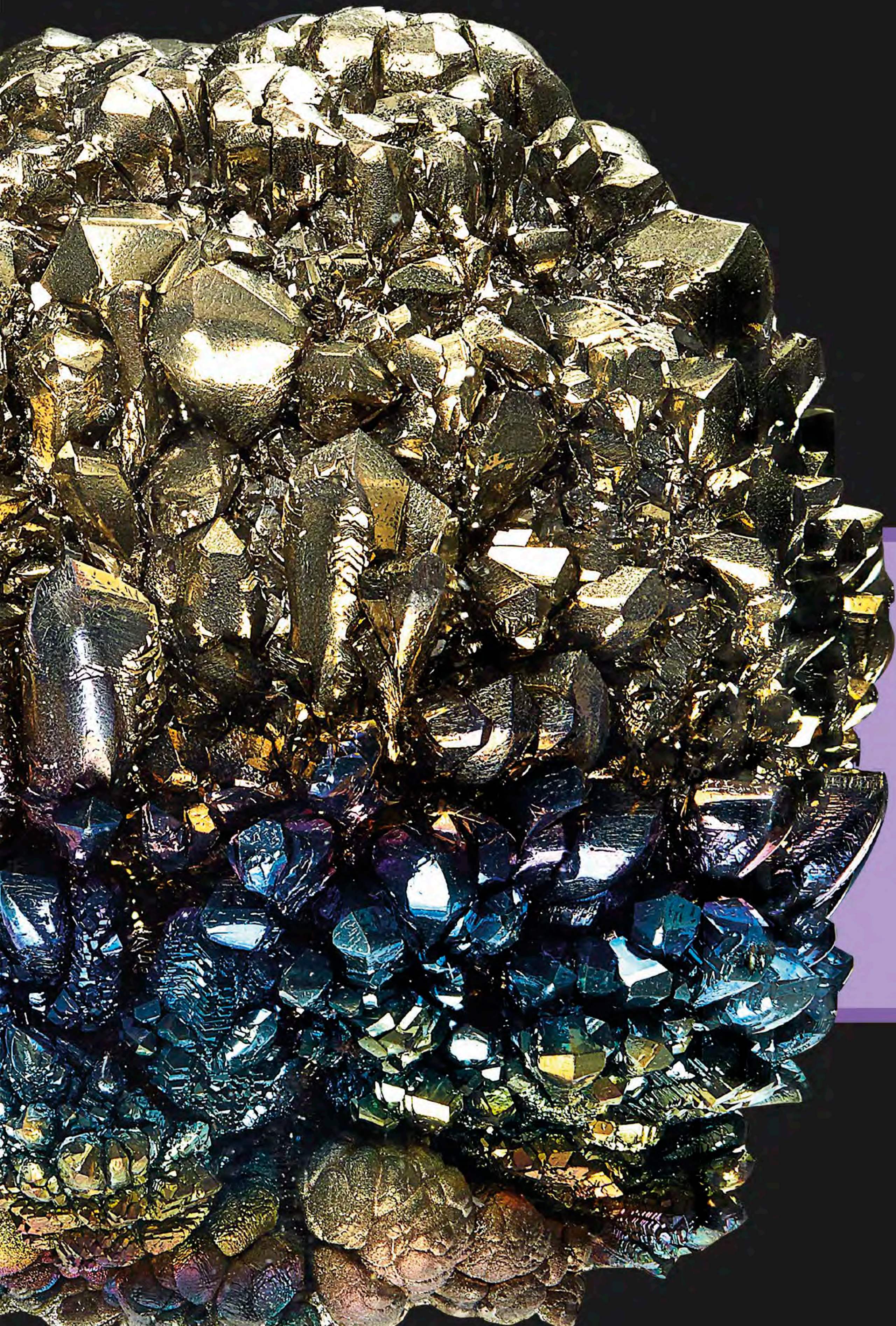

Europium's (Eu) colour changes when left in the air.

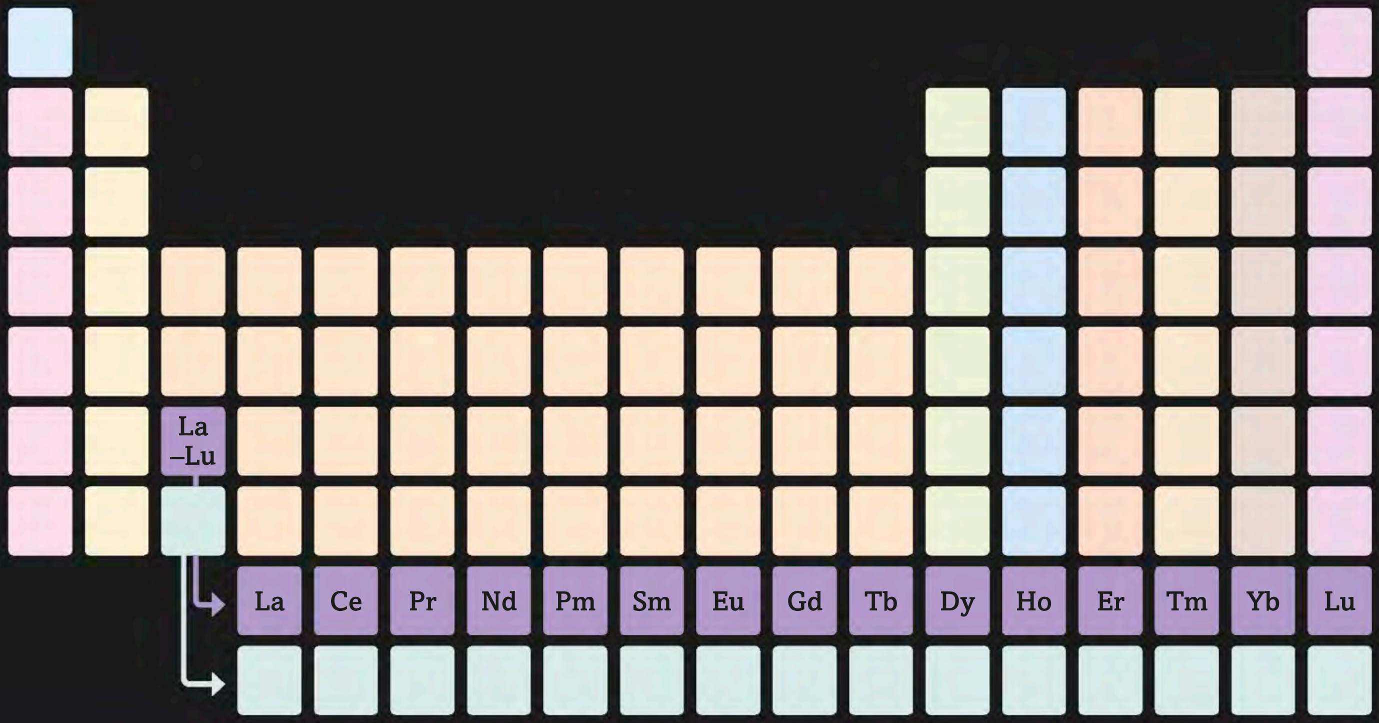

Lanthanides

This set is named after lanthanum, the first element in this series. The name "rare earth metals" is also given to these elements because they were discovered mixed together in complex minerals in Earth's crust, and were thought to be uncommon. However, they are actually not rare but abundant. These metals – between barium (Ba) and hafnium (Hf) – should fit between the alkaline earth metals and the transition metals, but they are normally shown underneath the main table to save on space.

Atomic structure
Atoms of every element in this group have two outer electrons. The lanthanides have large atoms, all with six electron shells.

Physical properties
The lanthanides are dense, shiny metals, which tarnish easily when exposed to air. They do not conduct electricity very well.

Chemical properties
These elements react slowly with oxygen (O) at room temperature, but the reactions speed up when heated.

Compounds
Many lanthanides form compounds with oxygen called oxides. These are often used in lasers and magnets.

57
La
Lanthanum

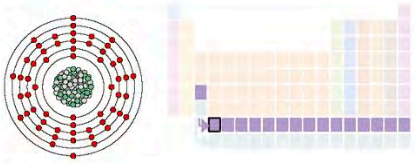

⊖ 57 ⊕ 57 ○ 82

State: Solid
Discovery: 1839

Forms

Lanthanum carbonate is used to treat patients with **kidney disease**.

Bastnasite

This reddish-brown mineral *is also found in other colours, including white, tan, and grey.*

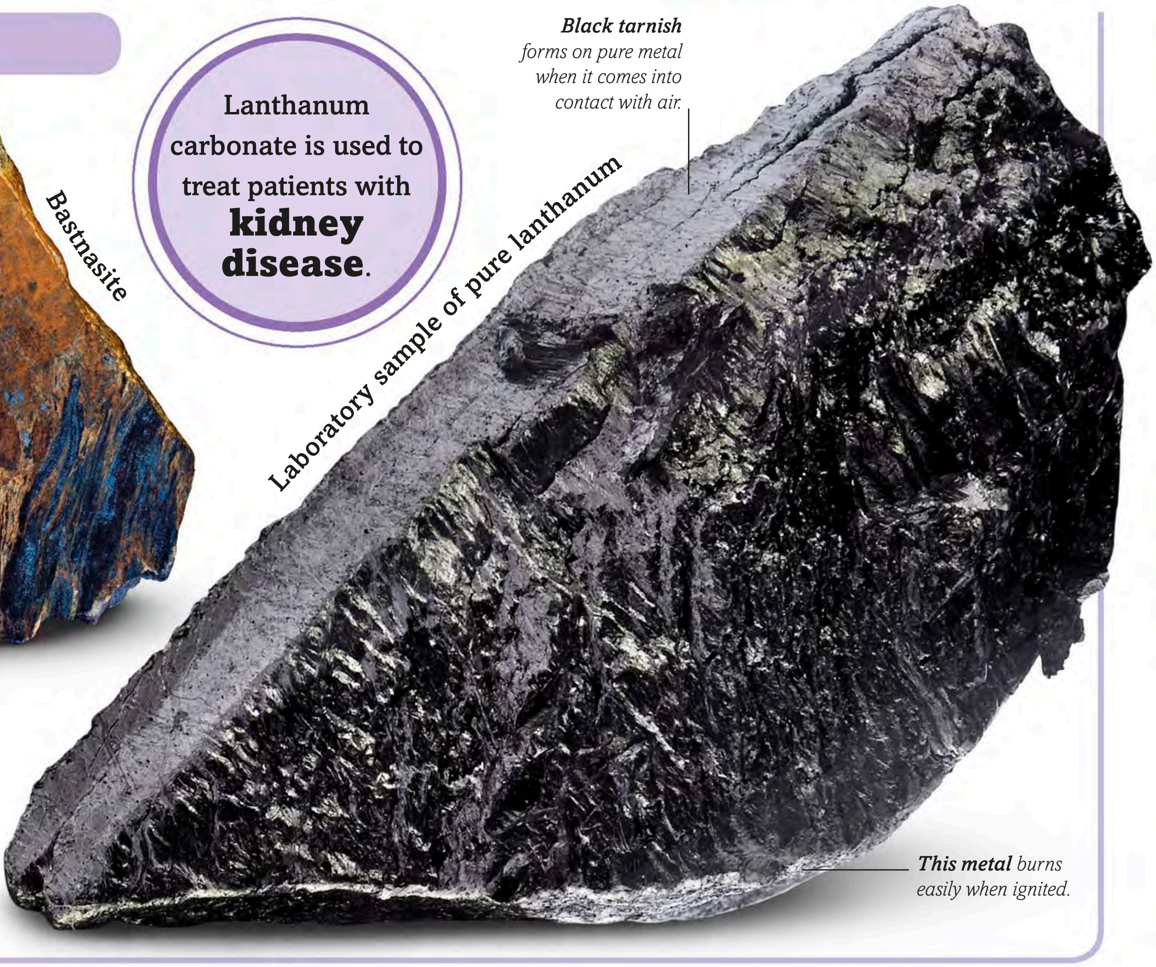

Laboratory sample of pure lanthanum

Black tarnish *forms on pure metal when it comes into contact with air.*

This metal *burns easily when ignited.*

Uses

Fluorescent lamp

This lamp *uses lanthanum to reduce the yellow colour in its light.*

Molten lanthanum

In its molten state*, lanthanum is used to smooth rough diamonds.*

Camera lens

This lens *can better focus light on an object due to the presence of lanthanum oxide in the glass.*

Although the word “lanthanum” means “to lie hidden”, it is more abundant than most metals. For example, it is three times more common than lead. This element was discovered in the mineral cerite in 1839. However, it took chemists almost another 100 years to find a way to purify the metal. Today, the mineral **bastnasite** is a source of **pure lanthanum**. The element’s applications range from its use in film studio lights and **lens-making** to refining petroleum.

58
Ce

Cerium

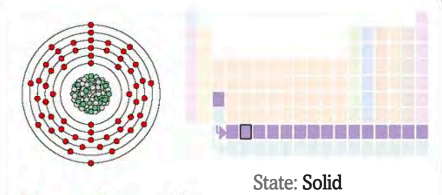

⊖ 58 ⊕ 58 ○ 82

State: Solid
Discovery: 1803

Cerium was the first of the lanthanides to be discovered. It is named after the dwarf planet Ceres, which was discovered two years before the element was isolated. Cerium is highly toxic when **pure**, but safer cerium compounds have some uses. The main use of cerium is in making phosphors, which are chemicals that produce lights of different colours. Phosphors are present in **flatscreen TVs** and bulbs.

Television

The inside of this screen *is coated with cerium-containing phosphors, which emit red, green, and blue light.*

Laboratory sample of pure cerium

The pure form *of the metal tarnishes on contact with air.*

Kitchen spatula

This red colour *comes from a compound called cerium sulfide.*

59
Pr

Praseodymium

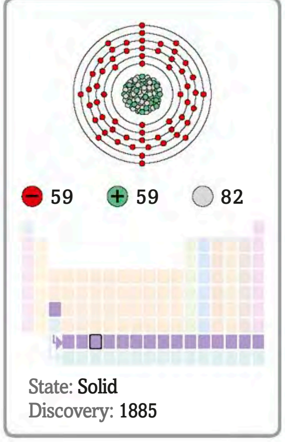

⊖ 59 ⊕ 59 ○ 82

State: Solid
Discovery: 1885

Part of this element's name comes from *prasinos*, the Greek word for "green". Normally a grey colour when pure, the element reacts slowly on contact with air to form a green coating. Praseodymium compounds give a yellow colour to glass and heat-resistant **ceramics**, and provide a green colour to some **artificial jewels**. This element also boosts the strength of magnets that contain it.

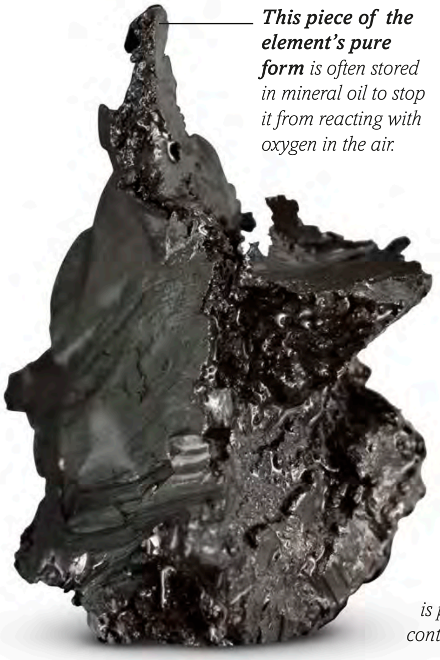

This piece of the element's pure form *is often stored in mineral oil to stop it from reacting with oxygen in the air.*

Laboratory sample of pure praseodymium

Yellow ceramic pot

This yellow colour *is produced by a solution containing praseodymium.*

Green cubic zirconia

This artificial gem *gets its green colour from tiny amounts of a compound of praseodymium and oxygen.*

60 Nd Neodymium

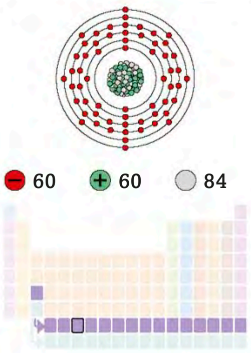

State: Solid
Discovery: 1885

Strong magnets made of neodymium can be used to lift thousands of times their own mass. This element was discovered in 1885 by the Austrian chemist Carl Auer von Welsbach, and it was originally used to colour glass. Small amounts of neodymium turn **glass** pinkish purple. Today, this element is also employed in lasers used in eye surgery.

Pink glass

***This glass** gets its colour from very tiny amounts of neodymium.*

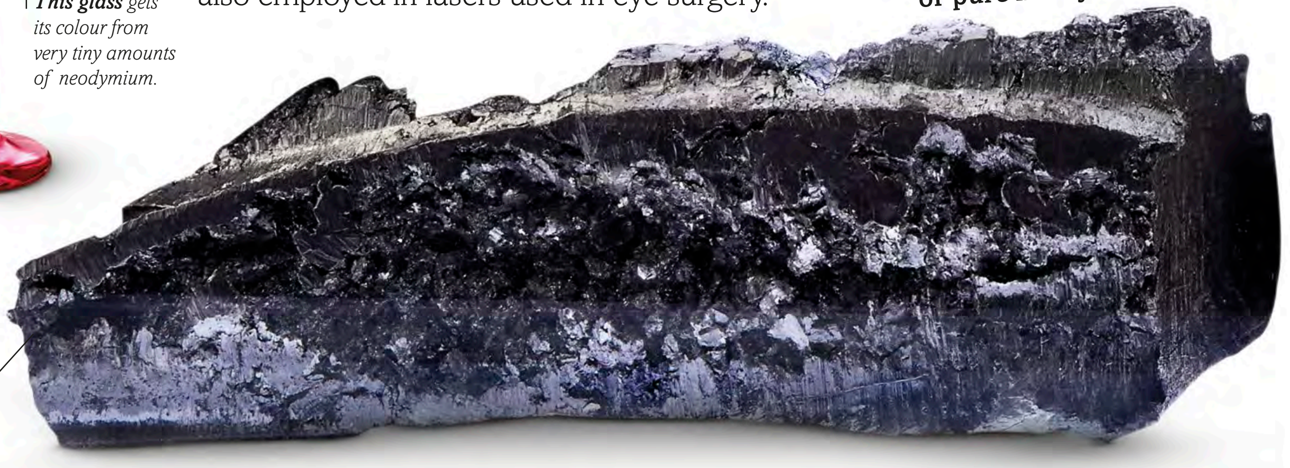

Laboratory sample of pure neodymium

***The pure element** turns black when it reacts with air.*

61 Pm Promethium

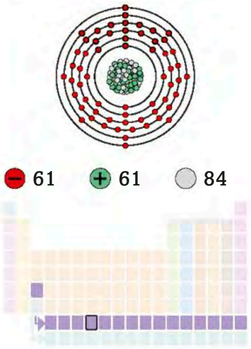

State: Solid
Discovery: 1945

***This paint** glows as a result of radioactive promethium.*

Promethium-rich paint in a tin seen from above

Missile

***This missile** uses radioactive promethium for electrical power.*

Promethium is the rarest lanthanide element. Any promethium that was in Earth's rocks decayed billions of years ago. Promethium is therefore produced artificially in nuclear reactors. Being very radioactive, it is used in some **missiles**, because it converts this radioactivity into electrical power. The addition of promethium also makes some **paints** glow in the dark.

62
Sm
Samarium

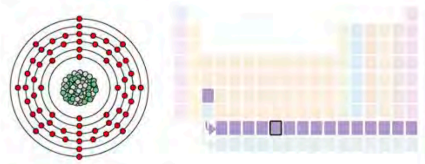

⊖ 62 ⊕ 62 ○ 88

State: Solid
Discovery: 1879

This element is named after the mineral samarskite from which it was first purified. However, another lanthanide-rich mineral called monazite is the main source of this element today. Samarium is mixed with cobalt to make permanent magnets that are often used in **electric guitars**.

Guitar pickups

These pickups *(components that sense vibrations produced by guitar strings) are made of samarium-cobalt magnets.*

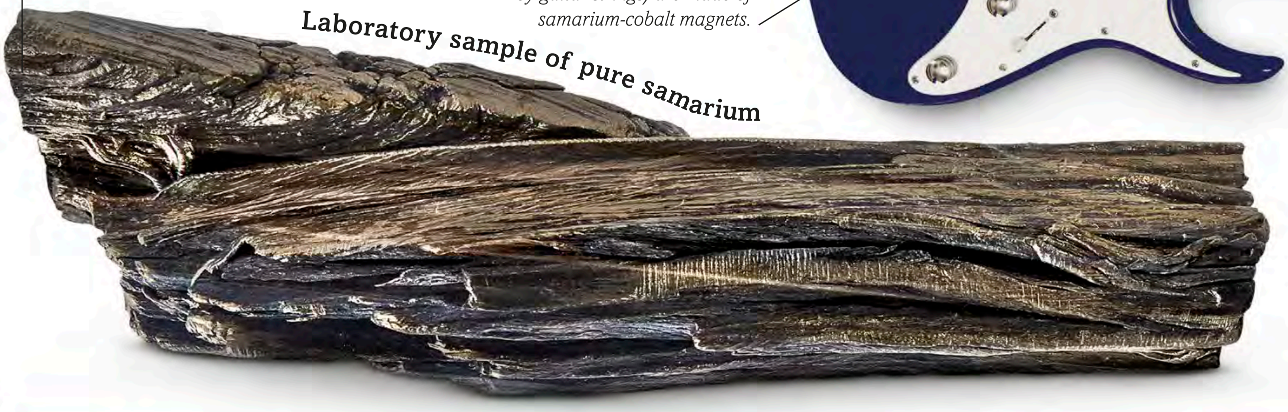

Laboratory sample of pure samarium

This silvery white metal *darkens on contact with air.*

63
Eu
Europium

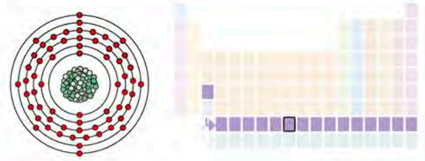

⊖ 63 ⊕ 63 ○ 89

State: Solid
Discovery: 1901

Europium was named after the continent of Europe. However, most of the world's supply of the element comes from the USA and China, where the mineral bastnasite is mined for the extraction of **pure europium**. A compound called europium oxide is used in euro and **British bank notes**. When placed under ultraviolet (UV) light, the compound gives off a red glow.

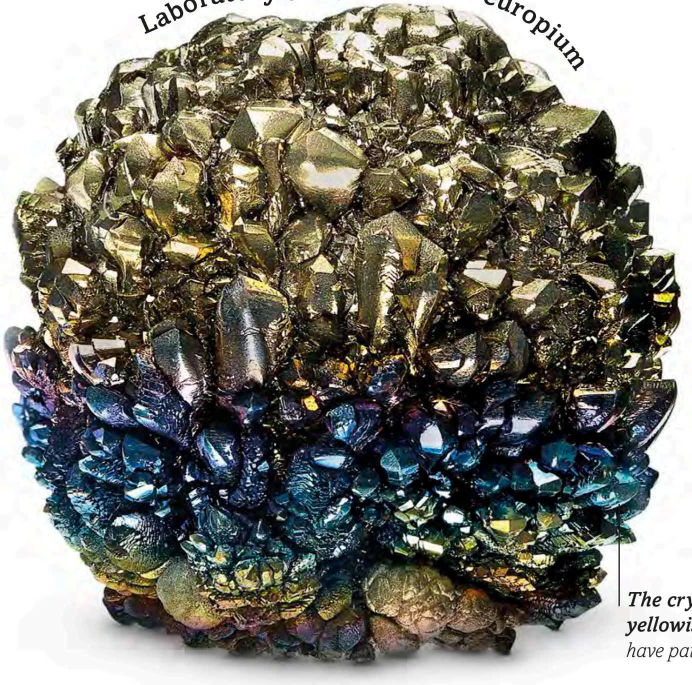

Laboratory sample of pure europium

The crystals of this yellowish metal *often have patches of dark oxides.*

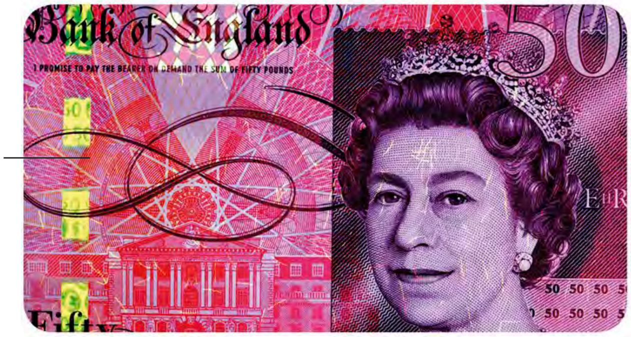
Bank of England
I PROMISE TO PAY THE BEARER ON DEMAND THE SUM OF FIFTY POUNDS
50
Fifty

This red glow *proves this note is real.*

Section of British note under UV light

64 Gd Gadolinium

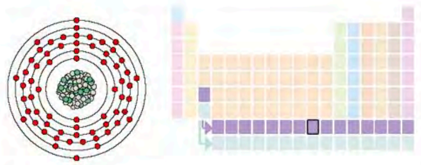

⊖ 64 ⊕ 64 ○ 93

State: Solid
Discovery: 1880

Gadolinium, and its mineral ore gadolinite, are named after the Finnish chemist Johan Gadolin, who discovered the element. Gadolinium compounds are used to obtain clear **MRI scans**. It is also used in electronics, and to make rust-resistant steel.

Laboratory sample of pure gadolinium

This soft, silvery metal *darkens when exposed to air.*

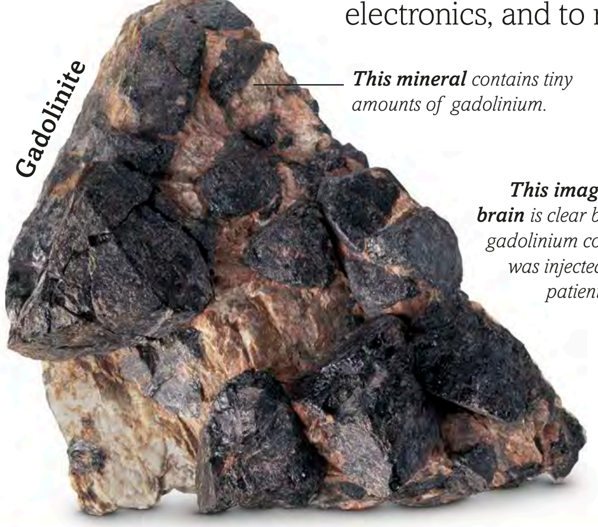

Gadolinite

This mineral *contains tiny amounts of gadolinium.*

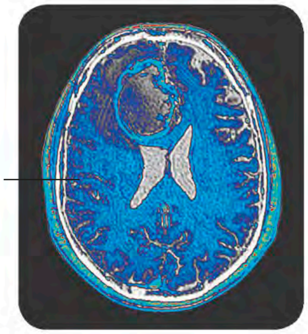

MRI scan of the human brain

This image of the brain *is clear because a gadolinium compound was injected into the patient's blood.*

65 Tb Terbium

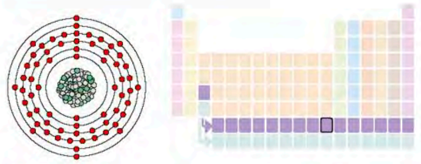

⊖ 65 ⊕ 65 ○ 94

State: Solid
Discovery: 1843

Terbium is named after the village of Ytterby in Sweden. It is a silvery metal that can be obtained from the ore monazite. This element has only a few uses. **Pure terbium** is added to other metals to make powerful magnets used in sound-producing devices, such as the **SoundBug™**. Its compounds are used to line **mercury lamps**.

SoundBug™ device

This device *uses magnets to turn any flat surface, like a window, into a loudspeaker.*

Laboratory sample of pure terbium

The pure metal *is soft enough to be cut with a knife.*

Mercury lamp

The mercury vapour in this lamp *produces ultraviolet light when electrified, and this is turned into a bright yellow glow by terbium.*

66
Dy

Dysprosium

⊖ 66 ⊕ 66 ○ 97

State: Solid
Discovery: 1886

Fergusonite

This mineral *contains tiny amounts of dysprosium.*

This pure metal *remains shiny at room temperature.*

Dysprosium reacts more easily with air and water than most other lanthanide metals. Although it was discovered in 1886, it took until the 1950s to purify it. This metal is often used with neodymium to produce magnets that are used in **car batteries**, wind turbines, and generators.

Laboratory sample of pure dysprosium

Some hybrid car batteries *contain dysprosium.*

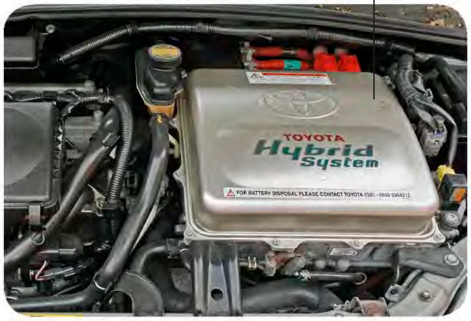

Hybrid car battery

67
Ho

Holmium

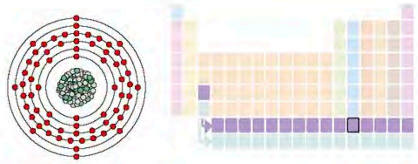

⊖ 67 ⊕ 67 ○ 98

State: Solid
Discovery: 1878

The Swedish chemist Per Teodor Cleve named holmium after the Swedish city of Stockholm. Pure holmium can produce a strong magnetic field and is therefore used in magnets. Its compounds are used to make lasers, and to colour glass and artificial jewels, such as **cubic zirconia**.

This artificial gemstone *is coloured red by small amounts of holmium.*

Red zirconia gemstone

Bright, silver shine

Laboratory sample of pure holmium

68

Er

Erbium

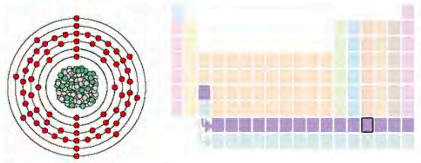

⊖ 68 ⊕ 68 ○ 99

State: **Solid**

Discovery: **1843**

This silvery element *slowly tarnishes on contact with air.*

Laboratory sample of pure erbium

This glass *contains erbium, which protects a welder's eyes from heat and bright light.*

Welding goggles

The rose pink finish *of this vase is from an erbium chloride glaze.*

Pink pottery

Like terbium and ytterbium, erbium is also named after the Swedish village of Ytterby, near which it was discovered. This element does not occur in its **pure form** in nature, but it can be obtained from the mineral monazite. Many erbium compounds are pink in colour and are used to colour **pottery** and glass.

69

Tm

Thulium

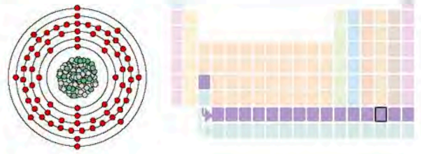

⊖ 69 ⊕ 69 ○ 100

State: **Solid**

Discovery: **1879**

Laboratory sample of pure thulium

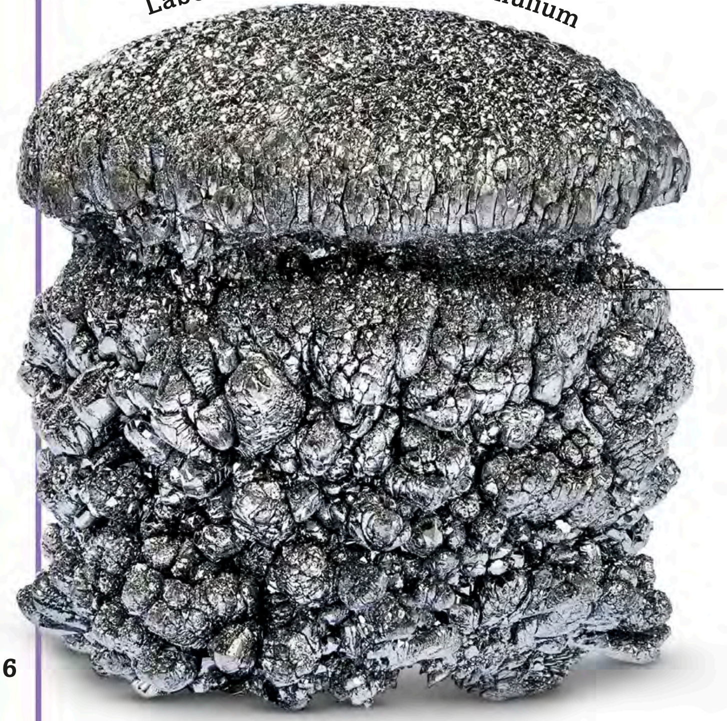

This soft metal *glows blue under ultraviolet (UV) light.*

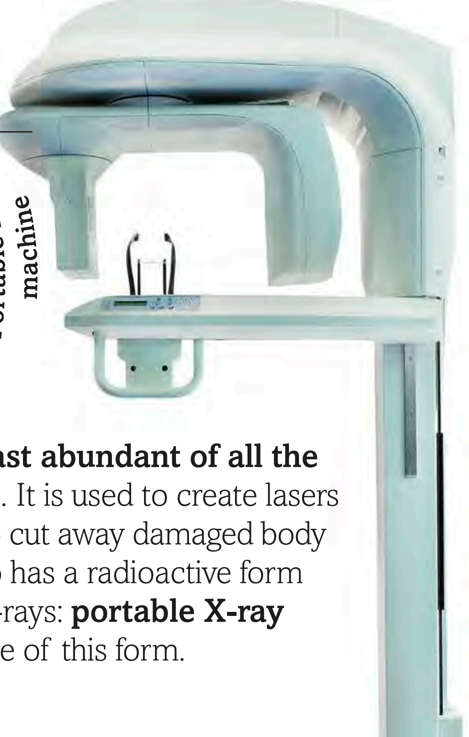

This machine *emits X-rays using a very small amount of thulium.*

Portable X-ray machine

Thulium is the least abundant of all the lanthanide metals. It is used to create lasers that surgeons use to cut away damaged body tissue. Thulium also has a radioactive form that can produce X-rays: **portable X-ray machines** make use of this form.

70
Yb

Ytterbium

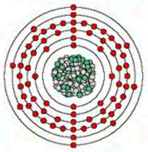

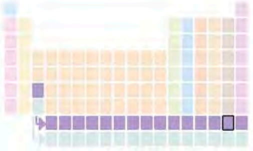

State: Solid
⊖ 70 ⊕ 70 ○ 103 Discovery: 1878

Ytterbium tends to be more reactive than other lanthanide metals. It is stored in sealed containers to stop the metal from reacting with oxygen. The **pure metal** has only a few uses. A small amount of ytterbium is used in making steel, while its compounds are used in some **lasers**.

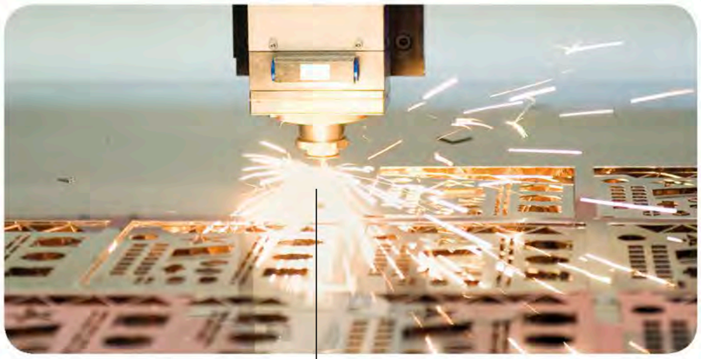

Laser cutting

An ytterbium laser *can cut through metals and plastics.*

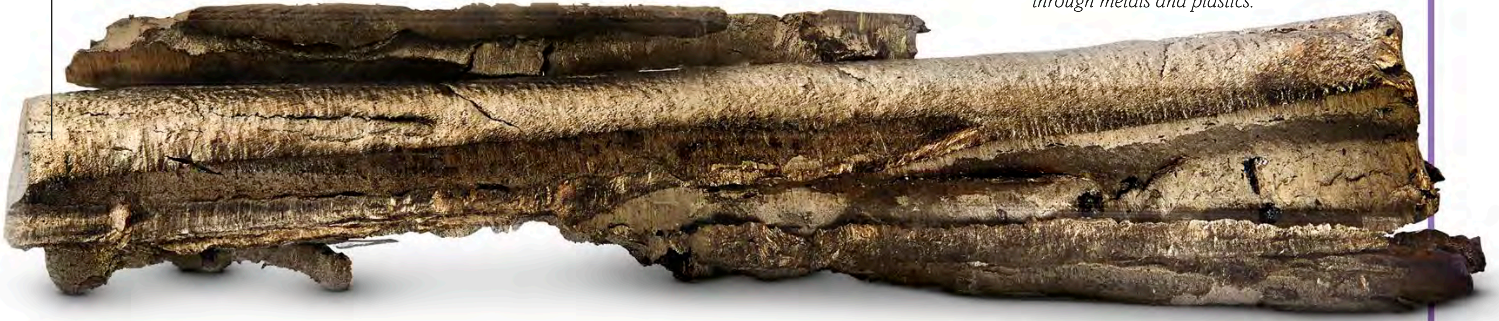

Laboratory sample of pure ytterbium

This bright, shiny metal *can be hammered into thin sheets.*

71
Lu

Lutetium

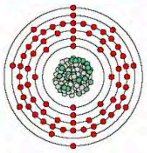

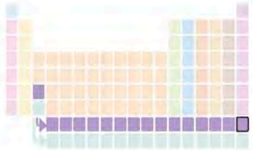

State: Solid
⊖ 71 ⊕ 71 ○ 104 Discovery: 1907

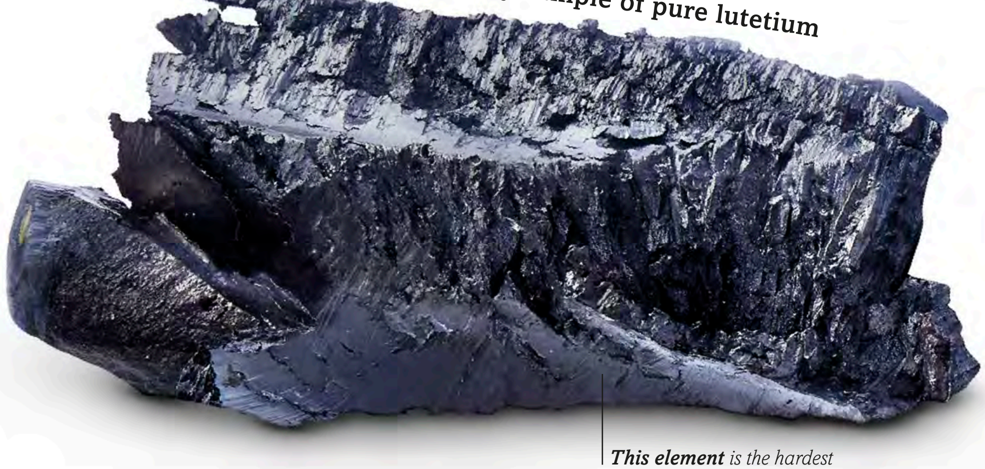

Laboratory sample of pure lutetium

This element *is the hardest and densest lanthanide metal.*

Lutetium was the last of the rare earth metals to be discovered. It is also the final member of the lanthanides. In its **pure form**, lutetium is very reactive and catches fire easily. It is rare and has few uses, mainly as a substance mixed with crude **oil**.

Oil refinery

Some oil refineries *use lutetium to break down crude oil to make fuels, such as petrol and diesel.*

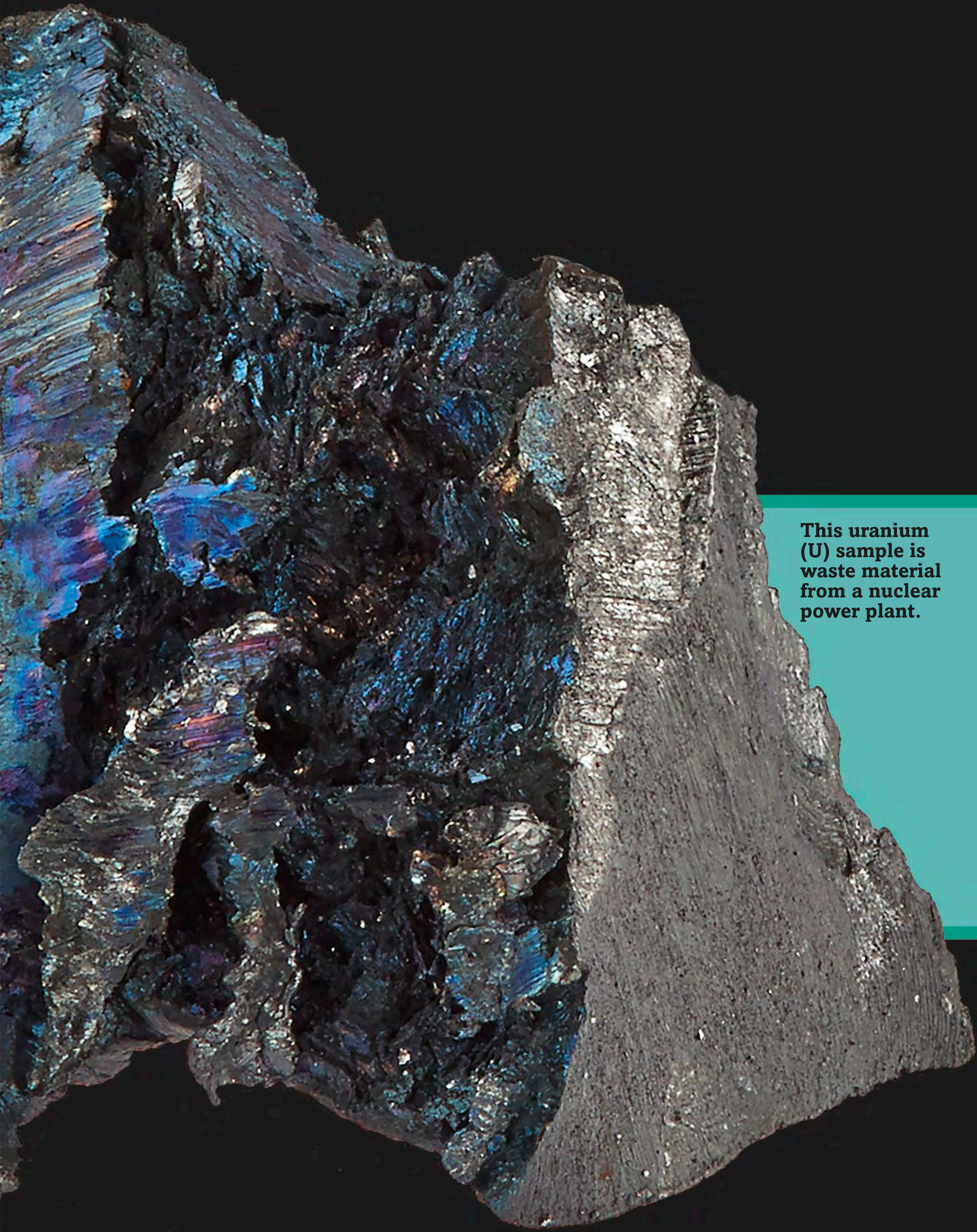

This uranium (U) sample is waste material from a nuclear power plant.

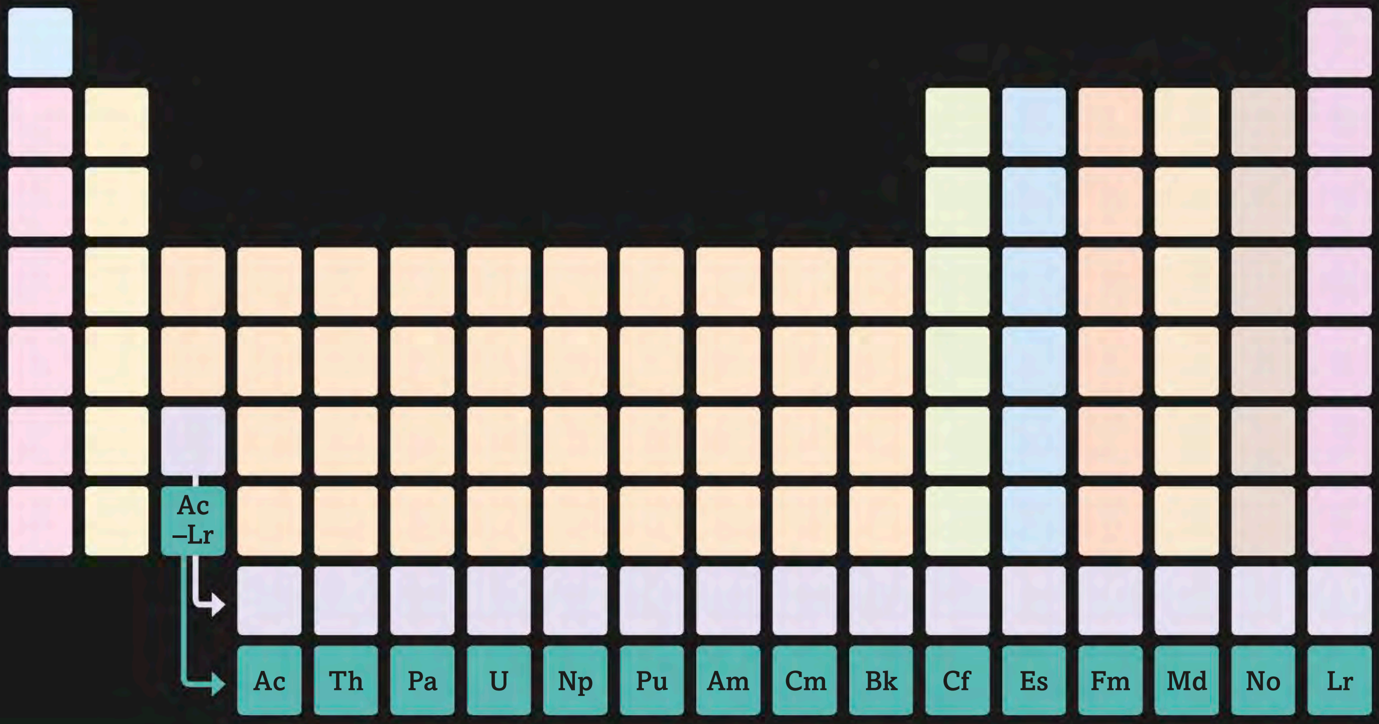

Actinides

These metals are named after actinium (Ac), the first member of the group. Although this group is often shown as the bottom row in the periodic table, to save space, they actually sit between radium (Ra), an alkali earth metal, and Rutherfodium (Rf), a transition metal. All the elements in this group are radioactive, and the final nine members are artificially produced in laboratories.

Atomic structure
All the elements in this group have two electrons in their outer shell. Their atoms all have seven electron shells.

Physical properties
Natural actinides are dense metals with high melting points. The physical properties of most of the artificial ones are unknown.

Chemical properties
The actinides are reactive metals and are never found in pure form in nature. They react easily with air, the halogens, and sulfur (S).

Compounds
Actinides form colourful compounds with halogens. Most actinide ores also contain compounds of oxygen (0) called oxides.

89 Ac Actinium

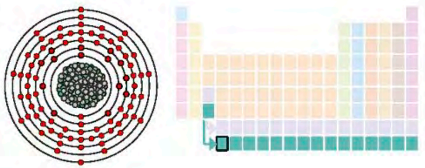

⊖ 89 ⊕ 89 ○ 138 State: Solid Discovery: 1899

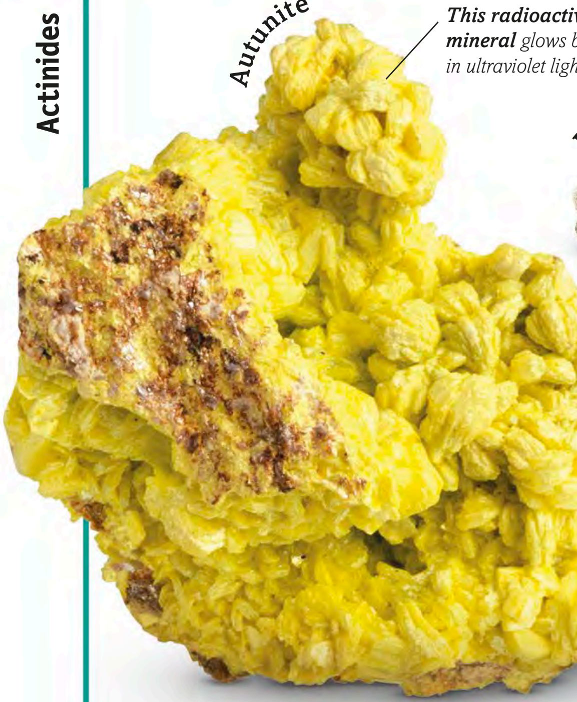

This radioactive mineral *glows brightly in ultraviolet light.*

This ore *contains uranium, which breaks down into actinium.*

This device *uses radioactive actinium to measure the amount of water.*

Neutron probe

Rare in nature, actinium is a metal formed by the decay of other radioactive elements. Its atoms are unstable and break down to make the elements francium and radon. Actinium is found in tiny amounts in uranium ores, such as **uranite**, and has limited applications. Its isotopes are used in radiation therapy to treat cancer.

90 Th Thorium

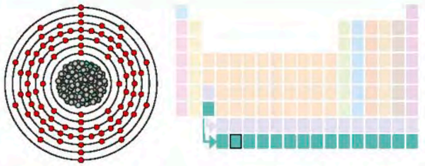

⊖ 90 ⊕ 90 ○ 142 State: Solid Discovery: 1829

The most common natural radioactive metal, thorium is used inside vacuum tubes to allow an electric current to flow. It can also undergo nuclear fission, a process in which atoms split in two and release energy. Scientists are exploring ways of making thorium-powered nuclear reactors that produce electricity.

This durable rock *made of solidified lava contains 12% thorium.*

This thorium coating *creates an electric current by releasing electrons.*

Thorianite

This ore *contains small crystals of thorium compounds.*

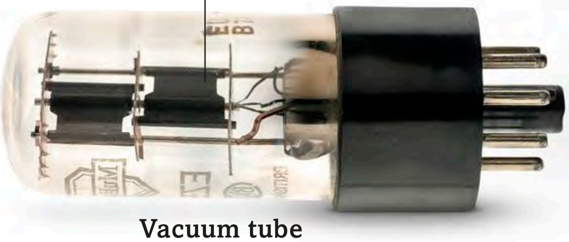

Vacuum tube

91
Pa Protactinium

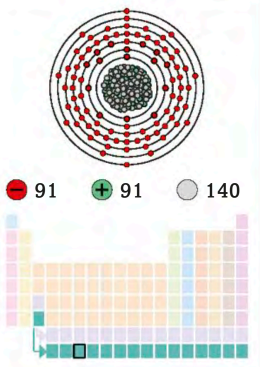

State: Solid
Discovery: 1913

This vibrant green radioactive mineral *contains tiny amounts of protactinium.*

This brittle, shiny ore *feels waxy.*

Torbernite

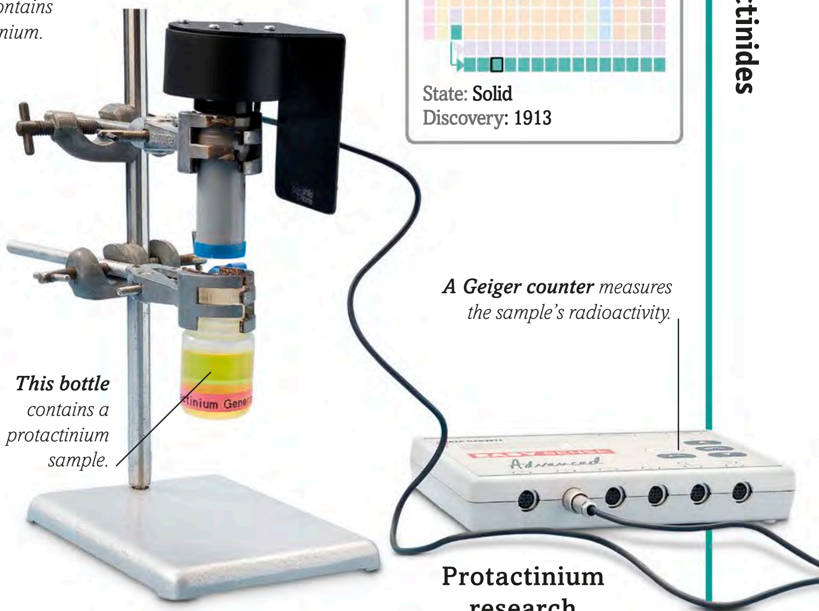

A Geiger counter *measures the sample's radioactivity.*

This bottle *contains a protactinium sample.*

Protactinium research

The name protactinium means "before actinium". This is because a uranium atom decays to form a protactinium atom, which then quickly breaks down into an actinium atom. Small quantities of protactinium are found in ancient sands and mud. Geologists use Geiger counters to carry out **research** to calculate how old the sands are.

These used nuclear fuel rods *contain protactinium.*

Nuclear waste

92 U Uranium

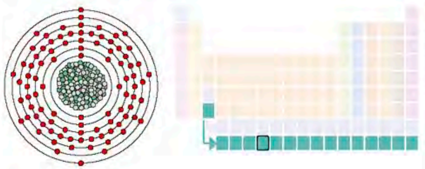

State: Solid
⊖ 92 ⊕ 92 ○ 146 Discovery: 1789

Named after the planet Uranus, uranium was the first known radioactive element. In the early 20th century, some manufacturers used uranium in **glass bowl** glazes, only to realize later that it was a harmful metal. An unstable form, called uranium-235, is used as fuel in nuclear reactors and in atomic bombs.

Chunk of pure uranium

This sample of pure uranium *is waste from a nuclear plant.*

These black sections *contain uranium dioxide, which is the main source of uranium.*

Uraninite

Uranium mixed into glass *makes this bowl glow bright green under ultraviolet (UV) lamps.*

Glass bowl

93 Np Neptunium

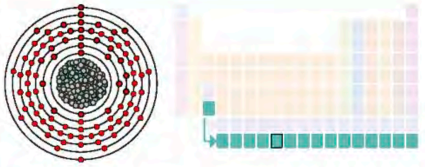

State: Solid
⊖ 93 ⊕ 93 ○ 144 Discovery: 1940

The radioactive elements *in this mineral decay to form neptunium.*

Uraninite

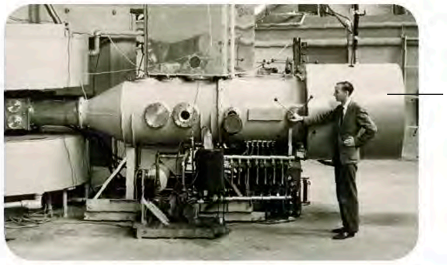

This cyclotron*, built in 1938, was used to discover neptunium.*

Cyclotron at the University of California, Berkeley, USA

Sitting next to uranium in the periodic table, neptunium was named after the planet Neptune. It exists in small amounts in radioactive ores, such as aeschynite. It forms during nuclear explosions and was first identified inside a machine called a **cyclotron**. There are no known uses for neptunium.

94
Pu

Plutonium

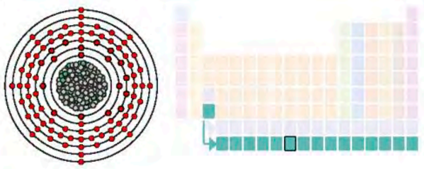

⊖ 94 ⊕ 94 ○ 239 State: Solid Discovery: 1940

Uraninite

Hardly any plutonium exists in nature: most of it has decayed into other elements over time. It was discovered during the development of nuclear bombs in World War II. Today, plutonium is used mostly as a nuclear fuel.

This ore *contains trace amounts of plutonium.*

This plutonium battery *was used in early pacemakers.*

CAUTION
RADIOACTIVE PLUTONIUM -238
LESS THAN 3 CURIES 1973
DO NOT DISCARD CONTACT
NUCLEAR BATTERY CORP
COLUMBIA, MARYLAND
DATE OF MANUFACTURE 1973

1970's pacemaker battery

This Martian rover *uses the heat given off by a supply of plutonium to generate electrical power.*

Curiosity Rover

95
Am

Americium

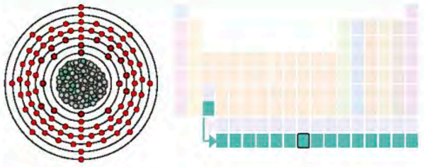

⊖ 95 ⊕ 95 ○ 148 State: Solid Discovery: 1944

Smoke detector component

This metallic element is not found in nature. Instead, it is produced inside nuclear reactors when uranium or plutonium atoms are bombarded with neutrons. Remarkably, americium is the most common radioactive element used in the home. Radioactivity emitted by americium atoms causes the air inside **smoke detectors** to conduct electricity. When smoke disrupts the electric current, an alarm goes off.

This smoke detector *contains tiny, harmless quantities of americium.*

96 Cm Curium

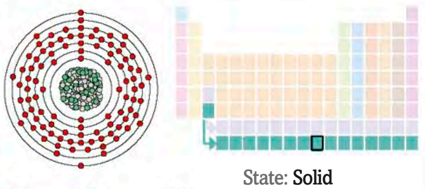

⊖ 96 ⊕ 96 ○ 151

State: Solid
Discovery: 1944

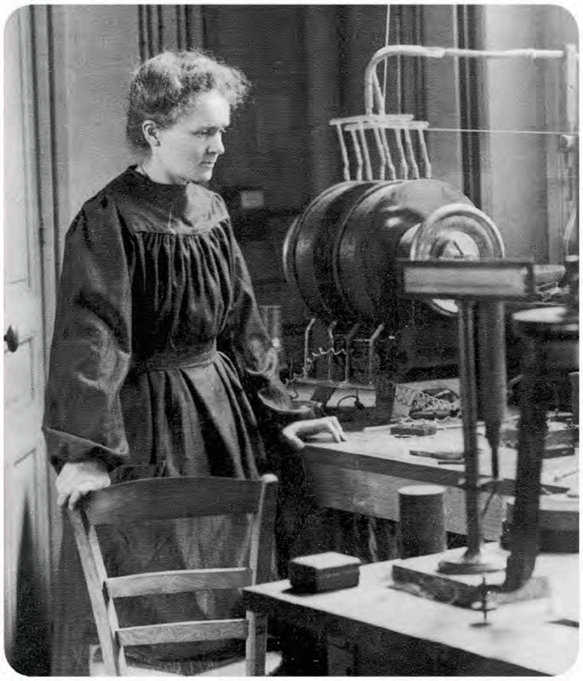

Marie Curie working in her laboratory

This lander studied the surface composition of the comet 67P.

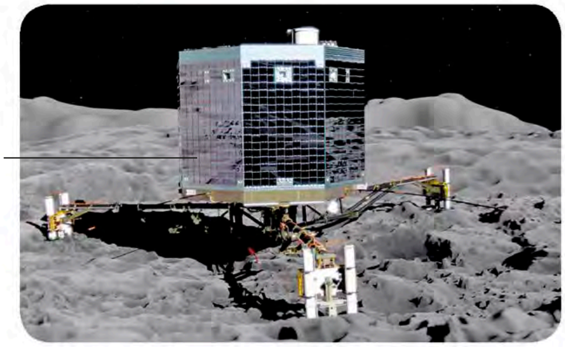

Philae lander

Curium is a silvery, radioactive metal that glows reddish purple in the dark. This element was discovered by the US scientist Glenn T Seaborg at the University of California. It was named after **Marie Curie**, the scientist who discovered the element polonium. Several space probes, such as the **Philae comet lander**, use X-ray devices containing curium to study their environment.

97 Bk Berkelium

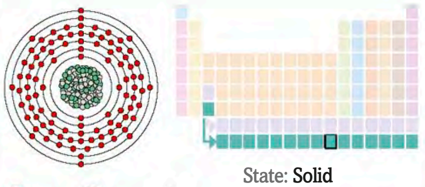

⊖ 97 ⊕ 97 ○ 150

State: Solid
Discovery: 1949

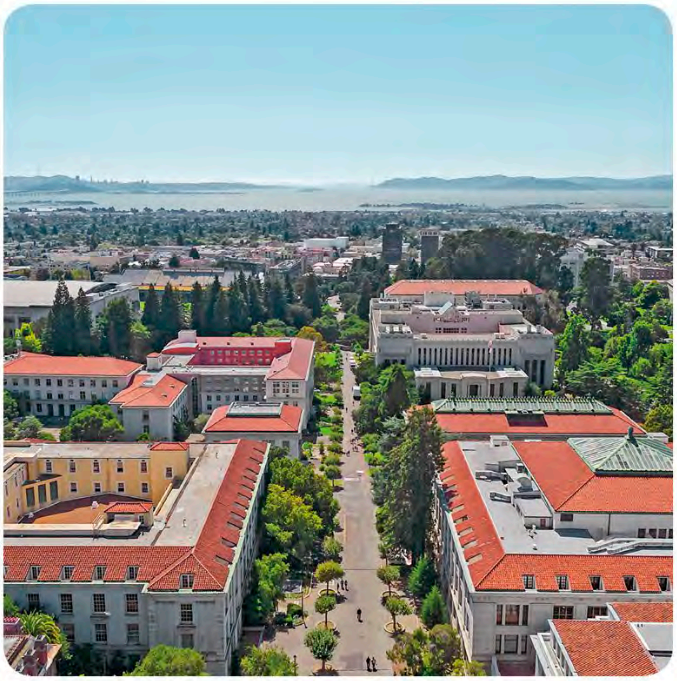

University of California, Berkeley campus, USA

This element was named after the city of Berkeley – home to the University of California – where this artificial element was discovered. It was first synthesized by **Glenn T Seaborg**. Berkelium has no uses other than the creation of heavier elements, such as tennessine.

Seaborg helped develop the atom bomb, but **opposed using it** in World War II.

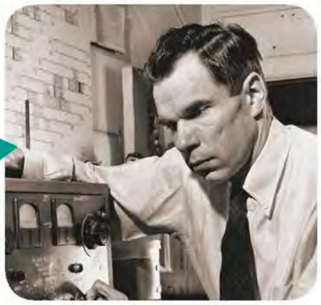

Glenn T Seaborg

98
Cf

Californium

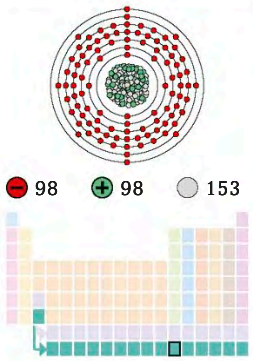

State: Solid
Discovery: 1950

This machine *uses californium to find water underground.*

Water detector

This isotope, or form, of californium *produces a lot of neutrons.*

Pellets of radioactive californium

Californium is named after the US state of California. This soft, silvery metal does not exist in nature and is made by smashing berkelium atoms with neutrons in a particle accelerator (a machine in which atoms are smashed together). This **radioactive element** is used in the treatment of cancer.

99
Es

Einsteinium

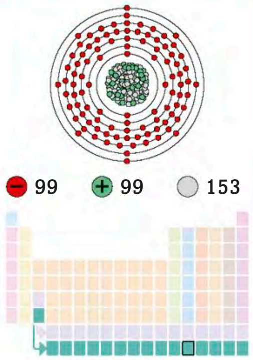

State: Solid
Discovery: 1952

Albert Einstein in his study

Only a **few milligrams** of einsteinium are made **every year**.

Einsteinium was discovered in the chemicals left over after the first hydrogen bomb test in 1952. The huge explosion fused smaller atoms together to make larger ones, including einsteinium. This element was named after the great German-born scientist **Albert Einstein**, and was found to be a silvery, radioactive metal that glows blue in the dark. It is only used for making heavier elements, such as mendelevium.

100 **Fm** Fermium

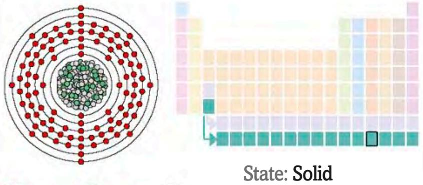

⊖ 100 ⊕ 100 ○ 157

State: Solid
Discovery: 1953

This artificial element was named after the Italian scientist Enrico Fermi. He built the first nuclear reactor in 1942, starting the American effort to build nuclear weapons during World War II. Fermium was first identified in the debris of an atom bomb test in 1953. This unstable element has no known uses beyond research.

Some scientists call Enrico Fermi the "father of the **atomic age**".

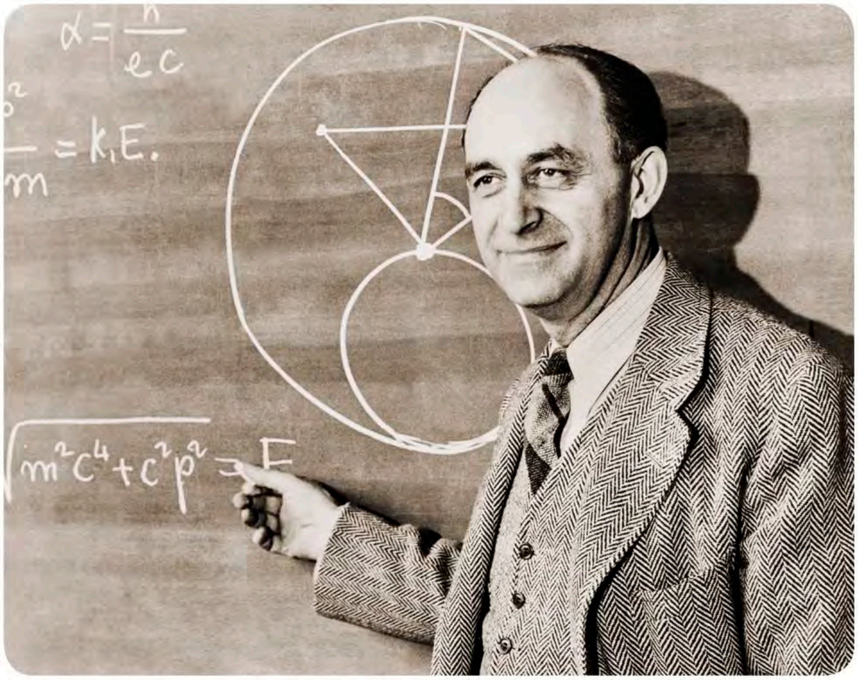

Enrico Fermi

101 **Md** Mendelevium

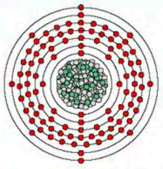

⊖ 101 ⊕ 101 ○ 157

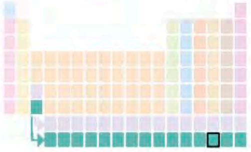

State: Solid
Discovery: 1955

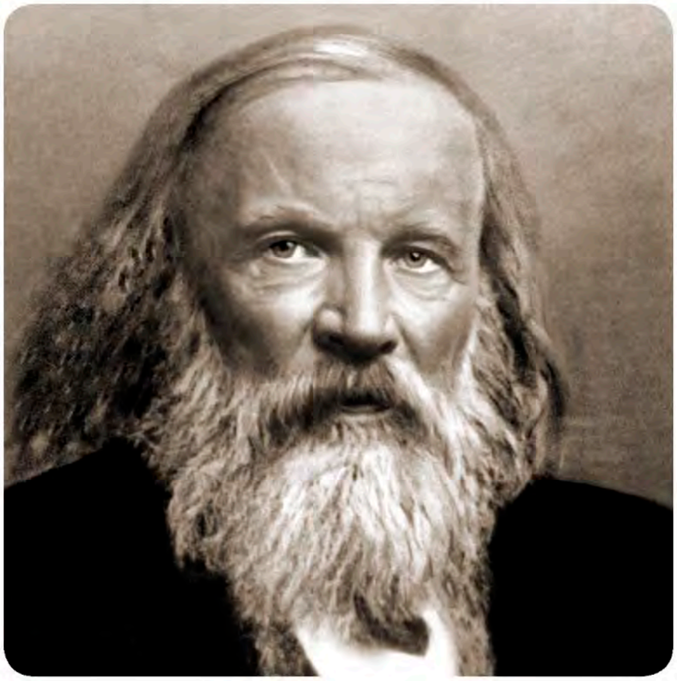

Dmitri Mendeleev

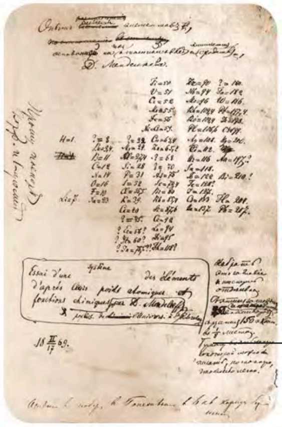

Mendeleev's periodic table

Mendeleev's notes *from 1869 show his method of arranging elements in columns and rows.*

Mendelevium is named after the Russian chemist Dmitri Mendeleev, who invented the periodic table. Mendelevium is produced in very small amounts by firing parts of helium atoms at einsteinium atoms in a particle accelerator (a machine in which atoms are smashed together).

102
No

Nobelium

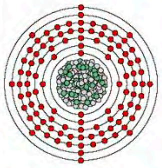

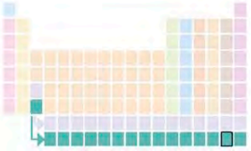

⊖ 102 ⊕ 102 ○ 157

State: Solid
Discovery: 1963

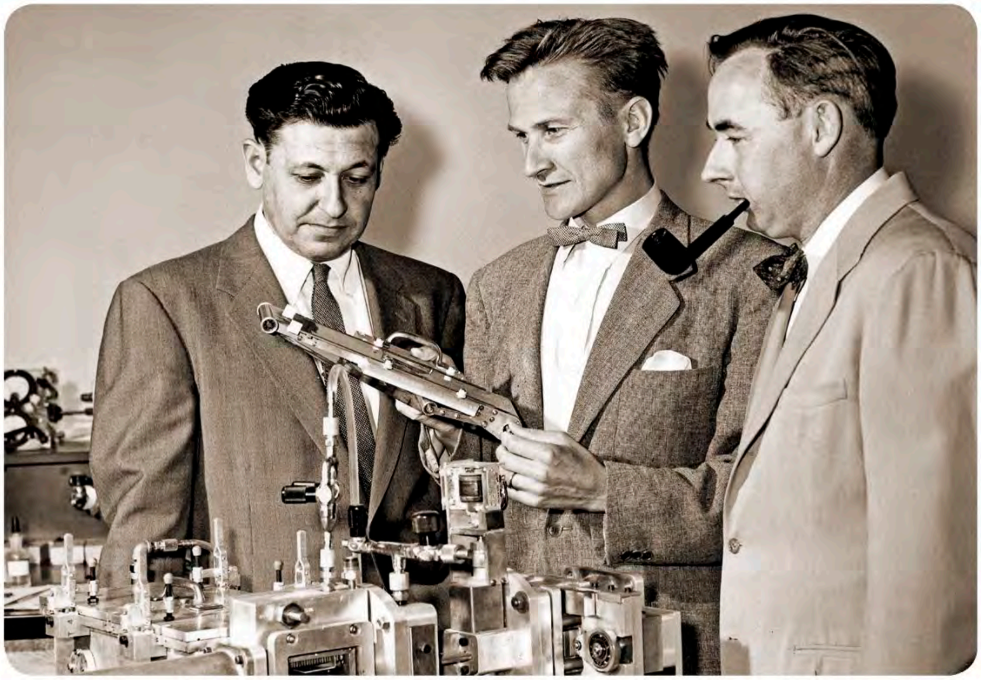

Albert Ghiorso, Torbjørn Sikkeland, and John R Walton

This artificial metal is named after the Swedish chemist Alfred Nobel, who started the Nobel Prize. It was discovered in 1963 by a team of scientists working in California, USA. This team included **Albert Ghiorso**, **Torbjørn Sikkeland**, and **John R Walton**. They used a particle accelerator to fire carbon atoms at curium atoms, creating nobelium atoms, which broke apart within minutes.

103
Lr

Lawrencium

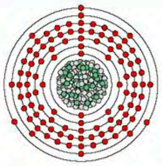

⊖ 103 ⊕ 103 ○ 163

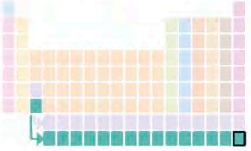

State: Solid
Discovery: 1965

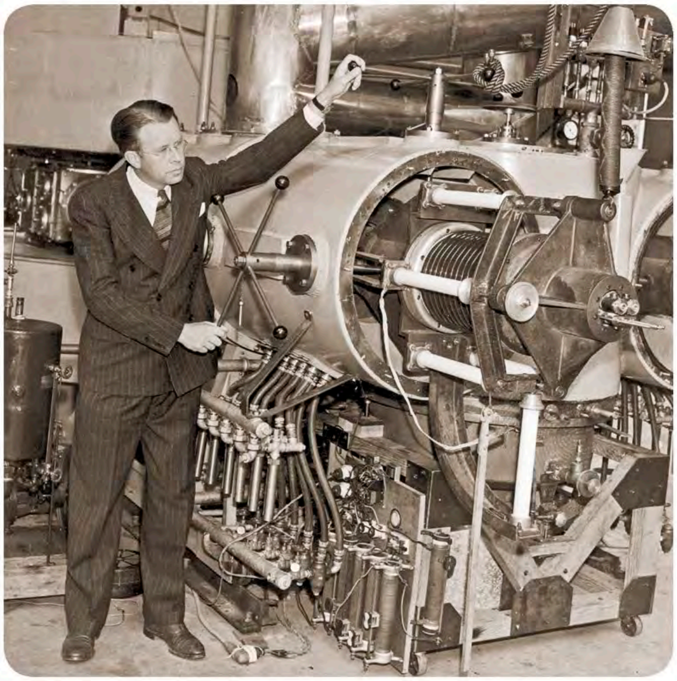

An early cyclotron

Lawrencium was produced at the **Berkeley lab** set up by Ernest Lawrence.

Lawrencium is named after the US scientist Ernest Lawrence, who developed the first cyclotron particle accelerator. This is a machine in which parts of atoms are smashed together by making them spin round in circles. Lawrencium atoms were produced in a similar machine by firing boron atoms at californium atoms.

Pure gallium (Ga) becomes liquid at 29°C (84.2°F).

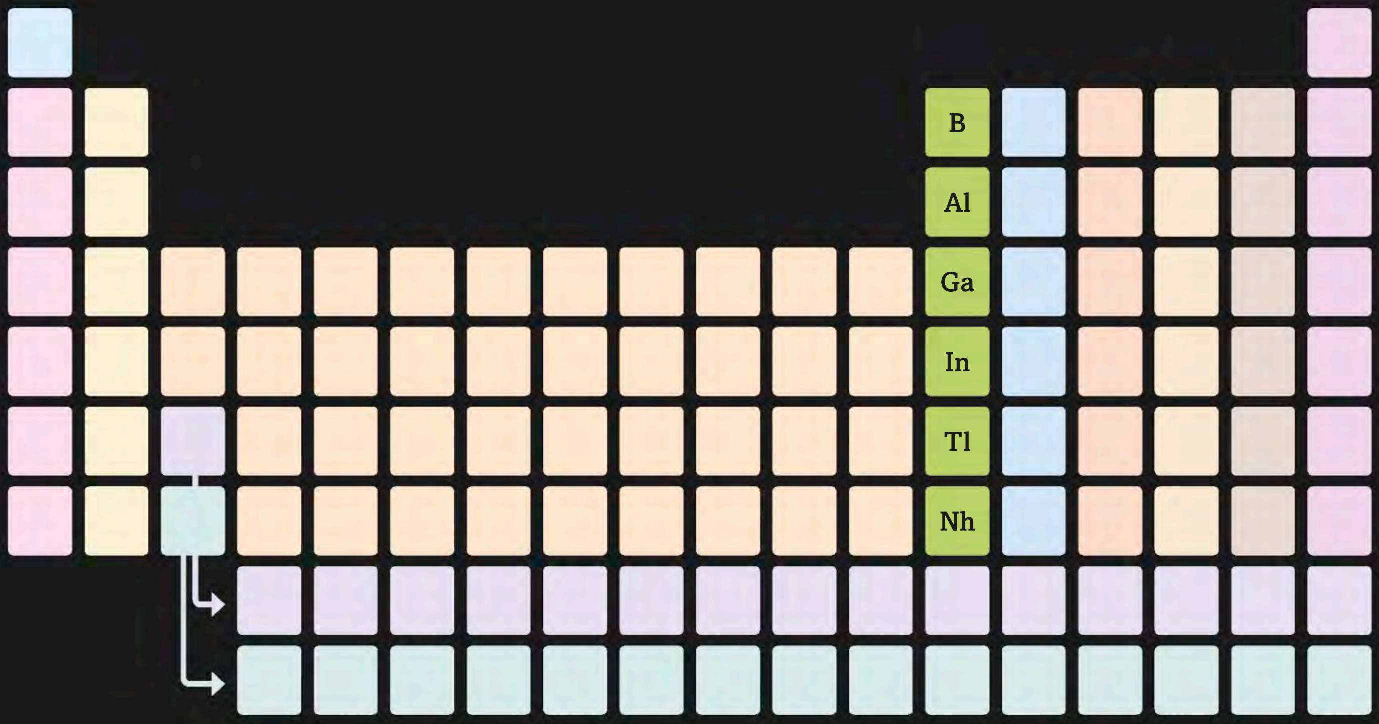

The Boron Group

This group contains five natural elements and one artificial element called nihonium (Nh). Although these elements are not very reactive, none of them are found in a pure form in nature. Boron (B), the first member, is a semi-metal (an element that has properties of both metals and non-metals), while the rest are metals. The second member, aluminium (Al), is the most common metal in Earth's rocks.

Atomic structure
Members of this group have three electrons in the outer shell of every atom. Some elements have unstable isotopes.

Physical properties
All elements, except boron, are shiny solids. Every member of this group is soft, except for boron (B), which is one of the hardest elements.

Chemical properties
Most of the elements don't react with water. Aluminium (Al) forms an oxide layer in water, and can react with it when this layer is corroded.

Compounds
They form compounds by losing electrons to other elements. All of them react with oxygen (O) by bonding to three oxygen atoms.

5
B

Boron

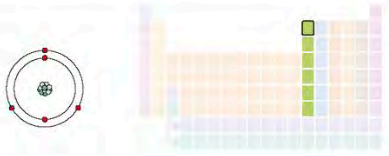

⊖ 5 ⊕ 5 ○ 6

State: Solid
Discovery: 1808

Forms

This translucent mineral *is found in dried lakes.*

This is a colourless form *of a compound of sodium and boron.*

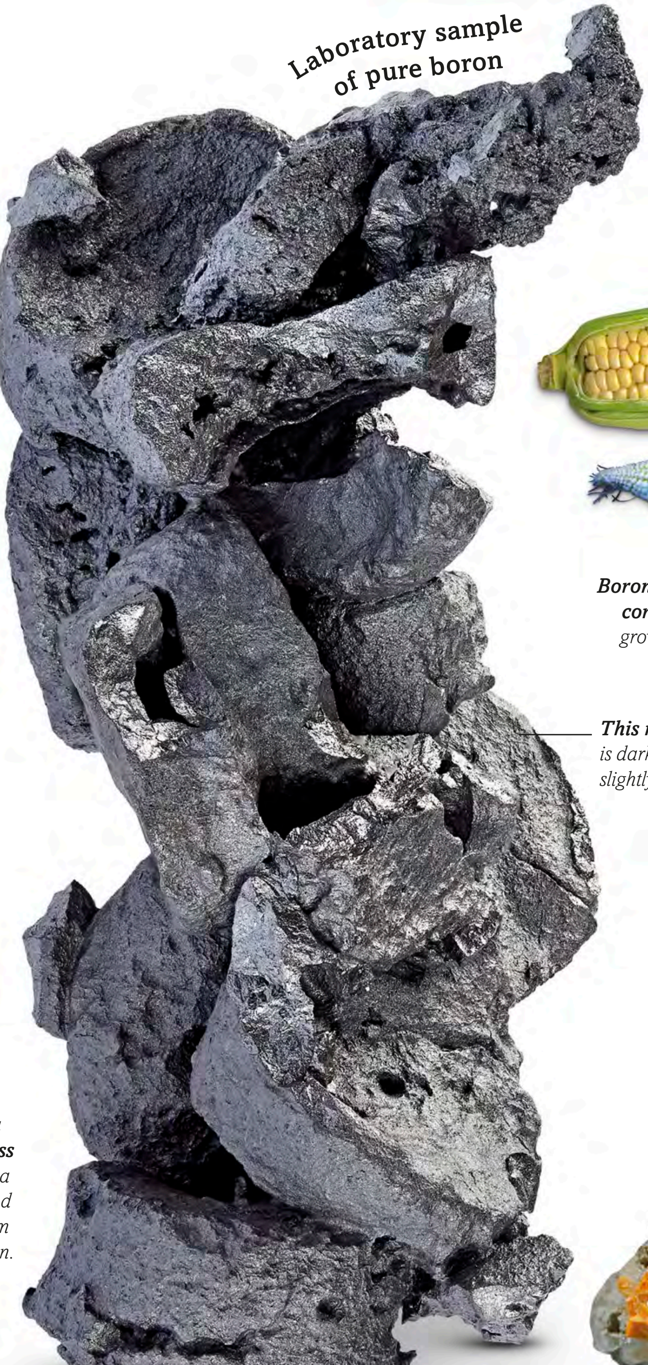

This metal *is dark and slightly shiny.*

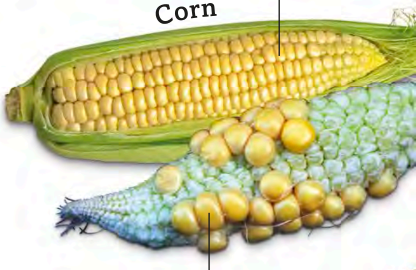

Boron-rich corn

Boron-deficient corn *does not grow properly.*

Colemanite

Some boron compounds are among the toughest artificial substances on Earth, with only diamond being harder. This element is a very hard material and becomes even harder when made to react with carbon or nitrogen. **Pure boron** can be extracted from various minerals, including **ulexite** and **kernite**. The demand for this element was once so high that people moved to live in the extreme heat of **Death Valley, USA**, to work in boron mines there. Compounds of boron in soil are essential for plants to grow healthily. We use boron in our homes every day.

This searing desert *is one of the main places on Earth where boron is found.*

Death Valley, USA

Needle- and leaf-shaped crystals

Uses

2 CUPS
16 OZ.
1 PINT LIQUID
1½
12
1 CUP
8
½
4

Measuring cup

This tough glass *contains boron oxide.*

Boric acid

These white crystals *are obtained from sodium borate.*

LCD screen

This screen *is composed of boron-rich glass, which makes it scratch-resistant.*

Modelling clay

This clay *is bouncy but also firm because it contains boron.*

Boron carbide is one of the **hardest** materials in use today.

IIB
CHURCHILL
IIB

Military tank

The protective body of this tank *contains boron carbide, a compound of boron and carbon.*

THÉNARD AND GAY-LUSSAC

The salt sodium borate, also called borax, was already in use 1,000 years ago. In 1808, the Frenchmen Joseph Louis Gay-Lussac and Louis Jacques Thénard isolated pure boron by heating borax with potassium.

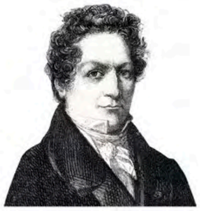

Louis Jacques Thénard
Born into a poor family, Thénard excelled as a scientist. He also discovered a compound called hydrogen peroxide.

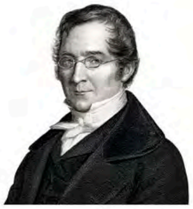

Joseph Louis Gay-Lussac
This French chemist is also remembered for discovering that the pressure of a gas goes up with its temperature.

Tough, heat-resistant glassware, such as **measuring cups**, are strengthened with boron. **Boric acid** is a natural antiseptic and can be used to treat minor cuts and scrapes. A flexible layer of boron-based glass fibres is used to toughen thin **LCD screens** for televisions and laptops. Even some kinds of **modelling clay** and bouncy silly putty contain boron compounds. Boron is named after a crumbly white salt called borax, which is used in detergents. The element is also present in a diverse range of objects, from insecticides to armour for **military tanks**.

13 **Al** Aluminium

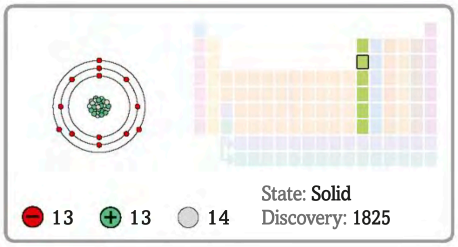

State: Solid
Discovery: 1825

Forms

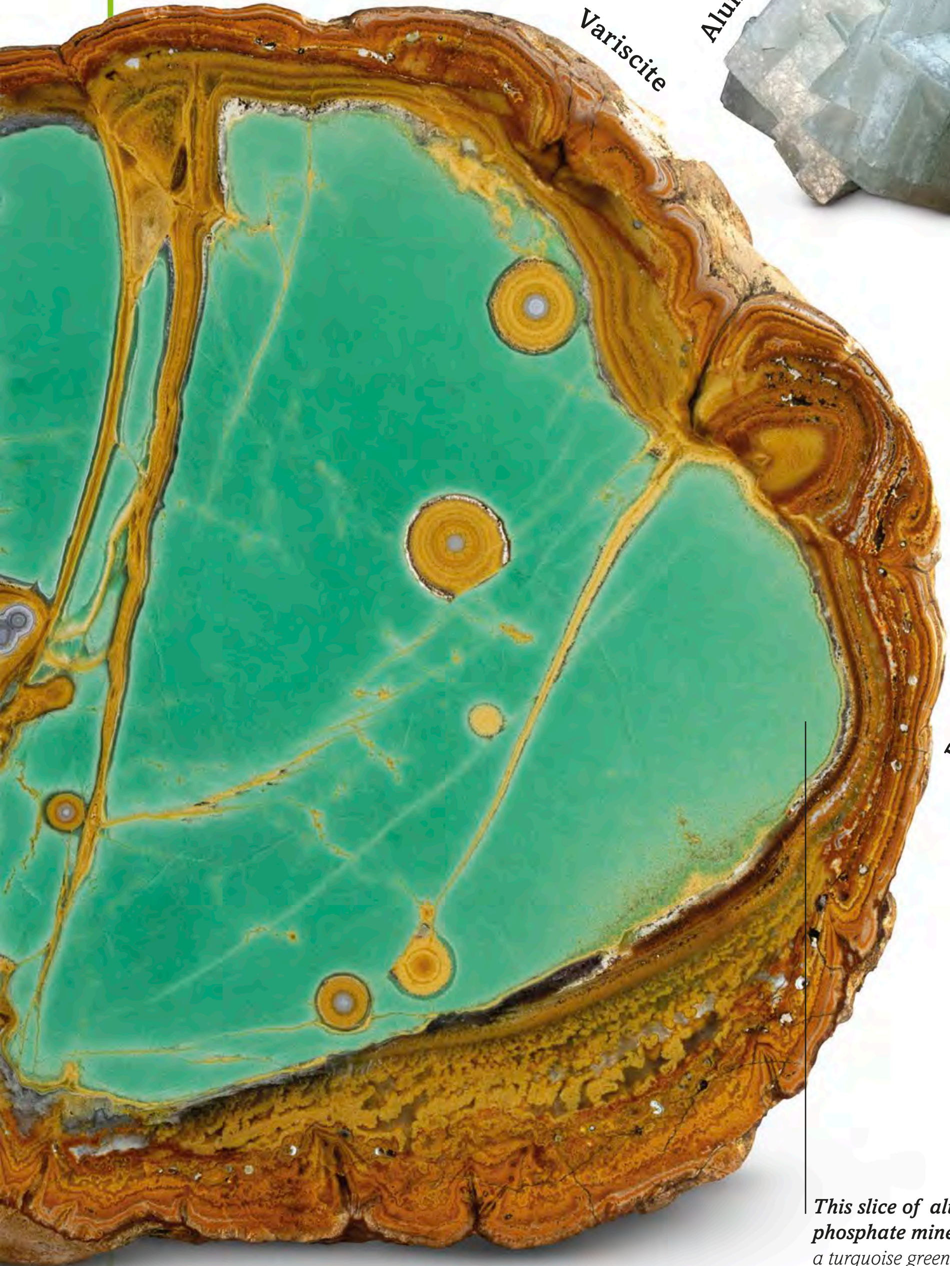
Variscite

This slice of ***aluminium phosphate mineral*** *has a turquoise green colour.*

Alum crystal

This crystal *contains aluminium and sulfur compounds.*

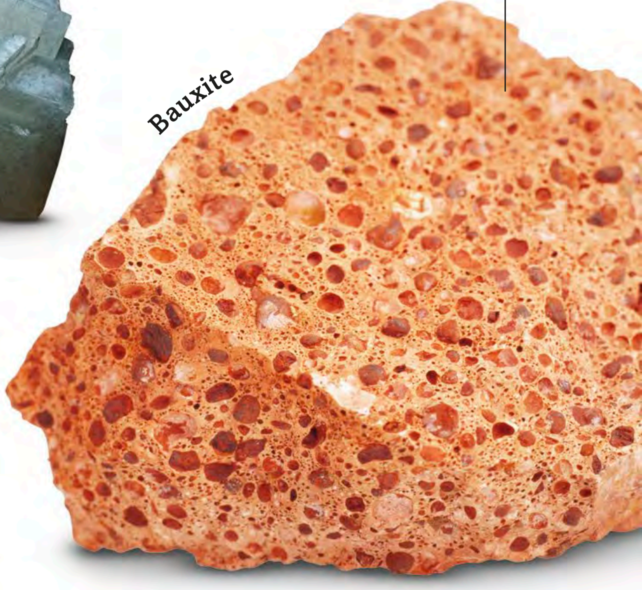
Bauxite

This aluminium ore *is the world's main source of the element.*

Pellets of pure aluminium refined in a laboratory

Reflective surface

Although aluminium is the most common metal in Earth's rocks, scientists did not discover it until the early 1800s. Even then, it took a further 80 years for scientists to work out how to use the ore **bauxite** to extract large amounts of **pure aluminium**. It can also be found in other minerals, including **variscite**. Today, aluminium is often recycled because producing it anew requires 15 times more energy. The metal makes a strong, shiny **foil** when rolled flat, and is useful for storing foods. A **fire protection suit** made from

Uses

Tennis racket

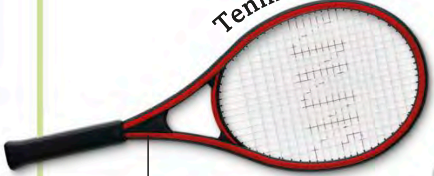

This aluminium frame *makes the racket light.*

Aluminium foil

This foil *does not break even as it is bent and twisted.*

Fire protection suit

This suit *protects against temperatures up to 1,000°C (1,800°F).*

Aluminium can

This can *is produced from recycled aluminium.*

RECYCLING ALUMINIUM

Aluminium is expensive to purify, so it is often recycled instead. Drinks cans are almost 100 per cent pure aluminium and can be shredded, melted down, and made into new cans.

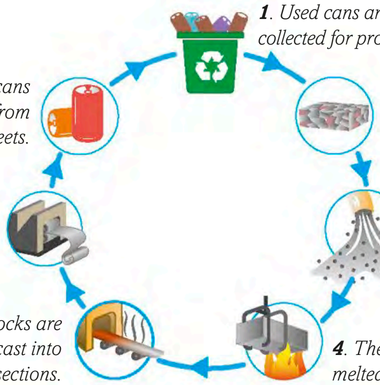

1. Used cans are collected for processing.

2. They are crushed into small bricks.

3. The blocks are shredded into tiny pieces.

4. The pieces are then melted into large blocks.

5. The blocks are then cast into smaller sections.

6. These are pressed into metal sheets.

7. New cans are made from these sheets.

Recycling one aluminium can saves enough energy to run a TV for three hours.

Smartwatch

The aluminium casing *protects the touchscreen.*

Esplanade Theatre, Singapore

Parts of this dome *are made of aluminium.*

Overhead cables

Aluminium cables *are lightweight.*

Boeing 737

KLM
KLM
PH-BXY
Royal Dutch Airlines
KLM

This plane's fuselage *is built from sheets of aluminium stretched around a frame.*

this foil reflects away heat. Aluminium is the most widely used metal after iron. It is very lightweight compared to iron's alloy steel and almost as strong. A dome made from aluminium, such as the one in the **Esplanade Theatre** in Singapore, can be much larger than a steel-based one, which would collapse under its own weight. Aluminium is also a good electrical conductor and so is used in **overhead cables**. Tough aluminium alloys are used to produce parts of some aircraft, including the **Boeing 737**.

JET TURBINE The curved blades of this jet engine are shaped very precisely to catch the air, and they are also strong enough to stay stiff when working at high temperatures. There are several tough metals that fit these requirements, but most are very dense, making them too heavy for an engine powering an aeroplane into the air. That leaves only one metal for the job: aluminium.

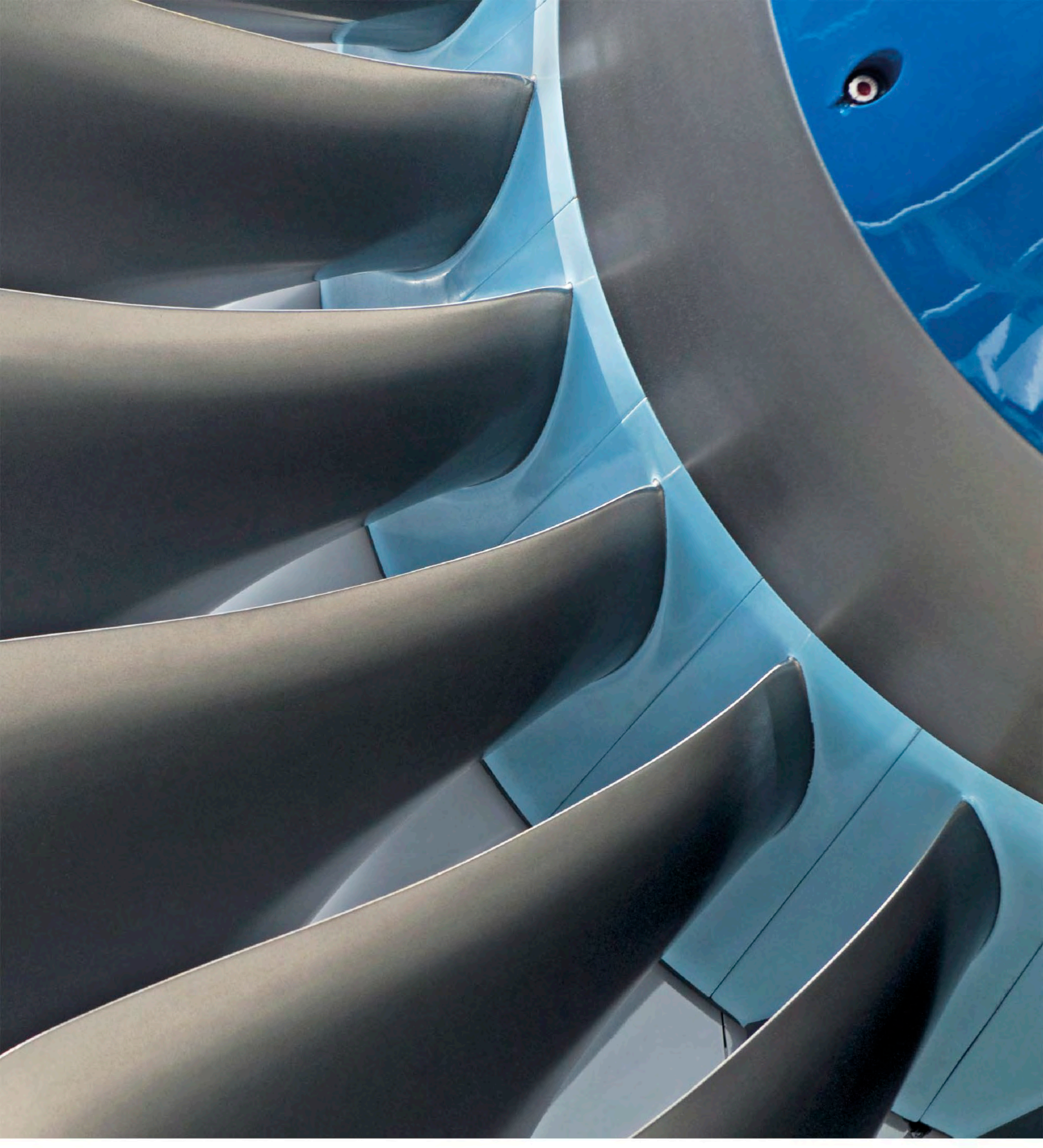

Aluminium is what makes high-speed, long-range air travel possible. Easily moulded, it is one-quarter the weight of steel, and it never rusts. Steel is stronger, but a plane made from it would be too heavy to fly. Instead aluminium is mixed with titanium and steel to produce tough yet lightweight alloys, which are used in the engines and bodies of jet aircraft. There is almost twice as much aluminium in Earth's rock as there is iron. However, purifying aluminium takes a lot of energy. Once pure, though, it can be recycled over and over again. So, one day these engine blades might transform into a fizzy drink can.

31 Ga Gallium

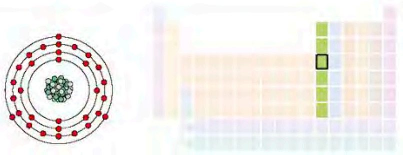

⊖ 31 ⊕ 31 ○ 39

State: Solid
Discovery: 1875

Forms

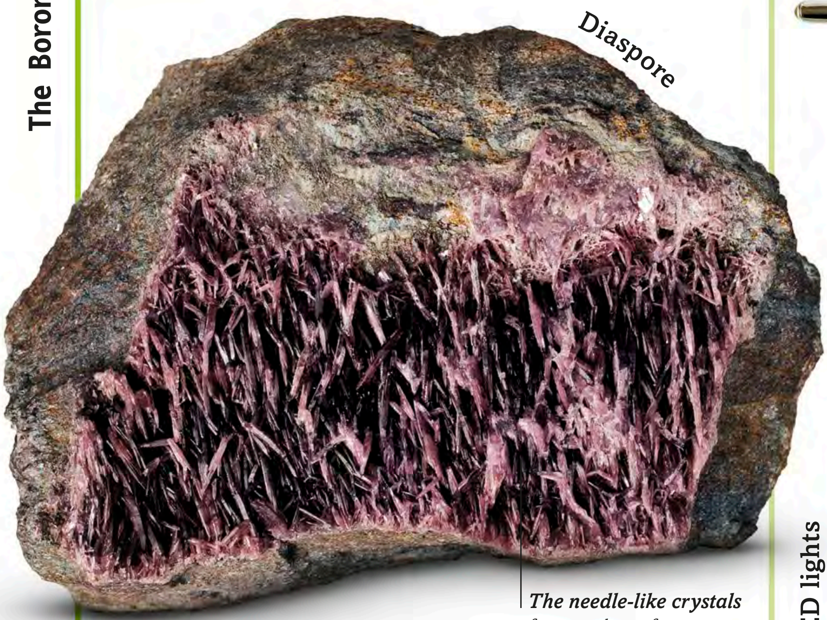

Diaspore

The needle-like crystals *form on the surface.*

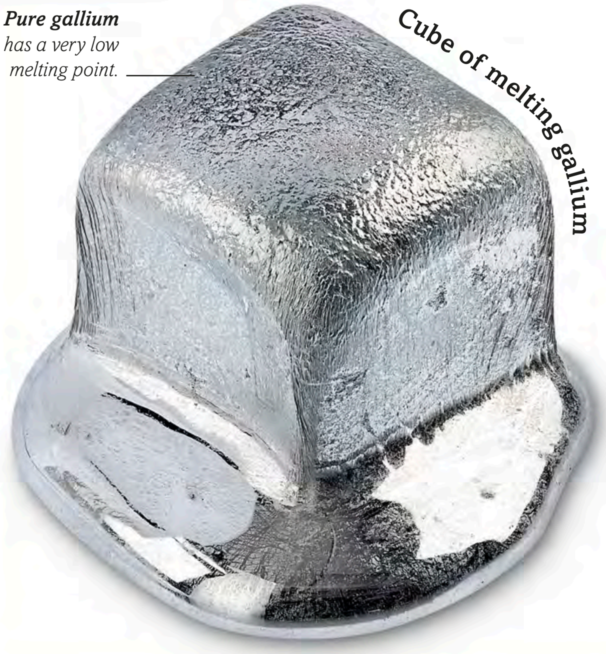

Cube of melting gallium

Pure gallium *has a very low melting point.*

Uses

36
38
40
42
35
37
39
41
°C

Thermometer

This medical thermometer *uses a gallium alloy instead of mercury.*

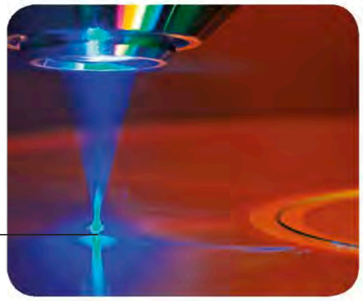

Blu-ray disc

A gallium laser *is used to read Blu-ray discs.*

Red LED lights

The red LED *gets its colour from gallium compounds.*

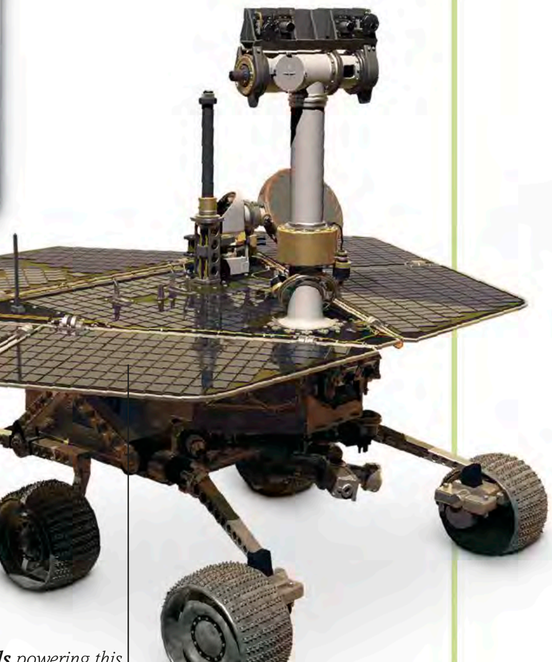

Opportunity rover

The solar panels *powering this rover, which is exploring Mars, contain gallium and arsenic.*

Gallium melts at just 29°C (84.2°F), which means it soon becomes liquid when held in the hand. This element is found in small amounts in ores of zinc and aluminium, such as **diaspore**. **Pure gallium** is isolated when the other elements from this ore are extracted.

Gallium has a number of uses. It is mixed with indium and tin to form a liquid alloy called galinstan, which can be used in **thermometers**. Gallium is also found in **Blu-ray lasers**, **LEDs**, and some solar panels, such as those on NASA's Mars **rovers**.

49
In

Indium

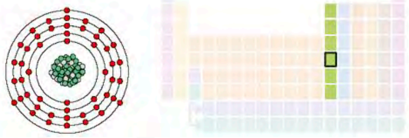

49 49 66

State: Solid
Discovery: 1863

Forms

When bent, indium produces a **"tin cry"** – a sound similar to a **scream**.

Sphalerite

This zinc mineral *is the main source of indium.*

Pure indium mould cast in a laboratory

Pure indium *is soft enough to draw lines on paper.*

Uses

This touchscreen *has a grid of very thin, transparent wires made of indium tin oxide.*

Touchscreen tablet

Welder's goggles

Protective goggles *coated with indium stop the heat damaging the welder's eyes.*

Transistor

The tiny electronic switches *inside this transistor contain indium.*

This glass coated with indium oxide *is shiny but still lets light through.*

Windows in a building

Indium is named after indigo, which is the colour of the light its atoms release when electrified. Its minerals are rare, and most of the metal is obtained from lead and zinc ores, such as **sphalerite**. **Pure indium** is very soft, and the element is mostly used in compound form. For example, a compound called indium tin oxide used on a **touchscreen** allows the computer to detect when a finger makes contact with the screen. Indium is also required in microchips, and to produce **welder's goggles** and **windows** that are heat- and glare-proof.

81
Tl
Thallium

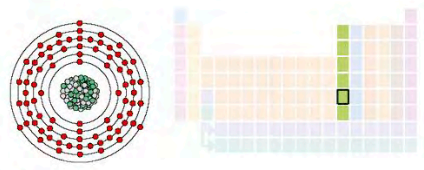

● 81 ⊕ 81 ○ 123
State: Solid
Discovery: 1861

Forms

This mineral contains iron, sulfur, and tiny amounts of thallium.

Pyrite

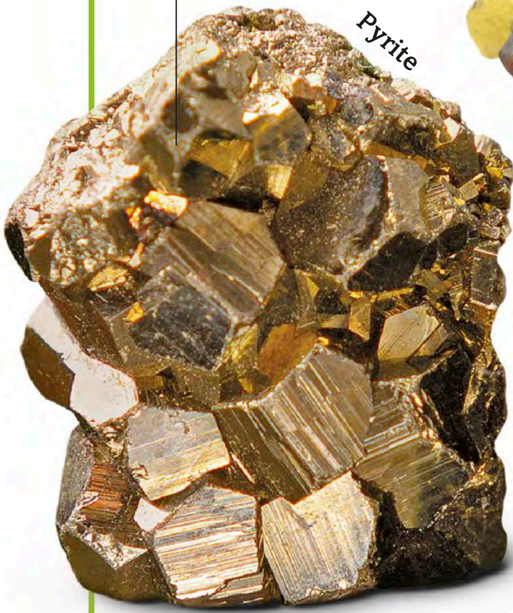

The grey crust on this aluminium mineral contains small amounts of thallium.

Thallium alum

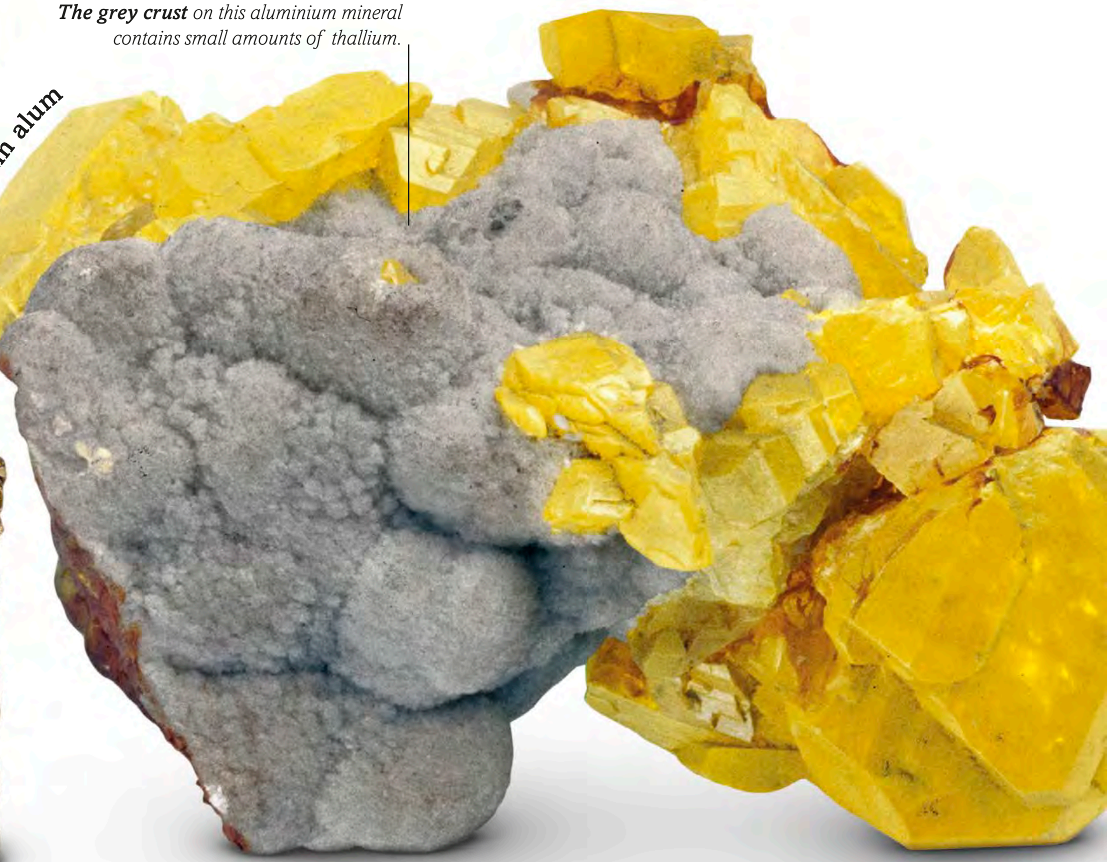

Laboratory sample of pure thallium in an airless vial

This soft and silvery metal is kept in a sealed glass tube as it is highly toxic and reacts easily with air.

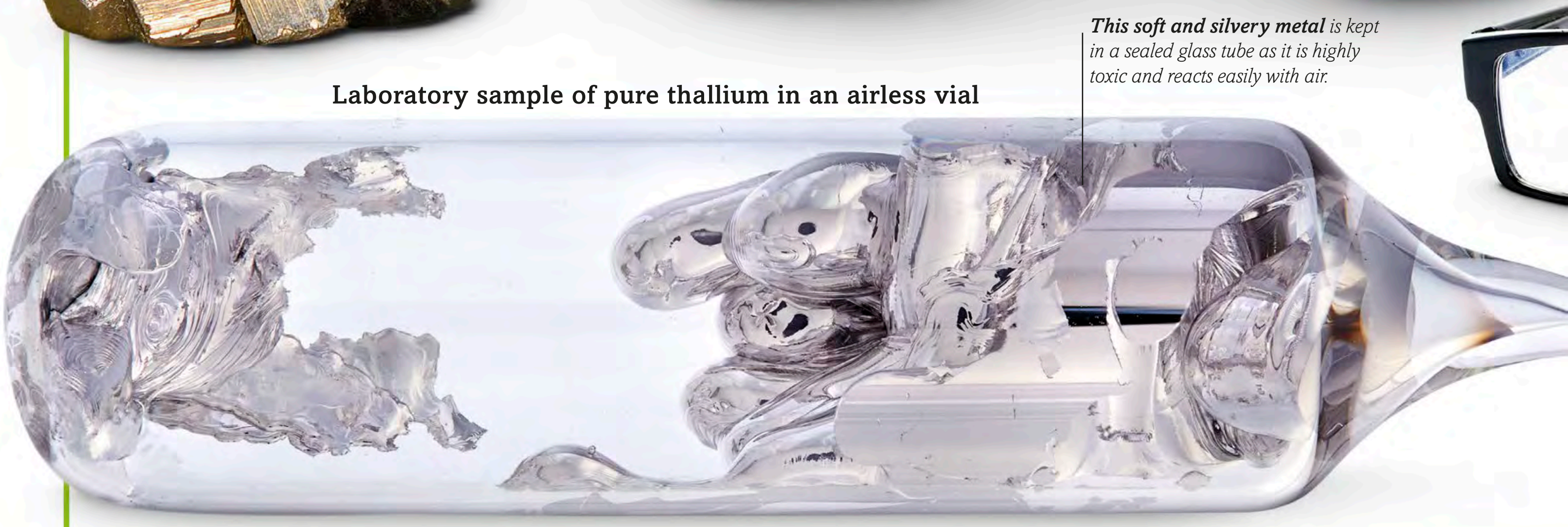

Thallium was named after the Greek word *thallos*, which means "green shoot": it was first identified from the colours in its flame, which includes a bright green light. Thallium was discovered in 1861 by William Crookes and Claude-Auguste Lamy. Although both chemists worked separately, they found the element in the same way – as a residue while making strong acids using the mineral **pyrite**. Thallium was later found to exist in larger amounts in other minerals, including **thallium alum**. **Pure thallium** is toxic and has to be handled

Uses

Heart function scan

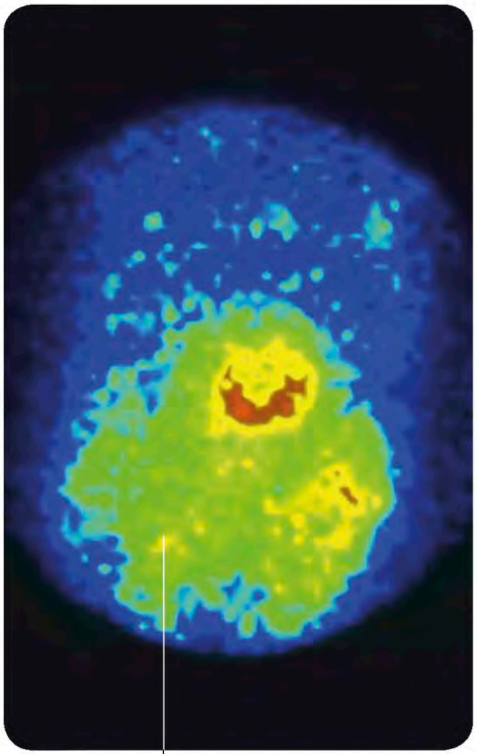

Blood injected with a thallium compound *shows up on a patient's heart scan.*

Spectacles

These thin lenses *contain strong, thallium-infused glass.*

Until the 1970s, thallium salts were commonly used as **ant poison**.

toxic and has to be handled with care when used. A chlorine compound of thallium is used in **scans** to study a patient's blood circulation. Thallium oxide also helps make glass stronger for use in **spectacles** and cameras.

113 Nh Nihonium

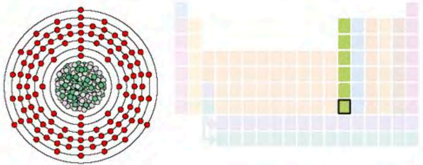

🔴 113 🟢 113 ○ 183

State: Solid
Discovery: 2004

Kozuka Morita (left), with a visiting official at the RIKEN Nuclear Research Centre, Wako, Japan

Nihonium was named after the Japanese word *nihon*, which means Japan. A metallic element, nihonium was first detected in 2003 by teams studying the artificial element moscovium, which has the atomic number of 115. They noticed that atoms of moscovium broke apart after only a few seconds into atoms of an element with an atomic number of 113. In 2004, Kozuka Morita and a team of scientists at the **RIKEN Nuclear Research Centre** in Japan isolated this element in a different way: they fused bismuth and zinc atoms together.

Glassy carbon (C) has a high resistance to rust.

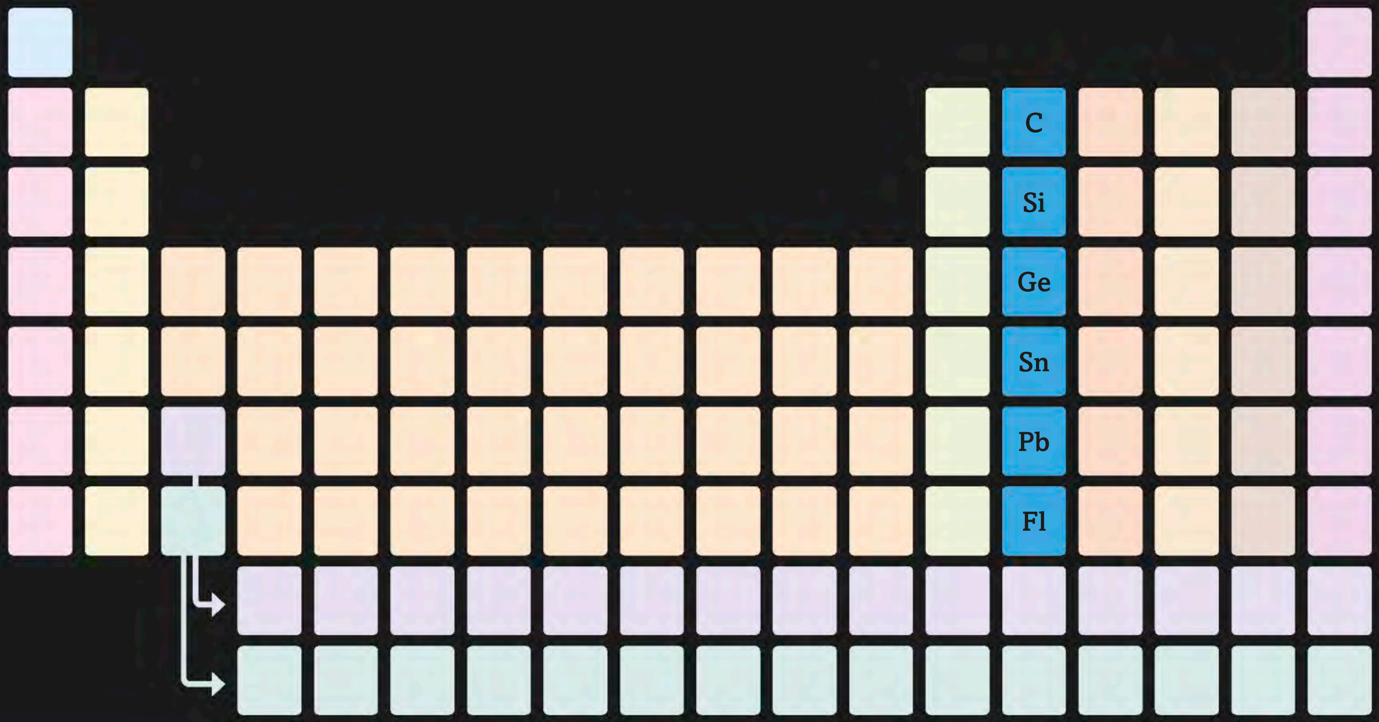

The Carbon Group

This group contains one non-metal, two semi-metals, and three metals. The non-metal carbon (C) is the main element in all living things. The semi-metals – silicon (Si) and germanium (Ge) – are elements that have the properties of both metals and non-metals, and they are essential in electronics. Two of the metals – tin (Sn) and lead (Pb) – have been used by humans for centuries. Flerovium (Fl), an artificial element, has no known uses.

Atomic structure
Members of this group have four electrons in the outer shell of each atom. These atoms can bond with up to four other atoms.

Physical properties
At room temperature, all natural elements in this set are solid. Flerovium (Fl) is an artificial element, and scientists think it may be a solid.

Chemical properties
All natural elements in this group can react with hydrogen (H). Carbon (C) and silicon (Si) can react with both metals and non-metals.

Compounds
These elements react with hydrogen to form compounds called hydrides. Each element can lose up to four electrons when forming compounds.

6 C Carbon

⊖ 6 ⊕ 6 ○ 6

State: Solid
Discovery: Prehistoric

Forms

Coal

***This compressed form of sooty carbon** is formed underground.*

Glassy carbon

Glossy surface

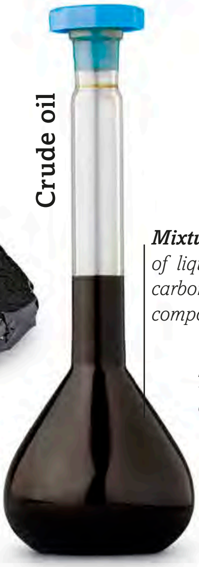

Crude oil

***Mixture** of liquid, carbon-rich compounds*

Raw diamond

***This colourless crystal** forms in magma deep underground.*

Laboratory sample of graphite

***The shiny, metallic surface** feels soft and slippery.*

Cut diamond

***This diamond's brightness** depends on its cut, which determines how many times light entering the gem will reflect inside.*

Carbon has the largest number of compounds of any element – with more than nine million known. Carbon is the fourth most common element in the Universe. Each carbon atom can bond to four others, allowing them to form chains and rings. Pure carbon exists in three forms on Earth – **graphite**, **diamond**, and buckminsterfullerene (a structure based on 60 interlinked carbon atoms). Diamond is the hardest substance in nature. It is often used in jewellery. The blades of some **saws** are coated with diamonds, and

Uses

Diamond blade saw

This sharp blade *is coated with tiny diamonds.*

Water purifier filter

This carbon filter *absorbs harmful chemicals in water.*

Battery

This battery *has a graphite core that carries electric current.*

LOTUS Sport
MAVIC
3G

Carbon-fibre bicycle

Sturdy and lightweight frame

This wheel *is composed of carbon fibre and some other materials, which make it rigid.*

This **strong** but **light** frame is made by fusing together carbon fibres.

Pencil "lead" containing graphite, not lead

Soft graphite *leaves a mark when pressed against paper.*

Kerosene lamp

This flame *is fuelled by carbon-rich oil.*

Polythene bag

This flexible plastic material *is composed of chains of carbon and hydrogen atoms.*

HARD AND SOFT CARBON

Diamond and graphite have different properties because of the arrangement of their atoms – tetrahedrons in diamond, and sheets of hexagons in graphite.

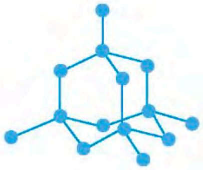

Hard as diamond
The tetrahedron, or pyramid structure, of atoms creates a rigid shape that is equally strong in all directions.

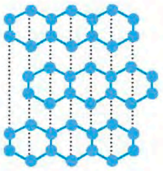

Soft as graphite
Graphite contains layers of carbon atoms that slide over each other easily, as there are only weak forces between them.

can cut into anything. Only a diamond can cut another diamond. Graphite is much softer, which is why it is used in **pencil "lead"**. It is also used in some batteries. Coal is currently the largest source of fuel for the generation of electricity, but its fumes are also known to have harmful environmental and health effects. Crude oil, natural gas, and **coal** are hydrocarbons (compounds containing only hydrogen and carbon) that occur in nature. They can be used as fuels and as plastics for objects such as **polythene bags**.

PINK DIAMOND With a mass of only just over 3 g (0.1 oz), this jewel – known as The Sweet Josephine – is one of the largest pink diamonds ever sold. Diamonds are normally colourless forms of pure carbon, and if there is any colour, it comes from tiny amounts of another substance. For example, boron makes the gem appear blue. Strangely, pink diamonds have no impurities, and no-one knows why they are pink.

The Sweet Josephine was cut from raw diamond that is more than 1.5 billion years old. This formed 150 km (93 miles) beneath Earth's surface and was then pushed up by a volcanic eruption, before eventually being dug out at a mine in Australia. Diamonds form when carbon is squeezed and heated to more than 1,000°C (1,832°F). This process rearranges the carbon atoms into a rigid crystal that makes diamond the world's hardest substance. The process also gives diamond the ability to bend light, a property that gives these jewels their glorious sparkle. With the right cut and polish, a diamond can be made into a beautiful gem that is prized throughout the world.

14 Si Silicon

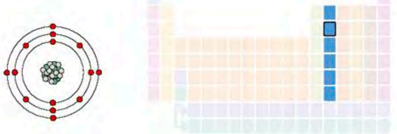

⊖ 14 ⊕ 14 ○ 14

State: Solid
Discovery: 1824

Forms

Fulgurite

This glassy mineral tube *is formed when quartz-rich sand is struck by lightning.*

Laboratory sample of pure silicon

The pure element *can shatter easily.*

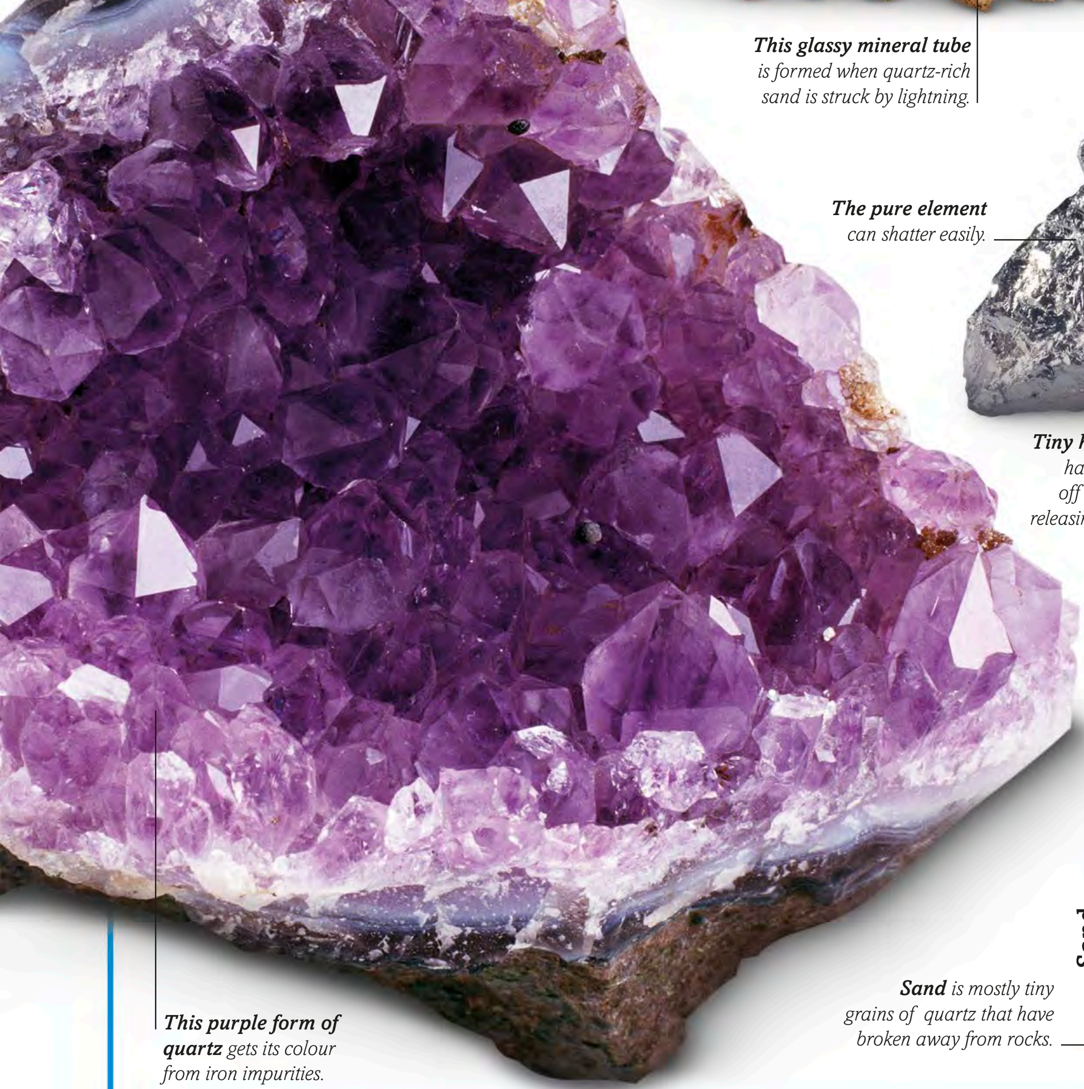

Amethyst

This purple form of quartz *gets its colour from iron impurities.*

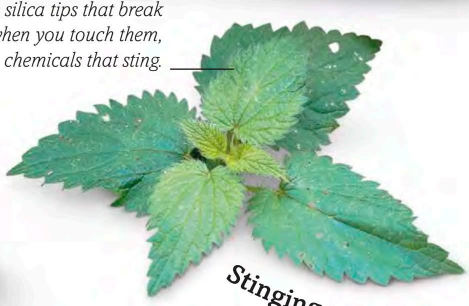

Stinging nettle

Tiny hairs on these leaves *have silica tips that break off when you touch them, releasing chemicals that sting.*

Sand

Sand *is mostly tiny grains of quartz that have broken away from rocks.*

About 90 per cent of the minerals that make up Earth's rocks contain silicon, a common element in our planet's crust. Nearly all silicon minerals are compounds of silicon and oxygen, known as silicates. The most common silicate is quartz, the mineral form of silicon dioxide, or silica. It is also the most common substance in **sand**. **Amethyst** is a type of quartz. Quartz deposits are widely found in rocks such as granite and sandstone. A valuable type of silica is **opal**, which is used as a gemstone. The clays used to make pottery and **ceramics** are also silicates.

Uses

Opal ring

Opal is a jewel *containing both silicon oxide and water molecules.*

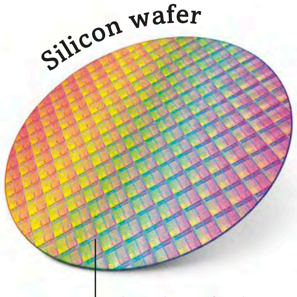
Silicon wafer

Thin slice of silicon

SILICON IN SPACE

Earth and Mars contain similar amounts of silicon. This element forms on the outer layer, or crust, of both planets. In contrast, the Moon is almost half silicon. Astronomers think this tells us that the Moon was formed from Earth's surface after an asteroid smashed into our planet about 4.4 billion years ago.

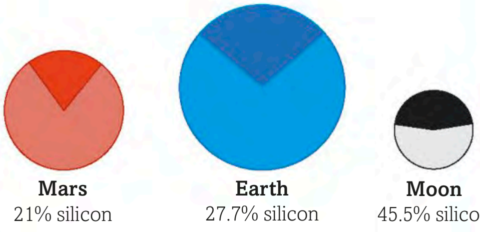

Smokescreen

Screens of silicon-based smoke *can be used in battles.*

Sunlight-powered plane

Solar panels

The engines *run on electricity produced using sunlight.*

Ceramic vase

Ceramic *is made of a silicon-based clay that has been shaped and permanently hardened by heat.*

Silicate aerogel in an experiment

Aerogel *conducts the heat from the flame poorly, preventing it from passing to the flower.*

Silicone baking mould

This silicone tray *is soft, but it does not melt in the oven.*

Silicone watch

This silicone band *is flexible and strong.*

One of the most important uses of silicon is in electronics. Thin slices called **silicon wafers** drive electronic circuits. This versatile element is also used to turn sunlight into electricity in solar panels. Artificial silica is used to create **aerogel**, a lightweight but tough substance that does not conduct heat well. It is used in fire-fighting suits, and prevents flames reaching a firefighter. Another silicon compound is silicone, which can be moulded into any shape, and is used in a wide range of products from **baking moulds** to **watches**.

32
Ge

Germanium

Forms

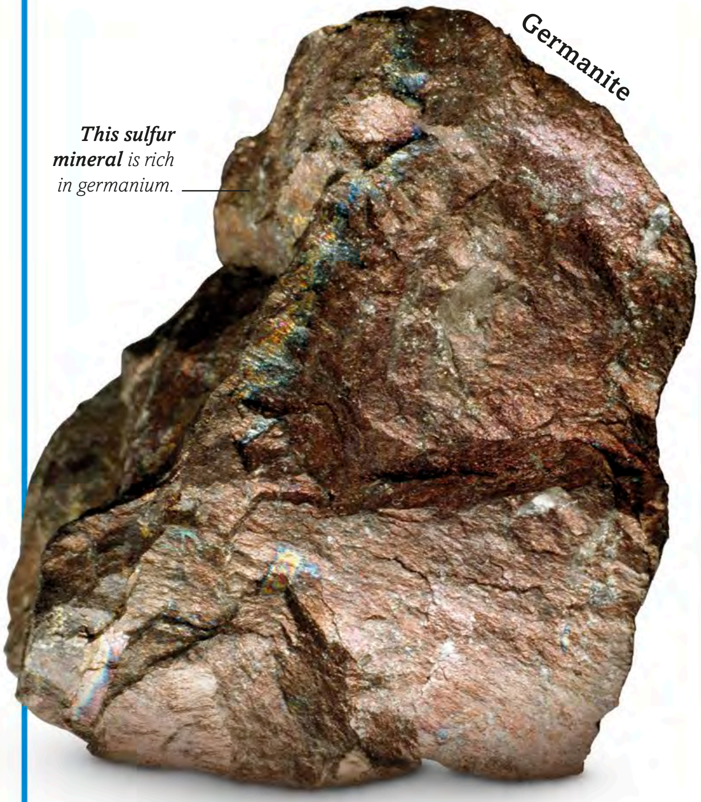

Germanite

This sulfur mineral *is rich in germanium.*

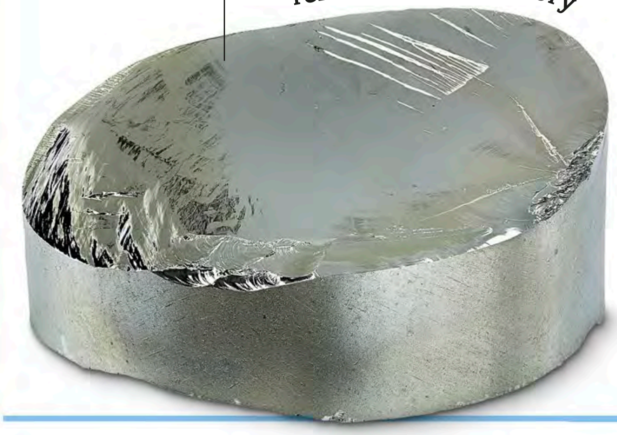

Disc of pure germanium refined in a laboratory

The pure form *is shiny like a metal, but brittle.*

Uses

Camera lens

The germanium oxide in the glass of this lens *bends surrounding light from a large area into the camera.*

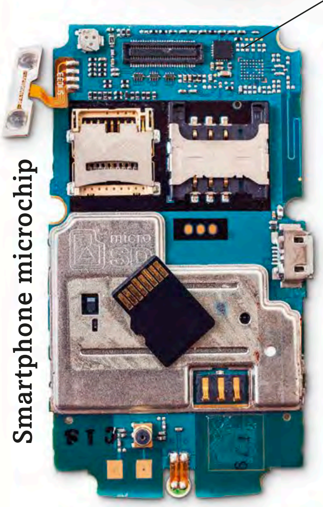

Smartphone microchip

This microchip *is made of silicon and germanium.*

Car with germanium sensor

This car contains a germanium-based sensor *that measures its distance from obstacles.*

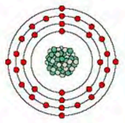

⊖ 32 ⊕ 32 ○ 41

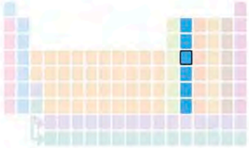

State: Solid
Discovery: 1886

Germanium is found in the **atmosphere** of **Jupiter**.

This semi-metal is named after the country Germany. It was discovered there in 1886 by chemist Clemens A Winkler, nearly 20 years after Russian chemist Dmitri Mendeleev predicted its existence and properties. **Germanite** is a mineral rich in germanium, but this element is mainly extracted from the ores of silver, copper, and lead. One of its compounds, germanium oxide, is used in wide-angle **camera lenses**. It is also used in some **microchips** and in a number of **car sensors** that aid in navigation.

50

Sn

Tin

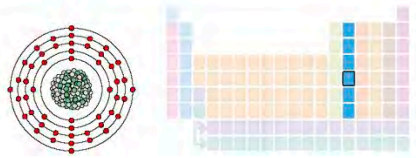

⊖ 50 ⊕ 50 ○ 69

State: **Solid**

Discovery: **c. 2100 BCE**

Forms

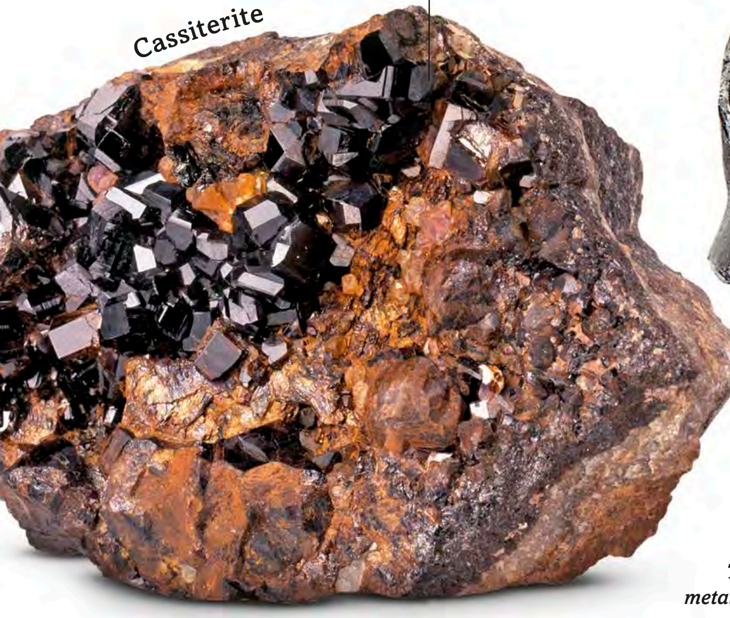

Cassiterite

The black colour *of the crystals is due to iron impurities.*

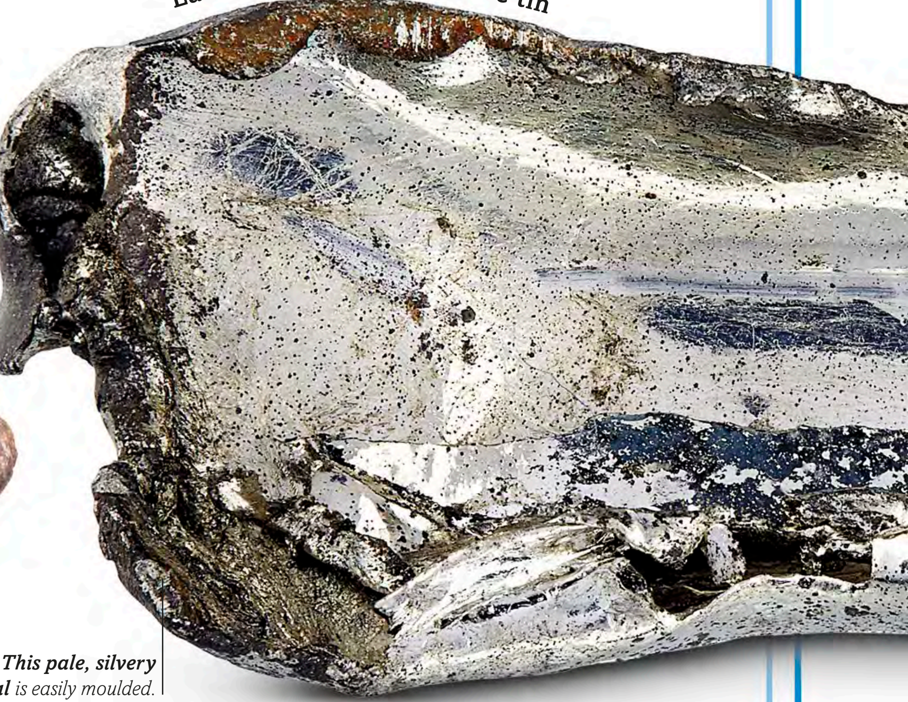

Laboratory sample of pure tin

This pale, silvery metal *is easily moulded.*

Uses

Watering can

Tin plating *over steel makes the can more resistant to corrosion.*

Pewter figurine

This shiny alloy *contains around 90% tin.*

Tin whistle

The tin plating *on this steel whistle prevents rusting.*

Pipe organ

The large pipes *of this organ are made of tin and lead.*

Tin was one of the first metals used by humans. As long as 5,000 years ago, tin was mixed with copper to make bronze, an alloy that was stronger than either pure metal. The ore **cassiterite** is the main source of **pure tin**. Uses for tin are many, including plating steel objects, such as **cans**, to stop them corroding. A compound called tin chloride is used for dyeing silks. This metal continues to be used in a variety of tough alloys, including **pewter**, soft solder, and bronze.

82

Pb

Lead

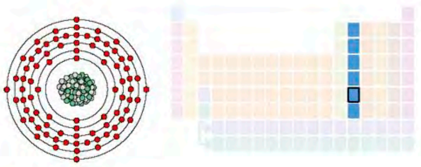

⊖ 82 ⊕ 82 ○ 126

State: Solid
Discovery: Ancient

Forms

Brown crocoite

This soft, crumbly mineral *is made of lead chromate.*

Galena

This mineral *has a bright, silvery shine.*

Anglesite

The prism-like crystals *of this mineral contain a lead-sulfur compound.*

Uses

This crystal glass *sparkles more than regular glass because it contains lead oxide.*

Lead crystal glass mug

Lead pipe

Rust-resistant pipe

For many years, lead and tin were thought to be different forms of **the same metal**.

The chemical symbol for lead, Pb, comes from the Latin word *plumbum*. This is where the word “plumber” comes from: in ancient Roman times, water pipes used in plumbing were made from this soft metal. Lead compounds are found in the minerals **crocoite**, **anglesite**, and **galena** – the main source of **pure lead**. Lead was used far more commonly in the past as an important ingredient in paints, hair-dyes, and insecticides. A common historical use was in **glassware**. It has limited applications today. Lead can

Pure strip of lead refined in a laboratory

This pure form of the metal has a dull grey colour.

These pale crystals on the galena are a calcium mineral.

Flashings, or "lead sheets", cover exposed corners on roofs to make them waterproof.

Flashings

because it absorbs radiation. It is also used in weights for diving, car batteries, and bendy **"flashings"** for sealing roofs. Lead fell out of favour because it turned out to be poisonous.

114
Fl
Flerovium

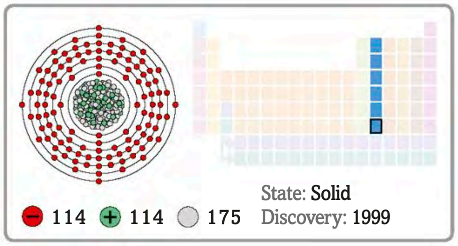

⊖ 114 ⊕ 114 ○ 175

State: Solid
Discovery: 1999

This machine produces flerovium by **smashing together** atoms of calcium and plutonium.

Particle accelerator at Joint Institute for Nuclear Research, Dubna, Russia

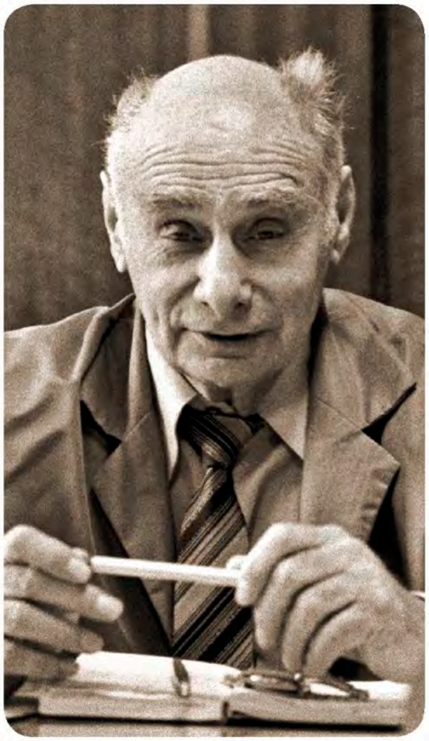

Georgy Flerov

Flerovium takes its name from the Russian scientist Georgy Flerov. He founded the **Joint Institute for Nuclear Research** in Dubna, Russia, where this element was first produced in a particle accelerator (a machine in which atoms are smashed together). Flerovium is highly radioactive and its atoms last for only a few seconds before breaking apart.

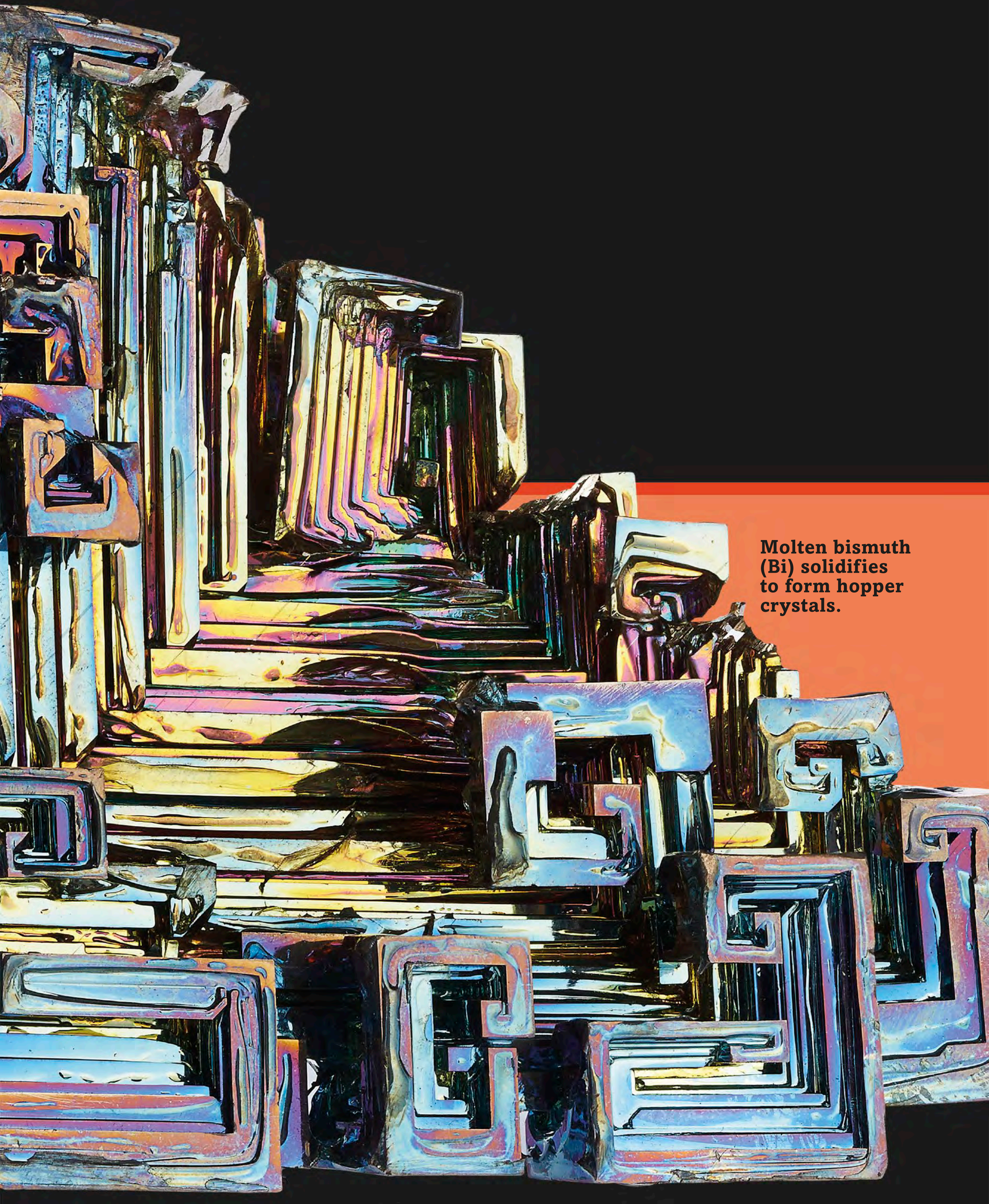

Molten bismuth (Bi) solidifies to form hopper crystals.

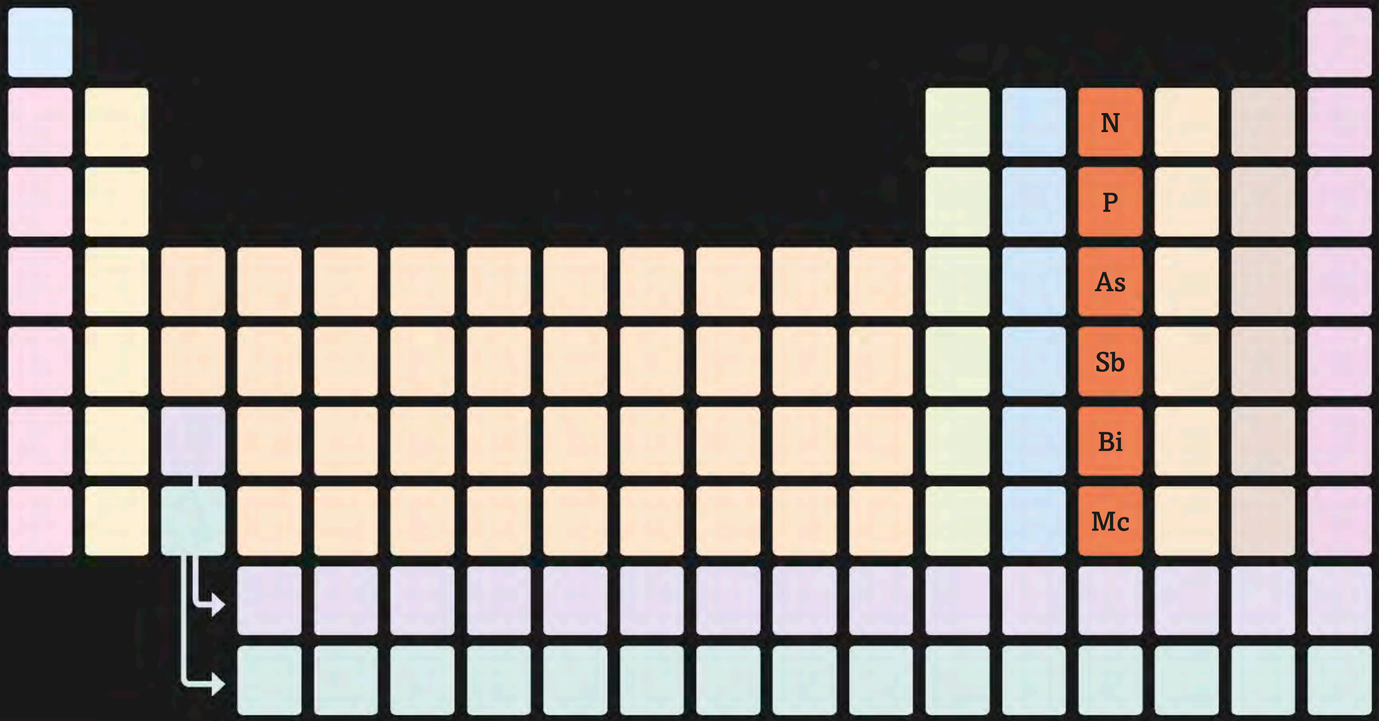

The Nitrogen Group

This group includes different types of natural element – non-metals, semi-metals, and dense metals – as well as moscovium (Mc), an artificial element. The group is also known as “pnictogens”. This derives from the Greek word *pnígein*, which means “to choke” and refers to the potential toxicity of nitrogen (N) in certain forms.

Atomic structure
Members of this group have atoms with five electrons in the outer shell. These atoms can form up to three bonds at the same time.

Physical properties
All the members are solids, except nitrogen (N). The density of elements increases down the group: bismuth (Bi) is 8,000 times denser than nitrogen.

Chemical properties
Phosphorus (P), which exists in two main forms, is reactive, but the others in this group are fairly stable.

Compounds
When reacting with three hydrogen (H) atoms, all members of the group form reactive gaseous compounds called hydrides.

7 N Nitrogen

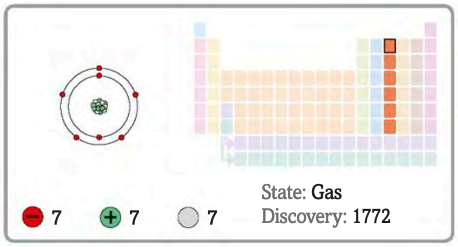

State: Gas
Discovery: 1772

Forms

Pure nitrogen in a glass sphere

Pure nitrogen gas *is trapped inside this glass sphere.*

Nitrogen gas gives off a purple glow when electrified.

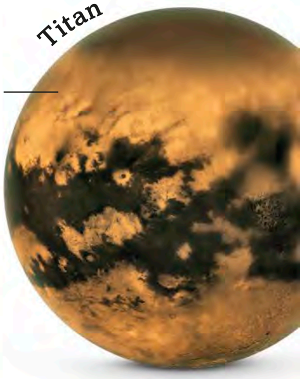

Saturn's largest moon *has an atmosphere containing 48% nitrogen.*

Liquid nitrogen

This clear and colourless liquid *forms when nitrogen is cooled to below –195°C (–319°F).*

This is a naturally *occurring form of sodium nitrate.*

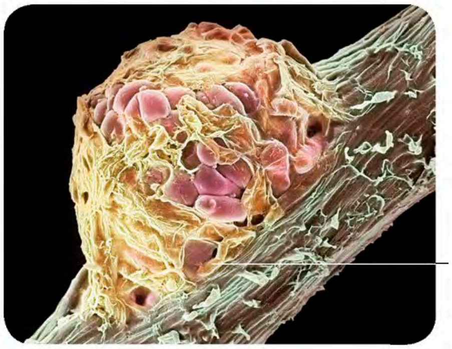

Microscopic image of root nodule

Bacteria living inside plant roots *can take nitrogen from the air for use by the plant.*

Nitrogen surrounds us all the time because it is the transparent gas that makes up nearly three-quarters of Earth's atmosphere. Since **pure nitrogen** does not react easily, its **liquid form** can be used to freeze and preserve items such as blood and tissue samples.

Nitratine is one of the few minerals rich in nitrogen. Some useful nitrogen compounds can be made by industrial processes. A group of nitrogen compounds is used in explosives, including **TNT** and **nitroglycerine**. When ignited, they explode because the bonds

Uses

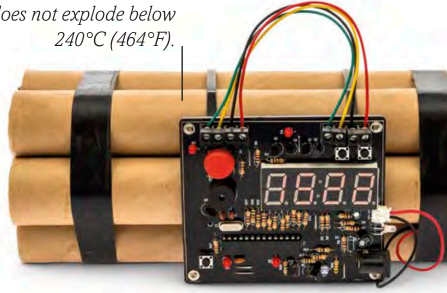

Trinitrotoluene (TNT) *does not explode below 240°C (464°F).*

TNT

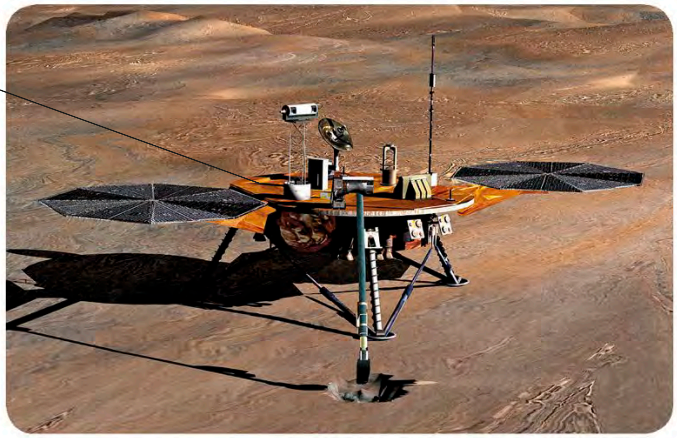

This lander *used 12 hydrazine-powered thrusters to land on Mars.*

Phoenix Mars Lander

Nitroglycerine *can be used to treat heart conditions.*

Nitroglycerine spray

Motorcycles *use nitromethane to fuel their powerful engines.*

Drag bike

These nitrogen-based azo dyes *are often used to colour textiles.*

Textile dye

NITROGEN CYCLE

Nitrogen is essential for life. The nitrogen cycle is a process that constantly recycles nitrogen between Earth's atmopshere and all living things.

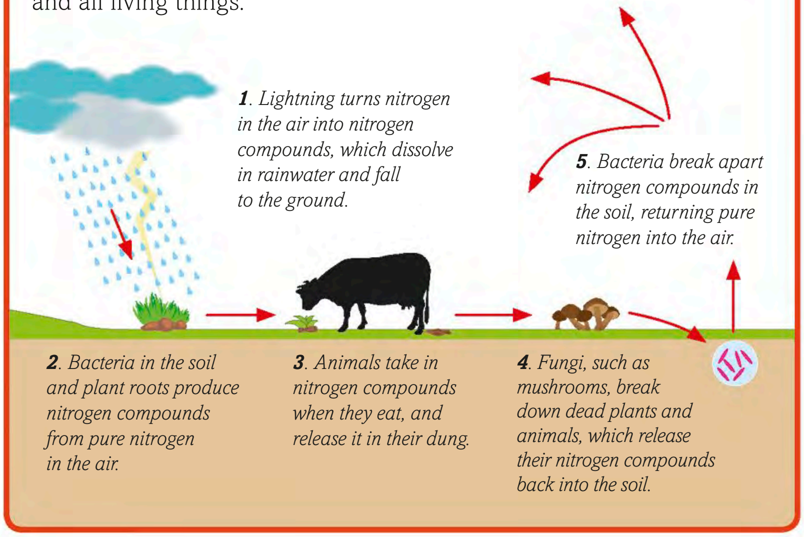

1. *Lightning turns nitrogen in the air into nitrogen compounds, which dissolve in rainwater and fall to the ground.*

2. *Bacteria in the soil and plant roots produce nitrogen compounds from pure nitrogen in the air.*

3. *Animals take in nitrogen compounds when they eat, and release it in their dung.*

4. *Fungi, such as mushrooms, break down dead plants and animals, which release their nitrogen compounds back into the soil.*

5. *Bacteria break apart nitrogen compounds in the soil, returning pure nitrogen into the air.*

Superstrong glues *contain small nitrogen compounds that link together.*

Super glue

These fertilizers *contain compounds called ammonium nitrates that promote growth in plants.*

Nitrogen fertilizer

between nitrogen atoms detach very quickly. Nitrogen fuels, such as nitromethane, are used in **drag bikes**, providing a lot more power than carbon-and-hydrogen only fuels, such as petrol. The compound hydrazine is used in thrusters on spacecraft, such as the **Phoenix Mars Lander**.

Some nitrogen compounds are put in **dyes** and **glues**. An industrial technique called the Haber process turns nitrogen and hydrogen gas into ammonia, a liquid commonly used to make **nitrogen fertilizers**. When mixed with soil, these fertilizers boost plant growth.

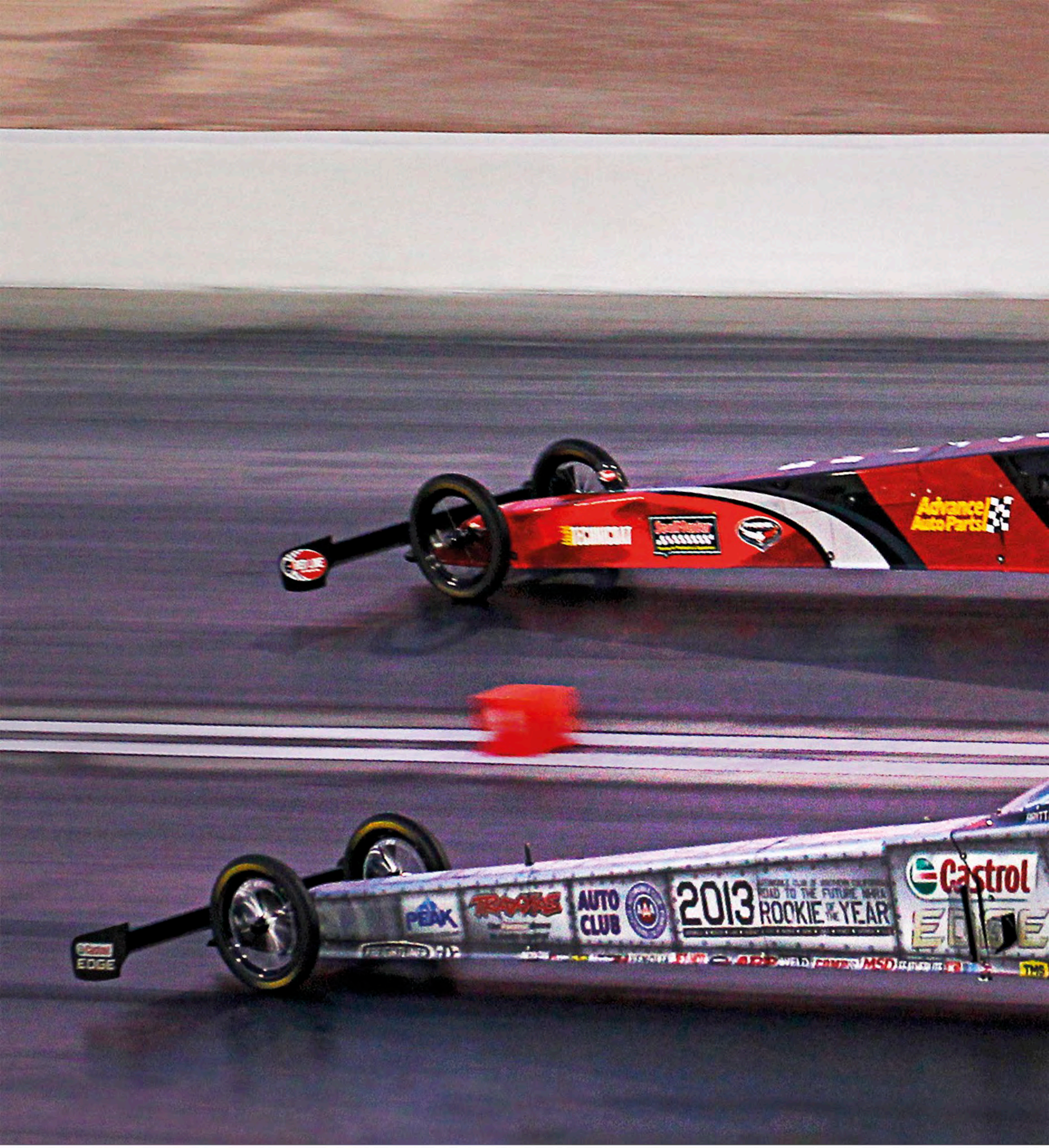

DRAG RACING

Zooming along a dead straight track, these all-powerful dragsters accelerate all the way to the finish line. They contain massive engines filled with an extra-powerful fuel called nitromethane, which is often shortened to “nitro”. Burning eight times faster than regular petrol used in most cars, this super fuel can push dragsters to speeds in excess of 480 km/h (300 mph).

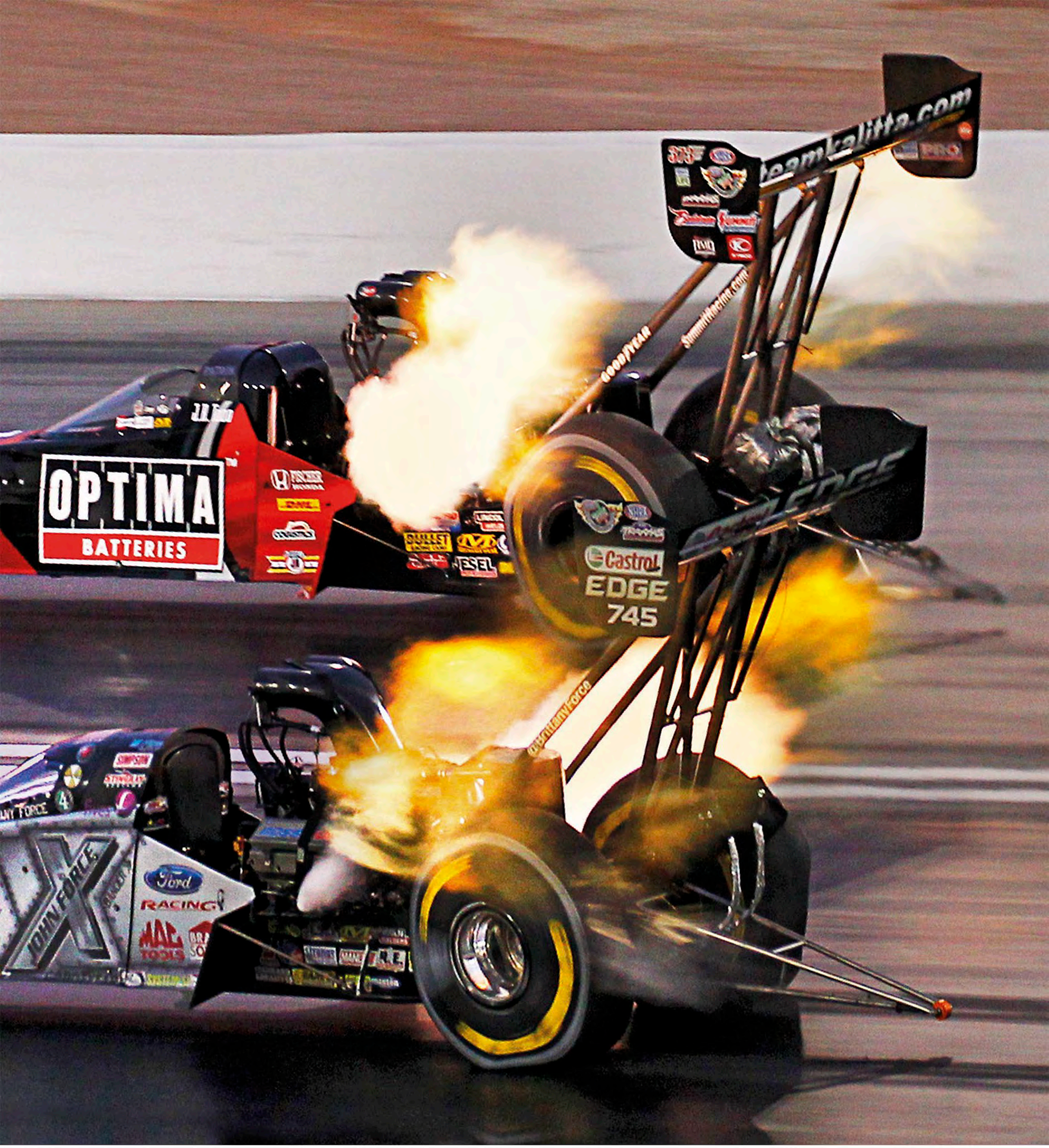

Nitromethane contains carbon, hydrogen, and nitrogen, but it is the latter element that really gives this fuel its immense power. During the process of combustion – when oxygen is mixed with the fuel in the race car's mighty engine – nitromethane burns so violently that nitrogen escapes from the fuel and returns to its pure form. This chain of events leads to a release of energy, which propels the dragsters to breakneck speeds. Although these races are an incredible spectacle to behold, driving using nitromethane can be dangerous due to the explosive nature of nitrogen when used in this way: drag racers are taking a risk to win.

15
P
Phosphorus

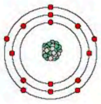

⊖ 15 ⊕ 15 ○ 16

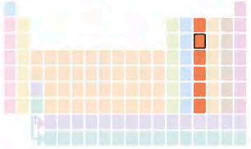

State: Solid
Discovery: 1669

Forms

***Deep phosphate mines** make up 80% of this tiny Pacific island.*

Phosphate mining in Nauru

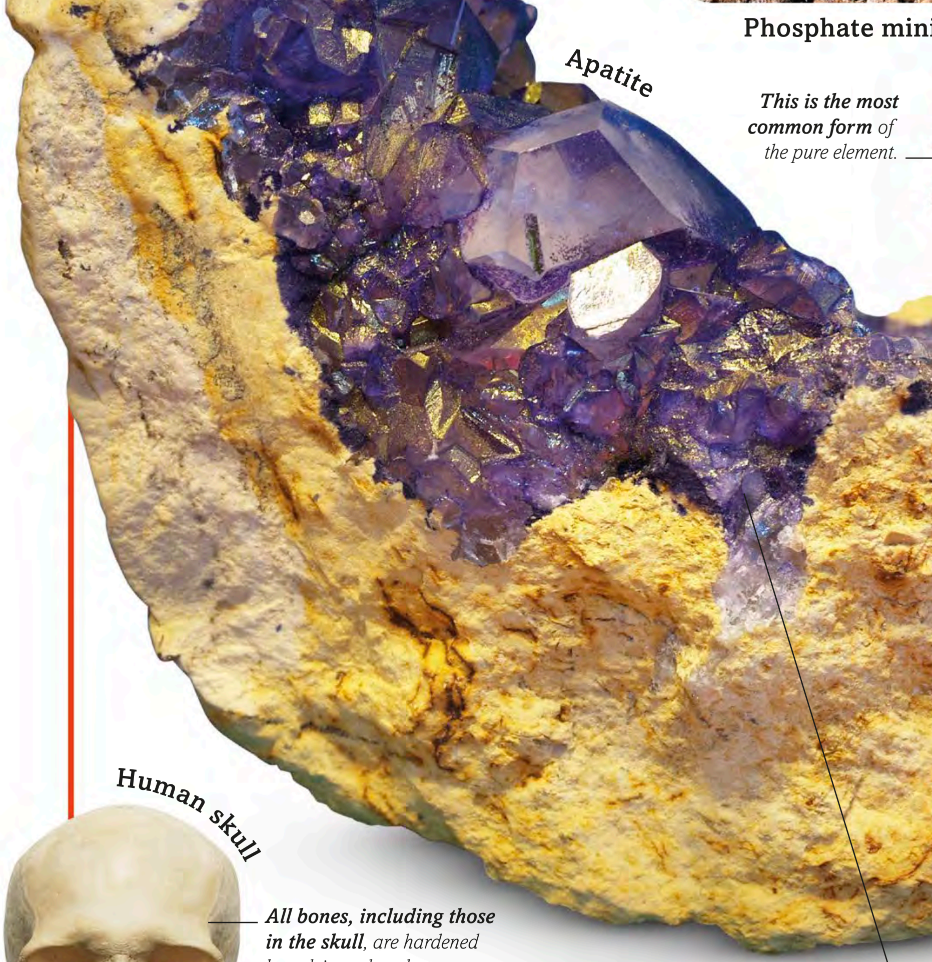

Apatite

***This purple colour** of the crystals is due to metal impurities.*

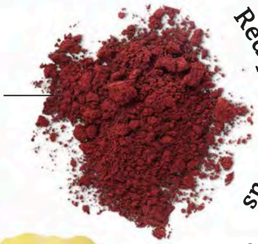

Red phosphorus

***This is the most common form** of the pure element.*

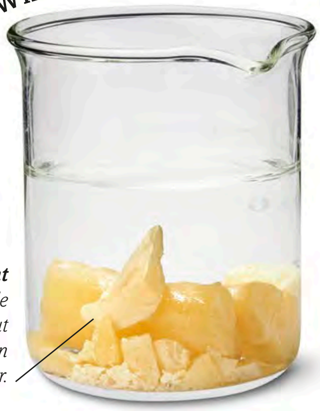

White phosphorus

***This element** remains stable in water, but catches fire on contact with air.*

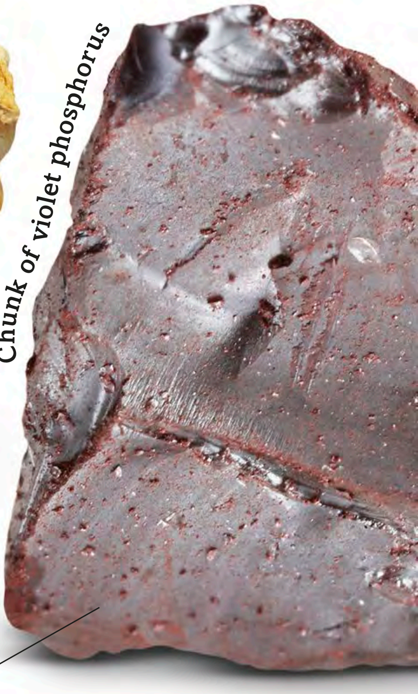

Chunk of violet phosphorus

***This form of phosphorus** is made in a laboratory by heating red phosphorus.*

Human skull

***All bones, including those in the skull**, are hardened by calcium phosphate.*

Tuna

***This fish** is rich in phosphorus.*

Phosphorus was accidently discovered by German alchemist Hennig Brand. In 1669, in his quest for the mythical Philosopher's Stone (a material some believed could turn any metal into gold), he boiled a large pot full of urine for days. This produced a mysterious glowing substance, which he called phosphorus, meaning "giver of light". Phosphorus is the first element to have a discoverer with a recorded name. It is never pure in nature, and occurs in different minerals. Phosphorus has several flammable,

Uses

This light, strong china *contains calcium phosphate.*

China tea set

LIFE'S BUILDING BLOCKS

DNA – short for deoxyribonucleic acid – is like a mini database packed with instructions telling your body how to work properly. It consists of a chain of molecules and looks like a twisted ladder called a double helix. The edges are sugar molecules linked together by units containing phosphorus.

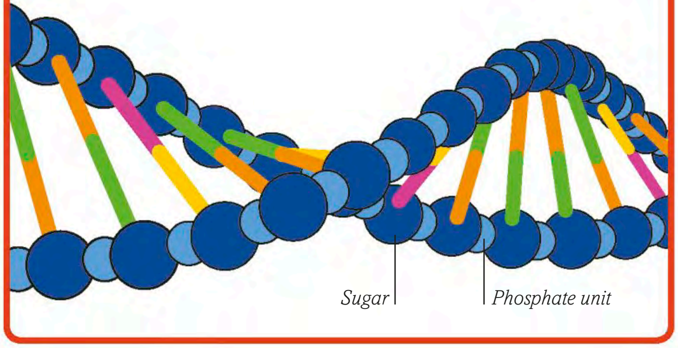

These flexible fibres *are composed of phosphate-rich glass.*

Fibre optics

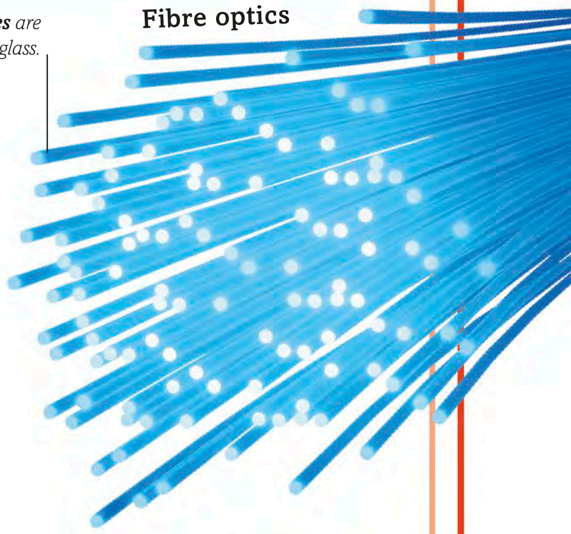

Safety match box

The striking surface on the sides of the box *contains phosphorus, which ignites the match upon contact.*

Spraying ammonium phosphate *over burning material cuts off its supply of oxygen. This extinguishes the fire.*

Fire extinguisher

Spraying crops with phosphates *kills pests, such as insects.*

Pesticides

Fertilizer

This fertilizer *contains ammonium phosphate to boost plant growth.*

solid forms, including **red**, **white**, black, and **violet**. The glow seen by Brand was caused by white phosphorus reacting with oxygen. Phosphorus is mainly found in **phosphate** minerals (in which phosphorus links to oxygen), such as **apatite**, its main ore.

Phosphates are present in **fine china**, and are an important ingredient in **fertilizers**. The strips on the sides of **safety match boxes** contain pure phosphorus. More complex phosphorus compounds used in **pesticides** are poisonous.

33
As

Arsenic

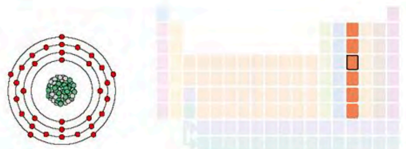

⊖ 33 ⊕ 33 ○ 42

State: Solid
Discovery: c. 1250

Forms

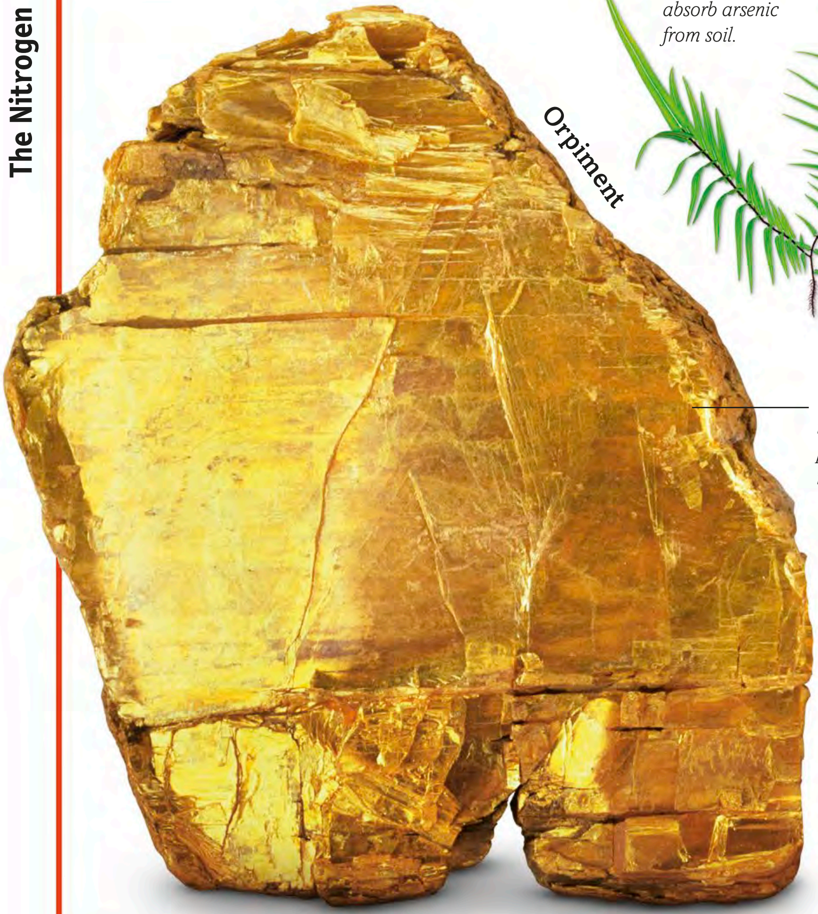

Orpiment

Until the 1800s, *this mineral was powdered for use in paints.*

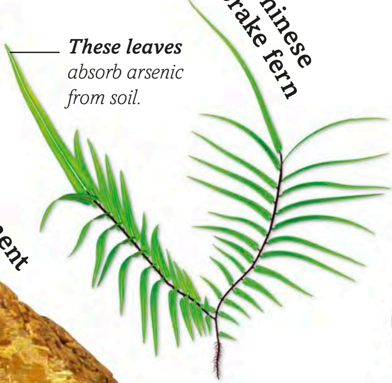

Chinese brake fern

These leaves *absorb arsenic from soil.*

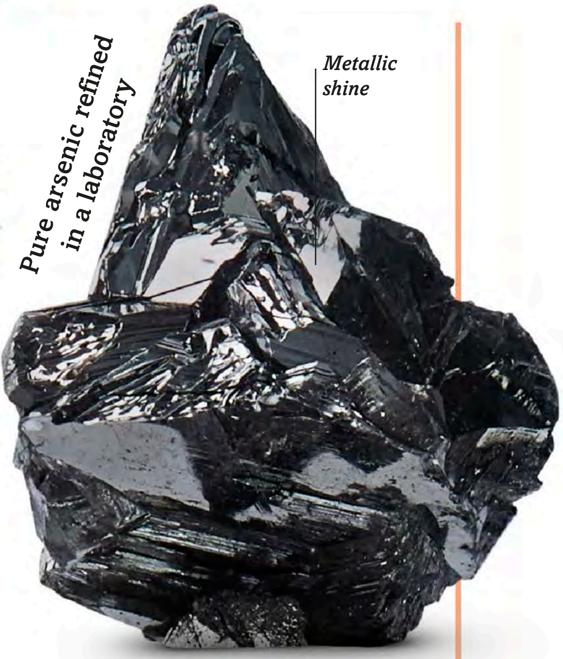

Pure arsenic refined in a laboratory

Metallic shine

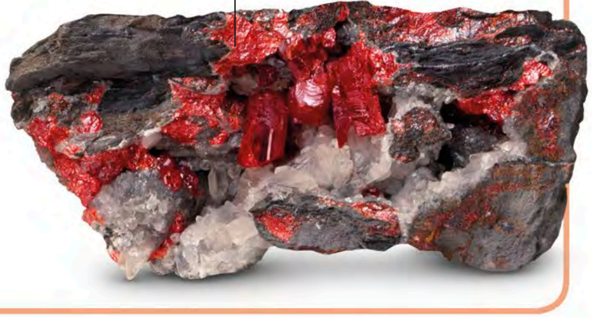

Realgar

This ore *is found in hot volcanic springs.*

Uses

Bullets

This tip *is an alloy of arsenic and lead.*

Rat poison

This poisonous arsenic compound *kills rats.*

When heated, arsenic **doesn't melt**, instead it **turns into a gas**.

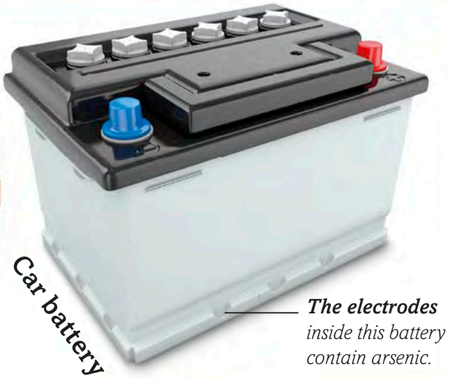

Car battery

The electrodes *inside this battery contain arsenic.*

Arsenic is often called the "king of poisons". Every form of arsenic – either pure or in a compound – is poisonous to animals. In fact, arsenic poisons have been used for centuries. This semi-metal is found in several minerals with striking colours, including **orpiment**. Naturally occurring **pure arsenic** has a shiny, grey colour. Arsenic compounds are used in making some **rat poisons**. The leading use of arsenic today is for strengthening lead. This is done by mixing arsenic with lead to create a tough alloy that is often used in **car batteries**.

51
Sb

Antimony

⊖ 51 ⊕ 51 ○ 71 State: Solid Discovery: c. 1600 BCE

Forms

Jamesonite

These needle-like crystals *contain antimony, lead, and iron.*

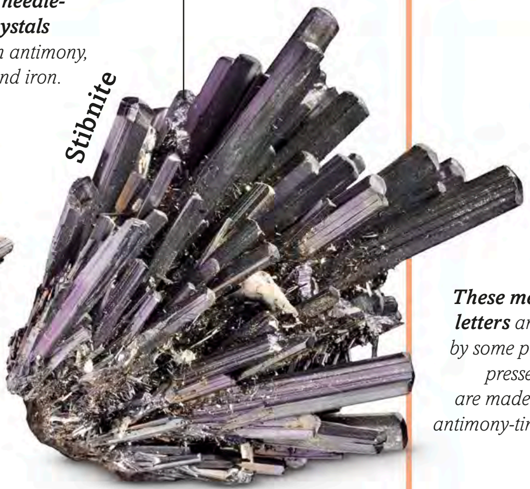

Stibnite

This mineral *tarnishes on contact with air.*

Pure antimony crystals refined in a laboratory

This silvery semi-metal *is hard but brittle.*

Uses

Printing press metal type

These metallic letters *are used by some printing presses, and are made of an antimony-tin alloy.*

Safety matches

These matches *with antimony in the tip burn brighter than the ones without it.*

Egyptian kohl

Kohl *is a dark eyeshadow.*

Kohl makes it **easier to see** in strong sunlight.

Antimony gets its name from the Greek word *anti-monos*, meaning “not alone”. This may refer to the fact that the element is never found pure in nature, but is always found combined with heavier metals, such as lead. The element’s symbol, Sb, comes from *stibium*, the Latin word for kohl, a form of eye make-up. The mineral ore **stibnite** is the largest source of **pure antimony**. Its pure form is mostly used to make hard alloys, such as that in the **metal type** used by some printers. Ancient **Egyptian kohl** was made from powdered stibnite.

83
Bi

Bismuth

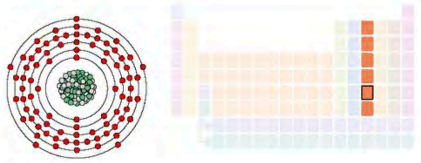

⊖ 83 ⊕ 83 ○ 126

State: Solid
Discovery: c. 1500

Forms

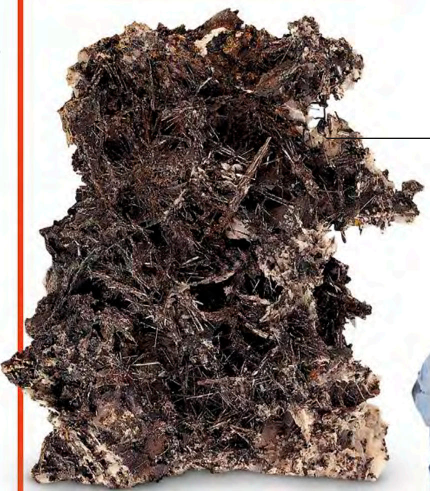

***This mineral** is the main source of pure bismuth.*

Bismuthinite

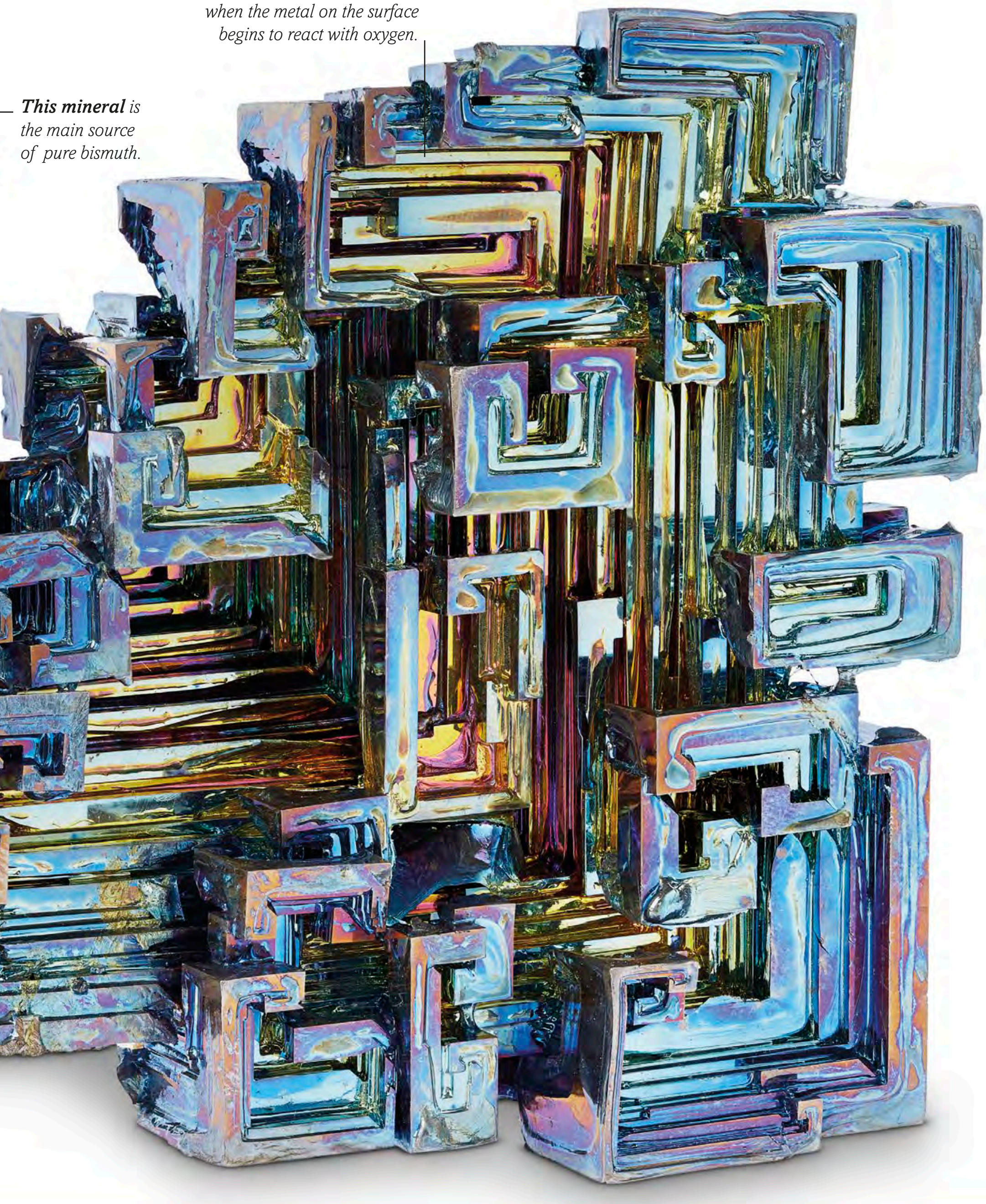

Hopper crystals refined in a laboratory

***These rainbow colours** form when the metal on the surface begins to react with oxygen.*

***These bismuth crystals** were produced in a laboratory.*

In Earth's crust, bismuth is about **twice as abundant as gold**.

Bismuth is a radioactive element but its atoms are relatively stable and last for millions of years. People have known about bismuth for centuries. The Incas of South America added it to weapons made of the alloy bronze to harden them, while ancient Egyptians used a bismuth mineral to make their cosmetics glittery. Pure bismuth forms an oxide in air that is seen as colourful crystals called **hopper crystals**. This element is very brittle and has few uses when not in a compound form. Yellow bismuth

Uses

***This cold box** uses a compound called bismuth telluride, which becomes cold when electrified, and keeps items cool.*

Portable refrigerator

***This nail polish** gives a pearly effect because of bismuth compounds.*

Yellow cosmetics

***This medicine** contains a bismuth compound that helps settle an upset stomach.*

Digestive medicine

Unlike most elements, bismuth's liquid state is **heavier** than its solid.

pigments are used in paints and **cosmetics**, while several bismuth compounds are also in **medicines**. An alloy of bismuth and tin is an ingredient in fire sprinklers.

115

Mc

Moscovium

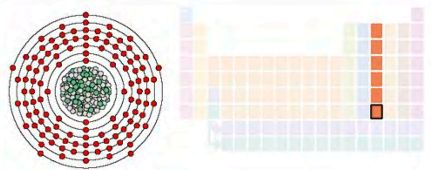

⊖ 115 ⊕ 115 ○ 174

State: **Solid (predicted)**

Discovery: **2004**

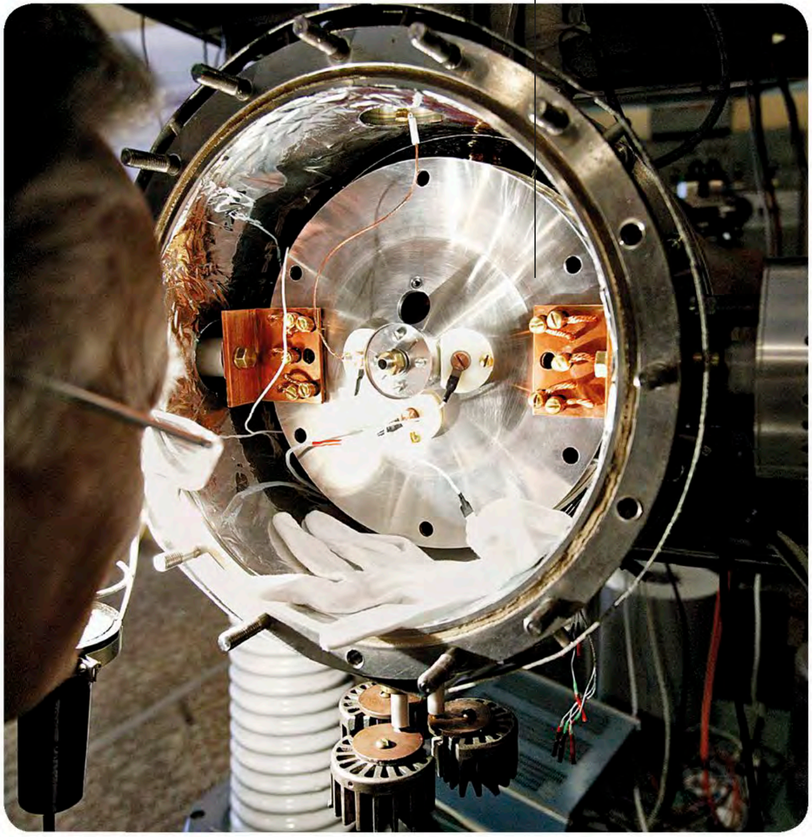

*This is one of the **machines** in this research centre.*

Joint Institute for Nuclear Research, Dubna, Russia

Only about a hundred or so atoms of this heavy, artificial element have been made. Moscovium was first created at the **Joint Institute for Nuclear Research** in Dubna, Russia. A team of Russian scientists, led by Yuri Oganessian, created this element by smashing americium atoms with parts of calcium atoms. It is named after the Russian capital city of Moscow. This element is extremely radioactive, and its atoms break up within a fraction of a second. Scientists think that moscovium would be a dense, metallic solid but with such small samples, they can only measure how big the atoms are before they break up.

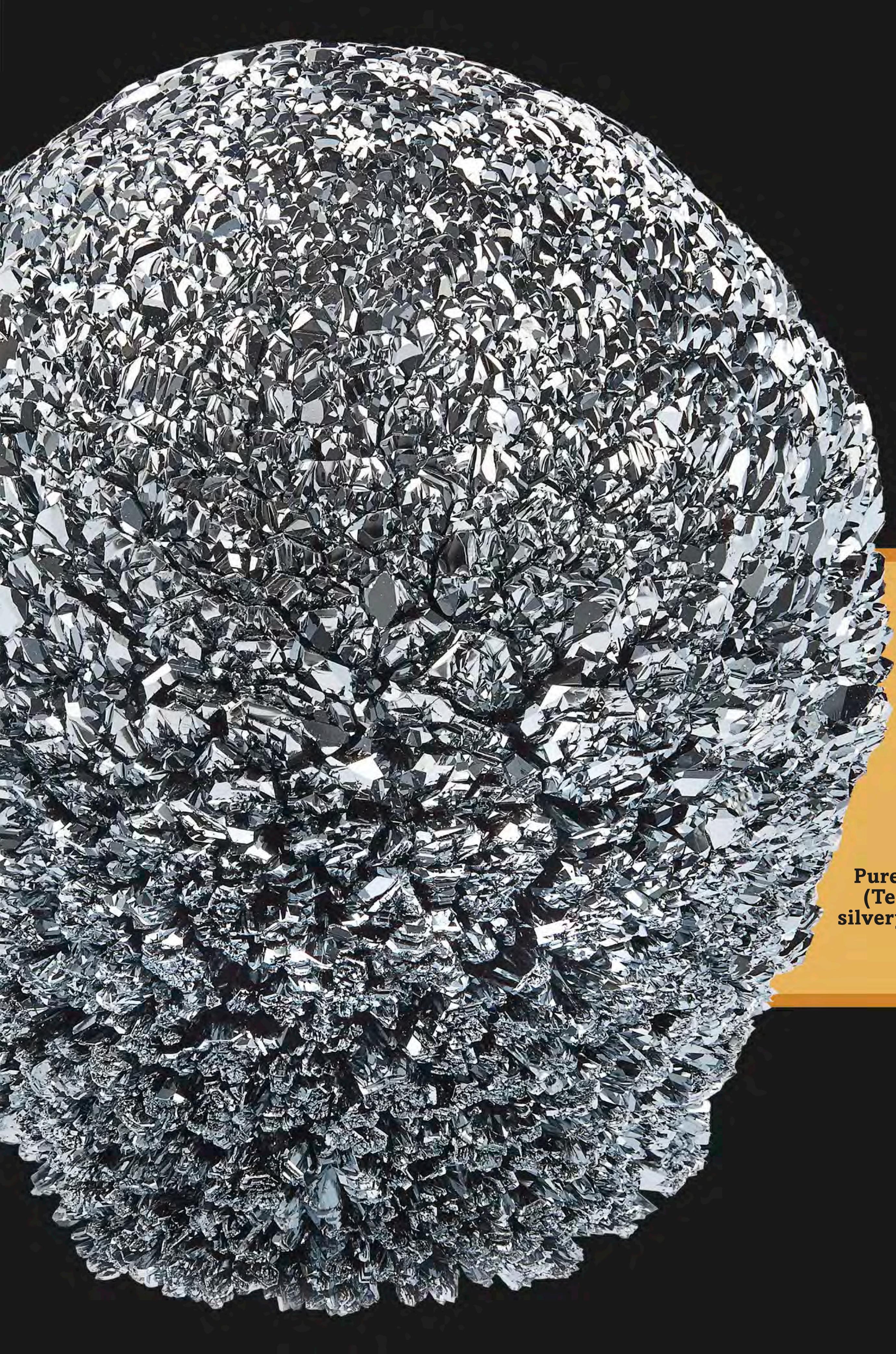

Pure tellurium (Te) can form silvery crystals.

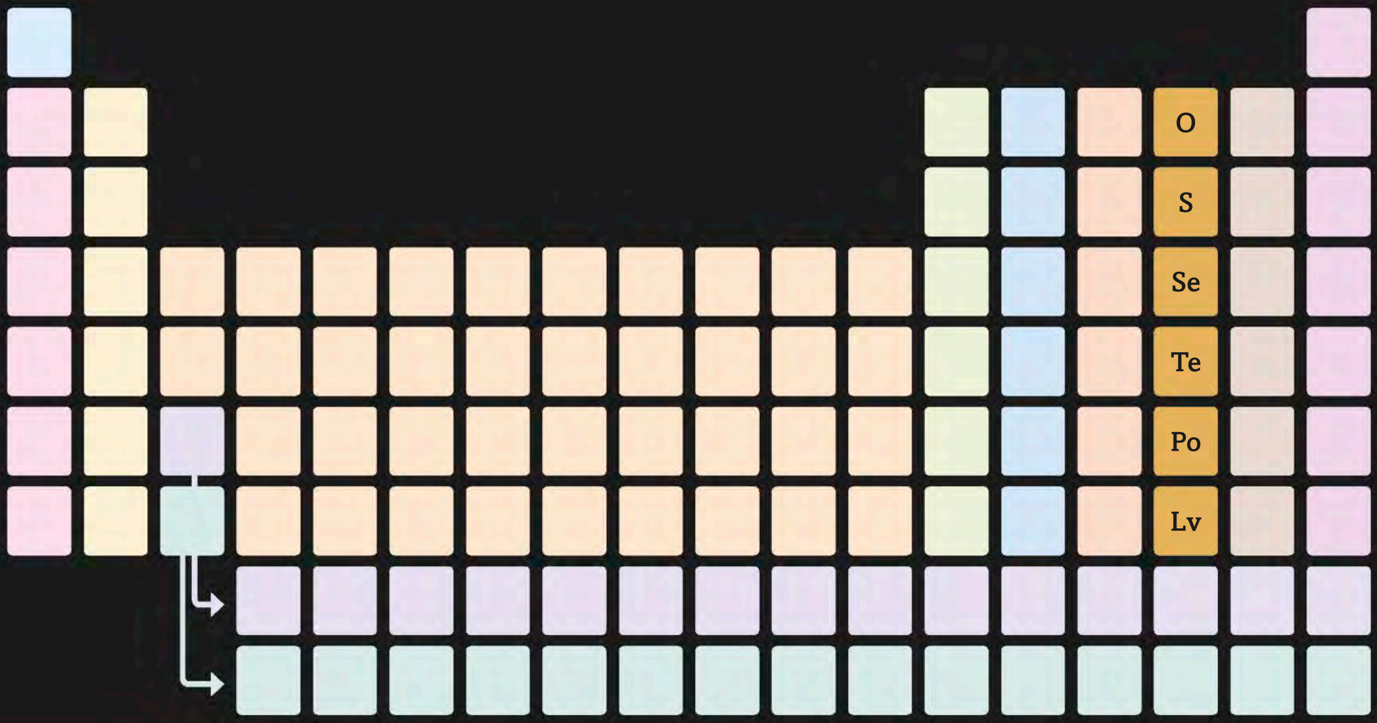

The Oxygen Group

This group does not include any natural metals. The first two members, oxygen (O) and sulfur (S), are non-metals widespread in nature. The remaining three natural elements are semi-metals. Only the artificial member, livermorium (Lv), is thought to be a metal, but chemists don't really know for sure.

Atomic structure
All members have six electrons in the outer shell of each atom. This electron structure makes these elements highly reactive.

Physical properties
The members of this group are solids, except oxygen (O), which is a gas at room temperature. The density of the elements increases down the group.

Chemical properties
The reactivity of these elements decreases down the group. Oxygen is always involved in the process of combustion.

Compounds
These elements can form compounds with each other. They all react with carbon (C) to form compounds, some with strong smells.

8

O

Oxygen

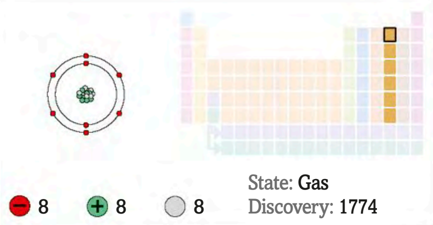

State: Gas
Discovery: 1774

Forms

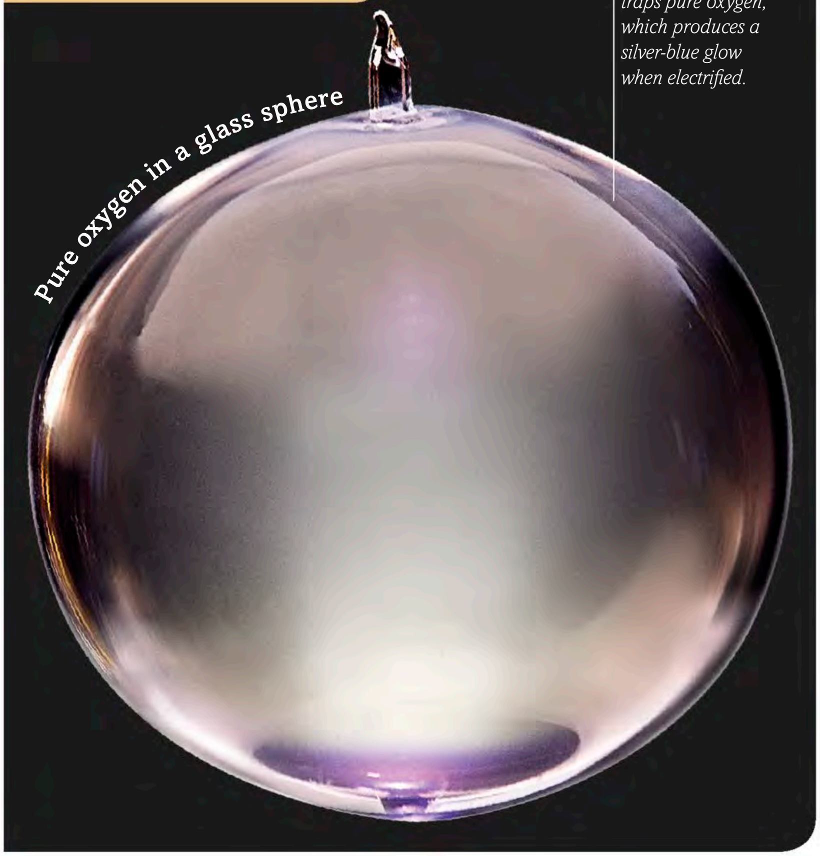

This glass sphere *traps pure oxygen, which produces a silver-blue glow when electrified.*

Streaks of light *adorn the sky as atoms of oxygen in the air are hit by a stream of particles blasting from the Sun.*

Northern lights

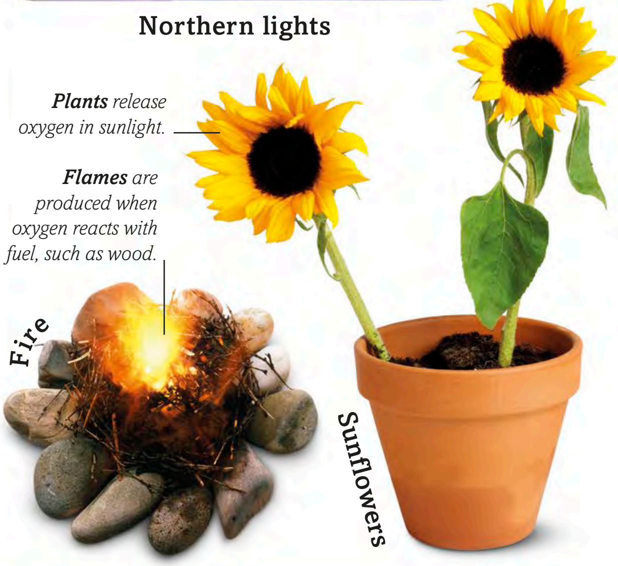

Plants *release oxygen in sunlight.*

Flames *are produced when oxygen reacts with fuel, such as wood.*

Each water *molecule has two atoms of hydrogen and one of oxygen.*

WHAT IS COMBUSTION?

Combustion is a chemical reaction that produces heat and light. Oxygen is essential for combustion to take place.

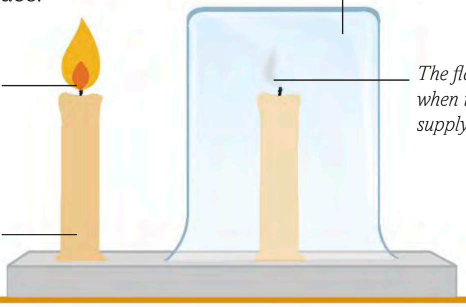

The flame is steady in the presence of oxygen.

Candle wax is made of a flammable carbon compound, which acts as a fuel.

A beaker is placed over a candle, starving the flame of oxygen.

The flame goes out when the oxygen supply is cut off.

Oxygen is the most common element in Earth's crust. Oxygen and its compounds make up half of all rocks and minerals on our planet. In the atmosphere, **pure oxygen** makes up around one-fifth of the air. This element is a transparent gas. Life on Earth depends on oxygen for survival. Animals breathe in air to collect the oxygen in it. Our bodies' cells then use that oxygen to break apart sugars to release energy, which powers our bodies. Another process that involves oxygen is the burning reaction called combustion, in which oxygen

Uses

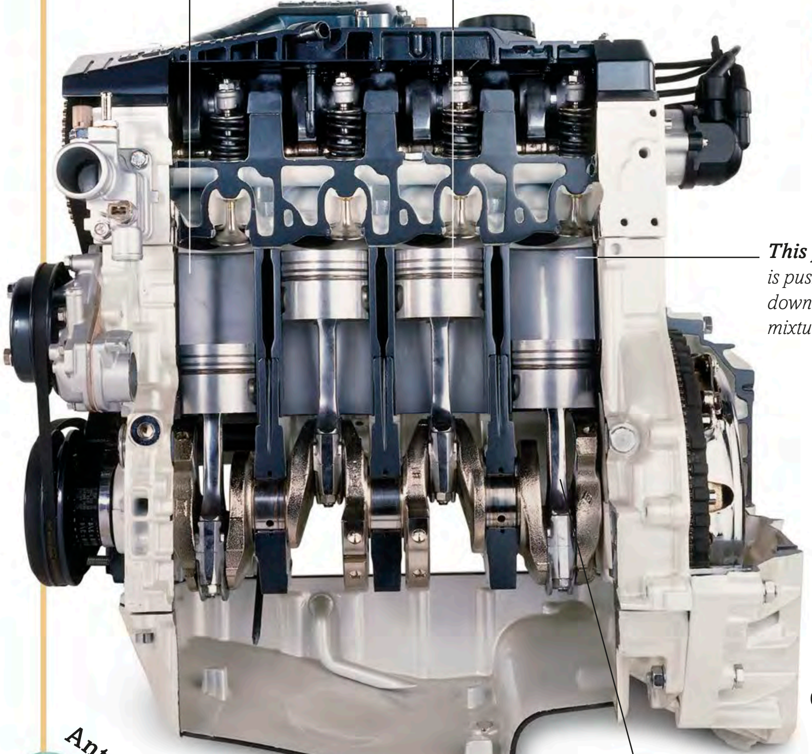

Internal combustion engine

This cylinder *takes in oxygen along with the fuel.*

The piston *then squeezes the mixture of oxygen and fuel.*

This piston *is pushed down when the mixture explodes.*

The up-and-down motion of the pistons *makes the wheels of a car spin.*

Steelmaking

Hot steel *may have impurities that are burned away using pure oxygen.*

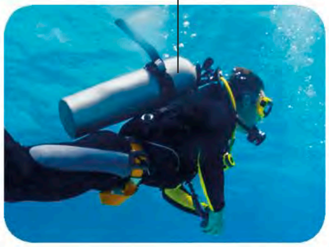

Scuba diver

The oxygen in this cylinder *can last for up to 30 minutes.*

Atlas V rocket

This rocket *carries liquid oxygen that mixes with the fuel to produce hot gases, which help the rocket lift off.*

Antifreeze

This liquid *contains an oxygen compound that prevents water from freezing in an engine.*

Oxygen is **colourless** as a gas, but looks **pale blue** as a liquid.

Mountaineer with oxygen supply

Mountaineers *have to adjust slowly to changing levels of oxygen as they climb higher.*

Medical oxygen cylinder

The oxygen in this cylinder *is given to patients in controlled amounts.*

Welding torch

This flammable gas-oxygen mixture *burns hot enough to melt metal.*

reacts with a fuel and produces **fire**. Oxygen is also used up when it reacts with other elements to form compounds called oxides. However, it is replenished by **plants** through a process called photosynthesis, which releases fresh oxygen. Car **engines** are powered by the combustion of petrol or other fuels. Oxygen is also useful in the **making of steel**. Tanks of oxygen let **mountaineers** breathe easily in environments that have low levels of this gas. Rockets, such as the **Atlas V**, carry liquid oxygen to burn fuel in the absence of air in space.

16 S Sulfur

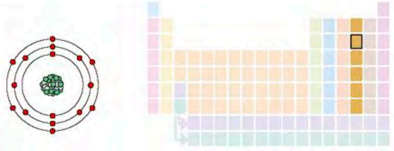

⊖ 16 ⊕ 16 ○ 16

State: Solid
Discovery: Prehistoric

Forms

Natural sulfur

These yellow sulfur crystals *often attach to volcanic mud.*

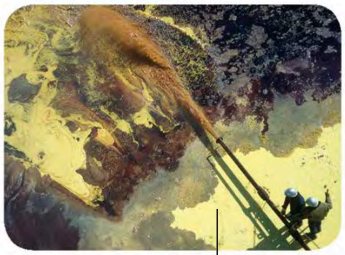

Liquid sulfur

Hot liquid sulfur *is pumped up from an underground mine.*

Celestine

This grey-blue crystal *contains a sulfur compound called strontium sulfate.*

Laboratory sample of pure sulfur

Soft, brittle granules

Skunk

Skunks *release a foul spray that contains three kinds of sulfur compounds.*

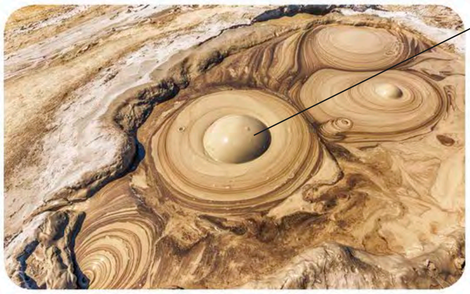

Mud volcano

Hydrogen sulfide bubbles *form in volcanic mud.*

Onions

Cut onions *release sulfur compounds that make our eyes water.*

Known since ancient times, sulfur is one of the few non-metals that can be found pure in nature. This yellow, crystalline element is found in large amounts near volcanic craters. Another name for sulfur is “brimstone”, which refers to the way its crystals burn, melting into a blood-red liquid. In some religions, brimstone is thought to be the fuel that burns in the underworld. **Pure sulfur** is extracted from underground deposits using hot water. The hot **liquid sulfur** is then pumped to the surface. This element is a common ingredient in many

Uses

Vulcanized rubber *– made by heating sulfur with natural rubber – is weatherproof.*

Vulcanized tyre

Preserved dried fruits

Some dried fruits *are preserved using powder containing a sulfur compound.*

When burned, sulfur in this candle *drives away pests.*

Sulfur candle

ACID RAIN

Sulfur dioxide gas produced by burning fuel dissolves in rain water, making a sulfur acid. This falls to the ground as acid rain.

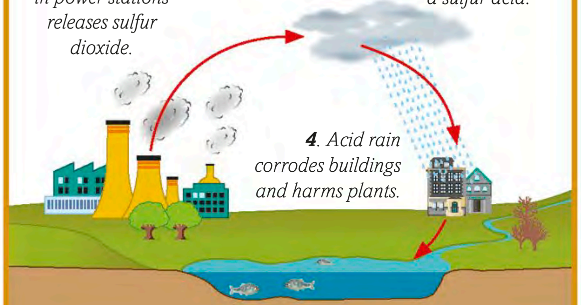
1. Burning coal in power stations releases sulfur dioxide.
2. The wind carries pollutants.
3. Sulfur dioxide mixes with water in clouds and forms a sulfur acid.
4. Acid rain corrodes buildings and harms plants.
6. It also makes rivers and lakes acidic.
5. Acid rain changes soil chemistry.

This plant *emits a rotting smell to attract meat-eating insects.*

Titan arum

Creams that contain sulfur compounds *can disinfect skin.*

Skin cream

Lead acid battery

This battery *contains strong sulfuric acid.*

Penicillin pills

Some antibiotics *contain sulfur compounds that kill harmful bacteria.*

Acid rain **destroys forests** by removing nutrients in the soil and in the leaves.

Sulfur acid rain damage

This limestone sculpture *has been weathered by acid rain.*

minerals, such as **celestine**. Many sulfur compounds smell bad. For example, the rotten-egg smell of volcanic pools is due to hydrogen sulfide gas. Other examples include **skunk** spray, the gaseous substance emitted by chopped **onions**, and the odour of the **titan arum** flower. There are many uses for this non-metal. Its compounds can harden natural rubber for use in **tyres**, preserve **dried fruits**, and make strong **battery** acids. The element has anti-bacterial properties and is used in antibiotic medicine, such as **penicillin**.

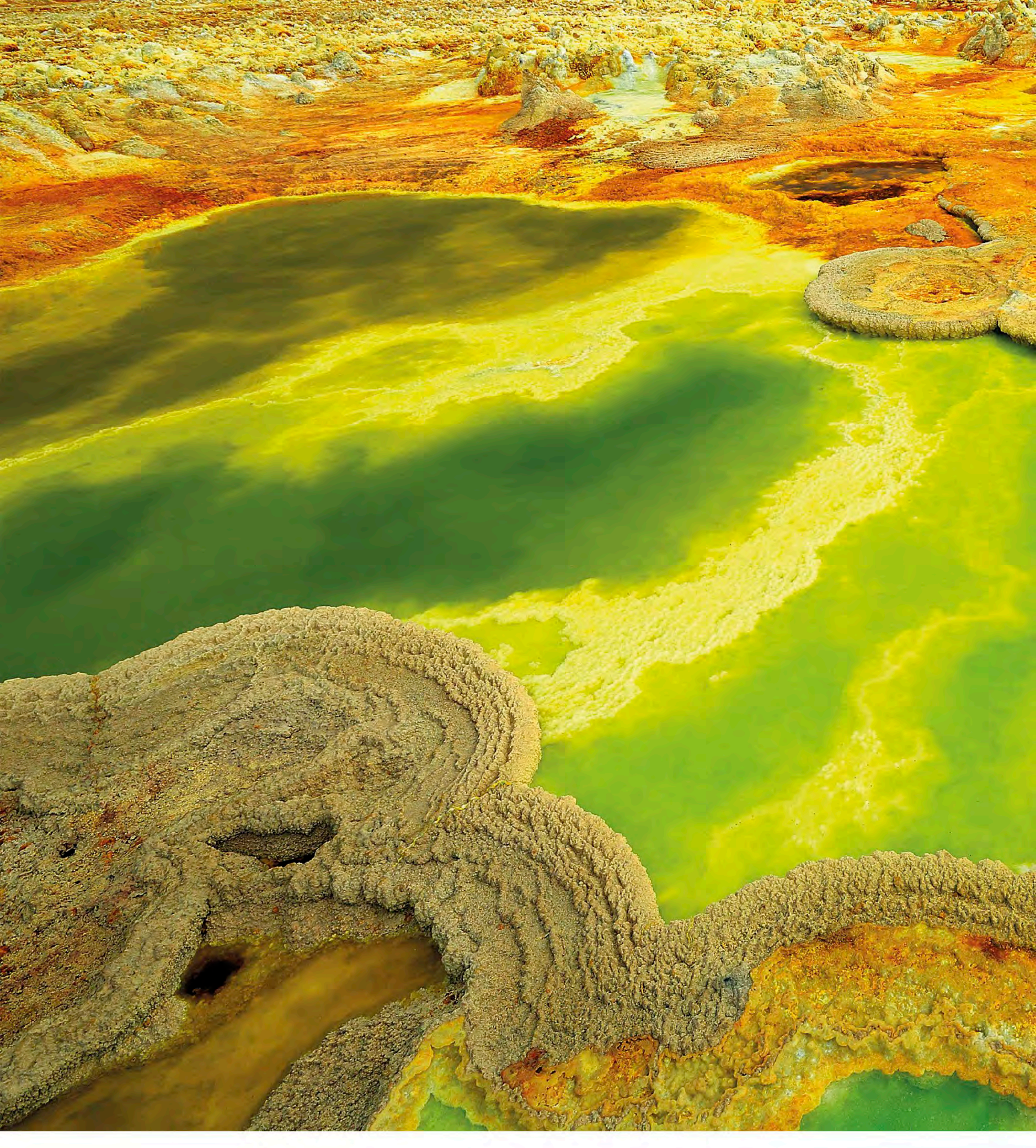

DANAKIL DEPRESSION This hot spring in Africa's Danakil Depression is surrounded by a yellow crust of pure sulfur. The sunken region between Ethiopia and Eritrea in East Africa is a wild volcanic area, packed with erupting craters, arid deserts, boiling mud, and pools with unusual colours caused by the presence of sulfur and many mineral salts.

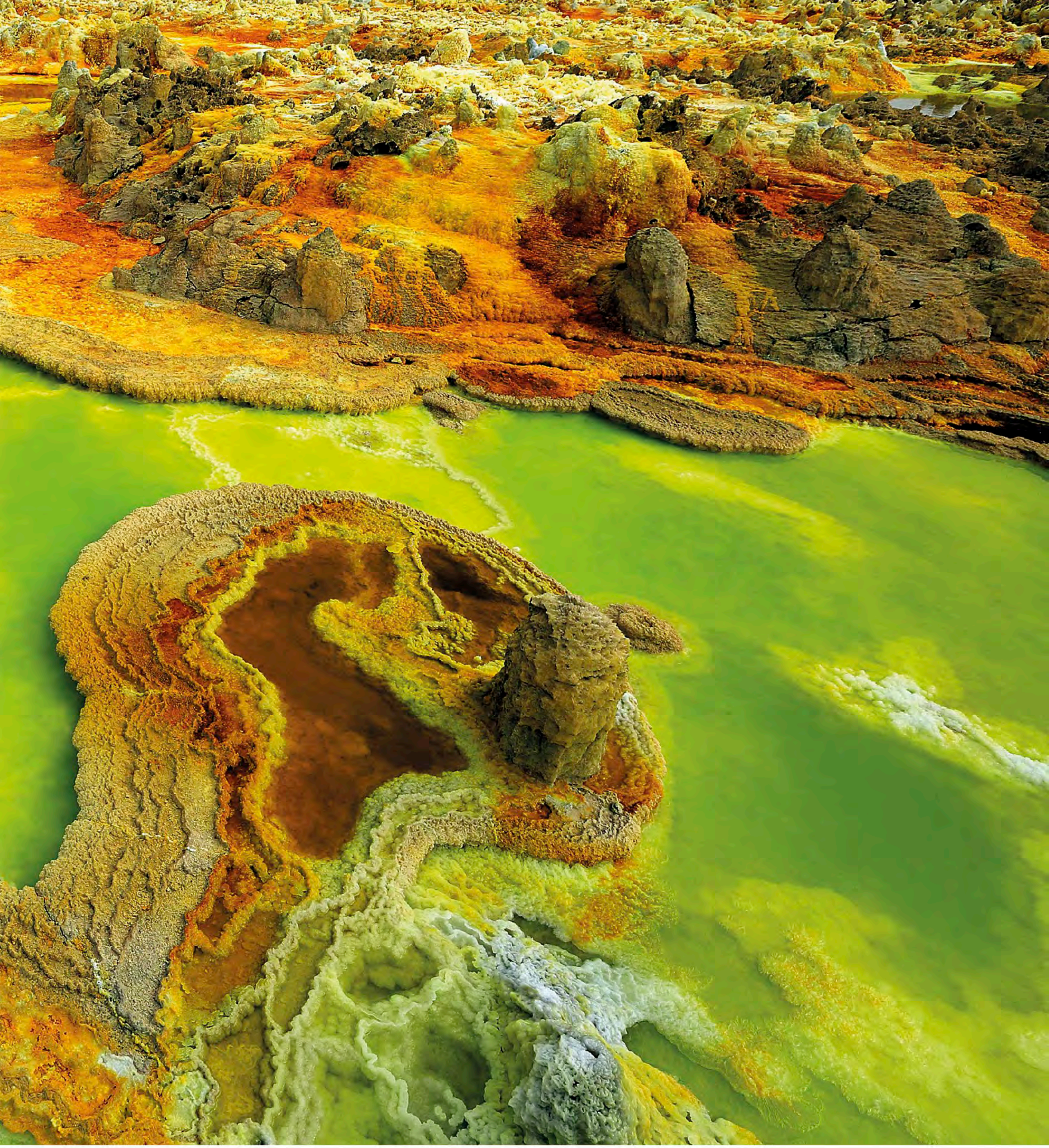

At more than 100 m (330 ft) below sea level, the Danakil Depression is one of the lowest points on Earth's surface. This area receives little or no rainfall and the weather is hot and dry, with temperatures soaring above 50°C (120°F). The scalding green water of the springs inside the depression contains pure sulfur as well as a toxic sulfur compound called sulfuric acid. As the water evaporates, sulfur deposits build up around the edges of the pools, making beautiful shapes across the vast landscape. Tourists visit to marvel at the remarkable sights at Danakil, even though the inhospitable conditions in the area give it the title of the "cruellest place on Earth".

34
Se

Selenium

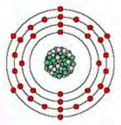

⊖ 34 ⊕ 34 ○ 45

State: Solid
Discovery: 1817

Forms

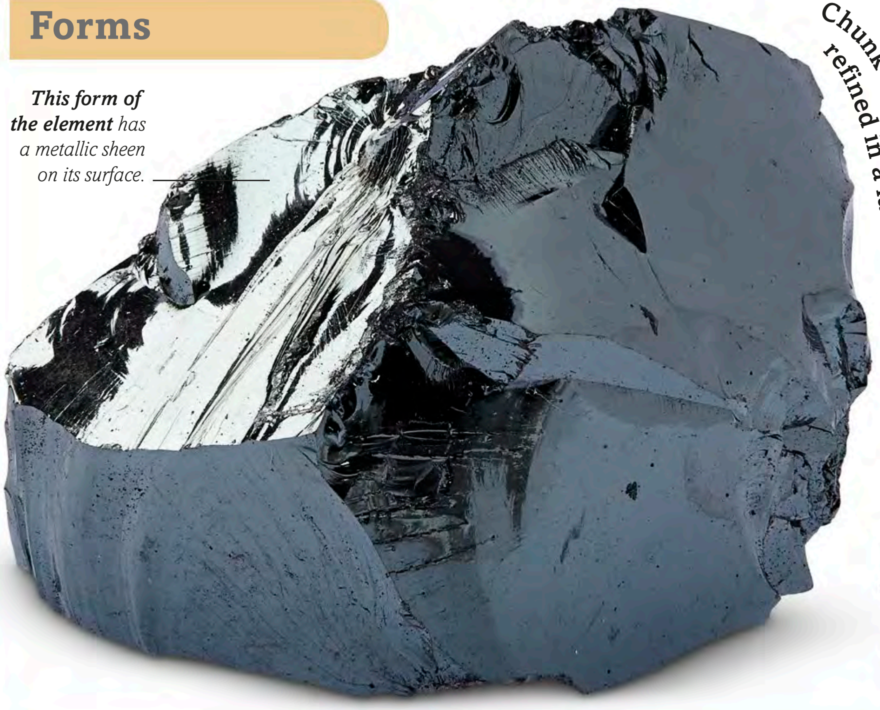

***This form of the element** has a metallic sheen on its surface.*

Chunk of pure grey selenium refined in a laboratory

Brazil nuts

***These nuts** are the richest food source of selenium.*

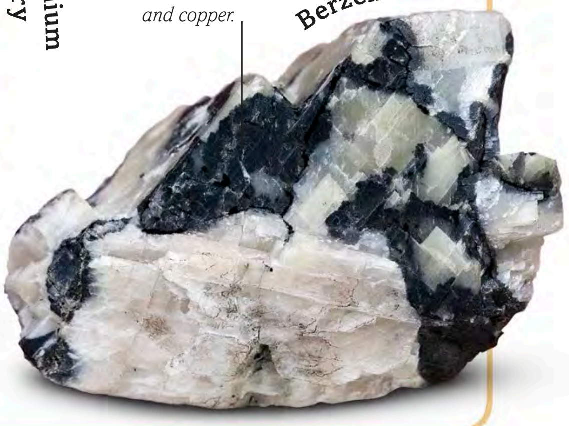

Berzelianite

***These dark areas** contain selenium and copper.*

Uses

Calculator

***This calculator** is powered by a solar cell made of selenium and nickel.*

Anti-dandruff shampoo

***A selenium compound** in this shampoo treats dandruff.*

Ceramic vase

***This bright colour** comes from the red selenium in the glaze.*

Photocopier

***Many office photocopiers** use selenium in powdered form.*

Selenium is named after Selene, the Greek goddess of the Moon. This element is a semi-metal and so has the properties of both metals and non-metals. Selenium has two main pure forms: **grey selenium**, which is a hard substance, and red selenium, which is a soft powder. The most common use of selenium is as an ingredient that provides colour in glass and **ceramics**. Selenium is sensitive to light, so it is used in solar cells that convert sunlight into electricity. It is also utilized in **photocopy machines**.

52
Te
Tellurium

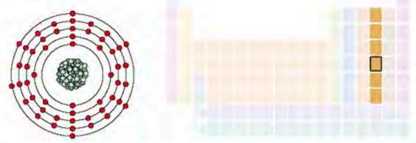

● 52 ⊕ 52 ○ 76

State: Solid
Discovery: 1783

Forms

Melonite

This ore *is a soft but dense solid.*

Sylvanite

This metallic crust *contains tellurium, gold, and silver.*

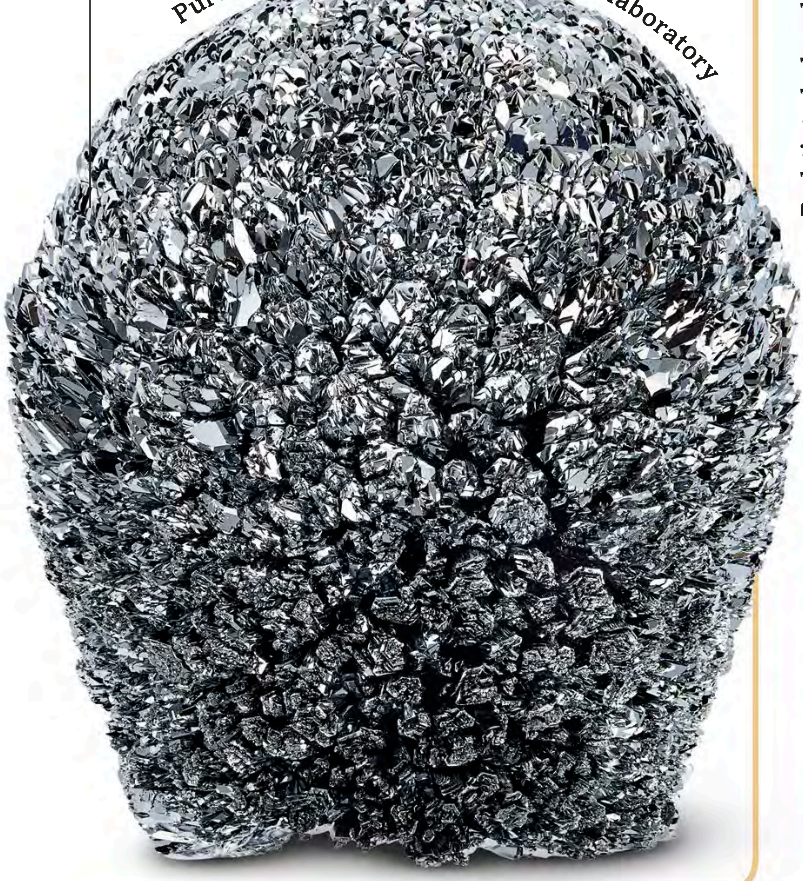

Pure tellurium crystals refined in a laboratory

This semi-metal *can form silvery white crystals.*

Uses

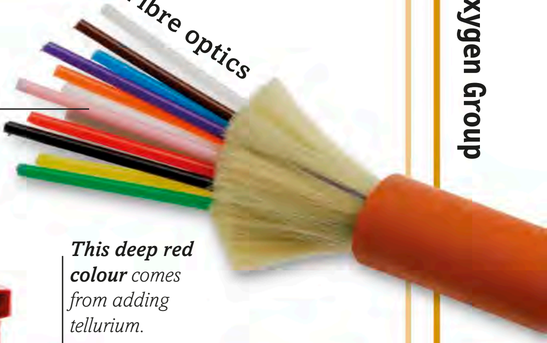

Fibre optics

These glass fibres *contain tellurium.*

Red-tinted glass bottle

This deep red colour *comes from adding tellurium.*

Tellurium **protects** bronze **from corroding** easily in air.

Topaz solar farm, California, USA

The solar panels *are linked to power cells containing tellurium.*

Tellurium is one of the 10 rarest elements on this planet. It gets its name from the Latin word *tellus*, which means Earth. This element is often found as a compound with another element, such as the metal nickel, as in the case of the ore **melonite**. Tellurium is also produced as a waste product when lead and copper are refined. **Pure tellurium** can take on two forms: a shiny, metallic solid or a brown powder. This element is mainly used in the glass of **fibre optics**, which carry high volumes of information much faster than copper cables.

84

Po

Polonium

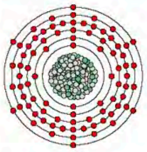

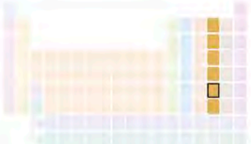

⊖ 84 ⊕ 84 ○ 125

State: Solid
Discovery: 1898

Forms

This uranium ore contains **0.0000001 per cent** polonium.

Uraninite

This mineral *contains uranium atoms that break apart to form atoms of polonium.*

Uses

Anti-static brush

This brush *is used to remove static electric charge from camera lenses and music records.*

Atom bomb

This bomb *can be detonated when the polonium inside it ignites.*

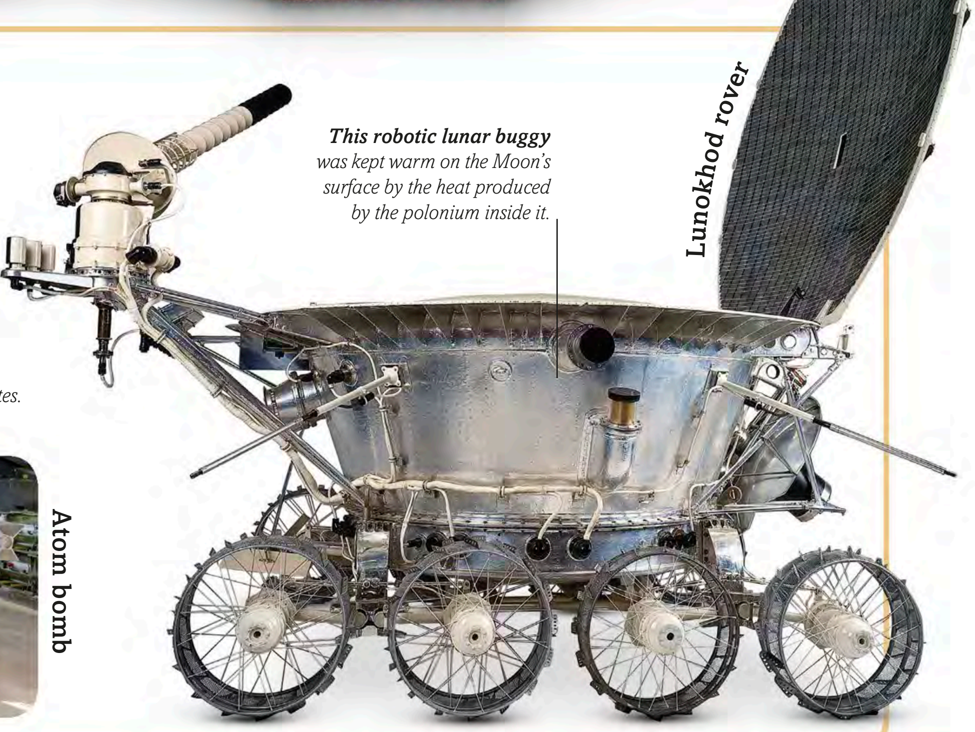

Lunokhod rover

This robotic lunar buggy *was kept warm on the Moon's surface by the heat produced by the polonium inside it.*

Polonium is very radioactive: 1 g (0.03 oz) of this metal quickly heats up to 500°C (932°F) because of the radiation it emits. This element was discovered by Marie and Pierre Curie in 1898. Marie named it after Poland, her homeland. It is rare in nature, and is normally produced in nuclear reactors. Despite its radioactivity, this element is used in a few ways. It can trigger the explosion of **atom bombs**. It heats and powers spacecraft, such as the Russian **Lunokhod rovers**, which landed on the Moon in the 1970s.

116

Lv

Livermorium

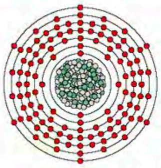

⊖ 116 ⊕ 116 ○ 177

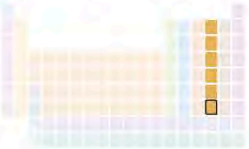

State: Solid
Discovery: 2000

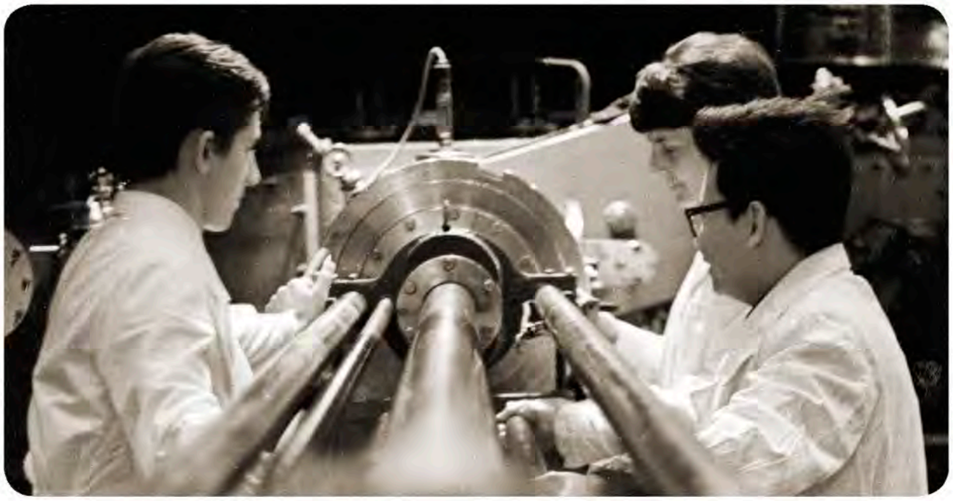

Heavy ion cyclotron,
Joint Institute of Nuclear Research, Dubna, Russia

Livermorium has been **named after** this **laboratory**.

Lawrence Livermore National Laboratory, California, USA

When livermorium atoms were first produced in 2000, they broke apart in a fraction of a second. The first successful attempt to create atoms of this element was made at the **Joint Institute of Nuclear Research** at Dubna, Russia. The team worked with material provided by the **Lawrence Livermore National Laboratory** in California, USA. This highly radioactive element was produced by firing calcium atoms at curium atoms in a particle accelerator (a machine in which atoms are smashed together).

This glass sphere contains pure iodine (I).

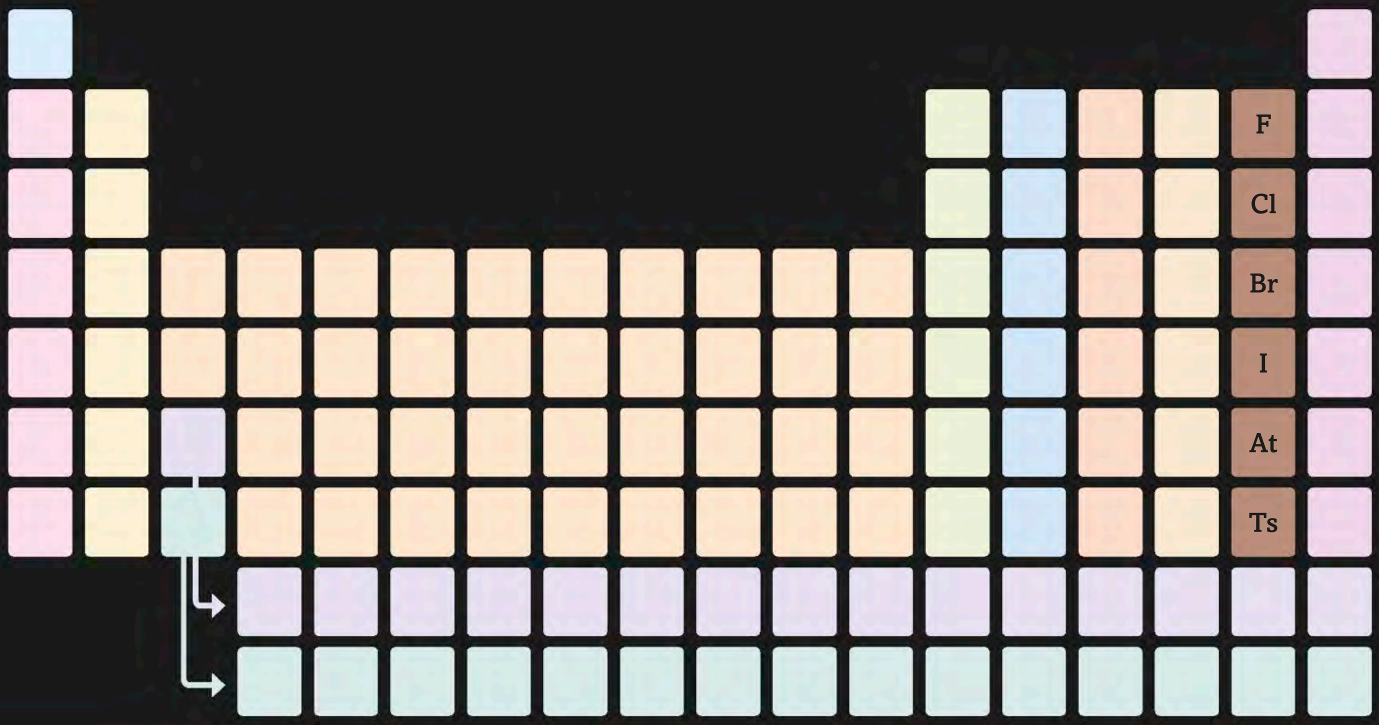

The Halogen Group

One of the most reactive groups in the periodic table, this set includes non-metals. The name “halogen” means “salt former”, which refers to the way that elements in this group react with metals to form salts, such as sodium chloride, widely known as common salt. Scientists don’t know much about tennessine (Ts), an artificial halogen.

Atomic structure
All members have seven electrons in the outer shell of each atom. There is space for one more electron in each outer shell.

Physical properties
Bromine (Br) is the only halogen that is liquid. Fluorine (F) and chlorine (Cl) are gases, while iodine and astatine (At) are solids.

Chemical properties
Every halogen atom receives one electron from other atoms to form a compound. Reactivity decreases down the group.

Compounds
The halogens react with hydrogen (H) to form acidic compounds. Halogen compounds are used in products such as bleach.

9

F

Fluorine

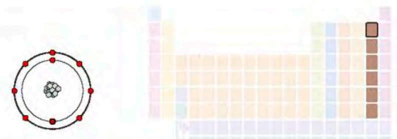

⊖ 9 ⊕ 9 ○ 10

State: Gas
Discovery: 1886

Forms

This soft mineral *is brittle, which means it breaks easily into lots of pieces.*

Cryolite

Laboratory sample

This sealed case *contains a mixture of fluorine and helium.*

> Topaz **means “fire”** in the ancient Indian language of Sanskrit.

These cubic crystals *are green due to impurities.*

Fluorite

Topaz

This precious gemstone *contains 20.7% fluorine.*

This highly reactive element is incredibly dangerous when pure: just a tiny amount added to the air can kill a person. A pale yellow gas, fluorine reacts with brick, glass, and steel, burning a hole straight through them. Because it is so dangerous, pure fluorine is often stored in nickel containers that can resist its attack. Minerals such as **cryolite** and **fluorite** contain this element. This gas and its less harmful compounds have a wide variety of uses. Hydrofluoric acid is a toxic liquid used to etch patterns on glass, as seen in some **glass vases**.

Uses

HENRI MOISSAN

In the early 1800s, chemists in Europe realized that minerals such as fluorite contained an unknown element. However, it took another 70 years before the French chemist Henri Moissan extracted pure fluorine following a series of dangerous experiments that even poisoned him several times.

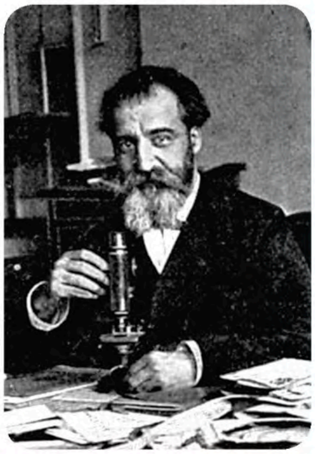

Etched glass vase

These patterns *form after the surface of the glass is burned by an acidic fluorine compound.*

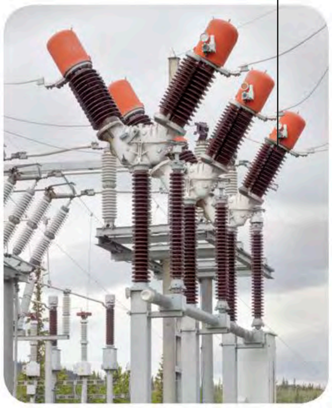

These circuit breakers*, containing a compound of fluorine and sulfur, cut electricity supply in an emergency.*

Circuit breakers

Ceramic pot

The shiny coating *on this ceramic dish is from a fluorine-containing glaze.*

Fluorine-rich plastic clothing *repels water.*

Waterproof clothing

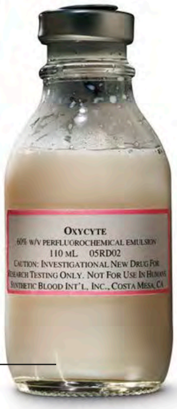

OXYCYTE
60% W/V PERFLUOROCHEMICAL EMULSION
110 mL 05RD02
CAUTION: INVESTIGATIONAL NEW DRUG FOR RESEARCH TESTING ONLY. NOT FOR USE IN HUMANS
SYNTHETIC BLOOD INT'L, INC., COSTA MESA, CA

Upon injection, this fluorine-rich liquid *heals damaged tissues by carrying oxygen to them.*

Oxycyte

PTFE was used by NASA to make **spacesuits** because it protected against heat.

This pan *has a PTFE coating that resists heat.*

Non-stick pan

Some toothpastes *contain fluorides that harden tooth enamel.*

Toothpaste

Some glazes used to coat ceramics contain fluorine minerals. When heated, these glazes release fluorine, which hardens the ceramic underneath. Another compound called polytetrafluoroethylene (PTFE) is commonly used to make **non-stick pans**: this material is slippery and prevents food that has burned while cooking from sticking to the pan. Thin fibres made of PTFE are also used to make lightweight, **waterproof clothing**. One of the most common uses of fluorine compounds is in **toothpaste**: they toughen teeth against decay.

17
Cl

Chlorine

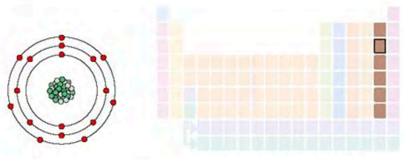

● 17 ⊕ 17 ○ 18

State: Gas
Discovery: 1774

Forms

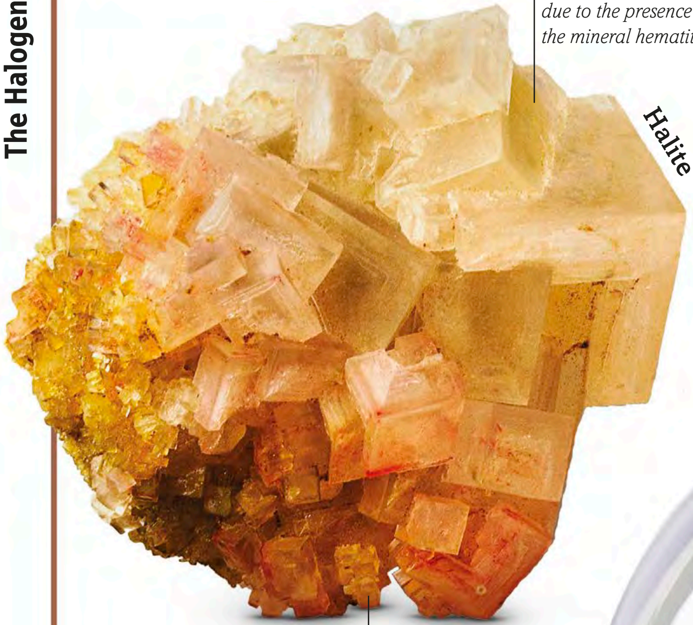

Halite

This orange colour *is due to the presence of the mineral hematite.*

Cube-shaped crystals

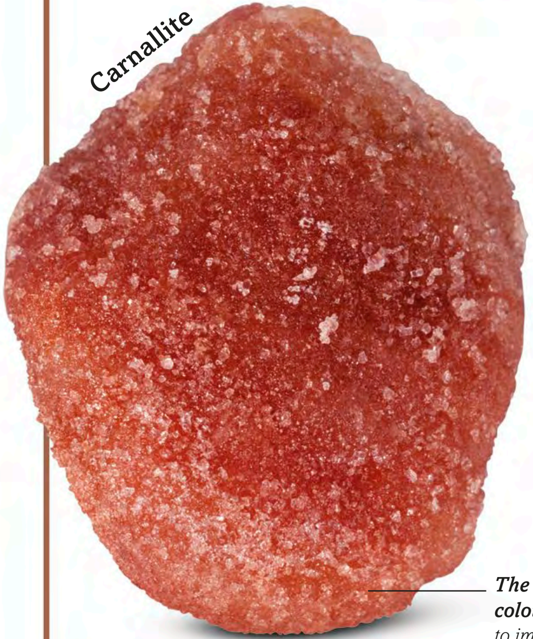

Carnallite

The intense red colour *is due to impurities.*

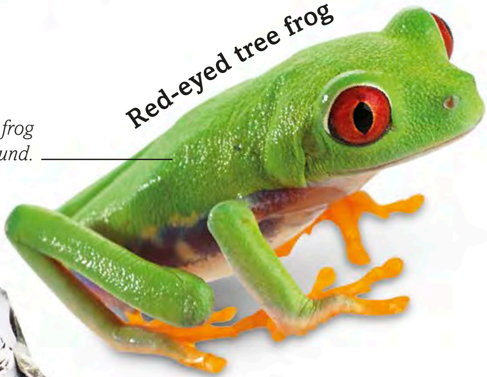

Red-eyed tree frog

The skin *of this tree frog contains a chlorine compound.*

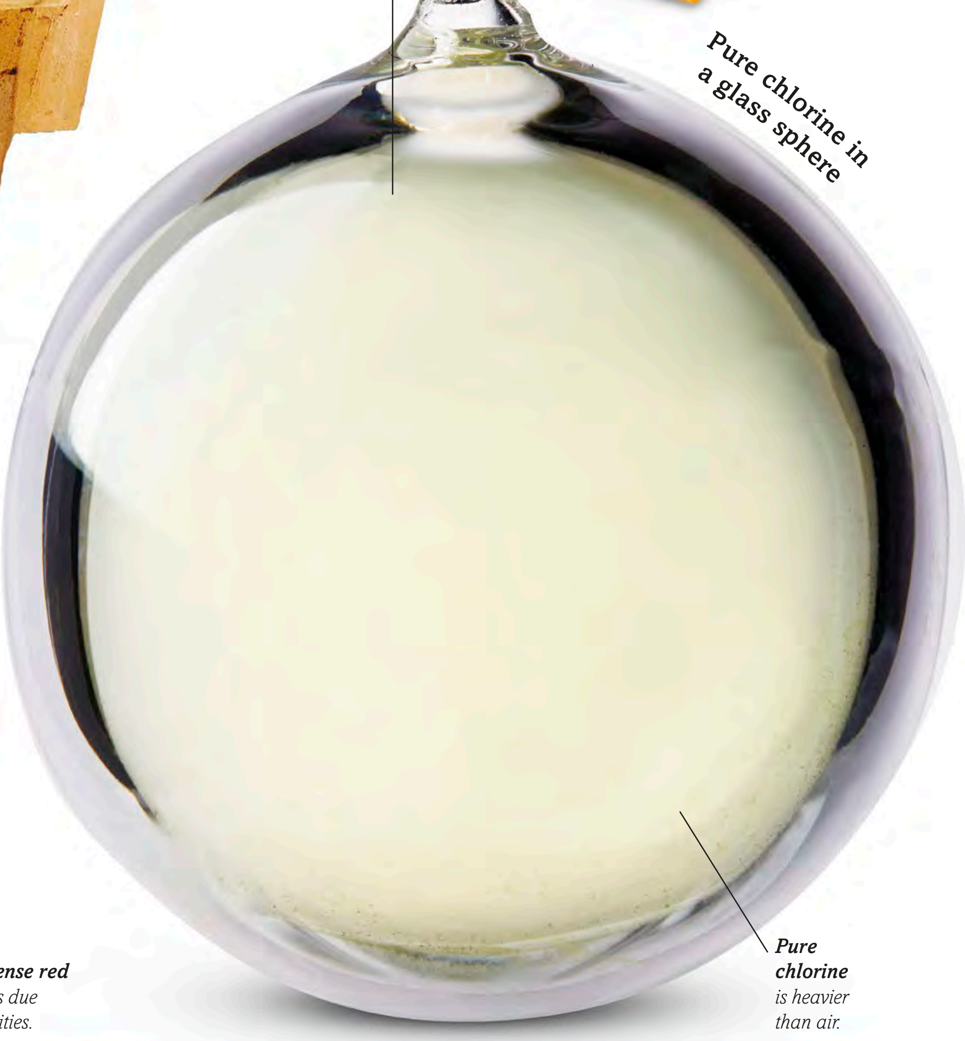

Pure chlorine in a glass sphere

This glass sphere *traps pure chlorine so as to prevent any reactions with air.*

Pure chlorine *is heavier than air.*

Chlorine is named after the Greek word *chlóros*, which means "pale green", a reference to the colour of this gaseous element. Chlorine is a highly reactive gas that forms a number of compounds, and does not exist pure in nature. The most common chlorine compound is sodium chloride, found in nature as the mineral **halite**. Chlorine compounds are important for the body and are used by muscles and nerves. They are also present in sweat. As it is poisonous in its pure form, chlorine gas was used as a weapon during World War I: soldiers had to wear masks for protection

Uses

Running shoes

The soles of some running shoes *contain chlorine compounds.*

Common salt

Sodium chloride *is used to season meals.*

CHLORINATION

The process of cleaning dirty water with chlorine is called chlorination. This method involves passing the water through a filter.

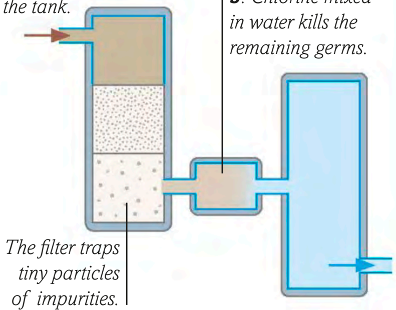

1. Dirty water enters the tank.

2. The filter traps tiny particles of impurities.

3. Chlorine mixed in water kills the remaining germs.

4. Clean water for drinking, cooking, and washing is released.

Chloroform

Breathing in this chlorine-infused liquid *makes people fall unconscious.*

PVC pipes

Strong water pipes *are composed of thick PVC.*

Safety goggles

This chlorine-rich plastic *is tough.*

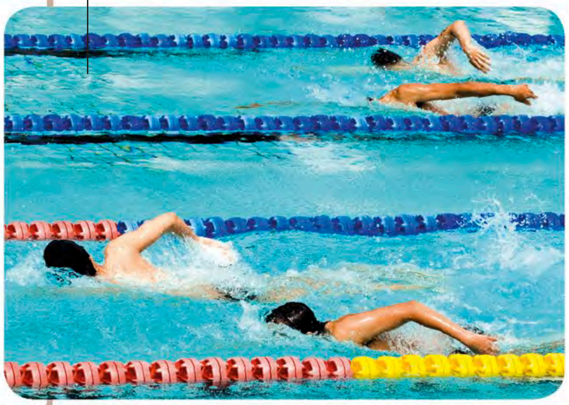

Swimming pool

The chlorine levels in this pool *need to be regulated carefully to maintain the quality of water.*

Bleach

This bleach cleaner *contains a compound called sodium hypochlorite.*

PVC suitcase

This case *is tough but flexible.*

against this weapon. Today, chlorine is used in many ways. Its compounds are present in everything from **running shoes** to **choloroform**. It reacts with hydrogen to make hydrochloric acid, an industrial cleaner. This corrosive liquid eats away at most metals, releasing hydrogen gas. A weaker chlorine acid is used to clean water in **swimming pools**, while **bleach** and other cleaners use chlorine compounds to kill germs. One of the most widely used plastics, **polyvinyl chloride (PVC)**, contains chlorine. It is a tough plastic, used to make many rigid objects.

SPIRO
NARVAL

OCEAN CLEAN UP

Chlorine is a common ingredient in cleaning products, and can be used for scrubbing everything from bathroom tiles to ocean floors. These divers are trying to remove harmful seaweed in the Mediterranean Sea using the power of chlorine. This green weed grows quickly and can potentially kill other sea plants by depriving them of their essential nutrients. Some fish are also poisoned if they eat this toxic weed.

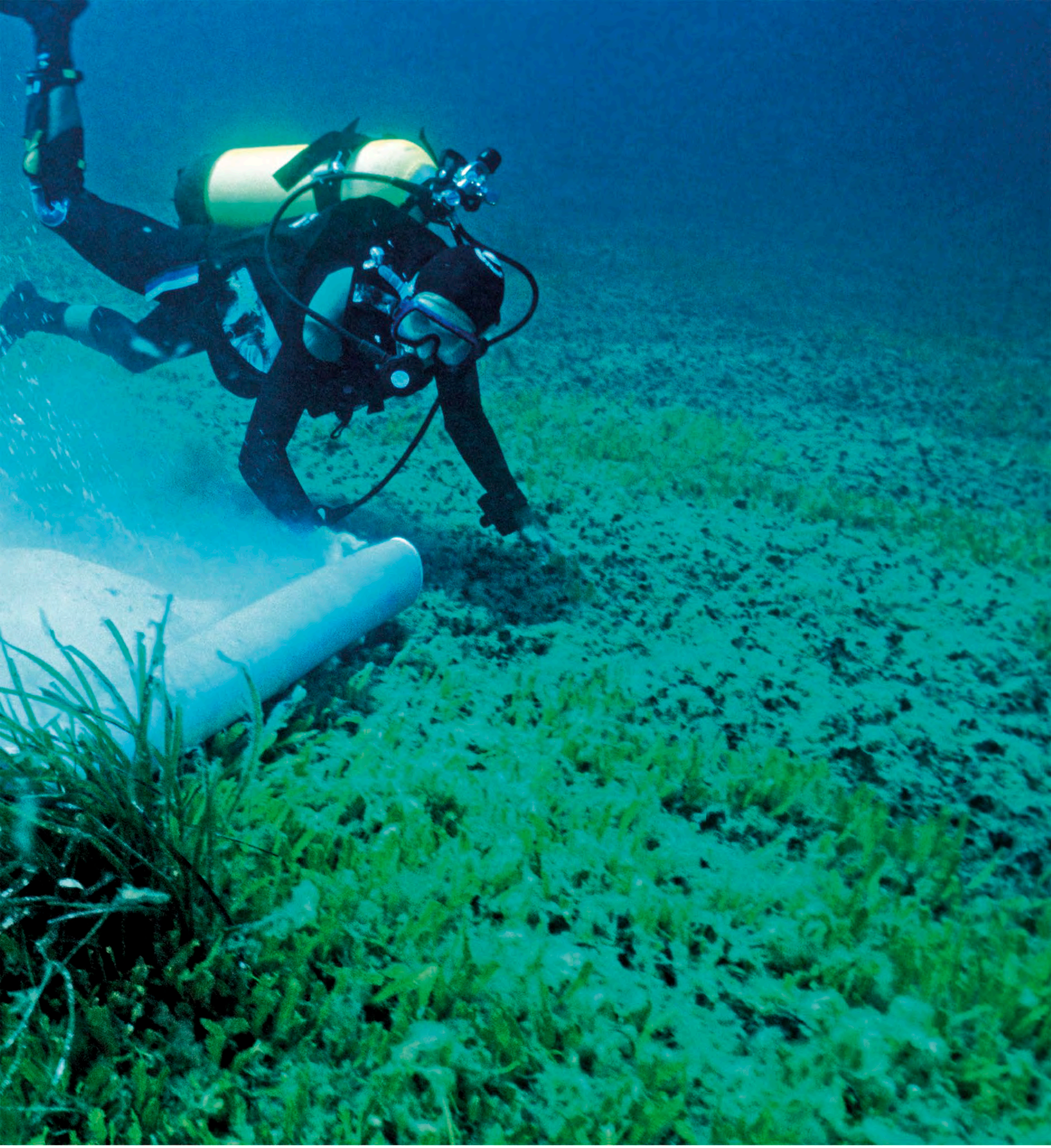

The two divers use chlorine twice in the cleaning process. First, they cover the thick seaweed with a sheet made of PVC, a tough plastic containing chlorine. Next, they pump a compound of sodium and chlorine called sodium hypochlorite under the sheet. This powerful liquid bleach kills the unwanted seaweed. Several weeks later, the divers return to remove the PVC sheets. The invading seaweed will not regrow, and the plants on the seabed will gradually return. Although chlorine is highly reactive and can damage skin and other body parts, divers are well-protected by their rubber wetsuits.

35 Br Bromine

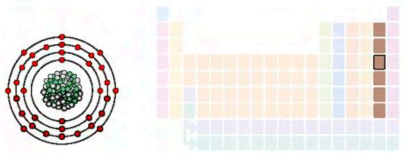

⊖ 35 ⊕ 35 ○ 45

State: **Liquid**
Discovery: **1826**

Forms

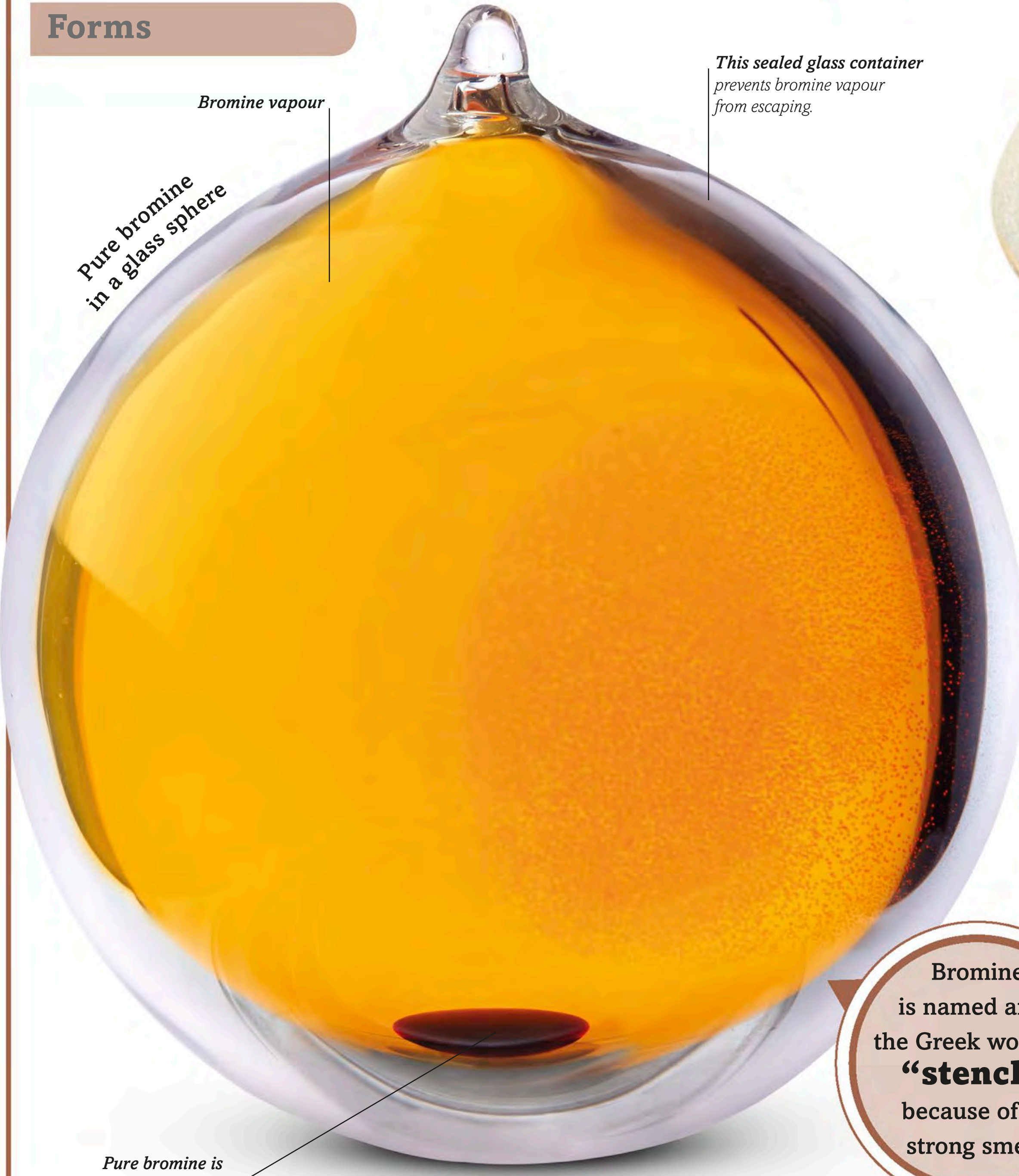

Pure bromine in a glass sphere

Bromine vapour

This sealed glass container *prevents bromine vapour from escaping.*

Pure bromine is *a red-brown liquid.*

Potassium bromide

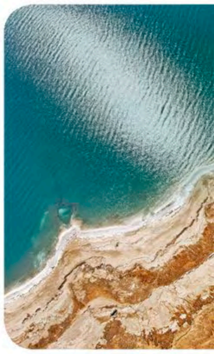

Bromine is named after the Greek word for **"stench"** because of its strong smell.

Bromine is the only non-metal that is a liquid at room temperature. A thick vapour given off by this liquid is dangerous if breathed in. **Pure bromine** is never found in nature. Its compounds can easily mix in water, and are found dissolved in seawater and extremely salty lakes, such as the **Dead Sea** in the Middle East. Solid bromine salts, including **potassium bromide**, collect as the water evaporates away, leaving behind crusts of white crystals. Bromine can then be extracted from the solid salts. A common

Uses

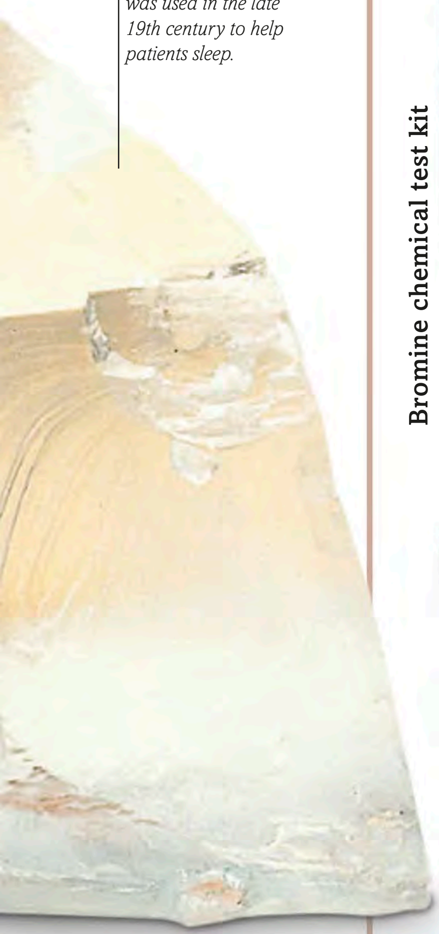

This bromine salt *was used in the late 19th century to help patients sleep.*

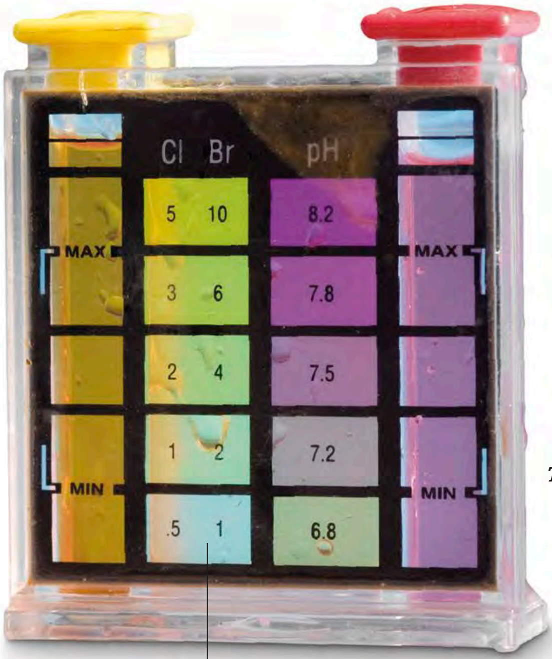

Bromine chemical test kit

The colours on this test kit *show how much bromine is in the water.*

Chubb Fire
HALON
Fire Extinguisher
Pull out pin
Free hose
Squeeze levers
Aim at fire base

Fire extinguisher

This fire extinguisher *uses a non-flammable, bromine-rich gas to extinguish fires.*

Photograph negative

This image *was created when light reacted with silver bromide.*

Bromine was used as a **weapon** in World War I.

Fireproof suit

The fabric *used to make this suit contains bromine compounds that make it fireproof.*

The Dead Sea Coast

Crusts of bromine salts *form along the Israeli shoreline.*

ANTOINE-JÉRÔME BALARD

In 1826, French chemist Antoine-Jérôme Balard discovered bromine. He heated seawater from salt marshes, and after most of the water had evaporated, he passed chlorine gas through what was left. The remaining liquid turned orange-red: this was bromine.

use of this element is as a disinfectant to clean water. It works better than chlorine in hot tubs as chlorine escapes into the air easily from the warm water. The concentration of bromine in swimming pools can be regulated using **chemical test kits**. Bromine compounds can be used in film photography in which images are printed using chemicals on **negatives**. Today, bromine is mainly used in **fireproof material**, such as firefighter suits or furnishings, because it doesn't catch fire easily.

53
I

Iodine

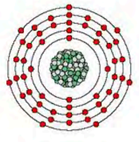

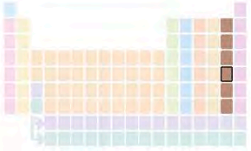

− 53 + 53 ○ 74

State: Solid
Discovery: 1811

Forms

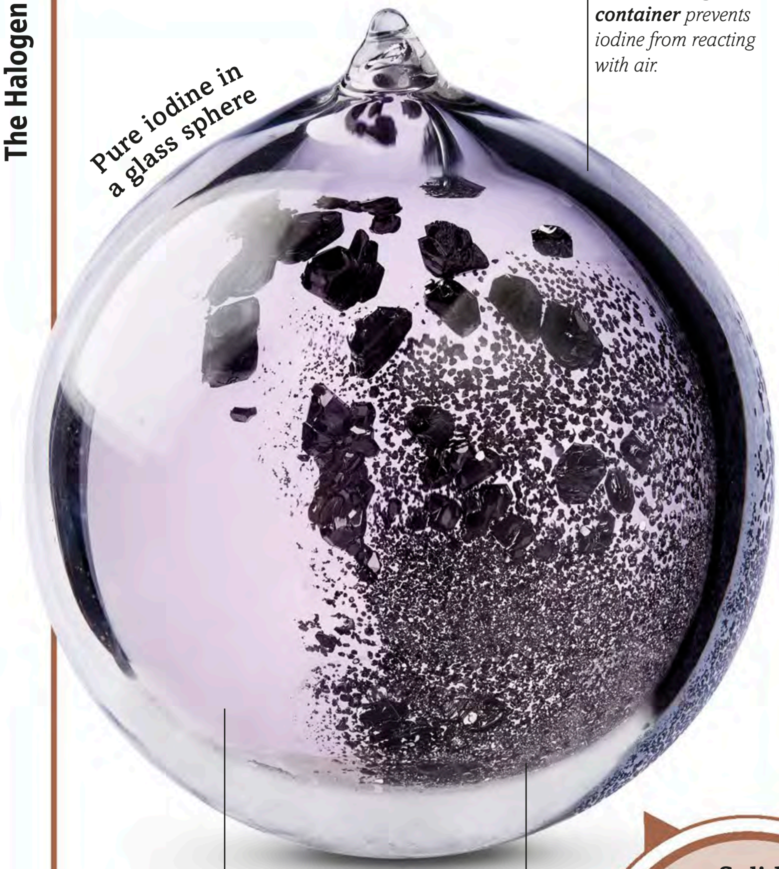

Pure iodine in a glass sphere

This sealed glass container *prevents iodine from reacting with air.*

Purple iodine vapour

Purple-black solid iodine

Solid iodine **does not melt** when heated, but turns into a vapour.

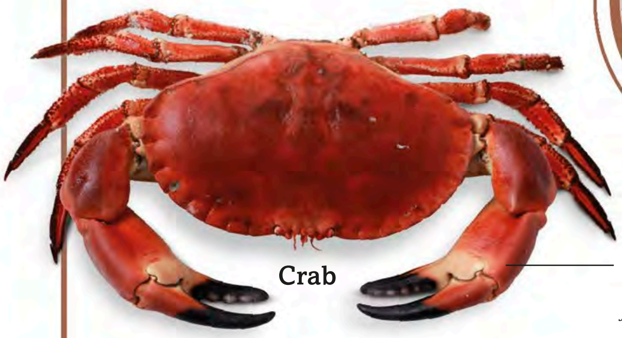

Crab

This crab *absorbs iodine from seawater.*

Uses

Printing ink

These coloured inks *are made using iodine compounds.*

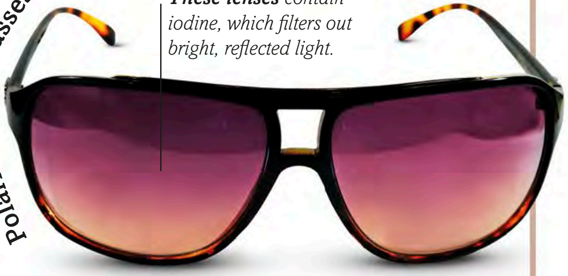

Polarizing sunglasses

These lenses *contain iodine, which filters out bright, reflected light.*

Candied cherries

The bright red colour *of these cherries is due to an iodine dye.*

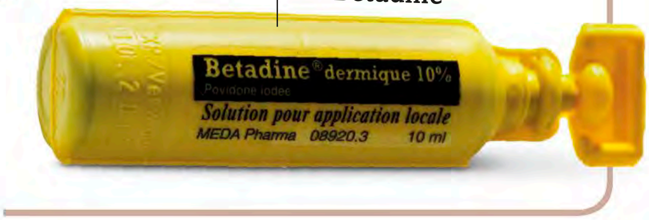

Betadine

This disinfectant *is applied to wounds to stop the spread of infection.*

Iodine is the only halogen that is solid at room temperature. The element forms a purple gas when heated, and is named after the Greek word *iodes*, which means "violet". Iodine was first discovered in seaweed, and many plants and animals in the sea have high levels of iodine. Seafood, including **crabs** and fish, provide the element in our diet. The human body needs small amounts of iodine to make an important substance called thyroxine, which helps us grow. Iodine is also used to make **printing ink**, red and brown food dyes, and disinfectants.

85

At

Astatine

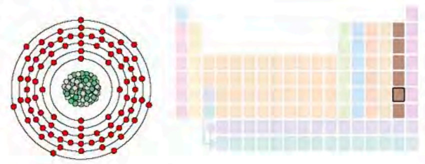

⊖ 85 ⊕ 85 ○ 125

State: Solid
Discovery: 1940

Atoms of astatine are unstable, and typically break down after just a few hours, into atoms of lighter elements, such as bismuth. This radioactive element itself forms in a similar way when atoms of a heavier element called francium break apart. Tiny amounts of this rare element are found in uranium ores, such as **uraninite**. The Italian physicist Emilio Segrè was one of the first scientists to isolate a sample of pure astatine. He was able to do so by using a particle accelerator: this is a machine that smashes together atoms and then studies the results.

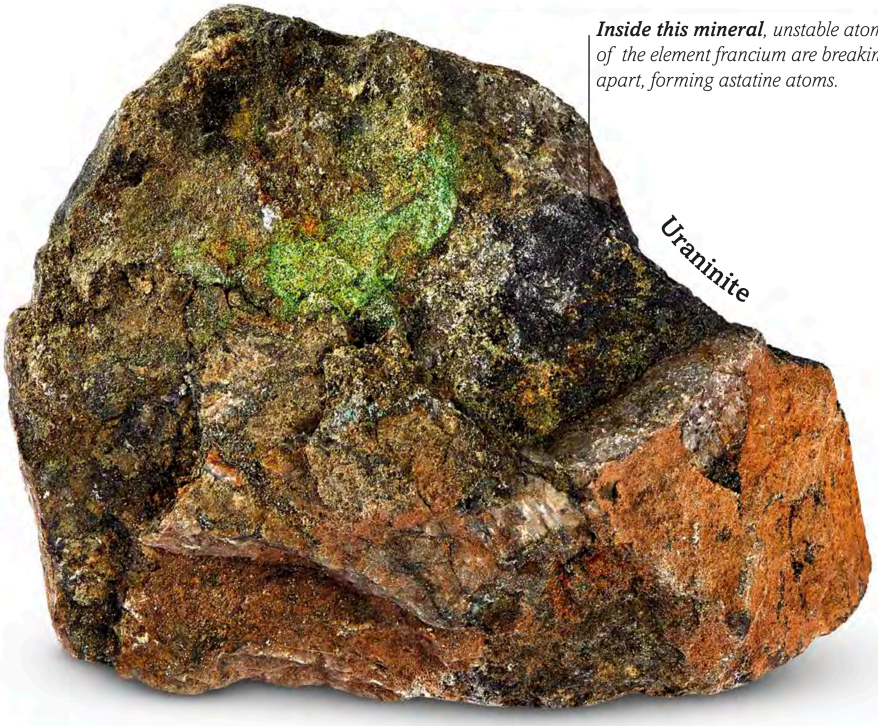

***Inside this mineral**, unstable atoms of the element francium are breaking apart, forming astatine atoms.*

117

Ts

Tennessine

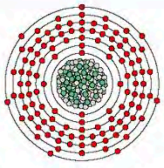

⊖ 117 ⊕ 117 ○ 177

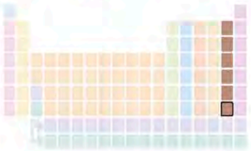

State: Solid
Discovery: 2011 (predicted)

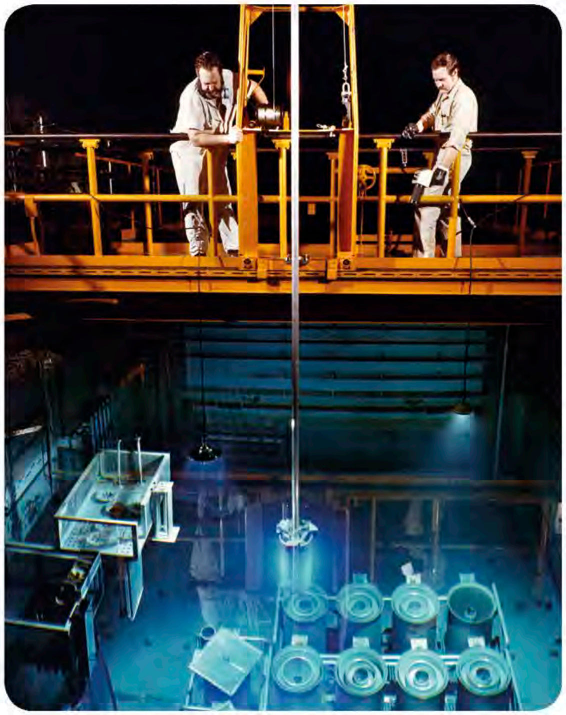

Nuclear Reactor, Oak Ridge National Laboratory, Tennessee, USA

Atoms of tennessine **existed for a few seconds** after they were formed.

Tennessine is the youngest element in the periodic table. It was produced in 2011, in the Russian city of Dubna. The element was named after the US state of Tennessee, home to the **Oak Ridge National Laboratory**, which houses one of the first, large-scale nuclear reactors ever built. Only a few atoms of this halogen element have ever been made. Even so, scientists have predicted it to be a semi-metal, not a non-metal like all the other halogens.

Krypton (Kr) becomes visible only when it is electrified.

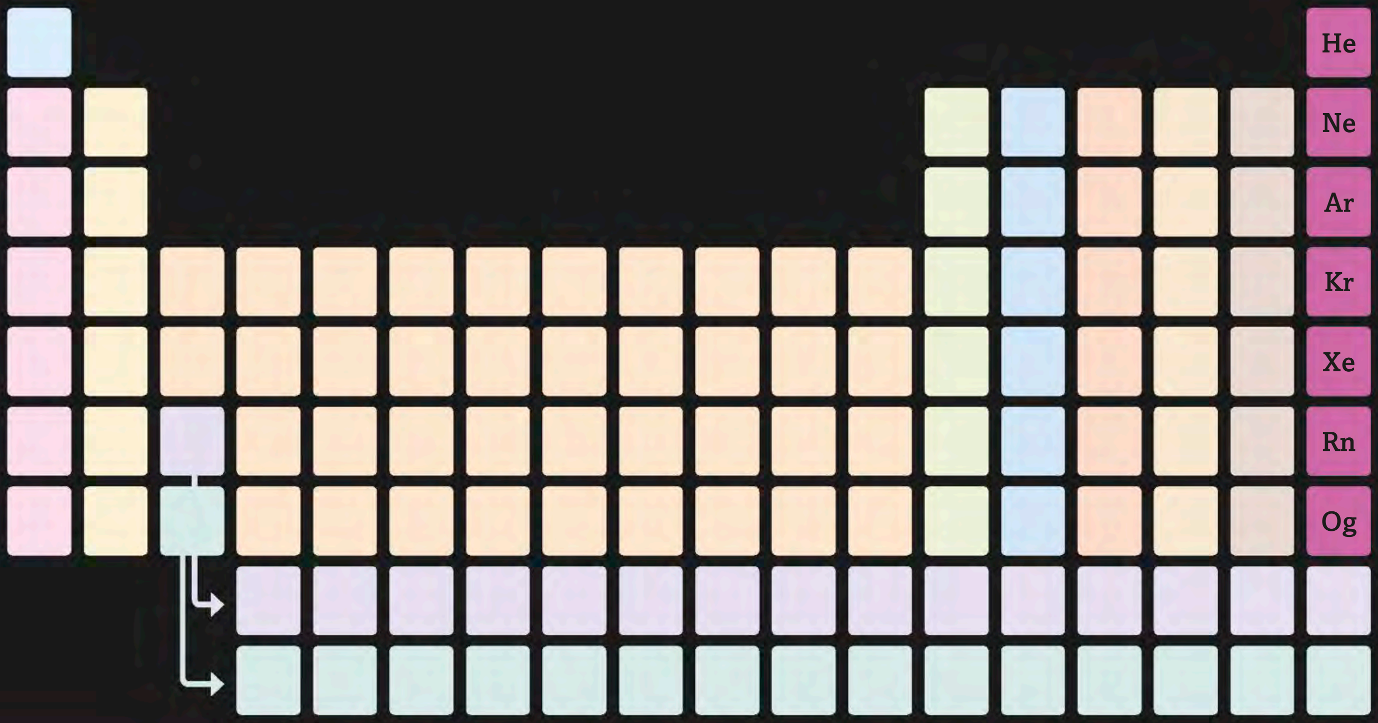

Noble Gases

The group on the far right of the periodic table belongs to the noble gases. These elements are described as "noble" because they do not react with the other "common" elements, such as oxygen (O). Their atoms never form bonds in nature, not even with atoms of their own kind, and so they are always gases at room temperature.

Atomic structure
Apart from a helium (He) atom that has two electrons in its outer shell, all other elements in this group have atoms with eight electrons.

Physical properties
All the members of this group are colourless gases. Going down the group, the density increases – radon (Rn) is 54 times denser than helium.

Chemical properties
Noble gases never react in nature. In the laboratory, heavier noble gases can be forced to form compounds with fluorine (F).

Compounds
These gases form no natural compounds. However, xenon (Xe), krypton (Kr), and argon (Ar) can be made to form compounds.

2 He Helium

⊖ 2 ⊕ 2 ○ 2

State: Gas
Discovery: 1895

Forms

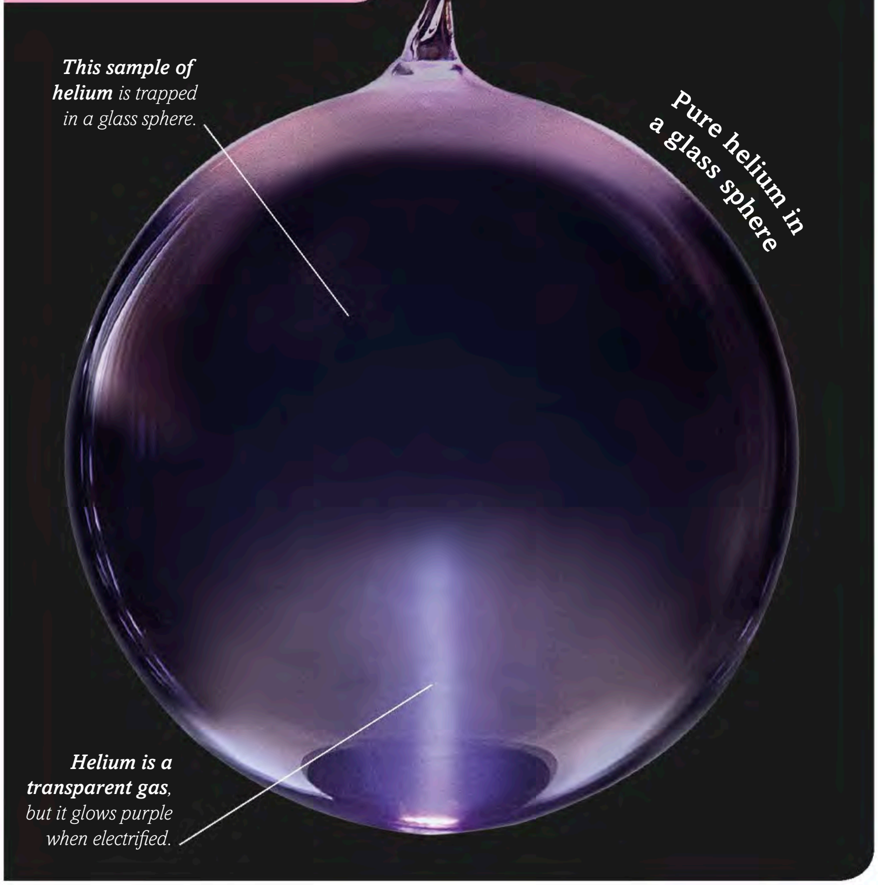

Pure helium in a glass sphere

This sample of helium *is trapped in a glass sphere.*

Helium is a transparent gas*, but it glows purple when electrified.*

Saturn

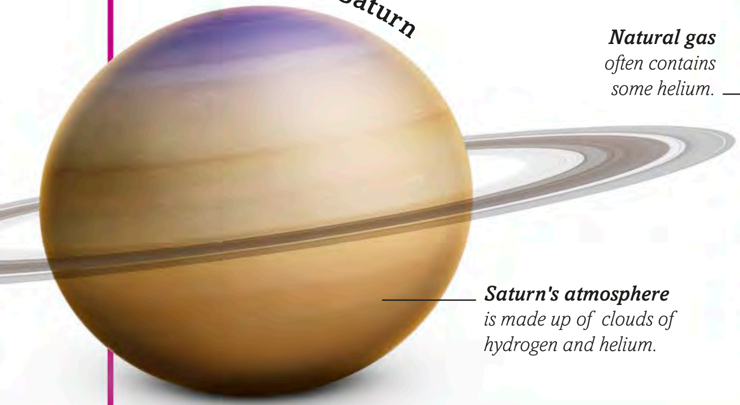

Saturn's atmosphere *is made up of clouds of hydrogen and helium.*

Natural gas *often contains some helium.*

Gas flare rig

Uses

A particle accelerator *is a machine that smashes atoms together: this one uses liquid helium to cool its parts.*

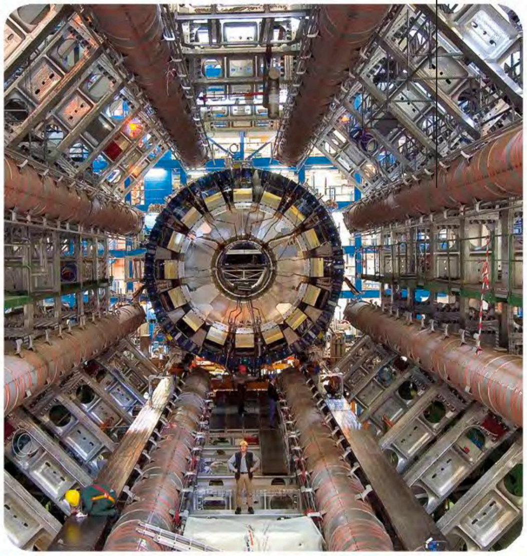

Large Hadron Collider, CERN, Switzerland

Helium-cooled MRI scanner

Helium is the second lightest element after hydrogen. This transparent gas was first discovered in 1868 by Sir William Ramsay, a Scottish chemist. Today, we know that a quarter of all the atoms in the Universe are helium. It is one of the main gases in the atmospheres of giant gas planets, such as **Saturn**. Being so light, however, helium is very rare on Earth: it escapes from our atmosphere into space. It was not until 1895 that chemists managed to collect a sample of helium gas coming from uraninite, a radioactive uranium mineral. Today, helium is

Airship

This airship *contains large quantities of helium to make itself lighter than air.*

Party balloon

This balloon *contains a mixture of helium and air.*

Helium-ion microscope

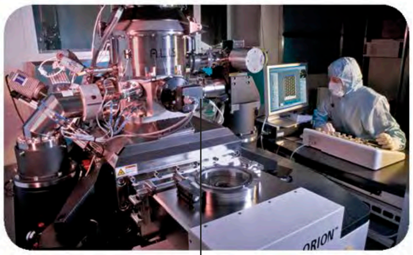

This powerful microscope *can zoom in to view much smaller details than most other microscopes.*

SUN GAS

In 1868, during a total solar eclipse (when the Moon passes directly in front of the Sun), helium was discovered in the cloud of gas seen around the Sun. The yellow colour of this cloud showed it contained an unknown gas, which was named after Helios, the Greek god of the Sun.

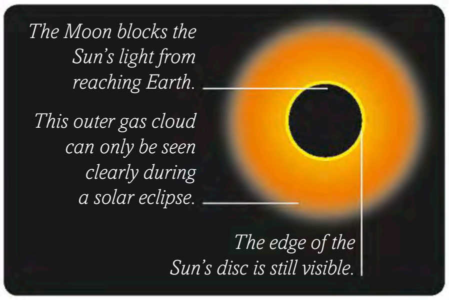

The Moon blocks the Sun's light from reaching Earth.

This outer gas cloud can only be seen clearly during a solar eclipse.

The edge of the Sun's disc is still visible.

This high-speed train *uses a pair of magnets: one to move forward and one to float over the track.*

Helium-cooled maglev train

The track is lined with a magnet *that repels the one on the train, making it float.*

Rocket helium tank

Helium in this container *will fill up a rocket's fuel tanks as they empty out during liftoff.*

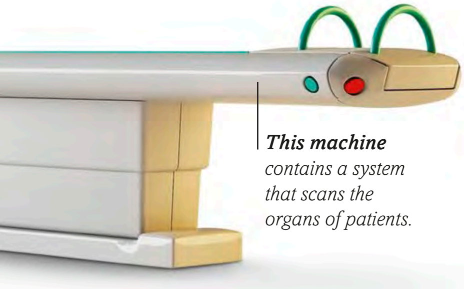

This machine *contains a system that scans the organs of patients.*

Helium has the **lowest melting point** of any element.

collected from underground reservoirs or is found mixed in natural **gas** and oil. Unlike hydrogen, which is very reactive, helium is a noble gas and does not react at all. This property makes it safe to use in objects such as **party balloons** and **airships**. To turn helium into a liquid, it must be cooled to an extreme temperature of –269°C (–452°F). Liquid helium is used to make things very cold, including the powerful magnets used to make **maglev trains** float along special tracks. **MRI scanners** also use liquid helium for cooling.

NEBULA This glowing nebula (cloud of gas and dust) is the Crescent Nebula. It is so vast that our entire Solar System would fit inside it seven times over. The nebula's light comes from a super-heated star at its centre. Known as WR 136, this star is 15 times heavier than our Sun and 250,000 times brighter. Its immense power comes from its fuel – helium.

Helium makes WR 136 hot and bright. The star once burned using hydrogen, like our Sun. Hydrogen atoms smashed together in the star's core until they became helium atoms, releasing energy in the process. However, the star ran out of hydrogen about 200,000 years ago. It began smashing together helium atoms instead, and ballooned into a gigantic red star, sending out a cloud of gas that spread around it. The star is producing a wind of electrified gases that hurtles out at 1,700 km (1,056 miles) every second. This wind continues to crash into the gas cloud, making it glow into the nebula we see. Eventually, WR 136 will run out of helium and its other fuels, and explode into an enormous fireball called a supernova.

10 **Ne** Neon

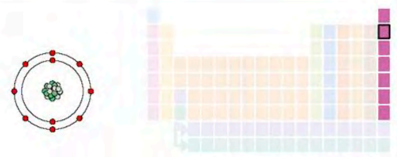

⊖ 10 ⊕ 10 ○ 10

State: Gas
Discovery: 1898

Forms

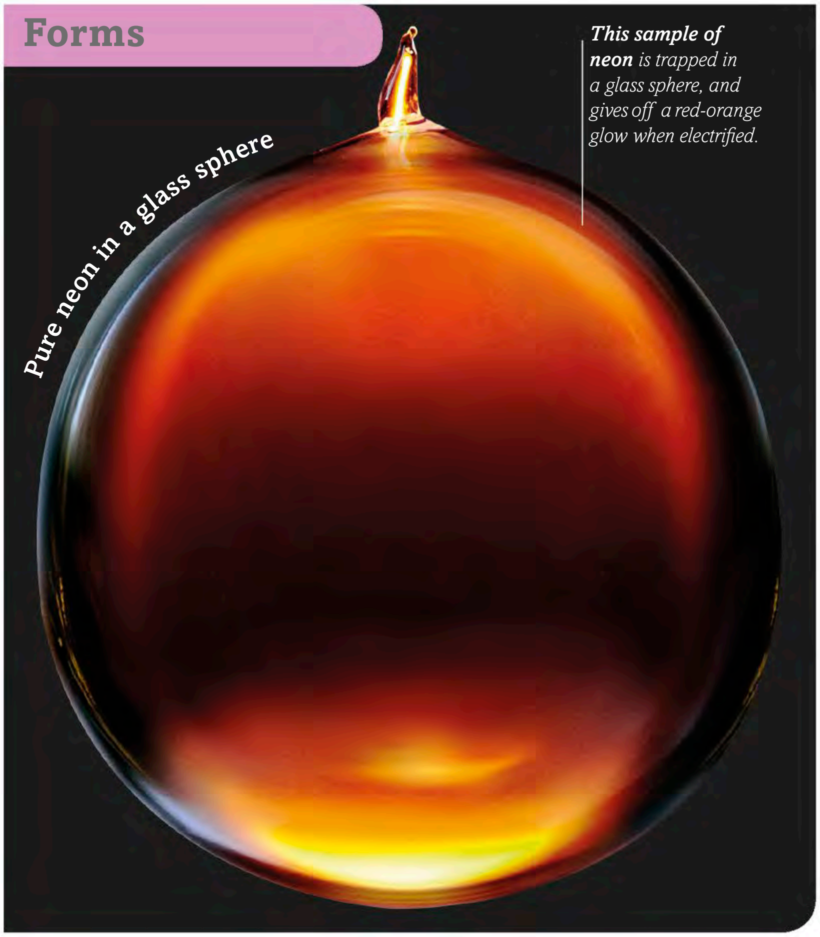

Pure neon in a glass sphere

This sample of ***neon*** *is trapped in a glass sphere, and gives off a red-orange glow when electrified.*

Volcanic eruption

Volcanic eruptions *release neon gas into the atmosphere.*

> **Neon lights** may refer to lighting produced using any noble gas.

Uses

A ***deep red beam*** *is emitted by this laser.*

Helium-neon laser

Glass tubes filled with neon *are used as brightly lit signs.*

Neon sign

Neon is a rare element: it makes up just 0.001 per cent of our atmosphere. Some of it was locked in Earth's rocks when the planet formed, and this is released into the air by **volcanic eruptions**. **Pure neon**, a transparent gas, is extracted by cooling air to a temperature of −189.34°C (−308.81°F), at which point the neon gas in the air turns to liquid. Neon can be mixed with helium to create research **lasers**. However, it is most commonly used in lighting, such as in **illuminated signs** or as bright warning beacons in the path of aircraft at airports.

18
Ar

Argon

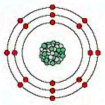

⊖ 18 ⊕ 18 ○ 22

State: Gas
Discovery: 1894

Forms

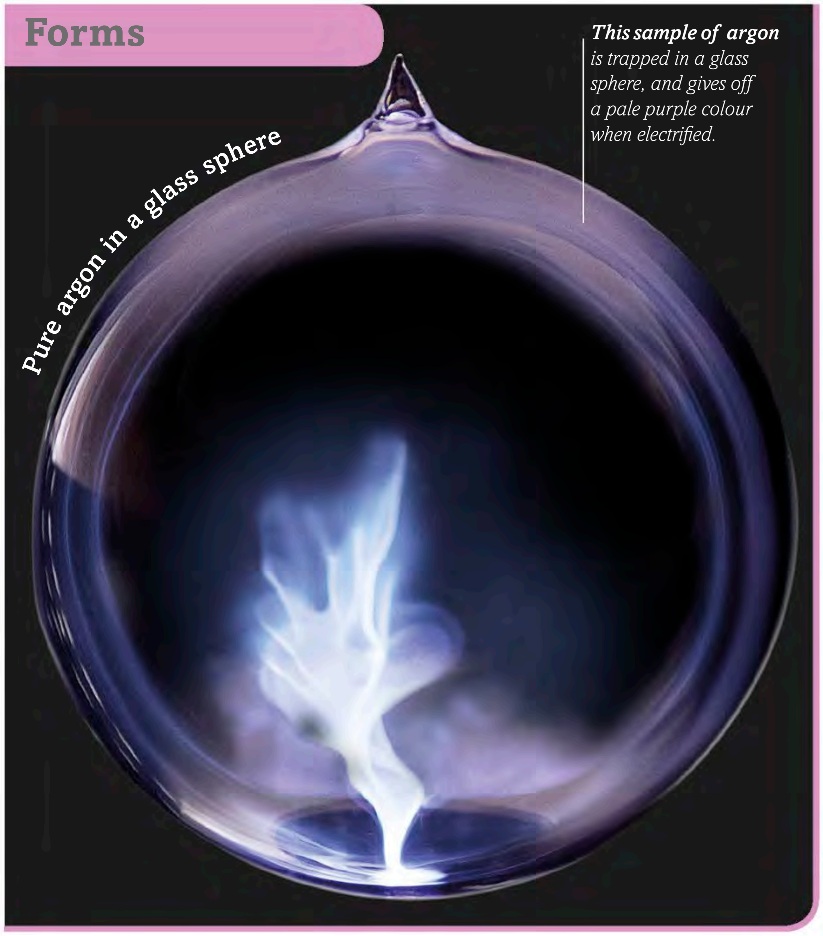

Pure argon in a glass sphere

This sample of argon *is trapped in a glass sphere, and gives off a pale purple colour when electrified.*

Argon-filled display

The Magna Carta*, a historical document, is stored in argon, which forces out oxygen and water vapour that would damage the parchment.*

Uses

Argon-filled suit

Some diving suits *are inflated with argon to retain heat in cold water.*

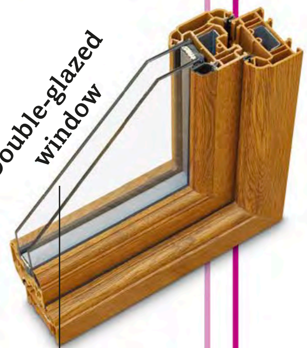

Double-glazed window

The space between the glass panes *is filled with argon to slow the loss of heat.*

Metal welding

Argon in this flame *prevents metals from reacting with oxygen*

Argon is the third most abundant gas in the atmosphere, after nitrogen and oxygen. It undergoes no reactions with any other element, and was named after the Greek word *argos*, meaning "idle". Argon does not conduct heat well so it is put in **double-glazed windows**, and in **diving suits** during cold, deep dives. Its lack of reactivity is useful. Argon is used in **museum displays** to protect delicate exhibits. It also stops metals reacting during hot **welding**. This element can also be useful in the production of titanium.

36 Kr Krypton

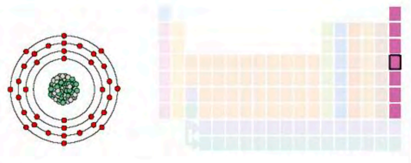

🔴 36 🟢 36 ⚪ 48

State: Gas
Discovery: 1898

Forms

Pure krypton in a glass sphere

*This sample of **krypton** is trapped in a glass sphere.*

*Krypton is a **transparent gas** but it glows with a blue-white light when electricity flows through it.*

*When this camera's **flash** activates, it's a result of krypton being electrified by the camera's battery.*

Digital camera

*Krypton-based **lasers** illuminate this building.*

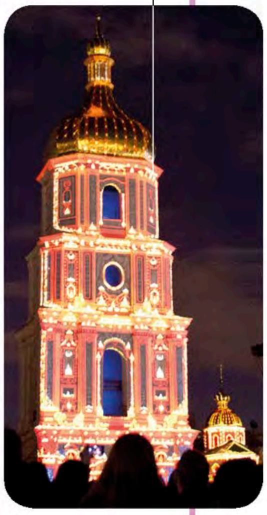

Laser lighting

Sir William Ramsey won a **Nobel Prize** in Chemistry for discovering the noble gases.

Uses

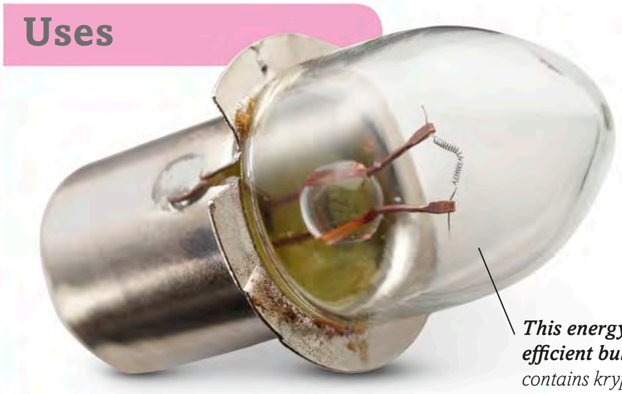

This energy-efficient bulb contains krypton.

Incandescent bulb

Plasma ball

This globe contains a mixture of noble gases, including krypton.

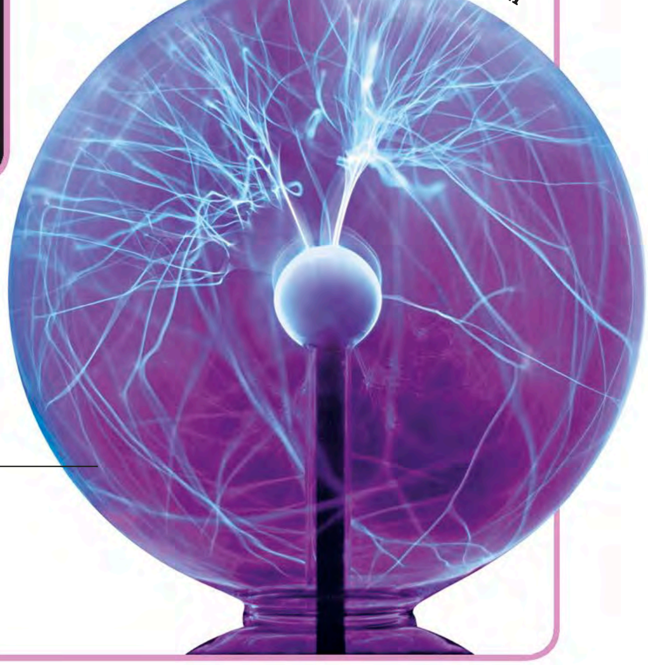

The word krypton means the "hidden one" in Greek. This element exists as an inert gas in nature, which means that it does not react with almost any other element. Krypton is not found in any minerals and only tiny amounts of it can be found in the air. **Pure krypton** produces a very bright white light when electrified with a current, which makes it ideal for use in **flash bulbs**. Krypton can react with the element fluorine to form the compound krypton fluoride, which is used to power some kinds of **laser**.

54
Xe

Xenon

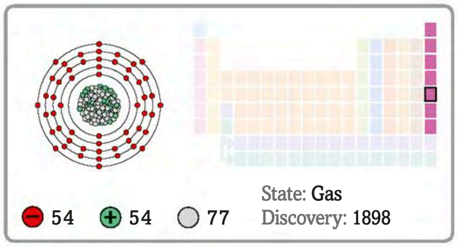

State: Gas
Discovery: 1898

Forms

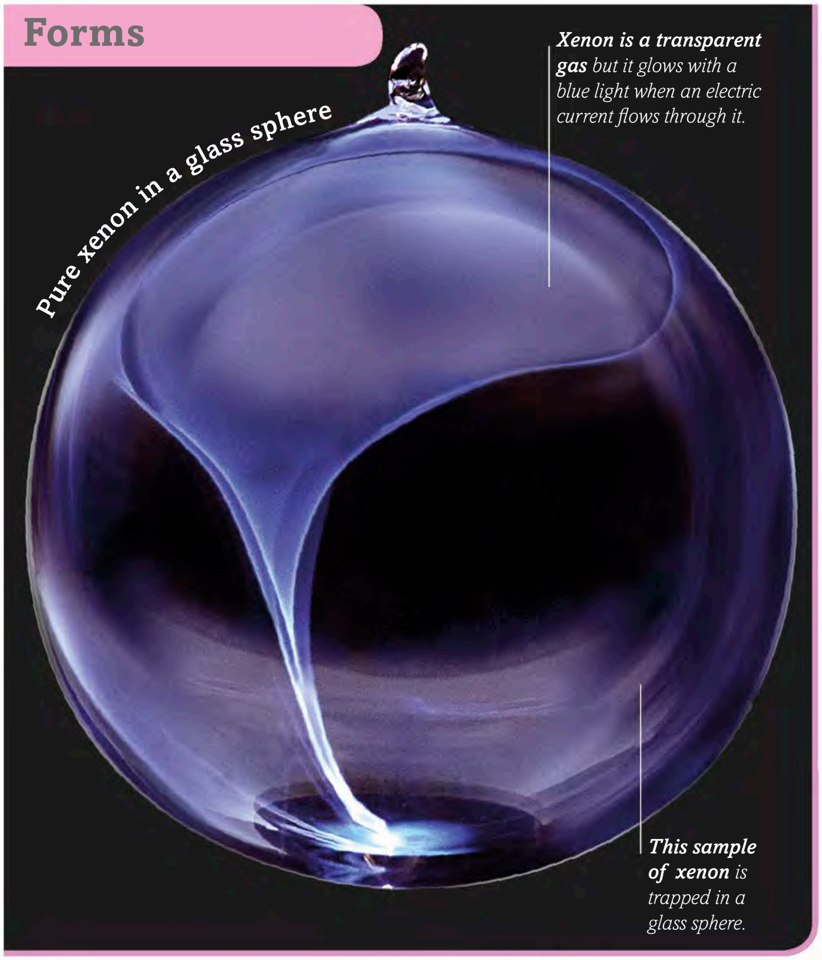

Pure xenon in a glass sphere

Xenon is a transparent gas *but it glows with a blue light when an electric current flows through it.*

This sample of xenon *is trapped in a glass sphere.*

Anaesthesia machine

The xenon produced by this device *makes a patient unconscious before an operation.*

Xenon is used in powerful **lasers** that can **kill bacteria**.

Uses

Car headlight

Xenon headlights *glow brighter than the more common halogen headlights.*

Xenon lamp

Xenon glow *disinfects the air in a room.*

High pressure inside this lamp *makes electrified xenon glow very brightly.*

Modern film projector lamp

Dawn space probe

Electrified xenon *blasts from the exhaust, pushing the probe forwards.*

Xenon is so rare that there is only one atom of this gaseous element for every 10 million atoms in the air. Like the other noble gases, xenon is colourless and odourless. It glows brightly when electrified, making it useful in very powerful lamps, such as those used in **film projectors** and **car headlights**. The gas is harmless when breathed in and can be used as an **anaesthetic**. When preparing food, **xenon lamps** can purify the air. To propel spacecraft, xenon is used in some rocket engines that produce streams of fast-moving, electrified atoms.

86 Rn Radon

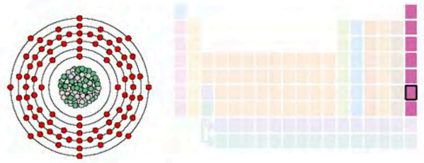

● 86 ⊕ 86 ○ 136

State: Gas
Discovery: 1900

This uranium mineral *releases radon gas, as radioactive metals in it break apart.*

These yellow crystals *belong to another uranium mineral called uranophane.*

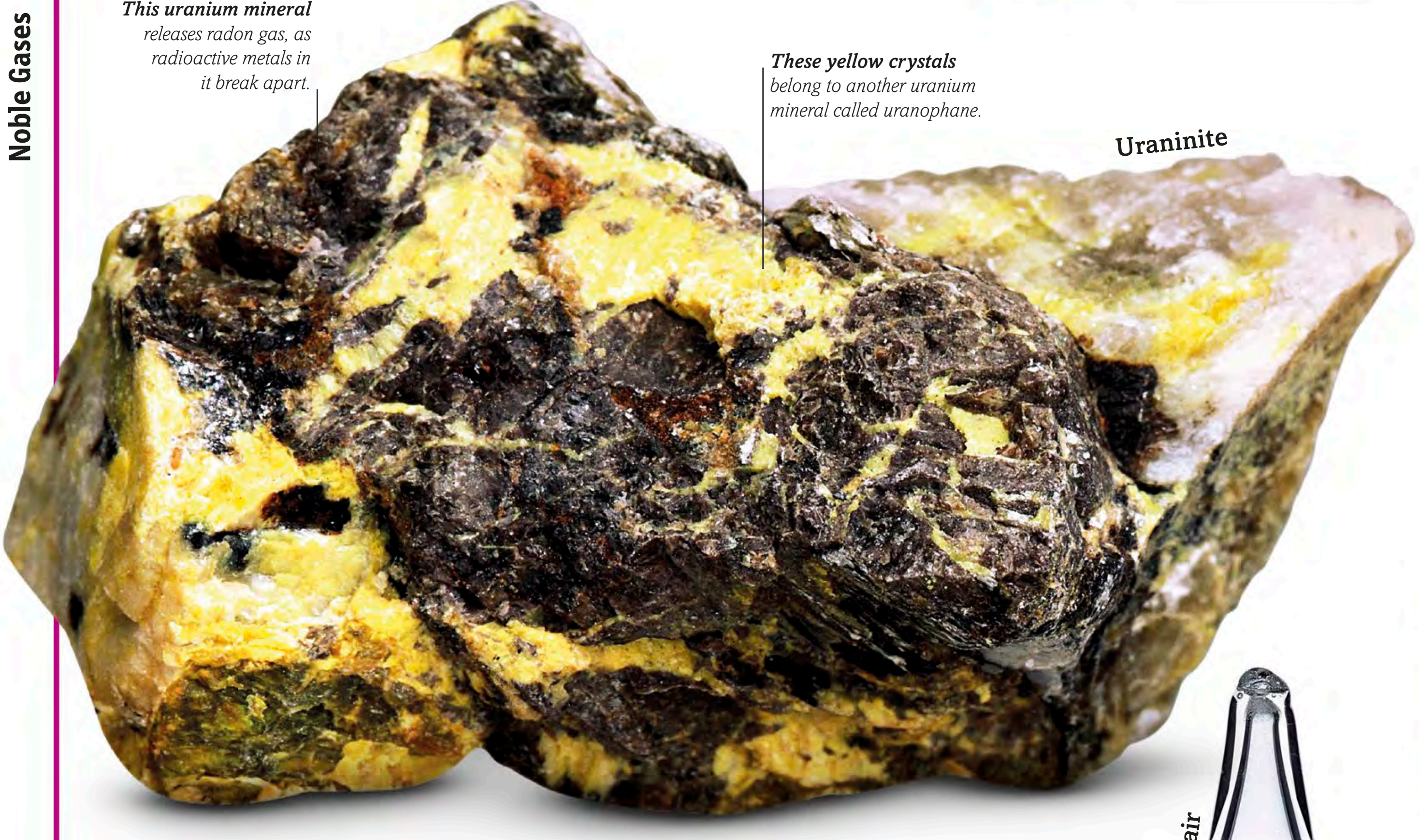

Uraninite

Radon is the only natural radioactive noble gas. This element is produced by the breakdown of uranium and other radioactive metals. Being a gas, radon escapes from minerals, such as **uraninite**, into the air. Radon is very radioactive and breathing it in can cause illness, such as lung cancer. In most places, the amount of radon in the air is incredibly tiny. However, its levels are high around **volcanic springs** and mud, where it bubbles out with other hot gases. Radon is also present in the water at **geothermal power plants**, which use heat energy from deep, volcanic rocks to make electricity. Radon is also more common in areas rich in granite rock. In these places, people use **test kits** to monitor their homes' radon levels.

It takes only **3.8 days** for half of radon's atoms to split into atoms of other elements.

Glass sphere containing radon and air

As it decays, *a compound called thorium dioxide emits radon.*

The muddy water *from volcanic springs contains radon.*

Hot springs in Lisvori, Greece

These pipes *draw water containing radon from deep under the ground, and this is then used to power the plant.*

Geothermal power plant

This kit *collects radon from the air so the amount of the gas in the area can be measured.*

Radon home test kit

RADO
Glenwo
3 Scienc
Glenwo
008807

118 Og Oganesson

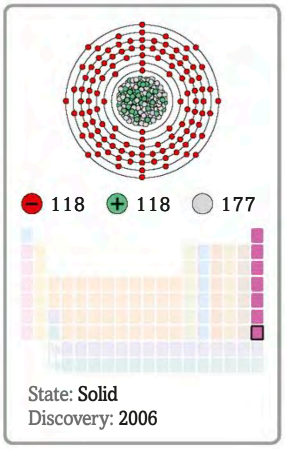

State: Solid
Discovery: 2006

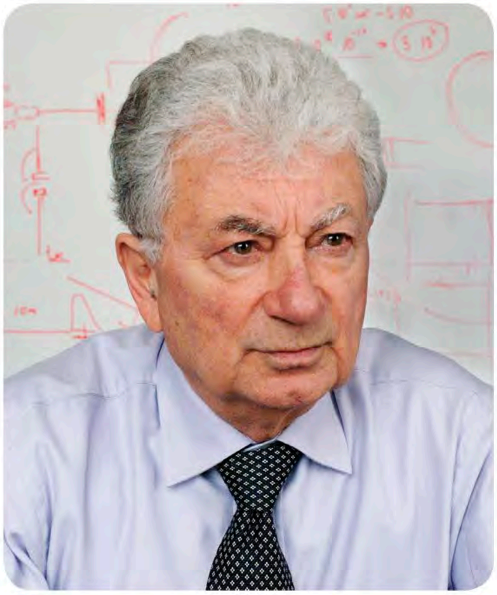

Yuri Oganessian

The heaviest element yet made is oganesson. Scientists think it would be solid at room temperature, but it may really be an unreactive, noble gas. However, only a few atoms have been created so far, so its properties are not well understood. Oganesson was first produced by a team of Russian and American scientists who smashed californium and calcium atoms together at the **Joint Institute for Nuclear Research** in Dubna, Russia. The element was named after **Yuri Oganessian**, the leader of the team.

Joint Institute For Nuclear Research, Russia

Glossary

Acid
A compound that contains hydrogen, and releases ions of hydrogen when it is dissolved in water. These ions make acids very reactive.

Actinide
A member of a group of radioactive metals with relatively large atoms.

Alchemist
People who experimented with chemicals before the science of chemistry began. Alchemists thought they could turn ordinary metals into gold.

Alkali
A compound that takes hydrogen ions from water molecules when it is dissolved in water. Alkalis react with acids.

Alkali metal
A group of metals that make alkalis when they react with water.

Alkaline earth metal
A group of metals that are found in nature, mainly in alkaline minerals.

Alloy
A material made by mixing a metal with tiny amounts of other metals or non-metals. Steel is a common alloy used in buildings and railways.

Artificial
Something that does not exist in nature. Several elements, including all that are heavier than uranium, are artificially produced by scientists in laboratories.

Atmosphere
The mixture of gases that surrounds a planet or moon. Earth's atmosphere is a mixture of nitrogen, oxygen, and argon, with tiny amounts of several other gases.

Atom
The smallest unit of an element. Atoms are composed of protons, neutrons, and electrons. The atoms of a particular element all have the same number of protons.

Atomic number
The number of protons in an atom of an element. Every element has a unique, unchanging atomic number.

Battery
A device containing chemicals that react to produce an electric current. There are two main kinds of battery: rechargeable and non-rechargeable.

Boiling point
The temperature at which a liquid gets hot enough to change into a gas.

Bond
The attraction between atoms that holds them together in an element or a compound.

Brittle
A way of describing a hard solid that shatters easily.

Carbonate
A compound that contains carbon and oxygen atoms, as well as atoms of other elements. Many minerals are carbonates.

Catalyst
A substance that speeds up a chemical reaction.

Chemist
A scientist who studies the elements, their compounds, and chemical reactions.

These vanadinite **crystals** contain the element vanadium.

Chemistry
The scientific study of the properties and reactions of the elements.

Chemical
Another word for a substance, generally meaning a compound made from several elements.

Combustion
A chemical reaction, involving oxygen, in which a fuel burns to produce heat and light in the form of flames.

Flames are the result of **combustion**.

Compound
A substance made of the atoms of one or more elements that are bonded together in a particular combination.

Conductor
A substance that lets heat or electricity flow easily through it.

Contract
To make or become smaller. Solids, liquids, and gases usually contract when the temperature goes down.

Corrosion
A chemical reaction that attacks a metal, or other solid object, usually due to the presence of oxygen and water.

Crystal
A naturally occurring solid substance whose atoms are arranged in a regular three dimensions pattern.

Decay
A process in which unstable atoms of radioactive elements break apart. The atoms of one element are transformed into those of another element during decay.

Density
The amount of matter held within a known volume of a material.

Dissolve
To become completely mixed into another substance. In most cases, a solid, such as salt, dissolves in a liquid, such as water.

Earthenware
A basic form of pottery, in which clay is heated to make it form hard structures.

Electrolysis
A process in which a compound is split into simpler substances using an electric current. Many elements, such as aluminium, are purified from their compounds in ores in this way.

Electrode
An electrical contact in an electric circuit. Electrodes can have a positive or negative charge.

Electron
A negatively charged particle inside an atom. Electrons orbit the atom's nucleus (or core) in layers called shells. They are also exchanged or shared by atoms to make bonds that hold molecules together.

Element
A pure substance that cannot be broken down into any simpler substances. Elements are the building blocks of matter. There are 118 known elements on Earth.

This **element** calcium is in crystal form.

Expand
To make or become bigger. Solids, liquids, and gases usually expand when the temperature rises.

Fission
A process in which the nucleus of an unstable atom splits in two when it collides with a neutron. It releases more neutrons that start the cycle again, and this continues as a chain reaction. Many unstable nuclei undergo fission spontaneously, without being hit by a neutron. Fission releases a lot of energy. The process is used to generate electricity in nuclear power plants, and it can trigger atom bomb explosions.

Flammable
A way of describing a material that can catch fire easily.

Fusion
A process in which small atoms, such as those of hydrogen, are fused together with such force that they merge to form larger atoms, while releasing a lot of energy. The Sun is powered by the fusion of hydrogen atoms into helium in its central core.

Gas
A state in which the particles of matter (atoms or molecules) are not attached to each other, and move freely. A gas can flow, take any shape, and fill any container.

Group
A set of elements in a column in the periodic table. Elements in a group have similar properties because each atom has the same number of electrons in its outer shell.

Halogen
A member of a group near the far right of the periodic table. Halogens form salts with metals. They are reactive, non-metallic elements.

Hardness
A measure of how easy it is to scratch or cut a substance with another substance.

Hydroxide
A type of compound containing hydrogen and oxygen, and normally a metallic element.

This is a chunk of ytterbium, a **lanthanide**.

Ion
An atom or a group of atoms that have an electric charge. While atoms have no overall charge, they become positive ions if they lose electrons or negative ions if they gain electrons.

Insulator
A substance that does not let heat or electricity flow easily through it.

Isotope
A form of an element with atoms that have the same number of protons but a different number of neutrons.

Lanthanide
A member of a set of metals with relatively large atoms. The elements in this series sit with the actinides, below the main part of the periodic table.

Laser
A beam of light with a single wavelength, in which the waves are all perfectly in step. Lasers are used in electronics and surgery.

Milk is a type of **mixture**.

LED
Short for light-emitting diode. An LED is a device that produces light when an electric current passes through it. The colour of its light depends on the compounds used in it.

Liquid
A state in which the particles of matter (atoms or molecules) are only loosely attached to each other, and move freely. A liquid can flow and take any shape, but has a fixed volume.

Mass
The amount of matter in a substance.

Matter
The material that makes up everything around us.

Maglev
Short for magnetic levitation. This refers to some kinds of high-speed train that use magnets to propel themselves while hovering over a track.

Magnet
A solid object that produces a magnetic field, which attracts certain materials to it and can attract or repel other magnets.

Magnetic
Relating to a magnet.

Magnetic field
The force field around a magnet.

Melting point
The temperature at which a solid gets hot enough to turn into a liquid.

Metal
A type of element that is likely to react by giving away the outermost electrons in its atoms. Most elements are metals, and they tend to be hard, shiny solids. Mercury is the only metal that is liquid at room temperature.

Mineral
A naturally occurring solid compound – or mixture of compounds – made up of different elements. Every mineral has particular characteristics, such as crystal shape and hardness. Minerals are mixed together to make the rocks in Earth's crust.

Mixture
A collection of substances that fill the same space but are not connected by chemical bonds. Examples of mixtures are seawater, milk, and mud. The contents of a mixture can be separated by a physical process, such as filtration.

Molecule
A single particle of a compound. Its two or more atoms are bonded together.

Neutron
A neutral particle in the nucleus of an atom. A neutron is about the same size as a proton but it does not have an electric charge.

Noble gases
A group of elements that are unreactive and generally form no compounds with the other elements. This is because the outermost shells in their atoms are filled with electrons. This group sits on the extreme right of the periodic table.

Non-metal
A type of element that is likely to react by acquiring electrons in the outermost shell of its atoms. Non-metals are usually crystalline solids, such as sulfur, or gases, such as oxygen. Bromine is the only non-metal that is liquid at room temperature.

Nucleus
An atom's core, which contains its protons and neutrons. Nearly all the mass of an atom is packed into its nucleus.

Ore
A rock or mineral from which a useful element such as a metal can be purified and isolated.

Oxide
A compound in which oxygen is bound to one or more other elements.

Particle
A basic unit of which substances are made. Sub-atomic particles are units of which atoms are made including protons, neutrons, electrons, and many other smaller ones.

Particle accelerator
A machine in which atoms or sub-atomic particles are made to collide at high speeds. These collisions are then studied by scientists. Particle accelerators are used to produce artificial elements as well as study particles smaller than atoms. A cyclotron is a type of particle accelerator.

These petal-like shapes may form when desert sand mixes with barite, an **ore** of barium.

Period
A set of elements in a row in the periodic table. Each atom of the elements in the first period have one electron shell. Atoms in the second period have two electron shells.

Periodic table
A table that identifies and classifies all known elements.

Photograph negative
A film or plate that has been exposed to light to show an image with reversed colours.

Photosynthesis
A complicated set of chemical reactions by which plants make their food. They use the energy of sunlight to convert water and carbon dioxide into sugar and oxygen.

Pollution
Harmful substances that are released into the environment. Pollution can be chemical – a gas, liquid, or solid added to the air, water, or soil.

Pressure
A measure of how hard a force pushes on a surface. Pressure depends upon the strength of the force and the area of the surface.

Proton
A positively charged particle in the nucleus of an atom. Protons attract electrons, and they circle the nucleus.

Radar
A system that detects the positions and speed of distant objects, such as aircraft.

Radiation
Energy released by atoms as light, infrared, ultraviolet (UV), and X-rays. The term "radiation" is also used to describe the rays given out by radioactive substances.

Radioactive
Describes a substance that contains unstable atoms, in which the nucleus breaks apart, or decays. When a nucleus decays, it releases at least one particle and its atomic number changes, so that the atom becomes a different element.

Reaction
A chemical process in which atoms or molecules form bonds with each other to form a new compound – or rearrange, to produce different compounds.

Renewable energy
A source of energy that will not run out, for example, wind.

Rust
The common name for the compounds that form when iron reacts with oxygen and water.

Salt
A compound that forms when an acid reacts with an alkali. Sodium chloride is the most familiar example of a salt.

Semi-metal
An element that has properties of both metals and non-metals.

Smelting
A chemical process that involves great heat in which a metal is extracted from its ore.

Solid
A state in which the particles of matter (atoms or molecules) are bound to each other, so they remain in fixed positions. A solid has a fixed shape and volume.

Stainless steel
An alloy of iron and carbon that also contains other metals, such as chromium, which stop the iron from rusting and increase its toughness.

Superconductor
A substance through which an electric current flows with no resistance; most substances resist the flow of electricity, and produce heat when current flows.

Synthetic
Another word for artificial. Out of the 118 known elements on Earth, more than 25 are synthetically made.

Toughness
A quality of a solid that shows how difficult it is to break. Steel is very tough – it may bend or twist, but is much harder to break.

Transition metal
A member of the set of metals that sit in the middle of the periodic table. Most metals belong to this set.

Toxic
Another word for poisonous or harmful.

Transparent
Another word for see-through. Glass, water, and air are transparent to light. Many materials are transparent to other forms of radiation.

Vacuum
An empty space that does not contain air or any other substance.

Vapour
A gas that can easily be changed back to a liquid, by cooling it or putting it under pressure.

Verdigris
The green-grey layer that forms on copper when it comes in contact with air.

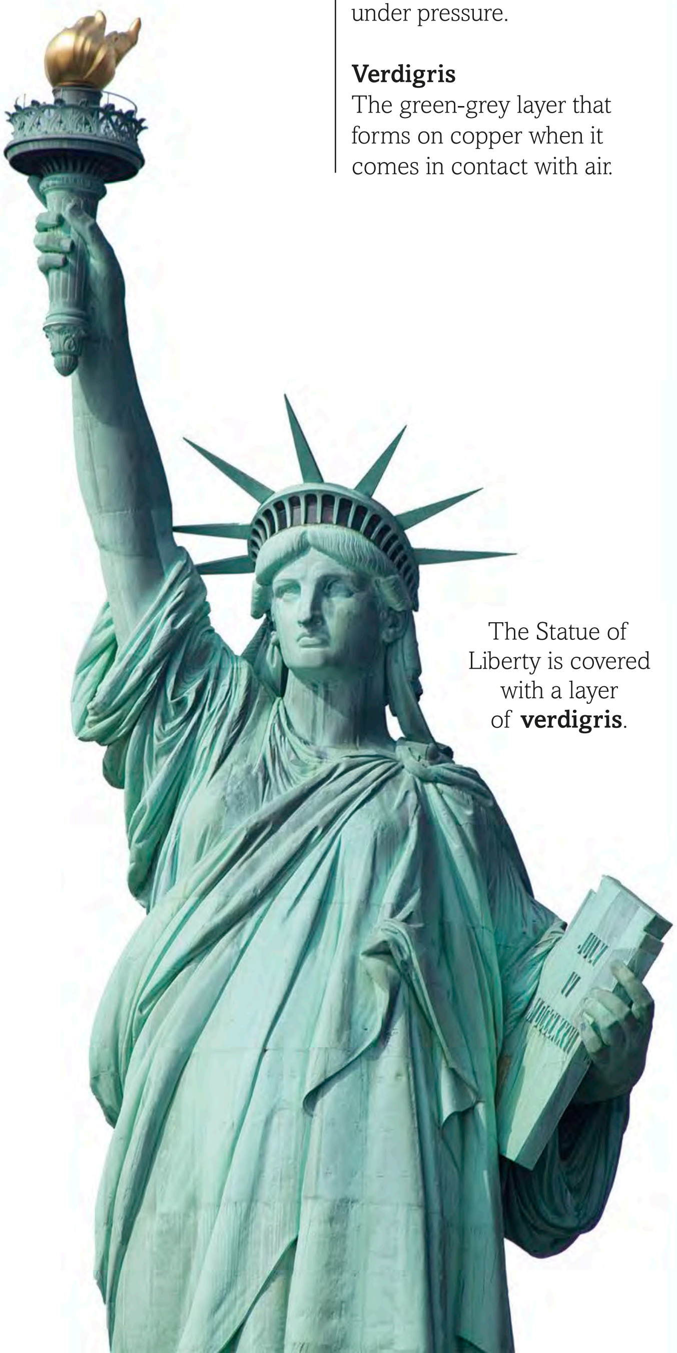

The Statue of Liberty is covered with a layer of **verdigris**.

Index

Page numbers in **bold** indicate a main entry.

A

acid rain 169
actinides 119
actinium 119, **120**, 121
aerogel 147
aeschynite 122
aircraft 54, 133, 134–135
airships 191
alchemy 9
alkali metals 23
alkaline earth metals 37
aluminium **132–133**, 134–135
Alvarez, Luis Walter 93
americium **123**
amethyst 146
anaesthesia machines 197
Andes Mountains, Peru 28–29
anglesite 150
antimony **161**
apatite 159
aphthitalite 30
aquamarine 38
aragonite 42
argenite 84
argon **195**
Armbruster, Peter 104
arsenic **160**
artificial joints 55, 64, 88
astatine **187**
astronauts 97
Atlas V (rocket) 167
atom bombs 174
atomic clocks 34
atoms 12–13

B

Badlands National Park, USA 92
Balard, Antoine–Jérôme 185
balloons 21, 191
bank notes 113
barite 48
barium 36, **48–49**
barometers 101
bastnasite 110
batteries 25, 59, 86
battery acid 169
bauxite 132
benitoite 48
berkelium **124**
beryllium **38–39**
Berzelius, Jacob 11
bicycles 78, 143
bismuth 152, **162–163**, 187
bleach 181
body armour 55
Bohr, Niels 104
bohrium **104**
bombs 21, 174
Boric acid 131
bornite 68
boron 129, **130–131**
Boyle, Robert 9
braggite 82
Brand, Hennig 158
brimstone 168
bromine **184–185**
brookite 55
Buddha 98
Bunsen, Robert 11, 34
buses 21

C

cadmium **86**
caesium 11, **34**
calaverite 96
calcite 42
calcium **42–43**, 44–45, 199
californium **125**, 199
cameras 40, 110, 148
cancer 86
cans 149
carbon 140–141, **142–143**, 145
carbon monoxide testers 82
cars 25, 79, 197
cassiterite 149
catalytic converters 82, 83
caves 43, 59
celestine 168
ceramics 47, 172, 179
cerium 11, **111**
chalcopyrite 68
chalk 43
Chandra X-ray Observatory (CXO) 93
chemical reactions 16–17
china 159
chlorine **180–181**, 182–183
chloroform 181
chlorophyll 41
chromite 57
chromium **57**
chrysoberyl 38
cinnabar 100
circuit boards 80, 85, 147
clay 131
Cleve, Per Teodor 115
clocks 34
clouds 85
coal 142
cobalt 52, **64–65**
coins 67, 83, 84
colloids 17
columbite 77
combustion 166
compasses 93
compounds 17
conductors 68, 75, 133
copernicium **107**
Copernicus, Nicolaus 107
copper **68–69**, 70–71
cosmetics 51, 162, 163
covalent bonds 17
crocoite 57, 150
Crookes, William 138
cryolite 178
cubic zirconia 115
Curie, Marie 51, 124, 174
Curie, Pierre 51, 174
curium **124**
cutlery 85
cyclotron particle accelerator 122, 127

D

Dalton, John 11
Damascus Steel 56
Danakil Depression 170–171
darmstadtium **106**
dating (rock) 33
Davy, Humphry 10, 46, 48
Dead Sea 184
death masks 97

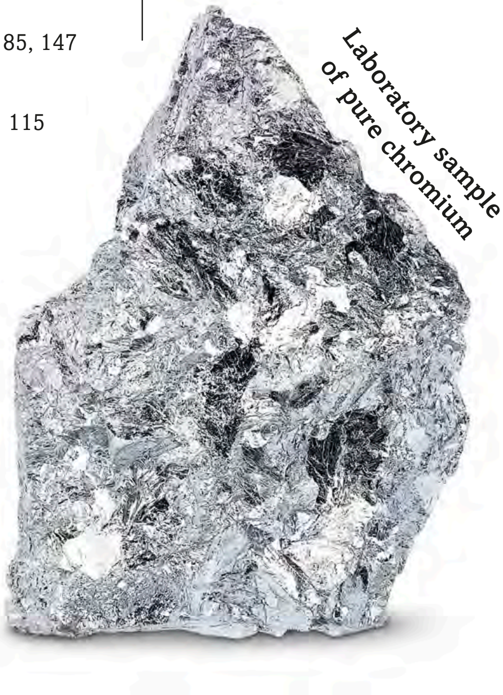
Laboratory sample of pure chromium

Diamond

Death Valley, USA 130
dental equipment 25, 76, 94
diamonds 142, 144–145
diaspore 136
diving clothing 183
DNA 159
dolomite 40
drag racing 155, 156–157
drill bits 89
drugs *see* medicines
dubnium **102**
dyes 155
dysprosium **115**

E

Earth's atmosphere 9
Einstein, Albert 125
einsteinium **125**
electrical equipment 68, 70–71, 80
electricity 47, 120
electrolysis 10
electromagnetism 13, 68
electrons 12
electronic devices 25
electroplating 69
engines 167
erbium **116**
Esplanade Theatre, Singapore 133
europium 108, **113**
euxenite 54
explosions 16
explosives 154

F

Fermi, Enrico 126
fermium **126**
fertilizers 155, 159
fibreglass 81
fibre optics 173
film projectors 197
fingerprinting 91
fire 167
fire protection suits 132–133, 185
fireworks 27, 41, 47
flares 47
Flerov, Georgy 151
flerovium **151**
flowers 168
fluorine **178–179**
fluorite 178
Fly Geyser, USA 44–45
francium **35**
fuel cell 95

G

Gadolin, Johan 74, 114
gadolinite 54
gadolinium **114**
Gahn, Johan Gottlieb 59
galena 150
gallium 128, **136**
galvanised steel 73, *73*
garnierite 66
gases 8, 11, 20, 189
Gay-Lussac, Joseph Louis 131
geiger counters 121
geothermal power plants 198
germanite 148
germanium **148**
Ghiorso, Albert 102
glass 31, 59, 112, 178
glucometer tests 83
glues 155
gold **96–97**, 98–99
Golden Buddha, Vietnam 98
graphite 142
greenockite 86
groups 14, 15
guitars 67
gunpowder 31

H

hafnium **87**
Hahn, Otto 105
halite 26, 180
halogens 177
hassium **104**
Heavy Ion Research Centre, Germany 104, 106
helicopters 38–39
helium **190–191**, 192–193
hemimorphite 72
Hofmann, Sigurd 106
holmium **115**
Holtermann Nugget 97
Hope, Thomas Charles 46
hopper crystals 162
hot springs 170–171
human body, elements in 9
hydrogen 18–19, **20–21**, 157
hydrogen bomb test 125

I

illuminated signs 194
indium **137**
ink 186
insulators 33
iodine 176, **186**
ionic bonds 17
iridium **92–93**
iron **60–61**
iron ore 62
isotopes 13, 65

J

jet engines 64, 134–135
jewellery 81, 94, 97
Joint Institute for Nuclear Research, Russia 151, 175, 199
Jupiter 20

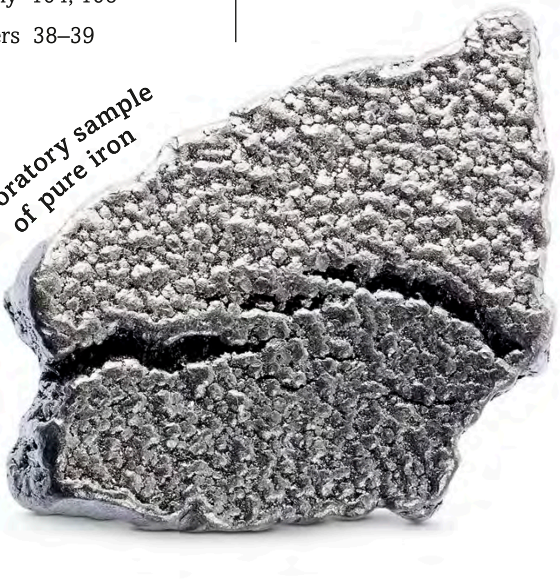

Laboratory sample of pure iron

K

kernite 130
Kirchhoff, Gustav 34
kohl 161
krypton 188, **196**

L

Lamy, Claude-Auguste 138
lanthanides 109
lanthanum 109, **110**
Large Hadron Collider, Switzerland 190
lasers 75, 117, 194, 196
Lavoisier, Antoine 10
Lawrence Livermore National Laboratory, USA 175
Lawrence, Ernest 127
lawrencium **127**
LCD screens 131
lead **150–151**
LED 74–75, 136
lepidolite 24, 32
leucite 32
life cycles 155
light bulbs 89, 101, 114
liquids 8, 11
lithium **24–25**
livermorium 165, **175**
lutetium **117**

M

magnesium 10, **40–41**
magnets 65, 67
malachite 68
manganese **58–59**
Marggraf, Andreas 72
medical imaging 33, 79, 114, 116
medical procedures 49
medicines 41, 95, 163, 169
Meitner, Lise 105
meitnerium **105**
melonite 173
melting points 89
Mendeleev, Dmitri 15, 126, 148
mendelevium 125, **126**
mercury **100–101**
meteorites 60, 92
microchips 87, 148
microscopes 81, 86, 91
military tanks 131
millerite 81
mining 29, 96, 130
missiles 112
mixtures 17
Moissan, Henri 179
molybdenite 78, 90
molybdenum **78–79**
monazite 74
Morita, Kozuka 139
moscovium 139, **163**
motorcycles 57
MRI scanners 191
mummification 27
musical instruments 67, 69, 83, 113

N

NASA (National Aeronautics and Space Administration) 74, 77, 93
nebula 21, 192–193
neodymium **112**
neon **194**
neptunium **122**
neutrons 13
Nevada Desert, USA 44–45
nickel **66–67**, 178
nickeline 66
night-vision equipment 33
nihonium **139**
niobium **77**
nitratine 154
nitrogen 152–153, **154–155**, 157
nitroglycerine 154
nitromethane 156
Nobel Prize in Chemistry 103

Nickel balls

Nobel, Alfred 127
nobelium **127**
noble gases 189
nuclear fuel 123
nuclear reactors 187
nucleus 13

O

Oak Ridge National Laboratory, USA 187
oceans 182–183
Oganessian, Yuri 163, 199
oganesson **199**
oil 117, 143
onions 168
Orion Nebula 21
orpiment 160
osmiridium 91
osmium **91**
oxygen 164–165, **166–167**, 189

P

paints 50, 65, 86, 112
palladium **82–83**
particle accelerators 103, 104, 106, 125, 127, 151, 175, 187, 190
pencils 143
pens 83
pentlandite 66, 80
Perey, Marguerite 35
periods 14, 15
perovskite 55
pesticides 159
petalite 24
petrol 59
pewter 149
Phoenix lander 155
phosphorus **158–159**
photocopy machines 172
photography 85, 94, 185
pink diamonds 144–145
pipes 150
plants 167
Plaster of Paris 43
platinum **94–95**
plutonium **123**
pnictogens 153
poisons 160
pollucite 32, 34
polonium **174**
polythene 143
pools 28–29, 44–45
potash 30
potassium 22, **30–31**
potassium bromide 184
pottery 49, 116
praseodymium **111**
printing presses 161
promethium **112**
protactinium **121**
protective clothing 183, 195
proton 13
PVC 181
pyrargyrite 84
pyrite 60, 138
pyrolusite 58

Q-R

quartz 146
radium **50–51**
radon 50, **198–199**
railway tracks 59
Ramsay, Sir William 190
reactions, chemical 16–17
rebreather 31
rhenium **90**

rhodium **81**
rhodochrosite 58
RIKEN Nuclear Research Centre, Japan 139
rock salt 29
rockets 21, 167
Röntgen, Wilhelm 106
roentgenium **106**
roofing materials 151
Roscoe, Henry 56
rovers (spacecraft) 136, 174
rubidium **32–33**
rust 17
Rutherford, Ernest 13, 102
rutherfordium **102**
ruthernium **80**

S

salt 26–27, 28–29
samarium **113**
sand 146
Saturn 190
scandium **54**
Seaborg, Glenn T 103, 124
seaborgium **103**
sealife 186
Segrè, Emilio 187
selenium **172**
serpentine 40
shells (atom) 12
shells (sea) 42
Sikkeland, Torbørn 127
silicon **146–147**
silver **84–85**
silver nitrate 85
skunks 168
smelting 61, 63
smithsonite 73
smoke detectors 123
sodalite 26
sodium **26–27**
sodium chloride 26, 177, 180
sodium hypochlorite 183
solar cells 172
solids 8, 11
solutions (mixture) 17
space probes 124, 155
spacecraft 21, 77, 174
spark plugs 49
spectacles 139
sperrylite 94
sphalerite 72, 137
stainless steel 57, 85
Statue of Liberty 69
steel 57, 85, 167
steelmaking 62–63
stibnite 161
strontianite 46
strontium **46–47**
suitcases 181
sulfur 10, 48, **168–169**, 170–171
Sun 20
sunglasses 85
sunscreen 55
Super Heavy Ion Linear Accelerator, USA 103
supernova 193
suspensions (mixture) 17
swimming pools 181
sylvite 30
symbols, elements 15

T

tantalite 88
tantalum **88**
technetium **79**
telescopes 39
televisions 70, 111
tellurium 164, **173**
tennessine **187**
terbium **114**
thallium **138–139**
Thénard, Louis Jacques 131
thermometers 101, 136
thorite 35
thorium 50, **120**
thulium **116**
tin **149**
titanium **55**
TNT 154
toasters 67
tools 56, 61, 142–143
toothpaste 47, 179
touchscreens 137
tractors 61
trains 191
transition metals 53
tungsten **89**
turbine, jet 134–135
Tutankhamun 97
tyres 169

U

ulexite 130
Ulloa, Antonio de 95
University of California, USA 124
uraninite 35, 50, 187, 198
uranite 120
uranium 118, **122**

V

vacuum tubes 120
vanadinite 56
vanadium **56**
variscite 132
Vauquelin, Louis Nicolas 39
volcanic activity 45, 194, 198

W

Walton, John R. 127
Wat Phrathat Doi Suthep, Thailand 97
watches 50–51, 55, 88, 147
water 20, 23, 166
waterproof clothing 179
weather stations 47
welding 137, 195
Welsbach, Carl Auer von 112
windows 137, 195
Winkler, Clemens A 148
wires, copper 70–71
witherite 48
Wöhler, Friedrich 75
wolframite 89

X

xenon **197**
xenotime 74
X-rays 49, 90, 93, 116

Y

ytterbium **117**
yttrium **74–75**

Z

zeolite 26
zinc **72–73**
zircon 76, 87
zirconium **76**

Laboratory sample of pure zinc

ACKNOWLEDGEMENTS

The publisher would like to thank the following people for their help with making the book: Agnibesh Das, John Gillespie, Anita Kakkar, Sophie Parkes, Antara Raghavan, and Rupa Rao for editorial assistance; Revati Anand and Priyanka Bansal for design assistance; Vishal Bhatia for CTS assistance; Jeffrey E Post, Ph D Chairman, Department of Mineral Sciences Curator, National Gem and Mineral Collection, National Museum of Natural History, Smithsonian; Kealy Gordon and Ellen Nanney from the Smithsonian Institution; Ruth O'Rourke for proofreading; Elizabeth Wise for indexing; and RGB Research Ltd (periodictable.co.uk), especially Dr Max Whitby (Project Director), Dr Fiona Barclay (Business Development), Dr Ivan Timokhin (Senior Chemist), and Michal Miškolci (Production Chemist).

The publisher would like to thank the following for their kind permission to reproduce their photographs:

(Key: a-above; b-below/bottom; c-centre; f-far; l-left; r-right; t-top)

9 Bridgeman Images: Golestan Palace Library, Tehran, Iran (cra). **Fotolia:** Malbert (ca). **Getty Images:** Gallo Images Roots Rf Collection / Clinton Friedman (fcla). **Wellcome Images http://creativecommons.org/licenses/by/4.0/:** (crb). **13 Getty Images:** Stockbyte (cra). **Science Photo Library:** Mcgill University, Rutherford Museum / Emilio Segre Visual Archives / American Institute Of Physics (crb). **15 Science Photo Library:** Sputnik (crb). **17 Alamy Stock Photo:** Dennis H. Dame (cr). **20 Dreamstime.com:** Alekc79 (cb). **NASA:** X-ray: NASA / CXC / Univ.Potsdam / L.Oskinova et al; Optical: NASA / STScI; Infrared: NASA / JPL-Caltech (cra). **21 Alamy Stock Photo:** Phil Degginger (cb); ULA (fcr). **NASA:** Bill Rodman (cla). **Science Photo Library:** U.S. Navy (crb). **24 Alamy Stock Photo:** PjrStudio (crb). **Dreamstime.com:** Titovstudio (ca). **naturepl.com:** Christophe Courteau (cr). **25 123RF.com:** Federico Cimino (cr). **Dreamstime.com:** Aleksey Boldin (cla); Bolygomaki (ca). **Getty Images:** Corbis (cla/mirror); Driendl Group (cb). **NASA:** (crb). **Science Photo Library. 26 123RF.com:** Stellargems (cra); Sara Winter (crb). **Dorling Kindersley:** Tim Parmenter / Natural History Museum, London (cr). **27 123RF.com:** Todsaporn Bunmuen (cl); Francis Dean (crb). **Alamy Stock Photo:** Artspace (cr); Hemis (c). **Dreamstime.com:** Abel Tumik (clb). **28-29 Alamy Stock Photo:** Hemis. **30 Alamy Stock Photo:** Siim Sepp (c). **31 123RF.com:** Petkov (ca). **Alamy Stock Photo:** Doug Steley B (cb). **Dorling Kindersley:** Dave King / The Science Museum, London (clb). **Dreamstime.com:** Mohammed Anwarul Kabir Choudhury (cr); Jarp3 (crb). **Getty Images:** John B. Carnett (c). **Science Photo Library:** CLAIRE PAXTON & JACQUI FARROW (cl). **32 123RF.com:** Dario Lo Presti (cra). **Getty Images:** De Agostini Picture Library (cr). **33 123RF.com:** Lenise Calleja (c/cracker); Chaiyaphong Kitphaephaisan (cr). **Alamy Stock Photo:** David J. Green (c). **Dreamstime.com:** Robert Semnic (cl). **Getty Images:** Stocktrek Images (cla). **Natural Resources Canada, Geological Survey of Canada:** (cb). **34 Dorling Kindersley:** Oxford University Museum of Natural History (cla). **Getty Images:** Ullstein Bild (clb); Universal Images Group (crb). **35 Alamy Stock Photo:** Universal Images Group North America LLC / DeAgostini (cla). **Getty Images:** Keystone-France (ca). **39 123RF.com:** Vladimir Kramin (crb). **Alamy Stock Photo:** Craig Wise (cla). **Dreamstime.com:** Studio306 (cla/sprinkler). **Getty Images:** fStop Images - Caspar Benson (cb). **NASA:** NASA / MSFC / David Higginbotham (c). **Science Photo Library:** David Parker (cr). **Wellcome Images http://creativecommons.org/licenses/by/4.0/:** Wellcome Library (cra). **40 Dorling Kindersley:** Colin Keates / Natural History Museum, London (cla). **41 123RF.com:** Thodonal (cb). **Alamy Stock Photo:** Mohammed Anwarul Kabir Choudhury (clb/cement); Dominic Harrison (cla); Phil Degginger (clb). **Dreamstime.com:** Nu1983 (cr); Marek Uliasz (cra). **Getty Images:** Yoshikazu Tsuno (crb). **Rex by Shutterstock:** Neil Godwin / Future Publishing (cl). **42 Alamy Stock Photo:** Phil Degginger (cl). **Dorling Kindersley:** Natural History Museum, London (clb); Holts Gems (cla). **43 123RF.com:** Oksana Tkachuk (c). **Alamy Stock Photo:** Ekasit Wangprasert (cb). **Dreamstime.com:** Waxart (cr). **44-45 Alamy Stock Photo:** Inge Johnsson. **47 123RF.com:** Anatol Adutskevich (cra); Paweł Szczepański (ca); Ronstik (crb). **Dorling Kindersley:** Durham University Oriental Museum (cla). **Dreamstime.com:** Showface (cr). **iStockphoto.com:** Lamiel (cl). **48-49 Alamy Stock Photo:** The Natural History Museum (cb). **49 123RF.com:** Roman Ivaschenko (cr); Wiesław Jarek (ca). **Getty Images:** DEA / S. VANNINI (c). **Science Photo Library:** ALAIN POL, ISM (crb). **51 Getty Images:** Heritage Images (cra). **Science Photo Library:** Public Health England (ca, crb); Public Health England (cb). **54 123RF.com:** Stocksnapper (cb). **Alamy Stock Photo:** Universal Images Group North America LLC / DeAgostin (cra). **Dreamstime.com:** Dimitar Marinov (crb). **55 123RF.com:** Leonid Pilnik (fcra); Sergei Zhukov (cr). **Alamy Stock Photo:** Military Images (cra); Hugh Threlfall (fcrb). **Dreamstime.com:** Flynt (crb). **56 123RF.com:** Mykola Davydenko (clb); Kaetana (crb). **Alamy Stock Photo:** Shawn Hempel (cl). **57 Alamy Stock Photo:** imageBROKER (crb). **Dorling Kindersley:** Natural History Museum, London (cr). **58 Alamy Stock Photo:** Vincent Ledvina (clb). **59 123RF.com:** Chaiyaphong Kitphaephaisan (c/rail); lightboxx (c); Tawat Langnamthip (crb). **Alamy Stock Photo:** Hemis (cb); B.A.E. Inc. (ca). **Dreamstime.com:** Nexus7 (cr). **Getty Images:** Michael Nicholson (cra). **60 123RF.com:** Serezniy (clb). **Getty Images:** Detlev van Ravenswaay (cl). **61 Alamy Stock Photo:** PhotoCuisine RF (c); SERDAR (l). **Dorling Kindersley:** Doubleday Holbeach Depot (cra). **Dreamstime.com:** Igor Sokolov (clb). **Science Photo Library:** Jim West (cr). **62-63 123RF.com:** Wang Aizhong. **64 Alamy Stock Photo:** Susan E. Degginger (crb); The Natural History Museum (cr). **65 Dorling Kindersley:** Rolls Royce Heritage Trust (cl). **Dreamstime.com:** Margojh (c). **Getty Images:** Pascal Preti (cb); Science & Society Picture Library (cla). **66 Alamy Stock Photo:** Alan Curtis / LGPL (ca). **67 123RF.com:** Psvrusso (ca); Евгений Косцов (crb). **Alamy Stock Photo:** INTERFOTO (fcla). **Dorling Kindersley:** National Music Museum (cla). **Getty Images:** Fanthomme Hubert (cra). **68 Alamy Stock Photo:** Jeff Rotman (crb). **Dorling Kindersley:** Natural History Museum, London (ca); Oxford University Museum of Natural History (clb). **69 123RF.com:** Dilyana Kruseva (cr); Vitaliy Kytayko (cla); Photopips (crb). **Alamy Stock Photo:** Paul Ridsdale Pictures (tc). **Dorling Kindersley:** University of Pennsylvania Museum of Archaeology and Anthropology (cb). **70-71 Alamy Stock Photo:** Novarc Images. **72 Alamy Stock Photo:** Phil Degginger (cla). **73 Alamy Stock Photo:** PjrStudio (clb). **Dreamstime.com:** Sean Pavone (cra). **NASA. 74 Dorling Kindersley:** Oxford University Museum of Natural History (cla). **75 123RF.com:** Belchonock (cr); Weerayos Surareangchai (cla). **Alamy Stock Photo:** dpa picture alliance (ca); Georgios Kollidas (cra); PNWL (cl). **Getty Images:** SSPL (cb). **Science Photo Library:** David Parker (crb). **76 123RF.com:** Okan Akdeniz (clb); Nevarpp (fclb); Andriy Popov (crb). **Dreamstime.com:** Ryan Stevenson (cb). **77 123RF.com:** Mohammed Anwarul Kabir Choudhury (cra); Vladimir Nenov (crb). **Alamy Stock Photo:** The Natural History Museum (cl). **NASA. 78 Alamy Stock Photo:** Oleksandr Chub (crb); The Natural History Museum (ca). **Science Photo Library. 78-79 Alamy Stock Photo:** Susan E. Degginger (c); epa european pressphoto agency b.v. (cb). **79 Science Photo Library:** David Parker (crb); Rvi Medical Physics, Newcastle / Simon Fraser (cr). **80 123RF.com:** Missisya (cb); Darren Pullman (clb). **Alamy Stock Photo:** GFC Collection (crb). **81 123RF.com:** Hywit Dimyadi (cra). **Dreamstime.com:** Shutterman99 (ca). **Getty Images:** Alain Nogues (crb). **82 Alamy Stock Photo:** Greenshoots Communications (ca); PjrStudio (crb). **Dreamstime.com:** Robert Chlopas (cr). **Science Photo Library. 83 123RF.com:** Dmitry Lobanov (crb); Jose Ignacio Soto (tr); Valerii Zan (cr). **Dreamstime.com:** Maloy40 (ca). **Getty Images:** Paul Taylor (cl). **84 Getty Images:** DEA / PHOTO 1 (clb); DEA / G.CIGOLINI (cl). **85 Alamy Stock Photo:** David J. Green (cra); Chromorange / Juergen Wiesler (crb). **Dorling Kindersley:** The University of Aberdeen (ca). **Dreamstime.com:** Stephanie Frey (cla); Gaurav Masand (cl). **Getty Images:** Science & Society Picture Library (clb). **Science Photo Library. 86 123RF.com:** Serhii Kucher (crb). **Alamy Stock Photo:** Ableimages (crb/micro). **Dreamstime.com:** Michal Baranski (cra). **Getty Images:** Lester V. Bergman (cra). **Science Photo Library. 87 Dreamstime.com:** Andrey Eremin (clb). **Science Photo Library. 88 123RF.com:** Ludinko (cra). **Getty Images:** Trisha Leeper (crb). **89 123RF.com:** Akulamatiau (crb); Anton Starikov (clb). **Dreamstime.com:** Homydesign (cb). **90 Alamy Stock Photo:** Antony Nettle (crb). **Dreamstime.com:** Farbled (c); Vesna Njagulj (clb). **91 Alamy Stock Photo:** Science (cra). **Dreamstime.com:** Reddogs (crb). **Science Photo Library:** Dr Gopal Murti (ca); Dirk Wiersma (cla). **92 Alamy Stock Photo:** Citizen of the Planet (crb). **Getty Images:** Yva Momatiuk and John Eastcott (cra). **Science Photo Library. 93 123RF.com:** Sergey Jarochkin (crb); mg154 (cl). **Alamy Stock Photo:** Pictorial Press Ltd (clb). **NASA:** CXC / NGST (ca). **94 Dorling Kindersley:** Natural History Museum, London (r). **Science Photo Library:** Natural History Museum, London (l). **95 Alamy Stock Photo:** Four sided triangle (c); I studio (ca); Friedrich Saurer (cb). **Dreamstime.com:** Adamanto (cr). **Getty Images:** PHAS (cra); Royal Photographic Society (cl). **Science Photo Library:** Dr P. Marazzi (fcr); National Physical Laboratory (C) Crown Copyright (cla); Sovereign / Ism (crb). **96 Science Photo Library:** Science Stock Photography (cra). **97 123RF.com:** Ratchaphon Chaihuai (clb). **Dorling Kindersley:** Alistair Duncan / Cairo Museum (ca); Barnabas Kindersley (cl). **Dreamstime.com:** Nastya81 (crb); Scanrail (c). **Getty Images:** Charles O'Rear (fcrb); John Phillips (cr). **magiccarpics.co.uk:** Robert George (cb). **NASA. 98-99 Alamy Stock Photo:** imageBROKER. **101 123RF.com:** Teerawut Masawat (cla). **Getty Images:** Science & Society Picture Library (ca); Science & Society Picture Library (cr). **Paul Hickson, The University of British Columbia:** (clb). **102 Getty Images:** Bettmann (cla). **Science Photo Library:** Ernest Orlando Lawrence Berkeley National Laboratory / Emilio Segre Visual Archives / American Institute Of Physics (clb). **103 Alamy Stock Photo:** Peter van Evert (crb); Randsc (clb). **Science Photo Library:** Lawrence Berkeley National Laboratory (cr). **104 Science Photo Library:** David Parker (clb); Wheeler Collection / American Institute Of Physics (cla); David Parker (cb). **105 Alamy Stock Photo:** imageBROKER (cla). **Science Photo Library:** Emilio Segre Visual Archives / American Institute Of Physics (cb). **106 Alamy Stock Photo:** Granger Historical Picture Archive (clb). **Science Photo Library:** David Parker (cla). **107 Alamy Stock Photo:** Sherab (cl). **Science Photo Library:** Dung Vo Trung / Look At Sciences (cr). **110 123RF.com:** Oleksandr Marynchenko (clb); Naruedom Yaempongsa (crb). **Alamy Stock Photo:** John Cancalosi (cla); Reuters (cb). **111 123RF.com:** Cobalt (cra); Veniamin Kraskov (cra/red). **Dreamstime.com:** Akulamatiau (cb). **Science Photo Library. 112 Alamy Stock Photo:** Everett Collection Historical (crb). **113 Alamy Stock Photo:** Ivan Vdovin (crb). **Fotolia:** Efired (cra). **114 Rex by Shutterstock:** (crb). **Science Photo Library:** Pr Michel Brauner, ISM (cra). **115 Alamy Stock Photo:** G M Thomas (cr). **Science Photo Library:** Patrick Llewelyn-Davies (crb). **116 123RF.com:** Vereshchagin Dmitry (br); Vitalii Tiahunov (cra); Vitalii Tiahunov (fcra). **117 123RF.com:** Preecha Bamrungrai (crb). **Dreamstime.com:** Hxdbzxy (cra). **120 Alamy Stock Photo:** Yon Marsh (br). **ESA:** (cra). **Science Photo Library:** Dirk Wiersma (clb). **121 Alamy Stock Photo:** Mike Greenslade (l). **Science Photo Library:** Trevor Clifford Photography (cra); Sputnik (crb). **122 Alamy Stock Photo:** Derrick Alderman (cr). **Science Photo Library:** J.C. REVY, ISM (cra); Lawrence Berkeley Laboratory (crb). **123 NASA:** NASA / JPL-Caltech / Malin Space Science Systems (cra). **Science Photo Library:** Thedore Gray, Visuals Unlimited (ca). **124 Alamy Stock Photo:** Randsc (crb). **Dreamstime.com:** Marcorubino (clb). **NASA. Science Photo Library:** Science Source (cla). **125 Alamy Stock Photo:** 501 collection (cla). **Getty Images:** George Rinhart (clb). **Science Photo Library:** US Department Of Energy (ca). **126 Science Photo Library:** American Institute Of Physics (cra); Sputnik (cb); Sputnik (clb). **127 Alamy Stock Photo:** Granger Historical Picture Archive (clb). **Science Photo Library:** Ernest Orlando Lawrence Berkeley National Laboratory / Emilio Segre Visual Archives / American Institute Of Physi (ca). **130 123RF.com:** Terry Davis (cla). **Alamy Stock Photo:** Chris stock photography (fcra). **130-131 Alamy Stock Photo:** Universal Images Group North America LLC / DeAgostini (cb). **131 123RF.com:** Sirichai Asawalapsakul (ca); Joerg Hackemann (cla); Wilawan Khasawong (ca/boric); Michał Giel (c/TV). **Alamy Stock Photo:** Chronicle (fcra). **Dorling Kindersley:** Tank Museum (crb). **Fotolia:** L_amica (c); Alex Staroseltsev (cr). **Getty Images:** Heritage Images (cra). **132 Science Photo Library:** Dirk Wiersma (ca). **133 123RF.com:** Destinacigdem (cla); Olaf Schulz (cb). **Dreamstime.com:** Apple Watch Edition™ is a trademark of Apple Inc., registered in the U.S. and other countries. (cr); Stepan Popov (cl); Simon Gurney (fcr); Zalakdagli (clb). **134-135 Getty Images:** Brasil2. **136 123RF.com:** Martin Lehmann (c). **Alamy Stock Photo:** BSIP SA (cra). **Getty Images:** Visuals Unlimited, Inc. / GIPhotoStock (cra/disc). **NASA. 137 123RF.com:** Norasit Kaewsai (cb/trans); Ouhdesire (clb); Dmytro Sukharevskyy (cb). **Dreamstime.com:** Christian Delbert (crb). **138 Dreamstime.com:** Monika Wisniewska (cl). **Getty Images:** Science & Society Picture Library (cr). **139 123RF.com:** Fotana (clb). **Alamy Stock Photo:** Stock Connection Blue (cla). **Getty Images:** The Asahi Shimbun (cr). **142 Alamy Stock Photo:** Pablo Paul (cra); WidStock (cla). **Dorling Kindersley:** Natural History Museum (crb); Natural History Museum (fcra). **143 123RF.com:** Oleksii Sergieiev (crb). **Alamy Stock Photo:** David J. Green (cla); Image.com (ca). **Dorling Kindersley:** National Cycle Collection (ca/cycle); The Science Museum, London (fcrb). **144-145 Bridgeman Images:** Christie's Images. **146 123RF.com:** Danilo Forcellini (fcrb). **Alamy Stock Photo:** Phil Degginger (cra); Perry van Munster (crb). **147 123RF.com:** Scanrail (ca). **Alamy Stock Photo:** MixPix (cla); Haiyin Wang (crb). **Dreamstime.com:** Halil I. Inci (fcrb). **Getty Images:** Handout (cr). **Science Photo Library:** Lawrence Berkeley National Laboratory (cb). **148 123RF.com:** Viktoriya Chursina (cra). **Dreamstime.com:** Bright (crb); Oleksandr Lysenko (cr). **Getty Images:** DEA / G. CIGOLINI (cla). **149 123RF.com:** Lapis2380 (cb). **150 Alamy Stock Photo:** Sarah Brooksby (cb). **Dorling Kindersley:** Natural History Museum, London (cla). **151 123RF.com:** Vira Dobosh (clb). **Science Photo Library:** Sputnik (cr); Sputnik (cb). **154 Science Photo Library:** Dr.Jeremy Burgess (clb). **155 123RF.com:** Mohammed Anwarul Kabir Choudhury (cra/color); Teerawut Masawat (cla); David Gilbert (ca). **Alamy Stock Photo:** Lyroky (cr); Tim Scrivener (crb). **NASA:** JPL (cra). **156-157 Getty Images:** Icon Sports Wire. **158 Dorling Kindersley:** Natural History Museum, London (cr/beaker). **Dreamstime.com:** Tomas Pavelka (c). **Getty Images:** Auscape (ca). **158-159 Science Photo Library. 159 123RF.com:** Action sports (cb); De2marco (c); Arina Zaiachin (clb); Mohammed Anwarul Kabir Choudhury (crb). **Getty Images:** Simone Brandt (cr). **160 123RF.com:** Maksym Yemelyanov (crb). **Alamy Stock Photo:** Andrew Ammendolia (fclb). **Dreamstime.com:** Jaggat (clb); Science Pics (ca). **161 Dreamstime.com:** Chris Boswell (cra). **Getty Images:** SuperStock (cb). **Science Photo Library:** Phil Degginger (cla); Phil Degginger (fcla). **162 Dorling Kindersley:** Harry Taylor (cla). **163 123RF.com:** Serezniy (cl). **Dreamstime.com:** Ericlefrancais (cla); Bert Folsom (clb). **Science Photo Library:** Sputnik (cr). **166 123RF.com:** Kameel (clb); Russ McElroy (cra). **167 123RF.com:** Rostislav Ageev (cra/diver). **Alamy Stock Photo:** PhotoAlto (crb); RGB Ventures / SuperStock (fcr). **Dreamstime.com:** Narin Phapnam (cb); Uatp1 (cl). **Getty Images:** STR (clb). **SuperStock:** Cultura Limited / Cultura Limited (cra). **168 123RF.com:** Cseh Ioan (clb). **Alamy Stock Photo:** Big Pants Productions (cb). **Science Photo Library:** Farrell Grehan (cra). **168-169 Alamy Stock Photo:** The Natural History Museum (ca). **169 123RF.com:** Lucian Milasan (crb); Nikkytok (c). **Alamy Stock Photo:** Krys Bailey (clb); Paul Felix Photography (cra). **Dreamstime.com:** Nfransua (ca); Kirsty Pargeter (cla); Olha Rohulya (cb). **170-171 Getty Images:** Kazuyoshi Nomachi. **172 123RF.com:** Maksym Bondarchuk (crb); Sauletas (cb). **Dreamstime.com:** Orijinal (clb). **Science Photo Library:** Dirk Wiersma (cra). **173 123RF.com:** Jiri Vaclavek (ca). **Alamy Stock Photo:** Hugh Threlfall (fcra); Universal Images Group North America LLC / DeAgostini (cla). **Getty Images:** Steve Proehl (crb). **174 Alamy Stock Photo:** Dan Leeth (clb). **175 Alamy Stock Photo:** CPC Collection (cb); Sputnik (cla). **178 Dorling Kindersley:** Natural History Museum, London (crb); Oxford University Museum of Natural History (cla). **179 123RF.com:** Kirill Krasnov (c); Chaovarut Sthoop (clb). **Alamy Stock Photo:** The Print Collector (cra); World History Archive (cl); World foto (cr). **Dreamstime.com:** Bogdan Dumitru (fcrb); Stephan Pietzko (cla). **Getty Images:** John B. Carnett (crb). **181 123RF.com:** Sergey Jarochkin (c); Dmitry Naumov (cr); Hxdbzxy (cb, cb/bleach). **Alamy Stock Photo:** Maksym Yemelyanov (crb). **Dorling Kindersley:** Thackeray Medical Museum (cl). **182-183 Science Photo Library:** Alexis Rosenfeld. **184-185 Getty Images:** George Steinmetz (cb). **Science Photo Library:** Charles D. Winters (c). **185 Dreamstime.com:** Jose Manuel Gelpi Diaz (crb); Larry Finn (ca). **Science Photo Library. 186 123RF.com:** Alexandr Malyshev (fcra); ń ńń (cra). **Alamy Stock Photo:** BSIP SA (crb). **187 Science Photo Library:** Union Carbide Corporation's Nuclear Division, courtesy EMILIO SEGRE VISUAL ARCHIVES, Physics Today Collection / AMERICAN INSTITUTE OF PHYSICS (clb). **190-191 Dreamstime.com:** Andrey Navrotskiy (b). **190 123RF.com:** Leonid Ikan (cb). **© CERN:** (cra). **191 Dreamstime.com:** Yinan Zhang (cb). **iStockphoto.com:** Gobigpicture (t). **Science Photo Library:** Brian Bell (ca); Patrick Landmann (crb). **192-193 Getty Images:** Rolf Geissinger / Stocktrek Images. **194 123RF.com:** Rainer Albiez (cra). **Alamy Stock Photo:** D. Hurst (crb). **Science Photo Library:** Andrew Lambert Photography (clb). **195 Dreamstime.com:** Stocksolutions (cr). **Getty Images:** Floris Leeuwenberg (cra); Mario Tama (clb). **Science Photo Library:** Crown Copyright / Health & Safety Laboratory (crb). **196 Dorling Kindersley:** Clive Streeter / The Science Museum, London (crb). **Dreamstime.com:** Liouthe (cra). **Getty Images:** Genya Savilov (fcra). **Science Photo Library:** (clb). **197 123RF.com:** Alexlmx (fcrb). **Alamy Stock Photo:** Alexandru Nika (crb). **Dreamstime.com:** Jultud (clb). **Getty Images:** Brand X Pictures (cra). **NASA:** JPL-Caltech (cb). **198 Science Photo Library:** Dirk Wiersma (ca). **199 123RF.com:** Nmint (cl). **Alamy Stock Photo:** ITAR-TASS Photo Agency (cra); Gordon Mills (cla); RGB Ventures / SuperStock (clb); ITAR-TASS Photo Agency (crb). **200 Alamy Stock Photo:** Shawn Hempel (bc). **205 Dorling Kindersley:** Natural History Museum (cla)

All other images © Dorling Kindersley
For further information see: **www.dkimages.com**

Electronic Projects for Beginners

F. G. RAYER, T.Eng. (CEI), Assoc. IERE

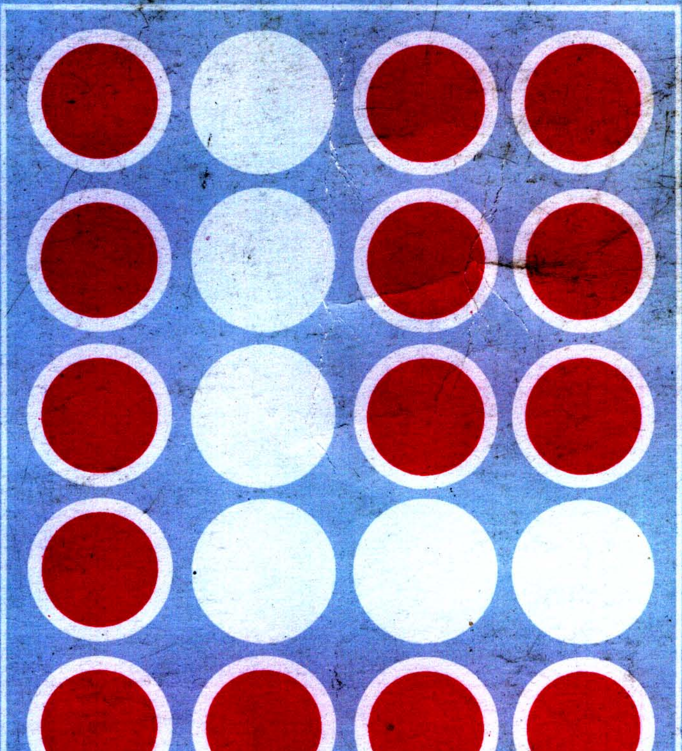

ELECTRONIC PROJECTS FOR BEGINNERS

ALSO BY THE SAME AUTHOR

No. BP32	How to Build Your Own Metal and Treasure Locators
No. BP37	50 Projects Using Relays, SCRs & Triacs
No. BP39	50 (FET) Field Effect Transistor Projects
No. BP43	How to Make Walkie-Talkies
No. BP57	How to Build Your Own Solid State Oscilloscope
No. BP67	Counter Driver and Numeral Display Projects
No. BP75	Electronic Test Equipment Construction
No. BP79	Radio Control for Beginners
No. BP84	Digital IC Projects
No. BP90	Audio Projects
No. BP93	Electronic Timer Projects
No. BP97	IC Projects for Beginners

ELECTRONIC PROJECTS FOR BEGINNERS

by

F.G. RAYER, T.Eng.(CEI), Assoc. IERE

BERNARD BABANI (publishing) LTD
THE GRAMPIANS
SHEPHERDS BUSH ROAD
LONDON W6 7NF
ENGLAND

PLEASE NOTE

Although every care has been taken with the production of this book to ensure that any projects, designs, modifications and/or programs etc. contained herein, operate in a correct and safe manner and also that any components specified are normally available in Great Britain, the Publishers do not accept responsibility in any way for the failure, including fault in design, of any project, design, modification or program to work correctly or to cause damage to any other equipment that it may be connected to or used in conjunction with, or in respect of any other damage or injury that may be so caused, nor do the Publishers accept responsibility in any way for the failure to obtain specified components.

Notice is also given that if equipment that is still under warranty is modified in any way or used or connected with home-built equipment then that warranty may be void.

©1978 BERNARD BABANI (publishing) LTD

ISBN 0 85934 054 6

First Published – August 1978
Reprinted – September 1980
Reprinted – June 1982
Reprinted – January 1984
Reprinted – April 1986
Reprinted – February 1988
Reprinted – April 1989
Reprinted – August 1991

Printed and bound in Great Britain by Mayhew McCrimmon Printers Ltd.

CONTENTS

Page
INTRODUCTION 7
  Conversion Table 8

SECTION 1: "NO SOLDERING" PROJECTS 9
  Theoretical Circuits 11
  Signal Tracer 13
  Multivibrator 14
  "C" Substitution Box 16
  "R" Substitution Box 18
  SCR Controlled Light 18
  One Transistor Intercom 21
  Tuned Ferrite Aerial 23
  Diode Radio 25
  One Transistor Amplifier 31
  Two Transistor Amplifier 32
  Push-Pull Output Stage 34
  Boost Speaker 37

SECTION 2: MISCELLANEOUS DEVICES 39
  Soldering 39
  One Transistor Audio Oscillator 40
  Multivibrator 41
  Glissade 43
  Morse Oscillator for Phones 45
  Two Door Siren 46
  UJT Oscillator 48
  UJT Continuity Tester 49
  UJT High Voltage Producer 50
  Signal Blinker 51
  Mains Wiring 52
  Dimmer/Heat Control 53
  Two Power 55
  Sensitive Touch Switch 56
  Water Level Indicator 58
  Optical Isolator 60
  LED Indicator 61
  "Magic Candle" 62
  "Nerve Tester" 63

Page
Burglar Alarm 64
Transistor Tester 65
Treasure Locator 67

SECTION 3: RADIO AND AUDIO FREQUENCY 71
MW Booster 71
Midget MW Diode Radio/Tuner 74
One Transistor VHF Receiver/Converter 75
Morse and Sideband Resolver 78
Easy Preamplifier 81
741 Preamplifier 83
Tone Control for Your Discs Etc. 85
Two Channel Mixer 86
¼ Watt IC Amplifier 88
Class A Amplifier 90
Two Stage Amplifier 92

SECTION 4: POWER SUPPLIES 95
Easy PSU 95
Easy Regulated PSU 97
Multi-Voltage Regulated PSU 98
One Transistor Adjustable PSU 99
Zener Regulated Supply 102
Motor PSU with Reverse 103
Wrong Polarity Protector 104
Transistor Lead-Out Positions 105

EQUIVALENT SEMICONDUCTORS 107

Page 7 & 8 missing --------->

SECTION I

"NO SOLDERING" PROJECTS

With these, connections are generally made with small terminals, or nuts and bolts. By this means components are connected together, and fixed to an insulated board or panel.

A project can be built using one size of bolt and nut throughout, but it will be found convenient to have two or three sizes. Small nuts and bolts can then be used for connections between components, while larger bolts with extra nuts, or terminals, can be used to connect external items, such as headphones.

For one size throughout, 6ba screws or bolts will be satisfactory, about ½ inch long. Only a few are needed for each project, but they are quite inexpensive in 50 or 100 packs. Some extra nuts will prove useful, and also a quantity of 6ba washers. If the insulated board or panel is not too thick, the ½ inch bolts with an extra nut each will do as terminals for connecting external items.

A 4ba bolt is *larger* than 6ba, and a few of these, ½ inch or more long, also with washers and nuts, can be used for terminals, if preferred. It is also possible to obtain terminal heads, so that bolts with nuts can be used as terminals.

The 8ba size is *smaller* than 6ba. These can be of advantage in a project where size is to be kept down, or for positions where several leads come close together (as at a transistor). Bolts, nuts and washers (8ba throughout) can be obtained together.

Naturally all sorts of nuts and bolts, to hand or obtained from constructional toys or old electronic equipment, can be used. Results are not going to depend on the sizes used, which are more a matter of convenience and appearance

There is also considerable latitude in the insulated panels or boards used to carry the components. Some projects need only a few holes in a small board. Here, it is convenient to use 1/16 inch thick or similar paxolin sheet, cutting a piece of

the wanted size and drilling holes for the bolts or terminals. A small hand drill is satisfactory for this purpose. The drills used can be 3/32 inch for 8ba bolts, 1/8 inch for 6ba, and 5/32 inch for 4ba, or any similar sizes giving clearance.

Where drilling is inconvenient, a ready-perforated board must be used. For projects where size is not important, perforated 1/8 inch thick hardboard is suitable – holes are generally at 1 inch intervals, but this is not important. For this type of board, the size of the holes will generally mean that 4ba nuts and bolts will be beeded.

Plain perforated board (that is, without any conductor foil) can be obtained with holes at 0.1 inch or 0.15 inch spacing but these holes are too small and too close together. However, such boards can be used instead of plain unperforated paxolin sheet by drilling the holes larger in just those places where bolts must be placed.

There is also a type of eyelet board, with holes at 5mm spacing and just large enough to take 8ba bolts. This is handy for small projects.

It is also relatively easy to prepare a fully perforated board or panel, which can be used for a number of projects in turn.

For a board up to about 4 x 6 inch paxolin 1/16 inch thick is adequate, with 1/8 inch sheet for larger boards. It is also possible to use 1/8 inch or similar PVC or other transparent sheet, which allows connections to be seen on both sides. Plain hardboard, or thin plywood, is also satisfactory. Hardboard or wood should be given a coat of shellac or similar varnish, before use, to keep out damp.

Where size is not important, the board can be drilled with holes spaced 1 inch from each other throughout. Do *not* mark off a grid with pencil lines, as this may cause leakage between adjacent terminals. An easy method is to mark such a grid on paper, secure this over the board or panel with tape, and drill all the holes. Make these of an appropriate size for the terminals or bolts.

For small items, the board will need to be prepared to suit.

But where a board is going to be a general purpose support for several projects in turn, it can be much larger. It is also handy to support it on two strips of wood, about 1 inch high, so that there is clearance underneath for bolts and connections.

It will be found that neat and efficient projects can be assembled on such boards, and that it is easy to check connections or components, or to make changes for experimental or other purposes.

Theoretical Circuits

A theoretical circuits shows an electronic device in simple form. Capacitors, resistors, transistors and other items are shown by their own symbols. Lines representing connections are drawn from symbol to symbol. The layout of actual components in the device may not be the same as the positions of the symbols in the theoretical circuit.

It will be found that it soon becomes easy to follow a theoretical circuit. An understanding of such circuits develops when they are compared with an actual layout which shows components and the connections.

The Signal Tracer in Fig. 1 is an example of this. The main item here is the transistor, which has emitter (E), base (B) and collector (C) wires. For the transistors shown, these wires come from the base as in the diagram, which is of the underside of the transistor. (Emitter, base and collector wires may be in different positions in other transistors.) It will be seen that the emitter E is connected to negative at the battery, and to a clip on a lentth of flexible wire.

The headphones – or single earpiece – go between collector C and battery positive. A headphone jack, marked P, is used to take a plug for the phones. Generally, an on-off switch would be present in one battery connection, but this is not necessary here as the circuit can be interrupted (to switch "off") by unplugging the phones. Exactly the same results would be obtained if two separate sockets were used, to take two plugs from the headphones.

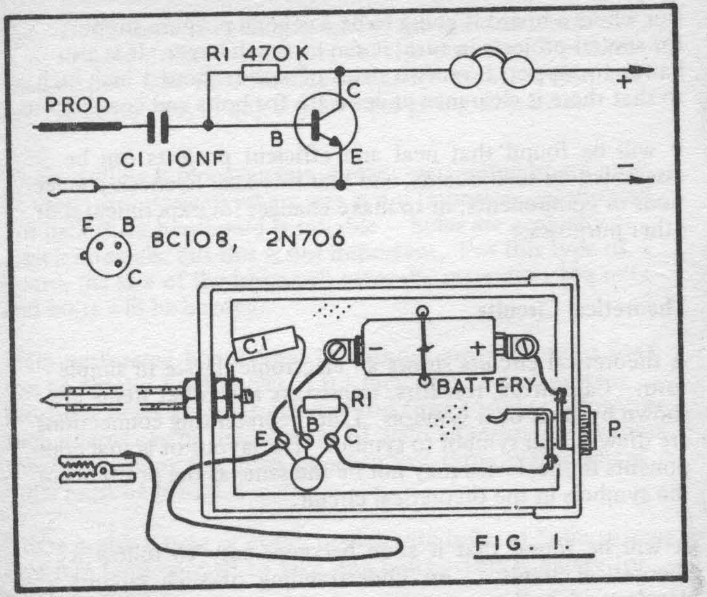

FIG. 1.

Resistor R1 is connected from collector C to base B, and it supplies a base current for the transistor, so that the latter can operate as an amplifier.

The signal to be amplified is obtained through capacitor C1. A capacitor has plates or foils separated from each other by an insulator. There is no direct current circuit through the capacitor, so placing the test prod on various points of a circuit being tested does not upset direct current working conditions.

It is clear, from Fig. 1, how the theoretical circuit is made up as a practical piece of equipment, with C1, R1, transistor, phones and battery connected together in the way required.

With later projects, the theoretical circuit and practical wiring diagram may be shown together in this way, or may appear as separate diagrams when they are larger. If the theoretical circuit and wiring diagram are compared, so that capacitors, resistors and other items on one are also found on the other,

it will soon be found that a theoretical circuit is readily followed and understood. It will be noted that the theoretical circuit could be built up in numerous ways, all equally satisfactory.

Signal Tracer

This locates and amplifies audio (sound) signals which may be present in a circuit. Construction is most easily carried out in a small plastic case (Fig. 1). An insulated board about 1¼ x 2 inch (or to suit the case) has three small bolts to secure emitter, base and collector leads of the transistor. Bolts also hold two small brackets, which press on the negative and positive ends of the 9v battery. String or elastic through holes in the board helps to secure the battery.

The prod is a long 6ba or 4ba bolt, or length of screwed rod. Lock one end of C1 lead with nuts, or loop it under the bolt head. Thread the clip flex through a hole in the case, and pass the prod through another hole, as shown. A washer and nut then fix the prod, with the board inside the case. The case lid can then be fitted, and the securing nut can be screwed on the phone outlet. Connections here can be tightly twisted round the outlet tags, and compressed with pliers. A piece of insulating sleeving is put on the test prod, and a clip with fixing screw is attached to the flexible lead.

Making Connections

For all the projects in this book, any thin insulated connecting wire can be used; or bare wire with insulated sleeving. It is handy to have two or more colours – such as red for positive, and black for negative circuits.

A loop is made at the bare end of the wire, and is put under the screw head, or under a washer. Where two or more leads go to the same bolt, it is convenient in some cases to have a washer or two extra, to go between loops. Each connection must be reasonably tight, and short circuits must not be allowed between wires or loops which are shown separated from each other.

Battery and external leads are generally of flexible wire. Positive and negative battery clips, with leads already attached, are available for transistor radio and similar batteries. Elsewhere it may be convenient to arrange contacts as in Fig. 1, or to twist wires to the battery tags.

The battery polarity (positive and negative) *must* always be correct, as shown.

When dealing with transistors, it will be found that some types have long wire leads, more than adequate for easy screwed connections. Others, such as the BC108 and similar types, have rather short wires, but can be held satisfactorily with 8ba bolts fairly near eath other. (With other transistor types not used in this section, soldering may be essential for proper connections to be made.)

Circuit Tracing

To use the signal tracer, connect the clip to the earth return line of the receiver or amplifier. The prod is then touched one by one on points where the audio signal should be present, working systematically through the audio amplifier, or from the first point where the signal can be heard. Take in each capacitor, joint, or stage in turn. When signals cease to be heard, the fault lies between this point and the previous point, where signals were heard.

Do not test any mains operated or high voltage equipment, as this is a job for an engineer who will follow the necessary and essential precautions for safety.

Medium or high impedance phones are most suitable for this tracer. A 600 ohm DC headset will be ideal. The tracer can be utilised as an amplifier for a crystal diode radio receiver

Multivibrator

You will be able to use this as an audio oscillator for numerous test purposes, or for Morse Code practice, or as a signal device in place of a bell or buzzer.

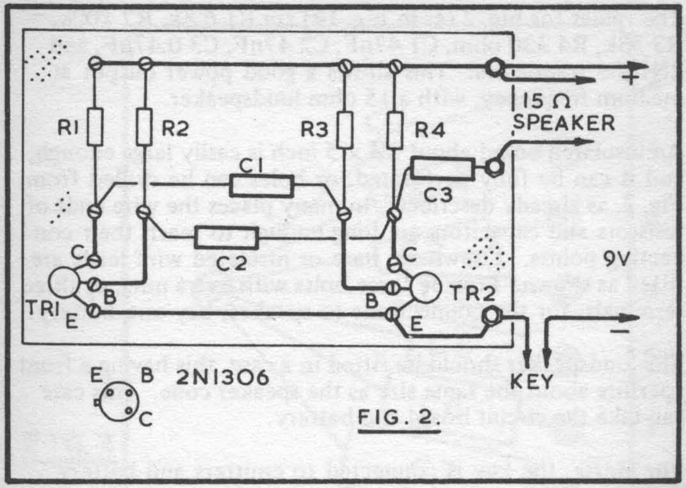

FIG. 2.

Fig. 2 shows the practical assembly of the device on an insulated board. A theoretical circuit is given in Fig. 14, and is not repeated here, so Figs. 2 and 14 should be looked at together.

R1 is the collector load resistor for transistor TR1. Collector current passes through this resistor, and an audio signal voltage is developed across it. Resistor R2 supplies base current for TR1.

TR2 has a similar collector resistor R4, and base resistor R3. Capacitor C1 couples the signal at the collector C of TR1 to the base B of TR2, while C2 similarly couples TR2 collector to TR1 base. Amplification in the transistors and feedback by the capacitors in this way result in the circuit oscillating. That is, each transistor goes rapidly in and out of conduction in turn, so that a voltage at audio frequency is produced across the collector load resistors. C3 is a coupling capacitor, allowing the signal voltage at the collector of TR2 to be taken to a loudspeaker or headphones.

With a circuit of this type, the component values, transistors, and battery voltage will not be critical. However, resistor values, capacitor values, battery voltage, transistors, and the loudspeaker or other sound-reproducing device will all have an influence on the frequency of oscillation, or tone produced.

The values for Fig. 2 (as in Fig. 14) are R1 6.8k, R2 100k, R3 33k, R4 330 ohm, C1 47nF, C2 47nF, C3 0.47μF, and 2N1306 transistors. This allows a good power output at medium frequency, with a 15 ohm loudspeaker.

An insulated board about 2½ x 5 inch is easily large enough, and it can be fully perforated, or holes can be drilled from Fig. 2, as already described. In many places the wire ends of resistors and capacitors are long enough to reach their connecting points. Elsewhere, bare or insulated wire leads are fitted as shown. Provide three bolts with extra nuts, or three terminals, for the connections to speaker, key and battery.

The loudspeaker should be fitted in a case, this having a front aperture about the same size as the speaker cone. This case can take the circuit board and battery.

For Morse, the key is connected to emitters and battery negative as shown, with flexible leads of convenient length. If the multivibrator is to be used as a warning device then these leads can instead go to a door bell push, or to a relay or SCR, used to complete the circuit, as with some later projects. In this way the multivibrator, with its speaker, can replace a buzzer or bell.

If the frequency of the multivibrator is to be changed, this can readily be done by altering the values of capacitors C1 and C2 in particular. The two capacitors need not be of the same value. Larger values reduce the frequency (lower the tone) while smaller capacitors increase the frequency. C3 also has some effect, due to loading from the speaker.

Such a circuit can be used with a battery voltage as low as 1.5v, for headphones. Transistor types are very uncritical. But if PNP transistors are fitted, the battery polarity must be reversed. The transistors shown are NPN.

"C" Substitution Box

When experimenting with circuits or checking a suspected capacitor, a box bringing into circuit a range of capacitor values may be used. In many circuit positions an exactly correct value is not required for testing by substitute, so that a very large number of capacitors need not be provided.

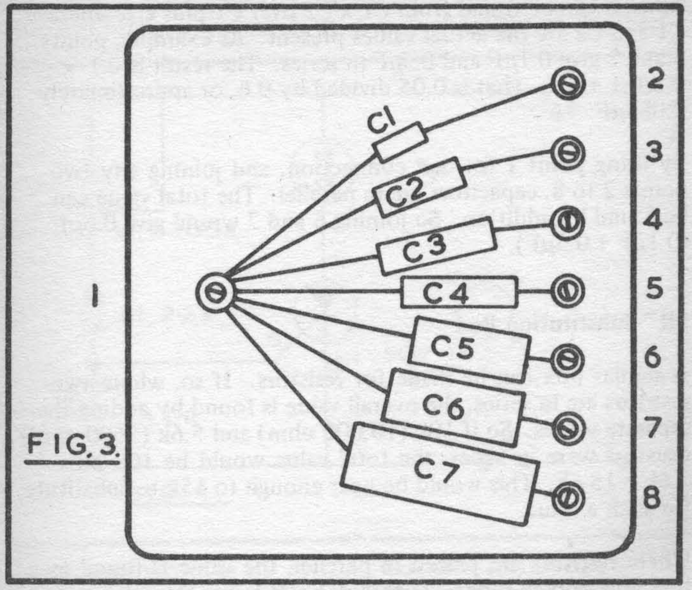

FIG. 3.

Such a box is easily made as in Fig. 3, using a plastic or other insulated case. Here, point 1 is common to all the capacitors, and any capacitor can be selected by connecting to points 2 to 8. Connections can be taken to 6ba bolts, using extra nuts. An alternative is to have eight sockets, with a pair of test leads equipped with plugs.

Suggested values are C1 100pF, C2 470pF, C2 1nF (1000pF or 0.001μF), C4 10nF (0.01μF), C5 0.1μF, C6 0.5μF, and C7 8μF. As the last capacitor will be polarised, mark its terminal to agree.

These values will be found sufficiently near for use in many coupling, decoupling, by-pass and similar circuit positions. The substitution box can be used in place of C1 or C2 in Fig. 2. But where a defective capacitor is found in a tuned circuit, it should be replaced by the exact value.

By ignoring terminal 1 and connecting leads to any pair of the terminals 2 to 8, two capacitors can be used in series. For those who like to work out the value of such combina-

tions, it can be found from C1 x C2 over C1 plus C2, where C1 and C2 are the actual values present. As example, points 6 and 7 give 0.1μF and 0.5μF in series. The result is 0.1 x 0.5/0.1 + 0.5. That is 0.05 divided by 0.6, or approximately 0.083μF.

By using point 1 for one connection, and joining any two points 2 to 8, capacitors are in parallel. The total value can be found by addition. So joining 6 and 7 would give 0.6μF (0.1μF + 0.5μF).

"R" Substitution Box

A similar box can be made for resistors. If so, where two resistors are in series, the overall value is found by adding the separate values. So if 10k (10,000 ohm) and 5.6k (5600 ohm) resistors were in series, the total value would be 10k plus 5.6k = 15.6k. This would be near enough to 15k to substitute for such a value.

Where resistors are placed in parallel, the value is found as for capacitors in series. That is, R1 x R2 over R1 + R2. So for the 10k and 5.6k resistors, this would be approximately 3.6k.

Naturally each resistor may be selected alone, as for the capacitors. If sufficient terminals are provided, a large number of resistors could be fitted. For most transistor circuits, values from about 150 ohm to 270k will fill most needs.

SCR Controlled Light

This device will switch on a light, when the level of general illumination falls below that for which triggering is set. It can thus be used for an automatic night light. Operation can be from a battery; or from a mains power supply, as described later.

Fig. 4 is the circuit, and introduces two new devices. One is the light-dependent resistor LDR. The resistance of this component depends on the strength of the light falling on it. The

resistance is low (some hundreds of ohms) with bright light, but rises to many thousands of ohms in dim light, or darkness.

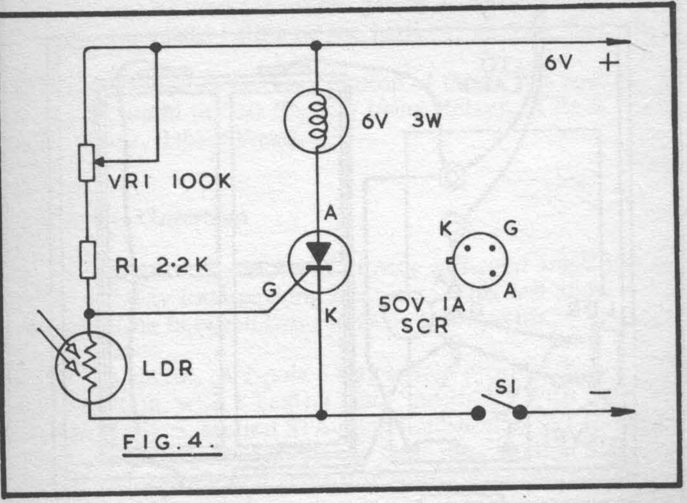

FIG. 4.

The second device is the silicon controlled rectifier SCR. Normally, this does not allow current to pass from its cathode K to its anode A. However, a small current applied to its gate G will trigger it so that it conducts from cathode to anode. It remains in this state even when gate current is removed. It is restored to the non-conducting state by removing the voltage from its cathode, by opening switch S1.

In this device the SCR is controlled by the LDR. With illumination present the resistance from gate to cathode is low, so that when S1 is closed, current does not flow from the anode circuit through the 6 volt 3 watt lamp. As illumination falls, the resistance of the LDR rises. This continues until a sufficient gate potential is present to operate the SCR, which moves into its avalanche or conducting state, lighting the lamp. The potentiometer VR1 enables the resistance from positive to gate to be adjusted, to set the illumination level at which the circuit operates.

A 50v 1 amp SCR is more than adequate for this circuit, and it resembles a large transistor externally, and can be mounted by its wire ends.

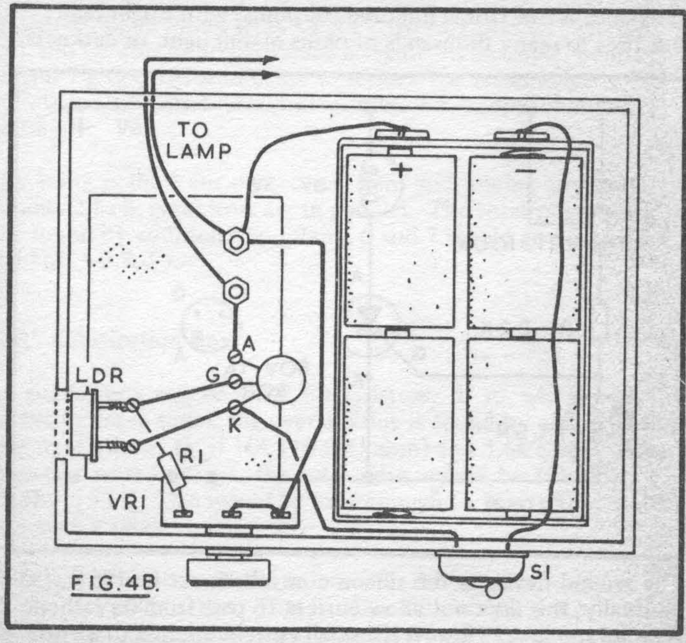

FIG.4B.

Fig. 4B shows constructional and wiring details. The SCR is mounted on a small insulated board. Clear bared wires are twisted tightly to the LDR, so that this also can be fitted by small bolts and nuts, as shown. The connections to VR1 are also tightly twisted and pressed together with pliers.

A battery holder for four 1.5v cells is shown, and connections to this can be by means of clips which have leads attached. The switch S1 is the small, low-voltage type having small screws, as used for toy lighting circuits and similar purposes.

A hole is cut in the box level with the LDR, which may be directed towards the window of the room, or positioned as found to give the results wanted. The bulb, in holder, can be positioned on top of the case. If light from it reaches the LDR this will not stop the circuit working, as the SCR can only be returned to the non-conducting state by opening the switch S1,

For an automatic night light or similar device, a case can be made from thin wood, with opening back or bottom. The device can also be used for various games, where momentarily interrupting the light falling on the LDR switches on the bulb.

If a detailed reference to the operation of the SCR is wanted, this can be found in "50 Projects Using Relays, SCRs & Triacs" (BP37, Babani Press).

One Transistor Intercom

A single high gain transistor will provide sufficient amplification for a 2-way loudspeaking intercom. This will allow communication between two rooms or other points.

Fig. 5 is the circuit. A 2-pole 3-way switch provides 2-way communication, with a central "Off" position. With the switch as in Fig. 5, section S1A takes loudspeaker SP1 to the base coupling capacitor C1. At the same time, section S1B connects the battery and transistor collector circuit to the

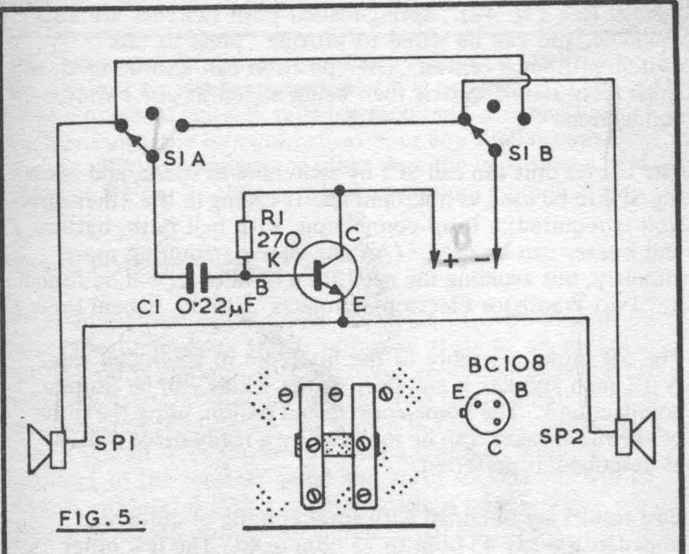

FIG.5.

speaker SP2. SP1 then acts as a microphone, and sounds are reproduced in SP2.

With the switch in the opposite position, S1A selects SP2 as microphone, and amplified signals pass via S1B to SP1, so that speech can be in the other direction.

The switch, amplifier, battery and SP2 are assembled as the master control unit, and twin leads run to SP1. The person at SP1 should say "Over" so that the other person can operate the switch, and he in turn should say "Over" to signify that SP1 is again acting as microphone.

A simplified home-made change over switch is shown in Fig. 5. This is two strips of metal, pivoted so that they can be swung from side to side, and bearing on two of the three round-headed screws. A strip of insulating material is fitted so that both strips move together. In Fig. 5, this switch is in its "Off" position.

Ready made 2-pole 2-way switches can be obtained and used instead (see Fig. 49). Spring-loaded push switches are also available, and can be wired to provide "press to talk". A switch without a central "Off" position can also be used, a single pole on-off switch then being added in one battery connection.

The master unit can call SP1 by switching to speak, and allowing SP2 to be used as microphone. If calling in the other direction is required, a third connection, with bell push, battery and buzzer can be used. (An alternative, requiring more circuitry, but avoiding the need for a third wire, will be found in "Two Transistor Electronic Projects", (BP30, Babani Press.)

Fig. 5B shows assembly of the intercom in a wooden case. A 3½ inch speaker is suitable, so the whole can be quite a compact unit. The home-constructed switch, using the bolts below the speaker, can be replaced by a ready-made switch, as described, if preferred.

Best results are obtained with speaker units of quite high impedance – say 45 ohm to 75 ohm or so. The few other components are assembled on a small insulated board, as for earlier circuits.

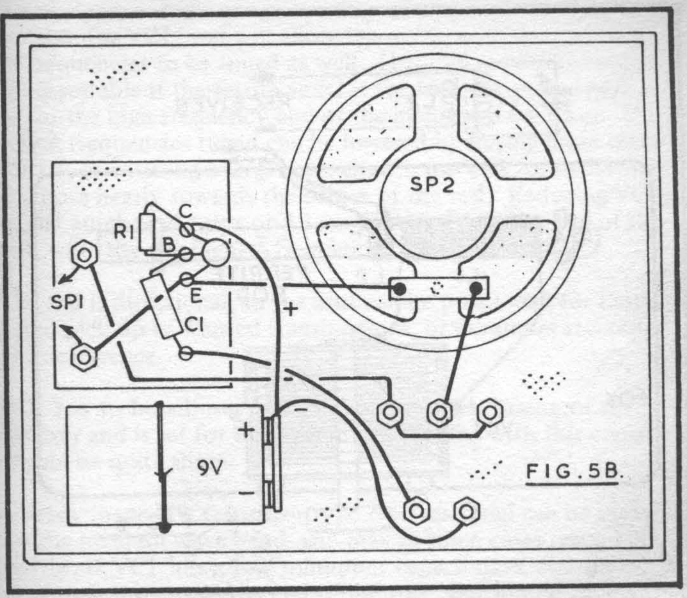

FIG.5B.

The speaker at the remote point also requires a small case. Dimensions are not important, and it is possible to utilise plastic boxes of suitable type for each unit. Connections between units are by twin bell wire, or any similar twin conductor. With a transistor of other than the type shown, the value of R1 can be chosen to give maximum amplification.

Tuned Ferrite Aerial

This simple device will allow "portable" reception with the old type of receiver which normally requires an external aerial, and it can also give improved reception in terms of freedom from interference on the 160 metre amateur band and adjacent frequencies. As shown in Fig. 6, it consists of a tuned circuit with ferrite rod, and a coupling winding to connect to the receiver aerial and earth sockets. It will be found very useful with a communications type receiver, where a long extended outdoor or similar aerial results in a high level of noise and interference.

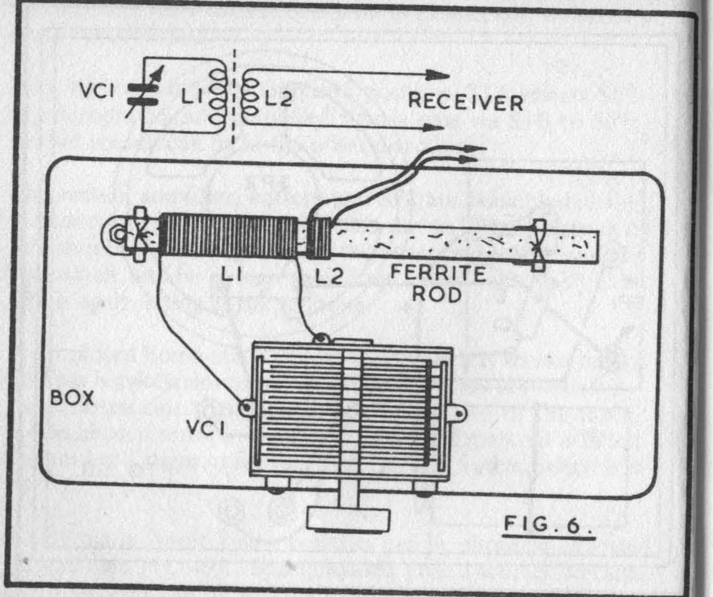

FIG. 6.

The device is constructed in a plastic lunch box about 6 x 4 x 4 inch or other *insulated* case. Some boxes of this kind, though strong, are of brittle material. These must be drilled without too much pressure, or they may crack. Drill points should also be sharp.

The rod is 6 inch long and 3/8 inch in diameter, and it is supported by two strips of insulating material, which are fixed with brackets. Each strip has a V-shaped notch, so that the rod can be held with thin string. L1 is the tuned winding, and it had sixty turns of 24 swg enamelled wire, side by side. Connections can be made to the frame (moving plates) and fixed plates of the capacitor VC1 by means of small bolts, or may be lightly twisted.

L2 is the coupling winding. Many communications type receivers have a low impedance input circuit, and twelve turns will be adequate for L2. The same wire can be used as for L1. With other receivers it may be worth while to experiment with the number of turns on L2, for best reception.

An air-spaced single gang capacitor of about 315pF is most suitable for VC1, and will allow the usual medium wave band of frequencies to be tuned as well. However, a smaller capacitor is suitable if the ferrite aerial is for 160m reception only, or for the high frequency end of the medium wave band. The actual frequencies tuned can be lowered by having more turns for L1, or by using a larger capacitance for VC1, or by having L1 more nearly towards the centre of the rod. Reducing VC1, or the number of turns on L1, or moving L1 to the end of the rod, will raise the highest frequency tuned.

The rod is directional, so the unit can be positioned for best signal pick up of wanted transmissions, or minimum reception of interference.

VC1 has to be adjusted in conjunction with tuning of the receiver and is set for best reception. Tuning with this control should be quite sharp.

A ready made MW transistor type receiver aerial can be useful for the medium wave band, and may in some cases reach 160 metres, if VC1 has a low minimum capacitance, and the winding is quite near the end of the rod. But should tuning not reach a high enough frequency, some turns will have to be removed.

Diode Radio

A lot of interest can be obtained from a diode radio receiver. It requires few components, no battery or other supply, and can give good headphone reception of a few of the more powerful transmitters. It can also be used as a tuner, as shown later, and will provide excellent audio quality for local station reception.

The circuit in Fig. 7 is intended for medium wave reception. That is, from about 200 metres to 500 metres, or 1500 kilohertz to 600 kilohertz.

The coil L1 is tuned by the variable capacitor VC1. At resonance with the wanted transmission, a signal voltage is developed across the coil. Detection or demodulation by the diode D1 allows the audio component of the signal to be

recovered, and taken to the phones or to an amplifier. Such "crystal sets" were used in very many homes, in the early days of radio.

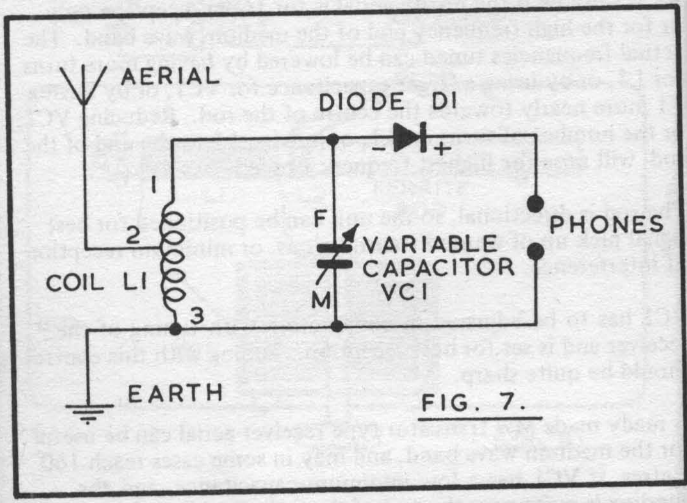

FIG. 7.

Fig. 7B shows the practical layout. It is convenient to use an insulated board larger than necessary, so that a transistor amplifier can be added later. A board about 5 x 8 inch is suggested, and it can be supported on 1 inch high runners (strips of wood) so that there is clearance for screws or terminals.

Coil L1 is wound with 26swg or similar wire, and has seventy turns, side by side. Secure the wire by passing it through two small holes, wind on about thirty five turns, and form a loop. Continue to wind in the same direction for a further thirty five turns, cut off the wire, and secure the end through two holes. The tube used can be 1½ inch to 2 inch in diameter, and can be paxolin, or can be made by winding thin card round a suitable object, smearing a little adhesive on to form a tube.

Point 1, the beginning of the coil, is taken to the fixed plates tag F of the variable capacitor, which is also connected to diode D1. The tapping 2 goes to the aerial terminal. End 3 is

taken to the moving plates connection or metal frame of the capacitor. Ends 1 and 2, and loop 3, are left long enough to reach these points, and the insulation is carefully scraped off the wire where connections are to be made. The capacitor VC1 can best be about 500pF, but this exact value is not essential.

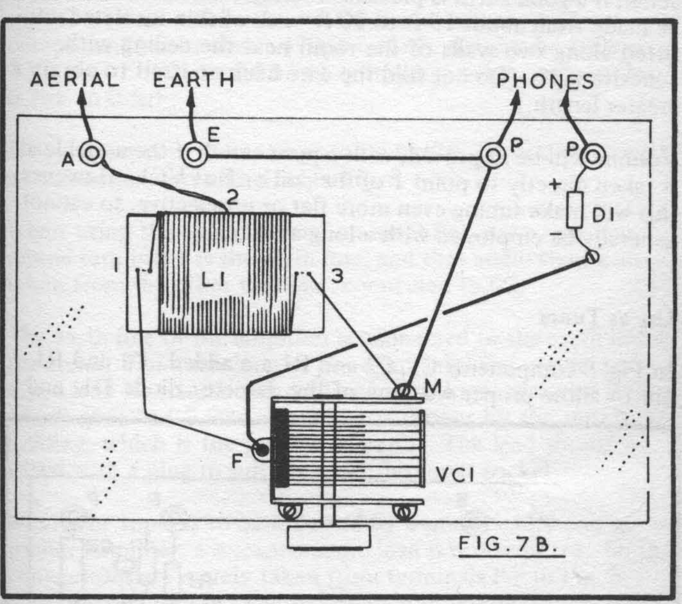

FIG.7B.

Four terminals, for Aerial, Earth, and Phones, will complete the receiver. Best results will be had by connecting a good headset, of about 500 ohm to 2,000 ohm. Very low impedance earpieces, such as are employed with some transistor receivers, are not suitable here.

With a simple receiver of this kind it is necessary to employ a reasonably effective aerial and earth. Only in some circumstances will volume be great enough with no earth, or with a short aerial. The earth connection can be taken to a metal spike in the ground, or to a metal object buried in soil, or to any other available earth other than a mains circuit earth. (No connection must be made to any mains plug or other mains circuit.) A cold water pipe may provide a good earth. Gas pipes should not be used.

For the aerial, a length of aerial wire some 30 ft to 75 ft or so in all may form the aerial itself and lead-in. If this can be suspended quite high, and clear of buildings or earthed objects, signal strength will be improved. However, satisfactory reception may be obtained from an indoor aerial, or shorter aerial, if a good earth is present. A simple indoor aerial can be made from about 15 ft to 20 ft or so of thin insulated wire, fitted along two walls of the room near the ceiling with adhesive tape. (Do not fold the wire back on itself to obtain a greater length.)

Volume will be improved, with a poor aerial, if the aerial lead is taken directly to point 1 of the coil or F at VC1. However, this will make tuning even more flat or unselective, so cannot generally be employed with a long aerial.

Use as Tuner

In Fig. 8 components C1, C2 and R1 are added. C1 and R1 are to allow proper working of the detector diode D1, and

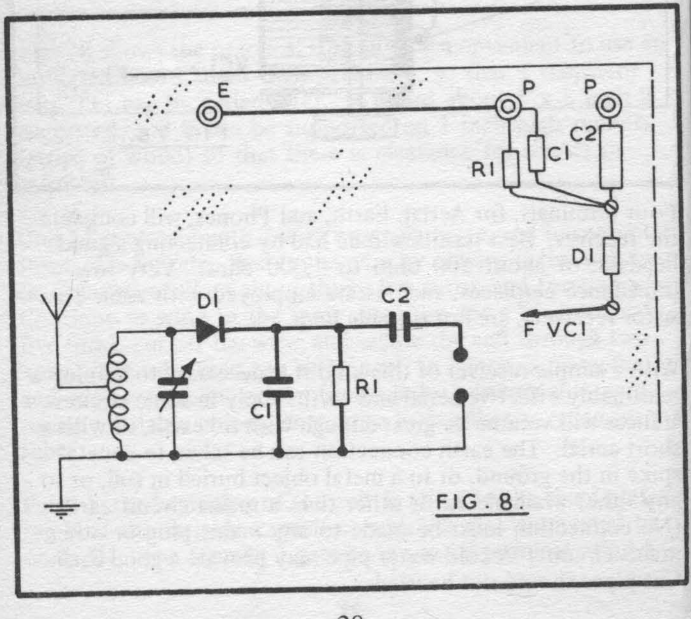

FIG. 8.

may be considered as replacing the headphones. C2 is a coupling capacitor passing audio signals to the following amplifier.

As with the previous circuit, any general purpose detection type diode is satisfactory for D1, such as the OA91 and many similar devices. The value of C1 is not critical, and this component can be about 470pF or 1nF. R1 can be 100k to 220k or so. C2 should be of quite large value, and can be 0.1μF to 0.5μF.

D1 can be moved from the position in Fig. 7B, and the few extra components added, as in Fig. 8.

When using the tuner to feed an amplifier, note that one phone terminal P is the earth line, and that audio signals are taken from the other terminal, connected to C2.

The earth line of the amplifier is connected to the earth line of the tuner. For a sensitive or powerful amplifier, a screened audio lead is recommended. This has an inner screened lead, which goes to C2, and a conductor former by the outer brading, which is the ground or earth. The lead should be fitted with a plug to suit the amplifier input socket.

Where the tuner is to be followed by a small headphone or similar amplifier, a screened audio lead is not required. So then connections are merely taken from terminals P-P in Fig. 8, remembering to join amplifier earth and tuner earth circuits as explained.

Medium and Long Wave

The addition of long wave coverage is worthwhile in some areas, and the coil can be modified as in Fig. 9 so that both MW and LW can be tuned.

A tube about 3 inch long and approximately 1½ inch in diameter will be necessary. The section L1 is wound as before. L2 consists of about two hundred turns of 34swg or 36swg wire, in two compact piles of one hundred turns each. All turns throughout are in the same direction. The junction of L1 and L2 goes to the waveband switch S1. The second

switch terminal goes to 4 and earth line. MW tuning is obtained with S1 closed, and LW tuning with S1 open.

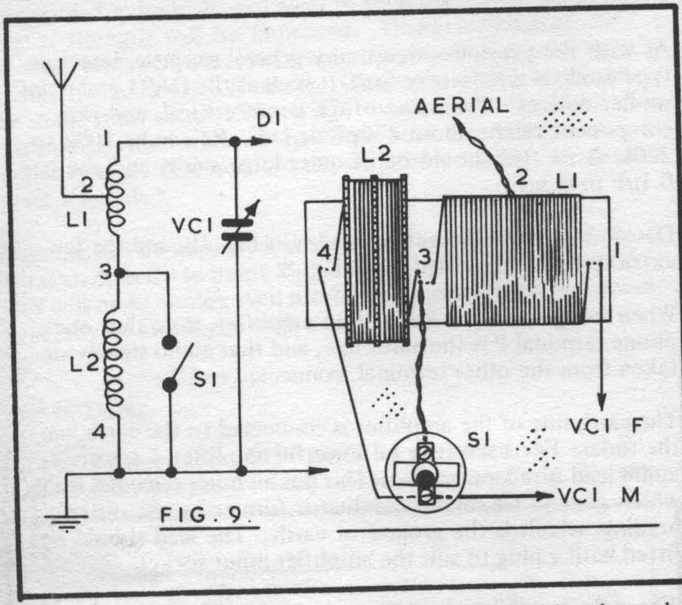

FIG. 9.

For accurate coverage of the usual MW and LW bands, and using a 1¼ inch diameter tube, L1 can be eighty-four turns of 32swg enamelled wire, with tapping 2 sixty turns from point 1. Leave a space of 5/8 inch and glue on a card disc cut to fit the tube. Also glue on two further discs, leaving spaces of ¼ inch between them. L2 is wound with 32swg double cotton covered wire, with one hundred and twenty turns in each slot (making 240 in all for L2). The coil can be fixed with two screws or small bolts, with spacers or extra nuts to hold it a little clear of the board. The windings should not be varnished, covered with adhesive, or wax.

With circuits of this kind there is no ganged tuning, or need for exact tuning of a particular frequency band. This means that coil winding details (wire gauge and covering, and diameter, or number of turns) can be varied to suit materials to hand. Ready-made coils can also be tried. Some commercially made coils with cores can give a slightly better performance than the air-cored home made type of coil. Coils can also be wound in a similar way to that in Fig. 6.

One Transistor Amplifier

Headphone volume obtained from the crystal diode receiver can be considerably boosted by adding a single transistor audio amplifier. Space for this can be found on the circuit board.

In Fig. 10, resistor R2 and the transistor have been added, with a switch S2 for on-off switching of the battery. R1, C1, D1 and C2 are already present, as shown in Fig. 8.

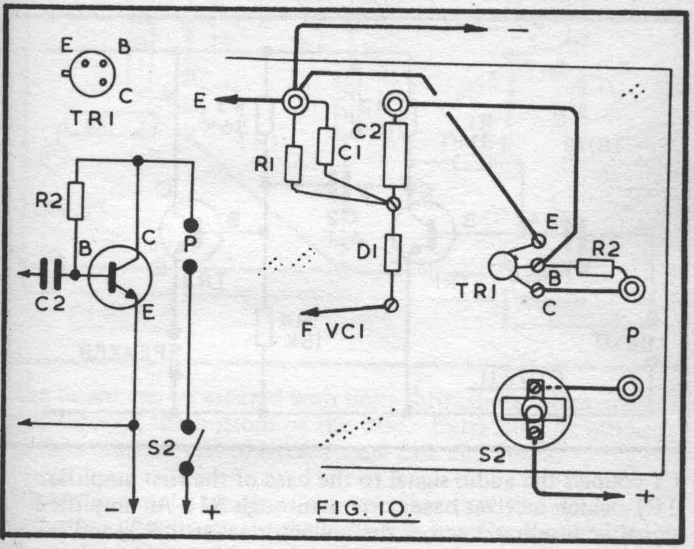

FIG. 10.

TR1, R2 and S2 are connected as shown. For TR1, a BC108, 2N706, or similar general purpose NPN transistor will be satisfactory. R2 can generally be 1 megohm with satisfactory results, but it can be worth while trying values from about 100k to 2.2 megohm here, as performance depends on the transistor, phones, and battery voltage.

Medium and high impedance phones, such as a 600 ohm DC headset, will be most suitable. The battery can be 4.5v to 9v.

If the phones are first connected to terminals P-P in Fig. 8, then taken to terminals P in Fig. 10, a very considerable increase in volume should be obtained.

If a PNP transistor is to be tried in this circuit, the polarity of the battery *must* be reversed.

Two Transistor Amplifier

Fig. 11 is the circuit for an easily made 2-transistor amplifier which can provide reasonable loudspeaker volume. The audio signal is taken to the input terminals, battery negative being the grounded side of this circuit.

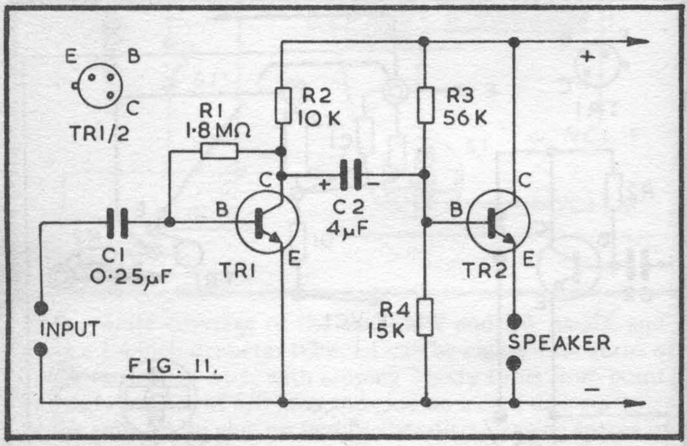

FIG. 11.

C1 couples the audio signal to the base of the first amplifier TR1, which receives base current through R1. An amplified signal is developed across the collector resistor R2, and is coupled to the base of the following amplifier TR2 by means of C2.

The two resistors R3 and R4 set the base operating conditions for TR2, and the output is taken from TR2 emitter.

Fig. 11B shows assembly of the components on a small insulated board. If preferred, connections between components can be on the underside of the board, in the way shown for later circuits. Fit red and black flexible leads for battery connections. The amplifier, with its battery, can be fixed in an insulated box or metal case. Sockets can then be used for input and output circuits, and an on-off switch can be placed in one battery lead.

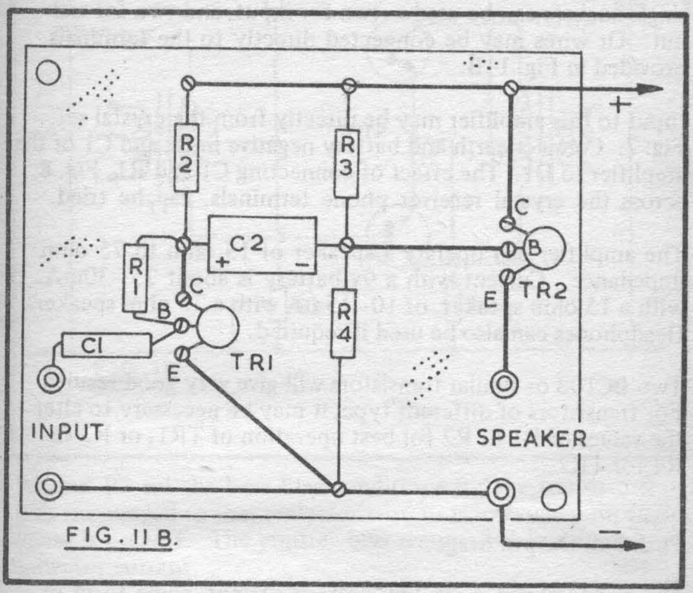

FIG. 11 B.

The board can be secured with bolts through its fixing holes, and through the bottom of the case. Extra nuts or other spacers must be placed between the case and board.

Whenever a metal box is used, remember that all connections and other circuits except the earth line must be insulated from the metal. The metal case itself is normally connected to the earth line, and a lead can be put under one of the fixing screws for this purpose. In Fig. 11, both input and speaker are returned to the earth line. So 3.5mm jack sockets of the type in which the outer or sleeve contact is common to the fixing bush could be used, as these points will be connected to battery negative and the earth line and metal box. But with some circuits it is necessary to use an insulated jack socket, or to insulate this from the metal. An example of this would

arise if the speaker were placed in the collector circuit of TR2, as the outer or sleeve circuit would then need to go to battery positive.

Instead of the type of socket which takes a jack plug, ordinary single sockets can be used – two for input, and two for output. Or wires may be connected directly to the terminals provided in Fig. 11B.

Input to this amplifier may be directly from the crystal set, Fig. 7. Connect earth and battery negative lines; and C1 of the amplifier to D1. The effect of connecting C1 and R1, Fig. 8, across the crystal receiver phone terminals, can be tried.

The amplifier can operate a speaker of 15 ohm to 75 ohm impedance. Current with a 9v battery is about 25–30mA with a 15 ohm speaker, or 10–15mA with a 75 ohm speaker. Headphones can also be used if required.

Two BC108 or similar transistors will give very good results. For transistors of different type, it may be necessary to alter the values of R1 or R2 for best operation of TR1, or R3 or R4 for TR2.

Push-Pull Output Stage

For good volume, with low battery current, some form of push-pull output stage is generally used. This has two transistors, and each amplifies one half of the audio signal. Base bias conditions of the two transistors are so arranged that only a low current flows during silent intervals. Overall battery drain is thus much lower than if a single transistor were used to obtain a similar level of volume.

Fig. 12 is a typical push-pull output stage. T1 is the input or driver transformer, and an audio signal is taken to its primary P. The secondary S is centre tapped. So for one half of the audio cycle one transistor base is driven positive and the other transistor base is driven negative, this being reversed for the remaining half cycle. Thus each transistor conducts alternately, and the centre-tapped output transformer T2 combines the collector currents, to operate the speaker connected to the secondary S.

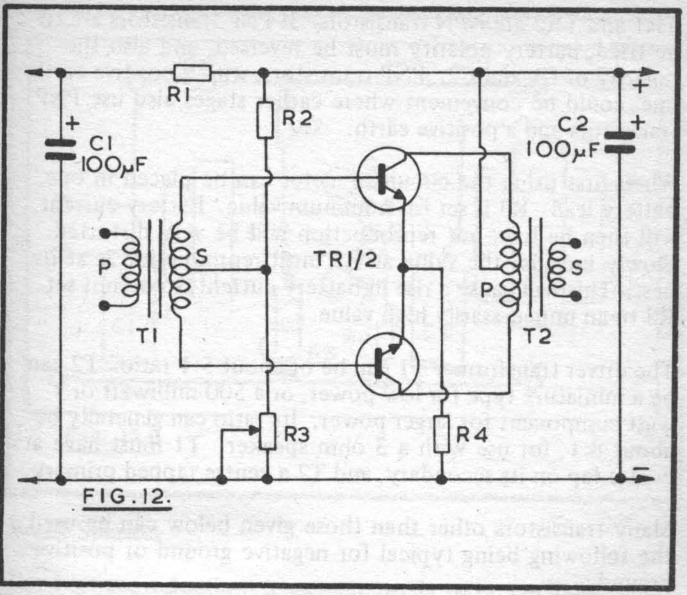

FIG. 12.

R2 and R3 set the base bias conditions for the transistors. Bias is arranged so that collector current is low, when no audio signal is present. The emitter bias resistor R4 limits peak collector current.

If R3 is a pre-set resistor, as in Fig. 12, operating conditions can be arranged to suit the transistors fitted. For best results TR1 and TR2 should be a matched pair. This means that they will have similar gain and current characteristics. However, two transistors of the same type number will generally be satisfactory.

It is assumed that the same battery will supply earlier stages of the amplifier, and current for this purpose is taken from C1 and R1. R1 and C1 prevent audio signals present in the positive line passing back to earlier amplifier stages. The by-pass capacitor C2 also helps in this respect. With simple circuits not giving much overall amplification, or small power outputs, R1, C1 and C2 might all be omitted.

TR1 and TR2 are NPN transistors. If PNP transistors are to be tried, battery polarity must be reversed, and also the polarity of C1 and C2. PNP transistors, with a positive earth line, could be convenient where earlier stages also use PNP transistors and a positive earth.

When first using the circuit, a meter can be placed in one battery lead. R3 is set for minimum value. Battery current will then be low, but reproduction will be very distorted. Slowly increase the value at R3 until reproduction is at its best. This will cause a rise in battery current, so do not set R3 to an unnecessarily high value.

The driver transformer T1 can be of about 5:1 ratio. T2 can be a miniature type for low power, or a 500 milliwatt or 1 watt component for larger power. Its ratio can generally be about 8:1, for use with a 3 ohm speaker. T1 must have a centre tap on its secondary, and T2 a centre tapped primary.

Many transistors other than those given below can be used, the following being typical for negative ground or positive ground.

Negative earth (NPN transistors). 2 x BC108. R1 1.2k. R3 100 ohm fixed or 250 ohm pre-set. R4 4.7 ohm.

Positive earth (PNP transistors). 2 x AC128 or 2 x AC142. R2 4.7k. R3 82 ohm or 250 ohm pre-set. R4 4.7 ohm.

R1 will generally be 1k to 2.2k, depending on the earlier amplifier. The primary of T1 can be connected to terminals P in Fig. 10; or alternatively to any audio source where a reasonably strong signal is present.

Where signals are too strong, an audio gain control needs to be fitted. In Fig. 8 this can be done by omitting R1, and disconnecting C2 from D1. Connect the outer tags of a 100k volume control or log pot from D1 to earth line. Take C2 to the centre tag or slider of the potentiometer. Any required audio level can then be taken off by adjusting the control knob.

Fig. 12B shows the board layout.

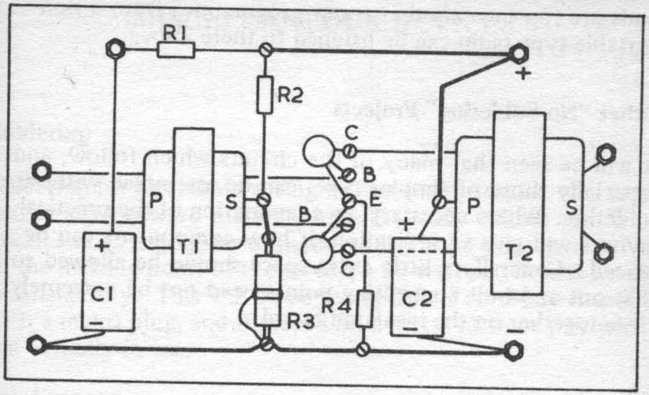

FIG. 12 B.

Boost Speaker

For best reproduction, a speaker needs to be enclosed in a cabinet or fitted to a baffle board. A cabinet can be made from wood, and can be about 6 x 10 inch and 4 inch deep, or as required for the speaker unit. Cut an aperture in the cabinet front to match the cone diameter, and glue thin fabric behind this opening. The speaker can be fixed with nuts and bolts, or short woodscrews driven in from behind.

Connect flexible leads, with separate plugs or a jack plug, to suit the equipment.

Small pocket receivers, which have a miniature speaker, can give much improved volume and reproduction when used with a larger speaker. If a headphone socket, operated from the receiver output stage, is present, it is only necessary to plug the speaker into this. A fairly large speaker, of sensitive type, will give best results. A unit 5 inch to 6 inch in diameter, or about 4 x 7 inch, will be convenient. The speaker has to be of approximately the correct impedance, which should be found in the receiver operating instructions. Where the speaker impedance is wrong, a matching transformer has to be employed between receiver and speaker.

Extension speakers are less frequently used today, but if two leads are run into another room programmes from a non-portable type radio can be listened to there also.

Other "No Soldering" Projects

It will be seen that many of the circuits which follow, and especially those of simpler type, can be assembled without soldering. Where necessary, an examination of the practical layouts will give an indication of how components can be placed. Generally a little extra space should be allowed so that nut and bolt connecting points need not be extremely close together on the insulated board.

SECTION 2
MISCELLANEOUS DEVICES

Soldering

Soldering for electronic wiring will be found to be a very easy and straightforward matter. An electrically heated iron is usual, and as large areas of metal do not have to be raised to the melting point of the solder, a 20 watt to 25 watt iron is satisfactory. This will be perfectly safe if correctly fitted with a mains plug, and if placed on a metal plate or tray when not actually in use.

It is best to use cored solder, as made for electronic wiring.

Leads, tags and other items to be joined should be clean and bright. They have to be raised in temperature to the melting point of the solder at the actual points where a joint is to be made, and the active core of the solder, or flux, must be present here when the joint is being made.

The iron should be tinned – that is, have melted solder on its tip – and have reached its correct temperature. Leads can be hooked or looped into position. The joint is then heated with the iron, and the solder is simultaneously applied to the joint, so that it melts and flows over the surfaces to be soldered. This usually takes place in a second or two. The solder and iron should be removed at once, as excessive heating may damage some components, and especially transistors or other semiconductors. The joint cools in a few seconds, and should be rigid, and solder having flown onto the items to be joined.

A probable cause of poor joints is insufficient heating of items to be soldered, or carrying solder to the joint on the iron, so that the activity of the flux is wasted.

Electronic items such as tags and the wire ends of components are generally designed for easy soldering. They will be copper, tinned copper, and similar metals and alloys which solder readily. For this reason, difficulties are unusual.

Some metals do not solder readily, or need special methods. Aluminium is one of these, and it is not used where popular cored solders intended for copper or general wiring are employed.

Make sure that heat from the iron does not damage furniture or other items, and keep the flexible cord away from the heated tip. Always switch the iron off at once when the work in progress is to be left.

One Transistor Audio Oscillator

This can be used instead of a bell or buzzer warning, or for code practice and similar purposes. It needs very few components, Fig. 13, and can be made to work satisfactorily with a wide range of transistors.

In Fig. 13, an NPN transistor is shown in the theoretical circuit, so battery polarity is as indicated, with negative to emitter. If a PNP transistor is fitted, battery polarity must be reversed.

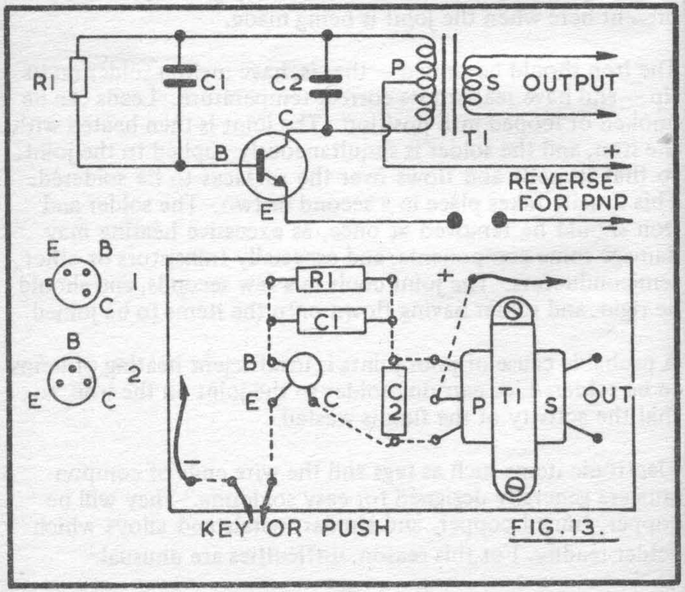

FIG.13.

(1) shows leads from a typical NPN transistor, such as the BC108, 2N706, and similar types. (2) is the base of the PNP AC128 transistor.

T1 can be a small speaker matching transformer, as used for push-pull stages. A centre tapped primary is necessary. The secondary S can be taken to a loudspeaker or headphones.

The values of R1, C1 and C2 can be varied considerably. C1 can be 0.1μF or so. R1 determines to some extent the power obtained, and can be about 22k, increased to 47k or 100k for lower power. The value of C2 has considerable influence on the note, or frequency of oscillation. Increasing the value of C2 lowers the pitch. C2 can be about 47nF to 4.7nF or thereabouts. A 9v battery can be used. Changes to the supply voltage will influence the note produced.

Fig. 13 also shows a layout for components, using either a fully perforated board, or a panel drilled to suit. Transformer T1 is fixed by screws. Small transformers of this type are usually wire ended, or have pins which pass down through holes in the board.

The wire ends of the resistors and capacitors are passed down through suitable holes, and bent over. The board is then turned so that the underside wiring can be carried out. In most places the wire ends of the components will be long enough to reach connecting points. Elsewhere, 24swg tinned copper wire, or similar connecting wire, will be used. Make the soldered joints as described, and snip off unwanted ends of wire. It is as well to leave the transistor to last, as this avoids heating it unnecessarily when soldering other items. Lengths of flex, anchored at holes in the board, are used for battery, key (or push) and speaker connections.

Multivibrator

Fig. 14 is the circuit of the multivibrator assembled in Fig. 2. Reference to Fig. 14B will show how this can also be prepared on a tag board.

The tag board has two rows of tags fitted to an insulated

board, and it is five tags long in Fig. 14B. Boards of this type can be obtained in many sizes. It is also easy to cut a section from a long tag board, to suit the needs of the item to be assembled. Such boards are very handy indeed for electronic assemblies, as the tags support resistors, capacitors and other items, and a neat layout is readily achieved.

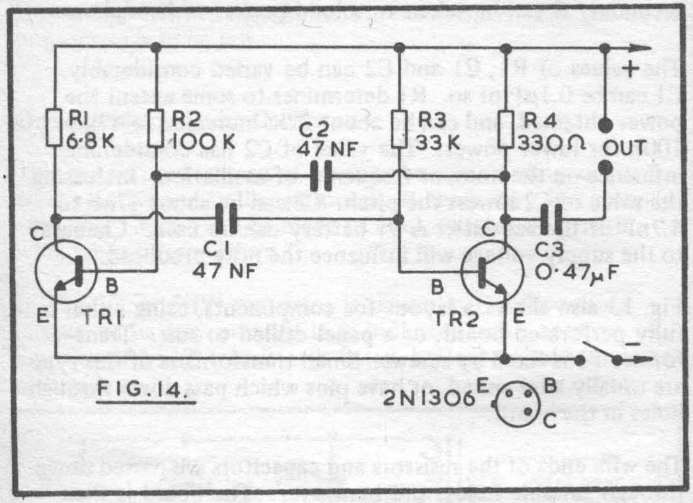

FIG. 14.

Tag strips are also useful for projects, and may have any number of tags required. Long tag strips can be cut down as necessary. Short tag strips have one tag for mounting the strip, longer strips having two tags formed as brackets for this purpose.

Place the components approximately as in Fig. 14B, and solder the wires to the tags. If any component values may be changed, do not twist the leads tightly round the tags before soldering, or removal will be difficult.

Wires which cross can be kept clear of each other, and then no insulated sleeving will be necessary. The transistors can be soldered on last, and can readily be changed without disturbing other parts.

The same layout can be used for the next project, though the component values are different.

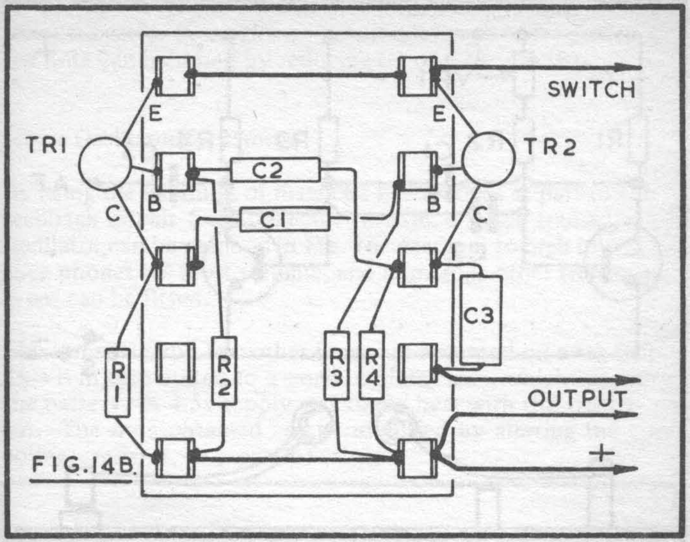

FIG. 14B.

Glissade

This device will produce swooping, rising and falling tones, for "effects" purposes, or can with additional skill be used to obtain musical notes through several octaves. The circuit is shown in Fig. 15, and reference can be made to Fig. 14B for a practical layout on a tag board.

In Fig. 15, suitable component values are:

R1	10k	C1	22nF
R2	12k	C2	22nF
R3	100k	C3	22nF
R4	10k	VR1	100k linear pot.

Transistors: 2 x BC108 or similar types.

The circuit is a wide frequency audio oscillator, in which the pitch can be controlled by VR1, and "on" switching is obtained by S. The oscillator is placed in a box, so that switch S and VR1 can be mounted on the top. The switch is a strip of springy metal, fitted with a spacer and bolt, so that it is just clear of a further bolt. A small downwards movement of the arm or knob completes the circuit.

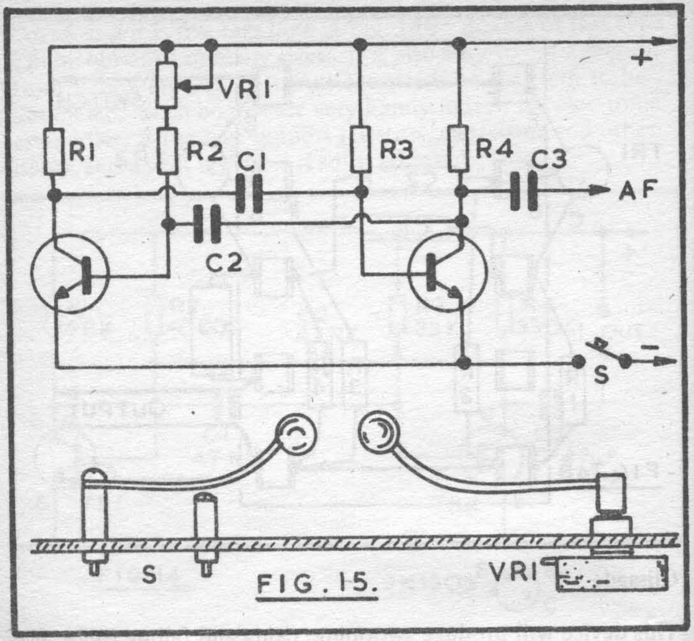

FIG. 15.

VR1 has a ¼ inch shaft coupler, into which is screwed a length of 4ba threaded rod. A knob is also fitted to the end of this rod. It can be swung backwards or forwards, to change the resistance of VR1. Connect VR1 so that movement of the knob towards the right raises the pitch.

Output from C3 can go to high resistance headphones, for amusement, but should be taken to an audio amplifier. Gain of the amplifier will have to be turned back somewhat, as the output from the glissade is at a quite high level.

Swooping tones are produced by holding down the switch S and swinging the control knob attached to VR1 from side to side. For notes which do not swoop to frequency, switch S must be released while VR1 is rotated. A slight quivering motion of the hand controlling VR1 will produce a vibrate effect. It is useful to have an ordinary switch in parallel with the knob control S, and to have the amplifier audio gain control near the left hand, so that fading effects can be produced as well.

About 4.5v to 9v can be used, this influencing volume. Should other transistor types give an unsuitable range of frequencies, the note can be raised by reducing C1 or C2 (or both).

Morse Oscillator for Phones

By using the windings of magnetic headphones as part of the feedback circuit from collector to base, a single transistor oscillator can be made as in Fig. 16. Medium to high impedance phones are most suitable, and numerous other transistor types can be fitted.

The transistor and few other items are mounted on a tag strip. This is in turn bolted to a small insulated box, which can carry the battery. A 4.5v supply was found best with this oscillator. The note obtained can be modified by altering the voltage, resistor, or capacitor values.

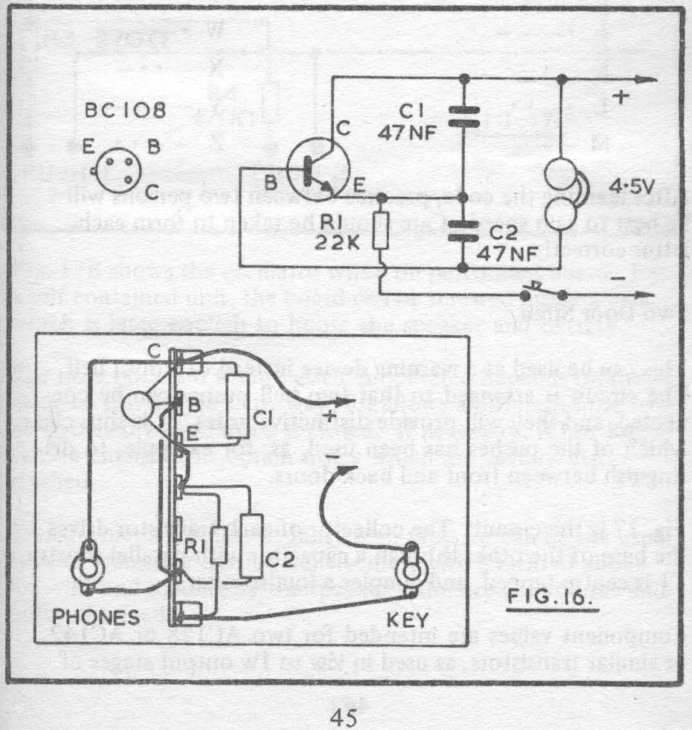

FIG.16.

For serious code practice a good Morse key is required. It is adjusted so that only a small movement of the wrist is needed to operate it. Each letter consists of short and long sounds, shown as (•) and (–). The (–) has three times the length of the (•)

A •–
B –•••
C –•–•
D –••
E •
F ••–•
G ––•
H ••••
I ••
J •–––
K –•–
L •–••
M ––
N –•
O –––
P •––•
Q ––•–
R •–•
S •••
T –
U ••–
V •••–
W •––
X –••–
Y –•––
Z ––••

After learning the code, practice between two persons will be best to gain speed. Care should be taken to form each letter correctly.

Two Door Siren

This can be used as a warning device instead of a door bell. The circuit is arranged so that two bell pushes can be connected, and they will provide distinctive notes. It is thus clear which of the pushes has been used, as, for example, to distinguish between front and back doors.

Fig. 17 is the circuit. The collector of each transistor drives the base of the other through a capacitor with parallel resistor. T1 is centre tapped, and couples a loudspeaker.

Component values are intended for two AC128 or AC142, or similar transistors, as used in ½w to 1w output stages of receivers and small amplifiers. Other transistors will of course operate in this circuit. The pitch obtained can be modified by changing C3. The circuit shown is for PNP transistors, and battery polarity has to be reversed for NPN transistors.

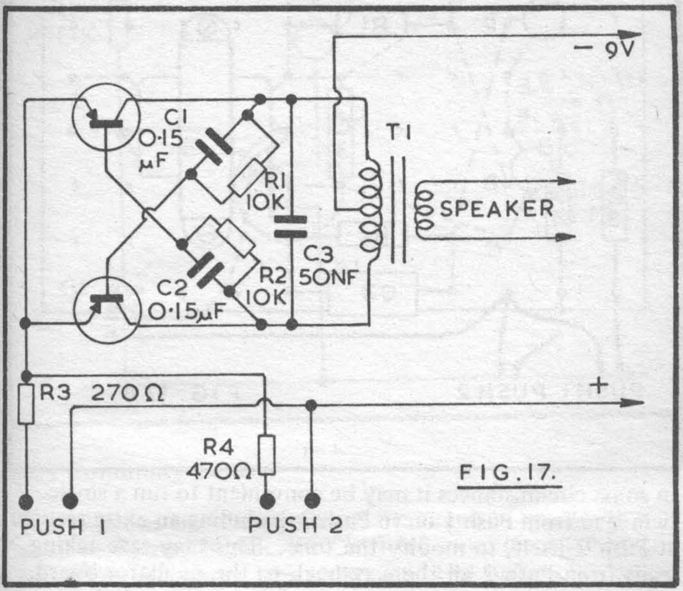

FIG. 17.

Fig. 17B shows the oscillator wired on perforated board. For a self contained unit, the board can be screwed inside a case which is large enough to house the speaker and battery.

The note obtained when Push 1 is operated depends on the value of R3. In the same way, R4 allows a different note to be obtained when Push 2 is closed. If necessary, R3 or R4 can be changed, to obtain notes which are readily distinguished.

Transformer T1 is best of fairly high ratio, or can be the usual type of transistor output transformer, with a ratio of about 8:1. T1 and loading by the speaker influence the volume and notes obtained.

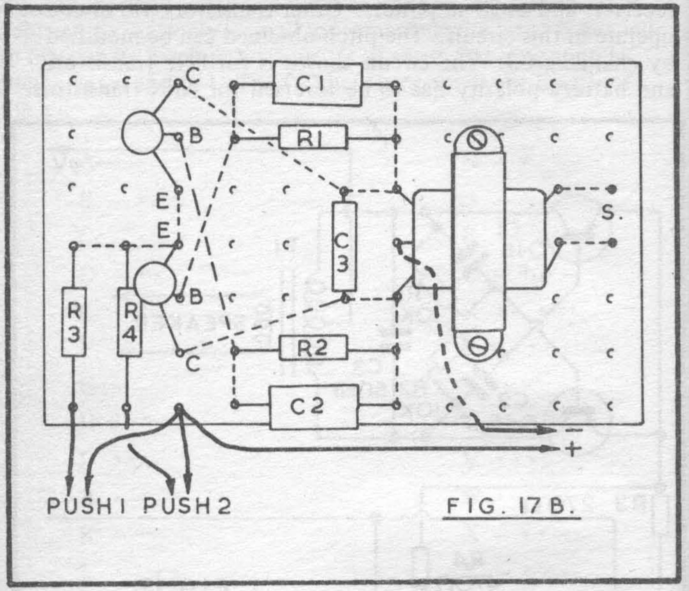

FIG. 17B.

In some circumstances it may be convenient to run a single twin lead from Push 1 on to Push 2, including an extra resistor at Push 2 itself, to modify the tone. This may save taking leads from Push 2 all the way back to the oscillator board.

UJT Oscillator

The uni-junction transistor is a device having emitter, Base 1, and Base 2, Fig. 18. B2 receives positive bias, in this case through the headphones. When forward bias on the emitter E is reached the UJT conducts.

In Fig. 18, the emitter potential rises as C1 charges through the resistor R1. When the UJT conducts, C1 is discharged, and the process is repeated. This can take place over a very wide range of frequencies, depending on the values used for R1 and C1. Here, these are chosen to produce an audio tone.

A battery supply of 3v to 9v will be convenient. For use of the UJT oscillator during code practice, place the key in one battery lead.

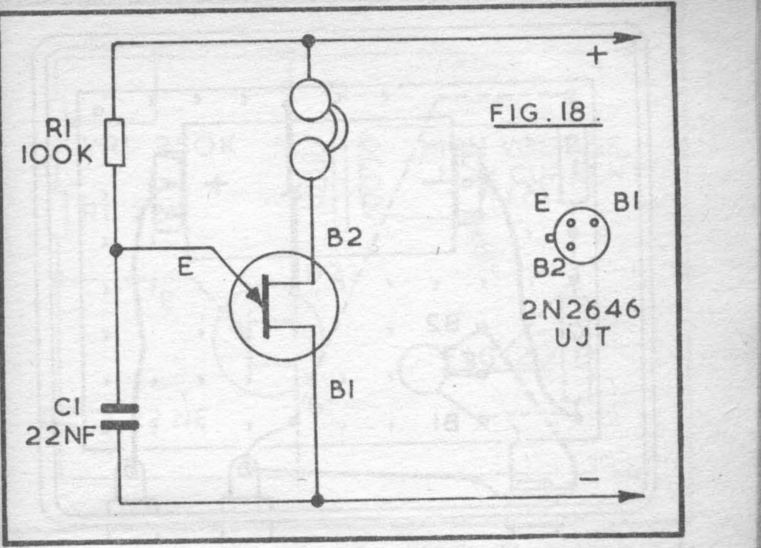

FIG. 18.

UJT Continuity Tester

Fig. 19 shows actual assembly of the circuit in Fig. 18 on a small perforated board. This is fitted in an insulated box, which carries a phone jack socket, and two sockets for Morse key or test leads. A 9v battery is used, secured with string or elastic.

For continuity tests, take prods or clips on the test leads to the suspected item. If the circuit is in order, the UJT oscillator will produce an audio tone. Oscillation will be obtained with a series resistance between the prods of up to 20k or more, so various components can also be tested, and some idea of the resistance in circuit can be gained from the sound obtained. The tester can also be used to check bulbs, lamps, fuses, extension and other leads, and similar items – these must of course be removed or disconnected from the equipment, before being tested.

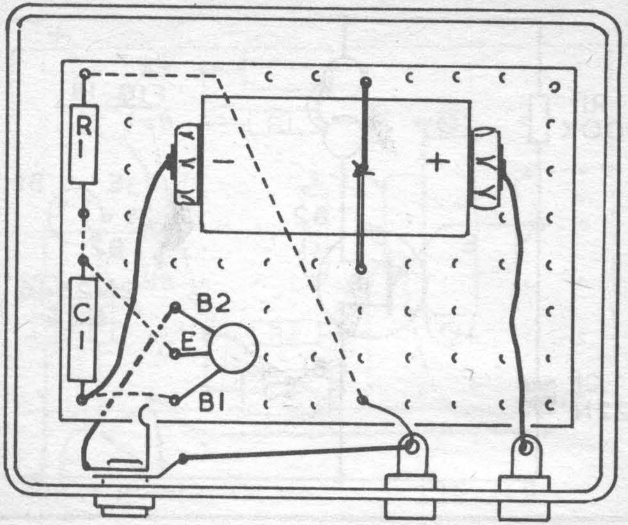

PHONES FIG. 19. KEY OR TEST LEADS

UJT High Voltage Producer

The UJT oscillator can be used to produce a high voltage, at very low current, by connecting a transformer as in Fig. 20. A loudspeaker transformer, as would be used in valve equipment, is suitable, and will have a ratio of about 45:1 or so. The usual secondary, or low-resistance winding, is connected to B2 and positive, so that a step-up of voltage is obtained. This should be found to light a small neon lamp.

VR1 is fitted to allow some adjustment, to find the most suitable frequency of oscillation. For maximum possible high voltage output, if required, operating voltage and current should be raised. No rating of the UJT must be exceeded. For the 2N2646, the supply should not exceed 30v, emitter current should not be over 50mA, and dissipation of power in the device should not be more than 300 milliwatts. (Dissipation can be found from Voltage x Current.)

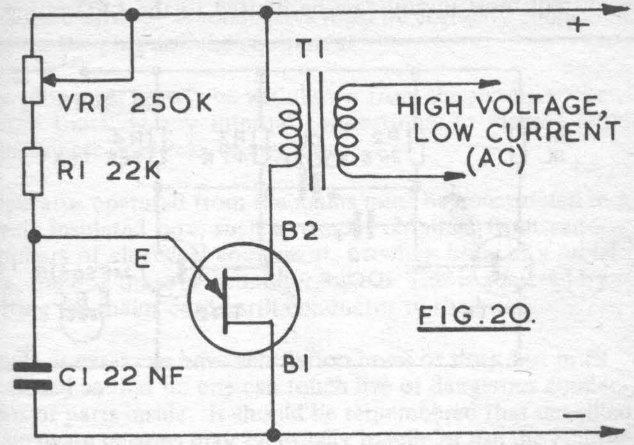

FIG.20.

A small mains transformer can be used for T1, with a low voltage winding, such as that for a 6.3v heater circuit, as primary, and the 200/250v winding as secondary. The high voltage output is alternating current.

Signal Blinker

Fig. 21 is a multivibrator in which the feedback capacitors C1 and C2 are of high value, so that conduction alternates between TR1 and TR2 at a low rate. A light emitting diode is placed in the collector circuit of TR2, and so flashes on and off.

This device can be used for various instruments and other equipment, and a flashing indicator draws attention to a greater extent than does one continuously illuminated. The LED can be located at any required point, with leads running to it.

Transistors other than the types shown can be utilised. R4 can usually be about 150 ohm for a 3v supply, 470 ohm to 1.2k for 6v, or 680 ohm to 1.5k for 9v. This depends to some

extent on the amount of light required when the indicator is operating. Lower resistor values at R4 will increase light, but current must in any case be limited to the LED rating.

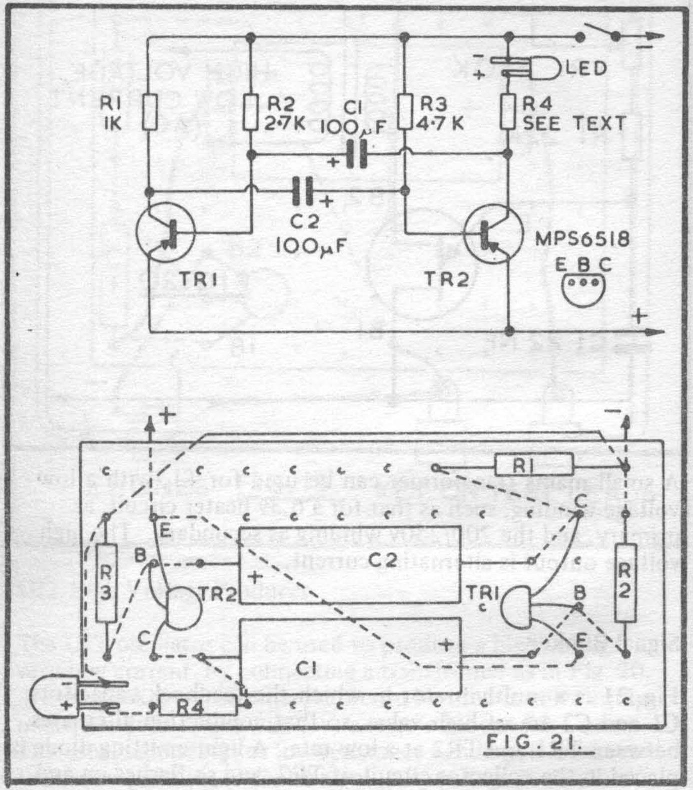

FIG. 21.

C1 and C2 need only be 6.4v working, and the device can be assembled on a very small perforated board. See p.61 for LED polarity.

Mains Wiring

All the devices described so far have operated from low voltage batteries, so than no shock hazard exists with them. However, it is necessary to employ mains power for some units, and proper care must then be taken to avoid any danger.

Power should be drawn from a 3-pin plug, and this should have a 3A or other low rating fuse installed. The live or L, neutral or N, and earth or E conductors must be correctly connected, both at the plug and the equipment.

The plug must always be withdrawn from the supply socket, before touching any internal connections, or making any adjustments or checks of any kind.

Apparatus operated from the mains must be constructed in a strong insulated box, such as can be obtained from various suppliers of electrical equipment; or when built in a metal box, the box must be soundly earthed. This is achieved by bolting the mains cord earth conductor to the box.

Boxes or cases can have ventilation holes or slots, but must be closed so that no one can touch live or dangerous connections or parts inside. It should be remembered that unskilled or unaware persons may eventually handle or use the equipment, so it must be arranged to present no unexpected hazards such as live terminals or parts, or openings for the fingers.

If these points are remembered, mains operated equipment can be safe to use and handle. No switch or other means of interrupting the circuit is put in the earth conductor, and any switch or equivalent device is placed in the live or L conductor. This leaves both earth and neutral permanently connected, for maximum safety.

Dimmer/Heat Control

It is often convenient to dim a filament lamp, or reduce the heat of a soldering iron to be left on for long periods, or the power of a low wattage heater. The silicon controlled rectifier circuit in Fig. 22 can be used for these purposes.

When the switch marked "Full" is closed, current for the load passes through this, so that full power is obtained.

When this switch is opened, current passes through the SCR. This device conducts only when the gate has been triggered, and the instantaneous point in the alternating current cycle

where this arises is determined by the setting of the potentiometer VR1. Triggering is advanced as the wiper of VR1 is moved towards R1. The average power supplied to the load can thus be adjusted between wide limits. Since the device is operating as a half wave controller, full power is not obtained with VR1 at maximum, so the switch is closed as described.

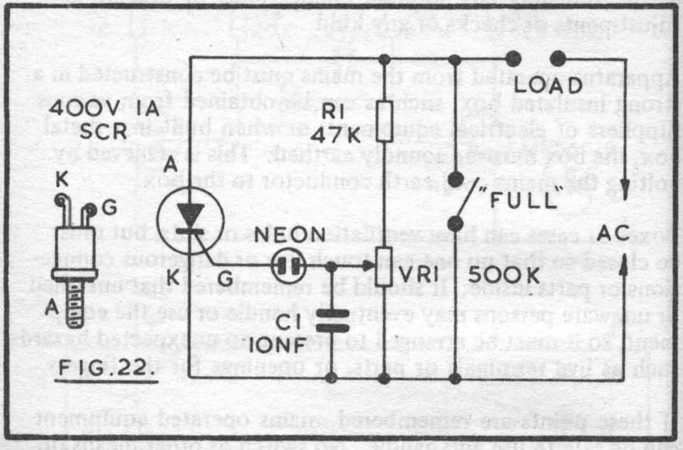

FIG.22.

Fig. 22B shows construction of the SCR controller in a metal box. The front of this carries a socket outlet, and the mains cord has a 3-pin plug, wired as explained earlier. It is thus easy to plug the controller into an outlet, and insert the plug of the lamp or other item to be used in the new outlet.

The mains cord is secured with a clamp, and passes through an insulated grommet. The tag strip supporting the small items is mounted by an unused tag. A strip of metal is cut to clamp round the body of the SCR, and is soldered to one tag, as shown. Run earth and neutral leads from the cord up through a hole, and connect them to the E and N terminals of the socket outlet. Also earth the case, at E. A well insulated wire runs from the L terminal of the outlet, to the "Full" switch. The outlet can be secured with bolts and nuts.

R1 should be 1 watt, and C1 is a 500v capacitor. The neon is a small type as used for indicators, etc.

After checking that the controller operates correctly, the back of the box should be screwed in place. A suitable size is about 4 x 5 x 2 inch deep.

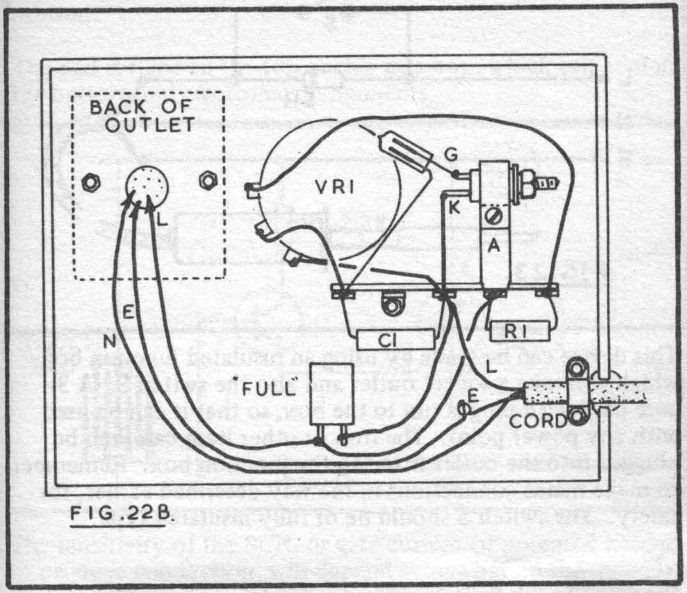

FIG.22B.

If a fully insulated box is used instead, then the bush of VR1 and of the switch should both be earthed by wires, except when these items are constructed wholly with insulated material. (This precaution is necessary to avoid any chance of metal parts outside the box becoming "live".)

Two Power

A circuit which will allow half or full power can be useful to reduce the heat of a soldering iron, or for dimming a filament lamp.

Fig. 23 is an easy circuit of this type. Full power is obtained when the switch S is closed. With S open, the silicon rectifier allows current to flow in one direction only, so that power is approximately halved. A 600v 1A or 1000v 1A rectifier is suitable.

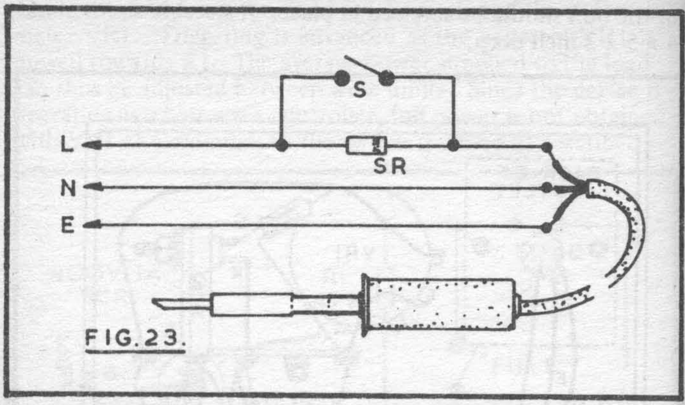

FIG.23.

This device can be made by using an insulated junction box which will take a socket outlet and also the switch S. A 3-core cord with plug is run to the box, so that it can be used with any power point. The iron or other item can then be plugged into the outlet fitted to the junction box. Remember to make mains connections in the way described earlier, for safety. The switch S should be of fully insulated type.

Sensitive Touch Switch

In Fig. 24, finger contact with the touch pad switches on a low voltage bulb. This circuit is suitable for a battery operated portable bedside light, and for some other devices.

Normally the silicon controlled rectifier is not conducting, so the lamp is extinguished. Finger contact with the pad provides positive base bias for the transistor so that it conducts. This moves the transistor emitter and SCR gate positive, triggering the SCR. Current then flows from cathode K to anode A, and through the bulb. The SCR remains in the conducting condition until the battery circuit is interrupted by opening the switch. A push switch, normally closed, is most suitable here. Pressing it then switches off the bulb.

This touch pad is easily made from a piece of circuit board with foil strips. It can be about 2 x 1 inch, or of any convenient size. Join together alternate foils – e.g., 1, 3, 5 and 7 – by soldering a connecting wire through the holes. This will provide one connection to the pad. The other foils – 2, 4 and 6 – are similarly joined together for the other pad connection.

The pad is fixed to the top or side of a box, which can contain the battery and electronic components.

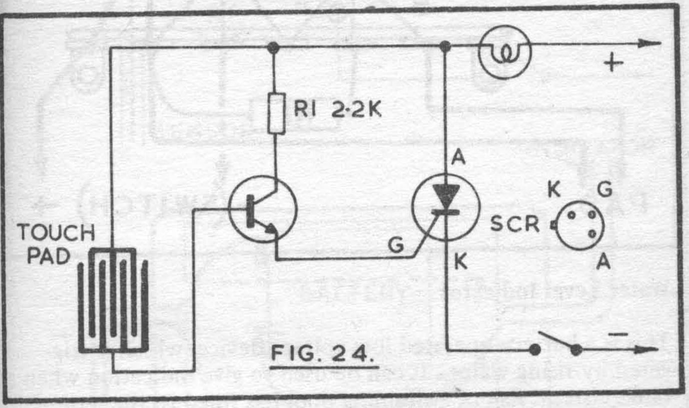

FIG. 24.

The sensitivity of the SCR, or gate current or potential needed to produce conduction, will depend somewhat on the individual device, while the gain of individual transistors will also be different. Numerous NPN types are suitable. If it is found that the switch is too sensitive, this can be corrected by placing a resistor from transistor base to SCR cathode or negative line. This resistor can be about 470k to 1 megohm or so. Excessive sensitivity will be apparent if the circuit is switched on by dust, damp air, or similar effects. The pad should be kept clean by wiping it with a dry cloth.

Fig. 24B shows assembly of the components on a tag strip. A small 50v 1A SCR is more than adequate. The bulb can be a screw in type, and can be 6.3v 0.15A, or 6.3v 0.3A. A 6v 3W (0.5A) bulb is also suitable, provided the battery is fairly large.

It is convenient to make a wooden box which will take the battery, and leave a little space for the tag strip assembly.

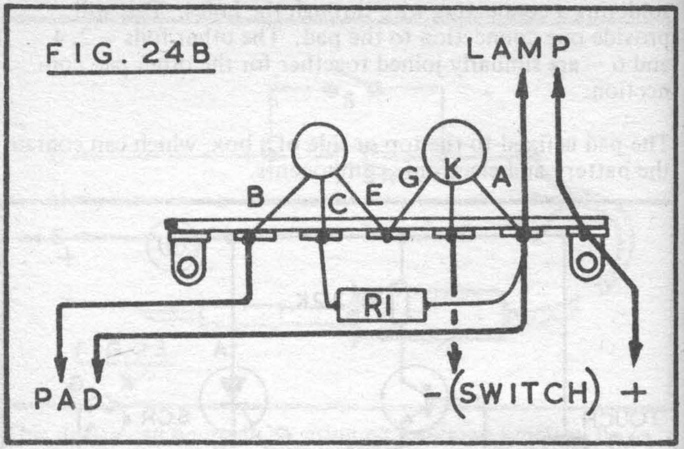

FIG. 24B

Water Level Indicator

This is a battery operated low voltage device, which is triggered by rising water. It can be used to give indication when a tank, cistern, fish or swimming pool has filled to the required level.

Fig. 25 shows the circuit, and also assembly on an insulated board. This also carries the battery, and it is fitted in a metal or plastic box.

In use, the base normally has no positive bias, as the sensor passes no current. When water reaches the sensor, it conducts, and the transistor passes collector current, to operate the indicator. R1 limits base current.

This circuit will operate satisfactorily with a wide range of PNP or NPN transistors, but battery polarity must be reversed for the PNP types.

The indicating or warning device can be one of the multivibrator or other audio oscillators shown earlier. These will give an audible warning, and will draw only a small current, so that a large transistor is not required.

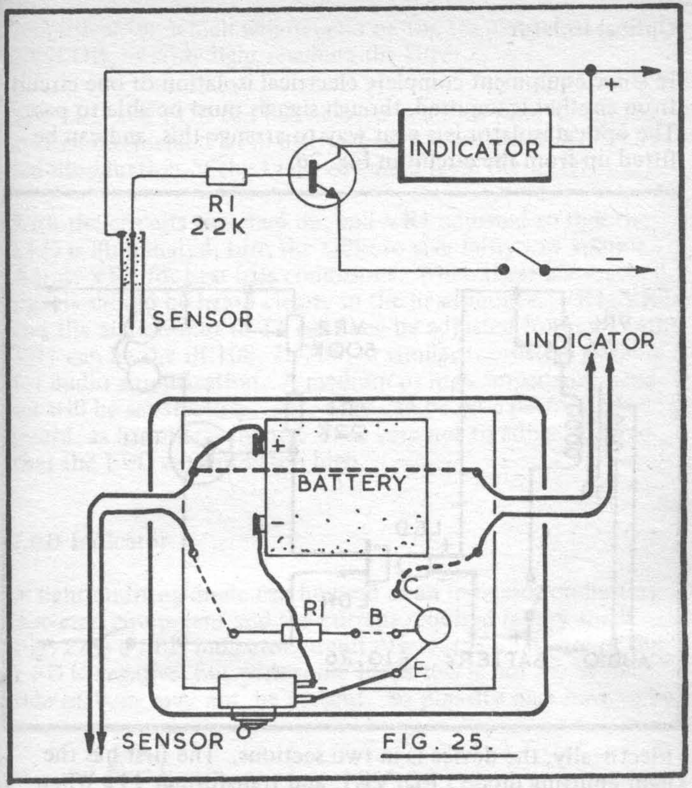

FIG. 25.

For indoor use or in a location protected from rain, the sensor can be two bare wires, about ¼ inch apart, bolted to a strip of insulated material. Where rain could operate the device, it needs to be covered. Alternatively, it can be made from a piece of metal tube, with a central wire, as in Fig. 25. The wire is bound with insulating tape for about half its length, to fit tightly in the tube. Flexible leads are soldered on, and the joints covered with tape. The sensor can be suspended inside the tank or pool, with its bottom at the required water level.

If the transistor is a high gain one, so that damp operates the indicator, sensitivity can be reduced by connecting a 100k resistor from base to emitter.

Optical Isolator

In some equipment complete electrical isolation of one circuit from another is required, though signals must be able to pass. The optical isolator is a neat way to arrange this, and can be fitted up from the circuit in Fig. 26.

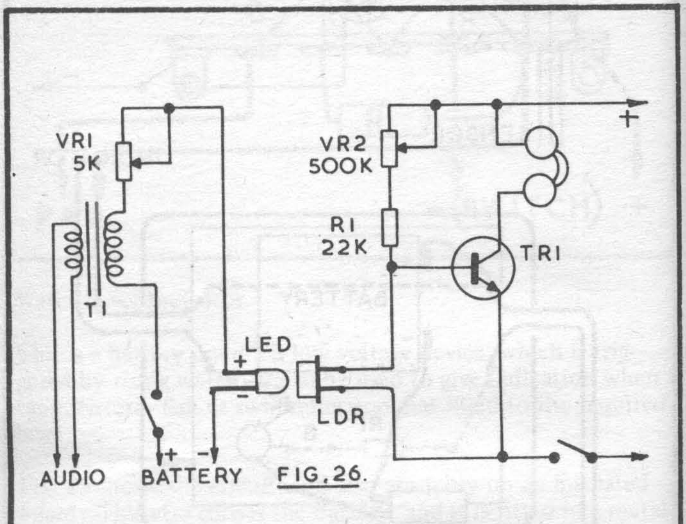

FIG.26.

Electrically, the device is in two sections. The first has the light emitting diode LED, VR1, and transformer T1. When the battery is switched on, current flowing through the LED is adjusted by VR1, so that the LED glows dimly. A signal voltage in the secondary of T1 then modulates the LED light intensity.

To operate this circuit from a radio having a 2/3 ohm or similar speaker, a speaker matching transformer with a ratio of about 45:1 can be used. (The actual ratio is not very important.) Connect the usual transformer secondary to the radio speaker, so that T1 provides a step up effect.

The second part of the equipment has the light dependent resistor LDR, forming part of the network providing TR1 with base bias. VR2 allows adjustment of conditions for best amplification, which will depend on the transistor, and also the LDR, or stray light reaching the latter.

The LED and LDR can be enclosed in a card tube, or should at least be shielded from light, which will swamp the modulated illumination of the LDR, obtained from the LED.

With the circuits switched on, and VR1 adjusted so that the LED is illuminated, turn the radio to give fairly low volume. Adjust VR2 for best bias conditions. When these are reached, signals should be heard clearly in the headphones. VR1, VR2 and the audio input to T1 can then be adjusted for best results. TR1 can be the BC108, 2N706, or similar transistors suitable for audio amplification. A medium or high impedance headset will be satisfactory. Assembly can be on a perforated board, as for other circuits. Take care not to adjust VR1 so that the LED current is too high.

LED Indicator

A light emitting diode can be used as an indicator on battery powered equipment and the current required is very small. Fig. 27 is a LED indicator circuit. Here, the long wire of the LED is negative, but with some LEDs this is not so. A flat side may, or may not, be present. So polarity may have to be

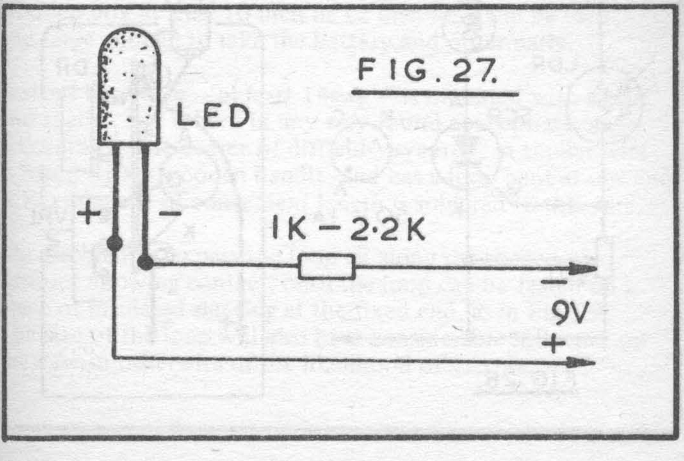

FIG.27.

found by trial – the correct connections are those allowing the LED to light. The series resistor may be about 150 ohm to 1k or more, depending on the actual voltage, and brightness required. Current drawn varies with different LEDs, and with voltage and series resistance, but can be about 5mA or less, without reducing the brightness too far.

The LED indicator with resistor is connected to the circuit where voltage is present when the equipment is on. LEDs are available in red, yellow, green and other colours, with grommets or clips for panel mounting. Different circuits may be identified by a range of colours.

"Magic Candle"

The triggering of a silicon controlled rectifier can be used for slave photoflash and similar purposes. In Fig. 28 it is employed to switch on a light when a match or lighter is brought near. Sensitivity is best in somewhat subdued light, and it will then be found that striking a match at some distance can bring on the light.

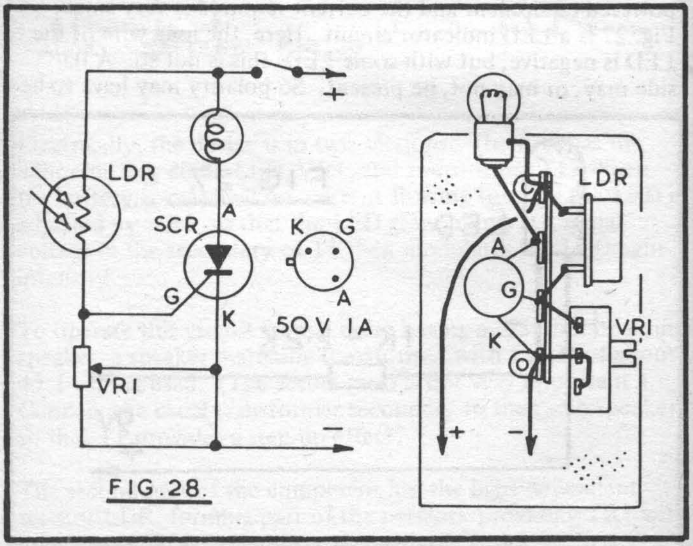

FIG.28.

The LDR has a high resistance when in darkness or subdued light, so that the gate potential of the SCR is not sufficiently high to trigger this device. When light reaches the LDR, its resistance falls, and the SCR gate voltage rises. The SCR then moves into its conducting condition, so that current flows through the bulb. The sensitivity of the circuit, or level of illumination needed to trigger the SCR, is set by VR1.

Components can be assembled on a small tag strip. This, with a battery, can occupy a tubular case, with the lamp on top. An opening is necessary opposite the LDR, and a hole is also made so that VR1 can be adjusted with a screwdriver. Here, a miniature pre-set 50k or 100k linear potentiometer is most suitable. It is necessary to have a small push switch to interrupt the battery current, as the SCR remains in the conducting condition until the circuit is momentarily broken. A 6V battery and bulb can be used. SCRs may fail to operate at all if attempts are made to use them with extremely low voltages, such as a single 1.5v cell.

"Nerve Tester"

This device tests the nerve, concentration, and steadiness of hand, and always provides amusement.

Connections and construction will be seen from Fig. 29. A wooden box at least 10 inch or 12 inch long will be used, and large enough to take the battery and other parts.

A stout bare wire – at least 14swg – is mounted with a bolt and spacer, and is bent in any way found appropriate, and according to the degree of difficulty wanted. A similar wire is fitted with a wooden handle, and has a loop bent at one end. A flexible lead of convenient length is soldered to this wire.

The problem is to pass the loop all along the shaped wire, without allowing contact, until the loop can be rested on a piece of insulated sleeving at the fixed end, as in Fig. 29. The size of the loop will also have considerable influence on the ease or otherwise of the likelihood of success.

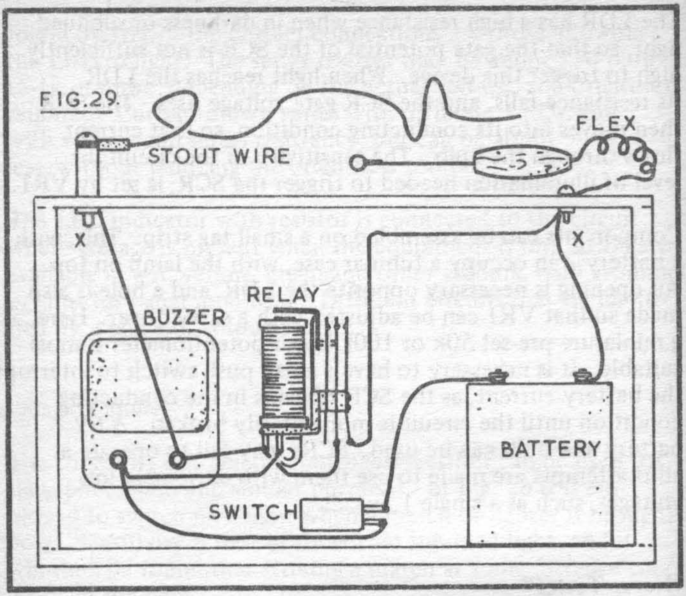

FIG.29.

If loop and wire touch, current flows from the battery to the buzzer, which sounds. Current also flows to the relay winding, so that the relay contacts close. Even a momentary contact between loop and wire will thus cause the buzzer to sound until the switch is used to break the circuit.

Any 4.5v, 6v or similar battery and buzzer will be satisfactory, with a relay which operates on a similar voltage. Connections are to the relay contacts which remain open until the relay is energised.

In use, the device can stand upright. A back is attached with small screws.

Burglar Alarm

The same circuit arrangement as in Fig. 29 will provide a burglar alarm operated by pressure mats. The latter are available in various sizes, and consist of a thin pad which can be placed under a carpet or ordinary mat.

Leads from the mat or mats will run to the terminals at X–X. The buzzer should be changed for a bell, which need not be fitted in the box, but can be situated where it is best heard. Momentary weight on any of the mats will then switch on the bell, which will be locked on by the relay as described, and will ring until the unit is switched off.

Transistor Tester

A very useful check of transistors can be made with the aid of a multi-range test meter, with a selection of base resistors to provide base current, as in Fig. 30. This method does not give an exact indication of transistor gain, but it does allow defective transistors to be eliminated, and provides a good idea of the performance to be obtained from surplus or other unknown transistors.

A 6-way switch allows the base to be open circuit, or to receive current through one of the resistors R1 to R5. These values are R1 330k, R2 110k, R3 33k, R4 11k and R5 3.3k. After allowing for the base-emitter potential, switch positions and corresponding base currents are near the following: 1, zero; 2, 10μA; 3, 30μA; 4, 100μA; 5, 0.3mA, and 6, 1mA.

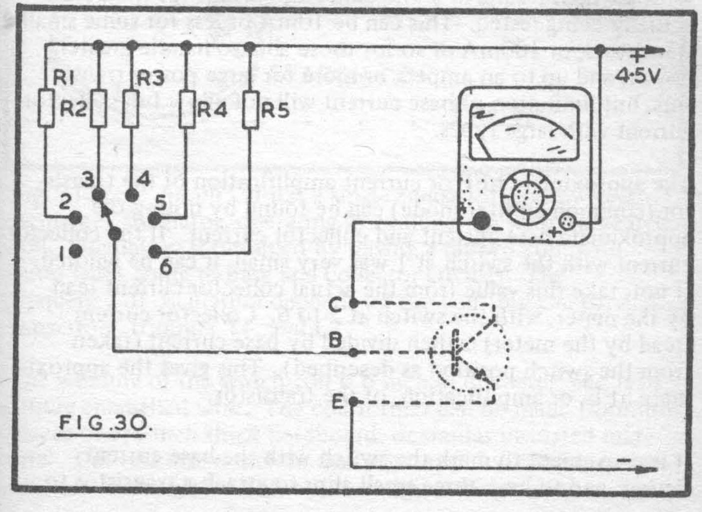

FIG.30.

Polarity is shown for NPN transistors. Reverse polarity of both meter and battery for PNP transistors.

For all tests, initially set the switch to 1, and the meter to its highest range (say 1 ampere). Afterwards, switch the meter to a lower range, after adjustment of the 6-way switch, as may prove to be necessary to obtain an easy reading.

With the switch at 1, a high reading on the meter shows that the transistor has a short circuit from emitter to collector, and is defective.

With germanium transistors, collector current (with the switch at 1) may be 0.1mA or so up to perhaps 0.5mA for small types, and up to some milliamperes for larger types. This small leakage current does not show that the transistor is defective.

In the case of silicon transistors, the leakage current (with switch at 1) will be zero or extremely small, if the device is in good condition.

The switch is then turned to one of the positions 2 to 6. If collector current does not rise, the transistor is defective. If current rises at each step, do not go beyond the switch setting which results in a safe collector current for the device actually being tested. This can be 10mA or less for some small transistors, or 100mA or so for those able to handle greater power, and up to an ampere or more for large power transistors, but limitation of base current will not allow full collector current with large types.

The approximate hFE or current amplification of the transistor (common emitter mode) can be found by noting the approximate base current and collector current. If the collector current with the switch at 1 was very small, it can be ignored. If not, take this value from the actual collector current read by the meter, with the switch at 2 to 6. Collector current (read by the meter) is then divided by base current (taken from the switch position as described). This gives the approximate hFE, or amplification, of the transistor.

It is convenient to mark the switch with the base current figures, and to have three small clips to attach a transistor to

the circuit. Switch, clips, and battery can all be fitted to a small box. A development of the circuit would be to include a current meter, with switch and shunts, to replace the external general purpose test instrument.

Treasure Locator

This locator is among the simplest that can be constructed, as it uses an ordinary small portable radio as part of the detection circuitry. No modifications or connections to the radio are needed.

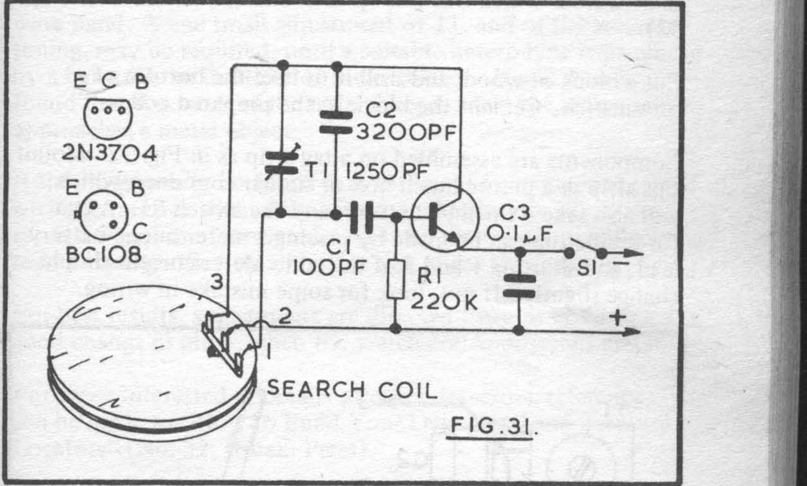

FIG.31.

Fig. 31 is the circuit. It has a single transistor as a low radio-frequency oscillator, operating in conjunction with the search coil. Oscillation is obtained by having a tapped coil, with outer leads going to base and collector circuits. The frequency of oscillation depends on the coil, and the parallel capacitors, trimmer T1, and C2.

The winding of the search coil is 6 inch in diameter, and is of 34swg enamelled wire. The coil former can be made from three pieces of 1/8 inch thick hardboard, or similar insulated material. Cut one disc 6 inch in diameter, and two others about 6½ inch in diameter. Smooth the edges with glasspaper, and

cement them securely together, with the small disc between the larger discs. This provides a former with a channel to take the winding. The adhesive should be allowed to dry with the discs under pressure, so that there is no gap into which turns can fall. The former should also be varnished, and allowed to dry, before winding.

The wire is passed up through a small hole, this being point 1. Wind on fifteen turns. Make a loop and pass this up through a hole, for point 2. Wind a further twenty-five turns in the same direction, finishing off at point 3. These leads are scraped and soldered to a small tag strip. This is an anchor point for insulated leads which will run up the handle of the locator (Fig. 32).

Cut a block of wood, and drill it to take the bottom of a broomstick. Cement the block to the prepared coil.

Components are assembled on a tag strip as in Fig. 32. Mount this strip in a plastic lunch box or similar container, which will also take a small 9v battery, and the switch S1. A test for oscillation can be made by placing a meter in one battery lead, and shorting 1 and 3 of the coil. Meter current should change slightly. If not, look for some mistake in wiring.

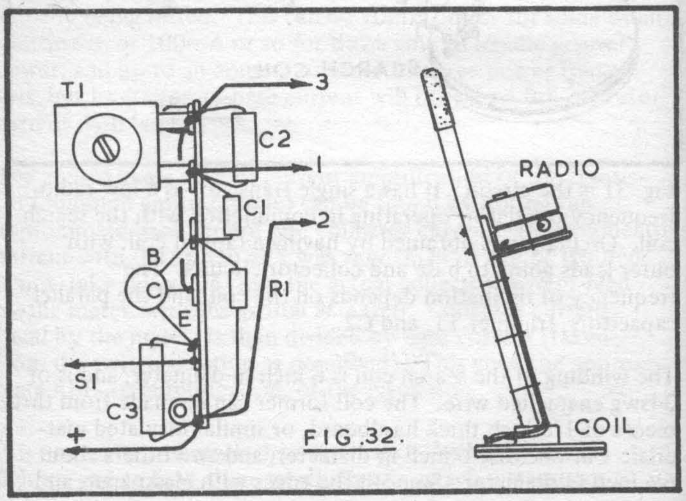

FIG.32.

A shelf is bolted to the handle, and the detector rests on this, with connections passing down to the search coil. A few loops of tape or string will secure these to the handle.

Operation in Great Britain can be on frequencies under 100kHz. Initially tune the radio receiver to 200kHz on long wave, and place it near the coil. Adjust T1 until a strong whistle is heard in the receiver speaker. This is the 2nd harmonic of the oscillator. T1 is then screwed down very slightly from this position.

The small portable radio is placed on top of the locator box, and can be held here with elastic. It is tuned to the medium wave band. Some small adjustment of T1, and of the receiver tuning, may be required, until a suitable heterodyne is produced by a harmonic or multiple of the oscillator. It will then be found that the pitch of the note changes when the search coil approaches a metal object.

It should generally be found that sufficient coupling is obtained with the receiver when they are situated as in Fig. 32. If not, run a few inches of insulated wire from the transistor collector tag up through the box and place this under the receiver.

For best results, adjustments are directed towards obtaining a good change in pitch, when the search coil approaches metal.

For those interested in details of other detectors, reference can be made to "How to Build Your Own Metal and Treasure Locators" (No. 32, Babani Press).

SECTION 3

RADIO AND AUDIO FREQUENCY

MW Booster

This unit can be employed in conjunction with any ordinary portable having a ferrite rod aerial, and will provide a considerable increase in range and sensitivity to weak signals. It has two functions – to allow the use of an external aerial and earth, and to amplify and couple signals to the portable receiver.

Tuning coverage is for medium waves. That is, about 200–550 metres, or 1500–550kHz. Many overseas and other stations can be picked up on this band, and distant reception can be particularly common on the higher frequencies, during the hours of darkness.

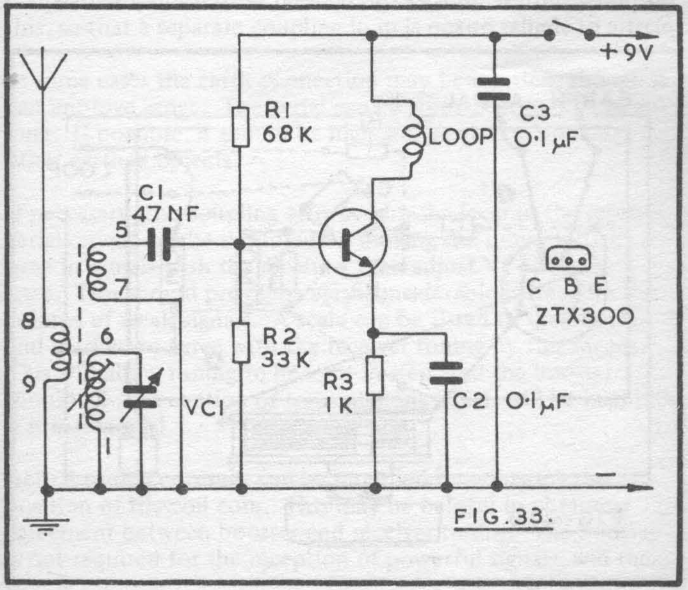
+9V
R1
68K
LOOP
C3
O·1 μF
C1
47 NF
5
7
8
6
9
1
VC1
R2
33K
R3
1K
C2
O·1 μF
C B E
ZTX300
–

FIG.33.

Fig. 33 is the circuit, and it uses a cored tuning coil with aerial coupling winding 8–9, tuned winding 6–1, and base coupling coil 5–7. These numbers are for the pins of the Denco (Clacton) "Blue" Range 2 component. It will be apparent that other coils, or surplus or home-wound inductors, can be substituted.

For the coil mentioned, VC1 is 315pF. However, this value is not too critical, provided the capacitor fitted allows tuning the wanted frequencies. A solid dielectric capacitor can be used, though the air spaced type is likely to be a little more efficient.

Coupling to the base is by C1, and resistors R1 and R2 set the base bias conditions, with R3 for emitter bias. These values will suit numerous RF transistors.

Fig. 33B shows assembly of the components on an insulated board. The coil is mounted by its threaded bush. When the board has been wired, it can be mounted by means of bolts with extra nuts. The case is preferably metal, for screening purposes, but the booster can be used satisfactorily with a plastic or similar box.

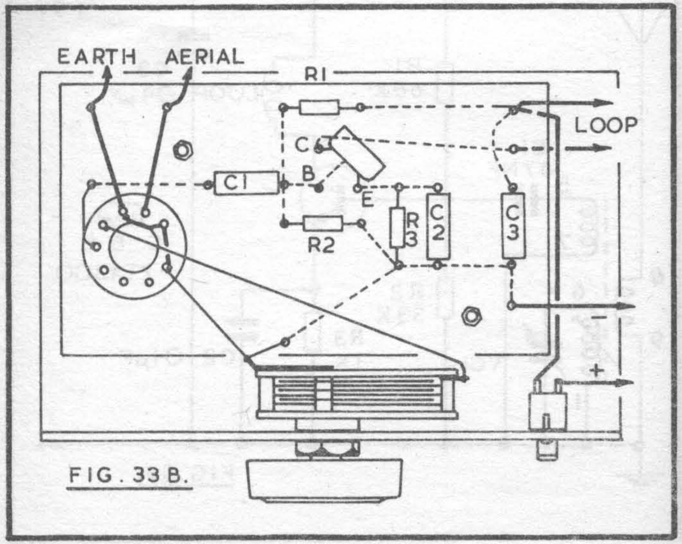

FIG. 33 B.

Fit a small slide switch to the box front, and aerial and earth sockets to the back, near the coil. A 9v battery is accommodated in the case.

Coupling to the transistor receiver is by means of a loop which is placed on the receiver ferrite rod, or near the rod. The loop can be about twenty turns of thin insulated wire, about ½ inch in diameter, and bound with cotton to keep the turns together The ends can be left long enough to reach the two connecting points in Fig. 33B.

If the case is metal, the booster and receiver can be near each other, so the loop leads will only need to be a few inches long. But when an insulated case is used, coupling from the receiver ferrite aerial or other circuits, to the booster coil, may cause instability. This will become apparent as whistles accompanying most signals. If this arises, the booster and receiver should be moved some inches apart. The aerial lead must not run near the loop connections.

Some receivers have an internal winding which is intended to couple an external aerial, and it may be possible to connect to this, so that a separate coupling loop is not required.

In some cases the earth connection may be omitted, though it can improve range. The aerial can be some 20ft to 40ft or so long. If possible, it should be high and clear of walls and other earthed objects.

If necessary, the coupling afforded by the loop to the receiver aerial circuit can be modified by moving the loop. Initially, tune in signals with the receiver, then adjust VC1 for resonance. This should provide a very considerable increase in the volume of weak signals. A scale can be fitted to the booster, and marked to agree with the receiver tuning by this means. This will allow tuning to be kept in step, and the booster should allow reception of transmissions which will be inaudible if it is removed.

Actual tuning coverage can be modified by adjusting the position of the coil core. This may be helpful in obtaining agreement between booster and receiver tuning. The booster is not required for the reception of powerful signals, and the

receiver automatic gain control circuit will tend to mask any increase in signal strength obtained with such transmissions.

Midget MW Diode Radio/Tuner

A small, compact tuner can be of utility in allowing radio reception through an amplifier, and a diode receiver can give excellent quality of reception. Such a unit can also be employed alone, with headphones, for personal listening.

Fig. 34 is a receiver/tuner which can be built in a very small plastic or other insulated box. VC1 is a compression type trimmer, fitted with a small knob for adjustment by hand. To do this, it is necessary to remove the short screw (generally 6ba) and replace this with a longer one, which projects to take the knob. The latter can be a terminal head, or the cap from an empty tube of adhesive, secured with lock nuts.

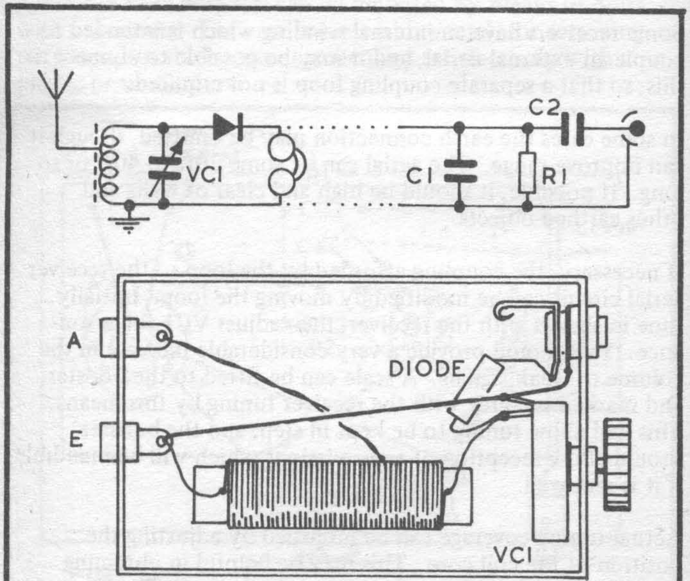

FIG.34.

For best efficiency and small size, the tuning coil is wound on a piece of ferrite rod. It consists of seventy turns of 24swg or similar enamelled wire, on a rod about 1½ inch long and 3/8 inch in diameter. Actual details will not be critical. A piece can be obtained from a longer rod by cutting a notch round the rod with a file and snapping at this point with the fingers. A few touches of adhesive will hold the ends of the winding secure. The loop for the aerial connection is about one-third the total number of turns from the diode end of the coil. (Moving this tapping towards the earthed end of the coil will increase selectivity or sharpness of tuning, but reduce volume.)

Aerial and earth sockets, and a 3.5mm jack outlet, complete the receiver. Headphones can be plugged directly in here.

The aerial and earth can be provided as explained for the diode radio in Fig. 7.

To use the radio as a tuner, C1, R1 and C2 may be added. Suitable values are C1 470pF, R1 220k, and C2 0.25μF. A screened lead with jack plug should be made up to connect the tuner to the amplifier input socket.

One Transistor VHF Receiver/Converter

The super-regenerative receiver is well known for its great sensitivity, which equals that of a multi-stage superhet. It has no alignment difficulties, and few components, so is an interesting project. Its main disadvantages are its background hiss, but this ceases when a signal is tuned in, and its lack of selectivity. The latter is however relatively unimportant on the VHF bands. The circuit here is particularly intended for 2m reception, but is easily modified. It is not intended for FM signals.

In Fig. 35 the detector is a VHF FET, and the field effect transistor will be found very satisfactory in circuits of this kind. T1 is the pre-set tuning capacitor, which allows adjustment over a fairly wide band of frequencies. Tuning is by VC1, which does not require a reduction drive.

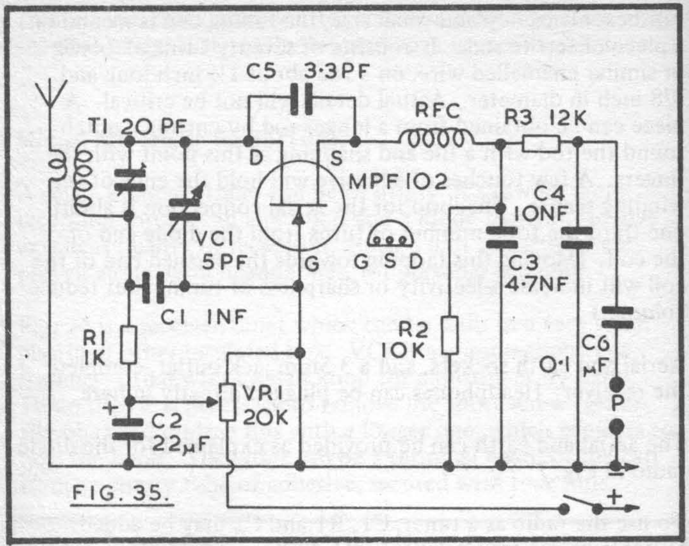

FIG. 35.

Control of super-regeneration is by means of the potentiometer VR1, which adjusts the FET drain potential. The value of VR1 is not very important, but a linear potentiometer is necessary.

C5 provides feedback from source to drain, for regeneration. R3 and the associated components form the quench network, with C6 feeding headphones, or an audio amplifier.

For best results, VR1 should enable the transistor to go smoothly into super-regeneration, which is shown by the hiss mentioned. Achieving this depends somewhat on the transistor, value of C5, adjustment of T1, and aerial loading. However, matters are not too critical.

Transistors unsuitable for VHF cannot be used. Individual FETs of the same type will also differ somewhat, so that in some cases it might be necessary to omit C5, or change the value. Alternatively, a small pre-set (not over 10pF) may be fitted. It will be found that a very small capacitance is required.

Another cause of failing to obtain super-regeneration is too tight aerial coupling, and this is easily corrected by moving the aerial coupling winding a little further from the larger winding.

The RF choke can be a VHF type, or can consist of seventy five turns of 36swg enamelled wire, side by side on a ¼ inch diameter insulated rod. Touches of adhesive may hold the end turns only.

Short leads are essential, and assembly on a tag board is shown in Fig. 35B. The variable capacitor is fitted first, and it can later attach the finished board to a panel, by means of an extra nut.

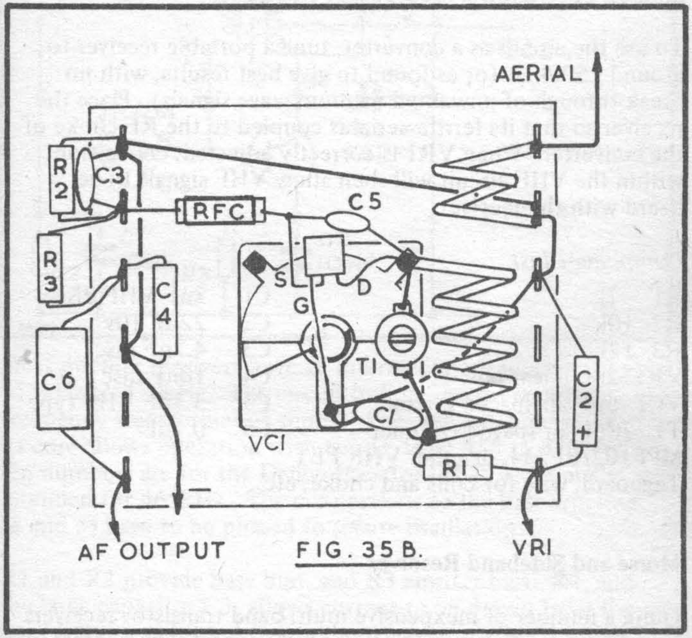

FIG. 35 B.

The larger coil has five turns of 18swg wire, and is ½ inch in diameter and ¾ inch long. The smaller coil has two turns. Wiring to VC1, T1, C5, the coil and C1 must be as short as can be arranged. Connections elsewhere are reasonably direct. A ceramic disc capacitor may be used at C1. T1 should be a low loss trimmer.

The board can be fitted in a case which will carry VR1, and large enough for the battery and on-off switch. The case can be used to support a telescopic aerial, some 30 inch or so when extended.

Frequencies tuned by VC1 can be raised by unscrewing T1, or by stretching the larger coil to separate the turns more, or by using a coil with fewer turns, or of smaller diameter. The receiver is most suitable for use over the 70MHz to 200MHz range.

For the single transistor alone, a sensitive pair of medium or high resistance phones will be needed. Audio ouptut from C6 can be taken to an amplifier, for more powerful reception with phones, or to allow a loudspeaker to be used.

To use the circuit as a converter, tune a portable receiver to around 1500kHz (or as found to give best results, with no break through of unwanted medium wave signals). Place the receiver so that its ferrite aerial is coupled to the RF choke of the converter. When VR1 is correctly adjusted, conversion within the VHF circuit will then allow VHF signals to be heard with the receiver.

Component List

R1	1k	C1	1nF VHF disc
R2	10k	C2	22μF 10v
R3	12k	C3	4.7nF disc
VR1	20k linear pot.	C4	10nF disc
VC1	5pF air-spaced variable	C5	3.3pF VHF type
T1	20pF air spaced trimmer	C5	0.1μF

MPF102/BF244, or other VHF FET
Tagboard, wire for coils and choke, etc.

Morse and Sideband Resolver

Quite a number of inexpensive multi-band transistor receivers have frequency coverage which includes one or more Amateur bands. That is, 1.8, 3.5, 7, 14, 21, or 28MHz. On these bands transmissions are often in Morse code or CW, or single sideband. CW or SSB can be resolved by the specialised Amateur receivers and transceivers, but not with the general purpose home receiver. To receive CW or SSB with the latter type of radio, it is necessary to add an oscillator. This either generates a heterodyne or beat note for code reception, or replaces the carrier which was eliminated from the single sideband transmission.

The circuit in Fig. 36 may be used for these purposes. It generates a carrier of about the same frequency as that used in the intermediate frequency stages of the receiver, exact adjustment of frequency being by VC1,

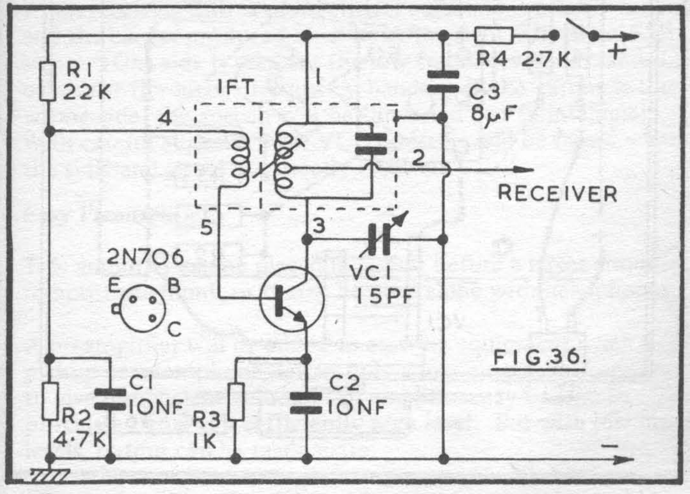

FIG.36.

Most popular receivers have an intermediate frequency of about 455kHz to 470kHz. The circuit in Fig. 36 uses an intermediate frequency transformer of similar type, and adjustment of its core allows operation with any IF from 455 to 470kHz. Pin numbers are for the Denco (Clacton) IFT13, normally intended for 465kHz. The connections to the base winding (4 and 5) have to be phased to secure oscillation.

R1 and R2 provide base bias, and R3 emitter bias. R4, and the large capacitor C3, allows current to be drawn from the same battery as is used to run the receiver, if wished. Alternatively, a separate 9v battery may be employed here.

Fig. 36B shows the practical layout of components, and wiring to them, using a perforated board. Connections and joints are kept close against the underside of the board. When wiring is finished, the board is mounted in a metal box, using ½ inch bolts with extra nuts for clearance.

VC1 can best have a dial with a central zero mark, this showing when the plates are half meshed.

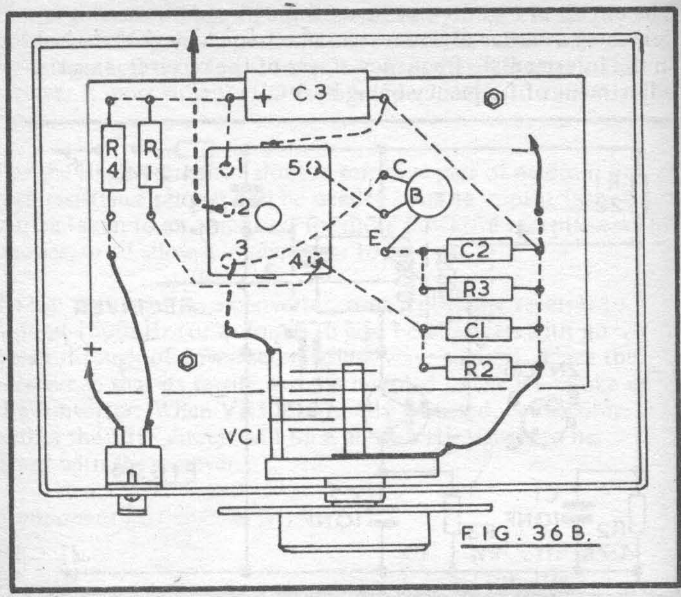

FIG. 36B.

An insulated lead is run from 2 on the IFT, and is placed near the receiver. This lead may be so positioned as to give best coupling into the IF stages of the receiver. If coupling is too weak, the signal from the CW/SSB Resolver will be inadequate. On the other hand, tight coupling will allow the carrier to swamp weak signals tuned in on the receiver. Fortunately the degree of coupling is not at all critical.

To set up the oscillator, first of all tune in any normal broadcast with the receiver. Set VC1 half closed and switch on. The core of the IFT should now be slowly adjusted until a loud heterodyne or beat note is heard accompanying the signal received. Set the core so that this note falls to about zero, or to a low pitch. It will then be possible to rotate VC1 either way from this central zero beat position, in each case producing a note which rises in frequency as VC1 is rotated.

To receive CW or Morse code, it is only necessary to tune in the code signal with the receiver, and then to adjust VC1 to obtain a note of the most suitable pitch. VC1 can be rotated either way from its central or zero beat position, to obtain best freedom from interference.

When receiving SSB, a more critical adjustment of VC1 is needed and the carrier produced must be at the right side of the SSB signal. (One side is used for the low frequency bands, and the other for the higher frequency bands.) If the carrier is the wrong side, the speech will be "inverted" or unintelligible. With careful adjustment of VC1, a setting will be found where the sideband signal is correctly resolved.

Easy Preamplifier

This amplifier can be placed in circuit before a larger amplifier, to boost the input, or it may be used alone with headphones.

A preamplifier will be added to existing equipment when a pickup or microphone having only a low output has been found to give insufficient volume. An amplifier may be able to provide a signal at a sufficiently high level. But with low input levels, output can be inadequate.

The stage in Fig. 37 can provide considerable gain. Audio input is to C1, and R1 provides base current for the transistor. Amplified signals are obtained across R2, and pass to the main amplifier (or phones) via the isolating capacitor C2.

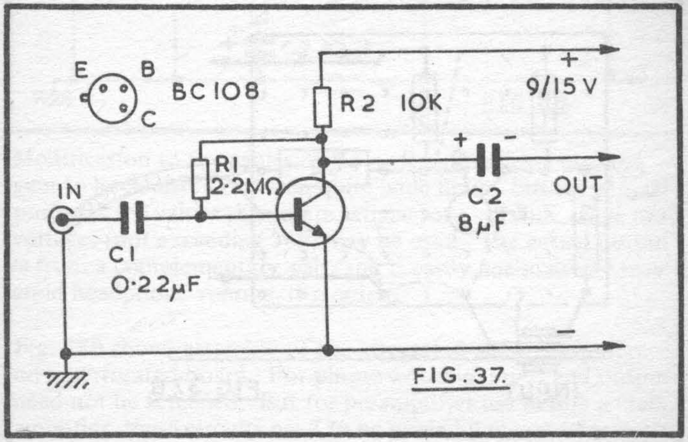

FIG.37.

With a BC108 or similar transistor, the supply can be about 9v to 15v. Numerous NPN transistors, intended for preamplifier use and of low noise type, will perform well. Current may be from a separate battery. Or it may in many cases be drawn from the main amplifier. To avoid instability or similar troubles, the positive supply should be from a well decoupled and smoothed point in the main amplifier. If this is not available, connect a 220μF capacitor from positive to negative in Fig. 37, and draw the positive supply through a 1k resistor. Negative is returned to the negative line of the main amplifier.

Should the main amplifier have a positive earth line, then the transistor can be a PNP type, R1 and R2 being changed to suit.

Fig. 37B shows assembly on a perforated board. When the circuit provides input to a main amplifier, it should generally be in a metal case. Input is by a screened lead, and a screened lead is also fitted for the audio circuit to the main amplifier. The case should be earthed to the negative line. This can be by means of one of the bolts securing the board.

Screening to avoid hum or instability will generally prove to be unnecessary when using the stage to operate headphones. So two sockets, or a jack socket, will be more convenient for the phone circuit.

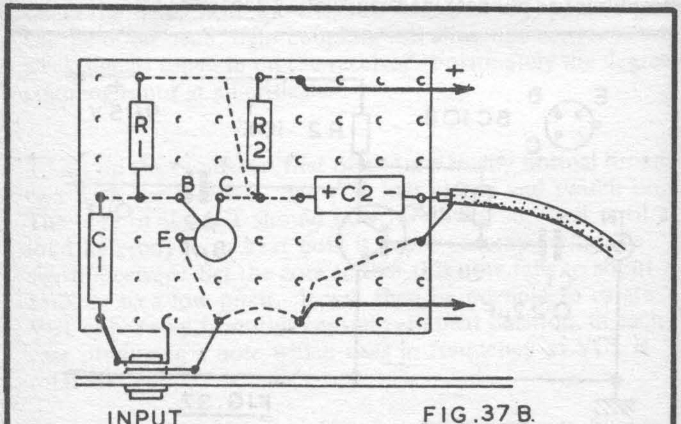

FIG.37B.

741 Preamplifier

Integrated circuits have several stages of amplification, formed as a single chip and enclosed in a small case. The 741 is an easily obtained audio IC which can be used in many circuits. The uA741CE or metal can 741 is convenient, as its leads may be spaced to suit any ready perforated board. The dual in line 741 is electrically similar.

Fig. 38 shows a preamplifier using the metal can 741. Input is via C1 to 3. As shown, 7 and 4 are positive and negative, while output is from 6. Negative feedback is applied to 2, via R4. C3 is an isolating capacitor.

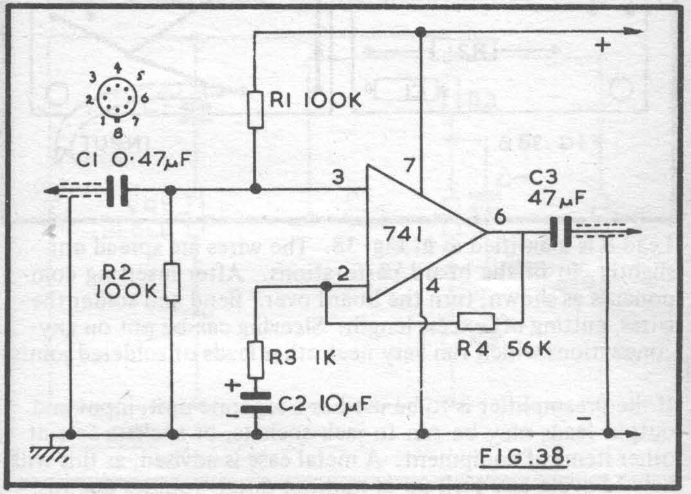

FIG.38.

Modification to the values of R3 or R4 allow feedback and gain to be adjusted between quite wide limits, but for general purposes the values shown are satisfactory. A wide range of voltages (not exceeding 36v) may be used. The actual output is from a complementary pair, and is easily adequate for very good headphone volume, if required.

Fig. 38B shows assembly of the integrated circuit amplifier on a perforated board. For phone working, input and output need not be screened. But for preamplifier use before a main amplifier, these circuits need to be made by means of screened leads.

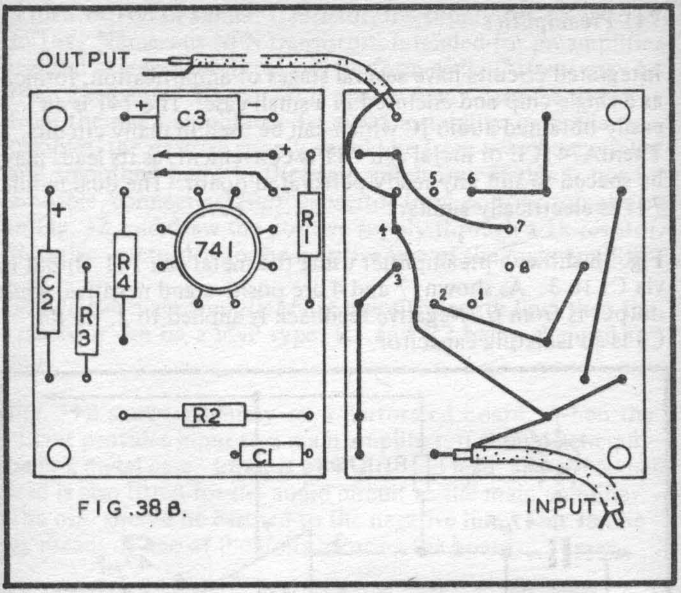

FIG.38B.

Lead 8 is identified as in Fig. 38. The wires are spread out slightly, to fit the board perforations. After inserting components as shown, turn the board over. Bend and solder the wires, cutting off excess length. Sleeving can be put on any connections which run very near other leads or soldered joints.

If the preamplifier is to be used as a separate unit, input and output leads may be run to jack sockets, or sockets to suit other items of equipment. A metal case is advised, as this will help to avoid any pick up of hum, or other troubles due to coupling to the board wiring. The board can be secured to the metal box or case with ½ inch 6ba bolts, and extra nuts to give a little clearance.

The amplifier may be run from its own battery; or from a supply used with a main amplifier. For the latter, it is wise to provide a decoupled point, in a similar manner to that described for the transistor preamplifier. This consists of a quite large capacitor (say 220μF to 470μF) from positive to negative across the preamplifier supply circuit, and a resistor of 1k to 2.2k or so from preamplifier positive to positive on the main amplifier.

Tone Control for Your Discs Etc.

The preamplifiers shown have a substantially flat response. That is, they do not purposely emphasize bass or treble, or have similar effects in modifying the audio frequencies present. In some circumstances it is necessary to boost treble or bass, or reduce the emphasis of these frequencies, in order to obtain the most satisfactory reproduction.

This may be achieved by making a fully adjustable tone control, and Fig. 39 is a suitable circuit. A volume control is also incorporated, as this has not been present in the preamplifiers.

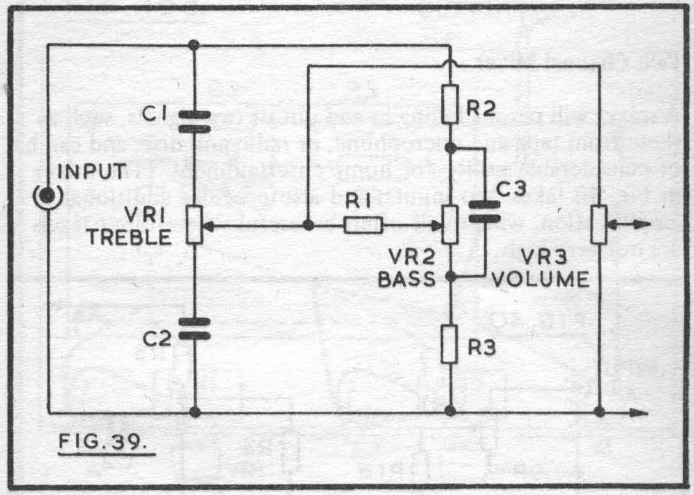

FIG.39.

The tone control is a passive circuit, needing no battery, and it can be placed between the source of audio signals (generally a microphone, pick-up, or radio tuner) and the amplifier or preamplifier. It may also be located between preamplifier and main amplifier, if wished.

Results are substantially flat with the treble control VR1 and base control VR2 in their central positions. Adjustment of VR1 and VR2 will provide separate boost or cut of treble, or boost or cut of bass frequencies. VR3 is the volume control.

The degree of control afforded by circuits of this type may be adjusted by alterations to the component values. In practice, it will be found that there is little point in having extreme levels of adjustment, which will never be required. The following values may be used: C1 and C2, 270pF; R1, 390k; R2 and R3, 100k; C3 680pF; VR1 and VR2, 2 megohm linear potentiometers; VR3, 2 megohm log potentiometer.

Components should be assembled in a metal case, for screening. The case is grounded to the earth line. Input and output sockets can be near the ends of the case, with VR1, VR2 and VR3 between. Components can be soldered directly to the potentiometer tags.

Two Channel Mixer

A mixer will permit fading in and out of two signals, such as these from tape and microphone, or radio and disc, and can be of considerable utility for home entertainment. The mixer in Fig. 40 takes two inputs, and also provides additional amplification, which will often be useful where input levels are not very high.

FIG.40.

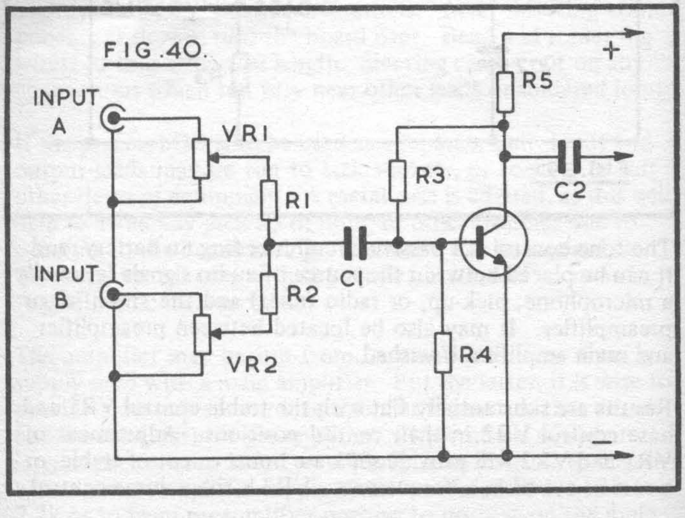

Input A has its individual volume control VR1, and Input B is similarly provided with VR2. Resistors R1 and R2 prevent short circuiting one input when the other potentiometer is set at zero. Coupling to the transistor base is by C1. R3 and R4 set base bias conditions, with R5 as the collector load and audio output from C2.

Component values are not very critical, though R3, R4 and R5 need to be selected to suit the transistor fitted. The mixer is connected and used in a similar way to that explained for the preamplifiers.

FIG.40B.

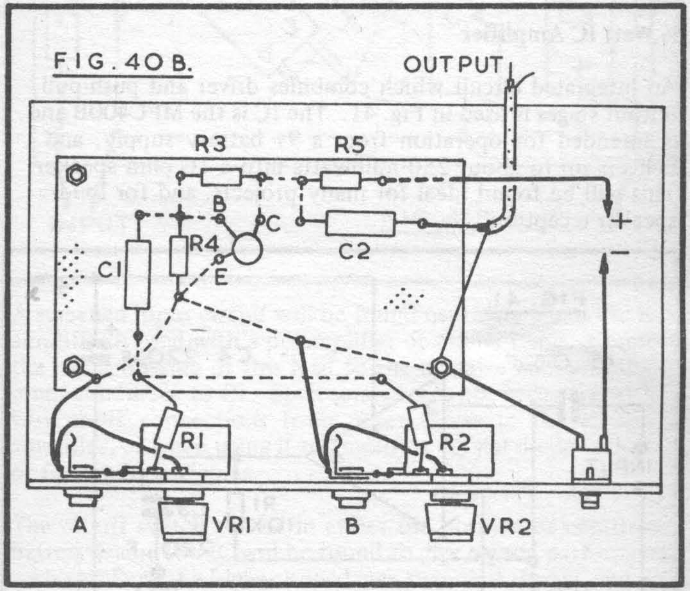

Fig. 40B is the layout, using a metal case or box. A screened output lead is provided. Input sockets A and B may be chosen to suit the type of plugs on existing equipment. The on-off switch is separate from the potentiometers, so that these can be left at pre-set levels if wished.

Component values may be chosen to suit low or high impedance inputs, if wished. Or the following may be used for general purposes:

Components List:

VR1, VR2, 1 megohm log potentiometers.
R1, R2, 200k
C1 0.5μF
R3 330k
R4 22k
R5 10k
C2 0.5μF
BC109

This is when using a 9v supply.

¼ Watt IC Amplifier

An integrated circuit which combines driver and push-pull output stages is used in Fig. 41. The IC is the MFC400B and is intended for operation from a 9v battery supply, and delivers up to about 250 milliwatts into a 16 ohm speaker. This will be found ideal for many projects, and for loudspeaker reception.

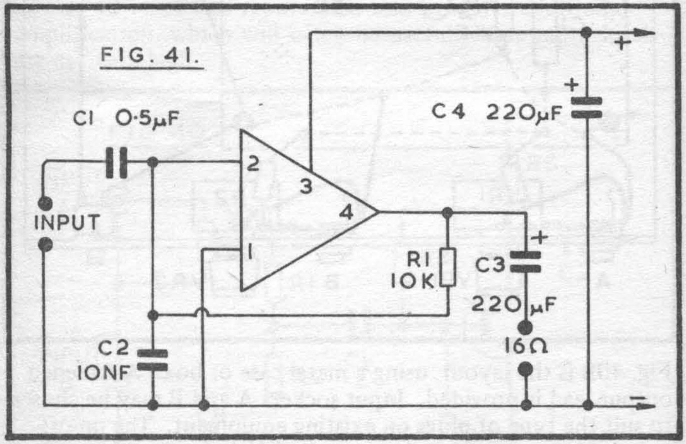
FIG. 41.

Audio input is to 2 of the IC, from the isolating capacitor C1. C2 is a by-pass capacitor to maintain stability. The IC has only four connections, 1 going to the negative line, and 3 to the positive supply. Audio output is from 4, and to the loudspeaker via C3. R1 provides negative feedback and bias to point 2.

The assembly of this amplifier on 0.15 inch matrix board is shown in Fig. 41B. Pin numbering is when looking at the top of the IC and the component side of the board.

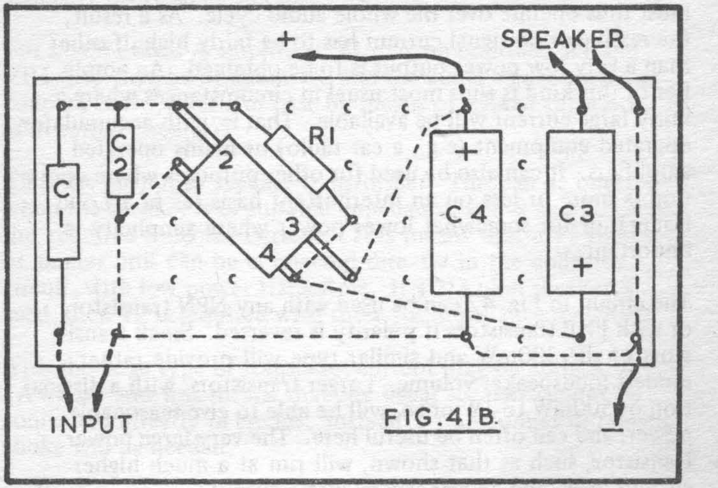

FIG. 41B.

A screened input circuit will be found necessary when the IC amplifier is used with a preamplifier or similar items. Connect the outer brading of this lead to the negative line, and the inner conductor to C1. Such screening is not necessary with very short connections from other stages to the IC amplifier, or when using it to amplify a crystal diode radio, or for similar purposes.

The on-off switch can be in either the positive or negative battery lead. The IC will be found to give a good performance with speakers of a higher impedance than that shown, and up to 75 ohm may be used (with some small loss in output). Units of lower impedance than 16 ohms should not be fitted.

The small size of this IC and the few extra components needed make is very suitable for miniature apparatus, probably with a 2½ inch speaker. It will also operate a large speaker well (see "Boost Speaker") and at a volume level adequate for many purposes.

Class A Amplifier

A Class A amplifier uses a single transistor, Fig. 42, which must thus operate over the whole audio cycle. As a result, the resting or no-signal current has to be fairly high, if other than a very low power output is to be obtained. An amplifier of this kind is thus most usual in circumstances where a fairly large current will be available. That is, with accumulator operated equipment (e.g., a car radio) or mains operated amplifiers. It can also be used for other purposes where operation is more or less on an intermittent basis (as in a Loud Hailer), or for somewhat lower power where simplicity is important.

The circuit in Fig. 42 can be used with any NPN transistor, or with PNP transistors if polarity is reversed. Small transistors of the 300mW and similar type will provide rather modest loudspeaker volume. Larger transistors, with a dissipation of 500mW to 1W or so, will be able to give reasonable power, and can often be useful here. The very large power transistor, such as that shown, will run at a much higher current level, and deliver several watts output.

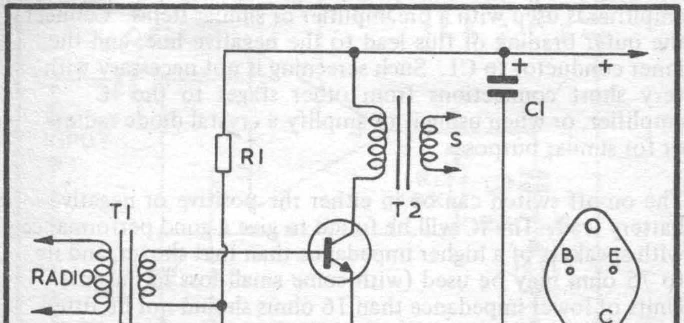

FIG. 42.

T1 is the coupling or driver transformer, which may receive audio signals from a radio receiver, smaller amplifier, or other equipment. A step down ratio of about 5:1 to 10:1 will generally prove to be suitable.

For smaller Class A amplifiers, T1 may be omitted. R1 and VR1 may then be of about 1k, and the audio signal can be coupled to the transistor base by means of a fixed capacitor, as in earlier circuits.

T2 is the output or speaker transformer. This can be omitted when the speaker is of suitable impedance, and is able to carry the required collector current. This means that a 75 ohm or similar unit can be connected directly in the collector circuit, with low power transistors. If a 2/3 ohm speaker is used, then T2 is required with these for impedance matching.

With the large type of transistor, current may be 500mA to 1A or so, and this means that the usual speaker cannot be connected directly in circuit. Instead, a transformer or tapped choke will be needed.

The circuit is often operated from 12v (a vehicle supply) but can be used with 9v, or more than 12v.

R1 and VR1 set bias conditions, and R2 limits peak collector current. For the 2N3055, R2 may be 1 ohm. R1 can be 100 ohm, and VR1 25 ohm. T2 can be 2:1 for a 3.5 ohm speaker. C1 is 3200μF. Begin with VR1 at minimum value, slowly increasing this until reproduction is satisfactory. (This will set collector current at about ½ ampere.)

With other transistors, adjustment can be made in the same way, by slowly increasing the base bias until best results are secured. Raising the value of VR1 in Fig. 42, or reducing the value of R1, will increase collector current, and thus raise the power which can be handled. However, current is limited by the power handling capacity of the transistor, so a meter should be put in one battery lead, and a check made that the power dissipation of the transistor is not exceeded. With small transistors, current can be limited to 50mA or so.

Fig. 42B shows assembly of the circuit in Fig. 42. The power

transistor is mounted on an aluminium plate, which acts as a heat sink to carry away heat from the transistor itself. Note that this plate is common to the collector. Base and Emitter pins pass through clearance holes. Smooth off any burr round the fixing and clearance holes, so that the transistor can rest flat on the metal. Where it is awkward to have the plate or case common to the collector, the usual insulation set can be fitted. This consists of bushes for the fixing bolts, and a thin mica washer to go under the transistor, so that electrical isolation is obtained.

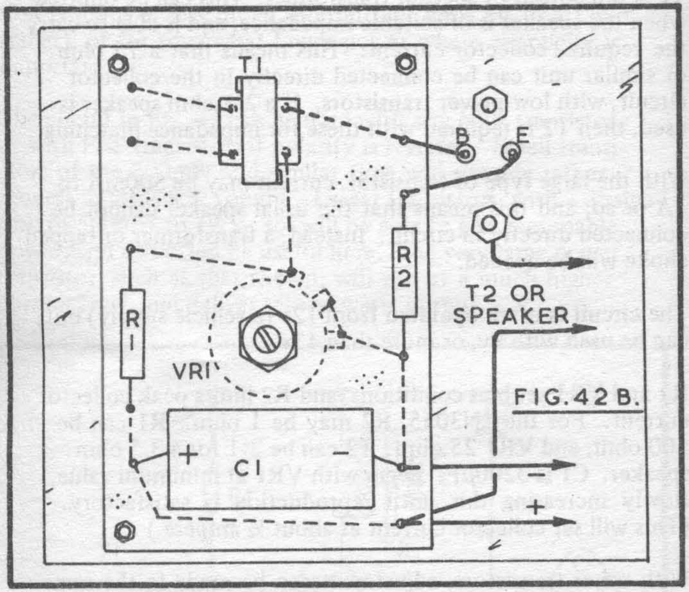

FIG.42B.

The insulated board carrying the other components is mounted by long bolts with extra nuts. A large Class A stage transistor of this type will grow very warm to the touch, if run at anything like maximum power rating. This does not indicate any defect.

Two Stage Amplifier

The circuit of a 2-stage amplifier is shown in Fig. 43, and this can be used for loudspeaker reproduction. Input to the driver stage is by the isolating capacitor C1. Operating conditions in this stage depend largely on the values of R2, R3, and the emitter bias resistor R5.

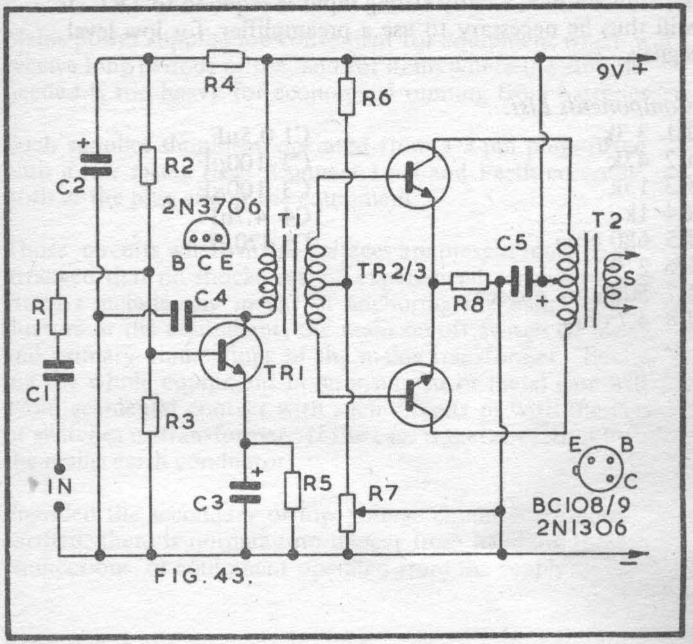

FIG. 43.

T1 is the driver transformer, with centre tapped secondary, so that TR2/3 may be operated in push-pull (see Fig. 12). T2 couples output to a 2/3 ohm speaker.

With the values shown, a BC109 will give very good results in the TR1 position, with 2 x BC108s for the output stage. R7 is set so that battery current is around 10mA, with no signal present. Current rises from 20mA to 60mA or so, according to volume, when the amplifier is driven.

A pair of 2N1306 transistors may be used for the output stage. Other transistors are also possible here, as operating conditions can be adjusted by R7.

The 2N3706 will operate in the TR1 position, though with

lower gain than the BC109. Numerous other small audio transistors can also be used here.

For full volume, a fairly strong input is required to TR1. It will thus be necessary to use a preamplifier, for low level signals.

Components List:

R1 3.3k
R2 47k
R3 15k
R4 1k
R5 680 ohm
R6 2.7k
R7 500 ohm pre-set
R8 4.7 ohm

C1 0.5μF
C2 100μF
C3 100μF
C4 4.7nF
C5 470μF

SECTION 4
POWER SUPPLIES

Mains power supplies are convenient for equipment which will receive long periods of use, and for items where the current needed is too heavy for economical running from batteries.

Such supplies should be operated from a 3-pin plug, fitted with a low rating fuse. Connect L, N and Earth correctly, both at the plug and at the equipment.

Those circuits where mains voltages are present must be so arranged that no shocks can be experienced. Mains voltage circuits include any means of anchoring the mains conductors in the equipment, the main on-off switch (if used), and primary connections of the mains transformer. Enclosing the whole equipment in an insulated or metal case will avoid accidental contact with such circuits or with the tags of switches or transformer. If the case is metal, earth it by the mains earth conductor.

Provided the secondary or low voltage circuit is correctly earthed, there is normally no danger from handling these connections, or equipment operated from the supply.

Easy PSU

A useful power supply suitable for many purposes can use the circuit in Fig. 44. Mains L (live) and N (neutral) conductors go to the primary P of the transforer. A switch is placed in the L circuit.

The secondary S can delivery any wanted low voltage, and also give isolation from the mains.

Almost all equipment will need direct current. The four separate rectifiers, or an equivalent single device, produce pulsating DC. The reservoir capacitor C1 smoothes this. Resistor R1, with the second capacitor C2, gives additional smoothing. Output from the supply is then substantially smooth DC, as would be obtained from a battery.

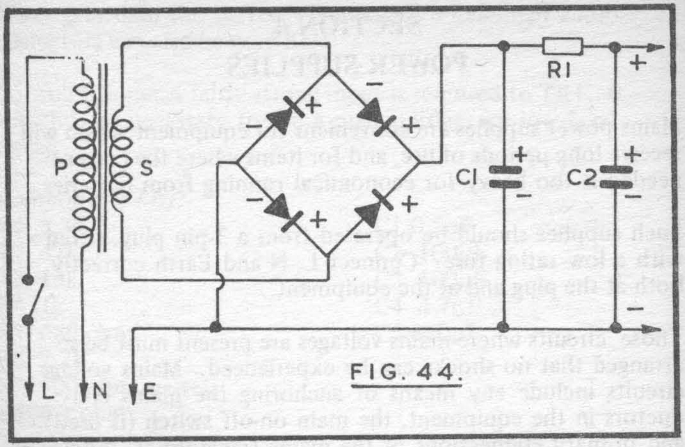

FIG. 44.

Sometimes R1 and C2 are omitted. Should hum accompany reproduction from an amplifier, this shows that smoothing is probably inadequate, and if so the cure is to add R1 and C2.

Any wanted voltage can be obtained by using a transformer with an appropriate secondary, with rectifiers and capacitors to suit.

C1 may charge up to approximately 1.4 times the RMS or usual voltage rating of the secondary S. Thus, if S were a 6.3v winding, about 8.8v would be expected across C1. But should R1 be present, a voltage drop will arise in it, the figure depending on the resistance value of R1, and current flowing.

For these reasons, the actual output may not be exactly the figure expected. This will often be unimportant, and such power supplies are largely used for all sorts of equipment.

For apparatus generally operated from a 9v or similar battery, a bell transformer with 3v, 5v and 8v output tappings is suitable. If R1 is 22 ohm, this will give around 9v at 100mA, with the 8v tap in use.

It is convenient to use 50v or 100v rectifiers, and the rating can be 1 ampere, with a 1 ampere transformer secondary. Four individual wire ended rectifiers can be mounted on a

tag strip, and connected as in Fig. 44. A combined rectifier will have AC, positive, and negative tags (see Fig. 45) and can be fitted instead.

The capacitors need to be of large value, and have a voltage rating higher than the actual voltage which will be present. The 12v capacitors could be fitted in the circuit described. C1 can be 2500μF or larger, and C2 can be of similar value.

Easy Regulated PSU

With the circuit in Fig. 44, the exact voltage will depend on the current drawn, rising when current is low. To avoid this, some means of regulation may be employed.

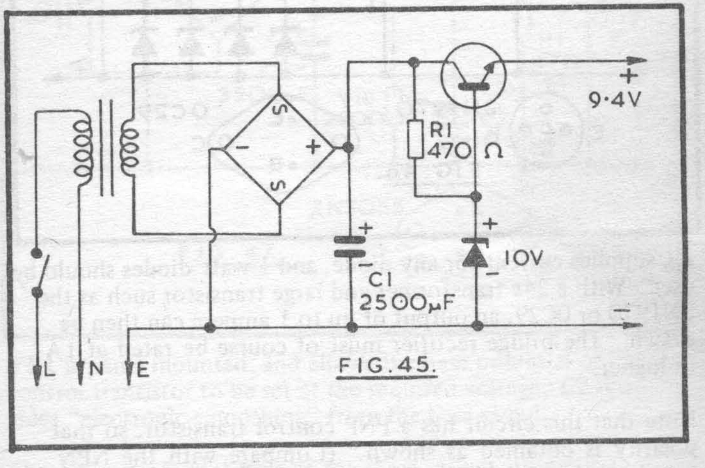

FIG.45.

Fig. 45 shows a typical regulation circuit, giving an output of approximately 9.4v. The output from the rectifier needs to be of somewhat higher voltage than required at the output, and a transformer with a 10v secondary is suitable. Current through R1 supplies the Zener diode, which holds the base of the transistor at virtually the same voltage for all current loads. Some voltage is lost in the transistor, so the output is nearly the same, for any current load. For up to about 200mA, a BFY51 or equivalent transistor is adequate, so this will do well for many purposes.

Multi-Voltage Regulated PSU

Fig. 46 is a somewhat similar circuit, but it has a switch, to select any one of the four Zener diodes. These may be chosen so that 6v, 9v, 12v and 15v will be available.

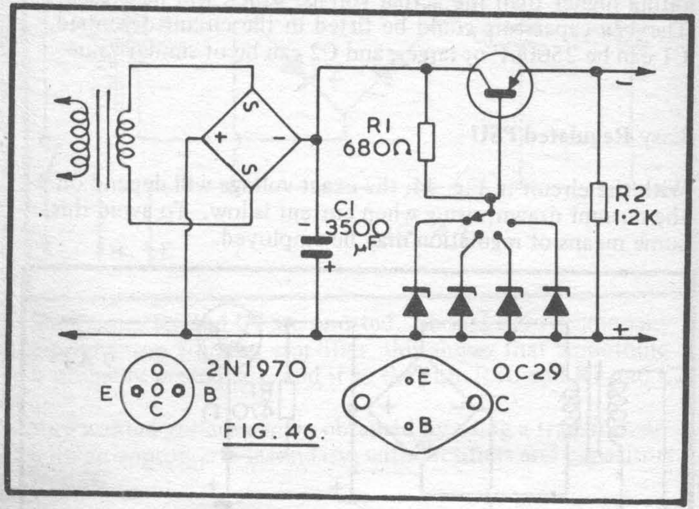

FIG. 46.

R1 supplies current for any diode, and 1 watt diodes should be used. With a 24v transformer and large transistor such as the 2N1970 or 0C29, an output of up to 1 ampere can then be drawn. The bridge rectifier must of course be rated at 1A or higher.

Note that this circuit has a PNP control transistor, so that polarity is obtained as shown. (Compare with the NPN transistor in the previous circuit.) However, either positive or negative can be grounded, and a separate earth terminal can be provided to allow this. A flying lead is placed between earth terminal and positive, or earth and negative, as required. The permanent negative earth connection shown in Fig. 45 must of course then be omitted.

More details of actual construction will be found later.

One Transistor Adjustable PSU

Fig. 47 is the circuit of a simple power supply unit which can be set to provide any output from zero to 12 volts. The transformer has a 17v secondary, rated at 1 ampere or higher. After rectification, DC is obtained from the reservoir capacitor C1.

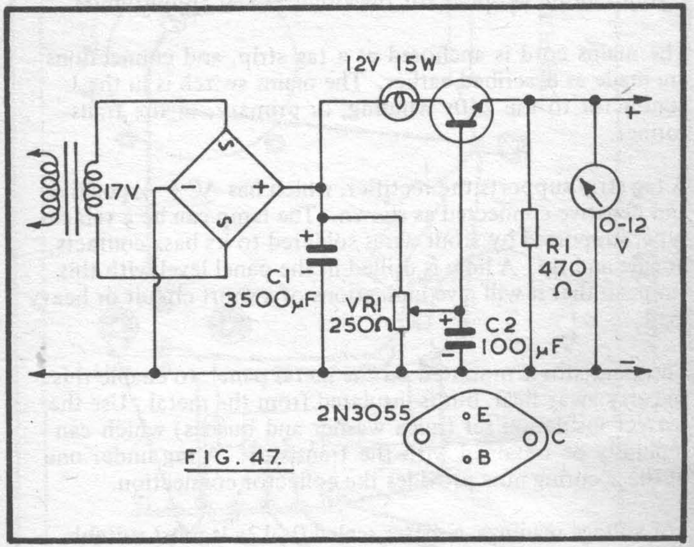

FIG. 47.

VR1 is panel mounted, and allows the base potential of the control transistor to be set at the required voltage. C2 provides "electronic smoothing" from the base circuit.

The 12 volt 15 watt bulb is for short circuit protection. With relatively low currents, little voltage is lost here. But if a short circuit arises in the output circuit, current is limited to a little over 1A by the bulb. This reduces the chances of any damage due to temporary shorts in leads or equipment.

R1 is a bleeder resistor, drawing some current even when no external load is connected. The meter is the usual high resistance type, scaled from zero to 12v. VR1 is adjusted until this instrument shows the wanted voltage.

The regulation of this supply is not so good as that of supplies using a number of transistors. But it will be found suitable for all ordinary purposes, and to supply power for circuits in this book, as only a small change in voltage arises, with changes to the load, or current drawn.

Fig. 47B shows practical construction of this PSU. A similar layout can be adopted for the other power supply units.

The mains cord is anchored at a tag strip, and connections are made as described earlier. The mains switch is in the L conductor to the 240v winding, or primary, of the transformer.

A tag strip supports the rectifier, which has AC (~), positive and negative connected as shown. The lamp can be a vehicle type, supported by stout wires soldered to its base contacts, or side and tip. A hole is drilled in the panel level with this lamp, so that it will give indications of a short circuit or heavy load.

The transistor is mounted on the metal panel, to enable this to carry away heat, but is insulated from the metal. Use the correct insulation set (mica washer and bushes) which can generally be obtained with the transistor. A tag under one of the securing nuts provides the collector connection.

For voltage readings, a meter scaled 0–12v is most suitable. Or a 1mA meter, with 12k 1% series resistor, may be used. If so, a new scale, marked 0–12, can be drawn for it.

A 3 watt linear potentiometer is used. Earth and negative are connected to the metal panel, for equipment with a negative ground line. The positive output terminal or socket is insulated from the metal.

The case can be metal, or wood. It should have a few ventilation holes. The transistor can be expected to run quite warm, at maximum current. Heating here rises as the output voltage is reduced by VR1, as in these circumstances more voltage is dropped in the transistor.

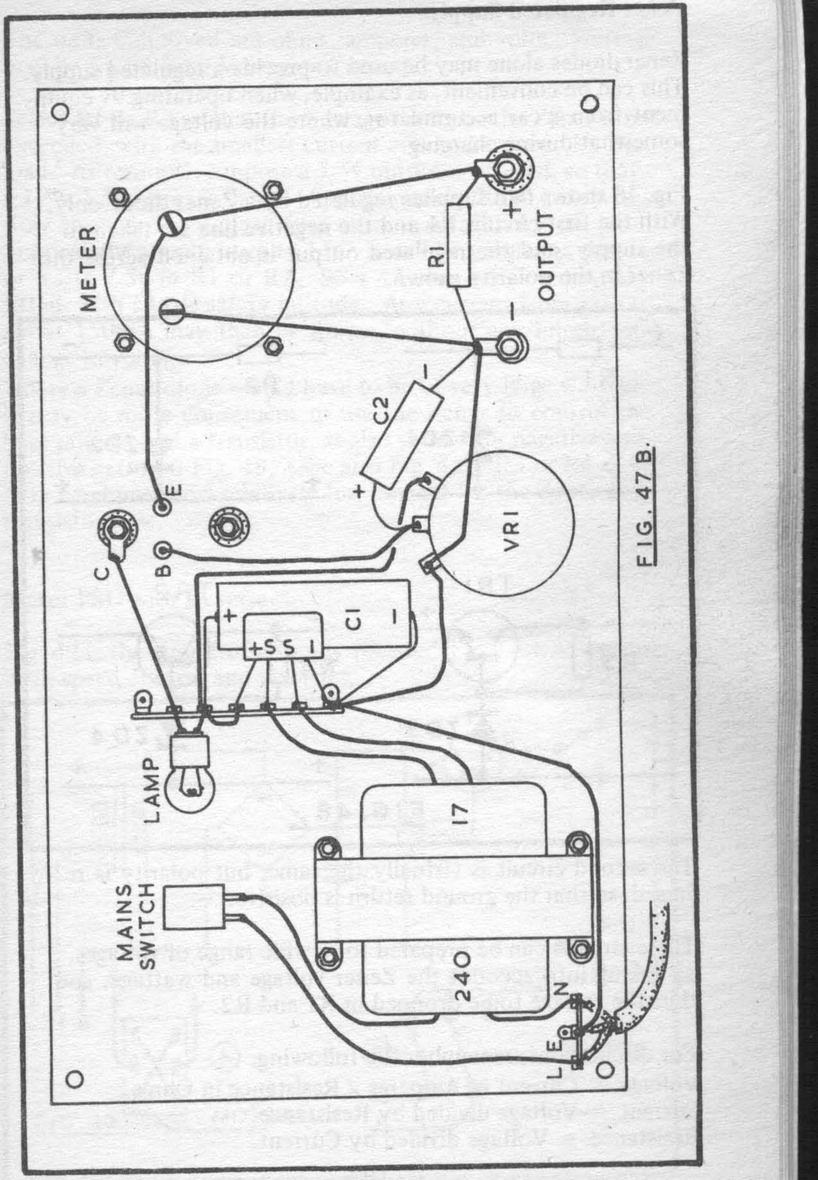

FIG. 47B.

Zener Regulated Supply

Zener diodes alone may be used to provide a regulated supply. This can be convenient, as example, when operating 9v equipment from a car accumulator, where the voltage will vary somewhat during charging.

Fig. 48 shows two supplies regulated by a Zener diode only. With the first circuit, R1 and the negative line are taken to the supply, and the regulated output is obtained across the Zener in the polarity shown.

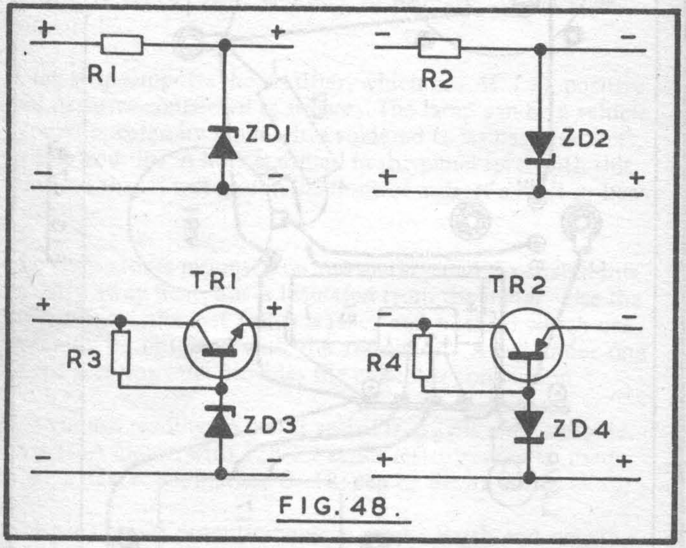

FIG. 48.

The second circuit is virtually the same, but polarity is reversed, so that the ground return is positive.

These circuits can be prepared for a wide range of voltages, by taking into account the Zener voltage and wattage, and thus the voltage to be dropped in R1 and R2.

For calculations, remember the following:
Voltage = Current in Amperes x Resistance in Ohms.
Current = Voltage divided by Resistance.
Resistance = Voltage divided by Current.

The units employed are ohms, amperes, and volts. Wattage may be found from Voltage x Current.

Select R1 or R2 so that the wattage rating of the Zener is not exceeded, with the smallest current ever to be drawn by the load. As example, suppose a 7.5v output is wanted, so that a 1½w 7.5v Zener is fitted. Approximately 200mA may thus flow through the Zener, if no current is taken by the load. Suppose the supply is up to 15v. It will be necessary to drop or waste 7.5v in R1 or R2. So a 42 ohm resistor may be fitted, with a little safety latitude. Any current from zero to about 150mA may then be drawn, without any important change in voltage.

Where a Zener diode would have to be of very large wattage, it may be more convenient to use the Zener to control the base potential of a transistor, as also shown for negative and positive earths in Fig. 48. (See also Fig. 45.) R3 or R4 then only need provide a relatively low current for the Zener and transistor base.

Motor PSU with Reverse

Fig. 49 is the circuit of a supply for model motors, and giving both speed control and reversing.

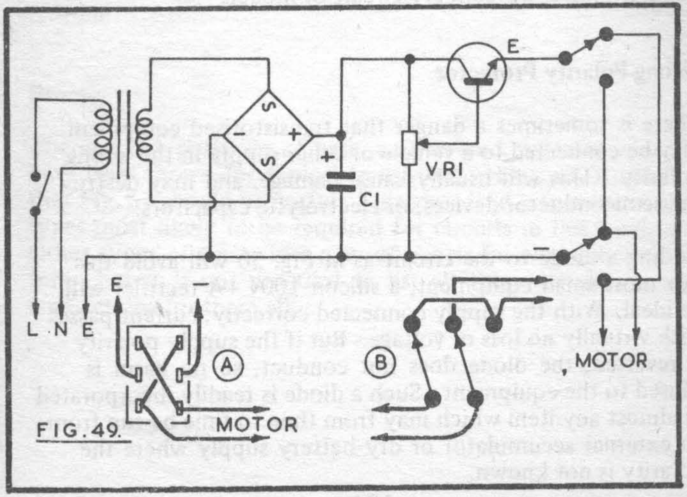

FIG. 49.

A transformer with a 12v secondary will be suitable for the usual model requiring up to 12v. The circuit is similar to that of the supply in Fig. 47. The potentiometer VR1 gives speed control. C1 can be around 2,000μF.

"A" is one type of ready-made switch, and it can be obtained with a central off position. "B" shows a home-built switch, similar to that described in Fig. 5. The switch must not short circuit the supply when changing over polarity, and one with central off is recommended.

Motors reversed in this way must be of the permanent magnet type, in which a change of polarity causes the armature to run in the opposite direction. Train and many other model motors are of this construction. Motors with a wound field will not be reversed by changing the polarity of supply.

A controller of this kind should be constructed as shown earlier, so that it is safe to use, and so that no joints in the mains circuits can be touched.

Earthing can only be omitted when a double-insulated transformer, of the type specifically made to avoid any need for earthing, has been used. The whole controller should be so constructed that, even if a fault develops, mains voltages cannot reach the secondary circuit or model.

Wrong Polarity Protector

There is sometimes a danger that transistorised equipment may be connected to a vehicle or other supply in the wrong polarity. This will usually cause damage, and may destroy the semiconductor devices, or electrolytic capacitors.

Adding a diode to the circuit as in Fig. 50 will avoid this. For most small equipment, a silicon 100v 1A rectifier will be ideal. With the supply connected correctly, current passes with virtually no loss of voltage. But if the supply polarity is reversed, the diode does not conduct, so no harm is caused to the equipment. Such a diode is readily incorporated in almost any item which may from time to time be run from an external accumulator or dry battery supply where the polarity is not known.

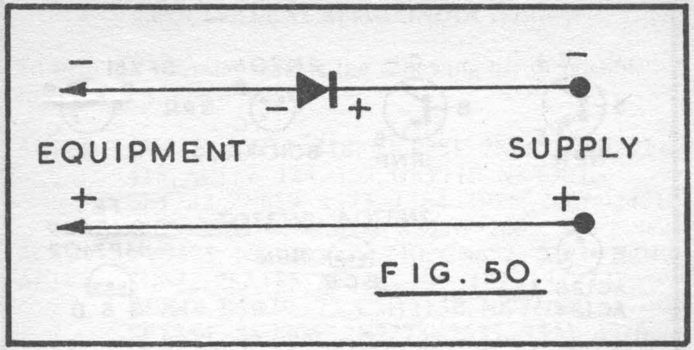

FIG. 50.

Component Tolerances

For circuits in this book, exact component values are generally not too important. Capacitor values such as 47nF (0.047μF) and 0.05μF may be regarded as interchangeable. In the same way, a 125μF capacitor may replace 100μF, or 200μF can be used instead of 220μF, and so on.

Resistors will generally be the standard 5% tolerance types. In some circuits exact resistor values are not at all critical. In others, and especially base, emitter and collector circuits, the values which provide best results can have much less latitude.

Bases

The lead-out positions of base, emitter and collector wires have been shown for typical transistors. However, it is felt that Fig. 51 should be useful. This gives lead information for types most likely to be required for circuits in this book. Other types of transistors may of course be used successfully, but it is not practical to list all these, or show connections for them all.

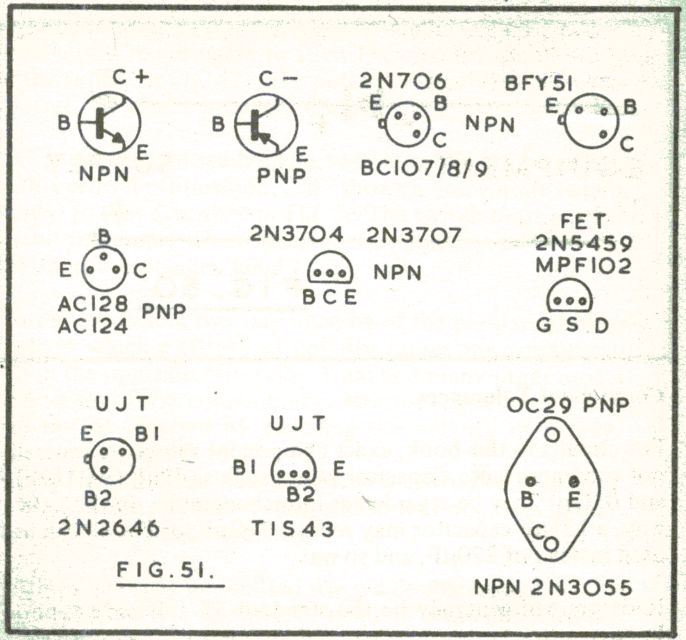
C +
B
E
NPN
C –
B
E
PNP
2N706
E
B
C
BC107/8/9
NPN
BFY51
E
B
C
B
E
C
AC128
AC124
PNP
2N3704
2N3707
B C E
NPN
FET
2N5459
MPF102
G S D
UJT
E
B1
B2
2N2646
UJT
B1
B2
E
TIS43
OC29 PNP
B
E
C
NPN 2N3055

FIG. 51.

EQUIVALENT SEMICONDUCTORS

To help our overseas readers the following list of possible equivalents is shown below: –

AC128	CV9778, OC74–318, NKT281, 2N467, 2SB222–415, AC117–124–153, GET110, NKT302, SFT232, 2N659–1373–1384–1926–2001–2431–4106, 2SB370, SK3004, RS276–2005.
ACY39	CV9475, NKT243, ACZ10, 2N652–2303–2801.
AF117	2N642, 2SA155, AF102–105–127–133–137, BFX48, BSW19–73, GM1213B, NKT603F, SFT354, SK3006, 2N2273–2635–3324–4916–5354, RS276–2003.
BC107	CV9780, AM251, BC147–167–207–317, MPS6566, SK3020–3122, ZTX107, TT107, RS276–2009.
BC108	CV10541, AM252, BC148–168–208–318, MPS6520, SK3020, ZTX108, TT108, RS276–2009.
BC109	CV10769–10806, AM253, BC149–169–209–319, MPS6521, SK3020, ZTX109, TT109, RS276–2009.
BC147	BC171–182/L–207–317, MPS6566, SK3020, RS276–2013,
BC148	BC172–183/L–208–318, MPS6520, SK3020, BS276–2009.
BC149	BC184/L–209–173–319, MPS6521, SK3020, RS276–2009.
BF200	BF162–173–225–271–311–314, SE5020/1/2/3/4, SK3018, 2N616–2708–3288–3693, RS276–2011.
BFY51	CV7723–7726–10827–11238, BFX51–68A, BFR19, BSX45/X, BSY46–85, SK3024, 40347, 2SC708, ME6102 2N2193/4–2218–2410–3053–3252–4046, RS276–2009.
BSY95A	CV7648–9379, BC108A, BSX51–73–91, BFX96, BSY20–58–62, ME9022, SK3122, 2SC321, 2N744–914–1708–2218–2369–5187, RS276–2011.
MPF102	E305, SK3116.
OA81	AA117–118–132–144, AAZ15, OA91–95–161, SD38, 1N38–56–476–618, 1S33.
OC26	CV8560, NKT452, 2N257–297–458–553–1021, 2SB3, AD130–131–138–140–149, OC28, SK3009, TF80, T1156–3027, 2N174–1073–2870–RS276–2006.

OC29	CV7083–8356, NKT402, 2N457, 2SB86–425, AD150, ASZ16, AUY21, OC28–35–36, SK3009, T13028, 2N174–1073–2870, RS276–2006.
OC44	CV5710–7003, OC170–171–410–613, 2N1303, 2SA15, AF101–117–126, ASY55, AC191, BC126, NKT211, SKT358, SK3005, 2N36–1191/2–1352–1373–1384, RS276–2003.
OC70	CV5457, OC303–602, NKT214, 2N279–1305, 2SB75, AC122–125–151–163, ACY27–34, ASY27, BC213–206A, BSX36, SK3003, 2N1193–1352–1375–2429, RS276–2004.
OC71	CV5712–7005–8344, OC70–304–604, NKT214, 2N280–1305, 2SB77, AC122–125–151–163, ACY35, BSX36, OC3041, SK3003, 2N1193–1352–1375–1384–2429, RS276–2004.
OC72	AC122/3–125/6–131–151 IV–152–162–191, BC126–213–206A, NKT211, SFT353, 2N282–1190–1352–1371–1384–1991–2431, CV5713–7006–8440, RS276–2005.
OC76	CV8314, OC307–6025, 2N284–1305, NKT212, 2SB89, AC128–131–151–152, ACY28, ASY58–76, BC126–213–260A, SK3004, 2N394–1191–1352–1371–1384–1991, RS276–2004.
ORP12	RS276–116, LDR03, RPY25.
1N4001	BY127, EM502, RS276–1101/1135, 1S100.
1N4002	BY127, EM502, RS276–1102/1136, 1S100.
2N1306	CV7350–10686, NKT736, 2N634–635–636–1891–1993–1994–1995, ASY26–29, BSX19–20, RS276–2001.
2N1613	CV7440–8843, BFY33–34–67, BSY44, 2N698–699–717–1711–1889–1890–1893, BFY41, BSX45/6, BSY53, MPS6530, 2SC708, 2N2101.

N.B.: Although equivalent semiconductors may have similar electrical properties, physical dimensions may be different and this must be born in mind if space is tight and for mounting details.

Remember polarities if replacing PNP with NPN types and vice versa. If in doubt always be advised by your dealer for suitable equivalent semiconductors.

Notes

Please note overleaf is a list of other titles that are available in our range of Radio and Electronic Books.

These should be available from most good Booksellers, Radio Component Dealers and Mail Order Companies.

However, should you experience difficulty in obtaining any title in your area, then please write directly to the publishers enclosing payment to cover the cost of the book plus adequate postage.

If you would like a copy of our latest catalogue of Radio & Electronics Books then please send a Stamped Addressed Envelope.

BERNARD BABANI (publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

160	Coil Design and Construction Manual	£1.95
202	Handbook of Integrated Circuits (IC's) Equivalents & Substitutes	£1.95
205	First Book of Hi-Fi Loudspeaker Enclosures	£0.95
208	Practical Stereo and Quadrophony Handbook	£0.75
214	Audio Enthusiasts Handbook	£0.85
219	Solid State Novelty Projects	£0.85
220	Build Your Own Solid State Hi-Fi and Audio Accessories	£0.85
221	28 Tested Transistor Projects	£1.25
222	Solid State Short Wave Receivers for Beginners	£1.95
223	50 Projects Using IC CA3130	£1.25
224	50 CMOS IC Projects	£1.35
225	A Practical Introduction to Digital IC's	£1.75
226	How to Build Advanced Short Wave Receivers	£1.95
227	Beginners Guide to Building Electronic Projects	£1.95
228	Essential Theory for the Electronics Hobbyist	£1.95
RCC	Resistor Colour Code Disc	£0.20
BP1	First Book of Transistor Equivalents and Substitutes	£1.50
BP2	Handbook of Radio, TV, Ind & Transmitting Tube & Valve Equivalents	£0.60
BP6	Engineers and Machinists Reference Tables	£0.75
BP7	Radio and Electronic Colour Codes and Data Chart	£0.40
BP14	Second Book of Transistor Equivalents & Substitutes	£1.75
BP24	52 Projects Using IC741	£1.75
BP27	Chart of Radio, Electronic, Semi-conductor and Logic Symbols	£0.50
BP28	Resistor Selection Handbook	£0.60
BP29	Major Solid State Audio Hi-Fi Construction Projects	£0.85
BP32	How to Build Your Own Metal and Treasure Locators	£1.95
BP33	Electronic Calculator Users Handbook	£1.50
BP34	Practical Repair and Renovation of Colour TVs	£1.95
BP36	50 Circuits Using Germanium, Silicon and Zener Diodes	£1.50
BP37	50 Projects Using Relays, SCR's and TRIACS	£1.95
BP39	50 (FET) Field Effect Transistor Projects	£1.75
BP42	50 Simple L.E.D. Circuits	£1.50
BP43	How to Make Walkie-Talkies	£1.95
BP44	IC 555 Projects	£1.95
BP45	Projects in Opto-Electronics	£1.95
BP47	Mobile Discotheque Handbook	£1.35
BP48	Electronic Projects for Beginners	£1.95
BP49	Popular Electronic Projects	£1.95
BP50	IC LM3900 Projects	£1.35
BP51	Electronic Music and Creative Tape Recording	£1.95
BP52	Long Distance Television Reception (TV-DX) for the Enthusiast	£1.95
BP53	Practical Electronics Calculations and Formulae	£2.95
BP54	Your Electronic Calculator and Your Money	£1.35
BP55	Radio Stations Guide	£1.75
BP56	Electronic Security Devices	£1.95
BP57	How to Build Your Own Solid State Oscilloscope	£1.95
BP58	50 Circuits Using 7400 Series IC's	£1.75
BP59	Second Book of CMOS IC Projects	£1.95
BP60	Practical Construction of Pre-amps, Tone Controls, Filters & Attenuators	£1.95
BP61	Beginners Guide To Digital Techniques	£0.95
BP62	Elements of Electronics – Book 1	£2.95
BP63	Elements of Electronics – Book 2	£2.25
BP64	Elements of Electronics – Book 3	£2.25
BP65	Single IC Projects	£1.50
BP66	Beginners Guide to Microprocessors and Computing	£1.95
BP67	Counter, Driver and Numeral Display Projects	£1.75
BP68	Choosing and Using Your Hi-Fi	£1.65
BP69	Electronic Games	£1.75
BP70	Transistor Radio Fault-Finding Chart	£0.50
BP71	Electronic Household Projects	£1.75
BP72	A Microprocessor Primer	£1.75
BP73	Remote Control Projects	£1.95
BP74	Electronic Music Projects	£1.75
BP75	Electronic Test Equipment Construction	£1.75
BP76	Power Supply Projects	£1.95
BP77	Elements of Electronics – Book 4	£2.95

BP78	Practical Computer Experiments	£1.75
BP79	Radio Control for Beginners	£1.75
BP80	Popular Electronic Circuits – Book 1	£1.95
BP81	Electronic Synthesiser Projects	£1.75
BP82	Electronic Projects Using Solar Cells	£1.95
BP83	VMOS Projects	£1.95
BP84	Digital IC Projects	£1.95
BP85	International Transistor Equivalents Guide	£2.95
BP86	An Introduction to BASIC Programming Techniques	£1.95
BP87	Simple L.E.D. Circuits – Book 2	£1.35
BP88	How to Use Op-Amps	£2.25
BP89	Elements of Electronics – Book 5	£2.95
BP90	Audio Projects	£1.95
BP91	An Introduction to Radio DXing	£1.95
BP92	Easy Electronics – Crystal Set Construction	£1.75
BP93	Electronic Timer Projects	£1.95
BP94	Electronic Projects for Cars and Boats	£1.95
BP95	Model Railway Projects	£1.95
BP96	C B Projects	£1.95
BP97	IC Projects for Beginners	£1.95
BP98	Popular Electronic Circuits – Book 2	£2.25
BP99	Mini-Matrix Board Projects	£1.95
BP100	An Introduction to Video	£1.95
BP101	How to Identify Unmarked IC's	£0.65
BP102	The 6809 Companion	£1.95
BP103	Multi-Circuit Board Projects	£1.95
BP104	Electronic Science Projects	£2.25
BP105	Aerial Projects	£1.95
BP106	Modern Op-Amp Projects	£1.95
BP107	30 Solderless Breadboard Projects – Book 1	£2.25
BP108	International Diode Equivalents Guide	£2.25
BP109	The Art of Programming the 1K ZX81	£1.95
BP110	How to Get Your Electronic Projects Working	£1.95
BP111	Elements of Electronics – Book 6	£3.50
BP112	A Z-80 Workshop Manual	£2.75
BP113	30 Solderless Breadboard Projects – Book 2	£2.25
BP114	The Art of Programming the 16K ZX81	£2.50
BP115	The Pre-Computer Book	£1.95
BP116	Electronic Toys Games & Puzzles	£2.25
BP117	Practical Electronic Building Blocks – Book 1	£1.95
BP118	Practical Electronic Building Blocks – Book 2	£1.95
BP119	The Art of Programming the ZX Spectrum	£2.50
BP120	Audio Amplifier Fault-Finding Chart	£0.65
BP121	How to Design and Make your Own P.C.B.s	£1.95
BP122	Audio Amplifier Construction	£2.25
BP123	A Practical Introduction to Microprocessors	£2.25
BP124	Easy Add-on Projects for Spectrum ZX81 & Ace	£2.75
BP125	25 Simple Amateur Band Aerials	£1.95
BP126	BASIC & PASCAL in Parallel	£1.50
BP127	How to Design Electronic Projects	£2.25
BP128	20 Programs for the ZX Spectrum & 16K ZX81	£1.95
BP129	An Introduction to Programming the ORIC-1	£1.95
BP130	Micro Interfacing Circuits – Book 1	£2.25
BP131	Micro Interfacing Circuits – Book 2	£2.25
BP132	25 Simple Shortwave Broadcast Band Aerials	£1.95
BP133	An Introduction to Programming the Dragon 32	£1.95
BP134	Easy Add-on Projects for Commodore 64 & Vic-20	£2.50
BP135	The Secrets of the Commodore 64	£2.50
BP136	25 Simple Indoor and Window Aerials	£1.95
BP137	BASIC & FORTRAN in Parallel	£1.95
BP138	BASIC & FORTH in Parallel	£1.95
BP139	An Introduction to Programming the BBC Model B Micro	£2.50
BP140	Digital IC Equivalents and Pin Connections	£3.95
BP141	Linear IC Equivalents and Pin Connections	£3.95
BP142	An Introduction to Programming the Acorn Electron	£1.95
BP143	An Introduction to Programming the Atari 600XL and 800XL	£2.50
BP144	Further Practical Electronics Calculations and Formulae	£3.75
BP145	25 Simple Tropical and M.W. Band Aerials	£1.95

BERNARD BABANI BP48

Electronic Projects for Beginners

- In this book the newcomer to electronics will find a wide range of easily made projects, many complete with actual component and wiring layouts. Furthermore, a number of projects have been arranged so that they can be constructed without any need for soldering and, thus, avoid the need for a soldering iron.
- This book which is written by the very experienced author Mr. F.G. Rayer is divided into four sections:–
  1. "No Soldering" Projects
  2. Miscellaneous Devices
  3. Radio and Audio Frequency
  4. Power Supplies
- An absolute 'must' for all beginners in electronics.

ISBN 0-85934-054-6
00195
£1.95
9 780859 340540

Microscope

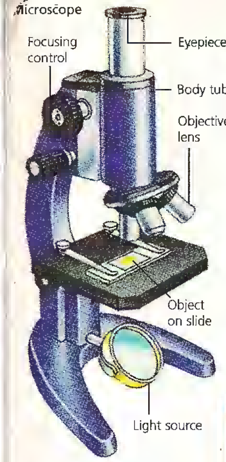
Focusing control
Eyepiece
Body tube
Objective lens
Object on slide
Light source

Light bulb

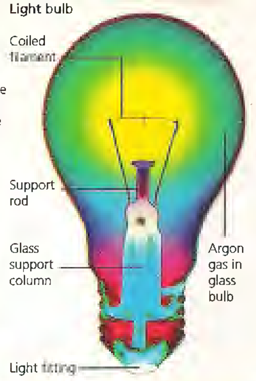
Coiled filament
Support rod
Glass support column
Argon gas in glass bulb
Light fitting

Human brain (cross section)

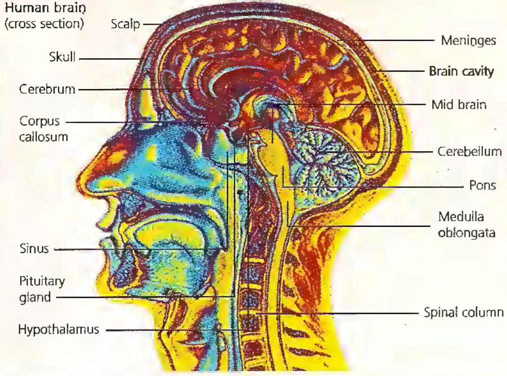
Scalp
Skull
Cerebrum
Corpus callosum
Sinus
Pituitary gland
Hypothalamus
Meninges
Brain cavity
Mid brain
Cerebellum
Pons
Medulla oblongata
Spinal column

Magnets

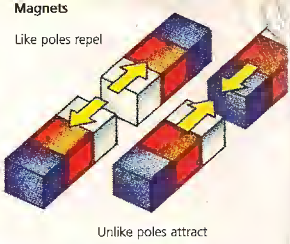
Like poles repel
Unlike poles attract

USBORNE

THE USBORNE ILLUSTRATED DICTIONARY OF SCIENCE

A COMPLETE REFERENCE GUIDE TO PHYSICS, CHEMISTRY AND BIOLOGY

WITH RECOMMENDED WEB SITES

Battery (cutaway)

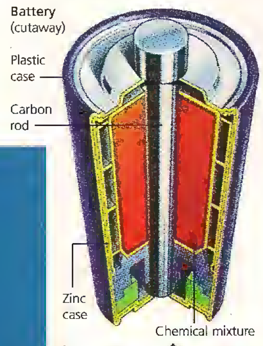
Plastic case
Carbon rod
Zinc case
Chemical mixture

Lenses

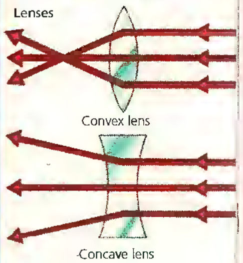
Convex lens
Concave lens

Human molar (cross section)

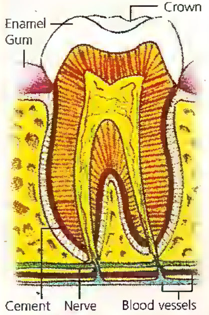
Crown
Enamel
Gum
Cement
Nerve
Blood vessels

Buttercup

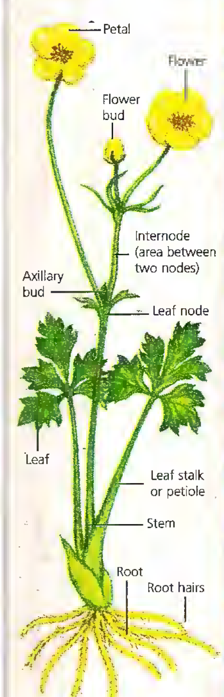
Petal
Flower
Flower bud
Internode (area between two nodes)
Axillary bud
Leaf node
Leaf
Leaf stalk or petiole
Stem
Root
Root hairs

Surface tension

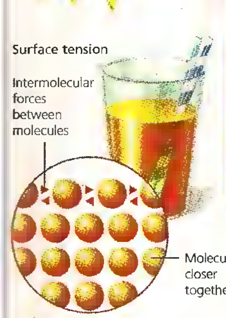
Intermolecular forces between molecules
Molecules closer together

Gears

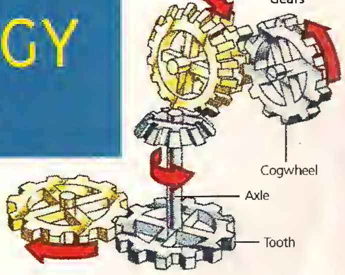
Cogwheel
Axle
Tooth

Bicycle dynamo (cutaway)

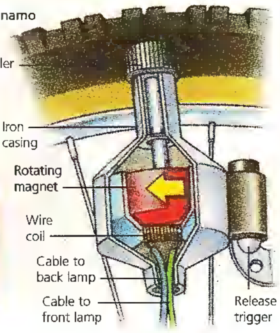
Roller
Iron casing
Rotating magnet
Wire coil
Cable to back lamp
Cable to front lamp
Release trigger

Atom model

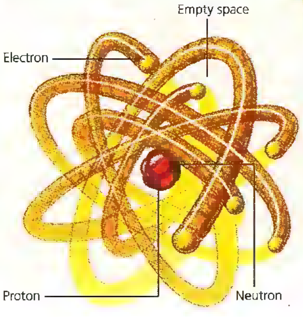
Empty space
Electron
Proton
Neutron

Human ear

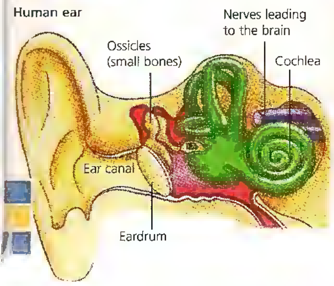
Ossicles (small bones)
Nerves leading to the brain
Cochlea
Ear canal
Eardrum

Crystals

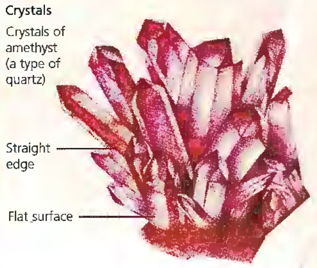
Crystals of amethyst (a type of quartz)
Straight edge
Flat surface

Usborne-recommended Web sites

At **www.usborne-quicklinks.com** we have created links to useful science resources on the Internet. There are links to homework and review sites, interactive science puzzles, tests and quizzes, online calculators and unit converters, sites where you can look up the formula for any chemical or view 3-D molecules, and "ask a scientist" sites where you can e-mail your questions to online experts. For quick and easy access to all the Usborne-recommended Web sites, go to **www.usborne-quicklinks.com** and enter the keywords "science dictionary".

The Web sites recommended in Usborne Quicklinks have been selected by Usborne editors as suitable, in their opinion, for children. The sites are regularly reviewed and updated. However, the content of a Web site may change at any time and Usborne Publishing is not responsible for the content on any Web site other than its own. We recommend that children are supervised while on the Internet, that they do not use Internet Chat Rooms, and that you use Internet filtering software to block unsuitable material. Please ensure that children read and follow the safety guidelines displayed on the Usborne Quicklinks Web site. For more information, see the Net Help area in Usborne Quicklinks.

For links to Usborne-recommended
science resources on the Internet, go to
www.usborne-quicklinks.com
and enter the keywords "science dictionary".

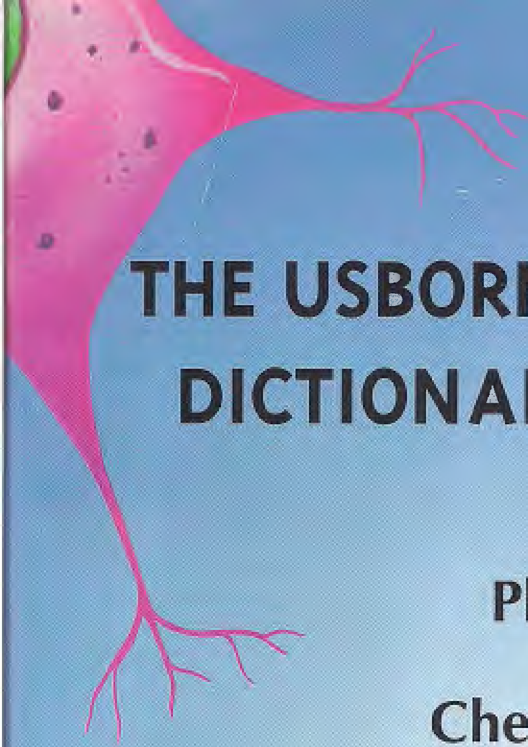

THE USBORNE ILLUSTRATED DICTIONARY OF SCIENCE

Physics 1

Chemistry 115

Biology 229

Glossary 344

Index 346

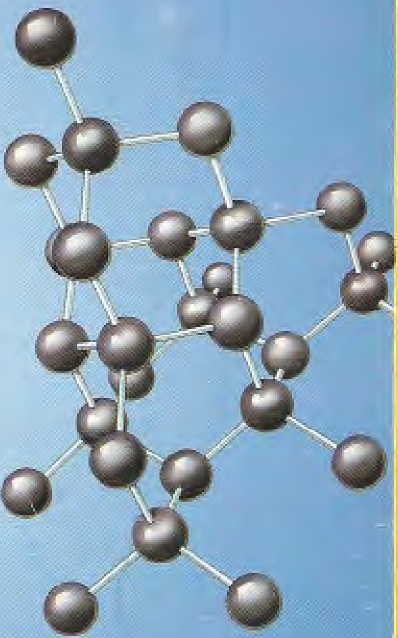

Corinne Stockley,
Chris Oxlade and Jane Wertheim
Revision editor: Kirsteen Rogers

Designers: Karen Tomlins and Verinder Bhachu
Digital illustrator: Fiona Johnson
Illustrators: Kuo Kang Chen and Guy Smith

Scientific advisors:
Dr. Tom Petersen, John Hawkins, Dr. John Durell
and Jerry McCoy (Physics) Dr. John Waterhouse,
Nick Christou, John Raffan, Rae Michaelis, Alan Alder and
Dr. Larry Scroggins (Chemistry) Dr. Margaret Rostron,
Dr. John Rostron and Dr. Jan Mercer (Biology)

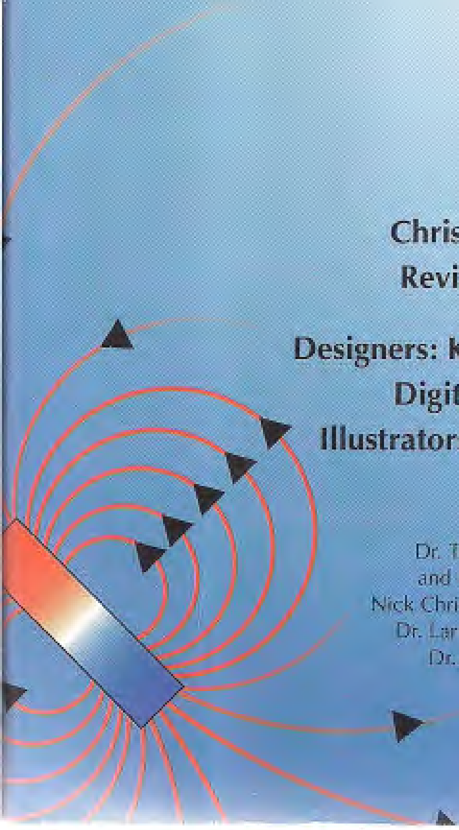

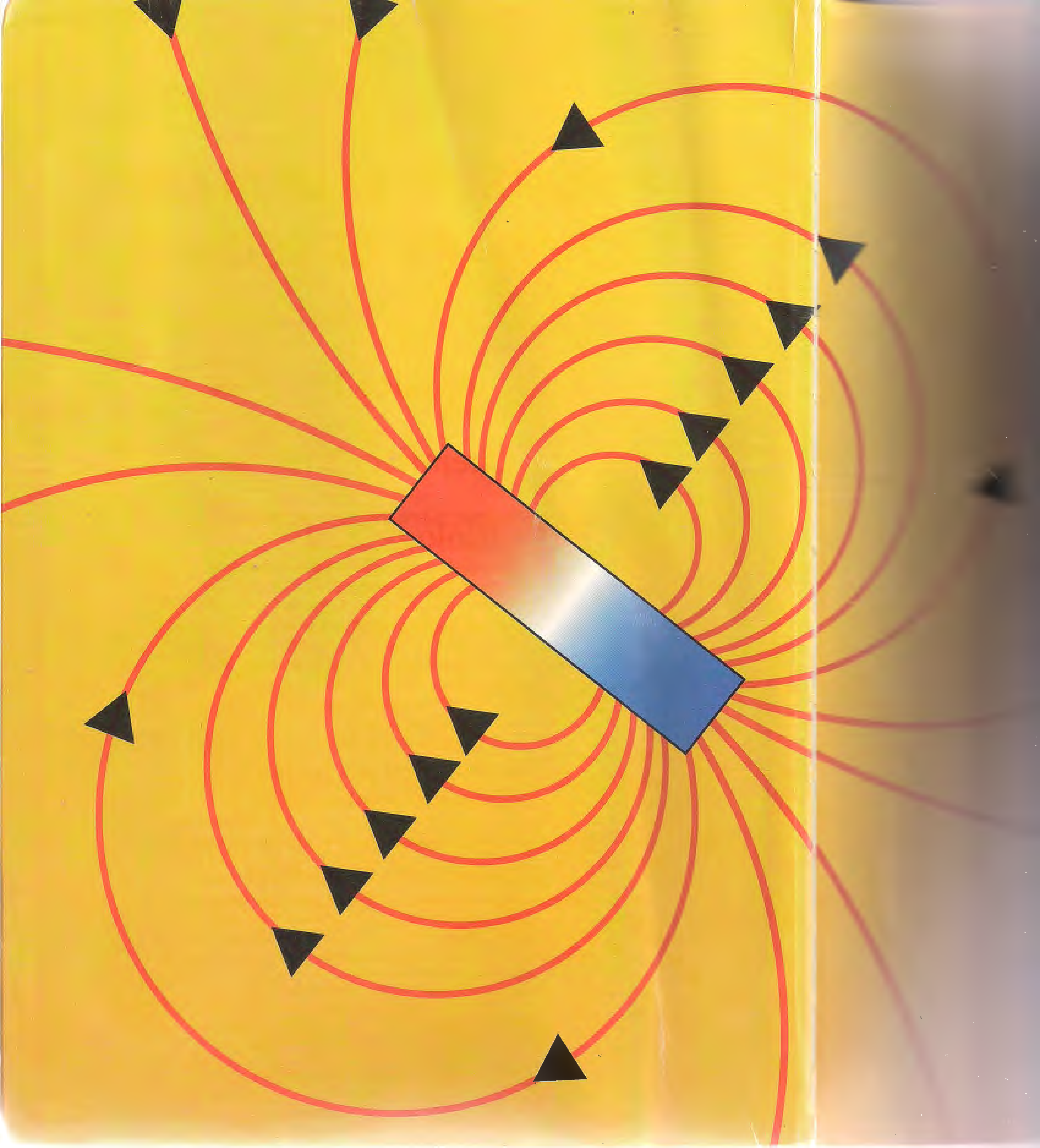

PHYSICS

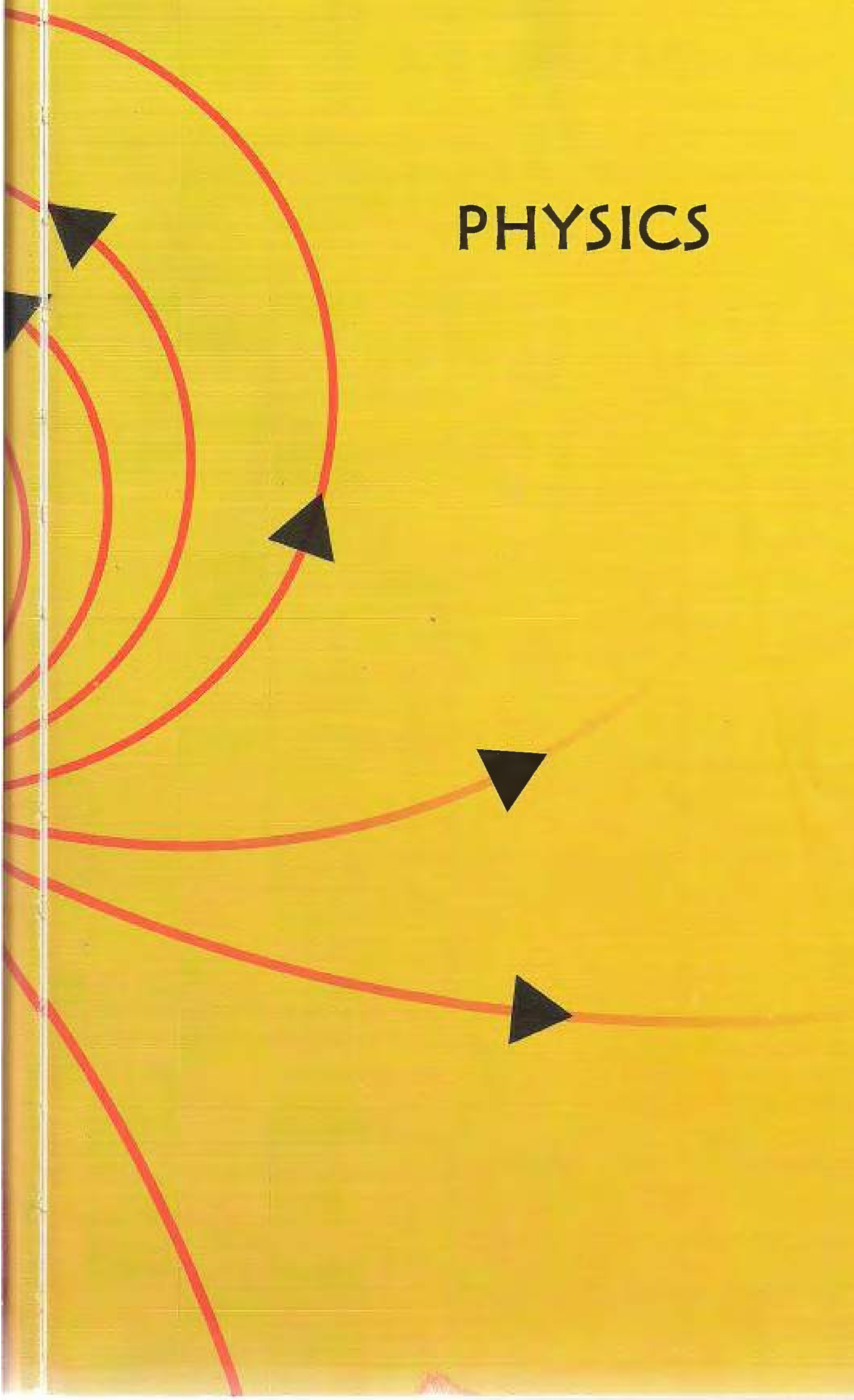

ABOUT PHYSICS

Physics is the study of the properties and nature of matter, the different forms of energy and the ways in which matter and energy, interact in the world around us. In this book, physics is divided into six colour-coded sections. The areas covered by these sections are explained below.

● Mechanics and general physics

Covers the main concepts of physics, e.g. forces, energy and the properties of matter.

● Electricity and magnetism

Explains the forms, uses and behaviour of these two linked phenomena.

● Heat

Explains heat energy in terms of its measurement and the effects of its presence and transference. Includes the gas laws.

● Atomic and nuclear physics

Examines atomic and nuclear structure and energy, radioactivity, fission and fusion.

● Waves

Looks at the properties and effects of wave energy and examines sound, electromagnetic and light waves in detail.

● General physics information

General material – charts and tables, also information on the treatment of experimental results.

CONTENTS

Mechanics and general physics

4 Atoms and molecules
6 Forces
8 Energy
10 Motion
12 Dynamics
14 Turning forces
16 Periodic motion
18 Gravitation
20 Machines
22 Molecular properties
24 Density

Heat

26 Temperature
28 Transfer of heat
30 Effects of heat transfer
32 Expansion upon heating
33 Behavior of gases

Waves

34 Waves
36 Reflection, refraction and diffraction
38 Wave interference
40 Sound waves
42 Perception of sound
44 Electromagnetic waves
46 Light
47 Reflection of light
50 Refraction of light
54 Optical instruments

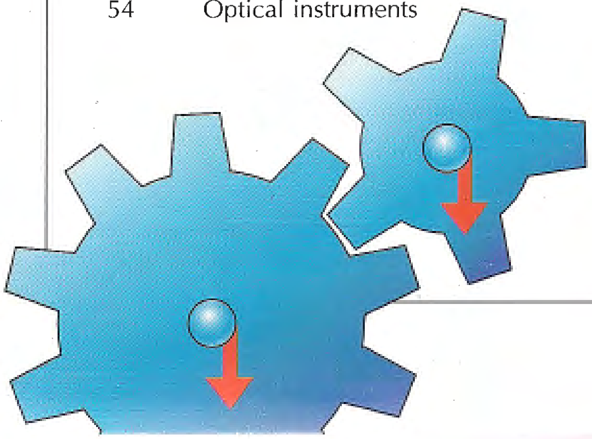

Electricity and magnetism

56 Static electricity
58 Potential and capacitance
60 Electric current
62 Controlling current
65 Semiconductors
66 Electrolysis
68 Cells and batteries
70 Magnets
72 Magnetic fields
74 Electromagnetism
77 Electric meters
78 Electromagnetic induction
80 Cathode rays

Atomic and nuclear physics

82 Atomic structure
84 Atomic and nuclear energy
86 Radioactivity
88 Detecting and measuring radioactivity
91 Uses of radioactivity
92 Nuclear fission and fusion
94 Power from nuclear reactions

General physics information

96 Quantities and units
98 Equations, symbols and graphs
100 Measurements
102 Accuracy and errors
104 Fields and forces
108 Vectors and scalars
109 Numbers
110 Circuit symbols
111 Transistors and gates
112 Properties of substances
113 Useful constants and values

ATOMS AND MOLECULES

The Ancient Greeks believed that all matter was made up of tiny particles which they called **atoms**. This idea has since been expanded and theories such as the **kinetic theory** have been developed which can be used to explain the physical nature and behavior of substances in much greater detail. Matter can exist in three different **physical states**. The state of a substance depends on the nature of the substance, its temperature and the pressure exerted on it. Changes between states are caused by changes in the pressure or temperature (see **changes of state**, page 30).

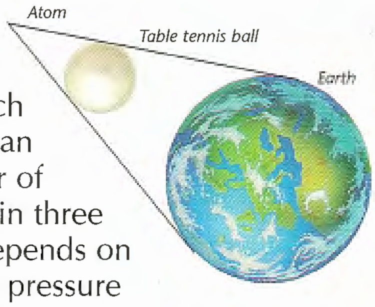

If atoms were the size of table tennis balls, by the same scale, table tennis balls would be as big as the Earth.

Atom

The smallest part of a substance which can exist and still retain the properties of the substance. The internal structure of the atom is explained on pages 82-83. Atoms are extremely small, having radii of about 10^{-10}m and masses of about 10^{-25}kg. They can form **ions*** (electrically charged particles) by the loss or gain of **electrons*** (see **ionization**, page 88).

Diagram showing relative sizes of some atoms

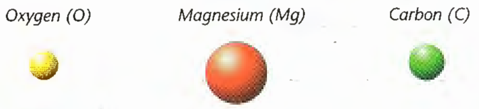

Molecule

The smallest naturally-occurring particle of a substance. Molecules can consist of any number of **atoms**, from one (e.g. neon) to many thousands (e.g. proteins), all held together by **electromagnetic forces***. All the molecules of a pure sample of a substance contain the same atoms in the same arrangement.

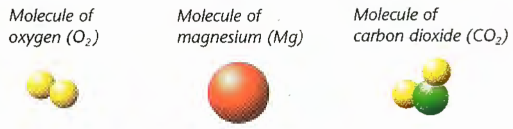
Molecule of oxygen (O_2)
Molecule of magnesium (Mg)
Molecule of carbon dioxide (CO_2)

Note that many substances do not have **molecules**, for example:

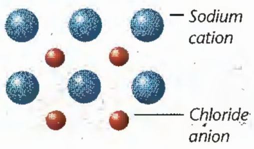

Ionic compound* *of* ***anions**** *and* ***cations****.

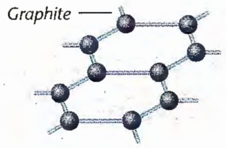

***Atomic lattice** of **atoms** all bonded together.*

Element

A substance which cannot be split into simpler substances by a chemical reaction. All **atoms** of the same element have the same number of **protons*** in their **nuclei*** (see **atomic number**, page 82).

Compound

A substance whose **molecules** contain the **atoms** of two or more **elements**, chemically bonded together, and which can thus be split into simpler substances. A **mixture** has no chemical bonding and is therefore not a compound.

***Element** 1*

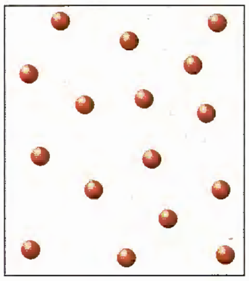

***Element** 2*

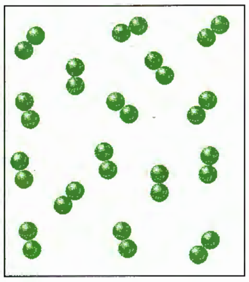

***Compound** of elements 1 and 2 – elements bonded together.*

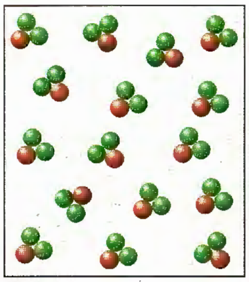

***Mixture** of elements 1 and 2 – no chemical bonding.*

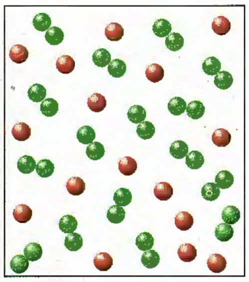

---

* **Anions**, **Cations**, 88 (**Ionization**); **Electromagnetic force**, 6; **Electrons**, 83; **Ionic compound**, 131; **Ions**, 88 (**Ionization**); **Nucleus**, **Protons**, 82.

Physical states

Solid state

A **state** in which a substance has a definite volume and shape and resists forces which try to change these.

Liquid state

A **state** in which a substance flows and takes up the shape of its containing vessel. It is between the **solid** and **gaseous** states.

Gaseous state

A **state** in which a substance expands to fill its containing vessel. Substances in this state have a relatively low density.

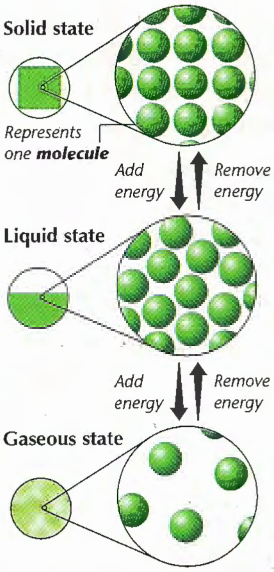

Molecules *vibrate about mean positions, having* ***molecular potential energy**** *and* ***vibrational kinetic energy****.

Average energy of molecule much less than that needed by it to break free from other molecules.

Energy added breaks down regular pattern – molecules can move around and thus have both ***translational*** *and* ***rotational kinetic energy****.

Average energy of molecule is just enough for it to break free from neighboring molecules, only to be captured by the next ones along.

Molecules have very large separation – they move virtually independently of each other – ***intermolecular forces**** *can be ignored.*

Average energy of molecule much greater than needed to break free from other molecules.

Gas

A substance in the **gaseous state** which is above its **critical temperature** and so cannot be turned into a liquid just by increasing the pressure – the temperature must be lowered first, to create a **vapor**.

Vapor

A substance in the **gaseous state** which is below its **critical temperature** (see **gas**) and so can be turned into a liquid by an increase in pressure alone – no lowering of temperature is required.

The kinetic theory

The **kinetic theory** explains the behavior of the different physical states in terms of the motion of **molecules.** In brief, it states that the molecules of **solids** are closest together, have least energy and so move the least, those of **liquids** are further apart with more energy, and those of **gases** are furthest apart with most energy. See above right.

Brownian motion

The observed random motion of small particles in water or air. It supports the kinetic theory, as it could be said to be due to impact with water or air **molecules.**

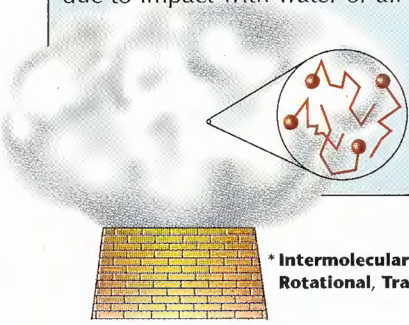

Brownian motion *of smoke particles as they are hit by* ***molecules*** *in the air.*

Diffusion

The mixing of two **gases**, **vapors** or **liquids** over a period of time. It supports the kinetic theory, since the particles must be moving to mix, and gases can be seen to diffuse faster than liquids.

Molecules *of two* ***gases*** *diffuse together over time.*

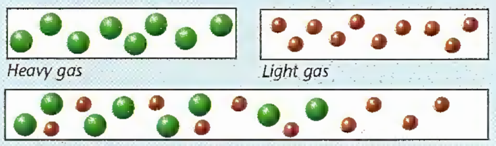

Light gas diffuses faster than heavy one.

Graham's law of diffusion

States that, at constant temperature and pressure, the rate of **diffusion** of a **gas** is inversely proportional to the square root of its density.

$$\text{Rate of diffusion} \propto \sqrt{\frac{1}{\text{density of gas}}}$$

---

* **Intermolecular forces**, 7; **Molecular potential energy**, 8;
Rotational, **Translational** and **Vibrational kinetic energy**, 9 (**Kinetic energy**).

FORCES

A **force** influences the shape and motion of an object. A single force will change its velocity (i.e. **accelerate*** it) and possibly its shape. Two equal and opposite forces may change its shape or size. It is a **vector quantity***, having both magnitude and direction, and is measured in **newtons**. The main types of force are **gravitational**, **magnetic**, **electric** and **strong nuclear**. See pages 104-107 for a comparison of the first three of these.

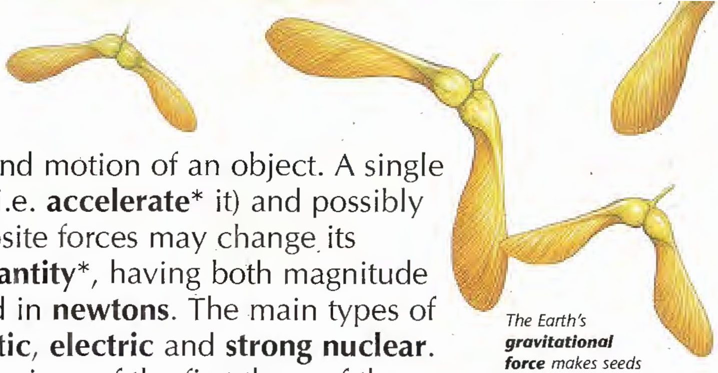

*The Earth's **gravitational force** makes seeds fall to the ground.*

Showing forces in diagrams

***Forces** are shown by arrowed lines (the length represents magnitude and the arrow represents direction).*

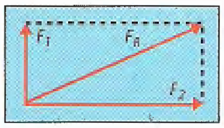

*Effect of F_1 and F_2 is the same as F_R (the **resultant force**). F_1 and F_2 are the **components** of F_R.*

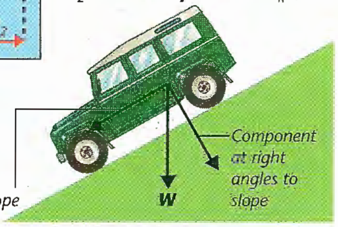
Force W resolved into two **components**
Component parallel to slope
W
Component at right angles to slope

Newton (N)

The **SI unit*** of force. One newton is the force needed to accelerate a mass of 1kg by 1m s^{-2}.

Force field

The region in which a force has an effect. The maximum distance over which a force has an effect is the **range** of the force. Force fields are represented by lines with arrows, called **field lines**, to show their strength and direction (see also pages 58 and 72).

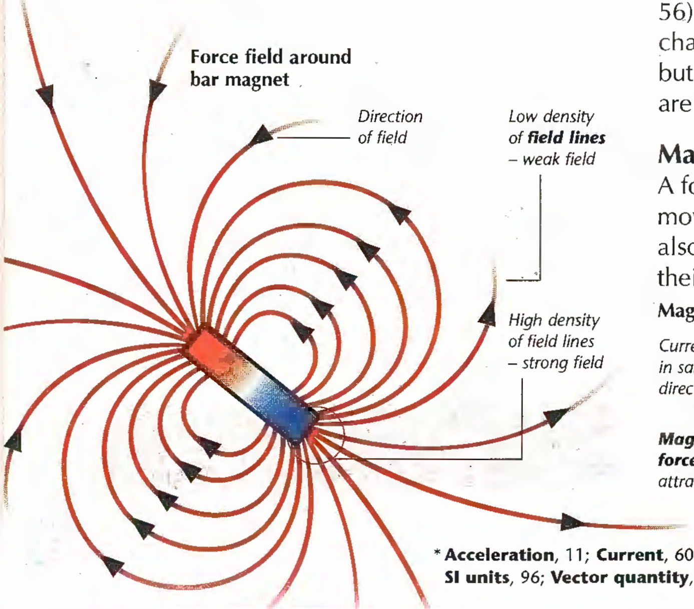

Force field around bar magnet

Direction of field
Low density of **field lines** – weak field
High density of field lines – strong field

Gravitational force or gravity

The force of attraction between any two objects which have mass (see also pages 18-19). It is very small unless one of the objects is very massive.

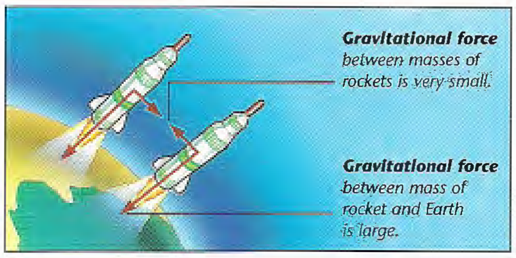
Gravitational force between masses of rockets is very small.
Gravitational force between mass of rocket and Earth is large.

Electromagnetic force

A combination of the **electric** and **magnetic forces**, which are closely related and difficult to separate.

Electric or electrostatic force

The force between two electrically-charged particles (see also page 56). It is repulsive if the charges are the same, but attractive if they are opposite.

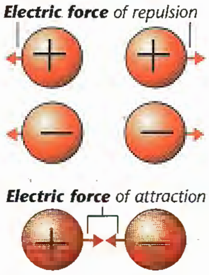
Electric force of repulsion
Electric force of attraction

Magnetic force

A force between two moving charges. These moving charges can be electric **currents*** (see also page 60) or **electrons*** moving around in their **electron shells***.

Magnetic forces in electric wires

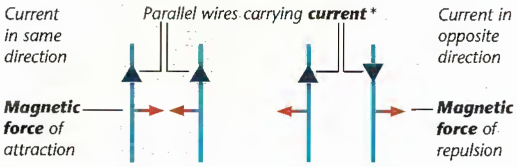
Current in same direction
Parallel wires carrying **current***
Current in opposite direction
Magnetic force of attraction
Magnetic force of repulsion

---

* **Acceleration**, 11; **Current**, 60; **Electrons**, **Electron shells**, 83; **SI units**, 96; **Vector quantity**, 108.

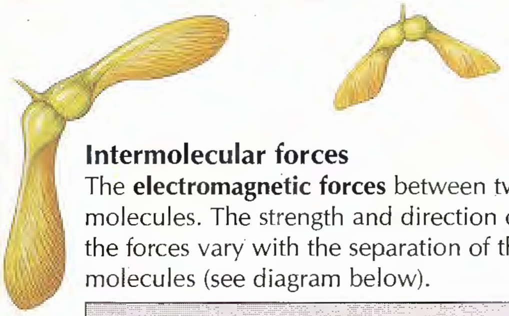

Intermolecular forces

The **electromagnetic forces** between two molecules. The strength and direction of the forces vary with the separation of the molecules (see diagram below).

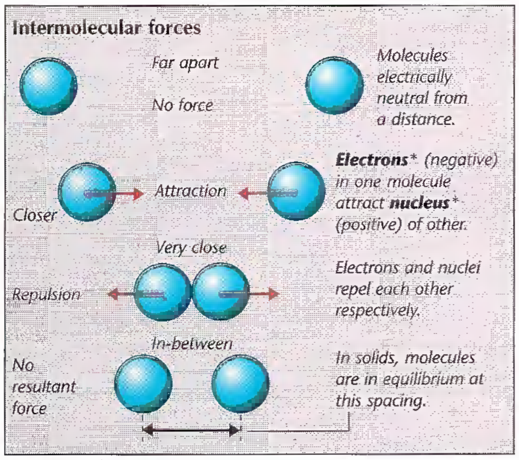
Intermolecular forces
Far apart
No force
Molecules electrically neutral from a distance.
Closer
Attraction
Electrons* (negative) in one molecule attract **nucleus*** (positive) of other.
Very close
Repulsion
Electrons and nuclei repel each other respectively.
In-between
No resultant force
In solids, molecules are in equilibrium at this spacing.

Tension

Equal and opposite forces which, when applied to the ends of an object, increase its length. They are resisted by the **intermolecular force** of attraction.

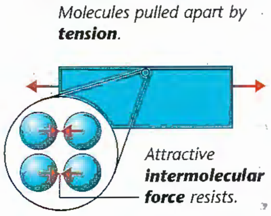
Molecules pulled apart by ***tension****.*
Attractive **intermolecular force** resists.

Compression

Equal and opposite forces which decrease the length of an object. They are opposed by the **intermolecular force** of repulsion.

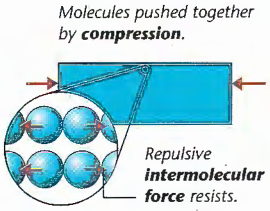
Molecules pushed together by ***compression****.*
Repulsive **intermolecular force** resists.

Contact force

The **intermolecular force** of repulsion between the molecules of two objects when they touch.

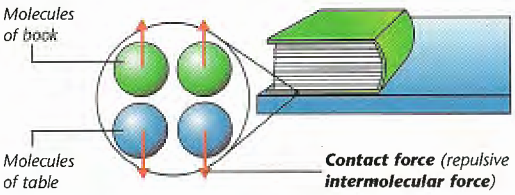
Molecules of book
Molecules of table
Contact force (repulsive **intermolecular force**)

Strong nuclear force

The force of attraction between all the particles of an atomic **nucleus*** (the **protons*** and **neutrons***). It prevents the **electric force** of repulsion between the protons from pushing the nucleus apart (see also page 84).

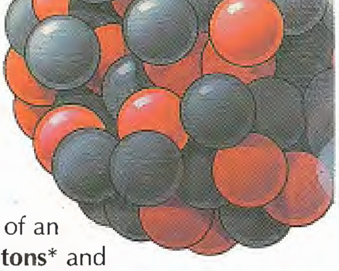
Particles in an atomic nucleus are held together by the ***nuclear force****.*

Frictional force or friction

The force which acts to oppose the motion of two touching surfaces over each other, caused by the **intermolecular force** of attraction between the molecules of the surfaces. There are two types, the **static** and the **kinetic frictional force**.

Static frictional force

The **frictional force** between two touching surfaces when a force is applied to one of them but they are not moving. The maximum value of the static frictional force occurs when they are on the point of sliding over each other. This is called the **limiting force**.

Kinetic frictional force or sliding frictional force

The **frictional force** when one surface is sliding over another at constant speed. It is slightly less than the **limiting force** (the maximum **static frictional force**).

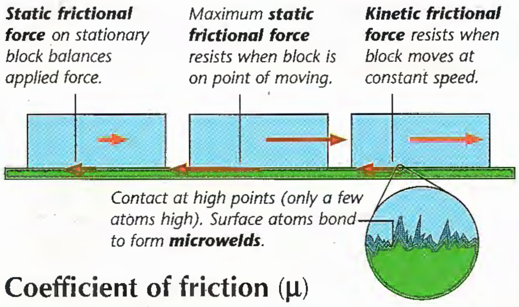
Static frictional force on stationary block balances applied force.
Maximum **static frictional force** resists when block is on point of moving.
Kinetic frictional force resists when block moves at constant speed.
Contact at high points (only a few atoms high). Surface atoms bond to form **microwelds**.

Coefficient of friction (μ)

The ratio of the **frictional force** between two surfaces to that pushing them together (the **normal contact force**). There are two values, the **coefficient of static friction** and the **coefficient of kinetic friction**.

Coefficient of friction $\mu = \dfrac{\textbf{frictional force (F)}}{\textbf{normal contact force (R)}}$

---

* **Electrons**, 83; **Neutrons**, **Nucleus**, **Protons**, 82.

ENERGY

Work is done when a force moves an object. **Energy** is the capacity to do work. When work is done on or by an object, it gains or loses energy respectively. Energy exists in many different forms and can change between them (energy **conversion** or **transformation**), but cannot be created or destroyed (**law of conservation of energy**). The **SI unit*** of energy and work is the **joule (J)**.

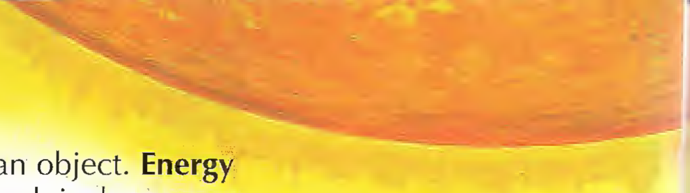

The energy from the Sun is the equivalent of that supplied by one million million million power stations.

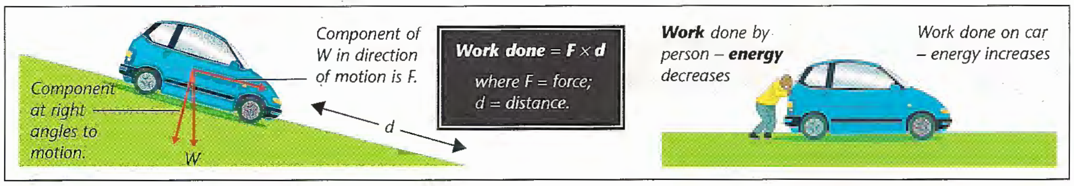
Component at right angles to motion.
W
Component of W in direction of motion is F.
d
Work done by person – energy decreases
Work done on car – energy increases

Work done = F × d

where F = force;
d = distance.

Potential energy (P.E.)

The energy of an object due to its position in a **force field***, which it has because work has been done to put it in that position. The energy has been "stored up". The three forms of potential energy are **gravitational potential energy**, **electromagnetic potential energy** and **nuclear potential energy** (depending on the force involved).

Gravitational potential energy

The **potential energy** associated with the position of an object relative to a mass which exerts a **gravitational force*** on it. If the object is moved further from the mass (e.g. an object being lifted on Earth), work is done on the body and its gravitational potential energy is raised.

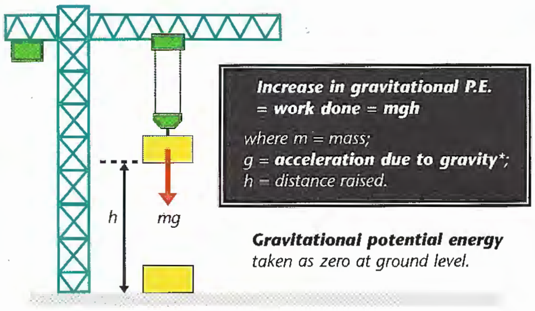

Increase in gravitational P.E. = work done = mgh

where m = mass;
g = **acceleration due to gravity***;
h = distance raised.

Gravitational potential energy *taken as zero at ground level.*

Nuclear potential energy

The **potential energy** stored in an atomic **nucleus***. Some nuclear potential energy is released during **radioactive decay***.

Electromagnetic potential energy

The **potential energy** associated with the position of a body in a **force field*** created by an **electromagnetic force***.

Molecular potential energy

The **electromagnetic potential energy** associated with the position of molecules relative to one another. It is increased when work is done against the **intermolecular force***.

Elastic potential energy or strain energy

An example of the **molecular potential energy**, stored as a result of stretching or compressing an object. It is the work done against the **intermolecular force***.

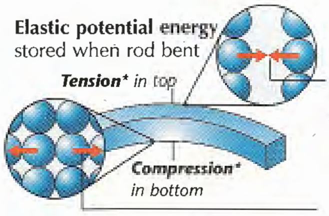

Elastic potential energy stored when rod bent

Attraction between particles (see ***intermolecular forces****, page 7).* ***Molecular potential energy*** *stored.*

Repulsion between particles (see ***intermolecular forces****, page 7).* ***Molecular potential energy*** *stored.*

Chemical energy

Energy stored in substances such as fuels, food, and chemicals in batteries. It is released during chemical reactions, e.g. as heat when a fuel burns, when the **electromagnetic potential energy** of the atoms and molecules changes.

Plants convert energy from sunlight into food – a store of ***chemical energy****.*

---

* **Acceleration due to gravity**, 18; **Compression**, 7; **Electromagnetic force**, **Force field**, **Gravitational force**, 6; **Intermolecular forces**, 7; **Nucleus**, 82; **Radioactive decay**, 87; **SI units**, 96; **Tension**, 7.

Kinetic energy (K.E.)

The energy associated with movement. It takes the form of **translational**, **rotational** and **vibrational energy**.

Kinetic energy of two objects linked by a spring

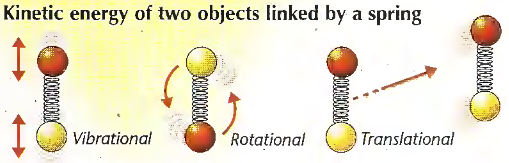
Vibrational
Rotational
Translational

$K.E. = \frac{1}{2} mv^2$ *where m = mass; v = velocity.*

Mechanical energy

The sum of the **kinetic energy** and **gravitational potential energy** of an object.

*The **mechanical energy** of a pendulum is constant (if resistive forces are neglected).*

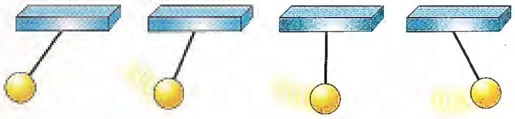

*All **gravitational potential energy*** | *Gravitational potential energy to **kinetic energy*** | *All kinetic energy (gravitational potential energy taken as zero here)* | *Kinetic energy to gravitational potential energy*

Internal or thermal energy

The sum of the **kinetic energy** and the **molecular potential energy** of the molecules in an object. If the temperature of an object increases, so does its internal energy.

Internal energy and temperature

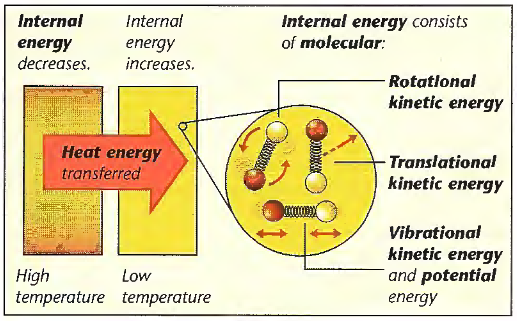
Internal energy decreases.
Internal energy increases.
Internal energy consists of molecular:
Rotational kinetic energy
Heat energy transferred
Translational kinetic energy
Vibrational kinetic energy and potential energy
High temperature
Low temperature

Heat energy or heat

The energy which flows from one place to another because of a difference in temperature (see pages 28-33). When **heat energy** is absorbed by an object, its **internal energy** increases (see diagram above).

Wave energy

The energy associated with wave action. For example, the energy of a water wave consists of the **gravitational potential energy** and **kinetic energy** of the water molecules.

Electric and magnetic energy

The types of energy associated with electric charge and moving electric charge (current). They are collectively referred to as **electromagnetic energy**.

Radiation

Any energy in the form of **electromagnetic waves*** or streams of particles. See also pages 29 and 86-87.

Power

The rate of doing work or the rate of change of energy. The **SI unit*** of power is the **watt** (**W**), which is equal to 1 joule per second.

Energy conversion in a power station

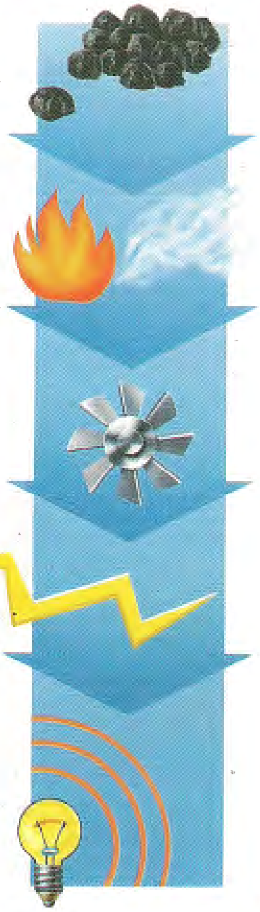

*Coal is the fossilized remains of plants that grew long ago (see also **fuel**, page 208). It is a store of **chemical energy** that came from the Sun.*

*Furnace in power station burns fuel and boils water. Here, **chemical energy** is converted to **internal energy** of steam.*

*Steam turns turbines. Internal energy of steam is converted to **rotational kinetic energy** of the turbine.*

*Generator converts kinetic energy to **electric energy**.*

*Appliances such as heaters, lamps and audio equipment convert electric energy into **heat energy**, light (**wave energy**) and sound (wave energy).*

---

* **Electromagnetic waves**, 44; **SI units**, 96.

MOTION

Motion is the change in position and orientation of an object. The motion of a **rigid** object (one which does not change shape) is made up of **translational motion**, or **translation**, i.e. movement of the **center of mass** from one place to another, and **rotational motion**, or **rotation**, i.e. movement around its center of mass. The study of the motion of points is called **kinematics**.

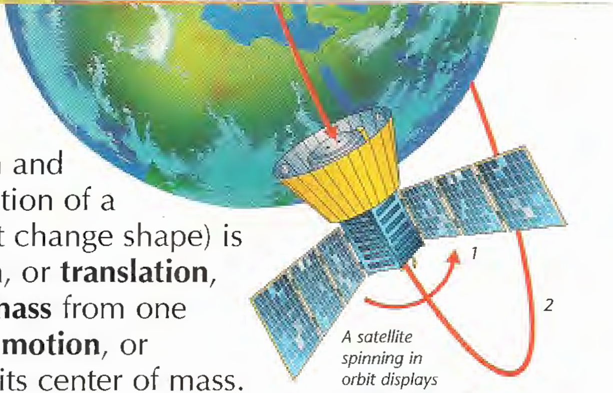

A satellite spinning in orbit displays ***rotational motion*** *(1) and* ***translational motion*** *(2).*

Linear motion

Linear or **rectilinear motion** is movement in a straight line and is the simplest form of **translational motion** (see introduction). The linear motion of any rigid object is described as the motion of its **center of mass**.

Center of mass

The point which acts as though the total mass of the object were at that point. The center of mass of a **rigid** object (see introduction) is in the same position as its **center of gravity** (the point through which the Earth's gravitational force acts on the object).

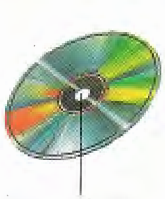

Center of mass *of uniform disk is at its center.*

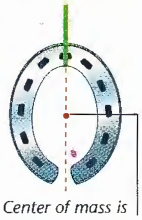

Center of mass is always under point of suspension.

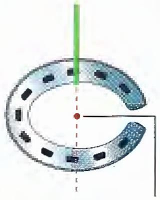

Center of mass is not always in object.

Displacement

The distance and direction of an object from a fixed reference point. It is a **vector quantity***. The position of an object can be expressed by its displacement from a specified point.

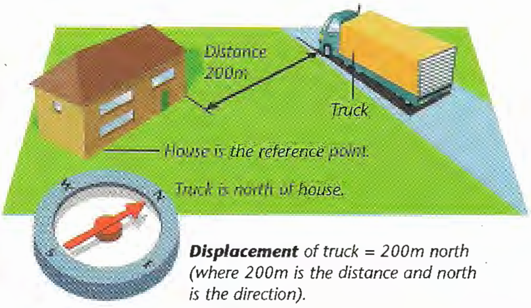
Distance 200m
Truck
House is the reference point.
Truck is north of house.

Displacement *of truck = 200m north (where 200m is the distance and north is the direction).*

Speed

The ratio of the distance traveled by an object to the time taken. If the speed of an object is constant, it is said to be moving with **uniform speed**. The **average speed** of an object over a time interval is the distance traveled by the object divided by the time interval. The **instantaneous speed** is the speed at any given moment.

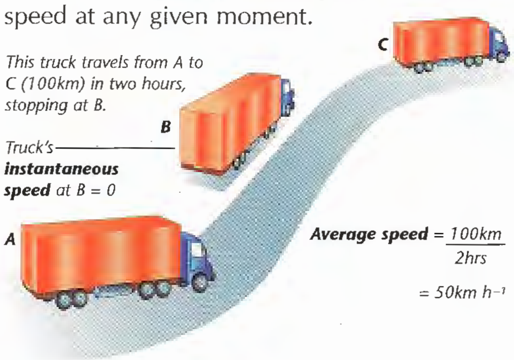

This truck travels from A to C (100km) in two hours, stopping at B.

Truck's ***instantaneous speed*** *at B = 0*

Average speed $= \frac{100\text{km}}{2\text{hrs}}$

$= 50\text{km h}^{-1}$

Velocity

The **speed** and direction of an object (i.e. its **displacement** in a given time). It is a **vector quantity***. **Uniform velocity**, **average velocity** and **instantaneous velocity** are all defined in a similar way to **uniform speed** etc. (see **speed**).

A displacement-time graph for an object which moves in a straight line from A to B and back to A (showing velocity calculation)

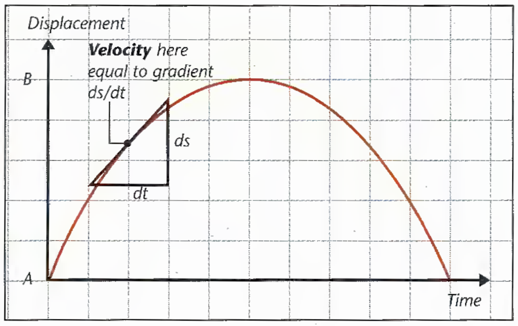

---

* **Vector quantity**, 108.

Relative velocity

The **velocity** which an object appears to have when seen by an observer who may be moving. This is known as the velocity of the object relative to the observer.

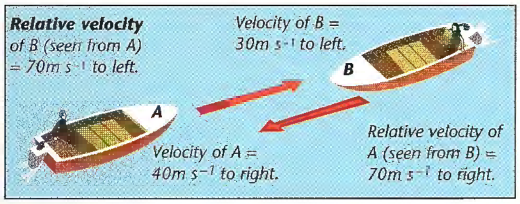
Relative velocity of B (seen from A) = 70m s⁻¹ to left.
Velocity of B = 30m s⁻¹ to left.
B
A
Velocity of A = 40m s⁻¹ to right.
Relative velocity of A (seen from B) = 70m s⁻¹ to right.

Acceleration

The ratio of the change in **velocity** of an object to the time taken. It is a **vector quantity***. An object accelerates if its **speed** changes (the usual case in **linear motion**) or its direction of travel changes (the usual case in **circular motion***). **Deceleration** in one direction is acceleration in the opposite direction (negative acceleration). An object whose velocity is changing the same amount in equal amounts of time is moving with **uniform acceleration**.

Graphs of velocity versus time showing acceleration

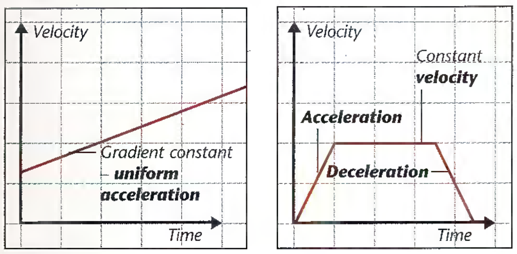

Distance traveled in equal time intervals increases.

Distance traveled in equal time intervals increases, remains constant, then decreases.

Rotational motion

The movement of an object about its **center of mass**. In rotational motion, each part of the object moves along a different path, so that the object cannot be considered as a whole in calculations. It must be split into small pieces and the **circular motion*** of each piece must be considered separately. From this, the overall motion of the object can be seen.

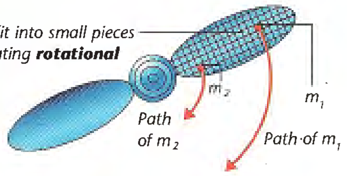
Object split into small pieces for calculating **rotational motion**
m_2
m_1
Path of m_2
Path of m_1

Equations of uniformly accelerated motion

Equations which are used in calculations involving **linear motion** with **uniform acceleration**. A **sign convention** must be used (see below). The equations use **displacement**, not distance, so changes of direction must be considered.

$$v = u + at$$

$$s = \tfrac{1}{2}(u + v)t$$

$$s = ut + \tfrac{1}{2}at^2$$

$$v^2 = u^2 + 2as$$

where t = time;
*u = initial **velocity** at time = 0;*
*v = final **velocity** after t;*
*s = **displacement** after t;*
*a = **acceleration** (constant).*

Sign convention

A method used to distinguish between motion in opposite directions. One direction is chosen as positive, and the other is then negative. The sign convention must be used when using the equations of motion (see above).

Sign convention

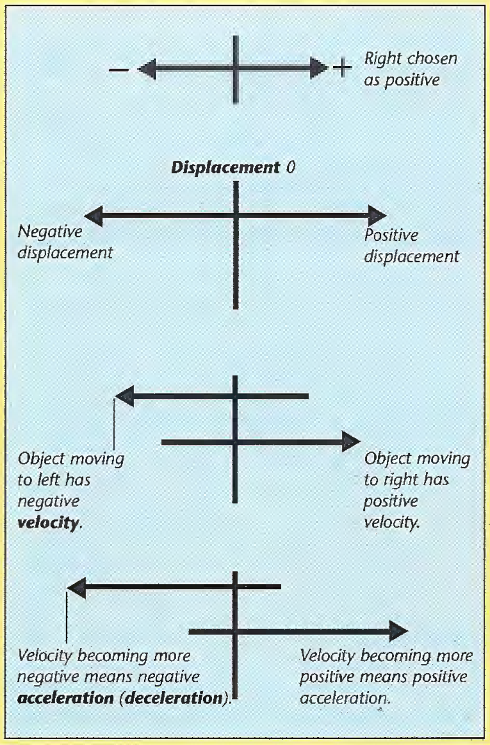
–
+
Right chosen as positive
Displacement 0
Negative displacement
Positive displacement
Object moving to left has negative **velocity**.
Object moving to right has positive velocity.
Velocity becoming more negative means negative **acceleration** (**deceleration**).
Velocity becoming more positive means positive acceleration.

* **Circular motion**, 17; **Vector quantity**, 108.

DYNAMICS

Dynamics is the study of the relationship between the motion of an object and the forces acting on it. A single force on an object causes it to change speed and/or direction (i.e. **accelerate***). If two or more forces act and there is no resultant force, the object does not accelerate, but may change shape.

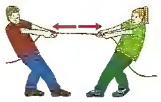

Two equal but opposite forces. No resultant force – no acceleration, but rope stretches.

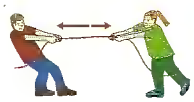

Forces not equal. Rope still stretches, but also accelerates to left due to resultant force.

Mass

A measurement of the **inertia** of an object. The force needed to accelerate an object by a given amount depends on its mass – a larger mass needs a larger force.

Momentum

The **mass** of an object multiplied by its **velocity***. Since velocity is a **vector quantity***, so is momentum. See also **law of conservation of linear momentum**.

$\text{Momentum} = mv$

where m = mass; v = velocity.

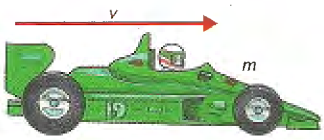

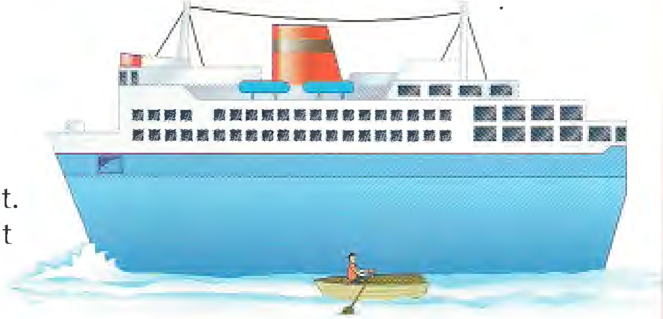

Inertia

The tendency of an object to resist a change of **velocity*** (i.e. to resist a force trying to **accelerate*** it). It is measured as **mass**.

*The large ship has much greater **inertia** (and therefore **mass**) than the little boat – a much larger force is needed to **accelerate*** it.*

Impulse

The force acting on an object multiplied by the time for which the force acts. From **Newton's second law**, impulse is equal to the change in **momentum** of an object. An equal change in momentum can be achieved by a small force for a long time or a large force for a short time.

$\text{Impulse} = Ft$

where F = force; t = time.

*Since force is rate of change of **momentum** (see **Newton's second law**) then:*

$\text{Impulse} = \text{change in momentum}$

*"Crumple zone" in the front of a car increases **collision** time – this makes force smaller.*

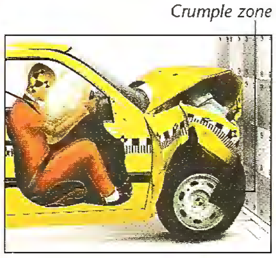

Newton's laws of motion

Three laws formulated by Newton in the late 1670s which relate force and motion.

Newton's first law

An object will remain at rest or in uniform motion unless acted upon by a force.

Forces on the object below are equal – no resultant force, so no acceleration.

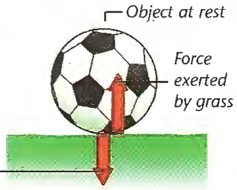

Newton's second law

If the **momentum** of an object changes, i.e. if it **accelerates***, then there must be a resultant force acting on it. Normally, the **mass** of the object is constant, and the force is thus proportional to the acceleration of the object. The direction of the acceleration is the same as the direction of the force.

$$\text{Force} = \frac{\text{change in momentum}}{\text{time}}$$

*If **mass** remains constant, then:*

$$\text{Force} = \text{mass} \times \text{acceleration}$$

---

* **Acceleration**, 11; **Vector quantity**, 108; **Velocity**, 10.

Collision

An occurrence which results in two or more objects exerting relatively large forces on each other over a relatively short time. This is not the everyday idea of a collision, because the objects do not necessarily have to be in contact.

Example of collision without contact

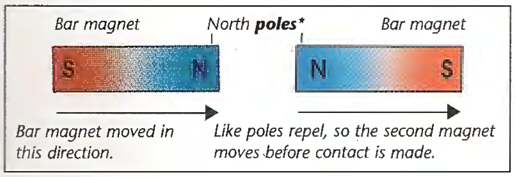
Bar magnet
North **poles***
Bar magnet
S
N
N
S
Bar magnet moved in this direction.
Like poles repel, so the second magnet moves before contact is made.

Law of conservation of linear momentum

If there is no external force on an object, then its linear **momentum** remains constant. If the system is considered just before and just after the **collision**, forces such as friction can be ignored.

Just before ***collision***

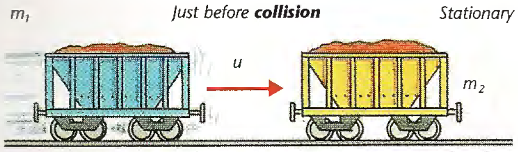

Total **momentum** = m_1u

Just after ***collision***

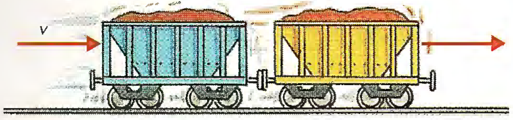

Total **momentum** = $(m_1 + m_2)v = m_1u$

Mass increases – **velocity*** decreases to conserve momentum.

Rocket engine

An engine which produces a high **velocity*** stream of gas through a nozzle by burning fuel held on board. The **mass** of gas is small, but its high velocity means it has a high **momentum**. The rocket gains an equal amount of momentum in the opposite direction (see **law of conservation of linear momentum**). Rocket engines are used in space because other engines require air.

Rocket engine

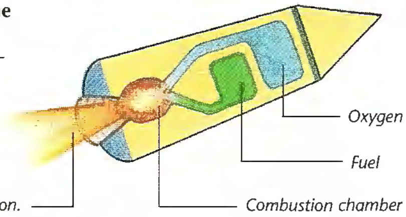

Stream of gas – ***momentum*** *is conserved so engine gains same amount of momentum as gas, but in opposite direction.*

Jet engine

An engine in which air is drawn in at the front to burn fuel, producing a high **velocity*** jet of gas. The principle is the same as that for the rocket engine, except that the gas is produced differently and the engine cannot be used in space because it requires air.

Jet engine

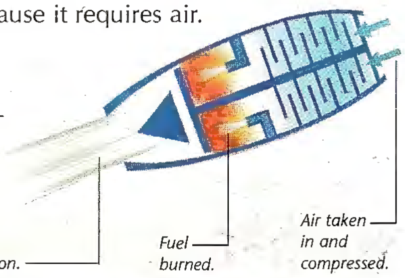

Stream of gas – ***momentum*** *is conserved so engine gains same amount of momentum as gas, but in opposite direction.*

Example of Newton's second law

A tennis ball hit by a racket undergoes a change of ***momentum***.

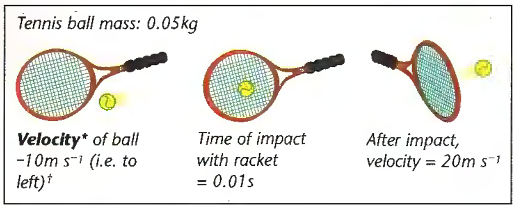

Tennis ball mass: 0.05kg

Velocity* *of ball* -10m s^{-1} *(i.e. to left)*†

Time of impact with racket = 0.01s

After impact, velocity = 20m s^{-1}

Resultant force found as follows:

$$\text{Force at impact} = \frac{\text{change in momentum}}{\text{time}} = \frac{(0.05 \times 20) - (0.05 \times -10)}{0.01} = 150\text{N}$$

Or:

$$\text{Force} = \text{mass} \times \text{acceleration} = \frac{\text{mass} \times \text{change in velocity}}{\text{time}}$$

$$= \frac{0.05 \times 30}{0.01} = 150\text{N}$$

Newton's third law

Forces always occur in equal and opposite pairs. Thus if object A exerts a force on object B, object B exerts an equal but opposite force on A. These forces do not cancel each other out, as they act on different objects.

Example of Newton's third law

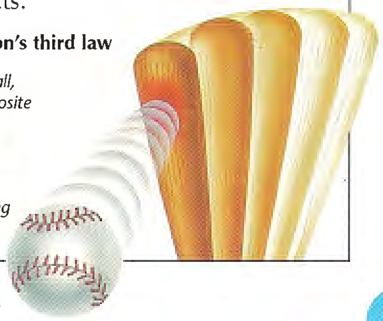

Bat exerts force on ball, accelerating it in opposite direction.

Ball exerts equal and opposite force on bat (felt as sudden slowing down of bat).

---

* **Pole**, 70; **Velocity**, 10.

† *Movement to right considered as positive (see* ***sign convention****, page 11).*

TURNING FORCES

A single force produces an **acceleration*** (see **dynamics**, page 12). In **linear motion***, it is a **linear acceleration.** In **rotational motion***, **angular acceleration*** (spinning faster or slower) is caused by a turning force or **moment** acting about the axis of rotation (the **fulcrum**).

Moment or torque

A measure of the ability of a force to rotate an object about an axis (the **fulcrum**). It is the size of the force multiplied by the perpendicular distance from the axis to the line along which the force acts (see diagram below). The **SI unit*** of moment is the **Newton meter** (Nm).

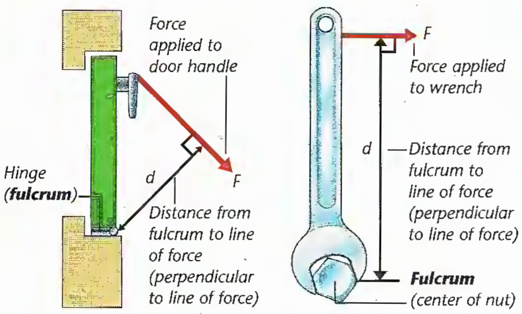
Force applied to door handle
Hinge (**fulcrum**)
d
F
Distance from fulcrum to line of force (perpendicular to line of force)
F
Force applied to wrench
d
Distance from fulcrum to line of force (perpendicular to line of force)
Fulcrum (center of nut)

In each case:

Moment = Fd

When considering moments, the axis about which they are taken must be stated and a **sign convention*** must be used to distinguish between clockwise and counterclockwise moments. The **resultant moment** is the single moment which has the same effect as all the individual moments acting together.

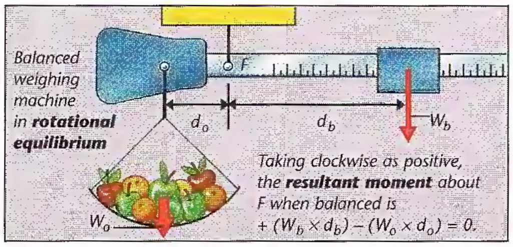

*Balanced weighing machine in **rotational equilibrium***

*Taking clockwise as positive, the **resultant moment** about F when balanced is* $+(W_b \times d_b) - (W_o \times d_o) = 0.$

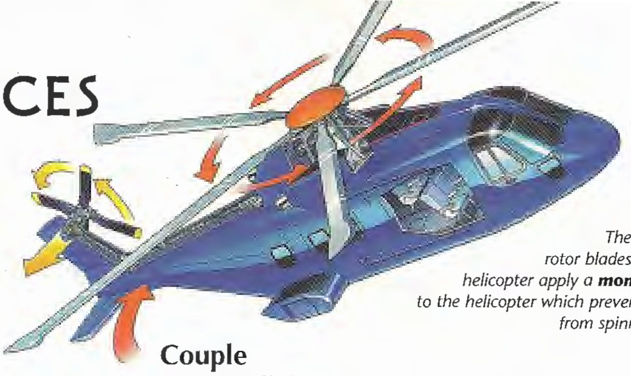

*The rear rotor blades of a helicopter apply a **moment** to the helicopter which prevents it from spinning.*

Couple

Two parallel forces which are equal and opposite but do not act along the same line. They produce a turning effect only, with no resultant acceleration of the **center of mass***. The **resultant moment** produced by a couple is the sum of the moments produced, and equals the perpendicular distance between the lines along which the forces act, multiplied by the size of one force.

*Equal and opposite forces (a **couple**) on a steering wheel cause it to turn.*

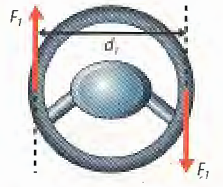

Moment of **couple** = $F_1 \times d_1$

In this case, one force applied to wheel by hand – the other by steering column.

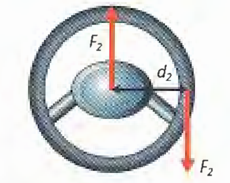

Moment of **couple** = $F_2 \times d_2$

Toppling

A condition which occurs if the vertical line through the **center of mass*** of an object does not pass through the base of the object. If this occurs, a **couple** of the weight and the **normal contact force*** rotates the object further over.

*Slightly tilted – **couple** will return it to level.*

***Toppling** – couple now acts to turn further over.*

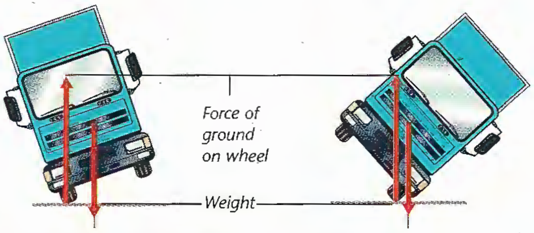
Force of ground on wheel
Weight

*Vertical line through **center of mass*** passes through base.*

Vertical line outside base

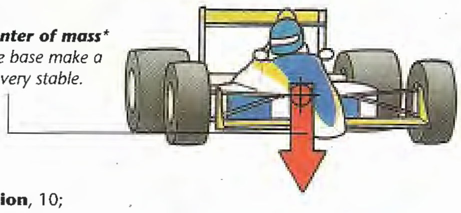

*A low **center of mass*** and wide base make a race car very stable.*

---

* **Acceleration**, 11; **Angular acceleration**, 17; **Center of mass**, **Linear motion**, 10; **Normal contact force**, 7 (**Coefficient of friction**); **Rotational motion**, **Sign convention**, 11; **SI units**, 96.

Equilibrium

When an object is not accelerating, it is said to be in **equilibrium**. It can be in **linear equilibrium** (i.e. the **center of mass*** is not accelerating) and/or **rotational equilibrium** (i.e. not accelerating about the center of mass). In addition, both cases of equilibrium are either **static** (not moving) or **dynamic** (moving).

Linear equilibrium

The state of an object when there is no acceleration of its **center of mass***, i.e. its speed and direction of motion do not change. The resultant force on the object when it is in linear equilibrium must be zero.

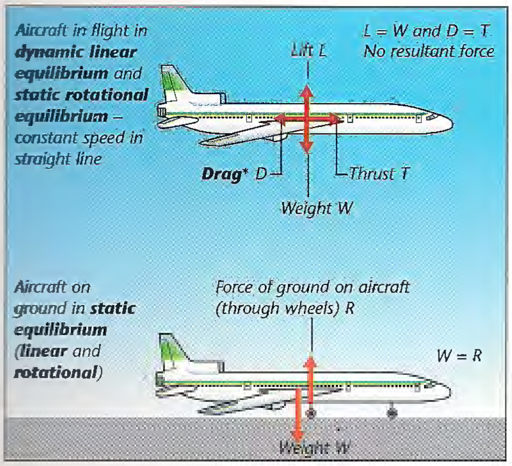
Aircraft in flight in **dynamic linear equilibrium** and **static rotational equilibrium** – constant speed in straight line
Lift L
L = W and D = T
No resultant force
Drag* D
Thrust T
Weight W
Aircraft on ground in **static equilibrium** (**linear** and **rotational**)
Force of ground on aircraft (through wheels) R
W = R
Weight W

Rotational equilibrium

The state of an object when there is no **angular acceleration***, i.e. it spins at constant **angular velocity***. If an object is in rotational equilibrium, the **resultant moment** (see **moment**) about any axis is zero.

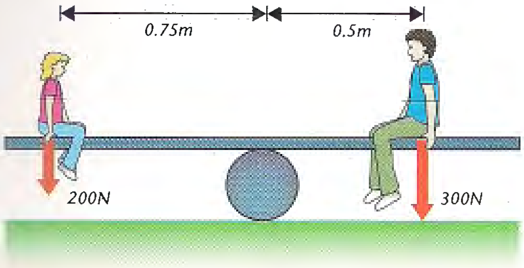
0.75m
0.5m
200N
300N

Beam in ***static rotational equilibrium****, since 200 × 0.75 = 300 × 0.5*

Stable equilibrium

A state in which an object moved a small distance from its equilibrium position returns to that position. This happens if the **center of mass*** is raised when the object is moved.

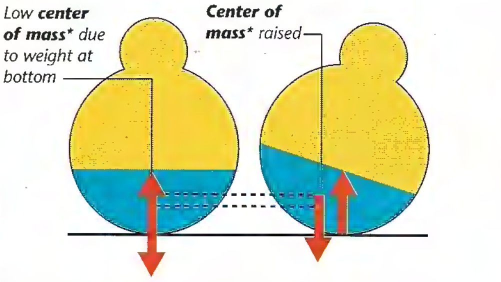
Low **center of mass*** due to weight at bottom
Center of mass* raised

Weight and force from ground form ***couple*** *to turn toy upright.*

Unstable equilibrium

A state in which an object moved a small distance from its equilibrium position moves further from that position. This happens if the **center of mass*** is lowered when the object is moved.

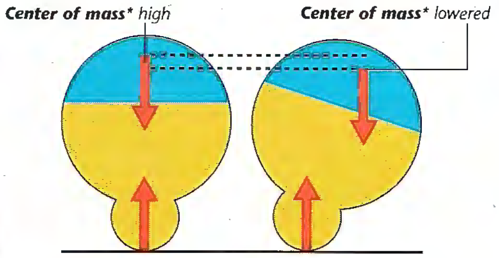
Center of mass* high
Center of mass* lowered

Weight and force form ***couple*** *which turns toy further over.*

Neutral equilibrium

A state in which an object moved a small distance from its equilibrium position remains in the new position. This happens if the **center of mass*** remains at the same height.

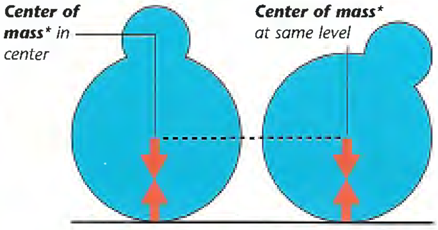
Center of mass* in center
Center of mass* at same level

Weight and force on same line – no ***couple*** *so toy stays in new position.*

---

* **Angular acceleration**, **Angular velocity**, 17;
Center of mass, 10; **Drag**, 19 (**Terminal velocity**).

PERIODIC MOTION

Periodic motion is any motion which repeats itself exactly at regular intervals. Examples of periodic motion are objects moving in a circle (**circular motion**), the swing of a pendulum and the vibration of molecules. **Wave motion*** consists of the periodic motion of particles or fields.

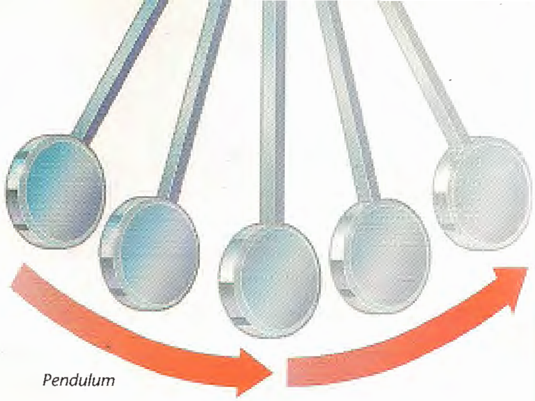

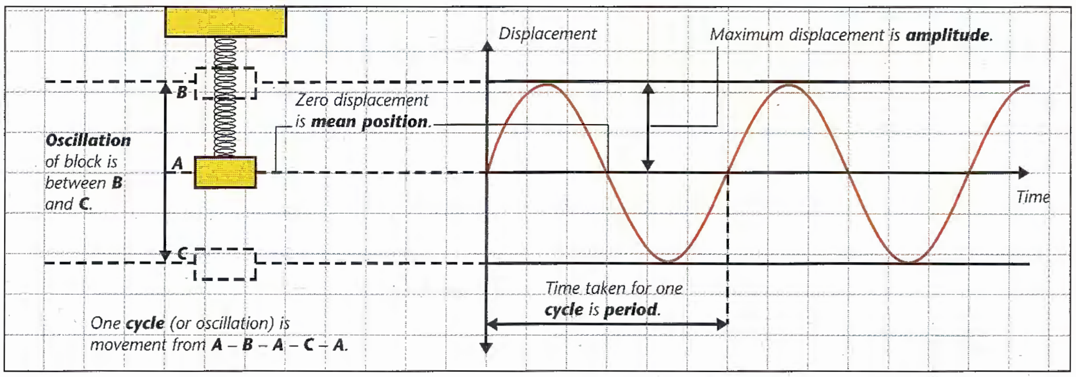
Oscillation of block is between B and C.
B
A
C
Zero displacement is mean position.
One cycle (or oscillation) is movement from A – B – A – C – A.
Displacement
Maximum displacement is amplitude.
Time
Time taken for one cycle is period.

Cycle

The movement between a point during a motion and the same point when the motion repeats. For example, one rotation of a spinning object.

Oscillation

Periodic motion between two extremes, e.g. a mass moving up and down on the end of a spring. In an oscillating system, there is a continuous change between **kinetic energy*** and **potential energy***. The total energy of a system (sum of its kinetic and potential energy) remains constant if there is no **damping**.

Period (T)

The time taken to complete one **cycle** of a motion, e.g. the period of rotation of the Earth about its axis is 24 hours.

Frequency (f)

The number of **cycles** of a particular motion in one second. The **SI unit*** of frequency is the **Hertz** (**Hz**), which is equal to one cycle per second.

$f = \frac{1}{T}$ where f = **frequency**; T = **period**.

Mean position

The position about which an object **oscillates**, and at which it comes to rest after oscillating, e.g. the mean position of a pendulum is when it is vertical. The position of zero displacement of an oscillating particle is usually taken as this point.

Amplitude

The maximum displacement of an **oscillating particle** from its **mean position**.

Damping

The process whereby **oscillations** die down due to a loss of energy, e.g. shock absorbers in cars cause oscillations to die down after a car has gone over a bump in the road.

Damping in an oscillating system

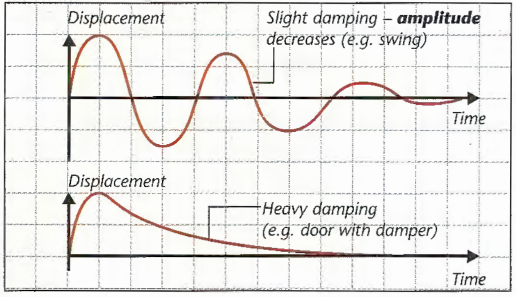

***Kinetic energy**, 9; **Potential energy**, 8; **SI units**, 96; **Wave motion**, 34.

Natural or **free oscillation**

The **oscillation** of a system when left after being started. The **period** and **frequency** of the system are called the **natural period** and **natural frequency** (these remain the same as long as the **damping** is not too great).

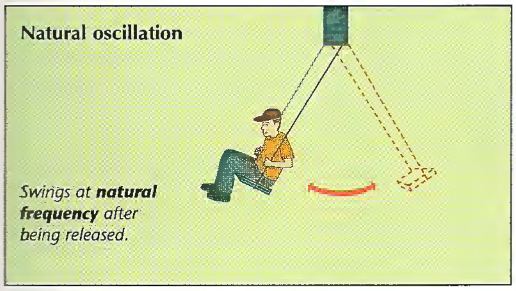
Natural oscillation
*Swings at **natural frequency** after being released.*

Forced oscillation

The **oscillation** of a system when given a repeated driving force (a force applied to the system) at regular intervals. The system is made to oscillate at the **frequency** of the driving force, irrespective of its **natural frequency**.

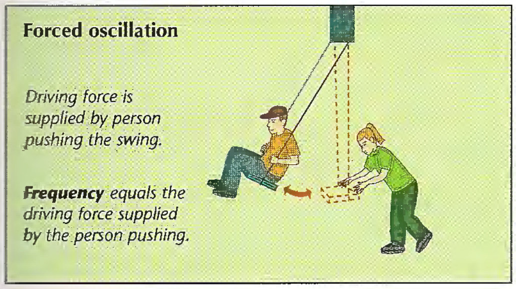
Forced oscillation
Driving force is supplied by person pushing the swing.
***Frequency** equals the driving force supplied by the person pushing.*

Resonance

The effect exhibited by a system in which the **frequency** of the driving force (a force applied to the system) is about the same as the **natural frequency** of the system. The system then has a large **amplitude**.

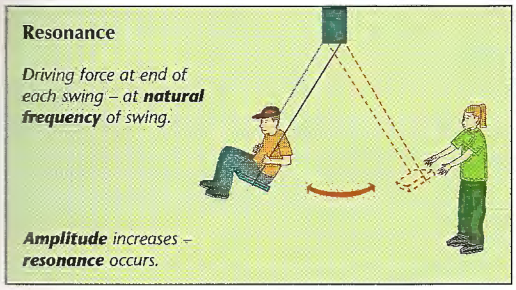
Resonance
*Driving force at end of each swing – at **natural frequency** of swing.*
***Amplitude** increases – **resonance** occurs.*

Circular motion

Uniform circular motion

The motion of an object in a circle at constant speed. Since the direction (and therefore the **velocity***) changes, the object is constantly **accelerating*** toward the center (**centripetal acceleration**), and so there is a force acting toward the center. Circular motion can be considered in terms of **angular velocity**.

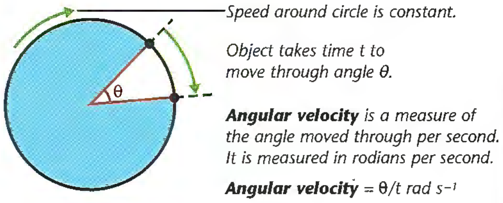
Speed around circle is constant.
θ
Object takes time t to move through angle θ.
***Angular velocity** is a measure of the angle moved through per second. It is measured in radians per second.*
***Angular velocity** = θ/t rad s⁻¹*

Centripetal acceleration (a)

The **acceleration*** of an object in circular motion (see above) acting toward the center of the circle.

Centripetal force

The force which acts on an object toward the center of a circle to produce **centripetal acceleration**, and so keeps the object moving in a circle.

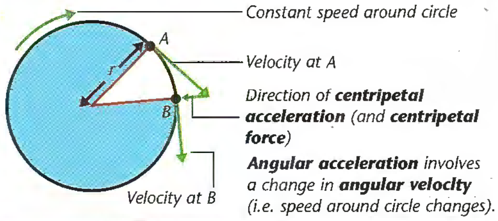
Constant speed around circle
A
r
B
Velocity at A
*Direction of **centripetal acceleration** (and **centripetal force**)*
***Angular acceleration** involves a change in **angular velocity** (i.e. speed around circle changes).*
Velocity at B

$$a = \frac{v^2}{r}$$

*where a = **centripetal acceleration**;*
v = speed around circle;
r = radius of circle.

Centripetal force has an equal and opposite reaction (see **Newton's second law**, page 12), called **centrifugal force**. It does not act on the object moving in the circle so is not considered in forces calculations.

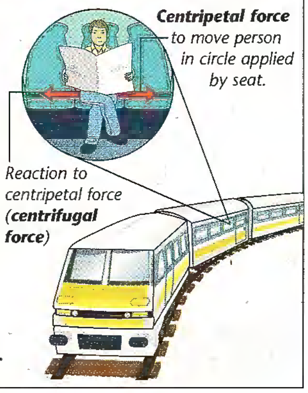
***Centripetal force** to move person in circle applied by seat.*
*Reaction to centripetal force (**centrifugal force**)*

---

* **Acceleration**, 11; **Velocity**, 10.

GRAVITATION

Gravitation is the effect of the **gravitational force*** of attraction (see also page 104) which acts between all objects in the universe. It is noticed with massive objects like the planets, which remain in orbit because of it. The gravitational force between an object and a planet, which pulls the object downward, is called the **weight** of the object.

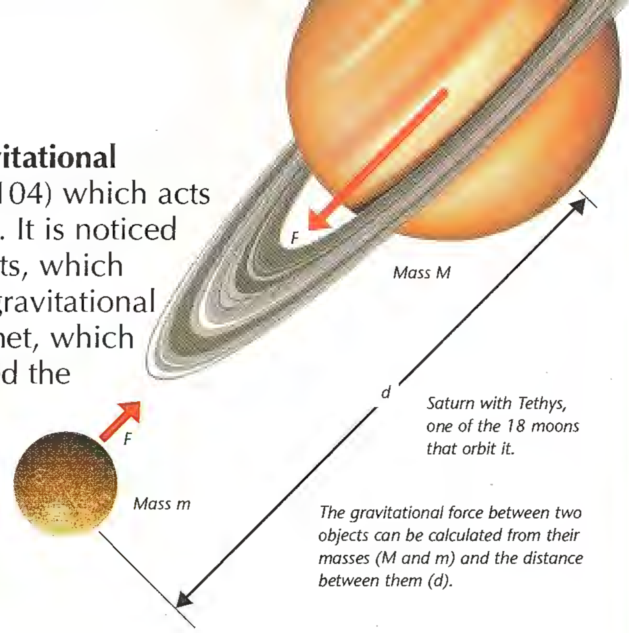

Saturn with Tethys, one of the 18 moons that orbit it.

The gravitational force between two objects can be calculated from their masses (M and m) and the distance between them (d).

Newton's law of gravitation

States that there is a gravitational force of attraction between any two objects with mass which depends on the masses of the objects and the distance between them. The **gravitational constant** (**G**) has a value of 6.7 × 10^{-11}Nm^2 kg^{-2}, and its small value means that gravitational forces are negligible unless one of the masses is very large.

$$F = G\frac{Mm}{d^2}$$ where G = **gravitational constant**.

Weight

The gravitational pull of a massive object (e.g. a planet) on another object. The weight of an object is not constant, but depends on the distance from, and mass of the planet. Hence, although the mass of an object is independent of its position, its weight is not.

Weight of mass of 100kg alters with position:

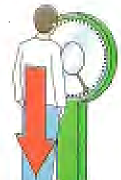

Weighing scales measure the force exerted on them. At the surface of the Earth, the weight of mass 100kg is 980N.

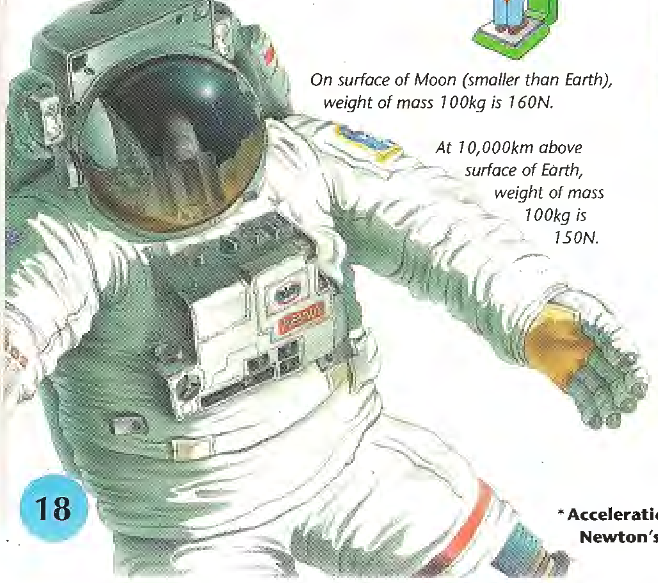

On surface of Moon (smaller than Earth), weight of mass 100kg is 160N.

At 10,000km above surface of Earth, weight of mass 100kg is 150N.

Acceleration due to gravity (g)

The **acceleration*** produced by the gravitational force of attraction. Its value is the same for any mass at a given place. It is about 9.8m s^{-2} on the Earth's surface, and decreases above the surface according to **Newton's law of gravitation**. The value of 9.8m s^{-2} is used as a unit of acceleration (the **g-force**).

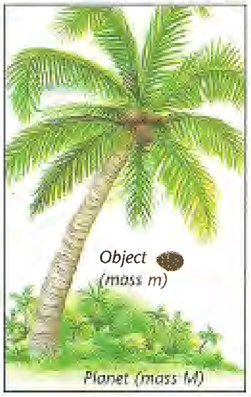

*From **Newton's law of gravitation** and **Newton's second law***:*

$$\text{Force on mass } (= mg) = G\frac{Mm}{d^2}$$

*So **acceleration due to gravity (g)**:*

$$g = G\frac{M}{d^2}$$

*During tight turns, pilots experience high **g-forces** (e.g. 5g – five times the normal) which can lead to blackout.*

***Acceleration**, 11; **Gravitational force**, 6; **Newton's second law**, 12.

Terminal velocity

The maximum, constant **velocity*** reached by an object falling through a gas or liquid. As the velocity increases, the resistance due to the air or liquid (**drag**) increases. Eventually, the drag becomes equal to the **weight** of the object, and its velocity does not increase any more.

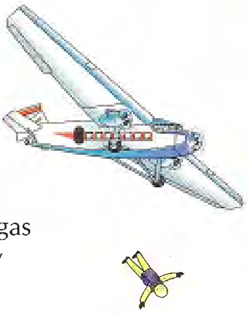

*Just after the sky diver jumps, **velocity*** = 0, so **drag** = 0 and **acceleration*** = g.*

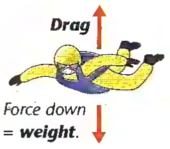
Drag
Force down = **weight**.

Velocity increases, drag increases, acceleration less than g.

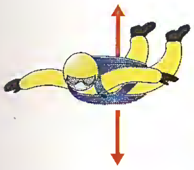

*At **terminal velocity**, drag same as weight, acceleration = 0.*

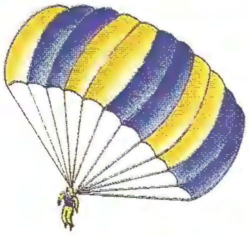

Parachute open, drag much greater, terminal velocity less.

Escape velocity

The minimum **velocity*** at which an object must travel in order to escape the gravitational pull of a planet without further propulsion. It is about 40,000km h^{-1} on Earth.

Free fall

The unrestricted motion of an object when it is acted upon only by the gravitational force (i.e. when there are no resistive or other forces acting, e.g. air resistance).

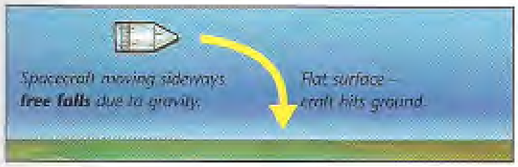
Spacecraft moving sideways **free falls** due to gravity.
Flat surface – craft hits ground.

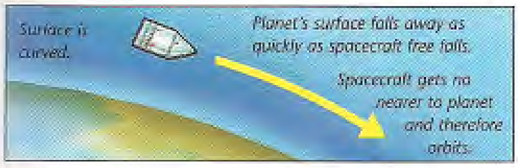
Surface is curved.
Planet's surface falls away as quickly as spacecraft free falls.
Spacecraft gets no nearer to planet and therefore orbits.

Weightlessness

The state in which an object does not exert any force on its surroundings.

True weightlessness

Weightlessness due to an object being in a gravity-free region.

Apparent weightlessness

The state of an object when it is as if there were no gravitational forces acting. This occurs if two objects **accelerate*** independently in the same way.

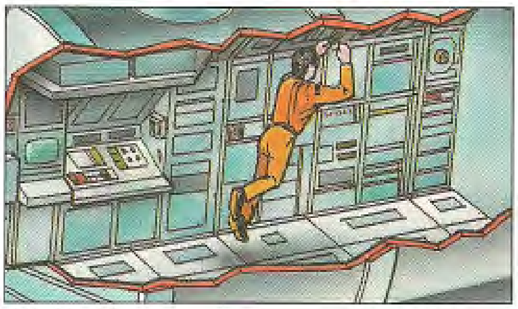

*Astronaut inside an orbiting spacecraft **free falls** in same way as spacecraft and is therefore **apparently weightless** inside it.*

Geo-stationary or parking orbit

The path of a satellite which orbits the Earth in the same direction as the rotation of the Earth so that it stays above the same place on the surface all the time. The satellite has a **period*** of 24 hours.

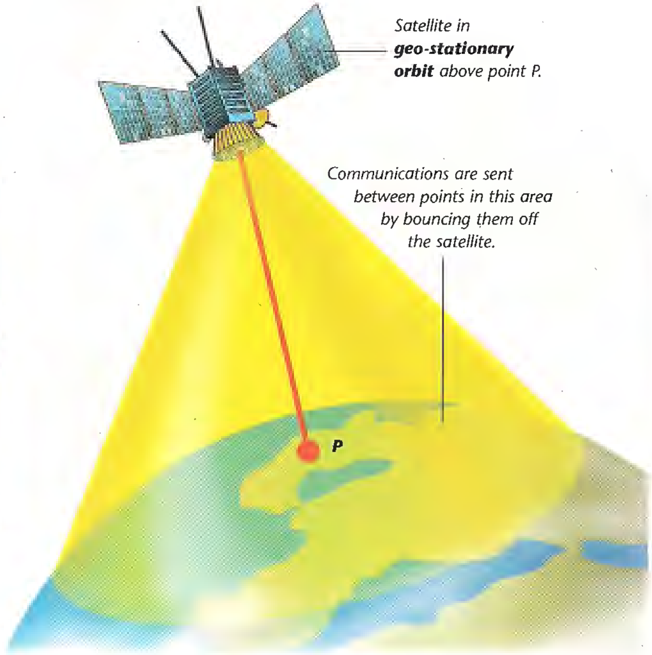
Satellite in **geo-stationary orbit** above point P.
Communications are sent between points in this area by bouncing them off the satellite.
P

---

* **Acceleration**, 11; **Period**, 16; **Velocity**, 10.

MACHINES

A **machine** is a device which is used to overcome a force called the **load**. This force is applied at one point and the machine works by the application of another force called the **effort** at a different point. For example, a small effort exerted on the rope of a **pulley** overcomes the weight of the object being raised by the pulley.

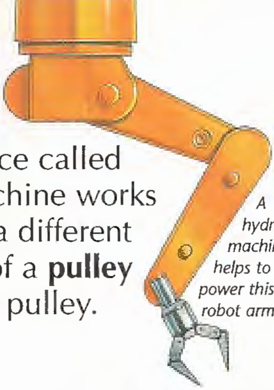

A hydraulic machine helps to power this robot arm.

Pulley system – an example of a machine

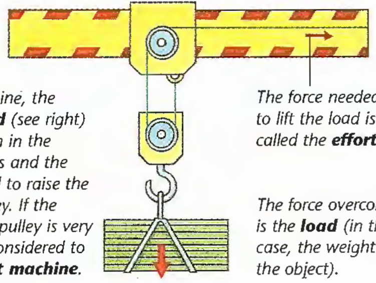

(See also page 21.)

*In this machine, the **useless load** (see right) is the friction in the pulley wheels and the force needed to raise the bottom pulley. If the mass of the pulley is very small, it is considered to be a **perfect machine**.*

*The force needed to lift the load is called the **effort**.*

*The force overcome is the **load** (in this case, the weight of the object).*

Useless load

The force needed to overcome the **frictional forces*** between the moving parts of a machine and to raise any of its moving parts.

Perfect machine

A theoretical machine, with a **useless load** of zero. Machines in which the useless load is negligible compared to the load can be considered as perfect machines.

Mechanical advantage (M.A.)

The load (L) divided by the effort (E). A mechanical advantage greater than one means that the load overcome is greater than the effort. The mechanical advantage of any given **perfect machine** remains the same as the load increases. The mechanical advantage of any given real machine increases slightly with load because **useless load** becomes less significant as the load increases.

$$\text{M.A.} = \frac{L}{E}$$

Diagram of screwjack showing effort, load and work done

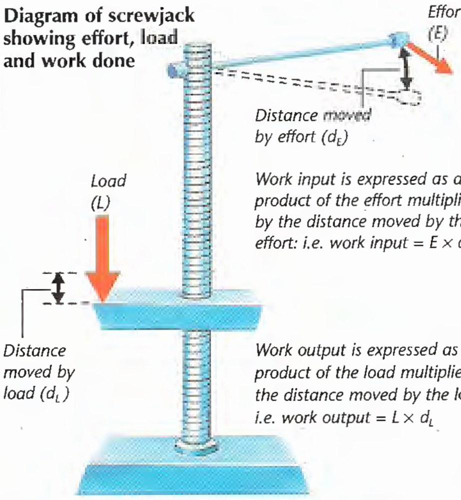

Distance moved by effort (d_E)

Work input is expressed as a product of the effort multiplied by the distance moved by the effort: i.e. work input = $E \times d_E$

Distance moved by load (d_L)

Work output is expressed as a product of the load multiplied by the distance moved by the load: i.e. work output = $L \times d_L$

Efficiency

The **work** done (force × distance – see page 8) on the load (work output) divided by the work done by the effort (work input), expressed as a percentage. All real machines have an efficiency of less than 100% due to **useless load**. **Perfect machines** are 100% efficient.

$$\text{Efficiency} = \frac{\text{work out}}{\text{work in}} \times 100$$

$$= \frac{L \times d_L}{E \times d_E} \times 100$$

$$= \text{M.A.} \times \frac{1}{\text{V.R.}} \times 100$$

$$\text{Efficiency} = \frac{\text{M.A.}}{\text{V.R.}} \times 100$$

*In a **perfect machine** (100% efficiency):* **M.A. = V.R.**

---

* **Frictional force**, 7.

Examples of machines

Hydraulic press

A large and small cylinder connected by a pipe and filled with fluid, used to produce large forces.

Hydraulic press

Volume of liquid moved $= a \times d_E = A \times d_L$ *so* **V.R.**$(d_E / d_L) = A / a$

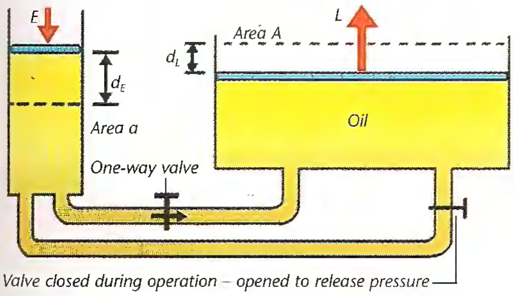

Valve closed during operation – opened to release pressure

Lever

Any rigid object which is pivoted about an axis called the **fulcrum** (**F**). The load and effort can be applied on either or the same side. There are three classes of lever, shown below.

Class 1

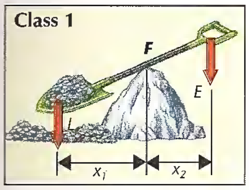

Fulcrum between effort and load.

Class 2

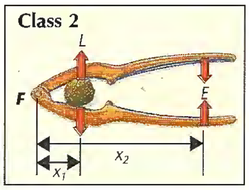

Load between effort and fulcrum.

Class 3

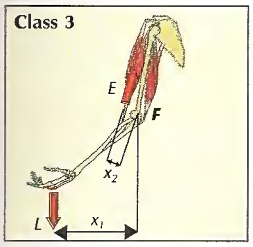

Effort between fulcrum and load.

For **equilibrium***:

$L \times x_1 = E \times x_2$

so $M.A. = \frac{L}{E} = \frac{x_2}{x_1}$

Thus $V.R. = \frac{x_2}{x_1}$

Velocity ratio** is calculated by considering **moments and assuming **M.A.** = **V.R.** (see **efficiency**).*

Gear

A combination of toothed wheels used to transmit motion between rotating shafts.

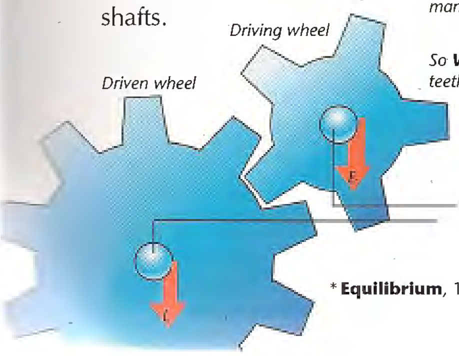

Twice as many teeth on the driven wheel means that the driving wheel must rotate twice as many times.

*So **V.R.** = number of teeth on driven wheel divided by number of teeth on driving wheel.*

Shafts of same diameter

Pulley system

A wheel (or combination of wheels) and a rope, belt or chain which transmits motion.

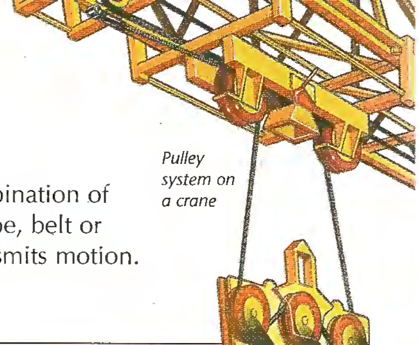

Pulley system on a crane

Pulley systems

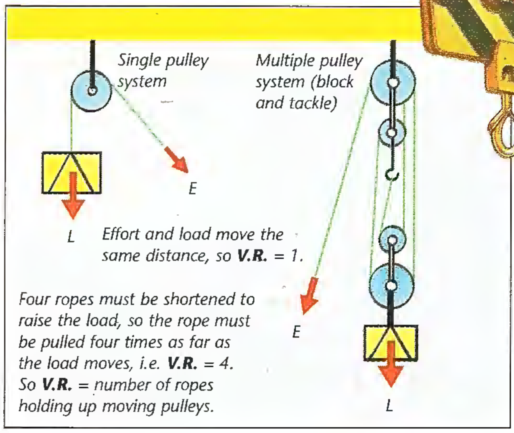

*Effort and load move the same distance, so **V.R.** = 1.*

*Four ropes must be shortened to raise the load, so the rope must be pulled four times as far as the load moves, i.e. **V.R.** = 4. So **V.R.** = number of ropes holding up moving pulleys.*

Inclined plane

A plane surface at an angle to the horizontal. It is easier to move an object up an inclined plane than to move it vertically upwards.

Inclined plane

$$V.R. = \frac{l}{h}$$

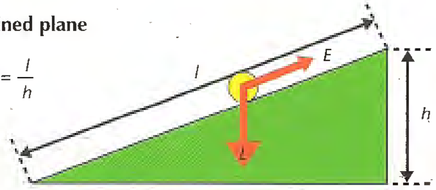

Screw jack

A system in which a screw thread is turned to raise a load (e.g. a car jack). The **pitch** is the distance between each thread on the screw.

Screw jack

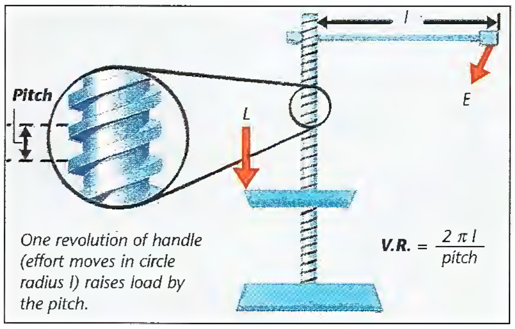

One revolution of handle (effort moves in circle radius l) raises load by the pitch.

$$V.R. = \frac{2 \pi l}{\text{pitch}}$$

---

* **Equilibrium**, 15; **Moment**, 14.

MOLECULAR PROPERTIES

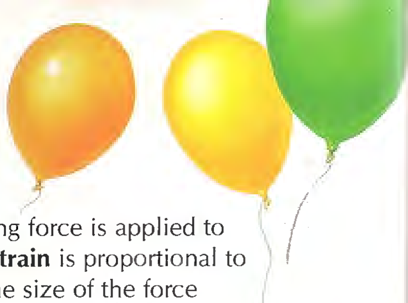

There are a number of properties of matter which can be explained in terms of the behavior of molecules, in particular their behavior due to the action of the forces between them (**intermolecular forces***). Among these properties, and explained on this double page, are **elasticity**, **surface tension** and **viscosity**. See also pages 4-5 and 24-25.

Elasticity

The ability of a material to return to its original shape and size after distorting forces (i.e. **tension*** or **compression***) have been removed. Materials which have this ability are **elastic**; those which do not are **plastic**.

*Cool wax is **plastic** (the seal leaves a permanent impression in the wax).*

Elasticity is a result of **intermolecular forces*** – if an object is stretched or compressed, its molecules move further apart or closer together respectively. This results in a force of attraction (in the first case) or repulsion (in the second), so the molecules return to their average separation when the distorting force is removed. This always happens while the size of the force is below a certain level (different for each material), but all elastic materials finally become plastic if the force exceeds this level (see **elastic limit** and **yield point**).

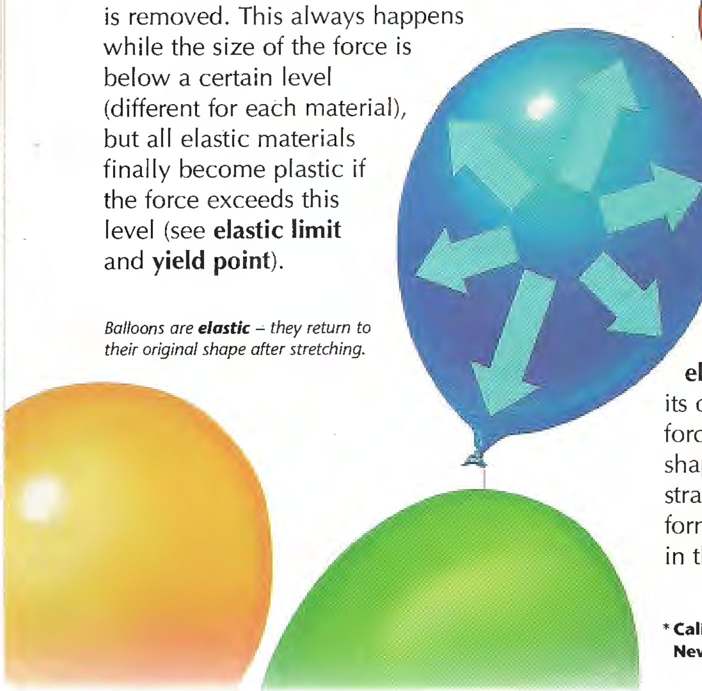

*Balloons are **elastic** – they return to their original shape after stretching.*

Hooke's law

States that,
when a distorting force is applied to an object, the **strain** is proportional to the **stress**. As the size of the force increases, though, the **limit of proportionality** (or **proportional limit**) is reached, after which Hooke's law is no longer true (see graph, page 23).

Strain and stress in a stretched wire

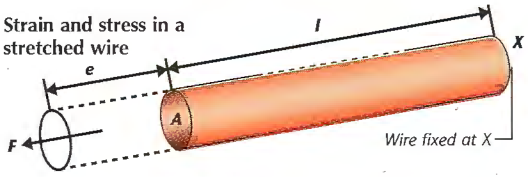

***Strain** is stated as change in length per unit length.*

$$\textbf{Strain} = \frac{e}{l}$$ *where e = change in length; l = original length.*

***Stress** is stated as force applied per unit area.*

$$\textbf{Stress} = \frac{F}{A}$$ *where F = force applied; A = cross-sectional area.*

***Spring balance** uses **Hooke's law** to measure force. Spring is extended in proportion to force applied.*

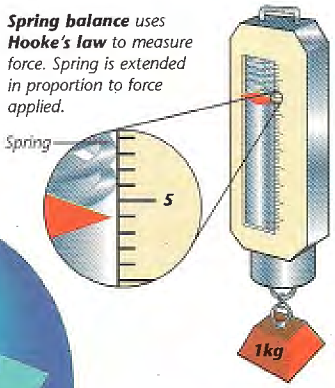

*Scale **calibrated*** so that length of spring gives size of force in **newtons***.*

*For an object in **tension*** or **compression**, **stress** divided by **strain** (see above) is always same figure for a given material (**Young's modulus** – see page 112) until **limit of proportionality** is reached.*

Elastic limit

The point, just after the **limit of proportionality** (see **Hooke's law**), beyond which an object ceases to be **elastic**, in the sense that it does not return to its original shape and size when the distorting force is removed. It does return to a similar shape and size, but has suffered a permanent strain (it will continue to return to this new form if forces are applied, i.e. it stays elastic in this sense).

---

* **Calibration**, 344; **Compression**, **Intermolecular forces**, 7; **Newton**, 6; **Tension**, 7.

Yield point

The point, just after the **elastic limit**, at which a distorting force causes a major change in a material. In a **ductile*** material, the internal structure changes – bonds between molecular layers break and the layers flow over each other. This change is called **plastic deformation** (the material becomes **plastic**). It continues as the force increases, and the material will eventually break. A **brittle** material, by contrast, will break at its yield point. The **yield stress** of a material is the value of the **stress** at its yield point. See graph below.

Stress/strain graph for a ductile* material

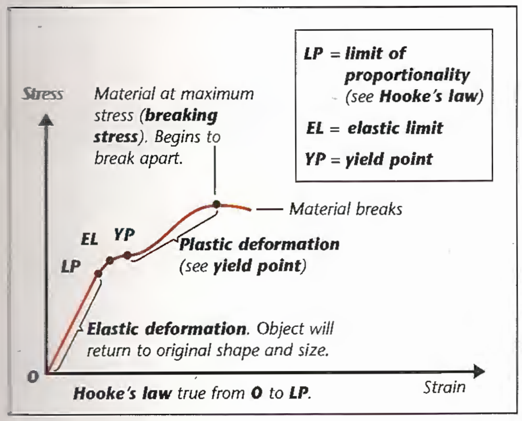

Viscosity

The ease of flow of a fluid. It depends on the size of the **frictional force*** between different layers of molecules as they slide over each other.

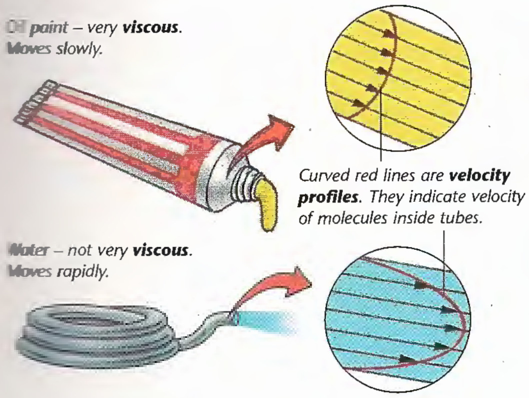

*Oil paint – very **viscous**. Moves slowly.*

*Curved red lines are **velocity profiles**. They indicate velocity of molecules inside tubes.*

*Water – not very **viscous**. Moves rapidly.*

*Outer layers of fluid slowed by **frictional force*** between them and sides of container. Working inwards, friction gradually loses effect. Effect lost much faster in water than oil paint.*

Surface tension

The skin-like property of a liquid surface resulting from **intermolecular forces*** which cause it to contract to the smallest possible area.

Two examples of surface tension

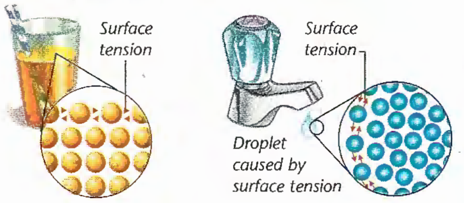
Surface tension
Surface tension
Droplet caused by surface tension

At the surface, the molecules are slightly further apart than the inner ones, and at a separation such that they attract each other (see **intermolecular forces**, page 7). The molecules cannot move closer because of equal forces on either side. The surface molecules are therefore in a constant state of tension, giving the surface elastic properties.

Water bug on water's surface

Adhesion

An **intermolecular force*** of attraction between molecules of different substances.

***Capillary action** or **capillarity** is a result of **adhesion** or **cohesion**.*

Upward **capillary action**

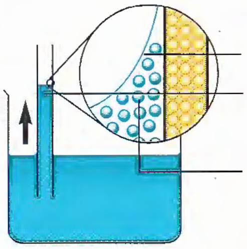

*Concave **meniscus***

*Water moves up fine bore glass tube (**capillary tube**).*

*Water molecules attracted to glass molecules because forces of **adhesion** are stronger than attraction between water molecules.*

Cohesion

An **intermolecular force*** of attraction between molecules of the same substance.

Downward **capillary action**

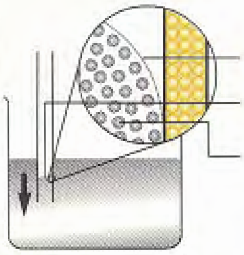

*Convex **meniscus***

Mercury moves down capillary tube.

*Mercury molecules attracted to each other because forces of **cohesion** are stronger than attraction between mercury and glass molecules.*

---

* **Ductile**, 344; **Frictional force**, **Intermolecular forces**, 7; **Meniscus**, 345.

DENSITY AND PRESSURE

The **density** (ρ) of an object depends on both the mass of its molecules and its volume (see formula, right). For example, if one substance has a higher density than another, then the same volumes of the substances have different masses (the first mass being greater than the second). Similarly, the same masses have different volumes.

Relative density or **specific gravity**

The density of a substance relative to the density of water (which is 1,000kg m^{-3}). It indicates how much more or less dense than water a substance is, so the figures need no units, e.g. 1.5 (one and a half times as dense). It is found by dividing the mass of any volume of a substance by the mass of an equal volume of water.

Eureka can

A can used to measure the volume of a solid object with an irregular shape, in order to calculate its density. The volume of water displaced is equal to the volume of the object. The density of the object is its mass divided by this volume.

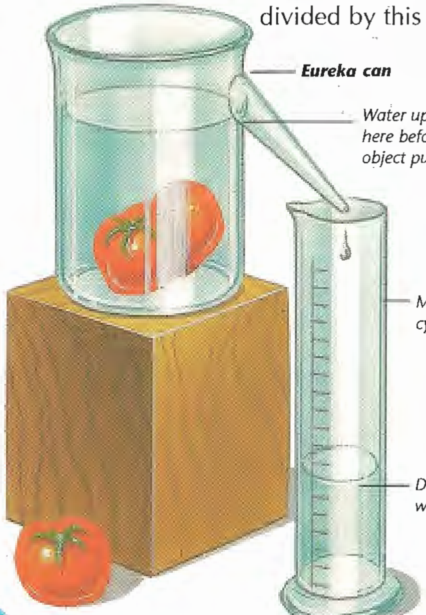

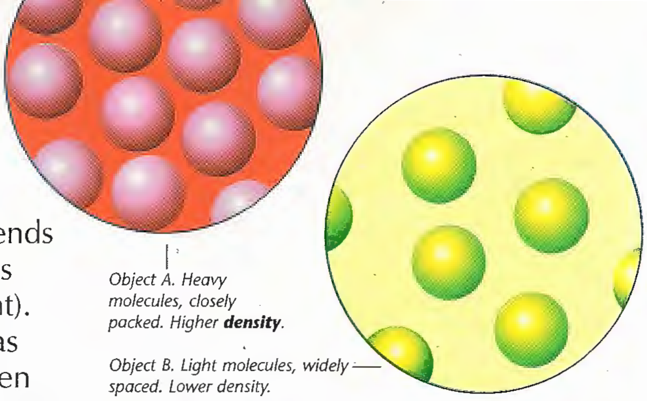

*Object A. Heavy molecules, closely packed. Higher **density**.*

Object B. Light molecules, widely spaced. Lower density.

$$\rho = \frac{m}{v} \quad \text{where } \rho = \text{density; } m = \text{mass; } v = \text{volume.}$$

Hence $m = \rho \times v$

v is the same for objects A and B, so object A has greater mass.

*The **SI unit*** of density is the kg m^{-3}.*

Density bottle

A container which, when completely full, holds a precisely measured volume of liquid (at constant temperature). It is used to measure the density of liquids (by measuring the mass of the bottle and liquid, subtracting the mass of the bottle and dividing by the volume of liquid).

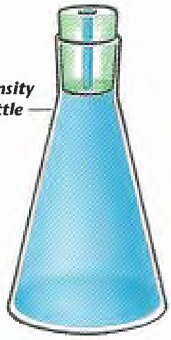

*Fine bore tube (**capillary tube**) in glass stopper. Bottle filled, stopper inserted, excess liquid rises through tube and runs out – ensures same volume each time.*

Hydrometer or **aerometer**

An instrument which measures the density of a liquid by the level at which it floats in that liquid. If the liquid is very dense, the hydrometer floats near the surface, as only a small volume of liquid needs to be displaced to equal the weight of the hydrometer.

Hydrometer

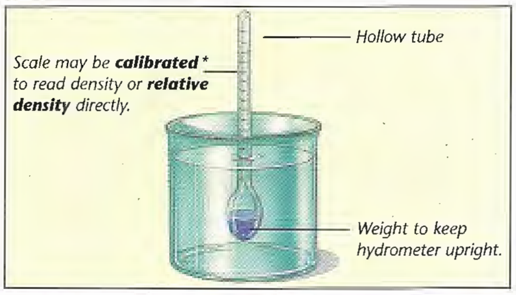

*Scale may be **calibrated*** to read density or **relative density** directly.*

Weight to keep hydrometer upright.

---

* **Calibration**, 344; **SI units**, 96.

Pressure

Pressure is the force, acting at right angles, exerted by a solid, liquid or gas on a unit area of a substance (solid, liquid or gas).

Pressure in a vessel of water

The greater the force on a fixed area, the greater the pressure.

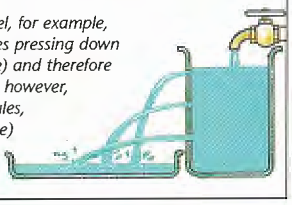

At the top of this water vessel, for example, there are few water molecules pressing down so there is little weight (force) and therefore little pressure. Further down, however, there are more water molecules, so there is more weight (force) and therefore greater pressure.

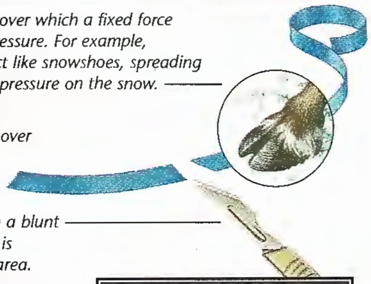

The greater the area over which a fixed force acts, the lower the pressure. For example, caribous' wide feet act like snowshoes, spreading the weight to reduce pressure on the snow.

The smaller the area over which a fixed force acts, the higher the pressure. A sharp knife cuts better than a blunt one because its force is applied to a smaller area.

The ***SI unit**** *of pressure is the* ***pascal*** *(Pa).*

$$\boxed{\textbf{Pressure} = \frac{\textbf{force}}{\textbf{area}}}$$

Barometer

An instrument used to measure **atmospheric pressure** – the pressure caused by the weight of air molecules above the Earth. There are several common types.

A ***barometer*** *for the home gives* ***pressure*** *readings and brief weather descriptions.*

Simple barometer

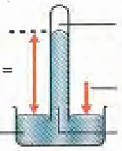
Torricellian vacuum *(no pressure acting down)*
Atmospheric pressure *= 760mm of mercury*
Atmospheric pressure
Mercury
Fixed diameter tube

Manometer

A U-shaped tube containing a liquid. It is used to measure difference in fluid pressures.

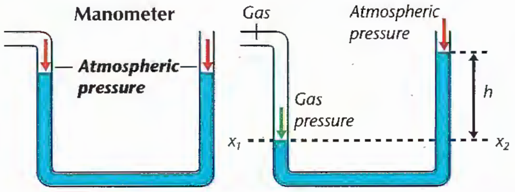

Pressures at x_1 *and* x_2 *(same level) must be the same.*
So pressure of gas = pressure at x_2
= atmospheric pressure + pressure of height (h) of liquid.

Objects in fluids

An object in a fluid experiences an upward force called the **buoyant force**. According to **Archimedes' principle**, this is equal to the weight of the fluid displaced by the object. The **principle of flotation** further states that, if the object is floating, the weight of displaced fluid (buoyant force) is equal to its own weight. It can be shown (see below) that whether an object sinks, rises or floats in a fluid depends entirely on density.

Archimedes' principle

Buoyant force = weight of fluid displaced

Principle of flotation

For a floating object: ***U = W*** — *where* ***U*** *= buoyant force;* ***W*** *= weight of object.*

Weight = mass (m) × acceleration due to gravity (g)
Mass = density (ρ) × volume (v)
So weight (of object or fluid displaced) = ρvg

Submarines demonstrate ***Archimedes' principle*** *and the* ***principle of flotation****. Altering air/water mix in ballast tanks alters density.*

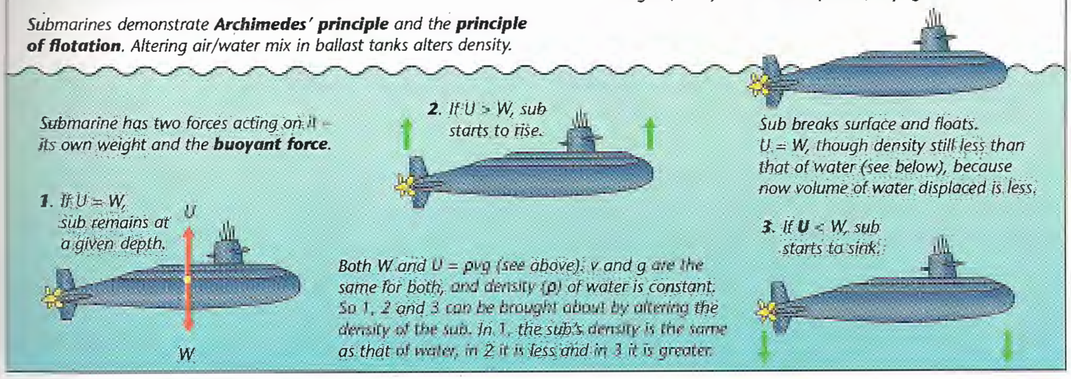

Submarine has two forces acting on it – its own weight and the ***buoyant force****.*

1. *If U = W, sub remains at a given depth.*

2. *If U > W, sub starts to rise.*

Both W and U = ρvg (see above); v and g are the same for both, and density (ρ) of water is constant. So 1, 2 and 3 can be brought about by altering the density of the sub. In 1, the sub's density is the same as that of water, in 2 it is less and in 3 it is greater.

Sub breaks surface and floats. U = W, though density still less than that of water (see below), because now volume of water displaced is less.

3. *If* ***U*** *< W, sub starts to sink.*

---

* **SI units**, 96.

TEMPERATURE

The **temperature** of an object is a measurement of how hot the object is. It is measured using **thermometers** which can be **calibrated*** to show a number of different temperature scales. The internationally accepted scales are the **absolute temperature scale** and the **Celsius scale**.

*On Venus, the **temperature** is approximately 480°C, or 753K. This is because thick clouds trap the Sun's radiation and prevent the heat from escaping.*

Thermometer

An instrument used to measure temperature. There are many different types and they all work by measuring a **thermometric property** – a property which changes with temperature. **Liquid-in-glass thermometers**, for example, measure the volume of a liquid (they are **calibrated*** so that increases in volume mark rises in temperature).

Liquid-in-glass thermometer

A common type of **thermometer** which measures temperature by the expansion of a liquid in a fine bore glass tube (**capillary tube**). A glass bulb holds a reservoir of the liquid, which is usually either mercury or colored alcohol. These substances are very responsive to temperature change – mercury is used for higher temperature ranges and alcohol for lower ones.

__Clinical thermometer__ (a type of __liquid-in-glass thermometer__). Used to measure body temperature, so has relatively small temperature range with intermediate graduations for accurate readings.

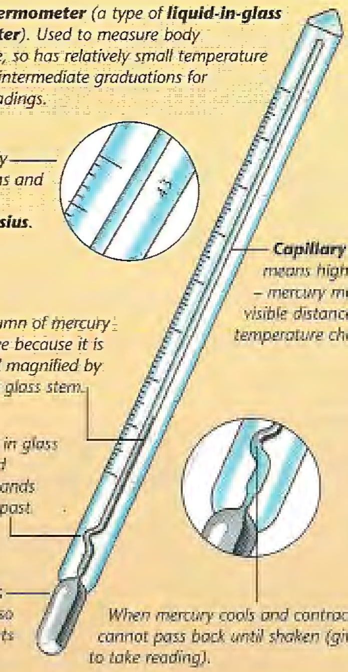

*Scale usually shows tenths and ends at 43 degrees **Celsius**.*

***Capillary tube** means high sensitivity – mercury moves a visible distance at each temperature change.*

Narrow column of mercury is easy to see because it is opaque and magnified by a triangular glass stem.

Constriction in glass tube. Heated mercury expands and pushes past.

Glass bulb is thin-walled so mercury heats up quickly.

When mercury cools and contracts, it cannot pass back until shaken (giving time to take reading).

Temperature scales

Fixed point

A temperature at which certain recognizable changes always take place (under given conditions), and which can thus be given a value against which all other temperatures can be measured. Examples are the **ice point** (the temperature at which pure ice melts) and the **steam point** (the temperature of steam above water boiling under **atmospheric pressure***). Two fixed points are used to **calibrate*** a thermometer – a **lower** and an **upper fixed point**. The distance between these points is known as the **fundamental interval**.

Using fixed points to calibrate* the Celsius scale on a thermometer

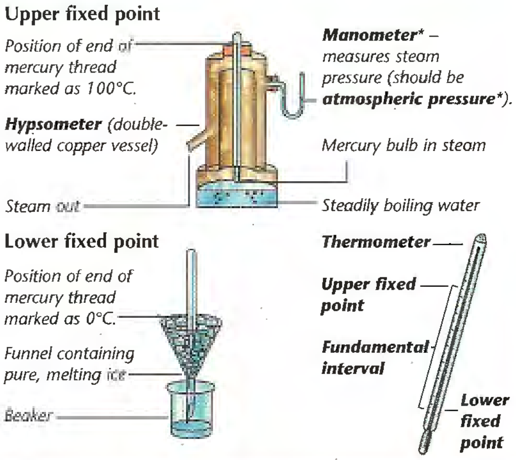

Upper fixed point
Position of end of mercury thread marked as 100°C.
Hypsometer (double-walled copper vessel)
Steam out
Manometer* – measures steam pressure (should be atmospheric pressure*).
Mercury bulb in steam
Steadily boiling water
Lower fixed point
Position of end of mercury thread marked as 0°C.
Funnel containing pure, melting ice
Beaker
Thermometer
Upper fixed point
Fundamental interval
Lower fixed point

* **Atmospheric pressure**, 25 (**Barometer**); **Calibration**, 344; **Manometer**, 25.

Maximum and **minimum thermometers**

Special **liquid-in-glass thermometers** which record the maximum or minimum temperature reached over a period of time. They contain a metal and glass **index** (see picture below) which is pushed up or pulled down (respectively) by the liquid **meniscus***. The index stays at the maximum or minimum position it reaches during the time the thermometer is left. It is reset using a magnet.

Maximum thermometer

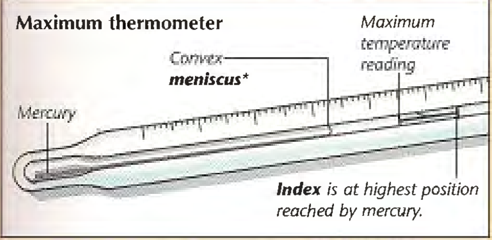
Convex meniscus*
Maximum temperature reading
Mercury
Index is at highest position reached by mercury.

Minimum thermometer

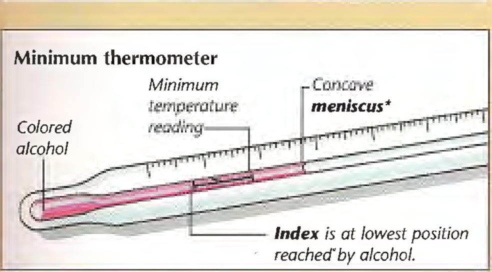
Minimum temperature reading
Concave meniscus*
Colored alcohol
Index is at lowest position reached by alcohol.

Other types of thermometer

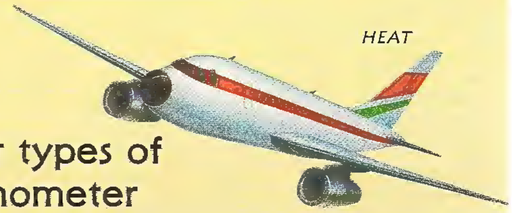

Resistance thermometer

Measures temperature from change in **resistance*** it causes in a coil of wire. Similar devices, e.g. under aircraft wings, use resistance change in **thermistors***.

*Aircraft have **thermistors*** under their wing surfaces to measure the air temperature.*

Liquid crystal thermometer

A thermometer containing liquid crystals that change color when they are heated.

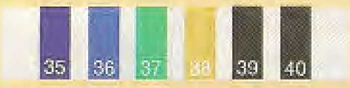
35
36
37
38
39
40

*A **liquid crystal thermometer** on the skin shows its temperature.*

Digital thermometer

A thermometer with a heat-sensitive electric component.

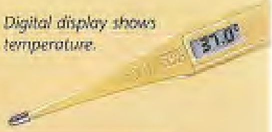

Digital display shows temperature.

Thermocouple

A device which uses the **e.m.f.*** produced across metal junctions to measure temperature difference.

***Thermocouple** (two metals, two junctions) used to find temperature X.*

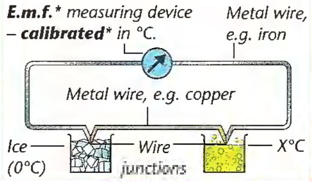
E.m.f.* measuring device – calibrated* in °C.
Metal wire, e.g. iron
Metal wire, e.g. copper
Ice (0°C)
Wire junctions
X°C

Absolute or **thermodynamic temperature scale**

A standard temperature scale, using units called **kelvins** (**K**). The zero value is given to the lowest possible temperature theoretically achievable, called **absolute zero**. It is impossible to have a lower temperature, as this would require a negative volume (see graph, right) which cannot exist.

Celsius scale (°C)

A standard temperature scale identical in graduations to the **absolute temperature scale**, but with the zero and one hundred degree values given to the **ice point** and **steam point** respectively (see **fixed point**).

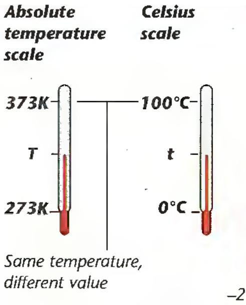
Absolute temperature scale
Celsius scale
373K
100°C
T
t
273K
0°C

Same temperature, different value

For conversion: $T = t + 273$

Volume – temperature graph for ideal gas*
(See also page 32.)

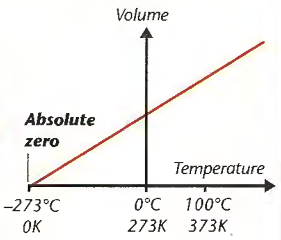

Fahrenheit scale (°F)

An old scale with the values 32°F and 212°F given to the **ice point** and **steam point** respectively (see **fixed point**). It is rarely used in scientific work.

* **Calibration**, 344; **Electromotive force (e.m.f.)**, 60; **Ideal gas**, 33; **Meniscus**, 345; **Resistance**, 62; **Thermistor**, 65.

TRANSFER OF HEAT

Whenever there is a temperature difference, **heat energy** (see page 9) is transferred by **conduction**, **convection** or **radiation** from the hotter to the cooler place. This increases the **internal energy*** of the cooler atoms, raising their temperature, and decreases the energy of the hotter atoms, lowering theirs. It continues until the temperature is the same across the region – a state called **thermal equilibrium**.

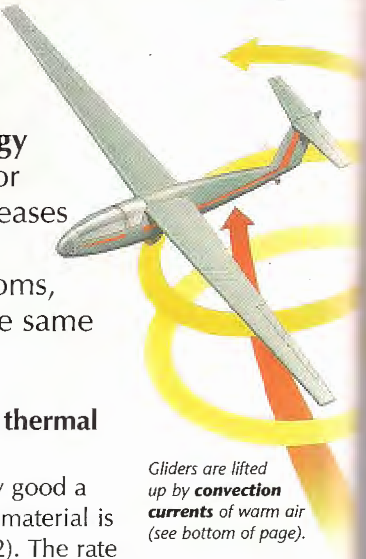

Gliders are lifted up by ***convection currents*** *of warm air (see bottom of page).*

Conduction or thermal conduction

The way in which heat energy is transferred in solids (and also, to a much lesser extent, in liquids and gases). In good **conductors** the energy transfer is rapid, occurring mainly by the movement of free **electrons*** (electrons which can move about), although also by the vibration of atoms – see **insulators** (bad conductors), below.

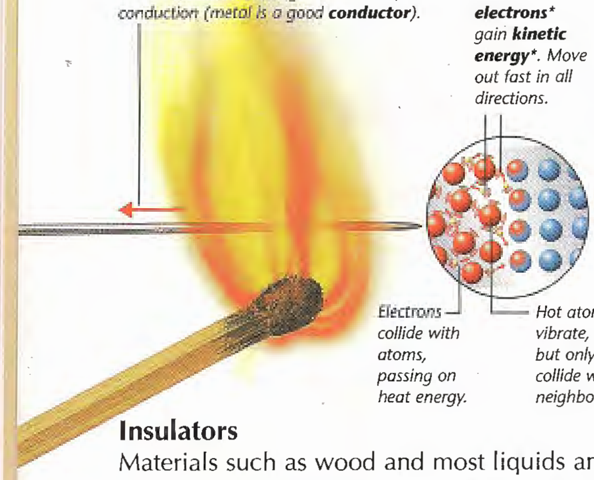

Heat is transferred along the needle by conduction (metal is a good ***conductor****).*

Heated ***electrons**** *gain* ***kinetic energy****. Move out fast in all directions.*

Electrons collide with atoms, passing on heat energy.

Hot atoms vibrate, but only collide with neighbors.

Insulators

Materials such as wood and most liquids and gases, in which the process of **conduction** is very slow (they are bad **conductors**). As they do not have free **electrons***, heat energy is only transferred by conduction by the vibration and collision of neighboring atoms.

Conductivity or thermal conductivity

A measure of how good a heat **conductor** a material is (see also page 112). The rate of heat energy transfer per unit area through an object depends on the conductivity of the material and the **temperature gradient**. This is the temperature change with distance along the material. The higher the conductivity and the steeper the gradient, the faster the energy transfer that takes place.

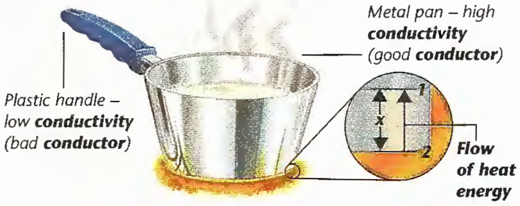

Metal pan – high ***conductivity*** *(good* ***conductor****)*

Plastic handle – low ***conductivity*** *(bad* ***conductor****)*

Flow of heat energy

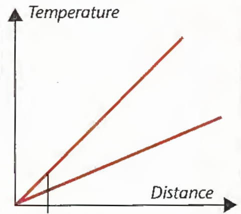

Steeper ***temperature gradient*** *means greater temperature difference over same distance, hence faster* ***conduction****.*

$$\textbf{Temperature gradient} = \frac{t_2 - t_1}{x}$$

where t_1, t_2 = temperatures at points 1 and 2; x = distance.

$$\textbf{Rate of transfer of heat energy per unit area} = k\,\frac{t_2 - t_1}{x}$$

where k = conductivity of metal.

Convection

A way in which heat energy is transferred in liquids and gases. If a liquid or gas is heated, it expands, becomes less dense and rises. Cooler, denser liquid or gas then sinks to take its place. Thus a **convection current** is set up. The picture on the right shows how convection currents cause daytime coastal breezes, a process which is reversed at night.

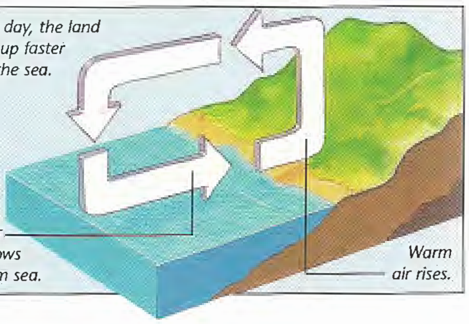

In the day, the land heats up faster than the sea.

Cooler air blows in from sea.

Warm air rises.

---

* **Electrons**, 83; **Internal energy**, **Kinetic energy**, 9.

Radiation

A way in which heat energy is transferred from a hotter to a cooler place without the **medium*** taking part in the process. This can occur though a vacuum, unlike **conduction** and **convection**. The term radiation is also often used to refer to the heat energy itself, otherwise known as **radiant heat energy**. This takes the form of **electromagnetic waves***, mainly **infra-red radiation***. When these waves fall on an object, some of their energy is absorbed, increasing the object's **internal energy*** and hence its temperature. See also **Leslie's cube**, right.

Use of radiation to supply hot water

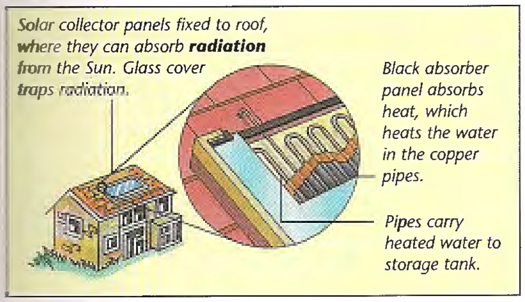
Solar collector panels fixed to roof, where they can absorb **radiation** from the Sun. Glass cover traps radiation.
Black absorber panel absorbs heat, which heats the water in the copper pipes.
Pipes carry heated water to storage tank.

Thermopile

A device for measuring **radiation** levels. It consists of two or more **thermocouples*** (normally over 50) joined end to end. Radiation falls on the metal junctions on one side and the temperature difference between these hot junctions and the cold ones on the other side produces **e.m.f.*** across the thermopile, the size of which indicates how much radiation has been absorbed.

Thermopile

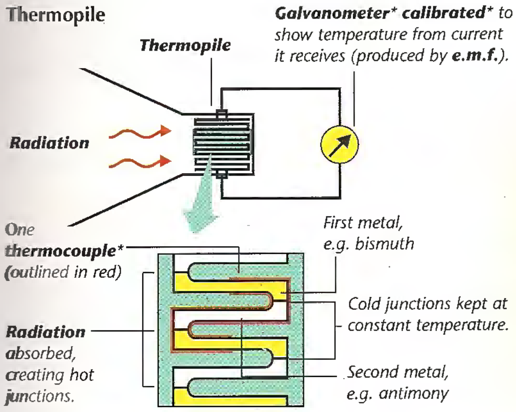
Thermopile
Galvanometer calibrated** to show temperature from current it receives (produced by **e.m.f.**).
Radiation
One thermocouple* (outlined in red)
First metal, e.g. bismuth
Cold junctions kept at constant temperature.
Radiation absorbed, creating hot junctions.
Second metal, e.g. antimony

Leslie's cube

A thin-walled, hollow cube (good **conductor**) with different outside surfaces. It is used to show that surfaces vary in their ability to **radiate** and absorb heat energy. Their powers of doing so are compared with an ideal called a **black body**, which absorbs all radiation that falls on it, and is also the best radiator.

Leslie's cube *used to compare powers of* ***radiation*** *– numbers show best (1) to worst (4) surface.*

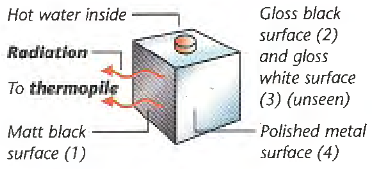
Hot water inside
Radiation
To thermopile
Matt black surface (1)
Gloss black surface (2) and gloss white surface (3) (unseen)
Polished metal surface (4)

Vacuum flask

A flask which keeps its contents at constant temperature. It consists of a double-walled glass container, with a vacuum between the walls (stopping heat energy transfer by **conduction** and **convection**) and shiny surfaces (minimizing transfer by **radiation**).

Vacuum flask

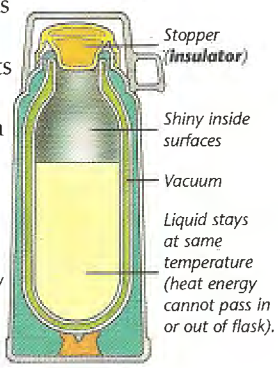
Stopper (insulator)
Shiny inside surfaces
Vacuum
Liquid stays at same temperature (heat energy cannot pass in or out of flask).

Greenhouse effect

The warming effect produced when **radiation** is trapped in a closed area, e.g. a greenhouse. The objects inside absorb the Sun's radiation and re-emit lower energy radiation which cannot pass back through the glass. Carbon dioxide in the atmosphere forms a similar barrier, and its level is increasing, hence the Earth is slowly getting warmer.

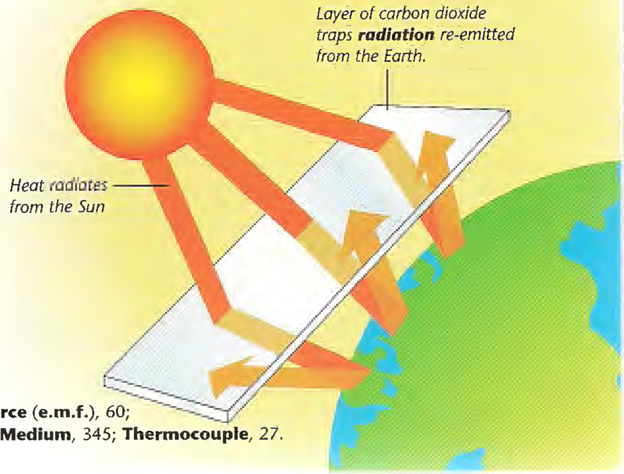
Layer of carbon dioxide traps **radiation** re-emitted from the Earth.
Heat radiates from the Sun

---

* **Calibration**, 344; **Electromagnetic waves**, 44; **Electromotive force (e.m.f.)**, 60; **Galvanometer**, 77; **Infra-red radiation**, 45; **Internal energy**, 9; **Medium**, 345; **Thermocouple**, 27.

EFFECTS OF HEAT TRANSFER

When an object absorbs or loses **heat energy** (see pages 28-29), its **internal energy*** increases or decreases. This results in either a rise or fall in temperature (the amount of which depends on the **heat capacity** of the object) or a **change of state**.

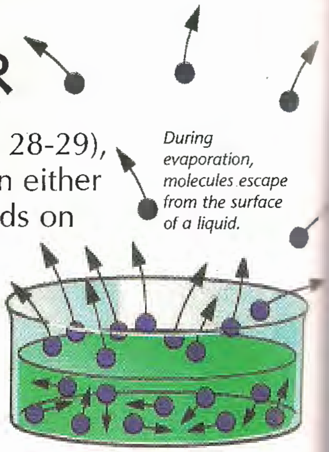

During evaporation, molecules escape from the surface of a liquid.

Changes of state

A **change of state** is a change from one **physical state** (the solid, liquid or gaseous state) to another (for more about **physical states**, see page 5). While a change of state is happening, there is no change in temperature. Instead, all the energy taken in or given out is used to make or break molecular bonds. This is called **latent heat** (**L**) – see graphs, page 31. The **specific latent heat** (**l**) of a substance is a set value, i.e. the heat energy taken in or given out when 1kg of the substance changes state.

Evaporation

The conversion of a liquid to a vapor by the escape of molecules from its surface. It takes place at all temperatures, the rate increasing with any one or a combination of the following: increase in temperature, increase in surface area or decrease in pressure. It is also increased if the vapor is immediately removed from above the liquid by a flow of air. The **latent heat** (see above) needed for evaporation is taken from the liquid itself which cools and in turn cools its surroundings.

Vaporization

The change of state from liquid to gaseous at a temperature called the **boiling point** (when the liquid is said to be **boiling**). The term is also used more generally for any change resulting in a gas or vapor, i.e. including also **evaporation** and **sublimation**.

Condensation

The change of state from gas or vapor to liquid.

Melting

The change of state from solid to liquid at a temperature called the **melting point** of the solid.

Freezing

The change of state from liquid to solid at the **freezing point** (the same temperature as the **melting point** of the solid).

Sublimation

The conversion of a substance from a solid directly to a gas, or vice-versa, without passing through the liquid state. Iodine and carbon dioxide are two substances that **sublime**.

Changes of state
Temperature remains constant (see graphs, page 31).

🟨 *Changes due to heating*

*Heat energy taken in, which would have raised the temperature, used instead (as **latent heat**) to break bonds.*

🟦 *Changes due to cooling*

*Heat energy which would have been lost (lowering temperature) used instead (as **latent heat**) to make bonds between molecules.*

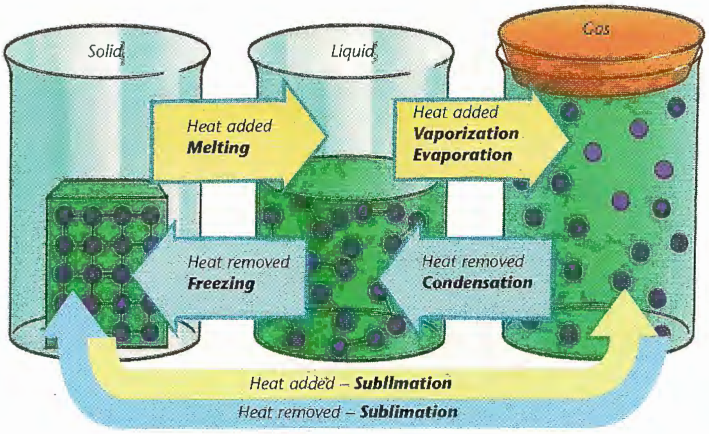
Solid
Liquid
Gas
Heat added **Melting**
Heat added **Vaporization Evaporation**
Heat removed **Freezing**
Heat removed **Condensation**
Heat added – **Sublimation**
Heat removed – **Sublimation**

*** Internal energy**, 9.

Graph showing increase of temperature with heat added

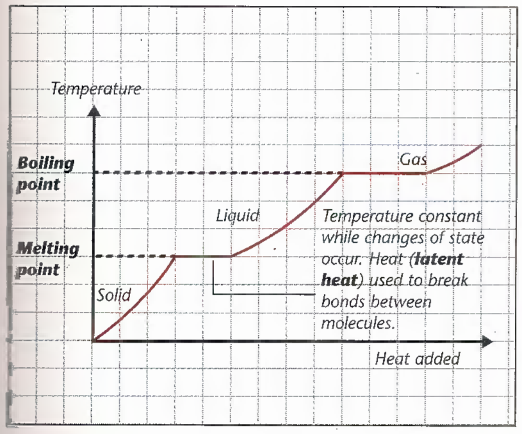

Graph showing decrease of temperature as object cools

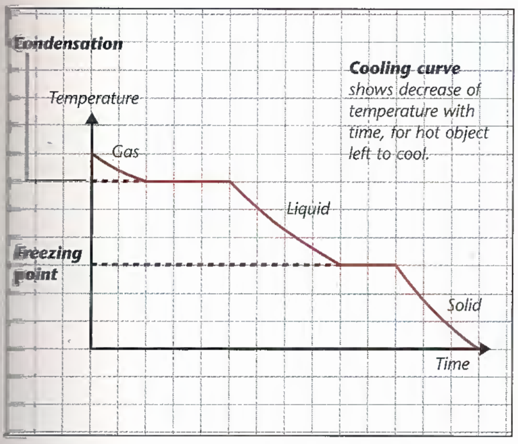

Specific latent heat of vaporization

The heat energy taken in when 1kg of a substance changes from a liquid to a gas at its **boiling point**. It is the same as the heat given out when the process is reversed

Specific latent heat of fusion

The heat energy taken in when 1kg of a substance changes from a solid to a liquid at its **melting point**. It is the same as the heat given out when the process is reversed. See also page 112.

$Q = ml$ *where Q = heat energy lost or gained by object; m = mass; l =* ***specific latent heat****.*

The ***SI unit**** *of* ***specific latent heat*** *is the joule per kilogram ($J\ kg^{-1}$).*

Heat capacity (C)

The heat energy taken in or given out by an object per temperature change by 1K. It is a property of the object and depends on both its mass and the material(s) of which it is made (as well as the temperature and pressure), hence its value is different for every object.

$Q = C(t_2 - t_1)$ *where Q = heat energy lost or gained; C = heat capacity; t_1 and t_2 = initial and final temperatures respectively.*

The ***SI unit**** *of* ***heat capacity*** *is the joule per kelvin ($J\ K^{-1}$).*

Specific heat capacity (c)

The heat energy taken in or given out per unit mass per unit temperature change. It is a property of the substance alone, i.e. there is a set value for each substance (though this changes with temperature and pressure). See also page 112.

$Q = mc\,(t_2 - t_1)$ *where m = mass; c =* ***specific heat capacity****; Q, t_1, t_2 as above.*

The ***SI unit**** *of* ***specific heat capacity*** *is the joule per kilogram per kelvin ($J\ kg^{-1}\ K^{-1}$).*

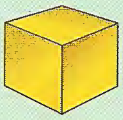

*Mass (m) of 2kg brass (**specific heat capacity** 380J $kg^{-1}\ K^{-1}$) heated for a set time. Temperature rises from 303K (t_1) to 307K (t_2).*

Q (heat gained) = 2 × 380 × (307–303)J

So Q = 3,040J

Thus same amount of heat energy taken in by 16kg of brass would raise temperature by 0.5K.

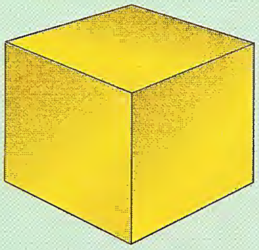

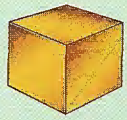

Same amount of heat energy given to mass of 2kg of copper causes temperature rise of 3.8K.

*Thus **specific heat capacity** of copper is 400J $kg^{-1}\ K^{-1}$.*

---

* **SI units**, 96.

EXPANSION DUE TO HEATING

Most substances expand when heated – their molecules move faster and further apart. The extent of this expansion (**expansivity**) depends on **intermolecular forces***. For the same amount of heat applied (at constant pressure), solids expand least, as their molecules are closest together and so have the strongest forces between them. Liquids expand more, and gases the most.

Expansion of solids on heating must be taken into account in building work.

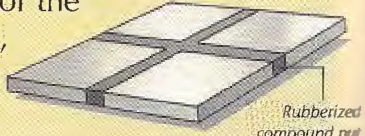

Rubberized compound put between paving stones

Bimetallic strip

A device which shows the expansion of solids due to heating. It is made up of two different strips of metal, joined along their (equal) length. When heated or cooled, both metals expand or contract (respectively), but at different rates, so the strip bends. Such strips are used in **thermostats**.

Thermostat (temperature regulator)

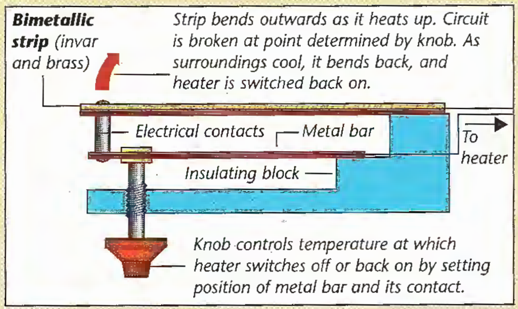

Strip bends outwards as it heats up. Circuit is broken at point determined by knob. As surroundings cool, it bends back, and heater is switched back on.

Knob controls temperature at which heater switches off or back on by setting position of metal bar and its contact.

Linear coefficient of expansion (α)

A measurement of the fraction of its original length by which a solid expands for a temperature rise of 1K.

Areal coefficient of expansion (β)

A measurement of the fraction of its original area by which a solid expands for a temperature rise of 1K.

For solids or liquids:

$$\text{Expansivity (linear, superficial or cubic)} = \frac{\text{change in (length, area or volume)}}{\text{original (length, area or volume)} \times \text{temperature rise}}$$

*Note the only relevant measurement for liquids is **cubic expansivity**. It is either **real** or **apparent** (see right), therefore so also is the change in volume in formula.*

For gases:

$$\text{Cubic expansivity} = \frac{\text{change in volume at constant pressure}}{\text{volume at 0°C (273K)} \times \text{temperature rise}}$$

Volume coefficient of expansion (γ)

A measurement of the fraction of its original volume by which a substance expands for a temperature rise of 1K. It is the same for all gases (at constant pressure) when they are assumed to behave as **ideal gases**. Since gases expand by very large amounts, the original volume is always taken at 0°C so that proper comparisons can be made (this is not necessar with solids or liquids as the changes are so small).

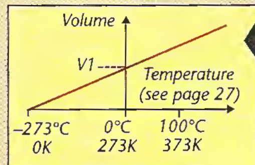

*Change in volume with temperature of **ideal gas** (constant pressure). **Law of volumes** (volume increases proportionally with **absolute temperature***) is obeyed.*

From zero volume, there are 273 temperature graduations to volume at 0°C (V1).

Graph rises proportionally, so for each graduation (kelvin rise), volume of gas increases by $^1/_{273}$ of volume V1.

*Thus, for an **ideal gas**:*

$$\text{Cubic expansivity} = \frac{1}{273}\,\text{K}^{-1}$$

Real or absolute cubic expansivity

An accurate measurement of the fraction of its volume by which a liquid expands for a temperature rise of 1K.

Apparent cubic expansivity

A measurement of the fraction of its volume by which a liquid apparently expands for a temperature rise of 1K. In fact, the heat applied also causes very slight expansion of the container, so its calibrated measurements are no longer valid.

Anomalous expansion

The phenomenon whereby some liquids contract instead of expanding when the temperature rises within a certain range (e.g. water between 0°C and 4°C).

---

* **Absolute temperature scale**, 27; **Intermolecular forces**, 7.

Behavior of gases

All gases behave in a similar way, and there are several **gas laws** which describe their behavior (see below and right). An **ideal gas** is a theoretical gas which, by definition, exactly obeys **Boyle's law** at all temperatures and pressures, but in fact also obeys the two other laws as well. When real gases are at normal temperatures and pressures, they show approximately ideal behavior (the higher the temperature and the lower the pressure, the better the approximation), hence the laws may be generally applied.

Boyle's law

The volume of a fixed mass of gas at constant temperature is inversely proportional to the pressure. For example, if the pressure on the gas increases, the volume decreases proportionally – the molecules move closer together. Note that the pressure exerted by the gas increases (the molecules hit the container walls more often).

Boyle's law

$$V \propto \frac{1}{P} \quad \text{or} \quad PV = \text{constant}$$

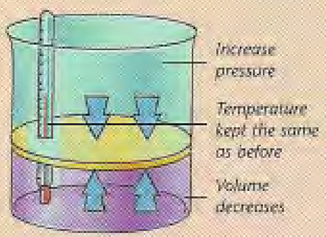

Key

P = *pressure*
V = *volume*
T = *temperature on* ***absolute scale****
R = ***gas constant****

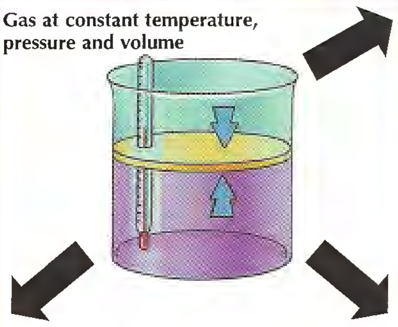

Gas at constant temperature, pressure and volume

The ***ideal gas equation****,* ***general gas equation*** *or* ***equation of state*** *links the temperature, pressure and volume. For one* ***mole**** *of gas:*

$$\frac{PV}{T} = R \quad \text{or} \quad PV = RT$$

Pressure law

The pressure of a fixed mass of gas at constant volume is proportional to the temperature on **the absolute scale***. For example, if the temperature increases but the volume is kept the same, the pressure inside the gas increases proportionally – the molecules move faster, and hit the container walls more often. Note that the pressure exerted on the gas to keep the volume constant must increase.

Pressure law

$$P \propto T \quad \text{or} \quad \frac{P}{T} = \text{constant}$$

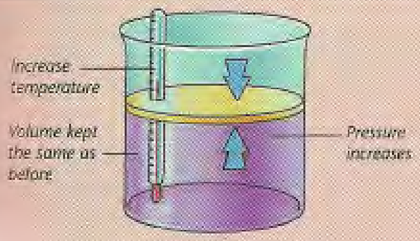

Law of volumes

The volume of a fixed mass of gas at constant pressure is proportional to the temperature on the **absolute scale***. For example, if the temperature increases and the pressure is kept the same, the volume increases proportionally (given an expandable container) – the molecules move faster and further apart. Note that the pressure exerted by the gas remains constant (the molecules hit the walls at the same frequency – they have more space, but greater energy).

Law of volumes

$$V \propto T \quad \text{or} \quad \frac{V}{T} = \text{constant}$$

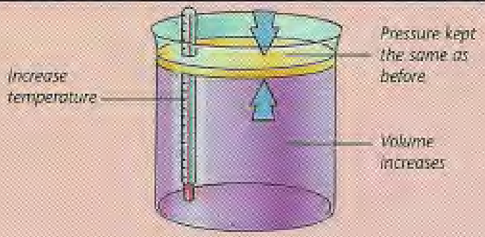

* **Absolute temperature scale**, 27; **Gas constant**, 113; **Mole**, 96.

WAVES

All **waves** transport energy without permanently displacing the **medium*** through which they travel. They are also called **traveling waves**, as the energy travels from a source to surrounding points (but see also **stationary wave**, page 43). There are two main types – **mechanical waves**, such as sound waves, and **electromagnetic waves** (see page 44). In all cases, the **wave motion** is regular and repetitive (i.e. **periodic motion** – see page 16) in the form of **oscillations** – regular changes between two extremes. In mechanical waves it is particles (molecules) that oscillate, and in electromagnetic waves it is electric and magnetic fields.

***Mechanical wave* (*transverse wave*)** *passes along string.*

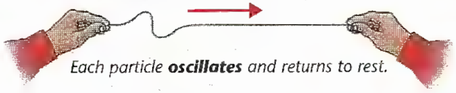
Each particle **oscillates** and returns to rest.

Displacement/time graph for oscillation of one particle

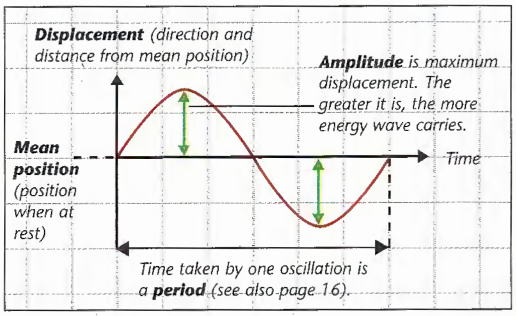

Transverse waves

Waves in which the oscillations are at right angles to the direction of energy (wave) movement, e.g. water waves (oscillation of particles) and all **electromagnetic waves*** (oscillation of fields – see introduction).

Crests or peaks

Points where a wave causes maximum positive displacement of the **medium***. The crests of some waves, e.g. water waves, can be seen as they travel.

Troughs

Points where a wave causes maximum negative displacement of the **medium***. The troughs of some waves, e.g. water waves, can be seen as they travel.

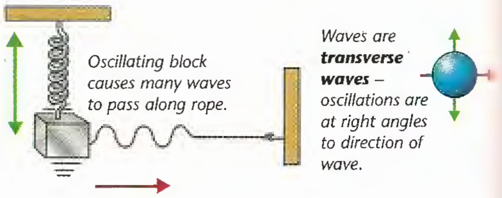
Oscillating block causes many waves to pass along rope.
Waves are **transverse waves** – oscillations are at right angles to direction of wave.

Displacement/distance graph for particles of section of rope at two "frozen" moments

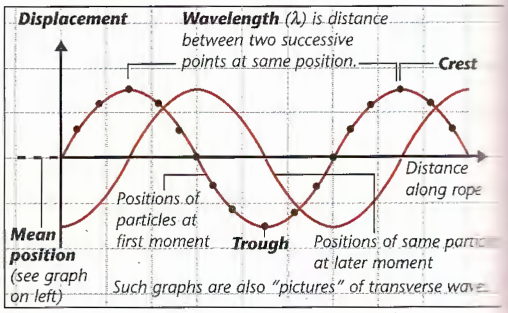

Wavefront

Any line or section taken through an advancing wave which joins all points that are in the same position in their oscillations. Wavefronts are usually at right angles to the direction of the waves and can have any shape, e.g. **circular** and **straight wavefronts**.

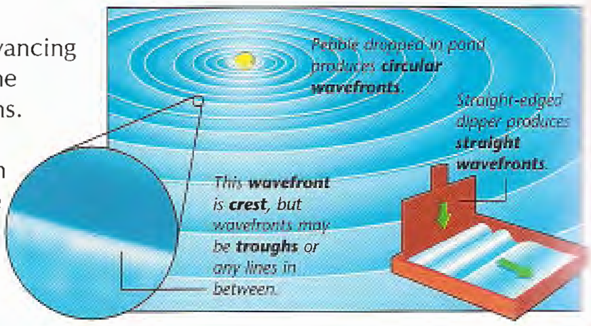
Pebble dropped in pond produces **circular wavefronts**.
Straight-edged dipper produces **straight wavefronts**.
This **wavefront** is **crest**, but wavefronts may be **troughs** or any lines in between.

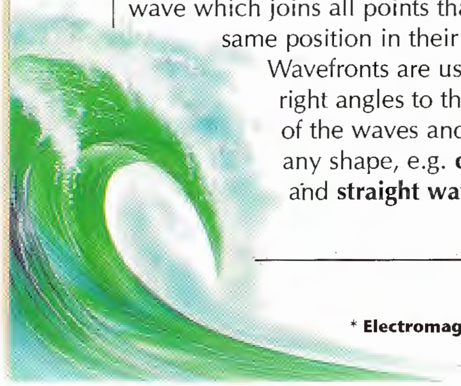

---

* **Electromagnetic waves**, 44; **Medium**, 345.

Longitudinal waves

Waves in which the oscillations are along the line of the direction of wave movement, e.g. sound waves. They are all **mechanical waves** (see introduction), i.e. it is particles which oscillate.

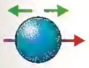

*In **longitudinal waves**, particles oscillate along line of wave direction.*

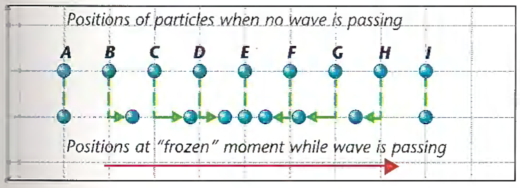
Positions of particles when no wave is passing
A
B
C
D
E
F
G
H
I
Positions at "frozen" moment while wave is passing

Graph of particles above at "frozen" moment

Graph not in this case "picture" of wave (see second graph, page 34).

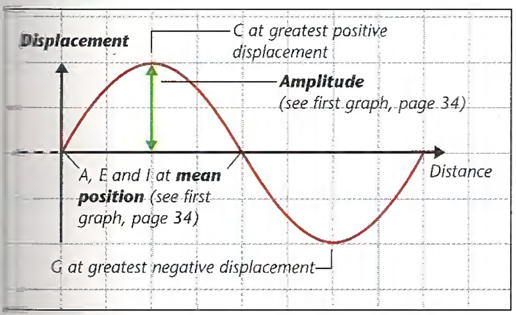

Compressions

Regions along a **longitudinal wave** where the pressure and density of the molecules are higher than when no wave is passing.

Rarefactions

Regions along a **longitudinal wave** where the pressure and density of the molecules are lower than when no wave is passing.

Graph of pressure or density versus distance for longitudinal wave shows compressions and rarefactions.

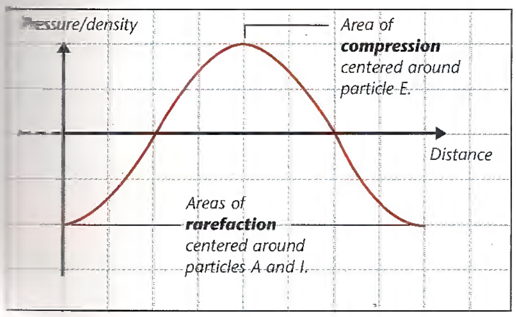

Wave speed

The distance moved by a wave per unit of time. It depends on the **medium*** through which the wave is traveling.

$$\text{Wave speed} = \frac{\text{distance moved by wave}}{\text{time}}$$

$$= \frac{\text{number of waves passing point} \times \text{wavelength}}{\text{time}}$$

$$= \text{frequency} \times \text{wavelength}$$

Hence:

$$v = f\lambda$$

where v = wave speed;
f = frequency; λ = wavelength.

Frequency (f)

The number of oscillations which occur in one second when waves pass a given point (see also page 16). It is equal to the number of **wavelengths** (see second graph, page 34) per second.

Attenuation

The gradual decrease in **amplitude** (see first graph, page 34) of a wave as it passes through matter and loses energy. The amplitudes of oscillations occurring further from the source are less than those of oscillations nearer to it. This can be seen as an overall **damping***.

Graph showing attenuated wave

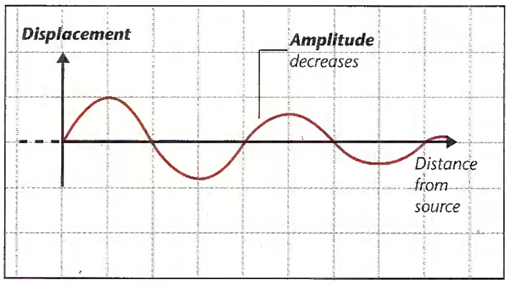

Wave intensity

A measurement of the energy carried by a wave. It is worked out as the amount of energy falling on unit area per second. It depends on the **frequency** and **amplitude** of the wave, and also on the **wave speed**.

Damping, 16; **Medium**, 345.

REFLECTION, REFRACTION AND DIFFRACTION

An obstacle or a change of **medium*** causes a wave to undergo **reflection**, **refraction** or **diffraction**. These are different types of change in wave direction and often also result in changes in the shape of the **wavefronts***. For more about the reflection and refraction of light waves, see pages 47-53.

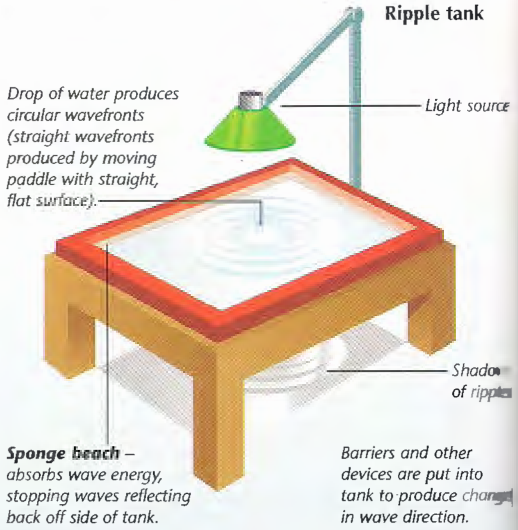
Ripple tank
Drop of water produces circular wavefronts (straight wavefronts produced by moving paddle with straight, flat surface).
Light source
Shado
of ripp
***Sponge beach** – absorbs wave energy, stopping waves reflecting back off side of tank.*
Barriers and other devices are put into tank to produce chang
in wave direction.

Ripple tank

A tank of water used to demonstrate the properties of water waves (see right).

Reflection

The change in direction of a wave due to its bouncing off a boundary between two **media***. A wave that has undergone reflection is called a **reflected wave**. The shape of its wavefronts depends on the shape of the **incident wavefronts** and the shape of the boundary. For more about the reflection of light waves, see pages 47-49.

Incident wave

A wave that is traveling toward a boundary between two **media***. Its wavefronts are called **incident wavefronts**.

Circular wavefronts hitting concave boundary produce straight wavefronts in this case (i.e. with ship at this distance).

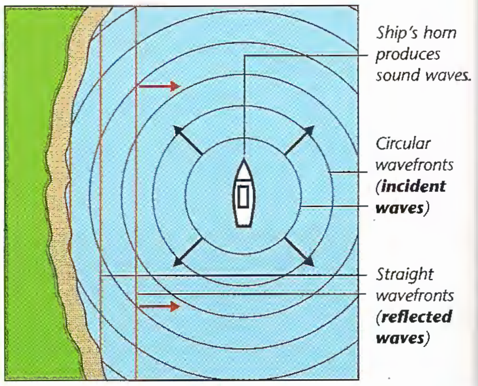
Ship's horn produces sound waves.
*Circular wavefronts (**incident waves**)*
*Straight wavefronts (**reflected waves**)*

Examples of reflected wave shapes

Circular wavefronts hitting straight boundary produce circular wavefronts.

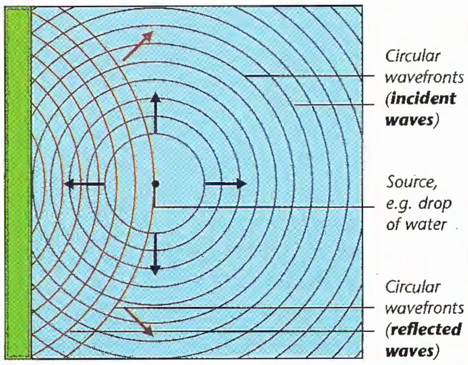
*Circular wavefronts (**incident waves**)*
Source, e.g. drop of water
*Circular wavefronts (**reflected waves**)*

Straight wavefronts hitting straight boundary produce straight wavefronts.

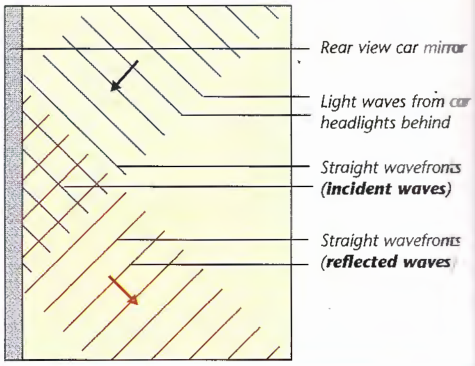
Rear view car mirror
Light waves from car headlights behind
*Straight wavefronts (**incident waves**)*
*Straight wavefronts (**reflected waves***

***Medium**, 345; **Wavefront**, 34.

Refraction

The change in direction of a wave when it moves into a new **medium*** which causes it to travel at a different speed. A wave which has undergone refraction is called a **refracted wave**. Its **wavelength*** increases or decreases with the change in speed, but there is no change in **frequency***. For more about the refraction of light waves, see pages 50-53.

Examples of refraction of waves in a ripple tank when moving into a new medium

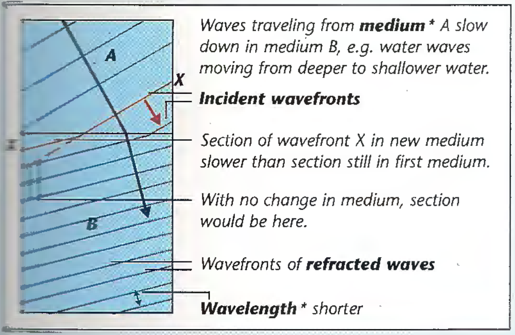
A
X
B
Waves traveling from **medium*** A slow down in medium B, e.g. water waves moving from deeper to shallower water.
Incident wavefronts
Section of wavefront X in new medium slower than section still in first medium.
With no change in medium, section would be here.
Wavefronts of **refracted waves**
Wavelength* shorter

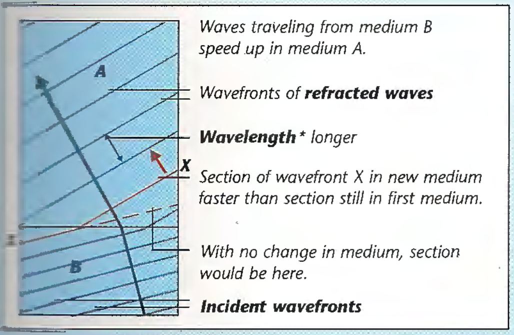
A
B
X
Waves traveling from medium B speed up in medium A.
Wavefronts of **refracted waves**
Wavelength* longer
Section of wavefront X in new medium faster than section still in first medium.
With no change in medium, section would be here.
Incident wavefronts

Other examples:

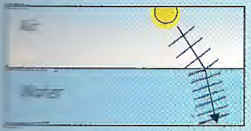

Waves slow down on entering a denser medium.

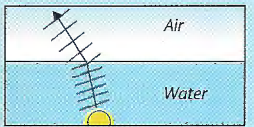

They speed up on entering a less dense medium.

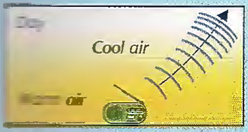
Day
Cool air
Warm air

Sound waves slow down on entering a cooler medium (cooler means denser).

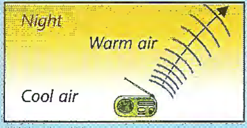
Night
Warm air
Cool air

They speed up on entering a warmer medium (warmer means less dense).

Refractive index (n)

A number which indicates the power of refraction of a given **medium*** relative to a previous medium. It is found by dividing the speed of the **incident wave** in the first medium by the speed of the **refracted wave** in the given medium (subscript numbers are used – see formula). The **absolute refractive index** of a medium is the speed of light in a vacuum (or, generally, in air) divided by the speed of light in that medium†.

$${}_1n_2 = \frac{v_1}{v_2}$$

where v_1, v_2 = speeds in first and second **media***.

This means **refractive index** of **medium*** 2 relative to medium 1.

Diffraction

The bending effect which occurs when a wave meets an obstacle or passes through an aperture. The amount the wave bends depends on the size of the obstacle or aperture compared to the **wavelength*** of the wave. The smaller the obstacle or aperture by comparison, the more the wave bends.

Diffraction of waves (sound waves) around obstacle

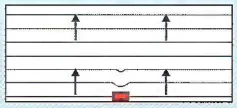

Obstacle small compared to **wavelength*** (wavelength of sound is about 2m) – a lot of diffraction, so no "shadow" formed.

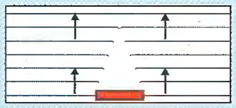

Obstacle about same size as wavelength – some diffraction, so small "shadow" formed, i.e. area through which no waves pass.

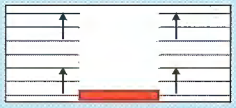

Obstacle large compared to wavelength – almost no diffraction, so large "shadow" formed.

Diffraction of water waves passing through aperture (slit)

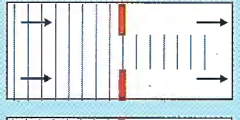

Aperture wide compared to **wavelength*** – little diffraction.

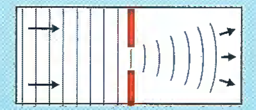

Aperture about same size as wavelength – some diffraction.

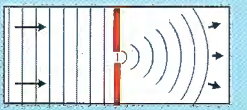

Aperture narrow compared to wavelength – a lot of diffraction.

Frequency, 35; **Medium**, 345; **Wavelength**, 34.

† For more about **refractive index** and light, see page 50.

WAVE INTERFERENCE

When two or more waves travel in the same or different directions in a given space, variations in the size of the resulting disturbance occur at points where they meet (see **principle of superposition**). This effect is called **interference**. When interference is demonstrated, e.g. in a **ripple tank***, sources which produce **coherent waves** are always used, i.e. waves with the same wavelength and frequency, and either **in phase** or with a constant **phase difference** (see **phase**). This ensures that the interference produces a regular, identifiable **interference pattern** of disturbance (see picture, page 39). The use of non-coherent waves would result only in a constantly-changing confusion of waves.

Phase

Two waves are **in phase** if they are of the same frequency and corresponding points are at the same place in their oscillations (e.g. both at **crests***) at the same instant. They are **out of phase** if this is not the case, and exactly out of phase if their displacements are exactly opposite (e.g. a crest and a **trough***). The **phase difference** between two waves is the amount, measured as an angle, by which a point on one wave is ahead of or behind the corresponding point on the other. For waves exactly out of phase, the phase difference is 180°; for waves in phase, it is 0°.

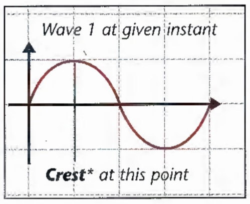
Wave 1 at given instant
Crest* at this point

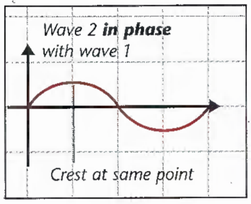
Wave 2 in phase with wave 1
Crest at same point

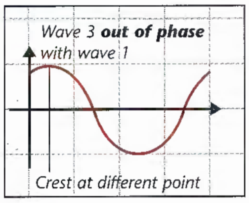
Wave 3 out of phase with wave 1
Crest at different point

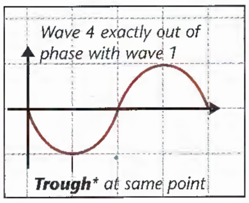
Wave 4 exactly out of phase with wave 1
Trough* at same point

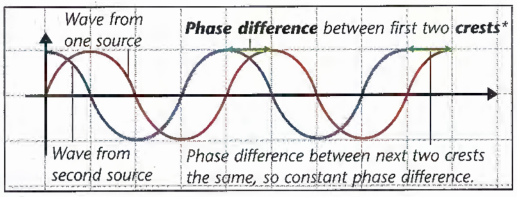
Wave from one source
Phase difference between first two crests*
Wave from second source
Phase difference between next two crests the same, so constant phase difference.

Young's slits

An arrangement of narrow, parallel slits, used to create two sources of **coherent** light (see introduction). They are needed because coherent light waves cannot be produced (for studying interference) as easily as other coherent waves, as light wave emission is usually random. The interference of the light **diffracted*** through the slits is seen on a screen as light and dark bands called **interference fringes**.

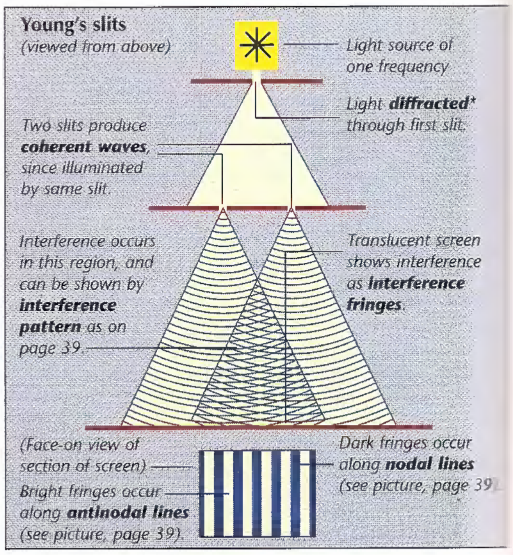
Young's slits
(viewed from above)
Light source of one frequency
Light diffracted* through first slit.
Two slits produce coherent waves, since illuminated by same slit.
Interference occurs in this region, and can be shown by interference pattern as on page 39.
Translucent screen shows interference as interference fringes.
(Face-on view of section of screen)
Bright fringes occur along antinodal lines (see picture, page 39).
Dark fringes occur along nodal lines (see picture, page 39)

Principle of superposition

States that when the **superposition** of two or more waves occurs at a point (i.e. two or more waves come together), the resultant displacement is equal to the sum of the displacements (positive or negative) of the individual waves.

---

* **Crests**, 34; **Diffraction**, 37; **Ripple tank**, 36; **Troughs**, 34.

Constructive interference

The increase in disturbance (reinforcement) which results from the **superposition** of two waves which are **in phase** (see **phase**).

Constructive interference

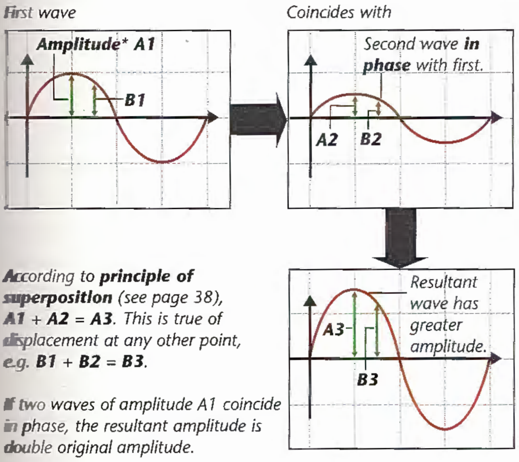
First wave

Coincides with

Second wave ***in phase*** *with first.*

Resultant wave has greater amplitude.

According to ***principle of superposition*** *(see page 38),* ***A1*** *+* ***A2*** *=* ***A3****. This is true of displacement at any other point, e.g.* ***B1*** *+* ***B2*** *=* ***B3****.*

If two waves of amplitude A1 coincide in phase, the resultant amplitude is double original amplitude.

Destructive interference

The decrease in disturbance which results from the **superposition** of two waves which are **out of phase**.

Destructive interference

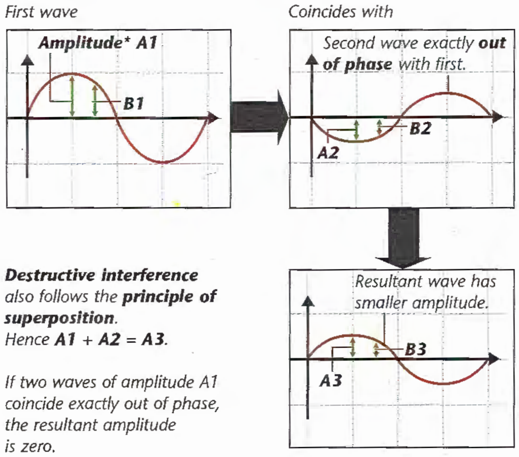
First wave

Coincides with

Second wave exactly ***out of phase*** *with first.*

Resultant wave has smaller amplitude.

Destructive interference *also follows the* ***principle of superposition****. Hence* ***A1*** *+* ***A2*** *=* ***A3****.*

If two waves of amplitude A1 coincide exactly out of phase, the resultant amplitude is zero.

Nodes or nodal points

Points at which destructive interference is continually occurring, and which are consequently regularly points of minimum disturbance, i.e. points where **crest*** meets **trough*** or **compression*** meets **rarefaction***. A **nodal line** is a line consisting entirely of nodes. Depending on waves, nodal lines may indicate, for example, calm water, soft sound or darkness (see also **Young's slits** picture, page 38).

Antinodes or antinodal points

Points at which **constructive interference** is continually occurring, and which are consequently regularly points of maximum disturbance, i.e. points where two **crests***, **troughs***, **compressions*** or **rarefactions*** meet. An **antinodal line** is a line consisting entirely of antinodes. Depending on waves, antinodal lines may indicate, for example, areas of rough water, loud sound or bright light (see also **Young's slits** picture).

Interference pattern at "frozen" moment
(not all antinodal/nodal lines shown).

*Two sources (****S1*** *and* ***S2****) produce* ***coherent waves****, in this case* ***in phase****.*

Nodal line (destructive interference)*. If waves are same amplitude, disturbance at all points along it is zero.*

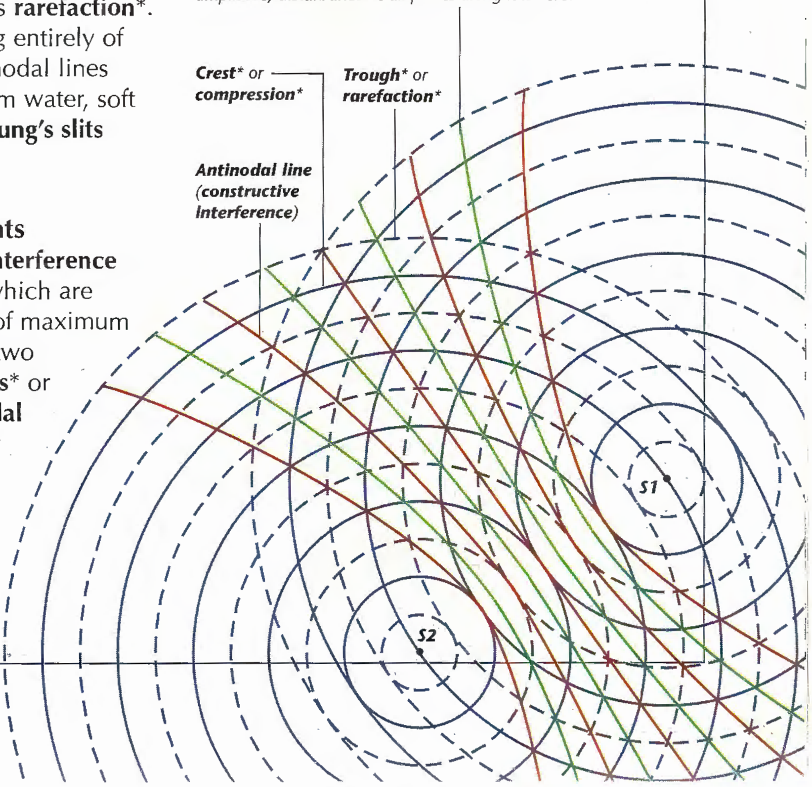
Crest* *or* ***compression****

Trough* *or* ***rarefaction****

Antinodal line (constructive interference)

---

* **Amplitude**, 34; **Compressions**, 35;
Crests, 34; **Rarefactions**, 35; **Troughs**, 34.

SOUND WAVES

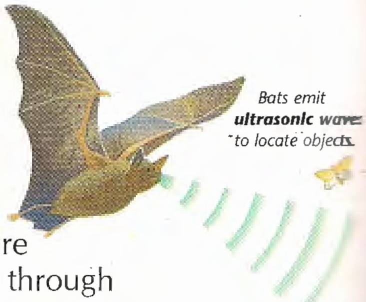

Bats emit ***ultrasonic waves*** *to locate objects.*

Sound waves, also called **acoustic waves**, are **longitudinal waves*** – waves which consist of particles oscillating along the same line as the waves travel, creating areas of high and low pressure (**compressions*** and **rarefactions***). They can travel through solids, liquids and gases and have a wide range of **frequencies***. Those with frequencies between about 20 and 20,000 **Hertz*** (the **sonic range**) can be detected by the human ear and are what is commonly referred to as sound (for more about perception of sound, see pages 42-43). Others, with higher and lower frequencies, are known as **ultrasound** and **infrasound**. The study of the behavior of sound waves is called **acoustics**.

Ultrasound

Sound composed of **ultrasonic waves** – waves with **frequencies*** above the range of the human ear, i.e. above 20,000 **Hertz***. Ultrasound has a number of uses.

Ultrasound *is used in* ***ultrasound scanning*** *of the human body (it uses* ***echoes*** *– see page 41).*

Bone, fat and muscle all reflect ***ultrasonic waves*** *differently. Reflected waves (****echoes****), e.g. from an unborn baby, are converted into electrical pulses which form an image (****scan****) on a screen.*

Scan of baby in mother's womb at 20 weeks

Infrasound

Sound composed of **infrasonic waves** – waves with **frequencies*** below the range of the human ear, i.e. below 20 **Hertz***. At present infrasound has few technical uses, as it can cause uncomfortable sensations in humans.

Sonic boom

A loud bang heard when a **shock wave** produced by an aircraft moving at **supersonic speed** passes a listener.

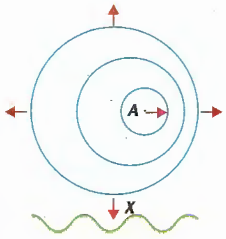

As aircraft (A) travels forward, it creates ***longitudinal waves**** *in air, i.e. areas of high and low pressure (****compressions**** *and* ***rarefactions****).*

Wavefronts* *can "get away" from aircraft and begin to disperse.*

Listener at X will hear waves as sound (a "whoosh" of air – as well as separate sound of engines).

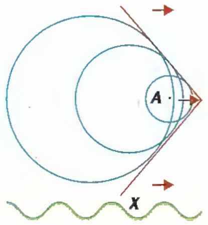

Supersonic *aircraft (A) overtakes its wavefronts while creating more so they overlap.*

*Causes large build up of pressure (****shock wave****) pushed in front of aircraft and unable to "get away". It is like bow wave of ship (if ship moving faster than water waves it creates).*

Listener at X will hear wave as sudden loud ***sonic boom****.*

Behavior of sound waves

Speed of sound

The speed at which sound waves move. It depends on the type and temperature of the **medium*** through which the sound waves travel. The speed of sound waves as they travel through dry air at 0°C is 331m s^{-1}, but this increases if the air temperature increases, or decreases if the air temperature goes down.

Subsonic speed

A speed below the **speed of sound** in the same **medium*** and under the same conditions.

Supersonic speed

A speed above the **speed of sound** in the same **medium*** and under the same conditions.

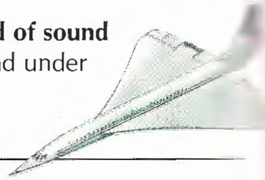

A supersonic passenger jet

---

* **Compressions**, **Frequency**, 35; **Hertz**, 16 (**Frequency**); **Longitudinal waves**, 35; **Medium**, 345; **Rarefactions**, 35; **Wavefront**, 34.

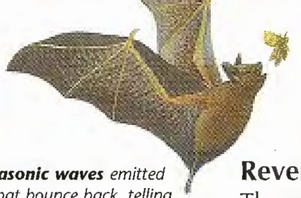

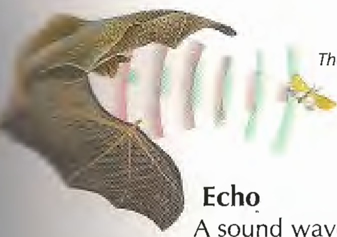

*The **ultrasonic waves** emitted by a bat bounce back, telling it the distance and size of the object. This technique is called **echolocation**.*

Echo

A sound wave which has been reflected off a surface, and is heard after the original sound. Echoes, normally those of **ultrasonic waves**, are often used to locate objects and determine their exact position (by measuring the time the echo takes to return to the source). This technique has a number of names, each normally used in a slightly different context, though the distinctions between them are unclear. **Ultrasound scanning** is one example. Others are **echo-sounding** and **sonar**, both of which have marine connotations (echo-sounding normally refers to using echoes to measure the depth of water below a ship, sonar to using them to detect objects under water). **Echolocation** usually describes the way animals use echoes to find prey or avoid obstacles in the dark.

Sonar (derived from **so**und **na**vigation and **r**anging)

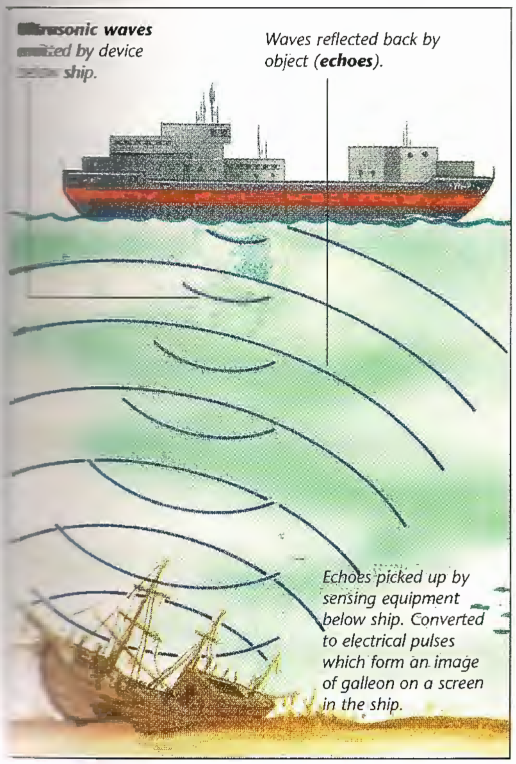

***Ultrasonic waves** emitted by device below ship.*

*Waves reflected back by object (**echoes**).*

Echoes picked up by sensing equipment below ship. Converted to electrical pulses which form an image of galleon on a screen in the ship.

Reverberation

The effect whereby a sound seems to persist for longer than it actually took to produce. It occurs when the time taken for the **echo** to return to the source is so short that the original and reflected waves cannot be distinguished. If the wave is reflected off many surfaces, then the sound is enhanced further.

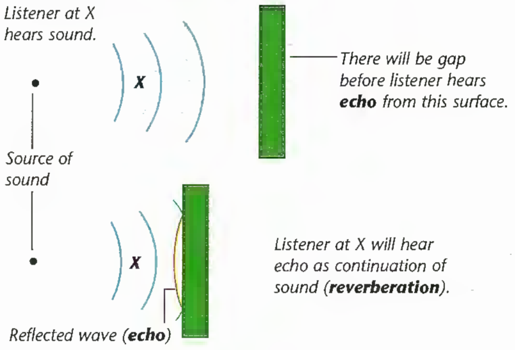

Listener at X hears sound.

Source of sound

*There will be gap before listener hears **echo** from this surface.*

*Listener at X will hear echo as continuation of sound (**reverberation**).*

*Reflected wave (**echo**)*

Doppler effect

The change in **frequency*** of the sound heard when either the listener or the source moves relative to the other. If the distance between them is decreasing, a higher frequency sound is heard than that actually produced. If it is increasing, a lower frequency sound is heard.

Doppler effect

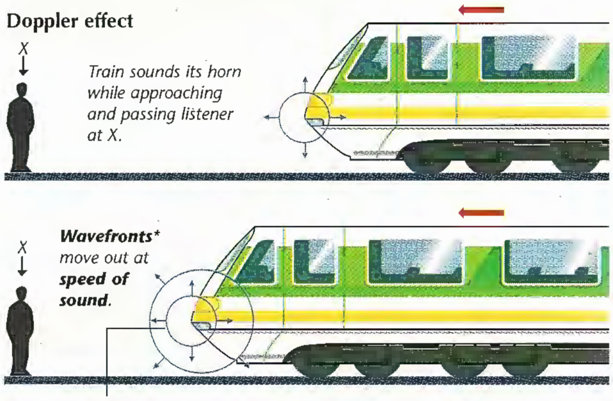

Train sounds its horn while approaching and passing listener at X.

Wavefronts move out at **speed of sound**.*

*Wavefronts closer together here because train moving forward while producing sound waves. Heard at X as sound of higher **frequency***. Lower frequency sound will be heard when train has passed.*

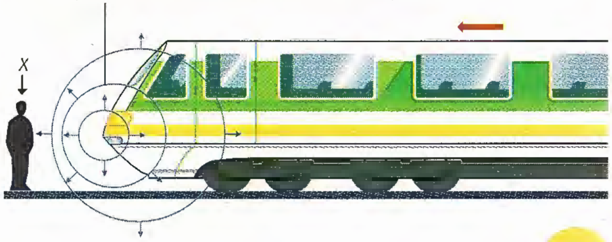

Frequency, 35; **Wavefront**, 34.

PERCEPTION OF SOUND

Sounds heard by the ear can be pleasant or unpleasant. When the waveform of a **sound wave** (see pages 40-41) repeats itself regularly, the sound is usually judged to be pleasant. However, when the waveform is unrepeated and irregular, the sound is thought of as a **noise**. Every sound has a particular **loudness** and **pitch** and many, especially musical sounds, are produced by **stationary waves**.

Loudness

The size of the sensation produced when sound waves fall on the ear. It is subjective, depending on the sensitivity of the ear, but is directly related to the **wave intensity*** of the waves. It is most often measured in **decibels** (**dB**), but also, more accurately, in **phons** (these take into account the fact that the ear is not equally sensitive to sounds of all **frequencies***).

*Aircraft taking off measures 110**dB**.*

Pitch

The perceived **frequency*** of a sound wave, i.e. the frequency heard as sound. A high pitched sound has a high frequency and a low pitched sound has a low frequency.

*The sound of a bird's song is high **pitched**. It has a high **frequency***.*

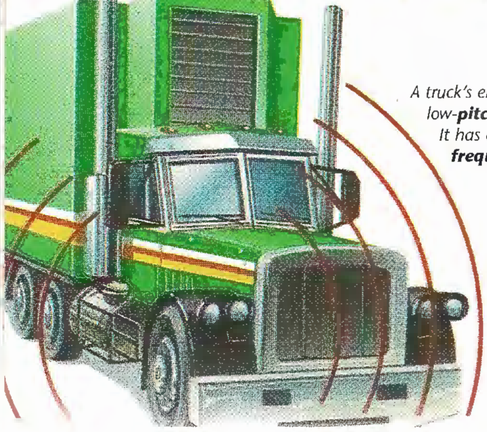

*A truck's engine has a low-**pitched** hum. It has a low **frequency***.*

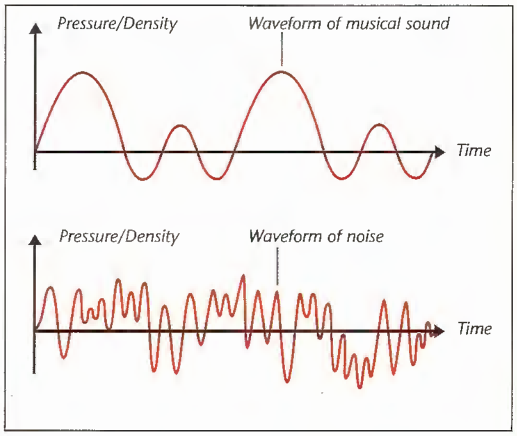
Pressure/Density
Waveform of musical sound
Time
Pressure/Density
Waveform of noise
Time

Beats

The regular variation in **loudness** with time which is heard when two sounds of slightly different **frequency*** are heard together. This is the result of **interference*** between the two waves. The **beat frequency** is equal to the difference in frequency between the two sounds (see diagram below). The closer together the frequency of the sounds, the slower the beats.

Beat frequency

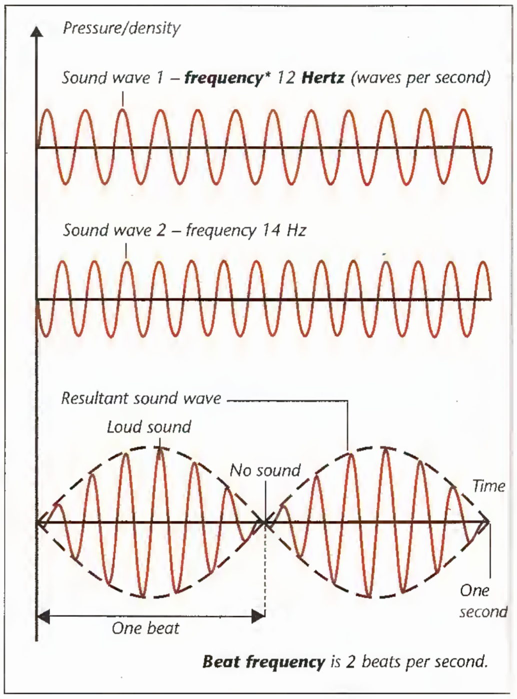
Pressure/density
Sound wave 1 – **frequency*** 12 **Hertz** (waves per second)
Sound wave 2 – frequency 14 Hz
Resultant sound wave
Loud sound
No sound
Time
One second
One beat
Beat frequency is 2 beats per second.

---

* **Frequency**, 35; **Interference**, 38; **Wave intensity**, 35.

Stationary or **standing wave**

A wave that does not appear to move. It is not in fact a true wave, but is instead made up of two waves of the same velocity and **frequency*** continuously moving in opposite directions between two fixed points (most commonly the ends of a plucked string or wire). The repeated crossing of the waves results in **interference*** – when the waves are **in phase***, the resultant **amplitude*** is large, and when they are **out of phase***, it is small or zero. At certain points (the **nodes**), it is always zero. The amplitude and frequency of a stationary wave in a string or wire determines those of the sound waves it produces in the air – the length and tension of the string or wire determine the range of frequencies, and hence the pitch of the sound produced.

Sonometer

*Apparatus used to demonstrate **stationary waves**. When plucked, wire vibrates and sound box amplifies sound caused by vibration.*

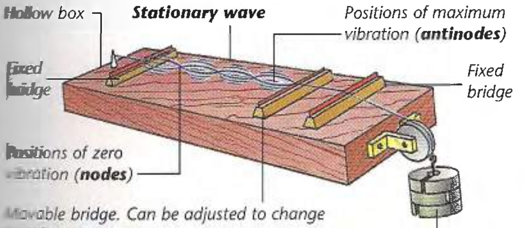
Hollow box
Stationary wave
Positions of maximum vibration (**antinodes**)
Fixed bridge
Fixed bridge
Positions of zero vibration (**nodes**)

*Movable bridge. Can be adjusted to change length of wire and so alter **pitch** of note.*

Weights. Can be adjusted to change tension of wire and so alter pitch of note.

Formation of stationary wave

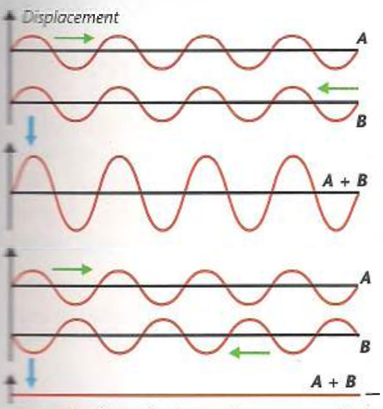

*Waves A and B (same **amplitude***) moving in opposite directions between two fixed points.*

Resultant wave at time t = 0. Amplitude doubled.

Resultant wave at time t = 1. – Amplitude zero.

*At time t = 2, resultant wave has same amplitude as at t = 0, but is transposed, i.e. **crests*** where there were **troughs*** and vice versa.*

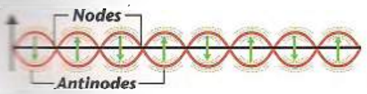

Resultant wave moves rapidly between position held at t = 0 and that held at t = 2. Wave observed appears stationary.

Amplitude, **Crests**, 34; **Frequency**, 35; **In phase**, 38 (**Phase**); **Interference**, 38; **Out of phase**, 38 (**Phase**); **Troughs**, 34.

Musical sounds

All music is based on some kind of **musical scale**. This is a series of **notes** (sounds of specific **pitch**), arranged from low to high pitch with certain **intervals** between them (a musical interval is a spacing in **frequency***, rather than time). The notes are arranged so that pleasant sounds can be obtained. What is regarded as a pleasant sound depends on the culture of the listener.

*Western **musical scale** is based on **diatonic scale** – consists of 8 **notes** (white notes on piano) ranging from lower to upper C.*

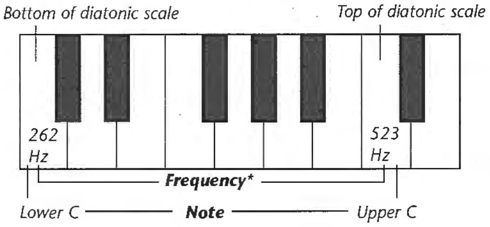

*Black notes have **frequencies*** between those of notes on **diatonic scale**. Together with these, they form **chromatic scale**.*

Modes of vibration

The same note played on different instruments, although recognizable as the same, has a distinct sound quality (**timbre**) characteristic to the instrument. This is because, although the strongest vibration is the same for each note whatever the instrument (its **frequency*** is the **fundamental frequency**), vibrations at other frequencies (**overtones**) are produced at the same time. The set of vibrations specific to an instrument are its modes of vibration.

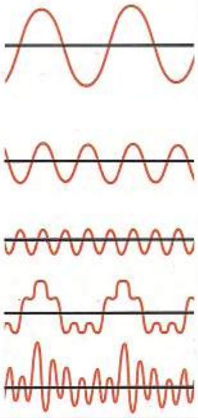

*Lowest **mode of vibration** (**fundamental frequency**) of note on given instrument. If frequencies of **overtones** are simple multiples of fundamental frequency, they are also called **harmonics**.*

1st overtone (2nd harmonic, i.e. frequency doubled. Fundamental frequency is 1st harmonic).

2nd overtone. This is 4th harmonic. In this case there is no 3rd harmonic.

Combined modes of vibration (i.e. all three together). Characteristic waveform of note for this instrument.

Same note played on different instrument may look like this.

ELECTROMAGNETIC WAVES

Electromagnetic waves are **transverse waves***, consisting of oscillating **electric** and **magnetic fields***. They have a wide range of **frequencies***, can travel through most **media***, including vacuums, and, when absorbed, cause a rise in temperature (see **infra-red radiation**). **Radio waves** and some **X-rays** are emitted when free **electrons*** are accelerated or decelerated, e.g. as a result of a collision. All other types occur when molecules change energy states (see page 84) and occur as pulses called **photons** (see **quantum theory**, page 84), rather than a continuous stream. For the wavelengths and frequencies of the different types of waves (the **electromagnetic spectrum**), see the table on page 113.

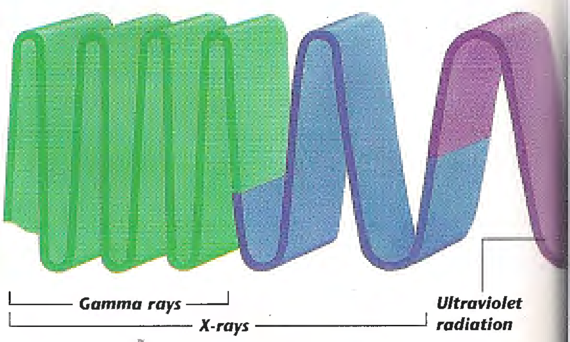

Electromagnetic spectrum** (range of electromagnetic waves) shown above. It is made up of **wavebands**, i.e. particular ranges of **frequencies and **wavelengths*** within which the waves all have the same characteristic properties.*

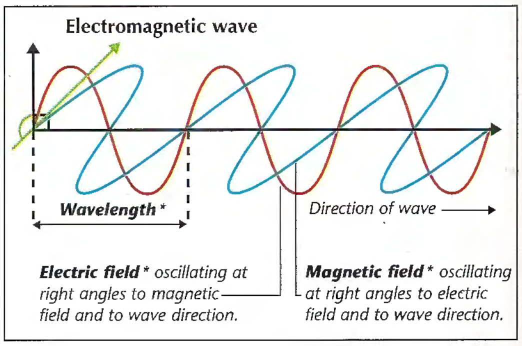
Electromagnetic wave
Wavelength*
Direction of wave
Electric field* oscillating at right angles to magnetic field and to wave direction.
Magnetic field* oscillating at right angles to electric field and to wave direction.

Gamma rays (γ-rays)

Electromagnetic waves emitted by **radioactive*** substances (see also page 86). They are in the same **waveband** and have the same properties as **X-rays**, but are produced in a different way, and are at the top end of the band with regard to energy.

X-rays

Electromagnetic waves which **ionize*** gases they pass through, cause **phosphorescence** and bring about chemical changes on photographic plates. They are produced in **X-ray tubes*** and have many applications.

***X-radiography** produces pictures (**radiographs**) of inside of body. **X-rays** pass through tissue but are absorbed by denser bones, so bones appear opaque.*

X-rays were passed through this hand to project a clear image of the bones onto a photographic plate.

Ultraviolet radiation (UV radiation)

Electromagnetic waves produced, for example, when an electric current is passed through **ionized*** gas between two **electrodes***. They are also emitted by the Sun, but only small quantities reach the Earth's surface. These small quantities are vital to life, playing the key part in **photosynthesis***, but larger amounts are dangerous. Ultraviolet radiation causes **fluorescence**, e.g. when produced in **fluorescent tubes***, and also a variety of chemical reactions, e.g. tanning of the skin.

Phosphorescence

A phenomenon shown by certain substances (**phosphors**) when they are hit by short **wavelength*** electromagnetic waves, e.g. **gamma rays** or **X-rays**. The phosphors absorb the waves and emit visible light, i.e. waves of longer wavelength. This emission may continue after the gamma or X-rays have stopped. If it only occurs briefly afterwards in rapid flashes, these are called **scintillations** (see also **scintillation counter**, page 90).

---

* **Electric field**, 58; **Electrode**, 66; **Electrons**, 83; **Fluorescent tube**, 80 (**Discharge tube**); **Frequency**, 35; **Ionization**, 88; **Magnetic field**, 72; **Medium**, 345; **Photosynthesis**, 254; **Radioactivity**, 86; **Transverse waves**, **Wavelength**, 34; **X-ray tube**, 80.

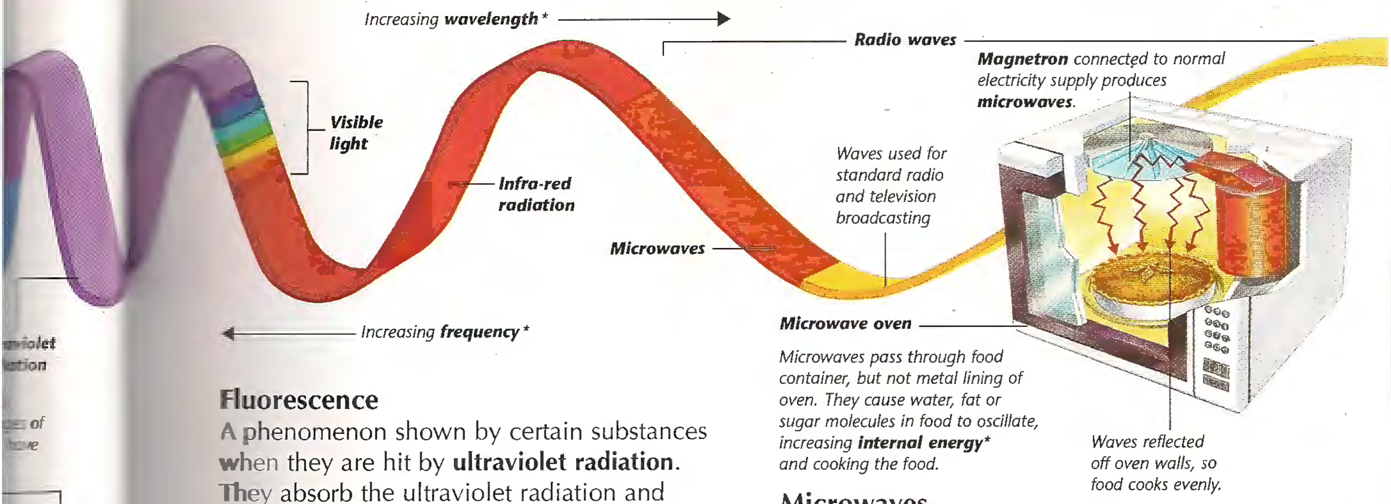
Increasing **wavelength** *
Radio waves
Magnetron connected to normal electricity supply produces **microwaves**.
Visible light
Infra-red radiation
Waves used for standard radio and television broadcasting
Microwaves
Increasing **frequency** *
Microwave oven
Microwaves pass through food container, but not metal lining of oven. They cause water, fat or sugar molecules in food to oscillate, increasing **internal energy*** and cooking the food.
Waves reflected off oven walls, so food cooks evenly.

Fluorescence

A phenomenon shown by certain substances when they are hit by **ultraviolet radiation**. They absorb the ultraviolet radiation and emit **visible light**, i.e. light waves of a longer **wavelength***. This emission stops as soon as the ultraviolet radiation stops.

Visible light

Electromagnetic waves which the eye can detect. They are produced by the Sun, by **discharge tubes*** and by any substance heated until it glows (emission of light due to heating is called **incandescence**). They cause chemical changes, e.g. on photographic film, and the different **wavelengths*** in the **waveband** are seen as different colors (see page 54).

Infra-red radiation (IR radiation)

The electromagnetic waves most commonly produced by hot objects and therefore those which are most frequently the cause of temperature rises (see introduction and **radiation**, page 29). They can be used to form **thermal images** on special infra-red sensitive film, which is exposed by heat, rather than light.

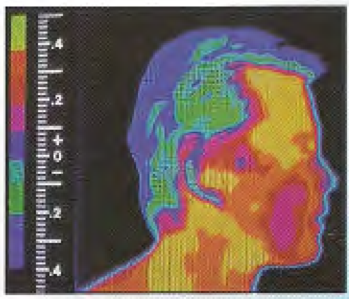

Thermal image *of a man's head*

Each different color represents a temperature difference of 0.1°C (see scale on left of picture).

The blue areas of the head are colder, and the yellow areas are hotter.

Microwaves

Very short **radiowaves** used in **radar** (**ra**dio **d**etection **a**nd **r**anging) to determine the position of an object by the time it takes for a reflected wave to return to the source (see also **sonar**, page 41 (**Echo**)). **Microwave ovens** use microwaves to cook food rapidly.

Radio waves

Electromagnetic waves produced when free **electrons*** in radio antennae are made to oscillate (and are hence accelerated) by an **electric field***. The fact that the frequency of the oscillations is imposed by the field means that the waves occur as a regular stream, rather than randomly.

How radio waves are used to communicate over long distances

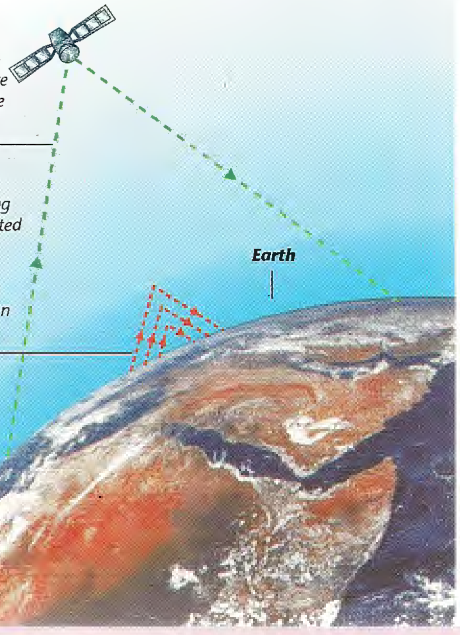
Radio waves with short **wavelengths*** can penetrate ionosphere, hence are used to communicate over long distances via satellites.
Radio waves with long **wavelengths*** reflected within ionosphere, hence are used to transmit information from place to place on same area of Earth's surface.
Earth
Ionosphere (region of **ionized*** gas around the Earth)

---

* **Discharge tube**, 80; **Electric field**, 58; **Electrons**, 83; **Frequency**, 35; **Internal energy**, 9; **Ionization**, 88; **Nuclear power station**, 94; **Wavelength**, 34.

LIGHT

Light consists of **electromagnetic waves*** of particular **frequencies*** and **wavelengths*** (see pages 44-45), but is commonly referred to and diagrammatically represented as **rays**. Such a ray is actually a line (arrow) which indicates the path taken by the light waves, i.e. the direction in which the energy is being carried.

Shadow

An area which light rays cannot reach due to an obstacle in their path. If the rays come from a point they are stopped by the obstacle, creating a complete shadow called an **umbra**.

Casting an umbra

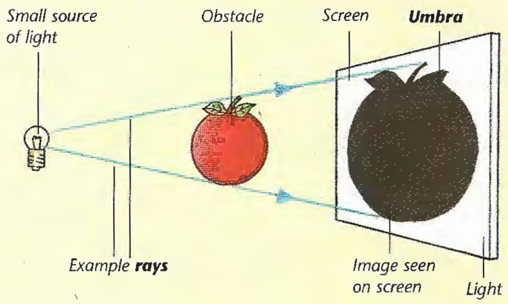
Small source of light
Obstacle
Screen
Umbra
Example **rays**
Image seen on screen
Light

If the light rays come from an extended source, a semi-shadow area called a **penumbra** is formed around the umbra.

Casting a penumbra

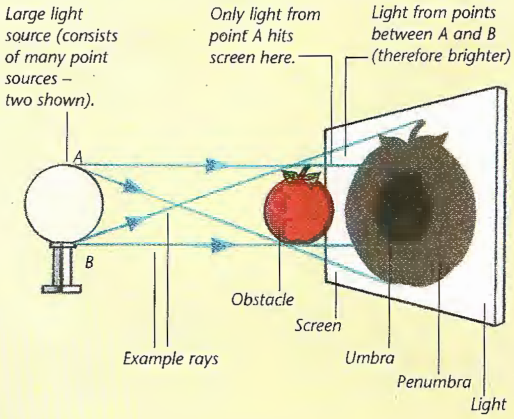
Large light source (consists of many point sources – two shown).
Only light from point A hits screen here.
Light from points between A and B (therefore brighter)
A
B
Obstacle
Screen
Example rays
Umbra
Penumbra
Light

Eclipse

The total or partial "blocking off" of light from a source. This occurs when an object casts a **shadow** by passing between the source and an observer. A **solar eclipse** is seen from the Earth when the Moon passes between the Earth and the Sun, and a **lunar eclipse** is seen when the Earth is between the Sun and the Moon.

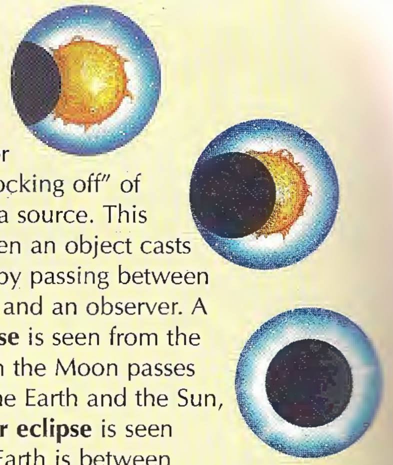

Moon blocking out Sun's light in an annular solar eclipse

Solar eclipse

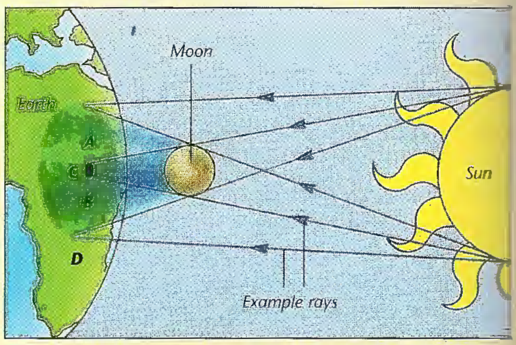
Moon
Earth
A
C
B
D
Sun
Example rays

The eclipse, as seen from positions A, B, C and D.

***Partial eclipse** seen from most places within circle, e.g. positions A and B. Crescent-shaped area of Sun still visible.*

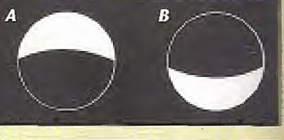

***Total eclipse** seen from position C. Sun completely obscured.*

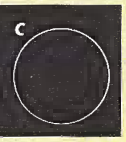

*No **eclipse** seen at any place outside circle.*

Annular eclipse

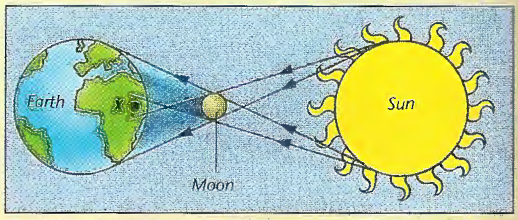
Earth
X
Sun
Moon

The eclipse, as seen from position X.

*An **annular eclipse** is a special **eclipse** which consists of a bright ring around a black area. It occurs when the Moon, Earth and Sun are particular distances apart.*

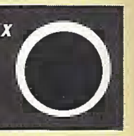

---

* **Electromagnetic waves**, 44; **Frequency**, 35; **Wavelength**, 34.

REFLECTION OF LIGHT

Reflection is the change in direction of a wave when it bounces off a boundary (see page 36). Mirrors are usually used to show the reflection of light (see below and also pages 48-49). It must be noted that when an object and its image are drawn in mirror (and **lens***) diagrams, the object is assumed to be producing light rays itself. In fact the rays come from a source, e.g. the Sun, and are reflected off the object.

Laws of reflection of light

1. The **reflected ray** lies in the same plane as the **incident ray** and the **normal** at the **point of incidence**.
2. The **angle of incidence** (**i**) = the **angle of reflection** (**r**).

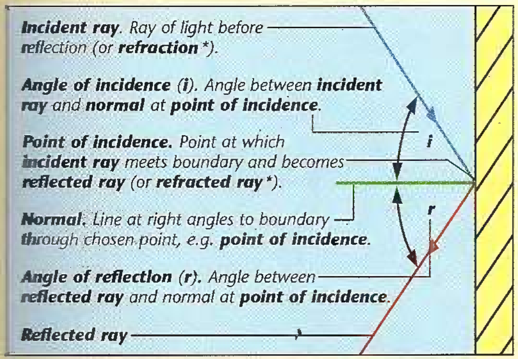
Incident ray. Ray of light before reflection (or **refraction** *).
Angle of incidence (i). Angle between **incident ray** and **normal** at **point of incidence**.
Point of incidence. Point at which **incident ray** meets boundary and becomes **reflected ray** (or **refracted ray** *).
Normal. Line at right angles to boundary through chosen point, e.g. **point of incidence**.
Angle of reflection (r). Angle between **reflected ray** and normal at **point of incidence**.
Reflected ray
i
r

Regular reflection

The reflection of parallel **incident rays** (see above) off a flat surface such that all the **reflected rays** are also parallel. This occurs when surfaces are very smooth, e.g. highly polished surfaces such as mirrors.

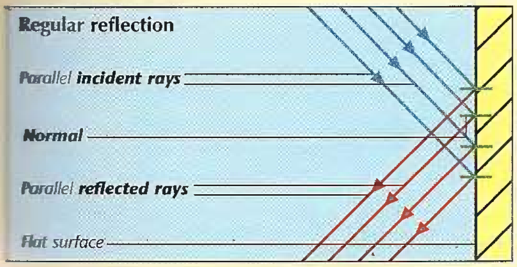
Regular reflection
Parallel **incident rays**
Normal
Parallel **reflected rays**
Flat surface

Diffuse reflection

The reflection of parallel **incident rays** (see left) off a rough surface such that the **reflected rays** travel in different directions and the light is scattered. This is the most common type of reflection as most surfaces are irregular when considered on a scale comparable to that of the **wavelength*** of light (see page 113).

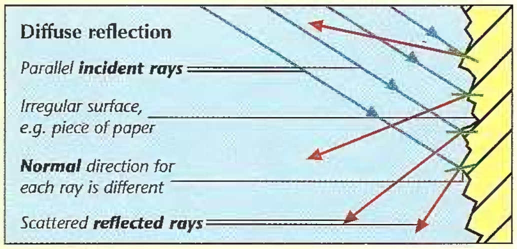
Diffuse reflection
Parallel **incident rays**
Irregular surface, e.g. piece of paper
Normal direction for each ray is different
Scattered **reflected rays**

Plane mirror

A mirror with a flat surface (see also **curved mirrors**, pages 48-49). The image it forms is the same size as the object, the same distance behind ("inside") the mirror as the object is in front, and **laterally inverted** (the left and right sides have swapped around).

Reflection in plane mirror

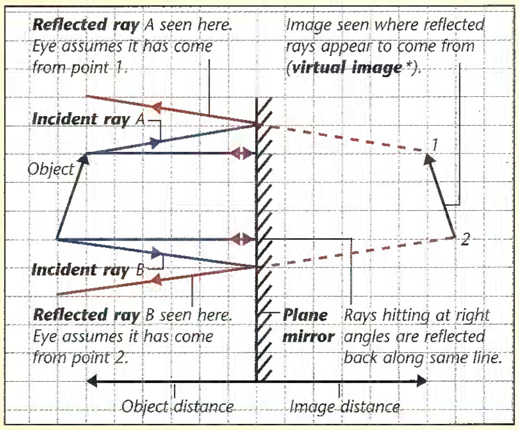
Reflected ray A seen here. Eye assumes it has come from point 1.
Image seen where reflected rays appear to come from (**virtual image** *).
Incident ray A
1
Object
2
Incident ray B
Reflected ray B seen here. Eye assumes it has come from point 2.
Plane mirror
Rays hitting at right angles are reflected back along same line.
Object distance
Image distance

Parallax

The apparent displacement of an observed object due to the difference between two points of view. For example, an object which is observed first with the left eye, and then with the right eye, appears to have moved. The first point of view is the left eye, and the second is the right eye. (See also **parallax error**, page 102.)

---

* **Lenses**, 52; **Refracted ray**, **Refraction**, 50; **Virtual image**, 49 (**Image**); **Wavelength**, 34.

Reflection of light (continued)

Light rays are reflected from curved surfaces, as from flat surfaces, according to the **laws of reflection of light** (see page 47). The images formed by reflection from **curved mirrors** are particularly easily observed. There are two types of curved mirror – **concave** and **convex mirrors**. For all diagrams showing reflection of light, the object is assumed to be the source of the light (see **reflection of light**, page 47) and certain points (see below), together with known facts about light rays passing through them, are used to construct the paths of the reflected rays.

Points used to construct paths of reflected rays
(See also page 52.)

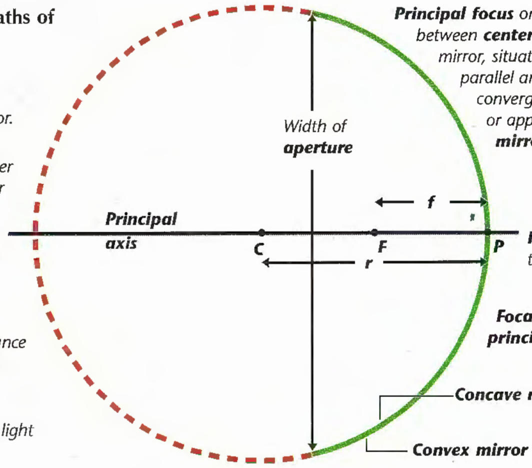
Width of
aperture
Principal axis
C
F
P
f
r

***Pole* (*P*).** *Center of curved mirror.*

***Center of curvature* (*C*).** *Center of sphere of which curved mirror is part. Any light ray passing through it (**concave mirrors**) or heading directly for it (**convex mirrors**) reflected back on itself.*

***Radius of curvature* (*r*).** *Distance between **center of curvature** and **pole**.*

***Aperture*.** *Area through which light passes to hit mirror.*

Principal focus *or* ***focal point* (*F*).** *Special point half-way between **center of curvature** and **pole** of curved mirror, situated on **principal axis**. All rays traveling parallel and close to axis reflected so that they converge at principal focus (**concave mirrors**) or appear to have diverged from it (**convex mirrors**).*

***Principal axis*.** *Straight line running through **center of curvature** and **pole**.*

***Focal length* (*f*).** *Distance between **principal focus** and **pole**.*

Concave mirror *reflects from this side.*

Convex mirror *reflects from this side.*

Concave or **converging mirror**

A mirror with a reflecting surface which curves inward (part of the inside of a sphere). When light rays parallel to the **principal axis** fall on such a mirror, they are reflected so that they converge at the **principal focus** in front of the mirror. The size, position and type of **image** formed depends on how far the object is from the mirror.

Convex or **diverging mirror**

A mirror with a reflecting surface which curves outwards (part of the outside of a sphere). When light rays parallel to the **principal axis** fall on such a mirror, they are reflected so that they appear to diverge from the **principal focus** behind ("inside") the mirror. The **images** formed are always upright and reduced, and **virtual images** (see **image**).

Concave mirror

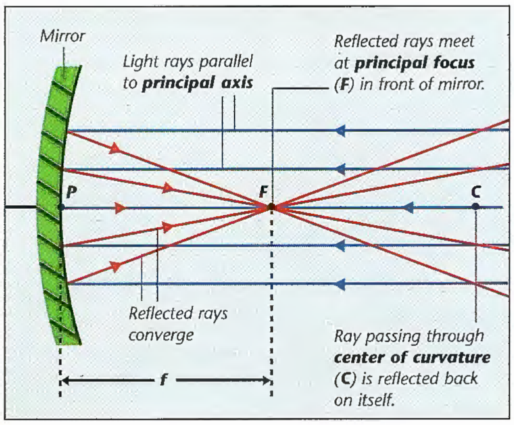
Mirror
Light rays parallel to **principal axis**
Reflected rays meet at **principal focus** (**F**) in front of mirror.
P
F
C
Reflected rays converge
Ray passing through **center of curvature** (**C**) is reflected back on itself.
f

Convex mirror

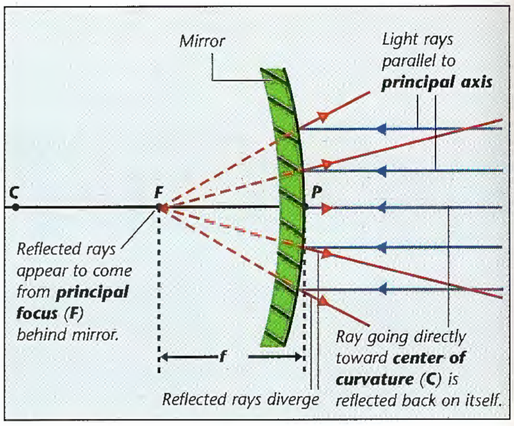
Mirror
Light rays parallel to **principal axis**
C
F
P
Reflected rays appear to come from **principal focus** (**F**) behind mirror.
f
Reflected rays diverge
Ray going directly toward **center of curvature** (**C**) is reflected back on itself.

Image

A view of an object as seen in a mirror. Just as an object is only seen because of light rays coming from it (see **reflection of light**, page 47), so too an image is seen where reflected rays (originally from the object) actually diverge from (**real image**), or appear to diverge from (**virtual image**).

Example of image formed by concave mirror *(object placed at* ***center of curvature****)*

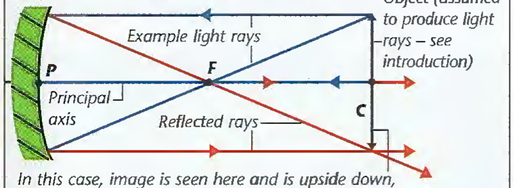
Object (assumed to produce light rays – see introduction)
Example light rays
P
F
Principal axis
Reflected rays
C

In this case, image is seen here and is upside down, same size as object and a ***real image*** *(reflected rays meet at points along it and then diverge).*

Image formed by convex mirrors

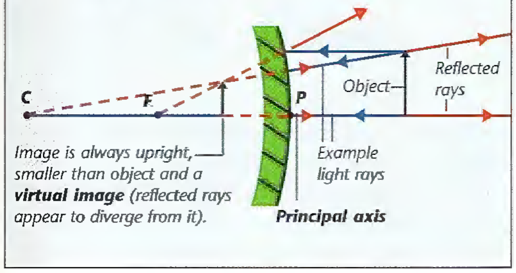
C
F
P
Object
Reflected rays
Image is always upright, smaller than object and a virtual image (reflected rays appear to diverge from it).
Example light rays
Principal axis

Mirror or lens formula

Gives the relationship between the distance of an object from the center of a curved mirror or **lens***, the distance of its **image** from the same point and the **focal length** of the mirror or lens. An image may be formed either side of a mirror or lens, so a **sign convention*** is used to give position.

Mirror formula: $\frac{1}{f} = \frac{1}{v} + \frac{1}{u}$

where f = **focal length**;
v = **image** distance (from **pole**);
u = object distance (from pole).

Real is positive sign convention for mirrors and lenses

1. *All distances are measured from the mirror as origin.*
2. *Distances of objects and* ***real images*** *are positive.*
3. *Distances of* ***virtual images*** *are negative.*
4. *The* ***focal lengths*** *of* ***convex mirrors*** *and* ***lenses**** *are positive. The focal lengths of* ***concave mirrors*** *and lenses are negative.*

Linear magnification

The ratio of the height of the image formed by a mirror or **lens*** to the object height.

$$\textit{Linear magnification} = \frac{\textbf{height of image}}{\textbf{height of object}}$$

Example of linear magnification

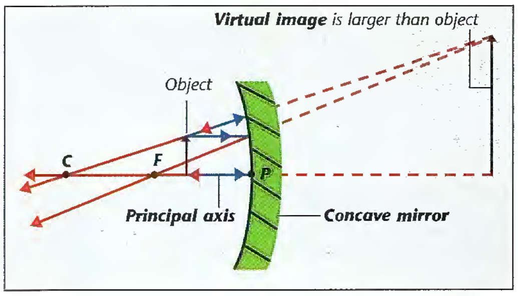
Virtual image is larger than object
Object
C
F
P
Principal axis
Concave mirror

Principle of reversibility of light

States that, for a ray of light on a given path due to reflection, **refraction*** or **diffraction***, a ray of light in the opposite direction in the same conditions will follow the same path. Light rays parallel to the **principal axis**, for example, are reflected by a **concave mirror** to meet at the **principal focus**. If a point source of light is placed at the principal focus, the rays are reflected parallel to the axis.

Spherical aberration

An effect seen when rays parallel to the **principal axis** (and different distances from it), hit a curved mirror and are reflected so that they intersect at different points along the axis, forming a **caustic curve**. The larger the **aperture**, the more this effect is seen. It is also seen in **lenses*** with large apertures.

Spherical aberration

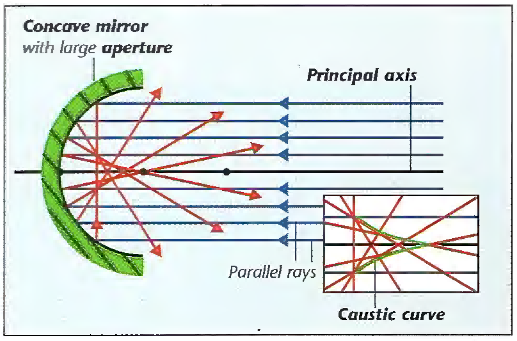
Concave mirror with large aperture
Principal axis
Parallel rays
Caustic curve

---

* **Diffraction**, 37; **Lens**, 52; **Refraction**, 37; **Sign convention**, 11.

REFRACTION OF LIGHT

Refraction is the change in direction of any wave as a result of its velocity changing when it moves from one **medium*** into another (see also page 37). When light rays (see page 46) move into a new medium, they are refracted according to the **laws of refraction of light**. The direction in which they are refracted depends on whether they move into a denser or less dense medium and are consequently slowed down or speeded up (see diagram below).

***Refraction** makes the end of this straw appear bent in the drink.*

Refraction at boundary between two media*

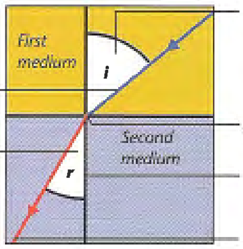

Incident ray**. Ray of light before refraction (or **reflection)*

***Angle of refraction** (**r**). Angle between **refracted ray** and **normal** at **point of incidence**.*

*If second medium denser, ray slows down and is refracted toward **normal** as here. If less dense, ray speeds up and is refracted away from normal.*

***Angle of incidence** (**i**). Angle between **incident ray** and **normal** at **point of incidence**.*

Point of incidence**. Point at which **incident ray** meets boundary and becomes **refracted ray** (or **reflected ray).*

***Normal**. Line at right angles to boundary through chosen point, e.g. **point of incidence**.*

Refracted ray

Laws of refraction of light

1. The **refracted ray** lies in the same plane as the **incident ray** and **normal** at the **point of incidence**.

2. (**Snell's law**). The ratio of the **sine*** of the **angle of incidence** to the sine of the **angle of refraction** is a constant for two given **media***.

This constant is the **refractive index** (**n** – see page 37). When referring to light, this is also known as the **optical density** and, as with refractive index in other cases, can also be calculated by dividing the velocity of light in one medium by its velocity in the second medium. See also **apparent depth** picture.

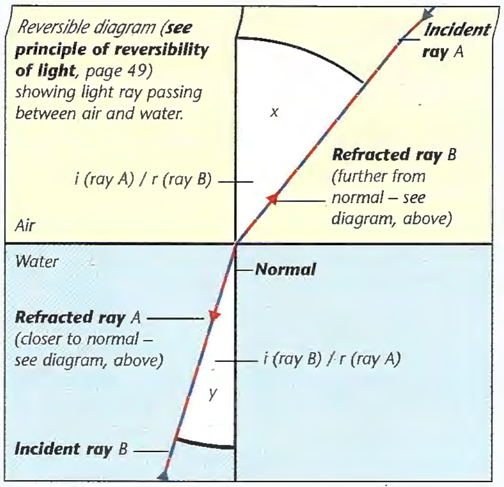

*Reversible diagram (**see principle of reversibility of light**, page 49) showing light ray passing between air and water.*

Incident ray A
x
Refracted ray B (further from normal – see diagram, above)
i (ray A) / r (ray B)
Air
Water
Normal
Refracted ray A (closer to normal – see diagram, above)
i (ray B) / r (ray A)
y
Incident ray B

*For either direction, **refractive index*** of second **medium*** relative to first is written $_1n_2$.*

*According to **Snell's law**:*

$$_1n_2 = \frac{\sin i}{\sin r}$$

Media may also be specified by subscript letters – $_an_w$ means refractive index of water relative to air and $_wn_a$ means that of air relative to water.

$$_an_w = \frac{\sin i\ (\text{ray A})}{\sin r\ (\text{ray B})} = \frac{\sin x}{\sin y}$$

$$_wn_a = \frac{\sin i\ (\text{ray B})}{\sin r\ (\text{ray B})} = \frac{\sin y}{\sin x}$$

$$\text{Thus } {}_an_w = \frac{1}{_wn_a}$$

*If no subscripts are given, the value is the **absolute refractive index***.*

* **Absolute refractive index**, 37 (**Refractive index**); **Medium**, 345; **Reflected ray**, **Reflection**, 47; **Sine**, 345.

Apparent depth

The position at which an object in one **medium*** appears to be when viewed from another medium. The brain assumes the light rays have traveled in a straight line, but in fact they have changed direction as a result of refraction. Hence the object is not actually where it appears to be.

Apparent depth

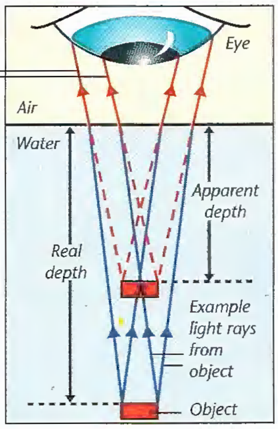
Rays refracted when they leave water
Brain assumes light rays have taken straight paths (red dotted lines), so object seen at X.
Eye
Air
Water
Apparent depth
Real depth
X
Example light rays from object
Object

*Real depth and apparent depth can also be used to calculate **refractive index***:*

$$ {}_a n_w = \frac{\text{real depth}}{\text{apparent depth}} $$

$$ \left({}_w n_a = \frac{1}{{}_a n_w} \right) $$

Critical angle (c)

The particular **angle of incidence** of a ray hitting a less dense **medium*** which results in it being refracted at 90° to the **normal**. This means that the refracted ray (**critical ray**) travels along the boundary, and does not enter the second medium.

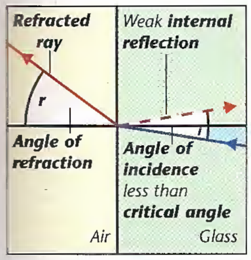
Refracted ray
r
Angle of refraction
Weak internal reflection
Angle of incidence less than critical angle
i
Air
Glass

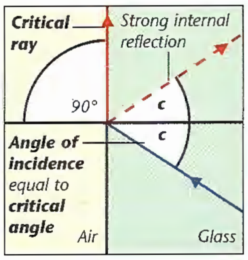
Critical ray
90°
Strong internal reflection
c
c
Angle of incidence equal to critical angle
Air
Glass

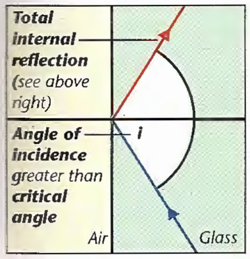
Total internal reflection (see above right)
Angle of incidence greater than critical angle
i
Air
Glass

Critical angle** can also be used to calculate **refractive index:*

$$ {}_g n_a = \sin c $$

$$ \left({}_a n_g = \frac{1}{{}_g n_a} \right) $$

(Note: sine of 90° is 1.)

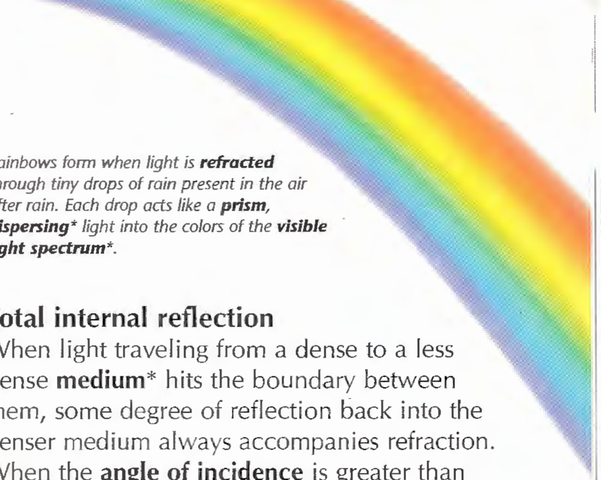

*Rainbows form when light is **refracted** through tiny drops of rain present in the air after rain. Each drop acts like a **prism**, **dispersing*** light into the colors of the **visible light spectrum***.*

Total internal reflection

When light traveling from a dense to a less dense **medium*** hits the boundary between them, some degree of reflection back into the denser medium always accompanies refraction. When the **angle of incidence** is greater than the **critical angle**, total internal reflection occurs, i.e. all the light is internally reflected.

***Optical fibers** transmit light by **total internal reflection**. Bundles of such fibers have a number of uses, e.g. in communications and in medicine (e.g. in **endoscopes**, used by doctors to see inside the body).*

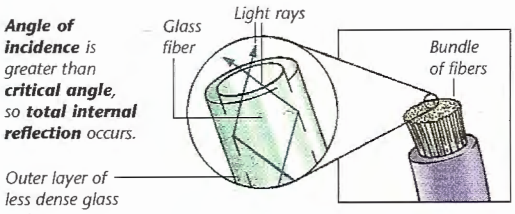
Angle of incidence is greater than critical angle, so total internal reflection occurs.
Glass fiber
Light rays
Bundle of fibers
Outer layer of less dense glass

Prism

A transparent solid which has two plane refracting surfaces at an angle to each other. Prisms are used to produce **dispersion*** and change the path of light by refraction and **total internal reflection.**

Prism refracting light ray

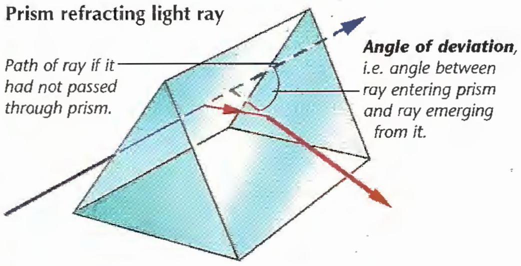
Path of ray if it had not passed through prism.
Angle of deviation, i.e. angle between ray entering prism and ray emerging from it.

Prism causing total internal reflection

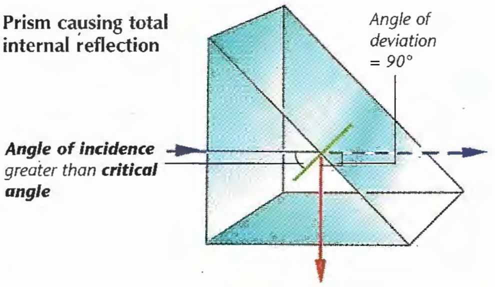
Angle of deviation = 90°
Angle of incidence greater than critical angle

---

* **Dispersion**, 54 (**Color**); **Medium**, 345; **Refractive index**, 37; **Visible light spectrum**, 54.

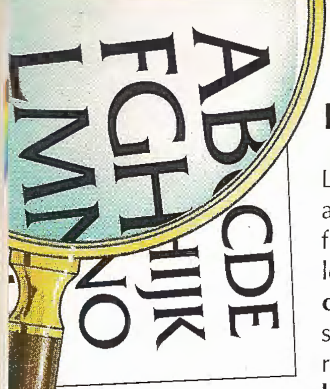

A magnifying glass is a ***converging lens*** used so that objects are made to look bigger than they really are.

Refraction of light (continued)

Light rays are refracted at curved surfaces, e.g. **lenses**, as at flat surfaces, according to the **laws of refraction of light** (see page 50). Unlike with flat surfaces, though, images are formed. There are two basic types of lens, **concave** and **convex lenses**, which can act as **diverging** or **converging lenses** depending on their **refractive index*** relative to the surrounding **medium***. For all diagrams showing image production by refraction, the object is assumed to be the light source (see **reflection of light**, page 47), and certain points (see below), together with known facts about light rays passing through them, are used to construct the paths of the refracted rays. The positions of objects and images can be determined using the **mirror (lens) formula***.

Points used to construct paths of refracted rays
(See also page 48.)

*All the lenses shown are considered to be thin lenses (i.e. the thickness of the lens is small compared to the **focal length**). Although light rays bend both on entering the lens and emerging from it, they are drawn as bending only once, at a vertical line running through **optical center** of the lens.*

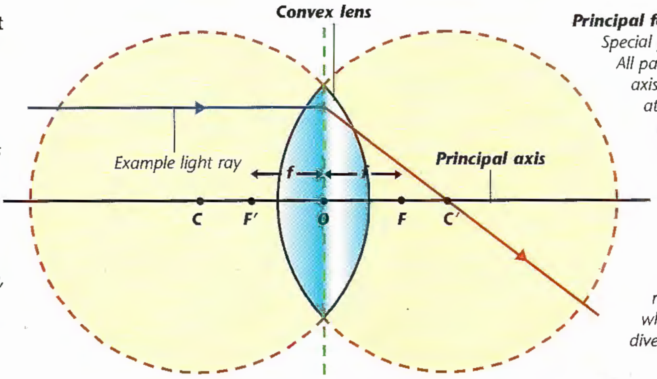

Convex lens
Example light ray
f
f
Principal axis
C
F′
O
F
C′

***Principal focus** or **focal point**. Special point on **principal axis**. All parallel rays traveling close to axis refracted so they converge at principal focus (**converging lens**) or appear to diverge from it (**diverging lens**). Since light can enter lens from either side, there are two principal foci – notation **F** is always given to principal focus at which rays converge or from which they appear to diverge (the other one is **F′**).*

***Optical center** (**O**). Center of lens. Light rays passing through it do not change direction.*

***Center of curvature**. Center of sphere of which lens surface is part. Since lens has two surfaces, there are two centers of curvature – notation **C** is always given to center of curvature on side of incident light ray (the other one is **C′**).*

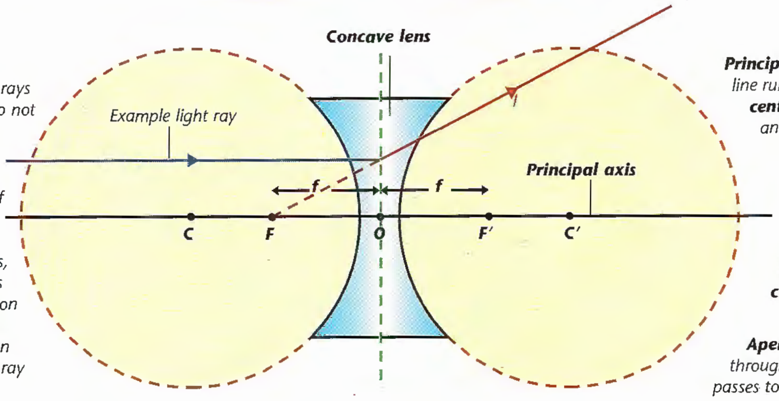

Concave lens
Example light ray
Principal axis
f
f
C
F
O
F′
C′

***Principal axis**. Straight line running through **centers of curvature** and **optical center**.*

***Focal length** (**f**). Distance between either **principal focus** and **optical center**.*

***Aperture**. Area through which light passes to hit lens.*

Converging lens

A lens which causes parallel rays falling on it to converge on the **principal focus** on the other side of the lens. Both **concave** and **convex lenses** can act as **converging lenses**, depending on the **refractive index*** of the lens relative to the surrounding **medium***. A glass convex lens in air acts as a converging lens, as shown in the diagram on the right.

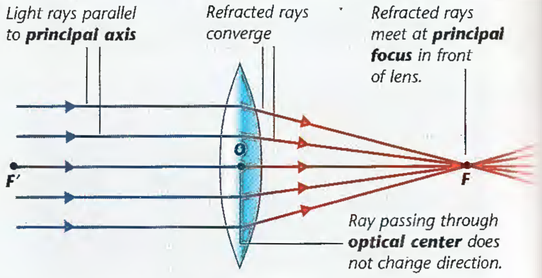

Light rays parallel to principal axis
Refracted rays converge
Refracted rays meet at principal focus in front of lens.
O
F′
F
Ray passing through optical center does not change direction.

*A glass convex lens in a **medium*** denser than glass acts as a **diverging lens**.*

* **Medium**, 345; **Mirror formula**, 49; **Refractive index**, 37.

Power (P)

A measure of the ability of a lens to converge or diverge light rays, given in **diopters** (when **focal length** is measured in meters). The shorter the focal length, the more powerful the lens.

$P = \frac{1}{f}$ where P = **power** of lens; f = **focal length**.

Convex lens

A lens with at least one surface curving outwards. A lens with one surface curving inwards and one outwards is convex if its middle is thicker than its outer edges (it is a **convex meniscus**). A glass convex lens in air acts as a **converging lens**. The size, position and type of image it forms (**real*** or **virtual***) depends on how far it is from the object.

Types of convex lens

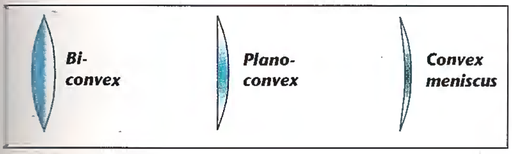
Bi-convex
Plano-convex
Convex meniscus

Convex lens

*Example of image formed by glass **convex lens** in air (object placed between **principal focus** and **optical center**).*

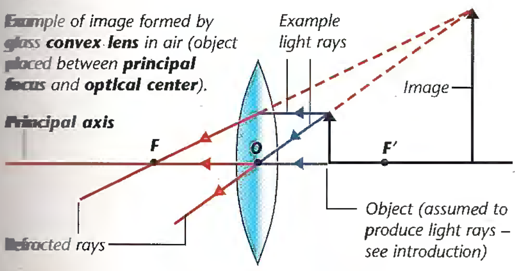
Example light rays
Image
Principal axis
F
O
F′
Object (assumed to produce light rays – see introduction)
Refracted rays

*In this case, image is seen behind the object and is upright, larger than object and a **virtual image***.*

Binoculars use lenses to magnify objects.

Concave lens

A lens which has at least one surface curving inwards. A lens with one surface curving inwards and one outwards is concave if its middle is thinner than its outer edges (it is a **concave meniscus**). A glass concave lens in air acts as a **diverging lens**. The position of the object in relation to the lens may vary, but the image is always of the same type.

Types of concave lens

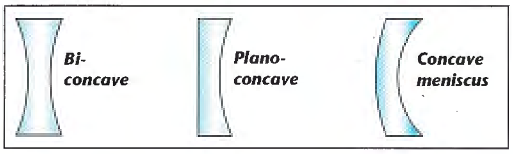
Bi-concave
Plano-concave
Concave meniscus

Concave lens

*Image formed by glass **concave lens** in air*

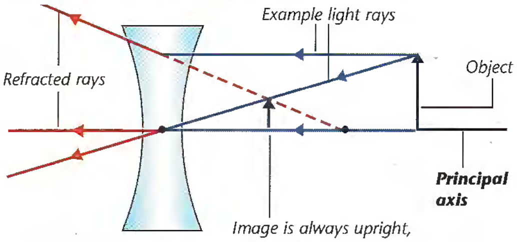
Example light rays
Object
Refracted rays
Principal axis

*Image is always upright, smaller than object and a **virtual image***.*

Diverging lens

A lens which causes parallel rays falling on it to diverge so that they appear to have come from the **principal focus** on the same side as the rays enter. Both **concave** and **convex lenses** can act as diverging lenses, depending on the **refractive index*** of the lens relative to the surrounding **medium***. A glass concave lens in air acts as a diverging lens, as shown in the diagram on the right.

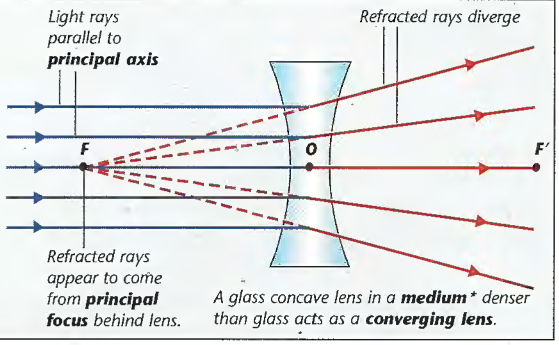
Light rays parallel to principal axis
Refracted rays diverge
F
O
F′
Refracted rays appear to come from principal focus behind lens.

*A glass concave lens in a **medium*** denser than glass acts as a **converging lens**.*

---

*Medium, 345; **Real image**, 49 (**Image**); **Refractive index**, 37; **Virtual image**, 49 (**Image**).

OPTICAL INSTRUMENTS

An **optical instrument** is one which acts on light, using one or more **lenses*** or **curved mirrors*** to produce a required type of image. Listed below are some of the more common optical instruments.

Camera

An optical instrument that is used to form and record an image of an object on film. The image is inverted and a **real image***.

Camera (reflex)

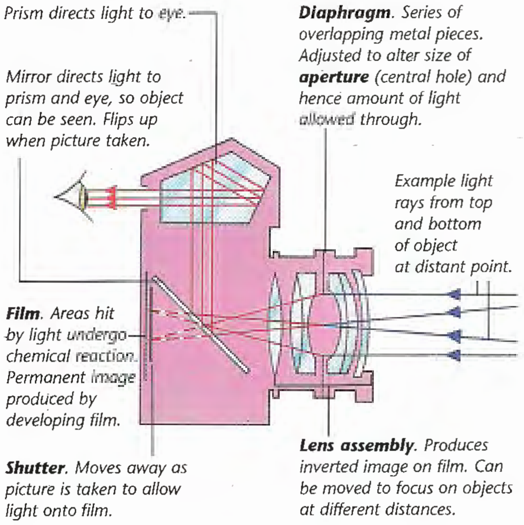

Prism directs light to eye.
Mirror directs light to prism and eye, so object can be seen. Flips up when picture taken.
***Film**. Areas hit by light undergo chemical reaction. Permanent image produced by developing film.*
***Shutter**. Moves away as picture is taken to allow light onto film.*
***Diaphragm**. Series of overlapping metal pieces. Adjusted to alter size of **aperture** (central hole) and hence amount of light allowed through.*
Example light rays from top and bottom of object at distant point.
***Lens assembly**. Produces inverted image on film. Can be moved to focus on objects at different distances.*

As before (see page 52), refraction by lenses shown as one change of direction only, this time on line through optical center of whole lens assembly.

Microscope

An optical instrument which magnifies very small objects. If it has only one **lens***, it is a **simple microscope** or **magnifying glass**. If it has more, it is a **compound microscope**.

Compound microscope

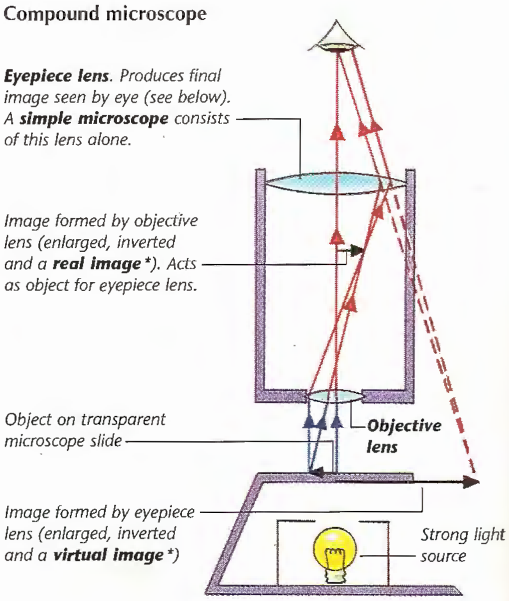

***Eyepiece lens**. Produces final image seen by eye (see below). A **simple microscope** consists of this lens alone.*
*Image formed by objective lens (enlarged, inverted and a **real image***). Acts as object for eyepiece lens.*
Object on transparent microscope slide
Objective lens
*Image formed by eyepiece lens (enlarged, inverted and a **virtual image***)*
Strong light source

Color

When all the different **wavelengths*** of **visible light** (see page 45) fall on the eye at the same time, white light is seen. However, white light can also undergo **dispersion**, whereby it is split into the **visible light spectrum** (i.e. its different wavelengths) by **refraction***. This may occur accidentally (see **chromatic aberration**), or it may be produced on purpose, e.g. with a **spectrometer**.

Inside a spectrometer

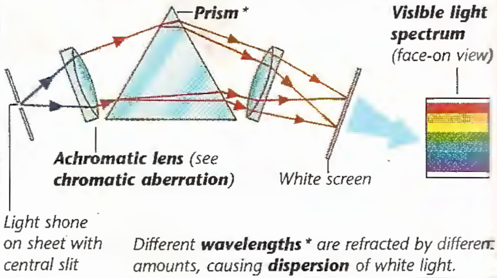

Prism*
***Visible light spectrum** (face-on view)*
***Achromatic lens** (see **chromatic aberration**)*
White screen
Light shone on sheet with central slit

*Different **wavelengths*** are refracted by differen amounts, causing **dispersion** of white light.*

Visible light spectrum

A display of the colors that make up a beam of white light. Each color band represents a very small range of **wavelengths*** – see **visible light**, page 45.

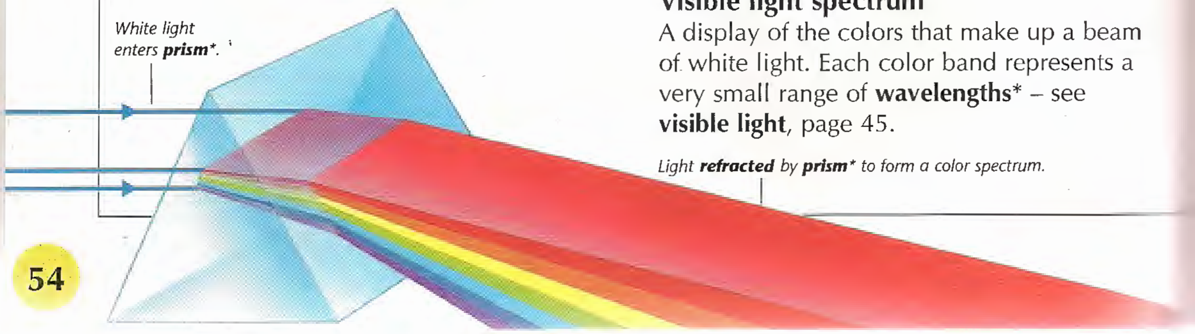

*White light enters **prism***.*

*Light **refracted** by **prism*** to form a color spectrum.*

Slide projector

An optical instrument which produces a magnified image of a slide.

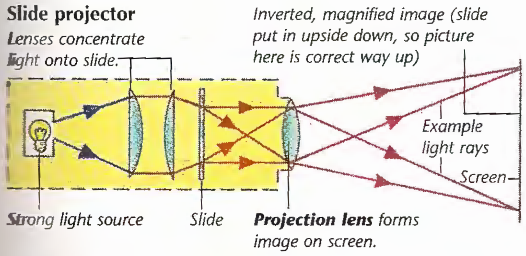
Slide projector
Lenses concentrate light onto slide.
Inverted, magnified image (slide put in upside down, so picture here is correct way up)
Example light rays
Screen
Strong light source
Slide
***Projection lens** forms image on screen.*

Telescope

An optical instrument used to make very distant (and therefore apparently very small) objects appear larger.

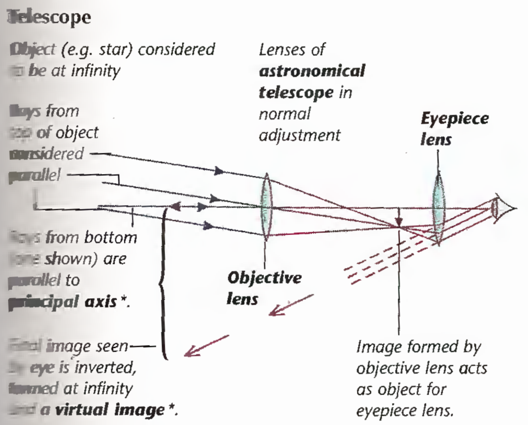
Telescope
Object (e.g. star) considered to be at infinity
*Lenses of **astronomical telescope** in normal adjustment*
Eyepiece lens
Rays from top of object considered parallel
*Rays from bottom (one shown) are parallel to **principal axis***.*
Objective lens
*Final image seen by eye is inverted, formed at infinity and a **virtual image***.*
Image formed by objective lens acts as object for eyepiece lens.

Visual angle

The angle, at the eye, of the rays coming from the top and bottom of an object or its image. The greater it is, the larger the object or image appears. Optical instruments which produce magnification, e.g. **microscopes**, do so by creating an image whose visual angle is greater than that of the object seen by the unaided eye. The **angular magnification** or **magnifying power** (see below) of such an instrument is a measurement of the amount by which it does so.

$$\textbf{Angular magnification} = \frac{\textbf{visual angle of image}}{\textbf{visual angle of object}}$$

Chromatic aberration or chromatism

The halo of colors (the **visible light spectrum** – see below, left) sometimes seen around images viewed through lenses. It results from **dispersion** (see **color**). To avoid this, good quality optical instruments contain one or more **achromatic lenses** – each consisting of two lenses combined so that any dispersion produced by one is corrected by the other.

***Achromatic lenses** in **compound microscopes** minimize **chromatic aberration**.*

Primary colors

Red, blue and green light – colors that cannot be made by combining other colored light. Mixed equally, they give white light. By mixing them in the right proportions, every color in the **visible light spectrum** can be produced. Note that these are the pure primary colors – those referred to in art (red, blue and yellow) only act as primary colors because the paints are impure.

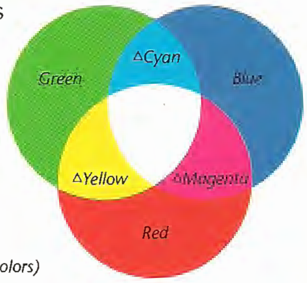
Green
△Cyan
Blue
△Yellow
△Magenta
Red

Primary colors

***Complementary** colors are any two that produce white light when mixed, e.g. red and cyan.*

*△ = **secondary colors** (combinations of primary colors)*

Color mixing

If white light is shone onto a pure colored filter, only light of the same color (range of **wavelengths***) as the filter passes through (the other colors are absorbed). This is **subtractive mixing** or **color mixing by subtraction**. If light of two different colors, filtered out in this way, is shone onto a white surface, a third color (a mixture of the two) is seen by the eye. This is **additive mixing** or **color mixing by addition**.

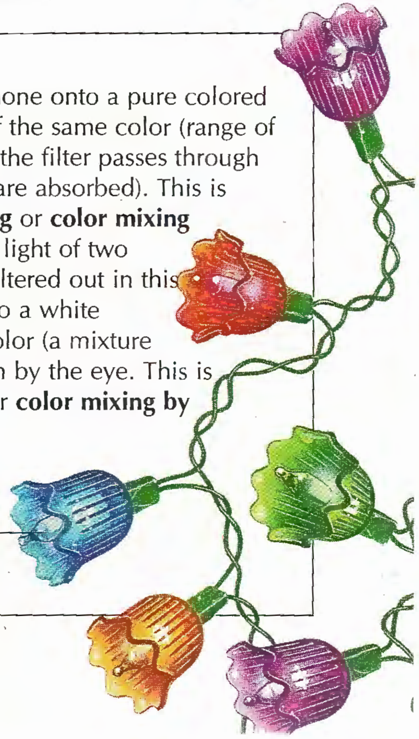

This light bulb looks blue because it only lets through blue light (all other colors have been absorbed by the blue coating).

Curved mirrors, 48; **Lenses**, **Principal axis**, 52; **Prism**, 51; **Real image**, 49 (**Image**); **Refraction**, 50; **Virtual image**, 49 (**Image**); **Wavelength**, 34.

STATIC ELECTRICITY

Electricity is the phenomenon caused by the presence or movement of charges (**electrons*** or **ions***) which exert an **electric force***. A material is said to have a negative electric charge if it has a surplus of electrons, and a positive electric charge if it has a deficit of electrons. An electric **current** (see page 60) is the movement of a charge through materials. (In a metal it is the electrons that move.) This can be contrasted with **static electricity**, which can be said to be electricity "held" by a material with electric charge.

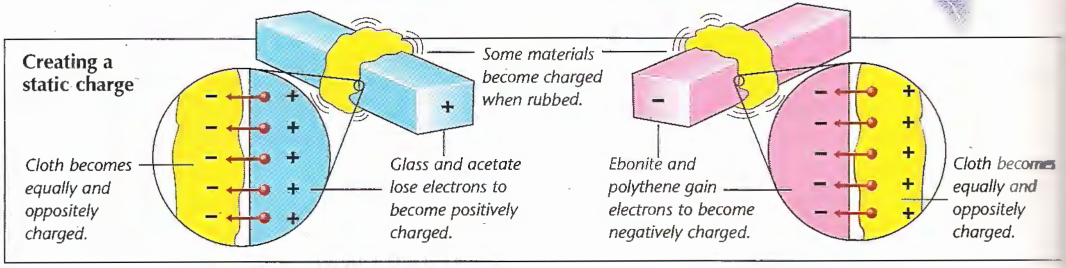

Creating a static charge

Cloth becomes equally and oppositely charged.
Some materials become charged when rubbed.
Glass and acetate lose electrons to become positively charged.
Ebonite and polythene gain electrons to become negatively charged.
Cloth becomes equally and oppositely charged.
+
–

First law of electrostatics

Like charges repel each other and unlike charges attract. A charged particle always attracts an uncharged **conductor** by **induction**.

Attraction and repulsion

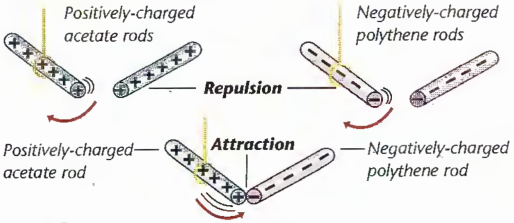

Positively-charged acetate rods
Repulsion
Negatively-charged polythene rods
Positively-charged acetate rod
Attraction
Negatively-charged polythene rod

Conductor

A material containing a large number of charges (e.g. electrons) which are free to move (see also **conductivity**, page 63). It can therefore **conduct** electricity (carry an electric **current** – see introduction). Metals, e.g. copper, aluminum and gold, are good conductors because they contain large numbers of electrons which are free to move.

Insulator

A material with very few or no charges (e.g. electrons) free to move (i.e. a bad **conductor**). Some insulators become electrically charged when rubbed. This is because electrons from the surface atoms are transferred from one substance to the next, but the charge remains on the surface.

Electroscope

An instrument for detecting small amounts of electric charge. A **gold leaf electroscope** is the most common type. When the leaf and rod become charged, they repel and the leaf diverges from the rod. The greater the charges, the larger the divergence of the leaf. A **condensing electroscope** contains a **capacitor*** between the cap and the case which increases the sensitivity.

Gold leaf electroscop

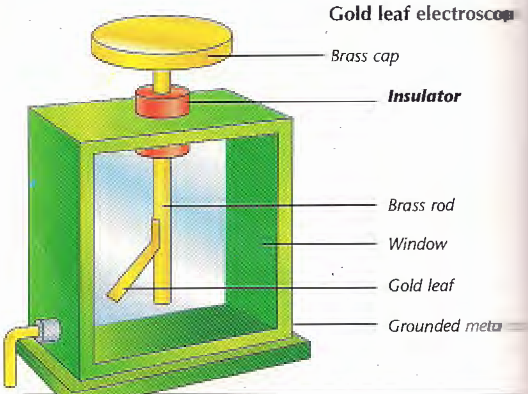

Brass cap
Insulator
Brass rod
Window
Gold leaf
Grounded meta

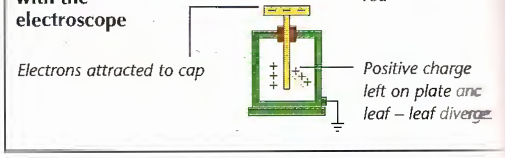

Detecting charge with the electroscope

Positively-charge rod
Electrons attracted to cap
Positive charge left on plate and leaf – leaf diverge

* **Capacitor**, 59; **Electric force**, 6; **Electrons**, 83; **Ions**, 88 (**Ionization**).

Induction or electrostatic induction

A process by which a **conductor** becomes charged with the use of another charge but without contact. Generally charges are induced in different parts of an object because of repulsion and attraction. By removing one type of charge the object is left permanently charged.

Charging a conductor by induction

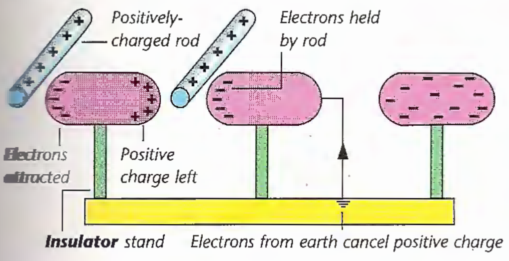
Positively-charged rod
Electrons held by rod
Electrons attracted
Positive charge left
Insulator stand
Electrons from earth cancel positive charge

Proof plane

A small disk made of a **conductor** mounted on a handle made of an **insulator**. It is used to transfer charge between objects.

Surface density of charge

The amount of charge per unit area on the surface of an object. It is greater where the surface is more curved, which leads to charge being concentrated at sharp points (see **point action**). Only a sphere has constant surface density of charge.

Variations in surface density of charge

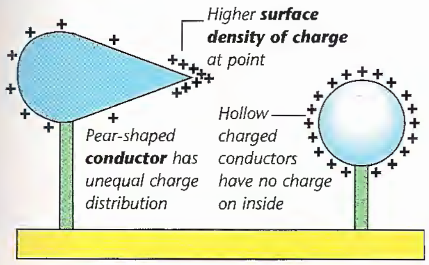
Higher ***surface density of charge*** at point
Pear-shaped ***conductor*** has unequal charge distribution
Hollow charged conductors have no charge on inside

Point action

The action which occurs around a sharp point on the surface of a positively-charged object. Positive ions in the air are repelled by the large charge at the point (see **surface density of charge**). These collide with air molecules and knock off electrons to produce more positive ions which are also repelled. The result is an **electric wind** of air molecules.

Lightning

The sudden flow of electricity from a cloud which has become charged due to the rubbing together of different particles, e.g. water droplets. A **lightning conductor** is used to help cancel the charge on the cloud by **point action** and to conduct the electricity down to earth so that it does not flow through buildings. The lightning strike is like the effect in a **discharge tube***.

Action of lightning conductor

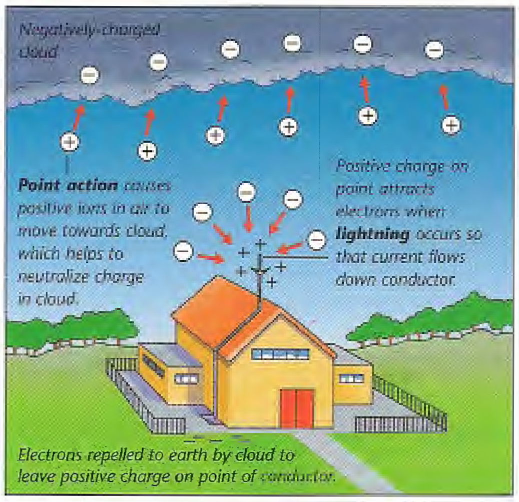
Negatively-charged cloud
Point action causes positive ions in air to move towards cloud, which helps to neutralize charge in cloud.
Positive charge on point attracts electrons when ***lightning*** occurs so that current flows down conductor.
Electrons repelled to earth by cloud to leave positive charge on point of conductor.

Van de Graaff generator

A machine in which positive charge from a point is transferred (by **point action**) to a moving band, collected by another point and deposited on a sphere-shaped conductor.

Van de Graaff generator

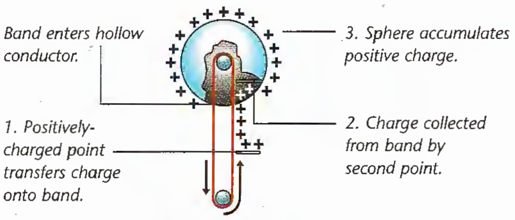
Band enters hollow conductor.
3. Sphere accumulates positive charge.
1. Positively-charged point transfers charge onto band.
2. Charge collected from band by second point.

Electrophorus

An instrument consisting of a negatively-charged **insulator** and a brass plate attached to an insulating handle. It is used to produce a number of positive charges from one negative charge.

***Discharge tube**, 80; **Mechanical energy**, 9.

POTENTIAL AND CAPACITANCE

A charge or collection of charges causes an **electric field**, i.e. a **force field*** in which charged particles experience an **electric force***. The intensity of an electric field at a point is the force per unit positive charge at that point, and the direction is the direction of the force on a positive charge at that point (see also pages 104-107). Charged objects in an electric field have **potential energy*** because of their charge and position. **Potential** itself is a property of the field (see below).

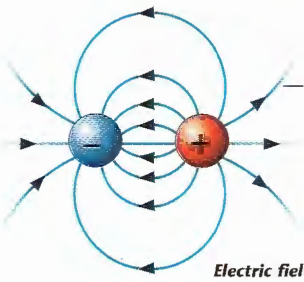

Electric field *of two opposite charges*

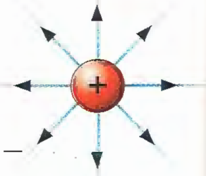

Electric field *of a single point charge*

$$F \propto \frac{q_1 q_2}{d^2}$$

where F = **electric force***; q_1, q_2 = size of charges; d = separation.

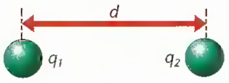

Electric force*

Potential

The **potential energy*** per unit charge at a point in an electric field, i.e. the work done in moving a unit positive charge to this point. The potential energy of a charge depends on the potential of its position and on its size. A positive charge tends to move towards points of lower potential. This is moving down the **potential gradient**. Potential cannot be measured, but the **potential difference** between two points can.

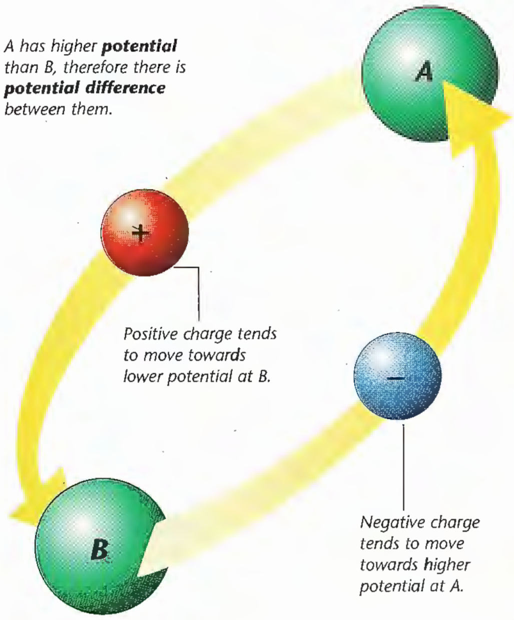

A has higher ***potential*** *than B, therefore there is* ***potential difference*** *between them.*

Positive charge tends to move towards lower potential at B.

Negative charge tends to move towards higher potential at A.

Potential difference

A difference in **potential** between two points, equal to the energy change when a unit positive charge moves from one place to another in an electric field. The unit of potential difference is the **volt** (potential difference is sometimes called **voltage**). There is an energy change of one joule if a charge of one **coulomb*** moves through one volt. A reference point (usually a connection to earth) is chosen and given a potential of zero.

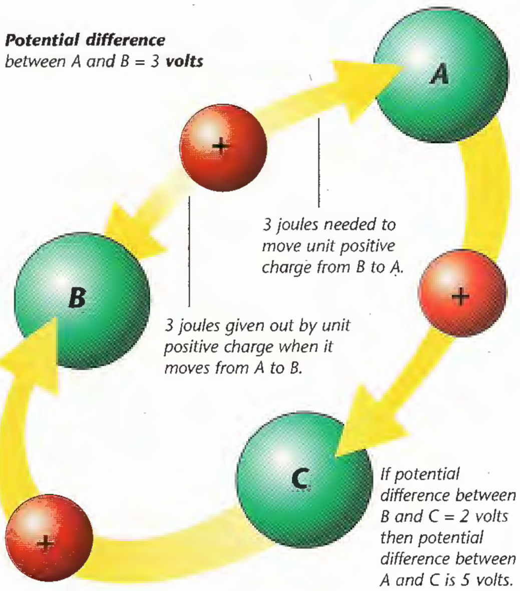

Potential difference *between A and B = 3* ***volts***

3 joules needed to move unit positive charge from B to A.

3 joules given out by unit positive charge when it moves from A to B.

If potential difference between B and C = 2 volts then potential difference between A and C is 5 volts.

Equipotential

A surface over which the **potential** is constant.

---

* **Coulomb**, 60; **Electric force**, **Force field**, 6; **Potential energy**, 8.

Capacitance

When a **conductor*** charge it undergo in **potential**. **Capac** the ratio of the charge by an object to its potential. An object **higher capacitan** **larger** charge to potential by the as an object with a capacitance.

where C = capacitance

Farad

The **unit of** capacitance an **object** whose potenti **volt when** given a

Capacitor

A device for storing of two parallel metal insulating material capacitance of a capacit dielectric used, so a suit the capacitance size required.

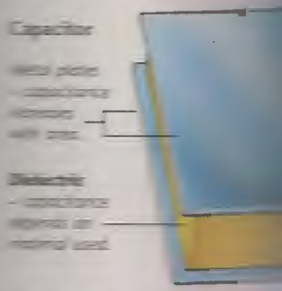

Dielectric constant

The ratio of the with a given **dielectric** the same capacitor the plates. The which the capacitan the given dielectric Note that measuring would produce a

Capacitance

When a **conductor*** is given a charge it undergoes a change in **potential**. **Capacitance** is the ratio of the charge gained by an object to its increase in potential. An object with a higher capacitance requires a larger charge to change its potential by the same amount as an object with a smaller capacitance.

$$C = \frac{Q}{V}$$

*where C = **capacitance**; Q = charge; V = **potential**.*

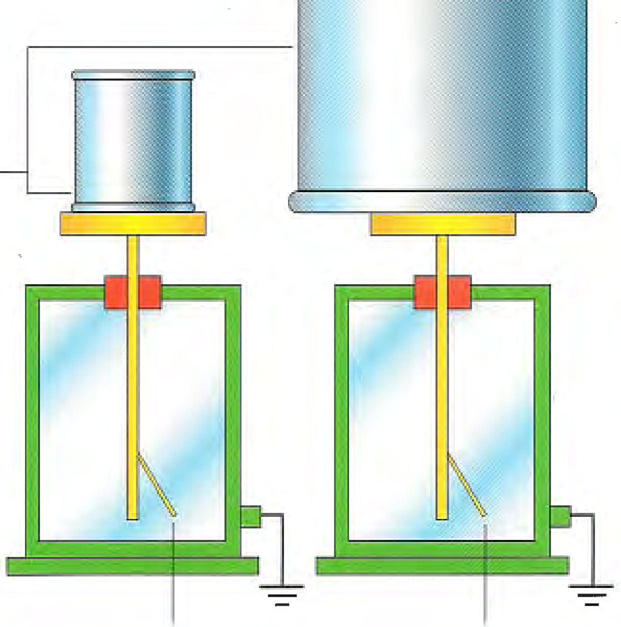

*Two different metal cans have different **capacitance**.*

*More charge (Q) must be given to larger can to give it same **potential** (V) – it has higher capacitance.*

Same leaf divergence shows same potential.

Farad

The unit of capacitance. It is the capacitance of an object whose **potential** is increased by one **volt** when given a charge of one **coulomb***.

Capacitor

A **device** for storing electric charge, consisting of **two** parallel metal plates separated by an insulating material called a **dielectric**. The capacitance of a capacitor depends on the dielectric used, so a dielectric is chosen to suit the capacitance needed and the physical size required.

Capacitor

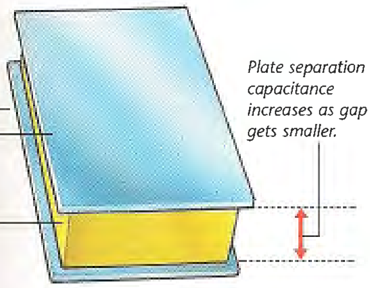

Metal plates – capacitance increases with area.

Dielectric – capacitance depends on material used.

Plate separation – capacitance increases as gap gets smaller.

Dielectric constant

The ratio of the capacitance of a **capacitor** with a given **dielectric** to the capacitance of the same capacitor with a vacuum between the plates. The value is thus the factor by which the capacitance is increased by using the given dielectric instead of a vacuum. (Note that measuring the dielectric against air would produce a very similar result.)

Electrolytic capacitor

A **capacitor** with a paste or jelly **dielectric** which gives it a very high capacitance in a small volume. Due to the nature of the dielectric, it must be connected correctly to the electricity supply.

Variable capacitor

A **capacitor** consisting of two sets of interlocking vanes, often with an air **dielectric**. The size of the interlocking area is altered to change the capacitance.

Variable capacitor

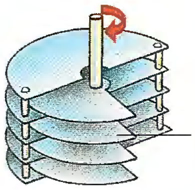

Variable capacitors are used in tuning circuits in radios.

Vanes swivel to change area between them.

Leyden jar

A **capacitor** consisting of a glass jar with foil linings inside and out. It was one of the first capacitors invented.

Paper capacitor

A **capacitor** made with two long foil plates separated by a thin waxed paper **dielectric**. **Polyester capacitors** are made in a similar way.

Paper capacitor

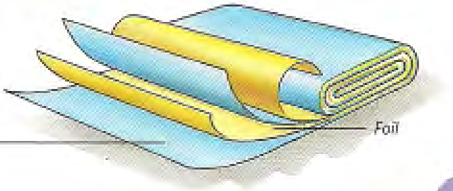

Conductor, 56; **Coulomb**, 60.

ELECTRIC CURRENT

An electric **current** (**I**) is the rate of flow of electric charge. In metal conductors, the charge which flows consists of electrons (negatively charged particles – see page 83), and these flow because in an **electric field*** there is a difference in **potential*** between two places. Therefore a **potential difference*** is needed to produce an electric current. A **circuit** is a closed loop, consisting of a source of potential difference and one or more components, around which the current flows.

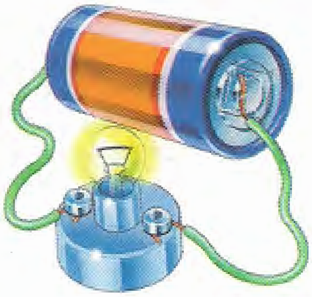

This battery (current source), wires and light bulb form a circuit.

Electromotive force (e.m.f.)

The **potential difference*** produced by a **cell***, **battery*** or **generator***, which causes current to flow in a circuit. A source of e.m.f. has two **terminals** (where wires are connected), between which it maintains a potential difference. A **back e.m.f.** is an e.m.f. produced by a component in the circuit which opposes the main e.m.f.

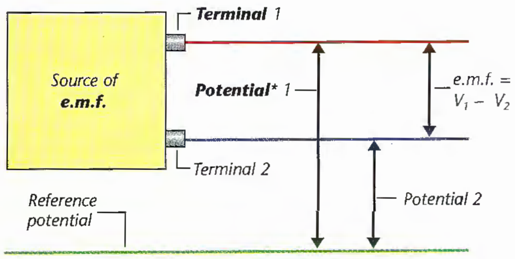
Terminal 1
Source of e.m.f.
Potential* 1
e.m.f. = $V_1 - V_2$
Terminal 2
Reference potential
Potential 2

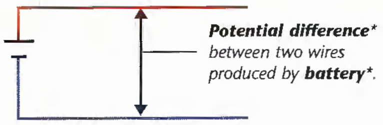
Potential difference* between two wires produced by battery*.

*One **terminal** connected to reference **potential****

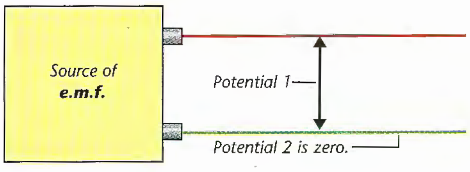
Source of e.m.f.
Potential 1
Potential 2 is zero.

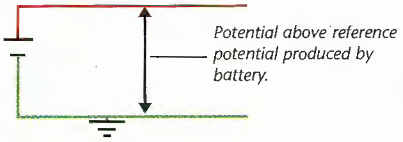
Potential above reference potential produced by battery.

Ampere or amp (A)

The **SI unit*** of current (see also page 96). One ampere is the current which, when flowing through two infinitely long wires one meter apart in a vacuum, produces a force of 2×10^{-7} newtons per meter of wire. Current is accurately measured by a **current balance**, which, by adapting the theory above, measures the force between two coils of wire through which current is flowing. **Ammeters*** are **calibrated*** using current balances.

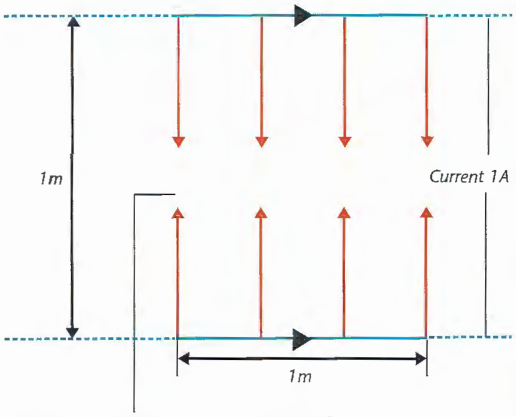
1m
Current 1A
1m

Force between each meter of wire = 2×10^{-7}N

Coulomb

The **SI unit*** of electric charge. It is equal to the amount of charge which passes a point in a conductor if one **ampere** flows through the conductor for one second.

$$Q = I \times t$$

where Q = charge past a point in **coulombs**; I = current; t = time.

---

* **Ammeter**, 77; **Battery**, 68; **Calibration**, 344; **Cell**, 68; **Electric field**, 58; **Generator**, 78; **Potential**, **Potential difference**, 58; **SI units**, 96.

Direct current (d.c.)

Current which flows in
Originally current was a
a point with higher pote
lower potential. Electron
other way, but the conv

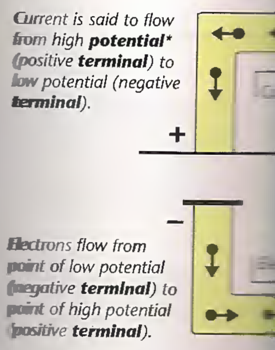
Current is said to flow from high potential* (positive terminal) to low potential (negative terminal).
+
–
Electrons flow from point of low potential (negative terminal) to point of high potential (positive terminal).

Alternating current a.

Current whose directi
at regular intervals. It is
alternating electrons
graph of current versus
waveform of the current
and electromotive forces
expressed as their root m
see picture, below

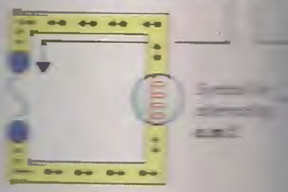

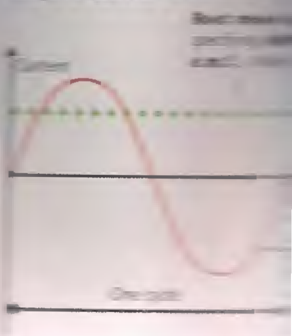

Direct current (d.c.)

Current which flows in one direction only. Originally current was assumed to flow from a point with higher **potential*** to a point with lower potential. Electrons actually flow the other way, but the convention has been kept.

*Current is said to flow from high **potential*** (positive **terminal**) to low potential (negative **terminal**).*

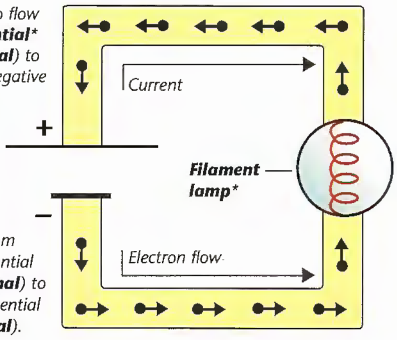

*Electrons flow from point of low potential (negative **terminal**) to point of high potential (positive **terminal**).*

Alternating current (a.c.)

Current whose direction in a circuit changes at regular intervals. It is caused by an alternating **electromotive force**. Plotting a graph of current versus time gives the waveform of the current. Alternating currents and electromotive forces are generally expressed as their **root mean square** values (see picture, below).

Electrons flow alternately one way then the other.

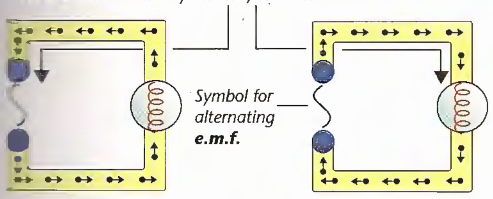

Graph of current versus time

***Root mean square**. Value used when specifying **alternating current** (or **e.m.f.**). Note that peak value is higher.*

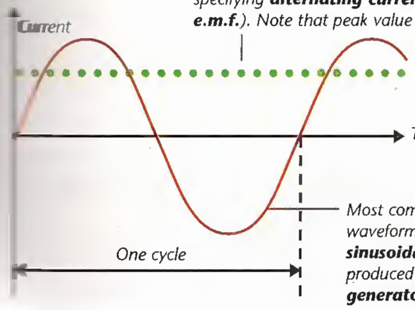

*Most common waveform (shape is **sinusoidal**) is that produced by **generators***.*

Electricity supply

Electricity for domestic and industrial use is produced at power stations by large **generators***. These produce **alternating current** at a frequency of 60Hz. Alternating current, unlike **direct current**, can be easily transformed (see **transformer**, page 79) to produce larger or smaller **potential differences***. This means that high voltages and thus low currents can be used for transmission, which considerably reduces power losses in the transmission cables.

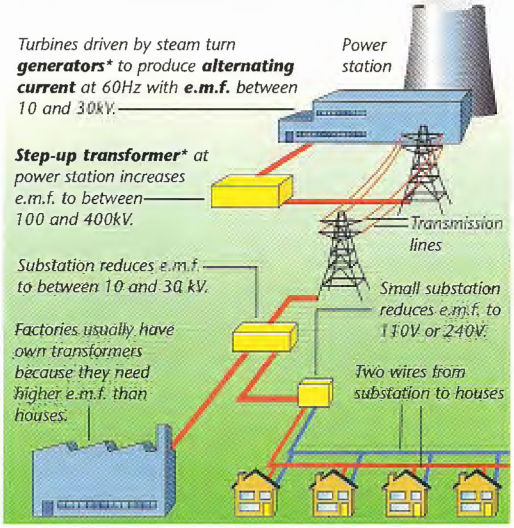

*Turbines driven by steam turn **generators*** to produce **alternating current** at 60Hz with **e.m.f.** between 10 and 30kV.*

Step-up transformer at power station increases e.m.f. to between 100 and 400kV.*

Substation reduces e.m.f. to between 10 and 30 kV.

Small substation reduces e.m.f. to 110V or 240V.

Factories usually have own transformers because they need higher e.m.f. than houses.

Two wires from substation to houses

All domestic electricity supplies consist of at least two wires from a substation along which **alternating current** flows. In some cases, one of the wires is connected to ground so that the **potential*** of the other alternates above and below ground. In some countries there is an additional wire connected to ground as a safety measure.

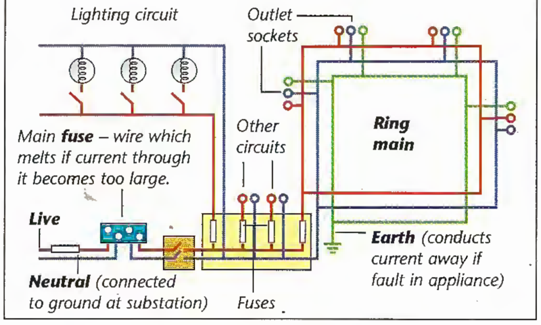

*Main **fuse** – wire which melts if current through it becomes too large.*

***Neutral** (connected to ground at substation)*

***Earth** (conducts current away if fault in appliance)*

Filament lamp, 64; **Generator**, 78; **Potential**, 58; **Potential difference**, 58; **Step-up transformer**, 79.

CONTROLLING CURRENT

The strength of a current flowing in a circuit depends on the nature of the components in the circuit as well as the **electromotive force***. The **resistance** of the components and the magnetic and electric fields they set up all affect the current in them.

Ohm's law

The current in an object at constant temperature is proportional to the **potential difference*** across its ends. The ratio of the potential difference to the current is the **resistance** of the object. The object must be at constant temperature for the law to apply since a current will heat it up and this will change its resistance (see also **filament lamp**, page 64). Ohm's law does not apply to some materials, e.g. **semiconductors***.

Ohm's law *states:*

$$\frac{V}{I} = R = \text{constant}$$

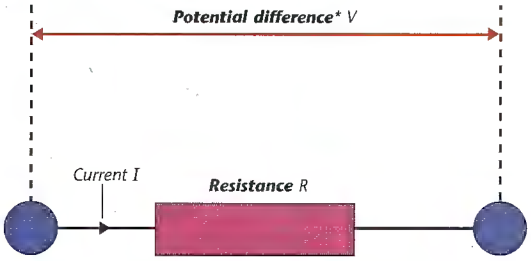

Example:

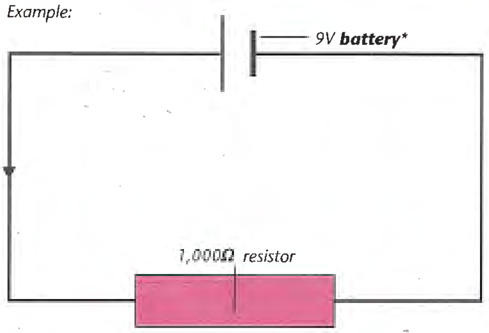

Current I through ***resistor*** $R = \frac{V}{R} = \frac{9}{1,000} = 0.009A = 9mA$

Resistance (R)

The ability of an object to resist the flow of current. The value depends on the **resistivity** of the substance from which the object is made, its shape and its size. The unit of resistance is the **ohm** (**Ω**). Electrons moving in the object hit atoms and give them energy, heating the object and using up energy from the source of **electromotive force***.

The rate at which electrical energy is changed, because of ***resistance****, to heat energy (i.e. the* ***power****) can be calculated thus:*

$$\text{Power} = IV = I^2R$$

where I = current;
V = ***potential difference****
across resistance; R = resistance.

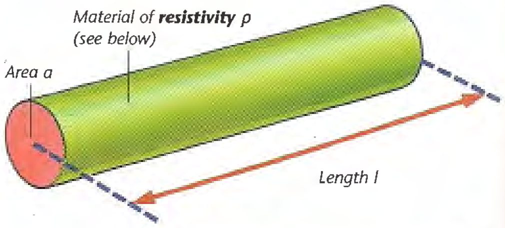

Resistance *is inversely proportional to area and proportional to length.*

$$R = \frac{\rho l}{a}$$

Resistivity (ρ)

The ability of a substance to resist current. Good **conductors*** have a low resistivity and **insulators*** have a high resistivity. It is the **reciprocal*** of the **conductivity** of the substance and depends on temperature. See also page 112.

Resistor

A device with a particular **resistance** value. Resistors can have values from less than one **ohm** up to many millions of ohms. The most common type is the **carbon resistor**, made from compressed carbon of known **resistivity**.

Carbon resistor

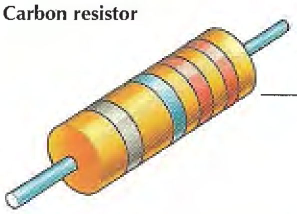

Color-coded stripes show ***resistance*** *value.*

Symbols for resis

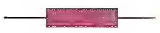

or

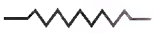

---

* **Battery**, 68; **Conductor**, 56; **Electromotive force (e.m.f.)**, 60; **Insulator**, 56; **Potential difference**, 58; **Power**, 9; **Reciprocal**, 345; **Semiconductors**, 65.

Conductivity

The ability of a substance to allow the flow of current (see also **conductor** and **insulator**, page 56). It is the inverse of the **resistivity**.

Internal resistance (r)

The **resistance** of a **cell*** or **battery*** to the current it causes. It is the resistance of the connections in the cell and some chemical effects (e.g. **polarization***). The current in a circuit may therefore be less than expected.

Internal resistance *is part of* ***resistance*** *of circuit.*

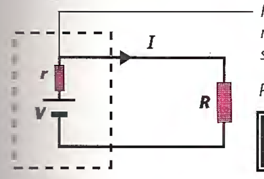

Internal resistance represented by resistor symbol in ***cell****.*

From ***Ohm's law****:*

$$V = I(R + r)$$

Variable resistor

A device whose **resistance** can be changed mechanically. It is either a coil of wire of a particular **resistivity** around a drum along which a contact moves (for high currents) or a carbon track with a moving contact. A variable resistor can be used as a **potential divider** if an extra contact is added. It is then a **potentiometer**.

Types of variable resistor

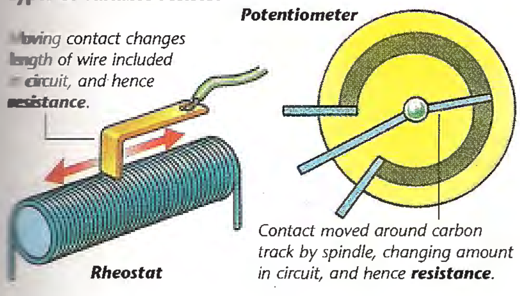

Moving contact changes length of wire included in circuit, and hence ***resistance****.*

Contact moved around carbon track by spindle, changing amount in circuit, and hence ***resistance****.*

Potential divider or voltage divider

A device used to produce a **potential difference*** from another, higher potential difference.

Circuit diagram of potential divider

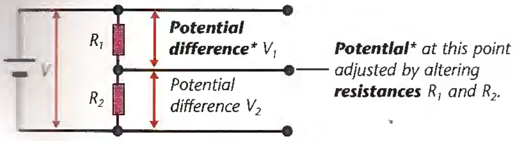

Potential difference* V_1

Potential difference V_2

Potential* *at this point adjusted by altering* ***resistances*** R_1 *and* R_2.

Wheatstone bridge

A circuit used to measure an unknown **resistance** (see diagram). When the **galvanometer*** indicates no current, the unknown value of one **resistor** can be calculated from the other three. The **meter bridge** is a version of the wheatstone bridge in which two of the resistors are replaced by a meter of wire with a high resistance. The position of the contact from the galvanometer on the wire gives the ratio R_3/R_4 in the circuit shown below.

Wheatstone bridge

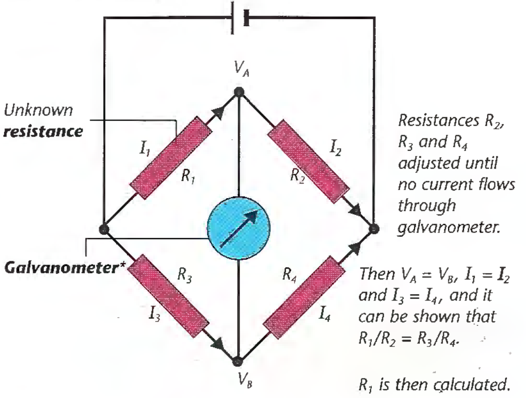

Unknown ***resistance***

Galvanometer*

Resistances R_2*,* R_3 *and* R_4 *adjusted until no current flows through galvanometer.*

Then $V_A = V_B$, $I_1 = I_2$ *and* $I_3 = I_4$*, and it can be shown that* $R_1/R_2 = R_3/R_4$.

R_1 *is then calculated.*

Kirchhoff's laws

Two laws which summarize conditions for the flow of current at an instant. The first states that the total current flowing towards a junction is equal to the total current flowing away from the junction. The second states that the sum of the **potential differences*** around a circuit, which for each **resistor** is the product of the current and the **resistance**, is equal to the **electromotive force*** applied to the circuit.

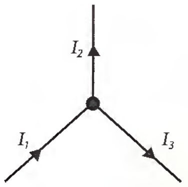

Kirchhoff's first law:

$$I_1 = I_2 + I_3$$

Kirchhoff's second law:

$$V = IR_1 + IR_2$$

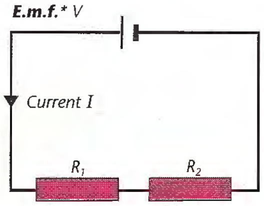

---

*Battery, Cell, 68; Electromotive force (e.m.f.), 60; Galvanometer, 77; Polarization, 68; Potential, Potential difference, 58.

Controlling current (continued)

Series

An arrangement of components in which all of the current passes through them one after the other.

Resistors* **in series**

Total resistance* $R_T = R_1 + R_2$

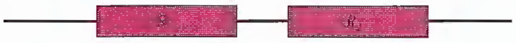

Capacitors* **in series**

Total capacitance* $\frac{1}{C_T} = \frac{1}{C_1} + \frac{1}{C_2}$

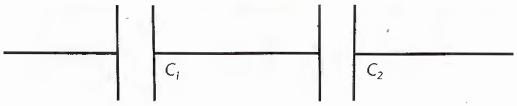

Parallel

An arrangement of components in which current divides to pass through all at once.

Resistors* **in parallel**

Total resistance* $\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2}$

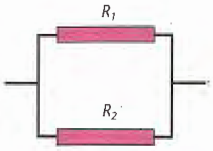

Capacitors* **in parallel**

Total capacitance* $C_T = C_1 + C_2$

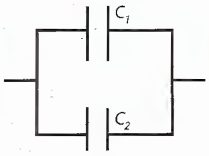

Filament lamp

A lamp consisting of a coil of tungsten wire (the **filament**) inside a glass bulb containing argon or nitrogen gas at low pressure. When current flows through the coil, it heats up rapidly and gives out light. Tungsten is used because it has a very high melting point and the bulb is gas-filled to reduce evaporation of the tungsten.

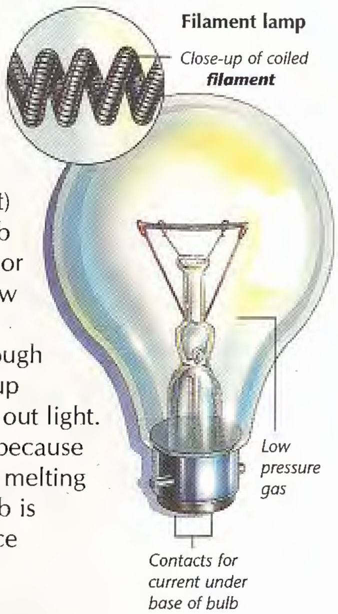
Filament lamp

Close-up of coiled ***filament***
Low pressure gas
Contacts for current under base of bulb

Switch

A device, normally mechanical (but see also **transistor**), which is used to make or break a circuit. A **relay*** is used when a small current is required to switch a larger current on and off.

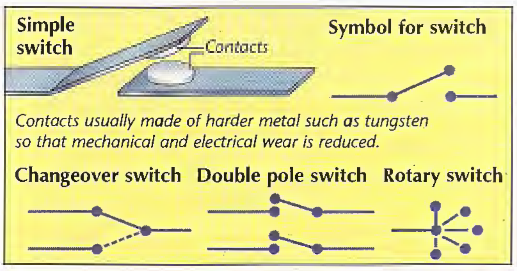

Contacts usually made of harder metal such as tungsten so that mechanical and electrical wear is reduced.

Impedance

The ratio of the **potential difference*** applied to a circuit to the **alternating current*** which flows in it. It is due to two things, the **resistance*** of the circuit and the **reactance**. The effect of impedance is that the **e.m.f.*** and current can be out of phase.

Reactance

The "active" part of **impedance** to **alternating current***. It is caused by **capacitance*** and **inductance** in a circuit which alter the **electromotive forces*** as the current changes.

Inductance

The part of the **impedance** of a circuit due to changing current affecting the **e.m.f.*** (see also **electromagnetic induction**, page 78). This happens in a device called an **inductor**.

Alternating potential difference*

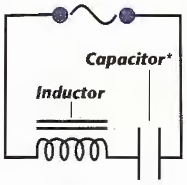

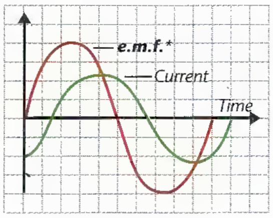

---

* **Alternating current**, 61; **Capacitance**, **Capacitor**, 59; **Electromotive force (e.m.f.)**, 60; **Potential difference**, 58; **Relay**, 75; **Resistance**, **Resistor**, 62.

SEMICON

Semiconductors ar
conductor and an
temperature or inc
are widely used in

Doping

The introduction of a sm
into a semiconductor. D
impurity used, the sem
either a **p-type** or **n-type**
types are used to make

Diode

A device made from one
semiconductor (see **dopi**
n-type semiconductor
very low **resistance*** in
is said to be **forward bia**
resistance in the other d

Construction of diode

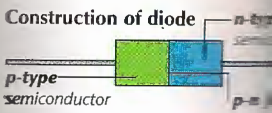

Forward biased *– diode has low* ***resistance****.

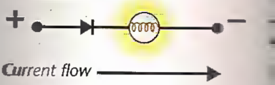

Current flow

Half-wave rectificati

The use of a **diode** to re
flowing in one directio
current*. Current only
the circuit.

Half-wave rectification

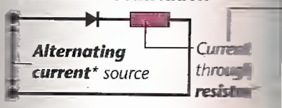

Full-wave rectificati

The conversion of alter
direct current*. It is use
is required from altern

Full-wave rectification

Alternating current* *source*

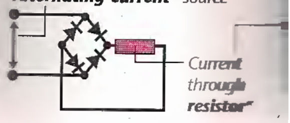

---

* **Alternating current**, 61; **Direc**
Resistance, **Resistivity**, **Resistor**

SEMICONDUCTORS

Semiconductors are materials whose **resistivity*** is between that of a **conductor** and an **insulator** (see page 56) and decreases with increasing temperature or increasing amounts of impurities (see **doping**, below). They are widely used in electronic circuits (see also page 111).

Doping

The introduction of a small amount of impurity into a semiconductor. Depending on the impurity used, the semiconductor is known as either a **p-type** or **n-type**. Combinations of these types are used to make **diodes** and **transistors**.

Diode

A device made from one piece of **p-type** semiconductor (see **doping**) and one piece of **n-type** semiconductor joined together. It has a very low **resistance*** in one direction (when it is said to be **forward biased**) and a very high resistance in the other direction (**reverse biased**).

Construction of diode

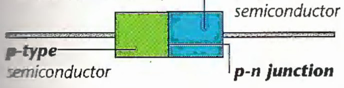
n-type semiconductor
p-type semiconductor
p-n junction

Symbol for diode

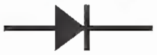

Forward biased *– diode has low* ***resistance****.*

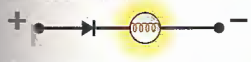

Current flow

Reverse biased *– diode has very high resistance.*

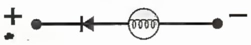

Negligible current flow

Half-wave rectification

The use of a **diode** to remove all the current flowing in one direction from **alternating current***. Current only flows one way around the circuit.

Half-wave rectification

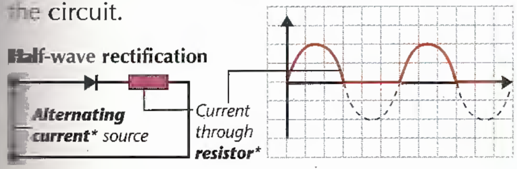
*Alternating current** source
Current through *resistor**

Full-wave rectification

The conversion of **alternating current*** to **direct current***. It is used when direct current is required from alternating current.

Full-wave rectification

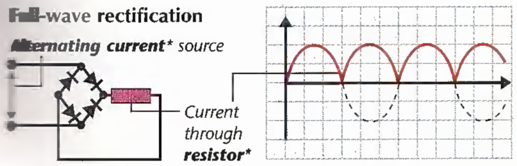
*Alternating current** source
Current through *resistor**

Light emitting diode (LED)

A **diode** with a higher **resistance*** than normal, in which light is produced instead of heat.

Numeric display of shaped LEDs

Symbol for light emitting diode

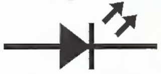

Thermistor

A semiconductor device whose **resistance*** varies with temperature, used in electronic circuits to detect temperature changes.

Transistor

A semiconductor, normally made from a combination of the two types of semiconductor. There are three connections, the **base**, **collector** and **emitter** (see diagrams below). The **resistance*** between the collector and emitter changes from very high to very low when a small current flows into the base. This small base current can therefore be used to control a much larger collector to emitter current.

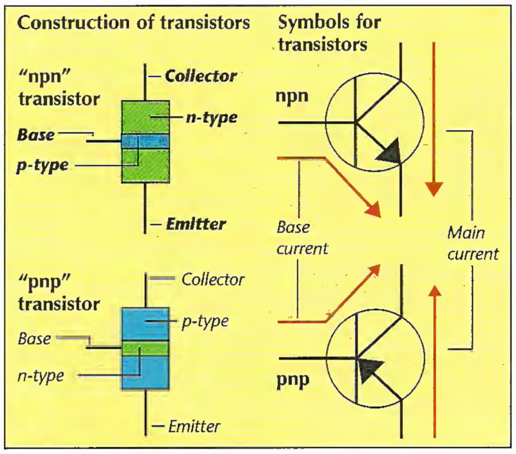
Construction of transistors
Symbols for transistors
"npn" transistor
Collector
n-type
Base
p-type
Emitter
npn
Base current
Main current
"pnp" transistor
Collector
p-type
Base
n-type
Emitter
pnp

---

*Alternating current**, 61; **Direct current**, 61; **Resistance**, **Resistivity**, **Resistor**, 62.

ELECTROLYSIS

Electrolysis is the process whereby electric current flows through a liquid containing **ions*** (atoms which have gained or lost an **electron*** to become charged) and the liquid is broken down as a result. The current is conducted by the movement of ions in the liquid, and chemicals are deposited at the points where the current enters or leaves the liquid. There are a number of industrial applications.

Electrolyte

A compound which conducts electricity when either molten or dissolved in water. All compounds made from ions or which split into ions when dissolved (**ionization***) are electrolytes. The concentration of ions in an electrolyte determines how well it conducts electricity.

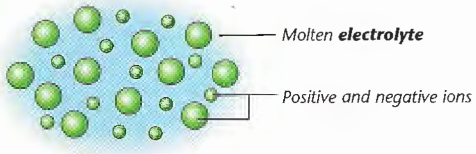
Molten **electrolyte**
Positive and negative ions

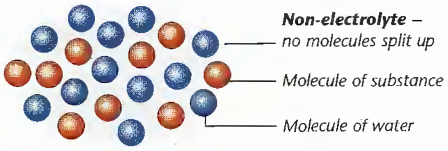
Non-electrolyte – no molecules split up
Molecule of substance
Molecule of water

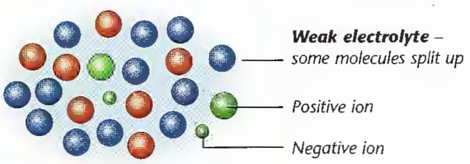
Weak electrolyte – some molecules split up
Positive ion
Negative ion

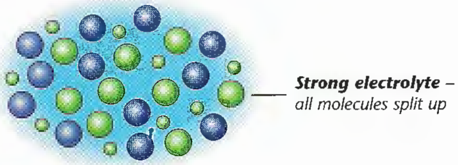
Strong electrolyte – all molecules split up

Electrode

A piece of metal or carbon placed in an **electrolyte** through which electric current enters or leaves during electrolysis. Two are needed – the **anode** (positive electrode) and the **cathode** (negative electrode). An **active electrode** is one which is chemically changed by electrolysis; an **inert electrode** is one which is not changed.

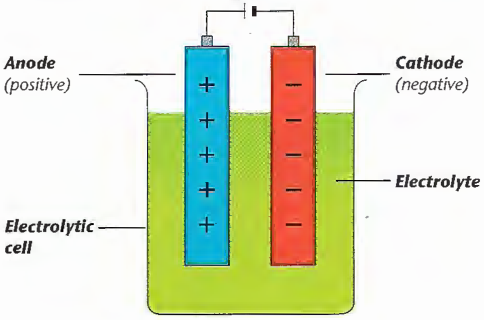
+ –
Anode (positive)
Cathode (negative)
Electrolyte
Electrolytic cell

Electrolytic cell

A vessel in which electrolysis takes place. It contains the **electrolyte** and the **electrodes**.

Ionic theory of electrolysis

A theory which attempts to explain what happens in the **electrolyte** and at the **electrodes** during electrolysis. It states that the **cations** (positive ions) are attracted towards the **cathode** and the **anions** (negative ions) towards the **anode**. There they gain or lose electrons respectively to form atoms (they are then said to be **discharged**). If there are two or more different anions, then one of them will be discharged in preference to the others. This is called **preferential discharge**.

Electrolysis of copper sulphate solution

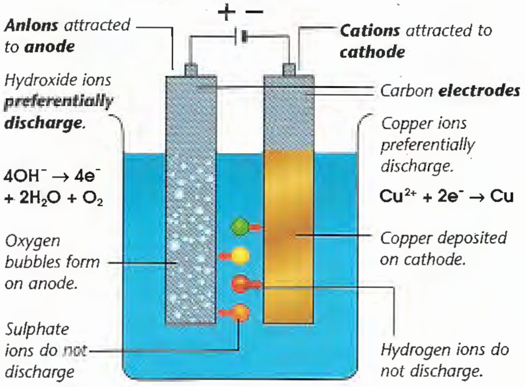
+ –
Anions attracted to **anode**
Hydroxide ions **preferentially discharge**.
$4OH^- \rightarrow 4e^- + 2H_2O + O_2$
Oxygen bubbles form on anode.
Sulphate ions do not discharge
Cations attracted to **cathode**
Carbon **electrodes**
Copper ions preferentially discharge.
$Cu^{2+} + 2e^- \rightarrow Cu$
Copper deposited on cathode.
Hydrogen ions do not discharge.

Faraday's laws of elect

Two laws which relate th
of electricity which passe
electrolyte to the masses
which are deposited. **Fara**
that the mass of the subst
proportional to the quanti
electrochemical equivale
the mass liberated by one
one second). **Faraday's se**
the mass of the substance
inversely proportional to
charge on its ion.

Electrolysis of copper sulphate s
electrodes (copper voltameter)

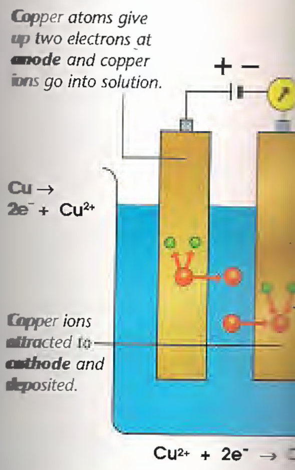
Copper atoms give up two electrons at anode and copper ions go into solution.
+ –
$Cu \rightarrow 2e^- + Cu^{2+}$
Copper ions attracted to cathode and deposited.
$Cu^{2+} + 2e^- \rightarrow$

Uses of electro

Electroplating or elec

The coating of a metal o
layer of another metal b
object forms the **cathode**
coating metal are in the
electrolyte.

Steel is cheap but ***corrodes**** *eas*
steel food cans are plated with a
fine layer of tin (which is less reacti
to prevent corrosion.

Electro-refining

A method of purifying m
Impure metal forms the
metal ions move to the c
pure metal. The impuriti
of the vessel.

---

* **Electrons**, 83; **Ions**, 88 (**Ionization**).

Ammeter, 77; **Corrosion**, 209.

Faraday's laws of electrolysis

Two laws which relate the quantity of electricity which passes through an **electrolyte** to the masses of the substances which are deposited. **Faraday's first law** states that the mass of the substance deposited is proportional to the quantity of electricity (the **electrochemical equivalent** of a substance is the mass liberated by one ampere flowing for one second). **Faraday's second law** states that the mass of the substance deposited is inversely proportional to the size of the charge on its ion.

Electrolysis of copper sulphate solution with copper electrodes (copper voltameter)

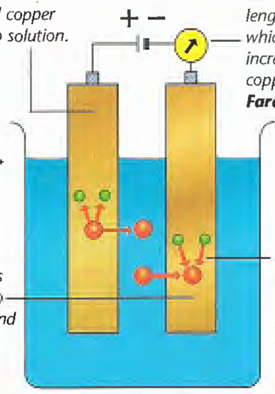
+ –
$Cu \rightarrow 2e^- + Cu^{2+}$
$Cu^{2+} + 2e^- \rightarrow Cu$

*Copper atoms give up two electrons at **anode** and copper ions go into solution.*

Ammeter measures current – if current or length of time for which it flows is increased, then more copper deposited – see **Faraday's first law**.*

*Copper ions attracted to **cathode** and deposited.*

*Each ion needs two electrons to be deposited on cathode. Aluminum ions (Al^{3+}) would need half as many electrons again – see **Faraday's second law**.*

Voltameter or coulometer

An **electrolytic cell** used for investigating the relationships between the amount of substance produced at the **electrodes** and the current which passes through the cell. For example, the **copper voltameter** (see below left) contains copper sulphate and copper electrodes.

Hoffmann voltameter

A type of **voltameter** used for collecting and measuring the volumes (and hence the masses) of gases liberated during electrolysis. For example, electrolysis of acidified water produces hydrogen and oxygen in a two to one ratio (note that this also indicates the chemical composition of water, i.e. H_2O).

Hoffmann voltameter

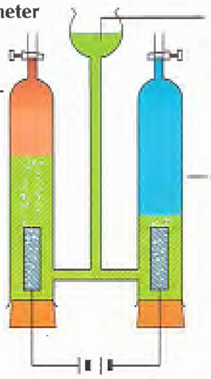

Water with small amount of sulphuric acid added (causes more hydrogen and hydroxide ions to be produced to speed up experiment).

*At **anode**:*
$4OH^- \rightarrow 4e^- + 2H_2O + O_2$

*At **cathode**:*
$2H^+ + 2e^- \rightarrow H_2$

One molecule of oxygen produced for every four electrons.

One molecule of hydrogen gas produced for every two electrons.

Uses of electrolysis

Electroplating or electrodeposition

The coating of a metal object with a thin layer of another metal by electrolysis. The object forms the **cathode**, and ions of the coating metal are in the **electrolyte**.

*Steel is cheap but **corrodes*** easily, so steel food cans are plated with a very fine layer of tin (which is less reactive) to prevent corrosion.*

Electro-refining

A method of purifying metals by electrolysis. Impure metal forms the **anode**, from which metal ions move to the **cathode** and form pure metal. The impurities fall to the bottom of the vessel.

Metal extraction

A process which produces metals from their molten ores by electrolysis. Very reactive metals are obtained by this process, e.g. sodium and aluminum.

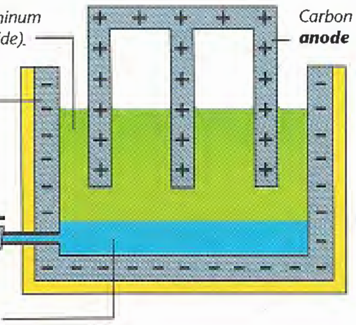

*Electrolysis of aluminum ore (aluminum oxide) **electrolyte***

Carbon **anode**

*Carbon **cathode***

Aluminum tapped off

Aluminum ions discharged at cathode to form aluminum atoms.

***Ammeter**, 77; **Corrosion**, 209.

CELLS AND BATTERIES

The Italian scientist Volta first showed that a **potential difference*** exists between two different metals when they are placed in certain liquids (**electrolytes***) and therefore that a **direct current*** can be produced from chemical energy. This arrangement is called a **cell**, **electrochemical** or **voltaic cell**. The potential difference (caused by chemical changes in the cell) is called an **electromotive force*** and its size depends on the metals used. A **battery** is two or more connected cells.

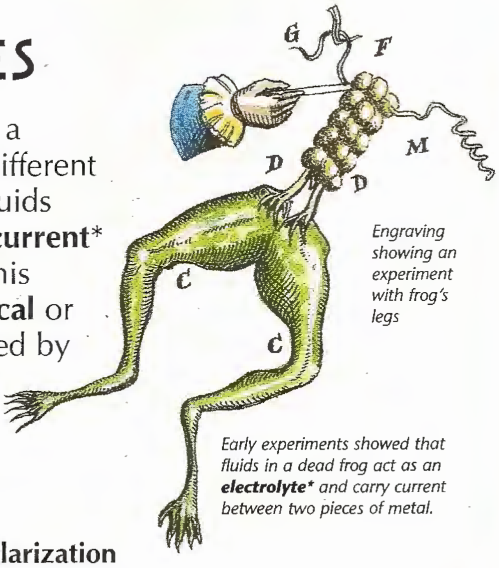

Engraving showing an experiment with frog's legs

*Early experiments showed that fluids in a dead frog act as an **electrolyte*** and carry current between two pieces of metal.*

Voltaic pile

The first battery made, consisting of a pile of silver and zinc disks separated by cardboard or cloth soaked in salt water. This arrangement is the same as a number of **simple cells** linked together.

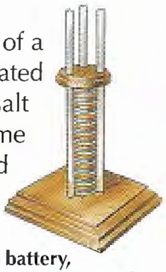

Volta's battery, known as a Voltaic pile

Close-up of a Voltaic pile

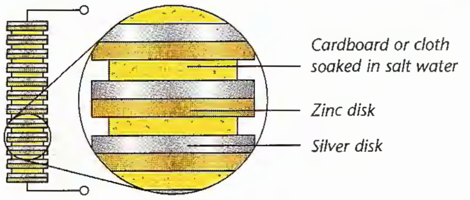
Cardboard or cloth soaked in salt water
Zinc disk
Silver disk

Simple cell

Two plates of different metals separated by a salt or acid solution **electrolyte*** (normally copper and zinc plates and dilute sulphuric acid). The simple cell only produces an **electromotive force*** for a short time before **polarization** and **local action** have an effect.

Action of simple cell

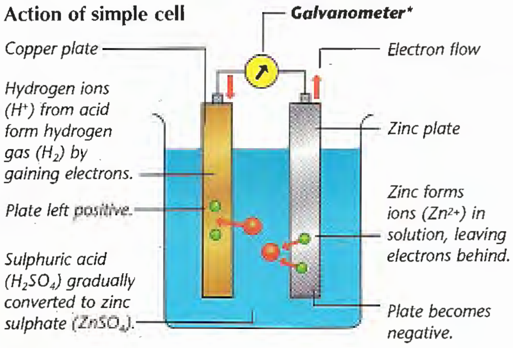
Galvanometer*
Copper plate
Electron flow
Hydrogen ions (H⁺) from acid form hydrogen gas (H₂) by gaining electrons.
Zinc plate
Plate left positive.
Zinc forms ions (Zn²⁺) in solution, leaving electrons behind.
Sulphuric acid (H₂SO₄) gradually converted to zinc sulphate (ZnSO₄).
Plate becomes negative.

Polarization

The formation of bubbles of hydrogen on the copper plate in a **simple cell**. This reduces the **electromotive force*** of the cell, both because the bubbles insulate the plate and also because a **back e.m.f.*** is set up. Polarization can be eliminated by adding a **depolarizing agent**, which reacts with the hydrogen to form water.

Polarization and local action in a simple cell

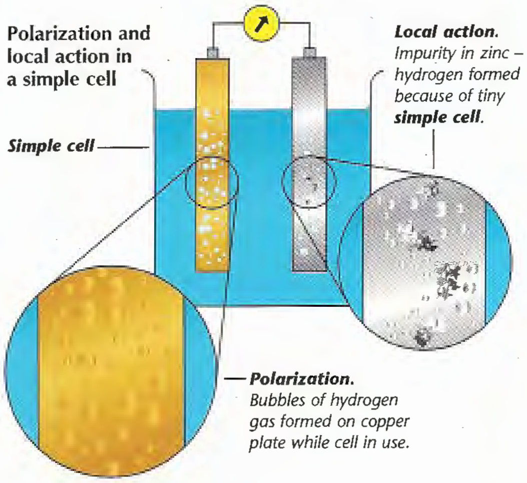
Simple cell
Local action. Impurity in zinc – hydrogen formed because of tiny simple cell.
Polarization. Bubbles of hydrogen gas formed on copper plate while cell in use.

Local action

The production of hydrogen at the zinc plate in a **simple cell**. Impurities (traces of other metals) in the zinc plate mean that tiny simple cells are formed which produce hydrogen due to **polarization**. Hydrogen is also produced as the zinc dissolves in the acid (even when the cell is not working). Local action can be prevented by coating the plate with an **amalgam***.

---

* **Amalgam**, 344; **Back e.m.f.**, 60 (**Electromotive force**); **Direct current**, 61; **Electrolyte**, 66; **Galvanometer**, 77; **Potential difference**, 58.

Capacity

The ability of a cell to p
a period of time. It is m
hours. For example, a 1
should produce one am

Leclanché cell

A cell in which **polariza**
manganese dioxide (a **d**
This removes hydrogen
is formed, but continues
excess hydrogen when t
The cell provides an **ele**
of 1.5V.

Leclanché cell

Carbon rod

Porous pot

*Carbon and manganese oxide (**depolarizing agent**)*

Ammonium chloride solution

Standard cell

A cell which produces a
and constant **electromot**
in laboratories for exper

Primary cell

Any cell which has a lim
chemicals inside it are e
cannot be replaced easi

Secondary cell

Also known as an **accu**
cell. A cell which can
connection to another
The main types are the
accumulator and the
alkaline cell.

Alkaline cell

A secondary cell conta
of potassium hydroxi
plates are normally ma
cadmium compounds
nickel-cadmium cell.
be left for months in a
without ill-effect.

---

* **Electromotive force** (**e.m.f.**

Capacity

The ability of a cell to produce current over a period of time. It is measured in **ampere hours**. For example, a 10 ampere-hour cell should produce one ampere for 10 hours.

Leclanché cell

A cell in which **polarization** is overcome by manganese dioxide (a **depolarizing agent**). This removes hydrogen more slowly than it is formed, but continues working to remove excess hydrogen when the cell is not in use. The cell provides an **electromotive force*** of 1.5V.

Leclanché cell

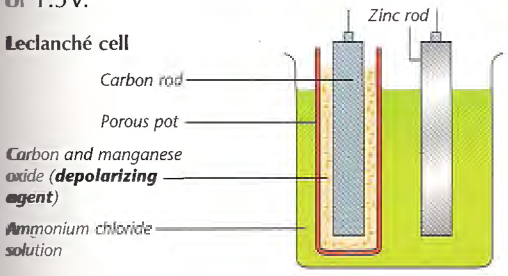
Zinc rod
Carbon rod
Porous pot
Carbon and manganese oxide (**depolarizing agent**)
Ammonium chloride solution

Standard cell

A cell which produces an accurately known and constant **electromotive force***. It is used in laboratories for experimental work.

Primary cell

Any cell which has a limited life because the chemicals inside it are eventually used up and cannot be replaced easily.

Dry cell

A version of the **Leclanché cell** in which the ammonium chloride solution is replaced by paste containing ammonium chloride, meaning that it is portable. The cell provides an **electromotive force*** of 1.5V. Dry cells deteriorate slowly due to **local action**, but still have a life of many months.

Dry cell

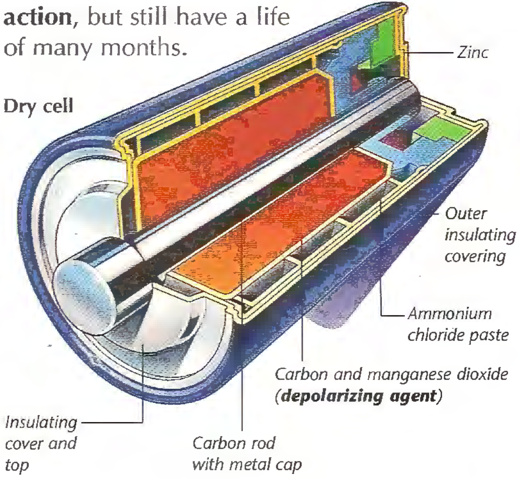
Zinc
Outer insulating covering
Ammonium chloride paste
Carbon and manganese dioxide (**depolarizing agent**)
Insulating cover and top
Carbon rod with metal cap

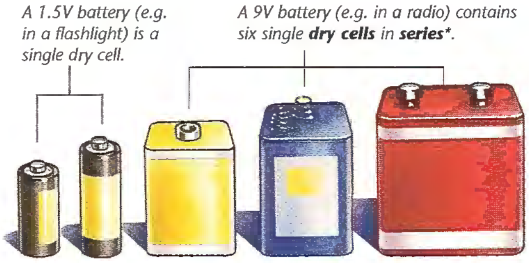

A 1.5V battery (e.g. in a flashlight) is a single dry cell.

*A 9V battery (e.g. in a radio) contains six single **dry cells** in **series***.*

Secondary cell

Also known as an **accumulator** or **storage cell**. A cell which can be recharged by connection to another source of electricity. The main types are the **lead-acid accumulator** and the **nickel-cadmiun alkaline cell**.

Alkaline cell

A secondary cell containing an **electrolyte*** of potassium hydroxide solution. The plates are normally made of nickel and cadmium compounds (it is then called a **nickel-cadmium cell**). Alkaline cells may be left for months in a discharged condition without ill-effect.

Lead-acid accumulator

A **secondary cell** containing a dilute sulphuric acid **electrolyte***, and plates made from lead and lead compounds. The cell can give out a very large current because it has a low **internal resistance***. It is mainly used in vehicles for starting and lighting.

Battery consisting of lead-acid accumulators

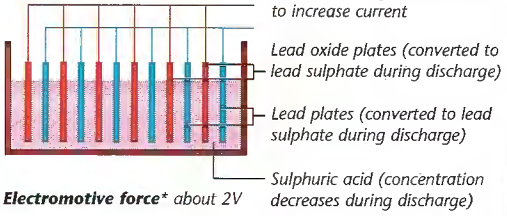
Plates have large surface area to increase current
Lead oxide plates (converted to lead sulphate during discharge)
Lead plates (converted to lead sulphate during discharge)
Sulphuric acid (concentration decreases during discharge)
Electromotive force* about 2V

---

* **Electromotive force (e.m.f.)**, 60; **Electrolyte**, 66; **Internal resistance**, 63; **Series**, 64.

MAGNETS

All **magnets** have a **magnetic field*** around them, and a **magnetic force*** exists between two magnets due to the interaction of their fields. Any material which is capable of being **magnetized** (can become a magnet) is described as **magnetic** (see **ferromagnetic**, below) and becomes magnetized when placed in a magnetic field. The movement of charge (normally **electrons***) also causes a magnetic field (see **electromagnetism**, pages 74-76).

Pole

A point in a magnet at which its **magnetic force*** appears to be concentrated. There are two types of pole – the **north** or **north seeking pole** and the **south** or **south seeking pole** (identified by allowing the magnet to line up with the Earth's **magnetic field***). All magnets have an equal number of each type of pole. The **first law of magnetism** states that unlike poles attract and like poles repel.

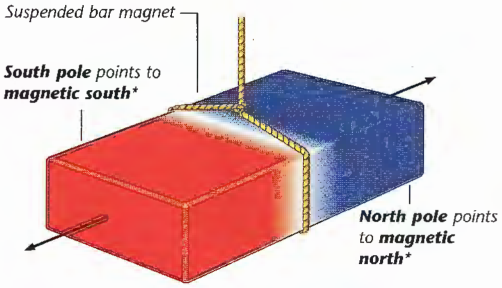
Suspended bar magnet
South pole points to **magnetic south***
North pole points to **magnetic north***

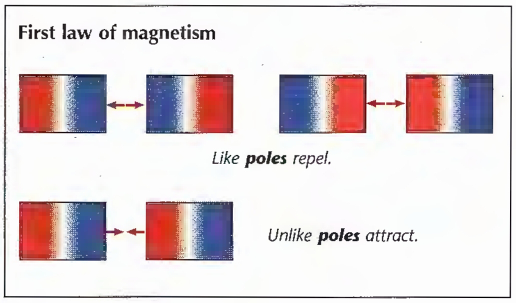

Magnetic axis

An imaginary line about which the **magnetic field*** of a magnet is symmetrical.

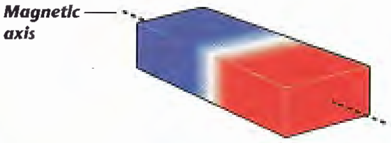

Ferromagnetic

Describes a material which is strongly magnetic (i.e. is magnetized easily). Iron, nickel, cobalt and alloys of these are ferromagnetic, and are described as either **hard** or **soft**. **Sintered** materials (made by converting various mixtures of powders of the above metals into solids by heat and pressure) can be made magnetically very hard or soft by changing the metals used.

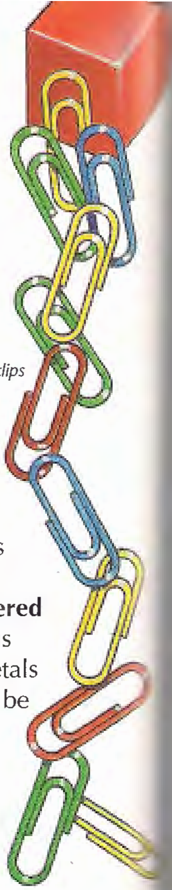

These paper clips have been temporarily ***magnetized****.*

Hard

Describes a **ferromagnetic** material which does not easily lose its magnetism after being magnetized, e.g. steel. Magnets made from these materials are called **permanent magnets**.

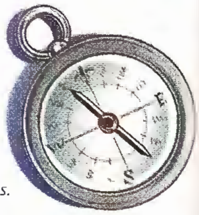

Hard ferromagnetic *materials are used as* ***permanent magnets****, e.g. as compass needles.*

Soft

Describes a **ferromagnetic** material which does not retain its magnetism after being magnetized, e.g. iron. Magnets made from these materials are called **temporary magnets**. **Residual magnetism** is the small amount of magnetism which can be left in magnetically soft materials.

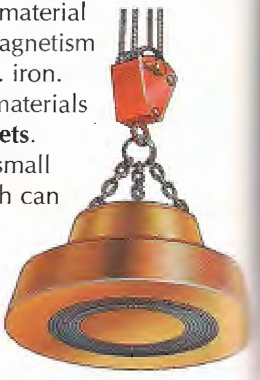

Soft ferromagnetic *materials are used as* ***cores**** *in* ***electromagnets****.*

Susceptibility

A measurement of the ability of a substance to become magnetized. **Ferromagnetic** materials have a high susceptibility.

---

* **Core**, **Electromagnet**, 74; **Electrons**, 83; **Magnetic field**, 72; **Magnetic force**, 6; **Magnetic north**, **Magnetic south**, 73.

Domain theory of ma

States that **ferromagneti** **dipoles** or **molecular m** with each other. These areas called **domains**, i in the same direction. A material becomes magn domains become **order**

Magnetizatior

When an object is mag become aligned (see **d** only happens when an **field*** and is called **ind**

Induced magnetism

Magnetic material outside ***magnetic field****.*

North end of ***dipoles*** *attracted to* ***south pole*** *of magnet – object becomes magnetized.*

Single touch

A method of magnetizi stroking it repeatedly w **permanent magnet** (see is induced in the object **field*** of the magnet.

Magnetism induced by single touch

Divided touch

A method of magnetizi stroking it repeatedly fro with the opposite **poles** **magnets** (see **hard**). Mag in the object from the m the magnets.

Magnetism induced by divide

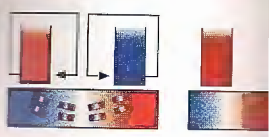

Alternating current, 61; **Magne**
Magnetic force, 6.

Domain theory of magnetism

States that **ferromagnetic** materials consist of **dipoles** or **molecular magnets**, which interact with each other. These are all arranged in areas called **domains**, in which they all point in the same direction. A ferromagnetic material becomes magnetized when the domains become **ordered** (i.e. aligned).

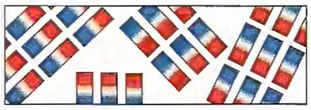

In a non-magnetized state, ***domains*** *are jumbled. The overall effect is that the domains cancel each other out.*

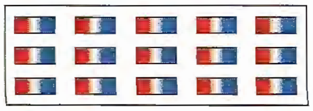

In a magnetized state, ***domains*** *are ordered. If ordered completely (as here) magnet is* ***saturated*** *– cannot become stronger.*

Magnetization

When an object is magnetized, all the **dipoles** become aligned (see **domain theory**). This only happens when an object is in a **magnetic field*** and is called **induced magnetism**.

Induced magnetism

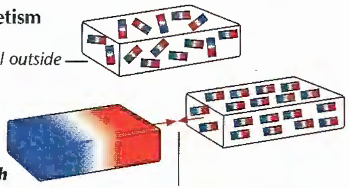

Magnetic material outside ***magnetic field*.***

North end of ***dipoles*** *attracted to* ***south pole*** *of magnet – object becomes magnetized.*

Magnetic force* *always attracts.*

Single touch

A method of magnetizing an object by stroking it repeatedly with the **pole** of a **permanent magnet** (see **hard**). Magnetism is induced in the object from the **magnetic field*** of the magnet.

Magnetism induced by single touch

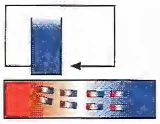

Divided touch

A method of magnetizing an object by stroking it repeatedly from the center out with the opposite **poles** of two **permanent magnets** (see **hard**). Magnetism is induced in the object from the **magnetic field*** of the magnets.

Magnetism induced by divided touch

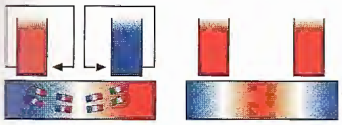

Consequent poles *are produced when like poles are used in* ***divided touch.***

Demagnetization

The removal of magnetism from an object. This can be achieved by placing the object in a changing **magnetic field***, such as that created by a coil carrying **alternating current***. Alternatively, the **dipoles** (see **domain theory**, above) can be excited to point in random directions by hammering randomly or by heating above 700°C.

Self-demagnetization

Loss of magnetism by a magnet because of the attraction of the **dipoles** (see **domain theory**) for the opposite **poles** of the magnet. It is reduced using pieces of soft iron (called **keepers**) arranged to form a closed loop of poles.

Self-demagnetization of bar magnet

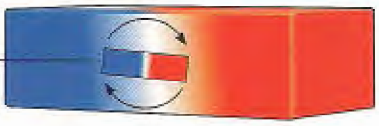

Dipoles *tend to turn.*

Reducing self-demagnetization

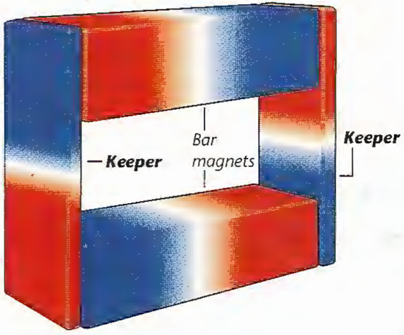

Poles *induced in keepers attract* ***dipoles.***

---

* **Alternating current**, 61; **Magnetic field**, 72; **Magnetic force**, 6.

MAGNETIC FIELDS

A **magnetic field** is a region around a **magnet** (see page 70) in which objects are affected by the **magnetic force***. The strength and direction of the magnetic field are shown by **magnetic field lines**.

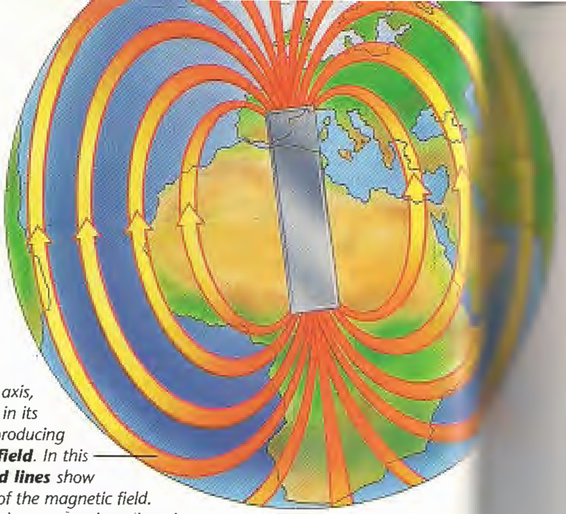

*As the Earth rotates on its axis, molten metal in its core moves, producing a **magnetic field**. In this diagram, **field lines** show the direction of the magnetic field. The lines are closest near the poles where the field is strongest.*

Magnetic field lines or flux lines

Lines which indicate the direction of the magnetic field around a magnet. They also show the strength of the field (see **magnetic flux density**, below). The direction of the field is the direction of the force on a **north pole***. Magnetic field lines are plotted by sprinkling iron filings around a magnet or by recording the direction of a **plotting compass** (a small compass with no directions marked on it) at various points.

Result of sprinkling iron filings around a magnet

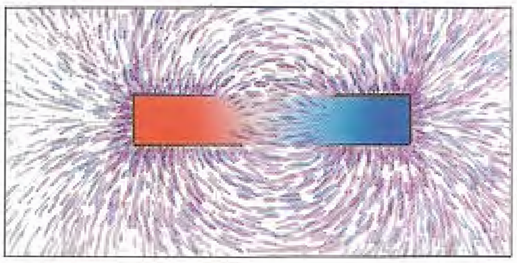

*Iron filings line up due to **induced magnetism***.*

Plotting the magnetic field lines around a bar magnet

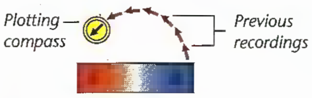
Plotting compass
Previous recordings

Complete plot of magnetic field lines
(Arrows indicate direction)

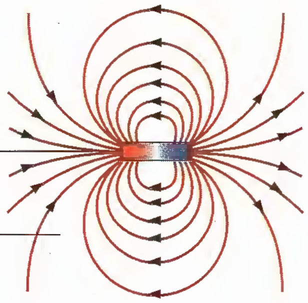
*High **magnetic flux density** – strong magnetic field*
*Low **magnetic flux density** – weaker magnetic field*

Magnetic flux density

A measurement of the strength of a magnetic field at a point. This is shown by the closeness of the **magnetic field lines** to each other. Magnetic flux density is normally highest around the **poles***.

Neutral point

A point of zero magnetism (the **magnetic flux density** is zero). It occurs where two or more magnetic fields interact with an equal but opposite effect. A bar magnet positioned along the **magnetic meridian**, with the **south pole*** pointing to the north, has two neutral points in line with its **magnetic axis***.

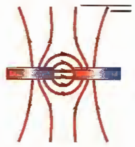
Magnetic field lines** created by opposite **poles placed together*

*No **neutral point***

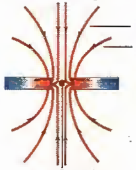
Magnetic field lines** created by like **poles placed together*

Neutral point

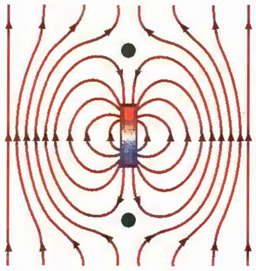

Neutral points** (marked with a dot) created by magnet with **south pole pointing north*

Diamagnetism

Magnetism displayed by some substances when placed in a strong magnetic field. A piece of diamagnetic material tends to spread **magnetic field lines** out and lines up with its long side perpendicular to them.

Paramagnetism

Magnetism displayed by some substances when placed in a strong magnetic field. A piece of paramagnetic material tends to concentrate **magnetic field lines** through it and lines up with its long side parallel to them. It is caused by **dipoles*** moving slightly towards alignment.

---

* **Dipole**, 71 (**Domain theory of magnetism**); **Electrons**, 83; **Induced magnetism**, 71; **Magnetic force**, 6; **Magnetic axis**, **Pole**, 70.

The Earth's m

The Earth has a magn
though there were a
center, lined up appr
geographic north and
the angle is constant
pole of a compass po
called **magnetic nort**
magnetic south.

Section through Earth's magnetic field

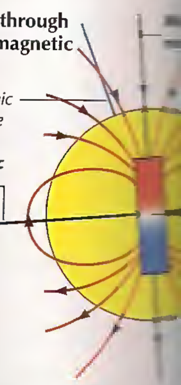
Geographic north pole
Magnetic equator

Magnetic meridian

The vertical plane
axis* of a magnet
magnetic field (i.e.
pointing to **magnetic**

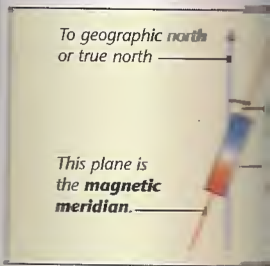
To geographic north or true north
*This plane is the **magnetic meridian.***

Permeability

A measure of the
"conduct" a magnet
more permeable
tends to be concent

*Soft iron has a higher **permeability** than air.*

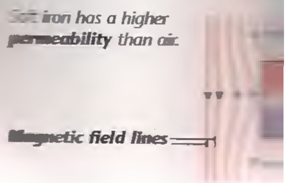
Magnetic field lines

Magnetic axis, **Pole** 7

The Earth's magnetism

The Earth has a magnetic field which acts as though there were a giant bar magnet in its center, lined up approximately between its geographic north and south poles, although the angle is constantly changing. The north pole of a compass points towards a point called **magnetic north**, its south pole to **magnetic south**.

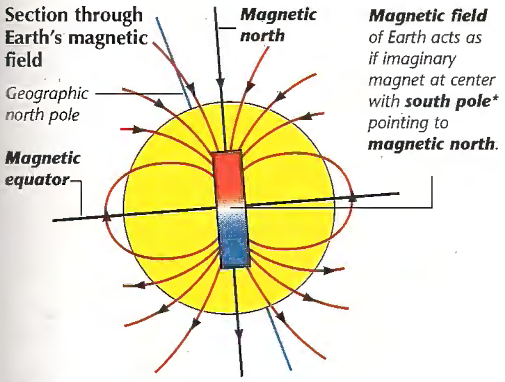
Section through Earth's magnetic field
Geographic north pole
Magnetic equator
Magnetic north
Magnetic field of Earth acts as if imaginary magnet at center with south pole* pointing to magnetic north.

Magnetic meridian

The vertical plane containing the **magnetic axis*** of a magnet suspended in the Earth's magnetic field (i.e. with its **north pole*** pointing to **magnetic north**).

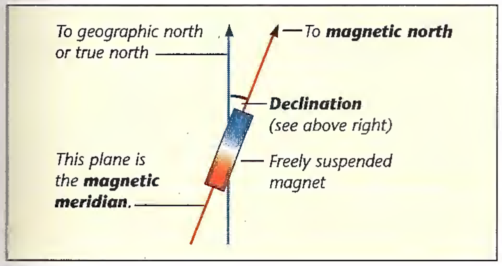
To geographic north or true north
To magnetic north
Declination (see above right)
This plane is the magnetic meridian.
Freely suspended magnet

Declination

The angle between a line taken to true north (the geographic north pole) and one taken along the **magnetic meridian** (towards **magnetic north**) at a point. The position of magnetic north is gradually changing and so the declination alters slowly with time.

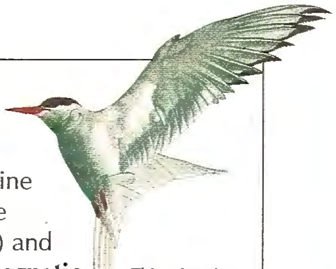

This migrating tern may use the Earth's magnetic field to guide it.

Isogonal lines

Lines joining places with equal **declination**. These are redrawn from time to time because of the changing direction of the Earth's magnetic field.

Inclination or dip

The angle between a horizontal line on the Earth's surface and the direction of the Earth's magnetic field at a point. It is measured using a **dip circle** (see picture, below).

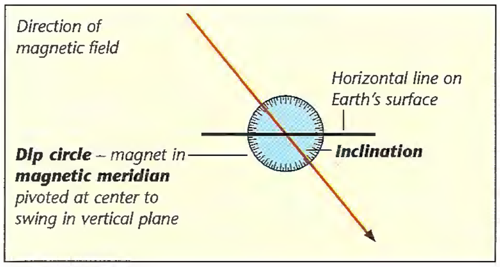
Direction of magnetic field
Horizontal line on Earth's surface
Dip circle – magnet in magnetic meridian pivoted at center to swing in vertical plane
Inclination

Isoclinal line

A line linking places with the same **inclination**.

Permeability

A measure of the ability of a substance to "conduct" a magnetic field. Soft iron is much more permeable than air, so the magnetic field tends to be concentrated through it.

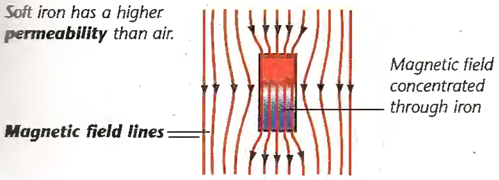
Soft iron has a higher permeability than air.
Magnetic field lines
Magnetic field concentrated through iron

Shielding or screening

The use of soft magnetic material to stop a magnetic field from reaching a point, effectively by "conducting" the field away. This is used in sensitive instruments, e.g. oscilloscopes.

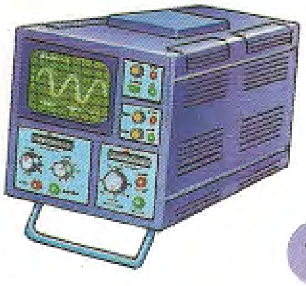

*Oscilloscope beams are **shielded** from unwanted magnetic fields by **mumetal**, a special alloy with a very high **permeability**.*

---

* **Magnetic axis**, **Pole**, 70.

ELECTROMAGNETISM

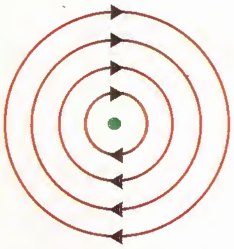

Cross-section of the **magnetic field** which would exist if there was a wire (shown in green) carrying current directly into the page.

An electric current flowing through a wire produces a **magnetic field** (see pages 72-73) around the wire, the shape of which depends on the shape of the wire and the current flowing. These magnetic fields can be plotted in the same way as for **permanent magnets***. This effect, called **electromagnetism**, is used in very powerful magnets and also to produce motion from an electric current.

Maxwell's screw rule

States that the direction of the magnetic field around a current-carrying wire is the way a screw turns when being screwed in the direction of the current.

Maxwell's screw rule

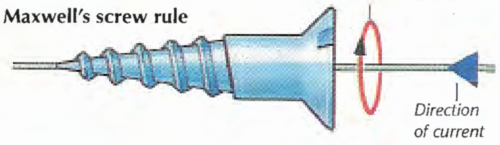
Direction of field
Direction of current

Right-hand grip rule

States that the direction of a magnetic field around a wire is that from the base to the tips of the fingers if the wire is gripped by the right hand with the thumb pointing in the direction of the current.

Right-hand grip rule

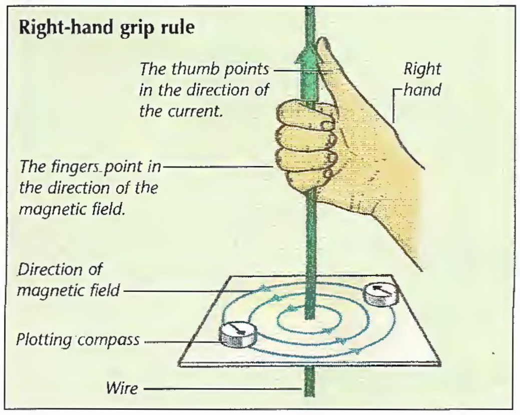
The thumb points in the direction of the current.
Right hand
The fingers point in the direction of the magnetic field.
Direction of magnetic field
Plotting compass
Wire

Coil

A number of turns of current-carrying wire, produced by wrapping the wire around a shaped piece of material (a **former**). Examples are a **flat coil** and a **solenoid**.

Flat coil or plane coil

A **coil** of wire whose length is small in comparison with its diameter.

Solenoid

A **coil** whose length is large in comparison with its diameter. The magnetic field produced by a solenoid is similar to that produced by a bar magnet. The position of the **poles*** depends on the current direction.

Solenoid

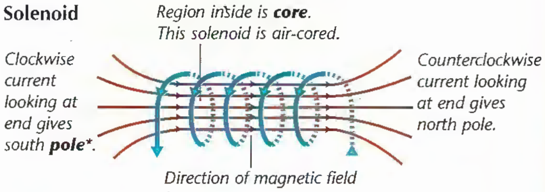
Region inside is **core**.
This solenoid is air-cored.
Clockwise current looking at end gives south **pole***.
Counterclockwise current looking at end gives north pole.
Direction of magnetic field

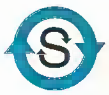

Arrowheads on an S point in a clockwise direction (clockwise current = south pole).

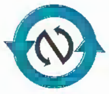

Arrowheads on an N point in an counterclockwise direction (counterclockwise current = north pole).

Core

The material in the center of a **coil** which dictates the strength of the field. Soft **ferromagnetic*** materials, most commonly soft iron, create the strongest magnetic field and are used in **electromagnets**.

Electromagnet

A **solenoid** with a **core** of soft, strongly **ferromagnetic*** material. This forms a magnet which can be switched on and off simply by turning the current on and off. Practical electromagnets are constructed so that two opposite **poles*** are close to each other, producing a strong magnetic field.

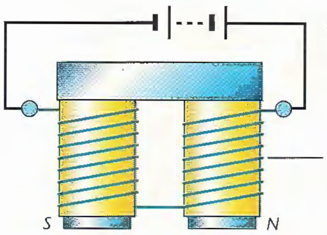

***Electromagnet** formed from two **solenoids** with iron **cores** and iron piece between ends.*

*Wire wound in opposite directions in each solenoid to produce opposite **poles***.*

---

* **Ferromagnetic**, 70; **Permanent magnets**, 70 (**Hard**); **Pole**, 70.

Applications of

Electromagnets have a la
applications, all of which
they attract metals when
on and therefore convert
to **mechanical energy***. I
following examples, soun
produced from the mecha

Electric buzzer

A device which produces
from **direct current***. A m
by an **electromagnet**, mo
in doing so breaks the cir
to the electromagnet. The
and the process is repeate
vibration of the arm produ
noise. In the **electric bell**,
to the arm repeatedly strik

Electric bell

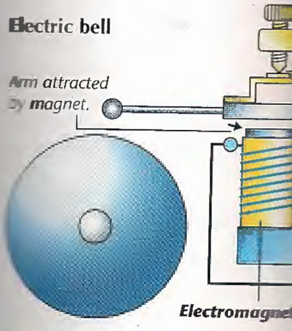
Arm attracted by magnet.
Electromagne

Earphone

A device used to transform
to sound waves. The **perm**
attracts the metal diaphrag
of this attraction is change
current (the incoming signa
the coils of the **electromag**
thus vibrates to produce so

Earphone

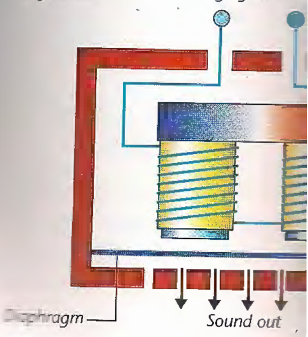
Changing current
Diaphragm
Sound out

Direct current, 61; **Electric energy**
Permanent magnets, 70 (**Hard**); **Po**

Applications of electromagnets

Electromagnets have a large number of applications, all of which use the fact that they attract metals when they are switched on and therefore convert **electric energy*** to **mechanical energy***. In two of the following examples, sound energy is produced from the mechanical energy.

Electric buzzer

A device which produces a buzzing noise from **direct current***. A metal arm is attracted by an **electromagnet**, moves towards it, and in doing so breaks the circuit carrying current to the electromagnet. The arm is thus released and the process is repeated. The resulting vibration of the arm produces a buzzing noise. In the **electric bell**, a hammer attached to the arm repeatedly strikes a bell.

Electric bell

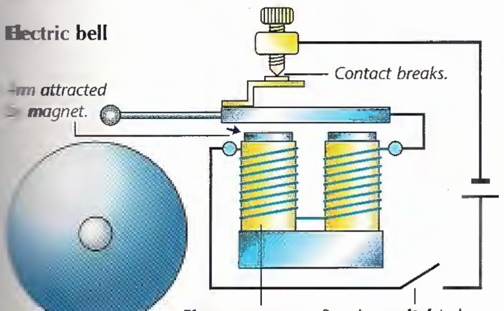
Contact breaks.
Arm attracted to magnet.
Electromagnet

*Pressing **switch*** closes circuit to activate magnet.*

Earphone

A device used to transform electrical signals to sound waves. The **permanent magnet*** attracts the metal diaphragm, but the strength of this attraction is changed as changing current (the incoming signals) flows through the coils of the **electromagnet**. The diaphragm thus vibrates to produce sound waves.

Earphone

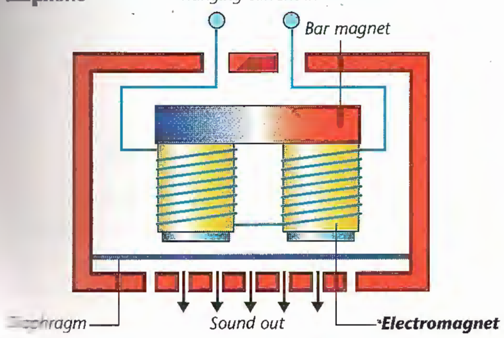
Changing current in
Bar magnet
Diaphragm
Sound out
Electromagnet

Lifting magnets

Large **electromagnets** which are used in steelworks to lift heavy loads. The activated electromagnet attracts steel, enabling it to be moved. The load is released when the current is switched off.

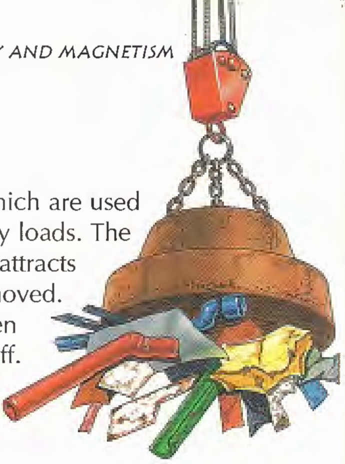

***Lifting magnet** used to move scrap metal from one place to another.*

Relay

A device in which a **switch*** is closed by the action of an **electromagnet**. A relatively small current in the **coil** of the electromagnet can be used to switch on a large current without the circuits being electrically linked.

Relay

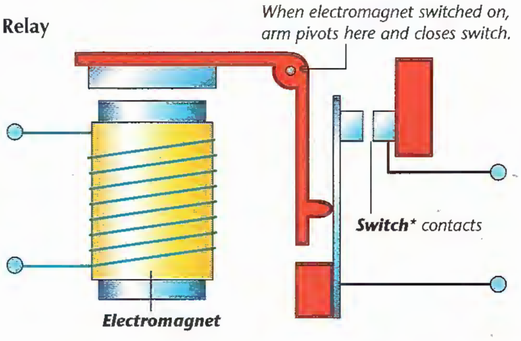
When electromagnet switched on, arm pivots here and closes switch.
Switch* contacts
Electromagnet

"Maglev" train

A train with **electromagnets** attached underneath, which runs on tracks with electromagnets on them. The magnets repel each other, so the train hovers just above the track. The reduced friction between the train and the track means that the train can travel faster.

Japanese "maglev" train

*The word "maglev" comes from **mag**netic **lev**itation.*

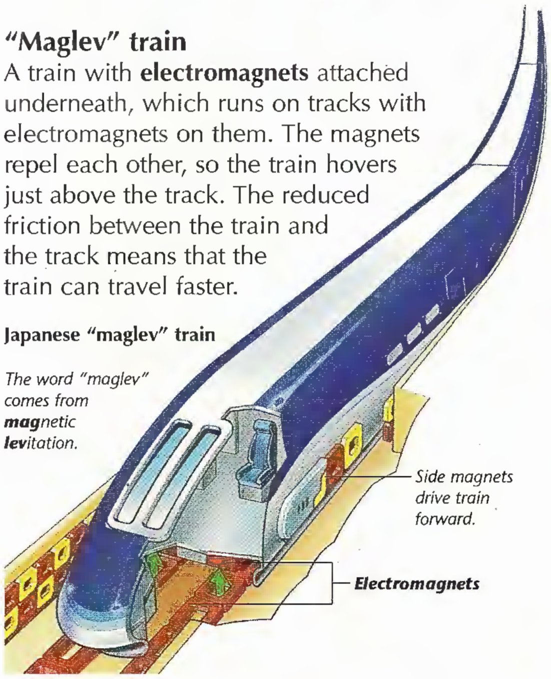
Side magnets drive train forward.
Electromagnets

Direct current, 61; **Electric energy**, 9; **Ferromagnetic**, 70; **Mechanical energy**, 9; **Permanent magnets**, 70 (**Hard**); **Pole**, 70; **Switch**, 64.

Electromagnets continued – the Lorentz force

The **Lorentz force** occurs when a current-carrying wire goes through a magnetic field. A force acts on the wire which can produce movement. This effect is used in **electric motors**, where **mechanical energy*** is produced from **electric energy***. The effect can also be used to measure current (see page 77), since the force depends on its magnitude.

Fleming's left-hand rule. See diagram, right. A right-hand rule can also be used. The thumb indicates force, but the meaning of the index and first fingers is reversed.

Fleming's left-hand rule

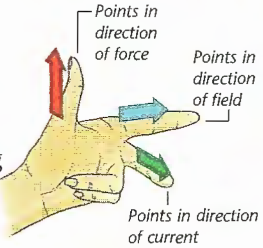
Points in direction of force
Points in direction of field
Points in direction of current

Electric motor

A device which uses the **Lorentz force** to transform **electric energy*** to **mechanical energy***. The simplest motor consists of a current-carrying, square-shaped **flat coil***, free to rotate in a magnetic field (see diagram below). Motors produce a **back e.m.f.*** opposing the e.m.f. which drives them. This is produced because once the motor starts, it acts as a **generator*** (i.e. the movement of the coil in the field produces an opposing current).

Simple electric motor

Commutator**. A ring split into two or more pieces, via which current enters and leaves the **coil of an **electric motor**. It ensures that the current enters the coil in the correct direction to make the motor rotate in one direction continuously.*

***Brushes**. Contacts, normally made of carbon, through which current enters the **commutator** in an **electric motor**.*

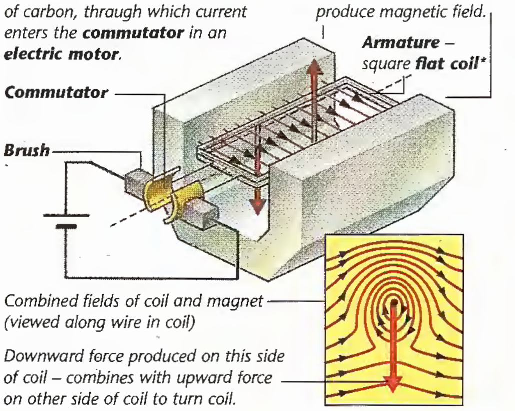
Poles of horseshoe magnet produce magnetic field.
Armature – square flat coil*
Commutator
Brush
Combined fields of coil and magnet (viewed along wire in coil)
Downward force produced on this side of coil – combines with upward force on other side of coil to turn coil.

Exploded view of a powerful electric motor

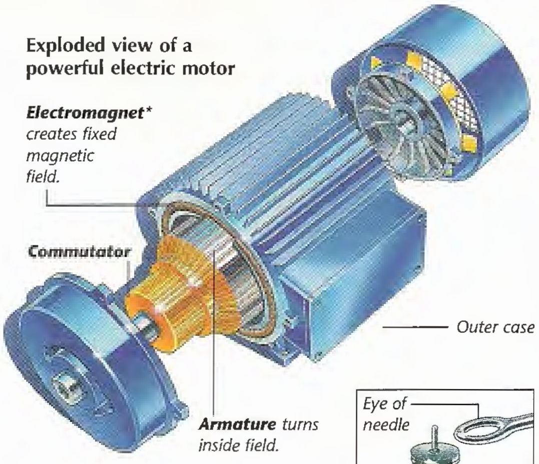
Electromagnet* creates fixed magnetic field.
Commutator
Outer case
Armature turns inside field.

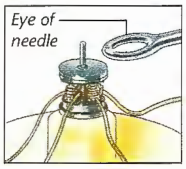
Eye of needle

Right: a Toshiba micro-motor 0.8mm (0.03in) wide.

Field windings

Sets of **coils*** around the outside of an **electric motor**, which take the place of a permanent magnet to produce a stronger magnetic field. This increases the power of the motor.

Loudspeaker

A device which uses the **Lorentz force** to transform electrical signals into **sound waves***. It consists of a **coil*** in a **radial magnetic field** (the direction of the field at any point is along a radius of this coil). As the current changes, the coil, which is attached to a paper cone, moves in and out of the field (see diagram). The paper cone vibrates the air, producing sound waves which depend on the strength and frequency of the current.

Loudspeaker

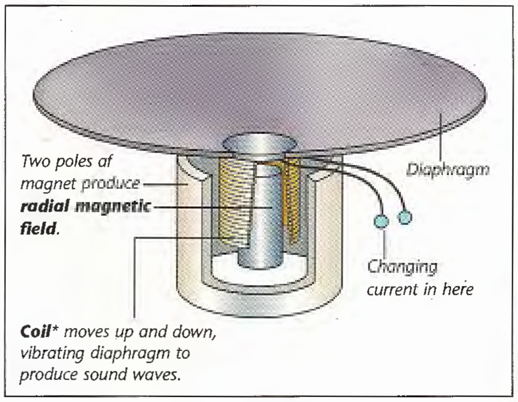
Two poles of magnet produce radial magnetic field.
Diaphragm
Changing current in here
Coil* moves up and down, vibrating diaphragm to produce sound waves.

---

*Back e.m.f., 60 (**Electromotive force**); **Coil**, 74; **Electric energy**, 9; **Electromagnet**, **Flat coil**, 74; **Generator**, 78; **Mechanical energy**, 9; **Sound waves**, 40.

ELECTRIC

Current can be det
a suspended magn
and observing its d
idea can be extend
a device (a **meter**
deflection indicates
strength of the curre
measuring device c
adapted to measure
difference*.

Galvanometer

Any device used to dete
by registering its magne
is a compass placed near
whether a current is pres
galvanometer uses the **L**
a deflection on a scale

Moving coil galvanometer

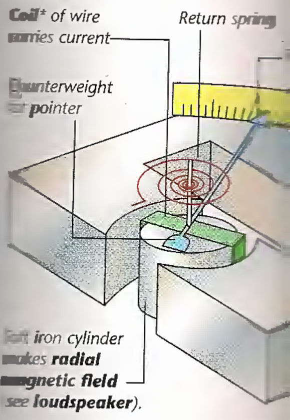
Coil* of wire carries current
Return spring
Counterweight pointer
iron cylinder makes radial magnetic field (see loudspeaker).

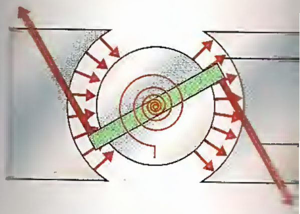

Moving iron meter

A **meter** in which the curre
induces magnetism in two
which attract or repel each
a deflection.

***Coil**, 74; **Direct current**, 61; **Pote**
Resistance, 62; **Series**, 64.

ELECTRIC METERS

Current can be detected by placing a suspended magnet near a wire and observing its deflection. This idea can be extended to produce a device (a **meter**) in which the deflection indicates on a scale the strength of the current. The current measuring device can then be adapted to measure **potential difference***.

Galvanometer

Any device used to detect a **direct current*** by registering its magnetic effect. The simplest is a compass placed near a wire to show whether a current is present. The **moving coil galvanometer** uses the **Lorentz force** to show a deflection on a scale (see diagram).

Moving coil galvanometer

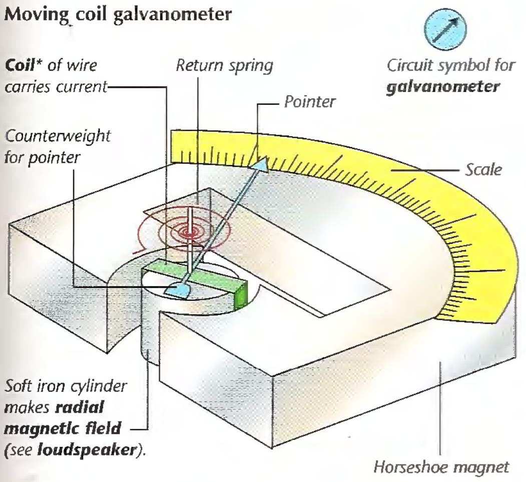
Coil* of wire carries current
Return spring
Circuit symbol for **galvanometer**
Pointer
Counterweight for pointer
Scale
Soft iron cylinder makes **radial magnetic field** (see **loudspeaker**).
Horseshoe magnet

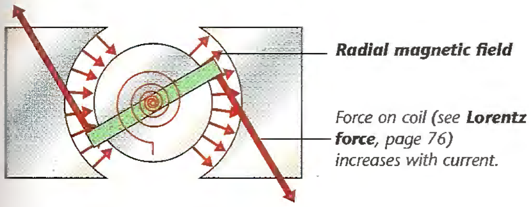
Radial magnetic field
Force on coil (see **Lorentz force**, page 76) increases with current.

Moving iron meter

A **meter** in which the current to be measured induces magnetism in two pieces of iron which attract or repel each other to produce a deflection.

Ammeter

A device used to measure current. It is a version of the **moving coil galvanometer**, designed so that a certain current produces a **full scale deflection**, i.e. the pointer moves to its maximum position. To measure higher currents, a **shunt** is added (see diagram below). A larger current now produces full scale deflection on the new scale.

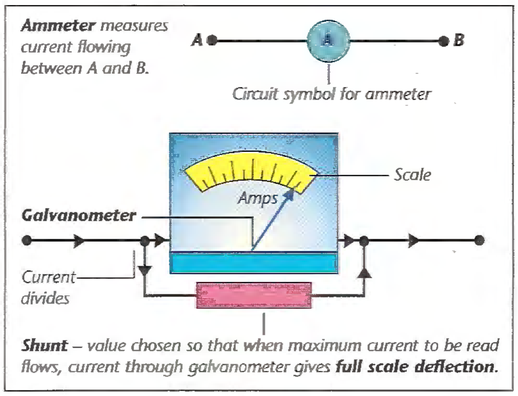
Ammeter measures current flowing between A and B.
A
A
B
Circuit symbol for ammeter
Scale
Amps
Galvanometer
Current divides
Shunt – value chosen so that when maximum current to be read flows, current through galvanometer gives **full scale deflection**.

Voltmeter

A device used to measure the **potential difference*** between two points. It is a **galvanometer** between the two points with a high **resistance*** in **series***. A certain potential difference produces the current for a **full scale deflection** (see **ammeter**). To measure higher potential differences, a **multiplier** is added (see diagram below).

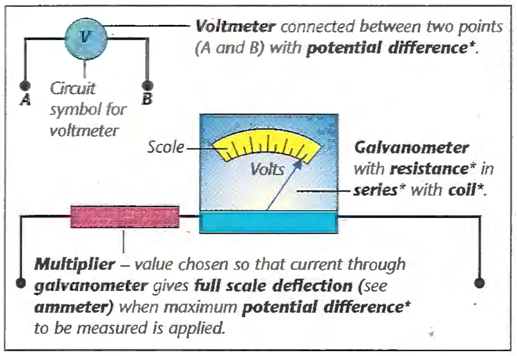
Voltmeter connected between two points (A and B) with **potential difference***.
V
A
B
Circuit symbol for voltmeter
Scale
Volts
Galvanometer with **resistance*** in **series*** with **coil***.
Multiplier – value chosen so that current through **galvanometer** gives **full scale deflection** (see **ammeter**) when maximum **potential difference*** to be measured is applied.

Multimeter

A **galvanometer** combined with the **shunts** (see **ammeter**) and **multipliers** (see **voltmeter**) necessary to measure currents and **potential differences***.

---

* **Coil**, 74; **Direct current**, 61; **Potential difference**, 58; **Resistance**, 62; **Series**, 64.

ELECTROMAGNETIC INDUCTION

Michael Faraday found that, as well as a current passing through a magnetic field producing movement (see **motor effect**, page 76), movement of a **conductor*** in a magnetic field produces an **electromotive force*** in the conductor. This effect, called **electromagnetic induction**, happens whenever a conductor is placed in a changing magnetic field.

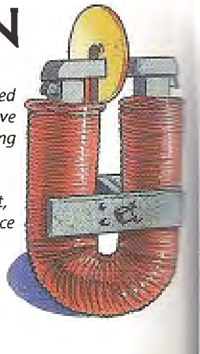

*Faraday induced an electromotive force by rotating a disk (the **conductor***) near a magnet, using this device called a disk **dynamo**.*

Faraday's law of induction

States that the size of an induced electromotive force in a **conductor*** is proportional to the rate at which the magnetic field changes.

Lenz's law

States that an induced electromotive force always acts to oppose the cause of it, e.g. in an **electric motor***, the e.m.f. produced because it acts as a **generator** opposes the e.m.f. driving the motor.

Fleming's right-hand rule or dynamo rule

The direction of an induced current can be worked out from the direction of the magnetic field and the movement by using the right hand (see diagram).

Fleming's right-hand rule

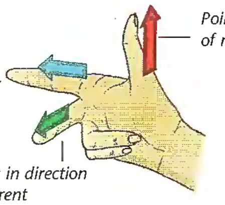
Points in direction of motion
Points in direction of force
Points in direction of current

Generator or dynamo

A device used to produce electric current from **mechanical energy***. In the simplest generator (see diagram, right), an alternating electromotive force is induced in a **coil*** as it rotates in a magnetic field. A generator for **direct currents*** has a **commutator***, as on an **electric motor***, which means the current always flows in the same direction.

Simple generator

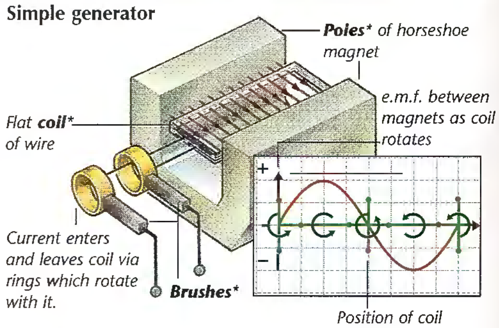
Poles* of horseshoe magnet
e.m.f. between magnets as coil rotates
Flat **coil*** of wire
+
–
Current enters and leaves coil via rings which rotate with it.
Brushes*
Position of coil

*A bicycle **dynamo** contains a **coil*** of wire that spins between two magnets.*

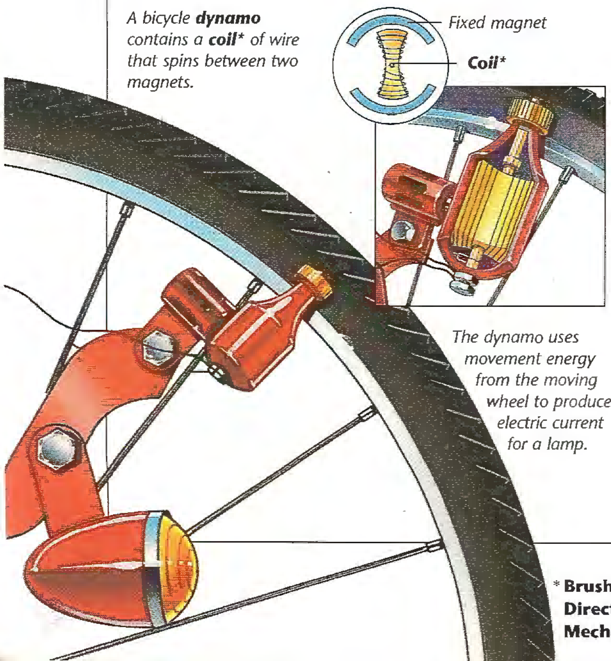
Fixed magnet
Coil*

The dynamo uses movement energy from the moving wheel to produce electric current for a lamp.

Mutual induction

The induction of an electromotive force in a **coil*** of wire by changing the current in a different coil. The changing current produces a changing magnetic field which induces a current in any other coil in the field. This was first demonstrated with **Faraday's iron ring**.

Faraday's iron ring

*Closing or opening **switch*** causes change in magnetic field in ring, which induces current in secondary circuit.*

Soft iron ring – "conducts" fields between coils.

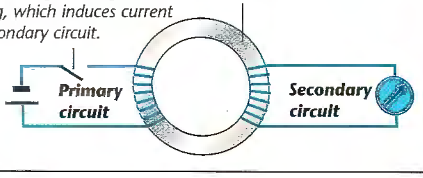
Primary circuit
Secondary circuit

* **Brushes**, 76; **Coil**, 74; **Commutator**, 76; **Conductor**, 56; **Direct current**, 61; **Electric motor**, 76; **Electromotive force** (**e.m.f.**), 60; **Mechanical energy**, 9; **Pole**, 70; **Switch**, 64.

Self-induction

The induction of an electromotive force in a **coil*** of wire due to the current inside it changing. For example, if the current in a coil is switched off, the resulting change in the magnetic field produces an electromotive force across the coil, in some cases much higher than that of the original.

Eddy current

A current set up in a piece of metal when a magnetic field around it changes, even though the metal may not be part of a circuit. Eddy currents can cause unwanted heat energy, e.g. in the iron core of a **transformer**. This can be prevented by laminating the iron core (see **transformers**, below).

Transformers

A **transformer** consists of two **coils*** of wire wound onto the same **core*** of soft **ferromagnetic*** material. It is used to change an alternating electromotive force in one of the coils to a different e.m.f. in the other coil, e.g. in electricity supply, see page 61. Hardly any energy is lost between the two circuits in a well-designed transformer.

Simple transformer

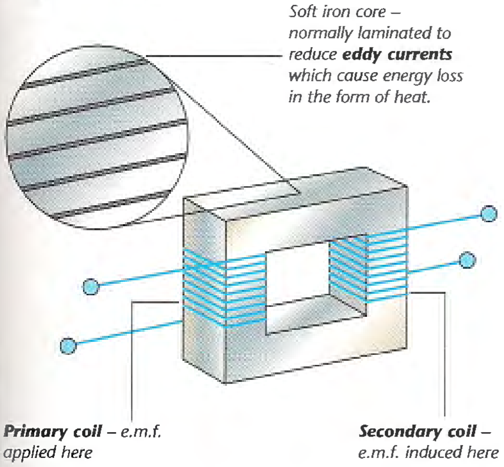
Soft iron core – normally laminated to reduce **eddy currents** which cause energy loss in the form of heat.
Primary coil – e.m.f. applied here
Secondary coil – e.m.f. induced here

Primary coil

The **coil*** in a transformer to which an alternating electromotive force is applied in order to produce an electromotive force in the **secondary coil**.

Secondary coil

The **coil*** in a transformer in which an alternating electromotive force is induced by the electromotive force applied to the **primary coil**. Some transformers have two or more secondary coils.

Turns ratio

The ratio of the number of turns in the **secondary coil** in a **transformer** to the number of turns in the **primary coil**. The turns ratio is also the ratio of the electromotive force in the secondary coil to that in the primary coil.

$$\text{Turns ratio} = \frac{N_2}{N_1} = \frac{V_2}{V_1}$$

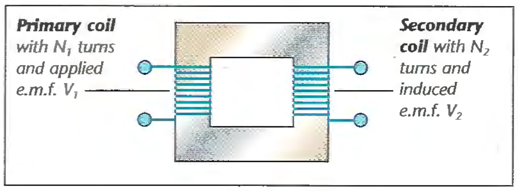
Primary coil with N_1 turns and applied e.m.f. V_1
Secondary coil with N_2 turns and induced e.m.f. V_2

Step-up transformer

A **transformer** in which the electromotive force in the **secondary coil** is greater than that in the **primary coil**. The **turns ratio** is greater than one.

Step-up transformer

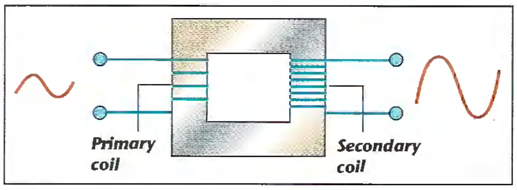
Primary coil
Secondary coil

Step-down transformer

A **transformer** in which the electromotive force in the **secondary coil** is less than that in the **primary coil**. The **turns ratio** is less than one.

Step-down transformer

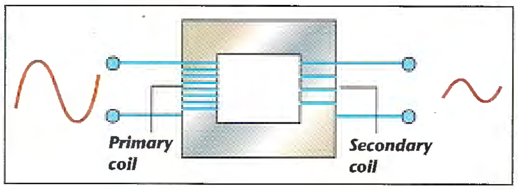
Primary coil
Secondary coil

---

* **Coil**, **Core**, 74; **Ferromagnetic**, 70.

CATHODE RAYS

A **cathode ray** is a continuous stream of **electrons** (negatively-charged particles – see page 83) traveling through a low pressure gas or a vacuum. It is produced when electrons are freed from a metal **cathode*** and attracted to an **anode***. Cathode rays have a number of applications, from the production of **X-rays*** to **television**. All of these involve the use of a shaped glass tube (called an **electron tube**) containing a low pressure gas or a vacuum for the rays to travel in. The rays are normally produced by an **electron gun**, which forms part of the tube.

Electron gun

A device which produces a continuous stream of electrons (a cathode ray). It consists of a heated **cathode*** which gives off electrons (this is called **thermionic emission**) and an **anode*** which attracts them to form a stream.

Typical experimental electron tube

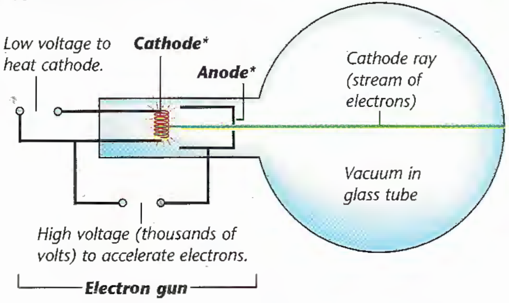
Low voltage to heat cathode.
Cathode*
Anode*
Cathode ray (stream of electrons)
Vacuum in glass tube
High voltage (thousands of volts) to accelerate electrons.
Electron gun

Maltese cross tube

An **electron tube** in which the cathode ray is interrupted by a cross which casts a "shadow" on a **fluorescent*** screen at the end of the tube. This shows that the electrons are moving in straight lines.

Maltese cross tube

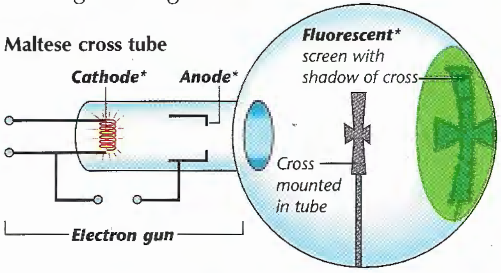
Cathode*
Anode*
Fluorescent* screen with shadow of cross
Cross mounted in tube
Electron gun

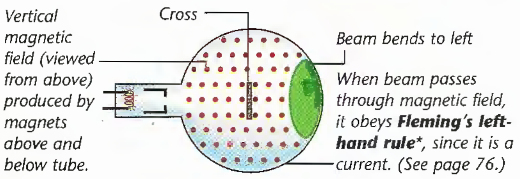
Vertical magnetic field (viewed from above) produced by magnets above and below tube.
Cross
Beam bends to left

When beam passes through magnetic field, it obeys ***Fleming's left-hand rule****, since it is a current. (See page 76.)

Discharge tube

A gas-filled glass tube in which **ions*** and electrons are attracted by the **electrodes*** and move towards them at high speed. As they do so, they collide with gas atoms, causing these atoms to split into more ions and electrons, and emit light at the same time. The color of the light depends on the gas used, e.g. neon produces orange light (used in advertising displays) and mercury vapor produces blue-green light (used for street lighting). Discharge tubes use up to five times less electricity than other lighting. A **fluorescent tube** is a discharge tube filled with mercury vapor, which emits **ultraviolet radiation***. This hits the inside of the tube, causing its coating of special powder to give out **visible light*** (see **fluorescence**, page 45).

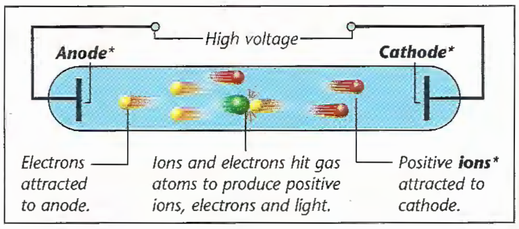
Anode*
High voltage
Cathode*
Electrons attracted to anode.
Ions and electrons hit gas atoms to produce positive ions, electrons and light.
Positive **ions*** attracted to cathode.

X-ray tube

A special electron tube used to produce a beam of **X-rays***. A cathode ray hits a tungsten target which stops the electrons suddenly. This causes X-rays to be emitted.

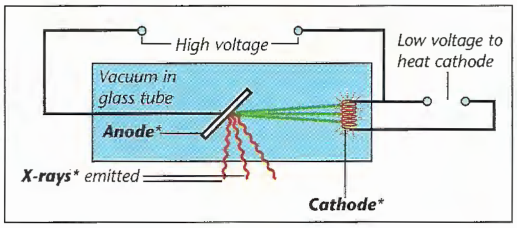
High voltage
Low voltage to heat cathode
Vacuum in glass tube
Anode*
X-rays* emitted
Cathode*

---

* **Anode**, **Cathode**, 66 (**Electrode**); **Fleming's left-hand rule** 76; **Fluorescence**, 45; **Ions**, 88 (**Ionization**); **Ultraviolet radiation**, 44; **Visible light**, 45; **X-rays**, 44.

The cathode ra

The **cathode ray oscillos**
instrument used to study
differences*. A cathode ra
gun produces a spot on a
In normal use, the ray is
across the back of the sc

Components of oscilloscope

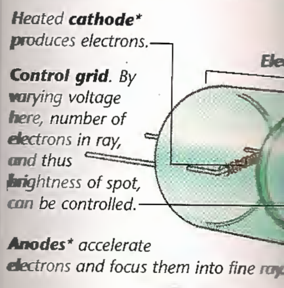
Heated **cathode*** produces electrons.
Control grid. By varying voltage here, number of electrons in ray, and thus brightness of spot, can be controlled.
Ele
Anodes* accelerate electrons and focus them into fine ray

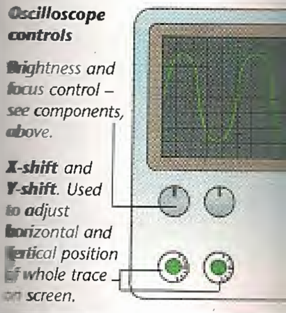
Oscilloscope controls
Brightness and **focus control** – see components, above.
X-shift and **Y-shift**. Used to adjust horizontal and vertical position of whole trace on screen.

Television

Television pictures are repr
an **electron tube** in which t
scans across the screen vary
according to the signal. Dif
light, according to the stren
given off from different part
produce a picture (see diag

Pixels (see diagram on right)

Anode, **Cathode**, 66 (**Electrode**); **F**
Potential difference, 58; **Primary**

The cathode ray oscilloscope

The **cathode ray oscilloscope** (**CRO**) is an instrument used to study currents and **potential differences***. A cathode ray from an **electron gun** produces a spot on a **fluorescent*** screen. In normal use, the ray is repeatedly swept across the back of the screen at a selected speed and so produces a visible trace across the front. If a signal is fed into the oscilloscope, the vertical position of the beam will change according to the strength of the signal, and the trace on the screen then shows this change over time.

Components of oscilloscope

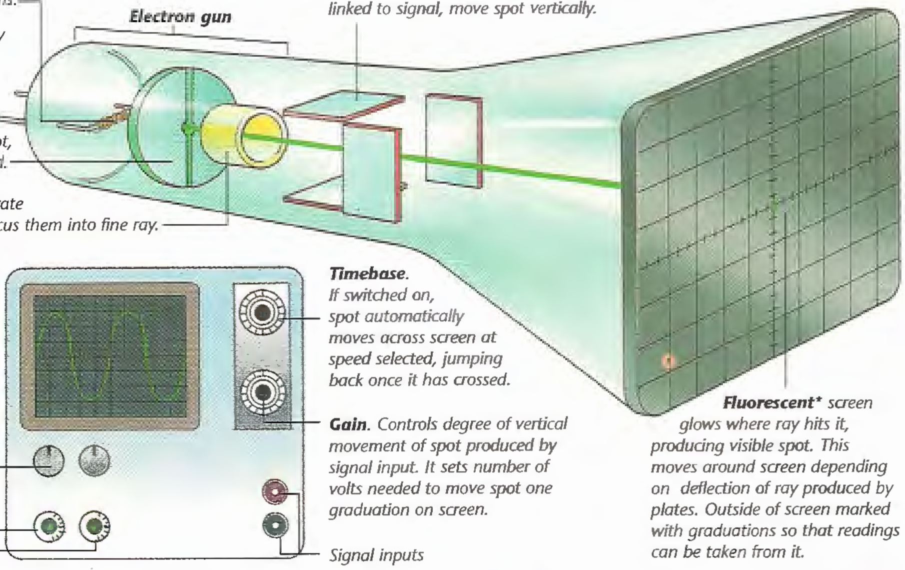

*Heated **cathode**** *produces electrons.*

***Control grid**. By varying voltage here, number of electrons in ray, and thus brightness of spot, can be controlled.*

Anodes* *accelerate electrons and focus them into fine ray.*

***Deflection system**. Two sets of plates which control position of spot on screen. **X-plates** used to move spot horizontally across screen under control of **timebase** (see below) and **Y-plates**, linked to signal, move spot vertically.*

Fluorescent* *screen glows where ray hits it, producing visible spot. This moves around screen depending on deflection of ray produced by plates. Outside of screen marked with graduations so that readings can be taken from it.*

Oscilloscope controls

Brightness and focus control – see components, above.

***X-shift** and **Y-shift**. Used to adjust horizontal and vertical position of whole trace on screen.*

***Timebase**. If switched on, spot automatically moves across screen at speed selected, jumping back once it has crossed.*

***Gain**. Controls degree of vertical movement of spot produced by signal input. It sets number of volts needed to move spot one graduation on screen.*

Signal inputs

Television

Television pictures are reproduced by using an **electron tube** in which the cathode ray scans across the screen varying in strength according to the signal. Different levels of light, according to the strength of the ray, are given off from different parts of the screen to produce a picture (see diagram on right).

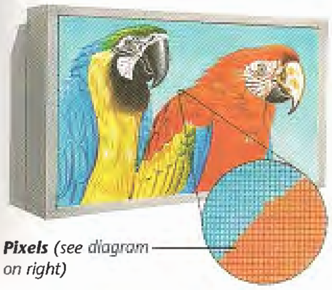

*Extra large high-definition televisions (HTDVs) have more **pixels** than ordinary TVs (see right), so the picture is sharper.*

***Pixels** (see diagram on right)*

Cutaway television cathode ray tube

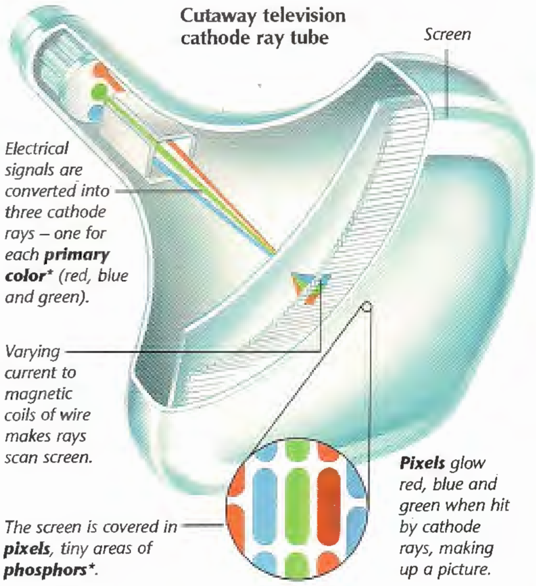

*Electrical signals are converted into three cathode rays – one for each **primary color*** (red, blue and green).*

Varying current to magnetic coils of wire makes rays scan screen.

*The screen is covered in **pixels**, tiny areas of **phosphors***.*

***Pixels** glow red, blue and green when hit by cathode rays, making up a picture.*

---

* **Anode**, **Cathode**, 66 (**Electrode**); **Fluorescence**, 45; **Phosphors**, 44 (**Phosphorescence**); **Potential difference**, 58; **Primary colors**, 55.

ATOMIC STRUCTURE

A great deal has been learned about the physical nature of atoms (see also page 4) since Greek philosophers first proposed that all matter was made of basic indivisible "building blocks". It is now known that an atom is not indivisible, but has a complex internal structure, consisting of many different smaller particles (**subatomic particles**) and a lot of empty space.

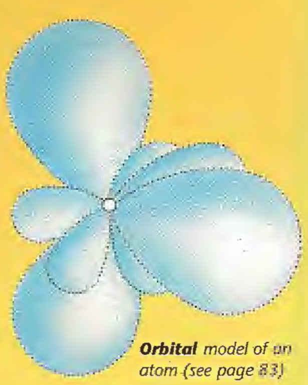

***Orbital** model of an atom (see page 83)*

Rutherford-Bohr atom

A "solar system" representation of an atom, devised by Ernest Rutherford and Niels Bohr in 1911. It is now known to be incorrect (**electrons** have no regular "orbits" – see **electron shells**).

Rutherford-Bohr atom model

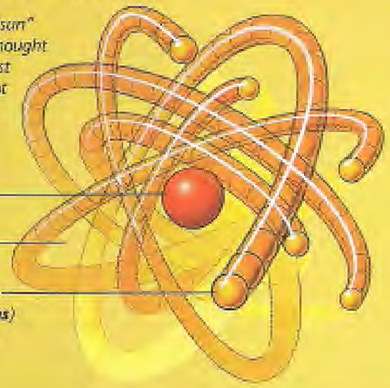

*Positively-charged "sun" (heavy **nucleus** – thought at the time to be just protons; neutron not yet discovered)*
Empty space
*Negatively-charged "planets" (**electrons**) kept in "orbits" by **electric force*** of attraction.*

Nucleus (pl. nuclei) or atomic nucleus

The central core of an atom, consisting of closely-packed **nucleons** (**protons** and **neutrons**).

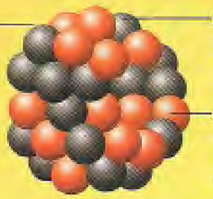

***Nucleus** – (almost all the mass of the atom, but very tiny – its radius is approximately $^1/_{10,000}$th that of the atom)*
***Neutron** (mass approximately 1,840 times that of an **electron**)*
***Proton** (mass approximately 1,836 times that of an **electron**)*

Protons

Positively-charged particles in the **nucleus**. The number of protons (**atomic number**) identifies the element and equals the number of **electrons**, so atoms are electrically neutral.

Neutrons

Electrically neutral particles in the **nucleus**. The number of neutrons in atoms of the same element can vary (see **isotope**).

Mass number (A)

The number of **protons** and **neutrons** (**nucleons**) in a **nucleus**. It is the whole number nearest to the **relative atomic mass** of the atom, and is important in identifying **isotopes**.

Atomic number (Z)

The number of **protons** in a **nucleus** (hence also the number of **electrons** around it). All atoms with the same atomic number are of the same element (see also **isotope**).

Neutron number (N)

The number of **neutrons** in a **nucleus**, calculated by subtracting the atomic number from the mass number. See also graph, page 87.

*The **mass** and **atomic numbers** are often written with the symbol of an element:*

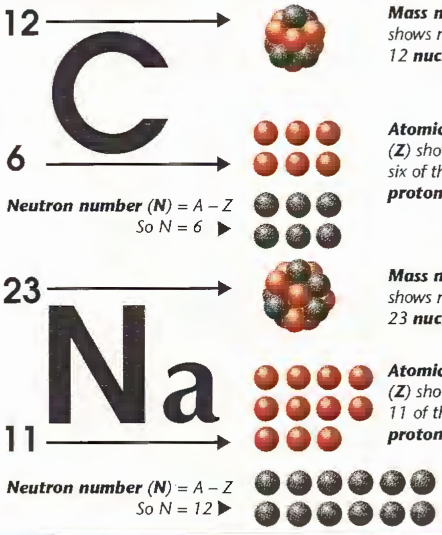

12
C
6
***Neutron number (N)** = A – Z*
So N = 6
23
Na
11
***Neutron number (N)** = A – Z*
So N = 12

***Mass number (A)** shows nucleus has 12 **nucleons**.*

***Atomic number (Z)** shows that six of these are **protons**.*

***Mass number (A** shows nucleus has 23 **nucleons**.*

***Atomic number (Z)** shows that 11 of these are **protons**.*

Electrons

articles with a negative ch
ery small mass. They mov
he **nucleus** in **electron she**
otons.

ectron shells

ons of space around a
aining moving **electro**
n have up to seven (from
alled the **K, L, M, N, O, P**
Each can hold up to a certai
ctrons (the first four, from
take up to 2, 8, 18 and
ectively).

The further away the sh
nucleus, the higher the
ctrons (the shell has a gi
el). The **outer shell** is the
electrons in it. If this is
octet (8 electrons), the
e (see page 85).

The positions of electro
s cannot be exactly dete
one time, but each she
itals, or **probability cloud**
is a region in which
ctrons are likely to be
see also illustration o

topes

rent forms of the same
atomic number, but
bers and hence different
are isotopes of every
only one natural
is **monoisotopic**
ally (see **radioisotope**

bers are used with names
***isotopes** are being speci*

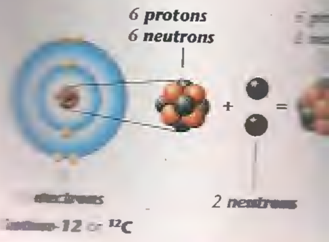

6 protons
6 neutrons
+
=
2 neutrons
-12 or ^{12}C

---

* **Electric force**, 6.

Electrons

Particles with a negative charge and very small mass. They move around the **nucleus** in **electron shells**. See also **protons**.

Electron shells

Regions of space around a **nucleus** containing moving **electrons**. An atom can have up to seven (from the inside, called the **K**, **L**, **M**, **N**, **O**, **P** and **Q shells**). Each can hold up to a certain number of electrons (the first four, from the inside, can take up to 2, 8, 18 and 32 electrons respectively).

The further away the shell is from the nucleus, the higher the energy of its electrons (the shell has a given **energy level**). The **outer shell** is the last shell with electrons in it. If this is full or has an **octet** (8 electrons), the atom is very stable (see page 85).

The positions of electrons in their shells cannot be exactly determined at any one time, but each shell consists of **orbitals**, or **probability clouds**. Each of these is a region in which one or two electrons are likely to be found at any time (see also illustration on page 82).

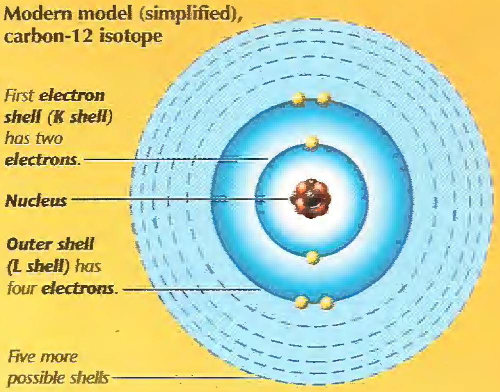

Modern model (simplified), carbon-12 isotope

*First **electron shell (K shell)** has two **electrons**.*
Nucleus
***Outer shell (L shell)** has four **electrons**.*
Five more possible shells

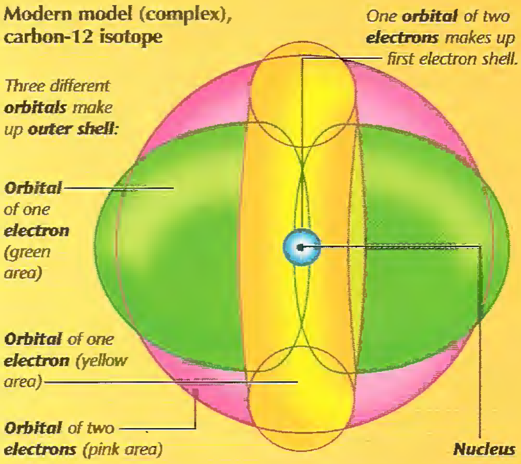

Modern model (complex), carbon-12 isotope

*One **orbital** of two **electrons** makes up first electron shell.*
*Three different **orbitals** make up **outer shell**:*
***Orbital** of one **electron** (green area)*
***Orbital** of one **electron** (yellow area)*
***Orbital** of two **electrons** (pink area)*
Nucleus

Isotopes

Different forms of the same element, with the same atomic number, but different **neutron numbers** and hence different **mass numbers**. There are isotopes of every element, since even if only one natural form exists (i.e. the element is **monoisotopic**), others can be made artificially (see **radioisotope**, page 86).

__Mass numbers__ are used with names or symbols when __isotopes__ are being specified:

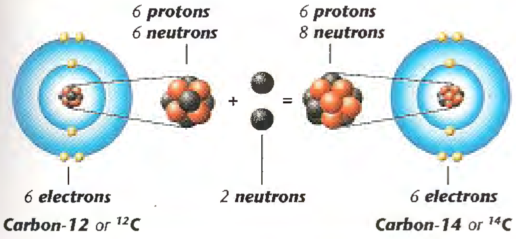

*6 **protons***
*6 **neutrons***
*6 **protons***
*8 **neutrons***
+
=
*6 **electrons***
*2 **neutrons***
*6 **electrons***
Carbon-12 *or* **^{12}C**
Carbon-14 *or* **^{14}C**

Relative atomic mass

Also called **atomic mass** or **atomic weight**. The mass of an atom in **unified atomic mass units** (**u**). Each of these is equal to $^1/_{12}$ of the mass of a carbon-12 atom (**isotope**). The relative atomic mass of a carbon-12 atom is thus 12u, but no other values are whole numbers, e.g. the relative atomic mass of aluminum is 26.9815u.

The relative atomic mass takes into account the various isotopes of the element, if these occur in a natural sample. Natural chlorine, for example, has three chlorine-35 atoms to every one of chlorine-37, and the relative atomic mass of chlorine (35.453u) is a proportional average of the two different masses of these isotopes.

ATOMIC AND NUCLEAR ENERGY

All things, whether large objects or minute particles, have a particular **energy state**, or level of **potential energy*** ("stored" energy). Moreover, they will always try to find their lowest possible energy state, called the **ground state**, which is the state of the greatest stability. In most cases, this involves recombining in some way, i.e. adding or losing constituents. In all cases it results in the release of the "excess" energy – in large amounts if the particles are atoms, and vast amounts if they are nuclei. The greater the **binding energy** of an atom or nucleus, the greater its stability, i.e. the less likely it is to undergo any change.

Binding energy (B.E.)

The energy input needed to split a given atom or nucleus into its constituent parts (see pages 82-83). The **potential energy*** of an atom or nucleus is less than the total potential energy of its parts when these are apart. This is because, when they came together, the parts found a lower (collective) **energy state** (see introduction and **nuclear force**), and so lost energy. The binding energy is a measure of this difference in potential energy – it is the energy needed to "go back the other way" – so the greater it is, the lower the potential energy of an atom or nucleus and the greater its stability. Binding energy varies from atom to atom and nucleus to nucleus.

Nuclear force

The strong force which holds the parts of a nucleus (**nucleons***) together and overcomes the **electric force*** of repulsion between the **protons***. Its effect varies according to the size of the nucleus (see graph, opposite) as the force only acts between immediately adjacent nucleons. The greater the attractive effect of the nuclear force, the higher the **binding energy** of the nucleus (i.e. the more energy was lost when the parts came together).

Quantum theory

States that energy takes the form of minute, separate pulses called **quanta** (sing. **quantum**), rather than a steady stream. The theory was originally limited to energy emitted by bodies (i.e. **electromagnetic wave*** energy), though all other kinds of energy (see pages 8-9) are now generally included. Electromagnetic quanta are now specified as **photons**. The theory further states that the amount of energy carried by a photon is proportional to the **frequency*** of the emitted electromagnetic radiation (see pages 44-45).

*Energy carried by **quantum (photon)**:*

$$E = hf$$

*where E = energy in joules; h = **Planck's constant** (6.63 × 10^{-34} J s^{-1}); f = **frequency*** in Hertz.*

Electron volt (eV)**. Unit of atomic energy, equal to energy gained by one electron moved through **potential difference of 1 V.*

$$1\,eV = 1.6 \times 10^{-19} J$$

***Megaelectron volt (MeV)**. Unit of nuclear energy, equal to 1 million eV.*

$$1\,MeV = 1.6 \times 10^{-13} J$$

Mass defect

The mass of an atom or nucleus is less than the sum of the masses of its parts when these are apart. The difference is the mass defect. It is the mass of the **potential energy*** lost when the parts came together (see **binding energy**, above, and formula, right).

*Einstein showed that energy has mass. Hence any loss of **potential energy*** also results in a loss of mass – the mass of the energy itself. **Einstein's mass-energy formula**:*

$$E = mc^2$$

where E = energy in joules; m = mass in kilograms; c = 3 × 10^8 (numerical value of speed of light in m s^{-1}).

*** Electric force**, 6; **Electromagnetic waves**, 44; **Frequency**, 35; **Nucleons**, 82 (**Nucleus**); **Potential difference**, 58; **Potential energy**, 8; **Protons**, 82.

Levels of nuclear stability

*Atoms with full **outer shells***.*

*Nuclei with **mass numbers*** between 40 and 120 are the most stable, and have the highest **B.E.** or **B.E. per nucleon** (see note under graph, below).*

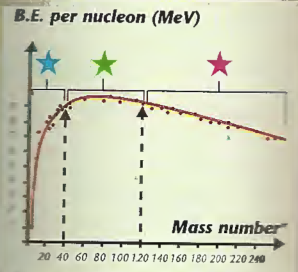

*Note: **binding energy per nucleon** (total B.E. divided by number of nucleons) is better representation of nuclear stability – total B.E. of one nucleus may be higher than that of another, but its B.E. per nucleon may be lower – see right.*

*Change in energy state of an atom or nucleus is called atomic or nuclear **transition**. Change resulting in change in chemical properties (i.e. a different element) is called **transformation** or **transmutation**.*

*As shown, finding higher B.E. (lower energy state) may occur with very little "help", e.g. pushing unstable atoms together, or in the case of **spontaneous radioactivity**. In other cases, energy may need to be added (heating atoms, nuclei or particles).*

Electric force, 6; **Electrons**, 83; **Outer shell**, 83 (**Electron shells**

Levels of nuclear stability

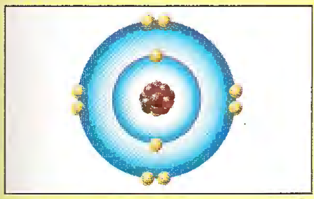

Atoms with full **outer shells***.

★ Nuclei with **mass numbers*** between 40 and 120 are the most stable, and have the highest **B.E.** or **B.E. per nucleon** (see note under graph, below).

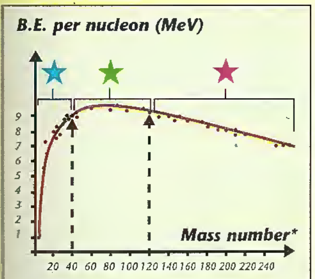

Note: **binding energy per nucleon** (total B.E. divided by number of nucleons) is better representation of nuclear stability – total B.E. of one nucleus may be higher than that of another, but its B.E. per nucleon may be lower – see right.

Change in energy state of an atom or nucleus is called atomic or nuclear **transition**. Change resulting in change in chemical properties (i.e. a different element) is called **transformation** or **transmutation**.

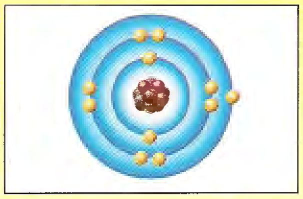

Few **electrons*** in outer shell, or nearly a full outer shell, i.e. just one or two electrons missing.

★ Nuclei with **mass numbers*** less than 40 or more than ★ 120 are less stable, and have low B.E. or B.E. per nucleon.

★ Nuclei with **mass numbers*** less than 40

Few **nucleons*** and hence relatively large surface area. Only a few nucleons (relative to whole number) have "pulling" **nuclear force** acting on them from neighbors on all sides, so overall effect of nuclear force is less.

OR

★ Nuclei with **mass numbers*** greater than 120

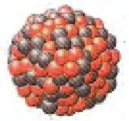

Very large number of nucleons. Means more **protons*** – **electric force*** of repulsion has more effect, overall effect of nuclear force is less.

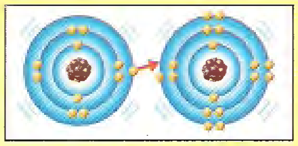

Putting unstable atoms together with other unstable atoms causes them to react – electrons are transferred, i.e. recombinations take place.

Result is always a release of energy ("excess" energy). Particles gain greater stability – total overall B.E. or B.E. per nucleon is higher than before. Energy released as heat and **kinetic energy*** of particles.

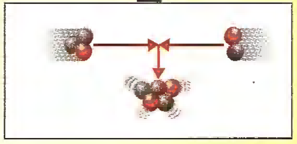

Heating nuclei gives them high **kinetic energy*** and means two will join when they collide (see **fusion**, page 93).

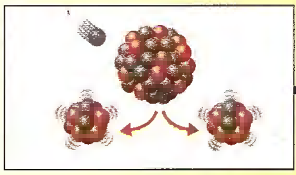

Hitting large nucleus with particle causes it to split (see **fission**, page 92).

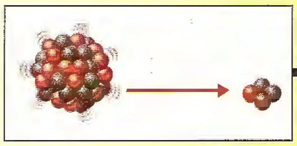

Some heavy nuclei release particles spontaneously (see **radioactivity**, page 86).

As shown, finding higher B.E. (lower energy state) may occur with very little "help", e.g. putting unstable atoms together, or in the case of spontaneous radioactivity. In other cases, energy may need to be added (heating atoms, nuclei or particles).

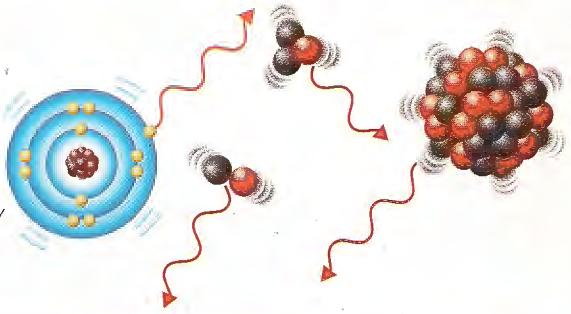

If not enough energy is added, energy is re-released as **photons** (see **quantum theory**). In atoms, electrons "fall" back and the type (**frequency***) of photons emitted depends on the shells between which they move. **X-rays*** (highest frequency, most energy) are emitted if they are the innermost shells, **UV radiation*** is emitted if they are the shells further out, and so on (see **electromagnetic spectrum**, page 44). With nuclei, **γ-rays*** are always emitted (more energy is involved).

Electric force, 6; **Electrons**, 83; **Frequency**, 35; **Kinetic energy**, 9; **Mass number**, 82; **Nucleons**, 82 (**Nucleus**);
Outer shell, 83 (**Electron shells**); **Protons**, 82; **Ultraviolet (UV) radiation**, **X-rays**, 44; **Gamma (γ) rays**, 44, 86.

RADIOACTIVITY

Radioactivity is a property of some unstable **nuclei** (see pages 82 and 84), whereby they break up spontaneously into nuclei of other elements and emit **radiation***, a process known as **radioactive decay**. There are three types of radiation emitted by radioactive elements: streams of **alpha particles** (called **alpha rays**); streams of **beta particles** (**beta rays**) and **gamma rays**. For more about the detection and uses of radiation, see pages 88-91.

Radioisotope or radioactive isotope

Any radioactive substance (all substances are effectively isotopes – see page 83). There are several naturally-occurring radioisotopes, most of which still exist because they have very long **half-lives** (e.g. uranium-238), though one, carbon-14, is continually produced by **cosmic rays** (see **background radiation**, page 88). Other radioisotopes are produced by **nuclear fission***, and more still are produced in research centers, where nuclei are hit by fast particles (e.g. **protons*** and **neutrons***). These are speeded up in **particle accelerators**, e.g. **cyclotrons** (see picture below).

Cut-away diagram of vacuum chamber (central part of cyclotron)

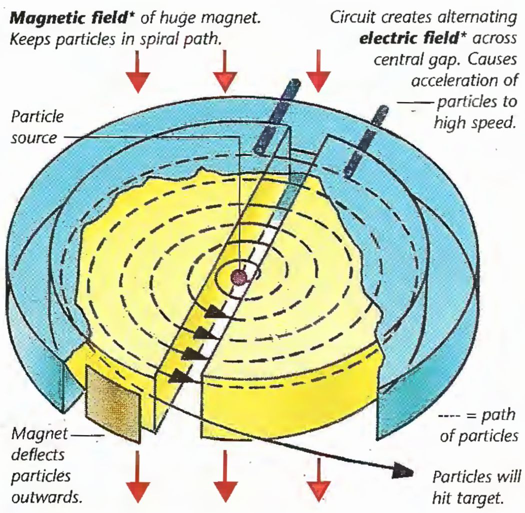
Magnetic field* of huge magnet. Keeps particles in spiral path.
*Circuit creates alternating **electric field*** across central gap. Causes acceleration of particles to high speed.*
Particle source
Magnet deflects particles outwards.
---- = path of particles
Particles will hit target.

Alpha
Beta
Gamma

The Greek letters used for the three types of radiation*

Alpha particles (α-particles)

Positively-charged particles ejected from some radioactive nuclei (see **alpha decay**). They are relatively heavy (two **protons*** and two **neutrons***), move relatively slowly and have a low penetrating power.

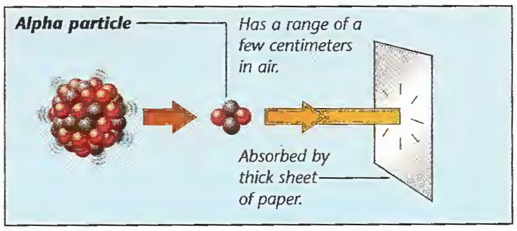
Alpha particle
Has a range of a few centimeters in air.
Absorbed by thick sheet of paper.

Beta particles (β-particles)

Particles ejected from some radioactive nuclei at about the speed of light. There are two types – **electrons*** and **positrons**, which have the same mass as electrons, but a positive charge. See **beta decay**, page 87.

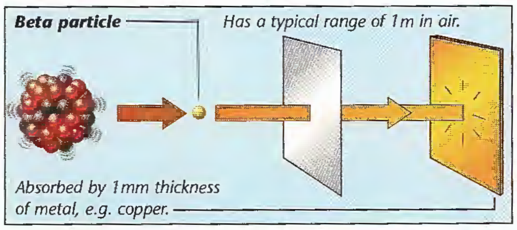
Beta particle
Has a typical range of 1m in air.
Absorbed by 1mm thickness of metal, e.g. copper.

Gamma rays (γ-rays)

Invisible **electromagnetic waves** (see also page 44). They have the highest penetrating power and are generally, though not always, emitted from a radioactive nucleus after an **alpha** or **beta particle**.

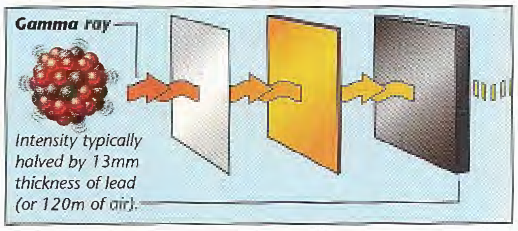
Gamma ray
Intensity typically halved by 13mm thickness of lead (or 120m of air).

---

* **Electric field**, 58; **Electrons**, 83; **Magnetic field**, 72; **Neutrons**, 82; **Nuclear fission**, 92; **Protons**, 82; **Radiation**, 9.

Radioactive decay

The spontaneous splitti
nucleus, which results
or **beta particles**, often
rays. When a nucleus e
i.e. undergoes a nuclea
energy is released (see
different nucleus (and a
this is also radioactive,
continues until a stable
is reached. Such a serie
called a **decay series, de**
series or **transformation**

Half-life (T½)

The time it takes for half
in a sample to undergo
hence for the radiation e
This is all that can be ac
is impossible to predict t
atom, since they decay i
randomly. The range of h
the half-life of strontium-
uranium-238 is 4.5 × 10

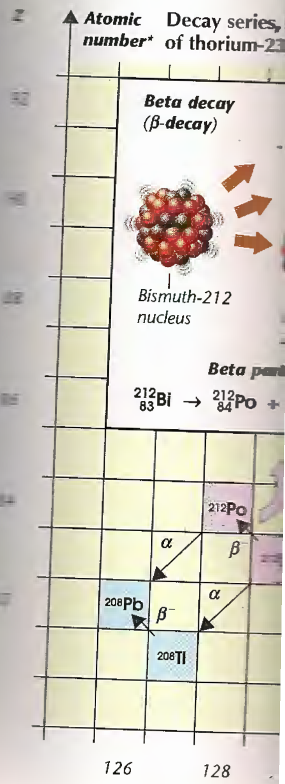
Atomic number*
Decay series, of thorium-23
Beta decay (β-decay)
Bismuth-212 nucleus
Beta pa
$^{212}_{83}Bi \rightarrow ^{212}_{84}Po +$
^{212}Po
α
$β^-$
^{208}Pb
^{208}Tl
126
128

Atomic number, 82; **Electrons**,
Neutron number, **Neutrons**, **Pro**

Radioactive decay

The spontaneous splitting up of a radioactive nucleus, which results in the ejection of **alpha** or **beta particles**, often followed by **gamma rays**. When a nucleus ejects such a particle, i.e. undergoes a nuclear **disintegration**, energy is released (see page 84), and a different nucleus (and atom) is formed. If this is also radioactive, the decay process continues until a stable (non-radioactive) atom is reached. Such a series of disintegrations is called a **decay series, decay chain, radioactive series** or **transformation series**.

Half-life (T½)

The time it takes for half the atoms on average in a sample to undergo **radioactive decay**, and hence for the radiation emitted to be halved. This is all that can be accurately predicted – it is impossible to predict the decay of any single atom, since they decay individually and randomly. The range of half-lives is vast, e.g. the half-life of strontium-90 is 28 years; that of uranium-238 is 4.5 × 10^9 years.

Alpha decay (α-decay)

The loss of an **alpha particle** by a radioactive nucleus. This decreases the **atomic number*** by two and the **mass number*** by four, and so a new nucleus is formed.

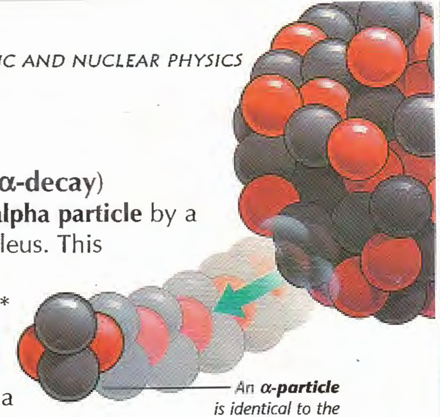

An ***α-particle*** *is identical to the nucleus of a helium atom.*

Beta decay (β-decay)

The loss of either kind of **beta particle** by a radioactive nucleus. The electron (**β^-** or **e^-**) is ejected (with another particle called an **antineutrino**) when a **neutron*** decays into a **proton***. The positron (**β^+** or **e^+**) is ejected (with another particle called a **neutrino**) when a proton decays into a neutron. Beta decay thus increases or decreases the **atomic number*** by one (the **mass number*** stays the same).

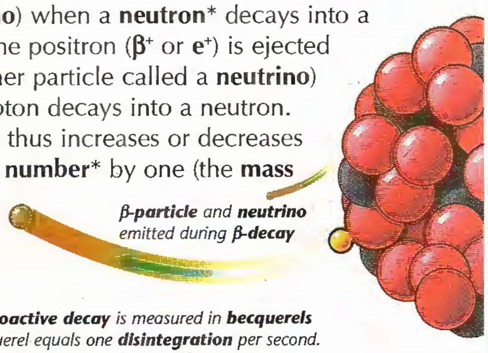

β-particle *and* ***neutrino*** *emitted during* ***β-decay***

The rate of ***radioactive decay*** *is measured in* ***becquerels (Bq)****. One becquerel equals one* ***disintegration*** *per second. An older unit, the* ***curie****, equals 3.7 × 10^{10} becquerels.*

Decay series, showing radioactive decay of thorium-232 to stable lead-208

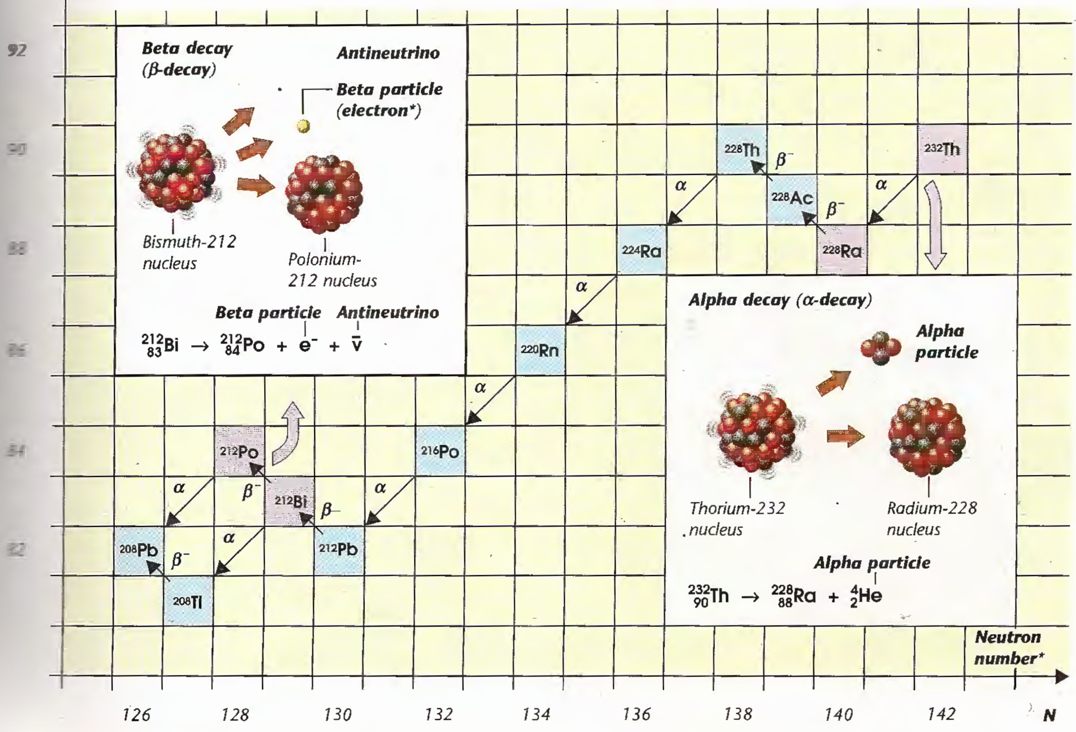
Z
Atomic number*
92
90
88
86
84
82
Beta decay (β-decay)
Antineutrino
Beta particle (electron*)
Bismuth-212 nucleus
Polonium-212 nucleus
Beta particle
Antineutrino
$^{212}_{83}Bi \rightarrow ^{212}_{84}Po + e^- + \bar{v}$
228Th
232Th
228Ac
228Ra
224Ra
220Rn
216Po
212Po
212Bi
212Pb
208Pb
208Tl
α
β⁻
Alpha decay (α-decay)
Alpha particle
Thorium-232 nucleus
Radium-228 nucleus
Alpha particle
$^{232}_{90}Th \rightarrow ^{228}_{88}Ra + ^4_2He$
Neutron number*
126
128
130
132
134
136
138
140
142
N

Atomic number, 82; **Electrons**, 83; **Mass number**, **Neutron number**, **Neutrons**, **Protons**, 82.

DETECTING AND MEASURING RADIOACTIVITY

There are a number of devices which detect and measure the radiation emitted by radioactive substances (**radioisotopes***). Some are used mainly in laboratories (to study artificially produced radioisotopes); others have a wider range of uses (e.g. as monitoring devices for safety purposes) and can also be used to detect **background radiation**. Most of the devices detect and measure the radiation by monitoring the **ionization** it causes – see **Geiger counter** and **pulse electroscope**, right, and **cloud** and **bubble chambers**, page 90.

*Radioactive substances (**radioisotopes***) have special hazard warning labels.*

Background radiation

Plants, rocks and animals contain carbon-14, a natural source of radiation.

Radiation present on Earth (in relatively small amounts), originating both from natural and unnatural sources. One notable natural source is carbon-14, which is taken in by plants and animals. This is constantly being produced from stable nitrogen-14 due to bombardment by **cosmic rays** (**cosmic radiation**) entering the atmosphere from outer space. These are streams of particles of enormously high energy. Unnatural sources of radiation include industry, medicine and weapons testing. The **background count** is a measure of the background radiation.

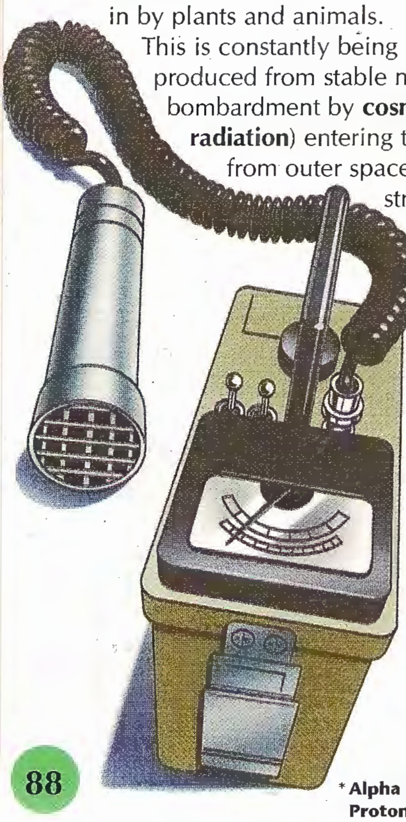

*A **Geiger counter**, used for measuring the **background** count.*

Ionization

The creation of **ions** (electrically-charged particles), which occurs when atoms (which are electrically neutral) lose or gain **electrons***, creating **cations** (positive ions) or **anions** (negative ions) respectively.

Ionization

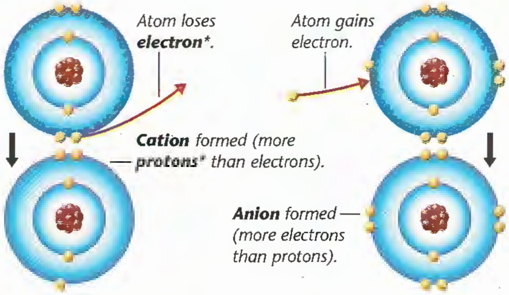

*Atom loses **electron****.
Atom gains electron.
Cation** formed (more **protons than electrons).*
***Anion** formed (more electrons than protons).*

In the case of radiation, **alpha** and **beta particles*** ionize the atoms of substances they pass through, usually creating cations. This is because their energy is so high that they cause one or more electrons to be "knocked out". For more about ions and ionization, see pages 130-131.

Ionization due to radiation

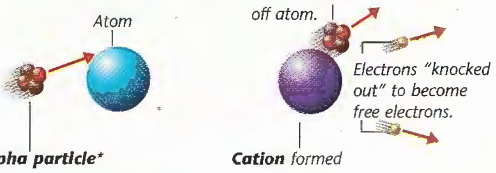

Atom
Alpha particle*
Particle bounces off atom.
Electrons "knocked out" to become free electrons.
***Cation** formed*

---

* **Alpha particles**, **Beta particles**, 86; **Electrons**, 83; **Protons**, 82; **Radioisotope**, 86.

Detection devi

Dosimeter or **film ba**
A device worn by all w
radioactive material. It
film (which radiation w
developed regularly an
darkening shows the **do**
wearer has been expose

Geiger counter

A piece of apparatus (s
consisting of a **Geiger-**
and/or **ratemeter** and
The tube is a gas-fille
electrodes* – its walls
and it has a central wi
apparatus indicates the

Geiger counter

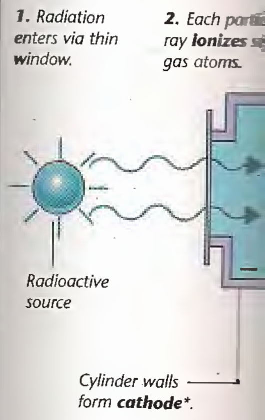

***1.** Radiation enters via thin window.*
***2.** Each parti ray **ionizes** gas atoms.*
Radioactive source
*Cylinder walls form **cathode****.

Pulse (Wulf) electros

A type of **gold-leaf elec**
The walls of a chambe
the cap form the **cathod**
side **anode*** is placed
leaf. This attracts **electr**
from the cap, leaving it
charged (the leaf moves
the rod, as they are bo
charged, but not enoug
the anode before the
source is introduced). T
indicates the presence
by beating forward and
each **ionization** it causes

---

* **Anode**, **Cathode**, 66 (**Electrode**
Gold-leaf electroscope, 56 (**Elect**

Detection devices

Dosimeter or film badge

A device worn by all who work with radioactive material. It contains photographic film (which radiation will darken). This is developed regularly and the amount of darkening shows the **dose** of radiation the wearer has been exposed to.

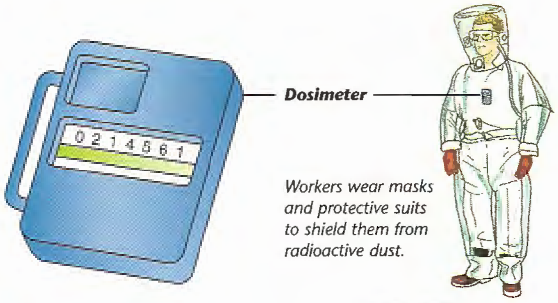

Workers wear masks and protective suits to shield them from radioactive dust.

Geiger counter

A piece of apparatus (see picture, page 88) consisting of a **Geiger-Müller tube**, a **scaler** and/or **ratemeter** and often a loudspeaker. The tube is a gas-filled cylinder with two **electrodes*** – its walls act as the **cathode***, and it has a central wire **anode***. The whole apparatus indicates the presence of radiation by registering pulses of current between the electrodes. These pulses result from the **ionization** the radiation causes in the gas (normally low pressure argon, plus a trace of bromine). A scaler is an electronic counter which counts the pulses and a ratemeter measures the count rate – the average rate of pulses in counts per second.

Geiger counter

1. Radiation enters via thin window.

*2. Each particle or ray **ionizes** several gas atoms.*

*3. Ions attracted to cathode, **electrons*** to anode.*

*4. Other atoms are hit on the way, creating **avalanche** of more ions and electrons.*

5. Electrons taken in at anode and "pulled" from cathode (to turn ions back into atoms).

6. Pulse of current (amplified because of avalanche) flows round circuit for each original particle or ray.

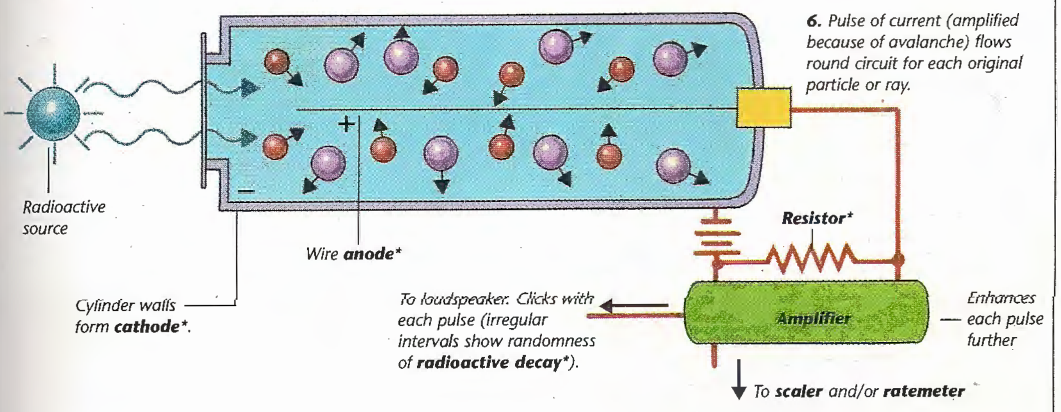

Radioactive source

*Wire **anode****

*Cylinder walls form **cathode****.

*To loudspeaker. Clicks with each pulse (irregular intervals show randomness of **radioactive decay***).*

Enhances each pulse further

*To **scaler** and/or **ratemeter***

Pulse (Wulf) electroscope

A type of **gold-leaf electroscope***. The walls of a chamber of air around the cap form the **cathode***, and a side **anode*** is placed close to the leaf. This attracts **electrons*** down from the cap, leaving it positively charged (the leaf moves away from the rod, as they are both negatively charged, but not enough to touch the anode before the radioactive source is introduced). The leaf indicates the presence of radiation by beating forward and back for each **ionization** it causes.

Pulse (Wulf) electroscope

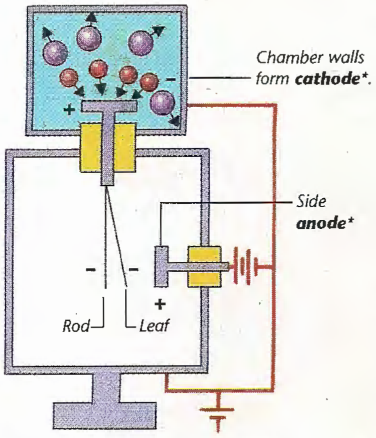

*Chamber walls form **cathode****.

*Side **anode****

*1. Radiation from radioactive source **ionizes** air in chamber.*

*2. **Avalanche** of ions and **electrons*** formed (see **Geiger counter**).*

3. Ions move to cathode, electrons enter cap and are pulled down leaf by anode.

*4. Leaf moves over to touch anode and pulse of current flows (see **Geiger counter**).*

5. Leaf becomes neutral and moves back to rod, aided by spring. Process starts again.

Anode, **Cathode**, 66 (**Electrode**); **Electrons**, 83;
Gold-leaf electroscope, 56 (**Electroscope**); **Radioactive decay**, 87; **Resistor**, 62.

Cloud chamber

A device in which the paths taken by **alpha** and **beta particles*** show up as tracks. This happens when the vapor in the chamber (alcohol or water vapor) is turned into **supersaturated** vapor by cooling (in one of two different ways – see below). A supersaturated vapor is vapor below the temperature at which it should condense, but which does not condense because there are no dust or other particles present for droplets to form around.

Wilson cloud chamber

Vapor cooled by sudden increase in volume (withdrawal of piston).

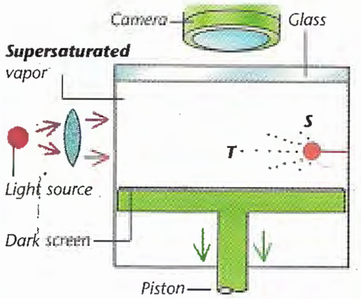

*1. Radiation from source (**S**) causes **ionization*** of vapor.*

2. Ions formed act like dust particles, i.e. vapor condenses on them.

*3. Tracks of liquid droplets (**T**) left where vapor has condensed (visible long enough to be photographed).*

Diffusion cloud chamber

*Vapor cooled by a base of dry ice (solid carbon dioxide). Vapor **diffuses*** downwards.*

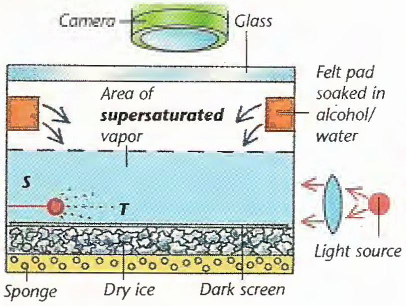
Camera
Glass
Felt pad soaked in alcohol/ water
Area of supersaturated vapor
S
T
Light source
Sponge
Dry ice
Dark screen

Cloud chamber** tracks (right) are produced at irregular intervals, showing random nature of **radioactive decay.*

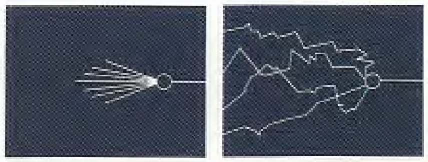

*Tracks made by heavy α-**particles*** are short, straight and thick.*

*Tracks made by light β-**particles*** are long, straggly and thin.*

Gamma rays do not create tracks themselves, but can knock **electrons*** out of single atoms. These then speed away and create tracks like β-particle tracks (see left).*

Bubble chamber

A device which, like a cloud chamber, shows particle tracks. It contains **superheated** liquid (usually hydrogen or helium) – liquid heated to above its boiling point, but not actually boiling because it is under pressure. After the pressure is suddenly lowered, nuclear particles entering the chamber cause **ionization*** of the liquid atoms. Wherever this occurs, the energy released makes the liquid boil, producing tracks of bubbles.

***Bubble chamber** tracks, showing paths taken by nuclear particles.*

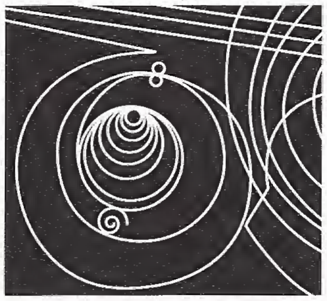

Bubble tracks are generally curved, because a magnetic field is set up to deflect particles. (This leads to better identification.)

Scintillation counter

A device which detects **gamma rays***. It consists of a **scintillation crystal** and a **photomultiplier** tube. The crystal is made of a **phosphor*** (e.g. sodium iodide). Phosphors emit light flashes (**scintillations**) when hit by radiation.

Scintillation counter

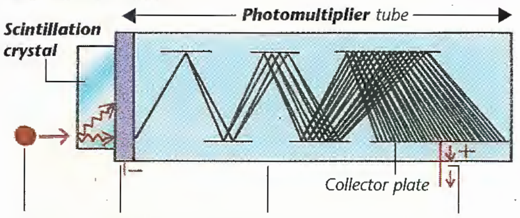

1. Radiation from source makes crystal emit light

*2. Photosensitive material with negative charge. Emits **electrons*** when hit by light flashes.*

*3. Main part of tube (**electron multiplier**). Electrons accelerated to far end by electric field, hitting metal plates and releasing more electrons.*

*4. Strong pulse of current hence produced for each original **gamma ray***. This is shown on **scaler** and/or **ratemeter** (see **Geiger counter**, page 89).*

---

* **Alpha particles**, **Beta particles**, 86; **Diffusion**, 5; **Electrons**, 83; **Gamma rays**, 86; **Ionization**, 88; **Phosphors**, 44 (**Phosphorescence**); **Radioactive decay**, 87.

USES OF

The radiation emitt
(radioactive substan
a wide variety of us
the fields of medici
archaeological rese

Radiology

The study of radioactivit
especially with regard to

Radiotherapy

The use of the radiation e
radioisotopes* to treat dis
cells are susceptible to ra
possible to destroy malig
by using carefully contro
radiation.

*This patient is undergoing **external beam radiotherapy**, where the radiation is emitted from a machine outside the body. Some types of cancer can be treated by **radioactive implants** inserted into the body.*

Radioactive tracing

A method of following the
substance through an obj
its concentration as it mo
by introducing a **radioiso**
substance and tracking the
The radioisotope used is c
the substance is said to be
medical diagnosis, for exa
of the radioisotope in an o
the presence of malignant
radioisotopes used always
lives* and decay into harm
substances.

Irradiation

Food, such as fruit and me
can be **irradiated** with **gamma rays***. The radiatio
delays ripening in fruit and vegetables, and destroys bacteria in meat, enabling it to keep fresh for longer.

---

* **Gamma rays**, 86; **Half-life**, 87; **Rad**

USES OF RADIOACTIVITY

The radiation emitted by **radioisotopes*** (radioactive substances) can be put to a wide variety of uses, particularly in the fields of medicine, industry and archaeological research.

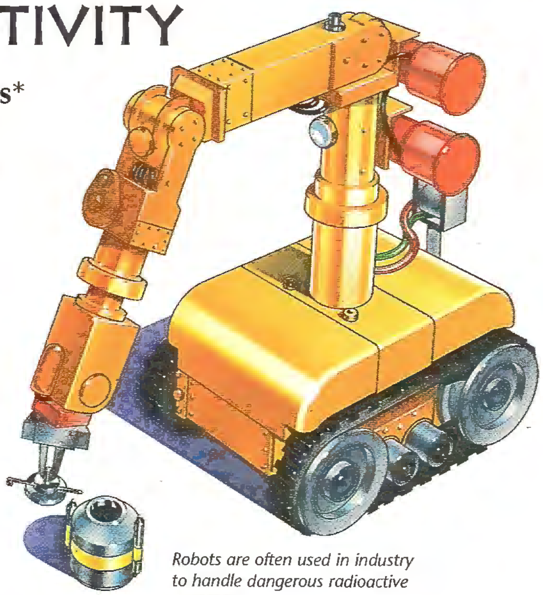

Robots are often used in industry to handle dangerous radioactive substances.

Radiology

The study of radioactivity and **X-rays***, especially with regard to their use in medicine.

Radiotherapy

The use of the radiation emitted by **radioisotopes*** to treat disease. All living cells are susceptible to radiation, so it is possible to destroy malignant (cancer) cells by using carefully controlled doses of radiation.

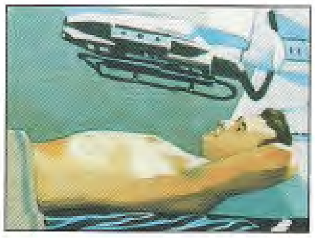

*This patient is undergoing **external beam radiotherapy**, where the radiation is emitted from a machine outside the body. Some types of cancer can be treated by **radioactive implants** inserted into the body.*

Radioactive tracing

A method of following the path of a substance through an object, and detecting its concentration as it moves. This is done by introducing a **radioisotope*** into the substance and tracking the radiation it emits. The radioisotope used is called a **tracer**, and the substance is said to be **labeled**. In medical diagnosis, for example, high levels of the radioisotope in an organ may indicate the presence of malignant (cancer) cells. The radioisotopes used always have short **half-lives*** and decay into harmless substances.

Irradiation

Food, such as fruit and meat can be **irradiated** with **gamma rays***. The radiation delays ripening in fruit and vegetables, and destroys bacteria in meat, enabling it to keep fresh for longer.

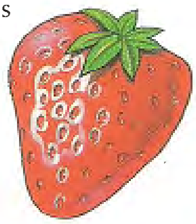

*After two weeks, this **irradiated** strawberry is still firm and fresh.*

Gamma radiography (γ-radiography)

The production of a **radiograph** (similar to a photograph) by the use of **gamma rays*** (see also **X-radiography**, page 44). This has many uses, including quality control in industry.

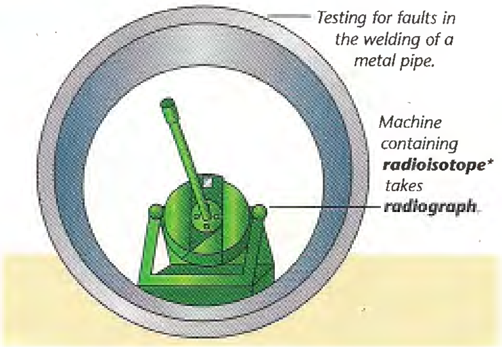

Testing for faults in the welding of a metal pipe.
*Machine containing **radioisotope*** takes **radiograph***

Radiocarbon dating or carbon dating

A way of calculating the time elapsed since living matter died. All living things contain a small amount of carbon-14 (a **radioisotope*** absorbed from the atmosphere), which continues to emit radiation after death. This emission gradually decreases (carbon-14 has a **half-life*** of 5,700 years), so the age of the remains can be calculated from its strength.

***Radiocarbon dating** showed that this insect trapped in amber is 5,000 years old.*

---

* **Gamma rays**, 86; **Half-life**, 87; **Radioisotope**, 86; **X-rays**, 44.

NUCLEAR FISSION AND FUSION

The central **nucleus** of an atom (see page 82) holds vast amounts of "stored" energy (see pages 84-85). **Nuclear fission** and **nuclear fusion** are both ways in which this energy can be released. They are both **nuclear reactions** (reactions which bring about a change in the nucleus).

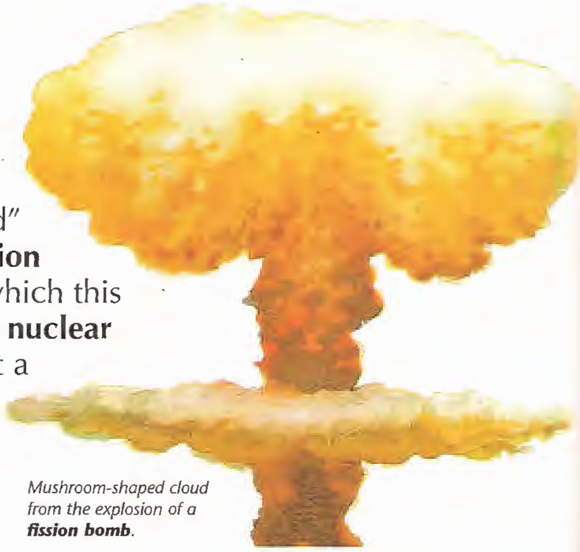

Mushroom-shaped cloud from the explosion of a ***fission bomb****.*

Nuclear fission

The process in which a heavy, unstable nucleus splits into two (or more) lighter nuclei, roughly equal in size, with the release of two or three **neutrons*** (**fission neutrons**) and a large amount of energy (see also page 84). The two lighter nuclei are called **fission products** or **fission fragments** and many of them are **radioactive***. Fission is made to happen (see **induced fission**) in **fission reactors*** to produce heat energy. It does not often occur naturally (**spontaneous fission**).

Induced fission of uranium-235

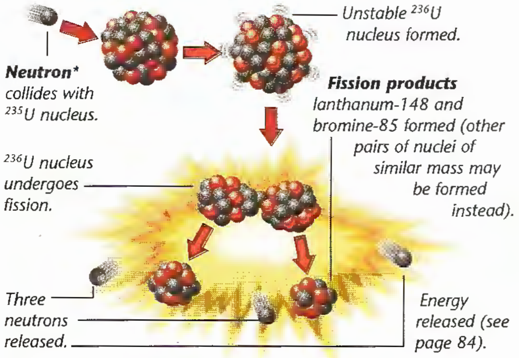
Unstable ^{236}U nucleus formed.
Neutron* collides with ^{235}U nucleus.
Fission products lanthanum-148 and bromine-85 formed (other pairs of nuclei of similar mass may be formed instead).
^{236}U nucleus undergoes fission.
Three neutrons released.
Energy released (see page 84).

Nuclear equation *for reaction, above (see* ***mass*** *and* ***atomic numbers****, page 82):*

$$^{235}_{92}U + ^{1}_{0}n \rightarrow ^{236}_{92}U \rightarrow ^{148}_{57}La + ^{85}_{35}Br + 3^{1}_{0}n + \text{energy}$$

Spontaneous fission

Nuclear fission which occurs naturally, i.e. without assistance from an outside agency. This may happen to a nucleus of a heavy element, e.g. the **isotope*** uranium-238, but the probability is very low compared to that of a simpler process like **alpha decay*** occurring instead.

Induced fission

Nuclear fission of a nucleus made unstable by artificial means, i.e. by being hit by a particle (often a **neutron***), which it then absorbs. Not all nuclei can be induced to fission in this way; those which can, e.g. those of the **isotopes*** uranium-235 and plutonium-239, are described as **fissile**. If there are lots of fissile nuclei in a substance (see also **thermal** and **fast reactor**, page 95), the neutrons released by induced fissions will cause more fissions (and neutrons), and so on. This is known as a **chain reaction**. A well-controlled chain reaction is allowed to occur in a **fission reactor***, but that occurring in a **fission bomb** is uncontrolled and extremely explosive.

Induced fission causing a chain reaction

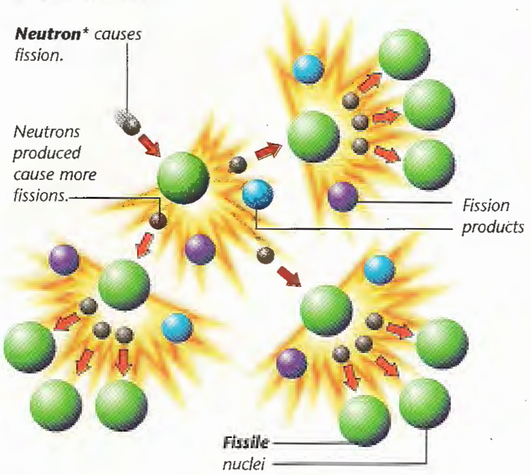
Neutron* causes fission.
Neutrons produced cause more fissions.
Fission products
Fissile nuclei

Critical mass

The minimum mass of a
needed to sustain a **chai**
induced fission). In small
the surface area to volum
and too many of the **neu**
the first fissions escape in
Nuclear fuel is kept in su

Fission bomb or atom

A bomb in which two **su**
(see above) are brought t
explosion. The resulting c
(see **induced fission**) rele
of energy.

Nuclear fusion

The collision and comb
nuclei to form a heavie
with the release of larg
(see also page 84). Unl
does not leave **radioact**
fusion requires tempera
degrees Celsius, to give
kinetic energy* for the
collide. (Because of the
fusion reactions are als
reactions.) It therefore o
the Sun (and stars like i
carried out with the aim
controlled, induced fusi

Example of nuclear fusion
(D-T reaction *– see also* ***fusion***

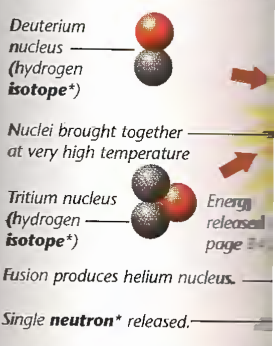
Deuterium nucleus (hydrogen **isotope***)
Nuclei brought together at very high temperature
Tritium nucleus (hydrogen **isotope***)
Energ releasec page
Fusion produces helium nucleus.
Single **neutron*** released.

Nuclear equation *for reaction*
(see ***mass and atomic numbers****,*

$$^{2}_{1}H + ^{3}_{1}H \rightarrow ^{4}_{2}He$$

---

* **Alpha decay**, 87; **Fission reactor**, 94;
Isotopes, 83; **Neutrons**, 82; **Radioactivity**, 86.

* **Fusion reactor**, 94; **Isotopes**, 83
energy, 9; **Neutrons**, 82; **Radioac**

Critical mass

The minimum mass of a **fissile** substance needed to sustain a **chain reaction** (see **induced fission**). In smaller **subcritical masses**, the surface area to volume ratio is too high, and too many of the **neutrons*** produced by the first fissions escape into the atmosphere. Nuclear fuel is kept in subcritical masses.

Fission bomb or atom bomb (A-bomb)

A bomb in which two **subcritical masses** (see above) are brought together by a trigger explosion. The resulting **chain reaction** (see **induced fission**) releases huge amounts of energy.

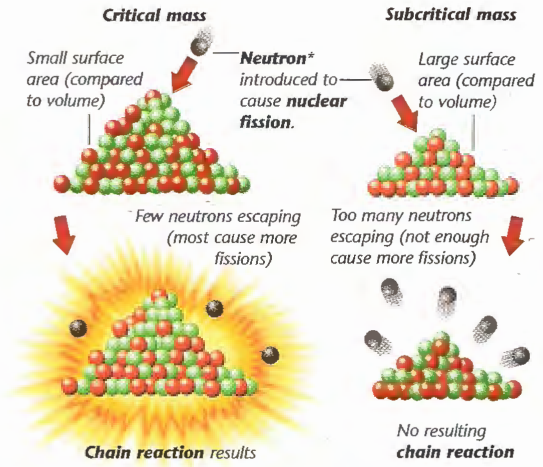
Critical mass
Subcritical mass
Small surface area (compared to volume)
Neutron* introduced to cause **nuclear fission**.
Large surface area (compared to volume)
Few neutrons escaping (most cause more fissions)
Too many neutrons escaping (not enough cause more fissions)
Chain reaction results
No resulting **chain reaction**

Nuclear fusion

The collision and combination of two light nuclei to form a heavier, more stable nucleus, with the release of large amounts of energy (see also page 84). Unlike **nuclear fission**, it does not leave **radioactive*** products. Nuclear fusion requires temperatures of millions of degrees Celsius, to give the nuclei enough **kinetic energy*** for them to fuse when they collide. (Because of the high temperatures, fusion reactions are also called **thermonuclear reactions**.) It therefore only occurs naturally in the Sun (and stars like it), but research is being carried out with the aim of achieving controlled, induced fusion in **fusion reactors***.

Example of nuclear fusion
*(**D-T reaction** – see also **fusion reactor**, page 94.)*

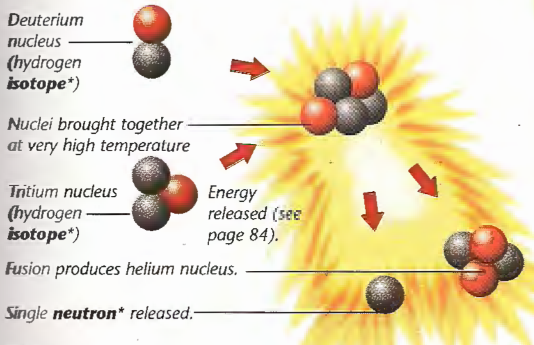
Deuterium nucleus (hydrogen **isotope***)
Nuclei brought together at very high temperature
Tritium nucleus (hydrogen **isotope***)
Energy released (see page 84).
Fusion produces helium nucleus.
Single **neutron*** released.

***Nuclear equation** for reaction above (see **mass and atomic numbers**, page 82):*

$$^{2}_{1}H + ^{3}_{1}H \rightarrow ^{4}_{2}He + ^{1}_{0}n + \text{energy}$$

Fusion bomb or hydrogen bomb (H-bomb)

A bomb in which uncontrolled **nuclear fusion** occurs in a mixture of tritium and deuterium (hydrogen **isotopes***). A trigger **fission bomb** creates the high temperature needed (fusion bombs are also called **fission-fusion bombs**). The energy released is about 30 times that released from a fission bomb of the same size.

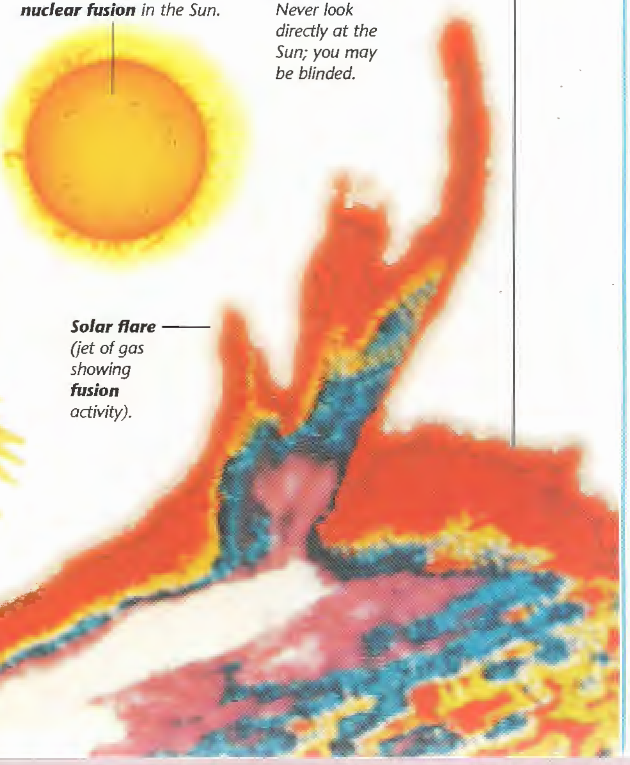
*Hydrogen undergoes **nuclear fusion** in the Sun.*

***WARNING:** Never look directly at the Sun; you may be blinded.*

Solar flare (jet of gas showing **fusion** activity).

Fusion reactor, 94; **Isotopes**, 83; **Kinetic energy**, 9; **Neutrons**, 82; **Radioactivity**, 86.

POWER FROM NUCLEAR REACTIONS

A **nuclear reactor** is a structure inside which nuclear reactions produce vast amounts of heat. There are potentially two main types of reactor – **fission reactors** and **fusion reactors**, though the latter are still being researched. All present-day **nuclear power stations** are built around a central fission reactor and each generates, per unit mass of fuel, far larger amounts of power (electricity) than any other type of power station.

Fission reactor

A **nuclear reactor** in which the heat is produced by **nuclear fission***. There are two main types in use in nuclear power stations – **thermal reactors** and **fast reactors** or **fast breeder reactors** (see page opposite), both of which use uranium as their main fuel.

The uranium is held in long cylinders packed in the **core** (center of the reactor). The rate of the **chain reaction*** (and hence the rate of power production) is closely controlled by **control rods**. The diagram below shows how a fission reactor can be used to generate power.

Schematic diagram of fission reactor and power station complex

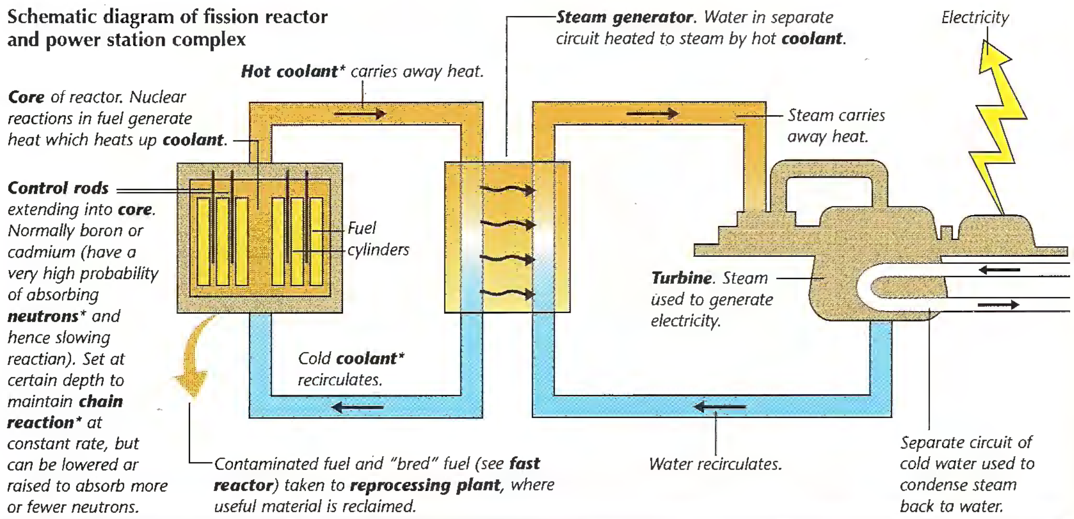

__Steam generator__. Water in separate circuit heated to steam by hot __coolant__.

Electricity

Hot coolant* *carries away heat.*

***Core** of reactor. Nuclear reactions in fuel generate heat which heats up **coolant**.*

Steam carries away heat.

Control rods** extending into **core**. Normally boron or cadmium (have a very high probability of absorbing **neutrons and hence slowing reaction). Set at certain depth to maintain **chain reaction*** at constant rate, but can be lowered or raised to absorb more or fewer neutrons.*

***Turbine**. Steam used to generate electricity.*

*Cold **coolant**** *recirculates.*

*Contaminated fuel and "bred" fuel (see **fast reactor**) taken to **reprocessing plant**, where useful material is reclaimed.*

Water recirculates.

Separate circuit of cold water used to condense steam back to water.

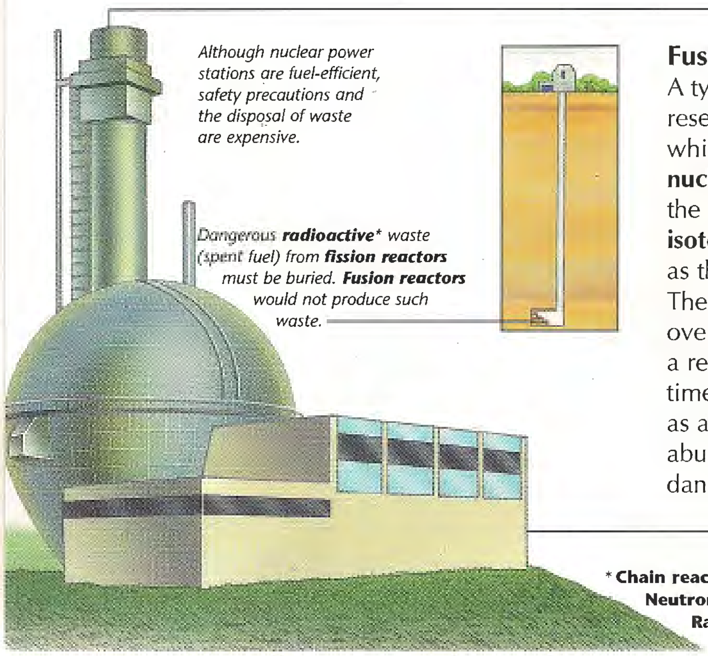

Although nuclear power stations are fuel-efficient, safety precautions and the disposal of waste are expensive.

*Dangerous **radioactive*** waste (spent fuel) from **fission reactors** must be buried. **Fusion reactors** would not produce such waste.*

Fusion reactor

A type of **nuclear reactor**, being researched but as yet undeveloped, in which the heat would be produced by **nuclear fusion***. This would probably be the fusion of the nuclei of the hydrogen **isotopes*** deuterium and tritium – known as the **D-T reaction** (see picture, page 93). There are several major problems to be overcome before a fusion reactor becomes a reality, but it would produce about four times as much energy per unit mass of fuel as a **fission reactor**. Also, hydrogen is abundant, whereas uranium is scarce, and dangerous and expensive to mine.

***Chain reaction**, 92 (**Induced fission**); **Coolant**, 344; **Isotopes**, 83; **Neutrons**, 82; **Nuclear fission**, 92; **Nuclear fusion**, 93; **Radioactivity**, 86.

Types of fission

Thermal reactor

A **fission reactor** contai
around the fuel cylinder
with light nuclei, such a
It is used to slow down
produced by the first fiss
fuel – the neutrons boun
nuclei (which themselve
absorb neutrons) and ev
to about 2,200m s^{-1}. Slo
improves their chances

Types of thermal reactor

Pressurized water reactor (PW

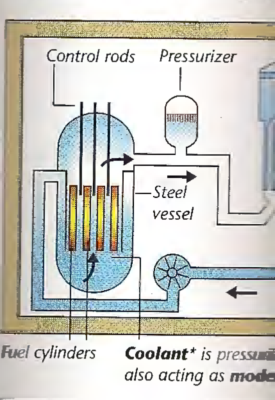

Fuel cylinders

Coolant* *is pressu*
*also acting as **mode***

Fast reactor or fast b

A **fission reactor**, inside
which cause the fission
remain as fast neutrons
2 × 10^7m s^{-1}). The fuel
enriched with extra nuc
(see **thermal reactor**) ar
Both of these will fissio
fast neutrons, unlike ura
far more likely to "cap
(becoming ^{239}U) and un
decay*. The final produ
however, is ^{239}Pu. Fast re
called "breeders" becau
of ^{239}U to ^{239}Pu is allowe
blanket of ^{238}U around t
more fuel is created and
reactors have a more co
at higher temperatures t
They are also more effici
greater proportion of the
becomes contaminated.

***Chain reaction**, 92 (**Induced fiss**
Isotopes, 83; **Neutrons**, 82; **Radi**

Types of fission reactor

Thermal reactor

A **fission reactor** containing a **moderator** around the fuel cylinders. This is a substance with light nuclei, such as graphite or water. It is used to slow down the fast **neutrons*** produced by the first fissions in the uranium fuel – the neutrons bounce off the light nuclei (which themselves are unlikely to absorb neutrons) and eventually slow down to about 2,200m s^{-1}. Slowing the neutrons improves their chances of causing further fissions (and continuing the **chain reaction***). Faster neutrons are likely to be "captured" by the most abundant nuclei – those of the **isotope*** uranium-238 (see **fast reactor**), whereas slow neutrons can travel on until they find uranium-235 nuclei. These will undergo fission when hit by neutrons of any speed, but make up a smaller percentage of the fuel (despite the fact that it is now often enriched with extra atoms of ^{235}U).

Types of thermal reactor

Pressurized water reactor (PWR)

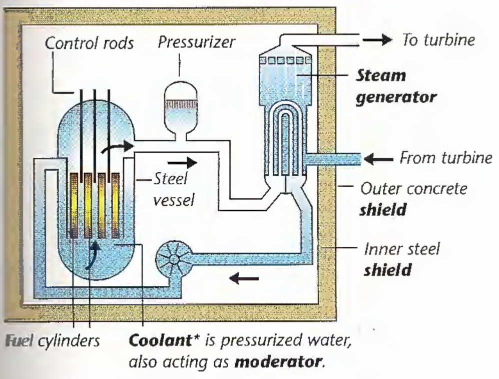
Control rods
Pressurizer
To turbine
Steam generator
From turbine
Steel vessel
Outer concrete **shield**
Inner steel **shield**
Fuel cylinders

Coolant* is pressurized water, also acting as **moderator**.

Advanced gas-cooled reactor (AGR)

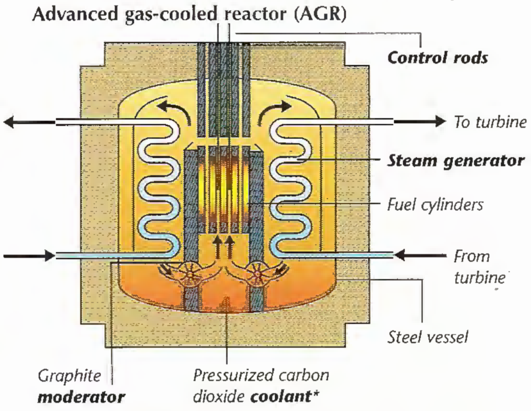
Control rods
To turbine
Steam generator
Fuel cylinders
From turbine
Steel vessel
Graphite **moderator**
Pressurized carbon dioxide **coolant***

Fast reactor or fast breeder reactor (FBR)

A **fission reactor**, inside which the **neutrons*** which cause the fission are allowed to remain as fast neutrons (traveling at about 2×10^7m s^{-1}). The fuel used is always enriched with extra nuclei of uranium-235 (see **thermal reactor**) and plutonium-239. Both of these will fission easily when hit by fast neutrons, unlike uranium-238, which is far more likely to "capture" the neutrons (becoming ^{239}U) and undergo **radioactive decay***. The final product of this decay, however, is ^{239}Pu. Fast reactors are also called "breeders" because this decay process of ^{239}U to ^{239}Pu is allowed to happen in a blanket of ^{238}U around the main fuel. Hence more fuel is created and can be stored. Fast reactors have a more compact core and run at higher temperatures than thermal reactors. They are also more efficient, using up a much greater proportion of their fuel before it becomes contaminated.

Fast reactor

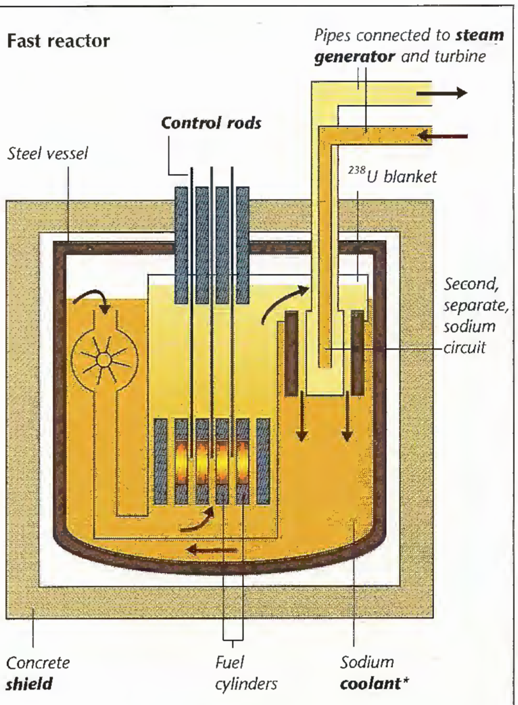
Pipes connected to **steam generator** and turbine
Control rods
Steel vessel
^{238}U blanket
Second, separate, sodium circuit
Concrete **shield**
Fuel cylinders
Sodium **coolant***

---

* **Chain reaction**, 92 (**Induced fission**); **Coolant**, 344; **Isotopes**, 83; **Neutrons**, 82; **Radioactive decay**, 87.

QUANTITIES AND UNITS

Physical quantities are such things as **mass***, **force*** and **current***, which are used in the physical sciences. They all have to be measured in some way and each therefore has its own **unit**. These are chosen by international agreement and are called **International System** or **SI units** – abbreviated from the French Système International d'Unités. All quantities are classified as either **basic quantities** or **derived quantities**.

Basic quantities

A set of quantities from which all other quantities (see **derived quantities**) can be defined (see table, below). Each basic quantity has its **basic SI unit**, in terms of which any other SI unit can be defined.

Basic quantity	Symbol	Basic SI unit	Abbreviation
Mass	*m*	*kilogram*	*kg*
Time	*t*	*second*	*s*
Length	*l*	*meter*	*m*
Current	*I*	*ampere*	*A*
Temperature	*T*	*kelvin*	*K*
Quantity of substance	–	*mole*	*mol*
Luminous intensity	–	*candela*	*cd*

Prefixes

A given SI unit may sometimes be too large or small for convenience, e.g. the meter is too large for measuring the thickness of a piece of paper. Standard fractions and multiples of the SI units are therefore used and written by placing a prefix before the unit (see table below). For example, the millimeter (mm) is equal to one thousandth of a meter.

Fractions and multiples in use

Fraction or multiple	Prefix	Symbol
10^{-9}	*nano-*	*n*
10^{-6}	*micro-*	*µ*
10^{-3}	*milli-*	*m*
10^{-2}	*centi-*	*c*
10^{-1}	*deci-*	*d*
10^{1}	*deca-*	*dc*
10^{2}	*hecto-*	*h*
10^{3}	*kilo-*	*k*
10^{6}	*mega-*	*M*
10^{9}	*giga-*	*G*

Basic SI units

Kilogram (kg)

The SI unit of mass. It is equal to the mass of an international prototype metal cylinder kept at Sèvres, near Paris.

Second (s)

The SI unit of time. It is equal to the duration of 9,192,631,770 **periods*** of a certain type of radiation emitted by the cesium-133 atom.

Meter (m)

The SI unit of length. It is equal to the distance light travels in a vacuum in $^{1}/_{299,792,458}$ of a second.

Ampere (A)

The SI unit of electric current (see also page 60). It is equal to the size of a current flowing through parallel, infinitely long, straight wires in a vacuum that produces a force between the wires of 2×10^{-7}N every meter.

Kelvin (K)

The SI unit of temperature. It is equal to $^{1}/_{273.16}$ of the temperature of the **triple point** of water (the point at which ice, water and steam can all exist at the same time) on the **absolute temperature scale***.

Mole (mol)

The SI unit of the quantity of a substance (note that this is different from mass because it is the number of particles of a substance). It is equal to the amount of substance which contains 6.023×10^{23} (this is **Avogadro's number**) particles (e.g. atoms or molecules).

Candela (cd)

The SI unit of intensity of light. It is equal to the strength of light from $^{1}/_{600,000}$ square meters of a **black body*** at the temperature of freezing platinum and at a pressure of 101,325N m^{-2}.

---

* **Absolute temperature scale**, 27; **Black body**, 29 (**Leslie's cube**); **Current**, 60; **Force**, 6; **Mass**, 12; **Period**, 16.

Derived quantities

Quantities other than ba
derived quantities. The
the **basic SI units** or oth
the quantity and are son

Derived quantity	Symb
Velocity	*v*
Acceleration	*a*
Force	*F*
Work	*W*
Energy	*E*
Power	*P*
Area	*A*
Volume	*V*
Density	*p*
Pressure	*P*
Period	*T*
Frequency	*f*
Impulse	–
Momentum	–
Electric charge	*Q*
Potential difference	*V*
Capacitance	*C*
Resistance	*R*

Derived quantities

Quantities other than **basic quantities** which are defined in terms of these or in terms of other derived quantities. The derived quantities have **derived SI units** which are defined in terms of the **basic SI units** or other derived units. They are determined from the defining equation for the quantity and are sometimes given special names.

Derived quantity	Symbol	Defining equation	Derived SI unit	Name of unit	Abbreviation
Velocity	v	$v = \frac{\text{change in displacement}}{\text{time}}$	$m\ s^{-1}$	–	–
Acceleration	a	$a = \frac{\text{change in velocity}}{\text{time}}$	$m\ s^{-2}$	–	–
Force	F	$F = \text{mass} \times \text{acceleration}$	$kg\ m\ s^{-2}$	newton	N
Work	W	$W = \text{force} \times \text{distance}$	N m	joule	J
Energy	E	Capacity to do work	J	–	–
Power	P	$P = \frac{\text{work done}}{\text{time}}$	$J\ s^{-1}$	watt	W
Area	A	Depends on shape (see page 101)	m^2	–	–
Volume	V	Depends on shape (see page 101)	m^3	–	–
Density	ρ	$\rho = \frac{\text{mass}}{\text{volume}}$	$kg\ m^{-3}$	–	–
Pressure	P	$P = \frac{\text{force}}{\text{area}}$	$N\ m^{-2}$	pascal	Pa
Period	T	Time for one cycle	s	–	–
Frequency	f	Number of cycles per second	s^{-1}	hertz	Hz
Impulse	–	$\text{Impulse} = \text{force} \times \text{time}$	N s	–	–
Momentum	–	$\text{Momentum} = \text{mass} \times \text{velocity}$	$kg\ m\ s^{-1}$	–	–
Electric charge	Q	$Q = \text{current} \times \text{time}$	A s	coulomb	C
Potential difference	V	$V = \frac{\text{energy transferred}}{\text{charge}}$	$J\ C^{-1}$	volt	V
Capacitance	C	$C = \frac{\text{charge}}{\text{potential difference}}$	$C\ V^{-1}$	farad	F
Resistance	R	$R = \frac{\text{potential difference}}{\text{current}}$	$V\ A^{-1}$	ohm	Ω

EQUATIONS, SYMBOLS AND GRAPHS

All **physical quantities** (see pages 96-97) and their units can be represented by **symbols** and are normally dependent in some way on other quantities. There is therefore a relationship between them which can be expressed as an **equation** and shown on a **graph**.

Equations

An **equation** represents the relationship between two or more physical quantities. This relationship can be expressed as a **word equation** or as an equation relating **symbols** which represent the quantities. The latter is used when a number of quantities are involved, since it is then easier to manipulate. Note that the meaning of the symbols must be stated.

Word equation

$$\text{Density} = \frac{\text{mass}}{\text{volume}}$$

Symbol equation

$$Q = m \times c \times (t_2 - t_1) \text{ or } Q = mc(t_2 - t_1)$$

where Q = heat energy lost or gained; m = mass; c = specific heat capacity; t_1 and t_2 = temperatures.

Graphs

A **graph** is a visual representation of the relationship between two quantities. It shows how one quantity depends on another. Points on a graph are plotted using the values for the quantities obtained during an experiment or by using the equation for the relationship if it is known. The two quantities plotted are called the **variables**.

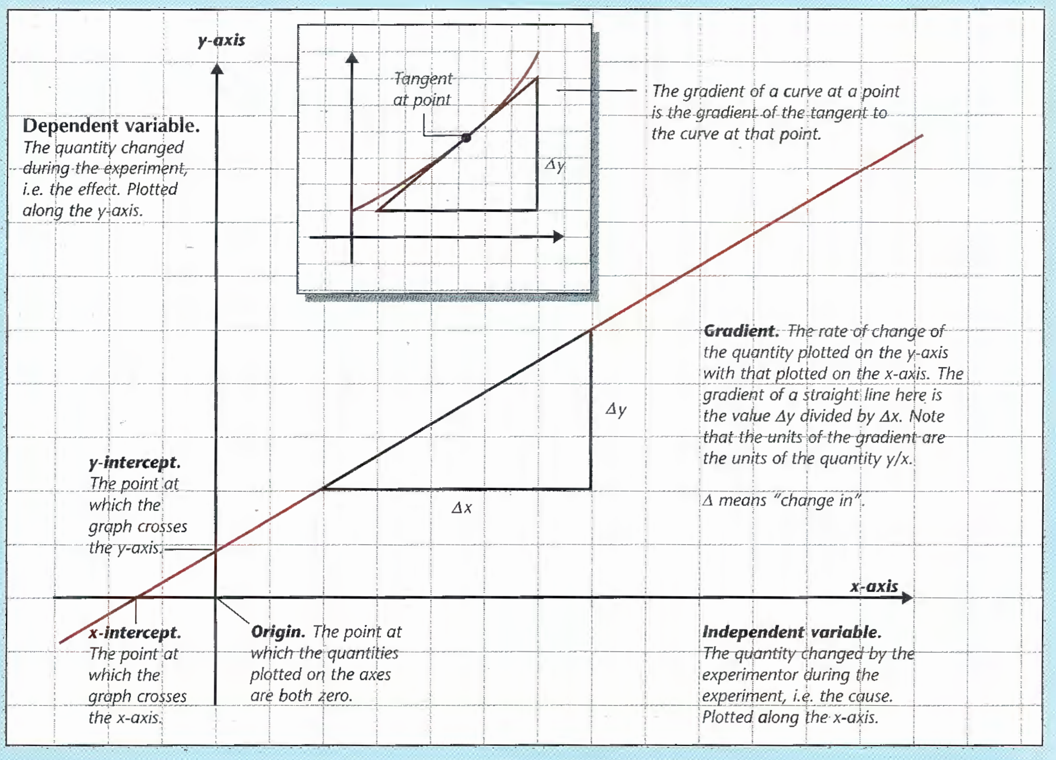

The gradient of a curve at a point is the gradient of the tangent to the curve at that point.

Dependent variable. *The quantity changed during the experiment, i.e. the effect. Plotted along the y-axis.*

Gradient. *The rate of change of the quantity plotted on the y-axis with that plotted on the x-axis. The gradient of a straight line here is the value Δy divided by Δx. Note that the units of the gradient are the units of the quantity y/x.*

Δ means "change in".

y-intercept. *The point at which the graph crosses the y-axis.*

x-intercept. *The point at which the graph crosses the x-axis.*

Origin. *The point at which the quantities plotted on the axes are both zero.*

Independent variable. *The quantity changed by the experimenter during the experiment, i.e. the cause. Plotted along the x-axis.*

Symbols

Symbols are used to rep
quantities. The value of
consists of a numerical v
Therefore any symbol re
number and a unit.

Symbols represent numb
m = 2.1kg, or s = 400J k

Current through resistor
ecessary to say I amps s
cluded).

Note that a symbol divid
umber, e.g. m = 2.1kg m

Plotting graphs

he quantity trolled during an periment should be otted along the axis, and the quantity hich changes as a esult along the y-axis.

2 Scales on the axes hould have values that e easy to find. (Avoid quares representing tiples of three.)

3 The axes should be eled by the symbol esenting the ntity (or name of uantity) and its e.g. length/mm.

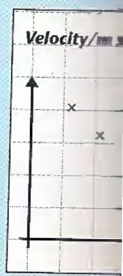

4 Points on the graph uld be marked in with a × or a ⊙.

5 smooth curve or ight line should be n which best fits points (this is ause physical ntities are normally ted in some definite Note that joining points up will not produce a smooth e. This is due to imental errors.

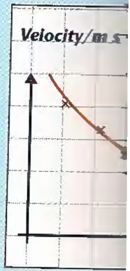

Symbols

Symbols are used to represent **physical quantities**. The value of a physical quantity consists of a numerical value and its unit. Therefore any symbol represents both a number and a unit.

Symbols represent number and unit, e.g. $m = 2.1\text{kg}$, or $s = 400\text{J kg}^{-1}\text{K}^{-1}$.

"Current through resistor = I" (i.e. it is not necessary to say I amps since the unit is included).

Note that a symbol divided by a unit is a pure number, e.g. $m = 2.1\text{kg}$ means that $m/\text{kg} = 2.1$.

This notation is used in tables and to label graph axes.

Any number in this column is a length in meters.

Any number in this column is a time squared measured in seconds squared.

l/m	t^2/s^2
0.9	3.6
1.0	4.0
1.1	4.4
1.2	4.8

Any number on this scale is a force in newtons.

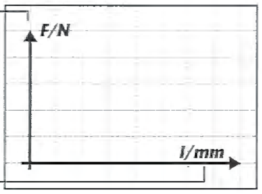

Any number on this scale is a length in millimeters.

Plotting graphs

1. The quantity controlled during an experiment should be plotted along the x-axis, and the quantity which changes as a result along the y-axis.

2. Scales on the axes should have values that are easy to find. (Avoid squares representing multiples of three.)

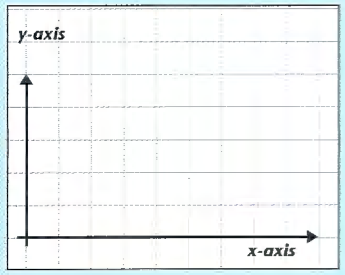

3. The axes should be labeled by the symbol representing the quantity (or name of the quantity) and its unit, e.g. length/mm.

4. Points on the graph should be marked in pencil with a × or a ⊙.

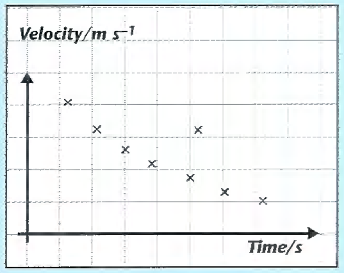

5. A smooth curve or straight line should be drawn which best fits the points (this is because physical quantities are normally related in some definite way). Note that joining the points up will not often produce a smooth curve. This is due to experimental errors.

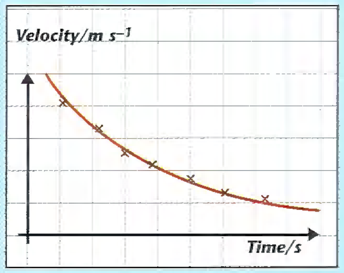

Information from graphs

A straight line graph which passes through the origin shows that the quantities plotted on the axes are proportional to each other (i.e. if one is doubled then so is the other).

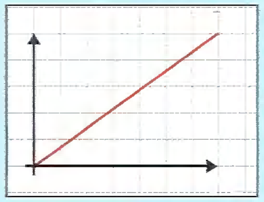

A straight line portion of a graph shows the region in which the relationship between two quantities is linear (i.e. one always changes by the same amount for a fixed change in the other).

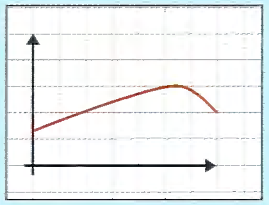

The amount of scatter of the points about the smooth curve gives an indication of the errors in the data due to inaccuracies in the procedure, the equipment and the measuring (this happens in any experiment).

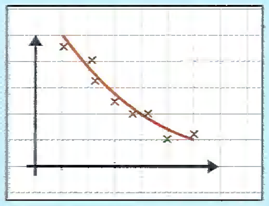

Individual points a long way from the curve are probably due to an error in measuring that piece of data in the experiment. However, the point should not be ignored – it should be checked and remeasured if possible.

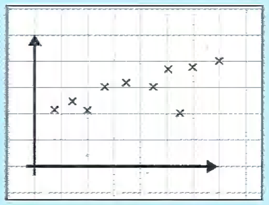

MEASUREMENTS

Measurement of length

The method used to measure a length depends on the magnitude of the length. A meter ruler is used for lengths of 50mm or more. The smallest division is normally 1mm and so lengths can be estimated to the nearest 0.5mm. For lengths less than 50mm, the error involved would be unacceptable (see also **reading error**, page 103). A **vernier scale** is therefore used. For the measurement of very small lengths (to 0.01mm) a **micrometer screw gauge** is used (see opposite).

Vernier scale

A short scale which slides along a fixed scale. The position on the fixed scale of the zero line of the vernier scale can be found accurately. It is used in measuring devices such as the **vernier slide callipers**.

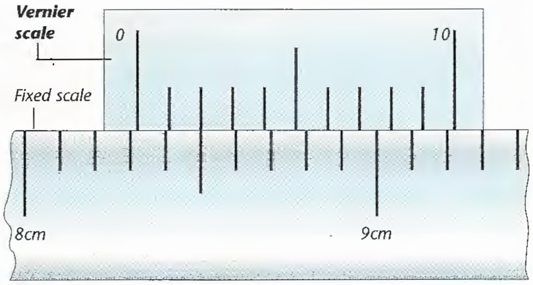

Method of reading position of zero line on vernier scale:

1. Read the position of the zero line approximately – in this case 8.3cm.

2. Find the position on the vernier scale where the marks coincide – in this case 2.

3. Add this to the previous figure – the accurate reading is 8.32cm.

Vernier slide callipers

An instrument containing a **vernier scale**, used to measure lengths in the range 10 to 100mm.

Method of measurement:

1. Close the jaws and check that the zero on the **vernier scale** coincides with the zero on the fixed scale. If not, note the reading (this is the **zero error***).

2. Close or open the jaws onto the object to be measured.

3. Lock the sliding jaw into position.

4. Record the reading on the scale.

5. Add or subtract the zero error (see 1) to get the correct reading.

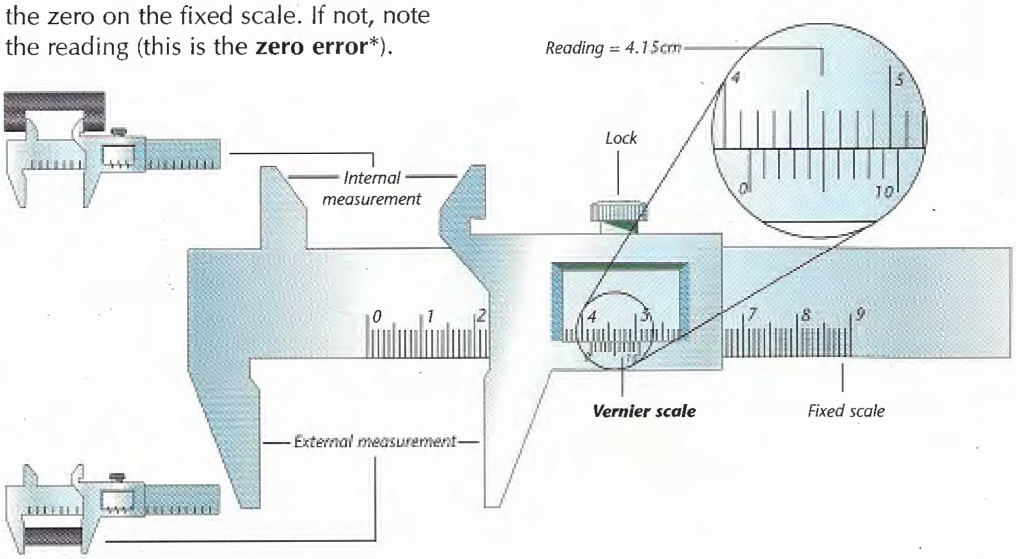

---

* **Zero error**, 102.

Micrometer screw ga

An instrument used for
accurate measuremen
to about 30mm.

Method of measurem

1. Determine the value
division on the spindle
(see diagram).

2. Using the ratchet, clo
instrument fully. The zer
should coincide with the
line. If not, note the zero

3. Using the ratchet, clo
object to be measured u

4. Note the reading of th
mark on the sleeve scale
.5mm).

Measurement

The volume of a liquid
space it takes up in its c
internal volume of the c
called its **capacity**. The
is the **liter** (**l**), equal to 1
1ml = 1cm³. The volume
measured using a gradu

The surface area and vo
regular shape are calcul
measurements of the obj

Regular shaped solid	Re
Measurements made using **vernier slide callipers** or **micrometer screw gauge**	
Volume V of solid calculated from	
Surface area A calculated from	

Micrometer screw gauge

An instrument used for accurate measurements up to about 30mm.

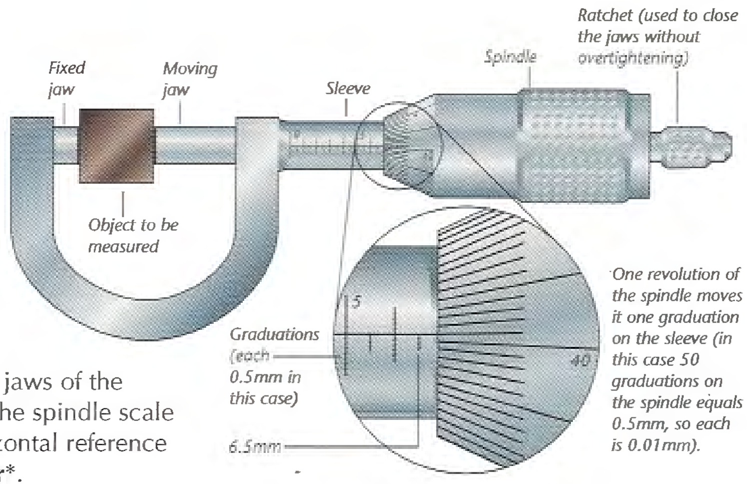
Fixed jaw
Moving jaw
Sleeve
Spindle
Ratchet (used to close the jaws without overtightening)
Object to be measured
5
40
Graduations (each 0.5mm in this case)
6.5mm
One revolution of the spindle moves it one graduation on the sleeve (in this case 50 graduations on the spindle equals 0.5mm, so each is 0.01mm).

Method of measurement:

1. Determine the value of a division on the spindle scale (see diagram).

2. Using the ratchet, close the jaws of the instrument fully. The zero on the spindle scale should coincide with the horizontal reference line. If not, note the **zero error***.

3. Using the ratchet, close the jaws on the object to be measured until it is gripped.

4. Note the reading of the highest visible mark on the sleeve scale (in this case 6.5mm).

5. Note the division on the spindle scale which coincides with the horizontal reference line (in this case 0.41mm).

6. Add the two readings and add or subtract the zero error (see 2) to get the correct reading (in this case 6.91mm).

Measurement of area and volume

The volume of a liquid is calculated from the space it takes up in its containing vessel. The internal volume of the containing vessel is called its **capacity**. The **SI unit*** of capacity is the **liter** (**l**), equal to $10^{-3}m^3$. Note that $1ml = 1cm^3$. The volume of a liquid is measured using a graduated vessel.

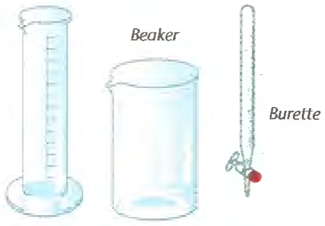
Measuring cylinder
Beaker
Burette

Examples of graduated vessels for measuring volume

The surface area and volume of a solid of regular shape are calculated from length measurements of the object (see below).

For solids of irregular shape, see **eureka can**, page 24.

Regular shaped solid	*Rectangular bar* 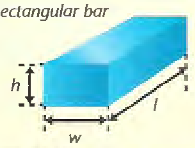	*Sphere* 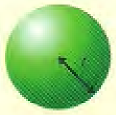	*Cylinder* 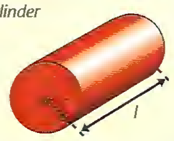
Measurements made using ***vernier slide callipers*** *or* ***micrometer screw gauge***	*h = height* *w = width* *l = length*	*r = radius*	*r = radius* *l = length*
Volume V of solid calculated from	$V = lwh$	$V = \frac{4}{3}\pi r^3$	$V = \pi r^2 l$
Surface area A calculated from	$A = \underset{\text{Top}}{2wl} + \underset{\text{Sides}}{2hl} + \underset{\text{Ends}}{2hw}$	$A = 4\pi r^2$	$A = \underset{\text{Curved surface}}{2\pi rl} + \underset{\text{Ends}}{2\pi r^2}$

---

* **SI units**, 96; **Zero error**, 102.

ACCURACY AND ERRORS

All experimental measurements are subject to some errors, other than those caused by carelessness (like misreading a scale). The most common errors which occur are **parallax errors**, **zero errors** and **reading errors**. When stating a reading, therefore, a number of **significant figures** should be quoted which give an estimate of the accuracy of the readings.

Parallax error

The error which occurs when the eye is not placed directly opposite a scale when a reading is being taken.

Parallax error reading a meter ruler

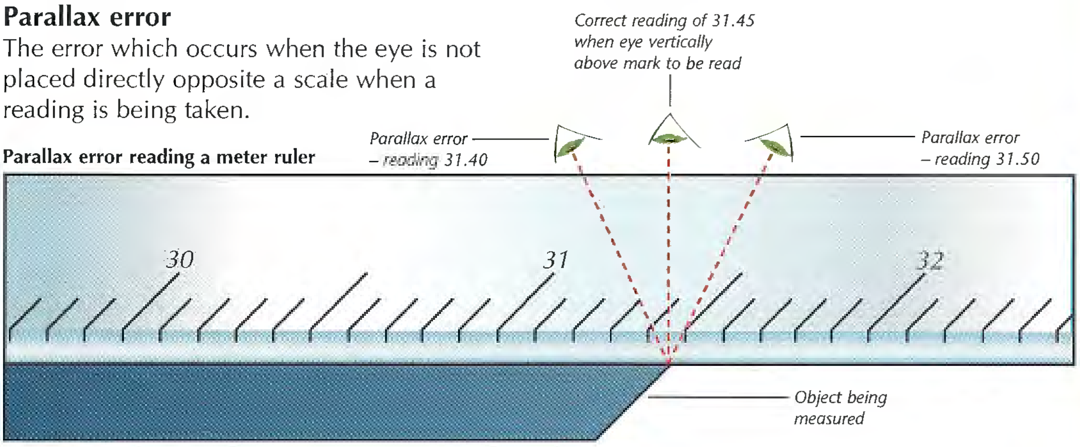

Correct reading of 31.45 when eye vertically above mark to be read

Parallax error – reading 31.40

Parallax error – reading 31.50

Object being measured

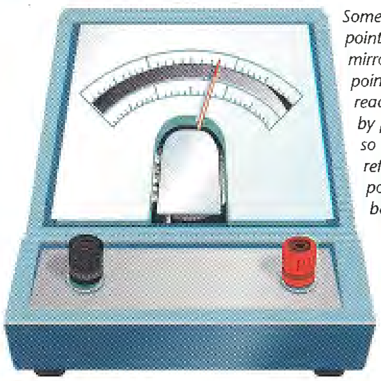

Some scales with pointers have a mirror behind the pointer. The correct reading is obtained by placing the eye so that the reflection of the pointer is hidden behind it.

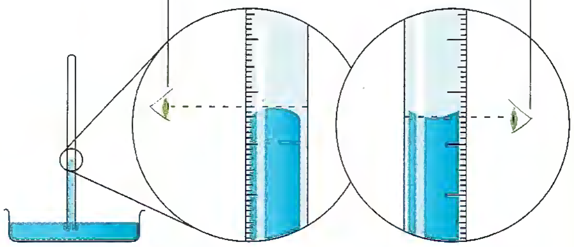

*To avoid parallax errors, readings of liquid levels must be taken with the eye lined up with the top or bottom of the **meniscus***.*

Zero error

The error which occurs when a measuring instrument does not indicate zero when it should. If this happens, the instrument should either be adjusted to read zero or the inaccurate "zero reading" should be taken and should be added to or subtracted from any other reading taken.

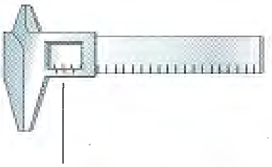

*Reading on **vernier slide callipers*** when closed (i.e. should read zero) is 0.2mm. This is **zero error**.*

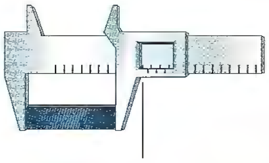

0.2mm must be subtracted from any reading (in this case, apparent reading is 53.9 but actual length is 53.9 – 0.2, i.e. 53.7mm).

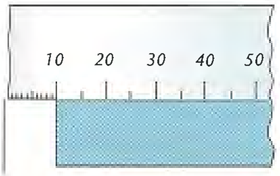

Zero error on meter ruler may be due to worn end. Should be solved by measuring from 10mm line and subtracting 10mm from all readings.

---

* **Meniscus**, 345; **Vernier slide callipers**, 100.

Reading error

The error due to the gu
in taking a reading fron
reading lies between th

*Reading on thermometer is betwe
estimate of next figure is half a di
36.85°C.*

Significant fig

The number of **signific**
is the number of figur
ignoring leading or tra
below) and disregard
decimal point. They gi
the accuracy of a read

*A reading of 3704mm has four
figures. It can be written as:*

3 704mr

3.704m

0.003 70

1st significant figure

*Note that the leading zeros here
but show the magnitude of the*

*The number of significant figures
accuracy of a reading or result.*

Smallest division on ammeter scale

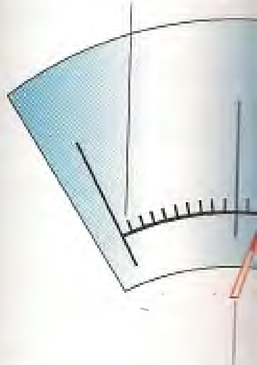

*The best guess is half a division, so
Three significant figures indicate tha
about 0.05A.*

*A reading with more figures, e.g. 1.2
than is possible on this scale.*

***Meniscus**, 345.

Reading error

The error due to the guesswork involved in taking a reading from a scale when the reading lies between the scale divisions.

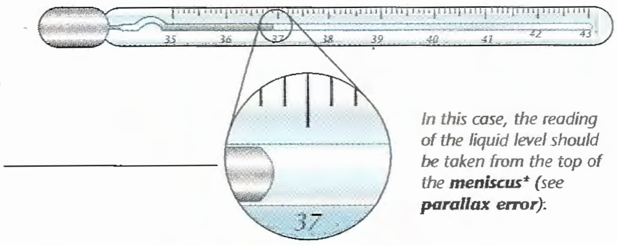

Reading on thermometer is between 36.8°C and 36.9°C. Best estimate of next figure is half a division to give a reading of 36.85°C.

*In this case, the reading of the liquid level should be taken from the top of the **meniscus*** *(see* ***parallax error****).*

Significant figures

The number of **significant figures** in a value is the number of figures in that value ignoring leading or trailing zeros (but see below) and disregarding the position of the decimal point. They give an indication of the accuracy of a reading.

A reading of 3704mm has four ***significant figures****. It can be written as:*

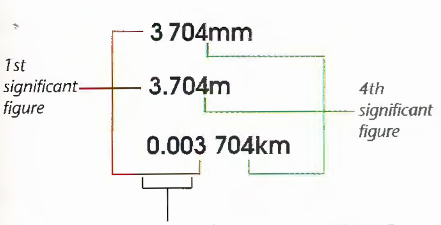

Note that the leading zeros here are not significant figures but show the magnitude of the reading.

The number of significant figures quoted is an indication of the accuracy of a reading or result.

Smallest division on ammeter scale = 0.1A

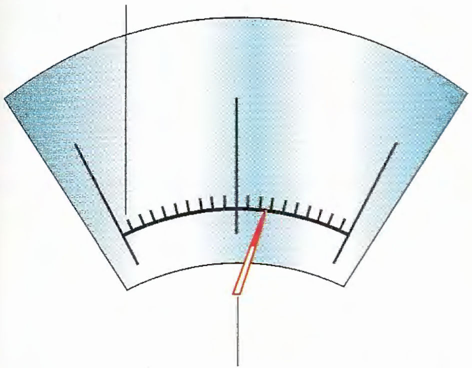

The best guess is half a division, so the reading is given as 1.25A. Three significant figures indicate that the reading is accurate to about 0.05A.

A reading with more figures, e.g. 1.2518A, implies more accuracy than is possible on this scale.

Rounding

The process of reducing the number of figures quoted. The last significant figure is dropped and the new last figure changed depending on the one dropped.

7.3925 *(quoted to 5 significant figures)*
= 7.393 *(rounded to 4 significant figures)*
= 7.39 *(rounded to 3 significant figures)*
= 7.4 *(rounded to 2 significant figures)*
= 7 *(rounded to 1 significant figure)*

0.08873 *(quoted to 4 significant figures)*
= 0.0887 *(rounded to 3 significant figures)*
= 0.089 *(rounded to 2 significant figures)*
= 0.09 *(rounded to 1 significant figure)*

Note that
29.000 *is quoted to 5 significant figures*
= 29.0 *(to 3 significant figures)*
= 29 *(to 2 significant figures)*
= 30 *(to 1 significant figure)*
In the last case here, the 0 is not a significant figure but must be included (see below).

For large numbers like 283,000 it is impossible to say how many of the figures are significant (the first three must be) because the zeros have to be included to show the magnitude. This ambiguity is removed by using the **exponential notation** (see page 109).

---

* **Meniscus**, 345.

FIELDS AND FORCES

This table is a comparison of the three forces normally encountered in physics (excluding the **nuclear force**). In fact, most of the forces dealt with in physics, e.g. the **contact force** between two objects, are examples of the **electromagnetic force** which is a combination of the **magnetic** and **electric forces**. For more about these and all other forces, see pages 6-7.

<table>
<tr><th></th><th>Gravitational force
(see also pages 6 and 18)</th><th>Electric force
(see also pages 6 a</th></tr>
<tr><td>Type of force
Note that a force can only exist between two masses, charges or currents, and that the size of the force is the same on both of them (see also Newton's third law, page 13). Note also that the forces only act between objects which are the same, e.g. there is a force between two masses, but not between a mass and a current.</td><td>Force acts between two objects with mass. It is always attractive.
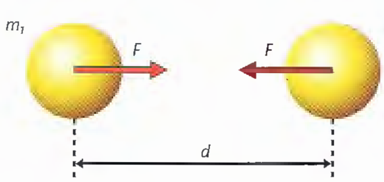
$F = G\frac{m_1m}{d^2}$
$G = 6.7 \times 10^{-11}Nm^2\ kg^{-2}$
G is the gravitational constant*. Its very small value means that the gravitational force is only noticeable when one of the objects is very large (e.g. a planet).</td><td>Force is between two
if the charges are of th
one negative and one
repulsive if the charg
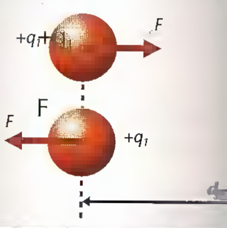
$F = K$
In air $K_E = 9 \times 10^9$Nm
This very large value m
to separate opposite ch</td></tr>
<tr><td>Description of force in terms of force field
The force field is the region around an object (mass, charge or current) in which its effects (gravitational, electric or magnetic) can be detected – see also page 6.</td><td>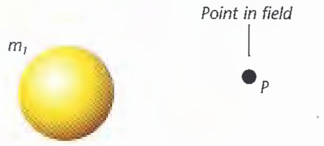
Mass m_1 produces a gravitational field in the space around it (see field intensity, page 106).
A second mass experiences a gravitational force when placed at any point (e.g. P) in the gravitational field of m_1.
A mass thus produces a gravitational field and is acted upon by a gravitational field.</td><td>q_1

Charge q_1 produces an
space around it (see fie
A second charge experi
force when placed at a
in the electric field of
A charge thus produce
is acted upon by an ele</td></tr>
<tr><td>Field direction
This is found by observing the effect of the force field on an object (mass, charge or current) placed in it.</td><td>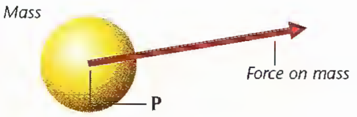
The direction of a gravitational field at a point P is the direction of the force on a mass placed at P.</td><td>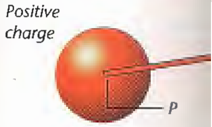
The direction of an elec
is the direction of the fo
charge placed at P.</td></tr>
</table>

---

* **Gravitational constant**, 18 (**Newton's law of gravitation**)

* **Electric field**, 58; **Fleming's lef**
Magnetic field, 72.

Electric force (see also pages 6 and 58)	**Magnetic force** (see also pages 6 and 70)
Force is between two charges. It is attractive if the charges are of the opposite sign, i.e. one negative and one positive, and repulsive if the charges are of the same sign.	Force is between two objects in which current is flowing. If the currents flow in the same direction, the force is attractive. If the currents flow in opposite directions, the force is repulsive.
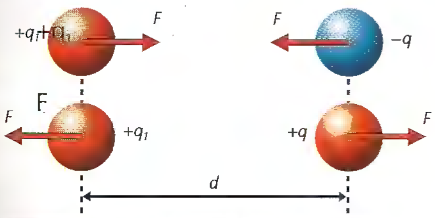	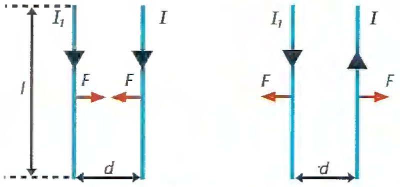
$F = K_E \frac{q_1 q}{d^2}$	$F = K_B \frac{I_1 I l}{d}$
In air $K_E = 9 \times 10^9 Nm^2 C^{-2}$	In air $K_B = 2.7 \times 10^{-7} N A^{-2}$
This very large value means it is difficult to separate opposite charges.	This small value indicates that the magnetic force is very small in comparison to the electric force.
q1 q1 Point in field P	I1 Point in field P
Charge q_1 produces an **electric field*** in the space around it (see **field intensity**, page 106). A second charge experiences an electric force when placed at any point (e.g. P) in the electric field of q_1. A charge thus produces an electric field and is acted upon by an electric field.	Current I_1 produces a **magnetic field*** in the space around it (see **field intensity**, page 106). A second current experiences a magnetic force when placed at any point P in the magnetic field of I_1. A current thus produces a magnetic field and is acted upon by a magnetic field.
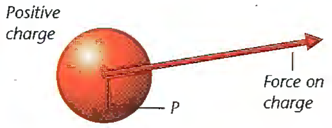	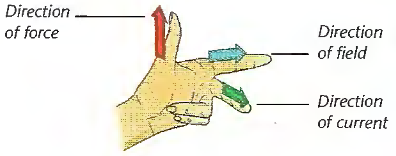 Direction of force Direction of field Direction of current
The direction of an electric field at a point P is the direction of the force on a positive charge placed at P.	The direction of a magnetic field at a point P is given by **Fleming's left hand rule***.

---

* **Electric field**, 58; **Fleming's left hand rule**, 76; **Magnetic field**, 72.

Fields and forces (continued)

Field intensity
This is found by measuring the effect of the force field on an object (mass, charge or current) placed in it.

Representation by field lines
Field lines (or **flux lines** or **lines of force** or **flux**) are used in all cases to represent the strength and direction of fields and to visualize them (see panel at bottom of page). Field lines never cross since the field would then have different directions at one point.

Potential energy (see also page 8)
This depends on **field intensity** and the object (its mass in a **gravitational field** or its charge in an **electric field***). The **potential*** at a point in a field is the energy per unit (of mass or charge) and depends upon the field only. Usually the only concern is the difference in potential, or **potential difference***, between two points. Potential can be defined by choosing a reference. The potential at a point is then the potential difference between the point and the reference point.

Gravitational force

To measure the field intensity g of a gravitational field due to a mass m_1 at a point P, a test mass m is placed at P and the gravitational force F on it is measured. Then:

$$g = \frac{\text{gravitational force } (F)}{\text{mass } (m)} \quad \text{or:} \quad F = mg$$

By comparison with the equation above for the gravitational force, the field intensity g at a distance d from a mass m_1 is:

$$g = G\frac{m_1}{d^2}$$

Gravitational field lines always end at a mass.

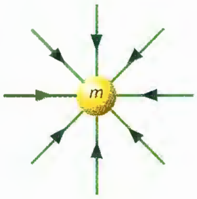

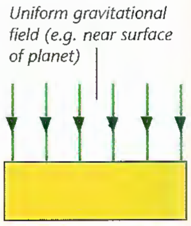

The gravitational potential difference between two points in a gravitational field is the work done against the forces of the field in moving a unit mass between the points.

$$\text{Gravitational potential difference} = \frac{\text{work done}}{\text{mass}}$$

Gravitational potential decreases as point moves along field line in direction of field (in direction of arrow).

Gravitational potential higher at P_1 than P_2.

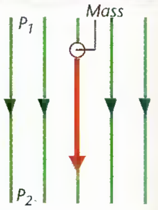

Electric force

To measure the field int
field at a point P due to
positive charge q is pla
electric force F on it is

$$E = \frac{\text{electric force } (F)}{\text{charge } (q)}$$

By comparison with the
the electric force, the fie
distance d from a charg

$$E = K_E\frac{q}{d}$$

Electric field lines alway
charge and end at an equ

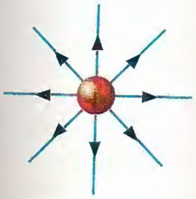

The electric potential diffe
two points in an electric f
done against the forces of
a unit positive charge betw

$$\text{Electric potential difference} =$$

Electric potential decrease
in direction of field (in
direction of arrow).

Electric potential higher
at point P_1 than at P_2.

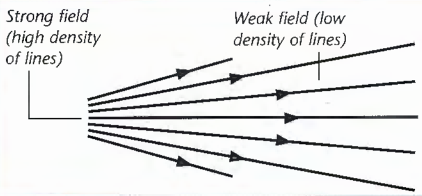
Strong field (high density of lines)
Weak field (low density of lines)

Uniform field has constant strength and direction.

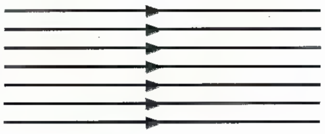

on-uniform field

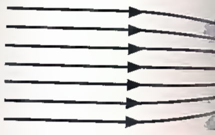

*** Electric field**, **Potential**, **Potential difference**, 58.

agnetic field lines, 72.

Electric force	Magnetic force
To measure the field intensity E of an electric field at a point P due to a charge q_1, a test positive charge q is placed at P and the electric force F on it is measured. Then: $E = \frac{\text{electric force }(F)}{\text{charge }(q)}$ or: $F = qE$ By comparison with the equation above for the electric force, the field intensity E at a distance d from a charge q_1 is: $E = K_E \frac{q_1}{d^2}$	To measure the field intensity B of a magnetic field due to I_1 at a point P, a conductor of length l carrying a current I is placed at P and the magnetic force F is measured. Then: $B = \frac{\text{magnetic force }(F)}{\text{current }(I) \times \text{length }(l)}$ or: $F = BIl$ By comparison with the equation above for the magnetic force, the field intensity B at a distance d from a current I_1 is: $B = K_B \frac{I_1}{d}$
Electric field lines always begin at a positive charge and end at an equal negative charge. 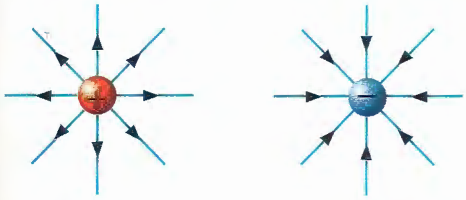	**Magnetic field lines*** have no beginning or end, but are always closed loops. This is because single north or south poles cannot exist. This is a fundamental difference compared to gravitational and electric fields. 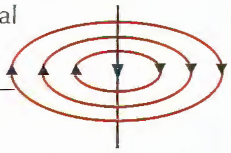 *Circular magnetic field lines around current-carrying wire.*
The electric potential difference between two points in an electric field is the work done against the forces of the field in moving a unit positive charge between them. $\text{Electric potential difference} = \frac{\text{work done}}{\text{charge}}$ Electric potential decreases in direction of field (in direction of arrow). Electric potential higher at point P_1 than at P_2. 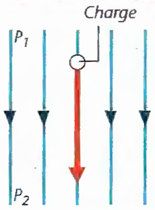 	Magnetic potential is much more difficult to define than for gravitational or electric fields because the field lines are circular. Note that if a point moves around a circular line in the diagram above, it returns to the same point, which must have the same potential, although it has moved along a field line. This means that magnetic potential is complicated to calculate.
Non-uniform field 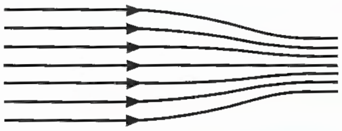	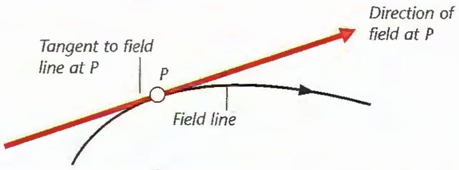 Tangent to field line at P P Direction of field at P Field line

---

* **Magnetic field lines**, 72.

VECTORS AND SCALARS

All quantities in physics are either **scalar** or **vector quantities**, depending on whether the quantity has direction as well as magnitude.

Scalar quantity

Any quantity which has magnitude only, e.g. mass, time, energy, density.

Vector quantity

Any quantity which has both magnitude and direction, e.g. force, displacement, velocity and acceleration. When giving a value to a vector quantity, the direction must be given in some way as well as the magnitude. Usually, the quantity is represented graphically by an arrowed line. The length of the line indicates the magnitude of the quantity (on some chosen scale) and the direction of the arrow indicates the direction of the quantity.

Parallelogram rule

A rule used when adding together two **vector quantities**. The two vectors are drawn from one point to form two sides of a parallelogram which is then completed. The diagonal from the original common point gives the sum of the two vectors (the **resultant**).

The parallelogram rule is used to help navigation at sea. The direction and speed of the tide must be taken into account as the second vector quantity to be added to the direction and speed of the boat.

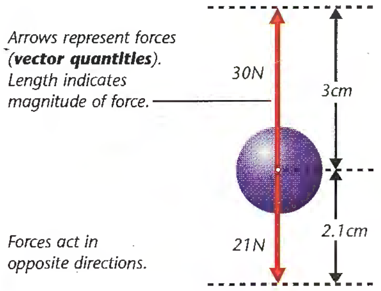
Arrows represent forces (**vector quantities**). Length indicates magnitude of force.
30N
3cm
Scale is 1cm = 10 newtons
2.1cm
21N
Forces act in opposite directions.

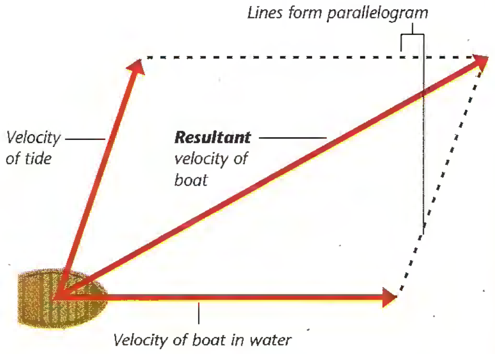
Lines form parallelogram
Velocity of tide
Resultant velocity of boat
Velocity of boat in water

Resolution

The process of splitting one **vector quantity** into two other vectors called its **components**. Normally, the two components are perpendicular to each other. Each component then represents the total effect of the vector in that direction.

Lift (**vector quantity**) from rotor of helicopter can be **resolved** into two **components**. The first acts upwards, in order to keep it airborne, and the second acts forwards, to move it along.

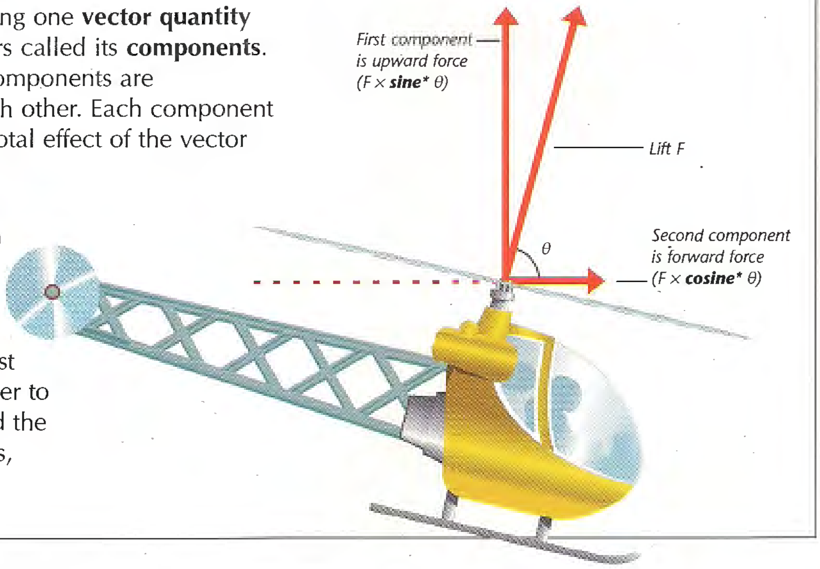
First component is upward force (F × **sine*** θ)
Lift F
θ
Second component is forward force (F × **cosine*** θ)

* **Cosine**, 344; **Sine**, 345.

NUMBERS

Very large or very s
a long time to write
notation is therefor
point is shown by w

1 000 000	$= 10^6$
100 000	$= 10^5$
10 000	$= 10^4$
1 000	$= 10^3$
100	$= 10^2$
10	$= 10^1$
1	$= 10^0$
0.1	$= 10^{-1}$
0.01	$= 10^{-2}$
0.001	$= 10^{-3}$
0.0001	$= 10^{-4}$
0.000 01	$= 10^{-5}$
0.000 001	$= 10^{-6}$

Scientific notation

A form of expressing num
number always has one d
point and is followed by
exponential notation to s
(see also **significant figure**

Order of magnitude

A value which is accura
of ten or so. It is impor
of the order of magnitud
quantities so that a figur

Typical orders of magnitude

Item
Earth
Car
Human
Bag of sugar
Orange
Golf ball
Table tennis ball
Proton
Electron

Item
Age of Earth
Time since emergence of man
Human life time
Time span of year
Time span of day
Time between heart beats
Camera shutter speed
Half-life of polonium-214
Time for light to travel 1m

† *In scientific work, the convention is*
used to make the numbers easier to
figures. The latter style is predomin

NUMBERS

Very large or very small numbers (e.g. 10 000 000 or 0.000 001) take a long time to write out and are difficult to read. The **exponential notation** is therefore used. In this notation, the position of the decimal point is shown by writing the power ten is raised to.

1 000 000 = 10^6 or "ten to the six"
100 000 = 10^5 or "ten to the five"
10 000 = 10^4 or "ten to the four"
1 000 = 10^3 or "ten to the three"
100 = 10^2 or "ten to the two"
10 = 10^1 or "ten to the one"
1 = 10^0 any number "to the nought" equals one
0.1 = 10^{-1} or "ten to the minus one"
0.01 = 10^{-2} or "ten to the minus two"
0.001 = 10^{-3} or "ten to the minus three"
0.0001 = 10^{-4} or "ten to the minus four"
0.000 01 = 10^{-5} or "ten to the minus five"
0.000 001 = 10^{-6} or "ten to the minus six"

Note that a negative exponent means "one over" so that $10^{-3} = {}^{1}/_{10^3} = {}^{1}/_{1\,000}$. This also applies to units, e.g. kg m^{-3} means kg/m^3 or kg per m^3.

Exponents are added when multiplying numbers, e.g. $10^5 \times 10^{-3}$ (= 100 000 × ${}^{1}/_{1\,000}$) = $10^{5-3} = 10^2 = 100$.

Scientific notation

A form of expressing numbers in which the number always has one digit before the decimal point and is followed by a power of ten in **exponential notation** to show its magnitude (see also **significant figures**, page 103).

Examples of numbers written in scientific notation[†]

56 342	*5.6342×10^4*
4 000	*4×10^3 (assuming 0s are not significant)*
569	*5.69×10^2*
23.3	*2.33×10^1*
0.98	*9.8×10^{-1}*
0.00211	*2.11×10^{-3}*

Order of magnitude

A value which is accurate to within a factor of ten or so. It is important to have an idea of the order of magnitude of some physical quantities so that a figure which has been calculated can be judged. For example, the mass of a person is about 60kg. Therefore a calculated result of 50kg or 70kg is quite reasonable, but a result of 6kg or 600kg is obviously not correct.

Typical orders of magnitude

Item	Mass/kg
Earth	*6×10^{24}*
Car	*5×10^3*
Human	*5×10^1*
Bag of sugar	*1*
Orange	*2×10^{-1}*
Golf ball	*5×10^{-2}*
Table tennis ball	*2×10^{-3}*
Proton	*2×10^{-27}*
Electron	*10^{-30}*

Item	Length/m
Radius of Milky Way galaxy	*10^{19}*
Radius of Solar System	*10^{11}*
Radius of Earth	*5×10^6*
Height of Mount Everest	*10^4*
Height of human	*2*
Thickness of paper	*10^{-4}*
Wavelength of light	*5×10^{-7}*
Radius of atom	*10^{-10}*
Radius of nucleus	*10^{-14}*

Item	Time/s
Age of Earth	*2×10^{17}*
Time since emergence of man	*10^{13}*
Human life time	*2×10^9*
Time span of year	*3×10^7*
Time span of day	*9×10^4*
Time between heart beats	*1*
Camera shutter speed	*10^{-2}*
Half-life of polonium-214	*1.5×10^{-4}*
Time for light to travel 1m	*3×10^{-9}*

Item	Energy/J
Energy given out by Sun per second	*10^{26}*
Energy released by San Francisco earthquake (1906)	*3×10^{17}*
Energy released by fission of 1g of uranium	*10^{11}*
Energy of lightning discharge	*10^9*
Energy of 1kW fire per hour	*4×10^6*
Kinetic energy of golf ball	*20*

[†] In scientific work, the convention is to print numbers up to 9999 closed up and without a comma. In numbers above this, small spaces are used to make the numbers easier to read at a glance. In non-technical writing, the convention is to add commas to numbers with four or more figures. The latter style is predominantly used in this book.

CIRCUIT SYMBOLS

This table shows the main symbols used to represent the various components used in electric circuits (see also pages 60-65).

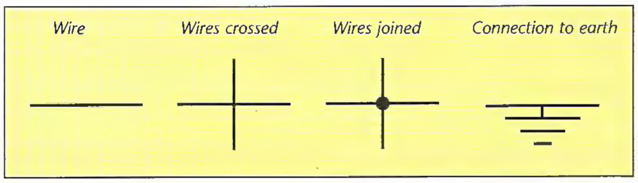
Wire
Wires crossed
Wires joined
Connection to earth

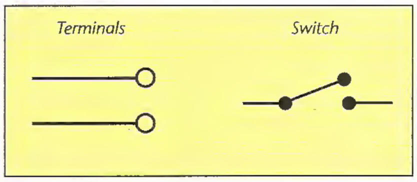
Terminals
Switch

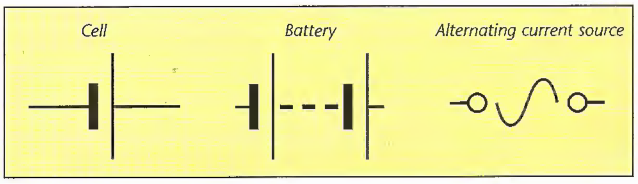
Cell
Battery
Alternating current source

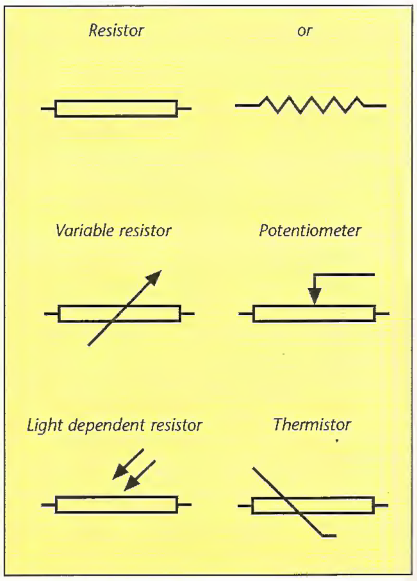
Resistor
or
Variable resistor
Potentiometer
Light dependent resistor
Thermistor

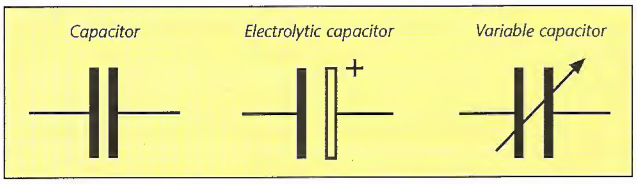
Capacitor
Electrolytic capacitor
+
Variable capacitor

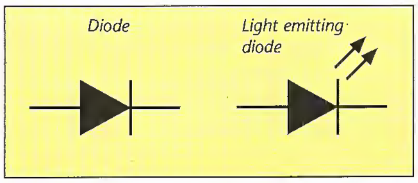
Diode
Light emitting diode

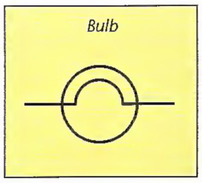

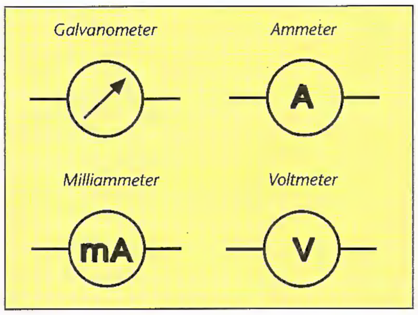
Galvanometer
Ammeter
A
Milliammeter
mA
Voltmeter
V

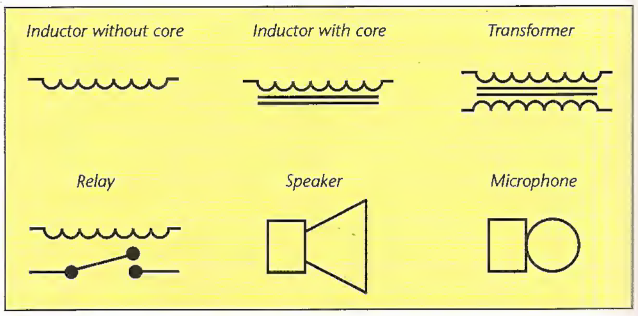
Inductor without core
Inductor with core
Transformer
Relay
Speaker
Microphone

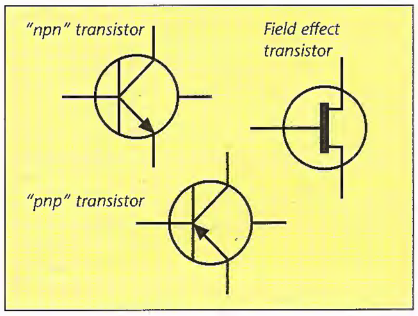
"npn" transistor
Field effect transistor
"pnp" transistor

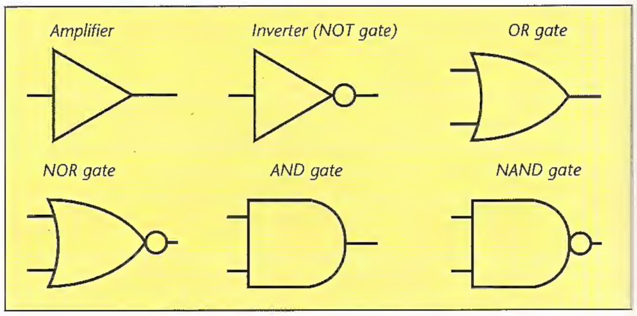
Amplifier
Inverter (NOT gate)
OR gate
NOR gate
AND gate
NAND gate

TRANSIST

Transistors* can be
microphone, and ar
use in complex circ
larger and slower va

Behavior of typical transisto
Potential difference* at collecto
Vb below about 0.6V means trans fully "off" – lamp goes out.
Vb in this range (about 0.6V to 0. gives Vc in range 0 to V. Transistor then amplifies signal at base.

Logic gates

The on and off states of a
indicate the numbers 1 an
circuits are therefore know
circuits are called **analog**
transistors with other com
make circuits which carry

Truth tables for basic logical ope

In	NOT
0	1
1	0

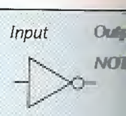
Input
Outp
NOT

A	B	OR	NOR
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

A	B	OR	NAND
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

Combinations of these gate
transistor circuits are used
circuits which can perform
operations, e.g. addition. T
integrated circuits, and ma
thousands of such compone
yet be built into a single sl

Potential difference, 58; **Relay**, 7

TRANSISTORS AND GATES

Transistors* can be used to amplify electrical signals, such as those from a microphone, and are also used as electronic switches. This has led to their use in complex circuits such as computers. They have replaced the much larger and slower valves and **relays***.

Behavior of typical transistor*

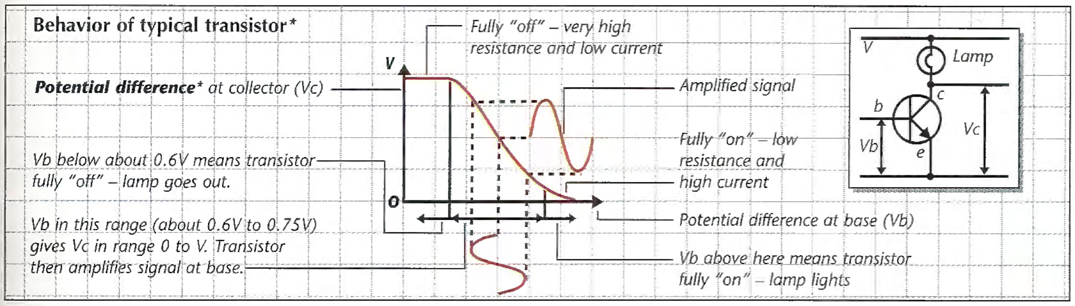
Fully "off" – very high resistance and low current
V
Potential difference* at collector (Vc)
Amplified signal
Fully "on" – low resistance and high current
Vb below about 0.6V means transistor fully "off" – lamp goes out.
0
Potential difference at base (Vb)
Vb in this range (about 0.6V to 0.75V) gives Vc in range 0 to V. Transistor then amplifies signal at base.
Vb above here means transistor fully "on" – lamp lights
V
Lamp
b
c
e
Vb
Vc

Logic gates

The on and off states of a transistor are used to indicate the numbers 1 and 0, respectively. The circuits are therefore known as **digital** (other circuits are called **analog**). Combinations of transistors with other components are used to make circuits which carry out logical operations.

Truth tables for basic logical operations

In	*NOT*
0	1
1	0

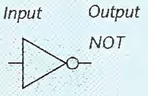
Input
Output
NOT

A	B	OR	NOR
0	0	0	1
0	1	1	0
1	0	1	0
1	1	1	0

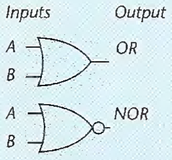
Inputs
Output
A
B
OR
A
B
NOR

A	B	OR	NAND
0	0	0	1
0	1	0	1
1	0	0	1
1	1	1	0

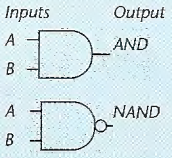
Inputs
Output
A
B
AND
A
B
NAND

Combinations of these gates and other transistor circuits are used to make complex circuits which can perform mathematical operations, e.g. addition. These are called **integrated circuits**, and may contain many thousands of such components and connections, yet be built into a single slice of silicon.

Computers

Integrated circuits mean that many thousands of logic gates can be put onto a single tiny component called a microchip. The CPU of a computer (see below) can be put on one chip.

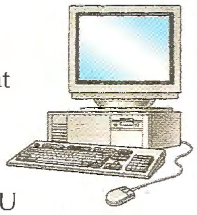

A personal computer with a CPU, monitor, keyboard and mouse

Typical computer system

*Disk and CD drives are examples of **input** and **output** devices. The disks can store many times more data than the **memory**, and also retain it when the computer is turned off (the information in the **read only memory** is lost).*

*Devices such as keyboard, mouse and screen are also **input** and **output** devices, from which data is put into the computer and to which it is sent. They are ways for the computer to link with the outside world.*

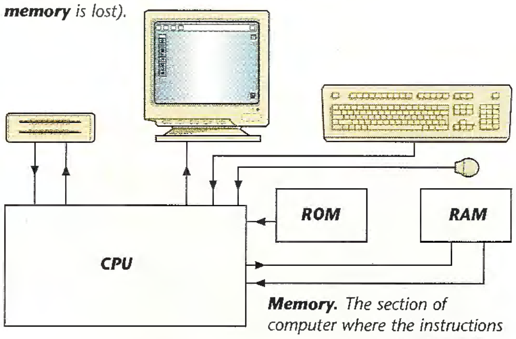

Central processing unit (CPU). *The center of the computer. It takes data from **memory** and **input** devices, performs operations on it (it can do this millions of times per second) and sends the results to memory or **output** devices.*

Memory. *The section of computer where the instructions (or program) for the **central processing unit** and the data are held. There are two types, **random access memory** (**RAM**), where data can be stored (written) and retrieved (read), and **read only memory** (**ROM**), from which "prerecorded" data can only be read.*

***Potential difference**, 58; **Relay**, 75; **Transistor**, 65.

PROPERTIES OF SUBSTANCES

Substance	Density† / 10^3kg m^{-3}	Young's modulus / 10^{10}N m^{-2}	Specific heat capacity† / J kg^{-1} K^{-1}	Specific latent heat of fusion / 10^4J kg^{-1}	Linear co-efficient of expansion / 10^{-6}K^{-1}	Thermal conductivity / W m^{-1} K^{-1}	Resistivity† / 10^{-8} ρ m
Aluminum	*2.70*	*7.0*	*908*	*40.0*	*25*	*242*	*2.67*
Antimony	*6.62*	*7.8*	*210*	*16.5*	*11*	*19*	*44*
Arsenic	*5.73*	–	*335*	–	*6.0*	–	*33.3*
Bismuth	*9.78*	*3.2*	*112*	*5.5*	*14*	*9*	*117*
Brass	*8.6 (approx)*	*9.0*	*389*	–	*19*	*109*	*8 (approx)*
Cadmium	*8.65*	*5.0*	*230*	*5.5*	*30*	*96*	–
Cobalt	*8.70*	–	*435*	*24.0*	*12*	*93*	*6.4*
Constantan	*8.90*	–	*420*	–	*16*	*23*	*49*
Copper	*8.89*	*11.0*	*385*	*20.0*	*16*	*383*	*1.72*
Gallium	*5.93*	–	*377*	–	*19*	*34*	*17.4*
Germanium	*5.40*	–	*324*	–	*5.7*	*59*	*4.6 x* 10^7
Gold	*19.3*	*8.0*	*128*	*6.7*	*14*	*300*	*2.20*
Iridium	*22.4*	–	*135*	–	*6.5*	*59*	*5.2*
Iron (cast)	*7.60*	*11.0*	*460*	*21.0*	*12*	*71*	*10.3*
Iron (wrought)	*7.85*	*21.0*	*460*	*21.0*	*12*	*71*	*10.3*
Lead	*11.3*	*1.6*	*127*	*2.5*	*29*	*36*	*20.6*
Magnesium	*1.74*	*4.1*	*1,030*	*30.0*	*26*	*154*	*4.24*
Mercury	*13.6*	–	*139*	*1.2*	*12*	*.9*	*95.9*
Molybdenum	*10.1*	–	*301*	–	*5.0*	*142*	*5.7*
Nickel	*8.80*	*21.0*	*456*	*29.0*	*13*	*59*	*6.94*
Palladium	*12.2*	–	*247*	*15.0*	*12*	*74*	*10.7*
Platinum	*21.5*	*17.0*	*135*	*11.5*	*9.0*	*71*	*10.5*
Selenium	*4.79*	–	*324*	*35.0*	*26*	*0.24*	10^{12} *(approx)*
Silicon (amorphous)	*2.35*	*11.3*	*706*	–	*2.5*	*175*	10^{10} *(approx)*
Silver	*10.5*	*7.7*	*234*	*10.5*	*19*	*414*	*1.63*
Steel (mild)	*7.80*	*22.0*	*450*	–	*12*	*46*	*15 (approx)*
Tantalum	*16.6*	*19.0*	*151*	–	*6.5*	*56*	*13.4*
Tellurium	*6.2*	–	*201*	–	*17*	*50*	*1.6 x* 10^5
Tin	*7.3*	*5.3*	*225*	*5.8*	*23*	*63*	*11.4*
Tungsten	*19.3*	*39.0*	*142*	–	*4.3*	*185*	*5.5*
Water	*1.00*	–	*4,200*	*33.4*	*33.4*	*0.2*	–
Zinc	*7.10*	*8.0*	*387*	*10.5*	*11*	*111*	*5.92*

---

† *Density, specific heat capacity and resistivity all change with temperature. Values quoted here are for room temperature, i.e. 18-22°C.*

Useful constan

Quantity

Speed of light i
Charge on electron
Mass of electron
Mass of proton
Mass of neutron
Avogadro's number
Faraday's constant
Gravitational constan
Gas constant

Values of comm

Quantity

Acceleration due to
Density of water
Density of mercury
Ice point (standard
Steam point
Standard atmospheri
Length of Earth day

The electromag

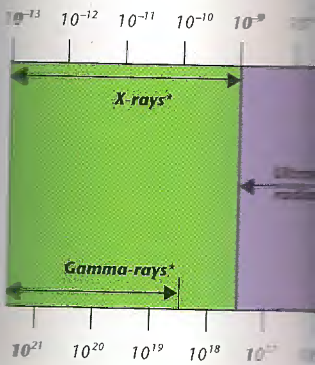

* **Electromagnetic spectrum,** 44;
Microwaves, Radio waves, 45;

Useful constants

Quantity	Symbol	Value
Speed of light in vacuum	c	2.998×10^8m s^{-1}
Charge on electron	e	1.602×10^{-19}C
Mass of electron	m_e	9.109×10^{-31}kg
Mass of proton	m_p	1.673×10^{-27}kg
Mass of neutron	m_n	1.675×10^{-27}kg
Avogadro's number	N_A	6.023×10^{23}mol^{-1}
Faraday's constant	F	9.65×10^4C mol^{-1}
Gravitational constant	G	6.670×10^{-11}N m^2 kg^{-2}
Gas constant	R	8.314J mol^{-1} K^{-1}

Values of common quantities

Quantity	Value
Acceleration due to gravity g (gravitational field strength)	9.81m s^{-2}
Density of water	1.00×10^3kg m^{-3}
Density of mercury	13.6×10^3kg m^{-3}
Ice point (standard temperature)	273K
Steam point	373K
Standard atmospheric pressure	1.01×10^5Pa
Length of Earth day	8.64×10^4s

The electromagnetic spectrum*

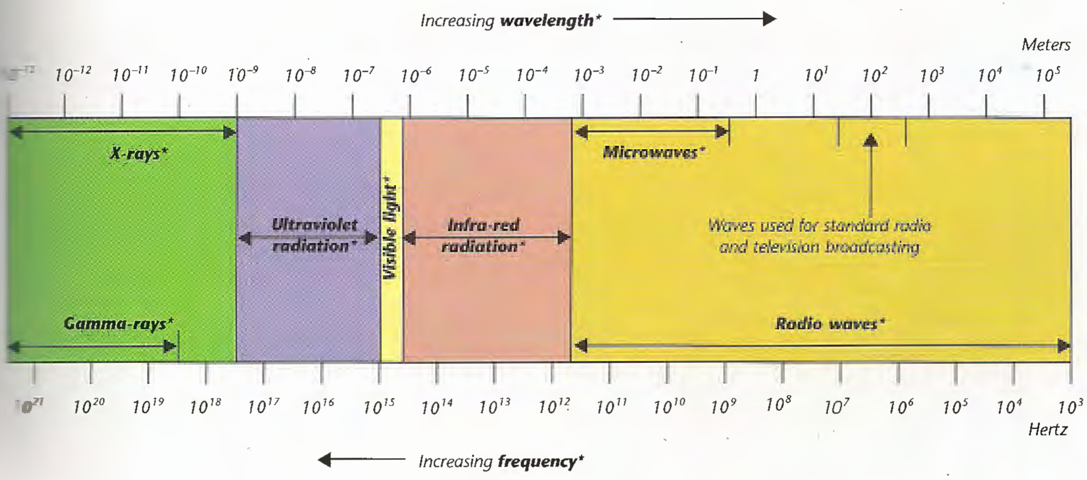
Increasing **wavelength*** →
Meters
10^{-12} 10^{-11} 10^{-10} 10^{-9} 10^{-8} 10^{-7} 10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 1 10^1 10^2 10^3 10^4 10^5
X-rays*
Gamma-rays*
Ultraviolet radiation*
Visible light*
Infra-red radiation*
Microwaves*
Waves used for standard radio and television broadcasting
Radio waves*
10^{21} 10^{20} 10^{19} 10^{18} 10^{17} 10^{16} 10^{15} 10^{14} 10^{13} 10^{12} 10^{11} 10^{10} 10^9 10^8 10^7 10^6 10^5 10^4 10^3
Hertz
← Increasing **frequency***

*__Electromagnetic spectrum__, 44; **Frequency**, 35; **Gamma rays**, 44; **Infra-red radiation**, 45; **Microwaves**, **Radio waves**, 45; **Ultraviolet radiation**, 44; **Visible light**, 45; **Wavelength**, 34; **X-rays**, 44.

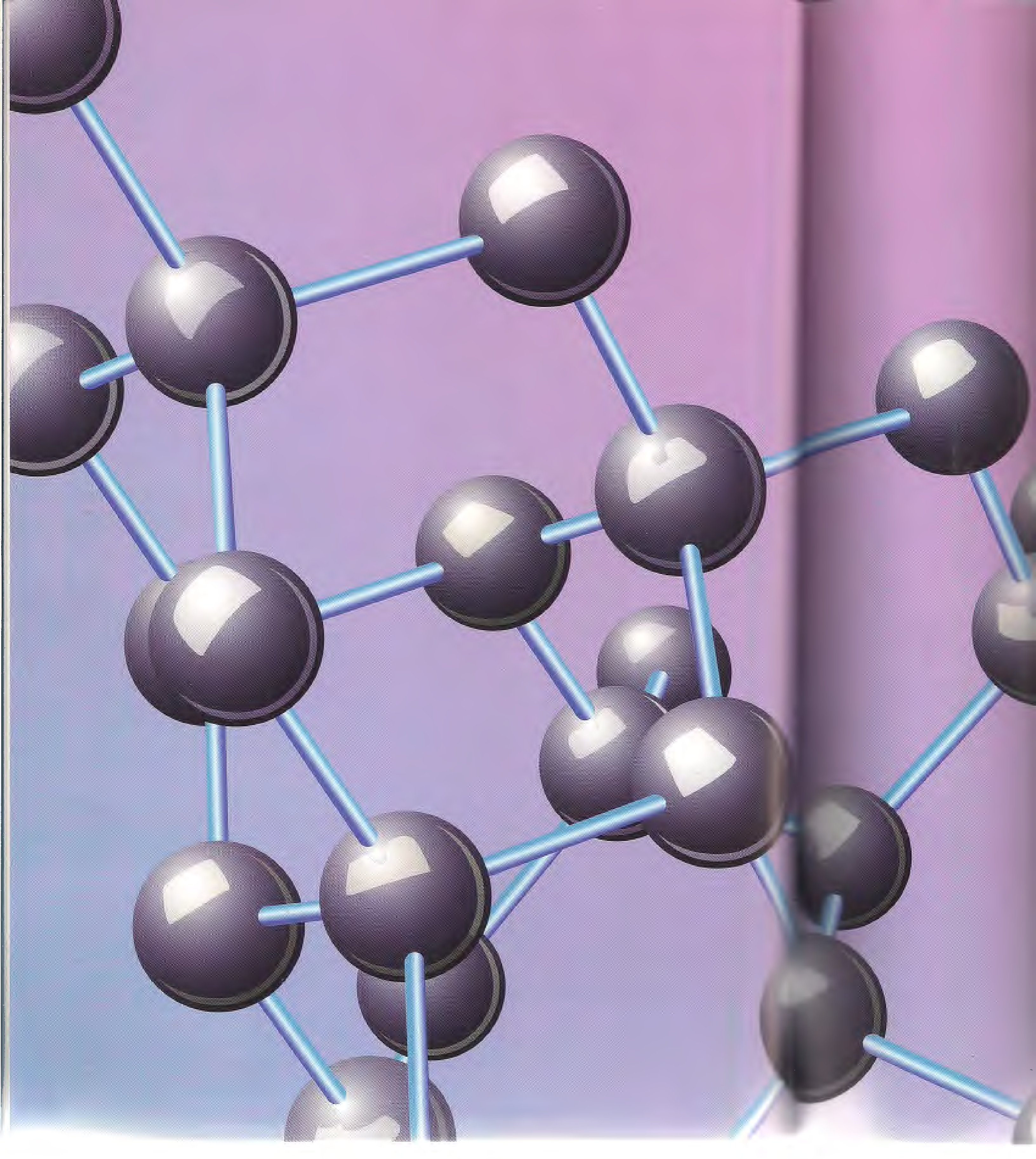

CHEMISTRY

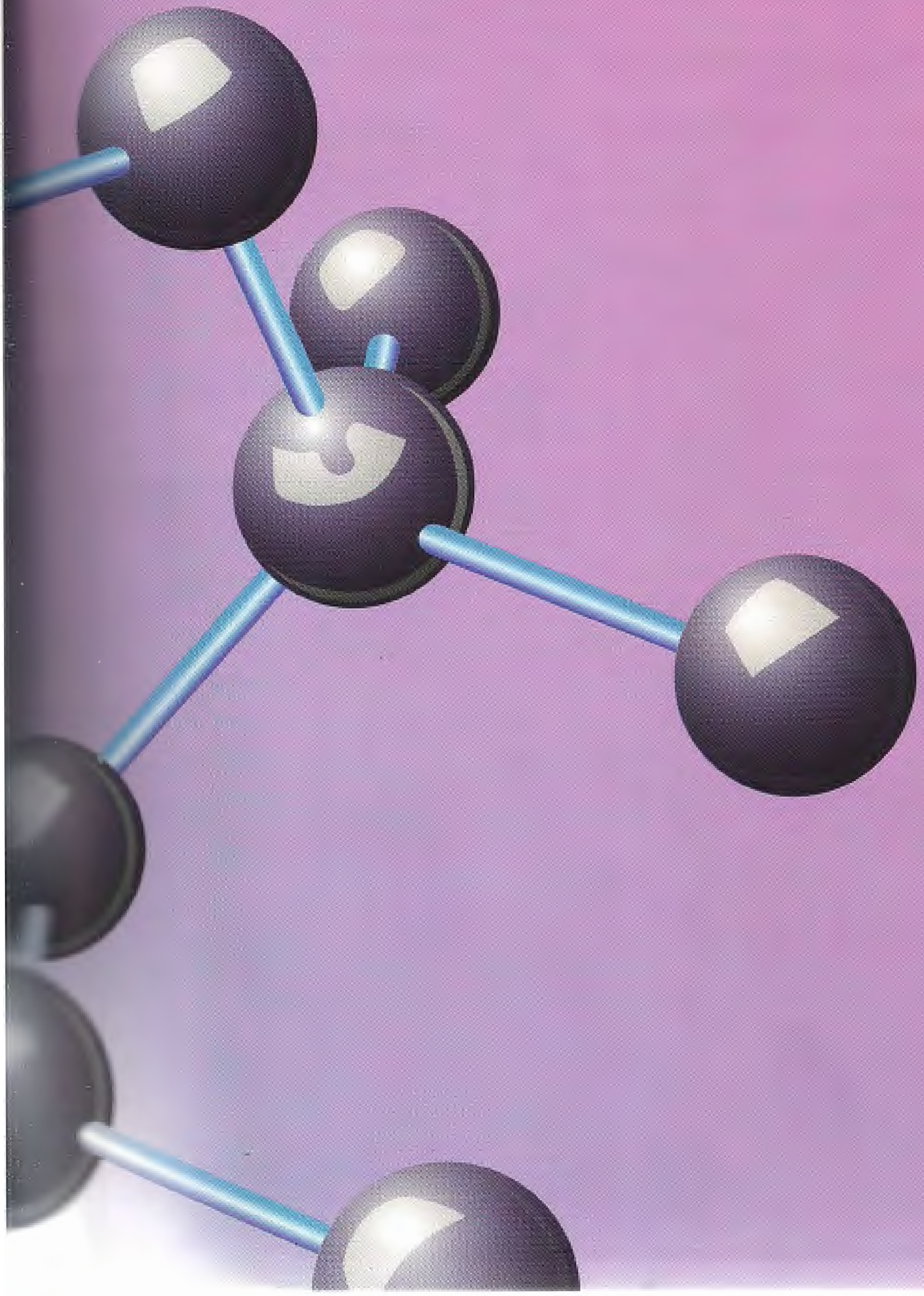

ABOUT CHEMISTRY

Chemistry is the study of the elements which form all existing substances. It covers their structure, how they combine to create other substances and how they react under various conditions. In this book, chemistry is divided into five color-coded sections. The areas covered by these sections are explained below.

Physical chemistry

Covers the structures, properties and behavior of substances. Includes the basic laws of chemistry.

Organic chemistry

Covers the carbon-chain compounds. Examines their structures and the various groups into which they fall.

Inorganic chemistry

Looks at the groups of elements in the periodic table, their properties, uses and compounds (except carbon compounds).

Environmental chemistry

Explains the interaction of naturally-occurring chemicals, and the effect of pollution.

General chemistry information

Charts and tables of properties, symbols and means of identification, plus information on apparatus, preparations, tests and forms of chemical analysis.

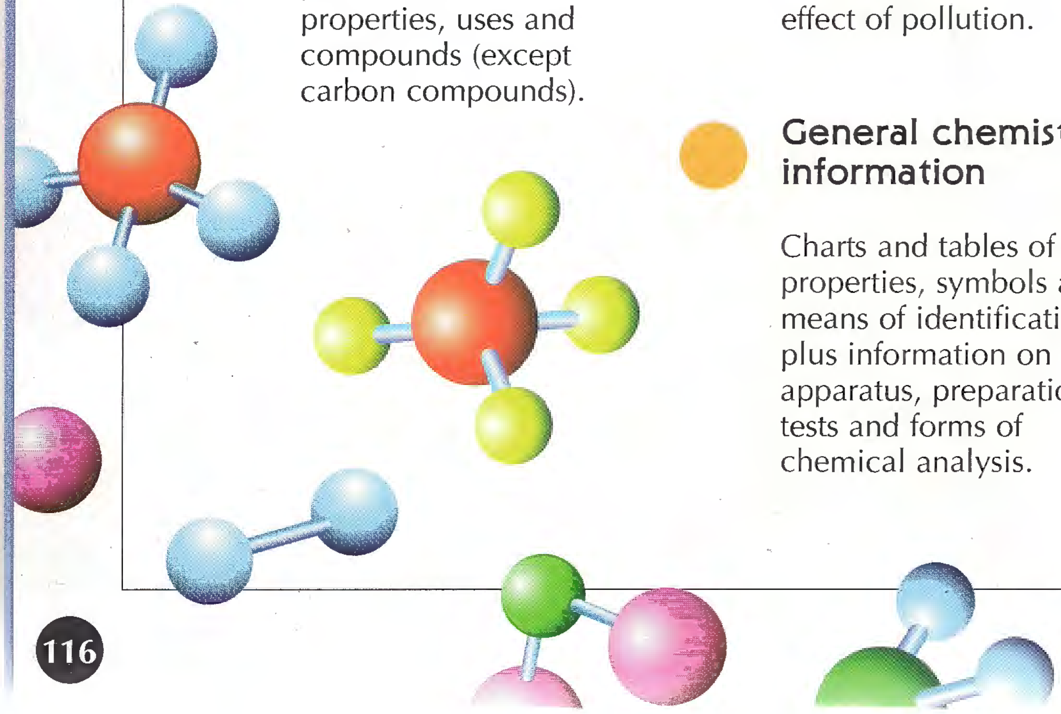

Physical ch

118 Physical c
119 Properties
120 States of n
122 Elements, mixtures
124 Atoms and
126 Atomic str
128 Radioacti
130 Bonding
135 Crystals
138 Measuring
140 Represen
142 Gas laws
144 Solutions
146 Energy an
148 Oxidation
150 Acids and
153 Salts
156 Electrolys
158 Reactivity
160 Rates of re
162 Reversible

Inorganic ch

164 The perio
166 Inorganic
167 Hydrogen
168 Group I,
170 Group II, metals
172 Transition
174 Iron, coppe
176 Group III
177 Group IV
178 Carbon
180 Group V
183 Group VI
184 Sulfur
186 Group VII,
189 Group VII

CONTENTS

Physical chemistry

118 Physical chemistry
119 Properties and changes
120 States of matter
122 Elements, compounds and mixtures
124 Atoms and molecules
126 Atomic structure
128 Radioactivity
130 Bonding
135 Crystals
138 Measuring atoms
140 Representing chemicals
142 Gas laws
144 Solutions and solubility
146 Energy and chemical reactions
148 Oxidation and reduction
150 Acids and bases
153 Salts
156 Electrolysis
158 Reactivity
160 Rates of reaction
162 Reversible reactions

Inorganic chemistry

164 The periodic table
166 Inorganic chemistry
167 Hydrogen
168 Group I, the alkali metals
170 Group II, the alkaline-earth metals
172 Transition metals
174 Iron, copper and zinc
176 Group III elements
177 Group IV elements
178 Carbon
180 Group V elements
183 Group VI elements
184 Sulfur
186 Group VII, the halogens
189 Group VIII, the noble gases

Organic chemistry

190 Organic chemistry
192 Alkanes
193 Alkenes
194 Alkynes
More homologous series
196 Alcohols
198 Petroleum
200 Polymers and plastics
202 Detergents
204 Food

Environmental chemistry

206 Water
208 Air and burning
210 Pollution

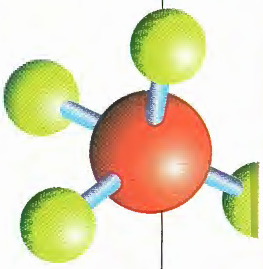

General chemistry information

211 The reactivity series
212 The properties of the elements
214 Naming simple organic compounds
216 The laboratory preparation of six common gases
218 Laboratory tests
220 Investigating substances
222 Qualitative and quantitative analysis
223 Apparatus
226 Chart of substances, symbols and formulas

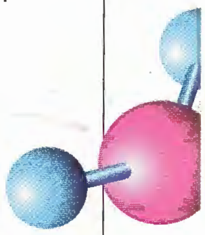

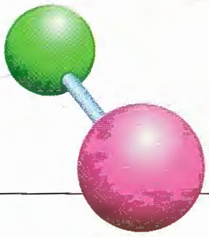

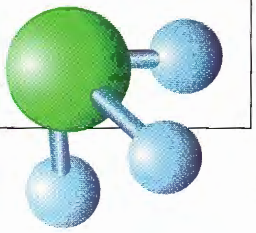

PHYSICAL CHEMISTRY

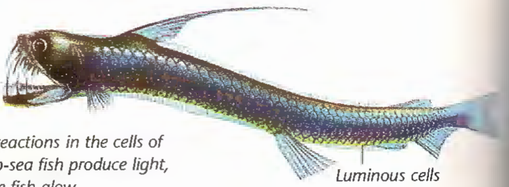
Chemical reactions in the cells of many deep-sea fish produce light, making the fish glow.

Physical chemistry is the study of the patterns of chemical behavior in **chemical reactions** under various conditions, which result from the **chemical** and **physical properties** of substances. Much of physical chemistry involves measurements of some kind. In the physical chemistry section of this book you can find out about the following areas:

1. Solids, liquids and gases, the changes between these states and the reasons for these changes in relation to the structure of a substance. (See **states of matter**, pages 120-121, **kinetic theory**, page 123 and **gas laws**, pages 142-143.)

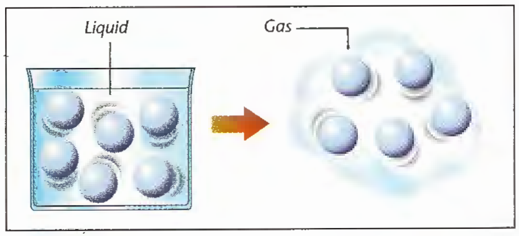

2. The physical and chemical composition of substances – their particles and bonding. (See **elements**, **compounds and mixtures**, pages 122-123, **atoms and molecules**, pages 124-125, **bonding**, pages 130-134 and **crystals**, pages 135-137.)

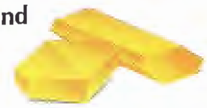
Sulfur crystals occur in two shapes.

3. The structure of the atom and its importance in the structure of substances (see **atomic structure** and **radioactivity**, pages 126-129).

Structure of a sodium atom

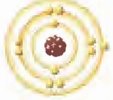

4. The measurement of quantities and the relationship between amounts of gases, liquids and solids (see **measuring atoms**, pages 138-139).

Measuring relative atomic mass

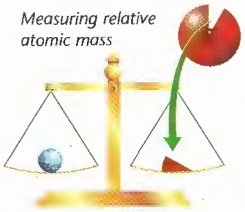

5. Representing chemicals and chemical reactions (see **representing chemicals**, pages 140-141).

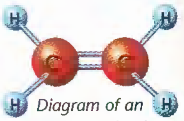
Diagram of an ethene molecule

Both these ways of representing an ethene molecule show that there are two carbon atoms and four hydrogen atoms.

C_2H_4 Molecular formula of ethene

6. How substances mix (see **solutions and solubility**, pages 144-145).

Salt is soluble – it dissolves in water to leave a clear solution.

7. Changes during chemical reactions (see **energy and chemical reactions**, pages 146-147 and **rates of reaction**, pages 160-161) and special reactions (see **oxidation and reduction**, pages 148-149 and **reversible reactions**, pages 162-163).

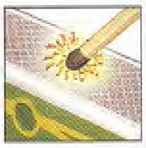
Chemicals on a match burn when they react with phosphorus on the box.

8. Special types of chemical behavior (see **acids and bases**, pages 150-152 and **salts**, pages 153-155).

Hydrochloric acid and sodium hydroxide (an alkaline solution) react together to form sodium chloride, a salt. The solutions have been colored with **litmus*** to show whether they are acidic, alkaline, or neutral.

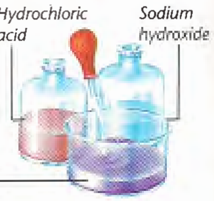

9. The action of electricity on substances and the production of electricity from reactions (see **electrolysis**, pages 156-157 and **reactivity**, pages 158-159).

Electrolysis of copper

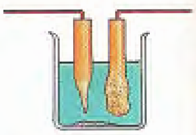

10. The different levels of reactivity shown by substances and the reasons for this (see **reactivity**, pages 158-159).

Stalagmites and stalactites form gradually as a result of a slow chemical reaction between calcium carbonate in limestone and carbonic acid in rainwater.

---

* **Litmus**, 152.

PROPERT

Physical properties.
All the properties of a
those which affect its
reactions. There are
qualitative properties
properties.

Qualitative properti

Descriptive properties
cannot be given a ma
are such things as sm

Some qualitative properties

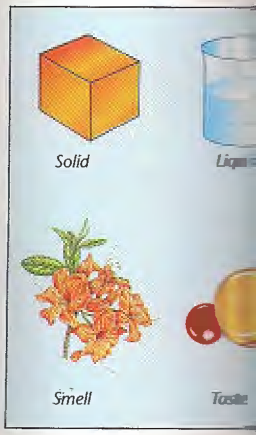

Quantitative prope

Properties which can
a specific mathemati
point, boiling point,
density*. Other exam

Some quantitative properties

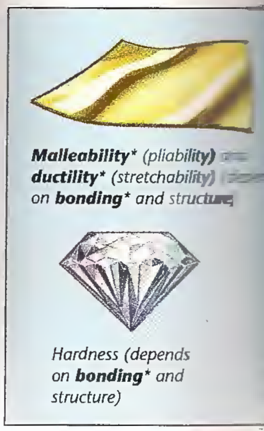
Malleability* (pliability)
ductility* (stretchability)
on **bonding*** and structure

Hardness (depends on **bonding*** and structure)

---

* **Bonding**, 130; **Conductivity**
Malleability, 345 (**Malleable**

PROPERTIES AND CHANGES

Physical properties

All the properties of a substance except those which affect its behavior in **chemical reactions**. There are two main types – **qualitative properties** and **quantitative properties**.

Qualitative properties

Descriptive properties of a substance which cannot be given a mathematical value. They are such things as smell, taste and color.

Some qualitative properties used to describe substances

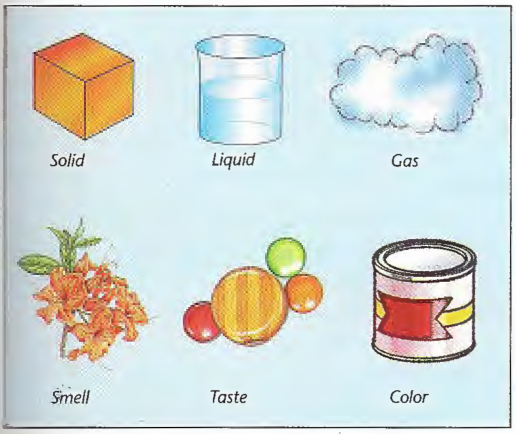

Quantitative properties

Properties which can be measured and given a specific mathematical value, e.g. melting point, boiling point, **mass***, **solubility*** and **density***. Other examples are shown below.

Some quantitative properties used to describe substances

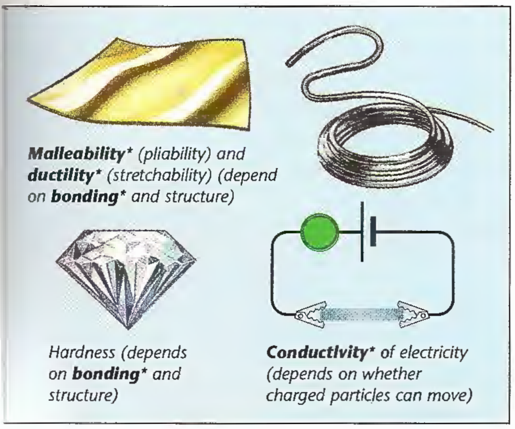

Malleability* *(pliability) and* ***ductility**** *(stretchability) (depend on* ***bonding**** *and structure)*

Hardness (depends on ***bonding**** *and structure)*

Conductivity* *of electricity (depends on whether charged particles can move)*

Physical change

A change which occurs when one or more of the **physical properties** of a substance is changed. It is usually easily reversed.

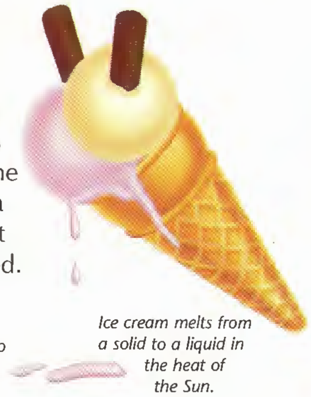

A ***physical change*** *from solid to liquid is caused by adding energy to the particles of the substance (see* ***kinetic theory****, page 123).*

Ice cream melts from a solid to a liquid in the heat of the Sun.

Chemical properties

Properties which cause specific behavior of substances during **chemical reactions**.

Chemical properties *depend on* ***electron configuration****,* ***bonding****, *structure and energy changes.*

Chemical reaction

Any change which alters the **chemical properties** of a substance or which forms a new substance. During a chemical reaction, **products** are formed from **reactants**.

Reactants

The substances present at the beginning of a **chemical reaction**.

Products

The substances formed in a **chemical reaction**.

The ***rusting**** *of iron is a chemical reaction. The reaction is quite slow – many reactions are much faster.*

Iron, water and oxygen from the air are the ***reactants****.*

Rust* *is the* ***product****.*

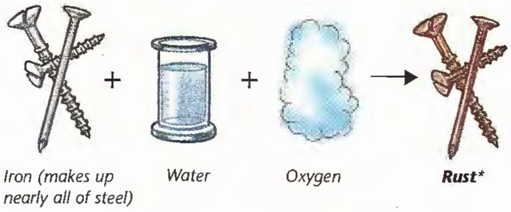
Iron (makes up nearly all of steel)
+
Water
+
Oxygen
Rust*

Reagent

A substance used to start a **chemical reaction**. It is also one of the **reactants**. Common reagents in the laboratory are hydrochloric acid, sulfuric acid and sodium hydroxide.

Bonding, 130; **Conductivity**, 63; **Density**, 24; **Ductility**, 344 (**Ductile**); **Electron configuration**, 127; **Malleability**, 345 (**Malleable**); **Mass**, 12; **Rust**, 174; **Rusting**, 209 (**Corrosion**); **Solubility**, 145.

STATES OF MATTER

A substance can be **solid**, **liquid** or **gaseous**. These are the **physical states** or **states of matter** (normally shortened to **states**). Substances can change between states, normally when heated or cooled to increase or decrease the energy of the particles (see **kinetic theory**, page 123).

Solid state

A state in which a substance has a definite volume and shape.

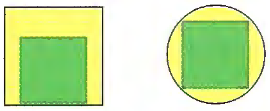

***Solid state** – volume and shape stay the same.*

Liquid state

A state in which a substance has a definite volume, but can change shape.

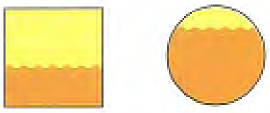

***Liquid state** – volume stays the same, but shape alters.*

Gaseous state

A state in which a substance has no definite volume or shape. It is either a **vapor** or a **gas**. A vapor can be changed into a liquid by applying pressure alone; a gas must first be turned into a vapor by reducing its temperature to below a level called its **critical temperature**.

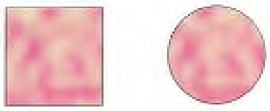

***Gaseous state** – volume and shape will alter.*

Phase

A separate part of a mixture of substances with different physical and chemical properties. A mixture of sand and water contains two phases, as does a mixture of oil and water.

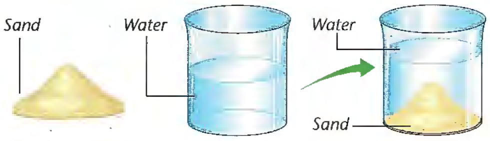

Fluid

A substance that will flow, i.e. is in either the **gaseous** or **liquid state**.

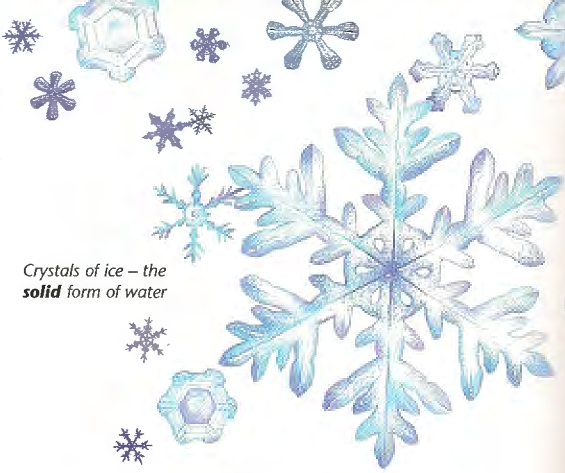

*Crystals of ice – the **solid** form of water*

Changes of state

A **change of state** is a **physical change*** of a substance from one state to another. It normally occurs because of a change in the energy of the particles, caused by heating or cooling (see **kinetic theory**, page 123).

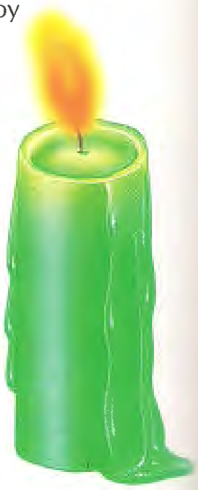

***Solid** wax becomes **molten** when heated.*

Molten

Describes the **liquid** state of a substance which is a **solid** at room temperature.

Solidification

The change of state from **liquid** to **solid** of a substance which is a solid at room temperature and atmospheric pressure.

Melting

The change of state from **solid** to **liquid**, usually caused by heating. The temperature at which a solid melts is called its **melting point** (see also pages 212-213), which is the same temperature as its **freezing point** (see **freezing**). At the melting point, both solid and liquid states are present. An increase in pressure increases the melting point. All pure samples of a substance at the same pressure have the same melting point.

*Ice (**solid** form of water) **melts** at 0°C or 273K. Adding substances such as orange juice to the water may lower its **melting point**.*

---

* **Physical change**, 119.

Freezing

The change of state fr
caused by cooling a li
at which a substance i
point, which is the sam
melting point (see **mel**

*Water **freezes** at 0°C or 273K. Added substances, such as salt in sea water may lower its **freezing point**.*

Fusion

The change of state fro
substance which is sol
and pressure. The subst
fused (or **molten**). A so
and then solidified into
described as fused.

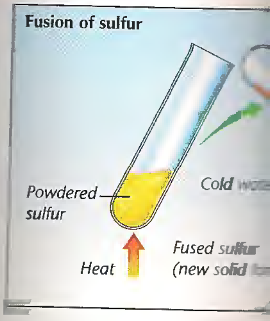

Boiling

A change of state from li
(**vapor**) at a temperature
point (see also pages 21
the formation of bubbles
liquid. All pure samples
liquid at the same pressu
have the same boiling point. An increase in pressure increases the boiling point. A decrease in pressure decreases the boiling point.

Water boils at 100°C or 373K.

Volatile, 345.

Freezing

The change of state from **liquid** to **solid**, caused by cooling a liquid. The temperature at which a substance freezes is the **freezing point**, which is the same temperature as the **melting point** (see **melting**).

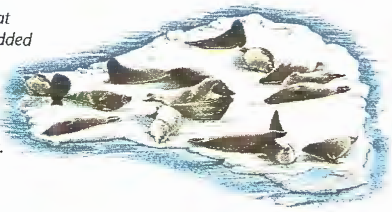

*Water **freezes** at 0°C or 273K. Added substances, such as salt in sea water may lower its **freezing point**.*

Fusion

The change of state from **solid** to **liquid** of a substance which is solid at room temperature and pressure. The substance is described as **fused** (or **molten**). A solid that has been fused and then solidified into a different form is also described as fused.

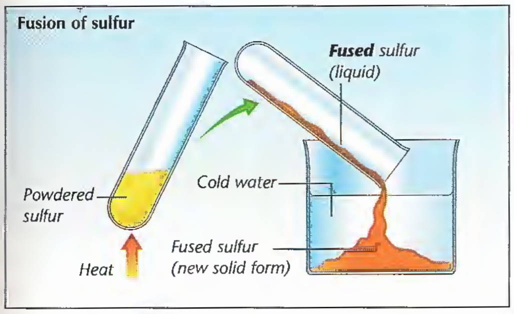

Boiling

A change of state from **liquid** to **gaseous** (**vapor**) at a temperature called the **boiling point** (see also pages 212-213). It occurs by the formation of bubbles throughout the liquid. All pure samples of the same liquid at the same pressure have the same boiling point. An increase in pressure increases the boiling point. A decrease in pressure decreases the boiling point.

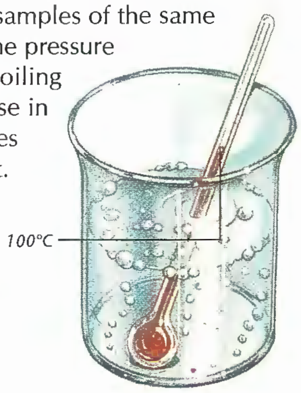

Water boils at 100°C or 373K.

Evaporation

A change of state from **liquid** to **gaseous** (**vapor**), due to the escape of molecules from the surface. (For more about this, see page 30.) A liquid which readily evaporates is described as **volatile***.

In rainforests where it is warm, raindrops quickly evaporate to form water vapor in the air.

Liquefaction

A change of state from **gaseous** (**gas**) to **liquid**, of a substance which is a gas at room temperature and pressure. It is caused by cooling (to form a **vapor**) and increasing pressure.

*Some **gases** are **liquefied** for transport.*

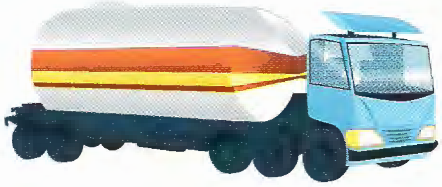

Condensation

A change of state from **gaseous** (**gas** or **vapor**) to **liquid**, of a substance which is a liquid at room temperature and pressure. It is normally caused by cooling.

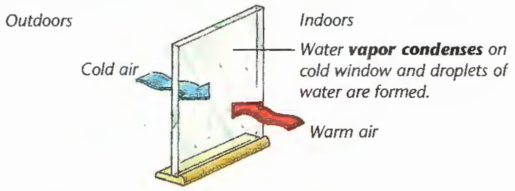
Outdoors
Indoors
Cold air
*Water **vapor condenses** on cold window and droplets of water are formed.*
Warm air

Sublimation

The change of state from **solid** to **gaseous** (**gas**, via **vapor**) on heating, and from gaseous directly to solid on cooling. At no stage is a liquid formed. See picture, page 162.

Vaporization

Any change resulting in a **gaseous state**, i.e. **boiling**, **evaporation** or **sublimation**.

---

* **Volatile**, 345.

ELEMENTS, COMPOUNDS AND MIXTURES

Elements, **compounds** and **mixtures** are the three main types of chemical substance. Most natural substances are made up of several compounds.

Element

A substance which cannot be split into a simpler substance by a chemical reaction. There are just over 100 known elements, classified in the **periodic table***, and most are solids or gases at room temperature. All atoms of the same element have the same number of **protons*** in their **nuclei*** (see **atomic number**, page 127).

Powdered sulfur
Iron filings

*Iron and sulfur are **elements** – they cannot be broken down into simpler substances.*

Compound

A combination of two or more **elements**, bonded together in some way. It has different physical and chemical properties from the elements it is made of. The proportion of each element in a compound is constant, e.g. water is always formed from two parts hydrogen and one part oxygen. This is shown by its chemical **formula***, H_2O. Compounds are often difficult to split into their elements and can only be separated by chemical reactions or **electrolysis***, a process in which an electric current is used to cause a chemical change.

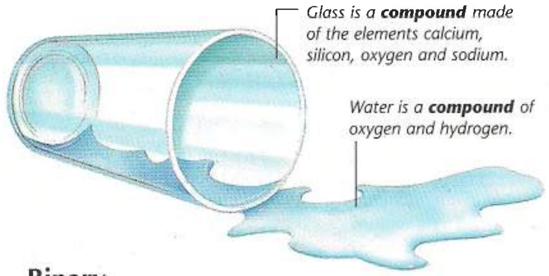

*Glass is a **compound** made of the elements calcium, silicon, oxygen and sodium.*

*Water is a **compound** of oxygen and hydrogen.*

Binary

Describes a **compound** composed of two **elements** only, e.g. carbon monoxide, which contains only carbon and oxygen.

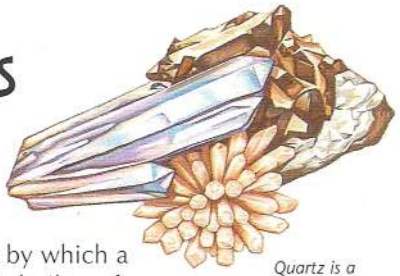

*Quartz is a **compound** of silicon and oxygen. The temperature and pressure at which **synthesis** takes place affects the structure of the mineral formed.*

Synthesis

The process by which a **compound** is built up from its **elements** or from simpler compounds by a sequence of chemical reactions, e.g. iron(III) chloride is made by passing chlorine gas over heated iron.

Mixture

A blend of two or more **elements** and/or **compounds** which are not chemically combined. The proportions of each element or compound are not fixed, and each keeps its own properties. A mixture can usually be separated into its elements or compounds fairly easily by physical means.

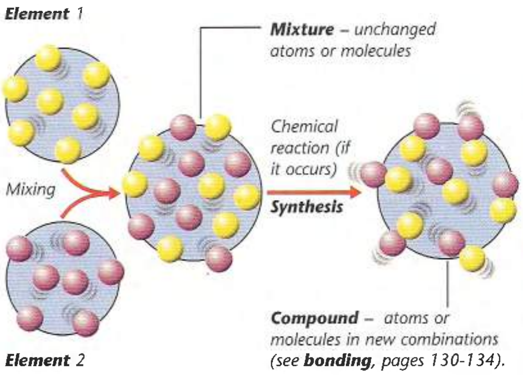

***Mixture** – unchanged atoms or molecules*

***Compound** – atoms or molecules in new combinations (see **bonding**, pages 130-134).*

Chemical symbol

A shorthand way of representing an **element** in **formulas** and **equations** (see pages 140-141). It represents one atom and usually consists of the first one or two letters of the name of the element, occasionally the Greek or Latin name. See pages 212-213 for a list of elements and their symbols, and pages 226-227 to match symbols to elements.

***Chemical symbol** S*

***Chemical symbol** Fe – ferrum is Latin for iron..*

* **Electrolysis**, 156; **Formulas**, 140; **Nucleus**, 126; **Periodic table**, 164; **Proton**, 126.

Homogeneous

Describes a substance where all the particles are in the same **phase*** e.g. **solutions*** (the physical and chemical properties throughout are the same).

Heterogeneous

Describes a substance where the particles are in more than one phas e.g. **suspensions*** (the properties of the solid particles are different from those of the liquid

Kinetic theor

The **kinetic theory** ex solids, liquids and ga **state*** between them, movement of the part are made (see diagram

*According to the **kinetic theor***

Particles in solids are closely packed together. They vibrate, but do not move about.

Heat gives particles enough energy to break the bonds that keep them together.

Particles in a liquid are quite close together, but are free to move about.

Eventually, heat gives particles enough energy to escape from the surface of a liquid to form a gas.

A gas consists of widely-spaced particles moving at high speeds.

The greater the speed and frequency with which molecules of a gas hit surfaces or each other, the higher its pressure.

* **Change of state**, **Fluid**, 12 **Phase**, 120; **Solution**, 14

Homogeneous

Describes a substance where all the particles are in the same **phase***, e.g. **solutions*** (the physical and chemical properties throughout are the same).

Homogeneous

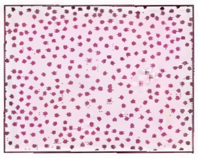

All particles are in the same ***phase****.*

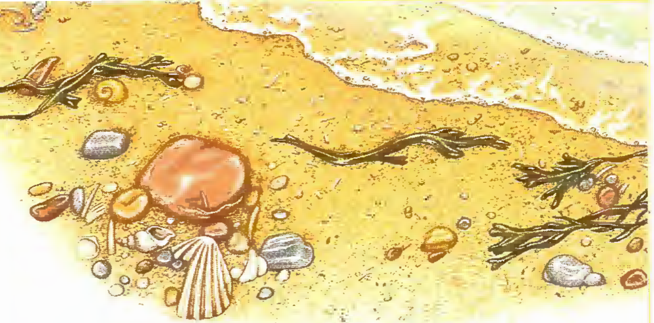

The sand on the seashore is a ***heterogeneous*** *mixture of tiny particles of quartz, seashell and organic matter.*

Heterogeneous

Describes a substance where the particles are in more than one **phase***, e.g. **suspensions*** (the properties of the solid particles are different from those of the liquid).

Heterogeneous

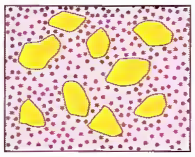

Particles are in different ***phases****.*

Pure

Describes a sample of a substance which consists only of one **element** or **compound**. It does not contain any other substance in any proportions. If the substance does contain traces of another element or compound, then it is described as **impure** and the other substance is called an **impurity**.

Kinetic theory

The **kinetic theory** explains the behavior of solids, liquids and gases, and **changes of state*** between them, in terms of the movement of the particles of which they are made (see diagram below).

According to the ***kinetic theory****:*

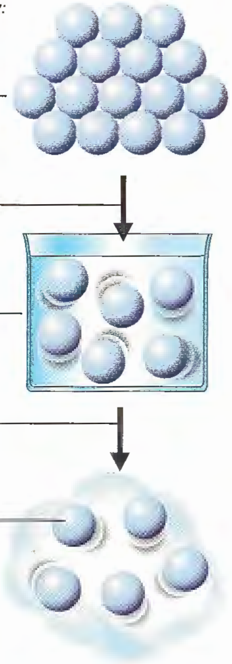

Particles in solids are closely packed together. They vibrate, but do not move about.

Heat gives particles enough energy to break the bonds that keep them together.

Particles in a liquid are quite close together, but are free to move about.

Eventually, heat gives particles enough energy to escape from the surface of a liquid to form a gas.

A gas consists of widely-spaced particles moving at high speeds.

The greater the speed and frequency with which molecules of a gas hit surfaces or each other, the higher its pressure.

Brownian motion

The random motion of small particles in water or air. It supports the kinetic theory, as it is clearly due to unseen impact with the water or air molecules.

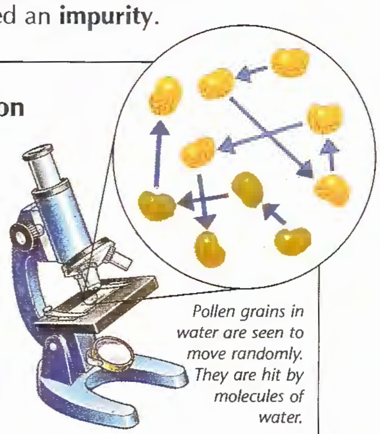

Pollen grains in water are seen to move randomly. They are hit by molecules of water.

Diffusion

The process by which two **fluids*** mix without mechanical help. The process supports the kinetic theory, since the particles must be moving to mix, and visible gases, such as bromine vapour (below), can be seen to diffuse faster than liquids. Only **miscible*** liquids diffuse.

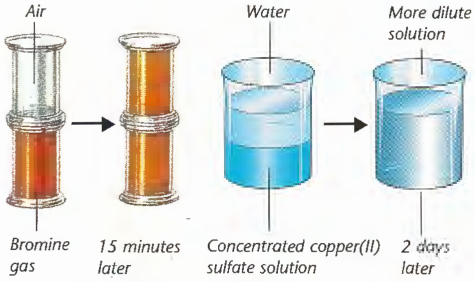

Change of state, **Fluid**, 120; **Miscible**, 145;
Phase, 120; **Solution**, 144; **Suspension**, 145.

ATOMS AND MOLECULES

Over 2,000 years ago, the Greeks decided that all substances consisted of small particles which they called **atoms**. Later theories extended this idea to include **molecules** – atoms joined together. **Inorganic*** molecules generally only contain a few atoms, but **organic*** molecules can contain hundreds of atoms.

Atom

The smallest particle of an element that retains the chemical properties of that element. The atoms of many elements are bonded together in groups to form particles called **molecules** (see also **covalent bonding**, page 132). Atoms consist of three main types of smaller particles – see **atomic structure**, page 126.

Molecule

The smallest particle of an element or compound that normally exists on its own and still retains its properties. Molecules normally consist of two or more **atoms** bonded together – some have thousands of atoms. **Ionic compounds*** consist of **ions*** (electrically charged particles) and do not have molecules.

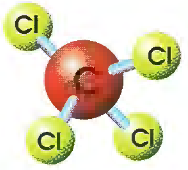

*Tetrachloromethane (CCl_4) **molecules** consist of one carbon and four chlorine **atoms**.*

*Neon **molecules** consist of a single neon **atom**.*

Atomicity

The number of **atoms** in a **molecule**, calculated from the **molecular formula*** of the compound.

*A **molecule** with an **atomicity** of one is described as **monatomic**.*

Helium

*A **molecule** with an **atomicity** of two is described as **diatomic**.*

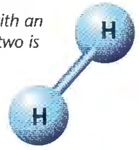

Hydrogen

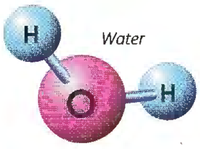

Water

*A **molecule** with an **atomicity** of three is described as **triatomic**.*

*A **molecule** with an **atomicity** of over three is described as **polyatomic**.*

Dalton's atomic theory

John Dalton's theory, published in 1808, attempts to explain how **atoms** behave. It is still generally valid. It states that:

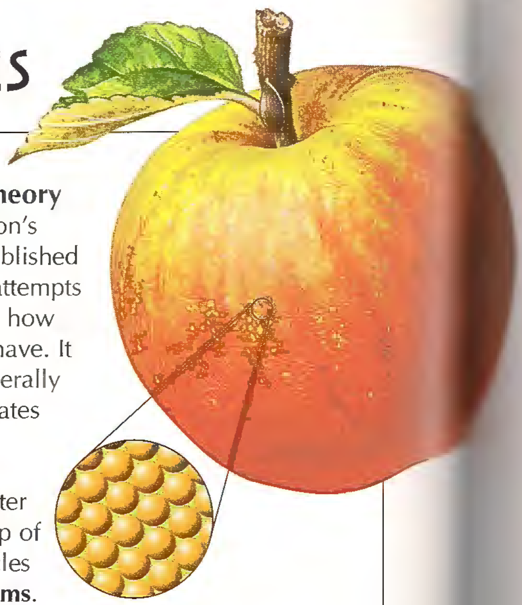

1. All matter is made up of tiny particles called **atoms**.

2. Atoms cannot be made, destroyed or divided. (This has since been disproved – see **radioactivity**, page 128.)

3. All atoms of the same element have the same properties and the same mass. (This has since been disproved – see **isotope**, page 127.)

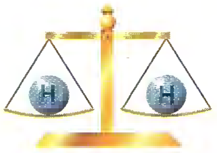

4. **Atoms** of different elements have different properties and different masses.

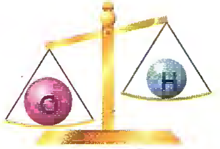

5. When compounds form, the **atoms** of the elements involved combine in simple whole numbers. (We now know, however, that large **organic*** molecules do not always combine in whole number ratios.)

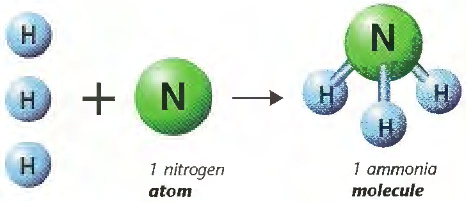

3 hydrogen **atoms**

1 nitrogen **atom**

1 ammonia **molecule**

---

* **Inorganic chemistry**, 166; **Ion**, 130; **Ionic compound**, 131; **Molecular formula**, 140; **Organic chemistry**, 190.

Dimer

A substance with **mole**
combination of two m
(a relatively small mol

*Nitrogen dioxide (**monomer***) combines to form dinitrogen tetraoxide (**dimer**).*

NO_2
Nitrog
dioxid

Basic laws of

Three laws of chemi
in the late eighteen
centuries. Two pre-
theory and the third
proportions) was de
laws were of great im
development of the

Law of constant co

States that all pure sa
chemical compound
elements combined i
by mass. It was devel
Joseph Proust, in 179

*All **molecules** of methane (see right) contain four hydrogen **atoms** (**relative atomic mass*** 1) and one carbon atom (relative atomic mass 12) – see below.*

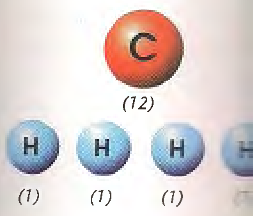

*All pure samples of a substance contain a whole number of **molecules** (i.e. parts of molecules do not exist in compounds).*

So all samples of methane contain carbon and hydrogen in the ratio 3:1 by mass.

Monomers, 200; **Organic** che
Relative atomic mass, **Relative**

Dimer

A substance with **molecules** formed from the combination of two molecules of a **monomer*** (a relatively small molecule).

*Nitrogen dioxide **(monomer*)** combines to form dinitrogen tetraoxide **(dimer)**.*

$NO_2(g)$	+	$NO_2(g)$	$\rightarrow$	$N_2O_4(g)$
Nitrogen dioxide		*Nitrogen dioxide*		*Dinitrogen tetraoxide*

Trimer

A substance with **molecules** formed from the combination of three molecules of a **monomer***.

Macromolecule

A **molecule** consisting of a large number of **atoms**. It is normally an **organic*** molecule with a very high **relative molecular mass***.

Basic laws of chemistry

Three laws of chemistry were put forward in the late eighteenth and early nineteenth centuries. Two pre-date **Dalton's atomic theory** and the third (the **law of multiple proportions**) was developed from it. These laws were of great importance in the development of the atomic theory.

Law of constant composition

States that all pure samples of the same chemical compound contain the same elements combined in the same proportions by mass. It was developed by a Frenchman, Joseph Proust, in 1799.

*All **molecules** of methane (see right) contain four hydrogen **atoms (relative atomic mass*** 1) and one carbon atom (relative atomic mass 12) – see below.*

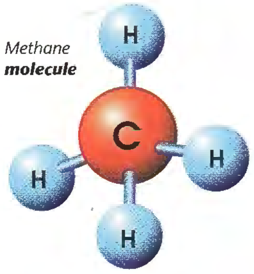

*Methane **molecule***

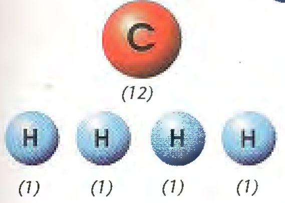

Proportion of carbon to hydrogen by mass = 12:4 = 3:1

*All pure samples of a substance contain a whole number of **molecules** (i.e. parts of molecules do not exist in compounds).*

So all samples of methane contain carbon and hydrogen in the ratio 3:1 by mass.

Law of conservation of mass

States that matter can neither be created nor destroyed during a chemical reaction. It was developed by a Frenchman, Antoine Lavoisier, in 1774.

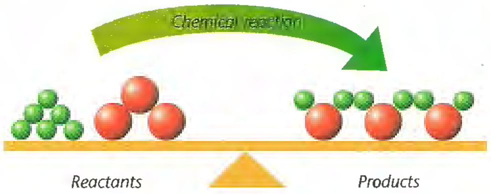

Law of multiple proportions

States that if two elements, A and B, can combine to form more than one compound, then the different masses of A which combine with a fixed mass of B in each compound are in a simple ratio. It is an extension of **Dalton's atomic theory**.

Example for one nitrogen atom:

Nitrogen dioxide, NO_2

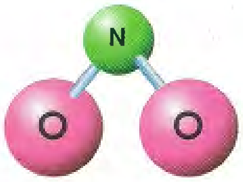

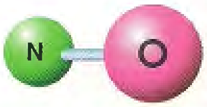

Nitrogen monoxide, NO

Dinitrogen oxide, N_2O

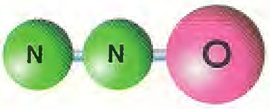

*Number of **atoms** of oxygen per atom of nitrogen are 2, 1 and ½ respectively.*

Masses of oxygen in ratio 4:2:1

---

* **Monomers**, 200; **Organic chemistry**, 190; **Relative atomic mass**, **Relative molecular mass**, 138.

ATOMIC STRUCTURE

Dalton's atomic theory (see page 124) states that the atom is the smallest possible particle. However, experiments have proved that it contains smaller particles, or **subatomic particles**. The three main subatomic particles are **protons** and **neutrons**, which make up the **nucleus**, and **electrons**, which are arranged around the nucleus.

Nucleus (pl. nuclei) or atomic nucleus

The structure at the center of an atom, consisting of protons and neutrons (usually about the same number of each) packed closely together, around which **electrons** move. The nucleus makes up almost the total mass of the atom, but is very small in relation to the total size.

Electron

A **subatomic particle** (see introduction) which moves around the **nucleus** of an atom within an **electron shell**. Its mass is very small, only $^{1}/_{1,836}$ that of a **proton**. An electron has a negative electrical charge, equal in size, but opposite to that of a proton. There are the same number of electrons as protons in an atom.

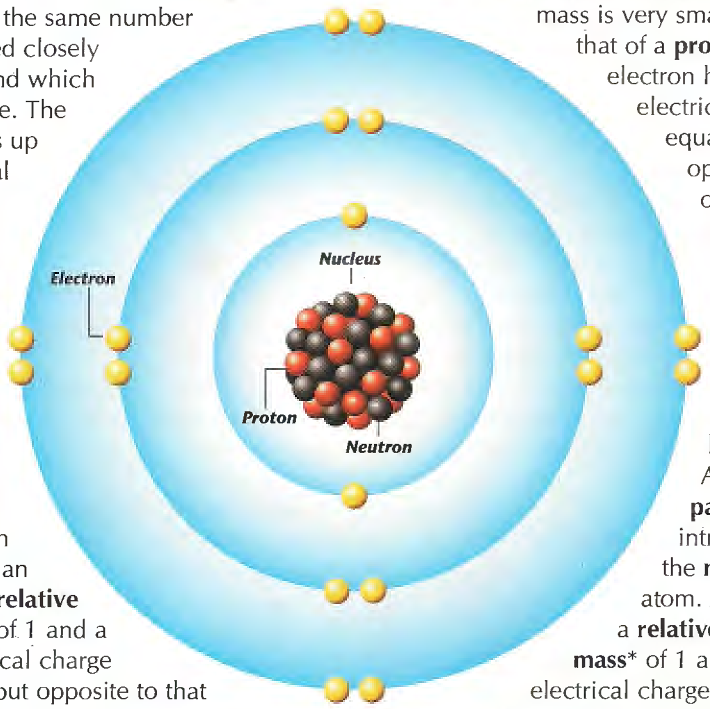

Proton

A **subatomic particle** (see introduction) in the nucleus of an atom. It has a **relative atomic mass*** of 1 and a positive electrical charge equal in size, but opposite to that of an **electron**. An atom has the same number of protons and electrons, making it electrically neutral.

Neutron

A **subatomic particle** (see introduction) in the **nucleus** of an atom. A neutron has a **relative atomic mass*** of 1 and no electrical charge. The number of neutrons in atoms of the same element can vary (see **isotope**, opposite).

Electron shell or shell

A region of space in which the **electrons** move around the **nucleus** of an atom. An atom can have up to seven shells, increasing in radius with distance from the nucleus, and each can hold up to a certain number of electrons. The model on the right is a simplified one – in fact, the exact positions of electrons cannot be determined at any one time, and each shell consists of **orbitals**.

The first three electron shells

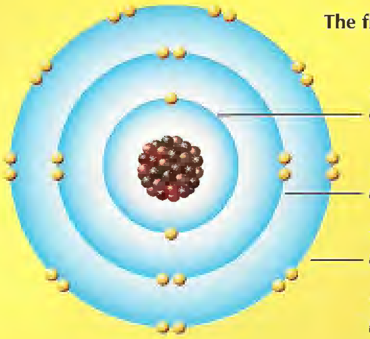

The first shell can hold one or two ***electrons.***

The second shell can hold up to eight electrons.

The third shell can hold up to 18 electrons.

The fourth shell can hold up to 32 electrons.

---

* **Relative atomic mass**, 138.

Orbital

A region in which there
two **electrons**. Each **elec**
one or more orbitals of

Outer shell

The last **electron shell** in
electrons. The number o
outer shell influences ho
and which **group** it is in
pages 164-165).

Electron configuratio

A group of numbers whi
arrangement of the **electr**
numbers are the number
electron shell, starting w

A sodium atom – electron con

First shell *contains two* ***electrons.***

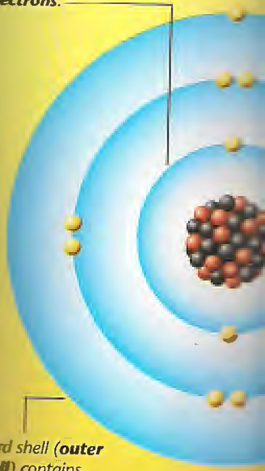

*…ird shell (****outer shell****) contains …e electron.*

Octet

A group of eight **electro**
shell. Atoms with an octet
are very stable and unreac
gases* (except helium) ha
atoms can achieve a stab
have an electron configu
of the nearest noble gas
electrons with other atom
bonding, page 132) or by
electrons (see **ionic bondin**

---

* **Noble gases**, 189.

Orbital

A region in which there can be either one or two **electrons**. Each **electron shell** consists of one or more orbitals of varying shapes.

Outer shell

The last **electron shell** in which there are **electrons**. The number of electrons in the outer shell influences how the element reacts and which **group** it is in (see **periodic table**, pages 164-165).

Electron configuration

A group of numbers which shows the arrangement of the **electrons** in an atom. The numbers are the numbers of electrons in each **electron shell**, starting with the innermost.

A sodium atom – electron configuration 2.8.1

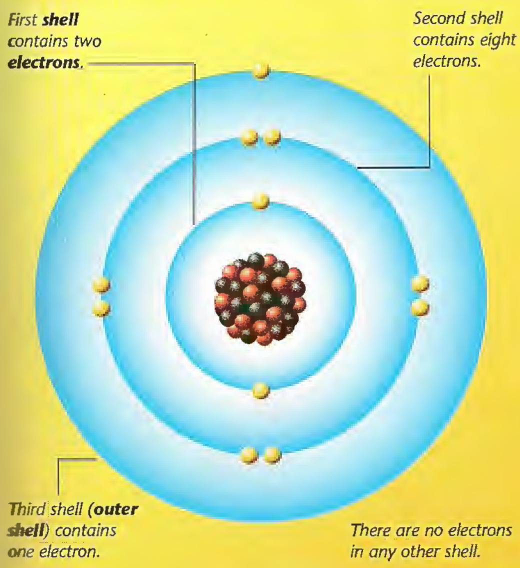
First **shell** contains two **electrons**.
Second shell contains eight electrons.
Third shell (**outer shell**) contains one electron.
There are no electrons in any other shell.

Octet

A group of eight **electrons** in a single **electron shell**. Atoms with an octet for the **outer shell** are very stable and unreactive. All **noble gases*** (except helium) have an octet. Other atoms can achieve a stable octet (and thus have an electron configuration similar to that of the nearest noble gas), either by sharing electrons with other atoms (see **covalent bonding**, page 132) or by gaining or losing electrons (see **ionic bonding**, page 131).

Atomic number

The number of **protons** in the **nucleus** of an atom. The atomic number determines what the element is, e.g. any atom with six protons is carbon, regardless of the number of **neutrons** and **electrons**.

Mass number

The total number of **protons** and **neutrons** in one atom of an element. The mass number of an element can vary because the number of neutrons can change (see **isotope**, below). The mass number is usually about twice the **atomic number**.

*The **atomic number** and **mass number** are often written with the symbol for the element.*

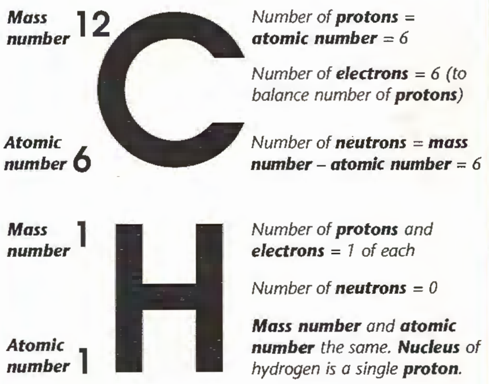
Mass number 12
C
Atomic number 6
Number of **protons** = **atomic number** = 6
Number of **electrons** = 6 (to balance number of **protons**)
Number of **neutrons** = **mass number** – **atomic number** = 6
Mass number 1
H
Atomic number 1
Number of **protons** and **electrons** = 1 of each
Number of **neutrons** = 0
Mass number and **atomic number** the same. **Nucleus** of hydrogen is a single **proton**.

Isotope

An atom of an element in which the number of **neutrons** is different from that in another atom of the same element. Isotopes of an element have the same **atomic number** but different **mass numbers**. Isotopes are distinguished by writing the mass number by the name or symbol of the element.

The three isotopes of carbon

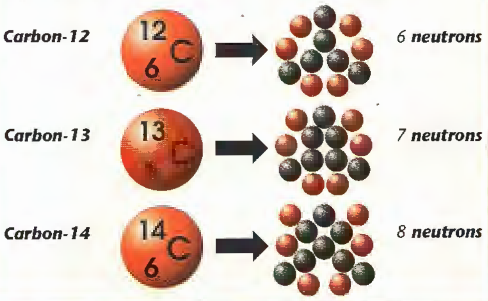
Carbon-12
12 6 C
6 neutrons
Carbon-13
13 6 C
7 neutrons
Carbon-14
14 6 C
8 neutrons

---

* **Noble gases**, 189.

RADIOACTIVITY

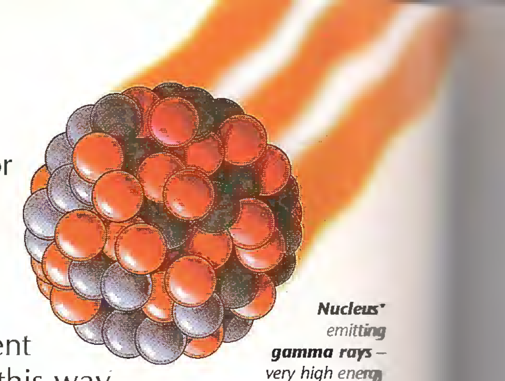
Nucleus* emitting **gamma rays** – very high energ **electromagnetic*** rays.

Radioactivity is a property of unstable **nuclei*** (for more about the reasons for their instability, see pages 84-85). It involves the nuclei breaking up spontaneously into nuclei of other elements and emitting rays or particles (**radiation**), a process known as **radioactive decay**. A radioactive element is one whose nuclei are gradually splitting up in this way.

Radioisotope or radioactive isotope

The general term for a radioactive substance, since all are **isotopes***. There are several naturally-occurring radioisotopes, such as carbon-14 and uranium-238, others are formed in a variety of ways. For more about this, see page 86.

Uranium with 146 ***neutrons**** *is written:* $^{238}_{92}U$
Mass number* — 238
Atomic number* — 92
It can also be written as U-238 or Uranium-238.

Alpha particle (α-particle)

One type of particle emitted from the **nucleus*** of a radioactive atom. It is like a helium nucleus, consisting of two **protons*** and two **neutrons***, has a **relative atomic mass*** of 4 and a charge of plus 2. It moves slowly and has a low penetrating power.

Beta particle (β-particle)

A fast-moving particle emitted from a radioactive **nucleus***. There are two different types of beta particle (for more about this, see page 86). They can penetrate objects which have a low density and/or thickness, such as paper.

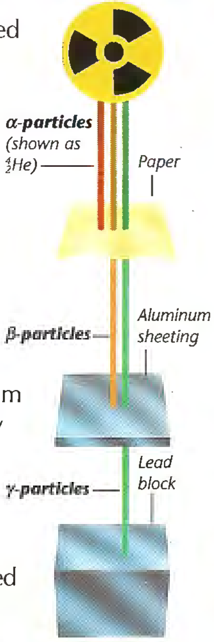
Source of radiation
α-particles (shown as ^4_2He)
Paper
β-particles
Aluminum sheeting
γ-particles
Lead block

Gamma rays (γ-rays)

Rays generally emitted after an **alpha** or **beta particle** from a radioactive **nucleus***. They take the form of waves (like light and X-rays) and have a high penetrating power, going through aluminum sheeting. They can be stopped by a thick block of lead.

Radioactive decay

The process whereby the **nuclei*** of a radioactive element undergo a series of **disintegrations** (a **decay series**) to become stable. For more about the different types of decay, see page 87.

Disintegration

The splitting of an unstable **nucleus*** into two parts, usually another nucleus and an **alpha** or **beta particle**. The **atomic number*** changes, so an atom of a new element is produced. If this is a stable atom, then no further disintegrations occur. If it is unstable, it disintegrates in turn and the process continues as a **decay series** until a stable atom is formed.

Disintegration of uranium-238 to thorium-234

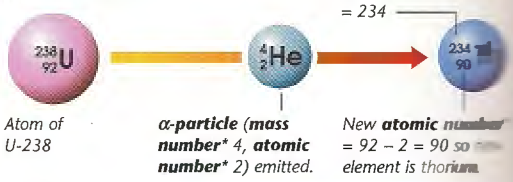
$^{238}_{92}U$
^4_2He
$^{234}_{90}$
New nucleus **mass number*** = 238 – 4 = 234
Atom of U-238
α-particle (**mass number*** 4, **atomic number*** 2) emitted.
New **atomic number** = 92 – 2 = 90 so element is thorium

Disintegration *is shown by a* ***nuclear equation***.

$$^{238}_{92}U \rightarrow ^{234}_{90}Th + ^{4}_{2}He$$

Decay series or radioactive series

The series of **disintegrations** involved when a radioactive element decays, producing various elements until one with stable atoms is formed.

Decay series for plutonium-242 to uranium-234
For another example, see page 87.

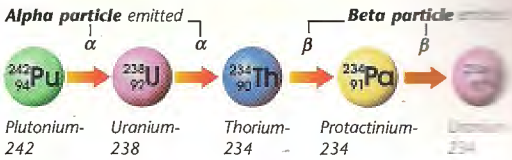
Alpha particle *emitted*
α
α
Beta particle
β
β
$^{242}_{94}Pu$
$^{238}_{92}U$
$^{234}_{90}Th$
$^{234}_{91}Pa$
Plutonium-242
Uranium-238
Thorium-234
Protactinium-234

---

* **Atomic number**, 127; **Electromagnetic waves**, 44; **Isotope**, **Mass number**, 127; **Neutron**, **Nucleus**, **Proton**, 126; **Relative atomic mass**, 138.

Becquerel

A unit of **radioactive de**
is equal to one nuclea
second. A **curie** equals

Half-life

The time taken for half
sample of a radioactive
radioactive decay. The
emitted is halved. The
widely, e.g. the half-life
is 4.5 thousand million
radium-221 is only 30 se

Uses of radioa

Nuclear fission

The division of a **nucle**
bombardment with a **ne**
splits, forming neutrons
elements, and releasing
energy. The release of
the fission of other atom
produces more neutrons
An element which can
described as **fissile**. Con
fission is used in **nuclea**
but uncontrolled fission
bombs*, is very explosi

Fission of uranium-235

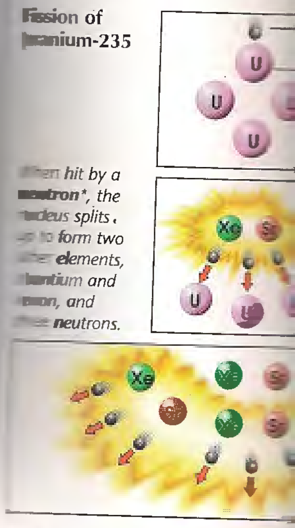

hit by a ***neutron****, *the nucleus splits, form two* ***elements****,* *ntium and on, and* ***neutrons***.

$^{235}_{92}U$ + ^1_0n → $^{90}_{38}Sr$ +

neutron* *(**mass number**** *1, atomic*

Atomic number, 127; **Fission bomb**
Mass number, 127; **Neutron**, 126;
Nuclear power station, 94; **Nucleus**

Becquerel

A unit of **radioactive decay**. One becquerel is equal to one nuclear **disintegration** per second. A **curie** equals 3.7 × 10^{10} becquerels.

Half-life

The time taken for half of the atoms in a sample of a radioactive element to undergo **radioactive decay**. The amount of radiation emitted is halved. The half-life varies widely, e.g. the half-life of uranium-238 is 4.5 thousand million years, but that of radium-221 is only 30 seconds.

Radioactive decay curve

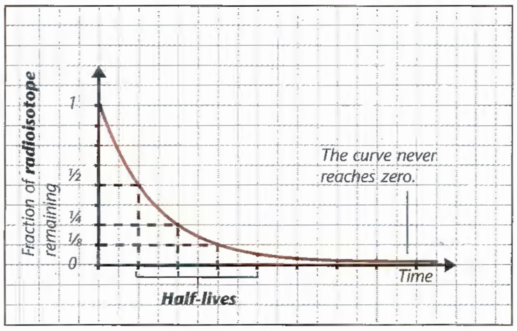

Uses of radioactivity

Nuclear fission

The division of a **nucleus***, caused by bombardment with a **neutron***. The nucleus splits, forming neutrons and nuclei of other elements, and releasing huge amounts of energy. The release of neutrons also causes the fission of other atoms, which in turn produces more neutrons – a **chain reaction**. An element which can undergo fission is described as **fissile**. Controlled nuclear fission is used in **nuclear power stations***, but uncontrolled fission, e.g. in **fission bombs***, is very explosive.

Fission of uranium-235

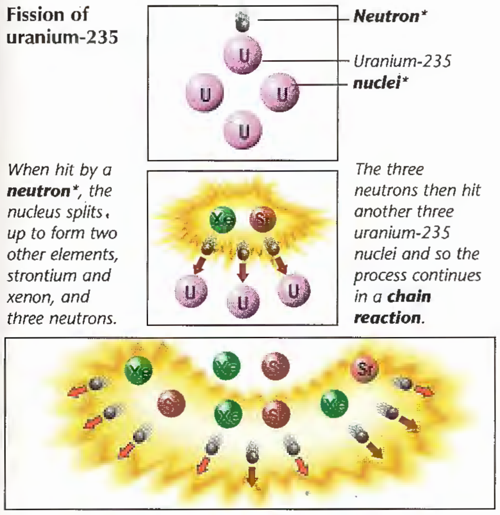

*When hit by a **neutron***, the nucleus splits, up to form two other elements, strontium and xenon, and three neutrons.*

*The three neutrons then hit another three uranium-235 nuclei and so the process continues in a **chain reaction**.*

$${}^{235}_{95}U + {}^{1}_{0}n \rightarrow {}^{90}_{38}Sr + {}^{143}_{54}Xe + 3\,{}^{1}_{0}n$$

*A **neutron*** (**mass number*** 1, **atomic number*** 0)*

Nuclear fusion

The combination of two **nuclei*** to form a larger one. It will only take place at extremely high temperatures and releases huge amounts of energy. Nuclear fusion takes place in the **fusion bomb***.

Radioactive tracing

A method of following a substance as it moves by tracking radiation from a **radioisotope** introduced into it. The radioisotope used is called a **tracer** and the substance is said to be **labeled**.

Radiocarbon dating or carbon dating

A method used to calculate the time elapsed since a living organism died by measuring the radiation it gives off. All living things contain a small amount of carbon-14 (a **radioisotope**) which gradually decreases after death.

Radiology

The study of radioactivity, especially with regard to its use in medicine (**radiotherapy**). Cancer cells are susceptible to radiation, so cancer can be treated by small doses.

Irradiation

The treating of food, such as fruit, with **gamma rays** to keep it fresh.

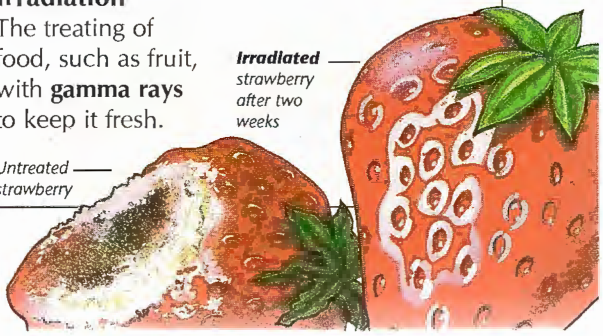
Irradiated strawberry after two weeks
Untreated strawberry

---

* **Atomic number**, 127; **Fission bomb**, **Fusion bomb**, 93; **Mass number**, 127; **Neutron**, 126; **Nuclear power station**, 94; **Nucleus**, 126.

BONDING

When substances react together, the tendency is always for their atoms to gain, lose or share **electrons*** so that they each acquire a stable (full) **outer shell*** of electrons. In doing so, these atoms develop some kind of attraction, or **bonding**, between them (they are held together by **bonds**). The three main types of bonding are **ionic bonding**, **covalent bonding** (see pages 132-133) and **metallic bonding** (see page 134). See also **intermolecular forces**, page 134.

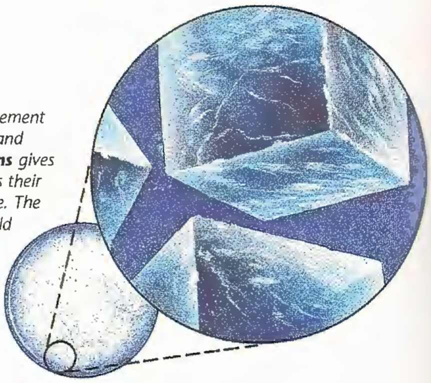

*The arrangement of sodium and chloride **ions** gives salt crystals their cubic shape. The ions are held together by **ionic bonds**.*

Valency electron

An electron, always in the **outer shell*** of an atom, used in forming a bond. It is lost by atoms in **ionic bonding** and **metallic bonding***, but shared with other atoms in **covalent bonding***.

Ions

An **ion** is an electrically charged particle, formed when an atom loses or gains one or more electrons to form a stable **outer shell***. All ions are either **cations** or **anions**.

Cation

An **ion** with a positive charge, formed when an atom loses electrons in a reaction (it now has more **protons*** than electrons). Hydrogen and metals tend to form cations. Their atoms have one, two or three electrons in their **outer shells***, and it is easier for them to lose electrons (leaving a stable shell underneath) than to gain at least five more.

*A magnesium atom has two electrons in its **outer shell***. These are lost to form an **ion** (**cation**) with a charge of +2. A magnesium ion (**cation**) is written Mg^{2+}.*

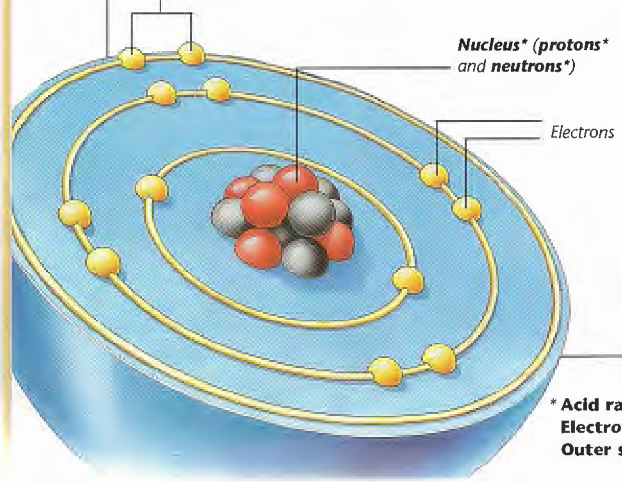

Anion

An **ion** with a negative charge, formed when an atom gains electrons in a reaction (it now has more electrons than **protons***). Non-metals tend to form anions. Their atoms have five, six or seven electrons in their **outer shells***, and it is easier for them to gain electrons (to acquire a stable shell) than to lose at least five. Some anions are formed by groups of atoms gaining electrons, e.g. **acid radicals***.

*A fluorine atom has seven electrons in its **outer shell***, so it gains one to form an **ion** (**anion**) with a charge of –1. A fluoride ion (**anion**) is written F^-.*

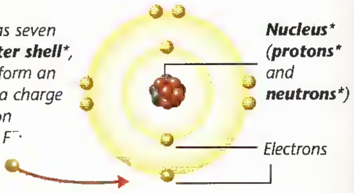

Ionization

The process of forming **ions**. This either happens when atoms lose or gain electrons or when a compound splits up into ions, e.g. hydrogen chloride forming a solution.

***Ionization** of hydrogen chloride in water, forming hydrogen ions and chloride ions.*

$HCl(g)$	$\rightarrow$	$H^+(aq)$	+	$Cl^-(aq)$
Covalent compound* of hydrogen chloride		*Separate ions produced in solution*		

---

* **Acid radical**, 153; **Covalent bonding**, **Covalent compounds**, 132; **Electron**, 126; **Metallic bonding**, 134; **Neutron**, **Nucleus**, 126; **Outer shell**, 127; **Proton**, 126.

Ionic bonding

hen two elements react
ns, the resulting **cations**
e opposite electrical ch
her. They stay together b
action. This type of bon
ic **bonding** and the elec
led **ionic bonds**. Eleme
eriodic **table*** tend to ex
ding, coming together
mpounds, e.g. sodium a
dium chloride) and mag
gen (magnesium oxide

m and chlorine react to form sod
***pound**.*

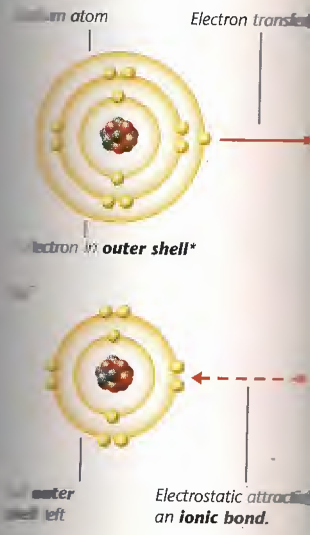

Electrostatic attrac
*an **ionic bond**.*

of sodium chloride is NaCl or

ectrovalency

he power of an **ion** to c
nother in **ionic bonding**
ize of the charge on the
mbine in such proport
harge of the compound

ents in Group 1 and Group 7
ovalent** (have an **electrovale
e a charge of +1 or –1.

ueous solution, 144; **Giant ionic**
ten, 120; **Outer shell**, 127; **Per**

Ionic bonding

When two elements react together to form **ions**, the resulting **cations** and **anions**, which have opposite electrical charges, attract each other. They stay together because of this attraction. This type of bonding is known as **ionic bonding** and the electrostatic bonds are called **ionic bonds**. Elements far apart in the **periodic table*** tend to exhibit this kind of bonding, coming together to form **ionic compounds**, e.g. sodium and chlorine (sodium chloride) and magnesium and oxygen (magnesium oxide).

Ionic compound

A compound whose components are held together by **ionic bonding**. It has no molecules, instead the **cations** and **anions** attract each other to form a **giant ionic lattice***. Ionic compounds have high melting and boiling points (the bonds are strong and hence large amounts of energy are needed to break them). They conduct electricity when **molten*** or in **aqueous solution*** because they contain charged particles (ions) which are free to move.

*Sodium and chlorine react to form sodium chloride, an **ionic compound**.*

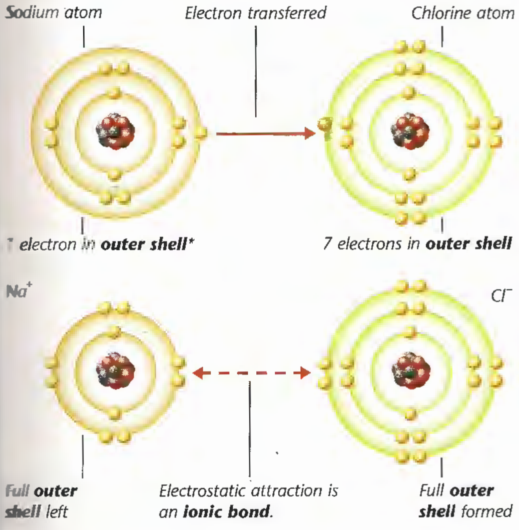
Sodium atom
Electron transferred
Chlorine atom
electron in **outer shell***
7 electrons in **outer shell**
Na⁺
Cl⁻
Full outer shell left
Electrostatic attraction is an **ionic bond**.
Full outer shell formed

Formula of sodium chloride is NaCl or Na^+Cl^-.

Model showing part of the giant ionic lattice of sodium chloride

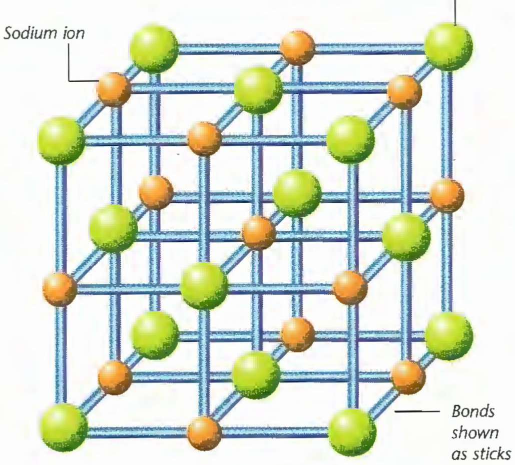
Chloride ion
Sodium ion
Bonds shown as sticks

*Note that there are no molecules – the formula gives the relative numbers of each type of ion in the **giant ionic lattice***. In this case, the formula NaCl indicates that the relative numbers of sodium and chloride ions are one to one.*

Electrovalency

The power of an **ion** to combine with another in **ionic bonding**. It is equal to the size of the charge on the ion. The ions combine in such proportions that the total charge of the compound is zero.

*Group 2 and Group 6 elements are **divalent** (have an **electrovalency** of two). Their ions each have a charge of +2 or –2.*

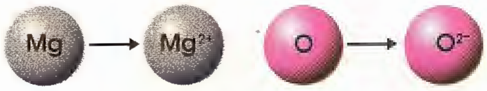

*Elements in Group 1 and Group 7 of the **periodic table*** are **monovalent** (have an **electrovalency** of one). Their ions each have a charge of +1 or –1.*

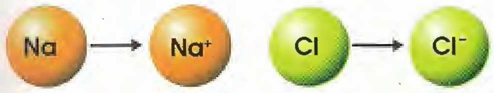

*Some Group 3 and Group 5 elements are **trivalent** (have an **electrovalency** of three). Their ions each have a charge of +3 or –3.*

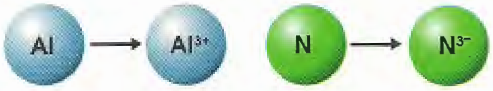

---

* **Aqueous solution**, 144; **Giant ionic lattice**, 137; **Molten**, 120; **Outer shell**, 127; **Periodic table**, 164.

Covalent bonding

Covalent bonding is the sharing of electrons between atoms in a molecule so that each atom acquires a stable **outer shell***. Electrons are shared in pairs called **electron pairs** (one pair being a **covalent bond**). Covalent bonds within a molecule are strong. **Covalent compounds** (compounds whose molecules have internal covalent bonds) are not normally so strongly held together. They are usually liquids or gases at room temperature because the forces between their molecules are **van der Waals' forces***. These weak forces need little energy to overcome them, so most covalent compounds have low melting and boiling points. They do not conduct electricity because there are no **ions*** present.

Single bond

A covalent bond that is formed when one pair of electrons is shared between two atoms.

A chlorine molecule has a **single bond**.

Chlorine atoms each have seven electrons in their **outer shell***.

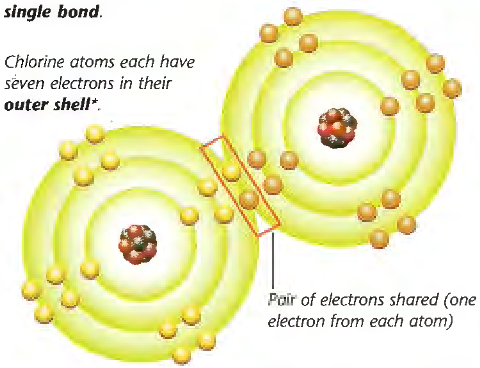

Pair of electrons shared (one electron from each atom)

Each atom now has a stable **outer shell*** of eight electrons (see **octet**, page 127).

A **single bond** is shown in a formula by a single line: Cl–Cl

Triple bond

A covalent bond that is formed when three pairs of electrons are shared between two atoms.

A nitrogen molecule has a **triple bond**.

Nitrogen atoms each have five electrons in their **outer shell***.

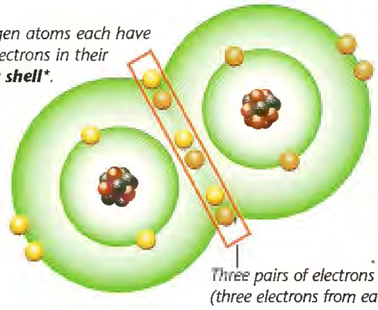

Three pairs of electrons shared (three electrons from each atom)

Each atom now has a stable **outer shell*** of eight electrons (see **octet**, page 127).

A **triple bond** is shown in a formula by a triple line: N≡N

Double bond

A covalent bond that is formed when two pairs of electrons are shared between two atoms.

An oxygen molecule has a **double bond**.

Oxygen atoms each have six electrons in their **outer shell***.

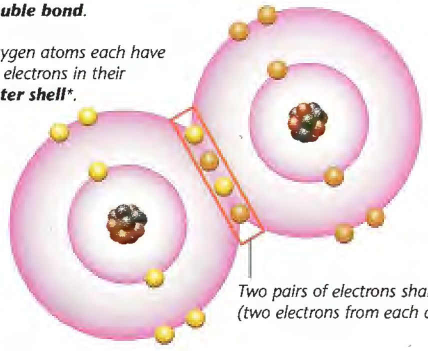

Two pairs of electrons shared (two electrons from each atom)

Each atom now has a stable **outer shell*** of eight electrons (see **octet**, page 127).

A **double bond** is shown in a formula by a double line: O=O

Dative covalent bond or coordinate bond

A covalent bond in which both electrons in the bond are provided by the same atom. It donates a **lone pair**.

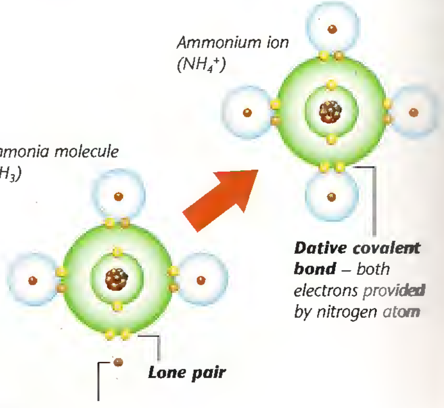

Ammonium ion (NH_4^+)

Ammonia molecule (NH_3)

Dative covalent bond – both electrons provided by nitrogen atom

Lone pair

Hydrogen ion (H^+)

*Ions, 130; **Outer shell**, 127; **van der Waals' forces**, 134.

Covalency

The maximum number
atom can form. It is eq
ydrogen atoms which
atom. The covalency of
onstant, but that of tra

drogen chloride (HCl)

drogen atom

ovalent elements have one or seven electrons in **outer shell***, e.g. hydrogen.

monia H_3)

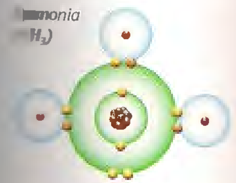

lent elements have either or five electrons in the **ter shell***, e.g. nitrogen.

Lone pair

pair of electrons in the
m which is not part of
ammonia picture on

ctronegativity

power of an atom to
tself in a molecule. If t
erent electronegativiti
nd is formed. Weakly e
ometimes called **elect**
um), as they form posi

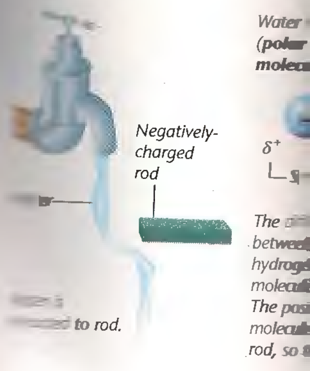

The
betwe
hydrog
molec
The pos
molecu
rod, so

ic compound, 131; **Isomers**, 191
shell, 127; **Polar solvent**, 144

Covalency

The maximum number of covalent bonds an atom can form. It is equal to the number of hydrogen atoms which will combine with the atom. The covalency of most elements is constant, but that of **transition metals*** varies.

Hydrogen chloride (HCl)

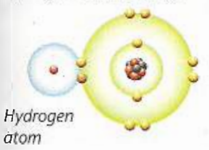

Monovalent** elements have either one or seven electrons in the **outer shell, e.g. hydrogen.*

Water (H_2O)

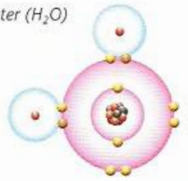

Divalent** elements have either two or six electrons in the **outer shell, e.g. oxygen.*

Ammonia (NH_3)

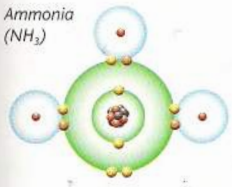

Trivalent** elements have either three or five electrons in the **outer shell, e.g. nitrogen.*

Methane (CH_4)

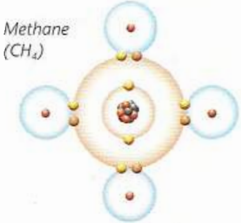

Tetravalent** elements have four electrons in the **outer shell, e.g. carbon.*

Lone pair

A pair of electrons in the **outer shell*** of an atom which is not part of a covalent bond (see ammonia picture on previous page).

Electronegativity

The power of an atom to attract electrons to itself in a molecule. If two atoms with different electronegativities are joined, a **polar bond** is formed. Weakly electronegative atoms are sometimes called **electropositive** (e.g. sodium), as they form positive ions fairly easily.

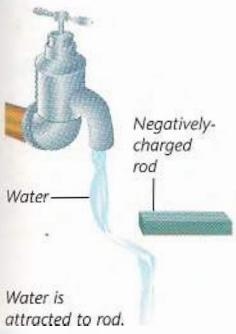

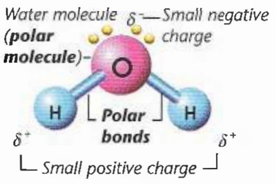

*The difference in **electronegativity** between oxygen (high) and hydrogen (low) causes water molecules to be **polar molecules**. The positive end of each water molecule is attracted to the charged rod, so the water "bends" toward it.*

Polar bond

A covalent bond in which the electrons spend a greater amount of time around one atom's **nucleus*** than the other. This effect is called **polarization**. It is caused by a difference in **electronegativity** between the atoms, the electrons being more attracted to one than the other.

Polar molecule

A molecule with a difference in electric charge between its ends, caused by an uneven distribution of **polar bonds**, and sometimes by **lone pairs**. Liquids with polar molecules may be **polar solvents*** and may dissolve **ionic compounds***. A **non-polar molecule** has no difference in charge at its ends.

Shapes of some simple polar and non-polar molecules

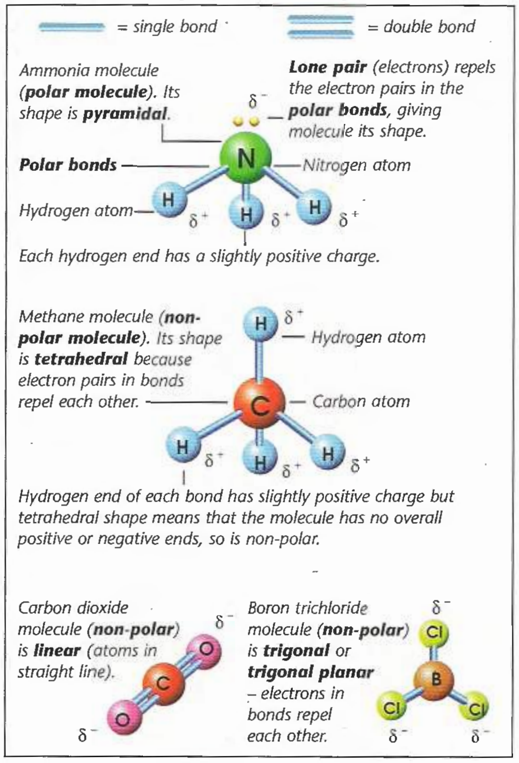

*Ammonia molecule (**polar molecule**). Its shape is **pyramidal**.*

***Lone pair** (electrons) repels the electron pairs in the **polar bonds**, giving molecule its shape.*

Each hydrogen end has a slightly positive charge.

*Methane molecule (**non-polar molecule**). Its shape is **tetrahedral** because electron pairs in bonds repel each other.*

Hydrogen end of each bond has slightly positive charge but tetrahedral shape means that the molecule has no overall positive or negative ends, so is non-polar.

*Carbon dioxide molecule (**non-polar**) is **linear** (atoms in straight line).*

*Boron trichloride molecule (**non-polar**) is **trigonal** or **trigonal planar** – electrons in bonds repel each other.*

Isomerism

The occurrence of the same atoms forming different arrangements in different molecules. The arrangements are **isomers***. They have the same **molecular formula*** but their other formulas may differ (see page 140).

---

* **Ionic compound**, 131; **Isomers**, 191; **Molecular formula**, 140; **Nucleus**, 126; **Outer shell**, 127; **Polar solvent**, 144; **Transition metals**, 172.

Metallic bonding

Metallic bonding is the attraction between particles in a **giant metallic lattice*** (i.e. in metals). The lattice consists of positive **ions*** of the metal with **valency electrons*** free to move between them. The free or **delocalized** electrons form the bonds between the ions and, because they can move, heat and electricity can be conducted through the metal. The forces between the electrons and ions are strong. This gives metals high melting and boiling points, since relatively large amounts of energy are needed to overcome them. For more about other types of bonding, see pages 130-133.

Delocalization

The sharing of **valency electrons*** by all the atoms in a molecule or **giant metallic lattice***. Delocalized electrons can belong to any of the atoms in the lattice and are able to move through the lattice, so the metal can conduct electricity and heat.

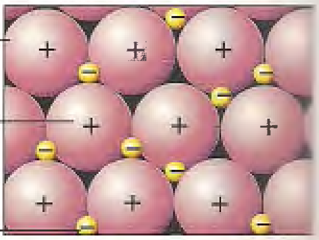
Giant metallic lattice*
*Metal **cations****
***Delocalized** electrons transmit electricity and heat.*
+
−

Intermolecular forces

van der Waals' forces

Weak attractive forces between molecules (**intermolecular forces***) caused by the uneven distribution and movement of electrons in the atoms of the molecules. The attractive force is approximately twenty times less than in **ionic bonding***. It is the force which holds **molecular lattices*** together, e.g. iodine, and solid carbon dioxide.

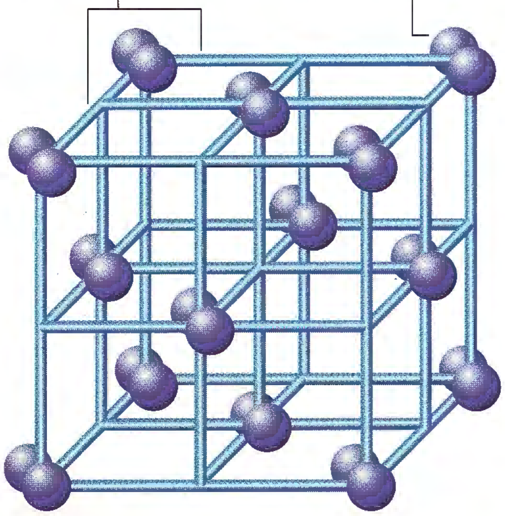
*Weak **van der Waals'** forces hold together this **molecular lattice*** of iodine.*
*Atoms of iodine held together in molecules by **covalent bonding****.*

Hydrogen bond

An attraction between a **polar molecule*** containing hydrogen and a **lone pair*** of electrons in another molecule. The **polar bonds*** mean that each hydrogen atom has a slightly positive charge and is therefore attracted to the electrons. Hydrogen bonding accounts for high melting and boiling points in water in relation to other substances with small, but **non-polar molecules***. Both the hydrogen bonds and the **van der Waals' forces** must be overcome to separate the molecules.

*Water molecule (**polar molecule***)*

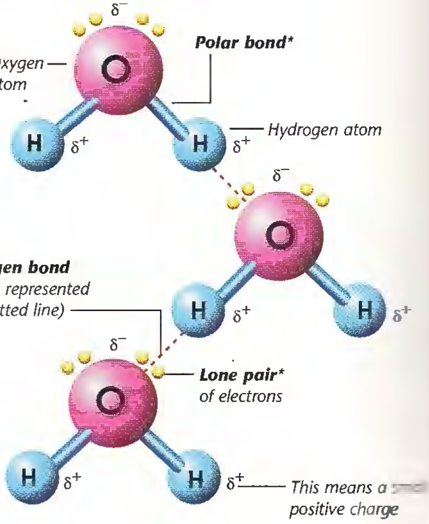
Oxygen atom
Polar bond*
Hydrogen atom
***Hydrogen bond** (usually represented by a dotted line)*
Lone pair of electrons*
This means a small positive charge
O
H
δ−
δ+

CRYSTALS

Crystals are solids wi
om regular arrange
e atoms, **ions*** or m
em can be of any t
dges of crystals are
bstances that form
crystalline. Solids wi
hich do not form cr

Crystallization

e process of forming cr
a number of ways, e.g.
ds, **subliming*** solids (so
), placing a **seed crys**
persaturated* solution o
stal in a **saturated*** solu
porating the solution. T
e most common. Either
porating means that the
te* decreases, so parti
tion and bond to the se
pended in the solution.
used to purify substanc

thods of crystallization

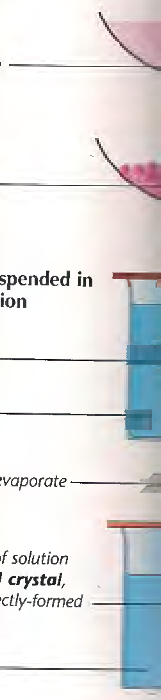
vent* allowed
aporate
ted solution*
crystals of
solute*
crystal suspended in
rated* solution
crystal
ted solution
** allowed to evaporate*
** coming out of solution*
*itself to **seed crystal**,*
ng large, perfectly-formed
liquid

---

* **Cation**, 130; **Covalent bonding**, 132; **Giant metallic lattice**, 137; **Intermolecular force**, 7; **Ion**, 130; **Ionic bonding**, 131
Lone pair, 133; **Molecular lattice**, 137; **Non-polar molecule**, 133 (**Polar molecule**); **Polar bond**, 133; **Valency electron**

ge plane, 136; **Formulas**,
153; **Saturated**, 145; **Solute**

CRYSTALS

Crystals are solids with regular geometric shapes, formed from regular arrangements of particles. The particles can be atoms, **ions*** or molecules and the bonding between them can be of any type or mixture of types. The edges of crystals are straight and the surfaces flat. Substances that form crystals are described as **crystalline**. Solids without a regular shape (i.e. those which do not form crystals) are described as **amorphous**.

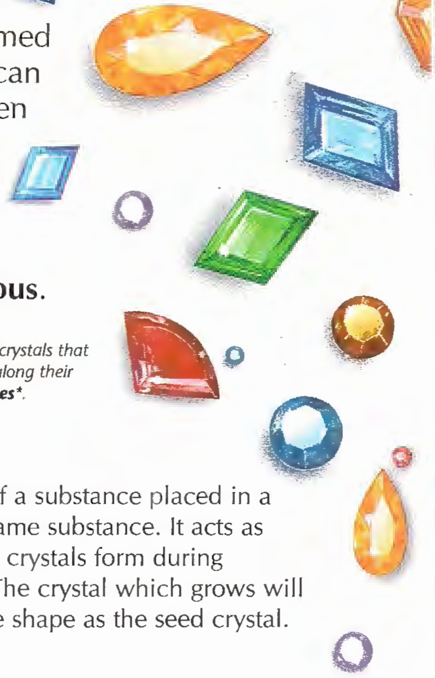

*Gemstones are crystals that have been cut along their **cleavage planes****.*

Crystallization

The process of forming crystals. It can happen in a number of ways, e.g. cooling **molten*** solids, **subliming*** solids (solid to gas and back), placing a **seed crystal** (see right) in a **supersaturated*** solution or placing a seed crystal in a **saturated*** solution and cooling or evaporating the solution. The last method is the most common. Either cooling or evaporating means that the amount of soluble **solute*** decreases, so particles come out of the solution and bond to the seed crystal, which is suspended in the solution. Crystallization can be used to purify substances – see page 221.

Methods of crystallization

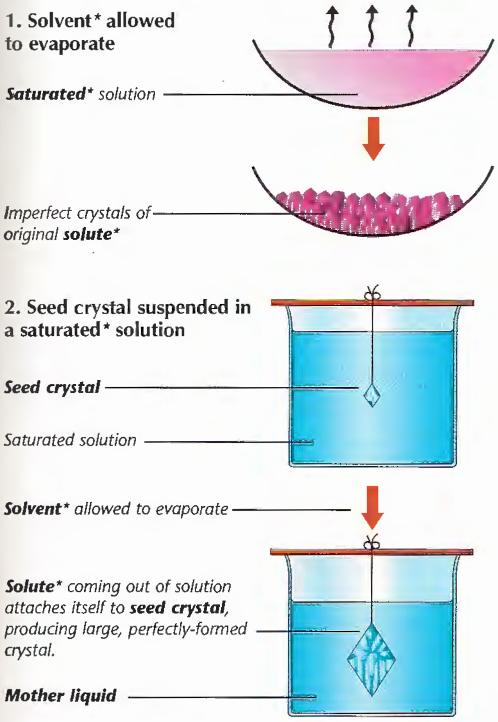
1. Solvent* allowed to evaporate
Saturated* solution
*Imperfect crystals of original **solute****
2. Seed crystal suspended in a saturated* solution
Seed crystal
Saturated solution
Solvent* allowed to evaporate
Solute* coming out of solution attaches itself to **seed crystal**, producing large, perfectly-formed crystal.
Mother liquid

Seed crystal

A small crystal of a substance placed in a solution of the same substance. It acts as a base on which crystals form during **crystallization**. The crystal which grows will take on the same shape as the seed crystal.

Mother liquid

The solution left after **crystallization** has taken place in a solution.

Water of crystallization

Water contained in crystals of certain **salts***. The number of molecules of water combined with each pair of **ions*** is usually constant and is often written in the chemical **formula*** for the salt. The water can be driven off by heating. Crystals which contain water of crystallization are **hydrated***.

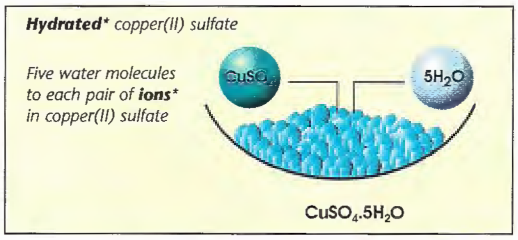

Hydrated* copper(II) sulfate

*Five water molecules to each pair of **ions*** in copper(II) sulfate*

$CuSO_4.5H_2O$

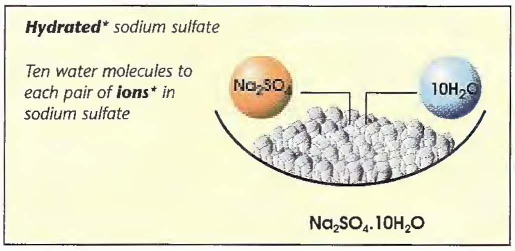

Hydrated* sodium sulfate

*Ten water molecules to each pair of **ions*** in sodium sulfate*

Na_2SO_4
$10H_2O$

$Na_2SO_4.10H_2O$

---

* **Cleavage plane**, 136; **Formulas**, 140; **Hydrated**, 154 (**Hydrate**); **Ion**, 130; **Molten**, 120; **Salts**, 153; **Saturated**, 145; **Solute**, **Solvent**, 144; **Sublimation**, 121; **Supersaturated**, 145.

Crystals continued – shapes and structures

Crystals (see page 135) exist in many different shapes and sizes. This is due to the different arrangement and bonding of the particles (atoms, molecules or **ions***). The arrangement in space of the particles and the way in which they are joined is called a **crystal lattice**. Different types of crystal lattice are shown on the opposite page. The shape of a particular crystal depends on its crystal lattice and how this lattice can be split along **cleavage planes**. The main crystal shapes are shown on the right – these are the basic shapes from which large crystals are built. A substance may have more than one crystalline form – see **polymorphism**, below.

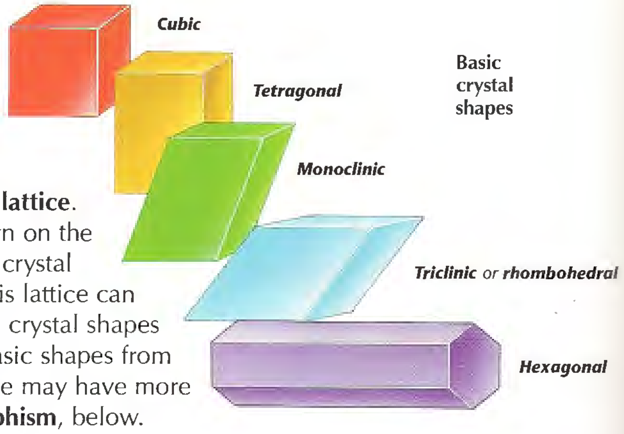

Basic crystal shapes

Polymorphism

The occurrence of two or more different crystals of the same substance, differing in shape and appearance. It is caused by different arrangements in the separate types. Changes between types often take place at a certain temperature called the **transition temperature**. Polymorphism in elements is called **allotropy**.

Allotropy

The occurrence of certain elements in more than one crystalline form. It is a specific type of **polymorphism**. The different forms are called **allotropes** and are caused by a change in arrangement of atoms in the crystals.

Monotropy

Polymorphism in which there is only one stable form. The other forms are unstable and there is no **transition temperature**.

Enantiotropy

Polymorphism in which there are two stable forms of a substance, one above its **transition temperature**, and one below.

Sulfur exhibits **enantiotropy** – it has two allotropes.

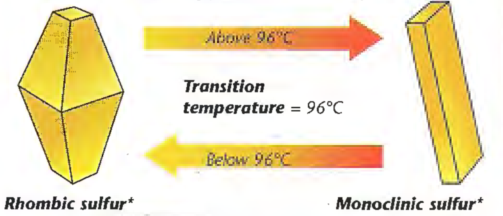

Transition temperature

The temperature at which a substance exhibiting **enantiotropy** changes from one form to another.

Isomorphism

The existence of two or more different substances with the same crystal structure and shape. They are described as **isomorphic**.

Cleavage plane

A plane of particles along which a crystal can be split, leaving a flat surface. If a crystal is not split along the cleavage plane, it shatters.

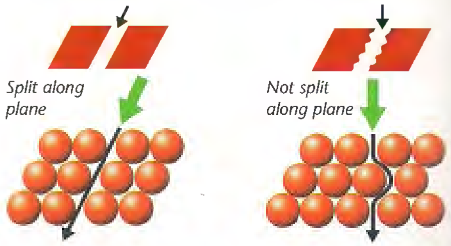

X-ray crystallography

The use of X-rays to work out crystal structure. Deflected X-rays produce a **diffraction pattern** from which the structure is worked out (see below).

X-ray crystallography

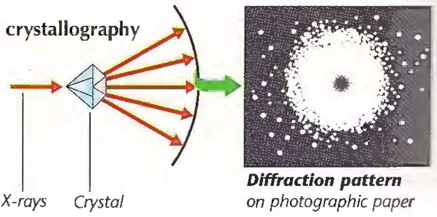

* **Ions**, 130; **Monoclinic sulfur**, **Rhombic sulfur**, 184.

Crystal lattices

Giant atomic lattice

A **crystal lattice** consisti
together by **covalent bon**
substances with giant at
extremely strong and hav
and boiling points.

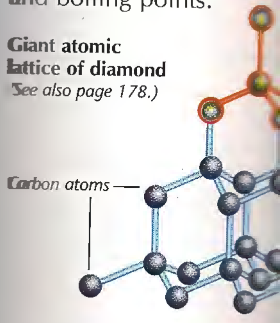

Giant atomic lattice of diamond
(See also page 178.)

Giant ionic lattice

A **crystal lattice** consistin
together by **ionic bonding**
chloride. The ionic bond
means that the substance
and boiling points.

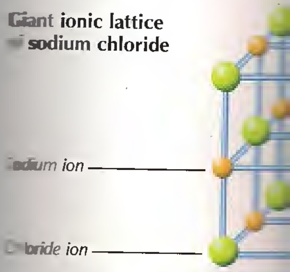

Giant ionic lattice of sodium chloride

Giant metallic lattice

A **crystal lattice** consistin
held together by **metallic**
The **delocalized*** electro
about, making a metal a
heat and electricity. The
slide over one another,
malleable* and **ductile***.

Giant metallic lattice of zinc

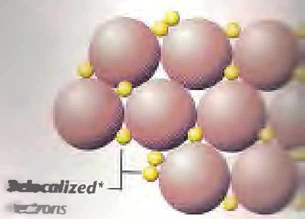

Cation, 130; **Covalent bonding**,
Ionic bonding, **Ionic compound**

Crystal lattices

Giant atomic lattice

A **crystal lattice** consisting of atoms held together by **covalent bonding***, e.g. diamond. Substances with giant atomic lattices are extremely strong and have very high melting and boiling points.

Giant atomic lattice of diamond
(See also page 178.)

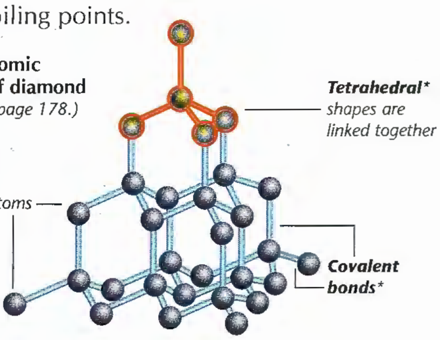
Tetrahedral* *shapes are linked together*
Carbon atoms
Covalent bonds*

Giant ionic lattice

A **crystal lattice** consisting of **ions*** held together by **ionic bonding***, e.g. sodium chloride. The ionic bonds are strong, which means that the substance has high melting and boiling points.

Giant ionic lattice of sodium chloride

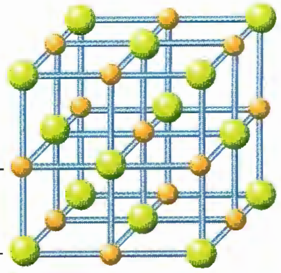
Sodium ion
Chloride ion

Giant metallic lattice

A **crystal lattice** consisting of metal atoms held together by **metallic bonding***, e.g. zinc. The **delocalized*** electrons are free to move about, making a metal a good conductor of heat and electricity. The layers of atoms can slide over one another, making metals **malleable*** and **ductile***.

Giant metallic lattice of zinc

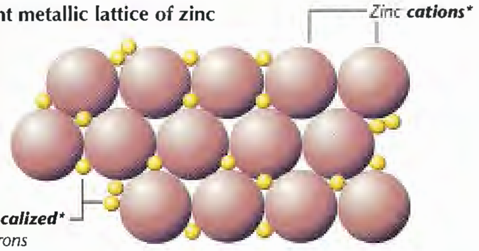
Zinc ***cations****
Delocalized* *electrons*

Molecular lattice

A **crystal lattice** consisting of molecules bonded together by **intermolecular forces** (see page 134), e.g. iodine. These forces are weak, so the crystal has low melting and boiling points compared with **ionic compounds*** and is easily broken. The **covalent bonds*** within the molecules themselves are stronger and break less easily.

Molecular lattice of solid iodine

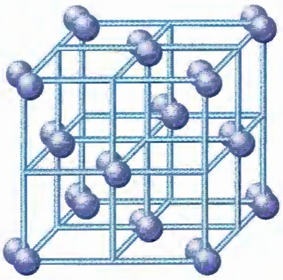
Iodine molecule

In crystals where the particles are all the same size, e.g. in a **giant metallic lattice**, various arrangements of the particles are possible. The most common are shown below.

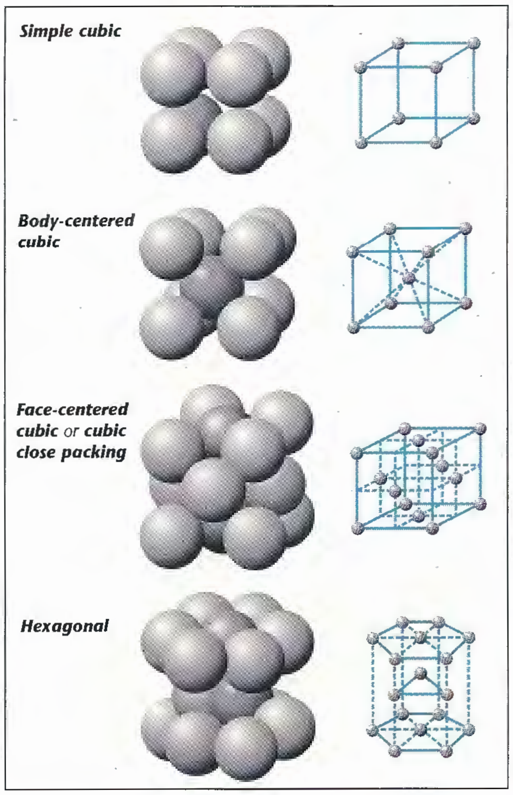
Simple cubic
Body-centered cubic
Face-centered cubic *or* ***cubic close packing***
Hexagonal

---

* **Cation**, 130; **Covalent bonding**, 132; **Delocalization**, 134; **Ductile**, 344; **Ion**, 130; **Ionic bonding**, **Ionic compound**, 131; **Malleable**, 345; **Metallic bonding**, 134; **Tetrahedral**, 133.

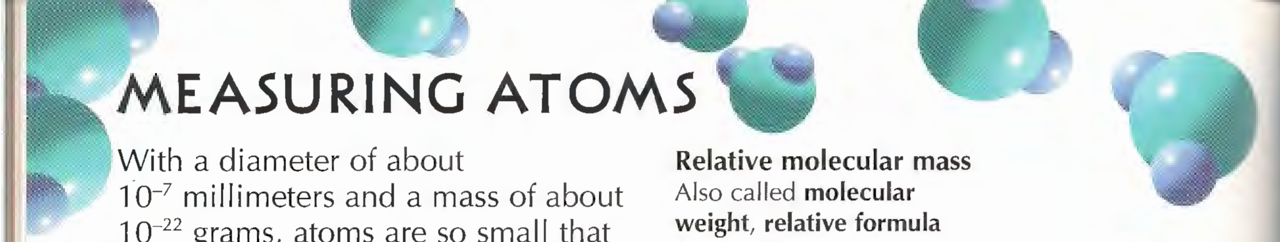

MEASURING ATOMS

With a diameter of about 10^{-7} millimeters and a mass of about 10^{-22} grams, atoms are so small that they are extremely difficult to measure. Their masses are therefore measured in relation to an agreed mass to give them a manageable value. Because there are many millions of atoms in a very small sample of a substance, the **mole** is used for measuring quantities of particles. The masses of atoms and molecules are measured using a machine called a **mass spectrometer**.

Relative atomic mass or atomic weight

The average mass (i.e. taking into account **relative isotopic mass** and **isotopic ratio**) of one atom of a substance divided by one twelfth the mass of a carbon-12 atom (see **isotope**, page 127). See page 83 for more about its units, and pages 212-213 for a table of relative atomic masses.

*Hydrogen has a **relative atomic mass** of approximately 1.*

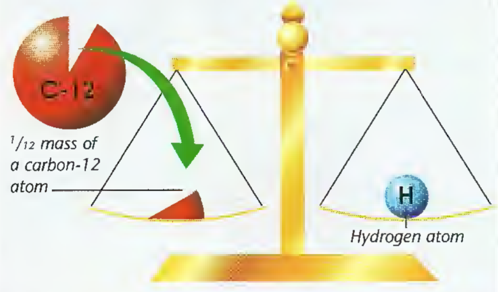

*Oxygen has a **relative atomic mass** of approximately 16.*

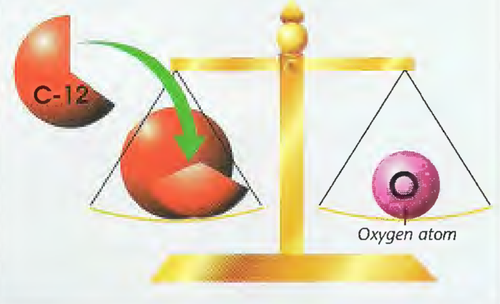

Relative molecular mass

Also called **molecular weight**, **relative formula mass** or **formula weight**. The mass of a molecule of an element or compound divided by one twelfth the mass of a carbon-12 atom (see **isotope**, page 127). It is the sum of the **relative atomic masses** of the atoms in the molecule.

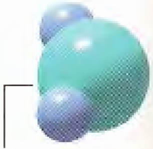

A water molecule contains one oxygen atom and two hydrogen atoms.

Relative molecular mass of water

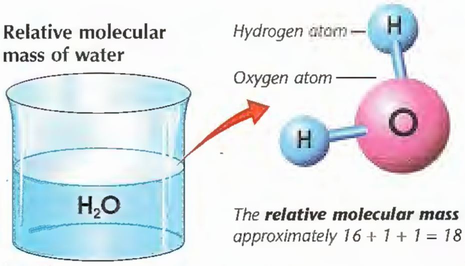

*The **relative molecular mass** is approximately 16 + 1 + 1 = 18*

Relative molecular mass** also applies to **ionic compounds, even though they do not have molecules.*

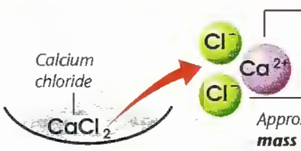

*Calcium (**relative atomic mass** 40)*

*Chlorine (**relative atomic mass** 35.5)*

*Approximate **relative molecular mass** = 40 + (2 × 35.5) = 111*

Relative isotopic mass

The mass of an atom of a specific **isotope*** divided by one twelfth the mass of a carbon-12 atom. It is nearly exactly the same as the **mass number*** of the isotope.

Isotopic ratio

The ratio of the number of atoms of each **isotope*** in a sample of an element. It is used with **relative isotopic masses** to calculate the **relative atomic mass** of an element.

Natural sample of chlorine contains about three times as many atoms of Cl-35 as Cl-37.

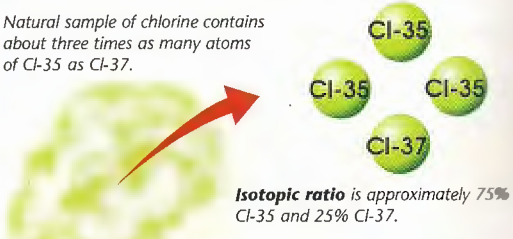

***Isotopic ratio** is approximately 75% Cl-35 and 25% Cl-37.*

*So **relative atomic mass** of chlorine = average **relative isotopic mass** = (3 × 35) + (1 × 37) ÷ 4 = 35.5*

* **Ionic compound**, 131; **Isotope**, **Mass number**, 127.

Mole (mol)

The **SI unit*** of the amou
(See also page 96.) One
same number of particle
in 12 grams of the carb

Avogadro's number

The number of particles
$6.023 \times 10^{23} mol^{-1}$.

*Each **mole** of copper contains **Avogadro's number** of atoms.*

*Each **mole** oxygen **Avogadro number** molecules.*

Molar mass

The mass of one **mole**
It is the **relative atomic**
of a substance expressed

***Relative atomic mass** = 23*

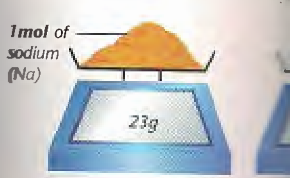

***Molar mass** 23g*

Molar volume

The volume of one **mole**
measured in cubic liters
solids and liquids vary
the same conditions ha
olume. The molar volu
s.t.p.* is 22.4 liters and
emperature and pressure
01,325 **pascals***) it is

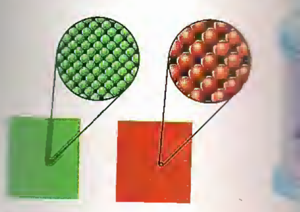

*solids and liquids, **molar volume** depends size and arrangement particles.*

Isotope, 127; **Pascal**, 97; **SI units**
s.t.p., 143; **Volumetric analysis**

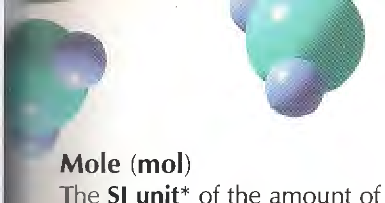

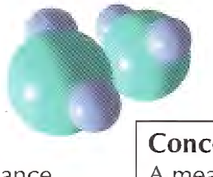

Mole (mol)

The **SI unit*** of the amount of a substance. (See also page 96.) One mole contains the same number of particles as there are atoms in 12 grams of the carbon-12 **isotope***.

Avogadro's number

The number of particles per **mole**, equal to $6.023 \times 10^{23} \text{mol}^{-1}$.

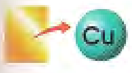

*Each **mole** of copper contains **Avogadro's number** of atoms.*

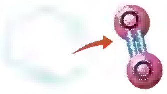

*Each **mole** of oxygen contains **Avogadro's number** of molecules.*

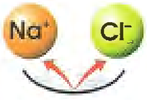

*A **mole** of sodium chloride contains 1**mol** Na^+ ions and 1mol Cl^- ions.*

Molar mass

The mass of one **mole** of a given substance. It is the **relative atomic** or **molecular mass** of a substance expressed in grams.

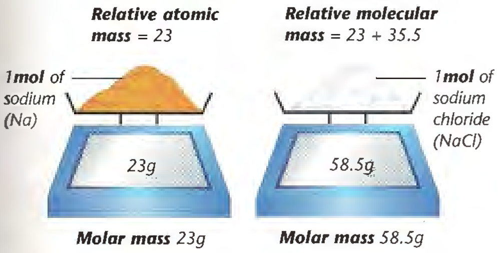
Relative atomic mass = 23
Relative molecular mass = 23 + 35.5
1**mol** of sodium (Na)
1**mol** of sodium chloride (NaCl)
23g
58.5g
Molar mass 23g
Molar mass 58.5g

Molar volume

The volume of one **mole** of any substance, measured in cubic liters. Molar volumes of solids and liquids vary, but all gases under the same conditions have the same molar volume. The molar volume of any gas at **s.t.p.*** is 22.4 liters and at **r.t.p.** (**room temperature and pressure**, i.e. 20°C and 101,325 **pascals***) it is 24dm^3.

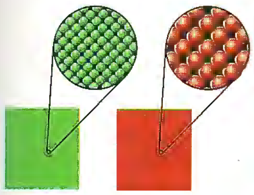

*In solids and liquids, **molar volume** depends on size and arrangement of particles.*

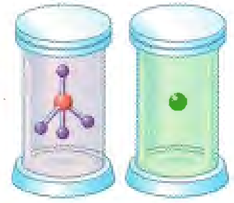

*All gases (at same temperature and pressure) have same **molar volume**. Their particles are not bonded together.*

Concentration

A measurement of the amount of a **solute*** dissolved in a **solvent***, expressed in **moles** per liter. **Mass concentration** is the mass of solute per unit volume, e.g. grams per liter.

Concentration** is the number of **moles** of **solute dissolved in each liter of **solvent***.*

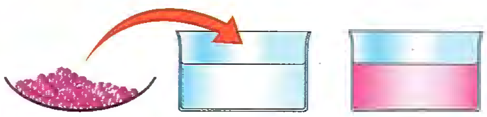
4mol of **solute***
2 liters of **solvent***
Concentration of 2 moles per liter

Molarity

A term sometimes used to describe the **concentration** when expressed in **moles** of **solute*** per liter of **solvent***. The molarity is also expressed as the **M-value**, e.g. a solution with a concentration of 3 moles per liter has a molarity of 3 and is described as a 3**M** solution.

*A **2M** copper(II) sulfate solution contains **2mol** of copper(II) sulfate in each liter.*

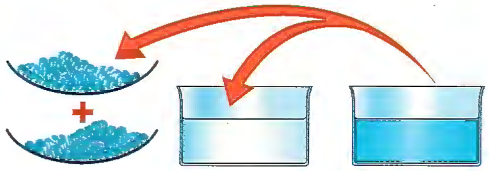
1mol $CuSO_4$ each
1liter water
2M solution

Molar solution

A solution that contains one **mole** of a substance dissolved in every liter of solution. It is therefore a 1M solution (see **molarity**).

*A **molar solution** (or 1M solution) of copper(II) sulfate contains 1**mol** of copper(II) sulfate in each liter.*

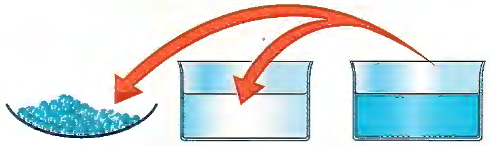
1mol copper(II) sulfate
1 liter water
1 liter molar solution of copper(II) sulfate

Standard solution

A solution of which the **concentration** is known. It is used for **volumetric analysis***.

---

* **Isotope**, 127; **Pascal**, 97; **SI units**, 96; **Solute**, **Solvent**, 144; **s.t.p.**, 143; **Volumetric analysis**, 222.

REPRESENTING CHEMICALS

Most chemicals are named according to the predominant elements they contain. Information about the chemical composition and structure of a compound is given by a **formula** (pl. **formulas**), in which the **chemical symbols*** for the elements are used. A chemical **equation** shows the reactants and products of a chemical reaction and gives information about how the reaction happens.

Formulas

Empirical formula

A formula showing the simplest ratio of the atoms of each element in a compound. It does not show the total number of atoms of each element in a **covalent compound***, or the **bonding** in the compound (see pages 130-134).

Displayed formula or full structural formula

A formula which shows the arrangement of the atoms in relation to each other in a molecule. All the bonds in the molecule are shown. In a displayed formula, single bonds are represented by a single line, double bonds by a double line, and so on.

Molecular formula

A formula representing one molecule of an element or compound. It shows which elements the molecule contains and the number of atoms of each in the molecule, but not the **bonding** of the molecule (see pages 130-134).

Diagram of ethene molecule, showing the types of bonds within the molecule

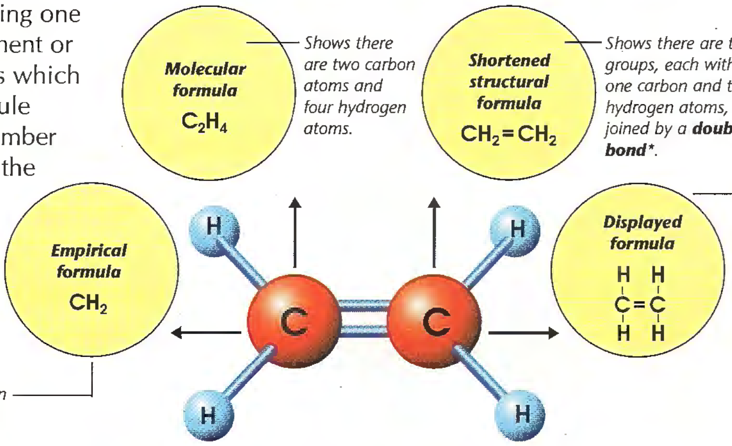

Shows there are two carbon atoms and four hydrogen atoms.

Shows there are two groups, each with one carbon and two hydrogen atoms, joined by a ***double bond****.*

Shows which atom is bonded to which, and the ***single*** *and* ***double bonds**** *present.*

Shows there are two hydrogen atoms to every carbon atom.

Shortened structural formula

A formula which shows the sequence of groups of atoms (e.g. a **carboxyl group***) in a molecule and the **bonding** (see pages 130-134) between the groups of atoms (shown as lines).

Stereochemical formula or 3-dimensional structural formula

A formula which uses symbols to show the 3-dimensional arrangement of the atoms and **bonds*** in a molecule. See **stereochemistry**, page 191, for the stereochemical formula of methane.

Percentage composition

The composition of a compound expressed in terms of the percentage of its mass taken up by each element.

Percentage composition *of carbon dioxide (CO_2)*

One carbon atom. ***Relative atomic mass**** *= 12*

Two oxygen atoms. ***Relative atomic mass*** *= 2 × 16 = 32*

Relative molecular mass* *of compound =*

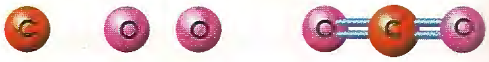

12 + (2 × 16) = 44

Percentage of oxygen = (32 ÷ 44) × 100 = 73%

Percentage of carbon = (12 ÷ 44) × 100 = 27%

Therefore, the ***percentage composition*** *of carbon dioxide = 27% carbon, 73% oxygen.*

*** Bonds**, 130; **Carboxyl group**, 195 (**Carboxylic acids**); **Chemical symbol**, 122; **Covalent compounds**, **Double bond**, 132; **Relative atomic mass**, **Relative molecular mass**, 138; **Single bond**, 132.

Names

Trivial name

An everyday name given It does not usually give about the composition or compound, e.g. salt (sod calcium carbonate).

Traditional name

A name which gives the elements of a substance, giving their quantities or structure of the substanc names are **systematic na**

Trivial name*: Green vitriol*
Traditional name*: Ferrous sulfate*

Iron sulfate ($FeSO_4$)

Equations

Word equation

An equation in which th involved in a reaction a names, e.g:

Sodium + Water → Sodium

However, the names ma the **formulas** of the subs opposite page).

$Na + H_2O \rightarrow$

Balanced equation

An equation in which the of each element involve the same on each side of it obeys the **law of conser** The numbers of molecu substance are shown by front of their **formula**, e.

$2Na + 2H_2O \rightarrow$

Aqueous solution, 144; **Law of c**
Oxidation number, **Oxidation st**

Names

Trivial name

An everyday name given to a compound. It does not usually give any information about the composition or structure of the compound, e.g. salt (sodium chloride), chalk (calcium carbonate).

Traditional name

A name which gives the predominant elements of a substance, without necessarily giving their quantities or showing the structure of the substance. Some traditional names are **systematic names**.

***Trivial name**: Green vitriol*
***Traditional name**: Ferrous sulfate*

Iron sulfate ($FeSO_4$)

Systematic name
Iron(II) tetraoxosulfate(VI)

Oxidation state of iron, i.e. Fe^{2+}*

Oxidation state of sulfur is +6*

This name is normally simplified to iron(II) sulfate.

Systematic name

A name which shows the elements a compound contains, the ratio of numbers of atoms of each element and the **oxidation number*** of elements with variable **oxidation states***. The **bonding** (see pages 130-134) can also be worked out from the name. In some cases the systematic name is simplified. Some systematic names are the same as **traditional names**. See also **naming simple organic compounds**, page 214.

Trivial name
Alcohol

Traditional name
Ethyl alcohol

Systematic name
*Ethanol (see **naming simple organic compounds**, page 214)*

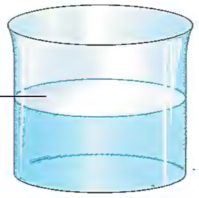

Equations

Word equation

An equation in which the substances involved in a reaction are indicated by their names, e.g:

Sodium + Water → Sodium hydroxide + Hydrogen

However, the names may be replaced by the **formulas** of the substances (see opposite page).

$$Na + H_2O \rightarrow NaOH + H_2$$

Balanced equation

An equation in which the number of atoms of each element involved in the reaction is the same on each side of the equation (i.e. it obeys the **law of conservation of mass***). The numbers of molecules of each substance are shown by the number in front of their **formula**, e.g:

$$2Na + 2H_2O \rightarrow 2NaOH + H_2$$

Ionic equation

An equation which only shows changes which occur to the ions in a reaction. (See example at bottom of page.)

State symbols

Letters written after the **formula** of a substance which show its **physical state*** in a reaction.

$$2Na(s) + 2H_2O(l) \rightarrow 2NaOH(aq) + H_2(g)$$

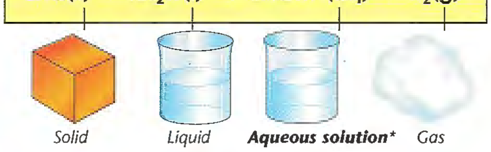

Spectator ion

An ion which remains the same after a chemical reaction.

*In the reaction below, Na^+, OH^-, H^+, Cl^- are all ions. Na^+, Cl^- appear on both sides of the equation – on this occasion they are **spectator ions**. Spectator ions are omitted from **ionic equations**.*

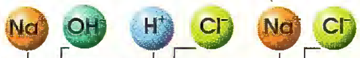

$$NaOH(aq) + HCl(aq) \rightarrow NaCl(aq) + H_2O(l)$$

***Ionic equation** is:* $OH^-(aq) + H^+(aq) \rightarrow H_2O(l)$

***Aqueous solution**, 144; **Law of conservation of mass**, 125; **Oxidation number**, **Oxidation state**, 149; **Physical states**, 120.

GAS LAWS

The molecules in a gas are widely spaced and move about in a rapid, chaotic manner (see **kinetic theory***). The combined volume of the gas molecules is much smaller than the volume the gas occupies, and the forces of attraction between the molecules are very weak. This is true for all gases, so they all behave in a similar way. Several **gas laws** describe this common behavior (see below).

Symbols used in **gas laws**
P = pressure
T = temperature in **kelvins**
V = volume
k = a **constant***

Gas at constant temperature, pressure and volume

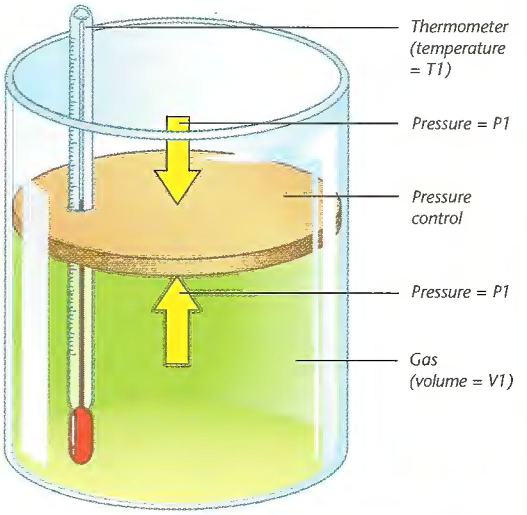

Thermometer (temperature = T1)
Pressure = P1
Pressure control
Pressure = P1
Gas (volume = V1)

Boyle's law

At constant temperature, the volume of a gas is inversely proportional to the pressure (the volume decreases as the pressure increases).

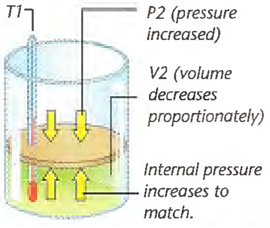

T1
P2 (pressure increased)
V2 (volume decreases proportionately)
Internal pressure increases to match.

$$P1 \times V1 = P2 \times V2$$
$$\text{or } PV = k$$

Law of volumes

At constant pressure, the volume is directly proportional to the temperature on the **absolute temperature scale** (the gas expands as the temperature increases).

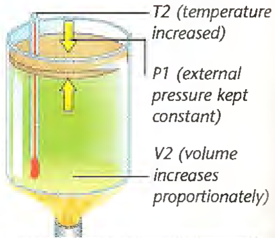

T2 (temperature increased)
P1 (external pressure kept constant)
V2 (volume increases proportionately)

$$\frac{V1}{T1} = \frac{V2}{T2} \quad \text{or} \quad \frac{V}{T} = k$$

Pressure law or Third gas law

At constant volume, the pressure is directly proportional to the temperature on the **absolute temperature scale** (the pressure increases with the temperature).

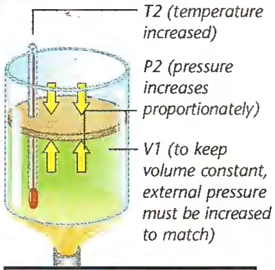

T2 (temperature increased)
P2 (pressure increases proportionately)
V1 (to keep volume constant, external pressure must be increased to match)

$$\frac{P1}{T1} = \frac{P2}{T2} \quad \text{or} \quad \frac{P}{T} = k$$

Ideal gas equation or General gas equation

An equation that shows the relationship between the pressure, volume and temperature of a fixed mass of gas.

$$\frac{P1 \times V1}{T1} = \frac{P2 \times V2}{T2}$$

For one **mole*** of gas: $PV = RT$
R is the **gas constant***. It is the same for all gases.

Ideal gas

A theoretical gas that behaves in an "ideal" way. Its molecules have no volume, do not attract each other, move rapidly in straight lines and lose no energy when they collide. Many real gases behave in approximately the same way as ideal gases when the molecules are small and widely spaced.

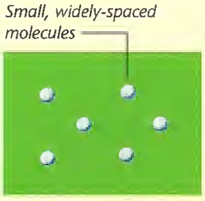

Small, widely-spaced molecules

Behaves like an **ideal gas.**

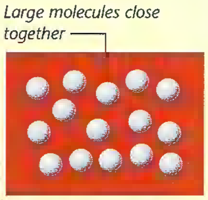

Large molecules close together

Does not behave like an **ideal gas.**

---

* **Constant**, 344; **Gas constant**, 113; **Kinetic theory**, 123; **Mole**, 139.

Partial pressure

The pressure that each
gases would exert if it
occupied by the

Dalton's law of

The total pressure
gases (which do not
the sum of the **partial**
the mixture.

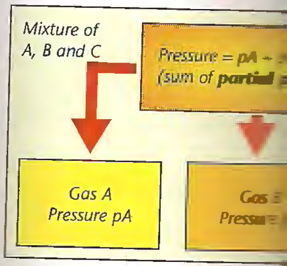

Mixture of A, B and C
Pressure = pA
(sum of
Gas A
Pressure pA

Graham's law of

If the temperature and
the rate of **diffusion*** of
proportional to the
The density of a gas is
heavy, and low if its
molecules move faster
so a gas with a high
slowly than a gas with a

Cotton ball soaked in ammonia solution

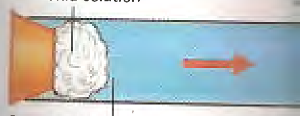

Ammonia gas
White ring
chloride forms

Light ammonia molecules **diffuse***
molecules. The two gases meet
the tube.

Rate of **diffusion** ∝
(proport

Relative vapor density

The density of a gas
of hydrogen. It is
density of a gas by the
Relative vapor density is
units.

Relative vapor density =

* **Celsius scale**, 27; **Diffusion**, 123;

Partial pressure

The pressure that each gas in a **mixture*** of gases would exert if it alone filled the volume occupied by the mixture.

Dalton's law of partial pressures

The total pressure exerted by a **mixture*** of gases (which do not react together) is equal to the sum of the **partial pressure** of each gas in the mixture.

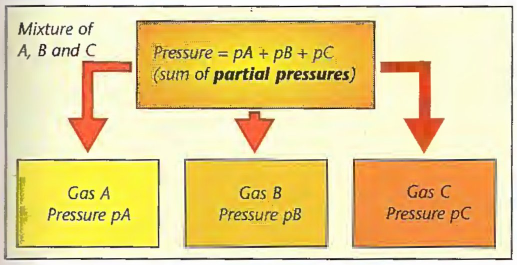
Mixture of A, B and C
Pressure = pA + pB + pC (sum of **partial pressures**)
Gas A Pressure pA
Gas B Pressure pB
Gas C Pressure pC

Graham's law of diffusion

If the temperature and pressure are constant, the rate of **diffusion*** of a gas is inversely proportional to the square root of its density. The density of a gas is high if its molecules are heavy, and low if its molecules are light. Light molecules move faster than heavy molecules, so a gas with a high density diffuses more slowly than a gas with a low density.

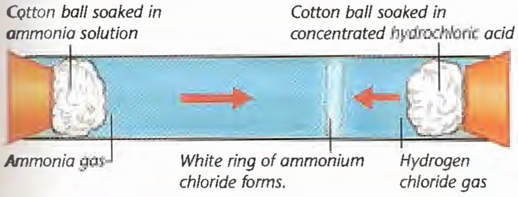
Cotton ball soaked in ammonia solution
Cotton ball soaked in concentrated hydrochloric acid
Ammonia gas
White ring of ammonium chloride forms.
Hydrogen chloride gas

*Light ammonia molecules **diffuse*** faster than hydrogen chloride molecules. The two gases meet nearer to the right-hand end of the tube.*

$$\text{Rate of diffusion} \propto \frac{1}{\sqrt{\text{density of gas}}}$$

(proportional)

Relative vapor density

The density of a gas relative to the density of hydrogen. It is calculated by dividing the density of a gas by the density of hydrogen. Relative vapor density is a ratio and has no units.

$$\textbf{Relative vapor density} = \frac{\text{density of the gas}}{\text{density of hydrogen}}$$

Gay-Lussac's law

When gases react together to produce other gases and all the volumes are measured at the same temperature and pressure, the volumes of the reactants and products are in a ratio of simple whole numbers.

$2CO(g) + O_2(g) \rightarrow 2CO_2(g)$

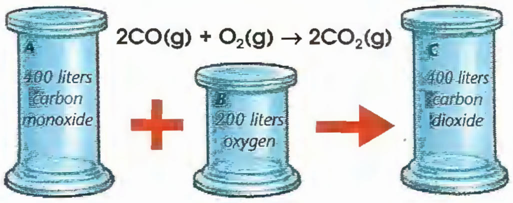
A
400 liters carbon monoxide
B
200 liters oxygen
C
400 liters carbon dioxide

*According to **Avogadro's law** (below), jars A and C above contain the same number of molecules.*

Avogadro's law or Avogadro's hypothesis

Equal volumes of all gases at the same temperature and pressure contain the same number of molecules.

s.t.p.

An abbreviation for **standard temperature and pressure**. These are internationally agreed standard conditions under which properties such as volume and density of gases are usually measured.

s.t.p. = temperature: 0°C or 273K (**kelvins**)
pressure: 101,325 **pascals***

Absolute temperature scale

A standard temperature scale, using units called **kelvins** (**K**). A kelvin is the same size as one degree **Celsius***, but the lowest point on the scale, zero kelvins or **absolute zero**, is equal to –273 degrees Celsius, a theoretical point where an **ideal gas** would occupy zero volume.

*To convert degrees **Celsius** to **kelvins**, add 273. To convert **kelvins** to degrees **Celsius**, subtract 273.*

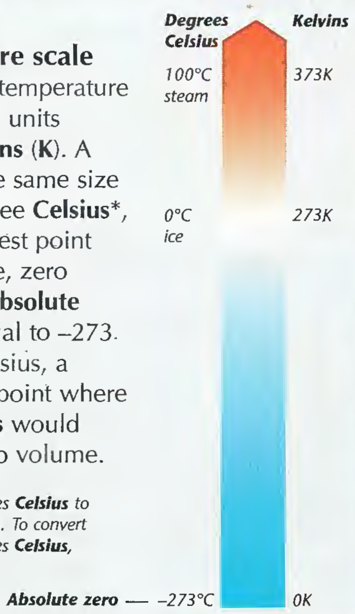
Degrees Celsius
Kelvins
100°C steam
373K
0°C ice
273K
Absolute zero
–273°C
0K

---

* **Celsius scale**, 27; **Diffusion**, 123; **Mixture**, 122; **Pascal**, 97.

SOLUTIONS AND SOLUBILITY

When a substance is added to a liquid, several things can happen. If the atoms, molecules or ions of the substance become evenly dispersed (**dissolve**), the **mixture*** is a **solution**. If they do not, the mixture is either a **colloid**, a **suspension**, or a **precipitate**. How well a substance dissolves depends on its properties, those of the liquid, and other factors such as temperature and pressure.

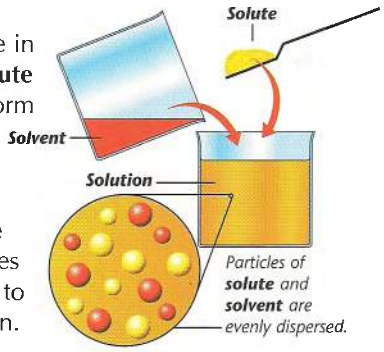
Solute
Solvent
Solution
Particles of **solute** and **solvent** are evenly dispersed.

Solvent
The substance in which the **solute** dissolves to form a solution.

Solute
The substance which dissolves in the **solvent** to form a solution.

Solvation
The process of **solvent** molecules combining with **solute** molecules as the solute dissolves. When the solvent is water, the process is called **hydration**. Whether or not solvation takes place depends on how much the molecules of the solvent and solute attract each other, and how strong the **bonds*** in the solute are.

Polar solvent
A liquid with **polar molecules***. Polar solvents generally dissolve **ionic compounds***. **Solvation** occurs because the charged ends of the solvent molecules attract the ions of the **giant ionic lattice***. Water is the most common polar solvent.

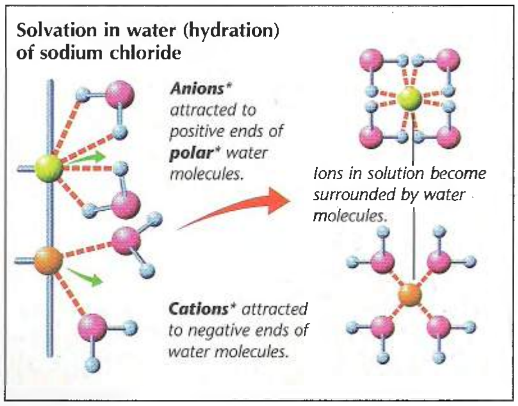
Solvation in water (hydration) of sodium chloride
Anions* attracted to positive ends of **polar*** water molecules.
Ions in solution become surrounded by water molecules.
Cations* attracted to negative ends of water molecules.

Non-polar solvent
A liquid with **non-polar molecules***. Non-polar solvents dissolve **covalent compounds***. The **solute** molecules are pulled from the **molecular lattice*** by the solvent molecules and **diffuse*** through the solvent. Many organic liquids are non-polar solvents.

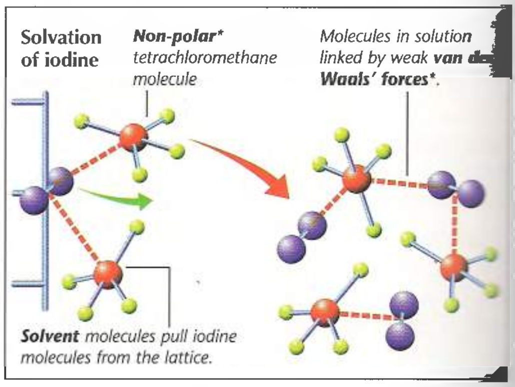
Solvation of iodine
Non-polar* tetrachloromethane molecule
Molecules in solution linked by weak **van de Waals' forces***.
Solvent molecules pull iodine molecules from the lattice.

Aqueous solvent
A **solvent** containing water. Water molecules are **polar***, so aqueous solvents are **polar solvents**.

Aqueous solution
A solution formed from an **aqueous solvent**. Aqueous solvents are **polar solvents** and form aqueous solutions. **Non-polar solvents** are **non-aqueous solvents** and form non-aqueous solutions.

Dilute
Describes a solution with a low **concentration*** of **solute**.

Dilute solution of copper(II) sulfate

Concentrated
Describes a solution with a high **concentration*** of **solute**.

Concentrated solution of copper(II) sulfate

*** Anion**, **Bonding**, **Cation**, 130; **Concentration**, 139; **Covalent compounds**, 132; **Diffusion**, 123; **Giant ionic lattice**, 137; **Ionic compound**, 131; **Mixture**, 122; **Molecular lattice**, 137; **Non-polar molecule**, 133 (**Polar molecule**); **van der Waals' forces**, 134.

Saturated
Describes a solution that
more **solute** at a given te
solute will remain as cry
temperature is raised, mo
until the solution become

Supersaturated
Describes a solution with
solute than a **saturated** so
temperature. It is formed
cooled below the tempera
would be saturated, and
for the solute to **crystalliz**
"extra" solute remains dis
is unstable – if crystals are
enters, the "extra" solute

Soluble
Describes a **solute** which
solvent. The opposite of so

Solubility
The amount of a **solute** wh
particular amount of **solve**
temperature.

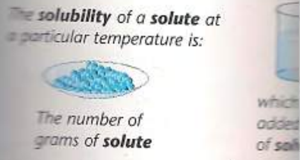
The **solubility** of a **solute** at a particular temperature is:
The number of grams of **solute**
which added of sol

The solubility of a solid usu
temperature, while the solu
decreases.

Sugar dissolves better in hot tea than in cold water.

The change of solubility wi
shown by a **solubility curve**.

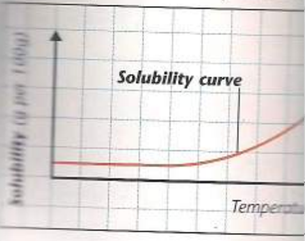

Crystallization, 135; **Diffusion**, 123

Saturated

Describes a solution that will not dissolve any more **solute** at a given temperature (any more solute will remain as crystals). If the temperature is raised, more solute may dissolve until the solution becomes saturated again.

Supersaturated

Describes a solution with more dissolved **solute** than a **saturated** solution at the same temperature. It is formed when a solution is cooled below the temperature at which it would be saturated, and there are no particles for the solute to **crystallize*** around, so the "extra" solute remains dissolved. The solution is unstable – if crystals are added or dust enters, the "extra" solute forms crystals.

Soluble

Describes a **solute** which dissolves easily in a **solvent**. The opposite of soluble is **insoluble**.

Solubility

The amount of a **solute** which dissolves in a particular amount of **solvent** at a known temperature.

*The **solubility** of a **solute** at a particular temperature is:*

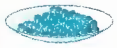

*The number of grams of **solute***

*which must be added to 100g of **solvent***

*to produce a **saturated** solution.*

The solubility of a solid usually increases with temperature, while the solubility of a gas decreases.

Sugar dissolves better in hot tea than in cold water.

Warm soft drinks have more bubbles than cold ones.

The change of solubility with temperature is shown by a **solubility curve**.

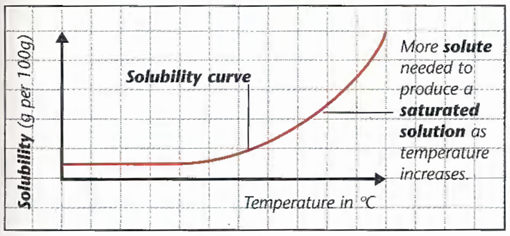

*More **solute** needed to produce a **saturated solution** as temperature increases.*

Precipitate

An **insoluble** solid (see **soluble**) formed when a reaction occurs in a solution.

***Precipitates** are named according to their appearance.*

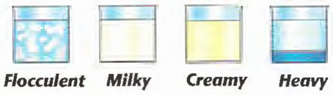

*The reaction below forms a dense white **precipitate** of silver chloride.*

*Dense **precipitate** of silver chloride*

$$AgNO_3(aq) + NaCl(aq) \rightarrow AgCl(\downarrow) + NaNO_3(aq)$$

Silver nitrate

Sodium chloride

*This symbol means **precipitate***

Miscible

Describes two or more liquids which **diffuse*** together. The opposite is **immiscible**.

Suspension

Fine particles of a solid (groups of atoms, molecules or ions) suspended in a liquid in which the solid does not dissolve.

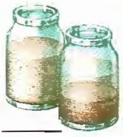

***Suspension** of sand and soil in water*

Particles settle to the bottom.

OR

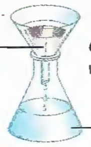

Particles can be filtered out.

Clear water

Colloid

A **mixture*** of extremely small particles of a substance dispersed in another in which it does not dissolve. The particles (groups of atoms, molecules or ions) are smaller than in a **suspension**.

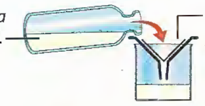

*Milk is a **colloid**.*

Particles pass through filter paper and do not settle.

***Emulsion**. A **colloid** consisting of tiny particles of one liquid dispersed in another liquid, e.g. mayonnaise.*

***Foam**. A **colloid** of small bubbles of gas dispersed in a liquid.*

***Mist**. A **colloid** consisting of tiny particles of a liquid dispersed in a gas.*

***Smoke**. A **colloid** consisting of tiny particles of a solid dispersed in a gas.*

---

* **Crystallization**, 135; **Diffusion**, 123; **Mixture**, 122.

ENERGY AND CHEMICAL REACTIONS

Nearly all chemical reactions involve a change in energy. Some reactions involve electrical energy or light energy, but almost all involve heat energy. The change in energy in a reaction results from the different amounts of energy involved when bonds are broken and formed. The study of heat energy in chemical reactions is called **thermochemistry**.

Enthalpy (H)

The amount of energy that a substance contains. It is impossible to measure directly, but its change during a reaction can be measured.

Enthalpy change of reaction or **heat of reaction** (ΔH)

The amount of heat energy given out or absorbed during a chemical reaction. If the reaction is a **change of state***, this amount is also known, particularly in physics, as the **latent heat** (see page 30). Hence the **molar enthalpy changes of fusion** and **vaporization** on page 147 are closely allied to the **specific latent heats** on page 31, though the quantities are different (each being more relevant to its science).

Enthalpy change = *total* **enthalpy** *of products* − *total* **enthalpy** *of reactants*

Enthalpy change of reaction of hydrogen and oxygen

$2H_2(g) + O_2(g) \rightarrow 2H_2O(g) \quad \Delta H = -488kJ$

Heat is given out, so it is an ***exothermic reaction.***

The value of ΔH *is only true for the number of* ***moles**** *and the* ***physical states**** *of the chemicals in the equation.*

J *stands for* ***joule****, a unit of energy.* ***kJ*** *stands for* ***kilojoule*** *(1,000* ***joules****).*

Energy level diagram

A diagram which shows the **enthalpy change of reaction** for a reaction.

Energy level diagram for above reaction of hydrogen and oxygen

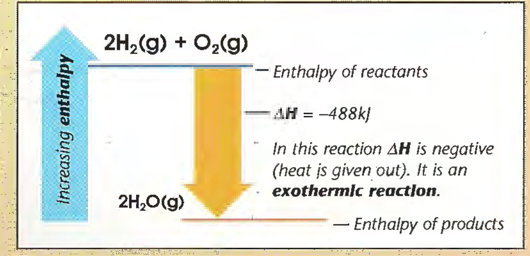
Increasing enthalpy
$2H_2(g) + O_2(g)$
Enthalpy of reactants
$\Delta H = -488kJ$
$2H_2O(g)$
Enthalpy of products

In this reaction ΔH *is negative (heat is given out). It is an* ***exothermic reaction.***

Standard enthalpy change of reaction ($\Delta H^{\ominus}$)

An **enthalpy change of reaction** measured at room temperature and pressure (**r.t.p.***). If solutions are used, their **concentration*** is 1M*.

Exothermic reaction

A chemical reaction during which heat is transferred to the surroundings.

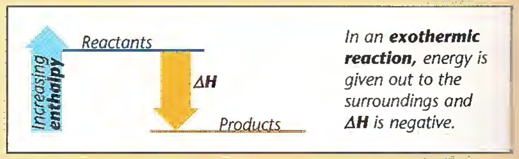
Increasing enthalpy
Reactants
ΔH
Products

In an ***exothermic reaction****, energy is given out to the surroundings and* ΔH *is negative.*

Endothermic reaction

A chemical reaction during which heat is absorbed from the surroundings.

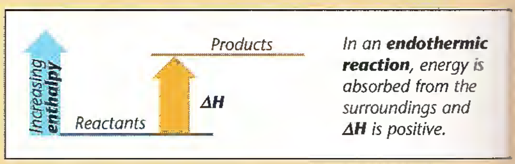
Increasing enthalpy
Products
ΔH
Reactants

In an ***endothermic reaction****, energy is absorbed from the surroundings and* ΔH *is positive.*

Bond energy

A measure of the strength of a **covalent bond*** formed between two atoms. Energy must be supplied to break bonds and is given out when bonds are formed. A difference in these energies produces a change in energy during a reaction.

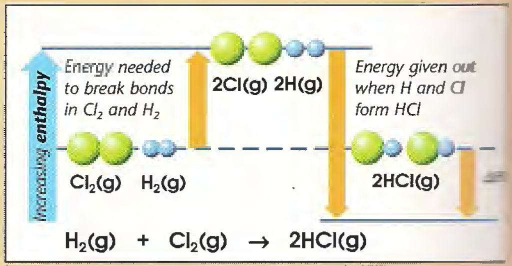
Increasing enthalpy
Energy needed to break bonds in Cl_2 and H_2
$2Cl(g)$ $2H(g)$
Energy given out when H and Cl form HCl
$Cl_2(g)$ $H_2(g)$
$2HCl(g)$

$H_2(g) + Cl_2(g) \rightarrow 2HCl(g)$

Law of conservation of energy

During a chemical reaction, energy cannot be created or destroyed. In a **closed system*** the amount of energy is constant.

---

* **Change of state**, 120; **Closed system**, 162; **Concentration**, 139; **Covalent bond**, 132; **Joule**, 97; **M-value**, 139 (**Molarity**); **Mole**, 139; **Physical states**, 120; **r.t.p.**, 139 (**Molar volume**).

Hess's law

States that the **enthalpy ch** occurs during a particula always the same, no matte in going from the reactants law can be illustrated by a right). Hess's law is used changes of reaction which directly, e.g. the **enthalpy** of methane.

Special enthalpy

Enthalpy change of co or heat of combustion

The amount of heat ene one **mole*** of a substance burned in oxygen. The for a substance is meas **calorimeter**.

Bomb calorimeter

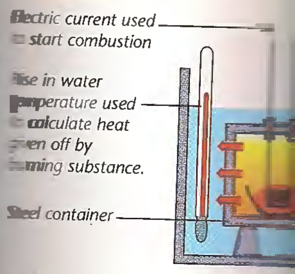
Electric current used to start combustion
Rise in water temperature used to calculate heat given off by burning substance.
Steel container

Enthalpy change of neu or heat of neutralization

The amount of heat energ when one **mole*** of hydro neutralized* by one mole OH⁻). If the acid and alk the heat of neutralization The **ionic equation*** for ne

$H^+(aq) + OH^-(aq) \rightarrow$

Hydrogen ion
Hydroxide ion

When a **weak acid*** or a we involved, the heat produced energy must be supplied to acid fully.

Crystal **lattice**, 136; **Ionic equation**, neutralization, 151; **Solvent**, 144;

Hess's law

States that the **enthalpy change of reaction** that occurs during a particular chemical reaction is always the same, no matter what route is taken in going from the reactants to the products. The law can be illustrated by an **energy cycle** (see right). Hess's law is used to find enthalpy changes of reaction which cannot be measured directly, e.g. the **enthalpy change of formation** of methane.

Energy cycle

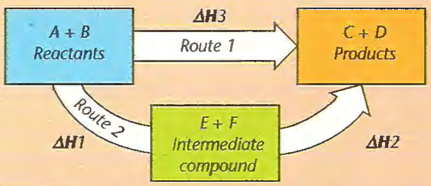
A + B
Reactants
ΔH3
Route 1
C + D
Products
Route 2
ΔH1
E + F
Intermediate compound
ΔH2

By ***Hess's law:***

$\Delta H1$ (for reaction $A + B \rightarrow E + F$) + $\Delta H2$ (for reaction $E + F \rightarrow C + D$) = $\Delta H3$ (for reaction $A + B \rightarrow C + D$)

Special enthalpy changes

Enthalpy change of combustion or heat of combustion

The amount of heat energy given out when one **mole*** of a substance is completely burned in oxygen. The heat of combustion for a substance is measured using a **bomb calorimeter**.

Bomb calorimeter

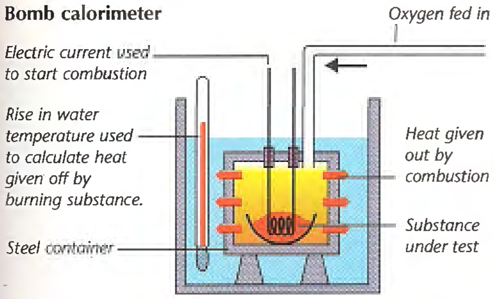
Oxygen fed in
Electric current used to start combustion
Rise in water temperature used to calculate heat given off by burning substance.
Heat given out by combustion
Substance under test
Steel container

Enthalpy change of neutralization or heat of neutralization

The amount of heat energy given out when one **mole*** of hydrogen ions (H^+) is **neutralized*** by one mole of hydroxide ions (OH^-). If the acid and alkali are fully **ionized***, the heat of neutralization is always –57kJ. The **ionic equation*** for neutralization is:

$H^+(aq) + OH^-(aq) \rightarrow H_2O(l)$ $\Delta H = -57kJ$

Hydrogen ion *Hydroxide ion* *Water molecule*

When a **weak acid*** or a **weak base*** is involved, the heat produced is less. Some energy must be supplied to ionize the acid fully.

Enthalpy change of solution or heat of solution

The amount of heat energy given out or taken in when one **mole*** of a substance dissolves in such a large volume of **solvent*** that further dilution produces no heat change.

Molar enthalpy change of fusion or molar heat of fusion

The amount of heat energy required to change one **mole*** of a solid into a liquid at its melting point. Energy must be supplied to break the bonds in the **crystal lattice*** of the solid.

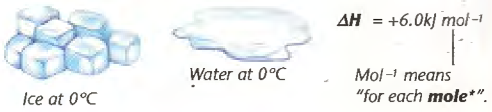
Ice at 0°C
Water at 0°C
ΔH = +6.0kJ mol⁻¹
Mol⁻¹ means "for each mole*".

Molar enthalpy change of vaporization or molar heat of vaporization

The heat energy needed to change one **mole*** of a liquid into a vapor at its boiling point.

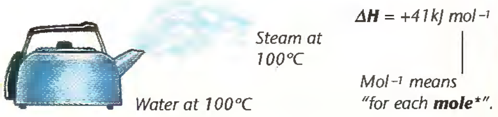
Steam at 100°C
Water at 100°C
ΔH = +41kJ mol⁻¹
Mol⁻¹ means "for each mole*".

Enthalpy change of formation or heat of formation

The heat energy given out or taken in when one **mole*** of a compound is formed from elements. For example:

$C(graphite) + O_2(g) \rightarrow CO_2(g)$ $\Delta H = -394kJ$

Carbon *Oxygen* *Carbon dioxide*

***Crystal lattice**, 136; **Ionic equation**, 141; **Ionization**, 130; **Mole**, 139; **Neutralization**, 151; **Solvent**, 144; **Weak acid**, **Weak base**, 152.

OXIDATION AND REDUCTION

The terms **oxidation** and **reduction** originally referred to the gain and loss of oxygen by a substance. They have now been extended to include the gain and loss of hydrogen and electrons. There is always a transfer of electrons in reactions involving oxidation and reduction, that is, the **oxidation state** of one or more of the elements is always changed.

Oxidation

A chemical reaction in which one of the following occurs:

1. An element or compound gains oxygen

$$2CuO(s) + C(s) \rightarrow CO_2(g) + 2Cu(s)$$

Oxidizing agent — *Element* ***oxidized*** — *Carbon gains oxygen*

2. A compound loses hydrogen

$$Cl_2(g) + H_2S(g) \rightarrow 2HCl(g) + S(s)$$

Oxidizing agent — *Compound* ***oxidized*** — *Hydrogen sulfide loses hydrogen*

3. An atom or ion loses electrons

$$Cl_2(g) + 2Na(s) \rightarrow 2Na^+Cl^-(s)$$

Oxidizing agent — *Atom* ***oxidized*** — *Sodium loses electrons*

A substance that undergoes oxidation is said to be **oxidized**, and its **oxidation state** is increased. Oxidation is the opposite of **reduction**.

Oxidizing agent

A substance which accepts electrons, and so causes the **oxidation** of another substance. The **oxidizing agent** is always **reduced** in a reaction.

Reduction

A chemical reaction in which one of the following occurs:

1. A compound loses oxygen

$$2CuO(s) + C(s) \rightarrow CO_2(g) + 2Cu(s)$$

Compound ***reduced*** — ***Reducing agent*** — *Copper(II) oxide loses oxygen*

2. A campound or element gains hydrogen

$$Cl_2(g) + H_2S(g) \rightarrow 2HCl(g) + S(s)$$

Element ***reduced*** — ***Reducing agent*** — *Chlorine gains hydrogen*

3. An atom or ion gains electrons.

$$Cl_2(g) + 2Na(s) \rightarrow 2Na^+Cl^-(s)$$

Atom ***reduced*** — ***Reducing agent*** — *Chlorine gains electron*

A substance that undergoes reduction is said to be **reduced**, and its **oxidation state** is decreased. Reduction is the opposite of **oxidation**.

Reducing agent

A substance which donates electrons, and so causes the **reduction** of another substance. The reducing agent is always **oxidized** in a reaction.

Redox

Describes a chemical reaction involving **oxidation** and **reduction**. The two processes always occur together because an **oxidizing agent** is always reduced during oxidation, and a **reducing agent** is always oxidized during reduction. In the example on the right, magnesium and chlorine undergo a redox reaction to form magnesium chloride.

$$Mg(s) + Cl_2(g) \rightarrow Mg^{2+} + 2Cl^-(s)$$

The simultaneous **oxidation** and **reduction** of the same element in a reaction is called **disproportionation**.

Redox reaction of magnesium and chlorine

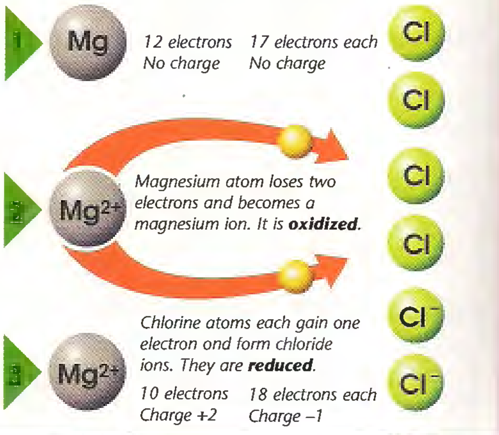

12 electrons No charge

17 electrons each No charge

Magnesium atom loses two electrons and becomes a magnesium ion. It is ***oxidized****.*

Chlorine atoms each gain one electron ond form chloride ions. They are ***reduced****.*

10 electrons Charge +2

18 electrons each Charge –1

Oxidation state

The number of electr
removed from, or add
forms a compound. The
element is usually equ
ion. An element's oxid
when it is **oxidized** an
is **reduced**.

Oxygen is ***reduced****.*

0

$$2Mg(s) + O_2(g$$

0

Magnesium is ***oxidized****.*

The rules below help to
oxidation state of an ele

O
δ⁻
Ca
+2
δ⁻
N
–3
δ⁻
Fe

1. The ***oxidation state*** *of a free element (one that is not part of a compound) is zero.*

2. The ***oxidation state*** *of an element in an* ***ionic compound**** *is equal to the electrical charge on its ion.*

Oxidation state *+2 (2 electrons removed)*

Oxidation states *of elements in* ***covalent compounds**** *are found by assuming the compound is* ***ionic*****, and working out the charge the ions would have. The most* ***electronegative**** *atom (the nearest to fluorine in the* ***periodic table*****) is assumed to gain the electrons.*

3. The sum of the ***oxidation states*** *of all the elements in a compound is zero.*

Iron(II) sulfate
Sum of ***oxidation states*** *= (+2)*

4. The ***oxidation state*** *of oxygen in a compound is normally –2, but in* ***peroxides****, e.g. hydrogen peroxide, it is –1.*

Covalent compounds, 132; **Electr**
Electronegativity, 133; **Giant i**

Oxidation state

The number of electrons which have been removed from, or added to, an atom when it forms a compound. The oxidation state of an element is usually equal to the charge on its ion. An element's oxidation state increases when it is **oxidized** and decreases when it is **reduced**.

*Oxygen is **reduced**.* ***Oxidation state** decreases.*

0 → −2

$$2Mg(s) + O_2(g) \rightarrow 2MgO(s)$$

0 → +2

*Magnesium is **oxidized**.* ***Oxidation state** increases.*

The rules below help to work out the oxidation state of an element:

*1. The **oxidation state** of a free element (one that is not part of a compound) is zero.*

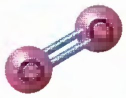

*The **oxidation state** of oxygen is 0. No electrons have been lost or gained.*

*2. The **oxidation state** of an element in an **ionic compound*** is equal to the electrical charge on its ion.*

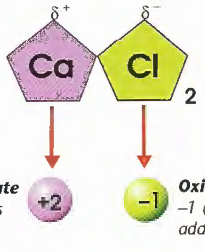

*Calcium chloride forms a **giant ionic lattice***.*

***Oxidation state** +2 (2 electrons removed)*

***Oxidation state** −1 (1 electron added)*

Oxidation states** of elements in **covalent compounds are found by assuming the compound is **ionic***, and working out the charge the ions would have. The most **electronegative*** atom (the nearest to fluorine in the **periodic table***) is assumed to gain the electrons.*

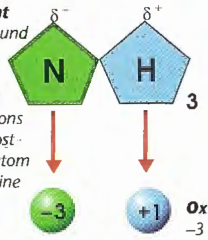

*Ammonia is a **covalent compound***.*

***Oxidation states** −3 and +1*

*3. The sum of the **oxidation states** of all the elements in a compound is zero.*

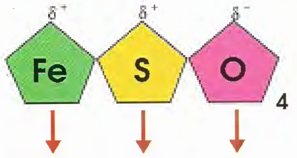

Iron(II) sulfate
*Sum of **oxidation states** = (+2) + (+6) + (4 × −2) = 0*

*4. The **oxidation state** of oxygen in a compound is normally −2, but in **peroxides**, e.g. hydrogen peroxide, it is −1.*

*5. The **oxidation state** of hydrogen is usually +1, except in metal **hydrides**, when it is −1.*

Oxidation number

A number that shows the **oxidation state** of an element in a compound. It is written in Roman numerals and placed in brackets after the name of the element. It is only included in the name of a compound when the element has more than one oxidation state.

Iron(III) chloride

***Oxidation number** of 3 and **oxidation state** of +3*

Lead(IV) oxide

***Oxidation number** of 4 and **oxidation state** of +4*

Redox potential

A measurement of the power of a substance to gain electrons in solution. A strong **reducing agent**, which readily loses electrons (which it can give to another substance), will have a high negative redox potential. A strong **oxidizing agent**, which easily gains electrons, will have a high positive redox potential. Redox potential is the same as **electrode potential***.

Redox series

A list of substances arranged in order of their **redox potentials**, the substance with the most negative redox potential being placed at the top. A substance usually **oxidizes** any substance above it in the series and **reduces** any substance below it. The further apart substances are in the series, the more easily they oxidize or reduce each other. The redox series is an extended version of the **electrochemical series***.

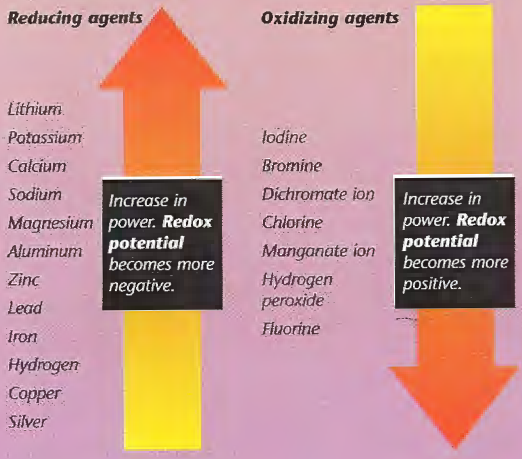

Reducing agents

Lithium
Potassium
Calcium
Sodium
Magnesium
Aluminum
Zinc
Lead
Iron
Hydrogen
Copper
Silver

*Increase in power. **Redox potential** becomes more negative.*

Oxidizing agents

Iodine
Bromine
Dichromate ion
Chlorine
Manganate ion
Hydrogen peroxide
Fluorine

*Increase in power. **Redox potential** becomes more positive.*

---

* **Covalent compounds**, 132; **Electrochemical series**, 159; **Electrode potential**, 158; **Electronegativity**, 133; **Giant ionic lattice**, 137; **Ionic compound**, 131; **Periodic table**, 164.

ACIDS AND BASES

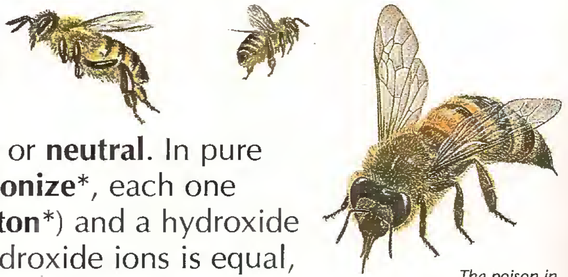

The poison in a bee's sting is an **acid**.

All chemicals are either **acidic**, **basic** or **neutral**. In pure water, a small number of molecules **ionize***, each one forming a hydrogen ion (a single **proton***) and a hydroxide ion. The number of hydrogen and hydroxide ions is equal, and the water is described as **neutral**. Some compounds dissolve in, or react with, water to produce hydrogen ions or hydroxide ions, which upset the balance. These compounds are either **acids** or **bases**.

Acid

A compound containing hydrogen which dissolves in water to produce hydrogen ions (H^+ – **protons***) in the solution. Hydrogen ions do not exist on their own in the solution, but join with water molecules to produce **hydronium ions**. These ions can only exist in solution, so an acid will only display its properties when it dissolves.

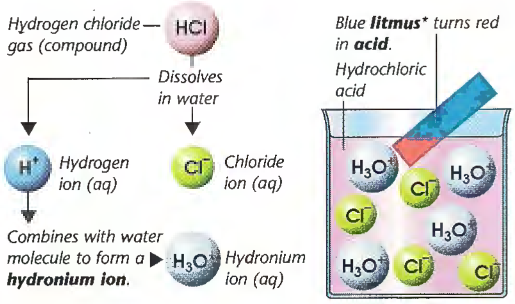
Hydrogen chloride gas (compound)
HCl
Dissolves in water
H^+
Hydrogen ion (aq)
Cl^-
Chloride ion (aq)
Combines with water molecule to form a **hydronium ion.**
H_3O^+
Hydronium ion (aq)
Blue **litmus*** turns red in **acid**.
Hydrochloric acid
H_3O^+
Cl^-

Types of acid

Citric and **ascorbic acid** in citrus fruits

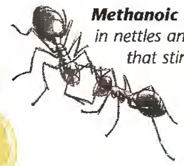

Methanoic acid in nettles and ants that sting

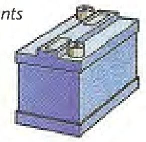

Sulfuric acid in **lead-acid accumulators***

Dilute acids have a sour taste, a **pH*** of less than 7 and turn blue **litmus*** red. They react with metals that are above hydrogen in the **electrochemical series*** to produce hydrogen gas.

$$\underset{\text{Acid}}{H_2SO_4(aq)} + \underset{\text{Metal}}{Mg(s)} \rightarrow \underset{\text{Salt*}}{MgSO_4(aq)} + \underset{\text{Hydrogen}}{H_2(g)}$$

Dilute **strong acids*** react with carbonates or hydrogencarbonates to produce carbon dioxide gas and are **neutralized** by **bases**.

Acidic

Describes any compound with the properties of an **acid**.

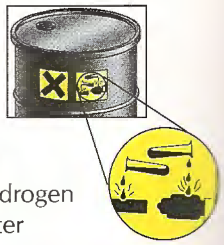

Some **acids** are corrosive and may have warning labels.

Hydronium ion (H_3O^+) or oxonium ion

An ion formed when a hydrogen ion attaches itself to a water molecule (see **acid**). When a reaction takes place in a solution containing hydronium ions, only the hydrogen ion takes part. Hence usually the hydronium ion can be considered to be a hydrogen ion.

Formation of hydronium ion (H_3O^+)

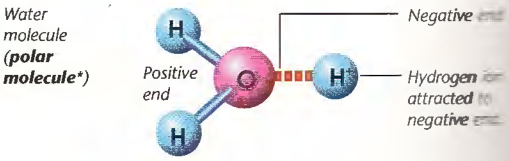
Water molecule (**polar molecule***)
Positive end
H
O
H
H^+
Negative end
Hydrogen ion attracted to negative end

Mineral acid

An acid which is produced chemically from a mineral, e.g. hydrochloric acid is produced from sodium chloride, and sulfuric acid is produced from sulfur.

Mineral acid	***Formula***
Hydrochloric	HCl
Sulfuric	H_2SO_4
Sulfurous	H_2SO_3
Nitric	HNO_3
Nitrous	HNO_2
Phosphoric	H_3PO_4

Organic acid	***Formula***
Ethanedioic (Oxalic)	(COOH
Methanoic (Formic)	HCOOH
Ethanoic (Acetic)	CH_3COOH

Organic acid

An **organic compound*** that is **acidic**. The most common ones are **carboxylic acids***.

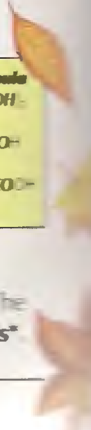

When leaves die and decompose, they form an **organic acid** called humic acid.

---

* **Carboxylic acids**, 195; **Electrochemical series**, 159; **Ionization**, 130; **Lead-acid accumulator**, 159; **Litmus**, 152; **Organic compounds**, 190; **pH**, 152; **Polar molecule**, 133; **Proton**, 126; **Salts**, 153; **Strong acid**, 152.

Base

A substance that will ne
accepting hydrogen ions
opposite of an acid. Bas
oxides and hydroxides
also a base. A substance
a base is described as ba
dissolves in water is an
produced when a base
ammonium **salt***.

*A wasp's sting contains an **alkali** and can be neutralized with an acid such as vinegar.*

*Household liquid cleaners contain **alkalis** that dissolve dirt.*

*Indigestion tablets contain **alkalis** such as magnesium hydroxide which **neutrallze** the acid produced by indigestion.*

Alkali

A **base**, normally a hydr
Group 1 or Group 2 of the
which is soluble in water
hydroxide ions (OH^-) in so
make a solution **alkaline**.

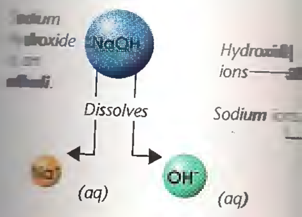
Sodium hydroxide
NaOH
Hydroxide ions
Sodium
Dissolves
Na^+
(aq)
OH^-
(aq)

Alkaline

scribes a solution forme
solves in water to form
tains more hydroxide
Alkaline solutions ha
7, turn red **litmus*** bl
cause they react with the
tions produced from str
a few metals, e.g. zinc
e off hydrogen gas.

$+ 2NaOH(aq) + 6H_2O(l) \rightarrow$
Sodium hydroxide
Water

radical, 153; **Antacid**, 344;
reaction, 162; **Salts**,

Base

A substance that will **neutralize** an **acid** by accepting hydrogen ions. It is the chemical opposite of an acid. Bases are usually metal oxides and hydroxides, although ammonia is also a base. A substance with the properties of a base is described as **basic**. A base which dissolves in water is an **alkali**. Ammonia is produced when a base is heated with an ammonium **salt***.

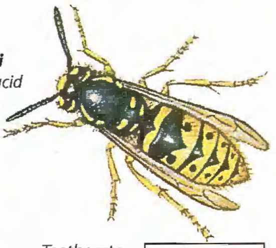

*A wasp's sting contains an **alkali** and can be neutralized with an acid such as vinegar.*

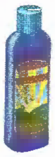

*Household liquid cleaners contain **alkalis** that dissolve dirt.*

*Indigestion tablets contain **alkalis** such as magnesium hydroxide which **neutralize** the acid produced by indigestion.*

*Toothpaste is a **base**. It neutralizes acids made in your mouth.*

Alkali

A **base**, normally a hydroxide of a metal in Group 1 or Group 2 of the **periodic table***, which is soluble in water and produces hydroxide ions (OH^-) in solution. These make a solution **alkaline**.

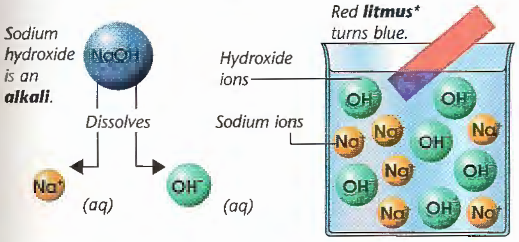
Sodium hydroxide is an **alkali**.
NaOH
Dissolves
Na⁺
(aq)
OH⁻
(aq)
Red **litmus*** turns blue.
Hydroxide ions
Sodium ions
OH⁻
Na⁺

Alkaline

Describes a solution formed when a **base** dissolves in water to form a solution which contains more hydroxide ions than hydrogen ions. Alkaline solutions have a **pH*** of more than 7, turn red **litmus*** blue, and feel soapy because they react with the skin. Alkaline solutions produced from **strong bases*** react with a few metals, e.g. zinc and aluminium, to give off hydrogen gas.

$$2Al(s) + 2NaOH(aq) + 6H_2O(l) \rightarrow 2NaAl(OH)_4(aq) + 3H_2(g)$$

$2Al(s)$	$2NaOH(aq)$	$6H_2O(l)$	$2NaAl(OH)_4(aq)$	$3H_2(g)$
Aluminium	*Sodium hydroxide*	*Water*	*Sodium aluminate*	*Hydrogen*

Amphoteric

Describes a substance that acts as an **acid** in one reaction, but as a **base** in another, e.g. zinc hydroxide.

Anhydride

A substance that reacts with water to form either an **acidic** or an **alkaline** solution (see **hydrolysis**, page 154). It is usually an oxide.

$$SO_2(g) + H_2O(l) \rightarrow H_2SO_3(aq)$$

$SO_2(g)$	$H_2O(l)$	$H_2SO_3(aq)$
*Sulfur dioxide (**anhydride**)*	*Water*	*Sulfurous acid*

Neutral

Describes a substance that does not have the properties of an **acid** or **base**. A neutral solution has an equal number of hydrogen and hydroxide ions. It has a **pH*** of 7 and does not change the color of **litmus***.

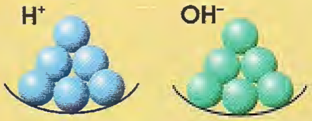

*A **neutral** solution contains an equal number of hydrogen and hydroxide ions.*

Neutralization

The reaction between an **acid** and a **base** to produce a **salt*** and water only. An equal number of hydrogen and hydroxide ions react together to form a **neutral** solution. The **acid radical*** from the acid and **cation*** from the base form a salt.

***Neutralization** is:*

ACID + BASE → SALT* + WATER

Bronsted-Lowry theory

Another way of describing **acids** and **bases**. It defines an acid as a substance which donates **protons***, and a base as one which accepts them.

*Ethanoic acid donates a **proton*** – it is an **acid**.*
*Water accepts a **proton*** – it is a **base**.*

$$CH_3COOH(aq) + H_2O(l) \rightleftharpoons H_3O^+(aq) + CH_3COO^-(aq)$$

*This sign means **reversible reaction****

Hydronium ion** donates **proton – it is an **acid**.*
*Ethanoate ion accepts **proton*** – it is a **base**.*

---

* **Acid radical**, 153; **Antacid**, 344; **Cation**, 130; **Litmus**, 152; **Periodic table**, 164; **pH**, 152; **Proton**, 126; **Reversible reaction**, 162; **Salts**, 153; **Strong base**, 152.

Acids and bases continued – strength and concentration

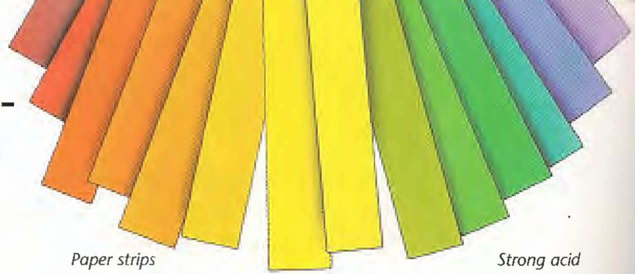

Paper strips impregnated with **universal indicator** can be used to test the strength of an acid.

Strong acid turns the paper red, and a strong alkaline solution turns it purple.

The **concentration*** of **acids** and **bases** (see previous two pages) depends on how many **moles*** of the acid or base are in a solution, but the strength depends on the proportion of their molecules which **ionize*** to produce **hydronium ions*** or hydroxide ions. A dilute **strong acid** can produce more hydrogen ions than a concentrated **weak acid**.

Strong acid

An acid that completely **ionizes*** in water, producing a large number of hydrogen ions in solution.

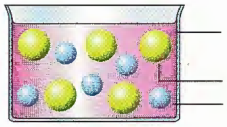
Acid radical (Cl^-)
H^+

Hydrochloric acid (**strong acid**). All hydrogen chloride molecules split up.

Weak acid

An acid that only partially **ionizes*** in water, i.e. only a small percentage of its molecules split into hydrogen ions and **acid radicals**.

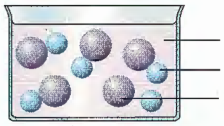
H^+
Acid radical (CH_3COO^-)

Ethanoic acid (**weak acid**). Only some molecules split up.

Strong base

A base that is completely **ionized*** in water. A large number of hydroxide ions are released to give a strongly alkaline solution.

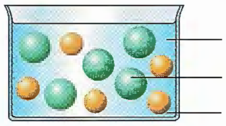

Sodium hydroxide (**strong base**). All the molecules are **ionized***.

Weak base

A base that is only partially **ionized*** in water. Only some molecules turn into hydroxide ions, giving a weakly alkaline solution.

Ammonia reacts slightly with water to give a low concentration of hydroxide ions:

$$NH_3(aq) + H_2O(l) \rightleftharpoons \underset{\text{Ammonium ion}}{NH_4^+(aq)} + \underset{\text{Hydroxide ion}}{OH^-(aq)}$$

*This sign means **reversible reaction****

pH

Stands for **power of hydrogen**, a measure of hydrogen ion **concentration*** in a solution.

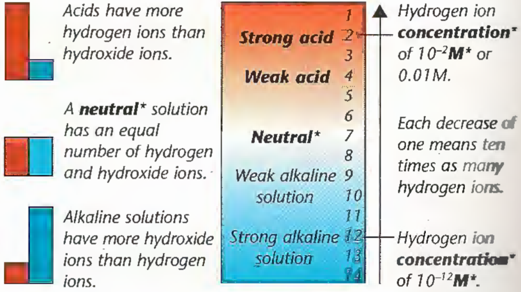
The pH scale
Acids have more hydrogen ions than hydroxide ions.
A **neutral*** solution has an equal number of hydrogen and hydroxide ions.
Alkaline solutions have more hydroxide ions than hydrogen ions.
1
Strong acid 2
3
Weak acid 4
5
6
Neutral* 7
8
Weak alkaline solution 9
10
11
Strong alkaline solution 12
13
14
Hydrogen ion **concentration*** of 10^{-2}**M*** or 0.01M.
Each decrease of one means ten times as many hydrogen ions.
Hydrogen ion **concentration*** of 10^{-12}**M***.

Indicator

A substance whose color depends on the **pH** of the solution it is in. Indicators can be used in solid or liquid form. Some common ones are shown at the bottom of this page.

Litmus

An **indicator** which shows whether a solution is acidic or alkaline. Acid turns blue litmus paper red, and alkaline solutions turn red litmus paper blue.

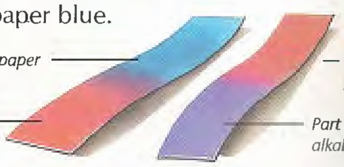
Blue litmus paper
Part dipped in acidic solution
Red
paper
Part dipped
alkaline

Universal indicator

An **indicator**, either in the form of paper or solution, which shows the **pH** of a solution with a range of colors.

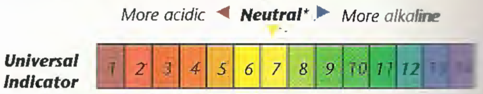
More acidic
Neutral*
More alkaline
Universal Indicator
1
2
3
4
5
6
7
8
9
10
11
12

Some other indicators

Methyl orange

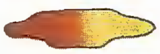

Red below 3, yellow above 4.5

Phenolphthalein

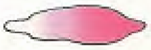

Colorless below 8.5, pink above 9.5

Bromoth

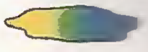

Yellow
blue

---

* **Concentration**, 139; **Hydronium ion**, 150; **Ionization**, 130; **M-value**, 139 (**Molarity**); **Mole**, 139; **Neutral**, 151; **Reversible reaction**, 162.

SALTS

All **salts** are **ionic**
contain at least o
(called the **acid ra**
can all be formed
of the hydrogen io
more other cations
below) or ammon
many industrial an

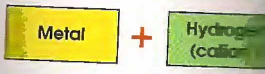
Metal
+
Hydroge
(catio

Acid radical

The **anion*** left after the
been removed from an

Acid	Radical
Hydrochloric	Cl^-
Sulfuric	SO_4^{2-}
Sulfurous	SO_3^{2-}
Nitric	NO_3^-
Nitrous	NO_2^-
Carbonic	CO_3^{2-}
Ethanoic	CH_3COO^-
Phosphoric	PO_4^{3-}

The radical name

Copper(II) sulfate

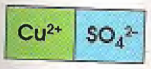
Cu^{2+}
SO_4^{2-}
Cation*
Acid radical

Basicity

The number of hydrogen
can be replaced to form a
hydrogen ions are necess

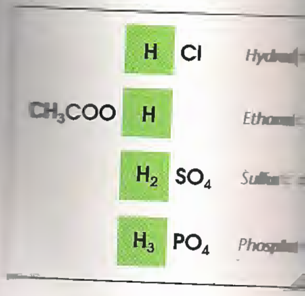
H Cl
Hydr
CH₃COO H
Etha
H₂ SO₄
Sulf
H₃ PO₄
Phosph

Cation, 130;
compound, 131.

SALTS

All **salts** are **ionic compounds*** which contain at least one **cation*** and one **anion*** (called the **acid radical**). Theoretically, they can all be formed by replacing one or more of the hydrogen ions in an acid by one or more other cations, e.g. metal ions (see below) or ammonium ions. Salts have many industrial and domestic uses.

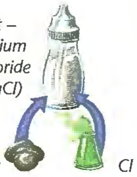

Salt – sodium chloride (NaCl)

Fertilizers – ammonium nitrate (NH_4NO_3)

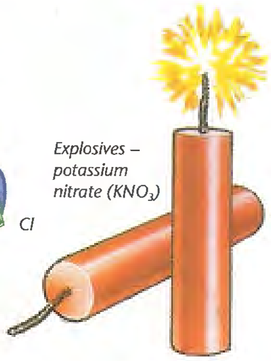

Explosives – potassium nitrate (KNO_3)

Metal + Hydrogen (cation*) Acid radical (anion*) → Metal (cation*) Acid radical (anion*) + Hydrogen

Acid *Salt*

Acid radical

The **anion*** left after the hydrogen ions have been removed from an acid. See table below.

Acid	*Radical*	*Radical name*
Hydrochloric	Cl^-	*Chloride*
Sulfuric	SO_4^{2-}	*Sulfate*
Sulfurous	SO_3^{2-}	*Sulfite*
Nitric	NO_3^-	*Nitrate*
Nitrous	NO_2^-	*Nitrite*
Carbonic	CO_3^{2-}	*Carbonate*
Ethanoic	CH_3COO^-	*Ethanoate*
Phosphoric	PO_4^{3-}	*Phosphate*

The radical name identifies the salt.

Copper(II) sulfate

Cu^{2+}	SO_4^{2-}
Cation*	***Acid radical***

Sodium chloride

Na^+	Cl^-
Cation*	***Acid radical***

Basicity

The number of hydrogen ions in an acid that can be replaced to form a salt. Not all the hydrogen ions are necessarily replaced.

	H	Cl	*Hydrochloric acid is **monobasic**.*
CH_3COO	H		*Ethanoic acid is **monobasic**.*
	H_2	SO_4	*Sulfuric acid is **dibasic**.*
	H_3	PO_4	*Phosphoric acid is **tribasic**.*

Normal salt

A salt containing only metal ions (or ammonium ions) and the **acid radical**, formed when all the hydrogen ions in an acid are replaced by metal ions (or ammonium ions).

Copper(II) sulfate and ammonium chloride (normal salts)

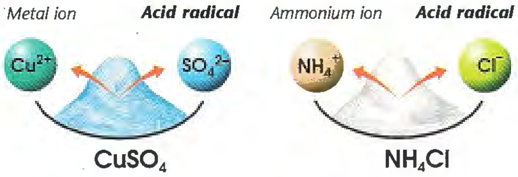

Acid salt

A salt containing hydrogen ions as well as metal ions (or ammonium ions) and the **acid radical**, formed when only some hydrogen ions in an acid are replaced by metal ions (or ammonium ions). Only acids with a **basicity** of two or more can form acid salts. Most acid salts are acidic, but some form alkaline solutions.

*Sodium hydrogensulfate (**acid salt**)*

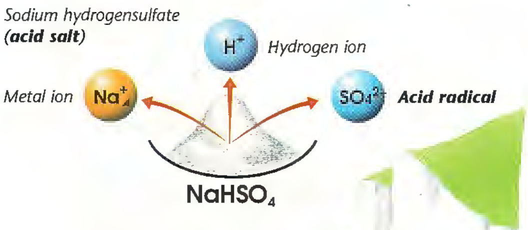

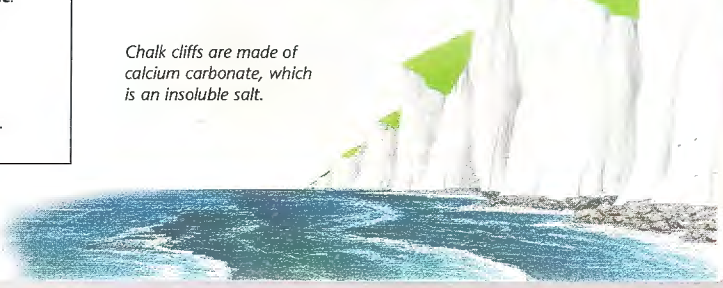

Chalk cliffs are made of calcium carbonate, which is an insoluble salt.

---

* **Anion**, **Cation**, 130;
Ionic compound, 131.

Salts (continued)

Basic salt

A salt containing a metal oxide or hydroxide, metal ions and an **acid radical***. It is formed when a **base*** is not completely **neutralized*** by an acid.

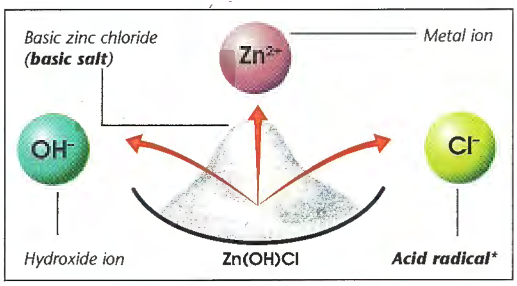
Basic zinc chloride
(basic salt)
Zn^{2+}
Metal ion
OH^-
Cl^-
Hydroxide ion
Zn(OH)Cl
Acid radical*

Double salt

A salt formed when solutions of two **normal salts*** react together. It contains two different **cations*** (either two different metal ions or a metal ion and an ammonium ion) and one or more **acid radicals***.

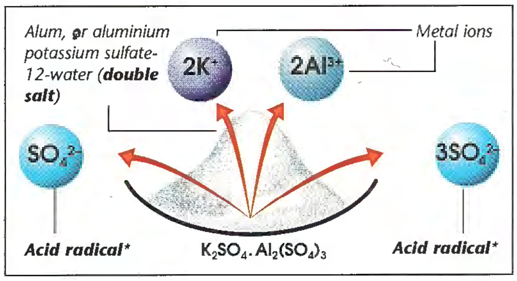
Alum, or aluminium potassium sulfate-12-water **(double salt)**
Metal ions
$2K^+$
$2Al^{3+}$
SO_4^{2-}
$3SO_4^{2-}$
Acid radical*
$K_2SO_4.Al_2(SO_4)_3$
Acid radical*

Complex salt

A salt in which one of the ions is a **complex ion**. This is made up of a central **cation*** linked (frequently by **dative covalent bonds***) to several small molecules (usually **polar molecules***) or ions.

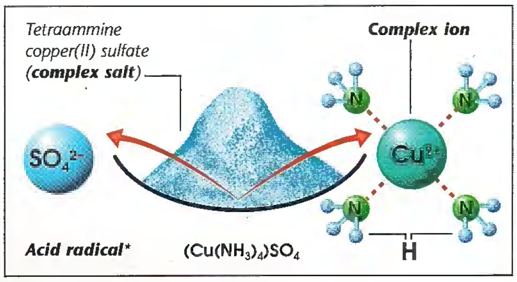
Tetraammine copper(II) sulfate **(complex salt)**
Complex ion
SO_4^{2-}
Cu^{2+}
N
N
N
N
H
Acid radical*
$(Cu(NH_3)_4)SO_4$

Anhydrate

A salt that does not contain **water of crystallization*** (it is **anhydrous**). The salt becomes a **hydrate** if it absorbs water.

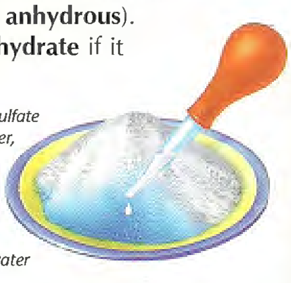

*When **anhydrous** copper(II) sulfate (a white powder) absorbs water, it becomes blue **hydrated** copper(II) sulfate crystals.*

*The water may be added (as here), but copper(II) sulfate is also **hygroscopic*** (absorbs water from the air).*

Hydrate

A salt that contains **water of crystallization*** (it is **hydrated**). The salt becomes an **anhydrate** if the water is removed.

$$\underset{\substack{\text{Anhydrous}\\\text{copper(II)}\\\text{sulfate}}}{CuSO_4} + 5H_2O \underset{\text{Water driven off by heating}}{\overset{\text{Water absorbed from the air}}{\rightleftharpoons}} \underset{\substack{\text{Hydrated}\\\text{copper(II)}\\\text{sulfate}}}{CuSO_4.5H_2O}$$

Dehydration

The removal of water from a substance. It is either removal of hydrogen and oxygen in the correct ratio to give water, or removal of water from a **hydrate** to give an **anhydrate**.

Hydrolysis

The chemical reaction of a compound with water to form another compound. When a salt reacts with water, the ions of the salt react with water molecules. This upsets the balance of hydrogen and hydroxide ions, and so gives an acidic or alkaline solution. A salt which has been made from the reaction between a **weak acid*** and a **strong base*** dissolves to give an alkaline solution. One which has been made from the reaction between a **strong acid*** and a **weak base*** dissolves to give an acidic solution.

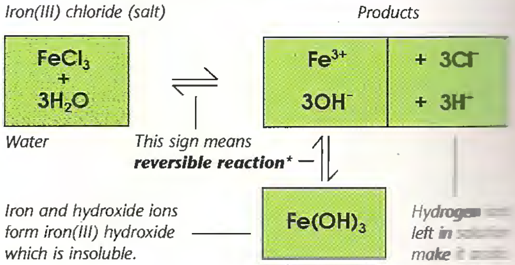
Iron(III) chloride (salt)
Products
$FeCl_3$ + $3H_2O$
Water
Fe^{3+} + $3Cl^-$
$3OH^-$ + $3H^-$
This sign means **reversible reaction***
Iron and hydroxide ions form iron(III) hydroxide which is insoluble.
$Fe(OH)_3$
Hydrogen
left in
make

---

* **Acid radical**, 153; **Base**, 151; **Cation**, 130; **Dative covalent bond**, 132; **Hygroscopic**, 206; **Neutralization**, 151; **Normal salt**, 153; **Polar molecule**, 133; **Reversible reaction**, 162; **Strong acid**, **Strong base**, 152; **Water of crystallization**, 135; **Weak acid**, **Weak base**, 152.

Salts can be made in a
the method depending
is soluble or insoluble in
below). Soluble salts are
solutions of the salts (ob
ways – see below) and
obtained in the form of

Solubility* of salts

Soluble salt
All ammonium, sodium and potassium salts
All nitrates
Chlorides ——— *EXCEPT*
Sulfates ——— *EXCE*
Ammonium Sodium Potassium ← *EXCEPT*

Soluble salts can be made
methods, which all produ
salt. This is partly evapora
crystallize*.

1. **Neutralization***, in whi
neutralized by an alkali.

ALKALI + ACID →

e.g. 2NaOH(aq) + H_2SO_4(aq) →
Sodium hydroxide *Sulfuric acid*

2. The action of an acid on
carbonate.

INSOLUBLE CARBONATE + ACID → SALT +

e.g. $MgCO_3$(s) + 2HCl(aq) → $MgCl_2$
Magnesium carbonate *Hydrochloric acid* *Magnes chlori*

3. The action of an acid on

INSOLUBLE BASE* + ACID →

e.g. CuO(s) + H_2SO_4(aq) →
Copper(II) oxide *Sulfuric acid*

130; **Base**, 151; **Cation**, 130;
Exothermic reaction, 146; **Ionic co**
Neutralization, 151; **Precipitate**, **Sol**

Preparation of salts

Salts can be made in a number of ways, the method depending on whether a salt is soluble or insoluble in water (see table below). Soluble salts are **crystallized*** from solutions of the salts (obtained in various ways – see below) and insoluble salts are obtained in the form of **precipitates***.

Solubility* of salts

Soluble salt	Insoluble salt
All ammonium, sodium and potassium salts	
All nitrates	
Chlorides —— *EXCEPT* →	*Silver and lead*
Sulfates —— *EXCEPT* →	*Barium and lead* *Calcium (slightly soluble)*
Ammonium *Sodium* ←—— *EXCEPT* — *Potassium*	*Most carbonates*

Soluble salts can be made by the following methods, which all produce a solution of the salt. This is partly evaporated and left to **crystallize***.

1. **Neutralization***, in which an acid is neutralized by an alkali.

ALKALI + ACID → SALT + WATER

e.g. $2NaOH(aq) + H_2SO_4(aq) \rightarrow Na_2SO_4(aq) + 2H_2O(l)$

Sodium hydroxide *Sulfuric acid* *Sodium sulfate* *Water*

2. The action of an acid on an insoluble carbonate.

INSOLUBLE CARBONATE + ACID → SALT + WATER + CARBON DIOXIDE

e.g. $MgCO_3(s) + 2HCl(aq) \rightarrow MgCl_2(aq) + H_2O(l) + CO_2(g)$

Magnesium carbonate *Hydrochloric acid* *Magnesium chloride* *Water* *Carbon dioxide*

3. The action of an acid on an insoluble **base***.

INSOLUBLE BASE* + ACID → SALT + WATER

e.g. $CuO(s) + H_2SO_4(aq) \rightarrow CuSO_4(aq) + H_2O(l)$

Copper(II) oxide *Sulfuric acid* *Copper(II) sulfate* *Water*

Double decomposition

A chemical reaction between the solutions of two or more **ionic compounds*** in which ions are exchanged. One of the new compounds formed is an insoluble salt, which forms a **precipitate***. Most insoluble salts and hydroxides are made by this method – the precipitate is filtered out and washed.

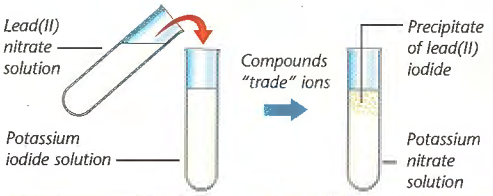
Lead(II) nitrate solution
Potassium iodide solution
Compounds "trade" ions
Precipitate of lead(II) iodide
Potassium nitrate solution

← **Anions*** *exchanged*

$Pb(NO_3)_2(aq) + 2KI(aq) \rightarrow PbI_2(\downarrow) + 2KNO_3(aq)$

← **Cations*** *exchanged*

Direct synthesis

A chemical reaction in which a salt is made directly from its elements. This method is used to make salts which react with water and therefore cannot be made by using solutions.

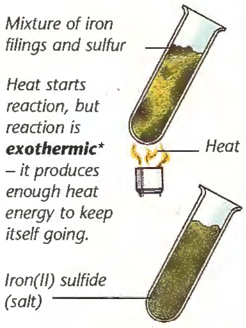
Mixture of iron filings and sulfur
Heat starts reaction, but reaction is **exothermic*** – it produces enough heat energy to keep itself going.
Heat
Iron(II) sulfide (salt)

$Fe(s) + S(s) \rightarrow FeS(s)$

Direct replacement

A reaction in which all or some of the hydrogen in an acid is replaced by another element, usually a metal. It is used to prepare soluble salts, except salts of sodium or potassium, both of which react too violently with the acid.

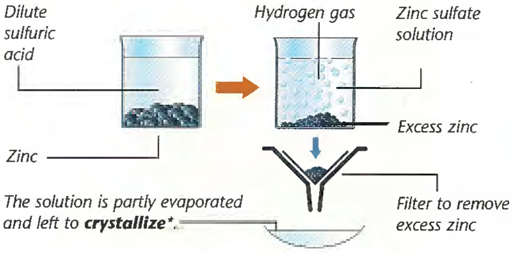
Dilute sulfuric acid
Zinc
Hydrogen gas
Zinc sulfate solution
Excess zinc
Filter to remove excess zinc

The solution is partly evaporated and left to ***crystallize****.

$Zn(s) + H_2SO_4(aq) \rightarrow ZnSO_4(aq) + H_2(g)$

---

* **Anion**, 130; **Base**, 151; **Cation**, 130; **Crystallization**, 135; **Exothermic reaction**, 146; **Ionic compound**, 131; **Neutralization**, 151; **Precipitate**, **Solubility**, 145.

ELECTROLYSIS

Electrolysis is a term describing the chemical changes which occur when an electric **current*** is passed through a liquid containing ions. Metals and graphite conduct electric current because some electrons are free to move through the **crystal lattice***, but **molten*** **ionic compounds*** or compounds which **ionize*** in solution conduct electric current by the movement of ions.

Electrolyte

A compound which conducts electricity when **molten*** or in **aqueous solution***, and decomposes during electrolysis. All **ionic compounds*** are electrolytes. They conduct electricity because when molten or in solution their ions are free to move. **Cations*** carry a positive charge and **anions*** a negative one. The number of ions in an electrolyte determines how well it conducts electricity.

Molten* *sodium chloride*

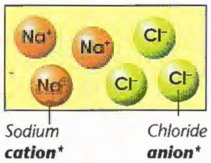
Na^+
Cl^-
Sodium **cation***
Chloride **anion***

Copper(II) sulfate solution (an ***aqueous solution****)*

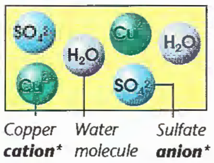
SO_4^{2-}
H_2O
Cu^{2+}
Copper **cation***
Water molecule
Sulfate **anion***

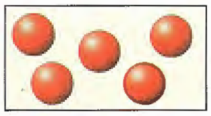

Non-electrolyte *– a compound which does not* ***ionize****.*

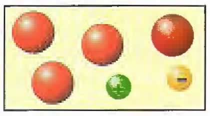

Weak electrolyte *– an* ***electrolyte*** *which is only partially* ***ionized****.*

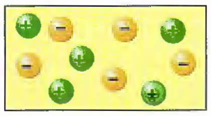

Strong electrolyte *– an* ***electrolyte*** *which is* ***ionized**** *completely.*

Electrode

A piece of metal or graphite placed in an **electrolyte** via which **current*** enters or leaves. There are two electrodes, the **anode** and **cathode**.

Inert electrode

An **electrode** that does not change during electrolysis, e.g. platinum. Some inert electrodes do react with the substances liberated.

Active electrode

An **electrode**, usually a metal, which undergoes chemical change during electrolysis.

Electrolytic cell

A vessel containing the **electrolyte** (either **molten*** or in **aqueous solution***) and the **electrodes.**

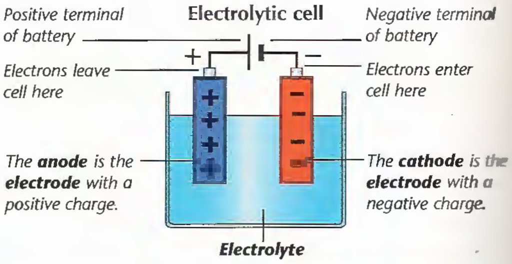
Electrolytic cell
Positive terminal of battery
Negative terminal of battery
Electrons leave cell here
Electrons enter cell here
+
−
The ***anode*** *is the* ***electrode*** *with a positive charge.*
The ***cathode*** *is the* ***electrode*** *with a negative charge.*
Electrolyte

Ionic theory of electrolysis

A theory which attempts to explain what happens in an **electrolytic cell** when it is connected to a supply of electricity. It states that **anions*** in the electrolyte are attracted to the **anode** (see **electrode**) where they lose electrons. The **cations*** are attracted to the **cathode** where they gain electrons. The ions which react at the electrodes are **discharged**. Electrons flow from the anode to the battery and from the battery to the cathode.

Electrolysis of molten* sodium chloride

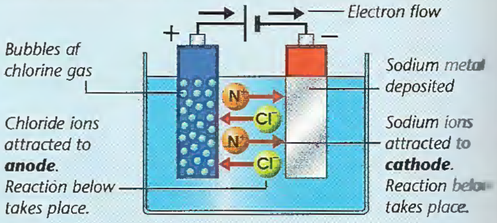
Electron flow
+
−
Bubbles of chlorine gas
Chloride ions attracted to **anode**.
Reaction below takes place.
Sodium metal deposited
Sodium ions attracted to **cathode**.
Reaction below takes place.
Na^+
Cl^-

Chloride ion loses an electron.
$Cl^- \rightarrow Cl + e^-$
Chlorine gas formed.
$Cl + Cl \rightarrow Cl_2$

Sodium ion gains an electron.
$Na^+ + e^- \rightarrow Na$
Sodium metal formed.

---

* **Anion**, 130; **Aqueous solution**, 144; **Cation**, 130; **Crystal lattice**, 136; **Current**, 159; **Ionic compound**, 131; **Ionization**, 130; **Molten**, 120.

Faraday's first law of
The mass of a substanc
chemical reactions at th
electrolysis is proportio
electricity passed throu

Amount of electricity = current ×

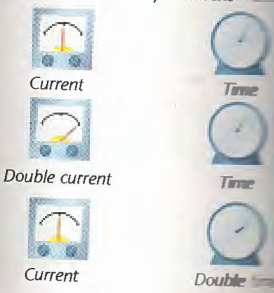
Current
Time
Double current
Time
Current
Double

Faraday's second law
When the same amount
passed through differen
number of **moles*** of eac
at the **electrodes** is inver
the size of the charge on

A copper ion must gain 2 electrons to form an atom.

$Cu^{2+} + 2e^- \rightarrow Cu$

If 1 ***faraday*** *(1* ***mole**** *of electrons) passes through copper(II) sulfate, two electrons are needed to turn each copper ion into an atom.*

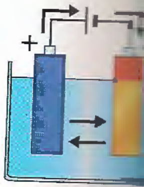

Other examples: *1 F produces one* *sodium atoms from sodium ions (Na^+).*

Voltameter or **coulome**
A type of **electrolytic cel**
the amount of a substanc
ctrolysis.

Coulomb (C)
The **SI unit*** of electric ch
of electricity passes a poi
ampere* flows for one sec

Faraday (F)
A unit of electric charge eq
coulombs. It consists of the
of electrons and therefore
of atoms from singly-charg

Ampere, 60; **Corrosion**, 209; **Mole**
Oxidation, 148; **Reactivity series**

Faraday's first law of electrolysis

The mass of a substance produced by chemical reactions at the **electrodes** during electrolysis is proportional to the amount of electricity passed through the **electrolyte**.

Amount of electricity = current × time

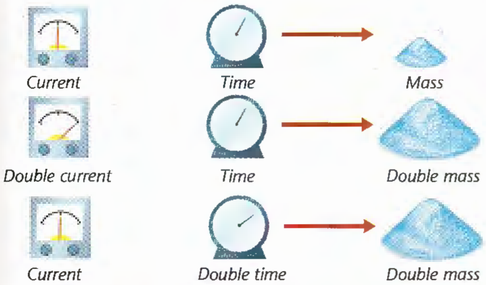
Current
Time
Mass
Double current
Time
Double mass
Current
Double time
Double mass

Faraday's second law of electrolysis

When the same amount of electricity is passed through different **electrolytes**, the number of **moles*** of each element deposited at the **electrodes** is inversely proportional to the size of the charge on its ion.

A copper ion must gain 2 electrons to form an atom.

$Cu^{2+} + 2e^- \rightarrow Cu$

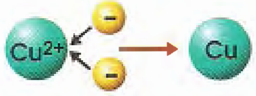

*If 1 **faraday** (1 **mole*** of electrons) passes through copper(II) sulfate, two electrons are needed to turn each copper ion into an atom.*

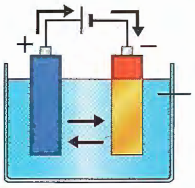

*Hence 1 **faraday** causes ½ **mole** of copper ions to be produced from atoms and deposited on the **cathode**. ½ is inversely proportional to 2 – the charge on a copper ion.*

Other examples:

1**F** produces one **mole*** sodium atoms from sodium ions (Na^+).	1**F** produces $^1/_3$ **mole*** aluminum atoms from its ions (Al^{3+}).

Voltameter or coulometer

A type of **electrolytic cell** used to measure the amount of a substance liberated during electrolysis.

Coulomb (C)

The **SI unit*** of electric charge. One coulomb of electricity passes a point when one **ampere*** flows for one second.

Faraday (F)

A unit of electric charge equal to 96,500 **coulombs**. It consists of the flow of one **mole*** of electrons and therefore liberates one mole of atoms from singly-charged ions.

Electrolysis in industry

Electro-refining

A method of purifying metals by electrolysis. Only the metal ions take part in electrolysis, the impurities are lost.

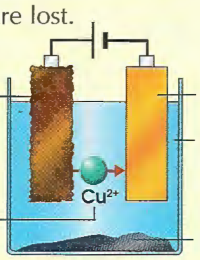
Impure copper **anode**
Pure copper **cathode**
Copper atoms give up electrons to form copper ions in the solution. These are attracted to the **cathode**.
Copper(II) sulfate solution
Cu2+
Impurities form a sludge.

Metal extraction

A process which produces metals from their **molten*** ores by electrolysis. Metals at the top of the **reactivity series*** are obtained in this way (see **aluminum**, page 176 and **sodium**, page 168).

Anodizing

The coating of a metal object with a thin layer of its oxide. Hydroxide ions are **oxidized*** at the metal **anode** in the electrolysis of dilute sulfuric acid, forming water and oxygen, which oxidizes the metal.

*These aluminum camping flasks have been **anodized** with aluminum oxide to prevent them from corroding.*

Electroplating

The coating of a metal object with a thin layer of another metal by electrolysis. The object forms the **cathode**, onto which metal ions in the **electrolyte** are deposited.

*This steel nail has been zinc plated to stop **corrosion*** (see **sacrificial protection**, page 159).*

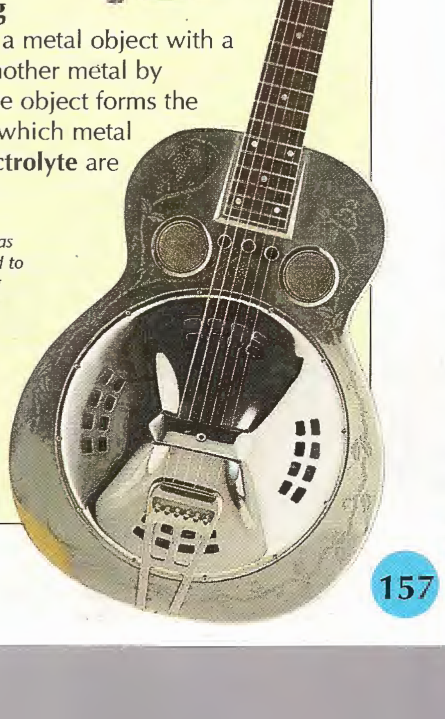

The metal front of this guitar has been plated with chrome by electrolysis.

---

* **Ampere**, 60; **Corrosion**, 209; **Mole**, 139; **Molten**, 120; **Oxidation**, 148; **Reactivity series**, 158; **SI units**, 96.

REACTIVITY

The **reactivity** of an element depends on its ability to gain or lose the electrons which are used for **bonding** (see pages 130-134). The more reactive an element, the more easily it will combine with others. Some elements are very reactive, others very unreactive. This difference can be used to produce electricity and protect metals from **corrosion***.

Displacement

A reaction in which one element replaces another in a compound. An element will only displace another lower than itself in the **reactivity series** (see right).

*Zinc **displaces** copper from copper(II) sulfate solution.*

$$CuSO_4(aq) + Zn(s) \rightarrow ZnSO_4(aq) + Cu(s)$$

Reactivity series or activity series

A list of elements (usually metals), placed in order of their reactivity. The series is constructed by comparing the reactions of the metals with other substances, e.g. acids and oxygen (for a summary of reactions, see page 211).

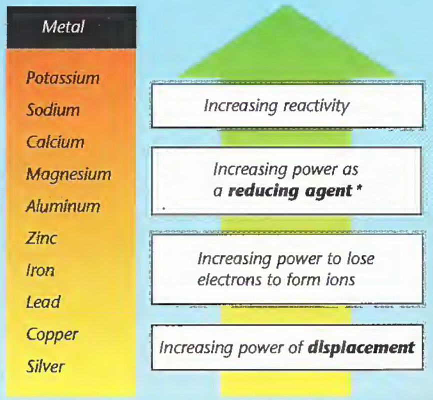
Metal
Potassium
Sodium
Calcium
Magnesium
Aluminum
Zinc
Iron
Lead
Copper
Silver
Increasing reactivity
Increasing power as a **reducing agent***
Increasing power to lose electrons to form ions
Increasing power of **displacement**

Half cell

An element in contact with water or an **aqueous solution*** of one of its compounds. Atoms on the surface form **cations***, which are released into the solution, leaving electrons behind. The solution has a positive charge and the metal a negative charge, so there is a **potential difference** between them.

Electrode potential (E)

The **potential difference** in a **half cell**. It is impossible to measure directly, so is measured relative to that of another half cell, normally a **hydrogen electrode** (see diagram). Electrode potentials show the ability to **ionize*** in **aqueous solution*** and are used to construct the **electrochemical series**.

Measuring the electrode potential of a metal

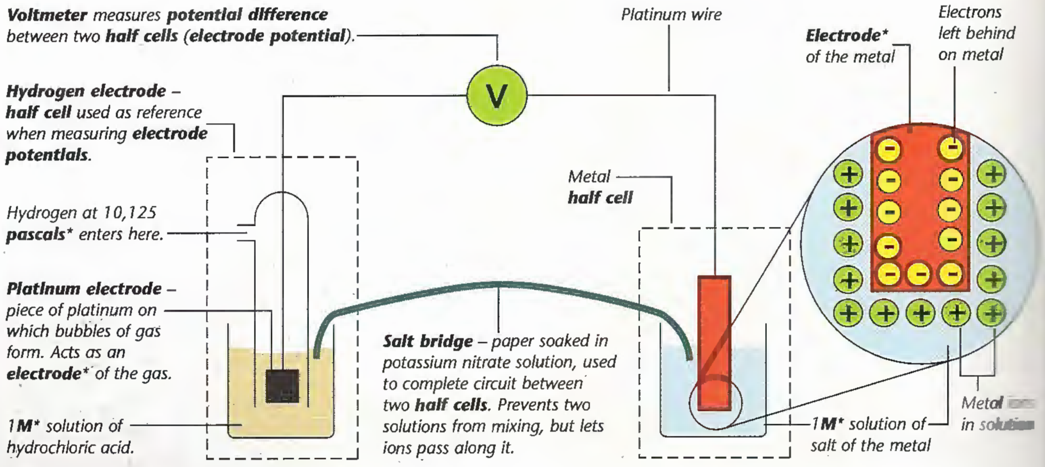
***Voltmeter** measures **potential difference** between two **half cells** (**electrode potential**).*
V
Platinum wire
Electrode* *of the metal*
Electrons left behind on metal
***Hydrogen electrode** – **half cell** used as reference when measuring **electrode potentials**.*
*Hydrogen at 10,125 **pascals*** enters here.*
*Metal **half cell***
Platinum electrode** – piece of platinum on which bubbles of gas form. Acts as an **electrode of the gas.*
***Salt bridge** – paper soaked in potassium nitrate solution, used to complete circuit between two **half cells**. Prevents two solutions from mixing, but lets ions pass along it.*
*1 **M*** solution of hydrochloric acid.*
*1 **M*** solution of salt of the metal*
Metal ions in solution

***Aqueous solution**, 144; **Cation**, 130; **Corrosion**, 209; **Electrode**, 156; **Ionization**, 130; **M-value**, 139 (**Molarity**); **Pascal**, 97; **Reducing agent**, 148.

Electrochemical serie

A list of the elements in
electrode potentials. Th
most negative electrode
at the top. The position
series shows how readil
aqueous solution*, and
of how reactive it is like

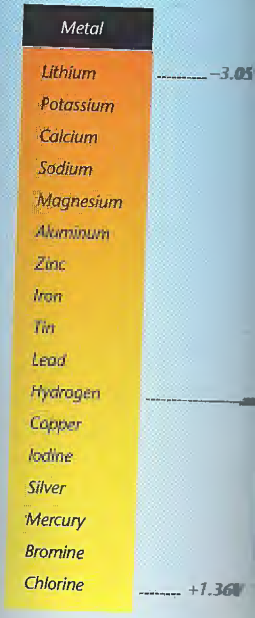
Metal
Lithium
Potassium
Calcium
Sodium
Magnesium
Aluminum
Zinc
Iron
Tin
Lead
Hydrogen
Copper
Iodine
Silver
Mercury
Bromine
Chlorine
–3.05
+1.36V

Cell or **electrochemical**

An arrangement of two **ha**
elements. The half cell wi
electrode potential forms
and the other forms the **po**
When these are connecte
between them. There are
primary cells, which canno
secondary cells, which ca
battery is a number of lin

Daniell cell (primary cell)

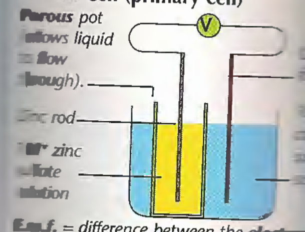
***Porous** pot allows liquid to flow through).*
V
Zinc rod
*1 **M*** zinc sulfate solution*

***E.m.f.** = difference between the **electro** **potentials** of zinc and copper = 1.*

Ampere, 60; **Aqueous solution**,
M-value, 139 (**Molarity**); **Rusting**

Electrochemical series

A list of the elements in order of their **electrode potentials**. The element with the most negative electrode potential is placed at the top. The position of an element in the series shows how readily it forms ions in **aqueous solution***, and is thus an indication of how reactive it is likely to be.

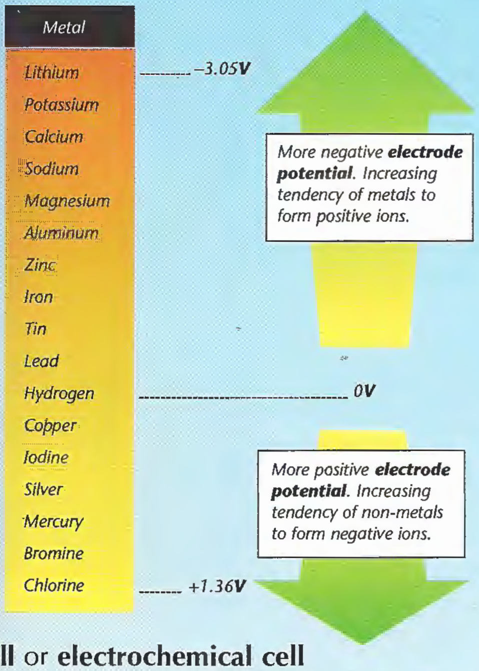
Metal
Lithium
Potassium
Calcium
Sodium
Magnesium
Aluminum
Zinc
Iron
Tin
Lead
Hydrogen
Copper
Iodine
Silver
Mercury
Bromine
Chlorine
−3.05V
0V
+1.36V

*More negative **electrode potential**. Increasing tendency of metals to form positive ions.*

*More positive **electrode potential**. Increasing tendency of non-metals to form negative ions.*

Cell or **electrochemical cell**

An arrangement of two **half cells** of different elements. The half cell with the most negative **electrode potential** forms the **negative terminal** and the other forms the **positive terminal**. When these are connected, a **current** flows between them. There are two types of cell – **primary cells**, which cannot be recharged, and **secondary cells**, which can be recharged. A **battery** is a number of linked cells.

Daniell cell (primary cell)

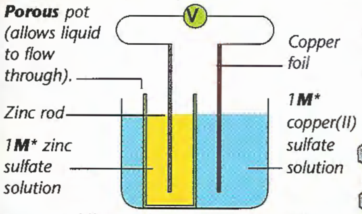
Porous pot (allows liquid to flow through).
V
Copper foil
Zinc rod
1**M*** zinc sulfate solution
1**M*** copper(II) sulfate solution

***E.m.f.** = difference between the **electrode potentials** of zinc and copper = 1.1V*

*A car battery consists of six **secondary cells** called **lead-acid accumulators***.*

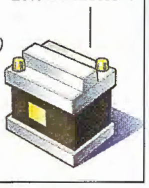

Potential difference or **voltage**

A difference in electric charge between two points, measured in **volts** (**V**) by an instrument called a **voltmeter**. If two points with a potential difference are joined, an electric **current**, proportional to the potential difference, flows between them.

Current

A flow of electrons (negatively-charged particles) through a material. The **SI unit*** of current is the **ampere*** (**A**), and current is measured using an **ammeter**. A current will flow in a loop, or **circuit**, between two points if there is a **potential difference** between them.

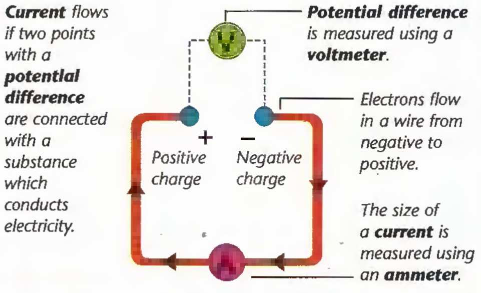
V
+
−
Positive charge
Negative charge
A

***Current** flows if two points with a **potential difference** are connected with a substance which conducts electricity.*

***Potential difference** is measured using a **voltmeter**.*

Electrons flow in a wire from negative to positive.

*The size of a **current** is measured using an **ammeter**.*

Electromotive force (**e.m.f.**)

The name given to the **potential difference** between the two terminals of a **cell** (i.e. the difference between the **electrode potentials** of the two **half cells**).

Sacrificial protection

Also known as **cathodic protection** or **electrical protection**. A method of preventing iron from **rusting*** by attaching a metal higher in the **electrochemical series** to it, which rusts instead.

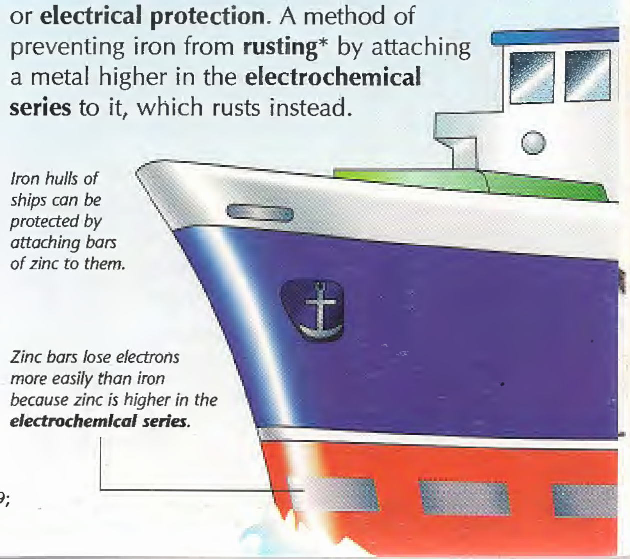

Iron hulls of ships can be protected by attaching bars of zinc to them.

*Zinc bars lose electrons more easily than iron because zinc is higher in the **electrochemical series**.*

---

* **Ampere**, 60; **Aqueous solution**, 144; **Lead-acid accumulator**, 69; **M-value**, 139 (**Molarity**); **Rusting**, 209 (**Corrosion**); **SI units**, 96.

RATES OF REACTION

The time it takes for a chemical reaction to finish varies from less than one millionth of a second to weeks or even years. It is possible to predict how long a particular reaction will take and how to speed it up or slow it down by altering the conditions under which it takes place. The efficiency of many industrial processes is improved by increasing the **rate of reaction**, e.g. by using high temperature and pressure, or a **catalyst**.

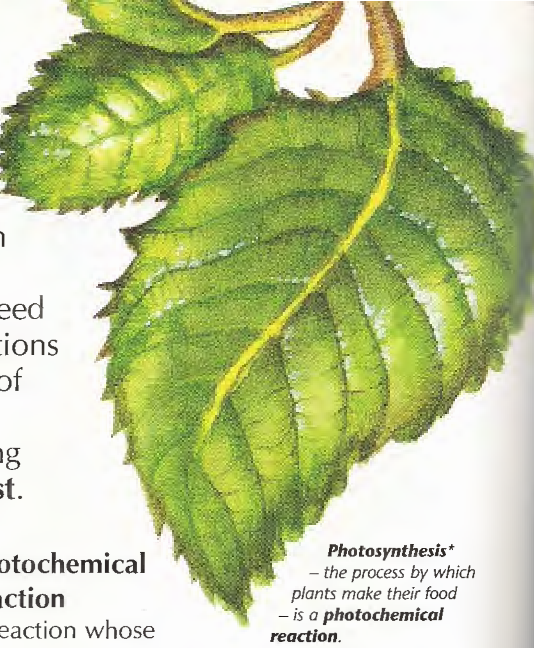

Photosynthesis* *– the process by which plants make their food – is a* ***photochemical reaction****.*

Rate of reaction

A measurement of the speed of a reaction. It is calculated by measuring how quickly reactants are used up or products are formed. The experimental method used to measure the rate of reaction depends on the **physical states*** of the reactants and products, and the data from such an experiment is plotted on a **rate curve**. The speed of a reaction varies as it proceeds. The rate at any time during the reaction is the **instantaneous rate**. The instantaneous rate at the start of the reaction is the **initial rate**. The **average rate** is calculated by dividing the total change in the amount of products or reactants by the time the reaction took to finish.

Rate curve – the mass of products with time

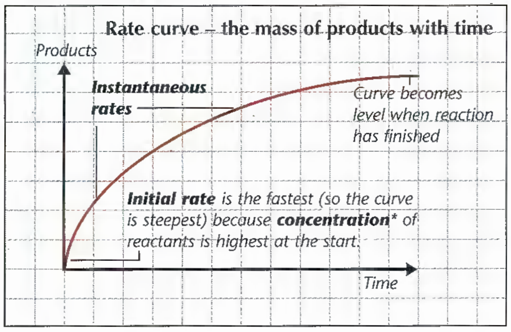

Collision theory

Explains why altering the conditions under which a reaction takes place affects its rate. For a reaction to take place between two particles, they must collide, so if more collisions occur, the **rate of reaction** increases. However, only some collisions cause a reaction, since not all particles have enough energy to react (see **activation energy**, right).

Photochemical reaction

A reaction whose speed is affected by the intensity of light, e.g. **photosynthesis***. Light gives reacting particles more energy and so increases the **rate of reaction**.

Photochemical reactions *occur in photography.*

Silver crystals form where light falls on the film, recording the picture.

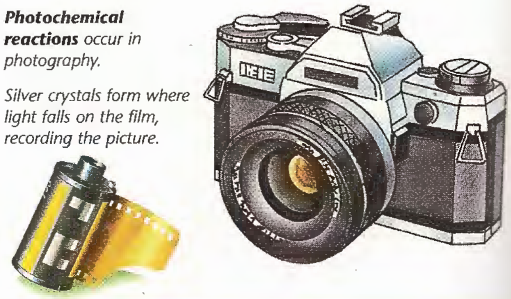

$$2AgCl(s) \rightarrow 2Ag(s) + Cl_2(g)$$

Activation energy (E)

The minimum energy that the particles of reactants must have for them to react when they collide (see **collision theory**). The **rate of reaction** depends on how many reacting particles have this minimum energy. In many reactions, the particles already have this energy and react immediately. In others, energy has to be supplied for the particles to reach the activation energy.

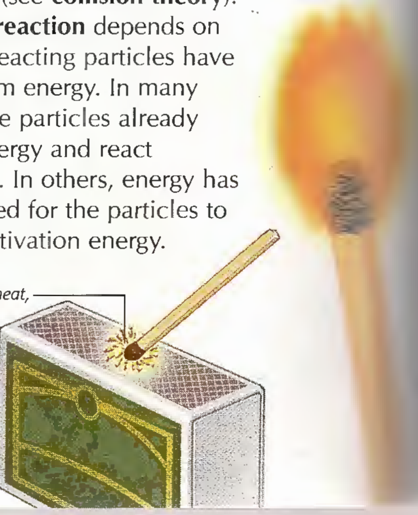

Friction produces heat, giving ***activation energy*** *to the particles in the match.*

---

* **Concentration**, 139; **Physical states**, 120; **Photosynthesis**, 209.

Changing rates of re
The **rate of reaction**
temperature is increas
gives more particles
the **activation energy**.

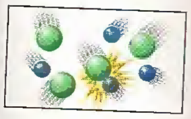

For reactions involvin
reaction will increase
increased. An increase
a gas increases the tem
decreases the volume
concentration* – see
142). The particles col
and with greater energ

Catalyst

A substance that increase
a chemical reaction, but
nchanged itself at the
This process is known as
work by lowering the acti
eaction. The catalyst use
written over the arrow
page 182). A catalyst wh
of one reaction may have

Decomposition of hydrogen

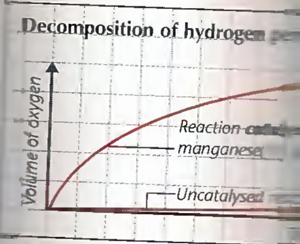

atalytic converter in
r's exhaust system
ains two metals,
num and rhodium,
ch act as ***catalysts****.*

carbon monoxide
ydrocarbons
to the metals and
together to form
dioxide and

ocatalysis

rocess in which one of
ction acts as a **catalyst**

ncentration, 139; **Physical sta**

Changing rates of reaction

The **rate of reaction** will increase if the temperature is increased. The heat energy gives more particles an energy greater than the **activation energy**.

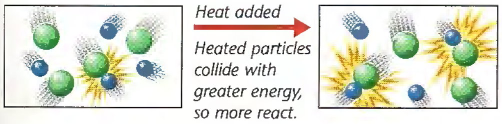

Heated particles collide with greater energy, so more react.

For reactions involving gases, the **rate of reaction** will increase if the pressure is increased. An increase in the pressure of a gas increases the temperature and decreases the volume (i.e. increases the **concentration*** – see also **gas laws**, page 142). The particles collide more often and with greater energy.

The **rate of reaction** will increase if the **concentration*** of one or more of the reactants is increased.

More molecules in the same space means more collisions.

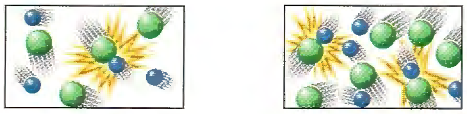

*Low **concentration****

*High **concentration****

The **rate of reaction** will increase if the surface area of a solid reactant is increased. Reactions in which one reactant is a solid can only take place at the surface of the solid.

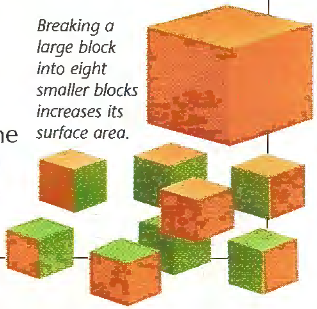

Breaking a large block into eight smaller blocks increases its surface area.

Catalyst

A substance that increases the rate of a chemical reaction, but is chemically unchanged itself at the end of the reaction. This process is known as **catalysis**. Catalysts work by lowering the **activation energy** of a reaction. The catalyst used in a reaction is written over the arrow in the equation (see page 182). A catalyst which increases the rate of one reaction may have no effect on another.

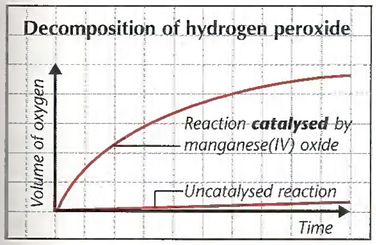

Hydrogen peroxide decomposes to form oxygen and water.

*The reaction speeds up when a **catalyst** is used.*

*A catalytic converter in a car's exhaust system contains two metals, platinum and rhodium, which act as **catalysts**.*

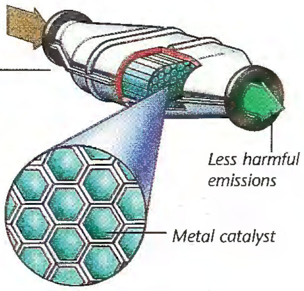

Toxic carbon monoxide and hydrocarbons cling to the metals and react together to form carbon dioxide and water.

Autocatalysis

A process in which one of the products of a reaction acts as a **catalyst** for the reaction.

Surface catalyst

A **catalyst** which attracts the reactants to itself. It holds them close to each other on its surface, so they react easily.

Homogenous catalyst

A **catalyst** in the same **physical state*** as the reactants.

Heterogenous catalyst

A **catalyst** in a different **physical state*** from that of the reactants.

Promoter

A substance which increases the power of a **catalyst**, so speeding up the reaction.

Inhibitor

A substance that slows a reaction. Some work by reducing the power of a **catalyst**.

Enzyme

A **catalyst** found in living things which increases the **rate of reaction** in a natural chemical process. For more about enzymes, see page 333.

*Spiders feed by secreting **enzymes** onto their prey. The enzymes speed up chemical reactions that break down the food.*

***Concentration**, 139; **Physical states**, 120.

REVERSIBLE REACTIONS

Many chemical reactions continue until one or all of the reactants are used up, and their products do not react together. When a reaction reaches this stage, it is said to have come to **completion**. Other reactions, however, never reach this stage. They are known as **reversible reactions**.

Reversible reaction

A chemical reaction in which the products react together to form the original reactants. These react again to form the products, and so on. The two reactions are simultaneous, and the process will not come to **completion** (see introduction) if it takes place in a **closed system**. At some stage during a reversible reaction, **chemical equilibrium** is reached.

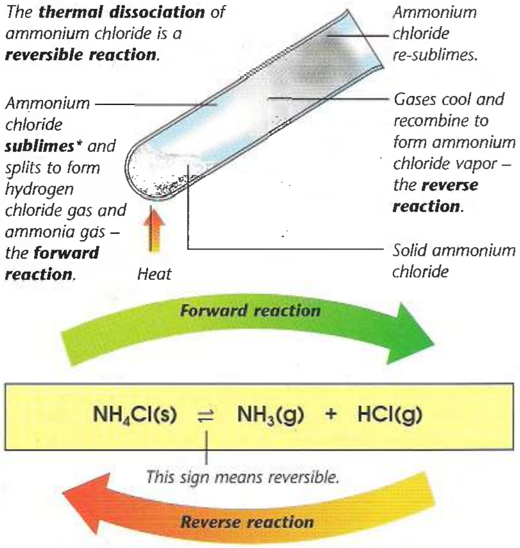
*The **thermal dissociation** of ammonium chloride is a **reversible reaction**.*

Ammonium chloride re-sublimes.
Ammonium chloride sublimes* and splits to form hydrogen chloride gas and ammonia gas – the forward reaction.
Gases cool and recombine to form ammonium chloride vapor – the reverse reaction.
Solid ammonium chloride
Heat
Forward reaction
Reverse reaction

$NH_4Cl(s) \rightleftharpoons NH_3(g) + HCl(g)$

This sign means reversible.

Forward reaction

The reaction in which products are formed from the original reactants in a **reversible reaction**. It goes from left to right in the equation.

Reverse reaction or backward reaction

The reaction in which the original reactants are reformed from their products in a **reversible reaction**. It goes from right to left in the equation.

*This tug of war represents a chemical reaction. The blue team are the reactants and the red team the products. If the red team wins, the reaction is a **forward reaction**. If the blue team wins, it is a **reverse reaction**.*

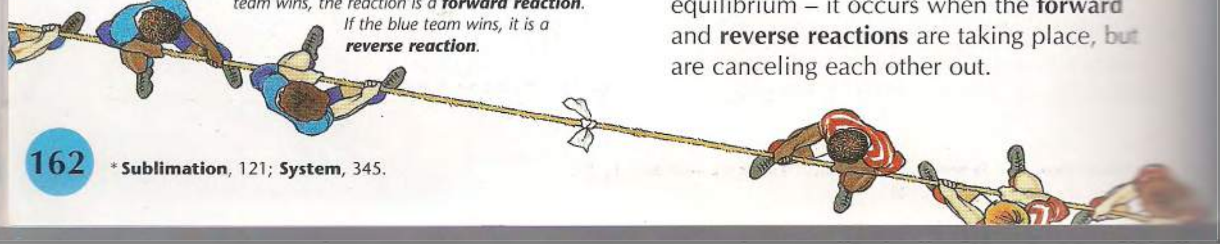

Dissociation

A type of **reversible reaction** in which a compound is divided into other compounds or elements. **Thermal dissociation** is dissociation caused by heating (the products formed recombine when cooled). Dissociation should not be confused with **decomposition**, in which a compound is irreversibly split up.

*Nitrogen dioxide undergoes **thermal dissociation** into nitrogen monoxide and oxygen.*

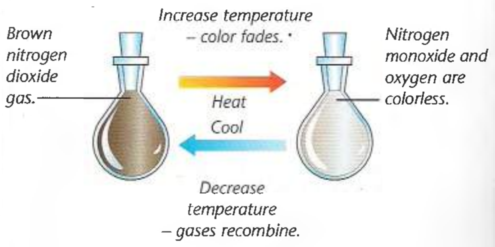
Brown nitrogen dioxide gas.
Increase temperature – color fades.
Heat
Cool
Decrease temperature – gases recombine.
Nitrogen monoxide and oxygen are colorless.

$2NO_2(g) \rightleftharpoons 2NO(g) + O_2(g)$

Closed system

A **system*** in which no chemicals can escape or enter. If a product of a **reversible reaction** escapes, for example into the atmosphere, the reaction can no longer move back the other way. A system from which chemicals can escape is an **open system**.

Equilibrium

The canceling out of two equal but opposite movements. For instance, a person walking up an escalator at the same speed as the escalator is moving down is in equilibrium. **Chemical equilibrium** is an example of equilibrium – it occurs when the **forward** and **reverse reactions** are taking place, but are canceling each other out.

***Sublimation**, 121; **System**, 345.

Chemical equilibriu

A stage reached in a re
in a **closed system** whe
reverse reactions take p
Their effects cancel eac
and the **concentrations***
and products no longer
equilibrium is a form of

The position o equilibrium

Any change of conditio
concentration* or press
reversible reaction alter
the **forward** or **reverse** r
the **chemical equilibriu**
restored, but with a diffe
reactants and products. T
position is said to have c

*First **equilibrium position***

*Alter conditions to favor **forward reaction** – **equilibrium position** is said to move right.*

*Alter conditions to favor **reverse reaction** – **equilibrium position** is said to move left.*

Le Chatelier's principle

A law stating that if chang
system* in **equilibrium**, th
itself to reduce the effects

1. Changing the pressure i
reactions involving gases
equilibrium position.

[illegible] reaction **A(g) + B(g) ⇌ AB(g**

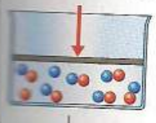
Molecule of A
Molecule of B
Molecule of AB

*[illegible] **pressure** – position moves [illegible] – more AB formed – i.e. [illegible]ber of molecules decreases [illegible] pressure again.*

Lowe left – num to r

[illegible]tration, 139; **Endothermic**
[illegible]ber process, 180; **System**, 345.

Chemical equilibrium

A stage reached in a **reversible reaction** in a **closed system** when the **forward** and **reverse reactions** take place at the same rate. Their effects cancel each other out, and the **concentrations*** of the reactants and products no longer change. Chemical equilibrium is a form of **equilibrium**.

*At start of reaction, higher **concentration*** of products than reactants.*

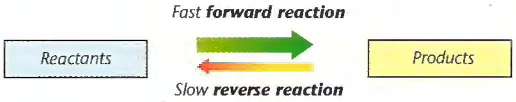

*At **chemical equilibrium**, products and reactants formed at same rate.*

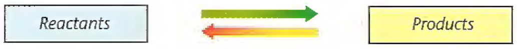

The position of chemical equilibrium

Any change of conditions (temperature, **concentration*** or pressure) during a **reversible reaction** alters the rate of either the **forward** or **reverse reaction**, destroying the **chemical equilibrium**. This is eventually restored, but with a different proportion of reactants and products. The **equilibrium position** is said to have changed.

*First **equilibrium position***

*Alter conditions to favor **forward reaction** – **equilibrium position** is said to move right.*

*Alter conditions to favor **reverse reaction** – **equilibrium position** is said to move left.*

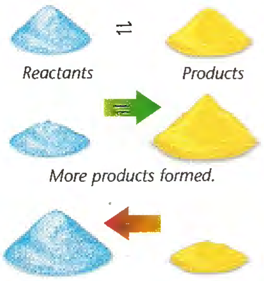

More products formed.

More reactants formed.

Le Chatelier's principle

A law stating that if changes are made to a **system*** in **equilibrium**, the system adjusts itself to reduce the effects of the change.

1. Changing the pressure in **reversible reactions** involving gases may alter the **equilibrium position**.

In the reaction $A(g) + B(g) \rightleftharpoons AB(g)$:

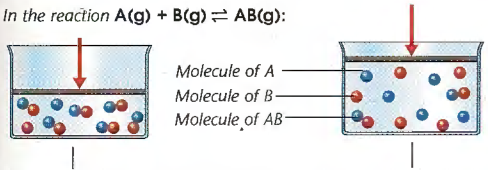

Raise pressure – position moves right – more AB formed – i.e. number of molecules decreases to lower pressure again.

Lower pressure – position moves left – more A and B formed – i.e. number of molecules increases to raise pressure again.

2. Changing the temperature in a **reversible reaction** also alters the **equilibrium position**. This depends on whether the reaction is **exothermic*** or **endothermic***. A reversible reaction which is exothermic in one direction is endothermic in the other.

*Ammonia is made by the **Haber process***.*

Exothermic*

$$N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g)$$

Nitrogen Hydrogen Ammonia

Endothermic*

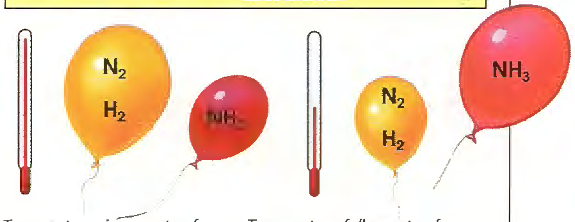

*Temperature rises – rate of **endothermic* reverse reaction** increases to absorb heat. Less ammonia formed – position moves left.*

*Temperature falls – rate of **exothermic* forward reaction** increases, giving out more heat energy. More ammonia produced – position moves right.*

3. Changing the **concentration*** of the reactants or products in a **reversible reaction** also changes the equilibrium position.

*Raise **concentration*** of reactants – rate of **forward reaction** increased.* OR *Lower **concentration** of products – rate of **reverse reaction** decreased.*

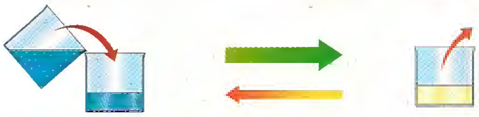

***Equilibrium position** moves right.*

*Lower **concentration** of reactants – rate of **forward reaction** decreased.* OR *Raise **concentration** of products – rate of **reverse reaction** increased.*

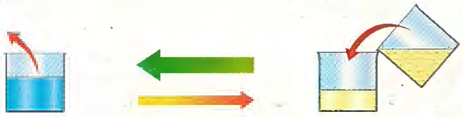

***Equilibrium position** moves left.*

---

* **Concentration**, 139; **Endothermic reaction**, **Exothermic reaction**, 146; **Haber process**, 180; **System**, 345.

THE PERIODIC TABLE

During the nineteenth century, many chemists tried to arrange the elements in an order which related to the size of their atoms and also showed regular repeating patterns in their behavior or properties. The most successful attempt was published by the Russian, Dimitri Mendeléev, in 1869, and still forms the basis of the modern **periodic table**.

Periodic table

An arrangement of the elements in order of their **atomic numbers***. Both the physical properties and chemical properties of an element and its compounds are related to the position of the element in the periodic table. This relationship has led to the table being divided into **groups** and **periods**. The arrangement of the elements starts on the left of period 1 with hydrogen and moves in order of increasing atomic number from left to right across each period in turn (see picture on the right).

Period

A horizontal row of elements in the **periodic table**. There are seven periods in all. Period 1 has only two elements – hydrogen and helium. Periods 2 and 3 each contain eight elements and are called the **short periods**. Periods 4, 5, 6 and 7 each contain between 18 and 32 elements. They are called the **long periods**. Moving from left to right across a period, the **atomic number*** increases by one from one element to the next. Each successive element has one more electron in the **outer shell*** of its atoms. All elements in the same period have the same number of shells, and the regular change in the number of electrons from one element to the next leads to a fairly regular pattern of change in the chemical properties of the elements across a period. For an example of such a gradual change in property, see below.

Periodic table

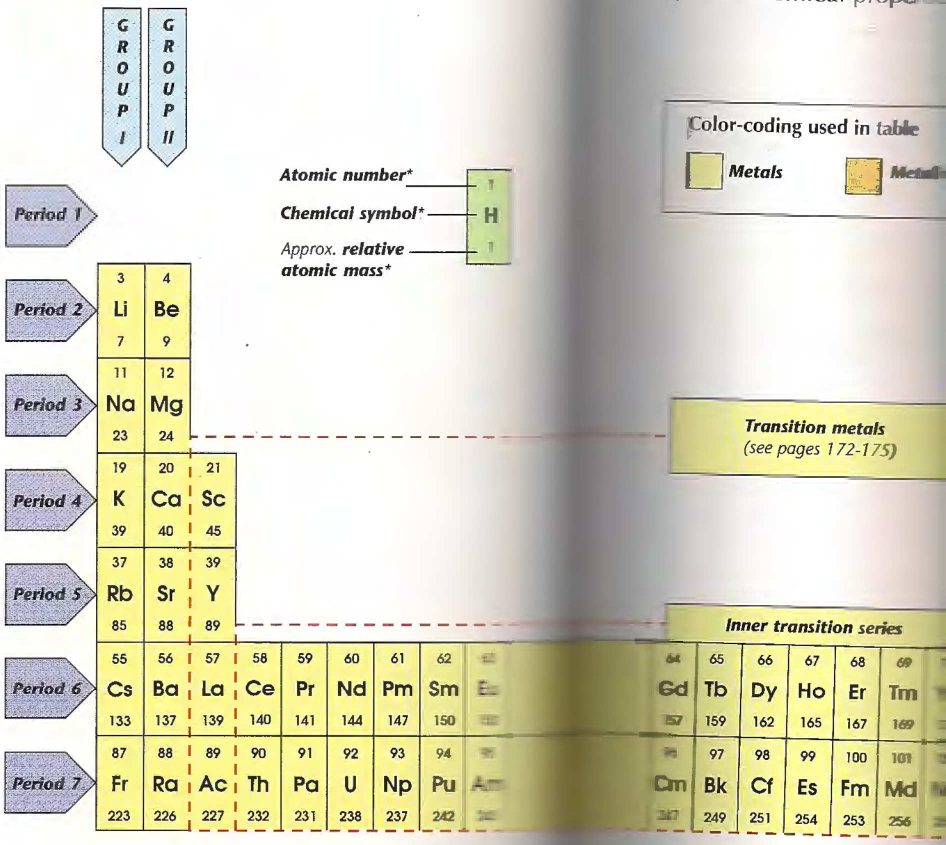
GROUP I
GROUP II
Period 1
Period 2
Period 3
Period 4
Period 5
Period 6
Period 7
Atomic number*
Chemical symbol*
Approx. relative atomic mass*
1 H 1
3 Li 7
4 Be 9
11 Na 23
12 Mg 24
19 K 39
20 Ca 40
21 Sc 45
37 Rb 85
38 Sr 88
39 Y 89
55 Cs 133
56 Ba 137
57 La 139
58 Ce 140
59 Pr 141
60 Nd 144
61 Pm 147
62 Sm 150
87 Fr 223
88 Ra 226
89 Ac 227
90 Th 232
91 Pa 231
92 U 238
93 Np 237
94 Pu 242
Transition metals
(see pages 172-175)
Inner transition series
64 Gd 157
65 Tb 159
66 Dy 162
67 Ho 165
68 Er 167
69 Tm 169
Cm 247
97 Bk 249
98 Cf 251
99 Es 254
100 Fm 253
101 Md 256
Color-coding used in table
Metals

Electron configuration* of elements across Period 2

All elements have the same ***outer shell****, but each successive element, going from left to right, has one more electron added to that she*

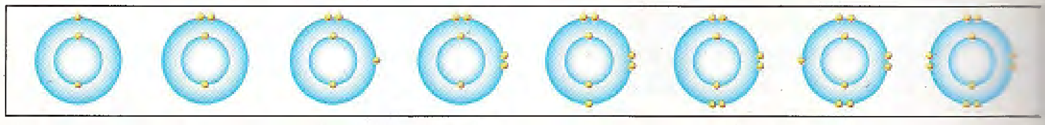

Strong ***reducing agents**** ——→ *Weak* ***reducing agents*** ——→ *Strong* ***oxidizing agents**** ——→

This shows a regular pattern of change across ***Period 2*** *in the ability of elements to* ***reduce**** *or* ***oxidize**** *other elements and compou*
(see also page 166). Neon is the exception – it is unreactive.

Group

A vertical column of ele
table. All groups are n
transition metal* grou
numerals, and some ha
in the same group have
electrons in their **outer** s
similar chemical prope

Metals and non-

Metal

An element with charac
properties that distinguis
non-metal. Elements on the
have metallic properties.
right, the elements gradu
metallic. Elements that are
metal or non-metal, but ha
properties, are called **metal**
to the right of metalloids a

---

* **Atomic number**, 127; **Chemical symbol**, 122; **Electron configuration**, **Outer shell**, 127; **Oxidation**, **Oxidizing agent**, **Reducing agent**, **Reduction**, 148; **Relative atomic mass**, 138.

ctivity, 28, 63; **Ductility**,
er shell, 127; **Physical states**

Group

A vertical column of elements in the **periodic table**. All groups are numbered (except for **transition metal*** groups) using Roman numerals, and some have names. Elements in the same group have the same number of electrons in their **outer shell***, and so have similar chemical properties.

Groups with alternative names

Group number	***Group name***
Group I	*The **alkali metals** (see pages 168-169)*
Group II	*The **alkaline-earth metals** (see pages 170-171)*
Group VII	*The **halogens** (see pages 186-188)*
Group VIII *(or **Group 0**)*	*The **noble gases** (see page 189)*

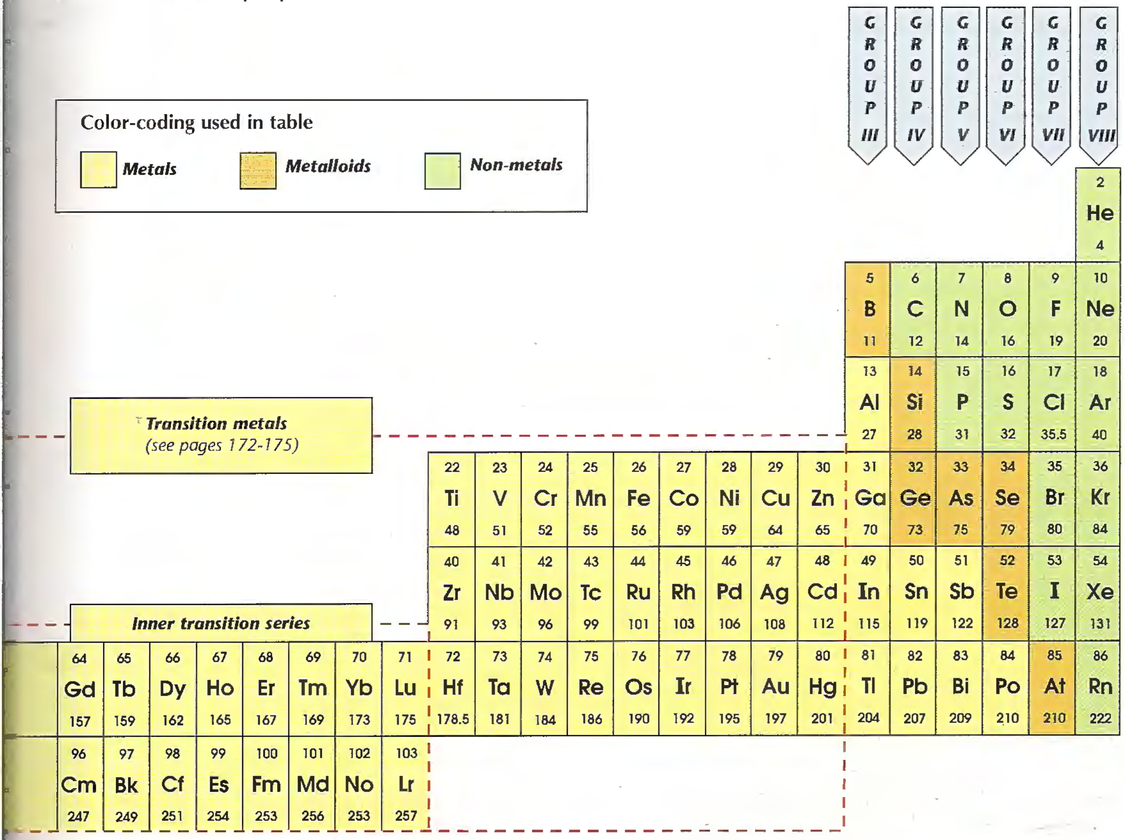

Color-coding used in table

Metals ***Metalloids*** ***Non-metals***

Transition metals
(see pages 172-175)

Inner transition series

																	GROUP III	***GROUP IV***	***GROUP V***	***GROUP VI***	***GROUP VII***	***GROUP VIII***
																						2 He 4
																	5 B 11	6 C 12	7 N 14	8 O 16	9 F 19	10 Ne 20
																	13 Al 27	14 Si 28	15 P 31	16 S 32	17 Cl 35.5	18 Ar 40
								22 Ti 48	23 V 51	24 Cr 52	25 Mn 55	26 Fe 56	27 Co 59	28 Ni 59	29 Cu 64	30 Zn 65	31 Ga 70	32 Ge 73	33 As 75	34 Se 79	35 Br 80	36 Kr 84
								40 Zr 91	41 Nb 93	42 Mo 96	43 Tc 99	44 Ru 101	45 Rh 103	46 Pd 106	47 Ag 108	48 Cd 112	49 In 115	50 Sn 119	51 Sb 122	52 Te 128	53 I 127	54 Xe 131
64 Gd 157	65 Tb 159	66 Dy 162	67 Ho 165	68 Er 167	69 Tm 169	70 Yb 173	71 Lu 175	72 Hf 178.5	73 Ta 181	74 W 184	75 Re 186	76 Os 190	77 Ir 192	78 Pt 195	79 Au 197	80 Hg 201	81 Tl 204	82 Pb 207	83 Bi 209	84 Po 210	85 At 210	86 Rn 222
96 Cm 247	97 Bk 249	98 Cf 251	99 Es 254	100 Fm 253	101 Md 256	102 No 253	103 Lr 257															

Metals and non-metals

Metal

An element with characteristic physical properties that distinguish it from a **non-metal**. Elements on the left of a **period** have metallic properties. Moving to the right, the elements gradually become less metallic. Elements that are not distinctly metal or non-metal, but have a mixture of properties, are called **metalloids**. Elements to the right of metalloids are non-metals.

Property	Metal	Non-metal
Physical state*	*Solids (except mercury)*	*Solid, liquid or gas (bromine is the only liquid).*
Appearance	*Shiny*	*Mainly non-shiny (iodine is one of the exceptions).*
Conductivity*	*Good*	*Poor (except graphite)*
Malleability*	*Good*	*Poor*
Ductility*	*Good*	*Poor*
Melting point	*Generally high*	*Generally low (except carbon)*
Boiling point	*Generally high*	*Generally low*

* **Conductivity**, 28, 63; **Ductility**, 344 (**Ductile**); **Malleability**, 345 (**Malleable**); **Outer shell**, 127; **Physical states**, 120; **Transition metals**, 172.

INORGANIC CHEMISTRY

Inorganic chemistry is the study of all the elements and their compounds except those compounds made of chains of carbon atoms (see **organic chemistry**, pages 190-205). The properties and reactions of inorganic elements and compounds follow certain patterns, or **trends**, in the **periodic table***. By looking up and down **groups*** and across **periods*** of the table, it is possible to predict the reactions of elements.

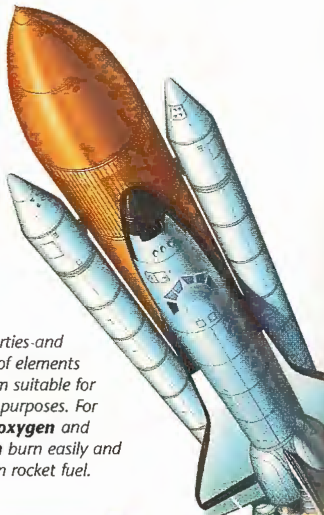

*The properties and reactions of elements make them suitable for particular purposes. For example, **oxygen** and **hydrogen** burn easily and are used in rocket fuel.*

Major periodic table trends

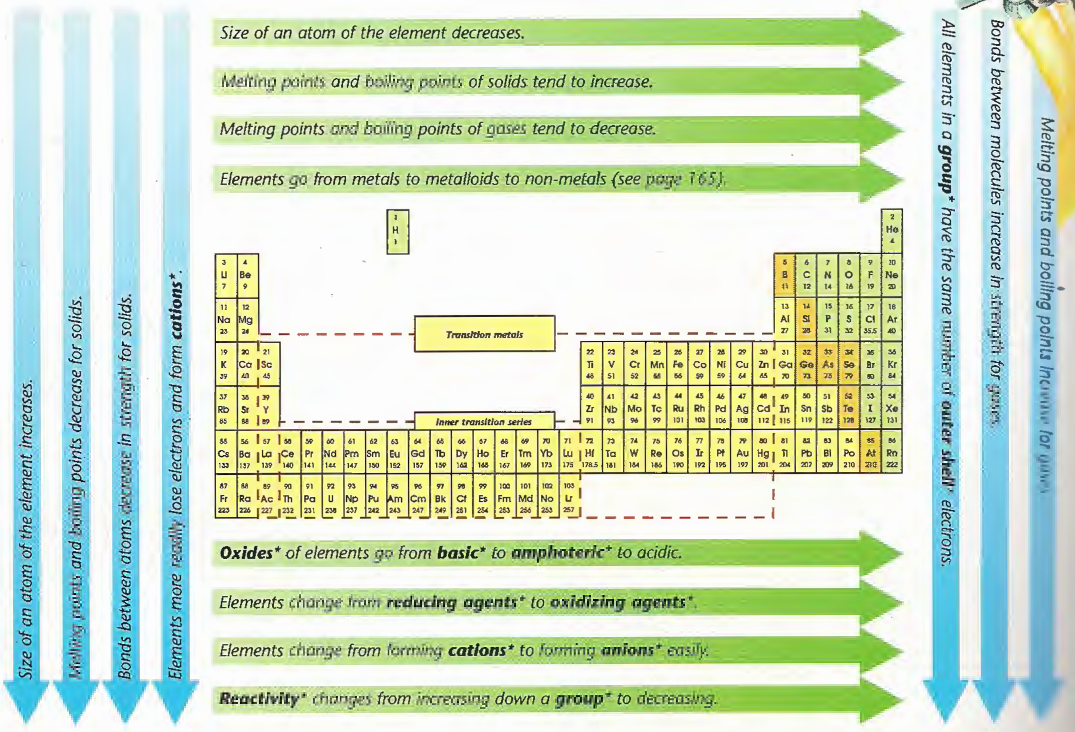

Size of an atom of the element increases.

Melting points and boiling points decrease for solids.

Bonds between atoms decrease in strength for solids.

*Elements more readily lose electrons and form **cations***.*

Size of an atom of the element decreases.

Melting points and boiling points of solids tend to increase.

Melting points and boiling points of gases tend to decrease.

Elements go from metals to metalloids to non-metals (see page 165).

__Oxides__ of elements go from __basic__* to __amphoteric__* to acidic.*

Elements change from __reducing agents__ to __oxidizing agents__*.*

Elements change from forming __cations__ to forming __anions__* easily.*

__Reactivity__ changes from increasing down a __group__* to decreasing.*

*All elements in a **group*** have the same number of **outer shell*** electrons.*

Bonds between molecules increase in strength for gases.

Melting points and boiling points increase for gases.

1 H 1
2 He 4
3 Li 7
4 Be 9
5 B 11
6 C 12
7 N 14
8 O 16
9 F 19
10 Ne 20
11 Na 23
12 Mg 24
13 Al 27
14 Si 28
15 P 31
16 S 32
17 Cl 35.5
18 Ar 40
19 K 39
20 Ca 40
21 Sc 45
22 Ti 48
23 V 51
24 Cr 52
25 Mn 55
26 Fe 56
27 Co 59
28 Ni 59
29 Cu 64
30 Zn 65
31 Ga 70
32 Ge 73
33 As 75
34 Se 79
35 Br 80
36 Kr 84
37 Rb 85
38 Sr 88
39 Y 89
40 Zr 91
41 Nb 93
42 Mo 96
43 Tc 99
44 Ru 101
45 Rh 103
46 Pd 106
47 Ag 108
48 Cd 112
49 In 115
50 Sn 119
51 Sb 122
52 Te 128
53 I 127
54 Xe 131
55 Cs 133
56 Ba 137
57 La 139
58 Ce 140
59 Pr 141
60 Nd 144
61 Pm 147
62 Sm 150
63 Eu 152
64 Gd 157
65 Tb 159
66 Dy 162
67 Ho 165
68 Er 167
69 Tm 169
70 Yb 173
71 Lu 175
72 Hf 178.5
73 Ta 181
74 W 184
75 Re 186
76 Os 190
77 Ir 192
78 Pt 195
79 Au 197
80 Hg 201
81 Tl 204
82 Pb 207
83 Bi 209
84 Po 210
85 At 210
86 Rn 222
87 Fr 223
88 Ra 226
89 Ac 227
90 Th 232
91 Pa 231
92 U 238
93 Np 237
94 Pu 242
95 Am 243
96 Cm 247
97 Bk 249
98 Cf 251
99 Es 254
100 Fm 253
101 Md 256
102 No 253
103 Lr 257
Transition metals
Inner transition series

Predicting reactions

Throughout the inorganic section of this book, each **group*** of elements has an introduction and chart which summarize some of the properties of the group's elements. Below the charts are blue boxes which highlight trends going down the group. After the introduction, more common group members are defined. Information on the other members of the group can often be determined from trends in **reactivity*** going down the group.

The following steps show how to predict the **reactivity*** of cesium with cold water. Cesium is in Group I (see pages 168-169).

1. The chart introducing Group I shows that the reactivity of the elements increases going down the group.

2. From the definitions of lithium, sodium and potassium it is found that all three elements react with water with increasing violence going down the group – lithium reacts gently, sodium reacts violently and potassium reacts very violently.

It is predicted that cesium, as it comes after potassium going down the group, will react extremely violently with wat

***Amphoteric**, 151; **Anion**, 130; **Basic**, 151 (**Base**); **Cation**, 130; **Group**, 165; **Outer shell**, 127; **Oxides**, 183; **Oxidizing agent**, 148; **Period**, 164; **Periodic table**, 164; **Reactivity**, 158; **Reducing agent**, 148.

HYDROG

Hydrogen (H_2),
first and lightest e
the most common
odourless, inflamm
on Earth in comp
natural gas and ste
of **water gas*** and
agent*, burns in a
with many substan
compounds of hy
Hydrogen is used
hydrogenation, pa
page 180), and as a

Hydrogen atom

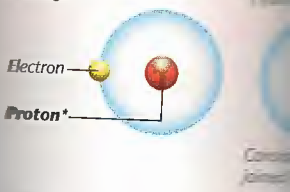

Electron
Proton*

***Hydrogen ion** (H^+). Consists of o*
lectron). Formed when a hydrogen
ydrogen ions combine with water
ons. Excess hydronium ions in a*

ydrogen reacts with sodium to

$2Na(s) + H_2(g)$

Sodium Hydrogen

*drogen is a **reducing agent***.*

$CuO(s) + H_2(g) \rightarrow$

Copper(II) oxide *Hydrogen*

Deuterium (**D** or ^2_1H)
isotope* of hydrogen
e **neutron***. It makes u
rogen. Water molecul
terium are called **deute**
avy water molecules. Hea
ear reactors to slow the

itium (**T** or ^3_1H)
isotope* of hydrogen
two **neutrons***. It is rare
ear reactors. It is **radioa**
particles*. **Tritiated wa**
molecules in which a
replaced by a **tritium**
rs to find how much

ic number, 127; **Beta particle**,
gen bond, 134; **Hydronium**
molecule, 133; **Proton**, 126;

HYDROGEN

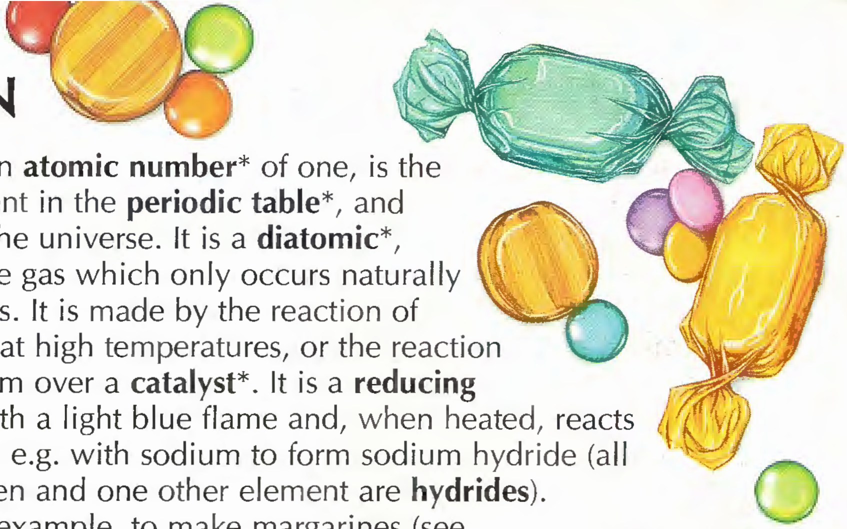

Hydrogen (**H_2**), with an **atomic number*** of one, is the first and lightest element in the **periodic table***, and the most common in the universe. It is a **diatomic***, odourless, inflammable gas which only occurs naturally on Earth in compounds. It is made by the reaction of natural gas and steam at high temperatures, or the reaction of **water gas*** and steam over a **catalyst***. It is a **reducing agent***, burns in air with a light blue flame and, when heated, reacts with many substances, e.g. with sodium to form sodium hydride (all compounds of hydrogen and one other element are **hydrides**). Hydrogen is used, for example, to make margarines (see **hydrogenation**, page 193) and ammonia (see **Haber process**, page 180), and as a rocket fuel. See also pages 217 and 218.

Sucrose ($C_{12}H_{22}O_{11}$), the sugar in sweets, is a compound of carbon, hydrogen and oxygen.

Hydrogen atom

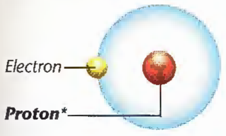

Hydrogen molecule

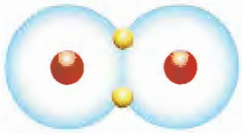

*Consists of two hydrogen atoms joined by a **covalent bond***.*

Hydrogen ion** (H^+). Consists of only one **proton (no electron). Formed when a hydrogen atom loses its electron. Hydrogen ions combine with water molecules to form **hydronium ions***. Excess hydronium ions in a solution make it acidic.*

Hydrogen reacts with sodium to make sodium hydride.

$2Na(s)$	+	$H_2(g)$	→	$2NaH(s)$
Sodium		*Hydrogen*		*Sodium hydride*

*Hydrogen is a **reducing agent***.*

$CuO(s)$	+	$H_2(g)$	→	$Cu(s)$	+	$H_2O(l)$
Copper(II) oxide		*Hydrogen*		*Copper*		*Water*

Deuterium (D or ${}^2_1H$)

An **isotope*** of hydrogen with one **proton*** and one **neutron***. It makes up 0.0156% of natural hydrogen. Water molecules containing deuterium are called **deuterium oxide** (**D_2O**) or **heavy water** molecules. Heavy water is used in nuclear reactors to slow the fast moving neutrons.

Tritium (T or ${}^3_1H$)

An **isotope*** of hydrogen with one **proton*** and two **neutrons***. It is rare but is produced by nuclear reactors. It is **radioactive***, emitting **beta particles***. **Tritiated water** contains some water molecules in which a hydrogen atom has been replaced by a **tritium** atom. It is used by doctors to find how much fluid a patient passes.

Hydrogen peroxide (H_2O_2)

A syrupy liquid. It is an oxide of hydrogen and a strong **oxidizing agent***. It is sold in solution as disinfectant and bleach.

Water (H_2O)

An oxide of hydrogen and one of the most common compounds on Earth. It is a colorless, odorless liquid which freezes at 0°C, boils at 100°C, has its greatest density (1g cm^{-3}) at 4°C and is the best solvent known. It is made of **polar molecules*** linked by **hydrogen bonds*** and is formed when hydrogen burns in oxygen. See also pages 206 and 218.

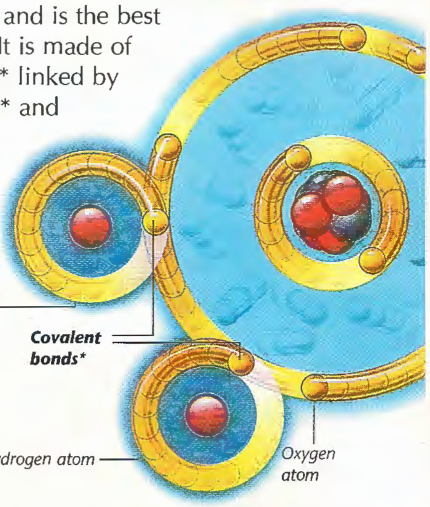

*Diagram of a **water** molecule (**polar molecule***)*

Hydroxide

A compound made of a **hydroxide ion** (**OH^-**) and a **cation***. Solutions containing more OH^- ions than H^+ ions are alkaline. Many hydroxides are not water-soluble, e.g. **lead(II) hydroxide** (**$Pb(OH)_2$**). However, the hydroxides of Group I elements and some others are water-soluble.

---

* **Atomic number**, 127; **Beta particle**, 128; **Catalyst**, 161; **Cation**, 130; **Covalent bond**, 132; **Diatomic**, 124; **Hydrogen bond**, 134; **Hydronium ion**, 150; **Isotope**, 127; **Neutron**, 126; **Oxidizing agent**, 148; **Periodic table**, 164; **Polar molecule**, 133; **Proton**, 126; **Radioactivity**, 128; **Reducing agent**, 148; **Water gas**, 179 (**Carbon monoxide**).

GROUP I, THE ALKALI METALS

The elements in **Group I** of the **periodic table*** are called **alkali metals** as they are all metals which react with water to form alkaline solutions. They all have similar chemical properties and their physical properties follow certain patterns. The chart below shows some of their properties.

Some properties of Group I elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Lithium	Li	6.94	2,1	INCREASING ↓	Silver-white metal	See below.
Sodium	Na	22.99	2,8,1		Soft, silver-white metal	See below.
Potassium	K	39.10	2,8,8,1		Soft, silver-white metal	See page 169.
Rubidium	Rb	85.47	Complex configuration but still one outer electron,		Soft, silver-white metal	To make special glass
Cesium	Cs	132.90			Soft metal with gold sheen	In photocells* and as a catalyst*
Francium	Fr	No known stable isotope*				

*The atoms of all Group I elements have one electron in their **outer shell****, hence the elements are powerful **reducing agents*** because this electron is easily lost in reactions. The resulting ion has a charge of +1 and is more stable because its new outer shell is complete (see **octet**, page 127). All Group I elements react in this way to form **ionic compounds***.*

*Going down the group, the reaction of the elements with water gets more violent, each forming an alkaline solution and hydrogen gas. The first three members tarnish in air and **rubidium** and **cesium** catch fire. All Group I elements are stored under oil because of their reactivity. They are soft enough to be cut easily with a knife.*

These two pages contain more information on **lithium**, **sodium**, **potassium** and their compounds. They are typical Group I elements.

Lithium (Li)

The least reactive element in Group I of the periodic table and the lightest solid element. Lithium is rare and is only found in a few compounds, from which it is extracted by **electrolysis***. It burns in air with a crimson flame. Lithium reacts vigorously with chlorine to form **lithium chloride** (**LiCl**) which is used in welding flux and air conditioners. A piece of lithium placed in water glides across the surface, fizzing gently.

$$2Li(s) + 2H_2O(l) \rightarrow 2LiOH(aq) + H_2(g)$$

Lithium | Water | Lithium hydroxide | Hydrogen

*After the reaction, the solution is strongly alkaline, due to the **lithium hydroxide** formed.*

Sodium (Na)

A member of Group I of the periodic table, found in many compounds. Its main ore is **rock salt** (containing **sodium chloride** – see also **potassium**). It is extracted from molten sodium chloride by **electrolysis***, using a **Downs' cell**. Sodium burns in air with an orange-yellow flame and reacts violently with non-metals and water (see equation for **lithium** and water, and substitute Na for Li). It is used in sodium vapor lamps and as a coolant in nuclear power stations.

Downs' cell (used to extract sodium from molten sodium chloride by electrolysis*)

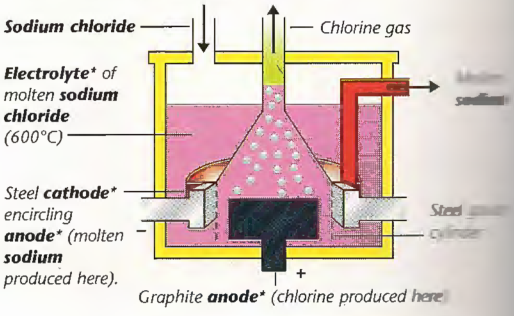
Sodium chloride
Chlorine gas
Electrolyte* *of molten* ***sodium chloride*** *(600°C)*
Steel ***cathode**** *encircling* ***anode**** *(molten* ***sodium*** *produced here).*
−
+
Graphite ***anode**** *(chlorine produced here*
Steel

---

* **Anode**, 156 (**Electrode**); **Catalyst**, 161; **Cathode**, 156 (**Electrode**); **Electrolysis**, **Electrolyte**, 156; **Electron configuration**, 127; **Ionic compound**, 131; **Isotope**, **Outer shell**, 127; **Periodic table**, 164; **Photocell**, 345; **Reducing agent**, 148; **Relative atomic mass**, 138.

Sodium hydroxide
A white, **deliquescen**
electrolysis* of **brine**
A **strong base***, it rea
sodium **salt*** and wate
soaps and paper.

Sodium carbonate
A white solid that diss
an alkaline solution.
washing soda (Na_2CO
also page 207), has
efflorescent* crystals
when ammonia, water
chloride react with ca
in the **Solvay process**.

***Washing soda** is used in the*
*of glass, as a **water softener****
in bath crystals.

Sodium bicarbonate
Also called **sodium hy**
bicarbonate of soda.
the **Solvay process** (see
In water it forms a weak

$2NaHCO_3(s) \rightarrow Na_2CO_3$

Sodium bicarbonate | Sodium carbonate

Sodium chloride (**NaCl**
A white solid which occu
in sea water and **rock salt**
see **sodium**). It forms **brine**
when dissolved in water
and is used to make
sodium hydroxide and
sodium carbonate.

***Sodium chloride** is used to*
preserve and flavor food.

Sodium nitrate (**NaNO**
A white solid used as a ferti
preserve meat.

acid, 344; **Deliquescent**, **Efflore**
Electrolysis, 156; **Hydrate**, 154;
Strong base, 152; **Water softener**

Sodium hydroxide (NaOH) or **caustic soda**

A white, **deliquescent*** solid, produced by **electrolysis*** of **brine** (see **sodium chloride**). A **strong base***, it reacts with acids to form a sodium **salt*** and water. It is used to make soaps and paper.

Sodium carbonate (Na_2CO_3)

A white solid that dissolves in water to form an alkaline solution. Its **hydrate***, called **washing soda** ($Na_2CO_3.10H_2O$ – see also page 207), has white, **efflorescent*** crystals and is made when ammonia, water and **sodium chloride** react with carbon dioxide in the **Solvay process**.

Washing soda *is used in the making of glass, as a* ***water softener**** *and in bath crystals.*

Sodium bicarbonate ($NaHCO_3$)

Also called **sodium hydrogencarbonate** or **bicarbonate of soda**. A white solid made by the **Solvay process** (see **sodium carbonate**). In water it forms a weak alkaline solution.

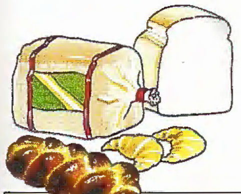

Sodium bicarbonate *is used in baking. The carbon dioxide gas it gives off when heated makes dough rise. It is also used as an* ***antacid**** *to relieve indigestion.*

$$\underset{\text{Sodium bicarbonate}}{2NaHCO_3(s)} \rightarrow \underset{\text{Sodium carbonate}}{Na_2CO_3(s)} + \underset{\text{Water}}{H_2O(l)} + \underset{\text{Carbon dioxide}}{CO_2(g)}$$

Sodium chloride (NaCl) or **salt**

A white solid which occurs in sea water and **rock salt** (see **sodium**). It forms **brine** when dissolved in water and is used to make **sodium hydroxide** and **sodium carbonate**.

Sodium chloride *is used to preserve and flavor food.*

Sodium nitrate ($NaNO_3$) or **Chile saltpeter**

A white solid used as a fertilizer and also to preserve meat.

Potassium (K)

A member of Group I of the periodic table. Potassium compounds are found in sea water and **rock salt** (containing **potassium chloride** – see also **sodium**). Potassium is extracted from molten potassium chloride by **electrolysis***. It is very reactive, reacting violently with chlorine and also with water (see equation for **lithium**, and substitute K for Li). It has few uses, but some of its compounds are important.

Potassium *reacting with water. It whizzes across the water giving off so much heat that the hydrogen produced bursts into flames.*

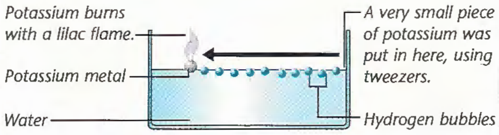
Potassium burns with a lilac flame.
Potassium metal
Water
A very small piece of potassium was put in here, using tweezers.
Hydrogen bubbles

Potassium hydroxide (KOH) or **caustic potash**

A white, **deliquescent*** solid. It is a **strong base*** which reacts with acids to form a potassium **salt*** and water. It is used to make soap (see page 202).

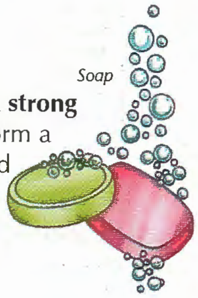

Potassium carbonate (K_2CO_3)

A white solid which is very water-soluble, forming an alkaline solution. It is used to make glass, dyes and soap.

Potassium chloride (KCl)

A white, water-soluble solid. Large amounts are found in sea water and **rock salt** (see **potassium**). It is used in fertilizers and to produce **potassium hydroxide**.

Potassium nitrate (KNO_3) or **saltpeter**

A white solid which dissolves in water to form a **neutral*** solution. It is used in fertilizers, explosives and to preserve meat.

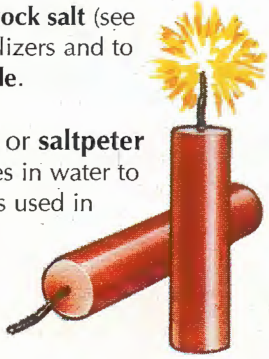

Gunpowder and some types of dynamite contain ***potassium nitrate****.*

Potassium sulfate (K_2SO_4)

A white solid, forming a **neutral*** solution in water. It is an important fertilizer.

---

* **Antacid**, 344; **Deliquescent**, **Efflorescent**, 206; **Electrolysis**, 156; **Hydrate**, 154; **Neutral**, 151; **Salts**, 153; **Strong base**, 152; **Water softeners**, 207.

GROUP II, THE ALKALINE-EARTH METALS

The elements in **Group II** of the **periodic table*** are called the **alkaline-earth metals**. The physical properties of the members of Group II follow certain trends, and, except **beryllium**, they all have similar chemical properties. They are very reactive, though less reactive than Group I elements. The chart below shows some of their properties. These two pages contain more information on **magnesium**, **calcium** and their compounds. Magnesium and calcium are typical Group II elements.

Some properties of Group II elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Beryllium	Be	9.01	2,2	INCREASING ↓	Hard, white metal	In light, corrosion-resistant alloys*
Magnesium	Mg	24.31	2,8,2		Silver-white metal	See below.
Calcium	Ca	40.31	2,8,8,2		Soft, silver-white metal	See right.
Strontium	Sr	87.62	Complex configuration, but still 2 outer electrons		Soft, silver-white metal	In fireworks
Barium	Ba	137.34			Soft, silver-white metal	In fireworks and medicine
Radium	Ra	Rare radioactive* metal			Soft, silver-white metal	An isotope* is used to treat cancer.

*The atoms of all Group II elements have two electrons in their **outer shell***, hence the elements are good **reducing agents*** because these electrons are fairly easily lost in reactions. Each resulting ion has a charge of +2 and is more stable because its new outer shell is complete (see **octet**, page 127). All Group II elements react this way to form **ionic compounds***, though some **beryllium** compounds have **covalent*** properties.*

*Going down the group, elements react more readily with both water and oxygen (see **magnesium** and **calcium**). They all **tarnish*** in air, but **barium** reacts so violently with both water and oxygen that it is stored under oil.*

Magnesium (Mg)

A member of Group II of the periodic table. It only occurs naturally in compounds, mainly in either **dolomite** ($CaCO_3.MgCO_3$ – a rock made of magnesium and **calcium carbonate**) or in **magnesium chloride** ($MgCl_2$), found in sea water. Magnesium is produced by the **electrolysis*** of molten magnesium chloride. It burns in air with a bright white flame.

$$2Mg(s) + O_2(g) \rightarrow 2MgO(s)$$

Magnesium Oxygen Magnesium oxide

***Magnesium** reacts rapidly with dilute acids:*

$$Mg(s) + 2HCl(aq) \rightarrow MgCl_2(aq) + H_2(g$$

Magnesium Hydrochloric acid Magnesium chloride Hydrogen

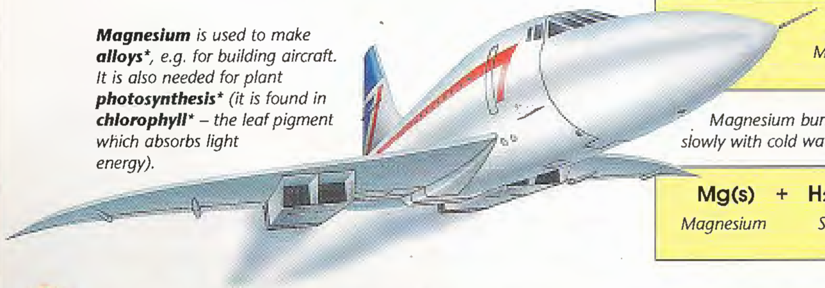

Magnesium** is used to make **alloys, e.g. for building aircraft. It is also needed for plant **photosynthesis*** (it is found in **chlorophyll*** – the leaf pigment which absorbs light energy).*

$$Mg(s) + Cl_2(g) \rightarrow MgCl_2(s$$

Magnesium Chlorine Magnesi chloride

Magnesium burns vigorously in chlorine (see above),
slowly with cold water and rapidly with steam (see below

$$Mg(s) + H_2O(g) \rightarrow MgO(s) + H_2(g$$

Magnesium Steam Magnesium oxide Hydrogen

* **Alloy**, 344; **Chlorophyll**, 255; **Covalent compounds**, 132; **Electrolysis**, 156; **Electron configuration**, 127; **Ionic compound**, 131; **Isotope**, **Outer shell**, 127; **Photosynthesis**, 254; **Periodic table**, 164; **Radioactivity**, 128; **Reducing agent**, 148; **Relative atomic mass**, 138; **Tarnish**, 345.

Magnesium hydroxi

A white solid that is
soluble in water. It is a
therefore **neutralizes***

***Magnesium hydroxide** is used*
for treating stomach upsets,

Magnesium sulfate

A white solid used in
constipation, in leather
fire-proofing.

Calcium (Ca)

A member of Group
periodic table. It occu
in many compound
those found in milk,
and in the Earth's cru
Calcium is extracted
its compounds by
electrolysis*. It burns
oxygen with a red fla
and reacts readily w
cold water and very
rapidly with dilute ac
(for equations see mag
and substitute Ca for
Calcium is used to m
high-grade steel and
production of uraniu

***Calcium** compounds are found*
and teeth.

Calcium hydroxide
slaked lime

A white solid which
water to form **limewat**
alkaline and is used to
(see page 218). Calciu
mortars and to remove

Calcium sulfate

A white solid that
occurs both as
anhydrite calcium
sulfate (**$CaSO_4$**) and
gypsum (**$CaSO_4.2H_2O$**
When heated, gypsum
forms plaster of Paris.

Antacid, 344; **Base**, 151; **Deli**
Electrolysis, 156; **Neutralizati**
Salts, 153; **Temporary hardness**

Magnesium hydroxide ($Mg(OH)_2$)

A white solid that is only slightly soluble in water. It is a **base*** and therefore **neutralizes*** acids.

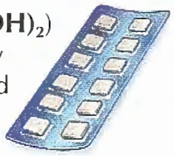

Magnesium hydroxide** is used in **antacids* *for treating stomach upsets, particularly indigestion.*

Magnesium sulfate ($MgSO_4$)

A white solid used in medicines for treating constipation, in leather processing and in fire-proofing.

Magnesium oxide (MgO)

A white solid which is slightly water-soluble. It is a **base***, forming magnesium **salts*** when it reacts with acids. It has a very high melting point and is used to line some furnaces.

$MgO(s)$	+	$2HCl(aq)$	$\rightarrow$	$MgCl_2(aq)$	+	$H_2O(l)$
Magnesium oxide		*Hydrochloric acid*		*Magnesium chloride*		*Water*

Calcium (Ca)

A member of Group II of the periodic table. It occurs naturally in many compounds, e.g. those found in milk, bones and in the Earth's crust. Calcium is extracted from its compounds by **electrolysis***. It burns in oxygen with a red flame and reacts readily with cold water and very rapidly with dilute acids (for equations see **magnesium** and substitute Ca for Mg). Calcium is used to make high-grade steel and in the production of uranium.

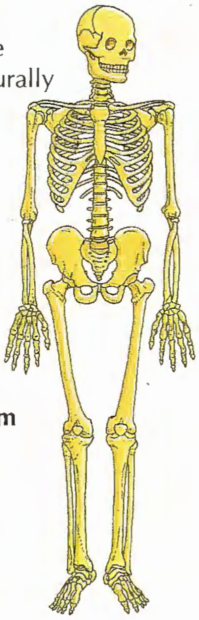

***Calcium** compounds are found in bones and teeth.*

Calcium hydroxide ($Ca(OH)_2$) or slaked lime

A white solid which dissolves slightly in water to form **limewater**. This is weakly alkaline and is used to test for carbon dioxide (see page 218). Calcium hydroxide is used in mortars and to remove excess acidity in soils.

Calcium sulfate

A white solid that occurs both as **anhydrite calcium sulfate** ($CaSO_4$) and **gypsum** ($CaSO_4.2H_2O$). When heated, gypsum forms plaster of Paris.

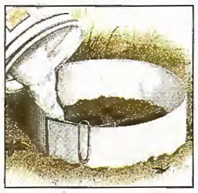

Plaster of Paris used to make a cast of an animal track

Calcium oxide (CaO) or quicklime

A white solid. It is a **base*** which is made by heating **calcium carbonate** in a lime kiln.

$CaCO_3(s)$	$\rightleftharpoons$	$CaO(s)$	+	$CO_2(g)$
Calcium carbonate	***Reversible reaction****	*Calcium oxide*		*Carbon dioxide*

***Calcium oxide, calcium carbonate** and **calcium hydroxide** are used to remove excess soil acidity.*

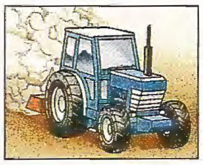

Calcium carbonate ($CaCO_3$)

A white, insoluble solid that occurs naturally as **limestone**, **chalk**, **marble** and **calcite**. It dissolves in dilute acids. Calcium carbonate is used to obtain **calcium oxide**, make cement and as building stone.

***Limestone** rock is corroded because rainwater containing dissolved carbon dioxide reacts with the limestone to form **calcium bicarbonate** (which itself dissolves slightly in water).*

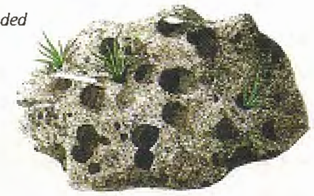

*The **calcium bicarbonate** formed when **limestone** dissolves in water causes **temporary hardness***.*

$CaCO_3(s)$	+	$H_2O(l)$	+	$CO_2(g)$	$\rightarrow$	$Ca(HCO_3)_2(aq)$
Calcium carbonate		*Water*		*Carbon dioxide*		*Calcium bicarbonate*

Calcium chloride ($CaCl_2$)

A white, **deliquescent***, water-soluble solid which is used as a **drying agent***.

---

* **Antacid**, 344; **Base**, 151; **Deliquescent**, 206; **Drying agent**, 344; **Electrolysis**, 156; **Neutralization**, 151; **Reversible reaction**, 162; **Salts**, 153; **Temporary hardness**, 207.

TRANSITION METALS

Transition metals have certain properties in common – they are hard, tough, shiny, **malleable*** and **ductile***. They **conduct*** heat and electricity, and have high melting points, boiling points and densities. Transition metals form **complex ions*** which are colored in solution. They also have more than one possible charge, e.g. Fe^{2+} and Fe^{3+}. Transition metals have many uses, some of which are shown on these two pages. (Information on iron, copper and zinc can be found on pages 174-175.) The members of the **inner transition series** (see page 165) are not shown here, as they are very rare and often unstable.

21 **Sc** Scandium 45	22 **Ti** Titanium 48	23 **V** Vanadium 51	24 **Cr** Chromium 52	25 **Mn** Manganese 55	26 **Fe** Iron 56	27 **Co** Cobalt 59	28 **Ni** Nickel 59	29 **Cu** Copper 64	30 **Zn** Zinc 65
39 **Y** Yttrium 89	40 **Zr** Zirconium 91	41 **Nb** Niobium 93	42 **Mo** Molybdenum 96	43 **Tc** Technetium 99	44 **Ru** Ruthenium 101	45 **Rh** Rhodium 103	46 **Pd** Palladium 106	47 **Ag** Silver 108	48 **Cd** Cadmium 112
57 **La** Lanthanum 139	72 **Hf** Hafnium 178.5	73 **Ta** Tantalum 181	74 **W** Tungsten 184	75 **Re** Rhenium 186	76 **Os** Osmium 190	77 **Ir** Iridium 192	78 **Pt** Platinum 195	79 **Au** Gold 197	80 **Hg** Mercury 201

Sc

Scandium
A very rare, light, silvery-white metal.

Ti

Titanium
A metal used to make strong, light, corrosion-resistant ***alloys**** *with high melting points, e.g. those used in aircraft wings, artificial hips, heart pacemakers, golf clubs and jewelry.*

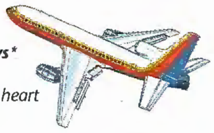

V

Vanadium
A rare, hard, white metal that is used to increase the strength and hardness of steel ***alloys**** *such as those used to make tools.* ***Vanadium pentoxide (V_2O_5)*** *is the* ***catalyst**** *used in the* ***contact process**** *to make sulfuric acid.*

Cr

Chromium
A hard, white metal found as ***chrome iron ore****. It is used as a corrosion-resistant coating on steel objects and in stainless steel. Chromium plating is used on car parts, bicycle handlebars and cutlery.*

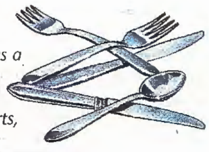

Mn

Manganese
A hard, brittle, reddish-white metal. It is found as ***pyrolusite (MnO_2)*** *and is used in many* ***alloys****, such as steels and bronzes.*

Fe

Iron
Has various uses, some of which are described on page 174.

Co

Cobalt
A hard, silvery-white, magnetic metal found combined with sulfur and arsenic. It is used in ***alloys****, e.g. with* ***iron*** *to make magnets. Its* ***radioisotope**** *is used to treat cancer.* ***Cobalt(II) chloride ($CoCl_2$)*** *is used to test for water (see page 218). Cobalt produces a blue color in glass and ceramics.*

Ni

Nickel
A magnetic metal that is found as ***nickel sulfide (NiS)****. It is used as a* ***catalyst****, in* ***alloys****, in* ***electroplating**** *and in rechargeable batteries. An* ***alloy**** *of nickel is used in coins and stainless steel.*

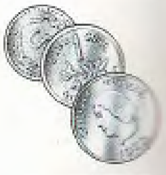

Cu

Copper
Has various uses, some of which are described on page 175.

Zn

Zinc
Has various uses, some of which are described on page 175.

Y

Yttrium
A metal used in crystals
to aluminum high-
transmission lines to increase

Zr

Zirconium
A rare metal used in
compounds and to

Nb

Niobium
A rare, gray metal.
used in some stainless
resistant to corrosion at high
alloys* *are used in jet engines*

Mo

Molybdenum
A hard, white metal
steels. It is used to make

Tc

Technetium
A metal which only
formed by uranium ***fission****
locate tumors.

Ru

Ruthenium
A hard, brittle metal. It
a ***catalyst****.*

Rh

Rhodium
A hard, silvery-white
used as a ***catalyst****, in*
high quality mirrors.

Pd

Palladium
A silvery-white metal used
and high-grade surgical
of palladium and ***platinum*** *reduce*
hydrocarbons* *in car exhaust.*

Ag

Silver
A soft, white metal
with other elements, e.g.
alloys*, in jewelry and coinage*
electroplated* *onto objects. Silver*
photography.

Cd

Cadmium
A soft, silvery-white metal
with ***zinc*** *and used to make*
alloys* *with low melting points.*
in ***control rods**** *in nuclear*
also in nickel-cadmium
batteries. Cadmium compounds are
yellow, orange and red pigments
plastics, paints and ceramics.

---

* **Alloy**, 344; **Catalyst**, 161; **Complex ion**, 154 (**Complex salt**); **Conductivity**, 28, 63; **Contact process**, 185; **Ductile**, 344; **Electroplating**, 157; **Malleable**, 345; **Radioisotope**, 128.

Alloy, **Amalgam**, 344;
plating, 157; **Fission**
rating, 199; **Oxidizing**

Y ***Yttrium***
A metal used in crystals for lasers, and added to aluminum high-voltage electricity transmission lines to increase ***conductivity****.*

Zr ***Zirconium***
A rare metal used in ***alloys****,* ***abrasives****, flame-proofing compounds and to absorb* ***neutrons**** *in nuclear reactors.*

Nb ***Niobium***
A rare, gray metal. Small quantities of it are used in some stainless steel to make it resistant to corrosion at high temperatures. Its ***alloys**** *are used in jet engines and rockets.*

Mo ***Molybdenum***
A hard, white metal that is used in ***alloys****, e.g. special steels. It is used to make ball bearings and lamp filaments.*

Tc ***Technetium***
A metal which only occurs as an unstable ***isotope**** *formed by uranium* ***fission****. It is used in medicine to locate tumors.*

Ru ***Ruthenium***
A hard, brittle metal. It is used in ***alloys**** *and as a* ***catalyst****.*

Rh ***Rhodium***
A hard, silvery-white metal found with ***platinum****. It is used as a* ***catalyst****, in* ***alloys**** *and in thin films to make high quality mirrors.*

Pd ***Palladium***
A silvery-white metal used in ***alloys****, telephone relays and high-grade surgical instruments.* ***Catalysts**** *made of palladium and* ***platinum*** *reduce the carbon monoxide and* ***hydrocarbons**** *in car exhaust.*

Ag ***Silver***
A soft, white metal sometimes found combined with other elements, e.g. sulfur. It is used, often in ***alloys****, in jewelry and coinage and is also* ***electroplated**** *onto objects. Silver* ***halides**** *are used in photography.*

Cd ***Cadmium***
A soft, silvery-white metal found with ***zinc*** *and used to make* ***alloys**** *with low melting points. It is used in* ***control rods**** *in nuclear reactors and also in nickel-cadmium rechargeable batteries. Cadmium compounds are used as yellow, orange and red pigments in plastics, paints and ceramics.*

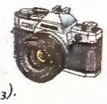

La ***Lanthanum***
*Similar in properties to aluminum, this is one of a group of rare metallic elements (****lanthanides****) with* ***atomic numbers**** *57-71. Camera lenses contain* ***lanthanum oxide*** *(La_2O_3).*

Hf ***Hafnium***
A metal used in ***control rods**** *in nuclear reactors to absorb* ***neutrons****, and in* ***alloys**** *to make cutting tools.*

Ta ***Tantalum***
A rare, pale gray metal used in electric lamp filaments and ***alloys****. Tantalum is also used in surgery to replace parts of the body, e.g. in skull plates and wire connecting the ends of nerves.*

W ***Tungsten***
A hard, gray metal that is resistant to corrosion. It is used in ***alloys**** *to make tools and lamp filaments.*

Re ***Rhenium***
A hard, heavy, gray metal used in ***thermocouples**** *and* ***catalysts****. It is used to make low-lead and lead-free* ***gasoline**** *with a high* ***octane rating****. An* ***alloy**** *of rhenium and* ***tungsten*** *is used in flash bulbs.*

Os ***Osmium***
A hard, white, crystalline metal, the densest element known. It is found with ***platinum*** *and used in* ***alloys**** *with platinum and* ***iridium****, e.g. in electrical contacts.* ***Osmium tetroxide*** *(OsO_4) is used to treat inflammatory arthritis.*

Ir ***Iridium***
A rare, hard, unreactive metal that looks like ***platinum*** *and is found with it. It is used in medicine in* ***radioactive**** *implants to control tumors and (with platinum) in heart pacemakers. It is also found in an* ***alloy**** *used for fountain pen nib-tips.*

Pt ***Platinum***
A hard, silvery-white metal used as a ***catalyst**** *and to make electrical contacts, jewelry, and various pin, plate and hinge devices for securing human bones. It is also used (with* ***iridium****) in wire* ***electrodes**** *in heart pacemakers.*

Au ***Gold***
A soft, shiny, yellow metal. It is very unreactive and usually found uncombined. South Africa and Russia have the most important gold deposits. Gold only reacts with very vigorous ***oxidizing agents**** *(such as chlorine) and certain combinations of acids. It is often used in* ***alloys**** *with* ***silver*** *or* ***copper*** *to give it more strength. These alloys are used in jewelry, coins and dentistry. Pure gold (24 carat gold) is also used in jewelry.*

Hg ***Mercury*** *or* ***quicksilver***
A poisonous, silvery-white, liquid metal mainly found as ***cinnabar*** *(HgS). It is used in thermometers, barometers, lamps and also in* ***amalgams**** *used by dentists.*

---

* **Abrasive**, **Alloy**, **Amalgam**, 344; **Atomic number**, 127; **Catalyst**, 161; **Conductivity**, 28, 63; **Control rods**, 94; **Electrode**, 156; **Electroplating**, 157; **Fission (nuclear)**, 129; **Gasoline**, 199; **Halides**, 186; **Hydrocarbons**, 190; **Isotope**, 127; **Neutron**, 126; **Octane rating**, 199; **Oxidizing agent**, 148; **Radioactivity**, 128; **Thermocouple**, 27.

IRON, COPPER AND ZINC

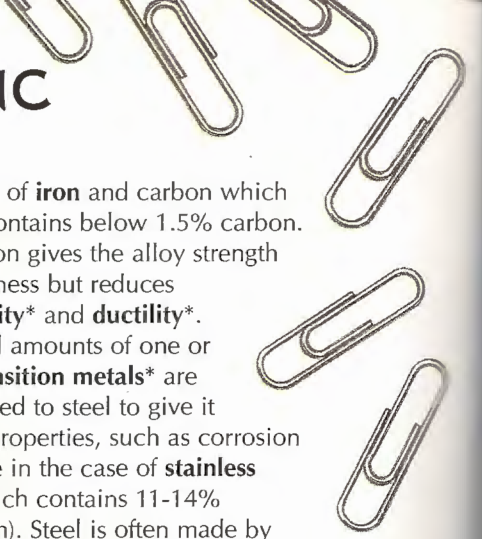

Iron (Fe)

A **transition metal*** in Period 4 of the **periodic table***. It is a fairly soft, white, magnetic metal which only occurs naturally in compounds. One of its main ores is **hematite** (Fe_2O_3), or **iron(III) oxide**, from which it is extracted in a **blast furnace**. Iron reacts to form both **ionic** and **covalent compounds*** and reacts with moist air to form **rust**. It burns in air when cut very finely into iron filings and also reacts with dilute acids. It is above hydrogen in the **electrochemical series***.

Extracting iron using a blast furnace

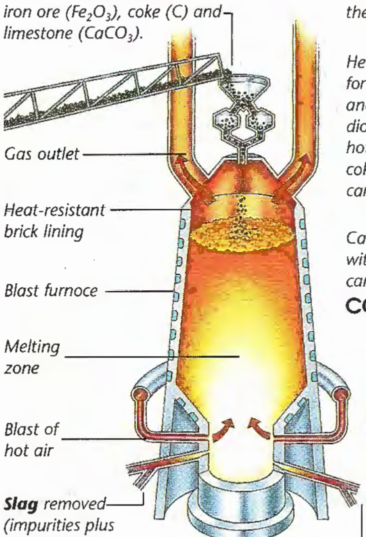
Raw materials fed in here are iron ore (Fe_2O_3), coke (C) and limestone ($CaCO_3$).
Gas outlet
Heat-resistant brick lining
Blast furnoce
Melting zone
Blast of hot air
Slag removed (impurities plus calcium oxide).

Reactions in the furnace:

Heated limestone forms calcium oxide and gives off carbon dioxide. Oxygen from hot air reacts with coke and also forms carbon dioxide.

Carbon dioxide reacts with coke to form carbon monoxide.

$$CO_2 + C \rightarrow 2CO$$

*Iron ore is **reduced*** to iron by carbon monoxide.*

$$Fe_2O_3 + 3CO \rightleftharpoons 2Fe + 3CO_2$$

Reversible reaction*

Molten iron tapped off here.

Iron made in the blast furnace is called **pig iron**. It contains about 5% carbon and 4% other impurities, such as sulfur. Most pig iron is converted to **steel**, although some is converted to **wrought iron** (by **oxidizing*** impurities) and some is melted down again along with scrap steel to make **cast iron**. Iron is a vital mineral in the human diet, as it is needed to make **hemoglobin***.

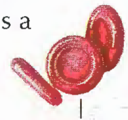

*Iron is found in **hemoglobin*** in red blood cells.*

***Wrought iron** is used to make crane hooks and anchor chains.*

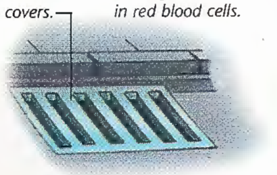

***Cast iron** is used to make drain covers.*

Steel

An **alloy*** of **iron** and carbon which usually contains below 1.5% carbon. The carbon gives the alloy strength and hardness but reduces **malleability*** and **ductility***. Measured amounts of one or more **transition metals*** are often added to steel to give it specific properties, such as corrosion resistance in the case of **stainless steel** (which contains 11-14% chromium). Steel is often made by the **basic oxygen process**. Scrap steel, molten iron and lime are put into a furnace, and oxygen is blasted onto the metal to **oxidize*** impurities.

***Steel** is used to make many objects. These steel paper clips contain about 0.08% carbon.*

Iron(II) or ferrous compounds

Iron compounds containing Fe^{2+} ions, e.g. **iron(II) chloride** ($FeCl_2$). Their solutions are green.

Iron(III) or ferric compounds

Iron compounds containing Fe^{3+} ions, e.g. **iron(III) chloride** ($FeCl_3$). Their solutions are yellow or orange.

Rust ($Fe_2O_3.xH_2O$) or hydrated iron(III) oxide

A brown solid formed when **iron**, water and air react together (see **corrosion**, page 209). The "x" in the formula shows that the number of water molecules varies. Iron and **steel** can be protected from rust by **galvanizing** – coating with a layer of **zinc** (see also **sacrificial protection**, page 159). The surface zinc **oxidizes*** in air, stopping the zinc and iron below from being oxidized. Galvanized cars remain rust-free longer than others.

*An ungalvanized car is only protected from rust by its coat of paint, and will rust more quickly than a galvanized one, though **phosphoric acid*** can be put on to stop the rust from spreading. Grease is put on engine parts to protect them from rust.*

***Alloy**, 344; **Covalent compounds**, 132; **Ductility**, 344 (**Ductile**); **Electrochemical series**, 159; **Hemoglobin**, 286 (**Red blood cells**); **Ionic compound**, 131; **Malleability**, 345 (**Malleable**); **Oxidation**, 148; **Periodic table**, **Phosphoric acid**, 182 (**Phosphorus pentoxide**); **Reduction**, 148; **Reversible reaction**, 162; **Transition metals**, 172.

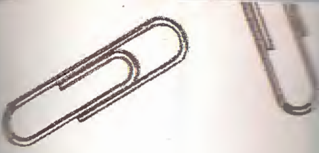

Copper (Cu)

A **transition metal*** in
table*. It is a red-
found naturally in
compounds are found
copper pyrites ((CuFe
($CuCO_3.Cu(OH)_2$). Co
the former by crushing
then roasting in a lim
silica. The iron comb
slag. The sulfur is rem
sulfur dioxide. The co
purified further by **elec**
unreactive metal and
slowly in air to form a
of **basic copper sulfate**
Copper is below hydro
electrochemical series*.
water, dilute acids or a
react with concentrated
sulfuric acid. (See also

***Copper** is a very good conductor
better), so it is used to make wires
Because it is soft but tough, it is
pipes for plumbing and central-*

*It is used in **alloys*** such
(copper plus **zinc**) and
plus tin) to make
and in **cupronickel**
nickel) to make*

*An **alloy*** of copper and gold is
jewelry. The greater the amount
less the number of carats the gold
less than 24 carats (pure gold).*

Zinc (Zn)

An element in Period
It is a silvery, soft metal
air. It is too reactive to
main ores are **zinc blende**
($ZnCO_3$) and **zincite**
extracted by roasting
oxide (ZnO) and then
it with coke. Zinc is ab
electrochemical series*.
with steam when red-h
used to coat iron and
galvanizing – see also
sacrificial protection,
in **alloys***, particularly

Alloy, 344; **Electrochemical series**
Petroleum, 198; **Periodic table**,
Transition metals, 172.

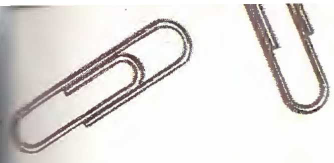

Copper (Cu)

A **transition metal*** in Period 4 of the **periodic table***. It is a red-brown, soft but tough metal found naturally in certain rocks. Its compounds are found in several ores, e.g. **copper pyrites** (($CuFe)S_2$) and **malachite** ($CuCO_3.Cu(OH)_2$). Copper is extracted from the former by crushing and removing sand and then roasting in a limited supply of air with silica. The iron combines with silica and forms **slag**. The sulfur is removed by burning to form sulfur dioxide. The copper produced is purified further by **electro-refining***. It is an unreactive metal and only **tarnishes*** very slowly in air to form a thin, green surface film of **basic copper sulfate** ($CuSO_4.3Cu(OH)_2$). Copper is below hydrogen in the **electrochemical series***. It does not react with water, dilute acids or alkalis. However, it does react with concentrated nitric or concentrated sulfuric acid. (See also page 219.)

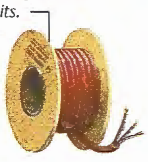

__Copper__ is a very good conductor of electricity (although silver is better), so it is used to make wires for electric circuits. Because it is soft but tough, it is also used to make pipes for plumbing and central-heating systems.

It is used in __alloys__ such as __brass__ (copper plus __zinc__) and __bronze__ (copper plus tin) to make "copper" coins, and in __cupronickel__ (copper plus nickel) to make "silver" coins.*

An __alloy__ of copper and gold is used to make jewelry. The greater the amount of copper, the less the number of carats the gold will be, i.e. less than 24 carats (pure gold).*

Copper(I) or cuprous compounds

Compounds containing Cu^+ ions, e.g. **copper(I) oxide** and **copper(I) chloride** (**CuCl**). **Copper(I)** compounds do not dissolve in water.

__Copper(I) oxide (Cu_2O)__ is used to make glass and paint.

Copper(II) or cupric compounds

Compounds that contain Cu^{2+} ions, e.g. **copper(II) sulfate** and **copper(II) chloride**. Copper(II) compounds dissolve in water to form light blue solutions, and are much more common than **copper(I) compounds**. **Copper(II) sulfate** (**$CuSO_4$**) has many uses, e.g. in dyeing and **electroplating***. It is also used in **Bordeaux mixture**, which kills molds growing on fruit and vegetables. (See also test for water, page 218.) **Copper(II) chloride** (**$CuCl_2$**) is used to remove sulfur from **petroleum***.

__Copper(II) chloride__ is used in fireworks to give a green color.

Zinc (Zn)

An element in Period 4 of the **periodic table***. It is a silvery, soft metal which **tarnishes*** in air. It is too reactive to occur naturally, and its main ores are **zinc blende** (**ZnS**), **calamine** (**$ZnCO_3$**) and **zincite** (**ZnO**). The zinc is extracted by roasting the ore to form **zinc oxide** (**ZnO**) and then **reducing*** it by heating it with coke. Zinc is above hydrogen in the **electrochemical series***. It reacts with oxygen, with steam when red-hot, and with acids. It is used to coat iron and steel to prevent **rust** (**galvanizing** – see also page 174 and **sacrificial protection**, page 159). It is also used in **alloys***, particularly **brass** (copper and zinc).

__Zinc oxide__ is used in a cream as a protection against skin irritation, e.g. diaper rash.

__Zinc__ is used in batteries.

---

* **Alloy**, 344; **Electrochemical series**, 159; **Electroplating**, **Electro-refining**, 157; **Petroleum**, 198; **Periodic table**, 164; **Reduction**, 148; **Tarnish**, 345; **Transition metals**, 172.

GROUP III ELEMENTS

The elements in **Group III** of the **periodic table*** are generally not as reactive as the elements in Groups I and II. Unlike those elements they show no overall trend in reactivity, and the first member of the group is a non-metal. The chart below shows some of their properties. More information on **aluminum** and its compounds can be found below the chart. Aluminum is the most widely used member of this group.

Some properties of Group III elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Boron	B	10.81	2,3	NO TREND ↓	Brown powder or yellow crystals	In **control rods***, glass and to harden steel
Aluminum	Al	26.98	2,8,3		White metal	See below.
Gallium	Ga	69.72	Complex configuration but still three outer electrons		Silver-white metal	In **semiconductors***
Indium	In	114.82			Soft, silver-white metal	In **control rods*** and transparent electrodes
Thallium	Tl	204.37			Soft, silver-white metal	In rat poison

Although all atoms of Group III elements have three outer electrons, they react to form different types of compound. Those of ***boron****, and some of* ***aluminum****, are* ***covalent*****. Other members of the group form mostly* ***ionic compounds*****.*

Aluminum (Al)

A member of Group III of the periodic table. It is the most common metal found on Earth, and occurs naturally in many compounds, e.g. **bauxite** (see **aluminum oxide**) from which it is extracted by **electrolysis***. It is hard, light, **ductile***, **malleable*** and a good conductor of heat and electricity. It reacts with the oxygen in air to form a surface layer of **aluminum oxide** which stops further corrosion. It also reacts with chlorine, dilute acids and alkalis.

Some uses of aluminum and its alloys*

Thin sheets of aluminum are used to wrap food, e.g. chocolate bars. It is also used to make soda cans.

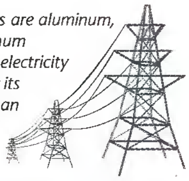

Powerlines are aluminum, as aluminum conducts electricity better for its weight than copper.

Its light weight makes it ideal for making many things, from aircraft to ladders and bicycles.

Aluminum oxide (Al_2O_3) or alumina

An **amphoteric***, white solid that is almost insoluble in water. It occurs naturally as **bauxite** ($Al_2O_3.2H_2O$ – see also **aluminum**) and as **corundum** (Al_2O_3) – an extremely hard crystalline solid. It is used in some cements and to line furnaces.

The extraction by electrolysis* of aluminum from bauxite

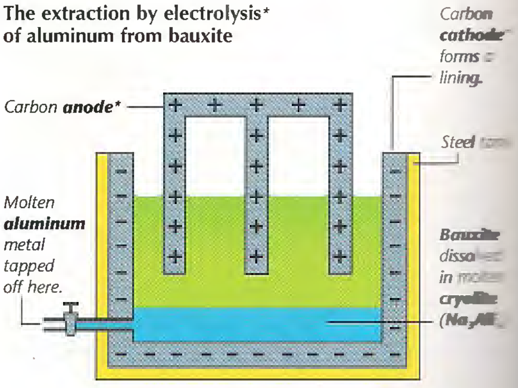
Carbon **cathode*** forms a lining.
Carbon **anode***
Steel
Molten **aluminum** metal tapped off here.
Bauxite dissolved in molten **cryolite** (Na_3AlF_6)

Aluminum hydroxide ($Al(OH)_3$)

A white, slightly water-soluble, **amphoteric*** solid, which is used in dyeing cloth, to make ceramics and as an **antacid***.

Aluminum sulfate ($Al_2(SO_4)_3$)

A white, water-soluble, crystalline solid used to purify water and make paper.

* **Alloy**, 344; **Amphoteric**, 151; **Anode**, 156 (**Electrode**); **Antacid**, 344; **Cathode**, 156 (**Electrode**); **Control rods**, 94; **Covalent compounds**, 132; **Ductile**, 344; **Electrolysis**, 156; **Electron configuration**, 127; **Ionic compound**, 131; **Malleable**, 345; **Periodic table**, 164; **Relative atomic mass**, 138; **Semiconductor**, 65.

GROUP IV

The elements in G
reactive and the
the group. For mo
below, **silicon** an

Some properties of Grou

Name of element	Chemical symbol
Carbon	C
Silicon	Si
Germanium	Ge
Tin	Sn
Lead	Pb

Silicon (Si)

A member of Group
It is a hard, shiny, gray
melting point. Silicon
common element in the
found in sand and rock
and **silicates**. When it
it reacts with some alk
otherwise it is genera

Silicon *is a* ***semiconductor******
is used to make silicon chips –
complete microelectronic circuits.

Silicon dioxide (SiO_2

Also called **silicon(IV)**
insoluble, white, crystal
many forms, such as **flint**
acidic and reacts with
Silicon dioxide has man
making of glass and ce

Sand *is impure* ***quartz****. Quartz crystals are used in watches.*

Silicates

Silicon compounds that
and oxygen, e.g. **calciu**
and make up most of the
used to make glass and

Covalent compounds, 132; **Elec**
Malleable, 345; **Metalloids**,
Semiconductor, 65; **Soft water**

GROUP IV ELEMENTS

The elements in **Group IV** of the **periodic table*** are generally not very reactive and the members show increasingly metallic properties going down the group. For more about the properties of these elements, see the chart below, **silicon** and **lead** (this page) and **carbon**, pages 178-179.

Some properties of Group IV elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Carbon	C	12.01	2,4	NO TREND ↓	Solid non-metal (see page 178)	See page 178.
Silicon	Si	28.09	2,8,4		Shiny, gray **metalloid*** solid	See below.
Germanium	Ge	72.59	Complex configuration but still four outer electrons		Grayish-white **metalloid*** solid	In transistors
Tin	Sn	118.69			Soft, silver-white metal	Tin plating, e.g. food containers
Lead	Pb	207.19			Soft, silver-gray metal	See below.

Although all atoms of Group IV elements have four outer electrons, they react to form different compound types. They all form ***covalent compounds****, but* ***tin*** *and* ***lead*** *form* ***ionic compounds**** *as well.*

Silicon (Si)

A member of Group IV of the periodic table. It is a hard, shiny, gray **metalloid*** with a high melting point. Silicon is the second most common element in the Earth's crust – it is found in sand and rocks as **silicon dioxide** and **silicates**. When it is ground into a powder it reacts with some alkalis and elements, otherwise it is generally unreactive.

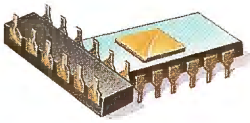

Silicon *is a* ***semiconductor**** *and is used to make silicon chips – complete microelectronic circuits.*

Silicon dioxide (SiO_2)

Also called **silicon(IV) oxide**, or **silica**. An insoluble, white, crystalline solid. It occurs in many forms, such as **flint** and **quartz**. It is acidic and reacts with concentrated alkalis. **Silicon dioxide** has many uses, e.g. in the making of glass and ceramics.

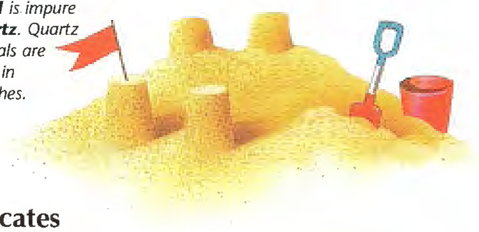

Sand *is impure* ***quartz****. Quartz crystals are used in watches.*

Silicates

Silicon compounds that also contain a metal and oxygen, e.g. **calcium metasilicate** (**$CaSiO_3$**), and make up most of the Earth's crust. They are used to make glass and ceramics.

Silicones

Complex, man-made compounds containing very long chains of **silicon** and oxygen atoms.

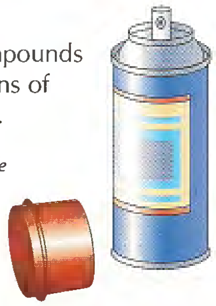

Silicones *are used in high-performance oils and greases and for non-stick surfaces. They are also used in waxes, polishes and varnishes, as they are water-repellant.*

Lead (Pb)

A member of Group IV of the periodic table. A soft, **malleable*** metal extracted from **galena** (**lead(II) sulfide**). It is not very reactive, though it **tarnishes*** in air, reacts slightly with **soft water*** and slowly with chlorine and nitric acid. It forms **ionic compounds*** called **lead(II)** or **plumbous compounds**, e.g. **lead(II) oxide (PbO)**, and **covalent compounds*** called **lead(IV)** or **plumbic compounds**, e.g. **lead(IV) oxide (PbO_2)**. Lead has many uses, e.g. in car batteries and roofing. It is used in hospitals to protect people from the harmful effects of X-rays.

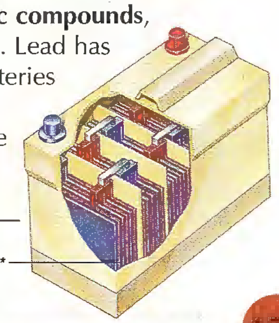

Car battery

Lead *and* ***lead(IV) oxide electrodes****

---

* **Covalent compounds**, 132; **Electron configuration**, 127; **Electrode**, 156; **Ionic compound**, 131; **Malleable**, 345; **Metalloids**, 165 (**Metal**); **Periodic table**, 164; **Relative atomic mass**, 138; **Semiconductor**, 65; **Soft water**, 207 (**Hard water**); **Tarnish**, 345.

CARBON

Carbon (C) is a member of **Group IV** of the **periodic table*** (see also chart, page 177). It is a non-metal and has several **allotropes***, including **diamond**, **graphite** and **buckminsterfullerene**, and an **amorphous*** (unstructured) form – **charcoal**. Carbon is not very reactive. It only reacts with steam when heated, and with hot, **concentrated*** sulfuric or nitric acids (see equation below). Carbon atoms can bond with up to four other atoms, including other carbon atoms. As a result, there are a vast number of carbon-based compounds (**organic compounds** – see page 190). Living tissue is made of carbon compounds, and animals break down these compounds to liberate energy (see **carbon cycle**, page 209).

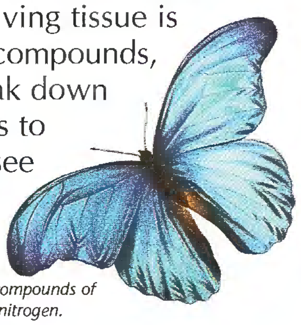

Animal and plant proteins are compounds of carbon, oxygen, hydrogen and nitrogen.

Equation for the reaction of carbon and nitric acid:

C	$+$	$4HNO_3$	$\rightarrow$	CO_2	$+$	$4NO_2$	$+$	$2H_2O$
Carbon		*Nitric acid*		*Carbon dioxide*		*Nitrogen dioxide*		*Water*

Carbon will burn in air when heated, to form carbon dioxide.

$C(s)$	$+$	$O_2(g)$	$\rightarrow$	$CO_2(g)$
Carbon		*Oxygen*		*Carbon dioxide*

If burned in a limited supply of air, ***carbon monoxide*** *forms.*

$2C(s)$	$+$	$O_2(g)$	$\rightarrow$	$2CO(g)$
Carbon		*Oxygen*		*Carbon monoxide*

Carbon is a ***reducing agent****. It reduces the* ***oxides**** *of any metal below zinc in the* ***reactivity series**** *of metals:*

C	$+$	$2PbO$	$\rightarrow$	CO_2	$+$	$2Pb$
Carbon		*Lead(II) oxide*		*Carbon dioxide*		*Lead*

Carbon is used in industry to reduce metal oxide ores to metals (see ***iron****, page 174).*

Diamond *is found in a rock called* ***kimberlite****. Rough diamonds are dull but can be cut to make glittering gems.*

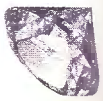

Diamond

A crystalline, transparent form of carbon. It is the hardest naturally occurring substance. All the carbon atoms are joined by strong **covalent bonds*** – accounting for its hardness and high melting point (3,750°C). Diamonds are used as **abrasives***, glass cutters, jewelry and on drill bits. **Synthetic diamonds** are made by subjecting **graphite** to high pressure and temperature, a very costly process.

The crystal structure of diamond

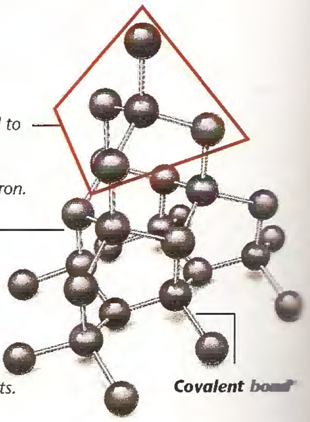

Each carbon atom is bonded to four other carbon atoms by ***covalent bonds**** *which are arranged to form a tetrahedron.*

Giant atomic lattice*

Diamond is harder and denser than ***graphite****.*

Diamond *and* ***graphite*** *both have high melting points.*

Covalent bond*

Graphite

A gray, crystalline form of carbon. The atoms in each layer are joined by strong **covalent bonds***, but the layers are only linked by weak **van der Waals' forces*** which allow them to slide over each other, making graphite soft and flaky. Graphite is the only non-metal to conduct electricity well. It also conducts heat. It is used as a lubricant, in **electrolysis*** (as **inert electrodes***), as contacts in electric motors, and in pencil leads.

The crystal structure of graphite

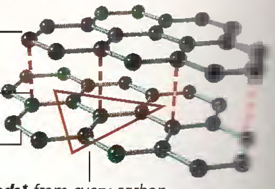

Layer of carbon atoms

van der Waals' forces* *link the layers.*

Covalent bond*

There are three ***covalent bonds**** *from every carbon atom to other carbon atoms in the same layer.*

---

* **Abrasive**, 344; **Allotropes**, 136 (**Allotropy**); **Amorphous**, 135; **Concentrated**, 144; **Covalent bond**, 132; **Electrolysis**, 156; **Giant atomic lattice**, 137; **Inert electrode**, 156; **Oxides**, 183; **Periodic table**, 164; **Reactivity series**, 158; **Reducing agent**, 148; **van der Waals' forces**, 134.

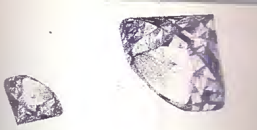

Buckminsterfull

A member of the fa
spherical crystalli
by condensing vap
Buckminsterfuller
dust between stars
rocks. Each molecu
in hexagons and pen
bonds* join hexago
fullerenes contain be
atoms.) Buckminste
but some of its compo
(substances which hav

The crystal structure of buckminsterfullerene

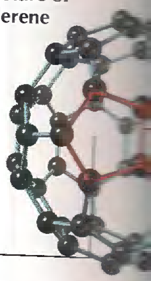

All molecules of buckminster-fullerene are ***icosahedral*** *– they have 20 hexagons and 12 pentagons.*

Pentagonal ring

Carbon dioxide (CO

A colorless, odorless
atmosphere (see **carbo**
It is made industrially
carbonate in a lime ki
for laboratory prepara
water to form **carbonic**

$CO_2(aq)$	$+$	$H_2O(l)$	$\rightleftharpoons$
Carbon dioxide		*Water*	***Reversible reaction***

Carbon dioxide is not
though it reacts with bo
and calcium hydroxide
(see page 218) and magn
ribbon burns in it.

Carbon dioxide *has many uses.* *used to make drinks fizzy.* *Carbon dioxide escapes when* *bottle or can is opened, as* *pressure is released.*

used in fire extinguishers. It is *than air, so forms a blanket* *the flames and does not* *air to reach the fire.*

---

phous, 135; **Anion**, **Cation**,
tor, 56; **Petroleum**, 198; **Red**
stance, 62; **Reversible reaction**

Buckminsterfullerene

A member of the family of **fullerenes** – spherical crystalline forms of carbon made by condensing vaporized **graphite** in helium. Buckminsterfullerene also occurs naturally in dust between stars and in some carbon-rich rocks. Each molecule has 60 atoms arranged in hexagons and pentagons. Double **covalent bonds*** join hexagons to hexagons. (Other fullerenes contain between 30 and 960 carbon atoms.) Buckminsterfullerene is an **insulator***, but some of its compounds are **superconductors** (substances which have no electrical **resistance***).

The crystal structure of buckminsterfullerene

*All molecules of buckminsterfullerene are **icosahedral** – they have 20 hexagons and 12 pentagons.*

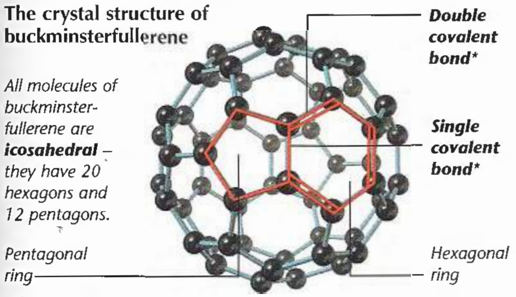

Coal

A hard, black solid formed over millions of years from the fossilized remains of plant material. It is mainly carbon but contains hydrogen, oxygen, nitrogen and sulfur as well. Three types of coal exist – **lignite**, **anthracite** and **bituminous coal**. Coal is used as a fuel in power stations, industry and homes. It was once an important source of chemicals (now mostly produced from **petroleum***). Heating coal in the absence of air (**destructive distillation**) produces **coal gas**, **coal tar** and **coke**, as well as ammonia, benzene and sulfur. Coke, which is brittle and porous, contains over 80% carbon and is used as a smokeless fuel (as is **charcoal**, another impure form of carbon).

Carbon fibers

Black, silky threads of pure carbon made from organic textile fibers. They are stronger and stiffer than other materials of the same weight, and are used to make light boats.

Carbon dioxide (CO_2)

A colorless, odorless gas found in the atmosphere (see **carbon cycle**, page 209). It is made industrially by heating calcium carbonate in a lime kiln (see also page 216 for laboratory preparation). It dissolves in water to form **carbonic acid** (**H_2CO_3**).

$$CO_2(aq) + H_2O(l) \rightleftharpoons 2H^+(aq) + CO_3^{2-}(aq)$$

Carbon dioxide *Water* ***Reversible reaction**** *Carbonic acid*

Carbon dioxide is not very reactive, though it reacts with both sodium and calcium hydroxide solutions (see page 218) and magnesium ribbon burns in it.

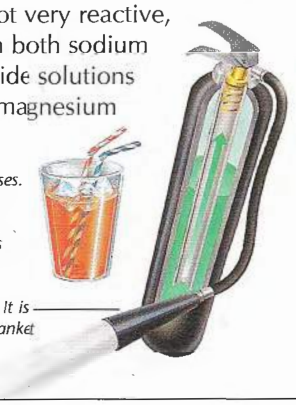

***Carbon dioxide** has many uses. It is used to make drinks fizzy. Carbon dioxide escapes when the bottle or can is opened, as the pressure is released.*

It is used in fire extinguishers. It is denser than air, so forms a blanket over the flames and does not allow air to reach the fire.

Carbon monoxide (CO)

A poisonous, colorless, odorless gas, made by passing **carbon dioxide** over hot carbon, and also by burning carbon fuels in a limited supply of air. It is not water-soluble, burns with a blue flame and is a **reducing agent*** (used to reduce metal oxide ores to metal – see **iron**, page 174). It is also used, mixed with other gases, in fuels, e.g. mixed with hydrogen in **water gas**, with nitrogen in **producer gas**, and with hydrogen (50%), methane and other gases in **coal gas**.

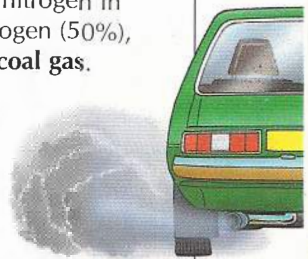

*If there is not enough oxygen, the **carbon monoxide** produced when fuel is burned is not changed to **carbon dioxide**. When a car engine runs in a closed garage, carbon monoxide accumulates.*

Carbonates

Compounds made of a metal **cation*** and a **carbonate anion*** (**CO_3^{2-}**), e.g. **calcium carbonate** (**$CaCO_3$**). Except Group I carbonates, they are insoluble in water and decompose upon heating. They all react with acids to give off **carbon dioxide**.

---

* **Amorphous**, 135; **Anion**, **Cation**, 130; **Covalent bond**, 132; **Insulator**, 56; **Petroleum**, 198; **Reducing agent**, 148; **Resistance**, 62; **Reversible reaction**, 162.

GROUP V ELEMENTS

The elements in **Group V** of the **periodic table*** become increasingly metallic going down the group (see chart below).

Some properties of Group V elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Nitrogen	N	14.00	2,5	INCREASING ↓	Colorless gas	See below.
Phosphorus	P	30.97	2,8,5		Non-metallic solid (see page 182)	See page 182.
Arsenic	As	74.92	Complex configuration but still five outer electrons		Three **allotropes*** (one is metallic)	In **semiconductors*** and **alloys***
Antimony	Sb	121.75			Silver-white metal	In type metal and other alloys
Bismuth	Bi	208.98			White metal with reddish tinge	In low melting point alloys and medicines

More information on **nitrogen**, **phosphorus** and their compounds can be found below and on pages 181-182. They are the two most abundant members of the group.

All the atoms of Group V elements have five electrons in their ***outer shell****. They all react to form* ***covalent compounds**** *in which they share three of these electrons with three from another atom, or atoms (see* ***octet****, page 127). Antimony, bismuth and nitrogen also form* ***ionic compounds****.*

Nitrogen (N_2)

A member of Group V of the periodic table. A colorless, odorless, **diatomic*** gas that makes up 78% of the atmosphere. It can be produced by **fractional distillation of liquid air*** (but see also page 217). Its **oxidation state*** in compounds varies from –3 to +5. It reacts with a few reactive materials to form **nitrides**.

$$6Li(s) + N_2(g) \rightarrow 2Li_3N(s)$$

Lithium *Nitrogen* *Lithium nitride*

Nitrogen is essential for all organisms as it is found in molecules in living cells, e.g. proteins (see also **nitrogen cycle**, page 209). It is used in the manufacture of ammonia (see **Haber process**, right) and nitric acid. **Liquid nitrogen**, which exists below –196°C, has many uses, including freezing food.

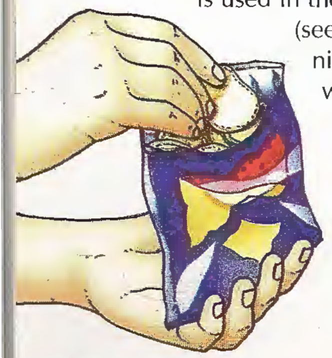

Packages of chips are filled with ***nitrogen*** *gas to keep them fresh longer (when air is left in the package, the chips go stale). The gas in the package also cushions the chips against damage during transportation.*

Haber process

This process is used to make **ammonia** from **nitrogen** and hydrogen which are reacted in a ratio of 1:3. Ammonia is produced as fast and economically as possible by using a suitable temperature, pressure and **catalyst*** (see below). The reaction is **exothermic*** and **reversible***.

$$N_2(g) + 3H_2(g) \underset{\substack{400°C \\ 250\text{ atmospheres}}}{\overset{\text{Iron catalyst*}}{\rightleftharpoons}} 2NH_3(g)$$

Nitrogen *Hydrogen* *Ammonia*

(Under these conditions 15% of the reactants combine to form ammonia.)

Haber process

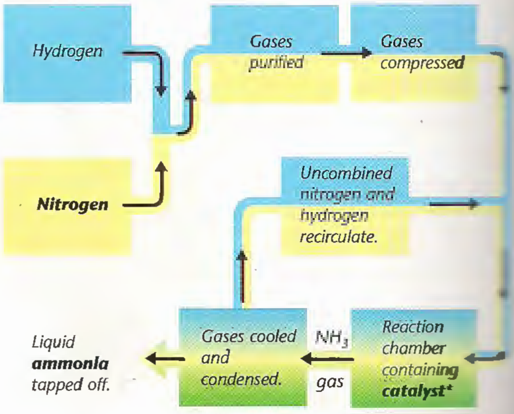
Hydrogen
Gases purified
Gases compressed
Nitrogen
Uncombined nitrogen and hydrogen recirculate.
Liquid ammonia tapped off.
Gases cooled and condensed.
NH_3 gas
Reaction chamber containing catalyst*

---

* **Allotropes**, 136 (**Allotropy**); **Alloy**, 344; **Catalyst**, 161; **Covalent compounds**, 132; **Diatomic**, 124; **Electron configuration**, **Exothermic reaction**, 146; **Fractional distillation of liquid air**, 183; **Ionic compound**, 131; **Outer shell**, 127; **Oxidation state**, 149; **Periodic table**, 164; **Relative atomic mass**, 138; **Reversible reaction**, 162; **Semiconductor**, 65.

Ammonia (NH_3)

A colorless, strong
dense than air and
made by the **Haber**
agent* and the only
alkaline solution in
known as **ammonia**
ammonium hydroxide
pure oxygen to give
water, and reacts w
give **ammonium chlo**

Ammonia *is used to make nitric acid, fertilizers, explosi household cleaners and plastics.*

Ammonium chloride
sal ammoniac

A white, water-solu
when **ammonia soluti**
with dilute hydrochl
sublimes* and **dissocia**
below and page 162
batteries which run

$$NH_4Cl(g) \underset{\text{Cool}}{\overset{\text{Heat}}{\rightleftharpoons}}$$

Ammonium chloride

Ammonium sulfate

A white, water-solub
produced by the reacti
sulfuric acid. It is a fer

Ammonium nitrate

A white, water-soluble crystalline solid formed when **ammonia solution** (see **ammonia**) reacts with dilute nitric acid. It gives off **dinitrogen oxide** when heated.

Ammonium nitrate *is used in explosives and fertilizers. It is also mixtures used to feed potted*

Covalent compounds, 132; **Oxidizing agent**, 148; **Red** **Reversible reaction**, 162; **Su**

Ammonia (NH_3)

A colorless, strong-smelling gas that is less dense than air and is a **covalent compound*** made by the **Haber process**. It is a **reducing agent*** and the only common gas to form an alkaline solution in water. This solution is known as **ammonia solution** (NH_4OH) or **ammonium hydroxide**. Ammonia burns in pure oxygen to give nitrogen and water, and reacts with chlorine to give **ammonium chloride**.

Ammonia *is used to make nitric acid, fertilizers, explosives, household cleaners and plastics.*

Ammonium chloride (NH_4Cl) or sal ammoniac

A white, water-soluble, crystalline solid made when **ammonia solution** (see **ammonia**) reacts with dilute hydrochloric acid. When heated it **sublimes*** and **dissociates*** (see equation below and page 162). It is used in the dry batteries which run many electrical appliances.

$$\underset{\text{Ammonium chloride}}{NH_4Cl(g)} \underset{\text{Cool}}{\overset{\text{Heat}}{\rightleftharpoons}} \underset{\text{Ammonia}}{NH_3(g)} + \underset{\text{Hydrogen chloride}}{HCl(g)}$$

Ammonium sulfate ($(NH_4)_2SO_4$)

A white, water-soluble, crystalline solid produced by the reaction of **ammonia** and sulfuric acid. It is a fertilizer.

Ammonium nitrate (NH_4NO_3)

A white, water-soluble, crystalline solid formed when **ammonia solution** (see **ammonia**) reacts with dilute nitric acid. It gives off **dinitrogen oxide** when heated.

Ammonium nitrate *is used in explosives and fertilizers. It is also found in mixtures used to feed potted plants.*

Dinitrogen oxide (N_2O)

Also called **nitrous oxide** or **laughing gas**. A colorless, slightly sweet-smelling, water-soluble gas. It is a **covalent compound*** formed by gently heating **ammonium nitrate**. It is used as an anaesthetic.

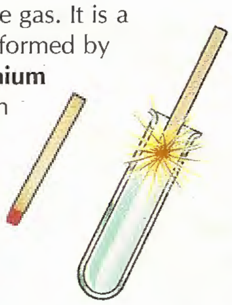

Dinitrogen oxide *supports the combustion of some reactive substances and relights a glowing splint.*

Nitrogen monoxide (NO)

Also called **nitric oxide** or **nitrogen oxide**. A colorless gas that is insoluble in water. It is a **covalent compound*** made when copper reacts with 50% concentrated nitric acid. It reacts with oxygen to form **nitrogen dioxide** and also supports the combustion of reactive elements.

Nitrogen dioxide (NO_2)

A very dark brown gas with a choking smell. It is a **covalent compound***.

$$\underset{\text{Copper}}{Cu} + \underset{\text{Concentrated nitric acid}}{4HNO_3} \rightarrow \underset{\text{Copper(II) nitrate}}{Cu(NO_3)_2} + \underset{\text{Water}}{2H_2O} + \underset{\text{Nitrogen dioxide}}{2NO_2}$$

Nitrogen dioxide is made when copper reacts with concentrated nitric acid and when some **nitrates*** are heated. It supports combustion and dissolves in water to give a mixture of nitric acid and **nitrous acid** (HNO_2). It is used as an **oxidizing agent***.

Nitrogen dioxide dimerizes *(two molecules of the same substance bond together) below 21.5°C to form* ***dinitrogen tetraoxide (N_2O_4)****, a colorless gas.*

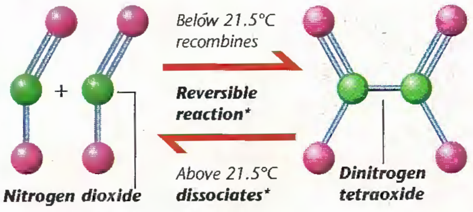

---

* **Covalent compounds**, 132; **Dissociation**, 162; **Nitrates**, 182; **Oxidizing agent**, 148; **Reducing agent**, 148; **Reversible reaction**, 162; **Sublimation**, 121.

Group V (continued)

Nitric acid (HNO_3) or nitric(V) acid

A light yellow, oily, water-soluble liquid. It is a **covalent compound*** containing nitrogen with an **oxidation state*** of +5. It is a very strong and corrosive acid which is made industrially by the three-stage **Ostwald process** (shown below).

Stage 1: Ammonia reacts with oxygen.

$$4NH_3 + 5O_2 \xrightarrow[\text{900°C}]{\text{Platinum-rhodium catalyst*}} 4NO + 6H_2O$$

Ammonia · *Oxygen* · *Nitrogen monoxide* · *Water*

Stage 2: Nitrogen monoxide cools and reacts with more oxygen to give nitrogen dioxide.

$$4NO + 2O_2 \rightarrow 4NO_2$$

Nitrogen monoxide · *Oxygen* · *Nitrogen dioxide*

*Stage 3: Nitrogen dioxide dissolves in water to form **nitric acid**.*

$$4NO_2 + 2H_2O + O_2 \rightarrow 4HNO_3$$

Nitrogen dioxide · *Water* · *Oxygen* · *Nitric acid*

Concentrated nitric acid is a mixture of 70% nitric acid and 30% water. It is a powerful **oxidizing agent***. **Dilute nitric acid** is a solution of 10% nitric acid in water. It reacts with **bases*** to give **nitrate salts*** and water. Nitric acid is used to make fertilizers and explosives.

Nitrates or nitrate(V) compounds

Sodium nitrate *(**$NaNO_3$**) is used to make gunpowder.*

Solid **ionic compounds*** containing the **nitrate anion*** (NO_3^-) and a metal **cation*** (see test for nitrate ion, page 218). Nitrogen in a nitrate ion has an **oxidation state*** of +5. Nitrates are **salts*** of **nitric acid** and are made by adding a metal oxide, hydroxide or carbonate to dilute nitric acid. All nitrates are water-soluble and most give off nitrogen dioxide and oxygen on heating (some exceptions are sodium, potassium and ammonium nitrates).

***Sodium** and **ammonium nitrates** are used as fertilizers.*

Nitrites or nitrate(III) compounds

Solid **ionic compounds*** that contain the **nitrite anion*** (NO_2^-) and a metal **cation***. They are usually **reducing agents***.

Phosphorus (P)

A non-metallic member of Group V (see chart, page 180). Phosphorus only occurs naturally in compounds. Its main ore is **apatite** ($3Ca_3(PO_4)_2.CaF_2$). It has two common forms. **White phosphorus**, the most reactive form, is a poisonous, waxy, white solid that bursts into flames in air. **Red phosphorus** is a dark red powder that is not poisonous and not very flammable.

*The minerals **apatite** (left) and turquoise (right) contain phosphorus.*

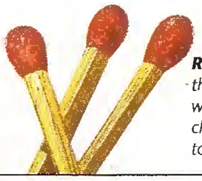

Phosphorus pentoxide (P_2O_5)

A white solid and **dehydrating agent***, made by burning **phosphorus** in air. It reacts vigorously with water to form **phosphoric acid** (H_3PO_4) and is used to protect against **rust***.

*Living organisms such as plants contain **phosphorus** compounds which promote healthy growth.*

***Red phosphorus** on the tip of a match reacts when struck against the chemicals on a matchbox to produce a flame.*

---

*Anion, 130; **Base**, 151; **Catalyst**, 161; **Cation**, 130; **Covalent compounds**, 132; **Dehydrating agent**, 344; **Ionic compound**, 131; **Oxidation state**, 149; **Oxidizing agent**, **Reducing agent**, 148; **Rust**, 174; **Salts**, 153.

GROUP

The elements
properties and
chart below sh

Some properties of	
Name of element	Chemical symbol
Oxygen	O
Sulfur	S
Selenium	Se
Tellurium	Te
Polonium	Po

More information
sulfur and their com
be found below and
184-185. They are
and have many uses.

Oxygen (O_2)

A colorless, odorless,
makes up 21% of the
most abundant element
and is vital for life (see
page 209). It supports
in water to form a **neutra**
a very reactive **oxidizing**
oxidizes iron to iron(II
produce oxygen by **pho**
it is obtained industria
distillation of liquid air.
e.g. in hospitals and to
See preparation of, and
pages 217 and 218.

Fractional distillation of liquid

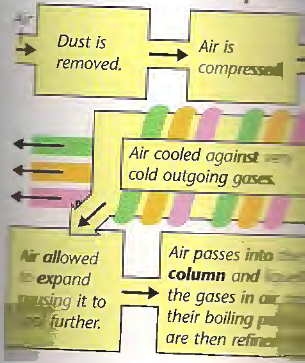

Acidic, 150; **Allotropes**, 136; **Alloy**,
figuration, 127; **Ionic compound**
Periodic table, 164; **Photocell**,

GROUP VI ELEMENTS

The elements in **Group VI** of the **periodic table*** show increasing metallic properties and decreasing chemical reactivity going down the group. The chart below shows some of the properties of these elements.

Some properties of Group VI elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Reactivity	Appearance	Uses
Oxygen	O	15.99	2,6	DECREASING ↓	Colorless gas (see below)	See below.
Sulfur	S	32.06	2,8,6		Yellow, non-metallic solid (see page 184)	See page 184.
Selenium	Se	78.96	Complex configuration but still six outer electrons		Several forms, metallic and non-metallic	In **photocells***
Tellurium	Te	127.60			Silver-white **metalloid*** solid	In **alloys***, colored glass, **semiconductors***
Polonium	Po	Radioactive* element			Metal	

More information on **oxygen**, **sulfur** and their compounds can be found below and on pages 184-185. They are found widely and have many uses.

*The atoms of all the elements in Group VI have six electrons in their **outer shell***. They need two electrons to fill their outer shell (see **octet**, page 127) and react with other substances to form both **ionic** and **covalent compounds***. The elements with the smallest atoms are most reactive as the atoms produce the most powerful attraction for the two electrons.*

Oxygen (O_2)

A colorless, odorless, **diatomic*** gas that makes up 21% of the atmosphere. It is the most abundant element in the Earth's crust and is vital for life (see **internal respiration**, page 209). It supports combustion, dissolves in water to form a **neutral*** solution and is a very reactive **oxidizing agent***, e.g. it oxidizes iron to iron(III) oxide. Plants produce oxygen by **photosynthesis*** and it is obtained industrially by **fractional distillation of liquid air**. It has many uses, e.g. in hospitals and to break down sewage. See preparation of, and test for, oxygen, on pages 217 and 218.

Fractional distillation of liquid air (see also page 220)

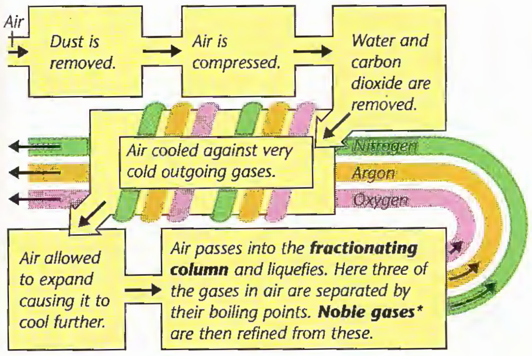

Dust is removed.

Air is compressed.

Water and carbon dioxide are removed.

Air cooled against very cold outgoing gases.

Air allowed to expand causing it to cool further.

*Air passes into the **fractionating column** and liquefies. Here three of the gases in air are separated by their boiling points. **Noble gases*** are then refined from these.*

Ozone (O_3)

A poisonous, bluish gas made of molecules which contain three **oxygen** atoms. It is an **allotrope*** of oxygen found in the upper atmosphere where it absorbs most of the Sun's harmful ultraviolet radiation (but see **ozone depletion**, page 210). It is produced when electrical sparks pass through air, e.g. when lightning occurs. Ozone is a powerful **oxidizing agent*** and is sometimes used to sterilize water.

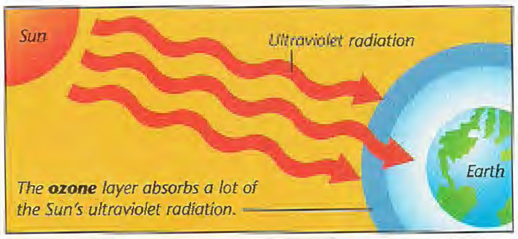

*The **ozone** layer absorbs a lot of the Sun's ultraviolet radiation.*

Oxides

Compounds of **oxygen** and one other element. Metal oxides are mostly **ionic compounds*** and **bases***, e.g. **calcium oxide** (**CaO**). Some metal and **metalloid*** oxides are **amphoteric***, e.g. **aluminum oxide** (Al_2O_3). Non-metal oxides are **covalent*** and often **acidic***, e.g. **carbon dioxide** (CO_2).

---

* **Acidic**, 150; **Allotropes**, 136; **Alloy**, 344; **Amphoteric**, **Base**, 151; **Covalent compounds**, 132; **Diatomic**, 124; **Electron configuration**, 127; **Ionic compound**, 131; **Metalloids**, 165; **Neutral**, 151; **Noble gases**, 189; **Outer shell**, 127; **Oxidizing agent**, 148; **Periodic table**, 164; **Photocell**, 345; **Photosynthesis**, 254; **Radioactivity**, 128; **Relative atomic mass**, 138; **Semiconductor**, 65.

SULFUR

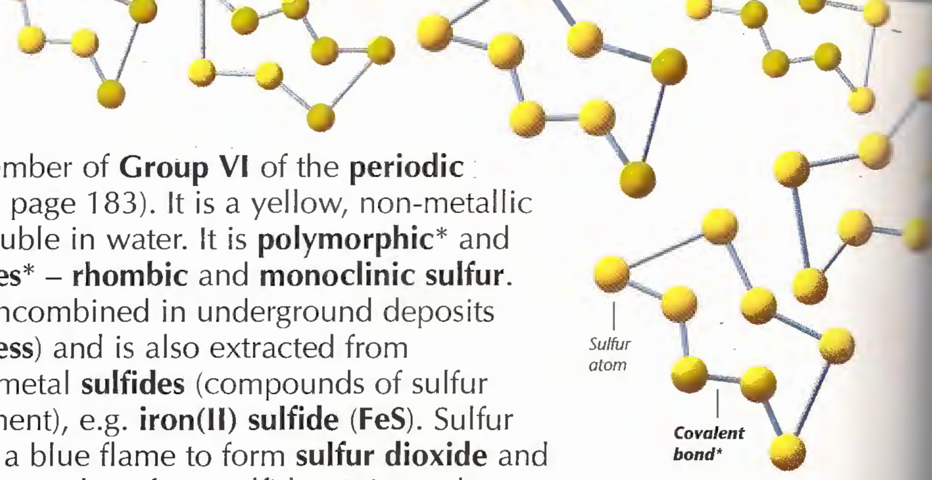

Both **rhombic** and **monoclinic sulfur** are made of puckered rings of eight sulfur atoms.

Sulfur (S) is a member of **Group VI** of the **periodic table*** (see chart, page 183). It is a yellow, non-metallic solid that is insoluble in water. It is **polymorphic*** and has two **allotropes*** – **rhombic** and **monoclinic sulfur**. Sulfur is found uncombined in underground deposits (see **Frasch process**) and is also extracted from **petroleum*** and metal **sulfides** (compounds of sulfur and another element), e.g. **iron(II) sulfide** (FeS). Sulfur burns in air with a blue flame to form **sulfur dioxide** and reacts with many metals to form sulfides. It is used to **vulcanize*** rubber, and to make **sulfuric acid**, medicines and **fungicides***.

Rhombic sulfur

Also called **alpha sulfur** (**α-sulfur**) or **orthorhombic sulfur**. A pale yellow, crystalline **allotrope*** of sulfur, the most stable form at room temperature.

Crystal of **rhombic sulfur** (see also page 136)

Molecular lattice* of sulfur rings (arranged differently from **monoclinic sulfur**)

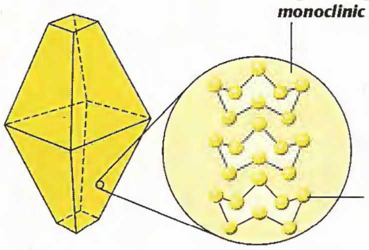

Rhombic sulfur is denser than monoclinic sulfur as the sulfur rings are more tightly packed together.

Monoclinic sulfur or beta sulfur (β-sulfur)

A yellow, crystalline **allotrope*** of sulfur. It is more stable than **rhombic sulfur** at temperatures over 96°C.

Crystal of **monoclinic sulfur** (long, thin and angular)

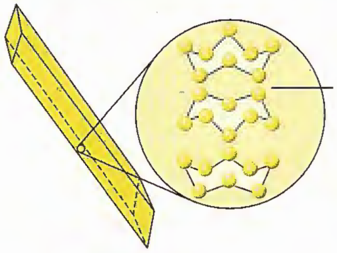

The sulfur rings are arranged in a **molecular lattice***, but in a different way from **rhombic sulfur**.

The allotropes* of sulfur

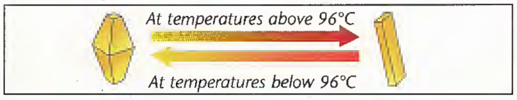

At temperatures above 96°C

At temperatures below 96°C

Plastic sulfur

A form of sulfur made when hot liquid sulfur is poured into water to cool it quickly. It can be kneaded and stretched into long fibers. It is not stable and hardens when rings of eight sulfur atoms reform (see above).

Flowers of sulfur

A fine, yellow powder formed when sulfur vapor is cooled quickly. The molecules are in rings of eight atoms.

Frasch process

The method used to extract sulfur from underground deposits by melting it. Sulfur produced this way is 99.5% pure.

Frasch process

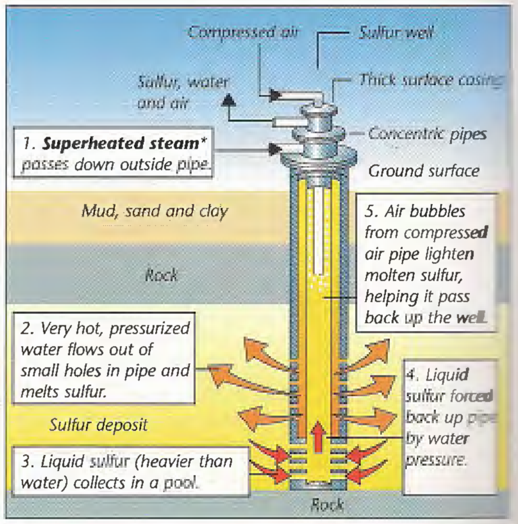

1. **Superheated steam*** passes down outside pipe.

2. Very hot, pressurized water flows out of small holes in pipe and melts sulfur.

3. Liquid sulfur (heavier than water) collects in a pool.

4. Liquid sulfur forced back up pipe by water pressure.

5. Air bubbles from compressed air pipe lighten molten sulfur, helping it pass back up the well.

---

* **Allotropes**, 136 (**Allotropy**); **Covalent bond**, 132; **Fungicide**, 344; **Molecular lattice**, 137; **Periodic table**, 164; **Petroleum**, 198; **Polymorphism**, 136; **Superheated steam**, **Vulcanization**, 345.

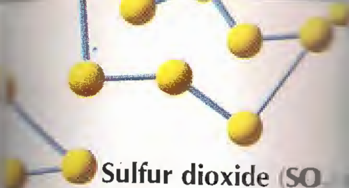

Sulfur dioxide (SO
A poisonous, choki
sulfurous acid when
It is a **covalent** comp
sulfur in air or add
It usually acts as a
It is used to make
fumigation*, as a blea
preservative for fru

Sulfur dioxide is used as an insecticide.

Sulfur trioxide (SO
A white, **volatile*** so
contact process (see
trioxide reacts very
form **sulfuric acid.**

Sulfurous acid (H
acid
A colorless, **weak aci**
dioxide dissolves in

Hydrogen sulfide (H
A colorless, poisono
eggs. It dissolves in
It is given off when o
when a dilute acid is

Sulfates or **sulfate(**
Solid **ionic compoun**
sulfate ion (SO_4^{2-}) and
occur naturally, e.g.
They are **salts*** of **sulfu**
adding **bases*** to dilute
sulfuric acid.

Sodium sulfate solution is used
"fix" photographs. This process
prints from going completely
when exposed to light.

Sulfites or **sulfate(IV)**
Ionic compounds* con
(SO_3^{2-}) and a metal **cati**
(Na_2SO_3). They are **salts**
and react with dilute **stro**
sulfur dioxide.

* **Base**, 151; **Bleach**, 344; **Catal**
Dibasic, 153; **Electrochemica**
Oxidizing agent, **Reducing**

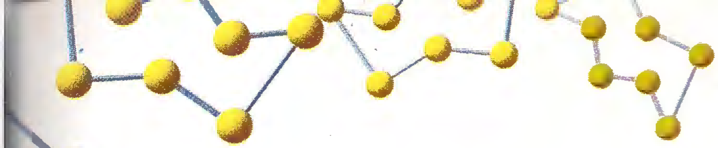

Sulfur dioxide (SO_2) or sulfur(IV) oxide

A poisonous, choking gas which forms **sulfurous acid** when dissolved in water. It is a **covalent compound*** made by burning sulfur in air or adding dilute acid to a **sulfite**. It usually acts as a **reducing agent***. It is used to make **sulfuric acid**, in **fumigation***, as a **bleach*** and as a preservative for fruit.

__Sulfur dioxide__ is used as an insecticide.

Sulfur trioxide (SO_3) or sulfur(VI) oxide

A white, **volatile*** solid that is formed by the **contact process** (see below, right). Sulfur trioxide reacts very vigorously with water to form **sulfuric acid**.

Sulfurous acid (H_2SO_3) or sulfuric(IV) acid

A colorless, **weak acid***, formed when **sulfur dioxide** dissolves in water.

Hydrogen sulfide (H_2S)

A colorless, poisonous gas, which smells like bad eggs. It dissolves in water to form a **weak acid***. It is given off when organic matter rots and when a dilute acid is added to a metal sulfide.

Sulfates or sulfate(VI) compounds

Solid **ionic compounds*** that contain a **sulfate ion** (SO_4^{2-}) and a **cation***. Many occur naturally, e.g. **calcium sulfate** ($CaSO_4$). They are **salts*** of **sulfuric acid**, made by adding **bases*** to dilute sulfuric acid.

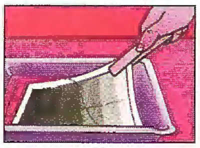

__Sodium sulfate__ solution is used to "fix" photographs. This process stops prints from going completely black when exposed to light.

Sulfites or sulfate(IV) compounds

Ionic compounds* containing a **sulfite ion** (SO_3^{2-}) and a metal **cation*** e.g. **sodium sulfite** (Na_2SO_3). They are **salts*** of **sulfurous acid**, and react with dilute **strong acids***, giving off **sulfur dioxide**.

Sulfuric acid (H_2SO_4) or sulfuric(VI) acid

An oily, colorless, corrosive liquid. It is a **dibasic*** acid, made by the **contact process** (see below). **Concentrated sulfuric acid** contains about 2% water, is **hygroscopic*** and a powerful **oxidizing** and **dehydrating agent***. **Dilute sulfuric acid**, a **strong acid***, contains about 90% water. It reacts with metals above hydrogen in the **electrochemical series*** to give the metal **sulfate** and hydrogen.

*__Concentrated sulfuric acid__ is an __oxidizing agent*__.*

$$Cu + 2H_2SO_4 \rightarrow CuSO_4 + SO_2 + 2H_2O$$

Copper — Concentrated sulfuric acid — Copper(II) sulfate — Sulfur dioxide — Water

*__Dilute sulfuric acid__ reacts with a __base*__ to give a __sulfate__.*

$$CuO(s) + H_2SO_4(aq) \rightarrow CuSO_4(aq) + H_2O(l)$$

Copper(II) oxide — Dilute sulfuric acid — Copper(II) sulfate — Water

__Sulfuric acid__ is used to make many things e.g. fertilizers, man-made fibers, detergents and paints.

The reaction between __concentrated sulfuric acid__ and water is very violent. To avoid accidents when the two are mixed, the acid is always added slowly to the water and not vice versa.

Contact process

The industrial process used to make **sulfuric acid**.

Contact process

*Dry and pure __sulfur dioxide__ and air are passed over a __catalyst*__ of vanadium pentoxide at 450°C.*

$$2SO_2(g) + O_2(g) \rightarrow 2SO_3(g)$$

Sulfur dioxide — Oxygen — Sulfur trioxide

__Sulfur trioxide__ is formed.

__Sulfur trioxide__ is absorbed by __concentrated sulfuric acid__, and __fuming sulfuric acid__, or __oleum__, is formed.

$$SO_3 + H_2SO_4 \rightarrow H_2S_2O_7$$

Sulfur trioxide — Concentrated sulfuric acid — Fuming sulfuric acid

__Fuming sulfuric acid__ is diluted to form __sulfuric acid__.

$$H_2S_2O_7 + H_2O \rightarrow 2H_2SO_4$$

Fuming sulfuric acid — Water — Sulfuric acid

---

* **Base**, 151; **Bleach**, 344; **Catalyst**, 161; **Cation**, 130; **Covalent compounds**, 132; **Dehydrating agent**, 344; **Dibasic**, 153; **Electrochemical series**, 159; **Fumigation**, 344; **Hygroscopic**, 206; **Ionic compound**, 131; **Oxidizing agent**, **Reducing agent**, 148; **Salts**, 153; **Strong acid**, 152; **Volatile**, 345; **Weak acid**, 152.

GROUP VII, THE HALOGENS

The elements in **Group VII** of the **periodic table*** are called the **halogens**, and their compounds and ions are generally known as **halides**. Group VII members are all non-metals and their reactivity decreases going down the group – the chart below shows some of their properties. For further information on group members, see below and pages 187-188. The power of Group VII elements as **oxidizing agents*** decreases down the group. They can all oxidize the ions of any members below them in the group. For example, **chlorine** displaces both **bromide** and **iodide anions*** from solution by oxidizing them to **bromine** and **iodine** molecules respectively. Bromine only displaces iodide anions from solution, and iodine cannot displace any halide anions from solution.

$$2KI(aq) + Br_2(l) \rightarrow 2KBr(aq) + I_2(s)$$

Bromine** displaces **iodide anions from potassium iodide. Each iodide anion loses an electron (is **oxidized***) when it is displaced by a **bromide anion***.*

Some properties of Group VII elements

Name of element	Chemical symbol	Relative atomic mass*	Electron configuration*	Oxidizing* power	Reactivity	Appearance
Fluorine	F	18.99	2,7	DECREASING ↓	DECREASING ↓	Pale yellow-green gas
Chlorine	Cl	35.45	2,8,7			Pale green-yellow gas
Bromine	Br	79.91	2,8,18,7			Dark-red fuming liquid
Iodine	I	126.90	2,8,18,18,7			Non-metallic black-gray solid
Astatine	At	No stable isotope*				

*The atoms of all the elements in Group VII contain seven electrons in their **outer shell*** and they all react to form both **ionic** and **covalent compounds***. The elements at the top of the group form more ionic compounds than those further down the group.*

***Fluorine** is never used in school laboratories as it is very poisonous and attacks glass containers. **Chlorine**, **bromine** and **iodine** do not react with glass, but chlorine is very poisonous, and so are the gases given off by the other two.*

Fluorine (F_2)

A member of Group VII of the periodic table. It is a **diatomic*** gas, extracted from **fluorospar** (CaF_2) and **cryolite** (Na_3AlF_6). It is the most reactive member of the group and is a very powerful **oxidizing agent***. It reacts with almost all elements. See pictures for some examples of its uses.

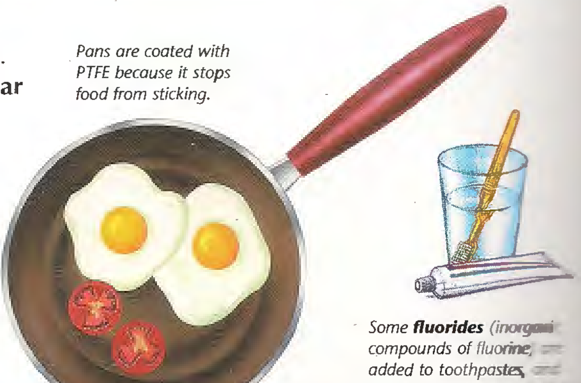

Pans are coated with PTFE because it stops food from sticking.

*Some **fluorides** (inorga compounds of fluorine, added to toothpastes, in some countries to water, to reduce tooth*

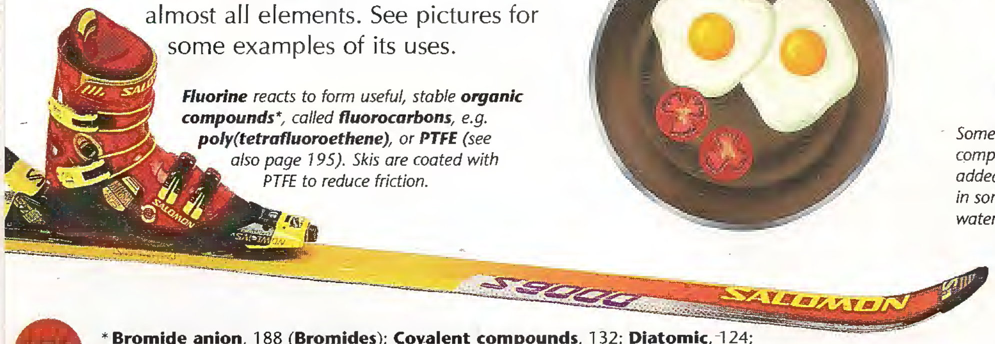

Fluorine** reacts to form useful, stable **organic compounds, called **fluorocarbons**, e.g. **poly(tetrafluoroethene)**, or **PTFE** (see also page 195). Skis are coated with PTFE to reduce friction.*

---

* **Bromide anion**, 188 (**Bromides**); **Covalent compounds**, 132; **Diatomic**, 124; **Electron configuration**, 127; **Iodide anion**, 188 (**Iodides**); **Ionic compound**, 131; **Isotope**, 127; **Organic compounds**, 190; **Outer shell**, 127; **Oxidation**, **Oxidizing agent**, 148; **Periodic table**, 164; **Relative atomic** mass

---

Chlorine (Cl_2)

A member of Grou
A poisonous, cho
very reactive and
compounds. **Sodiu**
important compou
and brine. Chlorin
electrolysis*, using
sodium, page 168
216). It is a very stro
Many elements react
chlorides (see equat

Chlorine** gas reacts with sod to form **sodium chloride** – common salt. Although chlorine gas is poisonous and sodium extremely reactive, both chemicals lose these dangerous properties when they join together to form sodium chloride. In the laboratory, this reaction takes place inside a **fume cupboard, so the harmful gas does not escape.*

$2Na(s) + Cl_2$
Hydrogen Ch

Chlorine** has many uses. It is (see **hydrogen chloride**), some **germicide in swimming pools. drinking water and disinfectants.*

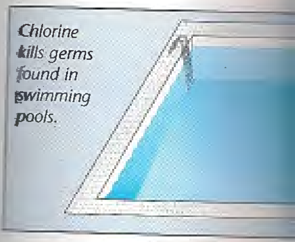

Chlorine kills germs found in swimming pools.

Sodium hypochlorite
or sodium chlorate(I)

A crystalline, white soli
stored dissolved in
water, and formed when
chlorine is added to a
cold, dilute sodium
hydroxide solution. It is
used in domestic **bleach***
and also to bleach paper
pulp white for writing.

---

* **Anion**, 130; **Base**, 151; **Bleach**,
Electrochemical series, 159; **Elec**
Germicide, 344; **Ionic compo**

Chlorine (Cl_2)

A member of Group VII of the periodic table. A poisonous, choking **diatomic*** gas which is very reactive and only occurs naturally in compounds. **Sodium chloride** (**NaCl**), its most important compound, is found in rock salt and brine. Chlorine is extracted from brine by **electrolysis***, using the **Downs' cell** (see **sodium**, page 168 and also **chlorine**, page 216). It is a very strong **oxidizing agent***. Many elements react with chlorine to form **chlorides** (see equation below).

*__Chlorine__ gas reacts with sodium to form __sodium chloride__ – common salt. Although chlorine gas is poisonous and sodium is extremely reactive, both chemicals lose these dangerous properties when they join together to form sodium chloride. In the laboratory, this reaction takes place inside a __fume cupboard*__, so the harmful gas does not escape.*

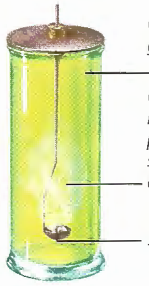
Chlorine gas
Cloud of minute pieces of sodium chloride
Sodium

$2Na(s)$	+	$Cl_2(g)$	→	$2NaCl(s)$
Hydrogen		*Chlorine*		*Hydrogen chloride*

*__Chlorine__ has many uses. It is used to make __hydrochloric acid__ (see __hydrogen chloride__), some organic solvents and also as a __germicide*__ in swimming pools. It is also used as a germicide in drinking water and disinfectants.*

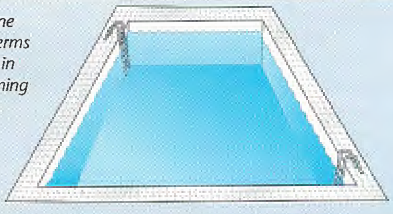

Chlorine kills germs found in swimming pools.

Sodium hypochlorite (**NaOCl**) or **sodium chlorate(I)**

A crystalline, white solid, stored dissolved in water, and formed when **chlorine** is added to a cold, dilute sodium hydroxide solution. It is used in domestic **bleach*** and also to bleach paper pulp white for writing.

Bleached writing paper

Chlorides

Compounds formed when **chlorine** combines with another element. Chlorides of non-metals (see **hydrogen chloride**) are **covalent compounds***, usually liquids or gases. Chlorides of metals, e.g. **sodium chloride** (**NaCl**), are usually solid, water-soluble, **ionic compounds*** made of a **chloride anion*** (**Cl^-**) and metal **cation***. See also page 218.

Hydrogen chloride (**HCl**)

A colorless, **covalent*** gas that forms ions when dissolved in a **polar solvent***. It is made by burning hydrogen in **chlorine**. It reacts with ammonia and dissolves in water to form **hydrochloric acid**, a **strong acid***. **Concentrated hydrochloric acid**, 35% hydrogen chloride and 65% water, is a fuming, corrosive and colorless solution. **Dilute hydrochloric acid**, about 7% hydrogen chloride and 93% water, is a colorless solution that reacts with **bases***, and with metals above hydrogen in the **electrochemical series***. Concentrated hydrochloric acid is used industrially to remove rust from steel sheets before they are **galvanized***.

__Concentrated hydrochloric acid__ is used to etch metals.

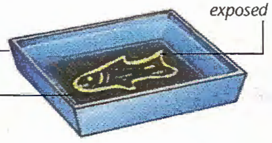
Bath of **concentrated hydrochloric acid**
Line of metal exposed to acid.
Resin* covering metal.

The metal exposed to the acid is eaten away, leaving a groove in the surface. When printing a picture, the groove is filled with ink.

Sodium chlorate ($NaClO_3$) or **sodium chlorate(V)**

A white, crystalline solid, formed when **chlorine** is added to warm concentrated sodium hydroxide, and also when **sodium hypochlorite** is warmed.

__Sodium chlorate__ kills weeds.

---

* **Anion**, 130; **Base**, 151; **Bleach**, 344; **Cation**, 130; **Covalent compounds**, 132; **Diatomic**, 124; **Electrochemical series**, 159; **Electrolysis**, 156; **Fume cupboard**, 224; **Galvanizing**, 174 (**Rust**); **Germicide**, 344; **Ionic compound**, 131; **Oxidizing agent**, 148; **Polar solvent**, 144; **Resins**, 345; **Strong acid**, 152.

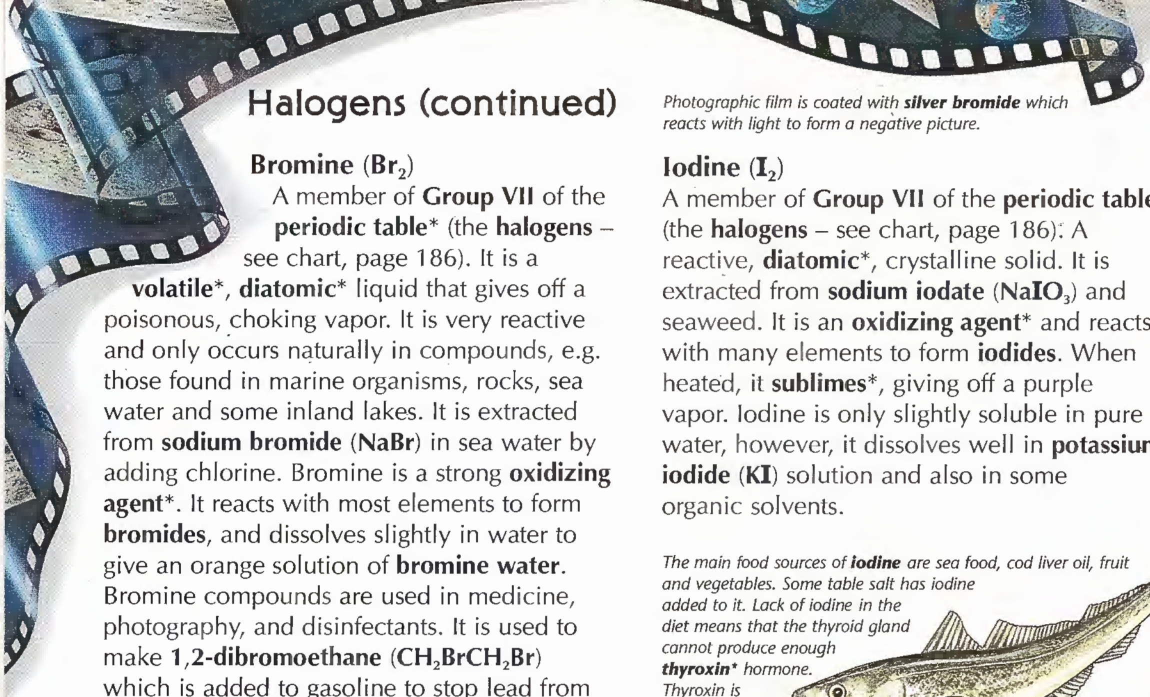

Photographic film is coated with **silver bromide** which reacts with light to form a negative picture.

Halogens (continued)

Bromine (Br_2)

A member of **Group VII** of the **periodic table*** (the **halogens** – see chart, page 186). It is a **volatile***, **diatomic*** liquid that gives off a poisonous, choking vapor. It is very reactive and only occurs naturally in compounds, e.g. those found in marine organisms, rocks, sea water and some inland lakes. It is extracted from **sodium bromide** (**NaBr**) in sea water by adding chlorine. Bromine is a strong **oxidizing agent***. It reacts with most elements to form **bromides**, and dissolves slightly in water to give an orange solution of **bromine water**. Bromine compounds are used in medicine, photography, and disinfectants. It is used to make **1,2-dibromoethane** (**CH_2BrCH_2Br**) which is added to gasoline to stop lead from accumulating in engines.

Bromides

Compounds of **bromine** and one other element. Bromides of non-metals are **covalent compounds*** (see **hydrogen bromide**). Bromides of metals are usually **ionic compounds*** made of **bromide anions*** (**Br^-**) and metal **cations***. Excepting **silver bromide** (**AgBr**), they are all water-soluble. See also page 218.

***Silver bromide** is used in photographic film. When exposed to light it decomposes to form silver.*

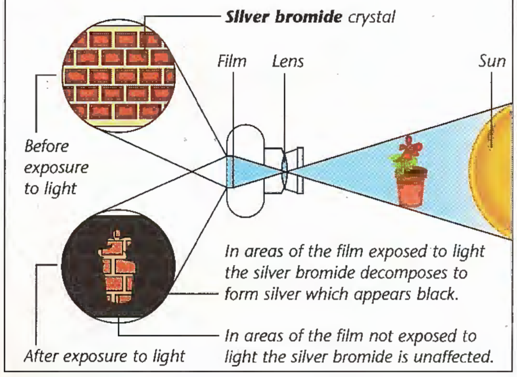
Silver bromide crystal
Film
Lens
Sun
Before exposure to light
In areas of the film exposed to light the silver bromide decomposes to form silver which appears black.
In areas of the film not exposed to light the silver bromide is unaffected.
After exposure to light

Hydrogen bromide (HBr)

A colorless, pungent-smelling gas, made by the reaction of **bromine** with hydrogen. Its chemical properties are similar to those of hydrochloric acid.

Iodine (I_2)

A member of **Group VII** of the **periodic table*** (the **halogens** – see chart, page 186). A reactive, **diatomic***, crystalline solid. It is extracted from **sodium iodate** (**$NaIO_3$**) and seaweed. It is an **oxidizing agent*** and reacts with many elements to form **iodides**. When heated, it **sublimes***, giving off a purple vapor. Iodine is only slightly soluble in pure water, however, it dissolves well in **potassium iodide** (**KI**) solution and also in some organic solvents.

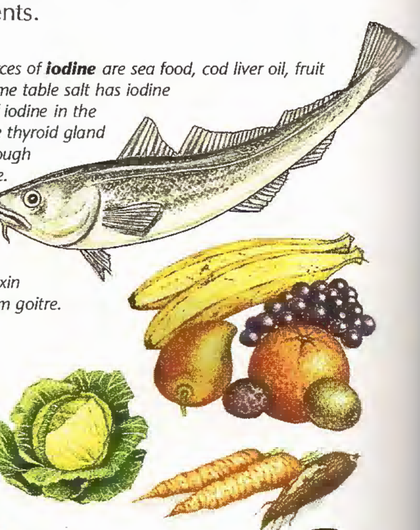

*The main food sources of **iodine** are sea food, cod liver oil, fruit and vegetables. Some table salt has iodine added to it. Lack of iodine in the diet means that the thyroid gland cannot produce enough **thyroxin*** hormone. Thyroxin is needed to regulate body metabolism. People with a thyroxin deficiency suffer from goitre.*

*Seaweed contains up to 0.5% **iodine** (by weight).*

***Tincture of iodine (iodine** dissolved in ethanol) is used as an antiseptic for cuts.*

Iodides

Compounds of **iodine** and one other element. Iodides of non-metals are **covalent compounds*** (see **hydrogen iodide**). Iodides of metals are usually **ionic***, made of **iodide anions*** (**I^-**) and metal **cations***. Except **silver iodide** (**AgI**), ionic iodides are water-soluble. See page 218.

Hydrogen iodide (HI)

A colorless gas with a pungent smell. It is a **covalent compound***, formed when hydrogen and **iodine** react. It dissolves in water to give a strongly **acidic*** solution called **hydroiodic acid** (its chemical properties are similar to those of hydrochloric acid).

*Acidic, 150; Anion, Cation, 130; Covalent compounds, 132; Diatomic, 124; Ionic compound, 131; Periodic table, 164; Oxidizing agent, 148; Sublimation, 121; Thyroxin, 336; Volatile, 345.

GROUP

The **noble gases**,
rare gases, make
the **periodic table**
Group 0. They
gases, obtained
distillation of liq
forms 0.9% of the
gases occur in e
They are all unre
atoms' **electron**
very stable (they
shell*). The light
form any compo
members form a

Helium (He)

The first member of
table. It is a colorless
gas found in the atm
200,000) and in some
USA. It is obtained
of liquid air* and is
having no known co
airships and balloons
dense than air and
also by deep-sea div
to avoid "the
bends".

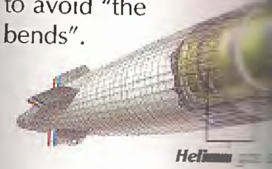

Neon (Ne)

A member of Group
periodic table. A col
odorless **monatomic***
found in the atmosp
part in 55,000). It is
by the **fractional disti**
of liquid air* and is
unreactive, having no
compounds. It is used
signs and fluorescen
as it emits an orange
when an electric disc
passes through it at lo
pressure.

***Electron configuration**,
Outer shell, 127; **Periodic**

GROUP VIII, THE NOBLE GASES

The **noble gases**, also called **inert** or **rare gases**, make up **Group VIII** of the **periodic table***, **also called Group 0**. They are all **monatomic*** gases, obtained by the **fractional distillation of liquid air***. **Argon** forms 0.9% of the air and the other gases occur in even smaller amounts. They are all unreactive because their atoms' **electron configuration*** is very stable (they all have a full **outer shell***). The lighter members do not form any compounds, but the heavier members form a few.

Helium (He)

The first member of Group VIII of the periodic table. It is a colorless, odorless, **monatomic*** gas found in the atmosphere (one part in 200,000) and in some natural gases in the USA. It is obtained by the **fractional distillation of liquid air*** and is completely unreactive, having no known compounds. It is used in airships and balloons, as it is eight times less dense than air and not inflammable, and also by deep-sea divers to avoid "the bends".

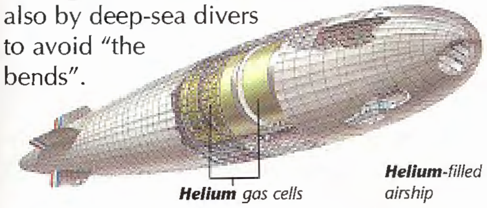
***Helium** gas cells*

***Helium**-filled airship*

Neon (Ne)

A member of Group VIII of the periodic table. A colorless, odorless **monatomic*** gas found in the atmosphere (one part in 55,000). It is obtained by the **fractional distillation of liquid air*** and is totally unreactive, having no known compounds. It is used in neon signs and fluorescent lighting as it emits an orange-red glow when an electric discharge passes through it at low pressure.

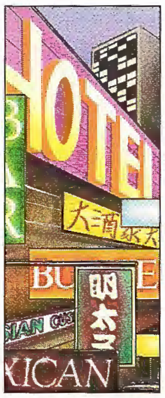

***Neon** signs*

Radon (Rn)

The last member of Group VIII of the periodic table. It is **radioactive***, occurring as a result of the **radioactive decay*** of radium.

Argon (Ar)

The most abundant member of Group VIII of the periodic table. It is a colorless, odorless, **monatomic*** gas that makes up 0.9% of the air. Obtained by the **fractional distillation of liquid air***, it is totally unreactive, having no known compounds. It is used in electric light bulbs and fluorescent tubes.

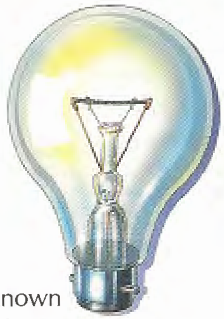
Electric light bulb

Krypton (Kr)

A member of Group VIII of the periodic table. It is a colorless, odorless, **monatomic*** gas found in the atmosphere (one part in 670,000). It is obtained by the **fractional distillation of liquid air*** and is unreactive, only forming one known compound, **krypton fluoride** (KrF_2). Krypton is used in some lasers and photographic flash lamps. It is also used in fluorescent tubes and in the stroboscopic lights which flank airport runways.

Xenon (Xe)

A member of Group VIII of the periodic table. A colorless, odorless **monatomic*** gas found in the atmosphere (0.006 parts per million). Obtained from the **fractional distillation of liquid air***, it is unreactive, forming only a very few compounds, e.g. **xenon tetrafluoride** (XeF_4). It is used to fill fluorescent tubes and light bulbs.

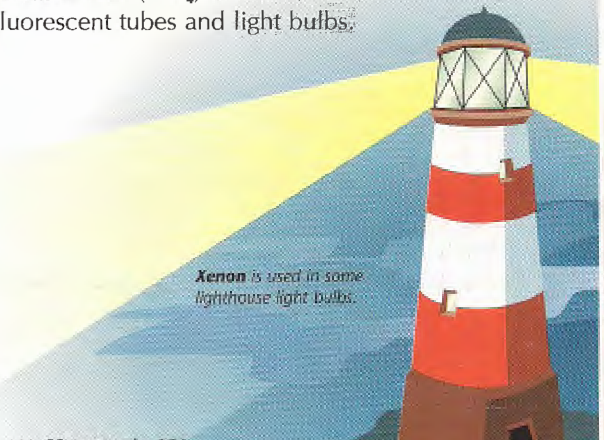
***Xenon** is used in some lighthouse light bulbs.*

---

* **Electron configuration**, 127; **Fractional distillation of liquid air**, 183; **Monatomic**, 124; **Outer shell**, 127; **Periodic table**, 164; **Radioactive decay**, **Radioactivity**, 128.

ORGANIC CHEMISTRY

Originally **organic chemistry** was the study of chemicals found in living organisms. However, it now refers to the study of all carbon-containing compounds, except the **carbonates*** and the **oxides*** of carbon. There are well over two million such compounds (**organic compounds**), more than all the other chemical compounds added together. This vast number of **covalent compounds*** is possible because carbon atoms can bond with each other to make a huge variety of **chains** and **rings**.

Some organic compounds are used to make paints.

Aliphatic compounds

Organic compounds whose molecules contain a **main chain** of carbon atoms. The chain may be **straight**, **branched** or even in **ring** form (though never a **benzene ring** – see **aromatic compounds**).

***Branched chain** of carbon atoms in a 3-methyl pentane molecule. In a branched chain, a carbon atom may be bonded to more than two other carbon atoms.*

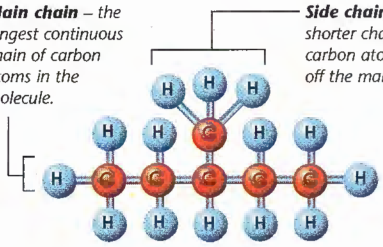

***Main chain** – the longest continuous chain of carbon atoms in the molecule.*

***Side chain** – a shorter chain of carbon atoms coming off the main chain.*

***Straight chain** of carbon atoms in a butan-1-ol molecule. No carbon atom is bonded to more than two other carbons.*

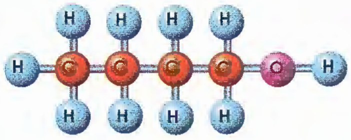

***Cyclohexane** molecule. An example of a molecule containing a ring of carbon atoms.*

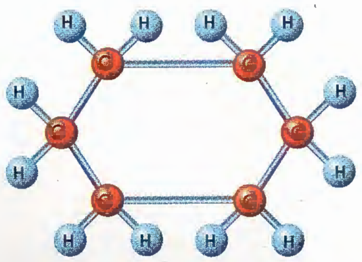

Aromatic compounds

Organic compounds whose molecules contain a **benzene ring**. A benzene ring has six carbon atoms but differs from an **aliphatic** ring because bonds between carbon atoms are neither **single** nor **double bonds*** but midway between, both in length and reactivity.

*There are two possible ways of representing a **benzene ring**.*

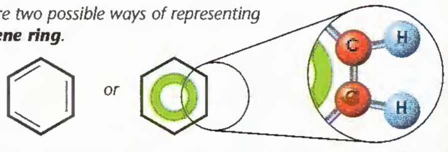

*The bonds linking the carbon atoms are midway between **single** and **double bonds*** because some electrons are free to move around the molecule.*

Hydrocarbons

Organic compounds that contain only carbon and hydrogen atoms.

*Two examples of **hydrocarbon** molecules:*

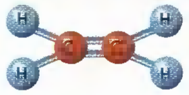

Ethene

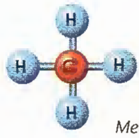

Methane

Functional group

An atom or group of atoms that gives a molecule most of its chemical properties. Organic molecules can have several such groups (see also pages 194-195).

*Most **functional groups** contain at least one atom that is not carbon or hydrogen.*

Ethanol molecule

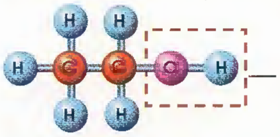

Functional group** of **alcohols, called a **hydroxyl group** (**–OH**)*

Ethene molecule

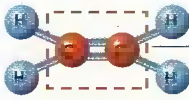

*Two carbon atoms joined by a **double** or **triple bond*** are also **functional groups**.*

---

* **Alcohols**, 196; **Carbonates**, 179; **Covalent compounds**, **Double bond**, 132; **Oxides**, 183; **Single bond**, **Triple bond**, 132.

Homologous series

A group of organic co
increase in size throu
a $-CH_2-$ group each
the **alkanes***) also have
e.g. the **alcohol*** hydr
Members of a series ha
properties but their ph
as they get larger. A ho
general formula for al

*The **general formula** for alcohols (where n stands for the number of*

First two members of the hom

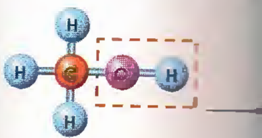

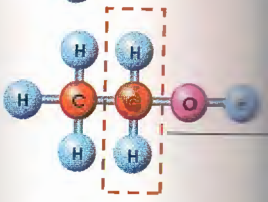

Saturated compounds

Organic compounds wh
have **single bonds*** be

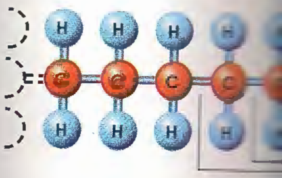

Unsaturated compoun

Organic compounds whos
least one **double** or **triple**

Polyunsaturated compo

A term used for compoun
have many **double** or trip
those found in soft marg

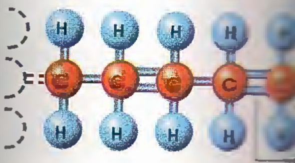

---

* **Alcohols**, 196; **Alkanes**, 192;
Stereochemical formula, **Struc**

Homologous series

A group of organic compounds which increase in size through the group by adding a $-CH_2-$ group each time. All series (except the **alkanes***) also have a **functional group**, e.g. the **alcohol*** **hydroxyl group** (**–OH**). Members of a series have similar chemical properties but their physical properties change as they get larger. A homologous series has a **general formula** for all its members.

*The **general formula** for **alcohols*** is $C_nH_{2n+1}OH$ (where n stands for the number of carbon atoms).*

First two members of the homologous series of alcohols*

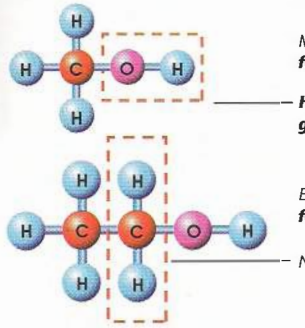

*Methanol (**structural formula*** CH_3OH)*

Hydroxyl functional group

*Ethanol (**structural formula*** CH_3CH_2OH)*

New $-CH_2-$ group added.

Saturated compounds

Organic compounds whose molecules only have **single bonds*** between atoms.

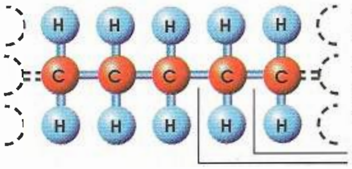

Part of a molecule found in butter

*Butter is a **saturated compound**.*

Single bonds only*

Unsaturated compounds

Organic compounds whose molecules have at least one **double** or **triple bond***.

Polyunsaturated compounds

A term used for compounds whose molecules have many **double** or **triple bonds***, e.g. those found in soft margarines.

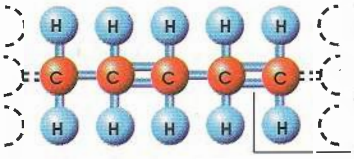

Part of a molecule found in margarine

*Margarines contain **polyunsaturated compounds**.*

Double bond*

Stereochemistry

The study of the 3-dimensional (3-D) structure of molecules. Comparing the 3-D structure of very similar organic molecules, e.g. **stereoisomers**, helps distinguish between them. The 3-D structure of a molecule is often shown by a **stereochemical formula*** – a diagram that shows how atoms are arranged in space.

Structural formula of methane*

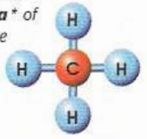

This simplified version of the molecule does not show the 3-D arrangement of the atoms.

Stereochemical formula of methane*

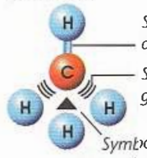

Symbol for bond along plane of page

Symbol for bond going into page

Symbol for bond coming out of page

The carbon-hydrogen bonds are arranged to form a tetrahedron.

Isomers

Two or more compounds with the same **molecular formula***, but different arrangements of atoms in their molecules. As a result, the compounds have different properties. There are two main types of isomer, **structural isomers** and **stereoisomers**.

Structural isomers

Compounds with the same **molecular formula***, but different **structural formulas***, i.e. the atoms are arranged in different ways.

*The **molecular formula*** C_2H_6O has two different **structural formulas***.*

Ethanol CH_3CH_2OH *Methoxymethane* CH_3OCH_3

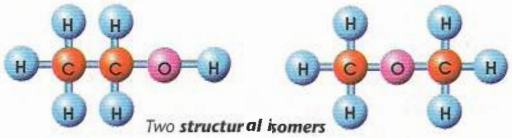

*Two **structural isomers***

Stereoisomers

Compounds with the same **molecular formula*** and grouping of atoms but a different 3-D appearance.

*The **molecular formula** C_4H_8 has two different **stereochemical formulas***.*

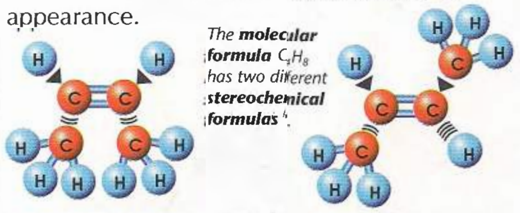

Cis *but-2-ene* *Two **stereoisomers*** Trans *but-2-ene*

---

* **Alcohols**, 196; **Alkanes**, 192; **Double bond**, 132; **Molecular formula**, 140; **Single bond**, 132; **Stereochemical formula**, **Structural formula** (**shortened**), 140; **Triple bond**, 132.

ALKANES

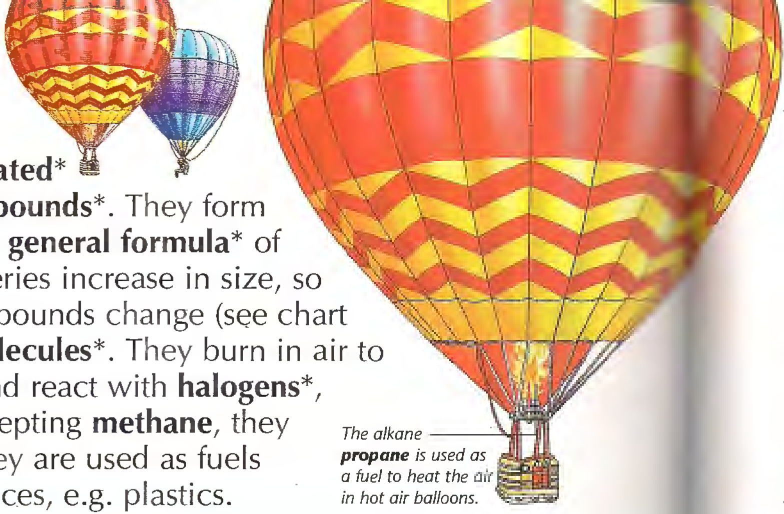

Alkanes, or **paraffins**, are all **saturated*** **hydrocarbons*** and **aliphatic compounds***. They form a **homologous series*** which has a **general formula*** of C_nH_{2n+2}. As the molecules in the series increase in size, so the physical properties of the compounds change (see chart below). Alkanes are **non-polar molecules***. They burn in air to form carbon dioxide and water, and react with **halogens***, otherwise they are unreactive. Excepting **methane**, they are obtained from **petroleum***. They are used as fuels and to make other organic substances, e.g. plastics.

The alkane ***propane*** *is used as a fuel to heat the air in hot air balloons.*

Some properties of alkanes				
Name of compound	Molecular formula*	Structural formula*	Physical state at 25°C	Boiling point (°C)
Methane	CH_4	CH_4	Gas	−161.5
Ethane	C_2H_6	CH_3CH_3	Gas	−88.0
Propane	C_3H_8	$CH_3CH_2CH_3$	Gas	−42.2
Butane	C_4H_{10}	$CH_3CH_2CH_2CH_3$	Gas	−0.5
Pentane	C_5H_{12}	$CH_3CH_2CH_2CH_2CH_3$	Liquid	36.0
Hexane	C_6H_{14}	$CH_3CH_2CH_2CH_2CH_2CH_3$	Liquid	69.0

The first part of the name indicates the number of carbon atoms in the molecule. The -ane ending means the molecule is an alkane (see page 214).

The next molecule in the series is always one $-CH_2-$ *group longer.*

Gradual change of state as molecules get longer.

The boiling points of the alkanes increase regularly as the molecules get longer. Melting points and densities follow the same trend, getting higher as the molecules increase in size.

Methane (CH_4)

The simplest alkane. It is a colorless, odorless, inflammable gas, which reacts with **halogens*** (see equation, below right) and is a source of hydrogen. **Natural gas** contains 99% methane.

Ethane (C_2H_6)

A member of the alkanes. A gas found in small amounts in **natural gas** (see **methane**), but mostly obtained from **petroleum***. Its properties are similar to those of methane. It is used to make other organic chemicals.

Propane (C_3H_8)

A member of the alkanes. A gas that is usually obtained from **petroleum***. Its properties are similar to **ethane**. It is bottled and sold as fuel for cooking and heating.

Alkanes are extracted from ***petroleum**** *and* ***natural gas*** *found deep under the ground.*

Cycloalkanes

Alkane molecules whose carbon atoms are joined in a ring, e.g. **cyclohexane** (see picture, page 190). Their properties are similar to those of other alkanes.

Substitution reaction

A reaction in which an atom or **functional group*** of a molecule is replaced by a different atom or functional group. The molecules of **saturated compounds***, e.g. alkanes, can undergo substitution reactions, but not **addition reactions** (see right).

Alkanes *react with* ***halogens**** *by undergoing a* ***substitution reaction****. Here is an example:*

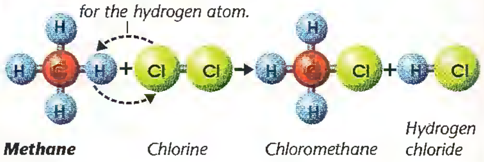

A chlorine atom is substituted for the hydrogen atom.

---

* **Aliphatic compounds**, 190; **Functional group**, 190; **General formula**, 191 (**Homologous series**); **Halogens**, 186; **Hydrocarbons**, 190; **Molecular formula**, 140; **Non-polar molecule**, 133 (**Polar molecule**, **Petroleum**, 198; **Saturated compounds**, 191; **Structural formula** (**shortened**), 140.

ALKENES

Alkenes, or **olefins**
compounds*. Alke
between carbon a
with the **general f**
physical propertie
molecules*. They
completely **oxidiz**
reactive than **alkar**
reactions, and son
and are used to m

Some properties of alkenes	
Name of compound	Molecu formul
Ethene	C_2H_4
Propene	C_3H
But-1-ene	C_4H
Pent-1-ene	C_5H

The number denotes the position of the ***double bond**** *in the molecule. Alkenes are named the same way as* ***alkanes****, but in -ene, not -ane (see page 21*

Ethene (C_2H_4) or ethy

The simplest alkene (see
colorless, sweet-smelli
addition reactions inc
polymerization* to form
known as **polythene**, (se
200). Ethene is used to
ethanol and many othe

Addition reaction

A reaction in which
together to produce a
One of the molecules
(have a **double** or **trip**

Addition reaction

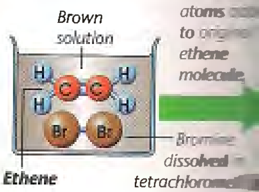

Brown solution
H
C
Br
Ethene
Two bromine atoms add to original ethene molecule
Bromine dissolved in tetrachlorome

The change of color is used as
compounds* *like alkenes.*

---

* **Addition polymerization**,
series); **Hydrocarbons**, 191
Saturated compounds, 191

ALKENES

Alkenes, or **olefins**, are **unsaturated*** **hydrocarbons*** and **aliphatic compounds***. Alkene molecules contain one or more **double bonds*** between carbon atoms. Those with only one form a **homologous series*** with the **general formula*** C_nH_{2n}. As the molecules increase in size, their physical properties change gradually (see below). Alkenes are **non-polar molecules***. They burn with a smoky flame and in excess oxygen are completely **oxidized*** to carbon dioxide and water. Alkenes are more reactive than **alkanes**, because of their double bond – they undergo **addition reactions**, and some form **polymers***. Alkenes are made by **cracking*** alkanes and are used to make many products including plastics and antifreeze.

Some properties of alkenes

Name of compound	Molecular formula*	Structural formula*	Physical state at 25°C	Boiling point (°C)
Ethene	C_2H_4	$CH_2{=}CH_2$	Gas	–104.0
Propene	C_3H_6	$CH_3CH{=}CH_2$	Gas	–47.0
But-1-ene	C_4H_8	$CH_3CH_2CH{=}CH_2$	Gas	–6.0
Pent-1-ene	C_5H_{10}	$CH_3CH_2CH_2CH{=}CH_2$	Liquid	30.0

*The number denotes the position of the **double bond*** in the molecule. Alkenes are named in the same way as **alkanes**, but end in -ene, not -ane (see page 214).*

*Each molecule is one –CH_2– group longer. The position of the **double bond*** is shown.*

Gradual change from gases to liquids to solids as the molecules get longer.

As the molecules get longer, the boiling points of the alkenes increase regularly. Melting points and densities follow the same trend.

Ethene (C_2H_4) or ethylene

The simplest alkene (see chart above) – it is a colorless, sweet-smelling gas which undergoes **addition reactions** including **addition polymerization*** to form **poly(ethene)**, commonly known as **polythene**, (see **homopolymer**, page 200). Ethene is used to make plastics, and also ethanol and many other organic chemicals.

Propene (C_3H_6) or propylene

A member of the alkenes. It is a colorless gas used to make propanone (also known as **acetone** – see **ketones**, page 194) and **poly(propene)**, also called **polypropylene**.

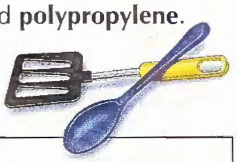

*Some kitchen tools are made from **poly(propene)**, the **polymer*** of **propene**.*

Addition reaction

A reaction in which two molecules react together to produce a single larger molecule. One of the molecules must be **unsaturated*** (have a **double** or **triple bond***).

Addition reaction

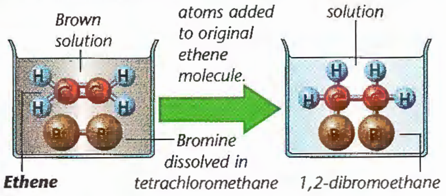
Brown solution
Two bromine atoms added to original ethene molecule.
Colorless solution
H
C
Br
Ethene
Bromine dissolved in tetrachloromethane
1,2-dibromoethane

*The change of color is used as a test for **unsaturated compounds*** like alkenes.*

Hydrogenation

An **addition reaction** in which hydrogen atoms are added to an **unsaturated compound*** molecule.

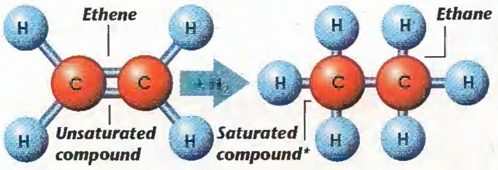

*This type of reaction is used in the margarine industry to harden animal and vegetable oils. (These oils are **unsaturated compounds***, but not alkenes).*

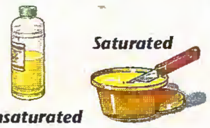

---

* **Addition polymerization**, 200; **Aliphatic compounds**, 190; **Cracking**, 198; **Double bond**, 132; **General formula**, 191 (**Homologous series**); **Hydrocarbons**, 191; **Molecular formula**, 140; **Non-polar molecule**, 133 (**Polar molecule**); **Oxidation**, 148; **Polymers**, 200; **Saturated compounds**, 191; **Structural formula** (shortened), 140; **Triple bond**, 132; **Unsaturated compounds**, 191.

ALKYNES

Alkynes, or **acetylenes**, are **unsaturated*** (each molecule has a carbon-carbon **triple bond***) and **aliphatic compounds***. They are **hydrocarbons*** and form a **homologous series*** with a **general formula*** C_nH_{2n-2}. Alkynes are named in the same way as **alkanes***, but end in -yne, not -ane (see page 214). They are **non-polar molecules*** with chemical properties similar to **alkenes***. They burn with a sooty flame in air, and a very hot flame in pure oxygen. Alkynes are produced by **cracking***. They are used to make plastics and solvents.

Structural formulae* of some alkynes	
Name of compound	Structural formula*
Ethyne	$CH \equiv CH$
Propyne	$CH_3C \equiv CH$
But-1-yne	$CH_3CH_2C \equiv CH$

Ethyne (C_2H_2) or acetylene

The simplest member of the alkynes. A colorless gas, less dense than air and with a slightly sweet smell. It is the only common alkyne. Ethyne undergoes the same reactions as the other alkynes but more vigorously, e.g. it reacts explosively with chlorine. It is used in oxy-acetylene welding torches as it burns with a very hot flame. Ethyne is made by **cracking*** and is used to make polyvinyl chloride (PVC) and other vinyl compounds.

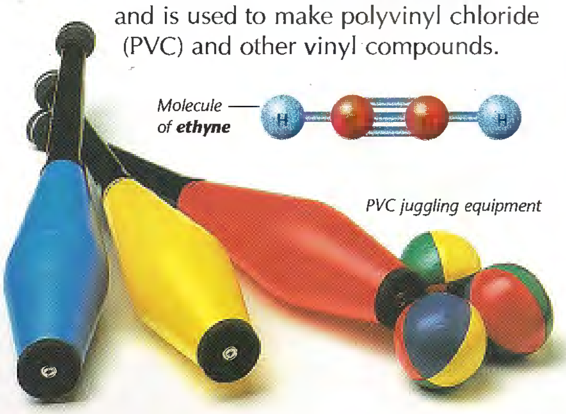

Molecule of **ethyne**

PVC juggling equipment

More homologous series

The following groups of organic compounds each form a **homologous series*** of **aliphatic compounds***. Each series has a particular **functional group*** and its members have similar chemical properties.

Aldehydes

Compounds that contain a **–CHO functional group***. They form a **homologous series*** with a **general formula*** $C_nH_{2n+1}CHO$, and are named like **alkanes*** but end in -al, not -ane (see page 215). They are colorless liquids (except **methanal**) and **reducing agents***, and undergo **addition***, **condensation*** and **polymerization reactions***. When **oxidized***, they form **carboxylic acids**.

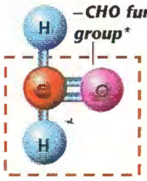

*Molecule of **methanal** (**HCHO**) or **formaldehyde**, the simplest **aldehyde**. It is a colorless, poisonous gas with a strong smell. It dissolves in water to make **formalin** – used to preserve biological specimens. It is also used to make **polymers*** and adhesives.*

Ketones

Compounds that contain a **carbonyl group** (a **–CO– functional group***). Ketones form a **homologous series*** with a complex **general formula***. They are named like **alkanes*** but end in -one, not just -e. Most are colorless liquids. They have chemical properties similar to **aldehydes** but are not **reducing agents***.

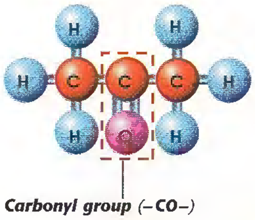

*Molecule of **propanone** (CH_3COCH_3) or **acetone**, the simplest **ketone***

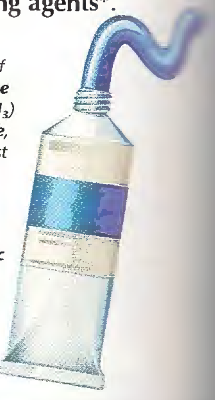

***Acrylic** paint*

***Propanone** is a colorless liquid that mixes with water. It is used to make **acrylic**, and as an organic solvent, e.g. as nail polish remover.*

Carboxylic acids

Compounds that co
(a **–COOH functio**
homologous series*
C_nH_{2n+1}COOH. The
(see page 215). Pu
acids*, they react
(see **condensation rea**

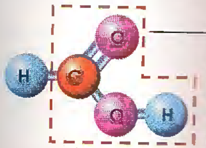

***Carboxyl group** (**–COOH**)*

*Molecule of **ethanoic***
*(CH_3COOH) or **acetic***

Vinegar, which
from grapes,
ethanoic acid.

Carboxyl

Dicarboxylic acids

Compounds that con
(see **carboxylic acids**

*Molecule of **ethanedioic acid** ($(COOH)_2$) or **oxalic acid** – a poisonous, **dicarboxylic acid** found in rhubarb leaves but not stalks.*

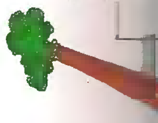

Two
gro

Esters

A **homologous series***
a **–COO– functional**
They are unreactive,
reacting a **carboxylic**
condensation reactio
vegetable oils and ani
and flowers their flav
used in perfumes and

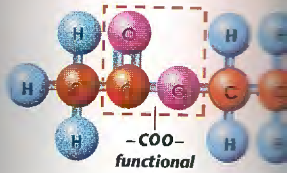

---

* **Addition reaction**, 193; **Aliphatic compounds**, 190; **Alkanes**, 192; **Alkenes**, 193; **Condensation reaction**, 197; **Cracking**, 19 **Functional group**, 190; **General formula**, 191 (**Homologous series**); **Hydrocarbons**, 190; **Non-polar molecule**, 133; **Oxidation**, 1 **Polymerization**, **Polymers**, 200; **Reducing agent**, 148; **Structural formula**, 140; **Triple bond**, 132; **Unsaturated compounds**, 191.

Alcohols, 196; **Functional**
Halogens, 186; **Ozone**, 210
Volatile, 345; **Weak acid**,

Carboxylic acids

Compounds that contain a **carboxyl group** (a **–COOH functional group***) and form a **homologous series*** with a **general formula*** **$C_nH_{2n+1}COOH$**. Their names end in -oic acid (see page 215). Pungent, colorless **weak acids***, they react with **alcohols*** to give **esters** (see **condensation reaction**, page 197).

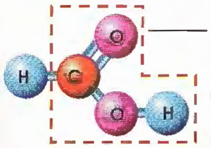

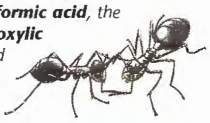

Molecule of ***methanoic acid*** *(**HCOOH**) or* ***formic acid****, the simplest* ***carboxylic acid*** *– a liquid found in ants and nettles.*

Carboxyl group (–COOH)

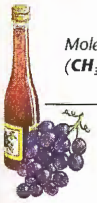

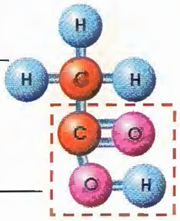

Molecule of ***ethanoic acid*** *(**CH_3COOH**) or* ***acetic acid***

Vinegar, which is made from grapes, contains ethanoic acid.

Carboxyl group

Dicarboxylic acids

Compounds that contain two **carboxyl groups** (see **carboxylic acids**) in each molecule.

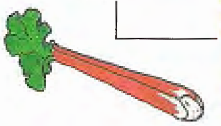

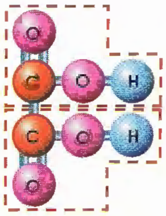

Molecule of ***ethanedioic acid*** *(**$(COOH)_2$**) or* ***oxalic acid*** *– a poisonous,* ***dicarboxylic acid*** *found in rhubarb leaves but not stalks.*

Two ***carboxyl groups***

Esters

A **homologous series*** of compounds containing a **–COO– functional group*** in every molecule. They are unreactive, colorless liquids made by reacting a **carboxylic acid** and **alcohol*** (see **condensation reaction**, page 197). Found in vegetable oils and animal fats, they give fruit and flowers their flavors and smells. They are used in perfumes and flavorings.

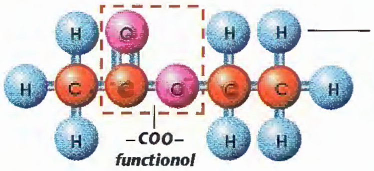

–COO– functional group*

Molecule of the ***ester ethyl ethanoate*** *(**$CH_3COOCH_2CH_3$**) or* ***ethyl acetate****. Some esters that smell like pears are used in sweets.*

Halogenoalkanes or alkyl halides

A **homologous series*** whose members contain one or more **halogen*** atoms (see also page 215). Most halogenoalkanes are colorless, **volatile*** liquids which do not mix with water. They will undergo **substitution reactions***. The most reactive contain iodine, and the least reactive contain fluorine.

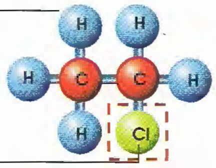

Molecule of ***chloroethane*** *(**CH_3CH_2Cl**), a* ***halogenoalkane****. Used to keep refrigerators cold (see* ***refrigerant****, page 345).*

The chlorine atom is the ***halogen* functional group****. It is called a* ***chloro group*** *(**–Cl**) (see page 215).*

Some important organic compounds have more than one **halogen*** atom in their molecules.

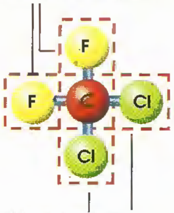

Fluoro groups (–F functional groups*)

Chloro groups (–Cl functional groups*)

Molecule of ***freon*** *(CCl_2F_2), a* ***chlorofluorocarbon*** *(a compound of chlorine, fluorine and carbon) once used as an aerosol propellant.*

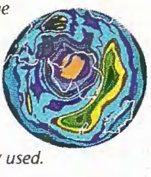

The orange patch on this satellite image of the Earth's atmosphere shows the hole in the ***ozone**** *layer. Freon is believed to contribute to this damage, so other propellants are now used.*

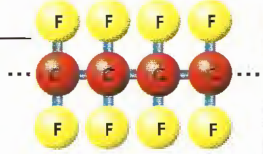

Molecule of ***poly(tetrafluoroethene)*** *(**PTFE**) (see also page 186)*

PTFE *is used as a non-stick coating on saucepans.*

Primary amines

Compounds that contain an **amino group** (**$-NH_2$ functional group***). They are **weak bases***, and have a fishy smell.

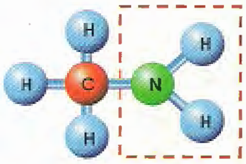

Methylamine *(**CH_3NH_2**), a* ***primary amine****.*

Amino group

Diamines

Compounds with two **amino groups** in each molecule.

---

* **Alcohols**, 196; **Functional group**, 191; **General formula**, 191 (**Homologous series**); **Halogens**, 186; **Ozone**, 210 (**Ozone depletion**); **Substitution reaction**, 192; **Volatile**, 345; **Weak acid**, **Weak base**, 152.

ALCOHOLS

Alcohols are organic compounds that contain one or more **hydroxyl groups** (**– OH functional groups***) in each molecule. The alcohols shown below in the chart are all members of a **homologous series*** of alcohols which are **aliphatic compounds*** with the **general formula*** $C_nH_{2n+1}OH$. As the molecules in the series increase in size, their physical properties change steadily. Some of the trends are shown in the chart below. As a result of their **hydroxyl groups**, alcohol molecules are **polar***, and have **hydrogen bonds***. Short-chain alcohols mix completely with water, but long-chain alcohols do not as their molecules have more $–CH_2–$ groups, making them less polar. Alcohols do not **ionize*** in water and are **neutral***. They burn, giving off carbon dioxide and water.

Some properties of alcohols

Name of compound	Structural formula*	Physical state at 25°C	Boiling point (°C)
Methanol	CH_3OH	Liquid	65.6
Ethanol	CH_3CH_2OH	Liquid	78.5
Propan-1-ol	$CH_3CH_2CH_2OH$	Liquid	97.2
Butan-1-ol	$CH_3CH_2CH_2CH_2OH$	Liquid	117.5

*Alcohols are named in the same way as **alkanes***, but end in -ol. The number in the name tells you which carbon atom the **hydroxyl group** is attached to (see opposite and page 214-215).*

The next member of the series (going down) is always a $–CH_2–$ group longer than the last.

The members gradually change to solids as the molecules get longer.

*Boiling points of alcohols increase as the molecules get longer. They have high boiling points in relation to their **relative molecular mass***, due to **hydrogen bonding***.*

Alcohols react with sodium:

$$\underset{\text{Ethanol}}{2CH_3CH_2OH} + \underset{\text{Sodium}}{2Na} \rightarrow \underset{\text{Sodium ethoxide}}{2CH_3CH_2ONa} + \underset{\text{Hydrogen}}{H_2}$$

Alcohols react with phosphorus halides to give **halogenoalkanes** (see page 195), and with **carboxylic acids*** to give **esters** (see **condensation reaction** and page 195).

Primary alcohols are **oxidized*** first to **aldehydes*** and then to **carboxylic acids***.

$$\underset{\text{Propan-1-ol}}{CH_3CH_2CH_2OH} \xrightarrow{\text{Acidified potassium permanganate catalyst*}} \underset{\text{Propanal}}{CH_3CH_2CHO} \rightarrow \underset{\text{Propanoic acid}}{CH_3CH_2COOH}$$

Secondary alcohols are **oxidized*** to **ketones** (see page 194).

$$\underset{\text{Propan-2-ol}}{CH_3CHOHCH_3} \xrightarrow{\text{Acidified potassium permanganate catalyst*}} \underset{\text{Propanone}}{CH_3COCH_3}$$

Ethanol (CH_3CH_2OH, often written C_2H_5OH)

Also called **ethyl alcohol**, or **alcohol**. An alcohol which is a slightly sweet-smelling water-soluble liquid with a relatively high boiling point. It burns with an almost colorless flame and is made by ethene reacting with steam. It is also produced by **alcoholic fermentation**.

__Ethanol__ is used as a solvent and in methylated spirits. It has many more uses including perfumes, paints, dyes, varnishes and alcoholic drinks.

---

* **Aldehydes**, 194; **Aliphatic compounds**, 190; **Alkanes**, 192; **Carboxylic acids**, 195; **Catalyst**, 161; **Functional group**, 190; **General formula**, 191 (**Homologous series**); **Hydrogen bond**, 134; **Ionization**, 130; **Neutral**, 151; **Oxidation**, 148; **Polar molecule**, 133; **Relative molecular mass**, 138; **Structural formula** (**shortened**), 140.

Alcoholic fermenta

The name of the
to produce **ethanol**
chemical in alcoho
from fruits or grain.
from fruit or grain
into ethanol by **enzy**
(**catalysts*** of the rea
Yeast is used in alco
because it has the
catalyses the chang

Laboratory fermentatio

Fermentation *mixture:* ***glucose****,* *water and yeast (ideal temperature is 37°C).*

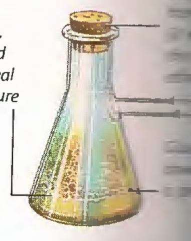

Yeast
high. Stronge
is made from
ethanol solution
from the water,
to make the

*__Enzyme*__*

$C_6H_{12}O_6$ →

Glucose solution from fruit or barley

Primary, seco

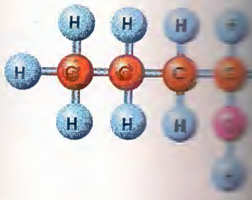

Molecule of __butan-1-ol,__
__alcohol__. The carbon atom
to the __hydroxyl group__ (see
introduction) has two hydrogen
atoms attached to it.

The
hydrog

---

* **Catalyst**, 161; **Distillation**,
Enzyme, 161; **Esters**, 195;
Glucose, 204; **Oxidation**,

Alcoholic fermentation

The name of the process used to produce **ethanol** (the potent chemical in alcoholic drinks) from fruits or grain. **Glucose*** from fruit or grain is converted into ethanol by **enzymes*** (**catalysts*** of the reactions in living cells). Yeast is used in alcoholic fermentation because it has the enzyme **zymase** which catalyses the change of glucose to ethanol.

Glucose* *in grapes is* ***fermented*** *to make wine.*

Laboratory fermentation

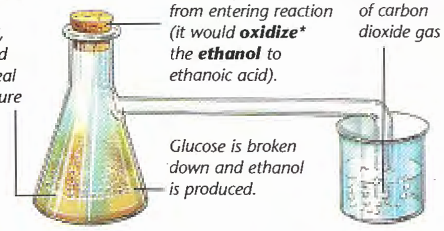

Fermentation *mixture:* ***glucose****, *water and yeast (ideal temperature is 37°C).*

Bung stops oxygen from entering reaction (it would ***oxidize**** *the* ***ethanol*** *to ethanoic acid).*

Bubbles of carbon dioxide gas

Glucose is broken down and ethanol is produced.

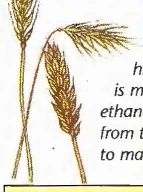

Yeast dies if ***ethanol*** *concentration gets too high. Stronger alcoholic drinks, e.g. whisky, which is made from grains, are made by* ***distilling**** *the ethanol solution. This process separates the ethanol from the water, and the concentrated alcohol is used to make the drinks more potent.*

$$C_6H_{12}O_6 \xrightarrow{\text{Enzyme*}} 2CH_3CH_2OH + 2CO_2$$

Glucose solution from fruit or barley	Ethanol	Carbon dioxide

Polyhydric alcohols

Alcohols whose molecules contain more than one **hydroxyl group** (see introduction).

Ethane-1,2-diol*, or* ***ethylene glycol*** *is a* ***diol*** *(contains two* ***hydroxyl groups****). Used as antifreeze.*

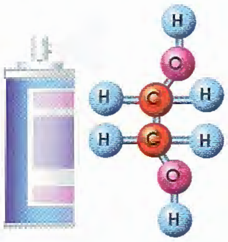

Propane-1,2,3-triol, glycerine*, or* ***glycerol****, is a* ***triol*** *(contains three* ***hydroxyl groups****). Used to make explosives.*

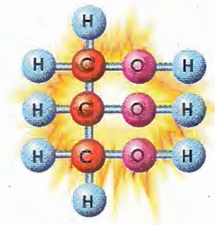

Condensation reaction

A type of reaction in which two molecules react together to form one, with the loss of a small molecule, e.g. water. (See also **condensation polymerization**, page 200.)

Example of a ***condensation reaction****:*

$$CH_3CH_2OH + CH_3COOH \rightarrow CH_3COOCH_2CH_3 + H_2O$$

Ethanol	Ethanoic acid	Ethyl ethanoate	Water molecule is lost

This reaction is also an **esterification reaction** as the product ethyl ethanoate is an **ester***. An alcohol and an organic acid always react to form an ester.

Primary, secondary and tertiary alcohols

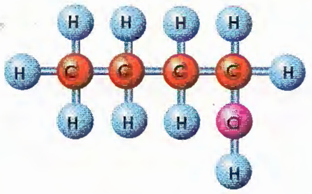

Molecule of ***butan-1-ol****, a* ***primary alcohol****. The carbon atom attached to the* ***hydroxyl group*** *(see introduction) has two hydrogen atoms attached to it.*

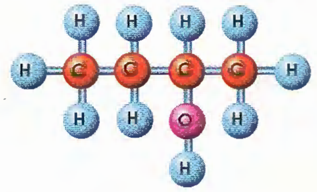

Molecule of ***butan-2-ol****, a* ***secondary alcohol****. The carbon atom attached to the hydroxyl group (see introduction) has one hydrogen atom attached to it.*

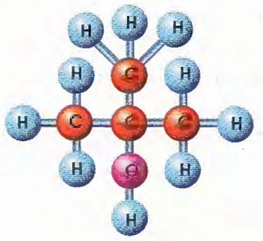

Molecule of ***2-methyl propan-2-ol****, a* ***tertiary alcohol****. The carbon atom attached to the hydroxyl group (see introduction) has no hydrogen atoms attached to it.*

The numbers in the names of the alcohols give the position of the carbon atom that the ***hydroxyl group*** *is bonded to. (See pages 214-215 for more information on naming alcohols.)*

---

* **Catalyst**, 161; **Distillation**, 220; **Enzyme**, 161; **Esters**, 195; **Glucose**, 204; **Oxidation**, 148.

PETROLEUM

Petroleum, or **crude oil**, is a dark, viscous liquid, usually found at great depths beneath the earth or sea-bed. It is often found with **natural gas***, which consists mainly of **methane***. Petroleum is formed over many thousands of years by the decomposition of animals and plants under pressure. It is a mixture of **alkanes*** which vary greatly in size and structure. Many useful products are formed by **refining** petroleum.

Refining

A set of processes which convert petroleum to more useful products. Refining consists of three main processes – **primary distillation**, **cracking** and **reforming**.

Primary distillation or fractional distillation of petroleum

A process used to separate petroleum into **fractions**, according to their boiling points (see also page 220). A **fractionating column** (see diagram) is kept very hot at the bottom, but it gets cooler towards the top. Boiled petroleum passes into the column as vapor, losing heat as it rises. When a fraction reaches a tray at a temperature just below its own boiling point, it condenses onto the tray. It is then drawn off along pipes. Fractions are distilled again to give better separations.

Fraction

A mixture of liquids with similar boiling points, obtained from **primary distillation**. **Light fractions** have low boiling points and short **hydrocarbon*** chains. **Heavy fractions** have higher boiling points and longer chains.

Cracking

A reaction which breaks large **alkanes*** into smaller alkanes and **alkenes***. The smaller alkanes are used as **gasoline**. Cracking occurs at high temperatures, or with a **catalyst*** (**catalytic cracking** or **"cat cracking"**).

$$\underset{\substack{\textit{Alkane}\\ \textit{(Nonane)}}}{C_9H_{20}} \rightarrow \underset{\substack{\textit{Alkane}\\ \textit{(Heptane)}}}{C_7H_{16}} + \underset{\substack{\textit{Alkene}\\ \textit{(Ethene*)}}}{C_2H_4}$$

Reforming

A process which produces **gasoline** from lighter **fractions** by breaking up **straight chain*** **alkanes*** and reassembling them as **branched chain*** molecules.

Primary distillation

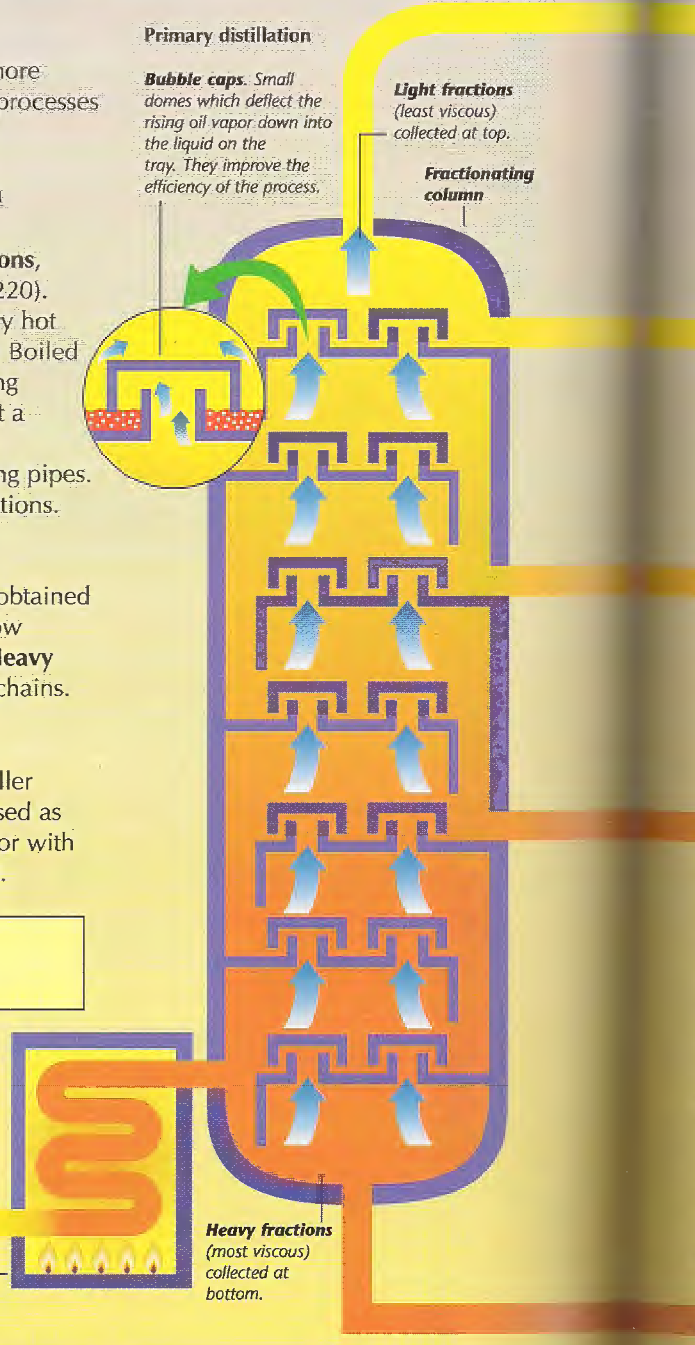

Bubble caps. *Small domes which deflect the rising oil vapor down into the liquid on the tray. They improve the efficiency of the process.*

Light fractions *(least viscous) collected at top.*
Fractionating column
Petroleum →
Furnace heats petroleum to 350°C
Heavy fractions *(most viscous) collected at bottom.*
Refinery gas
Gasoline
Kerosene
Diesel oil
Residue

Refinery gas

A gas which cons
of **methane***. Oth
fractions contain
butane (both **alkan**
are made into **liq**
petroleum gas (**LPG**

Gasoline or pe

A liquid **fraction**
primary distillati
of **alkanes*** with 5
atoms in their mo
has a boiling poin
40-150°C. See also
and **reforming**.

Kerosene or paraf

A liquid **fraction** o
primary distillation
of **alkanes*** with ab
atoms in their mole
boiling point range

Diesel oil or gas o

A liquid **fraction** ob
primary distillation.
alkanes* with about
carbon atoms in the
a boiling point of 25

Residue

The oil left after **prim**
distillation. It consis
hydrocarbons* of ver
molecular masses*,
containing up to 40
Its boiling point is grea
350°C. Some is used
which is used to heat
commercial buildings
to generate electricity.
re-distilled to form the
on the right.

*****Alkanes**, 192; **Alkenes**, 193; **Branched chain**, 190; **Catalyst**, 161; **Ethene**, 193; **Hydrocarbons**, 190; **Natural gas**, 192 (**Methane**); **Straight chain**, 190.

***Alkanes**, 192; **Branched cha**
Methane, 192; **Relative mole**

Refinery gas

A gas which consists mainly of **methane***. Other **light fractions** contain **propane** and **butane** (both **alkanes***) and are made into **liquefied petroleum gas** (**LPG**).

Liquefied petroleum gas *(see **refinery gas**) is used as bottled gas.*

Chemical feedstocks

Fractions of petroleum which are used in the production of organic chemicals. These fractions are mainly **refinery gas** and **naptha**, a part of the **gasoline** fraction.

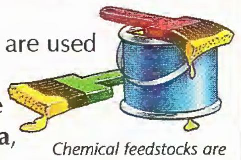

Chemical feedstocks are used to make paint.

Refinery gas

Gasoline or petrol

A liquid **fraction** obtained from **primary distillation**. It consists of **alkanes*** with 5 to 12 carbon atoms in their molecules and has a boiling point range of 40-150°C. See also **cracking** and **reforming**.

Octane rating

A measure of how well **gasoline** burns, measured on a scale of 0 to 100. It can be increased by using an **anti-knock agent** such as methyl-tertiary-butyl-ether ($C_5H_{12}O$).

Gasoline** used in cars has an **octane rating** of over 90. It consists mainly of **branched chain* alkanes*.

Gasoline

Kerosene or paraffin

A liquid **fraction** obtained from **primary distillation**. Kerosene consists of **alkanes*** with about 9-15 carbon atoms in their molecules. It has a boiling point range of 150-250°C.

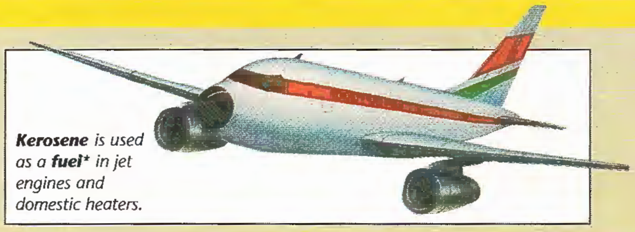

Kerosene** is used as a **fuel in jet engines and domestic heaters.*

Kerosene

Diesel oil or gas oil

A liquid **fraction** obtained from **primary distillation**. It consists of **alkanes*** with about 12-25 or more carbon atoms in their molecules. It has a boiling point of 250°C and above.

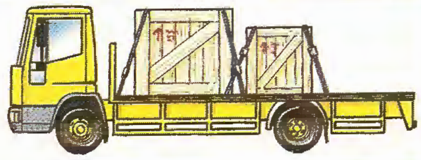

Diesel oil** is used as a **fuel in diesel engines.*

Diesel oil

Residue

The oil left after **primary distillation**. It consists of **hydrocarbons*** of very high **relative molecular masses***, their molecules containing up to 40 carbon atoms. Its boiling point is greater than 350°C. Some is used as **fuel oil**, which is used to heat homes and commercial buildings, as well as to generate electricity. The rest is re-distilled to form the substances on the right.

Lubricating oil

A mixture of non-**volatile*** liquids obtained from the distillation of **residue** in a vacuum.

Hydrocarbon waxes or paraffin waxes

Soft solids which are separated from **lubricating oil** after the distillation of **residue** in a vacuum.

Candles and polish

Bitumen or asphalt

A liquid left after the distillation of **residue** under vacuum. It is a tarry, black semi-solid at room temperature.

Road surfaces and roofing

Residue

*__Alkanes__, 192; **Branched chain**, 190; **Fuel**, 208; **Hydrocarbons**, 190; **Methane**, 192; **Relative molecular mass**, 138; **Volatile**, 345.

POLYMERS AND PLASTICS

Polymers are substances that consist of many **monomers** (small molecules) bonded together in a repeating sequence. They are very long molecules with a high **relative molecular mass***. Polymers occur naturally, e.g. **proteins***. There are also many **synthetic polymers**, e.g. **plastics**.

Monomers

Relatively small molecules that react to form polymers. For example, **ethene*** molecules are molecules which react together to form **polythene** (see also equation for **homopolymer**, below right).

Simplified picture of a polymerization reaction – a reaction in which monomers bond to form a polymer

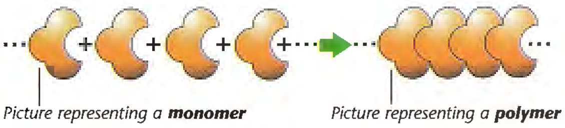

Picture representing a ***monomer*** *Picture representing a* ***polymer***

Synthetic polymers *such as* ***plastics*** *(see page 201) have many uses. Helmets for racing drivers are made of* ***thermosetting plastics*** *reinforced with synthetic fibers. Plastics used in the motorcycle bodywork reduce vehicle weight and help to save fuel.*

Addition polymerization

Polymerization reactions in which **monomers** bond to each other without losing any atoms. The polymer is the only product and has the same **empirical formula*** as the monomer. See also **addition reaction**, page 193.

Example of an addition polymerization reaction

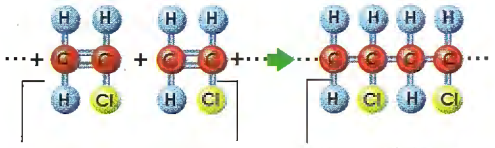
H H H H H H H H
C C C C C C C C
H Cl H Cl H Cl H Cl

Vinyl chloride ***monomer*** *Vinyl chloride* ***monomer*** ***Polyvinyl chloride (PVC)*** *polymer (****homopolymer****)*

Condensation polymerization

Polymerization reactions in which **monomers** form a polymer with the loss of small molecules such as water. See **condensation reaction**, page 197.

Homopolymer

A polymer made from a single type of **monomer**.

Reaction to produce the homopolymer polythene

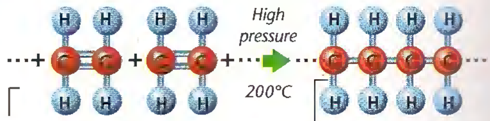
H H H H H H H H
C C C C C C C C
High pressure
200°C

All the ***monomers*** *are* ***ethene**** *molecules.* *This is an* ***addition polymerization*** *reaction.* *Part of a* ***polythene*** *molecule*

Copolymer

A polymer made from two or more different **monomers**. See **condensation polymerization** example below.

Depolymerization

The breakdown of a polymer into its original **monomers**. It occurs, for example, when **acrylic** is heated.

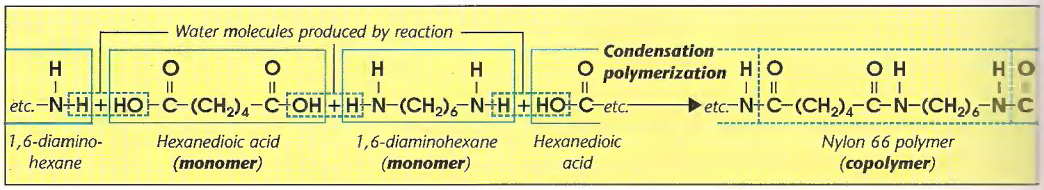
Water molecules produced by reaction
etc.–N–H + HO–C–(CH2)4–C–OH + H–N–(CH2)6–N–H + HO–C–etc.
Condensation polymerization
etc.–N–C–(CH2)4–C–N–(CH2)6–N–C
1,6-diamino-hexane
Hexanedioic acid (monomer)
1,6-diaminohexane (monomer)
Hexanedioic acid
Nylon 66 polymer (copolymer)

---

* **Empirical formula**, 140; **Ethene**, 193;
Proteins, 205; **Relative molecular mass**, 138.

Natural polym

Polymers that occ
rubber. Starch is
glucose*. For a p
see **starch**, page

Part of a rubber poly

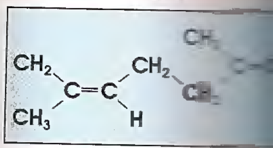
CH2
CH2
C=C
CH3
H

Rubber is extracted from ***lat***
then ***vulcanized**** *to prod*

Synthetic or man-

Polymers prepared
industry (not **natura**

Plastics

Synthetic polymers
are made from che
petroleum* and are
which are thermal
They are often not **bi**
off poisonous fumes
two types of plastic –
soften or melt on hea
thermosetting plastic
heating and do not
used in worktops).

Polyesters

Copolymers, formed
condensation polyme
and **dicarboxylic acid**
The monomers are
–COO– functional gr
found in **esters***.

Yachts have sails made of ***poly***
Some ***polyesters*** *are produced*
fibers which are used in clothing
and furnishing materials.

Nylons

A family of **polyamides**.
hard-wearing polyme
do not absorb water or
in fabrics, often mixed
condensation polymeri
for the manufacture of

---

* **Biodegradable**, 210; **Carboxyl**
Dicarboxylic acids, 195; **Diols**
Functional group, 190; **Gluc**

Natural polymers or biopolymers

Polymers that occur naturally, e.g. **starch** and **rubber**. Starch is made from **monomers** of **glucose***. For a picture of the starch polymer, see **starch**, page 204.

Part of a rubber polymer

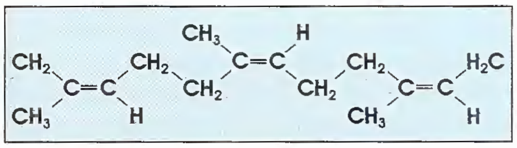

*Rubber is extracted from **latex*** tapped from the rubber tree. It is then **vulcanized*** to produce the rubber used in tires, hoses, etc.*

Synthetic or man-made polymers

Polymers prepared in the laboratory or in industry (not **natural polymers**), e.g. **nylons**.

Plastics

Synthetic polymers that are easily molded. They are made from chemicals derived from **petroleum*** and are usually durable, light solids which are thermal and electrical insulators. They are often not **biodegradable*** and give off poisonous fumes when burned. There are two types of plastic – **thermoplastics** which soften or melt on heating (e.g. **polythene**), and **thermosetting plastics** which harden upon heating and do not remelt (e.g. plastic used in worktops).

Polyesters

Copolymers, formed by the **condensation polymerization** of **diol*** and **dicarboxylic acid*** **monomers**. The monomers are linked by **–COO– functional groups***, as found in **esters***.

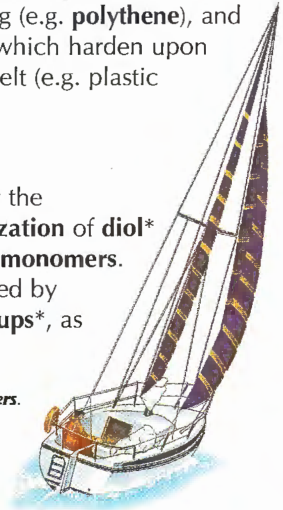

*Yachts have sails made of **polyesters**. Some **polyesters** are produced as fibers which are used in clothing and furnishing materials.*

Nylons

A family of **polyamides**. They are strong, hard-wearing polymers which stretch but do not absorb water or rot. They are used in fabrics, often mixed with other fibers. See **condensation polymerization** for the equation for the manufacture of nylon 66.

Polyamides

Copolymers formed by the **condensation polymerization** of a **dicarboxylic acid*** **monomer** with a **diamine*** **monomer**, e.g. **nylons**.

Polystyrene or poly(phenylethene)

A **homopolymer** formed by the **addition polymerization** of **styrene** (**phenylethene**).

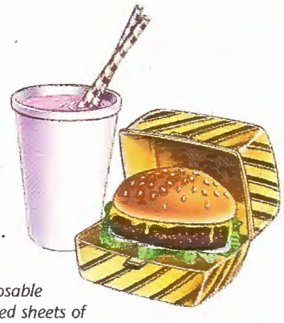

***Polystyrene** is used to make disposable knives, forks and cups. Air-expanded sheets of polystyrene are used in packaging and insulation.*

Polythene

Also called **poly(ethene)** or **poly(ethylene)**. A **homopolymer** formed by the **addition polymerization** of **ethene*** (see **homopolymer**, page 200). Polythene is produced in two forms (depending on the method used) – a soft material of low density, and a hard, more rigid, material of high density. Polythene has a **relative molecular mass*** of between 10,000 and 40,000 and is used to make many things, e.g. polythene bags (soft type), bowls used for washing (harder type).

Acrylic

Also called **poly(methylmethacrylate)** or **poly((1-methoxycarbonyl)-1-methylethene)**. A **homopolymer** formed by **addition polymerization**. It is often used as a glass substitute.

Methyl methacrylate**, the acrylic **monomer

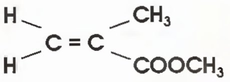

***Acrylic** is used to make outdoor signs.*

Polyvinyl chloride (PVC) or poly(chloroethene)

A hard-wearing **homopolymer** used to make many things, e.g. bottles and gloves. (See also **addition polymerization** picture, page 200).

*A **PVC** covering makes this underwater video camera waterproof.*

---

* **Biodegradable**, 210; **Carboxyl group**, 195 (**Carboxylic acids**); **Diamines**, 195; **Dicarboxylic acids**, 195; **Diols**, 197 (**Polyhydric alcohols**); **Esters**, 195; **Ethene**, 193; **Functional group**, 190; **Glucose**, 204; **Latex**, 345; **Petroleum**, 198; **Relative molecular mass**, 138; **Vulcanization**, 345.

DETERGENTS

Detergents are substances which, when added to water, enable it to remove dirt. They do this in three ways: by lowering the water's **surface tension*** so that it spreads evenly instead of forming droplets, by enabling grease molecules to dissolve in water, and also by keeping removed dirt suspended in the water. **Soap** is a type of detergent, but there are also many **soapless detergents**.

Detergent molecule

A large molecule consisting of a long **hydrocarbon*** chain with a **functional group*** at one end (making that end **polar***). The **non-polar*** chain is **hydrophobic** (repelled by water) and the polar end is **hydrophilic** (attracted to water). In water, these molecules group together to form **micelles**.

Simple representation of a detergent molecule

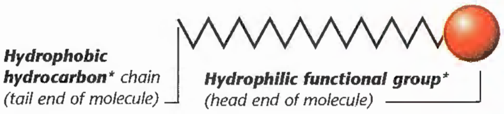
Hydrophobic hydrocarbon* chain (tail end of molecule)
Hydrophilic functional group* (head end of molecule)

Micelle

A spherical grouping of **detergent molecules** in water. Oils and greases dissolve in the **hydrophobic** center of the micelle. The picture below shows how micelles of dishwashing liquid remove grease.

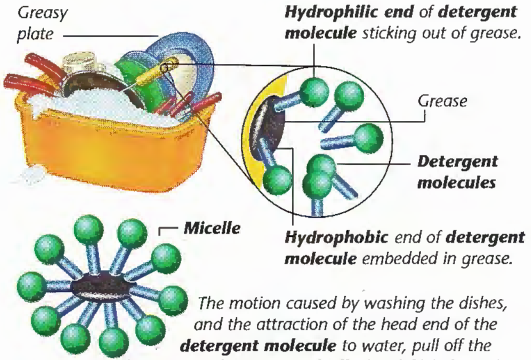
Greasy plate
Hydrophilic end of **detergent molecule** sticking out of grease.
Grease
Detergent molecules
Micelle
Hydrophobic end of **detergent molecule** embedded in grease.

The motion caused by washing the dishes, and the attraction of the head end of the ***detergent molecule*** *to water, pull off the detergent and grease. A* ***micelle*** *(see left) is formed when the grease is pulled off. Micelles then tend to keep grease suspended in solution.*

Soap

A type of detergent. It is the sodium or potassium **salt*** of a long-chain **carboxylic acid*** such as octadecanoic acid (see equation at bottom of page). It is made by reacting animal fats or vegetable oils (**esters***) with sodium hydroxide or potassium hydroxide solution (soap made with potassium hydroxide is softer). The process of making soap is **saponification**. Soap molecules form **micelles** in water. Soap produces a scum in **hard water*** whereas **soapless detergents** do not.

Saponification (soap-making)

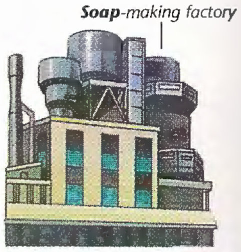
***Soap**-making factory*

Measured amounts of ***fats**** *and sodium hydroxide or potassium hydroxide solutions are continuously fed into a large, hollow, column-like structure. The column is at high temperature and pressure.*

Soap *and propane-1,2,3-triol are formed, and then the mixture is dissolved in salt water.*

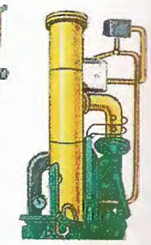

The final part of the process is ***fitting*** *or* ***finishing****. Any unreacted long-chain* ***carboxylic acids**** *are* ***neutralized**** *with alkali and the salt concentration is adjusted. The mixture is then centrifuged to separate out the soap.*

Saponification equation

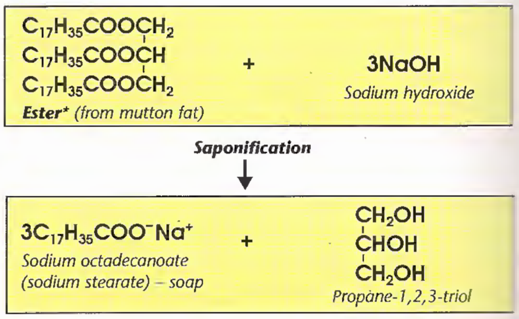

$$\begin{array}{l} C_{17}H_{35}COOCH_2 \\ C_{17}H_{35}COOCH \\ C_{17}H_{35}COOCH_2 \end{array} + 3NaOH \xrightarrow{\text{Saponification}} 3C_{17}H_{35}COO^-Na^+ + \begin{array}{l} CH_2OH \\ CHOH \\ CH_2OH \end{array}$$

Ester* *(from mutton fat)*
Sodium hydroxide
Sodium octadecanoate (sodium stearate) – soap
Propane-1,2,3-triol

All ***soap*** *molecules are sodium or potassium* ***salts**** *of long-chain* ***carboxylic acids****. In this example, the soap is a salt of* ***octadecanoic acid****.*

---

*Carboxylic acids**, 195; **Esters**, 195; **Fats**, 205 (**Lipids**); **Functional group**, 190; **Hard water**, 207; **Hydrocarbons**, 190; **Neutralization**, 151; **Non-polar molecule**, 133 (**Polar molecule**); **Salts**, 153; **Surface tension**, 23.

Soapless deterg
detergents

Types of deterge
by-products of **re**
They are used to
including **laundr**
and hair conditi
referred to as de
do not form a scu
lather better than
biodegradable (see

Example of a ***soapless***
not have an ***ionic**** *part –*

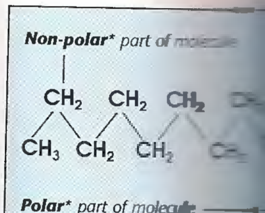
Non-polar* part of
CH₂ CH₂ CH₂
CH₃ CH₂ CH₂
Polar* part of molecu

Example of an ***ionic****
molecule – used in kitche

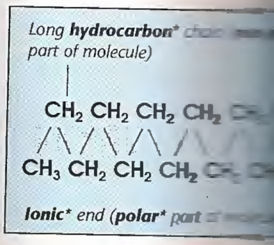
Long **hydrocarbon***
part of molecule)
CH₂ CH₂ CH₂ CH₂
CH₃ CH₂ CH₂ CH₂
Ionic* end (**polar*** part

Laundry powders

Soap or **soapless de**
clothes. They are b
alone, as they make
There are two main
– those used when
(usually soap powd
washing machines.
soapless detergent
added to keep the
brighten the appea
they also contain **enz**
biological laundry p
detergents. Enzyme
proteins* and loose

***Benzene ring**, 190 (**Aro**
Non-polar molecule, 133

Soapless detergents or **synthetic detergents**

Types of detergent made from by-products of **refining*** crude oil. They are used to make many products, including **laundry powders**, shampoos and hair conditioners, and are usually simply referred to as detergents. Soapless detergents do not form a scum in **hard water***, and lather better than **soaps**. If they are not **biodegradable** (see right), they pollute rivers.

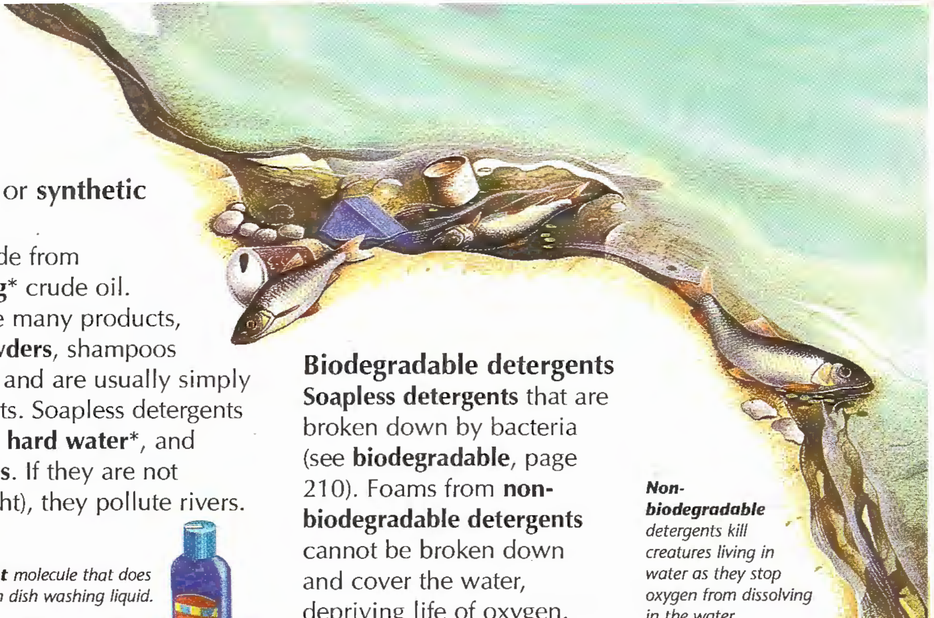

*Example of a **soapless detergent** molecule that does not have an **ionic*** part – used in dish washing liquid.*

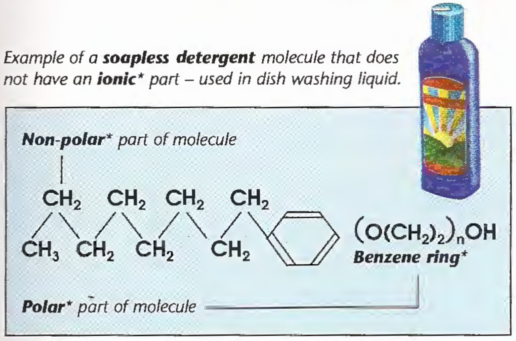
Non-polar* part of molecule
CH₂ CH₂ CH₂ CH₂
CH₃ CH₂ CH₂ CH₂
(O(CH₂)₂)ₙOH
Benzene ring*
Polar* part of molecule

*Example of an **ionic*** **soapless detergent** molecule – used in kitchen (laundry) soap.*

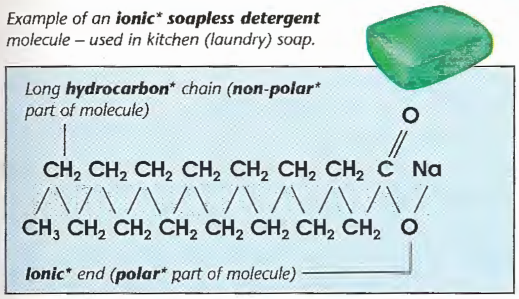
Long **hydrocarbon*** chain (**non-polar*** part of molecule)
O
CH₂ CH₂ CH₂ CH₂ CH₂ CH₂ CH₂ C Na
CH₃ CH₂ CH₂ CH₂ CH₂ CH₂ CH₂ CH₂ O
Ionic* end (**polar*** part of molecule)

Laundry powders

Soap or **soapless detergents** used to wash clothes. They are better for fabrics than water alone, as they make it easier to remove dirt. There are two main types of laundry powders – those used when hand-washing clothes (usually soap powders) and those used in washing machines. The latter are mostly soapless detergents with other substances added to keep the lather down and to brighten the appearance of the fabric. When they also contain **enzymes***, they are called **biological laundry powders**, or **enzyme detergents**. Enzymes help to break down **proteins*** and loosen dirt.

Biodegradable detergents

Soapless detergents that are broken down by bacteria (see **biodegradable**, page 210). Foams from **non-biodegradable detergents** cannot be broken down and cover the water, depriving life of oxygen.

***Non-biodegradable** detergents kill creatures living in water as they stop oxygen from dissolving in the water.*

Surfactants

Substances which lower the **surface tension*** of water. As a result of this property, detergents have many other uses, as well as removing dirt (see examples below).

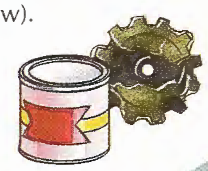

*Lubricating greases use **surfactants** to make them gel better.*

*Paints contain **surfactants** to ensure that the pigment is evenly mixed in, and that the paint gives a smooth finish and does not drip.*

***Surfactants** are added to cosmetics to make face powder cover well and evenly. They also ensure that cosmetic creams mix well with water and thicken properly.*

---

* **Benzene ring**, 190 (**Aromatic compounds**); **Enzyme**, 161; **Hard water**, 207; **Hydrocarbons**, 190; **Ionic compound**, 131; **Non-polar molecule**, 133 (**Polar molecule**); **Proteins**, 205; **Refining**, 198; **Surface tension**, 23.

FOOD

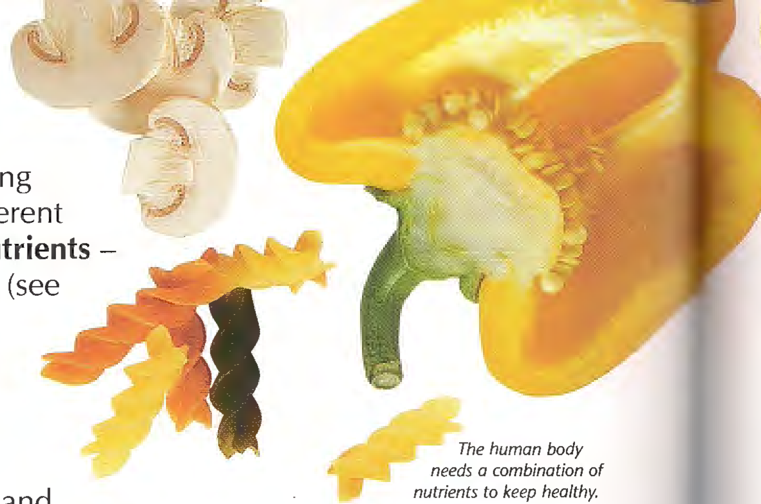

The human body needs a combination of nutrients to keep healthy.

In order to survive and grow, living organisms need a number of different substances. These include the **nutrients** – **carbohydrates**, **proteins** and **fats** (see **lipids**) – which are all **organic compounds*** made by plant **photosynthesis*** and taken in by animals. Also important are the **accessory foods** – water and **minerals**, needed by both plants and animals, and **vitamins**, needed by animals only. **Roughage**, or **fiber**, is also needed by many animals to help move food through the gut. Different animals need different amounts of these substances for a healthy diet. For more about minerals and roughage, see page 331.

Carbohydrates

Organic compounds* of varying complexity – the most complex, made of many individual units, being **polysaccharides** (see **starch**) and the simplest, made of just one unit, being **monosaccharides**. All have the **general formula*** $C_x(H_2O)_y$. Almost all living organisms use the monosaccharide **glucose** for energy.

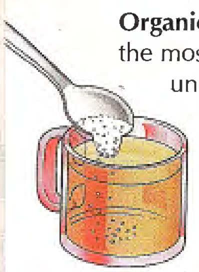

Sugar

*These foods contain **carbohydrates**.*

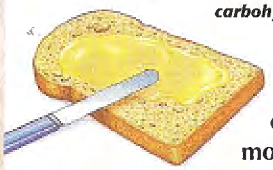

Bread

Pasta

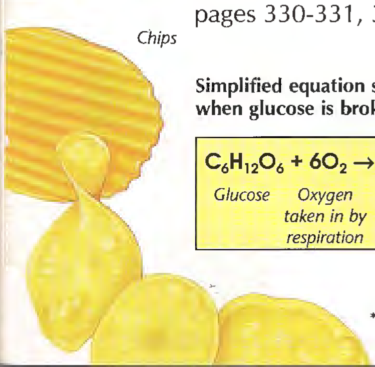

Chips

Glucose

A **monosaccharide** (see **carbohydrates**) with the **molecular formula*** $C_6H_{12}O_6$, the breakdown of which provides energy for plants and animals. Plants make their own by **photosynthesis***, storing it as **starch** until it is needed. Animals take in all forms of carbohydrate, break down the complex ones to glucose, and store this as the **polysaccharide** **glycogen**. For more about glucose, see pages 330-331, 334-335 and 338-339.

Simplified equation showing energy released when glucose is broken down in the body

$C_6H_{12}O_6$	$+ 6O_2 \rightarrow$	$6CO_2$	$+ 6H_2O$	+ ENERGY
Glucose	*Oxygen taken in by respiration*	*Carbon dioxide*	*Water*	*(measured in kJ)*

Sucrose

A **disaccharide**, i.e. a **carbohydrate**, composed of two **monosaccharide** units – in this case **glucose** and **fructose**. It is sweet-tasting, often used to sweeten food, and is commonly known as sugar. It has the **molecular formula*** $C_{12}H_{22}O_{11}$ and is obtained from sugar cane and sugar beet.

Starch

A **polysaccharide** (see **carbohydrates**) which is the storage form of **glucose** in plants. Like **glycogen** (see **glucose**), it is an example of a **natural polymer*** – the **monomers*** in this case being the glucose **monosaccharides**. Note that when these join, water molecules form at the links (see **condensation polymerization**, page 200).

Part of a starch molecule

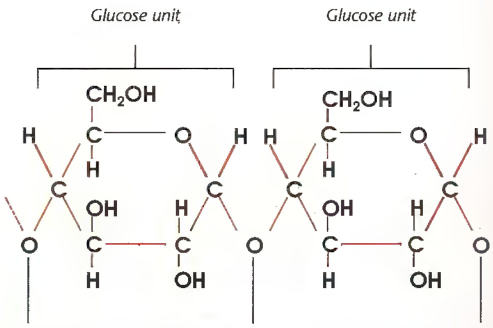

Before linking, these were OH on each molecule. H_2O formed and was "lost" into plant tissue.

Amino acids

Compounds whose carbon atom joined and an **amino** group* from amino acids.

*There are about 20 different natural **amino acids**. They all contain an **amino group*** and a **carboxyl group***.*

How proteins are

*Peanuts contain a lot of **protein**, so they are very nutritious.*

*1. Chewed peanuts go The **protein** they contain the stomach and the small*

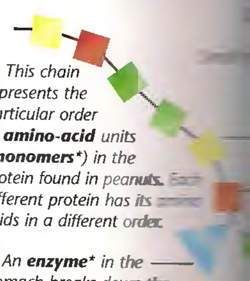

*2. This chain represents the particular order of **amino-acid** units (**monomers***) in the protein found in peanuts. different protein has its acids in a different order.*

*3. An **enzyme*** in the stomach breaks down the protein molecules (long **polypeptides**) into long enough to be called*

Enzymes are **catalysts*** up reactions in the body*

*4. An enzyme in the small intestine breaks the into molecules made of amino acids (**dipeptides**, into single amino acids.*

5. Amino acid molecules

6. In the body, certain proteins by joining amino

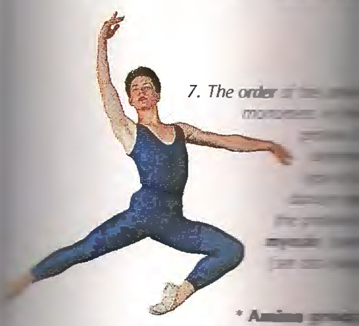

7. The order

---

* **General formula**, 191 (**Homologous series**); **Molecular formula**, 140; **Monomers**, 200; **Natural polymers**, 201; **Organic compounds**, 190; **Photosynthesis**, 209; **Polymer**, 200.

Catalysis, 161 (Cata
Organic solvent, 345

Amino acids

Compounds whose molecules contain a carbon atom joined to a **carboxyl group*** and an **amino group***. **Proteins** are made from amino acids. See also pages 328-329.

*There are about 20 different natural **amino acids**. They all contain an **amino group*** and a **carboxyl group***.*

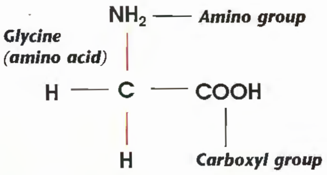

$H-\overset{NH_2}{\underset{H}{C}}-COOH$

How proteins are broken down in the body

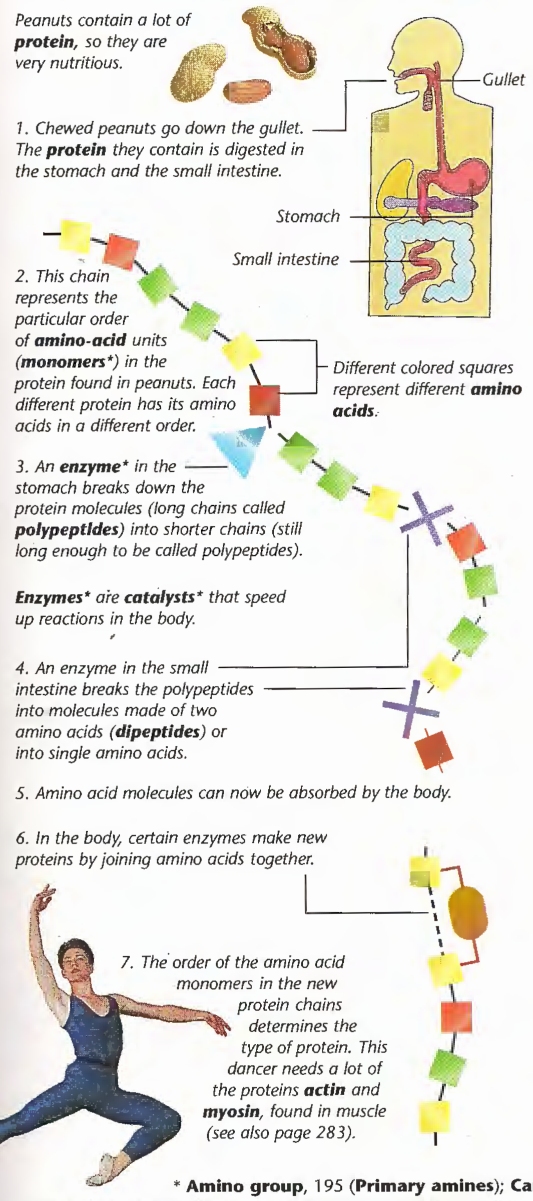

*Peanuts contain a lot of **protein**, so they are very nutritious.*

*1. Chewed peanuts go down the gullet. The **protein** they contain is digested in the stomach and the small intestine.*

*2. This chain represents the particular order of **amino-acid** units (**monomers***) in the protein found in peanuts. Each different protein has its amino acids in a different order.*

*Different colored squares represent different **amino acids**.*

*3. An **enzyme*** in the stomach breaks down the protein molecules (long chains called **polypeptides**) into shorter chains (still long enough to be called polypeptides).*

Enzymes* *are **catalysts*** that speed up reactions in the body.*

*4. An enzyme in the small intestine breaks the polypeptides into molecules made of two amino acids (**dipeptides**) or into single amino acids.*

5. Amino acid molecules can now be absorbed by the body.

6. In the body, certain enzymes make new proteins by joining amino acids together.

*7. The order of the amino acid monomers in the new protein chains determines the type of protein. This dancer needs a lot of the proteins **actin** and **myosin**, found in muscle (see also page 283).*

Proteins

Natural polymers* made from many **amino acid monomers*** joined together. The **relative molecular masses*** of proteins vary from 20,000 to several million. They are found mainly in meat, dairy food, nuts, cereal and beans. Animals need proteins for growth and repair of tissue. See also pages 330-331.

Vitamins

Organic compounds* found in small amounts in food. They are an essential part of the diet of animals. They are needed to help **enzymes*** **catalyse*** reactions in the body. See page 339 for a list of vitamins.

Example of a vitamin – vitamin C, also called ascorbic acid

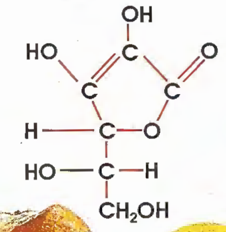

People who do not eat enough vitamin C get scurvy. Citrus fruits and vegetables are the main sources of this vitamin.

Lipids

A group of **esters***, including **fats** and waxes, found in living tissue (fats form a reserve energy source – see also pages 330-331). Insoluble in water but soluble in **organic solvents***, they are mostly solid or semi-solid and made of **saturated*** **carboxylic acids***, though a smaller group, the **oils**, are liquids and consist mainly of **unsaturated*** carboxylic acids.

Example of a reaction to make a fat

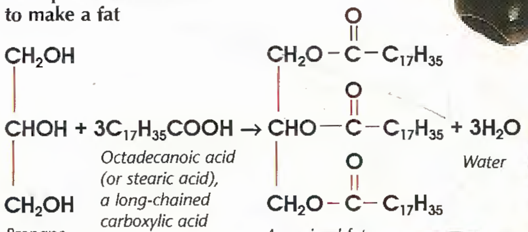

$$\begin{array}{l} CH_2OH \\ | \\ CHOH \\ | \\ CH_2OH \end{array} + 3C_{17}H_{35}COOH \rightarrow \begin{array}{l} CH_2O-\overset{O}{\overset{\|}{C}}-C_{17}H_{35} \\ | \\ CHO-\overset{O}{\overset{\|}{C}}-C_{17}H_{35} \\ | \\ CH_2O-\overset{O}{\overset{\|}{C}}-C_{17}H_{35} \end{array} + 3H_2O$$

Octadecanoic acid (or stearic acid), a long-chained carboxylic acid
Water
Propane-1,2,3-triol
An animal fat

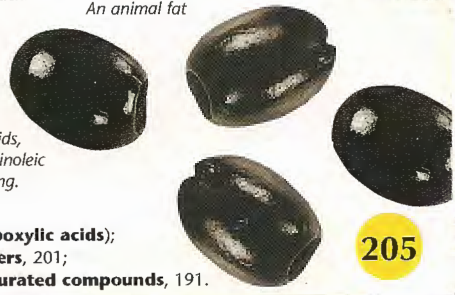

*Olives contain an **oil** which has a high proportion of **unsaturated*** fatty acids, such as oleic acid and linoleic acid. It is used in cooking.*

* **Amino group**, 195 (**Primary amines**); **Carboxyl group**, 195 (**Carboxylic acids**); **Catalysis**, 161 (**Catalyst**); **Enzyme**, 161; **Esters**, 195; **Monomers**, 200; **Natural polymers**, 201; **Organic solvent**, 345; **Relative molecular mass**, 138; **Saturated compounds**, **Unsaturated compounds**, 191.

WATER

Water (H_2O) is the most important compound on Earth. It is found on the surface and in the atmosphere, and is present in animals and plants. Vast amounts of water are used every day in the home and in industry, e.g. for manufacturing processes and the cooling of chemical plants. Water normally contains some dissolved gases, **salts*** and **pollutants***. See also page 167.

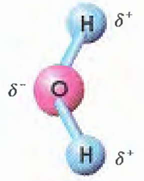

A molecule of water contains one oxygen atom and two hydrogen atoms.

*A water molecule is a **polar molecule****, which makes water a good **polar solvent****.*

Ice

The solid form of water. It has a **molecular lattice*** in which the molecules are further apart than in water. This is caused by **hydrogen bonds*** and means that ice is less dense than water, and that water expands when it freezes.

Cube of ice – the solid form of water

Water cycle

The constant circulation of water through the air, rivers and seas.

*Rainwater is relatively pure, but does contain some dissolved gases, e.g. carbon dioxide and sulfur dioxide (which produces **acid rain***).*

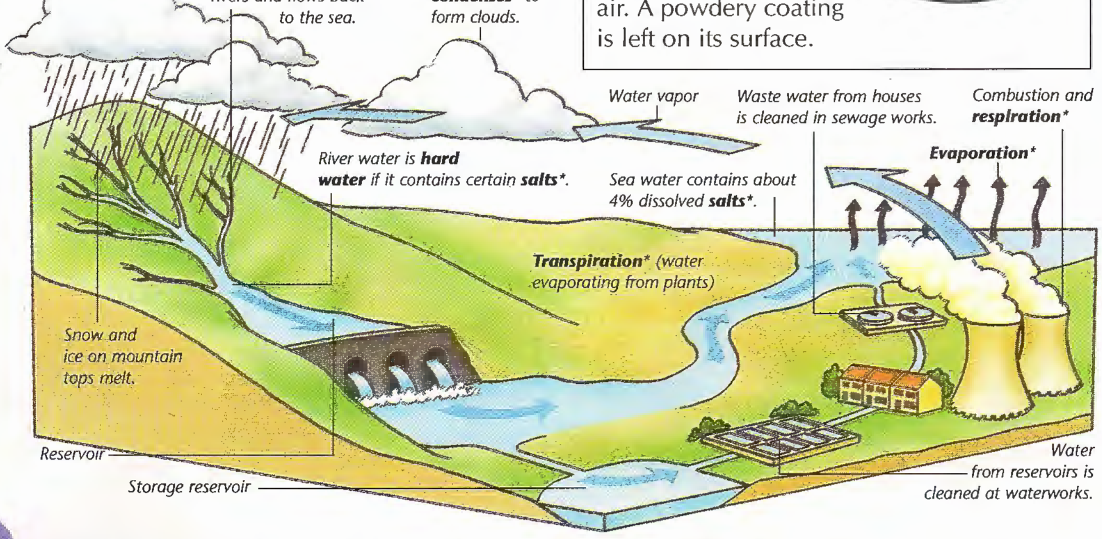
Rainwater runs off into rivers and flows back to the sea.
Water vapor **condenses*** to form clouds.
Water vapor
Waste water from houses is cleaned in sewage works.
Combustion and **respiration***
Evaporation*
River water is **hard water** if it contains certain **salts***.
Sea water contains about 4% dissolved **salts***.
Transpiration* (water evaporating from plants)
Snow and ice on mountain tops melt.
Reservoir
Storage reservoir
Water from reservoirs is cleaned at waterworks.

Atmospheric water

Over 70% of the Earth's surface is covered with water.

Humidity

The amount of water vapor in the air. It depends on the temperature and is higher (up to 4% of the air) in warm air than cold air.

Hygroscopic

Describes a substance which can absorb up to 70% of its own mass of water vapor. Such a substance becomes damp, but does not dissolve. Sodium chloride is an example of a hygroscopic substance.

Deliquescent

Describes a substance which absorbs water vapor from the air and dissolves in it, forming a **concentrated*** solution.

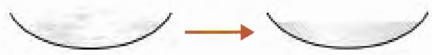

*Calcium chloride left open to the air absorbs water vapor and forms a **concentrated*** solution.*

Efflorescent

Describes a crystal which loses part of its **water of crystallization*** to the air. A powdery coating is left on its surface.

A white powder forms on sodium carbonate crystals.

---

* **Acid rain**, 210; **Concentrated**, 144; **Condensation**, **Evaporation**, 121; **Hydrogen bond**, 134; **Molecular lattice**, 137; **Polar molecule**, 133; **Polar solvent**, 144; **Pollutants**, 210; **Respiration**, 298; **Salts**, 153; **Transpiration**, 252; **Water of crystallization**, 135.

Water sup

Distilled wate

Water which
distillation*. It
some dissolved

Desalination

The treatment
dissolved **salts***.
ion exchange.

Purification

The treatment of
other harmful
that is safe to dri

At the waterworks,
reservoir trickles throu
of clean gravel and
activated carbon, to
particles of mud and

At a chlo
compounds
bacteria, th

Hard water

Water which conta
magnesium **salts***
rocks over which
calcium, page 17
contain these salts
are two types of h
hardness (which ca
removed relativel
permanent hardne
is more difficult to
Hard water does n
with soap and form
Soft water lathers
because it does not
with soap to form

Equation for the forma

Calcium and magnesium ions (in hard water) + Soap (sodi stearat

* **Crystals**, 135; **Distillatio**

Water supply

Distilled water

Water which has had **salts*** removed by **distillation***. It is very pure, but does contain some dissolved gases.

Desalination

The treatment of sea water to remove dissolved **salts***. It is done by **distillation*** or **ion exchange**.

Purification

The treatment of water to remove bacteria and other harmful substances, and produce water that is safe to drink.

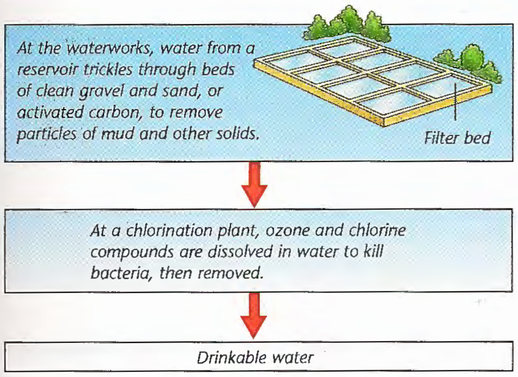

At the waterworks, water from a reservoir trickles through beds of clean gravel and sand, or activated carbon, to remove particles of mud and other solids.

At a chlorination plant, ozone and chlorine compounds are dissolved in water to kill bacteria, then removed.

Drinkable water

Hard water

Water which contains calcium and magnesium **salts*** that have dissolved from the rocks over which the water has flowed (see **calcium**, page 171). Water that does not contain these salts is called **soft water**. There are two types of hardness – **temporary hardness** (which can be removed relatively easily) and **permanent hardness** (which is more difficult to remove). Hard water does not lather with soap and forms a **scum**. Soft water lathers easily because it does not react with soap to form scum.

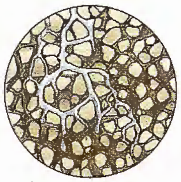

The types of mineral in water depend on the rocks it has flowed over.

Equation for the formation of scum

Calcium and magnesium ions (in hard water) + Soap (sodium stearate) → Scum (calcium and magnesium stearates) + Sodium ions

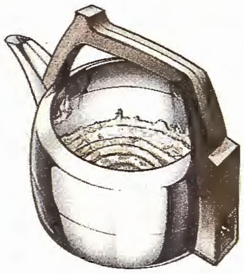

"Scale" forms in kettles which have been used to boil ***hard water.***

Temporary hardness

One type of water hardness, caused by the **salt*** calcium bicarbonate dissolved in the water. It can be removed by boiling, producing an insoluble white solid (calcium carbonate or "scale").

Permanent hardness

The more severe type of water hardness, caused by calcium and magnesium **salts*** (sulfates and chlorides) dissolved in the water. It cannot be removed by boiling, but can be removed by **distillation*** (producing **distilled water**) or by **water softening** (**ion exchange** or use of **water softeners**).

Ion exchange

A method of **water softening** (see **permanent hardness**). Water is passed over a material such as **zeolite** (sodium aluminum silicate), which removes calcium and magnesium ions and replaces them with sodium ions. Some organic **polymers*** are also used as ion exchange materials.

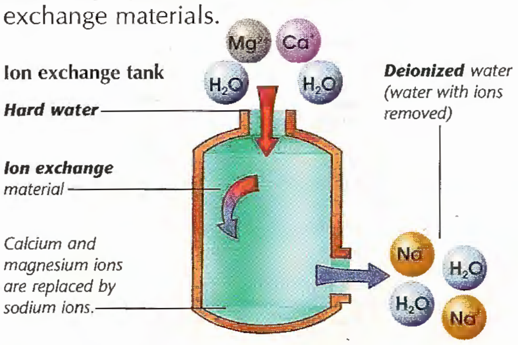

Ion exchange tank
Hard water
Ion exchange *material*
Calcium and magnesium ions are replaced by sodium ions.
Deionized *water (water with ions removed)*
Mg^{2+}
Ca^{+}
H_2O
H_2O
Na^{+}
H_2O
H_2O
Na^{+}

Water softeners

Substances used to remove **permanent hardness**. They react with the calcium and magnesium **salts*** to form compounds which do not react with soap.

Washing soda

The common name for the **hydrate*** of sodium carbonate (see also page 169). It is used as a **water softener** in the home.

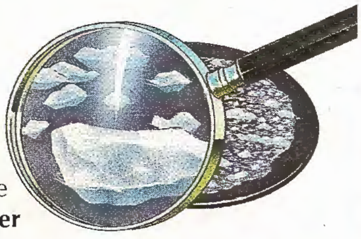

Crystals* *of washing soda*

---

* **Crystals**, 135; **Distillation**, 220; **Hydrate**, 154; **Polymers**, 200; **Salts**, 153.

AIR AND BURNING

Air is a mixture of gases, including oxygen, carbon dioxide and nitrogen, which surrounds the Earth and is essential for all forms of life. These gases can be separated by the **fractional distillation of liquid air***, and are used as raw materials in industry. Air also contains some water vapor and may contain **pollutants*** in some areas.

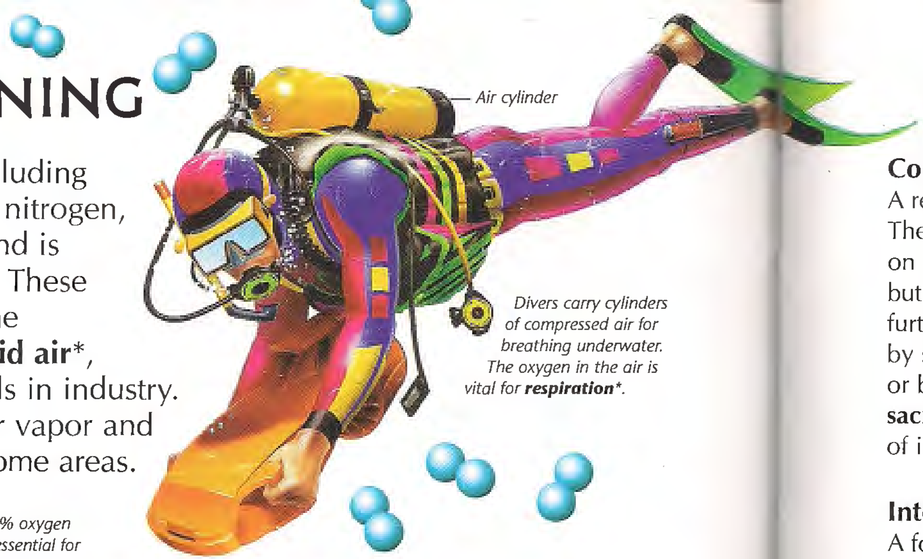

*Divers carry cylinders of compressed air for breathing underwater. The oxygen in the air is vital for **respiration***.*

Composition of dry air

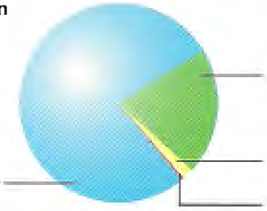

Combustion or burning

An **exothermic reaction*** between a substance and a gas. Combustion usually takes place in air, when the substance which burns combines with oxygen. Substances can also burn in other gases, though, e.g. chlorine. Combustion does not normally happen spontaneously. It has to be started by heating (see **activation energy**, page 160).

Natural gas *(mainly* ***methane****) burns in gas cookers, producing heat for cooking.*

$$CH_4(g) + O_2(g) \rightarrow CO_2(g) + 2H_2O(g) + \text{ENERGY}$$

$CH_4(g)$	$O_2(g)$	$CO_2(g)$	$2H_2O(g)$	ENERGY
Methane	*Oxygen from air*	*Carbon dioxide*	*Water vapor*	*for cooking*

Rapid combustion

Combustion in which a large amount of heat and light energy is given out.

Rapid combustion *can produce a large volume of gas as well as heat. This causes an explosion.*

Slow combustion

A form of **combustion** which takes place at low temperature. No **flames** occur. **Internal respiration** (see page 209) is a form of slow combustion.

Flame

A mixture of heat and light energy produced during **rapid combustion**.

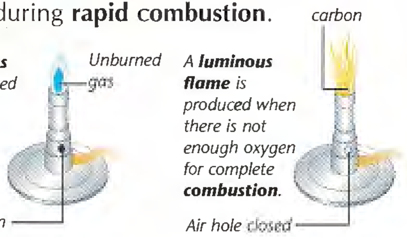

*A **non-luminous flame** is produced when there is enough oxygen for all of the substance to burn.*

*A **luminous flame** is produced when there is not enough oxygen for complete **combustion**.*

Fuel

A substance which is burned to produce heat energy. Most fuels used today are **fossil fuels**, which were formed from the remains of prehistoric animal and plant life.

*Wood is the oldest known **fuel**.*

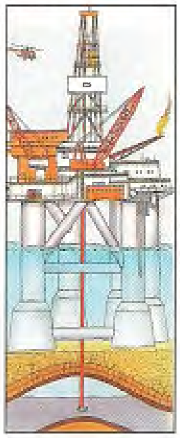

Fossil fuels*, such as* ***natural gas**** *and* ***petroleum**** *are extracted from deep under the ground.*

Calorific value

A measure of the amount of heat energy produced by a specific amount of a **fuel**. The table below shows the relative values for some common fuels.

Heat energy in ***kilojoules**** *per gram*

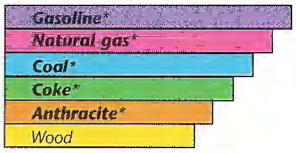

---

* **Anthracite**, 179 (**Coal**); **Coke**, 179 (**Coal**); **Exothermic reaction**, 146; **Fractional distillation of liquid air**, 183; **Gasoline**, 199; **Kilojoule**, 146; **Natural gas**, 192 (**Methane**); **Noble gases**, 189; **Petroleum**, 198; **Pollutants**, 210; **Respiration**, 298.

Corrosion

A reaction be
The metal is
on the surfac
but sometim
further corros
by stopping
or by preven
sacrificial pro
of iron is call

Internal resp

A form of **slo**
produces ene
with oxygen.
respiration,

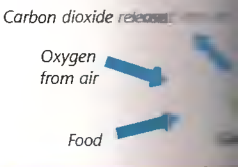

$C_6H_{12}O_6 + 6O_2$

Nitrogen cy

The constan
the air, anim

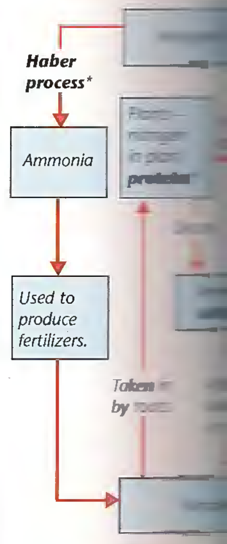

* **Carbohydrates**,
Petroleum, 198;

Corrosion

A reaction between a metal and the gases in air. The metal is **oxidized*** to form an oxide layer on the surface, usually weakening the metal, but sometimes forming a protective coat against further corrosion. Corrosion can be prevented by stopping oxygen from reaching the metal or by preventing electrons from leaving it (see **sacrificial protection**, page 159). The corrosion of iron is called **rusting** (see **rust**, page 174).

Internal respiration

A form of **slow combustion** in animals. It produces energy from the reaction of **glucose*** with oxygen. See also page 334 and **respiration**, page 298.

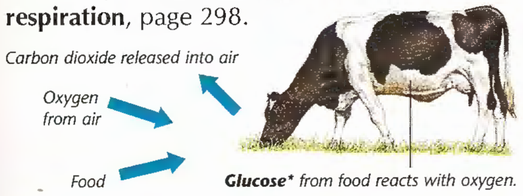
Carbon dioxide released into air
Oxygen from air
Food
Glucose* from food reacts with oxygen.

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O + \text{ENERGY}$$

Energy produced by reaction of glucose and oxygen

Nitrogen cycle

The constant circulation of nitrogen through the air, animals, plants and the soil.

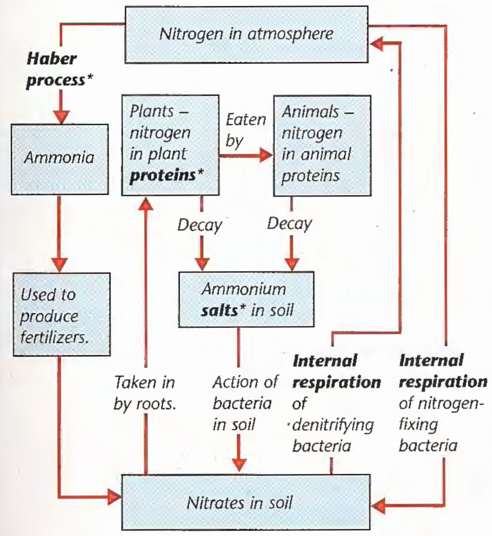
Nitrogen in atmosphere
Haber process*
Ammonia
Plants – nitrogen in plant **proteins***
Eaten by
Animals – nitrogen in animal proteins
Decay
Decay
Used to produce fertilizers.
Ammonium **salts*** in soil
Taken in by roots.
Action of bacteria in soil
Internal respiration of denitrifying bacteria
Internal respiration of nitrogen-fixing bacteria
Nitrates in soil

Photosynthesis

A **photochemical reaction*** in green plants. It involves the production of **glucose*** from carbon dioxide and water, using the energy from sunlight. Photosynthesis is chemically the opposite of **internal respiration**. See also pages 254-255.

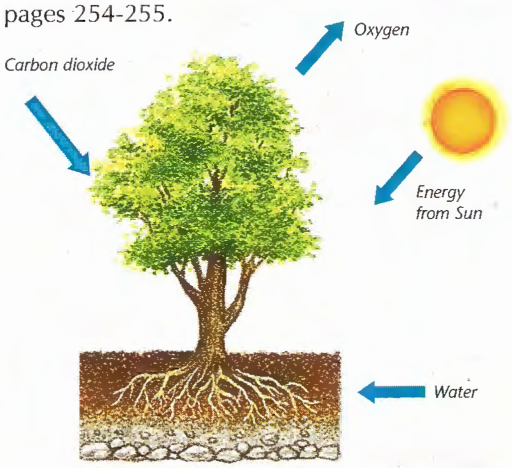
Oxygen
Carbon dioxide
Energy from Sun
Water

$$6CO_2 + 6H_2O \xrightarrow[\text{from Sun}]{\text{Energy}} C_6H_{12}O_6 + 6O_2$$

Carbon dioxide reacts with water, producing glucose.

Carbon cycle

The circulation of carbon through the air, animals, plants and the soil.

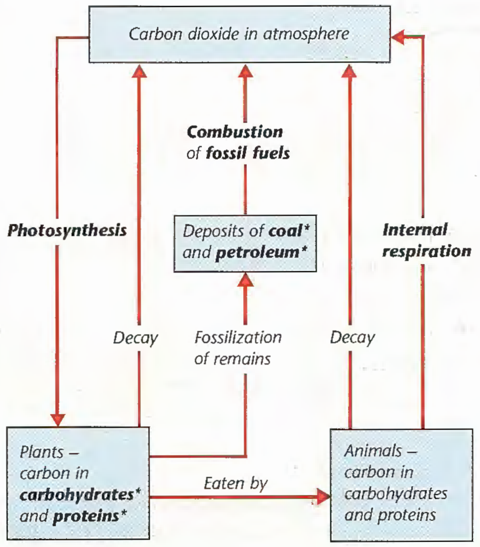
Carbon dioxide in atmosphere
Combustion of fossil fuels
Photosynthesis
Deposits of **coal*** and **petroleum***
Internal respiration
Decay
Fossilization of remains
Decay
Plants – carbon in **carbohydrates*** and **proteins***
Eaten by
Animals – carbon in carbohydrates and proteins

---

* **Carbohydrates**, 204; **Coal**, 179; **Glucose**, 204; **Haber process**, 180; **Oxidation**, 148; **Petroleum**, 198; **Photochemical reaction**, 160; **Proteins**, 205; **Salts**, 153.

POLLUTION

Pollution is the release into the land, atmosphere, rivers and oceans, of undesirable substances which upset the natural processes of the Earth. These substances are known as **pollutants**. The major sources and types of pollution are shown below.

Biodegradable

Describes a substance which is converted to simpler compounds by bacteria. Many plastics are not biodegradable (see also **biodegradable detergents**, page 203).

Smog

Fog mixed with dust and soot. It is acidic because of the sulfur dioxide produced when **fuels*** are burned in industrial cities.

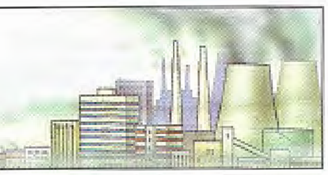

*Sulfur dioxide, produced by impurities in **fuels***, is the major cause of **acid rain**.*

Acid rain

Rainwater which is more acidic than usual. Rainwater normally has a **pH*** of between 5 and 6, due to dissolved carbon dioxide forming dilute carbonic acid. Sulfur dioxide and oxides of nitrogen, products of the combustion of **fuels***, react with water in the atmosphere to produce sulfuric and nitric acids with a pH of about 3.

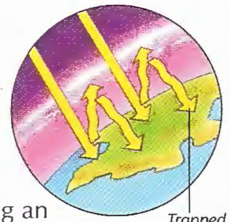

Trapped solar energy

Greenhouse effect

The trapping of solar energy in the atmosphere by carbon dioxide, causing an increase in temperature. The burning of **fuels*** creates more carbon dioxide, making the problem worse. See also page 29.

Ozone depletion

The thinning of the layer of **ozone*** gas in the upper atmosphere which protects the Earth from the Sun's harmful **ultraviolet radiation***. This effect is believed to be accelerated by chlorine acting as a **catalyst*** for the breakdown of oxygen to ozone. The chlorine comes from the decomposition of **chlorofluorocarbons*** (**CFCs**), chemicals used as a propellant in aerosols, as a coolant in refrigerators and in **polystyrene*** manufacture. International action is being taken to reduce the manufacture and use of CFCs, but many scientists believe that more could be done.

Thermal pollution

The effect of releasing warm water from factories and power stations into rivers and lakes. This causes a decrease in the oxygen dissolved in the water and affects aquatic life.

Eutrophication

An overgrowth of aquatic plants caused by an excess of nitrates, nitrites and phosphates from fertilizers in rivers. It results in a shortage of oxygen in the water, causing the death of fish and other water life.

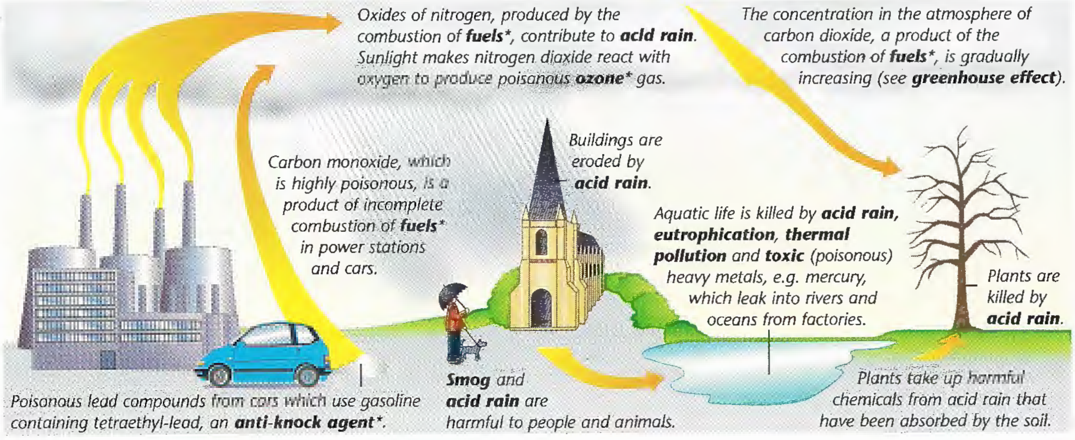

*Oxides of nitrogen, produced by the combustion of **fuels***, contribute to **acid rain**. Sunlight makes nitrogen dioxide react with oxygen to produce poisonous **ozone*** gas.*

*The concentration in the atmosphere of carbon dioxide, a product of the combustion of **fuels***, is gradually increasing (see **greenhouse effect**).*

*Carbon monoxide, which is highly poisonous, is a product of incomplete combustion of **fuels*** in power stations and cars.*

*Buildings are eroded by **acid rain**.*

*Aquatic life is killed by **acid rain**, **eutrophication**, **thermal pollution** and **toxic** (poisonous) heavy metals, e.g. mercury, which leak into rivers and oceans from factories.*

*Plants are killed by **acid rain**.*

*Poisonous lead compounds from cars which use gasoline containing tetraethyl-lead, an **anti-knock agent***.*

***Smog** and **acid rain** are harmful to people and animals.*

Plants take up harmful chemicals from acid rain that have been absorbed by the soil.

---

* **Anti-knock agent**, 199 (**Octane rating**); **Catalyst**, 161; **Chlorofluorocarbons**, 195; **Fuel**, 208; **Ozone**, 183; **Petroleum**, 198; **pH**, 152; **Polystyrene**, 201; **Purification**, 207; **Radioactivity**, 128; **Ultraviolet radiation**, 44.

THE RE

(showing

Metal	Symbol	Reaction with air	Reaction with water	Reaction with dilute strong acids*	Displacement* reactions	Reaction of carbon dioxide	Reaction of hydrogen with oxide	Action of heat on oxide	Action of heat on carbonate	Action of heat on nitrate	Symbol
Potassium	K	Burn strongly to form	React with	Explosive							

* **Displacement**, 158;

THE REACTIVITY SERIES

(showing ten metals – see also page 158)

Metal	Symbol	Reaction with air	Reaction with water	Reaction with dilute strong acids*	Displacement* reactions	Reaction of carbon dioxide	Reaction of hydrogen with oxide	Action of heat on oxide	Action of heat on carbonate	Action of heat on nitrate	Symbol
Potassium	**K**	*Burn strongly to form oxides.*	*React with cold water to produce hydrogen gas and hydroxide. Hydroxide dissolves in water to form alkaline solution. React with decreasing vigor down the series.*	*Explosive reaction to give hydrogen gas and* ***salt**** *solution.*	*All metals displace ions of metals below them from solution.*	*No reaction*	*No reaction*	*No reaction*	*No reaction*	*Decompose to form nitrite and oxygen.*	**K**
Sodium	**Na**										**Na**
Calcium	**Ca**	*Burn, when heated, to form oxides. Burn with decreasing vigor down the series.*		*React to give hydrogen gas and* ***salt**** *solution with decreasing vigor down the series.*					*Decompose to form oxide and carbon dioxide with increasing ease down the series.*	*Decompose to form oxide, oxygen and nitrogen dioxide with increasing ease down the series.*	**Ca**
Magnesium	**Mg**										**Mg**
Aluminum	**Al**		*No reaction with cold water. React with steam to form hydrogen gas and oxide. React with decreasing vigor down the series.*								**Al**
Zinc	**Zn**					*Oxide* ***reduced**** *to metal with increasing ease down the series. Carbon dioxide is formed.*					**Zn**
Iron	**Fe**										**Fe**
Lead	**Pb**	*Do not burn when heated, but form an oxide layer on surface.*	*No reaction*								**Pb**
Copper	**Cu**			*No reaction*			*Oxide* ***reduced**** *to metal with increasing ease down the series. Water is formed.*				**Cu**
Silver	**Ag**	*No reaction*						*Decomposes to form metal and oxygen only.*	*and carbon dioxide.*	*and nitrogen dioxide.*	**Ag**

***Displacement**, 158; **Reduction**, 148; **Salts**, 153; **Strong acid**, 152.

THE PROPERTIES OF THE ELEMENTS

Below is a chart giving information on the physical properties of the elements in the **periodic table** (see pages 164-165). The last eight elements (**atomic numbers*** 96-103 – see pages 165 and 226-227 for symbols and names) are not listed, as there is very little known about them – they all have to be made under special laboratory conditions and only exist for a fraction of a second. All the density measurements below are taken at room temperature except those of gases (marked with a †), which are measured at their boiling points. A dash (–) at any place on the chart indicates that there is no known value.

Element	Symbol	Atomic number*	Approximate relative atomic mass*	Density (g cm⁻³)	Melting point (°C) (brackets indicate approximations)	Boiling point (°C)
Actinium	***Ac***	*89*	*227*	*10.1*	*1,050*	*3,200*
Aluminum	***Al***	*13*	*27*	*2.7*	*660*	*2,470*
Americium	***Am***	*95*	*243*	*11.7*	*(1,200)*	*(2,600)*
Antimony	***Sb***	*51*	*122*	*6.62*	*630*	*1,380*
Argon	***Ar***	*18*	*40*	*1.4 †*	*–189*	*–186*
Arsenic	***As***	*33*	*75*	*5.73*	–	*613 **(sublimes*)***
Astatine	***At***	*85*	*210*	–	*(302)*	–
Barium	***Ba***	*56*	*137*	*3.51*	*714*	*1,640*
Beryllium	***Be***	*4*	*9*	*1.85*	*1,280*	*2,477*
Bismuth	***Bi***	*83*	*209*	*9.78*	*271*	*1,560*
Boron	***B***	*5*	*11*	*2.34*	*2,300*	*3,930*
Bromine	***Br***	*35*	*80*	*3.12*	*–7.2*	*58.8*
Cadmium	***Cd***	*48*	*112*	*8.65*	*321*	*765*
Calcium	***Ca***	*20*	*40*	*1.54*	*850*	*1,487*
Carbon	***C***	*6*	*12*	*2.25* ***(graphite*)***	*3,730 **(sublimes*)***	*4,830*
				3.51 ***(diamond*)***	*3,750*	–
Cerium	***Ce***	*58*	*140*	*6.78*	*795*	*3,470*
Cesium	***Cs***	*55*	*133*	*1.9*	*28.7*	*690*
Chlorine	***Cl***	*17*	*35.5*	*1.56 †*	*–101*	*–34.7*
Chromium	***Cr***	*24*	*52*	*7.19*	*1,890*	*2,482*
Cobalt	***Co***	*27*	*59*	*8.7*	*1,492*	*2,900*
Copper	***Cu***	*29*	*64*	*8.89*	*1,083*	*2,595*
Dysprosium	***Dy***	*66*	*162*	*8.56*	*1,410*	*2,600*
Erbium	***Er***	*68*	*167*	*9.16*	*1,500*	*2,900*
Europium	***Eu***	*63*	*152*	*5.24*	*826*	*1,440*
Fluorine	***F***	*9*	*19*	*1.11 †*	*–220*	*–188*
Francium	***Fr***	*87*	*223*	–	*(27)*	–
Gadolinium	***Gd***	*64*	*157*	*7.95*	*1,310*	*3,000*
Gallium	***Ga***	*31*	*70*	*5.93*	*29.8*	*2,400*
Germanium	***Ge***	*32*	*73*	*5.4*	*937*	*2,830*
Gold	***Au***	*79*	*197*	*19.3*	*1,063*	*2,970*
Hafnium	***Hf***	*72*	*178.5*	*13.3*	*2,220*	*5,400*
Helium	***He***	*2*	*4*	*0.147 †*	*–270*	*–269*
Holmium	***Ho***	*67*	*165*	*8.8*	*1,460*	*2,600*
Hydrogen	***H***	*1*	*1*	*0.07 †*	*–259*	*–252*
Indium	***In***	*49*	*115*	*7.3*	*157*	*2,000*
Iodine	***I***	*53*	*127*	*4.93*	*114*	*184*
Iridium	***Ir***	*77*	*192*	*22.4*	*2,440*	*5,300*
Iron	***Fe***	*26*	*56*	*7.85*	*1,535*	*3,000*
Krypton	***Kr***	*36*	*84*	*2.16 †*	*–157*	*–152*
Lanthanum	***La***	*57*	*139*	*6.19*	*920*	*3,470*
Lead	***Pb***	*82*	*207*	*11.3*	*327*	*1,744*
Lithium	***Li***	*3*	*7*	*0.53*	*180*	*1,330*
Lutetium	***Lu***	*71*	*175*	*9.84*	*1,650*	*3,330*

---

* **Atomic number**, 127; **Diamond**, **Graphite**, 178; **Relative atomic mass**, 138; **Sublimation**, 121.

Element
Magnesium
Manganese
Mercury
Molybdenum
Neodymium
Neon
Neptunium
Nickel
Niobium
Nitrogen
Osmium
Oxygen
Palladium
Phosphorus
Platinum
Plutonium
Polonium
Potassium
Praseodymium
Promethium
Protactinium
Radium
Radon
Rhenium
Rhodium
Rubidium
Ruthenium
Samarium
Scandium
Selenium
Silicon
Silver
Sodium
Strontium
Sulfur
Tantalum
Technetium
Tellurium
Terbium
Thallium
Thorium
Thulium
Tin
Titanium
Tungsten
Uranium
Vanadium
Xenon
Ytterbium
Yttrium
Zinc
Zirconium

---

* **Atomic number**, 127;
Relative atomic mass,

Element	Symbol	Atomic number*	Approximate relative atomic mass*	Density (g cm^{-3})	Melting point (°C) (brackets indicate approximations)	Boiling point (°C)
Magnesium	***Mg***	12	24	1.74	650	1,100
Manganese	***Mn***	25	55	7.2	1,240	2,100
Mercury	***Hg***	80	201	13.6	–38.9	357
Molybdenum	***Mo***	42	96	10.1	2,610	5,560
Neodymium	***Nd***	60	144	7.0	1,020	3,030
Neon	***Ne***	10	20	1.2 †	–249	–246
Neptunium	***Np***	93	237	20.4	640	–
Nickel	***Ni***	28	59	8.8	1,453	2,730
Niobium	***Nb***	41	93	8.57	2,470	3,300
Nitrogen	***N***	7	14	0.808 †	–210	–196
Osmium	***Os***	76	190	22.5	3,000	5,000
Oxygen	***O***	8	16	1.15 †	–218	–183
Palladium	***Pd***	46	106	12.2	1,550	3,980
Phosphorus	***P***	15	31	1.82 ***(white*)*** 2.34 ***(red*)***	44.2 ***(white)*** 590 ***(red)***	280 ***(white)*** –
Platinum	***Pt***	78	195	21.5	1,769	4,530
Plutonium	***Pu***	94	242	19.8	640	3,240
Polonium	***Po***	84	210	9.4	254	960
Potassium	***K***	19	39	0.86	63.7	774
Praseodymium	***Pr***	59	141	6.78	935	3,130
Promethium	***Pm***	61	147	–	1,030	2,730
Protactinium	***Pa***	91	231	15.4	1,230	–
Radium	***Ra***	88	226	5	700	1,140
Radon	***Rn***	86	222	4.4 †	–71	–61.8
Rhenium	***Re***	75	186	20.5	3,180	5,630
Rhodium	***Rh***	45	103	12.4	1,970	4,500
Rubidium	***Rb***	37	85	1.53	38.9	688
Ruthenium	***Ru***	44	101	12.3	2,500	4,900
Samarium	***Sm***	62	150	7.54	1,070	1,900
Scandium	***Sc***	21	45	2.99	1,540	2,730
Selenium	***Se***	34	79	4.79	217	685
Silicon	***Si***	14	28	2.35	1,410	2,360
Silver	***Ag***	47	108	10.5	961	2,210
Sodium	***Na***	11	23	0.97	97.8	890
Strontium	***Sr***	38	88	2.62	768	1,380
Sulfur	***S***	16	32	2.07 ***(rhombic*)*** 1.96 ***(monoclinic*)***	113 ***(rhombic)*** 119 ***(monoclinic)***	444 444
Tantalum	***Ta***	73	181	16.6	3,000	5,420
Technetium	***Tc***	43	99	11.5	2,200	3,500
Tellurium	***Te***	52	128	6.2	450	990
Terbium	***Tb***	65	159	8.27	1,360	2,800
Thallium	***Tl***	81	204	11.8	304	1,460
Thorium	***Th***	90	232	11.7	1,750	3,850
Thulium	***Tm***	69	169	9.33	1,540	1,730
Tin	***Sn***	50	119	7.3	232	2,270
Titanium	***Ti***	22	48	4.54	1,675	3,260
Tungsten	***W***	74	184	19.3	3,410	5,930
Uranium	***U***	92	238	19.1	1,130	3,820
Vanadium	***V***	23	51	5.96	1,900	3,000
Xenon	***Xe***	54	131	3.52 †	–112	–108
Ytterbium	***Yb***	70	173	6.98	824	1,430
Yttrium	***Y***	39	89	4.34	1,500	2,930
Zinc	***Zn***	30	65	7.1	420	907
Zirconium	***Zr***	40	91	6.49	1,850	3,580

* **Atomic number**, 127; **Monoclinic sulfur**, 184; **Red phosphorus**, 182 (**Phosphorus**);
Relative atomic mass, 138; **Rhombic sulfur**, 184; **White phosphorus**, 182 (**Phosphorus**).

NAMING SIMPLE ORGANIC COMPOUNDS

Simple **organic compounds*** (those with one or no **functional group***) can be named by following Stages 1 and 2.

Chart showing prefixes used to denote the number of carbon atoms in a chain

Number of carbon atoms in chain	*Prefix used*
One	**meth-**
Two	**eth-**
Three	**prop-**
Four	**but-**
Five	**pent-**
Six	**hex-**
Seven	**hept-**
Eight	**oct-**

Stage 1

Choose the sentence from a) to i) which describes the unidentified molecule, then go to the Stage 2 number indicated.

a) The molecule contains only carbon and hydrogen atoms and **single bonds***.	Go to 1
b) The molecule contains only carbon and hydrogen atoms and a **double bond***.	Go to 2
c) The molecule contains only carbon and hydrogen atoms and a **triple bond***.	Go to 3
d) The molecule contains carbon, hydrogen and a **hydroxyl group** (**–OH**).	Go to 4
e) The molecule contains carbon, hydrogen and a **–CHO group** at one end.	Go to 5
f) The molecule contains carbon, hydrogen and a **carbonyl group** (**–CO–**) between two carbons in the carbon chain.	Go to 6
g) The molecule contains carbon, hydrogen and **carboxyl group** (**–COOH**).	Go to 7
h) The molecule contains only carbon and hydrogen, but has a **side chain***.	Go to 8
i) The molecule contains carbon, hydrogen and one or more **halogen*** atoms.	Go to 9

Key to atoms

Carbon atom

Hydrogen atom

Other atoms as named

Stage 2

1. *The name of a molecule that contains only carbon and hydrogen atoms joined by **single bonds*** begins with the prefix for the number of carbons (see prefix chart, left) and ends in **-ane**.*
For example:

Methane

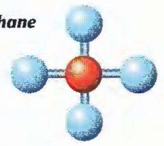

Ethane

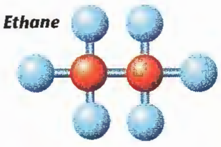

Propane

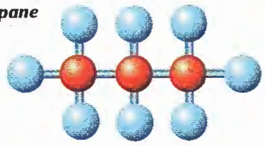

*These molecules are all **alkanes***.*

2. *The name of a molecule that contains only carbon and hydrogen atoms and has one **double bond*** begins with the prefix for the number of carbons (see prefix chart, left) and ends in **-ene**.*
For example:

Ethene

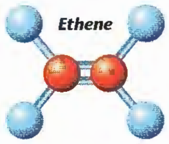

Propene

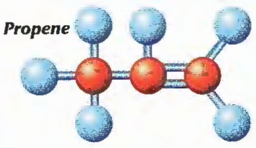

*These molecules are all **alkenes***.*

3. *The name of a molecule that contains only carbon and hydrogen atoms and has one **triple bond*** begins with the prefix for the number of carbons (see prefix chart, left) and ends in **-yne**.*
For example:

Ethyne

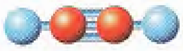

Propyne

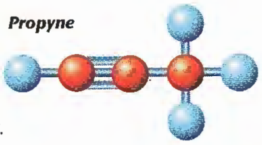

*These molecules are all **alkynes***.*

4a. *The name of a molecule that contains only carbon and hydrogen atoms and one **hydroxyl group** (**–OH**) begins with the prefix for the number of carbons (see prefix chart, left) and ends in **-ol**.*
For example:

Methanol

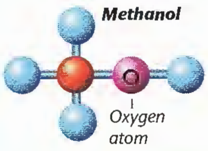

Ethanol

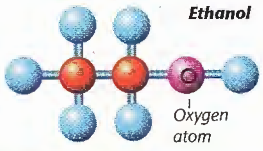

4b. *If the **–OH***
number of the
name. The carbon
molecule closest to

Octan-3-ol

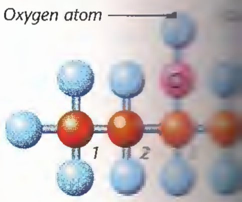

All molecules in

5. *The name of a*
atoms, and has a –
prefix for the number
*ends in **-al**. For*

Ethanal

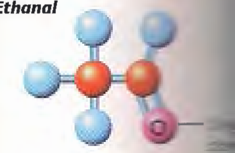

These molecules are

6. *The name of a*
atoms, and has a
carbon chain, begins
prefix chart, page 214

Propanone

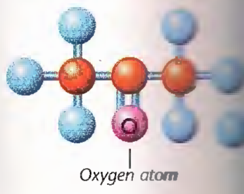

These molecules are all

7. *The name of a molecule*
*atoms and one **carboxyl***
the number of carbons (
***-oic acid**. For example:*

Methanoic acid

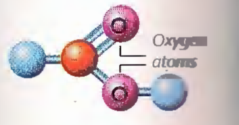

*All molecules in section **7***

---

* **Alkanes**, 192; **Alkenes**, 193; **Alkynes**, 194; **Double bond**, 132; **Functional group**, 190; **Halogens**, 186; **Organic compounds**, **Side chain**, 190; **Single bond**, **Triple bond**, 132.

* **Alcohols**, 196; **Halogenoalkanes**, 195;

__4b__. If the __–OH group__ is not at one end of the molecule, the number of the carbon to which it is attached is given in front of the name. The carbon atoms are always numbered from the end of the molecule closest to the –OH group. For example:

Butan-2-ol

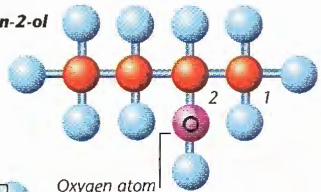

Octan-3-ol

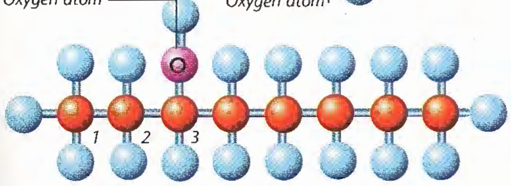

*All molecules in sections __4a__ and __4b__ are __alcohols*__.*

__5__. The name of a molecule that contains only carbon and hydrogen atoms, and has a __–CHO group__ ending the chain, begins with the prefix for the number of carbons (see prefix chart, page 214) and ends in __-al__. For example:

Ethanal

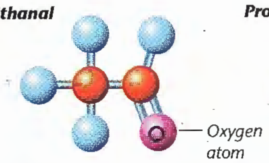

Propanal

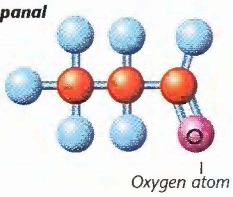

*These molecules are all __aldehydes*__.*

__6__. The name of a molecule that contains only carbon and hydrogen atoms, and has a __carbonyl group__ (__–CO–__) between the ends of the carbon chain, begins with the prefix for the number of carbons (see prefix chart, page 214) and ends in __-one__. For example:

Propanone

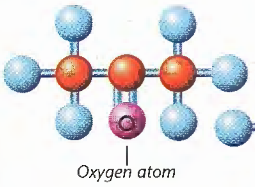

Butanone

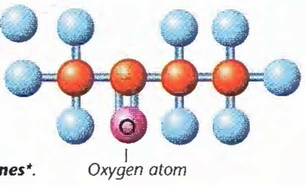

*These molecules are all __ketones*__.*

__7__. The name of a molecule that contains only carbon and hydrogen atoms and one __carboxyl group__ (__–COOH__) begins with the prefix for the number of carbons (see prefix chart, page 214) and ends in __-oic acid__. For example:

Methanoic acid

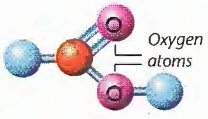

Ethanoic acid

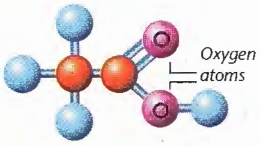

*All molecules in section __7__ are __carboxylic acids*__.*

*__8__. The name of a branched molecule begins with the name of the branch (__side chain*__). If this has only carbon and hydrogen atoms, its name begins with the prefix for the number of carbons in its chain (see prefix chart, page 214) and ends in __-yl__. The main chain is named afterward in the normal way (see 1). For example:*

2-methyl butane

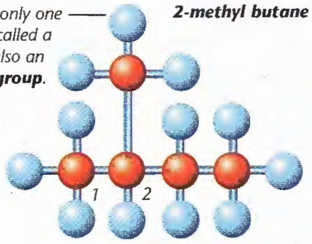

*This __side chain*__ has only one carbon atom, so it is called a __methyl group__. It is also an example of an __alkyl group__. Alkyl groups are any groups of carbon and hydrogen atoms that have a __general formula*__ of C_nH_{2n+1}.*

The figure at the beginning of the name gives the number of the carbon atom to which the side chain is joined. The carbon atoms are always numbered from the ends of the chain closest to the branch.

*__9__. The name of a molecule that contains carbon and hydrogen atoms and one or more __halogens*__ begins with the abbreviation for the halogen(s). (They are listed in alphabetical order if more than one.) The abbreviations for bromine, chlorine, fluorine and iodine are __bromo__, __chloro__, __fluoro__ and __iodo__ respectively.*

Chloromethane

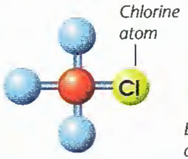

Bromoethane

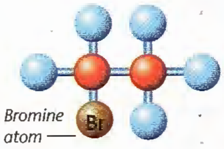

The end of the name is that which the molecule would have had if all the halogen atoms had been replaced by hydrogen atoms (see 1). With molecules of three carbon atoms or more, the name includes the number of the carbon atom to which the halogen is attached. The carbon atoms are always numbered from the end of the chain closest to the halogen(s). For example:

3-iodohexane

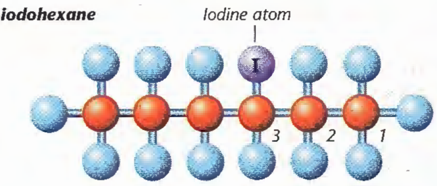

2-bromo,1-chloropentane

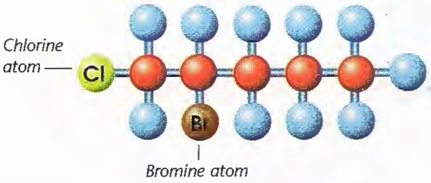

*All molecules in section __9__ are __halogenoalkanes*__.*

---

* **Alcohols**, 196; **Aldehydes**, 194; **Carboxylic acids**, 195; **General formula**, 191 (**Homologous series**); **Halogenoalkanes**, 195; **Halogens**, 186; **Ketones**, 194; **Side chain**, 190.

THE LABORATORY PREPARATION OF SIX COMMON GASES

Methods for preparing six gases – **carbon dioxide**, **chlorine**, **ethene**, **hydrogen**, **nitrogen** and **oxygen** – are described below.

Carbon dioxide (see page 179) is obtained from the reaction of hydrochloric acid with calcium carbonate (marble chips). A gas jar filled with water is placed on the beehive shelf over the mouth of the delivery tube. Gas produced by the reaction comes out of the delivery tube and displaces the water in the gas jar. This method of collecting a gas is called collecting gas **over water**.

$CaCO_3(s)$	+	$2HCl(aq)$	$\rightarrow$	$CaCl_2(aq)$	+	H_2O	+	$CO_2(g)$
Calcium carbonate		*Hydrochloric acid*		*Calcium chloride*		*Water*		*Carbon dioxide*

Preparing carbon dioxide

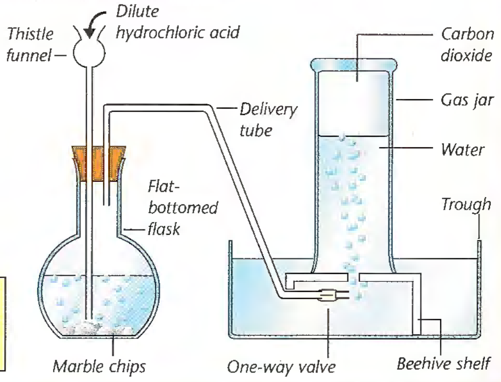
Dilute hydrochloric acid
Thistle funnel
Carbon dioxide
Gas jar
Delivery tube
Water
Flat-bottomed flask
Trough
Marble chips
One-way valve
Beehive shelf

Chlorine (see page 187) is prepared by **oxidizing*** concentrated hydrochloric acid using manganese(IV) oxide. This reaction is always done in a **fume cupboard***. The gas produced by the reaction contains some hydrogen chloride and water. The hydrogen chloride is removed by passing the stream of gas through water, and the water is removed by passing the gas through concentrated sulfuric acid. Finally the chlorine is collected in a gas jar. It displaces air from the gas jar as it is heavier. This method of gas collection is called collecting a gas by **upward displacement of air**.

Preparing chlorine

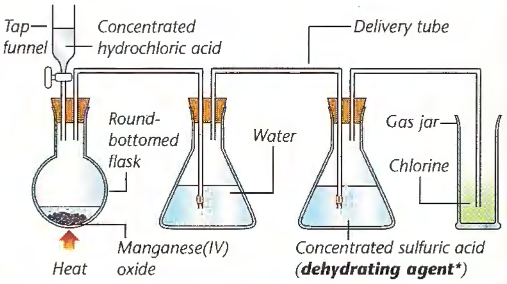
Tap funnel
Concentrated hydrochloric acid
Delivery tube
Round-bottomed flask
Water
Gas jar
Chlorine
Heat
Manganese(IV) oxide
Concentrated sulfuric acid
(**dehydrating agent***)

$MnO_2(s)$	+	$4HCl(aq)$	$\rightarrow$	$MnCl_2(aq)$	+	$Cl_2(g)$	+	$H_2O(l)$
Manganese(IV) oxide		*Hydrochloric acid*		*Manganese(IV) chloride*		*Chlorine*		*Water*

Ethene (see page 193) is prepared by dehydrating (removing water from) ethanol by reacting it with concentrated sulfuric acid. Aluminum sulfate is added to reduce frothing. The buffer flask ensures that any sodium hydroxide sucking back out of its flask does not mix with the acid. The sodium hydroxide removes acid fumes from the gas. Ethene is collected **over water** (see **carbon dioxide**, above).

Preparing ethene

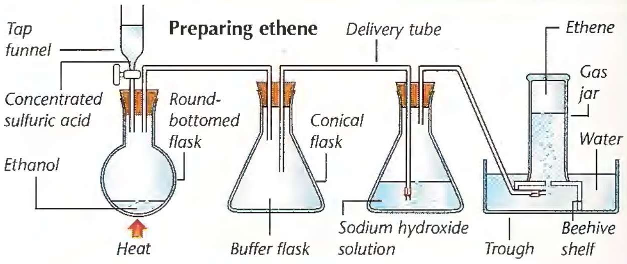
Tap funnel
Delivery tube
Ethene
Gas jar
Concentrated sulfuric acid
Round-bottomed flask
Conical flask
Water
Ethanol
Heat
Buffer flask
Sodium hydroxide solution
Trough
Beehive shelf

$CH_3CH_2OH(l)$	$\xrightarrow{\text{Concentrated sulfuric acid}}$	$C_2H_4(g)$	+	$H_2O(l)$
Ethanol		*Ethene*		*Water*

*__Dehydrating agent__, 344; **Fume cupboard**, 224; **Oxidation**, 148.

Hydrogen (see
the reaction
granulated z
usually added
hydrogen is
dioxide, page
needed, in
concentrated
by the **down**
pushes the air
it is lighter than

$Zn(s)$	+	$2HCl$
Zinc		*Hyd*

Nitrogen (see
prepared by
dioxide and ox
carbon dioxide
passing the air
hydroxide solu
removed by pass
heated copper.
collected **over**
dioxide, page 2
noble gases*

$CO_2(g)$ + $2NaOH$
Carbon dioxide

Oxygen (see page
when hydrogen
Manganese(IV)
catalyst* to speed
The gas is collec
carbon dioxide,
it must be dry, in
passed through
acid and is collec
displacement of
page 216).

$2H_2O_2(aq)$
Hydrogen peroxide

*__Catalyst__, 161; **Noble gases**

Hydrogen (see page 167) is obtained from the reaction of hydrochloric acid with granulated zinc. A little copper(II) sulfate is usually added to speed up the reaction. The hydrogen is collected **over water** (see **carbon dioxide**, page 216) unless dry hydrogen is needed, in which case it is passed through concentrated sulfuric acid and is collected by the **downward displacement of air** (it pushes the air down out of the gas jar as it is lighter than air).

$Zn(s)$	+	$2HCl(aq)$	$\rightarrow$	$ZnCl_2(aq)$	+	$H_2(g)$
Zinc		*Hydrochloric acid*		*Zinc chloride*		*Hydrogen*

Preparing hydrogen

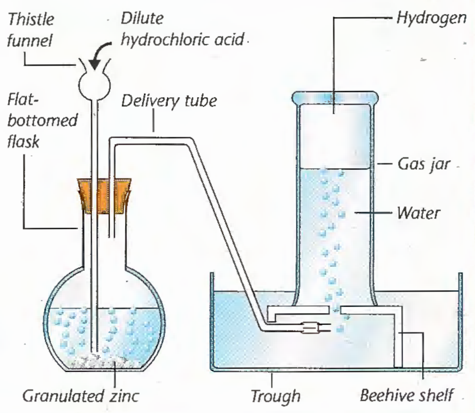
Thistle funnel
Dilute hydrochloric acid
Hydrogen
Flat-bottomed flask
Delivery tube
Gas jar
Water
Granulated zinc
Trough
Beehive shelf

Nitrogen (see page 180) is prepared by removing the carbon dioxide and oxygen from air. The carbon dioxide is removed by passing the air through sodium hydroxide solution. The oxygen is removed by passing the air over heated copper. The nitrogen is collected **over water** (see **carbon dioxide**, page 216). A residue of **noble gases*** remains in the nitrogen.

Preparing nitrogen

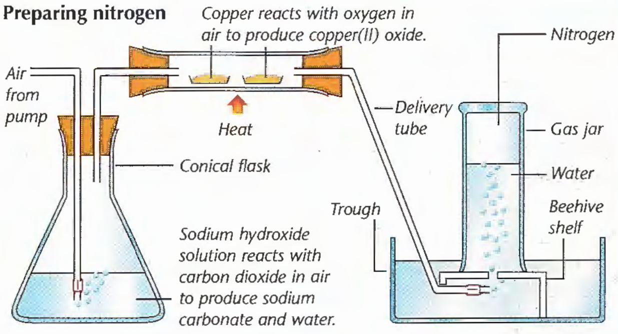
Copper reacts with oxygen in air to produce copper(II) oxide.
Nitrogen
Air from pump
Heat
Delivery tube
Gas jar
Conical flask
Water
Trough
Beehive shelf
Sodium hydroxide solution reacts with carbon dioxide in air to produce sodium carbonate and water.

$CO_2(g)$	+ $2NaOH(aq)$	$\rightarrow$	$Na_2CO_3(aq)$	+ $H_2O(l)$		
Carbon dioxide	*Sodium hydroxide*		*Sodium carbonate*	*Water*		

$O_2(g)$	+	$2Cu(s)$	$\rightarrow$	$2CuO(s)$
Oxygen		*Copper*		*Copper(II) oxide*

Oxygen (see page 183) is produced when hydrogen peroxide decomposes. Manganese(IV) oxide is used as a **catalyst*** to speed up this reaction. The gas is collected **over water** (see **carbon dioxide**, page 216) unless it must be dry, in which case it is passed through concentrated sulfuric acid and is collected by the **upward displacement of air** (see **chlorine**, page 216).

$$2H_2O_2(aq) \xrightarrow{\text{Manganese(IV) oxide}} 2H_2O(l) + O_2(g)$$

Hydrogen peroxide — *Water* — *Oxygen*

Preparing oxygen

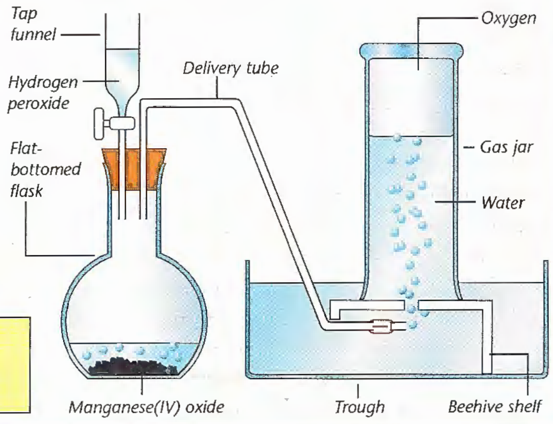
Tap funnel
Oxygen
Hydrogen peroxide
Delivery tube
Flat-bottomed flask
Gas jar
Water
Manganese(IV) oxide
Trough
Beehive shelf

---

*Catalyst, 161; Noble gases, 189.

LABORATORY TESTS

Various tests are used to identify substances. Some of the tests involve advanced machinery, others are simple laboratory tests and all are known collectively as **qualitative analysis**. Some of the more advanced tests are shown on page 222; these two pages cover simple laboratory tests leading to the identification of water, common gases, a selection of **anions*** and **cations*** (i.e. components of compounds) and some metals. The appearance or smell of a substance often gives clues to its identity – these can be confirmed by testing. If there are no such clues, then it is a matter of progressing through the tests, gradually eliminating possibilities (it is often a good idea to start with a **flame test**). Often more than one test is needed to identify an ion (anion or cation), as only one particular combination of results can confirm its presence (compare the tests and results for **lead**, **zinc** and **magnesium**).

Tests for water (H_2O)

Test	Results
*Add to **anhydrous*** copper(II) sulfate.*	*White copper(II) sulfate powder turns blue.*
*Add to **anhydrous*** cobalt(II) chloride.*	*Blue cobalt(II) chloride turns pink.*

Tests to identify gases

Gas	Symbol	Test	Results
Carbon dioxide	CO_2	*Pass into **limewater** (calcium hydroxide solution).*	*Turns limewater cloudy.*
Hydrogen	H_2	*Put a lighted splint into a sample of the gas.*	*Burns with a "popping" noise.*
Oxygen	O_2	*Put a glowing splint into a sample of the gas.*	*Splint relights.*

Tests for anions*

These tests are used to identify some of the **anions*** found in compounds.

Anion	Symbol	Test	Results
Bromide	Br^-	*Add silver nitrate solution to a solution of substance in dilute nitric acid.*	*Pale yellow precipitate that dissolves slightly in ammonia solution.*
Carbonate	CO_3^{2-}	*a) Add dilute hydrochloric acid to the substance.* *b) Try to dissolve the substance in water containing **universal indicator*** solution.*	*a) Carbon dioxide gas given off.* *b) If soluble, turns the indicator purple (compare **bicarbonate** test).*
Chloride	Cl^-	*Add silver nitrate solution to a solution of substance in dilute nitric acid.*	*Thick, white precipitate (which is soluble in ammonia solution).*
Hydrogencarbonate	HCO_3^-	*a) Add dilute hydrochloric acid to the substance.* *b) Try to dissolve the substance in water containing **universal indicator***.*	*a) Carbon dioxide gas evolved.* *b) Dissolves and turns indicator solution purple when boiled.*
Iodide	I^-	*Add silver nitrate solution to a solution of substance in dilute nitric acid.*	*Yellow precipitate that does not dissolve in ammonia solution.*
Nitrate	NO_3^-	*Add iron(II) sulfate solution followed by concentrated sulfuric acid to the solution.*	*Brown ring forms at the junction of the two liquids.*
Sulfate	SO_4^{2-}	*Add barium chloride solution to the solution.*	*White precipitate that does not dissolve in dilute hydrochloric acid.*
Sulfite	SO_3^{2-}	*Add barium chloride solution to the solution.*	*White precipitate that dissolves in dilute hydrochloric acid.*
Sulfide	S^{2-}	*Add lead(II) ethanoate solution to the solution.*	*Black precipitate.*

---

*Anhydrous, 154 (Anhydrate); Anion, Cation, 130; Universal indicator, 152.

Tests for

Most catio
by the same
identify pure
carry out a
gives a sele
can also be
reactions. A
listed in the
used to ide
metals are ins
cannot form

Cation	[illegible]
Aluminum	[illegible]
Ammonium	[illegible]
Calcium	Ca^{2+}
Copper(II)	Cu^{2+}
Iron(II)	Fe^{2+}
Iron(III)	Fe^{3+}
Lead(II)	Pb^{2+}
Magnesium	Mg^{2+}
Zinc	Zn^{2+}

*Cation, 130.

Tests for cations*

Most **cations*** in compounds can be identified by the same **flame tests** as those used to identify pure metals (see page 222 for how to carry out a flame test). The chart on the right gives a selection of flame test results. Cations can also be identified by the results of certain reactions. A number of these reactions are listed in the chart below. They cannot be used to identify pure metals, since many metals are insoluble in water and hence cannot form solutions.

Flame tests

Metal	Symbol	Flame color
Barium	***Ba***	*Yellow-green*
Calcium	***Ca***	*Brick red*
Copper	***Cu***	*Blue-green*
Lead	***Pb***	*Blue*
Lithium	***Li***	*Crimson*
Potassium	***K***	*Lilac*
Sodium	***Na***	*Orange-yellow*

Cation	Symbol	Test	Results
Aluminum	Al^{3+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.* *c) Compare with **lead** (see tests below).*	*a) White precipitate that dissolves as more sodium hydroxide solution is added.* *b) White precipitate that does not dissolve as more ammonia solution is added.* *c) –*
Ammonium	NH_4^+	*Add sodium hydroxide solution to a solution of the substance and heat gently.*	*Ammonia gas is given off. It has a distinctive choking smell.*
Calcium	Ca^{2+}	*a) See **flame test**.* *b) Add dilute sulfuric acid to a solution of the substance.*	*a) –* *b) White precipitate formed.*
Copper(II)	Cu^{2+}	*a) See **flame test**.* *b) Add dilute sodium hydroxide solution to a solution of the substance.* *c) Add dilute ammonia solution to a solution of the substance.*	*a) –* *b) Pale blue precipitate that dissolves as more sodium hydroxide is added.* *c) Pale blue precipitate, changing to deep blue solution as more ammonia solution is added.*
Iron(II)	Fe^{2+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.*	*a) Pale green precipitate formed.* *b) Pale green precipitate formed.*
Iron(III)	Fe^{3+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.*	*a) Red-brown precipitate formed.* *b) Red-brown precipitate formed.*
Lead(II)	Pb^{2+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.* *c) See also **flame test** to distinguish between lead and **aluminum**.*	*a) White precipitate that dissolves as more sodium hydroxide solution is added.* *b) White precipitate that does not dissolve as more ammonia solution is added.* *c) –*
Magnesium	Mg^{2+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.*	*a) White precipitate that does not dissolve as more sodium hydroxide solution is added.* *b) White precipitate that does not dissolve as more ammonia solution is added.*
Zinc	Zn^{2+}	*a) Add dilute sodium hydroxide solution to a solution of the substance.* *b) Add dilute ammonia solution to a solution of the substance.*	*a) White precipitate that dissolves as more sodium hydroxide solution is added.* *b) White precipitate that dissolves as more ammonia solution is added.*

*Cation, 150.

INVESTIGATING SUBSTANCES

The investigation of chemical substances involves a variety of different techniques. The first step is often to obtain a pure sample of a substance (impurities affect experimental results). Some of the separating and purifying techniques used to achieve this are explained on these two pages. A variety of different methods are then used to find out the chemical composition and the chemical and physical properties of the substance (**qualitative analysis**), and how much of it is present (**quantitative analysis**). For more information, see also pages 218-219 and 222.

Decanting

The process of separating a liquid from a solid that has settled, by pouring the liquid carefully out of the container.

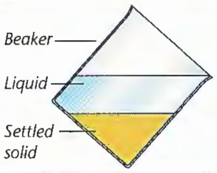
Beaker
Liquid
Settled solid

Filtering

The process of separating a liquid and a solid by pouring the mixture through a fine mesh. The mesh (usually filter paper) only lets liquid through.

Two methods of filtering

*Filtering through a **Buchner funnel** is the faster method, as the low pressure in the flask sucks the liquid rapidly through the filter paper.*

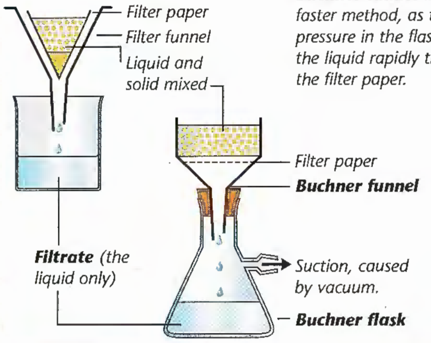
Filter paper
Filter funnel
Liquid and solid mixed
***Filtrate** (the liquid only)*
Filter paper
Buchner funnel
Suction, caused by vacuum.
Buchner flask

Centrifuging

*The process of separating different substances mixed in a liquid by spinning the test tube containing the liquid at high speed in a **centrifuge** (see picture below). Particles of different masses collect at different places in the test tube, the heaviest substance collecting at the bottom.*

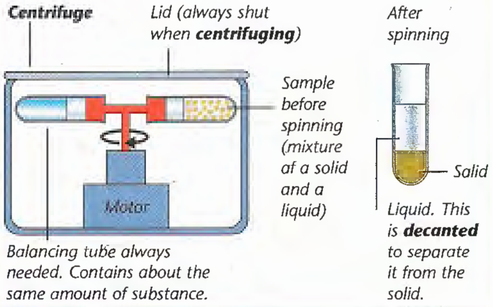
Centrifuge
*Lid (always shut when **centrifuging**)*
After spinning
Sample before spinning (mixture of a solid and a liquid)
Solid
Motor
Balancing tube always needed. Contains about the same amount of substance.
*Liquid. This is **decanted** to separate it from the solid.*

Distillation

*The process of separating a mixture of liquids, or a liquid from an impurity, by heating. The vapor of the liquid with the lowest boiling point comes off first and is condensed back to a liquid in a **Liebig condenser** (see picture below).*

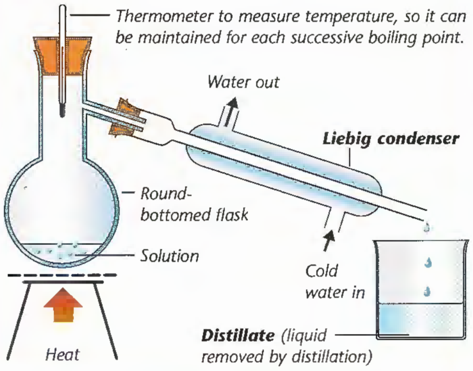
Thermometer to measure temperature, so it can be maintained for each successive boiling point.
Water out
Liebig condenser
Round-bottomed flask
Solution
Cold water in
Heat
***Distillate** (liquid removed by distillation)*

Fractional distillation

*A **distillation** process which separates two or more liquids with close boiling points, using a **fractionating column**. The vapor of the liquid with the lowest boiling point reaches the top of the column first. Small columns are used in laboratories (see picture below). Other columns are much larger and have many points at which different vapors are condensed and collected (see also pages 183 and 198).*

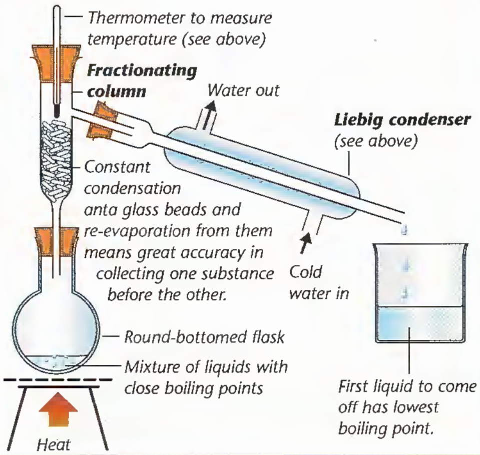
Thermometer to measure temperature (see above)
Fractionating column
Water out
***Liebig condenser** (see above)*
Constant condensation onto glass beads and re-evaporation from them means great accuracy in collecting one substance before the other.
Cold water in
Round-bottomed flask
Mixture of liquids with close boiling points
First liquid to come off has lowest boiling point.
Heat

Solvent extrac

The process of
solute* *by transf*
*original **solvent****
it is more soluble,
it can be easily
method of separa
when the solute
heated, and makes
particular propert
i.e. whether they
non-polar solvents
***extraction** is an*

Chromatograp

The process of
a mixture by the
*a medium (the **stati***
methods of chrom
*in a **solvent**** *(the*
chromatograph
they vary in their

Paper chromato

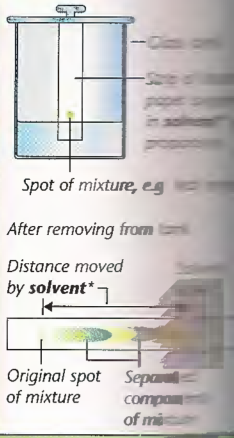
Spot of mixture, e.g
After removing from
*Distance moved by **solvent****
Original spot of mixture
Sep
compo
of mi

Desiccation

The process of remov
of crystallization*
*large glass **desiccators***
silica gel. Water is
bringing them into
anhydrous* *calcium*
in the case of liquids,

Desiccator

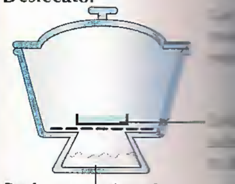

Drying agent* *not in contact with solid. Absorbs moisture from air, causing water to evaporate from*

* **Anhydrous**, 154 (**Anhy**
Non-polar solvent, Polar
Volatile, 345; **Water of**

Solvent extraction

*The process of obtaining a **solute*** by transferring it from its original **solvent*** to one in which it is more soluble, and from which it can be easily removed. It is a method of separation often used when the solute cannot be heated, and makes use of a particular property of the solvents, i.e. whether they are **polar** or **non-polar solvents***. **Ether extraction** is an example.*

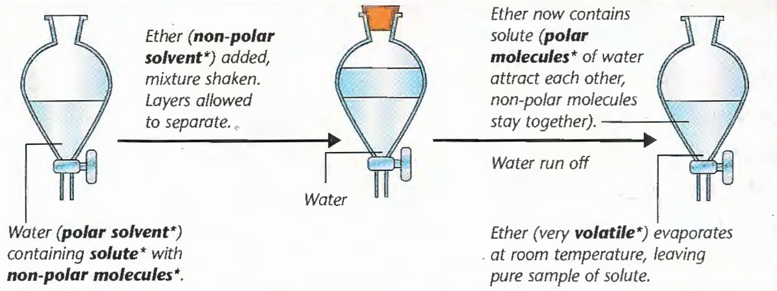
Ether (non-polar solvent*) added, mixture shaken. Layers allowed to separate.
Ether now contains solute (polar molecules* of water attract each other, non-polar molecules stay together).
Water run off
Water
Water (polar solvent*) containing solute* with non-polar molecules*.
Ether (very volatile*) evaporates at room temperature, leaving pure sample of solute.

Chromatography

*The process of separating small amounts of substances from a mixture by the rates at which they move through or along a medium (the **stationary phase**, e.g. blotting paper). Most methods of chromatography involve dissolving the mixture in a **solvent*** (the **eluent**), though it is vaporized in **gas chromatography**. Substances move at different rates because they vary in their **solubility*** and their attraction to the medium.*

Paper chromatography

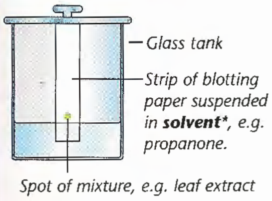
Glass tank
Strip of blotting paper suspended in solvent*, e.g. propanone.
Spot of mixture, e.g. leaf extract

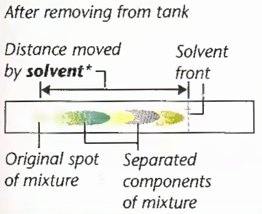
After removing from tank
Distance moved by solvent*
Solvent front
Original spot of mixture
Separated components of mixture

*Standard tables identify substances by **R_f value** – distance moved by substance over distance moved by **solvent***.*

*There are several methods of chromatography, including **column chromatography** (components in the mixture are separated in a column containing a solvent and a material that attracts molecules) and **gas chromatography** (vaporized mixture is separated as it passes along a heated column in a stream of gas).*

Crystallization

*The process of forming crystals from a solution, which can be used to produce a pure sample of a substance, as the impurities will not form crystals. To make pure crystals, a hot, **saturated*** solution of the substance is cooled and the crystals formed on cooling are removed by **filtering**. See also page 135.*

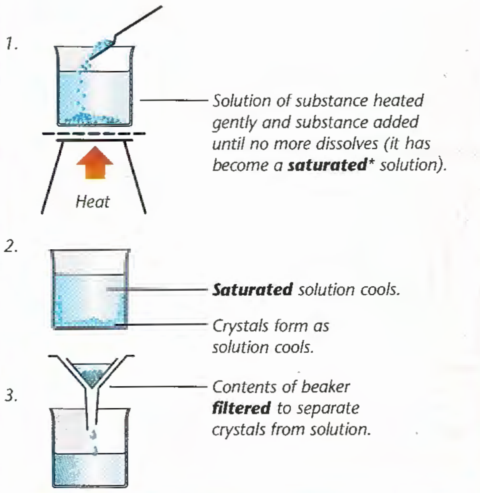
1.
Solution of substance heated gently and substance added until no more dissolves (it has become a saturated* solution).
Heat
2.
Saturated solution cools.
Crystals form as solution cools.
3.
Contents of beaker filtered to separate crystals from solution.

Desiccation

*The process of removing water mixed with a substance, or **water of crystallization*** from a substance. Solids are often dried in large glass **desiccators** that contain a **drying agent*** such as silica gel. Water is removed from most gases and liquids by bringing them into direct contact with a drying agent, e.g. **anhydrous*** calcium chloride (which absorbs the water and, in the case of liquids, is then **filtered** off).*

Desiccator

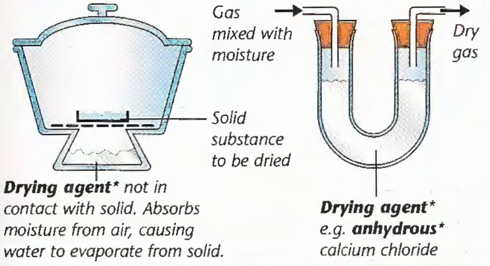
Gas mixed with moisture
Dry gas
Solid substance to be dried
Drying agent* not in contact with solid. Absorbs moisture from air, causing water to evaporate from solid.
Drying agent* e.g. anhydrous* calcium chloride

Melting point and boiling point tests

Tests are used to determine the purity of a sample. A pure sample of a substance has a particular known melting point and boiling point, and any impurities in a sample will alter these measurements.

Measuring the melting point

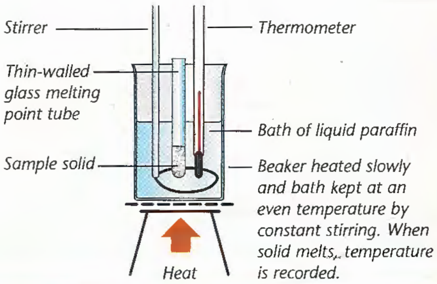
Stirrer
Thermometer
Thin-walled glass melting point tube
Bath of liquid paraffin
Sample solid
Beaker heated slowly and bath kept at an even temperature by constant stirring. When solid melts, temperature is recorded.
Heat

* **Anhydrous**, 154 (**Anhydrate**); **Drying agent**, 344; **Non-polar molecule**, 133 (**Polar molecule**); **Non-polar solvent**, **Polar solvent**, 144; **Saturated**, 144; **Solubility**, 145; **Solute**, **Solvent**, 144; **Volatile**, 345; **Water of crystallization**, 135.

QUALITATIVE AND QUANTITATIVE ANALYSIS

There are two types of analysis to investigate substances: **qualitative analysis** – any method used to study chemical composition – and **quantitative analysis** – any method used to discover how much of a substance is present in a sample. Below are some examples of both types of analysis.

Qualitative analysis

Below are some examples of qualitative analysis. The **flame test** and the tests on pages 218-219 are examples of qualitative analysis used in schools. The other methods described are more advanced.

Flame test

Used to identify metals. A substance is collected on the tip of a clean platinum or nichrome wire. This is held in a flame to observe the color with which the substance burns (see also page 219). Between tests, the wire is cleaned by dipping it in concentrated hydrochloric acid and then heating strongly.

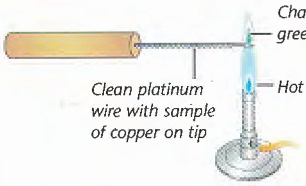
Change of flame color to green where copper burning
Clean platinum wire with sample of copper on tip
Hot bunsen flame

Mass spectroscopy

A method of investigating the composition of a substance, in particular the ***isotopes**** *it contains. It is also used as a method of quantitative analysis as it involves measuring the relative proportions of isotopes or molecules in the substance. The apparatus used is called a* ***mass spectrometer****.*

Mass spectrometer

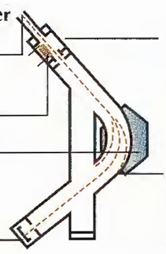
Vaporized sample of substance
Electron gun
Magnet
Ion detector
High energy electrons produced to **ionize*** the substance. The positive ions pass into an electric field which accelerates them.
A magnetic field deflects the ions of different masses by different amounts.

Nuclear magnetic resonance (n.m.r) spectroscopy

A method used for investigating the position of atoms in a molecule. Radio waves are passed through a sample of a substance held between the poles of a magnet. The amount of absorption reveals the positions of particular atoms within a molecule. This information is presented on a graph called a ***nuclear magnetic resonance (n.m.r.) spectrum****.*

N.m.r. spectrum of ethanol* (CH_3CH_2OH)

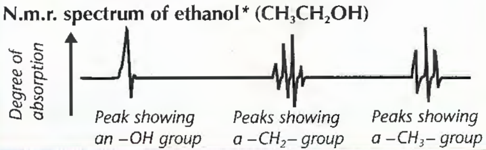

Quantitative analysis

Below are some examples of quantitative analysis. See also **mass spectroscopy.**

Volumetric analysis

A method of determining the concentration of a solution using ***titration****. This is the addition of one solution into another, using a* ***burette****. The concentration of one solution is known. The first solution is added from the burette until the* ***end point****, when all the second solution has reacted. (The end point is detected by using an* ***indicator****.) The volume of solution from the burette needed to reach the end point is called the* ***titer****. This, the volume of solution in the flask and the known concentration of one solution are used to calculate the concentration of the second solution.*

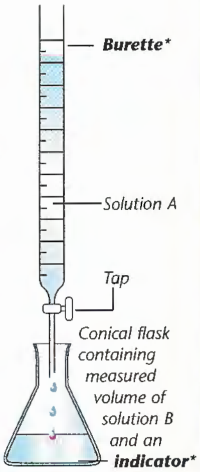
Burette*
Solution A
Tap
Conical flask containing measured volume of solution B and an **indicator***

Apparatus used for ***titrations***

Gravimetric analysis

A method of determining the amount of a substance present by converting it into another substance of known chemical composition that is easily purified and weighed.

Gravimetric analysis *can be used to measure the amount of lead in a sample of water containing a lead salt.*

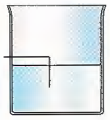

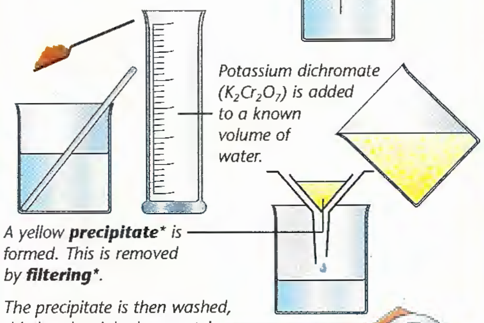
Potassium dichromate ($K_2Cr_2O_7$) is added to a known volume of water.

A yellow ***precipitate**** *is formed. This is removed by* ***filtering****.*

The precipitate is then washed, dried and weighed accurately.

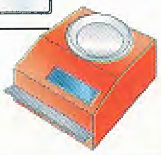

The concentration of the lead in the sample of water is calculated from the volume of water, the weight of lead chromate and the ***relative atomic mass**** *of lead.*

---

* **Burette**, 223; **Ethanol**, 196; **Filtering**, 220; **Ionization**, 130; **Indicator**, 152; **Isotope**, 127; **Precipitate**, 145; **Relative atomic mass**, 138.

APPARATUS

Apparatus is chemical equipment. The most common items are described and illustrated below and on pages 224-225. Simple 2-D diagrams used to represent them are also shown, together with approximate ranges of sizes.

Beaker

Used to hold liquids. Shows approximate volume.

Possible capacities: 5-5,000ml

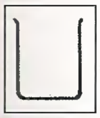

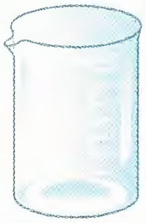

Beehive shelf

*Used to support a **gas jar*** while gas is being collected by the displacement of water. For examples of its use, see pages 216-217.*

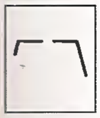

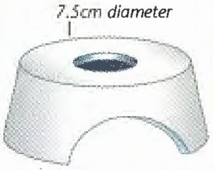

Bunsen burner

Used to provide heat for chemical reactions. Its adjustable air-hole allows some control of the flame temperature. If the hole is closed, the flame is yellow and cooler than the blue flame produced when the hole is open. See picture, page 208.

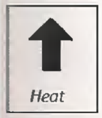

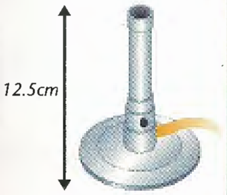

Burette

*Used to add accurate volumes of liquid during **titrations** (see **volumetric analysis**, page 222).*

Possible capacities: 10-100ml

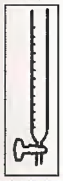

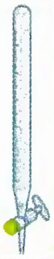

Condensers

Liebig condenser

*Used to condense vapors. Vapor passes through the central channel and is cooled by water flowing through the outer pipe. See **distillation**, page 220.*

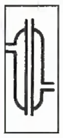

Possible lengths: 25-50cm

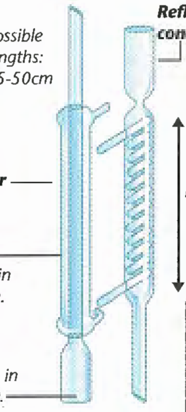
Reflux condenser
Liebig condenser
Length: 15.0cm
Water circulates in outer tube.
Vapor condenses in inner tube.

Reflux condenser

Used to return vapor to a liquid to prevent loss by evaporation.

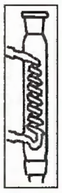

Crucible

*Used to hold small quantities of solids which are being heated strongly, either in a furnace or over a **bunsen burner**. They are made of porcelain, silica, fireclay, nickel or steel.*

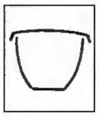

Possible diameters: 2.5-5.5cm

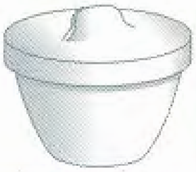

Crystallizing dish

Used to hold solutions which are being evaporated to form crystals. The flat bottom helps to form an even layer of crystals.

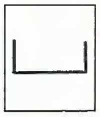

Possible capacities: 100-2,000ml

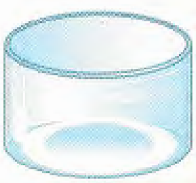

Delivery tube

A tube used to carry gases.

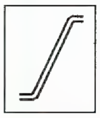

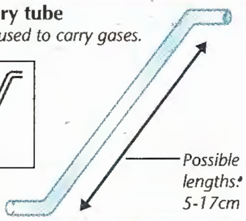

Desiccator

*A glass container used to dry solids. It contains a **drying agent***. See **desiccation**, page 221.*

22.0cm diameter

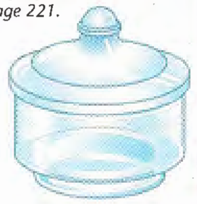

Evaporating basin

*Used to hold a solution whose **solvent*** is being separated from the **solute*** by evaporation (often using heat).*

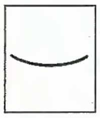

Possible capacities: 50-500ml

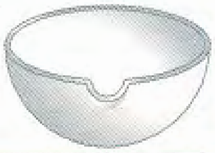

Filter paper

*Paper which acts as a strainer, only allowing liquids through, but no solid matter. Filter paper is graded according to how finely it is meshed, i.e. the size of particle it allows through. It is put in a **filter** or **Buchner funnel*** to give support as the liquid passes through, and the solid settles on the paper. See **filtering**, page 220.*

*Mesh of fine **filter paper**, magnified many times.*

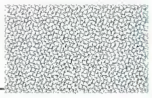

*Mesh of coarse **filter paper**, magnified many times.*

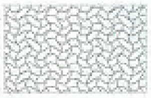

Holes between fibers allow tiny particles to pass through.

***Buchner funnel**, 224; **Drying agent**, 344;
Filter funnel, **Gas jar**, 224; **Solute**, **Solvent**, 144.

Flasks

Buchner flask

*Used when liquids are filtered by suction. See **filtering**, page 220.*

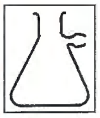

Possible capacities: 250-1,000ml

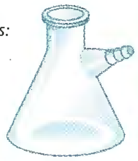

Conical flask

*Used to hold liquids when carrying out reactions and preparing solutions of known concentration. They are used in preference to beakers when it is necessary to have a container that can be stoppered. They have some volume markings but these are not as accurate as the markings on a **pipette** or **burette***.*

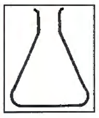

Possible capacities: 25-2,000ml

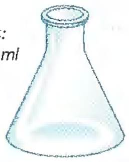

Flat-bottomed flask

Used to hold liquids when carrying out reactions where heating is not required (the flask stands on the workbench).

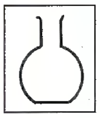

Possible capacities: 100-2,000ml

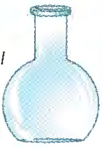

Round-bottomed flask

Used to hold liquids, especially when even heating is needed. Volume markings are approximate. It is held in position above the flame by a clamp.

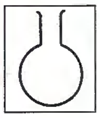

Possible capacities: 100-2,000ml

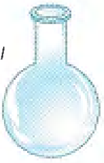

Volumetric flask

Used when mixing accurate concentrations of solutions. Each flask has a volume marking which is very exact and a stopper so that it can be shaken to mix the solution.

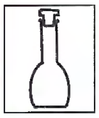

Possible capacities: 10-2,000ml

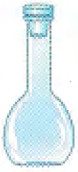

Fractionating column

*Used to separate components of a mixture by their boiling points. It contains glass balls or rings that provide a large surface area and thus promote condensation and re-evaporation. See **fractional distillation**, page 220.*

Possible lengths: 15-36cm

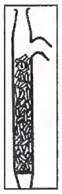

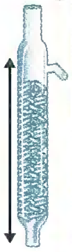

Fume cupboard

A glass panelled cupboard that contains an extractor fan and encloses an area of workbench. Dangerous experiments are carried out in a fume cupboard.

Funnels

Buchner funnel

*Used when liquids are filtered by suction. It has a flat, perforated plate, on which **filter paper*** is placed. See **filtering**, page 220.*

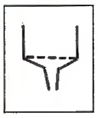

Possible capacities: 50-500ml

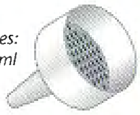

Tap funnel

For adding a liquid to a reaction mixture drop by drop. See pages 216-217.

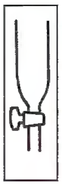

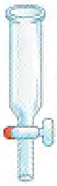

Filter funnel

*Used when separating solids from liquids by **filtering** (see page 220). **Filter paper*** is put inside the funnel.*

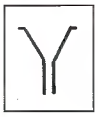

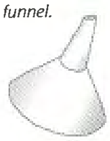

Thistle funnel

Used when adding a liquid to a reaction mixture.

Length: 30cm

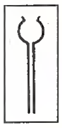

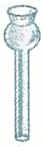

Separating funnel

*Used when separating **immiscible*** liquids. First the denser liquid is run off, then the less dense. See **solvent extraction**, page 221.*

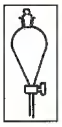

Possible capacities: 50-500ml

Gas jar

Used when collecting and storing gases. The jar can be sealed using a glass lid whose rim is coated with a thin layer of grease. See pages 216-217.

Possible heights: 15-30cm

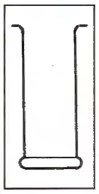

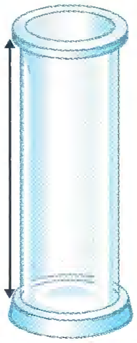

Gas syringe

Used to measure the volume of a gas. It is used both to receive gas and to inject gas into a reaction vessel.

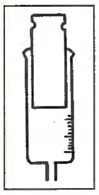

Capacity: 100ml

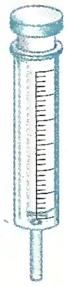

Gauze

Used to spread the heat from a flame evenly over the base of an object being heated. Made of iron, steel, copper or ceramics.

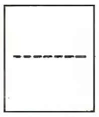

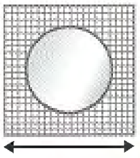

Length: 12.5cm

Measuring cylinder

Used to measure the approximate volume of liquids.

Possible capacities: 5-2,000ml

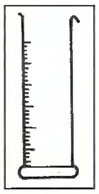

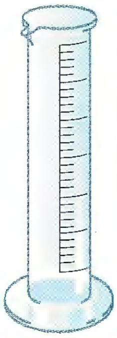

***Burette**, **Filter paper**, 223; **Immiscible**, 145 (**Miscible**).

Pipeclay triangle

Used to support **crucibles*** on **tripods** when they are being heated. They are made of iron or nickel-chromium wire enclosed in pipeclay tubes.

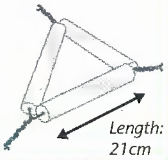

Pipettes

Pipette

Used to dispense accurate volumes of liquid. They come in different sizes for different volumes. The liquid is run out of the pipette until its level has dropped from one volume marking to the next.

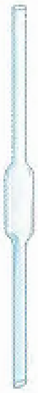

Possible capacities: 1-100ml

Dropping pipette or teat pipette

Used to dispense small volumes or drops of liquid. It does not provide an accurate measurement.

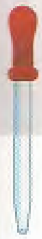

Possible capacities: 1-2ml

Stands and clamps

Used to hold apparatus, e.g. **round-bottomed flasks**, in position.

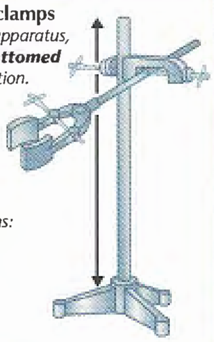

Possible lengths: 50-100cm

Spatula

Used to pick up small quantities of a solid.

Possible lengths: 10-20cm

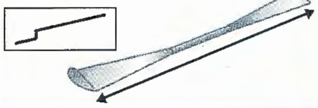

Test tube holder

Used to hold a test tube, e.g. when heating it in a flame, creating a chemical reaction within it, or transferring it from one place to another.

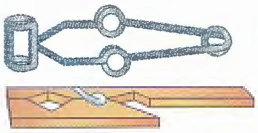

Test tube rack

Used to hold many test tubes upright.

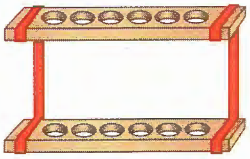

Thermometer

Used to measure temperature. They are filled either with alcohol or with mercury, depending on the temperature range for which they are intended.

Small temperature range: -10 to 50°C

Large temperature range: -10 to 400°C

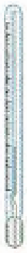

Tongs

Used to move hot objects.

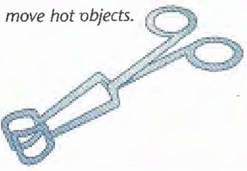

Top pan balances

Used for quick, accurate weighing.

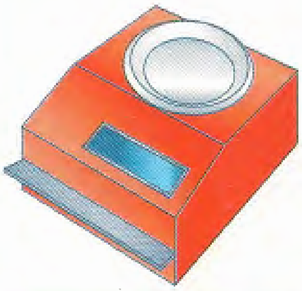

Trough

Used when collecting gas **over water** (see **carbon dioxide**, page 216). The water contained in a gas jar inverted in the trough is displaced into the trough. Troughs are also used when substances such as potassium are reacted (see picture, page 169). Possible diameters: 20-30cm.

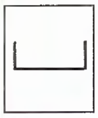

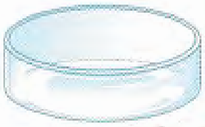

Tripod

Used with a **pipeclay triangle** or **gauze** when heating **crucibles***, **flasks**, etc.

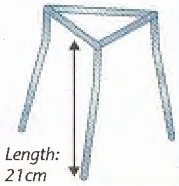

Tubes

Boiling tube

A thick-walled tube used to hold substances being heated strongly.

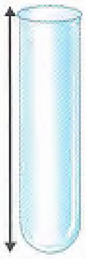

Possible length: 12.5cm

Test tube

A tube used to hold substances for simple chemical reactions not involving strong heating.

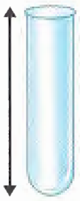

Possible length: 7.5cm

Ignition tube

A disposable tube used to hold small quantities of substances being melted or boiled.

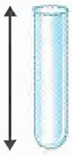

Possible length: 5.0cm

Watch glass

Used when evaporating small quantities. Possible diameters: 5-15cm

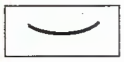

---

* **Crucible**, 223.

CHART OF SUBSTANCES, SYMBOLS AND FORMULAS

Below is a list of the symbols and formulas used in the chemistry section of this book. Each one is followed by the name of the substance it stands for[†]. Capital letters come alphabetically before small ones, i.e. each element is kept together with its compounds. For example, **CH_3OH** (methanol – a carbon compound) is found in an alphabetical list after **C** (carbon), before the **Ca** (calcium) list begins.

Symbol	Substance
$3Ca_3(PO_4)_2.CaF_2$	Apatite
Ac	Actinium
Ag	Silver
AgBr	Silver bromide
AgCl	Silver chloride
AgI	Silver iodide
$AgNO_3$	Silver nitrate
Al	Aluminum
$Al(OH)_3$	Aluminum hydroxide
Al_2O_3	Aluminum oxide
$Al_2O_3.2H_2O$	Bauxite
$Al_2(SO_4)_3$	Aluminum sulfate
Am	Americium
Ar	Argon
As	Arsenic
At	Astatine
Au	Gold
B	Boron
B_2O_3	Boron oxide
BCl_3	Boron trichloride
Ba	Barium
$BaCl_2$	Barium chloride
Be	Beryllium
Bi	Bismuth
Bk	Berkelium
Br/Br_2	Bromine
$-Br$	Bromo group
C	Carbon
C_2H_2	Ethyne
C_2H_4	Ethene
C_2H_5Br	Bromoethane
C_2H_5CHO	Propanal
C_2H_5Cl	Chloroethane
C_2H_5COOH	Propanoic acid
C_2H_5OH	Ethanol
C_2H_6	Ethane
C_3H_4	Propyne
C_3H_6	Propene
C_3H_6O	Propanone
C_3H_7OH	Propan-1-ol
C_3H_8	Propane
C_4H_6	But-1-yne
C_4H_8	But-1-ene
C_4H_9OH	Butan-1-ol
C_4H_{10}	Butane
C_5H_{10}	Pent-1-ene
C_5H_{12}	Pentane
$C_6H_8O_6$	Ascorbic acid
$C_6H_{12}O_6$	Glucose
C_6H_{14}	Hexane
C_7H_{16}	Heptane
C_8H_{18}	Octane
C_9H_{20}	Nonane
$C_{12}H_{22}O_{11}$	Sucrose
$C_{17}H_{35}COOH$	Octadecanoic acid
CCl_4	Tetrachloromethane
CH_2BrCH_2Br	1,2-dibromoethane
CH_2CHCl	Vinyl chloride
$-CH_3$	Methyl group
CH_3CCH	Propyne
CH_3CH_2CCH	But-1-yne
$CH_3CH_2CH_2CH_2OH$	Butan-1-ol
$CH_3CH_2CH_2OH$	Propan-1-ol
CH_3CH_2CHO	Propanal
CH_3CH_2Cl	Chloroethane
CH_3CH_2COOH	Propanoic acid
CH_3CH_2OH	Ethanol
CH_3CH_2ONa	Sodium ethoxide
CH_3CHO	Ethanal
$CH_3CHOHCH_3$	Propan-2-ol
CH_3Cl	Chloromethane
$CH_3COCH_2CH_3$	Butanone
CH_3COCH_3	Propanone
$CH_3COOCH_2CH_3$	Ethyl ethanoate
CH_3COOH	Ethanoic acid
CH_3NH_2	Methyl amine
CH_3OCH_3	Methoxymethane
CH_3OH	Methanol
CH_4	Methane
CHCH	Ethyne
CO	Carbon monoxide
$-CO-$	Carbonyl group
CO_2	Carbon dioxide
$-COOH$	Carboxyl group
$(COOH)_2$	Ethanedioic acid
$COOH(CH_2)_4COOH$	Hexanedioic acid
Ca	Calcium
$Ca_3(PO_4)_2$	Calcium phosphate
$CaCl_2$	Calcium chloride
$CaCO_3$	Calcium carbonate
$CaCO_3.MgCO_3$	Dolomite
CaF_2	Fluorospar
$Ca(HCO_3)_2$	Calcium hydrogencarbonate
CaO	Calcium oxide
$Ca(OH)_2$	Calcium hydroxide
$CaSiO_3$	Calcium metasilicate
$CaSO_4$	Calcium sulfate
$CaSO_4.2H_2O$	Gypsum
Cd	Cadmium
Ce	Cerium
Cf	Californium
Cl/Cl_2	Chlorine
$-Cl$	Chloro group
Cm	Curium
Co	Cobalt
$CoCl_2$	Cobalt(II) chloride
Cr	Chromium
Cs	Cesium
Cu	Copper
Cu_2O	Copper(I) oxide
CuCl	Copper(I) chloride
$CuCl_2$	Copper(II) chloride
$CuCO_3.Cu(OH)_2$	Malachite
$(CuFe)S_2$	Copper pyrites
$(Cu(NH_3)_4)SO_4$	Tetraammine copper(II) sulfate
$Cu(NO_3)_2$	Copper(II) nitrate
CuO	Copper(II) oxide
$CuSO_4$	Copper(II) sulfate
$CuSO_4.3Cu(OH)_2$	Basic copper sulfate
D	Deuterium
D_2O	Deuterium oxide
Dy	Dysprosium
Er	Erbium
Es	Einsteinium
Eu	Europium
F/F_2	Fluorine
$-F$	Fluoro group
Fe	Iron
Fe_2O_3	Haematite
$Fe_2O_3.xH_2O$	Rust
$FeCl_2$	Iron(II) chloride
$FeCl_3$	Iron(III) chloride
$Fe(OH)_3$	Iron(III) hydroxide
FeS	Iron(II) sulfide
$FeSO_4$	Iron(II) sulfate
Fm	Fermium
Fr	Francium
Ga	Gallium

[†] If you know a substance, but not its symbol, use the index (pages 346-381).

Symbol	Substance
Gd	Gadolinium
Ge	Germanium
H/H_2	Hydrogen
H_2CO_3	Carbonic acid
H_2O	Water
H_2O_2	Hydrogen peroxide
H_2S	Hydrogen sulfide
$H_2S_2O_7$	Fuming sulfuric acid
H_2SO_3	Sulfurous acid
H_2SO_4	Sulfuric acid
H_3PO_4	Phosphoric acid
HBr	Hydrogen bromide
HCl	Hydrogen chloride/ Hydrochloric acid
HCHO	Methanal
HCOOH	Methanoic acid
HI	Hydrogen iodide
HNO_2	Nitrous acid
HNO_3	Nitric acid
He	Helium
Hf	Hafnium
Hg	Mercury
HgS	Cinnabar
Ho	Holmium
I/I_2	Iodine
In	Indium
Ir	Iridium
K	Potassium
K_2CO_3	Potassium carbonate
$K_2Cr_2O_7$	Potassium dichromate
K_2SO_4	Potassium sulfate
$K_2SO_4.Al_2(SO_4)_3$	Aluminum potassium sulfate-12-water
KBr	Potassium bromide
KCl	Potassium chloride
KI	Potassium iodide
$KMnO_4$	Potassium permanganate
KNO_3	Potassium nitrate
KOH	Potassium hydroxide
Kr	Krypton
KrF_2	Krypton fluoride
La	Lanthanum
La_2O_3	Lanthanum oxide
Li	Lithium
Li_3N	Lithium nitride
LiCl	Lithium chloride
LiOH	Lithium hydroxide
Lr or Lw	Lawrencium
Lu	Lutetium
Md	Mendelevium
Mg	Magnesium
$MgCl_2$	Magnesium chloride
$MgCO_3$	Magnesium carbonate
MgO	Magnesium oxide
$Mg(OH)_2$	Magnesium hydroxide
$MgSO_4$	Magnesium sulfate
Mn	Manganese
$MnCl_2$	Manganese(IV) chloride
MnO_2	Pyrolusite/ Manganese(IV) oxide
Mo	Molybdenum
N/N_2	Nitrogen
N_2O	Dinitrogen oxide
N_2O_4	Dinitrogen tetraoxide
$-NH_2$	Amino group
$NH_2(CH_2)_6NH_2$	1,6-diaminohexane
NH_3	Ammonia
$(NH_4)_2SO_4$	Ammonium sulfate
NH_4Cl	Ammonium chloride
NH_4OH	Ammonia solution
NH_4NO_3	Ammonium nitrate
NO	Nitrogen monoxide
NO_2	Nitrogen dioxide
Na	Sodium
Na_2CO_3	Sodium carbonate
$Na_2CO_3.10H_2O$	Washing soda
Na_2SO_3	Sodium sulfite
Na_2SO_4	Sodium sulfate
Na_3AlF_6	Cryolite
$NaAl(OH)_4$	Sodium aluminate
NaBr	Sodium bromide
NaCl	Sodium chloride
$NaClO_3$	Sodium chlorate
$NaHCO_3$	Sodium hydrogencarbonate
$NaHSO_4$	Sodium hydrogensulfate
$NaIO_3$	Sodium iodate
$NaNO_2$	Sodium nitrite
$NaNO_3$	Sodium nitrate
NaOCl	Sodium hypochlorite
NaOH	Sodium hydroxide
Nb	Niobium
Nd	Neodymium
Ne	Neon
Ni	Nickel
NiS	Nickel sulfide
No	Nobelium
Np	Neptunium
O/O_2	Oxygen
O_3	Ozone
$-OH$	Hydroxyl group
Os	Osmium
OsO_4	Osmium tetroxide
P	Phosphorus
P_2O_5	Phosphorus pentoxide
Pa	Protactinium
Pb	Lead
PbI_2	Lead(II) iodide
$Pb(NO_3)_2$	Lead(II) nitrate
PbO	Lead(II) oxide
PbO_2	Lead(IV) oxide
$Pb(OC_2H_5)_4$	Tetraethyl-lead
$Pb(OH)_2$	Lead(II) hydroxide
PbS	Galena
Pd	Palladium
Pm	Promethium
Po	Polonium
Pr	Praseodymium
Pt	Platinum
Pu	Plutonium
Ra	Radium
Rb	Rubidium
Re	Rhenium
Rh	Rhodium
Rn	Radon
Ru	Ruthenium
S	Sulfur
SO_2	Sulfur dioxide
SO_3	Sulfur trioxide
Sb	Antimony
Sc	Scandium
Se	Selenium
Si	Silicon
SiO_2	Silicon dioxide
Sm	Samarium
Sn	Tin
Sr	Strontium
T	Tritium
Ta	Tantalum
Tb	Terbium
Tc	Technetium
Te	Tellurium
Th	Thorium
Ti	Titanium
Tl	Thallium
Tm	Thulium
U	Uranium
V	Vanadium
V_2O_5	Vanadium pentoxide
W	Tungsten
Xe	Xenon
$XeFe_4$	Xenon tetrafluoride
Y	Yttrium
Yb	Ytterbium
Zn	Zinc
$ZnCl_2$	Zinc chloride
$ZnCO_3$	Calamine
ZnO	Zincite/ Zinc oxide
$Zn(OH)_2$	Zinc hydroxide
$Zn(OH)Cl$	Basic zinc chloride
ZnS	Zinc blende
$ZnSO_4$	Zinc sulfate
Zr	Zirconium

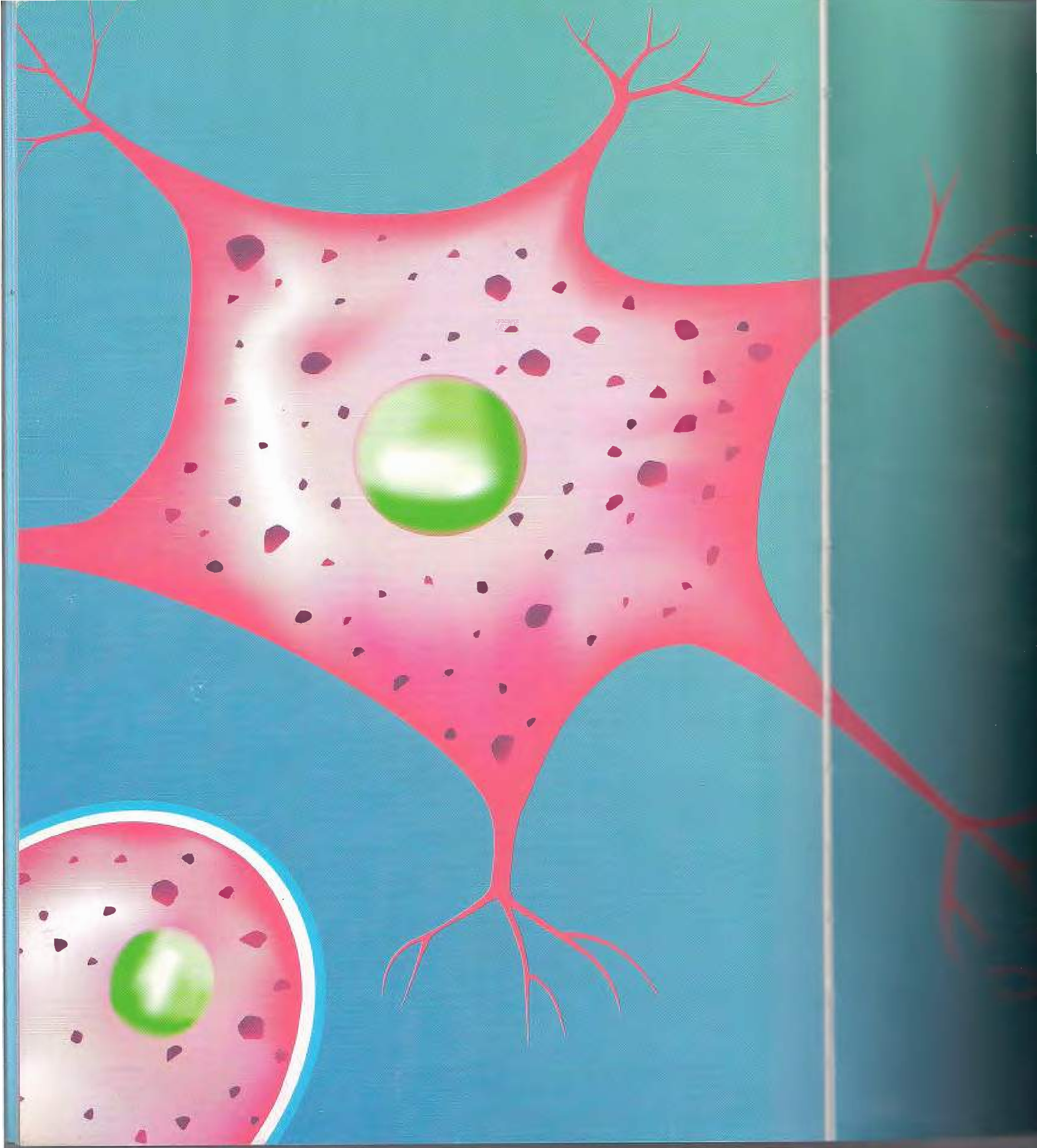

BIOLOGY

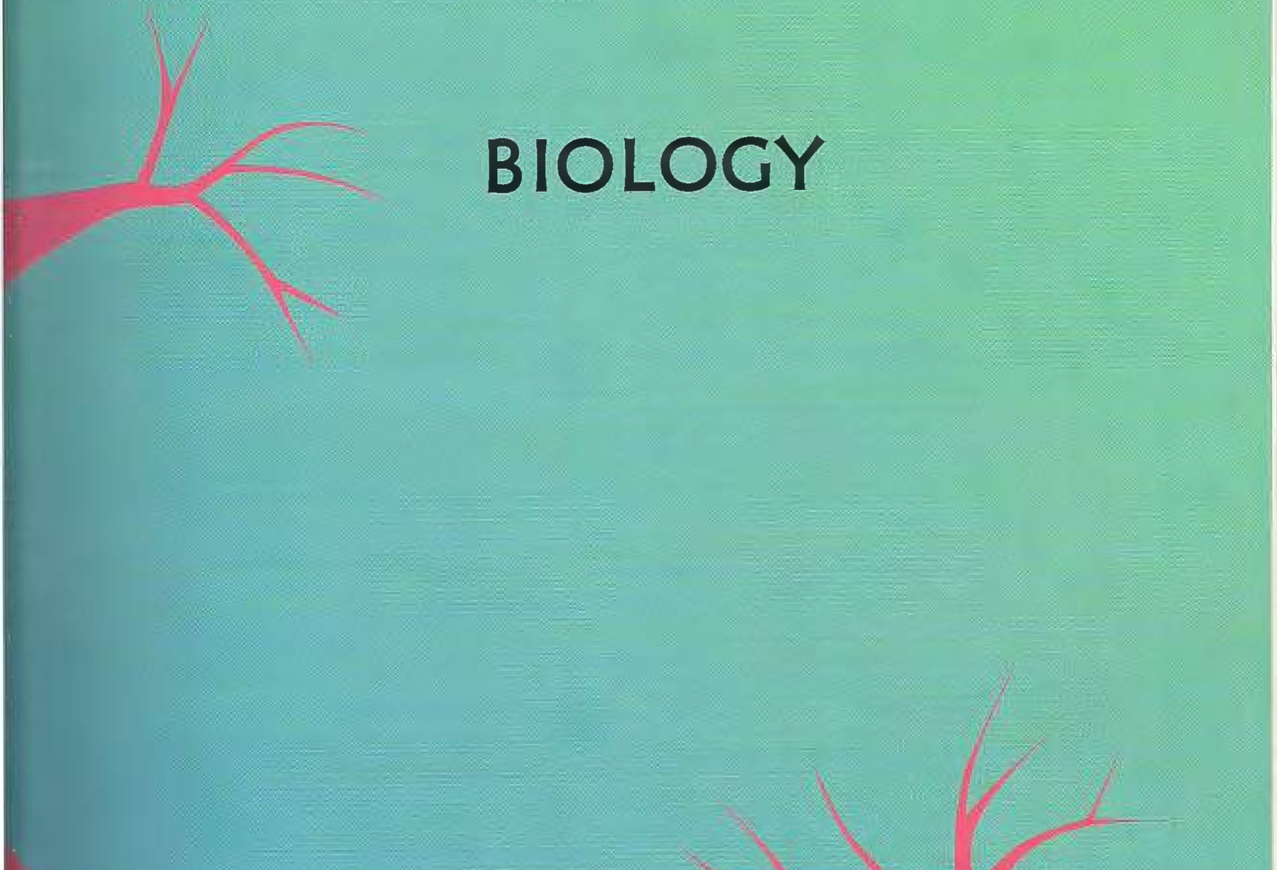

ABOUT BIOLOGY

Biology is the study of living things. It examines the structures and internal systems of different organisms and how these operate to sustain individual life, as well as looking at the complex web of relationships between organisms which ensure new life is created and maintained. In this book, biology is divided into six color-coded sections. The areas covered by these sections are explained below.

● Ecology and living things

Looks at the complex relationships between all living things, and their basic cellular structure.

● Zoology (humans)

Covers all the major terms of human biology. In many cases, these also apply to the vertebrates in general (see page 341).

● Botany

Covers the plant kingdom. Introduces the different types of plants, their main characteristics, internal structures and systems.

● Reproduction and genetics

Examines the different types of reproduction and introduces the branch of biology known as genetics.

● Zoology (animals)

Examines the component parts, systems and behavior typical of the major animal groups.

● General biology information

Covers subjects which relate to all living things. Includes tables of general information and classification charts.

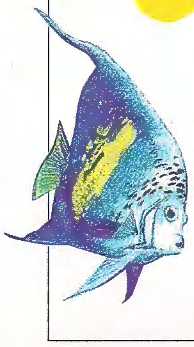

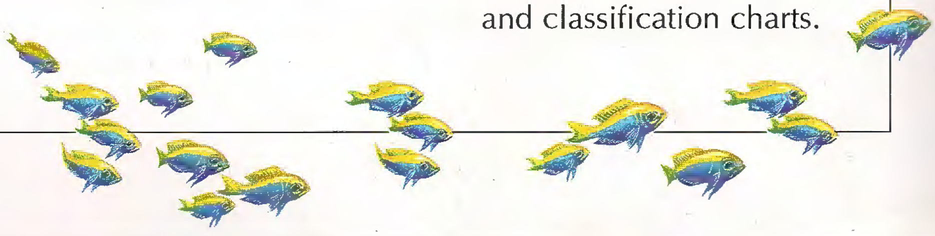

CONTENTS

Ecology and living things

232 Living things and their environment
234 Within an ecosystem
236 Life and life cycles
238 The structure of living things
240 Cell division

Botany

242 Vascular plants
244 Stems and roots
246 Inside an older plant
248 Leaves
251 Plant sensitivity
252 Plant fluid transportation
254 Plant food production
256 Flowers
258 Reproduction in a flowering plant
260 Seeds and germination
262 Fruit / Vegetative reproduction

Zoology (animals)

264 The body structure of animals
266 Animal body coverings
268 Animal movement
270 Animal feeding
272 Animal respiration
273 Animal excretion
274 Animal senses and communication
276 Animal reproduction and development

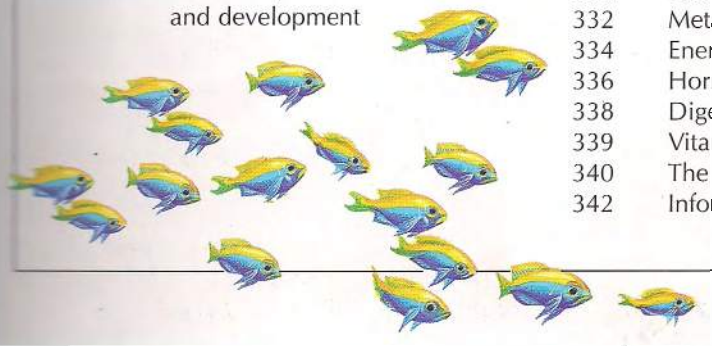

Zoology (humans)

278 The skeleton
280 Joints and bone
282 Muscles
284 Teeth
286 Blood
288 The circulatory system
290 The heart
292 Tissue fluid and the lymphatic system
294 The digestive system
296 Glands
298 The respiratory system
300 The urinary system
302 The central nervous system
304 The units of the nervous system
306 Nerves and nervous pathways
310 The skin
312 The eyes
314 The ears
316 The reproductive system
318 Development and reproduction

Reproduction and genetics

320 Types of reproduction
322 Cell division for reproduction
324 Genetics and heredity
327 Genetic engineering

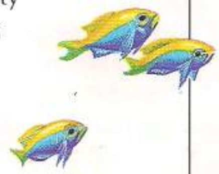

General biology information

329 Fluid movement
330 Food and how it is used
332 Metabolism
334 Energy for life and homeostasis
336 Hormones
338 Digestive juices and enzymes
339 Vitamins and their uses
340 The classification of living things
342 Informal group terms

LIVING THINGS AND THEIR ENVIRONMENT

The world can be divided into a number of different regions, each with its own characteristic plants and animals. All the plants and animals have become adapted to their own surroundings, or **environment** (see **adaptive radiation**, page 237), and their lives are linked in a complex web of interdependence. The environment is influenced by many different factors, e.g. temperature, water and light (**climatic factors**), the physical and chemical properties of the soil (**edaphic factors**), and the activities of living things (**biotic factors**). The study of the relationships between plants, animals and the environment is called **ecology**.

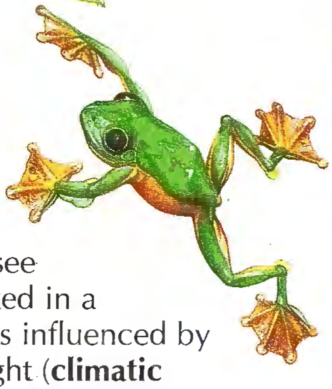

This tree frog's toes have adapted to help it cling to bark.

Biosphere

The layer of the Earth (including the oceans and the atmosphere) which is inhabited by living things. The biosphere's boundaries are the upper atmosphere (above) and the first layers of uninhabited rock (below).

Biomes

The main ecological regions into which the land surface can be divided. Each has its own characteristic seasons, day length, rainfall pattern and maximum and minimum temperatures. The major biomes (see map, above right) are **tundra**, **coniferous forest**, **deciduous forest**, **tropical forest**, **temperate grassland**, **savanna** (tropical grassland) and **desert**. Most are named after the dominant vegetation, since this determines all other living things found there. Each biome is a giant **habitat** (**macrohabitat**). Human activity, e.g. deforestation in tropical forests, has begun to have harmful effects on the habitats which exist within many biomes.

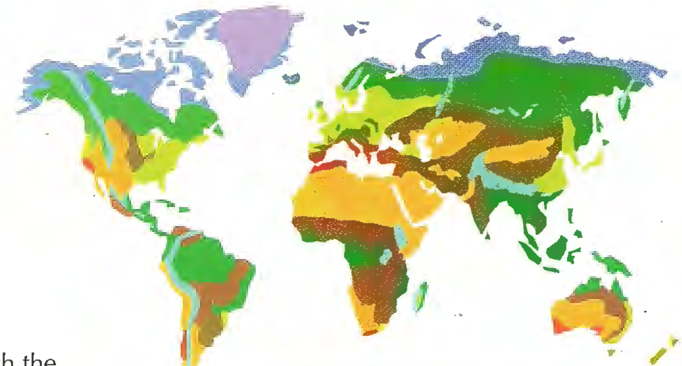

Map showing main world biomes

Key to biomes on map above

🟦 ***Tundra***
Very cold and windy. Most common plants: ***lichens*** *and small shrubs. Animals include musk ox.*

🟩 ***Coniferous forest***
Low temperatures all year. Dominant plants: ***conifers****, e.g. spruce. Most common large animals: deer.*

🟨 ***Deciduous forest***
Summers warm, winters cold. Dominant plants: ***deciduous*** *trees, e.g. beech. Many animals, e.g. foxes.*

🟩 ***Tropical forest***
High temperatures all year, heavy rainfall. Great variety of plants and animals, e.g. exotic birds.

🟧 ***Desert***
High temperatures (cold at night), very low rainfall. Typical plants: cacti. Animals include jerboas, scorpions.

🟫 ***Temperate grassland***
Open grassy plains. Hot summers and cold winters. Main plants: grasses. Animals include prairie dogs.

🟫 ***Savanna***
Main plants: grasses, but rainfall enough for trees. Typical animals: giraffe.

🟥 *Scrubland (****maquis****)*

🟦 *Mountains*

🟪 *Ice*

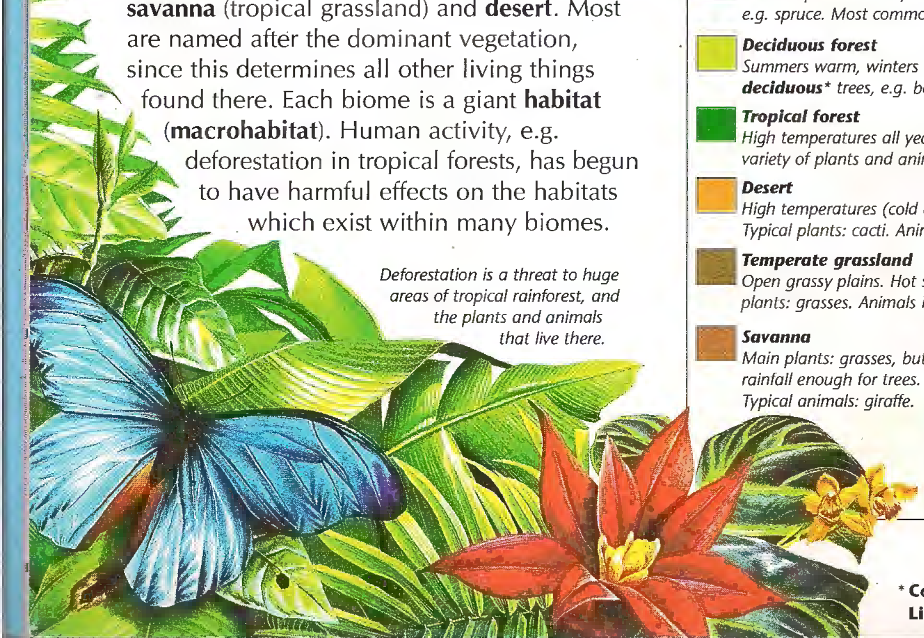

Deforestation is a threat to huge areas of tropical rainforest, and the plants and animals that live there.

---

* **Conifers**, 340; **Deciduous**, 236; **Lichens**, 342 (**Mutualists**).

Habitat

The natural home of a group of living things or a single living thing. Small habitats can be found within large habitats, e.g. a waterhole in the **savanna biome**. Very small specialized habitats are called **microhabitats**, e.g. a rotting acacia tree.

Community

The group of plants and animals found in one **habitat**. They all interact with each other and their environment.

Ecosystem

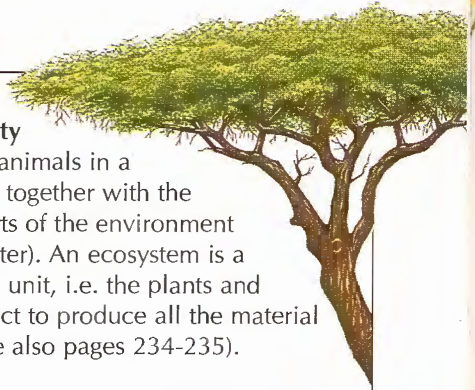

The **community** of plants and animals in a given **habitat**, together with the non-living parts of the environment (e.g. air or water). An ecosystem is a self-contained unit, i.e. the plants and animals interact to produce all the material they need (see also pages 234-235).

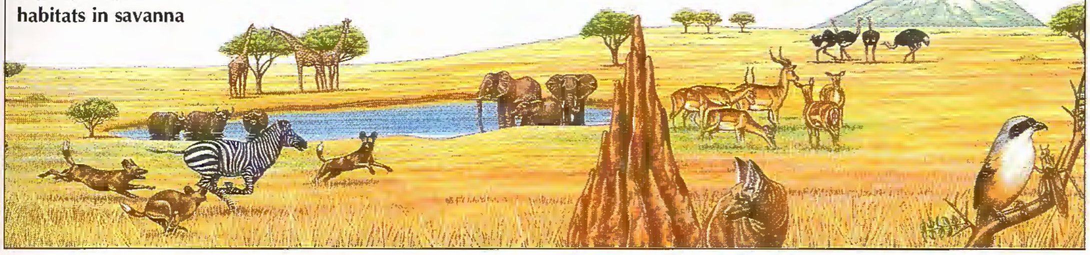

Waterhole and acacia tree habitats in savanna

***Ecosystem** includes environment, e.g. air and water.*

*The **community** includes antelopes and ostriches.*

Ecological succession

A process which occurs whenever a new area of land is colonized, e.g. a forest floor after a fire, a farm field which is left uncultivated or a demolition site which remains unused. Over the years, different types of plants (and the animals which go with them) will succeed each other, until a **climax community** is arrived at. This is a very stable **community**, one which will survive without change as long as the same conditions prevail (e.g. the climate).

Ecological succession in a disused field

***Pioneer community** (first **community**) of grasses, with insects, field mice, etc.*

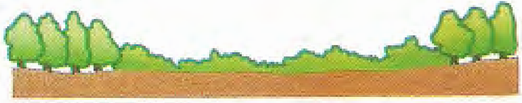

***Successional community** (intermediate **community**) of shrubs and bushes, with rabbits, thrushes, etc.*

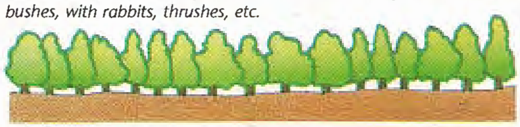

Climax community** of **deciduous trees, e.g. oak and beech, with foxes, badgers, warblers, etc.*

Ecological niche

The place held in an **ecosystem** by a plant or animal, e.g. what it eats and where it lives. **Gause's principle** states that no two species can occupy the same niche at the same time (if they tried, one species would die out or be driven away). For example, in the winter months, both the curlew and the ringed plover can be found living around the estuaries of Britain, and eating small creatures such as worms and snails. However, they actually occupy different niches. Curlews wade in the shallows, probing deep into the mud for food with their long beaks. Ringed plovers, by contrast, pick their food off the surface of the shore (their beaks are too short for probing). Hence both birds can survive in the same general area.

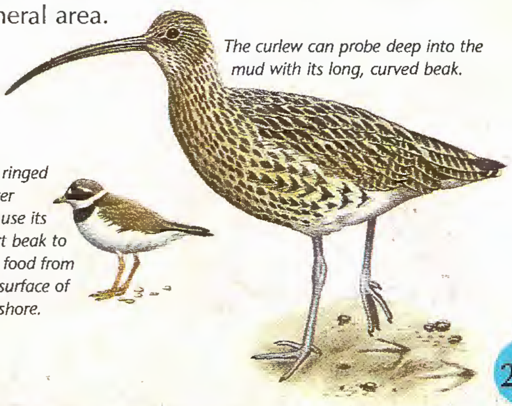

The curlew can probe deep into the mud with its long, curved beak.

The ringed plover can use its short beak to pick food from the surface of the shore.

---

* **Deciduous**, 236.

WITHIN AN ECOSYSTEM

An **ecosystem** consists of a group (**community***) of animals and plants which interact with each other and with their environment to produce a self-contained ecological unit.

Food web

The complex network of **food chains** in an ecosystem. A food chain is a linked series of living things, each of which is the food for the next in line. Plants make their food from non-living matter by **photosynthesis*** (they are **autotrophic**) and are always the first members of a chain. Animals cannot make their own food (they are **heterotrophic**) and so rely on the food-making activities of plants.

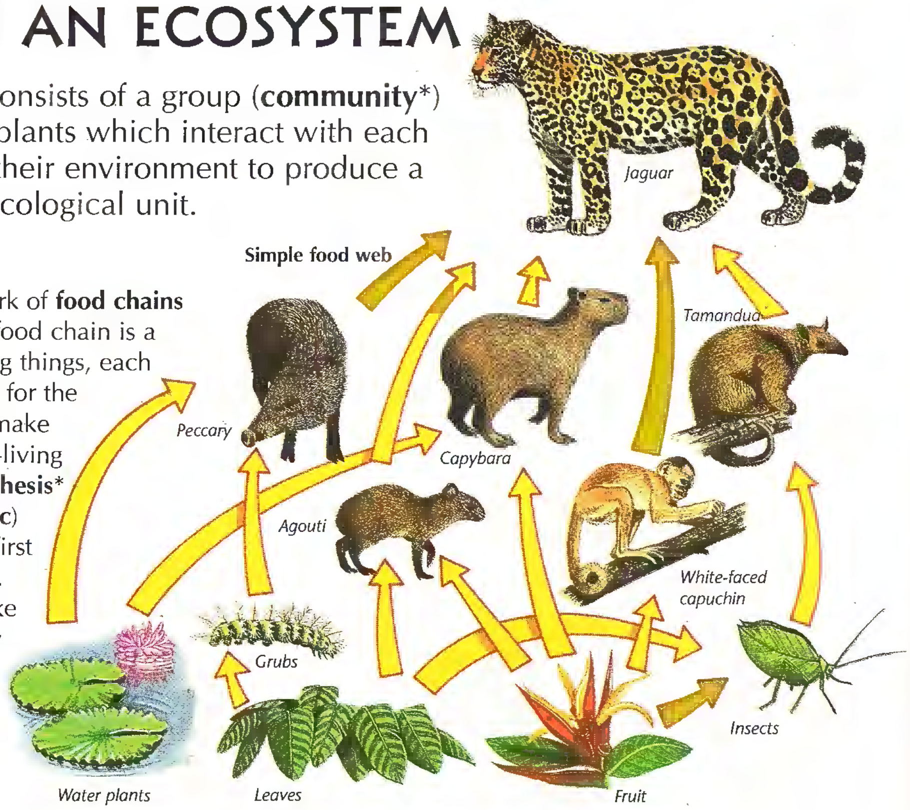

Simple food web

Jaguar
Tamandua
Peccary
Capybara
Agouti
White-faced capuchin
Grubs
Insects
Water plants
Leaves
Fruit

Carbon cycle

The constant circulation of the element carbon through living things and the atmosphere.

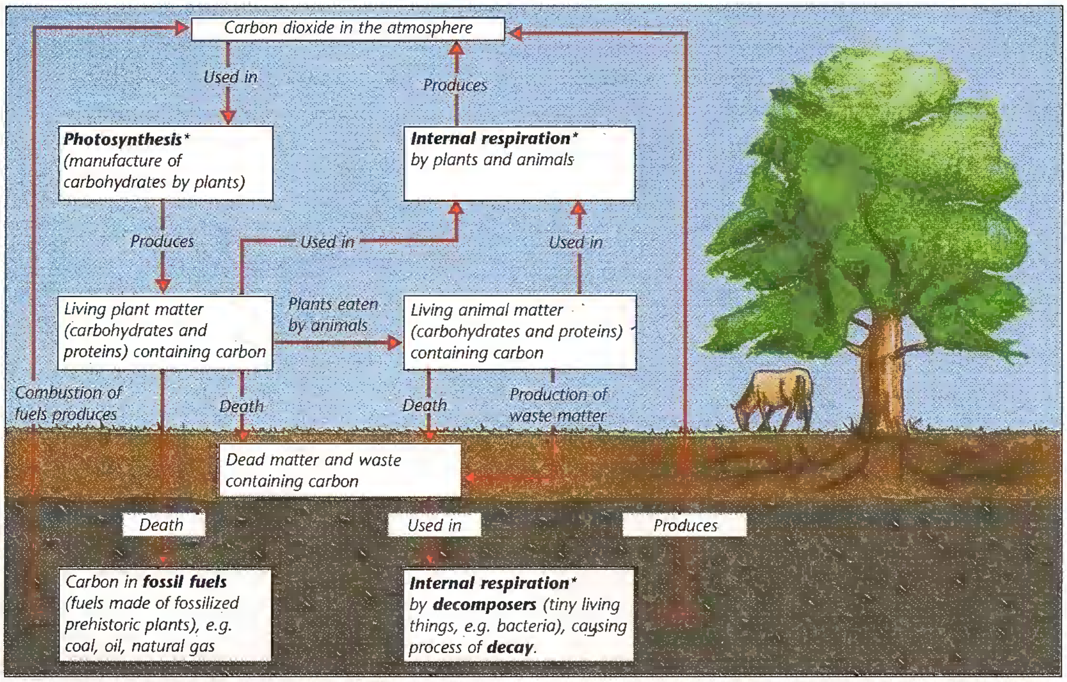

Carbon dioxide in the atmosphere
Used in
Produces
Photosynthesis* (manufacture of carbohydrates by plants)
Internal respiration* by plants and animals
Produces
Used in
Used in
Living plant matter (carbohydrates and proteins) containing carbon
Plants eaten by animals
Living animal matter (carbohydrates and proteins) containing carbon
Combustion of fuels produces
Death
Death
Production of waste matter
Dead matter and waste containing carbon
Death
Used in
Produces
Carbon in **fossil fuels** (fuels made of fossilized prehistoric plants), e.g. coal, oil, natural gas
Internal respiration* by **decomposers** (tiny living things, e.g. bacteria), causing process of **decay**.

* **Community**, 233; **Internal respiration**, 334; **Photosynthesis**, 254.

Trophic level or energy level

The level at which living things are positioned within a **food chain** (see **food web**). At each successive level, a great deal of the energy-giving food matter is lost. For example, a cow will break down well over half of the grass it eats (to provide energy). Hence only a small part of the original energy-giving material can be obtained from eating the cow (the part it used to build its own new tissue). This loss of energy means that the higher the trophic level, the fewer the number of animals, since they must eat progressively larger amounts of food to obtain enough energy. This principle is called the **pyramid of numbers**.

Pyramid of numbers

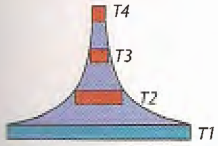

Number of individuals at each ***trophic level***

Pyramid of biomass

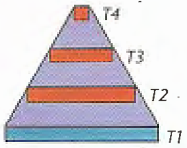

Total mass of individuals at each level (decrease is less extreme than left, since animals at higher levels tend to be larger).

Generalized food chain, showing trophic levels

***Producers** – green plants, which make their own food.* ***Trophic level T1**.*

***Primary consumers** or **first order consumers** – **herbivores** (plant-eating animals), e.g. rabbits. Energy-giving material obtained directly from **producers**.* ***Trophic level T2**.*

***Secondary consumers** or **second order consumers** – **carnivores** (flesh-eating animals), e.g foxes and owls, when they eat herbivores. Energy-giving material obtained from bodies of **primary consumers**. **Trophic level T3**.*

***Tertiary consumers** or **third order consumers** – carnivores, e.g. foxes and owls, when they eat smaller carnivores. Energy-giving material is obtained by most indirect method – from bodies of secondary consumers, i.e. animals which ate animals which ate **producers**. **Trophic level T4**.*

Notes:

*1. **Omnivores**, e.g. humans, eat plant and animal matter. They are thus placed on trophic level T2 at some times and on T3 (or T4) at others.*

2. Many carnivores, e.g. foxes, will eat both herbivores and smaller carnivores. They are thus on trophic level T3 at some times and on T4 at others.

Nitrogen cycle

The constant circulation of the element nitrogen through living things, the soil and the atmosphere.

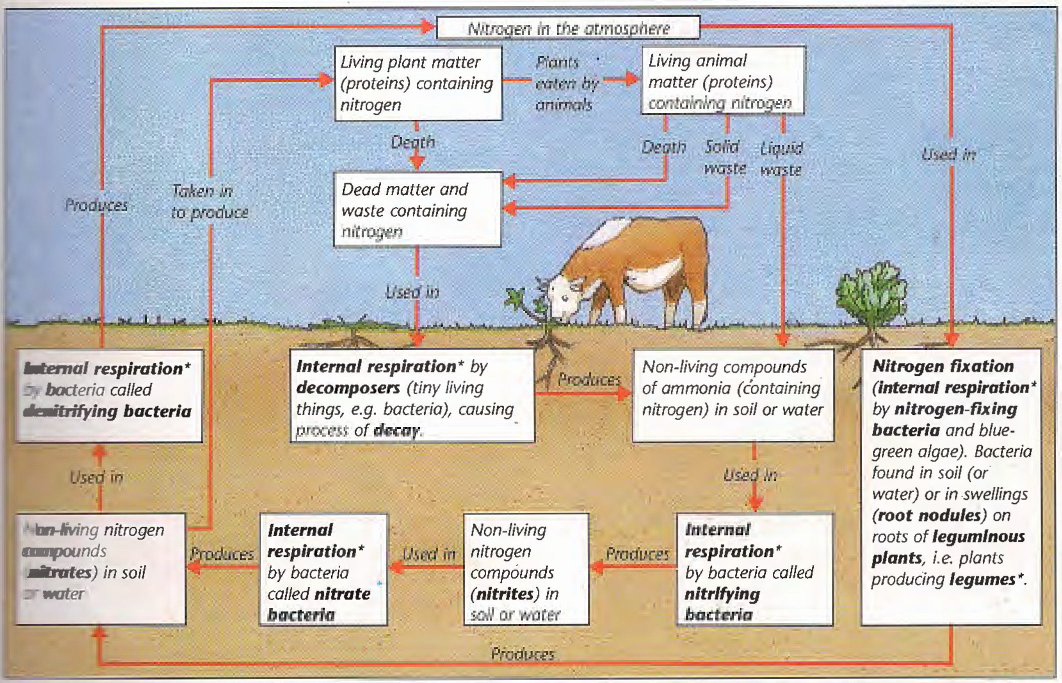
Nitrogen in the atmosphere
Living plant matter (proteins) containing nitrogen
Plants eaten by animals
Living animal matter (proteins) containing nitrogen
Death
Death
Solid waste
Liquid waste
Used in
Produces
Taken in to produce
Dead matter and waste containing nitrogen
Used in
Internal respiration* by bacteria called denitrifying bacteria
Internal respiration* by decomposers (tiny living things, e.g. bacteria), causing process of decay.
Produces
Non-living compounds of ammonia (containing nitrogen) in soil or water
Nitrogen fixation (internal respiration* by nitrogen-fixing bacteria and blue-green algae). Bacteria found in soil (or water) or in swellings (root nodules) on roots of leguminous plants, i.e. plants producing legumes*.
Used in
Used in
Non-living nitrogen compounds (nitrates) in soil or water
Produces
Internal respiration* by bacteria called nitrate bacteria
Used in
Non-living nitrogen compounds (nitrites) in soil or water
Produces
Internal respiration* by bacteria called nitrifying bacteria
Produces

---

*Internal respiration, 334; **Legume**, 262.

LIFE AND LIFE CYCLES

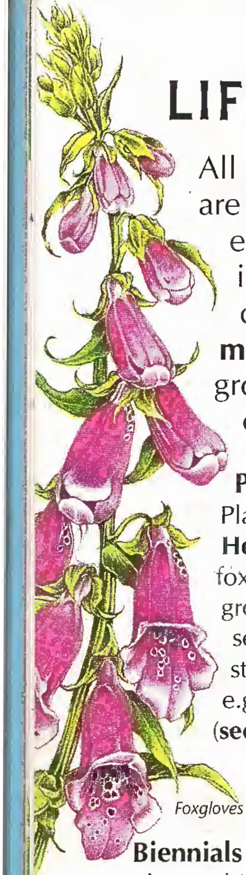

All living things show the same basic **characteristics of life**. These are respiration, feeding, growth, sensitivity (irritability), movement, excretion and reproduction. The **life cycle** of a plant or animal is the progression from its formation to its death, with all the changes this entails. (In some cases, these are drastic – see **metamorphosis**, page 277.) Below are some terms used to group plants and animals together according to their life cycle, or to describe characteristics of certain life cycles.

Perennials

Plants which live for many years. **Herbaceous perennials**, e.g. foxgloves, lose all the parts above ground at the end of each growing season, and grow new shoots at the start of the next. **Woody perennials**, e.g. trees, produce new growth (**secondary tissue***) each year from permanent stems.

Foxgloves are ***perennials****.*

Biennials

Plants which live for two years, e.g. carrots. In the first year, they grow and store up food. In the second, they produce flowers and seeds, and then die.

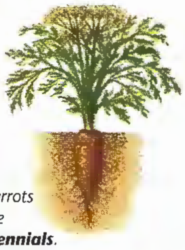

Carrots are ***biennials****.*

Annuals

Plants which live for one year, e.g. lobelias. In this time they grow from seed, produce flowers and seeds, and then die.

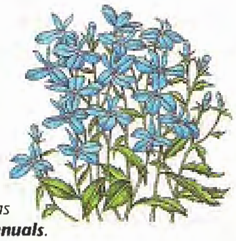

Lobelias are ***annuals****.*

Herbaceous

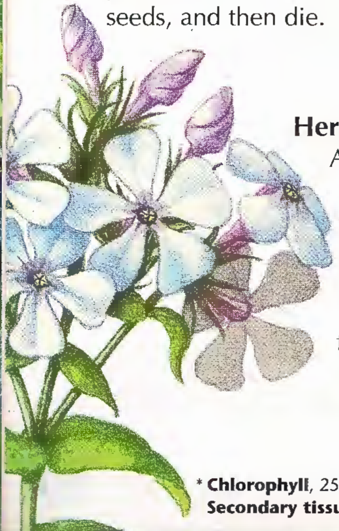

A term describing plants, e.g. phlox, which do not develop **secondary tissue*** above the ground, i.e. they are "like a herb", as distinct from shrubs and trees (**woody perennials**).

Phlox is a ***herbaceous*** *plant.*

Deciduous

A term describing **perennials** whose leaves lose their **chlorophyll*** and fall off at the end of each growing season, e.g. horse chestnuts.

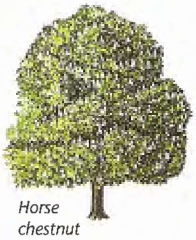

Horse chestnut

Evergreen

A term describing **perennials** which do not shed their leaves at the end of a growing season, e.g. firs.

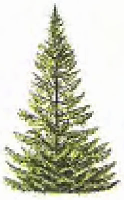

Grand fir

Ephemeral

Living for a very short time. Ephemeral plants are found in places which are hot and dry for most of the year (or for many years). The right growing conditions do not exist for long, so they must grow and produce seeds in a very short time. The only truly ephemeral animals are mayflies. Their adult life span is between a few minutes and one day.

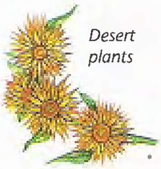

Desert plants

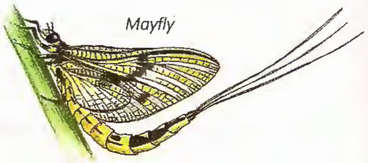

Mayfly

Anadromous

A term describing fish which live in the sea but swim upriver to breed, e.g. salmon. This is a form of **migration**, and the opposite is **catadromous** (going from river to sea).

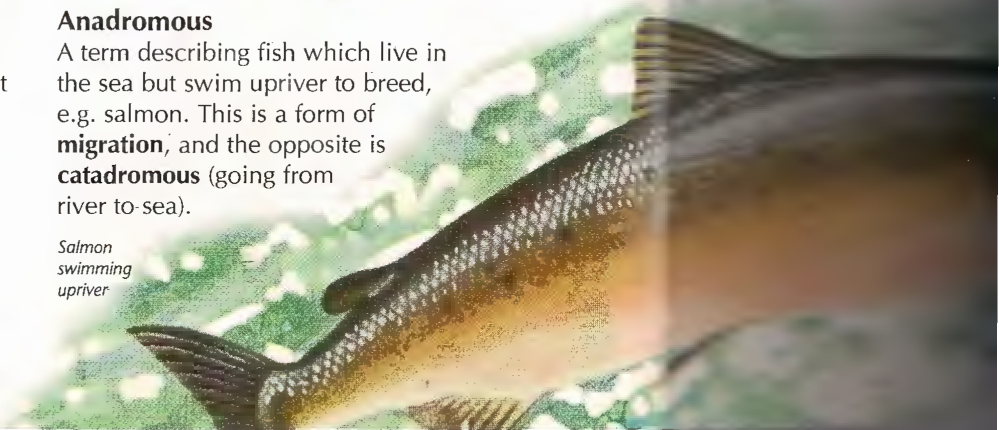

Salmon swimming upriver

---

* **Chlorophyll**, 255 (**Pigments**); **Secondary tissue**, 246.

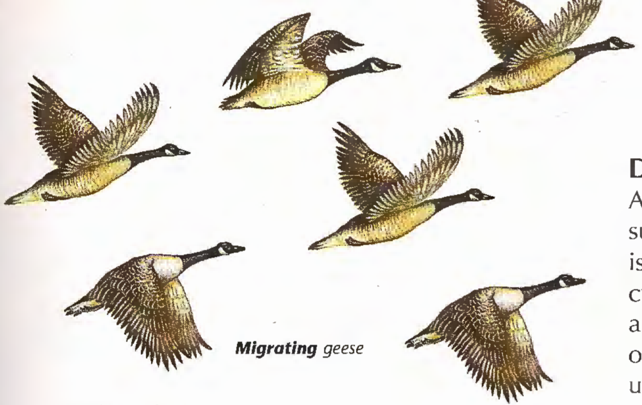

Migrating *geese*

Migration

Traveling seasonally from one region to another. This normally involves leaving an area in winter to find food elsewhere, and returning in the spring to breed. Migration is part of the life cycle of many animals, especially birds.

Dormancy

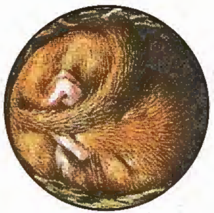

Dormouse in ***hibernation***

A period, or periods, of suspended activity which is a natural part of the life cycle of many plants and animals. Dormancy in plants occurs when conditions are unfavorable for growth (normally in winter). In animals, dormancy usually occurs because of food scarcity, and is either called **hibernation** or **aestivation**. Hibernation is dormancy in the winter (typical of many animals, e.g. some **mammals***), and aestivation is dormancy in drought conditions (occurs mainly in insects).

Life styles

The world has a vast diversity of living things, each one with its own style of life. This situation is a result of **genetic variation**. The living things can be grouped together according to shared characteristics, either by formal classification, which is based mainly on their inferred ancestry (see charts, pages 340-341) or by more informal groupings, based on general characteristics (see list, page 342).

Genetic variation

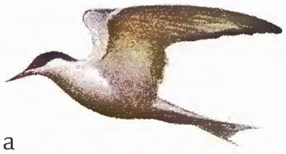

Birds' wings are shaped for flight.

The span of different forms of life, each of which is a variation and specifically designed to survive its environment. For example, fish and birds have streamlined shapes ideal for swimming and flying.

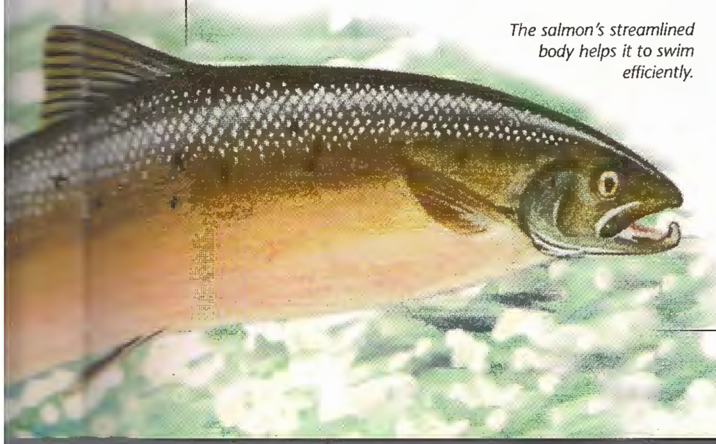

The salmon's streamlined body helps it to swim efficiently.

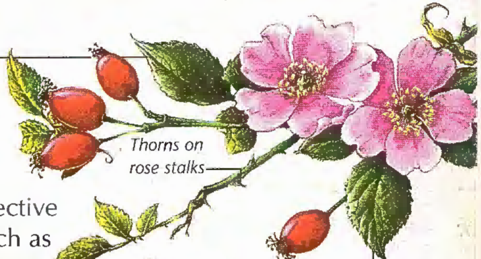

Thorns on rose stalks

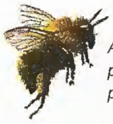

A bee's sting protects it from predators.

Many living things also possess protective measures such as thorns or poison stings. These protective designs become established in successive generations because those creatures with them are the most likely to survive long enough to breed (and perpetuate the designs). This corresponds with Darwin's theory of **natural selection**, (also called **Darwinism**), first expounded in the mid-nineteenth century.

Mimicry

A special type of genetic design, in which a plant or animal (the **mimic**) has a resemblance to another plant or animal (the **model**). This is used especially for protection (e.g. many unprotected insects have the coloring of those which sting), but also for other reasons (bee orchids are mimics for reproduction purposes – see page 259).

Model

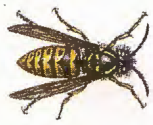

Wasp (protected with stinger)

Mimic

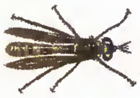

Hoverfly (unprotected)

---

* **Mammals**, 341.

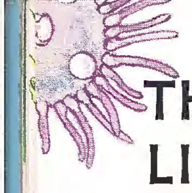

THE STRUCTURE OF LIVING THINGS

A living thing capable of a separate existence is called an **organism**. All organisms are made up of **cells** – the basic units of life, which carry out all the vital chemical processes. The simplest organisms have just one cell (they are **unicellular** or **acellular**), but very complex ones, e.g. humans, have many billions. They are **multicellular** and their cells are of many different types, each type specially designed for its own particular job. Groups of cells of the same type (together with non-living material) make up the different **tissues** of the organism, e.g. muscle tissue. Several different types of tissue together form an **organ**, e.g. a stomach, and a number of organs together form a **system**, e.g. a digestive system.

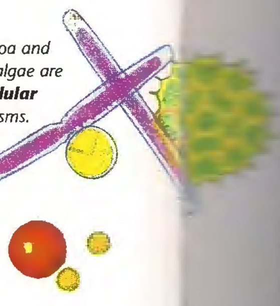

Protozoa and some algae are ***unicellular*** *organisms.*

The parts of a cell

All cells are made up of the same basic parts: the **cell membrane**, **cytoplasm** and **nucleus**. Each of these parts has a specific role to play.

Cell membrane

Also called the **plasma membrane** or **plasmalemma**. The outer surface of a cell. It is **semipermeable***, i.e. selective about which substances it allows through.

Cytoplasm

The material where all the chemical reactions vital to life occur (see **organelles**). It generally has a gel-like outer layer and a more liquid inner one (see **ectoplasm** and **endoplasm** – pictures, page 268).

Animal cell (cut away)

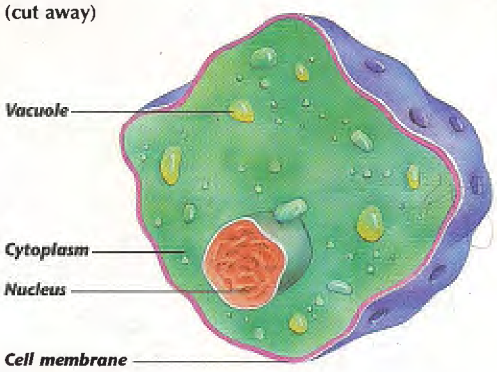

Plant cell (cut away)

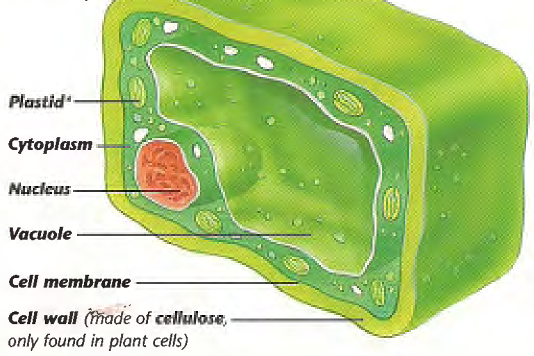

*Plastid**
Cytoplasm
Nucleus
Vacuole
Cell membrane
Cell wall (made of *cellulose*, only found in plant cells)

Nucleus (pl. nuclei)

The cell's control center. Its double-layered outer surface (**nuclear membrane**) encloses a gel-like fluid (**nucleoplasm** or **karyolymph**), which contains one or more **nucleoli*** and the genetic material **DNA***. This is held in **chromosomes*** – bodies which form a thread-like mass called **chromatin** when the cell is not dividing.

Vacuoles

Fluid-filled sacs in the **cytoplasm**. They are small and temporary in animal cells, and either transfer substances (see **Golgi complex** or contain fluid brought in (see **pinocytosis**, page 329). Most plant cells have one large, permanent vacuole, filled with **cell sap** (dissolved minerals and sugars).

---

* **Chromosomes**, 324; **DNA**, 324 (**Nucleic acids**); **Nucleoli**, **Plastids**, 240; **Semipermeable**, 329 (**Diffusion**).

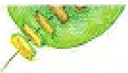

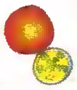

Organelles

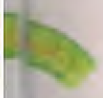

The **organelles** are tiny bodies in the **cytoplasm**. Each type (listed below and on page 240) has a vital role to play in the chemical reactions within the cell.

Lysosomes

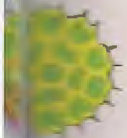

Round sacs containing powerful **enzymes***. They take in foreign bodies, e.g. bacteria, to be destroyed by the enzymes. Their outer skins do not usually let the enzymes out into the cell (to break down its contents), but if the cell becomes damaged, the skins disappear and the cell digests itself.

Ribosomes

Tiny round particles (most are attached to the **endoplasmic reticulum**). They are involved in building up proteins from amino acids (see page 330). Coded information (held by the **DNA** in the **nucleus**) is sent to the ribosomes in strands of a substance called **messenger RNA** (**mRNA**). These pass on the codes so that the ribosomes join the amino acids in the correct order to produce the right proteins. **RNA*** is present in at least two other forms in the cells. The ribosomes are made of **ribosomal RNA** (see **nucleoli***), and molecules of **transfer RNA** (**tRNA**) carry the amino acids to the ribosomes.

Endoplasmic reticulum or ER

A complex system of flat sacs, joining up with the **nuclear membrane** (see **nucleus**). It provides a large surface area for **enzyme*** reactions. **Rough ER** has **ribosomes** on the surface, where amino acids are combined to make proteins (see page 330). ER with no ribosomes is **smooth ER**. Smooth ER bears enzymes for many other cell processes.

Golgi complex

Also called a **Golgi apparatus**, **Golgi body** or **dictyosome**. A special set of membranous sacs, which collects, modifies and distributes the substances made by the ER (e.g. proteins). The substances fill the sacs, which gradually swell up at their outside edges until pieces "pinch off". These pieces, called **vesicles**, then travel out of the cell via the **cytoplasm** and **cell membrane**.

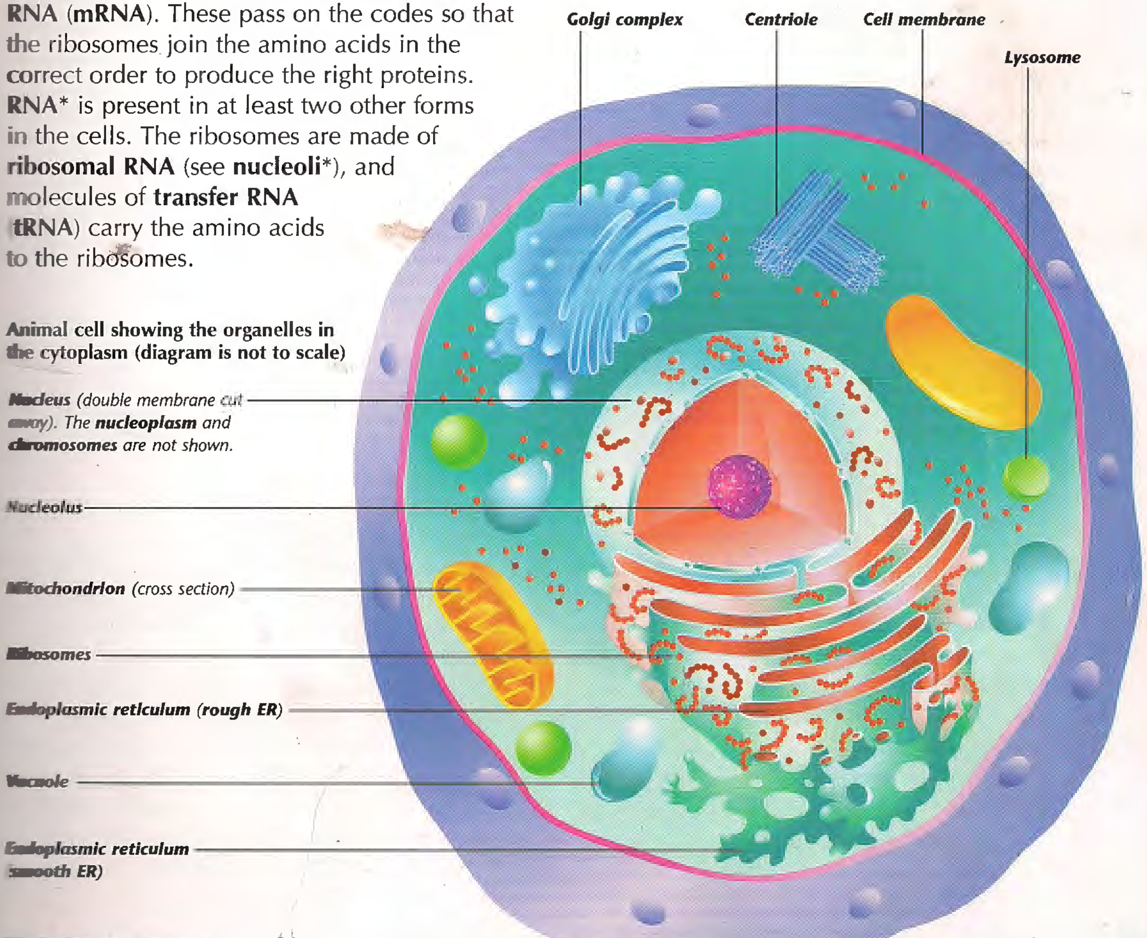
Golgi complex
Centriole
Cell membrane
Lysosome
Nucleus (double membrane cut away). The nucleoplasm and chromosomes are not shown.
Nucleolus
Mitochondrion (cross section)
Ribosomes
Endoplasmic reticulum (rough ER)
Vacuole
Endoplasmic reticulum (smooth ER)

Animal cell showing the organelles in the cytoplasm (diagram is not to scale)

Enzymes, 333; **Nucleoli**, 240; **RNA**, 324 (**Nucleic acids**).

Organelles (continued)

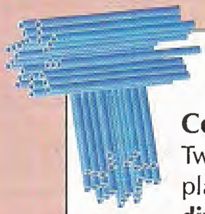

Centriole

Centrioles

Two bodies in animal and primitive plant cells which are vital to **cell division** (see right). In animal cells, they lie just outside the **nucleus***. Each lies in a dense area of **cytoplasm*** (**centrosome**) and is made up of two tiny cylinders, forming a +-shape or T-shape. Each cylinder is made up of nine sets of three tiny tubes (**microtubules**).

Nucleoli (sing. nucleolus)

One or more small, round bodies in the **nucleus***. They produce the component parts of the **ribosomes*** (made of **ribosomal RNA**), which are then transported out of the nucleus and assembled in the **cytoplasm***.

Nucleolus

Mitochondria (sing. mitochondrion)

Rod-shaped bodies with a double layer of outer skin. The inner layer forms a series of folds (**cristae**, sing. **crista**), providing a large surface area for the vital chemical reactions which go on inside the mitochondria (called the cell's "powerhouses"). They are the places where simple substances taken into the cell are broken down to provide energy. For more about this, see **aerobic respiration**, page 334.

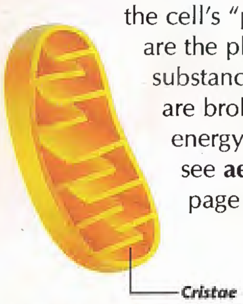

Mitochondrion

Plastids

Tiny bodies in plant cell **cytoplasm***. Some (**leucoplasts**) store starch, oil or proteins. Others – **chloroplasts*** – contain **chlorophyll*** (used in making food).

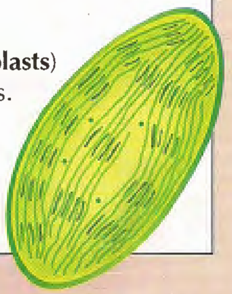

Plastid (chloroplast*)

Cell division

Cell division is the splitting up of one cell (the **parent cell**) into two identical **daughter cells**. There are two types of cell division, both involving the division of the **nucleus*** (**karyokinesis**) followed by the division of the **cytoplasm*** (**cytokinesis**). The first type of cell division (**mitosis**) is described on these two pages. It produces new cells for growth and also to replace the millions of cells which die each day (from damage, disease or simply because they are "worn out"). It is also the means of **asexual reproduction*** in many single-celled organisms. The second, special type of cell division produces the **gametes*** (sex cells) which will come together to form a new living thing. For more about this, see pages 322-323.

Mitosis

The division of the **nucleus*** when a plant or animal cell divides for growth or repair. It ensures that the two new nuclei (**daughter nuclei**) are each given the same number of **chromosomes*** (the bodies which carry the "coded" hereditary information). Each receives the same number of chromosomes as were in the original nucleus, called the **diploid number**. Every living thing has its own characteristic diploid number, i.e. all its cells (with the exception of the **gametes***) contain the same, specific number of chromosomes, grouped in identical pairs called **homologous chromosomes**. Humans have 46 chromosomes, in 23 pairs. Although mitosis is a continuous process, it can be divided for convenience sake into four phases. Before mitosis, however, there is always an **interphase**.

Interphase

The periods between cell divisions. Interphases are very active periods, during which the cells are not only carrying out all the processes needed for life, but are also preparing material to make "copies" of all their components (so both new cells formed after division will have all they need). Just before **mitosis** begins, the **chromatin*** threads in the **nucleus*** also duplicate, so that, after coiling up, each **chromosome*** will consist of two **chromatids** (see **prophase**). The **centriole** duplicates itself during interphase.

* **Asexual reproduction**, 321; **Chlorophyll**, 255 (**Pigments**); **Chloroplasts**, 254; **Chromatin**, 238 (**Nucleus**); **Chromosomes**, 324; **Cytoplasm**, 238; **Gametes**, 320; **Nucleus**, 238; **Ribosomes**, 239.

Phases of mitosis

1. Prophase

The **nuclear membrane*** begins to break down and the threads of **chromatin*** in the **nucleus*** coil up to form **chromosomes***. Each has already duplicated to form two identical, long coils (**chromatids**), joined by a sphere (**centromere**). The two **centrioles** move to opposite poles (ends) of the cell, as **spindle microtubules** form between them.

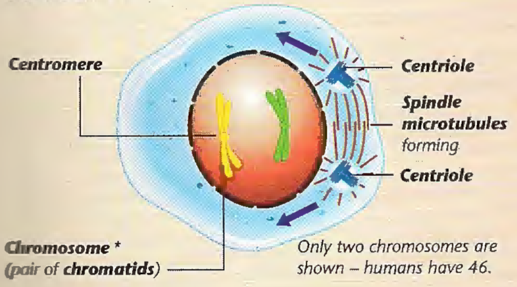
Centromere
Centriole
Spindle microtubules *forming*
Centriole
Chromosome* *(pair of* ***chromatids)***

Only two chromosomes are shown – humans have 46.

2. Metaphase

The **nuclear membrane*** disappears and the **spindle microtubules** surround the **chromosomes*** (paired **chromatids**). The chromosomes move towards its equator and become attached by their **centromeres** to the **spindle microtubules.**

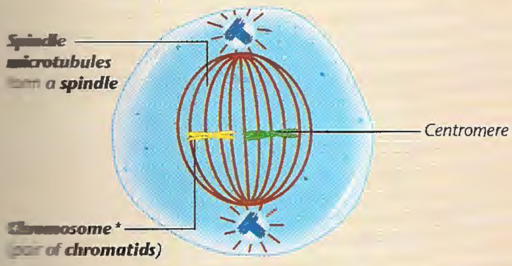
Spindle microtubules *form a* ***spindle***
Centromere
Chromosome* *(pair of* ***chromatids)***

3. Anaphase

The **centromeres** split and the two **chromatids** from each pair (now called **daughter chromosomes**) move to opposite poles of the **spindle**, seemingly "dragged" there by the contracting **spindle microtubules.**

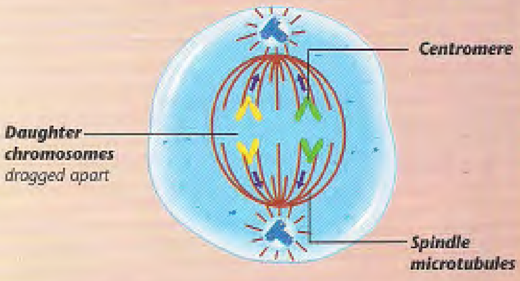
Centromere
Daughter chromosomes *dragged apart*
Spindle microtubules

4. Telophase

The **spindle microtubules** disappear and a new **nuclear membrane*** forms around each group of **daughter chromosomes.** This creates two new **nuclei*** (**daughter nuclei**), inside which the chromosomes uncoil and once again form a thread-like mass (**chromatin***).

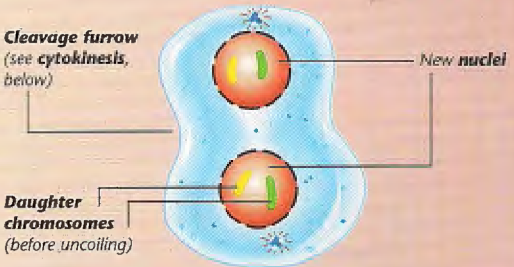
Cleavage furrow *(see* ***cytokinesis,*** *below)*
New ***nuclei***
Daughter chromosomes *(before uncoiling)*

Cytokinesis

The division of the **cytoplasm*** of a cell, which forms two new cells around the new **nuclei*** created by **mitosis** (or **meiosis***). In animal cells, a **cleavage furrow** forms around the cell's equator and then constricts as a ring until it cuts completely through the cell. In plant cells, a dividing line called the **cell plate** forms down the center of the cell, and a new **cell wall*** is built up along each side of it.

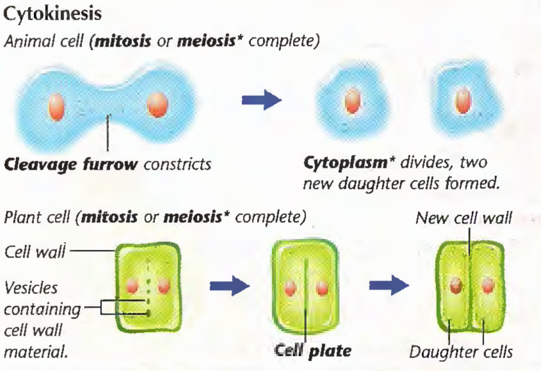
Cytokinesis

*Animal cell (**mitosis** or **meiosis*** complete)*

Cleavage furrow *constricts*
Cytoplasm* *divides, two new daughter cells formed.*

*Plant cell (**mitosis** or **meiosis*** complete)*

Cell wall
Vesicles containing cell wall material.
Cell plate
New cell wall
Daughter cells

---

* **Cell wall**, 238; **Chromatin**, 238 (**Nucleus**); **Chromosomes**, 324; **Cytoplasm**, 238; **Meiosis**, 322; **Nuclear membrane**, 238 (**Nucleus**).

VASCULAR PLANTS

With the exception of simple plants such as algae, fungi, mosses and liverworts (see classification chart, page 340), all plants are **vascular plants**. That is, they all have a complex system of special fluid-carrying tissue called **vascular tissue**. For more about how the fluids travel within the vascular tissue, see pages 252-253.

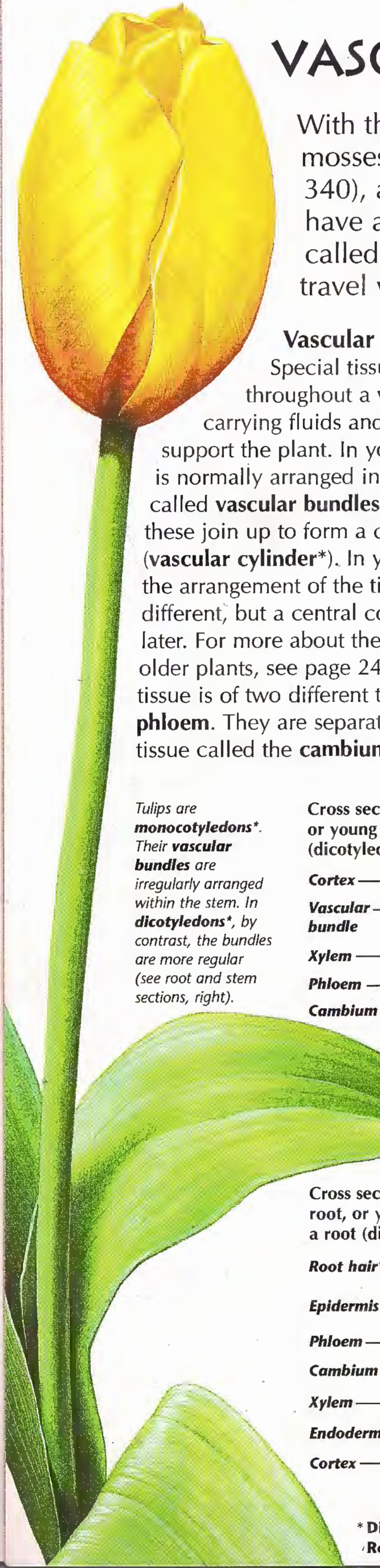

Vascular tissue

Special tissue which runs throughout a **vascular plant**, carrying fluids and helping to support the plant. In young stems, it is normally arranged in separate units called **vascular bundles**; in older stems, these join up to form a central core (**vascular cylinder***). In young roots, the arrangement of the tissue is slightly different, but a central core is also formed later. For more about the vascular tissue in older plants, see page 246. The vascular tissue is of two different types – **xylem** and **phloem**. They are separated by a layer of tissue called the **cambium**.

Tulips are ***monocotyledons****. Their* ***vascular bundles*** *are irregularly arranged within the stem. In* ***dicotyledons****, by contrast, the bundles are more regular (see root and stem sections, right).*

Cross section of young stem, or young part of a stem (dicotyledon*)

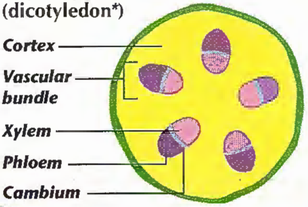
Cortex
Vascular bundle
Xylem
Phloem
Cambium

Cross section of young root, or young part of a root (dicotyledon*)

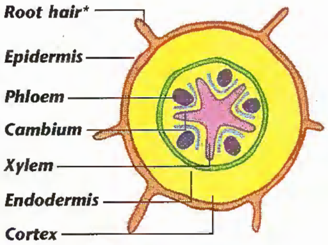
*Root hair**
Epidermis
Phloem
Cambium
Xylem
Endodermis
Cortex

Young stem, or young part of a stem (dicotyledon*)

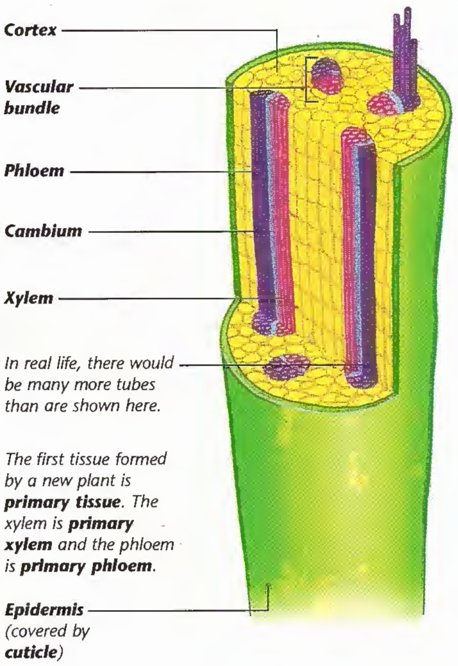
Cortex
Vascular bundle
Phloem
Cambium
Xylem
Epidermis (covered by *cuticle*)

In real life, there would be many more tubes than are shown here.

The first tissue formed by a new plant is ***primary tissue****. The xylem is* ***primary xylem*** *and the phloem is* ***primary phloem****.*

Longitudinal section of young root, or young part of a root (dicotyledon*)

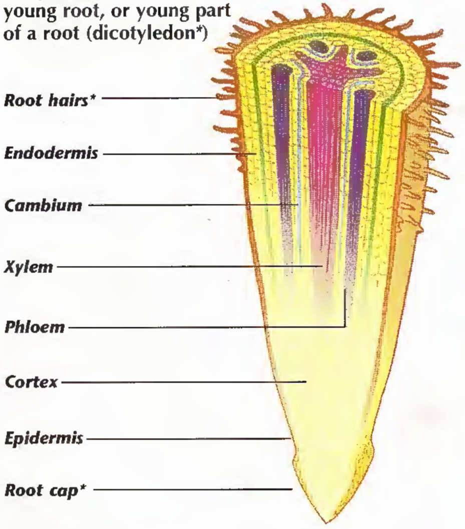
*Root hairs**
Endodermis
Cambium
Xylem
Phloem
Cortex
Epidermis
*Root cap**

---

* **Dicotyledon**, **Monocotyledon**, 261 (**Cotyledon**); **Root cap**, **Root hairs**, 245; **Vascular cylinder**, 246.

Constituents of vascular tissue

Xylem

A tissue which carries water up through a plant. In flowering plants it is made up of **vessels** or **tracheids**, with long, thin cells (**fibers**) providing support between them. Non-flowering plants have only tracheids. In older stems, the vessels become filled in, forming **heartwood***.

Vessels and **tracheids**

Tubes in the **xylem** which carry water. Their walls are strengthened with a hard substance called **lignin**. They occur as columns of cells whose contents have died. Vessels are shorter and wider than tracheids.

Cambium

A layer of narrow, thin-walled cells between the **xylem** on the inside and the **phloem** on the outside. The cells are able to divide, making more xylem and phloem. Such an area of cells is called a **meristem***.

Phloem

A tissue which distributes the food made in the leaves to all parts of the plant. It consists of fluid-carrying **sieve tubes**, each with a **companion cell** beside them, and other cells packed around them for support.

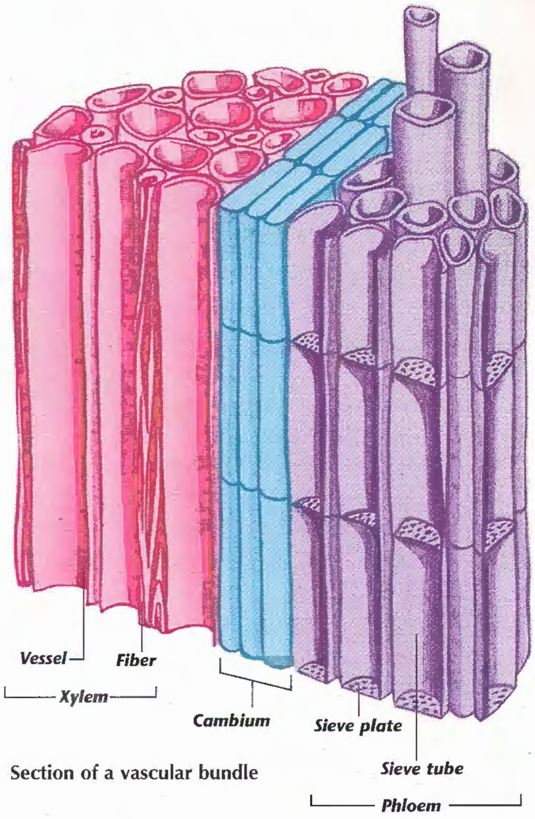

Section of a vascular bundle

Sieve tubes

Cells in long columns in the **phloem**. They are living cells with **cell walls*** and a thin layer of **cytoplasm** but no **nucleus***. The end walls between the cells, called **sieve plates**, have tiny holes in them to allow substances through.

Other tissues in vascular plants

Epidermis

A thin surface layer of tissue around all parts of a plant. In some areas, especially the leaves, it has many tiny holes, called **stomata***. In older stems, the epidermis is replaced by **phellem***. In older roots, it is replaced by **exodermis*** and then by phellem.

Cortex

A layer of tissue inside the **epidermis** of stems and roots. It consists mainly of **parenchyma**, a type of tissue with large cells and many air spaces. In some plants there is also some **collenchyma**, a type of supporting tissue with long, thick-walled cells. The cortex tends to get compressed and replaced by other tissues as a plant gets older.

Endodermis

The innermost layer of root **cortex**. Fluids which have seeped in between the cortex cells, instead of through them, are directed by its special **passage cells** into the central area of **vascular tissue**.

Pith or **medulla**

A central area of tissue found in stems, but not usually in roots. It is generally only called pith once the stem has developed a **vascular cylinder***. It is made up of **parenchyma** (see **cortex**), and is sometimes used to store food.

Cuticle

A thin outer layer of a waxy substance called **cutin** made by the **epidermis** above ground. It prevents too much water from being lost.

---

* **Exodermis**, 245 (**Piliferous layer**); **Heartwood**, 247; **Meristem**, 244; **Nucleus**, 238; **Phellem**, 247; **Stomata**, 249; **Vascular cylinder**, 246.

STEMS AND ROOTS

The **stem** and **roots** of a plant are its main supporting structures, as well as being important in transporting fluids (see pages 242-243 and 252-253). Their various parts are listed here. For more about the development of the stem and roots as plants get older, see pages 246-247.

Meristem

Any area in a plant from which new growth arises. The cells of a meristem are able to divide, producing new cells. A meristem found at the tip of the root (the **growing point**) or the stem (part of a **terminal bud**) is known as an **apical meristem**.

Stem attachments

Shoot

A new stem growing out of a seed or off the main stem of a plant.

Bud

A small outgrowth on a stem. It develops either into a new **shoot** or a flower.

Terminal bud

A **bud** growing at the end of a stem or **shoot**.

Axillary bud

Also called a **lateral bud** or **secondary bud**. A **bud** situated in an **axil** – the angle between a **shoot** or leaf stalk and the stem it is growing from.

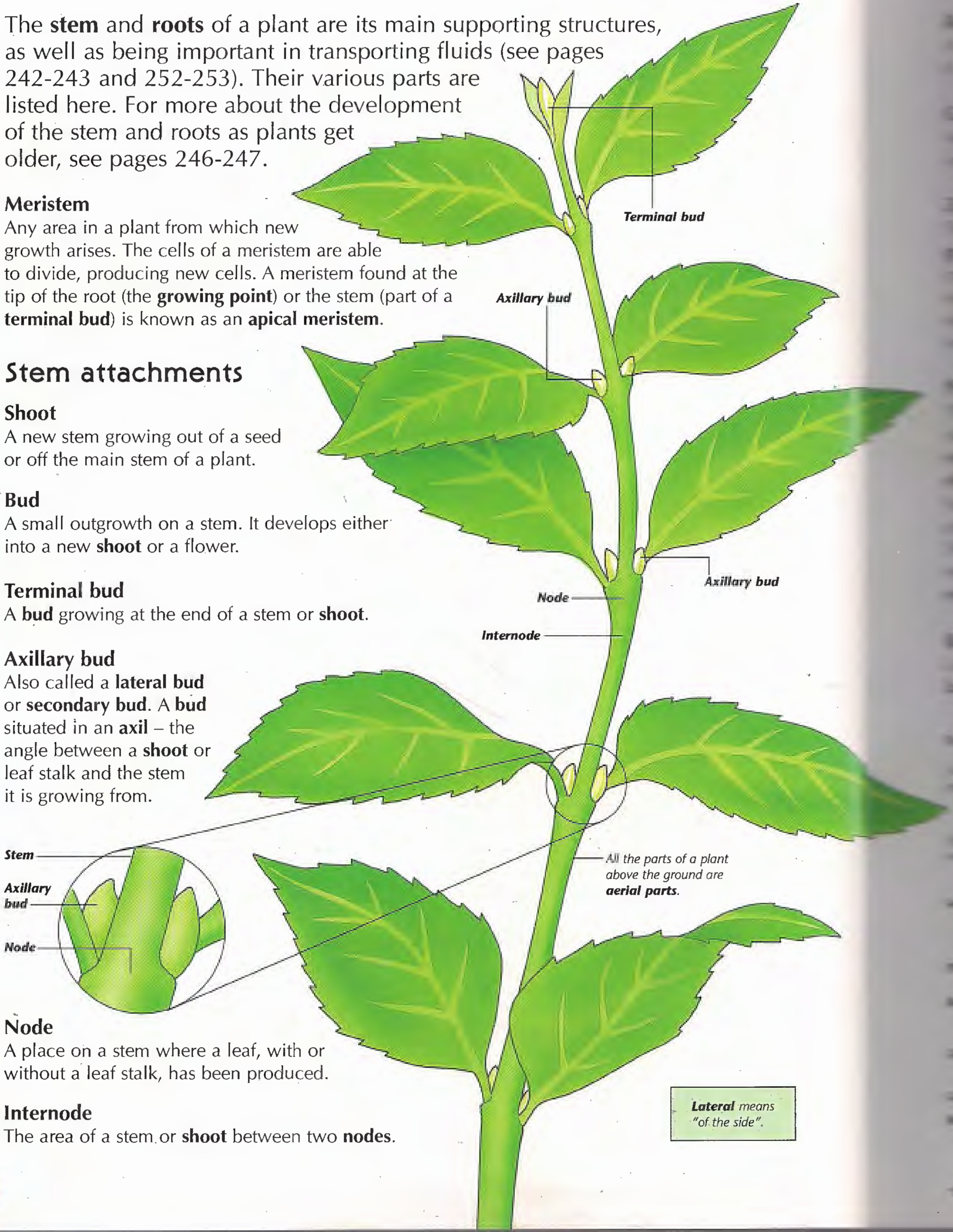

All the parts of a plant above the ground are ***aerial parts****.*

Node

A place on a stem where a leaf, with or without a leaf stalk, has been produced.

Internode

The area of a stem or **shoot** between two **nodes**.

Lateral *means "of the side".*

Parts of a root

Root cap

A layer of cells which protects the root tip as it is pushed down into the ground.

Growing point

An area just behind a root tip where the cells divide to produce new growth.

Zone of elongation

The area of new cells produced by the **growing point**, and located just behind it. The cells stretch lengthwise as they take in water, since their **cell walls*** are not yet hard. This elongation pushes the root tip further down into the soil.

Piliferous layer

The youngest layer of the **epidermis***, or outer skin, of a root. It is the area which produces **root hairs**. It is found just behind the **zone of elongation**. As the walls of the elongating cells harden, the outermost cells become the piliferous layer. The older piliferous layer (higher up the root) is slowly worn away, to be replaced by a layer of hardened cells called the **exodermis** (the outermost layer of the **cortex***).

Root hairs

Long outgrowths from the cells of the **piliferous layer**. They take in water and minerals.

Parts of a root

(See also root section, page 242.)

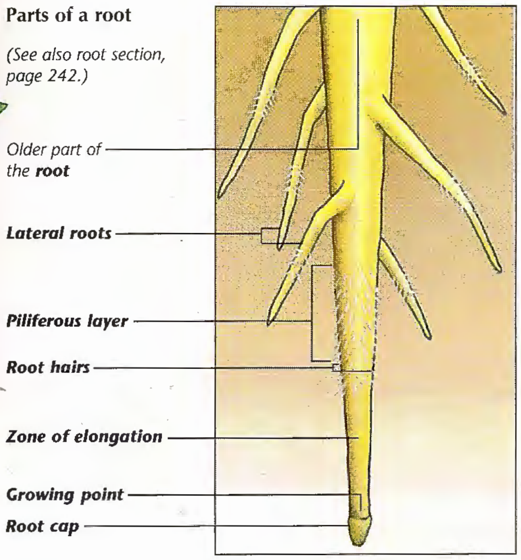
Older part of the root
Lateral roots
Piliferous layer
Root hairs
Zone of elongation
Growing point
Root cap

Types of roots

Tap root

A first root, or **primary root**, which is larger than the small roots, called **lateral roots** or **secondary roots**, which grow out of it. Many vegetables are swollen tap roots.

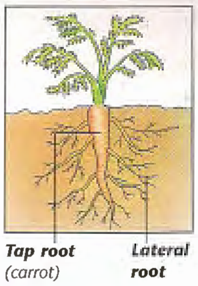
Tap root (carrot)
Lateral root

Fibrous roots

A system made up of a large number of equal-sized roots, all producing smaller **lateral roots**. The first root is not prominent, as it is in a **tap root** system.

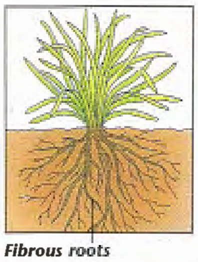
Fibrous roots

Adventitious roots

Roots which grow directly from a stem. Adventitious roots grow out of **bulbs*** (which are special types of stem), or from gardeners' cuttings.

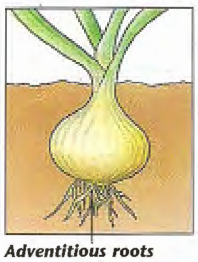
Adventitious roots

Aerial roots

Roots which grow from stems and do not normally grow into the ground. They can be used for climbing, e.g. in an ivy. Many absorb moisture from the air.

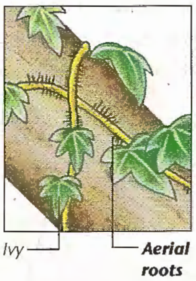
Ivy
Aerial roots

Prop roots

Special types of **aerial root**. They grow out from a stem and then down into the ground, which may be under water. Prop roots support a heavy plant, e.g. a mangrove.

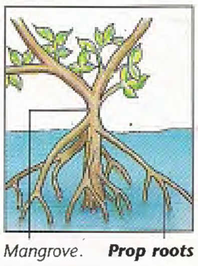
Mangrove
Prop roots

---

* **Bulb**, 263; **Cell wall**, 238; **Cortex**, **Epidermis**, 243.

INSIDE AN OLDER PLANT

A plant which lives for many years, such as a tree, forms **secondary tissue** as it grows. This consists of new layers of tissue to supplement the original tissue, or **primary tissue***. New supportive and fluid-carrying **vascular tissue*** is formed toward the center of the plant and new protective tissue is produced around the outside. The production of the new vascular tissue is called **secondary thickening**, and results in what is known as a **woody plant**.

New central tissue

Vascular cylinder

A vascular cylinder develops as the first step of **secondary thickening** in stems. More **cambium*** forms between the **vascular bundles***, and this then gives rise to more **xylem*** and **phloem***, forming a continuous cylinder.

Secondary thickening

The year-by-year production of more fluid-carrying **vascular tissue*** in plants which live for many years, resulting in a gradual increase in the diameter of the stem and roots. Each year, new layers of **xylem*** (**secondary xylem**) and **phloem*** (**secondary phloem**) are produced by the dividing cells of the **cambium*** between them. This process differs slightly between stems and roots, but the result throughout the plant is an ever-enlarging core of vascular tissue (which slowly "squeezes out" the **pith*** in stems). Most of this core is xylem, now also known as **wood**. The area of phloem does not widen much at all, because the xylem pushing outward wears it away.

Annual rings

The concentric circles which can be seen in a cross section of an older plant. Each ring is one year's new growth of **xylem***, and has two separate areas – **spring wood** and **summer wood**. Soft **spring wood** (or **early wood**) forms rapidly early in the growing season and has widely-spaced cells. Harder **summer wood** (or **late wood**) is produced later on. Its cells are more densely packed.

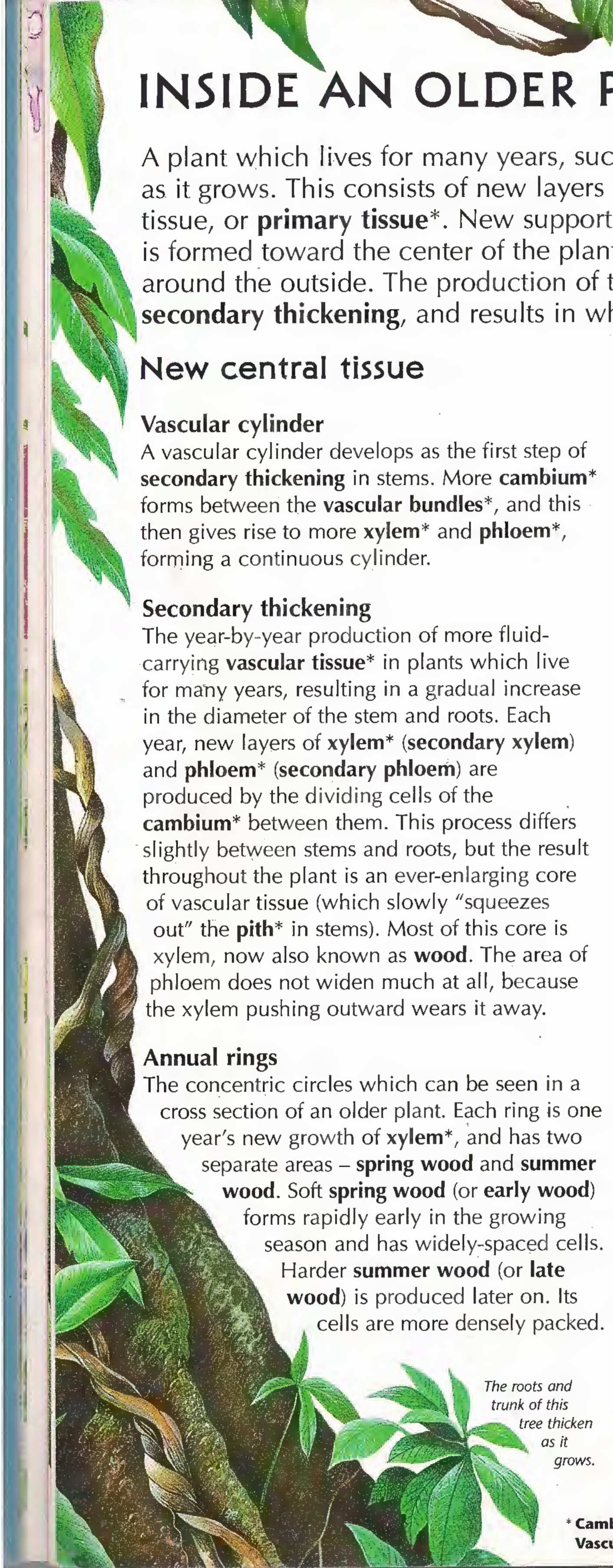

The roots and trunk of this tree thicken as it grows.

Secondary thickening in a stem

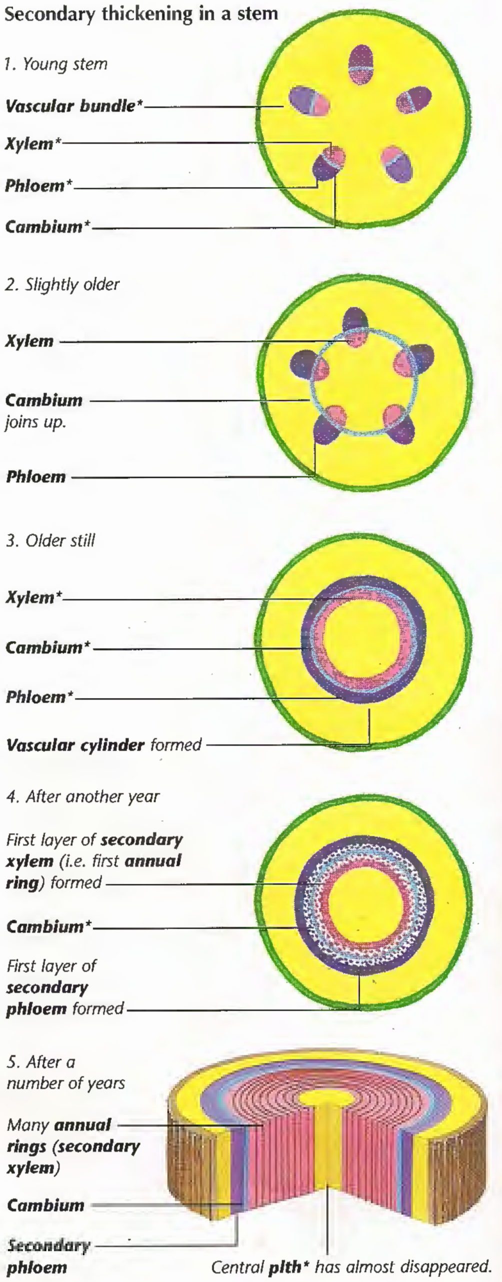
1. Young stem
Vascular bundle*
Xylem*
Phloem*
Cambium*
2. Slightly older
Xylem
Cambium joins up.
Phloem
3. Older still
Xylem*
Cambium*
Phloem*
Vascular cylinder formed
4. After another year
First layer of secondary xylem (i.e. first annual ring) formed
Cambium*
First layer of secondary phloem formed
5. After a number of years
Many annual rings (secondary xylem)
Cambium
Secondary phloem
Central pith* has almost disappeared.

---

* **Cambium**, 243; **Phloem**, 243; **Pith**, 243; **Primary tissue**, 242; **Vascular bundles**, 242 (**Vascular tissue**); **Xylem**, 243.

New outer tissue

As well as new **vascular tissue***, an older plant also forms extra areas of tissue around its outside to help protect it. These are called **phelloderm**, **phellogen** and **phellem** respectively (working from the inside). The three areas together are known as the **periderm**.

Phellogen or cork cambium

A cell layer which arises toward the outside of the stem and roots of older plants. It is a **meristem***, i.e. an area of cells which keep on dividing. It produces two new layers – **phelloderm** and **phellem**.

Phelloderm

A new cell layer produced by **phellogen** on its inside. It supplements the **cortex*** and is sometimes called **secondary cortex**.

Phellem or cork

A new cell layer produced by **phellogen** on its outside. The cells undergo **suberization**, i.e. become impregnated with a waxy substance called **suberin**. This makes the outer layer waterproof. The phellem cells slowly die and replace the previous outer cell layer (**epidermis*** in stems and **exodermis*** in roots). Dead phellem cells are called **bark**.

Tree (many years old)

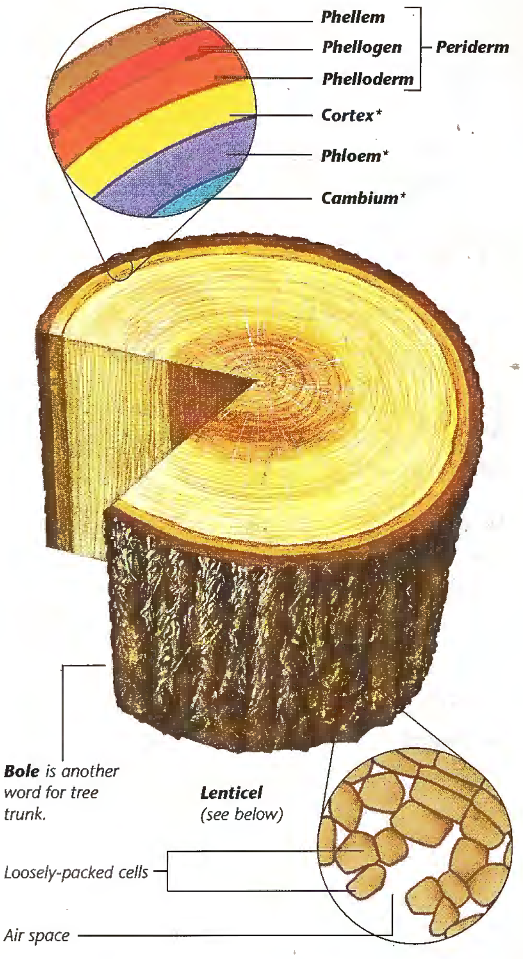

Bole *is another word for tree trunk.*

Lenticel *(see below)*

Bark *stops the tree from drying out and protects it from disease. It cannot grow or stretch, so it splits or peels as the trunk gets wider, and new bark grows underneath.*

Silver birch bark *English oak bark* *Scots pine bark* *Beech bark*

Lenticels

Tiny raised openings in the **phellem** through which an older plant exchanges oxygen and carbon dioxide. Inside them, a channel of loosely-packed cells allows the gases to move across the outer tissues to or from the **cortex***, which also has air spaces.

Types of wood

Heartwood

The oldest, central part of the **xylem*** in an older plant. The **vessels*** are filled in and no longer carry fluids, but they still provide support.

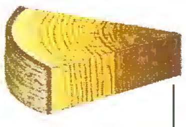

Heartwood

Sapwood

The outer area of **xylem*** in an older plant, whose **vessels*** still carry fluids. Sapwood also supports the tree and holds the tree's food reserves.

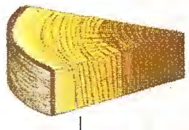

Sapwood

---

* **Cambium**, **Cortex**, **Epidermis**, 243; **Exodermis**, 245 (**Piliferous layer**); **Phloem**, 243; **Meristem**, 244; **Vascular tissue**, 242; **Vessels**, 243; **Xylem**, 243.

LEAVES

The **leaves** of a plant, collectively known as its **foliage**, are specially adapted to manufacture food. They do this by a process called **photosynthesis**. For more about this, see pages 254-255. There are many different shapes and sizes of leaves, but only two different types. **Simple leaves** consist of a single leaf blade, or **lamina**, and **compound leaves** are made up of small leaf blades called **leaflets**, all growing from the same leaf stalk. You can find out more about some of the different leaf shapes on page 250.

Simple leaf *(holly)*

Compound leaf
(horse chestnut)

Inside a leaf

Veins

Long strips of **vascular tissue*** inside a leaf (see picture, right), supplying it with water and minerals and removing the food made inside it. Some leaves have long, parallel veins, e.g. those of grasses, but most have a central vein inside a **midrib** (an extension of the leaf stalk), with many smaller branching veins.

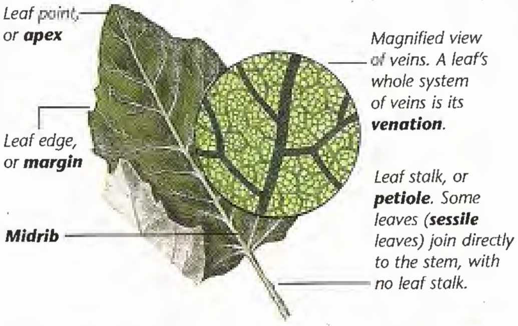
Leaf point, or **apex**
Leaf edge, or **margin**
Midrib

*Magnified view of veins. A leaf's whole system of veins is its **venation**.*

*Leaf stalk, or **petiole**. Some leaves (**sessile** leaves) join directly to the stem, with no leaf stalk.*

Spongy layer

A layer of irregular-shaped **spongy cells** and air spaces where gases circulate. The **spongy** and **palisade layers** together are the **mesophyll**.

Palisade layer

A cell layer just below the upper surface of a leaf. It is made up of regular, oblong-shaped **palisade cells**. These contain many **chloroplasts***.

Palisade cell

Leaf (cross section)

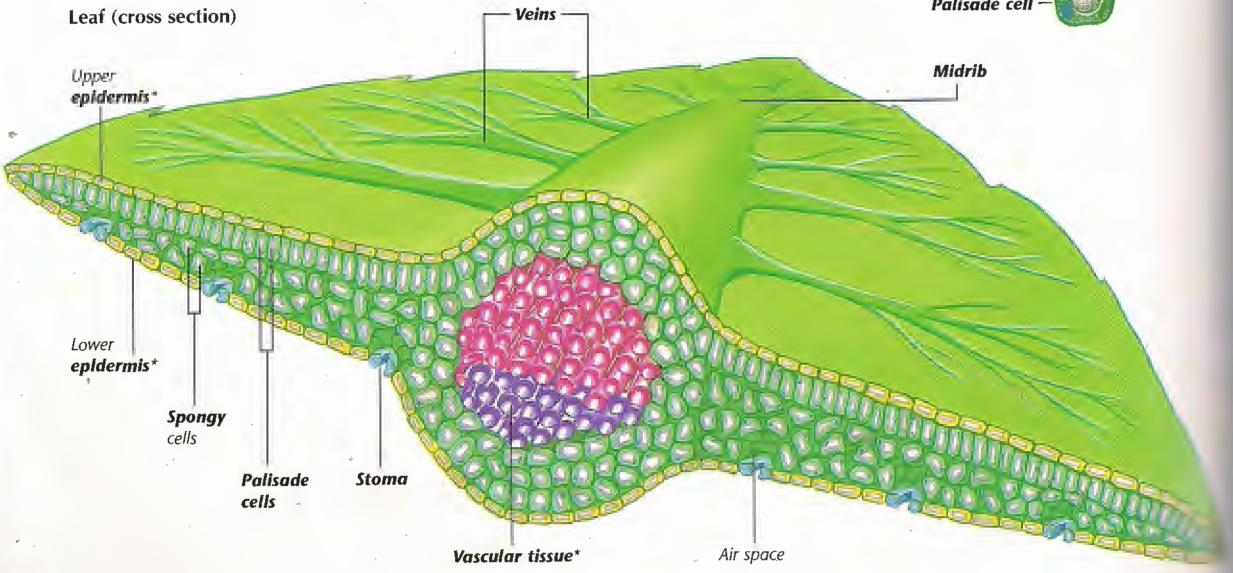
Veins
Midrib
Upper **epidermis***
Lower **epidermis***
Spongy cells
Palisade cells
Stoma
Vascular tissue*
Air space

---

* **Chloroplasts**, 254; **Epidermis**, 243; **Vascular tissue**, 242.

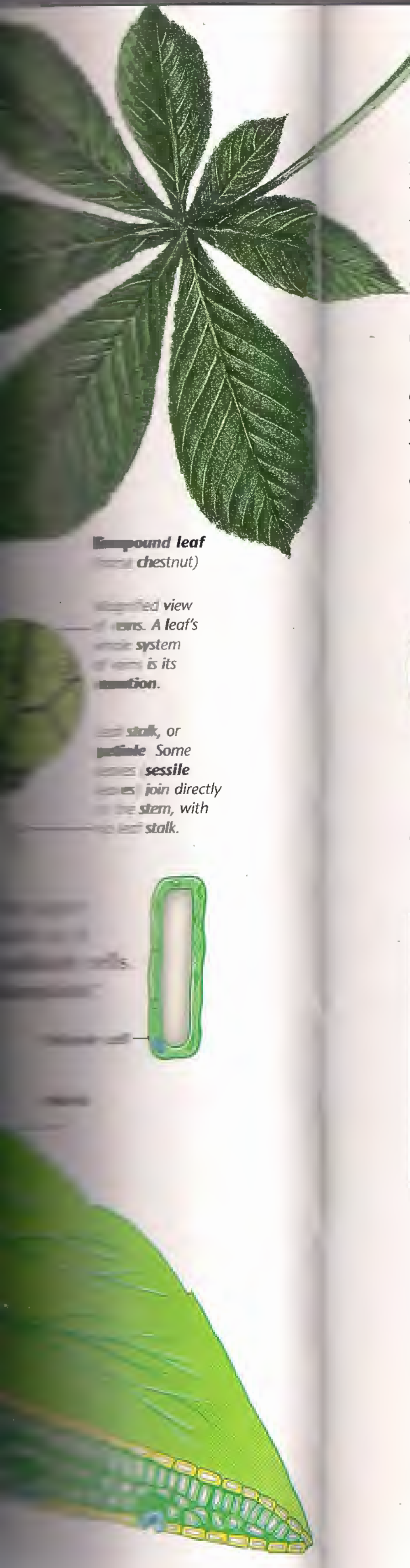
chestnut)
view
A leaf's
system
is its
stalk, or
Some
sessile
join directly
stem, with
stalk.

Stomata (sing. stoma)

Tiny openings in the **epidermis*** (outer skin), through which the exchange of water (**transpiration***) and gases takes place. Stomata are mainly found on the underside of leaves.

Guard cells

Pairs of crescent-shaped cells. The members of each pair are found on either side of a **stoma**, which they open and close by changing shape. This controls water and gas exchange. Guard cells are the only surface cells with **chloroplasts***.

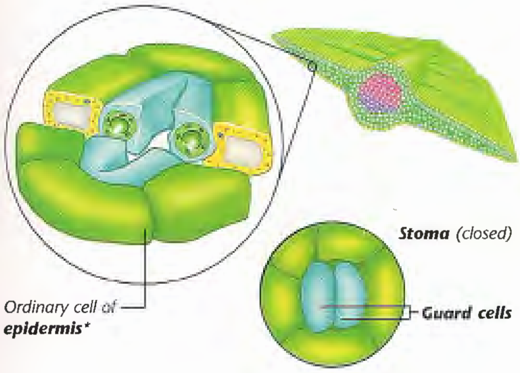
Close-up of **stoma** (open)
Leaf cross section
Stoma (closed)
Guard cells
Ordinary cell of **epidermis***

Leaf trace

An area of **vascular tissue*** which branches off that of a stem to become the central **vein** of a leaf.

Abscission layer

A layer of cells at the base of a leaf stalk which separates from the rest of the plant at a certain time of year (stimulated by a **hormone*** called **abscisic acid**). This makes the leaf fall off, forming a **leaf scar** on the stem.

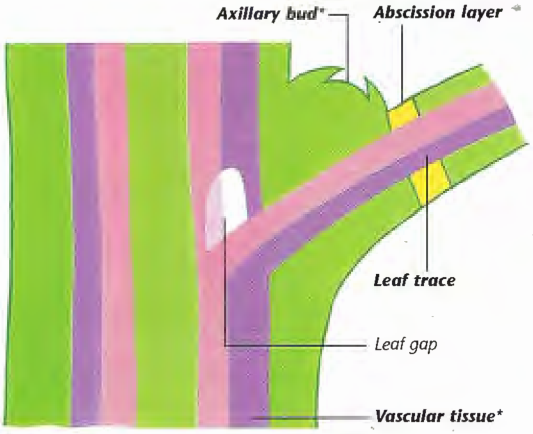
Axillary bud*
Abscission layer
Leaf trace
Leaf gap
Vascular tissue*

Special leaves

Stipule

A small, stalkless leaf at the base of a leaf stalk in many plants.

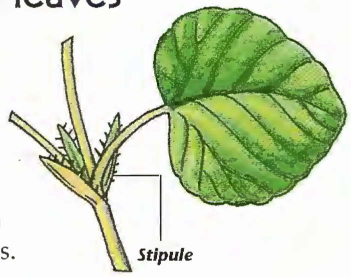

Tendril

A special thread-like leaf (or stem) which either twines around or sticks to a support.

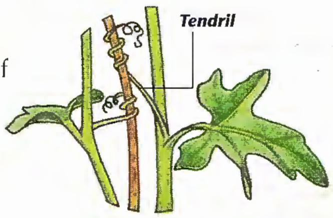

Bract

A leaf at the base of a flower stalk in many plants.

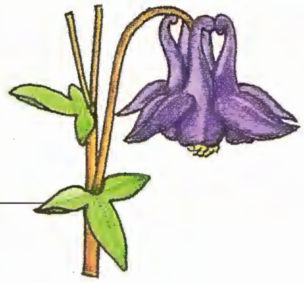

Spine

A specially modified leaf of a cactus. It has a reduced surface area to avoid losing much water.

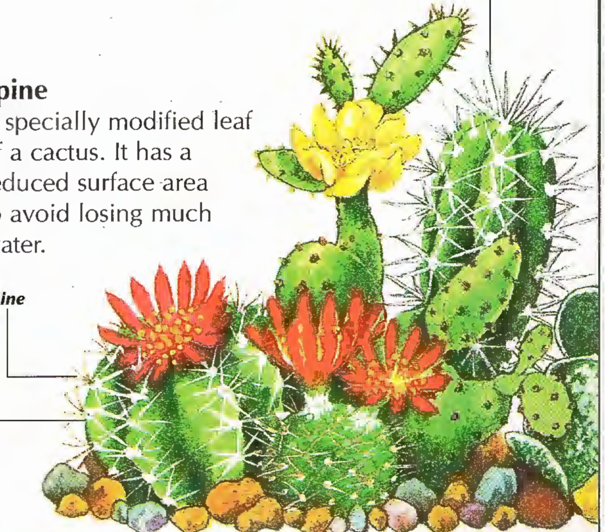

---

* **Axillary bud**, 244; **Chloroplasts**, 254; **Epidermis**, 243; **Hormones**, 336; **Transpiration**, 252; **Vascular tissue**, 242.

Types of compound leaf

Shown on this page are some types of **compound leaf** (leaves made up of **leaflets***), as well as some common leaf arrangements and leaf edges, or **margins**. The pictures are not to scale.

Trifoliate
Three **leaflets** grow from the same point.

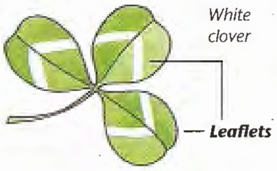
White clover
Leaflets

Ternate
Special type of **trifoliate** leaf. Each **leaflet** has three **lobes**.

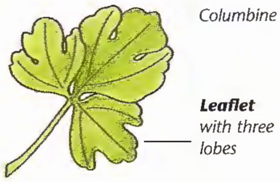
Columbine
Leaflet with three lobes

Palmate
The **leaflets** (five or more) radiate from one common point.

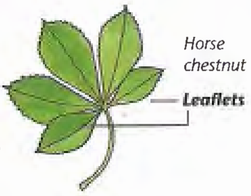
Horse chestnut
Leaflets

Pinnate
The **leaflets**, or **pinnae** (sing. **pinna**), are in **opposite** pairs.

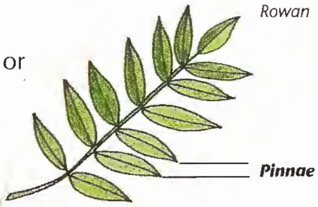
Rowan
Pinnae

Bipinnate/ tripinnate
A **pinnate** leaf with pinnate **leaflets**.

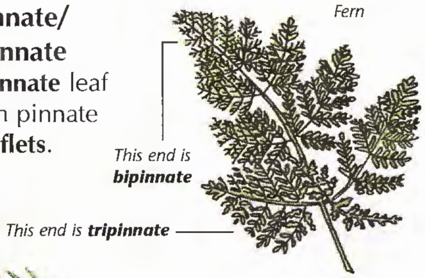
Fern
This end is **bipinnate**
This end is **tripinnate**

Palm trees have large **pinnate** leaves.

Leaf arrangements

Spiral
Leaves growing out from points forming a spiral around the stem.

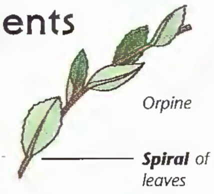
Orpine
Spiral of leaves

Opposite
Leaf pairs whose members grow from opposite stem sides.

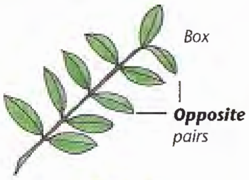
Box
Opposite pairs

Decussate
Opposite pairs, each pair at right angles to the one before.

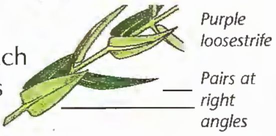
Purple loosestrife
Pairs at right angles

Rosette or **whorl**
A circle of leaves growing from one point.

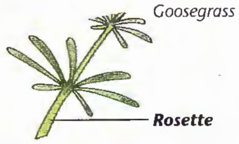
Goosegrass
Rosette

Basal rosette
A **rosette** growing at the base of a stem.

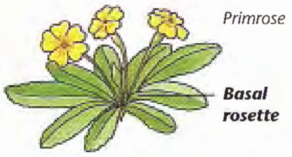
Primrose
Basal rosette

Perfoliate
Single or paired leaves whose bases are fused around the stem.

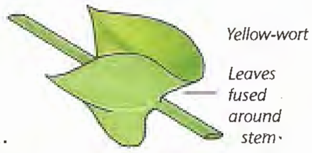
Yellow-wort
Leaves fused around stem

Leaf margins

Entire
The leaf **margin** has no indentations of any kind.

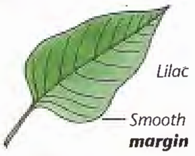
Lilac
Smooth **margin**

Serrate
The leaf **margin** has tiny jagged "teeth". May also be **lobed**.

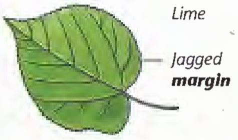
Lime
Jagged **margin**

Lobed
The leaf **margin** forms sections, or **lobes**. May also be **serrate**.

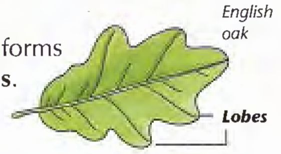
English oak
Lobes

---

* **Leaflets**, 248.

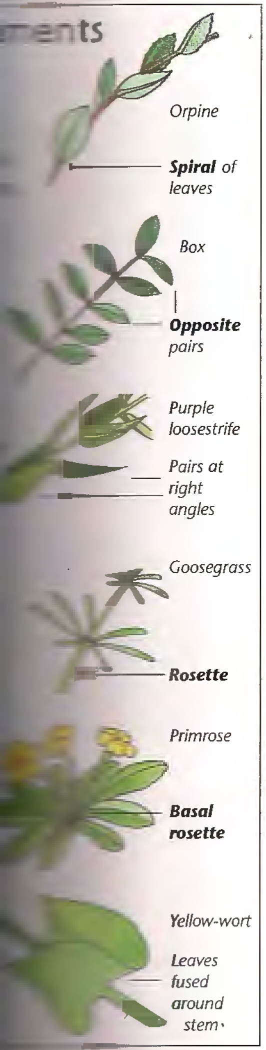
ments
Orpine
Spiral of leaves
Box
Opposite pairs
Purple loosestrife
Pairs at right angles
Goosegrass
Rosette
Primrose
Basal rosette
Yellow-wort
Leaves fused around stem

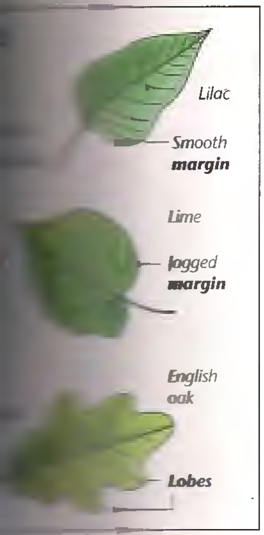
Lilac
Smooth margin
Lime
Jagged margin
English oak
Lobes

PLANT SENSITIVITY

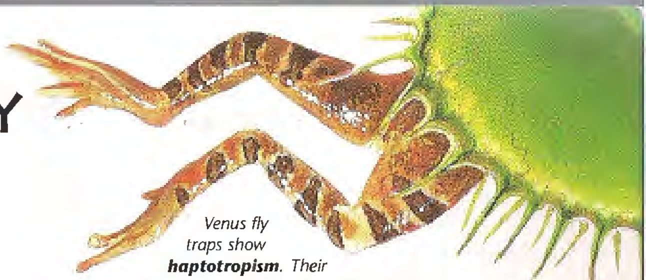
Venus fly traps show **haptotropism**. Their leaves respond to touch, snapping up animals such as insects and small frogs.

Plants have no nervous system, but they do still show **sensitivity**, i.e. they react to certain forms of stimulation. They do this by moving specific parts or by growing. This is called **tropism**. **Positive tropism** is movement or growth toward the stimulus and **negative tropism** is movement or growth away from it.

Hydrotropism

Response to water. For example, some roots may grow out sideways if there is more water in that direction.

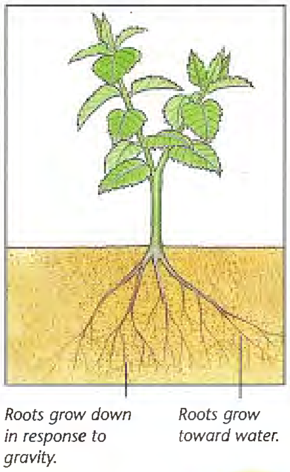
Roots grow down in response to gravity.
Roots grow toward water.

Geotropism

Response to the pull of gravity. This is shown by all roots, i.e. they all grow down through the soil.

Phototropism

Response to light. When the light is sunlight, the response is called **heliotropism**. Most leaves and stems show this by curving around to grow toward the light.

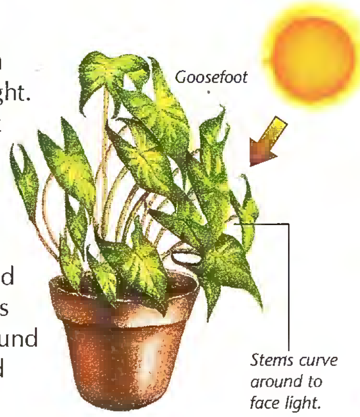
Goosefoot
Stems curve around to face light.

> **Growth hormones** or **growth regulators**
> Substances which promote and regulate plant growth. They are produced in **meristems*** (areas where cells are constantly dividing). **Auxins**, **cytokinins** and **gibberellins** are types of growth hormone.

Haptotropism or thigmotropism

Response to touch or contact. For example, the sticky hairs of a sundew plant curl around an insect when it comes into contact with them.

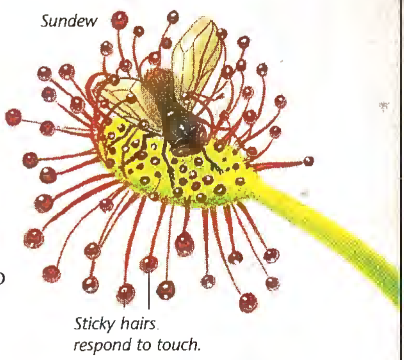
Sundew
Sticky hairs respond to touch.

Photoperiodism

The response of plants to the length of day or night (**photoperiods**), especially with regard to the production of flowers. It depends on a number of things, e.g. the plant's age and the temperature of its environment. **Long-night plants** only produce flowers if the night is longer than a certain length (called its **critical length**), **short-night plants** only if it is shorter. It is thought that a "message" to produce flowers is carried to the relevant area by a **hormone***, produced in the leaves when the conditions are right. This hormone has been called **florigen**. Some plants are **night-neutral plants**, i.e. their flowering does not depend on the length of night (see pictures, below).

These three plants each produce flowers according to different **photoperiods**.

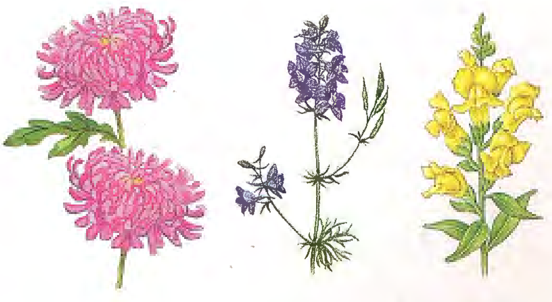
Chrysanthemum (**long-night plant**)
Larkspur (**short-night plant**)
Snapdragon (**night-neutral plant**)

---

* **Hormones**, 336; **Meristem**, 244.

PLANT FLUID TRANSPORTATION

The transportation of fluids in a plant is called **translocation**. The fluids travel within the **vascular tissue***, made up of **xylem*** and **phloem***. The xylem carries water (with dissolved minerals) from the roots to the leaves. The phloem carries food from the leaves to areas where it is needed.

Transpiration

The loss of water by evaporation, mainly through tiny holes called **stomata*** which are found on the underside of leaves.

Transpiration stream

A constant chain of events inside a plant. As the outer leaf cells lose water by **transpiration**, the concentration of minerals and sugars in their **vacuoles*** becomes higher than that of the cells further in. Water then passes outward by **osmosis***, causing more water to be "pulled" up through the tubes of the **xylem*** in the stem and roots (helped by **capillary action**). The roots then take in more water.

Capillary action

The way that fluids travel up narrow tubes (see also page 23). The molecules of the fluid are "pulled" upward by the attraction between them and the molecules of the tube.

Root pressure

A pressure which builds up in the roots of some plants. In all plants, water travels in from the soil and on through the layers of root cells by **osmosis***. In plants which develop root pressure, the pressure of this water movement is enough to force the water some way up into the tubes of the **xylem***. It is then "pulled" on upward by the **transpiration stream**. In other plants, the movement of water through root cells is all due to the "pull" of the transpiration stream.

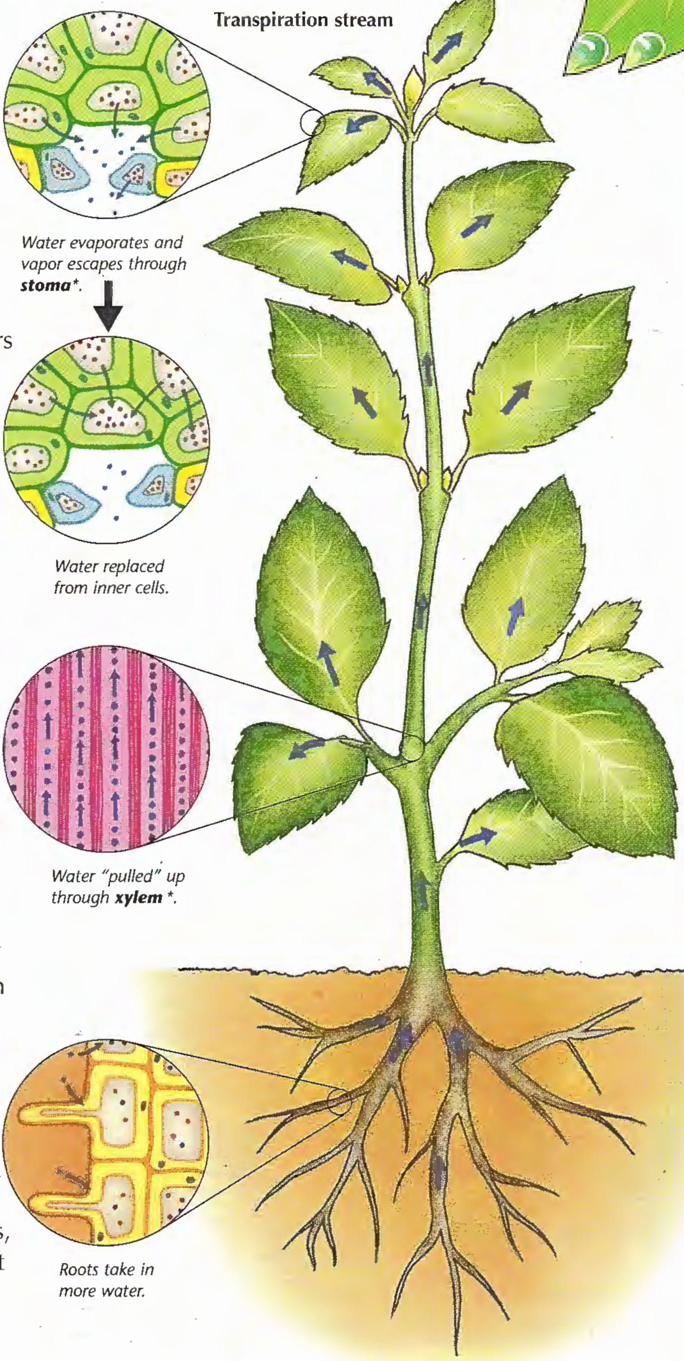

Transpiration stream

Water evaporates and vapor escapes through ***stoma****.*

Water replaced from inner cells.

Water "pulled" up through ***xylem*** *.*

Roots take in more water.

---

* **Osmosis**, 329; **Phloem**, 243; **Stomata**, 249; **Vacuoles**, 238; **Vascular tissue**, 242; **Xylem**, 243.

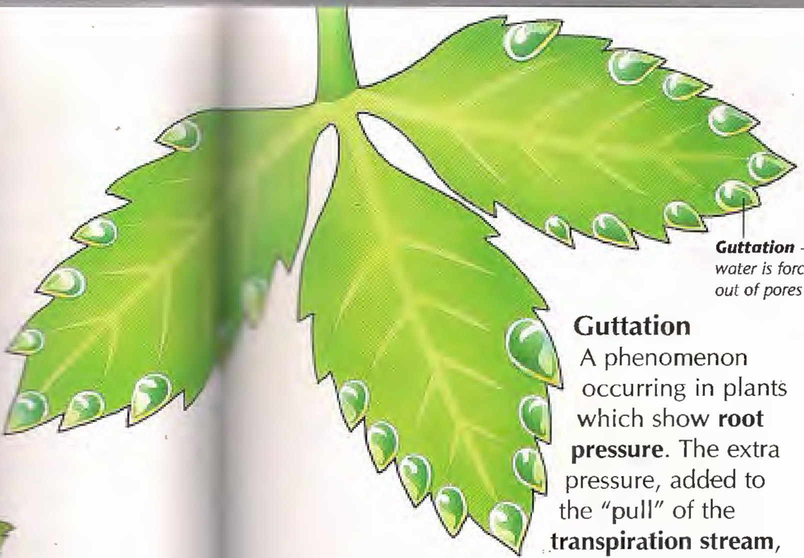
Guttation – water is forced out of pores

Guttation

A phenomenon occurring in plants which show **root pressure**. The extra pressure, added to the "pull" of the **transpiration stream**, may cause drops of water to be forced out of water-secreting areas of cells (**hydathodes**) via tiny pores at the tips or along the edges of the leaves.

Turgor

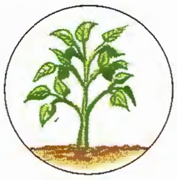
Healthy plant

The state of the cells in a healthy plant when its cells can take in no more water. Each cell is then said to be **turgid**. This means that water has passed by **osmosis*** into the **cell sap*** (dissolved minerals and sugars) in the cell's large central **vacuole***, and the vacuole has pushed as far out as it can go. The vacuole can push out no further because its outward pressure (called **turgor pressure**) is equalled by the opposing force of the rigid **cell wall*** (**wall pressure**). Turgid cells are important because they enable a plant to stand firm and upright.

Turgor

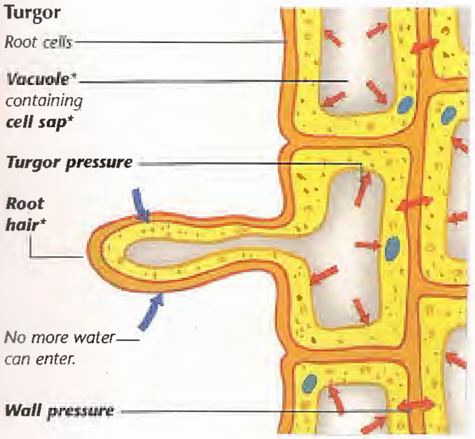
Root cells
Vacuole* containing **cell sap***
Turgor pressure
Root hair*
No more water can enter.
Wall pressure

Wilting

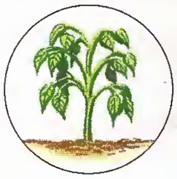
Wilting plant

A state of drooping, found in a plant subjected to certain conditions, such as excess heat. The plant is losing more water (by **transpiration**) than it can take in, and the **turgor pressure** (see **turgor**) of its cell **vacuoles*** drops. The cells become limp and can no longer support the plant, so it will droop.

Wilting

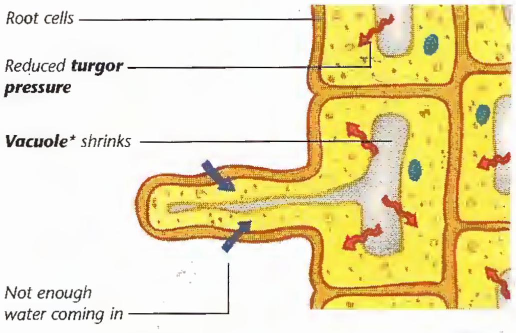
Root cells
Reduced **turgor pressure**
Vacuole* shrinks
Not enough water coming in

Plasmolysis

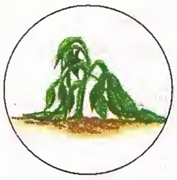
Dying plant

An extreme state in a plant, which may cause it to die. Such a plant is losing a large amount of water, often not only by **transpiration** in excess heat (see **wilting**), but also by **osmosis*** into very dry soil or soil with a very high concentration of minerals. The **vacuoles*** of the plant cells then shrink so much that they pull the **cytoplasm*** away from the **cell walls***.

Plasmolysis

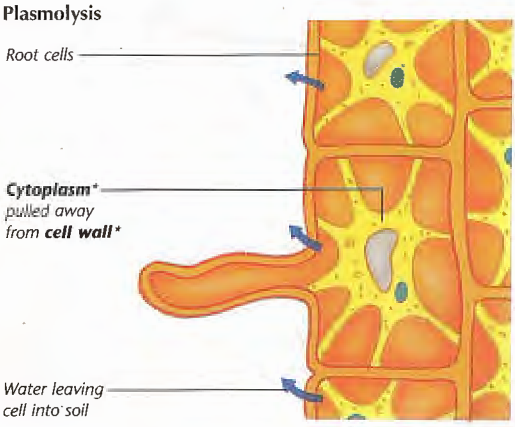
Root cells
Cytoplasm* pulled away from **cell wall***
Water leaving cell into soil

* **Cell sap**, 238 (**Vacuoles**); **Cell wall**, **Cytoplasm**, 238; **Osmosis**, 329; **Root hairs**, 245.

PLANT FOOD PRODUCTION

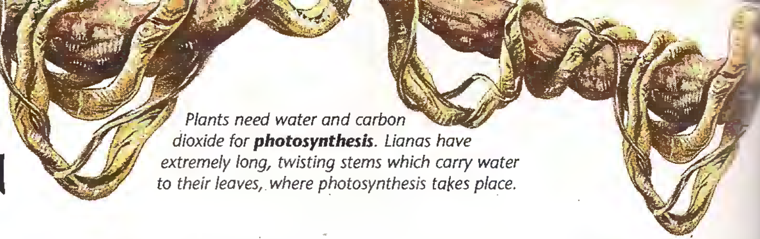
*Plants need water and carbon dioxide for **photosynthesis**. Lianas have extremely long, twisting stems which carry water to their leaves, where photosynthesis takes place.*

Most plants have the ability to make the food they need for growth and energy (unlike animals, which must take it in). The manufacturing process by which they make their complex food substances from other, simpler substances is called **photosynthesis**.

Photosynthesis

The series of chemical reactions (for basic equation, see page 209) by which green plants make their food. It occurs mainly in the **palisade cells***. Carbon dioxide is combined with water, using energy taken in from sunlight by **chloroplasts**. This produces oxygen as well as the plant's food (see diagram, page 255).

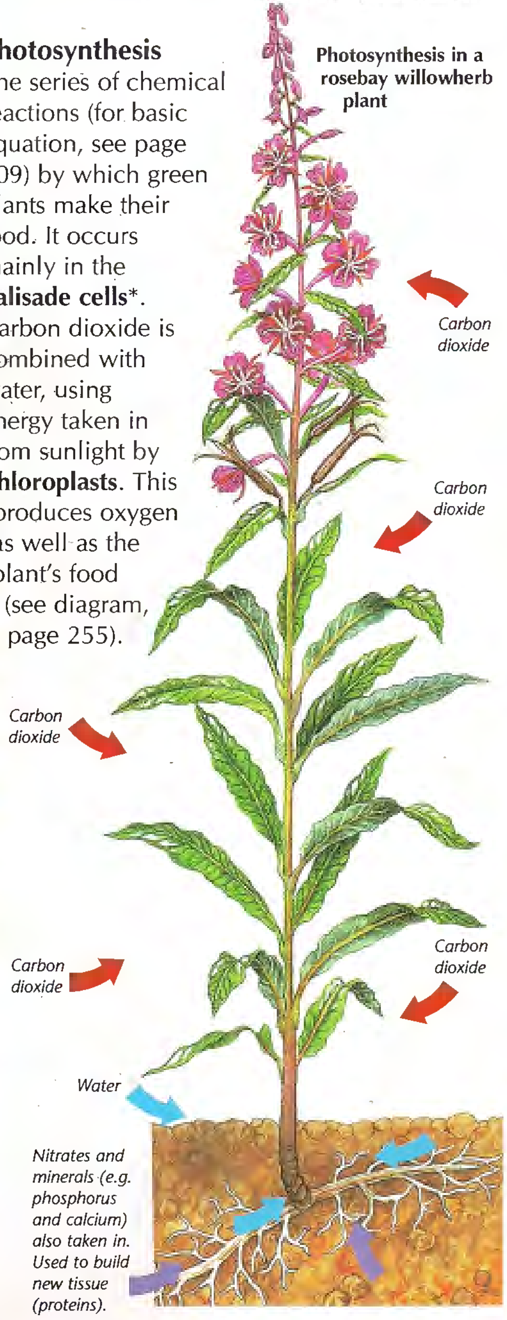
Photosynthesis in a rosebay willowherb plant
Carbon dioxide
Carbon dioxide
Carbon dioxide
Carbon dioxide
Carbon dioxide
Water
Nitrates and minerals (e.g. phosphorus and calcium) also taken in. Used to build new tissue (proteins).

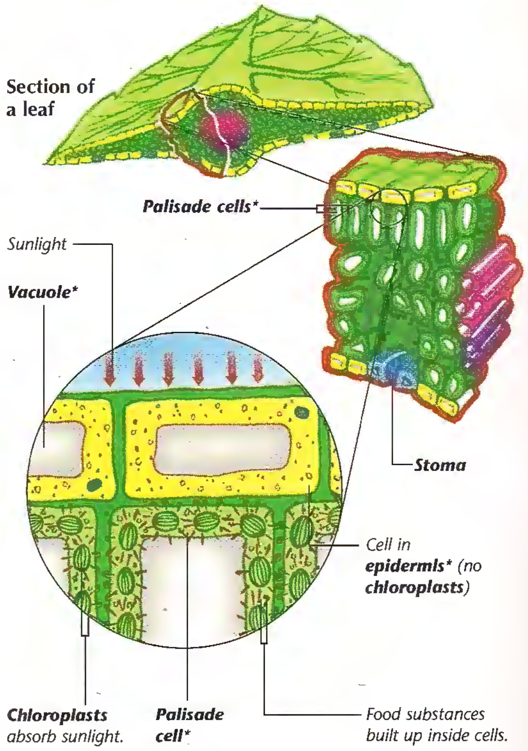
Section of a leaf
Palisade cells*
Sunlight
Vacuole*
Stoma
Cell in epidermis* (no chloroplasts)
Chloroplasts absorb sunlight.
Palisade cell*
Food substances built up inside cells.

Chloroplasts

Tiny bodies in plant cells (mainly in the leaves) which contain a green **pigment** called **chlorophyll**. This absorbs the Sun's light energy and uses it to "power" **photosynthesis**. Chloroplasts can move around inside a cell, according to light intensity and direction. See also page 240.

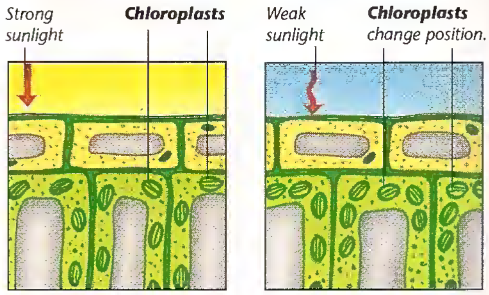
Strong sunlight
Chloroplasts
Weak sunlight
Chloroplasts change position.

---

* **Epidermis**, 243; **Internal respiration**, 334; **Palisade cells**, 248 (**Palisade layer**); **Vacuoles**, 238.

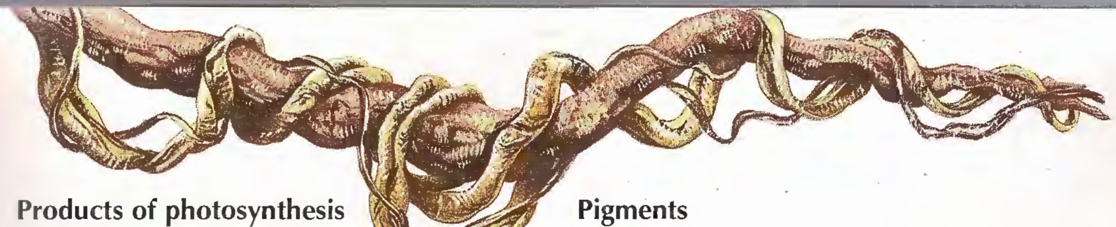

Products of photosynthesis

The process of photosynthesis works in co-ordination with that of **internal respiration***, the breakdown of food for energy. Photosynthesis produces oxygen and carbohydrates (needed for internal respiration), and internal respiration produces carbon dioxide and water (needed for photosynthesis).

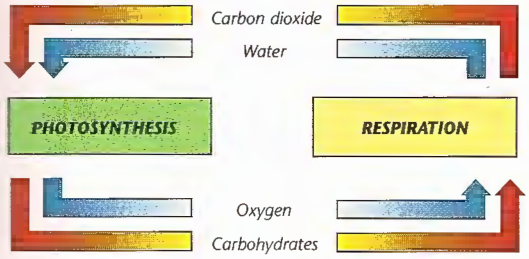
Carbon dioxide
Water
PHOTOSYNTHESIS
RESPIRATION
Oxygen
Carbohydrates

At most times, one of the two processes is occurring at a faster rate than the other. This means that excess amounts of its products are being produced, and not enough of the substances it needs are being made in the plant. In this case, extra amounts must be taken in and excess amounts given off or stored (see pictures 2 and 4 below).

Pigments

Substances which absorb light. White light is actually made up of a spectrum of many different colors. Each pigment absorbs some colors and reflects others.

Chlorophyll is a pigment found in all leaves. It absorbs blue, violet and red light, and reflects green light. This is why leaves look green.

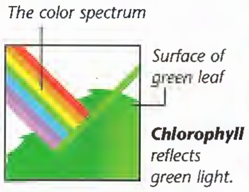
The color spectrum
Surface of green leaf

Chlorophyll reflects green light.

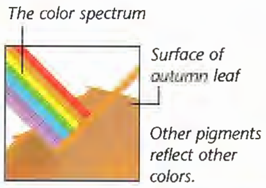
The color spectrum
Surface of autumn leaf

Other pigments reflect other colors.

Other pigments, such as **xanthophyll**, **carotene** and **tannin** are also present in leaves. They reflect orange, yellow and red light, but are masked by chlorophyll during the growing season. In autumn, the chlorophyll breaks down, and so the autumn colors appear. Plant pigments are used to give color to many things, e.g. paints and plastics.

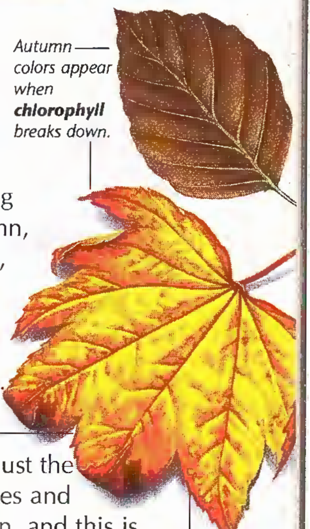

Autumn colors appear when **chlorophyll** breaks down.

Compensation points

Two points in a 24-hour period (normally around dawn and around dusk) when the processes of **photosynthesis** and **internal respiration*** (see above) are exactly balanced.

Photosynthesis is producing just the right amounts of carbohydrates and oxygen for internal respiration, and this is producing just the right amounts of carbon dioxide and water for photosynthesis.

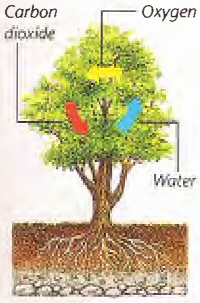
Carbon dioxide
Oxygen
Water

1. Around dawn (**compensation point**)

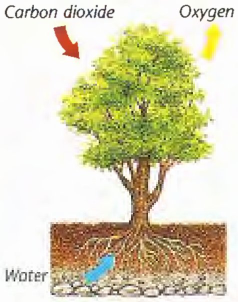
Carbon dioxide
Oxygen
Water

2. Midday (bright light, so faster **photosynthesis**)

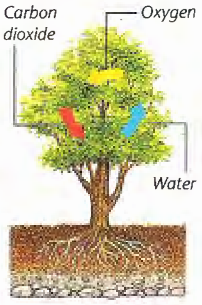
Carbon dioxide
Oxygen
Water

3. Around dusk (**compensation point**)

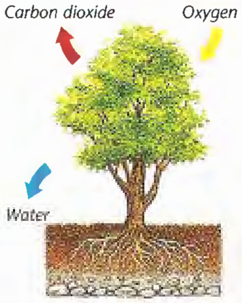
Carbon dioxide
Oxygen
Water

4. Midnight (no light, so no **photosynthesis**)

---

* **Internal respiration**, 334.

FLOWERS

The **flowers** of a plant contain its organs of **reproduction*** (producing new life). In **hermaphrodite** plants, e.g. buttercups and poppies, each flower has both male and female organs. **Monoecious** plants, e.g. corn, have two types of flowers on one plant – **staminate** flowers, which have just male organs, and **pistillate** flowers, which have just female organs. **Dioecious** plants, e.g. holly, have staminate flowers on one plant and pistillate flowers on a separate plant.

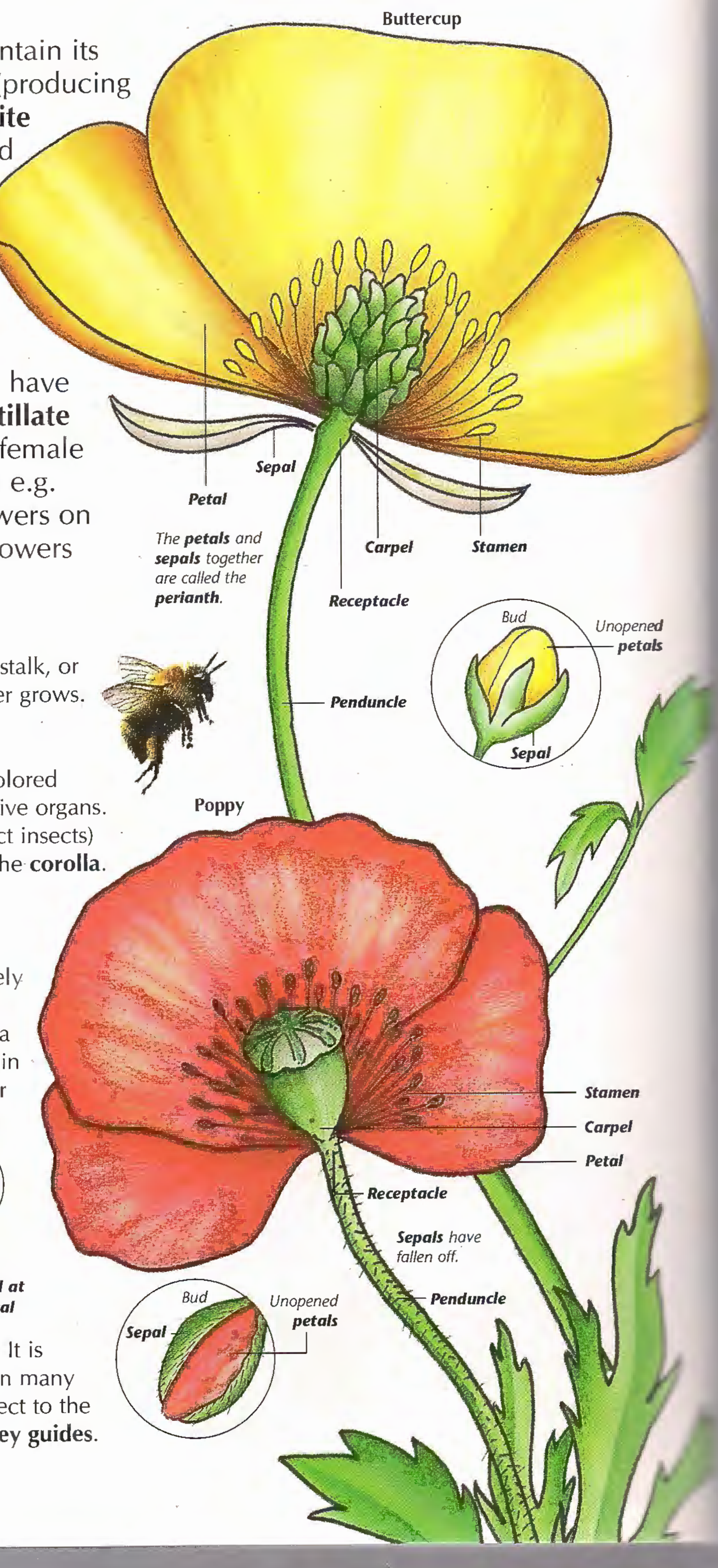
Buttercup
Petal
Sepal
Carpel
Stamen
Receptacle
Pendunde
Bud
Unopened petals
Sepal
Poppy
Stamen
Carpel
Petal
Receptacle
Pendunde
Bud
Sepal
Unopened petals

*The **petals** and **sepals** together are called the **perianth**.*

***Sepals** have fallen off.*

Receptacle

The expanded tip of the flower stalk, or **peduncle**, from which the flower grows.

Petals

The delicate, usually brightly colored structures around the reproductive organs. They are often scented (to attract insects) and are known collectively as the **corolla**.

Sepals

The small, leaf-like structures around a bud, known collectively as the **calyx**. In some flowers, e.g. buttercups, they remain as a ring around the opened **petals**; in others, e.g. poppies, they wither and fall off.

Nectar held at base of petal

Nectaries

Areas of cells at the base of the **petals** which produce a sugary liquid called **nectar**. This attracts insects needed for **pollination***. It is thought that the dark lines down many petals are there to direct an insect to the nectar, and they are called **honey guides**.

* **Pollination**, **Reproduction**, 258.

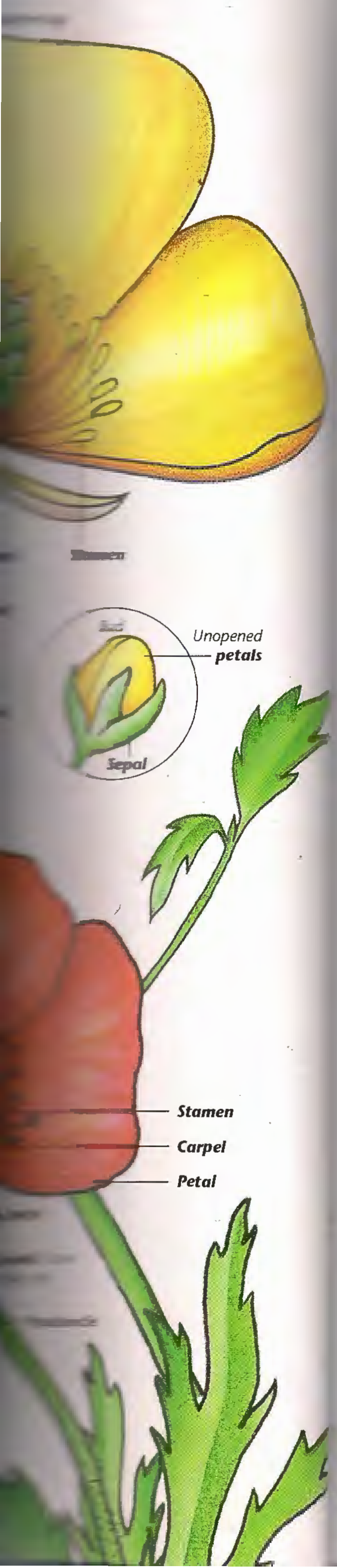
Unopened
petals
Sepal
Stamen
Carpel
Petal

The female organs

Carpel or pistil

A female reproductive organ, consisting of an **ovary**, **stigma** and **style**. Some flowers have only one carpel, others have several clustered together.

Ovaries

Female reproductive structures. Each is the main part of the **carpel** and contains one or more tiny bodies called **ovules***, each of which contains a female sex cell. An ovule is fixed by a stalk (**funicle**) to an area of the ovary's inside wall called the **placenta**. The stalk is attached to the ovule at a point called a **chalaza**.

Stigma

The uppermost part of a **carpel**, with a sticky surface to which grains of **pollen*** become attached during **pollination***.

Style

The part of a **carpel** which joins the **stigma** to the **ovary**. Many flowers have an obvious style, e.g. daffodils, but in others it is very short (e.g. buttercups) or almost non-existent (e.g. poppies).

Gynaecium

The whole female reproductive structure, made up of one or more **carpels**.

Buttercup carpels

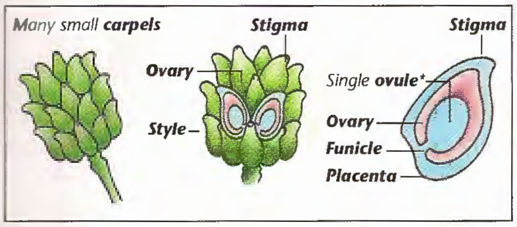
Many small carpels
Stigma
Ovary
Style
Stigma
Single ovule*
Ovary
Funicle
Placenta

Poppy carpel

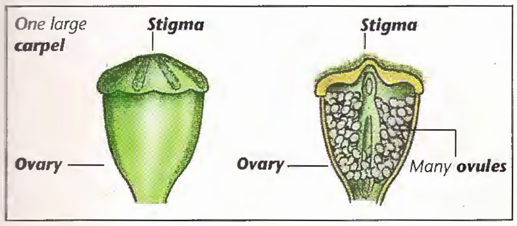
One large carpel
Stigma
Ovary
Stigma
Ovary
Many ovules

The male organs

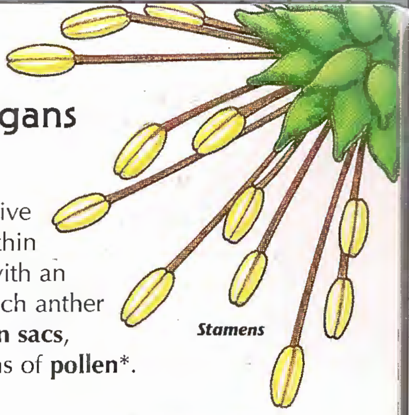

Stamens

The male reproductive organs. Each has a thin stalk, or **filament**, with an **anther** at the tip. Each anther is made up of **pollen sacs**, which contain grains of **pollen***.

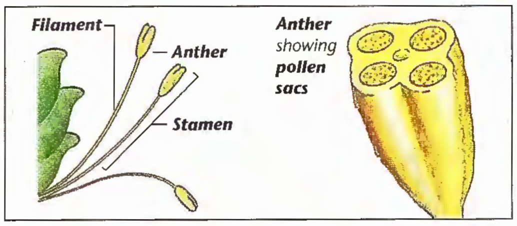
Filament
Anther
Stamen
Anther showing pollen sacs

Androecium

A collective term for all the male parts of a flower, i.e. all the **stamens**.

How the parts are arranged

Hypogynous flower

The **carpel** (or carpels) sit on top of the **receptacle**; all the other parts grow out from around its base. The position of the carpel is described as **superior**.

Tulip

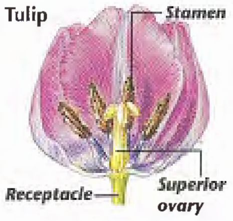
Stamen
Receptacle
Superior ovary

Perigynous flower

The **carpel** (or carpels) rest in a cup-shaped **receptacle**; all the other parts grow out from around its rim. The position of the carpel is described as **superior**.

Cherry

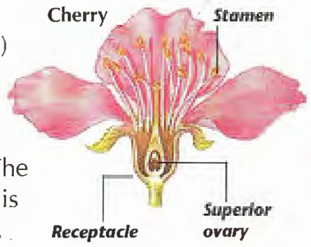
Stamen
Receptacle
Superior ovary

Epigynous flower

The flower parts grow from the top of a **receptacle** which completely encloses the **ovary** (or ovaries), but not the **stigma** and **style**. The position of the ovary is described as **inferior**.

Daffodil

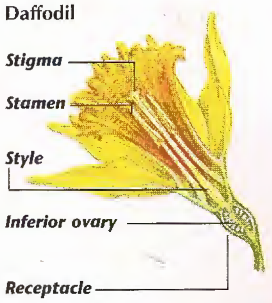
Stigma
Stamen
Style
Inferior ovary
Receptacle

---

* **Ovules, Pollen, Pollination**, 258.

REPRODUCTION IN A FLOWERING PLANT

Insects play a major part in the ***cross pollination*** *of some plant species (see page 259).*

Reproduction is the creation of new life. All flowering plants reproduce by **sexual reproduction***, when a male **gamete*** (sex cell) joins with a female gamete. In flowering plants, the male gametes (strictly speaking only male **nuclei***) are held in **pollen**, and the female gametes in **ovules**.

Pollen

Tiny grains formed by the **stamens*** (male parts) of flowers (see picture, right). Each grain is a special cell which has two **nuclei***. When a pollen grain lands on an **ovary*** (female body), one nucleus (the **generative nucleus**) divides into two, forming two **male nuclei** (reproductive bodies – see introduction).

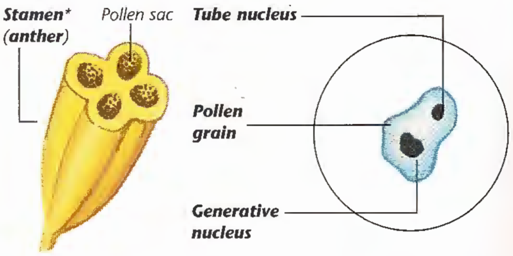

Stamen* (anther)
Pollen sac
Tube nucleus
Pollen grain
Generative nucleus

Ovules

The tiny structures inside a flower's female body, or **ovary***. They become seeds after **fertilization**. Each consists of an oval cell (the **embryo sac**), surrounded by layers of tissue called **integuments**, except at one point where there is a tiny hole (**micropyle**). Before fertilization, the embryo sac **nucleus*** undergoes several divisions (looked at in more detail on page 323 – under **gamete production, female**). This results in a number of new cells (some of which become part of the seed's food store), and two naked nuclei which fuse together. One of the new cells is the female **gamete*** (sex cell), or **egg cell**.

Pollination

The process by which a grain of **pollen** transfers its **male nuclei** (see **pollen**) into the **ovary*** of a flower. The grain lands on the **stigma***, and forms a **pollen tube**, under the control of the **tube nucleus** (the one which did not divide – see **pollen**). The tube grows down through the ovary tissue and enters an **ovule** via its **micropyle**. The two male nuclei then travel along it.

Fertilization

After **pollination**, one **male nucleus** (see **pollen**) fuses with the **egg cell** in the **ovule** to form a **zygote*** (the first cell of a new plant). The other joins with the two fused female nuclei to form a cell which develops into the **endosperm***.

Pollination in a poppy plant

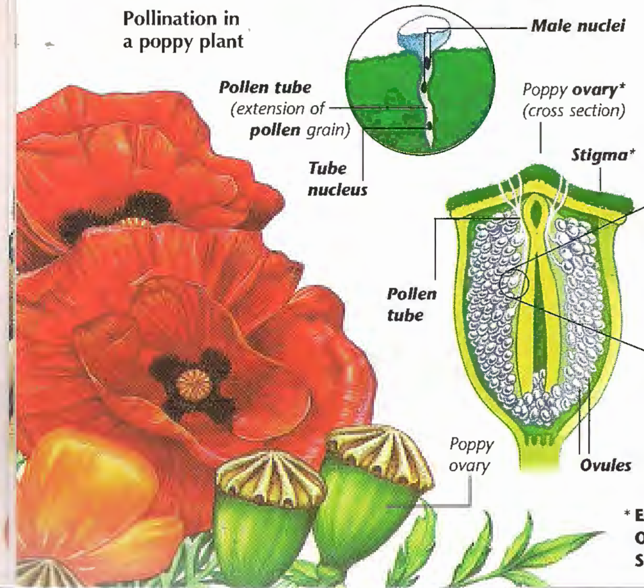

Male nuclei
Pollen tube
(extension of pollen grain)
Tube nucleus
Poppy ovary*
(cross section)
Stigma*
Pollen tube
Poppy ovary
Ovules

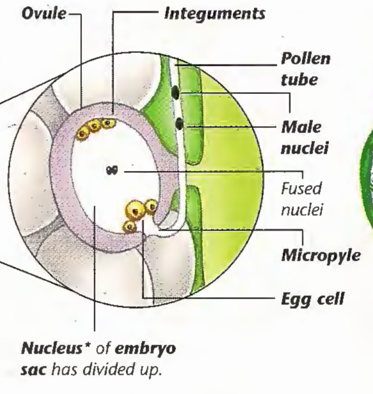

Ovule
Integuments
Pollen tube
Male nuclei
Fused nuclei
Micropyle
Egg cell

Nucleus* *of* ***embryo sac*** *has divided up.*

Cross section of ovary*

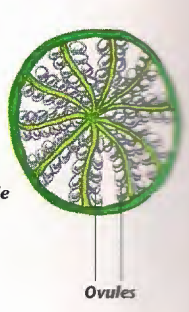

---

* **Endosperm**, 261; **Male nuclei**, 320 (**Gametes**); **Nucleus**, 238; **Ovaries**, 257; **Sexual reproduction**, 320; **Stamens**, 257; **Stigma**, 257; **Zygote**, 320.

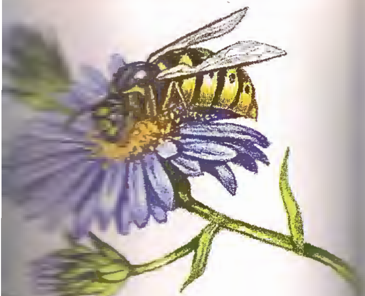

Insects play a major part in the ***cross pollination*** *of some plant species (see page 259).*

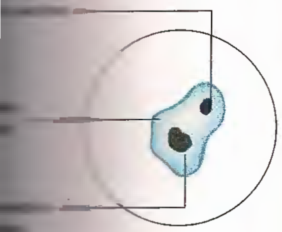

pollen

en) into the

nds on the

under the

one which did

be grows down

enters an **ovule**

nuclei then

eus (see

the **ovule** to

a new plant).

sed female

ops into the

Cross section of ovary*

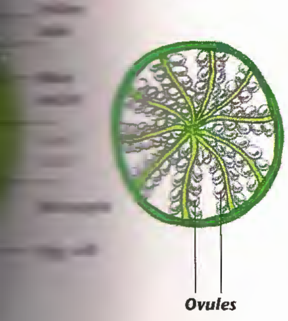

Nucleus, 238;

257;

Cross pollination

The **pollination** of one plant by **pollen** grains from another plant of the same type. (If the grains land on a different type of plant, they do not develop further, i.e. they do not produce **pollen tubes**.) The pollen may be carried by the wind, or by insects which drink the **nectar***.

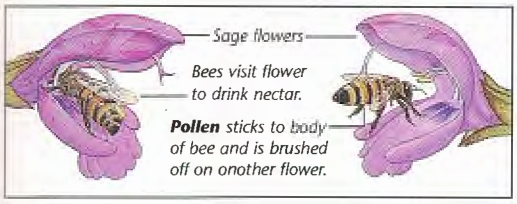
Sage flowers
Bees visit flower to drink nectar.
Pollen sticks to body of bee and is brushed off on another flower.

Self pollination

The **pollination** of a plant by its own **pollen** grains. For example, a bee orchid tries to attract male Eucera bees (for **cross pollination**) by looking and smelling like a female bee. But if it is not visited, its **stamens*** (male parts) bend over and transfer pollen to the **stigma*** of its **ovary*** (female body).

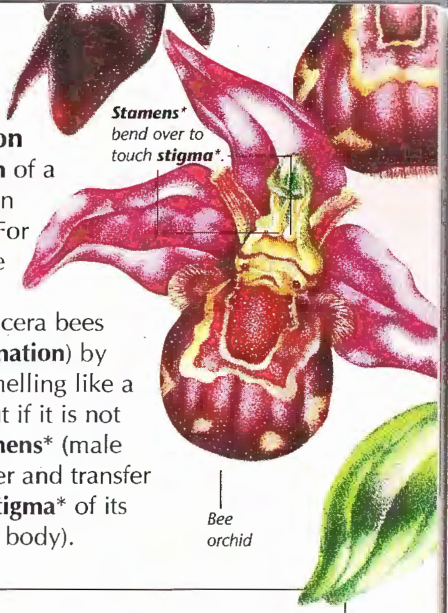
Stamens* bend over to touch **stigma***.
Bee orchid

Types and arrangements of flowers

Inflorescence

A group of flowers or **flowerheads** growing from one point.

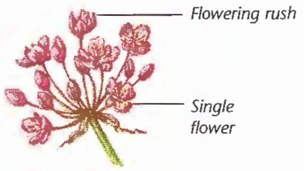
Flowering rush
Single flower

Flowerhead or composite flower

A cluster of tiny flowers, or **florets**.

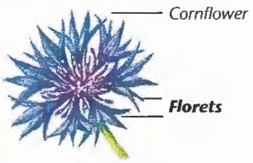
Cornflower
Florets

Umbellifer

An **inflorescence** with umbrella-shaped **flowerheads** (**umbels**).

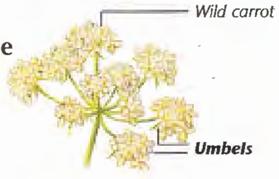
Wild carrot
Umbels

Ray florets

Florets with one long petal.

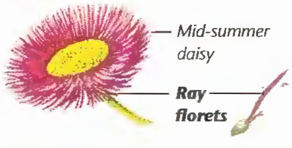
Mid-summer daisy
Ray florets

Disc florets

Florets whose petals are all the same size.

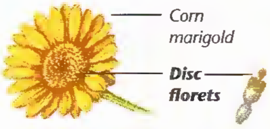
Corn marigold
Disc florets

Bell flower

Also called a **tubular** or **campanulate flower**. Its petals are joined to make a bell shape.

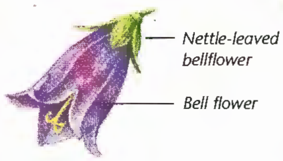
Nettle-leaved bellflower
Bell flower

Spurred flower

A flower with one or more petals extended backward to form **spurs**.

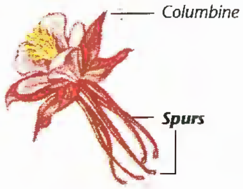
Columbine
Spurs

Lipped flower

A flower with two "lips" – an upper and lower one. The upper one often has a hood.

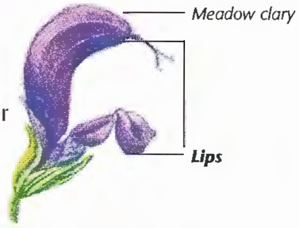
Meadow clary
Lips

Pea flower

A flower with an upper petal (the **standard**), two side petals (**wing petals**) and two lower petals forming the **keel** (which encloses the reproductive parts).

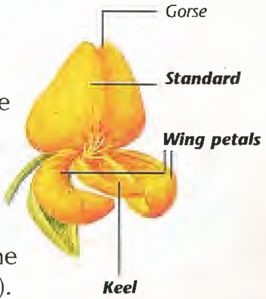
Gorse
Standard
Wing petals
Keel

*__Nectar__, 256 (**Nectaries**); **Ovaries**, 257; **Stamens**, 257; **Stigma**, 257.

SEEDS AND GERMINATION

After **fertilization*** in a flowering plant, an **ovule*** develops into a **seed**. This contains an **embryo**, i.e. a new developing plant, and a store of food. The **ovary*** ripens into a fruit, carrying the seed or seeds. You can find out more about different fruits on page 262.

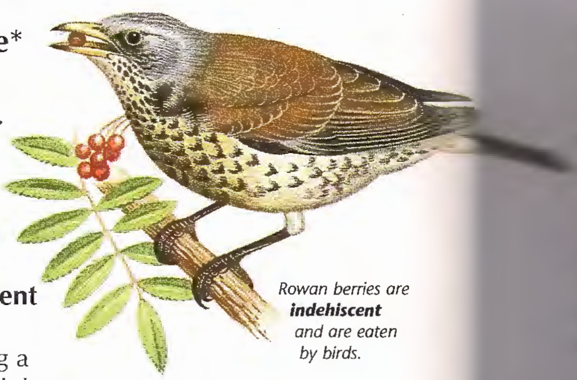

Rowan berries are ***indehiscent*** *and are eaten by birds.*

Dispersal or dissemination

The shedding of ripe seeds from the fruit of a parent plant. This happens in one of two main ways, depending on whether a fruit is **dehiscent** or **indehiscent**.

Dehiscent

A word describing a fruit from which the seeds are expelled before the fruit itself disintegrates. For example, a poppy capsule has holes in it, and the seeds are shaken out by the wind. Other fruits, e.g. pea pods, open spontaneously and "shoot" the seeds out. In many cases, the seeds may then be carried by wind, water or other means.

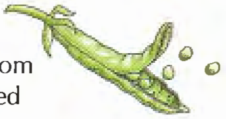

Pea pods burst open.

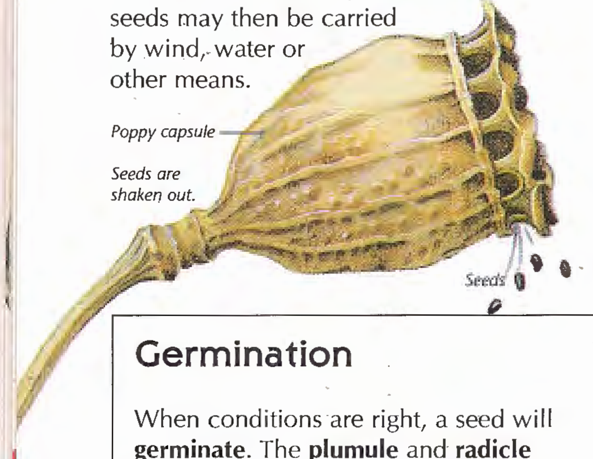

Seeds are shaken out.

Indehiscent

A word describing a fruit which becomes detached from the plant and disintegrates to free the seeds. For example, the "keys" of sycamores or the "parachutes" of dandelions are carried by the air, and hooked burrs catch on animal fur. The fruit then rots away in the ground to expose the seeds. Edible fruits may be eaten by animals, which then expel the seeds in their droppings.

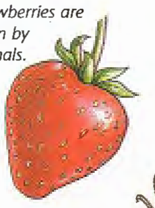

Strawberries are eaten by animals.

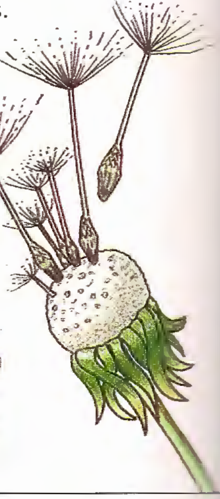

Dandelion "parachutes" are carried by the wind.

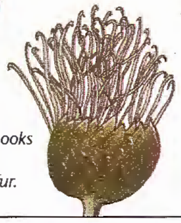

Burdock hooks catch on animals' fur.

Germination

When conditions are right, a seed will **germinate**. The **plumule** and **radicle** emerge from the seed coat, and begin to grow into the new plant, or **seedling**.

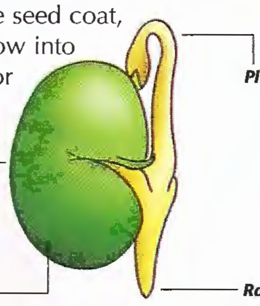

Hypogeal

A type of **germination**, e.g. in pea plants, in which the **cotyledons** remain below the ground within the **testa**, and the **plumule** is the only part to come above the ground.

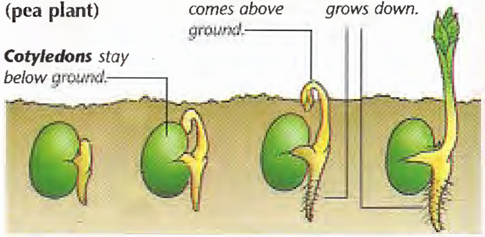

Hypogeal (pea plant)

Plumule *comes above ground.*

Radicle *grows down.*

Cotyledons *stay below ground.*

---

* **Fertilization**, 258; **Ovaries**, 257; **Ovules**, 258.

Parts of a seed

Hilum

A mark on a seed, showing where the **ovule*** was attached to the **ovary***.

Testa

The seed coat. It develops from the **integuments***.

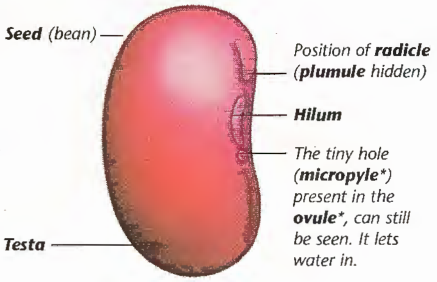
Seed (bean)
Position of **radicle** (**plumule** hidden)
Hilum
The tiny hole (**micropyle***) present in the **ovule***, can still be seen. It lets water in.
Testa

Plumule

The first bud, or **primary bud**, formed inside a seed. It will develop into the first shoot of the new plant.

Radicle

The first root, or **primary root**, of a new plant. It is formed inside a seed.

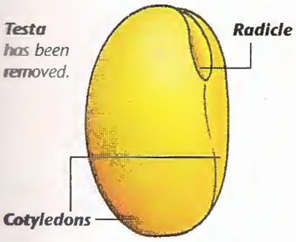
Testa has been removed.
Radicle
Cotyledons

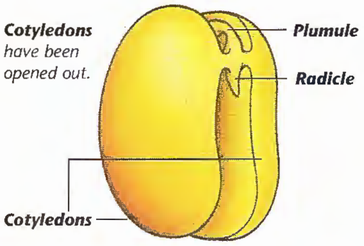
Cotyledons have been opened out.
Plumule
Radicle
Cotyledons

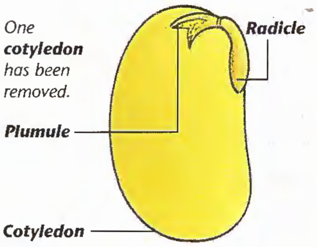
One **cotyledon** has been removed.
Radicle
Plumule
Cotyledon

Endosperm

A layer of tissue inside a seed which surrounds the developing plant and gives it nourishment. In some plants, e.g. pea, the **cotyledons** absorb and store all the endosperm before the seed is ripe; in others, e.g. grasses, it is not fully absorbed until after the seed **germinates**.

Cross section of a young bean

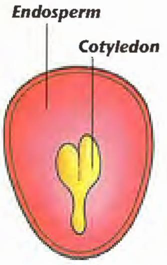
Endosperm
Cotyledon

Cross section of a mature bean

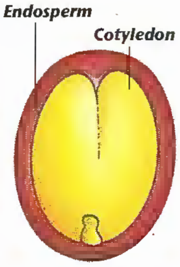
Endosperm
Cotyledon

Cotyledon or seed-leaf

A simple leaf which forms part of the developing plant. In some seeds, e.g. bean seeds, it absorbs and stores all the food from the **endosperm**. **Monocotyledons** are plants with one cotyledon, e.g. grasses; in **dicotyledons**, e.g. peas, there are two.

Epigeal

A type of **germination**, e.g. in bean plants, in which the **cotyledons** appear above the ground, below the first leaves – the true leaves.

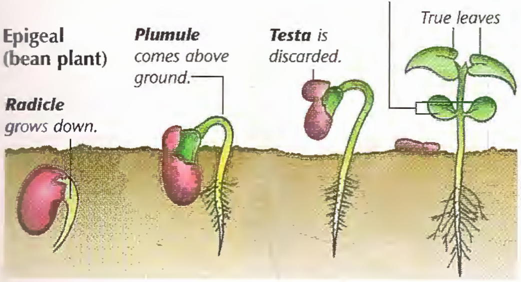
Epigeal (bean plant)
Radicle grows down.
Plumule comes above ground.
Testa is discarded.
Cotyledons
True leaves

Coleoptile

The first leaf of many **monocotyledons** (see **cotyledon**). It protects the first bud, and the first leaves emerge from it.

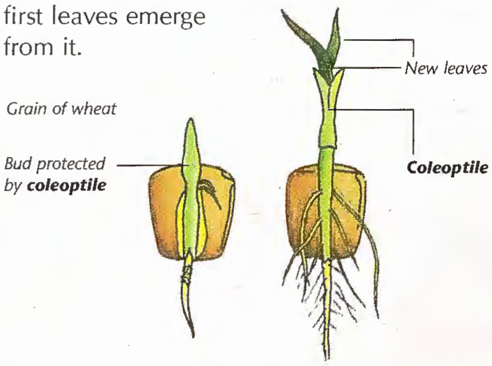
Grain of wheat
Bud protected by **coleoptile**
New leaves
Coleoptile

* **Integuments**, **Micropyle**, 258 (**Ovules**); **Ovaries**, 257.

FRUIT

A **fruit** contains the seeds of a plant. **True fruit** develop purely from the **ovary***, **false fruit** develop from the **receptacle*** as well (e.g. a strawberry). The outer wall of a fruit is called the **pericarp**. In some fruits, it is divided into an outer skin, or **epicarp**, a fleshy part, or **mesocarp**, and an inner layer, or **endocarp**. Listed below are the main types of fruit.

Legume or pod

A fruit with seeds attached to its inside wall. It splits along its length to open, e.g. a pea.

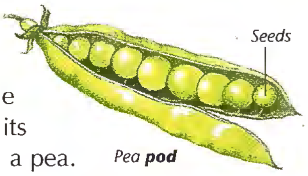

Nut

A dry fruit with a hard shell, which only contains one seed, e.g. a hazelnut or a walnut.

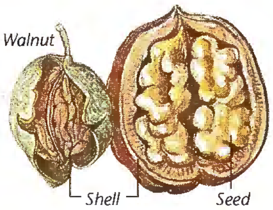

Drupe

A fleshy fruit with a hard seed in the middle, often known as a "stone", e.g. a plum.

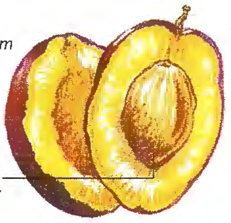

Grapefruit
Seed

Tomato seeds surrounded by juice

Berry

A fleshy fruit which contains many seeds, e.g. a tomato or a grapefruit. The "flesh" of citrus fruits is made up of tiny hairs, each one swollen up and full of juice.

Grain

Also called a **caryopsis** or **kernel**. A small fruit whose wall has fused with the seed coat, e.g. wheat.

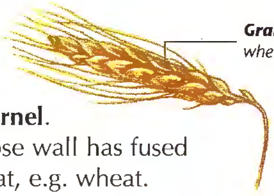

Pome

A fruit with a thick, fleshy, outer layer and a core, with the seeds enclosed in a capsule, e.g. an apple. Pomes are examples of **false fruits** (see introduction).

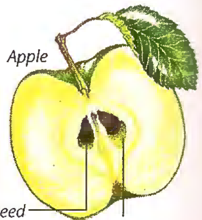

Achene

A small, dry fruit, with only one seed, e.g. a sycamore or buttercup fruit. A "winged" achene like a sycamore fruit is a **samara** or **key fruit**.

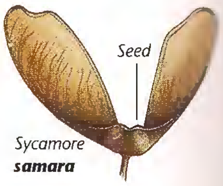

Artificial propagation

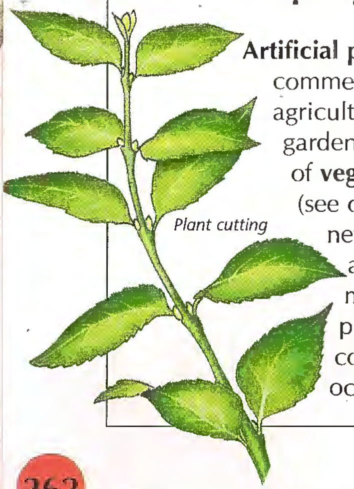

Artificial propagation is the commercial process, in agriculture and market gardening, which makes use of **vegetative reproduction** (see opposite). The fact that new plants need not always grow from seeds means that many more plants can be produced commercially than would occur naturally.

Cutting

A process in which a piece of a plant stem (the cutting) is removed from its parent plant and planted in soil, where it grows into a new plant. In some cases, it is first left in water for a while to develop roots.

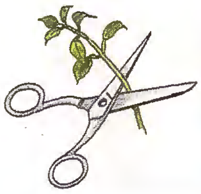

Taking a cutting

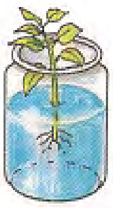

Cutting in water

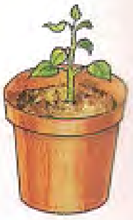

Cutting replanted

---

* **Ovaries**, 257; **Receptacle**, 256.

Tomato seeds surrounded by juice

Grains of wheat

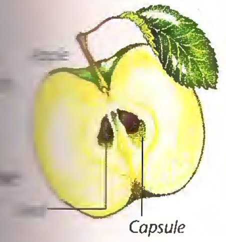
Capsule

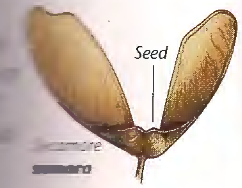

plant stem
parent plant
into a new
in water for

Cutting replanted

VEGETATIVE REPRODUCTION

As well as producing seeds, some plants have developed a special type of **asexual reproduction***, called **vegetative reproduction** or **vegetative propagation**, in which one part of the plant is able to develop unaided into a new plant.

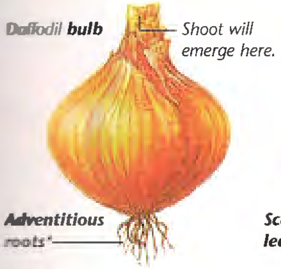
Daffodil **bulb**
Shoot will emerge here.
Adventitious roots*

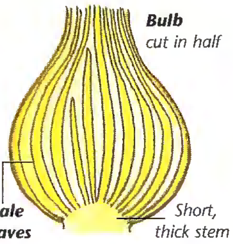
Bulb cut in half
Scale leaves
Short, thick stem

Bulb

A short, thick stem surrounded by scaly leaves (**scale leaves**) which contain stored food material. It is formed underground by an old, dying plant, and represents the first, resting, stage of a new plant, which will emerge as a shoot at the start of the next growing season. E.g. a daffodil bulb (see picture above).

Corm

A short, thick stem, similar to a **bulb**, except that the food store is in the stem itself. E.g. a crocus corm.

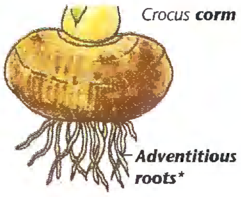
Crocus **corm**
Adventitious roots*

Rhizome

A thick stem, which has scaly leaves and grows horizontally underground. It produces roots along its length and also buds from which new shoots grow. Many grasses produce rhizomes, as well as other plants, e.g. mint and irises.

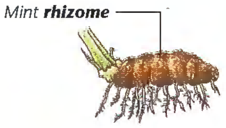
Mint **rhizome**

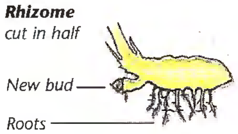
Rhizome cut in half
New bud
Roots

Stolon or runner

A stem which grows out horizontally near the base of some plants, e.g. the strawberry. The stolon puts down roots from points at intervals along this stem, and new plants grow at these points.

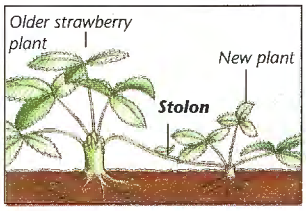
Older strawberry plant
New plant
Stolon

Tuber

A short, swollen, underground stem which contains stored food material and produces buds from which new plants will grow, e.g. a potato.

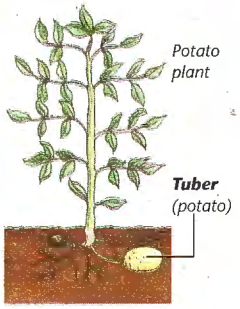
Potato plant
Tuber (potato)

Grafting

The process of removing a piece of a plant stem and re-attaching it elsewhere. The stem piece could be attached to a different part of the same plant (**autografting**), to another plant of the same species (**homografting**), or to a plant of a different species (**heterografting**). The piece removed is called the **scion**, and that to which it is attached is known as the **stock**.

Budding

A type of **grafting** where a bud and its adjacent stem are the parts grafted.

Grafting

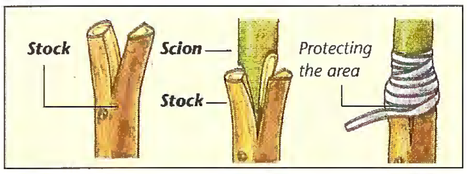
Stock
Scion
Stock
Protecting the area

Budding

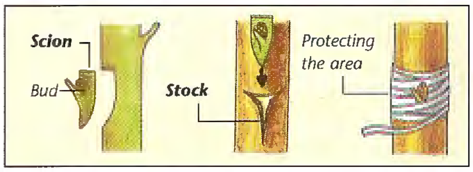
Scion
Bud
Stock
Protecting the area

---

* **Adventitious roots**, 245; **Asexual reproduction**, 321.

THE BODY STRUCTURE OF ANIMALS

Animals exist in a great variety of forms, from single-celled organisms to complex ones made of thousands of cells. The way they are **classified***, or divided into groups, depends to a large extent on how complex their bodies are. The two terms **higher animal** and **lower animal** are often used in this context. The higher an animal is, the more complex its internal organs are. In general, the distinguishing features of higher animals are **segmentation**, body cavities and some kind of skeleton.

Segmentation

The division of a body into separate areas, or **segments**, a step up in complexity from a simple undivided body. Generally, the more complex the animal, the less obvious its segments are. The most primitive form of segmentation is **metameric segmentation**, or **metamerism**. The segments (**metameres**) are very similar, if not identical. Each contains more or less identical parts of the main internal systems, which join up through the internal walls separating the segments. Such segmentation is found in most worms, for example. More complex segmentation is less obvious. In insects, for example, the body has three main parts – the head, **thorax** (upper body region) and **abdomen** (lower body region). Each of these is in fact a group of segments, called a **tagma** (pl. **tagmata**), but the segments are not divided by internal walls. They are simply visible as external markings.

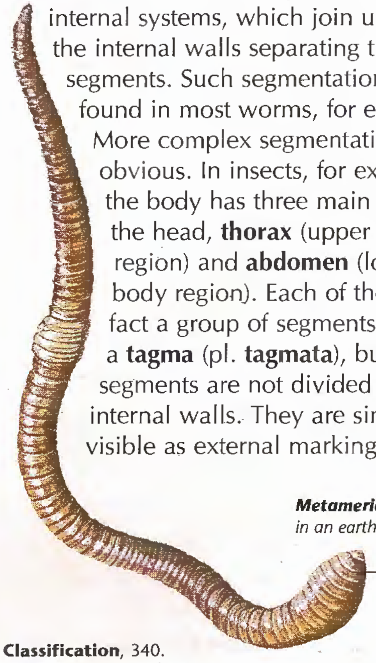

Metameric segmentation *in an earthworm*

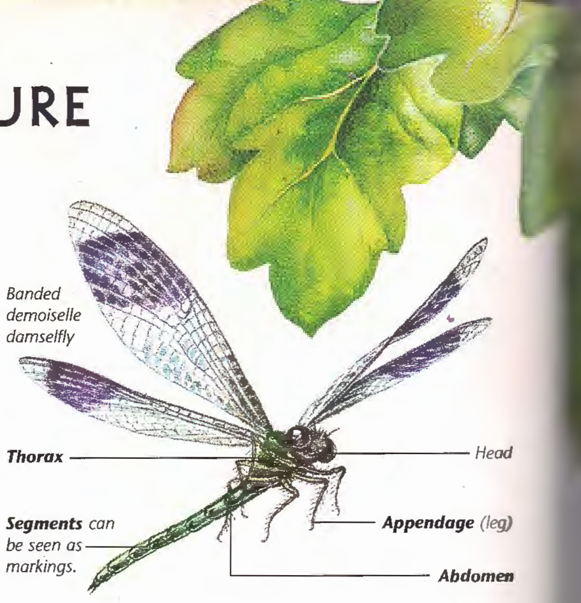

Banded demoiselle damselfly
Thorax
Head
Segments can be seen as markings.
Appendage (leg)
Abdomen

Appendage

A subordinate body part, i.e. one which projects from the body, such as an arm, leg, fin or wing.

Arrangement of the parts

Bilateral symmetry

An arrangement of body parts in which there is only one possible body division which will produce two mirror-image halves. It is typical of almost all freely-moving animals. The same state in flowers is called **zygomorphy** (e.g. in snapdragons).

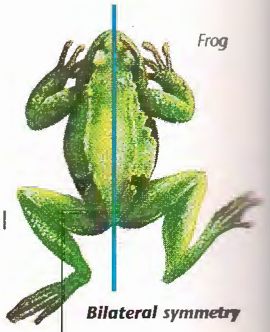

Only one division produces mirror-image halves.

Radial symmetry

A radiating arrangement of body parts around a central axis, e.g. in starfish. In such cases, there are two or more possible body divisions (sometimes in different planes) which will produce two mirror-image halves. The same state in flowers is called **actinomorphy** (e.g. in buttercups).

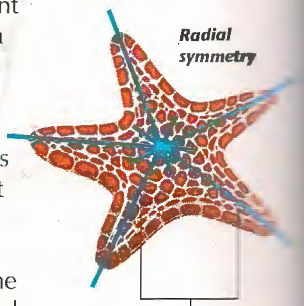

Several different divisions produce mirror-image hal

---

*Classification, 340.

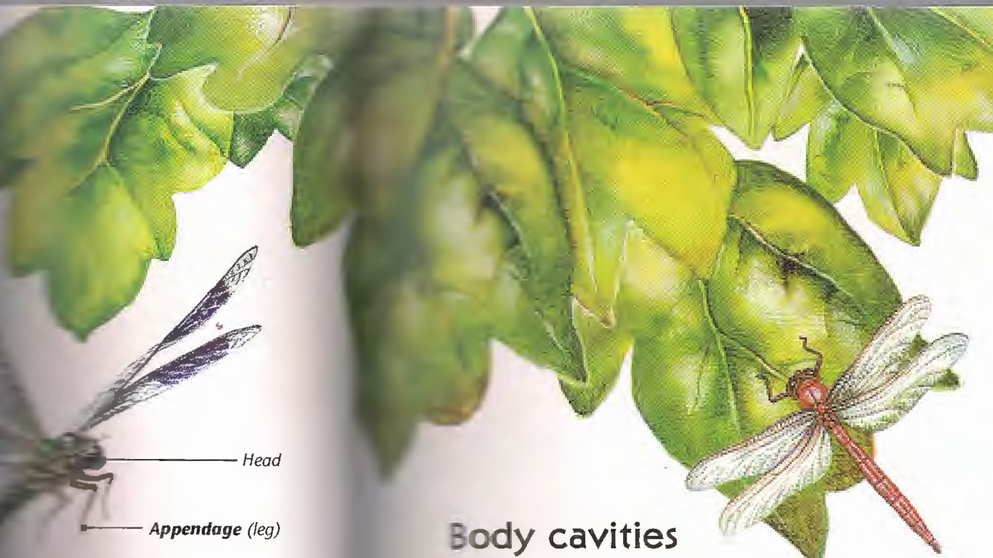
Head
Appendage (leg)
Abdomen

one which
an arm, leg,

the parts

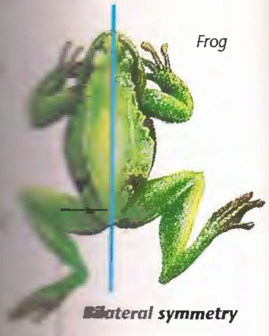

Bilateral symmetry

division produces
-image halves.

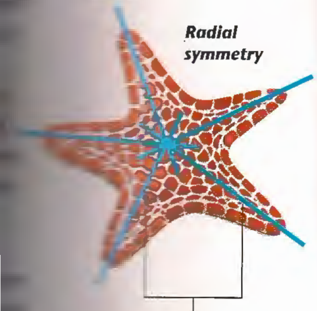

Radial symmetry

divisions
-image halves.

Body cavities

Almost all many-celled animals have a main fluid-filled body cavity, or **perivisceral cavity**, to cushion the body organs (very complex animals, e.g. humans, may have other smaller cavities as well). Its exact nature varies, but in most animals it is either a **coelom** or a **hemocoel**. In soft-bodied animals it is vital in movement, providing an incompressible "bag" for their muscles to work against. Such a system is called a **hydrostatic skeleton**.

Simplified cut-away of a peanut worm
(Not all body organs are shown.)

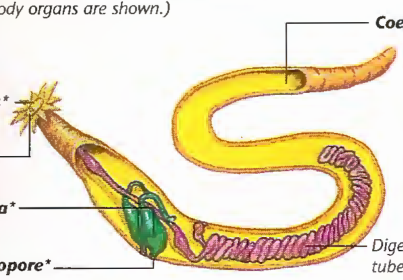
Coelom
Tentacles*
Mouth
Nephridia*
Nephridiopore*
Digestive tube

Coelom

The main body cavity (**perivisceral cavity**) of higher worms, **echinoderms***, e.g. starfish, and **vertebrates***, e.g. birds. It is fluid-filled to cushion the organs and is bounded by the **peritoneum**, a thin membrane which lines the body wall. In lower animals, e.g. many worms, the coelom assists in excretion. Their excretory organs, called **nephridia***, project into the coelom and remove fluid waste which has seeped into it. In higher animals, other more complex organs deal with these functions.

Simplified cut-away of a spider
(Not all body organs are shown.)

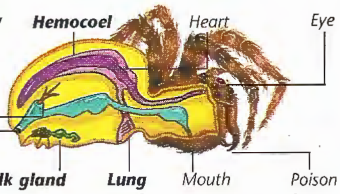
Hemocoel
Heart
Eye
Malpighian tubules*
Digestive tube
Silk gland
Lung book*
Mouth
Poison fang

Hemocoel

The fluid-filled main body cavity (**perivisceral cavity**) of **arthropods***, e.g. insects, and **molluscs***, e.g. snails. In molluscs, it is more of a spongy meshwork of tissue than a true cavity. Unlike a **coelom**, a hemocoel contains blood. It is an expanded part of the blood system, through which blood is circulated. In some animals, the hemocoel plays a part in excretion. In insects, for instance, water and fluid waste seep into it, and are then taken up by the **Malpighian tubules*** projecting into it.

Mantle cavity

A body cavity in shelled **molluscs***, e.g. snails. It lies between the **mantle** (a fold of skin lining the shell) and the rest of the body. Digestive and excretory wastes are passed into it, for removal from the body. In water-living molluscs, it also holds the **gills***; in land-living snails, it acts as a lung.

Simplified cut-away picture of a whelk
(Not all body organs are shown.)

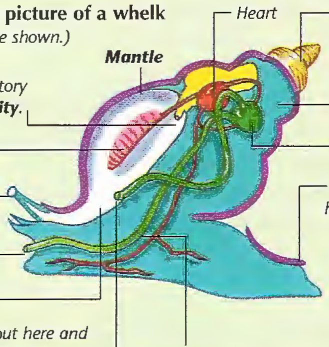
Heart
Shell
Mantle
Kidney discharges excretory waste into mantle cavity.
Spongy hemocoel meshwork
Gill*
Stomach
Mantle siphon (an inhalant siphon*)
Operculum – hard plate that covers shell opening when animal is inside.
Mouth
Mantle cavity
Digestive waste comes out here and leaves through a slit under the shell.
Digestive tube

* **Arthropods**, 341; **Echinoderms**, 341; **Gills**, 272; **Inhalant siphon**, 272 (**Siphon**); **Lung books**, **Malpighian tubules**, 273; **Molluscs**, 341; **Nephridiopore**, 273 (**Nephridia**); **Tentacles**, 274; **Vertebrates**, 341.

ANIMAL BODY COVERINGS

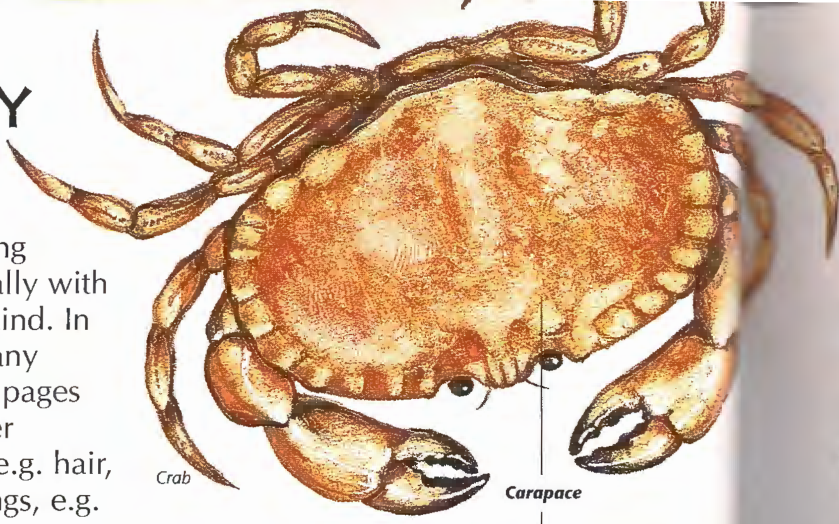

All animals have an enclosing outer layer, or "skin", normally with a further covering of some kind. In many cases, the skin has many layers, like human skin (see pages 310-311), and in most higher animals its covering is soft, e.g. hair, fur or feathers. Hard coverings, e.g. shells, are found in many lower animals and may form their only supporting framework, if they have no internal skeleton (**endoskeleton**). In such cases, the covering is called an **exoskeleton**. Some of the main body coverings are listed here.

Cuticle

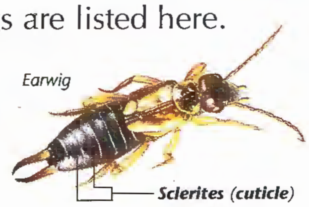

A non-living, waterproof, outer layer in many animals, secreted by the skin. In most **arthropods***, it hardens to form a supportive outer skeleton, or **exoskeleton**, e.g. the shells of crabs and the tough outer "coat" of insects. The term cuticle is in fact most often used to describe this insect "coat". It consists of a sugar-based substance (**chitin**) and a tough protein (**sclerotin**). It is often made up of **sclerites** – separate pieces joined by flexible, narrow areas. In other animals, e.g. earthworms, the cuticle remains a soft, waxy covering. (The term cuticle is sometimes used to mean the **stratum corneum*** in humans.)

Scales

There are two different types of scales. Those of bony fish, e.g. carp, are small, often bony plates lying within the skin. Those covering the limbs or whole bodies of many **reptiles*** (e.g. the legs of turtles) are thickened areas of skin.

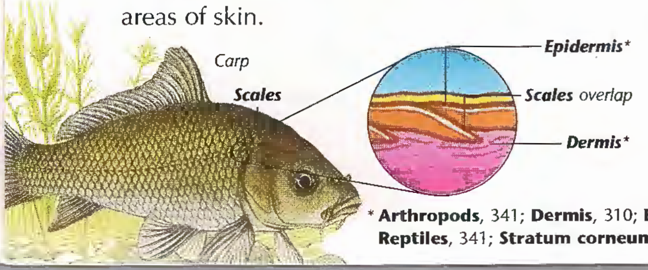

Carapace

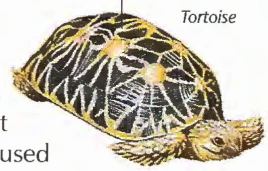

The shield-like shell of a crab, tortoise or turtle. In tortoises and turtles, it consists of bony plates fused together under a horny skin, but in crabs, it is a hardened **cuticle**.

Denticles or placoid scales

Sharp, backward-pointing plates, covering the bodies of cartilaginous fish, e.g. rays. They are similar to teeth, and stick out from the skin, unlike **scales**.

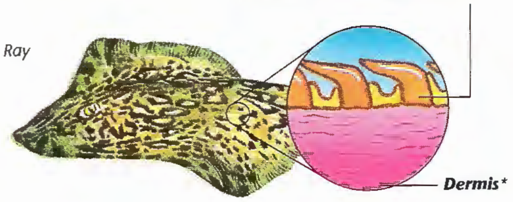

Elytra (sing. elytron)

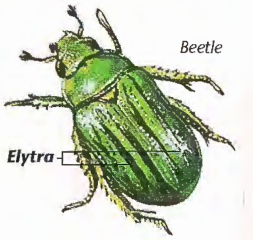

The front pair of wings of beetles and some bugs. They are designed to form a tough cover for the back pair of wings, used for flying.

Scute or scutum (pl. scuta)

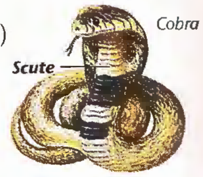

Any large, hard, external plate, especially those on the underside of a snake, used in movement.

* **Arthropods**, 341; **Dermis**, 310; **Epidermis**, 310; **Reptiles**, 341; **Stratum corneum**, 310.

Feathers

The insulating wa
body is made up o
known as its **plum**
light structure ma
substance called k
central shaft (or ra
filaments called ba
contour feathers
except the **down** fe
filaments called ba
hairs, feathers ha
attached to them,
which can fluff the
see **hair erector m**

Feathers

The insulating waterproof layer of a bird's body is made up of **feathers**, together known as its **plumage**. Each feather is a light structure made of a fibrous, horny substance called **keratin**. Each has a central **shaft** (or **rachis**) with thin filaments called **barbs**. The barbs of all **contour feathers**, i.e. all the feathers except the **down feathers**, have tiny filaments called **barbules**. Like body hairs, feathers have nerve endings attached to them, as well as muscles which can fluff them up to conserve heat (see **hair erector muscles**, page 311).

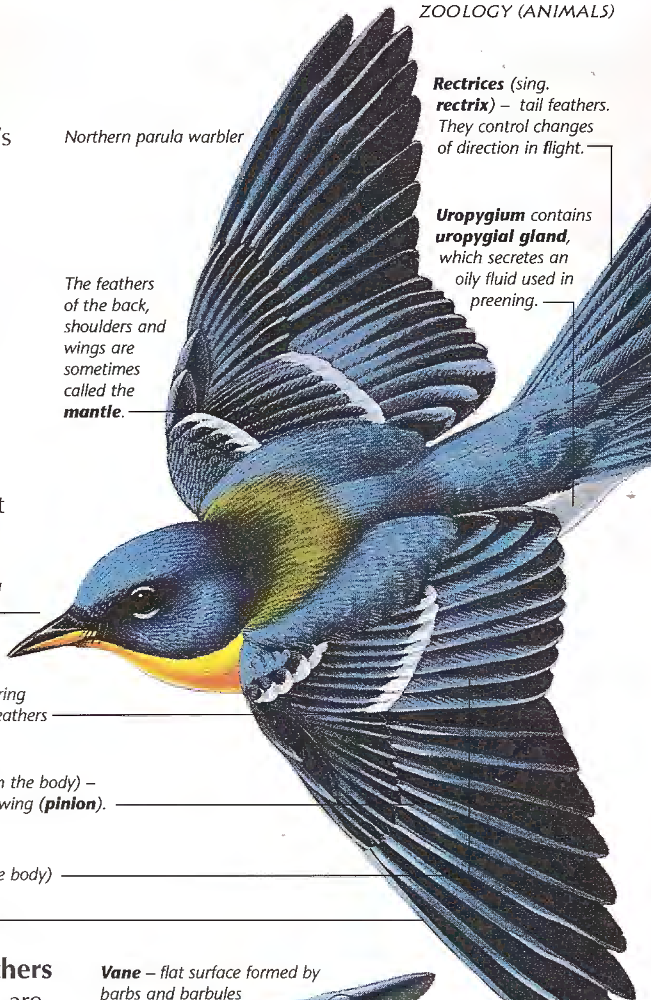
Northern parula warbler
Rectrices (sing. **rectrix**) – tail feathers. They control changes of direction in flight.
Uropygium contains **uropygial gland**, which secretes an oily fluid used in preening.
The feathers of the back, shoulders and wings are sometimes called the **mantle**.
Mandibles – upper and lower beak parts
Coverts – feathers covering bases of wing and tail feathers
Primaries (furthest from the body) – make up end section of wing (**pinion**).
Secondaries (nearer the body)

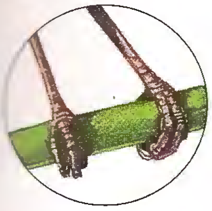

Birds' feet tend not to be **covered** in feathers, but are **protected** by small scales called **scutella** (sing. **scutellum**).

Remiges (sing. remix) or flight feathers

Those feathers of a bird's wings which are used in flight, consisting of the long, strong **primary feathers**, or **primaries**, and the shorter **secondary feathers**, or **secondaries**.

Down feathers or plumules

The fluffy, temporary feathers of all young birds, which have flexible **barbs**, but no true **barbules**. The adults of some types of bird keep some down feathers as an insulating layer close to the skin.

Feather follicles

Tiny pits in a bird's skin. Each one has a feather in it, just like a hair in a **hair follicle***. The cells at the base of the follicle grow up and out to form a feather, and then die away, leaving their hard, tough remains.

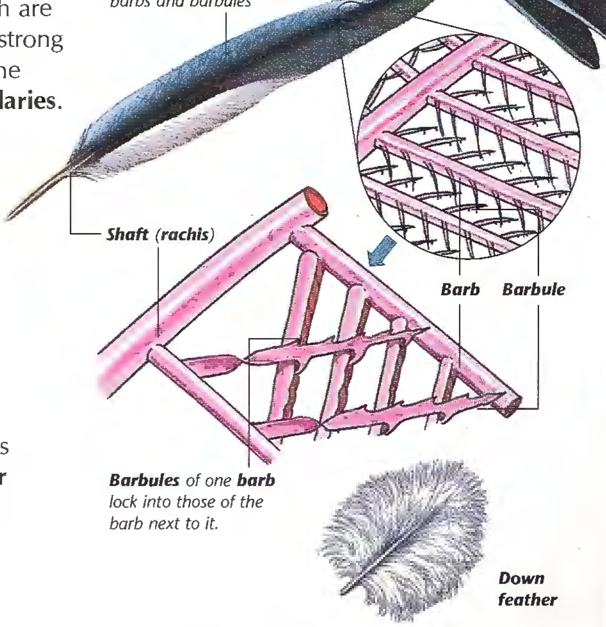
Vane – flat surface formed by barbs and barbules
Shaft (**rachis**)
Barb
Barbule
Barbules of one **barb** lock into those of the barb next to it.
Down feather

---

* **Hair follicles**, 311.

ANIMAL MOVEMENT

Most animals are capable of movement from place to place (**locomotion**) at least at some stage of their life (plants can only move individual parts – see **tropism**, page 251). The moving parts of animals vary greatly. Many animals have a system of bones and muscles similar to humans (see pages 278-283). Some of the parts used in animal movement are shown on these pages.

*All fish use **fins** to help them move from one place to another.*

Movement of simple animals

Pseudopodium (pl. pseudopodia)

An extension ("false foot") of the cell matter, or **cytoplasm***, of a single-celled organism. Such extensions are formed either in order for the organism to move or to enable it to engulf a food particle. The latter process is called **phagocytosis**.

Movement

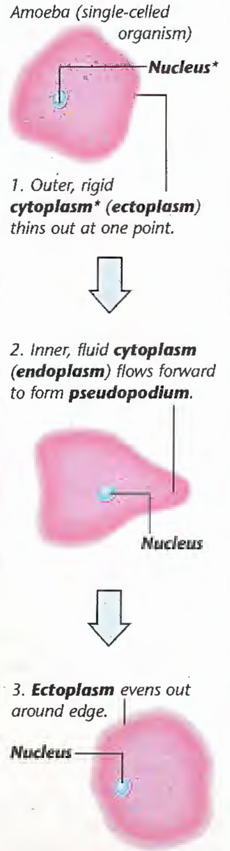
Amoeba (single-celled organism)
Nucleus*
Nucleus
Nucleus

*1. Outer, rigid **cytoplasm* (ectoplasm)** thins out at one point.*

*2. Inner, fluid **cytoplasm (endoplasm)** flows forward to form **pseudopodium**.*

*3. **Ectoplasm** evens out around edge.*

Organism has moved.

Phagocytosis

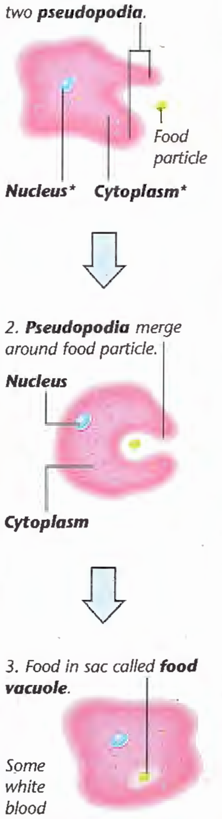
Food particle
Nucleus*
Cytoplasm*
Nucleus
Cytoplasm

*1. Amoeba forms two **pseudopodia**.*

*2. **Pseudopodia** merge around food particle.*

*3. Food in sac called **food vacuole**.*

*Some white blood cells use **phagocytosis** to engulf foreign bodies.*

Cilia (sing. cilium)

Tiny "hairs" on the outer body surfaces of many small organisms. They flick back and forth to produce movement. Cilia are also found lining the internal passages of more complex animals, e.g. human air passages (where they trap foreign particles).

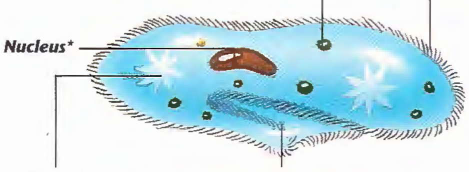
Paramecium (single-celled organism)
*Food in sac called **food vacuole***
Cilia
Nucleus*
Contractile vacuole*
***Cilia** inside channel called **oral groove** waft food particles inward.*

Flagella (sing. flagellum)

Any long, fine body threads, especially the one or more which project from the surface of many single-celled organisms. These lash backward and forward to produce movement. Organisms with flagella are **flagellate**.

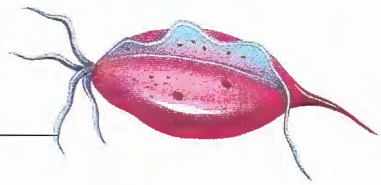
Trichomonas (single-celled organism)
Flagellum

Parapodia (sing. parapodium)

Paired projections from the sides of many aquatic worms, used to move them along. Each one ends in a bunch of bristles, or **chaetae** (sing. **chaeta**), which may also cover the body in some cases.

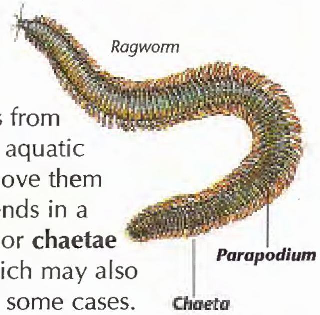
Ragworm
Parapodium
Chaeta

* **Contractile vacuoles**, 273; **Cytoplasm**, **Nucleus**, 238.

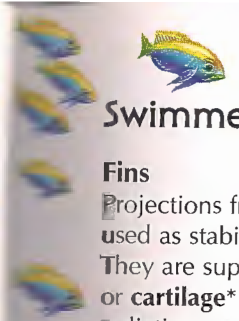

Swimmers

Fins

Projections from the
used as stabilizers ar
They are supported b
or **cartilage*** (depend
radiating out inside t
of fins, called **media**

Sheephead wrass
Caudal fin
***Anal fin**. If longer and extending under the belly, it is known as the **ventral fin**.*
Pectora
Pe

*In many fish, the **pelvic fins** are behind the **pectoral fins**.*

Median fins

The **fins** which run in
of the back and the b
eels, they form one lo
fin, but in most they a
dorsal, **caudal** (tail) a
The dorsal fin and the
of direction from side
helps to propel the fis

Paired fins

The **fins** of a fish whi
in **two** pairs: the **pect**
fins. They control mo

Swim bladder or ai

A long, air-filled pouc
The fish alters the am
bladder depending on
which it is swimming
the density of the fish
same as that of the
water, so it will not
sink if it stops
swimming.

***Median** or **medial** means "lying on the dividing line between the right and left sid*

***Dorsal** means "of the back o top surface".*

Cartilage, 281; **Digit**, 278 (
Habitat, 233; **Mammals**, 34

Swimmers

Fins

Projections from the body of a fish, which are used as stabilizers and to change direction. They are supported by **rays** – rods of bone or **cartilage*** (depending on the type of fish) radiating out inside them. Fish have two sets of fins, called **median** and **paired fins**.

Sheephead wrass

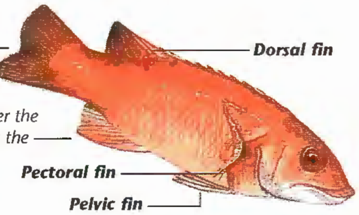
Caudal fin
Dorsal fin
Anal fin. If longer and extending under the belly, it is known as the **ventral fin**.
Pectoral fin
Pelvic fin

*In many fish, the **pelvic fins** are behind the **pectoral fins**.*

Median fins

The **fins** which run in a line down the center of the back and the belly. In some fish, e.g. eels, they form one long, continuous median fin, but in most they are divided into the **dorsal**, **caudal** (tail) and **anal** (or **ventral**) **fins**. The dorsal fin and the anal fin control changes of direction from side to side. The caudal fin helps to propel the fish through the water.

Paired fins

The **fins** of a fish which stick out from its sides in two pairs: the **pectoral fins** and the **pelvic fins**. They control movement up or down.

Swim bladder or air bladder

A long, air-filled pouch inside most bony fish. The fish alters the amount of air inside the bladder depending on the depth at which it is swimming. This keeps the density of the fish the same as that of the water, so it will not sink if it stops swimming.

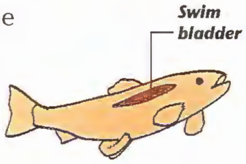
Swim bladder

***Median** or **medial** means "lying on the dividing line between the right and left sides".*

***Caudal** means "of the tail or hind part"; **caudate** means "having a tail".*

***Dorsal** means "of the back or top surface".*

***Ventral** means "of the front or lower surface".*

Flyers

Pectoralis muscles

Two large, paired chest muscles, found in many **mammals***, but especially highly developed in birds. Each wing has one **pectoralis major** and one **pectoralis minor**, attached at one end to the **keel**, a large extension of the breastbone. The muscles contract alternately to move the wings.

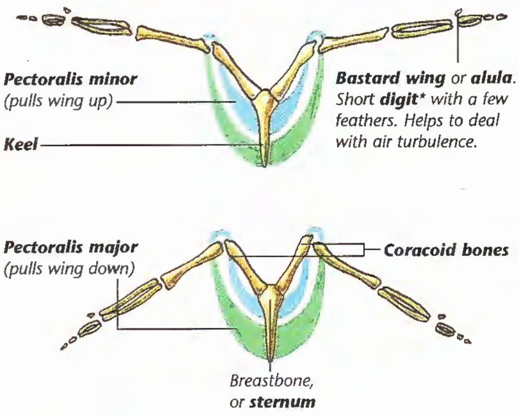
Pectoralis minor (pulls wing up)
Keel
Bastard wing or **alula**. Short **digit*** with a few feathers. Helps to deal with air turbulence.
Pectoralis major (pulls wing down)
Coracoid bones
Breastbone, or **sternum**

Walkers

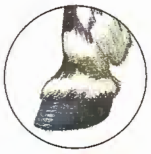

Unguligrade

Walking on hoofs at the tips of the toes, e.g. horses.

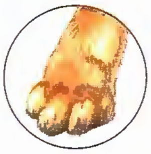

Digitigrade

Walking on the underside of the toes, e.g. dogs and cats.

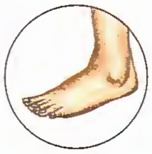

Plantigrade

Walking on the underside of the whole foot, e.g. humans.

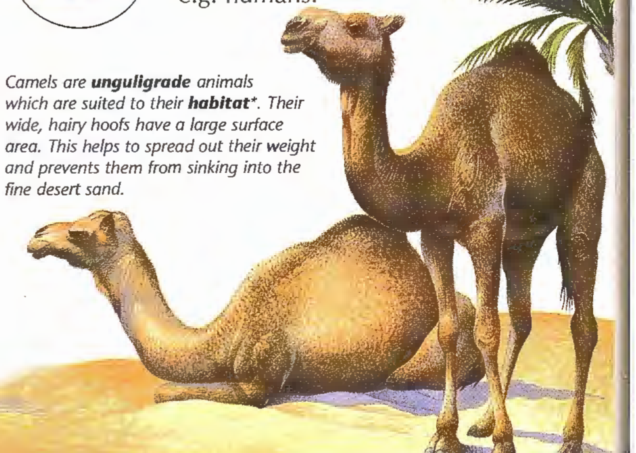

*Camels are **unguligrade** animals which are suited to their **habitat***. Their wide, hairy hoofs have a large surface area. This helps to spread out their weight and prevents them from sinking into the fine desert sand.*

---

* **Cartilage**, 281; **Digit**, 278 (**Phalanges**); **Habitat**, 233; **Mammals**, 341.

ANIMAL FEEDING

Different animals take in their food in many different ways, and with many different body parts. Some also have special internal mechanisms for dealing with the food (others have human-like **digestive systems** – see pages 294-295). Listed here are some of the main animal body parts involved in feeding and digestion.

Cnidoblasts or **thread cells**

Special cells found in large numbers on the **tentacles*** of **cnidarians***, e.g. sea anemones, used for seizing food. Each one contains a **nematocyst** – a long thread coiled inside a tiny sac. When a tentacle touches something, the threads shoot out to stick to it or sting it.

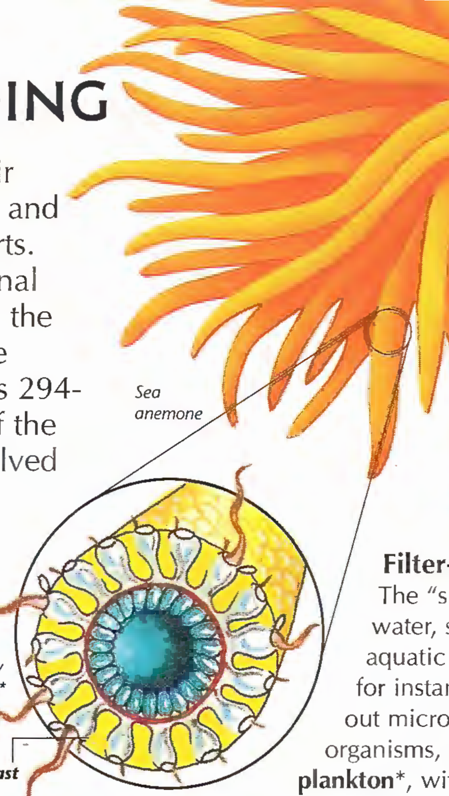

Diastema (pl. **diastemata**)

A gap between the front and back teeth of many plant-eaters. It is especially important in rodents, e.g. mice. They can draw their cheeks in through the gaps, so they do not swallow substances they may be gnawing.

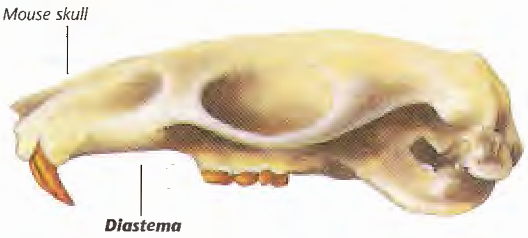

Carnassial teeth

The specially adapted last upper **premolar*** and first lower **molar*** of carnivorous mammals, used for shearing flesh.

Radula

The horny "tongue" of many **molluscs***, e.g. snails. It is covered by tiny teeth, which rasp off food.

Filter-feeding

The "sieving" of food from water, shown by many aquatic animals. Barnacles, for instance, sieve out microscopic organisms, or **plankton***, with bristly limbs called **cirri** (sing. **cirrus**).

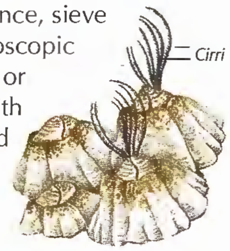

*Barnacles put out their **cirri** when they are covered with water.*

Some whales use frayed plates of horny **baleen**, or **whalebone**, hanging down from the top jaw. They sieve out small, shrimp-like animals called **krill**.

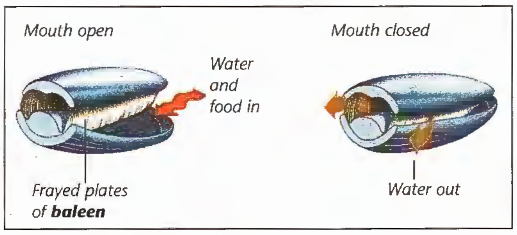

*A gray whale filtering sea water through its **baleen***

---

* **Cnidarians**, 341; **Molars**, 285; **Molluscs**, 341; **Plankton**, 342; **Premolars**, 285; **Tentacles**, 274.

Arthropod m

The mouths of **arthr**
made up of a numb
Depending on the a
these may look very
mouthparts, found i
mandibles, **maxillae**
and **labium**. The firs
many other arthropo
centipedes (some of
have two pairs of ma

*The **maxillae** of butterflies, moths and similar insects fit together to make a long sucking tube, or **proboscis**.*

Digestive stru

Crop

A thin-walled pouch,
esophagus*) in birds;
some worms, e.g. ear
insects, e.g. grasshop
crop before it goes in

Gizzard

A thick, muscular-wal
of the gullet (**esophag**
which have **crops**. The
instead food is groun
swallow pieces of gra
in other animals, the
gizzard, or hard, toot
to these walls, help to

***Cloaca**. Collecting chamber for body waste and reproductive bodies.*

* **Arthr**
Mamm

...ing

of food from

by many

Barnacles,

sieve

horny

down from

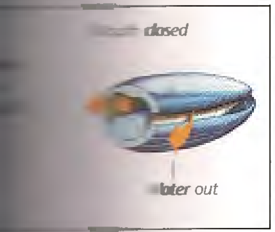

filtering sea
through
baleen

Arthropod mouthparts

The mouths of **arthropods***, e.g. insects, are made up of a number of different parts. Depending on the animal's feeding method, these may look very different. The basic mouthparts, found in all insects, are the **mandibles**, **maxillae** (sing. **maxilla**), **labrum** and **labium**. The first two are also found in many other arthropods, e.g. crabs and centipedes (some of these other arthropods have two pairs of maxillae).

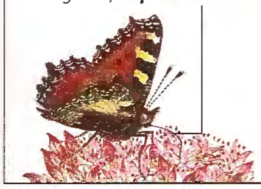

*The **maxillae** of butterflies, moths and similar insects fit together to make a long sucking tube, or **proboscis**.*

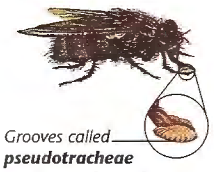

*The **labium** of houseflies is an extended pad-like sucking organ.*

*Grooves called **pseudotracheae** (sing. **pseudotrachea**)*

Typical arrangement of mouthparts (locust)

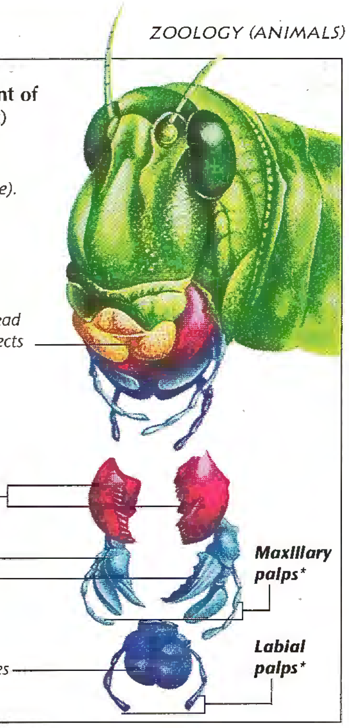

***Hypopharynx**. The tongue (not visible here). Used for sucking up liquids.*

***Labrum**. Upper lip. Hinged extension of head which covers and protects other mouthparts.*

***Mandibles**. Used for holding or biting food.*

***Maxillae**. In most cases manipulate food into mouth.*

***Labium**. Lower lip. In most cases manipulates food into mouth.*

Digestive structures

Crop

A thin-walled pouch, part of the gullet (**esophagus***) in birds; also a similar structure in some worms, e.g. earthworms, and some insects, e.g. grasshoppers. Food is stored in the crop before it goes into the **gizzard**.

Gizzard

A thick, muscular-walled pouch at the base of the gullet (**esophagus***) in those animals which have **crops**. These animals have no teeth, instead food is ground up in the gizzard. Birds swallow pieces of gravel to act as grindstones; in other animals, the muscular walls of the gizzard, or hard, tooth-like structures attached to these walls, help to grind up their food.

Pigeon

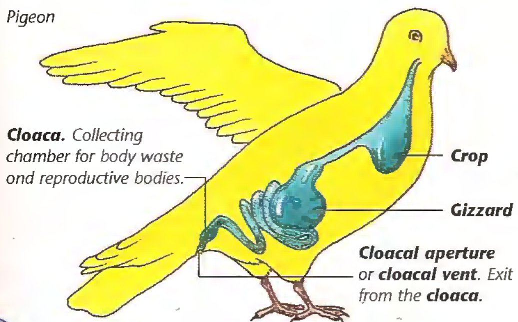

***Cloaca**. Collecting chamber for body waste ond reproductive bodies.*

***Cloacal aperture** or **cloacal vent**. Exit from the **cloaca**.*

Rumen

The large first chamber of the complex "stomach" of some plant-eating **mammals***, e.g. cows, into which food passes unchewed. It contains bacteria which can break down **cellulose***. Other animals pass this substance as waste, but these animals cannot afford to do this, as it makes up the bulk of their food (grass). Partially-digested food from the rumen is digested further in the second chamber, or **reticulum**, and then regurgitated to be chewed, when it is known as the **cud**. When swallowed again, it passes directly to the third and fourth chambers – the **omasum** and the **abomasum** (the true stomach) – for further processing.

*Animals that chew the cud (**ruminate**) are called **ruminants**.*

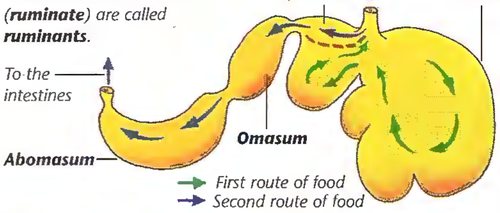

Cecum

Any blind-ended sac inside the body, especially one forming part of a digestive system. In many animals, e.g. rabbits, it is the site of an important stage of digestion (involving bacterial breakdown of **cellulose*** – see **rumen**). In others, e.g. humans (see **large intestine***), it has little use.

*__Arthropods__, 341; __Cellulose__, 238 (__Cell wall__); __Esophagus__, 294; __Mammals__, 341; __Palps__, 274.

ANIMAL RESPIRATION

The complex process of **respiration** consists of a number of stages (see introduction, page 298). Basically, oxygen is taken in and used by body cells in the breakdown of food, and carbon dioxide is expelled from the cells and the body. On these pages are some of the main animal respiratory organs.

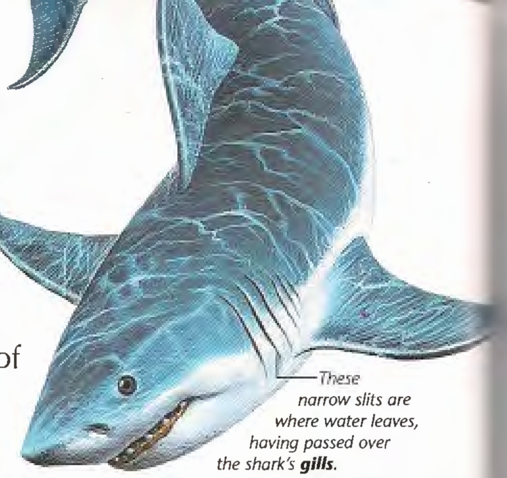

These narrow slits are where water leaves, having passed over the shark's **gills**.

Gills

Gills or **branchiae** (sing. **branchia**), are the breathing organs of most aquatic animals, containing many blood vessels. Oxygen is absorbed into the blood from the water passing over the gills. Carbon dioxide passes out the other way. There are two types of gills – **internal** and **external**.

Internal gills

Gills inside the body, found in various forms in all fish, most **mollusks***, e.g. limpets, and most **crustaceans** (a group of **arthropods*** which includes crabs). Most fish have four pairs of gills, with openings between them called **gill slits**. In more advanced fish, e.g. cod, they are covered by a flap called the **operculum**. In more primitive fish, e.g. sharks, they end in narrow openings in the skin on the side of the head. Each gill consists of a curved rod, the **gill bar** or **gill arch**, with many fine **gill filaments**, and even finer **gill lamellae** (sing. **lamella**) radiating from it. These all contain blood vessels.

Branchial means "of the gills."

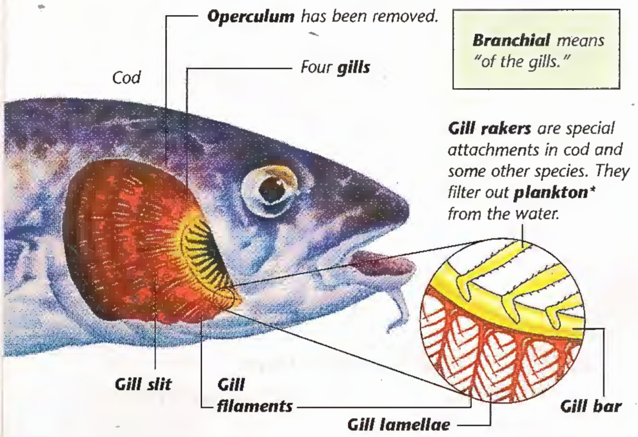
Operculum has been removed.
Cod
Four **gills**
Gill slit
Gill filaments
Gill lamellae
Gill bar

Gill rakers are special attachments in cod and some other species. They filter out **plankton*** from the water.

Breathing with gills

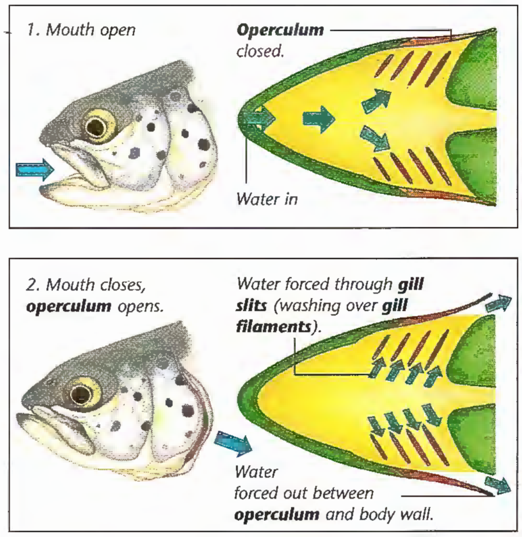
1. Mouth open
Operculum closed.
Water in
2. Mouth closes, **operculum** opens.
Water forced through **gill slits** (washing over **gill filaments**).
Water forced out between **operculum** and body wall.

External gills

Gills on the outside of the body, found in the young stages of most fish and **amphibians***, some older amphibians and the young aquatic stages of many insects (e.g. caddisfly **larvae*** and mayfly **nymphs***). Their exact form depends on the type of animal, but in many cases they are "frilly" outgrowths from the head, e.g. in young tadpoles.

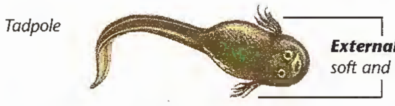

External gills are soft and "frilly".

Siphon

A tube carrying water to (**inhalant siphon**) or from (**exhalant siphon**) the **gills** of many lower aquatic animals, e.g. whelks (see picture, page 265). The exhalant siphon of **cephalopods** (**mollusks*** with **tentacles***), e.g. octopuses, is called the **hyponome***.

---

* **Amphibians**, 341; **Arthropods**, 341; **Hyponome**, 274; **Larva**, 277; **Mollusks**, 341; **Nymph**, 277; **Plankton**, 342; **Tentacles**, 274.

Other respirat

Spiracle
Any body opening thro
carbon dioxide are exc
blowhole). The term is
any of the tiny holes (a
sing. **stigma**) found in
e.g. insects.

Tracheae (sing. **trach**
Thin tubes leading in f
all **insects** (and the mo
They form an inner net
into narrower tubes cal
Oxygen from the air pa
walls to the body cells.
leaves via the same rou

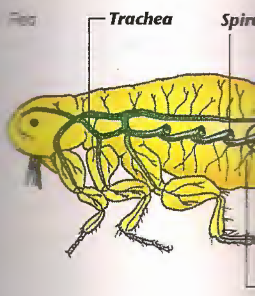
Trachea
Spir

Lung books or **book**
Paired breathing organs
which have four pairs)
advanced) spiders (whi
Each one has many blo
arranged like book pag
through slits (**spiracles**
book, and is absorbed
dioxide passes out the

Lung book

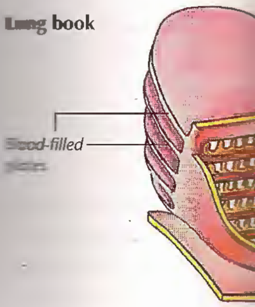
Blood-filled

* **Arthropods**, 341; **Cilia**, 268;
is also the word for the hum

Other respiratory organs

Spiracle

Any body opening through which oxygen and carbon dioxide are exchanged (e.g. a whale's blowhole). The term is used especially for any of the tiny holes (also called **stigmata**, sing. **stigma**) found in many **arthropods***, e.g. insects.

Tracheae (sing. trachea†)

Thin tubes leading in from the **spiracles** of all insects (and the most advanced spiders). They form an inner network, often branching into narrower tubes called **tracheoles**. Oxygen from the air passes through the tube walls to the body cells. Carbon dioxide leaves via the same route.

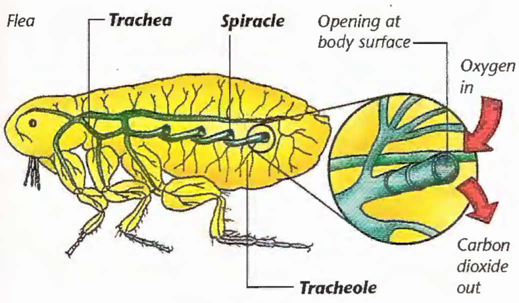
Flea
Trachea
Spiracle
Opening at body surface
Oxygen in
Carbon dioxide out
Tracheole

Lung books or book lungs

Paired breathing organs found in scorpions (which have four pairs) and some (less advanced) spiders (which have one or two). Each one has many blood-filled tissue plates, arranged like book pages. Oxygen comes in through slits (**spiracles**), one by each lung book, and is absorbed into the blood. Carbon dioxide passes out the same way.

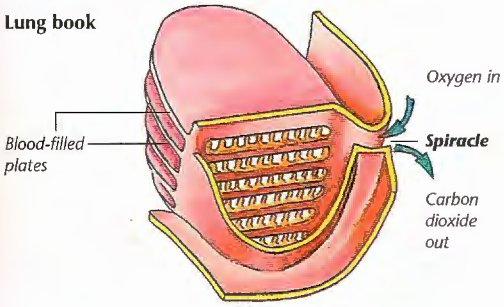
Lung book
Oxygen in
Blood-filled plates
Spiracle
Carbon dioxide out

Animal excretion

Excretion

The expulsion of waste fluid. It is vital to life as it gets rid of harmful substances. It is also vital to the maintenance of a balanced level of body fluids (see **homeostasis**, page 335).

Contractile vacuoles

Tiny sacs used for water-regulation in single-celled freshwater organisms. Excess water enters a vacuole via several canals arranged around it. At intervals, the vacuole expels its contents.

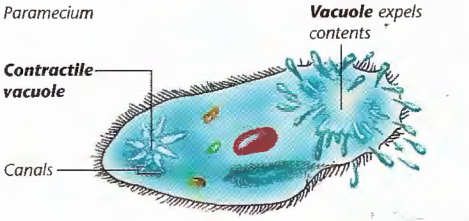
Paramecium
Vacuole expels contents
Contractile vacuole
Canals

Nephridia (sing. nephridium)

Waste-collecting tubes in many worms and the **larvae*** of many **mollusks***, e.g. mussels. In higher worms, they collect from the **coelom*** (see picture, page 265). Lower worms and mollusk larvae have more primitive **protonephridia**. The waste fluid enters these via hollow **flame cells** (**solenocytes**), which contain hair-like **cilia***. In both a nephridium and a protonephridium, the waste leaves through a tiny hole, or **nephridiopore**.

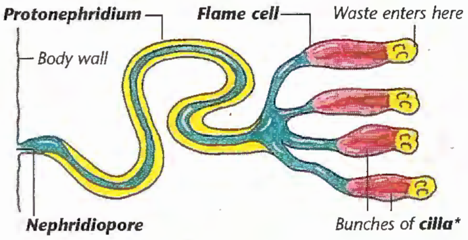
Protonephridium
Flame cell
Waste enters here
Body wall
Nephridiopore
Bunches of cilia*

Malpighian tubules

Long tubes found in many **arthropods***, e.g. insects. They carry dissolved waste from the main body cavity (**hemocoel***) into the rear of the gut. See picture, page 265.

---

* **Arthropods**, 341; **Cilia**, 268; **Coelom**, 265; **Hemocoel**, 265; **Larva**, 277; **Mollusks**, 341.

† *This is also the word for the human windpipe. See page 298.*

ANIMAL SENSES AND COMMUNICATION

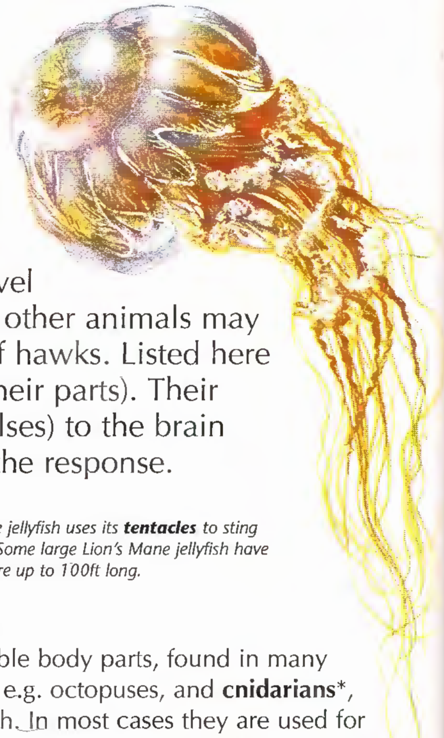

All animals show some **sensitivity** (**irritability**), i.e. response to external stimuli such as light and sound vibrations. Humans have a high overall level of sensory development, but individual senses in other animals may be even better developed, e.g. the acute vision of hawks. Listed here are some of the main animal sense organs (and their parts). Their responding parts send "messages" (nervous impulses) to the brain (or more primitive nerve center), which initiates the response.

*The Lion's Mane jellyfish uses its **tentacles** to sting and catch fish. Some large Lion's Mane jellyfish have tentacles that are up to 100ft long.*

Touch, smell and taste

Antennae (sing. antenna)

Whip-like, jointed sense organs on the heads of insects, centipedes and millipedes, and all **crustaceans** (a group of **arthropods*** which includes crabs and prawns). Insects, centipedes and millipedes have one pair, crustaceans have two. They respond to touch, temperature changes and chemicals (giving "smell" or "taste"). Some crustaceans also use them for swimming or to attach themselves to objects or other animals.

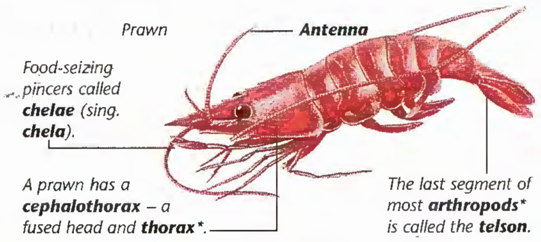

*Food-seizing pincers called **chelae** (sing. **chela**).*

*A prawn has a **cephalothorax** – a fused head and **thorax***.*

*The last segment of most **arthropods*** is called the **telson**.*

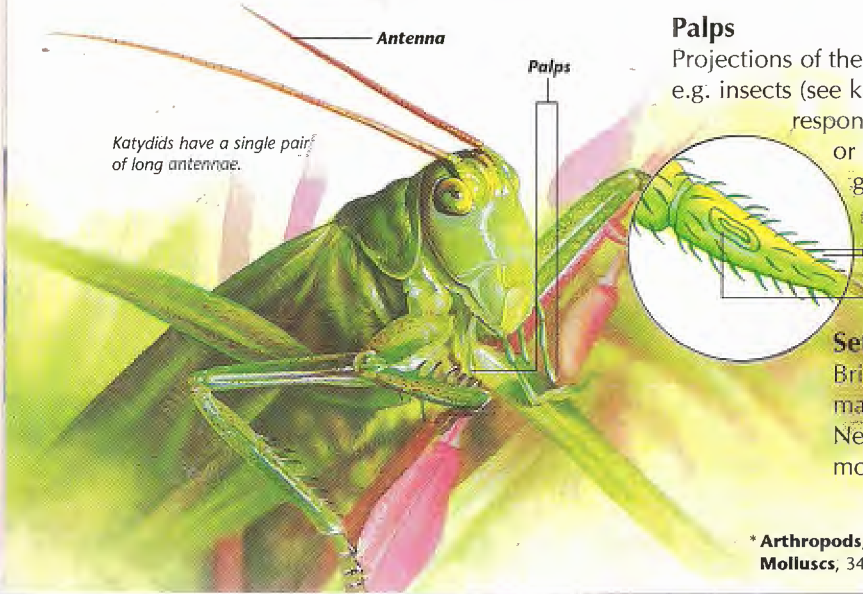

Katydids have a single pair of long antennae.

Tympanal organ *(see page 275)*

Tentacles

Long, flexible body parts, found in many **molluscs***, e.g. octopuses, and **cnidarians***, e.g. jellyfish. In most cases they are used for grasping food or feeling, though the longer of the two pairs found in land snails and slugs have eyes on the end.

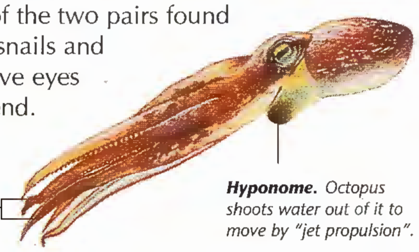

***Hyponome.** Octopus shoots water out of it to move by "jet propulsion".*

Vibrissae (sing. vibrissa) or whiskers

Stiff hairs standing out from the faces of many **mammals***, e.g. cats, around the nose. They are sensitive to touch.

Palps

Projections of the mouthparts of **arthropods***, e.g. insects (see katydid picture, left). They respond to chemicals (giving "smell" or "taste"). The term is also given to various touch-sensitive organs.

Setae (sing. seta)

Bristles produced by the skin of many **invertebrates***, e.g. insects. Nerves at their bases respond to movements of air or vibrations.

---

* **Arthropods**, **Cnidarians**, **Invertebrates**, **Mammals**, **Molluscs**, 341; **Thorax**, 264 (**Segmentation**).

Hearing and

Lateral lines

Two water-filled tube
of the body, just und
found in all fish and
spend most of their ti
toads. They enable th
water currents and p
use this information t

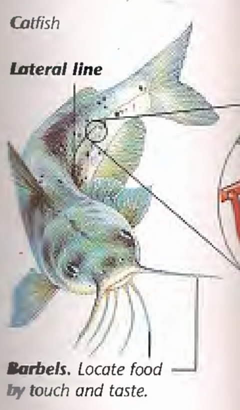

***Barbels.** Locate food by touch and taste.*

Tympanal organs o tympanum)

Sound detectors foun
legs in some insects,
the head in some lan
e.g. frogs. Each
is an air sac
covered by a thin
layer of tissue.
Sensitive fibers
in the organs
respond to
high-frequency
sound vibrations.

Statocysts

Tiny organs of balanc
invertebrates*, e.g. je
with tiny particles cal
sand grains. When th
grains move, stimulat
set off responses.

Haltères

Modified second pair
of wings on flies, whi
keep balance in flight.

* **Amphibians**, **Arthropods**,
Larynx, 298; **Pigments**, 255

Hearing and balance

Lateral lines

Two water-filled tubes lying along each side of the body, just under the skin. They are found in all fish and those **amphibians*** which spend most of their time in water, e.g. some toads. They enable the animals to detect water currents and pressure changes, and they use this information to find their way around.

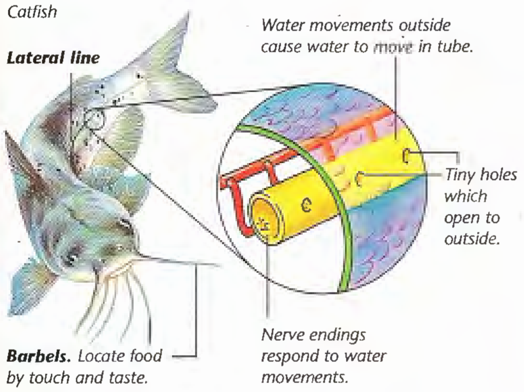
Catfish
Lateral line
Water movements outside cause water to move in tube.
Tiny holes which open to outside.
Barbels. Locate food by touch and taste.
Nerve endings respond to water movements.

Tympanal organs or tympani (sing. tympanum)

Sound detectors found on the lower body or legs in some insects, e.g. katydids, and on the head in some land **vertebrates***, e.g. frogs. Each is an air sac covered by a thin layer of tissue. Sensitive fibers in the organs respond to high-frequency sound vibrations.

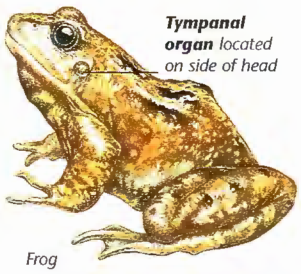
Tympanal organ located on side of head
Frog

Statocysts

Tiny organs of balance, found in many aquatic **invertebrates***, e.g. jellyfish. Each is a sac with tiny particles called **statoliths** inside, e.g. sand grains. When the animal moves, the grains move, stimulating sensitive cells which set off responses.

Haltères

Modified second pair of wings on flies, which keep balance in flight.

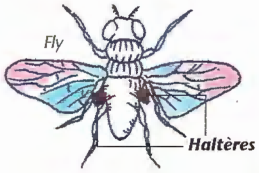
Fly
Haltères

Sight

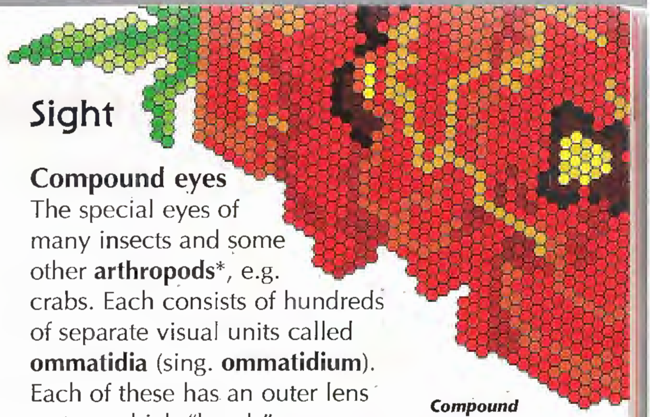

Compound eye view of a flower (mosaic image)

Compound eyes

The special eyes of many insects and some other **arthropods***, e.g. crabs. Each consists of hundreds of separate visual units called **ommatidia** (sing. **ommatidium**). Each of these has an outer lens system which "bends", or refracts, light onto a **rhabdom**, a transparent rod surrounded by light-responsive cells.

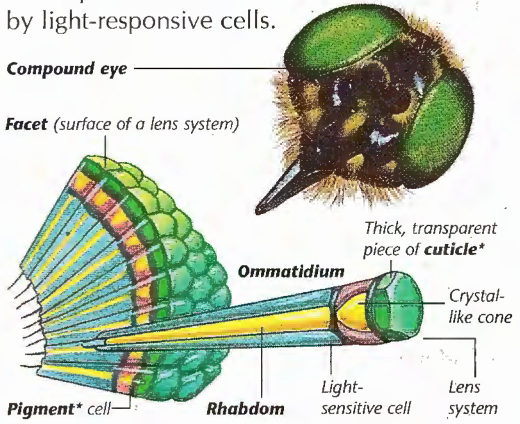
Compound eye
Facet (surface of a lens system)
Thick, transparent piece of cuticle*
Ommatidium
Crystal-like cone
Pigment* cell
Rhabdom
Light-sensitive cell
Lens system

After receiving information from all the ommatidia (each has a slightly different angle of vision so may record different light intensity or color), the brain assembles a complete **mosaic image**. This is enough for the animal's needs, but not as well defined as the image made by the human eye.

Communication

Pheromone

Any chemical made by an animal that causes responses in other members of the species, e.g. sexual attractants produced by many insects.

Stridulation

The rubbing together of body parts to make a shrill noise (often used to attract a mate). Katydids use their wing edges.

Syrinx (pl. syringes)

The vocal organ of birds, similar to the **larynx***, but found at the windpipe's base.

---

* **Amphibians**, **Arthropods**, 341; **Cuticle**, 266; **Invertebrates**, 341; **Larynx**, 298; **Pigments**, 255; **Vertebrates**, 341.

ANIMAL REPRODUCTION AND DEVELOPMENT

Chick hatching from *a* ***cleidoic egg***

Reproduction is the creation of new life. Most animals reproduce by **sexual reproduction***, the joining of a female sex cell, called an **ovum**, with a male sex cell, or **sperm**. Below are the main terms associated with the reproductive processes of animals.

Viviparous

A term describing animals such as humans, in which both the joining of the male and female sex cells (**fertilization**) and the development of the **embryo*** occur inside the female's body (the fertilization is **internal fertilization**), and the baby is born live.

New-born piglets suckling. Pigs are ***viviparous****.*

Oviparous

A term describing animals in which the development of the **embryo*** occurs in an **egg** which has been laid by the mother. In some cases, e.g. in birds, the male and female sex cells join inside the female's body (**internal fertilization**) and the egg already contains the embryo when laid. In other cases, e.g. in many fish, the many eggs each just contain an **ovum** (female sex cell) when laid, and the male then deposits **sperm** (male sex cells) over them (**external fertilization**).

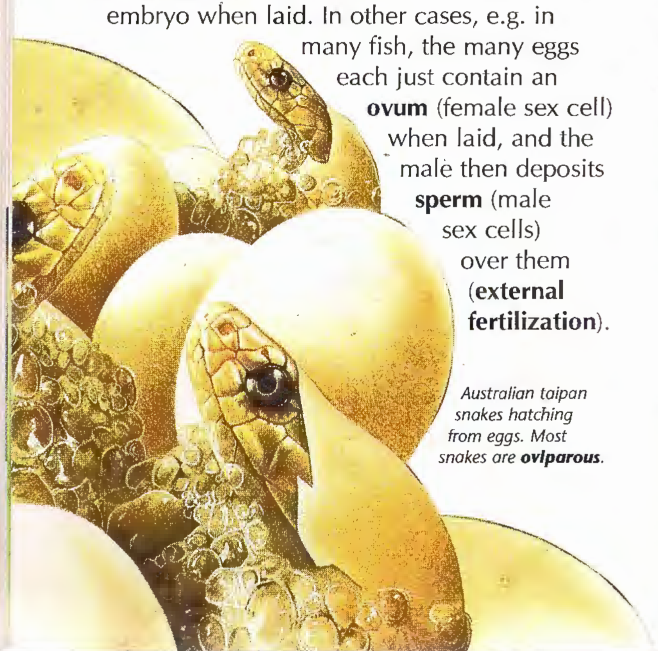

Australian taipan snakes hatching from eggs. Most snakes are ***oviparous****.*

Eggs

There are two main types of egg. **Cleidoic eggs** are produced by most egg-laying animals which live on land, e.g. birds and most **reptiles***, and also by a few aquatic animals, e.g. sharks. Such an egg largely isolates the **embryo*** from its surroundings, allowing only gases to pass through its tough shell (waste matter is stored). It contains enough food (**yolk**) for the complete development of the embryo, and the animal emerges as a tiny version of the adult. The other type of egg, produced by most aquatic animals, e.g. most fish, has a soft outer membrane, through which water and waste matter (as well as gases) can pass. The emerging young are not fully developed.

Cleidoic egg

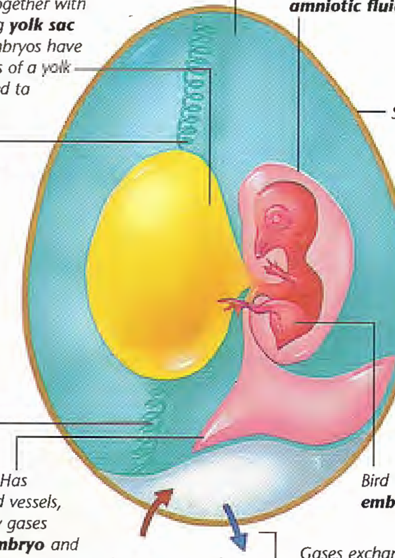

Albumen. *The "white" of the egg – providing protein and water.*

Amnion. *Thin layer of tissue, making* ***amniotic sac*** *which contains cushioning* ***amniotic fluid****.*

Yolk *(rich in phosphorus and fat). Gradually absorbed by* ***embryo****, together with surrounding* ***yolk sac*** *(human embryos have the remains of a yolk sac attached to them).*

Shell

Chalazae *(sing.* ***chalaza****). Twisted bands of* ***albumen****, holding* ***yolk*** *in place and acting as shock absorbers.*

Allantois. *Has many blood vessels, which carry gases between* ***embryo*** *and outside. Also present in early stages of human development.*

Bird ***embryo****

Oxygen in

Carbon dioxide out

Gases exchanged through shell and air space.

---

* **Embryo**, 320; **Reptiles**, 341; **Sexual reproduction**, 320.

Oviduct

Any tube in females
or **ova** (female sex c
outside. In some ani
are **fertilized** on the

Ovipositor

An organ extending
female insects, throu
many cases, it is lon
to pierce plant or an

Metamorphosis

The growth and de
involves intermedi
different from the
is a series of such
complete or partia
young form to the

Complete metamorphosi
and adult). The many insec
are called ***endopteryogot***

Male and female spotted
fritillary butterflies of south
Europe pair up and mate.
Female butterfly lays her eg
on small plant.

Larva *(pl.* ***larvae****) emerges*
egg. Has other names, e.g.
(beetles), ***maggot*** *(housefl*
caterpillar *(moths and*
butterflies). Sheds skin sever
times to allow for growth
(process called ***ecdysis****, or m*
common to all ***arthropods***

Final ***ecdysis*** *(see* ***larva****) re*
in ***pupa*** *(pl.* ***pupae****). Called*
chrysalis *in butterflies. Out*
skin is a hard protective cas
The cases surrounding moth
pupae have extra protection
in the form of a ***cocoon*** *of*
spun silk.

Hard case splits and mature
*adult (****imago****) emerges. The*
imago searches for a mate,
and the reproductive cycle is
then repeated.

Amphibians, **Arthropods**,

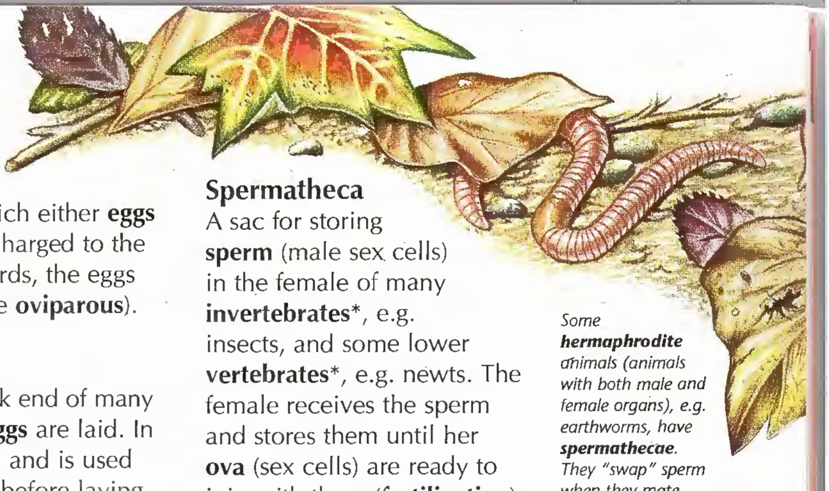

Oviduct

Any tube in females through which either **eggs** or **ova** (female sex cells) are discharged to the outside. In some animals, e.g. birds, the eggs are **fertilized** on the way out (see **oviparous**).

Ovipositor

An organ extending from the back end of many female insects, through which **eggs** are laid. In many cases, it is long and sharp, and is used to pierce plant or animal tissues before laying.

Spermatheca

A sac for storing **sperm** (male sex cells) in the female of many **invertebrates***, e.g. insects, and some lower **vertebrates***, e.g. newts. The female receives the sperm and stores them until her **ova** (sex cells) are ready to join with them (**fertilization**).

*Some **hermaphrodite** animals (animals with both male and female organs), e.g. earthworms, have **spermathecae**. They "swap" sperm when they mate.*

Metamorphosis

The growth and development of some animals involves intermediate forms which are very different from the adult form. Metamorphosis is a series of such changes, producing a complete or partial transformation from the young form to the adult.

All insects, most marine **invertebrates***, e.g. lobsters, and most **amphibians***, e.g. frogs, undergo some degree of metamorphosis (intermediate larval forms are common, e.g. legless **tadpoles** in frogs and toads). Below are examples of insect metamorphosis (two different kinds – **complete** and **incomplete metamorphosis**).

Complete metamorphosis *(two different forms between **egg** and adult). The many insects which undergo it, e.g. butterflies, are called **endopteryogotes**.*

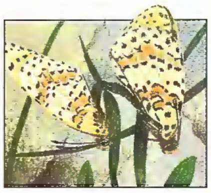

Male and female spotted fritillary butterflies of southern Europe pair up and mate. Female butterfly lays her eggs on small plant.

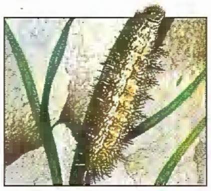

Larva *(pl. **larvae**) emerges from egg. Has other names, e.g. **grub** (beetles), **maggot** (houseflies), **caterpillar** (moths and butterflies). Sheds skin several times to allow for growth (process called **ecdysis**, or molting, common to all **arthropods***).*

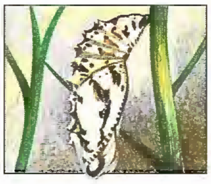

*Final **ecdysis** (see **larva**) results in **pupa** (pl. **pupae**). Called **chrysalis** in butterflies. Outer skin is a hard protective case. The cases surrounding moth pupae have extra protection in the form of a **cocoon** of spun silk.*

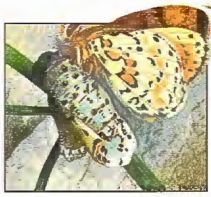

*Hard case splits and mature adult (**imago**) emerges. The imago searches for a mate, and the reproductive cycle is then repeated.*

Incomplete metamorphosis *(gradual development in stages). The insects which undergo it, e.g. locusts, are called **exopterygotes**.*

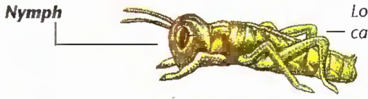

*Locust **nymphs** are called **hoppers**.*

***Nymph**. Emerges from egg. "Mini" version of adult insect, but resemblance only superficial, e.g. wings in very early stages of development or non-existent, many inner organs missing.*

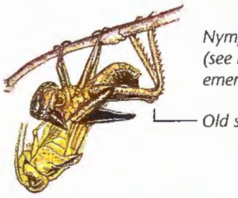

*Nymph undergoes several **ecdyses** (see **larva**), with some adult ports emerging each time.*

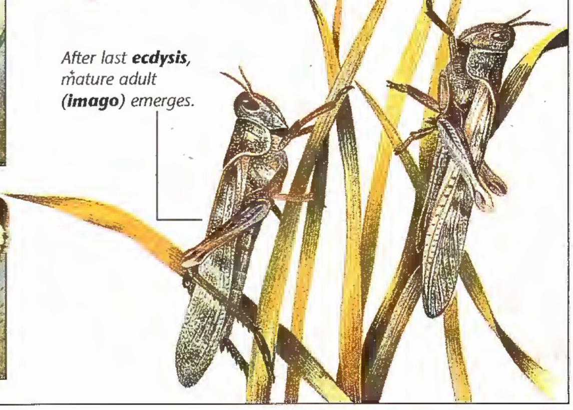

*After last **ecdysis**, mature adult (**imago**) emerges.*

* **Amphibians**, **Arthropods**, **Invertebrates**, **Vertebrates**, 341.

THE SKELETON

The human **skeleton** is a frame of over 200 bones which supports and protects the body organs (the **viscera**) and provides a solid base for the muscles to work against.

Cranium or skull

A case protecting the brain and facial organs. It is made of **cranial** and **facial bones**, fused at lines called **sutures**. The upper jaw, for instance, consists of two fused bones called **maxillae** (sing. **maxilla**).

Cranium

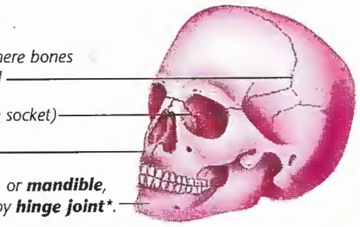
Suture where bones have fused
Orbit (eye socket)
Maxilla
Lower jaw, or **mandible**, attached by **hinge joint***.

Rib cage

A cage of bones forming the walls of the **thorax** or chest area. It is made up of 12 pairs of **ribs**, the **thoracic vertebrae** and the **sternum**. The ribs are joined to the sternum by bands of **cartilage*** called **costal cartilage**, but only the first seven pairs join it directly. The last five pairs are **false ribs**. The top three of these join the sternum indirectly – their costal cartilage joins that of the seventh pair. The bottom two pairs are **floating ribs**, only attached to the thoracic vertebrae at the back.

Rib cage

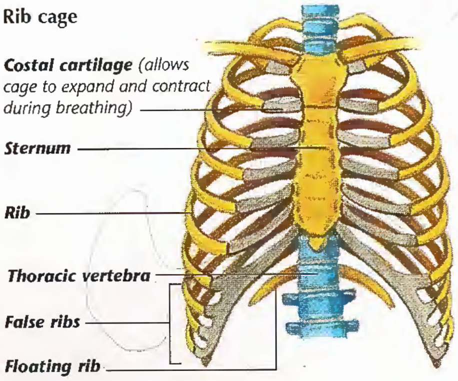
Costal cartilage (allows cage to expand and contract during breathing)
Sternum
Rib
Thoracic vertebra
False ribs
Floating rib

The bones of the skeleton

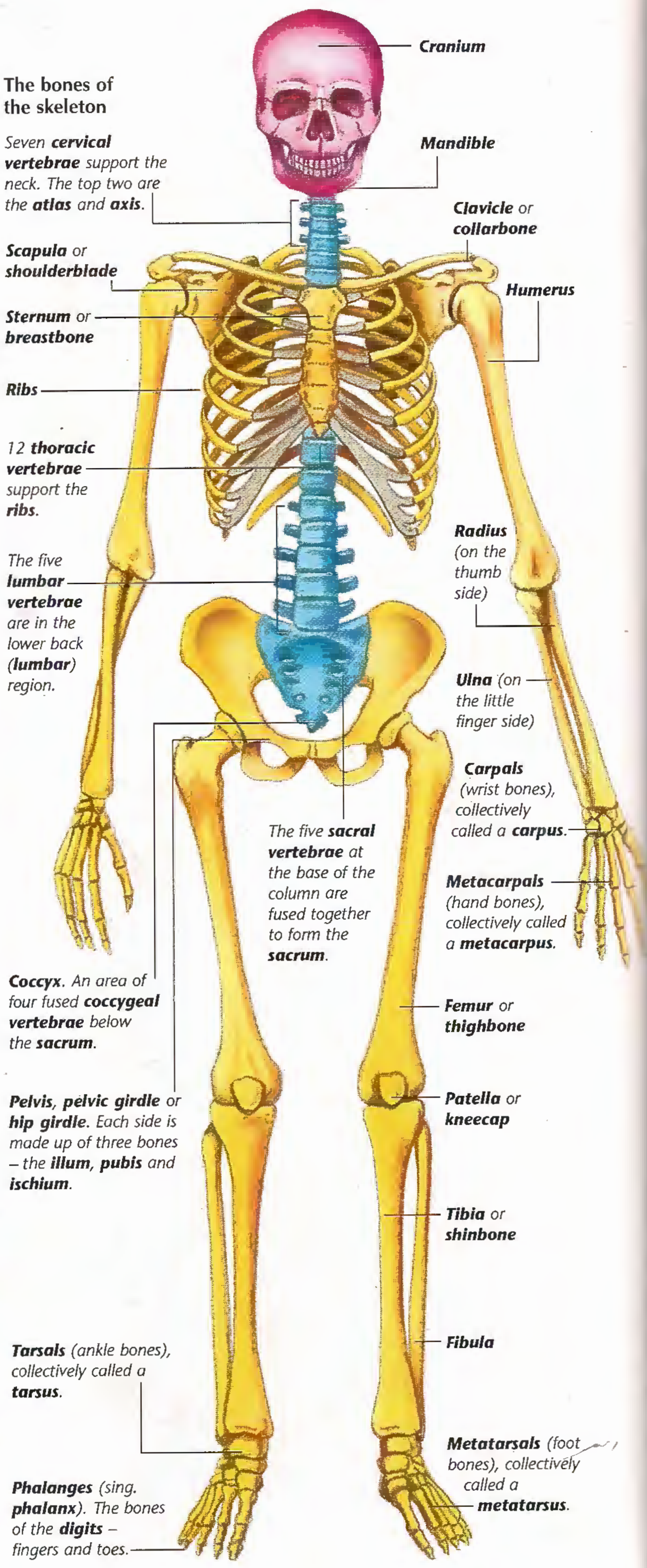
Cranium
Mandible
Clavicle or **collarbone**
Humerus
Scapula or **shoulderblade**
Sternum or **breastbone**
Ribs
Radius (on the thumb side)
Ulna (on the little finger side)
Carpals (wrist bones), collectively called a **carpus**.
Metacarpals (hand bones), collectively called a **metacarpus**.
Femur or **thighbone**
Patella or **kneecap**
Tibia or **shinbone**
Fibula
Metatarsals (foot bones), collectively called a **metatarsus**.
Tarsals (ankle bones), collectively called a **tarsus**.
Phalanges (sing. **phalanx**). The bones of the **digits** – fingers and toes.

*Seven **cervical vertebrae** support the neck. The top two are the **atlas** and **axis**.*

*12 **thoracic vertebrae** support the **ribs**.*

*The five **lumbar vertebrae** are in the lower back (**lumbar**) region.*

*The five **sacral vertebrae** at the base of the column are fused together to form the **sacrum**.*

***Coccyx**. An area of four fused **coccygeal vertebrae** below the **sacrum**.*

***Pelvis, pelvic girdle** or **hip girdle**. Each side is made up of three bones – the **ilium**, **pubis** and **ischium**.*

* **Cartilage**, 281; **Hinge joints**, 280.

Vertebral column

Also called the **spinal column**, **spine** or **backbone**. A flexible chain of 33 **vertebrae** which protects the **spinal cord***, supports the head and provides points of attachment for the **pelvis** and **rib cage**.

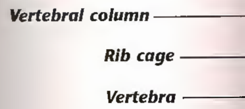
Vertebral column
Rib cage
Vertebra

Vertebrae (sing. vertebra)

The 33 bones of the **vertebral column**. A typical vertebra has a thick "chunk" (the **centrum** or **body**), vari
projections, or **proces**
(named below) and a central hole – the **vert**
foramen (pl. **foramina**
The foramina togethe
form the **neural**, **spina**
or **vertebral canal**, through which the **spinal cord*** runs.

Vertebral structu

The different verte
the skeleton on the
24 are movable an
disks of **cartilage***
fused together. The
structure describe
the top two, the **at**
(top vertebra) has a
skull which allows
axis has a "peg" (t
process) which fits
forms a **pivot joint**
allows the head to

* **Cartilage**, 281; **Ligaments**,

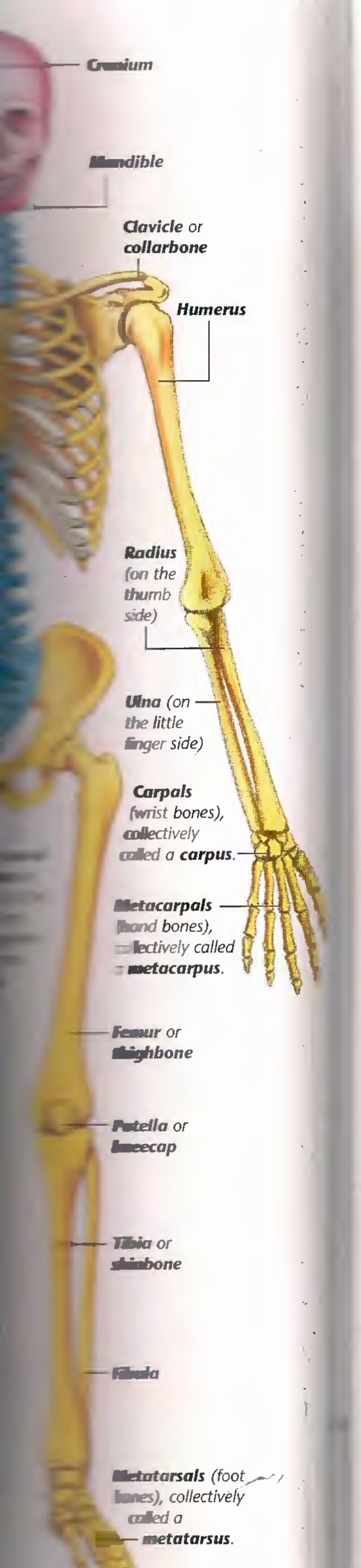
Cranium
Mandible
Clavicle or collarbone
Humerus
Radius (on the thumb side)
Ulna (on the little finger side)
Carpals (wrist bones), collectively called a carpus.
Metacarpals (hand bones), collectively called a metacarpus.
Femur or thighbone
Patella or kneecap
Tibia or shinbone
Fibula
Metatarsals (foot bones), collectively called a metatarsus.

Vertebral column

Also called the **spinal column**, **spine** or **backbone**. A flexible chain of 33 **vertebrae** which protects the **spinal cord***, supports the head and provides points of attachment for the **pelvis** and **rib cage**.

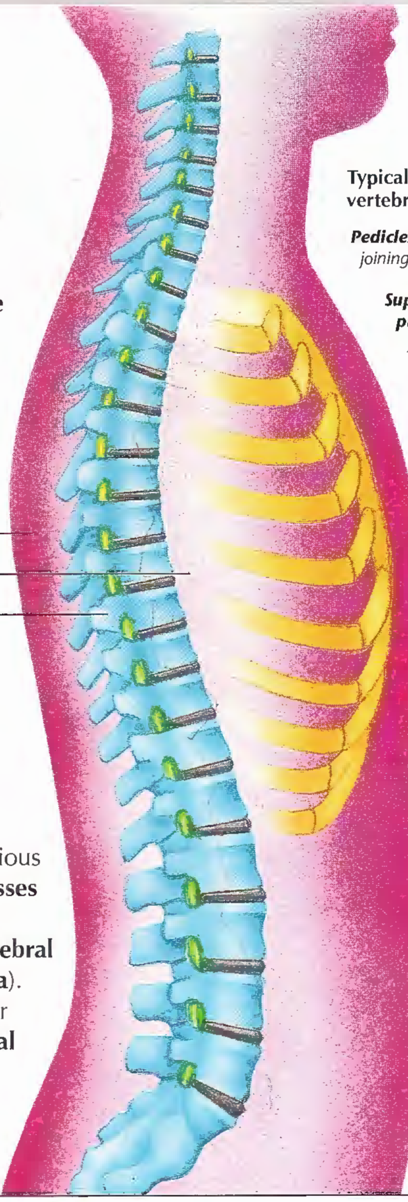
Vertebral column
Rib cage
Vertebra

Vertebrae (sing. vertebra)

The 33 bones of the **vertebral column**. A typical vertebra has a thick "chunk" (the **centrum** or **body**), various projections, or **processes** (named below) and a central hole – the **vertebral foramen** (pl. **foramina**). The foramina together form the **neural**, **spinal** or **vertebral canal**, through which the **spinal cord*** runs.

Typical vertebra (thoracic vertebra) seen from above

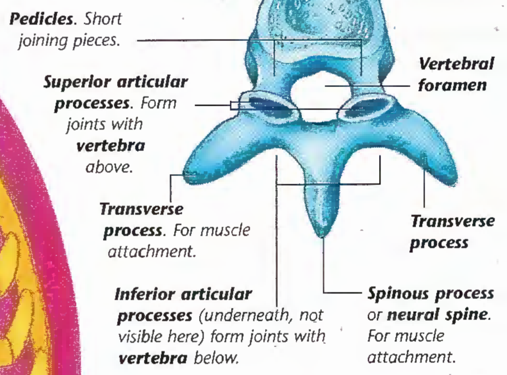
Centrum
Pedicles. Short joining pieces.
Vertebral foramen
Superior articular processes. Form joints with vertebra above.
Transverse process. For muscle attachment.
Transverse process
Inferior articular processes (underneath, not visible here) form joints with vertebra below.
Spinous process or neural spine. For muscle attachment.

Typical vertebra (thoracic vertebra) – seen from the side

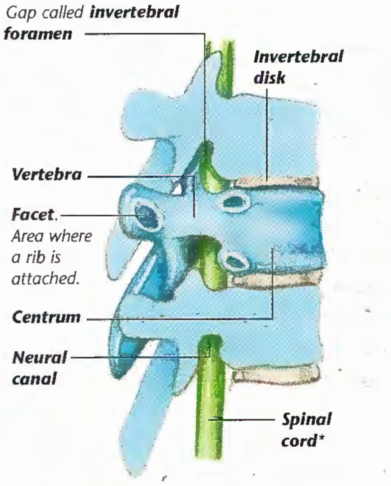
Gap called invertebral foramen
Invertebral disk
Vertebra
Facet. Area where a rib is attached.
Centrum
Neural canal
Spinal cord*

Vertebral structure

The different vertebrae are named around the skeleton on the opposite page. The top 24 are movable and linked by **invertebral disks** of **cartilage***. The bottom nine are fused together. They all have the typical structure described above right, except for the top two, the **atlas** and **axis**. The atlas (top vertebra) has a special joint with the **skull** which allows the head to nod. The axis has a "peg" (the **dens** or **odontoid process**) which fits into the atlas. This forms a **pivot joint**, a type of joint which allows the head to rotate.

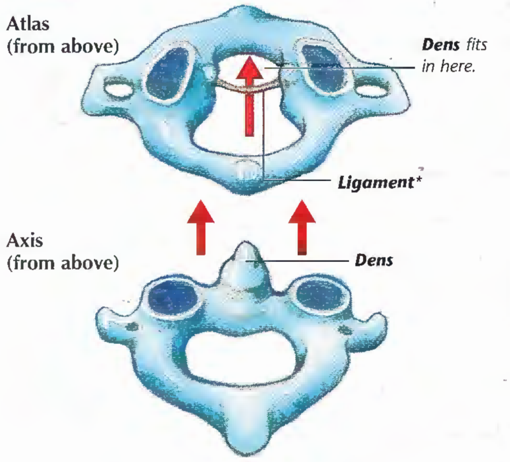
Atlas (from above)
Dens fits in here.
Ligament*
Axis (from above)
Dens

---

* **Cartilage**, 281; **Ligaments**, 280; **Spinal cord**, 302.

JOINTS AND BONE

The bones of the skeleton meet at many **joints**, or **articulations**. Some are **fixed joints**, allowing no movement, e.g. the **sutures*** of the skull. Most, however, are movable, and they give the body great flexibility. The most common are listed on this page.

Hinge joints

Joints (e.g. the knee joint) which work like any hinge. That is, the movable part (bone) can only move in one plane, i.e. in either of two opposing directions.

Hinge joint (knee joint)

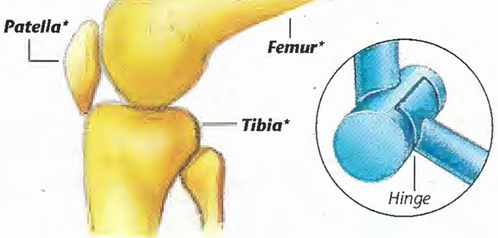

Gliding joints

Also called **sliding** or **plane joints**. Joints in which one or more flat surfaces glide over each other, e.g. those between the **carpals***. They are more flexible than **hinge joints**.

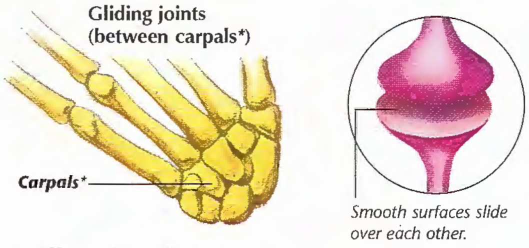
Gliding joints (between carpals*)
Carpals*
Smooth surfaces slide over each other.

Ball-and-socket joints

The most flexible joints (e.g. the hip joint). The movable bone has a rounded end which fits into a socket in the fixed bone. The movable bone can swivel, or move in many directions.

Ball-and-socket joint (hip joint)

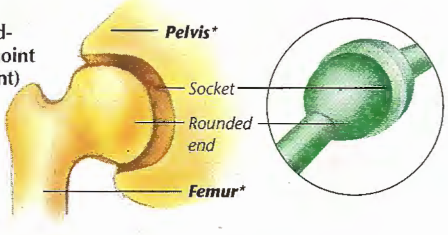

Connective tissue

There are many different types of **connective tissue** in the body. They all protect and connect cells or organs and have a basis of non-living material (the **matrix**) in which living cells are scattered. The difference between them lies in the nature of this material. The various types of tissue found at a joint, including **bone** itself, are all types of connective tissue. They all contain protein fibers and are either tough (containing **collagen** fibers) or elastic (containing **elastin** fibers).

The ease with which different types of tissue grow and repair depends largely on the amount of blood they contain. **Periosteum** is **vascular** (has a blood supply) and repairs itself quickly. **Cartilage** is **avascular** (has no blood supply) so takes longer to repair.

Periosteum

A thin layer of elastic connective tissue. It surrounds all bones, except at the joints (where **cartilage** takes over), and contains **osteoblasts** – cells which make new bone cells, needed for growth and repair.

Ligaments

Bands of connective tissue which hold together the bones of joints (and also hold many organs in place). Most are tough, though some are elastic, e.g. between **vertebrae***.

Synovial sac or synovial capsule

A cushioning "bag" of lubricating fluid (**synovial fluid**), with an outer skin (**synovial membrane**) of elastic connective tissue. Most movable joints, e.g. the knee, have such a sac lying between the bones. They are known as **synovial joints**.

---

* **Carpals**, **Femur**, **Patella**, **Pelvis**, 278; **Sutures**, 278 (**Cranium**); **Tibia**, 278; **Vertebrae**, 279.

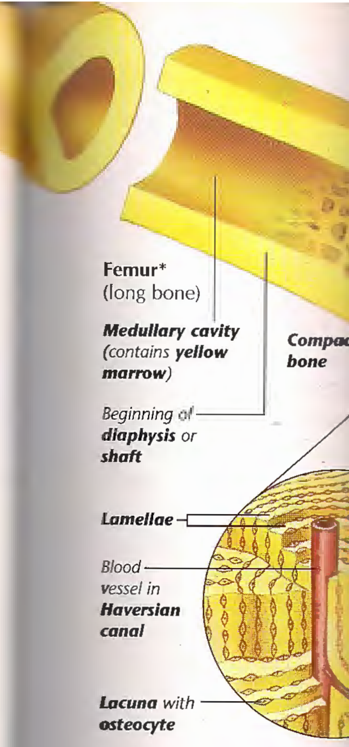
Femur* (long bone)
Medullary cavity (contains yellow marrow)
Compac bone
Beginning of diaphysis or shaft
Lamellae
Blood vessel in Haversian canal
Lacuna with osteocyte

Bone or osseous tiss

A type of tough connec
and resilient by large de
calcium compounds. T
osteocytes, are held in t
lacuna) within this non-
two types of bone: **spo**

Tendons or sinews

Bands of tough connect
joining muscles to bone
continuation of the men
the muscle, together wi
membranes of its bundl

Cartilage or gristle

A tough connective tiss
joints (**cartilaginous joi**
main cushion between
(e.g. **vertebrae***). In joir
synovial sacs, it covers
the bones and is called
cartilage. The end of th
the outer parts of the ea
of cartilage. Young skel
made of cartilage, thou
slowly turn to bone as
build up (a process call
ossification or **osteogen**

---

* **Femur**, **Fibula**, **Patella**, **Stern**
Vertebrae, 279.

Joints in
glide over
carpals*.
joints.

surfaces slide
other.

joint). The
which fits
movable
directions.

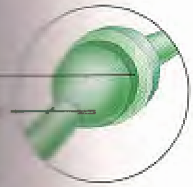

tissue. It
joints
contains
bone

hold
also
are

synovial
tissue.
have
They

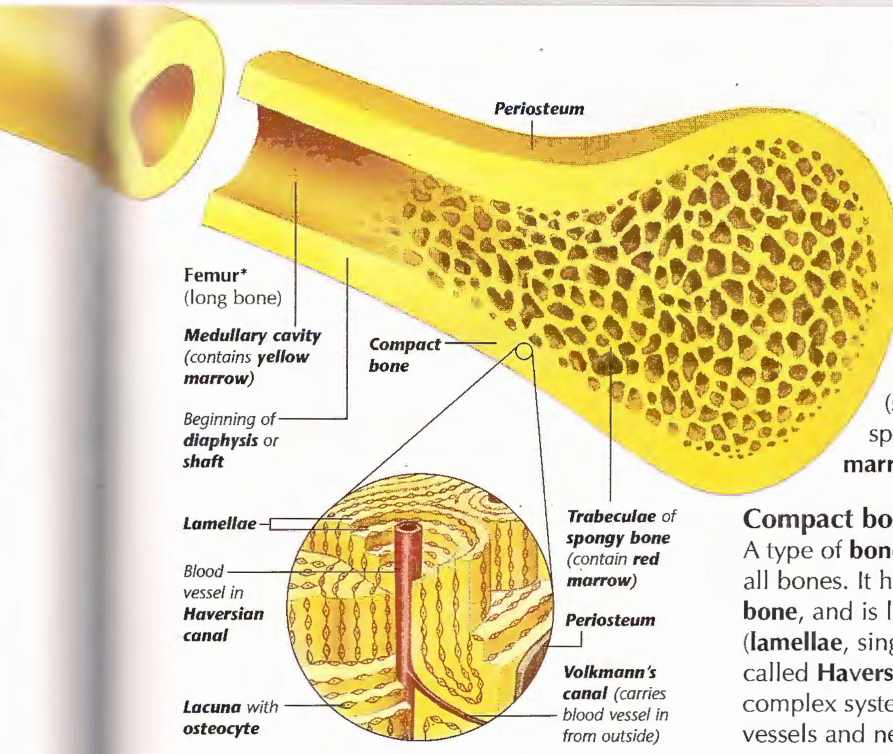
Periosteum
Femur* (long bone)
Medullary cavity (contains yellow marrow)
Compact bone
Beginning of diaphysis or shaft
Lamellae
Blood vessel in Haversian canal
Lacuna with osteocyte
Trabeculae of spongy bone (contain red marrow)
Periosteum
Volkmann's canal (carries blood vessel in from outside)

Spongy bone

A type of **bone** that is found in short and/or flat bones, e.g. the **sternum***, and fills the ends of long bones, e.g. the **femur***. It consists of a criss-cross network of flat plates called **trabeculae** (sing. **trabecula**), with many large spaces between them, filled by **red marrow** (see **bone marrow**).

Compact bone

A type of **bone** that forms the outer layer of all bones. It has far fewer spaces than **spongy bone**, and is laid in concentric layers (**lamellae**, sing. **lamella**) around channels called **Haversian canals**. These link with a complex system of tiny canals carrying blood vessels and nerves to the osteocytes.

Bone or osseous tissue

A type of tough connective tissue, made hard and resilient by large deposits of phosphorus and calcium compounds. The living bone cells, or **osteocytes**, are held in tiny spaces (**lacunae**, sing. **lacuna**) within this non-living material. There are two types of bone: **spongy** and **compact bone**.

Bone marrow

Two types of soft tissue. **Red marrow**, found in **spongy bone** (see **bone**), is where all new red (and some white) blood cells are made. **Yellow marrow** is a fat store, found in hollow areas (**medullary** or **marrow cavities**) in long bone **shafts**.

Tendons or sinews

Bands of tough connective tissue joining muscles to bones. Each is a continuation of the membrane around the muscle, together with the outer membranes of its bundles of fibers.

Cartilage or gristle

A tough connective tissue. In some joints (**cartilaginous joints**) it is the main cushion between the bones (e.g. **vertebrae***). In joints with **synovial sacs**, it covers the ends of the bones and is called **articular cartilage**. The end of the nose and the outer parts of the ears are made of cartilage. Young skeletons are made of cartilage, though these slowly turn to bone as minerals build up (a process called **ossification** or **osteogenesis**).

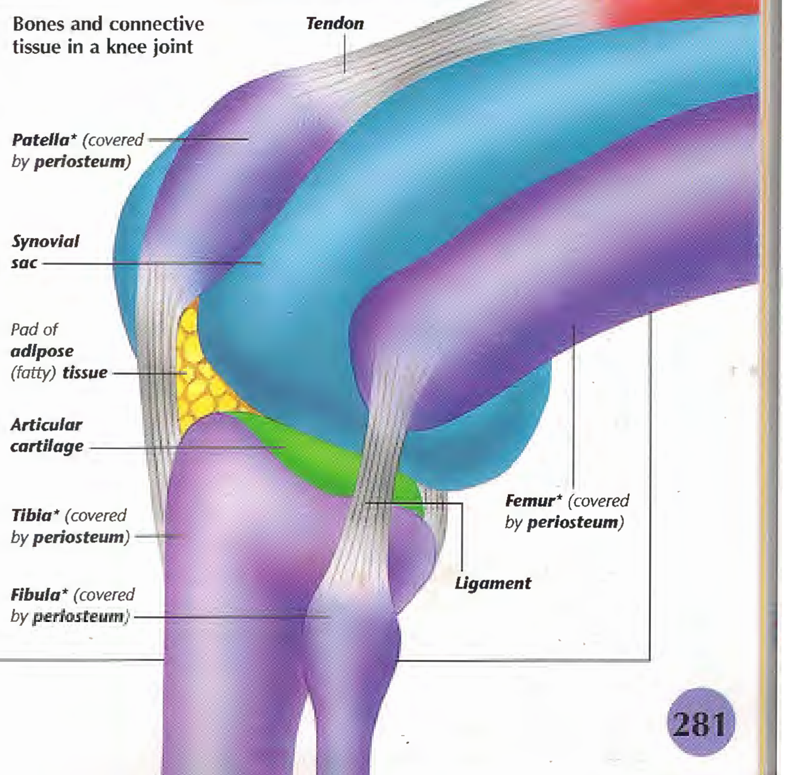
Bones and connective tissue in a knee joint

Tendon
Patella* (covered by periosteum)
Synovial sac
Pad of adipose (fatty) tissue
Articular cartilage
Tibia* (covered by periosteum)
Fibula* (covered by periosteum)
Femur* (covered by periosteum)
Ligament

---

* **Femur, Fibula, Patella, Sternum, Tibia**, 278; **Vertebrae**, 279.

MUSCLES

Muscles are areas of special elastic tissue (**muscle**) found all over the body. They may be either **voluntary muscles** (able to be controlled by conscious action) or **involuntary muscles** (not under conscious control). The main types of muscles are listed on this page.

Antagonistic pairs or opposing pairs

The pairs into which almost all muscles are arranged. The members of each pair produce opposite effects. In any given movement, the muscle which contracts to cause the movement is the **agonist** or **prime mover**. The one which relaxes at the same time is the **antagonist**.

Types of muscles

Skeletal muscles

All the muscles attached to the bones of the skeleton, which contract together or in sequence to move all the body parts. They are all **voluntary muscles** (see introduction) and are made of **striated muscle** tissue (see opposite). Some are named according to their position, shape or size, others are named after the movement they cause, e.g. **flexors** cause **flexion** (the bending of a limb at a joint), **extensors** straighten a limb.

Cardiac muscle

The muscle which makes up almost all of the wall of the heart. It is an **involuntary muscle** (see introduction) and is made of **cardiac muscle** tissue (see opposite).

Visceral muscles

The muscles in the walls of many internal organs, e.g. the intestines, and blood vessels. They are all **involuntary muscles** (see introduction) and consist of **smooth muscle** tissue (see opposite).

Example of an antagonistic pair (biceps and triceps)

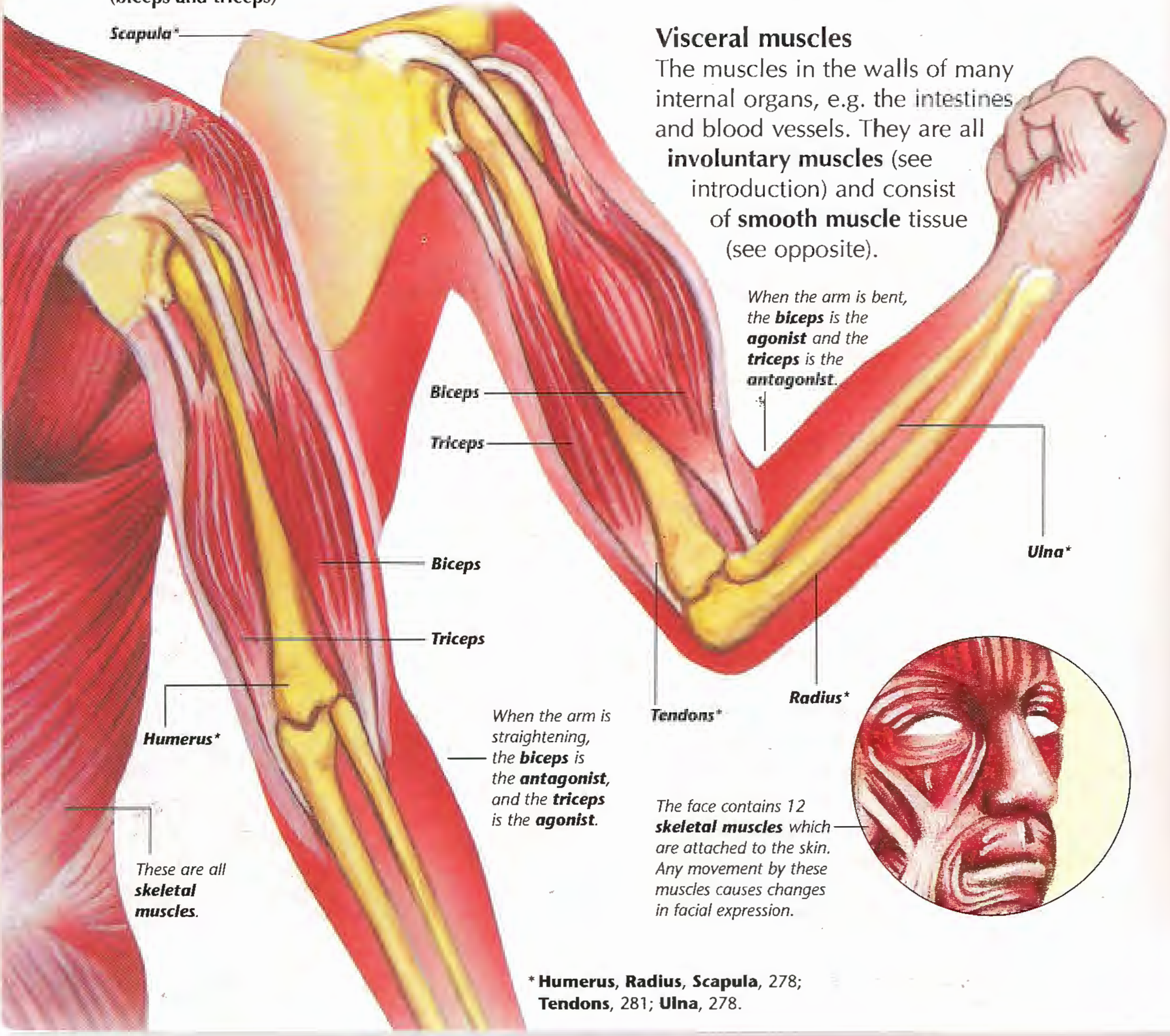

*When the arm is bent, the **biceps** is the **agonist** and the **triceps** is the **antagonist**.*

*When the arm is straightening, the **biceps** is the **antagonist**, and the **triceps** is the **agonist**.*

*These are all **skeletal muscles**.*

*The face contains 12 **skeletal muscles** which are attached to the skin. Any movement by these muscles causes changes in facial expression.*

* **Humerus**, **Radius**, **Scapula**, 278; **Tendons**, 281; **Ulna**, 278.

The structure

The different muscles
up of different kinds o
(groups of cells of diffe
tissue has many blood
food matter to be brok
energy, and nerves, wh
stimulate the muscles

Striated or striped m

The type of muscle tiss
up **skeletal muscles**. It
cells called **muscle fib**
in bundles called **fasci**
fiber has a striped (**stria**
appearance under a
microscope and is mad
of many smaller cylind
called **fibrils** or **myofib**
are the parts that contra
is stimulated by a nerve
interlocking **filaments**,
different types of protei
and **myosin** (thicker fila
slide past each other as

Relaxed)
Contracting)

Nervous stimu

Most muscles are stimu
impulses from nerves
body. For more about t

Muscle spindle

A group of **muscle fibe**
which has the end fibe
cell (**sensory neuron***)
The end fibers are part
dendron*). When the
are stimulated to send
"telling" it about the ne
The brain can then dete
needed for any further

Axon, 304; **Dendron**, 304 (**Den**
Motor neurons, **Sensory neur**

bones of
gether or in
parts. They are
duction) and
ue (see
rding to their
re named after
flexors cause
a joint),

most all of the
tary muscle
cardiac

The structure of muscle tissue

The different muscles in the body are made up of different kinds of muscle tissue (groups of cells of different types). The tissue has many blood vessels, bringing food matter to be broken down for energy, and nerves, which stimulate the muscles to act.

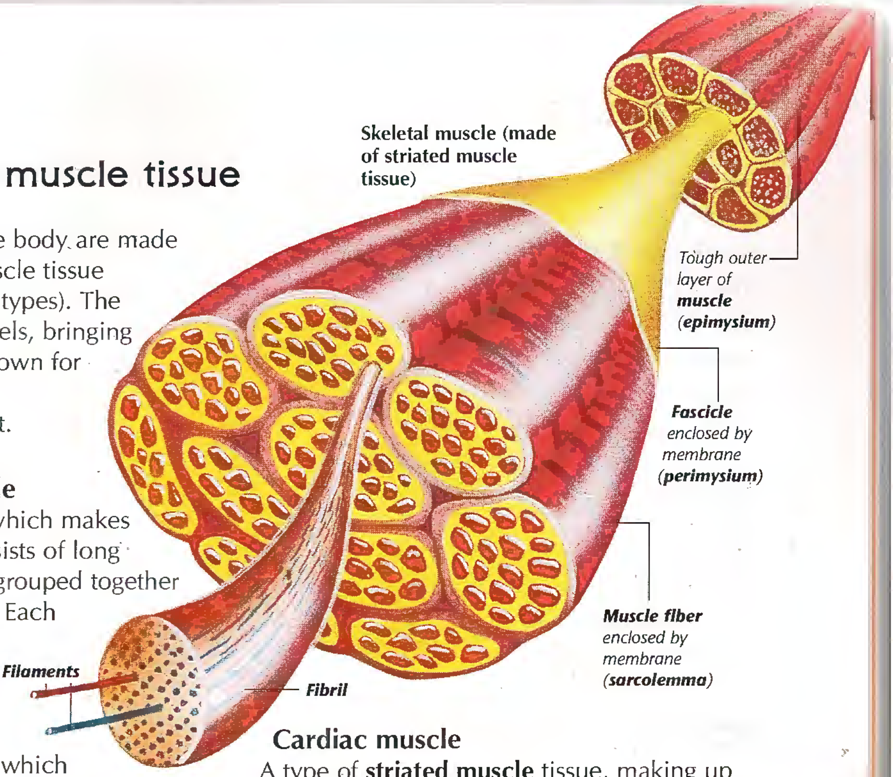
Skeletal muscle (made of striated muscle tissue)
Tough outer layer of **muscle** **(epimysium)**
Fascicle enclosed by membrane **(perimysium)**
Muscle fiber enclosed by membrane **(sarcolemma)**
Filaments
Fibril

Striated or striped muscle

The type of muscle tissue which makes up **skeletal muscles**. It consists of long cells called **muscle fibers**, grouped together in bundles called **fascicles**. Each fiber has a striped (**striated**) appearance under a microscope and is made of many smaller cylinders, called **fibrils** or **myofibrils**, which are the parts that contract when a fiber is stimulated by a nerve. The fibrils consist of interlocking **filaments**, or **myofilaments**, of two different types of protein – **actin** (thin filaments) and **myosin** (thicker filaments). These filaments slide past each other as a muscle contracts.

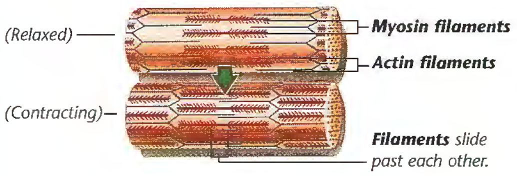
(Relaxed)
(Contracting)
Myosin filaments
Actin filaments
Filaments slide past each other.

Cardiac muscle

A type of **striated muscle** tissue, making up the **cardiac muscle** of the heart. Its constant rhythmical contractions are caused by stimulations from areas of the tissue which produce electrical impulses. Any nervous impulses just increase or decrease this heart rate.

Smooth muscle or visceral muscle

The type of muscle tissue which makes up the **visceral muscles**. It consists of short, spindle-shaped cells. The way it contracts is not yet fully understood, but it contains **actin** and **myosin**, and is stimulated by nerves.

Nervous stimulation

Most muscles are stimulated to move by impulses from nerves running through the body. For more about this, see pages 308-309.

Muscle spindle

A group of **muscle fibers** (see **striated muscle**) which has the end fibers of a sensory nerve cell (**sensory neuron***) wrapped around it. The end fibers are part of one main fiber (**dendron***). When the muscle stretches they are stimulated to send impulses to the brain, "telling" it about the new state of tension. The brain can then determine the changes needed for any further action.

Motor end-plate

The point where the end fibers of an "instruction-carrying" nerve cell (**motor neuron***) meet a **muscle fiber** (see **striated muscle**). The end fibers are branches from one main fiber (**axon***). This carries nervous impulses which make the muscle contract. Each impulse is duplicated and sent down each end branch, hence the whole muscle receives a multiplication of each impulse.

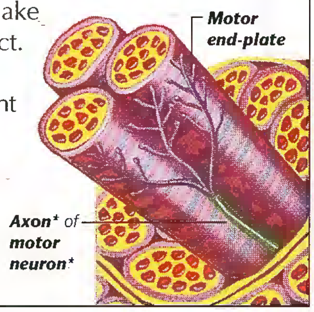
Motor end-plate
Axon* of **motor neuron***

---

* **Axon**, 304; **Dendron**, 304 (**Dendrites**); **Motor neurons**, **Sensory neurons**, 305.

TEETH

The **teeth** or **dentes** (sing. **dens**) help to prepare food for digestion by cutting and grinding it up. Each tooth is set into the jaw, which has a soft tissue covering called **gum** (**gingiva**). During their lives, humans have two sets of teeth (**dentitions**) – a temporary set, or **deciduous dentition**, made up of 20 **deciduous teeth** (also called **milk** or **baby teeth**), and a later **permanent dentition** (32 **permanent teeth**).

Parts of a tooth

Crown

The exposed part of a tooth. It is covered by **enamel**. It is the part most subject to damage or tooth decay.

Root

The part of a tooth that is fixed in a socket in the jaw. **Incisors** and **canines** have one root, **premolars** have one or two and **molars** have two or three. Each root is held in place by the tough fibers of a **ligament*** called the **periodontal ligament**. The fibers are fixed to the jawbone at one end, and to the **cement** at the other. They act as shock absorbers.

Dentine or ivory

A yellow substance which forms the second layer inside a tooth. It is not as hard as **enamel** but, like it, has many of the same constituents as bone. It also contains **collagen*** fibers and strands of **cytoplasm***. These run out from the **pulp** cells in the **pulp cavity**.

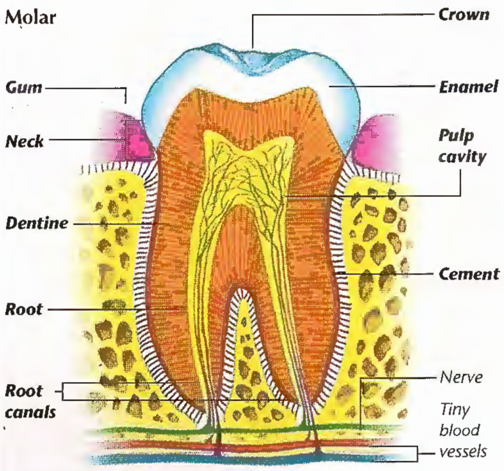

Neck or cervix

The part of a tooth just below the surface, lying between the **crown** and the **root**.

Enamel

A substance similar to bone, though it is harder (the hardest substance in the body) and has no living cells. It consists of tightly-packed crystals of **apatite**, a mineral which contains calcium, phosphorus and fluorine.

Cement or cementum

A bone-like substance, similar to **enamel** but softer. It forms the thin surface layer of the **root** and is attached to the jaw by the **periodontal ligament** (see **root**).

Pulp cavity

The central area of a tooth, surrounded by **dentine**. It is filled with a soft tissue called **pulp**, which contains blood vessels and nerve fiber endings. These enter at the base of a **root** and run up to the cavity inside **root canals**. The blood vessels supply food and oxygen to the living tissue, and the nerve fiber endings are **pain receptors***.

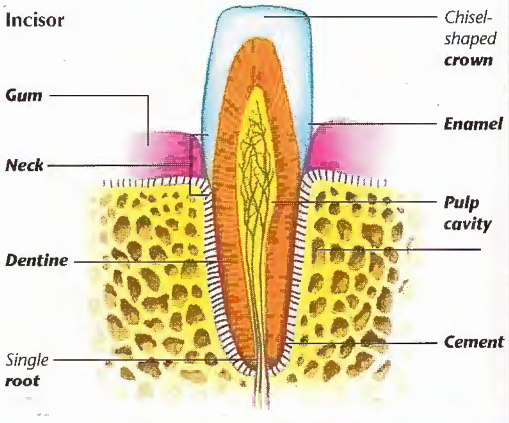

Types of teet

Incisors
Sharp, chisel-shaped t
and cutting. Each has
are four in each jaw, s
the mouth.

Canines or **cuspids**
Cone-shaped teeth (oft
teeth), used to tear foo
point (**cusp**) and one r
each jaw, one each sid
many mammals, they a

Premolars or **bicuspi**
Blunt, broad teeth, use
grinding (found in the
only). There are four in
each **canine**. Each has
cusps) and one **root**, e
premolars, which have

Molars
Blunt, broad teeth, sim
with a larger surface ar
used for crushing and g
has four surface points
have two **roots** each, a
three. There are six mo
three behind each pair
the third ones (at the ba
wisdom teeth.

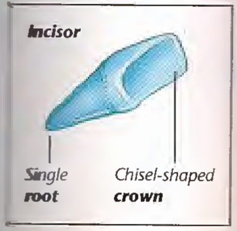

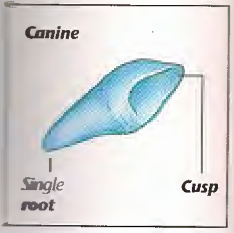

---

* **Collagen**, 280 (**Connective tissue**); **Cytoplasm**, 238; **Ligaments**, 280; **Pain receptors**, 311.

by
a soft
have two
made
later

surface,
root.

it is
body)
of tightly-
which
fluorine.

enamel
layer of
by the

by
called
and
the base
inside root
food and
nerve

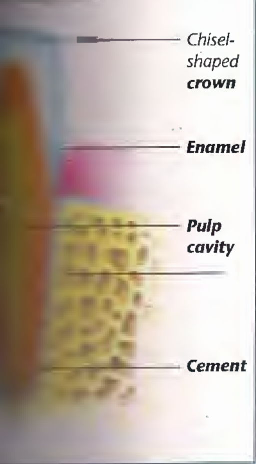
Chisel-shaped crown
Enamel
Pulp cavity
Cement

Types of teeth

Incisors

Sharp, chisel-shaped teeth, used for biting and cutting. Each has one root, and there are four in each jaw, set at the front of the mouth.

Canines or cuspids

Cone-shaped teeth (often called **eye** or **dog teeth**), used to tear food. Each has a sharp point (**cusp**) and one **root**. There are two in each jaw, one each side of the **incisors**. In many mammals, they are long and curved.

Premolars or bicuspids

Blunt, broad teeth, used for crushing and grinding (found in the permanent set of teeth only). There are four in each jaw, two behind each **canine**. Each has two sharp ridges (**cusps**) and one **root**, except the upper first premolars, which have two.

Molars

Blunt, broad teeth, similar to **premolars** but with a larger surface area. They are also used for crushing and grinding, and each has four surface points (**cusps**). Lower molars have two **roots** each, and upper ones have three. There are six molars in each jaw, three behind each pair of **premolars**, and the third ones (at the back) are known as **wisdom teeth**.

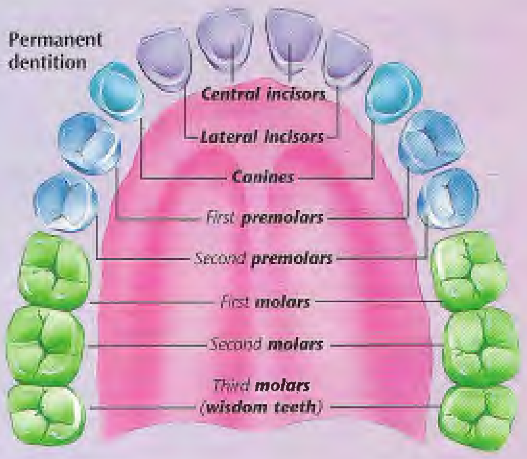
Permanent dentition
Central incisors
Lateral incisors
Canines
First premolars
Second premolars
First molars
Second molars
Third molars (wisdom teeth)

🟪 **Incisors** *(replace eight temporary incisors).*

🟦 **Premolars** *(replace eight temporary molars).*

🟦 **Canines** *(replace four temporary canines).*

🟩 **Molars** *(appear behind* ***premolars*** *and do not replace any deciduous teeth).*

Wisdom teeth

Four **molars** (the third ones in line), lying at the end points of the jaws. They appear last of all, when a person is fully mature (hence their name). Often there is no room for them to come through and they get stuck in the jawbone, or **impacted**. A few people never develop wisdom teeth.

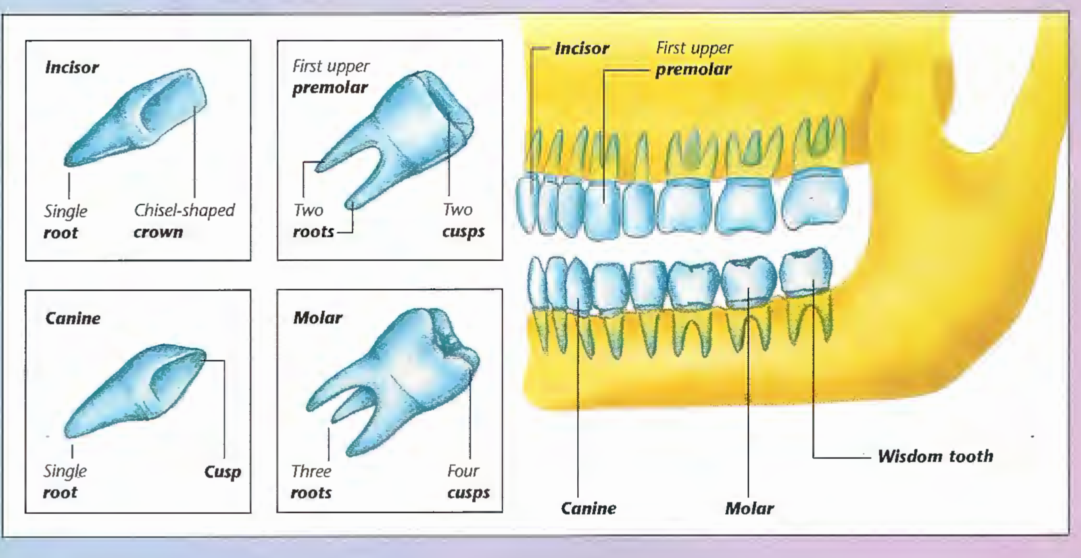
Incisor
Single root
Chisel-shaped crown
First upper premolar
Two roots
Two cusps
Canine
Single root
Cusp
Molar
Three roots
Four cusps
Incisor
First upper premolar
Canine
Molar
Wisdom tooth

BLOOD

Red blood cells

Blood is a vital body fluid, consisting of **plasma**, **platelets** and **red** and **white blood cells**. An adult human has about 5.5 liters (9.5 pints), which travel around in the **circulatory system*** – a system of tubes called **blood vessels**. The blood distributes heat and carries many important substances in its plasma. Old, dying blood cells are constantly being replaced by new ones in a process called **hemopoiesis**.

Blood constituents

Plasma

The pale liquid (about 90% water) which contains the blood cells. It carries dissolved food for the body cells, waste matter and carbon dioxide secreted by them, **antibodies** to combat infection, and **enzymes*** and **hormones*** which control body processes.

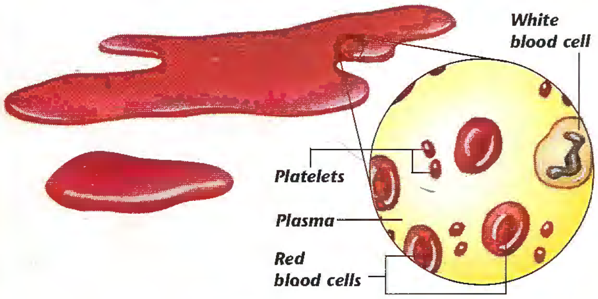
White blood cell
Platelets
Plasma
Red blood cells

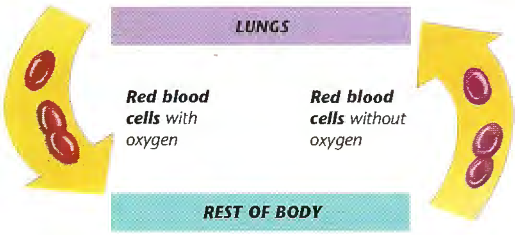
LUNGS
Red blood cells with oxygen
Red blood cells without oxygen
REST OF BODY

Red blood cells

Also called **red corpuscles** or **erythrocytes**. Red, disk-shaped cells with no **nuclei***. They are made in the **bone marrow*** and contain **hemoglobin** (an iron compound which gives blood a dark red color). This combines with oxygen in the lungs to form **oxyhemoglobin**, and the blood becomes bright red. The red cells pass the oxygen to the body cells (by **diffusion***) and then return to the lungs with hemoglobin.

Platelets or thrombocytes

Very small, disk-shaped bodies with no **nuclei***, made in the **bone marrow***. They gather particularly at an injured area, where they are important in the **clotting** of blood.

White blood cells

Also called **white corpuscles** or **leucocytes**. Large, opaque blood cells, important in body defense. There are several types. **Lymphocytes**, for example, are made in **lymphoid tissue*** and are found in the **lymphatic system*** as well as blood. They make **antibodies**. Other white cells – **monocytes** – are made in **bone marrow***. They "swallow up" foreign bodies, e.g. bacteria, in a process called **phagocytosis***. Many of them (called **macrophages**) leave the blood vessels. They either travel around (**wandering macrophages**) or become fixed in an organ, e.g. a **lymph node*** (**fixed macrophages**).

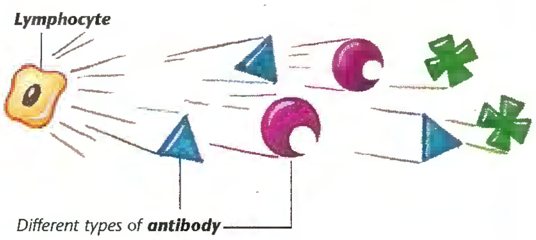
Lymphocyte
Different types of antibody

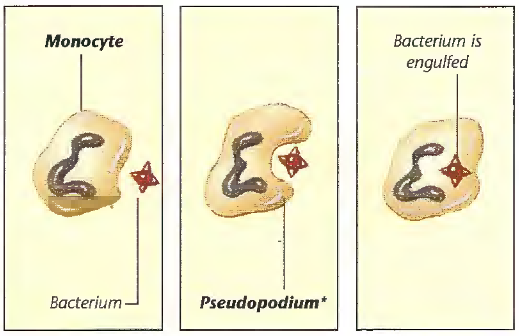
Monocyte
Bacterium
Pseudopodium*
Bacterium is engulfed

---

* **Bone marrow**, 281; **Circulatory system**, 288; **Diffusion**, 329; **Enzymes**, 333; **Hormones**, 336; **Lymphatic system**, **Lymph nodes**, 293; **Lymphoid tissue**, 293 (**Lymphoid organs**); **Nucleus**, 238; **Phagocytosis**, 268 (**Pseudopodium**).

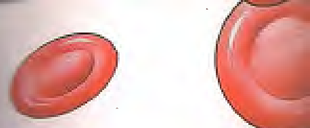

ABO blood groups

The main way of classi
with group A blood hav
red blood cells (and an
those with group B hav
antibodies). People wit
antigens (and neither ty
those with group O hav
both types of antibody).

Body defense

Antibodies

"Defense" proteins in
They are made by **lym**
blood cells) to comba
in the body. A differe
each antigen, and they
Anti-toxins neutralize
one joins with a toxin
antigen-antibody com
the bacteria or viruses
kill them by dissolving

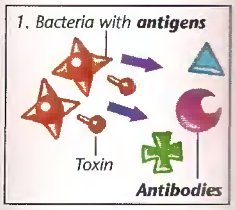
1. Bacteria with antigens
Toxin
Antibodies

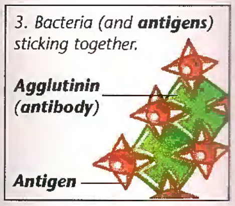
3. Bacteria (and antigens) sticking together.
Agglutinin (antibody)
Antigen

Antigens

Substances, mostly pr
the production of **anti**
them and any infectio
They may form part of
which enter the body,
(poisons) released by t
group antigens and an
are present in the body
to combat foreign bloo

---

* **Enzymes**, 333.

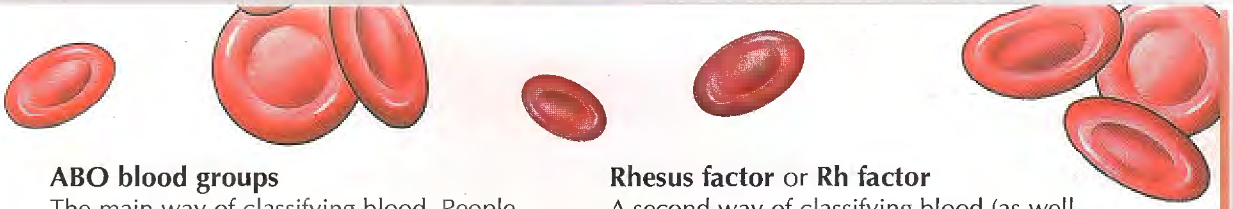

ABO blood groups

The main way of classifying blood. People with group A blood have A **antigen** on their **red blood cells** (and anti-B **antibodies**) and those with group B have B antigen (and anti-A antibodies). People with group AB have both antigens (and neither type of antibody) and those with group O have neither antigen (and both types of antibody).

Rhesus factor or Rh factor

A second way of classifying blood (as well as by **ABO blood group**). People whose **red blood cells** bear the **Rhesus antigen** are said to be **Rhesus positive**. Those without this antigen are **Rhesus negative**. Their blood does not normally contain anti-rhesus **antibodies**, but these would be produced if rhesus positive blood were to enter the body.

Body defense

Antibodies

"Defense" proteins in body fluids, e.g. **plasma**. They are made by **lymphocytes** (see **white blood cells**) to combat **antigens** (see below) in the body. A different antibody is made for each antigen, and they act in different ways. **Anti-toxins** neutralize toxins (poisons). Each one joins with a toxin molecule, making an **antigen-antibody complex**. **Agglutinins** stick the bacteria or viruses together, and **lysins** kill them by dissolving their outer membranes.

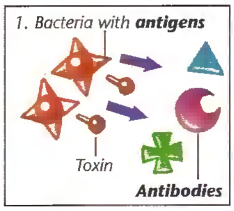
1. Bacteria with **antigens**
Toxin
Antibodies

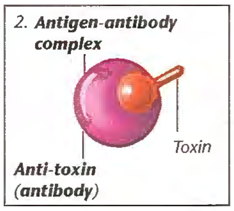
2. **Antigen-antibody complex**
Toxin
Anti-toxin (antibody)

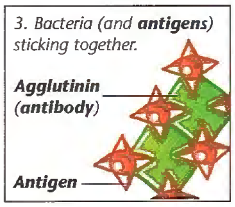
3. Bacteria (and **antigens**) sticking together.
Agglutinin (antibody)
Antigen

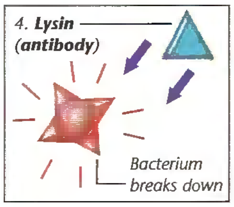
4. **Lysin (antibody)**
Bacterium breaks down

Antigens

Substances, mostly proteins, which cause the production of **antibodies** to combat them and any infection they may cause. They may form part of bacteria or viruses which enter the body, or they may be toxins (poisons) released by them. **ABO blood group** antigens and antibodies (see above) are present in the body from birth, ready to combat foreign blood group antigens.

Clotting or coagulation

The thickening of blood into a mass (**clot**) at the site of a wound. First, disintegrating **platelets** and damaged cells release a chemical called **thromboplastin**. This causes **prothrombin** (a **plasma** protein) to turn into **thrombin** (an **enzyme***). The thrombin then causes **fibrinogen** (another plasma protein) to harden into **fibrin**, a fibrous substance. A network of its fibers makes up the jelly-like clot.

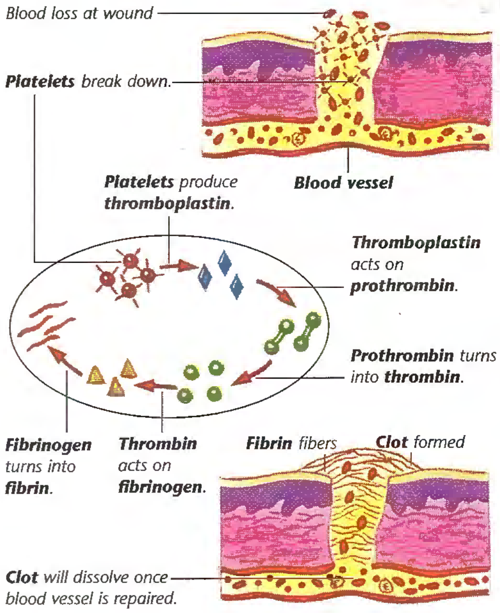
Blood loss at wound
Platelets break down.
Platelets produce **thromboplastin.**
Blood vessel
Thromboplastin acts on **prothrombin.**
Prothrombin turns into **thrombin.**
Fibrinogen turns into **fibrin.**
Thrombin acts on **fibrinogen.**
Fibrin fibers
Clot formed
Clot will dissolve once blood vessel is repaired.

Serum

A yellowy liquid consisting of the parts of the blood left after **clotting**. It contains many **antibodies** (produced to combat infections). When injected into other people, it can give temporary immunity to the infections.

---

* **Enzymes**, 333.

THE CIRCULATORY SYSTEM

The **circulatory** or **vascular system** is a network of blood-filled tubes, or **blood vessels**, of which there are three main types – **arteries**, **veins** and **capillaries**. A thin tissue layer called the **endothelium** lines arteries and veins, and is the only layer of capillary walls. Blood is kept flowing one way by the pumping of the heart, by muscles in artery and vein walls and by a decrease in pressure through the system (liquids flow from high to low pressure areas).

Passage of main substances in the circulatory system

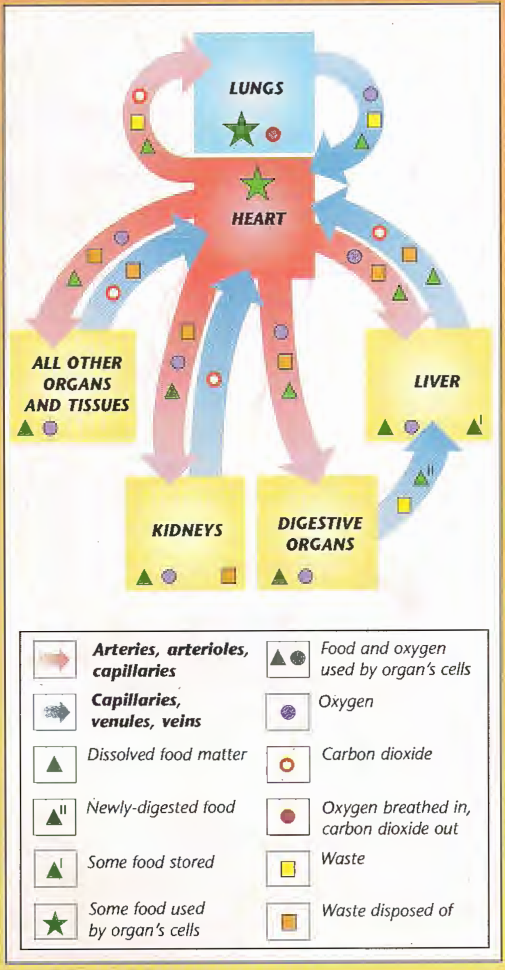
LUNGS
HEART
ALL OTHER ORGANS AND TISSUES
LIVER
KIDNEYS
DIGESTIVE ORGANS
Arteries, arterioles, capillaries
Capillaries, venules, veins
Dissolved food matter
Newly-digested food
Some food stored
Some food used by organ's cells
Food and oxygen used by organ's cells
Oxygen
Carbon dioxide
Oxygen breathed in, carbon dioxide out
Waste
Waste disposed of

Arteries

Wide, thick-walled blood vessels, making up the **arterial system** and carrying blood away from the heart. Smaller arteries (**arterioles**) branch off the main ones, and **capillaries** branch off the arterioles. Except in the **pulmonary arteries***, the blood contains oxygen (which makes it bright red). In all arteries it also carries dissolved food and waste, brought into the heart by **veins**, and there transferred to the arteries. These carry the food to the cells (via arterioles and capillaries) and the waste to the kidneys.

Artery

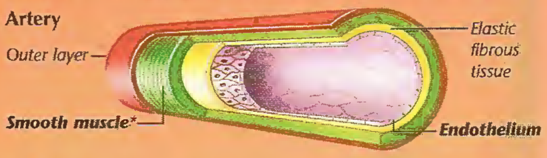
Outer layer
Smooth muscle*
Elastic fibrous tissue
Endothelium

Veins

Wide, thick-walled blood vessels, making up the **venous system** and carrying blood back to the heart. They contain valves to stop blood flowing backward due to gravity, and are formed from merging **venules** (small veins). These are formed in turn from merging **capillaries**. The blood contains carbon dioxide (except in the **pulmonary veins***) and waste matter, both picked up from body cells by the capillaries. The blood in the veins leading from the digestive system and liver also carries dissolved food. This is transferred to the arteries in the heart.

Vein

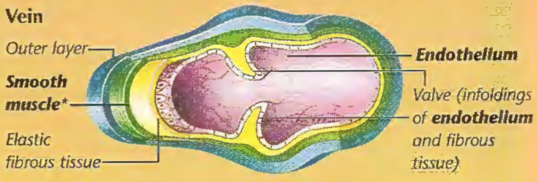
Outer layer
Smooth muscle*
Elastic fibrous tissue
Endothelium
Valve (infoldings of endothelium and fibrous tissue)

Capillaries

Narrow, thin-walled blood vessels, branching off **arterioles** (see **arteries**) to form a complex network. Oxygen and dissolved food pass out through their walls to the body cells, and carbon dioxide and waste pass in (see **tissue fluid**, page 292). The capillaries of the digestive organs and liver also pick up food. Capillaries finally join up again to form small veins (**venules**).

Capillary

Single layer (endothelium)

---

*__Hepatic portal vein__, 297 (**Liver**); **Pulmonary arteries**, 291 (**Pulmonary trunk**); **Pulmonary veins**, 291; **Smooth muscle**, 283.

The main art

Vascular *means "compos*
conducting vessels". In the
means having a blood sup

Avascular *means "contai*
vessels". In the case of an
no blood supply.

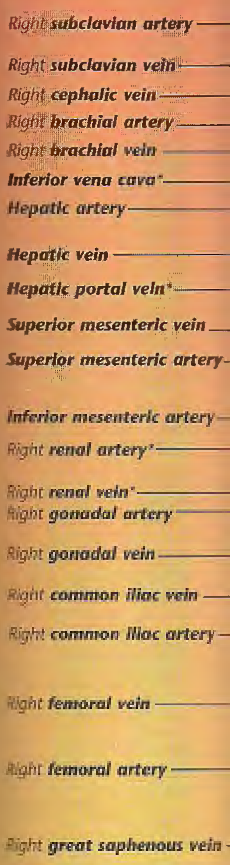
Right subclavian artery
Right subclavian vein
Right cephalic vein
Right brachial artery
Right brachial vein
Inferior vena cava*
Hepatic artery
Hepatic vein
Hepatic portal vein*
Superior mesenteric vein
Superior mesenteric artery
Inferior mesenteric artery
Right renal artery*
Right renal vein*
Right gonadal artery
Right gonadal vein
Right common iliac vein
Right common iliac artery
Right femoral vein
Right femoral artery
Right great saphenous vein

This is a diagram. The organs
circles) are not in the right pla

Key to organs

- ① Heart
- ② Lungs
- ③ Liver
- ④ Stomach
- ⑤ Spleen
- ⑥ P
- ⑦ I
- ⑧ K
- ⑨ G
  o
  p

Aorta, 291; **Hepatic porta**
Renal arteries, **Renal vei**

making up
blood away
arterioles)
capillaries branch
pulmonary
gen (which
it also carries
into the heart
the arteries.
via arterioles
the kidneys.

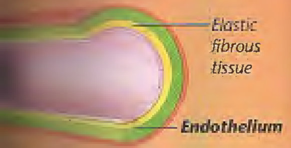
Elastic fibrous tissue
Endothelium

making up the
blood back to the
blood flowing
formed from
These are
capillaries. The
except in the
matter, both
capillaries. The
the digestive
dissolved food.
in the heart.

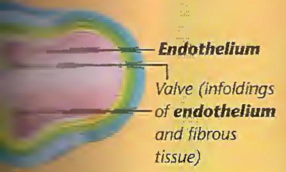
Endothelium
Valve (infoldings of endothelium and fibrous tissue)

branching
a complex
food pass out
cells, and carbon
tissue fluid, page
digestive organs
Capillaries finally
(venules).

Single layer (endothelium)

The main arteries and veins

Vascular *means "composed of or containing conducting vessels". In the case of animals, it means having a blood supply.*

Avascular *means "containing no conducting vessels". In the case of animals, it means having no blood supply.*

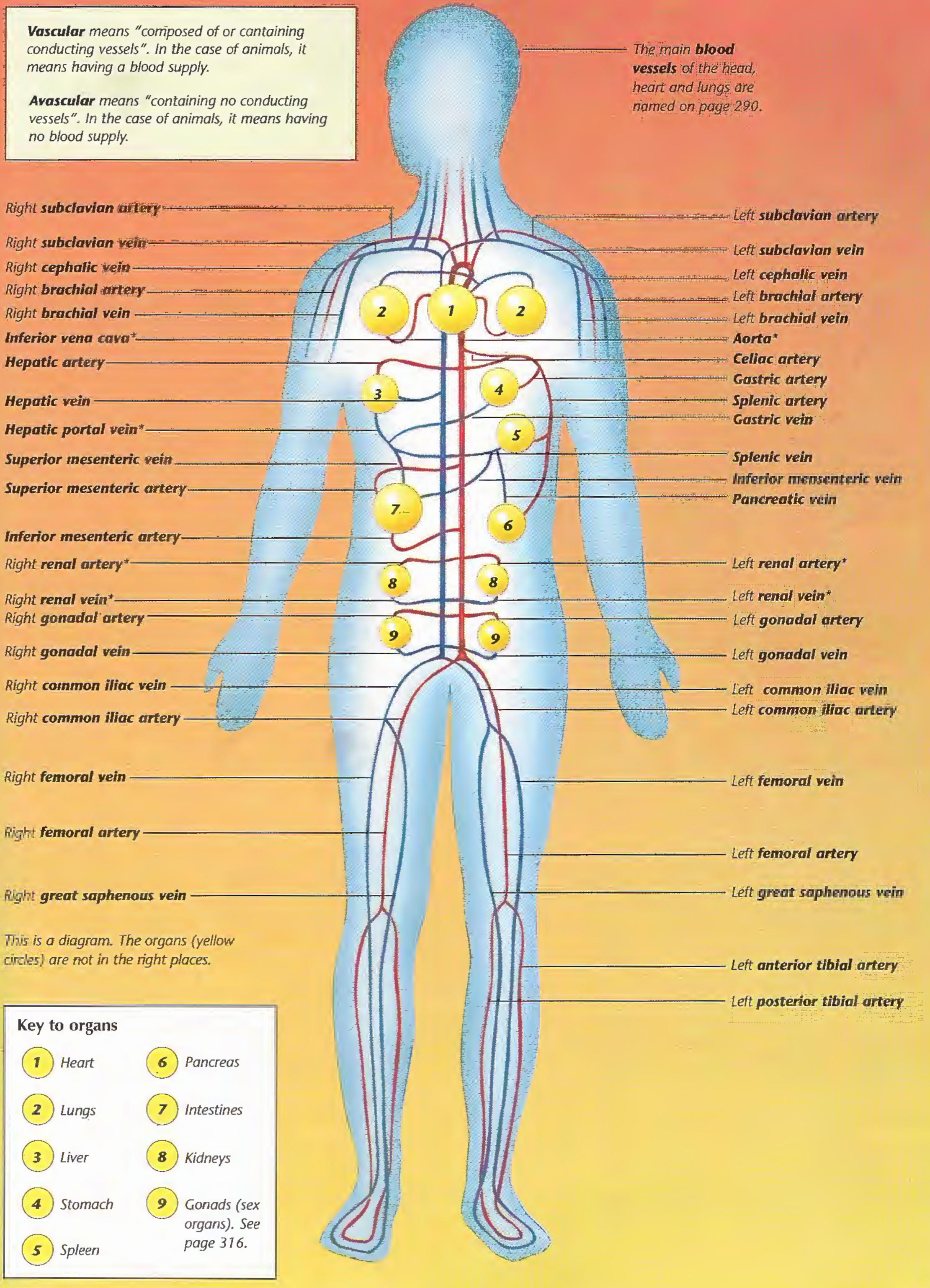
The main blood vessels of the head, heart and lungs are named on page 290.
Right subclavian artery
Right subclavian vein
Right cephalic vein
Right brachial artery
Right brachial vein
Inferior vena cava*
Hepatic artery
Hepatic vein
Hepatic portal vein*
Superior mesenteric vein
Superior mesenteric artery
Inferior mesenteric artery
Right renal artery*
Right renal vein*
Right gonadal artery
Right gonadal vein
Right common iliac vein
Right common iliac artery
Right femoral vein
Right femoral artery
Right great saphenous vein
Left subclavian artery
Left subclavian vein
Left cephalic vein
Left brachial artery
Left brachial vein
Aorta*
Celiac artery
Gastric artery
Splenic artery
Gastric vein
Splenic vein
Inferior mesenteric vein
Pancreatic vein
Left renal artery*
Left renal vein*
Left gonadal artery
Left gonadal vein
Left common iliac vein
Left common iliac artery
Left femoral vein
Left femoral artery
Left great saphenous vein
Left anterior tibial artery
Left posterior tibial artery
1
2
2
3
4
5
6
7
8
8
9
9

This is a diagram. The organs (yellow circles) are not in the right places.

Key to organs

1 Heart	6 Pancreas
2 Lungs	7 Intestines
3 Liver	8 Kidneys
4 Stomach	9 Gonads (sex organs). See page 316.
5 Spleen	

* **Aorta**, 291; **Hepatic portal vein**, 297 (**Liver**); **Inferior vena cava**, 291; **Renal arteries**, **Renal veins**, 300 (**Kidneys**).

THE HEART

The **heart** is a muscular organ which pumps blood around the blood vessels. (The heart and blood vessels together are the **cardiovascular system.**) It is surrounded by the **pericardial sac**. This consists of an outer membrane (the **pericardium**) and the cavity (**pericardial cavity**) between it and the heart. This cavity is filled with a cushioning fluid (**pericardial fluid**). The heart has four chambers – two **atria** and two **ventricles**, all lined by a thin tissue layer called the **endocardium**.

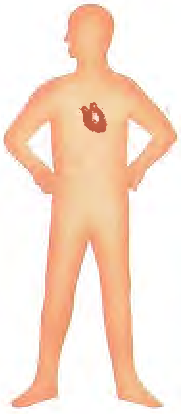

*Position of **heart***

The chambers of the heart

Atria (sing. atrium) or auricles

The two upper chambers. The left atrium receives **oxygenated** blood (blood with fresh oxygen – see also **hemoglobin***) from the lungs via the **pulmonary veins**. The right atrium receives **deoxygenated** blood from the rest of the body via the **superior** and **inferior vena cavae**. This is blood whose oxygen has been used by the cells and replaced by carbon dioxide.

Ventricles

The two lower chambers. The left ventricle receives blood from the left **atrium** and pumps it into the **aorta**. The right ventricle receives blood from the right atrium and pumps it via the **pulmonary trunk** to the lungs.

Key
→ ***Oxygenated** blood*
→ ***Deoxygenated** blood*

***Cardiac** means "of or near the heart".*
***Pulmonary** means "of the lungs".*

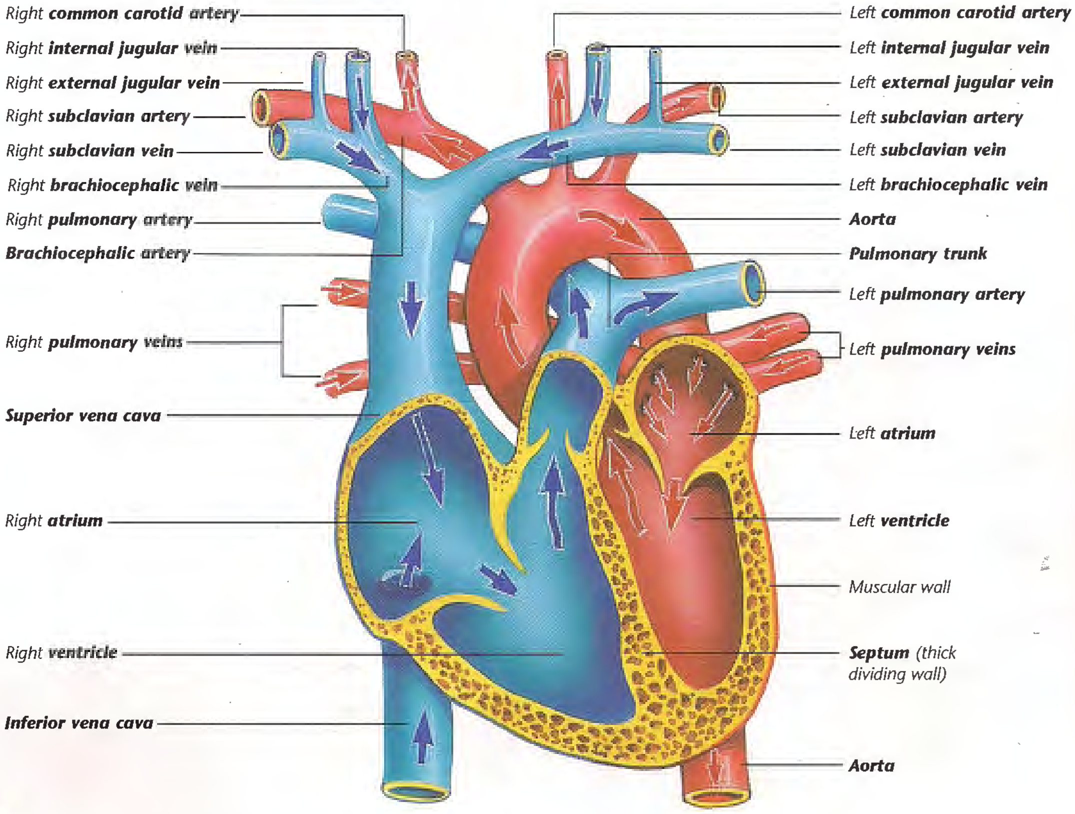
Right common carotid artery
Right internal jugular vein
Right external jugular vein
Right subclavian artery
Right subclavian vein
Right brachiocephalic vein
Right pulmonary artery
Brachiocephalic artery
Right pulmonary veins
Superior vena cava
Right atrium
Right ventricle
Inferior vena cava
Left common carotid artery
Left internal jugular vein
Left external jugular vein
Left subclavian artery
Left subclavian vein
Left brachiocephalic vein
Aorta
Pulmonary trunk
Left pulmonary artery
Left pulmonary veins
Left atrium
Left ventricle
Muscular wall
Septum (thick dividing wall)
Aorta

---

*Hemoglobin, 286 (Red blood cells).

The main ar

Aorta

The largest artery* i
blood with fresh oxy
ventricle to begin its

Pulmonary trunk

The artery* which c
oxygen out of the rig
the heart, it splits int
pulmonary arteries,

Superior vena cava

One of the two main
needing fresh oxyger
the right atrium. All
merge into it.

Inferior vena cava

One of the two main
needing fresh oxyger
the right atrium. All
merge into it.

Pulmonary veins

Four veins* which ca
oxygen to the left atr
veins come from the
pulmonary veins con

Semilunar valves

Two valves, so calle
because they have
crescent-shaped flap
is the aortic valve b
the left ventricle an
aorta. The other is t
pulmonary valve be
the right ventricle ar
pulmonary trunk.

Atrioventricular v
AV valves

Two valves, each be
atrium and its corre
ventricle. The left A
or mitral valve, is a
valve, i.e. it has two
flaps, or cusps. The
valve is a tricuspid v
i.e. it has three cusp

Arteries, Veins, 288.

The main arteries and veins

Aorta

The largest **artery*** in the body. It carries blood with fresh oxygen out of the left **ventricle** to begin its journey around the body.

Pulmonary trunk

The **artery*** which carries blood needing fresh oxygen out of the right **ventricle**. After leaving the heart, it splits into the right and left **pulmonary arteries**, one going to each lung.

Superior vena cava

One of the two main **veins***. It carries blood needing fresh oxygen from the upper body to the right **atrium**. All the upper body veins merge into it.

Inferior vena cava

One of the two main **veins***, carrying blood needing fresh oxygen from the lower body to the right **atrium**. All the lower body veins merge into it.

Pulmonary veins

Four **veins*** which carry blood with fresh oxygen to the left **atrium**. Two right pulmonary veins come from the right lung, and two left pulmonary veins come from the left lung.

Semilunar valves

Two valves, so called because they have crescent-shaped flaps. One is the **aortic valve** between the left **ventricle** and the **aorta**. The other is the **pulmonary valve** between the right ventricle and the **pulmonary trunk**.

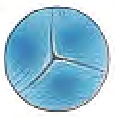
Closed flaps of ***pulmonary valve***

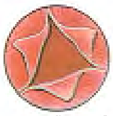
Open flaps of ***aortic valve***

Atrioventricular valves or AV valves

Two valves, each between an **atrium** and its corresponding **ventricle**. The left AV valve, or **mitral valve**, is a **bicuspid valve**, i.e. it has two movable flaps, or **cusps**. The right AV valve is a **tricuspid valve**, i.e. it has three cusps.

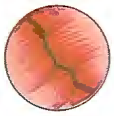
Closed ***cusps*** *of left* ***AV valve***

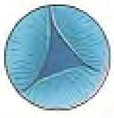
Open ***cusps*** *of right* ***AV valve***

The cardiac cycle

The **cardiac cycle** is the series of events which make up one complete pumping action of the heart, and which can be heard as the heartbeat (about 70 times a minute). First, both **atria** contract and pump blood into their respective **ventricles**, which relax to receive it. Then the atria relax and take in blood, and the ventricles contract to pump it out. The relaxing phase of a chamber is its **diastole phase**; the contracting phase is its **systole phase**. There is a short pause after the systole phase of the ventricles, during which all chambers are in diastole phase (relaxing). The different **valves** which open and close during the cycle are defined below left.

Cardiac cycle

***1. Atria** in **systole phase**, **ventricles** in **diastole phase**.*

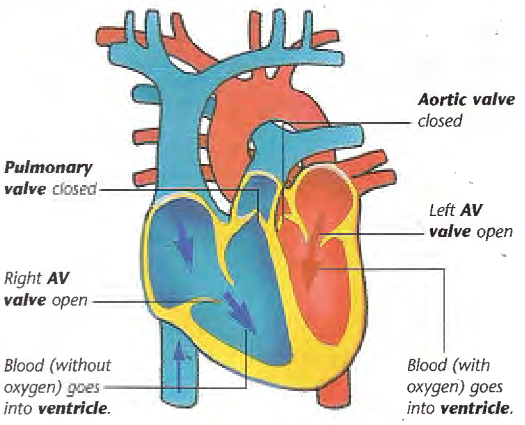
Aortic valve closed
Pulmonary valve closed
Left AV valve open
Right AV valve open
Blood (without oxygen) goes into ventricle.
Blood (with oxygen) goes into ventricle.

***2. Atria** in **diastole phase**, **ventricles** in **systole phase**.*

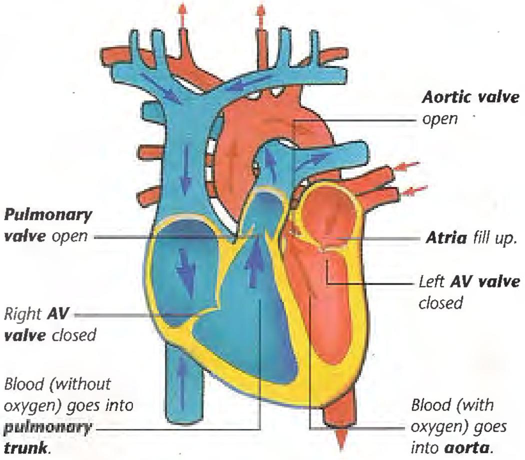
Aortic valve open
Pulmonary valve open
Atria fill up.
Left AV valve closed
Right AV valve closed
Blood (without oxygen) goes into pulmonary trunk.
Blood (with oxygen) goes into aorta.

---

* **Arteries**, **Veins**, 288.

TISSUE FLUID AND THE LYMPHATIC SYSTEM

The smallest blood vessels, called **capillaries***, are those in the most direct contact with the individual cells of the body, but even they do not touch the cells. The food and oxygen they carry finally reach the cells in **tissue fluid**, a substance which forms the link between the **circulatory system*** and the body's drainage system, known as the **lymphatic system**.

Tissue fluid

Also called **intercellular** or **interstitial fluid**. A fluid which surrounds the body cells. It seeps out from the blood through the walls of **capillaries*** (mainly at their high-pressure ends, after they have branched from **arterioles***) and is essentially **plasma***, though with fewer proteins. It carries oxygen and dissolved food to the body cells, and carbon dioxide and waste matter away from them. These latter substances enter the capillaries (mainly at their low-pressure ends, before they form **venules***).

The protein molecules not needed by the cells are too large to re-enter the capillaries. They pass, with some of the waste, into the **lymph capillaries** (see **lymph vessels**), whose walls are more easily penetrated.

The lymphatic system

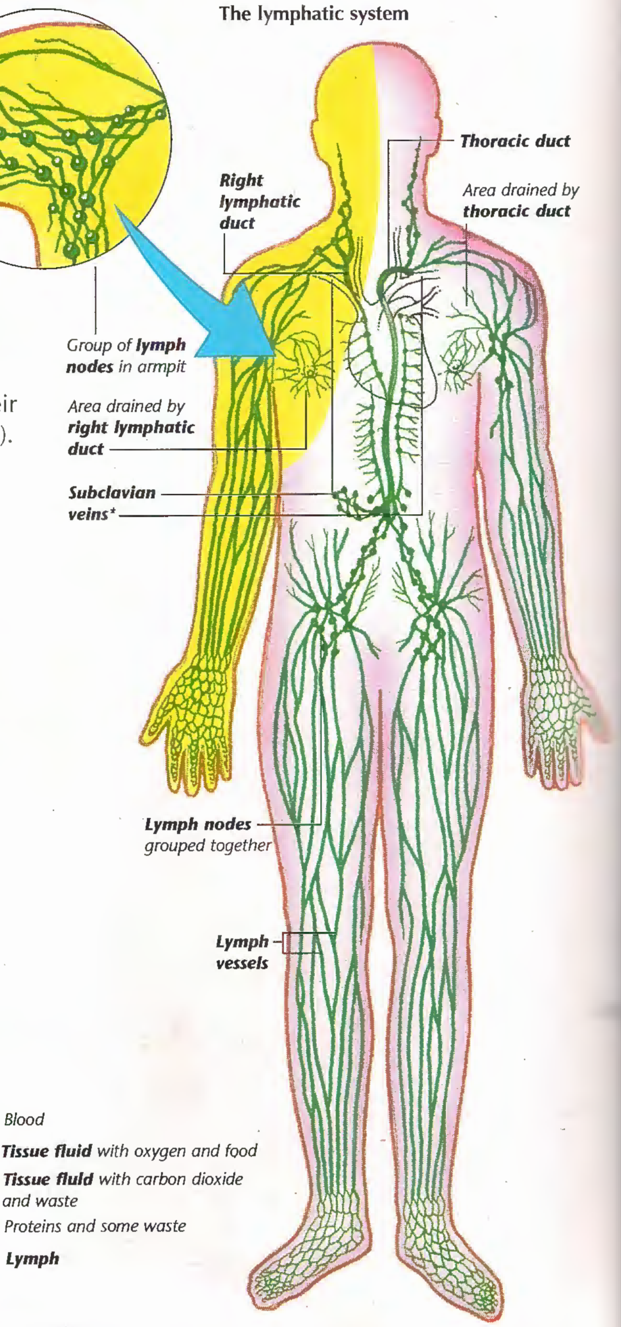
Thoracic duct
Right lymphatic duct
Area drained by thoracic duct
Group of lymph nodes in armpit
Area drained by right lymphatic duct
Subclavian veins*
Lymph nodes grouped together
Lymph vessels

Movement of substances in tissue fluid

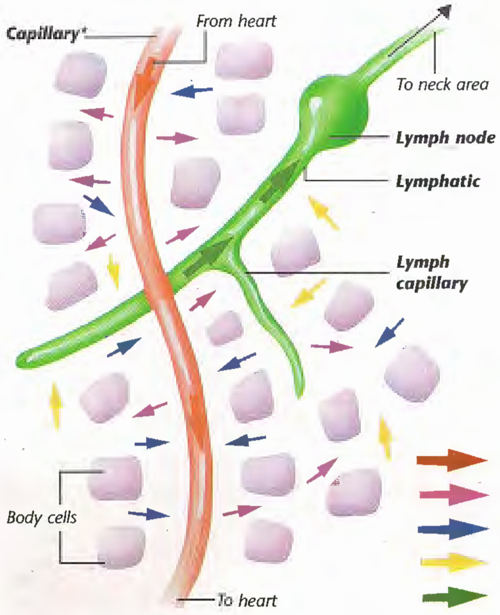
Capillary*
From heart
To neck area
Lymph node
Lymphatic
Lymph capillary
Body cells
To heart
Blood
Tissue fluid with oxygen and food
Tissue fluid with carbon dioxide and waste
Proteins and some waste
Lymph

---

* **Arterioles**, 288 (**Arteries**); **Capillaries**, 288; **Circulatory system**, 288; **Plasma**, 286; **Subclavian veins**, 289; **Venules**, 288 (**Veins**).

Lymphatic system

A system of tubes (**lym**
organs (**lymphoid orga**
recycling of body fluid
against disease. The ly
liquid **lymph** around th
back into the **veins***, ar
are the source of diseas

Lymph

The liquid in **lymph ve**
lymphocytes (see **lymp**
substances picked up fr
especially proteins suc
and **enzymes***) and als
lymph vessels).

Lymphoid org

The **lymphoid organs,**
are bodies connected
system. They are all m
of tissue (**lymphoid** or
and they all produce
disease-fighting white

Lymph nodes or lym

Small lymphoid organ
course of **lymph vesse**
e.g. in the armpits. Th
lymphocyte productio
contain a filter system
and foreign bodies. Th
by white blood cells (f

Spleen

The largest lymphoid organ, found just belo
the **diaphragm*** on the left side of the body. It holds an emergency store of red blood cell
and also contains white blood cells (**fixed macrophages***) which destroy foreign bodies, e.g. bacteria, and old blood cells.

Diaphragm, 298; **Endotheliu**
Lacteals, 295 (**Small intestine**)

most direct
not touch the
tissue fluid,
and the

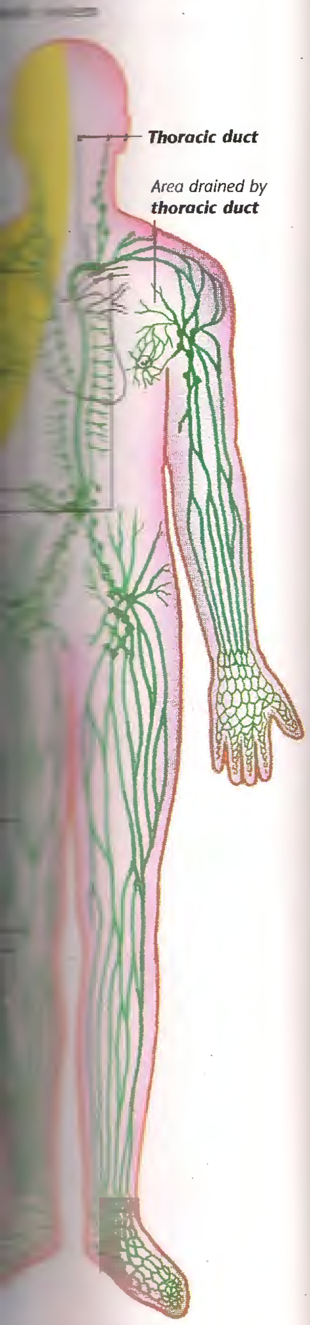

Lymphatic system

A system of tubes (**lymph vessels**) and small organs (**lymphoid organs**), important in the recycling of body fluids and in the fight against disease. The lymph vessels carry the liquid **lymph** around the body and empty it back into the **veins***, and the lymphoid organs are the source of disease-fighting cells.

Lymph

The liquid in **lymph vessels**. It contains **lymphocytes** (see **lymphoid organs**), some substances picked up from **tissue fluid** (especially proteins such as **hormones*** and **enzymes***) and also fat particles (see **lymph vessels**).

Lymph vessels or lymphatic vessels

Blind-ended tubes carrying **lymph** from all body areas toward the neck, where it is emptied back into the blood. They are lined with **endothelium***, and have valves to stop the lymph from being pulled back by gravity.

The thinnest lymph vessesls are **lymph capillaries**, and include the important **lacteals***, which pick up fat particles (too large to enter the bloodstream directly). The capillaries join to form larger vessels called **lymphatics**, which finally unite to form two tubes – the **right lymphatic duct** (emptying into the right **subclavian vein***) and the **thoracic duct** (emptying into the left **subclavian vein***).

Lymphoid organs

The **lymphoid organs**, or **lymphatic organs**, are bodies connected to the **lymphatic system**. They are all made of the same type of tissue (**lymphoid** or **lymphatic tissue**) and they all produce **lymphocytes*** – disease-fighting white blood cells.

Lymph nodes or lymph glands

Small lymphoid organs found along the course of **lymph vessels**, often in groups, e.g. in the armpits. They are the main sites of **lymphocyte** production (see above) and also contain a filter system which traps bacteria and foreign bodies. These are then engulfed by white blood cells (**fixed macrophages***).

Spleen

The largest lymphoid organ, found just below the **diaphragm*** on the left side of the body. It holds an emergency store of red blood cells and also contains white blood cells (**fixed macrophages***) which destroy foreign bodies, e.g. bacteria, and old blood cells.

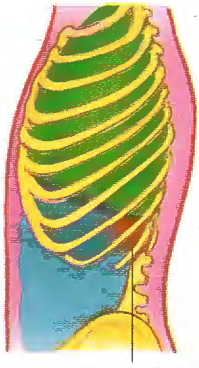

Position of **spleen**

Thymus gland

A lymphoid organ in the upper part of the chest. It is fairly large in children, reaches its maximum size at **puberty*** and then undergoes **atrophy**, i.e. it wastes away.

Tonsils

Four lymphoid organs: one **pharyngeal tonsil** (the **adenoids**) at the back of the nose, one **lingual tonsil** at the base of the tongue and two **palatine tonsils** at the back of the mouth.

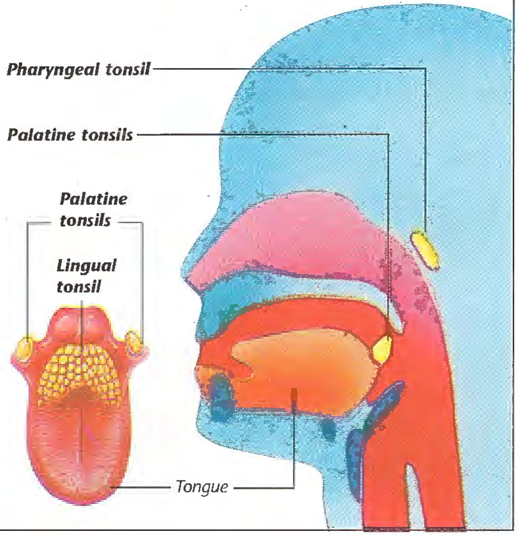

---

* **Diaphragm**, 298; **Endothelium**, 288; **Enzymes**, 333; **Fixed macrophages**, 286 (**White blood cells**); **Hormones**, 336; **Lacteals**, 295 (**Small intestine**); **Lymphocytes**, 286 (**White blood cells**); **Puberty**, 318; **Subclavian veins**, 289; **Veins**, 288.

THE DIGESTIVE SYSTEM

After food is taken in, or **ingested**, it passes through the **digestive system**, gradually being broken down into simple soluble substances by a process called **digestion** (see also pages 338-339). The simple substances are absorbed into the blood vessels around the system and transported to the body cells. Here they are used to provide energy and build new tissue. For more about all these different processes, see pages 330-335. The main parts of the digestive system are listed on these two pages. The pancreas and liver (see page 297) also play a vital part in digestion, forming the two main **digestive glands*** (producing **digestive juices***).

Position of ***digestive system***

Alimentary canal

Also called the **alimentary tract**, **gastrointestinal** (**GI**) **tract**, **enteric canal** or the **gut**. A collective term for all the parts of the digestive system. It is a long tube running from the mouth to the **anus** (see **large intestine**). Most of its parts are in the lower body, or **abdomen**, inside the main body cavity, or **perivisceral cavity***. They are held in place by **mesenteries** – infoldings of the cavity lining (the **peritoneum**).

Pharynx

A cavity at the back of the mouth, where the mouth cavity (**oral** or **buccal cavity**) and the **nasal cavities*** meet. When food is swallowed, the **soft palate** (a tissue flap at the back of the mouth) closes the nasal cavities and the **epiglottis*** closes the **trachea***.

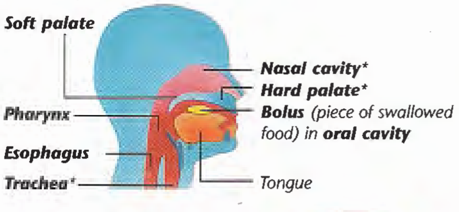

Bolus *(piece of swallowed food) in* ***oral cavity***

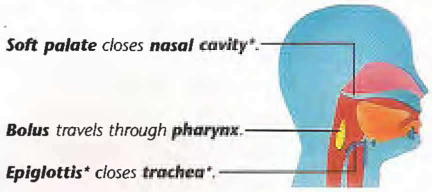

Soft palate *closes* ***nasal cavity****.

Bolus *travels through* ***pharynx***.

Epiglottis* *closes* ***trachea****.

Esophagus or gullet

The tube down which food travels to the **stomach**. A piece of swallowed food is a **bolus**.

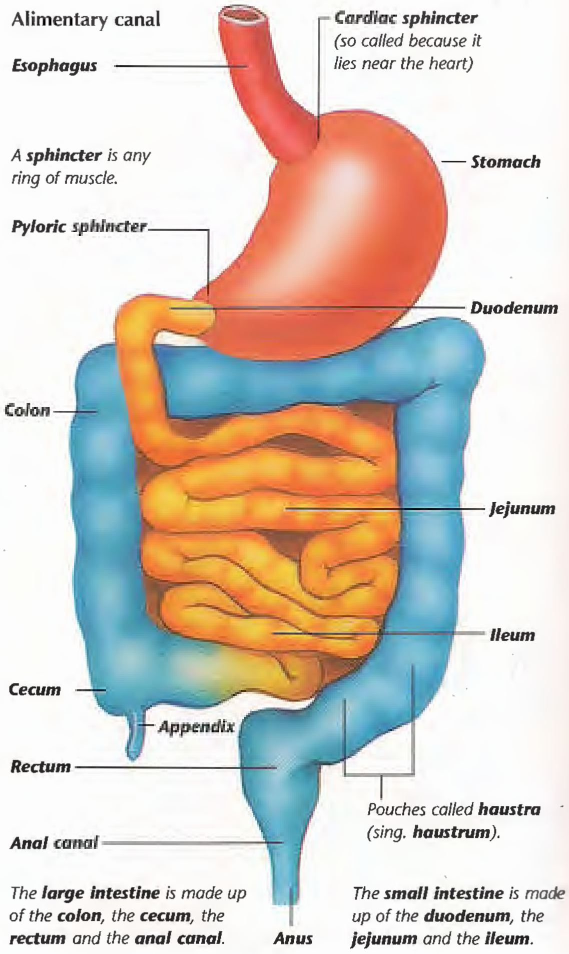

(so called because it lies near the heart)

A ***sphincter*** *is any ring of muscle.*

Pouches called ***haustra*** *(sing.* ***haustrum****).*

The ***large intestine*** *is made up of the* ***colon****, the* ***cecum****, the* ***rectum*** *and the* ***anal canal****.*

The ***small intestine*** *is made up of the* ***duodenum****, the* ***jejunum*** *and the* ***ileum****.*

---

* **Digestive juices**, 296 (**Digestive glands**); **Epiglottis**, 298; **Hard palate**, 307; **Nasal cavities**, 307 (**Nose**); **Perivisceral cavity**, 265; **Trachea**, 298.

Cardiac sphincter

Also called the **gastroe**
A muscular ring betwe
stomach. It relaxes to
food through.

Stomach

A large sac in which th
early stages of digestio
occur. Its lining has
many folds (**rugae**,
sing. **ruga**) which flatte
out to let it expand.
Some substances, e.g.
water, pass through its
into nearby blood vess
the semi-digested food
the **small intestine** (**duo**

Pyloric sphincter

Also called the **pyloric**
muscular ring between
small intestine. It relaxe
only after certain digest
occurred.

Small intestine

A coiled tube with three
the main site of digestio
ileum. Many tiny "finge
villus) project inward fr
contains **capillaries*** (tin
which most of the food
lymph vessel* called a **l**
combined fat particles
The remaining semi-liq
passes into the **large int**

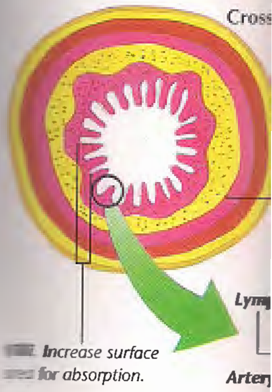

Increase surface for absorption.

* **Arteries**, 288; **Capillaries**, 288;
Epithelium, 310 (**Epidermis**); **E**

the
is a **bolus**.

because it
heart)

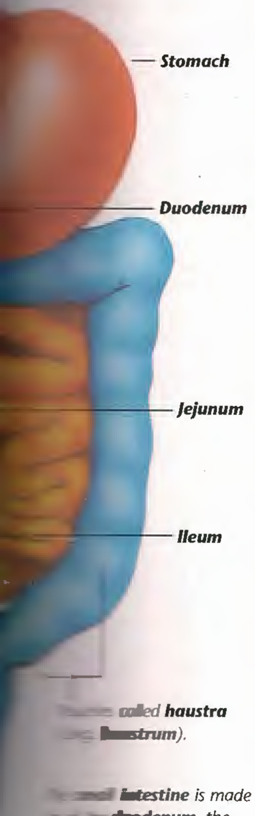
Stomach
Duodenum
Jejunum
Ileum
called haustra
intestine is made
duodenum, the
and the *ileum*.

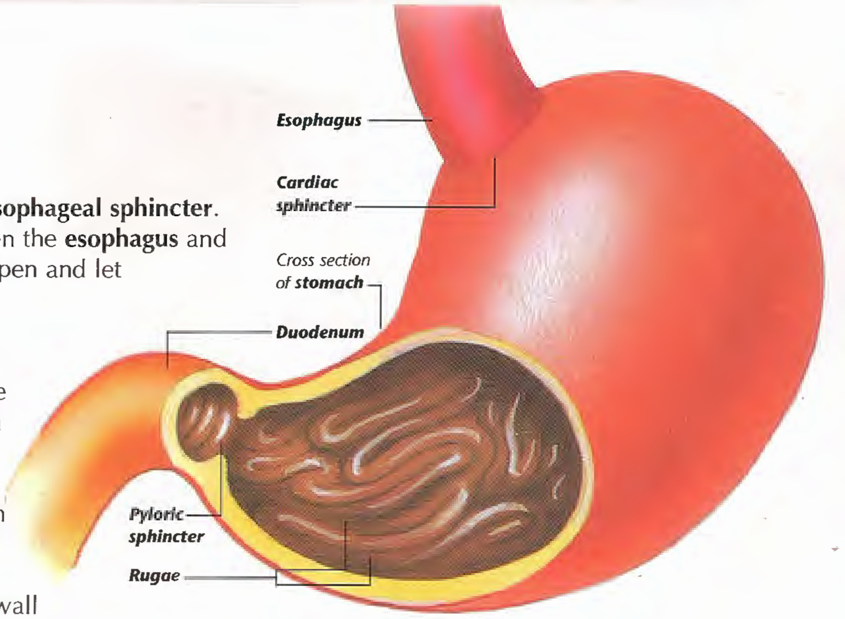
Esophagus
Cardiac sphincter
Cross section of stomach
Duodenum
Pyloric sphincter
Rugae

Cardiac sphincter

Also called the **gastroesophageal sphincter**. A muscular ring between the **esophagus** and **stomach**. It relaxes to open and let food through.

Stomach

A large sac in which the early stages of digestion occur. Its lining has many folds (**rugae**, sing. **ruga**) which flatten out to let it expand. Some substances, e.g. water, pass through its wall into nearby blood vessels, but almost all the semi-digested food (**chyme**) goes into the **small intestine** (**duodenum**).

Pyloric sphincter

Also called the **pyloric valve** or **pylorus**. A muscular ring between the **stomach** and the **small intestine**. It relaxes to let food through only after certain digestive changes have occurred.

Small intestine

A coiled tube with three parts – the **duodenum**, (the main site of digestion), **jejunum** and **ileum**. Many tiny "fingers" called **villi** (sing. **villus**) project inward from its lining. Each contains **capillaries*** (tiny blood vessels) into which most of the food is absorbed, and a **lymph vessel*** called a **lacteal**, which absorbs recombined fat particles (see **fats**, page 330). The remaining semi-liquid waste mixture passes into the **large intestine**.

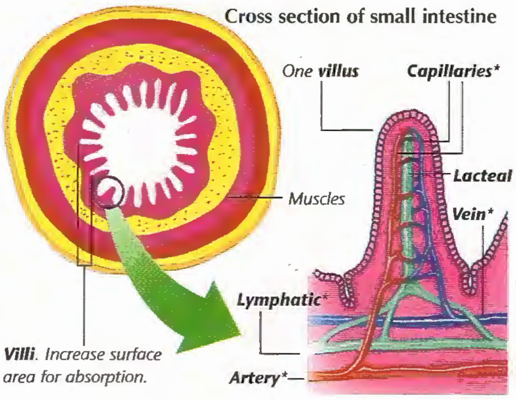
Cross section of small intestine
One villus
Capillaries*
Lacteal
Muscles
Vein*
Lymphatic*
Artery*

Villi. Increase surface area for absorption.

Large intestine

A thick tube receiving waste from the **small intestine**. It consists of the **cecum*** (a redundant sac), **colon**, **rectum** and **anal canal**. The colon contains bacteria, which break down any remaining food and make some important vitamins. Most of the water in the waste passes through the colon walls into nearby blood vessels. This leaves a semi-solid mass (**feces**), which is pushed out of the body (**defecation**) via the rectum, anal canal and **anus** – a hole surrounded by a muscular ring (the **anal sphincter**).

Appendix

A small, blind-ended tube off the **cecum** (see **large intestine**).

Mucous membrane or mucosa

A thin layer of tissue lining all digestive passages (also other passages, e.g. the air passages). It is a special type of **epithelium*** (a surface sheet of cells), containing many single-celled **exocrine glands***, called **mucous glands**. These secrete **mucus** – a lubricating fluid which, in the case of the digestive passages, also protects the passage walls against the action of **digestive juices***.

Peristalsis

The waves of contraction, produced by muscles in the walls of organs (especially digestive organs), which move substances along.

---

* **Arteries**, 288; **Capillaries**, 288; **Cecum**, 271; **Digestive juices**, 296 (**Digestive glands**); **Epithelium**, 310 (**Epidermis**); **Exocrine glands**, 296; **Lymphatic**, 293 (**Lymph vessels**); **Veins**, 288.

GLANDS

Glands are special organs (or sometimes groups of cells or single cells) which produce and secrete a variety of substances vital to life. There are two types of human glands – **exocrine** and **endocrine**.

Exocrine glands

Exocrine glands are glands which secrete substances through tubes, or **ducts**, onto a surface or into a cavity. Most body glands are exocrine, e.g. **sweat glands*** and **digestive glands**.

Digestive glands

Exocrine glands which secrete fluids called **digestive juices** into the digestive organs. The juices contain **enzymes*** which cause the breakdown of food (see chart, pages 338-339). Many of the glands are tiny, and set into the walls of the digestive organs, e.g. **gastric glands** in the stomach and **intestinal glands** (or **crypts of Lieberkühn**) in the small intestine. Others are larger and lie more freely, e.g. **salivary glands**. The largest are the **pancreas** and **liver**.

Salivary glands (secrete **saliva*** into mouth)

Only one side is shown – the three glands are duplicated on the other side.

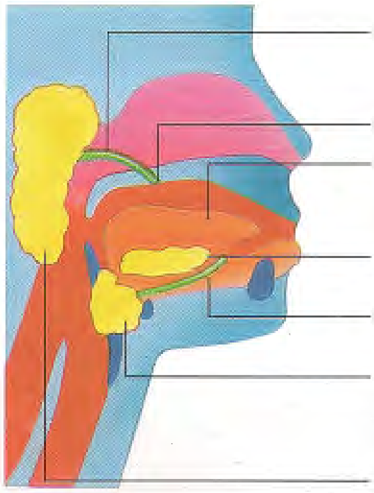
Duct called **Stensen's duct**
Opening into mouth
Tongue
Sublingual gland. Its ducts open into the floor of the mouth.
Duct called **Wharton's duct**
Submandibular gland or **submaxillary gland**
Parotid gland

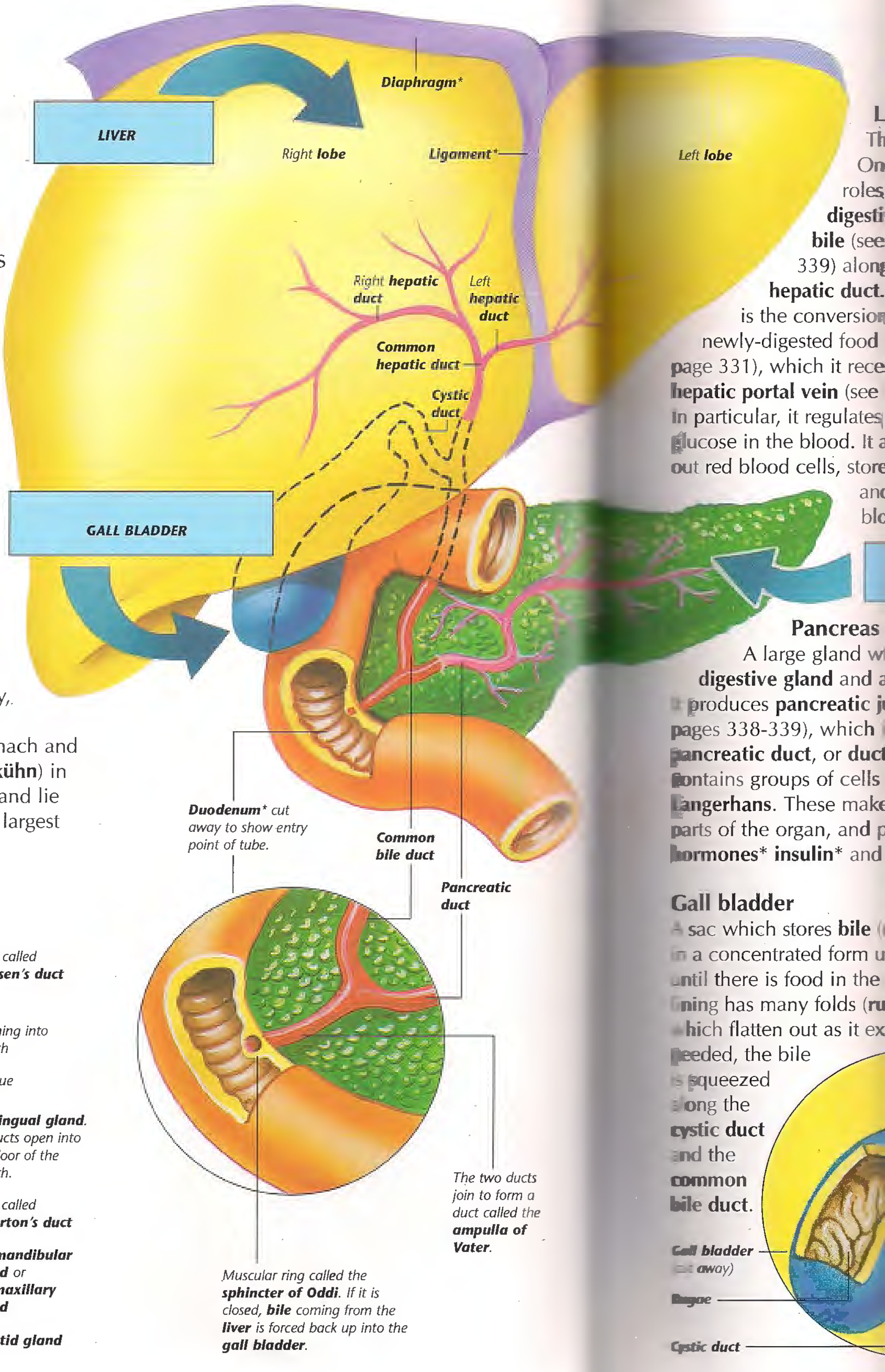
LIVER
Diaphragm*
Right lobe
Ligament*
Left lobe
Right hepatic duct
Left hepatic duct
Common hepatic duct
Cystic duct
GALL BLADDER
Duodenum* cut away to show entry point of tube.
Common bile duct
Pancreatic duct
The two ducts join to form a duct called the ampulla of Vater.
Muscular ring called the sphincter of Oddi. If it is closed, bile coming from the liver is forced back up into the gall bladder.
Gall bladder (cut away)
Rugae
Cystic duct

L
Th
On
roles
digesti
bile (see
339) along
hepatic duct.
is the conversion
newly-digested food
page 331), which it rece
hepatic portal vein (see
in particular, it regulates
glucose in the blood. It a
out red blood cells, store
and
blo

Pancreas

A large gland w
digestive gland and a
It produces pancreatic j
pages 338-339), which
pancreatic duct, or duct
contains groups of cells
Langerhans. These make
parts of the organ, and p
hormones* insulin* and

Gall bladder

A sac which stores **bile** (
in a concentrated form u
until there is food in the
lining has many folds (**ru**
which flatten out as it ex
needed, the bile
is squeezed
along the
cystic duct
and the
common
bile duct.

---

* **Diaphragm**, 298; **Duodenum**, 295 (**Small intestine**); **Enzymes**, 333; **Ligaments**, 280; **Saliva**, 338; **Sweat glands**, 311.

Duodenum, 295 (**Small intesti**
Hypothalamus, 303; **Insulin**, 3

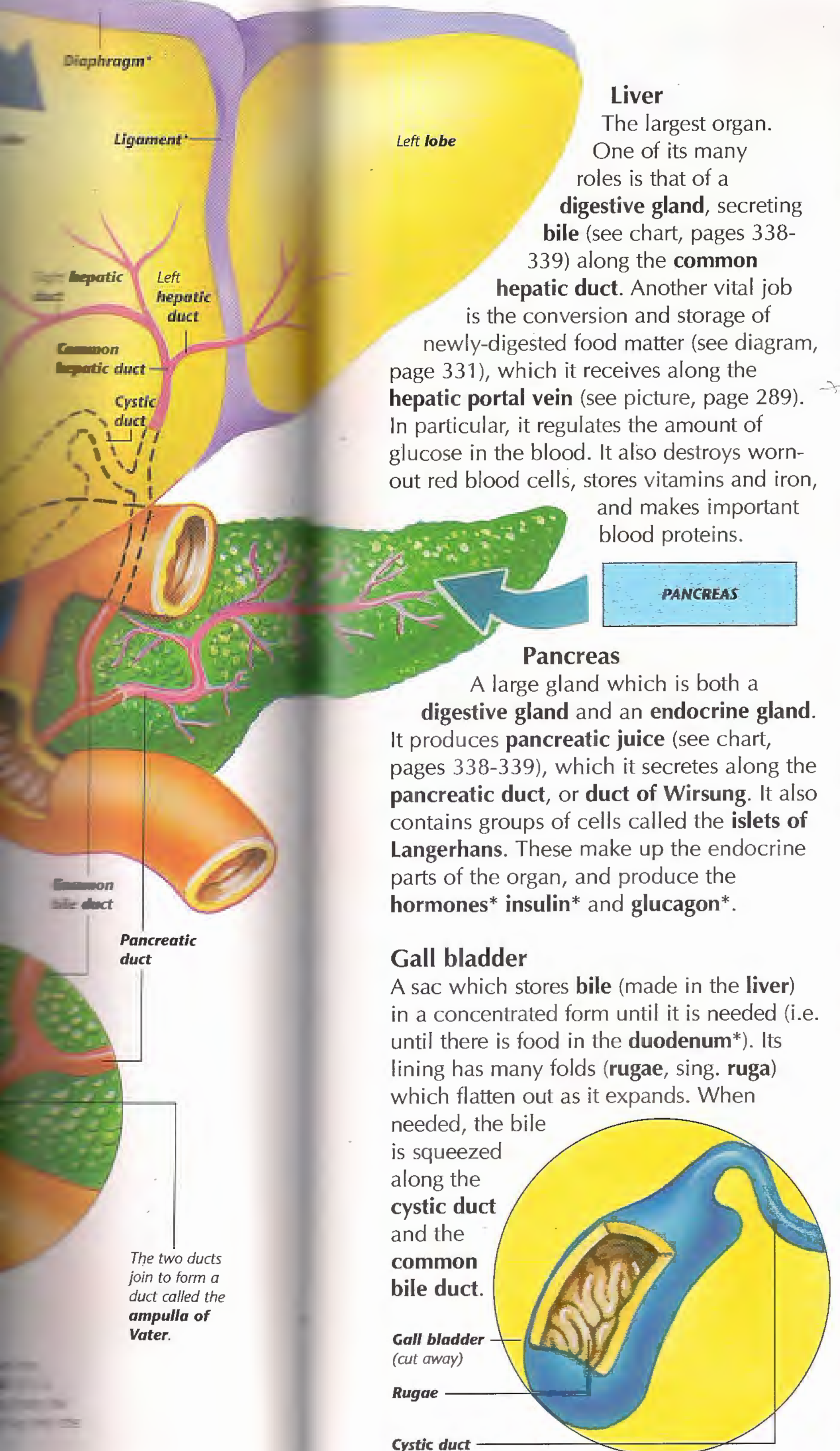
Diaphragm*
Ligament*
Left lobe
hepatic
Left hepatic duct
Common hepatic duct
Cystic duct
PANCREAS
Common bile duct
Pancreatic duct
The two ducts join to form a duct called the **ampulla of Vater**.
Gall bladder (cut away)
Rugae
Cystic duct

Liver

The largest organ. One of its many roles is that of a **digestive gland**, secreting **bile** (see chart, pages 338-339) along the **common hepatic duct**. Another vital job is the conversion and storage of newly-digested food matter (see diagram, page 331), which it receives along the **hepatic portal vein** (see picture, page 289). In particular, it regulates the amount of glucose in the blood. It also destroys worn-out red blood cells, stores vitamins and iron, and makes important blood proteins.

Pancreas

A large gland which is both a **digestive gland** and an **endocrine gland**. It produces **pancreatic juice** (see chart, pages 338-339), which it secretes along the **pancreatic duct**, or **duct of Wirsung**. It also contains groups of cells called the **islets of Langerhans**. These make up the endocrine parts of the organ, and produce the **hormones*** **insulin*** and **glucagon***.

Gall bladder

A sac which stores **bile** (made in the **liver**) in a concentrated form until it is needed (i.e. until there is food in the **duodenum***). Its lining has many folds (**rugae**, sing. **ruga**) which flatten out as it expands. When needed, the bile is squeezed along the **cystic duct** and the **common bile duct**.

Endocrine glands

Endocrine or **ductless glands** are glands which secrete substances called **hormones** directly into the blood (i.e. blood vessels in the glands). For more about hormones, see pages 336-337. The glands may be separate bodies (e.g. those below) or cells inside organs, e.g. in the sex organs.

Pituitary gland

Also called the **pituitary body** or **hypophysis**. A gland at the base of the brain, directly influenced by the **hypothalamus*** (see also **hormones**, page 336) and made up of an **anterior** (front) **lobe** (**adenohypophysis**) and a **posterior** (back) **lobe** (**neurohypophysis**). Many of its hormones are **tropic hormones**, i.e. they stimulate other glands to secrete hormones. It makes **ACTH**, **TSH**, **STH**, **FSH**, **LH**, **lactogenic hormone**, **oxytocin** and **ADH**.

Thyroid gland

A large gland around the **larynx***. It produces **thyroxin** and **thyrocalcitonin**.

Parathyroid glands

Two pairs of small glands embedded in the **thyroid gland**. They produce **PTH**.

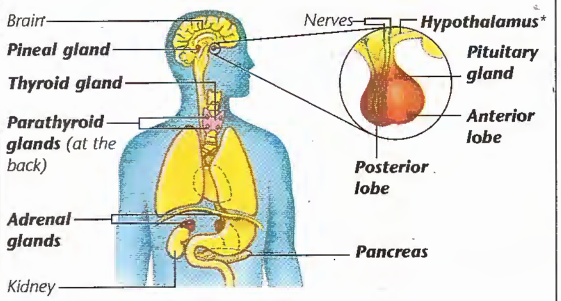
Brain
Pineal gland
Thyroid gland
Parathyroid glands (at the back)
Adrenal glands
Kidney
Nerves
Hypothalamus*
Pituitary gland
Anterior lobe
Posterior lobe
Pancreas

Adrenal glands or suprarenal glands

A pair of glands, one gland lying above each kidney. Each has an outer layer (**cortex**), producing **aldosterone**, **cortisone** and **hydrocortisone**, and an inner layer (**medulla**), producing **adrenalin** and **noradrenalin**.

Pineal gland

Also called the **pineal body**. A small gland at the front of the brain. Its role is not clear, but it is known to secrete **melatonin**, a hormone thought to influence **sex hormone*** production.

* **Duodenum**, 295 (**Small intestine**); **Glucagon**, **Hormones**, 336; **Hypothalamus**, 303; **Insulin**, 336; **Larynx**, 298; **Sex hormones**, 336.

THE RESPIRATORY SYSTEM

The term **respiration** covers three processes: **ventilation**, or breathing (taking in oxygen and expelling carbon dioxide), **external respiration** (the exchange of gases between the **lungs** and the blood – see also **red blood cells**, page 286) and **internal respiration** (food breakdown, using oxygen and producing carbon dioxide – see pages 334-335). Listed here are the component parts of the human **respiratory system**.

Position of ***respiratory system***

Trachea or windpipe

The main tube through which air passes on its way to and from the **lungs**.

Larynx

The "voice box" at the top of the **trachea**. It contains the **vocal cords** – two pieces of tissue folding inward from the trachea lining and attached to plates of **cartilage***. The opening between the cords is called the **glottis**. During speech, muscles pull the cartilage plates (and hence the cords) together, and air passing out through the cords makes them vibrate, producing sounds.

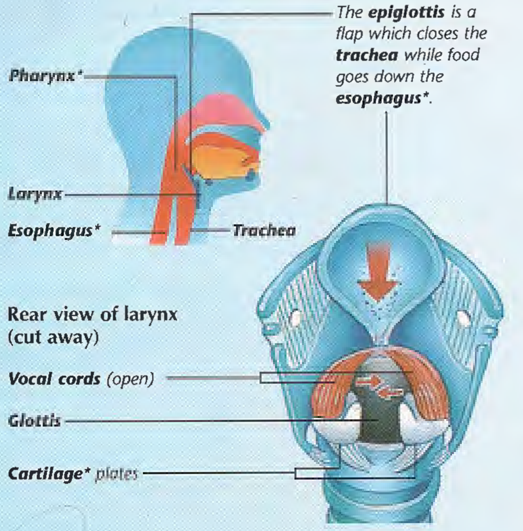

The ***epiglottis*** *is a flap which closes the* ***trachea*** *while food goes down the* ***esophagus****.*

Rear view of larynx (cut away)

Pharynx*
Larynx
Esophagus*
Trachea
Vocal cords (open)
Glottis
Cartilage* plates

Pleura (sing. pleuron) or pleural membrane

A layer of tissue surrounding each **lung** and lining the chest cavity (**thorax**). Between the pleura around a lung and the pleura lining the thorax there is a space (**pleural cavity**). This contains **pleural fluid**. The pleura and fluid-filled cavity make up a cushioning **pleural sac**.

Lungs

The two main breathing organs, inside which gases are exchanged. They contain many tubes (**bronchi** and **bronchioles**) and air sacs (**alveoli**).

Lung (blood vessels not shown)

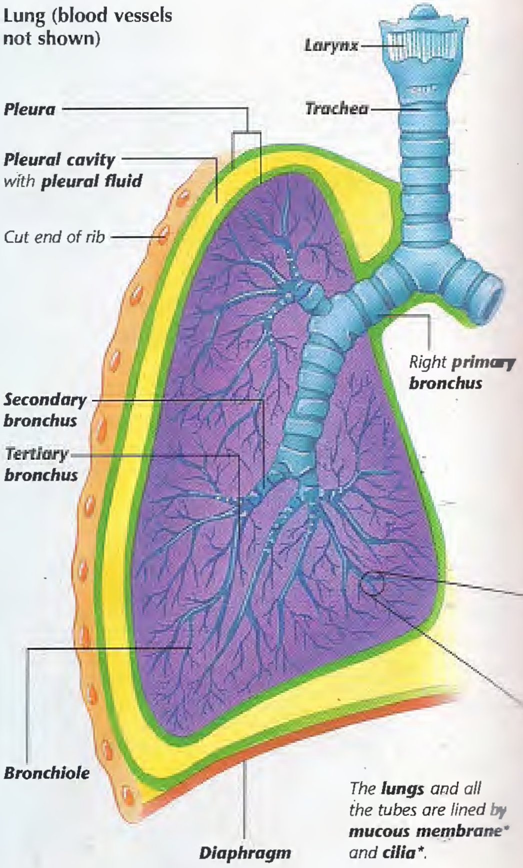

Larynx
Trachea
Pleura
Pleural cavity with pleural fluid
Cut end of rib
Right primary bronchus
Secondary bronchus
Tertiary bronchus
Bronchiole
Diaphragm

The ***lungs*** *and all the tubes are lined by* ***mucous membrane**** *and* ***cilia****.

Diaphragm or midriff

A sheet of muscular tissue which separates the chest from the lower body, or **abdomen**. At rest, it lies in an arched position, forced up by the abdomen wall below it.

---

* **Cartilage**, 281; **Cilia**, 268; **Esophagus**, 294;
Mucous membrane, 295; **Pharynx**, 294.

Bronchi (sing. **bronch**
The main tubes into wh
divides. The first two bra
and left **primary bronch**
a **lung** (via a hole called
pulmonary artery* brin
then branch into **second**
bronchi and **bronchiole**
blood vessels, both bran
pulmonary artery and m
pulmonary veins* (bloo

Bronchioles

The millions of tiny tube
accompanied by blood
of **tertiary bronchi** (see
smaller branches called
each one ending in a clu

Alveoli (sing. **alveolus**
The millions of tiny sacs
terminal bronchioles (se
They are surrounded by
blood vessels) whose blo
carbon dioxide. This pas
the capillary walls, and
the alveoli (to be breath
which has been breathe
passes into the capillarie
to merge together (even
pulmonary veins*).

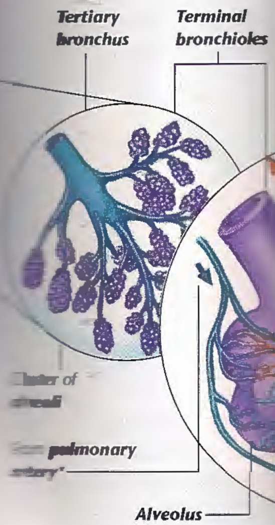

Tertiary bronchus
Terminal bronchioles
Cluster of alveoli
pulmonary artery*
Alveolus

Capillaries, 288; **Medulla**, 303;
Pulmonary arteries, 291 (**Pul**

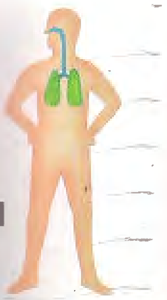

Position of **respiratory system**

s, inside which
contain many
es) and air

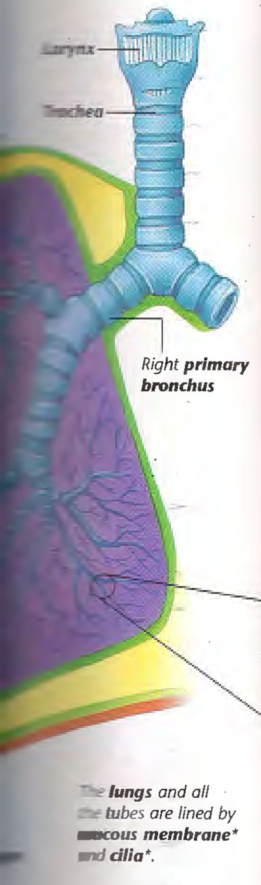

Larynx
Trachea
Right **primary bronchus**

lungs and all
tubes are lined by
cous membrane*
d cilia*.

separates the
domen. At
forced up by

Bronchi (sing. **bronchus**)
The main tubes into which the **trachea** divides. The first two branches are the right and left **primary bronchi**. Each carries air into a **lung** (via a hole called a **hilum**), alongside a **pulmonary artery*** bringing blood in. They then branch into **secondary bronchi**, **tertiary bronchi** and **bronchioles**, all accompanied by blood vessels, both branching from the pulmonary artery and merging to form **pulmonary veins*** (blood going out).

Bronchioles
The millions of tiny tubes in the **lungs**, all accompanied by blood vessels. They branch off **tertiary bronchi** (see **bronchi**) and have smaller branches called **terminal bronchioles**, each one ending in a cluster of **alveoli**.

Alveoli (sing. **alveolus**)
The millions of tiny sacs attached to **terminal bronchioles** (see **bronchioles**). They are surrounded by **capillaries*** (tiny blood vessels) whose blood is rich in carbon dioxide. This passes out through the capillary walls, and in through those of the alveoli (to be breathed out). The oxygen which has been breathed into the alveoli passes into the capillaries, which then begin to merge together (eventually forming **pulmonary veins***).

Bronchial means "of the **bronchi** or **bronchioles**."

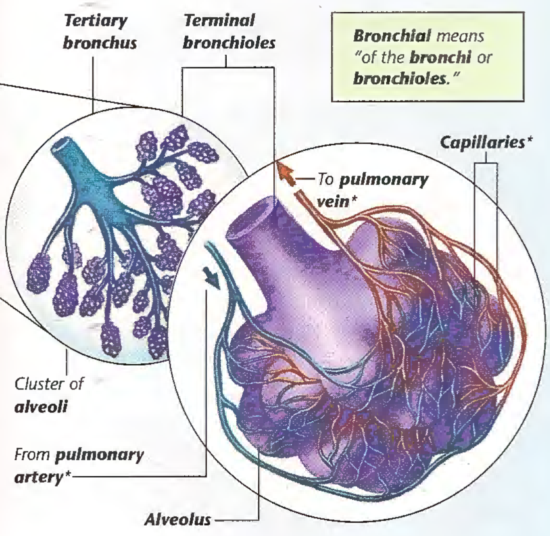

Tertiary bronchus
Terminal bronchioles
Capillaries*
To **pulmonary vein***
Cluster of **alveoli**
From **pulmonary artery***
Alveolus

Breathing

Breathing is made up of **inspiration** (breathing in) and **expiration** (breathing out). Both actions are normally automatic, controlled by nerves from the **respiratory center** in the **medulla*** of the brain. This acts when it detects too high a level of carbon dioxide in the blood.

Inspiration or **inhalation**
The act of breathing in. The **diaphragm** contracts and flattens, lengthening the chest cavity. The muscles between the ribs (**intercostal muscles**) also contract, pulling the ribs up and outward and widening the cavity. The overall expansion lowers the air pressure in the **lungs**, and air rushes in to fill them (i.e. to equalize internal and external pressure).

Inspiration

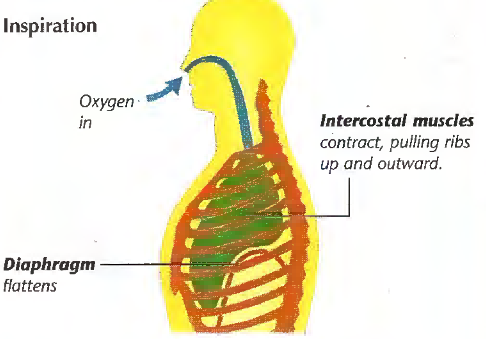

Oxygen in
Intercostal muscles contract, pulling ribs up and outward.
Diaphragm flattens

Expiration or **exhalation**
The act of breathing out. The **diaphragm** and **intercostal muscles** (see **inspiration**) relax, and air is forced out of the **lungs** as the chest cavity becomes smaller.

Expiration

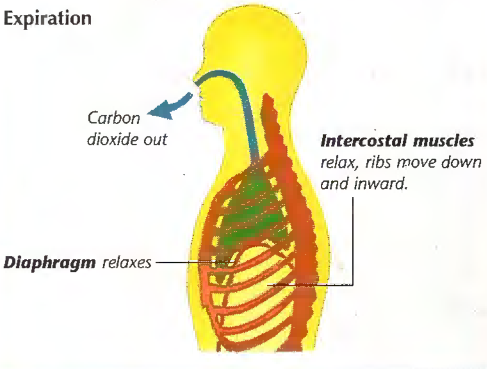

Carbon dioxide out
Intercostal muscles relax, ribs move down and inward.
Diaphragm relaxes

---

* **Capillaries**, 288; **Medulla**, 303;
Pulmonary arteries, 291 (**Pulmonary trunk**); **Pulmonary veins**, 291.

THE URINARY SYSTEM

The **urinary system** is the main system of body parts involved in **excretion**, which is the expulsion of unwanted substances. The parts are defined below. The lungs and skin are also involved in excretion (expelling carbon dioxide and sweat respectively).

Parts of the urinary system

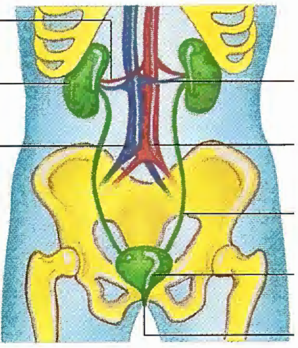

Renal means "of the **kidneys**".

Kidneys

Two organs at the back of the body, just below the ribs. They are the main organs of excretion, filtering out unwanted substances from the blood and regulating the level and contents of body fluids (see also **homeostasis**, page 335). Blood enters a kidney in a **renal artery** and leaves it in a **renal vein**.

Ureters

The two tubes which carry **urine** from the **kidneys** to the **bladder**.

Kidney (cut away)

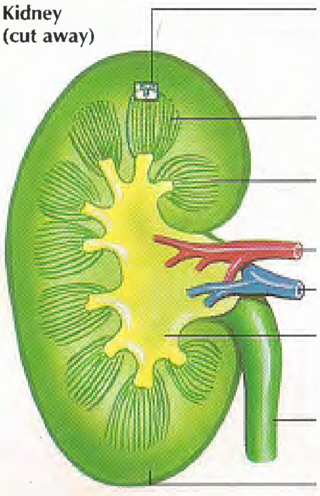

Area shown in picture, page 301.

*The light, inner part is the **medulla** or **renal medulla**. It consists of cone-shaped areas of tissue (**renal** or **medullary pyramids**).*

***Urine** collects in this cavity – the **pelvis** or **renal pelvis**.*

*The dark, outer part is the **cortex** or **renal cortex**.*

Bladder or urinary bladder

A sac which holds stored **urine**. Its lining has many folds (**rugae**, sing. **ruga**) which flatten out as it fills up, letting it expand. Two muscular rings – the **internal** and **external urinary sphincters** – control the opening from the bladder into the **urethra**. When the volume of urine gets to a certain level, nerves stimulate the internal sphincter to open, but the external sphincter is under conscious control (except in young children), and can be held closed for longer.

*Female **bladder** (cut away)*

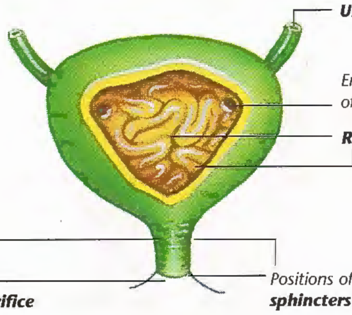

Urethra

The tube carrying **urine** from the **bladder** out of the body (in men, it also carries **sperm*** – see **penis**, page 316). The expulsion of urine is called **urination** or **micturition**.

Urea

A nitrogen-containing (**nitrogenous**) waste substance which is a product of the breakdown of excess **amino acids*** in the liver. It travels in the blood to the **kidneys**, together with smaller amounts of similar substances, e.g. creatinine.

Urine

The liquid which leaves the **kidneys**. Its main constituents are excess water, **urea** and minerals.

---

* **Amino acids**, 330 (**Proteins**); **Aorta**, **Inferior vena cava**, 291; **Sperm**, 320 (**Gametes**).

Inside a kidney

Diagram of a renal corpuscle (cut away)

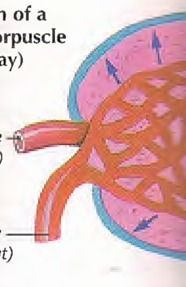

Nephrons

The tiny filtering units c
kidneys (there are abou
million per kidney). Eac
consists of a **renal corp**
and a **uriniferous tubul**

Renal corpuscles or Malpighian corpuscle

The bodies which filter
fluids out of the blood.
consists of a **glomerulu**
a **Bowman's capsule**.

Glomerulus

A ball of coiled-up **capi**
vessels) at the center of
The capillaries branch
entering the corpuscle
and re-unite to leave th
efferent arteriole.

Bowman's capsule

The outer part of each
a thin-walled sac aroun

Uriniferous tubules c

Long tubes, each one l
Bowman's capsule. Eac
– the **proximal convolu**
Henlé and the **distal co**
has many **capillaries*** (
twined around it. These
efferent arteriole (see g
re-unite to form larger
blood from the **kidney**.

Collecting duct or co

A tube which carries ur
uriniferous tubules into

* **ADH**, **Aldosterone**, 336; **Amin**
Capillaries, 288; **Hormones**, 3

Inside a kidney

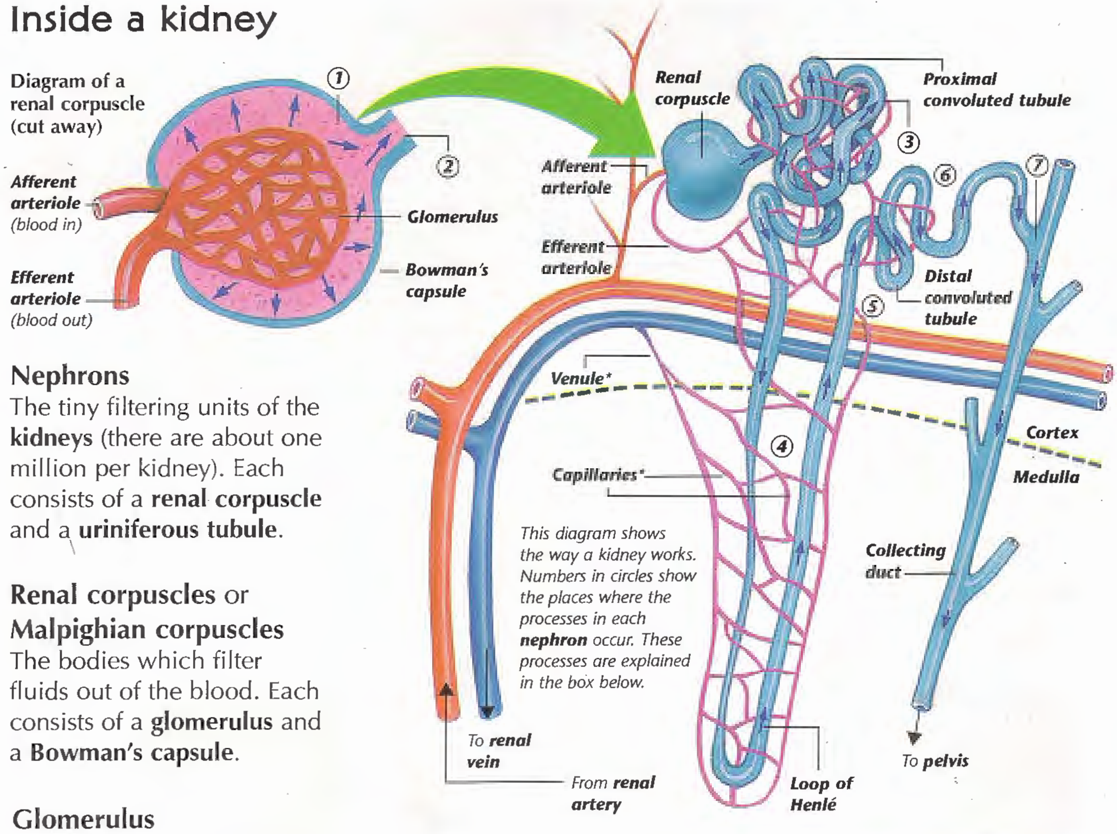
Diagram of a
renal corpuscle
(cut away)
①
②
Afferent arteriole (blood in)
Glomerulus
Efferent arteriole (blood out)
Bowman's capsule
Renal corpuscle
Proximal convoluted tubule
③
Afferent arteriole
Efferent arteriole
⑥
⑦
Distal convoluted tubule
⑤
Venule*
Cortex
④
Medulla
Capillaries*
Collecting duct
To renal vein
From renal artery
Loop of Henlé
To pelvis

This diagram shows the way a kidney works. Numbers in circles show the places where the processes in each **nephron** occur. These processes are explained in the box below.

Nephrons

The tiny filtering units of the **kidneys** (there are about one million per kidney). Each consists of a **renal corpuscle** and a **uriniferous tubule**.

Renal corpuscles or Malpighian corpuscles

The bodies which filter fluids out of the blood. Each consists of a **glomerulus** and a **Bowman's capsule**.

Glomerulus

A ball of coiled-up **capillaries*** (tiny blood vessels) at the center of each **renal corpuscle**. The capillaries branch from an **arteriole*** entering the corpuscle (an **afferent arteriole**) and re-unite to leave the corpuscle as an **efferent arteriole**.

Bowman's capsule

The outer part of each **renal corpuscle**. It is a thin-walled sac around the **glomerulus**.

Uriniferous tubules or renal tubules

Long tubes, each one leading from a **Bowman's capsule**. Each has three main parts – the **proximal convoluted tubule**, the **loop of Henlé** and the **distal convoluted tubule** – and has many **capillaries*** (tiny blood vessels) twined around it. These are branches of the **efferent arteriole** (see **glomerulus**) and they re-unite to form larger blood vessels carrying blood from the **kidney**.

Collecting duct or collecting tubule

A tube which carries **urine** from several **uriniferous tubules** into the **pelvis** of a **kidney**.

Key to kidney diagram above

1. ***Glomerular filtration**. As blood squeezes through the **glomerulus**, most of its water, minerals, vitamins, glucose, **amino acids*** and **urea** are forced into the **Bowman's capsule**, forming **glomerular filtrate**.*

2. ***Glomerular filtrate** moves into **proximal convoluted tubule**.*

3. ***Tubular reabsorption**. As **glomerular filtrate** runs through the **uriniferous tubule**, most vitamins, glucose and **amino acids*** are taken back into the blood in the twining **capillaries***.*

4. *Some minerals are also taken back. The **hormone* aldosterone*** controls reabsorption of more if needed.*

5. *Some water is also taken back. The **hormone* ADH*** controls reabsorption of more if needed.*

6. ***Tubular secretion**. Some substances, e.g. ammonia and some drugs, pass from the blood into the **uriniferous tubule**.*

7. *Resulting **urine** passes into **collecting duct**.*

***Distal** means "away from the point of origin or attachment".*

***Proximal** means "near the point of origin or attachment".*

---

* **ADH**, **Aldosterone**, 336; **Amino acids**, 330 (**Proteins**); **Arteriole**, 288 (**Arteries**); **Capillaries**, 288; **Hormones**, 336; **Venule**, 288 (**Veins**).

THE CENTRAL NERVOUS SYSTEM

The **central nervous system** (**CNS**) is the body's control center. It co-ordinates all its actions, both mechanical and chemical (working with **hormones***) and is made up of the **brain** and **spinal cord**. The millions of nerves in the body carry "messages" (nervous impulses) to and from these central areas (see pages 306-309).

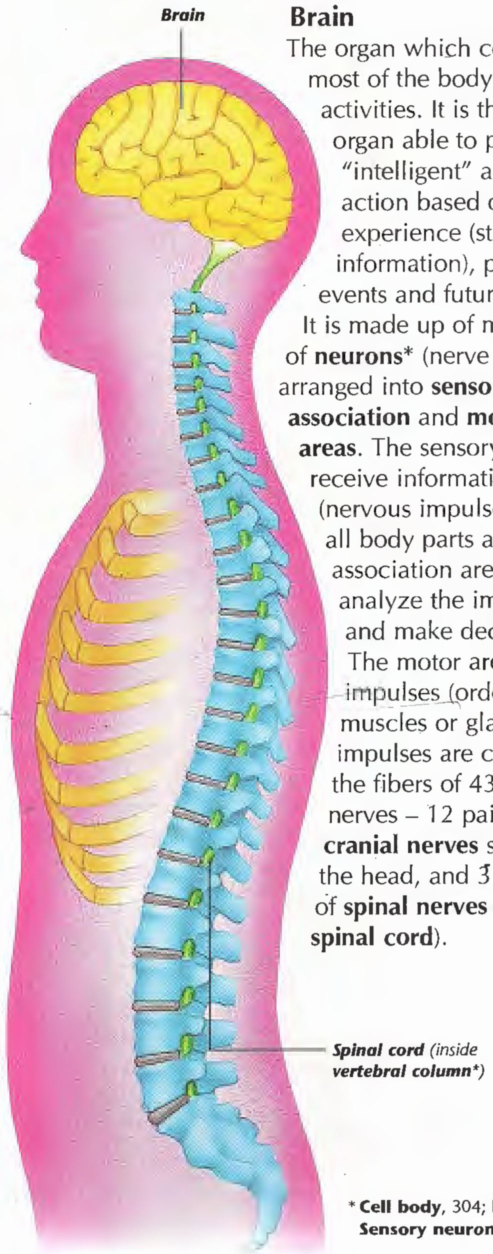

Brain

The organ which controls most of the body's activities. It is the only organ able to produce "intelligent" action – action based on past experience (stored information), present events and future plans. It is made up of millions of **neurons*** (nerve cells), arranged into **sensory**, **association** and **motor areas**. The sensory areas receive information (nervous impulses) from all body parts and the association areas analyze the impulses and make decisions. The motor areas send impulses (orders) to muscles or glands. The impulses are carried by the fibers of 43 pairs of nerves – 12 pairs of **cranial nerves** serving the head, and 31 pairs of **spinal nerves** (see **spinal cord**).

Spinal cord

A long string of nervous tissue running down from the **brain** inside the **vertebral column***. Nervous impulses from all parts of the body pass through it. Some are carried into or away from the brain, some are dealt with in the cord (see **involuntary actions**, page 309). 31 pairs of **spinal nerves** branch out from the cord through the gaps between the **vertebrae***. Each spinal nerve is made up of two groups of fibers: a **dorsal** or **sensory root**, made up of the fibers of **sensory neurons*** (bringing impulses in), and a **ventral** or **motor root**, made up of the fibers of **motor neurons*** (taking impulses out).

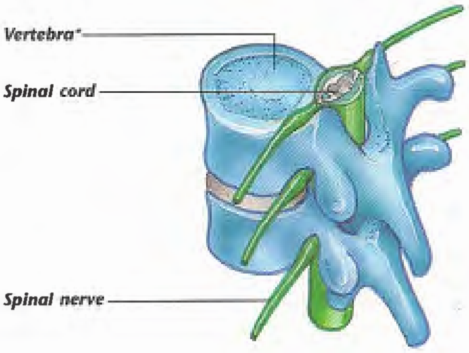

Neuroglia or glia

Stiffened cells which support and protect the nerve cells (**neurons***) of the central nervous system. Some produce a white, fatty substance called **myelin** (see also **Schwann cells**, page 304). This coats the long fibers found in the connective areas of the **brain** and the outer layer of the **spinal cord**, and leads to these areas being known as **white matter**. **Gray matter**, by contrast, consists mainly of **cell bodies*** and their short fibers, and its neuroglia do not produce myelin.

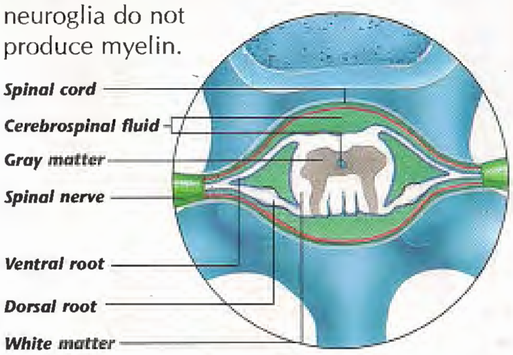

---

* **Cell body**, 304; **Hormones**, 336; **Motor neurons**, 305; **Neurons**, 304; **Sensory neurons**, 305; **Vertebrae**, **Vertebral column**, 279.

The parts of t

Cerebrum

The largest, most high
with many deep folds
two **cerebral hemisph**
corpus callosum (a ba
fibers*), and its outer
called the **cerebral co**
This contains all the
most important **senso**
association and **moto**
areas (see **brain**). It
controls most physica
activities and is the
center for mental
activities such as
decision-making,
speech, learning,
memory and imaginat

Cerebellum

The area which
coordinates muscle
movement and balanc
two things under the
overall control of the
cerebrum.

Midbrain or mesenc

An area joining the **di**
pons. It carries impuls
thalamus, and out fro
toward the **spinal cor**

Pons or pons Varoli

A junction of **nerve fi**
link between the parts
the **spinal cord** (via th

*In general terms, **cerebra***

Areas of cerebrum
🟥 ***Sensory areas**. Receive in*
1. General sensory area. R
*2. **Primary gustatory ar***
*3. **Primary auditory are***
*4. **Primary visual area.***
*5. **Primary olfactory are***

🟦 ***Motor areas**. Each tiny p*

🟨 ***Association areas**. Interp*
*6. **Visual association ar***
*7. **Auditory association***

Autonomic nervous system,
Nerve fibers, 304; **Pituitary**

The parts of the brain

Cerebrum

The largest, most highly developed area, with many deep folds. It is composed of two **cerebral hemispheres**, joined by the **corpus callosum** (a band of **nerve fibers***), and its outer layer is called the **cerebral cortex**. This contains all the most important **sensory**, **association** and **motor areas** (see **brain**). It controls most physical activities and is the center for mental activities such as decision-making, speech, learning, memory and imagination.

Cerebellum

The area which coordinates muscle movement and balance, two things under the overall control of the **cerebrum**.

Thalamus

The area which carries out the first, basic sorting of incoming impulses and directs them to different parts of the **cerebrum**. It also directs some outgoing impulses.

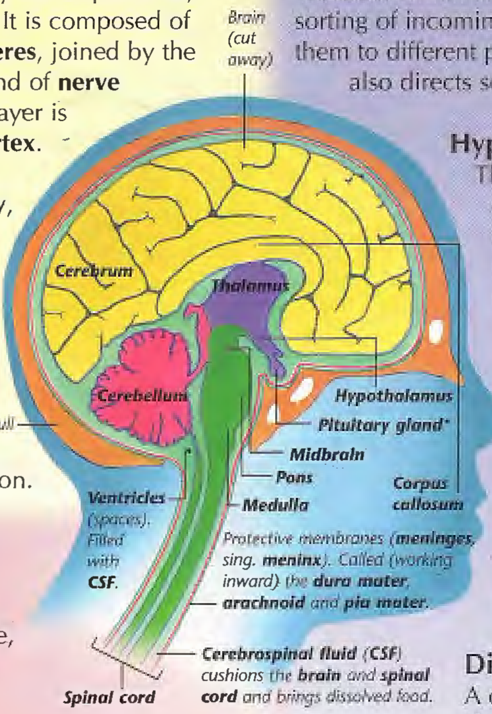
Brain (cut away)
Cerebrum
Thalamus
Cerebellum
Skull
Hypothalamus
Pituitary gland*
Midbrain
Pons
Corpus callosum
Medulla
Ventricles (spaces). Filled with CSF.
Protective membranes (meninges, sing. meninx). Called (working inward) the dura mater, arachnoid and pia mater.
Cerebrospinal fluid (CSF) cushions the brain and spinal cord and brings dissolved food.
Spinal cord

Hypothalamus

The master controller of most inner body functions. It controls the **autonomic nervous system*** (the nerve cells causing unconscious action, e.g. food movement through the intestines) and the action of the **pituitary gland***. Its activities are vital to **homeostasis*** – the maintenance of stable internal conditions.

Diencephalon

A collective term for the **thalamus** and **hypothalamus**.

Midbrain or mesencephalon

An area joining the **diencephalon** to the **pons**. It carries impulses in toward the **thalamus**, and out from the **cerebrum** toward the **spinal cord**.

Pons or pons Varolii

A junction of **nerve fibers*** which forms a link between the parts of the **brain** and the **spinal cord** (via the **medulla**).

Medulla or medulla oblongata

The area which controls the "fine tuning" of many unconscious actions (under the overall control of the **hypothalamus**). Different parts of it control different actions, e.g. the **respiratory center** controls breathing.

Brain stem

A collective term for the **midbrain**, **pons** and **medulla**.

*In general terms, **cerebral** means "of the **brain**".*

***Cephalic** means "of the head".*

Areas of cerebrum

🟥 ***Sensory areas**. Receive incoming impulses.*

1. *General sensory area. Receives impulses from muscles, skin and inner organs.*
2. ***Primary gustatory area**. Impulses from tongue.*
3. ***Primary auditory area**. Impulses from ears.*
4. ***Primary visual area**. Impulses from eyes.*
5. ***Primary olfactory area**. Impulses from nose.*

🟦 ***Motor areas**. Each tiny part sends out impulses to a specific muscle.*

🟨 ***Association areas**. Interpret impulses and make decisions. Some specific ones are:*

6. ***Visual association area**. Produces sight.*
7. ***Auditory association area**. Produces hearing.*

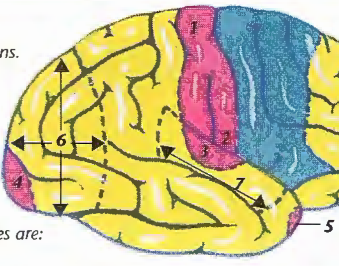

---

* **Autonomic nervous system**, 308; **Homeostasis**, 335; **Nerve fibers**, 304; **Pituitary gland**, 297.

THE UNITS OF THE NERVOUS SYSTEM

The individual units of both the brain and spinal cord (**central nervous system***) and the nerves of the rest of the body (**peripheral nervous system**) are the nerve cells, or **neurons**. They are unique in being able to transmit electrical "messages" (the vital nervous impulses) around the body. Each neuron consists of a **cell body**, an **axon** and one or more **dendrites**, and there are three types of neuron – **sensory**, **association** and **motor neurons**.

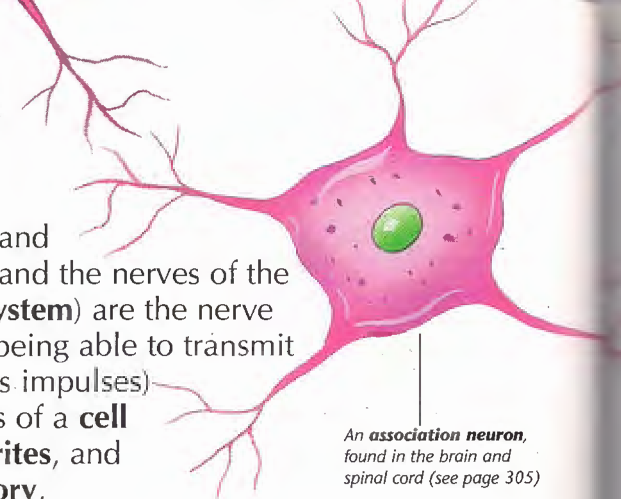

*An **association neuron**, found in the brain and spinal cord (see page 305)*

The parts of a neuron

Cell body or perikaryon

The part of a neuron containing its **nucleus*** and most of its **cytoplasm***. The cell bodies of all **association**, some **sensory** and some **motor neurons** lie in the brain and spinal cord. Those of the other sensory neurons are found in masses called **ganglia*** or as part of highly specialized **receptors*** in the nose and eyes. Those of the other motor neurons lie in **autonomic ganglia***.

Nerve fibers

The fibers (**axon** and **dendrites**) of a neuron. They are extensions of the **cytoplasm*** of the **cell body** and carry the vital nervous impulses. Most of the long nerve fibers which run out round the body (belonging to **sensory** or **motor neurons**) are accompanied by **neuroglial*** cells. These are called **Schwann cells** and they produce a sheath of **myelin*** around each fiber.

Dendrites

The **nerve fibers** carrying impulses toward a **cell body**. Most neurons have several short dendrites, but one type of **sensory neuron** has just one, elongated dendrite, often called a **dendron**. The endings of these dendrons form **receptors*** all over the body, and the dendrons themselves run inward to the cell bodies (which are found in **ganglia*** just outside the spinal cord).

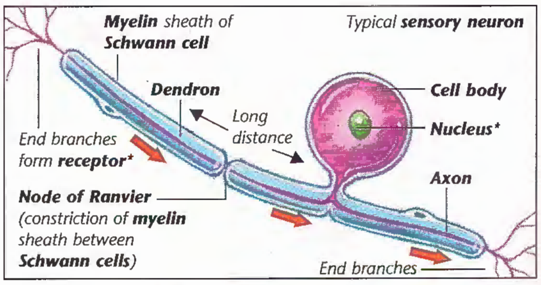
Typical **sensory neuron**
Myelin sheath of **Schwann cell**
Dendron
Long distance
Cell body
Nucleus*
End branches form **receptor***
Axon
Node of Ranvier (constriction of **myelin** sheath between **Schwann cells**)
End branches

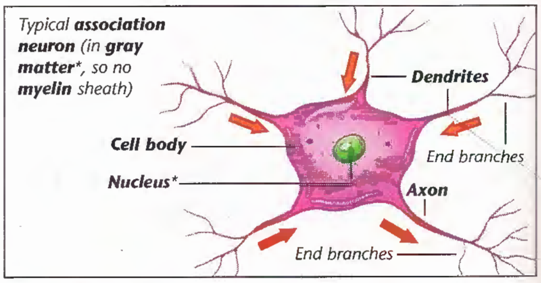
Typical **association neuron** (in **gray matter***, so no **myelin** sheath)
Dendrites
Cell body
End branches
Nucleus*
Axon
End branches

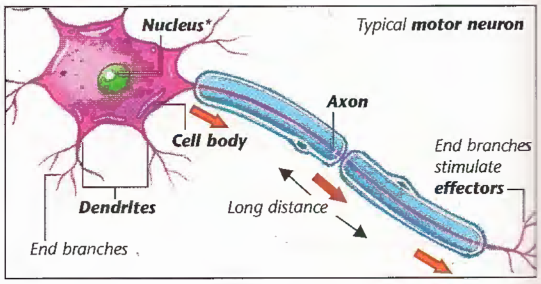
Nucleus*
Typical **motor neuron**
Axon
Cell body
End branches stimulate **effectors**
Dendrites
Long distance
End branches

Axon

The single long **nerve fiber** which carries impulses away from a **cell body**. The axons of all **association** and **sensory neurons** and some **motor neurons** lie in the brain and spinal cord. Those of the other motor neurons run out of the spinal cord to **autonomic ganglia***, or further to **effectors** (see **motor neurons**).

---

*Autonomic ganglia**, 309; **Central nervous system**, 302; **Cytoplasm**, 238; **Ganglia**, 306; **Grey matter**, **Myelin**, 302 (**Neuroglia**); **Nucleus**, 238; **Receptors**, 307.

Types of neur

Sensory neurons or **a**
The neurons which can
(nervous impulses) abo
The single **dendrites** (**d**
some sensory neurons
throughout the body, ar
endings fire off impulse
stimulated. For more **ab**
these endings (**receptor**
the different sensory ne
see pages 306-307.

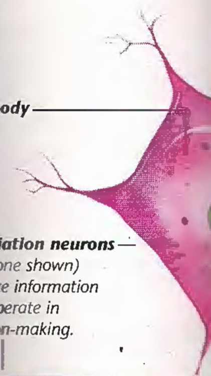

Association neurons — (only one shown) analyze information and **operate** in decision-making.

Association neurons

Also called **relay**, **intern**
connecting neurons, or
Special linking neurons,
vast numbers in the brai
cord. They are involved
impulses (from **sensory**
interpreting the sensory
and passing impulses to
neurons to initiate actio

Receptors, 307.

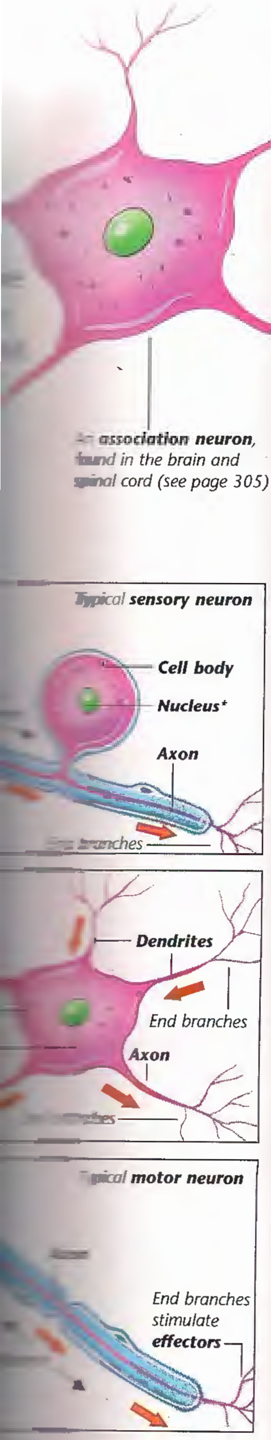
association neuron, in the brain and cord (see page 305)
sensory neuron
Cell body
Nucleus*
Axon
Dendrites
End branches
Axon
motor neuron
End branches stimulate effectors

carries
The axons of
and some
spinal cord.
run out of the
or further to

Types of neuron

Sensory neurons or afferent neurons

The neurons which carry "information" (nervous impulses) about sensations. The single **dendrites** (**dendrons**) of some sensory neurons run throughout the body, and their endings fire off impulses when stimulated. For more about these endings (**receptors**) and the different sensory neurons, see pages 306-307.

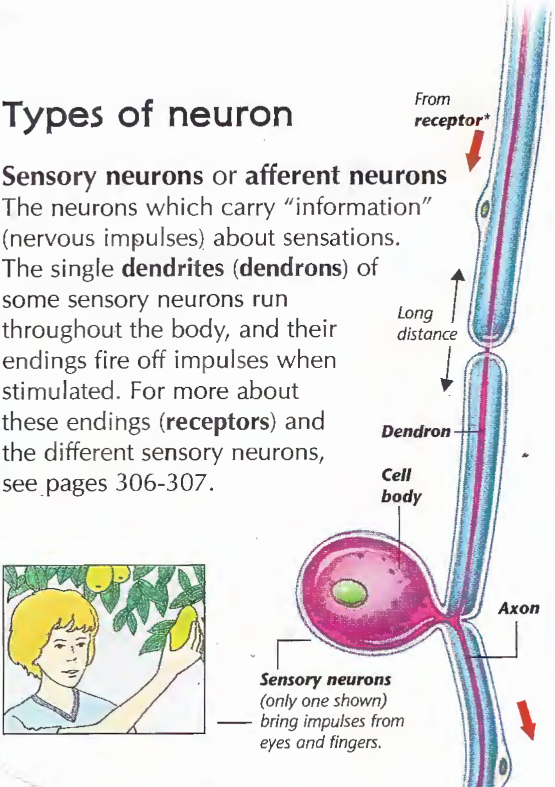
From receptor*
Long distance
Dendron
Cell body
Axon
Sensory neurons (only one shown) bring impulses from eyes and fingers.

Synapses

The tiny areas where the branching ends of the **axon** of one neuron meet the **dendrites** of the next. When an impulse reaches the end of the axon, a chemical called a **neurotransmitter** is released into the minute gap (**synaptic cleft** or **gap**) found at the junction. When enough of this has reached the other side, an impulse is sent on in the dendrites.

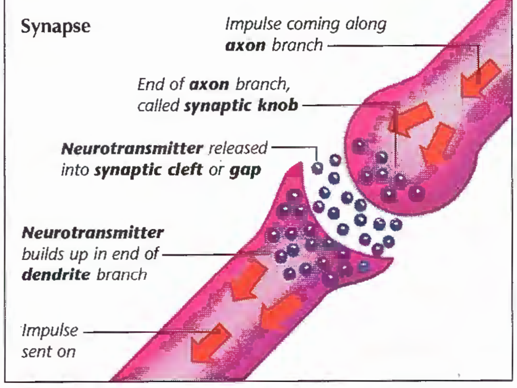
Synapse
Impulse coming along axon branch
End of axon branch, called synaptic knob
Neurotransmitter released into synaptic cleft or gap
Neurotransmitter builds up in end of dendrite branch
Impulse sent on

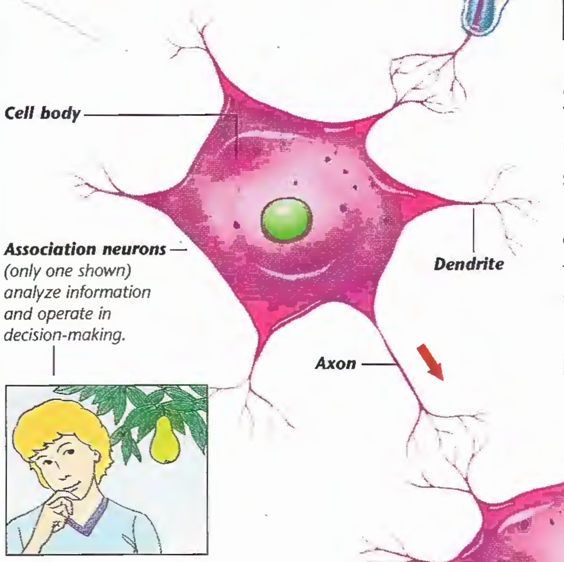
Cell body
Association neurons (only one shown) analyze information and operate in decision-making.
Dendrite
Axon

Motor neurons or efferent neurons

The neurons which carry "instructions" (nervous impulses) away from the brain and spinal cord. The ends of the **axons** of some motor neurons make connections with muscles or glands (called **effectors**), and the impulses they carry (passed onto them from **association neurons**) stimulate these organs into action. For more about the different motor neurons, see pages 308-309.

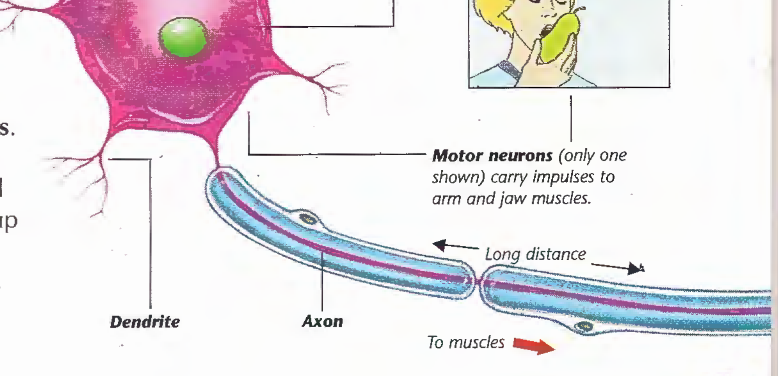
Cell body
Motor neurons (only one shown) carry impulses to arm and jaw muscles.
Long distance
Dendrite
Axon
To muscles

Association neurons

Also called **relay**, **internuncial** or **connecting neurons**, or **interneurons**. Special linking neurons, present in vast numbers in the brain and spinal cord. They are involved in picking up impulses (from **sensory neurons**), interpreting the sensory information, and passing impulses to **motor neurons** to initiate actions.

---

* **Receptors**, 307.

NERVES AND NERVOUS PATHWAYS

The **sensitivity** (**irritability**) of the body (its ability to respond to stimuli) relies on the transportation of "messages" (nervous impulses) by the fibers of nerve cells (**neurons***). The fibers which bring impulses into the brain and spinal cord are part of the **afferent system**. Those which carry impulses from the brain and cord are part of the **efferent system** (see pages 308-309). The fibers outside the brain and cord make up the **nerves** of the body, known collectively as the **peripheral nervous system** (**PNS**).

Nerves

Bundles of nerve fibers, blood vessels and **connective tissue***. Each nerve consists of several bundles (**fascicles**) of fibers and each fiber is part of a nerve cell (**neuron***). **Sensory nerves** have just the fibers (**dendrons***) of **sensory** (**afferent**) **neurons***, **motor nerves** have just the fibers (**axons***) of **motor** (**efferent**) **neurons***, and **mixed nerves** have both types of fiber.

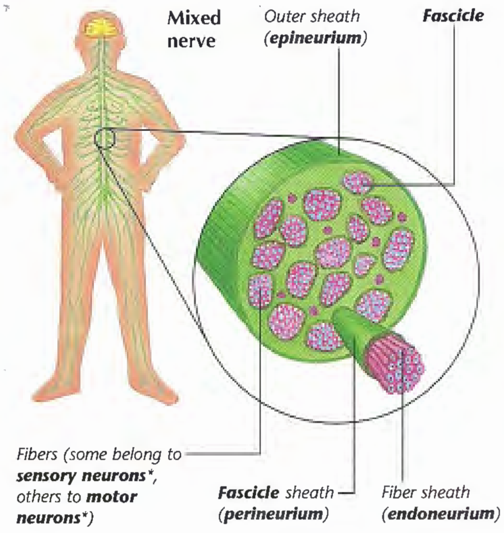
Mixed nerve
Outer sheath (***epineurium***)
Fascicle
Fibers (some belong to ***sensory neurons****, others to* ***motor neurons****)*
Fascicle *sheath* (***perineurium***)
Fiber sheath (***endoneurium***)

The afferent system

The **afferent system** is the system of nerve cells (**neurons***) whose fibers carry sensory information (nervous impulses) toward the spinal cord, up inside it and into the brain. The nerve cells involved are all the **sensory** (**afferent**) **neurons*** of the body. The impulses originate in **receptors** and are interpreted by the brain as sensations.

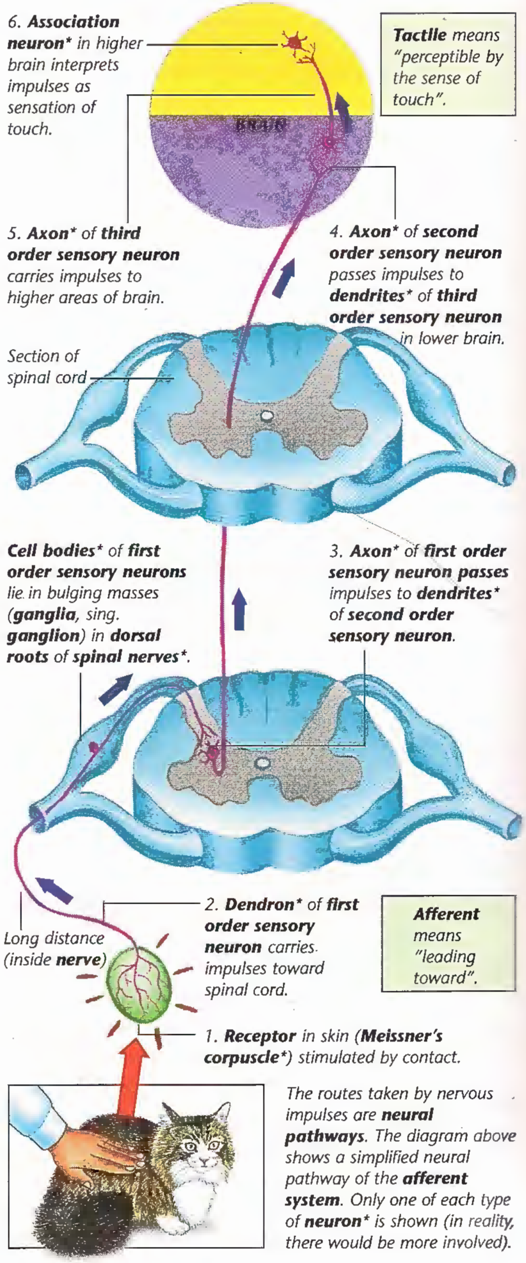
6. ***Association neuron**** *in higher brain interprets impulses as sensation of touch.*
BRAIN
5. ***Axon**** *of* ***third order sensory neuron*** *carries impulses to higher areas of brain.*
4. ***Axon**** *of* ***second order sensory neuron*** *passes impulses to* ***dendrites**** *of* ***third order sensory neuron*** *in lower brain.*
Section of spinal cord
Cell bodies* *of* ***first order sensory neurons*** *lie in bulging masses (****ganglia****, sing.* ***ganglion****) in* ***dorsal roots*** *of* ***spinal nerves****.*
3. ***Axon**** *of* ***first order sensory neuron passes*** *impulses to* ***dendrites**** *of* ***second order sensory neuron****.*
Long distance (inside ***nerve****)*
2. ***Dendron**** *of* ***first order sensory neuron*** *carries impulses toward spinal cord.*
1. ***Receptor*** *in skin (****Meissner's corpuscle*****) stimulated by contact.*

Tactile *means "perceptible by the sense of touch".*

Afferent *means "leading toward".*

The routes taken by nervous impulses are ***neural pathways****. The diagram above shows a simplified neural pathway of the* ***afferent system****. Only one of each type of* ***neuron**** *is shown (in reality, there would be more involved).*

---

* **Association neurons**, 305; **Axon**, **Cell body**, 304; **Connective tissue**, 280; **Dendron**, 304 (**Dendrites**); **Meissner's corpuscles**, 311; **Motor neurons**, 305; **Neurons**, 304; **Sensory neurons**, 305; **Spinal nerves**, 302 (**Spinal cord**).

Receptors

The parts of the **afferen**
nervous impulses whe
Most are either the sing
the long **dendron*** of a
neuron (see picture) or
endings. They are all en
tissue, and many have s
formed around them (e.
tongue). They are found
both near the surface (i
organs, **skeletal muscle**
inside (connected to inn
vessel walls, etc.).

Sense organs

The highly specialized s
body, each with many r
nose, **tongue**, eyes and
eyes and ears, see page

Divisions of the afferent syste

Central *a*
periphera
systems.
cells in bo
Afferent system. *Nerve cells bringing impulses in and up.*
Somatic afferent system. *Nerve cells bringing impulses from* ***receptors*** *near body surface and* ***skeletal muscles****.*

Association neurons, 305; **Ax**
Mucous membrane, 295; **Pha**

HWAYS

stimuli) the fibers the brain carry (see the nerves (PNS).

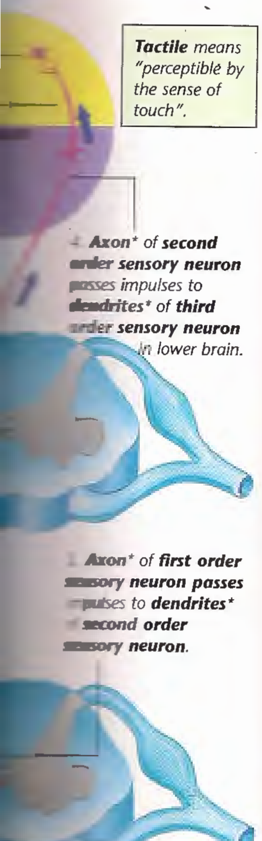

Tactile means "perceptible by the sense of touch".
Axon* of second sensory neuron passes impulses to dendrites* of third sensory neuron in lower brain.
Axon* of first order neuron passes to dendrites* second order neuron.

Afferent means "leading toward".

Meissner's by contact.

taken by nervous neural The diagram above implified neural the afferent one of each type is shown (in reality, be more involved).

Spinal cord).

Receptors

The parts of the **afferent system** which fire off nervous impulses when they are stimulated. Most are either the single branched ending of the long **dendron*** of a **first order sensory neuron** (see picture) or a group of such endings. They are all embedded in body tissue, and many have some kind of structure formed around them (e.g. a **taste bud** – see **tongue**). They are found all over the body, both near the surface (in the skin, **sense organs**, **skeletal muscles***, etc.) and deeper inside (connected to inner organs, blood vessel walls, etc.).

Sense organs

The highly specialized sensory organs of the body, each with many **receptors**. They are the **nose**, **tongue**, eyes and ears. For more about eyes and ears, see pages 312-315.

Divisions of the afferent system

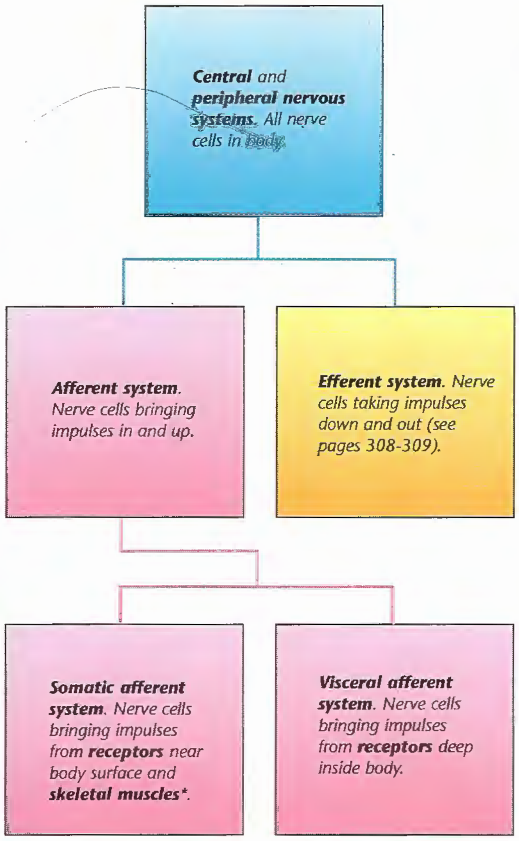

***Central** and **peripheral nervous systems**. All nerve cells in body.*

***Afferent system**. Nerve cells bringing impulses in and up.*

***Efferent system**. Nerve cells taking impulses down and out (see pages 308-309).*

Somatic afferent system**. Nerve cells bringing impulses from **receptors** near body surface and **skeletal muscles.*

***Visceral afferent system**. Nerve cells bringing impulses from **receptors** deep inside body.*

Nose

The organ of smell. Each of its two nostrils opens out into a **nasal cavity** which is lined with **mucous membrane*** and has many **olfactory hairs** extending from its roof. The hairs are the **dendrites*** of special **sensory neurons*** called **olfactory cells**. These are the **receptors** whose impulses are interpreted by the brain as sensations of smell (**olfactory sensations**).

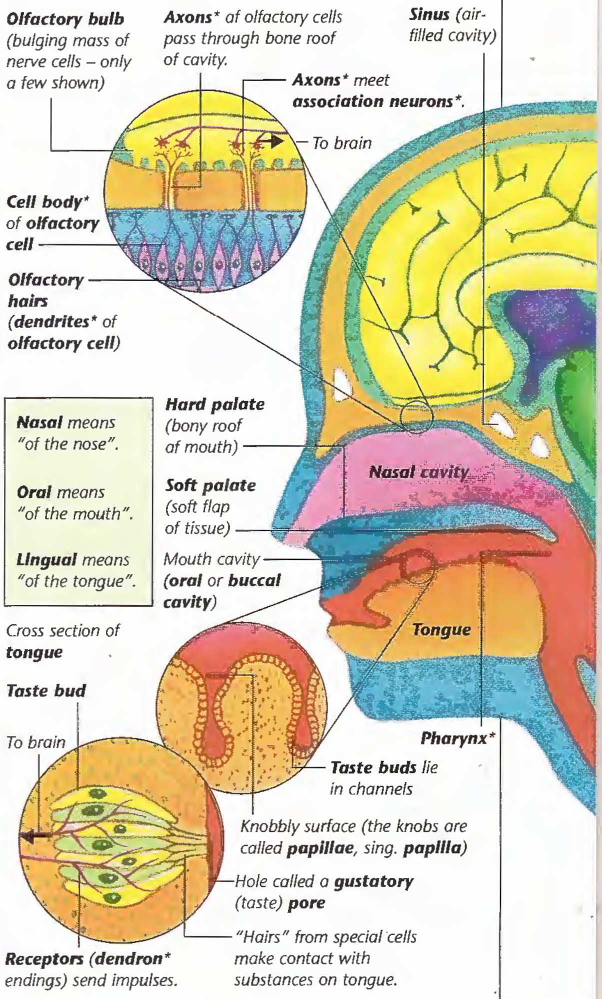

Olfactory bulb (bulging mass of nerve cells – only a few shown)
Axons* of olfactory cells pass through bone roof of cavity.
Sinus (air-filled cavity)
Axons* meet association neurons*.
To brain
Cell body* of olfactory cell
Olfactory hairs (dendrites* of olfactory cell)
Hard palate (bony roof of mouth)
Nasal cavity
Soft palate (soft flap of tissue)
Mouth cavity (oral or buccal cavity)
Tongue
Cross section of tongue
Taste bud
To brain
Pharynx*
Taste buds lie in channels
Knobbly surface (the knobs are called papillae, sing. papilla)
Hole called a gustatory (taste) pore
"Hairs" from special cells make contact with substances on tongue.
Receptors (dendron* endings) send impulses.

***Nasal** means "of the nose".*

***Oral** means "of the mouth".*

***Lingual** means "of the tongue".*

Tongue

The main organ of taste. It is a muscular organ which bears many **taste buds**. These tiny bodies contain the **receptors** whose impulses are interpreted by the brain as taste sensations (**gustatory sensations**).

---

* **Association neurons**, 305; **Axon**, **Cell body**, 304; **Dendron**, 304 (**Dendrites**); **Mucous membrane**, 295; **Pharynx**, 294; **Sensory neurons**, 305; **Skeletal muscles**, 282.

The efferent system

The **efferent system** is the second system of nerve cells (**neurons***) in the body (see also **afferent system**, pages 306-307). The fibers of its nerve cells carry nervous impulses away from the brain, down through the spinal cord and out around the body. The nerve cells involved are all the **motor** (**efferent**) **neurons*** of the body. The impulses they carry stimulate action in the surface muscles (**skeletal muscles***) or in the glands and internal muscles (in the walls of inner organs and blood vessels). All these organs are known collectively as **effectors**.

Divisions of the efferent system

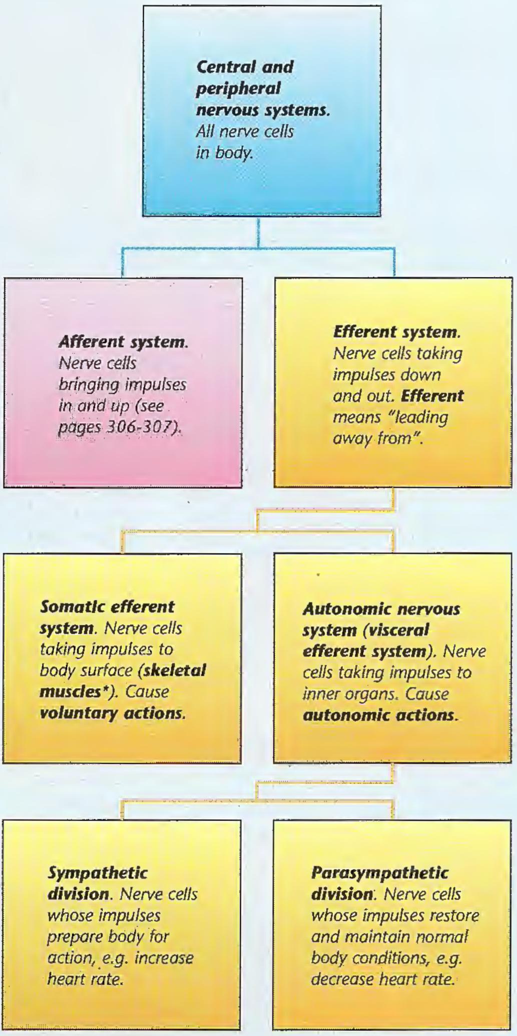

Central and peripheral nervous systems. *All nerve cells in body.*

Afferent system. *Nerve cells bringing impulses in and up (see pages 306-307).*

Efferent system. *Nerve cells taking impulses down and out.* ***Efferent*** *means "leading away from".*

Somatic efferent system. *Nerve cells taking impulses to body surface (**skeletal muscles***). Cause* ***voluntary actions.***

Autonomic nervous system (visceral efferent system). *Nerve cells taking impulses to inner organs. Cause* ***autonomic actions.***

Sympathetic division. *Nerve cells whose impulses prepare body for action, e.g. increase heart rate.*

Parasympathetic division. *Nerve cells whose impulses restore and maintain normal body conditions, e.g. decrease heart rate.*

The different actions

Voluntary actions

Actions which result from conscious activity by the brain, i.e. ones it consciously decides upon, e.g. lifting a cup. We are always aware of these actions, which involve **skeletal muscles*** only. The impulses which cause them originate in higher areas of the brain (especially the **cerebrum***) and are carried by nerve cells of the **somatic efferent system**.

Lifting a cup is a ***voluntary action.***

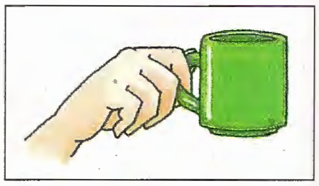

Simplified neural pathway* of voluntary action (somatic efferent system)

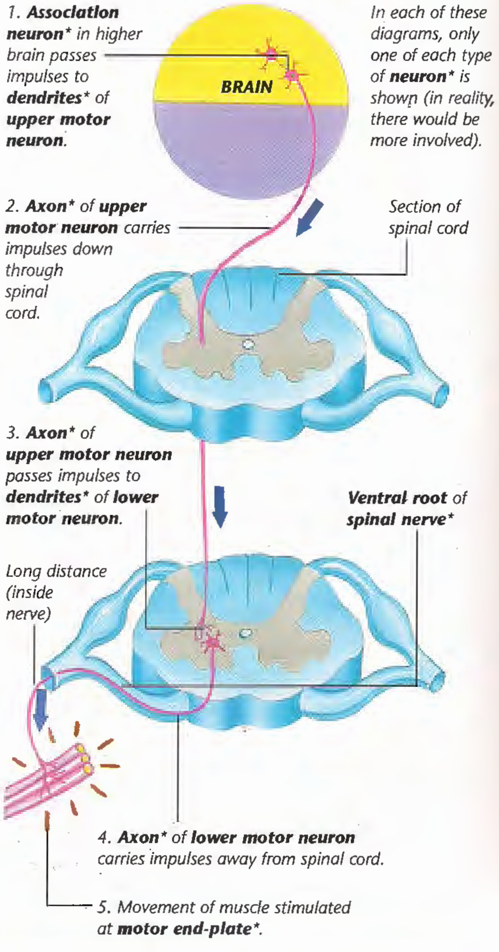
1. **Association neuron*** in higher brain passes impulses to **dendrites*** of **upper motor neuron**.
BRAIN
In each of these diagrams, only one of each type of **neuron*** is shown (in reality, there would be more involved).
2. **Axon*** of **upper motor neuron** carries impulses down through spinal cord.
Section of spinal cord
3. **Axon*** of **upper motor neuron** passes impulses to **dendrites*** of **lower motor neuron**.
Ventral root of **spinal nerve***
Long distance (inside nerve)
4. **Axon*** of **lower motor neuron** carries impulses away from spinal cord.
5. Movement of muscle stimulated at **motor end-plate***.

Involuntary actions

Automatic actions (one consciously decide up types. Firstly, there are of inner organs, e.g. th of which we are not no impulses which cause lower brain (especially and are carried by nerv **autonomic nervous sys** **autonomic actions**. The actions are **reflex actio**

The heartbeat is an ***autonomic action.***

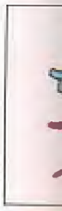

Simplified neural pathway* of autonomic action (sympathetic division of autonomic nervous system)

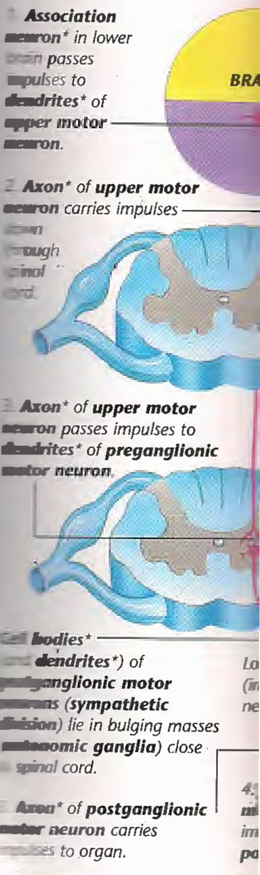
Association neuron* *in lower brain passes impulses to* ***dendrites**** *of* ***upper motor neuron.***
BRA
Axon* *of* ***upper motor neuron*** *carries impulses down through spinal cord.*
Axon* *of* ***upper motor neuron*** *passes impulses to* ***dendrites**** *of* ***preganglionic motor neuron.***
Cell ***bodies**** *(and* ***dendrites****) of* ***postganglionic motor neurons (sympathetic division)*** *lie in bulging masses* ***(autonomic ganglia)*** *close to spinal cord.*
Axon* *of* ***postganglionic motor neuron*** *carries impulses to organ.*
Lo (in ne
4. m im po

---

*Association neurons, 305; **Axon**, 304; **Cerebrum**, 303; **Dendrites**, 304; **Motor end-plate**, 283; **Motor neurons**, 305; **Neural pathways**, 306; **Neurons**, 304; **Skeletal muscles**, 282; **Spinal nerves**, 302 (**Spinal cord**).

Association neurons, 305; **Axc** **Motor end-plate**, 283; **Neural**

Involuntary actions

Automatic actions (ones the brain does not consciously decide upon). There are two types. Firstly, there are the constant actions of inner organs, e.g. the beating of the heart, of which we are not normally aware. The impulses which cause them originate in the lower brain (especially the **hypothalamus***) and are carried by nerve cells of the **autonomic nervous system**. They are called **autonomic actions**. The other involuntary actions are **reflex actions**.

The heartbeat is an ***autonomic action****.*

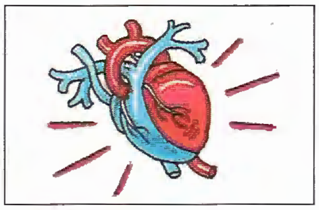

Simplified neural pathway* of autonomic action (sympathetic division of autonomic nervous system)

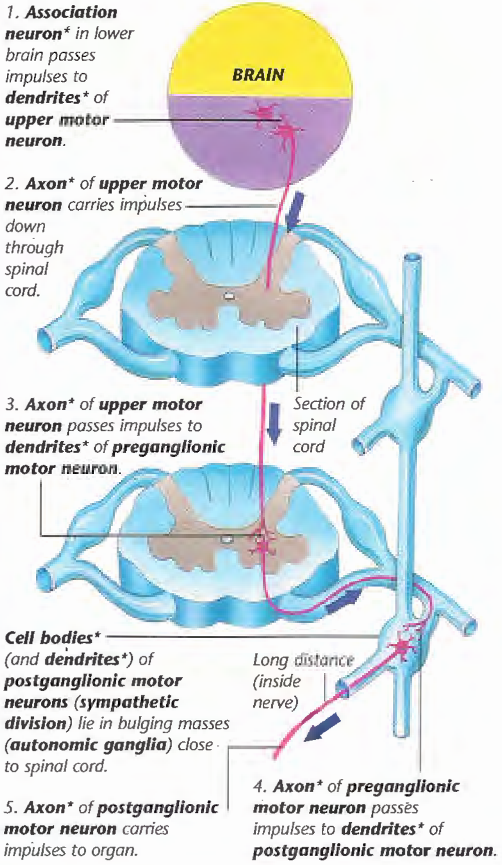
1. **Association neuron*** in lower brain passes impulses to **dendrites*** of **upper motor neuron**.
BRAIN
2. **Axon*** of **upper motor neuron** carries impulses down through spinal cord.
Section of spinal cord
3. **Axon*** of **upper motor neuron** passes impulses to **dendrites*** of **preganglionic motor neuron**.
Cell bodies* (and **dendrites***) of **postganglionic motor neurons** (**sympathetic division**) lie in bulging masses (**autonomic ganglia**) close to spinal cord.
Long distance (inside nerve)
4. **Axon*** of **preganglionic motor neuron** passes impulses to **dendrites*** of **postganglionic motor neuron**.
5. **Axon*** of **postganglionic motor neuron** carries impulses to organ.

Reflex actions

Involuntary actions of which we are aware. The term is most often used to refer to sudden actions of **skeletal muscles***, e.g. snatching the hand away from something hot. The impulses which cause such an action are carried by nerve cells of the **somatic efferent system** and the entire **neural pathway*** is a "short-circuited" one, called a **reflex arc**. In the case of **cranial reflexes** (those of the head, e.g. sneezing), this pathway involves a small part of the brain; with **spinal reflexes** (those of the rest of the body), the brain is not actively involved, only the spinal cord.

Pulling your hand away from an intense source of heat is a ***reflex action****.*

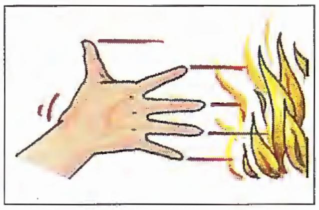

Simplified reflex arc (spinal reflex)

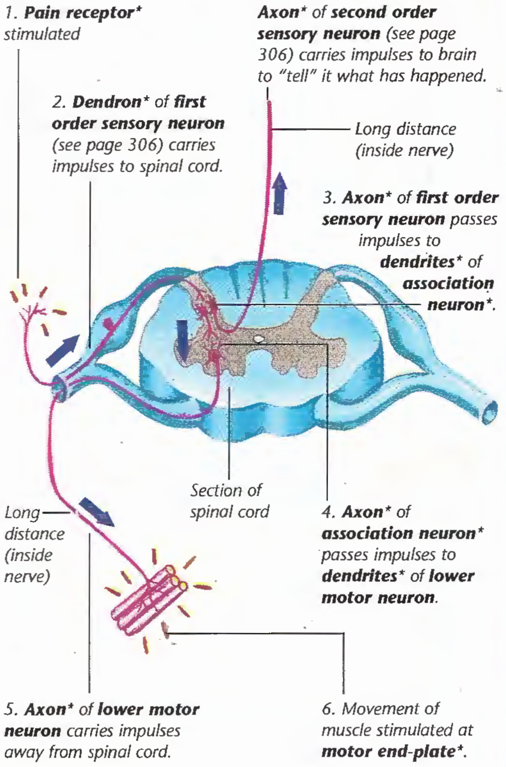
1. **Pain receptor*** stimulated
Axon* of **second order sensory neuron** (see page 306) carries impulses to brain to "tell" it what has happened.
2. **Dendron*** of **first order sensory neuron** (see page 306) carries impulses to spinal cord.
Long distance (inside nerve)
3. **Axon*** of **first order sensory neuron** passes impulses to **dendrites*** of **association neuron***.
Long distance (inside nerve)
Section of spinal cord
4. **Axon*** of **association neuron*** passes impulses to **dendrites*** of **lower motor neuron**.
5. **Axon*** of **lower motor neuron** carries impulses away from spinal cord.
6. Movement of muscle stimulated at **motor end-plate***.

---

* **Association neurons**, 305; **Axon, Cell body**, 304; **Dendron**, 304 (**Dendrites**); **Hypothalamus**, 303; **Motor end-plate**, 283; **Neural pathways**, 306; **Pain receptors**, 311; **Skeletal muscles**, 282.

THE SKIN

The **skin** or **cutis** is the outer body covering, made up of several tissue layers. It registers external stimulation, protects against damage or infection, prevents drying out, helps regulate body temperature, excretes waste (**sweat**), stores fat and makes **vitamin D***. It contains many tiny structures, each type with a different function. The entire skin (tissue layers and structures), is called the **integumentary system**.

The different layers

Epidermis

The thin outer layer of the skin which forms its **epithelium** (a term for any sheet of cells which forms a surface covering or a cavity lining). It is made up of several layers (**strata**, sing. **stratum**), shown in the picture, right.

*1. **Stratum corneum** (**horny** or **cornified layer**). Flat, dead cells filled with **keratin** (a fibrous, waterproofing protein). The cells are continually worn away or shed.*

*2. **Stratum granulosum** (**granular layer**). Flat, granulated cells. They slowly die away (there are no blood vessels in the epidermis to provide food and oxygen) and are pushed up to become part of the **stratum corneum**.*

*3. **Stratum germinativum**. Made up of two layers. The upper one (**stratum spinosum**) consists of new living cells. These push upward (and become part of the **stratum granulosum**) as more cells are made below them by the constantly-dividing cells of the lower layer (**stratum basale** or **Malphigian layer**).*

Epidermal layers

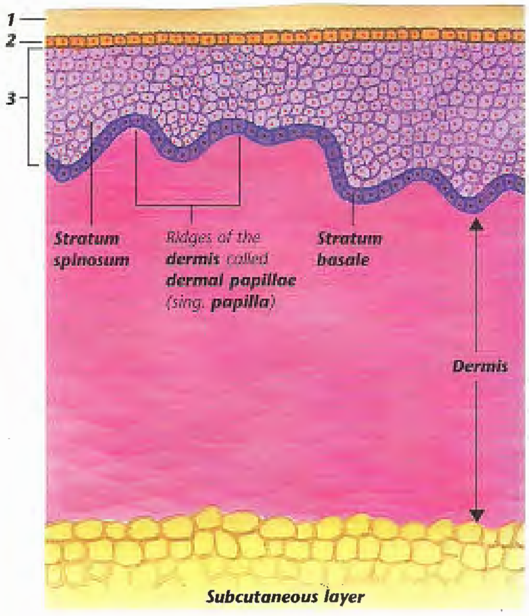

Dermis

The thick layer of **connective tissue*** under the **epidermis**, containing most of the embedded structures (see introduction). It also contains many **capillaries*** (tiny blood vessels) which supply food and oxygen.

Subcutaneous layer or superficial fascia

The layer of fatty tissue (**adipose tissue**) below the **dermis** (it is a fat store). Elastic fibers run through it to connect the dermis to the organs below, e.g. muscles. It forms an insulating layer. See also picture, right.

Melanin

A brown **pigment*** which shields against ultraviolet light by absorbing the light energy. It is found in all the layers of the **epidermis** of people from tropical areas, giving them dark skin. Fair-skinned people only have melanin in their lower epidermal layers, but produce more when in direct sunlight, causing a suntan.

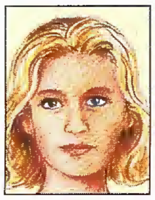

*People with **melanin** only in lower layers of the **epidermis** have fair skin.*

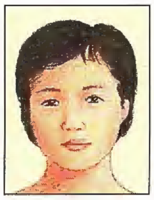

*The **pigment*** **carotene**, together with **melanin**, causes yellow skin.*

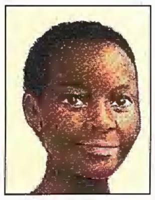

*Dark skin is caused by large amounts of **melanin** in all epidermal layers.*

---

* **Capillaries**, 288; **Connective tissue**, 280; **Pigments**, 255; **Vitamin D**, 339.

Structures in t

1. Meissner's corpusc
Bodies formed around
There are especially lar
fingertips and palms. T
receptors*, i.e. they ser
the brain when the skin
contact with an object.

2. Sebaceous glands
Exocrine glands* which
into **hair follicles**. They
an oil called **sebum** wh
waterproofs the hairs a
epidermis and keeps
them supple.

3. Hair erector musc
Special muscles, each
attached to a **hair follic**
When they contract (in
cold), the hairs straighte
traps more air and imp
insulation (especially in
animals with lots of hai
feathers or fur). It also c
"goose-pimples".

4. Hair follicles
Long, narrow tubes, eac
a **hair**. The hair grows a
added to its base from t
the follicle. Its older cel
forms inside them (see
corneum).

5. Pain receptors
Nerve fiber endings in
inner organs and in the
epidermis and the top
are the **receptors*** whic
when any stimulation (
touch) becomes excessi
causes a sensation of p

6. Hair plexuses or **r**
Special groups of nerve
forms a network around
is a **receptor***, i.e. it ser
impulses to the brain, i
the hair moves.

Capillaries, 288; **Exocrine gla**
Receptors, 307; **Urea**, 300.

Structures in the skin

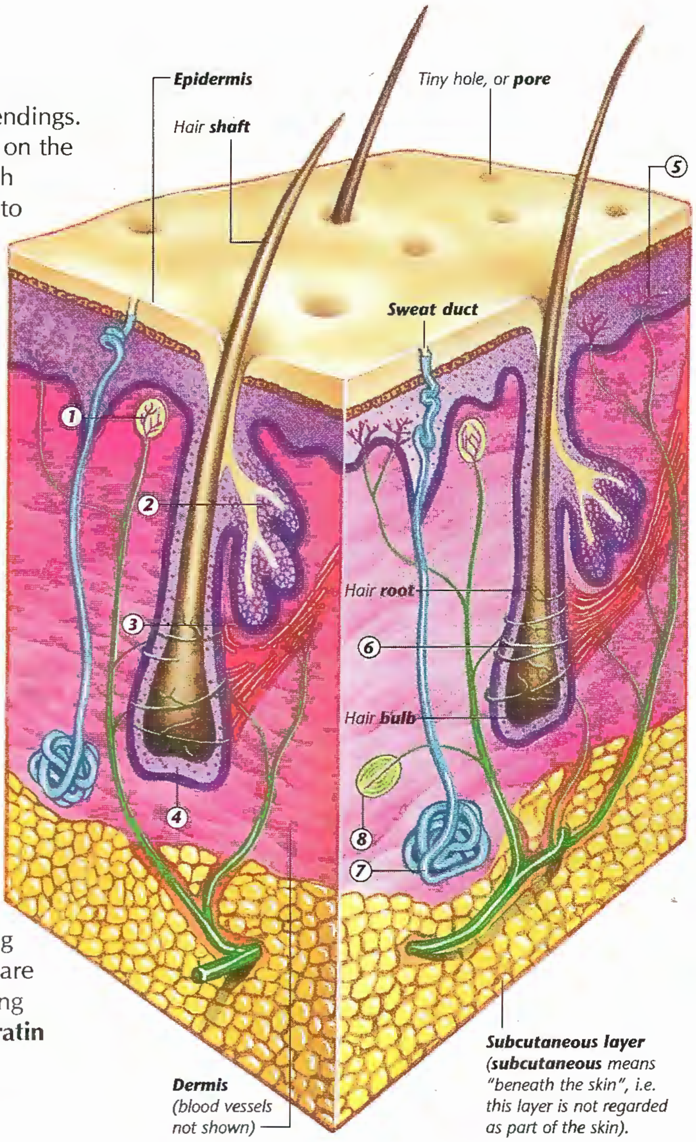
Epidermis
Hair shaft
Tiny hole, or pore
5
Sweat duct
1
2
3
Hair root
6
Hair bulb
4
8
7
Dermis (blood vessels not shown)
Subcutaneous layer (subcutaneous means "beneath the skin", i.e. this layer is not regarded as part of the skin).

1. Meissner's corpuscles

Bodies formed around nerve fiber endings. There are especially large numbers on the fingertips and palms. They are touch **receptors***, i.e. they send impulses to the brain when the skin makes contact with an object.

2. Sebaceous glands

Exocrine glands* which open into **hair follicles**. They produce an oil called **sebum** which waterproofs the hairs and **epidermis** and keeps them supple.

3. Hair erector muscles

Special muscles, each attached to a **hair follicle**. When they contract (in the cold), the hairs straighten. This traps more air and improves insulation (especially in animals with lots of hair, feathers or fur). It also causes "goose-pimples".

4. Hair follicles

Long, narrow tubes, each containing a hair. The hair grows as new cells are added to its base from the cells lining the follicle. Its older cells die as **keratin** forms inside them (see **stratum corneum**).

5. Pain receptors

Nerve fiber endings in the tissue of most inner organs and in the skin (in the **epidermis** and the top of the **dermis**). They are the **receptors*** which send impulses when any stimulation (e.g. pressure, heat, touch) becomes excessive. This is what causes a sensation of pain.

6. Hair plexuses or root hair plexuses

Special groups of nerve fiber endings. Each forms a network around a **hair follicle** and is a **receptor***, i.e. it sends nervous impulses to the brain, in this case when the hair moves.

7. Sweat glands or sudoriferous glands

Coiled **exocrine glands*** which excrete **sweat**. Each has a narrow tube (**sweat duct**) going to the surface. Sweat consists of water, salts and **urea***, which enter the gland from the cells and **capillaries*** (blood vessels).

8. Pacinian corpuscles

Special bodies formed around single nerve fiber endings, lying in the lower skin layers and the walls of inner organs. They are pressure **receptors***, i.e. they send impulses to the brain when the tissue is receiving deep pressure rather than light touch.

* **Capillaries**, 288; **Exocrine glands**, 296; **Receptors**, 307; **Urea**, 300.

THE EYES

The **eyes** are the organs of **sight**, sending nervous impulses to the brain when stimulated by light rays from external objects. The brain interprets the impulses to produce images. Each eye consists of a hollow, spherical capsule (**eyeball**), made up of several layers and structures. It is set into a socket in the skull (an **orbit**), and is protected by eyelids and eyelashes.

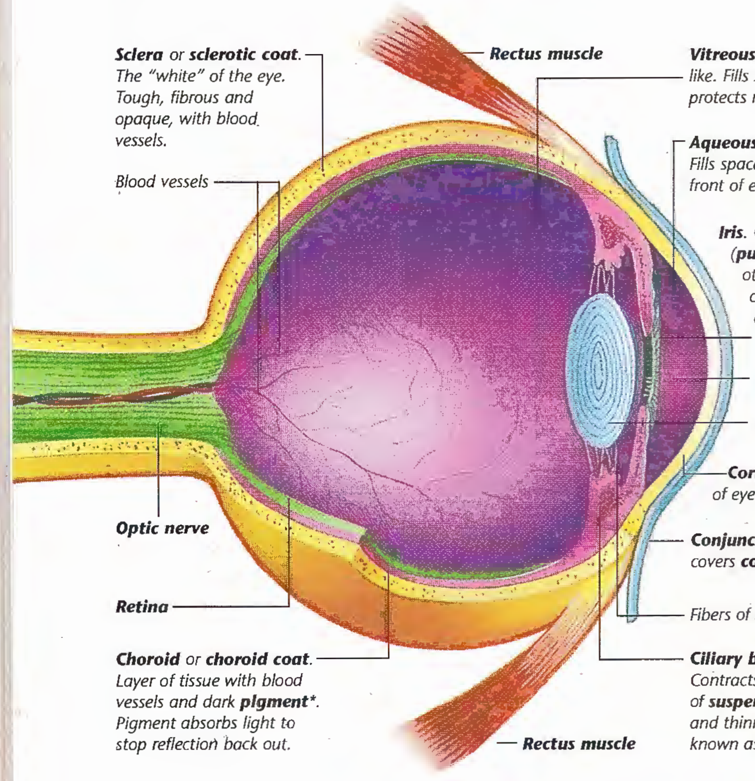
Blood vessels
Optic nerve
Retina
Rectus muscle
Lens
Rectus muscle

Sclera *or* ***sclerotic coat****. The "white" of the eye. Tough, fibrous and opaque, with blood vessels.*

Choroid *or* ***choroid coat****. Layer of tissue with blood vessels and dark* ***pigment*****. Pigment absorbs light to stop reflection back out.*

Vitreous humor*. Fluid similar to* ***aqueous humor****, but stiff and jelly-like. Fills space called* ***posterior cavity****. Keeps shape of eyeball, protects nervous layer (****retina****) and helps to "bend" (****refract****) light.*

Aqueous humor*. Watery liquid with sugars, salts and proteins. Fills space called* ***anterior cavity****. Protects* ***lens*** *and nourishes front of eye. Constantly drains away and is replaced.*

Iris*. Opaque disk of tissue, with blood vessels and central hole (****pupil****). Contains muscle fibers, some in concentric circles, others which radiate out from center to edge. Former contract to decrease pupil size (in bright light), latter contract to increase it (in dim light). Has various amounts of* ***pigment*****, producing eye colors.*

Pupil*. Central hole in* ***iris****.*

Cornea*. Transparent continuation of* ***sclera****. Protects front of eye and "bends" (****refracts****) light rays onto* ***lens****.*

Conjunctiva*. Thin* ***mucous membrane*****. Lines eyelids and covers* ***cornea****.*

Fibers of ***suspensory ligament*** *(see* ***lens****)*

Ciliary body*. Muscular ring (****smooth muscle*****) round* ***lens****. Contracts to make lens smaller (in diameter) and fatter (fibers of* ***suspensory ligament*** *relax). Relaxes to make lens larger and thinner (fibers tighten). Together with muscles of* ***iris****, known as* ***intrinsic eye muscles****.*

Lens

The transparent body whose role, like that of any lens, is to focus the light rays passing through it, i.e. "bend" (**refract**) them so that they come to a point, in this case on the **retina** (for more about lenses and refraction, see pages 50-53). A lens consists of many thin tissue layers and is held in place by the fibers of a **ligament*** called the **suspensory ligament**. These join it to the **ciliary body**, which can alter the lens shape so that light rays are always focused on the retina, whatever the distance of the object being looked at. This is known as **accommodation**. The rays form an upside-down image, but this is corrected by the brain.

Accommodation

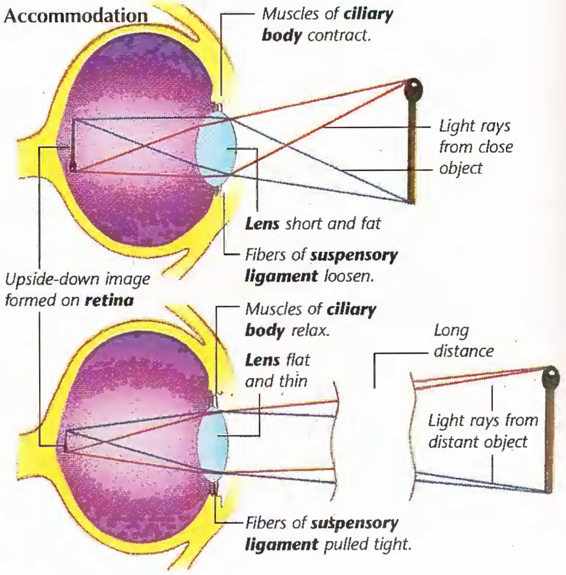
Muscles of **ciliary body** contract.
Light rays from close object
Lens short and fat
Fibers of **suspensory ligament** loosen.
Upside-down image formed on **retina**
Muscles of **ciliary body** relax.
Lens flat and thin
Long distance
Light rays from distant object
Fibers of **suspensory ligament** pulled tight.

---

* **Ligaments**, 280; **Mucous membrane**, 295; **Pigments**, 255; **Smooth muscle**, 283.

The inner nerv

Retina

The innermost layer of tis
the eyeball, made up of a
and a nervous layer consi
sensory nerve cells (**sens**
their fibers. These lie in c
nervous impulses to the
in the chains are **recepto**
fibers (**dendrons***) fire off
they are stimulated (by li
are called **rods** and **cone**
shapes. The receptors are
i.e. stimulated by light).

Close-up area of retina

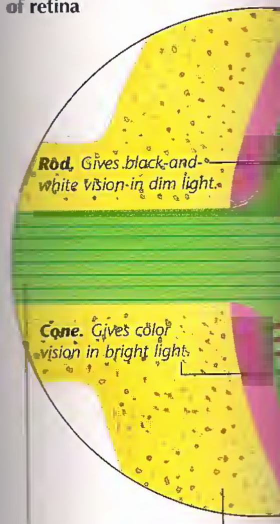
Rod. Gives black-and-white vision in dim light.
Cone. Gives color vision in bright light.
Sclera

Fibers form ***optic nerve****, leading to brain.*

Macula lutea or **yello**
An area of yellowish tiss
the **retina**. It has a smal
the **fovea** or **fovea centr**
highest concentration
and is the area of acutes
directly at a specific ob
focused on the fovea.

Blind spot or **optic di**
The point in the **retina**
leaves the eye. It has no
and so cannot send an

---

* **Dendron**, 304 (**Dendrites**);
Nasal cavity, 307 (**Nose**); **Pig**

The inner nervous layer

Retina

The innermost layer of tissue at the back of the eyeball, made up of a layer of **pigment*** and a nervous layer consisting of millions of sensory nerve cells (**sensory neurons***) and their fibers. These lie in chains and carry nervous impulses to the brain. The first cells in the chains are **receptors***, i.e. their end fibers (**dendrons***) fire off the impulses when they are stimulated (by light rays). These fibers are called **rods** and **cones** because of their shapes. The receptors are **photoreceptors** (i.e. stimulated by light).

Close-up area of retina

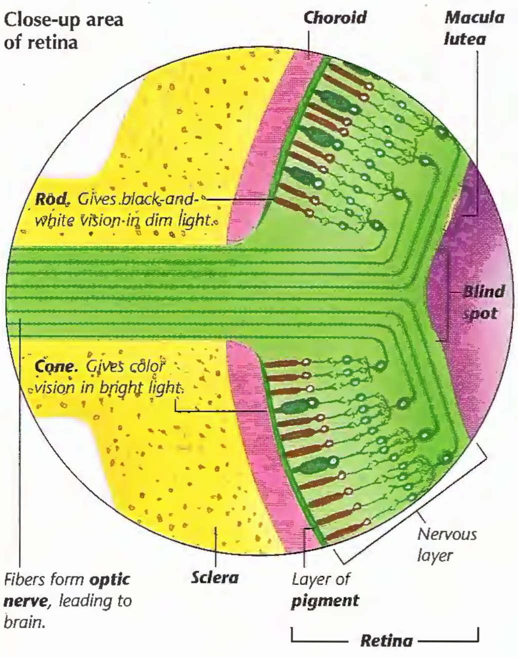
Choroid
Macula lutea
Rod. Gives black-and-white vision in dim light.
Blind spot
Cone. Gives color vision in bright light.
Nervous layer
Fibers form **optic nerve**, leading to brain.
Sclera
Layer of **pigment**
Retina

Macula lutea or yellowspot

An area of yellowish tissue in the center of the **retina**. It has a small central dip, called the **fovea** or **fovea centralis**. This has the highest concentration of **cones** (see **retina**) and is the area of acutest vision. If you look directly at a specific object, its light rays are focused on the fovea.

Blind spot or optic disk

The point in the **retina** where the **optic nerve** leaves the eye. It has no **receptors** (see **retina**) and so cannot send any impulses.

Structures around the eyeballs

Extrinsic eye muscles

The three pairs of muscles joining the eyeball to the eye socket (**orbit**). They contract to make the eyeball swivel around.

Extrinsic eye muscles

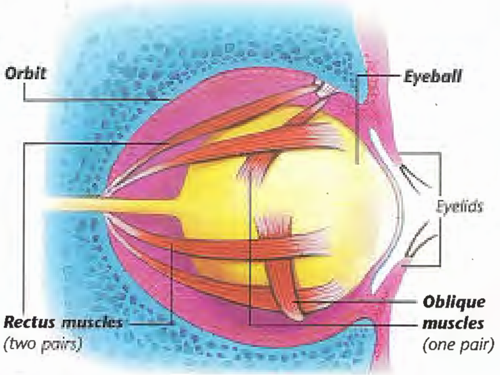
Orbit
Eyeball
Eyelids
Rectus muscles (two pairs)
Oblique muscles (one pair)

Lachrymal glands or tear glands

Two **exocrine glands***, one at the top of each eye socket (**orbit**). They secrete a watery fluid onto the lining of the upper eyelids via tubes called **lachrymal ducts**. The fluid contains salts and an anti-bacterial **enzyme***, and it washes over the surface of the eyes, keeping them moist and clean. It drains away via four **lachrymal canals**, two at the inside corner of each eye, which join to form a **nasolachrymal duct**. This empties into a **nasal cavity***.

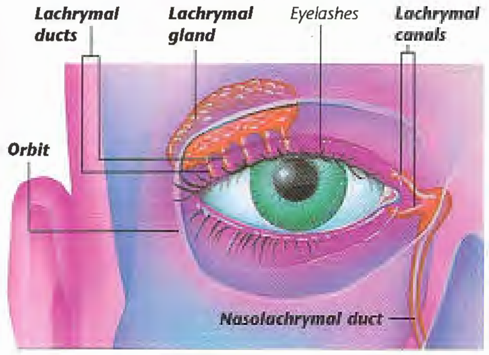
Lachrymal ducts
Lachrymal gland
Eyelashes
Lachrymal canals
Orbit
Nasolachrymal duct

Optic *means "of vision or the eye".*

Visual *means "perceptible by the sense of sight".*

---

* **Dendron**, 304 (**Dendrites**); **Enzymes**, 333; **Exocrine glands**, 296; **Nasal cavity**, 307 (**Nose**); **Pigments**, 255; **Receptors**, 307; **Sensory neurons**, 305.

THE EARS

The two **ears** are the organs of hearing and balance. Each one is divided into three areas – the **outer ear**, the **middle ear** and the **inner ear**.

Outer ear or external ear

An outer "shell" of skin and **cartilage*** (**pinna** or **auricle**), together with a short tube (**ear canal** or **external auditory canal**). The tube lining contains special **sebaceous glands*** (**ceruminous glands**) which secrete **cerumen** (ear wax).

Middle ear or tympanic cavity

An air-filled cavity which contains a chain of three tiny bones (**ear ossicles** or **auditory ossicles**) called the **malleus** (or **hammer**), **incus** (or **anvil**) and **stapes** (or **stirrup**).

Inner ear or internal ear

A connected series of cavities in the skull, with tubes and sacs inside them. The cavities (**cochlea**, **vestibule** and **semicircular canals**) are called the **bony labyrinth** and are filled with one fluid (**perilymph**). The tubes and sacs are filled with another fluid (**endolymph**) and are called the **membranous labyrinth**. They are the **cochlear duct**, **saccule**, **utricle** and **semicircular ducts**.

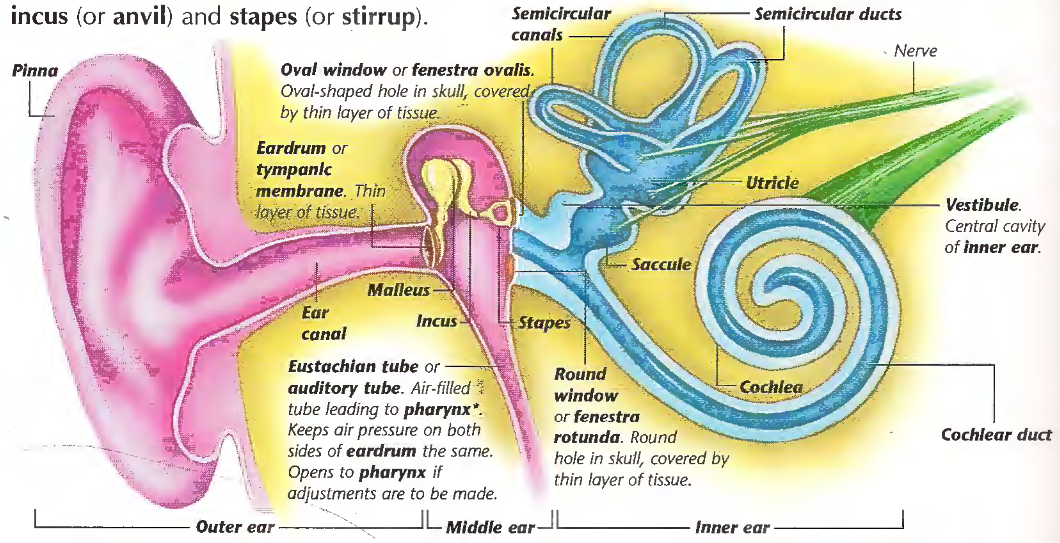
Semicircular canals
Semicircular ducts
Nerve
Pinna
Oval window or **fenestra ovalis.** Oval-shaped hole in skull, covered by thin layer of tissue.
Eardrum or **tympanic membrane.** Thin layer of tissue.
Utricle
Vestibule. Central cavity of **inner ear.**
Saccule
Malleus
Ear canal
Incus
Stapes
Eustachian tube or **auditory tube.** Air-filled tube leading to **pharynx***. Keeps air pressure on both sides of **eardrum** the same. Opens to **pharynx** if adjustments are to be made.
Round window or **fenestra rotunda.** Round hole in skull, covered by thin layer of tissue.
Cochlea
Cochlear duct
Outer ear
Middle ear
Inner ear

The inner ear and hearing

Cochlea

A spiralling tubular cavity, part of the **inner ear**. It contains **perilymph** (see **inner ear**) in two channels (continuous with each other), and also a third channel – the **cochlear duct**.

Cochlear duct

A spiralling tube within the **cochlea**, connected to the **saccule**. It contains **endolymph** (see **inner ear**) and a long body called the **organ of Corti**. This contains special hair cells whose hairs project into the endolymph and touch a shelf-like tissue layer (**tectorial membrane**). The bases of the cells are attached to nerve fibers (**dendron*** endings).

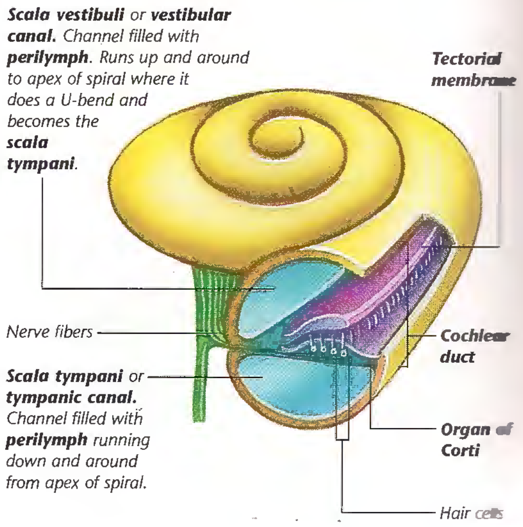
Scala vestibuli or **vestibular canal.** Channel filled with **perilymph.** Runs up and around to apex of spiral where it does a U-bend and becomes the **scala tympani.**
Tectorial membrane
Nerve fibers
Cochlear duct
Scala tympani or **tympanic canal.** Channel filled with **perilymph** running down and around from apex of spiral.
Organ of Corti
Hair cells

The inner ear a

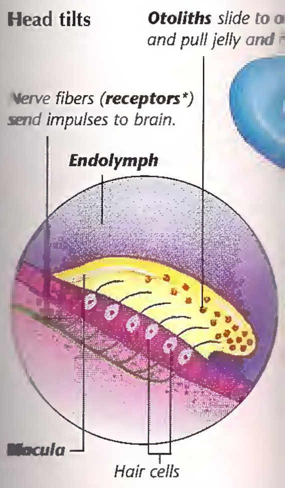
Head tilts
Otoliths slide to o
and pull jelly and
Nerve fibers (**receptors***) send impulses to brain.
Endolymph
Macula
Hair cells

Saccule (sacculus) and

Two sacs lying between t **ducts** and the **cochlear du** **endolymph** (see **inner ear** hair cells in patches in th cells have nerve fibers (**de** attached to them and hair jelly-like mass called a **m** This contains grains of ca **otoliths**). The maculae se information about forwar sideways or tilting motion

Sound waves (air vibrations) come **eardrum** vibrate.

Ear ossicles pick up vibrations and **window** (lever action magnifies vibrat

Vibrations of **oval window** cause **vestibule.**

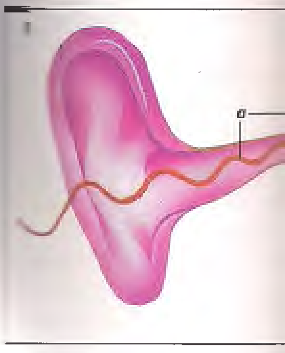

---

* **Cartilage**, 281; **Dendron**, 304 (**Dendrites**); **Pharynx**, 294; **Sebaceous glands**, 311.

Dendron, 304 (**Dendrites**); **Receptors**, 307.

The inner ear and balance

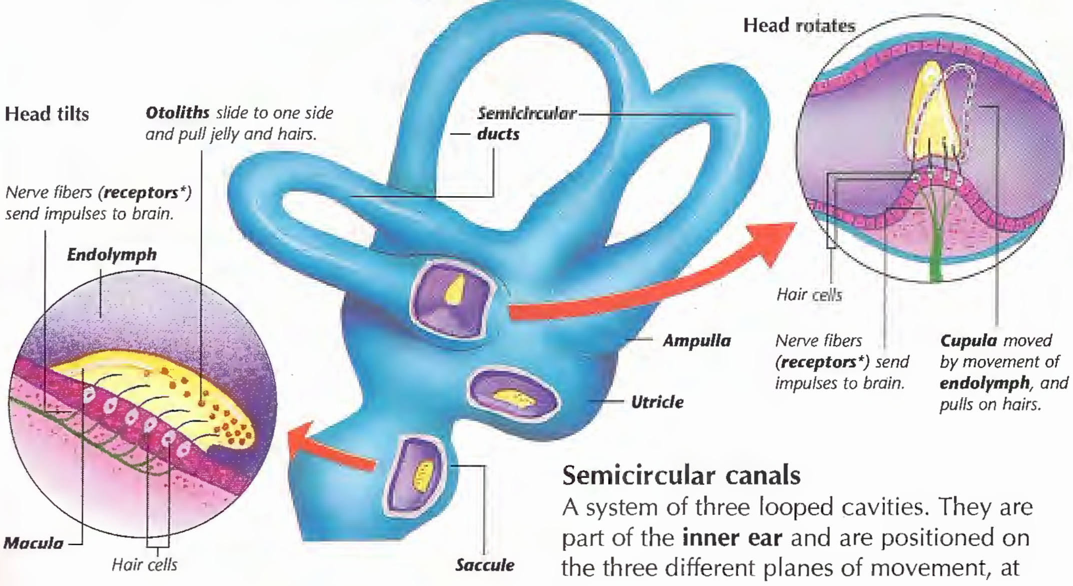
Head tilts
Otoliths slide to one side and pull jelly and hairs.
Nerve fibers (**receptors***) send impulses to brain.
Endolymph
Macula
Hair cells
Semicircular ducts
Saccule
Ampulla
Utricle
Head rotates
Hair cells
Nerve fibers (**receptors***) send impulses to brain.
Cupula moved by movement of **endolymph**, and pulls on hairs.

Saccule (sacculus) and utricle (utriculus)

Two sacs lying between the **semicircular ducts** and the **cochlear duct**. They contain **endolymph** (see **inner ear**) and have special hair cells in patches in their linings. These cells have nerve fibers (**dendron*** endings) attached to them and hairs embedded in a jelly-like mass called a **macula** (pl. **maculae**). This contains grains of calcium carbonate (**otoliths**). The maculae send the brain information about forward, backward, sideways or tilting motion of the head.

Semicircular canals

A system of three looped cavities. They are part of the **inner ear** and are positioned on the three different planes of movement, at right angles to each other.

Semicircular ducts

Three looped tubes inside the **semicircular canals**. Each contains **endolymph** (see **inner ear**) and a special sensory body, which lies across the basal swelling (**ampulla**, pl. **ampullae**) of the duct. The sensory bodies (**cupulae**, sing. **cupula**) work in a very similar way to **maculae** (see **saccule**) – each consists of a jelly-like mass (without **otoliths**) and hair cells. They send the brain information about rotation and tilting of the head.

*a) Sound waves (air vibrations) come in along **ear canal** and make **eardrum** vibrate.*

*b) **Ear ossicles** pick up vibrations and pass them to **oval window** (lever action magnifies vibrations about 20 times).*

*c) Vibrations of **oval window** cause waves in **perilymph** of **vestibule**.*

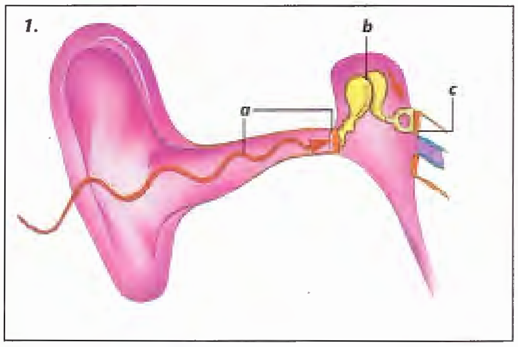

*d) Waves in **perilymph** of **scala vestibuli** cause waves in **endolymph** of **cochlear duct**.*

*e) Hairs move and cause nerve fibers (**receptors***) to send impulses to brain (which interprets them as sensation of hearing).*

f) Waves gradually fade out.

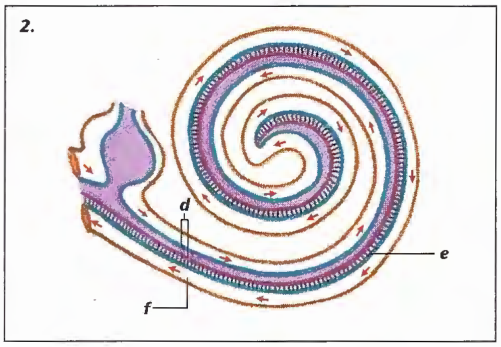

---

* **Dendron**, 304 (**Dendrites**); **Receptors**, 307.

THE REPRODUCTIVE SYSTEM

Reproduction is the process of producing new life. Humans reproduce by **sexual reproduction*** (described on pages 318-319) and the reproductive organs involved (making up the **reproductive system**) are called the **genital organs** or **genitalia**. They consist of the primary reproductive organs, or **gonads** (two **ovaries** in women, two **testes** in men) and a number of additional organs. In both women and men, cells in the gonads also act as **endocrine glands***, secreting many important **hormones***.

The male reproductive system

Testes (sing. testis) or testicles

The two male **gonads** (see introduction). They contain tube-like canals called **seminiferous tubules**, inside which the male **gametes*** (sex cells), called **sperm**, are made after **puberty*** (for more about how sperm are made, see pages 322-323). The testes lie in a sac (**scrotum**), which hangs below the abdomen (the temperature for sperm production must be slightly lower than body temperature). They also produce **hormones*** (**androgens** – see pages 336-337) after puberty.

Side view of male organs (only one testis shown)

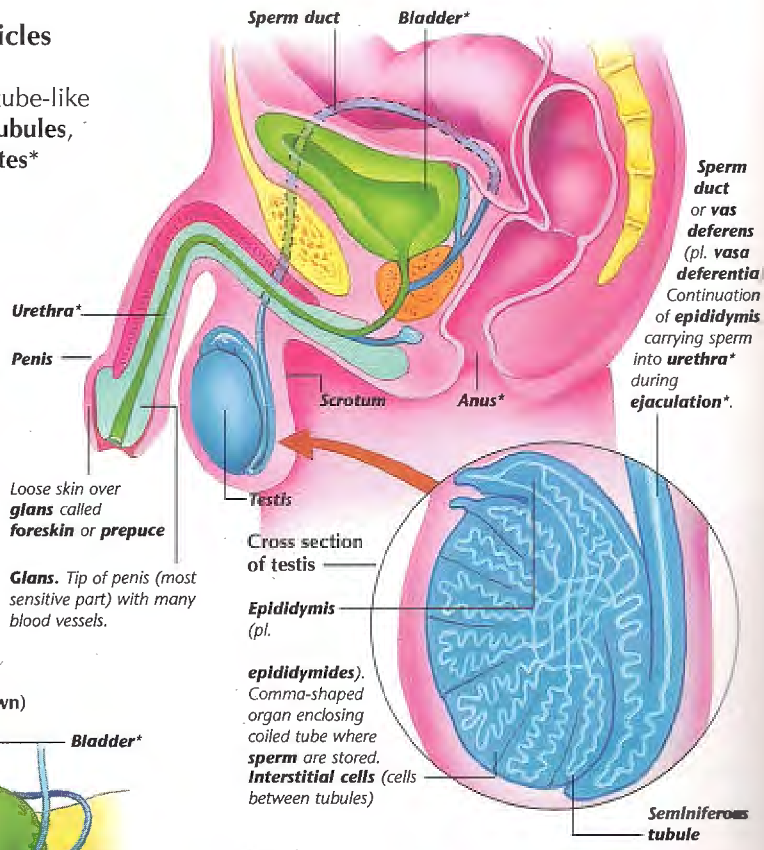

Sperm duct
Bladder*
Sperm duct or vas deferens (pl. vasa deferentia). Continuation of epididymis carrying sperm into urethra* during ejaculation*.
Urethra*
Penis
Scrotum
Anus*
Loose skin over glans called foreskin or prepuce
Glans. Tip of penis (most sensitive part) with many blood vessels.
Testis
Cross section of testis
Epididymis (pl. epididymides). Comma-shaped organ enclosing coiled tube where sperm are stored.
Interstitial cells (cells between tubules)
Seminiferous tubule

Ducts and glands (view from behind – testes not shown)

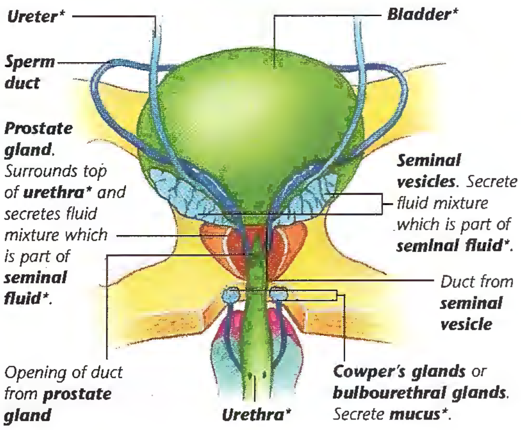

Ureter*
Bladder*
Sperm duct
Prostate gland. Surrounds top of urethra* and secretes fluid mixture which is part of seminal fluid*.
Seminal vesicles. Secrete fluid mixture which is part of seminal fluid*.
Duct from seminal vesicle
Opening of duct from prostate gland
Urethra*
Cowper's glands or bulbourethral glands. Secrete mucus*.

Penis

The organ through which **sperm** (see **testes**) are ejected (via the **urethra***) during **copulation***. It is made of soft, sponge-like **erectile tissue**, which has many spaces (**blood sinuses**), blood vessels and nerve fiber endings (**receptors***). When a man is sexually excited, the sinuses and blood vessels fill with blood (the blood vessels expand). This makes the penis stiff and erect.

---

*Anus, 295 (**Large intestine**); **Bladder**, 300; **Ejaculation**, 319 (**Copulation**); **Endocrine glands**, 297; **Gametes**, 320; **Hormones**, 336; **Mucus**, 295 (**Mucous membrane**); **Puberty**, 318; **Receptors**, 307; **Seminal fluid**, 319 (**Copulation**); **Sexual reproduction**, 320; **Ureters**, 300; **Urethra**, 300.

The female rep

Ovaries

The two female **gonads** (
They are held in place i
(below the kidneys) by **li**
attach them to the walls
female **gametes*** (sex ce
ovum), are produced reg
(in **ovarian follicles**) afte
about how ova are made

Inner parts of the female reproductive system

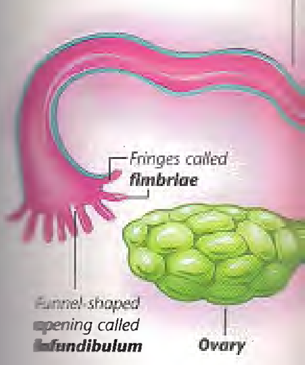

Fringes called fimbriae
Funnel-shaped opening called infundibulum
Ovary

Ovarian follicles

Areas of tissue appearing
in the **ovaries** after **pube**
contains a maturing **ovu**
The follicles gradually ge
secrete **hormones*** (see **e**
Each round of follicle pro
one fully mature follicle

Uterus or womb

The **hollow** organ, inside
baby (**fetus***) is held, or f
see **ovaries**) are discharg
cycle, page 318). It has a
membrane* (the **endome**
muscular wall with many

Vagina

The **muscular** canal leadi
out of the body. It carries
ovaries) and **endometriu**
during **menstruation***, re
during **copulation*** and s
canal. Its lining produces

*Anus, 295 (**Large intestine**); **Co**
Gametes, 320; **Hormones**, 336;
Ovulation, 318 (**Menstrual cycl**

The female reproductive system

Ovaries

The two female **gonads** (see introduction). They are held in place in the lower abdomen (below the kidneys) by **ligaments***. These attach them to the walls of the pelvis. The female **gametes*** (sex cells), called **ova** (sing. **ovum**), are produced regularly in the ovaries (in **ovarian follicles**) after **puberty***. For more about how ova are made, see pages 322-323.

Vulva or pudendum

A collective term for the outer parts of the female reproductive system – the **labia** and the **clitoris** (see picture, bottom right). The **labia** (**majora** and **minora**) are two folds of skin (one inside the other) which surround the openings from the **vagina** and the **urethra***. The clitoris is the most sensitive part. Like the **penis**, it is made of **erectile tissue** and has many **receptors***.

Inner parts of the female reproductive system

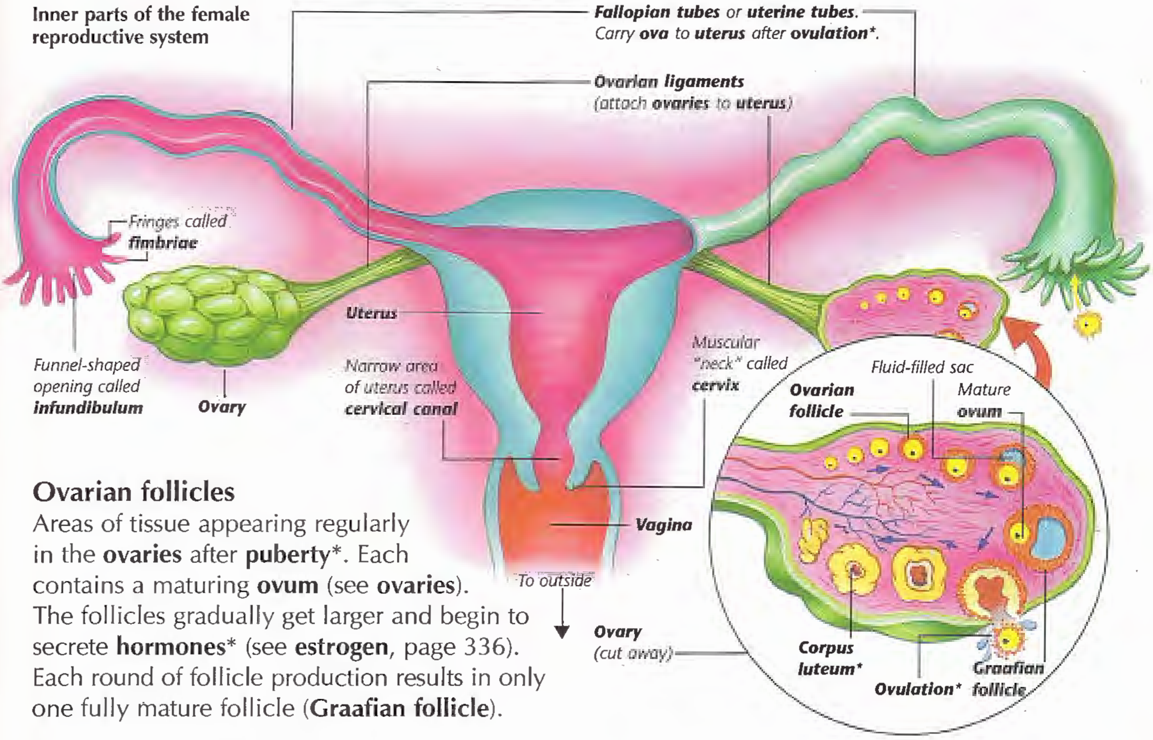
Fallopian tubes or **uterine tubes**. Carry **ova** to **uterus** after **ovulation***.
Ovarian ligaments (attach **ovaries** to **uterus**)
Fringes called **fimbriae**
Uterus
Funnel-shaped opening called **infundibulum**
Ovary
Narrow area of uterus called **cervical canal**
Muscular "neck" called **cervix**
Fluid-filled sac
Ovarian follicle
Mature **ovum**
Vagina
To outside
Ovary (cut away)
Corpus luteum*
Ovulation*
Graafian follicle

Ovarian follicles

Areas of tissue appearing regularly in the **ovaries** after **puberty***. Each contains a maturing **ovum** (see **ovaries**). The follicles gradually get larger and begin to secrete **hormones*** (see **estrogen**, page 336). Each round of follicle production results in only one fully mature follicle (**Graafian follicle**).

Uterus or womb

The hollow organ, inside which a developing baby (**fetus***) is held, or from which the **ova** (see **ovaries**) are discharged (see **menstrual cycle**, page 318). It has a lining of **mucous membrane*** (the **endometrium**), covering a muscular wall with many blood vessels.

Vagina

The muscular canal leading from the **uterus** out of the body. It carries away the **ova** (see **ovaries**) and **endometrium** (see uterus) during **menstruation***, receives the **penis** during **copulation*** and serves as the birth canal. Its lining produces a lubricating fluid.

Outer parts of the female reproductive system

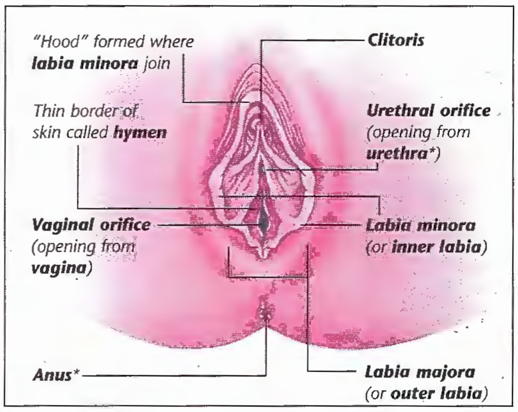
"Hood" formed where **labia minora** join
Clitoris
Thin border of skin called **hymen**
Urethral orifice (opening from **urethra***)
Vaginal orifice (opening from **vagina**)
Labia minora (or **inner labia**)
Anus*
Labia majora (or **outer labia**)

---

* **Anus**, 295 (**Large intestine**); **Copulation**, 319; **Corpus luteum**, 318 (**Menstrual cycle**); **Fetus**, 319 (**Pregnancy**); **Gametes**, 320; **Hormones**, 336; **Ligaments**, 280; **Menstruation**, 318 (**Menstrual cycle**); **Mucous membrane**, 295; **Ovulation**, 318 (**Menstrual cycle**); **Puberty**, 318; **Receptors**, 307; **Urethra**, 300.

DEVELOPMENT AND REPRODUCTION

Humans reproduce by **sexual reproduction***. The main processes this involves are described on these two pages, as well as the initial developments which allow it to happen.

Puberty

The point when the reproductive organs mature, and a person becomes capable of reproducing – roughly between the ages of 11 and 15 in girls, and 13 and 15 in boys. It involves a number of significant changes, all stimulated by **hormones*** (see **estrogen** and **androgens**, pages 336-337). All the new resulting features are called **secondary sex characters**, as distinct from the **primary sex characters** – the sex organs present from birth (see pages 316-317).

Changes at puberty

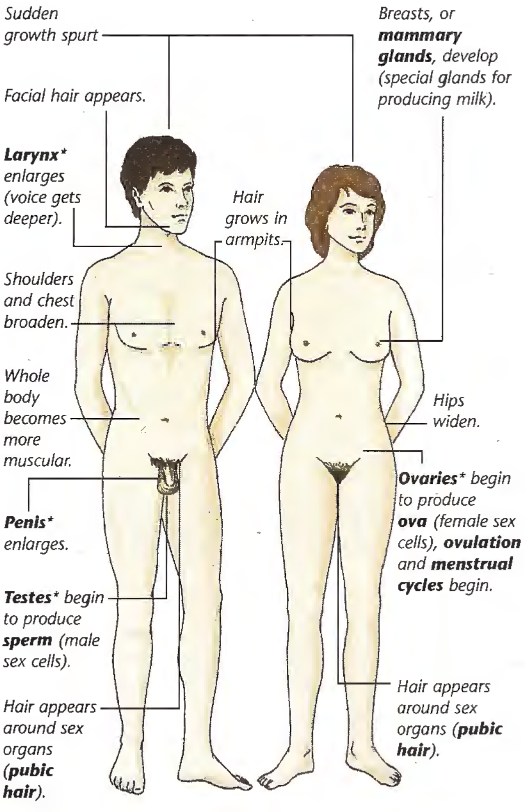
Sudden growth spurt
Facial hair appears.
Larynx* *enlarges (voice gets deeper).*
Shoulders and chest broaden.
Whole body becomes more muscular.
Penis* *enlarges.*
Testes* *begin to produce* ***sperm*** *(male sex cells).*
Hair appears around sex organs ***(pubic hair).***
Hair grows in armpits.
Breasts, or ***mammary glands,*** *develop (special glands for producing milk).*
Hips widen.
Ovaries* *begin to produce* ***ova*** *(female sex cells),* ***ovulation*** *and* ***menstrual cycles*** *begin.*
*Hair appears around sex organs (****pubic hair****).*

Menstrual cycle

A series of preparatory changes in the **uterus*** lining (**endometrium**), in case of **fertilization**. The lining gradually develops a new inner layer rich in blood vessels. If a fertilized **ovum** (female sex cell) does not appear, this new layer breaks down and leaves the body via the **vagina*** (**menstruation**). Each menstrual cycle lasts about 28 days and they occur continuously from **puberty** (usually between the ages of 11 and 15 – see left) to **menopause** (usually between 45 and 50), when ova production ceases. The events of the menstrual cycle run in conjunction with the **ovarian cycle** – the regular maturation of an ovum in an **ovarian follicle***, followed by **ovulation** (the release of the ovum into a **Fallopian tube***), and the breakdown of the **corpus luteum**. This body is formed from the burst **Graafian follicle*** (it does not break down if an ovum is fertilized). Both cycles are controlled by a group of **hormones*** (see pages 336-337).

Menstrual cycle / Ovarian cycle

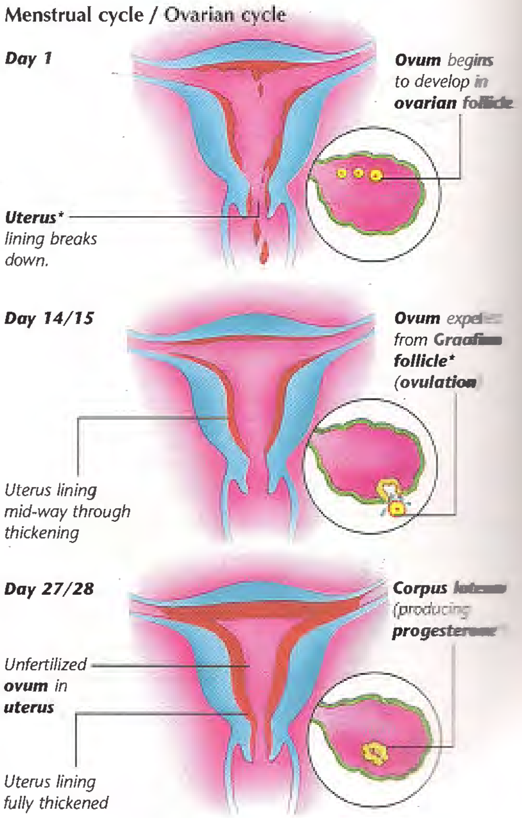
Day 1
Ovum *begins to develop in* ***ovarian follicle.***
Uterus* *lining breaks down.*
Day 14/15
Ovum *expelled from* ***Graafian follicle**** *(****ovulation****)*
Uterus lining mid-way through thickening
Day 27/28
Corpus luteum *(producing* ***progesterone****)*
Unfertilized ***ovum*** *in* ***uterus***
Uterus lining fully thickened

Copulation

Also called **coitus** or **sexu**
insertion of the **penis*** int
followed by rhythmical m
pelvis in one or both sexe
the male is **ejaculation** –
from the **urethra*** (in the
Semen consists of **sperm**
fluid mixture (**seminal flui**

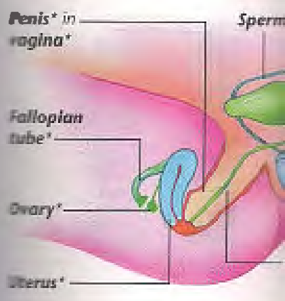
Penis* *in* ***vagina****
Sperm
*Fallopian tube**
*Ovary**
*Uterus**

Pregnancy

Pregnancy, or **gestation**
of carrying young. The
fertilization and giving
(**parturition**) is the **gesta**
about 9 months in hum
the new developing
individual in the **uterus***
is called a **fetus**, a term
usually used instead of
embryo* after about
two months of
pregnancy. A series
of powerful muscular
contractions called
labor occur just
before parturition.

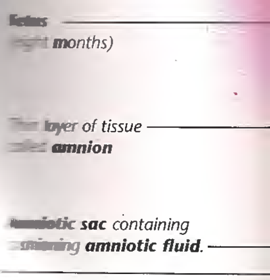
Fetus *(eight months)*
The layer of tissue called ***amnion***
Amniotic sac *containing cushioning* ***amniotic fluid.***

* **Fallopian tubes**, 317; **Graafian follicle**, 317 (**Ovarian follicles**); **Hormones**, 336; **Larynx**, 298; **Ovaries**, 317; **Penis**, 316; **Progesterone**, 336; **Sexual reproduction**, 320; **Testes**, 316; **Uterus**, **Vagina**, 317.

rmones 288; **Cleavage**, 320 (**Emb**
ostate **gland**, **Seminal vesicles**

Copulation

Also called **coitus** or **sexual intercourse**. The insertion of the **penis*** into the **vagina***, followed by rhythmical movements of the pelvis in one or both sexes. Its culmination in the male is **ejaculation** – the ejection of **semen** from the **urethra*** (in the penis) into the vagina. Semen consists of **sperm** (male sex cells) and a fluid mixture (**seminal fluid**).

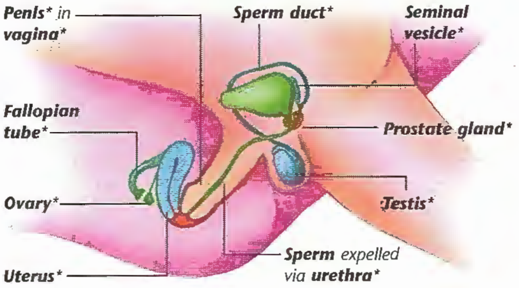
*Penis** in *vagina**
*Sperm duct**
*Seminal vesicle**
*Fallopian tube**
*Prostate gland**
*Ovary**
*Testis**
*Uterus**
*Sperm expelled via urethra**

Fertilization

A process which occurs after **ejaculation** if the **sperm** (male sex cells) meet an **ovum** (female sex cell) in a **Fallopian tube***. One sperm penetrates the ovum's outer skin (**zona pellucida**). Its **nucleus*** fuses with that of the ovum, and the first cell of a new baby (**zygote***) is formed. The new cell travels toward the **uterus***, undergoing many cell divisions (**cleavage***) as it does so. The ball of cells formed from these divisions then becomes embedded in the uterus wall (**implantation**), after which it is called an **embryo***.

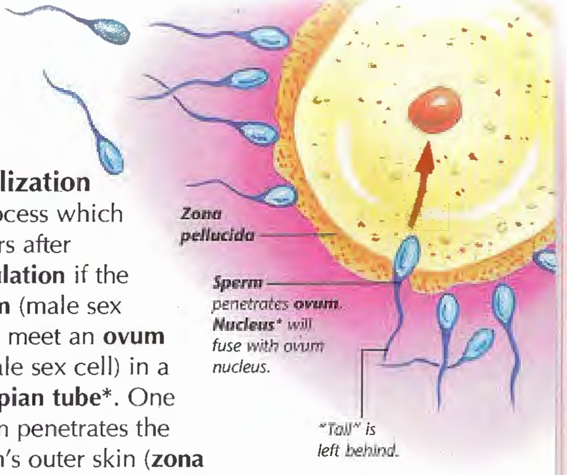
Zona pellucida
Sperm penetrates ovum. Nucleus will fuse with ovum nucleus.*
"Tail" is left behind.

Pregnancy

Pregnancy, or **gestation**, is the state of carrying young. The time between **fertilization** and giving birth (**parturition**) is the **gestation period** (about 9 months in humans) and the new developing individual in the **uterus*** is called a **fetus**, a term usually used instead of **embryo*** after about two months of pregnancy. A series of powerful muscular contractions called **labor** occur just before parturition.

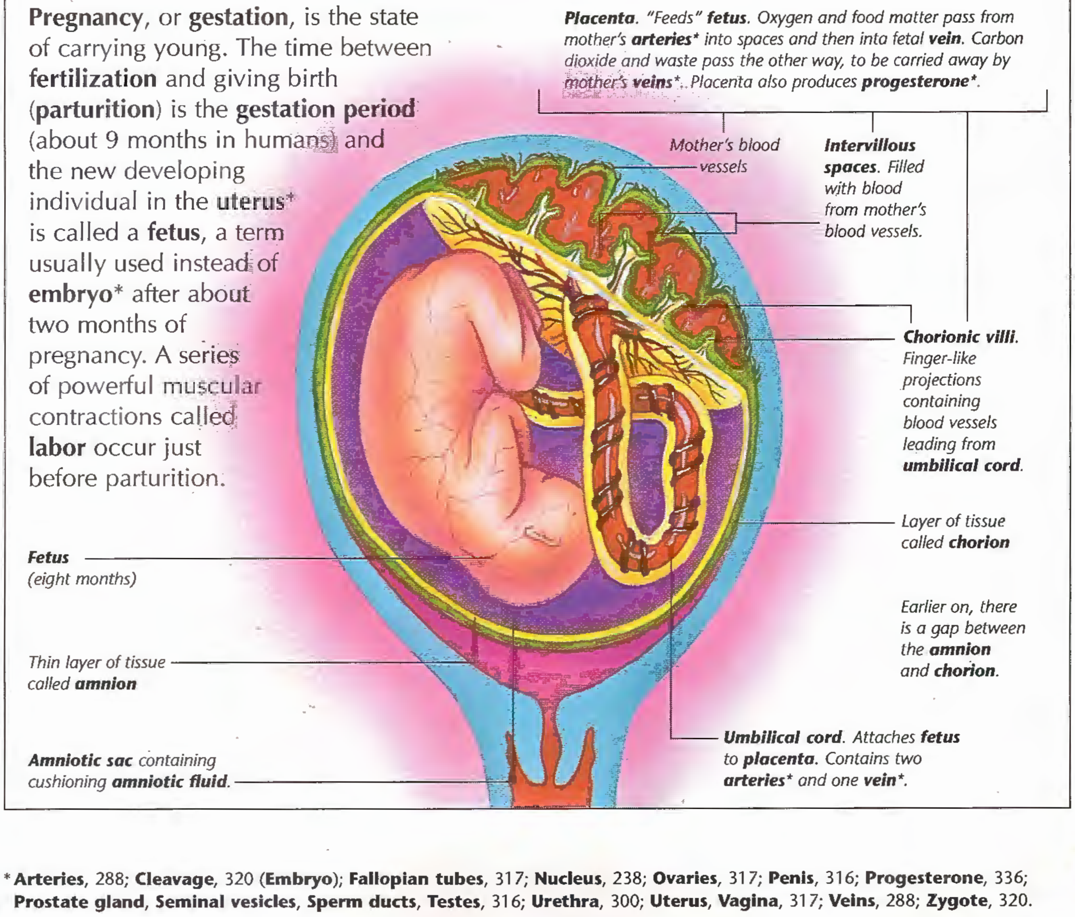
Placenta. "Feeds" fetus. Oxygen and food matter pass from mother's arteries into spaces and then into fetal vein. Carbon dioxide and waste pass the other way, to be carried away by mother's veins*. Placenta also produces progesterone*.*
Mother's blood vessels
Intervillous spaces. Filled with blood from mother's blood vessels.
Chorionic villi. Finger-like projections containing blood vessels leading from umbilical cord.
Layer of tissue called chorion
Fetus (eight months)
Earlier on, there is a gap between the amnion and chorion.
Thin layer of tissue called amnion
Umbilical cord. Attaches fetus to placenta. Contains two arteries and one vein*.*
Amniotic sac containing cushioning amniotic fluid.

* **Arteries**, 288; **Cleavage**, 320 (**Embryo**); **Fallopian tubes**, 317; **Nucleus**, 238; **Ovaries**, 317; **Penis**, 316; **Progesterone**, 336; **Prostate gland**, **Seminal vesicles**, **Sperm ducts**, **Testes**, 316; **Urethra**, 300; **Uterus**, **Vagina**, 317; **Veins**, 288; **Zygote**, 320.

TYPES OF REPRODUCTION

Reproduction is the creation of new life, a process which occurs in all living things. The two main types are **sexual** and **asexual reproduction**, but there is also a special case called **alternation of generations**.

Sexual reproduction

Sexual reproduction is the type of reproduction shown by all flowering plants and most animals. It involves the joining (**fusion**) of two **gametes** (sex cells), one male and one female. This process is called **fertilization**, and is further described on pages 258 (flowering plants), 319 (humans and similar animals) and 276 (other animals). The two gametes each have only half the number of **chromosomes*** (called the **haploid number***) as the plant or animal which produced them. This is achieved by a special kind of cell division (see pages 322-323) and ensures that when the gametes come together, the new individual produced has the correct, original number of chromosomes (called the **diploid number***).

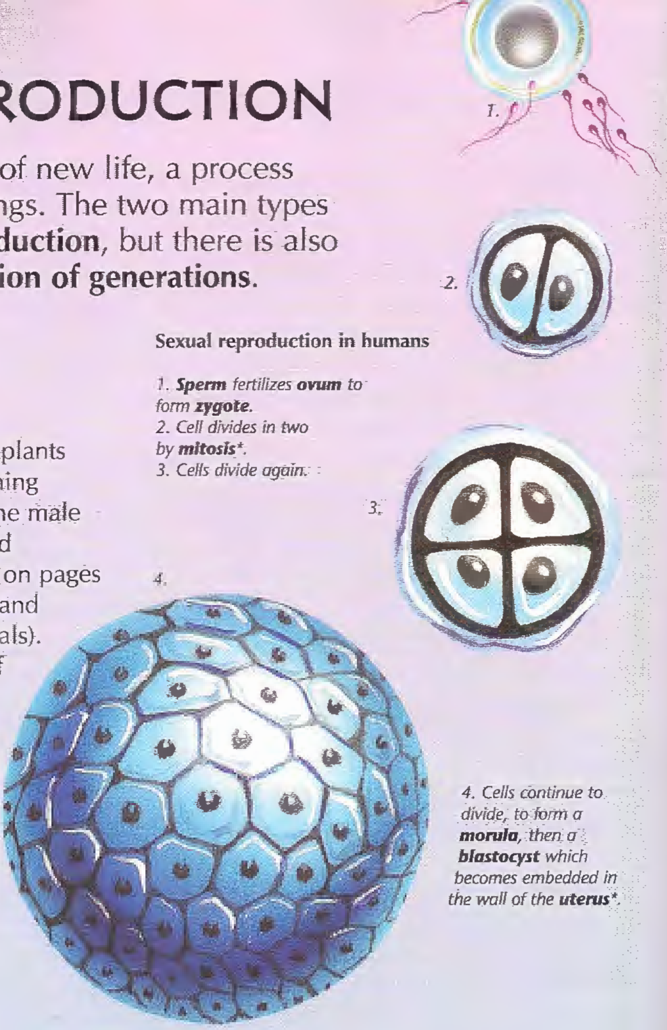

Sexual reproduction in humans

*1. **Sperm** fertilizes **ovum** to form **zygote**.*
*2. Cell divides in two by **mitosis***.*
3. Cells divide again.

*4. Cells continue to divide, to form a **morula**, then a **blastocyst** which becomes embedded in the wall of the **uterus***.*

Gametes or germ cells

The sex cells which join in **sexual reproduction** to form a new living thing. They are made by a special kind of cell division (see pages 322-323). In animals and simple plants, male gametes are known as **sperm**, short for **spermatozoa** (sing. **spermatozoon**) in animals and **spermatozooids** in simple plants. In flowering plants, they are just **nuclei*** (rather than cells) and are called **male nuclei** (see also pages 258 and 323). Female gametes are called **ova** (sing. **ovum**) or **egg cells** (egg cell is usually used in the case of plants). A sperm is smaller than an ovum and has a "tail" (**flagellum***).

Zygote

The first cell of a new living thing. It is formed when a male and female **gamete** join (see **sexual reproduction**).

Embryo

A new developing individual. It grows from one cell (the **zygote**) by a series of cell divisions (see pages 240-241) called **cleavage**. In humans, this first produces a ball of cells (**morula**) from the one original, and then a larger, hollow ball (**blastocyst**). After **implantation***, this is called the embryo. As it grows, the cells become **differentiated**, i.e. each develops into one kind of cell, e.g. a nerve cell.

*Human **embryo** (at eight weeks)*

---

* **Chromosomes**, 324; **Diploid number**, 240 (**Mitosis**); **Flagella**, 268; **Haploid number**, 322 (**Meiosis**); **Implantation**, 319 (**Fertilization**); **Nucleus**, 238; **Uterus**, 317.

Asexual reproc

Asexual reproduction i of reproduction, occurri plants and animals. The different types, e.g. **bina** organism dividing into t **vegetative reproducti** **sporulation**, but they al features. Firstly, only on and secondly, the new always genetically iden

Gemmation

Called **budding** in anima **asexual reproduction** o simple plants and anim It involves the formation cells which grows out of and develops into a new either breaks away from (in **colonial*** animals, e. stays attached (though s

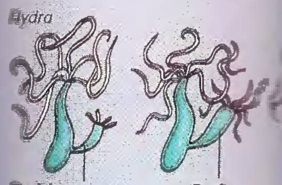

***Bud** forms on ent organism.*

***Bud** grows.*

Alternation of genera

A reproductive process simple animals and pla mosses. In the animals **sexual reproduction** a produced **asexually**. In the alternation is really of sexual reproduction (**gametophyte**) produce (**sporophyte**) by sexual then produces **spores** which grow into new g However, the spores are way as **gametes** (see pa they (and the gametoph the original number of c metophytes produce cell division (see pages no need to halve the ch

---

mosomes, 324; **Colonial**, **ploid number**, 322 (**Meiosis**

Asexual reproduction

Asexual reproduction is the simplest form of reproduction, occurring in many simple plants and animals. There are a number of different types, e.g. **binary fission** (a simple organism dividing into two identical ones), **vegetative reproduction***, **gemmation** and **sporulation**, but they all share two main features. Firstly, only one parent is needed and secondly, the new individual is always genetically identical to this parent.

Gemmation

Called **budding** in animals. A type of **asexual reproduction** occurring in many simple plants and animals, e.g. hydra. It involves the formation of a group of cells which grows out of the organism and develops into a new individual. It either breaks away from the parent or (in **colonial*** animals, e.g. corals) it stays attached (though self-contained).

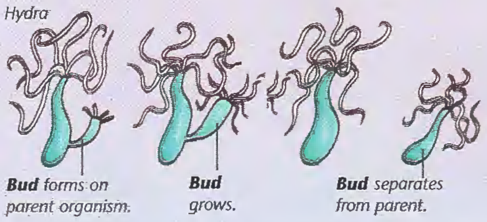
Hydra
Bud forms on parent organism.
Bud grows.
Bud separates from parent.

Sporulation

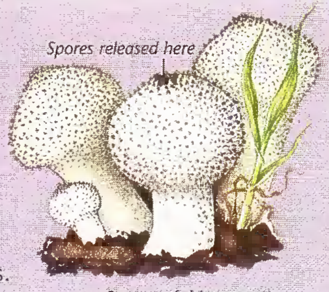
Spores released here

Spores of this common puffball (a complex fungus) disperse through a hole that forms in the ball.

The production of bodies called **spores** by simple plants, e.g. fungi and mosses. After dispersal by wind or water, these develop into new plants. There are two types of spores. One type is produced (e.g. in complex fungi, mosses and ferns) by a special kind of cell division (see pages 322-323) which is a feature of **sexual reproduction**. The new plants are not the same as the parent (see **alternation of generations**). Another kind of spore, however, is produced in plants such as simple fungi by ordinary cell division (see pages 240-241). The spores develop into plants which are identical to the parent (an important feature of **asexual reproduction**). Although only one parent is needed in both cases, true asexual reproduction really only occurs with the second type.

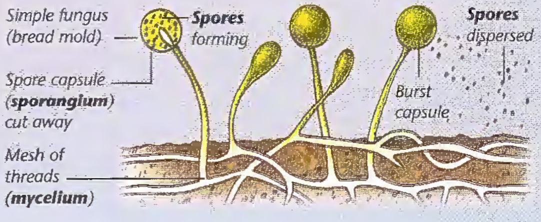
Simple fungus (bread mold)
Spores forming
Spore capsule (**sporangium**) cut away
Mesh of threads (**mycelium**)
Burst capsule
Spores dispersed

Alternation of generations

A reproductive process found in many simple animals and plants, e.g. jellyfish and mosses. In the animals, a form produced by **sexual reproduction** alternates with one produced **asexually**. In the plants, though, the alternation is really between two stages of sexual reproduction. One plant body (**gametophyte**) produces another (**sporophyte**) by sexual reproduction. This then produces **spores** (see **sporulation**) which grow into new gametophytes. However, the spores are made in the same way as **gametes** (see pages 322-323) and they (and the gametophytes) have only half the original number of **chromosomes***. The gametophytes produce gametes by ordinary cell division (see pages 240-241), as there is no need to halve the chromosomes again.

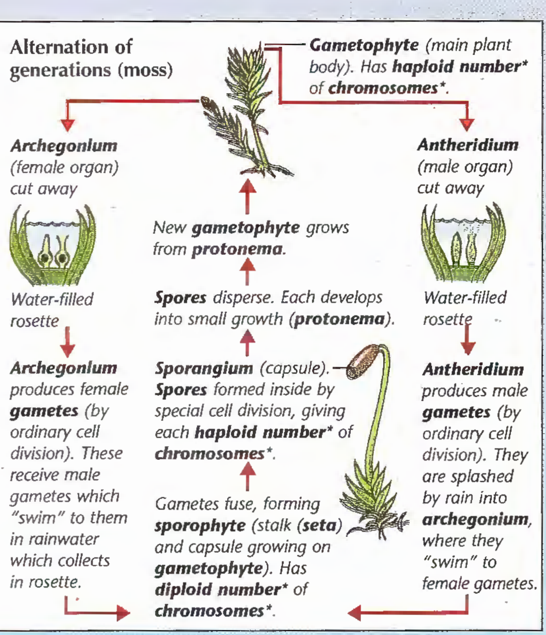
Alternation of generations (moss)
Gametophyte (main plant body). Has **haploid number*** of **chromosomes***.
Archegonium (female organ) cut away
Antheridium (male organ) cut away
New **gametophyte** grows from **protonema**.
Water-filled rosette
Water-filled rosette
Spores disperse. Each develops into small growth (**protonema**).
Archegonium produces female **gametes** (by ordinary cell division). These receive male gametes which "swim" to them in rainwater which collects in rosette.
Sporangium (capsule). **Spores** formed inside by special cell division, giving each **haploid number*** of **chromosomes***.
Gametes fuse, forming **sporophyte** (stalk (**seta**) and capsule growing on **gametophyte**). Has **diploid number*** of **chromosomes***.
Antheridium produces male **gametes** (by ordinary cell division). They are splashed by rain into **archegonium**, where they "swim" to female gametes.

* **Chromosomes**, 324; **Colonial**, 342; **Diploid number**, 240 (**Mitosis**); **Haploid number**, 322 (**Meiosis**); **Vegetative reproduction**, 263.

CELL DIVISION FOR REPRODUCTION

Many cells within a living thing can divide to produce new cells for growth and repair (see pages 240-241). There is, however, a second type of cell division, which happens specifically to produce the **gametes*** (sex cells) needed for **sexual reproduction*** (and also one of the two types of **spores***). The division of the **nucleus*** in this type of cell division is called **meiosis**. The production of gametes, including both the cell division and the subsequent maturing of the gametes, is called **gametogenesis**.

Meiosis

The division of the **nucleus*** when a cell divides to produce sex cells (see introduction). It can be split into two separate divisions – the **first meiotic division** (or **reduction division**) and the **second meiotic division** (each is followed by division of the **cytoplasm***). These can be divided into different phases (as in **mitosis***). Meiosis in general, and the first meiotic division in particular, ensures that each new **daughter nucleus** receives exactly half the number of **chromosomes*** as the original nucleus. The original number is the **diploid number** (see **mitosis**, page 240); the halved amount is the **haploid number**.

First meiotic division

These pictures show an animal cell, but only four **chromosomes*** are shown.

Prophase (early stage)

*Threads of **chromatin*** in **nucleus*** coil up to form **chromosomes***. Paired chromosomes (**homologous chromosomes**) line up side by side, forming pairs called **bivalents**. Each chromosome duplicates, becoming a pair of **chromatids** (each group of four chromatids now called a **tetrad**). **Centrioles*** move to opposite poles of cell.*

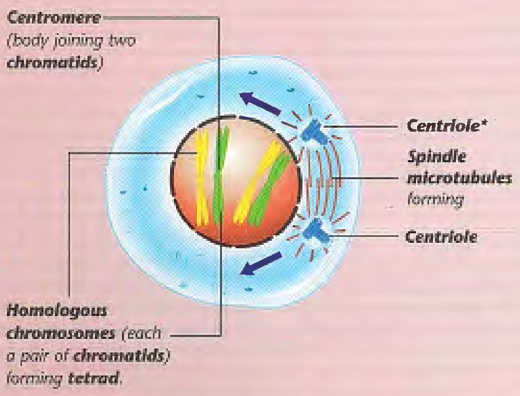
Centromere (body joining two **chromatids**)
Centriole*
Spindle microtubules forming
Centriole
Homologous chromosomes (each a pair of **chromatids**) forming **tetrad**.

Crossing over (occurs in early prophase)

Chromatids** of each **tetrad** cross over each other at places called **chiasmata** (sing. **chiasma**). Two chromatid pieces (one from each pair) break off and swap over. Causes mixing of **genes (helping to ensure new living things are never identical to parents, i.e. always a new variety of types).*

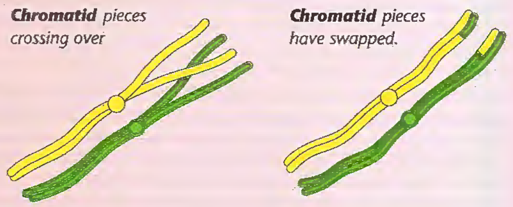
Chromatid pieces crossing over
Chromatid pieces have swapped.

Prophase (later stage)

***Homologous chromosomes** (each a pair of **chromatids**) move together to equator of cell.*

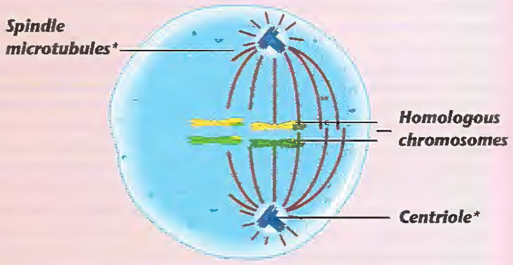
Spindle microtubules*
Homologous chromosomes
Centriole*

Metaphase

Nuclear membrane disappears, two **centrioles*** form a **spindle** (see **metaphase** of **mitosis**, page 241). **Chromosomes*** (pairs of **chromatids**) become attached to spindle by **centromeres**.*

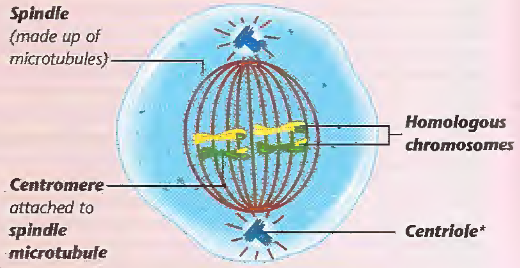
Spindle (made up of microtubules)
Homologous chromosomes
Centromere attached to **spindle microtubule**
Centriole*

Anaphase

***Homologous chromosomes** (eac
separate (see **Law of segregatio**
by **spindle microtubules**.*

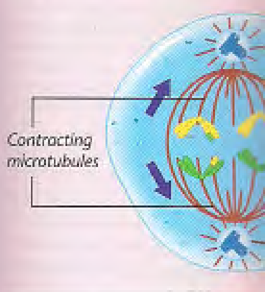
Contracting microtubules

Second meiot

The **second meiotic di**
cells produced by the
occurs in exactly the s
same phases, as **mitosi**
divides as part of cell
repair) and is followed
the division of the **cyto**

Gamete production (male)

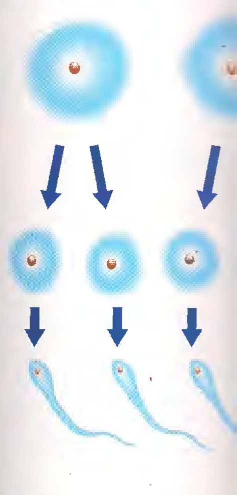

*Two cells formed from **first meio**
second meiotic division). In an
spermatids and mature into ma
sperm. In simple plants, four ce
into type of **spore*** involved in
In flowering plants, **nuclei*** of fo
(**mitosis***). Resulting cells (**polle**
(one later divides again to form*

---

* **Centrioles**, 240; **Chromatin**, 238 (**Nucleus**); **Chromosomes**, 324; **Cytoplasm**, 238; **Gametes**, 320; **Genes**, 325; **Mitosis**, 240; **Nuclear membrane**, 238 (**Nucleus**); **Sexual reproduction**, 320; **Spindle microtubules**, 241; **Spores**, 321 (**Sporulation**).

* **Alternation of generations**, 3
Cytoplasm, 238; **Gametes**, 320
Nuclear membrane, 238 (**Nuc**

Anaphase

Homologous chromosomes *(each still a pair of **chromatids**) separate (see **Law of segregation**, page 326), dragged apart by **spindle microtubules**.*

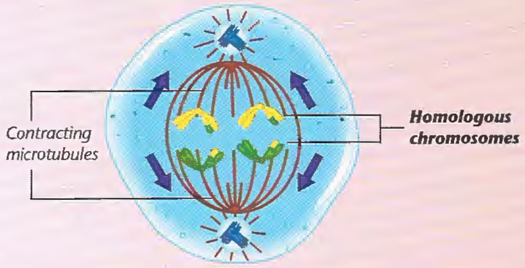
Contracting microtubules
Homologous chromosomes

Telophase

Spindle** disappears. Happens in conjunction with **cytokinesis** (division of **cytoplasm). Two new cells formed, each with half the original number of **chromosomes*** (each two **chromatids**). Brief **interphase*** (intervening period) usually follows, in which case **nuclear membranes*** form and chromosomes uncoil again to form thread-like mass (**chromatin***).*

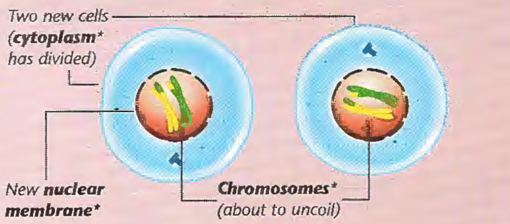
Two new cells (cytoplasm* has divided)
New nuclear membrane*
Chromosomes* (about to uncoil)

Second meiotic division

The **second meiotic division** happens in the cells produced by the **first meiotic division**. It occurs in exactly the same way, and with the same phases, as **mitosis*** (when the **nucleus*** divides as part of cell division for growth and repair) and is followed in the same way by the division of the **cytoplasm***.

The only difference is that each dividing nucleus now has only the **haploid number** of **chromosomes*** (see **meiosis**) so the resulting new sex cells (**gametes***) will also be haploid. The second division differs according to whether male or female gametes are to be produced, and the final maturing of the gametes after the second division is different in animals and plants (see text below).

Gamete production (male)

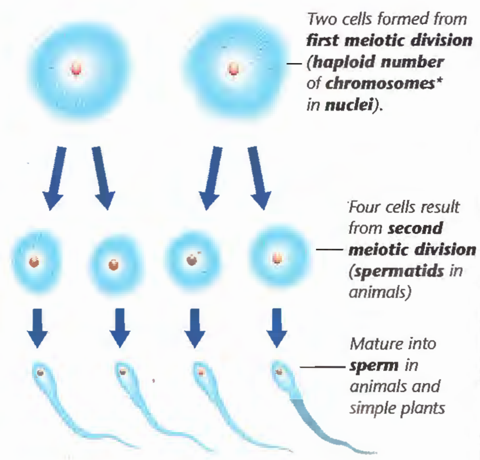
Two cells formed from first meiotic division (haploid number of chromosomes* in nuclei).
Four cells result from second meiotic division (spermatids in animals)
Mature into sperm in animals and simple plants

*Two cells formed from **first meiotic division** divide again (see **second meiotic division**). In animals, resulting four cells called **spermatids** and mature into male **gametes*** (sex cells), or **sperm**. In simple plants, four cells either develop into sperm or into type of **spore*** involved in **alternation of generations***. In flowering plants, **nuclei*** of four cells each divide again (**mitosis***). Resulting cells (**pollen*** grains) each have two nuclei (one later divides again to form two **male nuclei***).*

Gamete production (female)

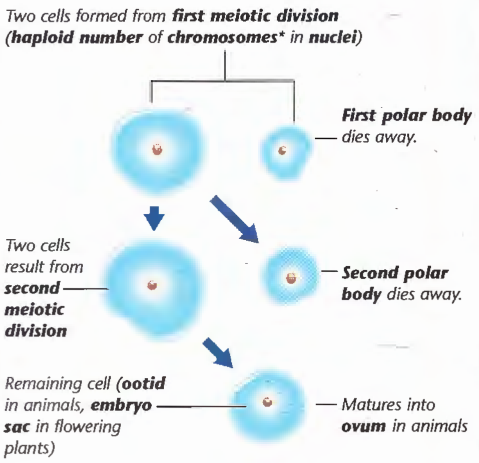
Two cells formed from first meiotic division (haploid number of chromosomes* in nuclei)
First polar body dies away.
Two cells result from second meiotic division
Second polar body dies away.
Remaining cell (ootid in animals, embryo sac in flowering plants)
Matures into ovum in animals

*One of two cells formed by **first meiotic division** dies away (called **first polar body**). Other divides again (see **second meiotic division**). Of two resulting cells, one (**second polar body**) dies away. In animals, other one called **ootid** and matures into female **gamete*** (sex cell), or **ovum**. In flowering plants, other one called **embryo sac** and its **nucleus*** divides three more times (by **mitosis***). Of eight new nuclei, six have cells form around them, two stay naked. One of six cells is female gamete, or **egg cell** (see **ovule**, page 258). Formation of egg cell in simple plants is very similar.*

* **Alternation of generations**, 321; **Centrioles**, 240; **Chromatin**, 238 (**Nucleus**); **Chromosomes**, 324; **Cytoplasm**, 238; **Gametes**, 320; **Interphase**, 240; **Male nuclei**, 320 (**Gametes**); **Mitosis**, 240; **Nuclear membrane**, 238 (**Nucleus**); **Pollen**, 258; **Spores**, 321 (**Sporulation**).

GENETICS AND HEREDITY

Genetics is a branch of biology. It is the study of **inheritance** – the passing of characteristics from one generation to the next. The bodies which are instrumental in this process are called **chromosomes**. Each chromosome is made up of **genes** – the "coded" instructions for the appearance and constituents of an organism. For more about inheritance, see page 326.

Chromosomes

Structures present at all times in the **nuclei*** of all cells, though they only become independently visible (as thread-like bodies of differing shapes and sizes) when a cell is dividing (and has been stained with a dye). Each one is made with a single molecule of **DNA** (see **nucleic acids**, below), plus proteins called **histones**. The DNA molecule is a chain of many connected **genes**.

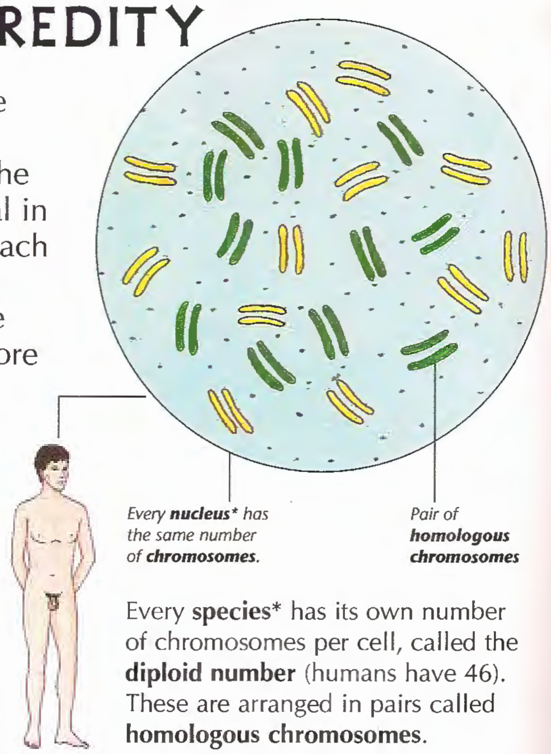
*Every **nucleus*** has the same number of **chromosomes**.*
*Pair of **homologous chromosomes***

Every **species*** has its own number of chromosomes per cell, called the **diploid number** (humans have 46). These are arranged in pairs called **homologous chromosomes**.

Nucleic acids

Two different acids, called **DNA** (**deoxyribonucleic acid**) and **RNA** (**ribonucleic acid**). Both are found in the **nuclei*** of all cells, hence their name (RNA is also found in the **cytoplasm*** – see **ribosomes**, page 239).

Each molecule of a nucleic acid is very large, and is composed of many individual units called **nucleotides**. A DNA molecule consists of two chains of nucleotides twisted around each other, forming a shape called a **double helix** (rather like a twisted ladder). An RNA molecule consists of one chain of nucleotides (and looks like a ladder cut in half lengthwise and twisted).

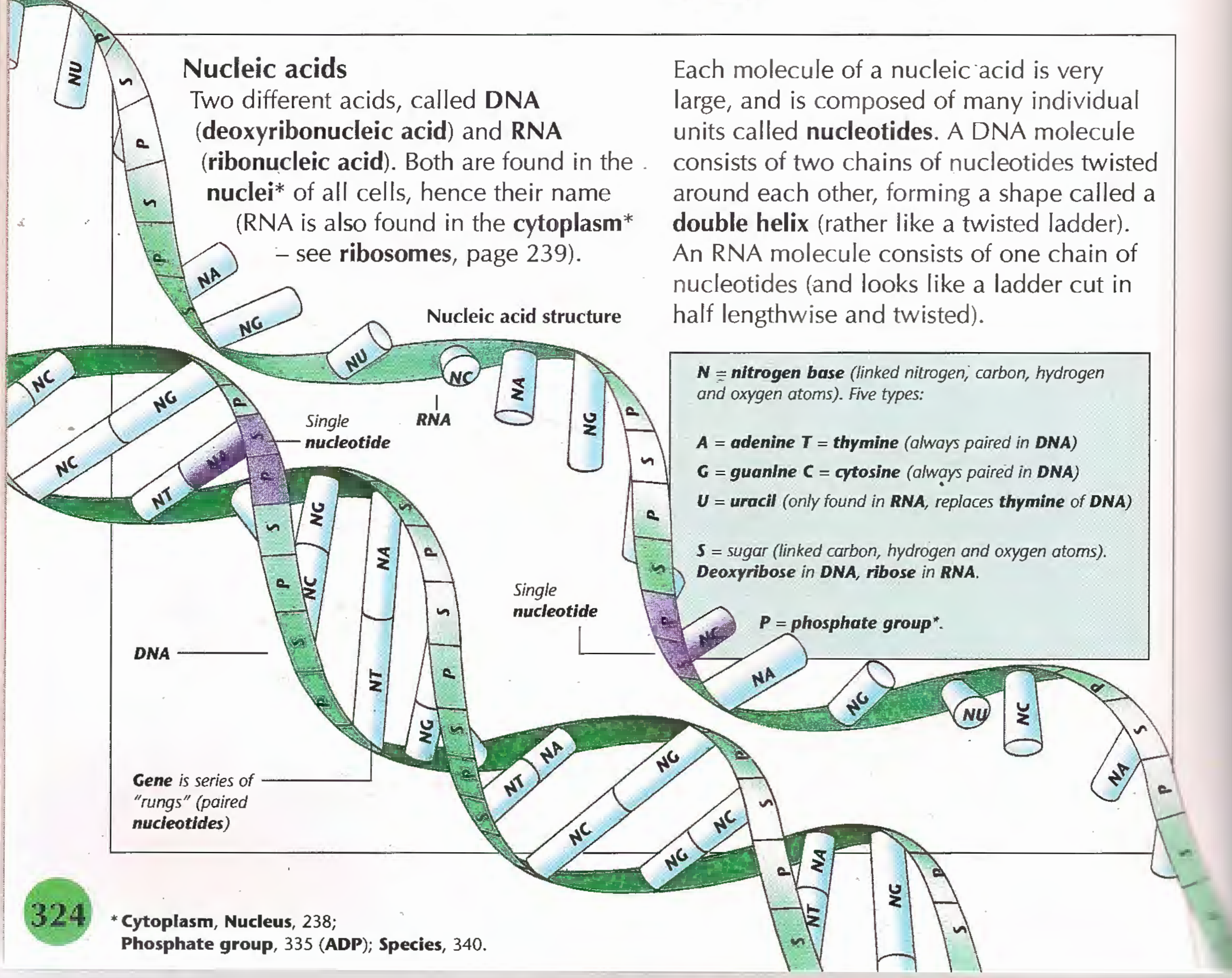
Nucleic acid structure
Single **nucleotide**
RNA
Single **nucleotide**
DNA
***Gene** is series of "rungs" (paired **nucleotides**)*
NU
NC
NG
NA
NT
P
S

***N = nitrogen base** (linked nitrogen, carbon, hydrogen and oxygen atoms). Five types:*

***A = adenine T = thymine** (always paired in **DNA**)*

***G = guanine C = cytosine** (always paired in **DNA**)*

***U = uracil** (only found in **RNA**, replaces **thymine** of **DNA**)*

***S** = sugar (linked carbon, hydrogen and oxygen atoms). **Deoxyribose** in **DNA**, **ribose** in **RNA**.*

P = phosphate group.*

---

* **Cytoplasm**, **Nucleus**, 238;
Phosphate group, 335 (**ADP**); **Species**, 340.

Genes

Sets of "coded" instruct
the **DNA** molecule of a
humans, each DNA mo
contain about 1,000 ge
connected series of abo
DNA "ladder". Since th
varies, each gene has a
relating to one specific
of the organism, e.g. its
the composition of a **ho**
exception of the **sex chr**
carried on paired **homol**
(see **chromosomes**) are
down the chromosomes
(one member of each pa
chromosome). These pai
same characteristic and
instructions. However, t
also be different, in whic
from one gene (the **dom**
out" those from the othe
unless **incomplete domin**
is shown. Two such non
called **alleles** or **allelom**

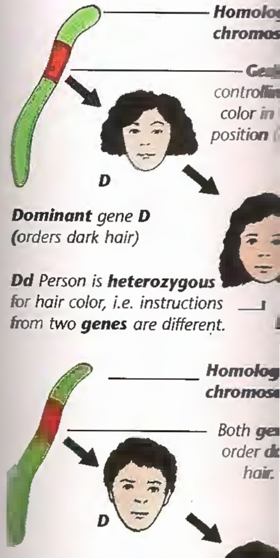

***Dominant** gene **D** (orders dark hair)*

***Dd** Person is **heterozygous** for hair color, i.e. instructions from two **genes** are different.*

*Person is **homozygous** for hair color, i.e. instructions from two genes are identical.*

*The two examples have different ge different sets of instructions (**DD** and **phenotype**, i.e. the resulting charac*

* **ABO blood groups**, 287; **Horm**

Genes

Sets of "coded" instructions which make up the **DNA** molecule of a **chromosome** (in humans, each DNA molecule is thought to contain about 1,000 genes). Each gene is a connected series of about 250 "rungs" on the DNA "ladder". Since the order of the "rungs" varies, each gene has a different "code", relating to one specific characteristic (**trait**) of the organism, e.g. its **blood group*** or the composition of a **hormone***. With the exception of the **sex chromosomes**, the genes carried on paired **homologous chromosomes** (see **chromosomes**) are also paired, and run down the chromosomes in the same order (one member of each pair on each chromosome). These paired genes control the same characteristic and may give identical instructions. However, their instructions may also be different, in which case the instructions from one gene (the **dominant** gene) will "mask out" those from the other (the **recessive** gene), unless **incomplete dominance** or **codominance** is shown. Two such non-identical genes are called **alleles** or **allelomorphs**.

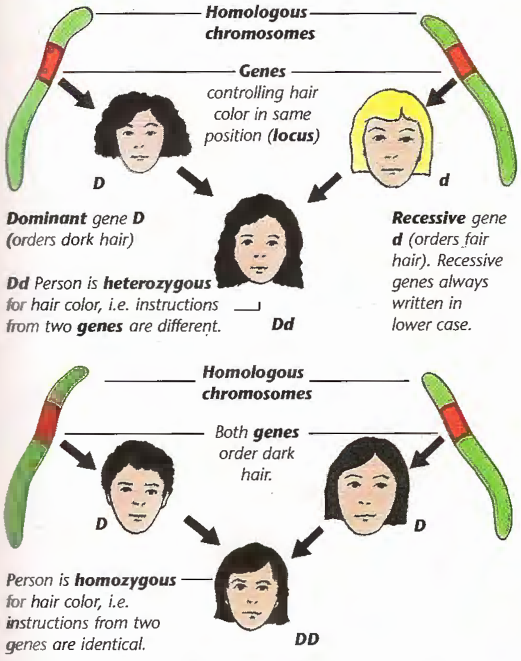
Homologous chromosomes
Genes controlling hair color in same position (**locus**)
D
d
Dominant gene **D** (orders dark hair)
Recessive gene **d** (orders fair hair). Recessive genes always written in lower case.
Dd Person is **heterozygous** for hair color, i.e. instructions from two **genes** are different.
Dd
Homologous chromosomes
Both **genes** order dark hair.
D
D
Person is **homozygous** for hair color, i.e. instructions from two genes are identical.
DD

*The two examples have different **genotypes** for hair color, i.e. different sets of instructions (**DD** and **Dd**), but are the same **phenotype**, i.e. the resulting characteristic is the same (dark hair).*

Incomplete dominance or blending

A situation where a pair of **genes** which control the same characteristic give different instructions, but neither is **dominant** (see **genes**) or obvious in the result. For example, a lack of dominance between a gene for red color and one for white results in the intermediate roan color of some cows.

***Incomplete dominance** can be used by gardeners to produce flowers of the same species but with a variety of colors. They do this by cross-pollinating flowers of different colors.*

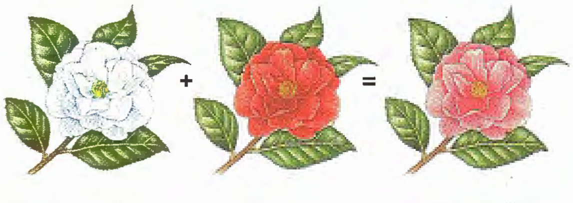

White camellia *Red camellia* *Pink camellia*

Codominance

A special situation where a pair of **genes** controlling the same characteristic give different instructions, neither is **dominant** (see **genes**), but both are represented in the result. The human **blood group*** AB, for example, results from equal dominance between a gene for group A and one for group B.

Sex chromosomes

One pair of **homologous chromosomes** (see **chromosomes**) in all cells (all the others are called **autosomes**). There are two different kinds of sex chromosomes, called the **X** and **Y chromosomes**. A male has one X and one Y. The Y chromosome carries the genetic factor (not a **gene** as such) determining maleness, thus all individuals with two X chromosomes are female.

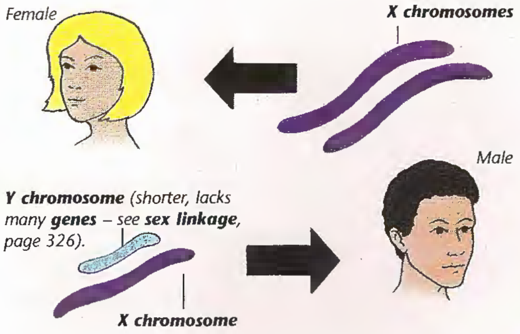
Female
X chromosomes
Male
Y chromosome (shorter, lacks many **genes** – see **sex linkage**, page 326).
X chromosome

---

* **ABO blood groups**, 287; **Hormones**, 336.

Inheriting genes

Every new organism inherits its **chromosomes*** (and **genes***) from its parents. In **sexual reproduction***, the **sperm*** and **ovum*** (sex cells) which come together to form this new individual have only half the normal number of chromosomes (the **haploid number** – see pages 322-323). This ensures that the **zygote*** (first new cell) formed from the two sex cells will have the normal number (see **chromosomes**, page 324). Two laws (**Mendel's laws**) point out genetic factors which are always true when cells divide to produce sex cells.

Law of segregation (Mendel's first law)

Homologous chromosomes* always separate when the **nucleus*** of a cell divides to produce **gametes*** (sex cells – see pages 322-323), hence so too do the paired **genes*** which control the same characteristic. The offspring thus always have paired genes (one member of each pair coming from each parent).

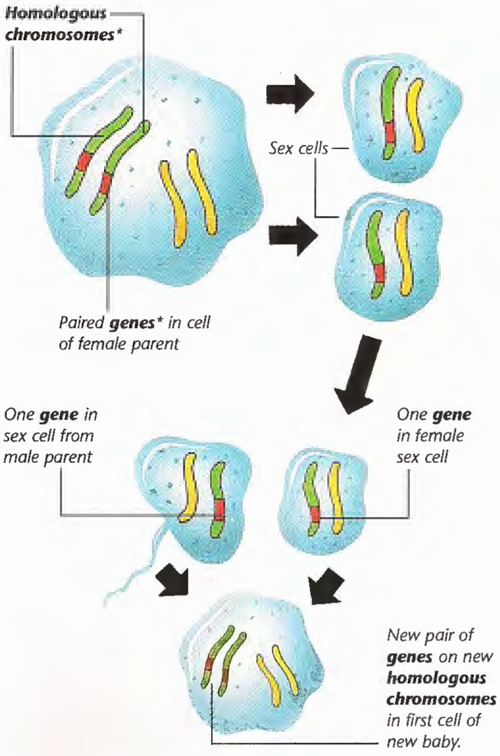

Homologous chromosomes*
Sex cells
Paired ***genes**** in cell of female parent
One ***gene*** in sex cell from male parent
One ***gene*** in female sex cell
New pair of ***genes*** on new ***homologous chromosomes*** in first cell of new baby.

Law of independent assortment (Mendel's second law)

Each member of a pair of **genes*** can join with either of the two members of another pair when a cell divides to form **gametes*** (sex cells). Hence all the different mixes are possible in a new individual.

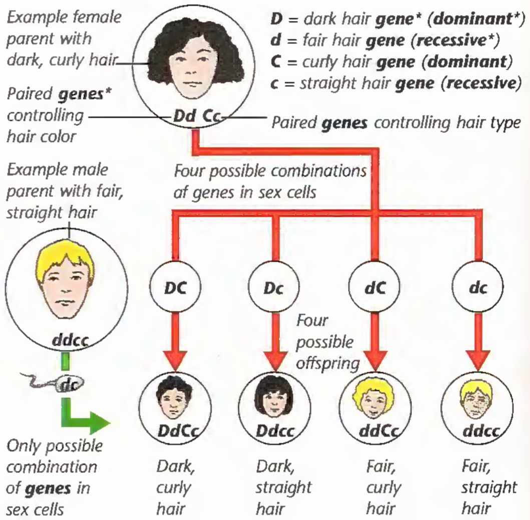

Example female parent with dark, curly hair
Paired ***genes**** controlling hair color
Dd Cc
D = dark hair ***gene* (dominant*)***
d = fair hair ***gene (recessive*)***
C = curly hair ***gene (dominant)***
c = straight hair ***gene (recessive)***
Paired ***genes*** controlling hair type
Example male parent with fair, straight hair
ddcc
dc
Only possible combination of ***genes*** in sex cells
Four possible combinations of genes in sex cells
DC
Dc
dC
dc
Four possible offspring
DdCc
Ddcc
ddCc
ddcc
Dark, curly hair
Dark, straight hair
Fair, curly hair
Fair, straight hair

Sex linkage

The two **sex (X) chromosomes*** in a female contain many paired **genes*** (like all **chromosomes***), but the **Y chromosome*** in a male lacks partners for most of the genes on its mate (the X). Thus any **recessive*** genes on the X will show up more often in males (see below). The unpaired genes on the X are called **sex-linked genes**.

*Example: A **gene*** related to sight is found on the **X chromosome****

C *= normal sight **gene*** (**dominant***)*
c *= color-blindness **gene** (**recessive***)*

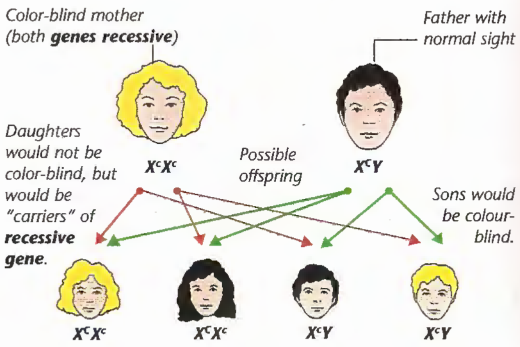

Color-blind mother (both ***genes recessive***)
Father with normal sight
X^cX^c
X^CY
Possible offspring
Daughters would not be color-blind, but would be "carriers" of ***recessive gene***.
Sons would be colour-blind.
X^CX^c
X^CX^c
X^cY
X^cY

---

* **Dominant**, 325 (**Genes**); **Homologous chromosomes**, 324 (**Chromosomes**); **Nucleus**, 238; **Ovum**, 320 (**Gametes**); **Recessive**, 325 (**Genes**); **Sexual reproduction**, 320; **Sperm**, 320 (**Gametes**); **X** and **Y chromosomes**, 325 (**Sex chromosomes**); **Zygote**, 320.

GENETIC

Genetic enginee

nucleus* in orde
used to create ne
medicine and ind
being discovered

Gene cloning

The main technique o
Desirable **genes*** are
by inserting DNA mo
the genes) into other
fast-breeding **bacteria**,
the DNA. Gene cloni
The most common me

Gene cloning

*1. DNA containing a particular desirable **gene***, known as **target DNA**, is removed from a donor cell.*

*2. Some bacteria contain a **pl***
– a ring of DNA separate from
*bacteria's **chromosomes***. Pl*
are capable of inserting themsel
into other organisms. Some als
*resistance to particular **antibi***
(drugs that destroy bacteria).
Plasmids can be obtained by bre
up bacteria that contain them.

*3. The strands of **plasmid** D*
*and **target DNA** are treated s*
that the ends are sticky. When
placed together and heated, th
plasmid DNA and target DNA
*join together. This is called **gene***
***splicing**. The new DNA is know*
*as **recombinant DNA**.*

4. The new plasmids insert themselves into new bacteria not resistant to the antibiotic.

5. As a colony of the bacteria grows, it is treated with this antibiotic. Any bacteria without the new plasmid are destroyed by the antibiotic. Bacteria containing it continue to multiply.

6. The resultant colony is now m
carrying the antibiotic-resistant p
(containing the desired gene). Th
many times, producing an enorm

* **Chromosomes**, 324; **DNA**, 32
Genes, 325; **Insulin**, 336; **Nuc**

GENETIC ENGINEERING

Genetic engineering is the deliberate alteration of **DNA*** within a cell **nucleus*** in order to modify an organism or population of organisms. It is used to create new products which are beneficial to science, agriculture, medicine and industry. New uses for genetically engineered organisms are being discovered all the time.

Gene cloning

The main technique of genetic engineering. Desirable **genes*** are duplicated artificially by inserting DNA molecules (containing the genes) into other organisms, such as fast-breeding bacteria, which then reproduce the DNA. Gene cloning is a complex process. The most common method is shown below.

Gene cloning

*1. DNA containing a particular desirable **gene***, known as **target DNA**, is removed from a donor cell.*

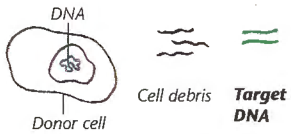

*2. Some bacteria contain a **plasmid** – a ring of DNA separate from the bacteria's **chromosomes***. Plasmids are capable of inserting themselves into other organisms. Some also give resistance to particular **antibiotics** (drugs that destroy bacteria). Plasmids can be obtained by breaking up bacteria that contain them.*

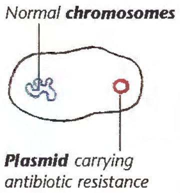

*3. The strands of **plasmid** DNA and **target DNA** are treated so that the ends are sticky. When placed together and heated, the plasmid DNA and target DNA join together. This is called **gene splicing**. The new DNA is known as **recombinant DNA**.*

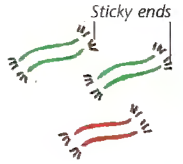

4. The new plasmids insert themselves into new bacteria not resistant to the antibiotic.

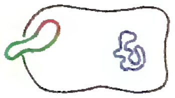

5. As a colony of the bacteria grows, it is treated with this antibiotic. Any bacteria without the new plasmid are destroyed by the antibiotic. Bacteria containing it continue to multiply.

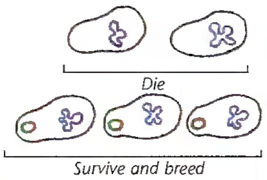

6. The resultant colony is now made up exclusively of bacteria carrying the antibiotic-resistant plasmids with the target DNA (containing the desired gene). This colony can now be multiplied many times, producing an enormous quantity of the gene.

Uses of genetic engineering

Pharming

The use of plants or animals to produce genetically modified pharmaceutical products. For example, a sheep has been genetically engineered to produce milk which contains **alpha-1 antitrypsin**, a drug which is beneficial to cystic fibrosis patients.

Protein manufacture

The production, in specially-created bacteria "factories", of medically useful proteins such as **insulin*** to help diabetics, and **antihemophilic globulin** to treat people with hemophilia.

Genetically engineered crops

Plants which have been bred with a greater resistance to disease, pesticides and weather, by inserting foreign **genes*** into their **nuclei***. An example of this is shown below.

Some flounders have an antifreeze chemical in their blood, which helps them to survive in freezing water.

*The **gene*** which gives their blood this property can be extracted and introduced into tomato plants. Tomatoes from such plants are now more able to withstand frost and snow. This makes them more readily available at the extremes of their normal growing season.*

***Chromosomes**, 324; **DNA**, 324 (**Nucleic acids**); **Genes**, 325; **Insulin**, 336; **Nucleus**, 238.

Uses of genetic engineering - continued

Animal cloning

Producing a genetically identical duplicate, or **clone**, of an animal. In 1997, scientists took a cell from a female sheep and placed its **chromosomes*** into another sheep's **ovum***, which had had its own chromosomes removed. The ovum was planted in the second sheep's womb, and five months later, a lamb, known to the world as Dolly, was born. This experiment proved it was possible to produce a complex living organism, without any kind of **sexual reproduction***.

Sheep created by sexual reproduction*

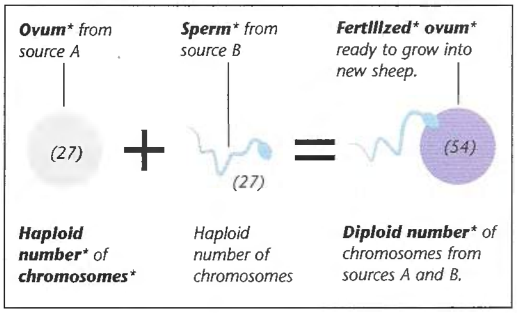
Ovum* from source A*
Sperm* from source B*
Fertilized* ovum* ready to grow into new sheep.*
(27)
+
(27)
=
(54)
Haploid number* of **chromosomes****
Haploid number of chromosomes
Diploid number* of chromosomes from sources A and B.*

Sheep created by gene cloning*

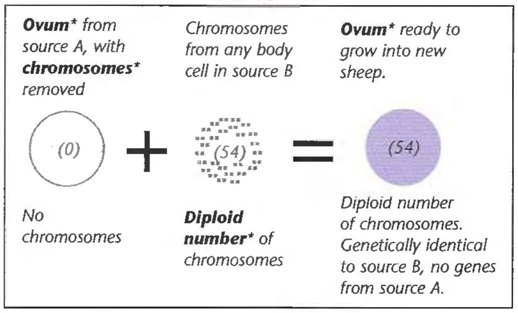
Ovum* from source A, with **chromosomes**** removed*
Chromosomes from any body cell in source B
Ovum* ready to grow into new sheep.*
(0)
+
(54)
=
(54)
No chromosomes
Diploid number* of chromosomes*
Diploid number of chromosomes. Genetically identical to source B, no genes from source A.

Genetics and the future

Human cloning

The theoretical creation of human life, using the same method that created Dolly the sheep. Genetic engineering techniques are now so advanced that any one of a human body's 100 million million cells could be used to create a new human. But, just as appearance, character and intellect are slightly different in identical twins, a **clone** of another human being would not be a replica, only a person with identical **genes***. A clone would also be a generation apart in age. Possible uses of this technology include the treatment of **infertile** couples – couples who are not naturally able to have children.

Genome mapping

Making a detailed list of the **nucleotides*** contained in the **genome** (genetic code) of any organism. Scientists have already mapped the genome of a yeast cell, and are currently mapping the three billion nucleotides contained in a human genome. They intend to complete this task early in the twenty-first century. The resulting map will enable them to identify every **gene*** in human **chromosomes*** and understand what each one does.

Genetic diagnosis

The identification of illness by examination of **genes***. Scientists can already identify some genetic disorders, which show up as irregularities in the **nucleotide*** sequence. For example, **Huntingdon's chorea** (an illness which causes gradual physical and mental deterioration) can now be detected in a **fetus***. Such research could also make it possible to identify a gene which makes people more susceptible to some cancers. Once identified, treatment could be applied to prevent the cancer from developing.

Organ modification

Introducing genes which encourage body organs to heal themselves. One new technique encourages the hearts of patients needing bypass surgery to grow new blood vessels themselves.

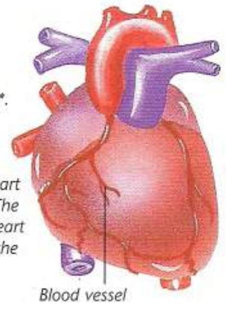
Blood vessel

*When the heart is growing inside an **embryo***, a particular **gene*** instructs it to construct its **arteries***. The instructions cease once the heart is fully grown. Scientists are developing a method of re-introducing this gene into the heart of a patient with blocked arteries. The gene would enable the damaged heart to grow new blood vessels around the constricted artery, so removing the need for a major heart operation.*

---

* **Arteries**, 288; **Chromosomes**, 324; **Diploid number**, 240 (**Mitosis**); **Embryo**, 320; **Fertilization**, 319; **Fetus**, 319 (**Pregnancy**); **Genes**, 325; **Gene cloning**, 327; **Haploid number**, 322 (**Meiosis**); **Nucleotides**, 324 (**Nucleic acids**); **Ovum**, 320 (**Gametes**); **Sexual reproduction**, 320; **Sperm**, 320 (**Gametes**).

FLUID M

The movement of
around the body,
movement in and
essential to the life
Food matter must
into the cells, and
harmful material m
move out. Most so
travel around the b
i.e. they (**solutes**) a
fluid (the **solvent** –

Diffusion

The movement of mole
substance from an area
higher concentration to
concentration is lower.
process (where the conc
solute is low, that of the
high, so its molecules w
way) and it ceases when
evenly distributed. Many
oxygen and carbon diox
and out of cells.

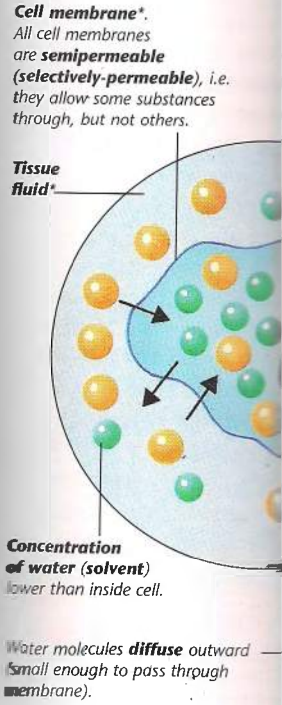
Cell membrane*. All cell membranes are **semipermeable** (**selectively-permeable**), i.e. they allow some substances through, but not others.*
Tissue fluid*
***Concentration of water** (**solvent**) lower than inside cell.*
*Water molecules **diffuse** outward (small enough to pass through membrane).*

---

* **ATP**, 335; **Cell membrane**, 238;
Tissue fluid, 292; **Vacuoles**, 238.

FLUID MOVEMENT

The movement of substances around the body, especially their movement in and out of cells, is essential to the life of an organism. Food matter must be able to pass into the cells, and waste and harmful material must be able to move out. Most solids and liquids travel around the body in **solutions**, i.e. they (**solutes**) are dissolved in a fluid (the **solvent** – normally water).

Diffusion

The movement of molecules of a substance from an area where they are in higher concentration to one where their concentration is lower. This is a two-way process (where the concentration of a **solute** is low, that of the **solvent** will be high, so its molecules will move the other way) and it ceases when the molecules are evenly distributed. Many substances, e.g. oxygen and carbon dioxide, diffuse into and out of cells.

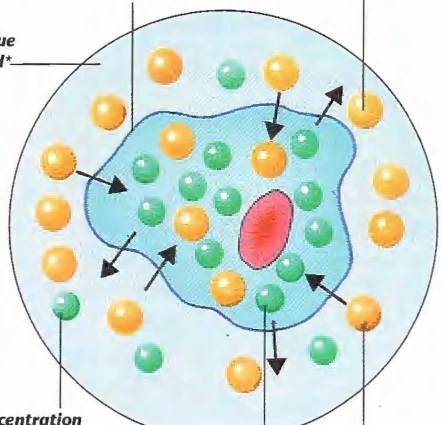

***Cell membrane**.** *All cell membranes are **semipermeable** (**selectively-permeable**), i.e. they allow some substances through, but not others.*
*Concentration of oxygen (**solute**) higher than inside cell.*
Tissue fluid*
***Concentration of water (solvent)** lower than inside cell.*
*Water molecules **diffuse** outward (small enough to pass through membrane).*
*Oxygen molecules **diffuse** into cell (small enough to pass through membrane).*

Osmosis

The movement of molecules of a **solvent** through a **semipermeable** membrane (see below, left) which lowers the concentration of a **solute** on the other side of the membrane, and evens out the concentrations either side. This is a one-way type of **diffusion**, occurring when the molecules of the solute cannot pass the other way. **Osmotic pressure** is the pressure which builds up in an enclosed space, e.g. a cell, when a **solvent** enters by osmosis.

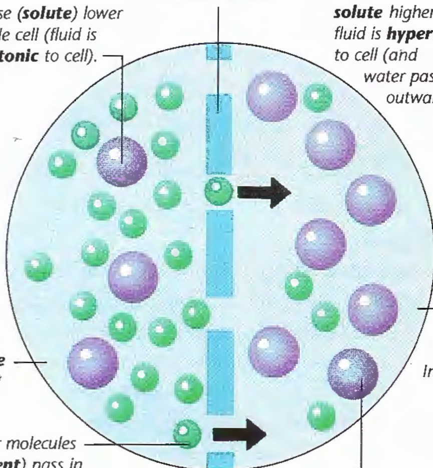

*Concentration of glucose (**solute**) lower outside cell (fluid is **hypotonic** to cell).*
Cell membrane*
*If concentration of **solute** higher outside, fluid is **hypertonic** to cell (and water passes outward).*
Tissue fluid*
Inside cell
*Water molecules (**solvent**) pass in (**osmosis**) until concentrations evened out (when fluid outside becomes **isotonic** to cell).*
Glucose molecules cannot leave.

Active transport

A process which occurs when substances have to be "pumped" in the opposite direction to that in which they would travel by **diffusion** (i.e. from low to high concentration), e.g. when cells take in large amounts of glucose for breakdown. It is not yet fully understood, but it is thought that special "carrier" molecules within the **cell membrane*** pick up molecules of solute, carry them through the membrane and release them. Energy is needed for this action (since it opposes the natural tendency). This is supplied in the form of **ATP***.

Pinocytosis

The taking in of a fluid droplet by inward-folding and separation of a section of **cell membrane*** (forming a **vacuole***). Most cells can do this.

* **ATP**, 335; **Cell membrane**, 238; **Tissue fluid**, 292; **Vacuoles**, 238.

FOOD AND HOW IT IS USED

Food is vital to all organisms, providing all the materials needed to be broken down for energy, to regulate cellular activities and to build and repair tissues (see pages 332-335). Of the various food substances, **carbohydrates**, **proteins** and **fats** are called **nutrients**, and **minerals**, **vitamins** (not needed by plants) and water are **accessory foods**. Plants build their own nutrients (by **photosynthesis***), and take in minerals and water; animals take in all the substances they need and break them down by digestion (see pages 338-339).

Carbohydrates

A group of substances made up of carbon, hydrogen and oxygen, which exist in varying degrees of complexity (see "terms used", page 339 and also page 204). In animals, complex carbohydrates are taken in and broken down by digestion (see page 338) into the simple carbohydrate **glucose**. The breakdown of glucose (**internal respiration***) provides almost all the energy for life's activities. Plants build up glucose from other substances (by **photosynthesis***).

Proteins

A group of substances made up of simpler units called **amino acids**. These contain carbon, hydrogen, oxygen, nitrogen and, in some cases, sulfur. Most protein molecules consist of hundreds, maybe thousands, of amino acids, joined together by links called **peptide links** into one or more chains called **polypeptides***. The many different types of protein each have a different arrangement of amino acids. They include the **structural proteins** (the basic components of new cells) and **catalytic proteins** (**enzymes***), which play a vital role in controlling cell processes.

Plants build up amino acids from the substances they take in (by **photosynthesis***), and then build proteins from them. Animals take in proteins and break them down into single amino acid molecules by digestion (see page 338). These are then transported in the blood to all the body cells and reassembled into the different proteins needed (see **ribosomes**, page 239 and also page 205).

Fats

A group of substances made up of carbon, hydrogen and a small amount of oxygen (see also **lipids**, page 205). Plants build fats from the substances they take in, and their seeds hold most as a store of food. This can be converted to extra **glucose** (see **carbohydrates**) to provide energy for the growing plant. Digestion of fats in animals produces **fatty acids** and **glycerol** (see page 338). If these need to be broken down for energy (as well as glucose), this occurs in the liver. This results in some products which the liver can convert to glucose, but others it cannot. These are instead converted elsewhere to a substance which forms a later stage of glucose breakdown.

Fatty acids and glycerol not needed for energy are immediately recombined to form fat particles and stored in various body areas, e.g. under the skin (see **subcutaneous layer**, page 310).

*Like all animals, humans cannot build thier own **nutrients**, and rely on the food they eat for energy. This comes either from plants, e.g. fruit and vegetables (see picture, left), or animals, e.g. meat and milk.*

---

* **Enzymes**, 333; **Internal respiration**, 334; **Photosynthesis**, 254; **Polypeptides**, 339.

Roughage or fiber

Bulky, fibrous food, e.
as lentils and beans.
cellulose, a **carbohyd**
walls*. Unlike most
cannot be digested by
including humans, bec
necessary **digestive en**
(Some animals, e.g.
enzyme, and others,
digest cellulose, do so
rumen, page 271.) The
bulky and coarse mean
gripped by intestinal
on through the diges

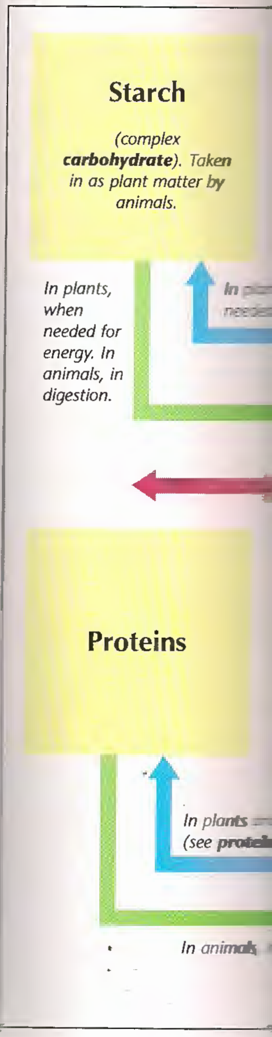
Starch
(complex **carbohydrate**). Taken in as plant matter by animals.
In plants, when needed for energy. In animals, in digestion.
Proteins
In plants
(see
In animals

Cell wall, 238; **Co-enzymes**, 333
Digestive enzymes, 338;

Roughage or **fiber**

Bulky, fibrous food, e.g. bran, and pulses such as lentils and beans. Much of it is made up of **cellulose**, a **carbohydrate** found in plant **cell walls***. Unlike most carbohydrates, cellulose cannot be digested by most animals, including humans, because they lack the necessary **digestive enzyme***, called **cellulase**. (Some animals, e.g. snails, do have this enzyme, and others, like cows, who must digest cellulose, do so in another way – see **rumen**, page 271.) The fact that roughage is bulky and coarse means that food can be gripped by intestinal muscles, and so moved on through the digestive system.

Vitamins

A group of substances vital to animals, though only needed in tiny amounts. The most important function of many vitamins is to act as **co-enzymes***, i.e. to help **enzymes*** catalyze chemical reactions. See page 339 for a list of vitamins and their functions.

Minerals

Natural inorganic substances, e.g. phosphorus and calcium. They form a vital part of plant and animal tissue, e.g. in bones and teeth. Many are found in **enzymes*** and **vitamins**. They include **trace elements**, e.g. copper and iodine, present in tiny amounts.

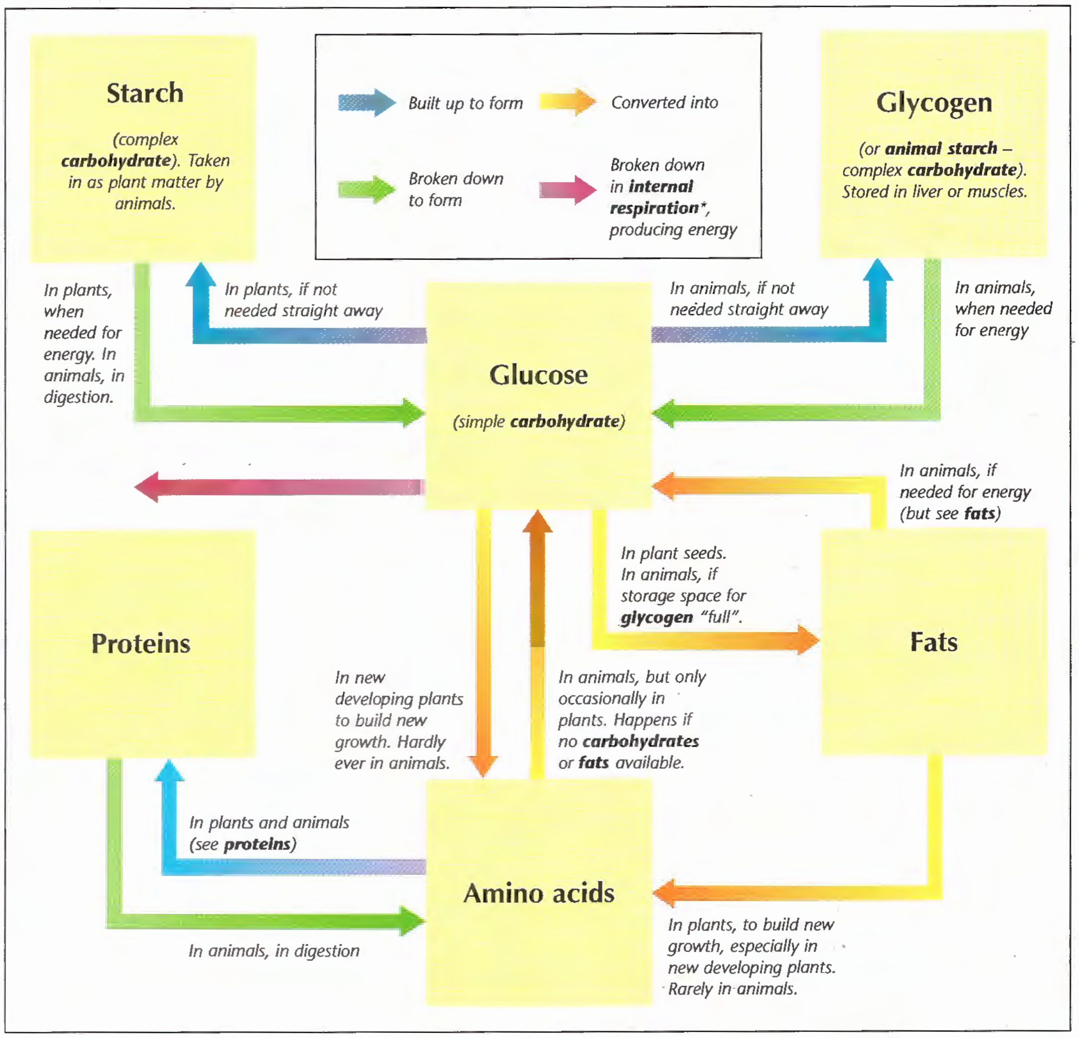
Starch
(complex carbohydrate). Taken in as plant matter by animals.
Built up to form
Converted into
Broken down to form
Broken down in internal respiration*, producing energy
Glycogen
(or animal starch – complex carbohydrate). Stored in liver or muscles.
In plants, when needed for energy. In animals, in digestion.
In plants, if not needed straight away
In animals, if not needed straight away
In animals, when needed for energy
Glucose
(simple carbohydrate)
In animals, if needed for energy (but see fats)
In plant seeds. In animals, if storage space for glycogen "full".
Proteins
In new developing plants to build new growth. Hardly ever in animals.
In animals, but only occasionally in plants. Happens if no carbohydrates or fats available.
Fats
In plants and animals (see proteins)
Amino acids
In animals, in digestion
In plants, to build new growth, especially in new developing plants. Rarely in animals.

---

* **Cell wall**, 238; **Co-enzymes**, 333 (**Enzymes**); **Digestive enzymes**, 338; **Internal respiration**, 334.

METABOLISM

Metabolism is a collective term for all the complex, closely-coordinated chemical reactions occurring inside an organism. These can be split into two opposing sets of reactions, called **catabolism** and **anabolism**. The rates of the reactions vary in response to variations in the organism's internal and external environments, and they play a major role in keeping internal conditions stable (see **homeostasis**, page 335).

Catabolism

A collective term for all the reactions which break down substances in the body (**decomposition reactions**). One example is digestion in animals, which breaks down complex substances into simpler ones (see chart, page 338). Another is the further breakdown of these simple substances in the cells (**internal respiration***). Catabolism always liberates energy (in digestion, most is lost as heat, but in internal respiration, it is used for the body's activities). This is despite the fact that, as with all chemical reactions, catabolism itself requires energy. The energy needed is taken from the much greater amount of energy produced during the reactions. The rest of this is released, hence the overall result is always an energy "profit".

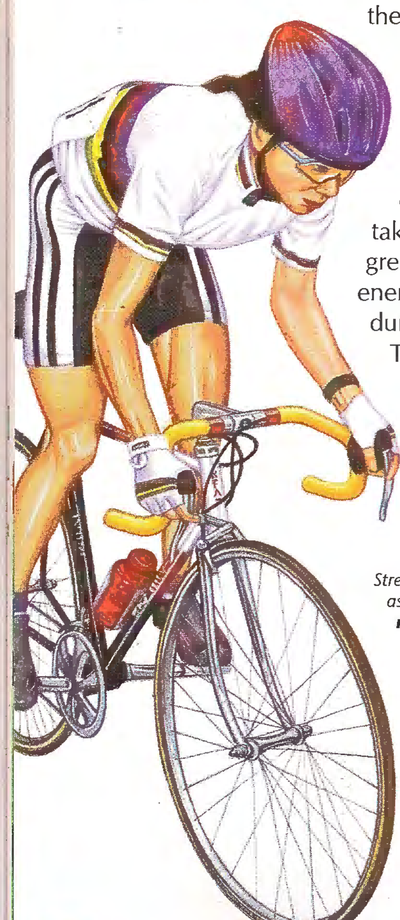

Strenuous exercise, such as cycling, can cause the ***metabolic rate*** *to increase to as much as 15 times the* ***basal metabolic rate****. The heart rate increases, and more oxygen is taken in. These changes enable food to be* ***catabolized*** *more quickly to produce the extra energy needed. One side effect is a rise in body temperature, which prompts the body to produce more sweat.*

Anabolism

A collective term for all the reactions which build up substances in the body (**synthesis reactions**). One example is the linking together of amino acids to form proteins (see page 330). Anabolism always needs energy to be taken in, since the small amount produced during the reactions is never enough (i.e. the overall result of anabolism is an energy "loss"). The extra energy is taken from the **catabolism** "profit".

Metabolic rate

The overall rate at which metabolic reactions occur in an individual. In human beings, it varies widely from person to person, and in the same individual under different conditions. It increases under stress, when the body temperature rises and during exercise, hence the true and accurate measurement of a person's metabolic rate is a measurement taken when the subject is resting, has a normal body temperature, and has not recently exercised. This is called the **basal metabolic rate** (**BMR**) and is expressed in **kilojoules** per square meter of body surface per hour (see measuring method and calculations, opposite).

People with high BMR can eat large amounts without putting on weight, because their **catabolism** of food matter (in the cells) happens so fast that not much fat is stored. This fast rate of reactions also often results in "excess" energy (i.e. energy not needed for **anabolism**), so they may appear to have a lot of "nervous energy". People with low BMR put on weight easily and often appear to have little energy.

The metabolic rate is influenced by a number of **hormones***, especially **STH**, **thyroxin**, **adrenalin** and **noradrenalin**. For more about these, see chart on pages 336-337.

---

*__Hormones__, 336; __Internal respiration__, 334.

Kilojoule

A unit of energy, spe
when referring to th
produced by the catab
when measuring a pe
rate (see **metabolic** ra
involved in measurin
known facts about th
produced by the brea
substances, with a me
consumption obtaine
conditions (see belo

To determine a person's basal
(BMR = kJ m^{-2} hr^{-1})

1. Facts known (determined
calorimeter*):*
a) If 1 liter of oxygen is used
are produced (i.e. enough heat
b) With fats, the result from
c) With proteins, the result from

2. First calculation:
Heat energy generated when food (in general) is broken do using 1 liter of oxygen = the average of the three figures above, i.e. 20.09kJ (provided subject measured has taken in equal amounts of all three foodstuffs).

3. Measure the oxygen used by the subject's body in a fixed time. Done using a spirometer (respirometer) – see picture, right.

A trace is drawn as the cylinder moves up and down.

The overall trend of the trace is
moves down as its volume of oxyg

Enzymes

Special proteins (catal
all living things and
reactions of life. They
they speed up reactio
being changed. Many
other substances, call
molecules are able to
of one reaction (catal
to the next reaction.

*__Catalyst__, 161.

Kilojoule

A unit of energy, specifically used in biology when referring to the amount of heat energy produced by the **catabolism** of food, and hence when measuring a person's **basal metabolic rate** (see **metabolic rate**). The calculations involved in measuring BMR combine certain known facts about the number of kilojoules produced by the breakdown of different substances, with a measurement of oxygen consumption obtained under controlled conditions (see below and right).

To determine a person's basal metabolic rate (BMR = kJ m^{-2} hr^{-1})

1. Facts known (determined using a piece of apparatus called a ***calorimeter****):*
a) If 1 liter of oxygen is used to break down carbohydrates, c.21.21kJ are produced (i.e. enough heat energy to heat c.5,050g water by 1°C).
b) With fats, the result from 1 liter oxygen is c.19.74kJ.
c) With proteins, the result from 1 liter oxygen is c.19.32kJ.

2. First calculation: Heat energy generated when food (in general) is broken down using 1 liter of oxygen = the average of the three figures above, i.e. 20.09kJ (provided subject measured has taken in equal amounts of all three foodstuffs).

3. Measure the oxygen used by the subject's body in a fixed time. Done using a spirometer (respirometer) – see picture, right.

Calculations (example):

1. (Measured) Subject used 1.5 liters oxygen in 5 minutes.

2. Hence he would use 18 liters oxygen in 1 hour (1.5 × 12).

3. (Known) 20.09kJ produced when food broken down by 1 liter oxygen.

4. Hence 361.62kJ produced if food broken down by 18 liters oxygen (20.09 × 18).

5. Hence 361.62kJ produced by breakdown of food in whole of subject's body in 1 hour (he would use 18 liters oxygen per hour – see point 2).

6. But BMR is measured in kJ per square meter of body surface per hour.

7. Standard chart used to work out body surface in square meters.

8. 361.62 divided by body surface (e.g. $2m^2$) = 180.81kJ m^{-2} hr^{-1} (BMR).

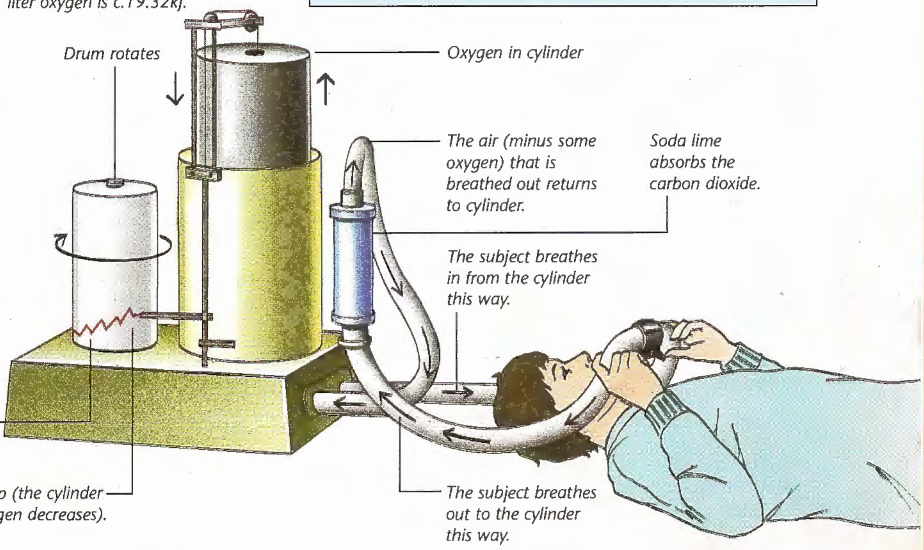
Drum rotates
Oxygen in cylinder
The air (minus some oxygen) that is breathed out returns to cylinder.
Soda lime absorbs the carbon dioxide.
The subject breathes in from the cylinder this way.
A trace is drawn as the cylinder moves up and down.
The overall trend of the trace is up (the cylinder moves down as its volume of oxygen decreases).
The subject breathes out to the cylinder this way.

Enzymes

Special proteins (**catalytic proteins**) found in all living things and vital to the chemical reactions of life. They act as **catalysts***, i.e. they speed up reactions without themselves being changed. Many enzymes are aided by other substances, called **co-enzymes**, whose molecules are able to "carry" the products of one reaction (catalyzed by an enzyme) on to the next reaction.

There are many different types of enzyme, e.g. **digestive enzymes**, which control the breakdown of complex food material into simple soluble substances (for more about these, see text and chart, pages 338-339), and **respiratory enzymes**, which control the further breakdown of these simple substances in the cells to liberate energy (i.e. **internal respiration** – see page 334).

***Catalyst**, 161.

ENERGY FOR LIFE AND HOMEOSTASIS

A living thing needs energy for its activities. This energy comes from a series of chemical reactions inside its cells, known as **internal respiration**, **tissue respiration** or **cellular respiration**. The cells contain various simple food substances, which are the results of digestive breakdown in animals (see pages 338-339) and **photosynthesis*** in plants. These substances all contain stored energy, which is released when internal respiration breaks them down. In almost all cases, glucose is the substance broken down (see **carbohydrates** and diagram, pages 330-331 and also equation, page 209). There are two kinds of respiration – **anaerobic** and **aerobic respiration**.

Anaerobic respiration

A type of **internal respiration** which does not need free oxygen (oxygen taken into the body). It takes place in the cells of all organisms and releases a small amount of energy. In most organisms, it consists of a chain of chemical reactions called **glycolysis**, which break down glucose into **pyruvic acid**. In normal circumstances this is then immediately followed by **aerobic respiration**, which breaks down this poisonous acid in the presence of oxygen. This breakdown releases the bulk of energy. In abnormal conditions, however, it may not be possible for the aerobic stage to follow immediately, in which case a further stage of anaerobic respiration occurs (see **oxygen debt**).

In some microscopic organisms, e.g. yeast and some bacteria, anaerobic respiration always runs through all its stages, providing enough energy for their needs without requiring oxygen.

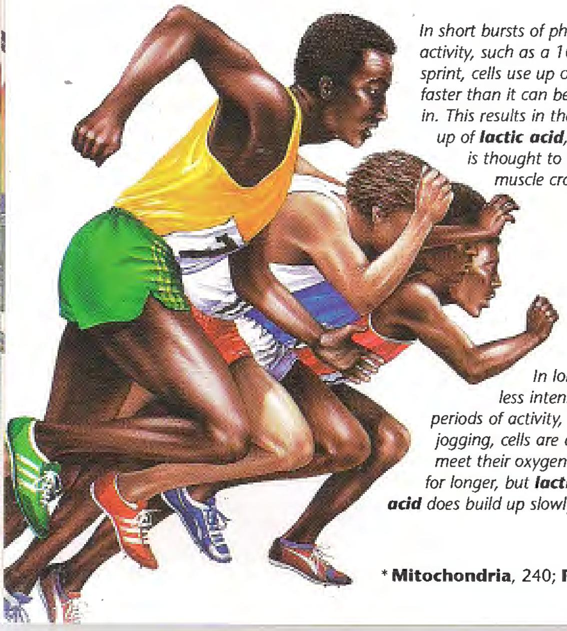

*In short bursts of physical activity, such as a 100m sprint, cells use up oxygen faster than it can be taken in. This results in the build up of **lactic acid**, which is thought to cause muscle cramps.*

*In longer, less intensive periods of activity, e.g. jogging, cells are able to meet their oxygen need for longer, but **lactic acid** does build up slowly.*

Aerobic respiration

A type of **internal respiration** which can only take place in the presence of free oxygen (oxygen taken into the body). It is the way in which most living things obtain the bulk of their energy and follows a stage of **anaerobic respiration**. Oxygen (brought by the blood) is taken into each cell and reacts in the **mitochondria*** with the **pyruvic acid** produced in anaerobic respiration. Carbon dioxide and water are the final products of the reactions, and chemical energy is released, which is then "stored" as **ATP**.

Aerobic respiration is an example of **oxidation** – the breakdown of a substance in the presence of oxygen.

Summary of aerobic stages of respiration reactions

$5O_2$	+	$2C_3H_4O_3$	$\rightarrow$	$6CO_2$	+	$4H_2O$
Oxygen		*Pyruvic acid*		*Carbon dioxide*		*Water*

Oxygen debt

A situation which occurs when extreme physical exercise is undertaken by an organism which shows **aerobic respiration**. Under these circumstances, the oxygen in the organism's cells is used up faster than it can be taken in. This means that there is not enough to break down the poisonous **pyruvic acid** produced in the first, **anaerobic**, stage of respiration. Instead, the acid undergoes further anaerobic reactions to convert it to **lactic acid** (much less harmful). This begins to build up and the organism is said to have acquired an oxygen debt. This is "paid off" later by taking in oxygen faster than usual to break down the lactic acid, by breathing heavily, for example.

---

* **Mitochondria**, 240; **Photosynthesis**, 254.

ADP (adenosine diphos
enosine triphosphat
o substances which co
uping called **adenosin**
and three **phosphate**
ese phosphate groups
sphorus, oxygen and
sphate group can com
stances (either by itse
sphate groups in a ch
When **aerobic respira**
emical energy released
ctions which result in a
ecules into ATP mole
a third phosphate grou
ergy taken in to effect
arded as being "stored
is a substance which
all cells. It is found in e
ntities in cells which r
muscle cells. When th
tions occur which con
. These reactions resu
ease of energy – the "st
power is supplied for

and ATP

ine + Ribose = Adenosine

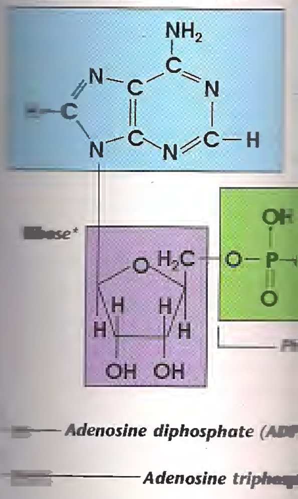

ersion of ATP into ADP

ATP	$\rightleftharpoons$	ADP +
enosine ***osphate***		***Adenosine diphosphate***

The reaction is reve

---

thalamus, 303; **Insulin**, 336;
bolic rate, 332; **Reversible re**

ADP (**adenosine diphosphate**) and **ATP** (**adenosine triphosphate**)

Two substances which consist of a chemical grouping called **adenosine**, combined with two and three **phosphate groups** respectively. These phosphate groups each consist of linked phosphorus, oxygen and hydrogen atoms. A phosphate group can combine with other substances (either by itself or linked with other phosphate groups in a chain).

When **aerobic respiration** occurs, the chemical energy released is involved in reactions which result in a conversion of ADP molecules into ATP molecules (by attachment of a third phosphate group in each case). The energy taken in to effect these reactions can be regarded as being "stored" in the form of ATP. This is a substance which can be easily stored in all cells. It is found in especially large quantities in cells which require a lot of energy, e.g. muscle cells. When the energy is needed, reactions occur which convert ATP back into ADP. These reactions result in an overall release of energy – the "stored" energy. In this way, power is supplied for the cell's activities.

ADP and ATP

Adenine + Ribose = Adenosine

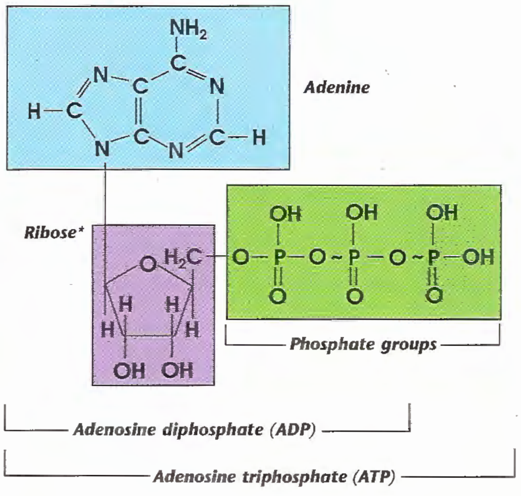

Conversion of ATP into ADP

ATP ⇌ **ADP** + **P** + **E**

Adenosine triphosphate | ***Adenosine diphosphate*** | ***Phosphate group*** | *Energy*

*The reaction is **reversible***.

Homeostasis

Crocodile gaping to lose heat

Homeostasis is the maintenance, by an organism, of a stable **internal environment**, i.e. a constant temperature, stable composition, level and pressure of body fluids, constant **metabolic rate***, etc. This is vital if the organism is to function properly.

Homeostasis requires the detection of any deviation from the norm (caused by new internal or external factors) and the means to correct such deviations, and is practiced most efficiently in birds and **mammals***, e.g. humans. Their detection of deviations is achieved by the **feedback** of information to controlling organs. The blood glucose level, for example, is constantly being detected by the pancreas (i.e. information is "fed back"). The correction of deviations is achieved by **negative feedback**, i.e. feedback which "tells" of deviations and results in a change of action. If the glucose level gets too high, for example, the pancreas reacts by producing more **insulin*** to reduce it (see also **antagonistic hormones**, page 336).

Most homeostatic actions are, like the insulin example, controlled by hormones, many of which are in turn controlled by the **hypothalamus*** in the brain. An example of the importance of the hypothalamus in homeostasis is the control of body temperature. All birds and mammals, e.g. humans, are **homiothermic** (warm-blooded), i.e. they can keep a constant temperature (about 37°C in humans) regardless of external conditions (the opposite is **poikilothermic**, or cold-blooded). A "thermostat" area of the hypothalamus, called the **preoptic area**, detects any changes in body temperature and sends impulses either to the **heat-losing center** or to the **heat-promoting center** (both also found in the hypothalamus). These areas then send out nervous impulses to cause various heat-losing or heat-promoting actions in the body.

*Penguins are **homiothermic**. They can generate enough heat to keep themselves and their eggs and young warm.*

---

* **Hypothalamus**, 303; **Insulin**, 336; **Mammals**, 341; **Metabolic rate**, 332; **Reversible reactions**, 162; **Ribose**, 324.

HORMONES

Hormones are special chemical "messengers" which control various activities inside an organism. These pages deal with the hormones produced by humans and their related groups. Plants also produce hormones (**phytohormones**), though these are not yet fully understood (see **abscission layer**, page 249, and **photoperiodism** and **growth hormones**, page 251).

Human hormones are secreted by **endocrine glands***. Some act only on specific body parts (**target cells** or **target organs**), others cause a more general response. The principal controller of hormone production is the **hypothalamus*** (part of the brain). It controls the secretions of many glands, mainly through its control of the **pituitary gland***, which itself controls many other glands. The hypothalamus "tells" the pituitary to produce its hormones by sending **regulating factors** to its **anterior lobe** and nervous impulses to its **posterior lobe**. Hormone secretion is vital to **homeostasis***.

Regulating factors

Special chemicals which control the production of a number of hormones, and hence many vital body functions. They are sent to the **anterior lobe** of the **pituitary gland*** by the **hypothalamus*** (part of the brain). There are two types – **releasing factors**, which make the gland secrete specific hormones, and **inhibiting factors**, which make it stop its secretion. For example, **FSHRF** (**FSH releasing factor**) and **LHRF** (**LH releasing factor**) cause the release of the hormones **FSH** and **LH** (see chart), and hence the onset of **puberty***. Many regulating factors are vital to **homeostasis***.

Antagonistic hormones

Hormones that produce opposite effects. **Glucagon** and **insulin** (see chart) are examples. When the blood glucose level drops too far, the pancreas produces glucagon to raise it again. A high glucose level causes the pancreas to produce insulin to lower the level (see also **homeostasis**, page 335).

Hormones	Where produc
ACTH (***adrenocorticotropic hormone***) *or* ***adrenocorticotropin***	***Pituitary gland*** *(page 297)* (ante
TSH (***thyroid-stimulating hormone***) *or* ***thyrotropin***	***Pituitary gland*** *(page 297)* (ante
STH (***somatotropic hormone***) *or* ***somatotropin*** *or* ***HGH*** (***human growth hormone***)	***Pituitary gland*** *(page 297)* (ante
FSH (***follicle-stimulating hormone***)	***Pituitary gland*** *(page 297)* (ante
LH (***luteinizing hormone***). *Also called* ***luteotropin*** *in women and* ***ICSH*** (***interstitial cell stimulating hormone***) *in men.*	***Pituitary gland*** *(page 297)* (ante
Lactogenic hormone *or* ***PR*** (***prolactin***)	***Pituitary gland*** *(page 297)* (ante
Oxytocin	***Hypothalamus*** *(page 303). Builds* ***gland*** (***posterior lobe***)
ADH (***anti-diuretic hormone***) *or* ***vasopressin***	***Hypothalamus*** *(page 303). Builds* ***gland*** (***posterior lobe***)
Thyroxin	***Thyroid gland*** *(page 297)*
TCT (***thyrocalcitonin***) *or* ***calcitonin***	***Thyroid gland*** *(page 297)*
PTH (***parathyroid hormone***) *or* ***parathyrin*** *or* ***parathormone***	***Parathyroid glands*** *(page 297)*
Adrenalin *or* ***adrenin*** *or* ***epinephrin*** ***Noradrenalin*** *or* ***norepinephrin***	***Adrenal glands*** *(page 297)* (***med*** *ngs. Secreted at times of excite*
Aldosterone	***Adrenal glands*** *(page 297)* (***corte***
Cortisone ***Hydrocortisone*** *or* ***cortisol***	***Adrenal glands*** *(page 297)* (***corte***
Estrogen *(female* ***sex hormone****)* ***Progesterone*** *(female* ***sex hormone****)*	*in* ***ovarian follicles*** *(page* ***corpus luteum*** *(page 318) in* ***ovar*** *organs, page 317). Also in* ***place*** *during pregnancy.*
Androgens *(male* ***sex hormones****),* *especially* ***testosterone***	*in* ***interstitial cells*** *in* ***testes*** *page 316).*
Gastrin	*stomach*
CCK (***cholecystokinin***)	*small intestine*
Secretin ***PZ*** (***pancreozymin***)	*small intestine*
Enterocrinin	*small intestine*
Insulin	*pancreas, when blood glucose level*
Glucagon	*pancreas, when blood glucose level*

* **Endocrine glands**, 297; **Homeostasis**, 335; **Hypothalamus**, 303; **Pituitary gland**, 297; **Puberty**, 318.

es		Where produced	Effects
hormone) or		**Pituitary gland** (page 297) (**anterior lobe**)	Stimulates production of hormones in **cortex** of **adrenal glands** (page 297).
hormone) or **thyrotropin**		**Pituitary gland** (page 297) (**anterior lobe**)	Stimulates production of **thyroxin** by **thyroid gland** (page 297).
e) or **somatotropin** **hormone**)		**Pituitary gland** (page 297) (**anterior lobe**)	Stimulates growth by increasing rate at which amino acids are built up to make proteins in cells.
hormone)		**Pituitary gland** (page 297) (**anterior lobe**)	In women, works with **LH** to stimulate development of **ova** in **ovarian follicles** (page 317) and secretion of **estrogen** by follicles in early stages of **menstrual cycle** (page 318). In men, causes formation of **sperm** (page 320).
Also called **ICSH** (**interstitial** e) in men.		**Pituitary gland** (page 297) (**anterior lobe**)	Stimulates **ovulation** (page 318), formation of **corpus luteum** (page 318) and its secretion of **estrogen** and **progesterone**. Works with **estrogen** and **progesterone** to stimulate thickening of lining of **uterus** (page 317). In men, causes production of **androgens**.
PR (**prolactin**)		**Pituitary gland** (page 297) (**anterior lobe**)	Works with **LH** to cause secretion of hormones by **corpus luteum** (page 318). Causes milk production after giving birth.
		Hypothalamus (page 303). Builds up in **pituitary gland** (**posterior lobe**)	Stimulates contraction of muscles of **uterus** (page 317) during labor and secretion of milk after giving birth.
e) or **vasopressin**		**Hypothalamus** (page 303). Builds up in **pituitary gland** (**posterior lobe**)	Increases amount of water re-absorbed into blood from **uriniferous tubules** (page 301) in kidneys.
		Thyroid gland (page 297)	Increases rate of food breakdown, hence increasing energy and raising body temperature. Works with **STH** in the young to control rate of growth and development. Contains iodine.
calcitonin		**Thyroid gland** (page 297)	Decreases level of calcium and phosphorus in blood by reducing their release from bones (where they are stored).
e) or **parathyrin**		**Parathyroid glands** (page 297)	Increases level of calcium in blood by increasing its release from bone (see above). Decreases phosphorus level.
epinephrin **ephrin**		**Adrenal glands** (page 297) (**medulla**). Also at nerve endings. Secreted at times of excitement or danger.	Stimulate liver to release more glucose into blood, to be broken down for energy. Stimulate increase in heart rate, faster breathing and blood vessel constriction.
		Adrenal glands (page 297) (**cortex**)	Increases amount of sodium and water in blood by causing re-absorption of more from **uriniferous tubules** (page 301) in kidneys.
tisol		**Adrenal glands** (page 297) (**cortex**)	Stimulate increase in rate of food breakdown for energy, and thus increase resistance to stress. Lessen inflammation.
hormone) **sex hormone**)		Mostly in **ovarian follicles** (page 317) and **corpus luteum** (page 318) in **ovaries** (female sex organs, page 317). Also in **placenta** (page 319) during pregnancy.	Estrogen activates development of **secondary sex characters** at **puberty** (page 318), e.g. breast growth. Both prepare **mammary** (milk) **glands** for milk production and work with **LH** to cause thickening of lining of **uterus** (page 317). Progesterone dominates toward end of **menstrual cycle** (page 318) and during pregnancy, when it maintains uterus lining and mammary gland readiness.
hormones),		Mostly in **interstitial cells** in **testes** (male sex organs, page 316).	Activate development and maintenance of **secondary sex characters** at **puberty** (page 318), e.g. beard growth.
		Cells in stomach	Stimulates production of **gastric juice** (page 338).
in)		Cells in small intestine	Stimulates opening of **sphincter of Oddi**, contraction of **gall bladder** and release of **bile** (all page 297) into **duodenum** (page 295).
		Cells in small intestine	Stimulate pancreas to produce **pancreatic juice** (page 338) and secrete it into **duodenum** (page 295).
		Cells in small intestine	Stimulates production of **intestinal juice** (page 338).
		Pancreas, when blood glucose level too high.	Stimulates liver to convert more glucose to glycogen for storage (page 331). Also speeds up transport of glucose to cells.
		Pancreas, when blood glucose level too low.	Stimulates faster conversion of glycogen to glucose in liver (page 331), and conversion of fats and proteins to glucose.

DIGESTIVE JUICES AND ENZYMES

All the **digestive juices*** of the human body (secreted into the intestines by **digestive glands***) contain **enzymes*** which control the breakdown of food into simple soluble substances. These are called **digestive enzymes** and can be divided into three groups. **Amylases** (or **diastases**) promote the breakdown of **carbohydrates*** (the final result being **monosaccharides** – see terms used, right). **Proteinases** (or **peptidases**) promote the breakdown of **proteins** into **amino acids*** by attacking the **peptide links** (see **proteins**, page 330). **Lipases** promote the breakdown of **fats** into **glycerol** and **fatty acids** (see **fats**, page 330). The chart below lists the different digestive juices of the body, together with their enzymes and the action of these enzymes.

Digestive juice: Saliva

Produced by**: **Salivary glands in mouth*

***Digestive enzyme**: **Salivary amylase** (or **ptyalin**)*

***Actions**: Starts breakdown of **carbohydrates* starch** and **glycogen** (**polysaccharides**). See page 331.*

***Products**: Some **dextrin** (shorter **polysaccharide**). See note 1.*

Digestive juice: Bile

Produced by**: Liver. Stored in **gall bladder, secreted into small intestine (see **CCK**, page 336).*

Constituents**: **Bile salts** and **bile acids

Actions**: Break up **fats (and intermediate compounds) into smaller particles, a process called **emulsification**.*

Digestive juice: Gastric juice

Produced by**: **Gastric glands in stomach lining. Secreted into stomach (see **gastrin**, page 336).*

***Digestive enzymes** (and one other constituent):*
1. ***Pepsin** (**proteinase**). See note 2.*
2. ***Rennin** (**proteinase**). Found only in the young.*
3. ***Hydrochloric acid***
4. ***Gastric lipase**. Found mainly in the young.*

Actions:
1. *Starts breakdown of **proteins*** (**polypeptides**).*
2. *Works (with calcium) to curdle milk, i.e. to act on its **protein** (**casein**). See note 3.*
3. *Activates **pepsin** (see note 2), curdles milk in adults (see note 3) and kills bacteria.*
4. *Starts breakdown of **fat*** molecules in milk.*

Products:
1. *Shorter **polypeptides***

2, 3. ***Curds**, i.e. milk solids*

4. *Intermediate compounds*

Digestive juice: Intestinal juice (or **succus entericus**)

Produced by**: **Intestinal glands in small intestine lining. Final secretion into small intestine (see **enterocrinin**, page 336).*

Digestive enzymes:
1. ***Maltase** (**amylase**)*
2. ***Sucrase** (or **invertase** or **saccharase**) (**amylase**)*
3. ***Lactase** (**amylase**)*
4. ***Enterokinase**. See note 2.*

Actions:
1. *Breaks down **maltose** (**disaccharide**).*
2. *Breaks down **sucrose** (**disaccharide**).*
3. *Breaks down **lactose** (**disaccharide**).*
4. *Completes breakdown of **proteins*** (**dipeptides**).*

Products:
1. ***Glucose** (or **dextrose**) (**monosaccharide**)*
2. ***Glucose** and **fructose** (**monosaccharides**)*
3. ***Glucose** and **galactose** (**monosaccharides**)*
4. ***Amino acids****

Digestive juice: Pancreatic juice

***Produced by**: Pancreas. Secreted into small intestine (see **secretin** and **PZ**, page 336).*

Digestive enzymes:
1. ***Trypsin** (**proteinase**). See note 2.*
2. ***Chymotrypsin** (**proteinase**). See note 2.*
3. ***Carboxypeptidase** (**proteinase**). See note 2.*
4. ***Pancreatic amylase** (or **amylopsin**)*
5. ***Pancreatic lipase***

Actions:

1, 2, 3. *Continue breakdown of **proteins*** (long and shorter **polypeptides**).*

4. *Continues breakdown of **carbohydrates***.*
5. *Breaks down **fat*** particles.*

Products:

1, 2, 3. ***Dipeptides** and some **amino acids***.*

4. ***Maltose** (**disaccharide**)*
5. ***Glycerol** and **fatty acids** (see **fats**, page 330).*

Notes

1. *Not much **dextrin** is produced at this stage, since food is not in the mouth long enough. Most carbohydrates pass through unchanged.*

2. ***Proteinases** are first secreted in inactive forms, to prevent them from digesting the digestive tube (made of **protein***, like most of the body). Once in the tube (beyond a protective layer of **mucous membrane***), these are converted into active forms. **Hydrochloric acid** changes **pepsinogen** (inactive) into **pepsin**, **enterokinase** changes **trypsinogen** into **trypsin**, and trypsin then changes **chymotrypsinogen** and **procarboxypeptidase** into **chymotrypsin** and **carboxypeptidase**.*

3. *The action of **rennin** and **hydrochloric acid** in curdling milk is vital, since liquid milk would pass through the system too fast to be digested.*

---

* **Amino acids**, 330 (**Proteins**); **Carbohydrates**, 330; **Digestive juices**, 296 (**Digestive glands**); **Enzymes**, 333; **Fats**, 330; **Gall bladder**, 297; **Gastric glands**, **Intestinal glands**, 296 (**Digestive glands**); **Mucous membrane**, 295; **Proteins**, 330; **Salivary glands**, 296.

Terms used

Polysaccharides

*The most complex **carbohy***

***monosaccharide** molecules.*

into the body are polysacch

polysaccharide in edible pl

in animal matter). For more

page 331.

Disaccharides

*Compounds of two **monosacch***

forming intermediate stages

***polysaccharides** or (in the*

taken into the body as such.

and sugar cane, lactose occurs

Vitamins and

Vitamin A (**retinol**)

***Sources**: Liver, kidneys, fish-liver*

*margarine, **pigment*** (**carote***

vegetables, especially tomatoes

to vitamin A in intestines).

***Uses**: Maintains general health*

aids growth, especially bones

dim light – involved in formation

***rhodopsin**), found in **rods** of re*

against infection.

Vitamin B complex

Group of at least 10 vitamins,

include:

***Thiamine** (or **aneurin**) (**B1**)*

***Riboflavin** (**B2**)*

Niacin** (or **nicotinic acid** or **nic

***Pantothenic acid** (**B5**)*

***Pyridoxine** (**B6**)*

Cyanocobalamin** (or **cobalam

***Folic acid** (**Bc** or **M**)*

Biotin** (sometimes called **vitamin

***Lecithin**.*

***Sources**: All found in yeast and*

wholewheat cereals and bread,

vegetables, e.g. beans (B12 not

products). B2 and B12 found espe

also found in eggs, nuts, fish, lean

B6, folic acid and biotin also made

***Uses**: Most needed for growth and*

tissues, e.g. muscles (B1, B6), nerv

B2, B3, B5, B6, B12) and hair (B2,

continuous function of body organs

except folic acid, biotin and lecithin

acting in breakdown of foods for

respiration). Many (especially B2*

*acting build-up of substances (**prot***

*regulatory or defence purposes. **B12***

*formation of blood cells, B5 and **B6***

*the chemicals (**neurotransmitte***

Amino acids, 330 (**Proteins**); **Car**

Epithelium, 310 (**Epidermis**); **Int**

Pigments, 255; **Proteins**, 330; **Pr**

Terms used

Polysaccharides
*The most complex **carbohydrates***. Each is a chain of **monosaccharide** molecules. Most carbohydrates taken into the body are polysaccharides, e.g. **starch** (the main polysaccharide in edible plants) and **glycogen** (the main one in animal matter). For more about starch and glycogen, see page 331.*

Disaccharides
*Compounds of two **monosaccharide** molecules, either forming intermediate stages in the breakdown of **polysaccharides** or (in the case of **sucrose** and **lactose**) taken into the body as such. (Sucrose is found in sugar beet and sugar cane, lactose occurs in milk.)*

Monosaccharides
*The simplest **carbohydrates***. Almost all result from **polysaccharide** breakdown, though **fructose** is taken into the body as such (e.g. in fruit juices), as well as resulting from **sucrose** breakdown. **Glucose** is the final result of all action on carbohydrates (fructose and **galactose** are converted to glucose in the liver).*

Polypeptides
*The complex form taken by all **proteins** ~~entering the~~ body. Each is a chain of hundreds (or thousands) of **amino acid*** molecules (see **proteins**, pages 330).*

Dipeptides
*Chains of two **amino acid** molecules, forming intermediate stages in the breakdown of **polypeptides**.*

Vitamins and their uses

Vitamin A (retinol)

Sources**: Liver, kidneys, fish-liver oils, eggs, dairy products, margarine, **pigment (**carotene**) in green and yellow fruit and vegetables, especially tomatoes and carrots (carotene converted to vitamin A in intestines).*

Uses**: Maintains general health of **epithelial cells (lining cells), aids growth, especially bones and teeth. Essential for vision in dim light – involved in formation of light-sensitive **pigment*** (**rhodopsin**), found in **rods** of **retina***. Aids in resistance against infection.*

Vitamin B complex

Group of at leost 10 vitamins, usually occurring together. Include:
***Thiamine** (or **aneurin**) (**B1**)*
***Riboflavin** (**B2**)*
***Niacin** (or **nicotinic acid** or **nicotinamide**) (**B3**)*
***Pantothenic acid** (**B5**)*
***Pyridoxine** (**B6**)*
***Cyanocobalamin** (or **cobalamin**) (**B12**)*
***Folic acid** (**Bc** or **M**)*
***Biotin** (sometimes called **vitamin H**)*
***Lecithin**.*

***Sources**: All found in yeast and liver. All except B12 found in wholewheat cereals and bread, wheatgerm and green vegetables, e.g. beans (B12 not found in any vegetable products). B2 and B12 found especially in dairy products. Most also found in eggs, nuts, fish, lean meat, kidneys and potatoes. B6, folic acid and biotin also made by bacteria in intestines.*

Uses**: Most needed for growth and maintenance of healthy tissues, e.g. muscles (B1, B6), nerves (B1, B3, B6, B12), skin (B2, B3, B5, B6, B12) and hair (B2, B5). Several also aid continuous function of body organs (B5, B6, lecithin). All except folic acid, biotin and lecithin are essential **co-enzymes, aiding in breakdown of foods for energy (**internal respiration***). Many (especially B2, B6, B12) also co-enzymes aiding build-up of substances (**proteins***) for growth and regulatory or defence purposes. B12 and folic acid vital to formation of blood cells, B5 and B6 vital to manufocture of nerve chemicals (**neurotransmitters***).*

Vitamin C (ascorbic acid)

***Sources**: Green vegetables, potatoes, tomatoes, citrus fruit, e.g. oranges, grapefruit, lemons.*

Uses**: Needed for growth and maintenance of healthy tissues, especially skin, blood vessels, bones, gums, teeth. Essential **co-enzyme in many metabolic reactions, especially **protein*** breakdown and build-up of **amino acids*** into new proteins (especially **collagen** – see **connective tissue**, page 280). Aids in resistance against infection and healing of wounds.*

Vitamin D (calciferol)

***Sources**: Liver, fish-liver oils, oily fish, dairy products, egg yolk, margarine, special substance (provitamin D3) in skin cells (converted to vitamin D when exposed to sunlight).*

Uses**: Essential for absorption of calcium and phosphorus, and their deposition in bones and teeth. May work with **PTH (**hormone**).*

Vitamin E (tocopherol)

***Sources**: Meat, egg yolk, leafy vegetables, nuts, dairy products, margarine, cereals, wholemeal bread, wheatgerm, seeds, seed and vegetable oils.*

***Uses**: Not yet fully understood. Protects membranes from some molecules which could bind and cause cancer.*

Vitamin K (phylloquinone or menaquinone)

***Sources**: Liver, fruit, nuts, cereols, tomatoes, green vegetables, especially cabbage, cauliflower, spinach. Also made by bacteria in intestines.*

Uses**: Essentiol for formation of **prothrombin in liver (needed to cause clotting of blood).*

* **Amino acids**, 330 (**Proteins**); **Carbohydrates**, 330; **Co-enzymes**, 333 (**Enzymes**); **DNA**, 324 (**Nucleic acids**); **Epithelium**, 310 (**Epidermis**); **Internal respiration**, 334; **Neurotransmitters**, 305 (**Synapse**); **Pigments**, 255; **Proteins**, 330; **Prothrombin**, 287 (**Clotting**); **PTH**, 336; **Retina**, 313.

THE CLASSIFICATION OF LIVING THINGS

Callicore cyllene

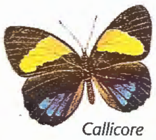
Callicore mengeli

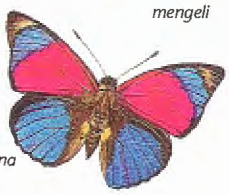
Agrias claudina

These butterflies are so rare, they do not have common names, only ***binomial*** *names, see below.*

Classification, also called **taxonomy**, is the grouping together of living things according to the characteristics they share. The main, formal type of classification (**classical taxonomy**) bases its groups primarily on structural characteristics. The resulting classification charts first list the largest groups – **Kingdoms** – and then list the smaller and smaller sub-divisions within these groups.

The first groups after the Kingdoms are called **phyla** (sing. **phylum**) in the case of animals, and **divisions** in the case of plants. After these come **classes**, **orders**, **families**, **genera** (sing. **genus**) and finally **species** – the smallest groupings. A species is defined as a group of interbreeding organisms, reproductively isolated from other groups. If this is impossible to establish, then a species is recognized on a **morphological** basis, i.e. by its external appearance.

Some divisions or phyla, especially those with only a few members, may not have all these groups, so the next group after a phylum may be an order, family, genus or even a species. There are also further "mid-way" groups in some cases, such as **sub-kingdoms**, **sub-phyla** or **sub-classes**.

There are areas which are still under dispute in both plant and animal classification. Most scientists recognize five main Kingdoms (see diagram, page 341), but some prefer to group living things into four Kingdoms: **animals** (including **Protista**), **plants** (including **fungi** and **algae**), **Monera** and **viruses**. The diagram below shows the divisions of the Plant Kingdom.

Nomenclature

The naming of organisms. Names of **species** are given in Latin so that all biologists world-wide can follow the same system. This is necessary as species are often known by various common names in different parts of the world. For example, one type of herring, *Alosa pseudoharengus*, has six different names throughout its geographical range.

Every organism has a two word name. This is called the **binomial system**. The first word is a **generic name** (from the **genus**) and the second identifies the species within that genus. The Latin names are governed by The International Commission for Zoological Nomenclature, at the Natural History Museum in London. Most names make direct reference to specific characteristics of the species, such as size, shape or habitat. The giant anteater, for example, is called *Myrmecophaga tridactyla* (*myrmeco* = ant, *phag* = eat, *tri* = three, and *dactyl* = fingers). This describes the food it eats and the three large digging claws at the end of each foreleg.

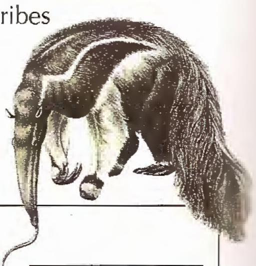
Giant anteater

Plant divisions

The Plant Kingdom is separated into the following major ***divisions****:*

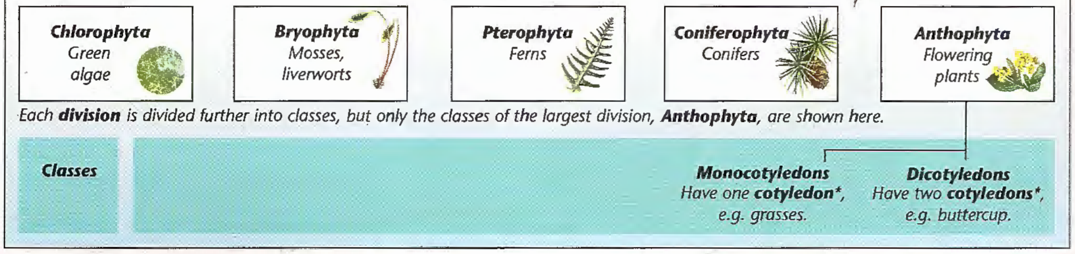
Chlorophyta
Green algae
Bryophyta
Mosses, liverworts
Pterophyta
Ferns
Coniferophyta
Conifers
Anthophyta
Flowering plants
Each ***division*** *is divided further into classes, but only the classes of the largest division,* ***Anthophyta****, are shown here.*
Classes
Monocotyledons
Have one **cotyledon***, e.g. grasses.
Dicotyledons
Have two **cotyledons***, e.g. buttercup.

---

* **Cotyledon**, 261.

Tracing Homo sapiens th

There are over one and a half

Kingdoms †

Phyla

Annelids
Worm-like, round segmented bodies

Sub-phyla

Classes

Orders

Bats
Mouse-like, have wings, nocturnal

Families

Genera

Species

† **Viruses** *are not included in cla when living in other cells.*

Other classifica

Classical taxonomy is o used to classify living thi method is **cladistic clas**

* **Cell wall**, 238; **Cellulose**, 331 **Nucleic acids**, 324; **Nucleus**,

Callicore mengeli

butterflies are so rare, not have common only ***binomial*** *see below.*

es of
at all biologists
system. This
known by
erent parts of
type of herring,
different
ical range.
ord name.
stem. The first
genus) and
es within that
erned by
for Zoological
istory Museum
direct reference
e species, such
giant anteater, for
haga tridactyla
= three, and

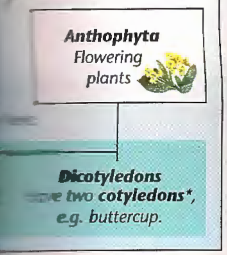

Tracing Homo sapiens through classical taxonomy

There are over one and a half million ***species*** *of animals. This is how humans fit into their own niche among them.*

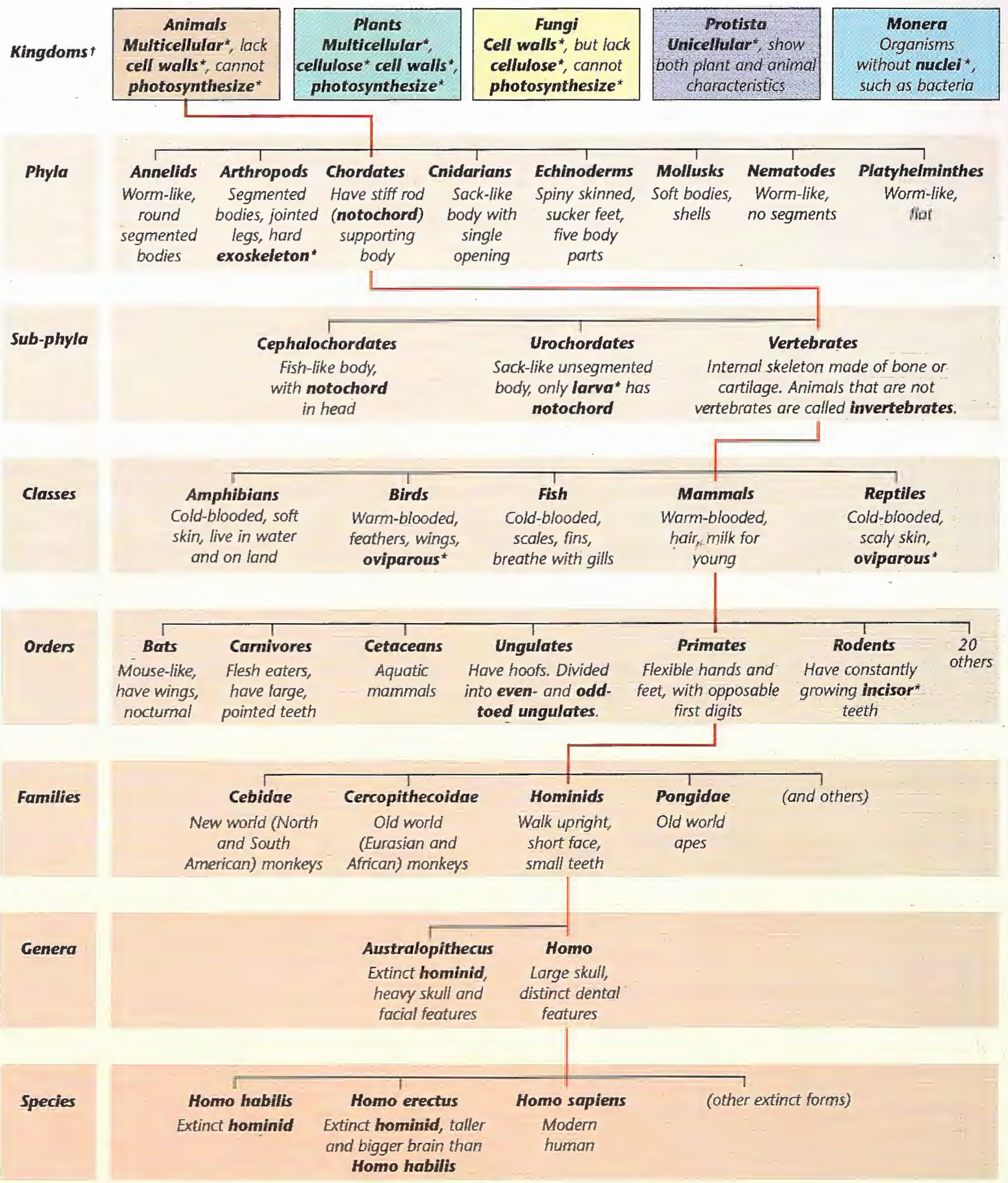

Kingdoms†

- **Animals** *Multicellular*, lack* ***cell walls****, cannot* ***photosynthesize****
- **Plants** *Multicellular*,* ***cellulose**** ***cell walls****,* ***photosynthesize****
- **Fungi** ***Cell walls****, but lack* ***cellulose****, cannot* ***photosynthesize****
- **Protista** ***Unicellular****, show both plant and animal characteristics*
- **Monera** *Organisms without* ***nuclei****, such as bacteria*

Phyla

- **Annelids** *Worm-like, round segmented bodies*
- **Arthropods** *Segmented bodies, jointed legs, hard* ***exoskeleton****
- **Chordates** *Have stiff rod* ***(notochord)*** *supporting body*
- **Cnidarians** *Sack-like body with single opening*
- **Echinoderms** *Spiny skinned, sucker feet, five body parts*
- **Mollusks** *Soft bodies, shells*
- **Nematodes** *Worm-like, no segments*
- **Platyhelminthes** *Worm-like, flat*

Sub-phyla

- **Cephalochordates** *Fish-like body, with* ***notochord*** *in head*
- **Urochordates** *Sack-like unsegmented body, only* ***larva**** *has* ***notochord***
- **Vertebrates** *Internal skeleton made of bone or cartilage. Animals that are not vertebrates are called* ***invertebrates****.*

Classes

- **Amphibians** *Cold-blooded, soft skin, live in water and on land*
- **Birds** *Warm-blooded, feathers, wings,* ***oviparous****
- **Fish** *Cold-blooded, scales, fins, breathe with gills*
- **Mammals** *Warm-blooded, hair, milk for young*
- **Reptiles** *Cold-blooded, scaly skin,* ***oviparous****

Orders

- **Bats** *Mouse-like, have wings, nocturnal*
- **Carnivores** *Flesh eaters, have large, pointed teeth*
- **Cetaceans** *Aquatic mammals*
- **Ungulates** *Have hoofs. Divided into* ***even-*** *and* ***odd-toed ungulates****.*
- **Primates** *Flexible hands and feet, with opposable first digits*
- **Rodents** *Have constantly growing* ***incisor**** *teeth*
- 20 others

Families

- **Cebidae** *New world (North and South American) monkeys*
- **Cercopithecoidae** *Old world (Eurasian and African) monkeys*
- **Hominids** *Walk upright, short face, small teeth*
- **Pongidae** *Old world apes*
- *(and others)*

Genera

- **Australopithecus** *Extinct* ***hominid****, heavy skull and facial features*
- **Homo** *Large skull, distinct dental features*

Species

- **Homo habilis** *Extinct* ***hominid***
- **Homo erectus** *Extinct* ***hominid****, taller and bigger brain than* ***Homo habilis***
- **Homo sapiens** *Modern human*
- *(other extinct forms)*

† ***Viruses*** *are not included in classical taxonomy. They consist of* ***nucleic acid**** *within a protein coat and can only "live" and reproduce when living in other cells. Therefore they are not classified as organisms, although they show many of the characteristics of living things.*

Other classification systems

Classical taxonomy is one of several systems used to classify living things. Another significant method is **cladistic classification**.

Cladistic classification

As with classical taxonomy, this system places groups into larger groups (called **clades**) based on shared characteristics. But clades also include the ancestral forms and all their descendants.

---

* **Cell wall**, 238; **Cellulose**, 331 (**Roughage**); **Exoskeleton**, 266; **Incisors**, 285; **Larva**, 277; **Multicellular**, 238; **Nucleic acids**, 324; **Nucleus**, 238; **Oviparous**, 274; **Photosynthesis**, 254; **Unicellular**, 238.

INFORMAL GROUP TERMS

Listed here are the main terms used to group living things together according to their general life styles (i.e. their ecological similarities – see also page 237). These are informal terms, as opposed to the formal terms of the classification charts (pages 340-341).

Plants

Xerophytes
Plants which can survive long periods of time without water, e.g. cacti.

Hydrophytes
Plants which grow in water or very wet soil, e.g. reeds.

Mesophytes
Plants which grow under average conditions of moisture.

Halophytes
Plants which can withstand very salty conditions, e.g. sea pinks.

Lithophytes
Plants which grow on rock, e.g. some mosses.

Epiphytes
Plants which grow on other plants, but only to use them for support, not to feed off them, e.g. some mosses.

Saprophytes
Plants which live on decaying plants or animals, feed off them, but are not the agents of their death, e.g. some fungi.

Animals

Predators
Animals which kill and eat other animals (their ***prey****), e.g. lions. Bird predators, e.g. hawks, are called* ***raptors****.*

Detritus feeders
Animals which feed on debris from decayed plant or animal matter, e.g. worms.

Scavengers
Large ***detritus feeders****, e.g. hyenas, which feed only on dead flesh (animal matter).*

Territorial
Holding and defending a ***territory*** *(an area of land or water) either singly or in groups, e.g. many fish, birds and mammals. This is usually linked with attracting a mate and breeding.*

Abyssal
Living at great depths in the sea, e.g. oarfish and gulper eels.

Demersal
Living at the bottom of a lake or the shallow sea, e.g. angler fish and prawns.

Sedentary
Staying mostly in one place (but not permanently attached), e.g. sea anemone.

Nocturnal
Active at night and sleeping during the day, e.g. owls and bats.

Plants and animals

Insectivores
Specialized organisms which eat mostly insects, e.g. pitcher plants (which trap and digest insects) and hedgehogs.

Parasites
Plants or animals which live in or on other living plants or animals (the ***hosts****), and feed off them, e.g. mistletoe and fleas. Not all are harmful to the host.*

Mutualists
*A pair of living things which associate closely with each other and derive mutual benefit from such close existence (****mutualism****).* ***Lichens****, normally found on bare rock, are an example. Each is in fact two plants (a fungus and an alga). The alga produces food (by* ***photosynthesis*****) for the fungus (which otherwise could not live on bare rock). The fungus uses its fine threads to hold the moisture the alga needs.*

Commensals
A pair of living things which associate closely with each other, one deriving benefit without affecting the other. One type of worm, for instance, is very often found in the same shell as a hermit crab. One of the most common examples of ***commensalism*** *is the existence of house mice wherever there are humans.*

Social or **colonial**
*Living together in groups. The two terms are synonymous in the case of plants, and refer to those which grow in clusters. In the case of animals, there is a difference of numbers between the terms. Lions, for example, are social, but their groups (****prides****) are not large enough to be called colonies. With true colonial animals, there is also a great difference in the level of interdependence between colony members. In a gannet colony, for example, this is relatively low (they only live close together because there is safety in numbers). In an ant colony, by contrast, different groups (****castes****) have very different jobs, e.g. gathering food or guarding the colony, so each member relies heavily on others. The highest level of colonial interdependence is shown by the tiny, physically inseparable, single-celled organisms which form one living mass, e.g. a sponge.*

Sessile
In the case of animals, this term refers to those which are not free to move around, i.e. they are permanently fixed to the ground or other solid object, e.g. barnacles. With plants, it describes those without stalks, e.g. stemless thistles.

Pelagic
Living in the main body of a lake or the sea, as opposed to at the bottom or at great depths. Pelagic creatures range from tiny ***plankton*** *through medium-sized fish and sharks to very large whales. The medium-sized and large ones are all animals, and are called* ***nekton*** *(from the Greek for "swimming thing"), as they swim.*

Plankton
Aquatic animals and plants, vast numbers of which drift in lakes and seas, normally near the surface (plant plankton is ***phytoplankton****, and animal plankton is* ***zooplankton****). Plankton is the food of many fish and whales and is thus vital to the ecological balance (****food chains*****) of the sea. Most are small.*

Littoral
Living at the bottom of a lake or the sea near the shore, e.g. crabs and seaweed.

Benthos
All ***abyssal****,* ***demersal*** *and* ***littoral*** *plants and animals, i.e. all those which live in, on or near the bottom of lakes or seas.*

* **Food chains**, 234 (**Food web**); **Photosynthesis**, 254.

DICTIONARY OF SCIENCE

Glossary and index

GLOSSARY

Abrasive
A material which wears away the surface of another material.

Adhesive
A substance which sticks to one or more other substances (see adhesion, page 23).

Alloy
*A mixture of two or more metals, or a metal and a non-metal. It has its own properties (which are metallic), independent of those of its constituents. For example, **brass*** is an alloy of copper and zinc, and **steel*** is an alloy of iron and carbon (different mixes give the steel different properties).*

Amalgam
*An **alloy** of mercury with other metals. It is usually soft and may even be liquid.*

Antacid
*A substance which counteracts excess stomach acidity by **neutralizing*** the acid. Examples are aluminum hydroxide and magnesium hydroxide.*

Bleach
*A substance used to remove color from a material or solution. Most strong **oxidizing*** and **reducing agents*** are good bleaches. The most common household bleach is a solution of sodium hypochlorite (also a highly effective **germicide**). The equation below shows the products of the reaction between sodium hypochlorite and a colored material.*

$$\underset{\text{Sodium hypochlorite}}{NaOCl} + \text{Colored material} \rightarrow \underset{\text{Sodium chloride}}{NaCl} + \text{Decolorized } \textbf{oxidized*} \text{ material}$$

Calibration
The "setting up" of a measuring instrument so that it gives the correct reading. The instrument is normally adjusted during manufacture so that it reads the correct value when it is measuring a known standard quantity, e.g. a balance would be adjusted to read exactly 1kg when a standard 1kg mass was on it.

Calorimetry
*The measurement of heat change during a chemical reaction or event involving heat transfer. For example, measuring the temperature rise of a known mass of a substance when it is heated electrically is used to find **specific heat capacity***, and the temperature rise of a mass of water can be used to calculate the energy produced by a fuel when it is burned (see **bomb calorimeter** diagram, page 147).*

Coefficient
*A **constant** for a substance, used to calculate quantities related to the substance by multiplying it by other quantities. For example, the force pushing two materials together multiplied by the **coefficient of friction*** for the surfaces gives the **frictional force***.*

Constant
*A numerical quantity that does not vary. For example, in the equation $E = mc^2$ (see also page 84), the quantity c (the speed of light in a vacuum) is the constant. E and m are **variables** because they can change.*

Coolant
*A fluid used for cooling in industry or in the home (see also **refrigerant**). The fluid usually extracts heat from one source and transfers it to another. In a **nuclear power station*** the coolant transfers heat from the nuclear reaction to the steam generator, where the heat is used to produce steam. This turns turbines and generates electricity.*

Cosine (of an angle)
The ratio of the length of the side adjacent to the angle to the length of the hypotenuse (the longest side) in a right-angled triangle. It depends on the angle.

Dehydrating agent
*A substance used to absorb moisture from another substance, removing water molecules if present, but also, importantly, hydrogen and oxygen atoms from the molecules of the substance. This leaves a different substance plus water (see also **drying agent**). Concentrated sulfuric acid is an example:*

$$\underset{\text{Glucose}}{C_6H_{12}O_6(s)} \xrightarrow{\text{Concentrated sulfuric acid}} \underset{\text{Carban}}{6C(s)} + \underset{\text{Water}}{6H_2O(l)}$$

Concentrated sulfuric acid can also be used as a drying agent if it does not react with the substance added to it. For example, it is used to dry samples of chlorine gas, i.e. remove surrounding molecules of water vapor (see page 216).

Drying agent
*A substance used to absorb moisture from another substance, but which only removes water molecules from in and around the substance, not separate hydrogen and oxygen atoms from its molecules. The substance itself is not changed (see also **desiccator**, page 221, and **dehydrating agent**). **Phosphorus pentoxide** (P_2O_5) is an example:*

$$\underset{\text{Hydrated* copper(II) sulfate}}{CuSO_4.5H_2O(s)} \xrightarrow{\text{Phosphorus pentoxide}} \underset{\text{Anhydrous* copper(II) sulfate}}{CuSO_4(s)} + \underset{\text{Water}}{5H_2O(l)}$$

Ductile
*Describes a substance which can be stretched. It is normally used of metals which can be drawn out into thin wire, e.g. copper. Different substances show varying degrees of **ductility** (see page 165). See also **yield point**, page 23.*

Fumigation
The killing of pests such as insects by poisonous gas, e.g. sulfur dioxide, or smoke.

Fungicide
A substance used to destroy harmful fungi, e.g. molds and mildews growing on crops.

Germicide
A substance used to destroy bacteria, especially those carrying disease (germs).

Graduations
Marks used for measurement, e.g.
*or a **gas syringe***.*

Inert
Describes an unreactive substance,
take part in chemical reactions.
inert) gases.*

Inversely proportional
When applied to two quantities,
is doubled, the other is halved.

Latex
A milky fluid produced by plants,
rubber tree, from which raw natural
also forms the basis of some
synthetic polymers.*

Malleable
Describes a substance which can
is normally used of substances
thin sheets, in particular many
Different substances show varying
page 165).

Mean
A synonym for average, i.e. the
by the number of values in the

Medium (pl. media)
Any substance through which a
glass is a medium when light travels

Meniscus
The concave or convex surface of
caused by the relative attraction of
those of the container (see also
*and **parallax error**, page 102).*

Ore
naturally-occurring mineral from
metal) is extracted, e.g. bauxite,

Organic solvent
organic liquid in which substances

Photocell or photoelectric cell
device used for the detection and

Proportional
applied to two quantities, this
relationship such that, for example,

Rate
amount by which one quantity
*e.g. **acceleration*** is the rate of*
that the second quantity is not
of Y against X is plotted, the
at a point is the gradient at that

Raw material
obtained from natural sources
coke and limestone are the raw
picture, page 174).

---

* **Anhydrous**, 154 (**Anhydrate**); **Brass**, 112, 175 (**Zinc**); **Coefficient of friction**, **Frictional force**, 7; **Hydrated**, 154 (**Hydrate**); **Neutralization**, 151; **Nuclear power station**, 94; **Oxidation**, **Oxidizing agent**, **Reducing agent**, 148; **Specific heat capacity**, 31; **Steel**, 174.

Acceleration, 11; **Chemical**
Evaporation, 121; **Gas syringe**,
SI units, 96; **Sublimation**, 121;

Graduations
Marks used for measurement, e.g. those on a **measuring cylinder*** or a **gas syringe***.

Inert
Describes an unreactive substance, i.e. one which does not easily take part in chemical reactions. Examples are the **noble** (or **inert**) **gases***.

Inversely proportional
When applied to two quantities, this means that, for example, if one is doubled, the other is halved.

Latex
A milky fluid produced by plants, particularly that produced by the rubber tree, from which raw natural rubber is extracted (and which also forms the basis of some **adhesives**). Also certain similar **synthetic polymers***.

Malleable
Describes a substance which can be molded into different shapes. It is normally used of substances which can be hammered out into thin sheets, in particular many metals and **alloys** of metals. Different substances show varying degrees of **malleability** (see page 165).

Mean
A synonym for average, i.e. the sum of a collection of values divided by the number of values in the collection.

Medium (pl. **media**)
Any substance through which a physical effect is transmitted, e.g. glass is a medium when light travels through it.

Meniscus
The concave or convex surface of a liquid, e.g. water or mercury. It is caused by the relative attraction of the molecules to each other and to those of the container (see also **adhesion** and **cohesion**, page 23 and **parallax error**, page 102).

Ore
A naturally-occurring mineral from which an element (usually a metal) is extracted, e.g. bauxite, which yields aluminum.

Organic solvent
An organic liquid in which substances will dissolve.

Photocell or **photoelectric cell**
A device used for the detection and measurement of light.

Proportional
When applied to two quantities, this means that they have a relationship such that, for example, if one is doubled, so is the other.

Rate
The amount by which one quantity changes with respect to another, e.g. **acceleration*** is the rate of change of **velocity*** with time. Note that the second quantity is not necessarily time in all cases. If a graph of Y against X is plotted, the rate of change of Y with respect to X at a point is the gradient at that point.

Raw material
A material obtained from natural sources for use in industry, e.g. iron **ore**, coke and limestone are the raw materials used to produce iron (see picture, page 174).

Reciprocal
The value obtained from a number when one is divided by it, i.e. the reciprocal number of x is $^1/x$. For example, the reciprocal of 10 is 0.1.

Refrigerant
A type of **coolant** used in refrigerators. It must be a liquid which **evaporates*** at low temperatures. The substances commonly used nowadays are the **chlorofluorocarbons***, although ammonia was widely used in the past.

Resins
Substances used as **adhesives**. They are often insoluble in water. **Natural resins** are organic compounds secreted by certain plants and insects. **Synthetic resins** are plastic materials produced by **polymerization***.

Sine (of an angle)
The ratio of the length of the side opposite to the angle to the length of the hypotenuse (the longest side) in a right-angled triangle. It depends on the angle.

Spectrum (pl. **spectra**)
A particular distribution of wavelengths and frequencies, e.g. the wavelengths in the **visible light spectrum*** range from 4×10^{-7}m to 7.5×10^{-7}m.

Superheated steam
Steam above a temperature of 100°C. It is obtained by heating water under pressure.

System
A set of connected parts which have an effect on each other and form a whole unit, e.g. a **digestive system*** or the substances involved in a reaction at **chemical equilibrium***.

Tangent (of an angle)
The ratio of the length of the side opposite to the angle to the length of the side adjacent to it in a right-angled triangle. It depends on the angle.

Tarnish
To lose or partially lose shine due to the formation of a dull surface layer, e.g. silver sulfide on silver or lithium oxide on lithium. Tarnishing is a type of **corrosion***.

Variable
A numerical quantity which can take any value. For example, in the equation $E = mc^2$ (see also page 84), E and m are variables since they can take any value (although the value of E depends on the value of m). The quantity c (the speed of light in a vacuum) is a **constant**.

Volatile
Describes a liquid that **evaporates*** easily, e.g. gasoline, or a solid that **sublimes*** easily, e.g. iodine.

Volume
A measurement of the space occupied by a body. See page 101 for calculations of volume. The **SI unit*** of volume is the cubic meter (m^3).

Vulcanization
The process of heating raw natural rubber (extracted from **latex**) with sulfur. Vulcanized rubber is harder, tougher and less temperature-sensitive than raw rubber (the more sulfur used, the greater the difference). This is because the sulfur atoms form cross-links between the chains of rubber molecules (see picture, page 201).

* **Acceleration**, 11; **Chemical equilibrium**, 163; **Chlorofluorocarbons**, 195; **Corrosion**, 209; **Digestive system**, 294; **Evaporation**, 121; **Gas syringe**, 224; **Measuring cylinder**, 224; **Noble gases**, 189; **Polymerization**, 200; **SI units**, 96; **Sublimation**, 121; **Synthetic polymers**, 201; **Velocity**, 10; **Visible light spectrum**, 54.

INDEX

The page numbers listed in the index are of three different types. Those printed in bold type (e.g. **92**) indicate in each case where the main definition(s) of a word (or words) can be found. Those in lighter type (e.g. 92) refer to supplementary entries. Page numbers printed in italics (e.g. *92*) indicate pages where a word (or words) can be found as a small print label to a picture. If a page number is followed by a word in brackets, it means that the indexed word can be found inside the text of the definition indicated. If it is followed by (**I**), the indexed word can be found in the introductory text on the page given. Bracketed singulars, plurals, symbols and formulas are given where relevant after indexed words. Synonyms are indicated by the word "see", or by an oblique stroke (/) if the synonyms fall together alphabetically.

α-decay, see Alpha decay
α-particles, see Alpha particles
α-rays, see Alpha rays
α-sulfur, see Alpha sulfur
β-decay, see Beta decay
β-particles, see Beta particles
β-rays, see Beta rays
β-sulfur, see Beta sulfur
γ-radiography, see Gamma radiography
γ-rays, see Gamma rays

1,2-dibromoethane (CH_2BrCH_2Br), 188 (Bromine), *193*
1,6-diaminohexane ($NH_2(CH_2)_6NH_2$), *200*
2-bromo,1-chloropentane, *215* (9)
2-methyl butane, *215* (8)
2-methyl propan-2-ol, 197
3-dimensional structural formula, see Stereochemical formula
3-iodohexane, *215* (9)
3-methyl pentane, *190*

A

Abdomen, 264 (Segmentation), 294 (Alimentary canal), 298 (Diaphragm)
Aberration,
  Chromatic, 55
  Spherical, 49
ABO blood groups, 287
Abomasum, 271 (Rumen)
A-bomb, see Atom bomb
Abrasive, 344
Abscisic acid, 249 (Abscission layer)
Abscission layer, 249
Absolute cubic expansivity, see Real cubic expansivity
Absolute refractive index, 37 (Refractive index), *50*
Absolute temperature scale, 27, 143
Absolute zero, 27 (Absolute temperature scale), 143 (Absolute temperature scale)
Abyssal, 342
Acceleration (a), 11, *97*
  Angular, *17*
  Centripetal, 17
  Linear, 14 (I)
  Uniform, 11 (Acceleration)
Acceleration due to gravity (g), 18, 113
Accelerators, Particle, 86 (Radioisotope)
Accessory foods, 204 (I), 330 (I)
Accommodation, 312 (Lens)
Accumulator, see Secondary cell
  Lead-acid, 69, *150*, *159*
Acellular, see Unicellular
Acetic acid, see Ethanoic acid
Acetone, see Propanone
Acetylene, see Ethyne
Acetylenes, see Alkynes
Achene, 262
Achromatic lens, *54* (Spectrometer), 55 (Chromatic aberration)
Acid(s), 150-152
  Abscisic, 249 (Abscission layer)
  Acetic, see Ethanoic acid
  Amino, 205, 239 (Ribosomes), 330 (Proteins), 331, 338, 339
  Ascorbic ($C_6H_8O_6$), *150*, *205*, 339 (Vitamin C)
  Bile, 338 (Bile)
  Carboxylic, 195, 202 (Soap), 205 (Lipids), *215* (7)
  Carbonic (H_2CO_3), 179 (Carbon dioxide), 210 (Acid rain)
  Deoxyribonucleic, see DNA
Acid(s) (cont'd)
  Dicarboxylic, 195
  Ethanedioic ($(COOH)_2$), *195*
  Ethanoic (CH_3COOH), *151*, *195*, *197*, *215* (7)
  Fatty, 330 (Fats), 338 (I, Pancreatic juice)
  Folic, 339 (Vitamin B complex)
  Formic, see Methanoic acid
  Fuming sulfuric ($H_2S_2O_7$), 185 (Contact process)
  Hexanedioic ($COOH(CH_2)_4COOH$), *200*
  Hydrochloric (HCl), *143*, *155*, *170*, *171*, 187 (Hydrogen chloride), 216 (Carbon dioxide, Chlorine), 217 (Hydrogen), 218 (Anions), 222 (Flame test), 338 (Gastric juice)
  Hydroiodic, 188 (Hydrogen iodide)
  Lactic, 334 (Oxygen debt)
  Methanoic (HCOOH), *150*, *195*, *215* (7)
  Nicotinic, see Niacin
  Nitric/Nitric(V) (HNO_3), *178*, 181 (Nitrogen dioxide), 182, 210 (Acid rain), 218 (Anions)
  Nitrous (HNO_2), 181 (Nitrogen dioxide)
  Nucleic, 324
  Octadecanoic ($C_{17}H_{35}COOH$), 202 (Soap), *205*
  Organic, 150
  Oxalic, see Ethanedioic acid
  Pantothenic, 339 (Vitamin B complex)
  Phosphoric (H_3PO_4), 182 (Phosphorus pentoxide)
  Propanoic (C_2H_5COOH), *196*
  Pyruvic, 334 (Anaerobic respiration)
  Ribonucleic, see RNA
  Stearic, see Octadecanoic acid
  Strong, 152
Acid(s) (cont'd)
  Sulfuric (H_2SO_4), *150*, *155*,
  (Acid rain), 216 (Ethene
  217 (Hydrogen), 218 (Nitra
  anion), 219 (Calcium cati
  344 (Dehydrating agent
  Sulfuric(IV), see Sulfurous ac
  Sulfuric(VI), see Sulfuric acid
  Sulfurous, *151*, 185
  Weak, 152
Acidic, 150
Acid radical, 153
Acid rain, *206*, 210
Acid salt, 153
Acoustics, 40 (I)
Acoustic waves, see Sound wa
Acrylic, *194*, 201
ACTH, 336
Actin, *205*, 283 (Striated muscl
Actinium (Ac), 164, 212
Actinomorphy, 264 (Radial sym
Action, Capillary, *23*, 252
Action, Local, 68
Action, Point, 57
Actions (human body), 308-30
  Autonomic, 309 (Involuntary actions)
  Involuntary, 309
  Reflex, 309
  Voluntary, 308
Activation energy (E), 160
Active electrode, 66 (Electrode
Active transport, 329
Activity series, see Reactivity se
Addition polymerization, 200
Addition reaction, 193
Additive mixing, 55 (Color mi
Adenine, 324
Adenohypophysis, see Anterior l
Adenoids, see Pharyngeal tonsil
Adenosine, 335 (ADP)
Adenosine diphosphate, see ADP
Adenosine triphosphate, see ATP
ADH, 336
Adhesion, 23
Adhesive, 344
Adipose tissue, *281*, 310 (Subcutaneous layer
ADP, 335
Adrenal glands, 297, *337*
Adrenalin/Adrenin, 336
Adrenocorticotropic hormone/ Adrenocorticotropin, see ACT
Advanced gas-cooled reactor (AGR), *95*
Advantage, Mechanical (M.A.), 2
Adventitious roots, 245
Aerial parts (plants), *244*
Aerial roots, 245
Aerobic respiration, 334
Aerometer, see Hydrometer
Aestivation, 237 (Dormancy
Afferent, *306*
Afferent arteriole, 301 (Glomerul

Acid(s) (cont'd)
  Sulfuric (H_2SO_4), *150*, *155*, **185**, 210 (Acid rain), 216 (Ethene), 217 (Hydrogen), 218 (Nitrate anion), 219 (Calcium cation), 344 (Dehydrating agent)
  Sulfuric(IV), see Sulfurous acid
  Sulfuric(VI), see Sulfuric acid
  Sulfurous, *151*, **185**
  Weak, **152**
Acidic, **150**
Acid radical, **153**
Acid rain, *206*, **210**
Acid salt, **153**
Acoustics, **40** (I)
Acoustic waves, see Sound waves
Acrylic, *194*, **201**
ACTH, **336**
Actin, *205*, **283** (Striated muscle)
Actinium (Ac), **164**, **212**
Actinomorphy, **264** (Radial symmetry)
Action, Capillary, *23*, **252**
Action, Local, **68**
Action, Point, **57**
Actions (human body), **308-309**
  Autonomic, **309** (Involuntary actions)
  Involuntary, **309**
  Reflex, **309**
  Voluntary, **308**
Activation energy (E), **160**
Active electrode, **66** (Electrode), 156
Active transport, **329**
Activity series, see Reactivity series
Addition polymerization, **200**
Addition reaction, **193**
Additive mixing, **55** (Color mixing)
Adenine, **324**
Adenohypophysis, see Anterior lobe
Adenoids, see Pharyngeal tonsil
Adenosine, **335** (ADP)
Adenosine diphosphate, see ADP
Adenosine triphosphate, see ATP
ADH, **336**
Adhesion, **23**
Adhesive, **344**
Adipose tissue, *281*, **310** (Subcutaneous layer)
ADP, **335**
Adrenal glands, **297**, 337
Adrenalin/Adrenin, **336**
Adrenocorticotropic hormone/ Adrenocorticotropin, see ACTH
Advanced gas-cooled reactor (AGR), *95*
Advantage, Mechanical (M.A.), **20**
Adventitious roots, **245**
Aerial parts (plants), *244*
Aerial roots, **245**
Aerobic respiration, **334**
Aerometer, see Hydrometer
Aestivation, **237** (Dormancy)
Afferent, *306*
Afferent arteriole, **301** (Glomerulus)
Afferent neurons, see Sensory neurons
Afferent system, **306-307**
  Somatic, **307**
  Visceral, **307**
Agent,
  Anti-knock, **199** (Octane rating)
  Dehydrating, **344**
  Depolarizing, **68** (Polarization)
  Drying, 221 (Desiccation), **344**
  Oxidizing, **148**
  Reducing, **148**
Agglutinins, **287** (Antibodies)
Agonist, **282** (Antagonistic pairs)
AGR, see Advanced gas-cooled reactor
Air, *183*, **208** (I), 211
  Downward displacement of, **217** (Hydrogen)
  Fractional distillation of (liquid air), 180 (Nitrogen), **183** (Oxygen), 189 (I), 208 (I)
  Upward displacement of, **216** (Chlorine), 217 (Oxygen)
Air bladder, see Swim bladder
Albumen, *276*
Alcohol(s), **196-197**, **196** (Ethanol), *214* (4a and b)
  Ethyl, see Ethanol
  Polyhydric, **197**
  Primary, 196, **197**
  Secondary, 196, **197**
  Tertiary, **197**
Alcoholic fermentation, **197**
Aldehydes, **194**, *215* (5)
Aldosterone, **336**
Algae (sing. alga), *238*, **340**
  Blue-green, 235
Alimentary canal/tract, **294**
Aliphatic compounds, **190**
Alkali, **151**
Alkali metals, 165, **168-169**
Alkaline, **151**
Alkaline cell, **69**
Alkaline-earth metals, 165, **170-171**
Alkanes, **192**, *214* (1)
Alkenes, **193**, *214* (2)
Alkyl group, 215 (8)
Alkyl halides, see Halogenoalkanes
Alkynes, **194**, *214* (3)
Allantois, *276*
Alleles/Allelomorphs, **325** (Genes)
Allotropes, **136** (Allotropy)
Allotropy, **136**
Alloy, **344**
Alpha decay (α-decay), **87**
Alpha-1 antitrypsin, **327** (Pharming)
Alpha particles (α-particles), **86**, **87** (Alpha decay), 90 (Cloud chamber), **128**
Alpha rays (α-rays), **86** (I)
Alpha sulfur (α-sulfur), see Rhombic sulfur
Alternating current (a.c.), **61**, 110
Alternation of generations, **321**
Alula, see Bastard wing
Alum, see Aluminum potassium sulfate-12-water
Alumina, see Aluminum oxide
Aluminum (Al), 112, *151*, 165, **176**, 211, 212, 219 (Cations)
Aluminum hydroxide ($Al(OH)_3$), **176**
Aluminum oxide (Al_2O_3), **176**, 183 (Oxides)
Aluminum potassium sulfate-12-water ($K_2SO_4.Al_2(SO_4)_3$), *154*
Aluminum sulfate ($Al_2(SO_4)_3$), **176**, 216 (Ethene)
Alveoli (sing. alveolus), **299**
Amalgam, **344**
Americium (Am), **164**, **212**
Amines, Primary, **195**
Amino acids, **205**, 239 (Ribosomes), **330** (Proteins), 331, 338, 339
Amino group ($-NH_2$), **195** (Primary amines, Diamines), 205 (Amino acids)
Ammeter, **77**, 110
Ammonia (NH_3), *132*, *133*, *143*, 151 (Base), *152*, 153 (I), *162*, *163*, 180 (Haber process), **181**, *182*, *209*, 219 (Cations)
Ammonia solution (NH_4OH), **181** (Ammonia), 218 (Anions), 219 (Cations)
Ammonium chloride (NH_4Cl), *143*, *153*, *162*, **181**
Ammonium hydroxide, see Ammonia solution
Ammonium nitrate (NH_4NO_3), *153*, **181**, 182 (Nitrates)
Ammonium sulfate ($(NH_4)_2SO_4$), **181**
Amnion, *276*, *319*
Amniotic fluid, *276* (Amnion), *319*
Amniotic sac, *276* (Amnion), *319*
Amoeba, *268*
Amorphous, **135** (I)
Amorphous silicon, **112**
Amp/Ampere (A), **60**, **96**, 159 (Current)
Ampere hours, **69** (Capacity)
Amphibians, 272 (External gills), 275 (Lateral lines), 277 (Metamorphosis), **341**
Amphoteric, **151**
Amplifier (symbol for), **110**
Amplitude, **16**, *34*
Ampulla (pl. ampullae), **315** (Semicircular ducts)
Ampulla of Vater, *296*
Amylase(s), **338** (I)
  Pancreatic, **338** (Pancreatic juice)
  Salivary, **338** (Saliva)
Anabolism, **332**
Anadromous, **236**
Anaerobic respiration, **334**
Anal canal, *294*, 295 (Large intestine)
Anal fin, **269** (Median fins)
Analog, **111** (Logic gates)
Anal sphincter, **295** (Large intestine)

Anaphase,
  (meiosis), **323**
  (mitosis), **241**
AND (logic gate/operation), 110, **111**
Androecium, **257**
Androgens, 316 (**Testes**), **336**, 337 (**LH**)
Aneurin, see **Thiamine**
Angle,
  Critical, **51**
  Visual, **55**
Angle of deviation, *51*
Angle of incidence (**i**), *47*, *50*
Angle of reflection (**r**), *47*
Angle of refraction (**r**), *50*
Angular acceleration, *17*
Angular magnification, **55** (**Visual angle**)
Angular velocity, *17*
Anhydrate, **154**
Anhydride, **151**
Anhydrite calcium sulfate ($CaSO_4$), **171** (**Calcium sulfate**)
Anhydrous, **154** (**Anhydrate**)
Animal(s), **264-277**, **341** (**I**)
  Higher, **264** (**I**)
  Lower, **264** (**I**)
Animal starch, see **Glycogen**
Anions, 66 (**Ionic theory of electrolysis**), **88** (**Ionization**), **130**, 218
Annelids, **341**
Annual rings, **246**, *247*
Annuals, **236**
Annular eclipse, *46*
Anode, **66** (**Electrode**), **156** (**Electrode**)
Anodizing, **157**
Anomalous expansion, **32**
Antacid, **344**
Antagonist, **282** (**Antagonistic pairs**)
Antagonistic hormones, **336**
Antagonistic pairs (muscles), **282**
Antennae (sing. **antenna**), **274**
Anterior cavity (eye), *312* (**Aqueous humor**)
Anterior lobe (pituitary gland), **297**
Anther, **257** (**Stamens**)
Antheridium, *321*
Anthophyta, **340**
Anthracite, **179** (**Coal**), *208*
Antibodies, **287**
Anti-diuretic hormone, see **ADH**
Antigen(s), **287**
  Rhesus, **287** (**Rhesus factor**)
Antigen-antibody complex, **287** (**Antibodies**)
Antihemophilic globulin, 327 (**Protein manufacture**)
Anti-knock agent, **199** (**Octane rating**)
Antimony (**Sb**), 112, 165, **180**, 212
Antineutrino, **87** (**Beta decay**)
Antinodal lines, *38*, **39** (**Antinodes**)
Antinodal points/Antinodes, **39**, *43*
Anti-toxins, **287** (**Antibodies**)
Anus, *294*, **295** (**Large intestine**)
Anvil (ears), see **Incus**
Antibiotics, *327* (**Gene cloning**)
Aorta, *289*, *290*, **291**
Aortic valve, **291** (**Semilunar valves**)
Apatite ($3Ca_3(PO_4)_2.CaF_2$), **182** (**Phosphorus**), 284 (**Enamel**)
Aperture,
  (camera), *54* (**Diaphragm**)
  (lens), *52*
  (mirror), *48*
Apex (leaves), *248*
Apical meristem, **244** (**Meristem**)
Apparatus (chemical), **223-225**
Apparatus, Golgi, see **Golgi complex**
Apparent cubic expansivity, **32**
Apparent depth, **51**
Apparent weightlessness, **19**
Appendage, **264**
Appendix, *294*, **295**
Aqueous humor, *312*
Aqueous solution, **144**
Aqueous solvent, **144**
Arachnoid, *303* (**Meninges**)
Archegonium, *321*
Archimedes' principle, **25**
Area (**A**), **97**
  (measurement of), **101**
  Surface (measurement of), **101**
Areal coefficient of expansion (β), **32**
Argon (**Ar**), 165, *183*, **189**, 212
Armature (electric motor), *76*
Aromatic compounds, **190**
Arsenic (**As**), 112, 165, **180**, 212
Arterial system, **288** (**Arteries**)
Arteriole(s), **288** (**Arteries**)
  Afferent, **301** (**Glomerulus**)
  Efferent, **301** (**Glomerulus**)
Artery(ies), **288**, *289*, **291**
  Brachial, *289*
  Brachiocephalic, *290*
  Celiac, *289*
  Common carotid, *290*
  Common iliac, *289*
  Femoral, *289*
  Gastric, *289*
  Gonadal, *289*
  Hepatic, *289*
  Inferior mesenteric, *289*
  Pulmonary, *290*, **291** (**Pulmonary trunk**)
  Renal, *289*, 300 (**Kidneys**)
  Splenic, *289*
  Subclavian, *289*, *290*
  Superior mesenteric, *289*
  Tibial, *289*
Arthropods, 265 (**Hemocoel**), **271**, 273 (**Spiracle, Malpighian tubules**), 274 (**Palps, Telson**), 275 (**Compound eyes**), *277* (**Larva**), **341**
Articular cartilage, **281** (**Cartilage**)
Articular processes (**Inferior, Superior**), *279*
Articulations, see **Joints**
Artificial propagation, **262**
Ascorbic acid ($C_6H_8O_6$), *150*, *205*, 339 (**Vitamin C**)
Asexual reproduction, **321**
Asphalt, see **Bitumen**
Association area(s), **302** (**Brain**), 303
  Auditory, **303**
  Visual, **303**
Association neurons, **305**
Astatine (**At**), 165, **186**, 212
Atlas, **279** (**Vertebral structure**)
Atmospheric pressure, **25** (**Barometer**), **113**
Atmospheric water, **206**
Atom(s), 4, **5**, 82-87, **124**, 125-134, 137-139
  Rutherford-Bohr, **82**
Atom bomb (**A-bomb**), see **Fission bomb**
Atomic energy, **84-85**
Atomicity, **124**
Atomic lattice, Giant, *4*, **137**
Atomic mass, Relative, **83**, **138**, 164-165, 212-213
Atomic mass units, Unified (**u**), **83** (**Relative atomic mass**)
Atomic nucleus, see **Nucleus**
Atomic number (**Z**), **82**, **127**, 164-165, 212-213
Atomic structure, **82-83**, **126-127**
Atomic theory, Dalton's, **124**
Atomic weight, see **Relative atomic mass**
ATP, 329 (**Active transport**), 334 (**Aerobic respiration**), **335**
Atria (sing. **atrium**), **290**
Atrioventricular valves, **291**
Attenuation, **35**
Auditory area, Primary, **303**
Auditory association area, **303**
Auditory ossicles, see **Ear ossicles**
Auditory tube, see **Eustachian tube**
Auricles,
  (ears), see **Pinnae**
  (heart), see **Atria**
Australopithecus, **341**
Autocatalysis, **161**
Autografting, **263** (**Grafting**)
Autonomic actions, **309** (**Involuntary actions**)
Autonomic ganglia, *309*
Autonomic nervous system, 303 (**Hypothalamus**), **308**
Autosomes, **325** (**Sex chromosomes**)
Autotrophic, **234** (**Food web**)
Auxins, **251** (**Growth hormones**)
Avalanche (particles), **89**
Avascular, 280 (**Connective tissue**), **289**
Average rate (of reaction), **160**
Average speed, **10** (**Speed**)
Average velocity, **10** (**Velocity**)
Avogadro's hypothesis/law, **143**
Avogadro's number, 95
  113, 139
AV valves, see **Atrio**
Axil, **244** (**Axillary bud**
Axillary bud, **244**
Axis,
  (bone), **279** (**Vertebral**
  Magnetic, **70**
  Principal (reflection,
  48, *52*
Axon, **304**

B

Baby teeth, see **Deci**
Backbone, see **Verte**
Back e.m.f., **60** (**Electro**
Background count, **88**
  radiation)
Background radiation,
Backward reaction, see
  reaction
Bacteria (sing. bacteri
  cloning)
  Denitrifying, *209*, **235**
  Nitrate, **235**
  Nitrifying, **235**
  Nitrogen-fixing, *209*,
  fixation)
Balance (equipment),
  Current, **60** (**Ampere**
  Spring, *22*
  Top pan, **225**
Balanced equation, **141**
Baleen, **270** (**Filter-feed**
Ball-and-socket joints,
Barbels, *275*
Barbs, **267** (**Feathers**)
Barbules, **267** (**Feathers**
Barium (**Ba**), 164, **170**,
  219 (**Flame tests**)
Barium chloride ($BaCl_2$
Bark, **247** (**Phellem**)
Barometer, **25**
  Simple, *25*
Basal metabolic rate, **332** (**Metabolic rate**
Basal rosette, **250**
Base (chemistry), 150
  Strong, **152**
  Weak, **152**
Base (transistor), **65**
Basic, **151** (**Base**)
Basic copper sulfate ($CuSO_4.3Cu(OH)_2$), **175**
Basicity, **153**
Basic oxygen process, **174**
Basic quantities, **96**
Basic salt, **154**
Basic SI units, **96**
Basic zinc chloride (
Bastard wing, *269*

Avogadro's number, 96 (**Mole**), **113**, **139**
AV valves, see **Atrioventricular valves**
Axil, **244** (**Axillary bud**)
Axillary bud, **244**
Axis,
  (bone), **279** (**Vertebral structure**)
  Magnetic, **70**
  Principal (reflection, refraction), *48*, *52*
Axon, **304**

B

Baby teeth, see **Deciduous teeth**
Backbone, see **Vertebral column**
Back e.m.f., **60** (**Electromotive force**)
Background count, **88** (**Background radiation**)
Background radiation, **88**
Backward reaction, see **Reverse reaction**
Bacteria (sing. **bacterium**), *327* (**Gene cloning**)
  Denitrifying, *209*, **235**
  Nitrate, **235**
  Nitrifying, **235**
  Nitrogen-fixing, *209*, **235** (**Nitrogen fixation**)
Balance (equipment),
  Current, **60** (**Ampere**)
  Spring, *22*
  Top pan, **225**
Balanced equation, **141**
Baleen, **270** (**Filter-feeding**)
Ball-and-socket joints, **280**
Barbels, *275*
Barbs, **267** (**Feathers**)
Barbules, **267** (**Feathers**)
Barium (**Ba**), 164, **170**, 212, **219** (**Flame tests**)
Barium chloride ($BaCl_2$), 218 (**Anions**)
Bark, **247** (**Phellem**)
Barometer, **25**
  Simple, *25*
Basal metabolic rate, **332** (**Metabolic rate**)
Basal rosette, **250**
Base (chemistry), 150 (**I**), **151**
  Strong, **152**
  Weak, **152**
Base (transistor), **65**
Basic, **151** (**Base**)
Basic copper sulfate ($CuSO_4.3Cu(OH)_2$), **175** (**Copper**)
Basicity, **153**
Basic oxygen process, **174** (**Steel**)
Basic quantities, **96**
Basic salt, **154**
Basic SI units, **96**
Basic zinc chloride ($Zn(OH)Cl$), *154*
Bastard wing, *269*
Battery, **68** (**I**), *69*, 110, **159** (**Cell**)
Bauxite ($Al_2O_3.2H_2O$), **176** (**Aluminum oxide**), 345 (**Ore**)
B.E., see **Binding energy**
Beaker, 101, **223**
Beat frequency, **42** (**Beats**)
Beats, **42**
Becquerel (**Bq**), *87*, **129**
Beehive shelf, **223**
Bell, Electric, **75** (**Electric buzzer**)
Bell flower, **259**
Benthos, **342**
Benzene ring, **190** (**Aromatic compounds**)
Berkelium (**Bk**), **165**
Berry, **262**
Beryllium (**Be**), 164, **170**, 212
Beta decay (β-decay), **87**
Beta particles (β-particles), **86**, 87 (**Beta decay**), 90 (**Cloud chamber**), **128**
Beta rays (β-rays), **86** (**I**)
Beta sulfur (β-sulfur), see **Monoclinic sulfur**
Biased (**Forward**, **Reverse**), **65** (**Diode**)
Bicarbonate of soda, see **Sodium bicarbonate**
Bicarbonates, 150 (**Acid**), **218** (**Anions**)
Biceps, *282*
Bi-concave (lens), *53*
Bi-convex (lens), *53*
Bicuspids, see **Premolars**
Bicuspid valve, **291** (**Atrioventricular valves**)
Biennials, **236**
Bilateral symmetry, **264**
Bile, 297 (**Liver**), 337 (**CCK**), **338**
Bile acids, **338** (**Bile**)
Bile duct, Common, *296*, **297** (**Gall bladder**)
Bile salts, **338** (**Bile**)
Bimetallic strip, **32**
Binary (compound), **122**
Binary fission, **321** (**Asexual reproduction**)
Binding energy (**B.E.**), **84**
Binding energy per nucleon, *85*
Binomial system, **340** (**Nomenclature**)
Biodegradable, **210**
Biodegradable detergents, **203**
Biological laundry powders, **203** (**Laundry powders**)
Biomass (ecology), **Pyramid of**, *235*
Biomes, **232**
Biopolymers, see **Natural polymers**
Biosphere, **232**
Biotic factors, **232** (**I**)
Biotin, **339** (**Vitamin B complex**)
Bipinnate, **250**
Birds, 233 (**Ecological niche**), **267**, **269** (**Pectoralis muscles**), **271**, **341**
Bismuth (**Bi**), *87*, 112, **165**, **180**, 212
Bitumen, **199**
Bituminous coal, **179** (**Coal**)
Bivalents, **322** (**Prophase**)
Black body, **29** (**Leslie's cube**)
Bladder, **300**, *316*
  Air, see **Swim bladder**
  Gall, **297**, 337 (**CCK**)
  Swim, **269**
Blast furnace, *174*
Blastocyst, **320** (**Embryo**)
Bleach, **344**
Blending, see **Incomplete dominance**
Blind spot, **313**
Blood, **286-287**
Blood cells (**Red**, **White**), **286**
Blood groups, **287** (**ABO blood groups**), 325 (**Codominance**)
Blood sinuses, **316** (**Penis**)
Blood vessels, **288** (**I**)
Blue-green algae, 235
BMR, see **Basal metabolic rate**
Body (vertebrae), see **Centrum**
Body-centered cubic (particle arrangement), *137*
Boiling, **30** (**Vaporization**), **121**
Boiling point, **30** (**Vaporization**), *31*, **121** (**Boiling**), 212-213
Boiling point test, **221**
Boiling tube, **225**
Bole, *247*
Bolus, **294** (**Esophagus**)
Bomb,
  Atom/Fission, **93**
  Fission-fusion, **93** (**Fusion bomb**)
  Fusion/Hydrogen, **93**
Bomb calorimeter, *147*
Bond(s), **130** (**I**)
  Coordinate, see **Dative covalent bond**
  Covalent, **132** (**I**)
  Dative covalent, **132**
  Double, **132**, 214 (**b** and **2**)
  Hydrogen, **134**
  Ionic, **131** (**I**)
  Polar, **133**
  Single, **132**, 214 (**a** and **1**)
  Triple, **132**, 214 (**c** and **3**)
Bond energy, **146**
Bonding, **130**, 131-134
  Covalent, **132-133**
  Ionic, **131**
  Metallic, **134**
Bone(s), **278-279**
  Coracoid, *269*
  Cranial, **278** (**Cranium**)
  Facial, **278** (**Cranium**)
Bone marrow, **281**
Bone (tissue), **280-281**
  Compact, **281**
  Spongy, **281**
Bony labyrinth, **314** (**Inner ear**)
Book lungs, see **Lung books**
Bordeaux mixture, **175** (**Copper(II) compounds**)
Boron (**B**), 165, **176**, 212
Boron trichloride (BCl_3), *133*
Bowman's capsule, **301**

Boyle's law, **33**, **142**
Brachial arteries, *289*
Brachial veins, *289*
Brachiocephalic artery, *290*
Brachiocephalic veins, *290*
Bract, **249**
Brain, **302**, *303*, 306, 308, 309
Brain stem, **303**
Branched chain (atoms), *190*
Branchiae (sing. **branchia**), see **Gills**
Branchial, **272**
Brass, **112**, *175*
Breaking stress, *23*
Breastbone, see **Sternum**
Breathing, **299**
Brine, **169** (**Sodium chloride**)
Brittle, **23** (**Yield point**)
Bromides, 186 (**I**), **188**, 218 (**Anions**)
Bromine (Br/Br_2), *92*, 165, 186, **188**, *193*, 212, *215* (**9**)
Bromine water, **188** (**Bromine**)
Bromoethane (C_2H_5Br), *215* (**9**)
Bromo group (–Br), *215* (**9**)
Bromothymol blue, *152*
Bronchi (sing. **bronchus**), *298*, **299**
  Primary, *298*, **299** (**Bronchi**)
  Secondary, *298*, **299** (**Bronchi**)
  Tertiary, *298*, **299** (**Bronchi**)
Bronchial, **299**
Bronchioles, *298*, **299**
  Terminal, **299** (**Bronchioles**)
Bronsted-Lowry theory, **151**
Bronze, *175*
Brownian motion, **5**, **123**
Brushes (electric motor), *76*
Bryophyta, **340**
Bubble caps, *198*
Bubble chamber, **90**
Buccal cavity, see **Oral cavity**
Buchner flask, *220*, **224**
Buchner funnel, *220*, **224**
Buckminsterfullerene, **179**
Bud(s) (plants), **244**
  Axillary/Lateral, **244**
  Primary, **261** (**Plumule**)
  Secondary, see **Axillary bud**
  Terminal, **244**
Buds, Taste, **307** (**Tongue**)
Budding,
  (artificial propagation), **263**
  (asexual reproduction), see **Gemmation**
Bulb,
  (hairs), *311*
  (light), **110**
  Olfactory (nose), *307*
  (plants), **263**
Bulbourethral glands, see **Cowper's glands**
Bunsen burner, **223**
Buoyant force, **25**
Burette, 101, **223**
Burning, see **Combustion**
But-1-ene (C_4H_8), **193**
But-1-yne (C_4H_6), **194**
Butan-1-ol (C_4H_9OH), *190*, **196**, **197**
Butan-2-ol, **197**, *215* (4b)
Butane (C_4H_{10}), **192**, 199 (**Refinery gas**)
Butanone ($CH_3COCH_2CH_3$), *215* (6)
Buzzer, Electric, **75**

C

Cadmium (**Cd**), 112, 165, **173**, 212
Calamine ($ZnCO_3$), **175** (**Zinc**)
Calciferol, see **Vitamin D**
Calcite, **171** (**Calcium carbonate**)
Calcitonin, see **TCT**
Calcium (**Ca**), 164, 170, **171**, 207, 211, 212, 219 (**Flame tests, Cations**)
Calcium bicarbonate ($Ca(HCO_3)_2$), *171*, 207 (**Temporary hardness**)
Calcium carbonate ($CaCO_3$), 141 (**Trivial name**), *153*, **171**, 179 (**Carbonates**), 207 (**Temporary hardness**), 216 (**Carbon dioxide**)
Calcium chloride ($CaCl_2$), *138*, *149*, **171**, *206*, *216*, 221 (**Desiccation**)
Calcium hydroxide ($Ca(OH)_2$), **171**
Calcium metasilicate ($CaSiO_3$), **177** (**Silicates**)
Calcium oxide (**CaO**), **171**, *174*, 183 (**Oxides**)
Calcium stearate, *207*
Calcium sulfate ($CaSO_4$), **171**, 185 (**Sulfates**)
  Anhydrite, **171** (**Calcium sulfate**)
Calibration, **344**
Californium (**Cf**), **165**
Callipers, Vernier slide, **100**
Calorific value, **208**
Calorimeter, Bomb, *147*
Calorimetry, **344**
Calyx, **256** (**Sepals**)
Cambium, *242*, **243**, 246 (**Vascular cylinder**), *247*
  Cork, see **Phellogen**
Camera, **54**
Candela (**cd**), **96**
Canines (teeth), **285**
Capacitance (**C**), **59**, **97**
Capacitor, **59**, 110
  Electrolytic, **59**, 110
  Paper, **59**
  Polyester, **59** (**Paper capacitor**)
  Variable, **59**, 110
Capacity,
  (electrochemical cells), **69**
  Heat (**C**), **31**
  Specific heat (**c**), **31**, 112
  (volume), **101**
Capillaries, **288**, 292 (**Tissue fluid**)
  Lymph, **293** (**Lymph vessels**)
Capillarity/Capillary action, *23*, **252**
Capillary tube, *23*, *24*, **26** (**Liquid-in-glass thermometer**)
Carapace, **266**
Carbohydrates, **204**, *209*, 254-255, **330**, 338, 339
Carbon (**C**), *83*, 88 (**Background radiation**), 91 (**Radiocarbon dating**), *127*, 138, *148*, 165, 174 (**Steel**), 177, **178**, 209, 211, 212, 214-215, 234, 330
Carbonates, 150 (**Acid**), 155 (**method 2**), **179**, 211, 218 (**Anions**)
Carbon cycle, **209**, **234**
Carbon dating, see **Radiocarbon dating**
Carbon dioxide (CO_2), *133*, *140*, *148*, 150 (**Acid**), *169*, *171*, *174*, *178*, **179**, 183 (**Oxides**), *197*, *204*, *208*, **209** (**Photosynthesis**), 210, 211, 216, 217 (**Nitrogen**), 218 (**Gases, Anions**), 234, 254-255, 288, 298-299
Carbon fibers, **179**
Carbonic acid (H_2CO_3), **179** (**Carbon dioxide**), 210 (**Acid rain**)
Carbon monoxide (**CO**), *174*, *178*, **179**, *210*
Carbon resistor, **62** (**Resistor**)
Carbonyl group (–CO–), **194** (**Ketones**), 214 (f), 215 (6)
Carboxyl group (–COOH), **195** (**Carboxylic acids**), 205 (**Amino acids**), 214 (g), 215 (7)
Carboxylic acids, **195**, 202 (**Soap**), 205 (**Lipids**), *215* (7)
Carboxypeptidase, **338** (**Pancreatic juice**)
Cardiac, **290**
Cardiac cycle, **291**
Cardiac muscle, **282**
Cardiac muscle (tissue), **283**
Cardiac sphincter, *294*, **295**
Cardiovascular system, **290** (**I**)
Carnassial teeth, **270**
Carnivores, **234** (**Secondary consumers**)
Carotene, 255 (**Pigments**), *310*, 339 (**Vitamin A**)
Carotid arteries, Common, *290*
Carpals, *278*, *280*
Carpel, *256*, **257**
Carpus, *278* (**Carpals**)
Cartilage, **281**
  Articular, **281** (**Cartilage**)
  Costal, **278** (**Rib cage**)
Cartilaginous joints, **281** (**Cartilage**)
Caryopsis, see **Grain**
Casein, **338** (**Gastric juice**)
Castes, **342** (**Social**)
Cast iron, **112**, **174** (**Iron**)
Catabolism, **332**
Catadromous, **236** (**Anadromous**)
Catalysis, **161** (**Catalyst**)
Catalyst, **161**, **333** (**Enzymes**)
  Heterogenous, **161**
  Homogenous, **161**
  Surface, **161**
Catalytic cracking/Cat cracking, **198** (**Cracking**)
Catalytic proteins, 330
  333 (**Enzymes**)
Caterpillar, *277* (**Larv**
Cathode, **66** (**Electrod**
  156 (**Electrode**)
Cathode ray oscillosc
  73, **81**
Cathode rays, **80-81**
Cathodic protection, see
  protection
Cations, 66, **88** (**Ioniz**
Caudal, **269**
Caudal fin, **269** (**Med**
Caudate, **269** (**Caudal**
Caustic curve, **49** (**Sph**
  aberration)
Caustic potash, see **Pot**
  hydroxide
Caustic soda, see **Sod**
CCK, **336**
Cebidae, **341**
Cecum, **271**, *294*, 295
Celiac artery, *289*
Cell(s) (biology), **238-241**
  Blood (**Red, White**), **28**
  Companion, **243** (**Ph**
  Daughter, **240** (**Cell**
  Egg, 258 (**Ovules**), 320
   323 (**Gamete prod**
  Flame, **273** (**Nephridi**
  Germ, see **Gametes**
  Guard, **249**
  Interstitial, *316*, 337
  Olfactory, **307** (**Nose**)
  Palisade, **248** (**Palisade**
  Parent, **240** (**Cell divisi**
  Passage, **243** (**Endoder**
  Schwann, **304** (**Nerve**
  Spongy, **248** (**Spongy**
  Target, **336** (**Hormones**
  Thread, see **Cnidoblasts**
Cell(s) (electrochemistry),
  110, **159**
  Alkaline, **69**
  Daniell, *159*
  Downs', *168*
  Dry, **69**
  Electrolytic, **66**, **156**
  Half, **158**
  Leclanché, **69**
  Nickel-cadmium, **69**
  Primary, **69**, **159** (**Cell**)
  Secondary, **69**, **159** (**Ce**
  Simple, **68**
  Standard, **69**
  Storage, see **Secondary**
  Voltaic, **68** (**I**)
Cell body, **304**, *305*
Cell division, **240-241**, 322-
Cell membrane, **238**, *239*, 3
Cell plate, **241** (**Cytokinesis**
Cell sap, **238** (**Vacuoles**), 253
Cellular respiration, see **Inter**
  respiration

- **Catalytic proteins**, 330 **(Proteins)**, **333 (Enzymes)**
- **Caterpillar**, *277* **(Larva)**
- **Cathode, 66 (Electrode), 156 (Electrode)**
- **Cathode ray oscilloscope (CRO)**, *73*, **81**
- **Cathode rays, 80-81**
- **Cathodic protection**, see **Sacrificial protection**
- **Cations**, 66, **88 (Ionization)**, **130**, 219
- **Caudal, 269**
- **Caudal fin, 269 (Median fins)**
- **Caudate, 269 (Caudal)**
- **Caustic curve, 49 (Spherical aberration)**
- **Caustic potash**, see **Potassium hydroxide**
- **Caustic soda**, see **Sodium hydroxide**
- **CCK, 336**
- **Cebidae, 341**
- **Cecum, 271**, *294*, 295 **(Large intestine)**
- **Celiac artery**, *289*
- **Cell(s)** (biology), **238-241**, 322-323
  - **Blood (Red, White), 286**
  - **Companion, 243 (Phloem)**
  - **Daughter, 240 (Cell division)**
  - **Egg**, 258 **(Ovules)**, 320 **(Gametes)**, 323 **(Gamete production, female)**
  - **Flame, 273 (Nephridia)**
  - **Germ**, see **Gametes**
  - **Guard, 249**
  - **Interstitial**, *316*, 337 **(Androgens)**
  - **Olfactory, 307 (Nose)**
  - **Palisade, 248 (Palisade layer)**
  - **Parent, 240 (Cell division)**
  - **Passage, 243 (Endodermis)**
  - **Schwann, 304 (Nerve fibers)**
  - **Spongy, 248 (Spongy layer)**
  - **Target, 336 (Hormones)**
  - **Thread**, see **Cnidoblasts**
- **Cell(s)** (electrochemistry), **68-69**, 110, **159**
  - **Alkaline, 69**
  - **Daniell**, *159*
  - **Downs'**, *168*
  - **Dry, 69**
  - **Electrolytic, 66, 156**
  - **Half, 158**
  - **Leclanché, 69**
  - **Nickel-cadmium, 69 (Alkaline cell)**
  - **Primary, 69, 159 (Cell)**
  - **Secondary, 69, 159 (Cell)**
  - **Simple, 68**
  - **Standard, 69**
  - **Storage**, see **Secondary cell**
  - **Voltaic, 68 (I)**
- **Cell body, 304**, *305*
- **Cell division, 240-241, 322-323**
- **Cell membrane, 238**, *239*, 329
- **Cell plate, 241 (Cytokinesis)**
- **Cell sap, 238 (Vacuoles), 253 (Turgor)**
- **Cellular respiration**, see **Internal respiration**
- **Cellulase, 331 (Roughage)**
- **Cellulose**, *238* **(Cell wall)**, 271 **(Rumen)**, **331 (Roughage)**
- **Cell wall**, *238*, 243 **(Sieve tubes)**
- **Celsius scale, 27**, 143 **(Absolute temperature scale)**
- **Cement/Cementum** (teeth), **284**
- **Center of curvature (C)**, *48*, *52*
- **Center of gravity, 10 (Center of mass)**
- **Center of mass, 10**
- **Central incisors**, *285*
- **Central nervous system, 302-303**
- **Central processing unit (CPU), 111**
- **Centrifugal force, 17 (Centripetal force)**
- **Centrifuge**, *220*
- **Centrifuging, 220**
- **Centrioles, 240**, 241, 322, 323
- **Centripetal acceleration, 17**
- **Centripetal force, 17**
- **Centromeres, 241, 322**
- **Centrosome**, 240 **(Centrioles)**
- **Centrum, 279 (Vertebrae)**
- **Cephalic, 303**
- **Cephalic veins**, *289*
- **Cephalochordates, 341**
- **Cephalopods**, 272 **(Siphon)**
- **Cephalothorax**, *274*
- **Cercopithecoidae, 341**
- **Cerebellum, 303**
- **Cerebral, 303**
- **Cerebral cortex, 303 (Cerebrum)**
- **Cerebral hemispheres, 303 (Cerebrum)**
- **Cerebrospinal fluid**, *303*
- **Cerebrum, 303**
- **Cerium (Ce), 164, 212**
- **Cerumen, 314 (Outer ear)**
- **Ceruminous glands, 314 (Outer ear)**
- **Cervical canal**, *317*
- **Cervical vertebrae**, *278*
- **Cervix,**
  - (teeth), see **Neck**
  - (uterus), **317**
- **Cesium (Cs)**, 164, 166, **168**, 212
- **Cetaceans, 341**
- **Chaetae** (sing. **chaeta**), **268 (Parapodia)**
- **Chain** (atoms), **190 (I)**
  - **Branched**, *190*
  - **Main**, *190*
  - **Side**, *190*, 214 **(h)**, 215 **(8)**
  - **Straight**, *190*
- **Chain, Decay**, see **Decay series**
- **Chain, Food, 234 (Food web)**
- **Chain reaction, 92 (Induced fission), 129 (Nuclear fission)**
- **Chalazae** (sing. **chalaza**),
  - (animal eggs), *276*
  - (flowering plants), **257 (Ovaries)**
- **Chalk**, 141 **(Trivial name), 171 (Calcium carbonate)**
- **Changeover switch**, *64*
- **Changes of state, 30-31, 120-121**
- **Characteristics of life, 236 (I)**
- **Charcoal, 179 (Coal)**
- **Charge** (electric), **56-59**, 97
  - **Surface density of, 57**
- **Chelae** (sing. **chela**), *274*
- **Chemical energy, 8**, *9*
- **Chemical equilibrium, 163**
- **Chemical feedstocks, 199**
- **Chemical properties, 119**
- **Chemical reactions, 119**, 146-149, 156-163, 166
- **Chemical symbols, 122**, 212-213, 226-227
- **Chiasmata** (sing. **chiasma**), **322 (Crossing over)**
- **Chile saltpeter**, see **Sodium nitrate**
- **Chitin, 266 (Cuticle)**
- **Chlorides**, *130*, *131*, *155*, **187**, 218 **(Anions)**
- **Chlorine (Cl/Cl_2)**, *131*, *132*, *138*, *146*, *148*, *149*, *160*, 165, *170*, **187**, *192*, *207*, 212, *215* **(9)**, 216, 218 **(Anions)**
- **Chloroethane (C_2H_5Cl)**, *195*
- **Chlorofluorocarbons (CFCs), 195**, 210 **(Ozone depletion)**, 345 **(Refrigerant)**
- **Chloro group (–Cl)**, *195*, 215 **(9)**
- **Chloromethane (CH_3Cl)**, *192*, *215* **(9)**
- **Chlorophyll**, *170*, 254 **(Chloroplasts)**, **255 (Pigments)**
- **Chlorophyta, 340**
- **Chloroplasts**, 240 **(Plastids)**, 249 **(Guard cells), 254**
- **Cholecystokinin**, see **CCK**
- **Chordates, 341**
- **Chorion**, *319*
- **Chorionic villi**, *319*
- **Choroid/Choroid coat**, *312*
- **Chromatic aberration, 55**
- **Chromatic scale**, *43*
- **Chromatids**, 240 **(Interphase), 241, 322-323**
- **Chromatin, 238 (Nucleus)**, 240 **(Interphase)**, 322 **(Prophase)**
- **Chromatism**, see **Chromatic aberration**
- **Chromatography, 221**
  - **Column, 221 (Chromatography)**
  - **Gas, 221 (Chromatography)**
  - **Paper**, *221*
- **Chrome iron ore, 172 (Chromium)**
- **Chromium (Cr)**, 165, **172**, 212
- **Chromosomes**, 238 **(Nucleus)**, 240 **(Mitosis)**, 241, 322, 323, **324**, 325, 326
  - **Daughter, 241 (Anaphase)**
  - **Homologous**, 240 **(Mitosis)**, 322, 323, **324 (Chromosomes)**, 326 **(Law of segregation)**
  - **Sex, 325**, 326 **(Sex linkage)**
  - **X, 325 (Sex chromosomes)**, 326 **(Sex linkage)**
  - **Y, 325 (Sex chromosomes)**, 326 **(Sex linkage)**
- **Chrysalis**, *277* **(Pupa)**
- **Chyme, 295 (Stomach)**

Chymotrypsin, 338 (Pancreatic juice)
Chymotrypsinogen, 338 (Note 2)
Cilia (sing. cilium), 268
Ciliary body, *312*
Cinnabar (HgS), 173 (Mercury)
Circuit(s), 60 (I), 159 (Current)
  Integrated, 111 (Logic gates)
  Primary, *78*
  Secondary, *78*
  (symbols), 110
Circular motion, 17
  Uniform, 17 (Circular motion)
Circular wavefronts, *34*
Circulatory system, 288-289
Cirri (sing. cirrus), 270 (Filter-feeding)
Cis but-2-ene, *191*
Clades, 341 (Cladistic classification)
Cladistic classification, 341
Clamps, 225
Classes, 340 (I)
Classical taxonomy, 340 (I)
Classification, 340 (I)
Clavicle, *278*
Cleavage, 320 (Embryo)
Cleavage furrow, 241 (Cytokinesis)
Cleavage plane, 136
Cleidoic eggs, 276 (Eggs)
Climatic factors, 232 (I)
Climax community, 233 (Ecological succession)
Clinical thermometer, *26*
Clitoris, 317 (Vulva)
Cloaca, *271*
Cloacal aperture/vent, *271*
Closed system, 162
Clot, 287 (Clotting)
Clotting, 287
Cloud chamber, 90
  Diffusion, *90*
  Wilson, *90*
Cnidarians, 270 (Cnidoblasts), 275 (Tentacles), 341
Cnidoblasts, 270
CNS, see Central nervous system
Coagulation, see Clotting
Coal, 179, *208*, *209*
  Bituminous, 179 (Coal)
  Destructive distillation of, 179 (Coal)
Coal gas, 179 (Carbon monoxide, Coal)
Coal tar, 179 (Coal)
Cobalamin, see Cyanocobalamin
Cobalt (Co), 112, 165, 172, 212
Cobalt(II) chloride ($CoCl_2$), 172 (Cobalt), 218 (Water)
Coccygeal vertebrae, *278*
Coccyx, *278*
Cochlea, 314
Cochlear duct, 314, *315*
Cocoon, *277* (Pupa)
Codominance, 325
Coefficient, 344
Coefficient of expansion,
  Areal (β), 32
  Linear (α), 32
  Volume (γ), 32
Coefficient of friction (μ), 7
Coefficient of kinetic friction, 7 (Coefficient of friction)
Coefficient of static friction, 7 (Coefficient of friction)
Coelom, 265
Co-enzymes, 333 (Enzymes), 339 (Vitamin B complex, Vitamin C)
Coherent waves, 38 (I)
Cohesion, 23
Coil (electromagnetism), 74
  Flat/Plane, 74
  Primary, 79
  Secondary, 79
Coitus, see Copulation
Coke, *174*, 179 (Coal), *208*
Coleoptile, 261
Collagen, 280 (Connective tissue), 339 (Vitamin C)
Collarbone, see Clavicle
Collecting duct/tubule, 301
Collector, 65 (Transistor)
Collenchyma, 243 (Cortex)
Collision, 13
Collision theory, 160
Colloid, 145
Colon, *294*, 295 (Large intestine)
Colonial, 342
Color(s), 54-55
  Complementary, *55*
  Primary, 55
  Secondary, *55*
Color mixing, 55
Column chromatography, 221
Combustion, 208
  Enthalpy change of/Heat of, 146
  Rapid, 208
  Slow, 208
Commensals, 342
Common bile duct, *296*, 297 (Gall bladder)
Common carotid arteries, *290*
Common hepatic duct, *296*, 297 (Liver)
Common iliac arteries, *289*
Common iliac veins, *289*
Community, 233
  Climax, 233 (Ecological succession)
  Pioneer, 233
  Successional, 233
Commutator, *76*
Compact bone, 281
Companion cells, 243 (Phloem)
Compass, Plotting, 72 (Magnetic field lines)
Compensation points, 255
Complementary colors, *55*
Complete metamorphosis, 277
Completion (chemical reaction), 162 (I)
Complex ion, 154 (Complex salt)
Complex salt, 154
Components, *6*, 108 (Resolution)
Composite flower, see Flowerhead
Compound(s), 4, 122
  Aliphatic, 190
  Aromatic, 190
  Copper(I), 175
  Copper(II), 175, 219 (Cations)
  Covalent, 132 (I)
  Cupric, see Copper(II) compounds
  Cuprous, see Copper(I) compounds
  Ferric, see Iron(III) compounds
  Ferrous, see Iron(II) compounds
  Ionic, *4*, 131
  Iron(II), 174, 219 (Cations)
  Iron(III), 174, 219 (Cations)
  Lead(II), 177 (Lead), 219 (Cations)
  Lead(IV), 177 (Lead)
  Nitrate(III), see Nitrites
  Nitrate(V), see Nitrates
  Organic, 190 (I), 214-215
  Plumbic, see Lead(IV) compounds
  Plumbous, see Lead(II) compounds
  Polyunsaturated, 191
  Saturated, 191
  Sulfate(IV), see Sulfites
  Sulfate(VI), see Sulfates
  Unsaturated, 191
Compound eyes, 275
Compound leaves, 248 (I), 250
Compound microscope, 54
Compression, 7
Compressions (waves), 35
Computers, 111
Concave lens, *52*, 53
Concave meniscus, 53 (Concave lens)
Concave mirror, 48
Concentrated, 144
Concentration, 139
  Mass, 139 (Concentration)
Condensation, 30, 121
Condensation polymerization, 200
Condensation reaction, 197
Condenser(s), 223
  Liebig, *220*, 223
  Reflux, 223
Condensing electroscope, 56 (Electroscope)
Conduction,
  (electricity), 56 (Conductor)
  Thermal (heat), 28
Conductivity,
  (electricity), 63
  Thermal (heat), 28, 112
Conductor(s),
  (electricity), 56
  Lightning, 57 (Lightning)
  Thermal (heat), 28 (Conduction)
Cones (eyes), 313 (Retina)
Conical flask, 224
Coniferous forest (biome), 232
Conifers, 340
Conjunctiva, *312*
Connecting neurons, see Association neurons
Connective tissue, 280
Consequent poles,
Conservation of energy,
  8 (I), 146
Conservation of
  Law of, 13
Conservation of mass,
Constant(s), 113, 344
  Dielectric, 59
  Gas, *33*, 113, *142*
  Gravitational (G),
    of gravitation),
  Faraday's, 113
  Planck's, *84*
Constantan, 112
Constant composition,
Constructive interference,
Consumers,
  Primary/First order,
  Secondary/Second order,
  Tertiary/Third order,
Contact force, 7
  Normal, 7 (Coefficient
Contact process, 185
Contour feathers, 267
Contractile vacuoles,
Control grid (oscilloscope),
Control rods, 94
Convection, 28
Convection current, 28
Converging lens, 52
Converging mirror, see
Conversion (energy), 8
Convex lens, *52*, 53
Convex meniscus, 53
Convex mirror, 48
Convoluted tubules
  301 (Uriniferous tubules
Coolant, *94*, *95*, 344
Cooling curve, *31*
Coordinate bond, see
  bond
Copolymer, 200
Copper (Cu), 112, *148*,
  175, *185*, 211, 212,
  219 (Flame tests,
Copper(I) chloride (CuCl),
  175 (Copper(I) compounds
Copper(I) compounds, 175
Copper(I) oxide (Cu_2O),
  compounds)
Copper(II) chloride ($CuCl_2$),
  175 (Copper(II) compounds
Copper(II) compounds, 175
  219 (Cations)
Copper(II) nitrate
Copper(II) oxide (CuO),
  217 (Nitrogen)
Copper(II) sulfate
  154, *155*, *157*, 175
  compounds), *185*, 217
  218 (Water)
  Basic
    175 (Copper)

Connective tissue, 280
Consequent poles, *71*
Conservation of energy, Law of, 8 (I), 146
Conservation of linear momentum, Law of, 13
Conservation of mass, Law of, 125
Constant(s), 113, **344**
  Dielectric, 59
  Gas, *33*, **113**, *142*
  Gravitational (G), **18 (Newton's law of gravitation)**, 113
  Faraday's, 113
  Planck's, *84*
Constantan, 112
Constant composition, Law of, 125
Constructive interference, 39
Consumers,
  Primary/First order, 235
  Secondary/Second order, 235
  Tertiary/Third order, 235
Contact force, 7
  Normal, 7 (Coefficient of friction)
Contact process, 185
Contour feathers, 267 (Feathers)
Contractile vacuoles, *268*, **273**
Control grid (oscilloscope), *81*
Control rods, 94 (Fission reactor), *95*
Convection, 28
Convection current, 28 (Convection)
Converging lens, 52
Converging mirror, see **Concave mirror**
Conversion (energy), **8 (I),** *9*
Convex lens, 52, 53
Convex meniscus, 53 (Convex lens)
Convex mirror, 48
Convoluted tubules (Distal, Proximal), 301 (Uriniferous tubules)
Coolant, *94*, *95*, **344**
Cooling curve, *31*
Coordinate bond, see **Dative covalent bond**
Copolymer, 200
Copper (Cu), 112, *148*, *158*, 165, 172, **175**, *185*, 211, 212, 217 **(Nitrogen)**, 219 **(Flame tests, Cations)**
Copper(I) chloride (CuCl), 175 (Copper(I) compounds)
Copper(I) compounds, 175
Copper(I) oxide (Cu_2O), 175 (Copper(I) compounds)
Copper(II) chloride ($CuCl_2$), 175 (Copper(II) compounds)
Copper(II) compounds, 175, 219 **(Cations)**
Copper(II) nitrate ($Cu(NO_3)_2$), *181*
Copper(II) oxide (CuO), *148*, *155*, *185*, *217* **(Nitrogen)**
Copper(II) sulfate ($CuSO_4$), *135*, *153*, *154*, *155*, *157*, **175 (Copper(II) compounds)**, *185*, 217 **(Hydrogen)**, 218 **(Water)**
  Basic ($CuSO_4.3Cu(OH)_2$), 175 (Copper)
Copper pyrites ((CuFe)S_2), 175 (Copper)
Copper voltameter, 67 (Voltameter)
Copulation, 319
Coracoid bones, *269*
Core,
  (electromagnets), **74**
  (nuclear reactor), **94 (Fission reactor)**
Cork, see **Phellem**
Cork cambium, see **Phellogen**
Corm, 263
Cornea, *312*
Cornified layer, see **Stratum corneum**
Corolla, 256 (Petals)
Corpus callosum, 303 (Cerebrum)
Corpuscles,
  Malpighian, see **Renal corpuscles**
  Meissner's, 311
  Pacinian, 311
  Red, see **Red blood cells**
  Renal, 301
  White, see **White blood cells**
Corpus luteum, 318 (Menstrual cycle), 337 **(LH, Lactogenic hormone, Estrogen)**
Corrosion, *157*, **209**
Cortex,
  (adrenal glands), **297**
  Cerebral, 303 (Cerebrum)
  (kidneys), *300*
  (plants), **243**
  Renal, see **Cortex** (kidneys)
  Secondary (plants), **247 (Phelloderm)**
Corti, Organ of, 314 (Cochlear duct)
Cortisol, see **Hydrocortisone**
Cortisone, 336
Corundum (Al_2O_3), 176 (Aluminum oxide)
Cosine, 344
Cosmic radiation, 88 (Background radiation)
Cosmic rays, 86 **(Radioisotope), 88 (Background radiation)**
Costal cartilage, 278 (Rib cage)
Cotyledon, 261
Coulomb (C), 60, 97, 157
Coulometer, see **Voltameter**
Couple (forces), **14**
Covalency, 133
Covalent bond, 132 (I)
  Dative, 132
Covalent bonding, 132-133
Covalent compounds, 132 (I)
Coverts, *267*
Cowper's glands, *316*
CPU, see **Central processing unit**
Cracking, 198
  Catalytic/Cat, 198 (Cracking)
Cranial bones, 278 (Cranium)
Cranial nerves, 302 (Brain)
Cranial reflexes, 309 (Reflex actions)
Cranium, 278
Creamy (precipitate), *145*
Crests (waves), **34**
Cristae (sing. **crista**), **240 (Mitochondria)**
Critical angle (c), 51
Critical length, 251 (Photoperiodism)
Critical mass, 93
Critical ray, 51 (Critical angle)
Critical temperature, 5 (Gas, Vapor), 120 (Gaseous state)
CRO, see **Cathode ray oscilloscope**
Crop (birds, etc.), **271**
Crossing over (chromosomes), **322**
Cross pollination, 259
Crown (teeth), **284**
Crucible, 223
Crude oil, see **Petroleum**
Crustaceans, 272 **(Internal gills),** *274* **(Antennae)**
Cryolite (Na_3AlF_6), *176*
Crypts of Lieberkühn, see **Intestinal glands**
Crystal(s), 135-137
  Scintillation, 90 (Scintillation counter)
  Seed, 135
Crystal lattices, 136 **(I), 137**
Crystalline, 135 **(I)**
Crystallization, 135, 221
  Water of, 135
Crystallizing dish, 223
Crystallography, X-ray, 136
Cubic (basic crystal shape), *136*
  Body-centered (particle arrangement), *137*
  Face-centered (particle arrangement), *137*
  Simple (particle arrangement), *137*
Cubic close packing (particle arrangement), *137*
Cubic expansivity,
  Absolute, see **Real cubic expansivity**
  Apparent, 32
  Real, 32
Cud, 271 (Rumen)
Cupric compounds, see **Copper(II) compounds**
Cupronickel, *175*
Cuprous compounds, see **Copper(I) compounds**
Cupulae (sing. **cupula**), **315 (Semicircular ducts)**
Curds, 338 (Gastric juice)
Curie (unit), *87*, **129 (Becquerel)**
Curium (Cm), 165
Current, Convection, 28 (Convection)
Current (electric) **(I)**, 56 **(I), 60-64,** 96 **(Ampere), 159**
  Alternating (a.c.), 61, 110
  Direct (d.c.), 61
  Eddy, 79
Current balance, 60 (Ampere)
Curvature,
  Center of (C), *48*, *52*
  Radius of (r), *48*
Curved mirrors, 48(I)

Cusp, 285 (Canines, Premolars, Molars)
Cuspids, see Canines
Cuticle,
  (animals), 266
  (plants), 243
Cutin, 243 (Cuticle)
Cutis, see Skin
Cutting, 263
Cyanocobalamin, 339 (Vitamin B complex)
Cycle(s),
  Carbon, 209, 234
  Cardiac, 291
  Energy, 147 (Hess's law)
  Life, 236 (I)
  Menstrual, 318, 337 (FSH, Estrogen)
  (motion), 16
  Nitrogen, 209, 235
  Ovarian, 318 (Menstrual cycle)
  Water, 206
Cycloalkanes, 192
Cyclohexane, *190*, 192 (Cycloalkanes)
Cyclotrons, 86 (Radioisotope)
Cystic duct, *296*, 297 (Gall bladder)
Cytokinesis, 241, 323 (Telophase)
Cytokinins, 251 (Growth hormones)
Cytoplasm, 238
Cytosine, 324

D

Dalton's atomic theory, 124
Dalton's law of partial pressures, 143
Damping, 16
Daniell cell, *159*
Darwinism, 237 (Genetic variation)
Dating (Carbon/Radiocarbon), 91, 129
Dative covalent bond, 132
Daughter cells, 240 (Cell division)
Daughter chromosomes, 241 (Anaphase)
Daughter nuclei, 241 (Telophase), 322 (Meiosis)
Decanting, 220
Decay (biological), 234, *235*
Decay, Radioactive, 87, 128
  Alpha (α-decay), 87
  Beta (β-decay), 87
Decay chain/series, 87 (Radioactive decay), 128
Decay curve, Radioactive, *129*
Deceleration, 11 (Acceleration)
Decibels (dB), 42 (Loudness)
Deciduous, 236
Deciduous dentition, 284 (I)
Deciduous forest (biome), 232
Deciduous teeth, 284 (I)
Declination, 73
Decomposers, 234, *235*
Decomposition, 162 (Dissociation)
  Double, 155
Decomposition reactions, 332 (Catabolism)
Decussate, 250
Defecation, 295 (Large intestine)
Deflection, Full-scale, 77 (Ammeter)
Deflection system (oscilloscope), *81*
Deformation,
  Elastic, *23*
  Plastic, 23 (Yield point)
Dehiscent, 260
Dehydrating agent, 344
Dehydration, 154
Deionized, *207*
Deliquescent, 206
Delivery tube, 223
Delocalization, 134, 137 (Giant metallic lattice)
Demagnetization, 71
Demersal, 342
Dendrites, 304
Dendron, 304 (Dendrites)
Denitrifying bacteria, 209, 235
Dens, see Dentes
Density (ρ), 24-25, 97, 112, 113, 212-213
  Magnetic flux, 72
  Optical, 50 (Snell's law)
  Relative, 24
  Relative vapor, 143
  Surface, of charge, 57
Density bottle, 24
Dentes (sing. dens),
  (teeth), 284 (I)
  (vertebrae), 279
Denticles, 266
Dentine, 284
Dentition, 284 (I)
  Deciduous, 284 (I)
  Permanent, 284 (I)
Deoxygenated, 290 (Atria)
Deoxyribonucleic acid, see DNA
Deoxyribose, 324
Dependent variable, *98*
Depolarizing agent, 68 (Polarization)
Depolymerization, 200
Derived quantities, 97
Derived SI units, 97
Dermal papillae (sing. papilla), *310*
Dermis, 310, *311*
Desalination, 207
Desert (biome), 232
Desiccation, 221
Desiccator, 221 (Desiccation), 223
Destructive distillation of coal, 179 (Coal)
Destructive interference, 39
Detergent(s), 202-203
  Biodegradable, 203
  Enzyme, see Biological laundry powders
  Non-biodegradable, 203 (Biodegradable detergents)
  Soapless/Synthetic, 203
Detergent molecule, 202, *203*
Detritus feeders, 342
Deuterium (D), *93*, 94 (Fusion reactor), 167
Deuterium oxide (D_2O), 167 (Deuterium)
Deviation, Angle of, *51*
Dextrin, 338 (Saliva)
Dextrose, see Glucose
Diamagnetism, 72
Diamines, 195
Diamond(s), *137*, 178, 212
  Synthetic, 178 (Diamond)
Diaphragm,
  (camera), *54*
  (human), 298, *299*
Diaphysis, *281*
Diastases, see Amylases
Diastema (pl. diastemata), 270
Diastole phase, 291 (Cardiac cycle)
Diatomic, *124*
Diatonic scale, *43*
Dibasic, *153*
Dicarboxylic acids, 195
Dicotyledons, 261 (Cotyledon), 340
Dictyosome, see Golgi complex
Dielectric, 59 (Capacitor)
Dielectric constant, 59
Diencephalon, 303
Diesel oil, 199
Differentiation, 320 (Embryo)
Diffraction, 37
Diffraction pattern, 136 (X-ray crystallography)
Diffuse reflection, 47
Diffusion, 5, 123, 329
  Graham's law of, 5, 143
Diffusion cloud chamber, *90*
Digestion, 271, 294 (I)
Digestive enzymes, 333 (Enzymes), 338
Digestive glands, 296, 338
Digestive juices, 296 (Digestive glands), 338
Digestive system, 294-295
Digital, 111 (Logic gates)
Digital thermometer, 27
Digitigrade, 269
Digits, *278*
Dilute, 144
Dimer, 125
Dimerization, 181 (Nitrogen dioxide)
Dinitrogen oxide (N_2O), *125*, 181
Dinitrogen tetraoxide (N_2O_4), *125*, *181*
Diode, 65, 110
  Light emitting (LED), 65, 110
Dioecious, 256 (I)
Diol, *197*
Diopters, 53 (Power)
Dip, see Inclination
Dipeptides, *205*, 339
Diploid number, 240 (Mitosis), 324 (Chromosomes)
Dipoles, 71 (Domain theory of magnetism)
Direct current (d.c.), 61
Direct replacement, 155
Direct synthesis, 155
Disaccharides, 204 (Sucrose
Discharge, 66 (Ionic theory o electrolysis), 156 (Ionic th electrolysis)
  Preferential, 66 (Ionic theor electrolysis)
Discharge tube, 80
Disintegration, 87 (Radioacti decay), 128
Disk(s),
  Invertebral, 279 (Vertebral
  Optic, see Blind spot
Disk florets, 259
Dispersal (seeds), 260
Dispersion, 54 (Color)
Displacement (chemistry), 15
Displacement (physics), 10, 3
Displacement of air,
  Downward, 217 (Hydrogen
  Upward, 216 (Chlorine), 217 (Oxygen)
Displayed formula, 140
Disproportionation, 148 (Re
Dissemination, see Dispersal
Dissociation, 162
  Thermal, 162 (Dissociation
Dissolve, 144 (I)
Distal, 301
Distal convoluted tubule, 301 (Uriniferous tubules
Distillate, *220*
Distillation, 220
  Destructive, 179 (Coal)
  Fractional, 220
  Fractional (of liquid air), 180 (Nitrogen), 183 (Oxyg 189 (I), 208 (I)
  Fractional (of petroleum/
    Primary, 198
Distilled water, 207
Divalent,
  (covalency), *133*
  (electrovalency), *131*
Diverging lens, 53
Diverging mirror, see Convex
Divided touch (magnets), 71
Divider, Potential/Voltage, 63
Divisions (classification), 340
DNA, 238 (Nucleus), 324 (Nu
  Recombinant, *327* (Gene cl
  Target, 327 (Gene cloning
Dog teeth, see Canines
Dolomite ($CaCO_3.MgCO_3$), 170 (Magnesium)
Domains, 71 (Domain theory magnetism)
Domain theory of magnetism,
Dominance, 325 (Genes), 32
  Incomplete, 325
Doping, 65
Doppler effect, 41
Dormancy, 237
Dorsal, 269

Direct synthesis, 155
Disaccharides, 204 (Sucrose), 339
Discharge, 66 (Ionic theory of electrolysis), 156 (Ionic theory of electrolysis)
  Preferential, 66 (Ionic theory of electrolysis)
Discharge tube, 80
Disintegration, 87 (Radioactive decay), 128
Disk(s),
  Invertebral, 279 (Vertebral structure)
  Optic, see **Blind spot**
Disk florets, 259
Dispersal (seeds), **260**
Dispersion, 54 (Color)
Displacement (chemistry), **158**
Displacement (physics), **10**, *34*
Displacement of air,
  Downward, 217 (Hydrogen)
  Upward, 216 (Chlorine), 217 **(Oxygen)**
Displayed formula, 140
Disproportionation, 148 (Redox)
Dissemination, see **Dispersal**
Dissociation, 162
  Thermal, 162 (Dissociation)
Dissolve, 144 (I)
Distal, 301
Distal convoluted tubule, 301 (Uriniferous tubules)
Distillate, *220*
Distillation, 220
  Destructive, 179 (Coal)
  Fractional, 220
  Fractional (of liquid air), 180 **(Nitrogen), 183 (Oxygen),** 189 **(I),** 208 **(I)**
  Fractional (of petroleum)/**Primary, 198**
Distilled water, 207
Divalent,
  (covalency), *133*
  (electrovalency), *131*
Diverging lens, 53
Diverging mirror, see **Convex mirror**
Divided touch (magnets), **71**
Divider, Potential/Voltage, 63
Divisions (classification), **340 (I)**
DNA, 238 **(Nucleus), 324 (Nucleic acids)**
  Recombinant, *327* **(Gene cloning)**
  Target, 327 **(Gene cloning)**
Dog teeth, see **Canines**
Dolomite ($CaCO_3.MgCO_3$), **170 (Magnesium)**
Domains, 71 (Domain theory of magnetism)
Domain theory of magnetism, 71
Dominance, 325 (Genes), 326
  Incomplete, 325
Doping, 65
Doppler effect, 41
Dormancy, 237
Dorsal, 269
Dorsal fin, 269 (Median fins)
Dorsal root, 302 (Spinal cord)
Dose (radiation), **89 (Dosimeter)**
Dosimeter, 89
Double bond, 132, 214 **(b** and **2)**
Double decomposition, 155
Double helix, 324 (Nucleic acids)
Double pole switch, *64*
Double salt, 154
Down feathers, 267
Downs' cell, *168*
Downward displacement of air, 217 (Hydrogen)
Drag, 19 (Terminal velocity)
Dropping pipette, 225
Drupe, 262
Dry cell, 69
Drying agent, 221 **(Desiccation), 344**
D-T reaction, *93*, **94 (Fusion reactor)**
Duct(s), 296 (Exocrine glands)
  Cochlear, 314, *315*
  Collecting, 301
  Common bile, *296*, **297 (Gall bladder)**
  Common hepatic, 297 (Liver)
  Cystic, *296*, **297 (Gall bladder)**
  Lachrymal, 313 (Lachrymal glands)
  Left hepatic, *296*
  Nasolachrymal, 313 (Lachrymal glands)
  Pancreatic, 296 (Pancreas)
  Right hepatic, *296*
  Right lymphatic, *292*, **293 (Lymph vessels)**
  Semicircular, *314*, **315**
  Sperm, *316*
  Stensen's, *296*
  Sweat, 311 (Sweat glands)
  Thoracic, *292*, **293 (Lymph vessels)**
  Wharton's, *296*
Ductile, 23 (Yield point), 344
Ductility, 344 (Ductile)
Ductless glands, see **Endocrine glands**
Duct of Wirsung, see **Pancreatic duct**
Duodenum, *294*, **295 (Small intestine)**, *296*
Dura mater, *303* **(Meninges)**
Dynamic equilibrium, 14
Dynamics, 12-13
Dynamo, see **Generator**
Dynamo rule, see **Fleming's right-hand rule**
Dysprosium (Dy), 165, 212

E

Ear(s), 314-315
  External, see **Outer ear**
  Inner/Internal, 314, 315
  Middle, 314
  Outer, 314
Ear canal, 314 (Outer ear)
Eardrum, *314*
Early wood, see **Spring wood**
Ear ossicles, 314 (Middle ear)
Earphone, 75
Ecdysis, *277* **(Larva)**
Echinoderms, 265 **(Coelom), 341**
Echo, 41
Echolocation, 41 (Echo)
Echo-sounding, 41 (Echo)
Eclipse, 46
  Annular, *46*
  Lunar, 46 (Eclipse)
  Partial, *46*
  Solar, 46 (Eclipse)
  Total, *46*
Ecological niche, 233
Ecological succession, 233
Ecology, 232 (I)
Ecosystem, 233, 234 **(I)**
Ectoplasm, *268*
Edaphic factors, 232 (I)
Eddy current, 79
Effectors, 305 **(Motor neurons), 308 (I)**
Efferent, 308
Efferent arteriole, 301 (Glomerulus)
Efferent neurons, see **Motor neurons**
Efferent system, 308-309
  Somatic, 308
  Visceral, see **Autonomic nervous**
Efficiency, 20
Efflorescent, 206
Effort, 20 (I)
Egg(s), 276
  Cleidoic, 276 (Eggs)
Egg cell, 258 **(Ovules), 320 (Gametes), 323 (Gamete production, female)**
Einsteinium (Es), 165
Einstein's mass-energy formula, *84*
Ejaculation, *316* **(Sperm duct), 319 (Copulation)**
Elastic, 22 (Elasticity)
Elasticity, 22
Elastic limit, 22
Elastic potential energy, 8
Elastin, 280 (Connective tissue)
Electrical protection, see **Sacrificial protection**
Electric bell, 75 (Electric buzzer)
Electric buzzer, 75
Electric charge (Q), 56-59, 97
Electric current (I), 56 **(I), 60-64,** 96 **(Ampere), 159**
Electric energy, 9
Electric field, 58 (I), 105, 107
Electric field lines, 107
Electric force, 6, 105, 107
Electricity, 56-69, 74-79
  Static, 56-57
Electricity supply, 61
Electric meters, 77
Electric motor, 76
Electric wind, 57 (Point action)
Electrochemical cell, see **Cell** (electrochemistry)

Electrochemical equivalent, 67 (Faraday's laws of electrolysis)
Electrochemical series, 159
Electrode, 66, 156
  Active, 66 (Electrode), 156
  Hydrogen, *158*
  Inert, 66 (Electrode), 156
  Platinum, *158*
Electrodeposition, see Electroplating
Electrode potential (E), 158
Electrolysis, 66-67, 156-157
  Faraday's laws of (First, Second), 67, 157
  Ionic theory of, 66, 156
Electrolyte, 66, 156
Electrolytic capacitor, 59, 110
Electrolytic cell, 66, 156
Electromagnetic energy, 9 (Electric energy)
Electromagnetic force, 6
Electromagnetic induction, 78-79
Electromagnetic potential energy, 8
Electromagnetic spectrum, *44*, 113
Electromagnetic waves, 34 (I), 44-45
Electromagnetism, 74-76
Electromagnets, 75-76
Electromotive force (e.m.f.), 60, 159
Electron(s), 80-81, 83, 86 (Beta particles), 87 (Beta decay), 88-90, 113, 126, 128 (Beta particle)
  Valency, 130
Electron configuration, 127
Electronegativity, 133
Electron gun, 80
Electron multiplier, *90*
Electron pairs, 132 (I)
Electron shells, 83, 126
Electron tube, 80 (I)
Electron volt (eV), 84
Electrophorus, 57
Electroplating, 67, 157
Electropositive, 133 (Electronegativity)
Electro-refining, 67, 157
Electroscope, 56
  Condensing, 56 (Electroscope)
  Gold leaf, 56 (Electroscope)
  Pulse (Wulf), 89
Electrostatic force see Electric force
Electrostatic induction, 57
Electrostatics, First law of, 56
Electrovalency, 131
Element(s), 4, 122, 164-165, 212-213
  Trace, 331 (Minerals)
Eluent, 221 (Chromatography)
Elytra (sing. elytron), 266
Embryo, 260 (I), 276, 319 (Fertilization), 320
Embryo sac, 258 (Ovules), 323 (Gamete production, female)
E.m.f., see Electromotive force
Emitter, 65 (Transistor)
Empirical formula, 140
Emulsification, 338 (Bile)
Emulsion, 145
Enamel (teeth), 284
Enantiotropy, 136
Endocardium, 290 (I)
Endocarp, 262 (I)
Endocrine glands, 297
Endodermis, 243
Endolymph, 314 (Inner ear)
Endometrium, 317 (Uterus), 318 (Menstrual cycle)
Endoneurium, *306*
Endoplasm, *268*
Endoplasmic reticulum, 239
Endopteryogotes, 277 (Complete metamorphosis)
Endoscopes, *51*
Endoskeleton, 266 (I)
Endosperm, 258 (Fertilization), 261
Endothelium, 288 (I)
Endothermic reaction, 146
End point, 222 (Volumetric analysis)
Energy, 8-9, 97, 146-147, 208-209, 330, 332-335
  Activation, 160
  Atomic, 84-85
  Binding (B.E.), 84
  Bond, 146
  Chemical, 8, *9*
  Elastic potential, 8
  Electric, 9
  Electromagnetic, 9 (Electric energy)
  Electromagnetic potential, 8
  Gravitational potential, 8
  Heat, 9, 28-32
  Internal, 9
  Kinetic (K.E.), 9
  Law of conservation of, 8 (I), 146
  Magnetic, 9
  Mechanical, 9
  Molecular potential, 8
  Nuclear, 84-85
  Nuclear potential, 8
  Potential (P.E.), 8, 106-107
  Radiant heat, 29 (Radiation)
  Rotational kinetic, 9 (Kinetic energy)
  Strain, see Elastic potential energy
  Thermal, see Internal energy
  Translational kinetic, 9 (Kinetic energy)
  Vibrational kinetic, 9 (Kinetic energy)
  Wave, 9
Energy cycle, 147 (Hess's law)
Energy level,
  (atoms), 83 (Electron shells)
  (ecology), see Trophic level
Energy level diagram, 146
Energy state, 84 (I)
Engine,
  Jet, 13
  Rocket, 13
Enteric canal, see Alimentary canal
Enterocrinin, 336
Enterokinase, 338 (Intestinal juice)
Enthalpy (H), 146
Enthalpy change of combustion, 147
Enthalpy change of formation, 147
Enthalpy change of fusion,
  Molar, 146 (Enthalpy change of reaction), 147
Enthalpy change of neutralization, 147
Enthalpy change of reaction (ΔH), 146
  Standard ($\Delta H^{\ominus}$), 146
Enthalpy change of solution, 147
Enthalpy change of vaporization,
  Molar, 146 (Enthalpy change of reaction), 147
Entire (leaves), 250
Environment, 232 (I)
  Internal, 335 (Homeostasis)
Environmental chemistry, 206-210
Enzyme(s), 161, 205, 239 (Lysosomes), 296 (Digestive glands), 333, 338
  Digestive, 333 (Enzymes), 338
  Respiratory, 333 (Enzymes)
Enzyme detergents, see Biological laundry powders
Ephemeral, 236
Epicarp, 262 (I)
Epidermis,
  (humans), 310, *311*
  (plants), 243
Epididymis (pl. epididymides), *316*
Epigeal, 261
Epiglottis, 294 (Pharynx), *298*
Epigynous flower, 257
Epimysium, *283*
Epinephrin, see Adrenalin
Epineurium, *306*
Epiphytes, 342
Epithelium, 310 (Epidermis)
Equation(s), 98, 140 (I), 141
  Balanced, 141
  General gas/Ideal gas, 33, 142
  Ionic, 141
  Nuclear, *92*, *93*, *128*
  Symbol, *98*
  Word, *98*, 141
Equation of state, see Ideal gas equation
Equations of uniformly accelerated motion, 11
Equator, Magnetic, *73*
Equilibrium, 15, 162
  Chemical, 163
  Dynamic, 15 (Equilibrium)
  Linear, 15
  Neutral, 15
  Rotational, 15
  Stable, 15
  Static, 15 (Equilibrium)
  Thermal, 28 (I)
  Unstable, 15
Equilibrium position, 163
Equipotential, 58
ER, see Endoplasmic reticulum
  Rough, 239 (Endoplasmic reticulum)
  Smooth, 239 (Endoplasmic reticulum)
Erbium (Er), 165, 212
Erectile tissue, 316 (Penis), 317 (Vulva)
Error(s), 102-103
  Parallax, 102
  Reading, 103
  Zero, 102
Erythrocytes, see Red blood c
Escape velocity, 19
Esophagus, 294, *295*
Esterification reaction, 197 (Condensation reactio
Esters, 195, 197 (Condensatio reaction)
Estrogen, 336, 337 (FSH, LH E
  (CH_3CHO), *215* (5)
Ethane (C_2H_6), 192, *193*, *214*
Ethane-1,2-diol, *197*
Ethanedioic acid ($(COOH)_2$),
Ethanoic acid (CH_3COOH), *1*
  195, *197*, *215* (7)
Ethanol (C_2H_5OH), *141*, *191*,
  214 (a), 216 (Ethene), *222*
Ethene (C_2H_4), *140*, *190*, 193,
  200, 201 (Polythene), *214*
Ether extraction, 221 (Solvent extraction)
Ethyl acetate, see Ethyl ethano
Ethyl alcohol, see Ethanol
Ethylene, see Ethene
Ethylene glycol, see Ethane-1
Ethyl ethanoate (CH_3COOH
  195, *197*
Ethyne (C_2H_2), 194, *214* (3)
Eureka can, 24
Europium (Eu), 164, 212
Eustachian tube, *314*
Eutrophication, 210
Evaporation basin, 223
Evaporation, 30, 121
Evergreen, 236
Excretion, 273, 300 (I)
Exhalant siphon, 272 (Sipho
Exhalation, see Expiration
Exocrine glands, 296
Exodermis, 245 (Piliferous la
Exopteryogotes, 277 (Incompl metamorphosis)
Exoskeleton, 266 (I)
Exothermic reaction, 146
Expansion, 32
  Anomalous, 32
Expansivity, 32 (I)
  Absolute cubic, see Real cu expansivity
  Apparent cubic, 32
  Real cubic, 32
Expiration, 299
Exponential notation, 109
Extensors, 282 (Skeletal musc
External auditory canal, see Ea
External beam radiotherapy,
External ear, see Outer ear
External fertilization, 276 (O
External gills, 272
External jugular veins, *290*
External respiration, 298 (I)

Error(s), **102-103**
  Parallax, **102**
  Reading, **103**
  Zero, **102**
Erythrocytes, see **Red blood cells**
Escape velocity, **19**
Esophagus, **294**, *295*
Esterification reaction, **197** (**Condensation reaction**)
Esters, **195**, 197 (**Condensation reaction**)
Estrogen, **336**, 337 (**FSH**, **LH**)**Ethanal** (**CH_3CHO**), *215* (5)
Ethane (C_2H_6), **192**, *193*, *214* (1)
Ethane-1,2-diol, *197*
Ethanedioic acid ($(COOH)_2$), *150*, *195*
Ethanoic acid (CH_3COOH), *150*, *151*, *195*, *197*, *215* (7)
Ethanol (C_2H_5OH), *141*, *191*, **196**, *197*, *214* (a), 216 (**Ethene**), *222*
Ethene (C_2H_4), *140*, *190*, **193**, *198*, *200*, **201** (**Polythene**), *214* (2), **216**
Ether extraction, **221** (**Solvent extraction**)
Ethyl acetate, see **Ethyl ethanoate**
Ethyl alcohol, see **Ethanol**
Ethylene, see **Ethene**
Ethylene glycol, see **Ethane-1,2-diol**
Ethyl ethanoate ($CH_3COOH_2CH_3$), *195*, *197*
Ethyne (C_2H_2), **194**, *214* (3)
Eureka can, **24**
Europium (**Eu**), **164**, **212**
Eustachian tube, *314*
Eutrophication, **210**
Evaporation basin, **223**
Evaporation, **30**, **121**
Evergreen, **236**
Excretion, **273**, **300** (**I**)
Exhalant siphon, **272** (**Siphon**)
Exhalation, see **Expiration**
Exocrine glands, **296**
Exodermis, **245** (**Piliferous layer**)
Exopterygotes, **277** (**Incomplete metamorphosis**)
Exoskeleton, **266** (**I**)
Exothermic reaction, **146**
Expansion, **32**
  Anomalous, **32**
Expansivity, **32** (**I**)
  Absolute cubic, see **Real cubic expansivity**
  Apparent cubic, **32**
  Real cubic, **32**
Expiration, **299**
Exponential notation, **109**
Extensors, **282** (**Skeletal muscles**)
External auditory canal, see **Ear canal**
External beam radiotherapy, *91*
External ear, see **Outer ear**
External fertilization, **276** (**Oviparous**)
External gills, **272**
External jugular veins, *290*
External respiration, **298** (**I**)
External urinary sphincter, **300** (**Bladder**)
Extrinsic eye muscles, **313**
Eye(s), **312-313**
  Compound, **275**
Eyeball, **312** (**I**)
Eye muscles,
  Extrinsic, **313**
  Intrinsic, *312* (**Ciliary body**)
Eyepiece lens,
  (microscope), *54*
  (telescope), *55*
Eye teeth, see **Canines**

F

Face-centered cubic (particle arrangement), *137*
Facet,
  (bones), *279*
  (compound eyes), *275*
Facial bones, **278** (**Cranium**)
Fahrenheit scale, **27**
Fallopian tubes, *317*, 319 (**Fertilization**)
False fruit, **262** (**I**)
False ribs, **278** (**Rib cage**)
Families, **340** (**I**)
Farad (F), **59**, 97
Faraday (F), **157**
Faraday's constant, **113**
Faraday's iron ring, **78** (**Mutual induction**)
Faraday's law of induction, **78**
Faraday's laws of electrolysis (**First**, **Second**), **67**, **157**
Fascicles,
  (muscles), **283** (**Striated muscle**)
  (nerves), **306**
Fast reactor/Fast breeder reactor (**FBR**), **95**
Fats, 204 (**I**), **205** (**Lipids**), 293 (**Lymph**), **295** (**Small intestine**), **330**, 331, 338
Fatty acids, **330** (**Fats**), 338 (**I**, **Pancreatic juice**)
Feather(s), **267**
  Contour, **267** (**Feathers**)
  Down, *267*
  Flight, see **Remiges**
  Primary, **267** (**Remiges**)
  Secondary, **267** (**Remiges**)
Feather follicles, **267**
Feces, **295** (**Large intestine**)
Feedback, **335** (**Homeostasis**)
  Negative, **335** (**Homeostasis**)
Feedstocks, Chemical, **199**
Femoral arteries, *289*
Femoral veins, *289*
Femur, *278*, *280*, *281*
Fenestra ovalis, see **Oval window**
Fenestra rotunda, see **Round window**
Fermentation, Alcoholic, **197**
Fermium (**Fm**), **165**
Ferric compounds, see **Iron(III) compounds**
Ferromagnetic, **70**
Ferrous compounds, see **Iron(II) compounds**
Ferrous sulfate, see **Iron(II) sulfate**
Fertilization, **320** (**Sexual reproduction**)
  (animals), **276**
  External, **276** (**Oviparous**)
  (flowering plants), **258**
  (humans), **319**
  Internal, **276** (**Viviparous**)
Fertilizers, *182*, 210 (**Eutrophication**)
Fetus, **319** (**Pregnancy**)
Fiber(s),
  Carbon, **179**
  Muscle, **283** (**Striated muscle**)
  Nerve, **304**
  Optical, *51*
  (plants), **243** (**Xylem**)
  (roughage), 204 (**I**), **331**
Fibrils, **283** (**Striated muscle**)
Fibrin, **287** (**Clotting**)
Fibrinogen, **287** (**Clotting**)
Fibrous roots, **245**
Fibula, *278*, *281*
Field(s), **104-107**
  Electric, **58** (**I**), 105, 107
  Force, **6**, **104-107**
  Gravitational, **104**, 106
  Magnetic, **72-73**, 105, 107
Field direction, **104-105**
Field effect transistor (symbol for), **110**
Field intensity, **106-107**
Field lines, **6** (**Force field**), **106-107**
  Electric, **107**
  Gravitational, **106**
  Magnetic, **72**, 107
Field windings (electric motor), **76**
Filament(s),
  Gill, **272** (**Internal gills**)
  (light bulb), **64**
  (muscles), **283** (**Striated muscle**)
  (plants), **257** (**Stamens**)
Filament lamp, **64**
Film (camera), *54*
Film badge, see **Dosimeter**
Filter-feeding, **270**
Filter funnel, **224**
Filtering, **220**
Filter paper, **223**
Filtrate, *220*
  Glomerular, **301** (**Glomerular filtration**)
Filtration, see **Filtering**
  Glomerular, **301**
Fimbriae, *317*
Finishing, see **Fitting**
Fin(s), **269**
  Anal, **269** (**Median fins**)
  Caudal, **269** (**Median fins**)
  Dorsal, **269** (**Median fins**)
  Median, **269**
  Paired, **269**
  Pectoral, **269** (**Paired fins**)
  Pelvic, **269** (**Paired fins**)
  Ventral, **269**

- **First law of electrostatics**, 56
- **First law of magnetism**, 70 (Pole)
- **First meiotic division**, 322 (Meiosis)
- **First order consumers**, see Primary consumers
- **First order sensory neuron**, *306*, *309*
- **First polar body**, 323 (Gamete production, female)
- **Fish**, 236 (Anadromous), 266 (Scales), *269*, *272*, *275* (Lateral lines), 341
- **Fissile**, 92 (Induced fission), 129 (Nuclear fission)
- **Fission, Binary**, 240 (Cell division)
- **Fission, Nuclear**, *85*, 92-93, 94-95, 129
  - **Induced**, 92
  - **Spontaneous**, 92
- **Fission bomb**, 93
- **Fission fragments**, see Fission products
- **Fission-fusion bomb**, 93 (Fusion bomb)
- **Fission neutrons**, 92 (Nuclear fission)
- **Fission products**, 92 (Nuclear fission)
- **Fission reactor**, 94
- **Fitting**, *202* (Saponification)
- **Fixed joints**, 280 (I)
- **Fixed macrophages**, 286 (White blood cells)
- **Fixed point** (Lower, Upper), 26
- **Flagella** (sing. flagellum), 268
- **Flagellate**, 268 (Flagella)
- **Flame**, 208
  - **Luminous**, *208*
  - **Non-luminous**, *208*
- **Flame cells**, 273 (Nephridia)
- **Flame tests**, 219, 222
- **Flask(s)**, 224
  - **Buchner**, *220*, 224
  - **Conical**, 224
  - **Flat-bottomed**, 224
  - **Round-bottomed**, 224
  - **Vacuum**, 29
  - **Volumetric**, 224
- **Flat coil**, 74
- **Fleming's right-hand rule**, 78
- **Flexion**, 282 (Skeletal muscles)
- **Flexors**, 282 (Skeletal muscles)
- **Flight feathers**, see Remiges
- **Flint**, 177 (Silicon dioxide)
- **Floating ribs**, 278 (Rib cage)
- **Flocculent** (precipitate), *145*
- **Florets**, 259 (Flowerhead)
  - **Disk**, 259
  - **Ray**, 259
- **Florigen**, 251 (Photoperiodism)
- **Flotation, Principle of**, 25
- **Flower(s)**, 256-259
  - **Bell**, 259
  - **Composite**, see Flowerhead
  - **Epigynous**, 257
  - **Hypogynous**, 257
  - **Lipped**, 259
  - **Pea**, 259
  - **Perigynous**, 257
  - **Spurred**, 259
- **Flowerhead**, 259
- **Flowers of sulfur**, 184
- **Fluid**, 120
  - **Amniotic**, *276* (Amnion), *319*
  - **Cerebrospinal**, *303*
  - **Intercellular/Interstitial**, see Tissue fluid
  - **Pericardial**, 290 (I)
  - **Pleural**, 298 (Pleura)
  - **Seminal**, 319 (Copulation)
  - **Tissue**, 292
- **Fluorescence**, 45
- **Fluorescent tube**, 80 (Discharge tube)
- **Fluorides**, *186*
- **Fluorine** (F/F_2), *130*, 165, 186, 212
- **Fluoro group** (–F), *195*, 215 (9)
- **Fluorocarbons**, *186*
- **Fluorospar** (CaF_2), 186 (Fluorine)
- **Flux density, Magnetic**, 72
- **Flux lines**, see Field lines
- **Foam**, 145
- **Focal length** (f), *48*, *52*
- **Focal point**, see Principal focus
- **Foliage**, 248 (I)
- **Folic acid**, 339 (Vitamin B complex)
- **Follicle(s)**,
  - **Feather**, 267
  - **Graafian**, 317 (Ovarian follicles), 318 (Menstrual cycle)
  - **Hair**, 311
  - **Ovarian**, 317, 337 (FSH, LH, Estrogen)
- **Follicle stimulating hormone**, see FSH
- **Food(s)**, 204-205, 330-331
  - **Accessory**, 204 (I), 330 (I)
- **Food chains**, 234 (Food web)
- **Food vacuole**, *268*
- **Food web**, 234
- **Foramen** (pl. foramina),
  - **Invertebral**, *279*
  - **Vertebral**, 279 (Vertebrae)
- **Force(s)**, 6-7, 97, 104-107
  - **Centrifugal**, 17 (Centripetal)
  - **Centripetal**, 17
  - **Compression**, 7
  - **Contact**, 7
  - **Electric**, 6, 105, 107
  - **Electromagnetic**, 6
  - **Electromotive** (e.m.f.), 60, 159
  - **Electrostatic**, see Electric force
  - **Frictional**, 7
  - **Gravitational**, 6, 104, 106
  - **Intermolecular**, 7, 134 (van der Waals' forces)
  - **Kinetic frictional**, 7
  - **Limiting**, 7 (Static frictional force)
  - **Lines of**, see Field lines
  - **Lorentz**, 76, 77
  - **Magnetic**, 6, 105, 107
  - **Newton's third law**, 13
  - **Normal contact**, 7 (Coefficient of friction)
  - **Nuclear**, 7, 84
  - **Resultant**, *6*
  - **Sliding frictional**, see Kinetic frictional force
- **Force(s)** (cont'd)
  - **Static frictional**, *7*
  - **Tension**, 7
  - **Turning**, 14-15
  - **Van der Waals'**, 134
- **Force field**, 6, 104-107
- **Forced oscillation**, 17
- **Foreskin**, *316*
- **Forest** (biomes),
  - **Coniferous**, 232
  - **Deciduous**, 232
  - **Tropical**, 232
- **Formaldehyde**, see Methanal
- **Formalin**, *194*
- **Formation, Enthalpy change of/ Heat of**, 147
- **Former** (electromagnets), 74 (Coil)
- **Formic acid**, see Methanoic acid
- **Formula** (pl. formulas), 140, 226-227
  - **3-dimensional structural**, see Stereochemical formula
  - **Displayed**, 140
  - **Einstein's mass-energy**, *84*
  - **Empirical**, 140
  - **Full structural**, see Displayed formula
  - **General**, 191 (Homologous series)
  - **Lens/mirror**, 49
  - **Molecular**, 140
  - **Shortened structural**, 140
  - **Stereochemical**, 140, 191 (Stereochemistry)
- **Formula mass, Relative/Formula weight**, see Relative molecular mass
- **Forward biased**, 65 (Diode)
- **Forward reaction**, 162
- **Fossil fuels**, 208 (Fuel), 234
- **Fovea/Fovea centralis**, 313 (Macula lutea)
- **Fractional distillation**, 220
- **Fractional distillation of liquid air**, 180 (Nitrogen), 183 (Oxygen), 189 (I), 208 (I)
- **Fractional distillation of petroleum**, see Primary distillation
- **Fractionating column**, *183*, 198 (Primary distillation), 220 (Fractional distillation), 224
- **Fractions** (petroleum distillation), 198
  - **Heavy**, 198 (Fraction)
  - **Light**, 198 (Fraction)
- **Fractions** (units), 96
- **Francium** (Fr), 164, 168, 212
- **Frasch process**, 184
- **Free fall**, 19
- **Free oscillation**, see Natural oscillation
- **Freezing**, 30, 121
- **Freezing point**, 30 (Freezing), 121 (Freezing)
- **Freon** (CCl_2F_2), *195*
- **Frequency** (f), 16, 35, 97
  - **Beat**, 42 (Beats)
  - **Fundamental**, 43 (Modes of vibration)
  - **Natural**, 17 (Natural oscillation)
- **Friction**, see Frictional
  - **Coefficient of**, 7
- **Frictional force**, 7
  - **Kinetic/Sliding**, 7
  - **Static**, 7
- **Fructose**, 204 (Sucrose), 338 juice), 339 (Monosacc
- **Fruit**, 262
  - **False**, 262 (I)
  - **Key**, see Samara
  - **True**, 262 (I)
- **FSH**, 336
- **FSHRF/FSH releasing factor**, 336 (Regulating factors
- **Fuel(s)**, 208
  - **Fossil**, 208 (Fuel), 234
- **Fuel oil**, 199 (Residue)
- **Fulcrum**, 14 (I, Moment), 21
- **Fullerenes**, 179 (Buckmin
- **Full scale deflection**, 77 (Amme
- **Full structural formula**, see Dis formula
- **Full-wave rectification**, 65
- **Fume cupboard**, 224
- **Fumigation**, 344
- **Fuming sulfuric acid** (H_2S_2O 185 (Contact process)
- **Functional group**, 190
- **Fundamental frequency**, 43 (M vibration)
- **Fundamental interval**, 26 (Fixed
- **Fungi** (sing. fungus), *320*, 340
- **Fungicide**, 344
- **Funicle**, 257 (Ovaries)
- **Funnel(s)**, 224
  - **Buchner**, *220*, 224
  - **Filter**, 224
  - **Separating**, 224
  - **Tap**, 224
  - **Thistle**, 224
- **Fuse** (electric), *61*
- **Fusion** (change of state), 121
  - **Molar enthalpy change of** Mo of, 146 (Enthalpy change of reaction), 147
  - **Specific latent heat of**, 31, 11 146 (Enthalpy change of rea
- **Fusion, Nuclear**, *85*, 93, 94 (Fus reactor), 129
- **Fusion** (reproduction), 320 (Sex reproduction)
- **Fusion bomb**, 93
- **Fusion reactor**, 94

G

- **Gadolinium** (Gd), 165, 212
- **Gain** (oscilloscope control), *81*
- **Galactose**, 338 (Intestinal juice), 339 (Monosaccharides)
- **Galena** (PbS), 177 (Lead)
- **Gall bladder**, 297, 337 (CCK)

Friction, see **Frictional force**
  Coefficient of, 7
Frictional force, 7
  Kinetic/Sliding, 7
  Static, 7
Fructose, 204 (**Sucrose**), 338 (**Intestinal juice**), 339 (**Monosaccharides**)
Fruit, 262
  False, 262 (**I**)
  Key, see **Samara**
  True, 262 (**I**)
FSH, 336
FSHRF/FSH releasing factor, 336 (**Regulating factors**)
Fuel(s), **208**
  Fossil, 208 (**Fuel**), 234
Fuel oil, 199 (**Residue**)
Fulcrum, 14 (**I, Moment**), 21 (**Lever**)
Fullerenes, 179 (**Buckminsterfullerene**)
Full scale deflection, 77 (**Ammeter**)
Full structural formula, see **Displayed formula**
Full-wave rectification, 65
Fume cupboard, 224
Fumigation, 344
Fuming sulfuric acid ($H_2S_2O_7$), **185** (**Contact process**)
Functional group, 190
Fundamental frequency, 43 (**Modes of vibration**)
Fundamental interval, 26 (**Fixed point**)
Fungi (sing. **fungus**), *320*, **340** (**I**)
Fungicide, 344
Funicle, 257 (**Ovaries**)
Funnel(s), **224**
  Buchner, *220*, **224**
  Filter, 224
  Separating, 224
  Tap, 224
  Thistle, 224
Fuse (electric), *61*
Fusion (change of state), **121**
  Molar enthalpy change of/Molar heat of, 146 (**Enthalpy change of reaction**), **147**
  Specific latent heat of, 31, 112, 146 (**Enthalpy change of reaction**)
Fusion, Nuclear, *85*, **93**, 94 (**Fusion reactor**), **129**
Fusion (reproduction), **320** (**Sexual reproduction**)
Fusion bomb, 93
Fusion reactor, 94

G

Gadolinium (**Gd**), **165, 212**
Gain (oscilloscope control), *81*
Galactose, 338 (**Intestinal juice**), **339** (**Monosaccharides**)
Galena (**PbS**), **177** (**Lead**)
Gall bladder, 297, 337 (**CCK**)
Gallium (**Ga**), 112, 165, **176**, 212
Galvanizing, *159*, **174**
Galvanometer, 77, 110
  Moving coil, 77 (**Galvanometer**)
Gametes, 320, 322, 323
Gametophyte, 321 (**Alternation of generations**)
Gamma radiography (**γ-radiography**), **91**
Gamma rays (**γ-rays**), **44**, *85*, **86**, 91, *113*, **128**
Ganglia (sing. **ganglion**), *306*
  Autonomic, *309*
Gas(**es**), **5, 33, 120** (**Gaseous state**), 198-199
  Coal, 179 (**Carbon monoxide, Coal**)
  Ideal, 33, 142
  Inert/Noble/Rare, 165, **189**, *208*, 217 (**Nitrogen**)
  Laughing, see **Dinitrogen oxide**
  Liquefied petroleum (**LPG**), **199** (**Refinery gas**)
  Natural, 192 (**Methane, Ethane**), 198 (**I**), *208*
  Producer, 179 (**Carbon monoxide**)
  Refinery, 199
  Water, 179 (**Carbon monoxide**)
Gas chromatography, 221 (**Chromatography**)
Gas constant, *33*, **113**, *142*
Gas-cooled reactor, Advanced (**AGR**), *95*
Gaseous state, 5, 120
Gas equation (**General/Ideal**), **33, 142**
Gas jar, 224
Gas laws, 33, 142-143
Gas oil, see **Diesel oil**
Gasoline, 188, **199**, *208*
Gas syringe, 224
Gastric artery, *289*
Gastric glands, 296 (**Digestive glands**), 338 (**Gastric juice**)
Gastric juice, 337 (**Gastrin**), **338**
Gastric lipase, 338 (**Gastric juice**)
Gastric vein, *289*
Gastrin, 337
Gastroesophageal sphincter, see **Cardiac sphincter**
Gastrointestinal tract, see **Alimentary canal**
Gates, see **Logic gates**
Gause's principle, 233 (**Ecological niche**)
Gauze, 224
Gay-Lussac's law, 143
Gear, 21
Geiger counter, 89
Geiger-Müller tube, 89 (**Geiger counter**)
Gemmation, 321
Gene(s), **325**
  Sex-linked, 326 (**Sex linkage**)
Gene cloning, 327
Gene splicing, 327 (**Gene cloning**)
Genera, 340 (**I**)
General formula, 191 (**Homologous series**)
General gas equation, see **Ideal gas equation**
Generative nucleus, 258 (**Pollen**)
Generator, 78
  Steam, *94, 95*
  Van de Graaff, 57
Generic name, 340 (**Nomenclature**)
Genetic diagnosis, 328
Genetic engineering, 327 (**I**)
Genetics, 324 (**I**)
Genetic variation, 237
Genital organs/Genitalia, 316 (**I**)
Genome, 328 (**Genome mapping**)
Genotypes, *325*
Geo-stationary orbit, 19
Geotropism, 251
Germanium (**Ge**), 112, 165, **177**, 212
Germ cells, see **Gametes**
Germicide, 344
Germination, 260
Gestation, see **Pregnancy**
Gestation period, 319 (**Pregnancy**)
g-force, 18 (**Acceleration due to gravity**)
Giant atomic lattice, *4*, **137**
Giant ionic lattice, *131*, **137**
Giant metallic lattice, 137
Gibberellins, 251 (**Growth hormones**)
Gill(s) (**External/Internal**), **272**
Gill bar/arch, 272 (**Internal gills**)
Gill filaments, 272 (**Internal gills**)
Gill lamellae (sing. **lamella**), **272** (**Internal gills**)
Gill rakers, *272*
Gill slits, 272 (**Internal gills**)
Gingiva, see **Gum**
Gizzard, 271
Gland(s), **296-297**
  Adrenal, 297, 337
  Bulbourethral, see **Cowper's glands**
  Ceruminous, 314 (**Outer ear**)
  Cowper's, *316*
  Digestive, 296, 338
  Endocrine/Ductless, 297
  Exocrine, 296
  Gastric, 296 (**Digestive glands**), 338 (**Gastric juice**)
  Intestinal, 296 (**Digestive glands**), 338 (**Intestinal juice**)
  Lachrymal, 313
  Lymph, see **Lymph nodes**
  Mammary, *318*, 337 (**Estrogen**)
  Mucous, 295 (**Mucous membrane**)
  Parathyroid, 297, 337 (**PTH**)
  Parotid, *296*
  Pineal, 297
  Pituitary, 297, *303*, 336 (**Hormones**), 337
  Prostate, *316*
  Salivary, 296 (**Digestive glands**), 338 (**Saliva**)
  Sebaceous, 311

Gland(s) (cont'd)
  Silk, *265*
  Sublingual, *296*
  Submandibular/Submaxillary, *296*
  Sudoriferous, see **Sweat glands**
  Suprarenal, see **Adrenal glands**
  Sweat, **311**
  Tear, see **Lachrymal glands**
  Thymus, **293**
  Thyroid, **297**, 337 (**TSH**, **Thyroxin**, **TCT**)
  Uropygial, *267* (**Uropygium**)
Glans, *316*
Glia, see **Neuroglia**
Gliding joints, **280**
Glomerular filtrate, **301** (**Glomerular filtration**)
Glomerular filtration, **301**
Glomerulus, **301**
Glottis, **298** (**Larynx**)
Glucagon, **336**
Glucose ($C_6H_{12}O_6$), 197 (**Alcoholic fermentation**), **204**, 209, **330** (**Carbohydrates**), 331, 334 (**Anaerobic respiration**), 338 (**Intestinal juice**), 339 (**Monosaccharides**), *344*
Glycerine/Glycerol,
  (biology), **330** (**Fats**), 338 (**I**, **Pancreatic juice**)
  (chemistry), see **Propane-1,2,3-triol**
Glycine, *205*
Glycogen, **204** (**Glucose**), **331**, 338 (**Saliva**), 339 (**Polysaccharides**)
Glycolysis, **334** (**Anaerobic respiration**)
Gold (**Au**), 112, 165, **173**, *175*, 212
Gold leaf electroscope, **56** (**Electroscope**)
Golgi complex/apparatus/body, **239**
Gonadal arteries, *289*
Gonadal veins, *289*
Gonads, **316** (**I**)
Graafian follicle, **317** (**Ovarian follicles**), 318 (**Menstrual cycle**)
Gradient,
  (graphs), *98*
  Potential, **58** (**Potential**)
  Temperature, **28** (**Conductivity**)
Graduations, **345**
Grafting, **263**
Graham's law of diffusion, **5**, **143**
Grain, **262**
Granular layer, see **Stratum granulosum**
Graphite, 95 (**Thermal reactor**), **178**, 212
Graphs, **98-99**
Grassland (biome), **232**
  Temperate, **232**
Gravimetric analysis, **222**
Gravitation, **18-19**
  Newton's law of, **18**
Gravitational constant (**G**), **18** (**Newton's law of gravitation**), 113
Gravitational field, **104**, 106
Gravitational field lines, **106**
Gravitational force, **6**, 104, 106
Gravitational potential energy, **8**
Gravity, see **Gravitational force**
  Acceleration due to (g), **18**, 113
  Center of, **10** (**Center of mass**)
  Specific, see **Relative density**
Gray matter, **302** (**Neuroglia**)
Great saphenous veins, *289*
Greenhouse effect, **29**, **210**
Green vitriol, see **Iron(II) sulfate**
Gristle, see **Cartilage**
Ground (wires), *61*, 110
Ground state, **84** (**I**)
Group 0 (elements), see **Group VIII**
Group I (elements), 164, **168-169**
Group II (elements), 164, **170-171**
Group III (elements), 165, **176**
Group IV (elements), 165, **177-179**
Group V (elements), 165, **180-182**
Group VI (elements), 165, **183-185**
Group VII (elements), 165, **186-188**
Group VIII (elements), 165, **189**
Growing point (roots), **245**
Growth hormone(s),
  Human, see **STH**
  (plants), **251**
Growth regulators, see **Growth hormones** (plants)
Grub, *277* (**Larva**)
Guanine, **324**
Guard cells, **249**
Gullet, see **Esophagus**
Gum, **284** (**I**)
Gustatory area, **Primary**, **303**
Gustatory pore, *307*
Gustatory sensations, **307** (**Tongue**)
Gut, see **Alimentary canal**
Guttation, **253**
Gynaecium, **257**
Gypsum ($CaSO_4.2H_2O$), **171** (**Calcium sulfate**)

H

Haber process, **180**
Habitat, **233**
Hafnium (**Hf**), 165, **173**, 212
Hair(s), 268 (**Cilia**), 274 (**Vibrissae**), **310-311**, *325*, *326*
  Olfactory, **307** (**Nose**)
  Pubic, *318*
  Root, **245**
Hair erector muscles, **311**
Hair follicles, **311**
Hair plexuses, **311**
Half cell, **158**
Half-life (T ½), **87**, **129**
Half-wave rectification, **65**
Halides, **186** (**I**)
  Alkyl, see **Halogenoalkanes**
Halogenoalkanes, **195**, **215** (**9**)
Halogens, 165, **186-188**, **214** (**I**), **215** (**9**)
Halophytes, **342**
Haltères, **275**
Hammer, see **Malleus**
Haploid number, **322** (**Meiosis**), 326 (**I**)
Haptotropism, **251**
Hard (magnets), **70**
Hard palate, *307*
Hard water, **207**
Harmonics, *43*
Haustra (sing. haustrum), *294*
Haversian canals, **281** (**Compact bone**)
H-bomb, see **Hydrogen bomb**
Heart, *288*, *289*, **290-291**
Heartwood, 243 (**Xylem**), **247**
Heat, see **Heat energy**
  Latent (**L**), **30**, *31*, 146 (**Enthalpy change of reaction**)
  Specific latent (**l**), **30**
Heat capacity (**C**), **31**
  Specific (c), **31**, 112
Heat energy, **9**, 28-33
  Radiant, **29** (**Radiation**)
Heat-losing center, **335** (**Homeostasis**)
Heat of combustion, see **Enthalpy change of combustion**
Heat of formation, see **Enthalpy change of formation**
Heat of fusion,
  Molar, see **Molar enthalpy change of fusion**
  Specific latent, **31**, 112, 146 (**Enthalpy change of reaction**)
Heat of neutralization, see **Enthalpy change of neutralization**
Heat of reaction, see **Enthalpy change of reaction**
Heat of solution, see **Enthalpy change of solution**
Heat of vaporization,
  Molar, see **Molar enthalpy change of vaporization**
  Specific latent, **31**, 146 (**Enthalpy change of reaction**)
Heat-promoting center, 335 (**Homeostasis**)
Heat transfer, **28-31**
Heavy (precipitate), *145*
Heavy fractions, **198** (**Fraction**)
Heavy water, see **Deuterium oxide**
Heliotropism, **251** (**Phototropism**)
Helium (**He**), *93*, 128 (**Alpha particle**), 165, **189**, 212
Hematite (Fe_2O_3), 174 (**Iron**)
Hemocoel, **265**
Hemoglobin, **286** (**Red blood cells**)
Hemopoiesis, **286** (**I**)
Henle, **Loop of**, **301** (**Uriniferous tubules**)
Hepatic artery, *289*
Hepatic duct(s),
  Common, *296*, **297** (**Liver**)
  Left, *296*
  Right, *296*
Hepatic portal vein, *289*, **297** (**Liver**)
Hepatic vein, *289*
Heptane (C_7H_{16}), *198*
Herbaceous, **236**
Herbaceous perennials, 236 (**Perennials**)
Herbivores, **235** (**Primary**
Hermaphrodite, **256** (**I**), 277 (**Spermatheca**)
Hertz (**Hz**), **16** (**Frequency**
Hess's law, **147**
Heterogeneous, **123**
Heterogeneous catalyst, **16**
Heterografting, **263** (**Grafting**
Heterozygous, *325*
Hexagonal (basic crystal sha
Hexagonal (particle arrange
Hexane (C_6H_{14}), **192**
Hexanedioic acid (**COOH(CH$_2$)$_4$COOH**
HGH, see **STH**
Hibernation, **237** (**Dormancy**
Higher animal, **264** (**I**)
Hilum,
  (lungs), **299** (**Bronchi**)
  (seeds), **261**
Hinge joints, **280**
Hip girdle, see **Pelvis** (skeleto
Histones, **324** (**Chromosomes**
Hoffman voltameter, **67**
Holmium (**Ho**), **165**, **212**
Homeostasis, 303 (**Hypothalam** 336 (**Hormones**)
Hominids, **341**
Homiothermic, **335** (**Homeosta**
Homo, **341**
  erectus, **341**
  habilis, **341**
  sapiens, **341**
Homogeneous, **123**
Homogeneous catalyst, **161**
Homografting, **263** (**Grafting**
Homologous chromosomes, 240 (**Mitosis**), 322, 323, 324 (**Chromosomes**), 326 (**L**… **of segregation**)
Homologous series, **191**, **194**
Homopolymer, **200**
Homozygous, *325*
Honey guides, **256** (**Nectaries**
Hooke's law, **22**
Hoppers, 277 (**Metamorphosis**
Hormone(s), 297 (**Endocrine glands**), **336-337**
  Adrenocorticotropic, see **AC**
  Antagonistic, **336**
  Anti-diuretic, see **ADH**
  Follicle-stimulating, see **FSH**
  Growth (plants), **251**
  Human growth, see **STH**
  Interstitial cell stimulation, se
  Lactogenic, **336**
  Luteinizing, see **LH**
  Parathyroid, see **PTH**
  Sex, 297 (**Pineal gland**), 336 (**E**
    Progesterone, **Androgens**

Heptane (C_7H_{16}), *198*
Herbaceous, 236
Herbaceous perennials, 236 (Perennials)
Herbivores, 235 (Primary consumers)
Hermaphrodite, 256 (I), 277 (Spermatheca)
Hertz (Hz), 16 (Frequency), 97
Hess's law, 147
Heterogeneous, 123
Heterogeneous catalyst, 161
Heterografting, 263 (Grafting)
Heterozygous, *325*
Hexagonal (basic crystal shape), *136*
Hexagonal (particle arrangement), *137*
Hexane (C_6H_{14}), 192
Hexanedioic acid ($COOH(CH_2)_4COOH$), *200*
HGH, see **STH**
Hibernation, 237 (Dormancy)
Higher animal, 264 (I)
Hilum,
  (lungs), **299 (Bronchi)**
  (seeds), **261**
Hinge joints, 280
Hip girdle, see **Pelvis** (skeleton)
Histones, 324 (Chromosomes)
Hoffman voltameter, 67
Holmium (Ho), 165, 212
Homeostasis, 303 **(Hypothalamus), 335,** 336 **(Hormones)**
Hominids, 341
Homiothermic, 335 (Homeostasis)
Homo, 341
  erectus, 341
  habilis, 341
  sapiens, 341
Homogeneous, 123
Homogeneous catalyst, 161
Homografting, 263 (Grafting)
Homologous chromosomes, 240 **(Mitosis)**, 322, 323, **324 (Chromosomes), 326 (Law of segregation)**
Homologous series, 191, 194-195
Homopolymer, 200
Homozygous, *325*
Honey guides, 256 (Nectaries)
Hooke's law, 22
Hoppers, 277 **(Metamorphosis)**
Hormone(s), 297 **(Endocrine glands), 336-337**
  Adrenocorticotropic, see **ACTH**
  Antagonistic, 336
  Anti-diuretic, see **ADH**
  Follicle-stimulating, see **FSH**
  Growth (plants), **251**
  Human growth, see **STH**
  Interstitial cell stimulation, see **LH**
  Lactogenic, 336
  Luteinizing, see **LH**
  Parathyroid, see **PTH**
  Sex, 297 **(Pineal gland), 336 (Estrogen, Progesterone, Androgens)**
Hormone(s) (cont'd)
  Somatotropic, see **STH**
  Thyroid-stimulating, see **TSH**
  Tropic, 297 (Pituitary gland)
Horny layer, see **Stratum corneum**
Hosts, 342 (Parasites)
Human growth hormone, see **STH**
Humans, 278-319, 341
Humerus, *278, 282*
Humidity, 206
Humor (Aqueous, Vitreous), *312*
Hydathodes, 253 (Guttation)
Hydrate, 154
Hydrated, 154 (Hydrate)
Hydrated iron(III) oxide, see **Rust**
Hydration, 144 (Solvation)
Hydraulic press, 21
Hydrides, *149* **(rule 5), 167 (I)**
Hydrocarbons, 191
Hydrocarbon waxes, 199
Hydrochloric acid (HCl), *143, 154, 170, 171*, **187 (Hydrogen chloride)**, 216 **(Carbon dioxide, Chlorine)**, 217 **(Hydrogen)**, 218 **(Anions)**, 222 **(Flame test)**, 338 **(Gastric juice)**
Hydrocortisone, 336
Hydrogen (H/H_2), 94 **(Fusion reactor)**, *132, 133*, 134 **(Hydrogen bond)**, 150, 151, 152, 153, 155, 164, **167**, *170*, 180 **(Haber process)**, 211, 212, 214-215, 217, 218 **(Gases)**, 330
Hydrogenation, 193
Hydrogen bomb (H-bomb), see **Fusion bomb**
Hydrogen bond, 134
Hydrogen bromide (HBr), 188
Hydrogencarbonates, see **Bicarbonates**
Hydrogen chloride (HCl), *130, 133, 143, 146, 148, 150, 162, 181*, **187**, *192*, 216 **(Chlorine)**
Hydrogen iodide (HI), 188
Hydrogen peroxide (H_2O_2), 167, 217 **(Oxygen)**
Hydrogen sulfide (H_2S), *148*, **185**
Hydroiodic acid, 188 (Hydrogen iodide)
Hydrolysis, 154
Hydrometer, 24
Hydronium ion (H_3O^+), 150, 152, 167
Hydrophilic, 202 (Detergent molecule)
Hydrophobic, 202 (Detergent molecule)
Hydrophytes, 342
Hydrostatic skeleton, 265 (Body cavities)
Hydrotropism, 251
Hydroxides, 151, 152, 155, **167**, 211
Hydroxyl group (–OH), 191 **(Homologous series), 196 (I),** 214 **(d and 4)**
Hygroscopic, 206
Hymen, *317*
Hypertonic, *329*
Hypogeal, 260
Hypogynous flower, 257
Hyponome, 272 (Siphon), *274*
Hypopharynx, *271*
Hypophysis, see **Pituitary gland**
Hypothalamus, 303, 335 **(Homeostasis),** 336 **(Hormones),** 337 **(Oxytocin, ADH)**
Hypotonic, *329*

I

Ice, *120, 147*, **206**
Ice point, 26 (Fixed point)
ICSH, see **LH**
Ideal gas, 33, 142
Ideal gas equation, 33, 142
Ignition tube, 225
Ileum, *294*, **295 (Small intestine)**
Iliac arteries, Common, *289*
Iliac veins, Common, *289*
Ilium, *278* **(Pelvis)**
Image(s), 49
  Mosaic, 275 (Compound eyes)
  Real, 49 (Image)
  Thermal, 45 (Infra-red radiation)
  Virtual, 49 (Image)
Imago, *277*
Immiscible, 145 (Miscible)
Impacted, 285 (Wisdom teeth)
Impedance, 64
Implantation, 319 (Fertilization)
Implants, Radioactive, *91*
Impulse, 12, *97*
Impure, 123 (Pure)
Impurity, 123 (Pure)
Incandescence, 45 (Visible light)
Incidence,
  Angle of (i), *47, 50*
  Point of, *47, 50*
Incident ray, *47, 50*
Incident wave, 36
Incident wavefronts, 36 (Incident wave)
Incisors, *284*, **285**
  Central, *285*
  Lateral, *285*
Inclination, 73
Inclined plane, 21
Incomplete dominance (genes), **325**
Incomplete metamorphosis, 277
Incus, 314 (Middle ear)
Indehiscent, 260
Independent assortment, Law of, 326
Independent variable, *98*
Index,
  Absolute refractive, 37 (Refractive index), *50*
  Refractive (n), 37, 50 **(Snell's law)**, *51*
  (thermometer device), **27 (Maximum and minimum thermometers)**
Indicator, 152
  Universal, 152

Indium (In), 165, 176, 212
Induced fission, 92
Induced magnetism, 71 (Magnetization)
Inductance, 64
Induction,
  Electromagnetic, 78-79
  Electrostatic, 57
  Faraday's law of, 78
  (magnetic), 71 (Magnetization)
  Mutual, 78
Inductor, 64 (Inductance), 110
Inert, 345
Inert electrode, 66 (Electrode), 156
Inert gases, see Noble gases
Inertia, 12
Inferior (ovary), 257 (Epigynous flower)
Inferior articular processes, *279*
Inferior mesenteric artery, *289*
Inferior mesenteric vein, *289*
Inferior vena cava, *289*, *290*, 291
Infertile, 328 (Human cloning)
Inflorescence, 259
Infra-red radiation (IR radiation), 45, *113*
Infrasonic waves, 40 (Infrasound)
Infrasound, 40
Infundibulum, *317*
Ingestion, 294 (I)
Inhalant siphon, *265* (Mantle siphon), 272 (Siphon)
Inhalation, see Inspiration
Inheritance, 324 (I)
Inhibiting factors, 336 (Regulating factors)
Inhibitor, 161
Initial rate (of reaction), 160
Inner ear, 314, 315
Inner transition series, 164-165, 172 (I)
Inorganic chemistry, 166, 167-189
In phase (waves), 38 (Phase)
Input (computers), 111
Insectivores, 342
Insects, 266, 271, 273, 274, 275, 277
Insoluble, 145 (Soluble)
Inspiration, 299
Instantaneous rate (of reaction), 160
Instantaneous speed, 10 (Speed)
Instantaneous velocity, 10 (Velocity)
Insulator(s),
  (electricity), 56
  (heat), 28
Insulin, 327 (Protein manufacture), 335 (Homeostasis), 336
Integrated circuits, 111 (Logic gates)
Integumentary system, 310 (I)
Integuments, 258 (Ovule), 261 (Testa)
Intensity,
  Field, 106-107
  Luminous, *96*
  Wave, 35
Intercellular fluid, see Tissue fluid
Intercostal muscles, 299 (Inspiration)
Interference, Wave, 38-39
  Constructive, 39
  Destructive, 39
Interference fringes, 38 (Young's slits)
Interference pattern, 38 (I), *39*
Intermolecular forces, 7, 134 (van der Waals' forces)
Internal ear, see Inner ear
Internal energy, 9
Internal environment, 335 (Homeostasis)
Internal fertilization, 276 (Viviparous)
Internal gills, 272
Internal jugular veins, *290*
Internal reflection, 51 (Total internal reflection)
Internal resistance (r), 63
Internal respiration, 209, 254 (Photosynthesis), 255 (Compensation points), 298 (I), 334
Internal urinary sphincter, 300 (Bladder)
International system units (SI units), 96-97
Interneurons, see Association neurons
Internode, 244
Internuncial neurons, see Association neurons
Interphase, 240, 323 (Telophase)
Interstitial cells, *316*, 337 (Androgens)
Interstitial cell stimulating hormone, see LH
Interstitial fluid, see Tissue fluid
Interval(s),
  Fundamental, 26 (Fixed point)
  (sound), 43 (Musical scale)
Intervillous spaces, *319*
Intestinal glands, 296 (Digestive glands), 338 (Intestinal juice)
Intestinal juice, 337 (Enterocrinin), 338
Intestine (Large, Small), *294*, 295
Intrinsic eye muscles, *312* (Ciliary body)
Inversely proportional, 345
Inversion, Lateral, 47 (Plane mirror)
Invertase, see Sucrase
Invertebral disks, 279 (Vertebrae)
Invertebral foramen, *279*
Invertebrates, 274 (Setae), 275 (Statocysts), 277 (Spermatheca, Metamorphosis), 341
Inverter (symbol for), 110
Involuntary actions, 309
Involuntary muscles, 282 (I), 282 (Cardiac muscle, Visceral muscles)
Iodides, 186 (I), 188, 218 (Anions)
Iodine (I/I_2), *137*, 165, 186, 188, 212, *215* (9)
  Tincture of, *188*
Ion(s), 88 (Ionization), 130
  Complex, 154 (Complex salt)
  Hydronium/Oxonium (H_3O^+), 150
  Spectator, 141
Ion exchange, 207
Ionic bonding, 131
Ionic bonds, 131 (I)
Ionic compound, *4*, 131
Ionic equation, 141
Ionic lattice, Giant, *131*, 137
Ionic theory of electrolysis, 66, 156
Ionization, 88, 130
Ionosphere, *45*
Iridium (Ir), 112, 165, 173, 212
Iris, *312*
Iron (Fe), 112, *149*, 165, 172, 174, 211, 212, 219 (Cations)
  Cast, 112, 174 (Iron)
  Pig, 174 (Iron)
  Wrought, 112, 174 (Iron)
Iron(II) chloride ($FeCl_2$), 174 (Iron(II) compounds)
Iron(II) compounds, 174, 219 (Cations)
Iron(II) sulfate ($FeSO_4$), *141*, *149*, 218 (Nitrate anion)
Iron(II) sulfide (FeS), *155*, 184 (I)
Iron(II) tetraoxosulfate(VI), see Iron(II) sulfate
Iron(III) chloride ($FeCl_3$), 122 (Synthesis), *154*, 174 (Iron(III) compounds)
Iron(III) compounds, 174, 219 (Cations)
Iron(III) hydroxide ($Fe(OH)_3$), *154*
Iron(III) oxide, see Hematite
  Hydrated, see Rust
IR radiation, see Infra-red radiation
Irradiation, 91, 129
Irritability, see Sensitivity
Ischium, *278* (Pelvis)
Islets of Langerhans, 297 (Pancreas)
Isoclinal line, 73
Isogonal lines, 73
Isomerism, 133
Isomers, 133 (Isomerism), 190
  Structural, 191
Isomorphism, 136
Isotonic, *329*
Isotope(s), 83, 127, 222 (Mass spectroscopy)
  Radioactive, see Radioisotope
Isotopic mass, Relative, 138
Isotopic ratio, 138
Ivory, see Dentine

J

Jar(s) (apparatus),
  Gas, 224
  Leyden, 59
Jejunum, *294*, 295 (Small intestine)
Jet engine, 13
Joints, 280-281
  Ball-and-socket, 280
  Cartilaginous, 281 (Cartilage)
  Fixed, 280 (I)
  Gliding, 280
  Hinge, 280
  Pivot, 279 (Vertebral structure)
  Plane/Sliding, see Gliding joints
  Synovial, 280 (Synovial sac)
Joule (J), 8 (I), 97
Jugular veins (External, Internal), *290*
Juice,
  Digestive, 296 (Digestive glands), 338
  Gastric, 337 (Gastrin), 338
  Intestinal, 337 (Enterocrinin
  Pancreatic, 297 (Pancreas), 336 (Secretin, PZ), 338

K

Karyokinesis, 240 (Cell division
Karyolymph, see Nucleoplasm
K.E., see Kinetic energy
Keel,
  (animals), 269 (Pectoralis mu
  (plants), 259 (Pea flower)
Keepers (magnets), 71 (Self-demagnetization)
Kelvins (K), 27 (Absolute tempe scale), 96, 143 (Absolute temperature scale)
Keratin, 267 (Feathers), 310 corneum), 311 (Hair follicle
Kernel, see Grain
Kerosene, 199
Ketones, 194, *215* (6)
Key fruits, see Samara
Kidneys, *288*, *289*, 300
Kilogram (kg), 96
Kilojoule (kJ), *146*, 333
Kinematics, 10 (I)
Kinetic energy (K.E.), 9
Kinetic friction, Coefficient of, 7 (Coefficient of friction
Kinetic frictional force, 7
Kinetic theory, 5, 123
Kingdoms, 340 (I)
  Animal, 340, 341
  Plant, 340, 341
Kirchhoff's laws, 63
Kneecap, see Patella
Krill, 270 (Filter-feeding)
Krypton (Kr), 165, 189, 212
Krypton fluoride (KrF_2), 189
K-shell, 83 (Electron shells)

L

Labeling, 91 (Radioactive tracing), 129
Labia (sing. labium),
  (humans), 317 (Vulva)
  (insects), *271*
Labor, 319 (Pregnancy)
Labrum, *271*
Lachrymal canals, 313 (Lachryma glands)
Lachrymal ducts, 313 (Lachryma glands)
Lachrymal glands, 313

- **Juice,**
  - **Digestive**, 296 (**Digestive glands**), **338**
  - **Gastric**, 337 (**Gastrin**), **338**
  - **Intestinal**, 337 (**Enterocrinin**), **338**
  - **Pancreatic**, 297 (**Pancreas**), 336 (**Secretin, PZ**), **338**

K

- **Karyokinesis**, **240** (**Cell division**)
- **Karyolymph**, see **Nucleoplasm**
- **K.E.**, see **Kinetic energy**
- **Keel**,
  - (animals), **269** (**Pectoralis muscles**)
  - (plants), **259** (**Pea flower**)
- **Keepers** (magnets), **71** (**Self-demagnetization**)
- **Kelvins** (**K**), **27** (**Absolute temperature scale**), **96**, **143** (**Absolute temperature scale**)
- **Keratin**, **267** (**Feathers**), **310** (**Stratum corneum**), **311** (**Hair follicles**)
- **Kernel**, see **Grain**
- **Kerosene**, **199**
- **Ketones**, **194**, *215* (**6**)
- **Key fruits**, see **Samara**
- **Kidneys**, *288*, *289*, **300**
- **Kilogram** (**kg**), **96**
- **Kilojoule** (**kJ**), *146*, **333**
- **Kinematics**, **10** (**I**)
- **Kinetic energy** (**K.E.**), **9**
- **Kinetic friction, Coefficient of**, **7** (**Coefficient of friction**)
- **Kinetic frictional force**, **7**
- **Kinetic theory**, **5**, **123**
- **Kingdoms**, **340** (**I**)
  - **Animal**, 340, **341**
  - **Plant**, **340**, 341
- **Kirchhoff's laws**, **63**
- **Kneecap**, see **Patella**
- **Krill**, **270** (**Filter-feeding**)
- **Krypton** (**Kr**), 165, **189**, 212
- **Krypton fluoride** (KrF_2), **189** (**Krypton**)
- **K-shell**, **83** (**Electron shells**)

L

- **Labeling**, **91** (**Radioactive tracing**), **129**
- **Labia** (sing. **labium**),
  - (humans), **317** (**Vulva**)
  - (insects), *271*
- **Labor**, **319** (**Pregnancy**)
- **Labrum**, *271*
- **Lachrymal canals**, **313** (**Lachrymal glands**)
- **Lachrymal ducts**, **313** (**Lachrymal glands**)
- **Lachrymal glands**, **313**
- **Lactase**, **338** (**Intestinal juice**)
- **Lacteals**, **293** (**Lymph vessels**), **295** (**Small intestine**)
- **Lactic acid**, **334** (**Oxygen debt**)
- **Lactogenic hormone**, **336**
- **Lactose**, **338** (**Intestinal juice**), 339 (**Disaccharides**)
- **Lacunae** (sing. **lacuna**), **281** (**Bone**)
- **Lamellae** (sing. **lamella**),
  - (bone), **281** (**Compact bone**)
  - **Gill**, **272** (**Internal gills**)
- **Lamina** (leaves), **248** (**I**)
- **Langerhans, Islets of**, **297** (**Pancreas**)
- **Lanthanides**, **173** (**Lanthanum**)
- **Lanthanum** (**La**), *92*, 164, **173**, 212
- **Lanthanum oxide** (La_2O_3), **173** (**Lanthanum**)
- **Large intestine**, *294*, **295**
- **Larva** (pl. **larvae**), *277*
- **Larynx**, 275 (**Syrinx**), **298**
- **Latent heat** (**L**), **30**, *31*, 146 (**Enthalpy change of reaction**)
  - **Specific** (**l**), **30**
- **Latent heat of fusion, Specific**, **31**, 112, 146 (**Enthalpy change of reaction**)
- **Latent heat of vaporization, Specific**, **31**, 146 (**Enthalpy change of reaction**)
- **Lateral**, **244**
- **Lateral bud**, see **Axillary bud**
- **Lateral incisors**, *285*
- **Lateral inversion**, **47** (**Plane mirror**)
- **Lateral lines**, **275**
- **Lateral roots**, **245** (**Tap root**)
- **Late wood**, see **Summer wood**
- **Latex**, *201*, **345**
- **Lattice(s), Crystal**, **136** (**I**), **137**
  - **Giant atomic**, *4*, **137**
  - **Giant ionic**, *131*, **137**
  - **Giant metallic**, **137**
  - **Molecular**, **137**
- **Laughing gas**, see **Dinitrogen oxide**
- **Laundry powders**, **203**
  - **Biological**, **203** (**Laundry powders**)
- **Law(s)**,
  - **Avogadro's**, **143**
  - **Boyle's**, **33**, **142**
  - (chemistry, basic), **125**
  - **Dalton's**, **143**
  - **Faraday's** (electrolysis), **67**, **157**
  - **Faraday's** (induction), **78**
  - **Gas**, **33**, **142-143**
  - **Gay-Lussac's**, **143**
  - **Graham's**, **5**, **143**
  - **Hess's**, **147**
  - **Hooke's**, **22**
  - **Kirchhoff's**, **63**
  - **Lenz's**, **78**
  - **Mendel's**, **326**
  - **Newton's**, **12-13**, **18**
  - **Ohm's**, **62**
  - **Pressure**, **33**, **142**
  - **Snell's**, **50**
  - **Third gas**, see **Pressure law**
- **Law of conservation of energy**, **8** (**I**), **146**
- **Law of conservation of linear momentum**, **13**
- **Law of conservation of mass**, **125**
- **Law of constant composition**, **125**
- **Law of diffusion, Graham's**, **5**, **143**
- **Law of electrostatics, First**, **56**
- **Law of gravitation, Newton's**, **18**
- **Law of independent assortment**, **326**
- **Law of induction, Faraday's**, **78**
- **Law of magnetism, First**, **70** (**Pole**)
- **Law of multiple proportions**, **125**
- **Law of partial pressures, Dalton's**, **143**
- **Law of segregation**, **326**
- **Law of volumes**, **33**, **142**
- **Laws of electrolysis, Faraday's**, **67**, **157**
- **Laws of motion, Newton's**, **12-13**
- **Laws of reflection of light**, **47**
- **Laws of refraction of light**, **50**
- **Lawrencium** (**Lr**, also **Lw**), **165**
- **Layer(s)** (plant tissue),
  - **Abscission**, **249**
  - **Palisade**, **248**
  - **Piliferous**, **245**
  - **Spongy**, **248**
- **Layer(s)** (skin),
  - **Cornified**, see **Stratum corneum**
  - **Granular**, see **Stratum granulosum**
  - **Horny**, see **Stratum corneum**
  - **Malpighian**, see **Stratum basale**
  - **Subcutaneous**, **310**, *311*
- **Lead** (**Pb**), 112, 165, **177**, *178*, 210, 211, 212, 219 (**Flame tests, Cations**), 222 (**Gravimetric analysis**)
- **Lead(II) compounds**, **177** (**Lead**), 219 (**Cations**)
- **Lead(II) ethanoate**, **218** (**Sulfide anion**)
- **Lead(II) hydroxide** ($Pb(OH)_2$), 167 (**Hydroxide**)
- **Lead(II) iodide** (PbI_2), *155*
- **Lead(II) nitrate** ($Pb(NO_3)_2$), *155*
- **Lead(II) oxide** (**PbO**), **177** (**Lead**)
- **Lead(II) sulfide**, see **Galena**
- **Lead(IV) compounds**, **177** (**Lead**)
- **Lead(IV) oxide** (PbO_2), **177** (**Lead**)
- **Lead-acid accumulator**, **69**, *150*, *159*
- **Leaflets**, 248 (**I**), **250**
- **Leaf scar**, **249** (**Abscission layer**)
- **Leaf trace**, **249**
- **Leaves**, **248-250**
  - **Compound**, 248 (**I**), **250**
  - **Scale**, **263** (**Bulb**)
  - **Simple**, **248** (**I**)
- **Le Chatelier's principle**, **163**
- **Lecithin**, **339** (**Vitamin B complex**)
- **Leclanché cell**, **69**
- **LED**, see **Light emitting diode**
- **Left-hand rule, Fleming's**, **76**
- **Legume**, **262**
- **Leguminous plants**, **235** (**Nitrogen fixation**)
- **Length** (**l**), *96*
  - **Focal** (**l**) *48*, *52*
  - (measurement of), **100**

Lens(es), 52 (I)
  Achromatic, *54* (Spectrometer), 55 (Chromatic aberration)
  Concave, *52*, 53
  Converging, 52
  Convex, *52*, *53*
  Diverging, 53
  (eye), 312
  Eyepiece, *54*, *55*
  Objective, *54*, *55*
  Projection, *55*
Lens assembly (camera), *54*
Lens formula, see Mirror formula
Lenticels, 247
Lenz's law, 78
Leslie's cube, 29
Leucocytes, see White blood cells
Leucoplasts, 240 (Plastids)
Lever, 21
Leyden jar, 59
LH, 336
LHRF/LH releasing factor, 336 (Regulating factors)
Lichens, 342 (Mutualists)
Lieberkühn, Crypts of, see Intestinal glands
Liebig condenser, *220*, 223
Life cycle, 236 (I)
Lifting magnet, 75
Ligament(s), 280
  Ovarian, *317*
  Periodontal, 284 (Root)
  Suspensory, 312 (Lens)
Light, 46-55
  Laws of reflection of, 47
  Laws of refraction of, 50
  Principle of reversibility of, 49, *50*
  Reflection of, 47-49
  Speed of, 113
  Visible, 45, 54-55 (Color), *113*
Light dependent resistor (symbol for), 110
Light emitting diode (LED), 65, 110
Light fractions (petroleum), 198 (Fraction)
Lightning, 57
Lightning conductor, 57 (Lightning)
Lignin, 243 (Vessels)
Lignite, 179 (Coal)
Limestone, 171 (Calcium carbonate), *174*
Limewater, 171 (Calcium hydroxide), 218 (Gases)
Limit,
  Elastic, 22
  Proportional, see Limit of proportionality
Limiting force, 7 (Static frictional force)
Limit of proportionality, 22 (Hooke's law), *23*
Linear (molecule), *133*
Linear acceleration, 14 (I)
Linear coefficient of expansion (α), 32, 112
Linear equilibrium, 15
Linear magnification, 49
Linear momentum, Law of conservation of, 13
Linear motion, 10
Lines of flux/force, see Field lines
Lingual, 307
Lingual tonsil, 293 (Tonsils)
Lipase(s), 338 (I)
  Gastric, 338 (Gastric juice)
  Pancreatic, 338 (Pancreatic juice)
Lipids, 205
Lipped flower, 259
Liquefaction, 121
Liquefied petroleum gas (LPG), 199 (Refinery gas)
Liquid air, Fractional distillation of, 180 (Nitrogen), 183 (Oxygen), 189 (I), 208 (I)
Liquid-in-glass thermometer, 26
Liquid nitrogen, 180 (Nitrogen)
Liquid state, 5, 120
Liter (l), 101
Lithium (Li), 164, 168, *180*, 212, 219 (Flame tests)
Lithium chloride (LiCl), 168 (Lithium)
Lithium hydroxide (LiOH), *168*
Lithium nitride (Li_3N), *180*
Lithophytes, 342
Litmus, 152
Littoral, 342
Liver, *289*, 294 (I), 297
Load, 20 (I)
  Useless, 20
Lobe(s),
  Anterior (pituitary gland), 297
  (leaves), 250 (Lobed)
  (liver), *296*, *297*
  Posterior (pituitary gland), 297
Lobed (leaves), 250
Local action, 68
Locomotion, 268 (I)
Locus, *325*
Logic gates, 111
Lone pair (electrons), 133
Longitudinal waves, 35
Long-night plants, 251 (Photoperiodism)
Long periods (element groups), 164 (Period)
Loop of Henlé, 301 (Uriniferous tubules)
Lorentz force, 76, *77*
Loudness, 42
Loudspeaker, 76, 110
Lower animal, 264 (I)
Lower fixed point, 26 (Fixed point)
Lower motor neuron, *308*, *309*
L-shell, 83 (Electron shells)
Lubricating oil, 199
Lumbar, *278* (Lumbar vertebrae)
Lumbar vertebrae, *278*
Luminous flame, *208*
Luminous intensity, *96*
Lunar eclipse, 46 (Eclipse)
Lung books, 273
Lungs, *289*, 298
Luteinizing hormone/Luteotropin, see LH
Lutetium (Lu), 165, 212
Lymph, 293
Lymphatic(s), 293 (Lymph vessels)
Lymphatic duct, Right, *292*, 293 (Lymph vessels)
Lymphatic organs, see Lymphoid organs
Lymphatic system, 293
Lymphatic tissue, see Lymphoid tissue
Lymphatic vessels, see Lymph vessels
Lymph capillaries, 293 (Lymph vessels)
Lymph glands, see Lymph nodes
Lymph nodes, 293
Lymphocytes, 286 (White blood cells), 293 (Lymphoid organs)
Lymphoid organs, 293
Lymphoid tissue, 293 (Lymphoid organs)
Lymph vessels, 293
Lysosomes, 239

M

Machine(s), 20-21
  Perfect, 20
Macrohabitat, 232 (Biomes)
Macromolecule, 125
Macrophages (Fixed, Wandering), 286 (White blood cells)
Macula (pl. maculae), 315 (Saccule)
Macula lutea, 313
Maggot, *277* (Larva)
Maglev train, 75
Magnesium (Mg), 112, *130*, *148*, *149*, 164, 170, 207, 211, 213, 219 (Cations)
Magnesium carbonate ($MgCO_3$), *155*
Magnesium chloride ($MgCl_2$), *155*, 170 (Magnesium), *171*
Magnesium hydroxide ($Mg(OH)_2$), *151*, 171
Magnesium oxide (MgO), *149*, *170*, 171
Magnesium stearate, *207*
Magnesium sulfate ($MgSO_4$), *150*, 171
Magnet(s), 70-79
  Lifting, 75
  Molecular, see Dipoles
  Permanent, 70 (Hard)
  Temporary, 70 (Soft)
Magnetic, 70 (I)
Magnetic axis, 70
Magnetic energy, 9
Magnetic equator, *73*
Magnetic field(s), 72-73, 105, 107
  Radial, 76 (Loudspeaker)
Magnetic field lines/flux lines, 72, 107
Magnetic flux density, 72
Magnetic force, 6, 105, 107
Magnetic meridian, 73
Magnetic north, 73
Magnetic south, 73
Magnetism, 70-79
  Domain theory of, 71
  First law of, 70 (Pole)
  Induced, 71 (Magnetization
  Residual, 70 (Soft)
Magnetization, 70 (I), 71
Magnetron, *45*
Magnification,
  Angular, 55 (Visual angle
  Linear, 49
Magnifying glass, see Simple microscope
Magnifying power, see Angular magnification
Magnitude, Order of, 109
Main chain (atoms), *190*
Malachite ($CuCO_3.Cu(OH)_2$), 175 (Copper)
Male nuclei, 258 (I, Pollen), 320 (Gametes), 323 (Gamet production, male)
Malleable, Malleability, 345
Malleus, 314 (Middle ear)
Malpighian corpuscles, see Renal corpuscles
Malpighian layer, see Stratum
Malpighian tubules, 265 (Hemocoel), 273
Maltase, 338 (Intestinal juice
Maltese cross tube, 80
Maltose, 338 (Pancreatic juice
Mammals, 269 (Pectoralis mus 271 (Rumen), 274 (Vibrissae
Mammary glands, *318*, 337 (Es
Mandible(s),
  (arthropods), *271*
  (birds), *267*
  (humans), *278*
Manganese (Mn), 165, 172, 21
Manganese(IV) chloride (MnCl
Manganese(IV) oxide (MnO_2), 216 (Chlorine), 217 (Oxyge
Man-made polymers, see Synth polymers
Manometer, 25
Mantle,
  (birds), *267*
  (snails, etc.), 265 (Mantle cavi
Mantle cavity, 265
Mantle siphon, *265*
Maquis, 232
Marble, 171 (Calcium carbona 216 (Carbon dioxide)
Margin (leaves), *248*, 250
Marrow, Bone 281
  Red, 281 (Bone marrow)
  Yellow, 281 (Bone marrow)
Marrow cavities, see Medullary cavities

Magnetic flux density, 72
Magnetic force, 6, 105, 107
Magnetic meridian, 73
Magnetic north, 73
Magnetic south, 73
Magnetism, 70-79
  Domain theory of, 71
  First law of, 70 (Pole)
  Induced, 71 (Magnetization)
  Residual, 70 (Soft)
Magnetization, 70 (**I**), **71**
Magnetron, *45*
Magnification,
  Angular, 55 (Visual angle)
  Linear, 49
Magnifying glass, see **Simple microscope**
Magnifying power, see **Angular magnification**
Magnitude, Order of, 109
Main chain (atoms), *190*
Malachite ($CuCO_3.Cu(OH)_2$), **175 (Copper)**
Male nuclei, 258 (**I, Pollen**), **320 (Gametes)**, 323 (**Gamete production, male**)
Malleable, Malleability, 345
Malleus, 314 (Middle ear)
Malpighian corpuscles, see **Renal corpuscles**
Malpighian layer, see **Stratum basale**
Malpighian tubules, 265 (**Hemocoel**), **273**
Maltase, 338 (Intestinal juice)
Maltese cross tube, 80
Maltose, 338 (Pancreatic juice)
Mammals, 269 (**Pectoralis muscles**), 271 (**Rumen**), 274 (**Vibrissae**), **341**
Mammary glands, *318*, 337 (**Estrogen**)
Mandible(s),
  (arthropods), *271*
  (birds), *267*
  (humans), *278*
Manganese (Mn), 165, **172**, 213
Manganese(IV) chloride ($MnCl_2$), *216*
Manganese(IV) oxide (MnO_2), **216 (Chlorine)**, **217** (**Oxygen**)
Man-made polymers, see **Synthetic polymers**
Manometer, 25
Mantle,
  (birds), *267*
  (snails, etc.), **265 (Mantle cavity)**
Mantle cavity, 265
Mantle siphon, *265*
Maquis, 232
Marble, 171 (Calcium carbonate), 216 (**Carbon dioxide**)
Margin (leaves), *248*, 250
Marrow, Bone 281
  Red, 281 (Bone marrow)
  Yellow, 281 (Bone marrow)
Marrow cavities, see **Medullary cavities**
Mass (m), **12**, 18 (**Weight**), 96
  Atomic, see **Relative atomic mass**
  Center of, 10
  Critical, 93
  Law of conservation of, 125
  Molar, 139
  Relative atomic, 83, 138, 164-165, 212-213,
  Relative formula, see **Relative molecular mass**
  Relative isotopic, 138
  Relative molecular, 138
  Subcritical, 93 (Critical mass)
Mass concentration, 139 (Concentration)
Mass defect, 84
Mass-energy formula, Einstein's, *84*
Mass number (A), **82, 127**
Mass spectrometer, 138 (**I**), **222 (Mass spectroscopy)**
Mass spectroscopy, 222
Matrix, 280 (Connective tissue)
Maxillae (sing. **maxilla**),
  (arthropods), *271*
  (humans), **278 (Cranium)**
Maximum thermometer, 27
Maxwell's screw rule, 74
Mean, 345
Mean position, 16, *34*
Measurements, 100-101
Measuring cylinder, 101, **224**
Mechanical advantage (M.A.), **20**
Mechanical energy, 9
Mechanical waves, 34 (**I**)
Median/Medial, 269
Median fins, 269
Medium (pl. **media**), **345**
Medulla,
  (adrenal glands), **297**
  (brain), **303**
  (kidneys), *300*
  (plants), see **Pith**
  Renal, see **Medulla** (kidneys)
Medulla oblongata, see **Medulla** (brain)
Medullary cavities, 281 (Bone marrow)
Medullary pyramids, see **Renal pyramids**
Megaelectron volt (MeV), **84**
Meiosis, 322
Meiotic division,
  First, 322
  Second, 322 (**Meiosis**), **323**
Meissner's corpuscles, 311
Melanin, 310
Melting, 30, 120
Melting point, 30 (Melting), *31*, **120** (**Melting**), 212, 213
Melting point test, 221
Membrane(s),
  Cell, 238, *239*, 329
  Mucous, 295
  Nuclear, 238 (Nucleus)
  Plasma, see **Cell membrane**
  Pleural, see **Pleura**
Membrane(s) (cont'd)
  Synovial, 280 (Synovial sac)
  Tectorial, 314 (Cochlear duct)
  Tympanic, see **Eardrum**
Membranous labyrinth, 314 (Inner ear)
Memory (computers), **111**
  Random access (RAM), 111 (Memory)
  Read only (ROM), 111 (Memory)
Menaquinone, see **Vitamin K**
Mendelevium (Md), 165
Mendel's laws, 326
Meninges (sing. **meninx**), *303*
Meniscus, 345
  Concave, 53 (Concave lens)
  Convex, 53 (Convex lens)
Menopause, 318 (Menstrual cycle)
Menstrual cycle, 318, 337 (**FSH, Estrogen**)
Menstruation, 318 (Menstrual cycle)
Mercury (Hg), 112, 113, 165, **173**, *210*, 213
Meridian, Magnetic, 73
Meristem, 244
  Apical, 244 (Meristem)
Mesencephalon, see **Midbrain**
Mesenteric arteries (Inferior, Superior), *289*
Mesenteric veins (Inferior, Superior), *289*
Mesenteries, 294 (Alimentary canal)
Mesocarp, 262 (I)
Mesophyll, 248 (Spongy layer)
Mesophytes, 342
Messenger RNA, 239 (Ribosomes)
Metabolic rate, 332
  Basal, 332 (Metabolic rate)
Metabolism, 332
Metacarpals, *278*
Metacarpus, *278* (**Metacarpals**)
Metal(s), **165**
  Alkali, 165, **168-169**
  Alkaline-earth, 165, **170-171**
  Transition, 165, **172-175**
Metal extraction, 67, 157
Metallic bonding, 134
Metallic lattice, Giant, 137
Metalloids, 165 (Metal)
Metameres, 264 (Segmentation)
Metameric segmentation/Metemerism, 264 (Segmentation)
Metamorphosis (Complete, Incomplete), 277
Metaphase,
  (meiosis), **322**
  (mitosis), **241**
Metatarsals, *278*
Metatarsus, *278* (**Metatarsals**)
Meter (m), 96
  Newton (Nm), 14 (Moment)
Meter bridge, 63 (Wheatstone bridge)
Meter(s) (electric), **77**
  Moving iron, 77
Methanal (HCHO), *194*
Methane (CH_4), *125*, *133*, *190*, *191*, **192**, 198 (**I**), *208*, *214* (**1**)

Methanoic acid ($HCOOH$), *150*, *195*, *215* (7)
Methanol (CH_3OH), *191*, 196, *214* (4a)
Methoxymethane (CH_3OCH_3), *191*
Methyl amine (CH_3NH_2) *195*
Methyl group ($-CH_3$), *215* (8)
Methyl methacrylate, *201*
Methyl orange, *152*
Methyl-tertiary-butyl-ether, 199 (Octane rating)
Micelle, 202
Microchip, 111 (Computers)
Microhabitat, 233 (Habitat)
Micrometer screw gauge, 101
Microphone (symbol for), 110
Micropyle, 258 (Ovules), *261*
Microscope, 54
  Compound, 54 (Microscope)
  Simple, 54 (Microscope)
Microtubules, 240 (Centrioles)
  Spindle (cell division) 241 (Prophase), 322 (Prophase)
Microwave ovens, 45 (Microwaves)
Microwaves, 45, *113*
Micturition, see Urination
Midbrain, 303
Middle ear, 314
Midrib (leaves), 248 (Veins)
Midriff, see Diaphragm
Migration, 237
Milk teeth, see Deciduous teeth
Milky (precipitate), *145*
Milliammeter (symbol for), 110
Mimic, 237 (Mimicry)
Mimicry, 237
Mineral acid, 150
Minerals, 330 (I), 331
Minimum thermometer, 27
Mirror(s),
  Concave/Converging, 48
  Convex, 48
  Curved, 48 (I)
  Diverging, see Convex mirror
  Plane, 47
Mirror formula, 49
Miscible, 145
Mist, 145
Mitochondria, *239*, 240, 334 (Aerobic respiration)
Mitosis, 240
Mitral valve, 291 (Atrioventricular valves)
Mixed nerves, 306 (Nerves)
Mixing, Color, 55
  Additive, 55 (Color mixing)
  Subtractive, 55 (Color mixing)
Mixture, 4 (Compound), 122
Model, 237 (Mimicry)
Moderator, 95 (Thermal reactor)
Modes of vibration, 43
Molars, 270 (Carnassial teeth), 285
Molar enthalpy change of fusion/Molar heat of fusion, 146 (Enthalpy change of reaction), 147
Molar enthalpy change of vaporization/Molar heat of vaporization, 146 (Enthalpy change of reaction), 147
Molarity, 139
Molar mass, 139
Molar solution, 139
Molar volume, 139
Mole (mol), 96, 139
Molecular formula, 140
Molecular lattice, 137
Molecular magnets, see Dipoles
Molecular mass, Relative, 138
Molecular potential energy, 8
Molecular properties, 22-23
Molecular weight, see Relative molecular mass
Molecule(s), 4, 5, 124, 125, 132-134, 138 (Relative molecular mass), 140, 142
  Detergent, 202, 203
  Non-polar, 133 (Polar molecule)
  Polar, 133
Mollusks, 265 (Hemocoel, Mantle cavity), 270 (Radula), 272 (Internal gills), 273 (Nephridia), 274 (Tentacles), 341
Molten, 120
Molybdenum (Mo), 112, 165, 173, 213
Moment(s), 14
  Principle of, 15 (Rotational equilibrium)
  Resultant, 14 (Moment)
Momentum, 12, *97*
  Law of conservation of (linear), 13
Monatomic, *124*
Monera, 340 (I)
Monobasic, *153*
Monoclinic (basic crystal shape), *136*
Monoclinic sulfur, *136*, 184, 213
Monocotyledons, 261 (Cotyledon), 340
Monocytes, 286 (White blood cells)
Monoecious, 256 (I)
Monoisotopic, 83 (Isotopes)
Monomers, 125, 200, 204 (Starch), *205*
Monosaccharides, 204 (Carbohydrates, Glucose), 339
Monotropy, 136
Monovalent,
  (covalency), *133*
  (electrovalency), *131*
Morphology, 340 (I)
Morula, 320 (Embryo)
Mosaic image, 275 (Compound eye)
Moss, *321*
Mother liquid, 135
Motion, 10-11
  Brownian, 5, 123
  Circular, 17
  Equations of uniformly accelerated, 11
  Linear, 10
  Newton's laws of, 12-13
  Periodic, 16-17
  Rectilinear, see Linear motion,
  Rotational, 10 (I), 11
Motion (cont'd)
  Translational, 10 (I)
  Uniform circular, 17 (Circular motion)
  Wave, 34 (I)
Motor (electric), 76
Motor areas, 302 (Brain), 303
Motor end-plate, 283
Motor nerves, 306 (Nerves)
Motor neuron(s), 305, 308 (Efferent system)
  Lower, *308*, *309*
  Postganglionic, *309*
  Preganglionic, *309*
  Upper, *308*, *309*
Motor root, see Ventral root
Moving coil galvanometer, 77 (Galvanometer)
Moving iron meter, 77
mRNA, see Messenger RNA
M-shell, 83 (Electron shells)
Mucosa, see Mucous membrane
Mucous glands, 295 (Mucous membrane)
Mucous membrane, 295
Mucus, 295 (Mucous membrane)
Multicellular, 238 (I)
Multimeter, 77
Multiple(s) (units), 96
Multiple proportions, Law of, 125
Multiplier, 77 (Voltmeter)
  Electron, *90*
Mumetal, *73*
Muscle(s), 282-283
  Cardiac, 282
  Extrinsic eye, 313
  Hair erector, 311
  Intercostal, 299 (Inspiration)
  Intrinsic eye, *312* (Ciliary body)
  Involuntary, 282 (I, Cardiac muscle, Visceral muscles)
  Oblique, *313*
  Pectoralis, 269
  Rectus, *313*
  Skeletal, 282, 308
  Visceral, 282
  Voluntary, 282 (I, Skeletal muscles)
Muscle fibers, 283 (Striated muscle)
Muscle spindle, 283
Muscle (tissue), 283
  Cardiac, 283
  Smooth, 283
  Striated/Striped, 283
  Visceral, see Smooth muscle
Musical scale, 43 (Musical sounds)
Musical sounds, 43
Mutual induction, 78
Mutualists, 342
M-value, 139 (Molarity)
Mycelium, *321*
Myelin, 302 (Neuroglia), 304 (Nerve fibers)
Myofibrils, see Fibrils
Myofilaments, see Filaments (muscles)
Myosin, *205*, 283 (Striated muscle)

N

NAND (logic operation), 110,
Naptha, 199 (Chemical feeds
Nasal, 307
Nasal cavities, 294 (Pharynx
  307 (Nose)
Nasolachrymal duct, 313 (Lac
  glands)
Natural frequency, 17 (Natural
  oscillation)
Natural gas, 192 (Methane, Et
  198 (I), *208*
Natural oscillation, 17
Natural period, 17 (Natural os
Natural polymers, 201, 204 (St
  205 (Proteins)
Natural resins, 345 (Resins)
Natural selection, 237 (Genetic
  variation)
Neck (bones of), *278*
Neck (teeth), 284
Nectar, 256 (Nectaries)
Nectaries, 256
Negative feedback, 335 (Home
Negative terminal, 159 (Cell
Negative tropism, 251 (I)
Nekton, 342 (Pelagic)
Nematocyst, 270 (Cnidoblasts
Nematodes, 341
Neodymium (Nd), 164, 213
Neon (Ne), 165, 189, 213
Nephridia, 265 (Coelom), 273
Nephridiopore, *265*, 273 (Neph
Nephrons, 301
Neptunium (Np), 164, 213
Nerve(s), 306
  Cranial, 302 (Brain)
  Mixed, 306 (Nerves)
  Motor, 306 (Nerves)
  Optic, *312*, *313*
  Sensory, 306 (Nerves)
  Spinal, 302 (Spinal cord)
Nerve fibers, 304
Nervous system, 302-309
  Autonomic, 303 (Hypothalam
    308
  Central, 302-303
  Peripheral, 304 (I), 306 (I)
Neural canal, 279 (Vertebrae)
Neural pathways, *306*, 308, 309
Neural spine, see Spinous proce
Neuroglia, 302
Neurohypophysis, see Posterior
Neuron(s), 304 (I)
  Afferent, see Sensory neurons
  Association/Connecting, 305
  Efferent, see Motor neurons
  First order sensory, *306*, *309*
  Internuncial, see Association n
  Lower motor, *308*, *309*
  Motor, 305, 308 (Efferent syste
  Postganglionic motor, *309*

N

NAND (logic operation), 110, **111**
Naptha, **199** (**Chemical feedstocks**)
Nasal, **307**
Nasal cavities, 294 (**Pharynx**), **307** (**Nose**)
Nasolachrymal duct, **313** (**Lachrymal glands**)
Natural frequency, **17** (**Natural oscillation**)
Natural gas, **192** (**Methane, Ethane**), 198 (**I**), *208*
Natural oscillation, **17**
Natural period, **17** (**Natural oscillation**)
Natural polymers, **201**, 204 (**Starch**), 205 (**Proteins**)
Natural resins, **345** (**Resins**)
Natural selection, **237** (**Genetic variation**)
Neck (bones of), *278*
Neck (teeth), **284**
Nectar, **256** (**Nectaries**)
Nectaries, **256**
Negative feedback, **335** (**Homeostasis**)
Negative terminal, **159** (**Cell**)
Negative tropism, **251** (**I**)
Nekton, **342** (**Pelagic**)
Nematocyst, **270** (**Cnidoblasts**)
Nematodes, **341**
Neodymium (**Nd**), **164**, **213**
Neon (**Ne**), 165, **189**, 213
Nephridia, 265 (**Coelom**), **273**
Nephridiopore, *265*, **273** (**Nephridia**)
Nephrons, **301**
Neptunium (**Np**), **164**, **213**
Nerve(s), **306**
  Cranial, **302** (**Brain**)
  Mixed, **306** (**Nerves**)
  Motor, **306** (**Nerves**)
  Optic, *312*, *313*
  Sensory, **306** (**Nerves**)
  Spinal, **302** (**Spinal cord**)
Nerve fibers, **304**
Nervous system, **302-309**
  Autonomic, 303 (**Hypothalamus**), **308**
  Central, **302-303**
  Peripheral, 304 (**I**), **306** (**I**)
Neural canal, **279** (**Vertebrae**)
Neural pathways, *306*, 308, 309
Neural spine, see **Spinous process**
Neuroglia, **302**
Neurohypophysis, see **Posterior lobe**
Neuron(s), **304** (**I**)
  Afferent, see **Sensory neurons**
  Association/Connecting, **305**
  Efferent, see **Motor neurons**
  First order sensory, *306*, *309*
  Internuncial, see **Association neurons**
  Lower motor, *308*, *309*
  Motor, **305**, 308 (**Efferent system**)
  Postganglionic motor, *309*
Neuron(s) (cont'd)
  Preganglionic motor, *309*
  Relay, see **Association neurons**
  Second order sensory, *306*, *309*
  Sensory, **305**, *306* (**Afferent system**)
  Third order sensory, *306*
  Upper motor, *308*, *309*
Neurotransmitters, **305** (**Synapses**), 339 (**Vitamin B complex**)
Neutral (substance), **151**
Neutral (wires), *61*
Neutral equilibrium, **15**
Neutralization, **151**
  Enthalpy change of/Heat of, **146**
Neutral point, **72**
Neutrino, **87** (**Beta decay**)
Neutron number (**N**), **82**
Neutrons, **82**, 113, **126**
  Fission, **92** (**Nuclear fission**)
Newton (**N**), **6**, **97**
Newton meter (**Nm**), **14** (**Moment**)
Newton's law of gravitation, **18**
Newton's laws of motion, **12-13**
Niacin, **339** (**Vitamin B complex**)
Niche, Ecological, **233**
Nickel (**Ni**), 112, 165, **172**, 213
Nickel-cadmium cell, **69** (**Alkaline cell**)
Nickel sulfide (**NiS**), **172** (**Nickel**)
Nicotinamide/Nicotinic acid, see **Niacin**
Night-neutral plants, **251** (**Photoperiodism**)
Niobium (**Nb**), 165, **173**, 213
Nitrate bacteria, **235**
Nitrates/Nitrate(V) compounds, *155*, **182**, *209*, 210 (**Eutrophication**), 211, 218 (**Anions**), *235*
Nitric acid (HNO_3), *178*, 181 (**Nitrogen dioxide**), **182**, 210 (**Acid rain**), 218 (**Anions**)
Nitric(V) acid, see **Nitric acid**
Nitric oxide, see **Nitrogen monoxide**
Nitrides, **180** (**Nitrogen**)
Nitrifying bacteria, **235**
Nitrites/Nitrate(III) compounds, **182**, 210 (**Eutrophication**), 211, *235*
Nitrogen (N/N_2), *132*, 165, **180**, *183*, *208*, 209 (**Nitrogen cycle**), 210 (**Acid rain**), 213, 217, 235, 330
  Liquid, **180** (**Nitrogen**)
Nitrogen base, **324**
Nitrogen cycle, **209**, **235**
Nitrogen dioxide (NO_2), *125*, *162*, *178*, **181**, *182*, *210*, 211
Nitrogen fixation, **235**
Nitrogen-fixing bacteria, **209**, **235** (**Nitrogen fixation**)
Nitrogen monoxide (**NO**), *162*, **181**, *182*
Nitrogenous, **300** (**Urea**)
Nitrogen oxide, see **Nitrogen monoxide**
Nitrous acid (HNO_2), **181** (**Nitrogen dioxide**)
Nitrous oxide, see **Dinitrogen oxide**
Nobelium (**No**), **165**
Noble gases, 165, **189**, *208*, **217** (**Nitrogen**)
Nocturnal, **342**
Nodal lines, *38*, **39** (**Nodes**)
Nodal points, see **Nodes**
Node(s),
  Lymph, **293**
  (plants), **244**
  (waves), **39**, **43** (**Stationary waves**)
Node of Ranvier, *304*
Nodules, Root, **235** (**Nitrogen fixation**)
Noise, **42** (**I**)
Nonane (C_9H_{20}), *198*
Non-aqueous solutions, **144** (**Aqueous solution**)
Non-aqueous solvents, **144** (**Aqueous solution**)
Non-biodegradable detergents, **203** (**Biodegradable detergents**)
Non-electrolyte, **66**, *156*
Non-linear (molecule), see **V-shaped**
Non-luminous flame, *208*
Non-metal, **165** (**Metal**)
Non-polar molecule, **133** (**Polar molecule**)
Non-polar solvent, **144**
NOR (logic operation), 110, **111**
Noradrenalin/Norepinephrin, **336**
Normal, *47*, *50*
Normal contact force, **7** (**Coefficient of friction**)
Normal salt, **153**
North/North seeking pole, **70** (**Pole**)
Nose, **307**
NOT (logic operation), 110, **111**
Notation (numbers), **109**
  Exponential, **109**
Notes, **43** (**Musical sounds**)
Notochord, **341** (**Chordates**)
Npn transistor, *65*, 110
N-shell, **83** (**Electron shells**)
N-type (semiconductor), **65** (**Doping**)
Nuclear energy, **84-85**
Nuclear equation(s), *92*, *93*, *128*
Nuclear fission, *85*, **92-93**, 94-95, **129**
Nuclear force, **7**, **84**
Nuclear fusion, *85*, **93**, 94 (**Fusion reactor**), **129**
Nuclear magnetic resonance (**n.m.r.**) **spectroscopy**, **222**
Nuclear magnetic resonance (**n.m.r.**) **spectrum**, **222**
Nuclear membrane, 238 (**Nucleus**)
Nuclear potential energy, **8**
Nuclear power stations, **94** (**I**)
Nuclear reactions, *85*, **92** (**I**), **94-95**, **129**
Nuclear reactor, **94**(**I**)
Nuclei, see **Nucleus**
Nucleic acids, **324**
Nucleoli (sing. **nucleolus**), *239*, **240**
Nucleon(s), **82** (**Nucleus**)
  Binding energy per, *84*

Nucleoplasm, 238 (Nucleus)
Nucleotides, 324 (Nucleic acids)
Nucleus (pl. **nuclei**) (atoms), **82**, 84-85, 92-95, **126**, 127 (**Atomic number**), 128-129
Nucleus (cells), **238**, 240 (**Mitosis**), 322 (**Meiosis**), 324 (**Chromosomes**)
  Daughter, 241 (Telophase), 322 (**Meiosis**)
  Generative, 258 (Pollen)
  Male, 258 (**I, Pollen**), **320 (Gametes)**, 323 (**Gamete production, male**)
  Tube, 258 (Pollination)
Number(s),
  Atomic (Z), 82, 127, 164-165, 212-213
  Avogadro's, 96 (**Mole**), **113, 139**
  Diploid, 240 (Mitosis), 324 (**Chromosomes**)
  Haploid, 322 (Meiosis), 326 (**I**)
  Mass (A), **82, 127**
  Neutron, 82
  (notation of), **109**
  Oxidation, 149
  Pyramid of (ecology), *234*
Nut, 262
Nutrients, 204 (I), 330 (I)
Nylons, 201
Nylon 66, *200*
Nymph, *277*

O

Objective lens,
  (microscope), *54*
  (telescope), *55*
Oblique muscles, *313*
Octadecanoic acid ($C_{17}H_{35}COOH$), 202 (**Soap**), *205*
Octan-3-ol, *215* (**4b**)
Octane rating, 199
Octet, 83 (Electron shells), **127**
Oddi, Sphincter of, *296*, 337 (**CCK**)
Odontoid process, see **Dentes** (vertebrae)
Ohm (Ω), 62 (Resistance), 97
Ohm's law, 62
Oil(s),
  Crude, see **Petroleum**
  Diesel, 199
  (food), **205 (Lipids)**
  Fuel, 199 (Residue)
  Gas, see **Diesel oil**
  Lubricating, 199
Olefins, see **Alkenes**
Oleum, see **Fuming sulfuric acid**
Olfactory area, Primary, 303
Olfactory bulb, *307*
Olfactory cells, 307 (Nose)
Olfactory hairs, 307 (Nose)
Olfactory sensations, 307 (Nose)
Omasum, 271 (Rumen)
Ommatidia (sing. **ommatidium**), **275 (Compound eye)**
Omnivores, 235
Ootid, 323 (Gamete production, female)
Open system, 162 (Closed system)
Operculum,
  (fish), **272 (Internal gills)**
  (snails, etc.), *265*
Opposing pairs, see **Antagonistic pairs**
Opposite (leaves), **250**
Optic, 313
Optical center (O), *52*
Optical density, 50 (Snell's law)
Optical fibers, *51*
Optical instruments, 54-55
Optic disk, see **Blind spot**
Optic nerve, *312*, *313*
OR (logic gate/operation), 110, **111**
Oral cavity, 294 (**Pharynx**), *307*
Oral groove, *268*
Orbit (eye), *278*, **312 (I)**, *313*
Orbit (satellite), **Geo-stationary/Parking, 19**
Orbital, 83 (Electron shells), **127**
Ordering, 71 (Domain theory of magnetism)
Order of magnitude, 109
Orders (classification), **340 (I)**
Ore, 345
  Chrome iron, 172 (Chromium)
Organ(s), **238 (I)** 264-265, *289*
  Digestive, 270-271, 294-295
  Excretory, 273, 300-301
  Genital, 316 (I)
  Lymphatic/Lymphoid, 293
  Reproductive, 277
  Respiratory, 272-273, 298-299
  Sense, 274-275, 307
  Target, 336 (Hormones)
  Tympanal, 275
Organelles, 239-240
Organic acid, 150
Organic chemistry, 190, 191-205
Organic compounds, 190 (I), 214-215
Organic solvent, 345
Organism, 238 (I)
Organ modification, 328
Organ of Corti, 314 (Cochlear duct)
Origin (graphs), *98*
Orthorhombic sulfur, see **Rhombic sulfur**
Oscillation(s), **16**, **34 (I)**
  Forced, 17
  Free/Natural, 17
Oscilloscope,
  Cathode ray (CRO), *73*, **81**
O-shell, 83 (Electron shells)
Osmium (Os), 165, **173**, 213
Osmium tetroxide (OsO_4), **173 (Osmium)**
Osmosis, 329
Osmotic pressure, 329 (Osmosis)
Osseous tissue, see **Bone**
Ossicles (Ear/Auditory), 314 (Middle ear)
Ossification, 281 (Cartilage)
Osteoblasts, 280 (Periosteum)
Osteocytes, 281 (Bone)
Osteogenesis, see **Ossification**
Ostwald process, 182 (Nitric acid)
Otoliths, 315 (Saccule)
Outer ear, 314
Outer shell, 83 (Electron shells), **127**
Out of phase (waves), **38 (Phase)**
Output (computers), **111**
Ova (sing. **ovum**), 276, **317 (Ovaries)**, 318 (**Menstrual cycle**), **320 (Gametes)**, 323 (**Gamete production, female**), 337 (**FSH**)
Oval window (ears), *314*
Ovarian cycle, 318 (Menstrual cycle)
Ovarian follicles, 317, 337 (**FSH, LH, Estrogen**)
Ovarian ligaments, *317*
Ovaries,
  (humans), **317**, 337 (**Estrogen**)
  (plants), **257**
Overtones, 43 (Modes of vibration)
Over water, 216 (Carbon dioxide, Ethene), 217 (**Nitrogen**)
Oviduct, 277
Oviparous, 276
Ovipositor, 277
Ovulation, *317* (**Fallopian tubes**), 318 (**Menstrual cycle**), 337 (**LH**)
Ovules, 257 (**Ovaries**), **258**
Oxalic acid, see **Ethanedioic acid**
Oxidation, 148-149, 334 (**Aerobic respiration**)
Oxidation number, 149
Oxidation state, 149
Oxides, 183, 209 (**Corrosion**), 211
Oxidizing agent, 148
Oxonium ion, see **Hydronium ion**
Oxygen (O/O_2), *132*, *149*, *162*, 165, *170*, *178*, *182*, **183**, *204*, *208*, *209*, 211, 213, *214*, *215*, 217, 218 (**Gases**), 254-255, 288-291, 298-299, 330, 333, 334
Oxygenated, 290 (Atria)
Oxygen debt, 334
Oxygen process, Basic, 174 (Steel)
Oxytocin, 336
Ozone (O_3), 183, *207*, 210 (**Ozone depletion**)

P

Pacinian corpuscles, 311
Pain receptors, 311
Paired fins, 269
Palate,
  Hard, *307*
  Soft, 294 (Pharynx), *307*
Palatine tonsils, 293 (Tonsils)
Palisade cells, 248 (Palisade
Palisade layer, 248
Palladium (Pd), 112, 165,
Palmate, 250
Palps, *271*, **274**
Pancreas, 297
Pancreatic amylase, 338 (Pa
juice)
Pancreatic duct, 297 (Panc
Pancreatic juice, 297 (Panc
336 (**Secretin, PZ**), 338
Pancreatic lipase, 338 (Panc
juice)
Pancreatic vein, *289*
Pancreozymin, see **PZ**
Pantothenic acid, 339 (Vitam
complex)
Paper capacitor, 59
Paper chromatography, 22
Papillae (sing. **papilla**), 307
Paraffin, see **Kerosene**
Paraffins, see **Alkanes**
Paraffin waxes, see **Hydroca**
waxes
Parallax, 47
Parallax error, 102
Parallel (components), **64**
Parallelogram rule, 108
Paramagnetism, 72
Paramecium, *268*, *273*
Parapodia (sing. **parapodium**
Parasites, 342
Parasympathetic division (ne
system), **308**
Parathormone/Parathyrin, se
Parathyroid glands, 297, 33
Parathyroid hormone, see **P**
Parenchyma, 243 (Cortex)
Parent cell, 240 (Cell divisio
Parking orbit, see **Geo-statio**
Parotid gland, *296*
Partial eclipse, *46*
Partial pressure(s), **143**
  Dalton's law of, 143
Particle accelerators, 86 (Ra
Parturition, 319 (**Pregnancy**
Pascal (Pa), *25*, **97**
Passage cells, 243 (Endoderm
Patella, *278*, *280*, *281*
P.E. see **Potential energy**
Pea flower, 259
Peaks (waves), see **Crests**
Pectoral fins, 269 (Paired fin
Pectoralis muscles, 269
Pedicles, *279*
Peduncle, 256 (Receptacle)
Pelagic, 342
Pelvic fins, 269 (Paired fins)
Pelvic girdle, see **Pelvis** (ske
Pelvis,
  Renal (kidneys), *300*
  (skeleton), *278*
Penis, 316, 319 (**Copulation**
Pent-1-ene (C_5H_{10}), 193

Palisade cells, **248** (**Palisade layer**)
Palisade layer, **248**
Palladium (**Pd**), 112, 165, **173**, 213
Palmate, **250**
Palps, *271*, **274**
Pancreas, **297**
Pancreatic amylase, **338** (**Pancreatic juice**)
Pancreatic duct, **297** (**Pancreas**)
Pancreatic juice, 297 (**Pancreas**), 336 (**Secretin**, **PZ**), **338**
Pancreatic lipase, **338** (**Pancreatic juice**)
Pancreatic vein, *289*
Pancreozymin, see **PZ**
Pantothenic acid, **339** (**Vitamin B complex**)
Paper capacitor, **59**
Paper chromatography, *221*
Papillae (sing. **papilla**), *307*
Paraffin, see **Kerosene**
Paraffins, see **Alkanes**
Paraffin waxes, see **Hydrocarbon waxes**
Parallax, **47**
Parallax error, **102**
Parallel (components), **64**
Parallelogram rule, **108**
Paramagnetism, **72**
Paramecium, *268*, *273*
Parapodia (sing. **parapodium**), **268**
Parasites, **342**
Parasympathetic division (nervous system), **308**
Parathormone/Parathyrin, see **PTH**
Parathyroid glands, **297**, 337 (**PTH**)
Parathyroid hormone, see **PTH**
Parenchyma, **243** (**Cortex**)
Parent cell, **240** (**Cell division**)
Parking orbit, see **Geo-stationary orbit**
Parotid gland, *296*
Partial eclipse, *46*
Partial pressure(s), **143**
  Dalton's law of, **143**
Particle accelerators, **86** (**Radioisotope**)
Parturition, **319** (**Pregnancy**)
Pascal (**Pa**), *25*, **97**
Passage cells, **243** (**Endodermis**)
Patella, *278*, *280*, *281*
P.E. see **Potential energy**
Pea flower, **259**
Peaks (waves), see **Crests**
Pectoral fins, **269** (**Paired fins**)
Pectoralis muscles, **269**
Pedicles, *279*
Peduncle, **256** (**Receptacle**)
Pelagic, **342**
Pelvic fins, **269** (**Paired fins**)
Pelvic girdle, see **Pelvis** (skeleton)
Pelvis,
  Renal (kidneys), *300*
  (skeleton), *278*
Penis, **316**, 319 (**Copulation**)
Pent-1-ene (C_5H_{10}), **193**
Pentane (C_5H_{12}), **192**
Penumbra, **46** (**Shadow**)
Pepsin, **338** (**Gastric juice**)
Pepsinogen, **338** (**Note 2**)
Peptidases, see **Proteinases**
Peptide links, **330** (**Proteins**), 338 (**I**)
Percentage composition, **140**
Perennials, **236**
  Herbaceous, **236** (**Perennials**)
  Woody, **236** (**Perennials**)
Perfect machine, **20**
Perfoliate, **250**
Perianth, *256*
Pericardial cavity, **290** (**I**)
Pericardial fluid, **290** (**I**)
Pericardial sac, **290** (**I**)
Pericardium, **290** (**I**)
Pericarp, **262** (**I**),
Periderm, **247** (**New outer tissue**)
Perigynous flower, **257**
Perikaryon, see **Cell body**
Perilymph, **314** (**Inner ear**)
Perimysium, *283*
Perineurium, *306*
Period (element group), **164**
  Long, **164** (**Period**)
  Short, **164** (**Period**)
Period (**T**) (quantity), **16**, *34*, 97
  Natural, **17** (**Natural oscillation**)
Periodic motion, **16-17**
Periodic table, **164-165**
Periodontal ligament, **284** (**Root**)
Periosteum, **280**
Peripheral nervous system, 304 (**I**), **306** (**I**)
Peristalsis, **295**
Peritoneum, **265** (**Coelom**), 294 (**Alimentary canal**)
Perivisceral cavity, **265** (**Body cavities**), 294 (**Alimentary canal**)
Permanent dentition, **284** (**I**)
Permanent hardness (water), **207**
Permanent magnets, **70** (**Hard**)
Permanent teeth, **284** (**I**)
Permeability, **73**
Peroxides, *149* (**rule 4**)
Petals, **256**,
  Wing, **259** (**Pea flower**)
Petiole, *248*
Petroleum, **198-199**, *208*, 209,
  Fractional distillation of, see **Primary distillation**
pH, **152**, 210 (**Acid rain**)
Phagocytosis, **268** (**Pseudopodium**)
Phalanges (sing. **phalanx**), *278*
Pharming, **327**
Pharyngeal tonsil, **293** (**Tonsils**)
Pharynx, **294**
Phase (cardiac cycle),
  Diastole, **291**
  Systole, **291**
Phase (cell division),
  (meiosis), **322-323**
  (mitosis), **241**
Phase (substances), **120**
  Stationary, **221** (**Chromatography**)
Phase (waves), **38**
Phase difference, **38** (**Phase**)
Phellem, **247**
Phelloderm, **247**
Phellogen, **247**
Phenolphthalein, *152*
Phenotypes, *325*
Phenylethene, see **Styrene**
Pheromone, **275**
Phloem, **243**, 246 (**Secondary thickening**), *247*
  Primary, *242* (**Primary tissue**)
  Secondary, **246** (**Secondary thickening**)
Phons, **42** (**Loudness**)
Phosphate groups, *324*, **335** (**ADP**)
Phosphates, **210** (**Eutrophication**)
Phosphorescence, **44**
Phosphoric acid (H_3PO_4), **182** (**Phosphorus pentoxide**)
Phosphors, **44** (**Phosphorescence**)
Phosphorus (**P**), 165, 180, **182**, 213
  Red, **182** (**Phosphorus**), 213
  White, **182** (**Phosphorus**), 213
Phosphorus pentoxide (P_2O_5), **182**
Photocell, **345**
Photochemical reaction, **160**
Photoelectric cell, see **Photocell**
Photomultiplier, **90** (**Scintillation counter**)
Photons, **84** (**Quantum theory**)
Photoperiodism, **251**
Photoperiods, **251** (**Photoperiodism**)
Photoreceptors, **313** (**Retina**)
Photosynthesis, 204, 209, **254-255**, 330
Phototropism, **251**
Phyla (sing. **phylum**), **340** (**I**)
Phylloquinone, see **Vitamin K**
Physical change, **119**
Physical chemistry, **118**, 119-163
Physical properties, **120**
Physical quantities, **96** (**I**), **98-99**
Physical states, 4 (**I**), **5**, **30-31**, **120-121**
Phytohormones, **336** (**Hormones**)
Phytoplankton, **342** (**Plankton**)
Pia mater, *303* (**Meninges**)
Pig iron, **174** (**Iron**)
Pigments, **255**
Piliferous layer, **245**
Pineal gland/body, **297**
Pinion, *267*
Pinnae (sing. **pinna**),
  (ears), **314** (**Outer ear**)
  (leaves), **250** (**Pinnate**)
Pinnate, **250**
Pinocytosis, **329**
Pioneer community, *233*
Pipeclay triangle, **225**
Pipette(s), **225**
  Dropping/Teat, **225**
Pistil, see **Carpel**
Pistillate, **256** (**I**)

Pitch,
  (mechanical), **21** (**Screw jack**)
  (sound), **43**
Pith, **243**
Pituitary gland/body, **297**, *303*, 336 (**Hormones**), 337
Pivot joint, **279** (**Vertebral structure**)
Pixels, **81**
Placenta,
  (humans), *319*, 337 (**Estrogen**)
  (plants), **257** (**Ovaries**)
Placoid scales, see **Denticles**
Planck's constant, *84*
Plane coil, see **Flat coil**
Plane joints, see **Gliding joints**
Plane mirror, **47**
Plankton, 270 (**Filter-feeding**), 272 (**Gill rakers**), **342**
Plano-concave, *53*
Plano-convex, *53*
Plant(s), 242-263, **340** (**I**)
  Leguminous, **235** (**Nitrogen fixation**)
  Long-night, **251** (**Photoperiodism**)
  Night-neutral, **251** (**Photoperiodism**)
  Short-night, **251** (**Photoperiodism**)
  Vascular, **242-243**
  Woody, **246** (**I**)
Plantigrade, **269**
Plasma, **286**
Plasmalemma/Plasma membrane, see **Cell membrane**
Plasmids, *327* (**Gene cloning**)
Plasmolysis, **253**
Plastic (quality), **22** (**Elasticity**)
Plastic(s) (substances), **201**
  Thermosetting, *200*, **201** (**Plastics**)
Plastic deformation, **23** (**Yield point**)
Plastic sulfur, **184**
Plastids, *238*, **240**
Platelets, **286**
Platinum (**Pt**), 112, 165, **173**, *182*, 213, 222 (**Flame test**)
Platinum electrode, *158*
Pleura, **298**
Pleural cavity, **298** (**Pleura**)
Pleural fluid, **298** (**Pleura**)
Pleural membrane, see **Pleura**
Pleural sac, **298** (**Pleura**)
Plexuses (**Hair/Root hair**), **311**
Plotting compass, **72** (**Magnetic field lines**)
Plumage, **267** (**Feathers**)
Plumbic compounds, see **Lead(IV) compounds**
Plumbous compounds, see **Lead(II) compounds**
Plumule(s),
  (birds), see **Down feathers**
  (plants), **261**
Plutonium (**Pu**), **92** (**Induced fission**), **95** (**Fast reactor**), *128*, **164**, **213**
Pnp transistor, *65*, 110
Pod, see **Legume**
Poikilothermic, **335** (**Homeostasis**)
Point(s),
  Antinodal, see **Antinodes**
  Boiling, **30** (**Vaporization**), *31*, **121** (**Boiling**), 212, 213
  Compensation, **255**
  End, **222** (**Volumetric analysis**)
  Fixed (**Lower**, **Upper**), **26**
  Focal, see **Principal focus**
  Freezing, **30** (**Freezing**), **121** (**Freezing**)
  Growing (roots), **245**
  Ice, **26** (**Fixed point**)
  Melting, **30** (**Melting**), *31*, **120** (**Melting**), 212, 213
  Neutral, **72**
  Nodal, see **Nodes**
  Steam, **26** (**Fixed point**)
  Triple, **96** (**Kelvin**)
  Yield, **23**
Point action, **57**
Point of incidence, *47*, *50*
Polar bodies, **323** (**Gamete production, female**)
Polar bond, **133**
Polarization (bonding), **133** (**Polar bond**)
Polarization (cells), **68**
Polar molecule, **133**
Polar solvent, **144**
Pole(s) (magnets), **70**
  Consequent, *71*
  North/North seeking, **70** (**Pole**)
  South/South seeking, **70** (**Pole**)
Pole(s) (mirrors), *48*
Pollen, 257 (**Stamens**), **258**, 323 (**Gamete production, male**)
Pollen sacs, **257** (**Stamens**)
Pollen tube, **258** (**Pollination**)
Pollination, **258**
  Cross, **259**
  Self, **259**
Pollutants, **210** (**I**)
Pollution, **210**
  Thermal, **210**
Polonium (**Po**), *87*, 165, **183**, 213
Poly((1-methoxycarbonyl)-1-methylethene), see **Acrylic**
Polyamides, **201**
Polyatomic, *124*
Poly(chloroethene), see **Polyvinyl chloride**
Polyesters, **201**
Polyester capacitor, **59** (**Paper capacitor**)
Poly(ethene), see **Polythene**
Poly(ethylene), see **Polythene**
Polyhydric alcohols, **197**
Polymerization, *200*
  Addition, **200**
  Condensation, **200**
Polymers, **200-201**
  Man-made, see **Synthetic polymers**
  Natural, **201**, **204** (**Starch**), 205 (**Proteins**)
  Synthetic, *200*, **201**
Poly(methylmethacrylate), see **Acrylic**
Polymorphism, **136**
Polypeptides, *205*, 330 (**Proteins**), **339**
Poly(phenylethene), see **Polystyrene**
Poly(propene), **193** (**Propene**)
Polypropylene, see **Poly(propene)**
Polysaccharides, **204** (**Carbohydrates**, **Glucose**, **Starch**), **339**
Polystyrene, **201**, 210
Poly(tetrafluoroethene) (**PTFE**), *186*, *195*
Polythene, 193, *200*, **201**
Polyunsaturated compounds, **191**
Polyvinyl chloride (**PVC**), 194 (**Ethyne**), *200*, **201**
Pome, **262**
Pons/Pons Varolii, **303**
Pore, *311*
  Gustatory, *307*
Porous, *159*
Portal vein, Hepatic, *289*, **297** (**Liver**)
Positive terminal, **159** (**Cell**)
Positive tropism, **251** (**I**)
Positrons, **86** (**Beta particles**), 87 (**Beta decay**)
Posterior cavity (eye), *312* (**Vitreous humor**)
Posterior lobe (pituitary gland), **297**
Postganglionic motor neuron, *309*
Potassium (**K**), 164, 168, **169**, 211, 213, 219 (**Flame tests**)
Potassium bromide (**KBr**), *186*
Potassium carbonate (**K_2CO_3**), **169**
Potassium chloride (**KCl**), **169**
Potassium dichromate (**$K_2Cr_2O_7$**), **222** (**Gravimetric analysis**)
Potassium hydroxide (**KOH**), **169**
Potassium iodide (**KI**), *155*, *186*, **188** (**Iodine**)
Potassium nitrate (**KNO_3**), *153*, *155*, **169**
Potassium permanganate (**$KMnO_4$**), *196*
Potassium sulfate (**K_2SO_4**), **169**
Potential, **58**, 106-107
  Electrode (**E**), **158**
  Redox, **149**
Potential difference, **58**, 97, 106-107, **159**
Potential divider, **63**
Potential energy (**P.E.**), **8**, 106-107
  Elastic, **8**
  Electromagnetic, **8**
  Gravitational, **8**
  Molecular, **8**
  Nuclear, **9**
Potential gradient, **58** (**Potential**)
Potentiometer, **63** (**Variable resistor**), **110**
Power (**P**), **9**, 97
  (lens), **53**
  Magnifying, see **Angular magnification**
Power of hydrogen, see **pH**
Power station, *61*
  Nuclear, **94** (**I**)
PR, see **Lactogenic horm**
Praseodymium (**Pr**), 164, 2
Precipitate, **145**
Predators, **342**
Preferential discharge, 66 (**of electrolysis**)
Prefixes (units), **96**
Preganglionic motor neur
Pregnancy, **319**
Premolars, 270 (**Carnassial**
Preoptic area, **335** (**Homeost**
Prepuce, see **Foreskin**
Pressure, **24-25**, 97, 142-1
  Atmospheric, **25** (**Barome**
  Dalton's law of partial, 14
  Osmotic, **329** (**Osmosis**
  Partial, **143**
  Room, **139** (**Molar volume**
  Root, **252**
  Standard, **143** (**s.t.p.**)
  Turgor, **253** (**Turgor**)
  Wall, **253** (**Turgor**)
Pressure law, **33**, **142**
Pressurized water reactor
Prey, **342** (**Predators**)
Prides, **342** (**Social**)
Primaries, see **Primary feath**
Primary alcohols, 196, **19**
Primary amines, **195**
Primary auditory area, **30**
Primary bronchi, *298*, **29**
Primary bud, **261** (**Plumule**
Primary cells, **69**, **159** (**Cell**
Primary circuit, *78*
Primary coil, **79**
Primary colors, **55**
Primary consumers, **235**
Primary distillation, **198**
Primary feathers, **267** (**Remi**
Primary gustatory area, **30**
Primary olfactory area, **30**
Primary phloem, *242* (**Prima**
Primary root, 245 (**Tap root**, **261** (**Radicle**)
Primary sex characters, **31**
Primary tissue, *242*
Primary visual area, **303**
Primary xylem, *242* (**Primary**
Primates, **341**
Prime mover, see **Agonist**
Principal axis (reflection, re *48*, *52*
Principal focus (**F**), *48*, *52*
Principle, Archimedes', **25**
Principle, Gause's, **233** (**Eco niche**)
Principle, Le Chatelier's, **16**
Principle of flotation, **25**
Principle of reversibility of l **49**, *50*
Principle of superposition, 3
Prism, **51**
Probability clouds, see **Orbit**
Proboscis, *271*

PR, see **Lactogenic hormone**
Praseodymium (**Pr**), **164**, **213**
Precipitate, **145**
Predators, **342**
Preferential discharge, **66** (**Ionic theory of electrolysis**)
Prefixes (units), **96**
Preganglionic motor neuron, *309*
Pregnancy, **319**
Premolars, 270 (**Carnassial teeth**), **285**
Preoptic area, **335** (**Homeostasis**)
Prepuce, see **Foreskin**
Pressure, **24-25**, 97, 142-143
  Atmospheric, **25** (**Barometer**), 113
  Dalton's law of partial, **143**
  Osmotic, **329** (**Osmosis**)
  Partial, **143**
  Room, **139** (**Molar volume**)
  Root, **252**
  Standard, **143** (**s.t.p.**)
  Turgor, **253** (**Turgor**)
  Wall, **253** (**Turgor**)
Pressure law, **33**, **142**
Pressurized water reactor (**PWR**), *95*
Prey, **342** (**Predators**)
Prides, **342** (**Social**)
Primaries, see **Primary feathers**
Primary alcohols, 196, **197**
Primary amines, **195**
Primary auditory area, **303**
Primary bronchi, *298*, **299** (**Bronchi**)
Primary bud, **261** (**Plumule**)
Primary cells, **69**, **159** (**Cell**)
Primary circuit, *78*
Primary coil, **79**
Primary colors, **55**
Primary consumers, **235**
Primary distillation, **198**
Primary feathers, **267** (**Remiges**)
Primary gustatory area, **303**
Primary olfactory area, **303**
Primary phloem, *242* (**Primary tissue**)
Primary root, 245 (**Tap root**), **261** (**Radicle**)
Primary sex characters, **318** (**Puberty**)
Primary tissue, *242*
Primary visual area, **303**
Primary xylem, *242* (**Primary tissue**)
Primates, **341**
Prime mover, see **Agonist**
Principal axis (reflection, refraction), *48*, *52*
Principal focus (**F**), *48*, *52*
Principle, Archimedes', **25**
Principle, Gause's, **233** (**Ecological niche**)
Principle, Le Chatelier's, **163**
Principle of flotation, **25**
Principle of reversibility of light, **49**, *50*
Principle of superposition, **38**
Prism, **51**
Probability clouds, see **Orbitals**
Proboscis, *271*
Procarboxypeptidase, **338** (**Note 2**)
Process (chemical),
  Basic oxygen, **174** (**Steel**)
  Contact, **185**
  Frasch, **184**
  Haber, **180**
  Ostwald, **182** (**Nitric acid**)
  Solvay, **169** (**Sodium carbonate**)
Process(es) (vertebrae), **279**
  Inferior articular, *279*
  Odontoid, see **Dens** (vertebrae)
  Spinous, *279*
  Superior articular, *279*
  Transverse, *279*
Producer gas, **179** (**Carbon monoxide**)
Producers, **235**
Products, **119**
  Fission, **92** (**Nuclear fission**)
Progesterone, 318 (**Menstrual cycle**), **336**, 337 (**LH**)
Projection lens (slide projector), *55*
Prolactin, see **Lactogenic hormone**
Promethium (**Pm**), **164**, **213**
Promotor, **161**
Proof plane, **57**
Propagation,
  Artificial, **262**
  Vegetative, see **Vegetative reproduction**
Propan-1-ol (C_3H_7OH), **196**
Propan-2-ol ($CH_3CHOHCH_3$), *196*
Propanal (C_2H_5CHO), *196*, *215*
Propane (C_3H_8), **192**, *214*
Propane-1,2,3-triol, *197*, *202*, *205*
Propanoic acid (C_2H_5COOH), *196*
Propanone (C_3H_6O), 193 (**Propene**), **194**, *196*, *215*
Propene (C_3H_6), **193**, *214*
Properties,
  Chemical, **119**
  Molecular, **22-23**
  Physical, 118, **119**, 212-213
  Qualitative, **119**
  Quantitative, **119**
  Thermometric, **26** (**Thermometer**)
Prophase,
  (meiosis), **322**
  (mitosis), **241**
Proportional, **345**
  Inversely, **345**
Proportionality,
  Limit of, **22** (**Hooke's law**), *23*
Proportional limit, see **Limit of proportionality**
Prop roots, **245**
Propylene, see **Propene**
Propyne (C_3H_4), **194**, *214*
Prostate gland, *316*
Protactinium (**Pa**), *128*, **164**, **213**
Protection,
  (biology), **237**
  Electrical/Sacrificial, **159**
Proteinases, **338** (**I**)
Protein manufacture, **327**
Proteins, 204 (**I**), **205**, *209*, 239 (**Ribosomes**), **330**, 331, 338, 339
  Catalytic, **330** (**Proteins**), **333** (**Enzymes**)
  Structural, **330** (**Proteins**)
Prothrombin, **287** (**Clotting**), **339** (**Vitamin K**)
Protista, **340**
Protonema, *321*
Protonephridia, **273** (**Nephridia**)
Protons, **82**, 113, **126**
Protozoa, **340** (**I**)
Proximal, **301**
Proximal convoluted tubule, **301** (**Uriniferous tubules**)
Pseudopodium (pl. **pseudopodia**), **268**
Pseudotracheae, *271*
P-shell, **83** (**Electron shells**)
Pterophyta, **340**
PTFE, see **Poly(tetrafluoroethene)**
PTH, **336**
Ptyalin, see **Salivary amylase**
P-type (semiconductor), **65** (**Doping**)
Puberty, **318**, 337 (**Estrogen**, **Androgens**)
Pubic hair, *318*
Pubis, *278* (**Pelvis**)
Pudendum, see **Vulva**
Pulley system, *20*, **21**
Pulmonary, **290**
Pulmonary arteries, *290*, **291** (**Pulmonary trunk**)
Pulmonary trunk, *290*, **291**
Pulmonary valve, **291** (**Semilunar valves**)
Pulmonary veins, *290*, **291**
Pulp, **284** (**Pulp cavity**)
Pulp cavity, **284**
Pulse (**Wulf**) **electroscope**, **89**
Pupa (pl. **pupae**), *277*
Pupil, *312* (**Iris**)
Pure, **123**
Purification, **207**
PVC, see **Polyvinyl chloride**
PWR, see **Pressurized water reactor**
Pyloric sphincter/valve/Pylorus, *294*, **295**
Pyramidal (molecule), *133*
Pyramid of biomass (ecology), *235*
Pyramid of numbers (ecology), *235*
Pyramids (**Renal/Medullary**), **300** (**Medulla**)
Pyridoxine, **337** (**Vitamin B complex**)
Pyrolusite (MnO_2), **172** (**Manganese**)
Pyruvic acid, **334** (**Anaerobic respiration**)
PZ, **336**

Q

Q-shell, **83** (**Electron shells**)
Qualitative analysis, 218 (**I**), 220 (**I**), **222**
Qualitative properties, **119**
Quanta (sing. **quantum**), **84** (**Quantum theory**)

Quantitative analysis, 220 (I), 222
Quantitative properties, 119
Quantities, 96-97, 113
  Basic, 96
  Derived, 97
  Physical, 96 (I), 98-99
  Vector, 108
  Scalar, 108
Quantum theory, 84
Quartz, *122*, 177 (Silicon dioxide)
Quicklime, see Calcium oxide
Quicksilver, see Mercury

R

Rachis, see Shaft (birds)
Radar, 45 (Microwaves)
Radial magnetic field, 72 (Loudspeaker)
Radial symmetry, 264
Radiant heat energy, 29 (Radiation)
Radiation, 9, 29, 86-87, 128 (I)
  Background, 88
  Cosmic, 88 (Background radiation)
  Infra-red (IR), 45, *113*
  Ultraviolet (UV), 44, *85*, *113*
Radicle, 261
Radioactive decay, 87, 128
Radioactive decay curve, *129*
Radioactive implants, *91*
Radioactive isotope, see Radioisotope
Radioactive series, see Decay series
Radioactive tracing, 91, 129
Radioactivity, 86-87, 88-91, 128-129
Radiocarbon dating, 91, 129
Radiograph,
  (gamma rays), 91 (Gamma radiography),
  (x-rays), *44*
Radiography,
  Gamma, 91
  (x-rays), *44*
Radioisotope, 86, 128
Radiology, 91, 129
Radiotherapy, 91, 129 (Radiology)
  External beam, *91*
Radio waves, 44 (I), 45, *113*
Radium (Ra), 87, 129 (Half-life), 164, 170, 213
Radius (bone), *278*, 282
Radius of curvature (r), *48*
Radon (Rn), 165, 189, 213
Radula, 270
Random access memory (RAM) (computers), 111 (Memory)
Range,
  (force field), 6
  Sonic, 40 (I)
Ranvier, Node of, *304*
Rapid combustion, 208
Raptors, 342 (Predators)
Rare earths, see Lanthanides
Rarefactions (waves), 35
Rare gases, see Noble gases
Rate(s), 160-161, 345
  Basal metabolic, 332 (Metabolic rate)
  Metabolic, 332
Rate curve, 160 (Rate of reaction)
Ratemeter, 89 (Geiger counter)
Rates of reaction, 160-161
  Average, 160
  Initial, 160
  Instantaneous, 160
Raw material, 345
Ray(s), Astral, *241*
Ray(s) (fish), 269 (Fins)
Ray(s) (light), 46 (I), 47-55
  Critical, 51 (Critical angle)
  Incident, *47*, *50*
  Reflected, *47*
  Refracted, *50*
Ray(s) (particles/waves),
  Alpha (α-rays), 86 (I)
  Beta (β-rays), 86 (I)
  Cathode, 80-81
  Cosmic, 86 (Radioisotope), 88 (Background radiation)
  Gamma (γ-rays), 44, *85*, 86, *113*, 128
Ray florets, 259
Reactance (electrical), 64
Reactants, 119
Reaction(s) (chemical), 119, 146-149, 156-163, 166
  Addition, 193
  Backward, see Reverse
  Condensation, 197
  Decomposition, 332 (Catabolism)
  Endothermic, 146
  Enthalpy change of (ΔH), 146
  Esterification, 197 (Condensation reaction)
  Exothermic, 146
  Forward, 162
  Heat of, see Enthalpy change of reaction
  Photochemical, 160
  Rates of, 160-161
  Reverse, 162
  Reversible, 162-163
  Standard enthalpy change of (ΔH^{θ}), 146
  Substitution, 192
  Synthesis, 332 (Anabolism)
Reaction(s) (nuclear), *85*, 92 (I), 94-95, 129
  Chain, 92 (Induced fission), 129 (Nuclear fission)
  D-T, *93*, 94 (Fusion reactor)
  Thermonuclear, 93 (Nuclear fusion)
Reactivity, 158-159
Reactivity series, 158, 211
Reactor, Nuclear, 94 (I)
  Advanced gas-cooled (AGR), *95*
  Fast/Fast breeder (FBR), 95
  Fission, 94
  Fusion, 94
  Pressurized water (PWR), *95*
  Thermal, 95
Reading error, 103
Read only memory (ROM) (computers), 111 (Memory)
Reagent, 119
Real cubic expansivity, 32
Real image, 49 (Image)
Real is positive sign convention, *49*
Receptacle, 256
Receptors, 305 (Sensory neurons), 307
  Pain, 311
Recessive, 325 (Genes), 326 (Sex linkage)
Reciprocal, 345
Recombinant DNA, *327* (Gene cloning)
Rectification, (Full-wave, Half-wave), 65
Rectilinear motion, see Linear motion
Rectrices (sing. rectrix), *267*
Rectum, *294*, 295 (Large intestine)
Rectus muscles, *313*
Red blood cells/corpuscles, 286
Red marrow, 281 (Bone marrow)
Redox, 148
Redox potential, 149
Redox series, 149
Red phosphorus, 182 (Phosphorus), 213
Reducing agent, 148
Reduction, 148-149
Reduction division, see First meiotic division
Refinery gas, 199
Refining, 198
Reflected ray, *47*
Reflected wave, 36 (Reflection)
Reflection, 36, 47-49
  Angle of (r), *47*
  Diffuse, 47
  Internal, 51 (Total internal reflection)
  Laws of, 47
  Regular, 47
  Total internal, 51
Reflex (Cranial, Spinal), 309 (Reflex actions)
Reflex actions, 309
Reflex arc, 309 (Reflex actions)
Reflux condenser, 223
Reforming, 198
Refracted ray, *50*
Refracted wave, 37 (Refraction)
Refraction, 37, 50-53, 312 (Lens)
  Angle of (r), *50*
  Laws of, 50
Refractive index (n), 37, 50 (Snell's law), *51*
  Absolute, 37 (Refractive index), *50*
Refrigerant, 345
Regular reflection, 46
Regulating factors, 336
Regulators, Growth, see Growth hormones
Relative atomic mass, 83, 138, 164-165, 212-213
Relative density, 24
Relative formula mass, see
  molecular mass
Relative isotopic mass, 13
Relative molecular mass, 13
Relative vapor density, 14
Relative velocity, 11
Relay, 75, 110
Relay neurons, see Associ
Releasing factor(s), 336 (Re
  factors)
  FSH, see FSHRF
  LH, see LHRF
Remiges (sing. remix), 26
Renal, 300
Renal arteries, *289*, 300 (K
Renal corpuscles, 301
Renal cortex, see Cortex (k
Renal medulla, see Medulla
Renal pelvis, see Pelvis (kid
Renal pyramids, *300* (Medul
Renal tubules, see Urinifero
Renal veins, *289*, 300 (Kidne
Rennin, 338 (Gastric juice)
Reprocessing plant, *94*
Reproduction, 320-323
  (animals), 276-277
  Asexual, 321
  (flowering plants), 256-25
  (humans), 316-319
  Sexual, 320
  Vegetative, 263
Reproductive system, 316-3
Reptiles, 266 (Scales), 27
Residual magnetism, 70 (Sof
Residue, 199
Resins, 345
  Natural, 345 (Resins)
  Synthetic, 345 (Resins)
Resistance (R), 62, 97
  Internal (r), 63
Resistance thermometer, 2
Resistivity (ρ), 62, 112
Resistor, 62, 110
  Carbon, 62 (Resistor)
  Light dependent (symbol
  Variable, 63, 110
Resolution, 108
Resonance, 17
Respiration, 272-273, 298-2
  Aerobic, 334
  Anaerobic, 334
  Cellular, see Internal respi
  External, 298 (I)
  Internal, 209, 254 (Photos
    255 (Compensation point
    298 (I), 334
  Tissue, see Internal respirat
Respiratory center, 299 (Bre
  303 (Medulla)
Respiratory enzymes, 333 (E
Respiratory system, 298-29
Resultant (vector quantities
  108 (Parallelogram rule
Resultant force, *6*

Relative formula mass, see **Relative molecular mass**
Relative isotopic mass, 138
Relative molecular mass, 138
Relative vapor density, 143
Relative velocity, 11
Relay, 75, 110
Relay neurons, see **Association neurons**
Releasing factor(s), **336 (Regulating factors)**
  FSH, see **FSHRF**
  LH, see **LHRF**
Remiges (sing. **remix**), **267**
Renal, 300
Renal arteries, *289*, 300 (**Kidneys**)
Renal corpuscles, 301
Renal cortex, see **Cortex** (kidneys)
Renal medulla, see **Medulla** (kidneys)
Renal pelvis, see **Pelvis** (kidneys)
Renal pyramids, *300* (**Medulla**)
Renal tubules, see **Uriniferous tubules**
Renal veins, *289*, 300 (**Kidneys**)
Rennin, 338 (Gastric juice)
Reprocessing plant, *94*
Reproduction, 320-323
  (animals), **276-277**
  Asexual, 321
  (flowering plants), **256-259**
  (humans), **316-319**
  Sexual, 320
  Vegetative, 263
Reproductive system, 316-317
Reptiles, 266 (**Scales**), 276 (**Eggs**), **341**
Residual magnetism, 70 (Soft)
Residue, 199
Resins, 345
  Natural, 345 (Resins)
  Synthetic, 345 (Resins)
Resistance (**R**), **62**, 97
  Internal (**r**), **63**
Resistance thermometer, 27
Resistivity (ρ), **62**, 112
Resistor, 62, 110
  Carbon, 62 (**Resistor**)
  Light dependent (symbol for), **110**
  Variable, 63, 110
Resolution, 108
Resonance, 17
Respiration, 272-273, 298-299
  Aerobic, 334
  Anaerobic, 334
  Cellular, see **Internal respiration**
  External, 298 (I)
  Internal, 209, 254 (**Photosynthesis**), 255 (**Compensation points**), 298 (**I**), **334**
  Tissue, see **Internal respiration**
Respiratory center, 299 (Breathing), 303 (**Medulla**)
Respiratory enzymes, 333 (Enzymes)
Respiratory system, 298-299
Resultant (vector quantities), **108 (Parallelogram rule)**
Resultant force, *6*
Resultant moment, 14 (Moment)
Reticulum, Endoplasmic, 239
Reticulum (stomach chamber), **271 (Rumen)**
Retina, 313
Retinol, see **Vitamin A**
Reverberation, 41
Reverse biased, 65 (Diode)
Reverse reaction, 162
Reversibility of light, Principle of, 49, *50*
Reversible reactions, 162-163
Rf value, 221 (Chromatography)
Rhabdom, 275 (Compound eye)
Rhenium (**Re**), 165, **173**, 213
Rheostat, *63*
Rhesus antigen, 287 (Rhesus factor)
Rhesus factor (**Rh factor**), **287**
Rhesus negative, 287 (Rhesus factor)
Rhesus positive, 287 (Rhesus factor)
Rhizome, 263
Rhodium (**Rh**), 165, **173**, *182*, 213
Rhodopsin, 337 (**Vitamin A**)
Rhombic sulfur, 136, **184**, 213
Rib(s), **278 (Rib cage)**
  False, 278 (Rib cage)
  Floating, 278 (Rib cage)
Rib cage, 278
Riboflavin, 339 (Vitamin B complex)
Ribonucleic acid, see **RNA**
Ribose, 324
Ribosomal RNA, 239 (Ribosomes)
Ribosomes, 239
Right-hand grip rule, 74
Right-hand rule, Fleming's, 78
Right lymphatic duct, *292*, **293 (Lymph vessels)**
Rigid, 10 (I)
Ring (atoms), *184*, **190 (I)**
  Benzene, 190 (Aromatic compounds)
Ring main (circuit), *61*
Ripple tank, 36
RNA, 239 (**Ribosomes**), **324** (**Nucleic acid**)
  Messenger, 239 (Ribosomes)
  Ribosomal, 239 (Ribosomes)
  Transfer, 239 (Ribosomes)
Rocket engine, 13
Rock salt, 168 (Sodium), 169 (Potassium)
Rodents, 341
Rods, 313 (Retina)
ROM, see **Read only memory**
Room temperature and pressure, see **r.t.p.**
Root(s) (hairs), *311*
Root(s) (plants), **244-245**
  Adventitious, 245
  Aerial, 245
  Fibrous, 245
  Lateral, 245 (Tap root)
  Primary, 245 (**Tap root**), **261** (**Radicle**)
  Prop, 245
  Secondary, see **Lateral roots**
  Tap, 245
Root(s) (spinal cord),
  Dorsal, 302 (Spinal cord)
  Motor, see **Ventral root**
  Sensory, see **Dorsal root**
  Ventral, 302 (Spinal cord)
Root(s) (teeth), **284**
Root canals, 284 (Pulp cavity)
Root cap, 245
Root hair(s), **245**
Root hair plexuses, see **Hair plexuses**
Root mean square (value), **61 (Alternating current)**
Root nodules, 235 (Nitrogen fixation)
Root pressure, 252
Rosette, 250
  Basal, 250
Rotary switch, *64*
Rotation, see **Rotational motion**
Rotational equilibrium, 15
Rotational kinetic energy, 9 (Kinetic energy)
Rotational motion, 10 (**I**), **11**
Roughage, 204 (**I**), **331**
Rough ER, 239 (Endoplasmic recticulum)
Round-bottomed flask, 224
Rounding (figures), **103**
Round window (ears), *314*
r.t.p., 139 (Molar volume)
Rubber, 201 (Natural polymers)
Rubidium (**Rb**), 164, **168**, 213
Rugae (sing. **ruga**), **295 (Stomach)**, 297 (**Gall bladder**), 300 (**Bladder**)
Rule,
  Dynamo, see **Fleming's right-hand rule**
  Fleming's right-hand, 78
  Maxwell's screw, 74
  Parallelogram, 108
  Right-hand grip, 74
Rumen, 271
Ruminants, *271*
Ruminate, *271*
Runner, see **Stolon**
Rust ($Fe_2O_3.xH_2O$), **174**
Rusting, 174 (**Rust**), **209** (**Corrosion**)
Ruthenium (**Ru**), 165, **173**, 213
Rutherford-Bohr atom, 82

S

Sac(s),
  Amniotic, *276* (**Amnion**), *319*
  Embryo, 258 (Ovules), **323 (Gamete production, female)**
  Pericardial, 290 (I)
  Pleural, 298 (Pleura)
  Pollen, 257 (Stamens)
  Synovial, 280
  Yolk, *276* (**Yolk**)
Saccharase, 338 (Intestinal juice)
Saccule/Sacculus, *314*, **315**
Sacral vertebrae, *278*

Sacrificial protection, 159
Sacrum, *278* (Sacral vertebrae)
Sal ammoniac, see Ammonium chloride
Saliva, *296*, 338
Salivary amylase, 338 (Saliva)
Salivary glands, 296 (Digestive glands), 338 (Saliva)
Salt(s), 153-155, 169 (Sodium chloride)
  Acid, 153
  Basic, 154
  Bile, 338 (Bile)
  Complex, 154
  Double, 154
  Normal, 153
  Rock, 168 (Sodium), 169 (Potassium)
Salt bridge, *158*
Saltpeter, see Potassium nitrate
  Chile, see Sodium nitrate
Samara, 262 (Achene)
Samarium (Sm), 164, 213
Sand, *123*, *177*
Sap, Cell, 238 (Vacuoles), 252 (Turgor)
Saphenous veins, Great, *289*
Saponification, 202 (Soap)
Saprophytes, 342
Sapwood, 247
Sarcolemma, *283*
Saturated (magnet), *71*
Saturated (solution), 145
Saturated compounds, 191
Savanna (biome), 232
Scalar quantity, 108
Scala tympani, *314*
Scala vestibuli, *314*
Scale(s) (animals), 266
  Placoid, see Denticles
Scale(s), Musical, 43
  Chromatic, *43*
  Diatonic, *43*
Scale(s) (temperature),
  Absolute temperature, 27, 143
  Celsius, 27, 143 (Absolute temperature scale)
  Fahrenheit, 27
  Thermodynamic temperature, see Absolute temperature scale
Scale, Vernier, 100
Scale (water), 207 (Temporary hardness)
Scale leaves, 263 (Bulb)
Scaler, 89 (Geiger counter)
Scan (ultrasound), *40*
Scandium (Sc), 164, 172, 213
Scanning, Ultrasound, *40*, 41 (Echo)
Scapula, *278*, *282*
Scavengers, 342
Schwann cells, 304 (Nerve fibers)
Scientific notation, 109
Scintillation(s), 44 (Phosphorescence), 90 (Scintillation counter)
Scintillation counter, 90
Scintillation crystal, 90 (Scintillation counter)
Scion, 263 (Grafting)
Sclera, *312*
Sclerites, 266 (Cuticle)
Sclerotic coat, see Sclera
Sclerotin, 266 (Cuticle)
Screening, see Shielding
Screw gauge, Micrometer, 101
Screw jack, 21
Screw rule, Maxwell's, 74
Scrotum, 316 (Testes)
Scum, 207
Scute/Scutum (pl. scuta), 266
Scutella (sing. scutellum), *267*
Sebaceous glands, 311
Sebum, 311 (Sebaceous glands)
Second (s), 96
Secondaries, see Secondary feathers
Secondary alcohols, 196, 197
Secondary bronchi, *298*, 299 (Bronchi)
Secondary bud, see Axillary bud
Secondary cell, 69, 159 (Cell)
Secondary circuit, *78*
Secondary coil, 79
Secondary colors, *55*
Secondary consumers, 235
Secondary cortex, 247 (Phelloderm)
Secondary feathers, 267 (Remiges)
Secondary phloem, 246 (Secondary thickening)
Secondary roots, see Lateral roots
Secondary sex characters, 318 (Puberty), 337 (Estrogen, Androgens)
Secondary thickening, 246
Secondary tissue, 246 (I)
Secondary xylem, 246 (Secondary thickening)
Second meiotic division, 322 (Meiosis), 323
Second order consumers, see Secondary consumers
Second order sensory neuron, *306*
Second polar body, 323 (Gamete production, female)
Secretin, 336
Sedentary, 342
Seeds, 260-261
Seed crystal, 135
Seed-leaf, see Cotyledon
Seedling, 260 (Germination)
Segmentation, 264
  Metameric, 264 (Segmentation)
Segments, 264 (Segmentation)
Segregation, Law of, 326
Selection, Natural, 237 (Genetic variation)
Selectively-permeable, see Semipermeable
Selenium (Se), 113, 165, 183, 213
Self-demagnetization, 71
Self-induction, 79
Self pollination, 259
Semen, 319 (Copulation)
Semicircular canals, *314*, 315
Semicircular ducts, *314*, 315
Semiconductors, 65
Semilunar valves, 291
Seminal fluid, 319 (Copulation)
Seminal vesicles, *316*
Seminiferous tubules, 316 (Testes)
Semipermeable, *329*
Sense organs, 307
Sensitivity,
  (animals), 274-275
  (humans), 306 (I)
  (plants), 251
Sensory areas, 302 (Brain), 303
Sensory nerves, 306 (Nerves)
Sensory neuron(s), 305, 306 (Afferent system)
  First order, *306*, *309*
  Second order, *306*, *309*
  Third order, *306*
Sensory root, see Dorsal root
Sepals, 256
Separating funnel, 224
Series,
  Activity, see Reactivity series
  (components), 64
  Decay, 87 (Radioactive decay), 128
  Electrochemical, 159
  Homologous, 191, 194-195
  Inner transition, 164-165, 172 (I)
  Radioactive, see Decay series
  Reactivity, 158, 211
  Redox, 149
  Transformation, see Decay series
Serrate, 250
Serum, 287
Sessile, *248*, 342
Seta, see Setae
Setae (sing. seta),
  (animals), 274
  (simple plants), *321* (Sporophyte)
Sex characters,
  Primary, 318 (Puberty)
  Secondary, 318 (Puberty), 337 (Estrogen, Androgens)
Sex chromosomes, 325, 326 (Sex linkage)
Sex hormones, 297 (Pineal gland), 336 (Estrogen, Progesterone, Androgens)
Sex linkage, 326
Sex-linked genes, 326 (Sex linkage)
Sexual intercourse, see Copulation
Sexual reproduction, 320
  (animals), 276 (I)
  (flowering plants), 258 (I)
  (humans), 318-319
Shadow, 46
Shaft,
  (bones), see Diaphysis
  (feathers), *267*
  (hairs), *311*
Shell (body covering), 266 (I, Cuticle, Carapace)
Shell, Electron, 83, 126
  Outer, 83 (Electron shells), 127
Shield (nuclear reactor), *95*
Shielding (magnetism)
Shinbone, see Tibia
Shock wave, *40*
Shoot, 244
Shortened structural
Short-night plants,
  251 (Photoperiodism
Short periods (element
  164 (Period)
Shoulderblade, see Scapula
Shunt, 77 (Ammeter)
Shutter (camera), *54*
Side chain (atoms),
  215 (8)
Sieve plates, 243 (Sieve
Sieve tubes, 243
Sign convention, 11
  Real is positive, *49*
Significant figures, 103
Silica, see Silicon dioxide
Silica gel, 221 (Desiccation
Silicates, 177
Silicon (Si), 112, 165,
Silicon dioxide (SiO_2),
Silicones, 177
Silicon(IV) oxide, see
Silk gland, *265*
Silver (Ag), 112, *160*,
  211, 213
Silver bromide (AgBr),
Silver chloride (AgCl),
Silver iodide (AgI), 188
Silver nitrate ($AgNO_3$),
  (Anions)
Simple barometer, *25*
Simple cell, 68
Simple cubic (particle arrangement), *137*
Simple leaves, 248 (I)
Simple microscope, 54
Sine, 345
Sinews, see Tendons
Single bond, 132, 214
Single touch (magnets), 71
Sintered, 70 (Ferromagnetic
Sinus(es), *307*
  Blood, 316 (Penis)
Sinusoidal (waveform), *67*
Siphon, 272
  Exhalant, 272 (Siphon)
  Inhalant, *265* (Mantle
  272 (Siphon)
  Mantle, *265*
SI units, 96-97
  Basic, 96
  Derived, 97
Skeletal muscles, 282,
Skeleton, 278-279
  Hydrostatic, 265 (Body
Skin, 266(I), 310-311
Skull, see Cranium
Slag, *174*, 175 (Copper)
Slaked lime, see Calcium
Slide callipers, Vernier,

Shielding (magnetism), **73**
Shinbone, see **Tibia**
Shock wave, *40*
Shoot, **244**
Shortened structural formula, **140**
Short-night plants, **251 (Photoperiodism)**
Short periods (element groups), **164 (Period)**
Shoulderblade, see **Scapula**
Shunt, **77 (Ammeter)**
Shutter (camera), *54*
Side chain (atoms), *190*, 214 (**h**), 215 (**8**)
Sieve plates, **243 (Sieve tubes)**
Sieve tubes, **243**
Sign convention, **11**
  Real is positive, *49*
Significant figures, **103**
Silica, see **Silicon dioxide**
Silica gel, 221 (**Desiccation**)
Silicates, **177**
Silicon (Si), 112, 165, **177**, 213
Silicon dioxide (SiO_2), **177**
Silicones, **177**
Silicon(IV) oxide, see **Silicon dioxide**
Silk gland, *265*
Silver (Ag), 112, *160*, 165, **173**, 211, 213
Silver bromide (AgBr), **188 (Bromides)**
Silver chloride (AgCl), *145*, *160*
Silver iodide (AgI), **188 (Iodides)**
Silver nitrate ($AgNO_3$), *145*, 218 **(Anions)**
Simple barometer, *25*
Simple cell, **68**
Simple cubic (particle arrangement), *137*
Simple leaves, **248 (I)**
Simple microscope, **54** (**Microscope**)
Sine, **345**
Sinews, see **Tendons**
Single bond, **132**, 214 (**a** and **1**)
Single touch (magnets), **71**
Sintered, **70 (Ferromagnetic)**
Sinus(es), *307*
  Blood, **316 (Penis)**
Sinusoidal (waveform), *61*
Siphon, **272**
  Exhalant, **272 (Siphon)**
  Inhalant, *265* **(Mantle siphon)**, **272 (Siphon)**
  Mantle, *265*
SI units, **96-97**
  Basic, **96**
  Derived, **97**
Skeletal muscles, **282**, 307, 308
Skeleton, **278-279**
  Hydrostatic, **265 (Body cavities)**
Skin, 266(I), **310-311**
Skull, see **Cranium**
Slag, *174*, 175 (**Copper**)
Slaked lime, see **Calcium hydroxide**
Slide callipers, **Vernier**, **100**
Slide projector, **55**
Sliding frictional force, see **Kinetic frictional force**
Sliding joints, see **Gliding joints**
Slow combustion, **208**
Small intestine, *294*, **295**
Smog, **210**
Smoke, **145**
Smooth ER, **239 (Endoplasmic reticulum)**
Smooth muscle, **283**
Snell's law, **50**
Soap, **202**, *207*
Soapless detergents, **203**
Social, **342**
Sodium (Na), *82*, *127*, *131*, *148*, 164, **168**, *196*, *207*, 211, 213, 219 (**Flame tests**)
Sodium aluminate ($NaAl(OH)_4$), *151*
Sodium aluminum silicate, see **Zeolite**
Sodium bicarbonate ($NaHCO_3$), *169*
Sodium bromide (NaBr), **188 (Bromine)**
Sodium carbonate (Na_2CO_3), **169**, *206*, 207, 217
Sodium chlorate ($NaClO_3$), **187**
Sodium chlorate(I), see **Sodium hypochlorite**
Sodium chlorate(V), see **Sodium chlorate**
Sodium chloride (NaCl), *131*, 141 (**Trivial name**), *145*, *153*, *156*, **169**, 187 (**Chlorine, Chlorides**), *344*
Sodium ethoxide (CH_3CH_2ONa), *196*
Sodium hydrogencarbonate, see **Sodium bicarbonate**
Sodium hydrogensulfate ($NaHSO_4$), *153*
Sodium hydroxide (NaOH), *151*, *155*, **169**, *202*, 216 (**Ethene**), 217, 219 (**Cations**)
Sodium hypochlorite (NaOCl), **187**, 344 (**Bleach**)
Sodium iodate ($NaIO_3$), **188 (Iodine)**
Sodium nitrate ($NaNO_3$), *145*, **169**, *182*
Sodium octadecanoate, *202*
Sodium stearate, *202*, *207*
Sodium sulfate (Na_2SO_4), *135*, *155*
Sodium sulfite (Na_2SO_3), **185 (Sulfites)**
Soft (magnets), **70**
Softeners, Water, **207**
Softening, Water, **207 (Permanent hardness)**
Soft palate, **294 (Pharynx)**, *307*
Soft water, **207 (Hard water)**
Solar eclipse, **46 (Eclipse)**
Solenocytes, see **Flame cells**
Solenoid, **74**
Solidification, **120**
Solid state, **5**, **120**
Solubility, **145**
Solubility curve, **145 (Solubility)**
Soluble, **145**
Solute, **144**, 329 (**I**)
Solution, **144 (I)**, 329 (**I**)
  Ammonia (NH_4OH), **181 (Ammonia)**, 218 (**Anions**), 219 (**Cations**)
  Aqueous, **144**
  Enthalpy change of/Heat of, **147**
  Molar, **139**
  Non-aqueous, 144 (**Aqueous solution**)
  Standard, **139**
Solvation, **144**
Solvay process, **169 (Sodium carbonate)**
Solvent, **144**, 329 (**I**)
  Aqueous, **144**
  Non-aqueous, 144 (**Aqueous solution**)
  Non-polar, **144**
  Organic, **345**
  Polar, **144**
Solvent extraction, **221**
Somatic afferent system, **307**
Somatic efferent system, **308**
Somatotropic hormone/Somatotropin, see **STH**
Sonar, **41 (Echo)**
Sonic boom, **40**
Sonic range, **40 (I)**
Sonometer, *43*
Sound(s), **40-43**
  Musical, **43**
  Speed of, **40**
Sound waves, **40-43**
South/South seeking pole, **70 (Pole)**
Spatula, **225**
Speaker, **76**, 110
Species, **340 (I)**
Specific gravity, see **Relative density**
Specific heat capacity (c), **30**, 112
Specific latent heat (l), **30**
Specific latent heat of fusion, **31**, 112, 146 (**Enthalpy change of reaction**)
Specific latent heat of vaporization, **31**, 146 (**Enthalpy change of reaction**)
Spectator ion, **141**
Spectrometer, **54 (Color)**
  Mass, 138 (**I**), **222 (Mass spectroscopy)**
Spectroscopy,
  Mass, **222**
  Nuclear magnetic resonance (n.m.r.), **222**
Spectrum, **345**
  Electromagnetic, *44*, **113**
  Nuclear magnetic resonance (n.m.r.), **222**
  Visible light, **54**
Speed, **10**
  Average, **10 (Speed)**
  Instantaneous, **10 (Speed)**
  Sub- and Supersonic, **40**
  Uniform, **10 (Speed)**
  Wave, **35**
Speed of light, **113**
Speed of sound, **40**

Sperm, 276, 316 (**Testes**), 319 (**Copulation**), **320** (**Gametes**), 323 (**I**), 337 (**FSH**)
Spermatheca, **277**
Spermatids, **323** (**Gamete production, male**)
Spermatozoa (sing. **spermatozoon**)/**Spermatozooids**, see **Sperm**
Sperm ducts, *316*
Spherical aberration, **49**
Sphincter, *294*
  Anal, **295** (**Large intestine**)
  Cardiac, *294*, **295**
  External urinary, **300** (**Bladder**)
  Gastroesophageal, see **Cardiac sphincter**
  Internal urinary, **300** (**Bladder**)
  Pyloric, *294*, **295**
Sphincter of Oddi, *296*, 337 (**CCK**)
Spinal canal, see **Neural canal**
Spinal column, see **Vertebral column**
Spinal cord, **302**
Spinal nerves, **302** (**Spinal cord**)
Spinal reflexes, **309** (**Reflex actions**)
Spindle (cell division), 241 (**Metaphase**), 322 (**Metaphase**)
Spindle, Muscle, **283**
Spindle microtubules (cell division), **241** (**Metaphase**), 322 (**Metaphase**)
Spine,
  (humans), see **Vertebral column**
  Neural, see **Spinous process**
  (plants), **249**
Spinous process, *279*
Spiracle, **273**
Spiral (leaves), **250**
Spleen, **293**
Splenic artery, *289*
Splenic vein, *289*
Spongy bone, **281**
Spongy cells, **248** (**Spongy layer**)
Spongy layer, **248**
Spontaneous fission, **92**
Sporangium, *321*
Spores, **321** (**Sporulation**)
Sporophyte, **321** (**Alternation of generations**)
Sporulation, **321**
Spring balance, *22*
Spring wood, **246** (**Annual rings**)
Spurred flower, **259**
Spurs, **259** (**Spurred flower**)
Stable equilibrium, **15**
Stainless steel, **174** (**Steel**)
Stamens, **257**
Staminate, **256** (**I**)
Stand(s) (equipment), **225**
Standard, **259** (**Pea flower**)
Standard cell, **69**
Standard enthalpy change of reaction ($\Delta H^{\ominus}$), **146**
Standard solution, **139**
Standard temperature and pressure, see **s.t.p.**
Standing wave, see **Stationary wave**
Stapes, **314** (**Middle ear**)
Starch, 201 (**Natural polymers**), **204**, **331**, 338 (**Saliva**), 339 (**Polysaccharides**)
  Animal, see **Glycogen**
State(s), Energy, **84** (**I**)
  Ground, **84** (**I**)
State(s), Oxidation, **149**
State(s), Physical, **4-5**, 30-31, **120-121**
  Changes of, **30**, **120-121**
  Equation of, see **Ideal gas equation**
  Gaseous, **5**, **120**
  Liquid, **5**, **120**
  Solid, **5**, **120**
States of matter, see **Physical states**
State symbols, **141**
Static electricity, **56-57**
Static equilibrium, **15** (**Equilibrium**)
Static frictional force, **7**
Static friction, Coefficient of, 7 (**Coefficient of friction**)
Stationary phase, **221** (**Chromatography**)
Stationary wave, **43**
Statocysts, **275**
Statoliths, **275** (**Statocysts**)
Steam, Superheated, **345**
Steam generator, *94*, *95*
Steam point, **26** (**Fixed point**)
Stearic acid, see **Octadecanoic acid**
Steel, 112, **174**
  Stainless, **174** (**Steel**)
Stem,
  Brain, **303**
  (plants), **244**
Stensen's duct, *296*
Step-down transformer, **79**
Step-up transformer, **79**
Stereochemical formula, **140**, 191 (**Stereochemistry**)
Stereochemistry, **191**
Stereoisomers, **191**
Sternum, *269*, *278*
STH, **336**
Stigma, see **Stigmata**
Stigmata (sing. **stigma**),
  (animals), **273** (**Spiracle**)
  (flowers), **257**, 258 (**Pollination**)
Stipule, **249**
Stirrup, see **Stapes**
Stock, **263** (**Grafting**)
Stolon, **263**
Stomach, *294*, **295**
Stomata (sing. **stoma**), **249**
Storage cell, see **Secondary cell**
s.t.p., **143**
Straight chain (atoms), *190*
Straight wavefronts, *34*
Strain, **22** (**Hooke's law**)
Strain energy, see **Elastic potential energy**
Strata (sing. **stratum**), **310** (**Epidermis**)
Stratum basale, *310* (**Stratum germinativum**)
Stratum corneum, *310*
Stratum germinativum, *310*
Stratum granulosum, *310*
Stratum spinosum, *310* (**Stratum germinativum**)
Stress, **22** (**Hooke's law**)
  Breaking, *23*
  Yield, **22** (**Yield point**)
Striated, **283** (**Striated muscle**)
Striated muscle, **283**
Stridulation, **275**
Striped muscle, see **Striated muscle**
Strong acid, **152**
Strong base, **152**
Strong electrolyte, *66*, *156*
Strontium (Sr), *129*, 164, **170**, 213
Structural formula,
  3-dimensional, see **Stereochemical formula**
  Full, see **Displayed formula**
  Shortened, **140**
Structural isomers, **191**
Structural proteins, **330** (**Proteins**)
Structure, Atomic, **82-83**, 126-127
Style (flowers), **257**
Styrene, **201** (**Polystyrene**)
Subatomic particles, **82** (**I**), **126** (**I**)
Sub-classes, **340** (**I**)
Subclavian arteries, *289*, *290*
Subclavian veins, *289*, *290*, 293 (**Lymph vessels**)
Subcritical mass, **93** (**Critical mass**)
Subcutaneous, 311 (**Subcutaneous layer**)
Subcutaneous layer, **310**, *311*
Suberin, **247** (**Phellem**)
Suberization, **247** (**Phellem**)
Sub-kingdoms, **340** (**I**)
Sublimation, **30**, **121**
Sublingual glands, *296*
Submandibular glands/Submaxillary glands, *296*
Sub-phyla, **340** (**I**)
Subsonic speed, **40**
Substitution reaction, **192**
Subtractive mixing, **55** (**Color mixing**)
Successional community, *233*
Succession, Ecological, **233**
Succus entericus, see **Intestinal juice**
Sucrase, **338** (**Intestinal juice**)
Sucrose ($C_{12}H_{22}O_{11}$), **204**, **338** (**Intestinal juice**), 339 (**Disaccharides**)
Sudoriferous glands, see **Sweat glands**
Sulfate(IV) compounds, see **Sulfites**
Sulfate(VI) compounds/Sulfates, *154*, **185**, 218 (**Anions**)
Sulfides, **184** (**I**), 218 (**Anions**)
Sulfites, **185**, 218 (**Anions**)
Sulfur (S), *121*, *136*, *148*, *155*, 165, 183, **184**, 213
  Alpha (α-sulfur), see **Rhombic sulfur**
  Beta (β-sulfur), see **Monoclinic sulfur**
  Flowers of, **184**
  Monoclinic, *136*, **184**, 213
Sulfur (cont'd)
  Orthorhombic, see
  Plastic, *184*
  Rhombic, *136*, **184**,
Sulfur(IV) oxide, see
Sulfur(VI) oxide, see
Sulfur dioxide (SO_2), *151*, 210 (**Smog**, **Acid rai**
Sulfuric acid (H_2SO_4), 210 (**Acid rain**), **216**, 218 (**Nitrate anion**), cation), 344 (**Dehydra**
  Fuming ($H_2S_2O_7$), **185** process)
Sulfuric(IV) acid, see
Sulfuric(VI) acid, see
Sulfurous acid (H_2SO_3),
Sulfur trioxide (SO_3), **185**
Summer wood, **246**
Superconductor, **179** (**Buckminsterf**
Superficial fascia, see layer
Superheated (liquid), **90** chamber)
Superior (ovary), **257** **flower, Perigynous**
Superior articular
Superior mesenteric
Superior mesenteric
Superior vena cava, *290*,
Superposition, **38** superposition)
Supersaturated,
  (solution), **145**
  (vapor), **90** (**Cloud**
Supersonic speed, **40**
Suprarenal glands, see
Surface area (measure
Surface catalyst, **161**
Surface density of charge,
Surface tension, **23**
Surfactants, **203**
Susceptibility, **70**
Suspension, **145**
Suspensory ligament, **312**
Sutures, **278** (**Cranium**)
Sweat, **311** (**Sweat glands**
Sweat duct, 311 (**Sweat**
Sweat glands, **311**
Swim bladder, **269**
Switch, **64**, 110
  Changeover, *64*
  Double pole, *64*
  Rotary, *64*
Symbol(s),
  Chemical, **122**, 212-213,
  (quantities and units), 110, 113
  State, **141**
Symbol equation, *98*
Symmetry (**Bilateral**,
Sympathetic division 308

Sulfur (cont'd)
  Orthorhombic, see **Rhombic sulfur**
  Plastic, *184*
  Rhombic, *136*, **184**, 213
Sulfur(IV) oxide, see **Sulfur dioxide**
Sulfur(VI) oxide, see **Sulfur trioxide**
Sulfur dioxide (SO_2), *151*, **185**, 210 (**Smog, Acid rain**)
Sulfuric acid (H_2SO_4), *150*, *155*, **185**, 210 (**Acid rain**), 216, 217, 218 (**Nitrate anion**), 219 (**Calcium cation**), 344 (**Dehydrating agent**)
  Fuming ($H_2S_2O_7$), **185** (**Contact process**)
Sulfuric(IV) acid, see **Sulfurous acid**
Sulfuric(VI) acid, see **Sulfuric acid**
Sulfurous acid (H_2SO_3), *150*, **185**
Sulfur trioxide (SO_3), **185**
Summer wood, **246** (**Annual rings**)
Superconductor, **179** (**Buckminsterfullerene**)
Superficial fascia, see **Subcutaneous layer**
Superheated (liquid), **90** (**Bubble chamber**)
Superior (ovary), **257** (**Hypogynous flower, Perigynous flower**)
Superior articular processes, *279*
Superior mesenteric artery, *289*
Superior mesenteric vein, *289*
Superior vena cava, *290*, **291**
Superposition, **38** (**Principle of superposition**)
Supersaturated,
  (solution), **145**
  (vapor), **90** (**Cloud chamber**)
Supersonic speed, **40**
Suprarenal glands, see **Adrenal glands**
Surface area (measurement of), **101**
Surface catalyst, ***161***
Surface density of charge, **57**
Surface tension, **23**
Surfactants, **203**
Susceptibility, **70**
Suspension, **145**
Suspensory ligament, **312** (**Lens**)
Sutures, **278** (**Cranium**)
Sweat, **311** (**Sweat glands**)
Sweat duct, **311** (**Sweat glands**)
Sweat glands, **311**
Swim bladder, **269**
Switch, **64**, 110
  Changeover, *64*
  Double pole, *64*
  Rotary, *64*
Symbol(s),
  Chemical, **122**, 212-213, 226-227
  (quantities and units), 96-97, **99**, 110, 113
  State, **141**
Symbol equation, *98*
Symmetry (**Bilateral, Radial**), **264**
Sympathetic division (nervous system), **308**
Synapses, **305**
Synaptic cleft/gap, **305** (**Synapses**)
Synaptic knob, *305*
Synovial capsule, see **Synovial sac**
Synovial fluid, **280** (**Synovial sac**)
Synovial joints, **280** (**Synovial sac**)
Synovial membrane, **280** (**Synovial sac**)
Synovial sac, **280**
Synthesis, **122**, **332** (**Anabolism**)
  Direct, **155**
Synthesis reactions, **332** (**Anabolism**)
Synthetic detergents, see **Soapless detergents**
Synthetic diamonds, **178** (**Diamond**)
Synthetic polymers, **201**
Synthetic resins, **345** (**Resins**)
Syrinx (pl. **syringes**), **275**
System, **238** (**I**), **345**
System (chemistry),
  Closed, **162**
  Open, **162** (**Closed system**)
System(s) (human body),
  Afferent, **306-307**
  Arterial, **288** (**Arteries**)
  Autonomic nervous, 303 (**Hypothalamus**), **308**
  Cardiovascular, **290** (**I**)
  Central nervous, **302-303**
  Circulatory, **288-289**
  Digestive, **294-295**
  Efferent, **308-309**
  Integumentary, **310** (**I**)
  Lymphatic, **293**
  Nervous, **302-309**
  Peripheral nervous, 304 (**I**), **306** (**I**)
  Reproductive, **316-317**
  Respiratory, **298-299**
  Somatic afferent, **307**
  Somatic efferent, **308**
  Urinary, **300-301**
  Vascular, see **Circulatory system**
  Venous, **288** (**Veins**)
  Visceral afferent, **307**
  Visceral efferent, see **Autonomic nervous system**
System(s) (physics),
  Deflection (oscilloscope), *81*
  Pulley, *20*, **21**
Systematic name, **141**
Systole phase, **291** (**Cardiac cycle**)

T

Tactile, **306**
Tagma (pl. **tagmata**), **264** (**Segmentation**)
Tangent, **345**
Tannin, **255** (**Pigments**)
Tantalum (**Ta**), 112, 165, **173**, 213
Tap funnel, **224**
Tap root, **245**
Target cells, **336** (**Hormones**)
Target DNA, *327* (**Gene cloning**)
Target organs, **336** (**Hormones**)
Tarnish, **345**
Tarsals, *278*
Tarsus, *278* (**Tarsals**)
Taste buds, **307** (**Tongue**)
Taxonomy, **340** (**I**)
  Classical, **340** (**I**)
TCT, **336**
Tear glands, see **Lachrymal glands**
Teat pipette, see **Dropping pipette**
Technetium (**Tc**), 165, **173**, 213
Tectorial membrane, **314** (**Cochlear duct**)
Teeth, **284-285**
  Baby, see **Deciduous teeth**
  Carnassial, **270**
  Deciduous, **284** (**I**)
  Dog/Eye, see **Canines**
  Milk, see **Deciduous teeth**
  Permanent, **284** (**I**)
  Wisdom, **285**
Telescope, **55**
Television, **81**
Tellurium (**Te**), 112, 165, **183**, 213
Telophase,
  (meiosis), **323**
  (mitosis), **241**
Telson, *274*
Temperate grassland (biome), **232**
Temperature, **26-27**, 96, 142-143
  Critical, **5** (**Gas, Vapor**), **120** (**Gaseous state**)
  Room, **139** (**Molar volume**)
  Standard, **143** (**s.t.p.**)
  Transition, **136**
Temperature gradient, **28** (**Conductivity**)
Temperature scale,
  Absolute, **27**, **143**
  Celsius, **27**, 143 (**Absolute temperature scale**)
  Fahrenheit, **27**
  Thermodynamic, see **Absolute temperature scale**
Temporary hardness (water), **207**
Temporary magnets, **70** (**Soft**)
Tendons, **281**
Tendril, **249**
Tension, **7**
  Surface, **23**
Tentacles, 270 (**Cnidoblasts**), **274**
Terbium (**Tb**), 165, **213**
Terminal(s), **60** (**Electromotive force**), 110, **159** (**Cell**)
Terminal bronchioles, **299** (**Bronchioles**)
Terminal bud, **244**
Terminal velocity, **19**
Ternate, **250**
Territorial, **342**
Territory, **342** (**Territorial**)
Tertiary alcohols, **197**
Tertiary bronchi, *298*, **299** (**Bronchi**)

- **Tertiary consumers, 235**
- **Test(s), 218-219**
  - **Boiling point, 221**
  - **Flame,** *219,* **222**
  - **Melting point, 221**
- **Testa, 261**
- **Testes** (sing. **testis**)/**Testicles, 316,** *318*
- **Testosterone, 336 (Androgens)**
- **Test tube, 225**
- **Test tube holder, 225**
- **Test tube rack, 225**
- **Tetraammine copper(II) sulfate ((Cu(NH$_3$)$_4$)SO$_4$),** *154*
- **Tetrachloromethane (CCl$_4$),** *124, 144, 193*
- **Tetrad, 322 (Prophase)**
- **Tetraethyl-lead (Pb(OC$_2$H$_5$)$_4$),** *210*
- **Tetragonal** (basic crystal shape), *136*
- **Tetrahedral** (molecule), *133*
- **Tetravalent** (covalency), *133*
- **Thalamus, 303**
- **Thallium (Tl),** 165, **176,** 213
- **Theory,**
  - **Bronsted-Lowry, 151**
  - **Collision, 160**
  - **Dalton's atomic, 124**
  - **Kinetic, 5, 123**
  - **Quantum, 84**
- **Theory of electrolysis, Ionic, 66, 156**
- **Theory of magnetism, Domain, 71**
- **Thermal conduction, 28**
- **Thermal conductivity, 28,** 112
- **Thermal dissociation, 162 (Dissociation)**
- **Thermal energy,** see **Internal energy**
- **Thermal equilibrium, 28 (I)**
- **Thermal images, 45 (Infra-red radiation)**
- **Thermal pollution, 210**
- **Thermal reactor, 95**
- **Thermionic emission, 80 (Electron gun)**
- **Thermistor,** 27, **65,** 110
- **Thermochemistry, 146 (I)**
- **Thermocouple, 27**
- **Thermodynamic temperature scale,** see **Absolute temperature scale**
- **Thermometer, 26, 225**
  - **Clinical,** *26*
  - **Digital, 27**
  - **Liquid-in-glass, 26**
  - **Maximum, 27**
  - **Minimum, 27**
  - **Resistance, 27**
- **Thermometric property, 26 (Thermometer)**
- **Thermonuclear reactions, 93 (Nuclear fusion)**
- **Thermopile, 29**
- **Thermoplastics, 201 (Plastics)**
- **Thermosetting plastics,** *200,* **201 (Plastics)**
- **Thermostat,** *32*
- **Thiamine, 339 (Vitamin B complex)**
- **Thighbone,** see **Femur**
- **Thigmotropism,** see **Haptotropism**
- **Third gas law,** see **Pressure law**
- **Third order consumers,** see **Tertiary consumers**
- **Third order sensory neuron,** *306*
- **Thistle funnel, 224**
- **Thoracic duct,** *292,* **293 (Lymph vessels)**
- **Thoracic vertebrae,** *278, 279*
- **Thorax, 264 (Segmentation), 278 (Rib cage)**
- **Thorium (Th),** *87, 128,* **164, 213**
- **Thread cells,** see **Cnidoblasts**
- **Thrombin, 287 (Clotting)**
- **Thrombocytes,** see **Platelets**
- **Thromboplastin, 287 (Clotting)**
- **Thulium (Tm), 165, 213**
- **Thymine, 324**
- **Thymus gland, 293**
- **Thyrocalcitonin,** see **TCT**
- **Thyroid gland, 297,** 337 **(TSH, Thyroxin, TCT)**
- **Thyroid-stimulating hormone/ Thyrotropin,** see **TSH**
- **Thyroxin,** *188,* **336,** 337 **(TSH)**
- **Tibia,** *278, 280, 281*
- **Tibial artery (Arterior, Posterior),** *289*
- **Timbre, 43 (Modes of vibration)**
- **Time (t), 96**
- **Timebase** (oscilloscope control), *81*
- **Tin (Sn),** 112, 165, **177,** 213
- **Tincture of iodine,** *188*
- **Tissue, 238 (I)**
  - **Adipose, 310 (Subcutaneous layer)**
  - **Connective, 280**
  - **Erectile, 316 (Penis),** 317 **(Vulva)**
  - **Lymphatic/Lymphoid, 293 (Lymphoid organs)**
  - **Osseous,** see **Bone** (tissue)
  - **Primary** (plant), *242*
  - **Secondary** (plant), **246 (I)**
  - **Vascular** (plant), **242**
- **Tissue fluid, 292**
- **Tissue respiration, see Internal respiration**
- **Titanium (Ti),** 165, **172,** 213
- **Titer, 222 (Volumetric analysis)**
- **Titration, 222 (Volumetric analysis)**
- **Tocopherol,** see **Vitamin E**
- **Tongs, 225**
- **Tongue,** *293,* **307**
- **Tonsil(s), 293**
  - **Lingual, 293 (Tonsils)**
  - **Palatine, 293 (Tonsils)**
  - **Pharyngeal, 293 (Tonsils)**
- **Top pan balances, 225**
- **Toppling, 14**
- **Torque,** see **Moment**
- **Torricellian vacuum,** *25*
- **Total eclipse,** *46*
- **Total internal reflection, 51**
- **Toxic,** *210*
- **Trabeculae** (sing. **trabecula**), **281 (Spongy bone)**
- **Trace elements, 331 (Minerals)**
- **Tracer, 91 (Radioactive tracing), 129 (Radioactive tracing)**
- **Trachea,** see **Tracheae**
- **Tracheae** (sing. **trachea**),
  - (animals), **273**
  - (humans), **298**
  - (plants), see **Vessels**
- **Tracheoles, 273 (Tracheae)**
- **Tracing, Radioactive,** 91, **129**
- **Traditional name, 141**
- **Train, Maglev, 75**
- **Trait, 325 (Genes)**
- ***Trans* but-2-ene,** *191*
- **Transfer RNA, 239 (Ribosomes)**
- **Transformation,**
  - (atomic/nuclear), *85*
  - (energy), see **Conversion**
- **Transformation series,** see **Decay series**
- **Transformer(s),** 79, 110
  - **Step-down, 79**
  - **Step-up, 79**
- **Transistor(s),** 65, **111**
  - **Field effect** (symbol for), **110**
  - **Npn,** *65,* 110
  - **Pnp,** *65,* 110
- **Transition** (atomic/nuclear), *85*
- **Transition metals,** 165, **172-175**
- **Transition series, Inner, 164-165, 172 (I)**
- **Transition temperature, 136**
- **Translation,** see **Translational motion**
- **Translational kinetic energy, 9 (Kinetic energy)**
- **Translational motion, 10 (I)**
- **Translocation, 252 (I)**
- **Transmutation,** see **Transformation** (atomic/nuclear)
- **Transpiration,** *206,* **252**
- **Transpiration stream, 252**
- **Transverse processes,** *279*
- **Transverse waves, 34**
- **Traveling waves, 34(I)**
- **Trends, 166 (I)**
- **Triatomic,** *124*
- **Tribasic,** *153*
- **Triceps,** *282*
- **Triclinic** (basic crystal shape), *136*
- **Tricuspid valve, 291 (Atrioventricular valves)**
- **Trifoliate, 250**
- **Trigonal/Trigonal planar** (molecule), *133*
- **Trimer, 125**
- **Triol,** *197*
- **Triple bond, 132, 214 (c and 3)**
- **Triple point, 96 (Kelvin)**
- **Tripod, 225**
- **Tritiated water, 167 (Tritium)**
- **Tritium (T),** *93,* **94 (Fusion reactor), 167**
- **Trivalent,**
  - (covalency), *133*
  - (electrovalency), *131*
- **Trivial name, 141**
- **tRNA,** see **Transfer RNA**
- **Trophic level, 235**
- **Tropical forest** (biome), **232**
- **Tropic hormones, 297 (Pit**
- **Tropism (Negative, Positive**
- **Trough** (equipment), **225**
- **Troughs** (waves), **34**
- **True fruit, 262 (I)**
- **True weightlessness, 19**
- **Truth tables, 111**
- **Trypsin, 338 (Pancreatic ju**
- **Trypsinogen, 338 (Note 2**
- **TSH, 336**
- **Tube**(s) (biology),
  - **Eustachian/Auditory,** *314*
  - **Fallopian,** *317,* 319 **(Fertil**
  - **Pollen, 258 (Pollination**
  - **Sieve, 243**
  - **Uterine,** see **Fallopian tubes**
- **Tube**(s) (chemistry/physics)
  - **Boiling, 225**
  - **Capillary,** *23, 24,* **26 (Liqui thermometer)**
  - **Delivery, 223**
  - **Discharge, 80**
  - **Electron, 80 (I)**
  - **Fluorescent, 80** (**Discharge**
  - **Geiger-Müller, 89 (Geiger**
  - **Ignition, 225**
  - **Maltese cross, 80**
  - **Test, 225**
  - **X-ray, 80**
- **Tube nucleus, 258 (Pollinati**
- **Tuber, 263**
- **Tubular reabsorption, 301**
- **Tubular secretion, 301**
- **Tubule**(s),
  - **Collecting,** see **Collecting d**
  - **Distal convoluted, 301 (Uri tubules)**
  - **Malpighian,** 265 **(Hemocoel**
  - **Proximal convoluted, 301 (Uriniferous tubules**
  - **Renal,** see **Uriniferous tub**
  - **Seminiferous, 316 (Testes**
  - **Uriniferous, 301,** 337 **(ADH, Aldosterone)**
- **Tundra** (biome), **232**
- **Tungsten (W),** 112, 165, **173,**
- **Turbine,** *9,* 94
- **Turgid, 253 (Turgor)**
- **Turgor, 253**
- **Turgor pressure, 253 (Turgor**
- **Turning forces, 14-15**
- **Turns ratio, 79**
- **Tympanic organs/Tympani** (sin **tympanum), 275**
- **Tympanic canal,** see **Scala ty**
- **Tympanic cavity,** see **Middle e**
- **Tympanic membrane,** see **Ear**

Trivalent,
  (covalency), *133*
  (electrovalency), *131*
Trivial name, **141**
tRNA, see **Transfer RNA**
Trophic level, **235**
Tropical forest (biome), **232**
Tropic hormones, **297** (**Pituitary gland**)
Tropism (**Negative**, **Positive**), **251** (**I**)
Trough (equipment), **225**
Troughs (waves), **34**
True fruit, **262** (**I**)
True weightlessness, **19**
Truth tables, **111**
Trypsin, **338** (**Pancreatic juice**)
Trypsinogen, **338** (**Note 2**)
TSH, **336**
Tube(s) (biology),
  Eustachian/Auditory, *314*
  Fallopian, *317*, 319 (**Fertilization**)
  Pollen, **258** (**Pollination**)
  Sieve, **243**
  Uterine, see **Fallopian tubes**
Tube(s) (chemistry/physics),
  Boiling, **225**
  Capillary, *23*, *24*, **26** (**Liquid-in-glass thermometer**)
  Delivery, **223**
  Discharge, **80**
  Electron, **80** (**I**)
  Fluorescent, **80** (**Discharge tube**)
  Geiger-Müller, **89** (**Geiger counter**)
  Ignition, **225**
  Maltese cross, **80**
  Test, **225**
  X-ray, **80**
Tube nucleus, **258** (**Pollination**)
Tuber, **263**
Tubular reabsorption, **301**
Tubular secretion, **301**
Tubule(s),
  Collecting, see **Collecting duct**
  Distal convoluted, **301** (**Uriniferous tubules**)
  Malpighian, 265 (**Hemocoel**), **273**
  Proximal convoluted, **301** (**Uriniferous tubules**)
  Renal, see **Uriniferous tubules**
  Seminiferous, **316** (**Testes**)
  Uriniferous, **301**, 337 (**ADH**, **Aldosterone**)
Tundra (biome), **232**
Tungsten (**W**), 112, 165, **173**, 213
Turbine, *9*, 94
Turgid, **253** (**Turgor**)
Turgor, **253**
Turgor pressure, **253** (**Turgor**)
Turning forces, **14-15**
Turns ratio, **79**
Tympanic organs/Tympani (sing. **tympanum**), **275**
Tympanic canal, see **Scala tympani**
Tympanic cavity, see **Middle ear**
Tympanic membrane, see **Eardrum**

U

Ulna, *278*, *282*
Ultrasonic waves, **40** (**Ultrasound**)
Ultrasound, **40**
Ultrasound scanning, *40*, 41 (**Echo**)
Ultraviolet radiation (**UV radiation**), **44**, *85*, *113*, 210 (**Ozone depletion**)
Umbellifer, **259**
Umbels, **259** (**Umbellifer**)
Umbilical cord, *319*
Umbra, **46** (**Shadow**)
Ungulates, **341**
Unguligrade, **269**
Unicellular, **238** (**I**)
Unified atomic mass units (**u**), **83** (**Relative atomic mass**)
Uniform acceleration, **11** (**Acceleration**)
Uniform circular motion, **17**
Uniform speed, **10** (**Speed**)
Uniform velocity, **10** (**Velocity**)
Units,
  International system/SI (**Basic**, **Derived**), **96-97**
  Unified atomic mass (**u**), **83** (**Relative atomic mass**)
Universal indicator, **152**
Unsaturated compounds, **191**
Unstable equilibrium, **15**
Upper fixed point, **26** (**Fixed point**)
Upper motor neuron, *308*, *309*
Upward displacement of air, **216** (**Chlorine**), **217** (**Oxygen**)
Uracil, **324**
Uranium (**U**), 92 (**Induced fission**), 94 (**Fission reactor**), 95 (**Thermal reactor**, **Fast reactor**), *128*, *129*, **164**, **213**
Urea, **300**
Ureters, **300**, *316*
Urethra, **300**, *316*, *319*
Urethral orifice, *300*, *317*
Urinary bladder, see **Bladder**
Urinary sphincters (**External**, **Internal**), **300** (**Bladder**)
Urinary system, **300-301**
Urination, **300** (**Urethra**)
Urine, **300**, *301*
Uriniferous tubules, **301**, 337 (**ADH**, **Aldosterone**)
Urochordates, **341**
Uropygial gland, *267* (**Uropygium**)
Uropygium, *267*
Useless load, **20**
Uterine tubes, see **Fallopian tubes**
Uterus, **317**, 318 (**Menstrual cycle**), 337 (**LH**, **Oxytocin**, **Estrogen**)
Utricle/Utriculus, *314*, **315**
UV radiation, see **Ultraviolet radiation**

V

Vacuole(s), **238**, *239*, 329 (**Pinocytosis**)
  Contractile, *268*, **273**
  Food, *268*
Vacuum flask, **29**
Vagina, **317**, 319 (**Copulation**)
Vaginal orifice, *317*
Valency, see **Covalency/Electrovalency**
Valency electron, **130**
Valve(s),
  Aortic, **291** (**Semilunar valves**)
  Atrioventricular/AV, **291**
  Bicuspid, **291** (**Atrioventricular valves**)
  Mitral, **291** (**Atroventricular valves**)
  Pulmonary, **291** (**Semilunar valves**)
  Pyloric, see **Pyloric sphincter**
  Semilunar, **291**
  Tricuspid, **291** (**Atrioventricular valves**)
Vanadium (**V**), 165, **172**, 213
Vanadium pentoxide (V_2O_5), **172** (**Vanadium**), 185 (**Contact process**)
Van de Graaff generator, **57**
Van der Waals' forces, **134**
Vane (feathers), *267*
Vapor, **5**, **120** (**Gaseous state**)
Vapor density, Relative, **143**
Vaporization, **30**, **121**
  Molar enthalpy change of/Molar heat of, 146 (**Enthalpy change of reaction**), **147**
  Specific latent heat of, **31**, 146 (**Enthalpy change of reaction**)
Variable(s), **345**
  Dependent, *98*
  Independent, *98*
Variable capacitor, **59**, 110
Variable resistor, **63**, 110
Vascular, 280 (**Connective tissue**), **289**
Vascular bundles, **242** (**Vascular tissue**)
Vascular cylinder, **246**
Vascular plants, **242-243**
Vascular system, see **Circulatory system**
Vascular tissue, **242**
Vas deferens (pl. **vasa deferentia**), see **Sperm ducts**
Vasopressin, see **ADH**
Vater, Ampulla of, *296*
Vector quantity, **108**
Vegetative reproduction/propagation, **263**
Vein(s) (blood vessels), **288**, *289*, **291**
  Brachial, *289*
  Brachiocephalic, *290*
  Cephalic, *289*
  Common iliac, *290*
  External jugular, *290*
  Femoral, *289*

Vein(s) (cont'd)
  Gastric, *289*
  Gonadal, *289*
  Great saphenous, *289*
  Hepatic, *289*
  Hepatic portal, *289*, **297** (**Liver**)
  Inferior mesenteric, *289*
  Internal jugular, *290*
  Pancreatic, *289*
  Pulmonary, *290*, **291**
  Renal, *289*, 300 (**Kidneys**)
  Splenic, *289*
  Subclavian, *289*, *290*, 293 (**Lymph vessels**)
  Superior mesenteric, *289*
Vein(s) (leaves), **248**
Velocity (v), **10**, 97
  Angular, *17*
  Average, **10** (**Velocity**)
  Escape, **19**
  Instantaneous, **10** (**Velocity**)
  Relative, **11**
  Terminal, **19**
  Uniform, **10** (**Velocity**)
Velocity profiles, *23*
Velocity ratio (**V.R.**), **20**
Vena cava (pl. **vena cavae**),
  Inferior, *289*, *290*, **291**
  Superior, *290*, **291**
Venation (leaves), *248*
Venous system, **288** (**Veins**)
Ventilation, **298** (**I**)
Ventral, **269**
Ventral fin, *269*
Ventral root, **302** (**Spinal cord**)
Ventricles,
  (brain), *303*
  (heart), **290**
Venules, **288** (**Veins**)
Vernier scale, **100**
Vernier slide callipers, **100**
Vertebrae (sing. **vertebra**), *278*, **279**, *302*
  Cervical, *278*
  Coccygeal, *278* (**Coccyx**)
  Lumbar, *278*
  Sacral, *278*
  Thoracic, *278*, *279*
Vertebral canal, see **Neural canal**
Vertebral column, **279**
Vertebral foramen, **279** (**Vertebrae**)
Vertebrates, 265 (**Coelom**), *277* (**Spermatheca**), **341**
Vesicle, **239** (**Golgi complex**)
Vessels,
  Blood, **288** (**I**)
  Lymphatic/Lymph, **293**
  (plants), **243**
Vestibular canal, see **Scala vestibuli**
Vestibule (ear), *314*
Vibration, Modes of, **43**
Vibrational kinetic energy, **9** (**Kinetic energy**)
Vibrissae (sing. **vibrissa**), **274**
Villi (sing. **villus**), **295** (**Small intestine**)
  Chorionic, *319*
Vinyl chloride (CH_2CHCl), *200*
Virtual image, **49** (**Image**)
Viruses, 340 (**I**), **341**
Viscera, **278** (**I**)
Visceral afferent system, **307**
Visceral efferent system, see **Autonomic nervous system**
Visceral muscle(s), **282**
Visceral muscle (tissue), see **Smooth muscle**
Viscosity, **23**
Visible light, **45**, 54, *113*
Visible light spectrum, **54**
Visual, **313**
Visual angle, **55**
Visual area, Primary, **303**
Visual association area, **303**
Vitamin(s), **204** (**I**), **205**, 330 (**I**), **331**, **339**
Vitamin A, **339**
Vitamin B complex, **339**
Vitamin C, *205*, **339**
Vitamin D, 310 (**I**), **339**
Vitamin E, **339**
Vitamin H, see **Biotin**
Vitamin K, **339**
Vitreous humor, *312*
Viviparous, **276**
Vocal cords, **298** (**Larynx**)
Volatile, **345**
Volkmann's canals, *281*
Volt(s) (V), **58** (**Potential difference**), *97*, **159** (**Potential difference**)
  Electron (eV), **84**
  Megaelectron (MeV), **84**
Voltage, see **Potential difference**
Voltage divider, see **Potential divider**
Voltaic cell, see **Cell**
Voltaic pile, **68**
Voltameter, **67**, **157**
  Copper, **67** (**Voltameter**)
  Hoffmann, **67**
Voltmeter, **77**, 110, *158*, 159 (**Potential difference**)
Volume(s) (V), 33, 97, 101, 142-143, **345**
  Law of, **33**
  Molar, **139**
Volume coefficient of expansion (γ), **32**
Volumetric flask, **224**
Voluntary actions, **308**
Voluntary muscles, **282** (**I**, **Skeletal muscles**)
V-shaped (molecule), *133*
Vulcanization, 201, **345**
Vulva, **317**

W

Wall pressure, **253** (**Turgor**)
Wandering macrophages, **286** (**White blood cells**)
Washing soda ($Na_2CO_3.10H_2O$), **169** (**Sodium carbonate**), **207**
Watch glass, **225**
Water (H_2O), 112, 113, *133*, *134*, *138*, **144** (**Solvation**), 150, 151, 154, 155, **167**, *182*, *185*, **206-207**, *209*, 211, 216-217, 218, **251** (**Hydrotropism**), 252-253, 254-255 (**Photosynthesis**), **273** (**Contractile vacuoles**), 329
  Atmospheric, **206**
  Bromine, **188** (**Bromine**)
  Distilled, **207**
  Hard, **207**
  Heavy, see **Deuterium oxide**
  Soft, **207** (**Hard**)
  Tritiated, **167** (**Tritium**)
Water cycle, **206**
Water gas, **179** (**Carbon monoxide**)
Water of crystallization, **135**
Water softeners, **207**
Water softening, **207** (**Permanent hardness**)
Water supply, **207**
Watt (W), **9** (**Power**), 97
Wave(s), **34-35**
  Acoustic, see **Sound waves**
  Coherent, **38** (**I**)
  Electromagnetic, 34 (**I**), **44-45**, *113*
  Incident, **36**
  Infrasonic, **40** (**Infrasound**)
  Longitudinal, **35**
  Mechanical, **34** (**I**)
  Radio, 44 (**I**), **45**, *113*
  Reflected, **36** (**Reflection**)
  Refracted, **37** (**Refraction**)
  Shock, *40*
  Sound, **40-43**
  Standing/Stationary, **43**
  Transverse, **34**
  Traveling, **34** (**I**)
Wavebands, *44* (**Electromagnetic spectrum**)
Wave energy, **9**
Wavefront(s), **34**
  Circular, *34*
  Incident, **36** (**Incident wave**)
  Straight, *34*
Wave intensity, **35**
Wave interference, **38-39**
Wavelength, *34*, *113*
Wave motion, **34** (**I**)
Wave speed, **35**
Waxes, **205** (**Lipids**)
  Hydrocarbon/Paraffin, **199**
Weak acid, **152**
Weak base, **152**
Weak electrolyte, *66*, *156*
Weight, **18**
  Atomic, see **Relative atomic mass**
  Formula/Molecular mass, see **Relative molecular mass**
Weightlessness (**Apparent**, **True**), **19**
Whalebone, see **Baleen**
Wharton's duct, *296*
Wheatstone bridge, **63**
Whiskers, see **Vibrissae**
White blood cells/c
White matter, **302**
White phosphorus, **182** (**Phosphorus**
Whorl, see **Rosette**
Wilson cloud chamber
Wilting, **253**
Windpipe, see **Trachea**
Wing(s), 267, **269**
  Bastard, *269*
Wing petals, **259** (**Pea**
Wires (symbols for), **11**
Wirsung, Duct of, see
Wisdom teeth, **285**
Womb, see **Uterus**
Wood, *208*, **246** (**Sec**
thickening)
  Spring/Early, **246**
  Summer/Late, **246**
Woody perennials, **236**
Woody plant, **246**
Word equation, *98*, **141**
Work, **8** (**I**), 97
Wrought iron, 112, **174**

X

Xanthophyll, **255** (**Pig**
X-axis (graphs), *98*
X chromosomes, **325** (**Sex chromosomes**), **326**
Xenon (Xe), *129*, 165, **18**
Xenon tetrafluoride (**XeF**
Xerophytes, **342**
X-intercept (graphs), *98*
X-plates (oscilloscope
system)
X-radiography, *44*
X-ray crystallography, **13**
X-rays, **44**, *85*, *113*
X-ray tube, **80**
X-shift (oscilloscope con
Xylem, *242*, **243**, 246
thickening), 252
  Primary, *242* (**Primary**
  Secondary, **246** (**Seco**
thickening)

Y

Y-axis (graphs), *98*
Y chromosomes, **325** (**Sex chromosomes**), 326
Yeast, **197** (**Alcoholic fe**
Yellow marrow, **281** (**Bone**
Yellowspot, see **Macula**
Yield point, **23**
Yield stress, **23** (**Yield poi**
Y-intercept (graphs), *98*

Whiskers, see **Vibrissae**
White blood cells/corpuscles, **286**
White matter, **302** (**Neuroglia**)
White phosphorus,
  182 (**Phosphorus**), 213
Whorl, see **Rosette**
Wilson cloud chamber, *90*
Wilting, **253**
Windpipe, see **Trachea** (humans)
Wing(s), 267, **269**
  Bastard, *269*
Wing petals, **259** (**Pea flower**)
Wires (symbols for), **110**
Wirsung, Duct of, see **Pancreatic duct**
Wisdom teeth, **285**
Womb, see **Uterus**
Wood, *208*, **246** (**Secondary thickening**)
  Spring/Early, **246** (**Annual rings**)
  Summer/Late, **246** (**Annual rings**)
Woody perennials, **236** (**Perennials**)
Woody plant, **246** (**I**)
Word equation, *98*, **141**
Work, **8** (**I**), 97
Wrought iron, 112, **174** (**Iron**)

X

Xanthophyll, **255** (**Pigments**)
X-axis (graphs), *98*
X chromosomes, **325** (**Sex chromosomes**), 326 (**Sex linkage**)
Xenon (**Xe**), *129*, 165, **189**, 213
Xenon tetrafluoride (**XeF_4**), **189** (**Xenon**)
Xerophytes, **342**
X-intercept (graphs), *98*
X-plates (oscilloscope), *81* (**Deflection system**)
X-radiography, *44*
X-ray crystallography, **136**
X-rays, **44**, *85*, *113*
X-ray tube, **80**
X-shift (oscilloscope control), *81*
Xylem, *242*, **243**, 246 (**Secondary thickening**), 252
  Primary, *242* (**Primary tissue**)
  Secondary, **246** (**Secondary thickening**)

Y

Y-axis (graphs), *98*
Y chromosomes, **325** (**Sex chromosomes**), 326 (**Sex linkage**)
Yeast, **197** (**Alcoholic fermentation**)
Yellow marrow, **281** (**Bone marrow**)
Yellowspot, see **Macula lutea**
Yield point, **23**
Yield stress, **23** (**Yield point**)
Y-intercept (graphs), *98*
Yolk, *276*
Yolk sac, *276* (**Yolk**)
Young's modulus, *22*, **112**
Young's slits, **38**
Y-plates (oscilloscope), *81*
Ytterbium (**Yb**), **165**, **213**
Yttrium (**Y**), 164, **173**, 213

Z

Zeolite, **207** (**Ion exchange**)
Zero error, **102**
Zinc (**Zn**), 112, *137*, *155*, *159*, 165, 173, **175**, 211, 213, 217 (**Hydrogen**), 219 (**Cations**)
Zinc blende (**ZnS**), **175** (**Zinc**)
Zinc chloride (**$ZnCl_2$**), **217**
  Basic, (**Zn(OH)Cl**), *154*
Zinc hydroxide (**$Zn(OH)_2$**), 151 (**Amphoteric**)
Zincite (**ZnO**), **175** (**Zinc**)
Zinc oxide (**ZnO**), **175** (**Zinc**)
Zinc sulfate (**$ZnSO_4$**), *155*, *158*, *159*
Zirconium (**Zr**), 165, **173**, 213
Zona pellucida, **319** (**Fertilization**)
Zone of elongation, **245**
Zooplankton, **342** (**Plankton**)
Zygomorphy, **264** (**Bilateral symmetry**)
Zygote, 258 (**Fertilization**), 319 (**Fertilization**), **320**
Zymase, **197** (**Alcoholic fermentation**)

ACKNOWLEDGEMENTS

Cover designers: Russell Punter and Zöe Wray
Additional text by Paul Dowswell
American editor: Carrie A. Seay
Additional design by Chris Scollen, Stephen Wright,
Anne Sharples, Sue Mims and Nerissa Davies.

Additional illustrations by:

Simone Abel, Victor Ambrus, Basil Arm, Dave Ashby, Iain Ashman, Mike Atkinson, Craig Austin (The Garden Studio), Graham Austin, Bob Bampton (The Garden Studio), John Barber, John Barker, Amanda Barlow, Terry Bave, David Baxter, Andrew Beckett, Joyce Bee, Stephen Bennett, Roland Berry, Andrzej Bielecki, Gary Bines, Val Biro, Blue Chip Illustration, Kim Blundell, Derick Bown, Isabel Bowring, Trevor Boyer, Wendy Bramall (Artist Partners), Derek Brazell, John Brettoner, Paul Brooks (John Martin Artists), Fiona Brown, Gerard Browne, Peter Bull, Mark Burgess, Hilary Burn, Andy Burton, Liz Butler, Martin Camm, Lynn Chadwick, Chris Chapman, Peter Chesterton, Paul Cooper, Sydney Cornford, Dan Courtney, Frankie Coventry (Artist Partners), Patrick Cox, Christine Darter, Kate Davies, Sarah De Ath (Linden Artists), Kevin Dean, Peter Dennis, David Downton, Richard Draper, Nick Dupays, Brian Edwards, Michelle Emblem (Middletons), Malcolm English, Caroline Ewen, Sandra Fernandez, James Field, Denise Finney, Dennis Finney, Wayne Ford, Don Forrest, Patience Foster, Sarah Fox-Davies, John Francis, Mark Franklin, Nigel Frey, Judy Friedlander, Peter Froste, Terry Gabby, Sheila Galbraith, Stephen Gardener, Peter Geissler, Nick Gibbard, Tony Gibson, William Giles, Mick Gillah, Victoria Goaman, David Goldston, Peter Goodwin, Victoria Gordon, Jeremy Gower, Terri Gower, Miranda Gray, Phil Green, Terry Hadler, Rys Hajdul, Edwina Hannam, Alan Harris, Brenda Haw, Tim Hayward, Bob Hersey, Nicholas Hewetson, Rosalind Hewitt, Keith Hodgson, Philip Hood, Adam Hook, Shirley Hooper, Chris Howell-Jones, Christine Howes, Carol Hughes (John Martin Artists), Rian Hughes, David Hurrell (Middletons), Gillian Hurry, John Hutchinson, Roy Hutchison (Middletons), Ian Jackson, Tony Jackson, Hans Jenssen, Chris Johnson, Elaine Keenan, Frank Kennard, Roger Kent, Aziz Khan, Colin King, Deborah King, Steven Kirk, Kim Lane, Jonathan Langley, Richard Lewington (The Garden Studio), Jason Lewis, Ken Lilly, Steve Lings (Linden Artists), Mick Loates (The Garden Studio), Rachel Lockwood, Tony Lodge, Kevin Lyles, Chris Lyon, Kevin Maddison, Janos Marffy, Andy Martin, Josephine Martin, Nick May, Rob McCaig, Joseph McEwan, David McGrail, Malcolm McGregor, Doreen McGuinness, Dee McLean (Linden Artists), David Mead, Jamie Medlin, Richard Millington, Annabel Milne, Sean Milne, David More (Linden Artists), Dee Morgan, Robert Morton (Linden Artists), David Mostyn, Paddy Mounter, David Nash, Susan Neale, Louise Nevett, Martin Newton, Barbara Nicholson, Louise Nixon, David Nockels (The Garden Studio), Richard Orr, Steve Page, David Palmer, Patti Pearce, Justine Peek, Liz Pepperell (The Garden Studio), Julia Piper, Gillian Platt (The Garden Studio), Maurice Pledger, Cynthia Pow (Middletons), Russell Punter, David Quinn, Charles Raymond (Virgil Pomfret Agency), Kim Raymond, Barry Raynor, Chris Reed, Phillip Richardson, Jim Robins, Allan Robinson, Michael Roffe, Michelle Ross, Peter Ross, Graham Round, Evie Safarewicz, Mike Saunders (Tudor Art), Jon Sayer, Peter Scanlan, John Scorey, Coral Sealey, John Shackell, Chris Shields (Wilcock Riley), John Sibbick (John Martin Artists), Penny Simon, Gwen Simpson, Chris Smedley, Graham Smith, Annabel Spencerley, Peter Stebbing, Sue Stitt, Roger Stewart, Ralph Stobart, Paul Sullivan, Alan Suttie, John Thompson-Steinkrauss (John Martin Artists), Sam Thompson, Stuart Trotter, Joyce Tuhill, Sally Voke (Middletons), Sue Walliker, Robert Walster, David Watson, Ross Watton, Phil Weare, Chris West, Wigwam Publishing Services, Sean Wilkinson, Adrian Williams, Adam Willis, Roy Wiltshire, Ann Winterbotham, Gerald Wood, James Woods (Middletons), Claire Wright, David Wright (Jillian Burgess), Jo Wright, Nigel Wright, Gordon Wylie, John Yates.

Photograph credits:

The ultrasound scan on page 40 is reproduced with the kind permission of Charlotte Tomlins.
© Alfred Pasieka / Science Photo Library (45, bottom left);
Digital Vision (45, bottom right; 93, bottom right; 206, top right).

Every effort has been made to trace and acknowledge ownership of copyright. The publishers will be glad to make suitable arrangements with any copyright holder whom it has not been possible to contact.

This edition first published in 2001 by Usborne Publishing Ltd, Usborne House,
83-85 Saffron Hill, London EC1N 8RT, England. www.usborne.com
Copyright © Usborne Publishing Ltd, 2001, 1999, 1988.
The name Usborne and the devices are Trade Marks of Usborne Publishing Ltd.
AE First published in America in 2001

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the publisher.
Printed in Great Britain.

Usborne Publishing is not responsible and does not accept liability for the availability or content of any Web site other than its own, or for any exposure to harmful, offensive, or inaccurate material which may appear on the Web. Usborne Publishing will have no liability for any damage or loss caused by viruses that may be downloaded as a result of browsing the sites it recommends.

USBORNE

THE USBORNE ILLUSTRATED DICTIONARY OF SCIENCE

With colorful labeled diagrams, clear definitions of key terms and concepts, extensive cross-referencing and a detailed index, this brightly illustrated dictionary is an essential study aid for home and school - and a helpful guide for reviewing for exams.

The structure of skin

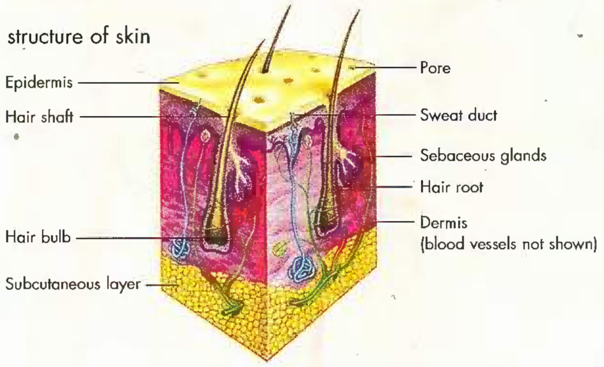
Epidermis
Hair shaft
Hair bulb
Subcutaneous layer
Pore
Sweat duct
Sebaceous glands
Hair root
Dermis
(blood vessels not shown)

In addition, this classic Usborne bestseller now has links to useful science resources on the Internet via the Usborne Quicklinks Web site at www.usborne-quicklinks.com

For more information about Usborne Publishing, visit www.edcpub.com or www.ubah.com

Published in the USA by
EDC PUBLISHING, 10302 E. 55th Place,
Tulsa, Oklahoma 74146, USA
EDUCATIONAL
DEVELOPMENT
CORPORATION
NOT FOR SALE IN CANADA

$29.95

JFMA JJASOND/04
ISBN 0-7945-0064-1
Made with paper from a sustainable source.

9 780794 500641

Ugly/Ground Plane Construction

The Method I Use in Building My Simple Receivers

Oct. 7, 2006 by Rick Andersen, KE3IJ

Occasionally I get emails from readers who want to know how I build these simple radio receiver circuits on an "L-shaped chassis" without cutting and scoring the copper, what I mean by "L-shaped chassis," or "what's all this about 10 megohm resistors?", etc. I tend to forget that not everybody has done this before, and I usually take it for granted that the reader is an accomplished "basement-tinkerer" and should therefore be able to just look at a schematic diagram and then come up with his own layout and method of construction. Sorry if I have annoyed anyone -- that wasn't my intention.

When I first design a circuit, I make what a friend of mine used to refer to as a "spider" (or, "spider-web")-- basically just construct the circuit by soldering the components together, lead by lead, into a reasonable facsimile of the schematic (no chassis or pc board necessary-- the stiffness of the soldered leads themselves hold the "spider" circuit together). This method is quick, and allows me to modify things fairly easily. It also grows rather quickly into a "haywire monstrosity" if I allow it.

Sometimes I do use a piece of copper-clad pc board as a ground plane, even during the "spider" stage, as the photo below illustrates:

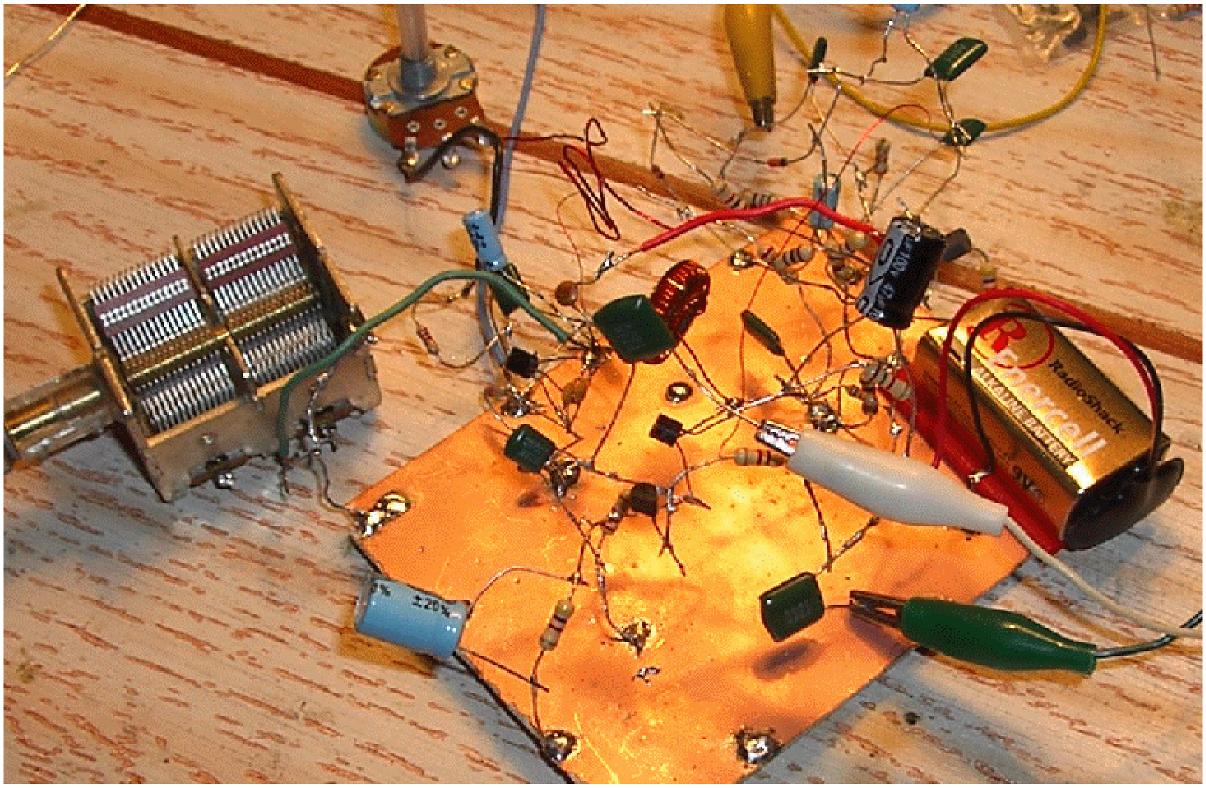

A "haywire" mess of my "AGC-80" Regen receiver in its early, experimental stage

To make things a bit clearer to you who have never built projects in this way before, I have attempted a free-hand diagram using Microsoft "Paint", which I converted to a .gif file and which you see below.

In the top-center of the diagram is the schematic for a simple common-emitter audio amplifier stage. The left side of the diagram illustrates how one might translate that schematic into an actual circuit on the copper-clad pc board. At right is a diagram of the "L-shaped Chassis." This is the style in which I build most of my circuits.

Basically, I use 2 pieces of 5x7 inch copper-clad pc board, available in the USA at **Radio Shack** [free plug]. One piece lies flat (the bottom piece), which is what I have shown in the diagram. I build my circuit on this bottom piece, using the copper floor as my ground connections ("ground plane"). Any resistors, caps, etc., that go to ground in the schematic, are literally soldered to this bottom ground plane and stood upright. At their top ends, the rest of the circuit is soldered, suspended up above the ground plane. Anywhere that I think I need some mechanical stability, I use a vertical 10 Megohm resistor, which acts to prop up and support the rest of the wiring above the ground plane. For most of the circuits I build, 10 Megohms is much higher a resistance value than any of the resistors in the circuit itself; therefore, the circuit pretty much ignores the 10M "posts", electrically, and they remain simply a mechanical "standoff insulator". The 2nd piece of copper clad board is drilled for switches, potentiometers, tuning caps, and earphone jacks, etc., and then is soldered at right angles to the flat bottom piece, making an "L-shaped chassis". It also helps to solder triangular pieces of copper-clad board as mechanical braces, at the corners where the front panel is soldered to the bottom ground plane.

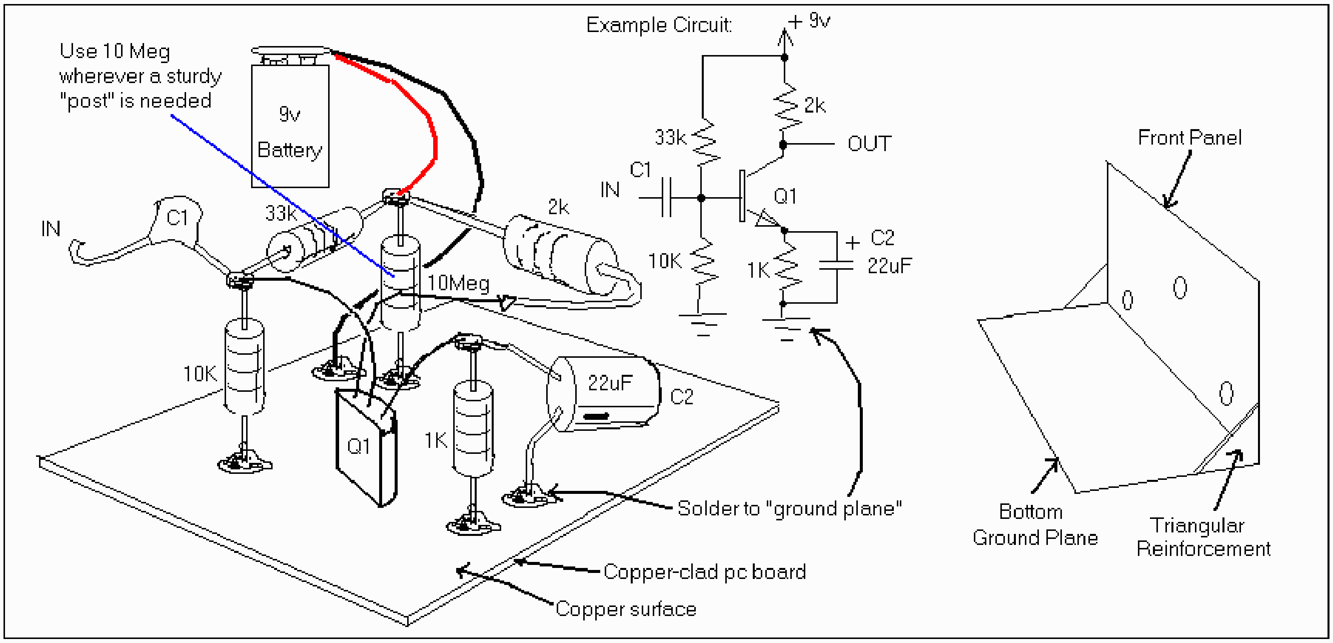
Use 10 Meg wherever a sturdy "post" is needed
9v Battery
IN
C1
33k
2k
10Meg
10K
Q1
1K
22uF
C2
Solder to "ground plane"
Copper-clad pc board
Copper surface
Example Circuit:
+ 9v
2k
33k
OUT
C1
IN
Q1
10K
1K
+ C2
22uF
Front Panel
Bottom Ground Plane
Triangular Reinforcement

Below is a photo of my 2006 creation, the "AGC-80", a Regen receiver for the 80 Meter ham band with audio-derived AGC. There's an article on this receiver elsewhere on this website; meanwhile, notice that it is built in the same way outlined above-- an "L-shaped" chassis made of a bottom piece of copper-clad board, soldered to a vertical piece of the same kind of board which is the front panel. [I also added a back wall to hold the antenna jack, and a right wall to hold the audio output jack, as well as a triangular reinforcement piece of copper-clad, at bottom-left between the front panel and bottom piece, although you can't see it in this photo.]

To the right of the AGC-80 is the infamous Radio Shack/Archer little beige Amplified Speaker that I'm always mentioning. I may knock Radio Shack for other reasons [for example, that they have pretty much left us hobbyists behind, and have become just another consumer electronics store] but I will always say good things about their little $12.95 beige-colored Amplified Speaker. It works like a charm for the kind of projects I build. I also run an audio cable out of the Radio Shack amp to a bigger speaker, for better-quality sound-- in the photo you can see a white "Optimus" computer speaker pressed into this service. While the Radio Shack's speaker does a fine job, the Optimus sounds quite a bit louder and way nicer. Of course, you may argue that a *real* Electronics Geek rolls his *own* audio amp rather than rely on a commercially-made, external one. Well, I'm often too impatient to build the audio power amp myself, after having spent hours or days getting the RF part of the radio to work the way I want. Also, homemade audio amps tend to oscillate (squeal, motorboat, howl) in mysterious ways once lovingly installed on the same circuit board as the RF circuitry.... After a while one grows tired of trying 50 ways to make it happy, and would rather just hook up an Amplified Speaker and be done with it.

To each his own.

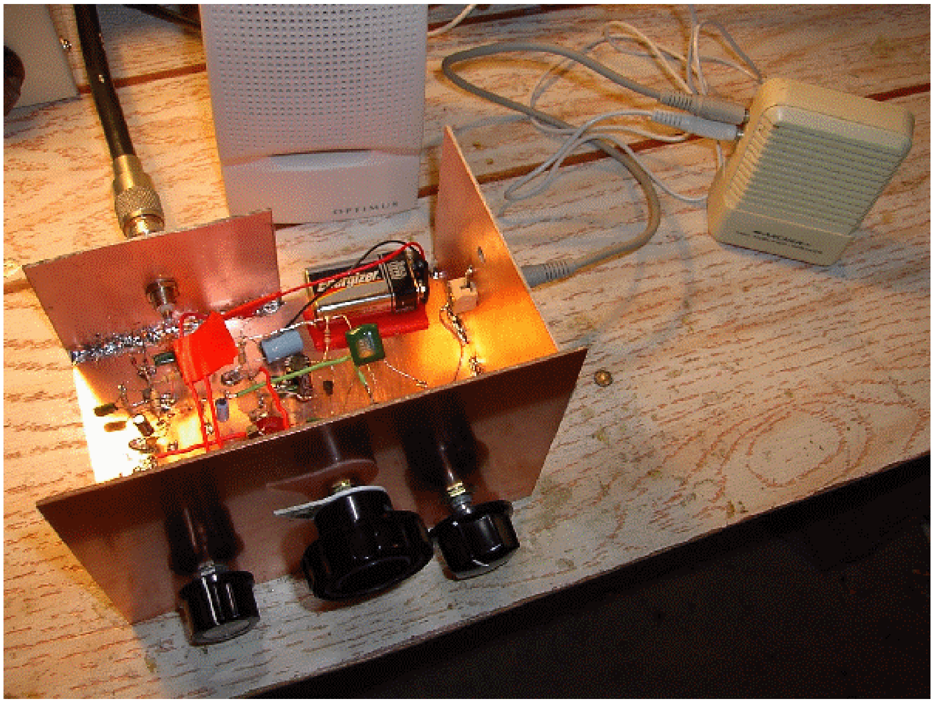

Please realize that you don't have to follow this method of construction (usually called "Ugly Construction" by hams), but I would advise you to make sure there is some equivalent of a copper ground plane in most of your radio projects... it helps to shield the circuit from outside influence (particularly when connected to earth ground) and contributes to stability.

I hope this has helped to give you a little better idea of the methods I use!
73 de Rick, KE3IJ

Back

Chemistry

Like the author's other companion books, **The Chemistry Companion** provides high quality information in unique one-page-per-topic presentations that do not overburden and distract readers with excessive details. The book offers concise summaries of general chemistry concepts, easily accessible in a convenient, reader-friendly format.

Suitable as an introduction or study guide, this companion presents the minimum of what readers need to know to understand the subject. It emphasizes the physics underlying chemistry. By looking at chemistry processes from a physics point of view, readers better appreciate what is happening from the chemical perspective that is usually found in traditional chemistry books.

The author focuses on the structure of matter, chemical components and bonds, the periodic table, states of matter, thermodynamics, reaction rates, carbon chemistry, biochemistry, and chemical, ionic, and electronic equilibria. Each topic is covered in a single-page outline format with just enough detail to enable a good understanding of the subject.

Features

- Provides a physical understanding of chemical concepts
- Presents clear explanations of difficult material, working through any inconsistencies in understanding
- Uses a convenient format for checking formulas and definitions
- Includes self-contained information on each page, assuming little prior knowledge

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K11517

ISBN: 978-1-4398-3088-8
90000
9 781439 830888

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press

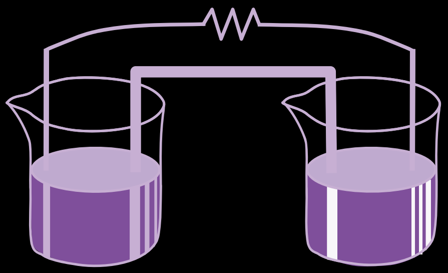

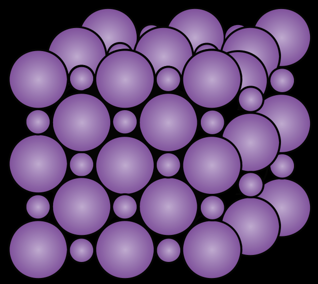

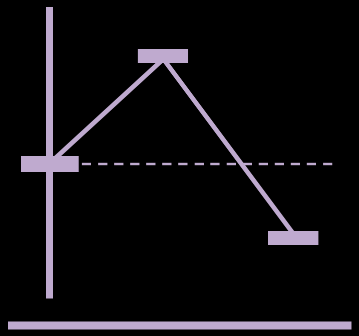

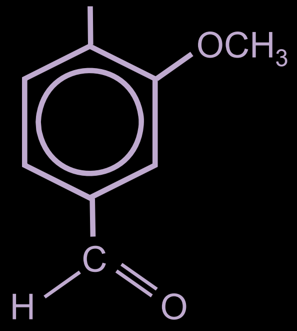

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press
Taylor & Francis Group

The Chemistry Companion

This page intentionally left blank

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20110517

International Standard Book Number-13: 978-1-4398-3089-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

This book is dedicated to Bill Cripps
— industrial chemist

This page intentionally left blank

Contents

Preface xiii

1. Structure of Matter 1
- 1.1 Atoms 2
- 1.2 Bohr Atom 3
- 1.3 Energy Levels 4
- 1.4 Schrödinger Equation 5
- 1.5 The Infinite Square Well 6
- 1.6 The Coulomb Potential 7
- 1.7 Covalent Bond 8
- 1.8 Ionisation Energy 9
- 1.9 Electron Affinity 10
- 1.10 Ionic Bond 11
- 1.11 Electronegativity 12
- 1.12 Metallic Bond 13
- 1.13 Electronic Structure of Solids 14

2. Chemical Components 15
- 2.1 Matter 16
- 2.2 Atomic Weight 17
- 2.3 Ions 18
- 2.4 Molecules 19
- 2.5 Mole 20
- 2.6 Compounds 21
- 2.7 Cations 22
- 2.8 Anions 23
- 2.9 Chemical Equation 24
- 2.10 Stoichiometry 25
- 2.11 Example 26

3. The Periodic Table 27
- 3.1 Electron Configuration 28
- 3.2 Periodic Law 29
- 3.3 Periodic Table 30
- 3.4 Groups 31
- 3.5 Energy Levels 32
- 3.6 Noble Gases 33
- 3.7 Atomic Size 34
- 3.8 Covalent Radii 35
- 3.9 Ionic Radii..36
- 3.10 Ionisation Energy 37
- 3.11 Electronegativity 38

4. Chemical Bonds 39
4.1 Chemical Bond 40
4.2 Lewis (Electron Dot) Formulae 41
4.3 Multiple Bonds 42
4.4 Lewis Single-Bonded Structures 43
4.5 Lewis Multiple-Bonded Structures 44
4.6 Lewis Exceptions to the Octet Rule 45
4.7 Oxidation Number 46
4.8 Oxidation Number Examples 47
4.9 Polar Bonds 48
4.10 Hybrid Orbitals 49
4.11 Polarisation 50
4.12 van der Waals Forces 51
4.13 Hydrogen Bond 52

5. States of Matter 53
5.1 Changes of State 54
5.2 Changes of State of Liquids/Gases 55
5.3 Phases of Matter 56
5.4 Gases 57
5.5 Solutions 58
5.6 Aqueous Solutions 59
5.7 Solubility of Solids 60
5.8 Solubility Equilibrium 61
5.9 Electrolytes 62
5.10 Osmosis 63
5.11 Solids 64
5.12 Crystalline Lattice Structures 65
5.13 Metallic Solids 66

6. Chemical Thermodynamics 67
6.1 Molecular Energy 68
6.2 Specific Heat Capacity 69
6.3 Enthalpy 70
6.4 Heat of Reaction 71
6.5 Heat of Reaction 72
6.6 Enthalpy of Formation 73
6.7 Entropy 74
6.8 Entropy Calculations 75
6.9 Gibbs Energy 76
6.10 Spontaneous Processes 77
6.11 Melting of Ice 78
6.12 Freezing of Water 79

6.13 Ice/Water Equilibrium 80
6.14 Chemical Equilibrium 81
6.15 Statistical Entropy 82

7. Rates of Reaction 83
7.1 Rates of Reaction 84
7.2 Collision Theory 85
7.3 Reaction Mechanism 86
7.4 Activation Energy 87
7.5 Nature of Reactants 88
7.6 Concentration 89
7.7 Rate Law 90
7.8 Rates of Reactions 91
7.9 Determination of Order 92
7.10 Determination of 1st Order Rate Law 93
7.11 Half-Life Method 94
7.12 Temperature 95
7.13 Catalysts 96

8. Chemical Equilibrium 97
8.1 Chemical Equilibrium 98
8.2 Law of Chemical Equilibrium 99
8.3 Equilibrium Constant 100
8.4 Le Chatelier’s Principle 101
8.5 Summary of Le Chatelier’s Principle 102
8.6 Equilibrium in Gaseous Systems 103
8.7 Solubility of Solids 104
8.8 Factors Affecting Equilibrium 105
8.9 Common Ion Effect 106
8.10 Precipitation 107
8.11 Complex Ions 108

9. Ionic Equilibrium 109
9.1 Electrolytes 110
9.2 Ionisation of Water 111
9.3 H^+ and OH^- Concentrations 112
9.4 Acids and Bases 113
9.5 Bronsted–Lowry 114
9.6 Strength of Acids 115
9.7 Acid–Base Reactions 116
9.8 Buffer Solutions 117
9.9 Indicators 118
9.10 Neutralisation 119

9.11 Titration 120
9.12 Choice of Indicator 121
9.13 Hydrolysis 122
9.14 Simultaneous Equilibria 123
9.15 Complex Ions 124

10. Electronic Equilibrium 125
10.1 Oxidation and Reduction 126
10.2 Redox Reaction 127
10.3 Single-Cell Redox Reaction 128
10.4 Standard Hydrogen Electrode 129
10.5 Standard Electrode Potentials 130
10.6 Spontaneous Redox Reactions 131
10.7 Oxidation Numbers 132
10.8 Balancing Redox Half-Reactions 133
10.9 Balancing Redox Reactions with Oxidation Numbers 134
10.10 Electrochemical Cell 135
10.11 EMF vs. Concentration 136
10.12 Electronic Equilibrium 137
10.13 Equilibrium Constant 138
10.14 Lead/Acid Battery 139
10.15 Dry Cell 140
10.16 Corrosion 141
10.17 Electrolysis 142

11. Carbon Chemistry....143
11.1 Carbon 144
11.2 Carbon Compounds 145
11.3 Functional Groups 146
11.4 Alkanes 147
11.5 Alkenes 148
11.6 Alkynes 149
11.7 Benzene 150
11.8 Alcohols 151
11.9 Ethers 152
11.10 Aldehydes 153
11.11 Ketones 154
11.12 Carboxylic Acids 155
11.13 Esters 156
11.14 Amides 157
11.15 Amines 158
11.16 Polymers 159
11.17 Physical and Chemical Properties 160

12. Biochemistry 161
12.1 Sugars 162
12.2 Polysaccharides 163
12.3 Lipids 164
12.4 Proteins 165
12.5 Nucleic Acids 166
12.6 DNA 167
12.7 Enzymes 168
12.8 ATP 169
12.9 Anaerobic Metabolism 170
12.10 Aerobic Metabolism 171
12.11 Cyclic Photophosphorylation 172
12.12 Non-Cyclic Photophosphorylation 173
12.13 Metabolism 174

This page intentionally left blank

Preface

This book is similar to previous Companion style books where each topic is covered in a single page outline format with enough detail to provide a good understanding of the subject.

This book emphasises the physics underlying chemistry, especially in the first chapter. I hope that by understanding what is happening from a physics point of view, the reader may better appreciate what is happening from the chemical perspective that is usually found in a traditional chemistry book.

I am indebted to Dr. Ray Sleet of the University of Technology, Sydney, an exemplary teacher who taught me that anything can be learned if it is broken down into manageable pieces with attainable goals. Many of the presentations in this book have their origin in Dr. Sleet's excellent undergraduate lectures.

I also thank Hilary Rowe for her persistence in getting this book into print and the editorial and production team at Taylor & Francis for their very professional and helpful approach to the whole publication process.

Tony Fischer-Cripps,
Killarney Heights, Australia

This page intentionally left blank

1. Structure of Matter

Summary

Energy levels Bohr H-like atom:

$$E_n = -\frac{m_e Z^2 {q_e}^4}{8{\varepsilon_o}^2 h^2 n^2}$$

Energy levels Schrödinger equation,
Coulomb potential:

$$E_n = -\frac{Z^2 {q_e}^4 m}{(4\pi\varepsilon_o)^2 2\hbar^2 n^2}$$

Covalent bond: the co-sharing of valence electrons.

Ionic bond: the electrostatic attraction between ions formed after electrons are transferred from one atom to another.

Metallic bond: free electrons in the valence band reduce the energy of the system and so act to hold atoms together.

The electronegativity describes the relative ability of an atom, when it combines with another atom, to become "more negative" by more strongly capturing a shared electron or electron pair.

1.1 Atoms

Sixth century BC Thales of Miletos proposes that the basic element from which all things are made is water. 450 BC Empedocles teaches that all matter is composed of earth, air, water and fire. Around 400 BC, Greek philosophers (Leucippus, **Democritus**, Epicurus) proposed that if one could divide a piece of matter again and again, eventually a limit would be reached where no further subdivision could take place, this limiting amount of matter was called the **atom**. About 300 BC, **Aristotle** rejects the atomistic view and argues that matter is based upon the four basic elements of Empedocles but adds the qualities of coldness, hotness, dryness and moistness. Aristotle's considerable reputation ensured that his ideas became embodied in religious teaching for many hundreds of years.

In the fifteenth century, new advances in physics suggested that matter was made from particles, in agreement with the ancient Greek atomists. Robert **Boyle** taught that matter consists of different types of **elements** that were composed of atoms of the same type. Different elements could join together in fixed proportions to form **compounds**. Later, in 1803, John **Dalton** proposed the **atomic theory of matter** that was based on quantitative experimental evidence from the weighing of different elements in combination. He created a scale atomic mass for the different elements that were then known. Dalton's reference atom was the lightest element known, hydrogen, which was assigned an atomic mass of one. Other elements were given atomic masses according to how heavy they were compared to a hydrogen atom.

In 1807, Humphry Davy decomposed potash into sodium and potassium metals using electrolysis. In 1832, Michael **Faraday** discovered a quantitative connection between electricity and the separation of compounds into elements in electrolysis. These observations suggested that atoms themselves contain electric charge. Experiments by William Crookes demonstrated visible "cathode rays" that emanated from a negatively charged electrode (cathode) and travelled towards the positive electrode (anode) in an evacuated tube.

In 1872, **Mendeleev** arranged elements in increasing order of atomic mass and discovered that the properties of certain elements were repeated at regular intervals. When elements were ordered in columns with the atomic mass going across from left to right, and similar chemical properties going down, a **periodic table** was formed whereby, using the known elements at the time, Mendeleev was able to predict the properties of some as-yet undiscovered elements from gaps in the table.

1.2 Bohr Atom

In 1897, **Thomson** demonstrated that the rays observed to be emitted from the cathodes of vacuum tubes were in fact charged particles which he called **electrons**. Thomson proposed that atoms consisted of a positively charged sphere within which were embedded negatively charged electrons.

Rutherford found in 1911 that the electrons were actually located at some distance from a central positively charged **nucleus**. He proposed that electrons orbited the nucleus and the electrostatic attraction between the nucleus and the electron was balanced by the centrifugal force arising from the orbital motion. However, a major problem with this was that if this were the case, then the electrons would continuously radiate all their energy as **electromagnetic waves** and very quickly fall into the nucleus.

$q_e = -1.6\times10^{-19}\,\mathrm{C}$

In 1913, **Bohr** postulated two important additions to Rutherford's theory of atomic structure:

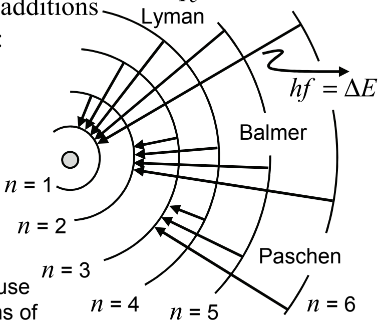

1. Electrons of mass m_e can orbit the nucleus at radius r with velocity v in what are called **stationary states** in which no emission of radiation occurs and in which the **angular momentum** L is constrained to have values:

$$L = m_e v r = \frac{nh}{2\pi}$$

The 2π appears because L is expressed in terms of ω rather than f.

2. Electrons can make transitions from one state to another accompanied by the emission or absorption of a single **photon** of energy $E = hf$, this being the absorption and emission spectra observed experimentally.

As in the Rutherford atom, the centrifugal force is balanced by Coulomb attraction:

$$\frac{1}{4\pi\varepsilon_o}\frac{q_e^2}{r^2} = \frac{m_e v^2}{r}$$

with the addition that: $m_e v r = \frac{nh}{2\pi}$

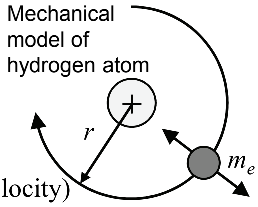

By summing the **kinetic energy** (from the orbital velocity) and the **potential energy** from the electrostatic force, the **total energy** of an electron at a given energy level n is given by:

$$E_n = -\frac{m_e Z^2 q_e^4}{8\varepsilon_o^2 h^2 n^2}$$

Note: **atomic number** Z = 1 for the hydrogen atom where the energy of the ground state is –13.6 eV. The energy levels for each state n rise as Z^2. Thus, according to the Bohr model, the energy level of the innermost shell for multi-electron atoms can in principle be several thousand eV.

1.3 Energy Levels

The stationary states or energy levels allowed by the Bohr model are called **electron shells** or orbitals, and are labelled K, L, M, N, etc. with K corresponding to n = 1. The number n is called the **principal quantum number**. According to the Bohr model, the electron energy only depends on n, but experiments show that in multi-electron atoms, electron shells consist of sub-levels (evidenced by fine splitting of spectral lines). For example, the L shell n =2 has two **sub-shells**, 2*s* and 2*p*.

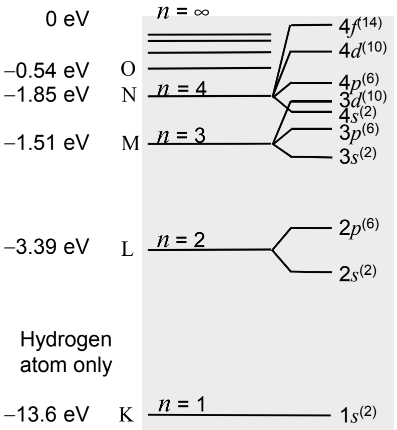

It is convenient to assign the energy at infinity as being 0 since as an electron moves closer to the nucleus, which is positively charged, its potential to do work is less and thus the energy levels for each shell shown are negative. In hydrogen, a single-electron atom, the energies for each shell are given by:

$$E = -\frac{13.6}{n^2} \quad \text{for hydrogen}$$

The electron-volt is a unit of energy.
1 eV = 1.602 × 10^{-19} J.

Sometimes the splitting of principal shells into sub-shells results in some overlap (e.g. 4*s* is lower in energy than 3*d*).

At each value of n the **angular momentum** can take on several distinct values. The number of values is described by the second quantum number l. The allowed values of l are 0, 1, … (n−1). Each value of l is indicated by a letter that indicates the sub-shell:

l = 0	*s*
l = 1	*p*
l = 2	*d*
l = 3	*f*
l = 4	*g*
l = 5	*h*

A third quantum number m describes the allowable changes in angle of the **angular momentum** vector in the presence of an electric field. It takes the values $-l$ to 0 to $+l$.

A fourth quantum number describes the **spin** of an electron where the spin can be either −1/2 or +1/2.

According to the **Pauli exclusion principle**, no electron in any one atom can have the same combination of quantum numbers. When all the electrons in an atom are in the lowest possible energy levels, the atom is said to be in its **ground state**. The outermost electrons in an atom are called the **valence electrons**.

For example, the 3*d* sub-shell can hold up to 10 electrons:

n = 3
thus: l = 0, 1, 2 (*s*, *p*, *d*)
and: $m = \underbrace{-2, -1, 0, 1, 2}$

5 values of m times 2 for spin, thus 10 possible electrons

1.4 Schrödinger Equation

The total energy E of an electron in an atom is the sum of the potential and kinetic energies. Expressed in terms of **momentum**, p, and mass of electron m, this is stated:

$$E = \frac{p^2}{2m} + V$$

The value of the potential function may depend on both position and time. The form of $V(x,t)$ is different for different arrangements of atoms (e.g. a single isolated atom, an atom in a regular array of a crystal).

Thus: $$hf = \frac{p^2}{2m} + V(x,t)$$ since $E = hf$

Let $$\left.\begin{aligned} p &= -i\hbar\frac{\delta\Psi}{\delta x} \\ E &= i\hbar\frac{\delta\Psi}{\delta t} \end{aligned}\right\}$$

Ψ is a variable, the form and value of which provide information about the motion of a wave/particle.

Thus: $$\boxed{-\frac{\hbar^2}{2m}\frac{\delta^2\Psi}{\delta x^2} + V(x,t)\Psi = i\hbar\frac{\delta\Psi}{\delta t}}$$ **Schrödinger equation**

The solution to the Schrödinger wave equation is the **wave function** Ψ. If V is a function of x only, then the wave equation can be separated into time-independent and time-dependent equations that can be readily solved.

$$-\frac{\hbar^2}{2m}\frac{\delta^2\psi}{\delta x^2}\overset{\downarrow}{+}V(\psi) = E\psi \qquad\qquad \overset{\downarrow}{\phi(t)} = e^{i\frac{E}{\hbar}t}$$

The resulting solutions of these equations, when multiplied together, give the **wave function**: $\Psi(x,t) = \psi(x)\phi(t)$

The wave function gives all the information about the motion of an electron in an atom. Ψ is a complex quantity, the magnitude of which $|\Psi|$ is interpreted as a **probability density function** which in turn can be used to determine the probability of an electron being at some position between x and Δx.

Quantum mechanics is concerned with determining the wave function (i.e. solving the Schrödinger equation) for particular potential energy functions such as those inside atoms. It is found that valid solutions to the time-independent wave equation occur only when the total energy is quantised. The solutions correspond to **stationary states**.

Solutions to the Schrödinger equation can be found for potential functions which are a function of both x and t. This enables time-dependent phenomena (e.g. the probability of transitions of electrons between energy levels in an atom) to be calculated and hence the intensity of spectral lines.

1.5 The Infinite Square Well

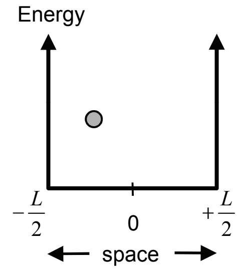

Consider an electron that is confined to be located in one-dimensional space $-L/2$ and $+L/2$ and where the **potential energy** of the electron is a constant (which can be conveniently set to zero). This is an example of an **infinite square well potential**. This potential is not usually found around electrons in an isolated atom, but does often represent that experienced by electrons in a solid and also in a chemical bond.

In the infinite square well potential, the electron cannot move more than a distance $x = \pm L/2$ from the centre position because it is constrained or bound by the infinite potential at the walls. In classical physics, the electron can have any value of total energy as it moves within this space.

In quantum physics, the allowed **stationary states** indicate the possible total energy of the electron. That is, the electron is moving as a particle whose probability of being in a particular position x is described by the standing wave patterns from the solutions to the **Schrödinger equation**. For the simple case of ***zero*** **potential energy** within the well, the energy levels are given by:

$$E = \frac{n^2\pi^2\hbar^2}{2L^2 m} \qquad n = 1,2,3,4\ldots,\ V(x) = 0$$

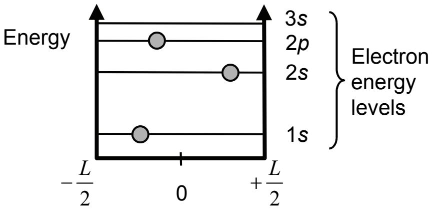

The electron can move anywhere within the confines ± L, but can only have kinetic energies allowable by the stationary states.

It can be seen that the energy increases as n increases. That is, the electron has a greater **kinetic energy** if it exists in a **stationary state** with a larger n. The minimum allowable energy is greater than the minimum potential energy (in this case, 0). This is the **zero point energy**.

Note also that the kinetic energy of the electron, for a given value of n, decreases as the length of the well L increases. That is, if the electron is given more room to move, then its total energy is lowered (in this case, the total energy is kinetic, since the potential energy was set to be 0). "Available space" is effectively a mechanism for providing a reduction in total energy of the system. This is important when the **valence electrons** of different atoms come near each other during the formation of **chemical bonds**.

1.6 The Coulomb Potential

In the **infinite square well potential**, the potential energy of the electron inside the well was constant, and independent of position. In an isolated **single-electron atom**, a more realistic scenario is the case where electrons are bound by the **Coulomb potential**.

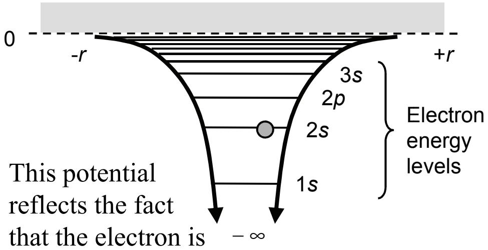

This potential reflects the fact that the electron is attracted more strongly by the nucleus at shorter distances *r* and so the potential energy is no longer constant – i.e. the potential varies with distance from the nucleus. Mathematically, the Coulomb potential is:

$$V(r) = -\frac{Zq_e^{\,2}}{4\pi\varepsilon_o r}$$

where Z is the **atomic number**, the number of protons in the nucleus.

By convention, an electron is assigned a zero total energy when it is at rest, at an infinite distance from the nucleus. As the negatively charged electron moves towards the positively charged nucleus (opposite to the direction of the electric field), it can acquire kinetic energy and/or do work and so its electrical potential is reduced. That is, the potential energy becomes negative. This is the same convention in electric field theory, where it is usual to consider a *positive* test charge moving in the *same* direction of the electric field – whereupon potential energy is reduced.

Note that in this potential, the energy (the potential energy) becomes more negative as *r* decreases. That is, electrons further away from the nucleus have a higher potential energy.

The total energy levels for an electron for this potential are given by the solution to the **Schrödinger equation** with $V(r)$ as above and are expressed:

$$E_n = -\frac{Z^2 q_e^{\,4} m}{(4\pi\varepsilon_o)^2 2\hbar^2 n^2} \qquad n = 1,2,3,4\ldots$$

This applies to an isolated **single electron atom** (e.g. hydrogen). Compare with Bohr model 1.2.

Here, the energy E becomes more negative as n decreases, and for a single electron ($Z = 1$) , and at $n = 1$, we obtain the total energy $E = -13.6$ eV. The Coulomb potential, as drawn above, also reflects the greater range of movement available to the electron as *r* increases.

When an electron is bound by a potential, it is forced into having discrete allowable energies. That is, the allowable states $E < 0$ are **bound states**. Above this level, the electron is free and can have any energy. This is indicated by the grey band in the figure above. Note that in this potential, the energy depends on the principle quantum number *n*. As *n* increases, the energy becomes less negative and is therefore at a higher potential.

1.7 Covalent Bond

Consider a **hydrogen atom**. Here we have one electron orbiting the nucleus. In the ground state, the electron is in one of the available 1*s* energy states.

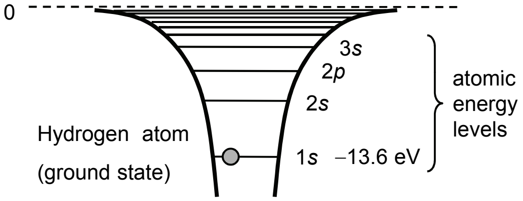
0
3*s*
2*p*
2*s*
Hydrogen atom
(ground state)
1*s* −13.6 eV
atomic energy levels

When a hydrogen atom comes into proximity with another hydrogen atom, we might be tempted to draw the energy levels like this:

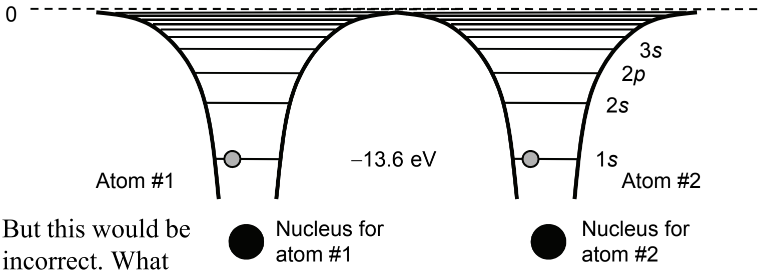
0
3*s*
2*p*
2*s*
−13.6 eV
1*s*
Atom #1
Atom #2
Nucleus for atom #1
Nucleus for atom #2

But this would be incorrect. What happens is that the electron in atom #2 is attracted by the positively charged nucleus of atom #1 and vice versa. Overall, the nucleus from atom #1 is attracted to both electrons. The nucleus from atom #2 is attracted to the same two electrons. Therefore, the two nuclei behave as if they were bonded together. The ***co***-sharing of these ***valence*** electrons and the resulting attraction of the two atoms is called a ***covalent*** **bond**.

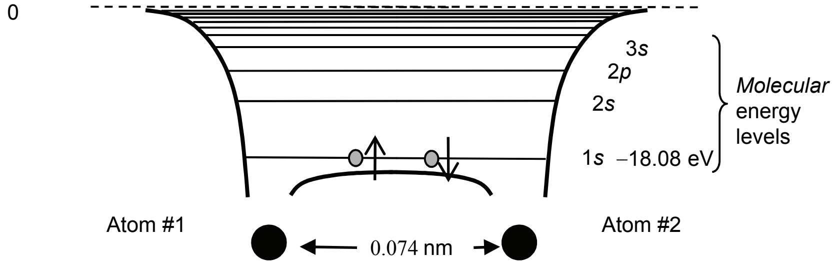
0
3*s*
2*p*
2*s*
1*s* −18.08 eV
Molecular energy levels
Atom #1
Atom #2
0.074 nm

When this attraction occurs, the sides of the "well" are reduced so the extent that the electrons are completely shared between the two atoms. The electrons (by the **Pauli exclusion principle**) have to have different spins ↑↓. Since this is energetically favourable (electrons have more space to move), then hydrogen naturally forms the **molecule** H_2. When a covalent bond is formed, energy is released (heat). To break the bond and separate the atoms, energy –(18.08–13.6) = –4.48 eV has to be supplied.

1.8 Ionisation Energy

In order for an electron to make a transition from a lower state to a higher state in an atom, it must absorb energy (via collision with a photon of sufficient energy, by heat, etc). The energy required to move the outermost (highest potential energy) electron in an isolated atom from its ground state to infinity is called the **first ionisation energy** and is usually expressed in units of eV. The resulting positively charged atom is called an **ion**.

The energy required to move the next outermost bound electron from its **ground state** to infinity is called the second ionisation energy, and so on.

Sometimes ionisation energies are expressed as kJ per mole. Since 1 eV is 1.602×10^{-19}J, then 1 eV per atom is 96.49 kJ per mole.

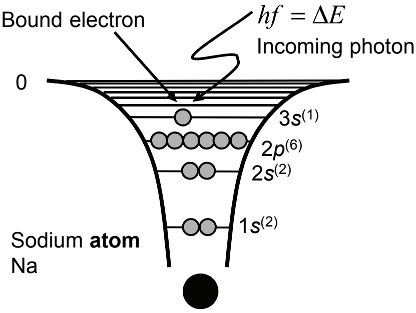

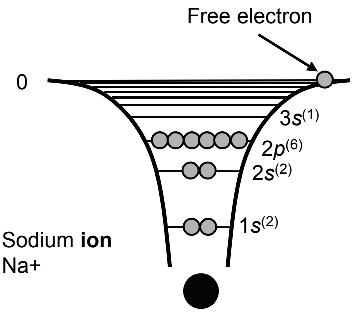

The incoming photon must have sufficient energy (in this example, > +5.1 eV for Na) to lift the electron to infinity distance from the nucleus for the atom to be ionised. If it does not, and if the incoming photon has sufficient energy to lift the electron to a higher energy level, then the atom may become "**excited**" rather than "**ionised**". If the incoming photon has insufficient energy to excite or ionise an atom, then it will just pass through or be scattered by the atom.

Important: We cannot just assign energies to these levels by calculation like we did with the hydrogen atom. The energies associated with each level in a multi-electron atom depend on the size of the atom, the atomic number, the degree of screening of outer electrons by the inner electrons, whether or not the atom is bonded with another, or exists in the gaseous phase, and most importantly, whether the level is filled or not. The ionisation energies are usually measured experimentally. The formula worked OK for hydrogen because we were dealing with just a single electron.

1.9 Electron Affinity

Consider a neutral chlorine atom. The first ionisation energy for Cl is 13 eV.

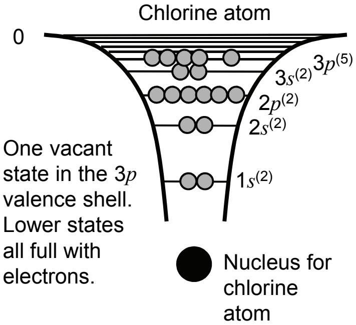
Chlorine atom
0
3s(2) 3p(5)
2p(2)
2s(2)
1s(2)
One vacant state in the 3p valence shell. Lower states all full with electrons.
Nucleus for chlorine atom

What happens when a free electron comes into contact with the neutral atom? The electron may occupy a vacancy in the 3*p* level, thus endowing the atom with a full outer energy shell. This condition is energetically desirable and is called a **noble gas configuration**. When the electron is bought from infinity into the 3*p* shell, its potential energy is lowered and experiments show that −3.7 eV is released (perhaps as heat). This energy that is released is called the **electron affinity** of the atom.

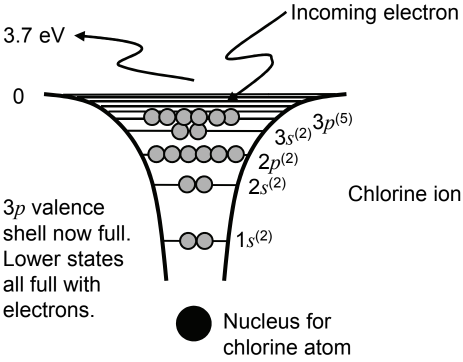
Incoming electron
3.7 eV
0
3s(2) 3p(5)
2p(2)
2s(2)
1s(2)
Chlorine ion
3p valence shell now full. Lower states all full with electrons.
Nucleus for chlorine atom

The **electron affinity** is the energy released when an electron is bought from infinity into a neutral atom. Like **ionisation energy**, it is measured experimentally (although usually indirectly). The incoming electron has to overcome the repulsion of the electrons already there, but, in the case of elements like Cl, the attainment of a **noble gas configuration** is a sufficient payoff.

1.10 Ionic Bond

Ionic bonds usually form between elements that have unpaired valence electrons. Consider the reaction between sodium and chlorine.

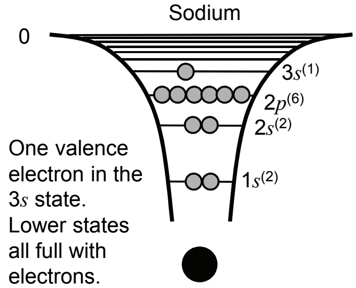
Sodium
0
3s(1)
2p(6)
2s(2)
1s(2)
One valence electron in the 3s state. Lower states all full with electrons.

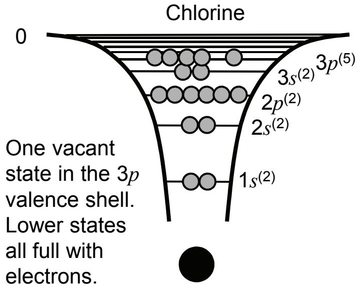
Chlorine
0
3s(2) 3p(5)
2p(2)
2s(2)
1s(2)
One vacant state in the 3p valence shell. Lower states all full with electrons.

The ionisation energy for Na is +5.1 eV. The electron affinity for Cl is –3.7 eV. Thus, if the lone electron in the 3*s* band in Na can climb over the energy barrier (–5.1 – –3.7 = –1.4 eV), then this will create an Na$^+$ sodium **ion** and a Cl$^-$ chlorine ion, leaving each with a net charge q_e. The resulting electrostatic (Coulomb) attraction is called an **ionic bond**.

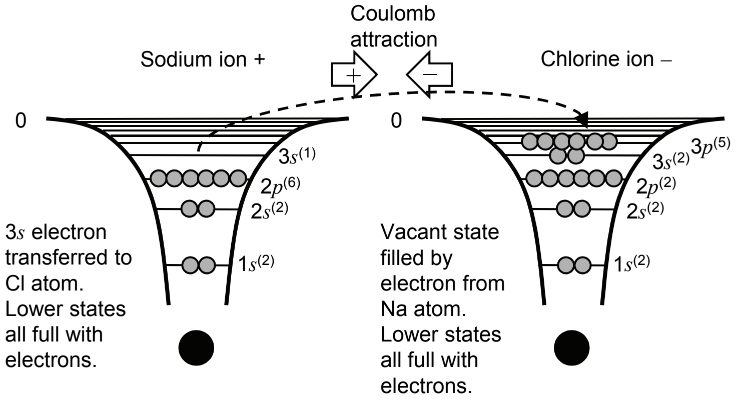
Coulomb attraction
Sodium ion +
Chlorine ion –
+
–
0
0
3s(1)
2p(6)
2s(2)
1s(2)
3s(2) 3p(5)
2p(2)
2s(2)
1s(2)
3s electron transferred to Cl atom. Lower states all full with electrons.
Vacant state filled by electron from Na atom. Lower states all full with electrons.

When the ions are formed, the attraction between them causes them to move towards each other and the electrical potential between them drops. The ions reach an **equilibrium distance** determined by the electrostatic attraction of their overall charge and the repulsion offered by their positively charged nuclei. Experiments show that the **bond energy** is –5.5 eV. This energy is lower than the –1.4 eV barrier and so an ionic bond between these two atoms is energetically favourable (since atomic systems tend settle to a state of minimum energy).

1.11 Electronegativity

It is energetically favourable for an atom to have completely filled outer energy shells (**noble gas configuration**). Metals tend to lose electrons to achieve this configuration because their few loosely bound outer shell valence electrons are easily removed (low **ionisation energy**). Non-metals, on the other hand, may have only a few vacancies in their outer shells and so prefer to gain electrons to attain a noble gas configuration.

When two atoms come together to form a **chemical bond**, the ability for one atom to gain, or pull, an electron from the other atom depends on both the ionisation energies and the electron affinities of the two atoms. A combined property that includes these two characteristics is called the **electronegativity** of the element.

Here's how it works in simplified terms. Hydrogen and chlorine can combine to form HCl by the formation of a covalent bond. An electron pair is shared between the H atom and the Cl atom. But is this sharing equal? The ionisation energy for H is +13.6 eV, and for Cl is +13 eV. The electron affinity of H is −0.75 eV and for Cl is −3.7 eV. Thus, for an electron to be transferred from Cl to H, a net energy of $13 - 0.75 = +12.25$ eV is required. For an electron to be transferred from H to Cl, a net energy of $13.6 - 3.7 = +9.9$ eV is required. Thus, in this covalent bond, the shared electron is biased towards being over near the Cl atom because less energy is required to transfer the electron from H to Cl compared to Cl to H. This unequal sharing makes the bond **polar** (since one end, the Cl end, has a net negative charge). Although we call the bond "covalent" it does have an ionic character as well. In general, there is a gradation of bond types from ionic to covalent depending on the nature of the atoms.

The **electronegativity** describes the relative ability for an atom, when it combines with another atom, to become "more negative" by more strongly capturing a shared electron or electron pair. It is measured by a variety of means, but at its simplest, depends on both the ionisation energy and the electron affinity of the atom. Electronegativities (no units) range from 4 (for Fl on the right-hand side of the periodic table) down to <1 (for elements on the left-hand side of the periodic table).

Two atoms with very different electronegativities are expected to form **ionic bonds**. Atoms with much the same electronegativities are expected to form **covalent bonds**.

An example is when an Na atom (with electronegativity 0.9) meets a Cl atom (electronegativity 3.0). The bond between them is expected to be predominantly ionic, with the chlorine atom becoming more negative. When H (with electronegativity 2.1) meets a Cl, the bond is expected to be predominantly covalent.

1.12 Metallic Bond

Consider what happens when we have two Li atoms close together. We might be first tempted to draw the potentials as:

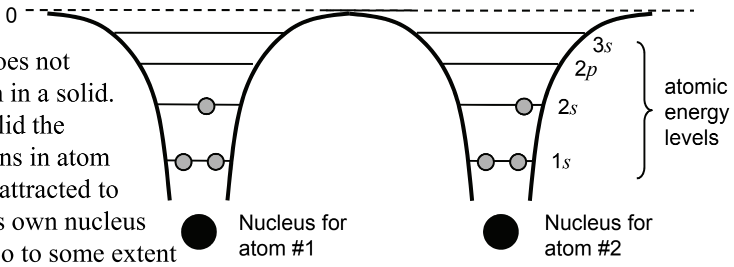

This does not happen in a solid. In a solid the electrons in atom #1 are attracted to both its own nucleus and also to some extent by the nucleus of atom #2. For example, the 1*s* electron orbiting the nucleus of atom #1 is also captured to some extent by the nucleus of atom #2. But, due to the **Pauli exclusion principle**, for each atom, we cannot have more than two electrons in the same energy level (i.e. the two 1*s* electrons for atom #1 as seen by atom #2 are no longer permitted to have energies at the 1*s* level because there are already two electrons from its own atom at that level).

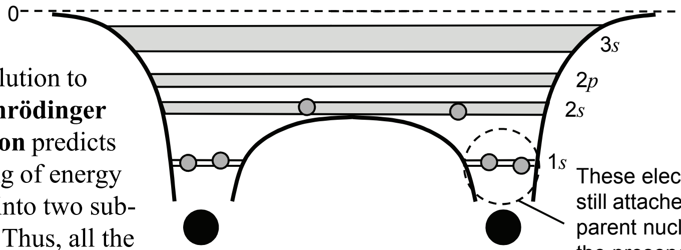

These electrons are still attached to their parent nucleus but the presence of the neighbour electrons causes their energy levels to shift.

The solution to the **Schrödinger equation** predicts splitting of energy levels into two sub-levels. Thus, all the electrons (say at the 1*s* level) for the pair of atoms can coexist by shifting their energies a little up and down so that their wave functions do not coincide and so the exclusion principle is satisfied. When there are many atoms present, the splitting of many levels into fine gradations creates a **band** of energies and if the energy barrier between the atoms is low enough for the upper bands to carry across to meet those of neighbouring atoms, the electrons in these bands become free to migrate from atom to atom. The electrons behave as if they were in a **potential well**. Because of the large range of movement of these electrons (compared to if they were still attached to their parent nuclei), their kinetic energy is lower (see Section 1.5) and a lowering of energy equates to stability – that is, the free electrons in the valence band reduce the energy of the system and so act to hold the two atoms together. This is called a **metallic bond.**

1.13 Electronic Structure of Solids

In a solid, the interacting potentials of many millions of relatively closely spaced atoms causes atomic energy levels to split into a very large number of sub-levels. The energy difference between each sub-level is so fine that each molecular level is considered to be virtually a continuous **band** of energies.

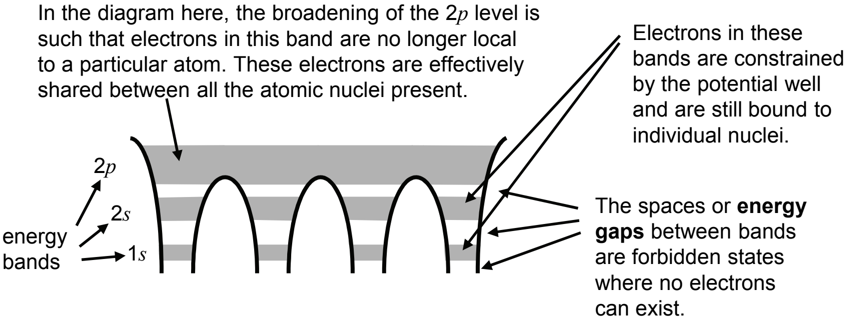
In the diagram here, the broadening of the 2*p* level is such that electrons in this band are no longer local to a particular atom. These electrons are effectively shared between all the atomic nuclei present.
Electrons in these bands are constrained by the potential well and are still bound to individual nuclei.
2*p*
2*s*
1*s*
energy bands
The spaces or **energy gaps** between bands are forbidden states where no electrons can exist.

If the highest energy band that contains electrons in the ground state (the **valence band**) in a solid is not completely full, then electrons within that band can easily move around from state to state *within the band*. Such movement can be readily obtained by applying an electric field to the solid. Such solids are thermal and electrical **conductors**.

If the valence band in a solid is full, and the next highest available band is positioned some distance away in terms of its energy levels, then the electrons within the topmost band cannot easily move from place to place or to the next highest band. Such materials are thermal and electrical **insulators**.

If the next highest available band is positioned fairly closely to the valence band, then even at room temperature, there may be sufficient thermal energy given to some electrons to be promoted to this higher level. The material becomes conducting and is a **semiconductor**. The band containing the conducting electrons is called the **conduction band**. In a conductor, the valence band is the conduction band. In a semiconductor, the conduction band (at 0K) is separated from the valence band (defined at 0K) by an **energy gap**.

Atoms in a solid generally form molecules which either arrange themselves in a regular pattern (**crystalline solids**) or the molecules do not repeat themselves in an orderly way (**amorphous solids**). When a solid forms, atoms are pushed and pulled around and settle into place when the interaction electron potentials reach a minimum level.

2. Chemical Components

Summary

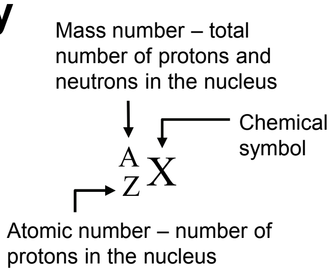
Mass number – total number of protons and neutrons in the nucleus
Chemical symbol
$^{A}_{Z}X$
Atomic number – number of protons in the nucleus

One atomic mass unit (amu) is 1/12th the mass of a carbon 12 atom. 1 amu = 1.6602×10^{-27} kg.

Avogadro's number: 6.022×10^{23} = 1 mole.

Atoms that lose or gain electrons are called ions:

- Cations (+) (electrons lost)
- Anions (−) (electrons gained)

A molecule is the smallest collection of atoms that is electrically neutral and can exist as a separate identifiable unit.

The molecular weight of a substance is the sum of the atomic weights of its constituents.

Stoichiometry is the process of accounting for the masses of atoms, molecules and compounds in chemical reactions.

2.1 Matter

Elements: cannot be reduced to simpler units by physical processes or chemical reactions. (Examples are oxygen, iron, carbon.) Generally, elements are substances which have only one kind of atom (as having a certain number of protons on the nucleus). There are just over 100 elements known, with about 95% of these occurring naturally, the remainder being created synthetically by nuclear processes.

Elements can be broadly divided into non-metals, semi-metals and metals.

Mixtures: can be separated into constituent parts by physical means. Mixtures are a product of mechanical or physical processes, such as metal alloys, suspensions, dispersions and colloids. **Homogenous** mixtures have uniform composition and appearance (examples are air, salty water) over a molecular length scale while **heterogeneous** mixtures have physically distinct regions. (Examples are fruit cake, concrete, fruit juice with pulp.)

Compounds: combinations of two or more elements in definite proportions where the composition of the compound is uniform throughout. Compounds have a chemical structure, the atoms being held in place by chemical bonds. Compounds keep their chemical identity when altered physically, and can only be reduced to their constitutive elements by chemical reactions. (Examples are water, salt, sulphuric acid.)

2.2 Atomic Weight

Except in the case of nuclear transformations, we can usually treat atoms as being much like the Bohr atom. Atoms are characterised by an **atomic number**, a **chemical symbol**, and the **mass number**.

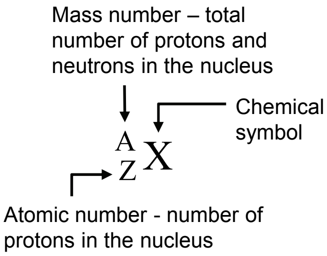
Mass number – total number of protons and neutrons in the nucleus
Chemical symbol
A
Z
X
Atomic number - number of protons in the nucleus

$m_e = 9.1096 \times 10^{-31}$ kg

$m_p = 1.6726 \times 10^{-27}$ kg

$m_N = 1.6749 \times 10^{-27}$ kg

The mass number is approximately equal to the atomic weight of the element. The **atomic weight** of an element is the mass (in relative terms) of an "average" atom of an element. The more precise term **atomic mass** is used to denote the absolute mass of a particular atom (in kg).

Since the **mass of an electron** m_e is very small nearly all of the mass of an atom is contributed by the **protons** and **neutrons** in the nucleus. Protons and neutrons have very nearly the same mass, so the number of neutrons in an atom is given by Z – A. However, it is found in nature that many atoms of the same element with atomic number Z have different mass numbers A because of having a different number of neutrons in the nucleus. Each type of atom A of an element is called a **nuclide**, or **isotope**.

By international agreement, one **atomic mass unit** (**amu**) is 1/12th the mass of a single carbon 12 atom. The atomic weight (or **relative atomic mass**) of an element is the ratio of the average mass per atom of the element to 1/12 of the mass of an atom of ^{12}C.

Note: The atomic weight of an element is not found from adding the masses of protons, neutrons and electrons. When an atom is formed, some of the mass is used as nuclear binding energy via the **Einstein** relationship $E = mc^2$.

The **carbon 12** isotope has 6 protons and 6 neutrons in the nucleus. The **carbon 13** atom has 6 protons and 7 neutrons in the nucleus. Both isotopes have 6 electrons. Experiments show that carbon in nature consists of 1.11% ^{13}C and 98.89% ^{12}C. The atomic weight of carbon is determined to be 12.011.

In chemistry, much of the arithmetic of chemical reactions involves the atomic weights for the atoms that take part. However, the actual nature of the chemical reactions is due to the number and arrangement of the outer-shell **valence electrons** in the atoms rather than the heavy nuclei. "Valence" is a word that means "power" and in some sense, it is the number and arrangement of valence electrons that give an atom "combining power" to form **chemical bonds** with other atoms.

2.3 Ions

Atoms that lose or gain electrons are called **ions**. When an atom loses one or more electrons, it is called a **cation**. When an atom gains one or more electrons, it is called an **anion**.

Examples of ions:

Atom	Z (atomic number)	No. electrons gained or lost	No. electrons remaining	Formula of ion
H	1	-1	0	H^+
Na	11	-1	10	Na^+
Ca	20	-2	18	Ca^{2+}
F	9	+1	10	F^-
S	16	+2	18	S^{2-}

When an ion is formed, the atom is no longer electrically neutral. The formula of the ion signifies the net electronic charge. Atoms may lose electrons, especially unpaired outer shell or valence electrons, by physical removal, absorption of **ionising radiation**, the electron being taken by another type of atom which has a stronger attraction for it, and so on.

Metallic elements tend to lose one or more electrons when they combine with other elements and so form cations. Non-metallic elements tend to gain electrons to form anions. When naming compounds, it is usual to list the more metallic element first. In **binary ionic compounds**, we list the metal cation first followed by the non-metal anion with an "ide" suffix. For covalent compounds, the more metallic element is listed first.

Sodium chloride
Carbon monoxide }
Carbon dioxide }
Nitrogen oxide

Where there is more than one possible compound, prefixes are used to distinguish them.

2.4 Molecules

Most elements bond with others to form **molecules**. A molecule is the smallest collection of atoms, that is electrically neutral, and can exist as a separate identifiable unit. The concept was first proposed by **Avogadro** in 1811, who, on the basis of observations made by Dalton and Gay-Lussac proposed that atoms combine to form molecules, *and* at the same temperature and pressure, equal volumes of *all* gases contain the same number of molecules. **Avogadro's law** provides the justification for a method of determining relative molecular weights →

because there was now a way to obtain equal numbers of different molecules by using different gases all at the same temperature, pressure and volume.

A molecule is most conveniently described in terms of a **molecular formula**.

H_2O

A water molecule consists of two H atoms and one O atom.

In some cases, such as in a crystalline solid, the concept of a molecule is not appropriate because the atoms which make up the substance are arranged in a regular repeating pattern. In this case, we speak of the **empirical formula**.

NaCl

A sodium chloride crystal consists of equal numbers of sodium and chlorine atoms arranged in a regular array or crystal lattice.

The **molecular weight** (or **relative molar mass**, or **relative molecular mass**) M_r of a substance is the sum of the atomic weights of its constituents.

If we had a certain mass of an element (or a molecule), say 12 grams of ^{12}C, how many atoms would this be? This number is called **Avogadro's number** N_A and has the value 6.0220943×10^{23}. N_A is found by experimental methods.

By international agreement, one **atomic mass unit** (**amu**) is 1/12th the mass of a carbon 12 atom.

12 g of $^{12}C = 6.0220943 \times 10^{23}$ atoms; therefore the mass of one ^{12}C atom is:

"But!", you might say, "How can you have 6.022... atoms if atoms are indivisible? Shouldn't this be a whole number?" Look at the 10^{23}. If you write Avogadro's number out in full, it *is* a whole number.

$$\begin{aligned}\text{mass} &= \frac{0.012}{6.02\times10^{23}}\text{kg}\\ &= 1.99\times10^{-23}\text{kg}\end{aligned}$$

It is convenient when working with quantities usually involved in chemical reactions to round down Avogadro's number to 6.02×10^{23}.

2.5 Mole

Chemical reactions usually involve large numbers of atoms and molecules. It is convenient to have a unit of measure that relates weight (i.e. mass in grams or kilograms, etc) of bulk chemicals with atomic or molecular weight. By definition, this unit of measure is called a **mole**, and it is the Avogadro number of atoms or molecules.

12 grams of ^{12}C contains 6.02×10^{23} atoms and is called 1 mole of ^{12}C

18 grams of H_2O also contains 6.02×10^{23} molecules and is called 1 mole of H_2O

The word "mole" is a shorthand way of saying "**Avogadro's number**". That is, it is easier to say "Consider one mole of sodium atoms" rather than "Consider one Avogadro's number of sodium atoms" or "Consider 6.02×10^{23} of sodium atoms". To find out the mass in grams of a mole of some atoms, we simply look up the **atomic weight**. To find out the mass in grams of a mole of some molecules, we simply use the molecular weight.

→ because the atomic weight (in amu) is expressed relative to the ^{12}C atom, and thanks to Avogadro, we know that 12 g of ^{12}C has N_A atoms.

Moles, why bother? It's convenient. One mole of *anything* is 6.02×10^{23} of the "anythings". When we are talking about atoms or molecules, we know that one mole has a mass equal to the atomic weight of the element, or the molecule because that's the way Avogadro's number was determined. It's the link between the *relative* atomic mass and the *actual* mass in grams.

Units of chemical accounting:

- One **atomic mass unit** (amu) is 1/12th the mass of a carbon 12 atom by international agreement. 1 amu is $1.66020943 \times 10^{-27}$ kg.
- The **atomic weight** (or **relative atomic mass**) of an element is the ratio of the average mass per atom of the element to 1/12 of the mass of an atom of ^{12}C.
- The total sum of the atomic weights for a molecule of substance is called the **molecular weight** or **relative molar mass** M_r.
- The molecular/atomic weight expressed in grams contains one mole of molecules/atoms. This is Avogadro's number 6.0220943×10^{23}.
- The **molar mass** M of an element is the atomic weight of the element expressed in g/mol. The molar mass of a compound is the sum of the atomic weights of the constituent elements expressed in g/mol.

2.6 Compounds

Compounds consist of two or more different elements which are chemically combined in a definite fixed ratio (as distinct from mixtures, which can have any proportions of ingredients). There are generally two types of compounds:

Molecular compounds are composed of atoms held together by chemical bonds in fixed proportion and are electrically neutral overall. For example, the water molecule H_2O is a molecular compound and the **molecular formula** shows the proportions of the atomic species within it.

Name	Molecular formula
Methane	CH_4
Carbon dioxide	CO_2
Ammonia	NH_3
Acetylene	C_2H_2
Ethylene	C_2H_4
Ethane	C_2H_6
Water	H_2O
Benzene	C_6H_6
Ethanol	C_2H_5OH
Naphthalene	$C_{10}H_8$
Aspirin	$C_9H_8O_4$

There is a distinction to be made between the **molecular formula** and the **empirical formula** of a compound. The empirical formula gives the ratio of component atoms in the lowest possible numerical terms. The empirical formula shows the ratio of component atoms as found to exist in practice. In many cases, the formulae are the same, but this cannot be always assumed. For example, the empirical formula for benzene is CH but the molecular formula is C_6H_6.

Ionic compounds are composed of charged atoms, or a charged group of atoms, of opposite sign that are held together by **ionic bonds**. When an ionic compound is formed from the joining of a charged group of atoms, it is said to be a **complex ion** (or a polyatomic ion, or a molecular ion). The ion with the positive charge in the compound is the **cation**. The ion with the negative charge is the **anion**. The ionic formula gives the ratio of anions and cations. The ionic formula is electrically neutral overall.

Name	Ions formed	Formula
Magnesium chloride	Mg^{2+} Cl^-	$MgCl_2$
Silver sulphate	Ag^+ SO_4^{2-}	Ag_2SO_4
Ammonium sulphate	NH_4^+ SO_4^{2-}	$(NH_4)_2SO_4$
Chromium (III) hydroxide	Cr^{3+} OH^-	$Cr(OH)_3$
Sodium chloride	Na^+ Cl^-	$NaCl$

2.7 Cations

+1

Name	Formula
Hydrogen	H^+
Lithium	Li^+
Sodium	Na^+
Potassium	K^+
Rubidium	Rb^+
Caesium	Cs^+
Silver	Ag^+
Copper	Cu^+
Mercury (I)	Hg_2^{2+}
Ammonium	NH_4^+

+2

Name	Formula
Beryllium	Be^{2+}
Magnesium	Mg^{2+}
Calcium	Ca^{2+}
Strontium	Sr^{2+}
Barium	Ba^{2+}
Lead	Pb^{2+}
Zinc	Zn^{2+}
Cadmium	Cd^{2+}
Nickel	Ni^{2+}
Manganese	Mn^{2+}
Tin (II)	Sn^{2+}
Iron (II)	Fe^{2+}
Mercury (II)	Hg^{2+}
Cobalt (II)	Co^{2+}
Chromium (II)	Cr^{2+}
Copper (II)	Cu^{2+}

+3 , +4

Name	Formula
Aluminium	Al^{3+}
Bismuth	Bi^{3+}
Iron (III)	Fe^{3+}
Cobalt (III)	Co^{3+}
Chromium (III)	Cr^{3+}
Tin (IV)	Sn^{4+}

In the naming of chemical compounds, it is customary to write the name of the most metallic element first, followed by the more non-metallic one. The second element or ion is given the *ide* suffix (e.g. sodium chloride).

Prefixes

1 – mono
2 – di
3 – tri
4 – tetra
5 – penta
6 – hexa

2.8 Anions

−1

Name	Formula
Fluoride	F^-
Chloride	Cl^-
Bromide	Br^-
Iodide	I^-
Hydroxide	OH^-
Nitrite	NO_2^-
Nitrate	NO_3^-
Chlorate	ClO_3^-
Perchlorate	ClO_4^-
Cyanide	CN^-
Permanganate	MnO_4^-
Thiocyanate	NCS^-
Bicarbonate	HCO_3^-
Bisulphate	HSO_4^-
Dihydrogen phosphate	$H_2PO_4^-$
Acetate	$CH_3CO_2^-$

−2

Name	Formula
Oxide	O^{2-}
Peroxide	O_2^{2-}
Sulphide	S^{2-}
Sulphate	SO_4^{2-}
Sulphite	SO_3^{2-}
Thiosulphate	$S_2O_3^{2-}$
Carbonate	CO_3^{2-}
Oxalate	$C_2O_4^{2-}$
Chromate	CrO_4^{2-}
Dichromate	$Cr_2O_7^{2-}$
Monohydrogen phosphate	HPO_4^{2-}

−3 , −4

Name	Formula
Nitride	N^{3-}
Phosphate	PO_4^{3-}
Hexacyanoferrate (III)	$Fe(CN)_6^{3-}$
Hexacyanoferrate (II)	$Fe(CN)_6^{4-}$

For polyatomic anions, the assignment of a suffix depends on the **oxidation number** (see Section 4.8) of the central non-metal atom.

Suffixes	
ide	
ite	ous
ate	ic

In some cases, the oxidation number of the metal is stated in roman numerals to distinguish compounds which would otherwise have the same name (e.g. iron (II) chloride and iron (III) chloride).

2.9 Chemical Equation

When atoms or molecules (the **reactants**) combine with other atoms or molecules to form compounds (the **products**), a **chemical reaction** is said to take place. Chemical reactions involve the breaking and formation of **chemical bonds**.

$$\text{Reactants} \rightarrow \text{Products}$$

This is the forward reaction

$$\text{Products} \rightarrow \text{Reactants}$$

This is the reverse reaction

Often one can tell if a chemical reaction has occurred by observation of products that are considerably different in appearance than the reactants. Examples are the formation of a **precipitate**, evolution of gas, change of colour, production of heat and so on. Often, chemical reactions proceed spontaneously, other times they have to be forced to proceed by the input of energy or the lowering of energy barriers.

Chemical equations are written using chemical formulae where the number of atoms or molecules is written before the formula:

$$2CO + O_2 \rightarrow 2CO_2$$

This equation says that two molecules of carbon monoxide react with one molecule of oxygen to product two molecules of carbon dioxide. This is a **balanced equation**, because the number of atoms in each reactant equals the number of atoms in each product. That is, we have two carbon atoms on the left-hand side and two carbon atoms on the right-hand side, and two oxygen atoms on the left side and two on the right side. We can also say that two **moles** of carbon monoxide combine with one mole of oxygen to form two moles of carbon dioxide. Since the mass of a mole of an atom or a molecule is the atomic or molecular weight expressed in grams, a balanced equation is balanced in terms of both number of atoms and mass on both sides of the equation. The net electric charge must also be the same on both sides of the equation.

Usually, reactions proceed until there is **chemical equilibrium**. Dynamic chemical equilibrium occurs when the rate of the forward reaction becomes equal to the rate of the reverse reaction:

$$\text{Reactants} \leftrightarrow \text{Products}$$

Dynamic equilibrium

By rate, we mean the rate of formation of products in units of mol/s.

2.10 Stoichiometry

Stoichiometry is the process of accounting for the masses of atoms, molecules and compounds in chemical reactions. The basic component of such accounting is the arrangement of chemical formulae in chemical equations. Stoichiometric calculations are performed in a logical sequence:

1. Write a balanced equation for the reaction.
2. Convert known masses of reactants and/or products into moles.
3. Calculate the number of moles of the reactants and/or products whose masses are unknown.
4. Calculate the masses of the unknown reactants and/or products.

Balancing an equation involves both a balance of atoms and also electric charge. This is particularly important in the case of ionic equations where spectator ions may be present and do not take part in the reaction. Some examples of unbalanced equations (with spectator ions) and the corresponding **net ionic equations** are:

$$\mathrm{Pb(NO_3)_2(aq) + KCl(aq) \rightarrow PbCl_2(s) + K^+ + NO^{3-}}$$

$$\mathrm{Pb^{2+}(aq) + 2Cl^-(aq) \rightarrow PbCl_2(s)}$$

$$\mathrm{Ni(OH)_2(s) + HNO_3(aq) \rightarrow NiNO_3(aq) + H_2O}$$

$$\mathrm{Ni(OH)_2(s) + 2H^+ \rightarrow Ni^{2+} + 2H_2O}$$

$$\mathrm{NaCO_3(aq) + HNO_3(aq) \rightarrow NaNO_3(aq) + H_2O + CO_2}$$

$$\mathrm{CO_3^{2-} + 2H^+ \rightarrow 2H_2O + CO_2}$$

Practical difficulties can arise when attempting to determine masses of reactants and products. For example, some reactants may be present in excess (more present than combine with one or other reactants). More than one reaction may take place at the same time, or a reaction may proceed in a series of intermediate steps. Products (such as gases) may escape and alter the equilibrium state of the reaction. The nature of the reactants may prevent mixing of the reactants and so there is incomplete yield of product.

2.11 Example

Acetylene gas is used together with oxygen in an "oxy" welding set. It is prepared from a reaction of calcium carbide and water:

$$\mathrm{CaC_2 + H_2O \rightarrow Ca(OH)_2 + C_2H_2} \quad \text{Unbalanced eqn.}$$

Balance this equation and determine the mass of acetylene that is produced from the reaction of 130 g CaC_2 and 100 g water.

$$\mathrm{CaC_2 + 2H_2O \rightarrow Ca(OH)_2 + C_2H_2} \quad \text{Balanced eqn.}$$

$$\begin{aligned}
\mathrm{CaC_2} &= 130\,\mathrm{g} \\
&= \frac{130}{40.1 + 2(12)} \\
&= 2.03\,\text{moles} \\
\mathrm{H_2O} &= 100\,\mathrm{g} \\
&= \frac{100}{2 + 16} \\
&= 5.56\,\text{moles}
\end{aligned}$$

But according to the balanced equation, for each mole of CaC_2, two moles of H_2O are required. Therefore, 2(2.03) = 4.06 moles of H_2O required, and so the mass of C_2H_2 formed would be:

$$\begin{aligned}
\mathrm{C_2H_2} &= 2.03\,\text{moles} \\
2.03 &= \frac{m}{2(12) + 2(1)} \\
&= 52.8\,\mathrm{g}
\end{aligned}$$

3. The Periodic Table

Summary

Atomic radius decreases →

Atomic radius increases ↓

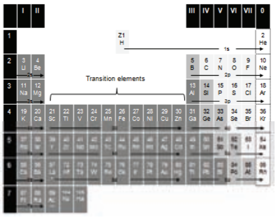
Transition elements

Bohr radius

$$r = n^2\left(\frac{\varepsilon_o h^2}{\pi m_e Z q_e^2}\right)$$

Ionisation energy increases →

Ionisation energy decreases ↓

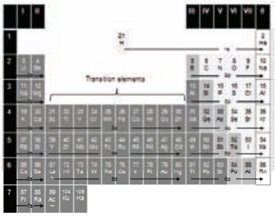
Transition elements

Noble gases

He	$1s^2$
Ne	He$2s^2 2p^6$
Ar	Ne$3s^2 3p^6$
Kr	Ar$3d^{10} 4s^2 4p^6$
Xe	Kr$4d^{10} 5s^2 5p^6$
Rn	Xe$5d^{10} 6s^2 6p^6$

Electronegativity increases →

Electronegativity decreases ↓

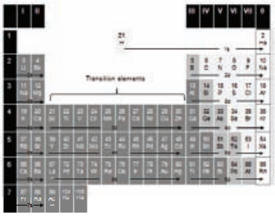
Transition elements

3.1 Electron Configuration

In a **multi-electron atom**, the electrons do not find themselves in a **Coulomb potential** of the form $1/r$ as in the **hydrogen atom** due to the **screening effect** of inner-shell electrons on the outer-shell electrons. In a multi-electron atom, the potential in which an electron finds itself (and hence its energy) depends on not only n (as in the single-electron atom), but also the second quantum number l. The arrangement of electrons in atoms in the **ground state** can be complicated, but follows certain rules:

1. Electrons occupy the lower energy levels first, before occupying higher levels.
2. Electrons at a particular energy level have opposite spin (**Pauli exclusion principle**).
3. Electrons tend to occupy an energy level as single electrons before pairing up (**Hund's rule**).

Examples of the ground-state electron configuration of the first few elements arranged in order of increasing **atomic number** Z

Hydrogen	$1s^1$	
Helium	$1s^2$	
Lithium	$1s^2 2s^1$	
Beryllium	$1s^2 2s^2$	
Boron	$1s^2 2s^2 2p^1$	
Carbon	$1s^2 2s^2 2p^2$	
Nitrogen	$1s^2 2s^2 2p^3$	
Oxygen	$1s^2 2s^2 2p^4$	
Fluorine	$1s^2 2s^2 2p^5$	
Neon	$1s^2 2s^2 2p^6$	
Sodium	$1s^2 2s^2 2p^6 3s^1$	
Magnesium	$1s^2 2s^2 2p^6 3s^2$	
Aluminium	$1s^2 2s^2 2p^6 3s^2 3p^1$	
Silicon	$1s^2 2s^2 2p^6 3s^2 3p^2$	
Phosphorus	$1s^2 2s^2 2p^6 3s^2 3p^3$	
Sulphur	$1s^2 2s^2 2p^6 3s^2 3p^4$	
Chlorine	$1s^2 2s^2 2p^6 3s^2 3p^5$	
Argon	$1s^2 2s^2 2p^6 3s^2 3p^6$	
Potassium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$	Note 4*s* has lower energy than 3*d* in K.
Calcium	$1s^2 2s^2 2p^6 3s^2 3p^6 4s^2$	Note 4*s* has lower energy than 3*d* in Ca.
Scandium	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^1 4s^2$	Keep sub-shells together even if no longer in increasing energy order.

Electrons spread out to occupy unfilled sub-shells before pairing up. For example, in oxygen, the arrangement of electrons is:

1*s*	2*s*	2*p*		
↑↓	↑↓	↑↓	↓	↓

When we write the electron configuration of elements, we keep the principal quantum numbers together even if they are not arranged in energy order. For example, in potassium, the 4*s* sub-shell actually has a lower energy than the 3*d* sub-shell, but we usually write the 3*d* along with the 3*s* and 3*p* sub-shells.

3.2 Periodic Law

In the nineteenth century, it was noticed that if elements were arranged in order of increasing **atomic weight**, certain properties of the elements tended to occur at periodic intervals. This is a striking observation. It is now known that **periodicity** occurs when the elements are arranged in increasing **atomic number** Z. Consider the first few elements and the property of reactivity:

1 Hydrogen	$1s^1$
2 Helium	$1s^2$
3 Lithium	$1s^22s^1$
4 Beryllium	$1s^22s^2$
5 Boron	$1s^22s^22p^1$
6 Carbon	$1s^22s^22p^2$
7 Nitrogen	$1s^22s^22p^3$
8 Oxygen	$1s^22s^22p^4$
9 Fluorine	$1s^22s^22p^5$
10 Neon	$1s^22s^22p^6$
11 Sodium	$1s^22s^22p^63s^1$
12 Magnesium	$1s^22s^22p^63s^2$
13 Aluminium	$1s^22s^22p^63s^23p^1$
14 Silicon	$1s^22s^22p^63s^23p^2$
15 Phosphorus	$1s^22s^22p^63s^23p^3$
16 Sulphur	$1s^22s^22p^63s^23p^4$
17 Chlorine	$1s^22s^22p^63s^23p^5$
18 Argon	$1s^22s^22p^63s^23p^6$
19 Potassium	$1s^22s^22p^63s^23p^64s^1$
20 Calcium	$1s^22s^22p^63s^23p^64s^2$
21 Scandium	$1s^22s^22p^63s^23p^63d^14s^2$

Another example of periodicity. Lithium is a highly reactive metal and so are sodium and potassium.

An example of periodicity. Helium is not very reactive (just about inert) and so are neon, and also argon.

Chemical reactivity is not the only property that tends to occur periodically. Other properties such as whether an element is a metal or a non-metal, whether they form certain compounds with other elements and so on can also occur periodically.

Initially it seems that the chemical reactivity of elements repeats in intervals of eight in atomic number Z. However, when the elements are arranged more completely we see that the periodicity actually depends upon the arrangement of outer-shell **valence electrons** – a not surprising observation in hindsight because it is the valence electrons that are involved in the formation of **chemical bonds**. For example, neon (10), argon (18), krypton (36), xenon (54) and radon (86) all have eight electrons in the outer shell and are inert. Na and K have one valence electron and easily form positive ions.

3.3 Periodic Table

Periodicity in chemical (and sometimes physical) properties can more readily be appreciated when elements are arranged in a table. Vertical columns in the table (**groups**) show elements with similar properties. Each horizontal row is called a **period**. The table also indicates (approximately) the way in which electron shells are occupied.

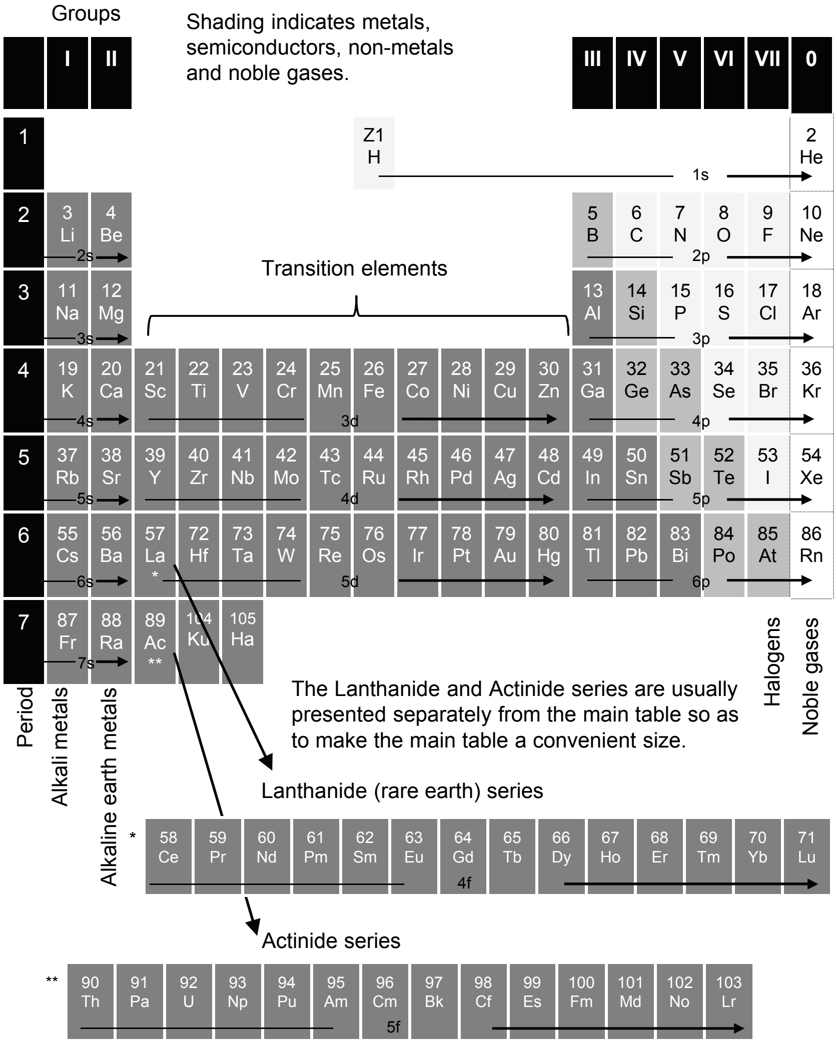

Shading indicates metals, semiconductors, non-metals and noble gases.

Transition elements

The Lanthanide and Actinide series are usually presented separately from the main table so as to make the main table a convenient size.

	I	II											III	IV	V	VI	VII	0
1							Z1 H											2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89 Ac **	104 Ku	105 Ha													

Lanthanide (rare earth) series

*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu

Actinide series

**	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

3.4 Groups

Hydrogen is placed in a separate position in the periodic table.

Hydrogen has one electron, which in the ground state is in the 1*s* orbital. This unpaired electron makes hydrogen a reactive element that is able to form compounds with many other elements, including itself, to form a gas H_2.

${}^{1}_{1}\mathrm{H}$	Protium 99.9% abundance
${}^{2}_{1}\mathrm{H}$	Deuterium (heavy hydrogen)
${}^{3}_{1}\mathrm{H}$	Tritium (radioactive)

Group I elements are the **alkali metals**. Having one weakly bound outer valence electron, they are highly reactive, and metallic in character. Elements in this group usually form M^+ ions which are water soluble.

Group II elements are the **alkaline earth metals**. Elements in this group usually form M^{2+} ions and are generally insoluble, occurring naturally as silicates, carbonates, sulphates and phosphates.

The **transition elements** generally occupy the positions between Groups II and III. All these elements are metals since their outermost shells contain only a few electrons. However, unlike alkali metals, transition metals are hard, brittle and have a high melting point (with the exception of mercury).

Group III elements are also considered metals (although boron has only semi-metallic properties). With the exception of boron, they form M^{3+} ions and are relatively soft.

Group IV elements range from non-metal, semi-metals (or semiconductors) to metals down the group with increasing size of atom (and hence screening effect of electrons). All have four electrons in their outer shell. **Carbon** is responsible for the formation of hydrocarbons and derivatives, the basis of life on Earth. **Silicon**, unlike carbon, tends to form bonds with oxygen and is the basis for most of the minerals of the Earth.

Group V elements range from non-metallic to metallic down the group with increasing atomic size and five valence electrons.

Group VI elements show little metallic character due to the increasing ionisation potentials as we go across the periodic table. Increasing atomic size going down the group confers some metallic properties to Se, Te and Po.

Group VII elements are referred to as the halogens. They all have a high tendency to complete their electron shells by forming salts. These elements are largely non-metallic, and mostly reactive.

Group 0 elements are the **noble gases**, so called because of their inertness, although they are able to form compounds with oxygen and fluorine under certain conditions. Argon was discovered in 1894.

3.5 Energy Levels

The ordering of the outer electron shells in terms of energy is the basis for interpretation of the **periodic table**. Calculations show that the ordering of the energy shells proceeds in the following sequence from lowest (more negative) to highest (less negative potential):

1*s*, 2*s*, 2*p*, 3*s*, 3*p*, 4*s*, 3*d*, 4*p*, 5*s*, 4*d*, 5*p*, 6*s*, 4*f*, 5*d*, 6*p*, 7*s*, 5*f*, 6*d*

From 1*s* to 3*p*, it is easy to understand because the electron shells are at an increasing radial distance from the nucleus (*n* quantum number dominates the energy distribution just as in the one-electron atom). The *n* quantum number determines the distance the electrons are from the nucleus. The *l* quantum number determines the shape of the electron shell. When $l = 0$, we have an *s* sub-shell which is spherical. When $l = 1$, we have a p sub-shell which is lobed. $l = 2$ gives *d* sub-shells which are four-lobed geometries. The significance of this is that in the fourth period, in K and Ca, the 4*s* sub-shell has a lower energy than the 3*d* sub shell. In these elements, electrons prefer to be in 4*s* rather than 3*d* in the ground state even though the radius of the 4*s* shell is larger than the 3*d* shell.

$l = 0$	*s*
$l = 1$	*p*
$l = 2$	*d*
$l = 3$	*f*
$l = 4$	*g*
$l = 5$	*h*

↓

This is a consequence of the increasing dominance of the *l* quantum number in the *d* sub-shell in determining the energy for a sub-shell. The *l* quantum number is connected with the angular momentum of the electrons and so, much like the case where the angular momentum for a rotating wheel is greater if the mass is concentrated at the outer edge compared to the case when the mass is evenly distributed, the concentration of mass of electrons in *d*-shaped shells results in a greater kinetic energy component to the total energy compared to spherical *s* shells.

For the first row of the **transition elements**, the 3*d* levels become occupied only after the 4*s* levels are filled (except for Cr). In these elements, the electrons in the 4*s* shell, being further from the nucleus (but having lower energy than the 3*d* electrons) shield those in the inner 3*d* shell and so the chemical properties for these elements are all very similar since in each case, the valence electrons are the 4*s* *outer* electrons while it is the number of inner 3*d* electrons that is changing. It is the outer **valence electrons** (with the highest principal quantum number) that interact with other atoms to form chemical bonds.

The ordering of shells given above is not the ordering of *all* the energy levels *within* an atom, only the *energy* ordering of the *outer* shells. For example, in K, the 4*s* shell is at a lower energy than the 3*d* shell ($3s^23p^64s^1$). By the time we get to Zn, the 3*d* shell is at a lower energy than the 4*s* shell ($3d^{10}4s^2$). There is no one sequence of energy shells that applies to all elements.

3.6 Noble Gases

The **noble gas** elements are extremely stable and generally do not form compounds with any other elements. The notable feature of noble gas elements is the interesting property of having full outer electron energy shells. Consider the second period elements. As we go from left to right, the number of electrons in the energy level with principal quantum number $n = 2$ is increasing. At neon, we have eight valance electrons: $2s^2$ $2p^6$. A group of eight (**octet**) valence electrons that completely fills an energy shell is very energetically stable and is called the **noble gas configuration**.

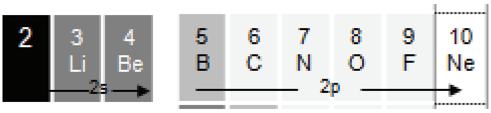

An atom with one more valence electron makes the element sodium, a highly reactive metal. Sodium is highly reactive because its single $3s^1$ valence electron is easily removed so that the sodium **cation** has a noble gas configuration in its outer shell. An atom with one electron less than sodium is the highly reactive gas fluorine. Fluorine readily attracts an electron from another atom to form an **anion** to achieve a **noble gas configuration**.

He	$\mathbf{1s^2}$
Ne	$1s^2\mathbf{2s^22p^6}$
Ar	$1s^22s^22p^6\mathbf{3s^23p^6}$
Kr	$1s^22s^22p^63s^23p^63d^{10}\mathbf{4s^24p^6}$
Xe	$1s^22s^22p^63s^23p^63d^{10}4s^24p^64d^{10}\mathbf{5s^25p^6}$
Rn	$1s^22s^22p^63s^23p^63d^{10}4s^24p^64d^{10}4f^{14}5s^25p^65d^{10}\mathbf{6s^26p^6}$

The noble gas configuration is where there is a completely filled energy shell and the next available higher energy level is an *s* shell. This is true for He where the 1*s* shell is filled and the next energy level is the 2*s* shell. For the other noble gases, we have completely filled *p* shells and the next highest energy level is the *s* sub-shell for the next quantum number *n*. There is a large energy gap between a *p* sub-shell and the next highest *s* shell and this gives the noble gas elements a high **ionisation energy**. Because all the occupied energy shells are filled, the electric charge distribution within noble gas elements is symmetric, resulting in no external electric field. The total angular momenta within these atoms adds up to zero, resulting in no external magnetic fields. These elements find it very difficult to form bonds with other atoms because they have little opportunity to interact electrostatically magnetically, or energetically.

3.7 Atomic Size

The size of an atom cannot be precisely stated since the location of the electrons can only be described statistically. **Atomic size**, for practical purposes, can be defined as the most probable distance from the nucleus to the most outer-shell electron but depends on whether the atom is free or combined with another.

Electrons are attracted to the nucleus by **electrostatic forces**, but are constrained to occupy defined orbits, or energy shells, by the principles of **quantum mechanics**. The electrostatic attraction between the nucleus and a near-shell electron is much larger than between the nucleus and an outer-shell electron, not only because of the increased distance, but also because of the **screening effect** that the inner electrons have on the outer electrons.

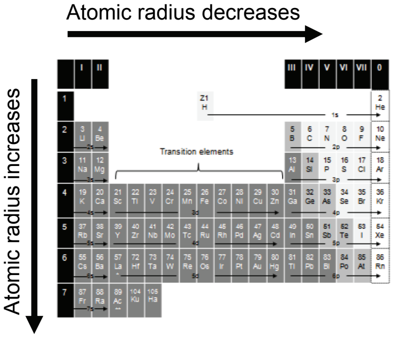

As we go across a **period** in the periodic table, the atomic number increases and so the magnitude of the positive nuclear charge also increases, and therefore so does the magnitude of the **electric field** within which the electrons find themselves. However, across any one period, the outer electrons exist in energy levels for one particular quantum number *n*. For example, in the 3rd period, the 3*s* and 3*p* levels are occupied. Because the electric field is also increasing as we go from left to right, so the electrons are acted upon by a larger Coulomb force and the **atomic radius** decreases across a period. However, because of the **screening effect** offered by the inner-shell electrons, the decrease in size is not as large as expected on the basis of the nuclear charge alone.

For a given vertical **group** of elements in the periodic table, the size of atoms increases as the atomic number Z increases (as we go down the column). We may therefore expect that the atomic radii would decrease but as we go down a group, electrons are filling states with a higher quantum number *n* and this more than offsets the effect of increasing Z.

Therefore **atomic radius** increases as we go down a group in the periodic table.

$$r = n^2\left(\frac{\varepsilon_o h^2}{\pi m_e Z q_e^2}\right)$$

The Bohr model of the atom shows how the atomic radius increases in proportion to n^2 and decreases (inversely) with Z.

3.8 Covalent Radii

When atoms form a **covalent bond**, the nuclei of the two atoms move close together (see Section 1.7) compared to two isolated atoms adjacent to each other.

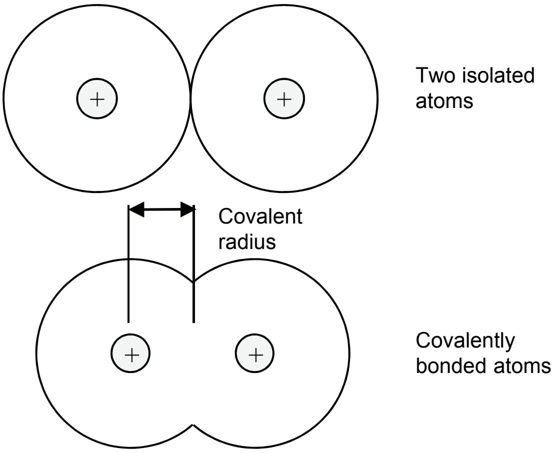

The **covalent radius** of an atom is defined as one half the distance between the nuclei of two identical *covalently bonded* atoms. The actual covalent radius of an atom in a molecule where the atoms are not identical depends on the atoms involved. The general term "**atomic radius**" usually means the covalent radius (where the two atoms are identical) and this serves as a useful benchmark for comparison with other atomic sizes.

For metallic atoms, the atomic or **metallic radius** is usually defined as half the distance between two nuclei of the atoms in the solid metallic state.

The shapes of the orbitals in which the valence electrons participate in bonding is important. The *s* orbital has a somewhat spherical shape centred on the nucleus, while the *p* orbitals have a dumbbell shape along the three coordinate axes. The shapes of the atomic orbitals involved in the formation of a particular bond determine the shape of the molecule.

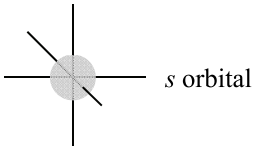

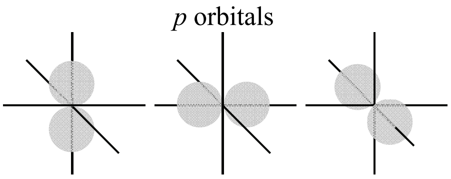

3.9 Ionic Radii

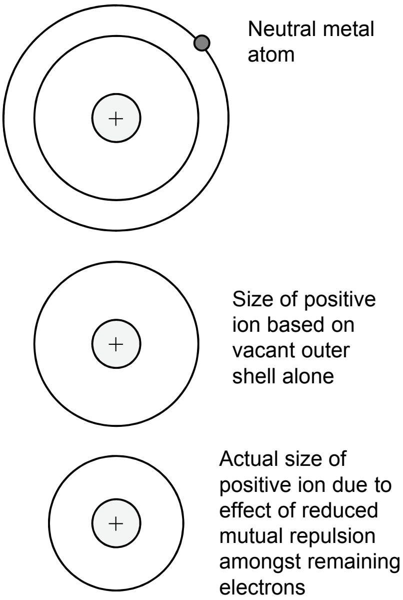
Neutral metal atom
+
Size of positive ion based on vacant outer shell alone
+
Actual size of positive ion due to effect of reduced mutual repulsion amongst remaining electrons
+

When a metal atom loses an electron to become a positively charged **ion**, a reduction in **atomic size** occurs. The electron that is lost is usually an outer-shell electron and so this electron shell becomes vacant and so the outermost electron shell is the next one closer in towards the nucleus. However, the reduction in size is not just due to an outer electron shell becoming vacant. When a positive ion is formed, the remaining electrons can bunch together a little more closely because of the reduced degree of mutual repulsion (since there are now fewer electrons surrounding the nucleus).

When a non-metal atom gains an electron to become a negative ion, this usually results in an outer-level electron shell accepting an additional electron – which would not ordinarily cause an appreciable increase in size, but the addition of this electron means that there is now an excess of negative charge in the outer shell and mutual repulsion causes these electrons to move apart and the atom “expands”.

In a sodium chloride ionic crystal, therefore, the positive sodium ions shrink and the negative chlorine ions expand such that the sodium ions are almost packed in between the spaces of the chlorine ions.

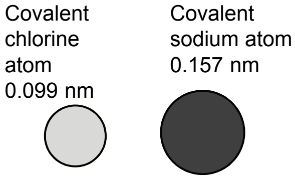
Covalent chlorine atom 0.099 nm
Covalent sodium atom 0.157 nm

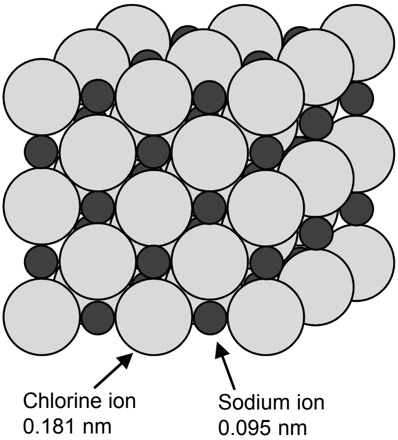
Chlorine ion 0.181 nm
Sodium ion 0.095 nm

3.10 Ionisation Energy

The **ionisation energy**, or **ionisation potential**, is the amount of energy needed to remove an electron from an isolated atom. The first ionisation energy is that needed to remove one outermost electron from an atom. Successive ionisation energies are those needed to remove further electrons from the atom.

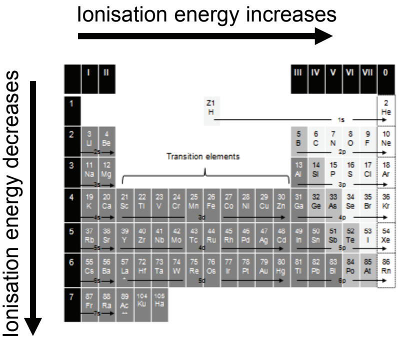

Successive ionisation energies always increase at each electron removal. As each electron is removed, the size of the atom decreases markedly and so the remaining electrons are closer to the positively charged nucleus, and so more energy is needed to remove them. Further, a particularly large increase in energy is needed to remove an electron from a filled energy level because a noble gas configuration of electrons is extremely stable.

Within a horizontal **period**, there is an increase in nuclear charge. As the **nuclear charge** increases (number of protons on the nucleus), the magnitude of the electric field within which the electrons find themselves also increases, and so generally speaking, the ionisation energy increases within a period. Also, the **atomic size** decreases from left to right and so the electrons for atoms towards the right are closer to the nucleus compared to those on the left, and so this also results in a general increase of **ionisation energy** from left to right.

Within a vertical **group**, there is generally a decrease of ionisation energy from top to bottom which is predicted on the basis of size of the atom alone. Any potential increase in ionisation energy on the basis of increased nuclear charge, going from top to bottom, is essentially cancelled out by the increased **screening effect** of the inner electrons on the outer electrons.

Generally speaking, elements with a low ionisation energy are on the left-hand side of the periodic table and those with a high ionisation energy are on the right. Noble gas configurations (eight valence electrons) can cause exceptions to these general rules.

3.11 Electronegativity

The **electronegativity** describes the relative ability for an atom, when it combines with another atom, to become “more negative” by more strongly capturing a shared electron or electron pair.

It is found that the more difficult it becomes for an atom to lose an electron, the easier it is for the atom to add an extra electron. That is, a higher **ionisation energy** also means a higher electronegativity.

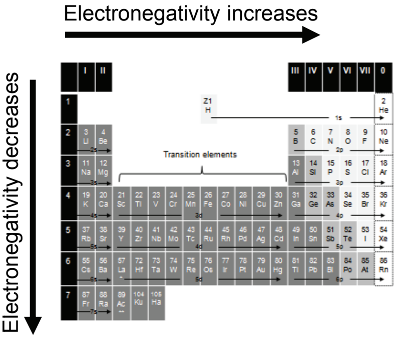

Some electronegativities within a **group** are:

Be	1.5
Mg	1.3
Ca	1.0
Sr	0.95
Ba	0.9
Ra	0.9

Some electronegativities within a **period** are:

Li	Be	B	C	N	O	F	Ne
1.0	1.5	2.0	2.5	3.0	3.5	4.0	-

The electronegativity of fluorine is so strong that in some circumstances, it is able to attract an electron from helium and so produce a compound that involves a noble gas.

The electronegativity of an atom influences the nature of the bonding between different types of atoms. Even when a bond is covalent, the shared electron spends more time nearer to the atom which has a higher electronegativity. The resulting molecule becomes polar.

4. Chemical Bonds

Summary

Lewis electron dot structures

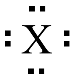

Van der Waals

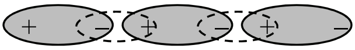

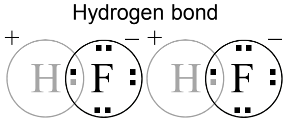

Oxidation numbers:

a) In free, uncombined, elements, the oxidation number of each atom is set to 0.

b) In compounds involving hydrogen, the oxidation number of hydrogen is 1+ .

c) In compounds involving oxygen, the oxidation number of O is usually 2–.

d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.

e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

Oxidation → Increase in Oxidation No. → Loss of Electrons

Reduction → Decrease in Oxidation No. → Gain of Electrons

4.1 Chemical Bond

Consider two hydrogen atoms that are initially a long way apart from each other. The two atoms are both electrically neutral, and have no ionic character. Why should they then be attracted to each other and form a stable H_2 molecule? Atoms tend to arrange themselves in the lowest possible energy state. In a **covalent bond** (like in H_2), the sharing of electrons provides the necessary reduction in total energy since the shared electrons, now having more space to move (over two atoms instead of one) have a lower **energy** than a single electron orbiting a single nucleus (see Section 1.5). This reduction in energy appears to us like a force of attraction acting between the atoms.

When two atoms form a **chemical bond**, they take up equilibrium positions according to the balance between long-range attractive and short-range repulsive forces $F(r)$ between them. At the **equilibrium position**, the potential energy $V(r)$ of the bond is a minimum.

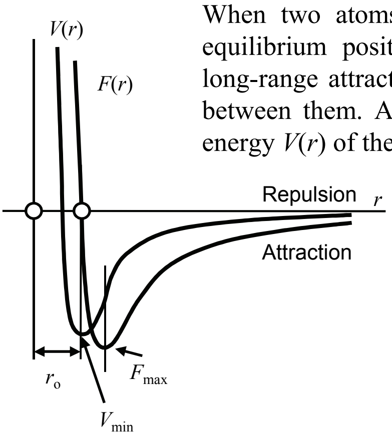

The potential energy function (V(r)) acting between two atoms can be very complex. Simple models are usually used, a very popular one being the **Lennard-Jones potential**:

$$V(r) = -\frac{A}{r^6} + \frac{B}{r^{12}}$$

In this potential, the negative term is the energy associated with the attractive forces and the positive term is the energy associated with the repulsive forces.

It is convenient to assign a potential energy of zero to widely spaced molecules or atoms so that when they approach, the potential energy becomes more negative. Since we generally assign a positive number to **work** done on a system (energy entering a system) and a negative number to work done by a system (energy leaving a system), work has to be done on the molecules or atoms to separate them. Thus, when the atoms are at the equilibrium position, energy is required to be done on the system to move one atom from the equilibrium distance to infinity against the force of attraction. This energy is called the **bond energy** (usually in kJ/mol).

To be consistent with the assignment of energy, we must therefore treat **attractive forces** as negative and repulsive forces as positive. The force F, with this sign convention, when multiplied by the distance of movement r, then results in an energy of the correct sign.

4.2 Lewis (Electron Dot) Formulae

Molecules and compounds form when bonds exist between atoms involving the transfer or sharing of valence electrons. The **Lewis symbol** for an atom is the element symbol surrounded by dots that represent the possible **valence electrons** for the atom.

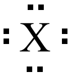

Lewis recognised that bonding or sharing of electrons occurred so that an element tended to acquire eight valence electrons (noble gas configuration): the **octet rule**. To represent a bond between two atoms, a **Lewis structure** is written so as to indicate the position of the valence electrons:

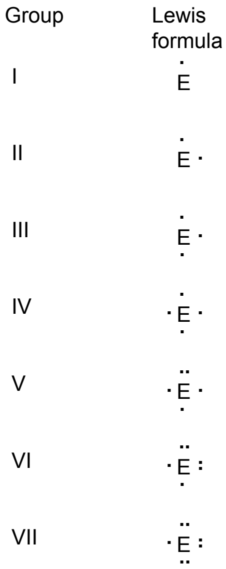

"E" denotes an element.

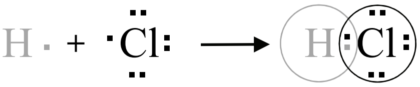

This **Lewis structure** shows how each element acquires a noble gas configuration when a chemical (in this case, ionic) bond is formed.

When determining a Lewis structure, the total number of dots that appear (the total number of **valence electrons**) is the sum of the number of valence electrons for each participating atom.

In some cases, it is not easily determined what the Lewis or **electron dot** formula for an atom should be. For example, in carbon, the electronic configuration is: $1s^2\ 2s^2\ 2p^2$ and we might be tempted to write the **electron dot** formula as:

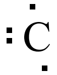

Although this arrangement can occur, when carbon usually combines with other atoms, the 2*s* orbitals combine with the 2*p* orbitals and the electrons are spread out over the resulting combined "*sp*" energy level giving four valence electrons. Carbon is **tetravalent**.

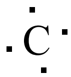

4.3 Multiple Bonds

Covalent bonds between atoms often involve sharing of more than one electron at a time. For example, carbon has a bonding capacity, or valence, of four, and so the carbon compound **ethane** is a stable molecule and is fairly unreactive.

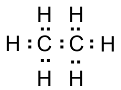

$$\mathrm{H}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\underset{..}{\overset{..}{\mathrm{C}}}}}:\underset{\mathrm{H}}{\overset{\mathrm{H}}{\underset{..}{\overset{..}{\mathrm{C}}}}}:\mathrm{H}$$

Ethylene, with molecular formula C_2H_4 should therefore leave two lone valence electrons:

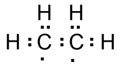

$$\mathrm{H}:\overset{\mathrm{H}}{\underset{\cdot}{\overset{..}{\mathrm{C}}}}:\overset{\mathrm{H}}{\underset{\cdot}{\overset{..}{\mathrm{C}}}}:\mathrm{H}$$

We would therefore expect ethylene to be a very reactive compound, in much the same way as Group I elements. However, this is not the case. Instead, the two carbon atoms arrange for the orbits of the two lone electrons to overlap and be shared. That is, two pairs of electrons are shared between the carbon atoms to form a **double bond**.

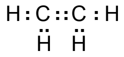

$$\mathrm{H}:\underset{\mathrm{H}}{\underset{..}{\mathrm{C}}}::\underset{\mathrm{H}}{\underset{..}{\mathrm{C}}}:\mathrm{H}$$

Triple bonds are also possible whereby three pairs of electrons are shared. The double and triple bonds offer even a greater range of movement of valence electrons (since more electrons are shared and so more electrons have more overall room to move) compared to a single bond. The strength of the bond is higher, and the C nuclei are closer together (shorter bond length).

Although the bond energy associated with a double bond is higher, the double bond results in a more reactive compound than a comparable single-bonded structure. This is because the double bond offers more opportunities for other more reactive elements than carbon to attach to these valence electrons – as if other elements can open up the bond and find two potential bonding sites rather than one. It is energetically more favourable to have two single bonds than one double bond.

Molecule	Structural formula	C – C length (Å)	C – C energy (kJ/mol)
C_2H_6 Ethane	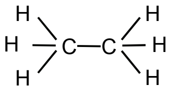	1.54	334
C_2H_4 Ethylene	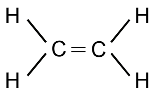	1.33	606
C_2H_2 Acetylene	$H—C\equiv C—H$	1.20	828

4.4 Lewis Single-Bonded Structures

Some examples of single-bonded structures are:

Molecular formula	Total No. valence electrons	Lewis formula	N	Shape	Note
F_2 Fluorine gas	14	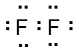	-	F — F	Linear
CH_4 Methane	8	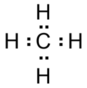	4	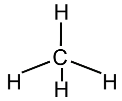	Tetrahedral
$SiCl_4$	32	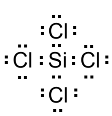	4	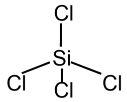	Tetrahedral
NH_3 Ammonia	8	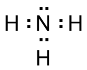	4	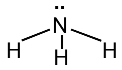	Pyramidal
H_2O Water	8	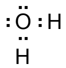	4	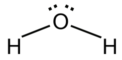	Bent
NH_4^+	8	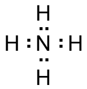	4	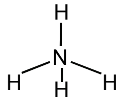	Tetrahedral
C_2H_6 Ethane	14	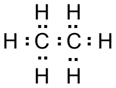	4,4	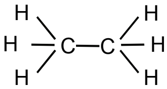	2 tetrahedral
H_2O_2	14	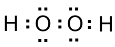	4,4	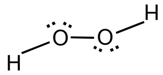	

Note that **lone pairs**, being closer to the central atom, have a greater repulsive effect than bonded pairs and so influence the shape of the molecule. Note that each atom is surrounded by eight electrons (except for H which requires two electrons for noble gas configuration).

4.5 Lewis Multiple-Bonded Structures

Examples of multiple-bonded structures:

Molecular formula	No. valence electrons	Lewis formula	N	Shape	Note
N_2 Nitrogen gas	10	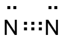	-	$N \equiv N$	Linear
C_2H_4 Ethylene	12	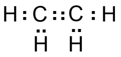	3,3	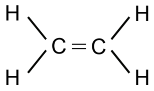	2 Tetrahedral
C_2H_2 Acetylene	10	H:C:::C:H	2,2	$H - C \equiv C - H$	Linear
CO_2 Carbon dioxide	16	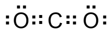	2	$O = C = O$	Linear
HCN Hydrogen cyanide	10	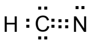	2	$H - C \equiv N$	Linear
$COCl_2$ Phosgene	24	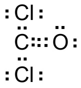	3	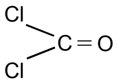	Planar
H_2CO Formaldehyde	12	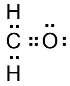	3	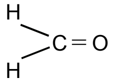	Planar

Note that each atom is surrounded by eight electrons (except for H which requires two electrons for noble gas configuration).

4.6 Lewis Exceptions to the Octet Rule

Examples of structures that do not obey the **octet rule** include:

(1) Atoms having more than eight electrons involved in bonding:

Molecular formula	No. valence electrons	Lewis formula	N	Shape	Note
PCl_5	40	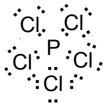	5	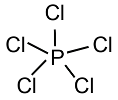	More than an octet
SF_4	43	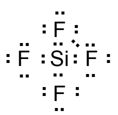	5	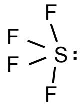	More than an octet with lone pair

When there are more than eight electrons involved, such as in PCl_5, the bonding usually involves electrons from the *d* sub-shell.

(2) Atoms having less than eight electrons involved in bonding:

BeH_2	4	H:Be:H	2	H — Be — H	Less than an octet
NO	11	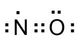		N = O	Less than an octet

When more than one valid structure can be written and the true structure cannot be written, the actual or true structure is called a **resonance hybrid** of the different structures. An example is NO. The unpaired electron (as shown above) can be written as attached to either the N or the O atom.

Example of **resonance structure**: CO_3^{2-}

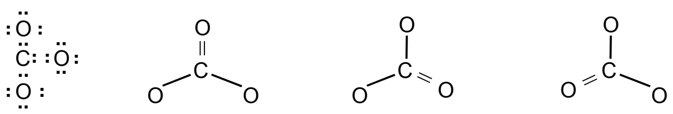

4.7 Oxidation Number

The **oxidation number**, or **oxidation state**, is a quantity that describes the number of electrons that appear to be gained or lost by an atom when a chemical bond is formed. That is, it is essentially the number of electrons that participate in the formation of the chemical bond. If we imagine all bonds to involve a transfer of electrons (even if in reality the bond may be covalent), the oxidation number is essentially the electronic charge that an atom appears to have after the bond has formed. The oxidation number is determined by a set of agreed rules. It is not strictly a physical quantity.

In a molecule or a compound, electrons that are shared between atoms are counted as belonging to the more **electronegative** atom. Electrons that are shared between two identical atoms are equally divided between the two atoms.

(a) In free, uncombined elements the oxidation number of each atom is set to zero. For example, the oxidation number of H in H_2 is zero.

(b) In compounds involving hydrogen, the oxidation number of hydrogen is 1+ except in the case of metal hydrides (where the hydrogen is bonded to an atom which is less electronegative) where it is 1−.

(c) In compounds involving oxygen, the oxidation number of O is usually 2−. The exceptions are the oxygen atom in peroxides and when oxygen is bonded with fluorine.

(d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.

(e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

Consider the compound H_2O. The oxygen atom, being more **electronegative** than the hydrogen, is seen to gain two electrons and so is given an oxidation number of 2− (since it seems to have acquired a net "negative" charge). Each hydrogen atom is seen to lose an electron and so the oxidation number of hydrogen in this compound is assigned as 1+.

When the oxidation number of an atom has increased, the atom is said to have been **oxidised** (as in the hydrogen above). This usually involves the loss of electrons. When the oxidation number of an atom has decreased, the atom is said to have been **reduced**, that is, its oxidation number has decreased. This usually involves a gain of electrons (as in the oxygen above).

The oxidation number is useful when naming some compounds. Prefixes like *hypo* and *per*, and suffixes like *ous*, *ic*, *ite*, are assigned according to the oxidation number of the central metal or non-metal atom in a compound.

4.8 Oxidation Number Examples

Example	Element	Oxidation state	Rule
S	S	0	a
H_2S	S	-2	b, e
Fe^{2+}	Fe	+2	d
Fe^{3+}	Fe	+3	d
MnO_4^-	Mn	+7	c, e
Mn^{2+}	Mn	+2	d
$Cr_2O_7^{2-}$	Cr	+6	c, e
Cr^{3+}	Cr	+3	d
NO_3^-	N	+5	c, e
NO	N	+2	c, e
NO_2	N	+4	c, e
SO_4^{2-}	S	+6	c, e
CO_2	C	+4	c, e
H_2O_2	O	-1	special case

When a substance is **oxidised**, it can be done so by either adding oxygen or removing hydrogen. Similarly, when a substance is **reduced**, it can be done so by either removing oxygen or adding hydrogen. For example, consider the production of methane CH_4 from carbon and hydrogen:

$$\mathrm{C + 2H_2 \rightarrow CH_4} \qquad \text{Oxidation No.}$$

$$\mathrm{2H_2 \rightarrow 4H^+ + 4e^-} \qquad \text{0 to 4+}$$

$$\mathrm{C + 4e^- \rightarrow C^{4-}} \qquad \text{0 to 4−}$$

From the point of view of the hydrogen, it has lost electrons and so is oxidised. From the point of view of the carbon, it has gained electrons and so is reduced. Of course in this compound, CH_4, the electrons are shared between the C and H atoms as covalent bonds, but from the perspective of oxidation numbers, the electrons are effectively transferred from H to C.

In some biochemical reactions, electrons (in the company of protons H^+) are transferred during oxidation and reduction reactions as a means of passing energy from one molecule to another. The addition of an electron via the addition of hydrogen stores energy in the compound being reduced. The energy is released when the compound is oxidised.

When CH_4 is combined with oxygen (e.g. when it is burnt), the C is oxidised to CO_2 and water: $\mathrm{CH_4 + O_2 \rightarrow CO_2 + 2H_2O}$

4.9 Polar Bonds

Although bonds between atoms may be classified as being **covalent** or **ionic**, in many different substances, there is a gradation from one type to the other. Unequal sharing of electrons in a covalent bond gives the bond an ionic character. Further, such a covalent molecule may acquire regions of net positive and negative charge and become **polarised**.

Separation of charges within a covalent molecule usually results in electrons spending more time on the atom which has the greatest **electronegativity**.

: C ::: O :

In CO, the shared electrons spend more time near O than C and so the centre of positive charge is offset from the centre of negative charge.

For example, in the molecule carbon monoxide, there is a difference in electronegativity between carbon (2.5) and oxygen (3.5) and so the shared electrons spend more time near the oxygen atom than the carbon atom. This results in a net separation of centres of **electric charge** within the molecule. The molecule is thus polarised and is a **dipole**. The molecule as a whole remains electrically neutral, but the distribution of charge within the molecule shifts so that the geometrical centre of positive charge is different than the centre of negative charge.

Although it is straightforward to determine if a diatomic molecule is polar or non-polar, by a consideration of the electronegativities of their atoms, the situation is more complicated for molecules with more than two atoms. For example, in CO_2, the individual bonds between the carbon and the oxygen molecules may be polar, but the molecule as a whole is non-polar because of the symmetry of the shape of the molecule.

:Ö::C::Ö: O = C = O μ = 0 debye

When CO_2 is placed in an electric field there is no net **dipole moment**.

Experiments show that **water** has a permanent dipole moment due to the presence of the electron pairs on the oxygen.

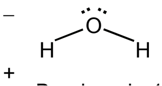

μ = 1.84 debye

Bond angle 104.5°

Polarization is responsible for **physical bonds** forming between molecules which in turn result in molecules forming liquids and solids.

4.10 Hybrid Orbitals

In a covalent bond, when valence electrons are shared between atoms, they do so by overlapping orbitals. Consider an oxygen atom: $1s^2$, $2s^2$, $2p^4$

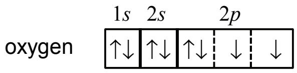

The shapes of molecules are difficult to predict on the basis of the orbitals of isolated atoms because the nature of these orbitals change when there are more than one electron in an atom and when atoms come together to form molecules.

Hybrid	Shape
sp	linear
sp^2	plane triangular
sp^3	tetrahedral

When a molecule of water forms, the 2*s* and the three 2*p* orbitals in the oxygen combine to form a new molecular orbital called the sp^3 **hybrid**. The superscript 3 indicates that three of the original *p* orbitals are involved in the hybrid orbital. It is this hybrid orbital $2sp^3$ that combines with the *s* orbitals of the H atoms.

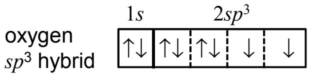

Each of the $2sp^3$ hybrid orbitals is at the same energy level.

When hybridization occurs, the rearrangement of electrons often results in there being more unpaired valence electrons than there were originally. For example, in carbon compounds, we have:

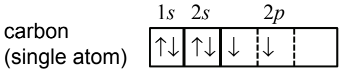

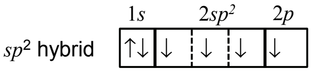

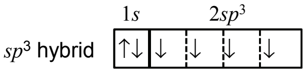

In CH_4, it is the hybridization of the C atomic energy levels into four identical $2sp^3$ orbitals that results in a regular tetrahedral geometry with bond angle 109.5°. In PCl_5 (more than eight electrons bonded), the hybrid orbitals are sp^3d.

In the double C=C bond of C_2H_4 one of the bonds is the end-to-end overlap of one of the sp^2 hybrid orbitals (this is called the σ **sigma bond**) along the line of the two C atoms (the other two sp^2 orbitals overlap with the *s* orbitals of the H atoms), while the other C bond is formed by the side-by-side overlap of the unhybridised (dumbbell shaped) 2*p* orbitals parallel to the plane of the C and H atoms (this is called the π **pi bond**).

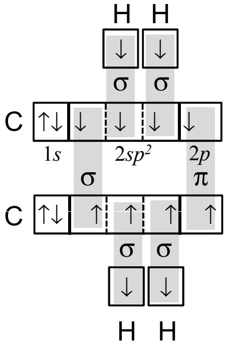

4.11 Polarisation

When an electric field *E* is applied to a molecule in a non-polar insulating material, the centre of charge of electrons moves left, the centre of charge of protons move right and the two centres of charge are then separated by distance *d*. The atom or molecule becomes **polarised** in the presence of *E*.

Polarisation in a material may occur due to several mechanisms, all of which may occur to some extent depending on the atom. Most often, polarisation occurs when an external electric field *E* is applied.

1. **Electronic polarisation**

   Small induced dipole moment arising from difference in the net centres of nucleus and electrons in an atom.

   $\varepsilon_r \approx 2-4$

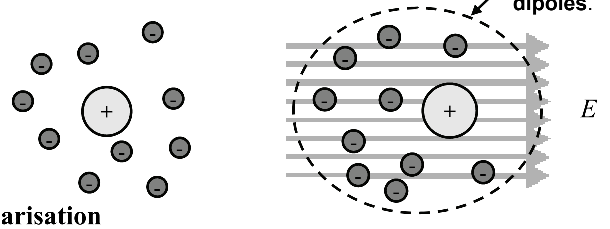

The entities with circles around them are **dipoles**.

2. **Ionic polarisation**

   Dipole moment created by shift of positive ions with respect to negative ions in unit cell.

   $\varepsilon_r \approx 6-10$

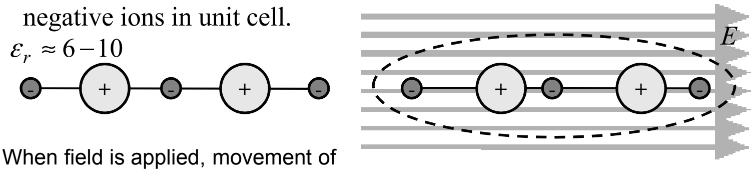

When field is applied, movement of ions produces a net dipole moment in a unit cell and hence a net polarisation.

3. **Dipolar polarisation**

   Net dipole moment created by alignment of molecule with external field due to presence of internal permanent dipoles from geometrical structure of molecule.

   $\varepsilon_r \approx 20-100$

Water molecule has a permanent dipole moment.

CO$_2$ molecule has no permanent dipole moment. No dipolar polarisation can occur.

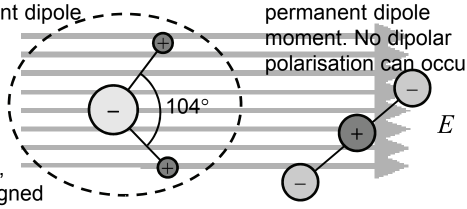

Note: Even though there is a permanent dipole moment, in the absence of an electric field, these dipole moments are not aligned (due to thermal agitation) and so the net polarisation over all molecules in the material is zero. When a field *E* is applied, molecular dipoles tend to align themselves with the field and there exists a net polarisation.

4.12 van der Waals Forces

The arrangement of electrons around atoms is not constant in time. Electrons move within their energy shells or levels with kinetic energy. An instantaneous movement of an electron within an atom may result in the atom becoming **polarised** for a short period of time. During this time, the polarisation in one atom may induce polarisation in a neighbour atom, resulting in a very brief electrostatic attraction between them. These instantaneous forces of attraction and repulsion are termed **London forces** after the scientist who proposed the explanation for them and are found in all covalent materials.

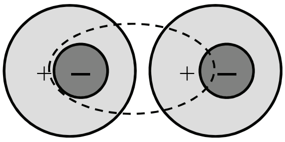

Instantaneous polarisation of one molecule results in induced dipole in neighbour molecule and a net instantaneous attraction between the two molecules.

In materials containing molecules that have permanent dipole moments, the molecules tend to align themselves (against the tendency of thermal agitation to keep them in random orientation). The electrostatic attraction between the polarised ends of the molecules draws them together. **Dipole-dipole interaction** may persist over many molecules and appears to us like a force holding the molecules together.

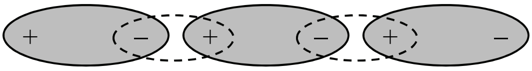

When considered over millions of atoms, the London and permanent dipole forces can be substantial and are called physical, as distinct from chemical, bonds between molecules. The intermolecular forces of this type are called **van der Waals forces**. In most cases, it is the van der Waals forces that bind atoms and molecules into liquids.

Intermolecular forces on this scale are responsible for deviations from ideal behaviour in liquids and gases. Van der Waals forces operate over a small scale in comparison to distances between molecules in a gas. It is only when gas molecules become closer together (such as under increased pressure) that these forces become strong enough to cause the molecules to then clump together and **condense** into a **liquid**. The forces become less effective when the temperature increases due to the tendency for an increase in random order associated with thermal agitation of molecules.

4.13 Hydrogen Bond

Hydrogen has one valence electron and so can readily form a single covalent bond. However, there are some compounds involving hydrogen in which the hydrogen atom can be said to be bonded to two atoms at the same time. Such compounds involve hydrogen bonding with small electronegative atoms.

Consider the **hydrogen fluoride** molecule:

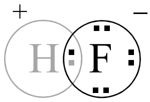

Because fluorine has a large **electronegativity**, the shared electron spends much of its time near the fluorine atom. The molecule, as shown above, is **polar**. Now, if there are other HF molecules nearby, then the positively charged H side of the molecule will be drawn into contact with the negatively charged F end of the other molecule. Two molecules thus join up so:

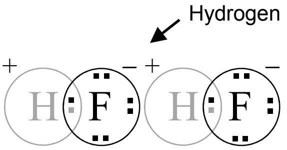

It looks to us like the H atom in the middle is holding the two outer F atoms in place like a tie bar. This electrostatic attraction is called a **hydrogen bond**. The hydrogen bond can only form between hydrogen and small electronegative atoms because the H "ion" is very small and only has space for one other atom to be located near enough to its polar end to form an **electrostatic** bond.

Hydrogen bonds also form between **water molecules**. Hydrogen bonding between water molecules gives rise to a characteristic crystal structure in **ice**, which in turn is reflected in the shape of **snowflakes**.

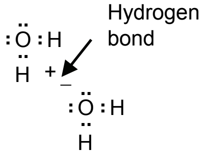

Hydrogen bonding is very important in biological processes. The bonds, having an energy of about 20 kJ, are stronger than those due to van der Waals forces (about 0.2 kJ) but are significantly weaker than those in chemical bonds (about 200 kJ). Certain biological processes rely on these bonds being of just the right strength for events to occur with ease (such as in the forming and breaking of the two halves of the molecular spiral in **DNA** – these hydrogen bonds involve H, O and N atoms).

5. States of Matter

Summary

Gases:

$$\frac{p_1V_1}{T_1} = \frac{p_2V_2}{T_2}$$ Combined gas law

$$p_1V_1 = p_2V_2$$ Boyle's law

$$\frac{V_1}{T_1} = \frac{V_2}{T_2}$$ Charles' law

Volume

Absolute temperature

$$pV = nRT$$

Absolute pressure

no. moles

Universal gas constant 8.3145 J/mol/K

Liquids: $$\text{Molarity} = \frac{\text{No. moles of solute}}{\text{Volume of solution in litres}}$$

Solids:

- Molecular solids - in which the binding forces between discrete atoms or molecules are due to van der Waals interactions between instantaneous electric dipoles.
- Ionic solids - in which the binding energy is due to Coulomb electrostatic forces between positively and negatively charged ions in the crystal.
- Covalent solids - in which the binding energy is due to shared valence electrons between atoms in the solid.
- Metallic solids - in which valence electrons are effectively shared amongst all the atoms in the solid.
- Amorphous solids – no long-range structure.

5.1 Changes of State

When a solid is heated, atoms/molecules within the solid absorb heat energy and this is stored as **internal energy** of vibration. As the temperature increases, the magnitude or amplitude of the vibration increases until the bonds holding the molecules together become temporarily ruptured and molecules flow over one another. This is called **melting**.

At the **melting point**, further heating results in no increase in temperature. The energy input in this case goes to breakage of bonds between the molecules and the material changes phase from solid to liquid.

Upon further heating, molecules of liquid increase their translational energy until they acquire sufficient kinetic energy to escape the liquid into the gaseous phase.

The presence of impurities lowers the melting point. A measurement of the melting point can be used as an index of the purity of a particular substance.

When a liquid cools, molecules begin to lose their internal energy of motion until eventually, intermolecular forces are able to hold them in place as a solid. The temperature at which this occurs is called the **freezing point** – which is usually equal to the melting point.

If the molecules in the cooling solid do not immediately lock into the pattern of the solid, then the temperature may fall below the official freezing point and the material remains in the liquid phase. This is called **supercooling**. Introduction of a seed crystal may cause the liquid to avoid the supercooling condition.

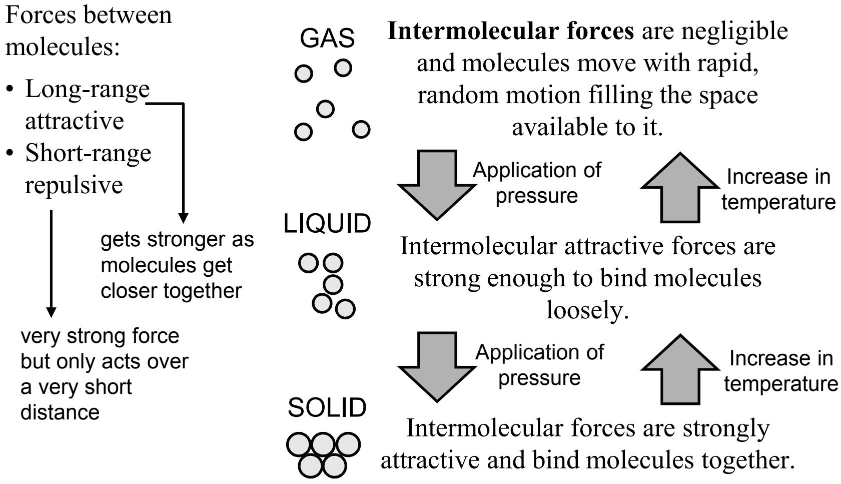

Forces between molecules:

- Long-range attractive
- Short-range repulsive

gets stronger as molecules get closer together
very strong force but only acts over a very short distance
GAS
LIQUID
SOLID
Application of pressure
Increase in temperature
Application of pressure
Increase in temperature

Intermolecular forces are negligible and molecules move with rapid, random motion filling the space available to it.

Intermolecular attractive forces are strong enough to bind molecules loosely.

Intermolecular forces are strongly attractive and bind molecules together.

5.2 Changes of State of Liquids/Gases

Closed system

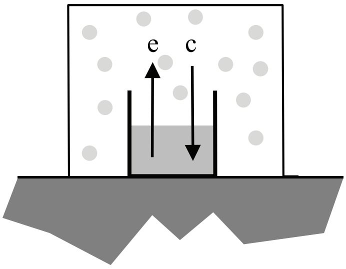

A dynamic equilibrium is set up where the rate of evaporation = the rate of condensation.

The **equilibrium vapour pressure** (**EVP**) or **saturated vapour pressure** is the pressure exerted by the vapour when in equilibrium with its liquid. The magnitude of the EVP depends on the nature of the liquid and the temperature.

- Nature of liquid. The stronger the **van der Waals forces** between molecules, the lower is their ability to escape the liquid and the lower the EVP.
- Temperature. Increase in temperature increases kinetic energy of molecules in the liquid and hence their ability to escape from the surface and hence the higher the EVP.

Open system

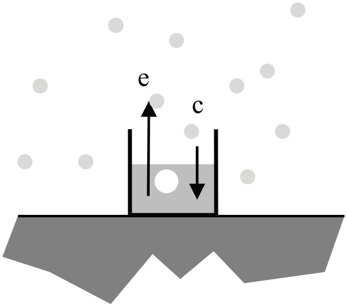

A liquid is said to boil at a temperature called the **boiling point**, at which the EVP of the liquid becomes equal to the prevailing atmospheric pressure. Vapour bubbles form in the liquid and rise to the surface.

The **normal boiling point** is the boiling point at 1 atm of pressure (101.3 kPa).

Increasing the prevailing atmospheric pressure increases the boiling point.

5.3 Phases of Matter

In a ***p–V* diagram**, the temperature is kept constant, volume decreased and pressure recorded.

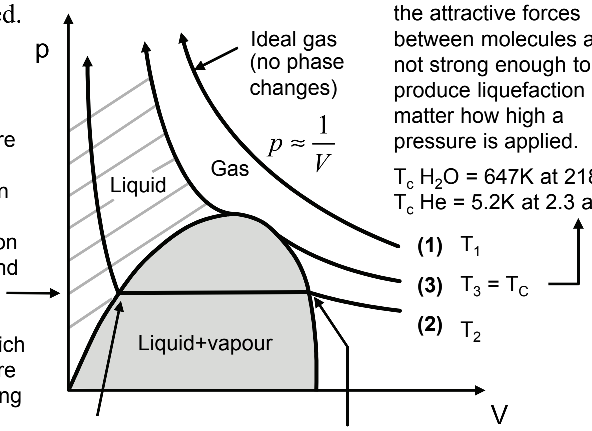
p
V
Liquid
Gas
Liquid+vapour
Ideal gas (no phase changes)
$p \approx \frac{1}{V}$
(1) T_1
(3) $T_3 = T_C$
(2) T_2

Critical Temperature T_c
There is, for each gas, a temperature above which the attractive forces between molecules are not strong enough to produce liquefaction no matter how high a pressure is applied.

T_c H_2O = 647K at 218 atm
T_c He = 5.2K at 2.3 atm

Vapour pressure
The partial pressure exerted by the vapour when it is in equilibrium with its liquid. It depends on the temperature and nature of the substance. The temperature at which the vapour pressure equals the prevailing atmospheric pressure is called the **boiling point**.

All gas condensed into liquid, attempts to further reduce volume produce large increase in pressure as liquid is compressed.

Gas starts condensing into liquid, no change in pressure as volume decreases.

In a **phase diagram**, we keep the volume *V* constant and plot pressure versus temperature.

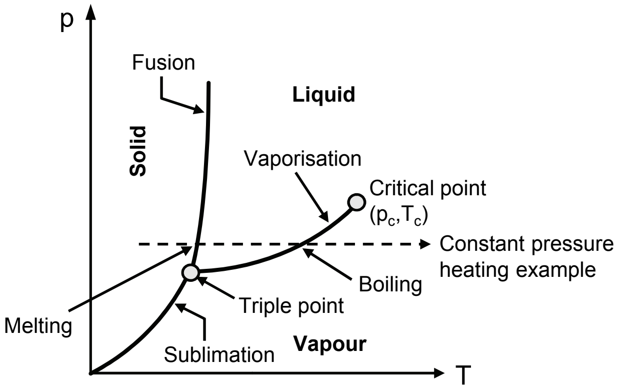
p
T
Fusion
Liquid
Solid
Vaporisation
Critical point (p_c,T_c)
Constant pressure heating example
Boiling
Triple point
Melting
Sublimation
Vapour

At each point (*p*,*T*) only a single phase can exist except on the lines where there is phase equilibrium. At the **triple point**, solid, liquid and vapour exist together in **equilibrium**.

5.4 Gases

The properties of a perfect or ideal gas are the most convenient to study. The properties of an ideal gas are:

- The molecules of the gas occupy a very small volume compared to the volume of the container.
- The molecules are very distant from one another and only interact elastically during collisions.

Real gases often behave like ideal gases at conditions in which no phase changes occur.

Macroscopic properties of a gas:

- Pressure
- Temperature
- Volume
- Mass

These quantities specify the **state** of a gas

Consider a mass (m) of gas:

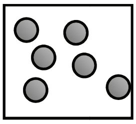

If p, V and T all vary, then:

$$\frac{p_1V_1}{T_1} = \frac{p_2V_2}{T_2} \quad \textbf{Combined gas law}$$

If temperature T is a constant:

$$p_1V_1 = p_2V_2 \quad \textbf{Boyle's law}$$

Note: These laws cannot be applied when the mass of gas changes during the process. Pressures and temperatures are absolute.

If pressure p is a constant:

$$\frac{V_1}{T_1} = \frac{V_2}{T_2} \quad \textbf{Charles' law}$$

Let us express the mass of a gas indirectly by specifying the number of moles. Experiment shows that Boyle's law and Charles' law lead to:

$$\underset{\text{Absolute pressure}}{p}\overset{\text{Volume}}{V} = \underset{\text{No. moles}}{n}\underset{\text{Universal gas constant 8.3145 J/mol/K}}{R}\overset{\text{Absolute temperature}}{T}$$

This equation links all the macroscopic quantities needed to describe the (steady) state of an *ideal* gas and is thus called an **equation of state**.

$$\text{No. moles} \longrightarrow n = \frac{\overset{\text{Mass in kg}}{m}}{M_m} \longleftarrow \text{Molar mass}$$

By using moles, we get the ideal gas equation with the universal gas constant R (units J/mol/K). Otherwise, value of R depends on the nature of the gas (i.e. no longer universal) and has units J/kg/K.

Example: Calculate the volume occupied by one mole of an ideal gas at 273 K at atmospheric pressure.

$$pV = nRT$$

$$101.3(V) = 1(8.314)(273)$$

$$V = 22.406\,\text{L}$$

5.5 Solutions

Homogenous mixtures of two or more gases, liquids, or solids are called **solutions**.

- Gases usually mix in any proportion. Air is a solution of predominantly nitrogen, oxygen and carbon dioxide.
- Liquid solutions can be made by dissolving a solid, liquid or gas (the **solute**) in a liquid (the **solvent**). When the solvent is water, the solution is called an **aqueous solution**.
- Solid solutions occur where atoms of one solid are randomly dispersed throughout the other solid. **Alloys** are common examples of solid solutions although some alloys are **compounds** (combine in definite proportions). Some alloys are also heterogeneous **mixtures**.

Intermediate between heterogeneous mixtures and homogenous solutions is the case where the dispersed solid particles are not broken up into molecular units and yet are not in collections large enough to be called a separate phase. In this case, the dispersion is called a **colloid**.

The strength, or **concentration**, of a solution can be described as:

- The ratio of the number of moles of one of the components (the solute or the solvent) relative to the total number of moles present is called the **mole fraction** of the component.
- The number of moles of solute per litre of final *solution* is called the **molarity** and given the symbol M (this is the most common method but depends on volume, which in turn is temperature dependent).
- The number of moles per kilogram of *solvent* is called the **molality** and is given the symbol m (has the advantage of not being dependent on temperature).
- The percent solute by mass or volume of the final solution.
- % by weight (**w/w**) is the number of grams of solute per 100 g of solution.
- % by volume (**w/v**) is the number of grams of solute per 100 mL of solution.

The limiting concentration of the solution to which a solute can be dissolved in a solvent is called the **solubility**. At this concentration, the solution is said to be **saturated**. Addition of excess solute results in no increase in concentration and a portion of the solute remains undissolved.

5.6 Aqueous Solutions

Many chemical reactions involve chemicals that exist as water, or **aqueous solutions**.

$$\text{Molarity} = \frac{\text{No. moles of solute}}{\text{Volume of solution in litres}}$$

Examples:

(1) Calculate the mass of solute needed to make up 1 litre solution of 0.5M sodium hydroxide:

$$\begin{aligned} \text{1 mole NaOH} &= 23+16+1 \\ &= 40\text{ g} \\ \text{mass required for 1 L solution at 0.5M} &= 40(0.5) \\ &= 20\text{ g} \end{aligned}$$

(2) Calculate the volume of 10M acid required to make up 250 mL of 0.4M solution of HCl:

$$\begin{aligned} 1\text{ L} &= 0.4\text{M} \qquad\qquad \text{Cl} = 35.5 \\ &= (0.4)(36.5)\text{ g} \qquad \text{H} = 1 \\ &= 14.6\text{ g} \\ 0.25\text{ L} &= 3.65\text{ g HCl required} \\ 10\text{M} &= 365\text{ g/L HCl} \\ 3.65\text{ g HCl} &= 10\text{ mL} \end{aligned}$$

It is often observed that for dilute aqueous solutions, the addition of the solute lowers the **freezing point** and raises the **boiling point** (reduction in vapour pressure) compared to pure water.

When an aqueous solution evaporates, a **hydrated ionic compound** may be formed. The hydrated compound contains water of crystallization:

$$\text{Co}^{2+}\text{(aq)} + 2\text{Cl}^-\text{(aq)} \xrightarrow{\text{evap}} \text{CoCl}_2 + 2\text{H}_2\text{O(s)}$$

5.7 Solubility of Solids

When a solute dissolves in a solvent, three distinct processes can be identified:

- Solute-solute interactions – whereby the solute atoms or molecules are separated from each other (such as the dissociation of Na^+ and Cl^- into ions). This is an **endothermic** process (heat is required to break the ionic bonds).
- Solvent-solvent interactions – whereby the solvent atoms or molecules may be separated from each other (such as when H_2O molecules are separated from one another to accommodate the incoming solute molecules). This is an **endothermic** process.
- Solute-solvent interactions – whereby processes occur between the solute and the solvent. These are usually **exothermic** (reduction in total energy).

Some ionic compounds are soluble in water, but insoluble in non-polar solvents (like benzene and carbon tetrachloride). When water is the solvent, a **hydration** process can occur which serves to insulate the positive and negative ions from each other and prevent them reforming into a solid. This is a solute-solvent interaction.

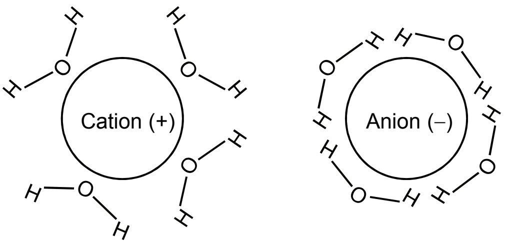

Hydration involves the formation of bonds and so is exothermic (heat is released as the total energy is lowered) and so the solute-solvent interactions dominate the solute-solute interactions. Non-polar solvents cannot hydrate the ions.

Some non-polar molecular solids are soluble in non-polar solvents but are insoluble in water (a polar solvent). Non-polar solute molecules are not hydrated by water, and so solute-solvent interactions are not significant here. If more energy is required to break the hydrogen bonds linking water molecules together than that required to break the bonds holding the components of the solute together, then the solute components remain insoluble. When placed in a non-polar solvent, solvent-solvent interactions may occur and the solvent molecules separate easily to accommodate solute molecules and the solute dissolves.

5.8 Solubility Equilibrium

If a solute is added to the solvent in increasing quantity, a point is reached where the atomic or molecular species being dissolved can no longer be accommodated in the solute. As more solute is added, the reverse process begins to occur so that some of the previously dissolved solute returns to the solute phase. This is called **precipitation**. This is most vividly illustrated when a solid solute is added to a liquid solvent. When the rate of precipitation becomes equal to the rate of **dissolution**, dynamic equilibrium is achieved and the solution is said to be **saturated**.

The **concentration** of a solute in a given solvent at **saturation** is called the **solubility** of the solute in the solvent.

Solubility depends on **temperature**. When a solute is dissolved in a solvent and heat is generated, the solubility generally decreases with increasing temperature. If heat is absorbed, the solubility generally increases with increasing temperature. Often, whether or not heat is absorbed or generated depends on the physical processes involved in any chemical reactions that may occur in the solution.

Solubility depends on **pressure**. For a non-reactive gas being dissolved in a liquid, the solubility *C* increases as the pressure *P* increases according to **Henry's law**: $C \propto P$

For example, dissolved CO_2 in a soft drink readily precipitates (observed as bubbles) when the pressure above the liquid is released (such as when the bottle is opened). An easy way to remember Henry's law is to think of it as the soft drink law.

For example, air dissolves in water to some extent and when it does so, heat is produced (since the gas "condenses" into a liquid form). The solubility of air in water thus decreases as the temperature increases. When water is heated, the dissolved air can be seen to "precipitate" as bubbles before the boiling point of water is reached.

One of the most common solvents is water, but not all compounds are soluble in water.

Compound	General rule	Exceptions
Nitrates	Soluble	
Chlorides	Soluble	AgCl, $PbCl_2$, Hg_2Cl_2
Sulphates	Soluble	$PbSO_4$, $BaSO_4$
Carbonates	Insoluble	Na_2CO_3, K_2CO_3, $(NH_4)2CO_3$
Hydroxides	Insoluble	NaOH, KOH, $Ba(OH)_2$, (NH_4OH)
Oxides	Insoluble	Na_2O, K_2O, BaO

5.9 Electrolytes

Some molecular and ionic compounds dissolve in water to give a solution that is able to conduct electric current. The compounds, when dissolved in water, dissociate into **ions**. The ions, being electrically charged, are thus able to move under the influence of an electric field and the movement of these ions is therefore an electric current.

- **Strong electrolytes** dissociate completely into ions. Many ionic compounds are strong electrolytes. Molecular compounds that are strong electrolytes are usually acids.
- **Weak electrolytes** only partially dissociate into ions. Usually, molecular compounds that are weak electrolytes are either an acid or a base.
- **Non-electrolytes** do not dissociate into ions when placed in solution. Molecular compounds that are neither acids or bases are usually non-electrolytes.

Molecular compounds that are **strong electrolytes**:

	Molecular formula	Ions formed in solution	
Hydrochloric acid	HCl	H^+, Cl^-	Strong acid
Nitric acid	HNO_3	H^+, NO_3^-	Strong acid
Sulphuric acid	H_2SO_4	H^+, SO_4^{2-}	Strong acid
Perchloric acid	$HClO_4$	H^+, ClO_4^-	Strong acid

Molecular compounds that are **weak electrolytes**:

	Molecular formula	Ions formed in solution	
Acetic acid	CH_3COOH	H^+, $CH_3CO_2^-$	Weak acid
Benzoic acid	C_6H_5COOH	H^+, $C_6H_5CO_2^-$	Weak acid
Ammonia	NH_3	NH_4^+, OH^-	Weak base

Molecular compounds that are **non-electrolytes**:

	Molecular formula	Ions formed in solution
Ethyl alcohol	C_2H_5OH	
Sucrose	$C_{12}H_{22}O_{11}$	

5.10 Osmosis

Osmosis is a movement of molecules through a **semi-permeable membrane**. A **semi-permeable membrane** is a sheet of substance that can allow some molecules through it but not others. Semi-permeable membranes can be things like **cell walls** in living tissue that contain openings that allow small molecules to pass through but not large molecules.

The concept of **osmosis** depends on the phenomenon of **diffusion**. Diffusion occurs when there is a **concentration gradient**. The concentration gradient drives the diffusion process. It is a fundamental consequence of the **second law of thermodynamics**. Diffusion happens in a wide variety of physical processes (e.g. the formation of the barrier potential at the p-n junction of a **semiconductor**).

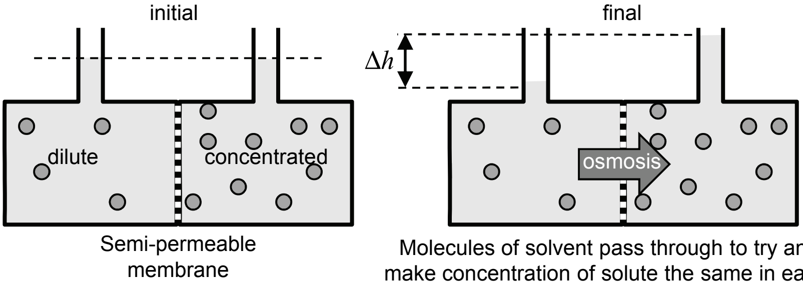

Molecules of solvent pass through to try and make concentration of solute the same in each.

Consider a **dilute solution** on one side of a semi-permeable membrane and a **concentrated solution** on the other side. In this example, the semi-permeable membrane allows molecules of **solvent** to pass through it, but not the **solute**. Be careful. In this case, the **concentration gradient** has to be looked at from the point of view of the solvent. In this case, molecules of solvent tend to pass from the dilute solution into the concentrated solution so as to attempt to achieve uniformity of concentration throughout. Molecules of solute from the concentrated solution would very much like to pass into the dilute side to achieve the same aim, but cannot pass through the membrane. The movement of solvent from one side to the other results in an increase in pressure on the concentrated side. The increase in pressure tends to inhibit the further flow of solvent and an equilibrium condition is reached where no more net movement occurs across the membrane. This pressure increase is called **osmotic pressure**.

For dilute solutions, the osmotic pressure π in atm (atmospheres) of non-electrolytes is given by $\pi = MRT$ where M is the molar concentration of the non-electrolyte, T is in K, and $R = 0.0821$ L atm/mol/K.

5.11 Solids

Generally speaking, there are two kinds of solids. **Crystalline solids** possess a repeating order of atoms, ions, complex ions, or molecules. The repeating pattern is called the **crystal lattice**. The repeating unit in the crystal lattice is called a **unit cell**. **Amorphous solids** have no repeating structure of atoms, ions or molecules.

Crystalline solids can be broadly classified as follows:

Molecular solids are those in which the binding forces between discrete molecules are due to instantaneous van der Waals interactions in addition to dipole-dipole attractions in polar molecules. Covalent bonds exist within the molecules between the atoms.

- Generally soft
- Gas, liquid or solid at room temperature
- Low melting point (about 20 ± 200°C)
- Not usually soluble in water
- Poor conductors of electricity
- Heat insulator

Examples: molecular crystals of noble gases, halogens. Compounds such as CO_2, CH_4, H_2O, C_6H_6, $C_{12}H_{22}O_{11}$ (sucrose) and most organic compounds.

Ionic solids are those in which the binding energy is due to Coulomb electrostatic forces between positively and negatively charged ions in the crystal. That is, ionic solids are composed of **cations** and **anions**.

- Generally hard
- Solid at room temperature
- High melting point (about 800°C)
- Usually soluble in water
- Poor conductors of electricity in solid form
- Heat insulator

Examples: cations from Group I (e.g. Na^+) or Group II (e.g. Mg^{2+}) with anions O^{2-}, S^{2-}, NO_3^-, CO_3^{2-}, ClO_4^-, SO_4^{2-}, CrO_4^{2-}, PO_4^{3-}, H^+, OH^-, MnO_4^-, CN^-.

Covalent solids are those in which the binding energy is due to shared valence electrons (**covalent bonds**) between atoms in the solid.

- Generally hard
- Solid at room temperature
- Very high melting point (about 1200°C)
- Insoluble in water
- Poor conductors of electricity
- Heat insulator

Examples : Si, C (**diamond**), Ge, SiO_2 (**quartz**), BN, SiC.

Metallic solids in which valence electrons are effectively shared amongst all the atoms in the solid.

Amorphous solids have no long-range regular repeating pattern of atoms or molecules. Examples are glass and most plastics. In these materials, there is an orderly structure in the neighbourhood of any one atom, but this is not regularly repeated throughout the material.

5.12 Crystalline Lattice Structures

For geometrical reasons, there are only fourteen types of lattices that satisfy symmetry operations such as translation, rotation, reflection and inversion. Each of these fourteen lattices is called a **Bravais lattice**. There are seven convenient **crystal systems** in the set of Bravais lattices: cubic, tetragonal, orthorhombic, trigonal, monoclinic, hexagonal and triclinic.

1. Cubic

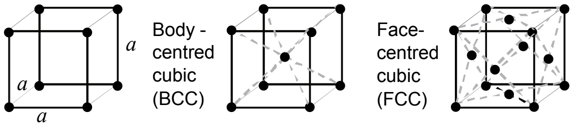
a
a
a
Body - centred cubic (BCC)
Face- centred cubic (FCC)

2. Tetragonal

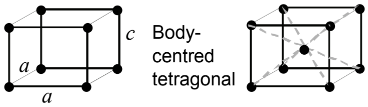
c
a
a
Body- centred tetragonal

3. Orthorhombic

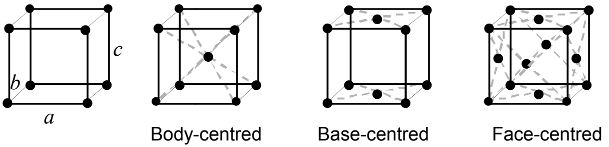
c
b
a
Body-centred orthorhombic
Base-centred orthorhombic
Face-centred orthorhombic

4. Triclinic

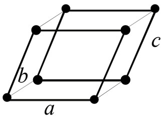

5. Monoclinic

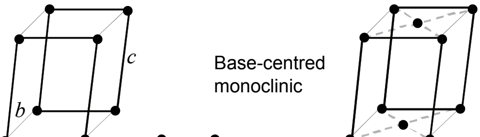
c
b
a
Base-centred monoclinic

6. Hexagonal

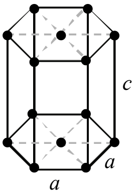

7. Trigonal

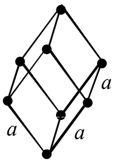

The particular geometry adopted by a solid depends on the relative sizes of the atoms or ions, their charge (for ions). Note: The lines shown in these images do not actually exist in a real material, but just serve to illustrate the overall shape of the lattice.

5.13 Metallic Solids

Metallic solids are those in which valence electrons are effectively shared amongst all the atoms in the solid. Atoms are held in place by the **metallic bond**.

- Can be either hard or soft
- Solid at room temperature (except Hg)
- Low to high melting point (room temperature and above)
- Insoluble in water
- Good conductors of electricity
- Heat conductor

Most metals crystallise into either body-centred cubic (BCC), hexagonal close-packed, and face-centred cubic (FCC). Common metals with **BCC** structure are Fe (**iron**), Cr (**chromium**), Mo (**molybdenum**), W (**tungsten**), V (**vanadium**). Common metals with **FCC** structure are Al (**aluminium**), Cu (**copper**), Au (**gold**), Pb (**lead**), Ni (**nickel**), Pt (**platinum**), Ag (**silver**). Common metals with close-packed hexagonal form are Cd (**cadmium**), Co (**cobalt**), Mg (**magnesium**), Ti (**titanium**), Zn (**zinc**). Metals with BCC structure have a higher yield strength than those with FCC structure.

At room temperature, iron has a BCC structure, but above 910°C, iron rearranges into an FCC structure (at which time it becomes non-magnetic). This change in structure takes place in the solid state and is called an **allotropic change**.

Solid solutions are very important in **metallurgy**, where they are called **alloys**. In a true solid solution, one component is randomly dispersed throughout the other component. Other types of alloys are molecular compounds (in fixed proportions) and heterogeneous mixtures. A particularly important alloy is that of iron and carbon, which makes **steel**.

Carbon Content:		
0.05–0.3%	Low-carbon steel	Mild steel, tough, ductile, easily forged and welded.
0.3–0.45%	Medium-carbon steel	Strong, hard, not so easily forged or welded.
0.45–0.75%	High-carbon steel	Very strong (high-strength steel) and hard. May be annealed for machining or heat treated for different degrees of hardness.
0.75–1.5%	Very high-carbon steel	

6. Chemical Thermodynamics

Summary

Internal energy: $\Delta U = nC_p(\Delta T)$

Work: $W = p\Delta V$

First Law: $Q_p = \Delta U + p\Delta V$

Enthalpy:
$$\begin{aligned}\Delta H &= \Delta U + p\Delta V \\ &= Q_p\end{aligned}$$

Entropy:
$$\Delta S = \frac{Q}{T}$$

$\Delta S_{\text{total}} > 0$ irreversible process
$\Delta S_{\text{total}} = 0$ reversible process
$\Delta S_{\text{total}} < 0$ impossible process

Gibbs energy: $\Delta G = \Delta H - T\Delta S$

Spontaneous reactions:

$$\Delta G_{\text{reaction}} = \Delta_{\text{f}} G^0{}_{\text{products}} - \Delta_{\text{f}} G^0{}_{\text{reactants}}$$

6.1 Molecular Energy

Molecules in a gas are capable of **independent motion**. Consider a diatomic gas molecule:

(a) the molecule itself can travel as a whole from one place to another

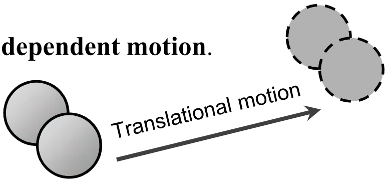

(b) the molecule can spin around on its axis

(c) the atoms within the molecule can vibrate backwards and forwards

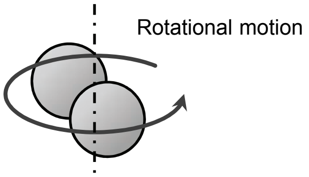

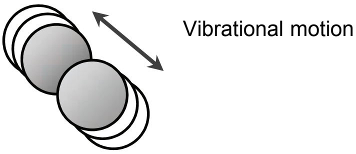

These **kinetic energy** terms represent the **internal energy** and have significance for the interpretation of the **specific heat** capacity of a substance. Changes in internal energy of a molecule manifest themselves physically as changes in **temperature** of the molecule.

The total energy of the molecule also contains some **potential energy** components. Firstly, there are **intermolecular forces** (such as **van der Waals forces**) that lead to the formation of liquids, and hydrogen bonds.

Secondly, there is the potential energy associated with the **chemical bond** between the atoms of the molecule, the **bond energy**.

These potential energy terms are associated with the *position* of the atoms within a molecule, and the position of molecules in relation to other molecules, as distinct from the *motion* of the atoms or molecules.

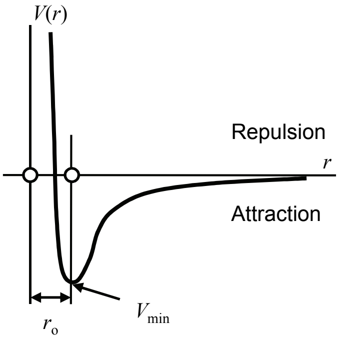

6.2 Specific Heat Capacity

The amount of energy (ΔQ) required to change the temperature of a mass of material is found to be dependent on the:

• Mass of the body (m)	kg
• Temperature increase (ΔT)	°C, K
• Nature of the material (c)	J/kg/K

$$\Delta Q = mc(\Delta T)$$

↓ Specific heat or **heat capacity**

The **specific heat** is the amount of heat required to change the temperature of 1 kg of material by 1°C

Material	c (kJ/kg/K)
water	4.186
steel	0.45
cast iron	0.54
aluminium	0.92

Molar specific heat, or **molar heat capacity** (C), is the amount of heat required to raise the temperature of 1 mole of the substance by 1°C.

$$\Delta Q = nC(\Delta T)$$

Experiments show that when a gas is heated at constant volume, the **molar heat capacity** C_v is always less than that if the gas is heated at constant pressure C_p. This is because for a constant volume process, there is no opportunity for there to be mechanical work done on or by the system.

For a constant pressure process, for a given temperature rise ΔT, there will always be a volume change ΔV. Therefore the energy into the system has to both raise the temperature *and* do work, thus C_p is always greater than C_v.

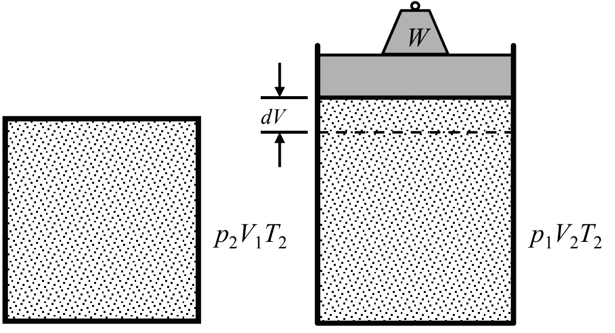

$$Q_v = nC_v(T_2 - T_1)$$

$$Q_p = nC_v(T_2 - T_1) + p(V_2 - V_1)$$
$$= nC_p(T_2 - T_1)$$

For a constant pressure process, a volume change involves a force acting through a distance and hence work is done on or by the system. Volume change in solids and liquids very small, and hence distinction between C_p and C_v not usually made.

↓ Q_p is a very important quantity in industrial processes and is given the name **enthalpy**.

6.3 Enthalpy

Consider a beaker of solution which is heated from an external source. That is, heat energy is transferred into the system across the boundary.

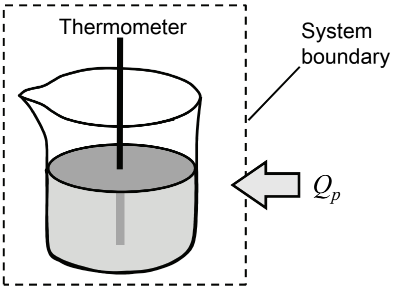

The temperature of the solution increases by an amount dependent on the **molar heat capacity** C_V of the solution. The change in **internal energy** of the system will be:

$$\Delta U = nC_v(\Delta T)$$

The solution may also expand by an amount ΔV and perform mechanical **work** (e.g. by a piston that fits into the top of the beaker and floats on the surface of the solution):

$$W = p\Delta V$$

The relationship between the heat entering the system, the work done by the system and the rise in internal energy is expressed by the **first law of thermodynamics**:

$$Q_p = \Delta U + p\Delta V$$

Q_p is positive when heat enters the system. W is positive when work is done by the system on the surroundings. ΔU is positive when ΔT is positive.

If heat energy Q_p is added to the system, then some of this heat goes into mechanical work and some into internal energy. The sum of the change in internal energy and the work done on or by the system is called the change in **enthalpy** ΔH of the system:

$$\begin{aligned}\Delta H &= \Delta U + p\Delta V \\ &= Q_p\end{aligned}$$

Note: $p\Delta V$ is usually small in comparison to ΔU in experiments in the laboratory whereupon ΔH is approximately equal to ΔU.

That is, the change in enthalpy of the system is the same as the amount of heat energy that has been put into it.

If the heating of the solution involves a **phase change**, then some of the heat energy input would be consumed by latent heat which would have to be accounted for.

If the heating of the solution initiates a **chemical reaction**, then heat energy may be released (e.g. as happens during combustion) or absorbed by the solution as the potential energy of the **chemical bonds** (chemical energy) changes. This is called the **heat of reaction**. The heat of reaction is a release, or absorption of potential energy of the chemical bonds.

6.4 Heat of Reaction

Consider the reaction between two chemical species in an aqueous solution in a beaker which is placed in a temperature controlled water bath so that the beaker is maintained at a constant temperature. That is, heat flow can occur across the system boundary but the solution is held **isothermal**. If, as a result of the reaction, heat energy is released, then this heat must pass out through the system boundary. The surrounding water bath ensures that the heat energy is taken away fast enough to prevent any temperature rise of the solution.

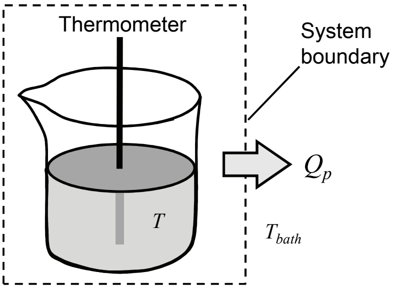

- The **internal energy** of the system remains the same after the reaction as indicated by no rise in temperature.
- **Energy** has come from the **heat of reaction** and been transferred into the water bath.
- If there has been a volume change, then mechanical **work** also would have been performed on (+ve) or by (-ve) the system.

In this example, there must have been a net release of bond energy as a result of the chemical reaction. Some of this energy has been carried away across the boundary by the water bath, and some (a small amount in this case) has gone into mechanical work against the pressure of the atmosphere. The change in **enthalpy** of the system is:

$$\Delta H = \Delta U + p\Delta V$$

We can assume that the mechanical work involved in this example is small compared to the heat lost by the system to the surroundings. Thus, $\Delta H = Q_p$. In this example, heat Q_p has passed out of the system and so the change in enthalpy is a decrease. When heat is passed out to the surroundings, the reaction is said to be **exothermic**. If the reaction involved an absorption of heat from the surroundings, then the change in enthalpy would be positive, and the reaction is called **endothermic**.

The heat of reaction represents the change in enthalpy in this isothermal system. The question now is, if there is a chemical reaction, *and* the temperature changes, how is the change in enthalpy calculated?

6.5 Heat of Reaction

Consider the reaction between two chemical species in an aqueous solution in an insulated beaker. A **thermometer** measures the temperature change during the reaction.

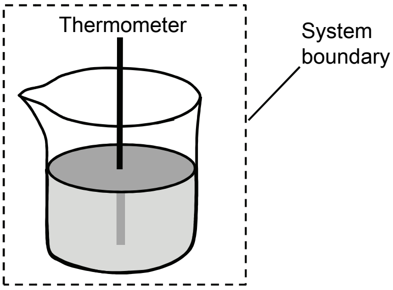

The reaction takes place inside the boundary of a system. For the present example, we will assume that the system boundary is **adiabatic**. That is, no heat flow will occur across the boundary, as if the beaker were fully insulated.

If, as a result of the reaction, the temperature of the solution rises, then:

- The **internal energy** of the system is higher after the reaction, as indicated by a rise in temperature.
- **Energy** has come from the **heat of reaction** and been transferred into internal energy of motion of the molecules in the solution.
- If there has been a volume change, then mechanical **work** also would have been performed on (+ve) or by (-ve) the system.

Since no external heat energy was involved (no heat flow across the system boundary), the heat energy needed to raise the temperature of the solution must have come from the net sum of energies from the breakage and formation of the **chemical bonds** involved in the reaction. In this example, there must have been a net release of bond energy as a result of the reaction. Some of this energy has gone into internal energy of motion of the molecules, while some (a small amount in this case) has gone into mechanical work against the pressure of the atmosphere. In this case, $\Delta H = \Delta U + p\Delta V$. If the work done is negligible, then $\Delta H \approx \Delta U$. In this case, the change in enthalpy is equal to the **heat of reaction**. Although no heat has crossed the system boundary, heat has appeared as a result of the chemical reaction – just as if it had come from outside the system. That is, in this *adiabatic* example, heat Q_p has entered the system from the chemical reaction and $\Delta H \approx \Delta U$.

6.6 Enthalpy of Formation

A chemical reaction is the process of breakage and formation of chemical bonds. Energy is absorbed when a bond is broken, and energy is released when a bond is formed. When a chemical reaction occurs, heat may be either given off, or absorbed to or from the surroundings. That is, the change in enthalpy may be either positive (**endothermic**) or negative (**exothermic**).

Liquid water molecules are made by the reaction between H_2 and O_2 gas molecules:

$$H_2(g)+\frac{1}{2}O_2(g)\rightarrow H_2O(aq)$$

The equation is written in terms of the formation of 1 mole of product. Under standard conditions (1 atm pressure and 25ºC) it is found that 286 kJ of heat energy is produced. That is, the change in enthalpy of the system is $\Delta H = -286$ kJ. This is the **heat of reaction**.

When the reactants are at standard conditions, and the products are also kept at standard conditions (e.g. by use of a water bath to maintain temperature), the heat of reaction is referred to as the **heat of formation**, or **enthalpy of formation**. Standard heats of formation are determined by experiment.

Name	Molecular Formula	Heat of Formation ΔH, kJ/mol
Methane	CH_4	-75
Carbon dioxide	CO_2	-394
Ammonia	NH_3	-46
Acetylene	C_2H_2	+227
Ethylene	C_2H_4	+227
Ethane	C_2H_6	-85
Water	H_2O	-286
Hydrochloric acid	HCl	-92
Sodium hydroxide	NaOH	-427
Aluminium oxide	Al_2O_3	-1670
Carbon monoxide	CO	-110

When the reaction proceeds in reverse, the sign of the heat of formation is reversed.

The reference point for heats of formation are elements at standard conditions. For example the heat of formation of $O_{2(g)}$ is set to zero.

Heats of any reaction can be obtained by addition of heats of formations of known compounds, or **Hess' law**. The change in enthalpy for the combustion of ethane is:

$$C_2H_6+\frac{7}{2}O_2\rightarrow 2CO_2+3H_2O$$

$$(C_2H_6(g)\rightarrow 2C(s)+3H_2(g)) \quad \Delta H=+85$$

$$2(C+O_2\rightarrow CO_2) \qquad \Delta H=2(-394)$$

$$3\left(H_2+\frac{1}{2}O_2\rightarrow H_2O\right) \qquad \Delta H=3(-286)$$

$$\Delta H=85-2(394)-3(286)=-1561\ \text{kJ/mol}$$

Note that the sign of ΔH for C_2H_6 has been reversed in this calculation since the reaction splits this compound rather than forms it. ΔH for the combustion of ethane is called the **calorific value**.

6.7 Entropy

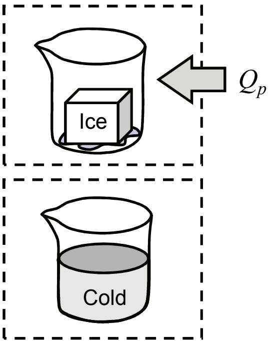

Consider a beaker of ice which is heated from an external source. The ice remains at 0°C (273K) and melts. The amount of energy Q_p needed is determined by the **latent heat of fusion** of ice.

The water molecules have a greater amount of disorder after melting compared to that before heating due to their increased velocity. A quantitative measure of disorder is **entropy**.

Absolute values of entropy S can be calculated from thermodynamic theory or measured in the laboratory. Of more interest to us in chemistry are changes in entropy as a result of physical changes of state, and chemical reactions. When a system undergoes a process at temperature T involving heat flow, the *change* in entropy of the **system** is:

$$\Delta S = \frac{Q}{T}$$

When a system absorbs heat, Q is positive and so the entropy increases, $\Delta S>0$. When a system rejects heat, Q is negative and the entropy decreases, $\Delta S<0$.

If the process occurs at a low temperature, then the change in entropy is greater than if the same process occurs at a higher temperature. At high temperatures, the entropy is already high and so the proportional change in entropy for a given heat flow Q is less compared to that which would occur at low temperatures.

For 1 mole ice:

$$\begin{aligned}\Delta S &= \frac{mL}{T}\\ &= \frac{0.018(335000)}{273}\\ &= 22.1\ \text{J/K}\end{aligned}$$

When a system undergoes a process and interacts with its surroundings (i.e. a non-isolated system), the total **entropy change** (of the universe) is:

$$\Delta S_{total} = \Delta S + \Delta S_{\text{surroundings}}.$$

The sign of ΔS_{total} signifies the presence of the following types of processes:

$\Delta S_{\text{total}} > 0$ irreversible process
$\Delta S_{\text{total}} = 0$ reversible process
$\Delta S_{\text{total}} < 0$ impossible process

Entropy is not energy. There is no "law of conservation of entropy". Irreversible processes *create* entropy.

When a phase change or a chemical reaction involves heat flow to or from the surroundings, entropy must be considered. The **total entropy** of the system and the surroundings cannot decrease. If the total entropy change is zero, then the process is reversible and so cannot be **spontaneous** in any one direction. For a reaction to proceed spontaneously from reactants to products, the total entropy ΔS_{total} must increase.

6.8 Entropy Calculations

Changes in entropy of a system at constant temperature are expressed as:

$$\Delta S = \frac{Q}{T}$$

When the temperature is not constant, we add up incremental changes in entropy for incremental transfers of heat:

$$dQ = mcdT$$

$$\begin{aligned}\Delta S &= \int_{Q_1}^{Q_2} \frac{1}{T} dQ \\ &= \int_{T_1}^{T_2} \frac{1}{T} mc\, dT \\ &= mc \ln \frac{T_2}{T_1}\end{aligned}$$

The **standard molar entropy** S^o is the entropy content of one mole of substance under standard conditions of 1 atm pressure and 25°C. Standard entropies are absolute values and represent the change in entropy to bring the substance from 0K to 273K in its standard state. Standard entropies are always positive. The zero point for entropy is 0K. This statement of zero point for entropy is called the **third law of thermodynamics**.

Element	S^o. (J/K/mol)
H_2(g)	131
O_2(g)	205
N_2(g)	192
Si(s)	19
C(s)	6
Fe(s)	27

Compound	S^o (J/K/mol)
NH_3(g)	192.8
H_2O(l)	70.0
H_2O(g)	188.8
CO_2(g)	213.8
CO(g)	197.7
CH_4(g)	186.2

For a chemical reaction or a physical process, the change in entropy can be calculated from:

$$\Delta S_{\text{reaction}} = S^o{}_{\text{products}} - S^o{}_{\text{reactants}}$$

For example:

$$H_2(g) + \frac{1}{2} O_2(g) \rightarrow H_2O(g)$$

$$\begin{aligned}\Delta S^o &= 188 - (131 + 0.5 \times 205) \\ &= -45.5 \text{ J/K/mol}\end{aligned}$$

Note that in this example, the change in entropy for this system is < 0 at the expense of an increase in entropy elsewhere in the universe. Hydrogen and oxygen will, under certain conditions, form water since it is energetically favourable to do so.

6.9 Gibbs Energy

It is a natural tendency of systems to attain the lowest possible potential energy, such as a ball rolling down a hill, electrons moving in an electric field, and so on. In an **exothermic** process, heat leaves a system ($\Delta H < 0$) and we might expect that such processes are always spontaneous and that **endothermic** processes are not. However, this is not observed. For example, when water evaporates spontaneously at room temperature, $\Delta H > 0$ and the water vapour has a higher energy than the liquid.

$$H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O(aq) \qquad \Delta H = -286\ \text{kJ/mol}$$

$$H_2(g) + \frac{1}{2}O_2(g) \rightarrow H_2O(g) \qquad \Delta H = -242\ \text{kJ/mol}$$

$$H_2O(aq) \rightarrow H_2O(g) \qquad \Delta H = 286 + -242 = +44\ \text{kJ/mol}$$

$+\Delta H$ signifies energy of products is greater than reactants (endothermic)

Whether or not a phase change or a chemical reaction proceeds spontaneously cannot be judged on the basis of energy alone. The change in entropy ΔS_{total} of both the system (the reaction) and the surroundings must also be considered. Whether or not the change in entropy or enthalpy is the more important factor in any particular process depends on the absolute temperature T.

Now, the total energy of a system is given by the **enthalpy** and this consists of the internal energy and the external work. Changes in total energy are expressed as: $\Delta H = \Delta U + p\Delta V$.

Part of the total energy H is unavailable for conversion into work. This part is connected with changes in **entropy**. Since the units of entropy are J/K, and the determining factor that quantifies the contribution of entropy in a particular process is T, then the unavailable energy can be calculated from the product TS. Thus, the total energy of a *system* can be written as:

$$H = G + TS$$

where G represents that part of the total energy that is freely available for conversion into work. This freely available energy is called the **Gibbs energy**. As the temperature increases, the amount of Gibbs energy decreases and the unavailable term TS increases. Changes in the amount of Gibbs energy in a constant temperature process are thus: $\Delta G = \Delta H - T\Delta S$.

The concept of Gibbs energy allows us to predict the direction of spontaneous chemical reactions by considering both energy and entropy. A spontaneous chemical reaction will proceed only when there is a decrease in available Gibbs energy: $\Delta G < 0$.

6.10 Spontaneous Processes

The concept of Gibbs free energy embodies both energy and entropy considerations in determining in which direction a chemical reaction proceeds. A decrease in Gibbs energy can arise due to a decrease in **enthalpy** or an increase in the product *TS*. In a particular chemical reaction, ΔH and ΔTS may have the same or opposite signs. For a constant pressure process, at constant temperature:

$$\Delta G = \Delta H - T\Delta S.$$

- If a reaction is exothermic $\Delta H < 0$ and the entropy change $\Delta S > 0$, then the free energy change $\Delta G < 0$ and the process is **spontaneous**.
- If a reaction is endothermic $\Delta H > 0$ and the entropy change $\Delta S < 0$, then the free energy change $\Delta G > 0$ and the process is not spontaneous.
- If the enthalpy change $\Delta H > 0$ and the entropy change $\Delta S > 0$, then the process is spontaneous at high temperature. When $\Delta H < 0$ and $\Delta S < 0$, the process is spontaneous at low temperature.

It is important to know that ΔS in the above equation is the entropy change of the *system*, not the total entropy change (of the universe). This enables us to conveniently predict and quantify the direction of a process based upon attributes of the system only.

In chemical reactions, the change in Gibbs energy depends on the physical states of the reactants and products, the temperature and pressure, and the concentration of the reactants and products. When reactants and products are at standard conditions (1 atm and 25°C), the free energy change is referred to as the **standard Gibbs energy change** ΔG^o.

The standard Gibbs energy change of the formation of a compound from its elements is called the **standard Gibbs energy of formation** $\Delta_f G^o$. Elements at their standard states have zero standard Gibbs energy of formation.

For a process or a chemical reaction, the change in Gibbs energy can be calculated from:

$$\Delta G_{\text{reaction}} = \Delta_f G^o{}_{\text{products}} - \Delta_f G^o{}_{\text{reactants}}$$

Compound	$\Delta_f G^o$. (kJ/mol)
$NH_3(g)$	−16.6
$H_2O(l)$	−237.2
$H_2O(g)$	−228.6
$CO_2(g)$	−394.3
$CO(g)$	−137.3
$CH_4(g)$	−50.8

6.11 Melting of Ice

Consider a mole of ice initially at 0°C. Now, let the temperature of the surroundings be 283K (+10°C).

For the system: $\Delta H = mL + mc\Delta T$

$$= 0.018(335000) + 0.018(4186)(10)$$

$$= 6780\text{ J} \quad \text{(endothermic)}$$

For the surroundings:

$$\Delta S_{\text{surroundings}} = \frac{-6780}{283} = -23.9\text{ J/K}$$

For the system: $dQ = mL + mcdT$

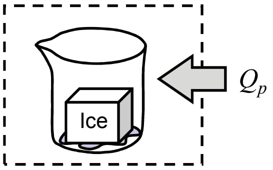

$$\Delta S = \frac{mL}{T_1} + \int_{Q_1}^{Q_2} \frac{1}{T} dQ$$

$$= \frac{mL}{T_1} + \int_{T_1}^{T_2} \frac{1}{T} mc\, dT$$

$$= \frac{0.018(335000)}{273} + 0.018(4186)\ln\frac{283}{273} = 24.8\text{ J/K}$$

Total change in entropy:

$$\Delta S_{\text{total}} = \Delta S + \Delta S_{\text{surroundings}}$$

$$= 24.8 - 23.9$$

$$= +0.9\text{ J/K}$$

This is an **irreversible process**. Even though the change in entropy is < 0 for the *surroundings*, the change in entropy of the system is greater, and so the change in **total entropy** is > 0.

$$\Delta G = \Delta H - T\Delta S$$

$$= 6780 - 283(24.8)$$

$$= 6780 - 7018$$

$$= -238\text{ J}$$

The **Gibbs free energy** change is < 0 so this process is spontaneous.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
$L_v = 22.57 \times 10^5$ J/kg
$L_f = 3.35 \times 10^5$ J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.12 Freezing of Water

Consider a mole of water, initially at 0°C. Now, let the temperature of the surroundings be 263K (−10°C).

For the system: $\Delta H = -mL + -mc\Delta T$

$$= -0.018(335000) + -0.018(2110)(10)$$
$$= -6.03 - 0.38$$
$$= -6410\ \text{J} \qquad \text{(exothermic)}$$

For the surroundings:

$$\Delta S_{\text{surroundings}} = \frac{6410}{263} = 24.4\ \text{J/K}$$

For the system: $dQ = -mL + mcdT$

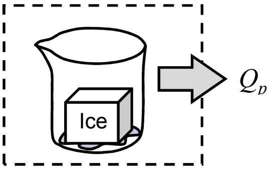

$$\Delta S = \frac{-mL}{T_1} + \int_{Q_1}^{Q_2} \frac{1}{T} dQ$$
$$= \frac{-mL}{T_1} + \int_{T_1}^{T_2} \frac{1}{T} mc\, dT$$
$$= \frac{-0.018(335000)}{273} + 0.018(2110)\ln\frac{263}{273} = -23.52\ \text{J/K}$$

Total change in entropy:

$$\Delta S_{\text{total}} = \Delta S + \Delta S_{\text{surroundings}}$$
$$= -23.52 + 24.4$$
$$= +0.88\ \text{J/K}$$
$$> 0$$

This is an **irreversible process**. Even though the change in entropy is < 0 for the *system*, the change in entropy of the surroundings is greater, and so the change in **total entropy** is > 0.

$$\Delta G = \Delta H - T\Delta S$$
$$= -6410 - 263(-23.5)$$
$$= -6410 + 6180$$
$$= -229\ \text{J}$$

This is a spontaneous process since the freezing of ice results in a decrease in **Gibbs free energy**.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
$L_v = 22.57 \times 10^5$ J/kg
$L_f = 3.35 \times 10^5$ J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.13 Ice/Water Equilibrium

Consider the conversion of one mole of water to ice at 273K:

For the system: $$\Delta H = -mL$$
$$= -0.018(335000)$$
$$= -6030\text{ J} \quad \text{(endothermic)}$$

For the surroundings:
$$\Delta S_{\text{surroundings}} = \frac{6030}{273} = 22.08\text{ J/K}$$

For the system: $$dQ = -mL$$
$$\Delta S = -\frac{mL}{T}$$
$$= \frac{-0.018(335000)}{273}$$
$$= -22.08\text{ J/K}$$

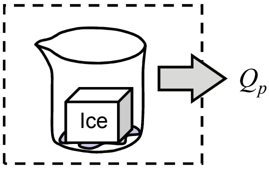

Total change in entropy:
$$\Delta S_{\text{total}} = \Delta S + \Delta S_{\text{surroundings}}$$
$$= -22.08 + 22.08$$
$$= 0\text{ J/K}$$

The total change of entropy is zero. This is a **reversible process**. Extraction of Q_p from the melted ice would cause it to re-freeze.

$$\Delta G = \Delta H - T\Delta S$$
$$= -6030 - 273(-22.08)$$
$$= 0\text{ J}$$

This process is at **phase equilibrium**.

Data:
c_{water} = 4186 J/kg/K
c_{Al} = 920 J/kg/K
L_v = 22.57 × 10^5 J/kg
L_f = 3.34 × 10^5 J/kg
ρ_{water} = 1000 kg/m^3
c_{ice} = 2110 J/kg/K

6.14 Chemical Equilibrium

A chemical reaction involves reactants and products. Whether or not a chemical reaction proceeds spontaneously depends on the sign of the change in free energy.

$$\Delta G_{\text{reaction}} = \Delta_{\text{f}} G^0{}_{\text{products}} - \Delta_{\text{f}} G^0{}_{\text{reactants}}$$

$$\text{Reactants} \rightarrow \text{Products}$$

If a chemical reaction proceeds from reactants to products spontaneously, we may at first expect the process to continue until the reactants are all converted to product. This is not observed. Usually, reactions proceed until there is **chemical equilibrium**:

$$\text{Reactants} \leftrightarrow \text{Products}$$

At equilibrium, the rate of forward reaction = rate of reverse reaction.

In chemical reactions, the **free energy** change depends on the temperature, physical state, pressure, and concentrations of the reactants and the products. That is, as the reaction proceeds, the reactants are consumed and the products are created – and the free energy change changes as the reaction proceeds in time. When the free energy change becomes equal to zero, the rate of forward reaction has become equal to the rate of the reverse reaction. This usually happens (for spontaneous reactions) when the concentration of the products is somewhat greater than those of the reactants – but it is important to note that this does not necessarily mean that all the products are used up. At **chemical equilibrium**, the forward and reverse reactions are both happening at the same rate.

Precisely what is meant by **rate of reaction** depends upon the context. Rates of reaction may be expressed as the change in concentration of one of the reactants per unit time where concentration is expressed in terms of moles per unit volume.

State	Concentration units
Gases	Partial pressure units
Liquids	Molarity (moles per litre)
Solids	Moles or mass units

The appropriate time period may be expressed in milliseconds, seconds, days or years depending on the reaction.

6.15 Statistical entropy

The change in entropy of a system can be calculated from macroscopic quantities, Q and T. Molecules of gas in a volume V at temperature T have a distribution of velocities, and hence, kinetic energies. The distribution of velocities over the total number of molecules was determined by Boltzmann:

The number of molecules N with an energy E was computed by Boltzmann using the **Maxwell velocity distribution** function:

$$N = Ce^{\frac{-E}{kT}}$$ Maxwell-Boltzmann *energy* distribution

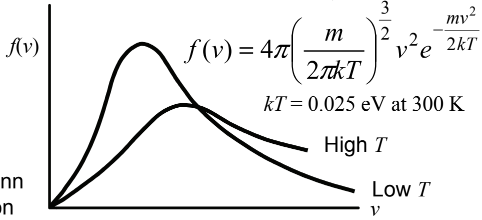

Consider a partitioned volume V containing N molecules. There is a discrete number of energies possible, and there is a finite number of ways that the molecules present can have a distribution of energies that matches the macroscopic properties of P, T and V. Each of these possibilities is called a **microstate**.

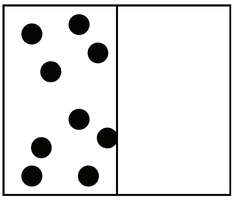

Now, one such possible state is if all the energy of the system were concentrated in one molecule and the others had zero energy. We know that this is very unlikely to ever happen, just as we know that the likelihood of throwing zero heads and 100 tails in 100 tosses of a coin is unlikely. The most likely scenario for our N molecules is that the energies will follow the Maxwell-Boltzmann distribution.

In mathematical symbols, the **statistical weight function** of the system $\Omega(E,V,N)$ is the number of possible microstates that are consistent with the observed **macrostate**. The lower the number of possible microstates, the more accurately we know information about the molecules. The entropy of the system is: $S = k \ln \Omega$.

If there were only one possible microstate, then S would be zero. As the number of possible microstates increases, the entropy becomes larger. When the partition is removed, the molecules rush to fill the entire volume. It is exceedingly unlikely that at any one instant, all the molecules will be located over on one half of the now larger volume. The number of possible microstates has increased. However, one might ask what if the container had only two molecules in it? In this case, it is likely that at any one time, the two molecules *will* find themselves over to one side of the container. The number of possible microstates, for the two-molecule scenario, is very small, and so is the entropy.

7. Rates of Reaction

Summary

Rates of reactions:

- The nature of the reactants
- The concentration of the reactants
- The temperature
- The presence of a catalyst

Rate law (concentrations):

$$aA + bB \rightarrow P$$

$$\frac{d[P]}{dt} = k[A]^n [B]^m$$

Rate law (temperature):

$$k = Ae^{\frac{-E}{RT}}$$ Arrhenius equation

Catalyst:

- Has no effect on the stoichiometry of the overall reaction.
- Does not affect the equilibrium position of the reaction.
- Affects the reaction mechanism so as to lower the activation energy, thus increasing the number of molecules that can make a productive collision.

7.1 Rates of Reaction

Consider the reaction between hydrogen gas and oxygen gas to produce liquid water:

$$H_2(g)+\frac{1}{2}O_2(g)\rightarrow H_2O(aq) \quad \Delta H=-286\text{ kJ/mol}$$

$$H_2(g)+\frac{1}{2}O_2(g)\rightarrow H_2O(aq)$$

$$\begin{aligned}\Delta S^{o}&=70-(131+0.5*205)\\&=-163.3\text{ J/K/mol}\end{aligned}$$

$$\begin{aligned}\Delta G&=\Delta H-T\Delta S\\&=-286000-273(-163.3)\\&=-286000+44580.9\\&\approx-240\text{ kJ/mol}\end{aligned}$$

The indications are that this reaction should proceed spontaneously when these two gases are mixed. While this might be so, experience indicates that this reaction does not proceed at all unless there is a spark or source of ignition present – in which case the reaction proceeds with explosive violence.

Clearly then, the **rates of reaction** are an important part of chemical reactions. While thermodynamics might provide information about the direction in which a reaction might proceed on its way to chemical equilibrium, more information is needed to determine at what rate the products are produced.

Associated with the rate of reaction is the **reaction mechanism**. The production of products and the consumption of reactants, while observable, only represents the net reaction in general. Chemical reactions may proceed in several steps. Intermediate species may form and be consumed, and these intermediate steps often influence the overall rate of the reaction as a whole.

The most important factors that govern the rate of a reaction are

- The nature of the reactants
- The concentration of the reactants
- The temperature
- The presence of a catalyst

The rate of reaction is the rate of change of **concentration** of either the reactants or the products.

7.2 Collision Theory

Chemical reactions involve the breaking and formation of bonds between atoms, ions or molecules. Elements or compounds that comprise the reactants form new compounds as products. The relationship between energy and entropy – the **Gibbs free energy** – determines which direction a chemical reaction will proceed. The question now is to ask just how a chemical reaction takes place.

The notion of chemical reactions only taking place between atoms or molecules when they contact each other in a certain way is not so obvious as it might seem and was only proposed in the early 1900s.

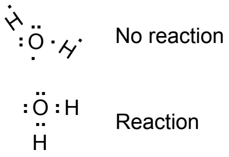

Central to the answer of this question is the fact that atoms, ions and molecules are in constant motion – especially when they are in liquid and gaseous form. Because they are in motion, they undergo collisions with other atoms and molecules.

When atoms, ions or molecules collide, a chemical reaction may occur. That is, chemical reactions occur as a result of collisions between molecules. This is known as **collision theory**.

During the collision, bonds are broken and formed as atoms and electrons are rearranged into new compounds. Several variables serve to determine how fast reactants are converted into products:

- Particles (either atoms, ions or molecules) have to collide in a particular way. That is, the orientation of the colliding particles has to be a certain way in order that the collision be effective at breaking bonds and forming new ones.
- The particles have to possess enough kinetic energy to allow them to interact to the extent that bonds may be broken and allow new ones to be formed.
- The number of collisions per unit time determines how fast the reaction proceeds and this is related to the concentration of the reactants.

The nature of the reactants, the temperature, and concentration and action of **catalysts** can be all explained by collision theory. Not all collisions result in a reaction. Only a small proportion are **productive collisions**.

7.3 Reaction Mechanism

We have agreed that for a chemical reaction to proceed, atoms, ions, or molecules must collide. Further, they must not just touch or brush past one another. They must collide with a certain amount of kinetic energy, and be in the right orientation with respect to one another in order for bonds to be broken and allow new ones to form. That is, there has to be a **productive collision**.

Consider the rusting of iron. Simply put, the reaction is:

$$4\mathrm{Fe(s)} + 3\mathrm{O_2(g)} \rightarrow 2\mathrm{Fe_2O_3(s)}.$$

This is a reasonably slow reaction. However, if one were to just take collision theory at face value, it would be luck indeed if four atoms of iron were to collide with three molecules of oxygen at the same time to form two molecules of iron oxide. In reality, reactions usually proceed as a series of smaller steps. The series of smaller steps is called the **reaction mechanism**.

$$\mathrm{Fe^{2+}} + 2\mathrm{OH^-} \rightarrow \mathrm{Fe(OH)_2}$$

$$4\mathrm{Fe(OH)_2} + \mathrm{O_2} + 2\mathrm{H_2O} \rightarrow \mathrm{Fe(OH)_3(s)}$$

$$\mathrm{Fe(OH)_3(s)} \rightarrow \mathrm{Fe_2O_3, H_2O} + 2\mathrm{H_2O}$$

The chances of a collision occurring between two atoms, ions or molecules is significantly greater than the chances of a collision between three or four of the required chemical species at the same time. Most reaction steps in chemical reactions involve two atoms or molecules. The slowest step in the reaction mechanism is the **rate-determining step**. The nature of the reaction steps cannot be deduced from the net reaction. The reaction steps can only be found by experiment.

Identification of the rate-determining step is important for establishing a quantitative rate law for the reaction as a whole. Often, the rates of reaction depend on the concentration of the reactants, but it is the concentration of the reactants within the rate-determining step that is they key issue. During the **elementary reactions**, or the **reaction steps**, intermediate chemical species may be formed which are used up in a later step and do not appear in the net equation.

Some reaction mechanisms are noted for their explosive properties. In these cases, the steps become a self-sustaining **chain reaction** (such as in **combustion**).

7.4 Activation Energy

A reaction step usually involves a two-particle collision (either atoms, ions or molecules). As two molecules approach each other, the outer electrons of one molecule will begin to feel the electrostatic repulsion of the other molecule. The kinetic energy of motion will be converted into electrical potential energy. The potential energy rises as the molecules come closer and closer together.

When all the kinetic energy has been expended, the molecules are in very close proximity and are in a somewhat unstable state. The molecules at this point form what is called an **activated complex**. The activated complex is a conglomeration of the participating atoms. At this point, the activated complex may split into the product molecules (the forward reaction) or bounce backwards and reform the original molecules (the reverse reaction).

The colliding molecules require sufficient **kinetic energy** to overcome mutual repulsion and allow the collision and formation of the activated complex. This is called the **activation energy** E_{act}. The activation energy depends on the nature of the reactants.

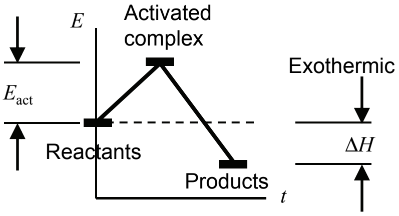

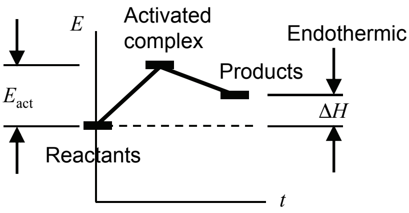

When a spark is introduced into a mixture of H_2 and O_2, the energy of the spark overcomes the activation energy and allows the reaction to proceed. The heat of the reaction furnishes the required activation energy for further reactions. That is, the H_2 gas undergoes **ignition**.

The activation energy is the difference in energy between the activated complex and the average energy of the reactants.

If a particular reaction has a large activation energy, then only a small proportion of collisions may result in enough kinetic energy to overcome it, and the reaction will be slow. If a particular reaction has a lower activation energy, then a greater proportion of collisions will have enough kinetic energy to overcome it and the reaction proceeds at a faster rate.

7.5 Nature of Reactants

The rate of a chemical reaction, that is, the rate of change of concentration of either the reactants or the products, depends on the nature of the reactants. When the permanganate ion is reduced by an Fe^{2+} ion in an acidic solution, the reaction is very fast and is limited by the time it takes to mix the solutions:

$$5Fe^{2+}(aq) + MnO_4^-(aq) + 8H^+(aq) \rightarrow 5Fe^{3+}(aq) + Mn^{2+}(aq) + 4H_2O \quad \text{fast}$$

But when the permanganate ion is reduced by oxalic acidic $H_2C_2O_4$, the reaction is fairly slow, taking several minutes.

$$5C_2O_4^{2-}(aq) + 2MnO_4^-(aq) + 16H^+(aq) \rightarrow 10CO_2(g) + 2Mn^{2+}(aq) + 8H_2O \quad \text{slow}$$

The main difference between the two reactions is the reducing agent. It could be argued that in the case of $H_2C_2O_4$, many more bonds must be broken compared to the case of the Fe^{2+} solution and this might influence the rate of the reaction. That is, the reaction mechanism for the second reaction may be quite different in character than that of the first reaction.

Experiments indicate that:

- Reactions that involve the breaking and formation of **covalent bonds** are usually slow at room temperature.
- Reactions that involve electron transfer, such as **ionic reactions** in aqueous solutions, are generally fast. Examples are precipitation reactions, **acid** on **base** reactions, most oxidation/reduction reactions.

Since the **activation energy** depends on the chemical species involved, then so does the rate of reaction. There is more chance of a successful or productive collision in the case of a low activation energy because a greater number of particles will have sufficient kinetic energy to overcome it.

It is not only the chemical nature of the reactants, but the physical nature as well. Since collision theory requires atoms or molecules to make contact to form the activated complex, any mechanism that affects the degree of contact between atoms or molecules will affect the rate of reaction. In homogenous systems, such as aqueous solutions, there is intimate contact, whereas in inhomogeneous systems, such as solid/liquids, or solid/gas, reactions only take place at the surface of contact of the atoms or molecules. Increasing the surface area by mixing or grinding increases the rate of reaction. A common example is that sawdust is almost explosive in air, while a block of wood burns relatively slowly in comparison.

7.6 Concentration

One of the most important effects on the rate of reaction is the **concentration** of one or more of the reactants. Generally speaking, increasing the concentration of a reactant increases the rate of the reaction (but not always). It is easy to accept that the chances of a successful, or productive, collision between two reacting particles is increased if there are more of them, per unit volume, involved in the process. Consider a reaction:

$$A + B \rightarrow AB.$$

If we only had one molecule of A and one of B, then we may have to wait some time before they encounter each other and collide in a way so as to produce the reaction. The collision rate, and hence the reaction rate, would be low.

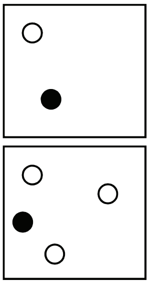

If we now increase the concentration of A, but putting more molecules of A in the reaction chamber, then there would be an increase in the chances of a successful collision – the reaction rate would increase. It is possible that doubling the concentration of one or other of the reactants would double the reaction rate.

The concentration of a species in moles per litre can be conveniently written using [ ] notation. Thus, the reaction rate is indicated by the rate of change of appearance of the product, or the rate of change of concentration of AB.

Reaction rate is expressed as moles/litre/second

$$\frac{d[\mathrm{AB}]}{dt} = k[\mathrm{A}][\mathrm{B}]$$

This equation is called a **rate law**. *k* is the constant of proportionality, or the **rate constant**. It should be noted that the rate law actually observed for a reaction cannot be necessarily deduced from the balanced net equation. This is because one of the intermediate reaction steps may be the rate-determining reaction, in which case the rate law for that step would be required. Unless more of the reactants are continually added, the rate of reaction will generally decrease with time as the reactants are consumed (and their concentrations decrease).

7.7 Rate Law

Consider the reaction: $A + A \rightarrow A_2$

$2A \rightarrow A_2$ Balanced equation

The **rate law** would be written:

$$\frac{d[A_2]}{dt} = k[A][A]$$

$$= k[A]^2$$ Rate law

Note that the exponent of [A] in the rate law is the same as the coefficient of A in the reactants in the balanced equation. This is a second-order reaction.

In general, for an equation of the form: $aA + bB \rightarrow P$ the rate law is written as:

$$\frac{d[P]}{dt} = k[A]^n[B]^m.$$

For a single-step reaction, the exponents *n* and *m* are equivalent to *a* and *b*. However, a chemical reaction nearly always proceeds via a **reaction mechanism**, that is, a series of reaction steps. Thus, in most cases, there is no connection between the exponents *n*, *m* and the coefficients *a*, *b* for the net reaction. In general, the rate law for the net reaction has to be determined by experiment.

The exponents *n* and *m* indicate the **order of the reaction** with respect to that reactant. For example, when $n = 2$ and $m = 1$, we say that the reaction is second order with respect to A and first order with respect to B. The overall order of the reaction is the sum $m+n$. When *n* or *m* equals zero, then the rate of reaction is independent of the concentration of the reactant and is a **zero-order reaction** with respect to that reactant.

The **rate constant**, *k*, is specific to the particular reaction and depends on the experimental conditions. The units of *k* depend on the order of the reaction. A high value of *k* indicates fast reaction. For first-order reactions, if $k > 10$ sec^{-1}, then the reaction is considered instantaneous.

The **rate law** shows how the instantaneous rate of a reaction changes with the concentration of the reactants, as long as other factors such as temperature, the nature of the reactants and the presence of catalysts remain constant.

The rate of a reaction with respect to either reactants or products is expressed in terms of the **rate constant** *k* and the **order** of the reaction.

7.8 Rates of Reactions

There are several methods of expressing the rates of a reaction:

Average rate of disappearance of a reactant R: $-\frac{[\mathrm{R}]_2 - [R]_1}{t_2 - t_1}$

Average rate of appearance of a product P: $\frac{[\mathrm{P}]_2 - [P]_1}{t_2 - t_1}$

Instantaneous rate of disappearance of a reactant: $-\frac{d[\mathrm{R}]}{dt}$

Instantaneous rate of appearance of a product: $\frac{d[\mathrm{P}]}{dt}$

Half-life (the time taken for half of a reactant to be consumed or product to be formed): $t_{1/2}$

The rate of formation of products P as a whole is equal to the negative of the rate of disappearance of the reactants R:

$$\frac{d[\mathrm{P}]}{dt} = -\frac{d[\mathrm{R}]}{dt}$$

But we must be careful in expressing the rate in terms of a single product or reactant. For example, consider the following reaction:

$$\mathrm{H_2(g) + I_2(g) \rightarrow 2HI(s).}$$

Two moles of HI are produced at the rate of disappearance of 1 mole of H_2 per second. Thus, the rate of change of concentration, from **stoichiometry**, is

$$\frac{d[\mathrm{HI}]}{dt} = -2\frac{d[\mathrm{H_2}]}{dt}.$$

The rate at which H_2 disappears is the same as the rate at which I_2 disappears, and is half the rate at which HI appears. We cannot determine the order of the reaction from this net equation; we can only determine the relative magnitudes of the rates of products and reactants, no matter what the order for each participant might be.

7.9 Determination of Order

One of the most common ways of determining the rate law for a reaction is to measure the **initial rate of reaction** (e.g. formation of products) while varying the concentration of one of the reactants and holding the concentration of the other reactants constant.

$$\frac{d[\mathrm{P}]}{dt} = k[\mathrm{A}]^n[\mathrm{B}]^m$$

Let us hold the concentration of B constant and express the rate as the rate of change of concentration of A:

$$-\frac{d[\mathrm{A}]}{dt} = k'[\mathrm{A}]^n$$
$$\log\frac{d[\mathrm{A}]}{dt} = -n\log[A] - \log k'$$

By plotting the log of the initial reaction rate (d[A]/d*t*) versus log of the initial concentration [A], then the slope of the resulting straight line will be the order *n*.

Similarly, we may hold the concentration of A constant and vary the concentration of B to obtain a value for *m*.

$$-\frac{d[\mathrm{B}]}{dt} = k''[\mathrm{B}]^m$$
$$\log\frac{d[\mathrm{B}]}{dt} = -m\log[B] - \log k''$$

It is important to know that the concentrations, and the rates, are determined at the initial conditions, that is, the rate of the reaction is measured as the slope of the tangent of the rate curve at $t = 0$.

The procedure can also be performed numerically, where we substitute values of reaction rate and concentration into the general rate law and solve simultaneous equations for the two unknowns *n* and *m*.

This method (either graphical or arithmetic) is called the **initial rate method**.

7.10 Determination of 1st Order Rate Law

An example for a simple **first-order reaction** is the decomposition of hydrogen peroxide to water and oxygen:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

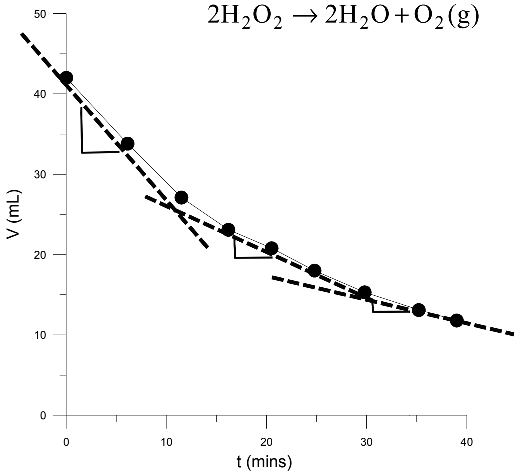

This is ordinarily a very slow reaction, but may be sped up by the addition of a catalyst. The concentration of H_2O_2 at any point is found by inhibiting the reaction by addition of dilute sulphuric acid, and the titration with $KMnO_4$. The volume of $KMnO_4$ used in the titration is an indication of the concentration of H_2O_2 left in the solution.

The slope of the tangent to the curve decreases as *t* increases, thus indicating that the rate of the reaction decreases with time. Since there is only one reactant, the decrease in reaction rate is therefore due to the decrease in concentration of H_2O_2. In general, the rate law would be written:

$$\frac{d[H_2O_2]}{dt} = k[H_2O_2]^n$$

For a first-order reaction, $n = 1$ and thus:

$$-\frac{d[H_2O_2]}{dt} = k[H_2O_2]$$

$$\frac{1}{[H_2O_2]}d[H_2O_2] = -kdt$$

$$\int_{[H_2O_2]_o}^{[H_2O_2]} \frac{1}{[H_2O_2]}d[H_2O_2] = -\int_0^t kdt$$

$$\ln[H_2O_2] - \ln[H_2O_2]_o = -kt$$

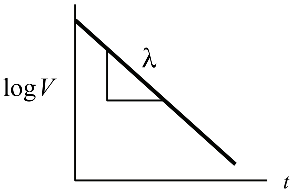

$[H_2O_2]$ is proportional to the titration volume in this experiment, and so a plot of log *V* versus *t* is a straight line if the reaction is first order, and the value of *k* can be found from the slope of a plot of log *V* versus *t*.

$$\ln[H_2O_2] = -kt + \ln[H_2O_2]_o$$

$$\log[H_2O_2] = -\frac{k}{2.303}t + \log[H_2O_2]_o \quad \text{(change of base to 10)}$$

7.11 Half-Life Method

The **half-life** $t_{1/2}$ of a reaction is the time for half of a reactant to be consumed, or half of a product to be formed. Consider previous case of a first-order reaction with one reactant:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

Proceeding as before, we obtain

$$-\frac{d[H_2O_2]}{dt} = k[H_2O_2]$$
$$\ln[H_2O_2] = -kt + \ln[H_2O_2]_o$$
$$[H_2O_2]_t = [H_2O_2]_o e^{-kt}$$

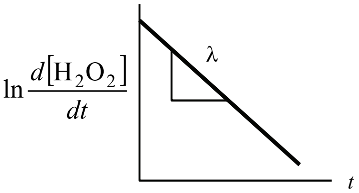

The half-life $t_{1/2}$ is the time for one half of the reactant to be consumed:

$$\frac{[H_2O_2]_t}{[H_2O_2]_o} = e^{-kt_{1/2}} = \frac{1}{2}$$
$$\ln\frac{1}{2} = -kt_{1/2}$$
$$t_{1/2} = \frac{0.69}{k}$$

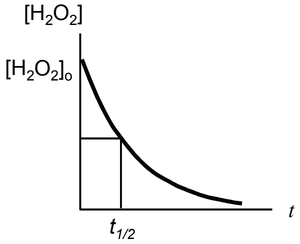

Either the rate constant *k* or the half-life (time measured for half the reactant to be consumed) can be determined from experimental data to give the other quantity. In a first-order reaction, the half-life is a constant and depends only on *k*.

7.12 Temperature

The rate of a chemical reaction depends on the frequency of collisions between atoms, ions or molecules, and also the fraction of which these particles have an energy greater than the activation energy. Increasing the concentration of reactants generally increases the frequency of collisions. Increasing the temperature also increases the collision frequency because the particles are moving faster and make more collisions per unit time. However, the most profound effect of an increase in temperature is to increase the fraction of particles with an energy above the activation energy.

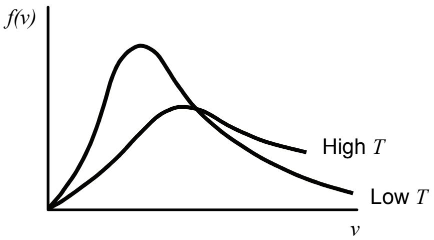

In gases, the distribution of velocities of molecules can be given by the Maxwell velocity distribution:

$$f(v) = 4\pi\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2 e^{-\frac{mv^2}{2kT}}.$$

At low temperature, a greater proportion of the molecules have a more narrow range of relatively low velocities compared to high temperature, where the range of velocities is more spread out, but shifted to higher values. Increasing the temperature therefore

- Shifts the molecular distribution of energies to a higher value
- Increases the proportion of molecules having higher energies

thus increasing the frequency and number of productive collisions per unit time and hence increasing the reaction rate.

The number of particles N with a total energy E was computed by Boltzmann using the **Maxwell velocity distribution** function:

$$N = Ce^{\frac{-E}{kT}}.$$

The rate constant for a specific reaction increases as the temperature increases, roughly in accordance with the **Arrhenius equation**:

$$k = Ae^{\frac{-E}{RT}},$$

where it can be seen that k is an exponential function of $1/T$. At room temperature, an increase of about 10°C results in an increase in reaction rate in the order of 2 to 3.

If E represents the activation energy, then this may be determined by experiment by plotting the logarithm of k versus $1/T$ over a range of temperatures.

7.13 Catalysts

The decomposition of hydrogen peroxide is a spontaneous, but slow reaction:

$$2H_2O_2 \rightarrow 2H_2O + O_2(g)$$

If powdered MnO_2 is added, the reaction proceeds much faster. Further, the MnO_2 molecules do not get used up during the reaction nor do they appear in the net equation. Substances which increase the rate of a reaction without appearing in the net reaction are called **catalysts** and the reaction is said to be **catalysed**.

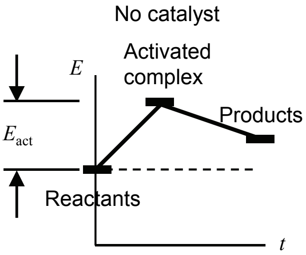

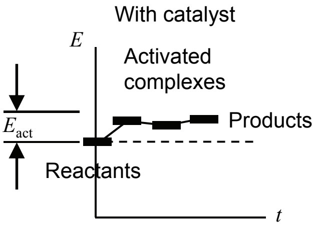

The role of a catalyst is to alter the reaction mechanism by insertion of additional steps so that the activation energy is reduced, allowing the reaction rate to increase.

While the catalyst may be consumed in one reaction step, and regenerated in another, it does not appear in the net reaction.

Catalysts may be solids, liquids or gases. When the reaction being catalysed is all of the same phase (e.g. all liquid), it is called **homogenous catalysis**. When the reaction occurs at the surface of a catalyst, such as over the platinum bed of a **catalytic converter** in an automobile exhaust system, it is called **heterogeneous catalysis**.

- A catalyst has no effect on the stoichiometry of the overall reaction.
- A catalyst does not affect the equilibrium position of the reaction. That is, the reaction proceeds to the same point of chemical equilibrium as it would without the catalyst. It just gets to equilibrium faster.
- A catalyst affects the reaction mechanism so as to lower the activation energy, thus increasing the number of molecules that can make a productive collision.

A negative catalyst decreases the rate of reaction and is called an **inhibitor**. The means by which this does so may not necessarily be to increase the activation energy, but is sometimes due to the inhibitor preventing another chemical present from acting as a catalyst. Auto catalysis, or **self-catalysis**, occurs when one of the products of a reaction acts as a catalyst for the reaction.

Enzymes act as catalysts in biological processes. Indeed, nearly all biological chemical reactions occur due to the action of catalysts.

8. Chemical Equilibrium

Summary

Law of chemical equilibrium:

$$K_c = \frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

Le Chatelier's principle:

> When a chemical system is at equilibrium, and that equilibrium is disturbed in some way, then the system will change in a way so as to counteract the disturbance until equilibrium is re-established.

Gaseous equilibrium:

$$K_p = \frac{K_c}{RT} = \frac{P_c}{P_a P_b}$$

Solid/solubility equilibrium:

$$K_{sp} = \left[\mathrm{A}^+(\mathrm{aq})\right]\left[\mathrm{B}^-(\mathrm{aq})\right]$$

Common ion effect:

> The presence of a common ion has reduced the solubility of a compound by several orders of magnitude.

8.1 Chemical Equilibrium

If a chemical reaction proceeds from reactants to products spontaneously, we may at first expect the process to continue until the reactants are all converted to product. This is not observed. Usually, reactions proceed until there is **chemical equilibrium**.

At equilibrium, the rate of forward reaction = rate of reverse reaction.

$$\text{Reactants} \leftrightarrow \text{Products}$$

As the reaction proceeds, the reactants are consumed and the products are created. The concentration of the reactants decreases (unless they are continuously replenished) and the concentration of the products increases. The rate of a chemical reaction depends on the concentration of the reactants. Therefore if the concentration of reactants is steadily decreasing, then the rate of conversion of reactants to products (in moles/litre/second) is also decreasing. At the same time, the rate of the conversion of products back to reactants is increasing because the concentration of the products is increasing. At chemical equilibrium, the rates of both forward and reverse reactions are equal.

$$A + B \leftrightarrow C + D$$

At equilibrium, we cannot tell from this equation how many moles of A and B exist, and how many moles of the product C or D exist.

The concentrations of A and B start off at some particular values, then decrease and level off with time. The concentrations of C and D start from zero and increase and level off with time. At equilibrium, the concentrations of A, B, C and D are all constant.

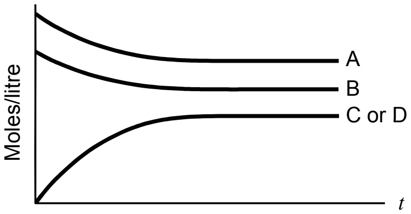

Although the balanced equation does not indicate the concentrations of reactants or products at any one time, they may be measured experimentally.

At equilibrium:

- The rate of formation of product C or D in moles per litre per second is equal to the rate of conversion of C and D back to molecules A and B.
- The concentrations of A, B and C, D are constant.

Chemical equilibrium is a **dynamic equilibrium**. Chemical activity still occurs at the equilibrium condition, but externally, we see no change in concentration of reactants or products.

It is important to know that **catalysts** do not alter the equilibrium state; they only serve to increase the rate of both forward and reverse reactions.

8.2 Law of Chemical Equilibrium

Consider the following single-step forward reaction:

$$\mathrm{A}+\mathrm{B}\rightarrow\mathrm{C}+\mathrm{D}$$

The rate law for this reaction is written:

$$\frac{d[\mathrm{C}+\mathrm{D}]}{dt}=k[\mathrm{A}][\mathrm{B}]$$

Now, consider the reverse reaction and its **rate law**:

$$\mathrm{C}+\mathrm{D}\rightarrow\mathrm{A}+\mathrm{B}$$

$$\frac{d[\mathrm{A}+\mathrm{B}]}{dt}=k'[\mathrm{C}][\mathrm{D}]$$

At equilibrium, the rate of forward reaction must equal the rate of the reverse reaction, and so:

$$k[\mathrm{A}][\mathrm{B}]=k'[\mathrm{C}][\mathrm{D}]$$

$$\frac{[\mathrm{C}][\mathrm{D}]}{[\mathrm{A}][\mathrm{B}]}=\frac{k}{k'}=K_c$$

That is, the ratio of the concentrations of the products over the reactants is a constant *at equilibrium*. A similar treatment can be written for more complex reaction steps.

More generally, it is found from experiment that for the general chemical equation $a\mathrm{A}+b\mathrm{B}\rightarrow c\mathrm{C}+d\mathrm{D}$ the **equilibrium constant** K_c is given by

$$K_c=\frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

law of chemical equilibrium

The magnitude of K_c indicates the extent of the reaction.

$K_c\leq10^{-5}$	Forward reaction is negligible
$10^{-5}\leq K_c\leq10^{5}$	Appreciable concentration of products and reactants
$K_c\geq10^{5}$	Reverse reaction is negligible and almost complete conversion of reactants to products

The value of the equilibrium constant depends on the temperature and the chemical species involved. However, it does not depend on the starting concentrations of either reactants or products. Whatever concentrations of A, B, C and D there are, when equilibrium is reached, the ratio of the concentrations expressed in this way will have the same value at that temperature. The value of *K* also depends on how the balanced equation is written (i.e. the choice of coefficients). The equilibrium constant must be associated with a particular balanced equation.

Note that the equilibrium constant does not provide any information on how long it takes for equilibrium to be established. It only provides information on the extent to which equilibrium favours the products over the reactants.

8.3 Equilibrium Constant

For any chemical reaction, the ratio of concentrations raised to the power of the coefficients can be written as:

$$\frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

This is more formally called the **mass-action expression**. It is only at chemical equilibrium that this expression becomes equal to a constant value K_c (the subscript c indicates concentrations in moles/litre).

K_c depends on the temperature. The units of K_c depend on the powers to which the concentrations are raised.

$$\mathrm{N_2O_4} \rightarrow 2\mathrm{NO_2(g)}$$

$$K_c = \frac{[\mathrm{NO_2}]^2}{[\mathrm{N_2O_4}]} = \mathrm{mol/L}$$

$$2\mathrm{CO}_{(g)} + \mathrm{O_2} \rightarrow 2\mathrm{CO_2}$$

$$K_c = \frac{[\mathrm{CO_2}]^2}{[\mathrm{CO}]^2[\mathrm{O_2}]} = \mathrm{L/mol}$$

A particular value of K_c depends on the form of the balanced equation.

$$\frac{1}{2}\mathrm{N_2O_4} \rightarrow \mathrm{NO_2(g)}$$

$$K_c' = \frac{[\mathrm{NO_2}]}{[\mathrm{N_2O_4}]^{1/2}} = \mathrm{mol}^{1/2}\ \mathrm{L}^{1/2}$$

$$= (K_c)^{1/2}$$

compare

Homogenous equilibrium occurs when all reactants and products are in the same physical state. **Heterogeneous equilibrium** occurs when reactants and products are in different physical states.

The concentrations of pure **solids** and **liquids** are constant. This is because if more *pure* solid or liquid is added, then the volume of that solid or liquid is also increased in the same proportion and so the number of moles per litre remains the same as before. The concentrations of these substances are removed from the equilibrium expression and are effectively incorporated into the value of K_c.

Gases cannot be removed from the equilibrium expression because their concentration changes during the reaction.

Writing a chemical equation in reverse inverts the value of K_c.

8.4 Le Chatelier's Principle

When a chemical system is at equilibrium, and that equilibrium is disturbed in some way, then the system will change in a way so as to counteract the disturbance until equilibrium is re-established. This is known as **Le Chatelier's principle**.

One way of disturbing the equilibrium condition is to add more of a reactant or a product. When more of a reactant is added, the forward reaction increases in rate (since reaction rate depends on the concentration) and more product is formed until the equilibrium condition, and the value of K_c, is restored. When more of product is added, then the concentration of the products increases, thus increasing the rate of the reverse reaction until the equilibrium condition is re-established. Similarly, if either a reactant or product is removed, then the forward or reverse reaction proceeds at a faster rate so as to restore the concentration of the removed species until the equilibrium condition is re-established.

Consider the effect of an increase in concentration of one reactant in the equilibrium expression:

$$N_2O_4 \rightarrow 2NO_2(g)$$

$$K_c = \frac{[NO_2]^2}{[N_2O_4]} = 0.87 \text{ at } 55^\circ\text{C}$$

Increasing the amount of N_2O_4 makes the denominator of the expression larger, and makes the computed value of K_c smaller. The concentration of NO_2 must increase until the original value of K_c is re-established. Note that at the new equilibrium, the concentrations of both products and reactants will have increased but their ratio, as expressed by the equilibrium expression, remains the same. Looked at another way, introducing more reactant tends to increase the rate of the forward reaction until the equilibrium constant is re-established back to its former level.

Another way of disturbing the equilibrium condition is to alter the temperature. The value of K_c at equilibrium depends very much on the temperature. In an exothermic reaction, increasing the temperature tends to favour the reverse reaction. A new equilibrium is reached with a new value of K_c.

8.5 Summary of Le Chatelier’s Principle

Summary of Le Chatelier’s Principle:

Change	**Effect on K_c**	**Shift in reaction**
Increase in concentration of one reactant	Constant	Forward reaction
Decrease in concentration of one reactant	Constant	Reverse reaction
Increase in concentration of one of the products	Constant	Reverse reaction
Decrease in concentration of one of the products	Constant	Forward reaction
Pressure increase with volume decrease	Constant	Decrease in number of gaseous molecules
Pressure decrease with volume increase	Constant	Increase in number of gaseous molecules
Temperature increase (exothermic reaction)	Decrease	Reverse reaction
Temperature increase (endothermic reaction)	Increase	Forward reaction
Temperature decrease (exothermic)	Increase	Forward reaction
Temperature decrease (endothermic)	Decrease	Reverse reaction
Catalyst added	Constant	No shift, but rate of both forward and reverse reactions increased

8.6 Equilibrium in Gaseous Systems

For reactions involving gases, the law of chemical equilibrium can be written in terms of **partial pressures**. The amount of gas in a mixture can be measured in terms of its partial pressure. The total of all the partial pressures is the total pressure in the system. Consider the reaction

$$\mathrm{A(g)} + \mathrm{B(g)} \rightarrow \mathrm{C(g)}$$

General gas equation: $PV = nRT$

Therefore, for A(g) $P_aV = nRT$ where P_a is the partial pressure of gas A in volume V

$$P_a = \frac{n}{V}RT$$

$$= [\mathrm{A}]RT$$ and similarly for B(g) and C(g).

Therefore,

$$K_c = \frac{[\mathrm{C}]}{[\mathrm{A}][\mathrm{B}]}$$

$$= RT\frac{P_c}{P_aP_b}$$

$$\frac{K_c}{RT} = \frac{P_c}{P_aP_b}$$ a constant.

$$= K_p$$

That is, a **gaseous equilibrium constant** K_p is the ratio of the partial pressures of the products over the partial pressures of the reactants.

Example: $N_2(g) + 3H_2(g) \rightarrow 2NH_3(g)$

$$K_p = \frac{P_{NH_3}^2}{(P_{N_2})(P_{H_2})^3}$$

$$K_c = \frac{[NH_3]^2}{[N_2][H_2]^3}$$

$$= \frac{(P_{NH_3}/RT)^2}{(P_{N_2}/RT)(H_2/RT)^3}$$

$$= \frac{P_{NH_3}^2}{(P_{N_2})(P_{H_2})^3}(RT)^2$$

$$= K_p(RT)^2$$

8.7 Solubility of Solids

Many chemical species dissolve in water. They can do so by a simple dissolution where the molecules just become separated from each other and are surrounded by water molecules, by dissociation into ions, and by chemical reaction with the water molecules.

Silver chloride, AgCl, has a low solubility in water. Those molecules that do dissolve, do so by forming ions. The **equilibrium constant** is therefore written as:

$$\mathrm{AgCl(s)} \leftrightarrow \mathrm{Ag^+(aq)} + \mathrm{Cl^-(aq)}$$

$$K_c = \frac{[\mathrm{Ag^+(aq)}][\mathrm{Cl^-(aq)}]}{[\mathrm{AgCl(s)}]}$$

Note that the chemical equation is written with the ions on the product side.

The concentration of a solid is a constant, and so is incorporated into the equilibrium constant to form a new constant, the **solubility product constant**, K_{sp}:

$$K_{sp} = [\mathrm{Ag^+(aq)}][\mathrm{Cl^-(aq)}] = 1.7\times10^{-10} \text{ at } 25^\circ\mathrm{C}$$

The **solubility product constant**, K_{sp}, applies to saturated aqueous solutions. The product of the ion concentrations is called the ion product. At equilibrium, the ion product equals K_{sp}. A low value of K_{sp} indicates that the concentration of ions in solution is low at equilibrium – in other words, the solubility of the compound is low.

The **solubility** of a compound in water is the number of moles of that compound that can be dissolved in 1 litre of water (at a specific temperature). The solubility of a compound can be obtained from the solubility product constant.

$$\begin{aligned} K_{sp} &= [\mathrm{Ag^+(aq)}][\mathrm{Cl^-(aq)}] \\ &= 1.7\times10^{-10} \text{ at } 25^\circ\mathrm{C} \\ s &= [\mathrm{Ag^+(aq)}] = [\mathrm{Cl^-(aq)}] \\ 1.7\times10^{-10} &= s^2 \\ s &= 1.3\times10^{-5}\,\mathrm{mol/L} \end{aligned}$$

From the balanced equation, 1 mole of AgCl will produce one mole of Ag^+ and 1 mole of Cl^- ions. Therefore in a 1 litre solution, there will be *s* moles of Ag^+ and *s* moles of Cl^- ions.

At equilibrium, the rate of **dissolution** is equal to the rate of **precipitation**.

In general, the solubility product constant is

$$\mathrm{M}_n\mathrm{X}_m \leftrightarrow n\mathrm{M}^{n+}\mathrm{(aq)} + m\mathrm{X}^{n-}\mathrm{(aq)}$$

$$K_{sp} = [\mathrm{M}^{n+}\mathrm{(aq)}]^n[\mathrm{X}^{n-}\mathrm{(aq)}]^m \quad \text{at solubility equilibrium}$$

Ionic solids (salts) and some molecular solids may dissociate into ions in aqueous solution.

8.8 Factors Affecting Equilibrium

A system tends towards a state of minimum **energy** and maximum **entropy**. The two tendencies can oppose or reinforce one another.

- If the forward reaction is exothermic $\Delta H < 0$ and the entropy change $\Delta S > 0$, then the forward reaction is favoured.
- If the forward reaction is endothermic $\Delta H > 0$ and the entropy change $\Delta S < 0$, then the reverse reaction is favoured.
- If the enthalpy change $\Delta H > 0$ and the entropy change $\Delta S > 0$, then the forward reaction is favoured at high temperature. When $\Delta H < 0$ and $\Delta S < 0$, the forward reaction is favoured at low temperature.

For a **solid** dissolving in a **liquid**, in saturated solution, the position of minimum energy can be on either side of the equation. The position of maximum entropy is on the dissolved side:

Solid ↔ Saturated solution
entropy gain →
← energy minimum →

For the dissolving of a **gas** in a **liquid**, entropy decreases as the gas dissolves and so the tendency is to oppose the dissolving process, however, when a gas dissolves, the gas molecules, being attracted to the liquid molecules, enter a state of lower potential energy and so this favours the dissolving process.

Gas ↔ Saturated solution
← entropy gain
energy minimum →

It is important to realise that the **equilibrium constant** is not the same thing as the **solubility product constant**. The solubility product constant is an experimentally derived value that applies to the solubility of a compound in **aqueous solution**. The equilibrium constant applies to any chemical reaction, not just those in aqueous (water) solution.

Many chemical reactions occur in dilute aqueous solutions. In these cases, for reactions that consume, or produce, water molecules, the concentration of the water is considered constant and is incorporated into the equilibrium constant. This is because the amount of water consumed or produced is usually a small fraction of the overall total amount of water present.

8.9 Common Ion Effect

The solubility of an ionic solid in a solution containing a common ion is greatly reduced. This is because the concentration of the common ion appears in the expression for the solubility product constant.

K_{sp} = 1.5×10^{-9} for $BaSO_4$ in water . The solubility of this salt in water can be calculated and compared with the solubility in 0.1M Na_2SO_4.

Solubility in water:

$$\begin{aligned} K_{sp} &= \left[\mathrm{Ba}^{2+}(\mathrm{aq})\right]\left[\mathrm{SO}_4^{2-}(\mathrm{aq})\right] \\ &= (s)(s) \\ s^2 &= 1.5\times10^{-9} \\ s &= 3.9\times10^{-5}\,\mathrm{mol/L} \end{aligned}$$

Solubility in 0.1M Na_2SO_4:

$$\begin{aligned} K_{sp} &= \left[\mathrm{Ba}^{2+}(\mathrm{aq})\right]\left[\mathrm{SO}_4^{2-}(\mathrm{aq})\right] \\ &= (s)(s+0.1) \\ &= 1.5\times10^{-9} \\ (s+0.1) &\approx 0.1 \quad \text{since } s \text{ will be} \ll 0.1 \\ 1.5\times10^{-9} &= s(0.1) \\ s &= 1.5\times10^{-8}\,\mathrm{mol/L} \end{aligned}$$

The presence of the **common ion** has reduced the solubility of this compound by several orders of magnitude.

8.10 Precipitation

The **solubility product constant** K_{sp} can be used to predict whether or not a solid precipitate will form as a result of a chemical reaction between solutions.

The product of the concentrations of the ions in a solution is called the **ion product**. When the ion product is greater than the solubility product constant, **precipitation** will occur until the concentration of the ion reduces to the extent that the ion product becomes equal to K_{sp}.

Consider the mixing of 1 litre of 0.5M $Ca(NO_3)_2$ and 1 litre of 0.2M Na_2SO_4. We begin by writing the ionic equation of interest:

$$CaSO_4 \leftrightarrow Ca^{2+} + SO_4^{2-}$$

Note that the equation has been written with the ions on the product side of the equation so that the ion product can be related to K_{sp} directly, and not the inverse.

The final volume of the mixed solution is 2 litres, and therefore the concentrations of the original solutions are:

$$\left[Ca^{2+}\right] = 0.25$$
$$\left[SO_4^{2-}\right] = 0.1$$
$$\begin{aligned} IP &= \left[Ca^{2+}(aq)\right]\left[SO_4^{2-}(aq)\right] \\ &= (0.25)(0.1) \\ &= 0.025 \end{aligned}$$

ion product

K_{sp} for $CaSO_4$ = 2.6×10^{-4}. Since the ion product is greater than K_{sp}, then $CaSO_4$ will form as a precipitate until the concentration of the Ca^{2+} and SO_4^{2-} ions fall until the ion product becomes equal to K_{sp}. That is, precipitation is a method of forcing out ions from the solution until the ion product becomes equal to K_{sp}, at which time the solid and the solution are in equilibrium. Precipitation has important economic consequences whereby the extraction of the desired product must be achieved as economically as possible. Precipitates are sometimes washed with a solution containing a **common ion** to further reduce their solubility.

8.11 Complex Ions

Some molecules or anions have a lone pair of electrons. This enables them to combine, or coordinate, with some transition metal ions to form a **complex ion**. For example, the molecule ammonia, NH_3, has a lone pair of electrons:

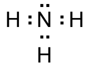

$$\begin{array}{c}\mathrm{H:\overset{\cdot\cdot}{\underset{\cdot\cdot}{N}}:H}\\ \mathrm{H}\end{array}$$

NH_3 can coordinate to Ag^+, Co^{3+}, Zn^{2+}, Ni^{2+} to form complex ions:

$$\mathrm{Ag(NH_3)_2^+, Co(NH_3)_4^{2+}, Zn(NH_3)_4^{2+}, Ni(NH_3)_6^{2+}}$$

The OH- ion also has a lone pair of electrons and can coordinate with Pb^{2+}, Zn^{2+}, Sn^{2+}, Al^{3+} :

$$\mathrm{Pb(OH)_4^{2-}, Zn(OH)_4^{2-}, Sn(OH)_4^{2-}, Al(OH)_4^{-}}$$

The **ligand** (the species with the lone pair) arranges itself around the metal ion and the electron pairs fill up vacant energy levels to create a noble gas configuration for the metal ion. The electron orbitals that are associated with isolated atoms are rearranged to give **hybrid orbitals** to accommodate the electron pairs. The overall charge on the complex ion remains unchanged.

The equilibrium condition between the reactants and products for complex ion formation is called the **stability constant** K_{stab}.

$$\mathrm{Ni^{2+} + 4CN^- \rightarrow Ni(CN)_4^{2-}}$$

Tetracyanonickelate

$$K_{stab} = \frac{\left[\mathrm{Ni(CN)_4^{2-}}\right]}{\left[\mathrm{Ni^{2+}}\right]\left[\mathrm{CN^-}\right]^4}$$

$$\mathrm{Pt^{2+} + 6NH_3 \rightarrow Pt(NH_3)_6^{2+}}$$

Hexaamineplatinum (II)

$$K_{stab} = \frac{\left[\mathrm{Pt(NH_3)_6^{2+}}\right]}{\left[\mathrm{Pt^{2+}}\right]\ \left[\mathrm{NH_3}\right]^6}$$

$$\mathrm{Zn^{2+} + 4OH^- \rightarrow Zn(OH)_4^{2-}}$$

Tetrahydroxozincate

$$K_{stab} = \frac{\left[\mathrm{Zn(OH)_4^{2-}}\right]}{\left[\mathrm{Zn^{2+}}\right]\left[\mathrm{OH^-}\right]^4}$$

9. Ionic Equilibrium

Summary

- Strong electrolytes dissociate completely into ions.
- Weak electrolytes only partially dissociate into ions.
- Non-electrolytes do not dissociate into ions when placed in solution.

Ion product of water:

$$K = \frac{\left[H^+\right]\left[OH^-\right]}{\left[H_2O\right]}$$

$$K_w = \left[H^+\right]\left[OH^-\right]$$

[H+] mol/L	**pH**	**pOH**
1	0	14
1×10^{-7}	7	7
1×10^{-14}	14	0

$$pH = -\log_{10}\left[H^+\right]$$

$$pOH = -\log_{10}\left[OH^-\right]$$

Acids – proton (H^+) donors

Bases – proton (H^+) acceptors

Acid equilibrium constant:

$$K_A = \frac{\left[H_3O^+\right]\left[A^-\right]}{\left[HA\right]}$$

Base equilibrium constant:

$$K_B = \frac{\left[BH^+\right]\left[OH^-\right]}{\left[B\right]}$$

Neutralisation: An acid on a base gives salt and water

Titration: The progressive addition of an acid to a base, or a base to an acid.

Hydrolysis: The chemical reaction of ions with water.

9.1 Electrolytes

When a substance dissolves, it may do so by simply separating from its neighbour molecules and become surrounded by water molecules, or the constituent molecules may break apart, or dissociate, and form negatively charged **anions** and positively charged **cations**. Such solutions conduct electricity by virtue of the movement of the ions under the influence of an applied electric field. Substances which dissolve to produce conducting solutions are called **electrolytes**. Substances which do not are called **non-electrolytes**.

- **Strong electrolytes** dissociate completely into ions.
- **Weak electrolytes** only partially dissociate into ions.
- **Non-electrolytes** do not dissociate into ions when placed in solution.

When an ionic substance dissociates into ions in an aqueous solution, the ions themselves are attracted to the polar water molecules and become **hydrated**. That is, a negative anion is surrounded by the positive (or H) end of water molecules and this hydrated ion can move around in solution under the influence of an electric field because as a whole, it is still negatively charged. Similarly, a positive cation is surrounded by the oxygen end of nearby water molecules and becomes hydrated, but the hydrated ion remains positively charged.

Both ionic and molecular compounds may be strong electrolytes. In the case of HCl, a molecular compound, we have

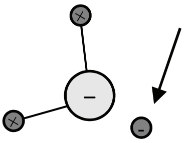

H^+ proton with positive charge showing one water molecule being attracted to it. Other water molecules may also be attracted to this proton.

$$\mathrm{HCl} + \mathrm{H_2O} \rightarrow \mathrm{H_3O^+} + \mathrm{Cl^-}$$

The **hydronium ion**, H_3O^+, is a hydrated H^+ ion. For convenience, we omit the water of hydration and just write

$$\mathrm{HCl} \rightarrow \mathrm{H^+} + \mathrm{Cl^-}$$

Examples of ionic compounds which are strong electrolytes are NaCl, $MgCl_2$, KBr, $CuSO_4$, $LiNO_3$. Examples of molecular covalent compounds which are strong electrolytes are HCl, HBr, HNO_3. Examples of weak electrolytes are $HC_2H_3O_2$, HNO_2, NH_3, HCN.

The degree to which dissociation occurs depends on energy changes (in particular, the energy of hydration) and entropy considerations as well as the nature of the solvent, the concentration of the solute, and the temperature. The dissociation of HCl is an example of **ionic equilibrium**.

9.2 Ionisation of Water

Experiments indicate that pure **water** is a very poor conductor of electricity, but it does conduct, indicating the presence of ions. Water is a weak **electrolyte**. Water dissociates into ions according to

$$\mathrm{H_2O} \leftrightarrow \mathrm{H^+} + \mathrm{OH^-}$$

The H^+ ions attach themselves to other water molecules to form the **hydronium** ion H_3O^+, but here we will simply write the H^+ ions as if they exist on their own.

The degree of dissociation is very small, and so the concentration of the H^+ and OH^- ions is also very small.

The concentration of water, expressed as moles/litre, can be readily established from its molecular weight and its density:

$$\begin{aligned}\text{m.w.}_{\mathrm{H_2O}} &= 16+2\\ &= 18\end{aligned}$$

One litre of water has a mass of 1000 g, therefore the number of moles *n* in a litre of water is:

$$\begin{aligned}n &= \frac{1000}{18}\\ &= 55.6\ \text{moles}\end{aligned}$$

That is, water has a concentration of 55.6M.

Since the concentration of the H^+ and OH^- ions in a litre of water is exceedingly small, the concentration of H_2O is virtually a constant, and so the **ion product** K_w is formed:

$$\begin{aligned}K &= \frac{[\mathrm{H^+}][\mathrm{OH^-}]}{[\mathrm{H_2O}]}\\ K_\mathrm{w} &= [\mathrm{H^+}][\mathrm{OH^-}]\end{aligned}$$

The concentration of H_2O (55.6M) is taken to be a constant and incorporated into K_w.

Experiments show that the **ion product of pure water** at 25°C is 1.0×10^{-14} mole^2/litre^2. The concentrations of H^+ and OH^- ions are 1.0×10^{-7} moles/litre respectively.

A solution containing equal concentrations of H^+ and OH^- ions at these concentrations are said to be **neutral**.

$$\begin{aligned}[\mathrm{H^+}] &= 1\times10^{-7}\ \text{mol/litre}\\ [\mathrm{OH^-}] &= 1\times10^{-7}\ \text{mol/litre}\end{aligned}$$

9.3 H^+ and OH^- Concentrations

Consider a 0.1M solution of HCl:

$$HCl \rightarrow H^+ + Cl^-$$

because HCl dissociates completely to H^+ and Cl^- ions.

The concentration of H^+ ions will be 0.1M and Cl^- ions 0.1M. We can ignore the small amount of H^+ ions from the dissociation of the water solvent. But, the equilibrium between the H^+ and the OH^- ions in the solution must still be established such that

$$\begin{aligned} 1\times10^{-14} &= \left[H^+\right]\left[OH^-\right] \\ &= 0.1\left[OH^-\right] \\ \left[OH^-\right] &= 1\times10^{-13}\,\text{mol/litre} \end{aligned}$$

That is, compared to neutral water, the concentration of the OH^- has been reduced from 1×10^{-7} to 1×10^{-13} due to a **common ion effect**.

Consider now an aqueous solution of 0.1M NaOH. Using the same calculation as above, we find that

$$NaOH \rightarrow Na^+ + OH^-$$

$$\left[H^+\right] = 1\times10^{-13}\,\text{mol/litre}$$

The added H^+ or OH^- ions suppress or inhibit the dissociation of the water molecules (in accordance with Le Chatelier).

Because the changes in concentration of these ions can cover such a large range of magnitudes, it is convenient to express them on a logarithmic scale – the "p" scale. The **pH** is the concentration of H^+ ions found from:

$$pH = -\log_{10}\left[H^+\right]$$

The **pOH** is the concentration of OH^- ions found from

$$pOH = -\log_{10}\left[OH^-\right]$$

The pH is a measure of the potential to combine with H^+ ions. A high pH means a high potential for H^+ combination.

Since the logarithm is the index to which the base must be raised to equal the number, the pH is the index of the H^+ concentration. For example, for a neutral solution,

$$\begin{aligned} pH &= -\log^{10}\left[H^+\right] = -\log_{10}\left(10^{-7}\right) \\ &= 7 \end{aligned}$$

It is easy to show that for a solution at equilibrium,

$$pH + pOH = 14$$

[H+] (mol/L)	**pH**	**pOH**
1	0	14
1×10^{-7}	7	7
1×10^{-14}	14	0

Say we have a 5M solution of HCl. What is the pH? $pH = -\log_{10} 5 = -0.7$

A negative pH is possible, but at high concentrations, the assumptions we make about the derivation of these equilibrium formulae breaks down.

9.4 Acids and Bases

A wide variety of substances can be generally classified into being either **acids**, **bases** or **salts**. Each of these has several distinctive characteristics when dissolved in water.

Acids	**Bases**
Electrical conductor	Electrical conductor
Sharp sensation on skin	Slippery feel
Sour taste	Bitter taste
Change the colour of litmus from blue to red	Change colour of litmus from red to blue.
React with some metals to form H_2 gas	
Dissolve carbonates to form CO_2 gas	

It was noticed that substances such as HCl, HNO_3, $HC_2H_3O_2$, H_2SO_4 and so on had the characteristics of acids. Substances such as NaOH, KOH, $Ca(OH)_2$, $Mg(OH)_2$ had the characteristics of bases. It was therefore postulated by **Arrhenius** (1859-1927) that acids are substances which dissolve in water to form H^+ ions, and bases are substances that dissolve in water to form OH^- ions:

Acid	Base
$\mathrm{HA} \rightarrow \mathrm{H}^+ + \mathrm{A}^-$	$\mathrm{BOH} \rightarrow \mathrm{B}^+ + \mathrm{OH}^-$

According to Arrhenius, all acids have the general formula HA and all bases have the general formula BOH.

A slightly more general definition of Arrhenius' scheme is that acids dissociate so as to increase the H^+ ions in water while bases dissociate and result in an increase in OH^- ions.

It was also known that when an acid was mixed with a base, the resulting solution lost its properties of either (except for being an electrical **conductor**) and formed a **salt** with the general form BA.

The degree of dissociation in water indicates the strength of the acid or the base:

$\mathrm{HCl} \rightarrow \mathrm{H}^+ + \mathrm{Cl}^-$ Strong acid (equilibrium favours products)

$\mathrm{HC_2H_3O_2} \rightarrow \mathrm{H}^+ + \mathrm{C_2H_3O_2^-}$ Weak acid (equilibrium favours reactants)

$\mathrm{NaOH} \rightarrow \mathrm{Na}^+ + \mathrm{OH}^-$ Strong base (equilibrium favours products)

$\mathrm{NH_3} + \mathrm{H_2O} \leftrightarrow \mathrm{NH_4^+} + \mathrm{OH}^-$ Weak base (equilibrium favours reactants)

$\mathrm{HCl} + \mathrm{NaOH} \rightarrow \mathrm{NaCl} + \mathrm{H_2O}$ acid + base = salt + water

The Arrhenius definition of acids and bases remains the simplest and most popular despite it having some limitations in only dealing with H^+ and OH^- ions as the cause of acidic or basic properties of substances.

9.5 Bronsted–Lowry

Reactions can take place in other solvents besides water. H^+ and OH^- ions may not be involved at all, yet the reactions still have the characteristics of acids and bases. The **Bronsted–Lowr**y definitions of acids and bases are

Acids – proton (H^+) donors
Bases – proton (H^+) acceptors

The Bronsted–Lowry classification is consistent with the Arrhenius definition of acids and bases but includes certain substances that act as bases which are not covered by the Arrhenius definition.

Consider the dissociation of HCl in water, and this time, we write the equation in full with the **hydronium ion**:

$$\mathrm{HCl + H_2O \rightarrow H_3O^+ + Cl^-}$$

Now, let's focus our attention on the water molecule. The water molecule accepts a H^+ ion, a proton, to become a hydronium ion. That is, the water molecule, according to Bronsted and Lowry, acts as a base. At equilibrium, the reverse reaction also occurs:

$$\mathrm{HCl + H_2O \leftarrow H_3O^+ + Cl^-}$$

Here, the hydronium ion donates a proton to the proceedings and the result is a water molecule. For all intents and purposes, the hydronium ion acts as an acid. In the Bronsted–Lowry scheme, the water molecule is the **conjugate base** of the hydronium ion. Or, the hydronium ion is the **conjugate acid** of the water molecule.

In the forward reaction, HCl donates a proton and becomes a Cl^- ion. HCl acts as an acid. In the reverse reaction, the Cl^- ion becomes HCl. That is, the Cl^- ion acts like a base, accepting a proton and becoming HCl. HCl is the **conjugate acid** of Cl^- and Cl^- is the **conjugate base** of HCl.

Every Bronsted–Lowry acid is accompanied by a Bronsted–Lowry base. The Bronsted–Lowry scheme has this extra dimension of creation of a **conjugate pair** compared to the Arrhenius theory:

$$\underset{\text{acid}}{\mathrm{HA}} + \underset{\text{base}}{\mathrm{H_2O}} \rightarrow \underset{\text{acid}}{\mathrm{H_3O^+}} + \underset{\text{base}}{\mathrm{A^-}}$$

In the forward reaction, HA is the acid and H_2O is the base. For the reverse reaction, H_3O^+ is the acid and A^- is the base.

The degree of dissociation indicates the strength of the acid or the base. A strong Bronsted–Lowry acid implies a weak conjugate base. One of the greatest strengths of this system is that acid-base reactions that do not involve the OH^- ion can be accommodated.

9.6 Strength of Acids

The strength of an acid is described by its tendency to dissociate into ions and produce H^+. The **pH** scale does not necessarily indicate whether a solution contains a strong acid (or base); it only indicates the concentration of H^+ (and OH^-) ions which depends on both the intrinsic strength of the acid and the molar concentration used to make the solution.

HCl is considered an intrinsically strong acid because it almost completely dissociates into H^+ and Cl^- ions:

$$\mathrm{HCl} \rightarrow \mathrm{H^+} + \mathrm{Cl^-}$$

The **equilibrium constant** K_{HCl} for this reaction is very large (≈10^7), favouring he products. Acetic acid is a weak acid:

$$\mathrm{HC_2H_3O_2} \leftrightarrow \mathrm{H^+} + \mathrm{C_2H_3O_2}^-$$

The equilibrium constant for this reaction is very small (about 1.8×10^{-5}), favouring the reactants.

where we always know that H^+ is actually hydrated

The general expression for the **acid equilibrium constant** is given as

$$\mathrm{HA} + \mathrm{H_2O} \leftrightarrow \mathrm{H_3O^+} + \mathrm{A^-}$$

$$K = \frac{\left[\mathrm{H_3O^+}\right]\left[\mathrm{A^-}\right]}{\left[\mathrm{HA}\right]\left[\mathrm{H_2O}\right]}$$

$$K_{\mathrm{A}} = \frac{\left[\mathrm{H_3O^+}\right]\left[\mathrm{A^-}\right]}{\left[\mathrm{HA}\right]}$$

since the concentration of H_2O is considered constant.

K_A provides a quantitative measure of the strength of an acid. That is, its tendency for dissociation to provide H^+ ions. pH is an indication of the concentration of H^+ ions in a particular solution. A highly concentrated solution of a weak acid may have a similar pH to a dilute solution of a strong acid.

Acid	**K_A**	**Conjugate base**
$HClO_4$	$\approx 10^{10}$	ClO_4^-
H_2SO_4	$\approx 10^{9}$	HSO_4^-
HCl	$\approx 10^{7}$	Cl^-
HSO_4	1.3×10^{-2}	SO_4^-
H_3PO_4	8×10^{-3}	$H_2PO_4^-$
$HC_2H_3O_2$	1.74×10^{-5}	$CH_3CO_2^-$
NH_4^+	6.3×10^{-10}	NH_3
NH_3	$\approx 10^{-30}$	NH_2^-

The strength of a base can be expressed in a similar manner by the **base equilibrium constant** K_B.

$$\mathrm{B} + \mathrm{H_2O} \leftrightarrow \mathrm{BH^+} + \mathrm{OH^-}$$

$$K_{\mathrm{B}} = \frac{\left[\mathrm{BH^+}\right]\left[\mathrm{OH^-}\right]}{\left[\mathrm{B}\right]}$$

These equilibrium constants can be expressed on a logarithmic scale:

$$\mathrm{p}K_{\mathrm{A}} = -\log_{10}\left[K_A\right]$$

$$\mathrm{p}K_{\mathrm{B}} = -\log_{10}\left[K_B\right]$$

For an acid and its conjugate base,

$$\mathrm{p}K_{\mathrm{A}} + \mathrm{p}K_{\mathrm{B}} = 14$$

9.7 Acid–Base Reactions

A few acids and bases are strong, but most are weak. Consider the dissociation of acetic acid $HC_2H_3O_2$, a weak acid with $K_A = 1.74 \times 10^{-5}$.

$$HC_2H_3O_2 + H_2O \leftrightarrow H_3O^+ + C_2H_3O_2^-$$

The concentration of H_3O^+ is found from:

$$K = \frac{\left[H_3O^+\right]\left[C_2H_3O_2^-\right]}{\left[HC_2H_3O_2\right]\left[H_2O\right]}$$

$$\left[H_3O^+\right] = 1.74 \times 10^{-5} \frac{\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]}$$

Strong acids:
HCl, HBr, HI, HNO_3,
H_2SO_4, $HClO_4$

If a solution is made with 0.1 moles of $HC_2H_3O_2$, then the concentration $[HC_2H_3O_2]$ is 0.1M. What then is the concentration of $[C_2H_3O_2^-]$? It must be small since the value of K_A implies a low incidence of dissociation. Now, let us add 0.1 moles of solid $NaC_2H_3O_2$. Since the salt, $NaC_2H_3O_2$ dissociates almost completely, the concentration of the acetate ion is now approximately 0.1M. Thus

$$\left[H_3O^+\right] = 1.74 \times 10^{-5} \frac{0.1}{0.1}$$

$$\mathrm{pH} = 4.75$$

If to this solution we now add a small amount of a strong base, say 0.001 moles (a drop) of 2M NaOH, then the NaOH reacts with the acetic acid molecules: $HC_2H_3O_2 + NaOH \leftrightarrow H_2O + NaC_2H_3O_2$

The quantity $[HC_2H_3O_2]$ is reduced from 0.1M to 0.099M. The quantity $[C_2H_3O_2^-]$ increases to 0.11M (since the $NaC_2H_3O_2$ dissociates). The pH of this new solution is

$$\mathrm{pH} = \log\left(1.74 \times 10^{-5} \frac{0.099}{0.11}\right) = 4.8$$

The $HC_2H_3O_2$ / $C_2H_3O_2^-$ solution is called a **buffer solution**, because of its ability to resist changes in pH. If a small amount of acid is added to the solution, this reacts with the acetate ion and the concentrations shift slightly the other way, and the pH remains almost unchanged.

If the concentration of the introduced acid or base becomes too large, then the **buffering capacity** of the solution is exhausted and the pH changes appreciably.

Note, without the common ion, addition of small amounts of acid or base can significantly alter the pH by several units.

9.8 Buffer Solutions

Buffer solutions have the important property of being able to maintain pH when disturbed by the addition of other acids or bases. Buffer solutions can be made in two ways:

1. Weak acid/salt of conjugate base (e.g. acetic acid and sodium acetate)

$$\mathrm{NaA_{(s)}} \rightarrow \mathrm{Na^+} + \mathrm{A^-}$$

$$\mathrm{HA} + \mathrm{H_2O} \leftrightarrow \mathrm{H_3O^+} + \mathrm{A^-}$$

$$\left[\mathrm{H_3O^+}\right] = K_A \frac{[\mathrm{HA}]}{\left[\mathrm{A^-}\right]}$$

$$\mathrm{pH} = -\log\left(K_A \frac{[\mathrm{HA}]}{\left[\mathrm{A^-}\right]}\right)$$

There are two equilibrium reactions going on here. One is the almost complete dissociation of the salt into its ions, and the other is the partial dissociation of the acid. The $[A^-]$ is dominated by the molarity of the salt solution. The other reaction is the acid equilibrium. Because of the presence of the common ion A^-, the equilibrium for this reaction lies on the reactants side of the equation and [HA] is therefore practically equal to the molarity of the acid.

Since [HA] is high (equilibrium to the left), any added OH^- from an introduced chemical will react with these ions:

$$\underset{\text{added}}{\mathrm{OH^-}} + \underset{\text{in buffer}}{\mathrm{HA}} \rightarrow \mathrm{A^-} + \mathrm{H_2O}$$

$$\underset{\text{added}}{\mathrm{H_3O^+}} + \underset{\text{in buffer}}{\mathrm{A^-}} \rightarrow \mathrm{HA} + \mathrm{H_2O}$$

The added H_3O^+ or OH^- ions will be prevented from affecting the pH by combing with either the HA or A^- ions already in the solution.

2. Weak base/salt of conjugate acid (e.g. ammonia/ammonium chloride)

$$\mathrm{BH^+X^-} \rightarrow \mathrm{BH^+} + \mathrm{X^-}$$

$$\mathrm{B} + \mathrm{H_2O} \leftrightarrow \mathrm{BH} + \mathrm{OH^-}$$

$$\left[\mathrm{OH^-}\right] = K_B \frac{[\mathrm{B}]}{\left[\mathrm{BH^+}\right]}$$

$$\mathrm{pOH} = -\log\left(K_B \frac{[\mathrm{B}]}{\left[\mathrm{BH^+}\right]}\right)$$

$$\mathrm{pH} = 14 - \mathrm{pOH}$$

Complete dissociation of the salt into its ions results in the $[BH^+]$ being dominated by the molarity of the salt solution. Because of the presence of the common ion BH^+, the equilibrium for the base reaction lies on the reactants side of the equation and [B] is therefore practically equal to the molarity of the base.

Any added H_3O^+ or OH^- ions will be prevented from affecting the pH by combing with either the B or BH^+ ions already in the solution.

9.9 Indicators

Indicators are either a weak acid or a weak base with the property of exhibiting a change in colour depending on the pH of the solution. The molecular form of the acid HA has one colour, and the ionic form, A^-, another colour.

$$\underset{\text{Colour 1}}{HA + H_2O} \leftrightarrow \underset{\text{Colour 2}}{H_3O^+ + A^-}$$

If a small amount of indicator is added to a solution containing a large concentration of H_3O^+, then the equilibrium for the indicator solution shifts towards the reactants and Colour 1 dominates. If the indicator is added to a solution of large concentration of OH^-, then the equilibrium shifts towards the products and Colour 2 dominates.

Typically, a few drops of indicator are added to a solution whose pH is to be determined, and the resulting colour of the solution indicates the pH when read off against a colour chart.

For the indicator:

$$HA + H_2O \leftrightarrow H_3O^+ + A^-$$

$$K_A = \frac{[H_3O^+][A^-]}{[HA]}$$

$$[H_3O^+] = K_A \frac{[HA]}{[A^-]}$$

Indicator	K_A	pH of colour change
Methyl orange	2×10^{-4}	3.7
Litmus	1×10^{-7}	7
Phenolphthalein	7×10^{-10}	9.1

An indicator changes colour when $[HA] = [A^-]$,

$$\therefore [H_3O^+] = K_A$$

Indicators are useful when they exist in a solution whose pH is slowly changing. If the pH, or the $[H_3O^+]$ of a solution changes so that [HA] becomes equal to $[A^-]$, and the $[H_3O^+]$ of the solution becomes equal to the K_A of the indicator and the colour changes.

It is important to know that an indicator does not necessarily change colour when $[H_3O^+] = [OH^-]$. Some indicators change colour on the basic side, while others change on the acidic side, and some at the neutral position, or **equivalence point**.

The **end point** occurs when the indicator changes colour. The reason why this is called the end point becomes clear when indicators are used in the process of **titration**.

9.10 Neutralisation

When an acid reacts with a base, the H^+ ions combine with the OH^- ions to form water, and the acidic and basic properties of the solution disappear along with the disappearing ions. This is **neutralisation**. That is, an acid on a base gives salt and water.

Consider 0.1M HCl added to 0.1M NaOH:

$$\mathrm{HCl} + \mathrm{NaOH} \rightarrow \mathrm{NaCl} + \mathrm{H_2O}$$

More specifically, in 0.1M solutions, we do not have HCl molecules and NaOH molecules. They are already dissociated into their constituent ions. The reaction is therefore actually

$$\mathrm{H_3O^+} + \mathrm{Cl^-} + \mathrm{Na^+} + \mathrm{OH^-} \rightarrow \mathrm{Na^+} + \mathrm{Cl^-} + 2\mathrm{H_2O}$$

The net reaction is therefore

$$\mathrm{H_3O^+} + \mathrm{OH^-} \rightarrow 2\mathrm{H_2O}$$

We know from experience that the products of this reaction are salt and water, or more specifically, Na^+ ions and Cl^- ions and water. That is, salty water, and not a mixture of acidic and basic solutions. But, how do we know this? Because in any aqueous solution at equilibrium, $K_w = 1 \times 10^{-14}$.

Equilibrium favours the products because the reverse reaction, the dissociation of water, has such a low **equilibrium constant** (or **ion product** in this case). The H_3O^+ ions combine with the OH^- ions to form water until such time as the product of the concentrations of H_3O^+ and OH^- ions fall to 1×10^{-14} moles/litre.

For strict neutralisation of a solution to occur, the initial concentrations of the H_3O^+ and OH^- ions must be similar – that is, the acid and the base must be of about the same strength. That is, if equal moles of acid and base are added, the final solution $[H_3O^+] = [OH^-] = 1 \times 10^{-7}$.

In a solution of acetic acid, very little dissociation occurs and so the reaction with a strong base in aqueous solution is:

$$\mathrm{HC_2H_3O_2} + \mathrm{Na^+} + \mathrm{OH^-} \rightarrow \mathrm{H_2O} + \mathrm{Na} + \mathrm{C_2H_3O_2^-}$$

$$\mathrm{HC_2H_3O_2} + \mathrm{OH^-} \rightarrow \mathrm{H_2O} + \mathrm{C_2H_3O_2^-}$$

The resulting solution is not neutral, but slightly basic in the sense that one of the products, $C_2H_3O_2{}^-$, is a Bronsted–Lowry base. In general,

$$\mathrm{HA} + \mathrm{OH^-} \rightarrow \mathrm{H_2O} + \mathrm{A^-}$$

An interesting application of neutralisation is the use of antacids for relief of heartburn. In order to raise the pH in the stomach, sodium bicarbonate can be used to react with HCl in the gastric juices to give H_2O and $CO_2(g)$.

9.11 Titration

The progressive addition of an acid to a base, or a base to an acid, is called **titration**. Typically, when a base of known concentration is added to an acid of unknown pH, we say that the acid is titrated with the base. A known number of moles of base is added to the acid, either in solid form, or more commonly, as a solution via a burette.

As the base is added, the concentration of $[H_3O^+]$ changes due to neutralisation. If the pH is monitored during the process, a titration curve can be plotted showing pH against moles of added base.

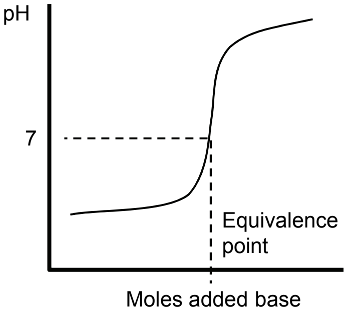

The moles of added base are measured usually by use of a burette. For the case of a strong acid and a strong base, there is a sharp change in pH at the equivalence point, where $[H_3O^+] = [OH^-]$. At or near this point, a very small addition of base results in a large increase in pH.

At the start of the titration, $[H_3O^+]$ is the molarity of the acid (for a strong acid). Before the equivalence point is reached, $[H_3O^+]$ is found from the initial moles of $[H_3O^+]$ less the moles of added OH^- (from the burette). At the equivalence point, $[H_3O^+] = [OH^-] = 1 \times 10^{-7}$. After the equivalence point, there is an excess of OH^-. Since we know how much OH^- is used to neutralise the acid, we know how much excess $[OH^-]$ exists after the equivalence point (from the burette reading) and the pH can be calculated from 14 – pOH.

The change in pH of the solution as the base is added can be measured using a pH meter, or more conventionally, using an **indicator**. When the indicator changes colour, the **end point**, the titration is finished.

9.12 Choice of Indicator

During a titration, the indicator to be used should have an **end point** which corresponds to the **equivalence point**. At first sight, this might sound like that the only useful indicator would be one that has an end point at pH = 7. However, the choice of indicator depends on the nature of the reactants.

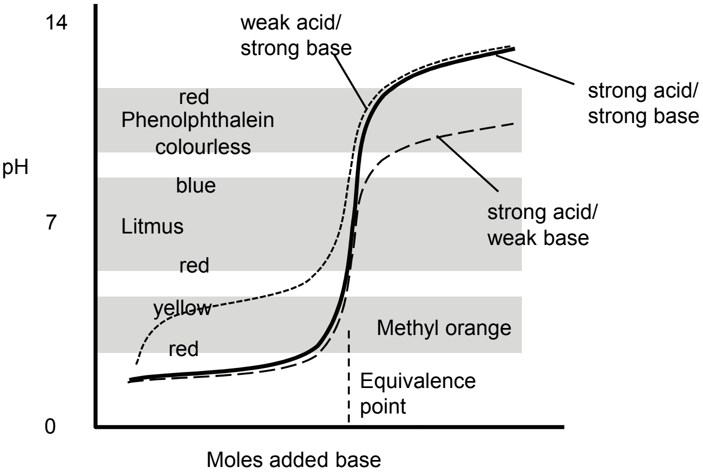

In the case of a strong acid being titrated with a strong acid, any of the three indicators above are suitable. However, for the case of a weak acid and a strong base, **Methyl orange** would not be suitable because it would change colour too early. **Phenolphthalein** would be a better choice because its end point (pH at which colour changes) more closely coincides with the equivalence point. For the case of a strong acid and a weak base, Methyl orange would be suitable, but not Phenolphthalein. **Litmus** is usually considered a poor choice because it changes colour over too wide a range of pH compared to other indicators. Weak acids are not usually used with weak bases in titration experiments.

A **universal indicator** is a mixture of other indicators designed so that different colours indicate different pH values.

In swimming pool maintenance, a few drops of phenol red (colour change from yellow to red over a pH range of 6.8 to 8.4) are used to measure pH. When the water is basic, the acid demand test involves putting a few drops of acid into the test mixture to determine how much acid is required to bring the solution back to the desired range of pH (usually 7.2 to 7.6).

9.13 Hydrolysis

Hydrolysis is the chemical reaction of ions with water. **Anions** may function as Bronsted-Lowry bases and receive a proton from the water molecule:

$$B^{n-} + H_2O \leftrightarrow OH^- + BH^{n-1}$$

Many elements occur in compounds as oxides and hydroxides. When hydrolysed, water molecules attach themselves to anions or cations formed by their dissociation in solution and these hydrolysed ions may act as acids or bases. The extent to which this occurs depends on the size of the atom and the charge.

Halide ions do not **hydrolyse** to any appreciable extent:

$$Cl^- + H_2O \leftarrow HCl + OH^- \qquad pH \approx 7$$

Oxide and sulphide ions hydrolyse:

$$O_2^{2-} + H_2O \rightarrow 2OH^- \qquad pH < 7$$

$$S^{2-} + H_2O \rightarrow HS^- + 2OH^-$$

Weak conjugate bases of strong acids do not hydrolyse:

$$ClO_4^- + H_2O \leftarrow HClO_4 + OH^- \qquad pH \approx 7$$

Strong conjugate bases of weak acids do hydrolyse:

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^- \qquad pH < 7$$

Cations may function as Bronsted-Lowry acids and donate proton to the water molecule:

$$A^{n+} + H_2O \leftrightarrow H_3O^+ + A^{n-1}$$

Hydrated cations of groups I and II do not hydrolyse. Hydrated transition metal ions (Al_3^+, NH_4^+) hydrolyse to give acid solutions.

$$[Fe(H_2O)_6]^{3+} + H_2O \rightarrow H_3O^+ + [Fe(H_2O)_5OH]^{2+}$$

The **salts** of strong acids to not hydrolyse. For example, NaCl dissolved in water produces

$$NaCl \rightarrow Na^+ + Cl^-$$

$$Na^+ + H_2O \rightarrow \text{no reaction}$$

$$pH = 7$$

$$Cl^- + H_2O \rightarrow \text{no reaction}$$

The salts of weak acids and strong bases do hydrolyse. For example, $NaC_2H_3O_2$ (sodium acetate) when dissolved in water produces

$$NaC_2H_3O_2 \rightarrow Na^+ + C_2H_3O_2{}^-$$

$$Na^+ + H_2O \rightarrow \text{no reaction}$$

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^- \qquad pH > 7$$

The salts of strong acids and weak bases do hydrolyse:

$$NH_4Cl \rightarrow NH^{4+} + Cl^-$$

$$Cl^- + H_2O \rightarrow \text{no reaction}$$

$$NH^{4+} + H_2O \rightarrow NH_3 + H_3O^+ \qquad pH < 7$$

The salts of weak acids and weak bases do hydrolyse but the resulting pH depends on the solubility constants K_A and K_B.

9.14 Simultaneous Equilibria

In the ideal case, only one equilibrium constant need be satisfied in a chemical reaction, but in practice, several equilibria are operating at the same time. For example, consider the hydrolysis reaction involving the dissociation of $(NH_4)_2S$ in solution:

$$NH_4^+ + H_2O \leftrightarrow NH_3 + H_3O^+$$
$$S^{2-} + H_2O \leftrightarrow HS^- + OH^-$$

There are two equilibria to be satisfied:

$$\begin{aligned} pK_{A\left(NH_4^+\right)} &= 9.2 \\ pK_{B\left(S^{2-}\right)} &= 0.2 \\ pK_{B\left(S^{2-}\right)} &> pK_{B\left(NH_4^+\right)} \\ &\therefore pH > 7 \end{aligned}$$

The quantitative analysis of hydrolysis involves the simultaneous equilibria of two reactions. For example consider the hydrolysis of the acetate ion:

$$C_2H_3O_2{}^- + H_2O \leftrightarrow HC_2H_3O_2 + OH^-$$

$$\begin{aligned} K_w &= \left[H_3O^+\right]\left[OH^-\right] \\ &= 1\times10^{-14} \\ K_h &= \frac{\left[OH^-\right]\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]} \end{aligned}$$

the **hydrolysis equilibrium constant**

$$\begin{aligned} \left[OH^-\right] &= \frac{K_w}{\left[H_3O^+\right]} \\ K_h &= \frac{K_w\left[HC_2H_3O_2\right]}{\left[C_2H_3O_2^-\right]} \\ &= \frac{K_w}{K_A} \end{aligned}$$

The hydrolysis equilibrium constant is expressed in terms of the associated acid (or base) equilibrium constant.

9.15 Complex Ions

Ligands are molecules or anions that have a lone pair of electrons. NH_3, H_2O and CN^- are examples. Some ligands combine with transition elemental metal ions – they become coordinated with the ion to form a **complex ion**.

$$[Cu(NH_3)_4]^{2+}$$
$$[Al(H_2O)_6]^{3+}$$
$$[Ag(CN)_2]^-$$

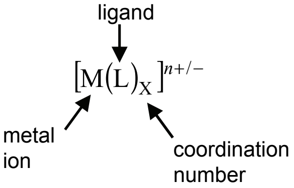

The **coordination number** usually equals twice the charge on the metal ion. The overall complex ion charge equals the sum of the cation and ligand charges. Complex ions may be anionic, cationic or neutral.

The properties of complex ions depend very much on the geometrical arrangement of molecules around the central atom. In general terms, the ligand provides electron pairs which find a place to attach themselves at available electron *d* orbitals around the central metal ion. However, this can happen in several ways. For example, consider the complex ion formed by Co and the ligands Cl^- and H_2O:

$[Co(Cl)_4]^{2-}$ coordination number 4

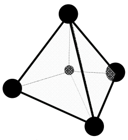

$[Co(H_2O)_6]^{2+}$ coordination number 6

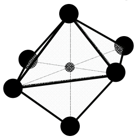

The lone pairs on the ligands face inwards towards and occupy *d* orbitals on the central metal ion. The reaction in water is

$$\underset{\text{blue}}{[Co(Cl)_4]^{2-}} + 6H_2O \leftrightarrow \underset{\text{pink}}{[Co(H_2O)_6]^{2+}} + 4Cl^-$$

Colours are often prominent in these compounds because the energy levels associated within the *d* orbitals are in the **visible light** range. Certain wavelength components of white light passing through solutions containing these ions are absorbed and the transmitted light that we see appears coloured.

10. Electronic Equilibrium

Summary

- **Oxidation** occurs at the **anode**. Neutral atoms being oxidised **lose** their electrons and become positively charged **cations** which tend to move towards the **cathode**.

$$4Fe(s) \rightarrow 2Fe_2^{2+}(aq) + 4e^-$$
$$Cu(s) \rightarrow Cu^{2+}(aq) + 2e^-$$
$$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^-$$
$$H_2^{2+}(g) \rightarrow 2H^+ + 2e^-$$

Oxidation occurs when the oxidation number of an atom is increased (loss of electrons).

- **Reduction** occurs at the **cathode**. Positively charged ions in solution take up the electrons coming from the oxidation reaction and become neutral atoms. The remaining negatively charged **anions** tend to move towards the **anode**.

$$3O_2 + 4e^- \rightarrow 2O_3^{2-}$$
$$2Ag^{2+} + 2e^- \rightarrow 2Ag(s)$$
$$2H^+ + 2e^- \rightarrow 2H_2(g)$$
$$Cu^{2+} + 2e^- \rightarrow Cu(s)$$

Reduction occurs when the oxidation number of an atom is decreased (gain of electrons).

10.1 Oxidation and Reduction

Oxidation and reduction reactions involve the **equilibrium** of a competition for electrons between atoms and ions. It is **electronic equilibrium**. To see how this comes about, we need to look at some numbers that quantify what drives this type of reaction.

The **oxidation number**, or **oxidation state**, is a quantity that describes the number of electrons that seem to be gained or lost by an atom when a chemical bond is formed. During a chemical reaction, **oxidation** occurs when the oxidation number of an atom is increased (loss of electrons). **Reduction** occurs when the oxidation number of an atom is decreased (gain of electrons). That is, the term reduction indicates a reduction in oxidation number.

Consider a simplified version of the oxidation of iron:

$$4\mathrm{Fe}(\mathrm{s})+3\mathrm{O}_2(\mathrm{g})\rightarrow 2\mathrm{Fe}_2\mathrm{O}_3(\mathrm{s})$$

The **oxidation half-reaction** is

$$4\mathrm{Fe}(\mathrm{s})\rightarrow 2\mathrm{Fe}_2^{2+}+4\mathrm{e}^-$$

In this half-reaction, the oxidation number of the Fe atoms increases from 0 to 2+ (loses electrons)

The **reduction half-reaction** is

$$3\mathrm{O}_2+4\mathrm{e}^-\rightarrow 2\mathrm{O}_3^{2-}$$

In this half-reaction, the oxidation number of the O atoms decreases from 0 to 2− (gains electrons).

The terms oxidation and reduction are historical in origin. They arose from the common observation that some reactants combine with oxygen (and so are said to be **oxidised**) while others lose oxygen atoms (and so are said to be **reduced**). Now it is recognised that it is the loss and gain of electrons that is the important feature of such reactions.

The **oxidising agent**, or **oxidant**, is the substance that is reduced. In the example above, oxygen is the oxidising agent since oxygen oxidises the iron. The **reducing agent**, or the **reductant**, is the substance that is oxidised. A **redox couple** is an oxidant and its complementary reductant. In the above example, the redox couples are written

$\mathrm{Fe}/\mathrm{Fe}_2^{2+}$ and $\mathrm{O}_2/\mathrm{O}_3^{2-}$

Oxidant/Reductant

(or Reduction/Oxidation)

The processes of oxidation and reduction do not occur in all chemical reactions. In many reactions, the oxidation numbers of the participating atoms do not change. Examples of **non-redox reactions** are some precipitation reactions, acids reacting with bases, acids and oxides, hydroxides, sulphides and carbonates. In **acid-base reactions**, it is more an exchange of protons that is important, not electrons.

10.2 Redox Reaction

Consider the redox reaction: $\mathrm{Cu(s)+2Ag^{2+}(aq)\rightarrow Cu^{2+}(aq)+2Ag(s)}$

We can separate out the oxidation and reduction processes physically using separate beakers of solution, a wire for electrons to flow along, and a salt bridge for ions to move across and electrodes where reactions take place.

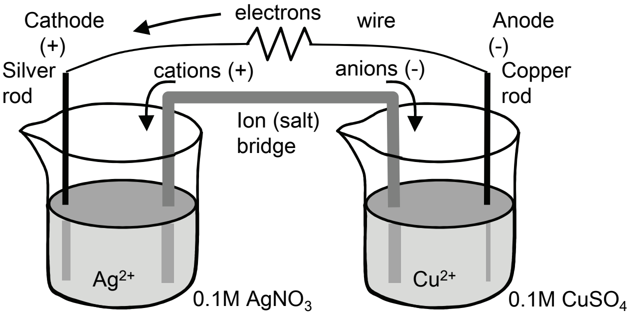

$\mathrm{2Ag^{2+}(aq)+2e^- \rightarrow 2Ag(s)}$

Reduction half-reaction

$\mathrm{Cu(s)\rightarrow Cu^{2+}(aq)+2e^-}$

Oxidation half-reaction

The salt bridge is a tube containing a salt solution (e.g. $NaNO_3$) and allows negatively charged anions to move into the anode solution and positively charged cations to move into the cathode solution while at the same time preventing bulk mixing of the solutions.

The electrode where reduction occurs is called the **cathode**. Silver has a greater affinity for **electrons** than copper. So, the Ag^{2+} ions in the solution combine with the free electrons in the cathode to form solid silver. This loss of free electrons at the cathode represents a net positive charge. The Ag^{2+} ions are pulling the electrons off the silver rod and forming solid silver at the cathode. But the consumption of Ag^{2+} ions in the solution results in a build-up of negative charge (the NO_3^- ions), and if this is not balanced in some way, the reaction will come to a stop as the ions in the solution repel any more electrons arriving at the cathode wanting to take part in the reaction.

The electrode where oxidation occurs is called the **anode.** The copper rod at the anode dissolves away to form Cu^{2+} ions and free **electrons**. These free electrons travel along the wire and are consumed at the cathode. The build-up of positively charge Cu^{2+} ions would eventually overcome the affinity for electrons by the silver and the reaction would come to a stop.

If the NO_3^- **anions** are allowed to flow from the cathode solution to the anode solution, and the Cu^{2+} **cations** flow from the anode solution to the cathode solution, then build-up of excess charge would be prevented and the reactions at each electrode would proceed spontaneously.

The redox reactions, arranged in this way, constitute an **electrochemical cell** (or **galvanic cell**) since a **voltage** is produced between the cathode and the anode. The sign of the anode and cathode is determined by the flow of electrons in the external wire, not the ion flow in the solutions.

10.3 Single-Cell Redox Reaction

Consider the redox reaction

$$\mathrm{Cu(s)} + 2\mathrm{Ag}^{2+}\mathrm{(aq)} \rightarrow \mathrm{Cu}^{2+}\mathrm{(aq)} + 2\mathrm{Ag(s)}$$

We have seen that this is an oxidation-reduction reaction whereby each half-reaction can be arranged to occur in separate solutions. This reaction can also occur in a single solution.

$$2\mathrm{Ag}^{2+}\mathrm{(aq)} + 2\mathrm{e}^{-} \rightarrow 2\mathrm{Ag(s)}$$

Reduction half-reaction

$$\mathrm{Cu(s)} \rightarrow \mathrm{Cu}^{2+}\mathrm{(aq)} + 2\mathrm{e}^{-}$$

Oxidation half-reaction

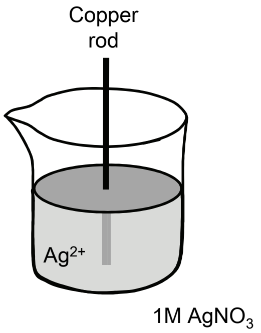

Redox reactions of this kind find application in the silvering of copper conductors in electrical switch gear whereby the oxidation (dulling) of copper due to atmospheric contaminants is to be avoided.

When a copper rod is dipped into a silver nitrate solution, silver metal is spontaneously deposited on the rod. The solution turns blue. Electron transfer occurs from the copper to the silver ions in the vicinity of the rod.

Another good example of a single-cell reaction is when a zinc rod is dipped into hydrochloric acid.

$$2\mathrm{H}^{+} + 2\mathrm{e}^{-} \rightarrow 2\mathrm{H}_2\mathrm{(g)}$$

Reduction half-reaction

$$\mathrm{Zn(s)} \rightarrow \mathrm{Zn}^{2+}\mathrm{(aq)} + 2\mathrm{e}^{-}$$

Oxidation half-reaction

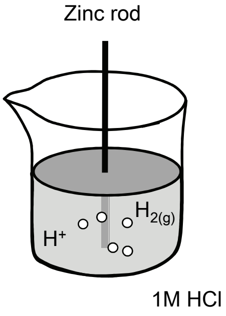

The metal dissolves in the acid, producing H_2 gas. The same thing happens with Mg, Al, Fe and Ni, but not with Cu, Ag and Au. The reason for some metals being able to be dissolved in HCl and others not can be found by a consideration of **standard electrode potentials**.

10.4 Standard Hydrogen Electrode

Consider the redox reaction

$$Cu^{2+}(aq) + H_2(g) \rightarrow Cu(s) + 2H^+(aq)$$

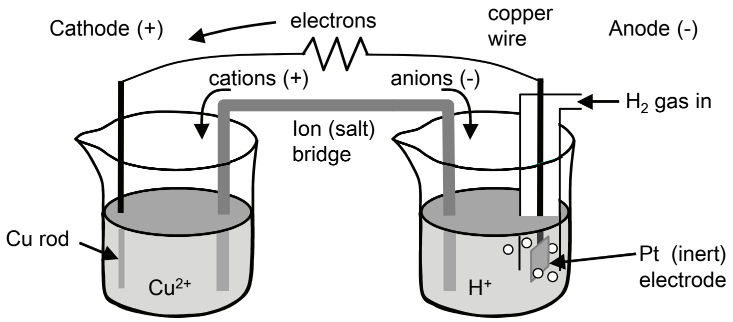

$$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s) \qquad H_2^{2+}(g) \rightarrow 2H^+ + 2e^-$$

Reduction half-reaction Oxidation half-reaction

Cu^{2+}/Cu Redox couple $H^+/H_2(g)$ Redox couple

When the H_2 gas is at 1 atm and the H^+ solution is 1 M, the H^+/H_2 redox couple is referred to as a **standard hydrogen electrode** and is assigned a 0V potential. The potential difference between the hydrogen standard cell and the other electrode is called the **standard electrode potential** E^o for the half-reaction occurring at that other electrode.

Standard electrode potentials are provided from the point of view of **oxidation** at the hydrogen standard electrode. That is, positive values of standard electrode potentials for a redox reaction indicate that the reaction will be one of **reduction** when the other half-cell is the standard cell. When the standard electrode potential for a redox reaction is listed as negative, it indicates that this reaction will be an **oxidation** reaction if the other electrode is the standard hydrogen electrode.

The standard electrode potential E^o for the Cu^{2+}/Cu half-reaction above is +0.34V and so the copper is reduced and the hydrogen is oxidised. The value of E^o for Zn is –0.76V and so if a zinc electrode and solution were used in place of the copper, the zinc would be oxidised and the hydrogen would be reduced.

10.5 Standard Electrode Potentials

Standard electrode potentials E^{o} for some **reduction half-reactions** where gases are at 1 atm and solutions are 1M:

$F_2(g) + 2e^- \rightarrow 2F^-$	+2.87
$O_3(g) + 2H^+(aq) + 2e^- \rightarrow 2H_2O$	+2.07
$MnO_4^{2+}(aq) + 8H^+(aq) + 5e^- \rightarrow Mn^{2-}(aq) + 4H_2O$	+1.51
$Cl_2(g) + 2e^- \rightarrow 2Cl^-(aq)$	+1.36
$O_2(g) + 4H^+(aq) + 4e^- \rightarrow 2H_2O$	+1.229
$Ag^+(aq) + e^- \rightarrow Ag(s)$	+0.799
$Fe^{3+}(aq) + e^- \rightarrow Fe^{2+}(aq)$	+0.771
$Cu^+(aq) + e^- \rightarrow Cu(s)$	+0.52
$Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$	+0.337
$Sn^{4+}(aq) + 2e^- \rightarrow Sn^{2+}(aq)$	+0.15
$S(aq) + 2H^+ + 2e^- \rightarrow H_2S(g)$	+0.141
$2H^+(aq) + 2e^- \rightarrow H_2^{2+}(g)$	0
$Pb^{2+}(aq) + 2e^- \rightarrow Pb(s)$	-0.126
$Sn^{2+}(aq) + 2e^- \rightarrow Sn(s)$	-0.136
$Fe^{2+}(aq) + 2e^- \rightarrow Fe(s)$	-0.440
$Zn^{2+}(aq) + 2e^- \rightarrow Zn(s)$	-0.763
$Al^{3+}(aq) + 3e^- \rightarrow Al(s)$	-1.66
$Mg^{2+}(aq) + 2e^- \rightarrow Mg(s)$	-2.37
$Na^+(aq) + e^- \rightarrow Na(s)$	-2.714
$Ca^{2+}(aq) + 2e^- \rightarrow Ca(s)$	-2.89
$K^+(aq) + e^- \rightarrow K(s)$	-2.925
$Li^+(aq) + e^- \rightarrow Li(s)$	-3.045

10.6 Spontaneous Redox Reactions

A very useful application of standard electrode potentials is the determination of whether a redox reaction will proceed spontaneously or not. Consider the copper and zinc redox couples:

$$\mathrm{Cu^{2+}/Cu} \qquad \mathrm{Zn^{2+}/Zn}$$

$$E^{\circ} = +0.34\ \mathrm{V} \qquad E^{\circ} = -0.76\ \mathrm{V}$$

There are four possible reactants and two possible reactions:

1 $\mathrm{Cu}^{2+}_{(ox)} + \mathrm{Zn}^{2+}_{(ox)}$ No reaction (simultaneous oxidation)

2 $\mathrm{Cu}^{2+}_{(ox)} + \mathrm{Zn}_{(red)}$ $\mathrm{Cu^{2+}(aq) + Zn(s) \rightarrow Cu(s) + Zn^{2+}(aq)}$

3 $\mathrm{Cu}_{(red)} + \mathrm{Zn}^{2+}_{(ox)}$ $\mathrm{Cu}_{(red)} + \mathrm{Zn}^{2+}_{(ox)} \rightarrow \mathrm{Cu^{2+}(aq) + Zn(s)}$

4 $\mathrm{Cu}_{(red)} + \mathrm{Zn}_{(red)}$ No reaction (simultaneous reduction)

The first and fourth possibilities cannot happen since we must have complementary oxidation and reduction processes happening. To determine which of the second and third possibilities will happen spontaneously when these chemicals are in an **electrochemical cell**, we add the **standard electrode potentials**. However, the standard electrode potentials are expressed in terms of reduction reactions, and so we must reverse the sign for a corresponding oxidation reaction.

For the second possibility, the electrode potential for the oxidation of copper in solution would be –0.34V. This is added to –0.76 for the reduction of zinc to obtain –1.1V. Since this total is a negative voltage for the cell as a whole, we say that this reaction will not proceed of its own accord.

For the third possibility, the electrode potential for the reduction of copper in solution is +0.34V. This is added to +0.76 for the oxidation of zinc to obtain +1.1V. Since this total is a positive value, we say that this reaction will proceed of its own accord and what's more, we would measure a **potential difference** of +1.1V across the terminals of the cell.

Standard electrode potentials for a redox reaction indicate the **reduction** potential of the half-reaction. Generally speaking, a redox reaction will proceed spontaneously if the **oxidant** (the one that is reduced) from the redox couple with the higher value of E° reacts with the **reductant** of the other.

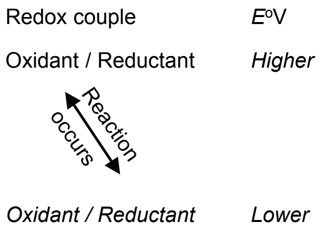

10.7 Oxidation Numbers

Although we can predict which redox reactions will occur spontaneously from a consideration of **standard electrode potentials**, we have yet to consider why the standard potentials are what they are. Consider the redox reaction

$$\mathrm{Cu(s) + 2Ag^{2+}(aq) \rightarrow Cu^{2+}(aq) + 2Ag(s)} \quad \text{Redox reaction}$$

$$\mathrm{2Ag^{2+}(aq) + 2e^{-} \rightarrow 2Ag(s)} \quad \text{Reduction half-reaction}$$

$$\mathrm{Cu(s) \rightarrow Cu^{2+}(aq) + 2e^{-}} \quad \text{Oxidation half-reaction}$$

From the above, we can conclude that silver ions have a stronger affinity for attracting **electrons** than do copper ions, and so the silver ions are reduced to a solid by taking up electrons and the solid copper is oxidised by releasing electrons, rather than the reverse. It is the relative affinity of ions to attract electrons that places a particular element in the list of standard potentials – which as will be remembered is presented from the point of view of reduction half-reactions.

This affinity for electrons by atoms is something we have met before: **electronegativity**. It was found that a practical measure of electronegativity is the method of electron counting by **oxidation numbers**. In a chemical reaction, when the oxidation number of an atom (i.e. its apparent charge) has increased, the atom is said to have been **oxidised**. When the oxidation number of an atom has decreased, the atom is said to have been **reduced**.

In the above reactions, the oxidation number of copper is initially 0. After reaction, the copper atoms have apparently lost two electrons and so its oxidation number is now 2+ (its apparent charge), and is thus an increase. The oxidation number of silver in the $\mathrm{Ag^{2+}}$ ions is initially 2+, and after reaction is 0. The oxidation number has decreased and the silver is therefore reduced. All this comes down to a competition for the attraction of electrons.

- **Oxidation** represents a loss of electrons or increase in oxidation number.
- **Reduction** represents a gain of electrons or decrease in oxidation number.

Oxidation numbers allow us to recognise oxidation-reduction reactions and are also useful for balancing the equations for these reactions. Changes in oxidation number do not necessarily indicate that an atom has gained or lost an electron, or has gained or lost net electric charge, even though we use them in this way for the purpose of establishing the numbers. Rather, they indicate how many electrons are involved for the atom in a chemical reaction.

10.8 Balancing Redox Half-Reactions

In redox reactions, both atoms and charges must be conserved. A convenient method of balancing redox equations is to use the equations for the half-reactions.

Consider the reaction between copper and nitric acid, HNO_3. The redox couples involved are

$Cu^{2+}/Cu(s)$

$E^{o} = +0.34V$

Since E^{o} for the nitric acid is more positive, then the copper must undergo oxidation:

$Cu(s) \rightarrow Cu^{2+}(aq) + 2e^{-}$

$H^{+}, NO_3^{-}/NO(g)$

$E^{o} = +0.96V$

and the NO_3 must undergo reduction:

$4H^{+} + NO_3^{-} + 3e^{-} \rightarrow NO(g) + H_2O$

Both these half-reactions are balanced, but they now need to be combined to form the net reaction, which is required also to be balanced. The procedure is to balance the charge. The lowest common factor in this example is 6, and so

$3Cu(s) \rightarrow 3Cu^{2+}(aq) + 6e^{-}$ $\quad$ $8H^{+} + 2NO_3^{-} + 6e^{-} \rightarrow 2NO(g) + 4H_2O$

These half-reactions can now be combined:

$$3Cu + 8H^{+} + 2NO_3^{-} \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O$$

Not all reactions can be balanced so easily. A general method is to

1. Balance for atoms other than O and H
2. Balance for O by adding water
3. Balance for H by adding H^{+}
4. Add OH^{-} ions to both sides for basic solution to form water with the H^{+} added from (3)
5. Balance for charge by adding electrons

It should be noted that balancing a half-reaction by adjusting coefficients does not alter the electrode potential for that half-reaction.

10.9 Balancing Redox Reactions with Oxidation Numbers

Another method of balancing redox reactions is by the use of oxidation numbers. Consider the reaction between copper and nitric acid HNO_3. The redox couples involved are:

$$\mathrm{Cu^{2+}/Cu(s)} \qquad \mathrm{NO_3^-/NO(g)}$$

$$E^{o} = +0.34\ \mathrm{V} \qquad E^{o} = +0.96\ \mathrm{V}$$

Since E^{o} for the nitric acid is more positive, then the copper must undergo oxidation. The unbalanced equation is:

$$\mathrm{Cu + NO_3^- \rightarrow Cu^{2+} + NO(g)}$$

Next, we write the oxidation numbers to each element in the reaction:

$$\begin{array}{l} \mathrm{Cu + NO_3^- \rightarrow Cu^{2+} + NO(g)} \\ 0 \quad 5+,2- \quad 2+ \quad 2+,2- \end{array}$$

Cu has undergone a change in oxidation number from 0 to 2+. An increase in oxidation number indicates oxidation. The change in oxidation number in this element is +2.

Rules for oxidation numbers

a) In free, uncombined, elements, the oxidation number of each atom is set to be 0.
b) In compounds involving hydrogen, the oxidation number of hydrogen is usually 1+.
c) In compounds involving oxygen, the oxidation number of O is usually 2–.
d) The sum of all the oxidation numbers of all atoms in an ion is equal (in both magnitude and sign) to the charge on the ion.
e) The sum of all the oxidation numbers of all atoms in a neutral molecule is 0.

N has undergone a change in oxidation number of 5+ to 2+. The change in oxidation number is –3. A decrease in oxidation number indicates reduction. There is no net change in oxidation number for the oxygen atoms.

We now introduce a further rule for oxidation numbers, and that is, the net *change* in oxidation number for the overall reaction is to be zero. That is, the overall *change* in oxidation number for the left side of the reaction has to be equal to the negative of the overall *change* oxidation number for the right side. The only way to do this here is to multiply the Cu atoms by 3 and the NO_3^- ions by 2. That is, operating on the changes in oxidation numbers, we have: $3(+2)+2(-3)=0$ Thus,

$$\mathrm{3Cu + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g).}$$

We need to add H_2O to get a balance for oxygen atoms:

$$\mathrm{3Cu + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O}$$

And finally, add H^+ ions to balance for H atoms:

$$\mathrm{3Cu + 8H^+ + 2NO_3^- \rightarrow 3Cu^{2+} + 2NO(g) + 4H_2O}$$

10.10 Electrochemical Cell

An **electrochemical cell**, **or galvanic cell**, is a redox reaction that proceeds spontaneously and has an **electrical potential difference** at its electrodes. This potential difference, when measured at a negligible current draw, is called the **emf** of the cell E_{cell}.

$$E_{\text{cell}} = \underset{\substack{\downarrow \\ \text{reduction}}}{E_{\text{cathode}}} - \underset{\substack{\downarrow \\ \text{oxidation}}}{E_{\text{anode}}}$$

When the half-cells are standard electrodes, the standard emf of the cell is given by

$$E^{\text{o}}_{\text{cell}} = E^{\text{o}}_{\text{cathode}} - E^{\text{o}}_{\text{anode}}$$

Example of electrochemical cell:

$$\begin{array}{ll} \text{Cu}^{2+}/\text{Cu} & \text{Fe}^{3+}(\text{aq})/\text{Fe}^{2+}(\text{aq}),\text{Pt} \\ E^{\text{o}} = +0.34\text{V} & E^{\text{o}} = +0.77\text{V} \end{array}$$

All at 1M concentrations

Since E^{o} for the iron reaction is higher than that for copper, then the Fe^{3+} is undergoing reduction (it is the oxidant) and so is the **cathode** for the cell. Copper is being oxidised and so is the **anode**.

$$\begin{aligned} E^{\text{o}}_{\text{cell}} &= E^{\text{o}}_{\text{cathode}} - E^{\text{o}}_{\text{anode}} \\ &= 0.77 - 0.34 \\ &= +0.43\text{V} \end{aligned}$$

Note: we do not have to "balance" the half-equations to determine the overall cell potential. For example, the **standard electrode potential** of the copper half cell is still +0.34 if we write

$$2\text{Cu}^{2+} + 4\text{e}^{-} \rightarrow 2\text{Cu}(\text{s})$$

The potentials do depend on the concentration of the species, but not the number of atoms that participate in the overall reaction.

In an electrochemical cell that does not have a connection between the anode and cathode, the redox reaction proceeds until the build-up of positive charge on the cathode prevents any more cations from moving from the anode and vice versa and the reaction stops. The system has acquired **electrical potential energy** as a result of the reaction proceeds. The reduction of energy of the system as a whole (i.e. the chemical reaction tending to send the system towards a lower energy state) appears as mechanical work or heat by virtue of the current flow.

10.11 EMF vs. Concentration

The potential of an electrochemical cell depends upon the concentration of the reactants. Standard potentials are given for 1M solutions. As the cell operates, the concentrations of the reactants change. The dependence of the cell potential on concentration is given by the **Nernst equation**.

$$p\mathrm{Ox} + ne^- \rightarrow q\mathrm{Red}$$

$$E = E^\circ + \frac{RT}{nF}\ln\frac{[\mathrm{Ox}]^p}{[\mathrm{Red}]^q}$$

$$= E^\circ + \frac{0.059}{n}\log_{10}\frac{[\mathrm{Ox}]^p}{[\mathrm{Red}]^q} \quad \text{at } 25^\circ\mathrm{C}$$

R - universal gas constant
T - temperature (K)
F - Faraday's constant 96487 C/mol

Consider the electrochemical cell at 25°C:

$$\mathrm{Pt}\,|\,\mathrm{Sn}^{4+}(0.5M),\mathrm{Sn}^{2+}(1.5M)\,||\,\mathrm{Fe}^{3+}(1.0M),\mathrm{Fe}^{2+}(0.5M)\,|\,\mathrm{Pt}$$

Anode Cathode

$$\mathrm{Sn}^{4+} + 2\mathrm{e}^- \leftrightarrow \mathrm{Sn}^{2+}$$

$$E_{\mathrm{Sn}} = E^\circ_{\mathrm{Sn}} + \frac{0.059}{2}\log_{10}\frac{\left[\mathrm{Sn}^{4+}\right]}{\left[\mathrm{Sn}^{2+}\right]}$$

$$= 0.15 + \frac{0.059}{2}\log_{10}\frac{[0.5]}{[1.5]}$$

$$= 9.14\mathrm{V}$$

$$\mathrm{Fe}^{3+} + \mathrm{e}^- \leftrightarrow \mathrm{Fe}^{2+}$$

$$E_{\mathrm{Fe}} = E^\circ_{\mathrm{Fe}} + \frac{0.059}{1}\log_{10}\frac{\left[\mathrm{Fe}^{4+}\right]}{\left[\mathrm{Fe}^{2+}\right]}$$

$$= 0.77 + \frac{0.059}{1}\log_{10}\frac{[1.0]}{[0.5]}$$

$$= 0.79\ \mathrm{V}$$

Since E_{Fe}>E_{Sn}, $\mathrm{Fe}^{3+}|\mathrm{Fe}^{2+}$ is the cathode and the cell reaction is

$$2\mathrm{Fe}^{3+} + \mathrm{Sn}^{2+} \rightarrow 2\mathrm{Fe}^{2+} + \mathrm{Sn}^{4+}$$

$$E_{\mathrm{cell}} = 0.79 + -0.14$$

$$= 0.65\mathrm{V}$$

When the electrode is a metal in contact with its cation (e.g. $\mathrm{Cu}^{2+}|\mathrm{Cu}$), then

$$\mathrm{Cu}^{2+} + 2\mathrm{e}^- \leftrightarrow \mathrm{Cu}_{(\mathrm{s})}$$

$$E = E^\circ + \frac{0.059}{2}\log_{10}\left[\mathrm{Cu}^{2+}\right]$$

When the electrode is a gas in contact with its ion (e.g. $\mathrm{H}^+|\mathrm{H}_2(\mathrm{g})$), then

$$2\mathrm{H}^+ + 2\mathrm{e}^- \leftrightarrow \mathrm{H}_2(\mathrm{g})$$

$$E = E^\circ + \frac{0.059}{2}\log_{10}\frac{\left[\mathrm{H}^+\right]^2}{\left(P_{H_2}\right)} \longrightarrow \text{Pressure in atm}$$

10.12 Electronic Equilibrium

Thus far we have considered the conditions under which an oxidation-reduction reaction will proceed. Consider an electrochemical cell:

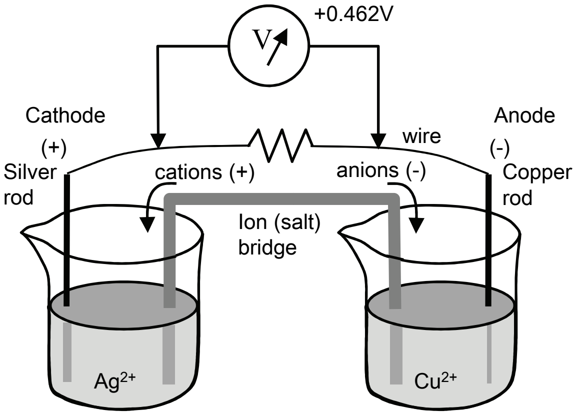

$2Ag^{2+} + 2e^{-} \rightarrow 2Ag(s)$ $Cu(s) \rightarrow Cu^{2+} + 2e^{-}$

Reduction half-reaction Oxidation half-reaction

How long does the spontaneous reaction proceed for? Does this cell "run down"? Does the voltmeter reading start to drop and then go to zero?

Initially, we know from our standard electrode potentials that the attainment of equilibrium, the objective of all chemical reactions, favours the forward reaction as written:

$$Cu(s) + 2Ag^{2+}(aq) \rightarrow Cu^{2+}(aq) + 2Ag(s)$$

As the reaction proceeds, the Ag^{2+} ions are being used up in the cathode solution, and there is a build-up of Cu^{2+} ions in the anode solution. Electrons lose potential energy as they pass through the wire and this energy is converted to heat in the resistor. When the electrons combine with the silver ions at the cathode, they are in a lower energy state compared to when they were bound to the copper. Attainment of a lower energy state is the driving force behind most spontaneous chemical reactions.

When the silver ions are in short supply, then not all the electrons coming from the wire will find a home to go to and there will be a slight increase in negative charge at the cathode. When there are no more silver ions to accept these electrons, the build-up of negative electric charge will be great enough to prevent any more from arriving. At this point, the electrical potential difference between the anode and cathode will be zero and the cell will be exhausted.

10.13 Equilibrium Constant

The potential of an electrochemical cell depends upon the concentration of the reactants. When the cell is exhausted, we reach a state of equilibrium and the **equilibrium constant** for the reaction is satisfied. Consider the redox reaction:

$$5\mathrm{Fe}^{2+} + \mathrm{MnO}_4^- + 8\mathrm{H}^+ \rightarrow 5\mathrm{Fe}^{3+} + \mathrm{Mn}^{2+} + 4\mathrm{H}_2\mathrm{O}$$

$$K_\mathrm{c} = \frac{\left[\mathrm{Fe}^{3+}\right]^5\left[\mathrm{Mn}^{2+}\right]}{\left[\mathrm{Fe}^{2+}\right]^5\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}$$

At equilibrium, $E = 0$, and so

$$E^\mathrm{o}_\mathrm{Fe} + 0.059\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]}{\left[\mathrm{Fe}^{2+}\right]} = E^\mathrm{o}_{\mathrm{MnO}_4^-} + \frac{0.059}{5}\log_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}$$

$$E^\mathrm{o}_{\mathrm{MnO}_4^-} - E^\mathrm{o}_\mathrm{Fe} = 0.059\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]}{\left[\mathrm{Fe}^{2+}\right]} - \frac{0.059}{5}\log_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}$$

$$\begin{aligned}
\Delta E^\mathrm{o} &= \frac{0.059}{5}\left(\log_{10}\frac{\left[\mathrm{Fe}^{3+}\right]^5}{\left[\mathrm{Fe}^{2+}\right]^5} - og_{10}\frac{\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}{\left[\mathrm{Mn}^{2+}\right]}\right)\\
&= \frac{0.059}{5}\log 10\frac{\left[\mathrm{Fe}^{3+}\right]^5\left[\mathrm{Mn}^{2+}\right]}{\left[\mathrm{Fe}^{2+}\right]^5\left[\mathrm{MnO}_4^-\right]\left[\mathrm{H}^+\right]^8}\\
&= \frac{0.059}{5}\log_{10}K_c
\end{aligned}$$

$$\begin{aligned}
K_c &= \log 10\frac{5\Delta E^\mathrm{o}}{0.059} & E^\mathrm{o}{}_{\mathrm{MnO}_4^-} &= +1.51\mathrm{V}\\
&= 5.1\times10^{62} & E^\mathrm{o}{}_\mathrm{Fe} &= +0.77\mathrm{V}
\end{aligned}$$

10.14 Lead/Acid Battery

A **battery** is a collection of cells arranged in series. In a motor vehicle **lead acid battery**, there are six cells, each producing about 2V, arranged in series to provide 12V for the vehicle. The redox couples involved are:

$$\begin{array}{ll} \mathrm{PbSO_4(s)/Pb(s),SO_4^{2-}} & \mathrm{PbO_2(s),H^+,SO_4^{2-}/PbSO_4(s)} \\ E^{\mathrm{o}} = -0.36\mathrm{V} & E^{\mathrm{o}} = +1.69\mathrm{V} \end{array}$$

Since the standard potential for the PbO_2 reaction is higher, then this species undergoes reduction and is the cathode.

Reduction (cathode) $$\mathrm{PbO_2(s) + 4H^+ + SO_4^{2-} + 2e^- \rightarrow PbSO_4(s) + 2H_2O}$$

Oxidation (anode) $$\mathrm{Pb(s) + SO_4^{2-} \rightarrow PbSO_4(s) + 2e^-}$$

$$\mathrm{PbO_2(s) + 4H^+ + 2SO_4^{2-} \rightarrow 2PbSO_4(s) + 2H_2O}$$ **Discharge** reaction

$$\begin{aligned} E^{\mathrm{o}}_{\mathrm{cell}} &= E^{\mathrm{o}}_{\mathrm{cathode}} - E^{\mathrm{o}}_{\mathrm{anode}} \\ &= 1.69 - -0.36 \\ &= 2.05\mathrm{V} \end{aligned}$$

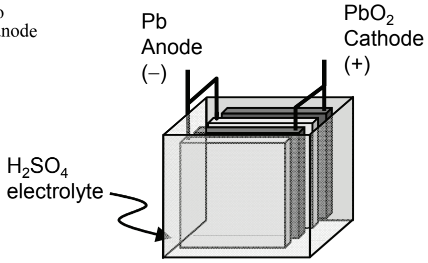

The concentration of H_2SO_4 in a motor vehicle battery is usually about 5M and so the actual cell potential is about 2.15V.

The **anode** (the negative plates) is solid lead, Pb, while the **cathode** (the positive plates) is **lead peroxide**, PbO_2, in a solution of **sulphuric acid**. The reaction product, **lead sulphate** $PbSO_4$, sometimes appears as a white precipitate on the battery terminals.

As the cell discharges, the concentration of H_2SO_4 decreases and the density of the solution also decreases. The density of the solution is thus a measure of the state of **charge** of the battery and is measured with a **hydrometer**.

When the battery is charged, the above reaction proceeds in reverse. Overcharging results in the evolution of gaseous H_2 at the anode and O_2 at the cathode, and the water in the electrolyte has to be replenished.

10.15 Dry Cell

In a conventional **dry cell**, or **Leclanché cell**, oxidation occurs at a zinc anode and reduction occurs at an inert carbon cathode. The electrolyte is not a liquid, but a paste of magnesium dioxide MnO_2, zinc chloride $ZnCl_2$, ammonium chloride NH_4Cl, and carbon black particles.

$Zn(s) \rightarrow Zn^{2+}(aq) + 2e^-$	Oxidation (anode −)
$2MnO_2(s) + H_2O + 2e^- \rightarrow Mn_2O_3(s) + 2OH^-$	Reduction (cathode +)
$NH_4^+(aq) + OH^-(aq) \rightarrow NH_3(g) + H_2O$	Acid/base absorbs the OH^- ions
$Zn^{2+}(aq) + 4NH_3(aq) \rightarrow [Zn(NH_3)_4]^{2+}(aq)$	Complex ion formation absorbs NH_3 gas
$Zn(s) + 2MnO_2(s) \rightarrow ZnMn_2O_4(s)$	Overall cell reaction

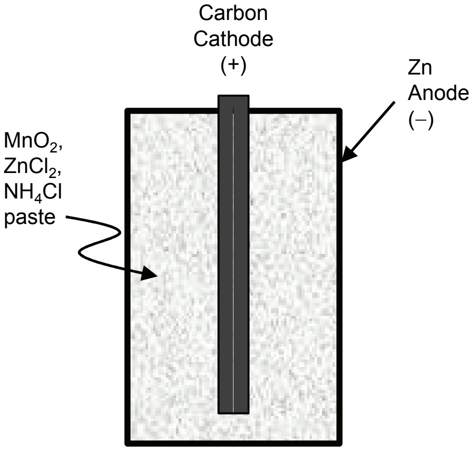

Modern **alkaline cells** use an alkaline electrolyte of KOH instead of the MnO_2, $ZnCl_2$, NH_4Cl paste of the zinc-carbon type of dry cell. In an alkaline cell, the half-reactions are

$$Zn(s) + 2OH^-(aq) \rightarrow ZnO(s) + H_2O(l) + 2e^-$$

$$2MnO_2(s) + H_2O(l) + 2e^- \rightarrow Mn_2O_3(s) + 2OH^-(aq)$$

10.16 Corrosion

Corrosion is the undesirable oxidation of a metal. The most well-known example is the **rusting of iron**. In simple terms, the complex redox reaction for the rusting of iron is written as

$$4\mathrm{Fe(s)} + 3\mathrm{O_2(g)} \rightarrow 2\mathrm{Fe_2O_3(s)}$$

In practice, the rusting of iron requires both air (O_2) and water, and the reaction proceeds as a series of steps.

$$\mathrm{Fe(s)} \rightarrow \mathrm{Fe_2^{2+}} + 2\mathrm{e^-} \qquad \mathrm{O_2} + 2\mathrm{H_2O} + 4\mathrm{e^-} \rightarrow 4\mathrm{OH^-}$$

Oxidation half-reaction   Reduction half-reaction

The OH^- ions react with the Fe^{2+} ions:

$$\mathrm{Fe^{2+}} + 2\mathrm{OH^-} \rightarrow \mathrm{Fe(OH)_2}$$

then $4\mathrm{Fe(OH)_2} + \mathrm{O_2} + 2\mathrm{H_2O} \rightarrow \mathrm{Fe(OH)_3(s)}$

and finally $\mathrm{Fe(OH)_3(s)} \rightarrow \mathrm{Fe_2O_3{,}H_2O} + 2\mathrm{H_2O}$

The brown coloured metal we usually associate with rust is the compound $\mathrm{Fe_2O_3{,}H_2O}$, an Fe(III) hydrated oxide

Iron and steel often rust badly because the rust has a lower density than the parent metal and tends to flake off, thus exposing more of the material to water and oxygen and promoting further rusting.

Water and oxygen are typically excluded by painting the iron article. However, a very effective way to reduce rusting of iron is to attach a more reactive metal (i.e. one with a lower E^o so that the iron undergoes reduction and the **sacrificial anode** (usually zinc) undergoes oxidation. This process is often called **cathodic protection**.

$\mathrm{Fe/Fe_2^{2+}}$ $\mathrm{Zn^{2+}/Zn}$

$E^o = -0.44\mathrm{V}$ $E^o = -0.76\mathrm{V}$

Iron or steel coated with zinc is said to be **galvanised** (since it acts like a galvanic cell).

Water and oxygen are typically excluded by painting the iron article. However, a very effective way to reduce rusting of iron is to attach a more reactive metal (i.e. one with a lower E^o so that the iron undergoes reduction and the sacrificial metal (usually zinc) undergoes oxidation.

Aluminium undergoes corrosion in the presence of atmospheric oxygen. A very thin (nm) tough layer of Al_2O_3 (**alumina**) appears within nanoseconds on the metallic surface which shields the parent metal from further corrosion. Alumina is a very hard ceramic material and is a non-conductor of electricity.

10.17 Electrolysis

An **electrolytic cell** is not a spontaneous redox reaction, and requires an electrical energy input to proceed.

Consider the redox reaction:

$$Cu^{2+}(aq) + 2Ag(s) \rightarrow Cu(s) + 2Ag^{2+}(aq)$$

Arranged in this way, the half-reactions are:

$$2Ag(s) \rightarrow 2Ag^{2+}(aq) + 2e^- \qquad Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$$

Oxidation half-reaction	Reduction half-reaction
-0.799V	+0.337V

The total potential is −0.462V and so this is not a spontaneous reaction. If a cell were constructed and connected with a wire between the two electrodes, the reaction would proceed in the reverse to that as shown above. However, if we apply an external **potential difference** to the electrodes, say a 1V potential, then

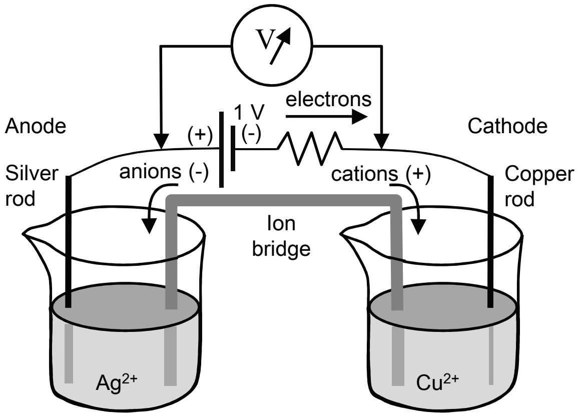

$$2Ag(s) \rightarrow 2Ag^{2+}(aq) + 2e^- \qquad Cu^{2+}(aq) + 2e^- \rightarrow Cu(s)$$

Oxidation half-reaction Reduction half-reaction

Arranged in this way, we have an **electrolytic cell** in which the copper rod is the cathode (+) and the silver rod is the anode (-). The reaction proceeds as written.

Note: **oxidation** occurs at the **anode** which in this case is the silver. **Reduction** occurs at the **cathode**, in this case the copper. The silver rod dissolves away and copper particles coat the copper rod.

The applied external potential has to overcome +0.462V in order for the reaction to proceed, and so the potential difference across the electrodes would be 1 – 0.462 = 0.538V.

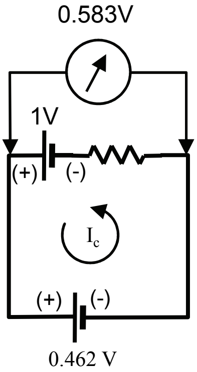

I_c indicates conventional current flow

11. Carbon Chemistry

Summary

Functional groups

H	Alkanes	Single C bond
C=C	Alkenes	Double C bond
C≡C	Alkynes	Triple C bond
OH	Alcohols	ROH
O	Ethers	ROR'
CHO	Aldehydes	RCHO
CO	Ketones	R(C=)OR
COOH	Acids	RCOOH
COO	Esters	RCOOR'
NH_2	Amines	RNH_2 R_2NH, R_3N

Alkyl groups

Methyl
Ethyl
Propyl
Butyl
Pentyl
Hexyl
Heptyl
Octyl
Nonyl
Decyl

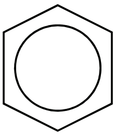

Benzene ring

11.1 Carbon

The element **carbon** has four outer valence electrons and is the first element of Group IV in the periodic table. The other elements in this group become more metallic in nature with increasing atomic number. At room temperature, carbon is relatively inert. At high temperatures, it can form covalent bonds with itself and many other elements. Carbon compounds are the basis for all life on Earth. It is interesting to note that **silicon**, the next element down in the group and also with four valence electrons, is the basis of the composition for nearly all of the Earth's minerals.

With four valence electrons, carbon has the choice of either losing or gaining four electrons to achieve a noble gas configuration but in practice, it tends to share its four electrons via covalent bonds. On the other hand, tin and lead, the heaviest elements in Group IV, with more loosely bound valence electrons, find it energetically favourable to share these electrons in the form of **metallic bonds**.

Carbon reacts with metals to form carbides: e.g. silicon carbide SiC; with non-metals to form molecular compounds: e.g. methane CH_4; with oxygen to form carbon monoxide CO and carbon dioxide CO_2; and with nitrogen to form cyanides (e.g. HCN).

Pure carbon forms three crystalline structures found naturally: **diamond**, **graphite,** and the **fullerenes**. These are **allotropes**. These have significantly different mechanical, thermal and electrical properties and differ only in the geometrical arrangement of their atoms in the crystal structures.

Diamond has a three-dimensional tetrahedral structure. Each carbon atom has four neighbours, and all the valence electrons are equally shared with no vacancies or lone electrons. Thus, diamond is a poor electrical conductor, has a high **hardness** and **elastic modulus**. Diamond is a good heat conductor due to the free passage of **phonons**.

Graphite has a two-dimensional hexagonal sheet structure with the sheets bonded together as layers by **van der Waals forces**. Each carbon atom has three neighbours. Thus, three valence electrons are involved in covalent bonding while the fourth is left in a relatively free state and is responsible for graphite being an electrical conductor.

The **fullerene** form of carbon consists of a three-dimensional arrangement of C atoms in a pentagon which together form a sphere or tube. The most famous is the C_{60} molecule, often referred to as a **buckyball**.

11.2 Carbon Compounds

Carbon forms a great number of compounds with hydrogen. These are called **hydrocarbons**. In addition, **hydrocarbon derivatives** are formed when other atoms in addition to hydrogen become involved in the bonding.

Central to this remarkable ability to form such a wide variety of compounds is the formation of chains of atoms. For example, consider the simple hydrocarbon **methane** CH_4:

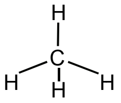

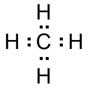

This molecule can be easily expanded to form **ethane** C_2H_6:

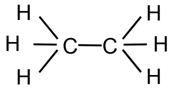

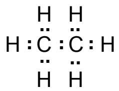

And so on for **propane**, C_3H_8, and **butane**, C_4H_{10}.

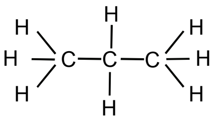

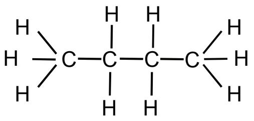

Other possibilities abound, including compounds involving a double carbon bond such as in **ethylene** and the triple bond in **acetylene**:

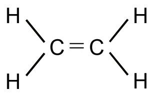

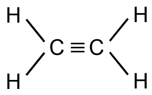

Compounds with single bonds between carbon atoms are **saturated hydrocarbons**. Compounds with double or triple bonds are called **unsaturated hydrocarbons**.

The reason for this is that saturated compounds are saturated with H atoms. When there is a double bond, the maximum possible number of H atoms that can attached to a C atom is decreased and so the C atoms is unsaturated.

Hydrocarbons with the same chemical formula but with different structures are possible. For example, **butene** C_4H_8, can exist in four different structural arrangements, each of which is called an **isomer**.

The substitution of other elements in place of one or more hydrogen atoms in carbon compounds is responsible for the existence of a large number of hydrocarbon derivatives. Alcohols are hydrocarbons with an attached OH **functional group**.

The ability of carbon to form chains, double bonds, isomers and contain functional groups is responsible for the great variety of carbon compounds which form the chemistry of life on Earth – or **organic chemistry**.

11.3 Functional Groups

Hydrocarbon derivatives are formed when other atoms in addition to hydrogen become involved in the bonding. In these cases, the core hydrocarbon is written with an attached **functional group**. The core or hydrocarbon compound (also called the **remainder** or **residue)** is written with the symbol R. A common functional group is the alcohol, or OH group.

methanol CH_3OH:

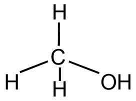

Alcohols can be written as ROH where R is the hydrocarbon residue, or the core hydrocarbon component, and OH is the alcohol functional group.

ethanol C_2H_5OH:

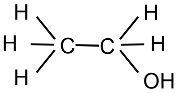

Functional groups

H	Alkanes
C=C	Alkenes
C≡C	Alkynes
OH	Alcohols
O	Ethers
CHO	Aldehydes
CO	Ketones
COOH	Acids
COO	Esters
NH_2	Amines

Functional groups determine the nature of the reactivity of the hydrocarbon involved. The chemistry of the compound is usually determined by the character of the functional group while the hydrocarbon residue is relatively inert.

Consider some compounds based upon the CH_3CH_2- residue:

Ethane: $CH_3CH_2 - H$

Ethanol: $CH_3CH_2 - OH$

Ethyl bromide: $CH_3CH_2 - Br$

The CH_3CH_2- residue is called the **ethyl group**. Similarly, the CH_3- is the **methyl group** and the $CH_3CH_2CH_2-$ is the **propyl group**.

These named **alkyl groups** form bonds with various different functional groups and the resulting compounds R–X are thus **derivatives** of the parent hydrocarbon (e.g. methane, ethane, etc).

11.4 Alkanes

Alkanes are **saturated hydrocarbons** since the bonds with the carbon atoms are single bonds and they contain only carbon and hydrogen. The names of the compounds are derived from the number of carbon atoms.

The first 10 alkanes

Methane	CH_4
Ethane	C_2H_6
Propane	C_3H_8
Butane	C_4H_{10}
Pentane	C_5H_{12}
Hexane	C_6H_{14}
Heptane	C_7H_{16}
Octane	C_8H_{18}
Nonane	C_9H_{20}
Decane	$C_{10}H_{22}$

Alkanes can have chain or branched chain (**aliphatic**), or cyclic (**alicyclic**) structures. As the complexity of the molecules increases, there are increased ways in which molecules can form structures with the same molecular formula.

Compounds having the same molecular formula but different structure are called **isomers**.

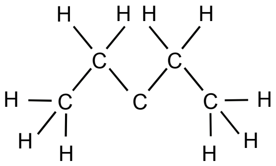

n-pentane (or normal pentane):

Alkanes are the principal components of **petroleum**. Low molecular weight compounds are generally gases at room temperature and pressure. Gasoline (petrol) and turpentine typically contain light fractions (C_7H_{16} to $C_{10}H_{22}$). Kerosene, diesel oil, lubricating oils, pitch, and so on consist of heavier compounds. Alkanes are relatively inert, but can undergo exothermic oxidation (**combustion**) which make them useful sources of energy.

Fuel	**Calorific value** (MJ/kg)
Petrol	45
Diesel	44.4
Ethanol	27.2

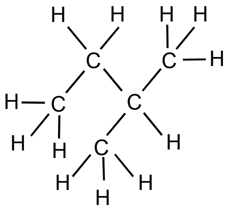

Alkanes with a chain structure have the molecular formula C_nH_{2n+2}.

iso-pentane (single branch at the end)

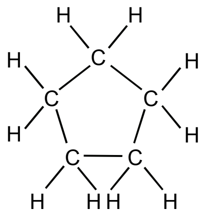

Alkanes having a cyclic or a ring structure have the formula C_nH_{2n}.

cyclo-pentane (ring structure)

Alkanes are sometimes called **paraffins** because they have little (parum) affinity for forming bonds with other atoms or molecules. When they do participate in reactions, this usually involves the **substitution** of a hydrogen atom with another atomic species.

11.5 Alkenes

Alkenes (also known as **olefins**) are unsaturated hydrocarbons because they contain a double bond between two of the available carbon atoms instead of single bonds. As a consequence, there are fewer than four atoms attached to each of the carbon atoms.

Alkenes in a chain configuration have the general formula C_nH_{2n}. Ethene or **ethylene** (C_2H_4) is the simplest alkene.

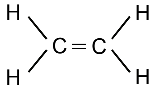

Bond length: 1.33 Å
Energy: 606 kJ/mol
ΔH = +52.5 kJ/mol

Note that the carbon double bond can occur in different places in the structure, leading to formation of isomers. In **propylene** (C_3H_6):

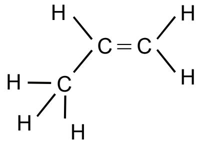

Ethene, propene, and butene are gases. Higher alkenes are generally liquids and those with greater than sixteen carbon atoms are generally waxy solids.

While alkenes are relatively inert, they are more reactive than alkanes because of the double bond. The carbon double bond has a higher bond strength compared to the single bond, but the double bond allows a greater potential for attachment for other atoms and this leads to greater reactivity compared to single-bonded structures. Chemical reactions involving alkenes are usually associated with the breaking of the double bond and the addition of other atoms.

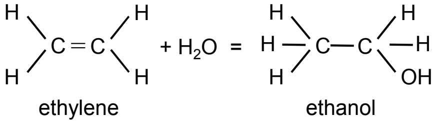

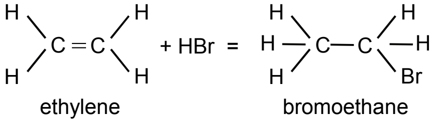

An important reaction involving alkenes is that of **polymerisation**, where a chain reaction occurs in which the C=C double bonds are successively broken and employed to form larger and larger single-bonded structures.

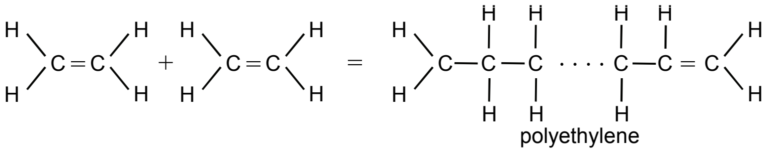

11.6 Alkynes

Alkynes are similar to alkenes except that they feature a carbon-carbon triple bond. Alkynes are unsaturated hydrocarbons and have the general formula C_nH_{2n-2}. Alkynes are also known as **acetylenes**, but the common use of this term refers to **ethyne**: C_2H_2.

C_2H_2 acetylene $H—C≡C—H$

Bond length: 1.2 Å
Bond energy: 828 kJ/mol
ΔH = +226.9 kJ/mol

The triple bond can occur anywhere in the chain in higher alhynes. Consider the compound 1-butyne. The 1- indicates that the triple bond is at the end of the chain:

$CH_3CH_2C≡CH$

The CH_3CH_2 is the ethyl group, and so this compound can also be named **ethylacetylene**.

Ethyne	$HC≡CH$
Propyne	$CH_3C≡CH$
1-Butyne	$CH_3CH_2C≡CH$
1-Pentyne	$CH_3(CH_2)_2C≡CH$
1-Hexyne	$CH_3(CH_2)_3C≡CH$
1-Heptyne	$CH_3(CH_2)_4C≡CH$
1-Octyne	$CH_3(CH_2)_5C≡CH$
1-Nonyne	$CH_3(CH_2)_6C≡CH$
1-Decyne	$CH_3(CH_2)_7C≡CH$

Alkynes have a similar chemistry to alkenes. The C≡CH unit is a linear structure and in acetylene, the molecules are therefore rod-like in shape.

The triple bond is capable of opening up and accommodating additional atoms during reactions. Additions may occur with halides, halogens, water, and ammonia.

The C≡CH unit, located at the end of the structure, imparts an acidic nature to alkynes and so acid-base reactions are also possible. Alkynes can also undergo **polymerisation** to give **polyacetylenes**.

One of the most common uses of acetylene is in gas welding. The combustion of acetylene in oxygen provides a flame temperature of about 3300°C making it one of the hottest flames available in industry.

$$2C_2H_2 + 5O_2 \rightarrow 4CO_2 + 2H_2O$$

Acetylene is a gas, but an unstable one, especially at moderate to high pressures. In welding gas bottle sets, acetylene is usually dissolved in liquid **acetone** or dimethylformamide and stored under pressure in a cylinder containing a porous medium, **agamassam**.

11.7 Benzene

An important class of carbon compounds arises when carbon atoms are arranged in a special ring formation.

The **benzene ring** C_6H_6 has all the atoms in a plane, with a 120° angle between the carbon atoms. The initial difficulty with this structure is that each carbon atoms has three neighbours sharing one electron, but there are four valence electrons in carbon. Evidently, there either has to be three double bonds present or some way of accommodating the fourth valence electron. If each second carbon atom had a double bond, then all available valence electrons would be shared, but experimental evidence indicates that the bond strength for the C-C bonds are equal around the ring, and that it is stronger than a single bond, but not as strong as a double bond.

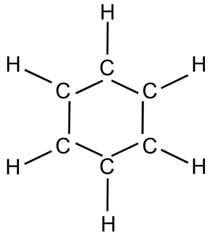

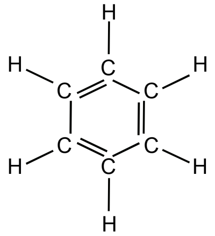

The bonding in a benzene ring is explained by the allowable movement of the **valence electrons** in the ring structure. Rather than forming single or double bonds, electrons in the ring form a resonance hybrid, or a superposition of single and double bonds, whereby they may have movement throughout the ring, much like valence electrons in a metal have movement throughout the **conduction band**.

To signify this, the benzene ring is often written in shorthand form using one of the following diagrams:

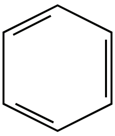

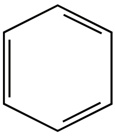

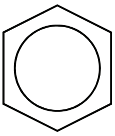

It should be always remembered that there are not single or double bonds, but a special bond somewhere between the two based upon electron movement around the ring.

Ring compounds of this type are called **aromatic compounds**. When they consist of C and H only, they are called **aromatic hydrocarbons**. Note: **aliphatic** compounds in a ring structure are not aromatic; they are simple ring structures (**alicyclic**). Aromatic compounds have mobile electrons within the ring and it is these that determine the chemistry of these compounds. In contrast to the addition reactions for alkenes with a chain structure, benzene undergoes substitution reactions in which H atoms are replaced by other atoms.

11.8 Alcohols

Hydrocarbon derivatives arise when functional groups combine with a residual hydrocarbon core. **Alcohols** are those compounds which have one or more OH (**hydroxyl**) groups attached, ROH.

Three types of alcohol are to be found depending on the number of C atoms attached to the C atom with the OH group. Primary (n) alcohols are written RCH_2OH.

$CH_3CH_2CH_2 — OH$ 1-propyl alcohol
n-propanol

Secondary (s) alcohols are written RR'CHOH:

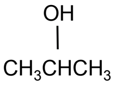

$$\overset{\overset{OH}{|}}{CH_3CHCH_3}$$

isopropyl alcohol
2-propanol
s-propanol

Tertiary (t) alcohols are RR'R"COH.

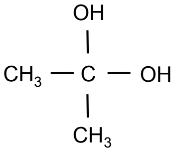

$$CH_3 — \underset{\underset{CH_3}{|}}{\overset{\overset{OH}{|}}{C}} — OH$$

t-butyl alcohol

Lower alcohols are soluble in water due to the presence of the polar OH group. In higher alcohols, the longer the carbon chain, the less water-like the compound is and behaves more like a hydrocarbon – volatile and relatively inert. **Pentanol** and higher compounds are insoluble in water.

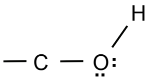

Alcohols can react with other atoms and can act as weak **bases** (due to the unshared electron pairs on the O atom) or weak **acids** – where the H atom on the OH group can act as a proton donor.

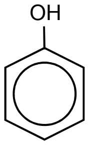

Phenol (hydroxybenzene) has the OH group attached to a benzene ring and forms the basis of synthesis for many other compounds.

Ethanol (C_2H_5OH) is "alcohol" to the common person (e.g. in alcoholic drinks). **Denatured ethanol** is the usual form used for cleaning and other uses. Additives render it unpalatable and, in some cases poisonous (such as when methanol is added to the ethanol to give **methylated spirits**).

Methanol (CH_3OH) is highly toxic and may cause blindness or death when only a few millilitres are ingested.

A common cleaning alcohol is **isopropyl alcohol**, or 2-propanol $(CH_3)_2CHOH$.

Ethylene glycol $C_2H_4(OH)_2$ is commonly used as an anti-freeze/ anti-boil additive in motor vehicle cooling systems. In its pure form it has a boiling point of 197.3 °C but has a lower heat capacity than water and so is used diluted. It does this by depressing the hydrogen bonding between water molecules so that, upon freezing, the temperature must be significantly lowered before ice crystals can form.

An interesting compound prepared from phenol, sodium hydroxide, and carbon dioxide is **aspirin**.

11.9 Ethers

Ethers have the general formula ROR' where R is an alkyl group. They can occur as aliphatic (chained), alicyclic, or aromatic (**benzene ring**) formations, or a combination. They are relatively unreactive hydrocarbon derivatives.

The lone pairs of electrons on the O atom allow ethers to form **hydrogen bonds** with water molecules and so they are generally partially soluble in water.

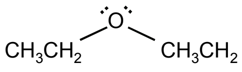

diethyl ether

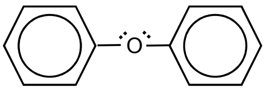

diphenyl ether

However, they cannot form hydrogen bonds between themselves and so generally have a lower boiling point than comparable alcohols.

In some respects, ethers are similar to alcohols in that they can be considered derivatives of water. In an alcohol, the R takes the place of one of the H atoms in the water molecule. In an ether, both of the H atoms are replaced by alkyl groups R and R'. Ethers are generally isomers of the corresponding alcohols. For example, ethanol, C_2H_5OH, has the same formula as dimethylether, but of course has quite different chemical and physical properties.

Perhaps one of the most well known ethers is **diethyl ether**, $CH_3CH_2OCH_2CH_3$, a low boiling point colourless liquid used as an anaesthetic. Diethyl ether can be formed from the removal of water from ethanol.

Dimethyl ether, CH_3OCH_3, (or **DME**) is a gas at room temperature is often used as an aerosol propellant since it can be stored as a liquid under pressure in equilibrium with the gas phase, evaporates readily and leaves the desired product as an aerosol mist for the application.

Methyl phenyl ether $CH_3OC_6H_5$ is found in aniseed oil.

Ethers are generally unreactive substances and can withstand attack from either acids or bases, but are more reactive than comparable alkanes.

Simple ethers have the alkyl groups on each end of the oxygen atom and if the groups are the same, are thus named dialkyl ether (such as in dimethylether). If one of the groups is a more complex group, then the structure of the name is alk-oxy-alkane. For example, $CH_3CH_2CH_2OCH_3$ is **methoxypropane**.

11.10 Aldehydes

Alcohols can be thought of as a step towards complete oxidation of the corresponding alkane. Primary alcohols can also undergo oxidation to form **aldehydes** and **carboxylic acids**, and secondary alcohols can oxidise to form **ketones**.

Aldehydes have a carbon-to-oxygen double bond (which itself is called the **carbonyl group**) with the general formula RCHO. They are usually prepared by **oxidation** of the corresponding alcohol.

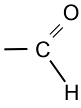

aldehyde group

Perhaps one of the most well known aldehyde is that made from **formaldehyde** dissolved in water to give **formalin**, a preservative fluid used for embalming biological tissues.

Consider the reaction between **methanol** and potassium dichromate:

$$3CH_3OH + Cr_2O_7^{2-} + 8H^+ \rightarrow 3CH_2O + 2Cr^{3+} + 7H_2O$$

In this equation, the Cr_2 ion has been reduced to the Cr^{3+} ion, and the methanol has lost two electrons (and two protons H^+) and hence has been oxidised to give CH_2O, formaldehyde.

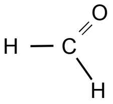

methanal (formaldehyde)

Methanal	HCHO	Formaldehyde
Ethanal	CH_3CHO	Acetaldehyde
Propanal	CH_3CH_2CHO	Propionaldehyde
Butanal	$CH_3CH_2CH_2CHO$	Butyraldehyde

Aldehydes often have a pleasant odour or flavour, such as in the benzene derivative vanillin:

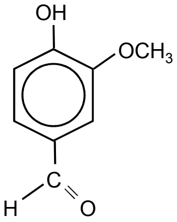

and cinnamaldehyde, the active ingredient in cinnamon.

To avoid confusion with alcohols (OH), the carbonyl group with an attached H is written CHO and not COH.

Cyclic forms also exist, such as **benzaldehyde**, an aromatic aldehyde:

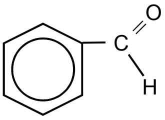

Aldehydes are closely related to **ketones**, but have the **carbonyl group** C=O at the end of a carbon chain rather than in between two C atoms. Lower aldehydes are soluble in water. Aldehydes can be further oxidised to form **carboxylic acids**.

11.11 Ketones

Ketones have a **carbonyl group** C=O attached to two other C atoms that form part of the same or different **alkyl groups**. (In an aldehyde, one of the other atoms is an H atom.) The general formula is RC(=O)R'.

The oxygen atom is more **electronegative** than attached carbon (i.e. the electron pairs that constitute the double bond are closer to the oxygen atom than the carbon atom), making this group polar. Unlike aldehydes, ketones do not have an H atom attached to the carbonyl group.

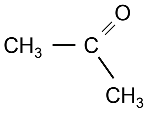

2-propanone
(acetone)

Acetone can be formed by the oxidation of 2-propanol according to

$$3CH_3CHOHCH_3 + Cr_2O_7^{2-} + 8H^+ \rightarrow$$

$$3CH_3COCH_3 + 2Cr^{3+} + 7H_2O$$

Aldehydes and **ketones** have many similarities and both are formed from the oxidation of an alcohol.

Ketones cannot be further oxidised to an acid since there is no H atom available on the functional group (as in the case of **aldehydes**). The absence of an H atom bonded to an oxygen atom means that both aldehydes and ketones do not form **hydrogen bonds** with each other and so generally have a lower boiling point than comparable alcohols. However, the lone pairs of electrons on the oxygen atom can form a hydrogen bond with an H atom in a water molecule and so lower ketones (and also aldehydes) are generally soluble in water.

The names of ketones (and aldehydes) are formed by the number of C atoms in the longest chain, and include the C atom in the carbonyl group.

Acetone CH_3COCH_3 is widely used as a relatively non-toxic solvent for various carbon compounds (e.g. paint, plastics, nail varnish, superglue, etc). It also has the useful property of being soluble in water and is commonly used as a drying agent for glassware in chemistry laboratories. It is very volatile and dries without leaving a residue.

Another common ketone used as a solvent for paints, varnish and gums is 2-butanone, or **methyl ethyl ketone** (**MEK**), or methyl acetone, $CH_3COCH_2CH_3$, a flammable clear liquid that is soluble in water. It has a characteristic sweet odour. As well as industrial uses as a solvent, it is also used as a polystyrene cement in plastic model kits.

Another significant ketone, with a cyclic (but not aromatic) structure, is **cyclohexanone** $(CH_2)_5CO$ which is used in the manufacture of **nylon**.

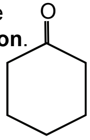

cyclohexanone

Ketones have significant biochemical functions and are produced when body fat is used as a source of energy when there is a shortage of **glucose** (such as during fasting).

11.12 Carboxylic Acids

Hydrocarbon derivatives that contain both the carbonyl (C=O) and hydroxyl (OH) functional groups are called **carboxylic acids** with the general formula RCOOH. When R is an aliphatic residue, the compound is called a **fatty acid**. When R is an aromatic ring, the acid is a **benzene ring derivative**.

The COOH group is called the **carboxyl group**. Carboxylic acids can be prepared from the oxidation of the corresponding alcohol.

$$5CH_3OH + 4MnO_4^- + 12H^+ \rightarrow$$

$$5HCOOH + 4Mn^{2+} + 11H_2O$$

In this reaction CH_3OH and H_2O produces HCOOH and four protons and four electrons. Since there is a loss of four electrons (compared to two for the oxidation of methanol to methanal, or formaldehyde) , formic acid is a more highly oxidised compound than **formaldehyde**.

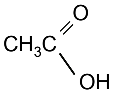
ethanoic acid
(acetic acid)

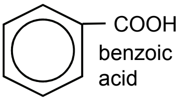
benzoic acid

A very common aliphatic acid is **acetic acid**, written CH_3COOH and is found in vinegar. Formic acid (from the latin formica for ant) in the found in venom of ants. Stearic acid is found in soaps, waxes and plant oils. Other carboxylic acids can be found in coconuts, chocolate,

Some carboxylic acids:

Formic	Methanoic	HCOOH
Acetic	Ethanoic	CH_3COOH
Propionic	Propionic	CH_3CH_2COOH
Butyric	Butanoic	CH_32CH_2COOH
Stearic	Octa decanoic	CH_316CH_2COOH

Similarly, ethanol can be oxidised to form **acetic acid**. Acetic acid can also be formed from the oxidation of acetaldehyde. Further oxidation of these compounds can be obtained via the rather energetic process of **combustion**, in which case the product is CO_2 and water.

Carboxylic acids are weak acids and dissociate only slightly, giving H^+ ions in water, and act as proton donors in reactions. The H^+ ions come from the COOH functional group. Lower carboxylic acids are soluble in water.

A common chemical demonstration reaction involves that between vinegar and **baking soda** to give sodium ethanoate, carbon dioxide and water:

$$CH_3COOH + NaHCO_3 \rightarrow 3CH_3COONa + CO_2 + 7H_2O$$

There is an ionic bond between the CH_3COO^- and Na^+ ions. By convention, the ionic compound is written with the metal ion last.

Carboxylic acids react with alcohols to form fats and oils (esters). **Fatty acids** are those needed by the body and are obtained from digestion of animal and vegetable oils and fats.

11.13 Esters

Carboxylic acids contain the functional group COOH. The H atom can be easily given up (proton donor): hence the acid classification.

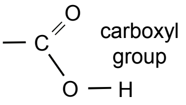

Derivatives of carboxylic acids are formed when the OH group is replaced by something else. When the OH group is replaced by OR, then an **ester** is formed. Esters are generally colourless liquids insoluble in water and have the general formula RCOOR'. For example, a reaction between methanol and acetic acid forms the ester **methyl acetate**:

$$CH_3OH + CH_3COOH \rightarrow CH_3COOCH_3 + H_2O$$

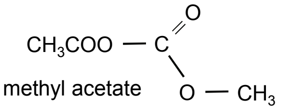

Note that the methyl group is the CH_3 at the end of the molecule. The "acetate" is from the carboxylic acid. Esters are written with the functional group first.

In this reaction, water is formed from the H atom in the incoming alcohol, and the OH group from the carboxylic acid, leaving behind an ester.

Esters are commonly used in food flavourings and perfumes due to their pleasant odour.

Methyl butyrate	Apple	$CH_3CH_2CH_2COOCH_3$
Ethyl butyrate	Strawberry	$CH_3CH_2CH_2COOCH_2CH_3$
Pentyl butyrate	Apricot	$CH_3CH_2CH_2COO(CH_2)_4CH_3$
Pentyl acetate	Banana	$CH_3COO(CH_2)_4CH_3$
Isoamyl acetate	Pear	$CH_3COOCH_2CH_2\,CH(CH_3)_2$
Octyl acetate	Orange	$CH_3COO(CH_2)_7CH_3$

Most naturally occurring fats and oils are very large esters made from organic acids and alcohols (usually **glycerol**). When there are no C=C bonds, the acid is said to be **saturated**. When the acid contains C=C bonds, it is **unsaturated**. When an acid has one C=C bond it is **monounsaturated**, and when there is more than one C=C bond, it is referred to as being **polyunsaturated**. These **fatty acids** form saturated, unsaturated and polyunsaturated fats and oils (oils are esters that are liquid at room temperature whereas waxes are solid).

Esters also undergo polymerisation to form common products such as **polyesters** used for clothing and recyclable polyethylene terephthalate (**PET**) drink bottles. An interesting example of an ester is nitroglycerine (which is somewhat incorrectly named).

11.14 Amides

Amides are derivatives of carboxylic acids in a similar manner to esters. When the OH group is replaced by NH_2, then an **amide** is formed. A special class of amides occurs when the OH group is replaced by NHR. Amides can be produced by the reaction between ammonia and an ester.

The reaction between methyl acetate and ammonia produces **ethanamide** (acetamide) and methanol:

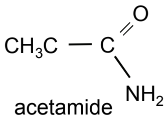
acetamide

$$CH_3COOCH_3 + NH_3 \rightarrow$$
$$CH_3CONH_2 + CH_3OH$$

In the NH_2 group, the electrons are more strongly drawn towards the more electronegative N atoms, leaving the H end with a positive charge and so amides are able to form **hydrogen bonds** with each other as they are attracted to the negatively charged lone pairs on neighbouring O atoms. This confers a relatively high melting point for these compounds. Amides have much the same solubility as comparable esters, with lower amides being soluble in water.

The most notable characteristic of amides are their mechanical strength and are often used, both in nature and industrially, as structural items.

In nature, amides for the basis of links between amino acids in **proteins**.

In industrial use, amides are joined together in long chains to produce polyamides, the most well known of which are **nylon** and **kevlar**.

Lysergic acid diethylamide (**LSD**) is also an amide.

Methanamide	$HCONH_2$
Ethanamide	CH_3CONH_2
Propanamide	$CH_3CH_2CONH_2$

The **amide link** joins amino acids together in biological processes to form **proteins**. **Amino acids** are carboxylic acids with an amine group of the form $H_2NCHRCOOH$. During polymerisation, water is eliminated between two acid groups to be joined together.

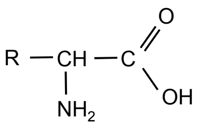
amino acid

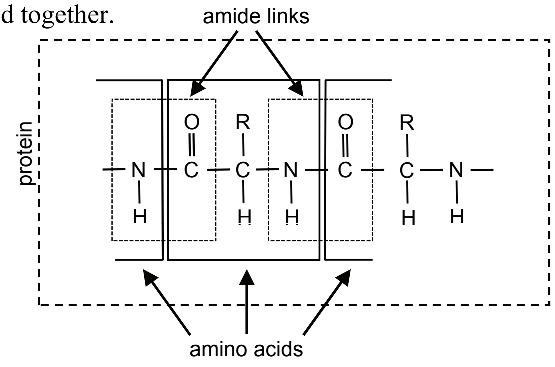

11.15 Amines

In an alcohol, the R takes the place of one of the H atoms in the water molecule to give the general formula ROH. Now consider the molecule **ammonia** NH_3.

If one or more H atoms are replaced by an alkyl group, then the resulting compound is called an **amine**. Primary amines RNH_2 have one R substitution, secondary amines R_2NH have two, and tertiary amines R_3N have all three H atoms replaced with alkyl groups. Aliphatic and aromatic structures are found.

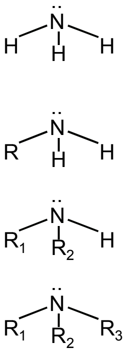

Ammonia	NH_3
Methylamine	CH_3NH_2
Ethylamine	$CH_3CH_2NH_2$
Propylamine	$CH_3CH_2CH_2NH_2$
Butylamine	$CH_3CH_2CH_2CH_2NH_2$
Aniline	$C_6H_5NH_2$

Just as in the case of ammonia, amines are bases due to the lone electron pair on the N atom. Primary and secondary amines can form hydrogen bonds with each other and all can form **hydrogen bonds** with water, thus making them soluble, the lower gaseous amines being very soluble.

Phenylamine, or **aniline**, is a primary amine where an H atom on the ammonia is replaced with a benzene ring. This has the interesting consequence of the lone pair of electrons on the N atom interacting with, and to some extent being absorbed into the ring, which in turn decreases its solubility and raises its boiling point compared to comparable aromatic compounds.

Ammonia is a colourless gas with a very sharp odour. Ammonium hydroxide in household cleaning products is a solution of ammonia gas in water where it slightly dissociates into the ammonium NH^{4+} and hydroxide OH^- ions.

Many primary aromatic amines are used as dyes.

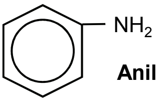

Aniline

Amines are important in biological functions and are produced from the breakdown of amino acids in the body. **Amino acids** are carboxylic acids with an amine group of the form $H_2NCHRCOOH$. Amines serve as neurotranmitters. A common natural amine is **histamine** which triggers an immune response. Anti-histamines suppress this activity in severe allergic reactions.

Methylphenethylamine (or **amphetamine**) is a powerful drug that modifies the action of natural neurotransmitters in the brain.

11.16 Polymers

The chemical and physical properties of hydrocarbon compounds and derivatives depend very much on the size of the molecule. Smaller groups such as methyl and ethyl compounds generally have a high solubility, a low boiling point. When these small molecular units, collectively called **monomers**, and joined together into long chains, the molecular size can increase greatly with the production of a variety of useful substances which are called **polymers**, or **plastics**.

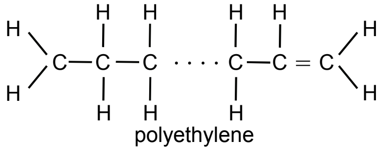
polyethylene

For example, in ethylene, the C=C bonds may be successively broken and joined to form larger and larger molecules of polyethylene. This process is usually performed with the help of a catalyst. The resulting compounds are called **addition polymers**. When other atoms or functional groups are involved, polymers of a great variety of properties are possible.

Another class of polymers are **condensation polymers**. In these, chains are formed by the expulsion of a water molecule from between the monomers to be joined. In **Nylon**, amide links are used to join monomers into large chain molecules.

One of the monomers used to make **Nylon** 6,6 is adipic (hexanedioic) acid:

$$HOOC—CH_2—CH_2—CH_2—CH_2—COOH$$

This is used with another monomer, 1,6 diaminehexane:

$$H_2N—CH_2—CH_2—CH_2—CH_2—CH_2—CH_2—NH_2$$

When these two monomers are chained together, water drops out as shown:

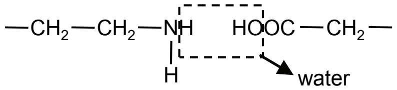

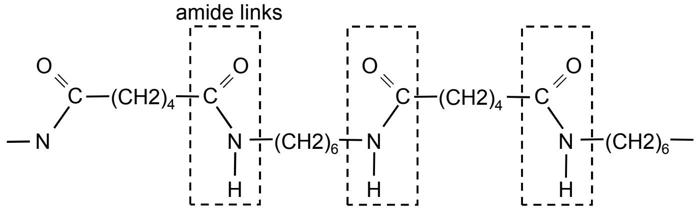

During polymerisation, these molecules continue to chain together, expelling a water molecule each time and forming amide linkages. **Kevlar** is a similar polymer but the amide links join benzene rings together.

11.17 Physical and Chemical Properties

Group	Boiling point	Solubility	Notes
Alkanes	Lower compounds are gases at room temp. Boiling point increases as size of molecule increases.	Do not form H bonds, insoluble in water.	Mainly unreactive with other compounds.
Alkenes	Similar to alkenes.	Similar to alkenes.	Slightly more reactive due to presence of C=C.
Alkynes			More reactive than alkenes due to C≡C.
Alcohols	H bonding causes boiling point to be higher than comparable alkanes.	Lower alcohols are soluble in water. Compounds become less water-like with increasing size.	OH group makes the molecules polar, and able to form H bonds.
Ethers	Cannot form H bonds between themselves so lower boiling point than comparable alcohol.	Can form H bonds with water molecules. Lower ethers soluble in water.	More reactive than comparable alkanes.
Aldehydes	Cannot form H bonds with each other and so generally have a lower boiling point than comparable alcohols.	Lower aldehydes soluble in water due to presence of lone pair electrons on O atom.	
Ketones	Similar to aldehydes.	Similar to aldehydes.	
Carboxylic acids	High boiling point due to H bonding.	Lower acids soluble.	Typically weak acids.
Esters	Lower boiling point due to lack of self-H bonds.	H bond with water confer solubility on lower esters.	
Amines	Reasonably high boiling point but lower than alcohols.	Lower amines soluble.	

12. Biochemistry

Summary

Glucose

$C_6H_{12}O_6$

Glycerol

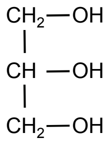

$$\begin{array}{l} CH_2-OH \\ \mid \\ CH-OH \\ \mid \\ CH_2-OH \end{array}$$

Amino acid

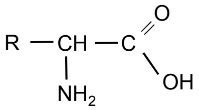

$$R-\underset{\displaystyle NH_2}{\underset{\mid}{CH}}-C\begin{array}{l}{/\!\!/}O\\ \backslash OH\end{array}$$

Base-pairing rules

A-T

G-C

Nucleotide

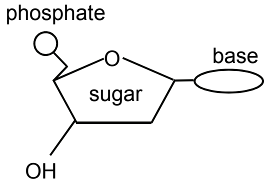

Anaerobic metabolism

$$C_6H_{12}O_6 + 2NAD^+ + 4ADP \rightarrow 2C_3H_4O_3 + 2NADH + 2H^+ + 4ATP$$

Aerobic metabolism

$$C_6H_{12}O_6 + 6O_2 + 36ADP + 36P \rightarrow 6CO_2 + 6H_2O + 36ATP$$

Photosynthesis

$$6CO_2 + 12H_2O + \text{energy} \rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O$$

$NAD^+ + 2H^+ + 2e^- \rightarrow NADH + H^+$ reduction

$NADH + H^+ \rightarrow NAD^+ + 2H^+ + 2e^-$ oxidation

12.1 Sugars

Sugars are part of a larger group of compounds called **carbohydrates**. Carbohydrates have the general formula $C_m(H_2O)_n$ and can be thought of as hydrated carbon compounds.

Glucose, $C_6H_{12}O_6$, an important sugar in biochemistry is an aldehyde .

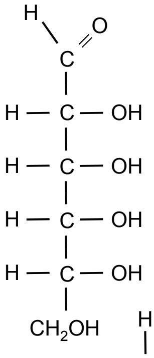

Fructose, an isomer of glucose, $C_6H_{12}O_6$, and is a ketone.

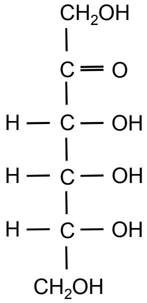

These compounds can readily form cyclic structures, which are often more common than the chain structures.

Glucose and fructose are examples of **monosaccharides**. **Sucrose**, or table sugar, is a **disaccharide** and is formed by the condensation (with the elimination of one water molecule) of one molecule of fructose and one of glucose. The reverse reaction, hydrolysis of sucrose, splits sucrose into glucose and fructose.

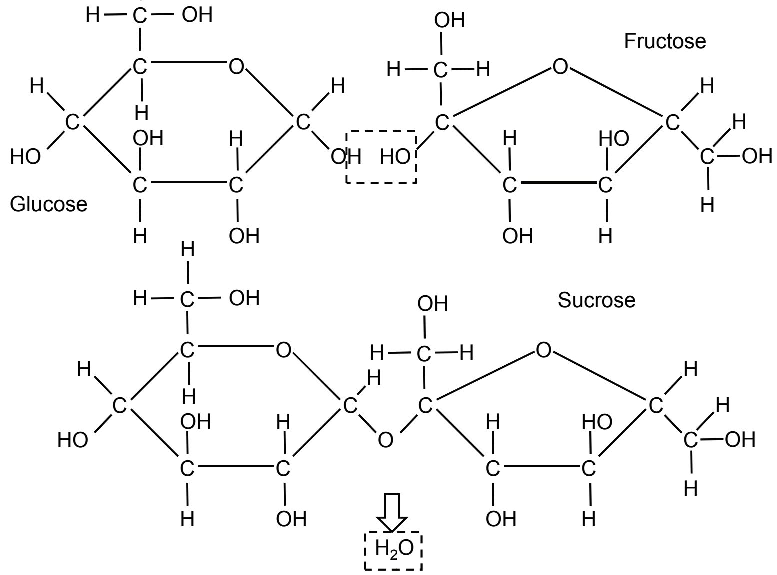

Glucose is the basic fuel for the process of **metabolism**. Glucose is recovered from fats and sugars in the diet and stored as **glycogen** in the body until converted back to glucose for use directly by cells.

12.2 Polysaccharides

Cellulose is a straight-chain polysaccharide $(C_6H_{10}O_5)_n$ consisting of thousands of glucose units. It is insoluble in water. The nature of particular forms of cellulose in organisms depends mainly upon the chain length of the molecule.

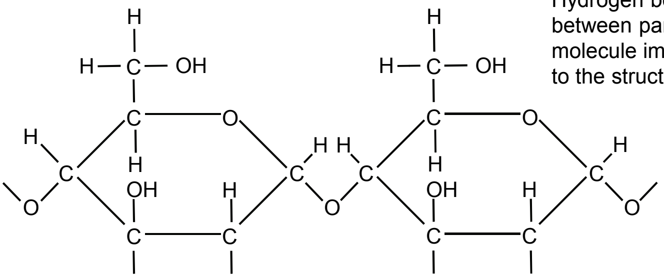

Cellulose

Hydrogen bonding between parts of the molecule impart strength to the structure.

Starch is similar to cellulose in that it is constructed from glucose monosaccharide units, but has a different orientation of linkage between the glucose elements, with more branching than cellulose. It is generally a mixture of glucose polymers of varying lengths, the shorter constituents (of molecular weight ≈4000) being water soluble.

Starch typically contains a mixture of linear (soluble) amylose and branched (relatively insoluble) amylopectin molecules. Complete hydrolysis gives glucose. Partial hydrolysis results in various starch sugars called dextrins.

Starch is broken down into glucose in the body by enzymes, (chiefly amylase, which is a constituent of saliva and pancreatic juice).

Glycogen is similar to the amylopectin form of starch in structure and is found in the liver and muscles of animals as an energy store. It is synthesised from glucose from digestion by enzymes in the liver where it is stored until released as **glucose** into the bloodstream as needed.

Cellulose acts as the main structural element of plants and forms the cell wall that gives plants their rigidity. Wood is about 50% cellulose. Mammals generally are incapable of breaking down cellulose to glucose although cows and sheep, and other ruminants, have the necessary bacteria in their digestive system that perform this function. Insects (e.g. termites) can also digest cellulose.

Starch is used as an energy store in plants. The glucose produced by photosynthesis is stored in plant tissue as starch until it is needed by the plant. Starch in plants, when eaten as food, is an important source of glucose for animals.

12.3 Lipids

Fats and oils are **esters**. Esters are derivatives of **carboxylic acids** where the H on the OH group is replaced by R so that the general formula is RCOOR'. Esters are usually derived from the reaction between acids and alcohols with the elimination of a water molecule.

Animal and plant fats are triple esters, or **triglycerides**, of **glycerol** (**glycerin**) $C_3H_8O_3$. Each of the three available OH groups in glycerol is available for combination with three molecules of **fatty acid**. Fatty acids typically have from 12 to 20 carbon atoms.

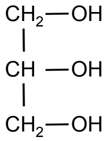

$$\begin{array}{l} CH_2-OH \\ | \\ CH-OH \\ | \\ CH_2-OH \end{array}$$

Glycerol
(propane 1,2,3 triol)

Fats, which are solid at room temperature, are made from **saturated fatty acids**, while **oils**, which are liquid at room temperature, are formed mainly from **unsaturated fatty acids**. Fats and oils from plants and animals typically contain a mixture of different types of esters, two or three fatty acids attached to a glycerol molecule.

Linolenic acid has the general formula

$$CH_3(CH_2)_4\underset{\uparrow}{C}H=CHCH_2CH=CH(CH_2)_7COOH$$

and is called an omega-6 fatty acid because the first C=C double bond is on the sixth carbon atom from the (left-hand) end. **Linoleic acid** is am omega-3 acid. (Note: normally numbering of atom positions is from the right.)

$$CH_3CH_2\underset{\uparrow}{C}H=CHCH_2CH=CHCH_2CH_2=CH(CH_2)_7COOH$$

Linoleic acid

Fats, along with oils and waxes, are one example of a broader group of **lipids**. Compared to carbohydrates, lipids generally contain a smaller proportion of O atoms. Lipids are relatively insoluble in water.

Fat in the body acts as a storage place for glucose. When fat reserves are called upon, the fats are converted back into glycerol and **fatty acids**. The glycerol in turn is converted into glucose by the liver.

Fats in the diet are also a source of **linolenic** $C_{17}H_{29}CO_2H$ and **linoleic** $C_{17}H_{31}CO_2H$ unsaturated fatty acids which are essential in animal dietary intake because they are unable to be synthesised by the body directly and are required for metabolism into a variety of other acids required by various bodily functions.

Fats also serve a structural role in the body by providing heat insulation and a barrier against mechanical shock for organs.

When fats hydrolyse, or react with water, in an alkaline solution, glycerol is formed along with the metal salt of the carboxylic acid. Such a reaction is called **saponification**, from which soaps (e.g. sodium stearate) is formed.

12.4 Proteins

Proteins are amino acids joined together by **amide** (or **peptide**) links and are classified as **polyamides**. The arrangement, or sequence, of amino acids within the protein structure determines the function of the protein. A typical protein may contain 100 or more amino acids.

Amino acids with the amine group attached to the same C atom as the COOH group are called **alpha-amino acids** and are the most important in biochemistry.

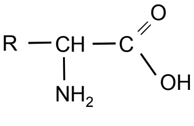

amino acid

The twenty or so amino acids found in proteins differ in the makeup of the R residual. In **glycine**, R is just an H atom, while in phenylaniline, R is a ring structure.

On one end of the chain, there is a free NH_2 group; this is the N terminal. At the other end of the chain, there is a COOH group, the C terminal. In between, water molecules have been eliminated to leave amino acid units, or residues.

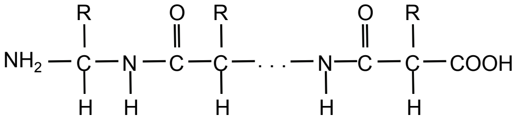

$$NH_2-\underset{H}{\overset{R}{C}}-\underset{H}{N}-\overset{O}{\overset{\|}{C}}-\underset{H}{\overset{R}{C}}-\cdots-\underset{H}{N}-\overset{O}{\overset{\|}{C}}-\underset{H}{\overset{R}{C}}-COOH$$

A typical **polypeptide** protein consists of between 100 and 500 or more amino acid residues. The sequence of amino acids is the **primary structure** of the molecule. The long chains of proteins themselves have a **secondary structure**, usually in the form of a helix or spiral. **Hydrogen bonds** between different amino acid groups are responsible for this structure. Often, there is more than one polypeptide chain present and they exist as intertwined helixes held together by hydrogen or ionic bonds leading to a **tertiary structure**.

When a protein undergoes complete hydrolysis, the amino acids are recovered.

The mechanical structure of proteins (i.e. the sequence of amino acids and hydrogen bonds) has great significance for living organisms. Proteins form the basis of skin, hair, muscle, tendons, and other tissues in the body.

Some proteins are enzymes and act as catalysts for chemical reactions which would otherwise proceed too slowly for use in organisms.

The secondary structure of proteins is stable within narrow temperature and pH ranges. When the secondary structure of a protein is disrupted, either by heat (such as by cooking), or immersion in acids or alkalis, the protein is said to be denatured and the physical properties change markedly.

12.5 Nucleic Acids

Carbohydrates, lipids and proteins make up the bulk of living organisms. A fourth major class of compounds comprise the nucleic acids **DNA** and **RNA**. Nucleic acids are long-chain polymers that consist of smaller units, **nucleotides**, joined together.

In DNA, there are four types of nucleotide. Each nucleotide consists of a 5-carbon sugar (**deoxyribose**), with an attached phosphate, and a nitrogen base. Each type is distinguished by the identity of the nitrogenous bases.

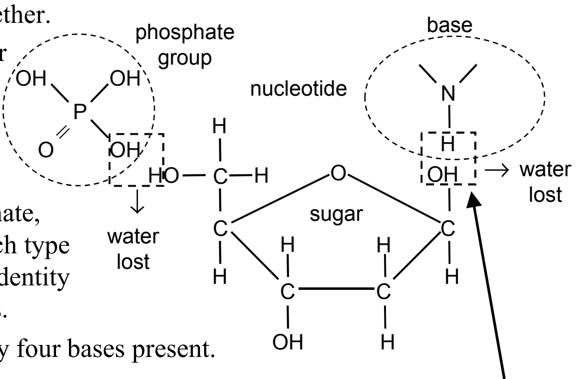
phosphate group
OH
OH
P
O
OH
HO
H
C
H
water lost
nucleotide
base
N
H
OH
water lost
O
sugar
C
C
H
C
H
C
H
OH
H

In DNA, there are only four bases present.

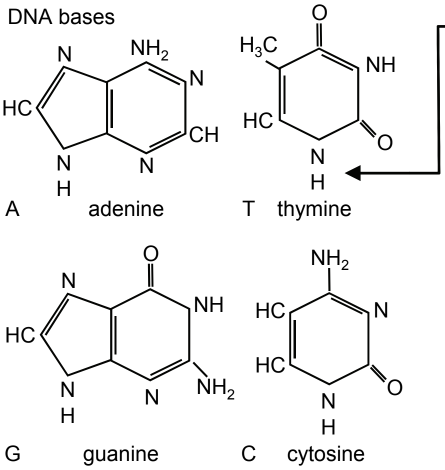
DNA bases
N
NH2
N
HC
CH
N
H
N
A adenine
H3C
O
NH
HC
O
N
H
T thymine
O
N
NH
HC
N
H
N
NH2
G guanine
NH2
HC
N
HC
O
N
H
C cytosine

The NH group on the base combines with the OH group on the sugar and a H_2O is lost.

The primary structure of DNA is a sequence of nucleotides bonded together as long chains. The phosphate group of one nucleotide bonds with the sugar of another (losing a molecule of water in the process).

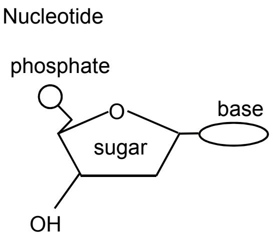
Nucleotide
phosphate
O
base
sugar
OH

A shorthand way of writing these groups is to just show the covalent bonds between the groups, and leave out the C and H atoms.

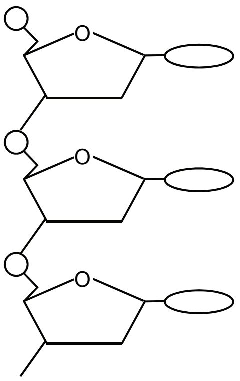

12.6 DNA

The primary structure of DNA is essentially the sequence of bases in the nucleotides in the chain. However, DNA does not ordinarily exist as a single chain, but is paired with an opposite chain so that the bases are paired and held together by **hydrogen bonds**.

The bases can only pair according to certain rules: A is always paired with T by a double hydrogen bond. G is always paired with C by a triple hydrogen bond.

The ladder-like structure is further characterised by being coiled up in a spiral, each chain with bases pointing inwards and bonding with the corresponding base on the other – the so-called **double-helix**. The nucleotides, identified by their bases, can appear in any particular order, but the sequence on one chain has to be reflected in the other by virtue of the **base-pairing rules** A-T and G-C.

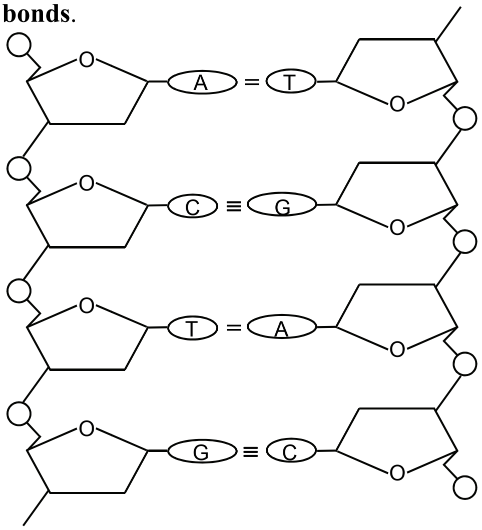

Sequences of nucleotides fall into functional groups called **genes**, which in turn lie along a single large DNA molecule called a **chromosome**. In most cells, chromosomes occur in duplicated pairs. When cell division by **mitosis** occurs, the DNA molecules are replicated by rupturing the hydrogen bonds between the two chains and then each chain forming bonds with new nucleotides according to base-pairing rules. When cell division is complete, the two new cells each contain paired chromosomes of the same genetic sequence as the parent cell.

Although there are only four bases in human DNA, the molecule typically consists of about 1500 nucleotides in the chain, allowing billions of combinations of base sequences possible in a single molecule.

RNA is similar to DNA but consists of ribose as the sugar and a single chain structure with the base **uracil** rather than thymine. There are different types of RNA; some act as messengers, carrying the DNA template to the site of protein synthesis (**ribosomes**) in the cell, while others transport amino acids to the ribosomes.

12.7 Enzymes

The many chemical reactions which take place in an organism to sustain life are collectively called **metabolism**. Typical reactions are **condensation** and **hydrolysis** reactions:

sugars	Condensation	polysaccharides (cellulose, starch)
fatty acids + glycerol	→	fats
	Hydrolysis	
amino acids	←	proteins

These reactions occur at temperatures of about 37°C under atmospheric pressure. The reaction rates would be too slow unless reaction pathways were altered by use of **catalysts**. Nearly every biological chemical reaction involves the use of a specific biological catalyst called an **enzyme**.

Enzymes are very large protein molecules of a specific shape. The enzyme molecule is typically much larger than the molecules actually involved in the chemical reaction. The reaction molecule (or molecules in the case of a condensation reaction), called the substrate, attaches itself to an **active site** on the enzyme molecule.

An **activated complex** is formed, and the reaction proceeds along a different pathway than would normally occur in the laboratory.

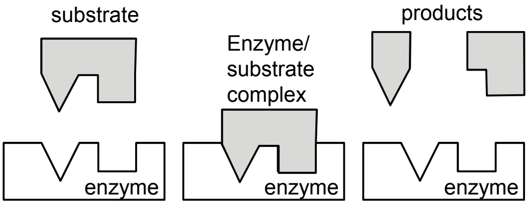

When the reaction is complete, the products detach from the enzyme and the enzyme is available to catalyse another reaction.

Enzymes lower the **activation energy** by increasing the collision frequency of molecules by virtue of bringing the molecules together spatially in an optimum manner, or altering the bond energies involved in the reaction by forming temporary bonds on the reacting molecules, and in some cases, applying mechanical stress to the reacting molecules to facilitate contact.

Enzyme inhibitors are used within the organism to control the activity of enzymes by masking or blocking the active sites on the enzyme to regulate reaction rates as a whole.

The importance of enzymes cannot be over-stated. In a living cell, the reagents of all the chemical reactions are not stored in isolated bottles and mixed when needed as in a laboratory, but exist all together in solution. The process of chemical reactions between these mixed reagents is orchestrated by the action of enzymes.

12.8 ATP

Organisms need nutrients to stay alive. Nutrients comprise carbohydrates, proteins and fats and, in animals, are usually ingested as food. Nutrients provide a source of energy, essential amino acids and essential fatty acids. Vitamins and other trace elements are also required for the proper functioning of enzymes.

Metabolism is the oxidation of glucose to carbon dioxide, water and energy:

$$C_6H_{12}O_6 + 6O_2 + 36ADP + 36P \rightarrow 6CO_2 + 6H_2O + 36ATP$$

Energy stored as phosphate bonds via the conversion of ADP (adenosinediphosphate) to ATP (adenosinetriphosphate)

The energy released by the oxidation process is stored as chemical potential energy in the formation of **ATP** molecules from **ADP** molecules. **Adenosinediphosphate** consists of a molecule of adenine, ribose, and two phosphate groups:

$$ADP + P + \text{energy} \leftrightarrow ATP + H_2O$$

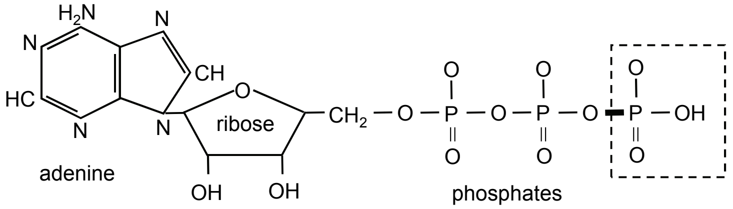

When energy is used to bind a third phosphate group to the end of the molecule, **adenosinetriphosphate** is formed. This is an endothermic process and is called **phosphorylation**. The energy used to create ATP is stored in the phosphate bond. Energy-rich ATP molecules are then hydrolysed to release their energy when used directly in other biological processes such as muscle action, protein synthesis and nerve impulse transmission.

Another important molecule involved in metabolism is the co-enzyme NAD (**nicotinamide adenine dinucleotide phosphate**). NAD takes part in oxidation-reduction reactions in metabolism by losing and accepting electrons ultimately necessary for conversion of ADP to ATP. The redox reactions are

$$NAD^+ + 2H^+ + 2e^- \rightarrow NADH + H^+ \quad \text{reduction}$$

$$NADH + H^+ \rightarrow NAD^+ + 2H^+ + 2e^- \quad \text{oxidation}$$

Reduction can occur by adding hydrogen.
Oxidation can occur by removing hydrogen.

When NAD^+ is reduced, it stores energy by storing an excited electron in the form of NADH and a H^+ ion. When $NADH+H^+$ is oxidised, it releases energy, usually to form ATP from ADP.

12.9 Anaerobic Metabolism

Anabolic processes build up low-energy reactants into high-energy products. Photosynthesis in plants is an anabolic process. **Catabolic** processes break down high-energy reactants to low-energy products, releasing energy in doing so. This energy is transferred to **ATP**.

Anaerobic metabolism involves two steps: (i) the breakdown of glucose from a **carbohydrate** into **pyruvic acid**:

$$C_6H_{12}O_6 + 2NAD^+ + 4ADP \rightarrow 2C_3H_4O_3 + 2NADH + 2H^+ + 4ATP$$

and (ii) the fermentation of pyruvic acid into either CO_2 and ethanol, or lactic acid. The ultimate aim is to produce ATP from ADP.

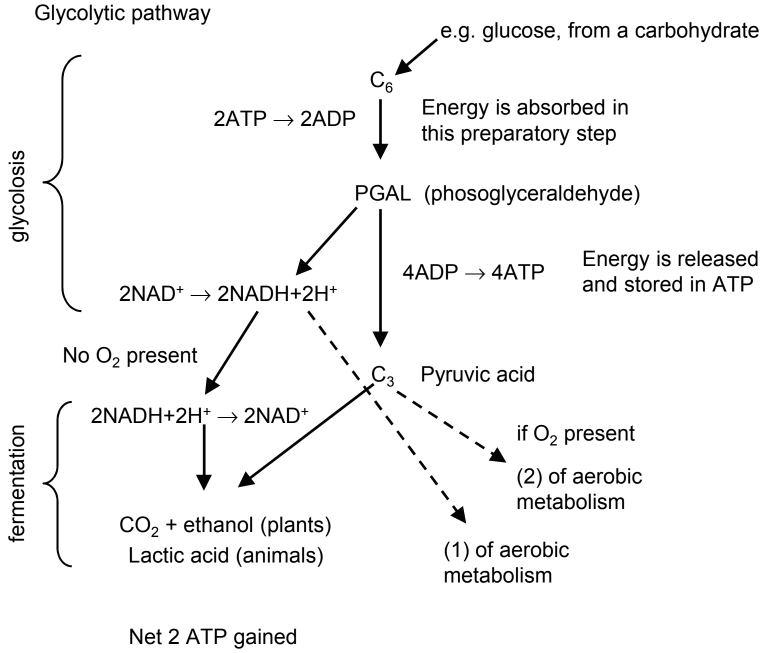
Glycolytic pathway
e.g. glucose, from a carbohydrate
glycolosis
C_6
2ATP → 2ADP
Energy is absorbed in this preparatory step
PGAL (phosoglyceraldehyde)
4ADP → 4ATP
Energy is released and stored in ATP
$2NAD^+ \rightarrow 2NADH+2H^+$
No O_2 present
C_3 Pyruvic acid
fermentation
$2NADH+2H^+ \rightarrow 2NAD^+$
if O_2 present
(2) of aerobic metabolism
CO_2 + ethanol (plants)
Lactic acid (animals)
(1) of aerobic metabolism
Net 2 ATP gained

Anaerobic metabolism occurs in all cells. It does result in a particularly high yield of ATP since the products, ethanol, lactic acid and pyruvic acid, still contain a substantial portion of the original energy from the glucose. While anaerobic metabolism can provide ATP for energy use (in muscles) when O_2 is in short supply, or none at all, it also provides a source of **pyruvic acid** for use in **aerobic respiration**.

12.10 Aerobic Metabolism

Aerobic metabolism, or **respiration**, is the process by which glycerol is oxidised with the production of ATP molecules from ADP molecules in the presence of molecular oxygen:

from glycolysis

$$\text{(1)} \quad 2NADH + 2H^+ + O_2 \rightarrow 2NAD^+ + 2H_2O \quad + \rightarrow 6\text{ ATP}$$

$$\text{(2)} \quad \underset{\text{Pyruvic acid}}{2C_3} + \overset{\text{Coenzyme A}}{2CoA} + 2NAD^+ \rightarrow \overset{C_2}{2AcetylCoA} + 2NADH + 2H^+ + 2CO_2$$

$$\text{(3)} \quad 2NADH + 2H^+ + O_2 \rightarrow 2NAD^+ + 2H_2O \quad + \rightarrow 6\text{ ATP}$$

$$\text{(4)} \quad 2AcetylCoA + \text{Krebs cycle} \rightarrow CO_2 + H_2O \quad + \rightarrow 2\text{ ATP directly} + NADH{+}H^+$$

(1) The NADH from **glycolysis** is oxidised, by the addition of O_2, to produce six ATP.

(2) **Pyruvic acid** C_3 is oxidised to an activated form of acetic acid C_2. NAD^+ is reduced to $NADH{+}H^+$.

(3) $NADH{+}H^+$ from (2) is oxidised by O_2, to form six ATP.

(4) **AcetylCoA** enters the **Krebs cycle** (or **citric acid cycle**). This consists of multiple reactions, the net effect of which is to form two molecules of ATP plus energy-rich NADH.

During the Krebs cycle, six molecules of $NADH{+}H^+$ and two molecules of a related compound, FADH, are formed. These electron-carrier molecules are energy-rich and can be further oxidised as in (1) and (3) which require O_2. The net result of (4) is the production of twenty two ATP molecules. **Anaerobic metabolism** produces two molecules of ATP, which added to the thirty four ATP molecules from aerobic metabolism makes thirty six molecules of ATP from the complete oxidation of one molecule of **glucose**.

Fats and **proteins** are also sources of ATP, by virtue of their breakdown into components which can be inserted into glycolytic and aerobic pathways.

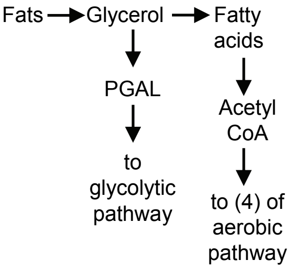

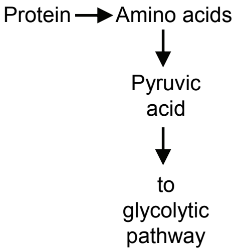

12.11 Cyclic Photophosphorylation

Ultimately, the energy contained within glucose used in animal metabolism comes from **photosynthesis** in plants. The overall reaction in plants, for the production of glucose, is

$$6CO_2 + 12H_2O + \overset{\overset{\text{light}}{\downarrow}}{\text{Energy}} \rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O$$

Note that O_2 is a product of this reaction.

Chlorophyll is the unique ingredient of plants that captures photons to promote an electron of low energy into an excited state. The excited electron is passed from one transfer molecule to another, losing energy at each step, until it returns to a chlorophyll molecule in an unexcited state. At each step, the energy lost is used to convert ADP to ATP:

$$\text{ADP} + \text{P} + \text{Energy} \leftrightarrow \text{ATP} + H_2O$$

As the excited electron is passed from molecule to molecule, *some* of the energy is used to create ATP. The process is not 100% efficient.

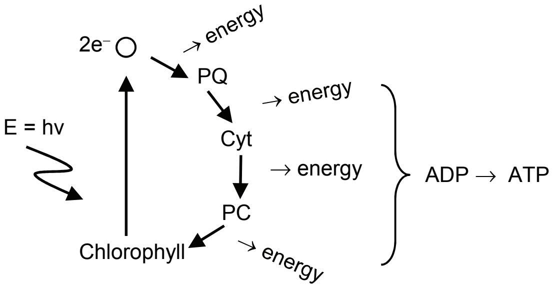

PQ, Cyt and PC are electron carrier molecules that undergo reduction (when they accept the electron) and oxidation (when they donate the electron).

Because electrons are donated by chlorophyll from the high-energy state and then to acceptance in a low-energy state, this synthesis of ATP is called **cyclic photophosphorylation**. The ATP, essentially created from ADP by the energy of sunlight, is ultimately used to create high-energy glucose (which is stored in the plant as starch) from low-energy CO_2. Animals eat the plants and convert starch into glucose, which is then used to create ATP for their own metabolic processes.

In this mode of photophosphorylation, electrons are transported around in a cycle; no oxygen or NADH+H^+ is produced.

12.12 Non-Cyclic Photophosphorylation

Another way of producing ATP used by plants has a beginning similar to cyclic photophosphorylation, but employs a more complicated path for extraction of energy from excited electrons. This non-cyclic photophosphorylation is used by modern green plants :

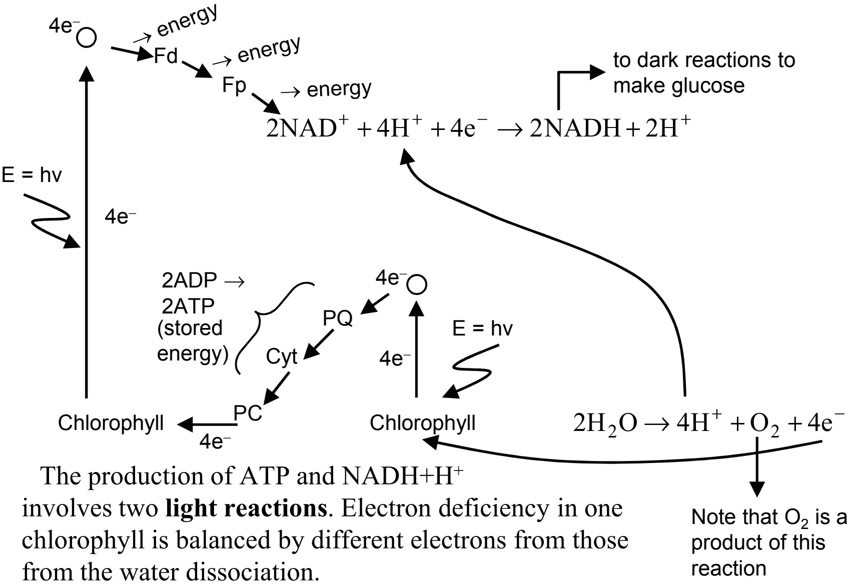
4e⁻
→ energy
Fd
→ energy
Fp
→ energy
to dark reactions to make glucose
$2NAD^+ + 4H^+ + 4e^- \rightarrow 2NADH + 2H^+$
E = hv
4e⁻
2ADP →
2ATP
(stored
energy)
4e⁻
PQ
Cyt
PC
E = hv
4e⁻
Chlorophyll
4e⁻
Chlorophyll
$2H_2O \rightarrow 4H^+ + O_2 + 4e^-$
Note that O_2 is a product of this reaction

The production of ATP and NADH+H$^+$ involves two **light reactions**. Electron deficiency in one chlorophyll is balanced by different electrons from those from the water dissociation.

Additional reactions take place in which the energy obtained from the light reactions is converted into carbohydrates for storage. That is, CO_2 is reduced, and (if the energy is coming from NADH+H$^+$) the NDAH+H$^+$ is oxidised with the net result being glucose. These are known as **dark reactions** since no light is required.

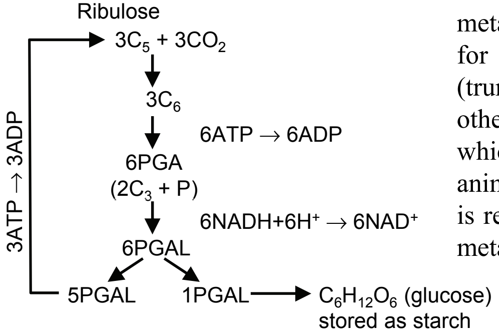
Ribulose
$3C_5 + 3CO_2$
$3C_6$
6ATP → 6ADP
6PGA
$(2C_3 + P)$
6NADH+6H$^+$ → 6NAD$^+$
6PGAL
5PGAL
1PGAL
$C_6H_{12}O_6$ (glucose) stored as starch
3ATP → 3ADP

The resulting glucose is used in metabolism by the plant cells for making cellulose/starch (trunks and branches, etc) and other sugars (such as in fruit), which may in turn be eaten by animals from which the glucose is recovered and used in animal metabolism.

12.13 Metabolism

Summary:

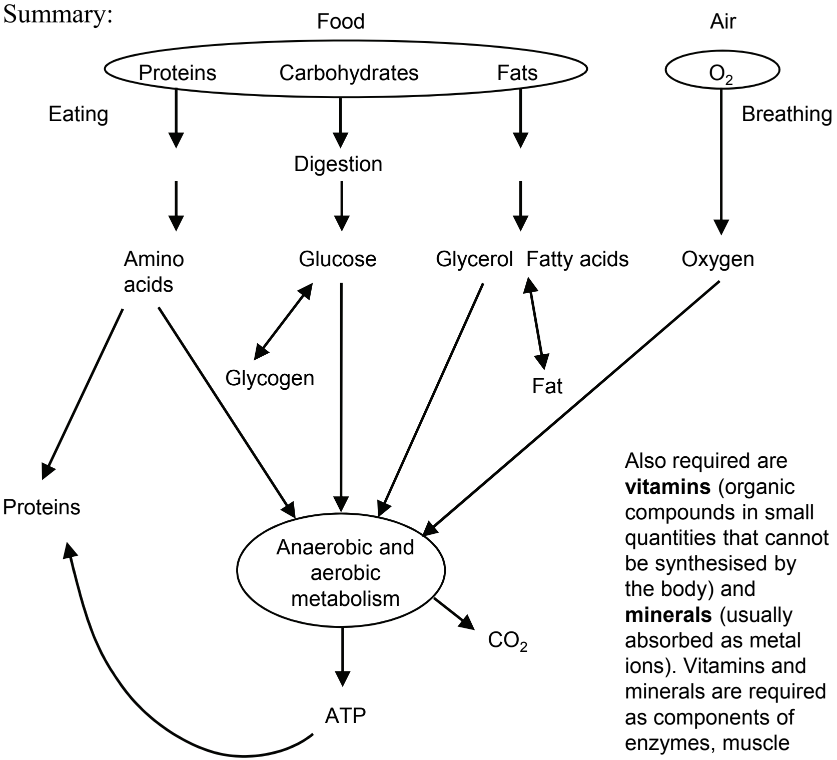
Food
Air
Proteins
Carbohydrates
Fats
O_2
Eating
Breathing
Digestion
Amino acids
Glucose
Glycerol
Fatty acids
Oxygen
Glycogen
Fat
Proteins
Anaerobic and aerobic metabolism
CO_2
ATP

Also required are **vitamins** (organic compounds in small quantities that cannot be synthesised by the body) and **minerals** (usually absorbed as metal ions). Vitamins and minerals are required as components of enzymes, muscle action, bones, teeth, haemoglobin, etc.

ATP is used to provide the energy source for conversion of amino acids into proteins required by the body for the production of skin, hair, cells, enzymes, muscles, nerves, signalling, anaerobic respiration, as well as for energising DNA replication. It also provides energy for muscle action, vision, brain activity, and nearly all biological energy transformations. Glycogen and fats can be stored in the body and sent back into the metabolism pathway as glucose and glycerol for oxidation into CO_2 and energy when there is no food being digested.

This page intentionally left blank

Chemistry

Like the author's other companion books, **The Chemistry Companion** provides high quality information in unique one-page-per-topic presentations that do not overburden and distract readers with excessive details. The book offers concise summaries of general chemistry concepts, easily accessible in a convenient, reader-friendly format.

Suitable as an introduction or study guide, this companion presents the minimum of what readers need to know to understand the subject. It emphasizes the physics underlying chemistry. By looking at chemistry processes from a physics point of view, readers better appreciate what is happening from the chemical perspective that is usually found in traditional chemistry books.

The author focuses on the structure of matter, chemical components and bonds, the periodic table, states of matter, thermodynamics, reaction rates, carbon chemistry, biochemistry, and chemical, ionic, and electronic equilibria. Each topic is covered in a single-page outline format with just enough detail to enable a good understanding of the subject.

Features

- Provides a physical understanding of chemical concepts
- Presents clear explanations of difficult material, working through any inconsistencies in understanding
- Uses a convenient format for checking formulas and definitions
- Includes self-contained information on each page, assuming little prior knowledge

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K11517

ISBN: 978-1-4398-3088-8
90000
9 781439 830888

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press

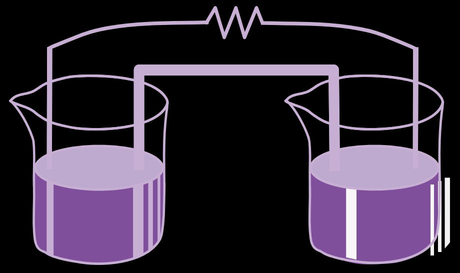

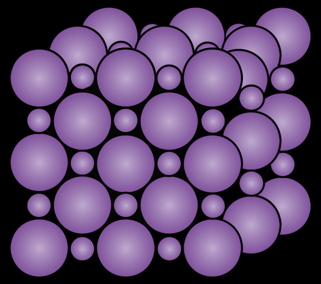

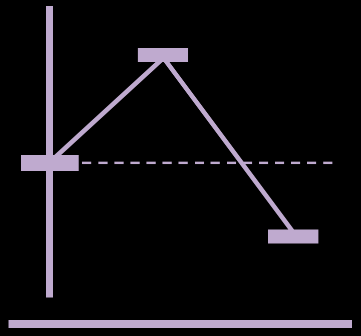

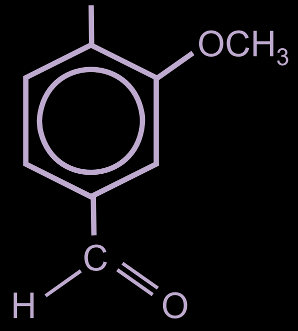

The Chemistry Companion

A C FISCHER-CRIPPS

CRC Press
Taylor & Francis Group

United States Patent [19]

Rogers, Sr.

[11] **4,292,804**

[45] **Oct. 6, 1981**

[54] METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS

[76] Inventor: **Leroy K. Rogers, Sr.,** #5 Capistrano Ct., Ft. Myers, Fla. 33908

[21] Appl. No.: **158,303**

[22] Filed: **Jun. 10, 1980**

[51] **Int. Cl.**^{3} **F15B 11/06**

[52] **U.S. Cl.** **60/407;** 91/187; 91/275

[58] **Field of Search** 60/407, 412; 91/187, 91/275, 364

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,881,399	5/1975	Sagi et al.	91/187 X
3,885,387	5/1975	Simington	60/407 X
4,018,050	4/1977	Murphy	60/412 X

Primary Examiner—Allen M. Ostrager
Attorney, Agent, or Firm—Burns, Doane, Swecker & Mathis

[57] ABSTRACT

The present invention relates to a method and apparatus for operating an engine having a cylinder and a piston reciprocable therein on compressed gas. The apparatus comprises a source of compressed gas connected to a distributor which distributes the compressed gas to the cylinder. A valve is provided to selectively admit compressed gas to the cylinder when the piston is in an approximately top dead center position. In one embodiment of the present invention the timing of the opening of the valve is advanced such that the compressed gas is admitted to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases. In a further embodiment of the present invention a valve actuator is provided which increases the length of time over which the valve remains open to admit compressed gas to the cylinder as the speed of the engine increases. A still further embodiment of the present invention relates to an apparatus for adapting a conventional internal combustion engine for operation on compressed gas.

22 Claims, 8 Drawing Figures

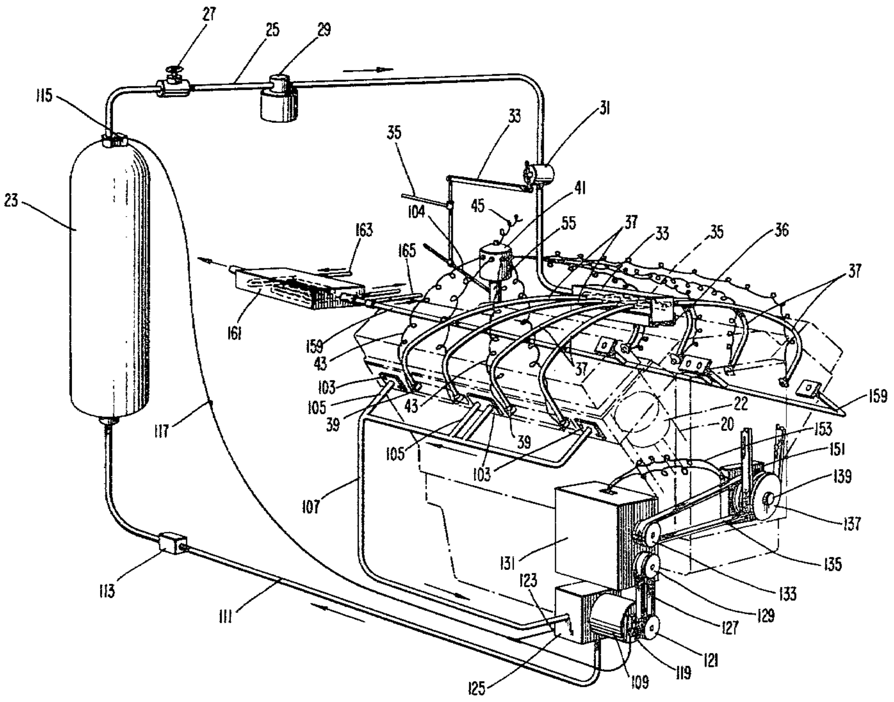

U.S. Patent Oct. 6, 1981 Sheet 1 of 3 4,292,804

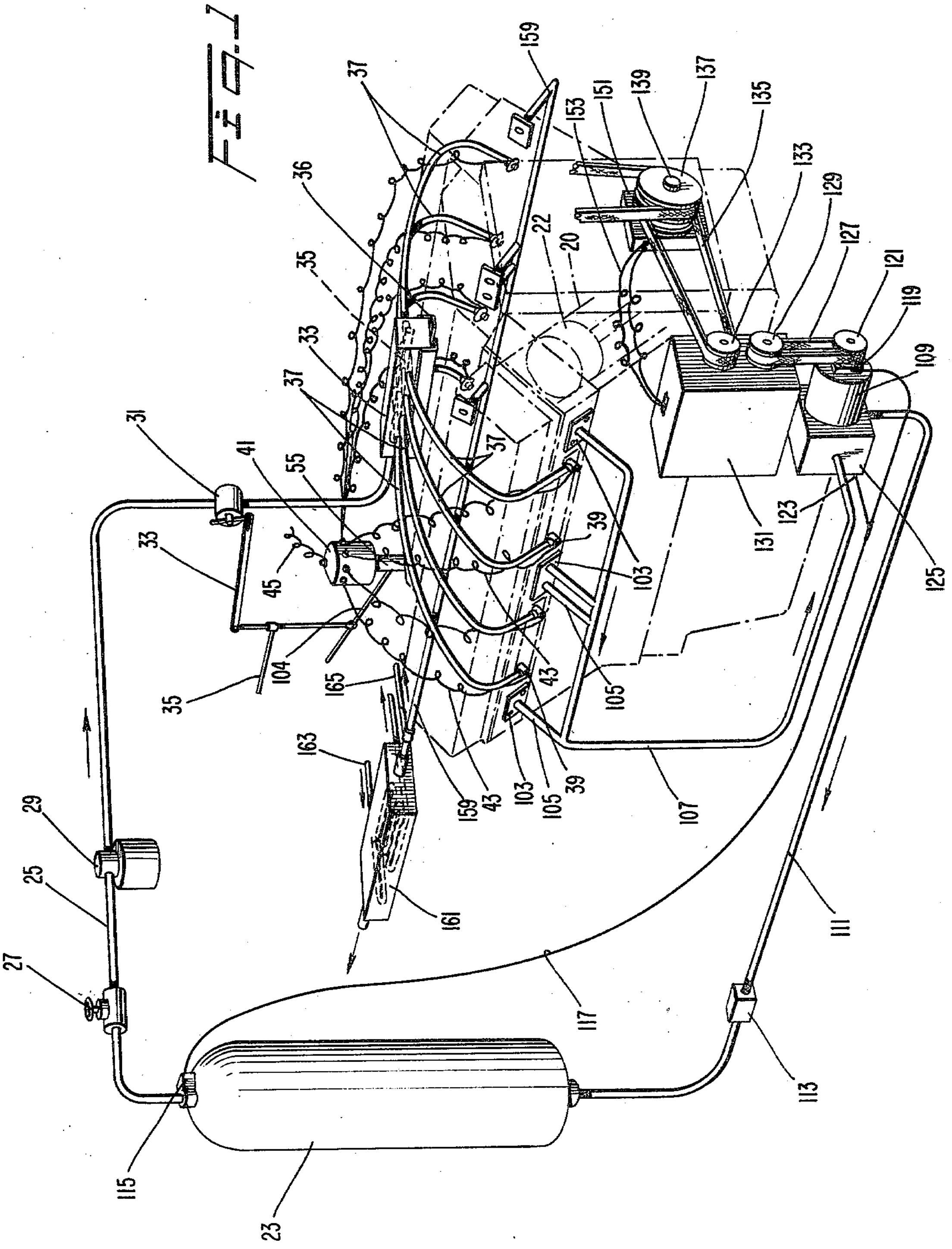
FIG. 1

159
37
36
35
33
37
37
31
41
55
45
33
35
104
165
163
29
25
27
159
43
103
105
39
107
161
117
111
113
115
23
22
20
153
151
139
137
135
133
129
127
121
119
109
131
123
125
103
39
105
43

Fig. 6

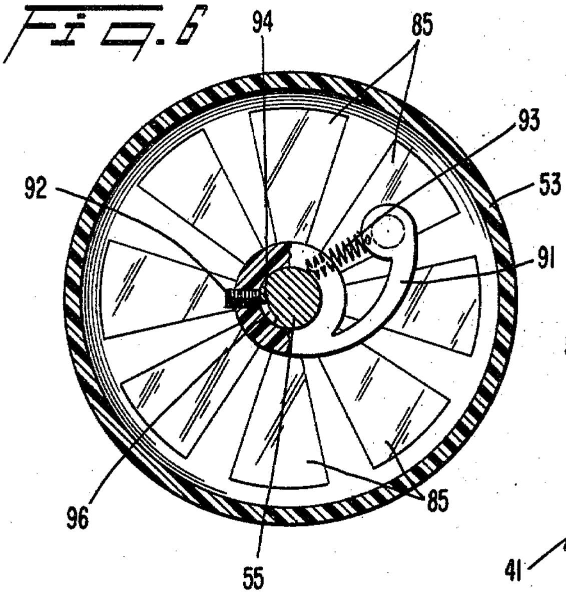
94
85
93
92
53
91
96
85
55

Fig. 7

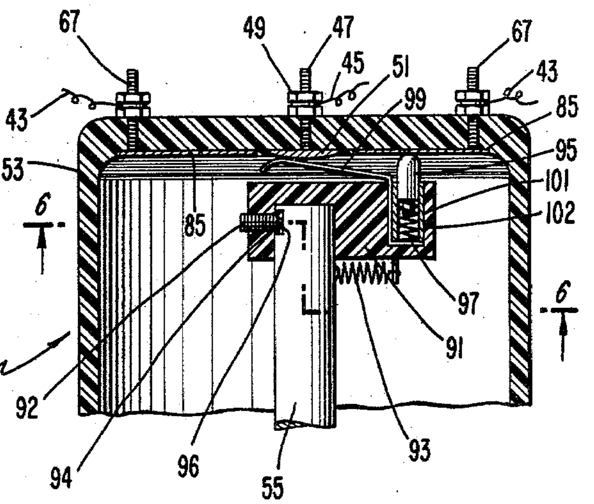
67
49
47
45
51
67
43
43
99
85
53
95
101
102
6
85
6
97
91
41
92
93
94
96
55

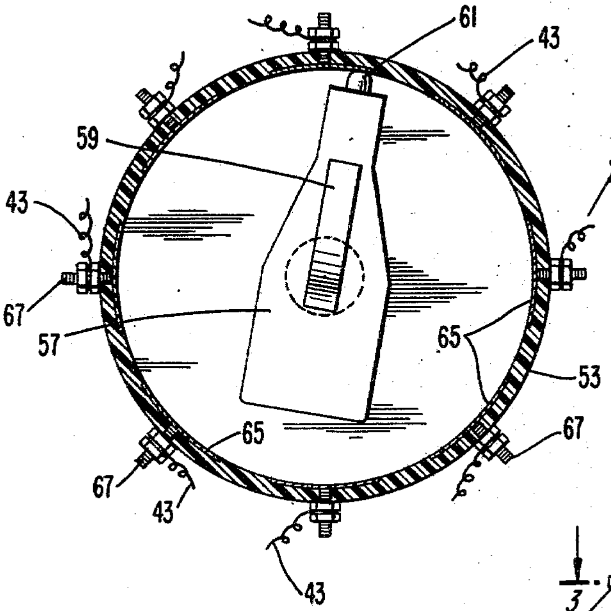
61
43
59
43
67
57
65
53
65
67
67
43
43

Fig. 3

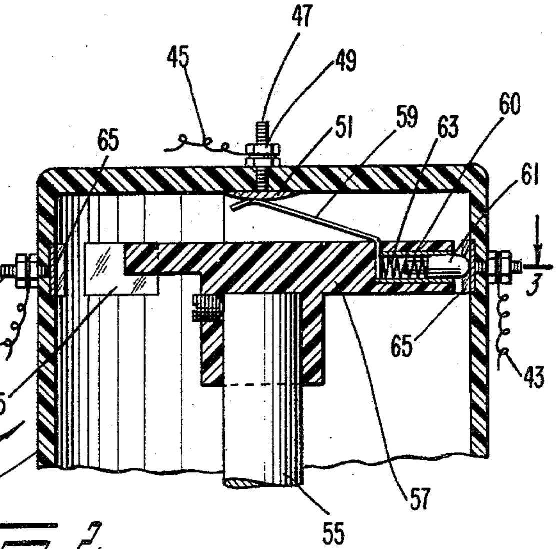
47
45
49
51
59
63
60
65
61
3
3
67
65
43
43
65
41
53
55
57

Fig. 2

Fig. 5

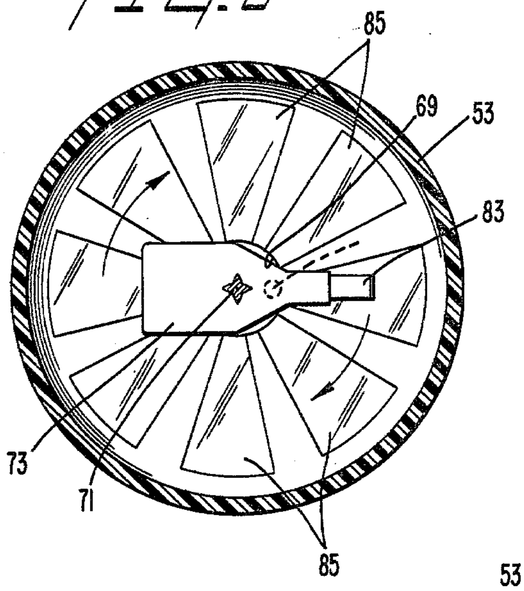
85
69
53
83
73
71
85

Fig. 4

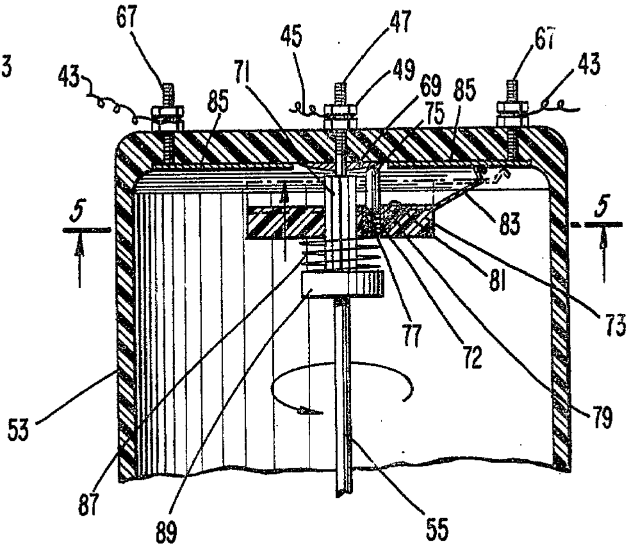
67
45
47
67
43
71
49
69
85
43
85
75
5
5
83
81
73
77
72
53
79
87
89
55

Fig. 6

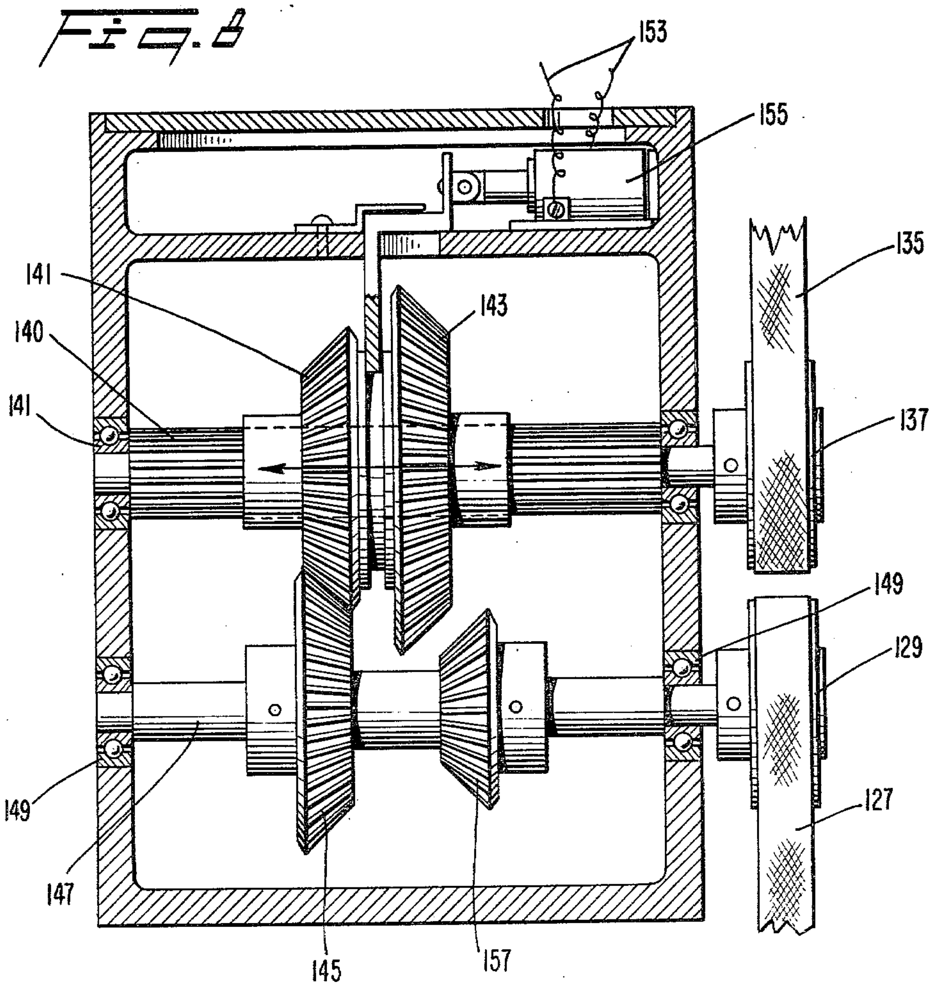
153
155
135
141
143
140
141
137
149
129
149
127
147
145
157

METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS

BACKGROUND AND SUMMARY OF THE PRESENT INVENTION

The present invention relates to a method and apparatus for operating an engine using a compressed gas as the motive fluid. More particularly, the present invention relates to a apparatus for adapting a pre-existing internal combustion engine for operation on a compressed gas.

Air pollution is one of the most serious problems facing the world today. One of the major contributors to air pollution is ordinary internal combustion engine which are used in most motor vehicles today. Various devices, including many items mandated by legislation, have been proposed in an attempt to limit the pollutants which an internal combustion engine exhausts to the air. However, most of these devices have met with limited success and are often both prohibitively expensive and complex. A clean alternative to the internal combustion engine is needed to power vehicles and other machinery.

A compressed gas, preferably air, would provide an ideal motive fluid for a engine since it would eliminate the usual pollutants exhausted from an internal combustion engine. An apparatus for converting an internal combustion engine for operation on compressed air is disclosed in U.S. Pat. No. 3,885,387 issued May 27, 1975 to Simington. The Simington patent discloses an apparatus including a source of compressed air and a rotating valve actuator which opens and closes a plurality of mechanical poppet valves. The valves deliver compressed air in timed sequence to the cylinders of an engine through adapters located in the spark plug holes. However, the output speed of an engine of this type is limited by the speed of the mechanical valves and the fact that the length of time over which each of the valves remains open cannot be varied as the speed of the engine increases.

Another apparatus for converting an internal combustion engine for operation on steam or compressed air is disclosed in U.S. Pat. No. 4,102,130 issued July 25, 1978 to Stricklin. The Stricklin patent discloses a device which changes the valve timing of a conventional four stroke engine such that the intake and exhaust valves open once for every revolution of the engine instead of once every other revolution of the engine. A reversing valve is provided which delivers live steam or compressed air to the intake valves and is subsequently reversed to allow the exhaust valves to deliver the expanded steam or air to the atmosphere. A reversing valve of this type however does not provide a reliable apparatus for varying the amount of motive fluid injected into the cylinders when it is desired to increase the speed of the engine. Further, a device of the type disclosed in the Stricklin patent requires the use of multiple reversing valves if the cylinders in a multi-cylinder engine were to be fired sequentially.

Therefore, it is an object of the present invention to provide a reliable method and apparatus for operating an engine or converting an engine for operation with a compressed gas.

A further object of the present invention is to provide a method and apparatus which is effective to deliver a constantly increasing amount of compressed gas to an engine as the speed of the engine increases.

A still further object of the present invention is to provide a method and apparatus which will operate an engine using compressed gas at a speed sufficient to drive a conventional automobile at highway speeds.

It is still a further object of the present invention to provide a method and apparatus which is readily adaptable to a standard internal combustion engine to convert the internal combustion engine for operation with a compressed gas.

Another object of the invention is to provide a method and apparatus which utilizes cool expanded gas, exhausted from a compressed gas engine, to operate an air conditioning unit and/or an oil cooler.

These and other objects are realized by a method and apparatus according to the present invention for operating an engine having at least one cylinder and a reciprocating piston therein using compressed gas as a motive fluid. The apparatus includes a source of compressed gas and a distributor connected with the source of the compressed gas for distributing the compressed gas to the at least one cylinder. A valve is provided for admitting the compressed gas to the cylinder when the piston is in approximately a top dead center position within the cylinder. An exhaust is provided for exhausting the expanded gas from the cylinder as the piston returns to approximately the top dead center position.

In a preferred embodiment of the present invention a device is provided for varying the duration of each engine cycle over which the valve remains open to admit compressed gas to the cylinder dependent upon the speed of the engine. In a further preferred embodiment of the present invention, an apparatus for advancing the timing of the opening of the valve is arranged to admit the compressed gas to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases.

Further features of the present invention include a valve for controlling the amount of compressed gas admitted to the distributor. Also, a portion of the gas which has been expanded in the cylinder and exhausted through the exhaust valve is delivered to a compressor to be recompressed and returned to the source of compressed gas. A gear train is selectively engagable to drive the compressor at different operating speed depending upon the pressure maintained at the source of compressed air and/or the speed of the engine. Still further, a second portion of the exhaust gas is used to cool a lubricating fluid for the engine or to operate an air conditioning unit.

In a preferred embodiment of the present invention, the valve for admitting compressed gas to the cylinder is electrically actuated. The device for varying the duration of each engine cycle over which the intake valve remains open as the speed of the engine increase comprises a rotating element whose effective length increases as the speed of the engine increases such that a first contact on the rotating element is electrically connected to a second contact for a longer period of each engine cycle. The second contact actuates the valve whereby the valve remains in an open position for a longer period of each engine cycle as the speed of the engine increases.

Still further features of the present invention include an adaptor plate for supporting the distributor above an intake manifold of a conventional internal combustion engine after a carburetor has been removed to allow air to enter the cylinders of the engine through the intake manifold and conventional intake valves. Another adaptor plate is arranged over an exhaust passageway of the internal combustion engine to reduce the cross-sectional area of the exhaust passageway.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of a method and apparatus for operating an engine according to the present invention will be described with reference to the accompanying drawings wherein like members bear like reference numerals and wherein:

FIG. 1 is a schematic representation of an apparatus according to the present invention arranged on an engine;

FIG. 2 is a side view of one embodiment of a valve actuator according to the present invention;

FIG. 3 is a cross-sectional view taken along the line 3—3 in FIG. 2;

FIG. 4 is a cross-sectional view of a second embodiment of a valve actuator according to the present invention;

FIG. 5 is a view taken along the line 5—5 in FIG. 4;

FIG. 6 is a cross-sectional view of a third embodiment of a valve actuator according to the present invention;

FIG. 7 is a view taken along the line 7—7 in FIG. 6;

FIG. 8 is a cross-sectional view of a gearing unit to drive a compressor according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

With reference to FIG. 1, an engine block **21** (shown in phantom) having two banks of cylinders with each bank including cylinders **20** having pistons **22** reciprocable therein (only one of which is shown in phantom) in a conventional manner. While the illustrated engine is a V-8 engine, it will be apparent that the present invention is applicable to an engine having any number of pistons and cylinders with the V-8 engine being utilized for illustration purposes only. A compressed gas tank **23** is provided to store a compressed gas at high pressure. It may also be desirable to include a small electric or gas compressor to provide compressed gas to supplement the compressed gas held in the tank **23**. In a preferred embodiment, the compressed gas is air which can be obtained from any suitable source.

A line **25** transports the gas withdrawn from the tank **23** when a conventional shut off valve **27** is open. In addition, a solenoid valve **29** preferably operated by a suitable key operated switch (not shown) for the engine is also arranged in the line **25**. In normal operation, the valve **27** is maintained open at all times with the solenoid valve **29** operating as a selective shut off valve to start and stop the engine **21** of the present invention.

A suitable regulating valve **31** is arranged downstream from the solenoid valve **29** and is connected by a linkage **33** to a throttle linkage **35** which is operator actuated by any suitable apparatus such as a foot pedal (not shown). The line **25** enters an end of a distributor **33** and is connected to an end of a pipe **35** which is closed at the other end. A plurality of holes, which are equal to the number of cylinders in the engine **21**, are provided on either side of the pipe **35** along the length of the pipe **35**.

When the present invention is used to adapt a conventional internal combustion engine for operation on compressed gas, an adaptor plate **36** is provided to support the distributor **33** in spaced relation from the usual intake opening in the intake manifold of the engine after a conventional carburetor has been removed. In this way, air is permitted to enter the internal combustion engine through the usual passageways and to be admitted to the cylinders through suitable intake valves (not shown). The adaptor plate **36** is secured to the engine block **21** and the distributor **33** by any suitable apparatus, e.g., bolts.

Each of the holes in the pipe **35** is connected in fluid-tight manner to a single line **37**. Each line **37** carries the compressed gas to a single cylinder **20**. In a preferred embodiment, each of the lines **37** is $\frac{1}{2}$ inch high pressure plastic tubing attached through suitable connectors to the distributor **33** and the pipe **35**. Each of the lines **37** is connected to a valve **39** which is secured in an opening provided near the top of each of the cylinders **20**. In the case of a conversion of a standard internal combustion engine, the valves **39** can be conveniently screwed into a tapped hole in the cylinder **20** typically provided for a spark plug of the internal combustion engine. In a preferred embodiment, the valves **39** are solenoid actuated valves in order to provide a fast and reliable opening and closing of the valves **39**.

Each of the valves **39** is energized by a valve actuator **41** through one of a plurality of wires **43**. The valve actuator **41** is driven by a shaft of the engine similar to the drive for a conventional distributor of an internal combustion engine. That is, a shaft **55** of the valve actuator **41** is driven in synchronism with the engine **21** at one half the speed of the engine **21**.

A first embodiment of the valve actuator **41** (FIGS. 2 and 3) receives electrical power through a wire **45** which is energized in a suitable manner by a battery, and a coil if necessary (not shown) as is conventional in an internal combustion engine. The wire **45** is attached to a central post **47** by a nut **49**. The post **47** is connected to a conducting plate **51** arranged within a housing **53** for the valve actuator **41**. Within the housing **53**, the shaft **55** has an insulating element **57** secured to an end of the shaft **55** for co-rotation therewith when the shaft **55** is driven by the engine **21**. A first end of a flexible contact **59** is continuously biased against the conducting plate **51** to receive electricity from the battery or another suitable source. A second end of the contact **59** is connected to a conducting sleeve **60** which is in constant contact with a spring biased contact **61** which is arranged within the sleeve **60**. The contact **61** is biased by a spring **63** which urges the contact **61** towards a side wall of the housing **53**.

With reference to FIG. 3, a plurality of contacts **65** are spaced from one another and are arranged around the periphery of the housing **53** at the same level as the spring biased contact **61**. Each contact **65** is electrically connected to a post **67** which extends outside of the housing **53**. The number of contacts **65** is equal to the number of cylinders in the engine **21**. One of the wires **43**, which actuate the valves **39**, is secured to each of the posts **67**.

In operation, as the shaft **55** rotates in synchronism with the engine **21**, the insulating element **57** rotates and electricity is ultimately delivered to successive ones of the contacts **65** and wires **43** through the spring biased contact **61** and the flexible contact **59**. In this way, each of the electrical valves **39** is actuated and opened in the proper timed sequence to admit compressed gas to each of the cylinders **20** to drive the pistons **22** therein on a downward stroke.

The embodiment illustrated in FIGS. 2 and 3 is effective to actuate each of the valves 39 to remain open for a long enough period of time to admit sufficient compressed gas to each of the cylinders 20 of the engine 21 to drive the engine 21. The length of each of the contacts 65 around the periphery of the housing 53 is sufficient to permit the speed of the engine to be increased when desired by the operator by moving the throttle linkage 35 which actuates the linkage 33 to further open the regulating valve 31 to admit more compressed gas from the tank 23 to the distributor 33. However, it has been found that the amount of air admitted by the valves 39 when using the first embodiment of the valve actuator 41 (FIGS. 2 and 3) is substantially more than required to operate the engine 21 at an idling speed. Therefore, it may be desirable to provide a valve actuator 41 which is capable of varying the duration of each engine cycle over which the solenoid valves 39 are actuated, i.e., remain open to admit compressed gas, as the speed of the engine 21 is varied.

A second embodiment of a valve actuator 41 which is capable of varying the duration of each engine cycle over which each of the valves 39 remains open to admit compressed gas to the cylinders 20 dependent upon the speed of the engine 21 will be described with reference to FIGS. 4 and 5 wherein members corresponding to those of FIGS. 2 and 3 bear like reference numerals. The wire 45 from the electrical source is secured to the post 47 by the nut 49. The post 47 has a annular contact ring 69 electrically connected to an end of the post 47 and arranged within the housing 53. The shaft 55 rotates at one half the speed of the engine as in the embodiment of FIGS. 2 and 3.

At an upper end of the shaft 55, a splined section 71 slidably receives an insulating member 73. The splined section 71 of the shaft 55 positively holds the insulating member 73 for co-rotation therewith but permits the insulating member 73 to slide axially along the length of the splined section 71. Near the shaft 55, a conductive sleeve 72 is arranged in a bore 81 in an upper surface of the insulating element 73 generally parallel to the splined section 71. A contact 75, biased towards the annular contact ring 69 by a spring 77, is arranged within the conductive sleeve 72 in contact therewith. The conductive sleeve 72 also contacts a conductor 79 at a base of the bore 81.

The conductor 79 extends to the upper surface of the insulating element 73 near an outer periphery of the insulating element 73 where the conductor 79 is electrically connected to a flexible contact 83. The flexible contact 83 selectively engages a plurality of radial contacts 85 arranged on an upper inside surface of the housing 53. A weak spring 87 arranged around the splined section 71 engages a stop member 89 secured on the shaft 55 and the insulating element 73 to slightly bias the insulating element 73 towards the upper inside surface of the housing 53 to ensure contact between the flexible contact 83 and the upper inside surface of the housing 53. As best seen in FIG. 5, the radial contacts 85 on the upper inside surface of the housing 53 are arranged generally in the form of radial spokes extending from the center of the housing 53 with the number of contacts being equal to the number of cylinders 20 in the engine 21. The number of degrees covered by each of the radial contacts 85 gradually increases as the distance from the center of the upper inside surface of the housing 53 increases.

In operation of the device of FIGS. 4 and 5, as the shaft 55 rotates, electricity flows along a path through the wire 45 down through post 47 to the annular contact member 69 which is in constant contact with the spring biased contact 75. The electrical current passes through the conductive sleeve 72 to the conductor 79 and then to the flexible contact 83. As the flexible contact 83 rotates along with the insulating member 73 and the shaft 55, the tip of the flexible contact 83 successively engages each of the radial contacts 85 on the upper inside of the housing 53. As the speed of the shaft 55 increases, the insulating member 73 and the flexible contact 83 attached thereto move upwardly along the splined section 71 of the shaft 55 due to the radial component of the splines in the direction of rotation under the influence of centrifugal force. As the insulating member 73 moves upwardly, the flexible contact 83 is bent such that the tip of the contact 83 extends further radially outwardly from the center of the housing 53 (as seen in phantom lines in FIG. 4). In other words, the effective length of the flexible contact 83 increases as the speed of the engine 21 increases.

As the flexible contact 83 is bent and the tip of the contact 83 moves outwardly, the tip remains in contact with each of the radial contacts 85 for a longer period of each engine cycle due to the increased angular width of the radial contacts with increasing distance from the center of the housing 53. In this way, the length of time over which each of the valves 39 remains open is increased as the speed of the engine is increased. Thus, a larger quantity of compressed gas or air is injected into the cylinders as the speed increases. Conversely, as the speed decreases and the insulating member 73 moves downwardly along the splined section 71, a minimum quantity of air is injected into the cylinder due to the shorter length of the individual radial contact 85 which is in contact with the flexible contact 83. In this way, the amount of compressed gas that is used during idling of the engine 21 is at a minimum whereas the amount of compressed gas which is required to increase the speed of the engine 21 to a level suitable to drive a vehicle on a highway is readily available.

With reference to FIGS. 6 and 7, a third embodiment of a valve actuator 41 according to the present invention includes an arcuate insulating element 91 having a first end pivotally secured by any suitable device such as screw 92 to the shaft 55 for co-rotation with the shaft 55. The screw 92 is screwed into a tapped hole in the insulating element 91 such that a tab 94 at an end of the screw 92 engages a groove 96 provided in the shaft 55. In this way, the insulating element 91 positively rotates with the shaft 55. However, as the shaft 55 rotates faster, a second end 98 of the insulating element 91 is permitted to pivot outwardly under the influence of centrifugal force because of the groove 96 provided in the shaft 55. A spring 93 connected between the second end 98 of the element 91 and the shaft 55 urges the second end of the element 91 towards the center of the housing 53.

A contact 99 similar to the contact 59 (FIG. 2) is arranged such that one end of the contact 99 is in constant contact with the conducting plate 51 located centrally within the housing 53. The other end of the contact 99 engages a conductive sleeve 101 arranged in bore 102. A contact element 95 is arranged in the conductive sleeve 101 in constant contact with the sleeve 101. The bore 102 is arranged generally parallel to the shaft 55 near the second end of the arcuate insulating element **91**. The contact **95** is biased by a spring **97** towards the upper inside surface of the housing **53** for selective contact with each of the plurality of radial contacts **85** which increase in arc length towards the outer peripheral surface of the housing **53** (FIG. **6**).

In operation of the device of FIGS. **6** and **7**, as the shaft **55** rotates the arcuate insulating element **91** rotates with the shaft **55** and the second end **98** of the insulating element **91** tends to pivot about the shaft **55** due to centrifugal force. Thus, as the effective length of the contact **95** increases, i.e., as the arcuate insulating element **91** pivots further outwardly, the number of degrees of rotation over which the contact **95** is in contact with each of the radial contacts **85** on the upper inside surface of the housing **53** increases thereby permitting each of the valves **39** to remain open for a longer period of each engine cycle to admit more compressed gas to the respective cylinder **20** to further increase the speed of the engine **21**.

With reference to FIG. **1**, a mechanical advance linkage **104** which is connected to the throttle linkage **35**, advances the initiation of the opening of each valve **39** such that compressed gas is injected into the respective cylinder further before the piston **22** in the respective cylinder **20** reaches a top dead center position as the speed of the engine is increased by moving the throttle linkage **35**. The advance linkage **104** is similar to a conventional standard mechanical advance employed on an internal combustion engine. In other words, the linkage **104** varies the relationship between the angular positions of a point on the shaft **55** and a point on the housing **53** containing the contacts. Alternatively, a conventional vacuum advance could also be employed. By advancing the timing of the opening of the valves **39**, the speed of the engine can more easily be increased.

The operation of the engine cycle according to the present invention will now be described. The compressed gas injected into each cylinder of the engine **21** drives the respective piston **22** downward to drive a conventional crankshaft (not shown). The movement of the piston downwardly causes the compressed gas to expand rapidly and cool. As the piston **22** begins to move upwardly in the cylinder **20** a suitable exhaust valve (not shown) arranged to close an exhaust passageway is opened by any suitable apparatus. The expanded gas is then expelled through the exhaust passageway. As the piston **22** again begins to move downwardly a suitable intake valve opens to admit ambient air to the cylinder. The intake valve closes and the ambient air is compressed on the subsequent upward movement of the piston until the piston reaches approximately the top dead center position at which time the compressed gas is again injected into the cylinder **20** to drive the piston **22** downward and the cycle begins anew.

In the case of adapting a conventional internal combustion engine for operation on compressed gas, a plurality of plates **103** are preferably arranged over an end of the exhaust passageways in order to reduce the outlet size of the exhaust passageways of the conventional internal combustion engine. In the illustrated embodiment, a single plate having an opening in the center is bolted to the outside exhaust passageway on each bank of the V-8 engine while another single plate having two openings therein is arranged with one opening over each of the interior exhaust passageways on each bank of the V-8 engine. A line **105** is suitably attached to each of the adaptor places to carry the exhaust to an appropriate location. In a preferred embodiment, the exhaust lines **105** are 1½" plastic tubing.

In a preferred embodiment, the exhaust lines **105** of one bank of the V-8 engine are collected in a line **107** and fed to an inlet of a compressor **109**. The pressure of the exhaust gas emmanating from the engine **21** according to the present invention is approximately 25 p.s.i. In this way, the compressor **109** does not have to pull the exhaust into the compressor since the gas exhausted from the engine **21** is at a positive pressure. The positive pressure of the incoming fluid increases the efficiency and reduces wear on the compressor **109**. The exhaust gas is compressed in the compressor **109** and returned through a line **111** and a check valve **113** to the compressed gas storage tank **23**. The check valve **113** prevents the flow of compressed gas stored in the tank **23** back towards the compressor **109**.

A suitable pressure sensor **115** is arranged at an upper end of the tank **23** and sends a signal along a line **117** when the pressure exceeds a predetermined level and when the pressure drops below a predetermined level. The line **117** controls an electrically actuated clutch **119** disposed at a front end of the compressor **109**. The clutch **119** is operative to engage and disengage the compressor **109** from a drive pulley **121**. Also, the signal carried by the line **117** actuates a suitable valve **123** arranged on a compressor housing **125** to exhaust the air entering the compressor housing **125** from the line **107** when the clutch **119** has disengaged the compressor **109** from the drive pully **121**.

In a preferred embodiment, when the pressure is the tank **23** reaches approximately 600 p.s.i., the clutch **119** is disengaged and the compressor **109** is deactivated and the valve **123** is opened to exhaust the expanded gas delivered to the compressor **109** from the line **107** to the atmosphere. When the pressure within the tank **23** drops below approximately 500 p.s.i., the sensor **115** sends a signal to engage the clutch **119** and close the valve **123**, thereby operating the compressor **109** for supplying the tank **23** with compressed gas.

The pulley **121** which drives the compressor **109** through the clutch **119** is driven by a belt **127** which is driven by a pulley **129** which operates through a gear box **131**. With reference to FIGS. **1** and **8**, a second pulley **133** on the gear box is driven by a belt **135** from a pulley **137** arranged on a drive shaft **139** of the engine **21**. The pulley **137** drives a splined shaft **140** which has a first gear **141** and a second larger gear **143** arranged thereon for rotation with the splined shaft **140**. The splined shaft **140** permits axial movement of the gears **141** and **143** along the shaft **140**.

In normal operation (as seen in FIG. **8**), the first gear **141** engages a third gear **145** arranged on a shaft **147** which drives the pulley **129**. The shafts **140** and **147** are arranged in suitable bearings **149** arranged at each end thereof. When the speed of the engine **21** drops below a predetermined level, a suitable sensor **151** responsive to the speed of the drive shaft **139** of the engine **21** generates a signal which is transmitted through a line **153** to a solenoid actuator **155** arranged within the gear box **131**. The solenoid actuator **155** moves the first and second gears **141**, **143** axially along the splined shaft **140** to the right as seen in FIG. **8** such that the second, larger gear **143** engages a fourth smaller gear **157** which is arranged on the shaft **147**. The ratio of the second gear **143** to the fourth gear **157** is preferably approximately 3 to 1.

In this way, when the speed of the engine **21** drops below the predetermined level as sensed by the sensor **151** (which predetermined level is insufficient to drive the compressor **109** at a speed sufficient to generate the 500-600 pounds of pressure which is preferably in the tank **23**), the solenoid actuator **155** is energized to slide the gears **143**, **141** axially along the splined shaft **140** so that the second, larger gear **143** engages the fourth, smaller gear **157** to drive the pulley **129** and hence the compressor **109** at a higher rate of speed to generate the desired pressure. When the speed of the engine increases above the predetermined level, in a preferred embodiment approximately 1500 rpm, the solenoid actuator **155** is deactivated by the sensor **151** thereby moving the gears **143** and **141** to the left as seen in FIG. **8** such that the first gear **141** re-engages with the third gear **145** to effectuate a 1 to 1 ratio between the output shaft **139** of the engine **21** and the pulley **129**.

The other bank of the V-8 engine has its exhaust ports arranged with adapter plates **103** similar to those on the first bank. However, the exhaust from this bank of the engine **21** is not collected and circulated through the compressor **109**. In a preferred embodiment, a portion of the exhaust is collected in a line **159** and fed to an enlarged chamber **161**. A second fluid is fed through a line **163** into the chamber **161** to be cooled by the cool exhaust emmanating from the engine **21** in the line **159**. The second fluid in the line **163** may be either transmission fluid contained in a transmission associated with the engine **21** or a portion of the oil used to lubricate the engine **21**. A second portion of the exhaust from the second bank of the V-8 engine is removed from the line **159** in a line **165** and used as a working fluid in an air conditioning system or for any other suitable use.

It should be noted that the particular arrangement utilized for collecting and distributing the gas exhausted from the engine **21** would be determined by the use for which the engine is employed. In other words, it may be advantageous to rearrange the exhaust tubing such that a larger or smaller percentage of the exhaust is routed through the compressor **109**. It should also be noted that since the exhaust lines **105** are plastic tubing, a rearrangement of the lines for a different purpose is both simple and inexpensive.

In operation of the engine of the present invention, the engine **21** is started by energizing the solenoid valve **29** and any suitable starting device (not shown), e.g., a conventional electric starter as used on an internal combustion engine. Compressed gas from the full tank **23** flows through the line **25** and a variable amount of the compressed gas is admitted to the distributor **33** by controlling the regulator valve **31** through the linkage **33** and the operator actuated throttle linkage **35**. The compressed gas is distributed to each of the lines **37** which lead to the individual cylinders **20**. The compressed gas is admitted to each of the cylinders **20** in timed relationship to the position of the pistons within the cylinders by opening the valves **39** with the valve actuator **41**.

When it is desired to increase the speed of the engine, the operator moves the throttle linkage **35** which simultaneously admits a larger quantity of compressed gas to the distributor **33** from the tank **23** by further opening the regulator valve **31**. The timing of the valve actuator **41** is also advanced through the linkage **104**. Still further, as the speed of the engine **21** increases, the effective length of the rotating contact **83** (FIG. **4**) or **95** (FIG. **6**) increases thereby electrically contacting a wider portion of one of the stationary radial contacts **85** to cause each of the valves **39** to remain open for a longer period of each engine cycle to admit a larger quantity of compressed gas to each of the cylinders **20**.

As can be seen, the combination of the regulating valve **31**, the mechanical advance **104**, and the valve actuator **41**, combine to produce a compressed gas engine which is quickly and efficiently adaptable to various operating speeds. However, all three of the controls need not be employed simultaneously. For example, the mechanical advance **104** could be utilized without the benefit of one of the varying valve actuators **41** but the high speed operation of the engine may not be as efficient. By increasing the duration of each engine cycle over which each of the valves **39** remains open to admit compressed gas to each of the cylinders **20** as the speed increases, conservation of compressed gas during low speed operation and efficient high speed operation are both possible.

After the compressed gas admitted to the cylinder **20** has forced the piston **22** downwardly within the cylinder to drive the shaft **139** of the engine, the piston **22** moves upwardly within the cylinder **20** and forces the expanded gas out through a suitable exhaust valve (not shown) through the adapter plate **103** (if employed) and into the exhaust line **105**. The cool exhaust can then be collected in any suitable arrangement to be compressed and returned to the tank **23** or used for any desired purpose including use as a working fluid in an air conditioning system or as a coolant for oil.

When using the apparatus and method of the present invention to adapt a ordinary internal combustion engine for operation with compressed gas it can be seen that considerable savings in weight are achieved. For example, the ordinary cooling system including a radiator, fan, hoses, etc. can be eliminated since the compressed gas is cooled as it expands in the cylinder. In addition, there are no explosions within the cylinder to generate heat. Further reductions in weight are obtained by employing plastic tubing for the lines which carry the compressed gas between the distributor and the cylinders and for the exhaust lines. Once again, heavy tubing is not required since there is little or no heat generated by the engine of the present invention. In addition, the noise generated by an engine according to the present invention is considerably less than that generated by an ordinary internal combustion engine since there are no explosions taking place within the cylinders.

The principles of preferred embodiments of the present invention have been described in the foregoing specification. However, the invention which is intended to be protected is not to be construed as limited to the particular embodiments disclosed. The embodiments are to be regarded as illustrative rather than restrictive. Variations and changes may be made by others without departing from the spirit of the invention. Accordingly, it is expressly intended that all such variations and changes which fall within the spirit and the scope of the present invention as defined in the appended claims be embraced thereby.

What is claimed is:

1. An apparatus for operating an engine having at least one cylinder and a reciprocating piston therein comprising:

a source of compressed gas;

distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;

valve means for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder;

altering means for increasing the duration of each engine cycle over which the valve means admits compressed gas to the at least one cylinder as the speed of the engine increases; and

exhaust means for exhausting gas as the piston subsequently approaches approximately the top dead center position.

2. The apparatus of claim **1** further comprising control means for controlling the amount of compressed gas admitted to the distributor means.

3. The apparatus of claim **1** wherein the valve means is a solenoid valve secured in an opening in the cylinder above the level of the piston at the top dead center position.

4. The apparatus of claims **1** or **2** further comprising means for advancing the timing of the valve means as the speed of the engine increases such that compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

5. The apparatus of claim **4** wherein the means for advancing the timing comprises a mechanical linkage connected to an operator actuated accelerator linkage.

6. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is compressed in a compressor driven by an output shaft of the engine and is returned to the source of compressed gas.

7. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used to cool transmission fluid for a transmission associated with the engine.

8. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used as a working fluid in an air conditioning system.

9. The apparatus of claim **6** further comprising first gearing means interposed between the output shaft of the engine and the compressor for increasing the speed at which the compressor is driven.

10. The apparatus of claim **6** further comprising clutch means attached to the compressor both for disengaging the compressor from the output shaft of the engine when a first predetermined pressure at the source of compressed gas is exceeded and for engaging the compressor with the output shaft of the engine when the pressure at the source of compressed gas drops below a second predetermined pressure.

11. The apparatus of claim **9** further comprising means for both disengaging the first gearing means when a predetermined speed of the engine is exceeded and engaging a second gearing means for driving the compressor at a speed slower than the first gearing means when the predetermined speed of the engine is exceeded.

12. The apparatus of claim **1** wherein the valve means is electrically actuated and wherein the altering means comprises:

a rotating member timed with the at least one cylinder and arranged within a housing;

first and second contacts arranged on a first end of the rotating member and on an inside surface of the housing, respectively;

means for increasing the distance of the first contact from the rotational axis of the rotating member as the speed of the engine increases such that the first contact moves radially outwardly within the housing; and

said second contact presenting a longer arc length to the first contact as the distance of the first contact from the rotational axis of the rotating member increases.

13. The apparatus of claim **12** wherein the rotating member comprises an arcuate arm and wherein the means for increasing the distance of the first contact comprises pivotally mounting a second end of the arcuate arm about the axis of rotation of the rotating member and spring means for biasing the first end of the arcuate arm towards a radially inward position whereby the first end of the arcuate arm pivots radially outwardly as the speed of the engine increases.

14. The apparatus of claim **12** wherein the rotating member is axially slidably received on a rotating shaft for co-rotation therewith, said shaft having splines with a radial component in the direction of rotation, and wherein the first contact comprises a flexible contact located on an upper surface of the rotating member, said flexible contact being biased against the inside surface of the housing which carries the second contacts whereby as the speed of the engine increases the rotating member is urged axially along the splined shaft towards the inside surface of the housing such that the flexible contact is forced radially outwardly along the inside surface.

15. The apparatus of claim **12** wherein the second contact comprises of radially extending conductor arranged on an upper inside surface of the housing, said conductor increasing in arc length as the conductor extends radially outwardly from a central portion of the housing.

16. An apparatus for adapting an internal combustion engine for operation with compressed gas, the internal combustion engine having at least one cylinder, a piston reciprocable within the at least one cylinder, intake and exhaust means disposed in the at least one cylinder, and a tapped hole in the at least one cylinder adapted to receive a spark plug, the apparatus comprising:

a source of compressed gas;

distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;

valve means arranged in the tapped hole for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder; and

altering means for increasing the duration of each engine cycle over which the valve means remains open to admit the compressed gas as the speed of the engine increases.

17. An apparatus as in claim **16** further comprising first adapter plate means for supporting the distributor means above an intake manifold of the engine, which adaptor plate means allows ambient air to enter through the intake manifold.

18. The apparatus of claim **16** further comprising second adapter plate means for reducing the exit area of the exhaust means.

19. A method of operating an engine on compressed gas, said engine having at least one cylinder and a piston reciprocable therein comprising the steps of:

delivering compressed gas from a source to a distributor;

distributing the compressed gas to the at least one cylinder;

admitting compressed gas to the at least one cylinder through an intake valve when the piston is at approximately a top dad center position;

increasing the duration of each engine cycle over which compressed gas is admitted to the at least one cylinder as the engine speed increases; and

exhausting the remaining gas when the piston subsequently reaches approximately the top dead center position.

20. The method of claim **19** further comprising the step of controlling the amount of compressed gas which is delivered to the distributor.

21. The method of claim **19** further comprising the step of advancing the timing of the opening of the intake valve as the speed of the engine increases.

22. An apparatus for operating an engine having at least one cylinder and a piston reciprocable therein on compressed gas comprising:

a source of compressed gas;

distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;

electrically actuated valve means secured in an opening in the at least one cylinder for selectively admitting compressed gas to the at least one cylinder when the piston is in approximately a top dead center position; and

means for advancing the timing of the valve means as the speed of the engine increases whereby compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

* * * * *

Fuel from 'Burning Water'
provided by The Reclamation Project
*** Working Plans ***
01/09/02

UltimateMallStore Disclaimer: We have indication these plans work when built, but every person is unique this depends on your ability to make it work,also we are not responsible for any accidents or injuries,etc so please be careful ,if you choose to try it, use common sense and start SMALL, like a lawnmower engine, please report any successes you might have.

The following file was recieved by Dave the Gravman via FAX, he kindly sent it to me, thanks Dave!

- **Convert your engine to burn hydrogen & oxygen**
- **Make vapor from water on demand & pollution-free**
- **Convert your vehicle to a ZPV: Zero Pollution Vehicle**
- **Freedom from Gasoline**
- **Simple and inexpensive conversion**
- **On Demand vapor rate via throttle**
- **Easy do it yourself public domain plans**
- **Help clean the air while you save money**

FIGURE 1

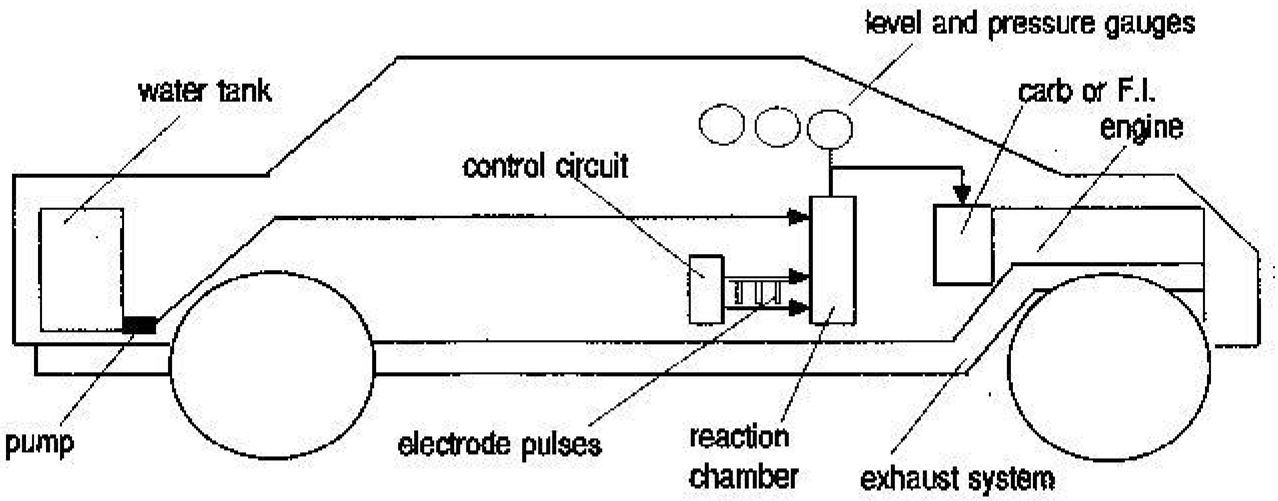
level and pressure gauges
water tank
carb or F.I.
engine
control circuit
pump
electrode pulses
reaction chamber
exhaust system

These plans can be used to run your car, truck, RV, motorcycle, airplane, etc. from tap water.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

Are you fed up with ga$oline price$ and THE POLLUTION?

Would you like to do something about it besides complain and wait?

Are you still CHOKING on the whole idea of fossil fuel CON$UMPTION? WHY?

Wouldn't you like to free yourself from centrally-controlled or imported fuel?

Do you have a 2nd vehicle you don't use every day?

Now there is something we can actually do about it, as individuals willing to help clean the environment, and travel at lower cost to both you and your environment. Do-it-yourself plans allow the individual (that's you and me, folks) to make a difference.

This is the easiest and lowest cost way to convert your car to run on (relatively) free energy. Now with existing technology, anyone can stand up and make a difference by reducing the local automotive pollution, eliminate ga$oline expen$e$, help restore our atmosphere, and breathe a little easier. You will be making use of your entire existing system, except the fuel tank and catalytic converter.

THE PLAN:

Know the Truth and set yourself FREE. Set a good example for the World you choose to create. Exercise your own Free Will. Live cleaner and healthier.

Build and install a low-cost alternative method for running your vehicle (internal combustion) engine on TAP WATER, using off-the-shelf components. This is simply an efficient way to convert ordinary tap water into gaseous Hydrogen and Oxygen, and then burn these vapours in the engine, instead of that $melly, $tinky, expen$ive 'other $tuff.

This 'minisystem' runs easily from your existing battery and electrical system, and it plugs into your carburetor with simple off-the-shelf fittings. You will be installing a plastic water tank, a control circuit, a reaction chamber, a hi-pressure carb/FI fitting, and 3 gauges (see Figure 1), and then hooking into your existing carb/FI.

The SIMPLICITY comes from being an 'on-demand' system requiring no fancy storage, or plumbing. You crank the gas pedal or throttle and you electrically create more vapour for immediate consumption, on demand; Lo - Hi Flow Rate as needed from idle - max power. The only real change is that you are using tap water as fuel, instead of the traditional petroleum-based fuel. Given, a choice, which way would you choose?

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

FAQs

Q: Does it really work?

A: Yes; this is well-established technology dating back to stainless steel. But be sure to follow these instructions using the proper mechanical and electrical assembly techniques, as it incorporates the best qualities of several techniques.

Q: How does it qualify as 'free energy'?

A: If you're paying someone for the water you use, then it is not strictly 'free'. But, the alternative is to keep buying into the expen$ive ga$oline and re$ultant hydrocarbon pollution.

Q: ls it Safe?

A: Technically, it is safer than running on fossil fuel, because you are no longer choking on your own emissions (health-wise), but in general it is practically as safe as your current gasoline arrangement. You will be installing a few simple safety devices, using current automotive standards.

Q: What kind of performance can I expect?

A: Properly adjusted, your modified vapour-only fuel system will run cooler, and at a modestly higher power level. The mileage performance expected from this design ranges from 50-300 mpg, depending on your adjusting skills.

Q: Can I do the modification myself?

A: Why not? If you know someone with basic mechanical and/or electrical skills, you can even delegate some of the construction. If you are using a fuel-injected engine, you may have to get a mechanic's opinion.

0: What is the environmental impact that my vehicle will have?

A: It will be producing H20 steam and unburnt 02, hence it will be cleaning the environment, rather than dumping nauseous toxins into it. Plus you will be helping to save our dwindling supply of atmospheric oxygen. Any excess vapour in the reaction becomes either steam or oxygen. You can also expect to be receiving more than casual interest from those around you.

Q: Is this really a steam engine?

A: No; really. Exceedingly hi temp & pressure are not used. This is strictly an internal combustion engine (burning orthohydrogen) with residual steam in the exhaust as a by product.

Gasoline as a fuel is NOT NECESSARY; it is optional.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

1. ORIGIN - In the 19th century, the gasoline portion of the refining process, was first considered to be a 'waste' product of extracting the purified crude oil. Later on, it was discovered that it could be $old as fuel, instead of just dumping it back in the hole as had been the tradition.

2. CONSUMPTION RATE - The gasoline consumption rate for every mass-produced car has been carefully 'designed in' as a market asset. As an indication, simply observe how quickly and closely ALL the local different gas stations adjust their prices. Even the hybrid cars which use electric motors still consume a designed amount of gasoline, and their price tags are prohibitively high.

3. EFFICIENCY - There is a lot of thermochemical energy in gasoline, but there is even more energy in water. The DOE has quoted about 40%, 50 it is probably much more than that. Most people are unaware that 'internal combustion' is DEFINED as: a thermo-vapour process; as in 'no liquid in the reaction'; AND most of the gasoline in a standard internal combustion engine, is ACTUALLY CONSUMED cooked and finally broken down) IN THE CATALYTIC CONVERTER, which happens AFTER the fuel has been not-so-burnt in the engine. Sadly, this means that most of the fuel we use in this way, is used only to cool down the combustion process, rather than using a cleaner and more efficient means to do so.

4. ADDITIVES - Also sadly, we are told by 'authorities' that some of the many gasoline additives are in the mix to increase performance; but because of its current overly-complex molecular structure, the real built-in function of the gasoline formula is to slow down the combustion so that only so much is actually consumed in the cylinder, and the liquid balance goes to the catalytic converter. As a further insult, the additives are also there to clog and prevent the use of the Pogue-style carburetors, designed to get 200-300 mpg.

5. PROFIT - Is the Pope catholic? does the bear poop in the woods? Of course the oil companies are making a huge profit. It is by design. What do you suppose the Gulf War was about? Just look at where the crude oil comes from and where the money is flowing. Rest assured that the oil companies do NOT want you to know how to make use of this water-fuel technology. They have been making money on our ignorance, dis-empowerment, and willingness to follow along in a mindless 'comfort-zone' of toxic waste, suppression, resentment, and apathy: WHO NEEDS IT?

Let us proceed to set a good example and do it right, do it clean.

Exceedingly simple

Water is pumped as needed to replenish and maintain the liquid level in the chamber. The electrodes are vibrated with a 0.5-SA electrical pulse which breaks 2(H20) --- (2H2 + 02).

When the pressure reaches say 30-60 psi, you turn the key and go. You step on the pedal, you send more energy to the electrodes, and thus more vapour to the cylinders; i.e. fuel vapour on demand.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

You set the idle - max flow rate to get the most efficient use of power, and you're off to the races.

In the BIG picture, your Free Energy is coming from the tap water, in an open system, as the latent energy in the water is enough to power the engine, and hence drive the alternator and whatever belt-driven accessories; AND the alternator is efficient enough to run the various electrical loads (10-20 amps), including the additional low current to run this vapour reaction. No extra batteries are required.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

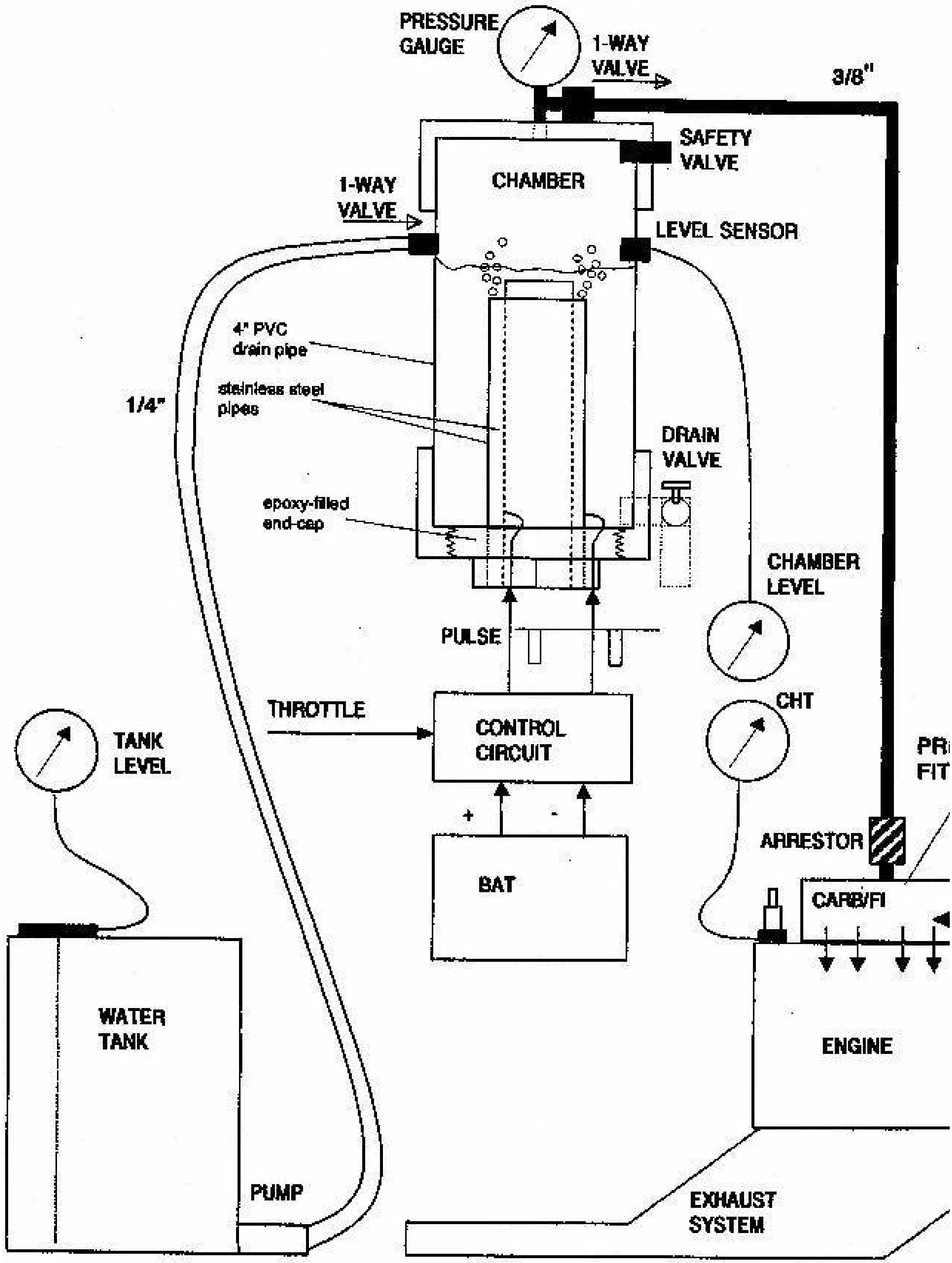
PRESSURE GAUGE
1-WAY VALVE
3/8"
SAFETY VALVE
CHAMBER
1-WAY VALVE
LEVEL SENSOR
4" PVC drain pipe
stainless steel pipes
1/4"
DRAIN VALVE
epoxy-filled end-cap
CHAMBER LEVEL
PULSE
CHT
THROTTLE
CONTROL CIRCUIT
TANK LEVEL
+
-
PR
FIT
ARRESTOR
BAT
CARB/FI
WATER TANK
ENGINE
PUMP
EXHAUST SYSTEM

FIGURE 2 BLOC DIAGRAM

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

OVERVIEW - Here is the suggested sequence of steps:

1. Install the CHT (or EGT) gauge and measure your current operating temp range (gasoline), for comparison.

2. Build & test the controller to verify the correct pulse output.

3. Build the reaction chamber & test it w. the controller (i.e pressure out).

4. Install the tank, controller, chamber, and pressure fittings.

5. Run engine & Adjust the control circuit as necessary for best performance.

6. Install the stainless steel valves and get the pistons/cylinders coated with ceramic.

7. Coat the exhaust system with ceramic without the catalytic converter OR let it rust out and replace the whole dang thang with stainless steel pipe sections.

YOU WILL NEED:

- **- plastic water tank with pump and level sensor.**
- **- control circuit, wiring, connectors, and epoxy.**
- **- reaction chamber with electrodes and fittings.**
- **- 3/8" stainless steel flex-tubing, fittings and clamps.- carb/FI vapor-pressure fitting kit. - pressure, CHT (or EGT), & level gauges.**
- **- stainless steel valves.**
- **- copper mesh junction.**
- **- ceramic surface treatment for cylinders & pistons.**
- **- stainless steel or ceramic treated exhaust assembly.**

BASIC TOOLS

- **- drill, screwdriver and pliers**
- **- hole cutter**
- **- wire-wrap, solder-iron and clippers**
- **- DVM and oscilloscope.**

REACTION CHAMBER

Construct as shown in Figure 2. Use a section of 4" PVC waste pipe with a threaded screw-cap fitting on one end and a standard end-cap at the other.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

Make sure to drill-and-epoxy or tap threads through the PVC components for all fittings. Set and control the water level in the chamber so that it well submerses the pipe electrodes; yet leave some headroom to build up the hydrogen/oxygen vapour pressure.

Use stainless steel wires inside the chamber or otherwise use a protective coating; use insulated wires outside. Ensure that the epoxy perfects the seal, or otherwise lay down a bead of water-proof silicone that can hold pressure.

The screw fitting may require soft silicone sealant, or a gasket; its purpose is to hold pressure and allow periodic inspection of the electrodes. No leaks, no problems. Make sure you get a symmetric 1-5mm gap between the 2 stainless steel pipes. The referenced literature suggests that the closer to 1mm you get, the better. You WILL want to get your chamber level sensor verified BEFORE you epoxy the cap on.

Make your solder connections at the wire/electrode junctions nice, smooth, and solid; then apply a water-proof coating, e.g. the epoxy you use for joining the pipes to the screw cap.

This epoxy must be water-proof and be capable of holding metal to plastic under pressure.

You WILL want to get your chamber level sensor verified BEFORE you epoxy the cap on.

CONTROL CIRCUIT

Figure 3

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

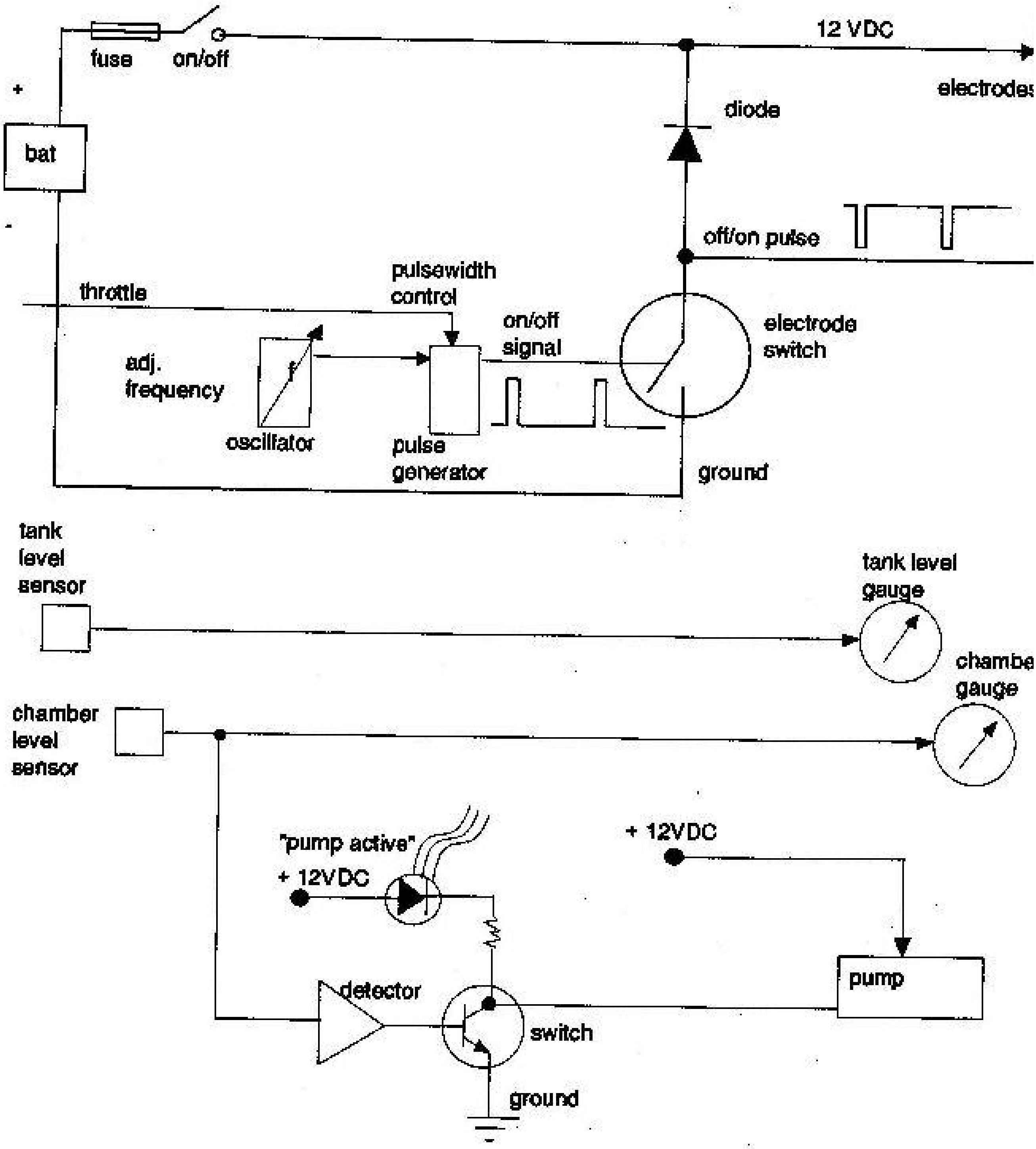
fuse
on/off
12 VDC
+
electrode
diode
bat
-
off/on pulse
throttle
pulsewidth
control
on/off
signal
electrode
switch
adj.
frequency
f
oscillator
pulse
generator
ground
tank
level
sensor
tank level
gauge
chambe
gauge
chamber
level
sensor
"pump active"
+ 12VDC
+ 12VDC
detector
switch
pump
ground

Figures 3 & 4 show a simple circuit to control and drive this mini-system. You are going to make a 'square-pulse' signal that 'plays' the electrodes like a tuning fork; which you can watch on an oscilloscope. The premise given by the literature is:

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

The faster you want do go down the road, the 'fatter' you make the pulses going into the reaction chamber. Duty cycle will vary with the throttle in the vicinity of 90%Mark 10%Space (Off/On).

NOTE : Figures 4 and 5 have been revised, click to see the large version

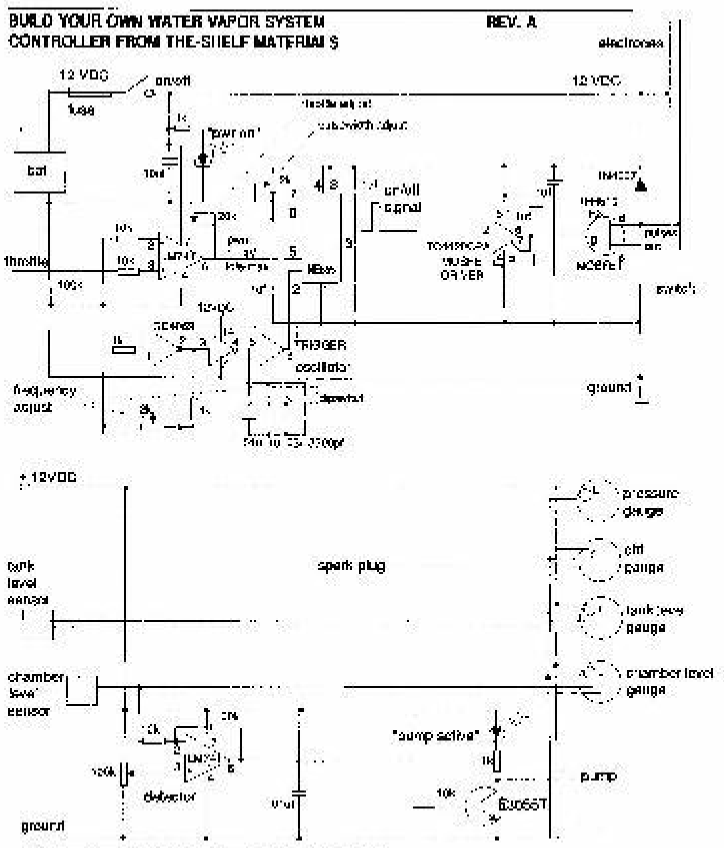
BUILD YOUR OWN WATER VAPOR SYSTEM
CONTROLLER FROM THE-SHELF MATERIALS
REV. A
electronics
12 VDC
on/off
fuse
12 VDC
throttle adjust
pulse width adjust
"pwr on"
bat
throttle
10k
LM741
on/off signal
TC4420CA MOSFET DRIVER
1N4007
IRF612
MOSFET
pulses out
switch
TRIGGER
oscillator
CD4093
frequency adjust
ground
+12VDC
tank level sensor
spark plug
pressure gauge
ctrl gauge
tank level gauge
chamber level sensor
chamber level gauge
"pump active"
pump
detector
2N3055T
ground

REVISION "A"

FIGURE 4 E-SCHEMATIC

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

BUILD YOUR OWN WATER VAPOR SYSTEM
CONTROLLER FROM THE-SHELF MATERIALS

REV. A

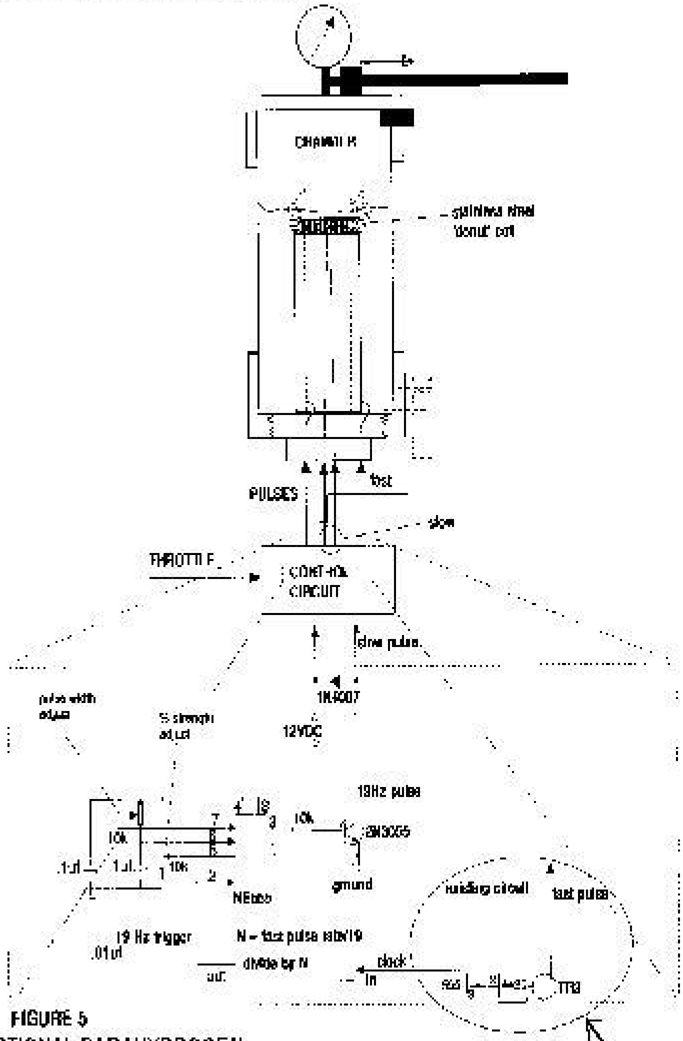
CHAMBER
stainless steel
'donut' coil
PULSES
fast
slow
THROTTLE
CONTROL
CIRCUIT
slow pulse
1N4007
12VDC
pulse width
adjust
% strength
adjust
19Hz pulse
10k
2N3055
10k
.1uf
1uf
10k
1
2
7
3
NE555
ground
fast pulse
19 Hz trigger
.01uf
N = fast pulse rate/19
divide by N
out
clock
in
555
TTL

FIGURE 5
OPTIONAL PARAHYDROGEN
ENHANCEMENT ! REVISION "A"

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

electro

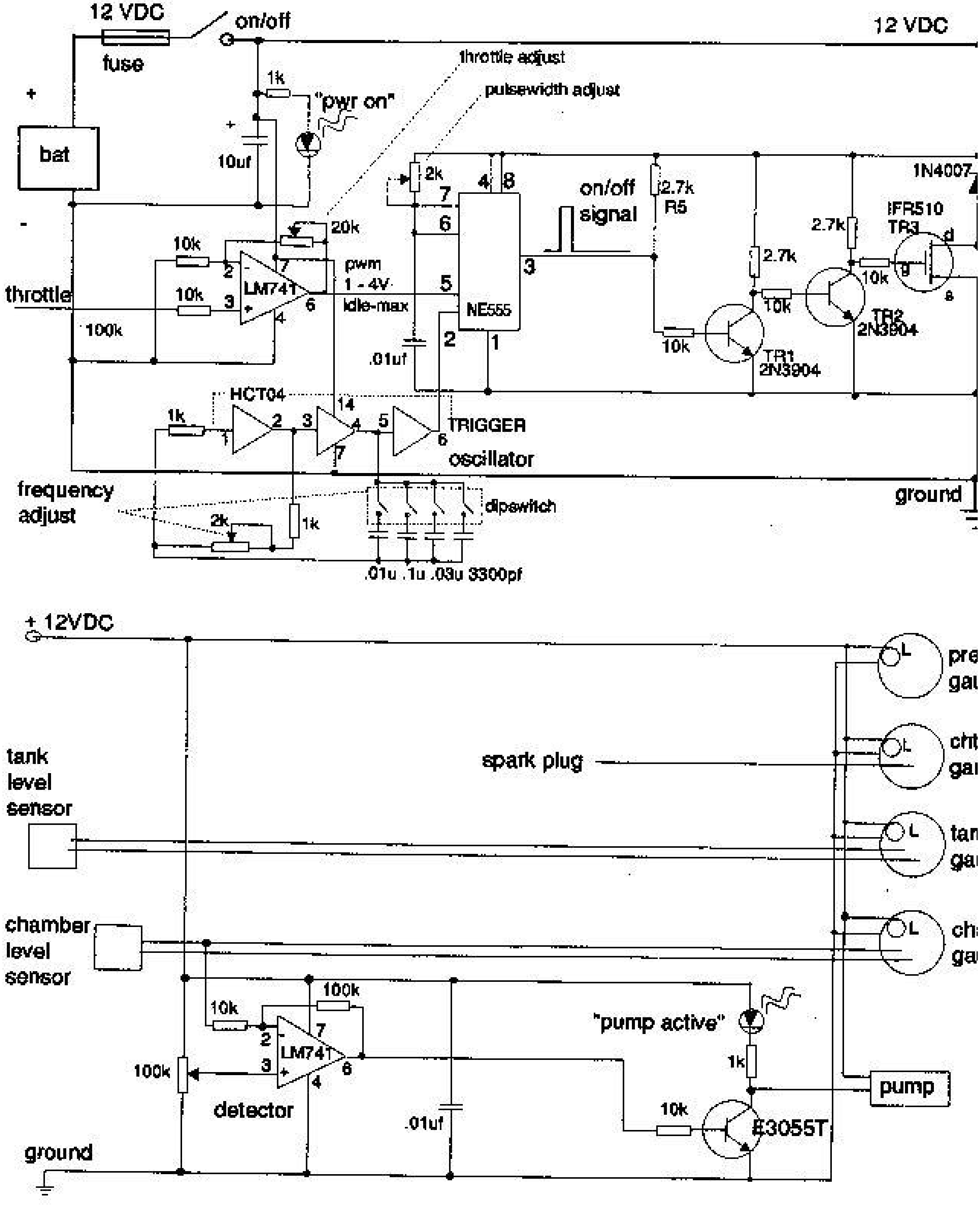
12 VDC
on/off
fuse
12 VDC
+
bat
-
1k
"pwr on"
+
10uf
throttle adjust
pulsewidth adjust
2k
7
4
8
on/off
signal
2.7k
R5
1N4007
IFR510
TR3
d
g
s
2.7k
2.7k
10k
10k
20k
10k
-
2
7
LM741
10k
3
+
6
4
pwm
1 - 4V
idle-max
5
6
NE555
3
2
1
.01uf
throttle
100k
10k
TR1
2N3904
TR2
2N3904
HCT04
14
1k
1
2
3
4
5
6
7
TRIGGER
oscillator
ground
frequency
adjust
2k
1k
dipswitch
.01u .1u .03u 3300pf
+ 12VDC
tank
level
sensor
spark plug
chamber
level
sensor
10k
100k
2
-
7
LM741
3
+
6
4
100k
detector
.01uf
"pump active"
1k
10k
E3055T
pump
ground
L
L
L
L
pre
gau
cht
gal
tan
gal
ch
ga

FIGURE 4   E-SCHEMATIC

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

There is nothing sacred about how the pulse waveform is generated; there are many ways to generate pulses, and the attached diagrams show a few. Figure 4 gives the NE555-circuit approach from the referenced patent. The output switching transistor must be rated for

1-5 amps @ I2VDC (in saturation).

Go with a plan that works for you or your friendly neighborhood technoid or mechanic, and go get all the circuit elements from your local electronics store, such as Radio-Shack or Circuits-R-Us, including the circuit board, IC sockets, and enclosure/box.

DigiKey has better selection, service, and knowledge; plus they have no minimum order crap. Be sure to use a circuit board with a built-in ground plane, and to accommodate room for mounting 2 or 3 of the gauges. Mounting the reaction chamber in the engine compartment will require running a stub to your pressure gauge where you can watch it.

You can easily make 30-gauge wire-wrap connections between the socket pins and through-hole discrete components having wire leads. Also make sure to get spec sheets on any IC you use. More details of the best circuits to use will be announced pending prototype testing. You WILL want to get your chamber level sensor verified BEFORE you epoxy the cap on.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

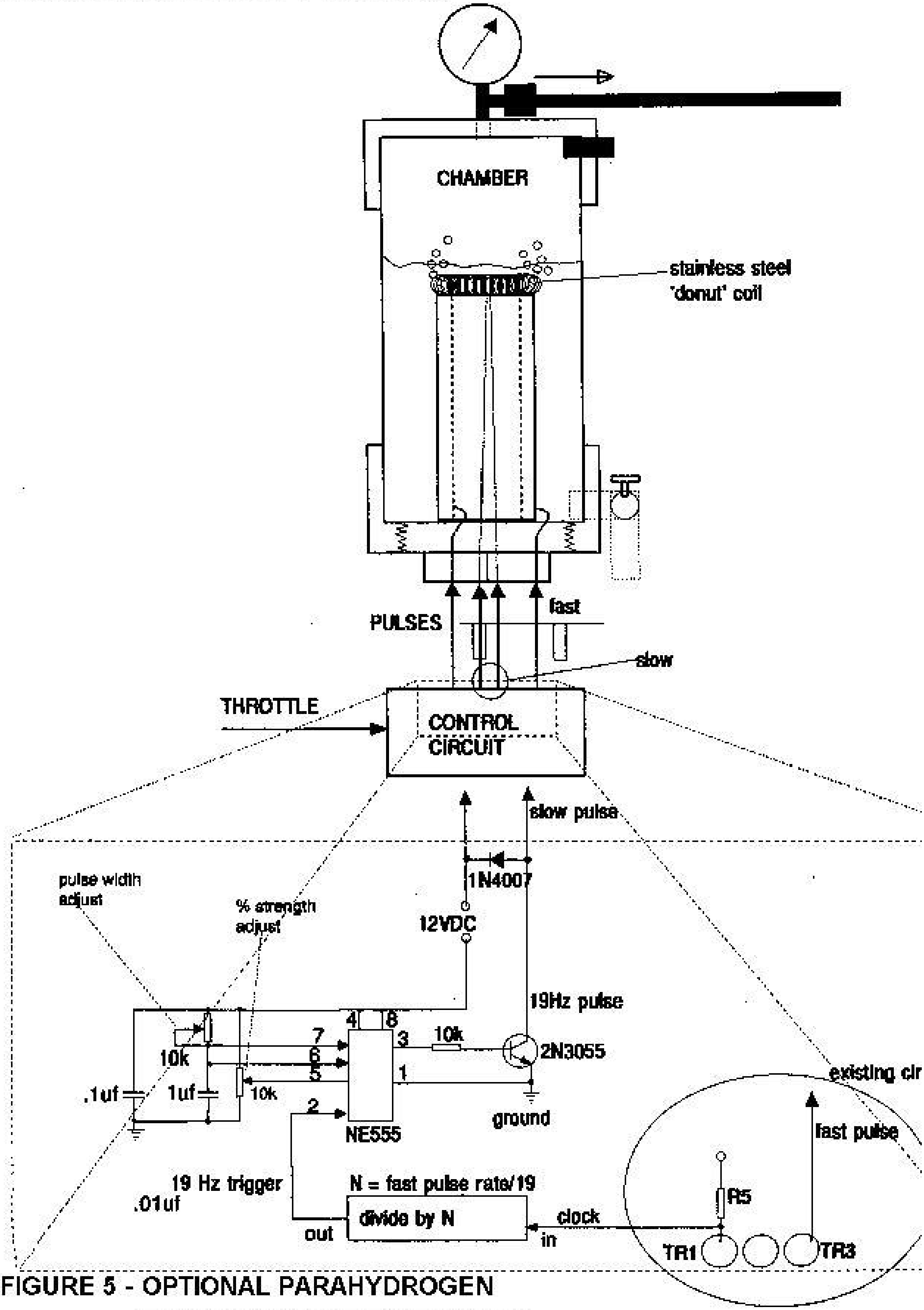
CHAMBER
stainless steel 'donut' coil
PULSES
fast
slow
THROTTLE
CONTROL CIRCUIT
slow pulse
1N4007
12VDC
pulse width adjust
% strength adjust
19Hz pulse
10k
2N3055
.1uf
1uf
10k
10k
4
8
7
6
5
3
1
2
NE555
ground
existing cir
fast pulse
19 Hz trigger
.01uf
N = fast pulse rate/19
divide by N
out
clock
in
R5
TR1
TR3

FIGURE 5 - OPTIONAL PARAHYDROGEN

ENHANCEMENT !

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

THROTTLE CONTROL

If you have a throttle position sensor, you should be able to access the signal from the sensor itself OR from the computer connector. This signal is input to the circuit as the primary control (i.e. throttle level pulse width = vapour rate).

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

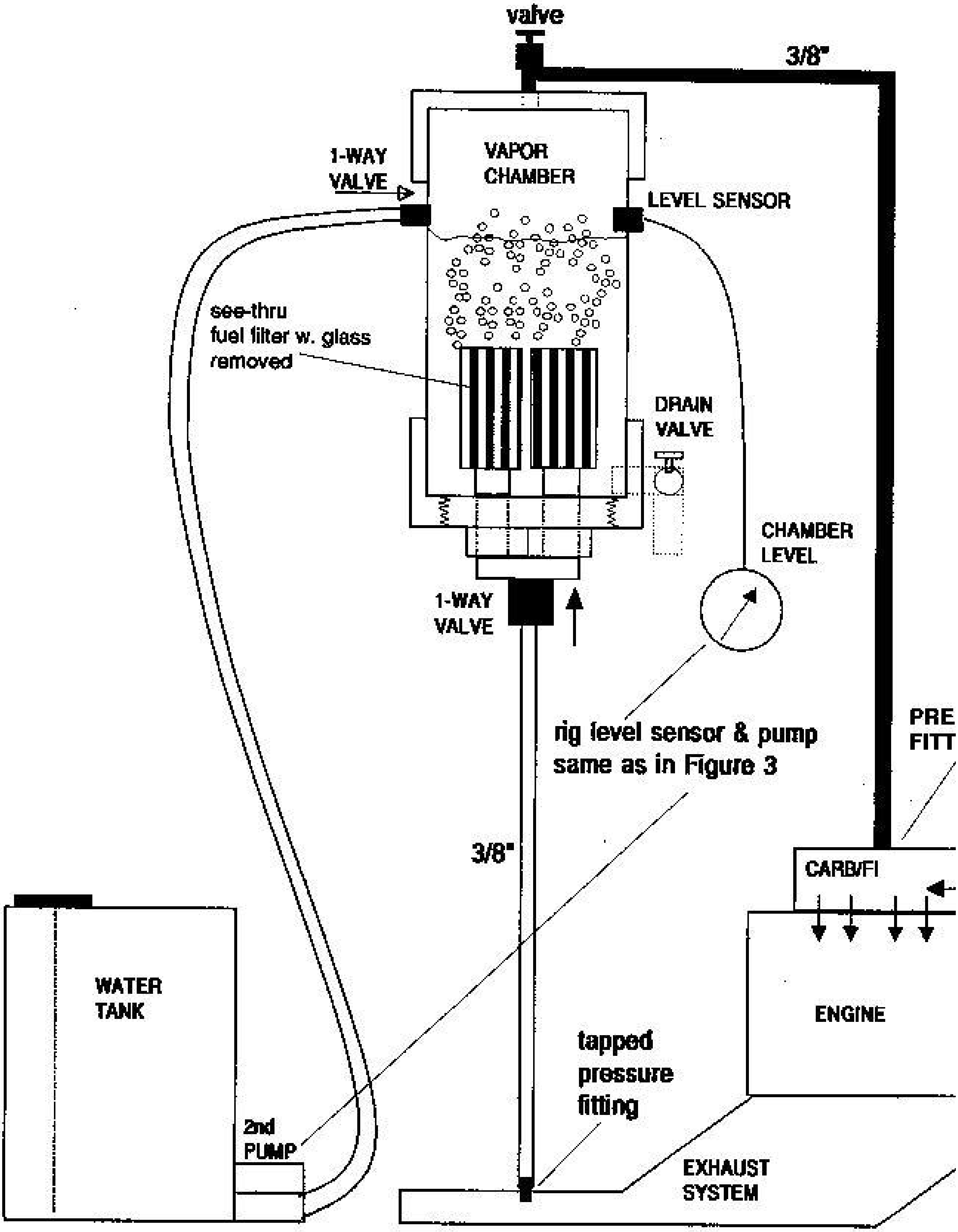
valve
3/8"
1-WAY
VALVE
VAPOR
CHAMBER
LEVEL SENSOR
see-thru
fuel filter w. glass
removed
DRAIN
VALVE
CHAMBER
LEVEL
1-WAY
VALVE
rig level sensor & pump
same as in Figure 3
PRE
FITT
3/8"
CARB/FI
WATER
TANK
ENGINE
tapped
pressure
fitting
2nd
PUMP
EXHAUST
SYSTEM

FIGURE 6 OPTIONAL WATER COOLING

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

If you don't have such a signal available, you will have to rig a rotary POT (variable resistor) to the gas linkage (i.e. coupled to something at the gas pedal or throttle cable running to the carb or Fl. If you make the attachment at the cart/Fl, be sure to use a POT that can handle the engine temp cycles. Don't use a cheezy-cheapy POT; get one rated for long life and mechanical wear; mount it securely to something sturdy and stationary that will not fall apart when you step on the gas.

CONTROL RANGE

The full throttle RANGE (idle-max) MUST control the vapour rate, i.e. pulse-width (duty). The resistor values at the throttle signal must allow the throttle signal voltage, say 1-4 Volt swing, to drive the VAPOR RATE. You will be using this voltage swing to generate a 10% ON 'square' pulse.

The patent implies using a 'resonant' pulse in the 10-250 KHz frequency range; but it is not explicitly stated so. In this circuit, you will simply tune to whatever frequency makes the most efficient vapour conversion. You will have to get into the specs for each IC you use, to insure you connect the right pins to the right wires, to control the frequency and pulse width.

You can use spare sockets to try out different discrete component values. Just keep the ones that are spec-compatible in the circuit, and get the job done.

You crank up the throttle signal and put more electrical energy (fatter pulses) into the electrodes; verify you can get 10% duty on the scope (2 - 100 usec on the horizontal time-base). Your averaging DVM will display the 90%-I 0% DC voltage across the output transistor (Vce or Vds or Output to Ground). Set and connect DVM in the supply current and measure .5 - 5 amps, without blowing the DVM fuse. Now verify that you got everything you wanted.

Verify your wiring connections using your DVM as a continuity detector. Check your wiring 1 at a time and yellow line your final schematic as you go. You can best use board-mount miniature POTs for anything you want to set-and-forget. The LEDs are there to give you a quick visual check of normal vs abnormal operation of your new creation. You WILL want to get your chamber level sensor verified BEFORE you epoxy the cap on (see Figures 2 & 4).

Figure 2 also shows that fittings are required to the carb/FI I. There are ready-made kits (such as by Impco) available for making your pressure fittings to the carburetor or fuel-injector as the case may be. You will necessarily be sealing the built in vents and making a 1-way air-intake.

The copper mesh comprises the inadvertent backfire' protection for the reaction chamber. Make sure that all vapor/duct junctions are air-tight and holding full pressure without leakage. Your new 'system' is considered successful and properly adjusted when you get the full power range at lower temp and minimum vapor flow without blowing the pressure safety valve.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

CHT (or EGT)

Monitor your engine temp with the CHT (cylinder head temp) or EGT (exhaust gas temp) instead of your original engine temp indicator (if any). Your existing gauge is TOO SLOW for this application and will not warn you against overheating until after you have burnt something.

Make sure that your engine RUNS NO HOTTER than in the gasoline arrangement. VDO makes a CHT gauge with a platinum sensor that fits under your spark plug against the cylinder head (make sure it is REAL CLEAN before you reinstall your spark plug (as this is also an electrical ground).

ENGINE/EXHAUST TREATMENT

Get the valves replaced with stainless steel ones AND get the pistons/cylinders ceramic-treated ASAP when you have successfully converted and run your new creation. Do not delay as these items WILL RUST, either by sheer use or by neglect (i.e. letting it sit). You could make max use of your current exhaust system by using it with your new deal until it rusts through, then have your mechanic or welder friend to fit a stainless steel exhaust pipe (no catalytic converter is required). But it could be ea$ier to send your existing exhaust system out for the ceramic treatment, and then simply re-attach it to the exhaust ports.

GENERAL

1. Do not discard or remove any of the old gasoline set up components, e.g. tank, carb/FI, catalytic converter, unless necessary. Better to always leave an easy way to revert back to something that at least runs, just in case. Some people are leaving their gasoline set up completely intact, and switching back and forth at will, just to have a backup plan.

2. Set your throttle circuit so that you get minimum vapour flow at IDLE, and maximum vapour flow at FULL POWER without blowing the pressure relief valve. In this way, you control how 'lean' your mixture is by the strength of the pulse (i.e. 'fatness' at the optimum pulse frequency).

If you just don't get enough power (at any throttle setting), it means that you need to

- **(1) change the pulse frequency,**
- **(2) change the gap between the electrodes,**
- **(3) change the size (bigger) electrodes, OR**
- **(4) make a higher output pulse voltage (last resort).**

Always use an output transistor, such as a MOSFET, that is rated for the voltage and current you need to get the job done. OK so you might have to play around with it some. Isn't that where all the Fun is anyhow?

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

If you get ANY engine knock our loud combustions (not compensated by adjusting the timing), it means that you need to install an additional coil in the chamber, and drive the coil with an additional pulse signal (about 19 Hz on the .lsec time base) (see Figure 5).

Here,you will be slowing down the burn rate just enough so that the vapours burn throughout the power stroke of the piston. Be sure to include a board-mount POT to set the correct strength of this 2nd pulse signal into the coil. This is a stainless steel coil of about 1500 turns (thin wire) that you can arrange like a donut around the center pipe (but NOT touching either electrode), directly over the circular 1-5mm gap. You want NO KNOCKING at any power/throttle setting; smooth power only, but also no excess hydrogen leftover from the combustion.

5. Build the canister(s) as tall as you can without compromising your ability to mount them conveniently near the dash panel, or in the engine compartment, as the case may be. This way, you can always make the electrodes bigger, if necessary without undue hardship. Remember that anything in the engine compartment should be mounted in a bullet-proof, vibration and temperature tolerant fashion.

6. If you have to drill a through-hole for wiring or plumbing through metal, make sure to also install a grommet for protection against chafing. Always watch your chamber pressure range from

IDLE (15-25 psi) - FULL POWER (30-60 psi) - Set your safety-pressure relief-valve to 75 psi and make sure it's rated for much higher.

7. Shut OFF the power switch and pull over if there is any MALFUNCTION of the system. Your engine will last longest when it still develops FULL POWER+ at some minimum temperature that we are sure you can find, by leaning back the Royal Vapour Flow and/or by making use of the water-vapour cooling technique (see Figure 7). Keep good mpg performance records, and periodic maintenance/inspection. Keep it clean; save some money; clean the air; heal the planet; happy motoring; tell a friend; enjoy your freedom and self-empowerment; haul ass.

8. There lacks documented material for perfecting this vapour system through a fuel injector; there may be some details you will discover on your own as working prototypes progress. For example, you may be restricted to inject the hydrogen/oxygen vapour WITHOUT ANY water vapour, as it may rust the injectors. If engine temp and CHT is a problem, then you will want to re-think your plan, e.g. ceramic-coating the injectors. There is always 'replacing the Fl system with a Carb'.

9. If you install the water-vapour system (for lower operating temp/stress), you will want to lean the mixture (vapour/air) for minimum vapour flow rate to achieve any given throttle position (idle - max). Make sure that you get a minimum flow for IDLE an a modestly sufficient flow for MAX, that does the cooling job without killing the combustion.

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

10. If you cannot find stainless steel pipe combinations that yield the 1-5mm gap, you can always regress back to alternating plates of +/- electrodes.

11. If you are concerned about the water freezing in your system, you can (a) add some 98% isopropyl alcohol and re-adjust the pulse frequency accordingly; or (b) install some electric heating coils.

12. Do not let ANYONE ever compromise your dream, your freedom, your independence, your truth.

REFERENCES

1. Stephen Chambers 'Apparatus for Producing Orthohydrogen and/or Parahydrogen' US Patent 6126794

2. Stanley Meyer 'Method for the Production of a Fuel Gas' US Patent 4936961

3. Creative Science & Research, 'Fuel From Water', fuelless.com

4. Carl Cella 'A Water-Fuelled Car' Nexus Magazine Oct-Nov 1996

5. Peter Lindemann's Free Energy site and his article from NEXUS 'Where in the World is All the Free Energy'

6. George Wiseman's Eagle Research
'The Gas-Saver and HyCO Series'

7. C. Michael Holler 'The Dromedary Newsletter' and 'SuperCarb Techniques'

8. Stephen Chambers 'Prototype Vapor Fuel System'

DISCLAIMER: The author of this document assumes no liability for the use or misuse of this information; which is made available as public-domain information, for the purposes of education, ecology, health, well-being, freedom, liberty, and pursuit of happiness.

COMMON LAW COPYRIGHT #285714: All rights to the use and duplication of these plans are hereby reserved for the People, in their efforts to heal and restore the environment. Dare to express your uniqueness and environmental ideals. This technology is an exercise in responsible self-determination.

Thank you for your purchase, please leave me your positive feedback and after I will leave you the same feedback enjoy!

Thank you Take care Best Regards U.M.S

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

Create PDF with GO2PDF for free, if you wish to remove this line, click here to buy Virtual PDF Printer

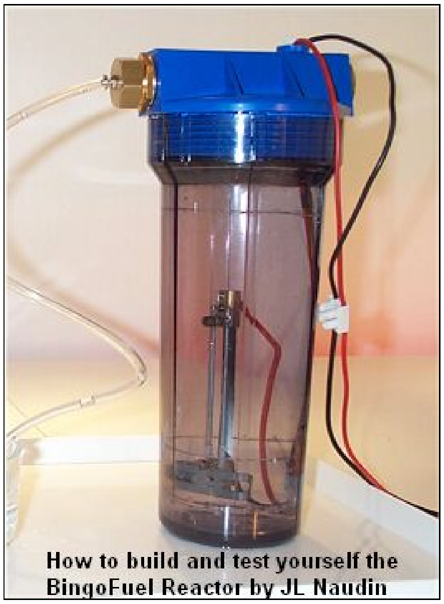
How to build and test yourself the
BingoFuel Reactor by JL Naudin

How to build yourself the BingoFuel Reactor v1.1

*created on April 2, 2003 - **JLN Labs** - Last update April 22, 2003*

Cliquez ici pour la version Française

Toutes les informations et schémas sont publiés gratuitement (freeware) et sont destinés à un usage personnel et non commercial
All informations and diagrams are published freely (freeware) and are intended for a private use and a non commercial use.

You will find in this page all the details for building and testing yourself the BingoFuel Reactor v1.1.

The BingoFuel Reactor v1.1 uses common parts which can be found easily in any plumbing shops, no machining and special tools are required. The main part used is a simple water filtration unit with its anti-scale cartdridge (see the photo below).

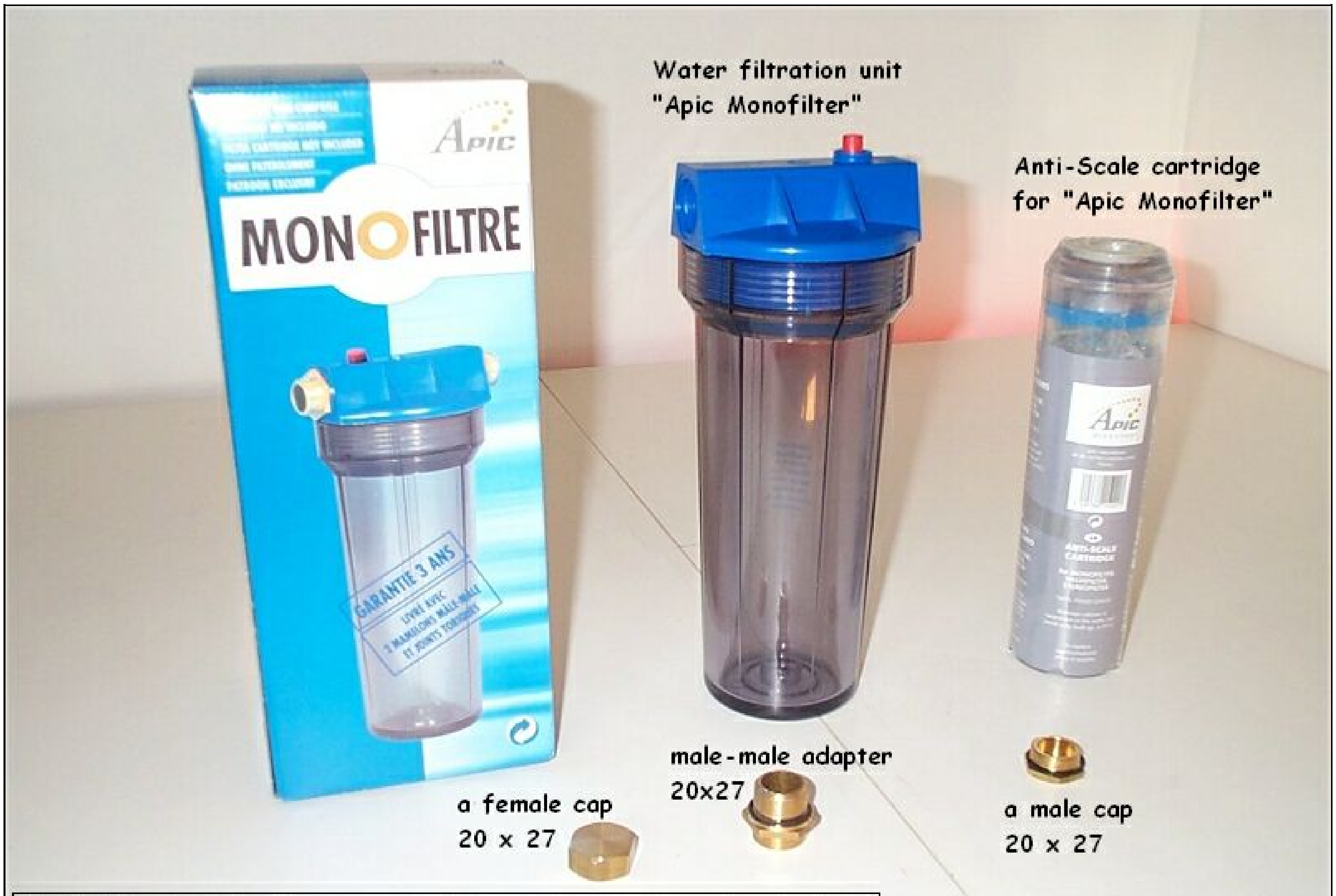
MONOFILTRE
Apic
Water filtration unit
"Apic Monofilter"
Anti-Scale cartridge
for "Apic Monofilter"
male-male adapter
20x27
a female cap
20 x 27
a male cap
20 x 27

BingoFuel Reactor v1.1 - The components for the reactor and and reaction chamber
(c) 2003 by Jean-Louis Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

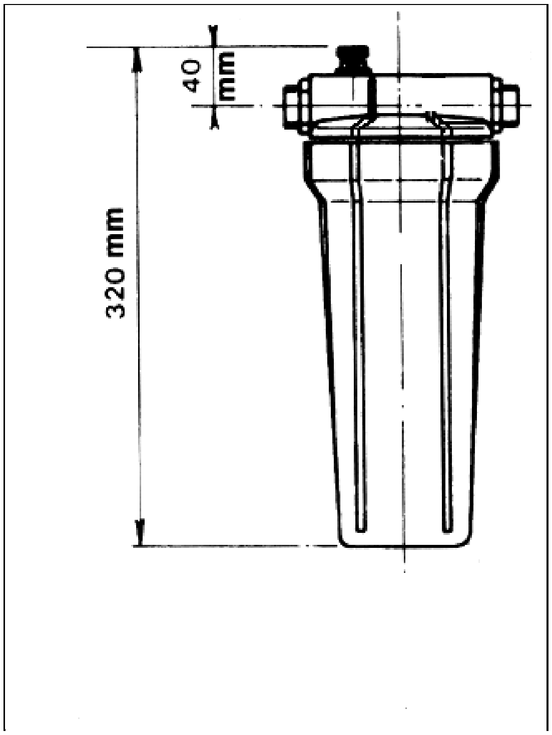

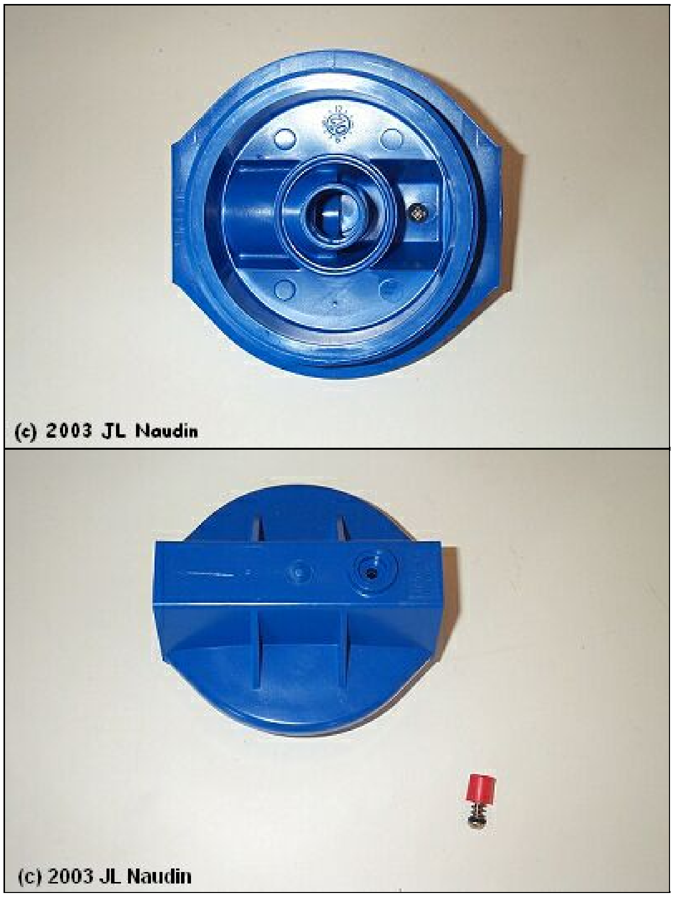

(c) 2003 JL Naudin

(c) 2003 JL Naudin

Step 1 : remove the purge valve (the red button) from the head of the water filtration unit.

(c) 2003 JL Naudin

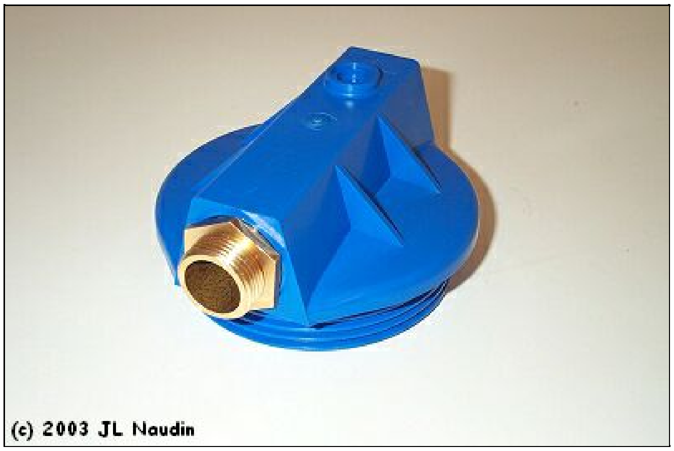

Step 2 : put the 20x27 brass cap at the input and the 20x27 female-female adapter on the device at the output
as shown in the pictures above

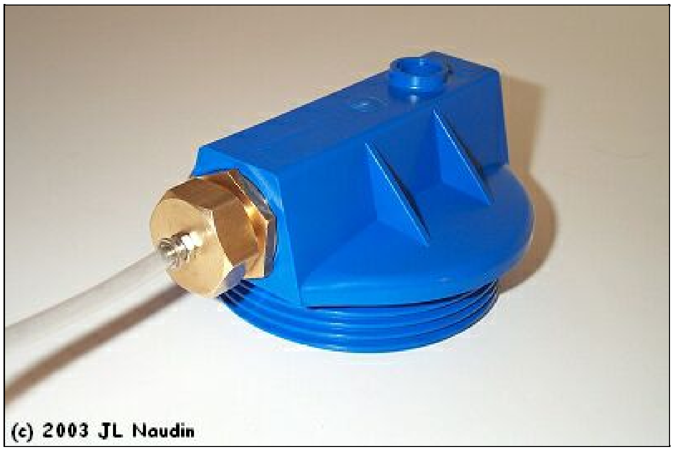

Step 3 : drill a small hole into the 20x27 female cap and put a silicon tube with an adapter on the cap,
and then screw the cap on the output of the head of the device

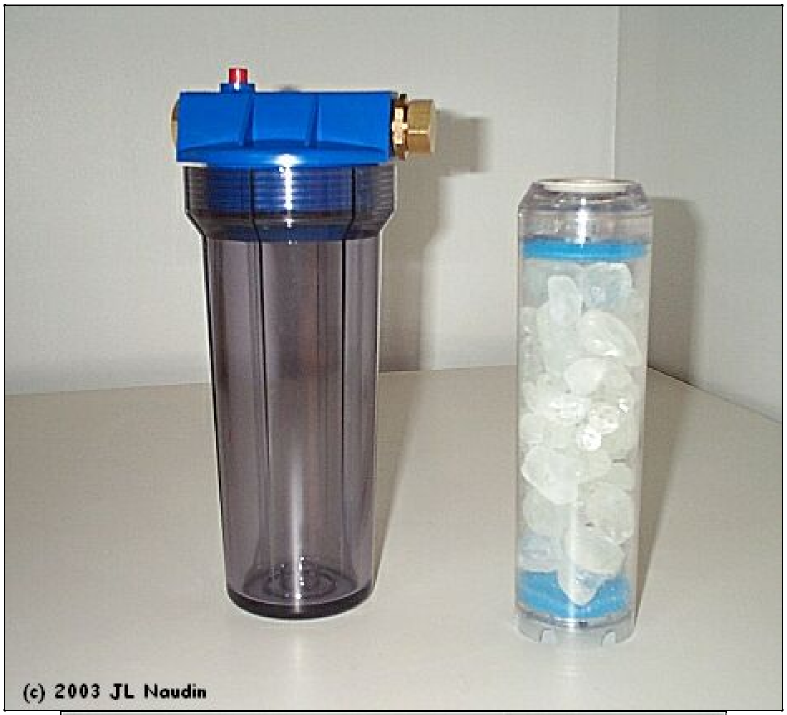
(c) 2003 JL Naudin

(c) 2003 JL Naudin

Step 4 : Take the anti-scale cartridge and remove the bottom plastic grid and then all the crystals inside the cartdridge.

Below, all the components required for building the plasma reaction chamber.

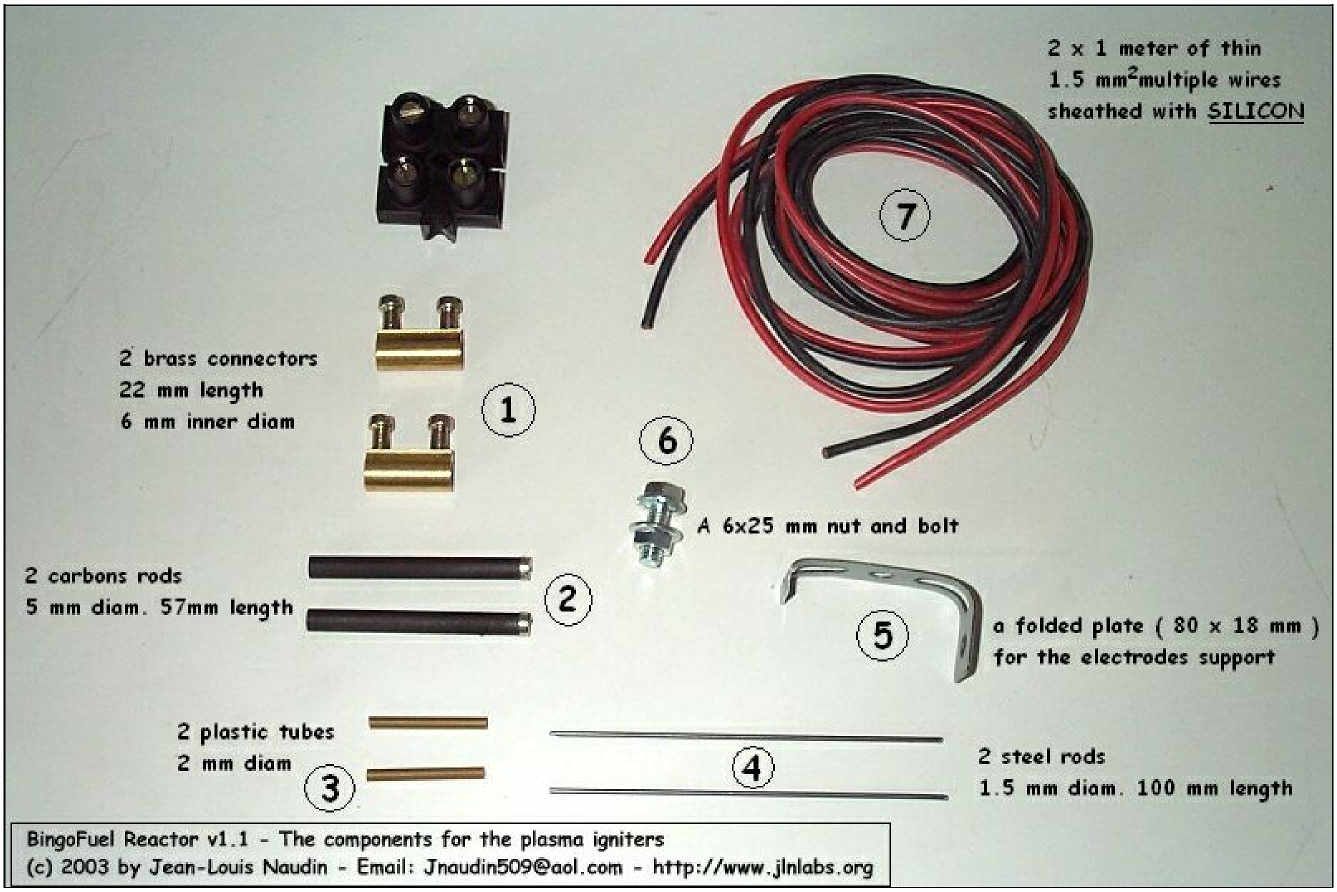
2 x 1 meter of thin
1.5 mm² multiple wires
sheathed with SILICON
7
2 brass connectors
22 mm length
6 mm inner diam
1
6
A 6x25 mm nut and bolt
2 carbons rods
5 mm diam. 57mm length
2
5
a folded plate (80 x 18 mm)
for the electrodes support
2 plastic tubes
2 mm diam
3
4
2 steel rods
1.5 mm diam. 100 mm length
BingoFuel Reactor v1.1 - The components for the plasma igniters
(c) 2003 by Jean-Louis Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Note : You must use two multiple thin wires cables (1.5 mm^2) sheated with silicon due to the overheating of the wire
produced by the strong current flow.

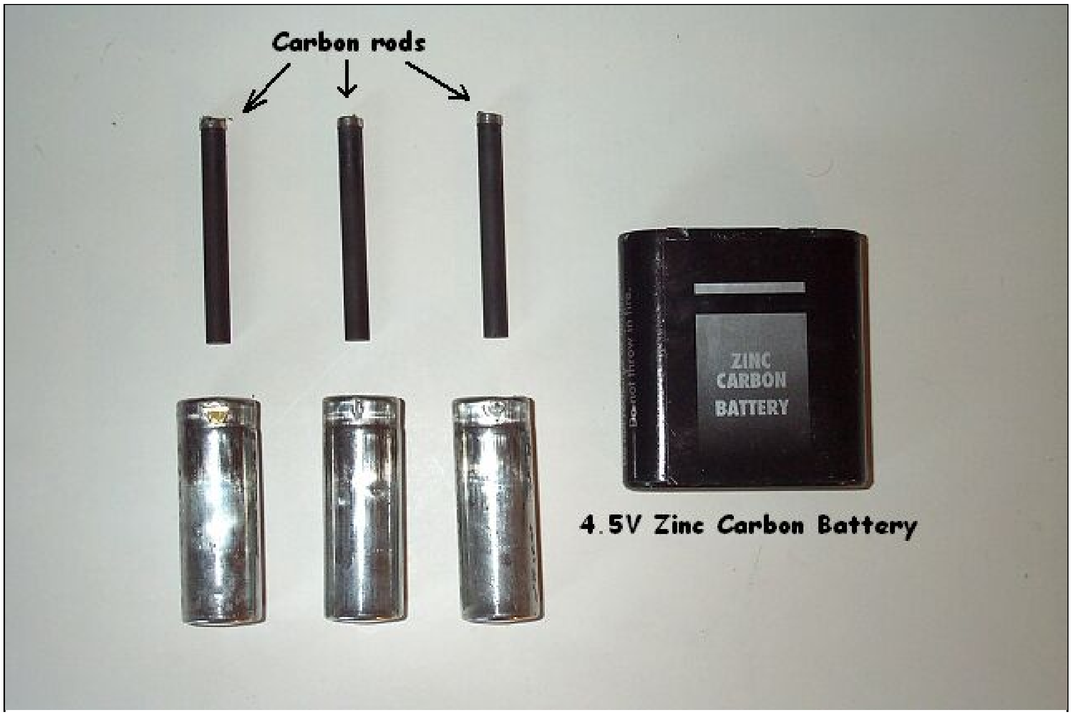
Carbon rods
ZINC
CARBON
BATTERY
4.5V Zinc Carbon Battery

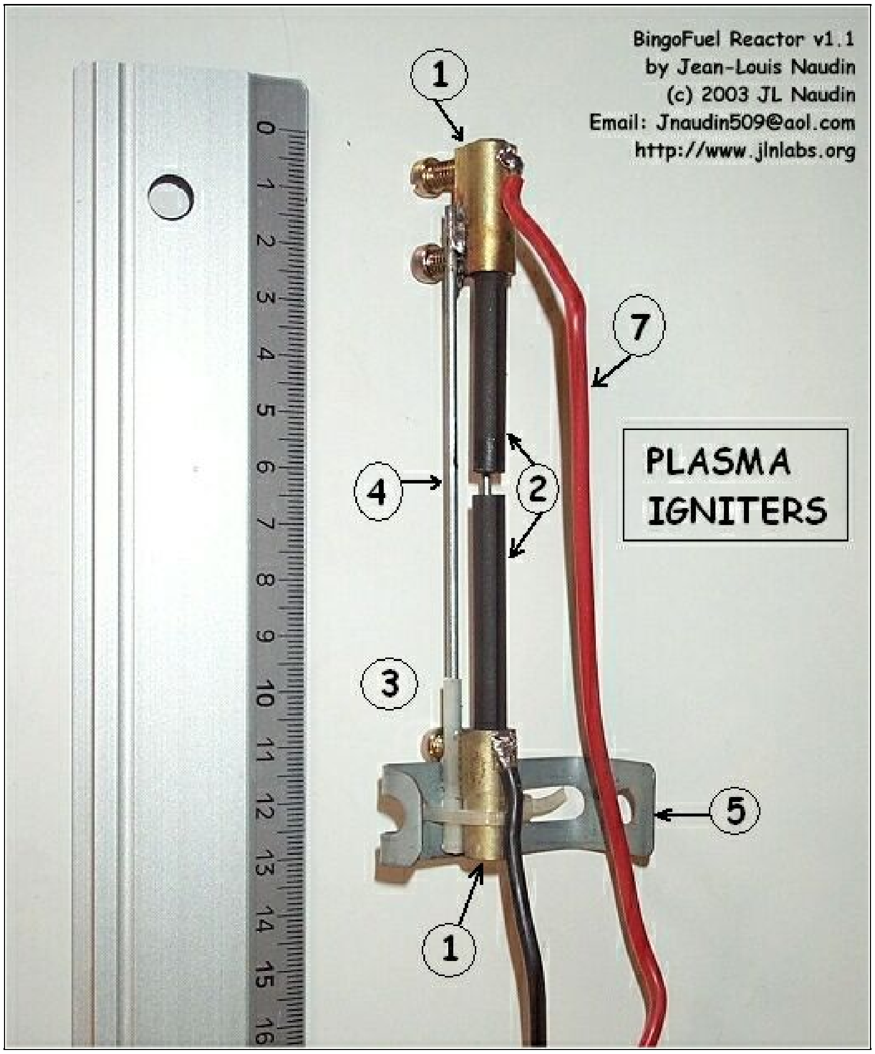
BingoFuel Reactor v1.1
by Jean-Louis Naudin
(c) 2003 JL Naudin
Email: Jnaudin509@aol.com
http://www.jlnlabs.org
PLASMA IGNITERS
1
2
3
4
5
7
1

Step 5 : Mount all the components as shown in the photo above

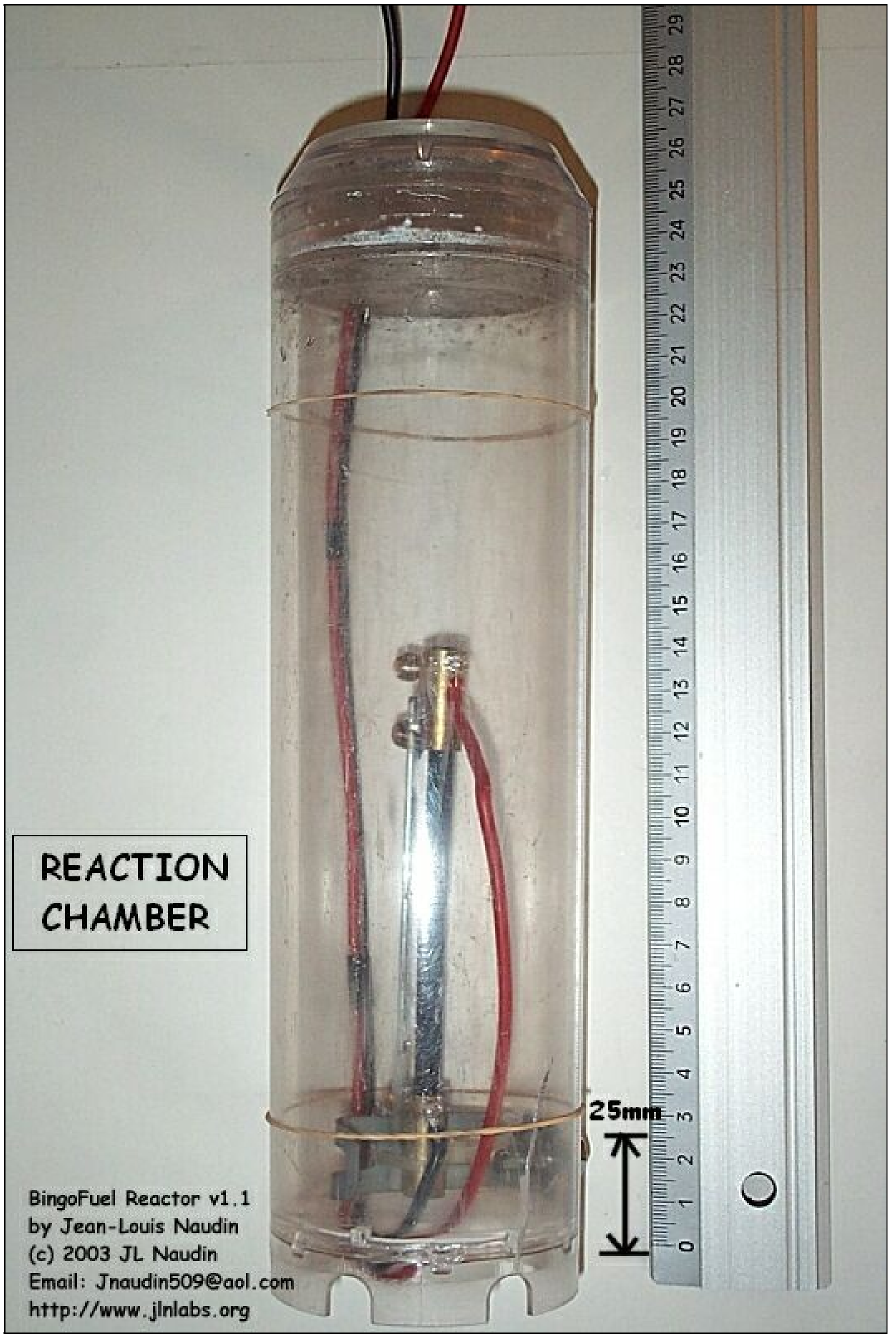
REACTION CHAMBER
25mm
BingoFuel Reactor v1.1
by Jean-Louis Naudin
(c) 2003 JL Naudin
Email: Jnaudin509@aol.com
http://www.jlnlabs.org

Step 6 : Drill a 6 mm diameter hole at 25 mm from the bottom of the plastic cartridge, then fix the plasma igniters with a 6x25 mm nut and a bolt as shown in the photo above

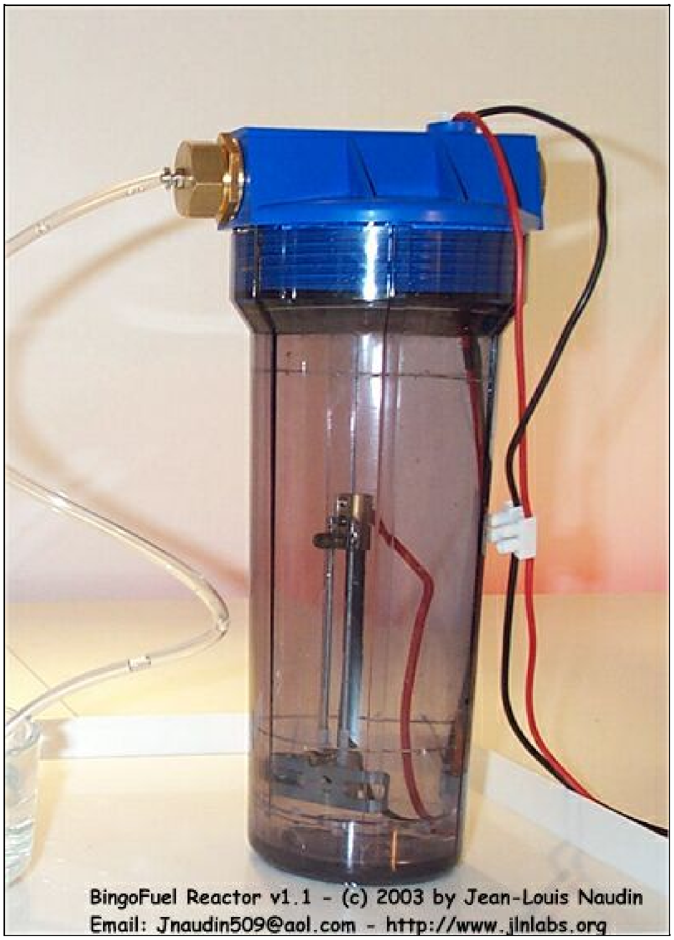
BingoFuel Reactor v1.1 - (c) 2003 by Jean-Louis Naudin
Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Step 7 : Place the reaction chamber into the transparent tank then fill the reactor with ordinary tap water

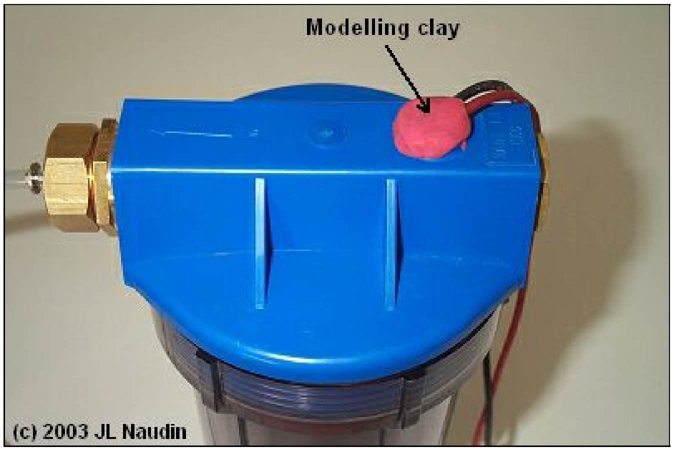
Modelling clay
(c) 2003 JL Naudin

Step 8 : Put modelling clay to close the hole around the electrical wires.

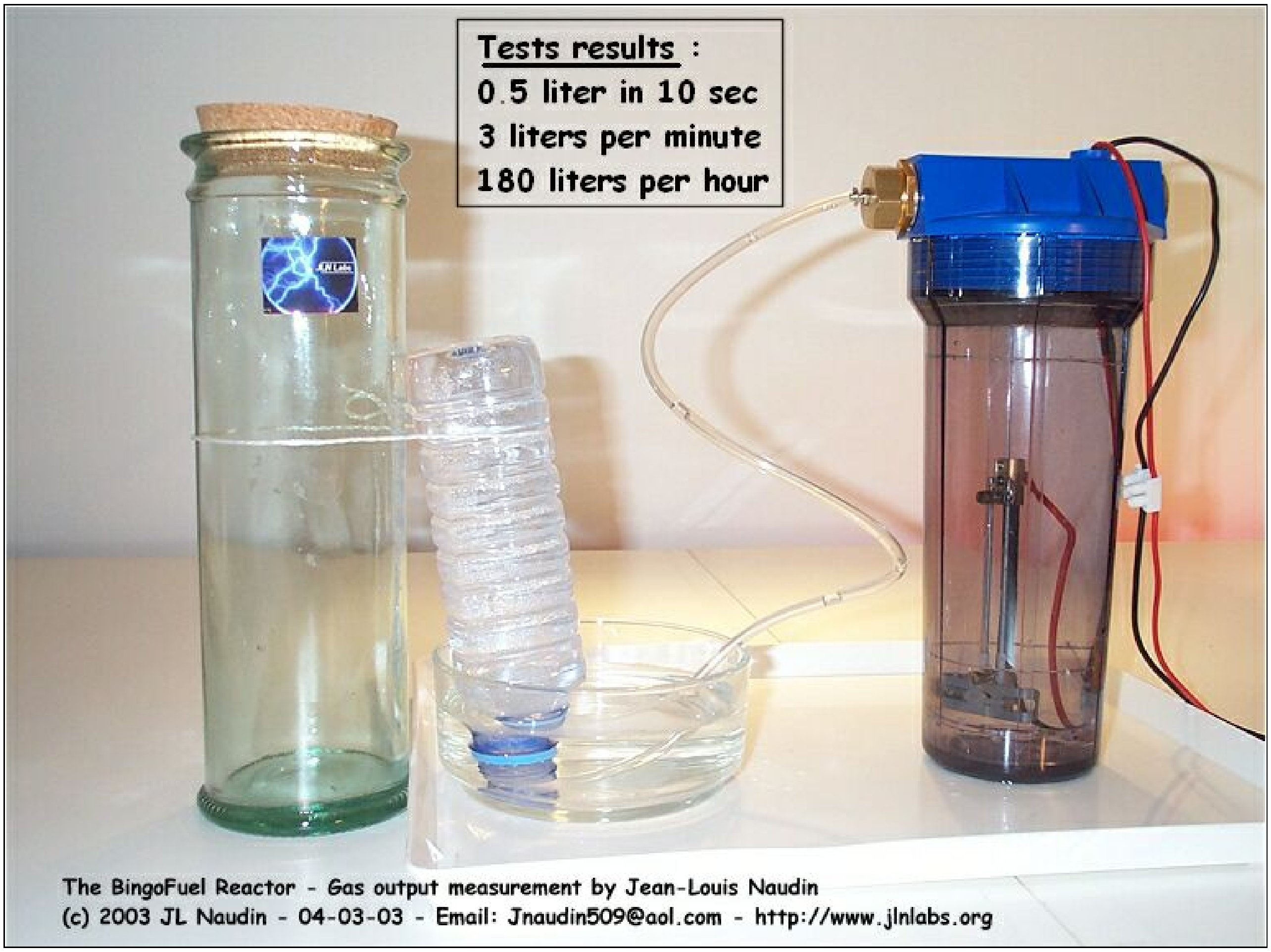

The BingoFuel Reactor - Gas output measurement by Jean-Louis Naudin
(c) 2003 JL Naudin - 04-03-03 - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Notes from Jean-Louis Naudin : The BingoFuel Reactor v1.1 is only a demonstrator, this is a proof of concept device which shows that the device works really. This version is able to produce a great amount of synthetic gas during some minutes. A new version of the BingoFuel Reactor is under development, this version (v2.0) will be able to run for a long time without tuning and adjustment...

Disclaimer : The author assumes no liability for any incidental, consequential or other liability from the use of this information. All risks and damages, incidental or otherwise, arising from the use or misuse of the information contained herein are entirely the responsibility of the user. Although careful precaution has been taken in the preparation of this material. **Be Carefull**, you must conduct this test in **a well ventiled room or better in open air**, you must not smoke during the test. **This experiment is not intended for the inexperienced.** User of this document should be very carefull to try anything out ! If you do it, the risk of any results is just yours. I take no responsibility of anything that might happen, let it be of a wrong information or anything else.

See also :

(c) 2003 JL Naudin

-
CFR v1.2 for BingoFuel production
(c) 2003 by JL Naudin - May 13, 2003
The CFR for producing BingoFuel...

Email : JNaudin509@aol.com

---

return to the BingoFuel project home page

Free web stats

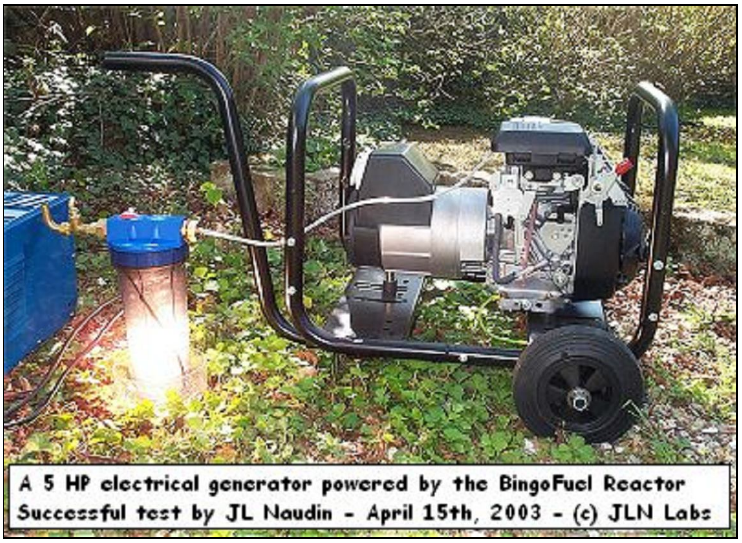

A 5 HP electrical generator powered by the BingoFuel Reactor
Successful test by JL Naudin - April 15th, 2003 - (c) JLN Labs

A 5HP Electrical Generator fully powered with the BingoFuel Reactor

*created on April 2, 2003 - **JLN Labs** - Last update April 18, 2003*

☐ **Cliquez ici pour la version Française** ☐

Toutes les informations et schémas sont publiés gratuitement (freeware) et sont destinés à un usage personnel et non commercial

All informations and diagrams are published freely (freeware) and are intended for a private use and a non commercial use.

On April 15th, 2003, an Electrical Generator powered with a 5HP (160 cm^3) 4-stroke combustion engine (a Honda GC160) has been tested successfully with the *BingoFuel Reactor*. The 5HP combustion engine has been fully powered with synthetic gas produced by the *BingoFuel Reactor*.

The Electrical Generator tested with the *BingoFuel Reactor* is a Ranger2500 from SDMO (see below) :

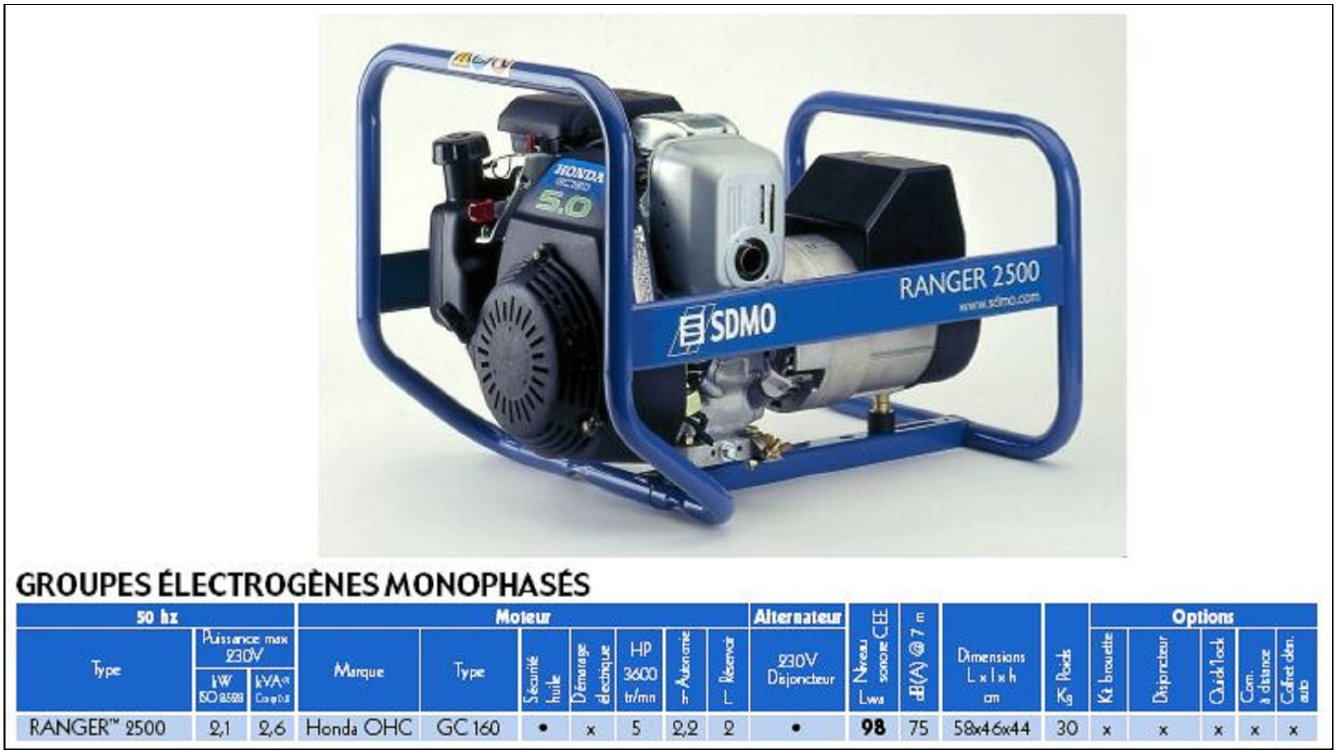

GROUPES ÉLECTROGÈNES MONOPHASÉS

50 hz			Moteur							Alternateur					Options				
Type	Puissance max 230V		Marque	Type	Sécurité huile	Démarrage électrique	HP 3600 tr/mn	h Autonomie	L Réservoir	230V Disjoncteur	Niveau sonore CEE Lwa	dB(A) @ 7 m	Dimensions L x l x h cm	Kg Poids	Kit brouette	Disjoncteur	Quick lock	Com. à distance	Coffret dém. auto
	kW ISO 8528	kVA Cos φ 0.8																	
RANGER™ 2500	2,1	2,6	Honda OHC	GC 160	•	x	5	2,2	2	•	**98**	75	58x46x44	30	x	x	x	x	x

	ENGINE SPECIFICATIONS
MODEL	**HONDA GC160**
Type	4-stroke, overhead camshaft single cylinder
Total Displacement	160 cm^3 (9.8 cu in)
Bore & Strike	64 x 50 mm (2.5 x 2.0 in)
Max Horsepower (Gross)	3.7 kw [1](5.0hp) at 3,600 rpm
Max Torque (Gross)	10.3 N·m (1.05 kgf·m, 7.6 lbf·ft) at 2,500 rpm
Compression Ratio	8.5: 1
Fuel Consumption	313 g/kWh (230 g/HPh, 0.51 lb/HPh)
Cooling System	Forced-air
Ignition System	Transistorized magneto ignition
Ignition Timing	20° B.T.D.C
Spark Plug	BPR6ES (NGK)
Carburetor	Horizontal type, butterfly valve
Air Cleaner	Dry (paper) type
Governor	Centifugal mechanical governor
Lubricating System	Splash
Oil Capacity	0.58 lt (0.61 US qt, 0.55 Imp qt)
Recommended operating ambient tempature	-15°C to 40°C (5°F to 104°F)
Starting System	Recoil starter
Stopping System	ignition primary circuit ground
Fuel Used	Automotive unleaded gasoline (minimum 86 pump octane)
Fuel tank capacity	2.0 lt (0.53 US gal, 0.44 imp gal)
P.T.O. Shaft Rotation	Counterclockwise (viewed from P.T.O. side)

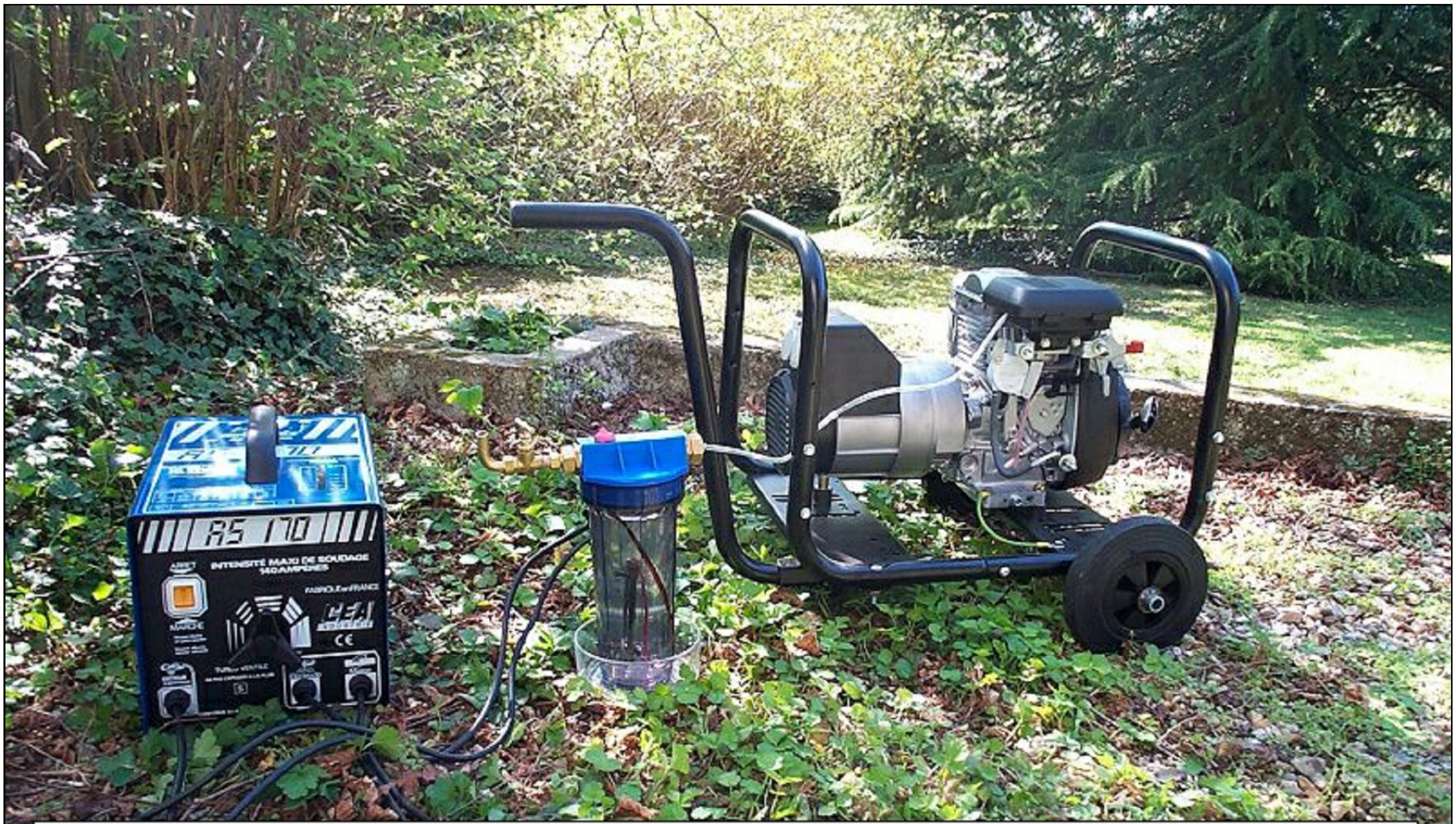

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

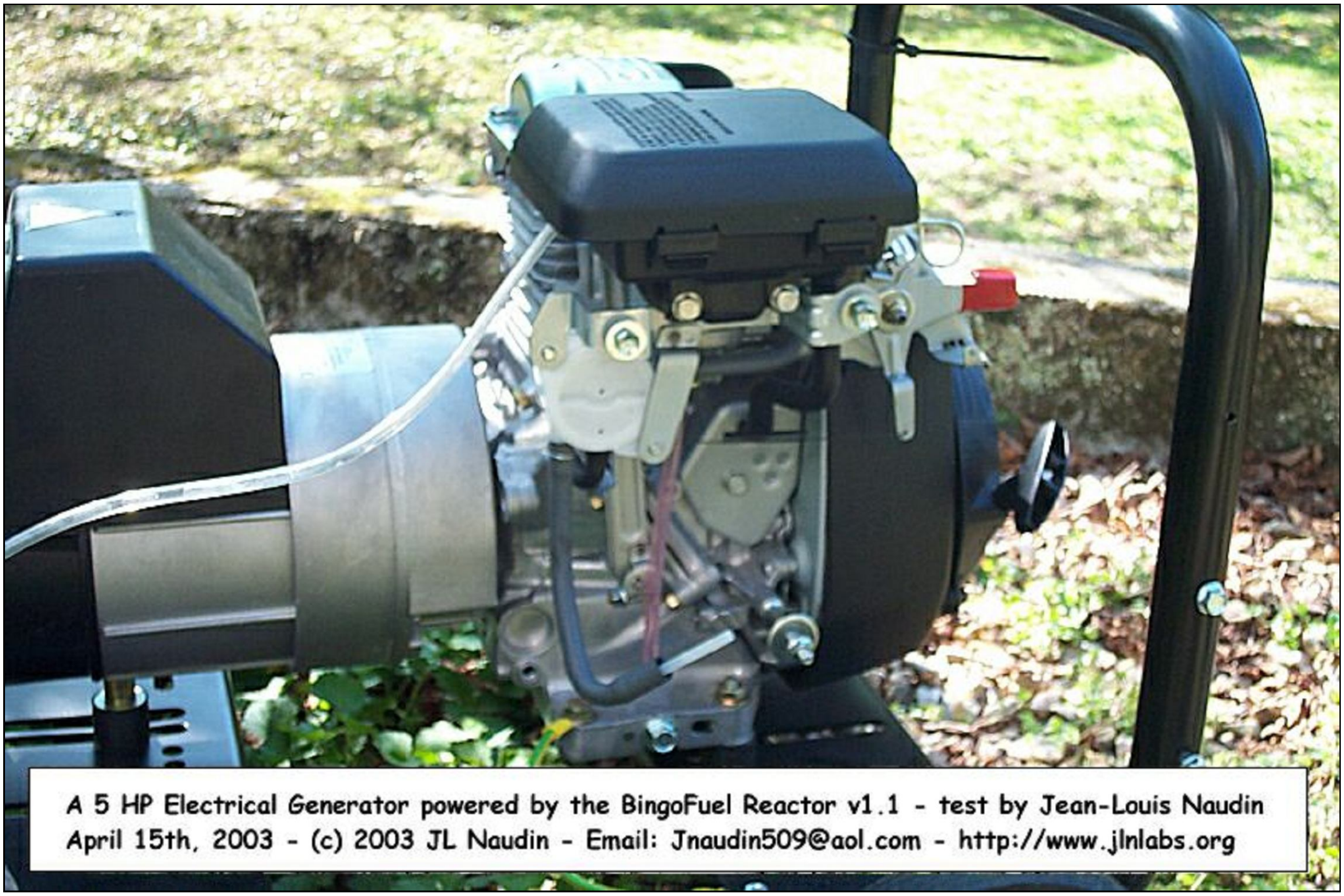
A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Photo above: The fuel tank (not used here) has been completly removed for this test.

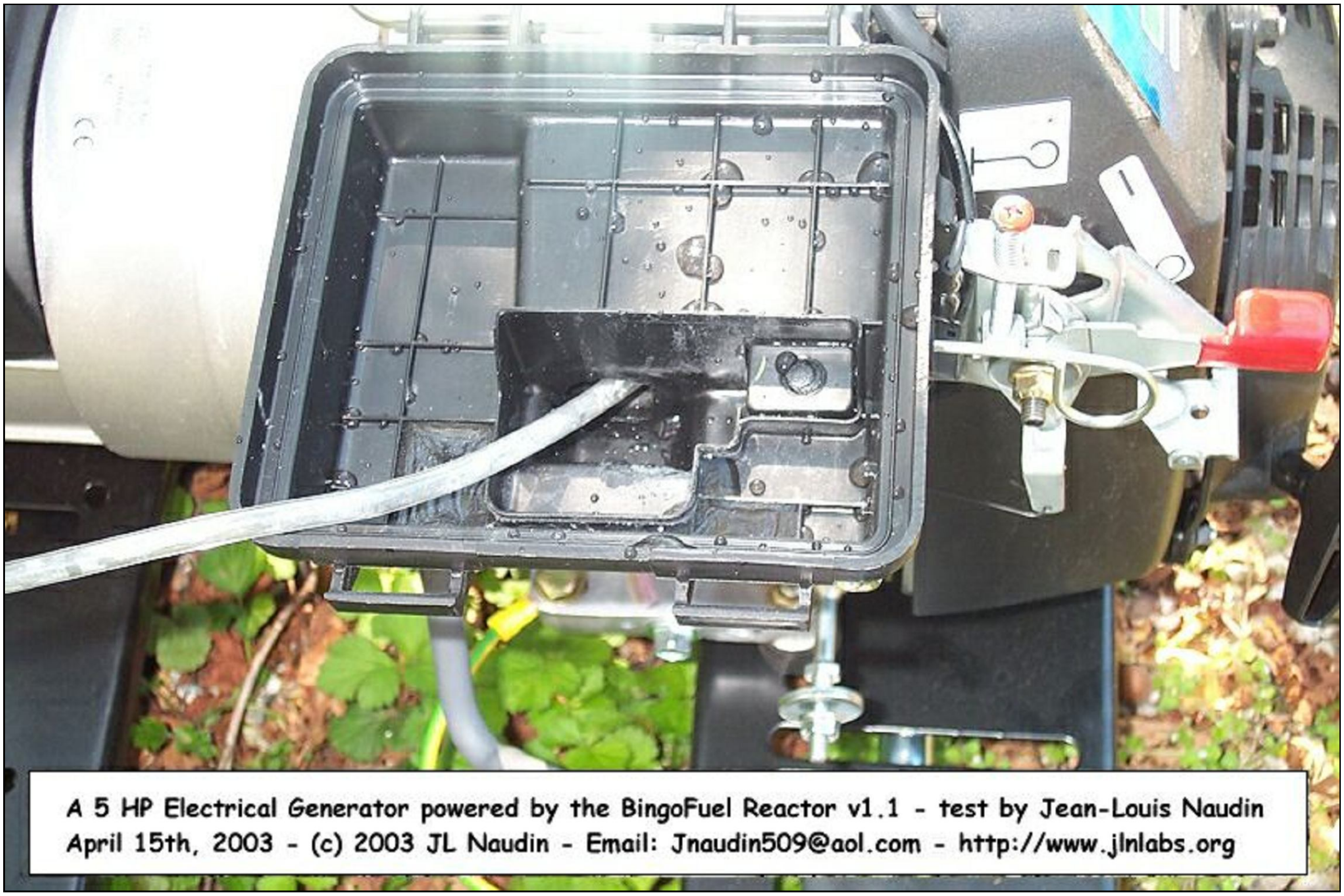
A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Photo above: The air filter has been removed and the synthetic gas output is directly placed at the carburettor input.

Tests results with the BingoFuel Reactor v1.1

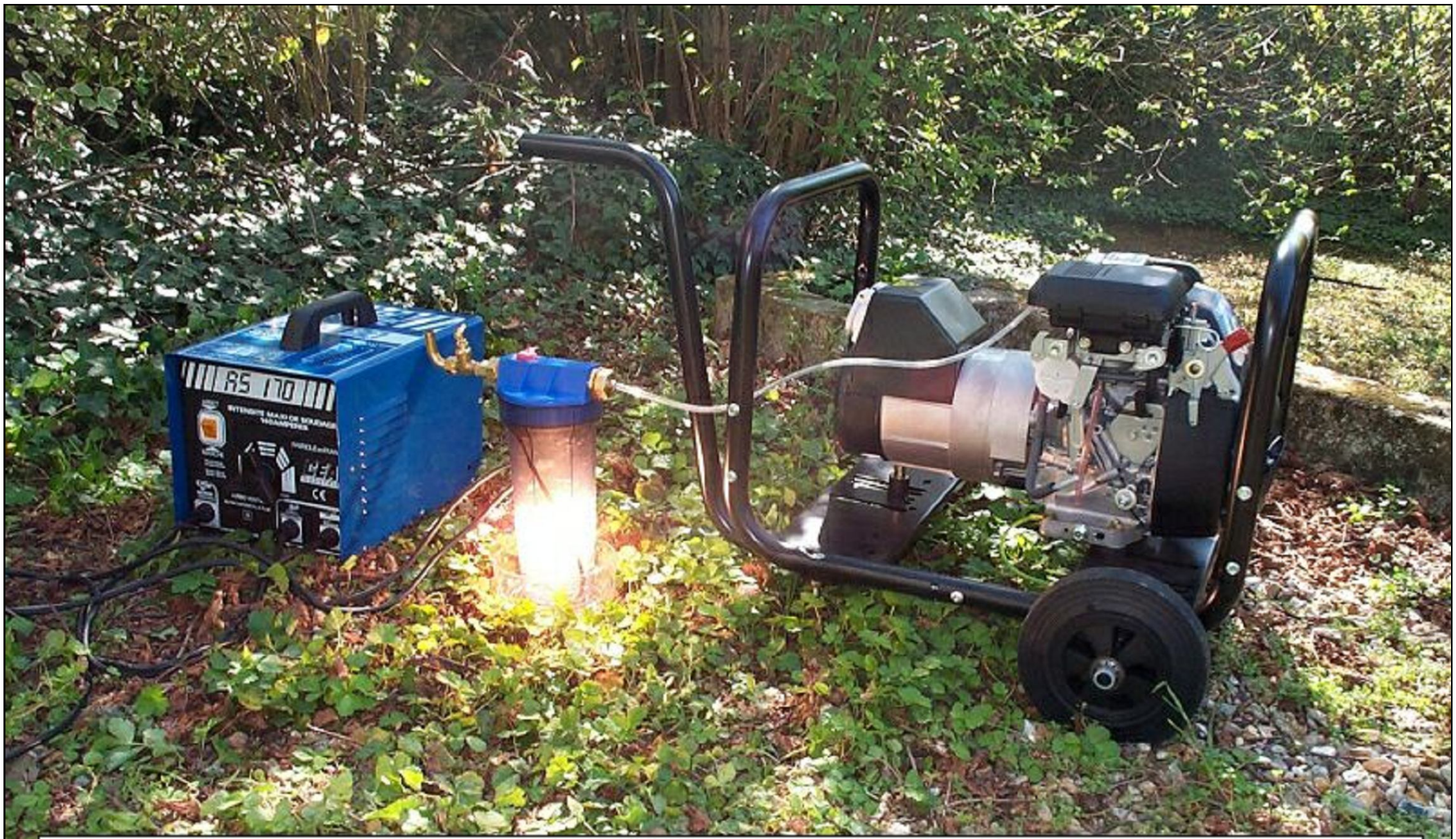

A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

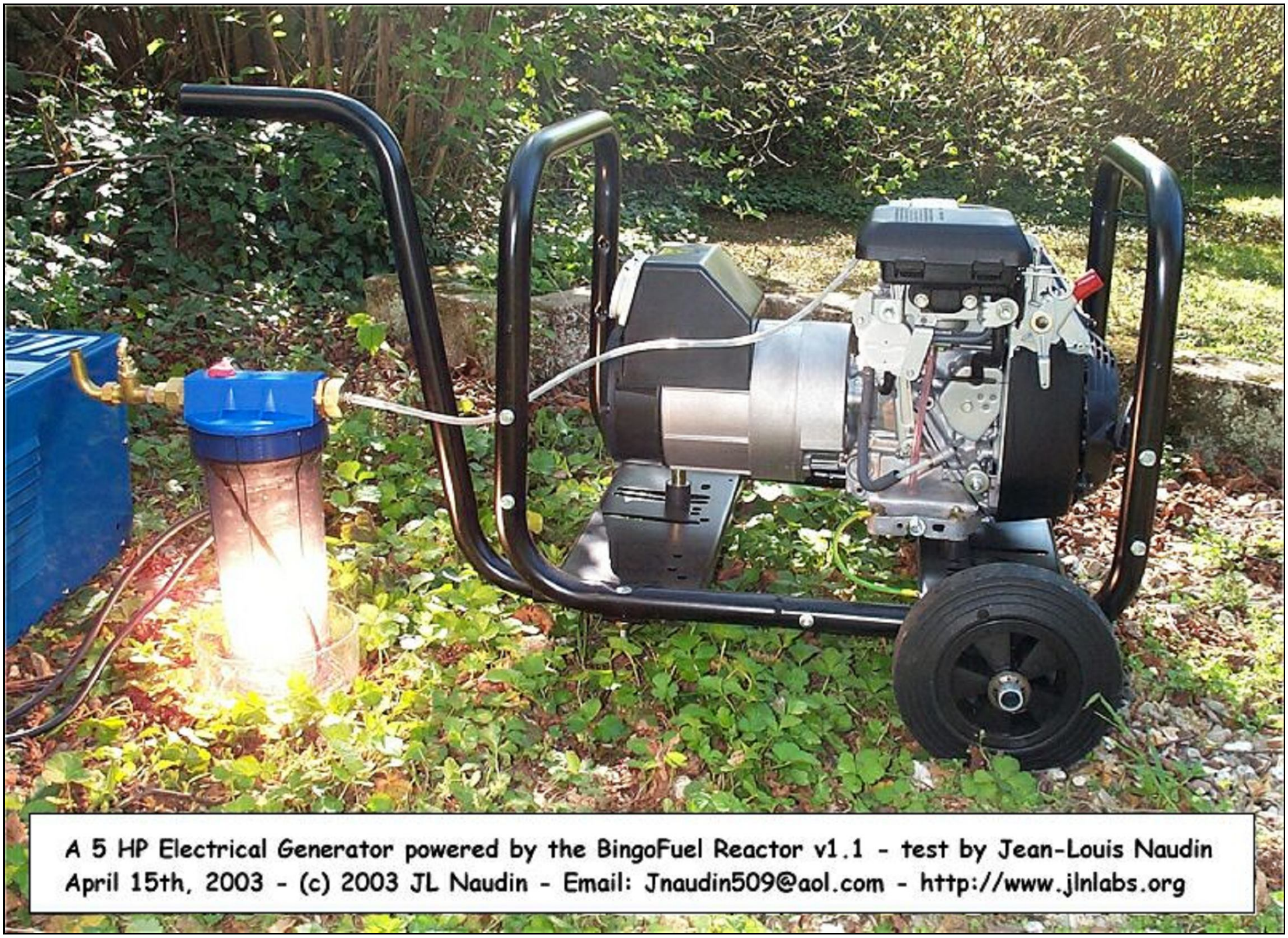
A 5 HP Electrical Generator powered by the BingoFuel Reactor v1.1 - test by Jean-Louis Naudin
April 15th, 2003 - (c) 2003 JL Naudin - Email: Jnaudin509@aol.com - http://www.jlnlabs.org

Two tests runs have been conducted successfully, in all the cases the 5HP engine has worked without any problem.

See the video of the 5HP engine fully powered with the BingoFuel Reactor

To see the video, the free downloadable RealPlayer is required

You may download free the RealPlayer 8 Basic at : http://proforma.real.com/real/player/blackjack.html

RealOne Player
Fichier
A 5HP Electrical Generator powe...
0:00 / 1:32

A 5 HP Power Generator fully powered by the BingoFuel Reactor v 1.1

April 15, 003

(c) 2003 by JL Naudin

Email: Jnaudin509@aol.com

http://www.jlnlabs.org

[Click on the picture above to see the video (937 Kb)](#)

Comments from JL Naudin : These tests are very encouraging and confirms fully that the synthetic gas generated by the *BingoFuel Reactor* can be used as fuel for a common combustion engine...

See also the previous tests :

- 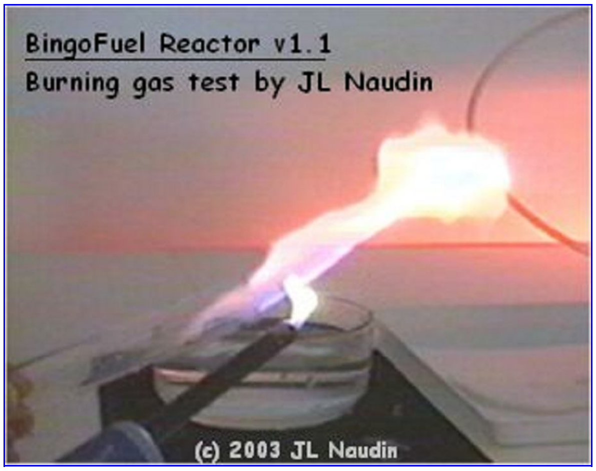
BingoFuel Reactor v1.1
Burning gas test by JL Naudin
(c) 2003 JL Naudin
  Burning gas tests generated by the *BingoFuel Reactor*

Email : **JNaudin509@aol.com**

---

return to the BingoFuel project home page

Weboscope free

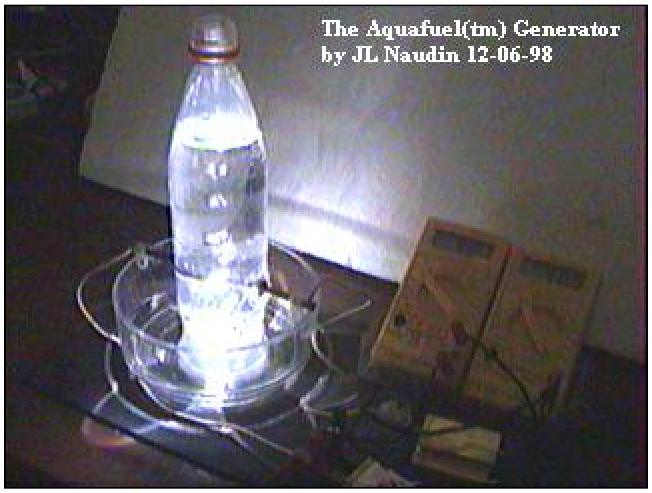

The AquaFuel® generator

How to generate gas from water for use as Fuel

created on 12-06-98 - ***JLN Labs*** - last update on 04-03-2003

Toutes les informations et schémas sont publiés gratuitement (freeware) et sont destinés à un usage personnel et non commercial
All informations and diagrams are published freely (freeware) and are intended for a private use and a non commercial use.

Disclaimer: The author assumes no liability for any incidental, consequential or other liability from the use of this information. All risks and damages, incidental or otherwise, arising from the use or misuse of the information contained herein are entirely the responsibility of the user. Although careful precaution has been taken in the preparation of this material.

On 12-06-98, I have built and reproduced successfully a very simple AquaFuel™ generator:
The AquaFuel™ is fully based on the Hilliary Eldridge patent US 603,058 "Electrical Retort" granted on April 26, 1898.

This is a non fossil combustible gas which is produced by an electric discharge of carbon arcs within distilled, fresh, salt or other types of water, thus being essentially composed of Hydrogen, Oxygen, Carbon and their compounds.

Today, there are a lot of names for this synthetic gas such as Aquafuel™, Aqualene™, Magnegas™, TrueFuel™, Carbo-hydrogen™....

AquaFuel: An example of the emerging new energies and the new methods for their scientific study

Author: Ruggero Maria Santilli
Comments: 22 pages, Texture
Report-no: TTL-98-004
Subj-class: General Physics
Journal-ref: Hadronic Journal Supplement, Vol. 13, pp. 1-22, 1998

In this paper we initiate studies of the emerging new forms of energy by using as a representative example the new combustible gas called AquaFuel, discovered and patented by William H. Richardson, jr., whose rights are now owned by Toups Technology Licensing, Inc. (TTL), of Largo, Florida. In essence, AquaFuel is a new energy converter capable of transforming Carbon and water into a new combustible gas via an electric discharge. We show that AquaFuel can be produced easily, safely and rapidly in large amounts, and exhibits greatly reduced emission pollutants as compared to fossil fuels of current use. Despite its simplicity, the chemical and physical characteristics of AquaFuel are largely unknown at this writing. We then review nine basic experimental measurements which are necessary for a scientific appraisal of AquaFuel. We outline the limitations of quantum mechanics and chemistry for the treatment of {\it new} forms of energy, namely, energies which by definition should be {\it beyond} said theories. We finally point out the availability of broader theories specifically constructed for the study of new energies and point out available applications.

Link to this paper : http://arxiv.org/abs/physics/9805031

This generator produces a mixture of carbon monoxide and hydrogen (COH_2) and this is a gas wich burns very cleanly in oxygen or air, and it can be used as fuel for an internal combustion engine. When burned, COH_2 produces carbon dioxide and water vapor, so it generates very little, if any, pollution to the environnement.

Below, an analysis of this gas conducted by the NASA :

Hydrogen	46.483 %
Carbon Dioxide	9.329
Ethylene	0.049
Ethane	0.005
Acetylene	0.616
Oxygen	1.164
Nitrogen	3.818
Methane	0.181
Carbon Monoxide	38.370
Total	**100.015**

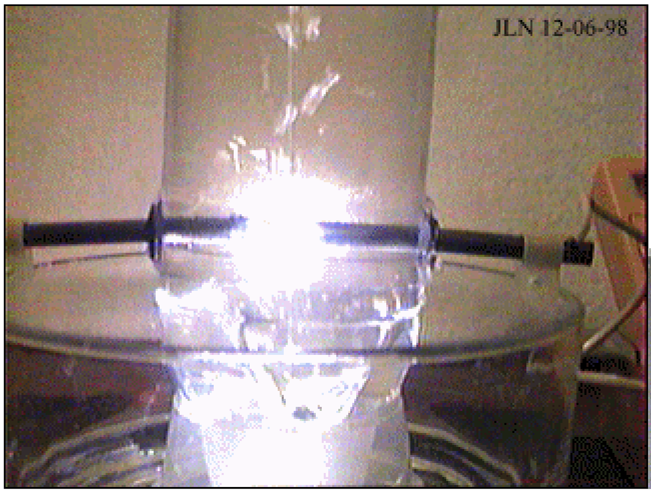

he AquaFuel(tm) Generator
Jean-Louis Naudin 12-08-98

This simple experiment is only for testing purposes and only for a proof of the concept. This small generator can't be used for a long working period and it must be used only for demonstration.

You need to get very few materials and it is very simple to build and test it....

Be Carefull, this device generates an explosive gas, *you must conduct this test in* ***a well ventiled room or better in open air****, you must not smoke during the test.. Don't forget that the carbon monoxide (CO) is a very toxic gas, so* ***never*** *breath this gas* ***before*** *burning it.* ***This experiment is not intended for the inexperienced.*** *User of* *this document should be very carefull to try anything out ! If you do it the risk of any results is just yours. I take no responsibility of anything that might happen, let it be of a wrong information or anything else.*

You need only to get :
- A little plastic soda bottle,
- two carbon rods (70mm length, 6mm diam)
- one 1 ohm 50Watts resistor
- a DC Power supply which is able to deliver 35v / 10A
- some wires, plugs and silicon cement.

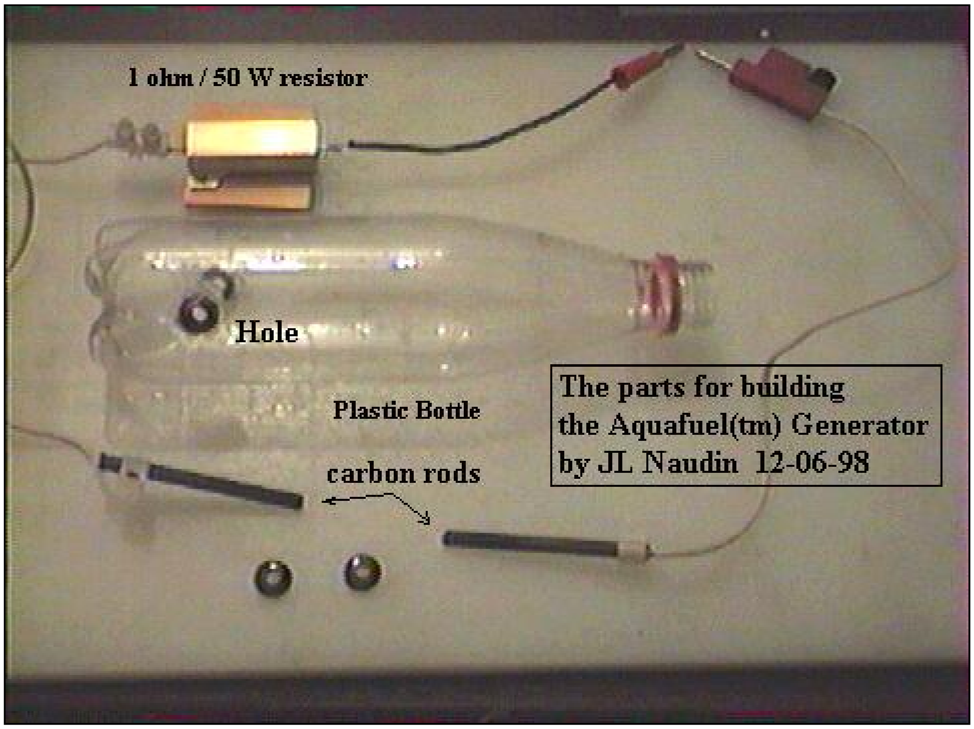
1 ohm / 50 W resistor
Hole
Plastic Bottle
carbon rods
The parts for building
the Aquafuel(tm) Generator
by JL Naudin 12-06-98

Very few material is needed.....

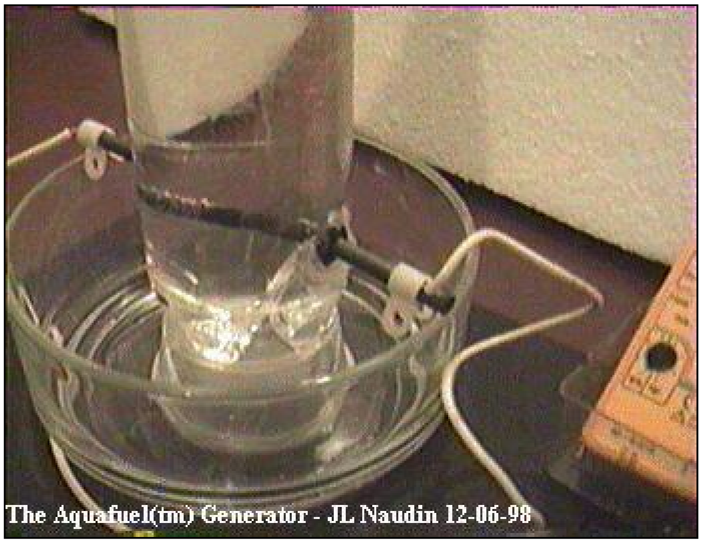
The Aquafuel(tm) Generator - JL Naudin 12-06-98

1) Drill two opposite holes (10mm diam) at 60mm from the bottom of the bottle and insert the carbon rods with a rubber washer and glue the washer with silicon cement. I suggest you to make one of the carbon rod into round shape. The two carbon rods must be just slightly in contact before you switch on (see below).

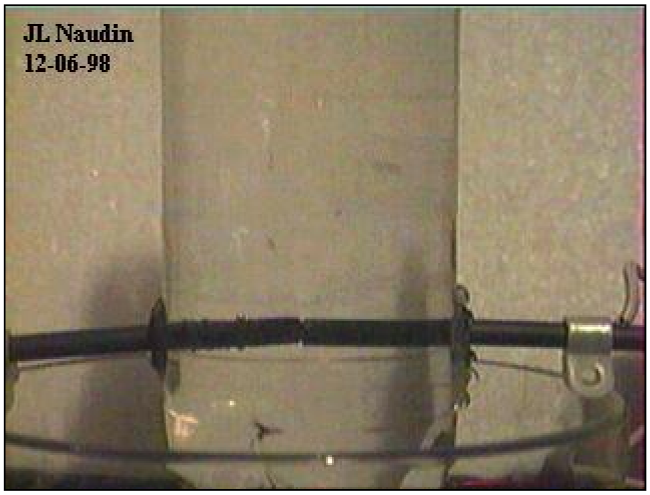
JL Naudin
12-06-98

2) Connect the 1 ohm 50W resistor in serie with one of carbon rod and one pole of your DC power supply (set for 34V/15A DC), the other pole of your power supply is connected to the other carbone rod. You may add some additional multimeters for measuring the Current and the Voltage input. Fill up your generator with only distilled or fresh water.

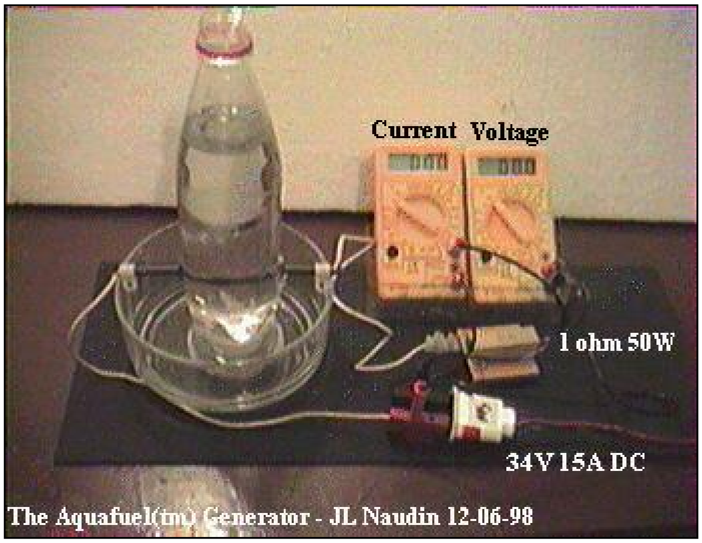
Current
Voltage
1 ohm 50W
34V 15A DC
The Aquafuel(tm) Generator - JL Naudin 12-06-98

3) Now, your are ready to produce AquaFuel™......

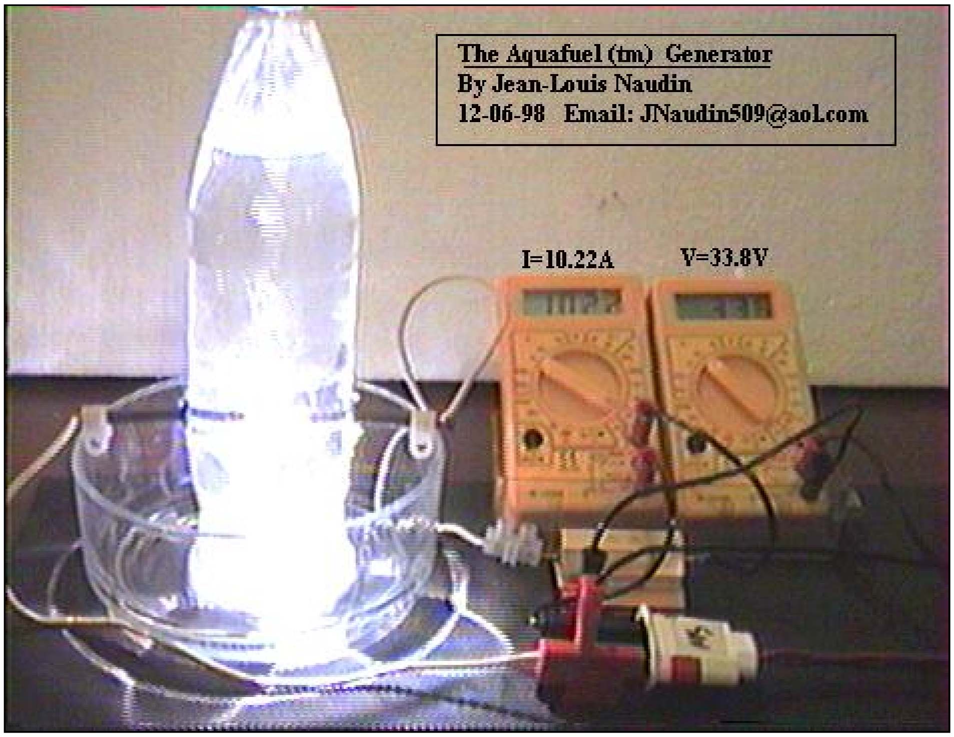
The Aquafuel (tm) Generator
By Jean-Louis Naudin
12-06-98 Email: JNaudin509@aol.com
I=10.22A
V=33.8V

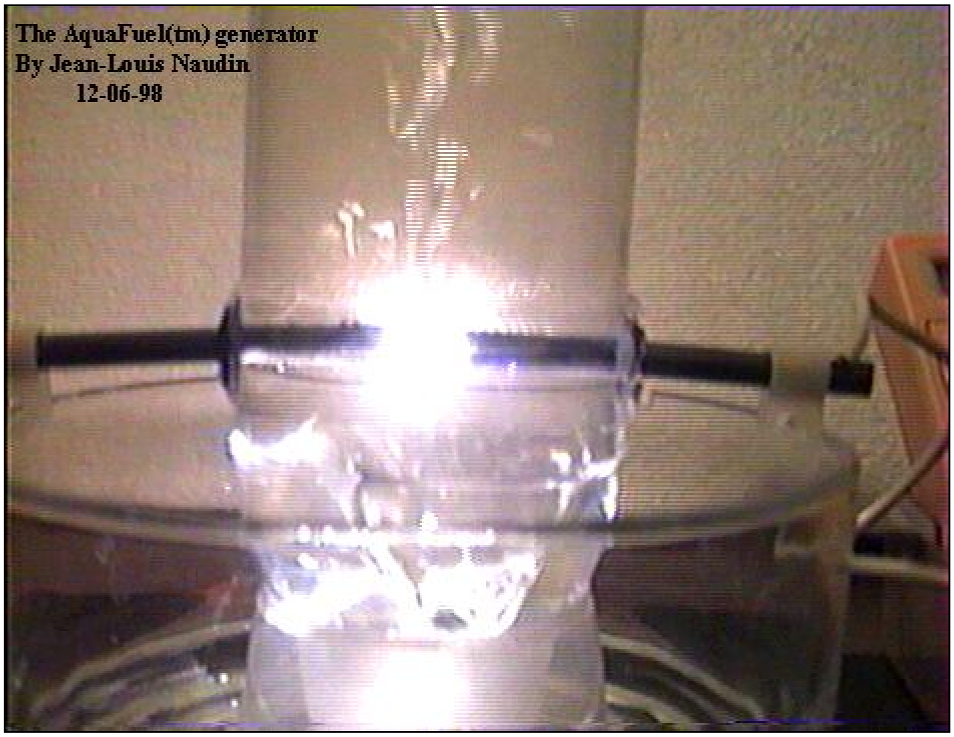
The AquaFuel(tm) generator
By Jean-Louis Naudin
12-06-98

The AquaFuel(tm) Generator
By Jean-Louis Naudin - 12-08-98

[Click here to see some animated videos of the AquaFuel(tm) generator in action]

You will see below some videos which show my AquaFuel(tm) in action and some personal comments.......

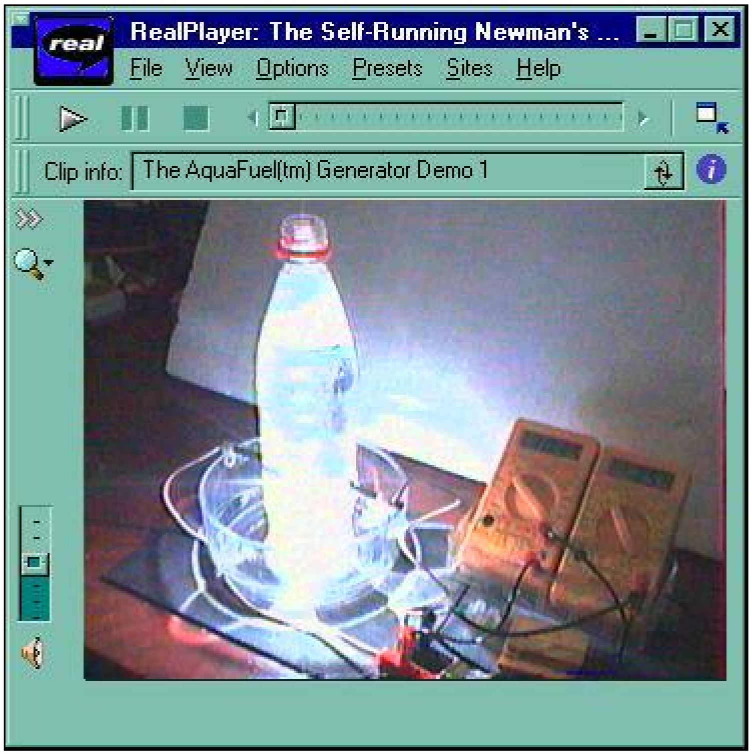
RealPlayer: The Self-Running Newman's ...
File
View
Options
Presets
Sites
Help

Clip info: The AquaFuel(tm) Generator Demo 1

click here to see the [AquaFuel(tm) demo 1] (277kb)

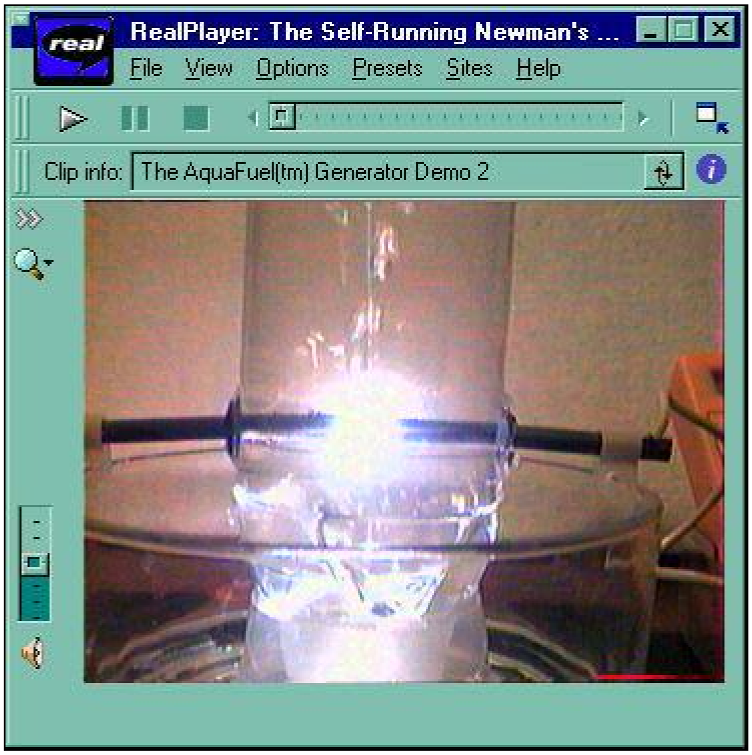
RealPlayer: The Self-Running Newman's ...
File
View
Options
Presets
Sites
Help

Clip info: The AquaFuel(tm) Generator Demo 2

click here to see the [AquaFuel(tm) demo 2] (90kb)

(if you don't have the RealPlayer 5.0, you may download it freely at :
http://www.real.com/products/player/)

Some reference documents and web link :

United States Patent 6,183,604 - Santilli - February - 6, 2001

Durable and efficient equipment for the production of a combustible and non-pollutant gas from underwater arcs and method therefor

Abstract
A system for producing a clean burning combustible gas comprising an electrically conductive first electrode and an electrically conductive second electrode. A motor coupled to the first electrode is adapted to move the first electrode with respect to the second electrode to continuously move the arc away from the plasma created by the arc. A water tight container for the electrodes is provided with a quantity of water within the tank sufficient to submerge the electrodes.

Inventors: Santilli; Ruggero Maria (Palm Harbor, FL)
Assignee: Hadronic Press, Inc. (Palm Harbor, FL)

Field of the Invention

The present invention relates to durable and efficient equipment for the production of a combustible and non-pollutant gas from underwater arcs and method therefor and more particularly pertains to producing a combustible gas from the underwater arcing of electrodes moving with respect to each other.

Click here to see the full patent

Link to the MagneGas™ technology web site

United States Patent 6,217,713 - Lee , et al. - April 17, 2001

Process for producing aquafuel by using carbon fiber bundle electrodes

Abstract
The invention concerns a process for producing aquafuel by replacing conventional inflexible carbon bars with thin, flexible and tough carbon fiber bundles as consumptive electrodes which thereby can be sustainedly fed and can produce aquafuel continuously. Such carbon fiber bundle electrodes can be prepared by pultrusion, and electrodes may be further carbonized or graphitized in order to increase the conductivity and gas productivity thereof.

Inventors: Lee; Chi-Young (Hsin-Chu Hsien, TW); Chen; Swe-Kai (Hsin-Chu, TW); Tai; Nyan-Hwa (Hsin-Chu, TW)
Assignee: National Science Council of Republic of China (Taipei, TW)

Field of The Invention
The invention relates to a process for producing aquafuel by using carbon fiber bundle electrodes, and in particular, to a process for producing aquafuel by auto-feeding flexible carbon fiber bundle electrode in an electrolytic reaction, wherein said flexible carbon fiber bundle electrodes are formed by impregnating flexible, tough carbon fibers with a resin and said electrode can be used sustainedly in the electrolytic reaction of water under low voltage for a long period of time.

Click here to see the full patent

- **US 603,058 - Electrical Retort by Hilliary Eldridge April 26, 1898 (*the TRUE original Aquafuel patent*)**
- **US 5,159,900 - Method and means of generating gas from water for use as a fuel** *(Aquafuel Patent)*
- **US 5,417,817 - Biomass gasification process and apparatus**

- **US 5,435,274 - Electrical Power Generation Without Harmful Emissions** *(Aquafuel Patent)*
- **US 5,692,459 - Pollution Free Vehicle Operation**
- **US 5,792,325 - Electric Arc Material Processing System** *(Aquafuel Patent)*
- **US 5,826,548 - Power Generation Without Harmful Emissions** *(Aquafuel Patent)*
- **US 6,299,738 - Fuel Gas production by underwater arcing** *(Aquafuel Patent)*
- **EP1227142 - Method and system for producing hydrogen from solid carbon materials and water**
- **Infinite Energy Vol.2, No9, 1996**
- **Infinite Energy Vol.2, No10, 1996**
- **Infinite Energy Vol.2, No11, 1997**
- **Infinite Energy Vol.4, No19, 1998**
- **Aqualux Corporation web site** ***(disabled)***
- **MagneGas™ technology web site** *from Hadronic Press Inc*
- **A new fuel produced from water and carbon** *by Alexander's Gaz and Oil production*
- **Le gaz à l'eau** ***(gaz de houille)***
- **The scientific research undelying Santilli's Plasm-Arc-Flow reactors, Magnegaz(Tm), and Magneliquids(Tm)** *from IBR*
- **AquaFuel** *by the Morgan Energy Group*
- **AquaFuel Vs Diesel**, *analysis the Morgan Energy Group*
- **AquaFuel Vs Gasoline**, *analysis the Morgan Energy Group*
- **The TrueFuel™ : Technical informations** *by Advanced Energy Research Corporation*
- *The Aquafuel™ is a trademark from Aqualux Corporation*
- *The Magnegas™ is a trademark from Hadronic Press Inc*
- *The TrueFuel™ is a trademark from Advanced Energy Research Corporation*
- *The Carbo-Hydrogen™ is a trademark from DW Energy Research*

Email : **JNaudin509@aol.com**

Return to the JLN Labs home page

# 359

Creative
SCIENCE
& Research

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Information Big Brother & the Oil Companies do not want you to know!

Free Electricity

From the Phone Co.

Time is running out

www.FuellessPower.com ● www.Fuelless.com

Cover Page

Creative SCIENCE & Research

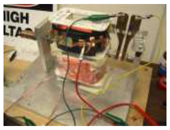

359

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Question: Is there enough useable energy from my phone line to light a 100 watt light bulb?

Answer: Yes, but only for a short time. If you are looking for a way to light up your home then build our fuelless engine.(see our web site at: www.fuellesspower.com or write us for a catalog.) This engine is powerful enough to run a 10 kw 120 vac 60 hz generator which can run your home.

You can still use your incoming phone line as a useful light source. You can get enough light to light up a room or two by using a diode and a large 100 volt capacitor bank to collect the 50 vdc. What you then want to do, is to use a high efficient lighting source, such as ultra bright white diodes, White neon lighting or flouresent bulbs. (great for emergancy lighting in any storm.)

The 50 vdc current you have collected will not light Neon or Flourescent bulbs until the voltage is much higher. 300 to 5,000 vac, You can buy a 15,000 volt neon transformer and pulse the 50 vdc into the primary using a wind up spring pulse switch (food timers) or build a low milliamp transistor type pulse generator, this will turn the DC to AC which you can then collect from the secondary out put of your transformer. You can also build our #363-a step up DC (No amperage loss! 100% eff. Using Photo caps) Or just for a quick fun experiment charge 20 - 50 volt capacitors, with each capacitor having about 4700 uf (micro farads), simply fully charge each capacitor by hand then connect them in series using Radio Shack allegator clips (wire), just like batteries, Negative to Positive, **Warning! Use rubber gloves when working with High Volatge, High Voltage can kill!**

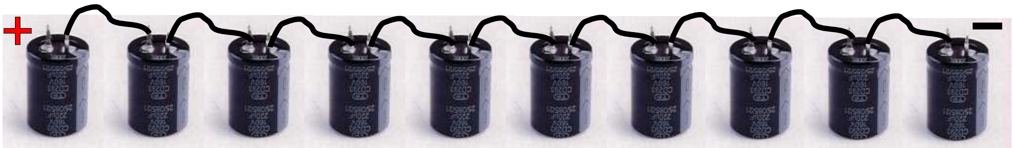

Once you are done with connecting them in series, use a small low amp (Miliamp) isolation transformer which you can buy from any electronic supplier or Radio Shack store. (Search for Electronic Suppliers on web search engine) connect the negative to the primary and the positive you will need to pulse the + on and off very quikly! the faster the pulse the better the output of energy! Connect the secondary leads from the isolation transformer to the ends of a neon or flourescent light bulb, as you pulse the positive to the primary of the transformer the bulb will light very brightly!

15,000 v Neon Transformer

Or you can pulse this homemade Tesla Coil to light a bulb!

Again, there are many ways to get free electricity from your phone lines. (also great for out door lighting.)

Updated 2004: The below method seems to be the best way so far to collect and store the 50 vdc energy. This method should be done at night or when the phone line is not in use.

To obtain 1 amp at 15 vdc we use a 10 volt capacitor to step down the 50 VDC that is coming from the lines. This capacitor is a very special capacitor, try to get the same type if they still sell them. I think it was made in Japan. I wish we had them in stock we would sell them. Photo caps are very special and might work as well.

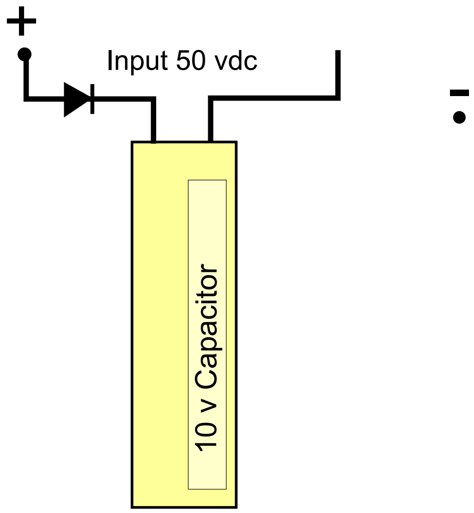

Electrolytic Capacitor
10 v x 1000 uf 105 C
Made by Elna part # 8643

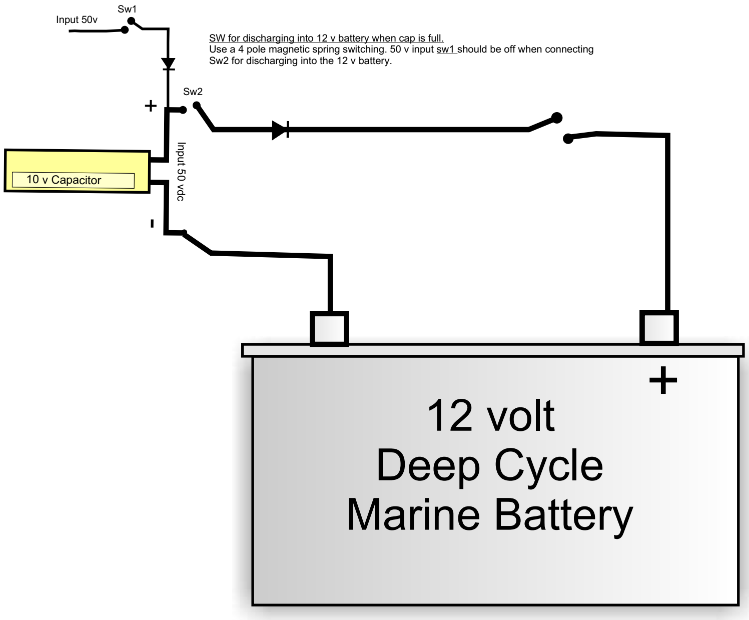

SW for discharging into 12 v battery when cap is full.
Use a 4 pole magnetic spring switching. 50 v input sw1 should be off when connecting Sw2 for discharging into the 12 v battery.

Power diode works best!

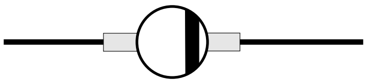

G1 G4g

Let cap charge every 4 to 6 seconds, you may want to use a spring wind up timer for switching or use transistors as dc pulser to control 4 pdt relays which all can be powered from free energy from the earth method, *(See our free energy from the earth plans #459 $29.95 plus $3.95 shipping.)* Or you can use small magnetic reed switches (about 10) placed and glued around a kitchen timer. Glue a small 1/4 diameter neodymium magnet to the rotating handle.

Copyright 2002 - 2004 Creative Science & Research

High Voltage Care

Be careful, high voltage or even low voltage with high amperage can kill you. Do not attempt to work on this project unless you are a trained professional in high and low - high amperage..

Remember to wear RUBBER GLOVES and RUBBER SOLE SHOES. Using rubber platex gloves from a dollar store or such, work good for up to 50 volts, any higher voltage you should wear a thicker rubber pair of gloves....

WARNING: discharge all capacitors before you touch and work on them or around them.. If you do not know what a electrolytic capacitor is then we suggest you buy a beginners book in electronics that you can get from from any Radio Shack or electronic store or book store.

You can also use the earth ground to gain more electrical energy into your capacitor by using one leg (the Negative) from the the telephone out put! A diode in the right direction of this leg will do wonders. it's that easy.

The following information is confidential and has copyright and patent pending: Capacitors store and discharge free energy! Radiant energy is stored in between the capacitors plates and the electrons rest on the metal plates themselves. Radiant energy is also created at discharge, but in reality is taking the energy from a very special place that we do not want to disclose at this time.

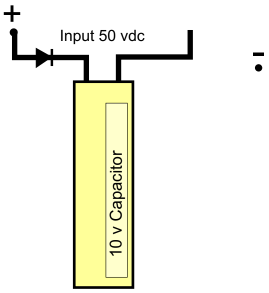

Electrolytic Capacitor
10 v x 1000 uf 105 C
Made by Elna part # 8643

Power diode works best!

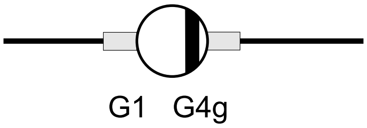

Copyright 2002 - 2004 Creative Science & Research

Part 2

We are going to show you several different methods that you can use to get free electricity from your phone line.

The amount of electricity will vary and depend on which set up you use. most of these set ups can be used at night time while your phone is not in use. using it this way will be the strongest.... by using these methods it will keep the phone 1/2 way off the hook ? (hard to explain)
One of the methods that we will explain is a bit crude....
If you should get an incoming phone call in the middle of the night or day it will charge the capacitors faster and with more amperage input, but with one disadvantage, the phone will not ring! But we have discovered a way past this, but with a slight amperage loss. By using what we invented called a ceramic permanent magnet capacitor blocker filter.. you will see the drawings on the following pages.

I'm sure there are much better ways to charge and gain more voltage and amperage from your phone line than what we have shown you here. but I am sure you will understand that our time is limited, it's a fun project, but we have been developing and researching much better technology than this! If you find a better way then please contact us so we can put it in our plans and share it to the world, we can mentionyour name as being the inventor if you like.

(Timer Switching 4 PDT relays)

The Electricity you will get can be constant when hooked up to multible capacitor banks (Differant caps than the one shown on page 2). Using one 9 volt battery capacitor timer switches, to fire each cap bank at it's proper time. or by using earth energy. (use 2 - 63 v x 4700 uf electrolytic capacitors)

We have actually lit a 100 watt 115 volt AC light Bulb with this set up. 2 caps connected in series. Charge each cap by hand, then connect together in series. But the main purpose is to use this system as a 12 volt deep cycle battery charger that will run a DC to AC 115 volt inverter to use as usable household energy to power a light bulb, an ac motor, TV, VCR etc...

Copyright 2002 - 2004 Creative Science & Research

You can Produce up to 3 amps every 30 seconds by charging a 35 volt x 4,700 uF electrolytic capacitor bank. Connecting large uf capacitors in parallel increases the amperage but slows down the charge time. You can also use them in series to step up the voltage to get 100 volts or 150 vdc if you use three fully charged caps connected in series. The Fuelless engine can be designed as a small 2 hp electric motor and made to run on .05 ma x 300 input and can run from the phone line using our Com /SW method, (See Our Plans) you would need six fully charged caps 50 to 63 v x 4700 uf caps, to run this type of High Effiecient free energy motor. the output will be greater than the input!

You can also use a DC to AC inverter (not the type to run your home)right from the phone wall outlet and then by using a step down transformer you can get 12 volts ac which you will then need to change to dc using a one way diode. The amperage will be greater and can charge a 12 volt marine battery.

Notice:

Be sure to be neat! Put all Components in a safe and convenient plastic box. Use PC board, make it professional looking don't be sloppy! You must keep it safe were Kids as well as some Adults will not get shocked or hurt.

IS IT LEGAL?

We don't Know?
When we first put it in our catalog we were confident that it was. Now we are not so sure? But, it's your phone and you are paying for it! Once it passes into your home it should be yours, as long as you hook it up right and it does not kick back up into the phone line and cause any harm to the phone companies generator lines or breaker boxes etc... that's why we use one way diodes. We have knocked out our phone lines for about an hour or so. when pulsing into the earth you sometimes get a kick back of more current and voltage than what is being collected or inputed.

Remember: it's your phone service! you pay for it every month. The electric you pay for is used to run your telephone, if you wish to use it for something else there should be no law against it.... WE HOPE!!!!????

Copyright 2002 - 2004 Creative Science & Research

OBJECTIVE;

All the methods described in these plans are ways to get cheap Free energy for your home, (#359 is not intended for your entire home).

There are millions of poor people all around the world who can't afford electricity but may have a phone, (this system is much better than using candles). It's fast and cheap to put together for emergancy low light conditions. If one can afford it, we recomend building our Fuelless engine to run a car altntr, or 10 kw home generator system and stop paying a high electric bill, our system can pay for itself. Someday I hope we will be allowed to manufacture our free energy motors then you can simply go down to your local hardware store and buy a free energy motor and generator to power your entire home, but that may not be soon. The only way the open public is going to be able to get a hold of a free energy electric motor is to build one for thereself. It's really not that hard! The cost to build our 2 hp (Fuelless) Free Energy Electric Motor is only about $195. You can use the 2 hp to run a 80 amp car alternator to charge a 12 volt marine deep cycle battery system connected to a 500 to 5000 watt inverter (Which you can buy or make, see our #579 plans $40) to run your lighting, fans, VCR, TV, Refrig, Microwave etc...

Collecting Free Energy from your phone line is also a fun science project! Your friends will love it. Most people do not even know they have 50 vdc running to there phone line, and when the phone rings it sends 3 times more amperage to your charging system. Set your phone up were it will ring itself every 2 minutes for one hour or until your batteries are fully charged. It's basically free! The phone company is not going to charge you extra for it.

Use 4 - 12 v deep cycle marine batteries, (car batteries will not work) and connect them in parralel to increase your amp hr. then connect your inverter or inverters to your fully charged battery bank and enjoy free lighting, free electricity etc... Notice: Do not let your batteries go under a 25% charge, the batteries will last years if you take care of them by not under charging or over charging them. Use a digital volt meter. If you need more amperage, again use our Fuelless Engine or our free energy from the earth.

PO BOX 557 New Albany, IN. 47151 USA

Copyright 2002 - 2004 Creative Science & Research

Use a DC Volt meter to find the negative and the positive. WARNING! Do not ground the positive to the earth, it will knock out your phone line for about 1 hr!

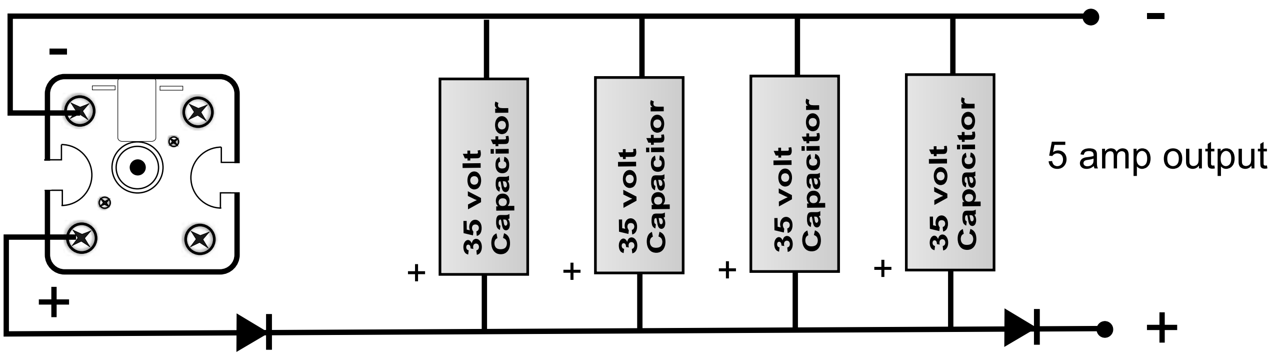
-
-
+
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
+
+
+
+
5 amp output
+

Using both the negative and the positive of the phone line seems to work best. Use 4 - 35 v x 4700 uf electrolytic capacitors. This can be used as a battery charger using the diode as shown or a com / sw switching method as seen in our Fuelless Engine plans or you can simply use a hand wind timer with a strong magnet glued or taped to the handle and using multiple reed switches. Or power a low volt, low milliamp 4 PDT relay using the earth power (see plans #459 $29.95) or our #396 homemade batteries $16.95

Many homes have 2 phone lines, you can use both lines to get 100 vdc with double the amperage output! Use 4 - 150 v x 4700 uf capacitors instead, you can then light a 100 watt bulb, or pulse the 100 volts into a 15,000 volt neon transformer to get a 14,000 vac out put to power several neon bulb tubes or flourescent bulbs.

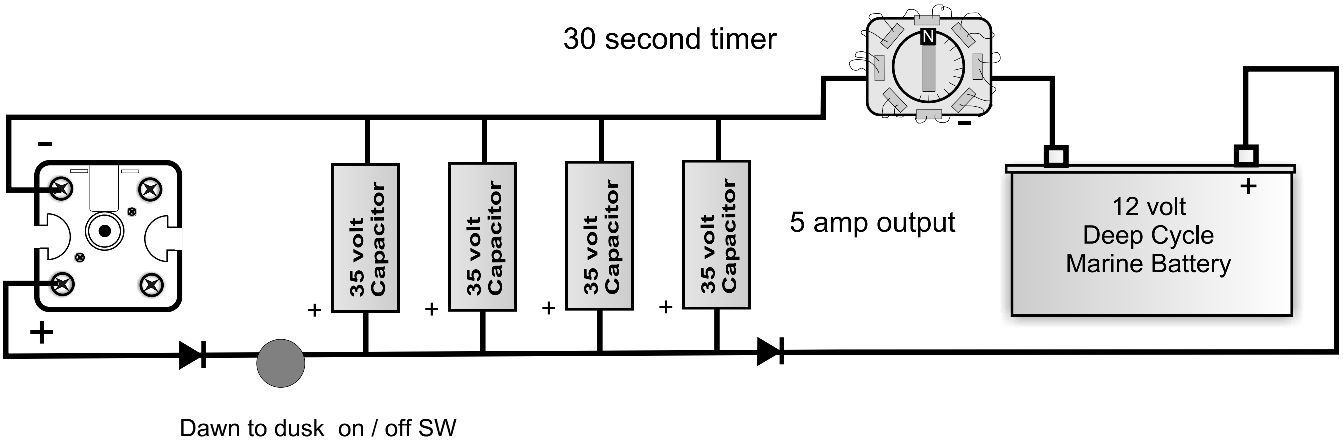
30 second timer
N
-
-
+
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
+
+
+
+
5 amp output
+
12 volt
Deep Cycle
Marine Battery
Dawn to dusk on / off SW

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

Copyright 2002 - 2004 Creative Science & Research

Grounding to Earth

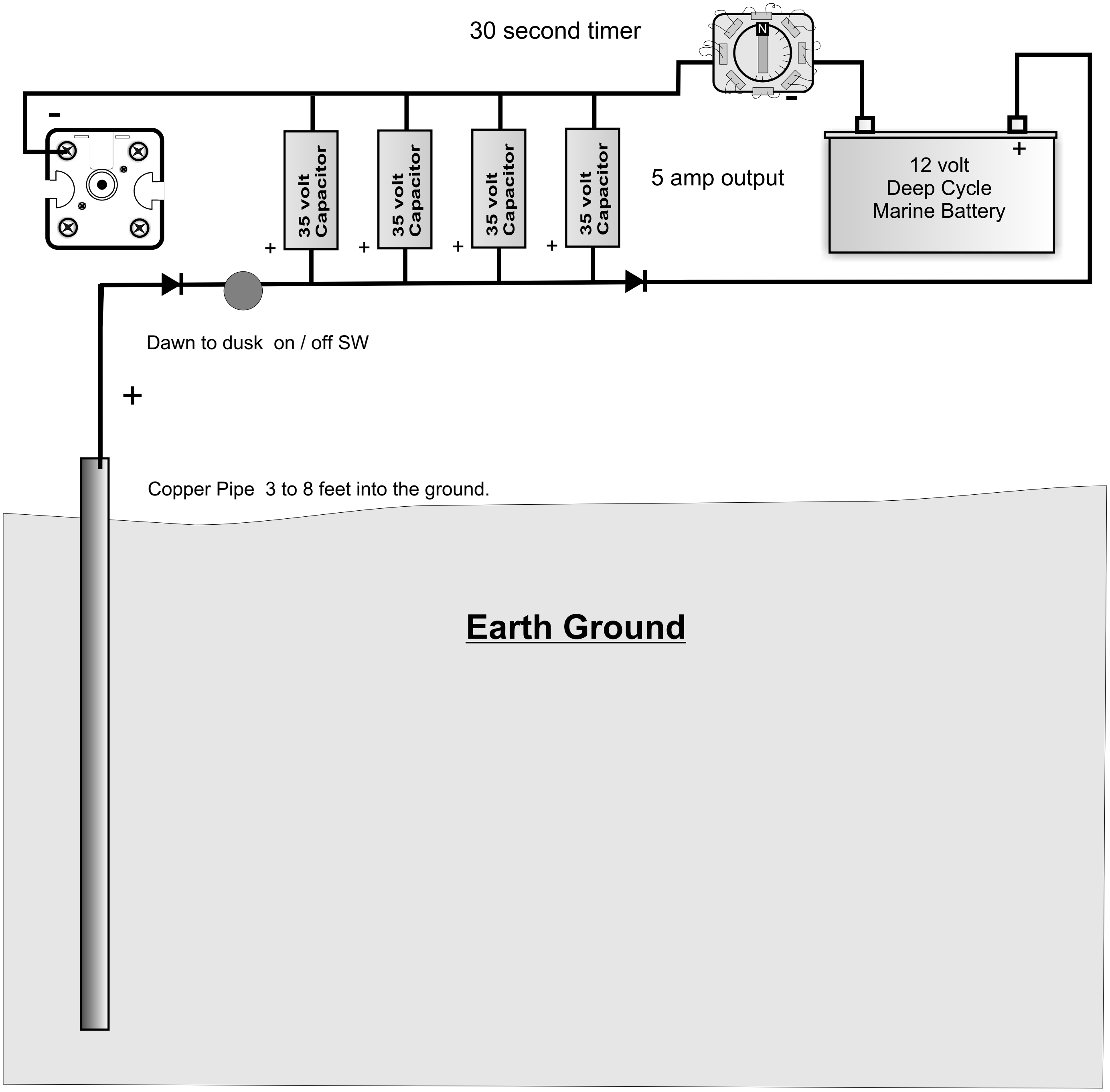
30 second timer
N
-
-
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
35 volt Capacitor
+
+
+
+
5 amp output
+
12 volt
Deep Cycle
Marine Battery
Dawn to dusk on / off SW
+
Copper Pipe 3 to 8 feet into the ground.
Earth Ground

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

Copyright 2002 - 2004 Creative Science & Research

Grounding to Earth

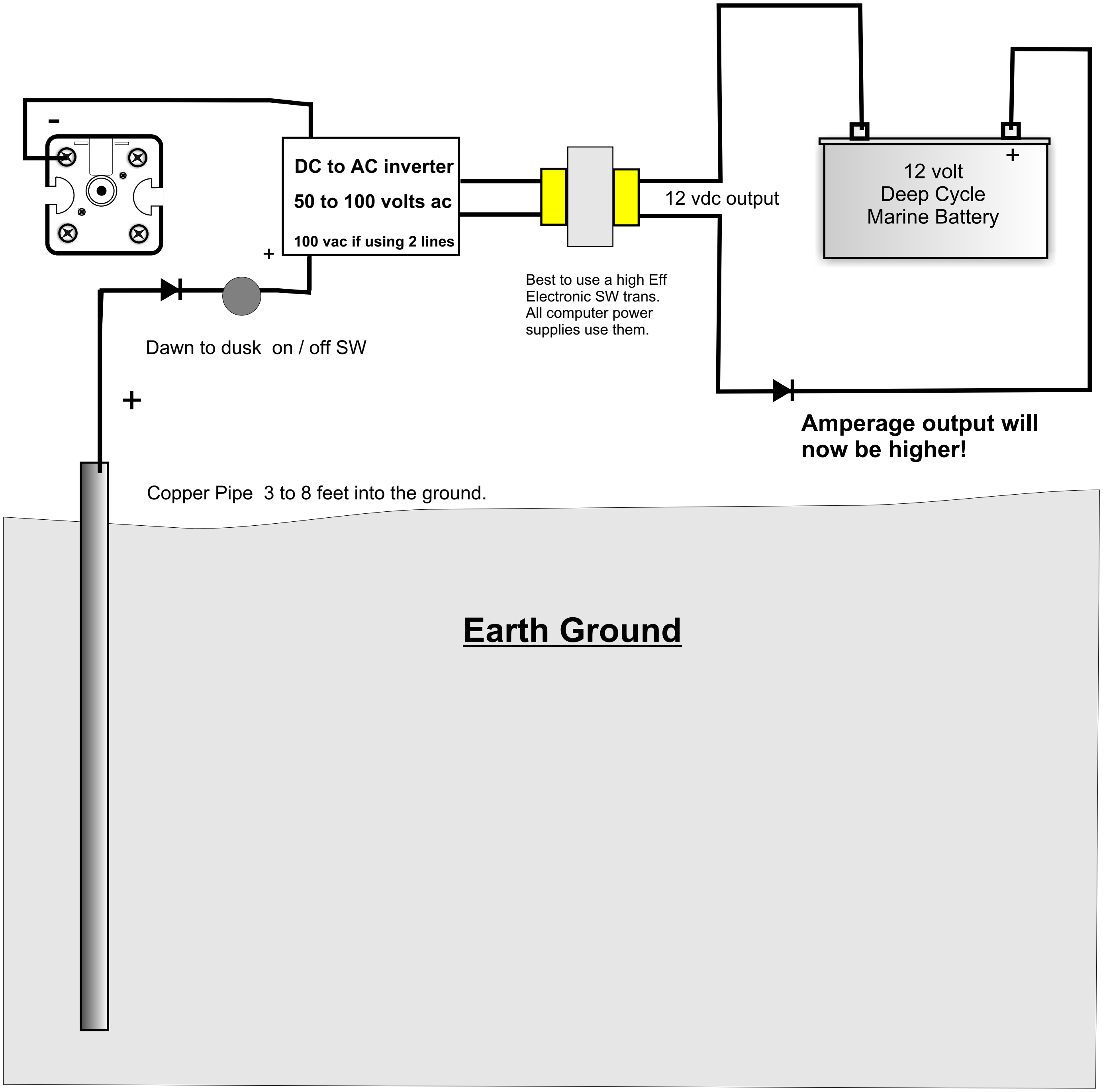
-
DC to AC inverter
50 to 100 volts ac
100 vac if using 2 lines
+
12 vdc output
12 volt
Deep Cycle
Marine Battery
+
Best to use a high Eff
Electronic SW trans.
All computer power
supplies use them.
Dawn to dusk on / off SW
+
Amperage output will
now be higher!
Copper Pipe 3 to 8 feet into the ground.
Earth Ground

If you wish to build your own hand spring timer and relays you can buy material at **www.smallparts.com**

Creative SCIENCE & Research

359

PO BOX 557 New Albany, IN. 47151 USA

Copyright 2002 - 2004 Creative Science & Research

A Simple but effective Inverter

A simple but very effective inverter that you can build if you know how to handle a soldering gun. Parts can be purchased at any radio shack store or electronic supply house. Search the web. When I say search the web, you would not believe how many people do not know what I am talking about so I will explain it. You can find just about anything you need on the web. If you have AOL.com then that is a search engine as well as Google.com, Yahoo.com etc.. They are all powerful search engines, you simply type in a word in the search the web white address bar, for example: Looking for capacitors, Type in capacitors or electronic parts and supplies. You can also buy beginners books in electronics from any Radio shack or on line store to quickly help you learn how to put this inverter together, some suppliers sell HV inverter kits.

The 50 vdc must be changed to ac voltage current or it will not work (step up voltage) into a step up transformer or diode / capacitor step up multiplier circuit.

High Voltage Power supply / DC to AC Inverter

Rated for 1.5 to + 9vdc if using 50 vdc or 100 find cap and transistors rated for that voltage or more, or you will burnout trans

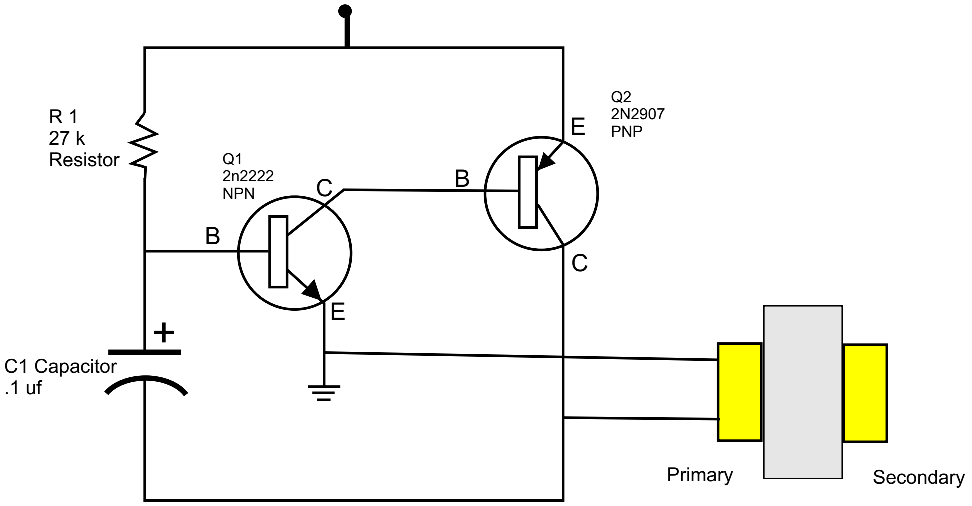

T1 = 6.3 v x 120v transformer This circuit will produce about 220 volts. Use a 1m resisitor for higher volatges.

We found this cicuit in an old electronics book, it is for low voltages but you can replace the parts with higher rated parts so they will work on higher amperages or higher voltages.

Catalog #27

Creative SCIENCE & Research

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151   USA

Information Big Brother & the Oil Companies do not want you to know!

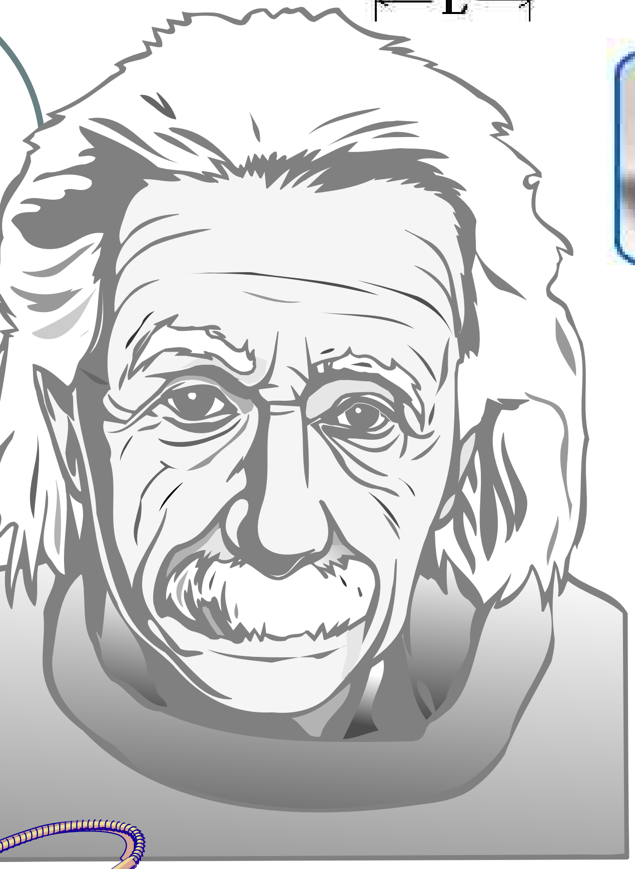

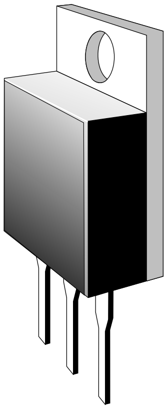

2004

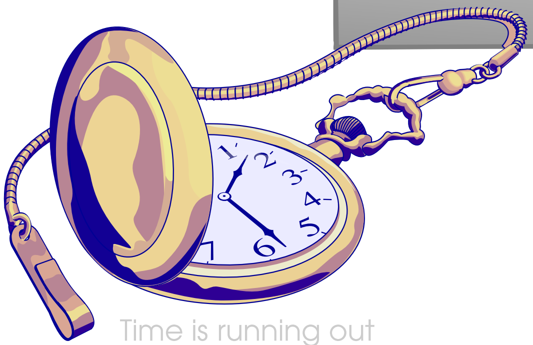

Time is running out

Search for:

FREE ENERGY ▼ Find

www.fuellesspower.com

Copyright 2002 - 2004 Creative Science & Research

INTRODUCTION

CREATIVE SCIENCE AND RESEARCH "Since 1993"

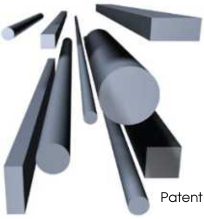

Patent Pending 2003

Free Energy using Carbon rod,
as well as special Molded Soft Iron Ductal Cores

The reason we made this **catalog** was to raise money for research and development and to also share to the World our findings and discoveries on **The Fuelless Engine.**
All of our research and findings on the topic of free energy have guided us to many other ideas and devices along the way which are included in this catalog, but **none can compare** to the **Fuelless Engine!** If the whole world uses this **free energy engine** we will all be able to free ourselves from the Gas, Oil, and Electric companies that now hold us prisoner. By the year 2020 every home should have their own power plant, and we will do everything in our power to make this happen.

This engine will run forever with a once a year maintenance check, and will provide **FREE ENERGY** for the rest of your life.

If you are a skeptic and find Free Energy and Perpetual motion a little hard to believe then consider this, The earth is FREE ENERGY in a perpetual motion state. Hydrogen and oxygen atoms are FREE ENERGY constantly being reused over and over. A coil of wire induced by a DC current and then turned off will produce **Free energy from a collapsing magnetic field.** HV Capacitor's play a very important roll as well..... Solar Cells are free energy devices but they have there limitations, they can only operate when sunlight is present, but our devices are far greater than solar cell technology and can operate 24 hrs a day 7 days a week.

We take what we do very seriously! We would not sell anything that we did not first test ourselves. I personally believe 100% in what we sell. **Our plans are Guaranteed to work or your money back!**

Free Energy is everywhere, don't be left in the dark.

Thank you
David Waggoner
David Waggoner
Owner / Research and Development

Creative Science & Research
PO BOX 557
New Albany, IN. 47151 USA

sales@fuellesspower.com

www.fuellesspower.com

www.fuelless.com

www.fuellessengine.com

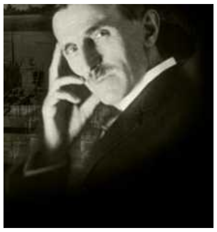

Nikola Tesla

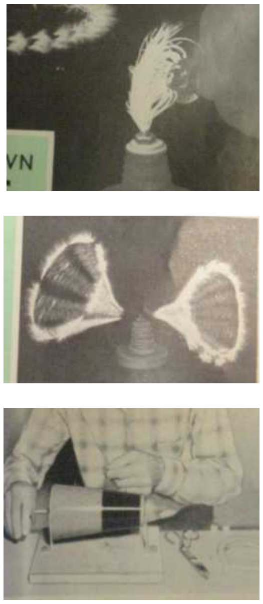

Tesla Coil

The Fuelless Engine

Copyright 2002 - 2004 Creative Science & Research

Free Energy ● ***Free Electricity for your home or car!***

FREE POWER! For your home, garage or car!

The US Government nor the Oil Companies will not let us or anyone else manufacture these powerful engines right now, (They say it could destroy our economy?)
But if we can slowly introduce these devices to the world I think we can convert every home and automobile to run on free energy by the year 2020, I believe it could be done safely and would not hurt our economy at all. That's where you come in, We need backyard researchers just like yourself, to build our device(s), use them, and keep them confidential! We have made it as easy for you to build using our easy to follow plans.

2′ x 2′ x 1′ **350 Hp**

Many years of research went into designing this engine. There is no other engine like it. You can scale these plans up or down to build any size horsepower that you want..

THIS ENGINE CREATES IT'S OWN POWER, *It does not run on any type of gasoline, oil or any other type of combustible fuel. It has a high free energy output! Guys, if you do not get these plans now you will be kicking yourself in the pants tomorrow! This is the engine that every free energy researcher is talking about!*

EASY TO BUILD STEP BY STEP PLANS / BLUE PRINTS

MANY MAGAZINE ARTICLES HAVE BEEN WRITTEN ON OUR ENGINE! You may have read about us in the 1996 July Issue of Exotic Research magazine. Popular Science has done many articles on free energy in the early 1940's and 50's, I'm sure before BIG BROTHER asked them to stop!

Plans Only $16.95

1 - 50 hp Order # 362

GUARANTEED TO WORK OR YOUR MONEY BACK!

Video Only $29.95

Order # 361

50 to 350 hp plans only $40.00 Order#362rc

Creative Science & Research P.O. BOX 557 New Albany, IN. 47151
Once we teach you the secrete in building this engine You can then design and build this engine to any HP you desire.

Copyright 2002 - 2004 Creative Science & Research

All Our Plans are Now on CD!

FREE ENERGY MOTORS & GENERATORS

*** CONFIDENTIAL INFORMATION!**
For Selected Customers Only!

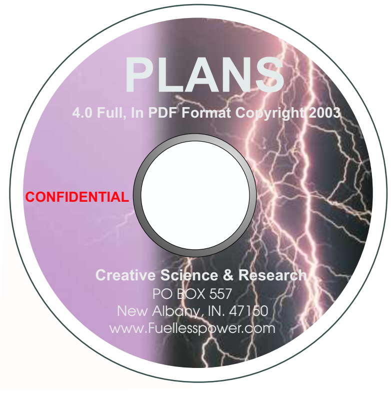
PLANS
4.0 Full, In PDF Format Copyright 2003
CONFIDENTIAL
Creative Science & Research
PO BOX 557
New Albany, IN. 47150
www.Fuellesspower.com

VISA
MasterCard
DISCOVER
AMERICAN EXPRESS
PLANS
PRIORITY MAIL

As requested by many of our customers!

* Information The Oil Companies and Big Brother Do Not Want You To Know!

We are now offering our entire web site and catalog of Free Energy Plans for only **$300**

You save over $200 (a $500.00 value)
Order # CD Package # 1

*** LOADED with information! Packed full of color photo's, More information! etc....**
Plans are sent to you on a CD Computer Disk, which you can then view or print from your computer.

PDF format for Adobe Reader Computer Program. If you do not have Adobe Reader 4.0
It is FREE for downloading at www.adobe.com

* Please add $3.95 for US Priority Shipping which includes online tracking

*** Guaranteed To Work Or Your Money Back!**

Other CD Packages Available: See our web site order form, Bottom page
www.FuellessPower.com

Please Send a Check or Money Order To:

Creative Science & Research
PO BOX 557 New Albany, IN. 47151 USA

www.fuellesspower.com or www.fuelless.com

NOTICE: Most of our plans can are now available by E-mail Download

612:195

THE GOLDEN BOOK OF CHEMISTRY EXPERIMENTS

HOW TO SET UP A HOME LABORATORY—OVER 200 SIMPLE EXPERIMENTS

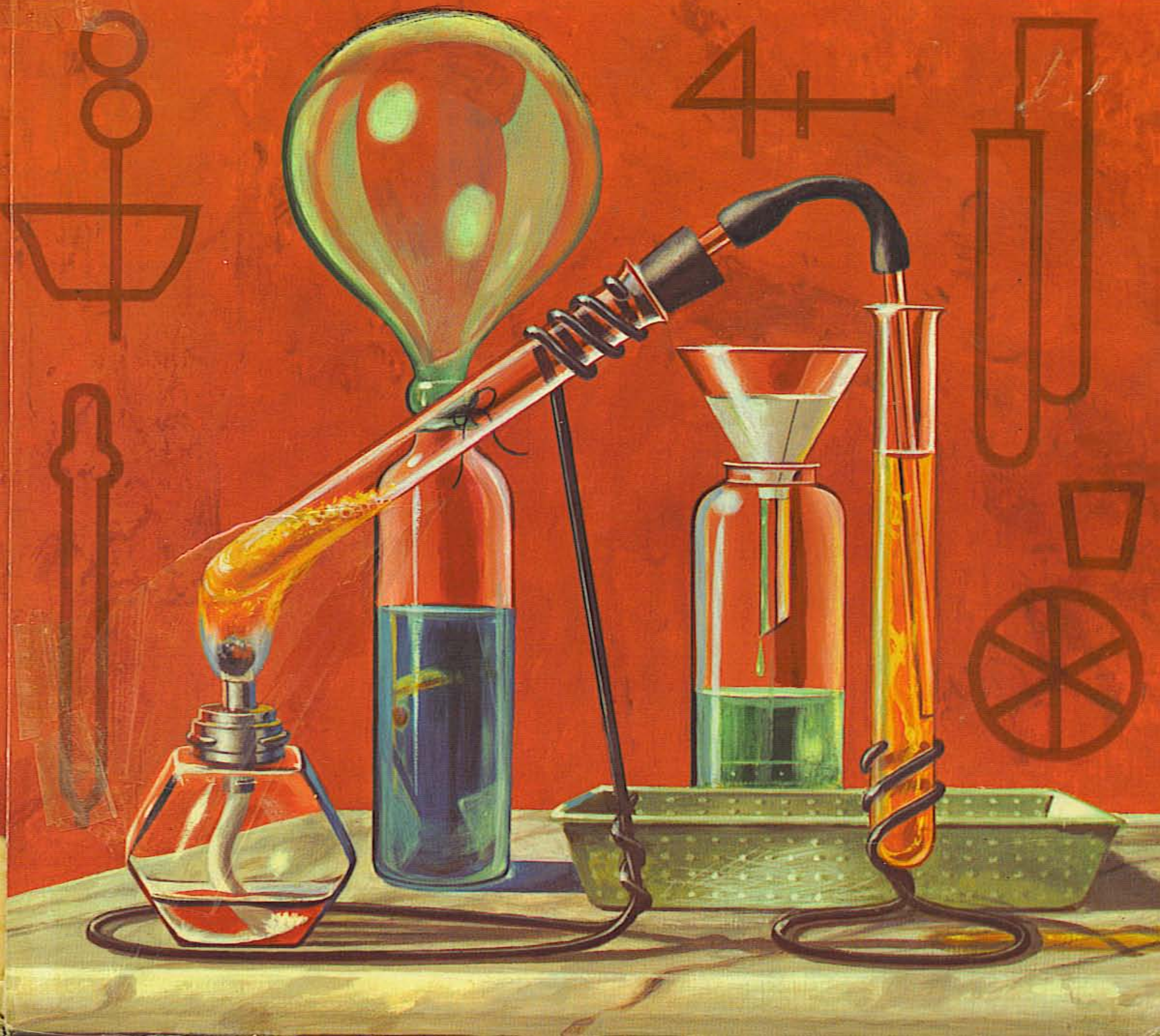

THE GOLDEN BOOK OF Chemistry Experiments

How to Set Up a Home Laboratory— Over 200 Simple Experiments

BY ROBERT BRENT

ILLUSTRATED BY HARRY LAZARUS

GOLDEN PRESS NEW YORK

© Copyright 1960 by Golden Press, Inc. All rights reserved. Printed in U.S.A. Published by Golden Press, Inc., Rockefeller Center, New York 20, N. Y.

Words Used by Chemists

Acid: a hydrogen-containing compound that releases hydrogen ions in solution.

Alloy: a material made up by combining two or more metals.

Analysis: breaking down a compound into two or more substances.

Anhydrous: free from water.

Atom: the smallest unit of an element that can enter into the making of a chemical compound.

Atomic weight: the weight of an atom compared with the weight of an oxygen atom set at 16.

Base: a compound containing the hydroxide group (OH).

Catalyst: a substance that helps in a chemical reaction without itself being changed.

Chemical change: a change of a substance into another substance having different properties.

Chemistry: a branch of science dealing with the compositions of substances and the changes that can be made in them.

Combustion: burning; a chemical change that produces heat and light.

Compound: a substance consisting of two or more different kinds of atoms in definite proportions by weight.

Crystal: a solid in which atoms or molecules are arranged in a definite pattern.

Density: the weight of a liquid or a solid in grams per cm^3 or milliliter.

Distillate: a liquid that has been turned into vapor and again cooled into a liquid.

Distillation: the process of producing a distillate.

Ductile: capable of being drawn out into a wire.

Electrolysis: breaking down a substance by passing an electric current through it.

Electrolyte: a substance that, when in a solution or when melted, will conduct an electric current.

Element: a substance that contains only one kind of atoms.

Equation: a complete description of a chemical reaction by the use of symbols, formulas, and signs.

Evaporation: the changing of a substance into vapor; also the process of removing water by heating.

Filtrate: a liquid obtained by filtration.

Filtration: the process of straining a liquid from a solid through porous material, usually filter paper.

Formula: a group of symbols and numbers giving the composition of a compound.

Hydrate: a compound containing loosely bound water of hydration (water of crystallization) that can be driven off by heating.

Hydroxide: a compound that contains the hydroxyl (OH) radical.

Ion: an electrically charged atom or group of atoms (radical).

Malleable: capable of being hammered or rolled into a thin sheet.

Matter: anything that takes up space and has weight.

Metal: an element that is a good conductor of electricity, has luster, and whose oxide forms a base with water.

Metalloid: an element that has properties of both metals and nonmetals.

Mixture: a mingling of substances not combined chemically.

Molecular weight: the sum of the atomic weights of the atoms that make up a molecule of a compound.

Molecule: the smallest unit of a compound that can exist in the free state.

Neutralization: the reaction of an acid and a base to give a salt and water.

Nonmetal: an element that is a poor conductor of electricity, does not have luster, and whose oxide forms an acid when combined with water.

Organic chemistry: the chemistry of the carbon compounds.

Oxidation: the process by which a substance combines with oxygen.

Precipitate: an insoluble solid formed in a solution by chemical reaction.

Radical: a group of atoms that behave chemically as a single atom.

Reaction: a chemical change.

Reduction: removal of oxygen; the opposite of oxidation.

Salt: compound (other than water) formed by the reaction of an acid and a base.

Saturated solution: a solution that contains the maximum amount of solute under the conditions.

Solubility: the number of grams of a solute needed to make a saturated solution in 100 grams of solvent.

Solute: the substance dissolved in a solvent.

Solution: a non-settling mixture of a solute in a solvent.

Solvent: a liquid in which a solute is dissolved.

Sublimation: a process by which a solid is turned into vapor and again cooled into a solid without passing through a liquid stage.

Subscript: a small numeral indicating the number of atoms of a certain element in the formula of a compound.

Substance: any specific kind of matter whether element, compound, or mixture.

Symbol: a letter or two letters representing one atom of an element.

Synthesis: the making up of a compound from simpler compounds or from elements; the opposite of analysis.

Valence: the number of hydrogen atoms which one atom of an element can displace or with which it can unite.

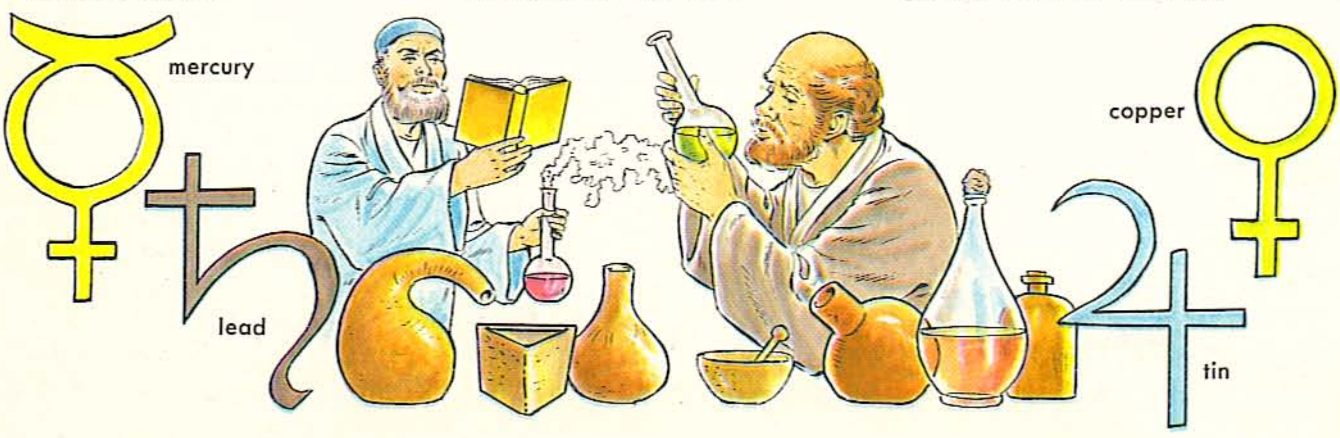

Contents

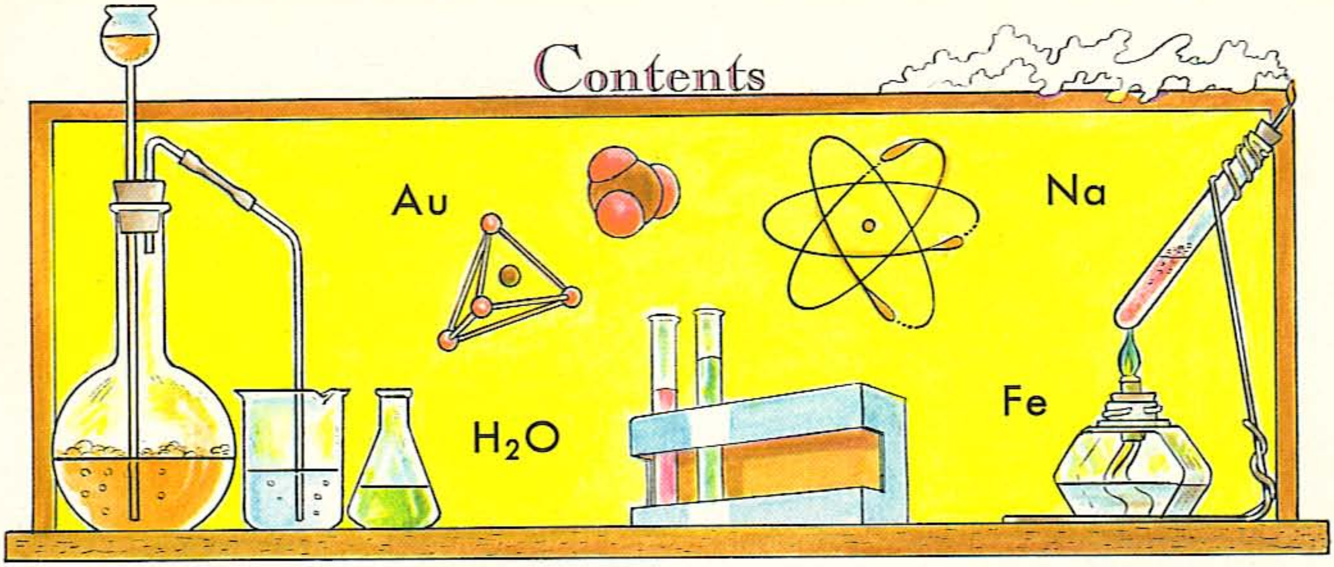

WHAT CHEMISTRY IS

Words Used by Chemists 2
The Importance of Chemistry 4
Chemists of the Past 6

YOUR HOME LABORATORY

Equipment for Chemistry 9
Setting up Your Home Laboratory 10
Making Apparatus for Experiments 12
Scientific Measurements 14
Correct Laboratory Techniques 16

THE SCIENTIFIC APPROACH

Mr. Faraday's Candle 18
You — Scientist! 20
Elements, Compounds, and Mixtures 22

WATER AND GASES

Water — Our Most Important Compound 24
Oxygen — The Breath of Life 26
Hydrogen — Lightest of All 28
Carbon Dioxide 30
Nitrogen and Its Compounds 32
Chlorine — Friend and Foe 34

CHEMICAL FORMULAS

Chemical Shorthand 36
The Periodic Table of the Elements 38

ACIDS, BASES, AND SALTS

The Mysteries of Solutions 40
Working with Acids 42
Working with Bases 43
Salts — Chemicals of Many Uses 46

NONMETALS

Iodine — Violet or Brown? 48
Sulfur and Its Compounds 50
Silicon — The Element You Step on 54
Boron — Future Rocket-Power Element? 56

METALS

Sodium and Potassium 58
Calcium — for Building 60
Let's Compare Two Metals 62
Aluminum — in Abundance 64
Manganese — Metal of Many Colors 66
We Live in an Age of Iron 68
Copper — Yesterday, Today 70
Silver — One of the "Noble" Metals 72

MORE ABOUT FORMULAS

Valences and Formulas 74

ORGANIC CHEMISTRY

Carbon — Element of a Million Compounds. 76
The Chemistry of Carbon Compounds 78
The Formulas of Carbon Compounds 80
A Lot of Hydrocarbons 82
Carbohydrates — Sweet and Bland 84
Many Kinds of Alcohols 88
Carboxylic Acids 90
Fats and Oils for Energy 92
Soap and Soap Making 94
Proteins—The Body-Building Foods 96
Colloidal Dispersions 100
Natural and Artificial Fibers 102
Plastics — A Modern Giant 104

CHEMICAL MATHEMATICS

Working out Chemical Equations 106

THE FUTURE OF CHEMISTRY

What's Ahead in Chemistry? 109

Where to Get Chemicals and Equipment 110
Common Chemicals and Their Formulas 111
Index 112

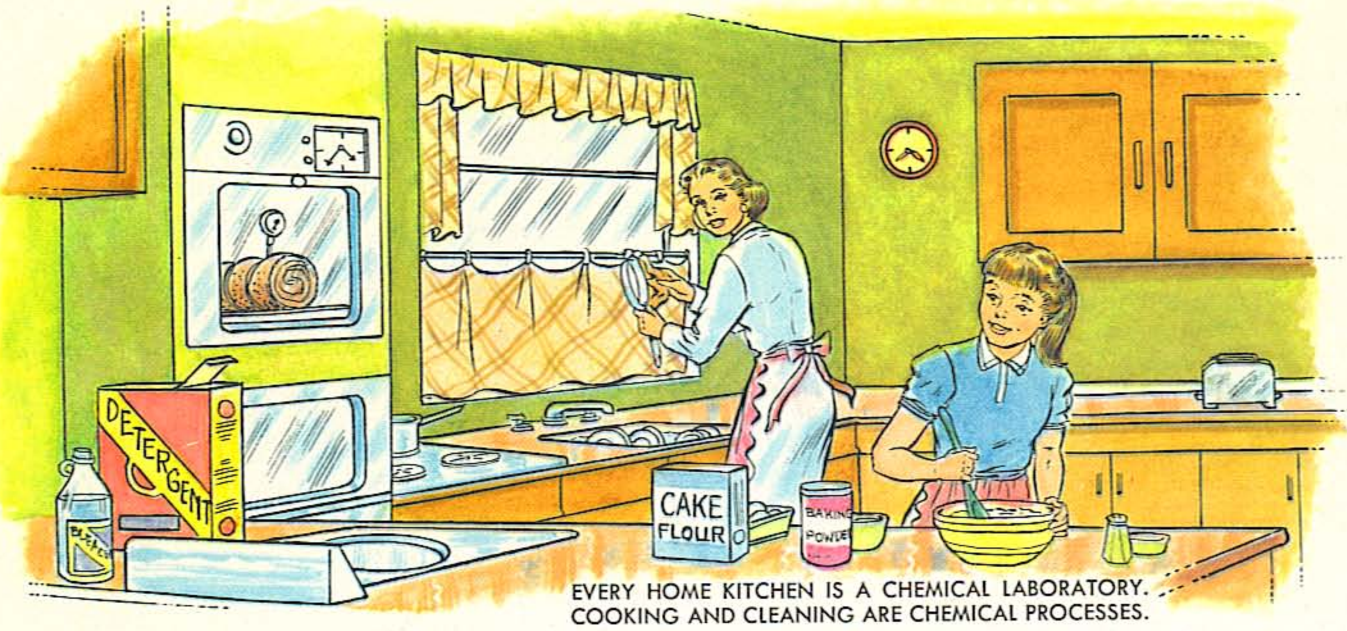

EVERY HOME KITCHEN IS A CHEMICAL LABORATORY.
COOKING AND CLEANING ARE CHEMICAL PROCESSES.

The Importance of Chemistry

THERE IS HARDLY a boy or a girl alive who is not keenly interested in finding out about things. And that's exactly what chemistry is: FINDING OUT ABOUT THINGS — finding out what things are made of and what changes they undergo.

What things? Any thing! Every thing!

Take a look around you. All the things you see — and lots of things you can't see — have to do with the science of chemistry.

Let's start with yourself. The air you breathe is a mixture of chemical substances and the process of breathing is a chemical reaction. The foods you eat are all chemical products and the ways in which your body turns them into muscles and bones and nerves and brain cells are some of the greatest of all chemical mysteries.

The clothes you wear, the books you read, the medicine you take, the house in which you live — all these are products of chemistry. So is the family car — the metal in it, the rubber on which it rolls, the gas that moves it.

Nature itself is a tremendous chemical laboratory. Everything in nature is forever passing through chemical changes. Here on earth, plants and animals

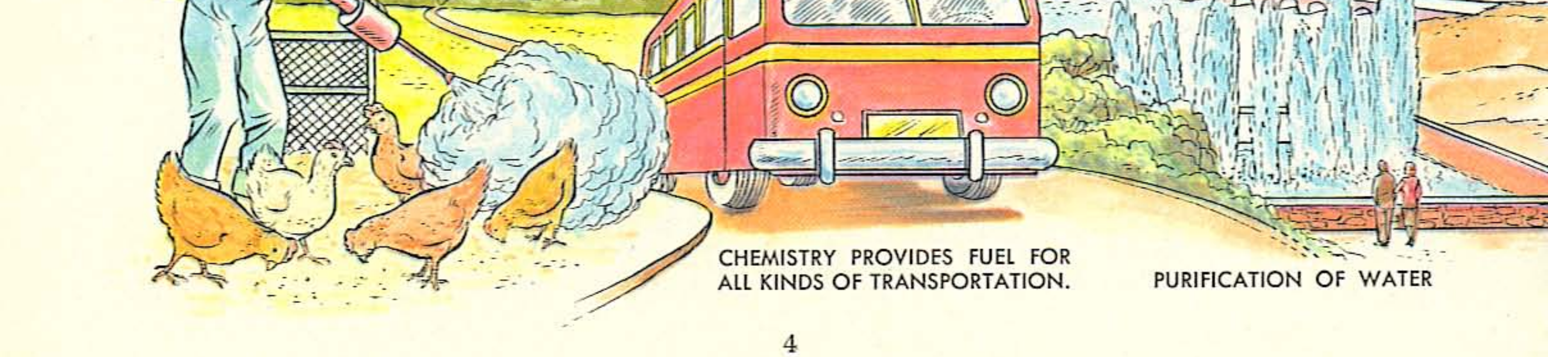

INSECT SPRAYS MEAN HEALTHIER LIVESTOCK.

CHEMISTRY PROVIDES FUEL FOR ALL KINDS OF TRANSPORTATION.

PURIFICATION OF WATER

grow, die, and decay; rocks crack and crumble under the influence of air and water. In the universe, new stars are formed, others fade. The sun that gives us heat and light and energy is a flaming furnace of chemical processes that will eventually burn itself out, billions of years from now.

Chemistry is one of the most important of all sciences for human welfare.

Chemistry means the difference between poverty and starvation and the abundant life. The proper use of chemistry makes it possible for farmers to feed the world's ever-increasing population, for engineers to develop new means of transportation and communication that will bring the peoples of the world closer together, for doctors to cure the diseases of mankind, for manufacturers to produce the thousands of items that are necessary for better and richer living.

And this is only the beginning.

Within recent years, scientists have succeeded in penetrating into the innermost secrets of chemical substances and have begun to make use of the tremendous force that lies hidden in them. This atomic power opens up amazing possibilities for the future.

You will live in a world in which chemistry will become ever more important. To understand that world it is necessary to understand the truths and laws on which modern chemistry is based and to learn how chemists of the past unraveled them.

This book will help you get this insight — not alone by your reading it, but also by your conscientiously doing the experiments described and learning what each of them has to tell you.

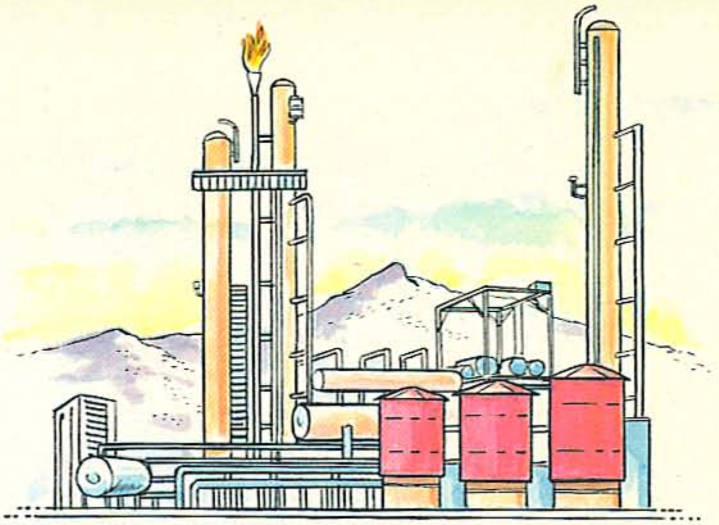

OIL IS THE BASIS FOR COUNTLESS CHEMICAL PRODUCTS.

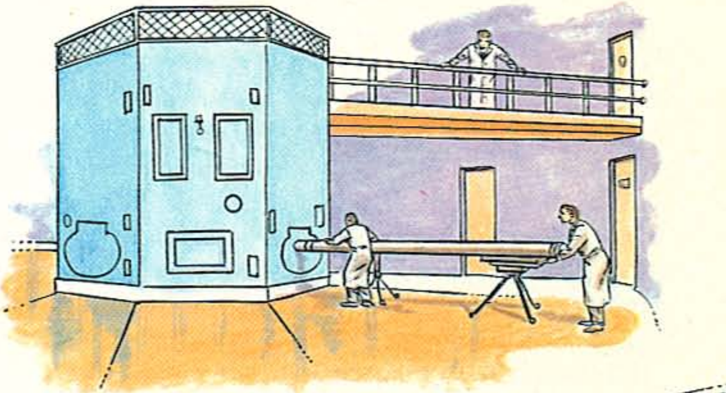

ATOMIC ENERGY FOR THE FUTURE

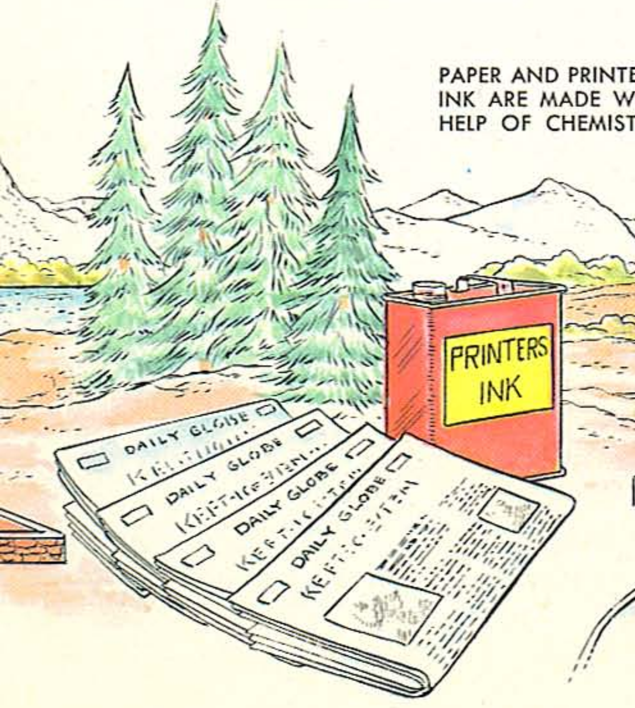

PAPER AND PRINTERS' INK ARE MADE WITH HELP OF CHEMISTRY.

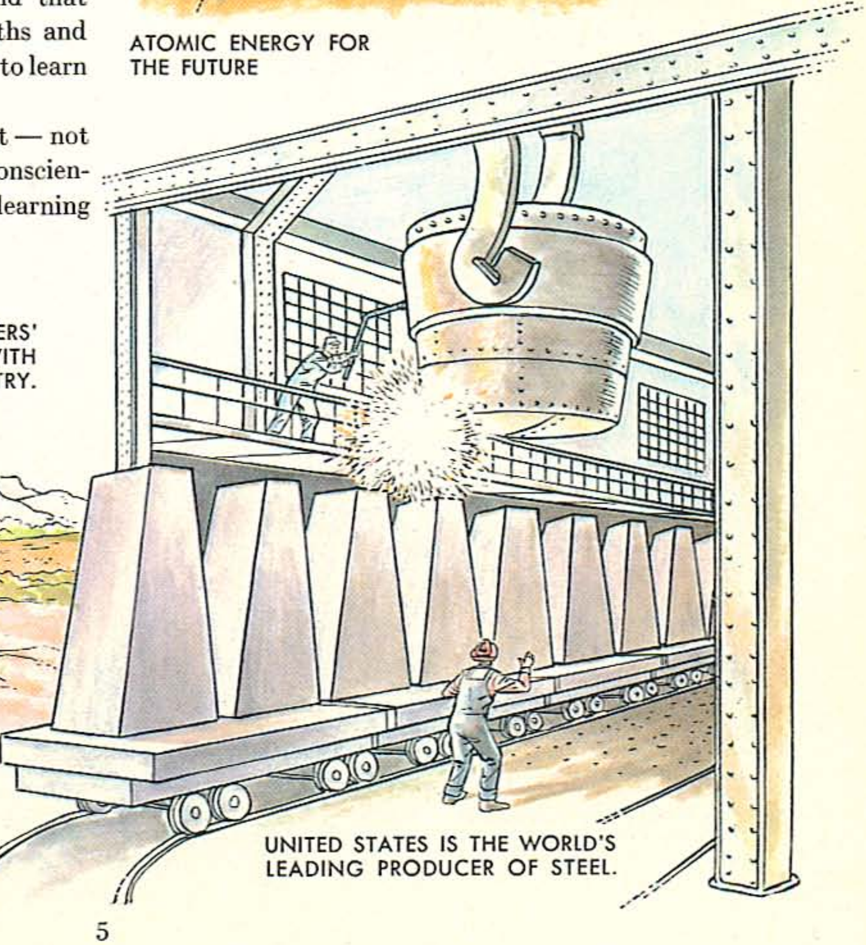

UNITED STATES IS THE WORLD'S LEADING PRODUCER OF STEEL.

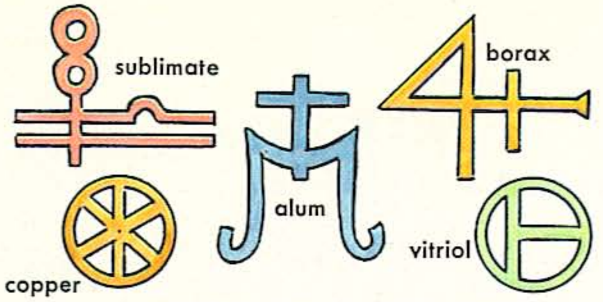

SYMBOLS USED BY ALCHEMISTS

Chemists of the Past

Many thousand years ago, an early ancestor of yours pushed a stick into the hot lava flowing from an erupting volcano. The stick burst into fire. He held it up as a torch. It gave off light and heat and finally turned into ashes.

This ancient man might be considered the world's first chemist. He had actually taken a substance called wood and had, by a chemical process called combustion or burning, turned it into something else.

The discovery of the use of fire was the first great step leading toward modern chemistry. Fire made it possible to turn raw foodstuffs into edible meals, to bake shaped clay into pottery, to make glass, to drive metals out of their ores.

For thousands of years people were chiefly interested in the results of what they did — they didn't care about what happened or why it happened. It was only about 2,500 years ago that philosophers began to wonder about what things were made of and what happened when a thing changed into something else.

Around 400 B. C., in Greece, a thinker by the name of Empedocles came up with an idea that seemed to make sense. He explained that everything in the world was made from just four things which he called "elements": fire, water, air, and earth. Think of that burning stick mentioned above. It gave off fire — so, obviously, the stick had to contain fire. It sizzled — which meant there was water in it. It smoked — and smoke would be some kind of air. It left ashes — and ashes are earth, as certainly everyone should know.

Everyone — except another Greek, Democritus, born around the time when Empedocles died. He had a different notion — that all matter was made up of tiny particles which he called *atomos* — something that cannot be cut further.

But Democritus didn't get very far with his idea. The greatest Greek philosopher of the day, Aristotle, held out for the four elements. And because of his great reputation this false idea governed the thinking of scientists for two thousand years — because no one dared suggest that he knew better than the great Aristotle!

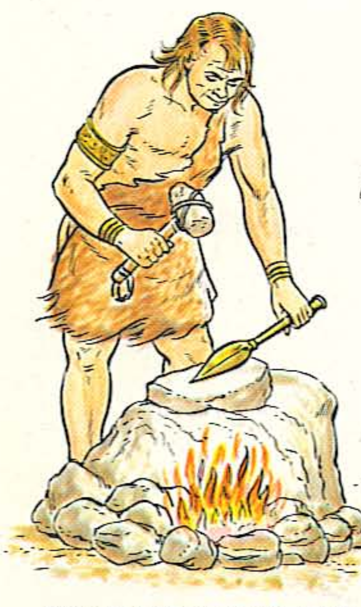

BRONZE-AGE MAN WAS ONE OF THE EARLIEST CHEMISTS.

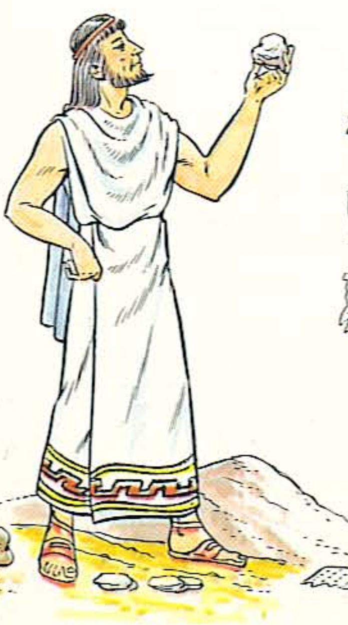

DEMOCRITUS INSISTED THAT MATTER CONSISTS OF ATOMS.

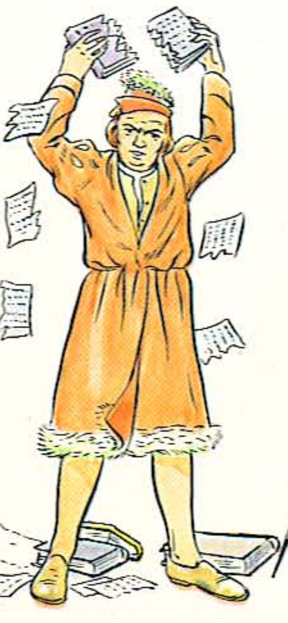

PARACELSUS TOLD HIS PUPILS TO USE EXPERIMENTS.

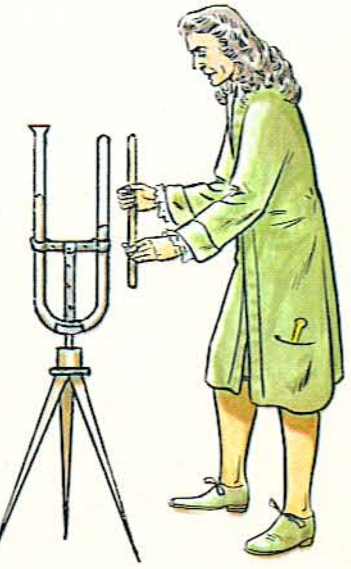

BOYLE INVESTIGATED GASES AND BROKE OLD TRADITIONS.

In the meantime, scientists of Arabia began work in a subject they called alchemy — from Arabic *al*, the, and *kimia*, pouring together. They mixed things and boiled and distilled and extracted in the hope, some day, of finding a way of making GOLD! They discovered a great number of things not previously known, developed many sound laboratory methods, and gave the science of chemistry its name — but they never created the slightest speck of gold. Neither did a great number of European alchemists.

For hundreds of years chemistry made little headway. Then, in 1525, a Swiss doctor and scientist spoke up. He had the imposing name of Theophrastus Bombastus Paracelsus von Hohenheim. He challenged his students to tear up their books with the old theories that had been developed through reasoning only and to find out for themselves *through experiments* whether a scientific theory was right or wrong. But only a few people paid attention to him.

More than a hundred years passed before an Englishman, Robert Boyle, in 1661, succeeded in killing off the old idea of the four elements. He did it by establishing that there are many elements — substances that cannot be formed by other substances and cannot be broken into other substances.

Another hundred years went by. Then, at the time of the American Revolution, the day finally dawned for modern chemistry.

A Swede, Karl Scheele, and an Englishman, Joseph Priestley, discovered oxygen, and a Frenchman, Antoine Laurent Lavoisier, explained the true nature of burning and made up the first scientific listing of all known elements — twenty-eight at the time.

Within a few years, more elements were found. With the help of electricity, an English chemist, Humphry Davy, in a single year brought to light six new metals — among them sodium, potassium, calcium, and magnesium.

Twenty years later, in 1828, another important break-through occurred. A German chemist, Friedrich Wohler, working in his laboratory produced a chemical, urea, that had never before been made outside the body of a living animal.

More and more things were happening. New elements were discovered, new chemicals created. The advances in chemistry greatly influenced industry, agriculture and medicine.

And then, in 1898, the Polish-born Marie Curie and her French husband, Pierre, discovered the "miracle element," radium. This opened up a whole new age in chemistry.

Within the last fifty years, chemistry has moved forward with giant steps. But not a single one of these steps would have been possible without the dedicated work of the chemists of the past who laid the foundation on which modern chemistry rests.

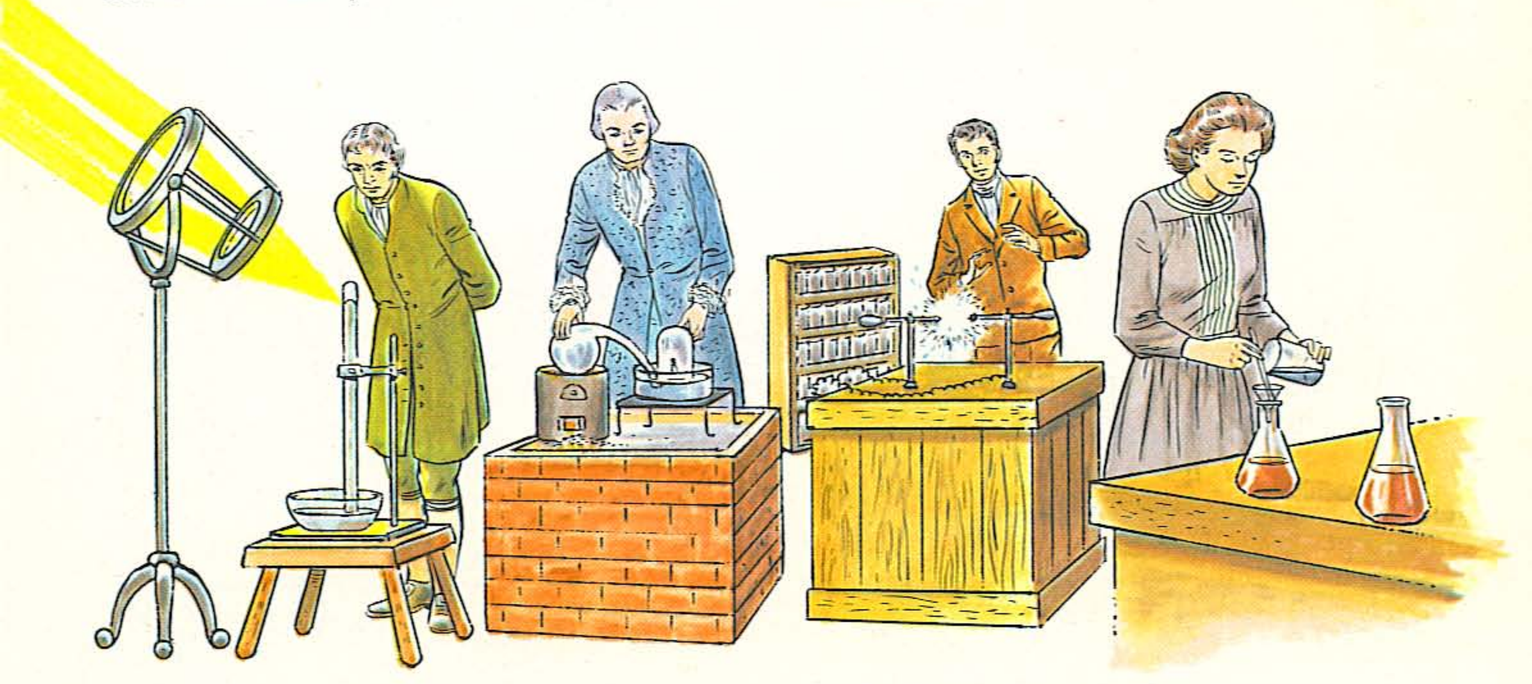

PRIESTLEY USED HEAT OF SUN TO PRODUCE OXYGEN.

LAVOISIER GAVE THE RIGHT EXPLANATION OF BURNING.

DAVY BROUGHT ELECTRICITY INTO CHEMICAL RESEARCH.

MARIE CURIE AND HER HUSBAND DISCOVERED RADIUM.

REGULAR LABORATORY WARE

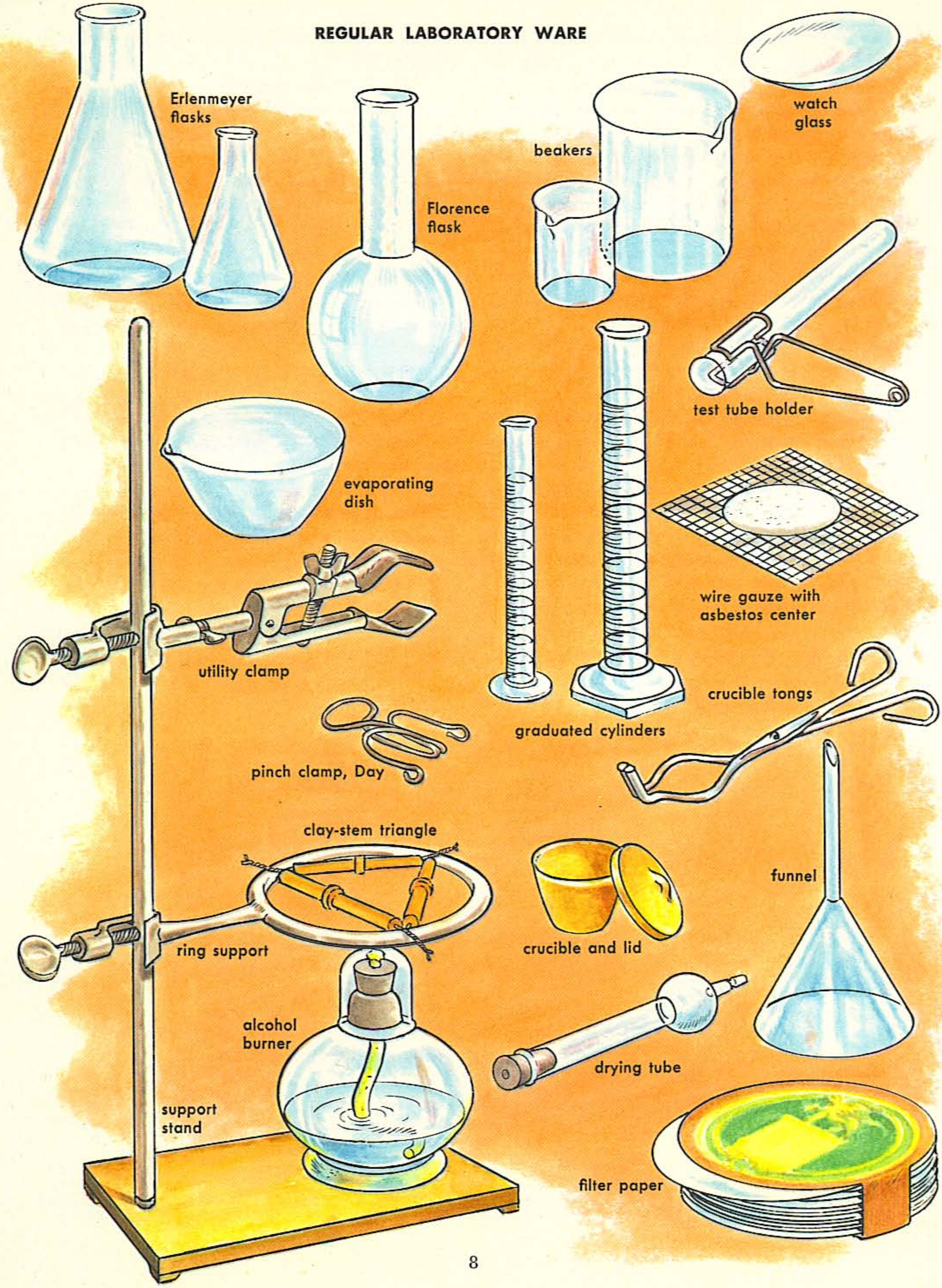
Erlenmeyer flasks
Florence flask
beakers
watch glass
test tube holder
evaporating dish
wire gauze with asbestos center
utility clamp
graduated cylinders
crucible tongs
pinch clamp, Day
clay-stem triangle
funnel
crucible and lid
ring support
alcohol burner
drying tube
support stand
filter paper

Equipment for Chemistry

Some of the greatest discoveries in chemistry were made by scientists who had no special equipment but simply used whatever was at hand.

In your home lab experiments it will pay you to follow the example of these early chemists. Put your imagination to work. Use whatever suitable equipment you can find around the house (as suggested in column to the right) and buy only what is absolutely necessary (as shown below). Some items may be purchased in a local drugstore or scientific supply shop. If not, you can buy them from one of the suppliers listed on page 110.

Later on — if you really get excited about chemistry — you may want to use your pocket money for some of the lab equipment shown on page 8.

LABORATORY WARE FOR HOME LAB

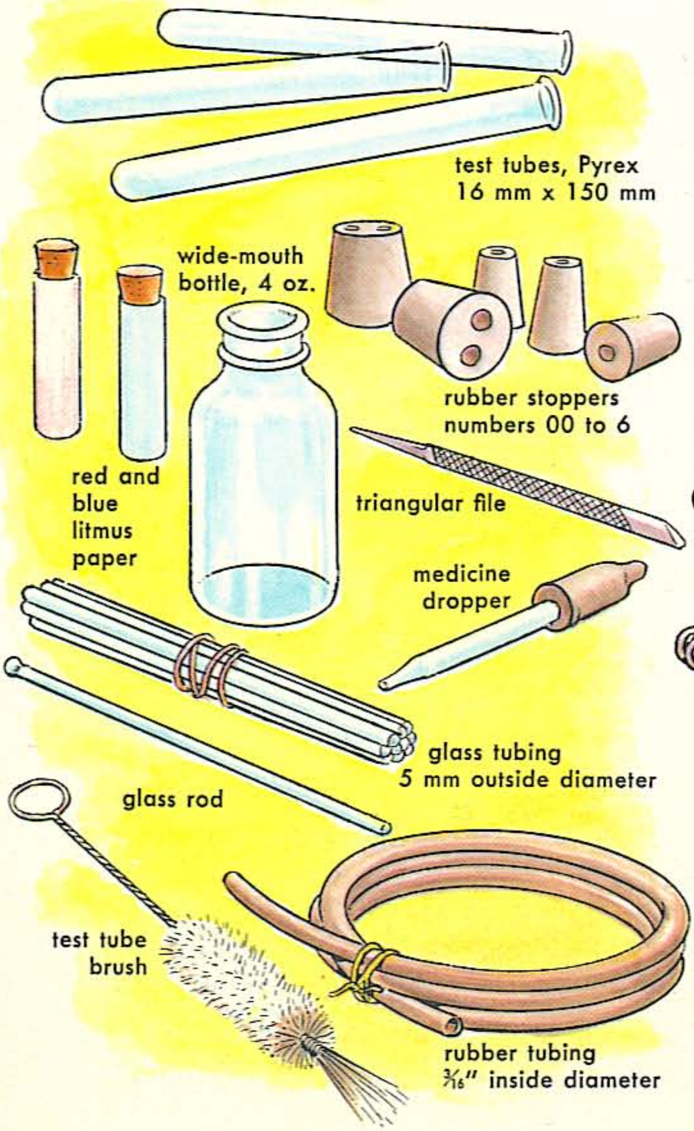
test tubes, Pyrex
16 mm x 150 mm
wide-mouth bottle, 4 oz.
rubber stoppers numbers 00 to 6
red and blue litmus paper
triangular file
medicine dropper
glass tubing 5 mm outside diameter
glass rod
test tube brush
rubber tubing 3/16" inside diameter

IMPROVISED EQUIPMENT FOR HOME LAB

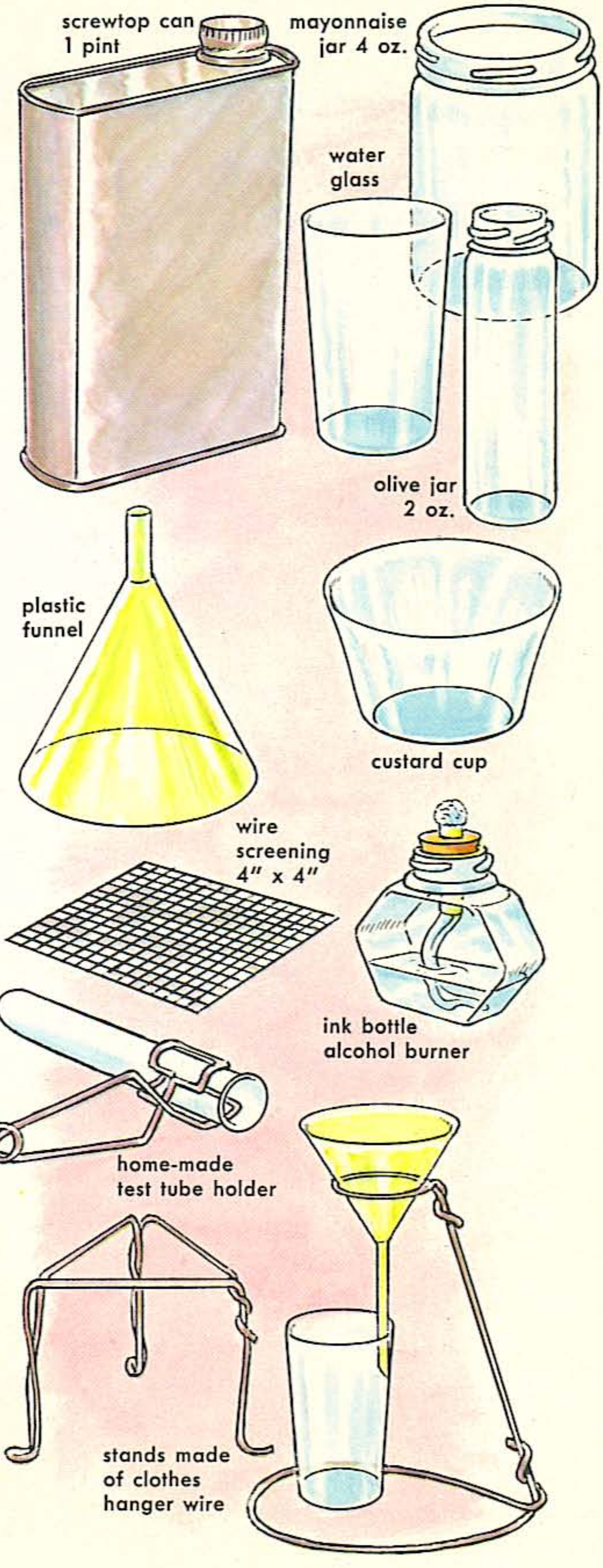
screwtop can 1 pint
mayonnaise jar 4 oz.
water glass
olive jar 2 oz.
plastic funnel
custard cup
wire screening 4" x 4"
ink bottle alcohol burner
home-made test tube holder
stands made of clothes hanger wire

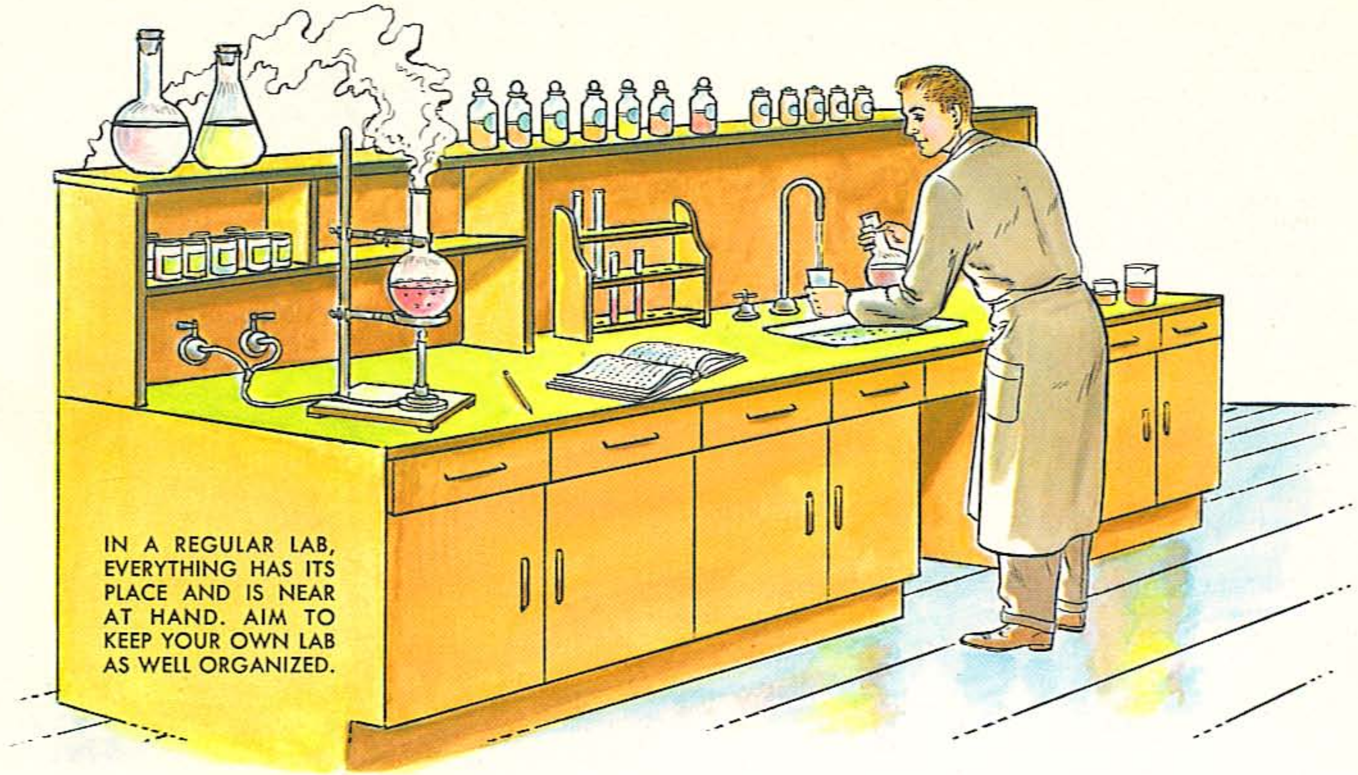

IN A REGULAR LAB, EVERYTHING HAS ITS PLACE AND IS NEAR AT HAND. AIM TO KEEP YOUR OWN LAB AS WELL ORGANIZED.

Setting Up Your Home Laboratory

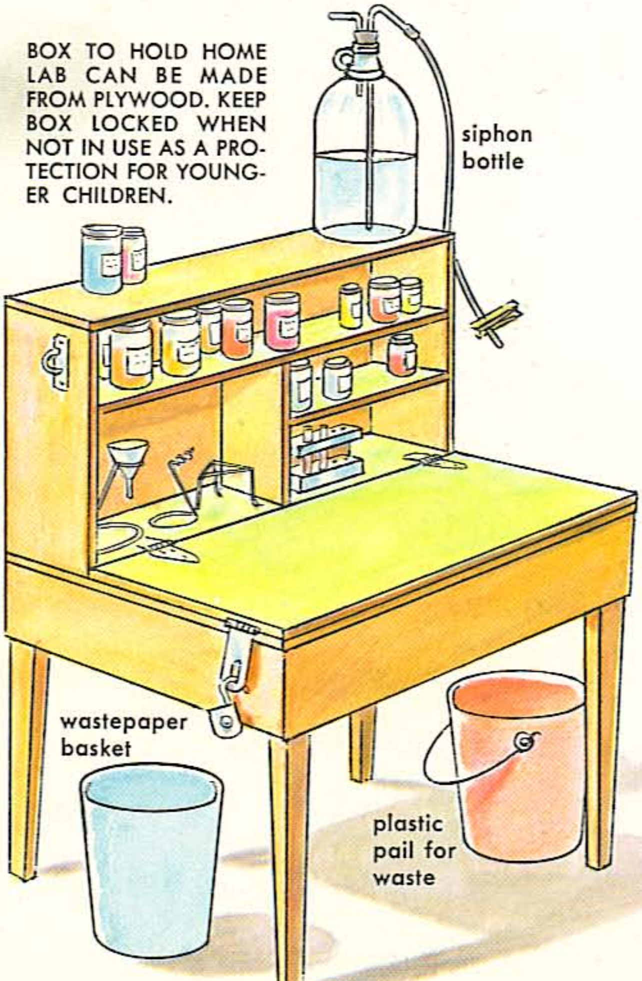

BOX TO HOLD HOME LAB CAN BE MADE FROM PLYWOOD. KEEP BOX LOCKED WHEN NOT IN USE AS A PROTECTION FOR YOUNGER CHILDREN.

It is possible that you may be permitted to work at the kitchen table when this is not in use. But it is far better if you have a place where you will not be disturbed and where you can store your equipment — a corner in your room, or in the basement or the garage.

These are the things you'll need in your lab:

Work Table. An old, sturdy table will do. Cover it with a plastic top to protect the wood.

Water Supply. If you have a faucet nearby, fine. Otherwise, make a siphon bottle (page 11).

Waste Disposal. If you can dump your waste directly into the kitchen *drain* (NOT into the sink), you are all right. If not, collect it in a plastic pail to be thrown out when you're finished.

Source of Heat. In the regular laboratory, special gas burners are used. In the home lab, you can use a burner for denatured alcohol. Have a shallow metal pan under the burner for fire safety.

Storage. If there's no one around to disturb your chemicals and equipment, an open shelf is OK. Otherwise, use a box that can be locked up.

Containers. Keep chemicals in glass jars and bottles. LABEL THEM ALL CLEARLY.

Stands. Make your own test tube stand as well as stands for holding glassware for heating.

IF YOU DO NOT HAVE RUNNING WATER IN YOUR LAB, MAKE SIPHON FROM A 1-GALLON BOTTLE. START SIPHON BY BLOWING THROUGH L-SHAPED GLASS TUBE. SPRING CLOTHES PIN REGULATES WATER FLOW.

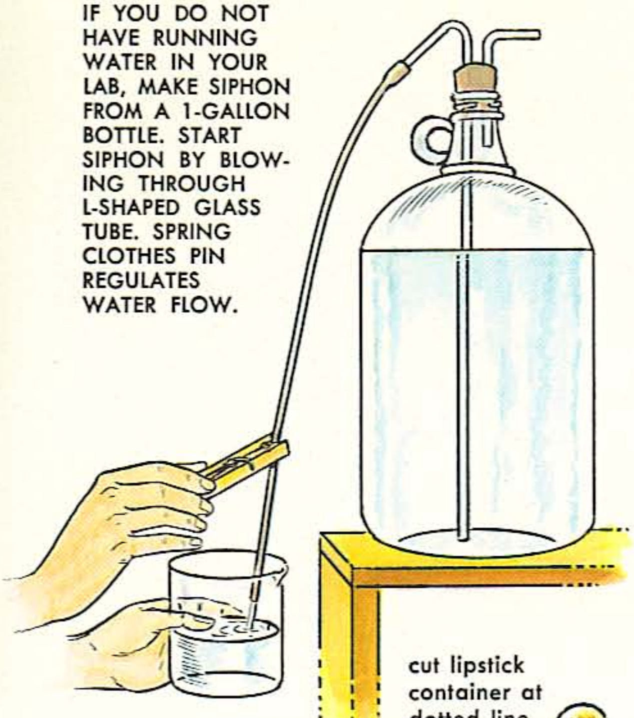

YOU CAN MAKE A SIMPLE TEST TUBE STAND FROM A SALT BOX. CUT OUT SIDE AND USE CUT-OUT SECTION TO KEEP TEST TUBES LINED UP STRAIGHT.

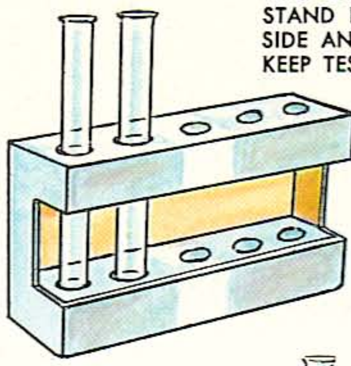

REGULAR LAB STAND FOR TEST TUBES CAN BE PRODUCED FROM STRIPS OF ¼-INCH WOOD. FOLLOW PATTERNS BELOW TO MAKE TWO SIDES AND THREE HORIZONTAL PIECES. DRILL HOLES. ASSEMBLE WITH THIN TACKS.

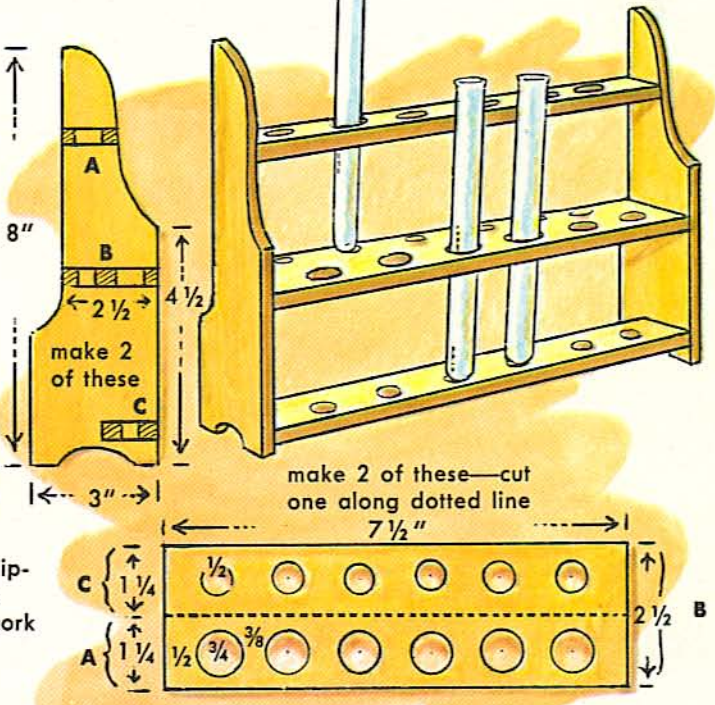
8″
A
B
C
2½
4½
make 2 of these
3″
make 2 of these—cut one along dotted line
7½″
C
1¼
A
1¼
½
½
¾
⅜
2½
B

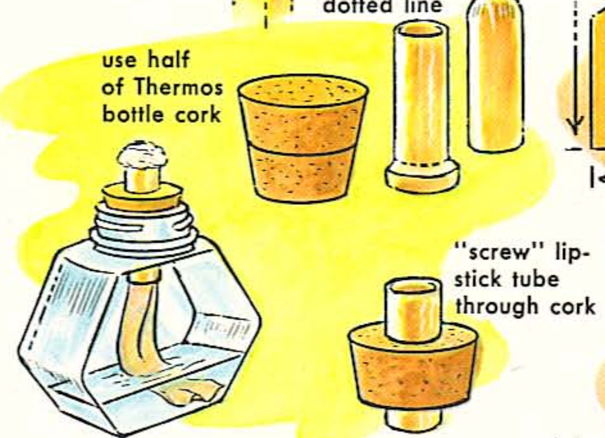
cut lipstick container at dotted line
use half of Thermos bottle cork
"screw" lipstick tube through cork

MAKE AN ALCOHOL BURNER FROM INK BOTTLE, THERMOS BOTTLE CORK, AND LIPSTICK CONTAINER. BUY WICK AT A HARDWARE STORE.

USE A PAIR OF PLIERS FOR MAKING THESE STANDS FROM CLOTHES HANGER WIRE. ALSO MAKE THE FUNNEL STAND ILLUSTRATED ON PAGE 9.

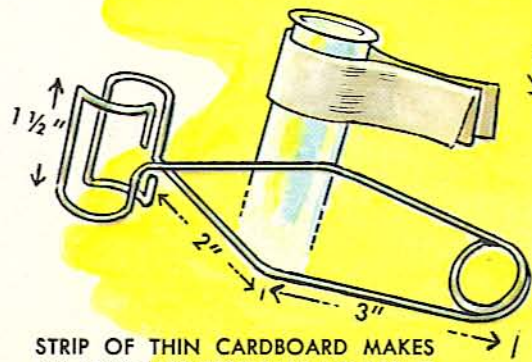

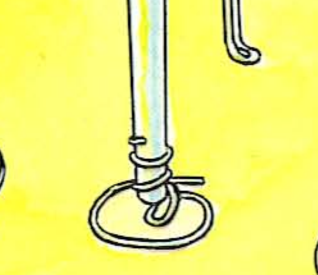

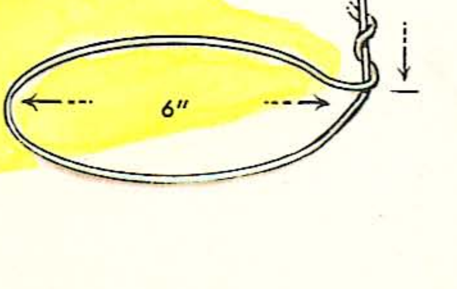

STRIP OF THIN CARDBOARD MAKES AN ADEQUATE TEST TUBE HOLDER. IF YOU LIKE, YOU CAN MAKE A HOLDER FROM CLOTHES HANGER WIRE AS SHOWN HERE.

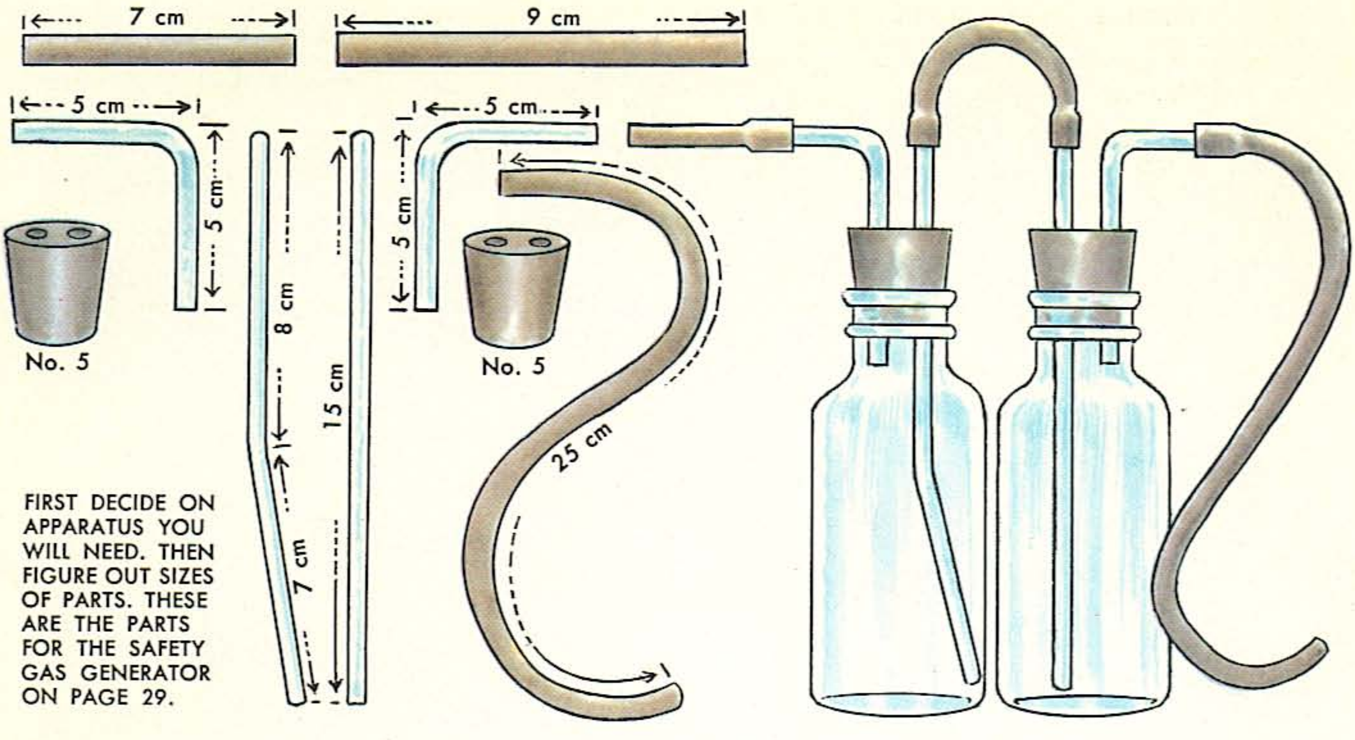

FIRST DECIDE ON APPARATUS YOU WILL NEED. THEN FIGURE OUT SIZES OF PARTS. THESE ARE THE PARTS FOR THE SAFETY GAS GENERATOR ON PAGE 29.

Making Apparatus for Experiments

Most of your chemical experiments you will perform in test tubes and jars. But occasionally you will need an apparatus — a device consisting of bottles and stoppers, glass and rubber tubing.

A good chemist takes pride in his apparatus. He makes it with great care — not just for looks but, more important, for safety. An apparatus that leaks flammable gas can be very dangerous.

Before you start to put an apparatus together, make a simple drawing of it so that you will know what it will consist of. Then get out the various parts you will need to put it together.

To make an apparatus, you need to know how to cut a glass tube, how to bend it, and how to draw it to a jet point. See page 13.

It is wise to use glass tubes of one diameter only, with rubber tubing to fit. Glass tubes of an *outside* diameter of 6 millimeters fit snugly into the holes in the usual rubber stoppers. Rubber tubing of an *inside* diameter of $\frac{3}{16}$" fits over the 6mm glass tubes.

To determine the right size stoppers to use in the bottles of your apparatus, measure the mouths of the bottles against the stoppers shown below in actual size. Order stoppers by number. Keep a selection of different sizes on hand.

Follow the safety precautions on page 16.

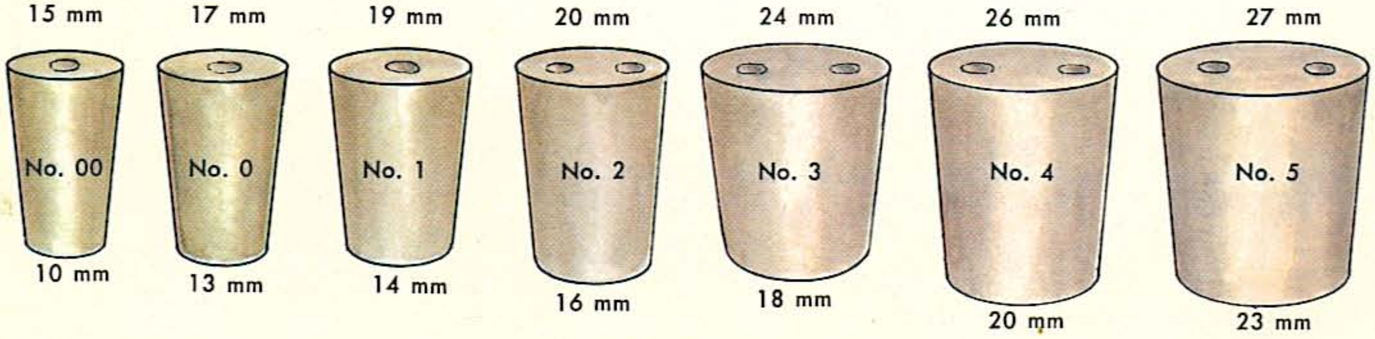

THESE ARE THE ACTUAL SIZES OF RUBBER STOPPERS. BY MEASURING THEM AGAINST YOUR LAB WARE YOU WILL KNOW WHICH TO ORDER. No. 0 FITS THE 16 mm TEST TUBE. No. 5 FITS 4-OZ. WIDE-MOUTH BOTTLE.

CUTTING GLASS TUBES

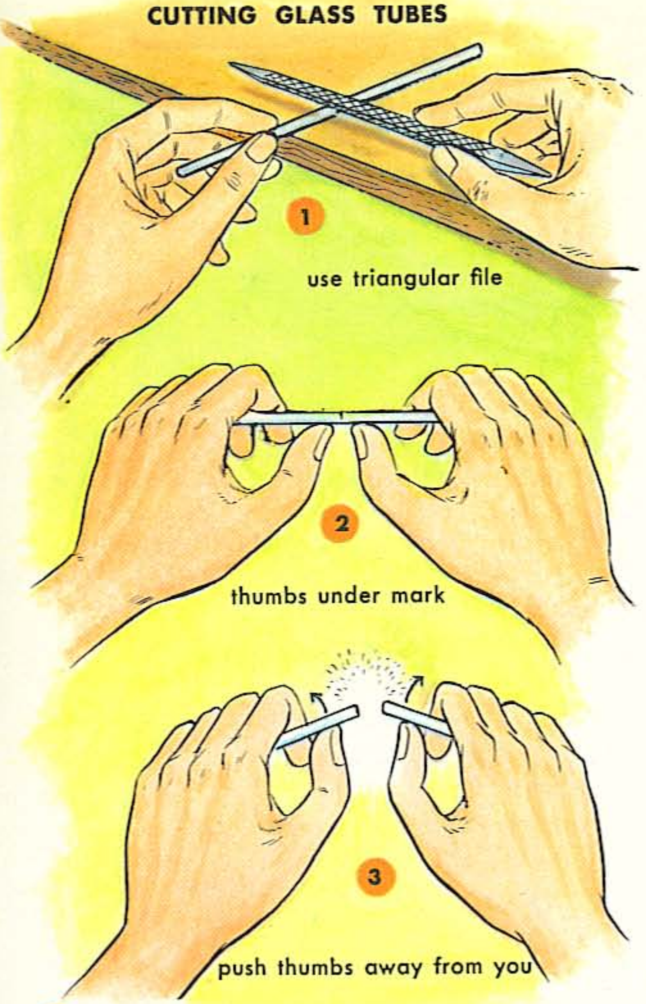
1
use triangular file
2
thumbs under mark
3
push thumbs away from you

1. MAKE SINGLE SCRATCH WITH FILE—DO NOT "SAW."
2. HOLD TUBE WITH BOTH HANDS, THUMBS BELOW MARK.
3. SNAP TUBE INTO TWO PIECES WITH A QUICK JERK.

FIRE GLAZING CUT EDGES

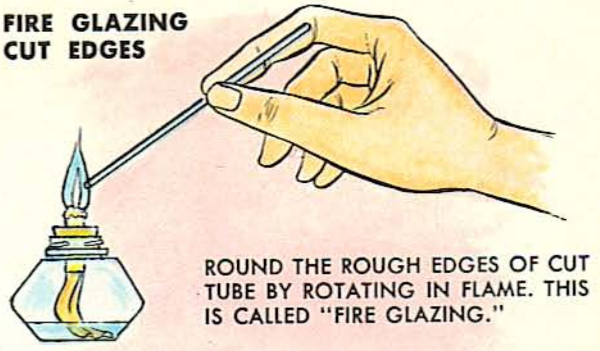

ROUND THE ROUGH EDGES OF CUT TUBE BY ROTATING IN FLAME. THIS IS CALLED "FIRE GLAZING."

MAKING JET POINT

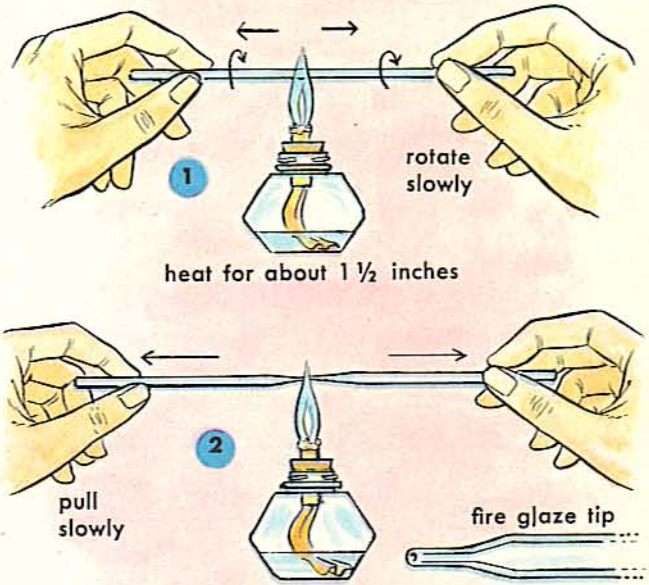
1
rotate slowly
heat for about 1 ½ inches
2
pull slowly
fire glaze tip

1. HEAT SECTION OF TUBE EVENLY WHERE YOU WANT TO FORM JET POINT, BY ROLLING IT IN THE FLAME.
2. WHEN SOFT, PULL SLOWLY. CUT APART WHEN HARD.

BENDING GLASS TUBES

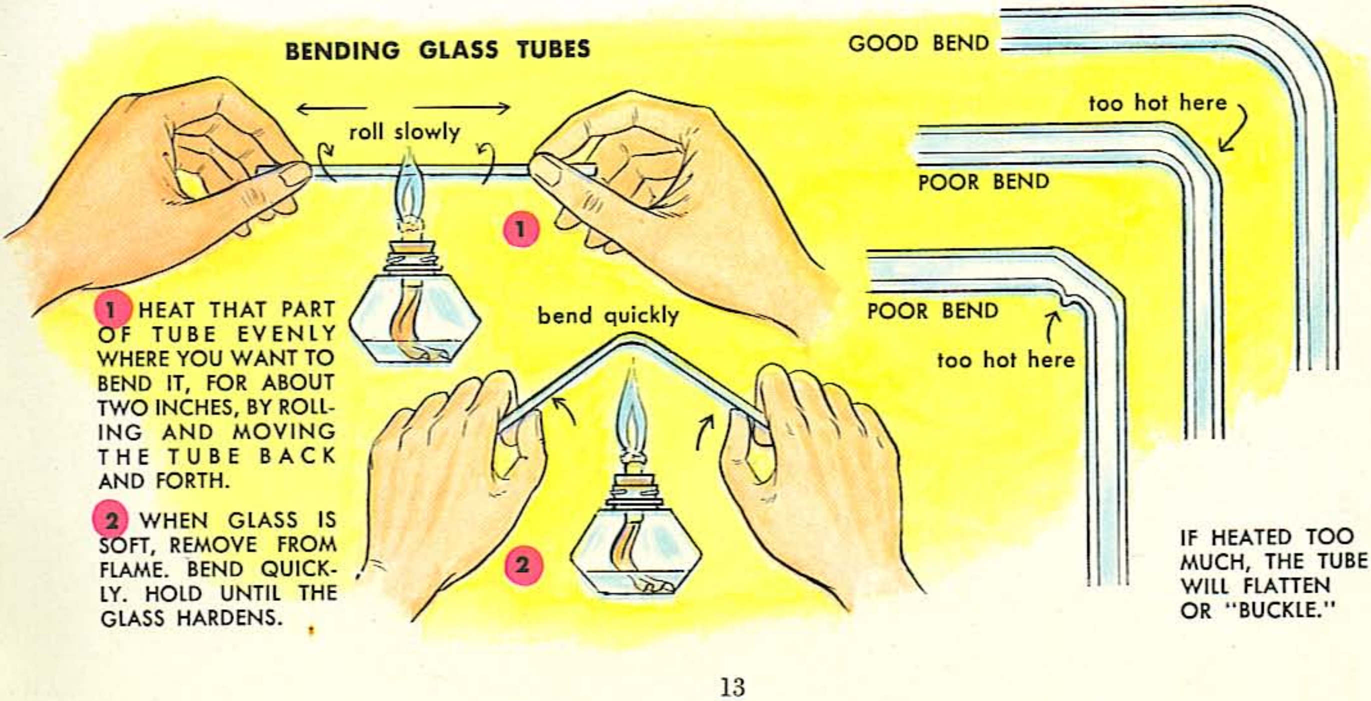
roll slowly
1
bend quickly
2
GOOD BEND
too hot here
POOR BEND
POOR BEND
too hot here

1. HEAT THAT PART OF TUBE EVENLY WHERE YOU WANT TO BEND IT, FOR ABOUT TWO INCHES, BY ROLLING AND MOVING THE TUBE BACK AND FORTH.
2. WHEN GLASS IS SOFT, REMOVE FROM FLAME. BEND QUICKLY. HOLD UNTIL THE GLASS HARDENS.

IF HEATED TOO MUCH, THE TUBE WILL FLATTEN OR "BUCKLE."

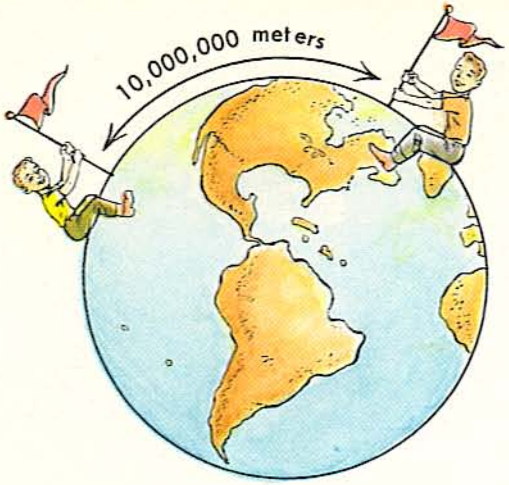

Scientific Measurements

In science, the metric system is preferred over our usual system. It is much easier to work with when once you have learned it — for instead of dividing or multiplying by 12 or 32 or 16 to go from one unit to the next, you simply move the decimal point. Just remember these two things:

1. That the names of the basic units are meter for lengths, liter for volumes, grams for weights — abbreviated to m, l, and g (without a period after them).

2. That 1000 of a kind are called kilo; 100, hekto; 10, deca; 1/10 is called deci; 1/100, centi; 1/1000, milli.

METRIC UNITS OF LENGTH

1000 meters (m) = 1 kilometer (km)
1 meter (m) = 1000 millimeters (mm)
1 meter (m) = 39.37 inches
2.540 centimeters (cm) = 1 inch

METRIC UNITS OF VOLUME

1 liter (l) = 1000 cubic centimeters (cm^3 or cc)
1 liter (l) = 1000 milliliters (ml)
1 liter (l) = 1.06 quarts (liquid)
0.946 liter (l) = 1 quart (liquid)

METRIC UNITS OF WEIGHT

1000 grams (g) = 1 kilogram (kg)
1 gram (g) = 1000 milligrams (mg)
1 gram (g) = 0.035 ounces avoirdupois
28.350 grams (g) = 1 ounce avoirdupois

1 cubic centimeter =
1 cm^3 = 1 cc =
1 milliliter =
1 ml

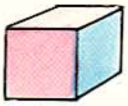

Making a Model of a Cubic Centimeter

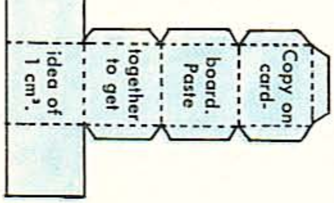

BOTTOM LINE OF FIGURE TO THE RIGHT IS 1 DECIMETER (1 dm) OR 10 CENTIMETERS (10 cm) OR 100 MILLIMETERS (100 mm).

FRONT SURFACE OF FIGURE IS 1 SQUARE DECIMETER (1 dm^2) OR 100 SQUARE CENTIMETERS (100 cm^2).

VOLUME OF WHOLE FIGURE IS 1 CUBIC DECIMETER (1 dm^3) OR 1000 CUBIC CENTIMETERS (1000 cm^3 OR 1000 cc) OR 1000 MILLILITERS (1000 ml) OR 1 LITER (1 l). 1 l WATER WEIGHS 1 KILOGRAM (1 kg).

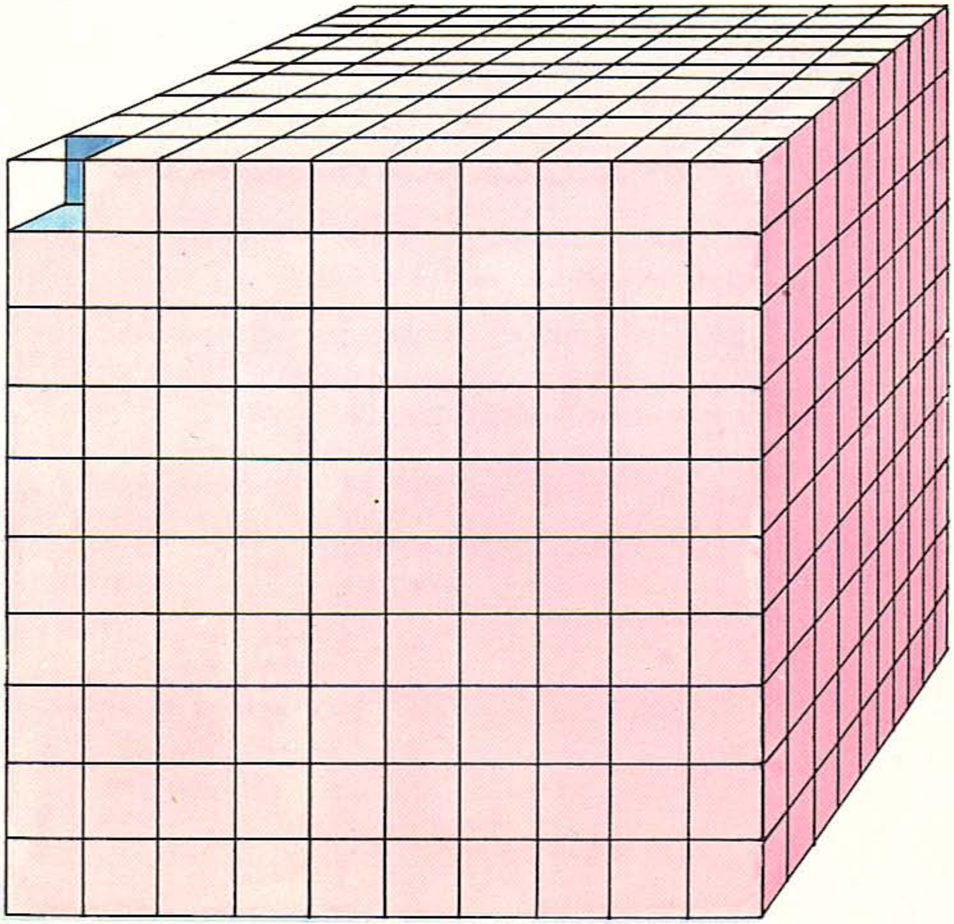

1 cubic decimeter = 1 liter = 1000 cubic centimeters = 1000 milliliters

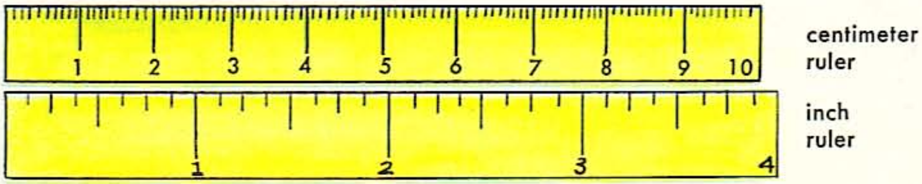

MAKING A HAND BALANCE

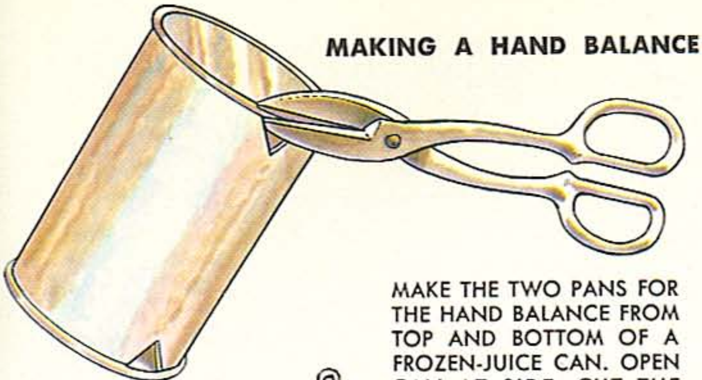

MAKE THE TWO PANS FOR THE HAND BALANCE FROM TOP AND BOTTOM OF A FROZEN-JUICE CAN. OPEN CAN AT SIDE. CUT THE PANS WITH TIN SNIPS.

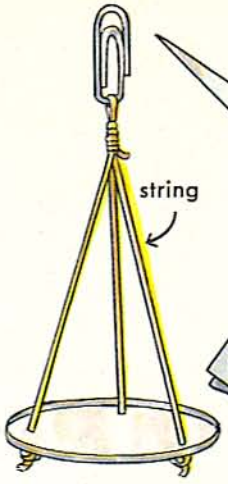

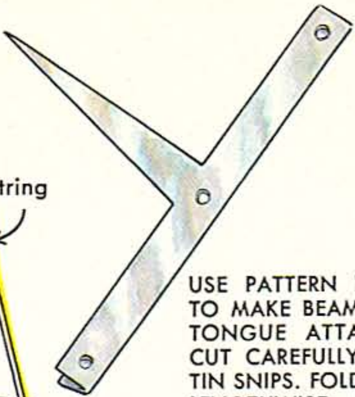

USE PATTERN BELOW TO MAKE BEAM WITH TONGUE ATTACHED. CUT CAREFULLY WITH TIN SNIPS. FOLD BEAM LENGTHWISE. PUNCH HOLES WITH NAIL.

FOR MANY EXPERIMENTS YOU NEED TO WEIGH CHEMICALS. FOR THIS, MAKE A HAND BALANCE.

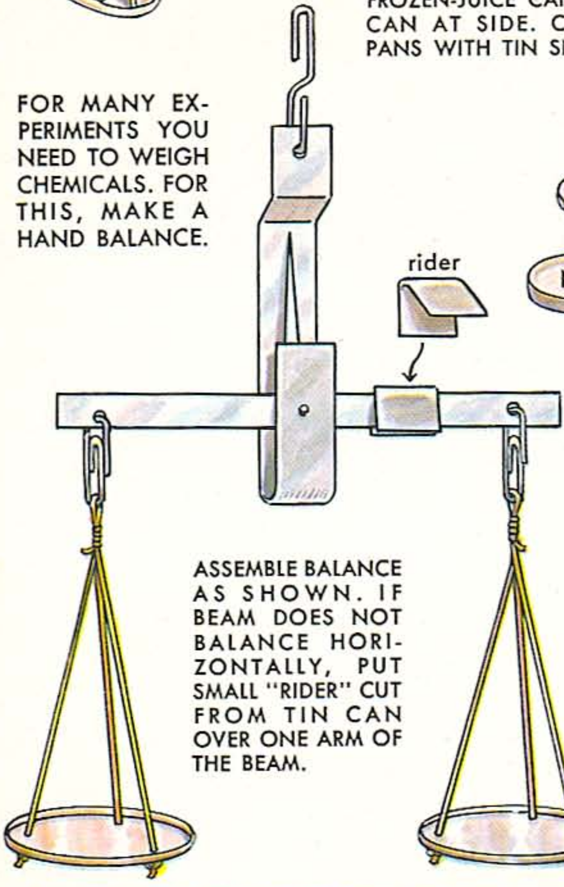

ASSEMBLE BALANCE AS SHOWN. IF BEAM DOES NOT BALANCE HORIZONTALLY, PUT SMALL "RIDER" CUT FROM TIN CAN OVER ONE ARM OF THE BEAM.

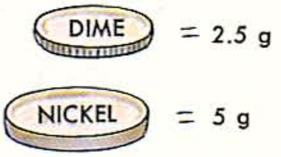
DIME
= 2.5 g
NICKEL
= 5 g

YOU CAN USE COINS TO WEIGH 2.5 g, 5 g, 7.5 g, 10 g, ETC. FOR SMALLER WEIGHTS, CUT A TIN STRIP TO WEIGH THE SAME AS A NICKEL. THEN CUT IT IN FOUR 1 g AND TWO .5 g WEIGHTS.

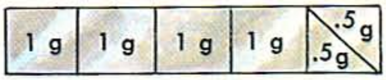
1 g
1 g
1 g
1 g
.5 g
.5 g

MAKING A GRADUATE

A GRADUATE IS USED FOR MEASURING LIQUIDS.

YOUR 6" TEST TUBE HOLDS 22 ml. YOU CAN USE IT FOR ROUGH MEASUREMENTS: TUBE NOT QUITE FULL IS 20 ml, NOT QUITE HALF FULL 10 ml. FOR SMALLER AMOUNTS, COPY THE RULER TO THE RIGHT AND ATTACH IT TO THE SIDE OF A TEST TUBE WITH SCOTCH TAPE FOR MEASURING ml's.

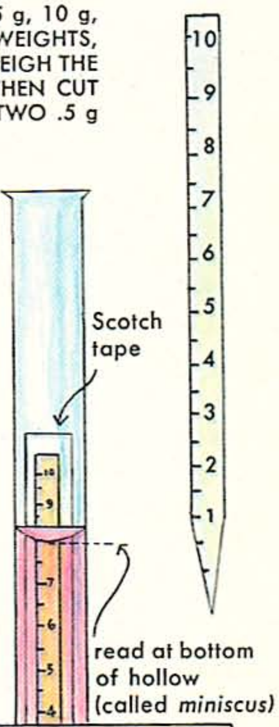
10
9
8
7
6
5
4
3
2
1
Scotch tape
read at bottom of hollow (called *miniscus*)

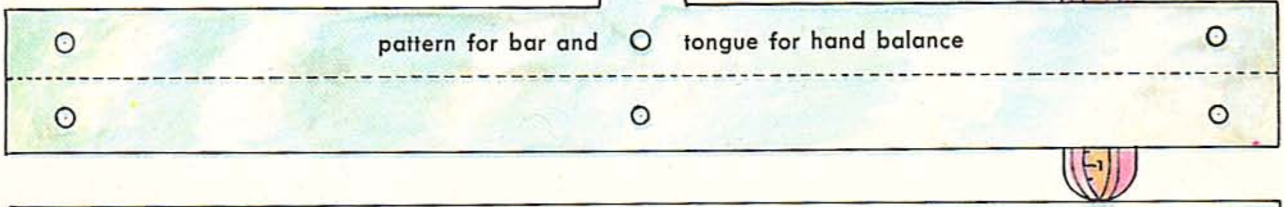
tongue
pattern for bar and tongue for hand balance

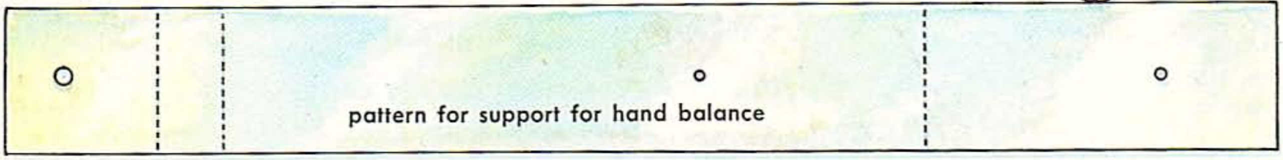
pattern for support for hand balance

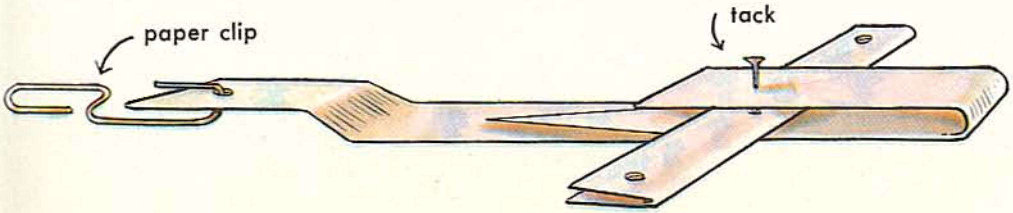
paper clip
tack

COPY THE PATTERNS ABOVE ONTO A PIECE OF PAPER. TRANSFER DESIGNS TO PIECE OF TIN CAN. CUT OUT AND BEND AS SHOWN TO THE LEFT.

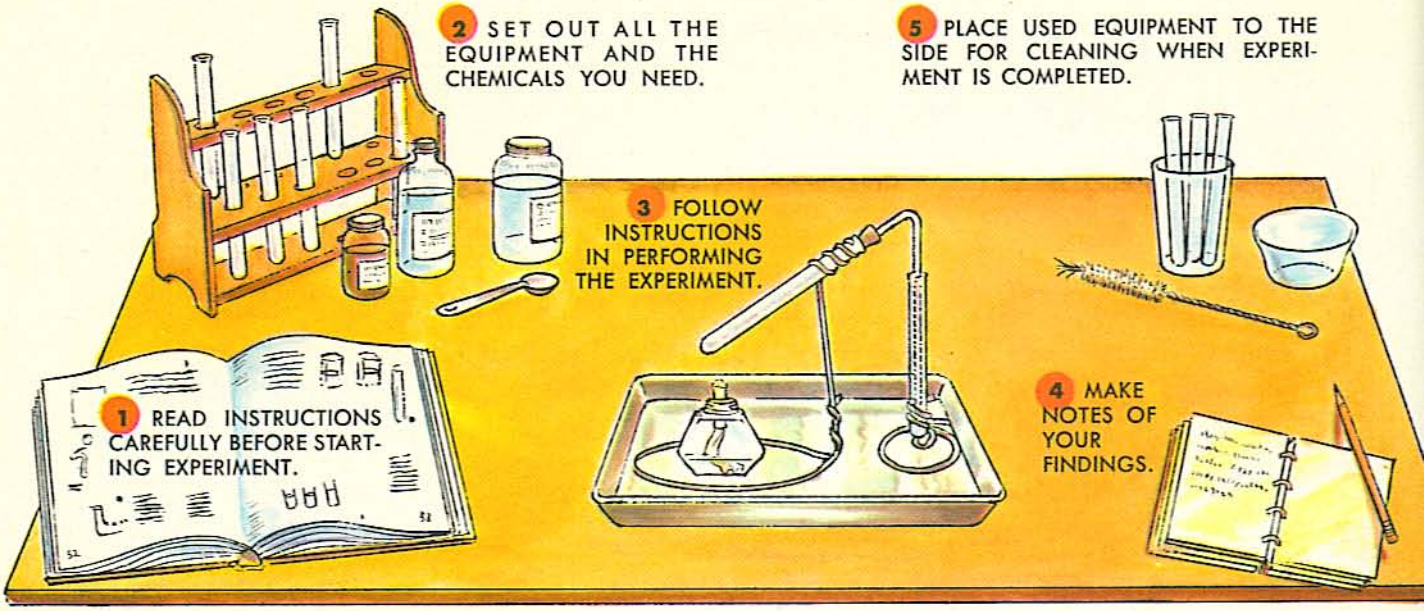

1 READ INSTRUCTIONS CAREFULLY BEFORE STARTING EXPERIMENT.
2 SET OUT ALL THE EQUIPMENT AND THE CHEMICALS YOU NEED.
3 FOLLOW INSTRUCTIONS IN PERFORMING THE EXPERIMENT.
4 MAKE NOTES OF YOUR FINDINGS.
5 PLACE USED EQUIPMENT TO THE SIDE FOR CLEANING WHEN EXPERIMENT IS COMPLETED.

Correct Laboratory Techniques

In your home laboratory, three considerations are of the greatest importance: SAFETY, NEATNESS, and EXACTNESS.

SAFETY — All the experiments in this book are safe when done in the correct laboratory way as shown on these pages.

Treat chemicals with respect. Never taste anything unless specifically told to do so. If there are younger children in the family, lock up your chemicals when you are not working with them.

Protect your clothes with a plastic apron.

Be careful with fire. When you use your alcohol burner, have a metal pan under it for safety.

NEATNESS — Get the habit of lining up equipment and chemicals you need on one side and placing used items on the opposite side — keeping the space between them clear for your experiments.

Put chemicals away and clean glassware as soon as you have finished an experiment.

EXACTNESS — Label all bottles and jars containing chemicals clearly and correctly.

Where amounts of chemicals are not given, use the smallest amount that will tell you what you want to know.

Observe the chemical reactions carefully and make complete notes of them as you go along.

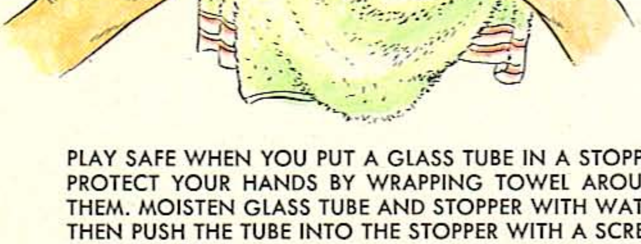

PLAY SAFE WHEN YOU PUT A GLASS TUBE IN A STOPPER. PROTECT YOUR HANDS BY WRAPPING TOWEL AROUND THEM. MOISTEN GLASS TUBE AND STOPPER WITH WATER, THEN PUSH THE TUBE INTO THE STOPPER WITH A SCREWDRIVER MOTION.

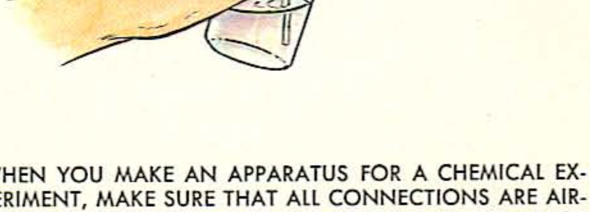

WHEN YOU MAKE AN APPARATUS FOR A CHEMICAL EXPERIMENT, MAKE SURE THAT ALL CONNECTIONS ARE AIRTIGHT. USE THE RIGHT SIZE STOPPER FOR MOUTH OF CONTAINER, GLASS TUBES THAT FIT SNUGLY INTO THE STOPPER HOLES, AND TIGHT-FITTING RUBBER TUBING.

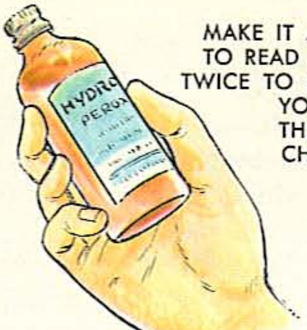

MAKE IT A HABIT TO READ A LABEL TWICE TO BE SURE YOU HAVE THE RIGHT CHEMICAL.

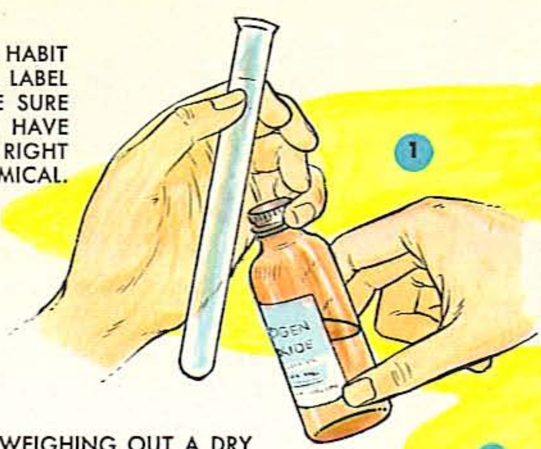

TRAINED CHEMISTS NEVER PUT STOPPER OF BOTTLE ON DESK—THEY KEEP IT IN ONE HAND.

1 TAKE STOPPER OFF BOTTLE WITH YOUR LEFT HAND.

2 KEEP STOPPER IN THE LEFT HAND WHILE YOU POUR FROM THE BOTTLE IN YOUR RIGHT, WITH THE LABEL FACING UP. THEN REPLACE THE STOPPER.

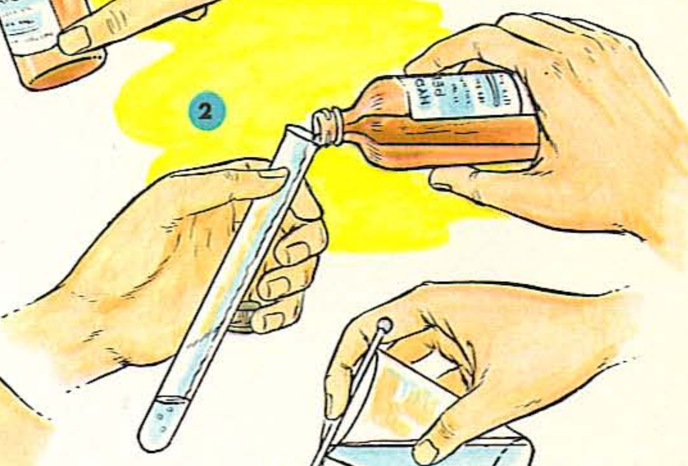

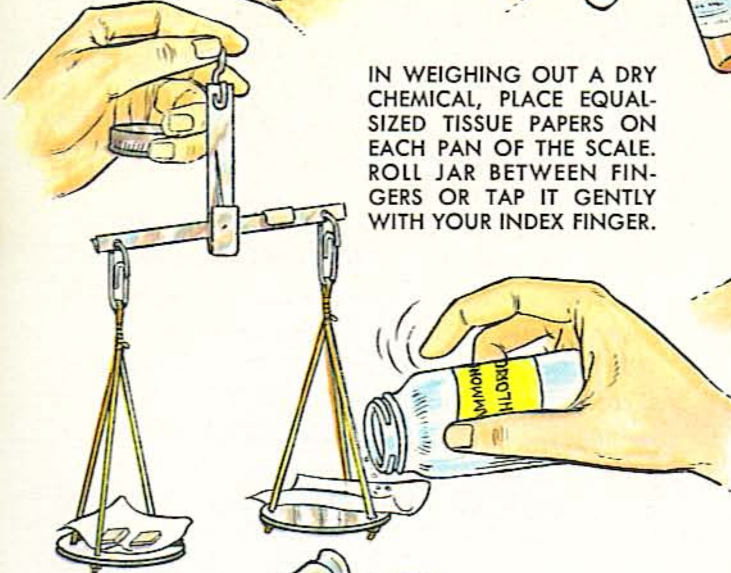

IN WEIGHING OUT A DRY CHEMICAL, PLACE EQUAL-SIZED TISSUE PAPERS ON EACH PAN OF THE SCALE. ROLL JAR BETWEEN FINGERS OR TAP IT GENTLY WITH YOUR INDEX FINGER.

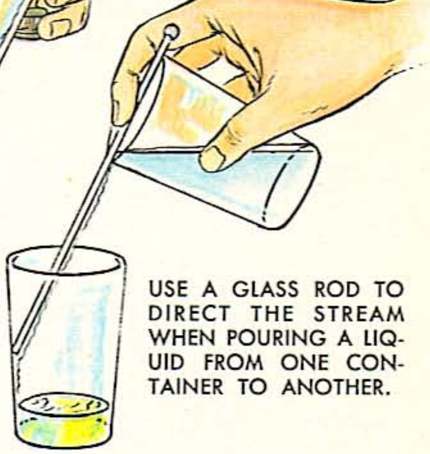

USE A GLASS ROD TO DIRECT THE STREAM WHEN POURING A LIQUID FROM ONE CONTAINER TO ANOTHER.

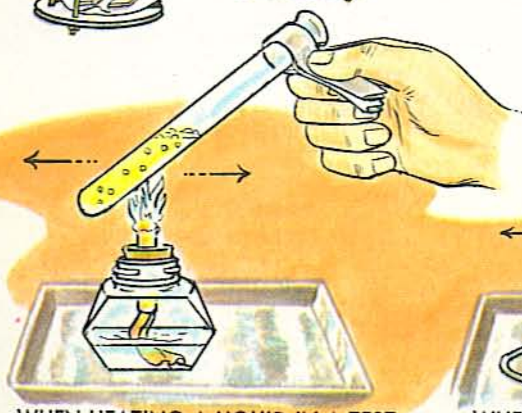

WHEN HEATING A LIQUID IN A TEST TUBE, HOLD TUBE WITH A HOLDER. KEEP THE TUBE MOVING. DO NOT HEAT BOTTOM OF TUBE—CONTENTS MAY "BUMP" AND SQUIRT OUT. NEVER POINT MOUTH OF TEST TUBE TOWARD YOURSELF OR TOWARD ANYONE ELSE.

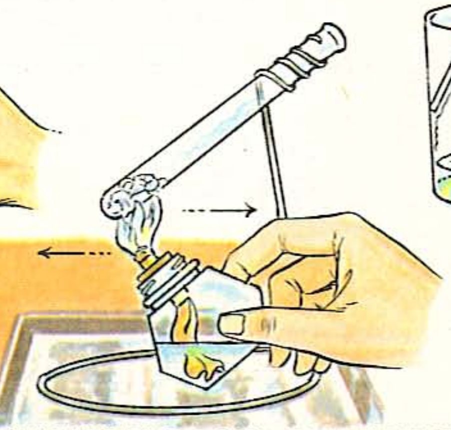

WHEN HEATING A SOLID IN A TEST TUBE, PLACE THE TUBE IN A STAND AND MOVE THE FLAME OF YOUR BURNER BACK AND FORTH TO HEAT THE CONTENTS EVENLY.

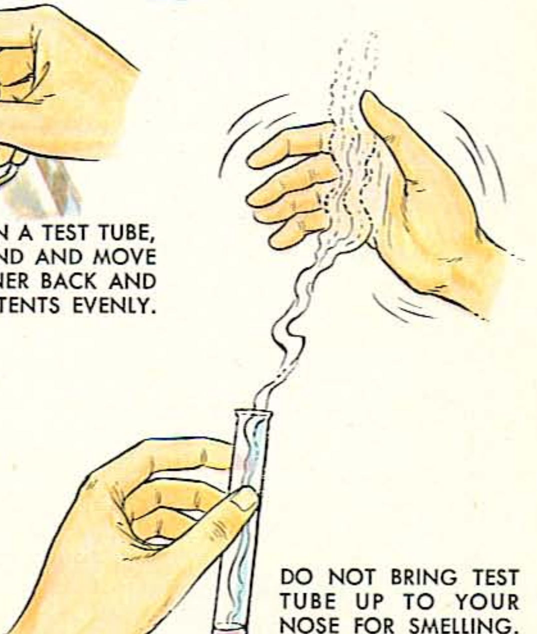

DO NOT BRING TEST TUBE UP TO YOUR NOSE FOR SMELLING. INSTEAD, WAFT THE ODORS TOWARD YOU WITH YOUR HAND.

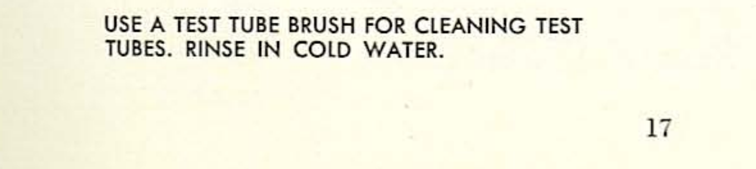

USE A TEST TUBE BRUSH FOR CLEANING TEST TUBES. RINSE IN COLD WATER.

Mr. Faraday's Candle

In the winter of 1859, Michael Faraday, a great British scientist, gave a number of lectures for young people. The talks dealt with one subject only: the features or "phenomena" of — a candle!

"There is not a law," Faraday told his listeners, "under which any part of this universe is governed which does not come into play and is touched upon in these phenomena. There is no better, there is no more open door by which you can enter into the study of natural philosophy than by considering the phenomena of a candle." He then set out to prove his point by lighting a candle and demonstrating all the processes involved.

In burning a candle you start with a SOLID substance that turns, first, into a LIQUID, then into a GAS (or, more correctly, into a gas-like vapor). The melted candle grease is held in a level position by GRAVITY yet seems to defy gravity by rising in the wick by a force called CAPILLARY ACTION. In burning, the candle produces ENERGY in the form of LIGHT and HEAT. At the same time, it goes into CHEMICAL REACTIONS that reveal what it is made of.

As you enter the study of chemistry, you can do no better than to repeat for yourself some of the experiments that Mr. Faraday demonstrated to his young audience.

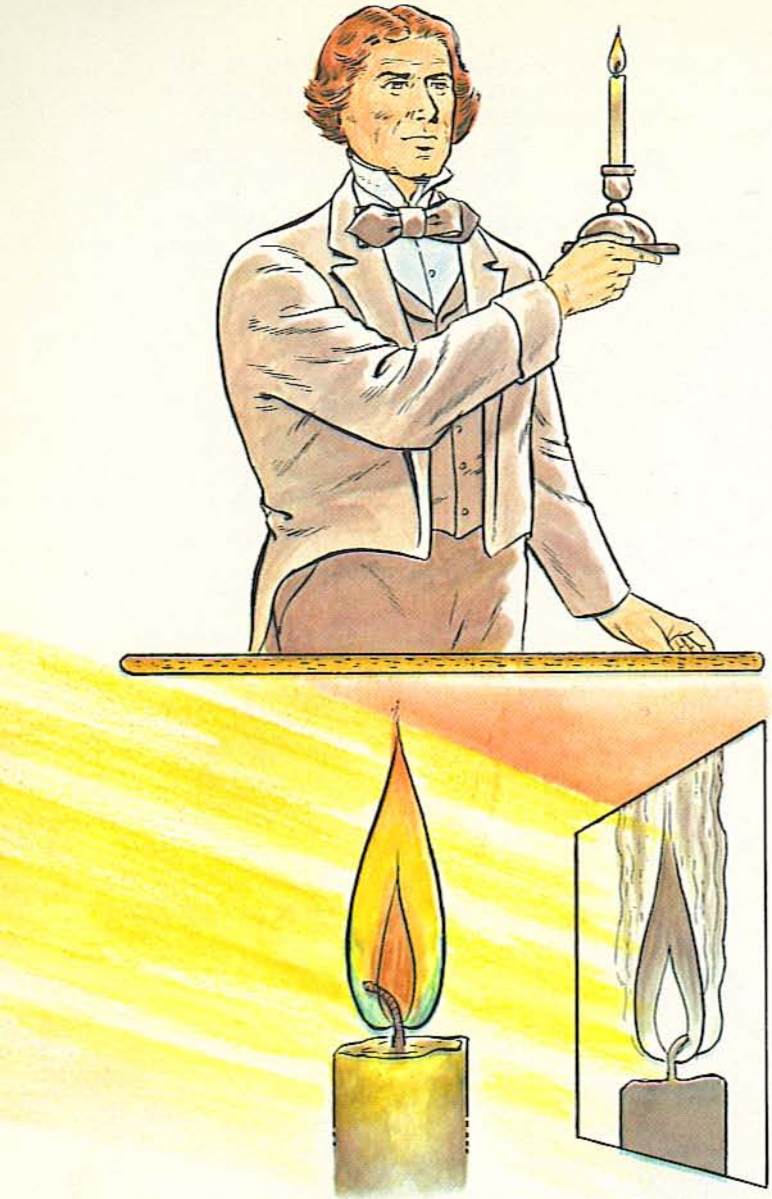

PLACE A BURNING CANDLE IN THE SUN AND CATCH THE SHADOW ON A PIECE OF WHITE PAPER. YOU WILL DISCOVER THAT IT IS THE BRIGHTEST PART OF THE FLAME THAT CASTS THE DARKEST SHADOW.

CANDLE FLAME IS BURNING VAPOR

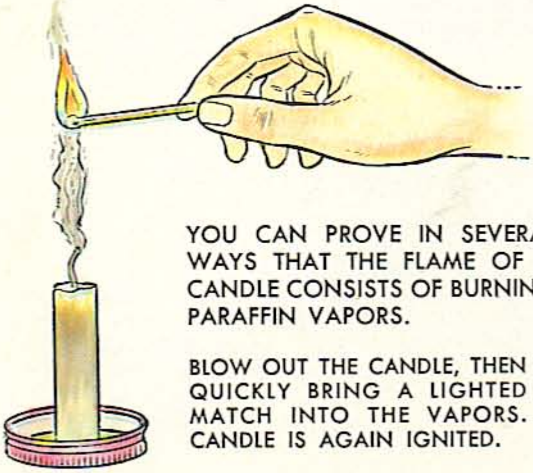

YOU CAN PROVE IN SEVERAL WAYS THAT THE FLAME OF A CANDLE CONSISTS OF BURNING PARAFFIN VAPORS.

BLOW OUT THE CANDLE, THEN QUICKLY BRING A LIGHTED MATCH INTO THE VAPORS. CANDLE IS AGAIN IGNITED.

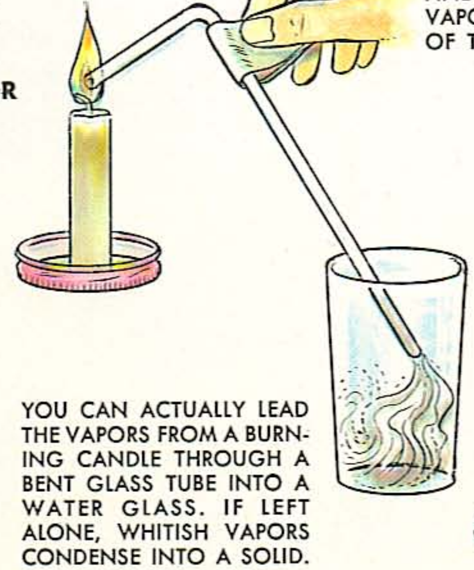

YOU CAN ACTUALLY LEAD THE VAPORS FROM A BURNING CANDLE THROUGH A BENT GLASS TUBE INTO A WATER GLASS. IF LEFT ALONE, WHITISH VAPORS CONDENSE INTO A SOLID.

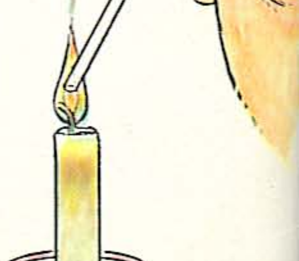

MAKE A GAS WORKS FROM A CANDLE BY HOLDING A GLASS TUBE, DRAWN TO A POINT, IN THE FLAME AND LIGHTING THE VAPORS AT THE END OF TUBE.

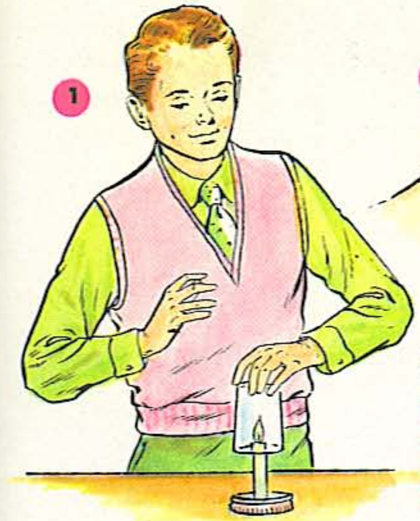

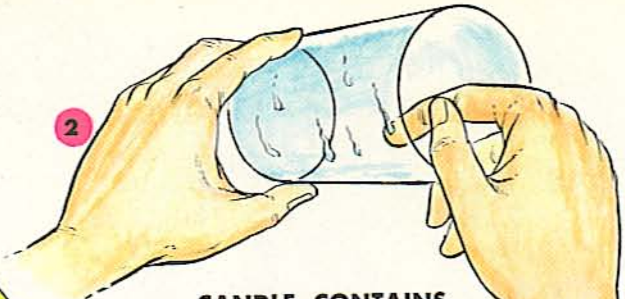

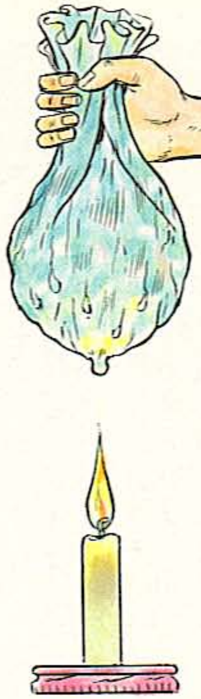

CANDLE CONTAINS HYDROGEN

A CANDLE FORMS WATER WHEN IT BURNS. ONLY HYDROGEN BURNING IN OXYGEN FORMS WATER.

① HOLD A COLD GLASS FOR A FEW MOMENTS OVER A BURNING CANDLE. DEW FORMS ON THE INSIDE.

② BY WIPING FINGER INSIDE THE GLASS YOU CAN MAKE DEW FLOW TOGETHER INTO WATER DROPS.

CRUSH A FEW ICE CUBES AND SPRINKLE THEM WITH SALT. WRAP IN ALUMINUM FOIL. HOLD OVER BURNING CANDLE. WATER DROPS WILL FORM ON THE OUTSIDE OF THE FOIL.

CANDLE CONTAINS CARBON

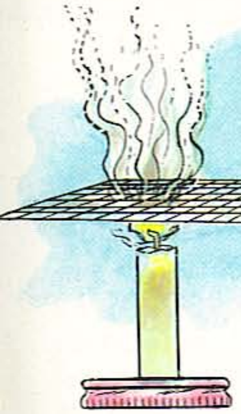

PLACE PIECE OF WIRE SCREENING OVER FLAME. SCREENING COOLS THE FLAME AND SOOT FORMS. THE SOOT IS CARBON.

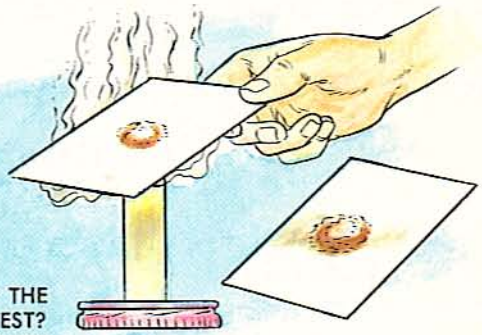

WHAT PART OF THE FLAME IS HOTTEST? TO FIND OUT, PUSH PIECE OF CARDBOARD SIDEWAYS INTO FLAME. OUTSIDE OF FLAME SCORCHES A SOOTY RING.

ANOTHER WAY TO SHOW THAT CANDLE CONTAINS CARBON IS TO PROVE THAT CARBON DIOXIDE IS FORMED WHEN CANDLE BURNS.

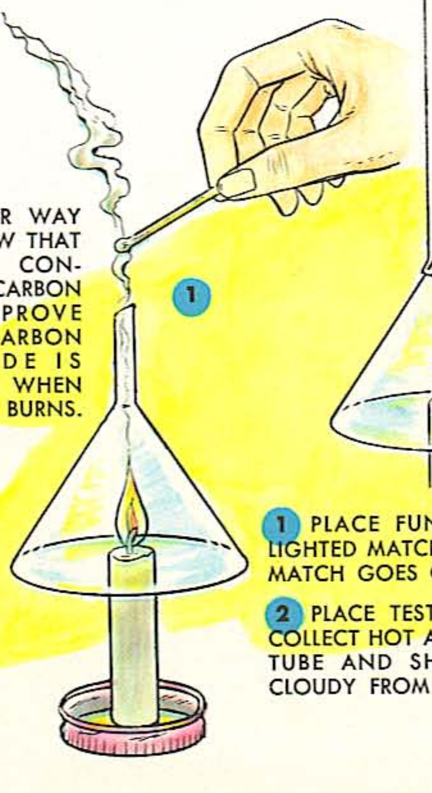

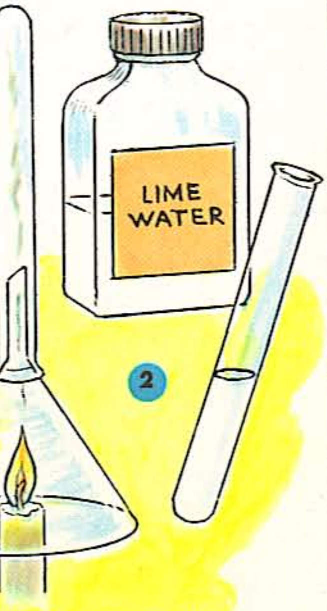
LIME WATER
2

① PLACE FUNNEL OVER FLAME. HOLD LIGHTED MATCH IN HOT AIR FROM FLAME. MATCH GOES OUT IN THIS AIR.

② PLACE TEST TUBE OVER FUNNEL AND COLLECT HOT AIR. POUR LIME WATER INTO TUBE AND SHAKE. LIME WATER GETS CLOUDY FROM CARBON DIOXIDE.

WHAT YOU FOUND OUT

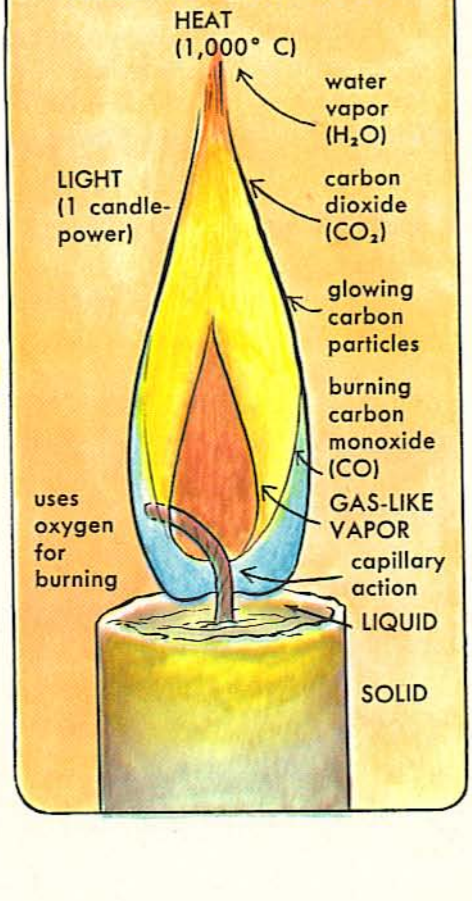
HEAT (1,000° C)
water vapor (H_2O)
LIGHT (1 candle-power)
carbon dioxide (CO_2)
glowing carbon particles
burning carbon monoxide (CO)
uses oxygen for burning
GAS-LIKE VAPOR
capillary action
LIQUID
SOLID

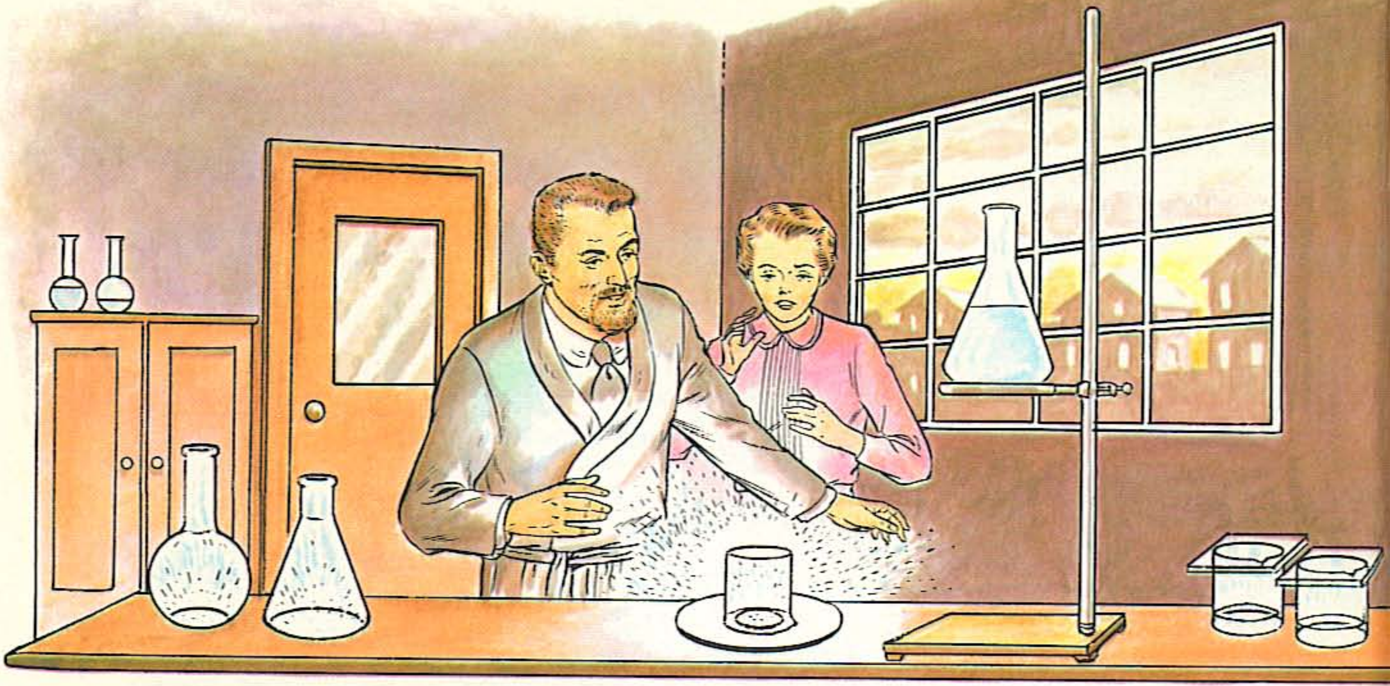

You—Scientist!

In 1896, a young Polish chemist, Marie Curie, and her French husband, Pierre, decided to find out why a certain uranium ore called pitchblende gave off rays that were much stronger than the uranium content of the ore could explain.

They secured a whole ton of powdered ore from a mine in northern Bohemia and set to work. First the powder had to be boiled with strong acids to extract the mysterious substance hidden in it. Then the solution had to be filtered and boiled down. What remained had to be purified by various processes which the Curies had to invent themselves.

After two years of back-breaking work they reached their goal. One night they went to the shed in which they had been working. They opened the door and stepped in without putting on the lights. All around them, the containers that held the solutions of the new substance glowed in the dark! They had discovered a new element — radium — a million times more active than uranium.

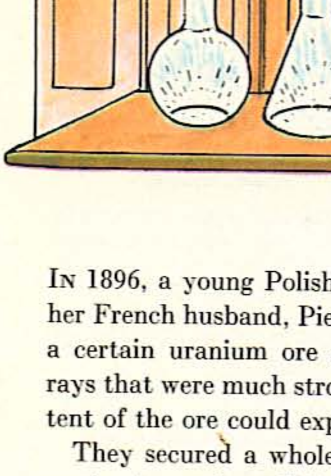

SOLUTION—STIR WATER INTO THE MIXTURE OF SALT AND DIRT. THE WATER WILL DISSOLVE THE SALT BUT NOT THE DIRT. YOU NOW HAVE THE SALT IN "WATERY SOLUTION."

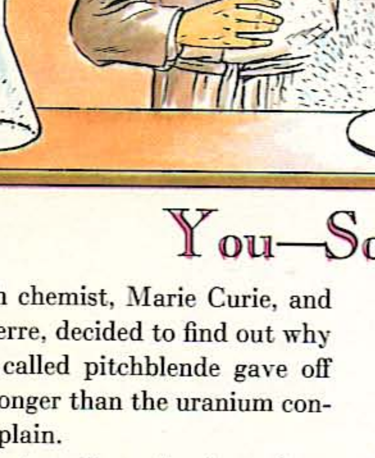

DECANTATION—LET DIRT-MIXED SALT SOLUTION STAND UNTIL MOST OF THE DIRT HAS SETTLED. THEN POUR OFF THE LIQUID. THIS PROCESS IS CALLED "DECANTATION."

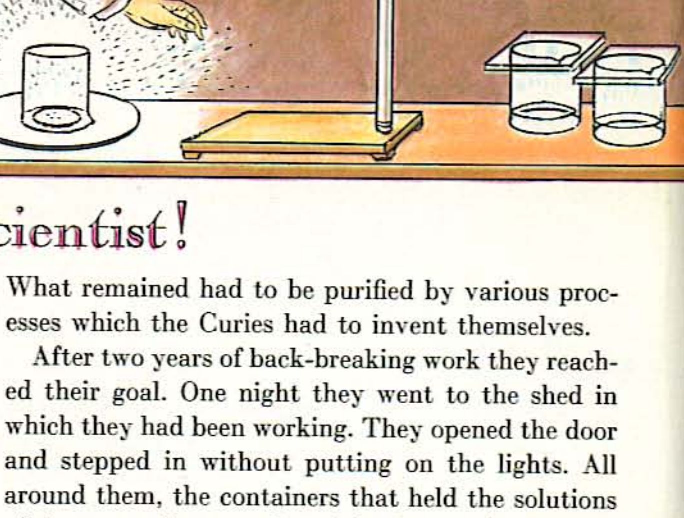

FILTRATION 1—THE LIQUID IS PROBABLY STILL MURKY. TO CLEAR IT, IT NEEDS TO BE FILTERED BY LETTING IT RUN THROUGH FILTER PAPER (PAPER TOWELING WILL DO).

Why tell again the story of the discovery of radium? Because it contains all the features that show the nature of the true scientist.

Curiosity first. The Curies were curious about the mystery that lay in that greyish-black powder. They became obsessed with a desire to find out — not in the hope of gaining money or fame but to establish a scientific truth.

Before starting their work, the Curies gathered all the known facts about the material with which they were to work. To this knowledge they added their own imagination, figuring out the method they had to use to arrive at the result they were seeking.

For the next two years they literally slaved in the drafty shed that was their laboratory.

After they had made their discovery, the Curies made their method of extracting radium known to the world so that other scientists could check and test what they had done.

As an example of the scientific method the Curies used, let us follow in their footsteps — but with a much simpler problem:

1 MIX THOROUGHLY ONE TABLESPOON OF DIRT AND ONE TEASPOON OF ORDINARY TABLE SALT. NOW DECIDE THAT YOU WANT TO EXTRACT THE SALT FROM THIS MIXTURE AS EARNESTLY AS THE CURIES DECIDED TO EXTRACT THE MYSTERIOUS SUBSTANCE FROM PITCHBLENDE —WITH THE EXCEPTION THAT YOU KNOW WHAT YOU ARE AFTER.

2 GET THE FACTS TOGETHER. DIRT IS "DIRTY," SALT IS WHITE. DIRT PARTICLES ARE OF MANY DIFFERENT SHAPES, SALT CONSISTS OF TINY CUBES. DIRT DOES NOT DISSOLVE IN WATER, SALT DOES.

3 NEXT FIGURE OUT A SUITABLE WAY OF SEPARATING THE TWO SUBSTANCES. ON THE BASIS OF WHAT YOU KNOW YOU SHOULD BE ABLE TO SEPARATE THEM WITH A PAIR OF TINY TWEEZERS—BUT IT WOULD PROBABLY TAKE YOU A YEAR TO DO IT. OR YOU COULD DISSOLVE THE SALT IN WATER AND SEPARATE THE SOLUTION FROM THE INSOLUBLE DIRT.

4 YOU DECIDE ON THE SECOND WAY, USING THE STEPS SHOWN ON THE BOTTOM OF THESE PAGES. IN DOING THIS, YOU DO WHAT THE CURIES DID IN EXTRACTING RADIUM AND LEARN, IN THE PROCESS, THE IMPORTANT LABORATORY TECHNIQUES OF SOLUTION, DECANTATION, FILTRATION, EVAPORATION, AND CRYSTALLIZATION.

5 FINALLY, YOU CHECK THE RESULT. THE WHITE SUBSTANCE LEFT AFTER EVAPORATION SHOULD BE SALT—BUT IS IT? IT LOOKS LIKE SALT. IT TASTES LIKE SALT. BY CHEMICAL TESTS YOU CAN PROVE THAT IT *IS* SALT.

By using the same procedure in all other experiments in this book you are learning the methods that real scientists follow in their work — you are becoming a scientist yourself.

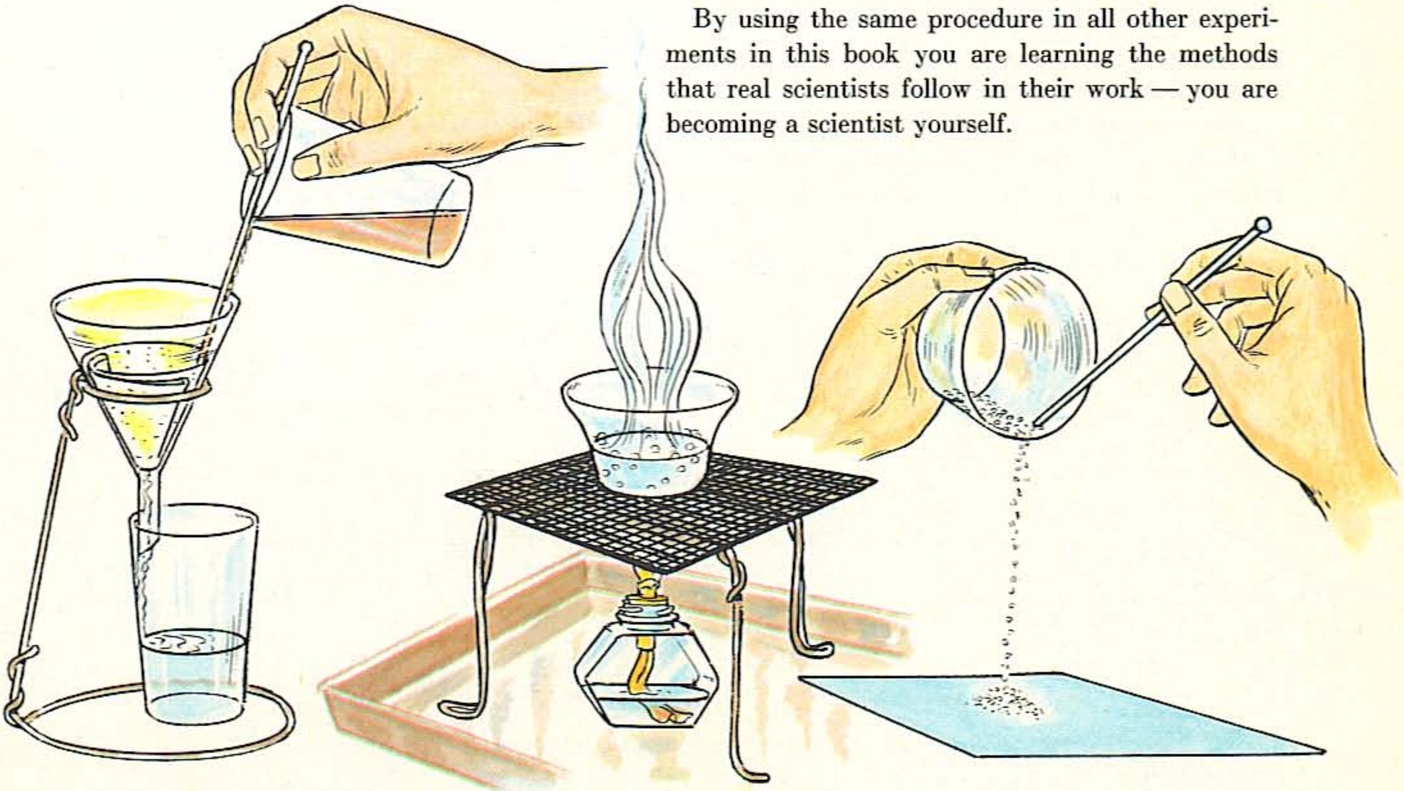

FILTRATION 2—FOLD FILTER PAPER AS SHOWN ON OPPOSITE PAGE AND FIT IT IN FUNNEL. POUR LIQUID ONTO FILTER PAPER. CLEARED LIQUID IS CALLED "FILTRATE."

EVAPORATION—THE FILTRATE CONTAINS THE SALT. THE SALT CAN NOW BE FREED BY REMOVING THE WATER BY BOILING IT AWAY. THIS IS KNOWN AS "EVAPORATION."

CRYSTALLIZATION—AS WATER IS REMOVED, THERE IS TOO LITTLE OF IT LEFT FOR THE SALT TO STAY IN SOLUTION. THE SALT MAKES ITS APPEARANCE AS TINY CRYSTALS.

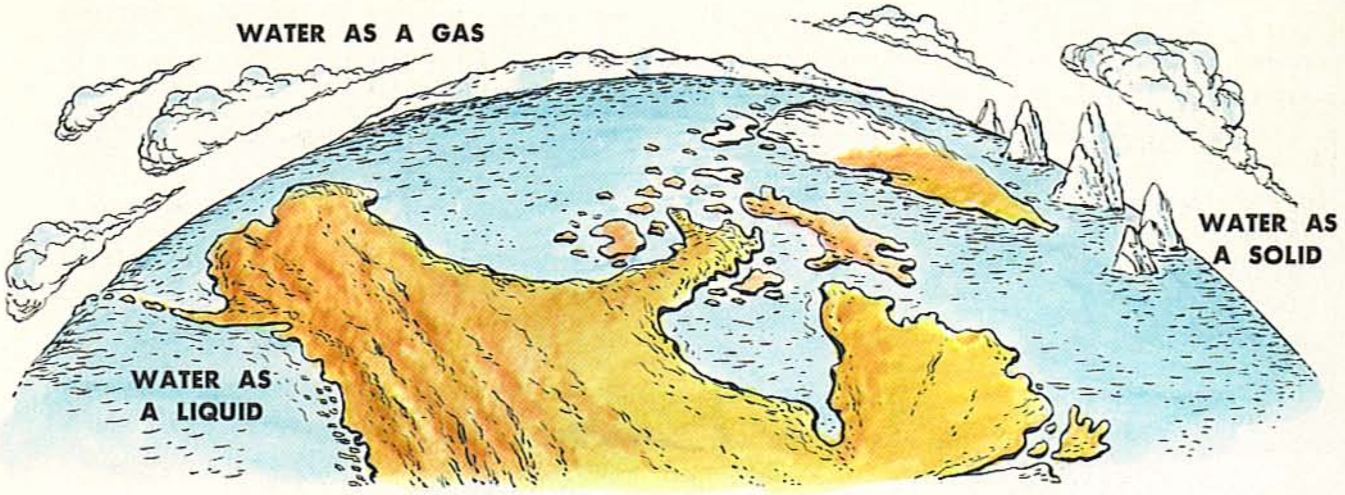

Elements, Compounds, and Mixtures

In all your experiments in chemistry, you will be dealing with "matter."

Matter is anything that takes up room and has weight (or "mass"). An iron bar is matter — it takes up room and is heavy, as you very well know. Water is matter — it takes up room when you fill a pail with it, and a full pail weighs plenty. The air around you is matter — it takes up lots of room; it may not seem very heavy, yet the earth's atmosphere presses down on every square inch of your body with a weight of almost fifteen pounds.

Matter has three distinct forms. Iron, for instance, is a SOLID. Water is a LIQUID. Air has the form of a GAS.

If you should take iron and divide it again and again until you couldn't divide it any further, every tiny particle would still be iron. A thing that consists of one kind of matter only is called an ELEMENT.

Take water, on the other hand. You will learn to break water up into two kinds of matter — each of them an element. A thing in which two or more elements are combined chemically is called a COMPOUND. In a compound the proportions of the different elements that make it up are always exactly the same.

Air also consists of different kinds of matter, but they are not combined chemically — they are simply mixed together. When you make a MIXTURE, you can mix the ingredients together in any proportions that suit you.

TWO ELEMENTS AND A MIXTURE

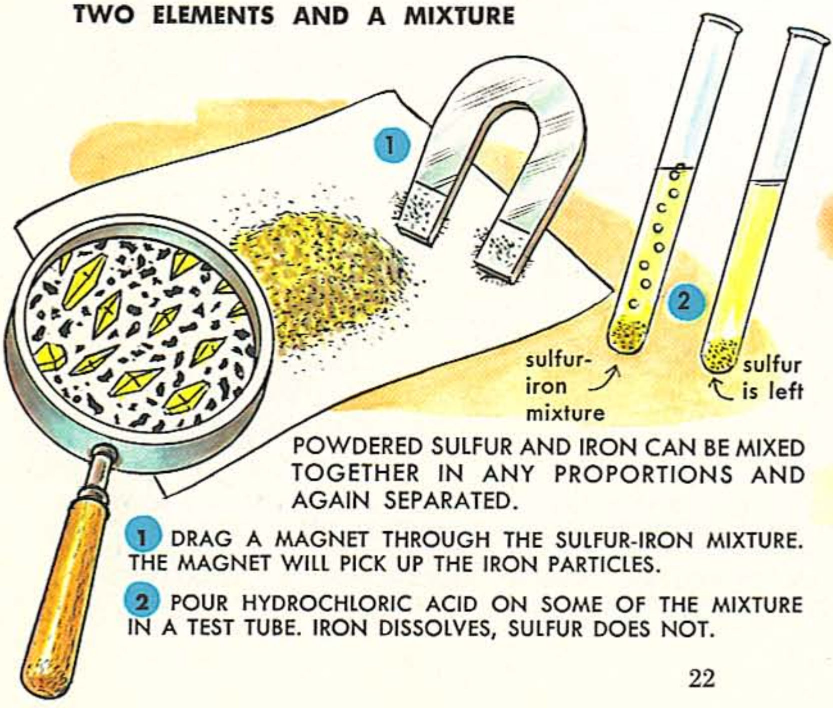

POWDERED SULFUR AND IRON CAN BE MIXED TOGETHER IN ANY PROPORTIONS AND AGAIN SEPARATED.

① DRAG A MAGNET THROUGH THE SULFUR-IRON MIXTURE. THE MAGNET WILL PICK UP THE IRON PARTICLES.

② POUR HYDROCHLORIC ACID ON SOME OF THE MIXTURE IN A TEST TUBE. IRON DISSOLVES, SULFUR DOES NOT.

MAKING A COMPOUND

MIX TOGETHER 2 g OF FLOWERS OF SULFUR AND 3.5 g OF IRON FILINGS. PLACE MIXTURE IN A DAMAGED TEST TUBE. HEAT. SHORTLY A CHEMICAL REACTION TAKES PLACE. THE MIXTURE GLOWS AND BLACK IRON SULFIDE FORMS. THIS CANNOT BE SEPARATED INTO SULFUR AND IRON AS IN PREVIOUS TESTS.

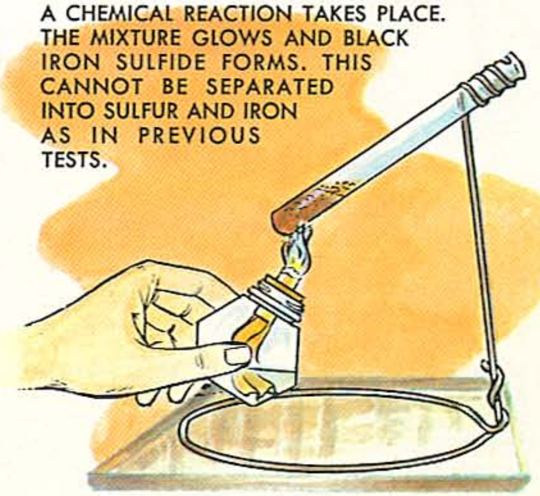

ELEMENTS ARE SUBSTANCES THAT CONSIST OF ONE KIND OF MATTER ONLY. THEY CAN BE DIVIDED INTO METALS, METALLOIDS (METAL-LIKE), NONMETALS. SEVERAL OF THE NONMETALS ARE GASES.

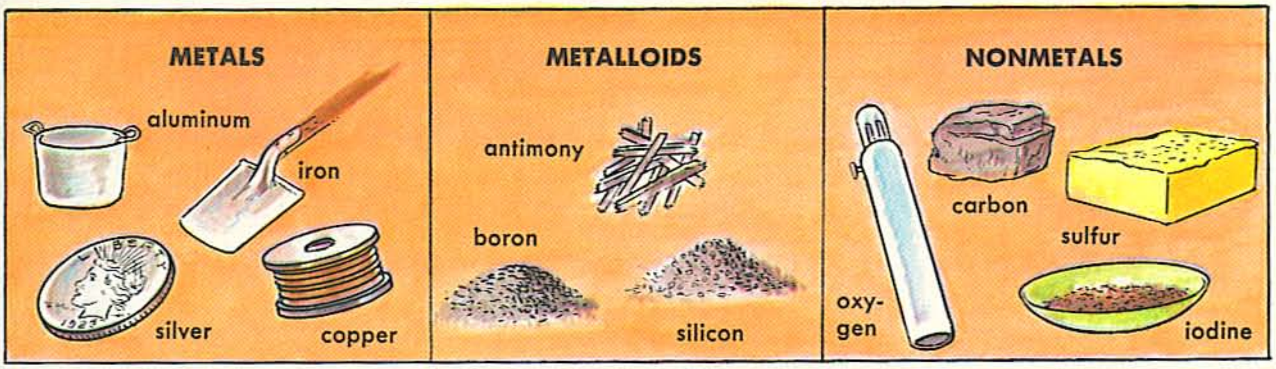
METALS
aluminum
iron
silver
copper
METALLOIDS
antimony
boron
silicon
NONMETALS
carbon
sulfur
oxygen
iodine

COMPOUNDS—INORGANIC. ALL COMPOUNDS CONSIST OF TWO OR MORE ELEMENTS. INORGANIC COMPOUNDS (WITH A FEW EXCEPTIONS) ARE THOSE THAT DO NOT CONTAIN THE ELEMENT CARBON.

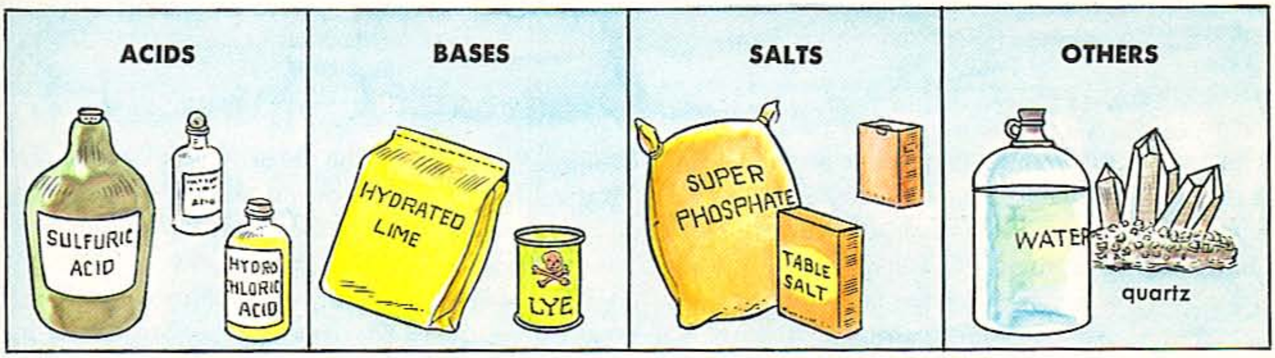
ACIDS
SULFURIC ACID
HYDROCHLORIC ACID
BASES
HYDRATED LIME
LYE
SALTS
SUPER PHOSPHATE
TABLE SALT
OTHERS
WATER
quartz

CARBON COMPOUNDS—ORGANIC. ORIGINALLY, COMPOUNDS MADE BY LIVING THINGS (PLANTS AND ANIMALS) WERE CALLED "ORGANIC." TODAY ORGANIC CHEMISTRY COVERS THE CARBON COMPOUNDS.

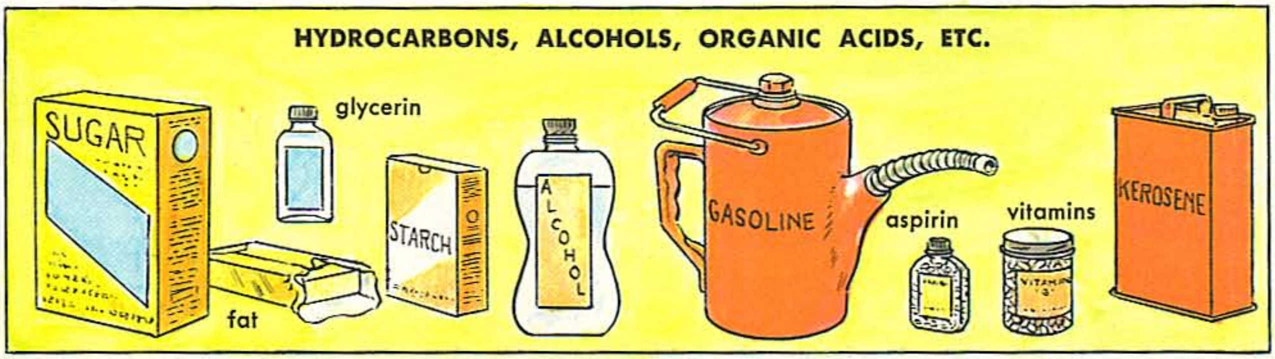
HYDROCARBONS, ALCOHOLS, ORGANIC ACIDS, ETC.
SUGAR
glycerin
fat
STARCH
ALCOHOL
GASOLINE
aspirin
vitamins
KEROSENE

MIXTURES CAN CONSIST OF ELEMENTS OR COMPOUNDS. SOME MIXTURES ARE COARSE. SOME (COLLOIDS) CONTAIN TINY PARTICLES. STILL OTHERS (SOLUTIONS) ARE OF SAME STRUCTURE THROUGHOUT.

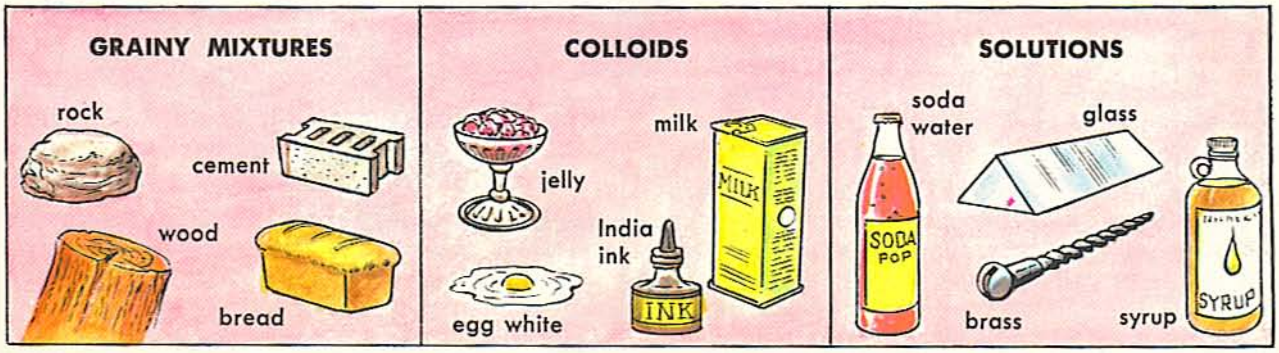
GRAINY MIXTURES
rock
cement
wood
bread
COLLOIDS
milk
MILK
jelly
India ink
INK
egg white
SOLUTIONS
soda water
SODA POP
glass
brass
syrup
SYRUP

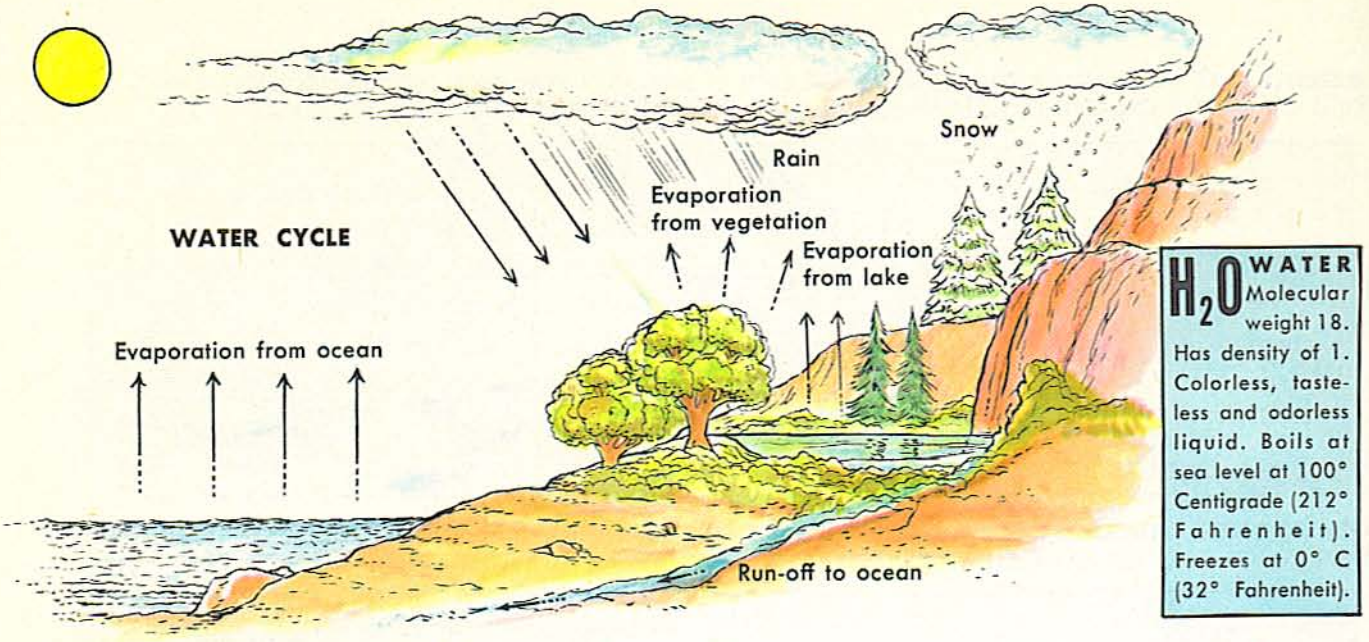

H_2O WATER Molecular weight 18. Has density of 1. Colorless, tasteless and odorless liquid. Boils at sea level at 100° Centigrade (212° Fahrenheit). Freezes at 0° C (32° Fahrenheit).

Water—Our Most Important Compound

Yes, water is the most important of all chemical compounds. Without it, there would be no life — all human beings and all animals would thirst to death, and all plant life would wilt and die.

Fortunately, water is also the most common compound in the world. Almost three quarters of the earth's surface is covered by water. This water is forever traveling. It is turned into invisible vapor by evaporation from oceans and lakes and growing things. When cooled, the vapor forms clouds of tiny water drops. Further cooling makes the drops fall to earth as rain or snow that fill up rivers and lakes and oceans and continue the water cycle.

Chemists use nature's method to produce chemically pure water. They turn ordinary tap water into steam by boiling, then turn the steam back into water by cooling. This process is called distillation and the water is called distilled water.

WATER AS A SOLVENT

THE MOST IMPORTANT FUNCTION OF WATER IN CHEMICAL EXPERIMENTS IS AS A SOLVENT—THAT IS, A LIQUID IN WHICH CHEMICALS MAY BE DISSOLVED. FIND OUT BY AN EASY EXPERIMENT WHETHER HEATING THE WATER HELPS IN DISSOLVING A CHEMICAL.

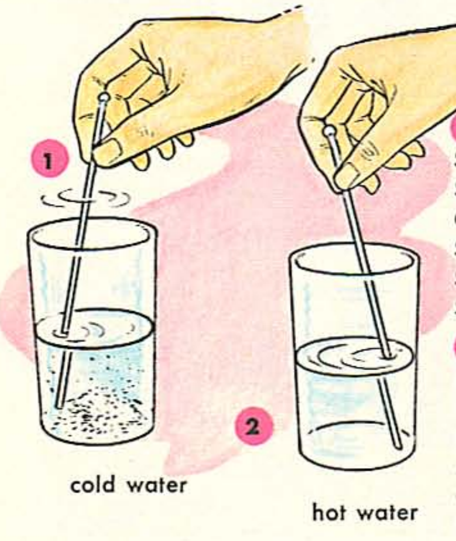

1 DROP 1 TABLESPOON WASHING SODA IN ½ GLASS OF COLD WATER. STIR. PART OF THE SODA DISSOLVES SLOWLY.

2 REPEAT WITH HOT WATER. SODA DISSOLVES QUICKLY. HOT WATER IS USUALLY FASTER THAN COLD FOR PREPARING A SOLUTION.

WATER AS A CATALYST

WATER HELPS BRING ABOUT MANY CHEMICAL REACTIONS WITHOUT ITSELF ENTERING INTO THEM. A SUBSTANCE THAT ACTS THIS WAY IS CALLED A CATALYST.

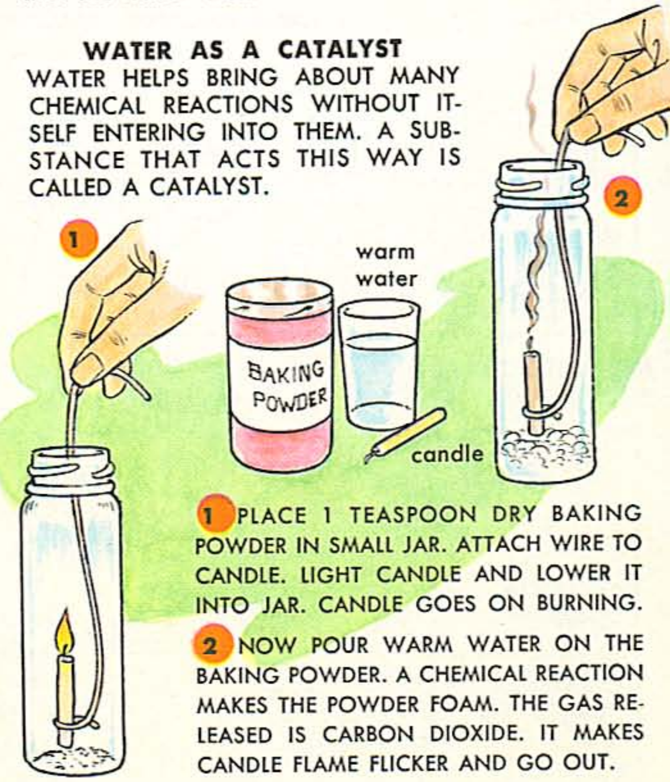

1 PLACE 1 TEASPOON DRY BAKING POWDER IN SMALL JAR. ATTACH WIRE TO CANDLE. LIGHT CANDLE AND LOWER IT INTO JAR. CANDLE GOES ON BURNING.

2 NOW POUR WARM WATER ON THE BAKING POWDER. A CHEMICAL REACTION MAKES THE POWDER FOAM. THE GAS RELEASED IS CARBON DIOXIDE. IT MAKES CANDLE FLAME FLICKER AND GO OUT.

ELECTROLYSIS OF WATER

ELECTRICITY CAN BE USED TO BREAK WATER APART INTO THE TWO ELEMENTS OF WHICH IT CONSISTS —THE GASES HYDROGEN AND OXYGEN.

YOU CAN GET THE REQUIRED ELECTRICITY FROM THREE OR FOUR ORDINARY FLASHLIGHT BATTERIES. YOU WILL ALSO NEED TWO PIECES OF INSULATED COPPER WIRE AND TWO "ELECTRODES" MADE FROM CARBON RODS.

Making Electrodes

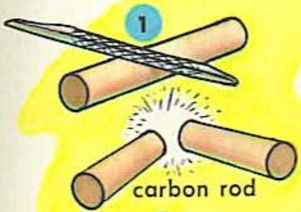

1 SCORE THE MIDDLE OF THE CARBON ROD FROM AN OLD FLASHLIGHT BATTERY, USING A FILE. BREAK THE ROD INTO TWO PIECES.

2 BARE THE WIRE FOR 2" AT EACH END OF TWO 18" LENGTHS OF INSULATED WIRE. TIE ONE BARED WIRE AROUND END OF EACH OF CARBON ROD HALVES.

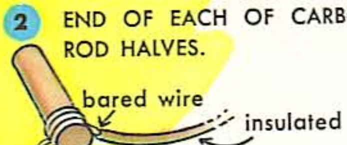

3 BIND ELECTRICIAN'S TAPE FIRMLY AROUND CARBON RODS SO THAT NO WIRE IS EXPOSED.

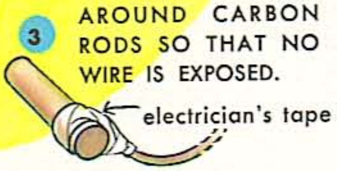

Setting up Electrolysis

WATER IS A POOR CONDUCTOR OF ELECTRICITY—SO YOU DISSOLVE 1 TABLESPOON OF WASHING SODA IN 1 PINT OF WATER AND FILL A WATER GLASS AND TWO TEST TUBES WITH THIS SOLUTION. THEN SET UP THE APPARATUS AS SHOWN AT RIGHT.

MATERIALS FOR EXPERIMENTS

AN ORDINARY FLASHLIGHT BATTERY WILL GIVE YOU MATERIALS YOU NEED FOR EXPERIMENTS ON THIS AND SEVERAL FOLLOWING PAGES.

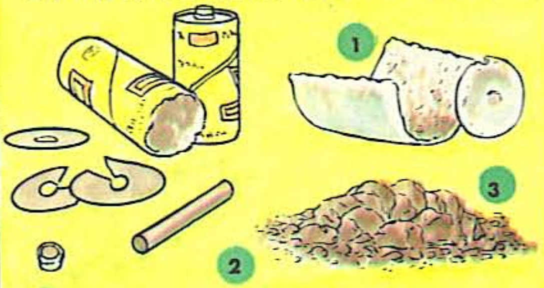

1 OPEN UP BATTERY CASE CAREFULLY WITH A CAN OPENER AND CLEAN THE ZINC CASING.

2 SCRAPE CARBON ROD CLEAN WITH DULL KNIFE.

3 DRY OUT THE MOIST BLACK POWDER, WHICH IS MOSTLY MANGANESE DIOXIDE. STORE IN JAR. THROW REMAINING PARTS OF THE BATTERY AWAY.

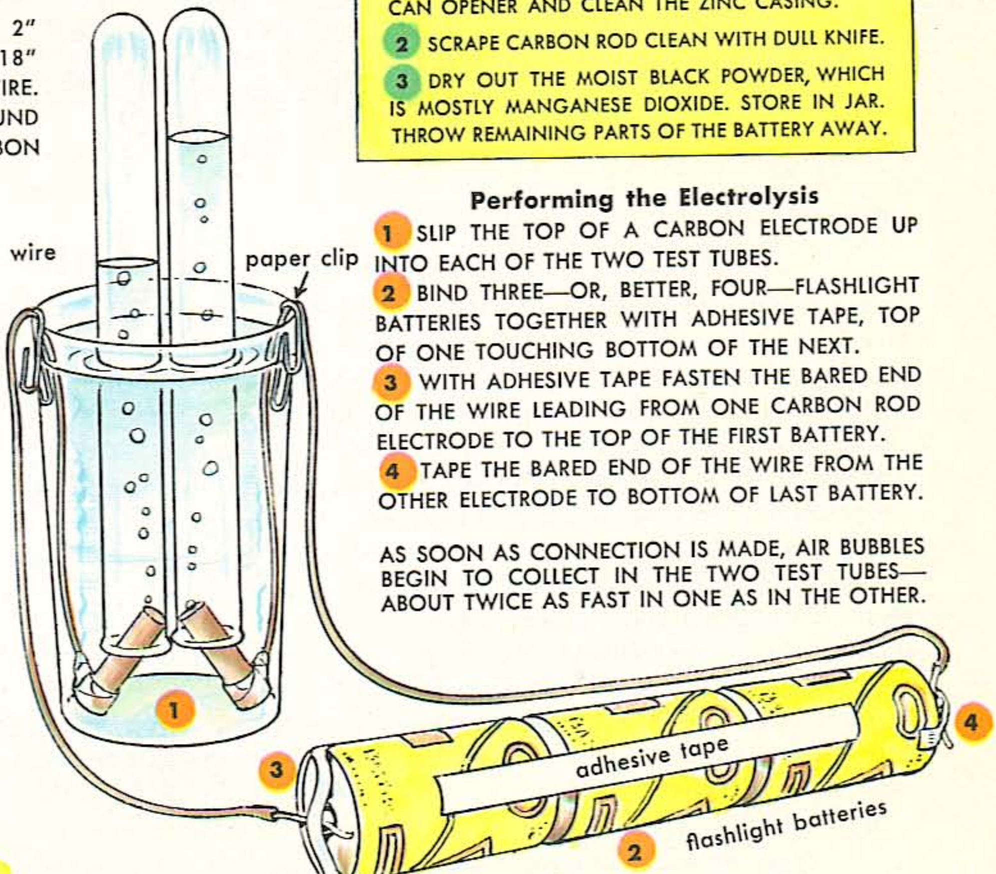

Performing the Electrolysis

1 SLIP THE TOP OF A CARBON ELECTRODE UP INTO EACH OF THE TWO TEST TUBES.

2 BIND THREE—OR, BETTER, FOUR—FLASHLIGHT BATTERIES TOGETHER WITH ADHESIVE TAPE, TOP OF ONE TOUCHING BOTTOM OF THE NEXT.

3 WITH ADHESIVE TAPE FASTEN THE BARED END OF THE WIRE LEADING FROM ONE CARBON ROD ELECTRODE TO THE TOP OF THE FIRST BATTERY.

4 TAPE THE BARED END OF THE WIRE FROM THE OTHER ELECTRODE TO BOTTOM OF LAST BATTERY.

AS SOON AS CONNECTION IS MADE, AIR BUBBLES BEGIN TO COLLECT IN THE TWO TEST TUBES—ABOUT TWICE AS FAST IN ONE AS IN THE OTHER.

TEST FOR HYDROGEN

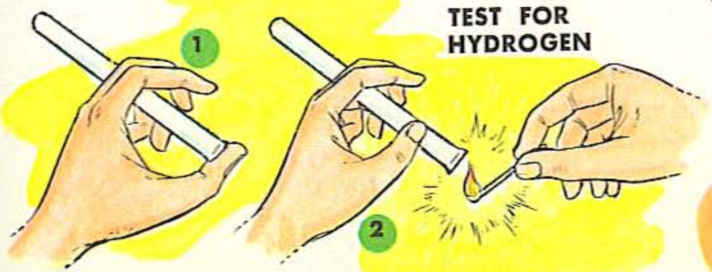

1 WITH YOUR THUMB, CLOSE THE MOUTH OF THE TEST TUBE FIRST FILLED WITH GAS. LIFT THE TUBE OUT OF THE WATER, MOUTH DOWN.

2 BRING LIGHTED MATCH TO THE MOUTH OF THE TUBE. CONTENTS BURN WITH A SOFT "POP!" THIS IS THE TEST FOR HYDROGEN.

TEST FOR OXYGEN

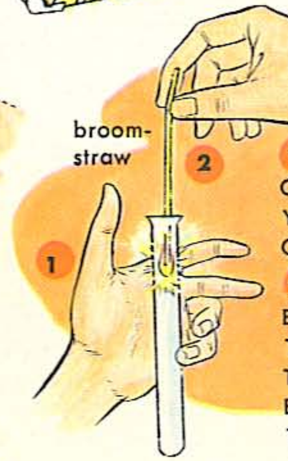

1 WHEN SECOND TUBE IS FULL OF GAS, CLOSE ITS MOUTH WITH YOUR THUMB. LIFT THE TUBE OUT OF THE WATER WITH MOUTH UP.

2 LIGHT A BROOMSTRAW. BLOW OUT THE FLAME. BRING THE GLOWING END DOWN IN THE TEST TUBE. GLOWING EMBER BURSTS INTO BRIGHT FLAME. THIS IS TEST FOR OXYGEN.

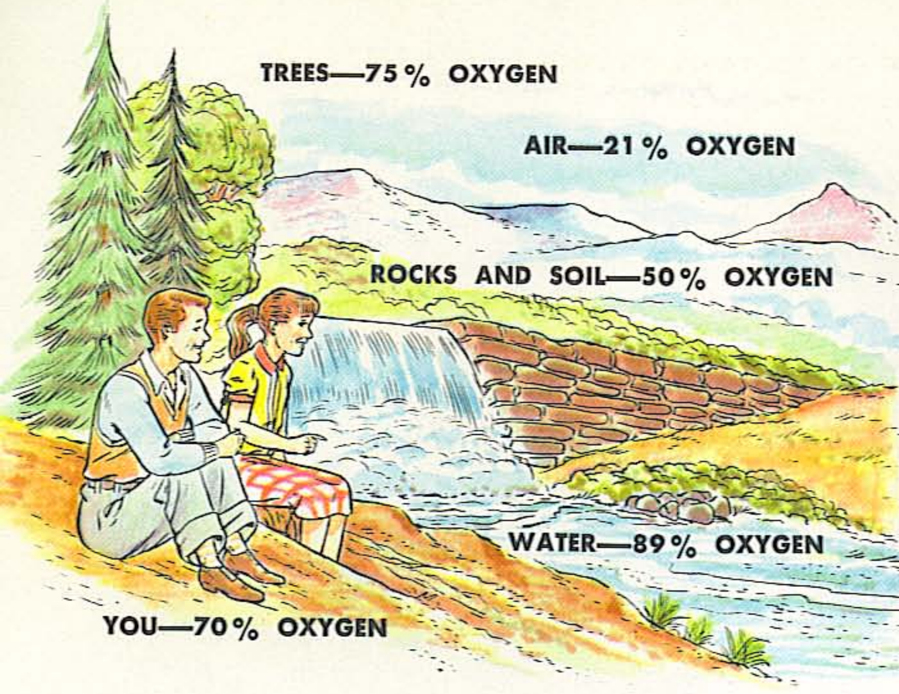

O **OXYGEN** Element 8. Atomic wt. 16. Colorless, odorless gas, supporting combustion (burning). 1.1 weight of air. Slightly soluble in water—3 volumes in 100 volumes at 20° C. Oxygen is the most common element on earth.

Oxygen—The Breath of Life

If you could hold your breath for a few minutes so that no air could get into your lungs, you would die.

For thousands of years, people have known that no human being can live without air. But it was not until Karl Scheele, a Swedish chemist, in 1772, and Joseph Priestley, an Englishman, in 1774, discovered and described oxygen that people knew that it is the oxygen in the air that is important to life.

Both of these scientists discovered that things burn more fiercely in pure oxygen than they do in the mixture of oxygen and other gases called "air."

In the lab, oxygen is produced by driving it out of certain oxygen-containing compounds. A good one to use in the home lab is hydrogen peroxide. You can get it at a drug store in a 3% solution. Hydrogen peroxide is related to water.

Water, as you know, consists of 2 parts of hydrogen to 1 part of oxygen. You could write it: Hydrogen 2 — Oxygen 1. That's pretty much what chemists do — except that they abbreviate the names to initials, use small numbers, and don't bother about the number 1. The formula becomes H_2O.

Hydrogen peroxide contains 2 parts of hydrogen to every 2 parts of oxygen. How would you write it in chemical language?

H_2O_2? You're perfectly right!

H_2O_2 becomes water (H_2O) and gives off oxygen (O) when you throw a catalyst into it. For a catalyst, you can use the manganese dioxide from an old flashlight battery (page 25).

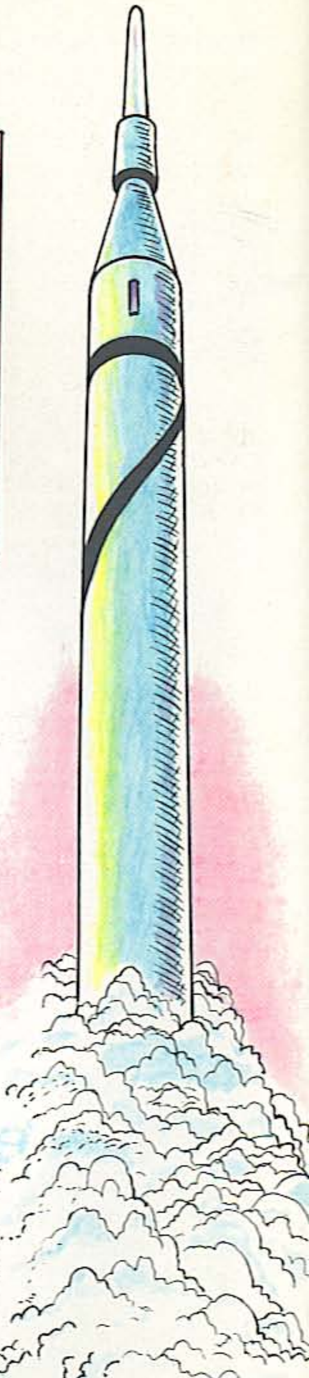

IT'S A LONG STEP FROM THE DISCOVERY OF OXYGEN IN 1772 TO ITS PRESENT-DAY USE IN INDUSTRY AND HOSPITALS, AIRPLANES AND SPACE SHIPS, AND FOR SENDING SATELLITES INTO ORBIT.

THE ATMOSPHERE CONTAINS OXYGEN

1 FASTEN A SMALL CANDLE TO MIDDLE OF PIE PLATE WITH CANDLE DRIPPINGS. FILL PIE PLATE WITH WATER. LIGHT CANDLE. PLACE AN EMPTY JAR OVER CANDLE.

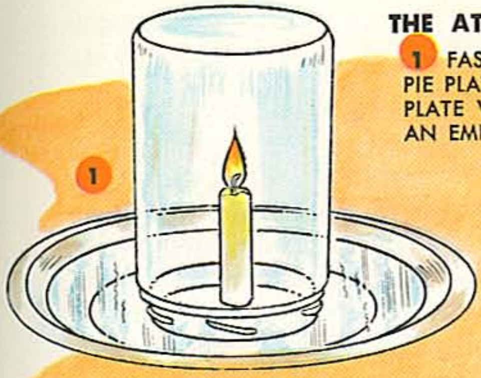

2 A MOMENT LATER, CANDLE GOES OUT. WATER RISES IN JAR TO REPLACE OXYGEN USED.

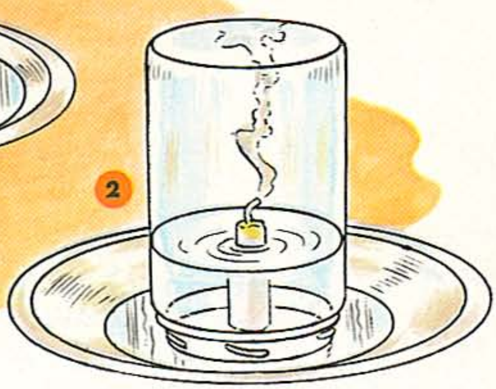

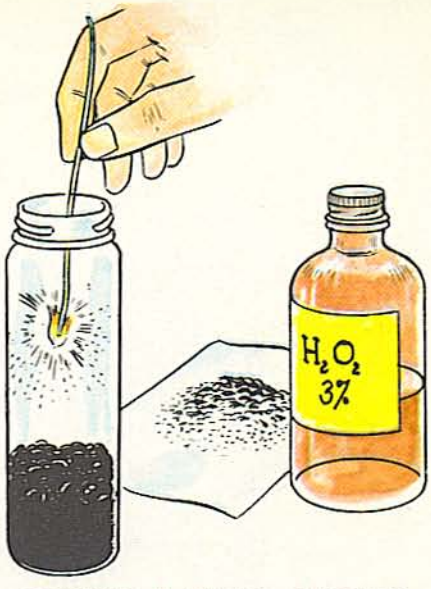

MAKING A SMALL AMOUNT OF OXYGEN

FILL JAR ¼ FULL OF 3% HYDROGEN PEROXIDE. ADD PINCH OF MANGANESE DIOXIDE FROM FLASHLIGHT BATTERY (SEE PAGE 25). TEST FOR OXYGEN WITH GLOWING BROOMSTRAW. EMBER GLOWS BRIGHTER AND MAY BURST INTO FLAME.

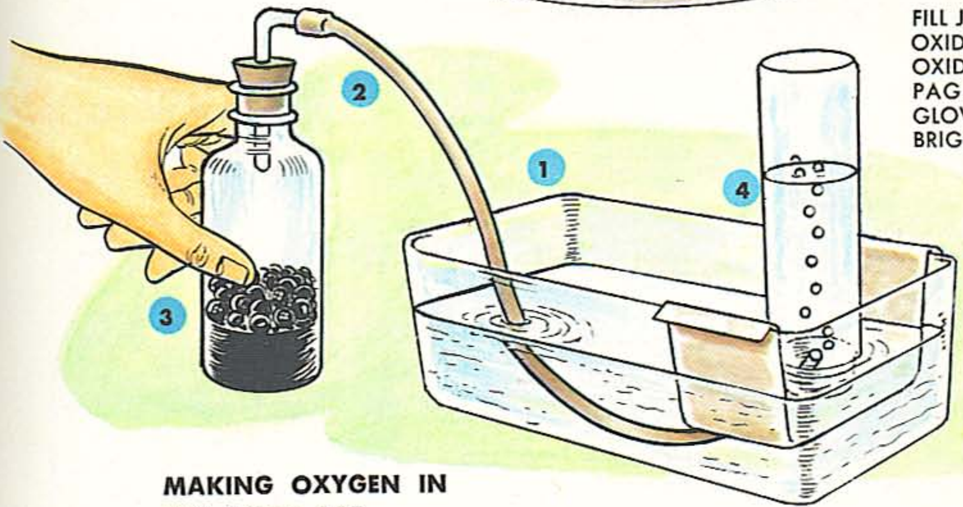

MAKING OXYGEN IN THE HOME LAB

1 TO COLLECT OXYGEN, YOU NEED A "PNEUMATIC TROUGH." THIS IS A DEEP, WATER-FILLED TRAY WITH METAL "BRIDGE."

2 FIT BOTTLE WITH STOPPER WITH L-SHAPED GLASS TUBE AND RUBBER TUBE LONG ENOUGH TO REACH HOLE OF BRIDGE.

3 FILL BOTTLE ¼ FULL OF 3% HYDROGEN PEROXIDE. ADD ⅛ TEASPOON OF MANGANESE DIOXIDE. PUT THE STOPPER IN.

4 FILL JAR WITH WATER AND PLACE IT UPSIDE DOWN ON THE BRIDGE IN SUCH A WAY THAT THE OXYGEN BUBBLES INTO IT AND FILLS IT BY FORCING OUT AND REPLACING THE WATER

5 WHEN JAR IS FULL OF OXYGEN, SLIDE A GLASS PLATE UNDER OPENING (OR PUT STOPPER IN IT). TURN JAR RIGHT SIDE UP—QUICKLY, TO PREVENT THE OXYGEN FROM ESCAPING.

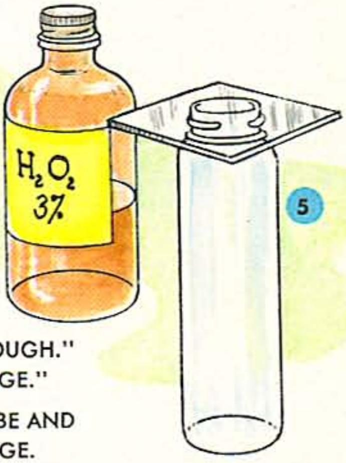

oxygen is slightly heavier than air —so keep mouth of jar up

MANY MATERIALS BURN IN OXYGEN

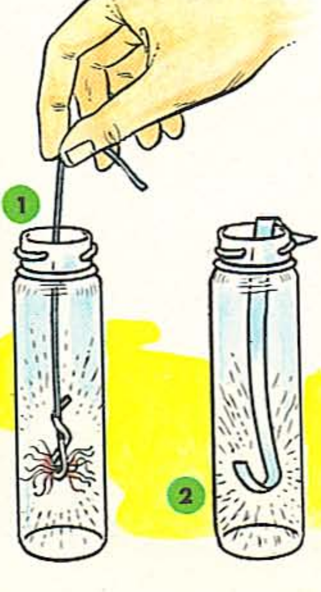

1 ATTACH TUFT OF STEEL WOOL TO WIRE. HEAT TO RED HEAT OVER ALCOHOL BURNER. LOWER INTO JAR OF OXYGEN. IRON BURSTS INTO FLAME.

2 PLACE SMALL PIECE OF SULFUR IN CROOK OF BENT STRIP OF TIN CUT FROM CAN. IGNITE SULFUR WITH MATCH. LOWER INTO JAR OF OXYGEN. SULFUR BURNS WITH A BRILLIANT, BLUE LIGHT.

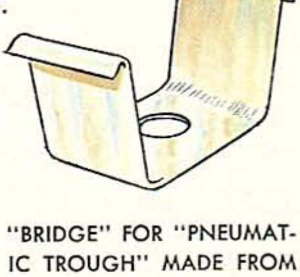

"BRIDGE" FOR "PNEUMATIC TROUGH" MADE FROM 2½" STRIP OF TIN CAN.

Hydrogen—Lightest of All

HYDROGEN is the lightest element in existence—1/14 the weight of air. For this reason one of its early uses was for filling balloons. The first man-carrying gas balloon was sent up by the Frenchman, Jacques Charles, in 1783. The danger of using an explosive gas for this purpose was demonstrated in 1937 in the *Hindenburg* disaster, when the hydrogen-filled Zeppelin dirigible exploded on arriving at Lakehurst, New Jersey, after a trip across the Atlantic Ocean. Thirty-six people lost their lives.

Hydrogen is one of the most important of all the elements. It is found in all living things — your own body is approximately 10 per cent hydrogen. Water, as you know, is part hydrogen. So is the food you eat, the milk you drink, the clothes you wear, and such common, everyday things as gasoline and fuel oil and cooking gas.

In the home lab, you can make hydrogen by adding strips of zinc from a flashlight battery to hydrochloric acid which consists of hydrogen (H) and another gas called chlorine (Cl). The zinc forms a compound ($ZnCl_2$) with the chlorine and sets the hydrogen free (H_2).

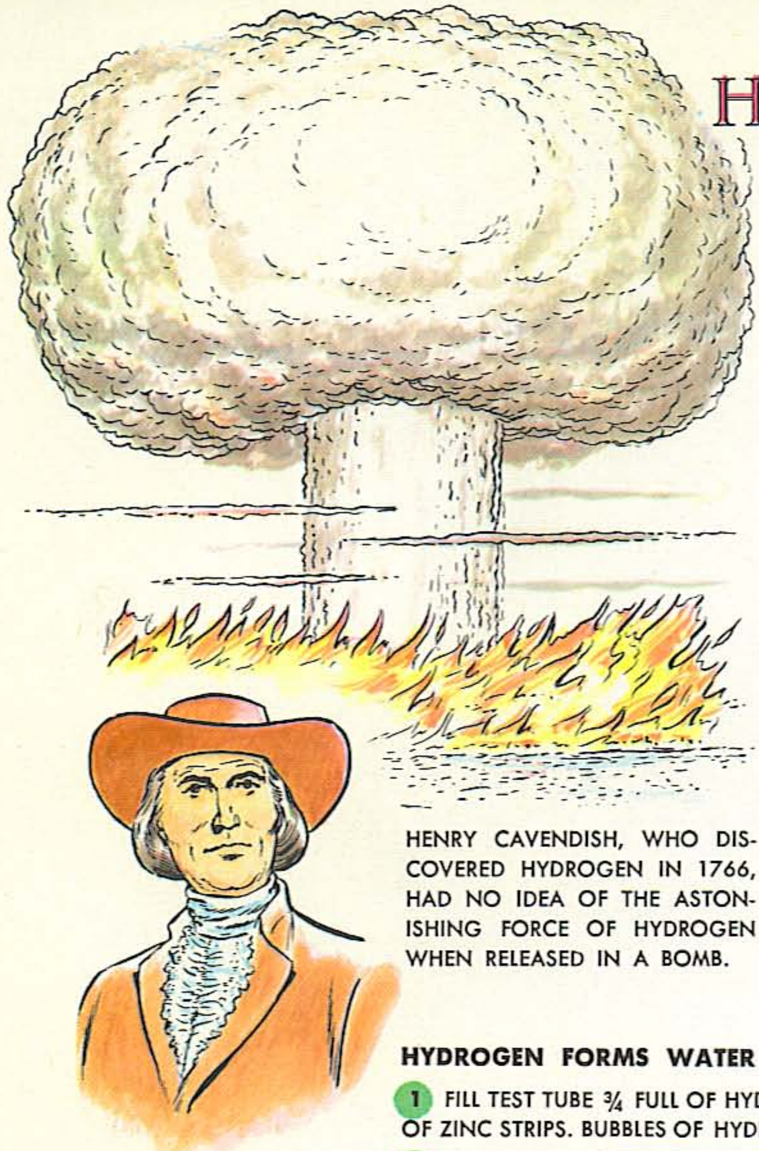

HENRY CAVENDISH, WHO DISCOVERED HYDROGEN IN 1766, HAD NO IDEA OF THE ASTONISHING FORCE OF HYDROGEN WHEN RELEASED IN A BOMB.

HYDROGEN FORMS WATER WHEN IT BURNS

1 FILL TEST TUBE ¾ FULL OF HYDRO-CHLORIC ACID. ADD A COUPLE OF ZINC STRIPS. BUBBLES OF HYDROGEN FORM IMMEDIATELY.

2 CLOSE TEST TUBE WITH RUBBER STOPPER WITH GLASS TUBE DRAWN TO JET POINT. COVER APPARATUS WITH A TOWEL.

3 PLACE EMPTY TEST TUBE OVER GLASS TUBE. AFTER 1 MINUTE, TEST THIS TUBE FOR HYDROGEN WITH LIGHTED MATCH. IF TUBE "BARKS," PUT IT BACK. AFTER ANOTHER MINUTE, TRY AGAIN. WHEN SOFT "POP" TELLS YOU GAS IS PURE, LIGHT JET.

4 HOLD A COLD GLASS OVER HYDROGEN FLAME. DEW COVERING THE INSIDE OF THE GLASS SHOWS THAT WATER IS FORMING.

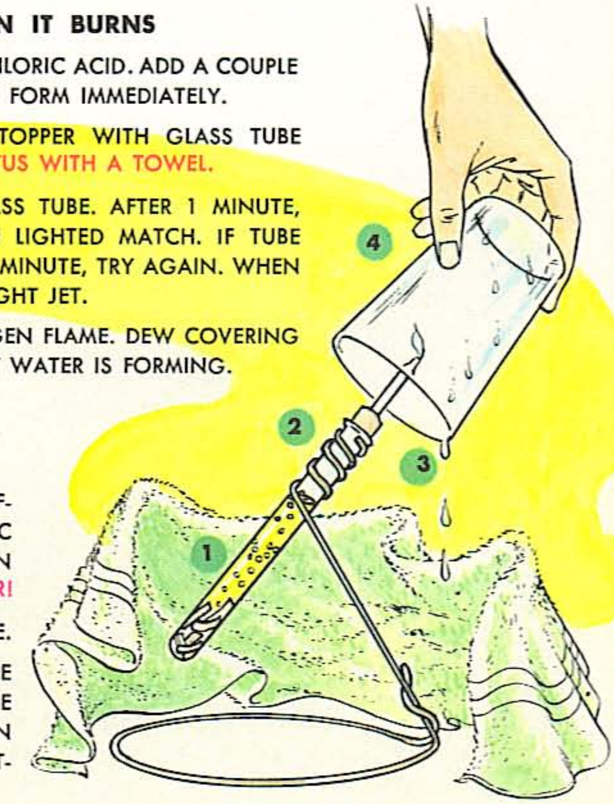

HYDROGEN IS LIGHTEST GAS KNOWN

1 FILL A POP BOTTLE ¼ FULL OF HALF-AND-HALF MIXTURE OF HYDROCHLORIC ACID AND WATER. DROP IN HALF A DOZEN ZINC STRIPS. LET NO FLAME COME NEAR!

2 FIT BALLOON ON MOUTH OF BOTTLE.

3 WHEN BALLOON IS INFLATED, TIE OPENING WITH STRING AND REMOVE FROM BOTTLE. IF PERMITTED, BALLOON WILL RISE TO CEILING INDOORS. OUTDOORS, IT WILL SOAR UP IN THE SKY.

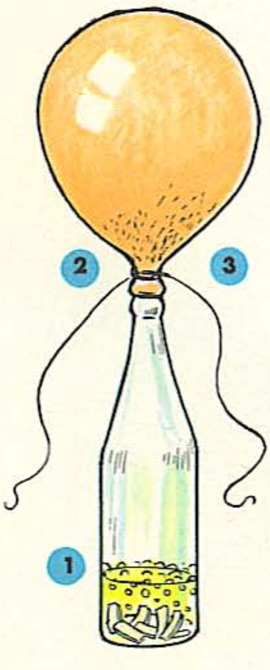

MAKING HYDROGEN IN THE LAB

1 MAKE THE SAFETY GAS GENERATOR DESCRIBED ON PAGE 12.

2 SET UP PNEUMATIC TROUGH AS ON PAGE 27.

3 FILL BOTTLE **A** ⅔ FULL OF MIXTURE OF EQUAL PARTS OF HYDROCHLORIC ACID AND WATER.

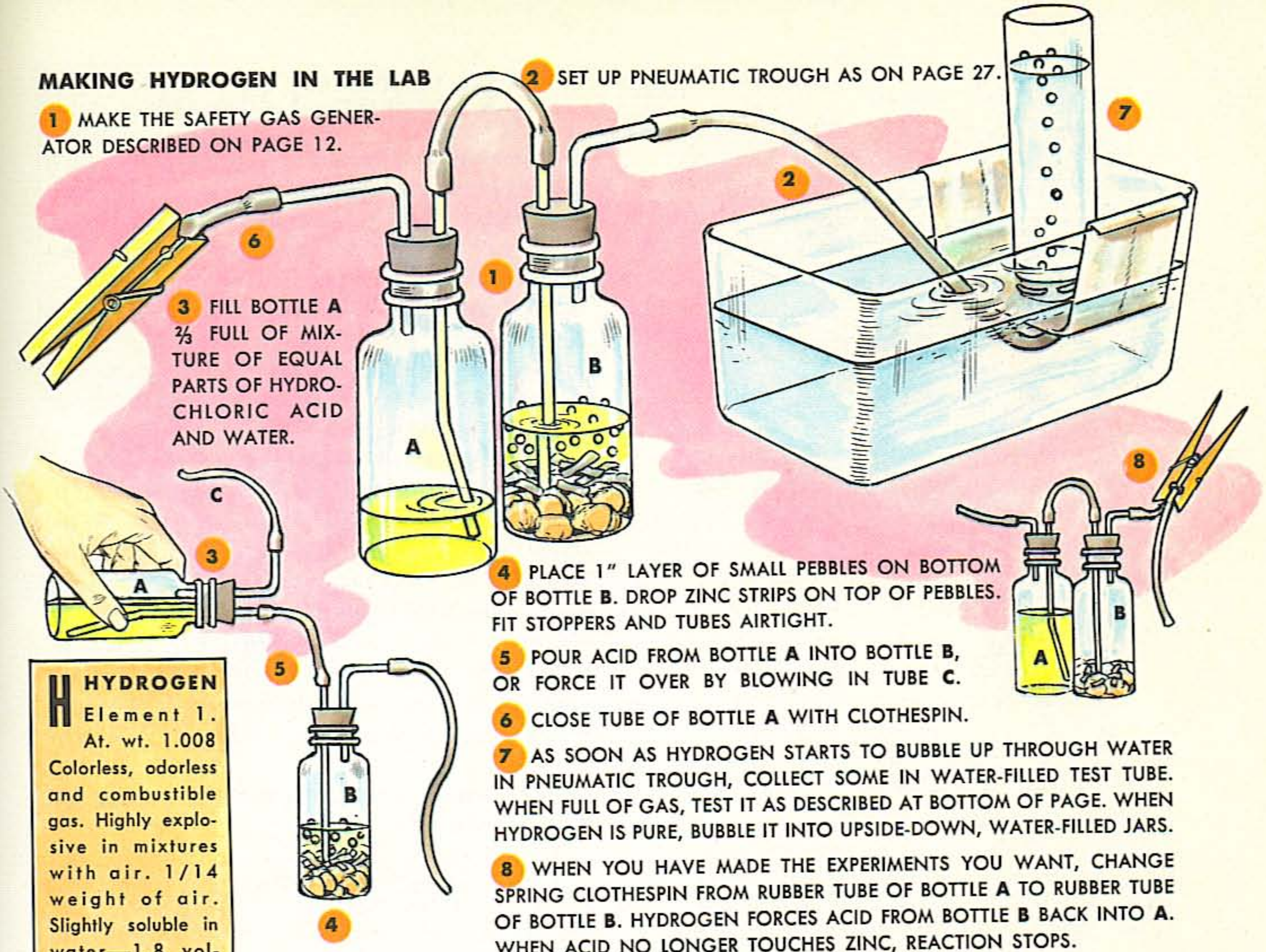

4 PLACE 1" LAYER OF SMALL PEBBLES ON BOTTOM OF BOTTLE **B**. DROP ZINC STRIPS ON TOP OF PEBBLES. FIT STOPPERS AND TUBES AIRTIGHT.

5 POUR ACID FROM BOTTLE **A** INTO BOTTLE **B**, OR FORCE IT OVER BY BLOWING IN TUBE **C**.

6 CLOSE TUBE OF BOTTLE **A** WITH CLOTHESPIN.

7 AS SOON AS HYDROGEN STARTS TO BUBBLE UP THROUGH WATER IN PNEUMATIC TROUGH, COLLECT SOME IN WATER-FILLED TEST TUBE. WHEN FULL OF GAS, TEST IT AS DESCRIBED AT BOTTOM OF PAGE. WHEN HYDROGEN IS PURE, BUBBLE IT INTO UPSIDE-DOWN, WATER-FILLED JARS.

8 WHEN YOU HAVE MADE THE EXPERIMENTS YOU WANT, CHANGE SPRING CLOTHESPIN FROM RUBBER TUBE OF BOTTLE **A** TO RUBBER TUBE OF BOTTLE **B**. HYDROGEN FORCES ACID FROM BOTTLE **B** BACK INTO **A**. WHEN ACID NO LONGER TOUCHES ZINC, REACTION STOPS.

H **HYDROGEN**

Element 1.
At. wt. 1.008
Colorless, odorless and combustible gas. Highly explosive in mixtures with air. 1/14 weight of air. Slightly soluble in water—1.8 volumes in 100 volumes at 20° C.

PLAYING SAFE WITH HYDROGEN

IN MIXTURES WITH AIR, HYDROGEN IS HIGHLY EXPLOSIVE. FOLLOW SAFETY RULES BELOW.

- MAKE ONLY SMALL AMOUNTS OF HYDROGEN IN THE HOME LAB. A 4-OZ. GENERATOR BOTTLE WILL GIVE YOU ALL THE HYDROGEN YOU NEED. MAKE ALL CONNECTIONS AIRTIGHT.
- TEST HYDROGEN FOR PURITY BY COLLECTING A TEST TUBE FULL OF IT AND BRINGING A LIGHTED MATCH TO MOUTH OF TUBE, AS SHOWN ON PAGE 25. HYDROGEN MIXED WITH AIR EXPLODES WITH A SHARP "BARK." PURE HYDROGEN BURNS WITH A QUIET "POP."
- KEEP FLAME AWAY FROM YOUR MAIN GENERATOR BOTTLE.
- IGNITE HYDROGEN ONLY FROM TEST TUBE GENERATOR DESCRIBED ON OPPOSITE PAGE, AND THEN ONLY AFTER YOU HAVE TESTED IT FOR PURITY.

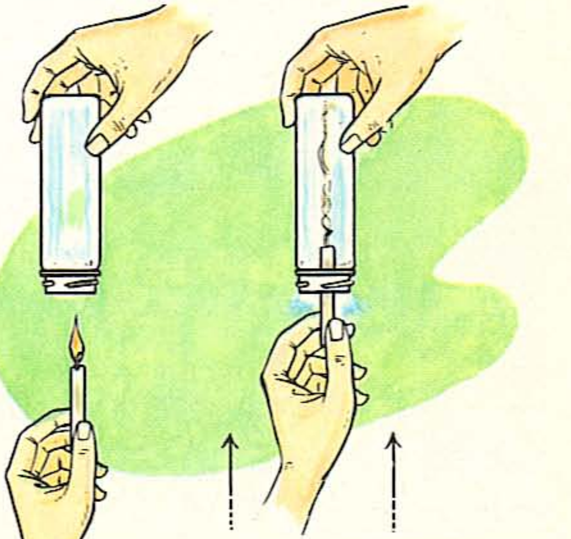

WHEN YOU KNOW FROM TESTING SAMPLES OF GAS COLLECTED IN TEST TUBES THAT HYDROGEN IS PURE, FILL SMALL JAR WITH IT. LIFT JAR OUT OF WATER, MOUTH DOWN. BRING LIGHTED CANDLE UP INTO JAR. HYDROGEN BURNS AT MOUTH OF JAR. CANDLE GOES OUT.

Carbon Dioxide

You HAVE already learned in experimenting with a burning candle that when something containing carbon burns in the air, a gas, carbon dioxide (CO_2), is formed. This is one of the most important gases for human life. The reason is that green plants, in sunlight, are able to take the carbon out of the carbon dioxide in the air and, by combining it with oxygen and hydrogen from water and with various minerals in the soil, produce all the vegetable matter that humans and animals eat.

You cannot see the CO_2 in the air — but you can see it when it has been cooled and compressed into a solid block of "dry ice." When dissolved in water (H_2O), carbon dioxide (CO_2) forms a weak acid (H_2CO_3). You know the taste of this acid from soda water — the bubbles are CO_2 being set free.

Carbonic acid combines with many metals to make "carbonates." You can drive the CO_2 out of most carbonates with the help of a weak acid — even with vinegar, which is diluted acetic acid.

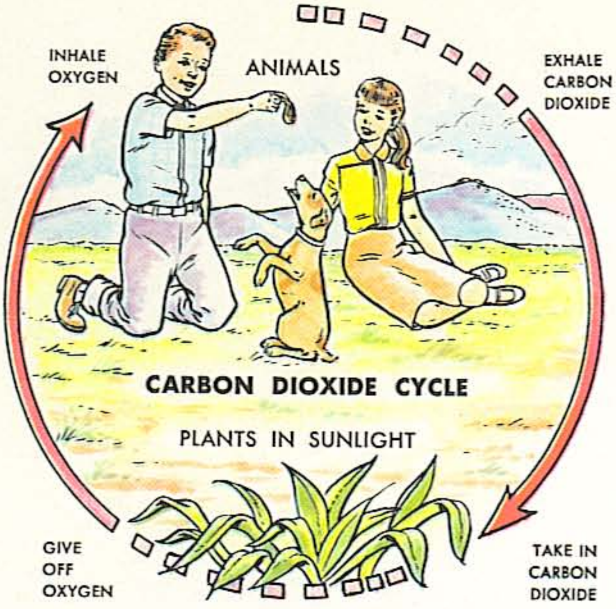

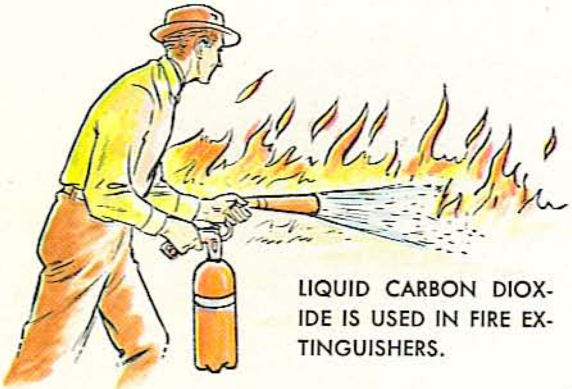

LIQUID CARBON DIOXIDE IS USED IN FIRE EXTINGUISHERS.

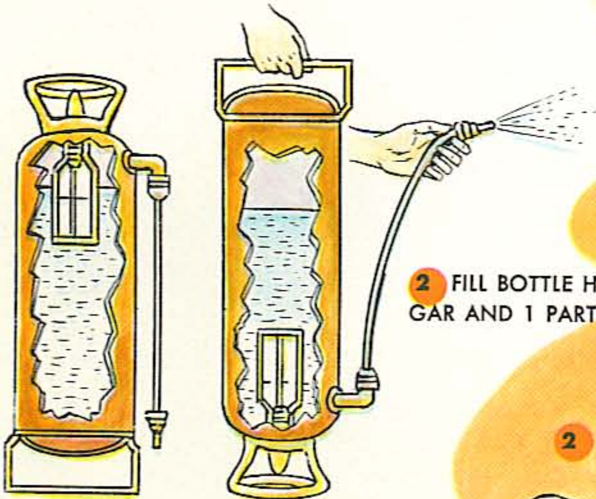

CHEMICAL FIRE EXTINGUISHERS CONTAIN SOLUTION OF BAKING SODA AND A BOTTLE OF SULFURIC ACID. WHEN TURNED UPSIDE DOWN, THE CHEMICALS MIX AND FORM CARBON DIOXIDE WHICH FORCES OUT THE WATER.

MAKING A FIRE EXTINGUISHER MODEL

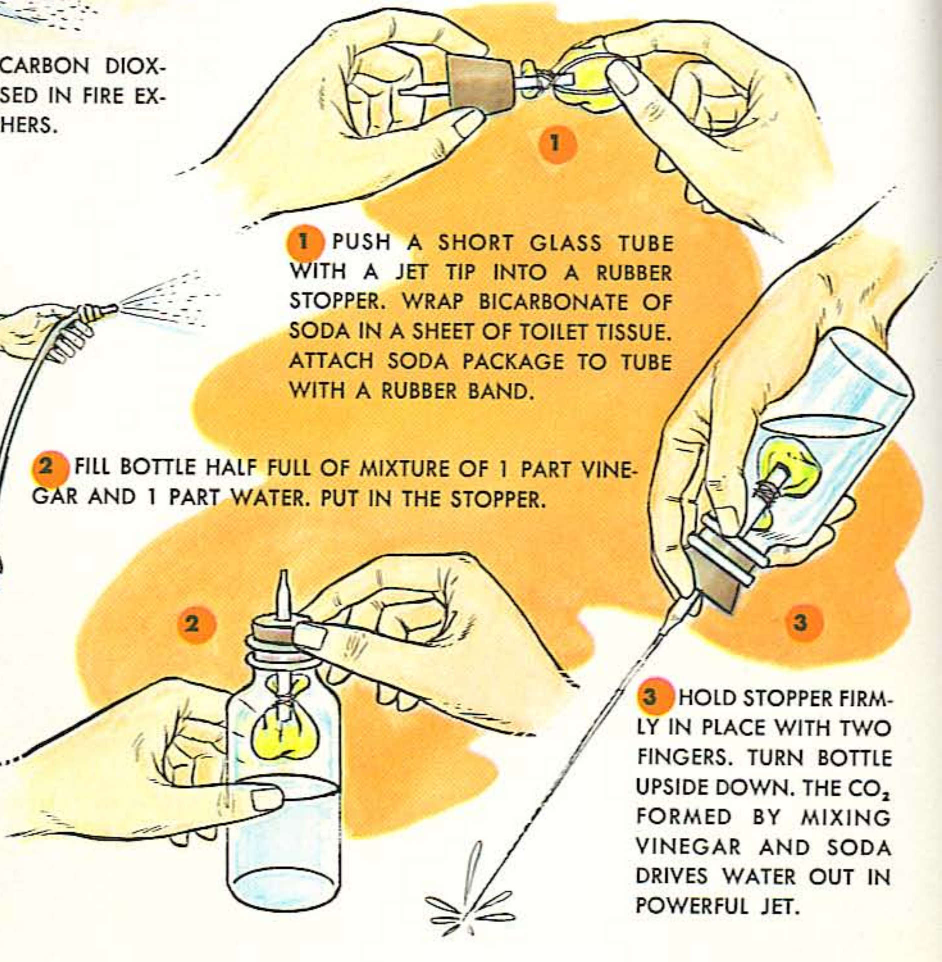

1 PUSH A SHORT GLASS TUBE WITH A JET TIP INTO A RUBBER STOPPER. WRAP BICARBONATE OF SODA IN A SHEET OF TOILET TISSUE. ATTACH SODA PACKAGE TO TUBE WITH A RUBBER BAND.

2 FILL BOTTLE HALF FULL OF MIXTURE OF 1 PART VINEGAR AND 1 PART WATER. PUT IN THE STOPPER.

3 HOLD STOPPER FIRMLY IN PLACE WITH TWO FINGERS. TURN BOTTLE UPSIDE DOWN. THE CO_2 FORMED BY MIXING VINEGAR AND SODA DRIVES WATER OUT IN POWERFUL JET.

CO_2 CARBON DIOXIDE

Compound.

Molecular wt. 44. Colorless, odorless gas. Does not burn. Does not support combustion (burning). 1.529 weight of air. Fairly soluble in water—88 volumes in 100 volumes at 20° C.

FEATURES OF CARBON DIOXIDE

CO_2 IS HEAVIER THAN AIR AND DOES NOT SUPPORT BURNING. YOU CAN PROVE BOTH POINTS:

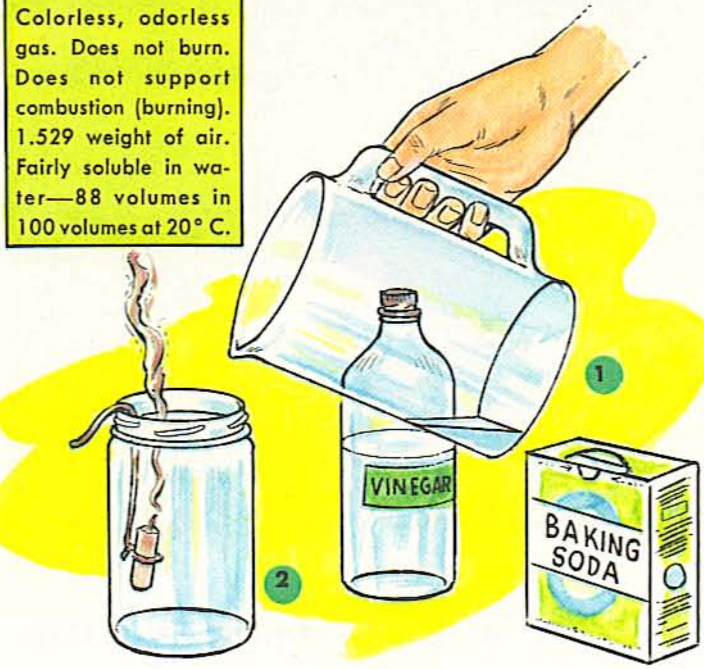

① PLACE 1 TEASPOON OF BAKING SODA IN A PITCHER. POUR A SMALL AMOUNT OF WHITE VINEGAR OVER THE SODA.

② HANG A LIGHTED CANDLE IN A JAR BY A WIRE. POUR THE CARBON DIOXIDE FORMED IN THE PITCHER INTO THE JAR THE WAY YOU WOULD POUR WATER. WHEN THE CARBON DIOXIDE REACHES THE TOP OF THE CANDLE, THE FLAME GOES OUT.

MAKING LIME WATER

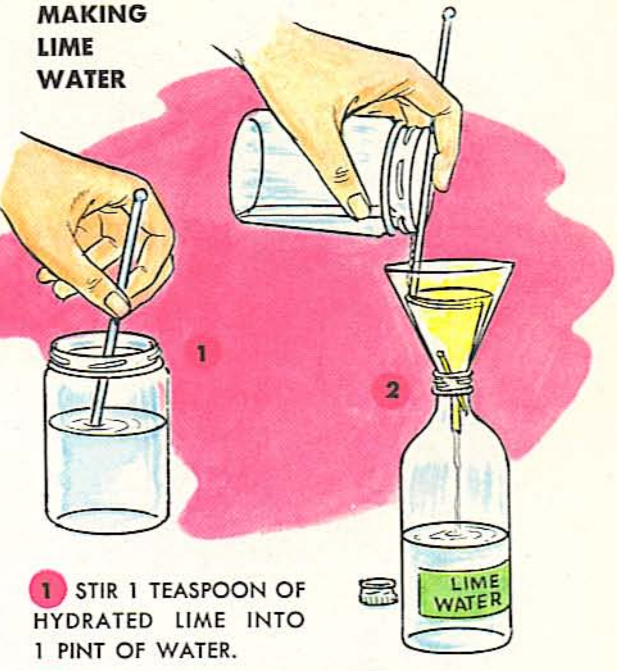

① STIR 1 TEASPOON OF HYDRATED LIME INTO 1 PINT OF WATER.

② LET STAND UNTIL LIME SINKS TO THE BOTTOM. FILTER LIQUID INTO A BOTTLE. CLOSE BOTTLE TIGHTLY.

BURNING PRODUCES CARBON DIOXIDE

HANG BURNING CANDLE IN JAR CONTAINING A FEW ml LIME WATER. COVER TOP WITH A GLASS PLATE. WHEN CANDLE HAS GONE OUT, SHAKE LIME WATER UP WITH THE AIR. MILKINESS PROVES THAT CO_2 HAS BEEN PRODUCED.

MAKING CO_2 FROM MARBLE

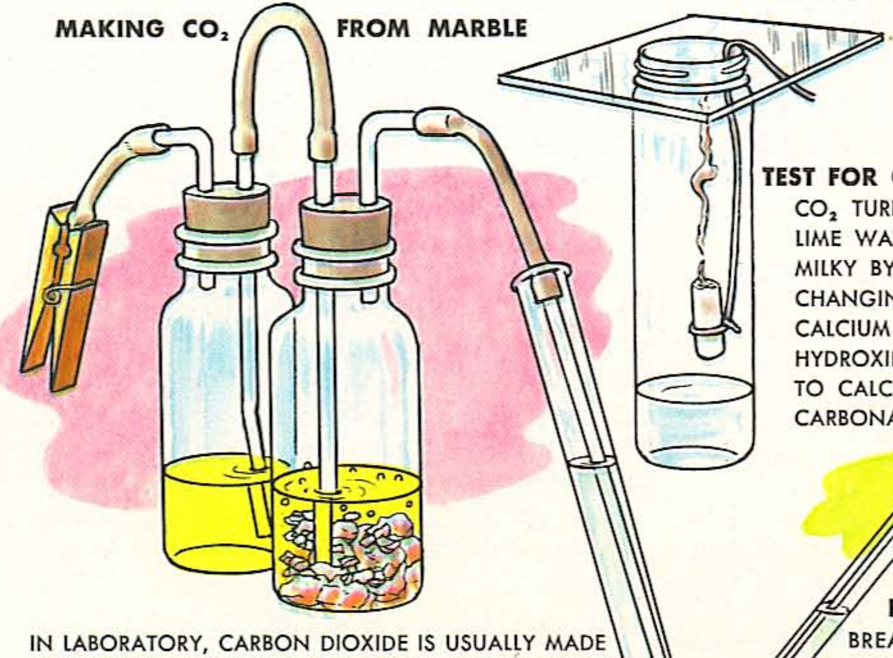

IN LABORATORY, CARBON DIOXIDE IS USUALLY MADE FROM MARBLE CHIPS AND DILUTED HYDROCHLORIC ACID (1 PART ACID TO 1 PART WATER). PLACE CHIPS ON TOP OF PEBBLES IN YOUR GAS GENERATOR, THEN PROCEED AS FOR MAKING HYDROGEN (PAGE 29).

TEST FOR CO_2

CO_2 TURNS LIME WATER MILKY BY CHANGING CALCIUM HYDROXIDE TO CALCIUM CARBONATE.

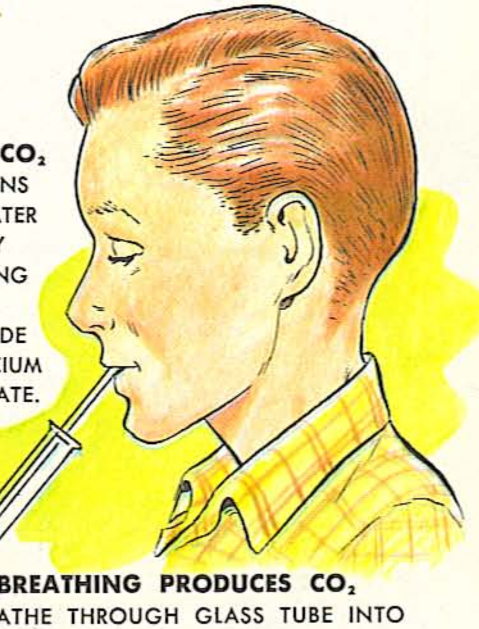

BREATHING PRODUCES CO_2

BREATHE THROUGH GLASS TUBE INTO LIME WATER IN TEST TUBE. LIME WATER GETS MILKY. THIS SHOWS THAT BREATHING IS A BURNING PROCESS.

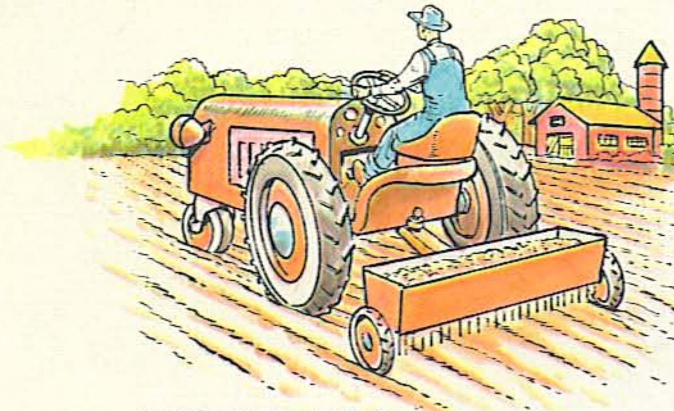

NITROGEN GOES INTO BOTH FERTILIZERS AND EXPLOSIVES.

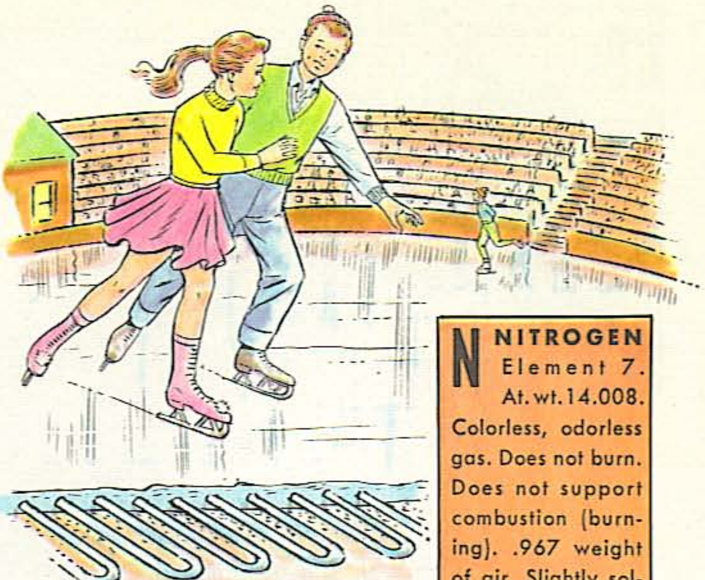

DRY AMMONIA GAS IS USED IN THE LARGE-SCALE PRODUCTION OF ICE.

N NITROGEN Element 7. At. wt. 14.008. Colorless, odorless gas. Does not burn. Does not support combustion (burning). .967 weight of air. Slightly soluble in water—1.5 volumes in 100 vols. at 20°C.

Nitrogen and Its Compounds

When you burn anything in the air, only about one-fifth of the air goes into chemical combination with what you are burning. The rest (except for a small fraction) does not enter into the process. It is a gas called nitrogen (N) — the most abundant free element on earth.

Nitrogen is what you might call a "lazy" element. It does not help in burning nor does it burn if you try to ignite it. It is only at high temperatures and under great pressures that a chemist can make nitrogen combine with another element, hydrogen, to form ammonia gas (NH_3), from which other nitrogen compounds can be made.

Yet, in nature, tiny bacteria on the roots of certain plants can take nitrogen from the air and make it combine with oxygen and minerals in the soil into "nitrates." And that is of tremendous importance to all of us — for all plants need nitrates if they are to thrive. If plants do not get nitrates naturally, the farmer must add them to his soil in the form of some kind of fertilizer.

You will not have much satisfaction out of working with nitrogen itself, but you will find it interesting to deal with some of its compounds — especially with ammonia gas (NH_3). You will also want to have a look at one of the half dozen combinations nitrogen makes with oxygen, the brown gas called nitrogen dioxide (NO_2).

NITROGEN FROM THE ATMOSPHERE

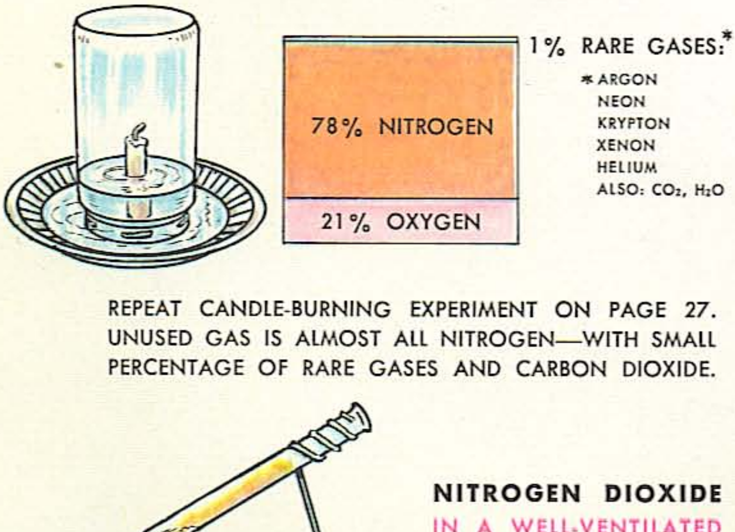

REPEAT CANDLE-BURNING EXPERIMENT ON PAGE 27. UNUSED GAS IS ALMOST ALL NITROGEN—WITH SMALL PERCENTAGE OF RARE GASES AND CARBON DIOXIDE.

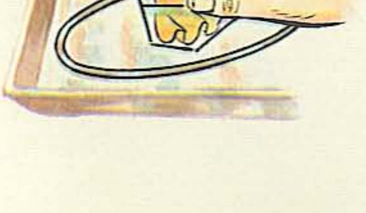

NITROGEN DIOXIDE

IN A WELL-VENTILATED ROOM, HEAT EQUAL AMOUNTS OF SALTPETER AND SODIUM BISULFATE IN DRY TEST TUBE. IN A MOMENT, A BROWN GAS FORMS. IT IS NITROGEN DIOXIDE. DO NOT INHALE —GAS IS VERY IRRITATING.

NH_3 AMMONIA Compound. Molecular weight 17. Colorless gas with strong, penetrating odor. .596 weight of air. Highly soluble in water—70,000 vols. in 100 vols. at 20°C.

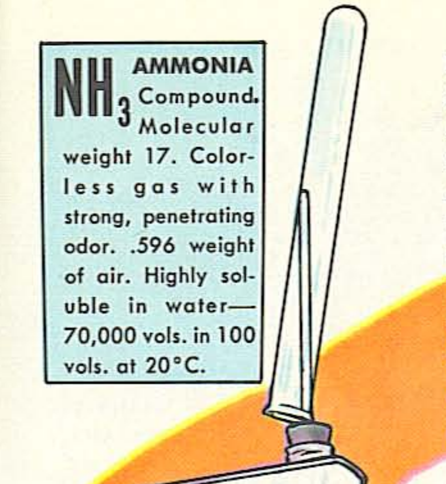

PRODUCING AMMONIA

SIMPLEST WAY OF PRODUCING AMMONIA IS TO GET IT FROM ITS SOLUTION AS HOUSEHOLD AMMONIA.

FILL PINT CAN ONE QUARTER FULL OF HOUSEHOLD AMMONIA. FIT STOPPER WITH 6" GLASS TUBE IN OPENING. PLACE TEST TUBE OVER GLASS TUBE. HEAT CAN OVER LOW FLAME. TEST TUBE IS FULL OF AMMONIA WHEN MOIST, RED LITMUS PAPER HELD AT ITS MOUTH TURNS BLUE.

SOLUBILITY OF AMMONIA

REMOVE A FILLED TEST TUBE FROM GAS GENERATOR CAN, MOUTH DOWN. CLOSE MOUTH OF TUBE WITH THUMB. OPEN TUBE UNDER WATER. AMMONIA DISSOLVES EASILY, WATER RUSHES IN AND FILLS TUBE.

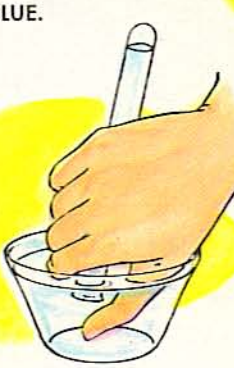

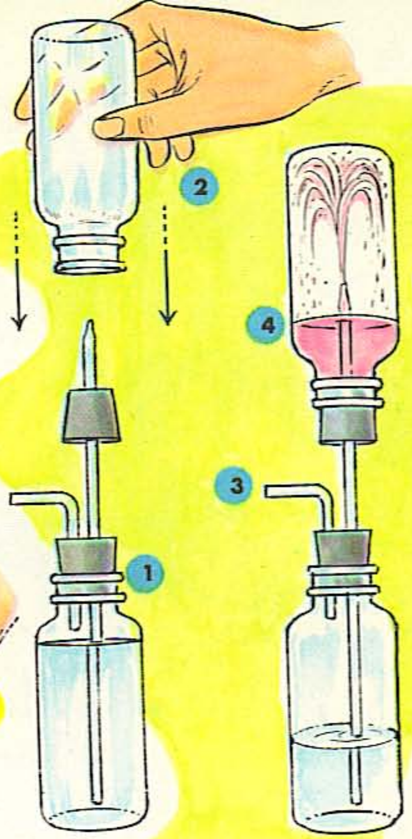

THE AMMONIA FOUNTAIN

AMMONIA'S EXTRAORDINARY SOLUBILITY CAN BE SHOWN IN A SPECTACULAR DEMONSTRATION.

① MAKE UP APPARATUS AS SHOWN IN ILLUSTRATION. FILL IT WITH WATER. ADD 5 DROPS OF PHENOLPHTHALEIN SOLUTION.

② FILL DRY, EMPTY BOTTLE WITH AMMONIA FROM GENERATOR CAN. KEEPING BOTTLE UPSIDE DOWN, PLACE IT FIRMLY ON TOP STOPPER OF APPARATUS.

③ BLOW INTO L-SHAPED GLASS TUBE TO DRIVE A FEW DROPS OF WATER UP INTO THE UPPER BOTTLE.

④ SUDDENLY, WATER SPURTS FROM LOWER BOTTLE UP INTO UPPER BOTTLE IN A FOUNTAIN THAT TURNS PINK AS AMMONIA REACTS ON PHENOLPHTHALEIN.

MOIST, RED LITMUS PAPER TURNS BLUE IN AMMONIA.

MAKING AMMONIA FROM SAL AMMONIAC

① ON A PIECE OF PAPER, MIX 1 PART OF SAL AMMONIAC WITH 2 PARTS OF HYDRATED LIME. ADD A FEW DROPS OF WATER. DROP MIXTURE INTO A TEST TUBE. PROVIDE TUBE WITH STOPPER AND L-SHAPED GLASS TUBE. THEN HEAT OVER LOW FLAME.

② COLLECT AMMONIA IN DRY TEST TUBE. TEST IT WITH LITMUS PAPER AND FOR SOLUBILITY.

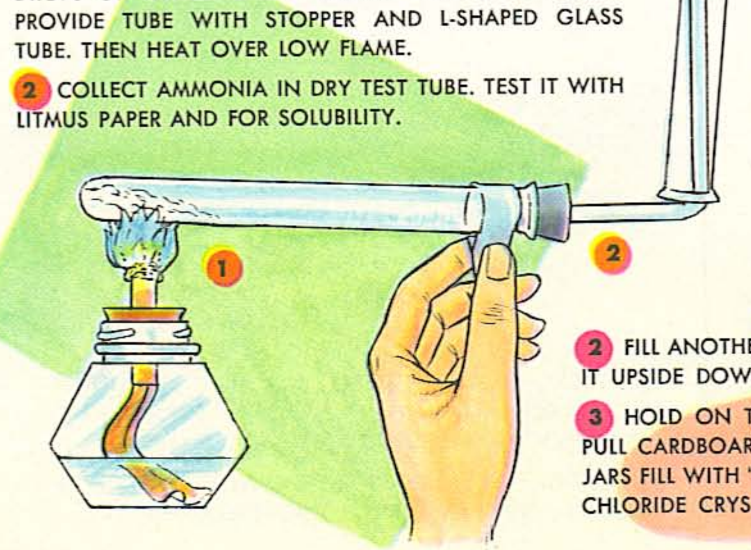

THE WHITE SMOKE MYSTERY

① MOISTEN INSIDE OF JAR WITH SMALL AMOUNT OF HYDROCHLORIC ACID. POUR EXCESS ACID BACK INTO ITS BOTTLE. COVER JAR WITH SQUARE OF CARDBOARD.

② FILL ANOTHER JAR WITH AMMONIA. PLACE IT UPSIDE DOWN ON CARDBOARD.

③ HOLD ON TO AMMONIA-FILLED JAR AND PULL CARDBOARD AWAY. IMMEDIATELY, BOTH JARS FILL WITH "SMOKE" OF TINY AMMONIUM CHLORIDE CRYSTALS.

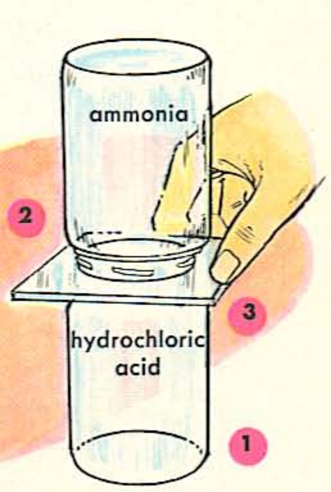

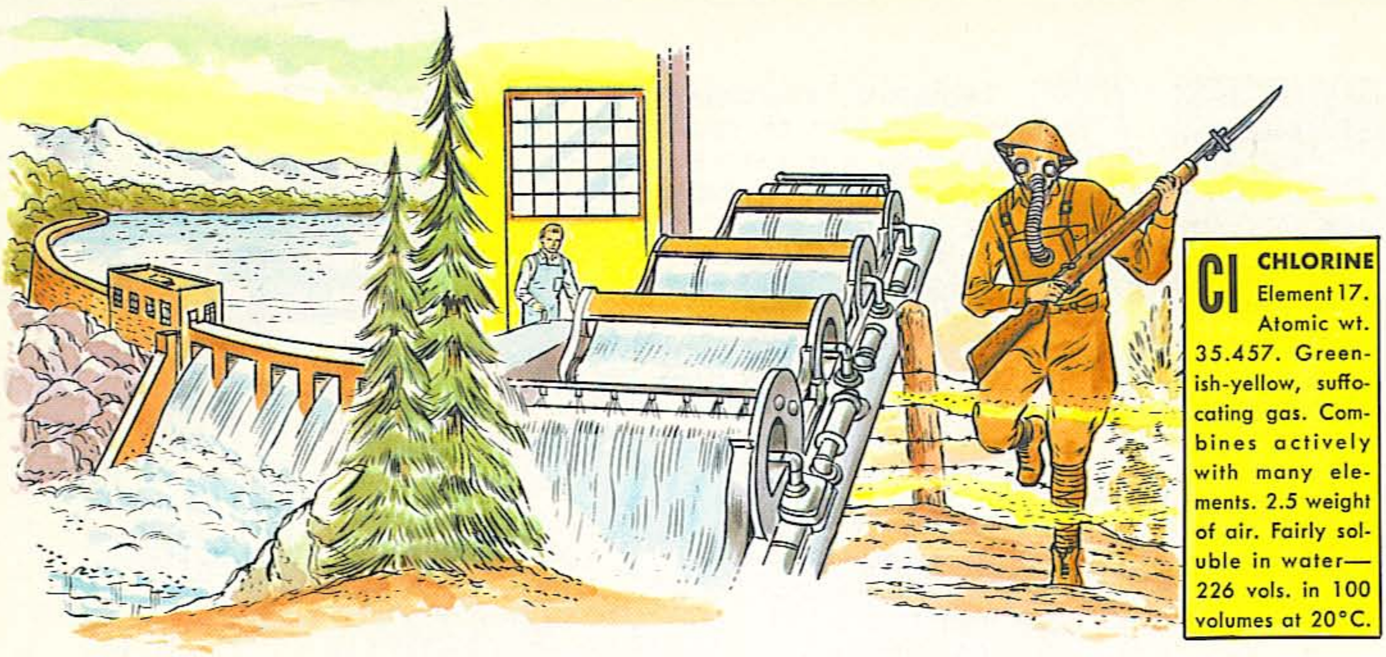

Cl **CHLORINE** Element 17. Atomic wt. 35.457. Greenish-yellow, suffocating gas. Combines actively with many elements. 2.5 weight of air. Fairly soluble in water—226 vols. in 100 volumes at 20°C.

Chlorine—Friend and Foe

CHLORINE IS a gas of great importance. We wouldn't be certain of safe drinking water in our cities if it weren't for chlorine — a small amount of it in the water kills the dangerous germs that may lurk in it. Chlorine is also used extensively in bleaching.

Chlorine is a friendly gas when it is used correctly. But it is dangerous when used improperly because it affects the lungs. As a "poison gas" it caused many casualties in World War I.

You can produce chlorine as a greenish-yellow gas by driving it out of one of its compounds — hydrochloric acid (HCl), which consists of hydrogen (H) and chlorine (Cl), or a common laundry bleach ("Clorox" or others), which is a solution of sodium hypochlorite (NaClO).

Have a bottle of diluted household ammonia (90% water, 10% household ammonia) on hand. Sniff this if you get too strong a whiff of chlorine.

NOTE: Perform these experiments out-of-doors or before an open window. Be careful not to breathe fumes.

MAKING TEST PAPER FOR CHLORINE

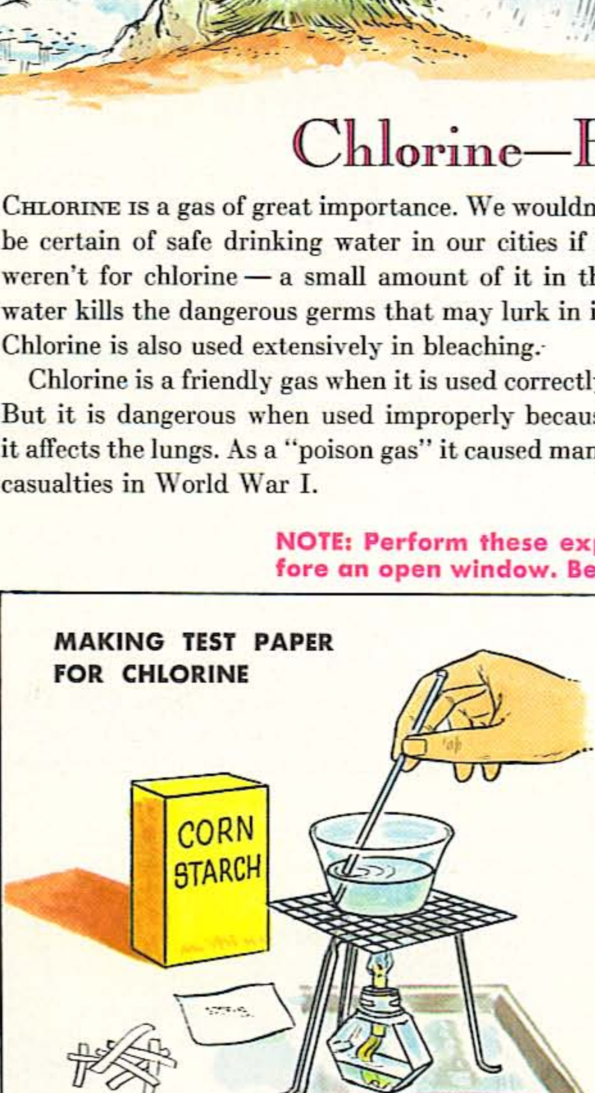

MIX .5 g (1/4 TEASPOON) STARCH WITH 30 ml WATER. BRING TO BOIL. DISSOLVE IN MIXTURE A SMALL AMOUNT OF POTASSIUM IODIDE (AS MUCH AS TWO GRAINS OF RICE). DIP STRIPS OF FILTER PAPER IN MIXTURE; THEN DRY THEM.

CHLORINE FROM HYDROCHLORIC ACID

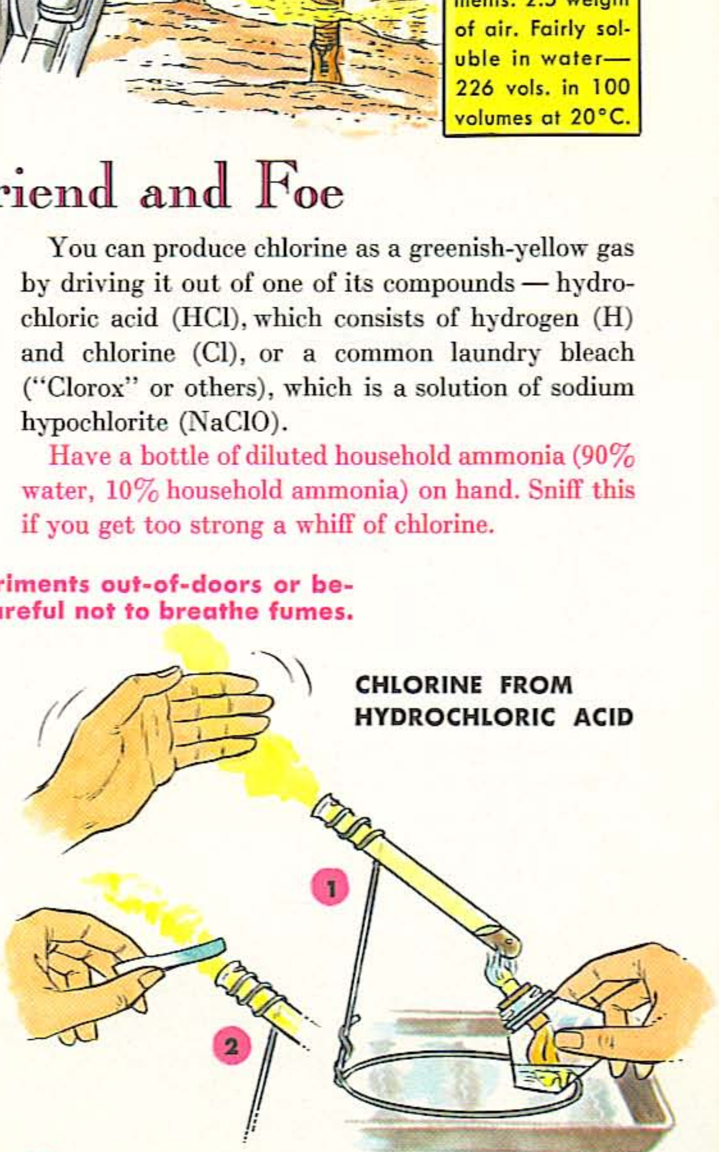

1 Put .5 g (1/8 TEASPOON) MANGANESE DIOXIDE INTO TEST TUBE. ADD 3 ml (1/8 TEST TUBE) UNDILUTED HYDROCHLORIC ACID. HEAT GENTLY. CHLORINE FORMS. WAFT A LITTLE CAREFULLY TOWARD YOU FOR A SNIFF.

2 TEST GAS BY HOLDING MOISTENED STARCH-IODIDE PAPER AT MOUTH OF TUBE. PAPER TURNS BLUE.

MAKING CHLORINE IN THE HOME LAB

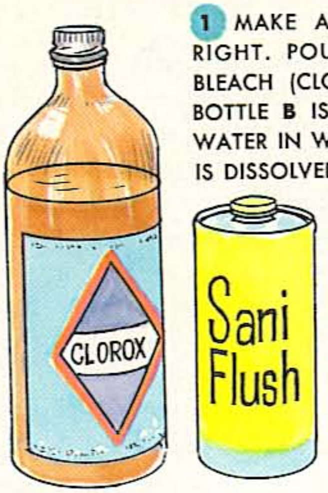

1 MAKE APPARATUS SHOWN AT RIGHT. POUR 1 INCH OF LIQUID BLEACH (CLOROX) INTO BOTTLE **A**. BOTTLE **B** IS EMPTY. BOTTLE **C** HAS WATER IN WHICH ½ TEASPOON LYE IS DISSOLVED.

2 TAKE STOPPER OUT OF BOTTLE **A**. DROP IN ½ TEASPOON SODIUM BISULFATE (SANI-FLUSH). REPLACE THE STOPPER.

3 CHLORINE GAS FORMS AND FILLS **B**.

4 LYE WATER IN BOTTLE **C** ABSORBS EXCESS OF CHLORINE GAS.

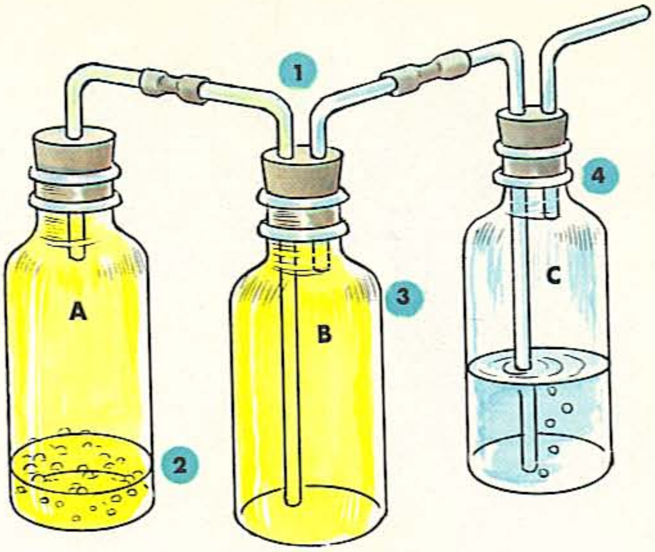

WHEN REACTION SLOWS ADD MORE SODIUM BISULFATE

EXPERIMENTS WITH CHLORINE

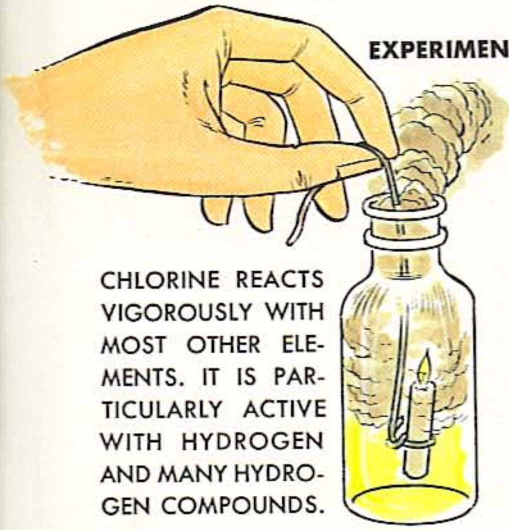

CHLORINE REACTS VIGOROUSLY WITH MOST OTHER ELEMENTS. IT IS PARTICULARLY ACTIVE WITH HYDROGEN AND MANY HYDROGEN COMPOUNDS.

LOWER A BURNING CANDLE INTO A BOTTLE OF CHLORINE GAS. A DENSE SMOKE OF CARBON IS FORMED. THE CHLORINE COMBINES WITH THE HYDROGEN OF THE CANDLE AND SETS THE CARBON IN IT FREE AS SOOT.

CHLORINE WILL COMBINE DIRECTLY WITH SEVERAL METALS. IRON ACTUALLY BURNS IN CHLORINE GAS!

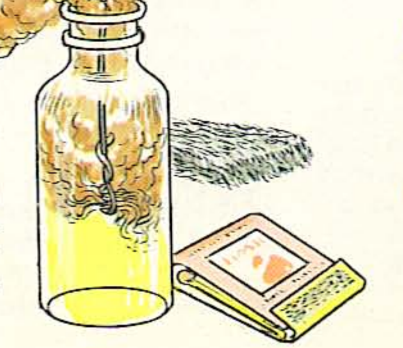

FASTEN A SMALL WAD OF STEEL WOOL TO A PIECE OF WIRE. HEAT IT WITH A MATCH AND LOWER IT INTO CHLORINE-FILLED BOTTLE. A HEAVY BROWN SMOKE OF IRON CHLORIDE POURS OUT.

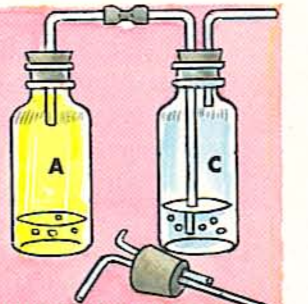

NOTE: EACH TIME YOU REMOVE THE GAS-COLLECTING BOTTLE **B** FOR EXPERIMENT, CONNECT BOTTLES **A** AND **C** TO PREVENT CHLORINE FROM GETTING OUT IN THE ROOM.

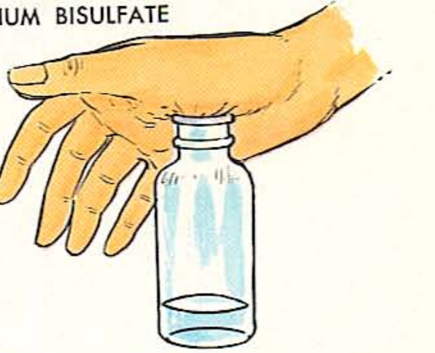

TO SHOW THE SOLUBILITY OF CHLORINE, POUR A SMALL AMOUNT OF WATER INTO A CHLORINE-FILLED BOTTLE. CLOSE THE BOTTLE MOUTH WITH YOUR PALM. SHAKE. THE CHLORINE DISSOLVES AND THE BOTTLE STICKS TO YOUR PALM FROM THE SUCTION CREATED.

CHLORINE HAS GREAT USE IN BLEACHING COTTON AND LINEN AND WOOD PULP. YET IT IS NOT THE CHLORINE THAT PERFORMS THE BLEACHING.

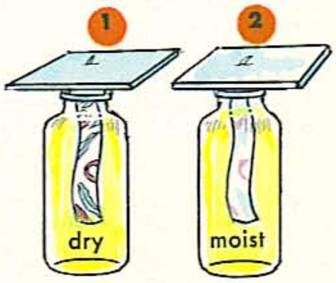

1 FILL A BOTTLE WITH CHLORINE GAS. HANG IN IT (FROM A CORK OR FROM A PIECE OF CARDBOARD) A STRIP OF DRY, BRIGHTLY COLORED COTTON CLOTH. NOTHING HAPPENS. COLOR OF CLOTH IS NOT AFFECTED.

2 MOISTEN THE CLOTH AND AGAIN HANG IT IN THE CHLORINE. SOON THE COLORS FADE—ONLY TRULY "FAST" COLORS REMAIN. CHLORINE, IN CONTACT WITH WATER, COMBINES WITH THE HYDROGEN AND LIBERATES OXYGEN. THE LIBERATED OXYGEN DOES THE BLEACHING.

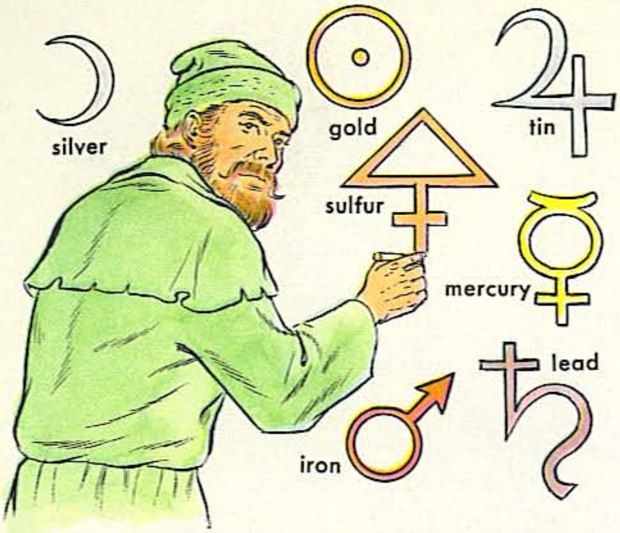

THE ALCHEMISTS USED FANCIFUL FIGURES TO REPRESENT THE CHEMICALS WITH WHICH THEY WORKED.

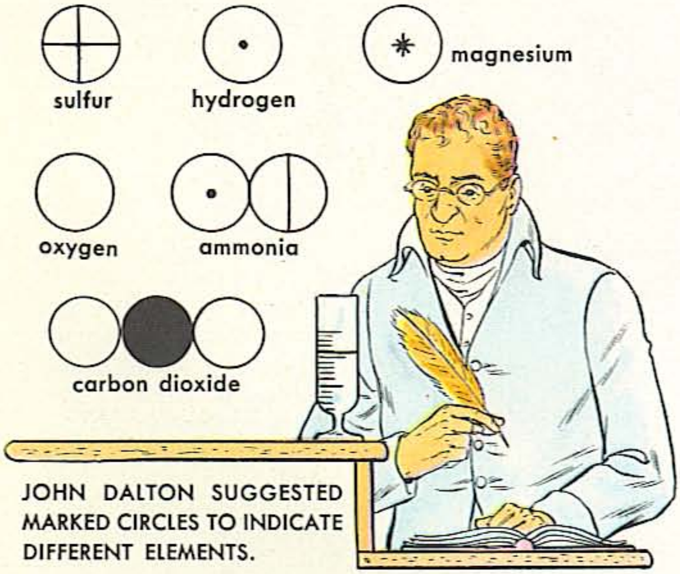

JOHN DALTON SUGGESTED MARKED CIRCLES TO INDICATE DIFFERENT ELEMENTS.

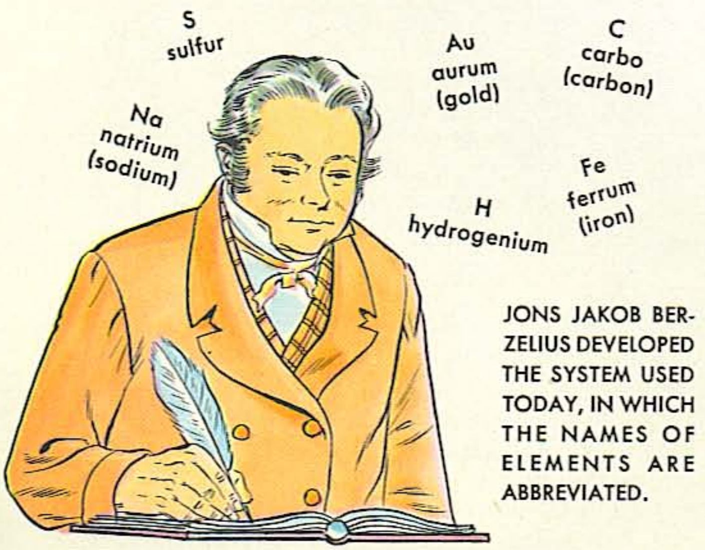

JONS JAKOB BERZELIUS DEVELOPED THE SYSTEM USED TODAY, IN WHICH THE NAMES OF ELEMENTS ARE ABBREVIATED.

Chemical Shorthand

So FAR YOU have experimented with oxygen and hydrogen, carbon dioxide and nitrogen, and chlorine; you have also separated water into the two elements of which it consists, and have combined the two elements iron and sulfur into a chemical compound. In taking notes of your experiments you are certain to have learned that it is much quicker to write "H" than "hydrogen," and easier to write "CO_2" than "carbon dioxide." Before long, it will seem the simplest and most logical thing in the world to use these abbreviations of the names of the different elements rather than the full names.

Yet it took chemists hundreds of years before they settled on this uniform method of writing out their chemical formulas.

In the early days of chemistry no one bothered to do much writing about it. But it became necessary for the alchemists to write down their experiments — how else could they retrace their steps in case they actually hit upon the gold they were seeking? They invented a whole line of complicated symbols that only they could understand.

As chemists delved deeper and deeper into the mysteries of matter it became more and more important for them to write out their experiments in such a way that all other chemists would know what they were trying to explain.

The first to invent a usable system was John Dalton, an English scientist. The invention was almost forced upon him.

In his study of chemistry he had become convinced that all chemical reactions could be explained in terms of the tiniest possible part of one element reacting with the tiniest possible part of another. These particles he called "atoms." The smallest possible part of the compound that resulted he called a "compound atom"— today we call it a "molecule."

To explain his "atomic theory" Dalton made use of circles, each with a marking to indicate a specific element. These circles served to explain Dalton's theory but they were too difficult to work with to show complicated.chemical reactions.

A Swedish chemist, Jons Jakob Berzelius, worked out a simpler system — the same system scientists use today.

For his symbols he took the first letter of the Latin name of each element — C for "carbo," S for "sulfur." Where two names started with the same letter, he added a small letter to one of the symbols to distinguish the two elements from each other — he used Ca for "calcium," for instance, to distinguish it from carbon (C).

But Berzelius went an important step further.

By then the French chemist, Joseph Louis Proust, had discovered that whenever elements form compounds these are always of a very definite composition — the "Law of Definite Composition." Water molecules, for example, always contain the same number of hydrogen and oxygen atoms. And Dalton had found that when two elements combine in different ways they do this in simple proportions — the "Law of Multiple Proportions." One atom of carbon and one atom of oxygen make carbon monoxide; one atom of carbon and two atoms of oxygen make carbon dioxide.

To describe these things in a simple way Berzelius made each of his symbols stand not only for a specific element but also for its relative weight as compared to the weight of other elements—its "atomic weight." To show the composition of a compound he simply put together the symbols for the elements into a "formula"—CO, HCl, FeS, and so on. "CO" then not only meant that one atom of carbon and one atom of oxygen combine to make one molecule of carbon monoxide, but also that 12 weight units of carbon (12 being the atomic weight of carbon) combine with 16 weight units of oxygen (16 being the atomic weight of oxygen) to form 28 weight units of the compound carbon monoxide.

When a compound contained several atoms of the same element Berzelius indicated this by placing a number in front of the symbol. It was later found necessary to change this to a smaller number, called a "subscript," placed at the lower right of the symbol — H_2O, CO_2.

In recent years it has been necessary to change Dalton's idea of an atom as being the smallest indivisible part of an element. Nowadays we have machines, such as the cyclotron, that can bombard, or "smash" atoms into still smaller parts—neutrons, and electrically charged protons and electrons. According to today's atomic theory protons and neutrons form the nucleus of the atom and electrons whirl around the nucleus with such tremendous speed that they seem to form a "shell" around it.

But even with our new idea of an atom, Dalton's main theory is still useful for explaining chemical reactions, and Berzelius' method is still the simplest "shorthand" method any scientist has ever devised for writing them down.

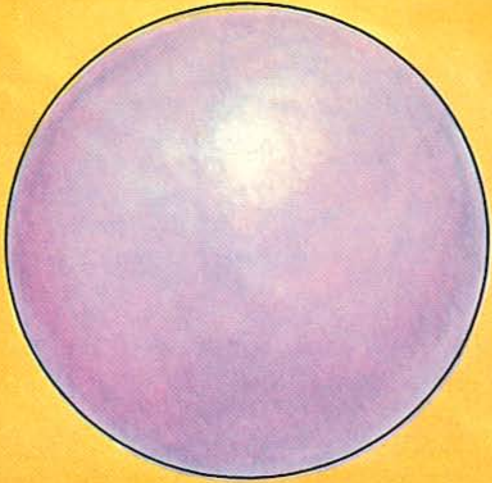

AN ATOM MIGHT LOOK LIKE A BALL SUCH AS THIS IF YOU ENLARGED IT A BILLION TIMES. THE "SHELL" IS NOT SOLID—IT CONSISTS OF ELECTRONS MOVING SO FAST THAT THEY SEEM TO FORM A SOLID SHELL.

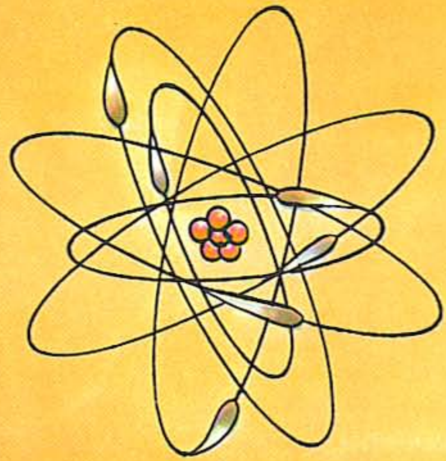

IF YOU COULD SLOW DOWN AN ENLARGED CARBON ATOM YOU MIGHT SEE TWO OF ITS ELECTRONS TRAVELING AROUND THE NUCLEUS IN AN "INNER SHELL" AND FOUR MORE WHIRLING AROUND IN AN "OUTER SHELL."

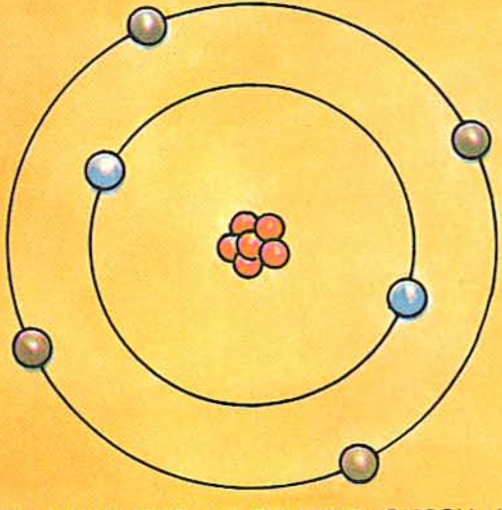

IF YOU COULD HALT AN ENLARGED CARBON ATOM COMPLETELY, IT WOULD LOOK A LOT LIKE OUR SOLAR SYSTEM, WITH A "SUN" (PROTONS AND NEUTRONS) IN THE CENTER AND "PLANETS" (ELECTRONS) AROUND IT.

The Periodic Table of the Elements

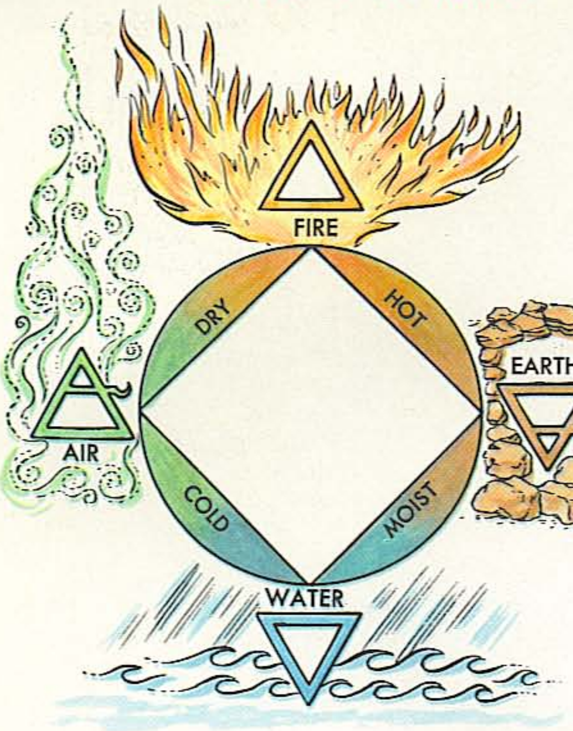

From the earliest times people have tried to explain what "matter" was made of. Most early philosophers agreed that "matter" was made up of what they called "elements." But their idea of an "element" was quite different from what we mean by that word today.

The early Greek philosophers thought the entire universe was composed of only four basic substances: fire, earth, water, and air. This explanation made sense at the time and was not seriously challenged for many centuries.

The old Romans actually knew nine of the substances we call elements today. They called them, of course, by their Latin names (the same we use today in chemical symbols): *carbo* (carbon — C), *sulfur* (S), *aurum* (gold — Au), *argentum* (silver —

FOR MORE THAN A THOUSAND YEARS PHILOSOPHERS INSISTED THAT ALL SUBSTANCES WERE MADE UP OF FOUR ELEMENTS: FIRE THAT WAS DRY AND HOT, EARTH THAT WAS HOT AND MOIST, WATER THAT WAS MOIST AND COLD, AIR THAT WAS COLD AND DRY. WE KNOW BETTER NOW!

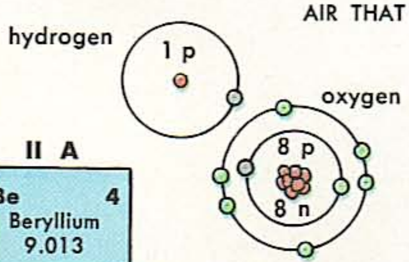

THE MODERN PICTURE OF AN ATOM HAS A NUCLEUS IN THE CENTER, CONSISTING OF PROTONS (p) AND NEUTRONS (n), WITH ELECTRONS IN RINGS AROUND IT.

THE PERIODIC TABLE OF THE ELEMENTS

	O	I A	II A	III A	IV A	V A	VI A	VII A	
1		**H** 1 Hydrogen 1.008							
2	**He** 2 Helium 4.003	**Li** 3 Lithium 6.940	**Be** 4 Beryllium 9.013						
3	**Ne** 10 Neon 20.183	**Na** 11 Sodium 22.991	**Mg** 12 Magnesium 24.32						
4	**A** 18 Argon 39.944	**K** 19 Potassium 39.1	**Ca** 20 Calcium 40.08	**Sc** 21 Scandium 44.96	**Ti** 22 Titanium 47.9	**V** 23 Vanadium 50.95	**Cr** 24 Chromium 52.01	**Mn** 25 Manganese 54.94	**Fe** 26 Iron 55.85
5	**Kr** 36 Krypton 83.8	**Rb** 37 Rubidium 85.48	**Sr** 38 Strontium 87.63	**Y** 39 Yttrium 88.92	**Zr** 40 Zirconium 91.22	**Nb** 41 Niobium 92.91	**Mo** 42 Molybdenum 95.95	**Tc** 43 Technetium 99	**Ru** 44 Ruthenium 101.1
6	**Xe** 54 Xenon 131.3	**Cs** 55 Cesium 132.91	**Ba** 56 Barium 137.36	57-71 Lanthanons	**Hf** 72 Hafnium 178.50	**Ta** 73 Tantalum 180.95	**W** 74 Tungsten 183.86	**Re** 75 Rhenium 186.22	**Os** 76 Osmium 190.2
7	**Rn** 86 Radon 222	**Fr** 87 Francium 223	**Ra** 88 Radium 226.05	89-103 Actinons					

🟪 INERT GASES
🟨 ALKALI METALS
🟦 ALKALINE EARTH METALS

La 57 Lanthanum 138.92	**Ce** 58 Cerium 140.13	**Pr** 59 Praseodymium 140.92	**Nd** 60 Neodymium 144.27	**Pm** 61 Promethium 145	**Sm** 62 Samarium 150.35
Ac 89 Actinium 227	**Th** 90 Thorium 232.05	**Pa** 91 Protactinium 231	**U** 92 Uranium 238.07	**Np** 93 Neptunium 237	**Pu** 94 Plutonium 242

ROWS RUNNING FROM LEFT TO RIGHT ARE CALLED PERIODS. COLUMNS RUNNING FROM TOP TO BOTTOM ARE CALLED GROUPS. THE ELEMENTS WITHIN A GROUP HAVE MANY TRAITS IN COMMON.

Ag), *ferrum* (iron — Fe), *cuprum* (copper — Cu), *stannum* (tin — Sn), *plumbum* (lead — Pb), *hydrargyrum* (mercury—Hg).

By 1800, thirty-four elements had been discovered. Within the next ten years, thirteen more had been added and had been given made-up Latin names — among them *natrium* (sodium — Na), *kalium* (potassium — K), and *aluminium* (aluminum — Al). By the beginning of the twentieth century, eighty-four elements were known.

Today the number has reached 102 — the last ten man-made, produced by splitting the atoms of other elements. Within a short time, Element 103 will probably be discovered.

In this table you will find listed the 102 elements that are known today. Each element is described by its chemical symbol, its atomic number, its full name, and its atomic weight.

MANY SCIENTISTS HAD NOTICED THAT IF YOU LINE UP THE ELEMENTS ACCORDING TO ATOMIC WEIGHTS, CERTAIN CHEMICAL TRAITS OCCUR PERIODICALLY. THE RUSSIAN SCIENTIST, DMITRI MENDELEEFF, ON THIS BASIS DISCOVERED THE PERIODIC LAW AND DEVELOPED THE PERIODIC TABLE.

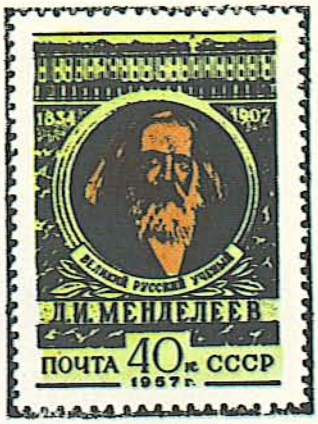

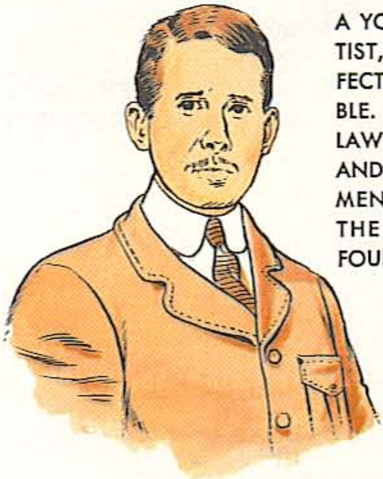

A YOUNG ENGLISH SCIENTIST, HENRY MOSELEY, PERFECTED THE PERIODIC TABLE. HE DISCOVERED THE LAW OF ATOMIC NUMBERS AND ARRANGED THE ELEMENTS ACCORDING TO THE ELECTRIC CHARGE FOUND IN THE NUCLEUS.

magnesium

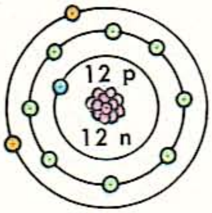

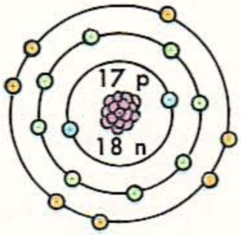

chlorine

THE NUMBER OF PROTONS IN AN ATOM IS ITS ATOMIC NUMBER. AN ATOM ALWAYS HAS THE SAME NUMBER OF PROTONS AND ELECTRONS. HYDROGEN IS THE SIMPLEST OF ALL ATOMS.

VIII		I B	II B	III B	IV B	V B	VI B	VII B
				B 5 Boron 10.82	C 6 Carbon 12.011	N 7 Nitrogen 14.008	O 8 Oxygen 16	F 9 Fluorine 19
				Al 13 Aluminum 26.98	Si 14 Silicon 28.09	P 15 Phosphorus 30.975	S 16 Sulfur 32.066	Cl 17 Chlorine 35.457
Co 27 Cobalt 58.94	Ni 28 Nickel 58.71	Cu 29 Copper 63.54	Zn 30 Zinc 65.38	Ga 31 Gallium 69.72	Ge 32 Germanium 72.6	As 33 Arsenic 74.91	Se 34 Selenium 78.96	Br 35 Bromine 79.916
Rh 45 Rhodium 102.91	Pd 46 Palladium 106.4	Ag 47 Silver 107.88	Cd 48 Cadmium 112.41	In 49 Indium 114.82	Sn 50 Tin 118.7	Sb 51 Antimony 121.76	Te 52 Tellurium 127.61	I 53 Iodine 126.91
Ir 77 Iridium 192.2	Pt 78 Platinum 195.09	Au 79 Gold 197	Hg 80 Mercury 200.61	Tl 81 Thallium 204.39	Pb 82 Lead 207.21	Bi 83 Bismuth 209	Po 84 Polonium 210	At 85 Astatine 210

🟥 HEAVY METALS
🟪 METALLOIDS
🟩 NON-METALS
🟧 RARE EARTH METALS
🟨 UNSTABLE ELEMENTS

Eu 63 Europium 152	Gd 64 Gadolinium 157.26	Tb 65 Terbium 158.93	Dy 66 Dysprosium 162.51	Ho 67 Holmium 164.94	Er 68 Erbium 167.27	Tm 69 Thulium 168.94	Yb 70 Ytterbium 173.04	Lu 71 Lutetium 174.99
Am 95 Americium 243	Cm 96 Curium 247	Bk 97 Berkelium 249	Cf 98 Californium 249	E 99 Einsteinium 254	Fm 100 Fermium 255	Mv 101 Mendelevium 256	No 102 Nobelium 251	? 103

The Mysteries of Solutions

FROM THE earliest days, scientists experimenting with chemistry have worked with solutions. The liquid they used for making a solution (usually water) they called the "solvent." The chemical dissolved was the "solute."

When chemists began to use electricity as one of their tools, they discovered that different solutions behaved in different ways. The solution in water of a great number of chemicals — sugar among them — did not let electricity pass through. They were "non-conductors." Some chemicals, on the other hand, conducted electricity very easily. They were good conductors — "electrolytes."

In 1874 a Swedish scientist named Svante Arrhenius developed a theory to help explain the mysterious behavior of solutions. He was only 25 years old at the time.

His idea was that when a chemical that conducts electricity is dissolved in water, each molecule is broken up — "dissociated" — into electrically charged atoms or groups of atoms. These atoms or groups of atoms Arrhenius called "ions" from a Greek word that means "to wander." His new theory came to be called "Arrhenius' theory of ionization."

When table salt (sodium chloride, NaCl), for instance, is dissolved in water, it ionizes into positively charged sodium ions (Na^+) and negatively charged chlorine ions (Cl^-). These ions "wander" about in all directions until an electric current is applied to the solution. When that happens, the negative ions rush to the positive pole, the positive ions to the negative pole. It is the ions that conduct the current through the solution.

The reason that non-conductors do not conduct electricity is that they do not dissociate into ions.

Arrhenius' theory of ionization helped explain a great number of things that have puzzled chemists. His theory has been modified somewhat over the years but in most respects holds true today.

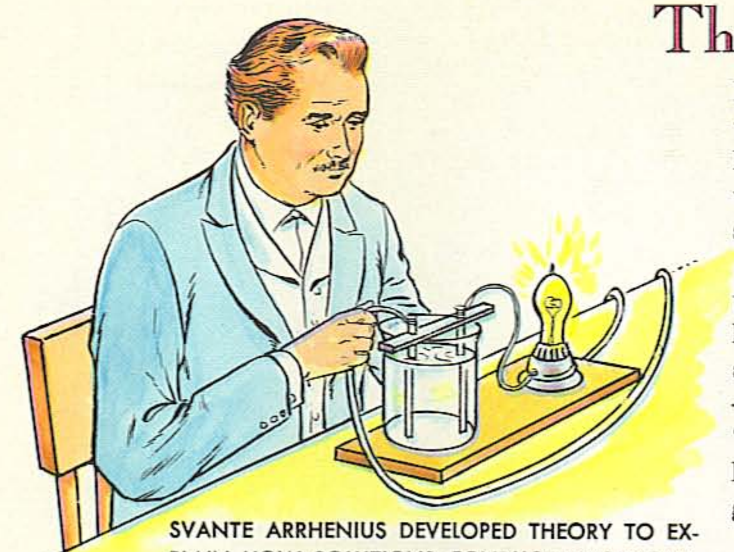

SVANTE ARRHENIUS DEVELOPED THEORY TO EXPLAIN HOW SOLUTIONS CONDUCT ELECTRICITY.

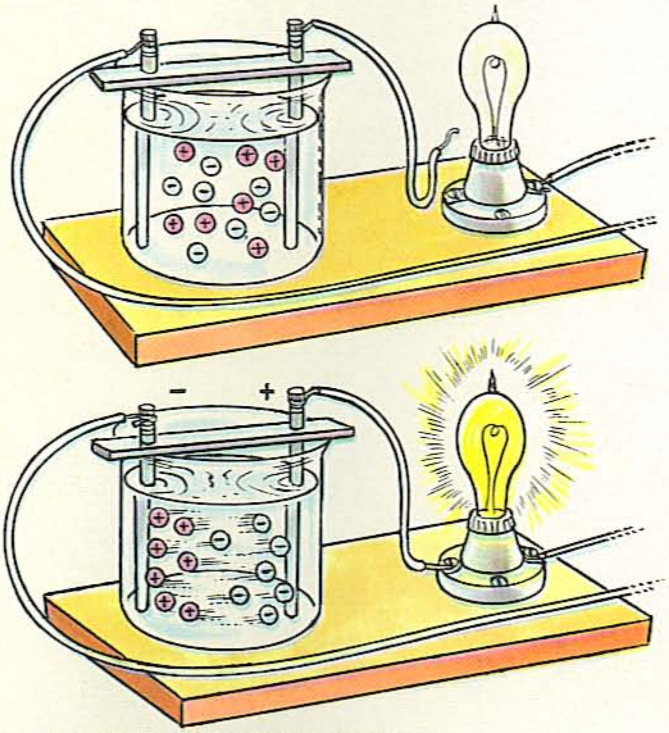

IN HIS EARLY EXPERIMENTS, SVANTE ARRHENIUS USED A SIMPLE SET-UP. YOU CAN EASILY REPEAT SOME OF HIS EXPERIMENTS IN YOUR OWN LAB, USING FLASHLIGHT BATTERIES.

TESTING CONDUCTIVITY OF SOLUTIONS

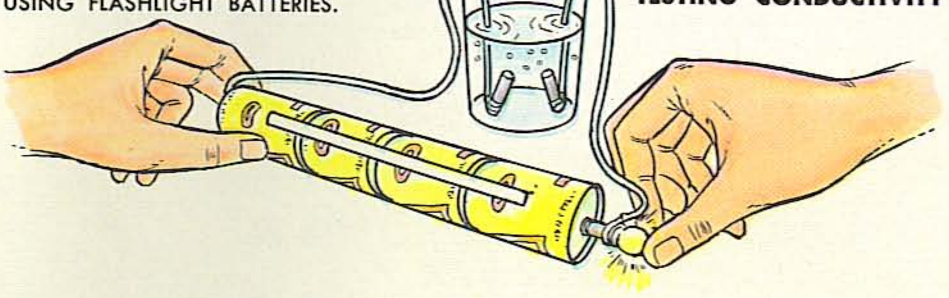

SET UP THE SAME APPARATUS AS ON PAGE 25. ADD FLASHLIGHT BULB TO END OF ONE WIRE. TRY DIFFERENT SOLUTIONS IN GLASS. SOME CONDUCT ELECTRICITY AND BULB LIGHTS UP. OTHERS DO NOT AND THE BULB DOES NOT LIGHT UP.

SATURATED SOLUTIONS

A SATURATED SOLUTION IS ONE IN WHICH NO MORE OF THE CHEMICAL WILL GO IN SOLUTION AT THAT PARTICULAR TEMPERATURE.

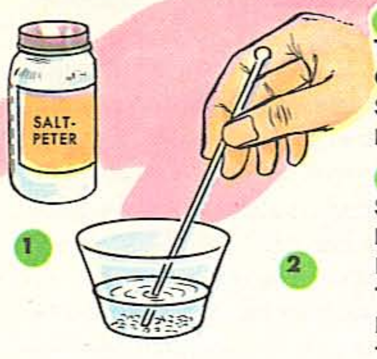

1 POUR 20 ml WATER OF ROOM TEMPERATURE INTO A CUSTARD CUP. ADD 6 g SALTPETER (POTASSIUM NITRATE). STIR. ALL THE SALTPETER DISSOLVES.

2 ADD 3 g MORE SALTPETER. STIR. SOME OF THE ADDED SALTPETER DOES NOT DISSOLVE. CLEAR LIQUID IS SATURATED AT ROOM TEMPERATURE. (AT 20°C., 6.3 g KNO_3 MAKES SATURATED SOLUTION IN 20 ml WATER.)

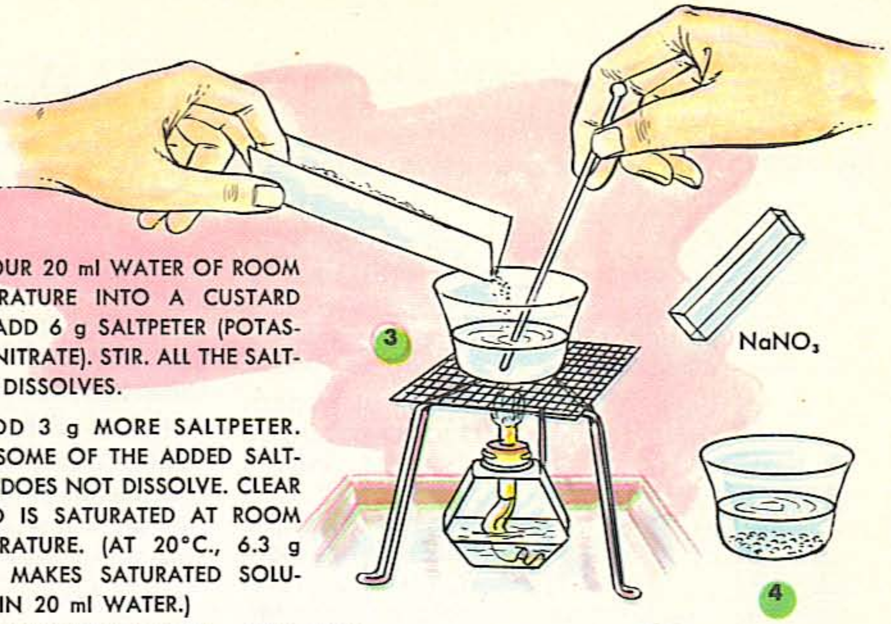

3 PLACE CUSTARD CUP OVER ALCOHOL BURNER. ADD 10 g MORE SALTPETER. SOON ALL SALTPETER IS DISSOLVED. AT HIGHER TEMPERATURES IT TAKES MORE SOLUTE TO MAKE A SATURATED SOLUTION. (AT BOILING, 20 ml H_2O DISSOLVES 49 g SALTPETER.)

4 TAKE SOLUTION OFF FIRE. AS IT COOLS, MUCH OF THE SALTPETER COMES OUT AS CRYSTALS BY SLOW CRYSTALLIZATION. LIQUID IS AGAIN A SOLUTION SATURATED AT ROOM TEMPERATURE.

BEHAVIOR OF SOLUTIONS

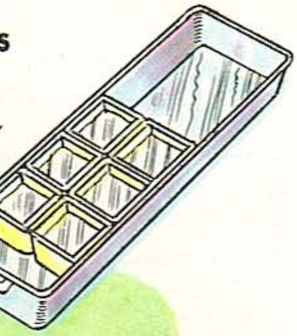

SOLUTION HAS LOWER FREEZING POINT THAN THE SOLVENT USED.

IN TRAY WITH INDIVIDUAL ICE CUBE CUPS, POUR WATER INTO EACH CUP. IN ONE, DISSOLVE 1 PINCH OF SALT, IN NEXT 2 PINCHES, AND SO ON. LEAVE ONE WITHOUT SALT. PLACE IN FREEZING COMPARTMENT. CUPS LEAST SALTED FREEZE FIRST.

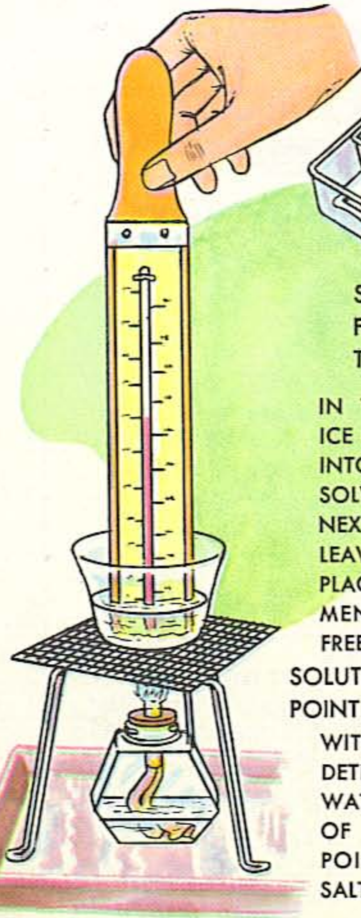

SOLUTION HAS HIGHER BOILING POINT THAN THE SOLVENT USED.

WITH CANDY THERMOMETER, DETERMINE AT WHAT POINT WATER BOILS. ADD 1 PINCH OF SALT. WHAT IS BOILING POINT NOW? ADD MORE SALT. READ AGAIN.

CRYSTALLIZATION

YOU CAN FOLLOW CRYSTALLIZATION OF $MgSO_4$. IN TEST TUBE, HEAT MIXTURE OF 5 ml WATER AND 1 TEASPOON EPSOM SALT UNTIL SALT DISSOLVES. POUR HOT SOLUTION OVER PANE OF GLASS CLEANED WITH DETERGENT. CRYSTALS MAKE NEEDLE-LIKE NETWORK.

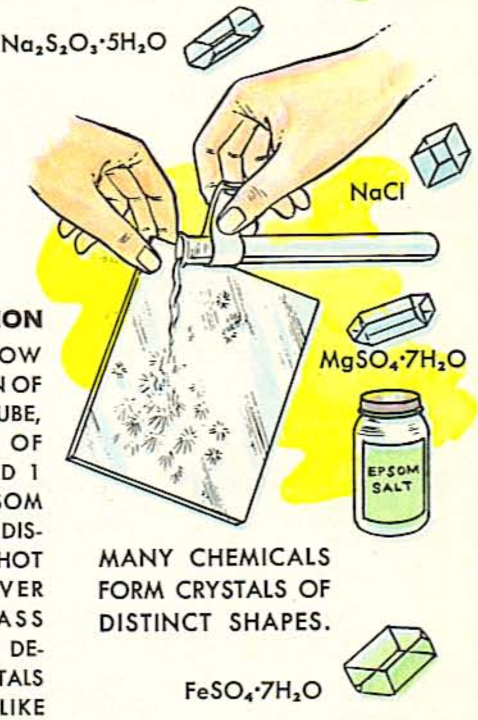

MANY CHEMICALS FORM CRYSTALS OF DISTINCT SHAPES.

MAKING SOLUTIONS

MAKE 50 ml GRADUATE FIRST: MEASURE 50 ml WATER INTO A NARROW JAR, USING 10 ml TEST TUBE GRADUATE SHOWN ON PAGE 15. MAKE A MARK AT 50 ml LEVEL.

10% (10 PER CENT) SOLUTION: MEASURE 40 ml WATER INTO A CUSTARD CUP. ADD 5 g OF THE CHEMICAL. STIR. (TO MAKE IT DISSOLVE QUICKER, YOU MAY WANT TO HEAT THE WATER SLIGHTLY.) POUR SOLUTION INTO 50 ml GRADUATE. ADD WATER TO THE 50 ml MARK.

2% SOLUTION: MEASURE 40 ml WATER INTO CUSTARD CUP. ADD 1 g OF THE CHEMICAL. STIR TO DISSOLVE. POUR INTO 50 ml GRADUATE. ADD WATER TO 50 ml.

HOW DO YOU KNOW AN ACID?

1. ACIDS TASTE SOUR.

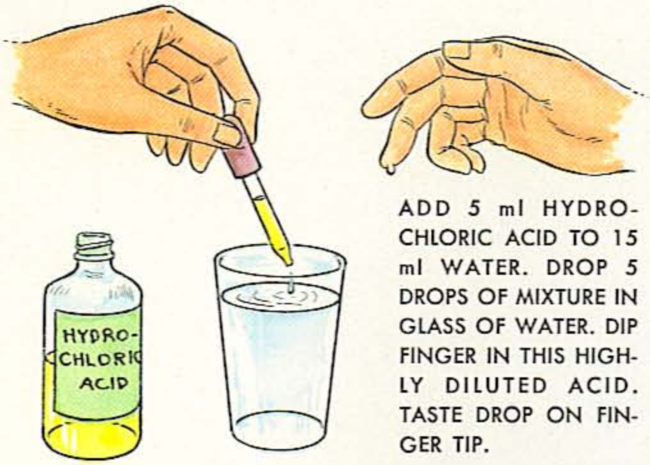

ADD 5 ml HYDROCHLORIC ACID TO 15 ml WATER. DROP 5 DROPS OF MIXTURE IN GLASS OF WATER. DIP FINGER IN THIS HIGHLY DILUTED ACID. TASTE DROP ON FINGER TIP.

2. ACIDS ACT WITH INDICATORS.

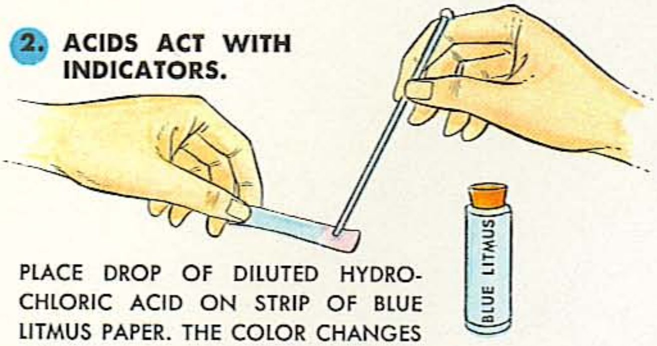

PLACE DROP OF DILUTED HYDROCHLORIC ACID ON STRIP OF BLUE LITMUS PAPER. THE COLOR CHANGES TO RED.

3. ACIDS ACT WITH METALS.

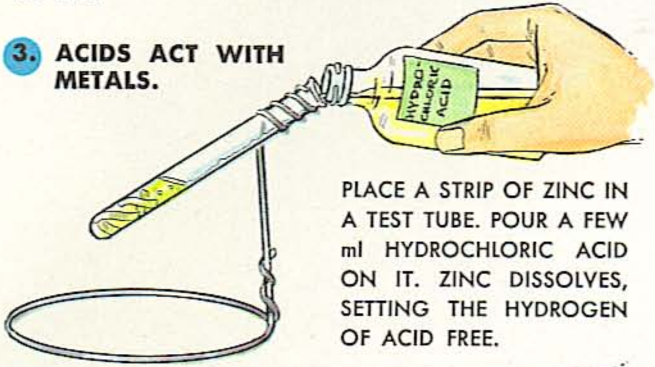

PLACE A STRIP OF ZINC IN A TEST TUBE. POUR A FEW ml HYDROCHLORIC ACID ON IT. ZINC DISSOLVES, SETTING THE HYDROGEN OF ACID FREE.

4. ACIDS NEUTRALIZE BASES.

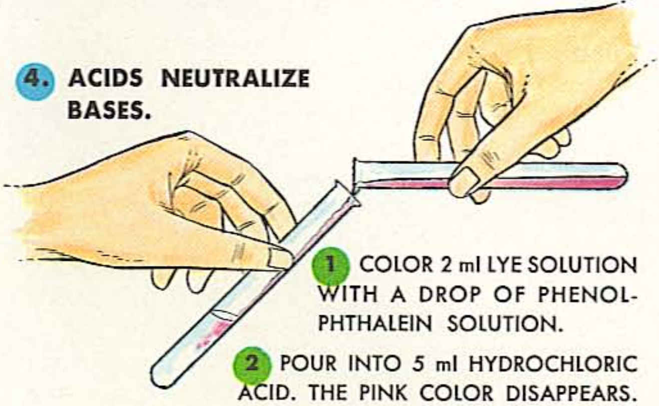

① COLOR 2 ml LYE SOLUTION WITH A DROP OF PHENOLPHTHALEIN SOLUTION.

② POUR INTO 5 ml HYDROCHLORIC ACID. THE PINK COLOR DISAPPEARS.

Working With Acids

ACIDS have many traits in common. They taste sour. They change the color of certain plant substances—which are called "indicators." They contain hydrogen (H) that can be replaced by a metal. They neutralize bases.

But what *is* an acid? Earlier, the "acidic" traits were used to define an acid. But with the modern understanding of the atom, a different definition is used. You will remember that the nucleus of an atom contains positively charged protons. Acids in solution liberate protons as ions (H^+). And so we say that an acid is a substance that will give up — or "donate" — protons to another substance. Acids are "proton donors." The foremost acids used in industry are sulfuric acid (H_2SO_4), nitric acid (HNO_3), and hydrochloric acid (HCl).

The first two — sulfuric acid and nitric acid — should NEVER be used in the home lab. They are much too DANGEROUS. They destroy the skin and might blind you if you got them in the eyes. (Wherever a chemical experiment would ordinarily call for sulfuric acid, this book uses sodium acid sulfate — $NaHSO_4$, sodium bisulfate, "Sani-Flush"; wherever

(CONTINUED ON PAGE 44)

HOME-MADE INDICATORS

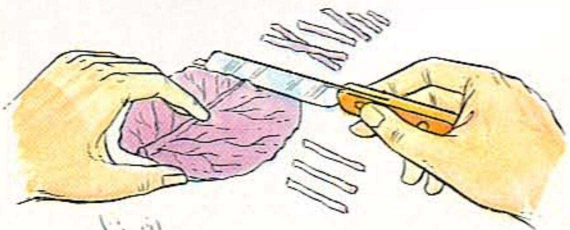

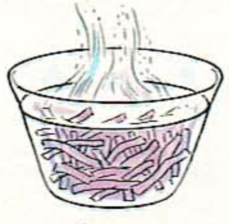

CUT UP OR GRATE A RED CABBAGE LEAF. DROP IN HOT WATER. STEEP FOR ½ HOUR. POUR OFF LIQUID. USE AS INDICATOR.

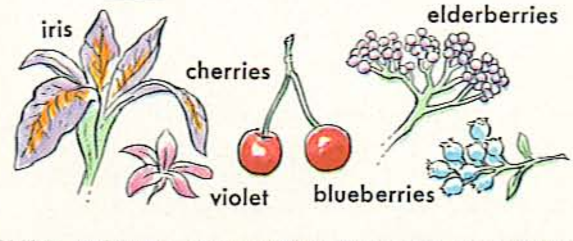

MANY FLOWERS AND FRUITS CONTAIN COLORING MATTER WHICH YOU CAN EXTRACT WITH HOT WATER AND USE AS AN INDICATOR FOR ACIDS AND BASES.

Working With Bases

BASES taste brackish. They change the color of "indicators." They contain a combination of oxygen and hydrogen atoms called "hydroxyl" (OH). They neutralize acids.

But what *is* a base? When a base is dissolved in water it liberates negatively charged hydroxyl ions (OH^-). When a base is neutralized, these ions take on — or "accept"— positively charged protons from another substance. A base is a substance that will accept and combine with protons from another substance. Bases are "proton acceptors." The most important bases are sodium hydroxide ("lye," NaOH), ammonium hydroxide ("ammonia," NH_4OH), and calcium hydroxide ("slaked lime," $Ca(OH)_2$).

The first of these — sodium hydroxide — is used in many households to clean sluggish drains and to keep sinks from stopping up ("Drano"). USE IT WITH GREAT CARE in your experiments. Do not touch lye flakes with your fingers and do not get the solution on your skin — it dissolves the natural oil. It is particularly dangerous to get lye in your eyes. If you get lye on you, dilute it quickly with LOTS OF WATER.

(CONTINUED ON PAGE 45)

LABORATORY INDICATORS

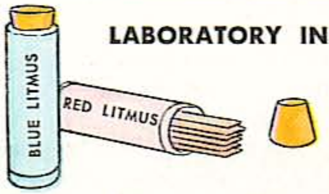

LITMUS PAPER IS MOST COMMONLY USED INDICATOR. AN ACID TURNS BLUE LITMUS RED. BASES TURN RED LITMUS BLUE.

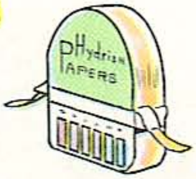

pHYDRION PAPER IS MORE EXACT INDICATOR FOR ACIDS AND BASES.

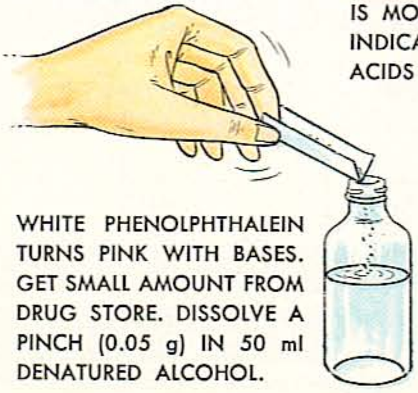

PHENOLPHTHALEIN

WHITE PHENOLPHTHALEIN TURNS PINK WITH BASES. GET SMALL AMOUNT FROM DRUG STORE. DISSOLVE A PINCH (0.05 g) IN 50 ml DENATURED ALCOHOL.

HOW DO YOU KNOW A BASE?

1. **BASES TASTE BRACKISH.**

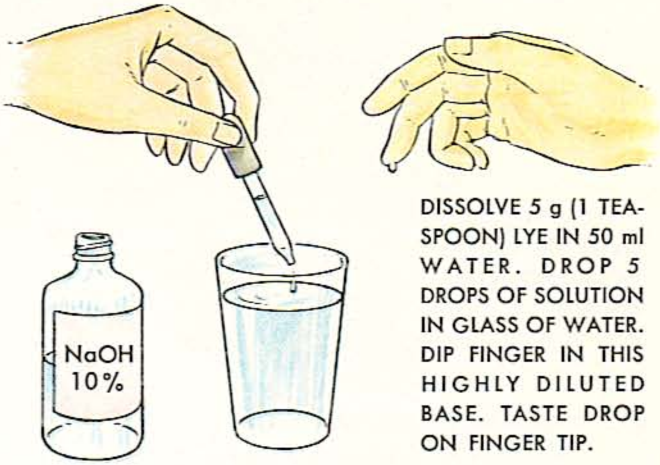

DISSOLVE 5 g (1 TEASPOON) LYE IN 50 ml WATER. DROP 5 DROPS OF SOLUTION IN GLASS OF WATER. DIP FINGER IN THIS HIGHLY DILUTED BASE. TASTE DROP ON FINGER TIP.

2. **BASES ACT WITH INDICATORS.**

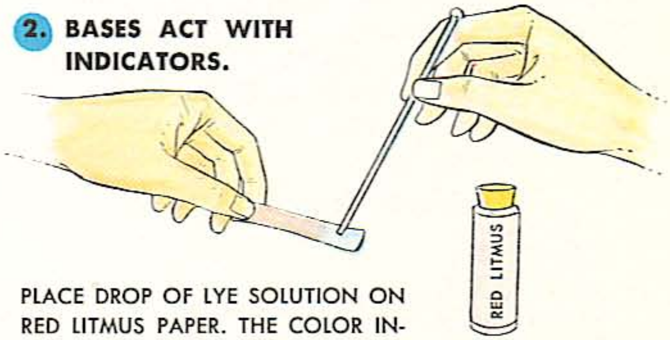

PLACE DROP OF LYE SOLUTION ON RED LITMUS PAPER. THE COLOR INSTANTLY CHANGES TO BLUE.

3. **BASES ACT WITH FAT.**

ADD TINY LUMP OF FAT TO 5 ml LYE SOLUTION. HEAT GENTLY. FAT DISSOLVES TO FORM SOAP.

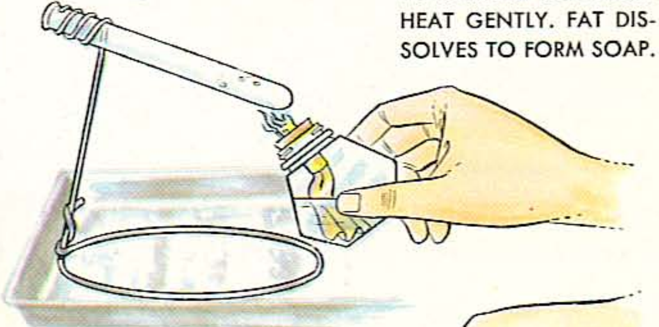

4. **BASES NEUTRALIZE ACIDS.**

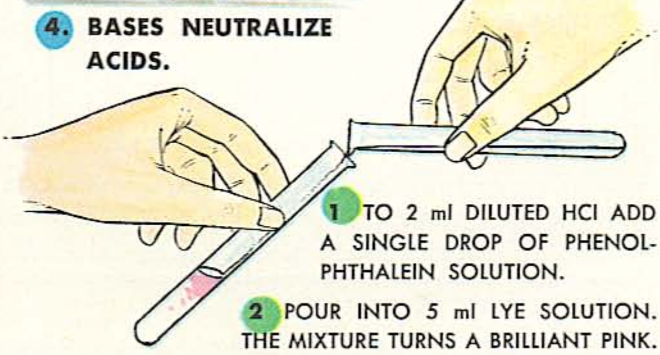

1 TO 2 ml DILUTED HCl ADD A SINGLE DROP OF PHENOLPHTHALEIN SOLUTION.

2 POUR INTO 5 ml LYE SOLUTION. THE MIXTURE TURNS A BRILLIANT PINK.

HOUSEHOLD ITEMS CONTAINING ACIDS

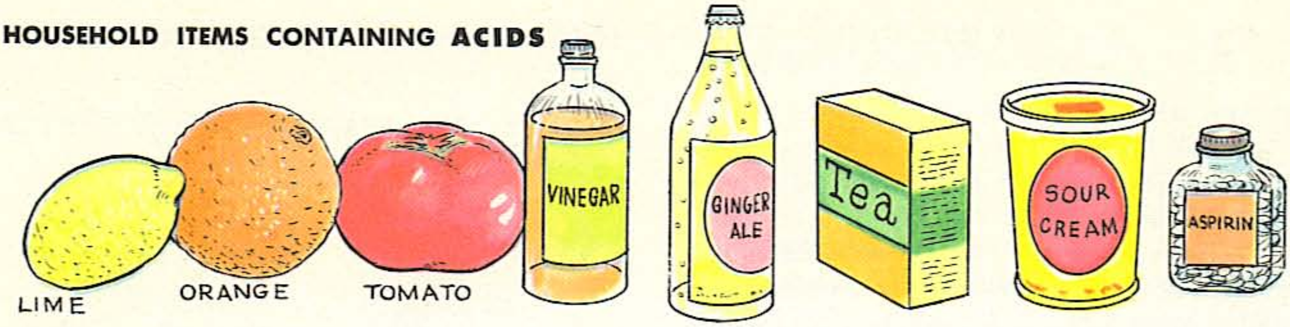

Acids—Continued

nitric acid would be called for, this book produces it in a mixture of a nitrate, KNO_3, and sodium bisulfate.)

Hydrochloric acid is used in many households under the name of "muriatic acid." Whenever you use hydrochloric acid in an experiment; USE IT WITH GREAT CARE. If any of it gets on you, dilute it quickly with LOTS OF WATER. Or neutralize it with bicarbonate of soda (but not if in the eyes).

ACID FROM NON-METALLIC OXIDE

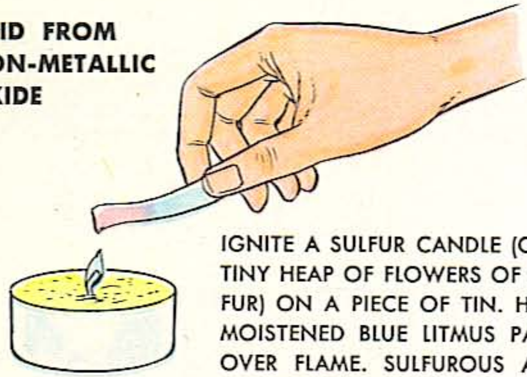

IGNITE A SULFUR CANDLE (OR A TINY HEAP OF FLOWERS OF SULFUR) ON A PIECE OF TIN. HOLD MOISTENED BLUE LITMUS PAPER OVER FLAME. SULFUROUS ACID FORMED TURNS IT RED.

ACID FROM A SALT

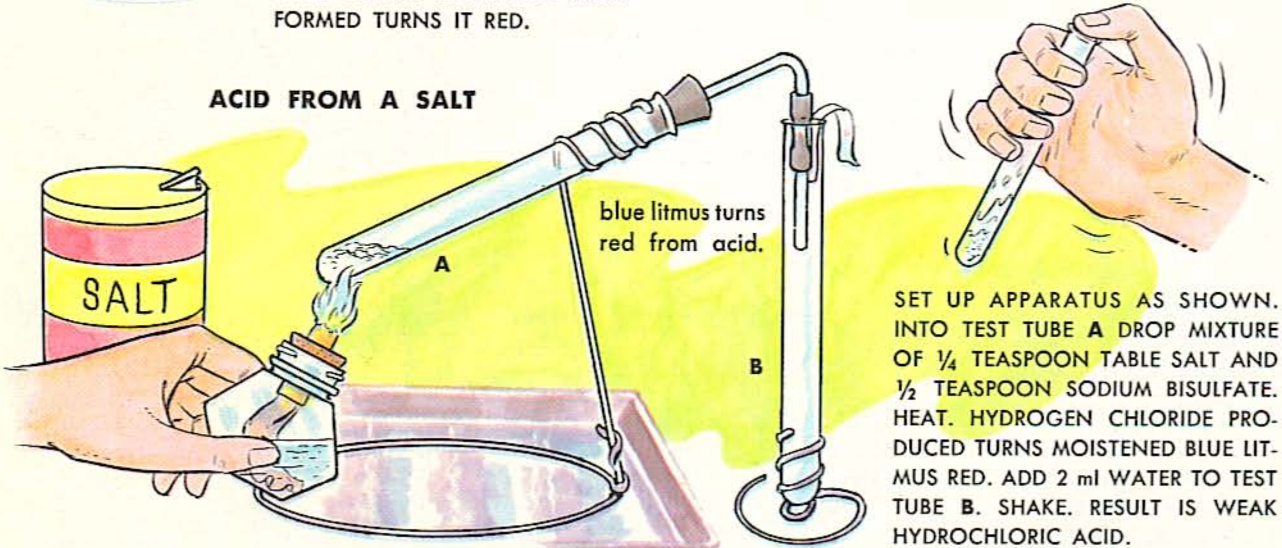
SALT
A
B
blue litmus turns red from acid.

SET UP APPARATUS AS SHOWN. INTO TEST TUBE **A** DROP MIXTURE OF ¼ TEASPOON TABLE SALT AND ½ TEASPOON SODIUM BISULFATE. HEAT. HYDROGEN CHLORIDE PRODUCED TURNS MOISTENED BLUE LITMUS RED. ADD 2 ml WATER TO TEST TUBE **B**. SHAKE. RESULT IS WEAK HYDROCHLORIC ACID.

pH SYSTEM IS A WAY OF DESCRIBING THE RELATIVE ACIDITY OR ALKALINITY OF A SOLUTION. PURE WATER IS NEUTRAL WITH pH7. THE LOWER THE NUMBER BELOW 7, THE MORE ACID THE SOLUTION. THE HIGHER

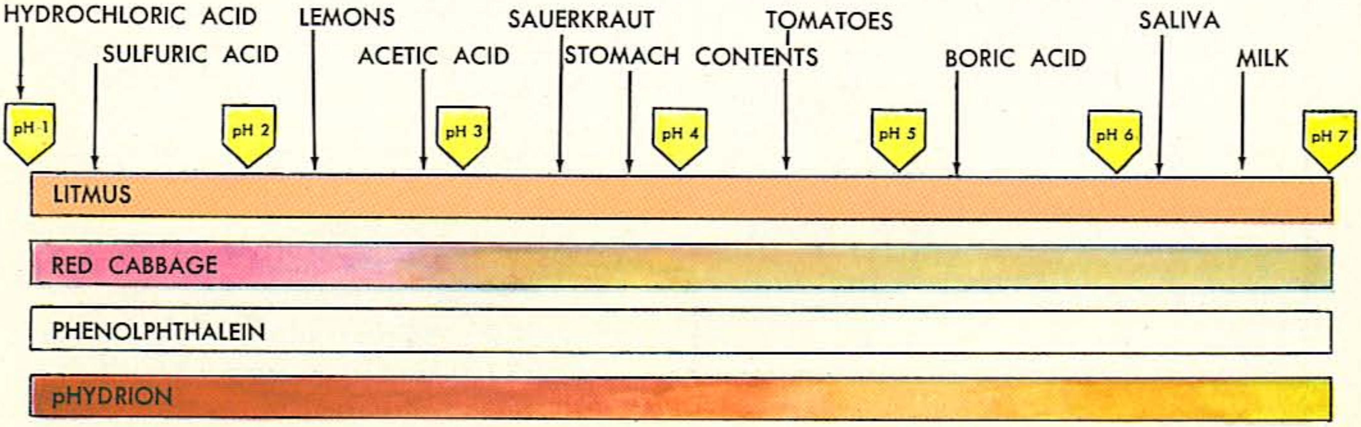

HOUSEHOLD ITEMS CONTAINING BASES

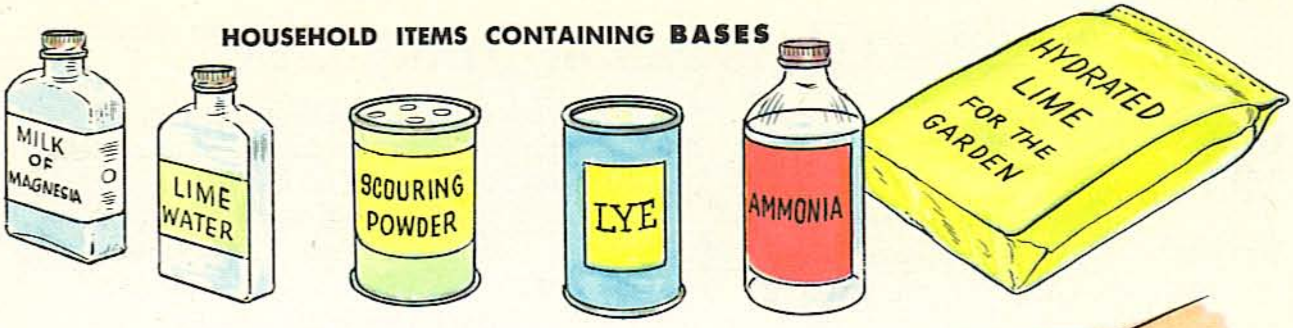

Bases—Continued

You can also neutralize it with vinegar (but not if in the eyes).

Ammonia is a common household cleaning liquid. Ammonia should also be handled with care and should be washed off quickly if you get it on you. Also watch your nose when you work with ammonia. It has a very strong smell.

Calcium hydroxide is a white powder. You will use it in a great number of experiments.

BASE FROM METALLIC OXIDE

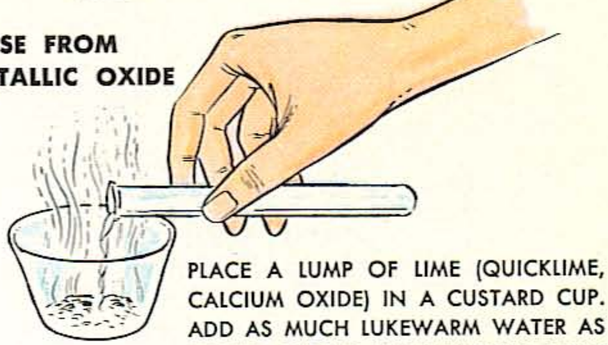

PLACE A LUMP OF LIME (QUICKLIME, CALCIUM OXIDE) IN A CUSTARD CUP. ADD AS MUCH LUKEWARM WATER AS IT WILL ABSORB. LIME HEATS UP, GIVES OFF STEAM, CRUMBLES INTO POWDER OF SLAKED LIME (CALCIUM HYDROXIDE).

BASE FROM A SALT

IN A CUSTARD CUP, DISSOLVE 1 TEASPOON SAL SODA (WASHING SODA, SODIUM CARBONATE) IN 50 ml WATER. HEAT SLIGHTLY. ADD SLAKED LIME MIXED WITH WATER. STIR. CHEMICAL REACTION PRODUCES SODIUM HYDROXIDE AND CALCIUM CARBONATE. FILTER. CLEAR LIQUID CONTAINS THE SODIUM HYDROXIDE (LYE). THE CALCIUM CARBONATE IS HELD BACK BY THE FILTER.

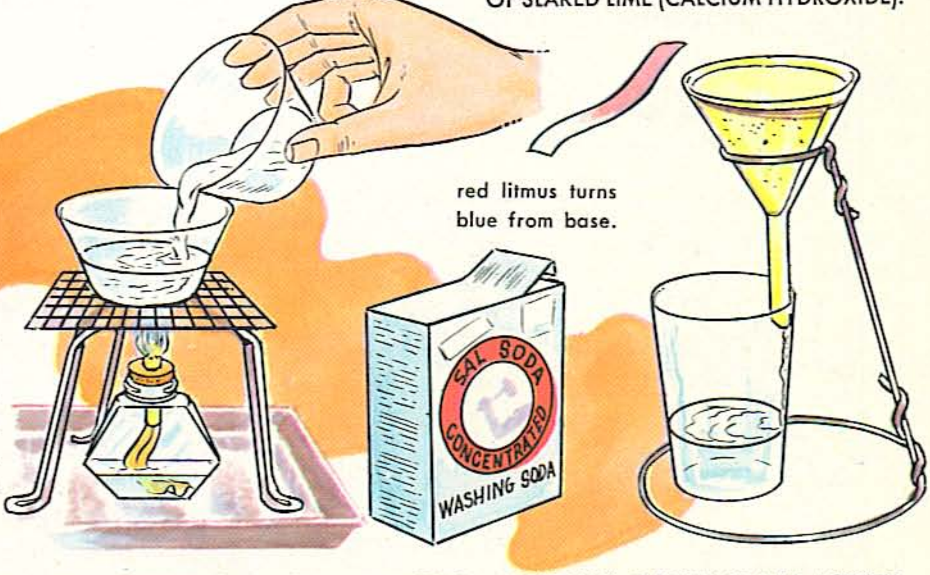

red litmus turns blue from base.

THE NUMBER ABOVE 7, THE MORE ALKALINE THE SOLUTION. WHEN YOU KNOW AT WHAT pH AN INDICATOR CHANGES COLOR, YOU CAN DETERMINE THE ACIDITY OR ALKALINITY OF THE SOLUTION YOU ARE TESTING.

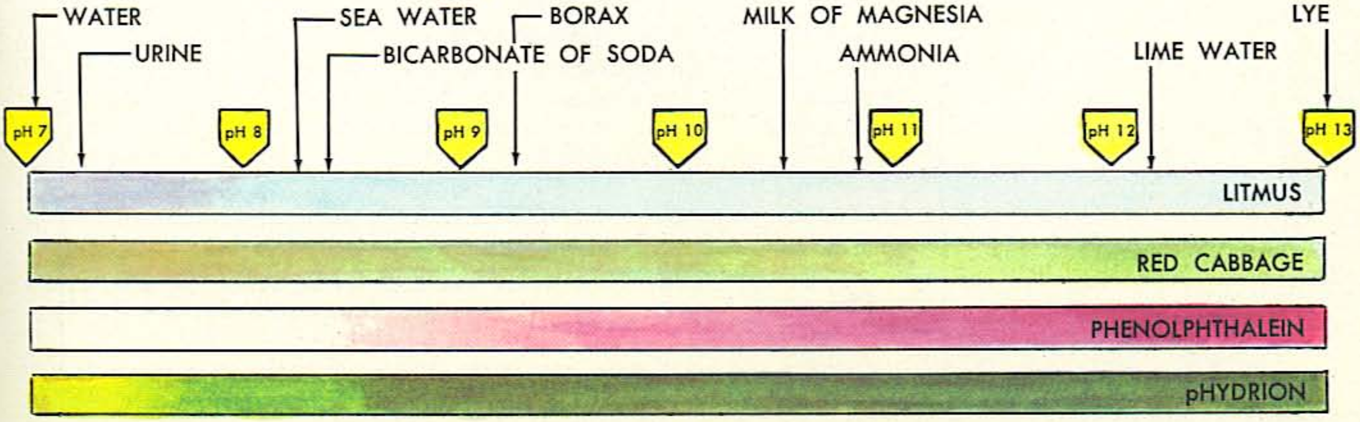

Salts—Chemicals of Many Uses

NEUTRALIZATION IS USED EXTENSIVELY IN CHEMICAL ANALYSIS IN A TECHNIQUE CALLED **TITRATION.**

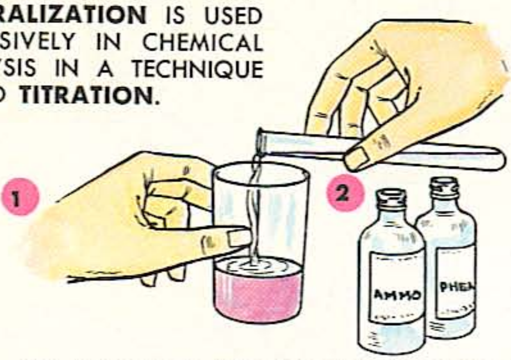

TO DETERMINE THE UNKNOWN STRENGTH OF A BASE, THE CHEMIST DROPS INTO IT FROM A LONG TUBE—A BURETTE—AS MUCH ACID OF KNOWN STRENGTH AS IS NECESSARY TO NEUTRALIZE IT. BY CHECKING ACID USED HE FIGURES STRENGTH OF BASE.

① FOR A TRY AT TITRATION, MIX A FEW ml OF HOUSEHOLD AMMONIA WITH 40 ml WATER. ADD A DROP OF PHENOLPHTHALEIN. THIS WILL COLOR THE MIXTURE A DEEP PINK.

② POUR 10 ml DILUTED HYDROCHLORIC ACID INTO MEASURING TUBE. POUR SOME OF THIS ACID INTO THE AMMONIA UNTIL COLOR HAS ALMOST VANISHED.

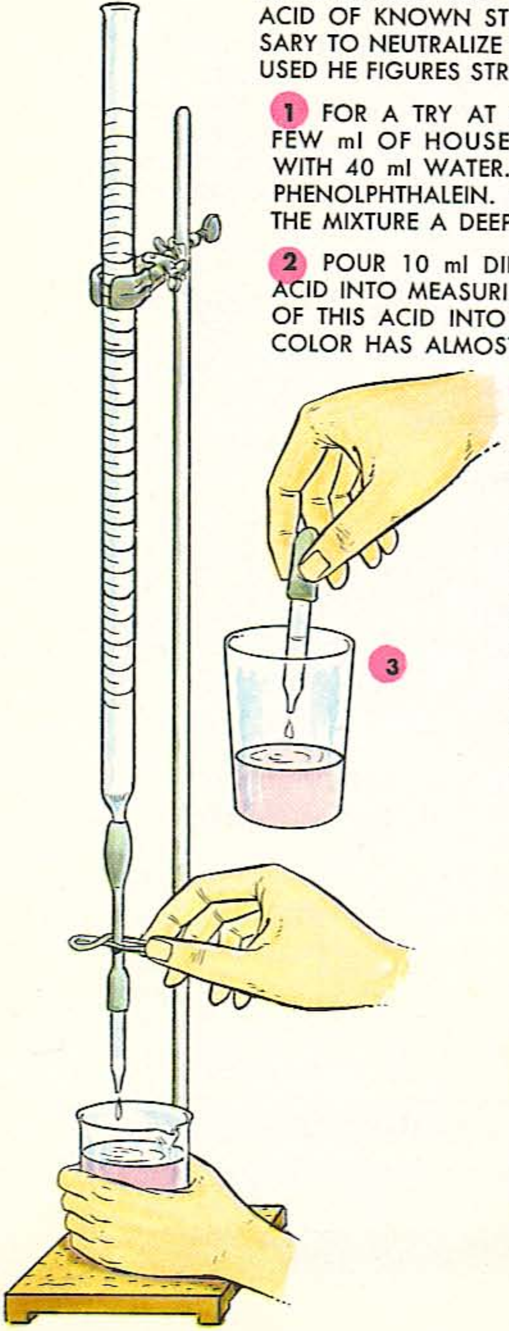

③ PICK UP A FEW ml OF THE MEASURED ACID IN AN EYE DROPPER (PIPETTE). DROP ACID SLOWLY INTO THE AMMONIA MIXTURE UNTIL COLOR IS COMPLETELY GONE. RETURN ACID NOT USED TO MEASURING TUBE. YOU NOW KNOW HOW MANY ml ACID YOU HAD TO USE TO NEUTRALIZE THE AMMONIA.

WHAT HAPPENS when you neutralize an acid with a base or a base with an acid? The hydrogen atoms (H^+ ions) of the acid combine with the hydroxyl groups (OH^- ions) of the base to form water, and the metal atoms of the base combine with what remains of the acid to form a salt. Or simply:

BASE plus ACID turns into
WATER plus SALT

This, for example, is what happens when you neutralize sodium hydroxide with hydrochloric acid:

$$\mathrm{NaOH} + \mathrm{HCl} \rightarrow \mathrm{HOH} + \mathrm{NaCl}$$

The result is water and sodium chloride — ordinary table salt which has given its name to other substances of a similar nature.

Of all the salts used in industry, table salt ($NaCl$) and washing soda (Na_2CO_3) are of greatest importance. Numerous other chemicals are produced from them. Our way of life would be completely disrupted if our country's industry did not have enough of these two salts.

Many other salts are necessary for our well-being. You'll probably find at least half a dozen different salts used daily in your home — in cooking and baking, in gardening, for cleaning.

In your chemical experiments you'll be working with two classes of salts: *normal salts* (such as $NaCl$, Na_2CO_3, KI) which contain no free hydrogen or hydroxyl ions, and *acid salts* (such as $NaHSO_4$, $NaHCO_3$) which contain replaceable hydrogen.

Some of these salts dissolve easily in water — all the nitrates (salts of nitric acid) and most of the chlorides (salts of hydrochloric acid). Many salts, on the other hand, are insoluble — most of the carbonates (salts of carbonic acid) and most sulfides (salts of hydrosulfuric acid).

HOW THE NAMES OF SALTS ARE MADE UP

THE ACID	FORMULA AND NAME OF ACID		FORMULA AND NAME OF SALT	
SULFURIC ACID	H_2SO_4	HYDROGEN SULFATE	Na_2SO_4	SODIUM SULFATE
NITRIC ACID	HNO_3	HYDROGEN NITRATE	$NaNO_3$	SODIUM NITRATE
CARBONIC ACID	H_2CO_3	HYDROGEN CARBONATE	Na_2CO_3	SODIUM CARBONATE
ACETIC ACID	$HC_2H_3O_2$	HYDROGEN ACETATE	$NaC_2H_3O_2$	SODIUM ACETATE
HYDROCHLORIC ACID	HCl	HYDROGEN CHLORIDE	$NaCl$	SODIUM CHLORIDE
HYDROSULFURIC ACID	H_2S	HYDROGEN SULFIDE	Na_2S	SODIUM SULFIDE
SULFUROUS ACID	H_2SO_3	HYDROGEN SULFITE	Na_2SO_3	SODIUM SULFITE
NITROUS ACID	HNO_2	HYDROGEN NITRITE	$NaNO_2$	SODIUM NITRITE
CHLOROUS ACID	$HClO_2$	HYDROGEN CHLORITE	$NaClO_2$	SODIUM CHLORITE

REMEMBER: **-IC** ACIDS FORM **-ATE** SALTS;
HYDRO- -IC ACIDS FORM **-IDE** SALTS; **-OUS** ACIDS FORM **-ITE** SALTS

HOUSEHOLD ITEMS CONTAINING SALTS

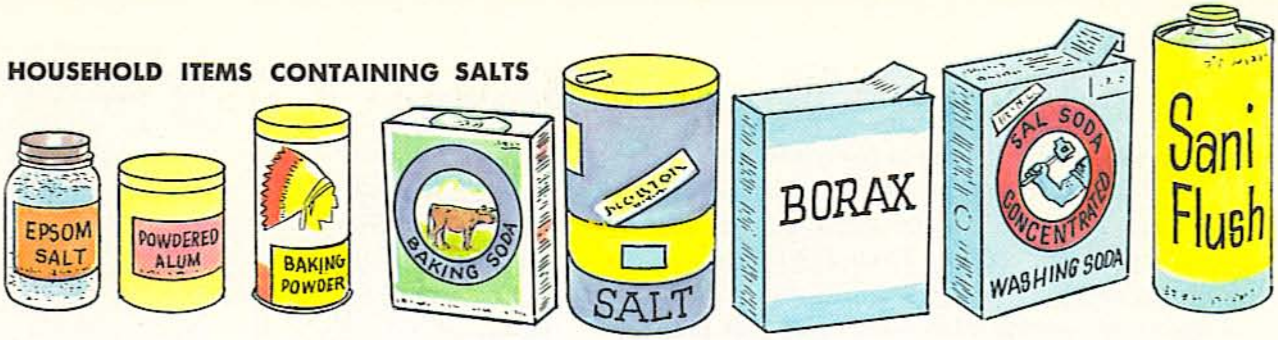

DIFFERENT WAYS OF PRODUCING SALTS

SALT FROM METAL AND ACID

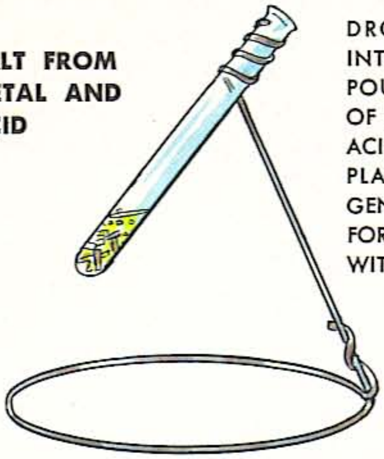

DROP ZINC STRIPS INTO A TEST TUBE. POUR IN A COUPLE OF ml HYDROCHLORIC ACID. THE ZINC DISPLACES THE HYDROGEN OF THE ACID TO FORM A SALT ($ZnCl_2$) WITH THE CHLORINE.

SALT FROM METAL OXIDE AND ACID

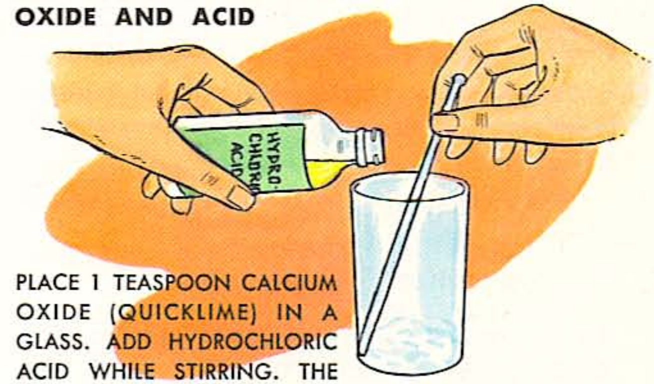

PLACE 1 TEASPOON CALCIUM OXIDE (QUICKLIME) IN A GLASS. ADD HYDROCHLORIC ACID WHILE STIRRING. THE QUICKLIME DISSOLVES IN THE ACID, FORMING CALCIUM CHLORIDE AND WATER.

SALT FROM ANOTHER SALT AND ACID

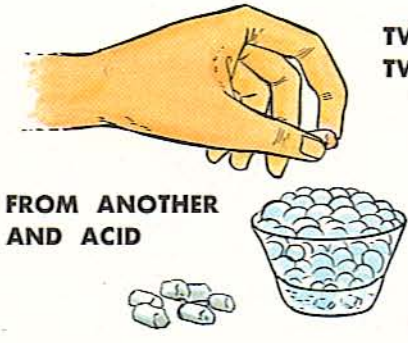

DROP PIECES OF CHALK, MARBLE, OR OYSTER SHELLS (ALL OF THEM CALCIUM CARBONATES) IN A FEW ml HYDROCHLORIC ACID. RESULT IS CALCIUM CHLORIDE AND CARBONIC ACID (WHICH BREAKS UP INTO CARBON DIOXIDE AND WATER).

TWO SALTS FROM TWO OTHER SALTS

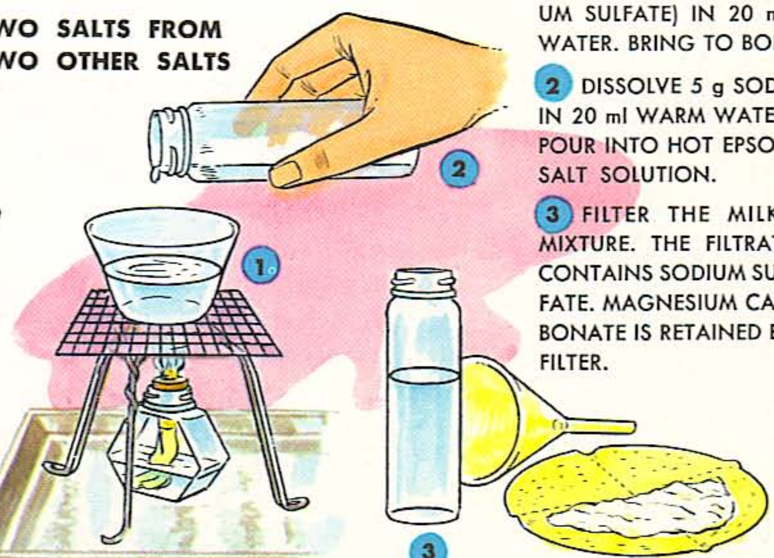

1 DISSOLVE 5 g EPSOM SALTS (MAGNESIUM SULFATE) IN 20 ml. WATER. BRING TO BOIL.

2 DISSOLVE 5 g SODA IN 20 ml WARM WATER. POUR INTO HOT EPSOM SALT SOLUTION.

3 FILTER THE MILKY MIXTURE. THE FILTRATE CONTAINS SODIUM SULFATE. MAGNESIUM CARBONATE IS RETAINED BY FILTER.

There are many ways of producing a salt in addition to neutralization.

When you made iron sulfide directly from the two elements iron and sulfur, you produced a salt:

$$Fe + S \rightarrow FeS$$

When you caused zinc metal to react with hydrochloric acid, you made a salt:

$$Zn + 2HCl \rightarrow ZnCl_2 + H_2$$

When you made sodium hydroxide, you used a base and a salt to form a new base and a new salt:

$$Ca(OH)_2 + Na_2CO_3 \rightarrow 2NaOH + CaCO_3$$

A salt and an acid often form another salt and another acid:

$$CaCO_3 + 2HCl \rightarrow$$
$$CaCl_2 + H_2CO_3 \enspace (H_2O + CO_2)$$

Two soluble salts may also form two other salts—one of them insoluble:

$$Na_2CO_3 + MgSO_4 \rightarrow MgCO_3 + Na_2SO_4$$

Iodine—Violet or Brown?

IODINE
Element 53.
At. wt. 126.91.
Gray-black crystals of a peculiar odor. Sublimes with violet color. Combines directly with metals and nonmetals. It has a density of 4.9.

IODINE IS an interesting element to experiment with. It is easily driven out of its compounds as beautiful, violet fumes that turn into grayish-black, metallic-looking crystals on cooling. These crystals can be further purified by turning them into vapor again, and again cooling them into crystal form. This process is called "sublimation."

You are probably familiar with the 2% alcoholic solution of iodine known as "tincture of iodine." It is found in almost every home medicine cabinet and is used as a disinfectant for wounds. Iodine has many other uses — in photography and in the preparation of various medicines and dyes.

Iodine has the bad habit of staining practically everything with which it comes in contact with a brown stain that won't come off in washing. That's why it is advisable to have sodium thiosulfate — photographer's fixing salt, "hypo"— around when you work with iodine. Hypo in solution forms a colorless compound with iodine.

IODIZED SALT
TINCTURE OF IODINE 2%

Be careful not to breathe fumes.

MAKING IODINE

① IN A PYREX CUSTARD CUP MIX TOGETHER 2 g POTASSIUM IODIDE, 2 g MANGANESE DIOXIDE, 4 g SODIUM BISULFATE. HEAT MIXTURE GENTLY. SOON VIOLET FUMES EMERGE.

② DROP HALF A DOZEN ICE CUBES INTO A JAR. ADD A LITTLE WATER. PLACE JAR AS A LID ON TOP OF CUSTARD CUP. THE VIOLET FUMES SETTLE ON BOTTOM OF JAR AS GRAYISH-BLACK, SHINY IODINE CRYSTALS.

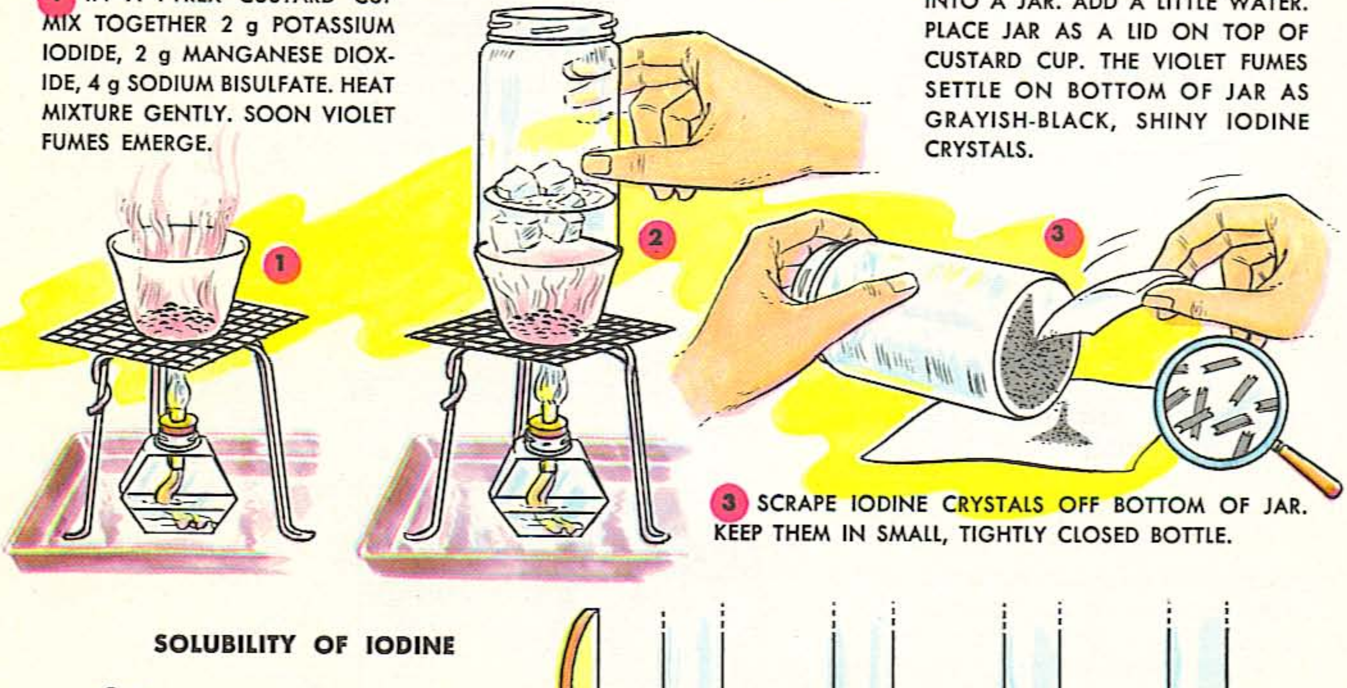

③ SCRAPE IODINE CRYSTALS OFF BOTTOM OF JAR. KEEP THEM IN SMALL, TIGHTLY CLOSED BOTTLE.

SOLUBILITY OF IODINE

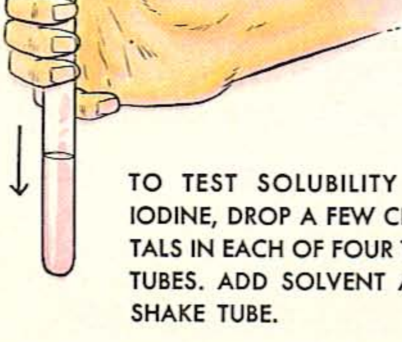

TO TEST SOLUBILITY OF IODINE, DROP A FEW CRYSTALS IN EACH OF FOUR TEST TUBES. ADD SOLVENT AND SHAKE TUBE.

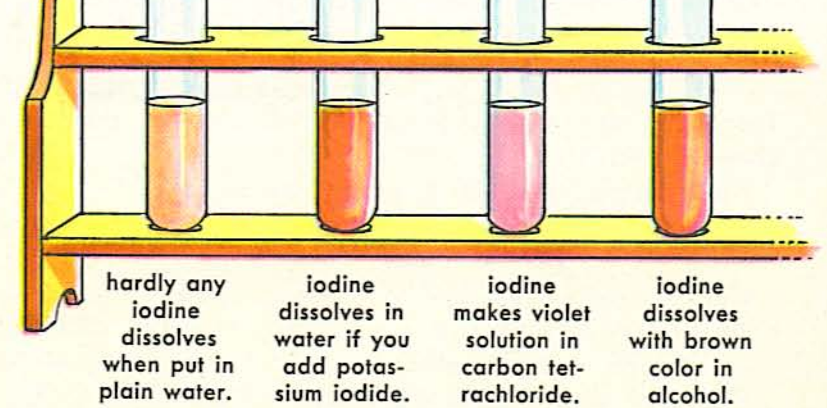

hardly any iodine dissolves when put in plain water.	iodine dissolves in water if you add potassium iodide.	iodine makes violet solution in carbon tetrachloride.	iodine dissolves with brown color in alcohol.

IODINE FREED BY CHLORINE

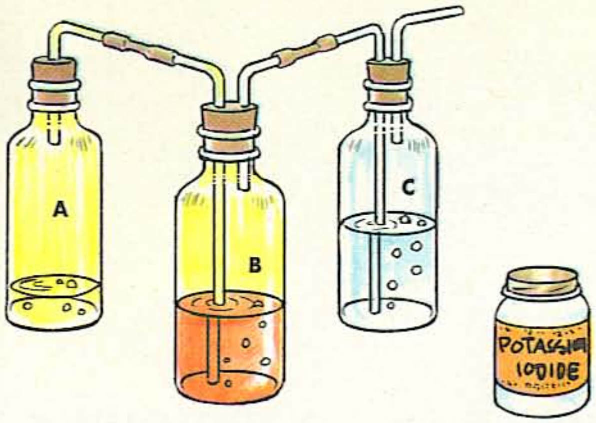

SET UP APPARATUS AS DESCRIBED ON PAGE 35 WITH THIS EXCEPTION: IN BOTTLE **B**, MAKE SOLUTION OF ½ g POTASSIUM IODIDE IN 40 ml WATER. AS CHLORINE BUBBLES THROUGH THIS SOLUTION IT TURNS BROWN FROM THE FREED IODINE. WITH MORE CHLORINE IT CLEARS AGAIN WHEN COLORLESS IODIC ACID FORMS.

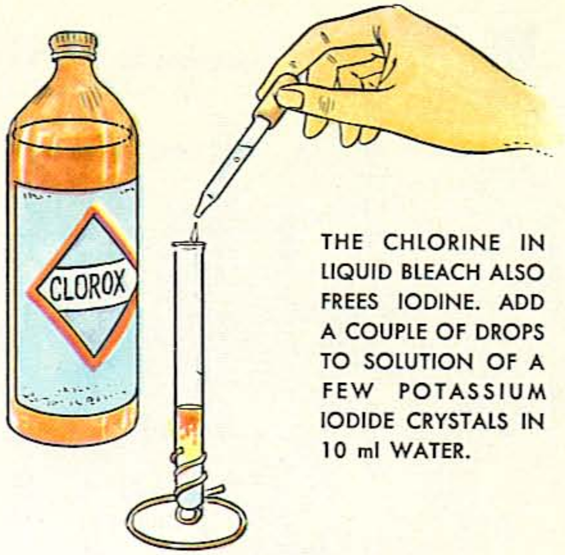

THE CHLORINE IN LIQUID BLEACH ALSO FREES IODINE. ADD A COUPLE OF DROPS TO SOLUTION OF A FEW POTASSIUM IODIDE CRYSTALS IN 10 ml WATER.

IODINE BY OXIDATION

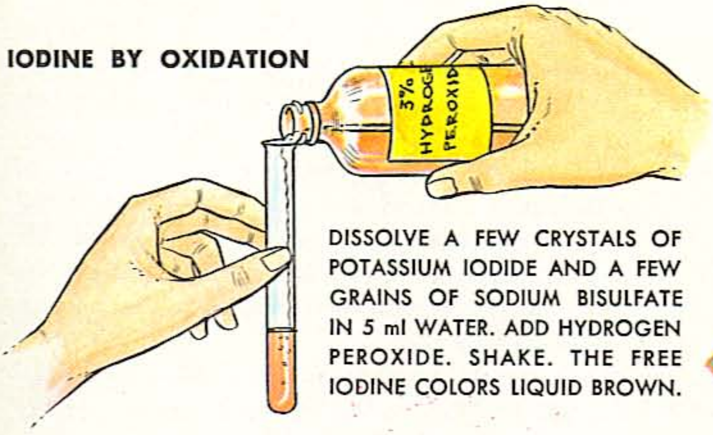

DISSOLVE A FEW CRYSTALS OF POTASSIUM IODIDE AND A FEW GRAINS OF SODIUM BISULFATE IN 5 ml WATER. ADD HYDROGEN PEROXIDE. SHAKE. THE FREE IODINE COLORS LIQUID BROWN.

MAKING HYDROGEN IODIDE

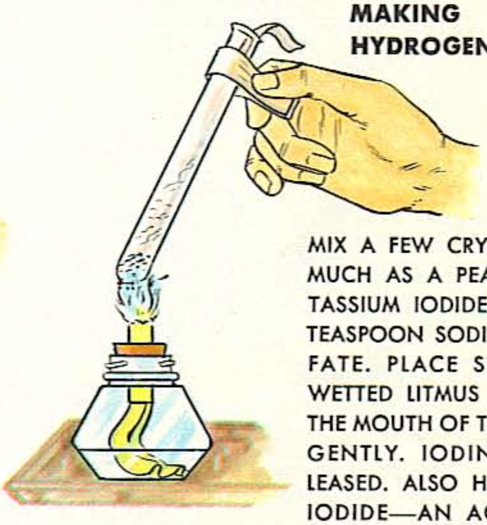

MIX A FEW CRYSTALS (AS MUCH AS A PEA) OF POTASSIUM IODIDE WITH ¼ TEASPOON SODIUM BISULFATE. PLACE STRIPS OF WETTED LITMUS PAPER AT THE MOUTH OF TUBE. HEAT GENTLY. IODINE IS RELEASED. ALSO HYDROGEN IODIDE—AN ACID THAT TURNS BLUE LITMUS RED.

STARCH TEST FOR IODINE

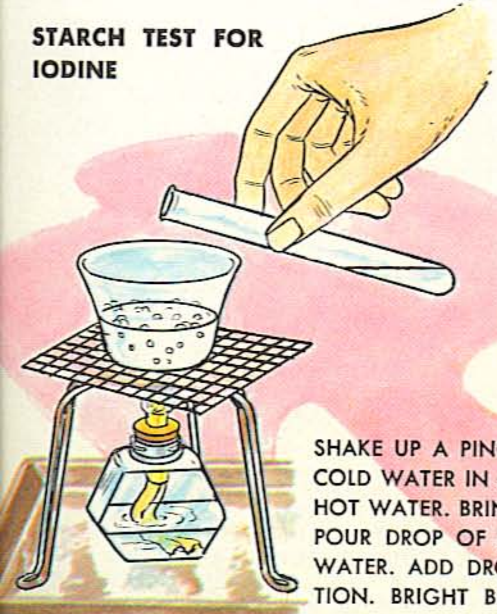

SHAKE UP A PINCH OF STARCH WITH COLD WATER IN A TEST TUBE. ADD TO HOT WATER. BRING TO A BOIL. COOL. POUR DROP OF MIXTURE INTO 10 ml WATER. ADD DROP OF IODINE SOLUTION. BRIGHT BLUE COLOR RESULTS.

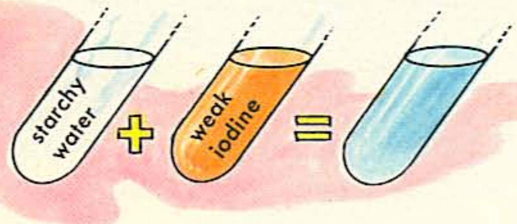

REMOVING IODINE STAIN

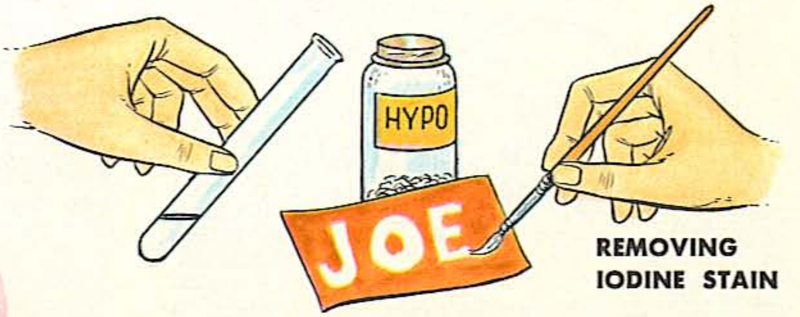

PAINT PAPER WITH IODINE. DISSOLVE A FEW CRYSTALS OF SODIUM THIOSULFATE ("HYPO") IN WATER. PAINT WITH THIS SOLUTION OVER THE BROWN COLOR. YOU WILL GET WHITE LETTERS AS HYPO FORMS COLORLESS COMPOUND WITH IODINE.

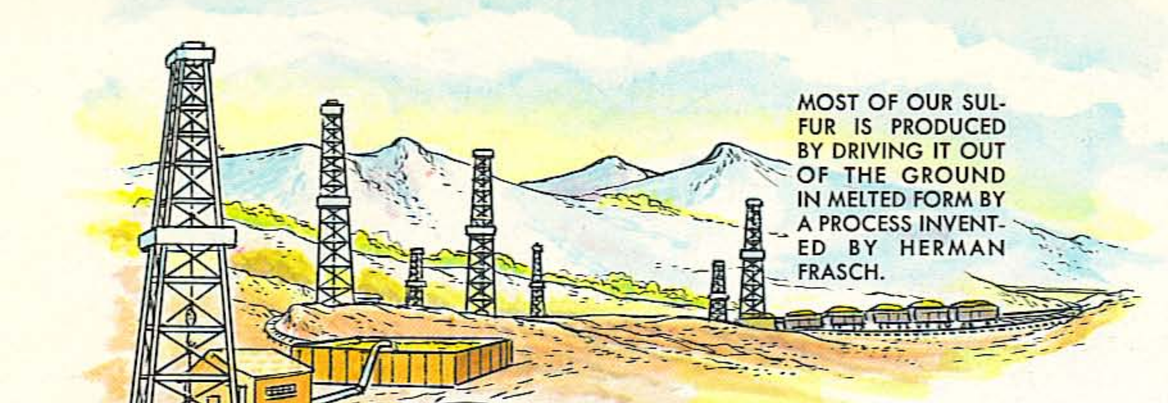
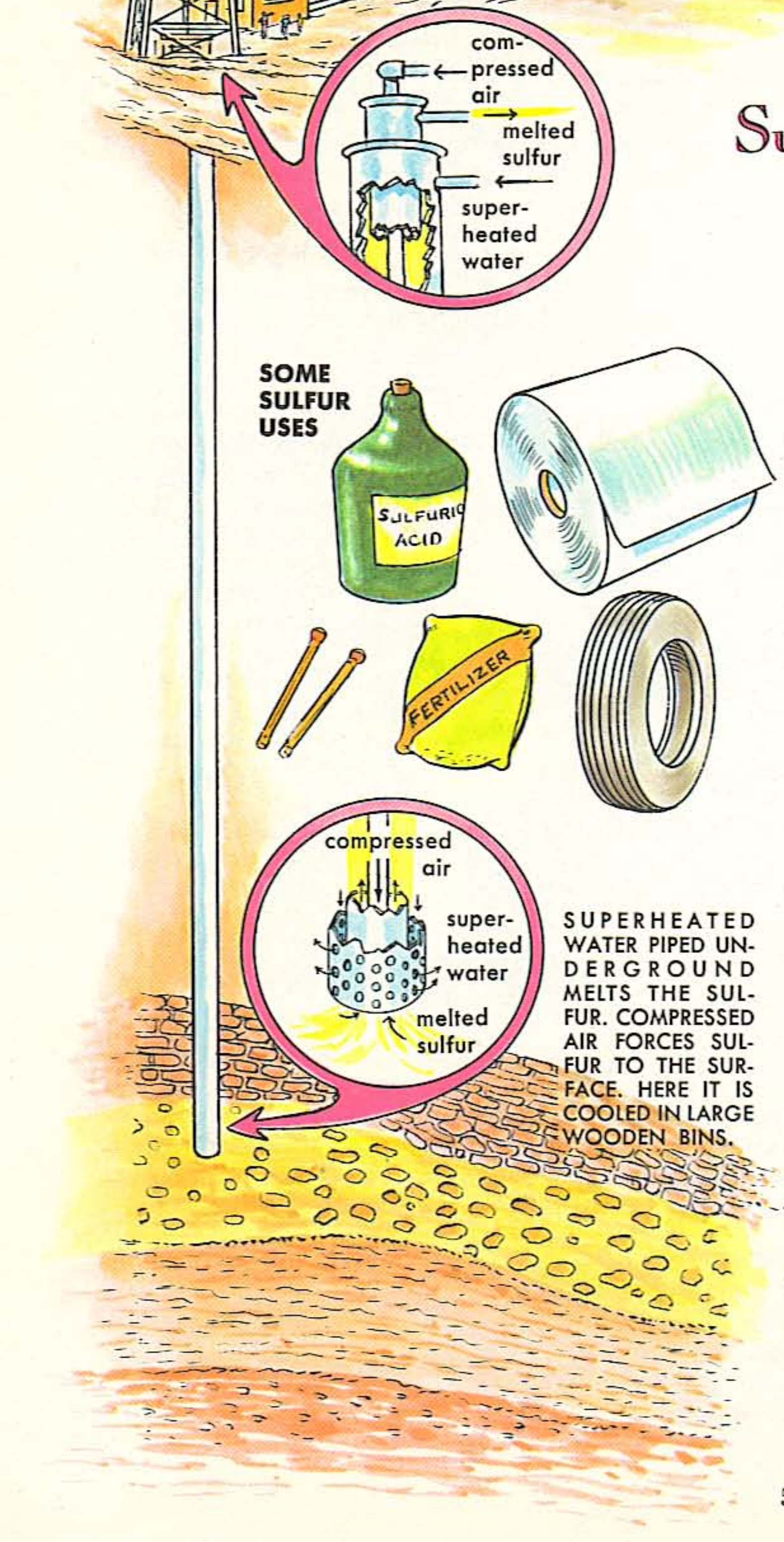
com-pressed air
melted sulfur
super-heated water
SOME SULFUR USES
SULFURIC ACID
FERTILIZER
compressed air
super-heated water
melted sulfur

MOST OF OUR SULFUR IS PRODUCED BY DRIVING IT OUT OF THE GROUND IN MELTED FORM BY A PROCESS INVENTED BY HERMAN FRASCH.

SUPERHEATED WATER PIPED UNDERGROUND MELTS THE SULFUR. COMPRESSED AIR FORCES SULFUR TO THE SURFACE. HERE IT IS COOLED IN LARGE WOODEN BINS.

S **SULFUR**
Element 16.
Atomic wt.: 32.066. Density: 2.07. Yellow crystals. Insoluble in water. Melts at 119°C. Boils at 444°C. Burns in air with blue flame.

Sulfur and Its Compounds

In the old days, sulfur was called "brimstone" ("burning stone" — from an old word, *brennen*, to burn). When it burned with a blue flame and a suffocating smell, people were certain that the devil himself was around.

Until fairly recently, most sulfur came from the volcanic Italian island of Sicily. But today, America produces most of the world's sulfur. About a hundred years ago, big deposits were found in Louisiana, several hundred feet underground. The problem of getting it up was solved in 1894 in a very clever way by a young German emigrant, Herman Frasch. He piped superheated water underground to melt the sulfur, then forced the melted sulfur to the top with compressed air.

Sulfur itself is used for many purposes. By a process called "vulcanization" it turns sticky, gummy raw rubber into elastic rubber usable for automobile tires and other rubber products. Sulfur also goes into such things as matches and gunpowder and medical preparations.

But by far the greatest use of sulfur is in the preparation of sulfuric acid (H_2SO_4). This acid enters into the

(CONTINUED ON PAGE 52)

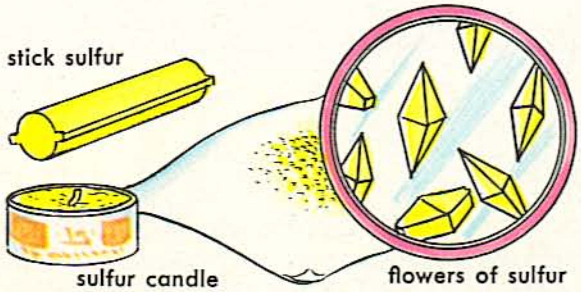
stick sulfur
sulfur candle
flowers of sulfur

SULFUR CAN USUALLY BE BOUGHT IN THREE DIFFERENT FORMS: AS STICK SULFUR, SULFUR CANDLES, AND AS A POWDER (FLOWERS OF SULFUR). UNDER MICROSCOPE, SULFUR POWDER PROVES TO BE RHOMBIC CRYSTALS.

MELTING SULFUR

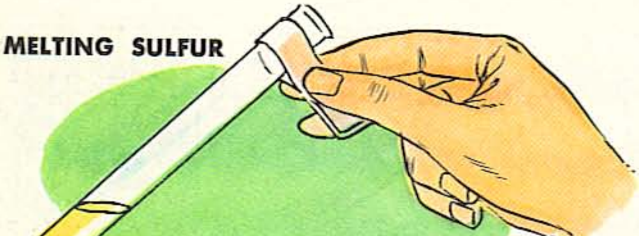

WHEN YOU MELT SULFUR, IT GOES THROUGH FOUR STAGES:

1. IT FIRST MELTS INTO A WATERY, STRAW-COLORED LIQUID.
2. IT NEXT BECOMES SLOW-FLOWING, CARAMEL-BROWN.
3. IT TURNS ALMOST SOLID.
4. IT BECOMES LIQUID AGAIN AND BOILS WITH YELLOW VAPOR.

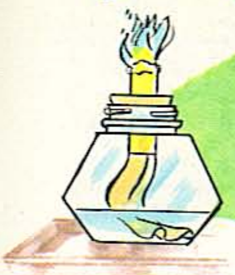

MAKING MONOCLINIC CRYSTALS OF SULFUR

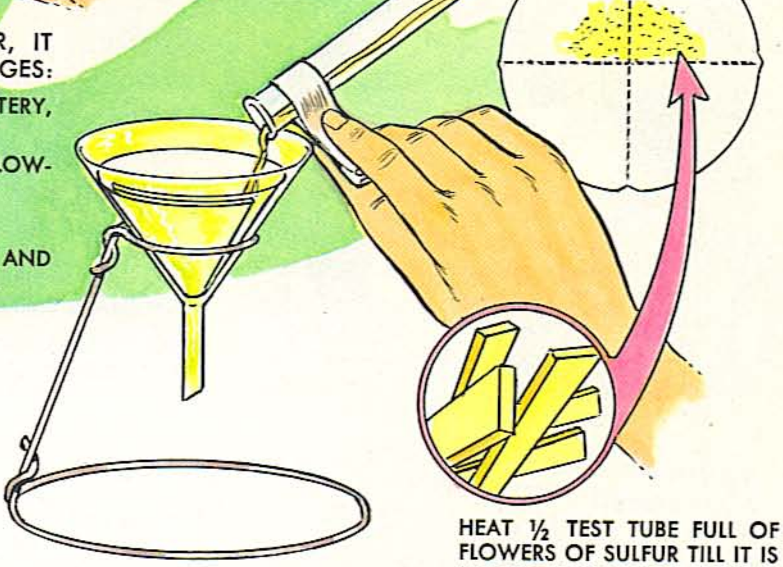

HEAT ½ TEST TUBE FULL OF FLOWERS OF SULFUR TILL IT IS MELTED WITH LIGHT COLOR. POUR MELTED SULFUR INTO A DRY FILTER. AS SOON AS CRUST FORMS ON TOP, OPEN UP FILTER PAPER. YOU WILL SEE THAT SULFUR HAS FORMED TINY NEEDLE-LIKE CRYSTALS.

CASTING WITH SULFUR

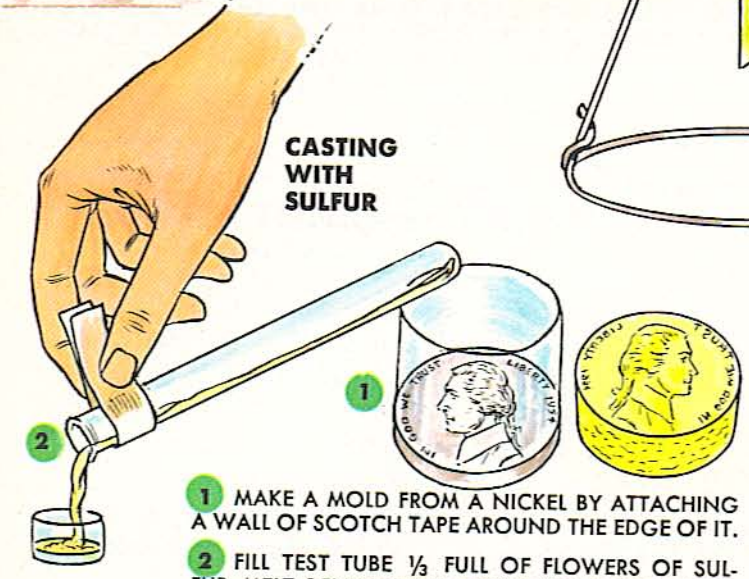

1 MAKE A MOLD FROM A NICKEL BY ATTACHING A WALL OF SCOTCH TAPE AROUND THE EDGE OF IT.

2 FILL TEST TUBE ⅓ FULL OF FLOWERS OF SULFUR. MELT GENTLY HIGH ABOVE FLAME. POUR INTO MOLD. WHEN COOLED YOU HAVE A PERFECT CAST.

PLASTIC SULFUR

1 MELT ½ TEST TUBE POWDERED SULFUR. CONTINUE HEATING. SOON IT NO LONGER FLOWS. YOU CAN TURN TUBE UPSIDE DOWN WITHOUT ANYTHING COMING OUT.

2 HEAT THE THICKENED SULFUR FURTHER UNTIL IT FLOWS FREELY AGAIN. THEN POUR THE DARK FLUID INTO COLD WATER. IT TURNS INTO A PLASTIC MASS. IN A FEW DAYS THIS AGAIN BECOMES YELLOW SULFUR.

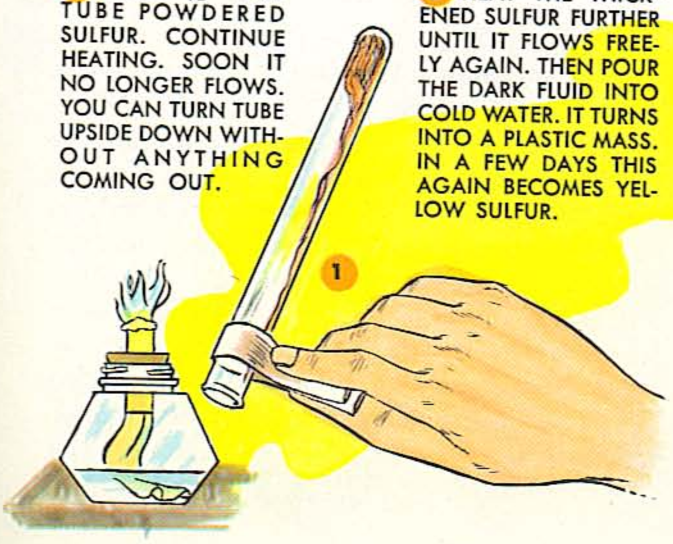

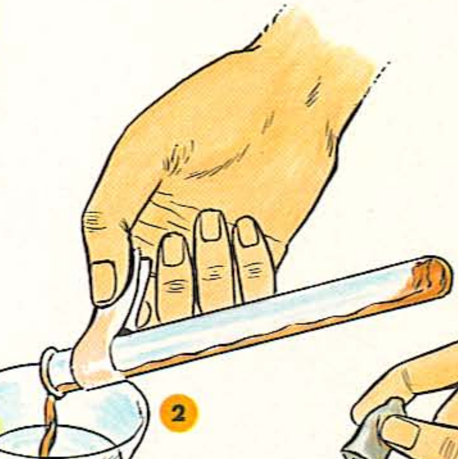

PRECIPITATED SULFUR

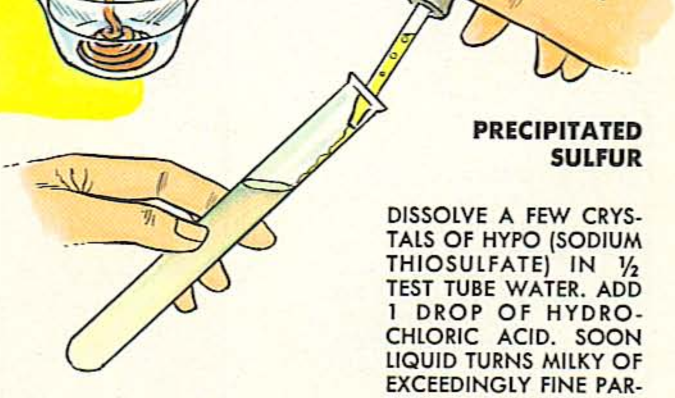

DISSOLVE A FEW CRYSTALS OF HYPO (SODIUM THIOSULFATE) IN ½ TEST TUBE WATER. ADD 1 DROP OF HYDROCHLORIC ACID. SOON LIQUID TURNS MILKY OF EXCEEDINGLY FINE PARTICLES OF SULFUR.

SULFUR DIOXIDE FOR BLEACHING

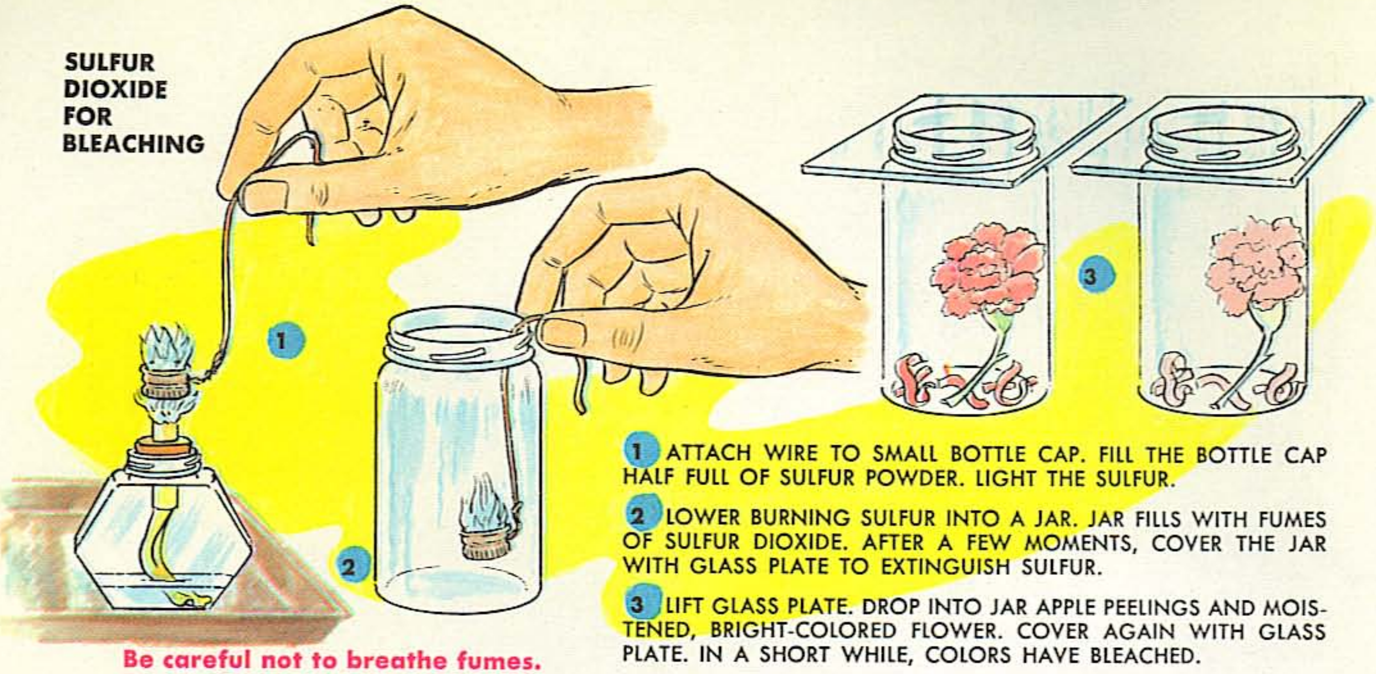

1 ATTACH WIRE TO SMALL BOTTLE CAP. FILL THE BOTTLE CAP HALF FULL OF SULFUR POWDER. LIGHT THE SULFUR.

2 LOWER BURNING SULFUR INTO A JAR. JAR FILLS WITH FUMES OF SULFUR DIOXIDE. AFTER A FEW MOMENTS, COVER THE JAR WITH GLASS PLATE TO EXTINGUISH SULFUR.

3 LIFT GLASS PLATE. DROP INTO JAR APPLE PEELINGS AND MOISTENED, BRIGHT-COLORED FLOWER. COVER AGAIN WITH GLASS PLATE. IN A SHORT WHILE, COLORS HAVE BLEACHED.

Be careful not to breathe fumes.

MAKING SULFUROUS ACID

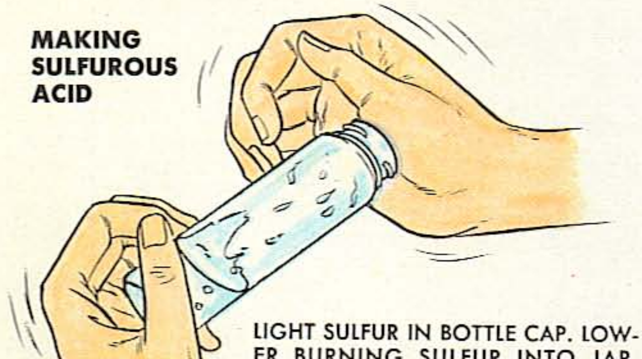

LIGHT SULFUR IN BOTTLE CAP. LOWER BURNING SULFUR INTO JAR. WHEN JAR IS FULL OF FUMES, REMOVE SULFUR. ADD A FEW ml WATER. SHAKE. AS SO_2 DISSOLVES IN WATER IT FORMS A WEAK ACID—SULFUROUS ACID, H_2SO_3. TEST FOR ACID WITH BLUE LITMUS PAPER.

SULFUR DIOXIDE FROM A SALT

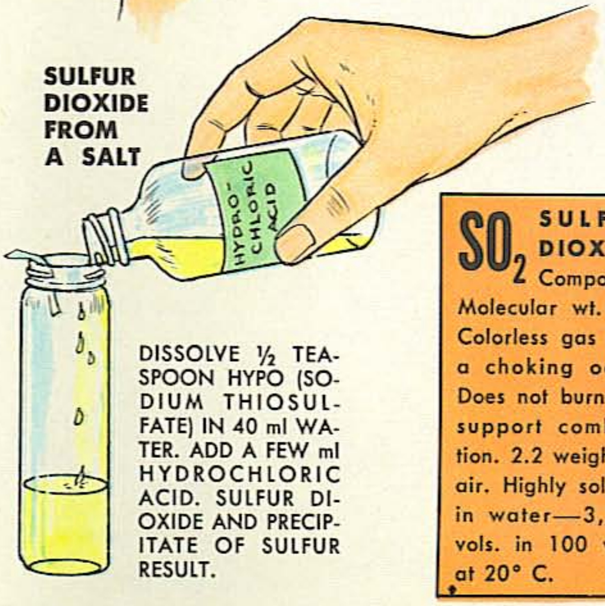

DISSOLVE ½ TEASPOON HYPO (SODIUM THIOSULFATE) IN 40 ml WATER. ADD A FEW ml HYDROCHLORIC ACID. SULFUR DIOXIDE AND PRECIPITATE OF SULFUR RESULT.

SO_2 SULFUR DIOXIDE Compound. Molecular wt. 64. Colorless gas with a choking odor. Does not burn nor support combustion. 2.2 weight of air. Highly soluble in water—3,937 vols. in 100 vols. at 20° C.

Sulfur—Continued

production — directly or indirectly — of practically every manufactured article we use today. It is used in refining gasoline, in making steel and paper, fibers and films, plastics and explosives, and thousands of other chemicals.

Sulfur Dioxide — The first step in making sulfuric acid from sulfur is to burn the sulfur.

When burning in the air, each atom of sulfur takes on two atoms of oxygen to make one molecule of sulfur dioxide gas (SO_2).

By a special, complicated process, sulfur dioxide can be forced to take on another oxygen atom and form sulfur trioxide (SO_3). With water, this makes sulfuric acid:

$$H_2O + SO_3 \rightarrow H_2SO_4$$

Hydrogen Sulfide — Many sulfur compounds have unpleasant, penetrating smells. Some of these compounds have very complex molecules — just imagine a skunk producing a chemical with this formula: $CH_3CH_2CH_2CH_2SH$! The smell of rotten eggs, on the other hand, comes from the simple compound hydrogen sulfide (H_2S).

Hydrogen sulfide is used in chemical analysis to determine what metals are found in a certain substance. It combines with metals into salts (sulfides) that can be distinguished from each other by their colors and by the way they react with acids and other chemicals.

NOTE: Perform these experiments out-of-doors or before an open window. Be careful not to breathe fumes.

H_2S HYDROGEN SULFIDE Compound. Molecular wt. 34. Colorless gas with odor of rotten eggs. Burns with a blue flame to form SO_2. 1.17 weight of air. Fairly soluble in water—258 vols. in 100 vols. at 20° C.

HYDROGEN SULFIDE

IS AN IMPORTANT LAB TOOL FOR CHEMICAL ANALYSIS.

HYDROGEN SULFIDE HAS SMELL OF ROTTEN EGGS

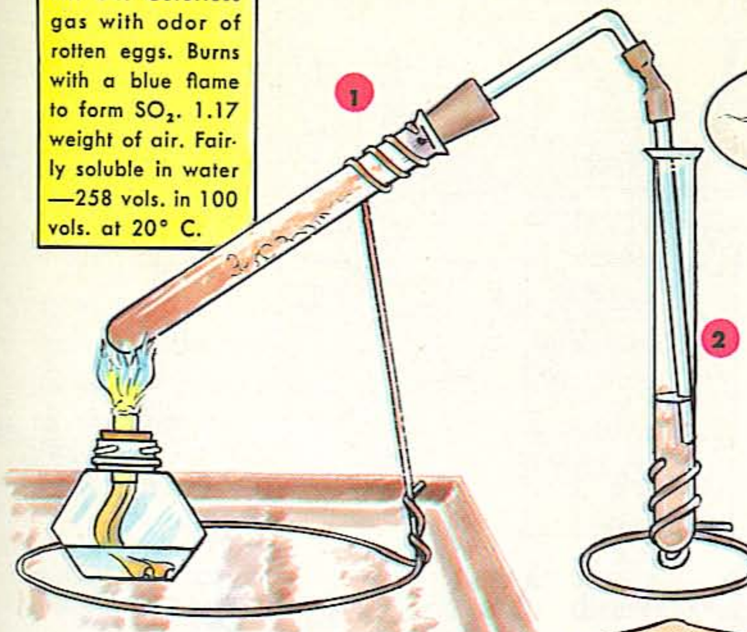

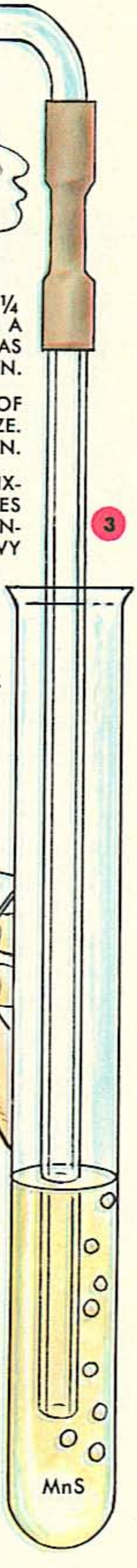

① DROP INTO A DRY TEST TUBE ¼ TEASPOON POWDERED SULFUR AND A LUMP OF CANDLE WAX AS LARGE AS A PEA. SET UP APPARATUS AS SHOWN.

② MAKE SOLUTION IN TEST TUBE OF CHEMICAL YOU WANT TO ANALYZE. LEAD GLASS TUBE INTO THE SOLUTION.

③ HEAT TEST TUBE WITH SULFUR MIXTURE. HYDROGEN SULFIDE BUBBLES INTO TEST SOLUTION. IF THIS CONTAINS SALT OF ONE OF THE HEAVY METALS, A PRECIPITATE WILL FORM.

HYDROGEN SULFIDE FROM FeS

MAKE IRON SULFIDE AS DESCRIBED ON PAGE 22. BREAK THE TEST TUBE (IN A PAPER BAG). CRUSH THE FeS WITH A HAMMER. DROP SMALL PIECES IN ANOTHER TEST TUBE. BY ADDING HYDROCHLORIC ACID YOU MAKE H_2S.

EXPERIMENTS WITH H_2S

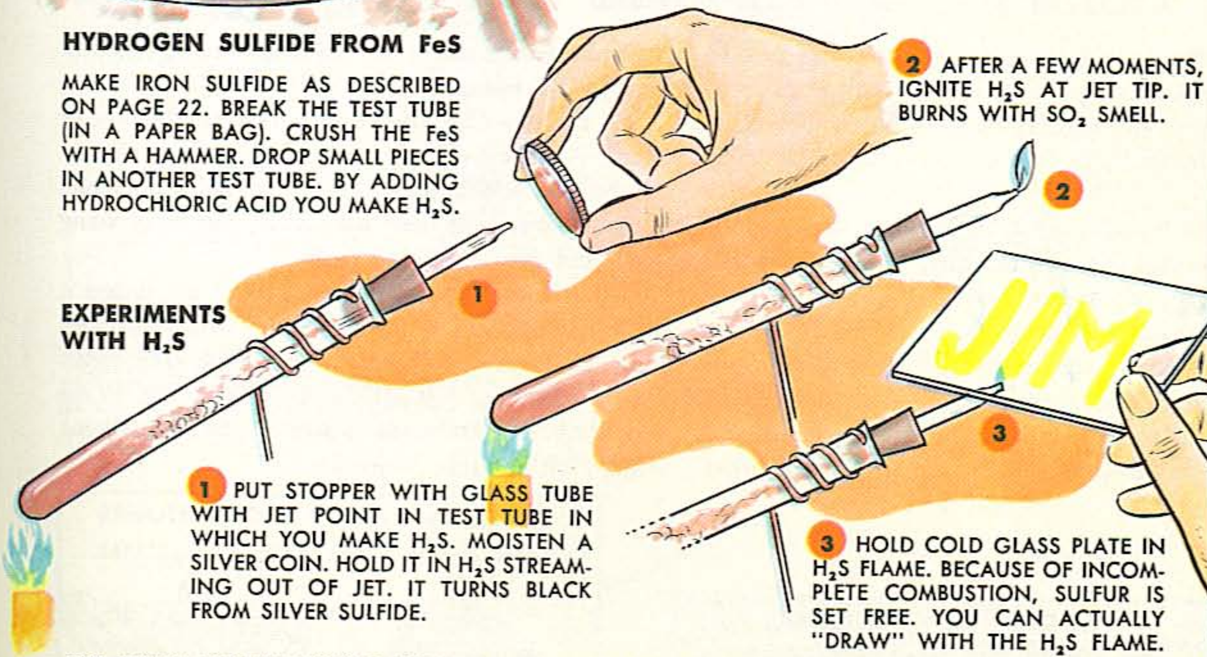

① PUT STOPPER WITH GLASS TUBE WITH JET POINT IN TEST TUBE IN WHICH YOU MAKE H_2S. MOISTEN A SILVER COIN. HOLD IT IN H_2S STREAMING OUT OF JET. IT TURNS BLACK FROM SILVER SULFIDE.

② AFTER A FEW MOMENTS, IGNITE H_2S AT JET TIP. IT BURNS WITH SO_2 SMELL.

③ HOLD COLD GLASS PLATE IN H_2S FLAME. BECAUSE OF INCOMPLETE COMBUSTION, SULFUR IS SET FREE. YOU CAN ACTUALLY "DRAW" WITH THE H_2S FLAME.

THE COLOR OF THE SULFIDE FORMED WHEN YOU LEAD H_2S INTO A SOLUTION CONTAINING A SALT OF A HEAVY METAL WILL HELP TELL YOU WHAT METAL IS FOUND IN THE SALT.

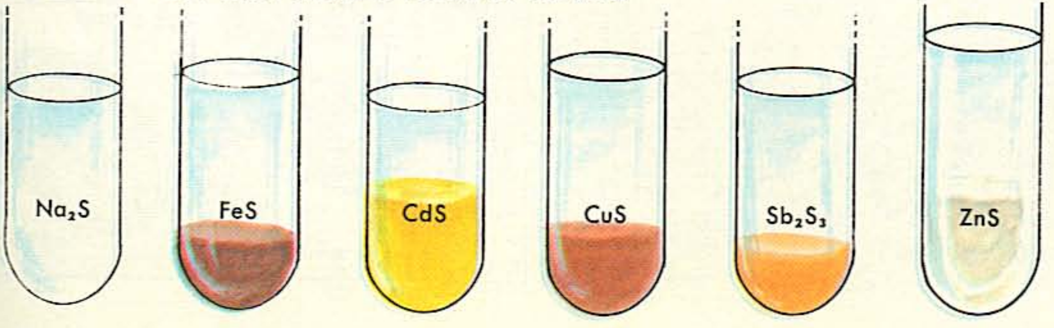

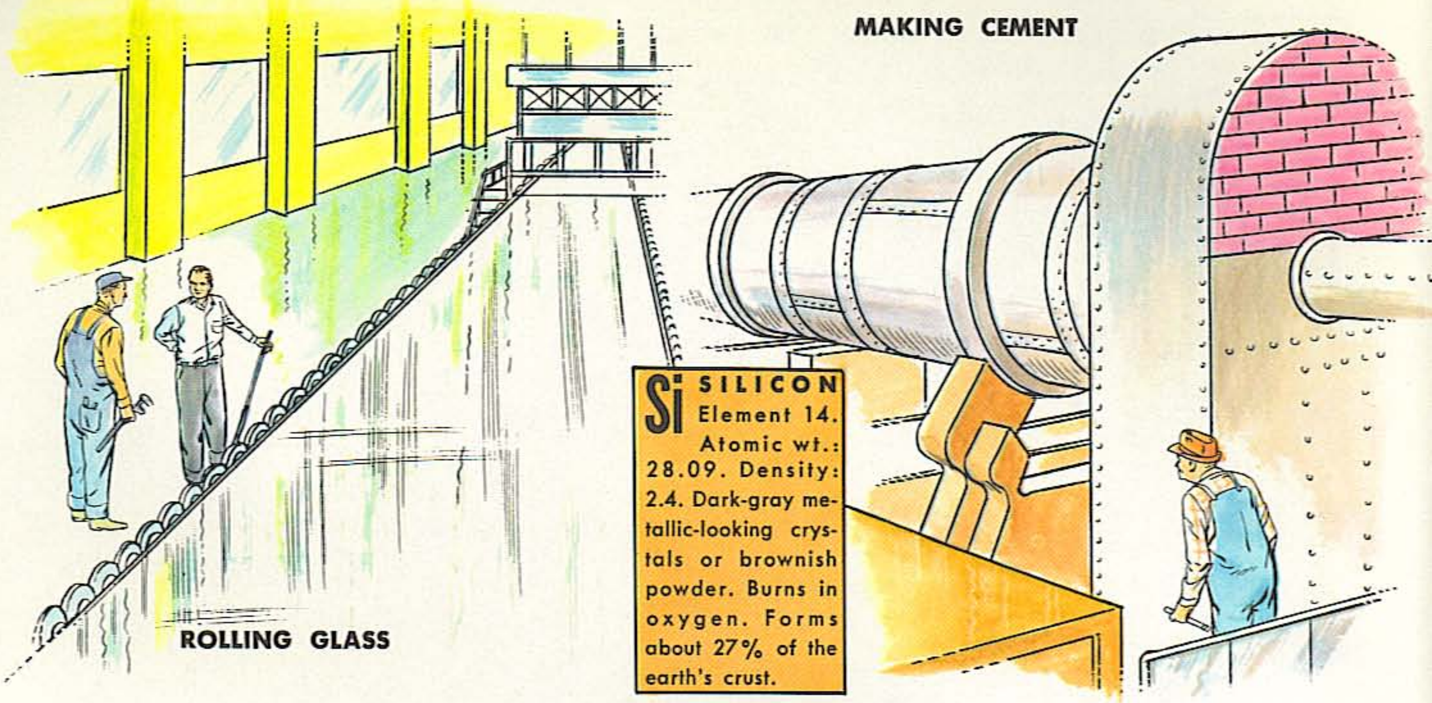

Si **SILICON** Element 14. Atomic wt.: 28.09. Density: 2.4. Dark-gray metallic-looking crystals or brownish powder. Burns in oxygen. Forms about 27% of the earth's crust.

Silicon—The Element You Step On

SILICON (from the Latin *silex*, flint) is the second most abundant element on earth — after oxygen. Whether you are walking on sand or clay, rock or cement, almost half of what you're stepping on is silicon.

Silicon is found in nature in combination with oxygen (mostly the dioxide, SiO_2) and in different silicates (salts of various silicic acids).

With few exceptions, silicon compounds are insoluble in water. And that is a good thing for all of us. The glass of our windows and the glasses from which we drink are silicates. So are the glazes on our cups and the enamel on our bathtubs. Most glass and many glazes are made by fusing together sand (SiO_2), limestone, and soda.

The silicates of sodium and potassium dissolve in water. A concentrated solution of sodium silicate (Na_2SiO_3) is sold in hardware stores under the name of waterglass. It is used as a glue, for fireproofing wood and for preserving eggs.

Within recent years, chemists have developed a whole line of new silicon compounds called silicones. Some of them are oil-like. Some look like putty ("Silly Putty"). Still others are rubber-like. Paper and cloth can be made water-repellent by being treated with suitable silicones.

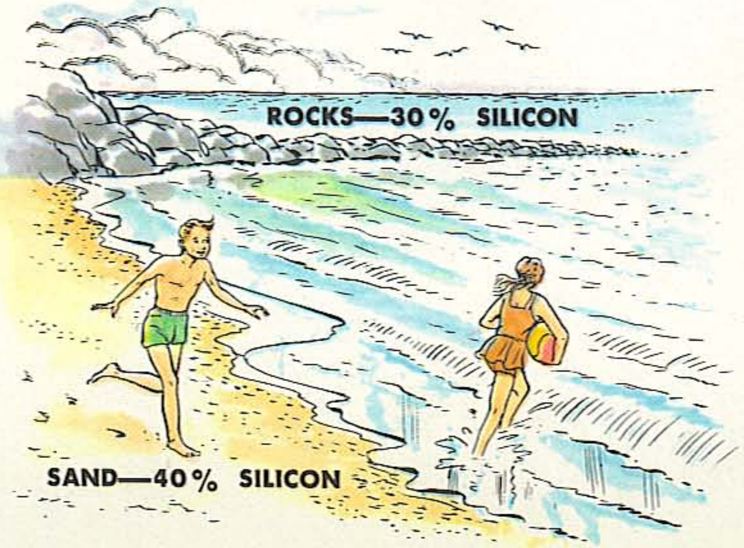

MINERALS AND PRECIOUS STONES

MANY ARE SILICON OXIDES OR SILICATES.

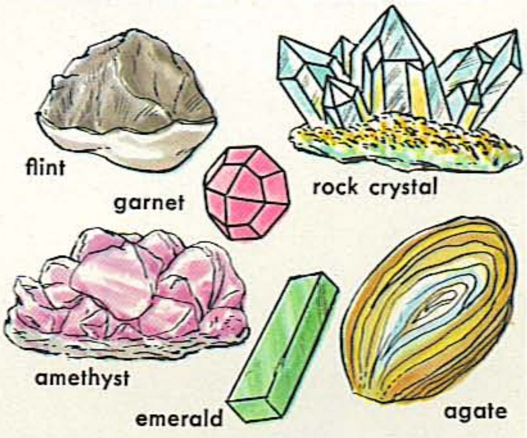

MAKING SILICIC ACID

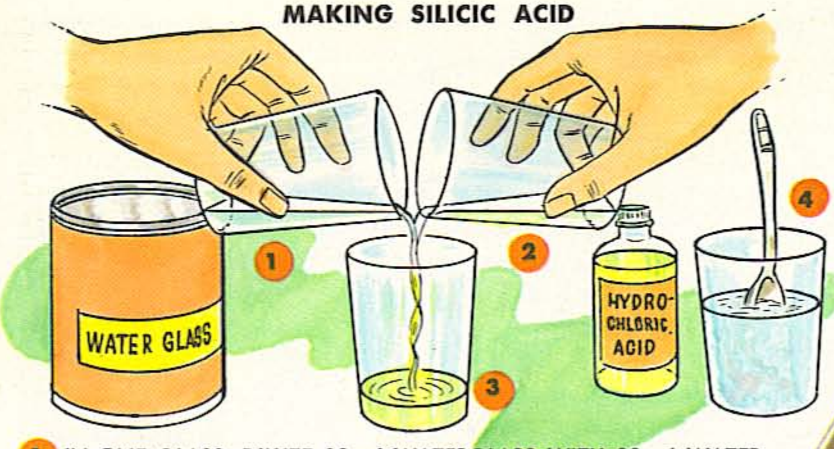

1 IN ONE GLASS, DILUTE 20 ml WATERGLASS WITH 20 ml WATER.

2 IN ANOTHER, MIX 10 ml HYDROCHLORIC ACID AND 10 ml WATER.

3 POUR THE TWO MIXTURES AT ONE TIME INTO A THIRD GLASS.

4 STAND SPOON UPRIGHT IN THE MIXTURE WHICH, ALMOST IMMEDIATELY, TURNS INTO A JELLY ("GEL") SO STIFF THAT SPOON STANDS BY ITSELF AND YOU CAN TURN THE GLASS UPSIDE DOWN.

MAKING SILICON DIOXIDE

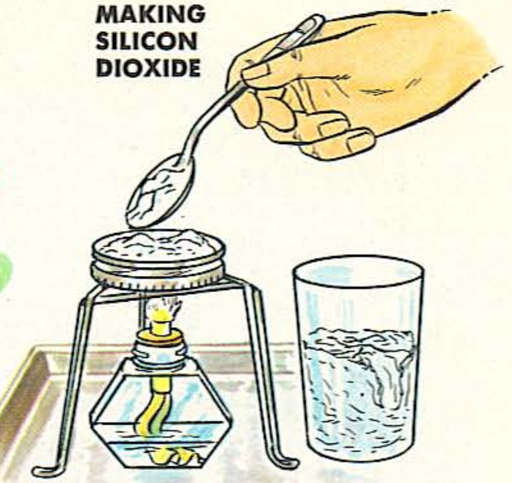

PLACE SOME OF THE GEL ON A METAL JAR LID. HEAT. THE SILICIC ACID (H_2SiO_3) GIVES UP WATER (H_2O) AND TURNS INTO A GRAYISH-WHITE POWDER OF SILICON DIOXIDE (SiO_2).

WEAKNESS OF SILICIC ACID

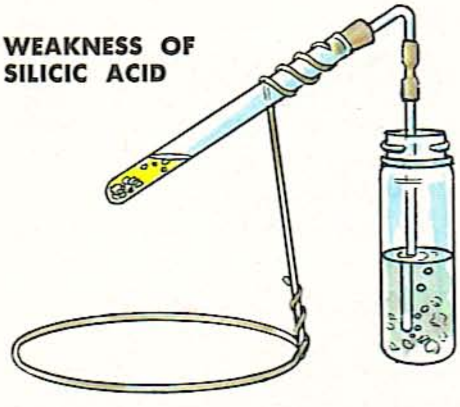

SILICIC ACID IS SO WEAK THAT CARBONIC ACID (H_2CO_3) DRIVES IT OUT OF WATERGLASS. MAKE THE CO_2 BY POURING HYDROCHLORIC ACID OVER MARBLE CHIPS.

MAKING WATER-GLASS

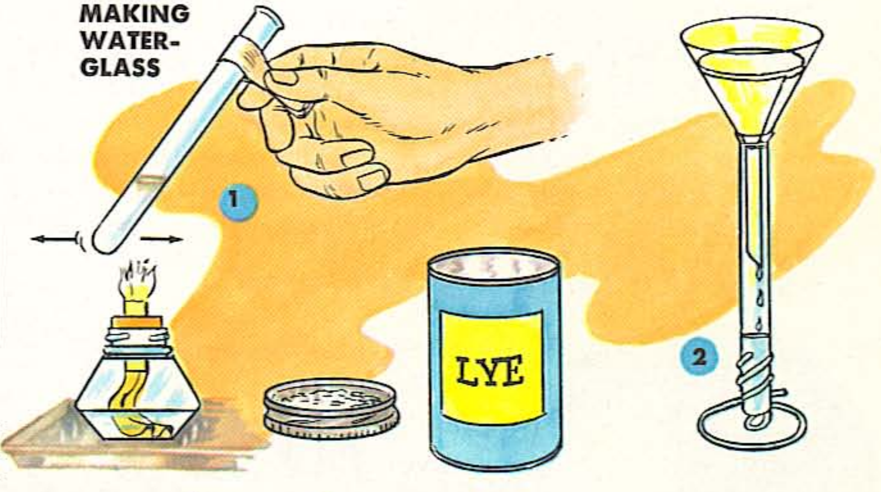

1 IN A TEST TUBE, MIX 1 g OF THE SILICON DIOXIDE YOU MADE, 2 g LYE (NaOH), AND 5 ml WATER. HEAT CAREFULLY, MOVING TUBE.

2 AFTER FILTERING, YOU WILL HAVE A CLEAR SOLUTION OF SODIUM SILICATE (Na_2SiO_3).

MAKING SILICATES

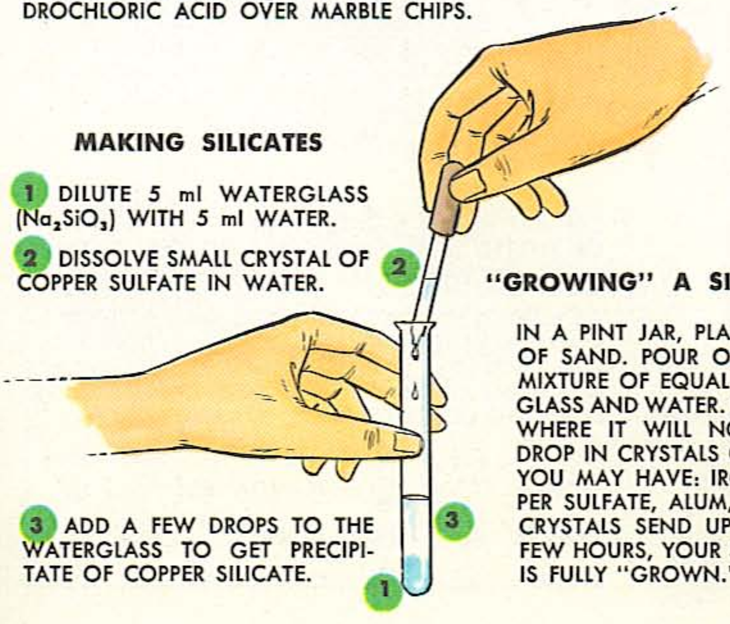

1 DILUTE 5 ml WATERGLASS (Na_2SiO_3) WITH 5 ml WATER.

2 DISSOLVE SMALL CRYSTAL OF COPPER SULFATE IN WATER.

3 ADD A FEW DROPS TO THE WATERGLASS TO GET PRECIPITATE OF COPPER SILICATE.

"GROWING" A SILICON "JUNGLE"

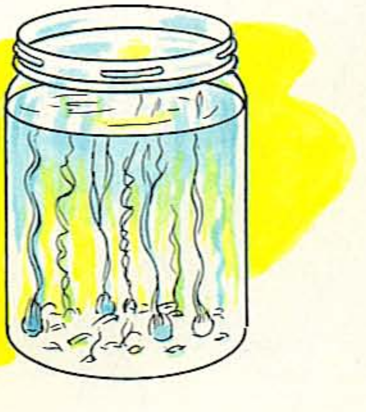

IN A PINT JAR, PLACE ½-INCH LAYER OF SAND. POUR ON TOP OF THIS A MIXTURE OF EQUAL PARTS OF WATERGLASS AND WATER. PLACE IT IN A SPOT WHERE IT WILL NOT BE DISTURBED. DROP IN CRYSTALS OF VARIOUS SALTS YOU MAY HAVE: IRON SULFATE, COPPER SULFATE, ALUM, EPSOM SALT. THE CRYSTALS SEND UP "SHOOTS." IN A FEW HOURS, YOUR SILICATE "JUNGLE" IS FULLY "GROWN."

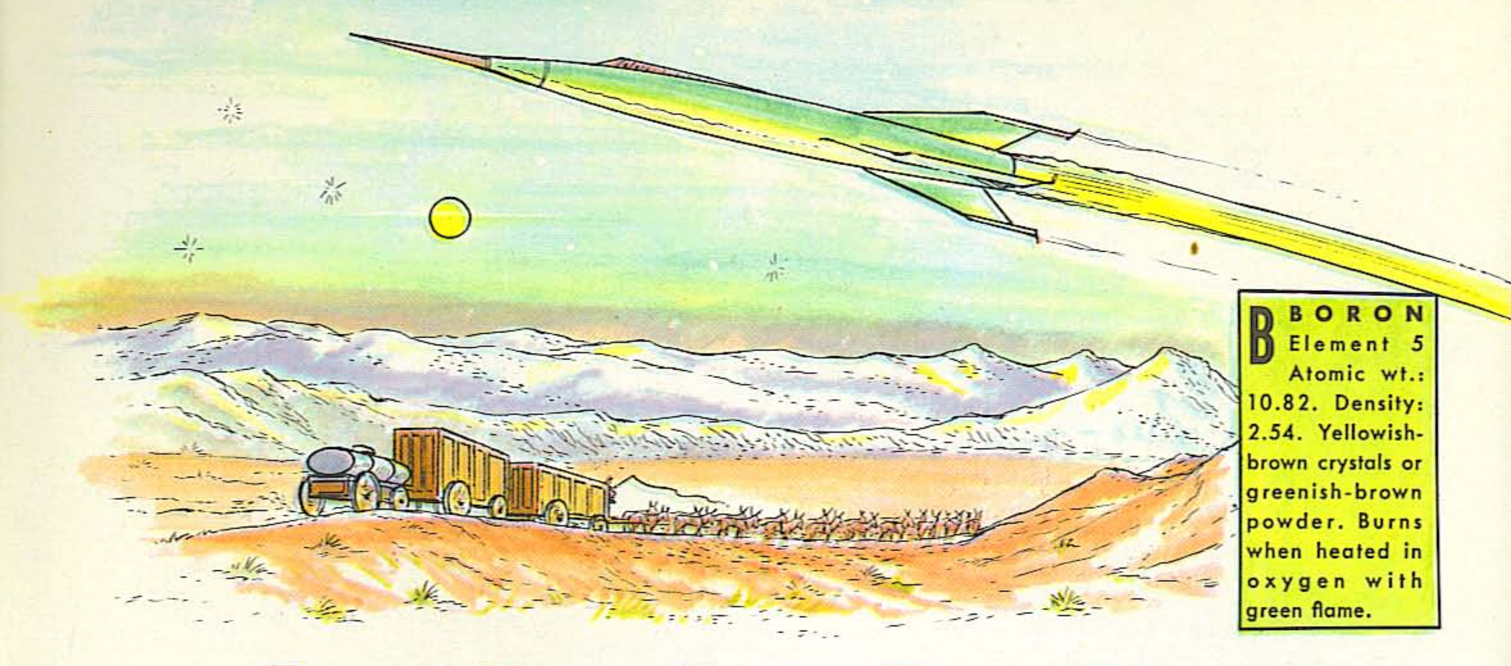

B **BORON**
Element 5
Atomic wt.: 10.82. Density: 2.54. Yellowish-brown crystals or greenish-brown powder. Burns when heated in oxygen with green flame.

Boron—Future Rocket-Power Element?

Less than a hundred years ago, a mineral called borax, containing the element boron, was carted out of Death Valley in California by twenty-mule teams — about the slowest transportation you can think of. Someday, boron may be put in zip-fuels for space missiles — the fastest form of transportation imaginable. Boron has the ability (as does carbon) to combine with hydrogen in a number of ways. When these boranes or boron hydrides burn, they develop a tremendous amount of power.

Boron can be isolated as a hard, brownish-black powder. Its carbon compound, boron carbide (B_4C), is almost as hard as diamond.

But you are probably more familiar with boron

BORAX BEAD TEST IN CHEMICAL ANALYSIS

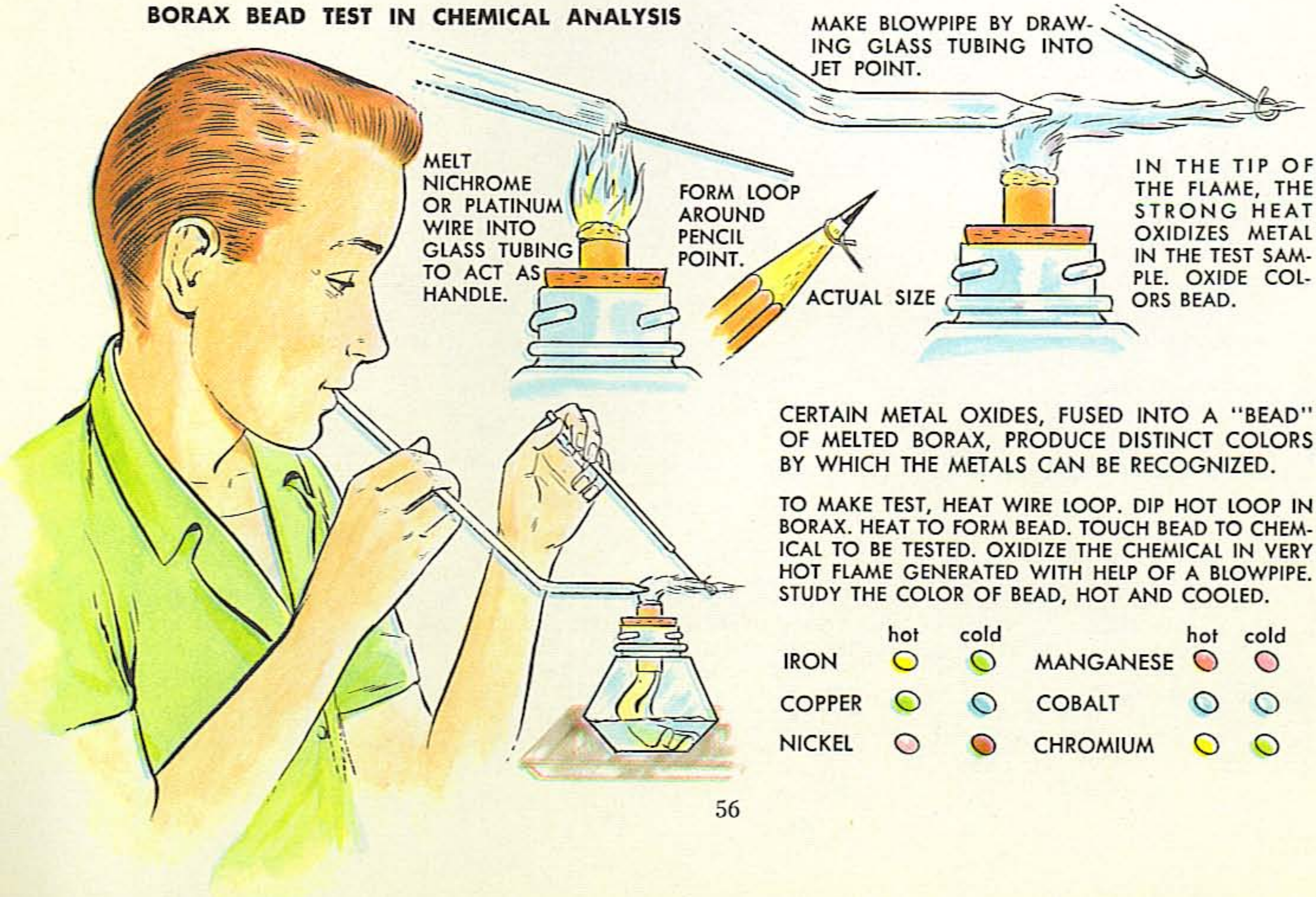

MAKE BLOWPIPE BY DRAWING GLASS TUBING INTO JET POINT.

MELT NICHROME OR PLATINUM WIRE INTO GLASS TUBING TO ACT AS HANDLE.

FORM LOOP AROUND PENCIL POINT.

ACTUAL SIZE

IN THE TIP OF THE FLAME, THE STRONG HEAT OXIDIZES METAL IN THE TEST SAMPLE. OXIDE COLORS BEAD.

CERTAIN METAL OXIDES, FUSED INTO A "BEAD" OF MELTED BORAX, PRODUCE DISTINCT COLORS BY WHICH THE METALS CAN BE RECOGNIZED.

TO MAKE TEST, HEAT WIRE LOOP. DIP HOT LOOP IN BORAX. HEAT TO FORM BEAD. TOUCH BEAD TO CHEMICAL TO BE TESTED. OXIDIZE THE CHEMICAL IN VERY HOT FLAME GENERATED WITH HELP OF A BLOWPIPE. STUDY THE COLOR OF BEAD, HOT AND COOLED.

	hot	cold		hot	cold
IRON	🟡	🟢	MANGANESE	🟣	🔴
COPPER	🟢	🔵	COBALT	🔵	🔵
NICKEL	🟣	🟤	CHROMIUM	🟡	🟢

through two of its compounds which are found in almost every household: boric acid (H_3BO_3), used as a mild antiseptic, and borax (sodium tetraborate, $Na_2B_4O_7 \cdot 10H_2O$), used for cleaning purposes and as a water softener.

Borax has a great number of uses outside the home. It is used for soldering, for producing certain kinds of soap, and for making other boron compounds.

The glass industry uses large quantities of borax for making boron-aluminum-silicate glass. You know this kind of glass by its trade name, Pyrex. Kitchen utensils and laboratory ware made of Pyrex glass have the great advantage over ordinary glass that they can be placed directly on the fire and do not break so easily when they are subjected to sudden heating or cooling.

MAKING BORIC ACID

1 IN A CUSTARD CUP, ADD 6 g BORAX TO 15 ml WATER. BRING TO BOIL. STIR UNTIL BORAX IS DISSOLVED.

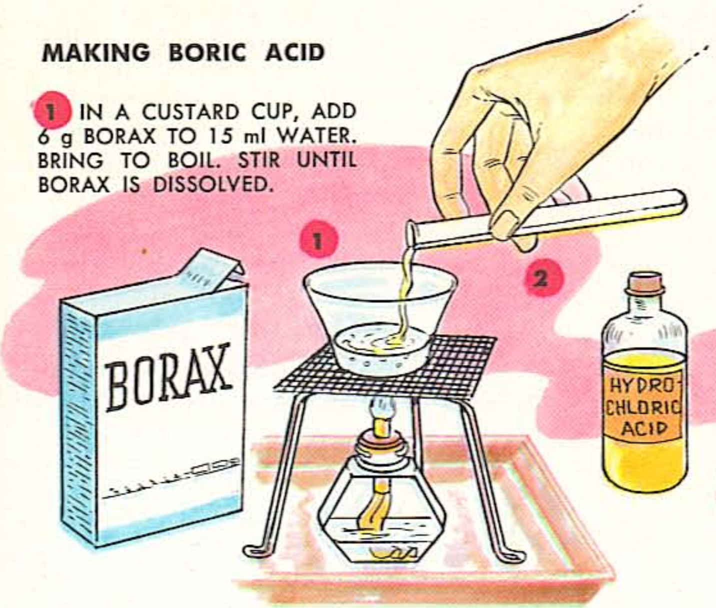

2 ADD 4 ml HYDROCHLORIC ACID TO HOT BORAX SOLUTION. STIR. REMOVE FROM FIRE. BORIC ACID CRYSTALLIZES OUT AS SOLUTION COOLS.

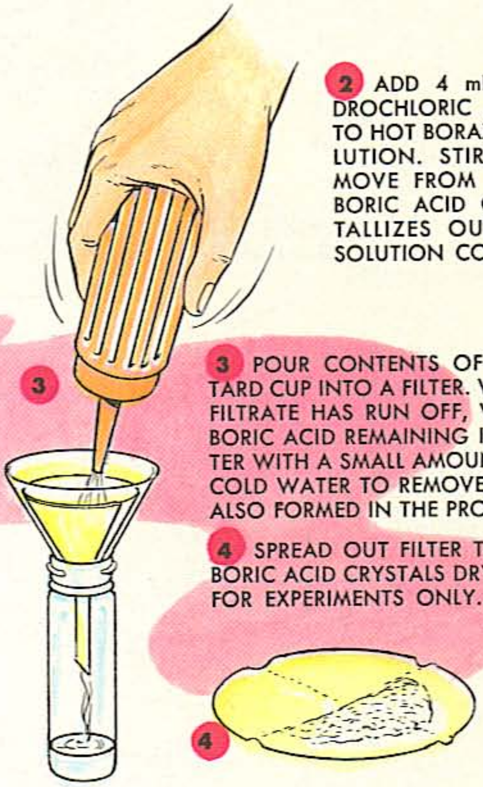

3 POUR CONTENTS OF CUSTARD CUP INTO A FILTER. WHEN FILTRATE HAS RUN OFF, WASH BORIC ACID REMAINING IN FILTER WITH A SMALL AMOUNT OF COLD WATER TO REMOVE NaCl ALSO FORMED IN THE PROCESS.

4 SPREAD OUT FILTER TO LET BORIC ACID CRYSTALS DRY. USE FOR EXPERIMENTS ONLY.

TESTS FOR BORIC ACID

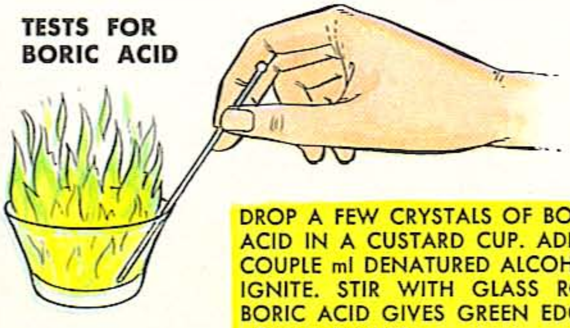

DROP A FEW CRYSTALS OF BORIC ACID IN A CUSTARD CUP. ADD A COUPLE ml DENATURED ALCOHOL. IGNITE. STIR WITH GLASS ROD. BORIC ACID GIVES GREEN EDGES TO THE FLAMES.

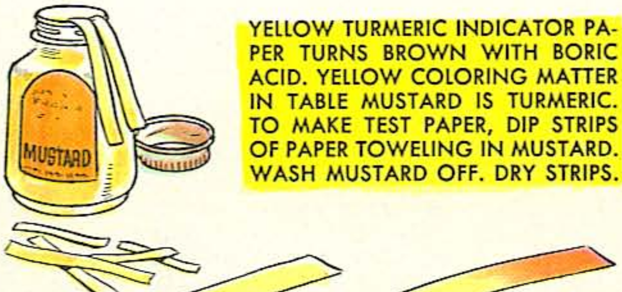

YELLOW TURMERIC INDICATOR PAPER TURNS BROWN WITH BORIC ACID. YELLOW COLORING MATTER IN TABLE MUSTARD IS TURMERIC. TO MAKE TEST PAPER, DIP STRIPS OF PAPER TOWELING IN MUSTARD. WASH MUSTARD OFF. DRY STRIPS.

MAKING BORIC OXIDE

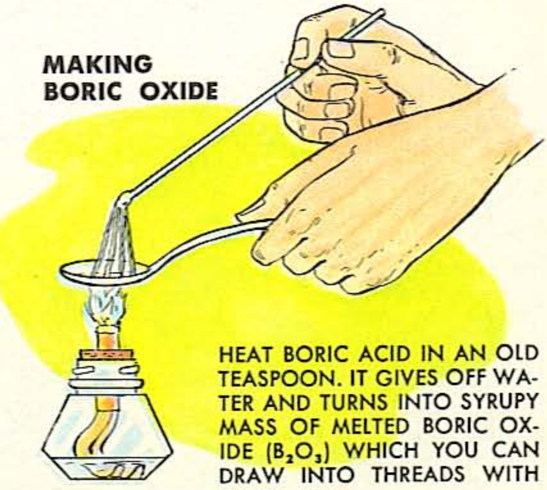

HEAT BORIC ACID IN AN OLD TEASPOON. IT GIVES OFF WATER AND TURNS INTO SYRUPY MASS OF MELTED BORIC OXIDE (B_2O_3) WHICH YOU CAN DRAW INTO THREADS WITH GLASS ROD.

Sodium and Potassium

The salts of sodium and potassium have been used for thousands of years in making soap and glass and for a great number of other purposes.

Sodium chloride (NaCl) is the most common sodium salt — it is the chemical that makes ocean water "salty." Plants growing in the ocean take up so much of the sodium that people along the seacoasts of the world used to burn dried seaweed to secure "soda ash" (sodium carbonate, Na_2CO_3). Inland plants, on the other hand, pick up potassium from the soil. Inland people boiled out wood ashes in large pots to get "potash" (potassium carbonate, K_2CO_3).

In 1807, the British scientist, Humphry Davy, succeeded in isolating the metals found in these salts. They proved to be wax-soft and silvery. He called them sodium (from soda ash) and potassium (from potash). These are still their English names. But in chemical formulas they are referred to as natrium (Na) and kalium (K) — from abbreviations of the Arabic names of the ashes: *natrun* and *al qili* (alkali).

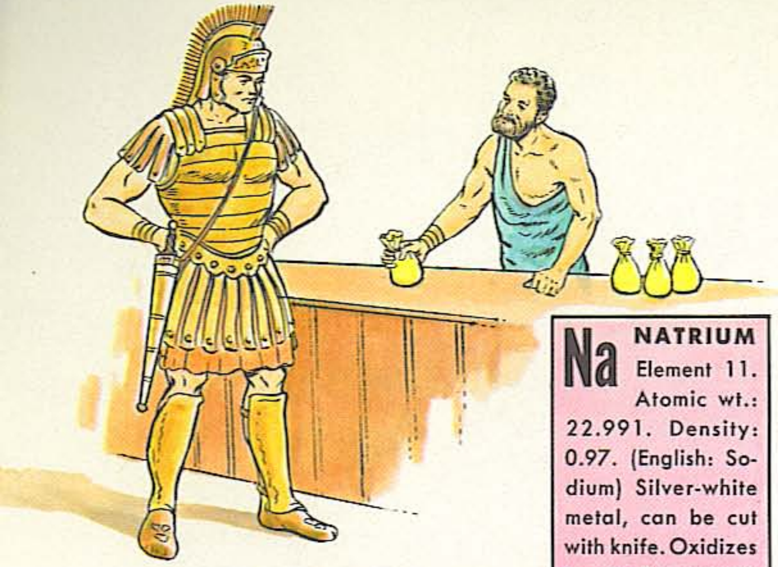

"SALARY" COMES FROM *SALARIUM* —THE WAGES PAID IN SALT TO ROMAN SOLDIERS.

Na **NATRIUM** Element 11. Atomic wt.: 22.991. Density: 0.97. (English: Sodium) Silver-white metal, can be cut with knife. Oxidizes in air. Reacts with water. Burns with yellow flame.

CRYSTALLIZING SALT BY EVAPORATION

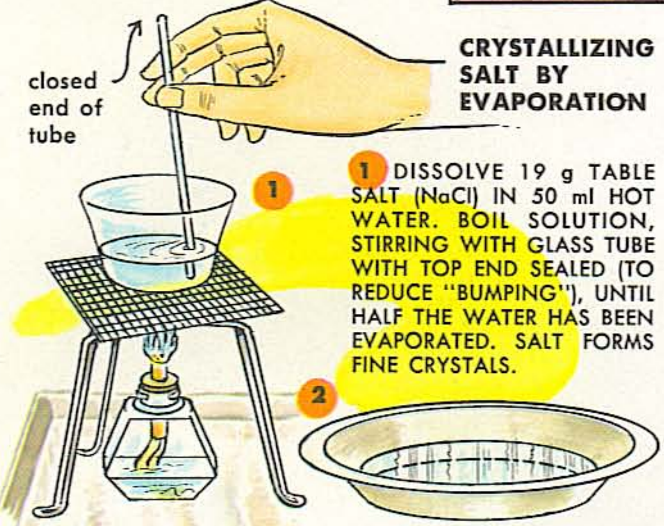

1 DISSOLVE 19 g TABLE SALT (NaCl) IN 50 ml HOT WATER. BOIL SOLUTION, STIRRING WITH GLASS TUBE WITH TOP END SEALED (TO REDUCE "BUMPING"), UNTIL HALF THE WATER HAS BEEN EVAPORATED. SALT FORMS FINE CRYSTALS.

2 POUR CLEAR LIQUID INTO LARGE PIE PLATE. PLACE IN SUNNY WINDOW FOR WATER TO EVAPORATE SLOWLY. THE CRYSTALS FORMED WILL BE MUCH LARGER.

MAKING NORMAL SALT FROM ACID SALT

SODIUM SULFATE IS PRODUCED BY HEATING SODIUM ACID SULFATE WITH SODIUM CHLORIDE.

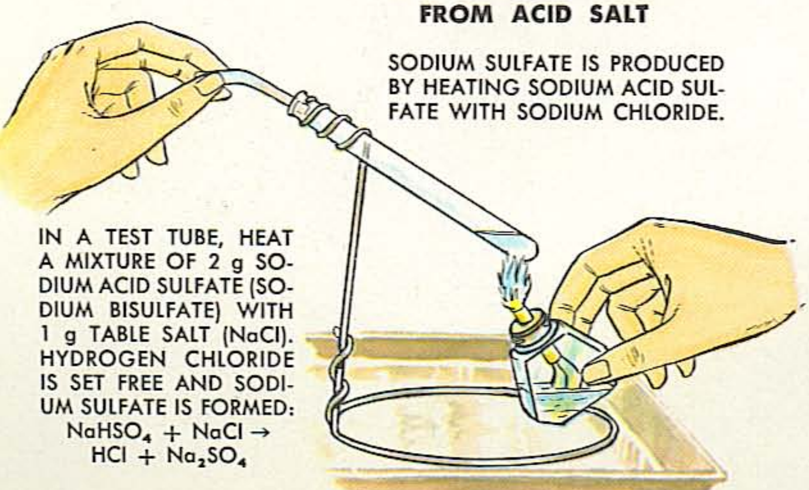

IN A TEST TUBE, HEAT A MIXTURE OF 2 g SODIUM ACID SULFATE (SODIUM BISULFATE) WITH 1 g TABLE SALT (NaCl). HYDROGEN CHLORIDE IS SET FREE AND SODIUM SULFATE IS FORMED:

$$NaHSO_4 + NaCl \rightarrow HCl + Na_2SO_4$$

MAKING ACID SALT FROM NORMAL SALT

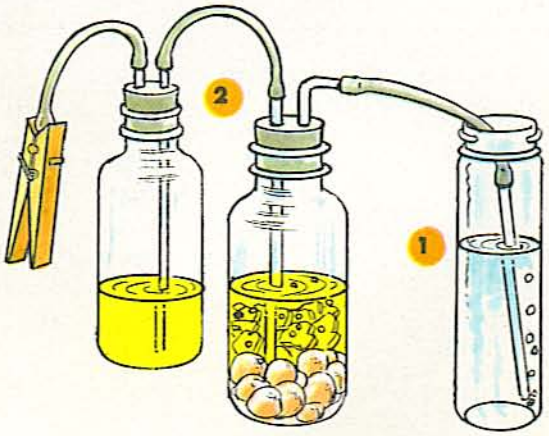

THE ACID CARBONATE ($NaHCO_3$) IS MADE BY LEADING CO_2 TO NORMAL CARBONATE (Na_2CO_3).

1 MAKE SATURATED SOLUTION BY SHAKING 3 TEASPOONS WASHING SODA IN 30 ml COOL, BUT NOT COLD, WATER. FILTER IT.

2 SET UP APPARATUS FOR MAKING CO_2 AS SHOWN ON PAGE 31. LEAD CO_2 INTO SODA SOLUTION FOR 10 MINUTES. THEN SET ASIDE. SHORTLY $NaHCO_3$ CRYSTALS APPEAR.

K KALIUM Element 19. Atomic wt.: 39.100. Density: 0.87. (English: Potassium) Silver-white metal, so soft it can be cut with knife. Oxidizes in air. Reacts with water. Burns with violet flame.

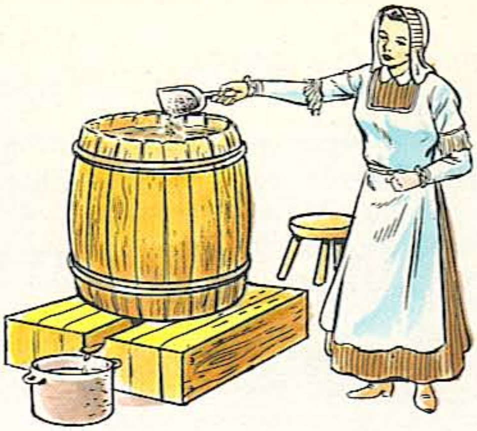

PIONEER WOMEN MADE POTASH FROM WOOD ASHES.

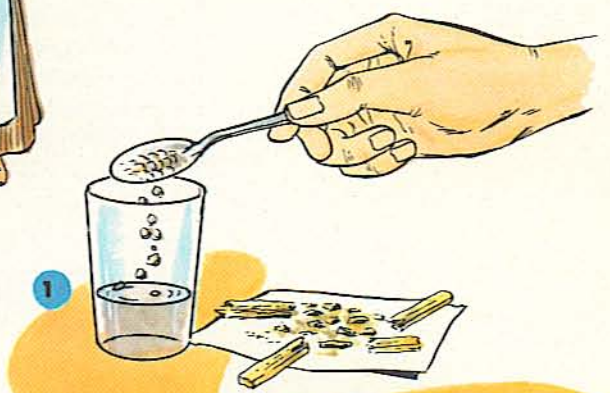

NITRATE TO NITRITE

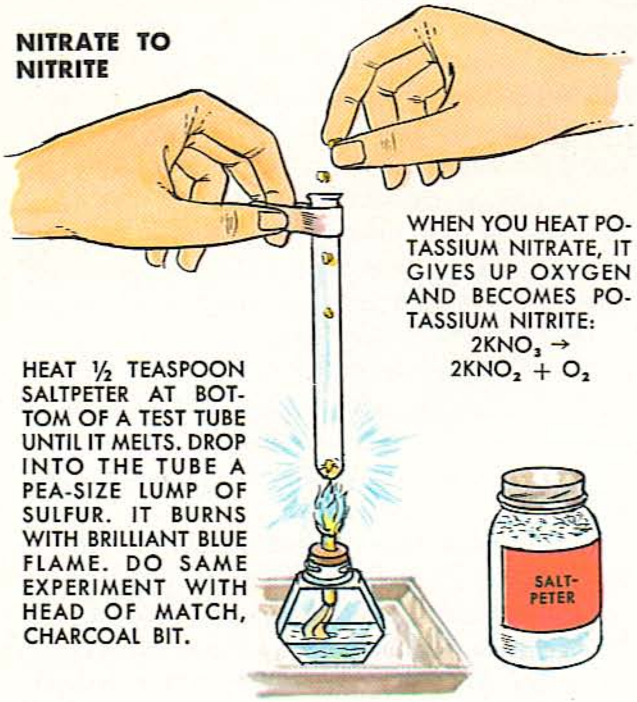

WHEN YOU HEAT POTASSIUM NITRATE, IT GIVES UP OXYGEN AND BECOMES POTASSIUM NITRITE:

$$2KNO_3 \rightarrow 2KNO_2 + O_2$$

HEAT ½ TEASPOON SALTPETER AT BOTTOM OF A TEST TUBE UNTIL IT MELTS. DROP INTO THE TUBE A PEA-SIZE LUMP OF SULFUR. IT BURNS WITH BRILLIANT BLUE FLAME. DO SAME EXPERIMENT WITH HEAD OF MATCH, CHARCOAL BIT.

MAKING POTASH

1 STIR UP SEVERAL TEASPOONS OF FRESH WOOD ASHES WITH WARM WATER. SKIM OFF WOOD REMAINS.

2 FILTER THE MIXTURE OF ASHES AND WATER. COLLECT FILTRATE IN A CUSTARD CUP. EVAPORATE MOST OF WATER. THEN COOL TO PERMIT K_2CO_3 TO CRYSTALLIZE OUT.

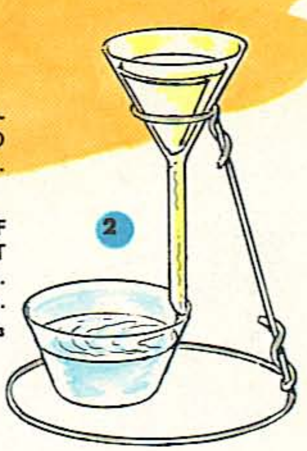

FLAME COLOR TEST FOR SODIUM AND POTASSIUM

THE COMPOUNDS OF CERTAIN METALS GIVE DISTINCT COLORS TO A FLAME. DIP NICHROME WIRE IN HCl TO CLEAN IT. HEAT IT. THEN DIP LOOP IN COMPOUND AND HOLD IN FLAME.

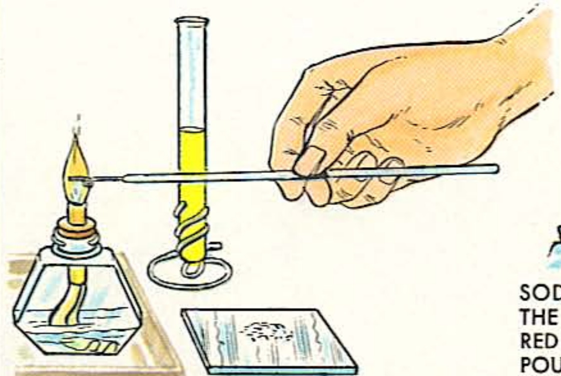

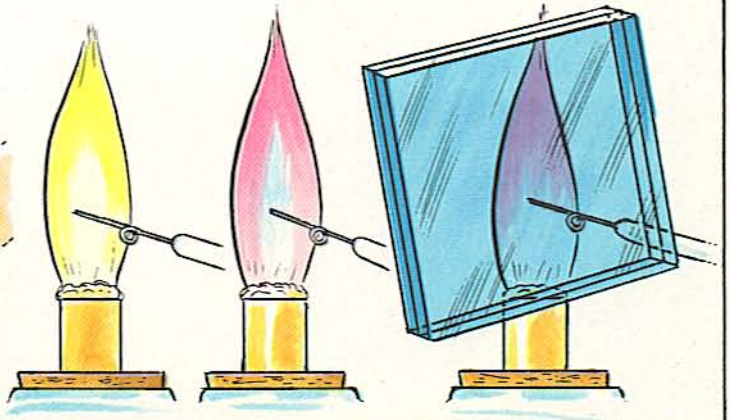

SODIUM COMPOUNDS GIVE THE FLAME A BRIGHT YELLOW-RED COLOR. POTASSIUM COMPOUNDS GIVE VIOLET FLAME.

TO SEE VIOLET COLOR OF POTASSIUM IN MIXTURE WITH Na, USE BLUE GLASS TO SCREEN OUT YELLOW OF Na.

Calcium— for Building

Ca **CALCIUM**
Element 20
Atomic wt.: 40.08. Density: 1.54. Silver-white metal with bright luster. Reacts with moist air and water to form hydroxide. When heated burns with brilliant light.

STAND UP STRAIGHT. You can do it because your bones contain calcium. Tell a mason to put up a brick house. He can do it with mortar containing calcium. Tell a master builder to build a monument. He will make it from marble — calcium again. Tell a hen to "go lay an egg." She can do it if she gets enough calcium in her feed to make the shell.

Calcium carbonate ($CaCO_3$) is the starting point for most calcium compounds — and for other chemicals as well. It is found in nature in cliffs and mountain ranges in the form of chalk and limestone and marble. And it makes up the shells of clams and mussels and billions of tiny sea creatures.

Calcium carbonate is almost insoluble in water. But if the water contains carbon dioxide, some goes in solution as calcium bicarbonate ($Ca(HCO_3)_2$). This explains the formations in our famous limestone caves. Rainwater containing carbon dioxide seeped through the ground and dissolved a small amount of limestone. In falling from the cave ceiling and drying, the drops gave up H_2O and CO_2 and left $CaCO_3$ behind. The minute deposits of falling drops during thousands of years created the stalactites hanging from the roof of the caves and the stalagmites rising from the floor.

A widespread mineral called gypsum is the sulfate of calcium. In this, each molecule of sulfate has two molecules of water attached to it ($CaSO_4 \cdot 2H_2O$). When gypsum is heated, it loses three quarters of its water and becomes plaster of Paris ($2CaSO_4 \cdot H_2O$). When you mix plaster of Paris and water, it again takes on the full amount of H_2O and hardens into a hydrate similar to the original gypsum.

The name of calcium was given to the metal hidden in limestone by its discoverer, Humphry Davy. It comes from *calx*, the old Latin name for lime.

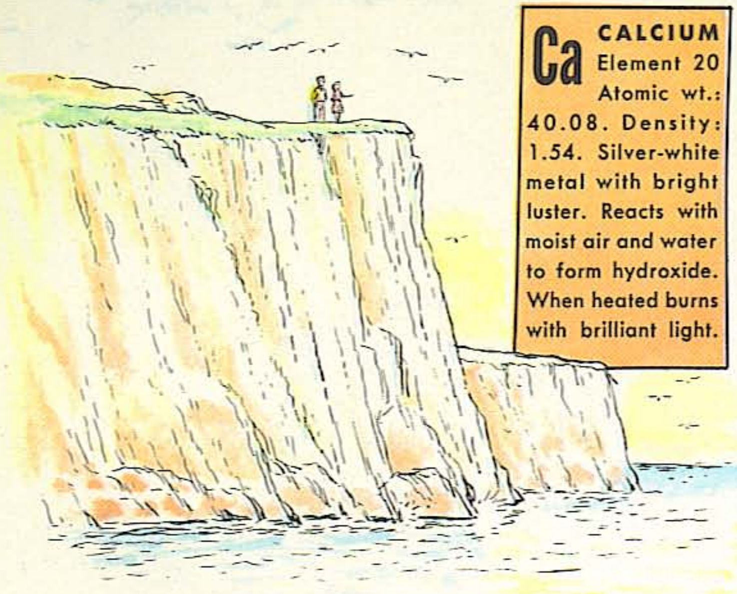
"THE WHITE CLIFFS OF DOVER" CONSIST OF ALMOST PURE CALCIUM CARBONATE IN THE FORM OF CHALK.

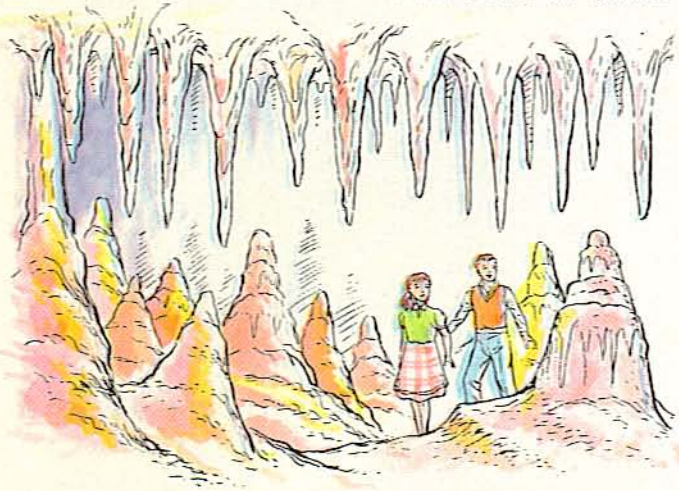
STALACTITES AND STALAGMITES ARE UNDERGROUND DEPOSITS OF $CaCO_3$.

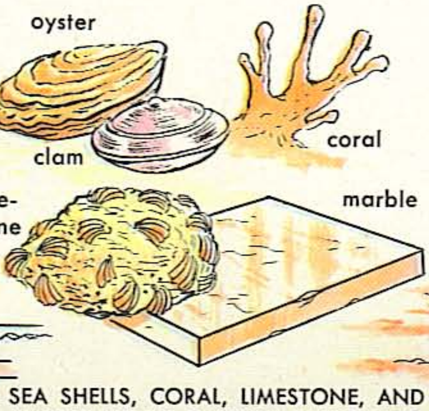

SEA SHELLS, CORAL, LIMESTONE, AND MARBLE ARE ALL CALCIUM CARBONATE.

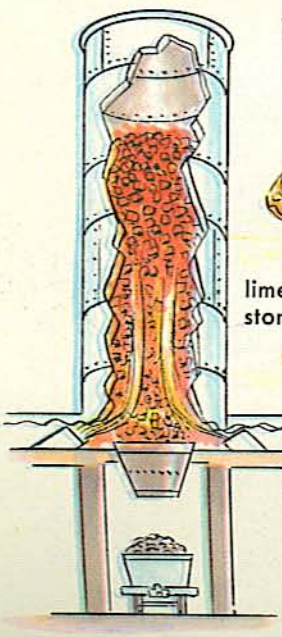
WHEN LIMESTONE IS HEATED IN KILNS, IT LOSES CARBON DIOXIDE AND TURNS INTO QUICKLIME—CALCIUM OXIDE.

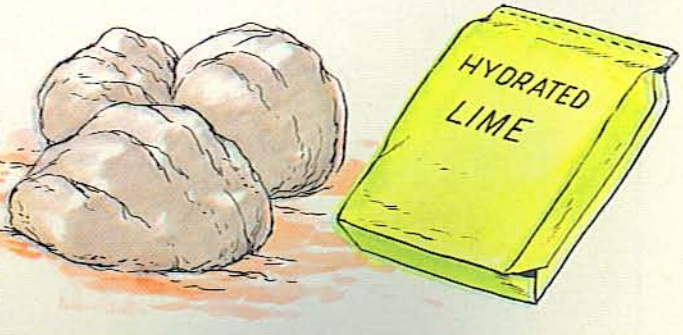

WHEN WATER IS ADDED TO LUMPS OF QUICKLIME (CaO), THEY CRUMBLE INTO A WHITISH POWDER OF SLAKED OR HYDRATED LIME ($Ca(OH)_2$). (SEE ALSO PAGE 45.)

TESTING HARDNESS OF WATER

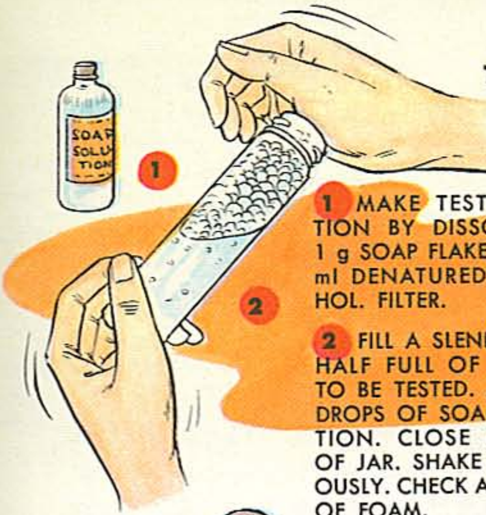
SOAP SOLUTION
1
2

1 MAKE TEST SOLUTION BY DISSOLVING 1 g SOAP FLAKES IN 20 ml DENATURED ALCOHOL. FILTER.

2 FILL A SLENDER JAR HALF FULL OF WATER TO BE TESTED. ADD 10 DROPS OF SOAP SOLUTION. CLOSE MOUTH OF JAR. SHAKE VIGOROUSLY. CHECK AMOUNT OF FOAM.

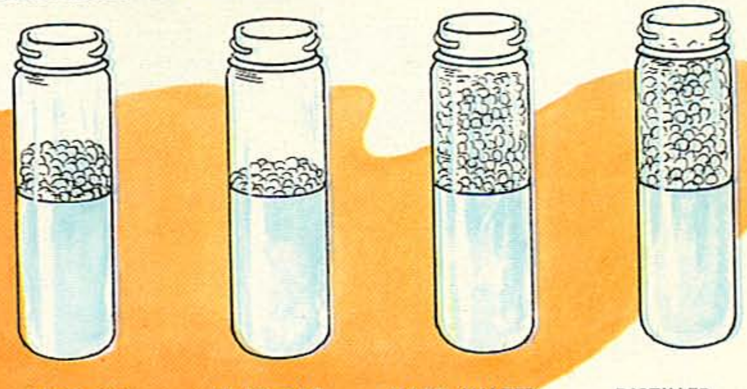

SOFT TAP WATER GIVES FAIR AMOUNT OF FOAM.

HARD TAP WATER MAKES VERY LITTLE FOAM.

HARD WATER SOFTENED WITH WASHING SODA MAKES FOAM.

DISTILLED WATER MAKES LARGE AMOUNT OF FOAM.

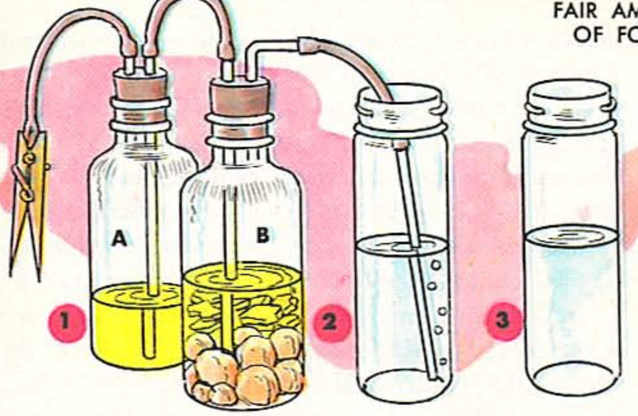

MAKING "HARD" WATER

1 SET UP GAS GENERATOR AS SHOWN ON PAGE 31. BOTTLE **A** CONTAINS HYDROCHLORIC ACID. IN BOTTLE **B**, PLACE MARBLE CHIPS ON TOP OF PEBBLES. POUR ACID ON MARBLE CHIPS TO MAKE CO$_2$.

2 LEAD THE CARBON DIOXIDE INTO LIME WATER. IT TURNS MILKY THROUGH FORMATION OF CaCO$_3$.

3 CONTINUE LEADING CO$_2$ INTO MILKY SOLUTION. MILKINESS DISAPPEARS. INSOLUBLE CaCO$_3$ HAS BEEN TURNED INTO SOLUBLE Ca(HCO$_3$)$_2$. THIS IS THE SUBSTANCE THAT MAKES MOST HARD WATER "HARD."

DISTILLING WATER

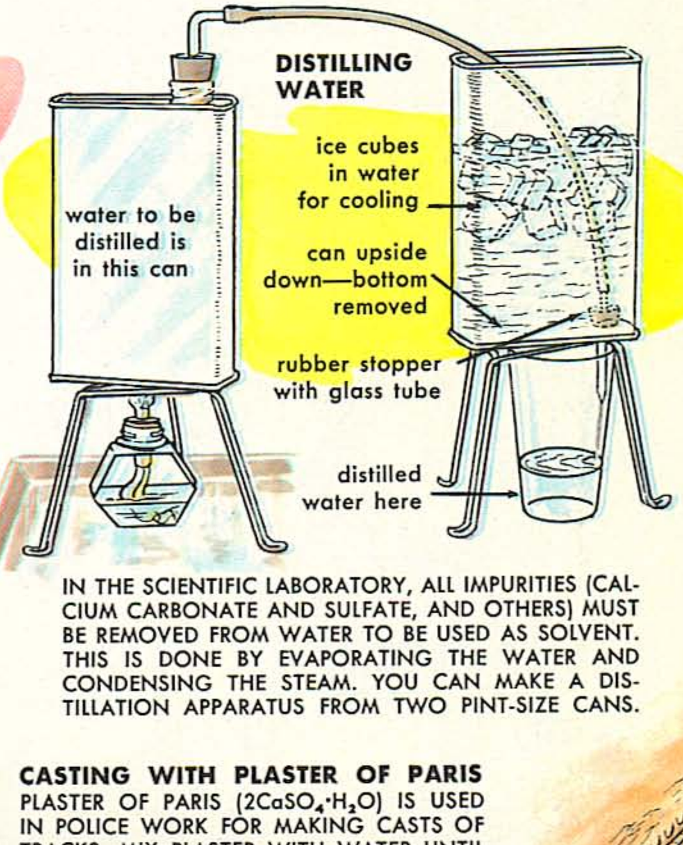
water to be distilled is in this can
ice cubes in water for cooling
can upside down—bottom removed
rubber stopper with glass tube
distilled water here

IN THE SCIENTIFIC LABORATORY, ALL IMPURITIES (CALCIUM CARBONATE AND SULFATE, AND OTHERS) MUST BE REMOVED FROM WATER TO BE USED AS SOLVENT. THIS IS DONE BY EVAPORATING THE WATER AND CONDENSING THE STEAM. YOU CAN MAKE A DISTILLATION APPARATUS FROM TWO PINT-SIZE CANS.

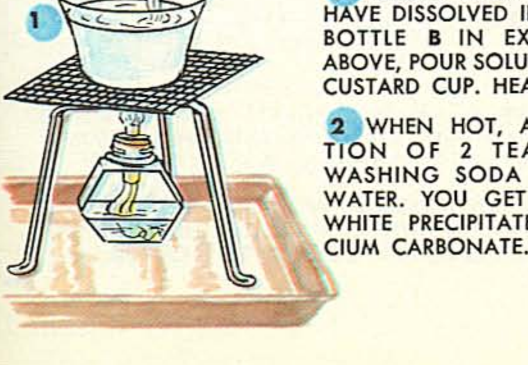

PRECIPITATED CALCIUM CARBONATE

1 WHEN MARBLE CHIPS HAVE DISSOLVED IN ACID IN BOTTLE **B** IN EXPERIMENT ABOVE, POUR SOLUTION INTO CUSTARD CUP. HEAT.

2 WHEN HOT, ADD SOLUTION OF 2 TEASPOONS WASHING SODA IN 50 ml WATER. YOU GET A HEAVY WHITE PRECIPITATE OF CALCIUM CARBONATE.

CASTING WITH PLASTER OF PARIS

PLASTER OF PARIS (2CaSO$_4$·H$_2$O) IS USED IN POLICE WORK FOR MAKING CASTS OF TRACKS. MIX PLASTER WITH WATER UNTIL IT HAS CONSISTENCY OF MEDIUM CREAM. POUR IN TRACK. LEAVE TO SET ONE HOUR.

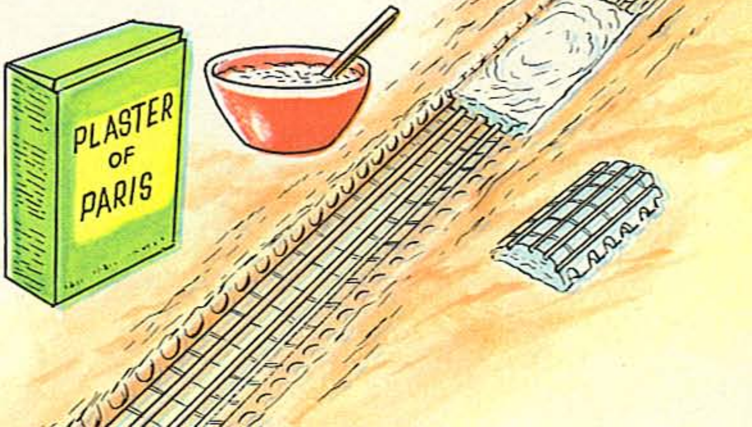
PLASTER OF PARIS

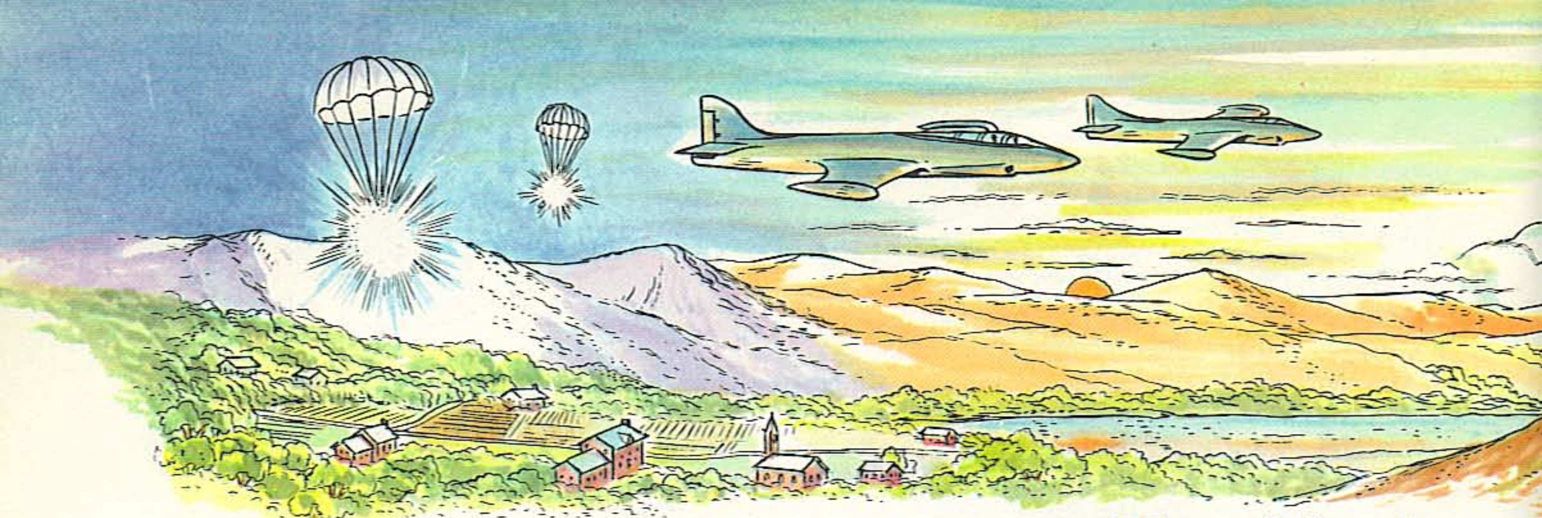

Let's Compare Two Metals

TAKE A LOOK at the periodic table of elements on pages 38-39. In column IIA you find the metal magnesium, in column IIB the metal zinc. The fact that the two families in which they are found both have the Roman numeral II would indicate that they are related. But the fact that they are in separate "subgroups" would suggest that they differ in certain respects. That is exactly the case.

In their compounds they are very much alike. One atom combines with one atom of oxygen to form the oxide (MgO and ZnO), and one atom replaces two atoms of hydrogen in forming a salt ($MgCl_2$ and $ZnCl_2$, for instance). But in some of their reactions they do not behave alike — as you will learn.

Before World War II, magnesium had little use — mainly in flash photography because it burns with a blinding, white light. But the metal became important when lightweight planes were needed — melted together with other metals it forms an "alloy"

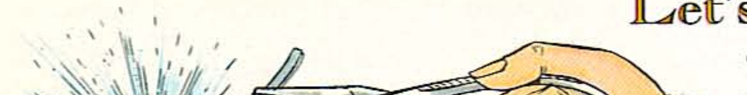

HOLD 2-INCH PIECE OF MAGNESIUM RIBBON WITH A PAIR OF PLIERS. IGNITE IT. IT BURNS WITH A BRILLIANT, WHITE FLAME. MIX ASHES (MgO) WITH WATER. TEST MIXTURE WITH RED LITMUS PAPER.

CUT SLIVER OF ZINC. HOLD IT IN FLAME. IT BURNS WITH BLUISH-GREEN FLAME TO ZINC OXIDE. ZnO IS YELLOW WHEN HOT, WHITE WHEN COOL.

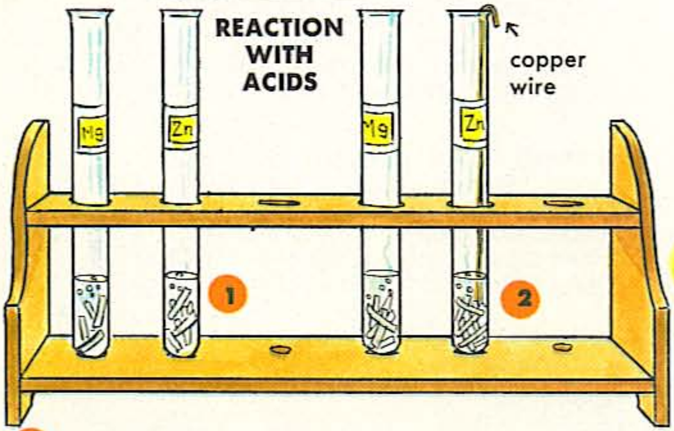

① BOTH METALS REACT WITH WEAK ACIDS, EVEN WITH VINEGAR—Mg WITH COLD VINEGAR, Zn WITH HOT.

② POUR SOLUTION OF 1 g SODIUM BISULFATE IN 10 ml WATER ON Mg AND Zn. Mg REACTS FAST, Zn SLOWLY. NOW TOUCH ZINC WITH A COPPER WIRE. REACTION SPEEDS UP, CAUSED BY ELECTRIC PROCESS.

REPLACEMENT OF COPPER

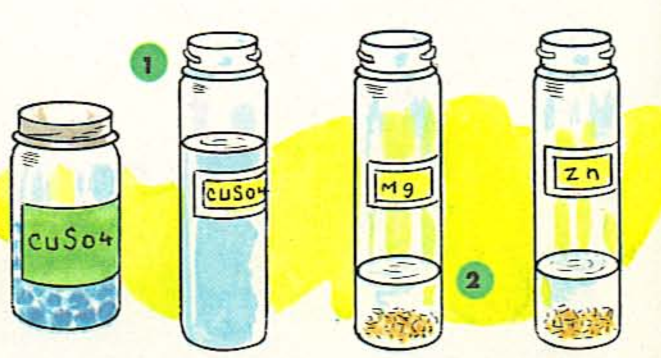

① DISSOLVE 4 g COPPER SULFATE IN 40 ml WATER. POUR HALF OF THE SOLUTION OVER STRIPS OF MAGNESIUM, THE OTHER HALF OVER SLIVERS OF ZINC.

② COPPER IS FORCED OUT AND Mg AND Zn GO INTO SOLUTION. IF ENOUGH METAL IS USED, THE BLUE COLOR DISAPPEARS. $MgSO_4$ AND $ZnSO_4$ ARE COLORLESS.

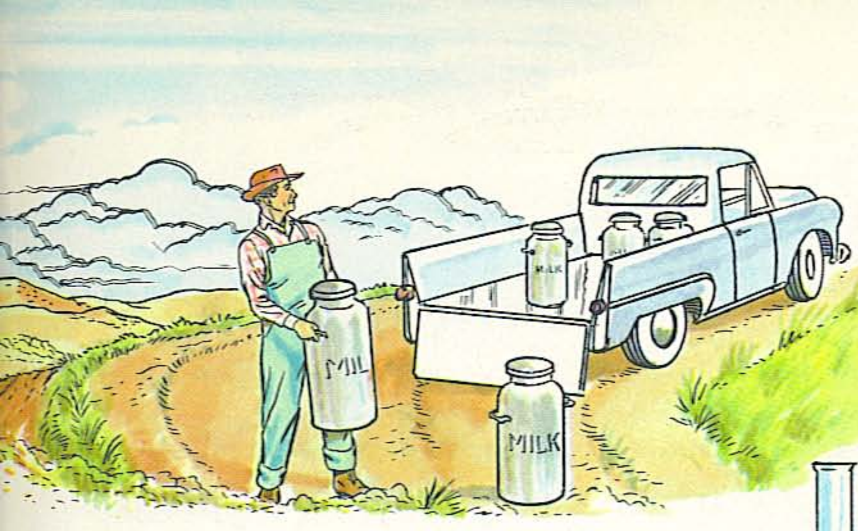

Mg **MAGNESIUM** Element 12. Atomic wt.: 24.32. Density: 1.75. Silver-white metal. Ductile, malleable. Reacts with boiling water. Burns in air with very brilliant white light.

Zn **ZINC** Element 30. Atomic wt.: 65.38. Density: 7.1. Bluish-white metal. Ductile and malleable. Distils when heated to boiling. Can be made to burn with bluish flame.

that is light yet very strong. Some magnesium compounds are used in medicine: milk of magnesia ($Mg(OH)_2$) and Epsom salt ($MgSO_4 \cdot 7H_2O$).

Zinc has been used for ages to coat iron pails and pipes to prevent them from rusting — "galvanized iron." Zinc is also a part of many alloys (German silver and brass) and is important in the making of dry-cell batteries.

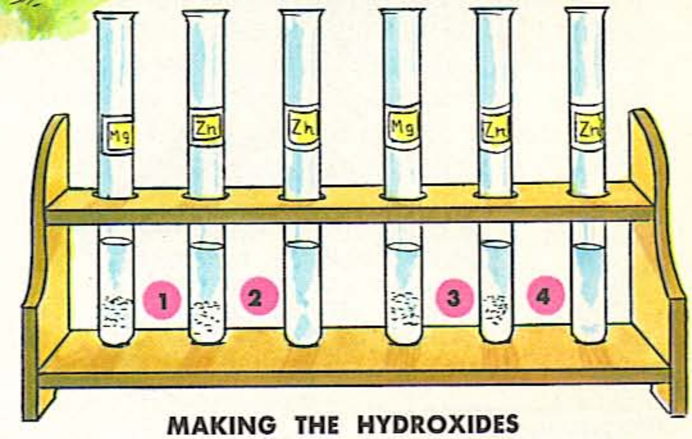

MAKING THE HYDROXIDES

1 ADD SODIUM HYDROXIDE SOLUTION TO SOLUTION OF MAGNESIUM SULFATE. WHITE $Mg(OH)_2$ FORMS.

2 ADD SMALL AMOUNT OF NaOH SOLUTION TO DILUTED TINNERS' FLUID ($ZnCl_2$). $Zn(OH)_2$ IS FORMED. ADD MORE NaOH. PRECIPITATE DISSOLVES WITH FORMATION OF SODIUM ZINCATE (Na_2ZnO_2).

3 ADD AMMONIA (AMMONIUM HYDROXIDE) TO MAGNESIUM SULFATE SOLUTION. AGAIN $Mg(OH)_2$ FORMS.

4 ADD SMALL AMOUNT OF AMMONIA TO DILUTED TINNERS' FLUID. $Zn(OH)_2$ FORMS. ADD MORE. THE $Zn(OH)_2$ DISSOLVES, FORMING COMPOUND WITH NH_3.

MAKING THE CARBONATES

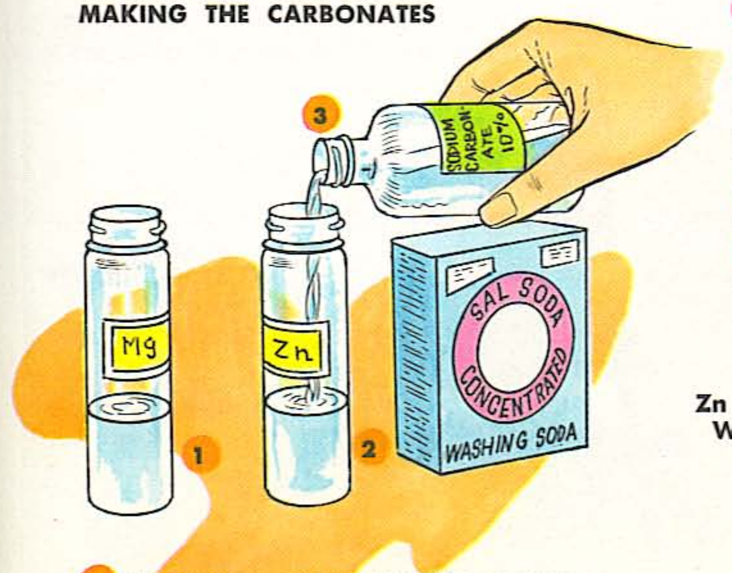

1 DISSOLVE 2 g EPSOM SALT (MAGNESIUM SULFATE, $MgSO_4 \cdot 7H_2O$) IN 20 ml WATER.

2 GET FROM HARDWARE STORE SMALL BOTTLE OF "TINNERS' FLUID." THIS IS A STRONG SOLUTION OF ZINC CHLORIDE. DILUTE 5ml OF FLUID WITH 15 ml WATER.

3 MAKE SOLUTION OF 5 g WASHING SODA (SODIUM CARBONATE) IN 50 ml WATER. ADD SOME OF THIS SOLUTION TO THE OTHER TWO. IN BOTH JARS YOU WILL GET A HEAVY WHITE PRECIPITATE. IN THE Mg JAR, THIS IS NORMAL MAGNESIUM CARBONATE ($MgCO_3$). IN Zn JAR, CO_2 IS SET FREE AND BASIC ZINC CARBONATE ($Zn(OH)_2, ZnCO_3$) RESULTS.

Zn AND Mg WITH H_2S

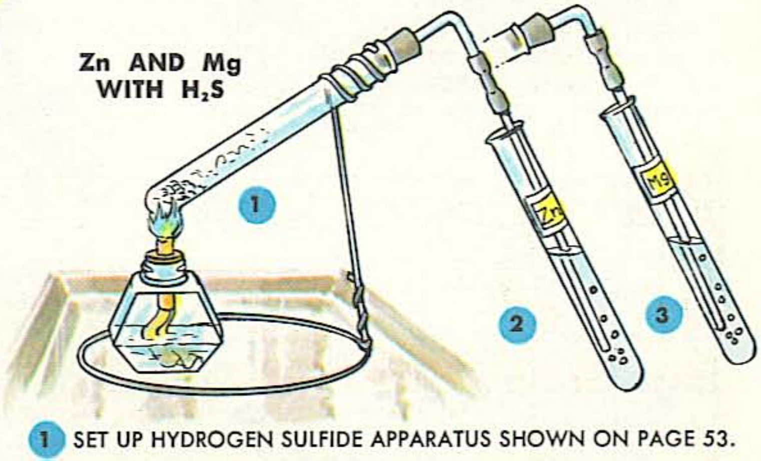

1 SET UP HYDROGEN SULFIDE APPARATUS SHOWN ON PAGE 53.

2 LEAD HYDROGEN SULFIDE (H_2S) INTO DILUTED TINNERS' FLUID ($ZnCl_2$). YOU GET A WHITE PRECIPITATE OF ZnS.

3 LEAD H_2S INTO SOLUTION OF EPSOM SALT ($MgSO_4$). HERE ALSO YOU GET WHITE PRECIPITATE. BUT NOT OF MAGNESIUM SULFIDE. THIS REACTS WITH THE WATER TO MAKE $Mg(OH)_2$.

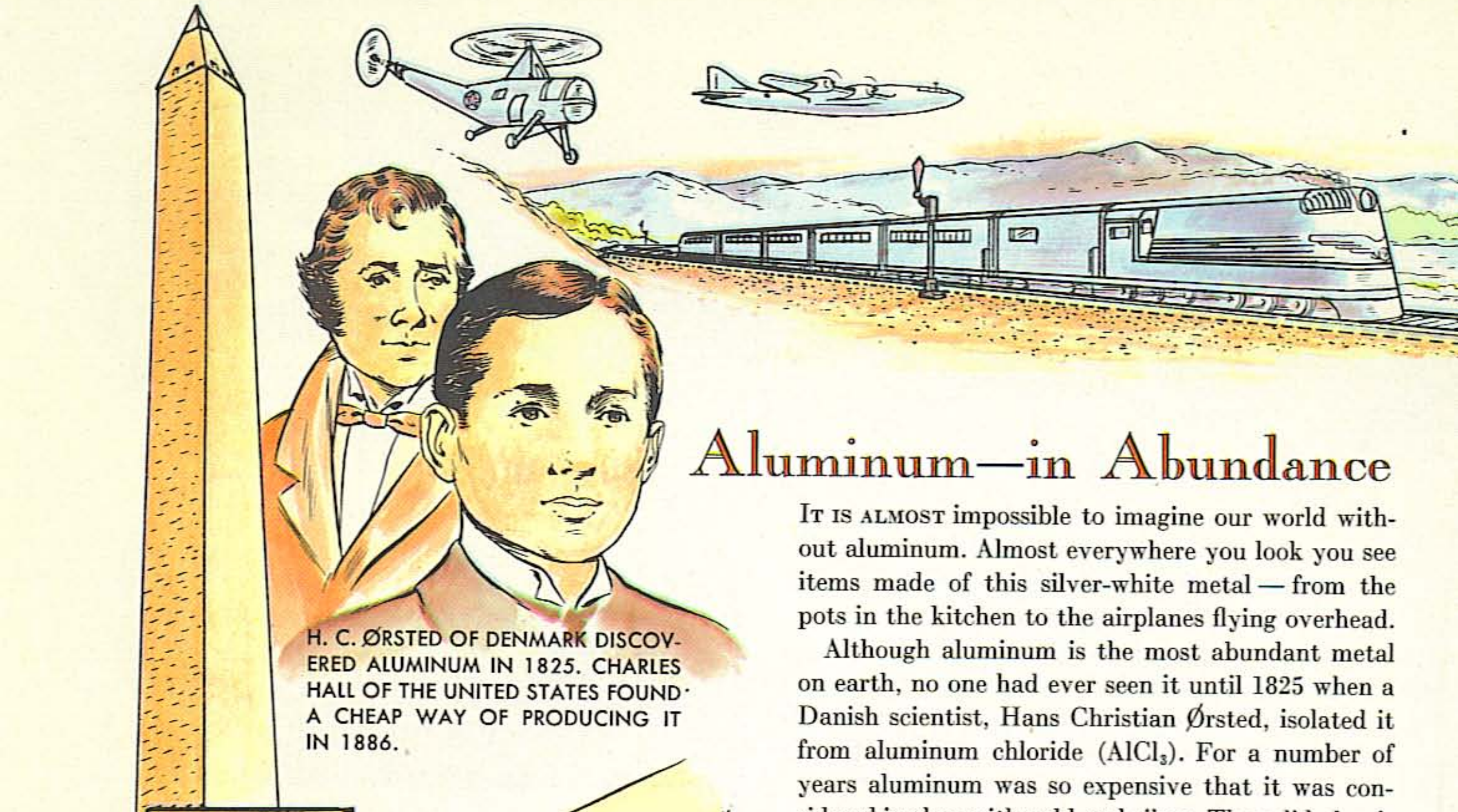

Aluminum—in Abundance

It is almost impossible to imagine our world without aluminum. Almost everywhere you look you see items made of this silver-white metal — from the pots in the kitchen to the airplanes flying overhead.

Although aluminum is the most abundant metal on earth, no one had ever seen it until 1825 when a Danish scientist, Hans Christian Ørsted, isolated it from aluminum chloride ($AlCl_3$). For a number of years aluminum was so expensive that it was considered in class with gold and silver. The solid aluminum cap placed on top of the Washington Monument in 1884 was first put on public display so that everyone could have a look at such a great rarity. Two

H. C. ØRSTED OF DENMARK DISCOVERED ALUMINUM IN 1825. CHARLES HALL OF THE UNITED STATES FOUND A CHEAP WAY OF PRODUCING IT IN 1886.

Al ALUMINUM Element 13. Atomic wt.: 26.98. Density: 2.70. Silver-white metal; ductile, malleable, able to take a high polish. Amphoteric. Will burn in oxygen with white flame.

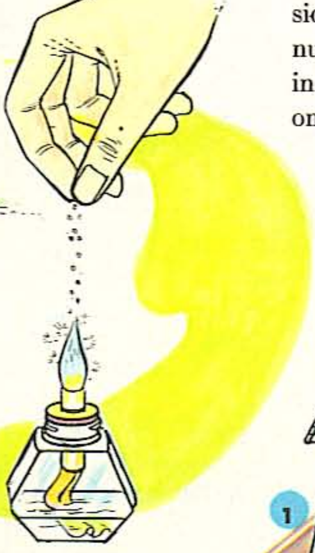

MAKE A SMALL AMOUNT OF ALUMINUM POWDER BY FILING IT OFF AN OLD ALUMINUM POT. SPRINKLE IN FLAME TO MAKE SPARKS OF BURNING ALUMINUM.

GROWING ALUM CRYSTALS

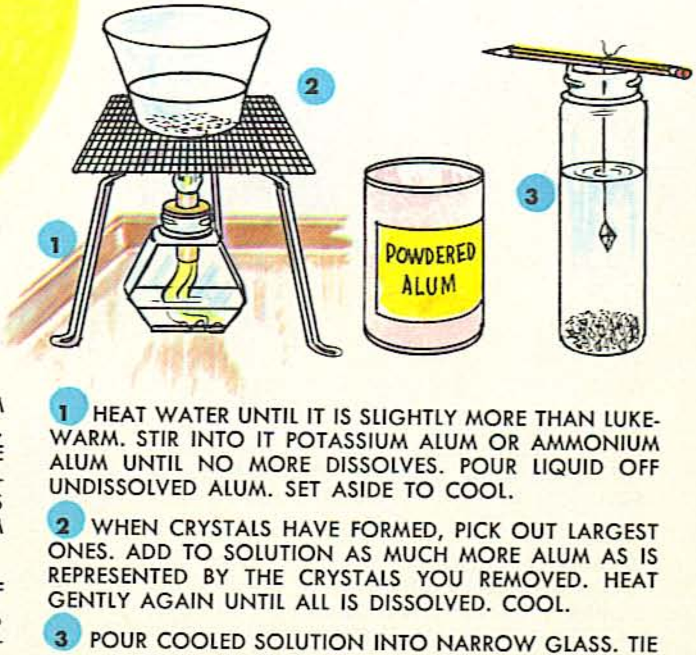

1 HEAT WATER UNTIL IT IS SLIGHTLY MORE THAN LUKEWARM. STIR INTO IT POTASSIUM ALUM OR AMMONIUM ALUM UNTIL NO MORE DISSOLVES. POUR LIQUID OFF UNDISSOLVED ALUM. SET ASIDE TO COOL.

2 WHEN CRYSTALS HAVE FORMED, PICK OUT LARGEST ONES. ADD TO SOLUTION AS MUCH MORE ALUM AS IS REPRESENTED BY THE CRYSTALS YOU REMOVED. HEAT GENTLY AGAIN UNTIL ALL IS DISSOLVED. COOL.

3 POUR COOLED SOLUTION INTO NARROW GLASS. TIE THREAD TO LARGEST CRYSTAL YOU PICKED. HANG THIS IN SOLUTION FROM A PENCIL. PLACE IN QUIET SPOT. LET THE CRYSTAL GROW FOR A WEEK OR MORE.

DISSOLVING ALUMINUM

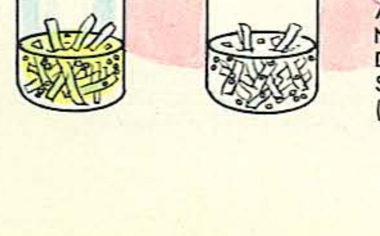

1 CUT ALUMINUM FOIL IN SMALL STRIPS. DROP THEM IN A LITTLE DILUTED HYDROCHLORIC ACID. HYDROGEN IS RELEASED; ALUMINUM CHLORIDE IS FORMED.

2 DROP STRIPS OF ALUMINUM FOIL IN 10% NaOH SOLUTION. HYDROGEN IS FREED AND SODIUM ALUMINATE ($NaAlO_2$) IS FORMED.

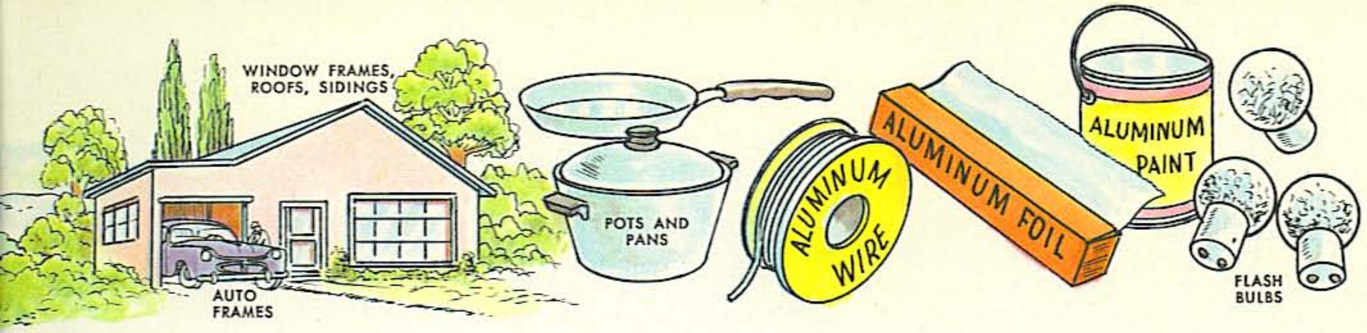

years later, a 22-year-old American chemist, Charles Martin Hall, invented a way of producing aluminum cheaply from aluminum oxide (Al_2O_3). Since then aluminum has become one of the most popular of all metals — mostly because of its lightness.

The mineral bauxite ($AlHO_2, Al(OH)_3$) is our main source of aluminum. But aluminum is also found in nature as oxide and in many complex silicates. Clay, for instance, is an aluminum silicate.

Two things about aluminum will interest you as a chemist. One is that aluminum is an "amphoteric" element, which means that it can form not only a base ($Al(OH)_3$), but also an acid ($HAlO_2$). The other is that aluminum sulfate ($Al_2(SO_4)_3$) has the ability to combine with potassium sulfate (K_2SO_4) and ammonium sulfate ($(NH_4)_2SO_4$) into beautiful cubic crystals of double salts called "alums"— $KAl\,(SO_4)_2 \cdot 12H_2O$ and $NH_4Al(SO_4)_2 \cdot 12H_2O$.

MAKING ALUMINUM HYDROXIDE

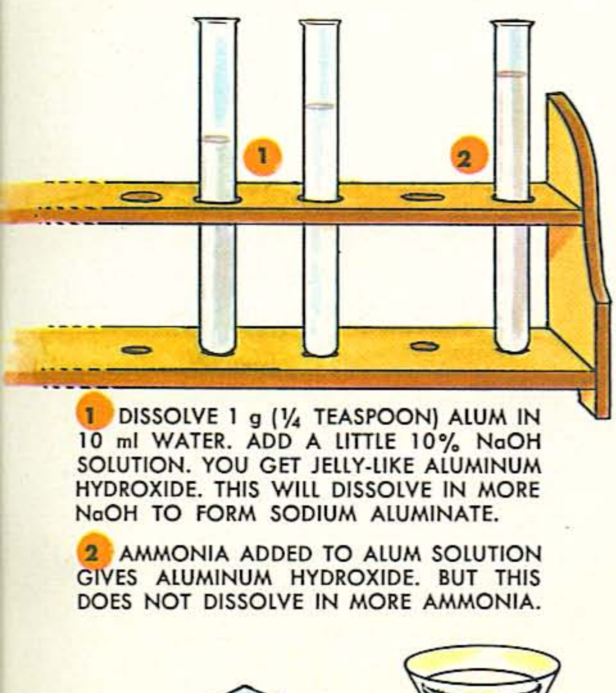

1 DISSOLVE 1 g (¼ TEASPOON) ALUM IN 10 ml WATER. ADD A LITTLE 10% NaOH SOLUTION. YOU GET JELLY-LIKE ALUMINUM HYDROXIDE. THIS WILL DISSOLVE IN MORE NaOH TO FORM SODIUM ALUMINATE.

2 AMMONIA ADDED TO ALUM SOLUTION GIVES ALUMINUM HYDROXIDE. BUT THIS DOES NOT DISSOLVE IN MORE AMMONIA.

CLEARING WATER WITH ALUM

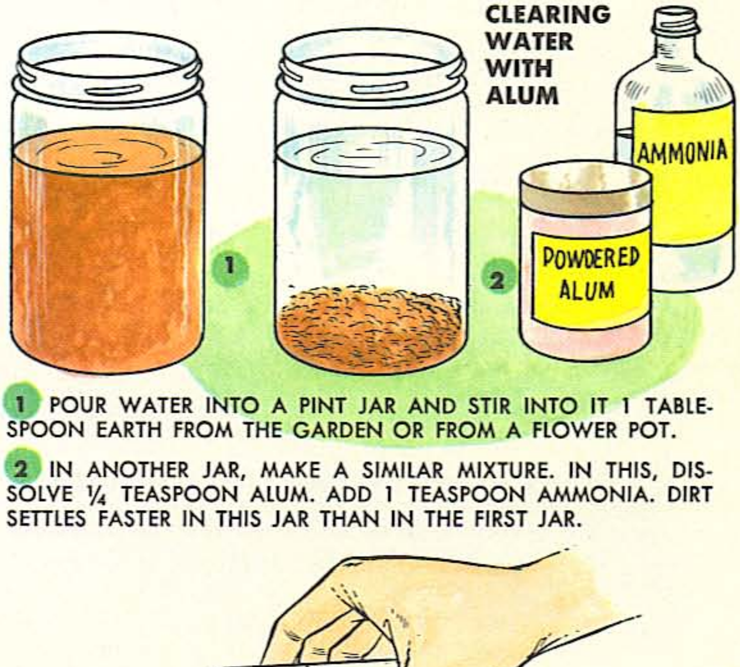

1 POUR WATER INTO A PINT JAR AND STIR INTO IT 1 TABLESPOON EARTH FROM THE GARDEN OR FROM A FLOWER POT.

2 IN ANOTHER JAR, MAKE A SIMILAR MIXTURE. IN THIS, DISSOLVE ¼ TEASPOON ALUM. ADD 1 TEASPOON AMMONIA. DIRT SETTLES FASTER IN THIS JAR THAN IN THE FIRST JAR.

ALUM AND SODA

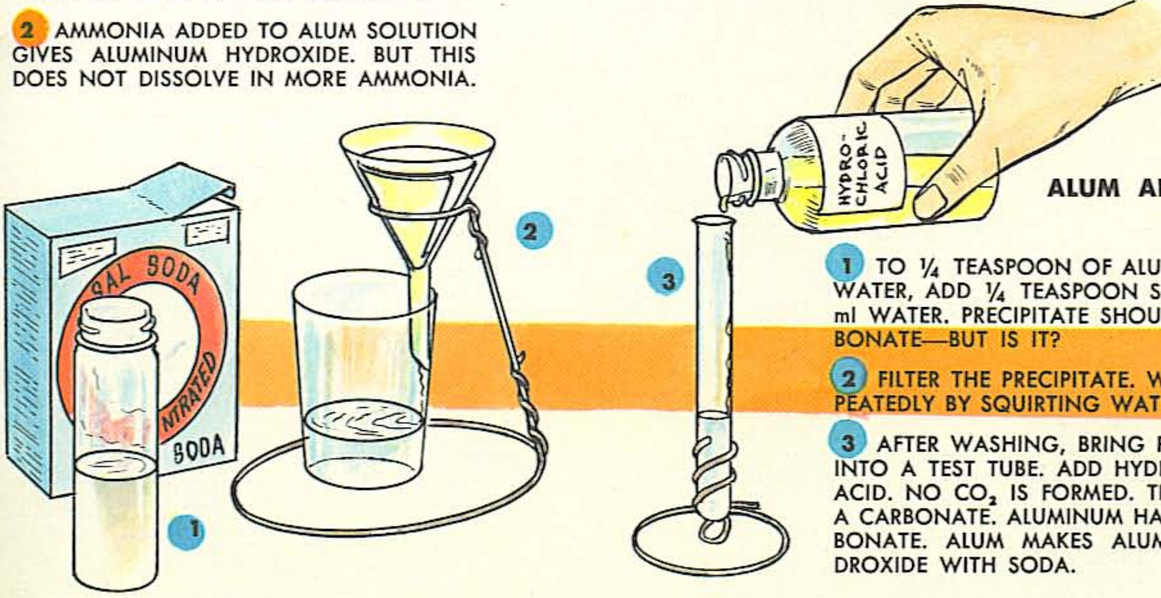

1 TO ¼ TEASPOON OF ALUM IN 20 ml WATER, ADD ¼ TEASPOON SODA IN 20 ml WATER. PRECIPITATE SHOULD BE CARBONATE—BUT IS IT?

2 FILTER THE PRECIPITATE. WASH IT REPEATEDLY BY SQUIRTING WATER INTO IT.

3 AFTER WASHING, BRING PRECIPITATE INTO A TEST TUBE. ADD HYDROCHLORIC ACID. NO CO_2 IS FORMED. THIS IS NOT A CARBONATE. ALUMINUM HAS NO CARBONATE. ALUM MAKES ALUMINUM HYDROXIDE WITH SODA.

MANGANESE DIOXIDE TO MANGANESE SULFATE

MANGANESE SULFATE TO MANGANESE HYDROXIDE

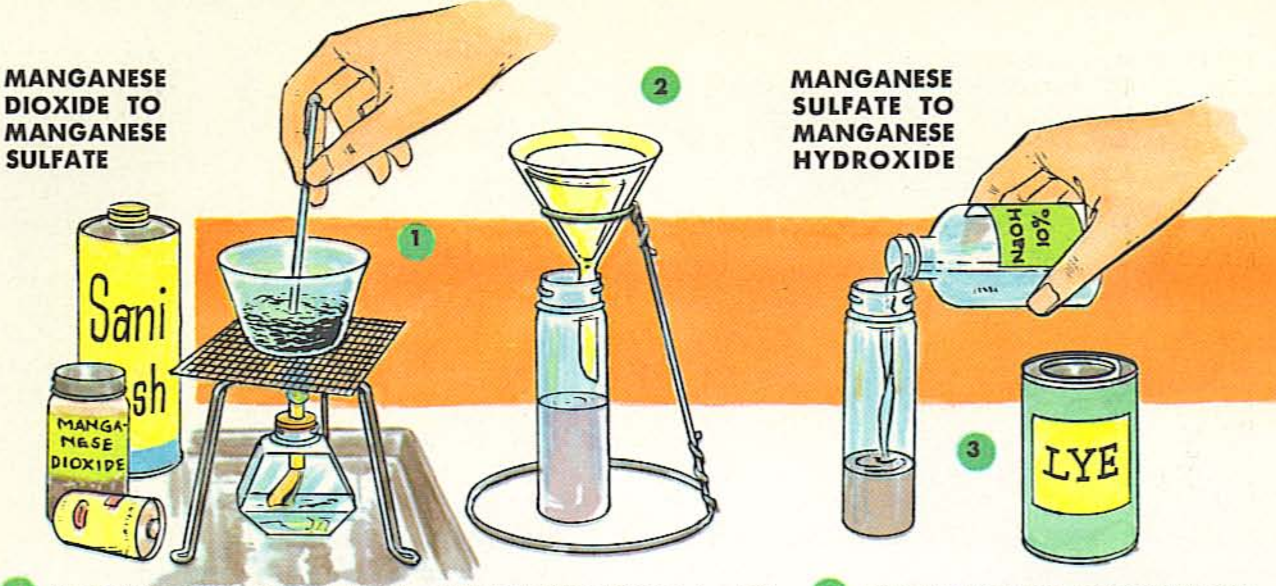

1 IN A PYREX CUSTARD CUP, MIX 2 g MANGANESE DIOXIDE, 6 g SODIUM BISULFATE, AND 10 ml WATER. HEAT MIXTURE GENTLY. IT WILL BUBBLE VIGOROUSLY BECAUSE OXYGEN IS SET FREE.

2 AFTER A FEW MINUTES, ADD 30 ml WATER. FILTER. FILTRATE CONTAINS MANGANESE SULFATE ($MnSO_4$) AND SODIUM SULFATE.

3 INTO HALF OF THE MANGANESE SULFATE SOLUTION YOU HAVE MADE, POUR 10% SOLUTION OF NaOH UNTIL NO MORE PRECIPITATE FORMS. WHITISH $Mn(OH)_2$ OXIDIZES INTO BROWN MnO(OH).

Manganese—Metal of Many Colors

METALLIC MANGANESE has no use by itself. But add up to 15 per cent of it to steel and the result is an alloy — "manganese steel" — so hard that it is suitable for machine parts that are exposed to a lot of rough wear.

The most common ore from which manganese is extracted goes under the name of "pyrolusite." This is nothing but your old friend manganese dioxide (MnO_2) which you found in your flashlight battery and have already used in a great number of your chemical experiments.

The compounds of manganese come in almost any color you can think of: black and brown, white and pink and red, violet and green. In working with these compounds, your fingers and glassware may get brown. You can get rid of this stain easily with diluted hydrochloric acid. Rinse thoroughly with water afterwards.

EXPERIMENTS WITH POTASSIUM PERMANGANATE

$KMnO_4$ WILL GIVE YOU AN IDEA OF SMALLNESS OF MOLECULE.

1 DISSOLVE ½ g POTASSIUM PERMANGANATE IN 50 ml WATER. THIS GIVES A SOLUTION OF 1 TO 100, OR 1/100.

2 DILUTE 5 ml OF THIS SOLUTION WITH 45 ml WATER. YOU NOW HAVE A SOLUTION OF 1 TO 1,000, OR 1/1,000.

3 AGAIN, 5 ml TO 45 ml WATER FOR SOLUTION 1/10,000.

4 AGAIN, 5 ml TO 45 ml WATER FOR SOLUTION 1/100,000.

5 AGAIN, 5 ml TO 45 ml WATER FOR SOLUTION 1/1,000,000. COLOR YOU STILL SEE IS CAUSED BY THE PRESENCE OF MORE THAN 600,000,000,000,000,000 MOLECULES OF $KMnO_4$.

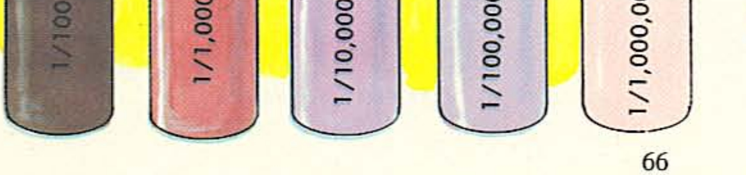

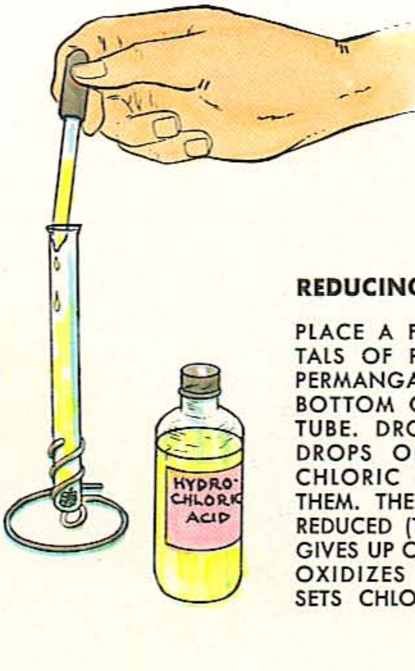

REDUCING $KMnO_4$

PLACE A FEW CRYSTALS OF POTASSIUM PERMANGANATE ON BOTTOM OF A TEST TUBE. DROP A FEW DROPS OF HYDROCHLORIC ACID ON THEM. THE $KMnO_4$ IS REDUCED (THAT IS, IT GIVES UP OXYGEN). IT OXIDIZES HCl AND SETS CHLORINE FREE.

THE EXPERIMENTS ALONG THE TOP OF THESE PAGES SHOW HOW IT IS POSSIBLE TO MOVE FROM ONE COMPOUND TO ANOTHER.

MANGANESE SULFATE TO MANGANESE CARBONATE

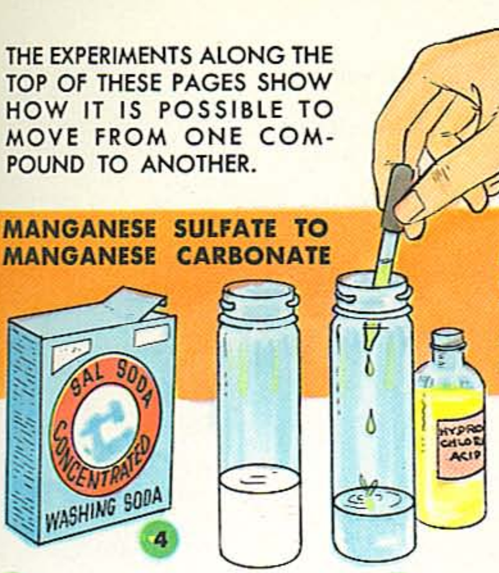

MANGANESE CARBONATE TO MANGANESE CHLORIDE

MANGANESE CHLORIDE TO MANGANESE SULFIDE

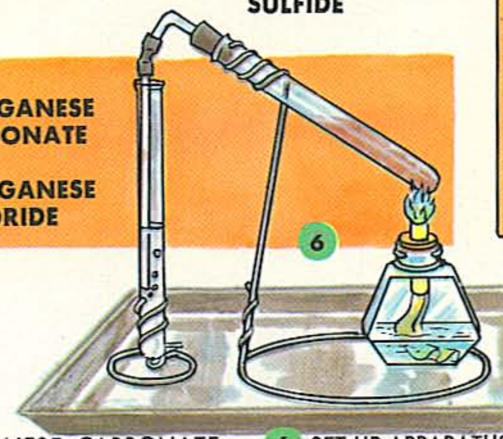

Mn **MANGANESE**
Element 25.
Atomic wt.: 54.94.
Density: 7.44. Silvery-gray metal with reddish tinge. Reacts with water. Its compounds with oxygen range from bases to acids.

4 INTO SECOND HALF OF SOLUTION, POUR SOLUTION OF 4 g SODIUM CARBONATE IN 10 ml WATER. WHITE PRECIPITATE IS $MnCO_3$.

5 LET MANGANESE CARBONATE SETTLE. POUR LIQUID OFF PRECIPITATE. ADD HYDROCHLORIC ACID BY THE DROP UNTIL DISSOLVED. RESULT IS MANGANESE CHLORIDE ($MnCl_2$).

6 SET UP APPARATUS FOR MAKING HYDROGEN SULFIDE (SEE PAGE 53). POUR SOLUTION OF $MnCl_2$ INTO TEST TUBE. DILUTE IT IF NECESSARY. LEAD H_2S INTO IT. YOU GET MANGANESE SULFIDE.

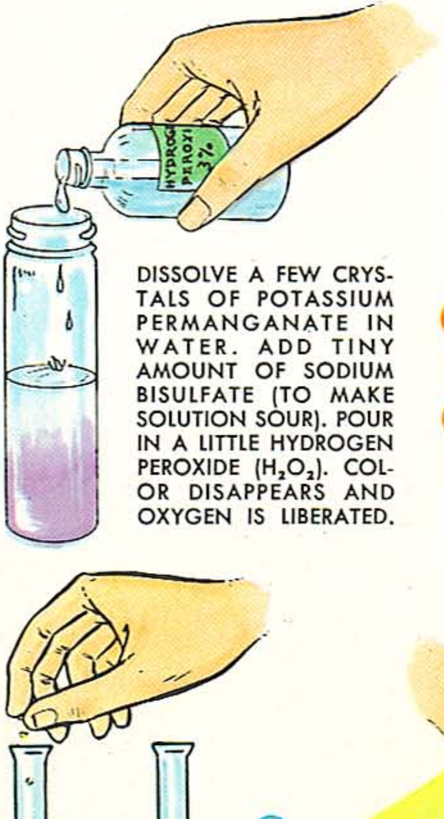

DISSOLVE A FEW CRYSTALS OF POTASSIUM PERMANGANATE IN WATER. ADD TINY AMOUNT OF SODIUM BISULFATE (TO MAKE SOLUTION SOUR). POUR IN A LITTLE HYDROGEN PEROXIDE (H_2O_2). COLOR DISAPPEARS AND OXYGEN IS LIBERATED.

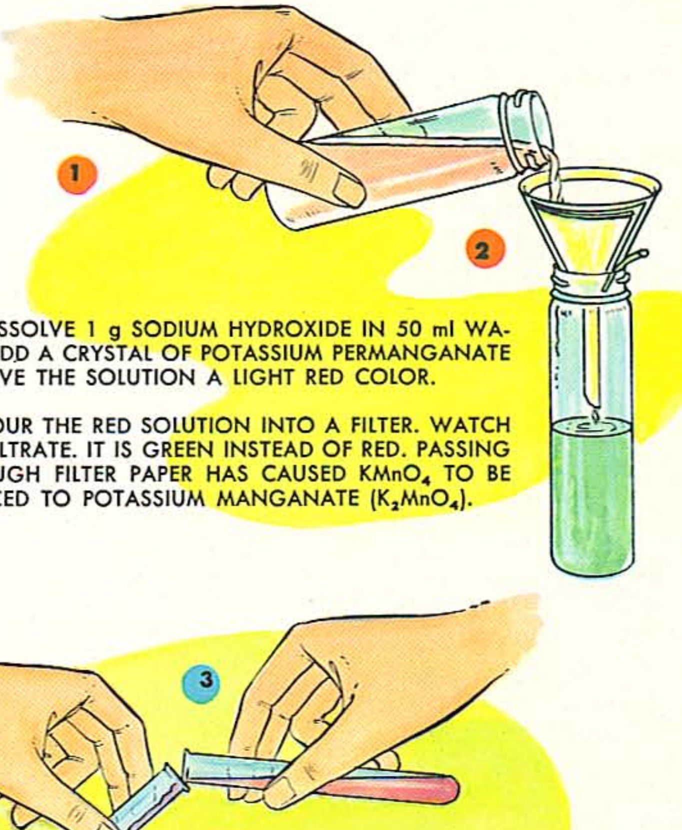

1 DISSOLVE 1 g SODIUM HYDROXIDE IN 50 ml WATER. ADD A CRYSTAL OF POTASSIUM PERMANGANATE TO GIVE THE SOLUTION A LIGHT RED COLOR.

2 POUR THE RED SOLUTION INTO A FILTER. WATCH THE FILTRATE. IT IS GREEN INSTEAD OF RED. PASSING THROUGH FILTER PAPER HAS CAUSED $KMnO_4$ TO BE REDUCED TO POTASSIUM MANGANATE (K_2MnO_4).

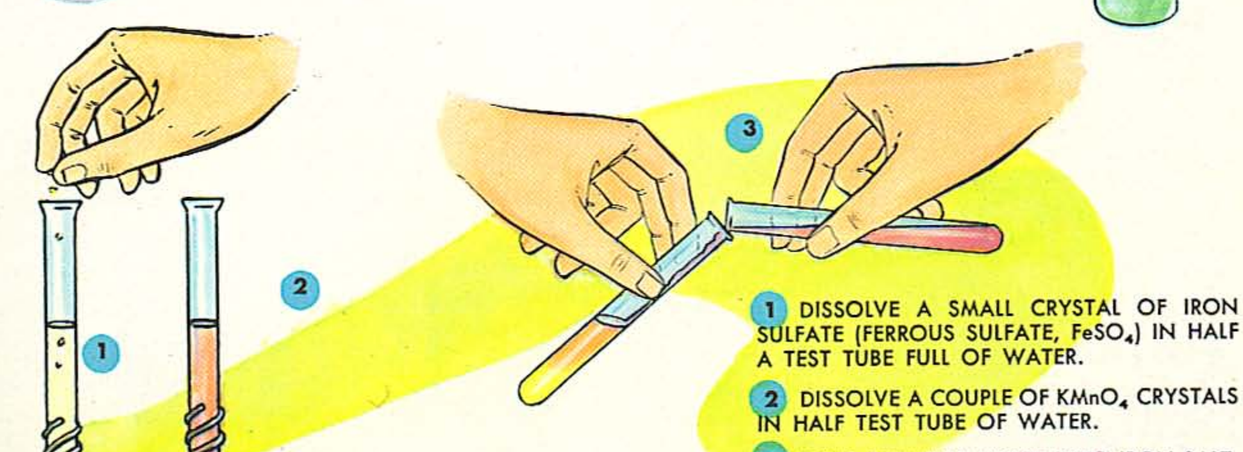

1 DISSOLVE A SMALL CRYSTAL OF IRON SULFATE (FERROUS SULFATE, $FeSO_4$) IN HALF A TEST TUBE FULL OF WATER.

2 DISSOLVE A COUPLE OF $KMnO_4$ CRYSTALS IN HALF TEST TUBE OF WATER.

3 POUR PERMANGANATE INTO IRON SALT. GREEN FERROUS SULFATE IS OXIDIZED TO BROWN FERRIC SULFATE ($Fe_2(SO_4)_3$).

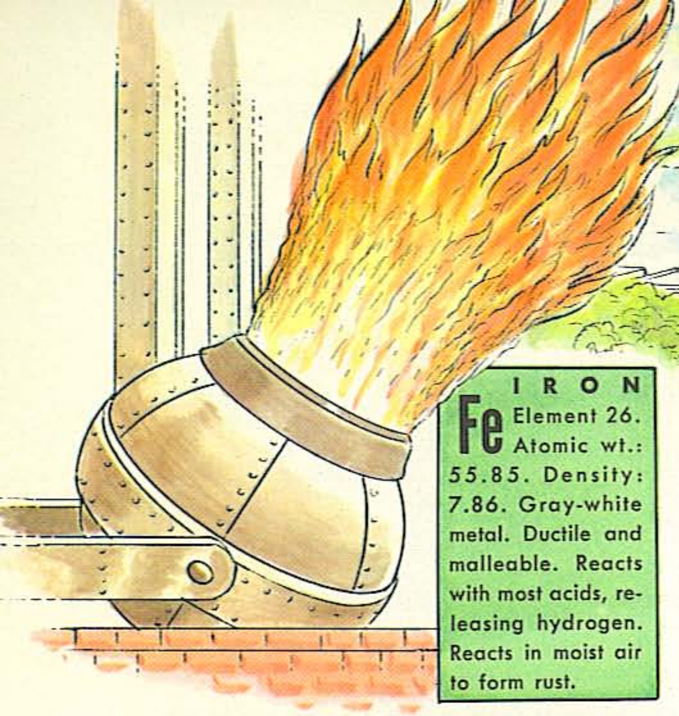

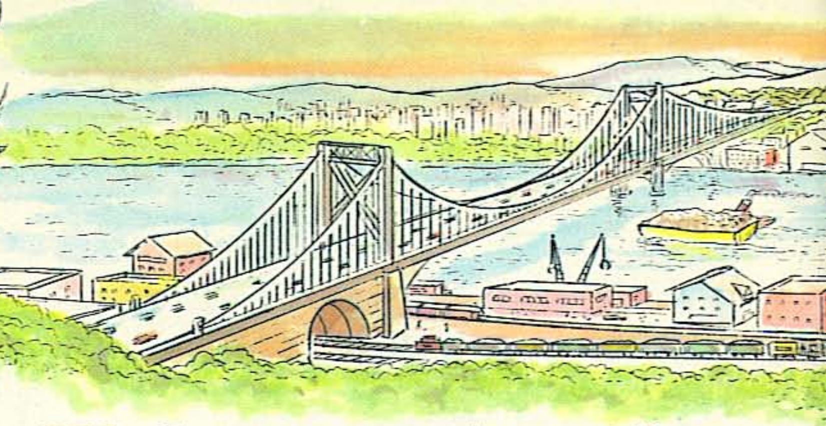

Fe **IRON**
Element 26.
Atomic wt.: 55.85. Density: 7.86. Gray-white metal. Ductile and malleable. Reacts with most acids, releasing hydrogen. Reacts in moist air to form rust.

We Live in an Age of Iron

IRON METAL has the peculiar quality of being magnetic — that is, of being attracted and influenced by a force called magnetism. If you should walk around your home and touch different things with a magnet, you would be surprised at the large number of them that would prove to contain iron. They would range in size from the car in the garage and the refrigerator and stove in the kitchen to the nails in the walls and the needles and pins in your mother's sewing box.

The moment you step outdoors and look around, you will be even more amazed. Skyscrapers and

TWO KINDS OF IRON COMPOUNDS

IRON FORMS TWO KINDS OF COMPOUNDS. IN FERROUS SALTS, EACH IRON ATOM HAS REPLACED TWO HYDROGEN ATOMS. IN FERRIC SALTS, EACH IRON ATOM HAS REPLACED THREE HYDROGEN ATOMS. GREEN FERROUS SALTS EASILY OXIDIZE INTO RED-BROWN FERRIC SALTS.

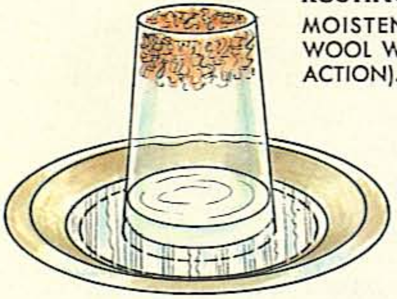

RUSTING OF IRON

MOISTEN A WAD OF FINE STEEL WOOL WITH VINEGAR (TO SPEED UP ACTION). WEDGE IT IN BOTTOM OF A GLASS. INVERT GLASS IN PIE PLATE OF WATER. IN A FEW DAYS, WATER HAS RISEN IN GLASS. IRON HAS REACTED WITH OXYGEN AND MOISTURE TO FORM RUST—$(Fe_2O_3)_2 \cdot 3H_2O$.

MAKING A FERRIC SALT

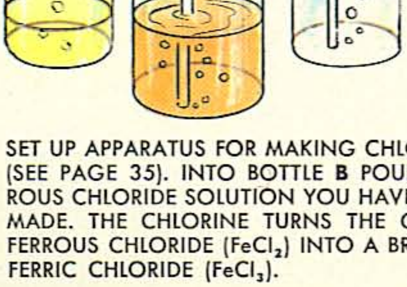

SET UP APPARATUS FOR MAKING CHLORINE (SEE PAGE 35). INTO BOTTLE **B** POUR FERROUS CHLORIDE SOLUTION YOU HAVE JUST MADE. THE CHLORINE TURNS THE GREEN FERROUS CHLORIDE ($FeCl_2$) INTO A BROWN FERRIC CHLORIDE ($FeCl_3$).

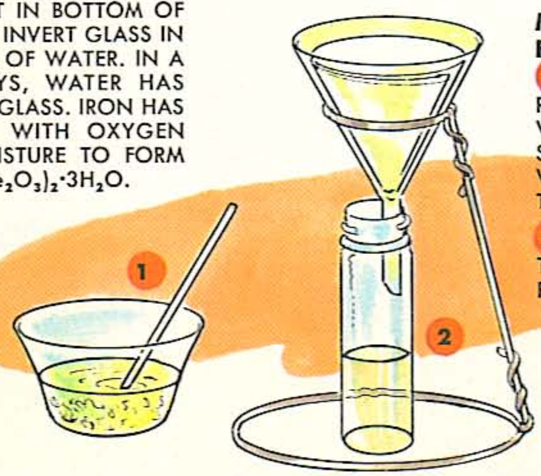

MAKING A FERROUS SALT

1 POUR HYDROCHLORIC ACID OVER STEEL WOOL. HYDROGEN IS SET FREE AS STEEL WOOL DISSOLVES. FILTER THE SOLUTION.

2 LIGHT-GREEN FILTRATE CONTAINS FERROUS CHLORIDE ($FeCl_2$).

TEST FOR IRON SALTS

1 IN ONE TEST TUBE, DILUTE SOME FERRIC CHLORIDE SOLUTION WITH WATER.

2 IN ANOTHER, DILUTE SOME OF THE FERROUS CHLORIDE SOLUTION WITH WATER.

3 TO EACH, ADD A FEW DROPS OF SOLUTION OF ¼ TEASPOON POTASSIUM FERROCYANIDE IN 50 ml WATER. FERRIC SALT MAKES A DEEP BLUE PRECIPITATE OF PRUSSIAN BLUE. FERROUS SALT MAKES LIGHT BLUE PRECIPITATE.

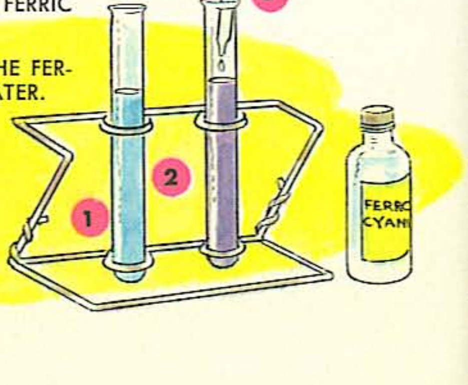

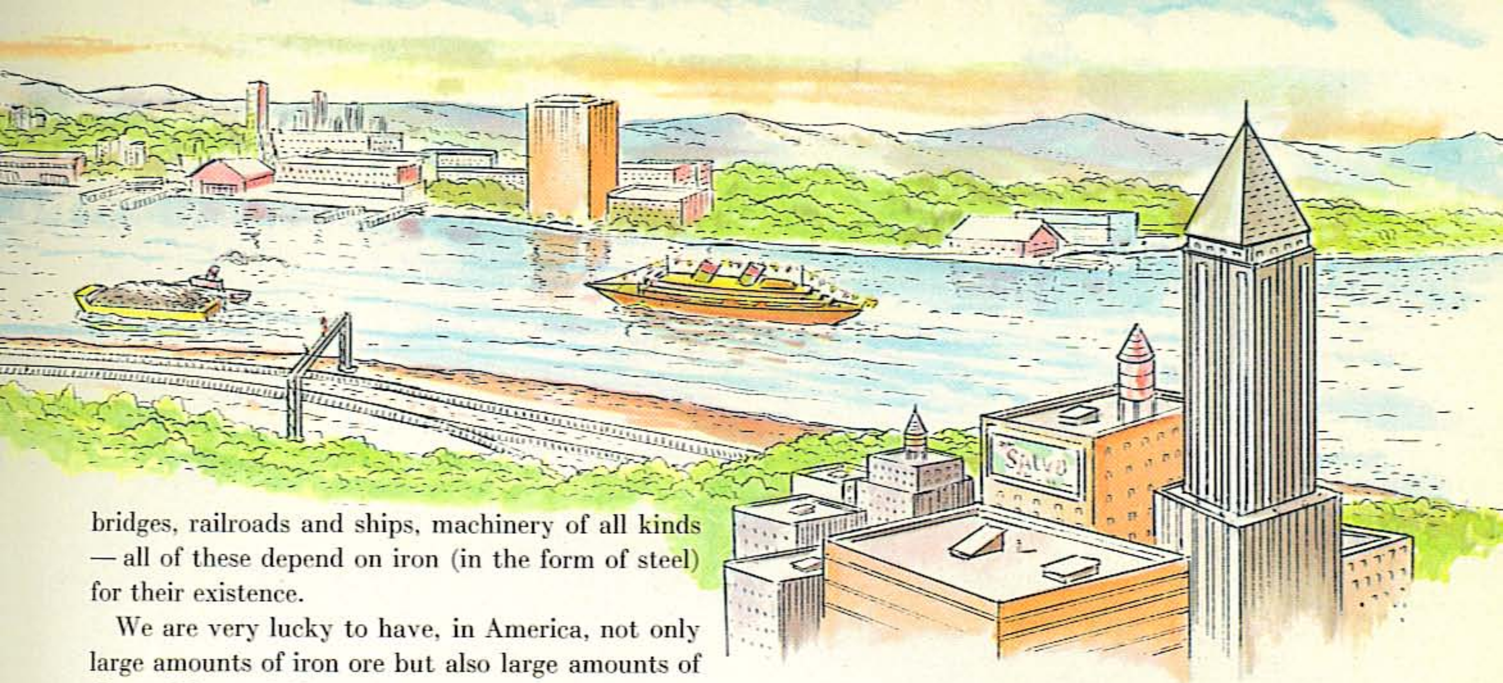

bridges, railroads and ships, machinery of all kinds — all of these depend on iron (in the form of steel) for their existence.

We are very lucky to have, in America, not only large amounts of iron ore but also large amounts of the coal from which to make the coke that goes into iron production.

The iron is driven out of its ore (mostly Fe_2O_3) in huge furnaces. Each furnace can make as much as 1,000 tons of iron at one time from 2,000 tons of ore, 1,000 tons of coke, and 500 tons of limestone. A blast of hot air is forced through the mixture. The coke burns with great heat to carbon dioxide. This, with more coke, forms carbon monoxide, and this, in turn, reduces the iron oxide to metallic iron. In chemical language, this is what happens:

$$C + O_2 \rightarrow CO_2 \text{ plus heat}$$
$$CO_2 + C \rightarrow 2CO$$
$$Fe_2O_3 + 3CO \rightarrow 2Fe + 3CO_2$$

At the same time, the limestone combines with various impurities to form a glass-like compound called "slag." This is removed when the white-hot iron is poured out into moulds and cooled into bars of "pig iron."

The pig iron is brittle because it contains close to 5 per cent carbon. To turn it into steel, the carbon must be burned out until only from .5 to 1.5 per cent remains. This is done either by the Bessemer process (named for an Englishman, Henry Bessemer) or by the open-hearth process. The finished steel is molded into "ingots" and shipped to manufacturing plants all over the country.

In chemical experiments, the most commonly used iron compound is the iron sulfate (ferrous sulfate, $FeSO_4 \cdot 7H_2O$) — also called "green vitriol" and "copperas." Don't let the last name mislead you— it has nothing to do with copper but comes from an old French word, *couperose.*

FROM FERROUS SALT TO FERRIC

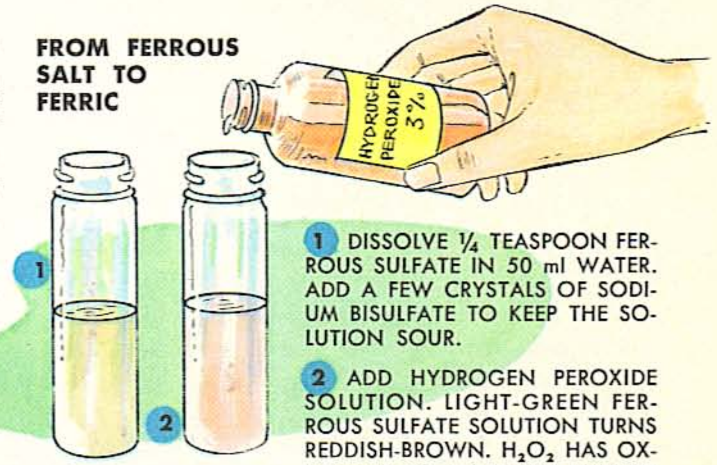

① DISSOLVE ¼ TEASPOON FERROUS SULFATE IN 50 ml WATER. ADD A FEW CRYSTALS OF SODIUM BISULFATE TO KEEP THE SOLUTION SOUR.

② ADD HYDROGEN PEROXIDE SOLUTION. LIGHT-GREEN FERROUS SULFATE SOLUTION TURNS REDDISH-BROWN. H_2O_2 HAS OXIDIZED $FeSO_4$ TO FERRIC SULFATE ($Fe_2(SO_4)_3$).

IRON HYDROXIDES AND CARBONATE

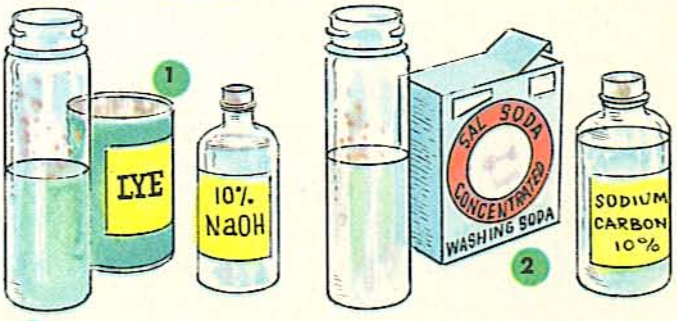

① TO SOLUTION OF ¼ TEASPOON FERROUS SULFATE IN 50 ml WATER, ADD SOLUTION OF SODIUM HYDROXIDE. PURE FERROUS HYDROXIDE IS WHITE. BECAUSE OF IMPURITIES, YOU GET DIRTY-GREEN PRECIPITATE OF $Fe(OH)_2$, SOON OXIDIZING TO BROWN FERRIC HYDROXIDE.

② TO ANOTHER PORTION OF FERROUS SULFATE SOLUTION ADD SODIUM CARBONATE SOLUTION. PURE CARBONATE MADE WITH NO OXYGEN PRESENT IS WHITE— BUT YOU GET MUDDY, WHITISH-GREEN PRECIPITATE OF FERROUS CARBONATE, EVENTUALLY TURNING INTO FERRIC HYDROXIDE.

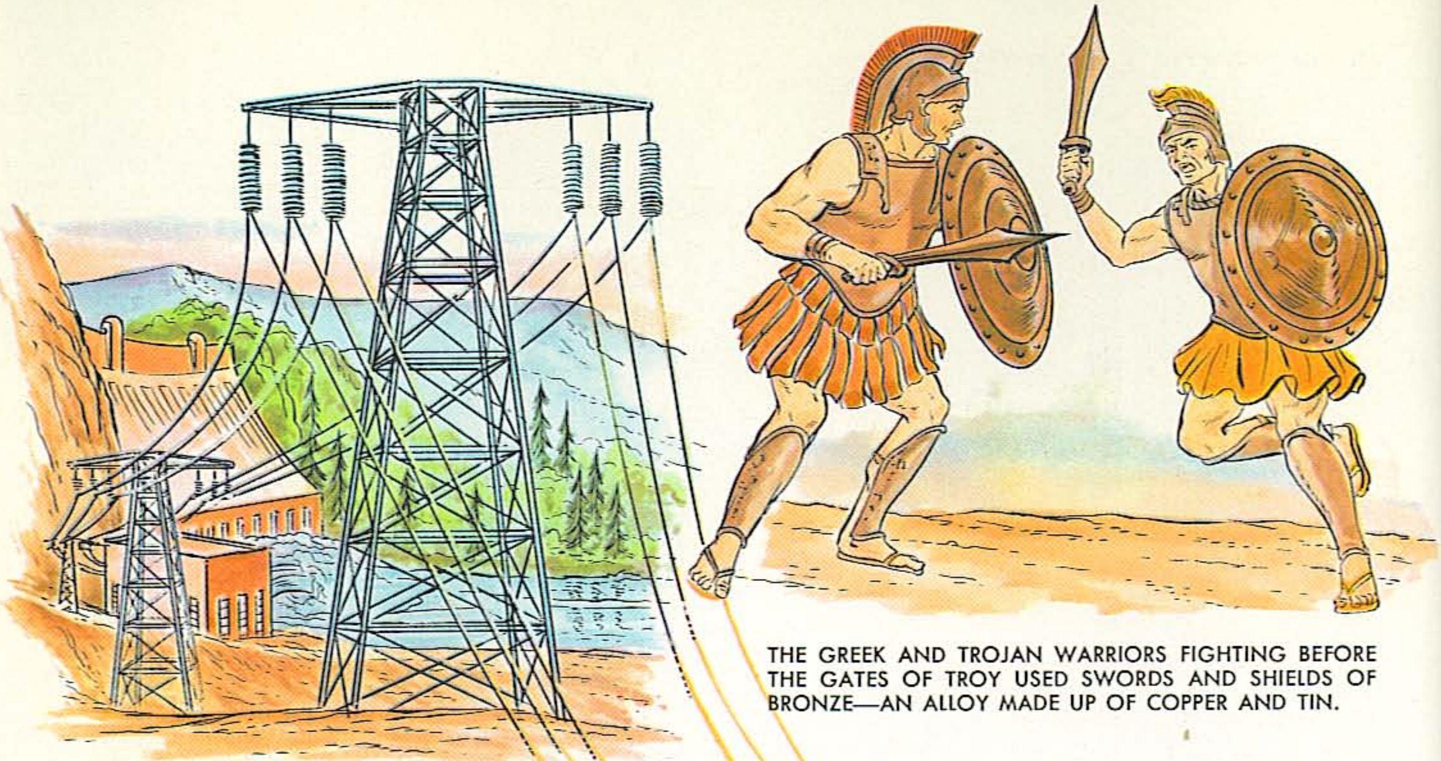

THE GREEK AND TROJAN WARRIORS FIGHTING BEFORE THE GATES OF TROY USED SWORDS AND SHIELDS OF BRONZE—AN ALLOY MADE UP OF COPPER AND TIN.

Copper—Yesterday, Today

COPPER IS ONE of the few metals found free in nature. That is why it was used long before historic times for weapons and utensils. The main trouble with it was its softness. This was remedied when some early coppersmith discovered that copper and tin (also found free in nature) melted together formed an alloy that was much harder than either of the two metals. This alloy gave its name to more than two thousand years of human history — the period called the "Bronze Age."

A great number of weapons from the Bronze Age have been found in Greece. When they were dug out of the ground, they were covered with a green "rust." This deposit was called verdigris — literally "green of Greece" (from old French, *vert de Grèce*). It consists of basic cupric carbonate — the same compound you will see on a bronze statue or a copper-clad church spire exposed to wind and weather.

Copper became especially valuable less than a hundred years ago when a satisfactory method for producing a steady flow of electricity was invented. After silver, copper is the best conductor of electricity. Today, the most important use for copper is for electrical purposes. It serves to bring the current from the place where it is produced to the place where it is to be used (although, within recent years, some aluminum has taken its place for high-tension wires). You will find copper in the wiring in your own home and in every electrical gadget you use.

Copper makes two kinds of salts. In cuprous salts, one copper atom has taken the place of one hydrogen atom; in cupric salts, one copper atom has taken the place of two hydrogen atoms. Cuprous salts (such as cuprous chloride, $CuCl$) are colorless, while cupric salts (such as cupric sulfate, $CuSO_4 \cdot 5H_2O$) are bright blue in color.

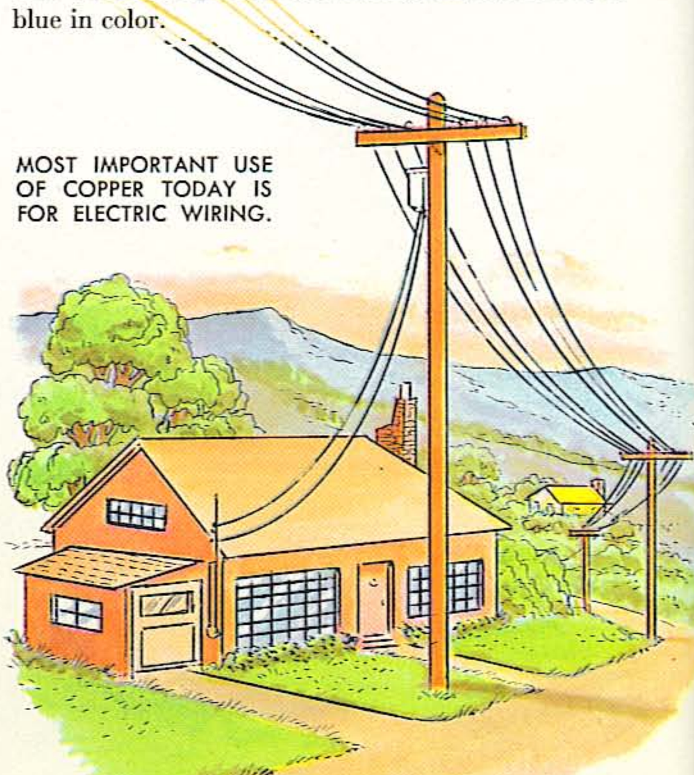

MOST IMPORTANT USE OF COPPER TODAY IS FOR ELECTRIC WIRING.

MAKING COPPER COMPOUNDS

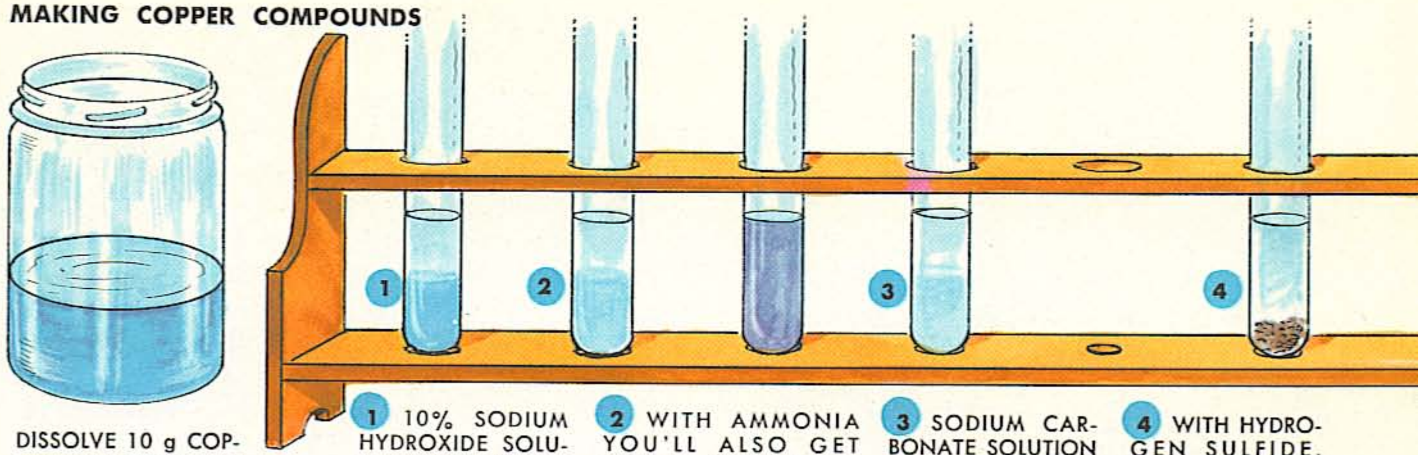

DISSOLVE 10 g COPPER SULFATE IN 100 ml WATER. POUR 10 ml INTO EACH OF FOUR TEST TUBES.

1 10% SODIUM HYDROXIDE SOLUTION PRECIPITATES DIRTY-BLUE CUPRIC HYDROXIDE ($Cu(OH)_2$).

2 WITH AMMONIA YOU'LL ALSO GET $Cu(OH)_2$, BUT THIS DISSOLVES IN MORE AMMONIA WITH DEEP BLUE COLOR.

3 SODIUM CARBONATE SOLUTION GIVES BLUE-GREEN CUPRIC CARBONATE PRECIPITATE.

4 WITH HYDROGEN SULFIDE, BROWNISH-BLACK PRECIPITATE OF CUPRIC SULFIDE.

Cu COPPER Element 29. Atomic wt.: 63.54. Density: 8.97. Soft metal of reddish color. Easily rolled and drawn into wire. Next to silver, is the best conductor of electricity.

FROM CUPRIC SALT TO CUPROUS SALT

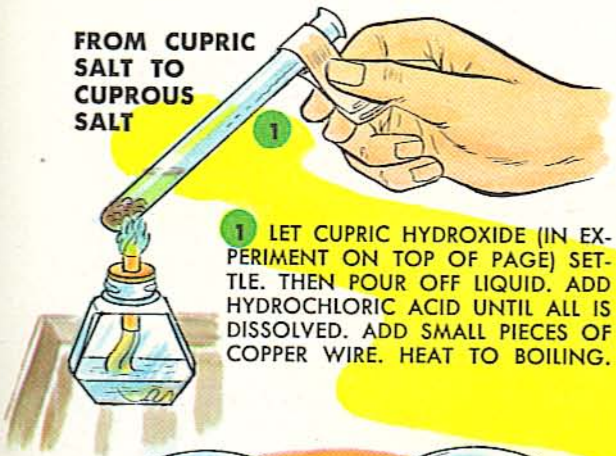

1 LET CUPRIC HYDROXIDE (IN EXPERIMENT ON TOP OF PAGE) SETTLE. THEN POUR OFF LIQUID. ADD HYDROCHLORIC ACID UNTIL ALL IS DISSOLVED. ADD SMALL PIECES OF COPPER WIRE. HEAT TO BOILING.

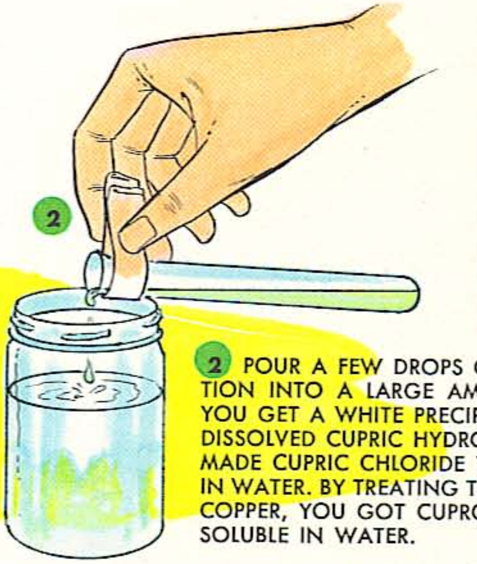

2 POUR A FEW DROPS OF THE HOT SOLUTION INTO A LARGE AMOUNT OF WATER. YOU GET A WHITE PRECIPITATE. WHEN YOU DISSOLVED CUPRIC HYDROXIDE IN HCl, YOU MADE CUPRIC CHLORIDE WHICH IS SOLUBLE IN WATER. BY TREATING THIS WITH METALLIC COPPER, YOU GOT CUPROUS CHLORIDE, INSOLUBLE IN WATER.

COPPER SULFATE IN CHEMICAL ANALYSIS

WATERFREE (ANHYDROUS) CUPRIC SULFATE SHOWS IF WATER IS PRESENT IN A LIQUID BEING TESTED.

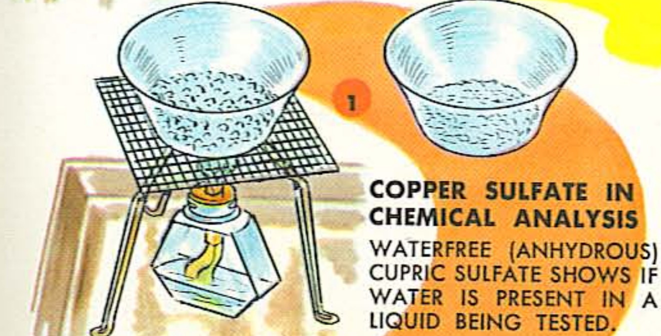

1 CRUSH A FEW CUPRIC SULFATE CRYSTALS. HEAT WHILE STIRRING UNTIL THEY HAVE TURNED INTO A WHITE POWDER.

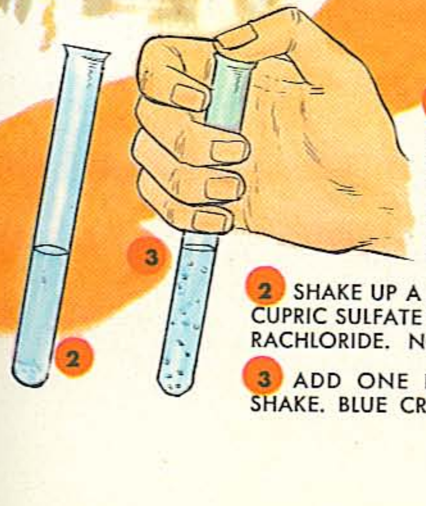

2 SHAKE UP A LITTLE ANHYDROUS CUPRIC SULFATE WITH CARBON TETRACHLORIDE. NOTHING HAPPENS.

3 ADD ONE DROP OF WATER. SHAKE. BLUE CRYSTALS FORM.

REPLACING COPPER WITH IRON

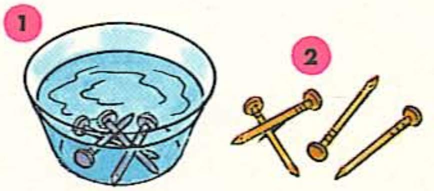

1 DROP SEVERAL CLEAN NAILS INTO A SOLUTION OF COPPER SULFATE. LEAVE FOR HALF AN HOUR.

2 NAILS ARE NOW COATED WITH METALLIC COPPER AND THE SOLUTION CONTAINS FERROUS SULFATE.

METALS CAN BE ARRANGED IN A REPLACEMENT SERIES. ANY METAL IN THE SERIES WILL DRIVE OUT ANOTHER METAL BELOW IT AND TAKE ITS PLACE IN THE SALT.

REPLACEMENT SERIES

POTASSIUM
SODIUM
CALCIUM
MAGNESIUM
ALUMINUM
ZINC
CHROMIUM
IRON
NICKEL
TIN
LEAD
COPPER
MERCURY
SILVER
PLATINUM
GOLD

Silver—One of the "Noble" Metals

SILVER — LIKE COPPER and gold — is found free in nature and was therefore known to man long before he learned to extract metals from their ores.

Pure silver has one drawback — it is almost as soft as copper. That's why most silver is alloyed with copper to make it harder. Sterling silver — a famous alloy used for jewelry — contains 7.5 parts copper to 92.5 parts silver. Only 9/10 of American silver coins is silver — the rest is copper.

When you take a snapshot, you get yourself involved in a series of complicated, chemical processes — all of them having to do with silver. It is hard to believe that the blacks and grays you see when you look at a photograph are various concentrations of metallic silver.

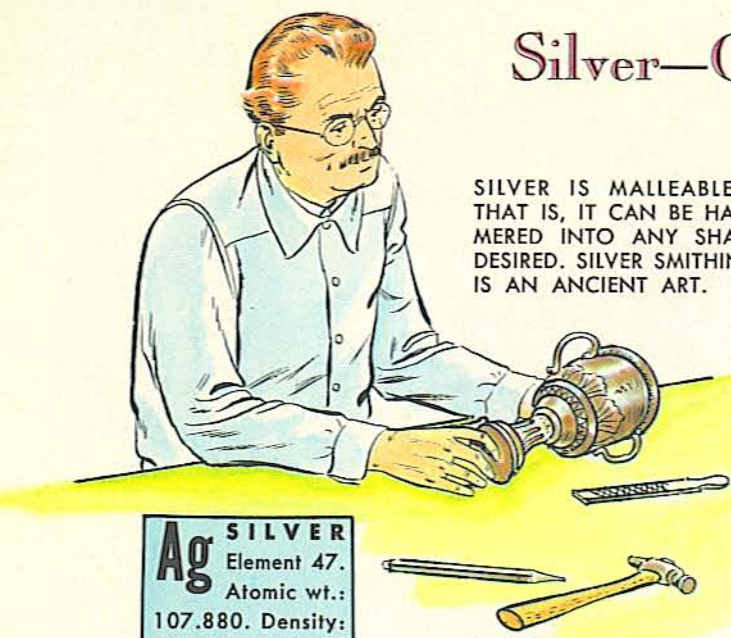

SILVER IS MALLEABLE—THAT IS, IT CAN BE HAMMERED INTO ANY SHAPE DESIRED. SILVER SMITHING IS AN ANCIENT ART.

Ag SILVER Element 47. Atomic wt.: 107.880. Density: 10.54. Soft, white metal with bright luster. Easily hammered out and drawn into wire. Best known conductor of electricity.

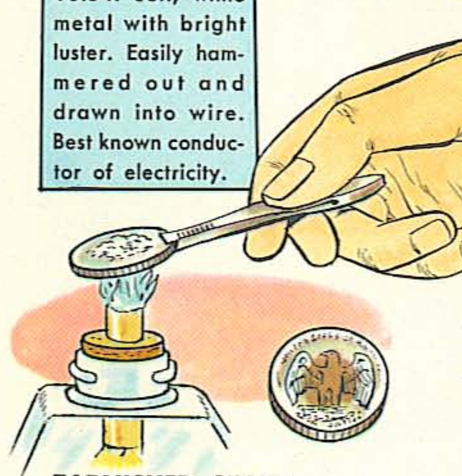

TARNISHED SILVER

SILVER TARNISHES WHEN IT IS EXPOSED TO SULFUR. PLACE A FEW CRYSTALS OF SODIUM THIOSULFATE ("HYPO") ON A SILVER COIN. HEAT UNTIL HYPO MELTS. WASH. HYPO HAS LEFT STAIN OF BROWN-BLACK SILVER SULFIDE.

REMOVING TARNISH FROM SILVER

1 LINE BOTTOM OF CUSTARD CUP WITH ALUMINUM FOIL. PLACE TARNISHED COIN ON FOIL. POUR CUP HALF FULL OF WATER. ADD ¼ TEASPOON SODIUM CARBONATE. BOIL GENTLY. TARNISH VANISHES.

2 YOU CAN USE THIS METHOD FOR CLEANING SILVERWARE. PLACE SILVER TO BE CLEANED IN ALUMINUM TRAY. ADD WATER AND SODA. BRING TO A BOIL. THE SILVER BECOMES SHINY AGAIN.

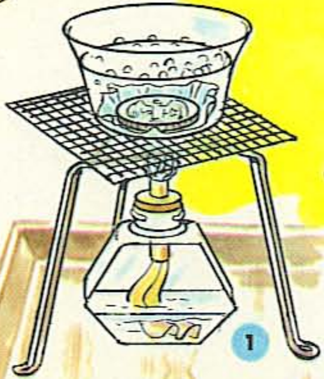

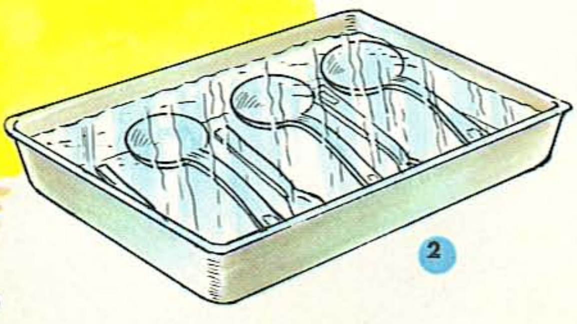

SILVER COMPOUNDS

GET 5 g SILVER NITRATE IN YOUR LOCAL DRUG STORE. DISSOLVE IN 50 ml WATER.

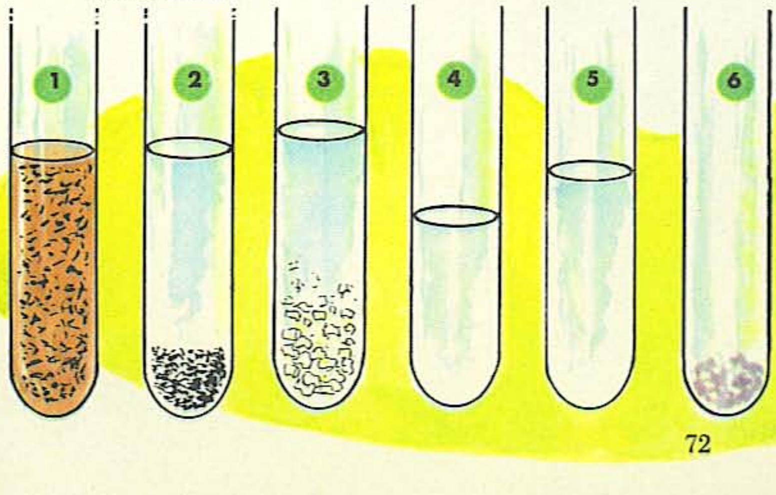

1 TO 5 ml SILVER NITRATE ($AgNO_3$) SOLUTION, ADD SODIUM HYDROXIDE SOLUTION. YOU GET DARK-BROWN PRECIPITATE—NOT OF HYDROXIDE, BUT OF SILVER OXIDE.

2 TO 5 ml $AgNO_3$ SOLUTION, ADD AMMONIA. PRECIPITATE OF SILVER OXIDE DISSOLVES WHEN YOU ADD MORE AMMONIA.

3 TO 5 ml $AgNO_3$ SOLUTION, ADD TABLE SALT ($NaCl$) SOLUTION. CHEESELIKE PRECIPITATE IS SILVER CHLORIDE ($AgCl$).

4 TO PART OF $AgCl$ PRECIPITATE, ADD AMMONIA. SILVER CHLORIDE DISSOLVES.

5 TO ANOTHER PART OF $AgCl$, ADD SODIUM THIOSULFATE SOLUTION. $AgCl$ DISSOLVES.

6 PLACE REMAINING $AgCl$ IN THE SUN. IT TURNS VIOLET FROM METALLIC SILVER.

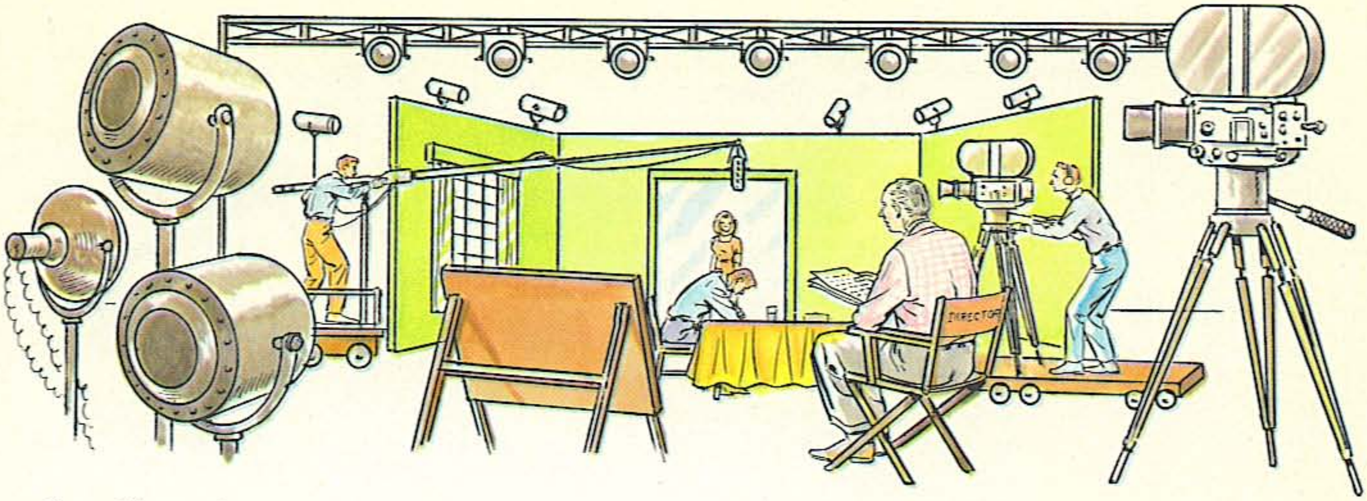

In making a photographic film, the manufacturer spreads an emulsion of gelatin that contains silver bromide (AgBr) over a transparent sheet of cellulose acetate. When the silver bromide is exposed to light, a certain amount of it gives up metallic silver ($AgBr \rightarrow Ag + Br$). More of this silver is brought out in the developing bath. When fully developed, the film is placed in a fixing bath which removes all unexposed silver bromide. After washing and drying, you have a photographic negative in which the white parts you photographed appear black and the black parts appear white.

To make a natural-looking picture, you place the negative on a piece of photographic paper and go through a similar procedure, as above, of exposing, developing, fixing, washing, and drying.

MORE THAN 150 TONS OF SILVER ARE USED EACH YEAR IN MAKING FILM FOR THE MOVIES.

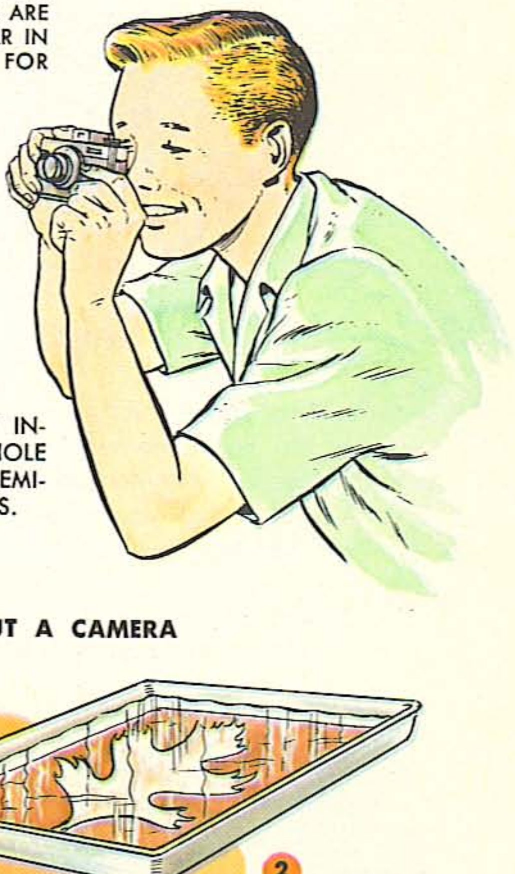

PHOTOGRAPHY INVOLVES A WHOLE SERIES OF CHEMICAL PROCESSES.

PHOTOGRAPHING WITHOUT A CAMERA

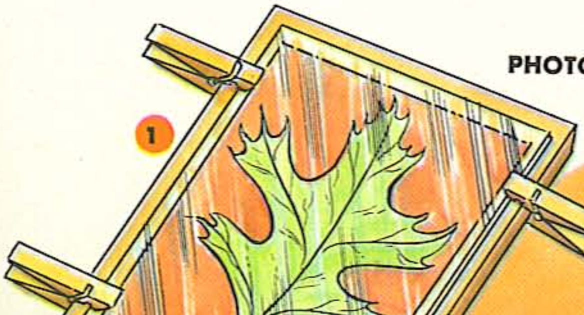

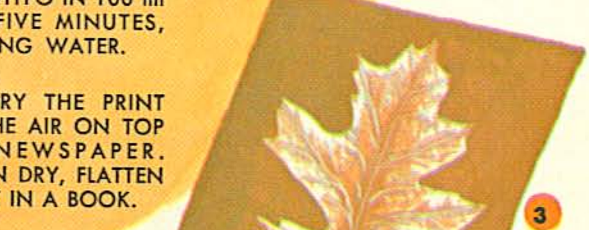

2 FIX THE LEAF PRINT IN A SOLUTION OF 10 g HYPO IN 100 ml WATER. AFTER FIVE MINUTES, WASH IN RUNNING WATER.

3 DRY THE PRINT IN THE AIR ON TOP OF NEWSPAPER. WHEN DRY, FLATTEN PRINT IN A BOOK.

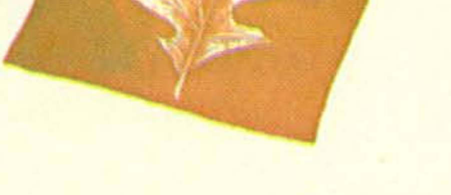

1 FROM A COMMERCIAL PHOTOGRAPHER, GET A FEW SHEETS OF "PRINTING-OUT PAPER," A SLOW PHOTOGRAPHIC PAPER. IN THE SHADE, PLACE SHEET ON PIECE OF PLYWOOD, SENSITIZED SIDE UP. ON TOP OF IT, LAY A LEAF AND A SHEET OF GLASS. HOLD IN POSITION WITH SPRING CLOTHES PINS. EXPOSE TO SUN UNTIL PAPER IS BLACKISH-VIOLET.

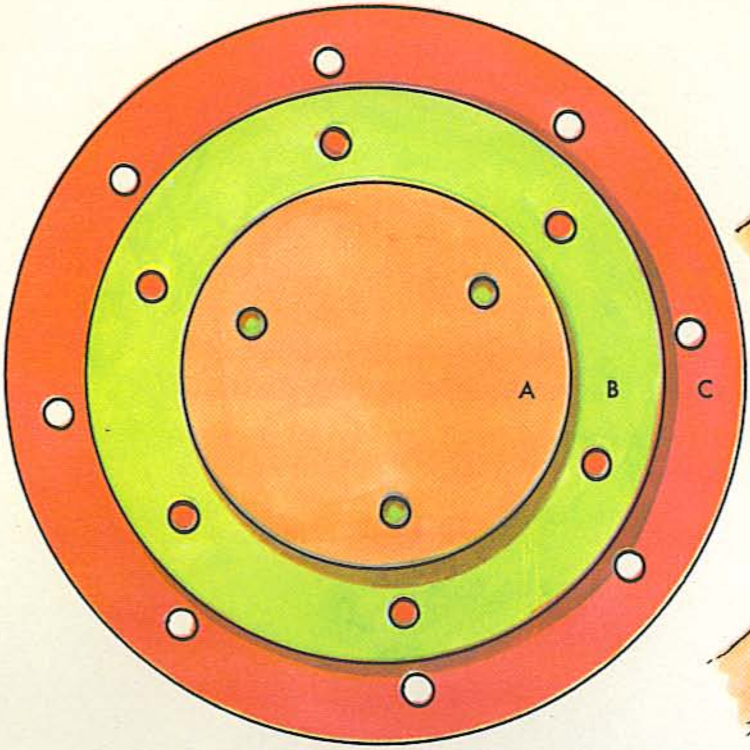

TRACE EACH OF THE CIRCLES SHOWN ABOVE ONTO CARDBOARD. PUNCH HOLES AS INDICATED. USE AS PATTERNS FOR CUTTING CIRCLES OF CONSTRUCTION BOARD.

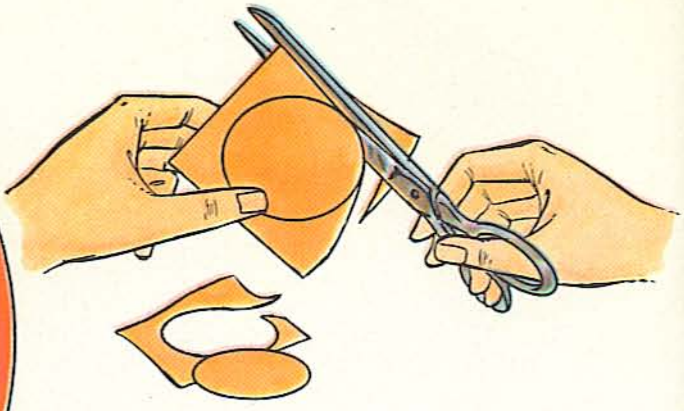

USING PATTERNS AT LEFT, CUT OUT SEVERAL CIRCLES OF CONSTRUCTION BOARD IN VARIOUS COLORS.

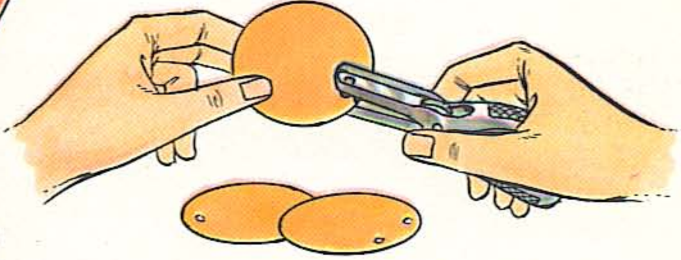

PUNCH THE HOLES NECESSARY TO INDICATE VALENCES.

Valences and Formulas

As you have studied the chemical formulas in the text, you will have noticed that one atom of hydrogen combines with one atom of chlorine (HCl), two hydrogen atoms with one atom of oxygen (H_2O), and three hydrogen atoms with one atom of nitrogen (NH_3).

The capacity of one atom to hold on to other atoms is called its valence (from Latin *valentia*, strength). No atom has a lower valence than the hydrogen atom, so we use hydrogen as our starting point and give it a valence of 1. Two hydrogen atoms combine with one oxygen atom — that gives oxygen a valence of 2. Nitrogen has a valence of 3. Two oxygen atoms combine with one carbon atom to make CO_2. Carbon has a valence of 4.

The chart on page 75 shows some of the common

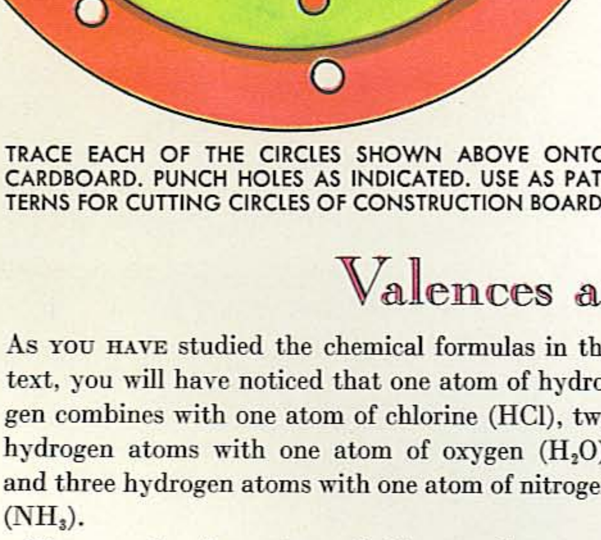

THESE DIAGRAMS SHOW WHAT HAPPENS WHEN YOU BURN CARBON AND TEST FOR CO_2. ONE CARBON ATOM (WITH FOUR POSITIVE VALENCES) COMBINES WITH TWO ATOMS OF OXYGEN (EACH WITH TWO NEGATIVE VALENCES) TO FORM ONE MOLECULE OF CO_2 (ARROW POINTING UP INDICATES THAT THIS IS A GAS). ONE MOLECULE CARBON DIOXIDE COMBINES WITH ONE MOLECULE CALCIUM HYDROXIDE TO FORM ONE MOLECULE OF CALCIUM CARBONATE (ARROW POINTING DOWN INDICATES THAT IT IS A PRECIPITATE) AND ONE MOLECULE OF WATER.

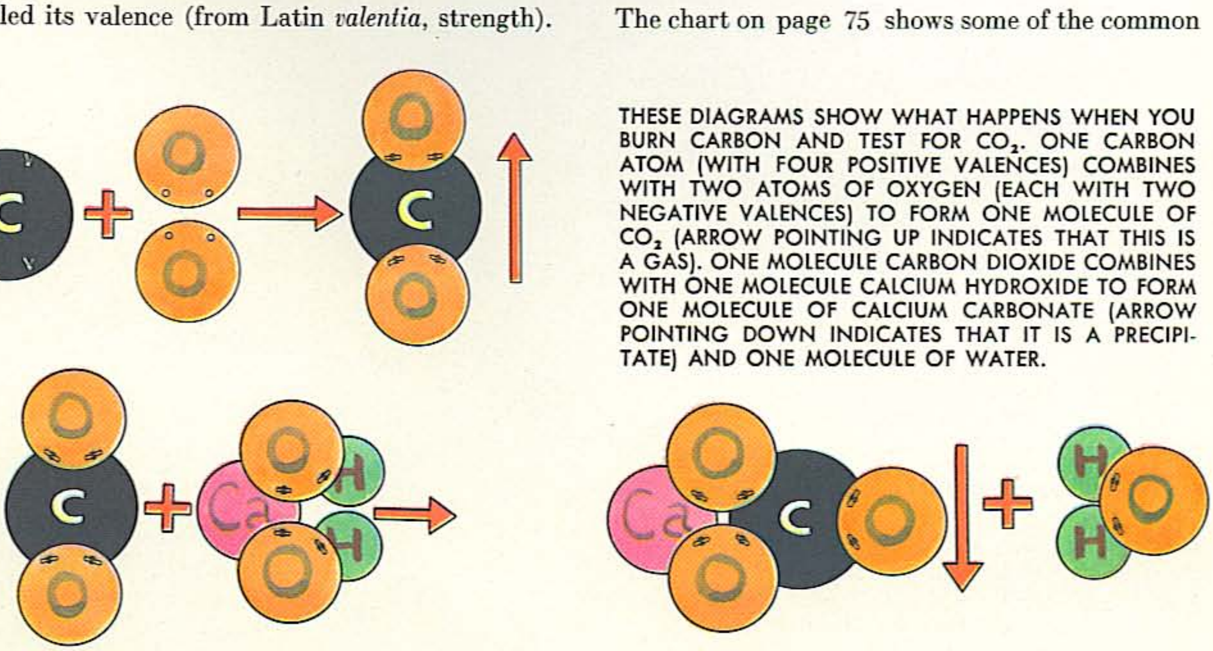

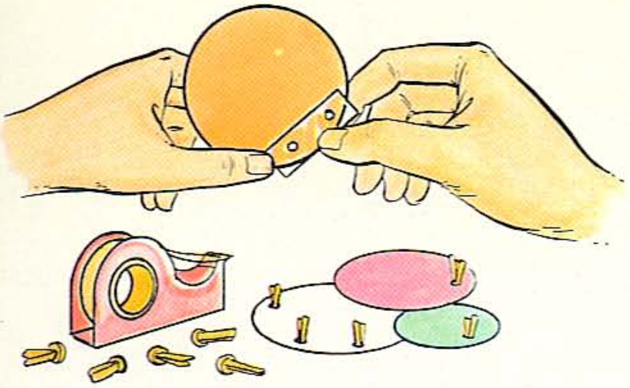

PUT ¼" BRASS CLIPS IN HOLES SHOWING POSITIVE VALENCES. HOLD THEM IN PLACE WITH SCOTCH TAPE.

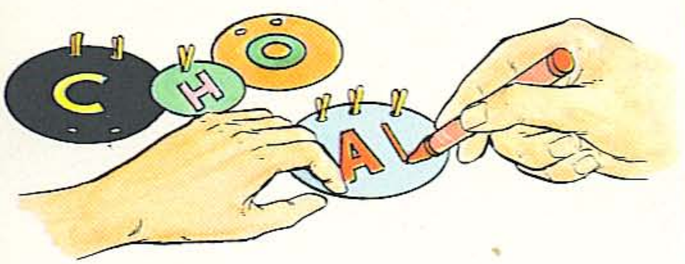

WRITE THE NAMES OF THE ELEMENTS WITH CRAYONS.

SOME COMMON VALENCES

Positive Valences			Negative Valences		
Item	Valence	Circle	Item	Valence	Circle
Ag	+1	A	Cl	−1	A
Al	+3	C	I	−1	A
B	+3	C	N	−3	B
C	+4	C	O	−2	B
Ca	+2	B	S	−2	C
Cu	+1+2	B			
Fe	+2+3	B			
H	+1	A			
K	+1	A			
Mg	+2	B			
Mn	+2	B	CO_3	−2	B
Na	+1	A	NO_3	−1	A
S	+4+6	C	OH	−1	A
Si	+4	C	SO_3	−2	B
NH_4	+1	A	SO_4	−2	B

valences for making up formulas. Most of the items are elements, but some of them are "radicals" — that is, groups of atoms that hang together in chemical reactions, such as the ammonium radical (NH_4) that behaves as a metal, and the sulfate radical (SO_4) that goes into the making of salts.

Notice that some valences have plus (+) signs, others have minus (—) signs. When you make up the formula for a compound, there must be the same number of pluses and minuses. Hydrogen with one plus (H^+) and oxygen with two minuses (O^{--}) would not fit together — you need H_2 to combine with O. Similarly, C with + 4 (C^{++++}) takes two O, each with —2 (O^{--}), in order to balance.

To get a clear understanding of chemical formulas, make yourself a set of atom models as shown on these pages. With these models you will be able to figure out how compounds are made up and what happens in the various chemical reactions you will cause in your experiments.

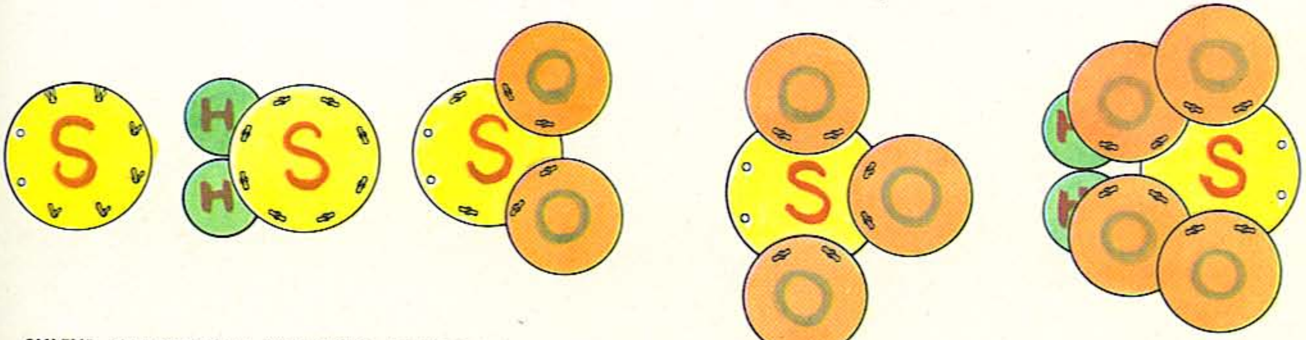

SULFUR HAS SEVERAL VALENCES. IT HAS A VALENCE OF −2 IN H_2S, OF +4 IN SO_2, AND OF +6 IN SO_3 AND IN SULFURIC ACID (H_2SO_4). IN MAKING THE CIRCLE FOR SULFUR, YOU CAN SHOW THESE VALENCES WITH TWO EMPTY HOLES AND SIX BRASS CLIPS.

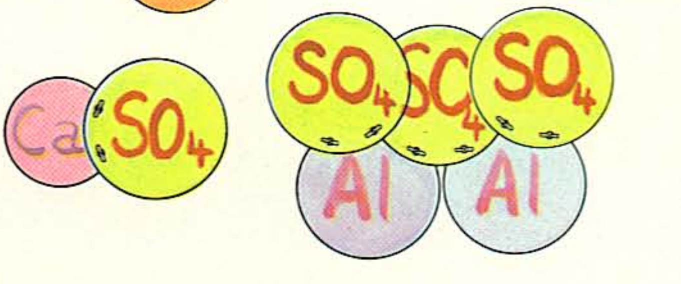

INSTEAD OF USING ONE SULFUR CIRCLE AND FOUR OXYGEN CIRCLES TO INDICATE A SULFATE, YOU CAN MAKE UP A SINGLE CIRCLE TO STAND FOR THE SULFATE RADICAL (SO_4) WITH TWO NEGATIVE VALENCES.

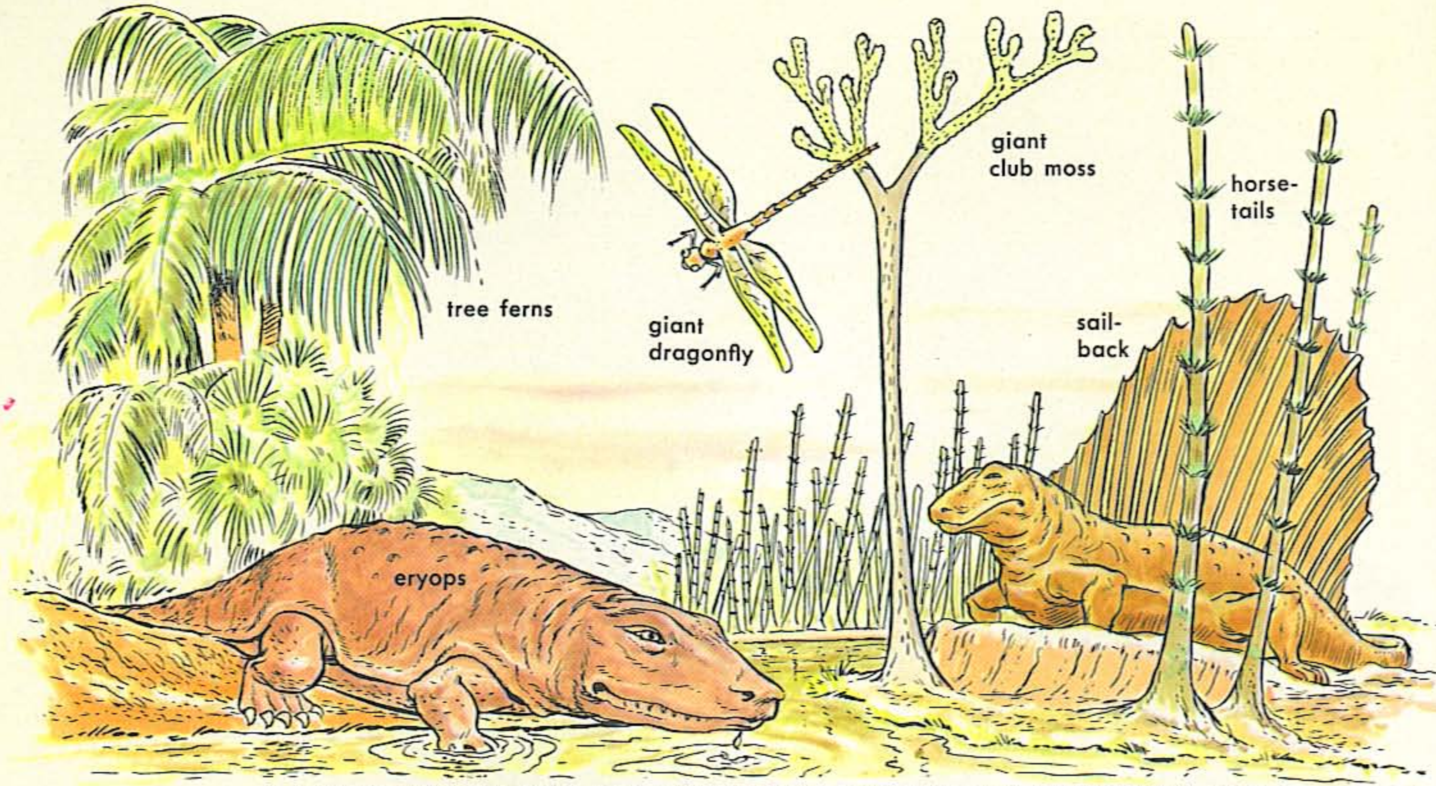

A LANDSCAPE IN THE COAL AGE, APPROXIMATELY THREE HUNDRED MILLION YEARS AGO.

Carbon—Element of a Million Compounds

To THE OLD ROMANS, *carbo* meant coal — a black rock that would burn. To the modern chemist, carbon is an element found in all living things — plants and animals — and in many dead things. It is hidden in the whitest sugar and the reddest rose and the greenest apple, in hundreds of thousands of compounds produced by nature and in many thousands more created in the laboratory.

The soot from a smoking candle is almost pure carbon. So is also the graphite that forms the "lead" of your pencil and the diamond in the jeweler's window. The coal that we use for fuel contains from 80 to 90 per cent carbon — the other 10 to 20 per cent is made up of various substances from which a great number of important and valuable chemical compounds are made.

All the coal we mine deep underground today is made up of the remains of plants that grew around three hundred million years ago — huge tree ferns, giant club mosses and horsetails. They thrived in the hot, humid climate, died and tumbled to the ground. During the ages they were covered by other dead trees and by layers upon layers of mud. Eventually, pressure and heat turned them into coal.

PRESSURE AND HEAT TURNED TREES AND OTHER PLANTS INTO THE COAL WE USE TODAY.

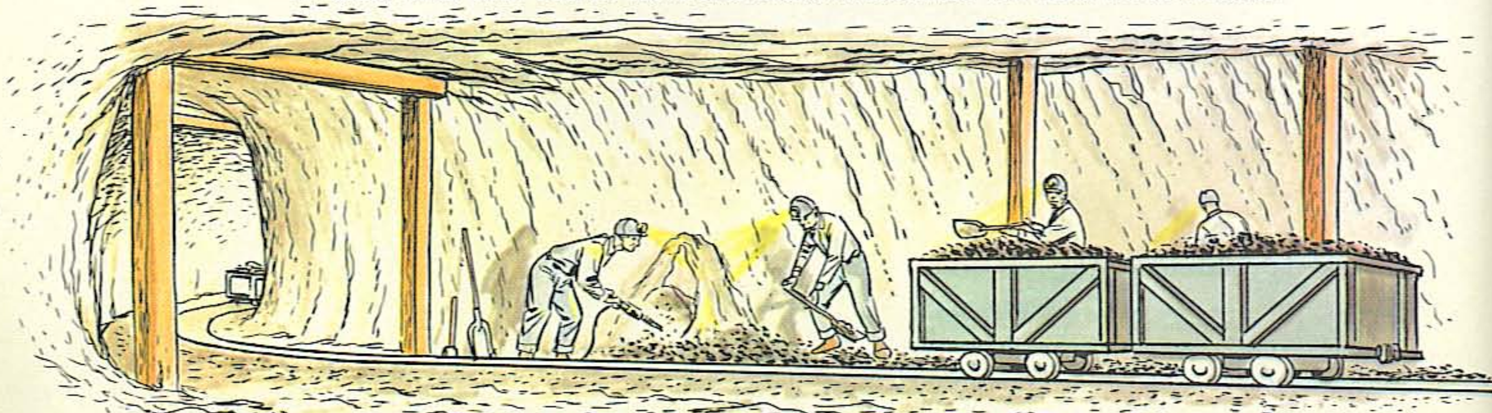

C CARBON Element 6. Atomic wt.: 12.011. Found in nature as diamond (density 3.52), as graphite (density 2.25), and as coal. Diamond burns in oxygen, coal in air.

THREE FORMS OF CARBON

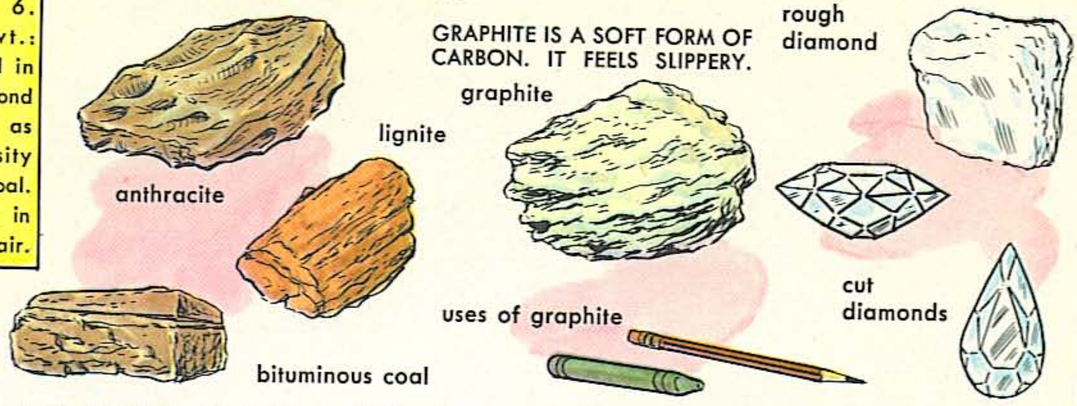

GRAPHITE IS A SOFT FORM OF CARBON. IT FEELS SLIPPERY.

ANTHRACITE IS ALSO CALLED HARD COAL. BITUMINOUS COAL IS SOFT COAL. LIGNITE IS BROWN COAL. GRAPHITE IS A SOFT FORM OF CARBON. IT FEELS SLIPPERY. DIAMOND IS THE HARDEST SUBSTANCE KNOWN.

DESTRUCTIVE DISTILLATION

IN REGULAR DISTILLATION (SEE PAGE 61), A CHEMICAL IS PURIFIED. IN DESTRUCTIVE OR DRY DISTILLATION, THE SUBSTANCE IS BROKEN INTO SEVERAL DIFFERENT CHEMICALS.

for dry distillation of wood, whittle twig into slivers, or use wooden matches without heads.

for dry distillation of coal, crush lump of bituminous coal into powder.

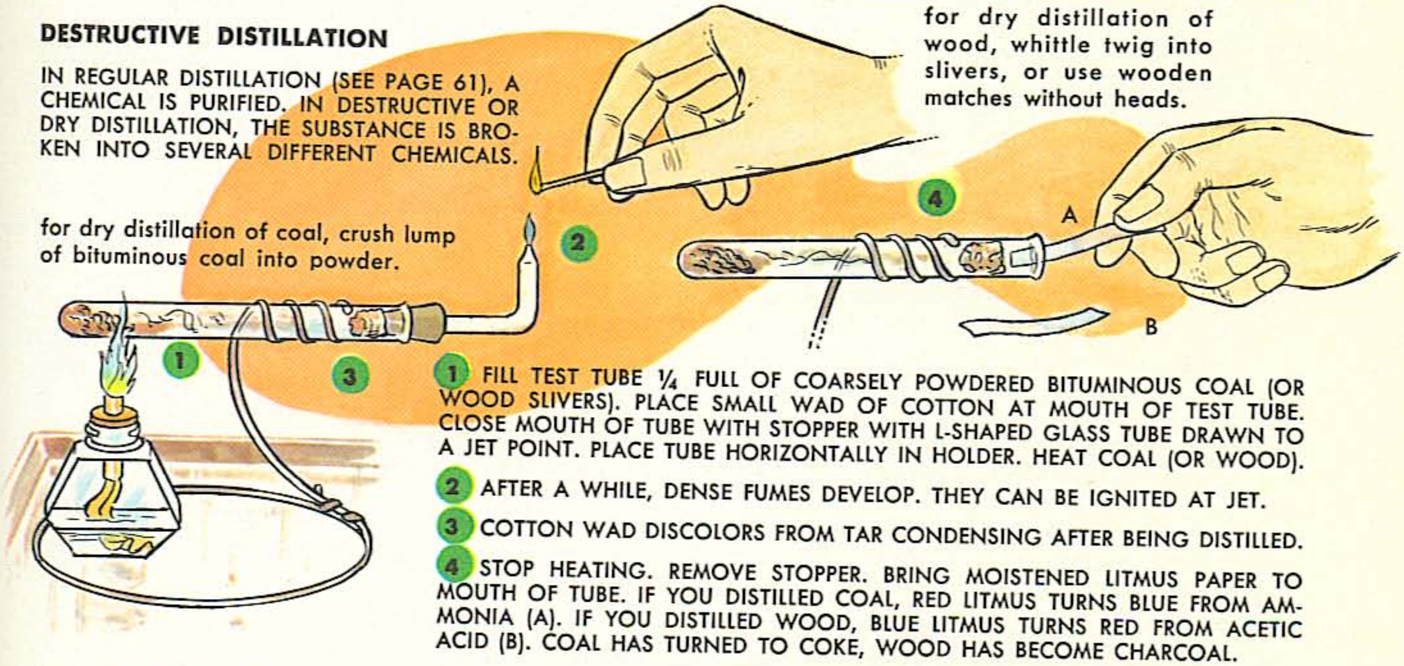

① FILL TEST TUBE ¼ FULL OF COARSELY POWDERED BITUMINOUS COAL (OR WOOD SLIVERS). PLACE SMALL WAD OF COTTON AT MOUTH OF TEST TUBE. CLOSE MOUTH OF TUBE WITH STOPPER WITH L-SHAPED GLASS TUBE DRAWN TO A JET POINT. PLACE TUBE HORIZONTALLY IN HOLDER. HEAT COAL (OR WOOD).

② AFTER A WHILE, DENSE FUMES DEVELOP. THEY CAN BE IGNITED AT JET.

③ COTTON WAD DISCOLORS FROM TAR CONDENSING AFTER BEING DISTILLED.

④ STOP HEATING. REMOVE STOPPER. BRING MOISTENED LITMUS PAPER TO MOUTH OF TUBE. IF YOU DISTILLED COAL, RED LITMUS TURNS BLUE FROM AMMONIA (A). IF YOU DISTILLED WOOD, BLUE LITMUS TURNS RED FROM ACETIC ACID (B). COAL HAS TURNED TO COKE, WOOD HAS BECOME CHARCOAL.

CARBON IN SUGAR

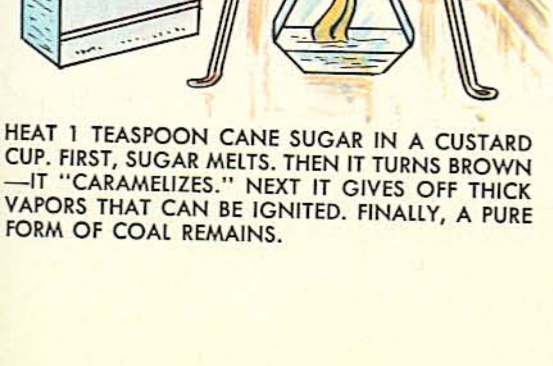

HEAT 1 TEASPOON CANE SUGAR IN A CUSTARD CUP. FIRST, SUGAR MELTS. THEN IT TURNS BROWN —IT "CARAMELIZES." NEXT IT GIVES OFF THICK VAPORS THAT CAN BE IGNITED. FINALLY, A PURE FORM OF COAL REMAINS.

YOU CAN PROVE PRESENCE OF CARBON IN THE FOOD YOU EAT BY HEATING SMALL SAMPLES OF CHEESE, BREAD, MEAT, SUGAR. BE SURE TO DO THIS OUTDOORS TO PREVENT EXPERIMENTS FROM SMELLING UP THE WHOLE HOUSE.

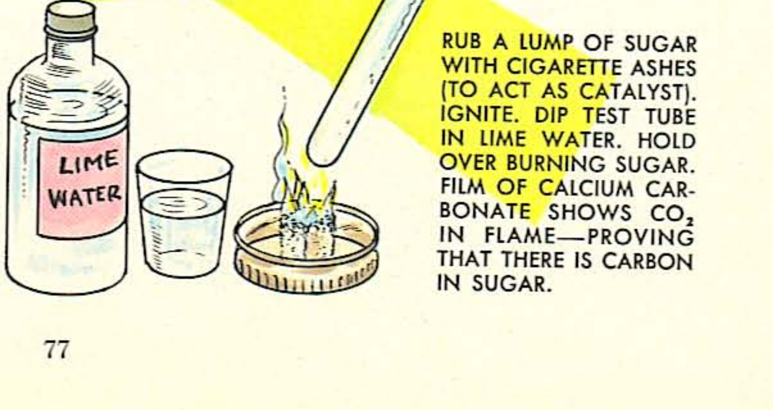

RUB A LUMP OF SUGAR WITH CIGARETTE ASHES (TO ACT AS CATALYST). IGNITE. DIP TEST TUBE IN LIME WATER. HOLD OVER BURNING SUGAR. FILM OF CALCIUM CARBONATE SHOWS CO_2 IN FLAME—PROVING THAT THERE IS CARBON IN SUGAR.

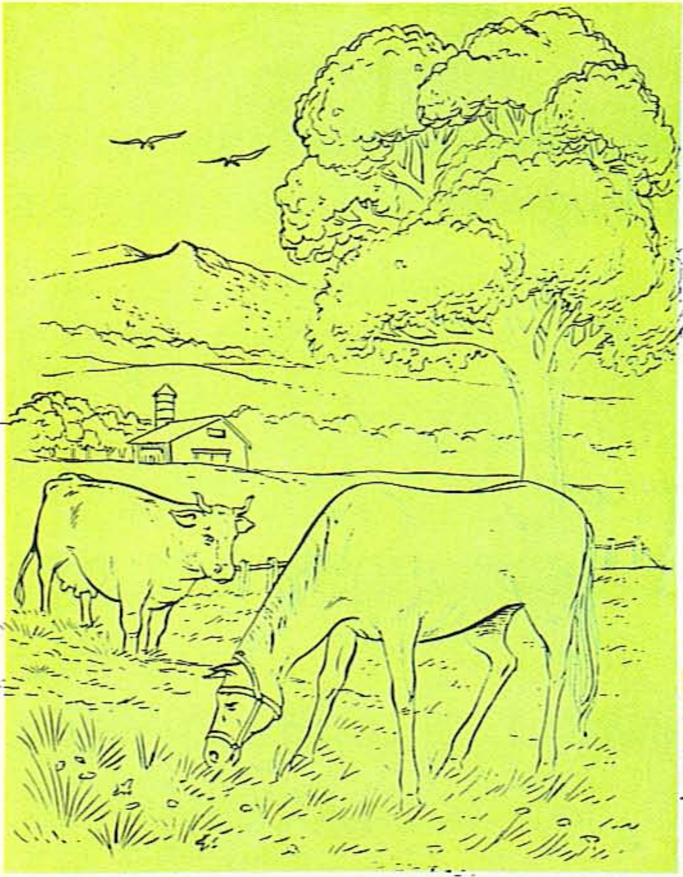

IT WAS ONCE BELIEVED THAT ORGANIC COMPOUNDS COULD BE PRODUCED ONLY BY LIVING ORGANISMS.

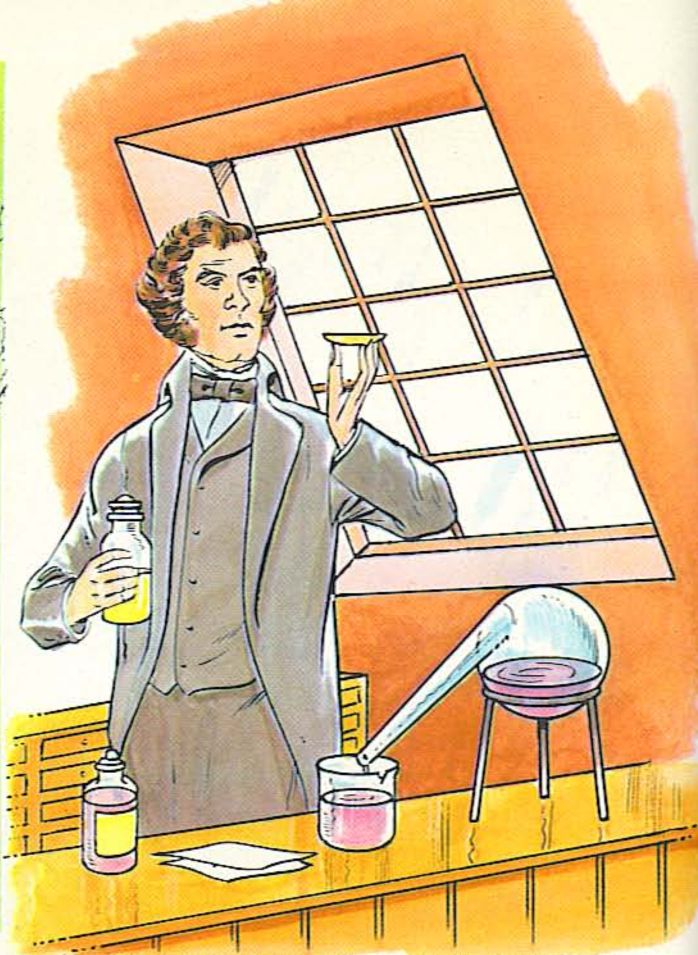

IN 1828, FRIEDRICH WÖHLER SUCCEEDED IN MAKING AN ORGANIC COMPOUND ARTIFICIALLY IN HIS LABORATORY.

The Chemistry of Carbon Compounds

THE CHEMISTS of about two hundred years ago divided all compounds very neatly into two groups — organic and inorganic. The organic compounds were those produced by living organisms — that is, plants and animals. The inorganic compounds were made up of dead things — rocks and minerals, water and various gases. No organic compound, these chemists insisted, could ever be produced artificially — they required the force we call "life" for their creation. And then, in 1828, a German chemist, Friedrich Wohler, completely upset this idea.

In his laboratory, Wöhler had mixed ammonium sulfate ($(NH_4)_2SO_4$) and potassium cyanate (KCNO), expecting to get ammonium cyanate. After evaporating, he analyzed the compound he had made. To his amazement he discovered that it was not ammonium cyanate at all, but urea — a compound produced in the kidneys of living animals, including man. The atoms of the ammonium cyanate molecule had rearranged themselves into a urea molecule.

NH_4 CNO had turned into $(NH_2)_2CO$.

A few years later, another organic compound — acetic acid — was made artificially. And then the lid really blew off. More and more products of plant and animal life were put together — synthesized — in the laboratory. And as if this were not enough, chemists began producing organic compounds that were not even found in nature.

It became clear that the old meaning of organic chemistry no longer was right. And so, the definition was changed. Today, organic chemistry is defined as "the chemistry of the carbon compounds." This definition is almost, but not 100 per cent, correct. The metallic carbonates, for instance, are still considered to be inorganic compounds, and carbon dioxide and carbonic acid are regarded as being both organic and inorganic.

You may think it odd that a whole branch of chemistry should deal with the compounds of a single element. But you will not be surprised at all when you start experimenting with a few of the close to 1,000,000 carbon compounds.

HYDROCARBONS CONTAIN TWO ELEMENTS ONLY: CARBON AND HYDROGEN. HYDROCARBONS WITH FEW ATOMS TO THEIR MOLECULES ARE GASES. OTHERS WITH MANY ATOMS ARE LIQUIDS AND SOLIDS.

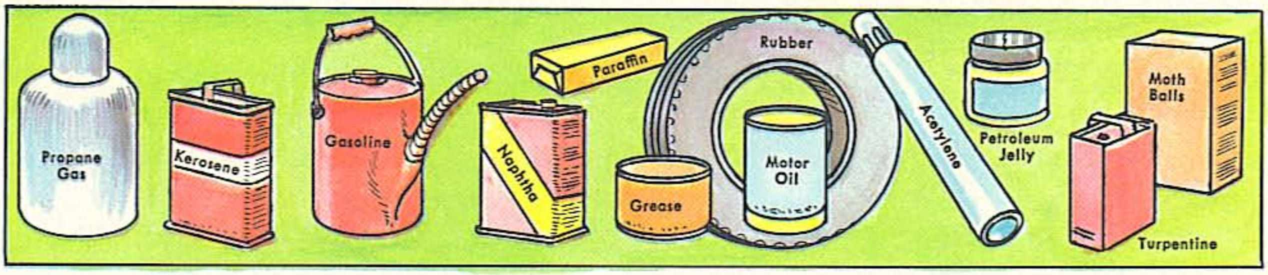
Propane Gas
Kerosene
Gasoline
Paraffin
Naphtha
Grease
Rubber
Motor Oil
Acetylene
Petroleum Jelly
Moth Balls
Turpentine

ALCOHOLS MAY BE CONSIDERED HYDROCARBONS IN WHICH A HYDROGEN ATOM IS REPLACED BY OH.

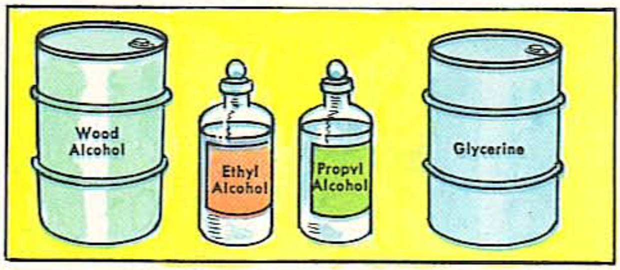
Wood Alcohol
Ethyl Alcohol
Propyl Alcohol
Glycerine

CARBOHYDRATES ARE IN MANY OF OUR MOST VALUABLE FOODSTUFFS AS STARCH AND SUGARS.

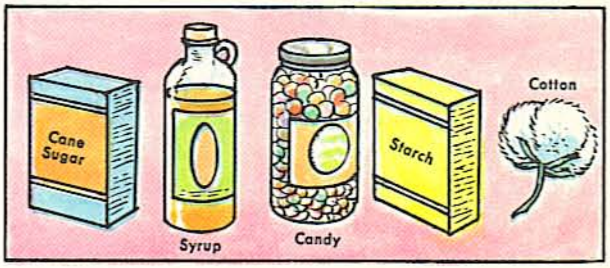
Cane Sugar
Syrup
Candy
Starch
Cotton

ESTERS IN ORGANIC CHEMISTRY CAN BE COMPARED TO SALTS IN INORGANIC CHEMISTRY. FATS AND OILS ARE THE MOST IMPORTANT ESTERS. THESE ARE THE "SALTS" OF GLYCERINE AND FATTY ACIDS.

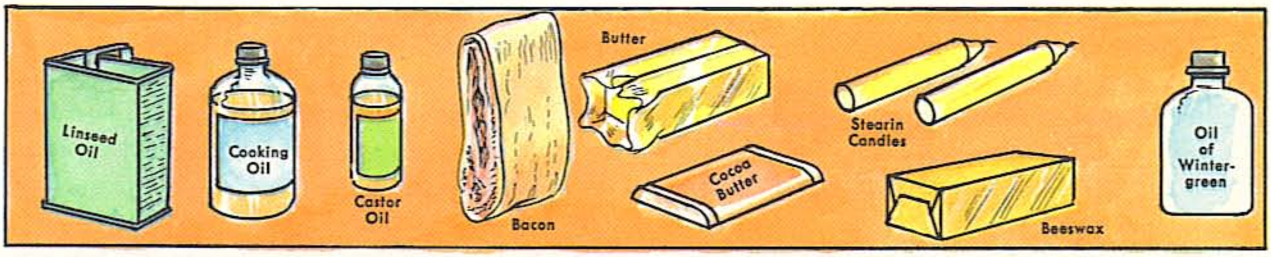
Linseed Oil
Cooking Oil
Castor Oil
Bacon
Butter
Cocoa Butter
Stearin Candles
Beeswax
Oil of Wintergreen

CARBOXYLIC ACIDS ARE NAMED FOR THE CARBOXYL GROUP—COOH—FOUND IN THEIR FORMULAS.

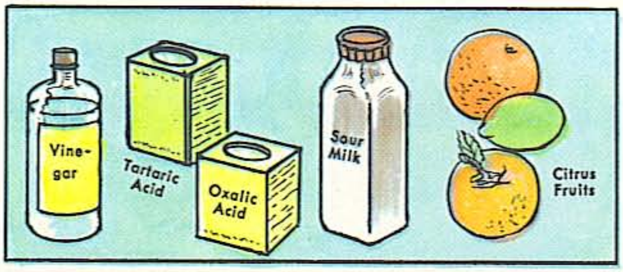
Vinegar
Tartaric Acid
Oxalic Acid
Sour Milk
Citrus Fruits

PROTEINS ARE COMPLEX COMPOUNDS THAT CONTAIN CARBON, HYDROGEN, OXYGEN, NITROGEN.

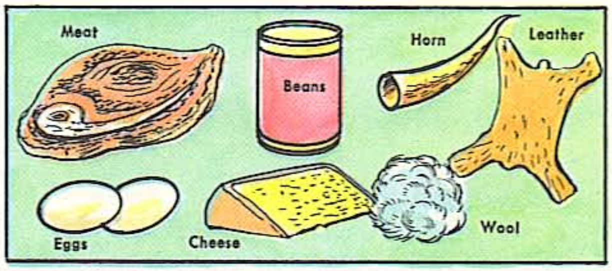
Meat
Beans
Horn
Leather
Eggs
Cheese
Wool

OTHER CARBON COMPOUNDS—IN ADDITION TO THE MAIN GROUPS ILLUSTRATED ABOVE, THERE ARE NUMEROUS OTHER KINDS OF CARBON COMPOUNDS. MANY HAVE VERY COMPLICATED FORMULAS.

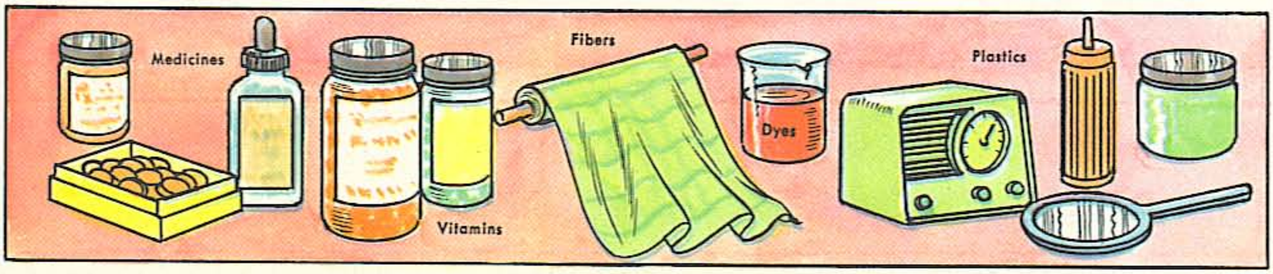
Medicines
Vitamins
Fibers
Dyes
Plastics

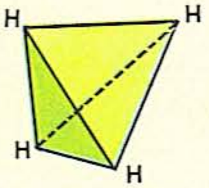

methane, CH_4

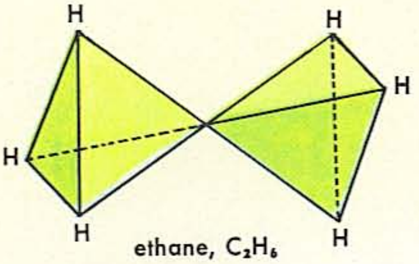

ethane, C_2H_6

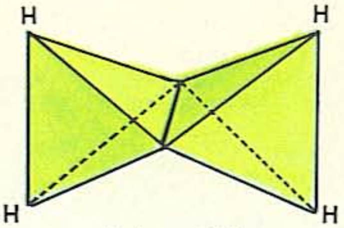

ethylene, C_2H_4

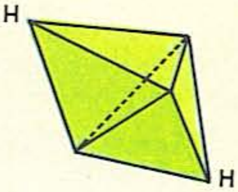

acetylene, C_2H_2

The Formulas of Carbon Compounds

How is it possible for carbon to make so many different compounds of such tremendous variety? That was one of the great questions facing chemists during the last century.

It was easy enough to explain carbon dioxide. Carbon has a valence of 4, oxygen of 2 — the formula had to be CO_2. It was also easy to explain the molecule of the simple hydrocarbon methane (CH_4). But how explain compounds consisting of two atoms of carbon and six of hydrogen (C_2H_6, ethane), or two atoms of carbon and four of hydrogen (C_2H_4, ethylene), or two of carbon and only two of hydrogen (C_2H_2, acetylene)?

A German chemist and professor, Friedrich August Kekulé, came up with the solution. The answer was quite simple:

While the atoms of most elements "hook on" to the atoms of other elements according to their valences, the atoms of carbon "hook on" to each other as well. To understand this, write out carbon atoms with four lines to indicate the valence bonds, but arrange the lines in these three different ways:

$$\equiv C- \quad -C\equiv \quad =C=$$

Then hook them together, two by two, in these three different ways:

$$\equiv C-C\equiv \quad =C=C= \quad -C\equiv C-$$

Now add a hydrogen atom to each of the free bonds — and there you have the formulas for the three hydro-carbons — ethane (C_2H_6), ethylene (C_2H_4), and acetylene (C_2H_2):

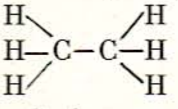

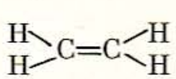

$$H-C\equiv C-H$$

So far so good. But there were still many carbon compound formulas that would not line up in this kind of arrangement. C_6H_6, for instance — benzene, an important hydrocarbon obtained by distillation of coal.

Again, it was Kekulé who offered the explanation. This time it came to him in a dream. He had been

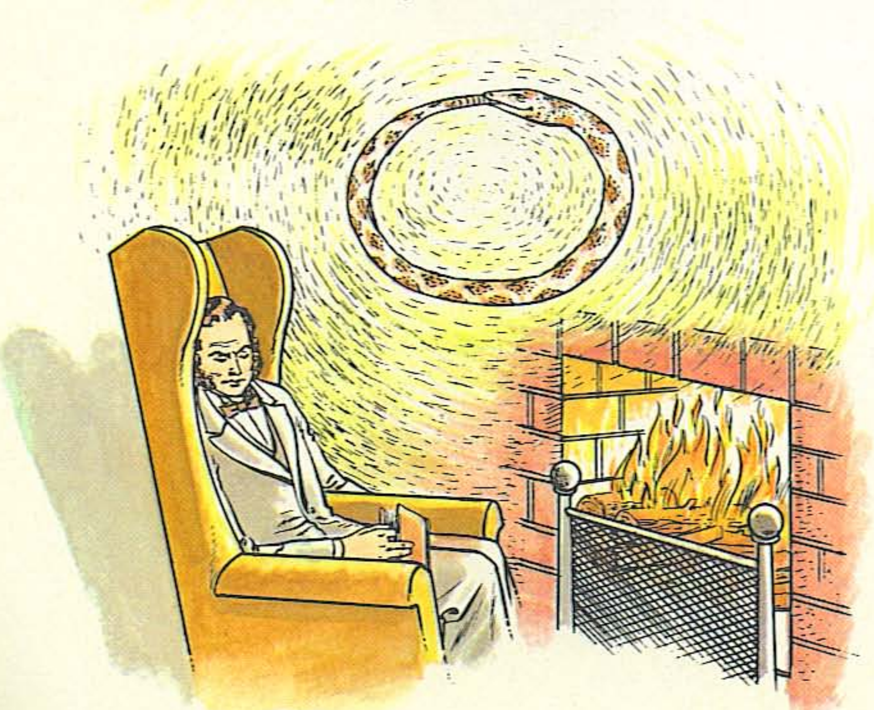

KEKULÉ'S BENZENE RING

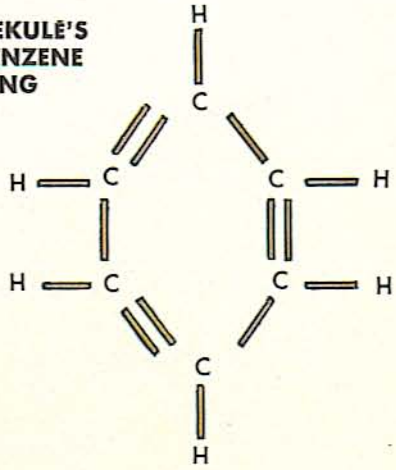

AUGUST KEKULE HIT UPON THE STRUCTURE OF THE BENZENE MOLECULE IN A DREAM. A SNAKE SEEMED TO WHIRL IN A RING BEFORE HIS EYES. BY ARRANGING THE SIX CARBON ATOMS IN A RING, THE PROBLEM WAS SOLVED.

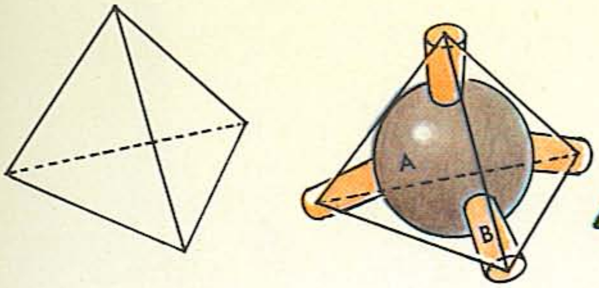

IN A CARBON ATOM (A), THE FOUR VALENCE BONDS POINT AWAY FROM THE CENTER (NUCLEUS) TOWARD THE CORNERS OF A TETRAHEDRON—A FIGURE MADE UP OF FOUR TRIANGLES. IN MODELS OF CARBON ATOMS, VALENCE BONDS ARE SHOWN BY RODS (B).

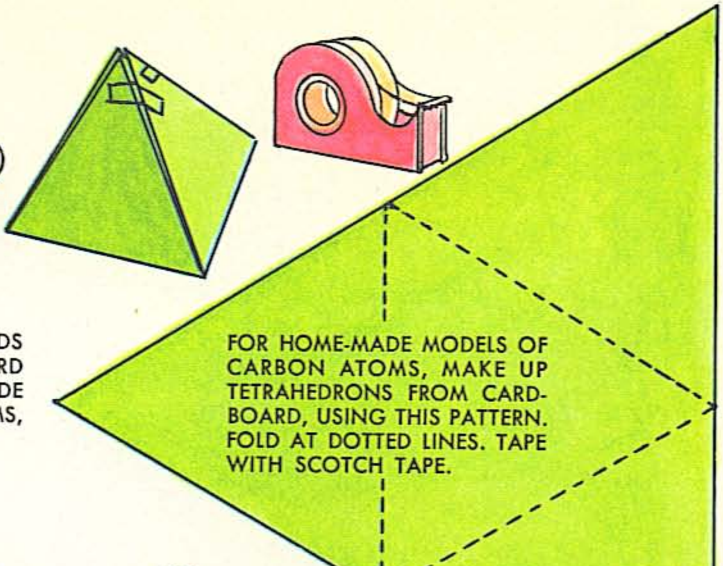

FOR HOME-MADE MODELS OF CARBON ATOMS, MAKE UP TETRAHEDRONS FROM CARDBOARD, USING THIS PATTERN. FOLD AT DOTTED LINES. TAPE WITH SCOTCH TAPE.

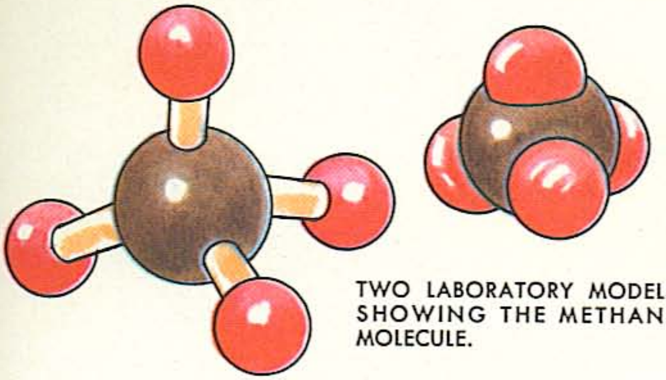

TWO LABORATORY MODELS SHOWING THE METHANE MOLECULE.

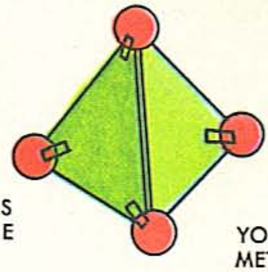

small circles represent hydrogen atoms

YOUR OWN MODEL OF METHANE MOLECULE.

working all day with long lines of organic formulas. In the evening he dozed before the fire. In his dream, the lines of formulas turned into snakes, twisting and twining — until suddenly one of the snakes grasped its own tail and whirled around in a ring. This dream gave Kekule the clue: the carbon atoms in benzene hang together in a ring, each atom using three of its bonds to hang on to the atoms next to it, with one bond free to hook onto a hydrogen atom.

Starting from these very simple formulas, modern scientists can figure out the most complicated chemical formulas.

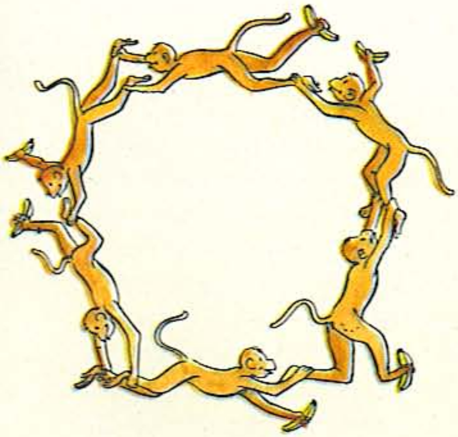

YOU CAN THINK OF THE BENZENE RING AS SIX MONKEYS HANGING ON TO EACH OTHER WITH ONE OR TWO HANDS, HOLDING BANANAS IN THEIR FREE HANDS.

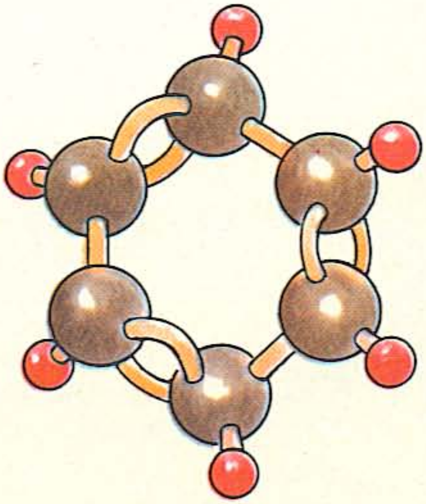

THIS IS THE WAY THE BENZENE MOLECULE LOOKS WHEN IT IS CONSTRUCTED FROM PARTS USED TO MAKE UP LABORATORY MODELS FOR DEMONSTRATION.

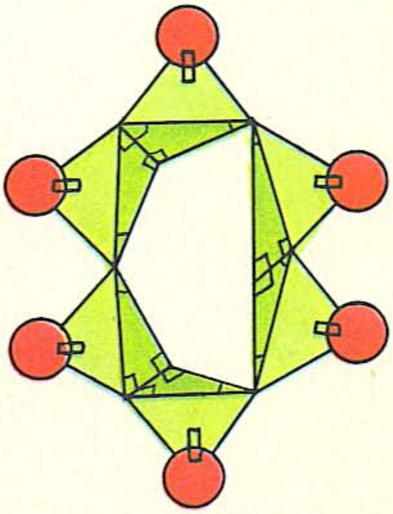

THIS IS HOW THE BENZENE MOLECULE WILL LOOK WHEN YOU P
TOGETHER FROM HOME-MADE
BON ATOMS. YOU CAN DO
WITH SCOTCH TAPE.

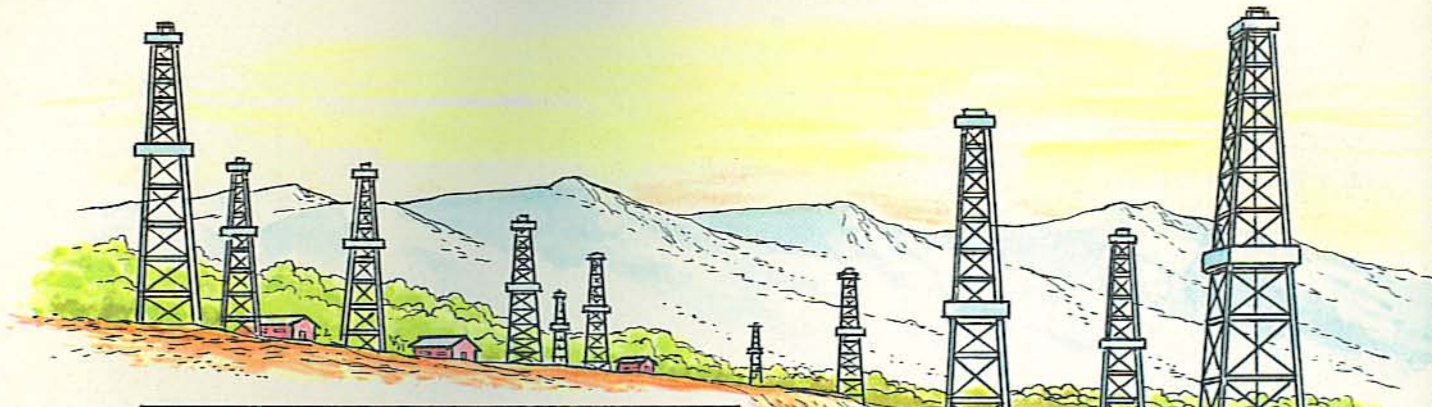

A HUNDRED YEARS AFTER THE FIRST OIL WELL IN PENNSYLVANIA, OIL FIELDS ARE FOUND IN SEVERAL STATES.

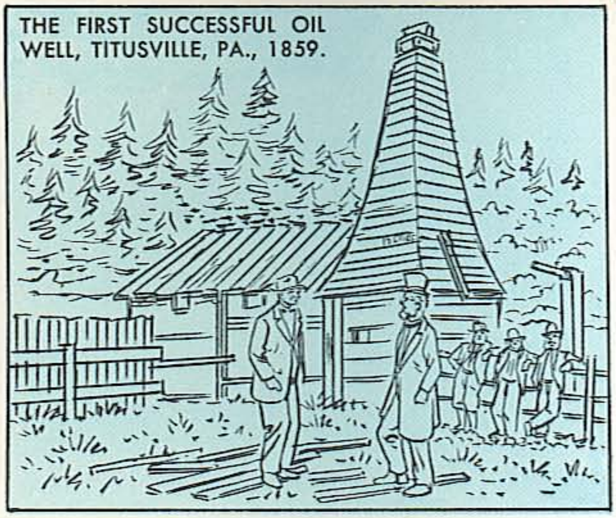

THE FIRST SUCCESSFUL OIL WELL, TITUSVILLE, PA., 1859.

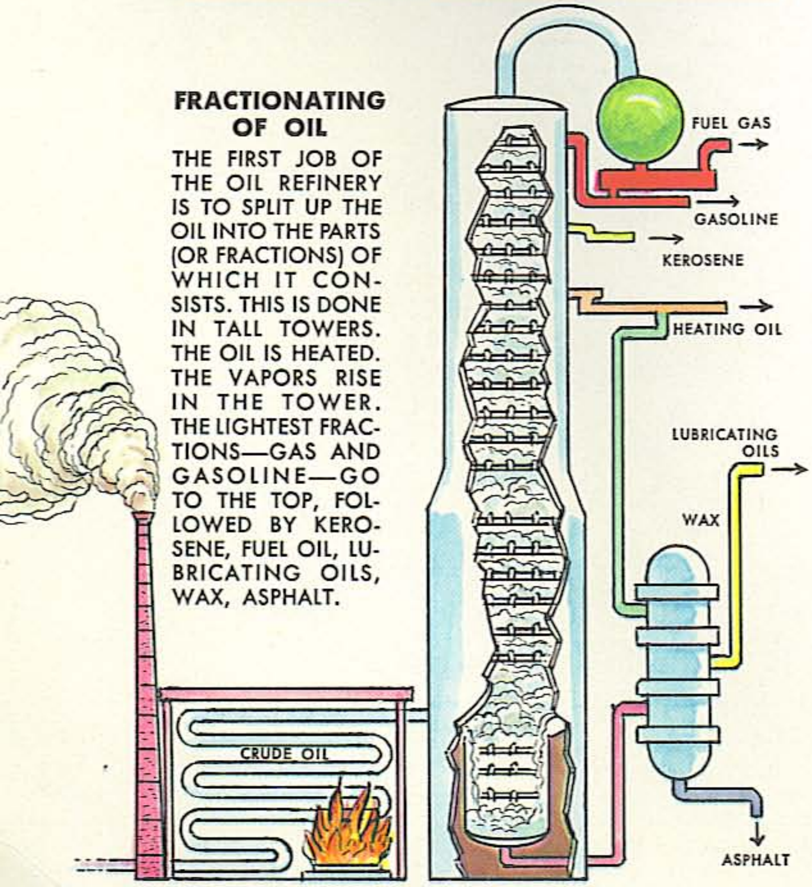

FRACTIONATING OF OIL

THE FIRST JOB OF THE OIL REFINERY IS TO SPLIT UP THE OIL INTO THE PARTS (OR FRACTIONS) OF WHICH IT CONSISTS. THIS IS DONE IN TALL TOWERS. THE OIL IS HEATED. THE VAPORS RISE IN THE TOWER. THE LIGHTEST FRACTIONS—GAS AND GASOLINE—GO TO THE TOP, FOLLOWED BY KEROSENE, FUEL OIL, LUBRICATING OILS, WAX, ASPHALT.

A Lot of Hydrocarbons

THE FAMILY CAR stops at the service station.

"Fill 'er up!" The attendant pours what may be a hexane-septane-octane-nonane mixture into the gas tank. "Check your oil, mister?" Into the engine goes another hydrocarbon mixture — possibly along the line of $C_{20}H_{42}$-$C_{21}H_{44}$-$C_{22}H_{46}$. And so you take off—on hydrocarbon tires. Yes, gasoline, motor oil, and rubber are all hydrocarbons — compounds that contain only hydrogen and carbon.

A great number of hydrocarbons come from petroleum (crude oil). Coal and natural gas provide several others. Many more are produced by nature — natural rubber, turpentine, camphor, to mention a few. Even the red coloring of tomatoes and the yellow of carrots are hydrocarbons.

One of the remarkable things about hydrocarbons is that it is possible to combine some of those with small molecules into others with larger ones (as in making synthetic rubber), as well as to "crack" those with large molecules into others with smaller ones (as when a heavy oil is "cracked" into gasoline). But that is only the beginning. By replacing one or more hydrogen atoms with hydroxyl groups (OH) or carboxyl groups (COOH) or chlorine atoms (Cl), for instance, it is possible to build up more complicated compounds — which can then be built up further and further. And that is exactly what chemists are doing today — giving us medicines and dyes, plastics and explosives, and countless other things.

METHANE—CH_4

1 HAMMER LUMPS OF BITUMINOUS COAL INTO A COARSE POWDER. FILL FUNNEL WITH IT. BRING FUNNEL INTO LARGE JAR.

2 TURN JAR UPSIDE DOWN. FILL JAR WITH WATER. PLACE A WATER-FILLED TEST TUBE OVER FUNNEL. IN A FEW DAYS, TUBE IS FILLED WITH METHANE.

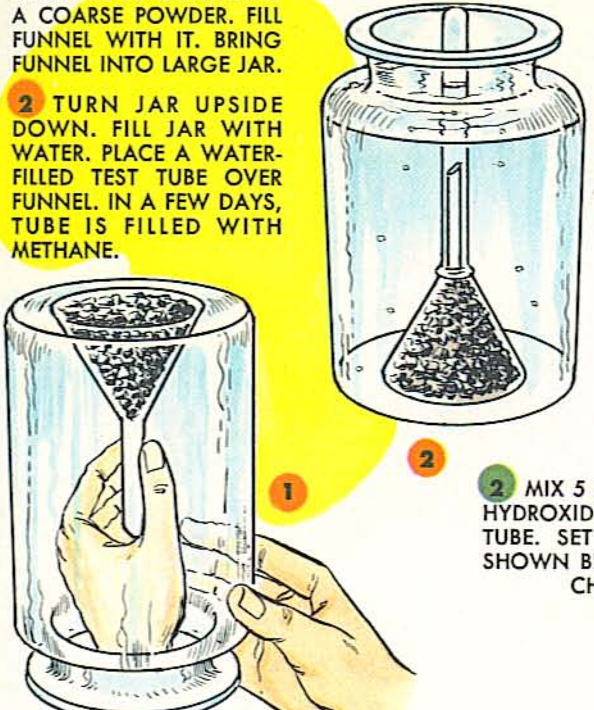

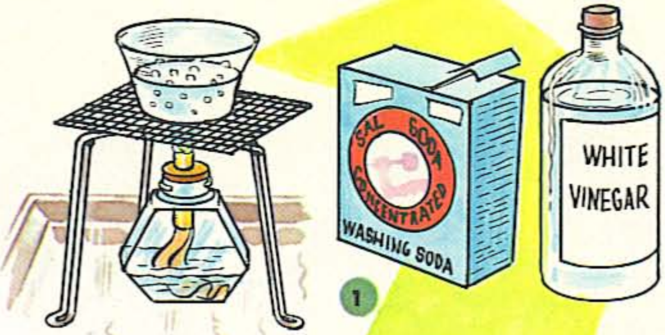

IN THE LABORATORY, METHANE IS MADE BY HEATING WATER-FREE SODIUM ACETATE WITH "SODA LIME."

1 TO MAKE SODIUM ACETATE, ADD WASHING SODA TO ½ CUSTARD CUP VINEGAR UNTIL NO MORE CO_2 IS GIVEN OFF. EVAPORATE MIXTURE AT LOW HEAT TO GET WHITE POWDER OF SODIUM ACETATE.

2 MIX 5 g SODIUM ACETATE (CH_3COONa), 5 g SODIUM HYDROXIDE, AND 5 g CALCIUM OXIDE. DROP INTO TEST TUBE. SET UP APPARATUS FOR COLLECTING GAS AS SHOWN BELOW. HEAT TO MAKE METHANE:

$$CH_3COONa + NaOH \rightarrow CH_4 + Na_2CO_3$$

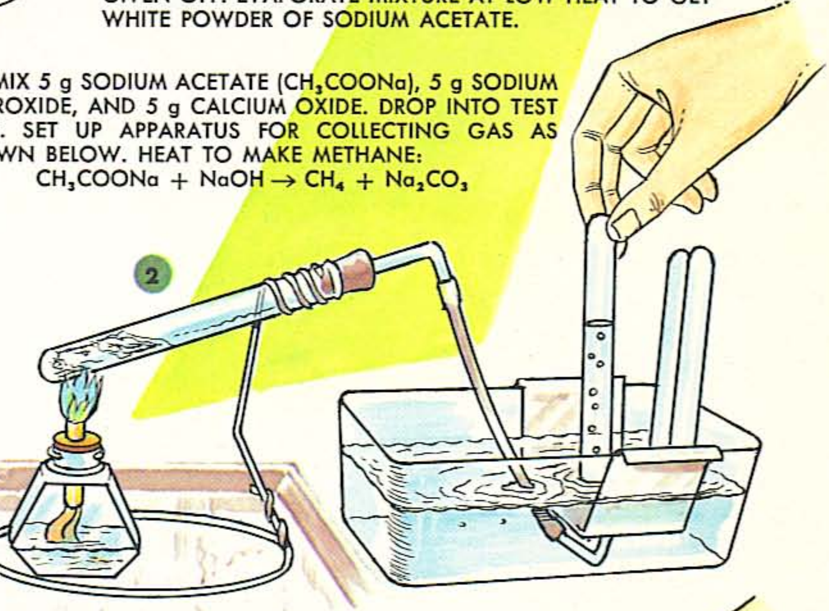

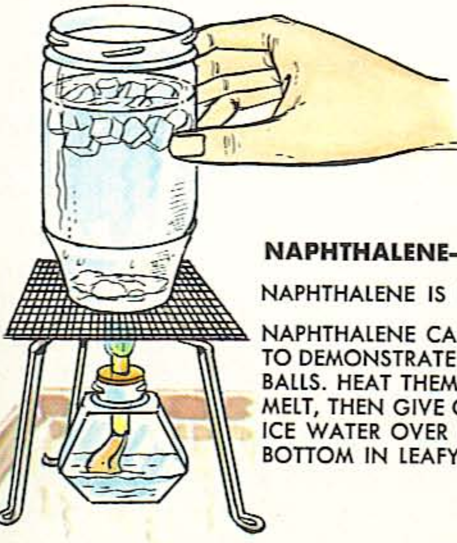

NAPHTHALENE—$C_{10}H_8$

NAPHTHALENE IS USED IN MAKING MOTH BALLS.

NAPHTHALENE CAN BE PURIFIED BY SUBLIMATION. TO DEMONSTRATE THIS, CRUSH A COUPLE OF MOTH BALLS. HEAT THEM IN A CUSTARD CUP. FIRST THEY MELT, THEN GIVE OFF VAPOR. PUT JAR FILLED WITH ICE WATER OVER CUP. NAPHTHALENE SETTLES ON BOTTOM IN LEAFY CRYSTALS.

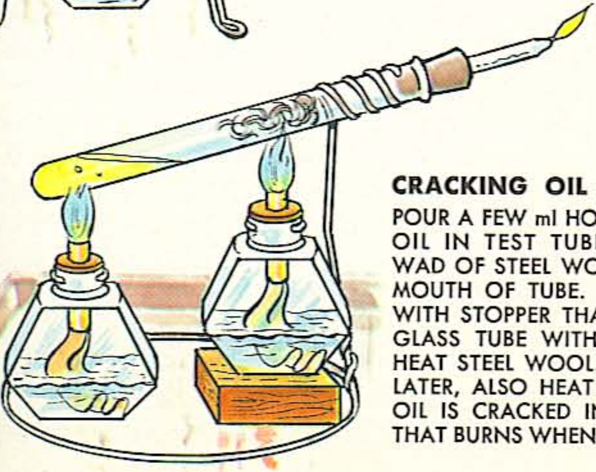

CRACKING OIL

POUR A FEW ml HOUSEHOLD OIL IN TEST TUBE. PLACE WAD OF STEEL WOOL NEAR MOUTH OF TUBE. CLOSE IT WITH STOPPER THAT HAS A GLASS TUBE WITH JET TIP. HEAT STEEL WOOL. A LITTLE LATER, ALSO HEAT THE OIL. OIL IS CRACKED INTO GAS THAT BURNS WHEN IGNITED.

TURPENTINE—$C_{10}H_{16}$

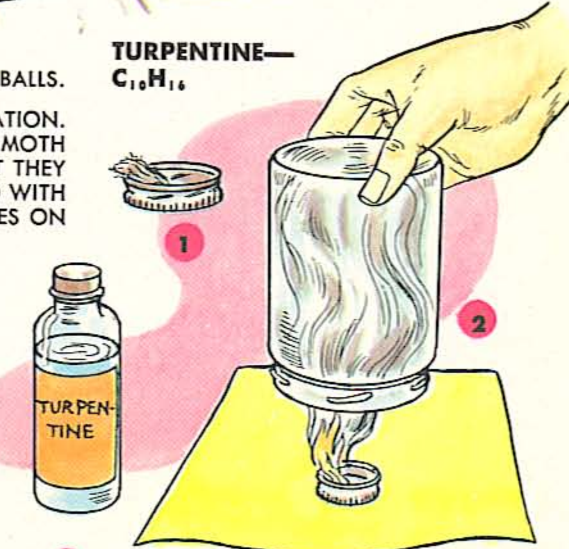

1 POUR A LITTLE TURPENTINE INTO A BOTTLE CAP. PLACE A SMALL WICK IN IT.

2 PLACE BOTTLE CAP ON PIECE OF PAPER. IGNITE TURPENTINE. IT BURNS INCOMPLETELY, GIVING OFF A BLACK SMOKE OF CARBON WHICH YOU CAN COLLECT IN A JAR.

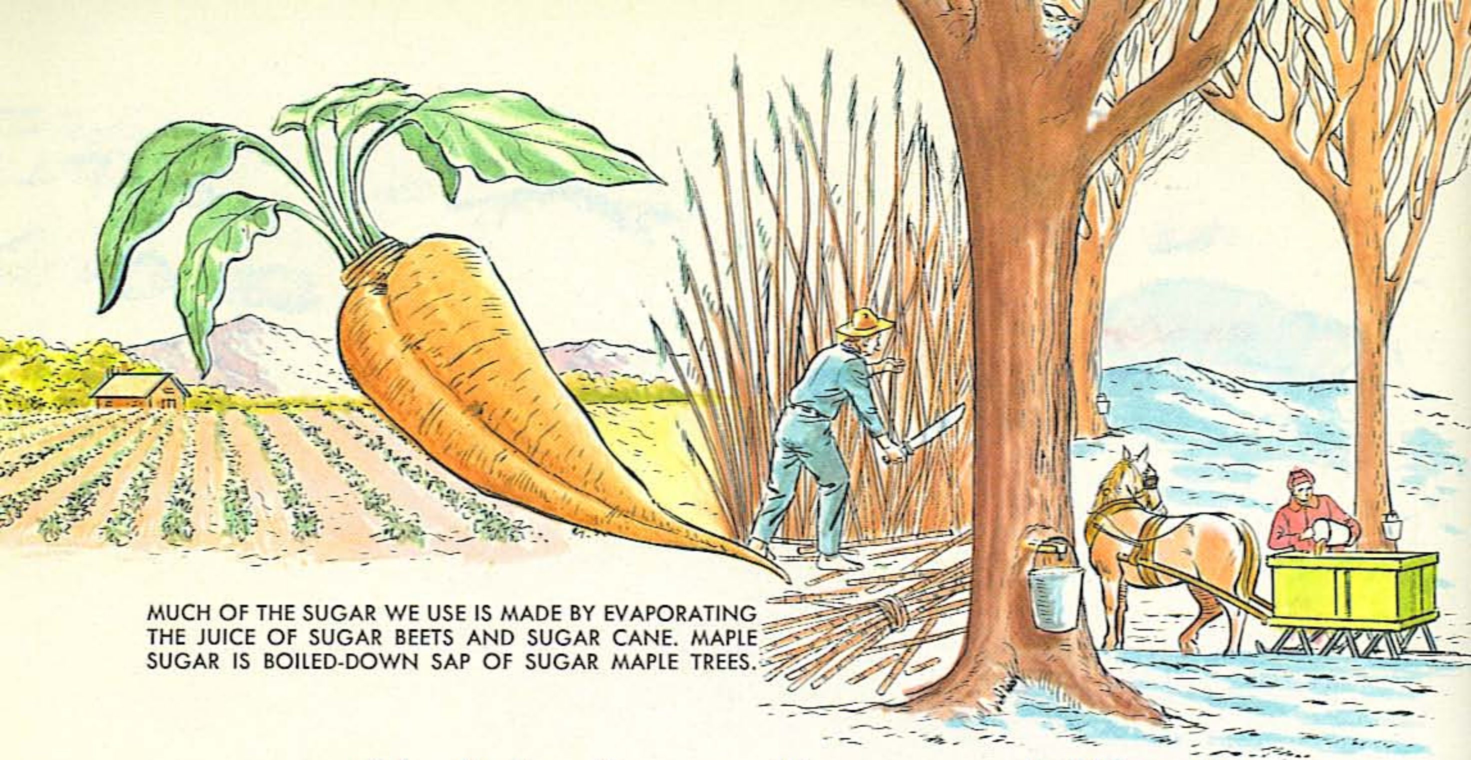

MUCH OF THE SUGAR WE USE IS MADE BY EVAPORATING THE JUICE OF SUGAR BEETS AND SUGAR CANE. MAPLE SUGAR IS BOILED-DOWN SAP OF SUGAR MAPLE TREES.

Carbohydrates—Sweet and Bland

USUALLY, when we talk about "hydrates" we mean chemicals that contain water. But when we talk about carbohydrates we mean organic compounds of carbon, hydrogen, and oxygen in which the proportion between hydrogen and oxygen is the same as in water (H_2O) — that is, twice as much hydrogen as oxygen. And so we find carbohydrates that have 22 atoms of hydrogen and 11 atoms of oxygen to 12 atoms of carbon ($C_{12}H_{22}O_{11}$), or 12 hydrogen and 6 oxygen to 6 carbon ($C_6H_{12}O_6$), or 10 and 5 to 6 carbon atoms ($C_6H_{10}O_5$).

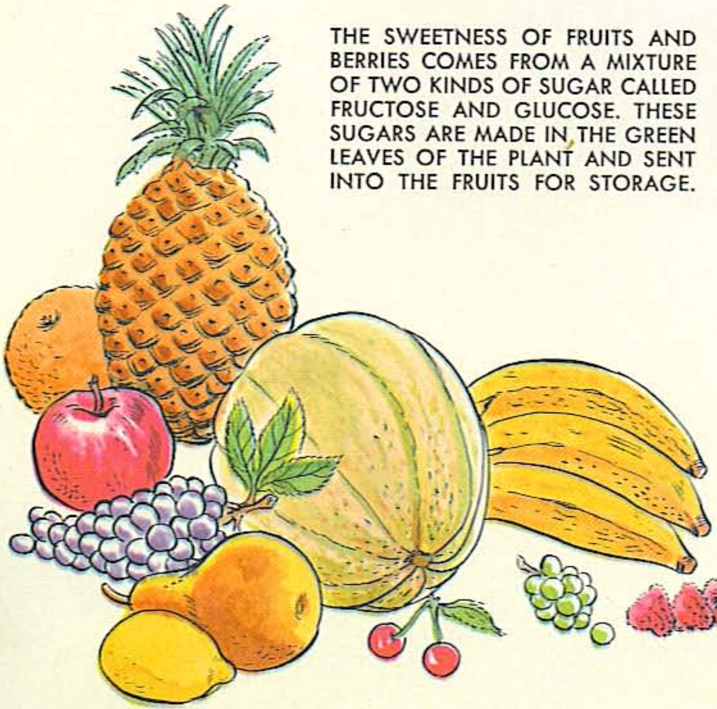

THE SWEETNESS OF FRUITS AND BERRIES COMES FROM A MIXTURE OF TWO KINDS OF SUGAR CALLED FRUCTOSE AND GLUCOSE. THESE SUGARS ARE MADE IN THE GREEN LEAVES OF THE PLANT AND SENT INTO THE FRUITS FOR STORAGE.

Carbohydrates are produced by plants by a remarkable process called photosynthesis — "putting things together with the help of light." When green leaves are exposed to sunlight, the chlorophyll in them combines the hydrogen from water with carbon dioxide from the air, while setting oxygen free — along this line:

$$6H_2O + 6CO_2 + \text{sunlight} \rightarrow C_6H_{12}O_6 + 6O_2 \uparrow$$

Carbohydrates are of tremendous importance to all of us. They make up a large part of our food supply in the form of sugars and starches. Another carbohydrate called cellulose helps to clothe us (cotton, linen) and shelter us (wood).

SUGARS — Most of our sugar comes from sugar beets or sugar cane. The juice is pressed out, cleared, filtered, and evaporated. The result is pure, white crystals of a sugar with the chemical name sucrose ($C_{12}H_{22}O_{11}$).

Another sugar called glucose ($C_6H_{12}O_6$) is found in ripe fruits, often in the company of still another sugar of the same formula called fructose ($C_6H_{12}O_6$). These two sugars can be made in the laboratory by treating the more complicated sucrose with an acid. The sucrose picks up water and splits into glucose and fructose by a process known as inversion:

$$\underset{\text{(sucrose)}}{C_{12}H_{22}O_{11}} + H_2O \rightarrow \underset{\text{(glucose)}}{C_6H_{12}O_6} + \underset{\text{(fructose)}}{C_6H_{12}O_6}$$

(CONTINUED ON PAGE 86)

TEST FOR GLUCOSE SUGAR

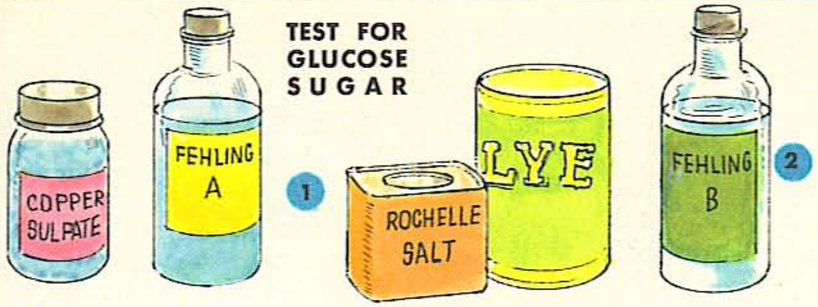

A GERMAN SCIENTIST, HERMAN FEHLING, THOUGHT UP THE TEST FOR GLUCOSE THAT HAS HIS NAME. FOR THIS TEST, TWO SOLUTIONS ARE REQUIRED. THEY ARE MIXED JUST BEFORE USE.

① FEHLING A. DISSOLVE 5 g COPPER SULFATE IN 70 ml WATER.

② FEHLING B. DISSOLVE 7 g SODIUM HYDROXIDE IN 70 ml WATER. IN THIS SOLUTION, DISSOLVE 25 g ROCHELLE SALT (SODIUM-POTASSIUM TARTRATE) FROM YOUR LOCAL DRUG STORE.

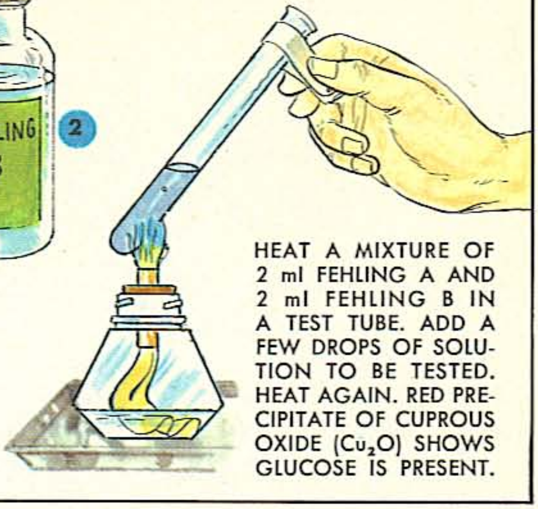

HEAT A MIXTURE OF 2 ml FEHLING A AND 2 ml FEHLING B IN A TEST TUBE. ADD A FEW DROPS OF SOLUTION TO BE TESTED. HEAT AGAIN. RED PRECIPITATE OF CUPROUS OXIDE (Cu_2O) SHOWS GLUCOSE IS PRESENT.

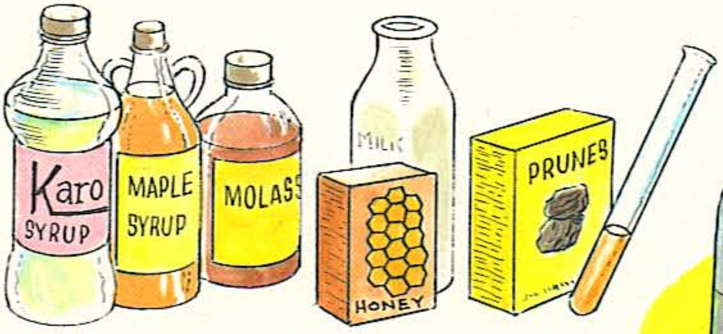

USE FEHLING TEST TO FIND OUT IF DIFFERENT SWEET-TASTING FOODS CONTAIN GLUCOSE SUGAR: CORN SYRUP, MAPLE SYRUP, MOLASSES, HONEY. ALSO TRY JUICES OF VARIOUS FRUITS: PRUNES, ORANGES, LEMONS, BERRIES. SEVERAL CONTAIN GLUCOSE AND GIVE RED PRECIPITATE. SUGAR IN MILK (LACTOSE) GIVES Cu_2O PRECIPITATE.

TEST CANE SUGAR WITH FEHLING. YOU DO NOT GET RED PRECIPITATE. CANE SUGAR IS NOT GLUCOSE BUT ANOTHER SUGAR CALLED SUCROSE.

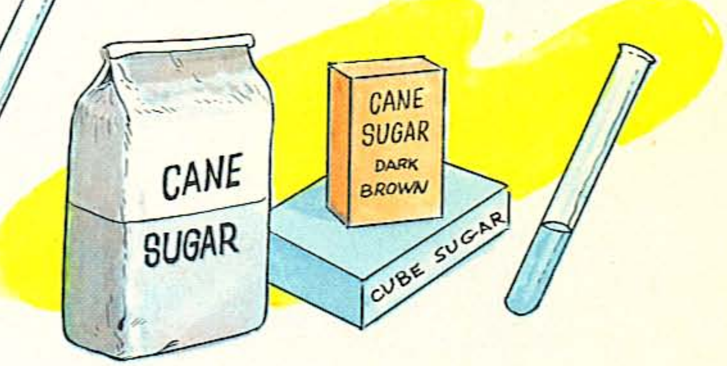

SUCROSE TO GLUCOSE

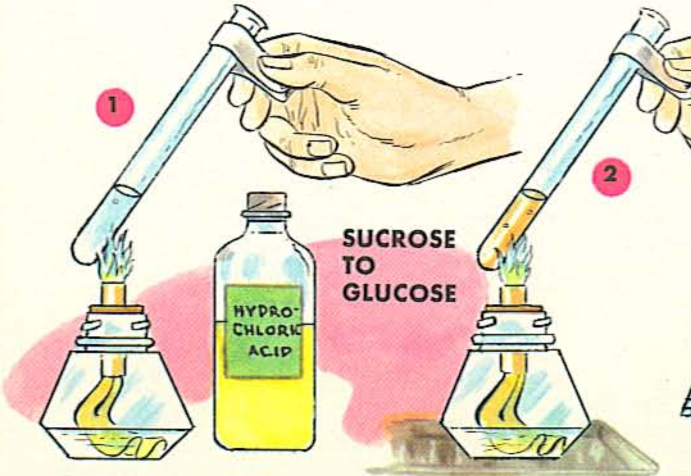

① DISSOLVE 1 g CANE SUGAR IN 10 ml WATER IN A TEST TUBE. ADD 10 DROPS HYDROCHLORIC ACID. HEAT GENTLY FOR A FEW MINUTES WITHOUT BOILING.

② HEAT FEHLING SOLUTION IN ANOTHER TEST TUBE. ADD A FEW ml SUGAR SOLUTION. HEAT AGAIN. YOU GET RED PRECIPITATE. GLUCOSE HAS BEEN FORMED.

MAKING SUGAR CANDY

IT IS EASY TO MAKE LOLLIPOPS. OVER LOW HEAT AND WHILE STIRRING, DISSOLVE ½ CUP SUGAR IN 2 TABLESPOONS WATER AND 2 TABLESPOONS LIGHT SYRUP. THEN CONTINUE HEATING WITHOUT STIRRING UNTIL A SAMPLE DROPPED INTO COLD WATER FORMS BRITTLE THREAD. SPOON OUT TABLESPOONFULS ONTO A SHEET OF GREASED ALUMINUM FOIL. PUSH STICK IN EACH BLOB. REMOVE WHEN COLD.

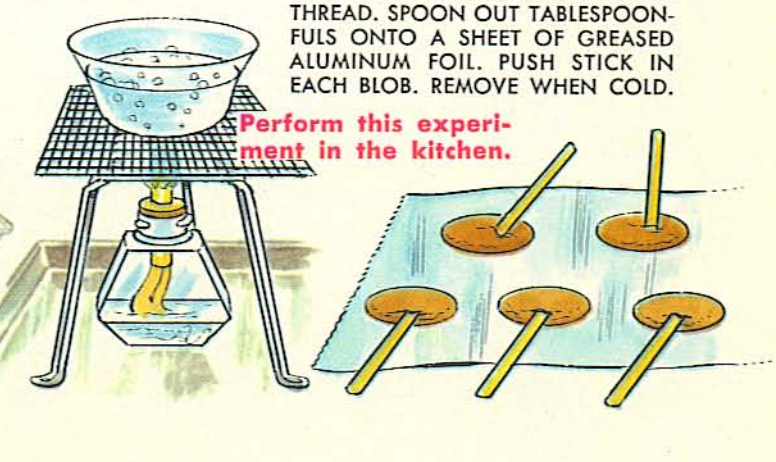

Perform this experiment in the kitchen.

Carbohydrates—Continued

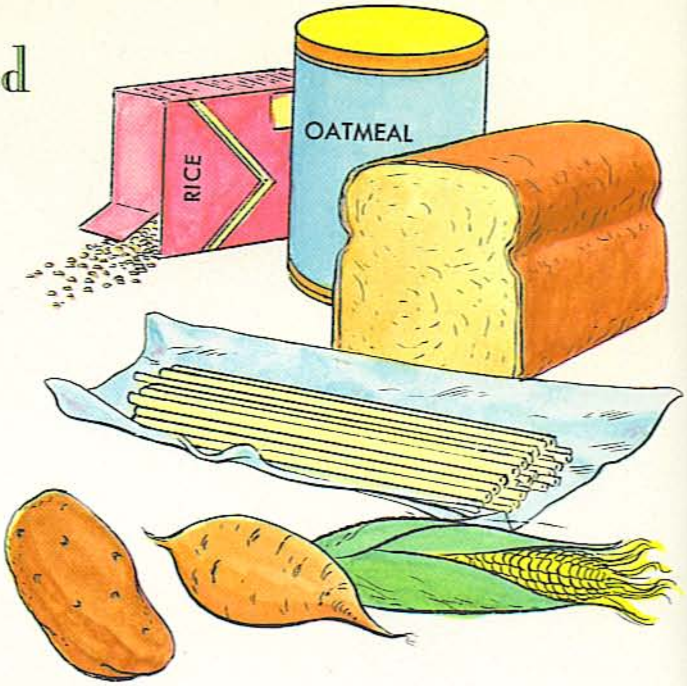

STARCHES — Starch is distributed in most plant parts. It is a carbohydrate with very large molecules. Take a look at its formula: $(C_6H_{10}O_5)_x$. At first glance it looks quite simple. But note that little x — it stands for "any number of times." A single molecule of starch may weigh 6,000 times as much as a single molecule of glucose.

You can break this polysaccharide ("many-sugar") into the monosaccharide ("single-sugar") glucose by treating it with an acid.

CELLULOSE is the building material of the plant world. It makes up the cell walls of leaves and stalks, wood and fibers. Cotton is 95 per cent cellulose. The paper on which this book is printed is specially treated cellulose. So is the cellophane around your candy and the rayon that goes into ladies' dresses. For more about cellulose in natural fibers and rayon, see pages 102-103.

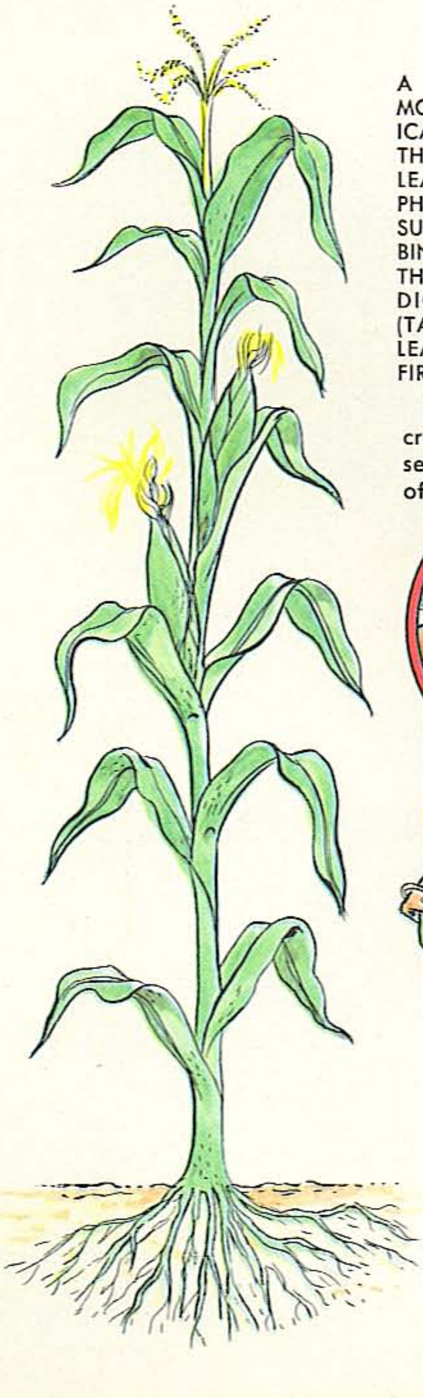

A GROWING PLANT IS THE MOST ASTONISHING CHEMICAL FACTORY ON EARTH. THE GREEN SUBSTANCE IN LEAVES—CALLED CHLOROPHYLL—WITH THE HELP OF SUNLIGHT IS ABLE TO COMBINE WATER (TAKEN IN BY THE ROOTS) WITH CARBON DIOXIDE FROM THE AIR (TAKEN IN THROUGH THE LEAVES) TO FORM SUGAR FIRST AND THEN STARCH.

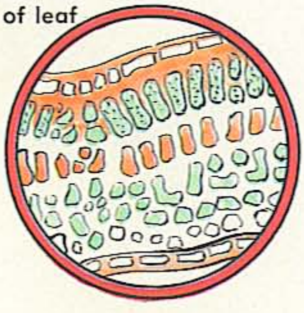

EXPERIMENTS WITH PHOTOSYNTHESIS

1 POT UP A NASTURTIUM OR GERANIUM PLANT AND PLACE IT IN THE DARK FOR A COUPLE OF DAYS. THEN FASTEN STRIPS OF BLACK PAPER ACROSS BOTH SIDES OF ONE OR MORE LEAVES. NOW EXPOSE THE GROWING PLANT TO THE SUNLIGHT FOR TWO HOURS.

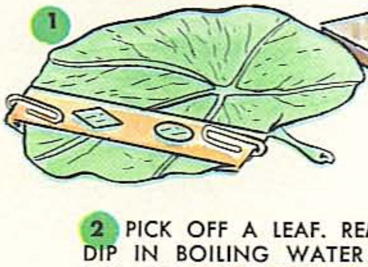

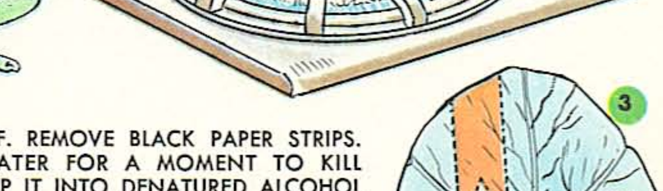

2 PICK OFF A LEAF. REMOVE BLACK PAPER STRIPS. DIP IN BOILING WATER FOR A MOMENT TO KILL THE LEAF. THEN DROP IT INTO DENATURED ALCOHOL IN A CUSTARD CUP. PLACE CUSTARD CUP IN A POT OF BOILING WATER. AS ALCOHOL GETS HOT, IT EXTRACTS THE CHLOROPHYLL FROM THE LEAF. KEEP LEAF IN ALCOHOL UNTIL ALL CHLOROPHYLL IS OUT.

3 PLACE LEAF IN IODINE TEST SOLUTION. PARTS EXPOSED TO SUN TURN BLUE. THIS PROVES PRESENCE OF STARCH. UNEXPOSED PARTS BECOME BROWN.

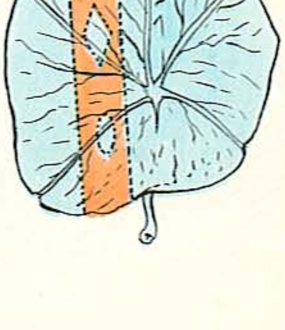

PREPARING STARCH

IT IS A SIMPLE MATTER TO GET STARCH FROM POTATOES.

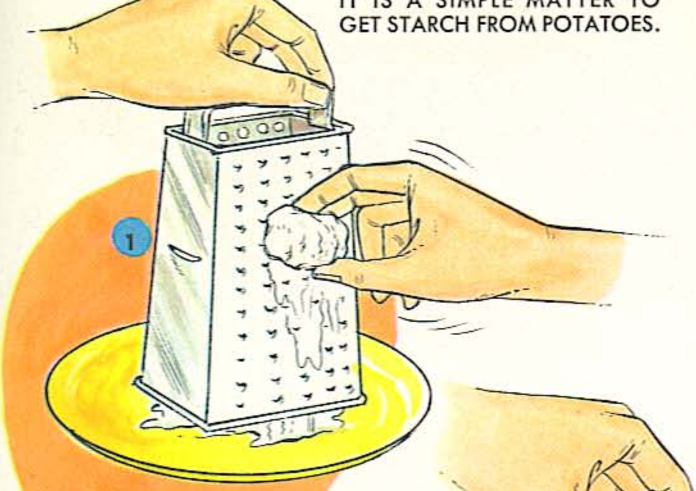

TEST FOR STARCH

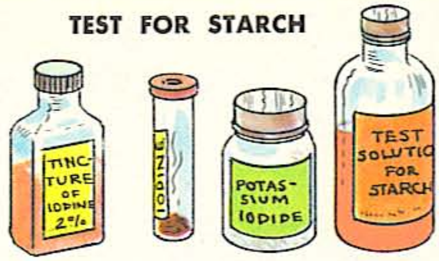

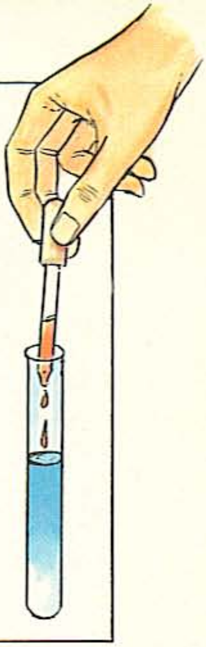

SOLUTION FOR TESTING STARCH CAN BE MADE BY DILUTING 5 ml REGULAR TINCTURE OF IODINE WITH 45 ml WATER.

OR YOU CAN MAKE IT BY DISSOLVING A FEW IODINE CRYSTALS (FROM EXPERIMENT ON PAGE 48) IN A SOLUTION OF 1 g POTASSIUM IODIDE IN 50 ml WATER. IODINE GIVES A BLUE COLOR TO STARCH.

1 PEEL A COUPLE OF POTATOES. GRATE THEM UP ON FINE SIDE OF GRATER.

2 PLACE POTATO PULP ON DOUBLE LAYER OF CHEESE CLOTH. FOLD UP AND TIE. DIP BAG IN POT OF WATER. SQUEEZE. DIP AND SQUEEZE UNTIL MOST STARCH IS OUT.

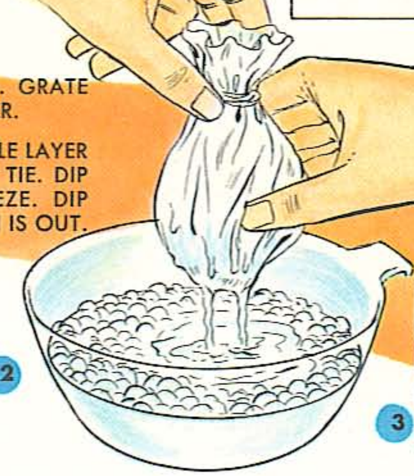

3 LET STARCH SETTLE IN POT. POUR OFF MOST OF WATER. POUR REST INTO A JAR FOR FURTHER SETTLING.

4 AGAIN, POUR WATER OFF UNTIL VERY LITTLE IS LEFT. POUR THE REST ONTO A PLATE. PLACE IN A WARM SPOT TILL STARCH IS DRY.

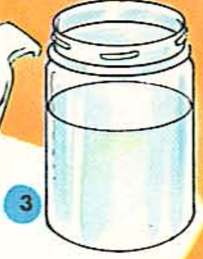

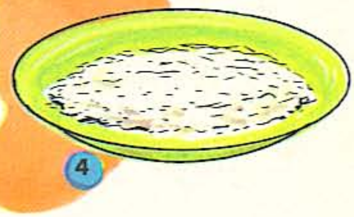

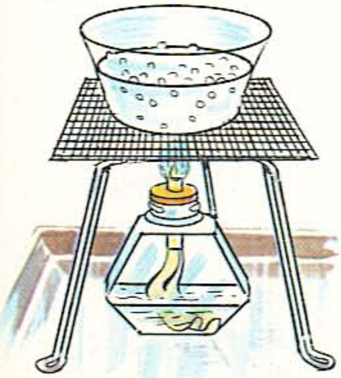

EXPERIMENTS WITH STARCH

TO MAKE 1 PER CENT STARCH SOLUTION, STIR ½ g STARCH IN 5 ml COLD WATER. ADD 45 ml MORE WATER. BRING TO BOIL WHILE STIRRING, THEN COOL SOLUTION.

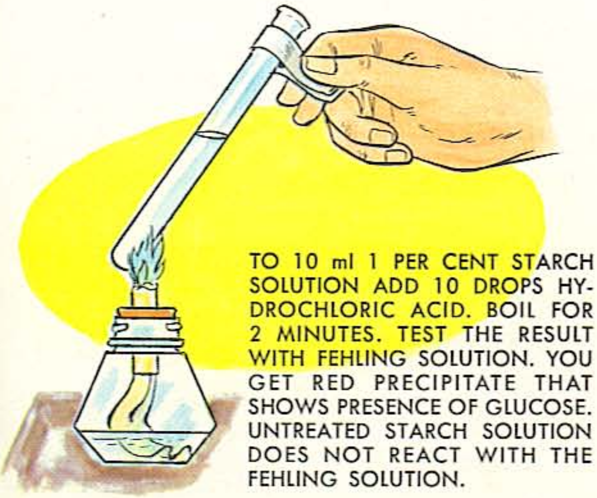

TO 10 ml 1 PER CENT STARCH SOLUTION ADD 10 DROPS HYDROCHLORIC ACID. BOIL FOR 2 MINUTES. TEST THE RESULT WITH FEHLING SOLUTION. YOU GET RED PRECIPITATE THAT SHOWS PRESENCE OF GLUCOSE. UNTREATED STARCH SOLUTION DOES NOT REACT WITH THE FEHLING SOLUTION.

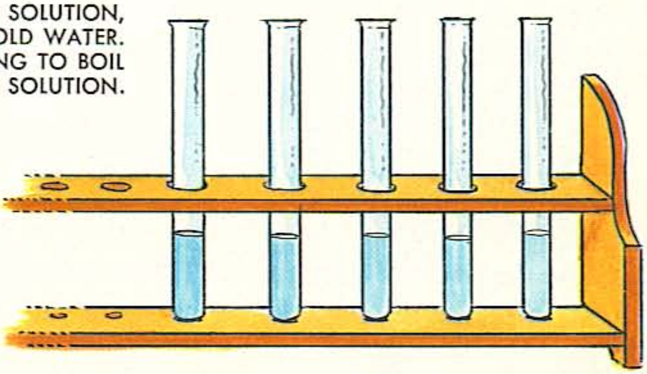

LINE UP FIVE TEST TUBES, EACH CONTAINING 5 ml WATER AND 1 DROP IODINE TEST SOLUTION. IN ANOTHER TEST TUBE, ADD 2 DROPS OF SALIVA (SPITTLE) TO 5 ml STARCH SOLUTION. PLACE THIS IN GLASS OF WARM (NOT HOT) WATER. WITH 2-MINUTE INTERVALS, DROP 3 DROPS SALIVA-STARCH MIXTURE INTO A TEST TUBE WITH IODINE SOLUTION. SHAKE. COLOR GETS LESS AND LESS BLUE. SALIVA DIGESTS THE STARCH AND TURNS IT INTO A SUGAR, MALTOSE.

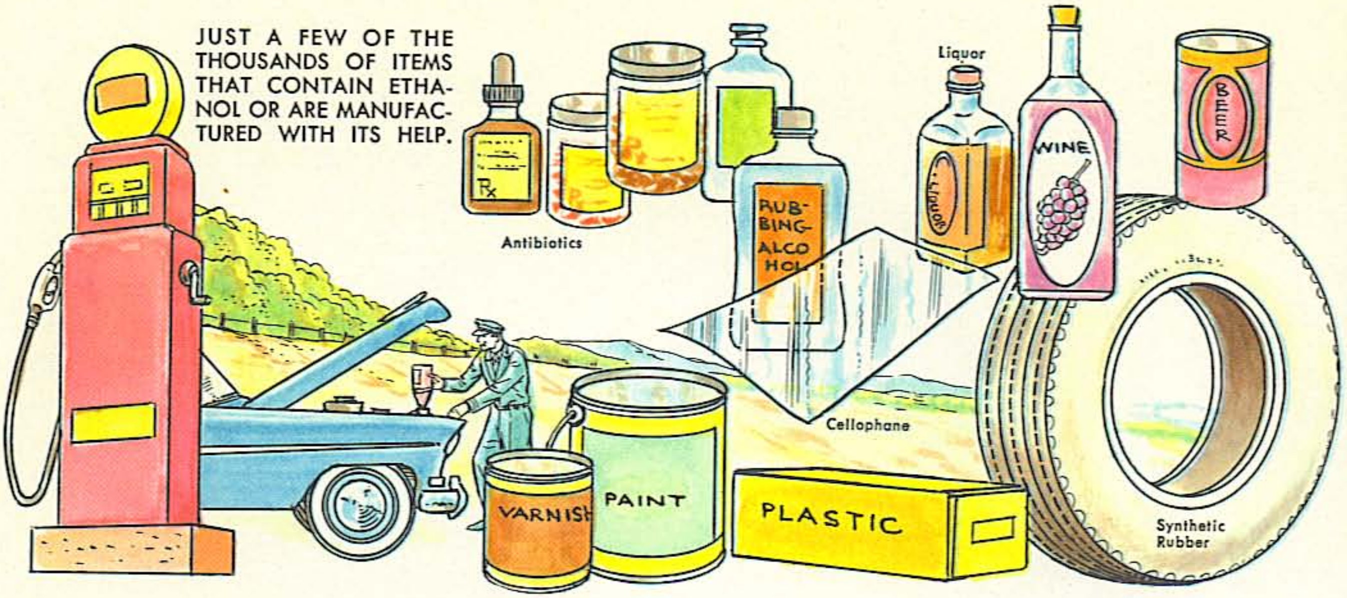
JUST A FEW OF THE THOUSANDS OF ITEMS THAT CONTAIN ETHANOL OR ARE MANUFACTURED WITH ITS HELP.

Rx
Antibiotics
RUBBING ALCOHOL
Liquor
LIQUOR
WINE
BEER
Cellophane
VARNISH
PAINT
PLASTIC
Synthetic Rubber

Many Kinds of Alcohols

To MOST PEOPLE, alcohol is the strong stuff in beer, wine, and hard liquor. But to a chemist, this is just one of many alcohols.

Alcohols may be considered hydrocarbons in which one or more hydrogen (H) atoms are replaced by hydroxyl (OH) groups. Their names are made up from the names of the hydrocarbons to which they are related by giving these an "-ol" ending. In this way, CH_4, methane, becomes CH_3OH, methanol (also called methyl alcohol); C_2H_6, ethane, becomes C_2H_5OH, ethanol (also known as ethyl or grain alcohol); and so on. Methanol (CH_3OH) was originally called wood alcohol because it was made by the destructive distillation of wood. It is very poisonous and is therefore used to "denature" ethanol, making this unfit for drinking.

Ethanol (C_2H_5OH) is produced today, to a great extent, in the same way in which it was made thousands of years ago, by a process called fermentation. In this, the tiny plant cells of yeast are made to grow in the solution of a simple sugar, such as glucose ($C_6H_{12}O_6$). In growing, the yeast cells give off a substance called zymase. This acts as a catalyst and turns the glucose into ethanol and carbon dioxide:

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 \uparrow$$

The ethanol is finally separated from the watery liquid by distillation.

Glycerol ($C_3H_5(OH)_3$) is still another alcohol which you probably know better under the name of glycerin. Glycerol may be considered a product of propane (C_3H_8) in which not one but three H atoms have been replaced by OH.

THE "FAMILY TREE" OF ETHANOL—WITH SOME OF ITS CHILDREN, GRANDCHILDREN, AND GREAT-GRANDCHILDREN.

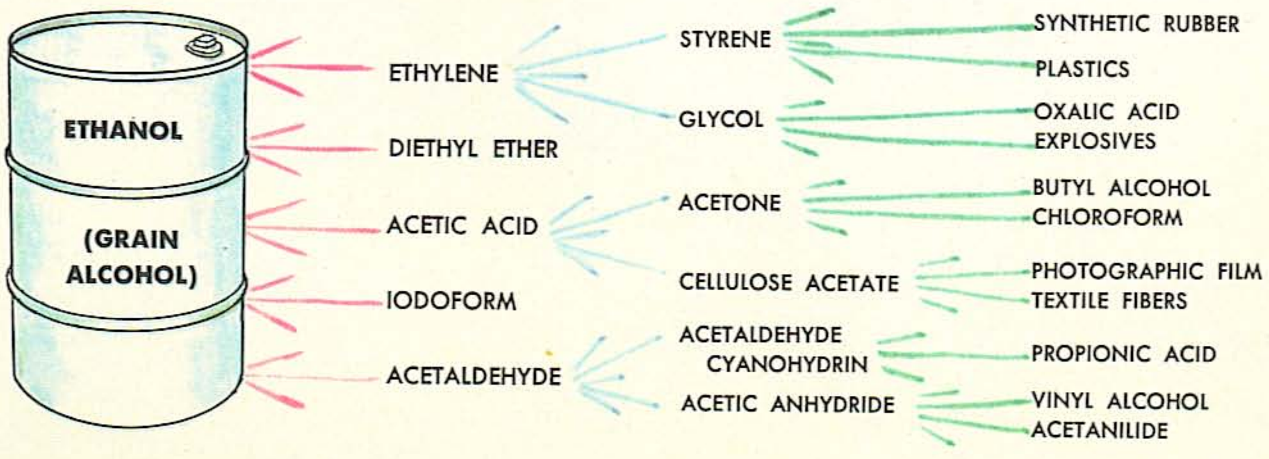
ETHANOL
(GRAIN ALCOHOL)
ETHYLENE
DIETHYL ETHER
ACETIC ACID
IODOFORM
ACETALDEHYDE
STYRENE
GLYCOL
ACETONE
CELLULOSE ACETATE
ACETALDEHYDE CYANOHYDRIN
ACETIC ANHYDRIDE
SYNTHETIC RUBBER
PLASTICS
OXALIC ACID
EXPLOSIVES
BUTYL ALCOHOL
CHLOROFORM
PHOTOGRAPHIC FILM
TEXTILE FIBERS
PROPIONIC ACID
VINYL ALCOHOL
ACETANILIDE

METHYL ALCOHOL—METHANOL

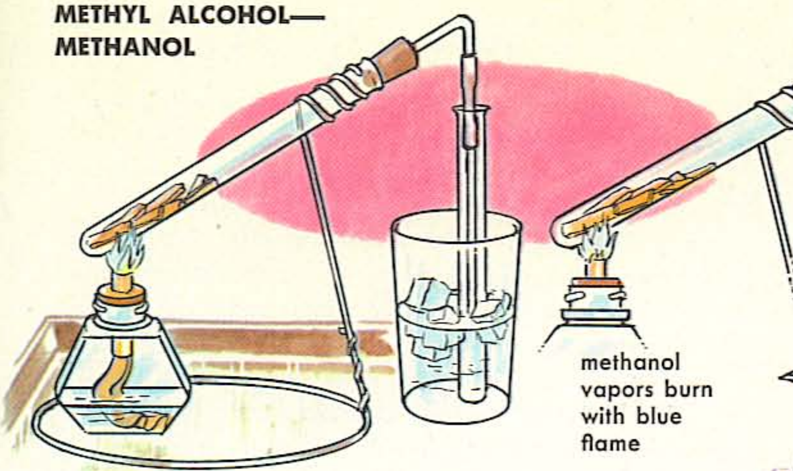
methanol vapors burn with blue flame

METHANOL CAN BE PRODUCED BY DRY DISTILLATION OF WOOD. FILL A TEST TUBE ONE THIRD FULL OF SLIVERS OF WOOD. HEAT. LEAD VAPORS THROUGH L-SHAPED GLASS TUBE INTO TEST TUBE IN MIXTURE OF WATER AND ICE.

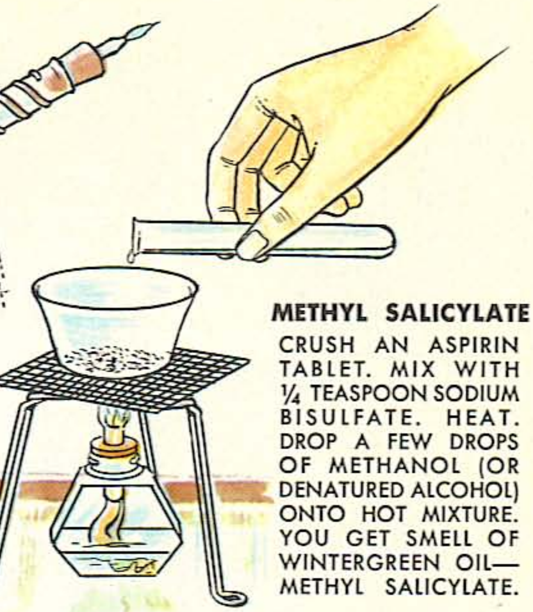

METHYL SALICYLATE

CRUSH AN ASPIRIN TABLET. MIX WITH 1/4 TEASPOON SODIUM BISULFATE. HEAT. DROP A FEW DROPS OF METHANOL (OR DENATURED ALCOHOL) ONTO HOT MIXTURE. YOU GET SMELL OF WINTERGREEN OIL—METHYL SALICYLATE.

ETHYL ALCOHOL—ETHANOL

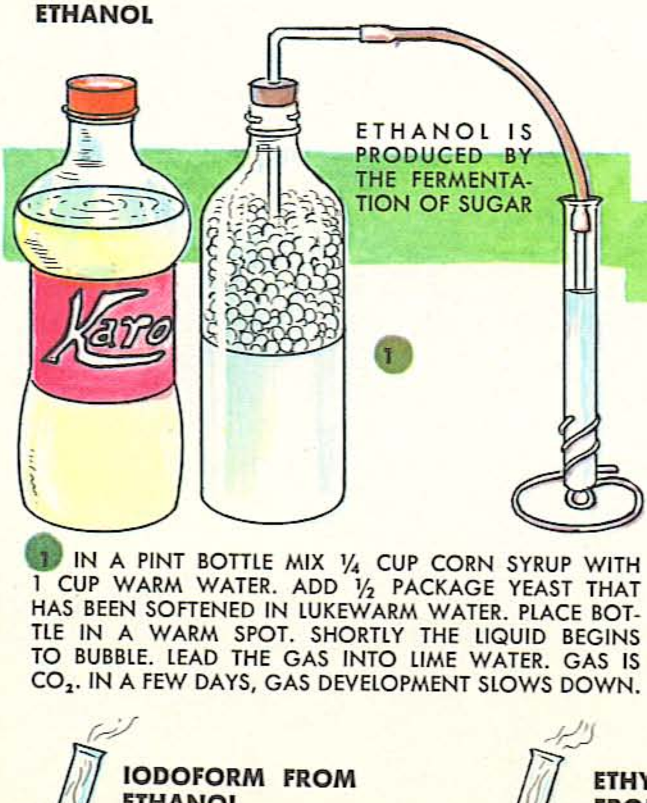

ETHANOL IS PRODUCED BY THE FERMENTATION OF SUGAR

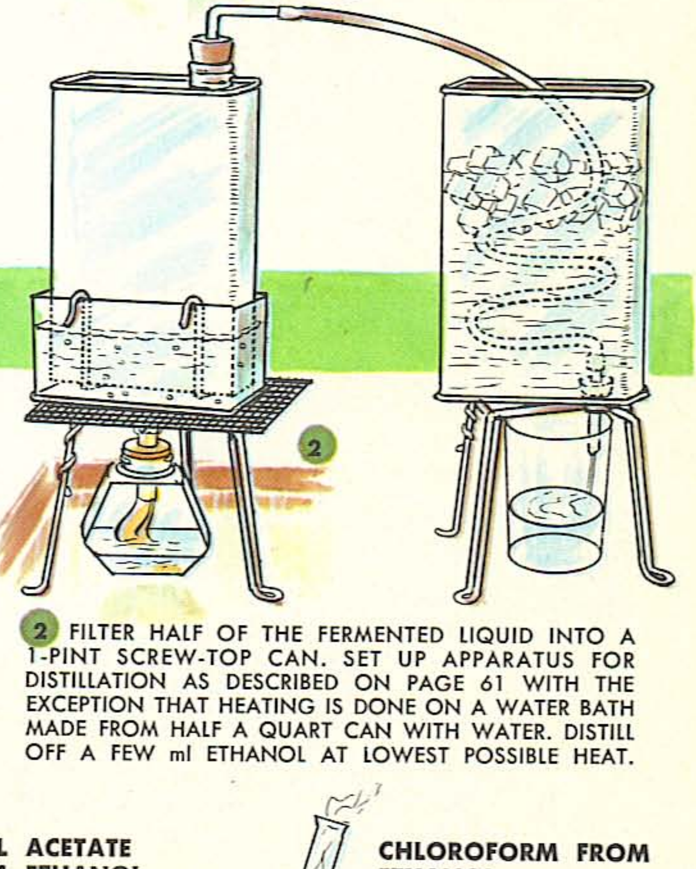

1 IN A PINT BOTTLE MIX 1/4 CUP CORN SYRUP WITH 1 CUP WARM WATER. ADD 1/2 PACKAGE YEAST THAT HAS BEEN SOFTENED IN LUKEWARM WATER. PLACE BOTTLE IN A WARM SPOT. SHORTLY THE LIQUID BEGINS TO BUBBLE. LEAD THE GAS INTO LIME WATER. GAS IS CO_2. IN A FEW DAYS, GAS DEVELOPMENT SLOWS DOWN.

2 FILTER HALF OF THE FERMENTED LIQUID INTO A 1-PINT SCREW-TOP CAN. SET UP APPARATUS FOR DISTILLATION AS DESCRIBED ON PAGE 61 WITH THE EXCEPTION THAT HEATING IS DONE ON A WATER BATH MADE FROM HALF A QUART CAN WITH WATER. DISTILL OFF A FEW ml ETHANOL AT LOWEST POSSIBLE HEAT.

IODOFORM FROM ETHANOL

TO A SOLUTION OF 1 g POTASSIUM IODIDE IN 5 ml WATER ADD IODINE CRYSTALS TO GET DARK BROWN COLOR. ADD 5 ml ETHANOL. ADD 10% NaOH SOLUTION UNTIL COLOR DISAPPEARS. HEAT GENTLY TWO MINUTES. LET COOL. THE YELLOW PRECIPITATE IS IODOFORM—CHI_3.

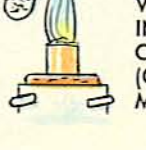

ETHYL ACETATE FROM ETHANOL

IN A TEST TUBE, MIX 3 ml ETHANOL WITH 2 g SODIUM BISULFATE AND 3 ml WHITE VINEGAR. HEAT IT GENTLY. SNIFF CAREFULLY. THE SOUR SMELL OF VINEGAR HAS TURNED INTO THE FRUITY SMELL OF ETHYL ACETATE ($CH_3COOC_2H_5$). IT IS A MUCH-USED SOLVENT.

CHLOROFORM FROM ETHANOL

MIX 5 ml ETHANOL WITH 5 ml SODIUM HYPOCHLORITE SOLUTION ("CLOROX"). HEAT MIXTURE GENTLY FOR A FEW MOMENTS WITHOUT BOILING. THEN SNIFF CAREFULLY. YOU GET THE PECULIAR SWEETISH ODOR OF CHLOROFORM. THE C_2H_5OH HAS BEEN TURNED INTO $CHCl_3$.

Carboxylic Acids

CAN YOU THINK of anything more refreshing than a glass of cold lemonade on a hot summer's day? Or anything better than cranberry sauce for adding a tangy taste to the Thanksgiving dinner?

The tartness of lemonade and cranberry sauce comes from organic acids.

These acids are found ready-made in nature in great numbers. Some of them occur as free acids (citric acid, tannic acid, malic acid), others as esters (products of acids and alcohols, such as fats and oils and the flavors of many fruits and the odors of many flowers). Still other of these organic acids are produced by the action of bacteria (acetic acid from wine or cider, lactic acid when milk turns sour, butyric acid in rancid butter).

Some organic acids can be extracted directly from the plant parts in which they are found. But to get them in pure and concentrated form it is usually necessary to turn them into sodium or calcium salts and then free the acids from the salts with a stronger acid. Many of the acids which were formerly obtained from plant parts can now be made artificially in the laboratory.

Organic acids have one thing in common. They all contain a combination of one carbon atom, one oxygen atom, and one hydroxyl group (OH). This COOH combination, called a carboxyl group (from a joining-up of the words *carb*on and hydr*oxyl*), has given the organic acids their scientific name, carboxylic acids. When these acids form salts it is the H in the carboxyl group that is replaced by a metal, as, for instance, when CH_3COOH (acetic acid) forms CH_3COONa (sodium acetate).

ACETIC ACID IS WHAT MAKES VINEGAR TASTE SOUR. VINEGAR MEANS "SOUR WINE." THAT IS WHAT IT USED TO BE.

TANNIC ACID, USED FOR TANNING, IS FOUND IN THE BARK OF A NUMBER OF TREES AND IN GALL APPLES ON OAK TREES.

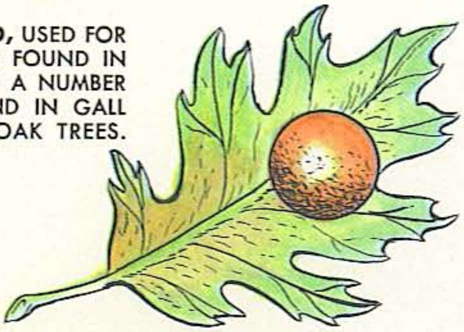

SALICYLIC ACID GOT ITS NAME FROM *SALIX,* LATIN FOR WILLOW. IT WAS FIRST MADE FROM BARK OF WILLOW TREES.

CITRIC ACID GIVES THE TART TASTE TO CITRUS FRUITS—LEMONS, LIMES, GRAPEFRUITS, AND ORANGES.

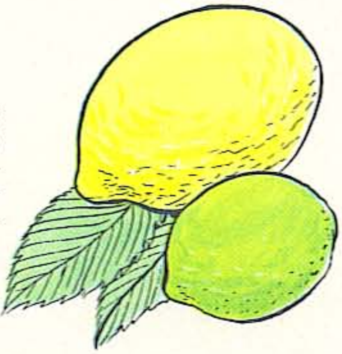

MALIC ACID IS FOUND IN MANY UNRIPE FRUITS—GREEN APPLES, PLUMS, CURRANTS, AND A GREAT NUMBER OF OTHERS. NAME COMES FROM LATIN *MALUS*—APPLE TREE.

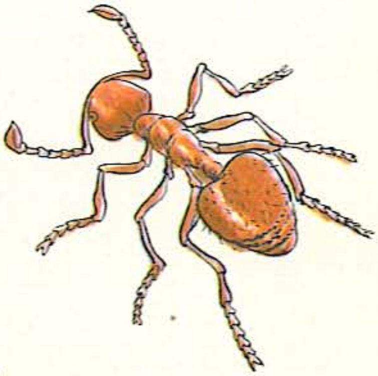

OXALIC ACID ORIGINALLY CAME FROM THE WOOD SORREL PLANT—*OXALIS*. NOW IT IS MADE ARTIFICIALLY.

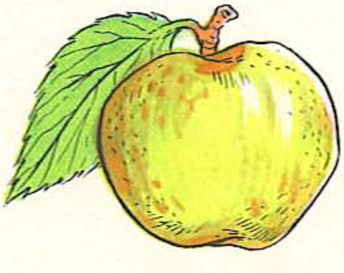

FORMIC ACID IS THE HIGHLY IRRITATING ACID THAT ANTS *(FORMICA)* PUMP INTO YOU WHEN THEY BITE YOU.

ACETIC ACID

YOU HAVE ALREADY MADE SODIUM ACETATE (ON PAGE 83).

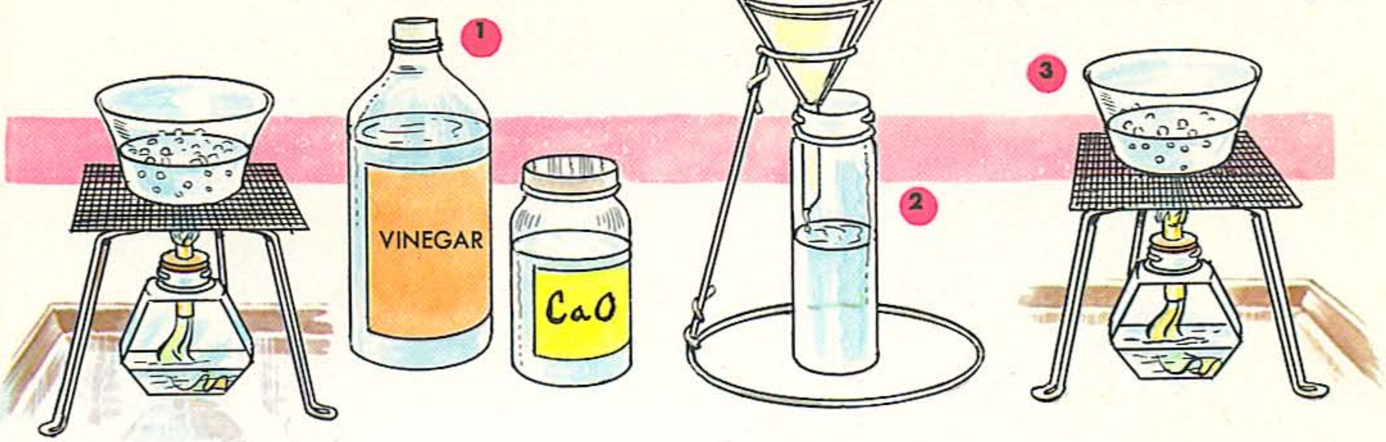

VINEGAR IS DILUTED ACETIC ACID. SEVERAL OF ITS SALTS—ACETATES—CAN BE MADE FROM VINEGAR. USE LIME FOR MAKING THE CALCIUM SALT—$(CH_3COO)_2Ca$.

1 WARM 50 ml WHITE VINEGAR IN A CUSTARD CUP. ADD CALCIUM OXIDE UNTIL NO MORE DISSOLVES.

2 FILTER SOLUTION TO REMOVE UNDISSOLVED CALCIUM OXIDE. FILTRATE CONTAINS CALCIUM ACETATE.

3 EVAPORATE SOLUTION UNTIL ALMOST DRY. DO NOT OVERHEAT—IF YOU DO, THE ACETATE BREAKS UP INTO CALCIUM CARBONATE AND ACETONE (CH_3COCH_3).

YOU CAN AGAIN DRIVE ACETIC ACID OUT OF ITS CALCIUM SALT.

MIX CALCIUM ACETATE WITH AN EQUAL AMOUNT OF SODIUM BISULFATE. PLACE IN DRY TEST TUBE. HEAT GENTLY. YOU GET SHARP ODOR OF ACETIC ACID. MOISTENED BLUE LITMUS PAPER AT MOUTH OF TUBE TURNS RED.

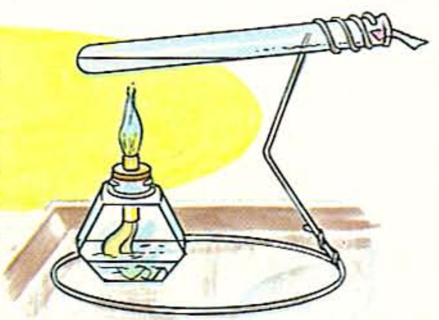

TANNIC ACID

TANNIC ACID IS FOUND IN TEA.

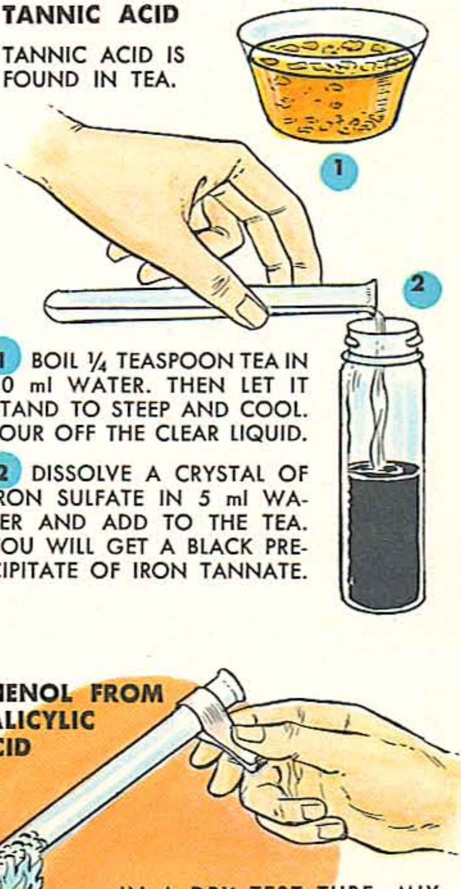

1 BOIL ¼ TEASPOON TEA IN 50 ml WATER. THEN LET IT STAND TO STEEP AND COOL. POUR OFF THE CLEAR LIQUID.

2 DISSOLVE A CRYSTAL OF IRON SULFATE IN 5 ml WATER AND ADD TO THE TEA. YOU WILL GET A BLACK PRECIPITATE OF IRON TANNATE.

SALICYLIC ACID

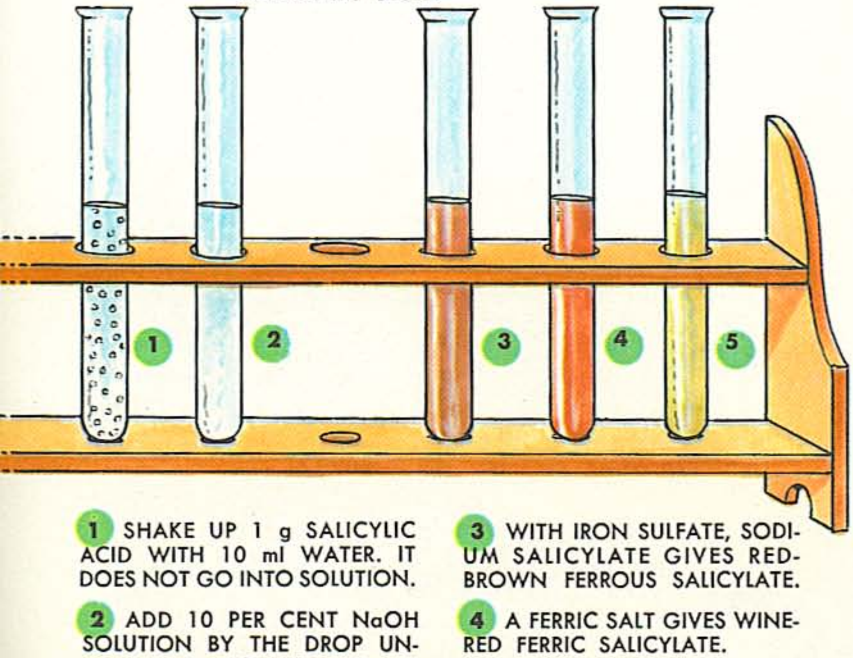

1 SHAKE UP 1 g SALICYLIC ACID WITH 10 ml WATER. IT DOES NOT GO INTO SOLUTION.

2 ADD 10 PER CENT NaOH SOLUTION BY THE DROP UNTIL ALL SALICYLIC ACID IS DISSOLVED. YOU NOW HAVE A SODIUM SALICYLATE SOLUTION.

3 WITH IRON SULFATE, SODIUM SALICYLATE GIVES RED-BROWN FERROUS SALICYLATE.

4 A FERRIC SALT GIVES WINE-RED FERRIC SALICYLATE.

5 COPPER SULFATE GIVES THE GREEN COPPER SALICYLATE.

PHENOL FROM SALICYLIC ACID

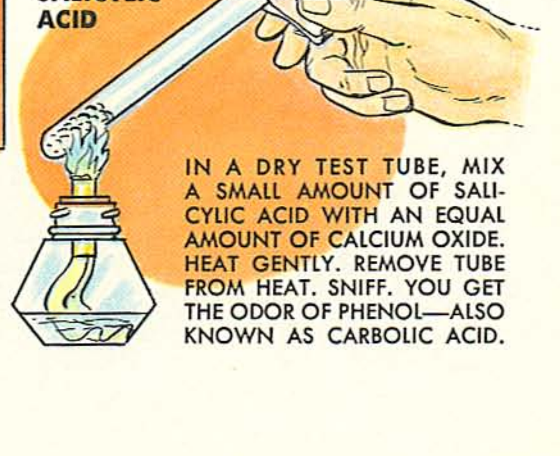

IN A DRY TEST TUBE, MIX A SMALL AMOUNT OF SALICYLIC ACID WITH AN EQUAL AMOUNT OF CALCIUM OXIDE. HEAT GENTLY. REMOVE TUBE FROM HEAT. SNIFF. YOU GET THE ODOR OF PHENOL—ALSO KNOWN AS CARBOLIC ACID.

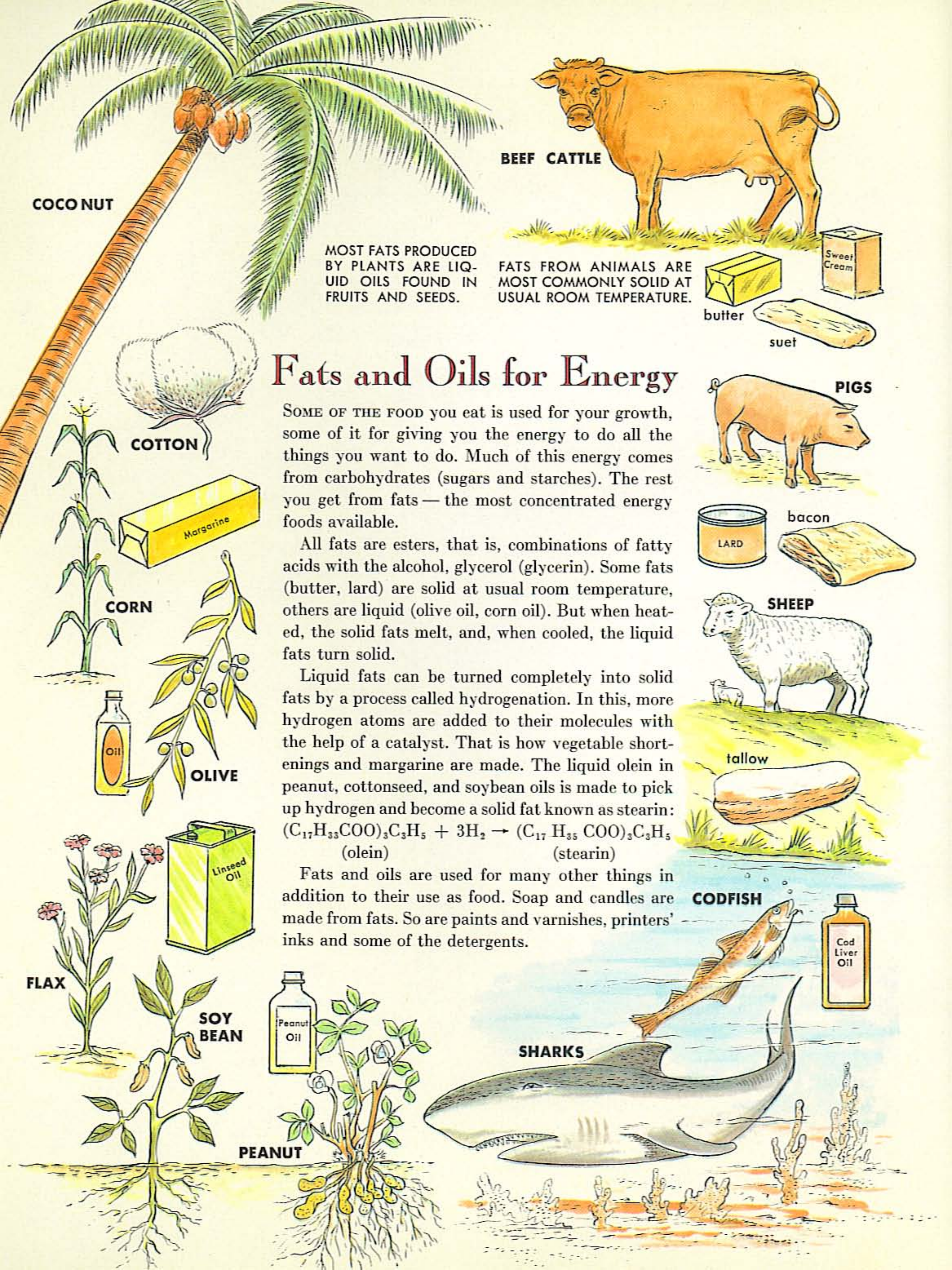

MOST FATS PRODUCED BY PLANTS ARE LIQUID OILS FOUND IN FRUITS AND SEEDS.

FATS FROM ANIMALS ARE MOST COMMONLY SOLID AT USUAL ROOM TEMPERATURE.

COCONUT
BEEF CATTLE
Sweet Cream
butter
suet
COTTON
Margarine
PIGS
LARD
bacon
CORN
SHEEP
Oil
OLIVE
tallow
Linseed Oil
CODFISH
Cod Liver Oil
FLAX
SOY BEAN
Peanut Oil
SHARKS
PEANUT

Fats and Oils for Energy

Some of the food you eat is used for your growth, some of it for giving you the energy to do all the things you want to do. Much of this energy comes from carbohydrates (sugars and starches). The rest you get from fats — the most concentrated energy foods available.

All fats are esters, that is, combinations of fatty acids with the alcohol, glycerol (glycerin). Some fats (butter, lard) are solid at usual room temperature, others are liquid (olive oil, corn oil). But when heated, the solid fats melt, and, when cooled, the liquid fats turn solid.

Liquid fats can be turned completely into solid fats by a process called hydrogenation. In this, more hydrogen atoms are added to their molecules with the help of a catalyst. That is how vegetable shortenings and margarine are made. The liquid olein in peanut, cottonseed, and soybean oils is made to pick up hydrogen and become a solid fat known as stearin:

$$\underset{\text{(olein)}}{(C_{17}H_{33}COO)_3C_3H_5} + 3H_2 \rightarrow \underset{\text{(stearin)}}{(C_{17}H_{35}COO)_3C_3H_5}$$

Fats and oils are used for many other things in addition to their use as food. Soap and candles are made from fats. So are paints and varnishes, printers' inks and some of the detergents.

EXTRACTING FAT

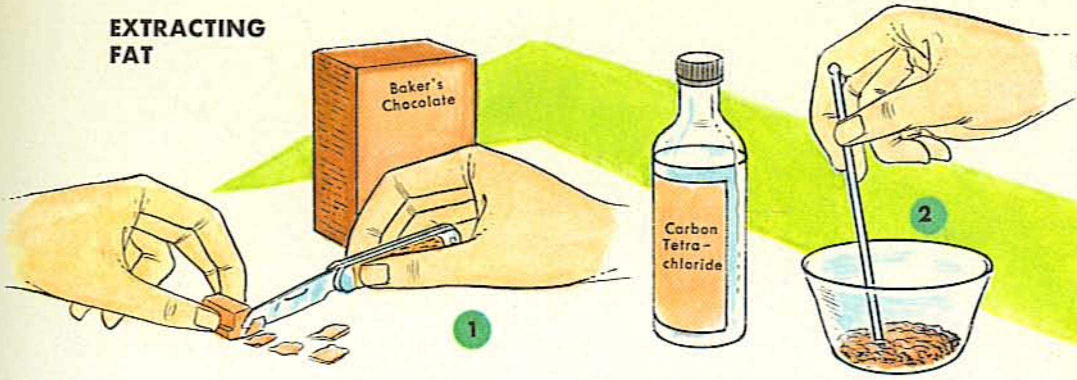

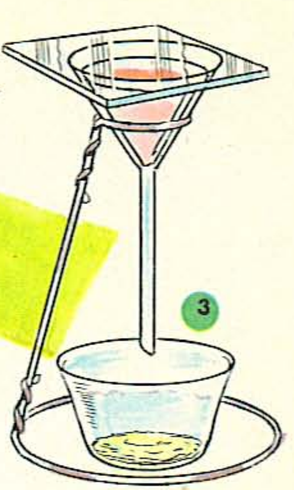

① SHAVE A SMALL SQUARE OF BAKER'S CHOCOLATE OR BITTER CHOCOLATE INTO FINE BITS.

② IN A CUSTARD CUP, POUR CARBON TETRACHLORIDE OVER THE CUT-UP CHOCOLATE AND STIR.

③ FILTER CHOCOLATE-TETRACHLORIDE MIXTURE. LET FILTRATE STAND UNTIL CARBON TETRACHLORIDE HAS EVAPORATED AND YELLOW-WHITE COCOA BUTTER IS LEFT.

Be careful not to breathe fumes.

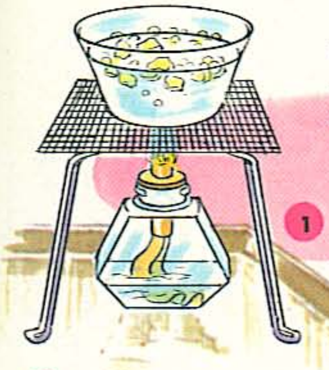

RENDERING FAT

"RENDERING" IS THE MOST COMMON METHOD OF EXTRACTING FAT.

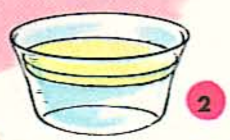

① CUT UP A SMALL AMOUNT OF SUET—THE FAT FROM A PIECE OF BEEF. DROP IT INTO HOT WATER. BOIL WATER FOR TEN MINUTES OR MORE.

② REMOVE THE RENDERED-OUT SUET. PLACE CUSTARD CUP IN REFRIGERATOR. AFTER COOLING YOU CAN LIFT OFF THE FAT AS A SOLID DISK.

TEST FOR FAT

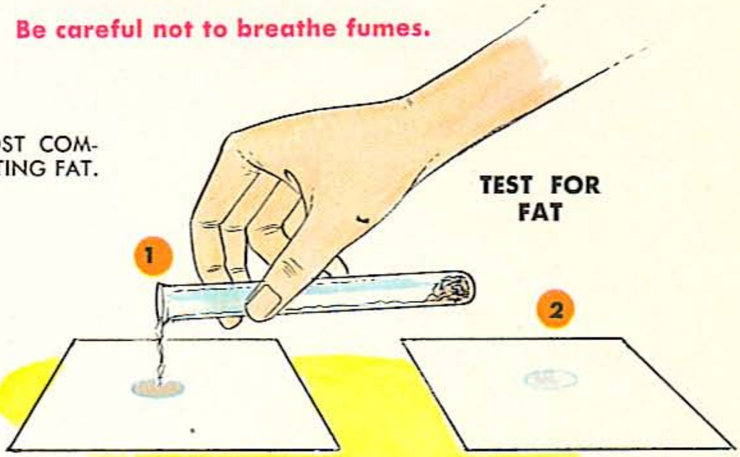

① CRUSH A COUPLE OF PEANUTS. DROP THEM IN A TEST TUBE. COVER THEM WITH CARBON TETRACHLORIDE AND LET STAND ABOUT 5 MINUTES. POUR A FEW DROPS ON A PIECE OF PAPER. LET CARBON TETRACHLORIDE EVAPORATE.

② LOOK AT THE PAPER AGAINST THE LIGHT. THE ALMOST TRANSPARENT "GREASE SPOT" IS A TEST FOR FAT.

TEST FOR GLYCEROL (GLYCERIN)

IN A DRY TEST TUBE ADD ¼ TEASPOON SODIUM BISULFATE TO 1 ml VEGETABLE OIL AND HEAT GENTLY. WAFT THE IRRITATING ODOR TOWARD YOU AND SNIFF CAUTIOUSLY. THE SMELL IS FROM ACROLEIN WHICH IS PRODUCED BY BREAKING DOWN THE GLYCERIN IN THE FAT.

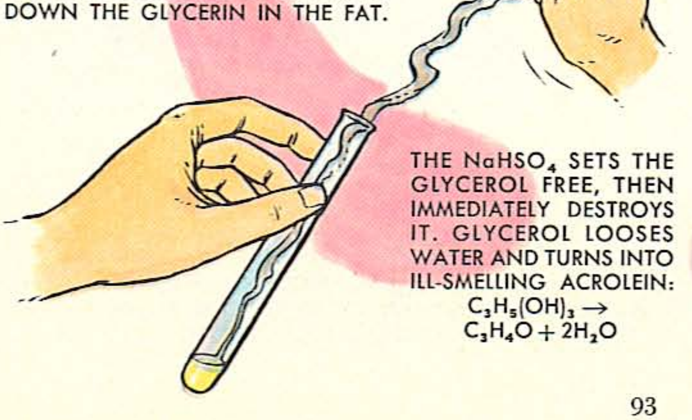

THE $NaHSO_4$ SETS THE GLYCEROL FREE, THEN IMMEDIATELY DESTROYS IT. GLYCEROL LOOSES WATER AND TURNS INTO ILL-SMELLING ACROLEIN:

$$C_3H_5(OH)_3 \rightarrow C_3H_4O + 2H_2O$$

FATTY ACIDS

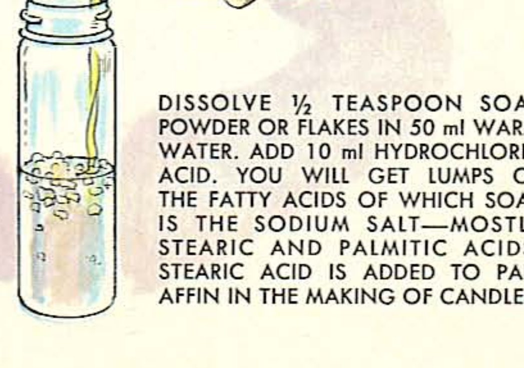

DISSOLVE ½ TEASPOON SOAP POWDER OR FLAKES IN 50 ml WARM WATER. ADD 10 ml HYDROCHLORIC ACID. YOU WILL GET LUMPS OF THE FATTY ACIDS OF WHICH SOAP IS THE SODIUM SALT—MOSTLY STEARIC AND PALMITIC ACIDS. STEARIC ACID IS ADDED TO PARAFFIN IN THE MAKING OF CANDLES.

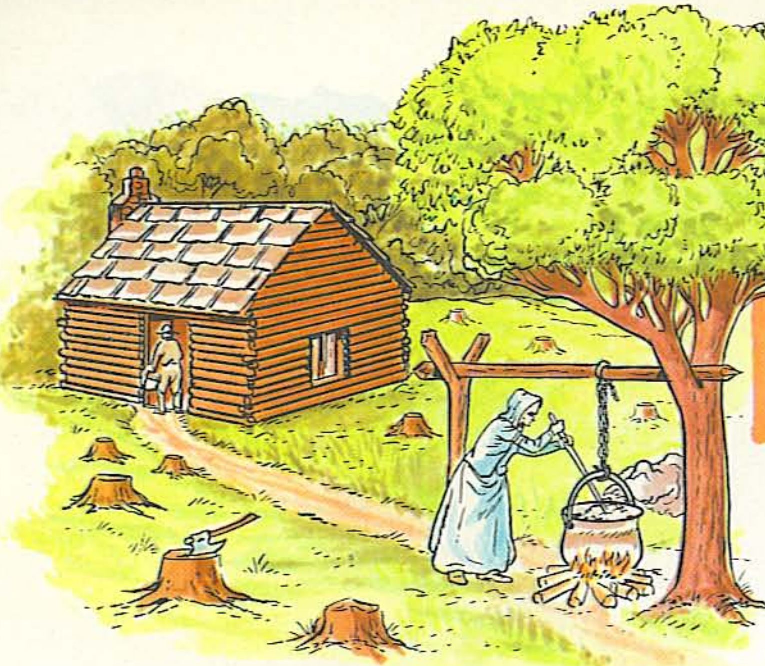

IN THE OLD-FASHIONED SOAP KETTLE, ONLY A FEW GALLONS OF SOAP COULD BE MADE AT ONE TIME.

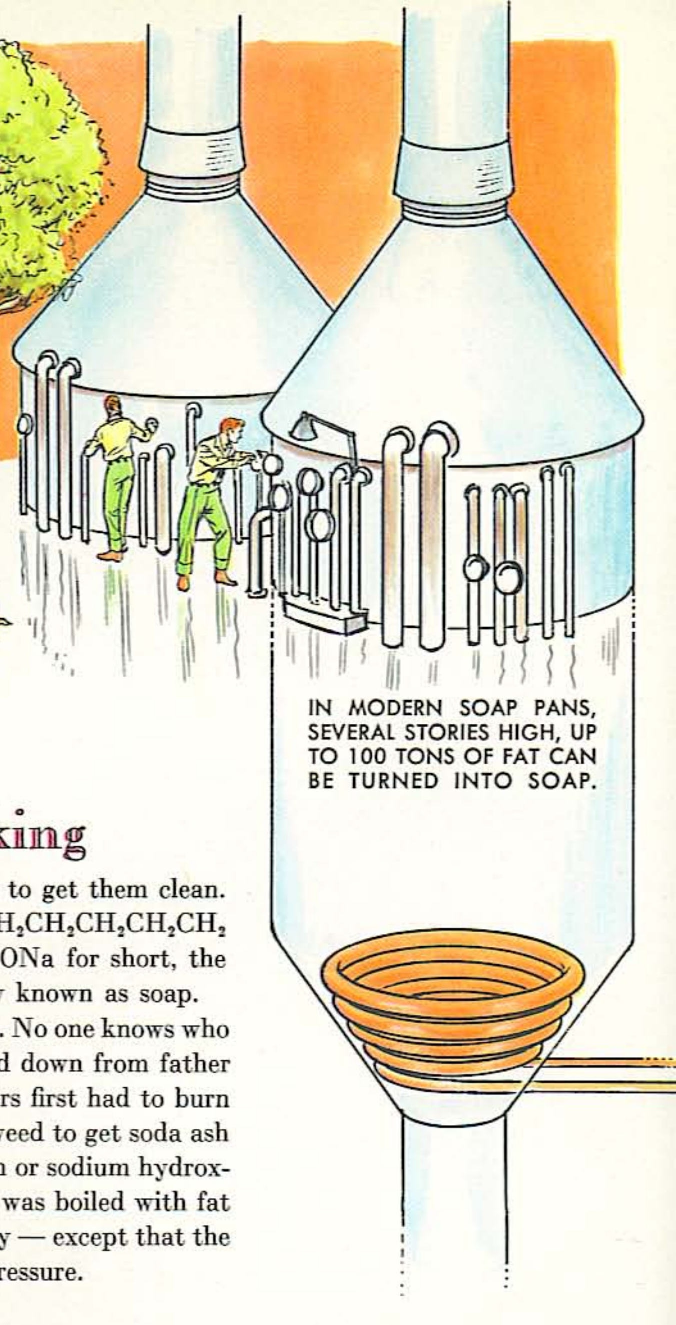

IN MODERN SOAP PANS, SEVERAL STORIES HIGH, UP TO 100 TONS OF FAT CAN BE TURNED INTO SOAP.

Soap and Soap Making

WHENEVER YOUR HANDS get dirty, it is an easy matter to get them clean. All you need is water and plenty of $CH_3CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2CH_2COONa$ — $C_{17}H_{35}COONa$ for short, the sodium salt of stearic acid, a substance more generally known as soap.

Soap has been used for cleaning for thousands of years. No one knows who invented it — but the method for making it was passed down from father to son, from mother to daughter. The early soap makers first had to burn wood to get potash (K_2CO_3—see page 59) or dried seaweed to get soda ash (Na_2CO_3). This was treated with lime to make potassium or sodium hydroxide (KOH or NaOH — see page 45), and this, in turn, was boiled with fat to make soap. Very much the same method is used today — except that the boiling is done in tremendous soap pans under steam pressure.

STRONG SOAP BUBBLES RESULT WHEN YOU ADD GLYCERIN TO THE SOAP SOLUTION. HERE IS A RECIPE: 5 g SOAP, 100 ml WATER, AND 10 ml GLYCERIN.

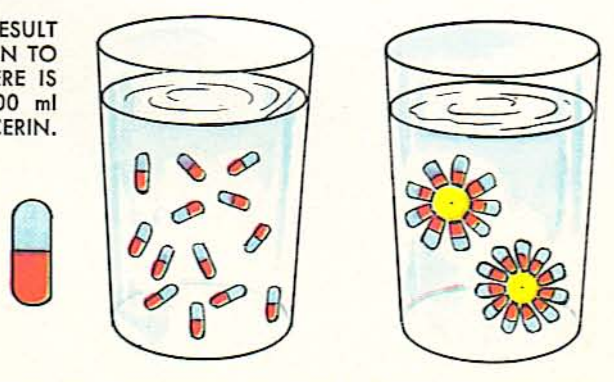

THIS IS THE WAY SCIENTISTS BELIEVE THAT SOAP ACTS: ONE END OF THE SOAP MOLECULE IS SOLUBLE IN WATER, THE OTHER END IN OIL. WHEN OIL IS SHAKEN UP IN SOAPY WATER, THE OIL DROPS ARE SURROUNDED BY THE SOAP MOLECULES DIPPING THE OIL-DISSOLVING ENDS INTO THE OIL. THE WATER-SOLUBLE ENDS HOLD THE OIL DROPLETS SUSPENDED.

MAKING SOAP

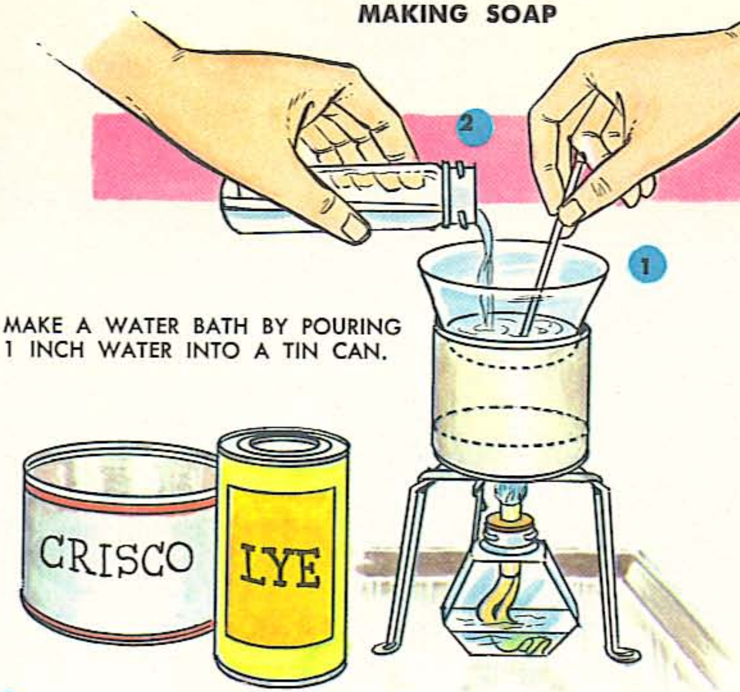

MAKE A WATER BATH BY POURING 1 INCH WATER INTO A TIN CAN.

① MELT 10 g SHORTENING ("CRISCO" OR SIMILAR PRODUCT) IN A CUSTARD CUP ON THE WATER BATH.

② DISSOLVE 5 g SODIUM HYDROXIDE (NaOH) IN 15 ml WATER. ADD 15 ml DENATURED ALCOHOL (TO SPEED UP THE ACTION). POUR THIS SOLUTION INTO THE MELTED SHORTENING WHILE STIRRING. CONTINUE HEATING AND STIRRING UNTIL A SMALL SAMPLE DISSOLVES COMPLETELY IN ½ TEST TUBE WATER. THE SOAP MAKING (SAPONIZATION) IS THEN COMPLETED.

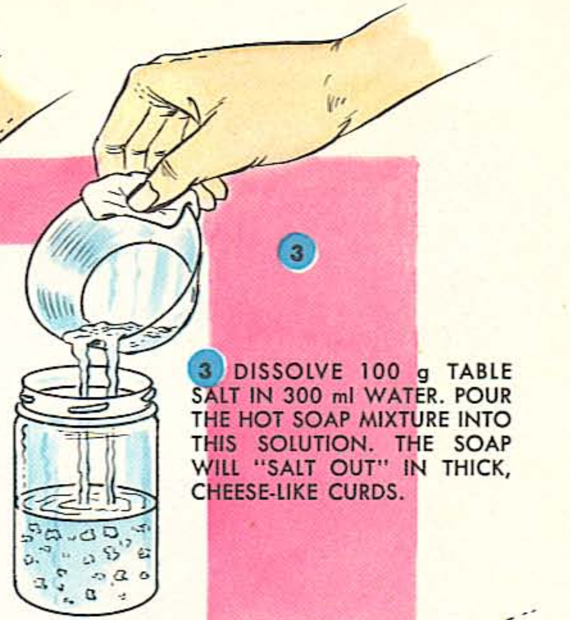

③ DISSOLVE 100 g TABLE SALT IN 300 ml WATER. POUR THE HOT SOAP MIXTURE INTO THIS SOLUTION. THE SOAP WILL "SALT OUT" IN THICK, CHEESE-LIKE CURDS.

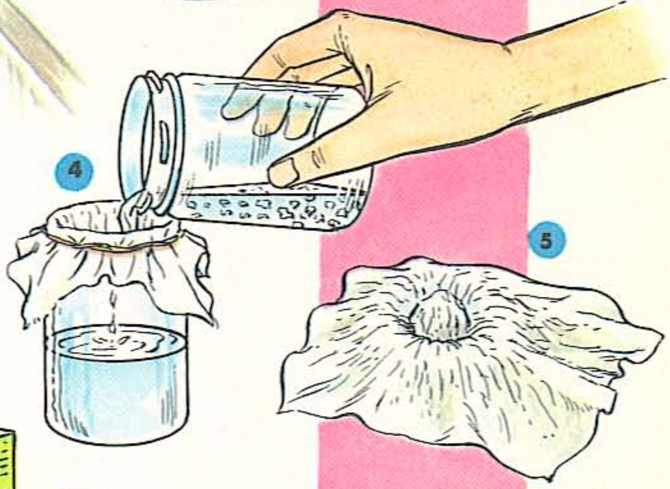

④ TIE A PIECE OF CHEESE CLOTH OVER THE TOP OF A JAR. POUR THE SALT SOLUTION WITH THE SOAP CURDS INTO CHEESE CLOTH AND LET SALT SOLUTION DRAIN OFF. WASH THE SOAP BY POURING TWO TEST TUBES OF ICE-COLD WATER THROUGH IT TO REMOVE MOST OF THE SALT THAT'S STILL ON IT.

⑤ FINALLY, SQUEEZE OUT THE WATER AND SPREAD OUT THE CHEESE CLOTH TO LET THE SOAP DRY.

TESTING SOAP AND DETERGENTS

JoY
Soap Flakes

DISSOLVE 1 g OF YOUR HOMEMADE SOAP IN 50 ml LUKEWARM WATER. ALSO MAKE SOLUTIONS IN 50 ml WATER OF 1 g TOILET SOAP, 1 g SOAP FLAKES, 1 g SOAP POWDER, 1 g POWDERED DETERGENT, AND 1 ml LIQUID DETERGENT.

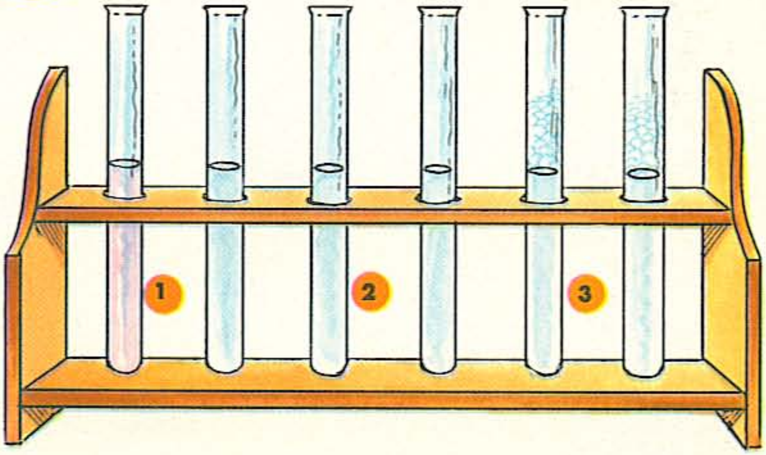

① POUR 10 ml OF THE SOAP AND DETERGENT SOLUTIONS INTO SEPARATE TEST TUBES. TEST EACH SOLUTION FOR ACID AND BASE WITH LITMUS PAPER AND PHENOLPHTHALEIN.

② SHAKE 5 DROPS OF OIL INTO EACH SOLUTION. NOTE THE DIFFERENCE IN THE WAY THE SOLUTIONS MAKE EMULSION WITH OIL.

③ AGAIN, POUR 10 ml OF EACH SOLUTION INTO SEPARATE TEST TUBES. ADD 5 ml LIMEWATER TO EACH. SHAKE AND NOTICE THE DIFFERENCE IN THE AMOUNT OF FOAM MADE BY EACH SOLUTION IN THIS "HARD" WATER.

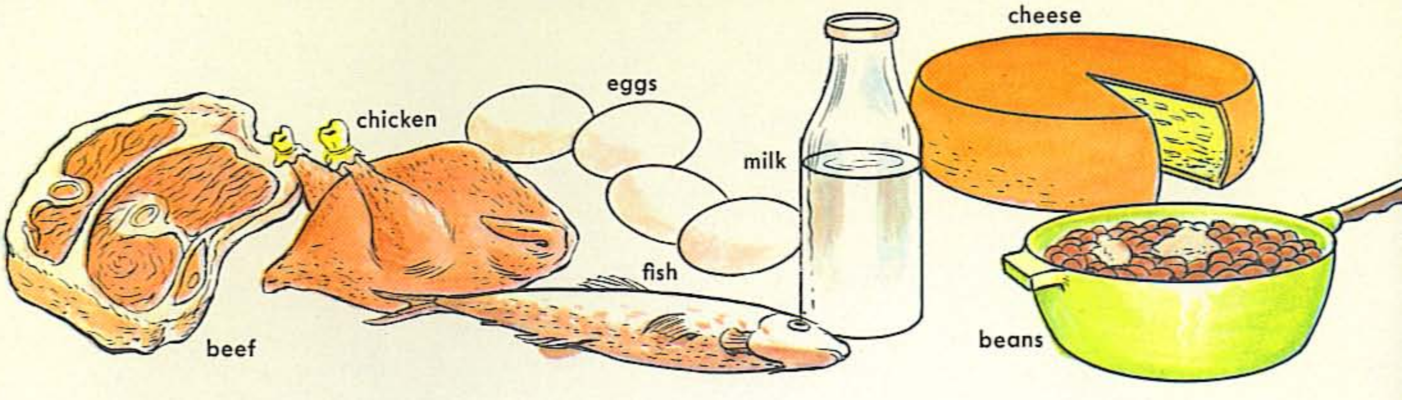

Proteins—the Body-Building Foods

At almost every meal, we look forward especially to the proteins: ham and eggs for breakfast, hamburgers or frankfurters for lunch, steak or chicken for dinner. We drink milk mostly for the sake of its proteins. Even many of our desserts are protein products — from ice cream to Jell-O.

While most other foodstuffs, such as carbohydrates and fats, consist of carbon, hydrogen, and oxygen, the proteins also contain nitrogen and, for the most part, sulfur. Their molecules are "giants" compared with the molecules of other chemical compounds. One of them, albumin in egg, has this estimated formula: $C_{696}H_{1125}O_{200}N_{190}S_{18}$.

Not all proteins are edible. You would hardly think of eating hair and nails, furs and feathers — yet these are all proteins.(CONTINUED ON PAGE 99)

PROTEIN IN EGG

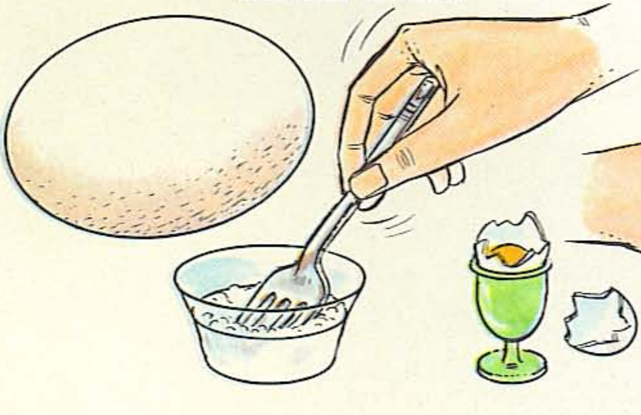

CRACK AN EGG. SEPARATE WHITE FROM YOLK BY LETTING WHITE FLOW INTO A CUP WHILE RETAINING YOLK IN EGG SHELL. BEAT WHITE WITH FORK.

THE PROTEIN IN EGG IS CALLED ALBUMIN

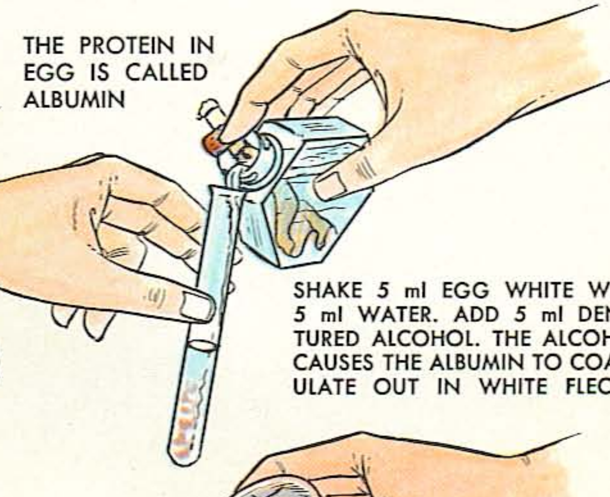

SHAKE 5 ml EGG WHITE WITH 5 ml WATER. ADD 5 ml DENATURED ALCOHOL. THE ALCOHOL CAUSES THE ALBUMIN TO COAGULATE OUT IN WHITE FLECKS.

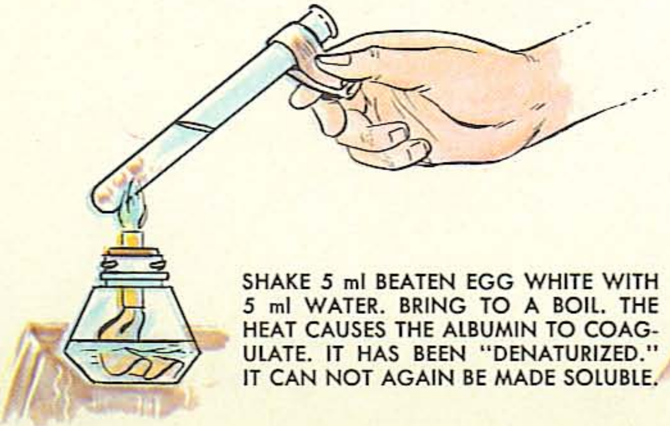

SHAKE 5 ml BEATEN EGG WHITE WITH 5 ml WATER. BRING TO A BOIL. THE HEAT CAUSES THE ALBUMIN TO COAGULATE. IT HAS BEEN "DENATURIZED." IT CAN NOT AGAIN BE MADE SOLUBLE.

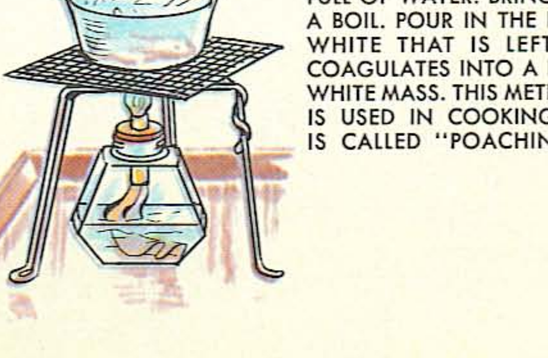

FILL CUSTARD CUP HALF FULL OF WATER. BRING TO A BOIL. POUR IN THE EGG WHITE THAT IS LEFT. IT COAGULATES INTO A FIRM WHITE MASS. THIS METHOD IS USED IN COOKING. IT IS CALLED "POACHING."

THE ITEMS ON THE TOP OF THESE TWO PAGES ALL CONTAIN PROTEINS.

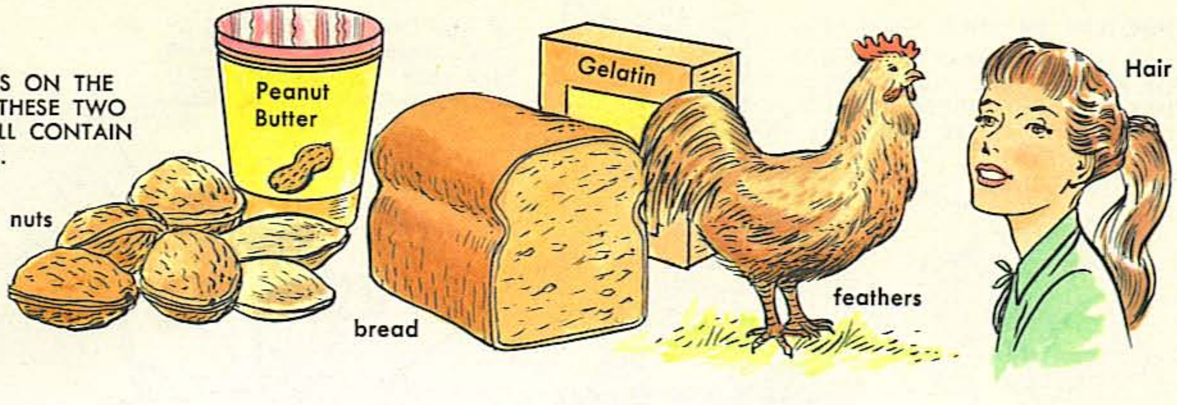

WHAT DOES ALBUMIN CONSIST OF?

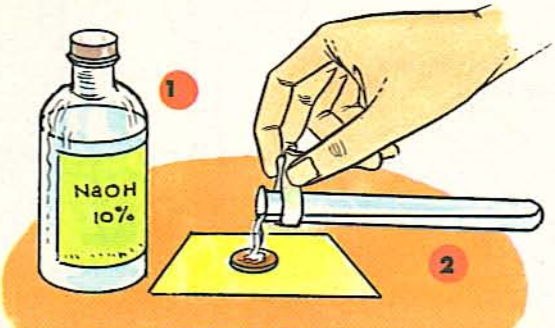

1 DROP A SMALL PIECE OF COAGULATED EGG WHITE INTO A TEST TUBE. COVER IT WITH 5 ml 10% NaOH SOLUTION. HEAT. WHITE GOES IN SOLUTION.

2 POUR A FEW DROPS OF THE EGG WHITE SOLUTION ONTO A BRIGHT SILVER COIN. IN A FEW MINUTES SILVER COIN TURNS BROWNISH-BLACK FROM SILVER SULFIDE, PROVING THAT ALBUMIN CONTAINS SULFUR.

1 PLACE A SMALL PIECE OF COAGULATED EGG WHITE ON A PIECE OF TIN. HEAT. VAPORS SMELL OF AMMONIA AND TURN WETTED RED LITMUS PAPER BLUE. AMMONIA IS NH_3. ALBUMIN MUST CONTAIN N AND H.

ALBUMIN IS FOUND IN EGGS, BLOOD, MILK, AND GRAIN.

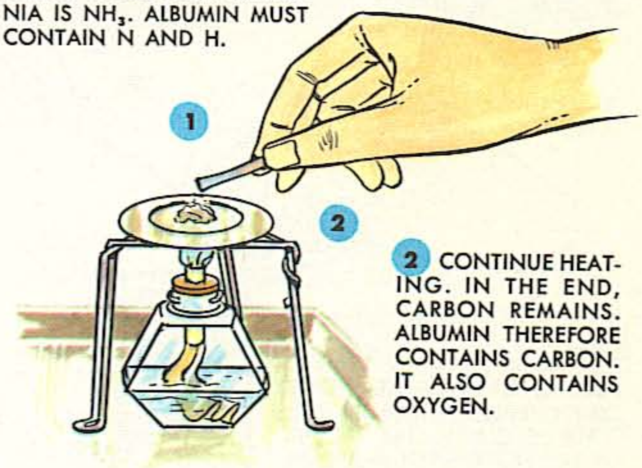

2 CONTINUE HEATING. IN THE END, CARBON REMAINS. ALBUMIN THEREFORE CONTAINS CARBON. IT ALSO CONTAINS OXYGEN.

TEST FOR SOLID WHITE PROTEIN

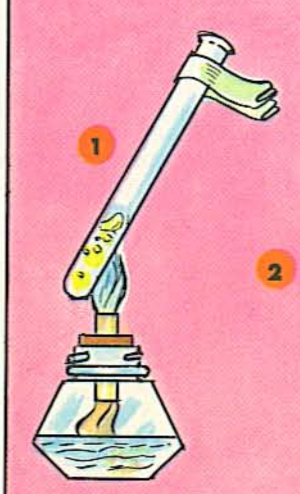

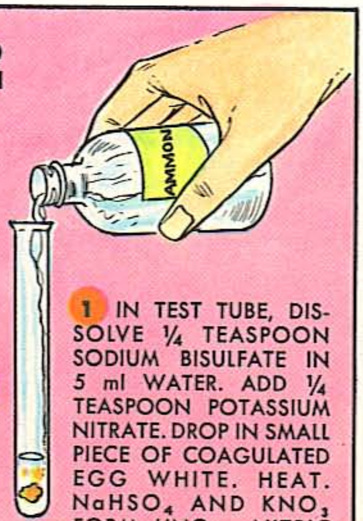

1 IN TEST TUBE, DISSOLVE ¼ TEASPOON SODIUM BISULFATE IN 5 ml WATER. ADD ¼ TEASPOON POTASSIUM NITRATE. DROP IN SMALL PIECE OF COAGULATED EGG WHITE. HEAT. $NaHSO_4$ AND KNO_3 FORM HNO_3—NITRIC ACID. THIS COLORS THE ALBUMIN YELLOW.

2 ADD HOUSEHOLD AMMONIA. THE YELLOW ALBUMIN TURNS BRIGHT ORANGE.

TEST CHEESE, WOOL, CHICKEN, LIMA BEANS THE SAME WAY.

WHAT DOES EGG YOLK CONTAIN?

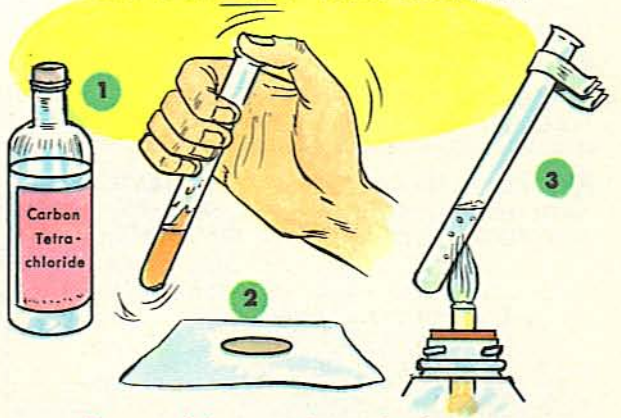

Be careful not to breathe fumes.

1 SHAKE 5 ml OF THE YOLK WITH 5 ml CARBON TETRACHLORIDE TO FIND OUT IF IT CONTAINS FAT.

2 POUR A LITTLE OUT ON PAPER. LET CARBON TETRACHLORIDE EVAPORATE. GREASE SPOT REMAINS.

3 HEAT THE MIXTURE. YOU GET A WHITE COAGULATION. YOLK AND WHITE BOTH CONTAIN ALBUMIN.

PROTEIN IN MILK

MILK IS AN IMPORTANT SOURCE OF PROTEIN. THE PROTEIN IN MILK IS CALLED CASEIN. CHEESE IS SPECIALLY TREATED CASEIN.

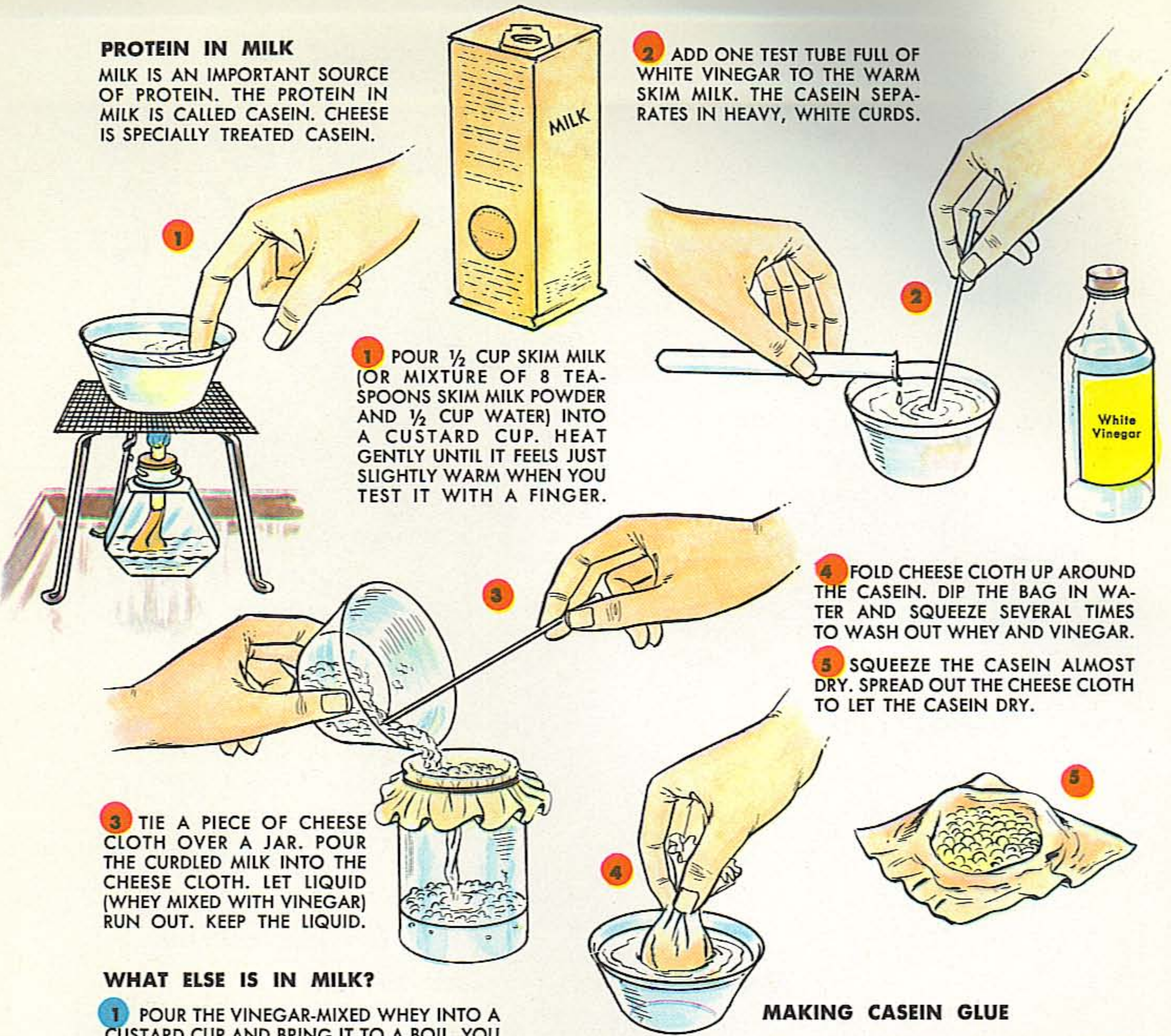

1 POUR ½ CUP SKIM MILK (OR MIXTURE OF 8 TEASPOONS SKIM MILK POWDER AND ½ CUP WATER) INTO A CUSTARD CUP. HEAT GENTLY UNTIL IT FEELS JUST SLIGHTLY WARM WHEN YOU TEST IT WITH A FINGER.

2 ADD ONE TEST TUBE FULL OF WHITE VINEGAR TO THE WARM SKIM MILK. THE CASEIN SEPARATES IN HEAVY, WHITE CURDS.

3 TIE A PIECE OF CHEESE CLOTH OVER A JAR. POUR THE CURDLED MILK INTO THE CHEESE CLOTH. LET LIQUID (WHEY MIXED WITH VINEGAR) RUN OUT. KEEP THE LIQUID.

4 FOLD CHEESE CLOTH UP AROUND THE CASEIN. DIP THE BAG IN WATER AND SQUEEZE SEVERAL TIMES TO WASH OUT WHEY AND VINEGAR.

5 SQUEEZE THE CASEIN ALMOST DRY. SPREAD OUT THE CHEESE CLOTH TO LET THE CASEIN DRY.

WHAT ELSE IS IN MILK?

1 POUR THE VINEGAR-MIXED WHEY INTO A CUSTARD CUP AND BRING IT TO A BOIL. YOU WILL SEE TINY WHITE FLECKS. THESE ARE ALBUMIN COAGULATED OUT BY THE HEAT.

2 FILTER THE WHEY. TEST THE FILTRATE WITH FEHLING SOLUTION (SEE PAGE 85). MILK SUGAR GIVES RED Cu_2O PRECIPITATE.

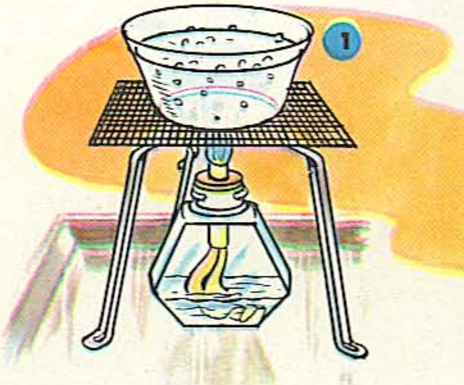

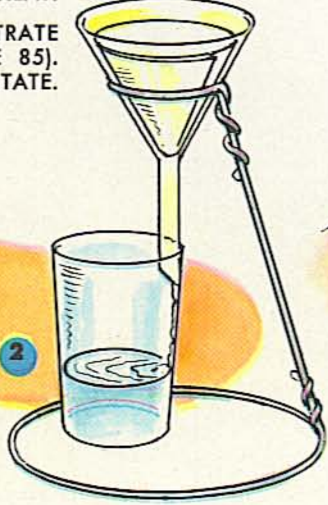

MAKING CASEIN GLUE

SOFTEN 4 g CASEIN WITH 4 ml WATER. SHAKE UP 1 g CALCIUM OXIDE IN 4 ml WATER. POUR THE CALCIUM OXIDE MIXTURE INTO THE CASEIN WHILE STIRRING. THE RESULTING SMOOTH PASTE IS AN EXCELLENT GLUE FOR PAPER AND FOR WOOD.

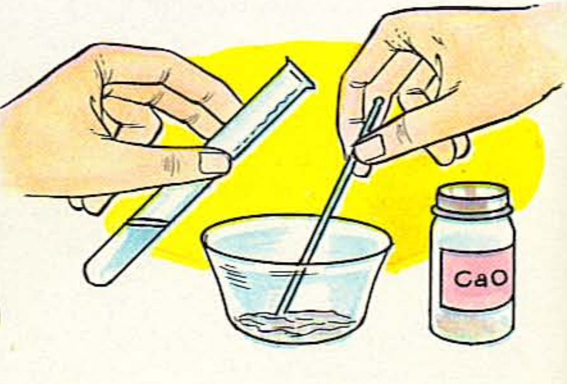

GELATIN IS A PROTEIN

GELATIN IS MADE FROM ANIMAL BONES AND HIDES.

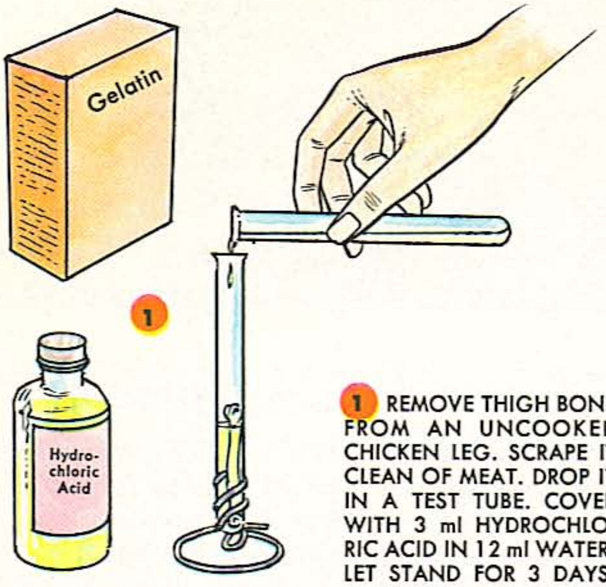

1 REMOVE THIGH BONE FROM AN UNCOOKED CHICKEN LEG. SCRAPE IT CLEAN OF MEAT. DROP IT IN A TEST TUBE. COVER WITH 3 ml HYDROCHLORIC ACID IN 12 ml WATER. LET STAND FOR 3 DAYS.

2 THE DILUTED HYDROCHLORIC ACID DISSOLVES THE CALCIUM SALTS IN THE BONE, LEAVING A FLEXIBLE SUBSTANCE CALLED OSSEIN. WASH OFF THE ACID. EXTRACT THE GELATIN IN THE OSSEIN BY BOILING IN WATER.

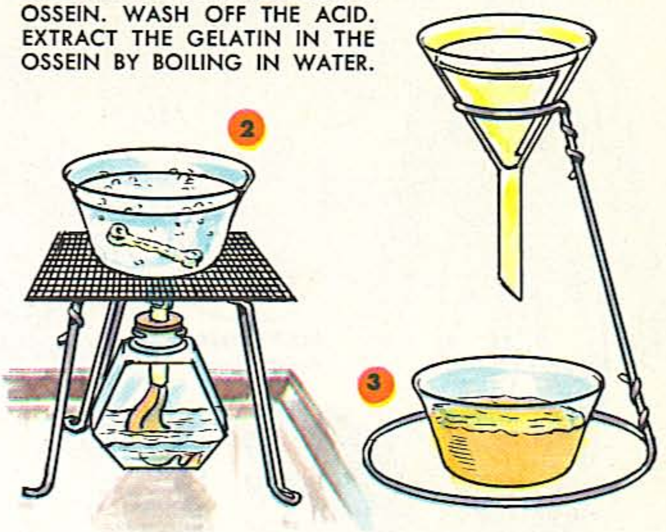

3 FILTER THE SOLUTION. ON COOLING, IT BECOMES JELLY-LIKE. IT HAS TURNED FROM "SOL" TO "GEL."

Proteins—Continued

You are certain to be familiar with three common, pure proteins: albumin in eggs, casein in milk, and gelatin.

ALBUMIN — Egg white contains around 13 per cent albumin — from Latin *albus*, white.

When you shake up egg white with water, you get what looks like an almost clear solution. But this is not a "true" solution such as you get when you dissolve salt or sugar — it is another kind of "solution" called a "colloidal dispersion." For more about colloidal dispersions, see pages 100-101.

As long as egg white is kept cool, it stays transparent and almost liquid. But what happens when you heat it? You know from frying or boiling an egg: It hardens — coagulates — into a solid white mass which you can not again "dissolve" in water. The chemist's term for this change is "denaturation" — the egg white has changed its nature.

CASEIN — Casein is another protein that goes into your diet. Some of the casein you drink (milk), some of it you eat (ice cream and cheese).

In cheese making, the casein is separated from the liquid part of the milk — the whey. It is then pressed and stored until ripe. The flavors of cheeses are caused mostly by esters created during the ripening.

GELATIN — Gelatin is a protein made from animal skins and bones, horns and hooves.

Gelatin behaves in a peculiar manner with water. In cold water it merely swells, but in hot water it "dissolves" readily, forming a colloidal dispersion. As long as you keep this dispersion warm, it remains in a liquid form that is called a "sol." When cooled, it turns into a jelly-like form called a "gel."

TEST FOR LIQUID PROTEINS

MIX 5 ml OF LIQUID TO BE TESTED WITH 5 ml 10% SOLUTION OF NaOH. ADD TWO DROPS OF A 2% SOLUTION OF COPPER SULFATE (1 g IN 50 ml WATER). LIQUID WILL TURN REDDISH-VIOLET.

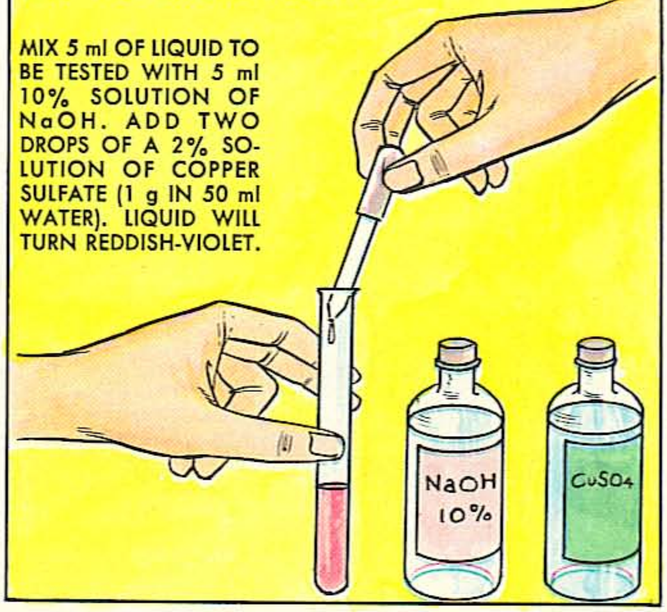

BLOOD
solid in liquid

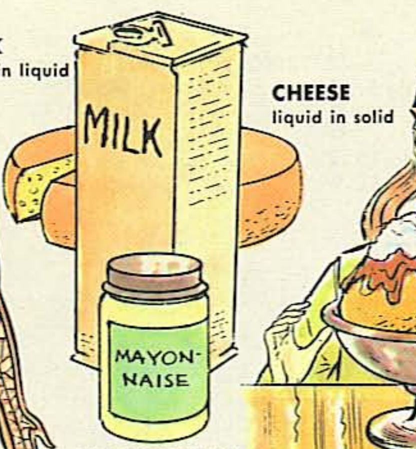
MILK
solid in liquid
MILK
CHEESE
liquid in solid
MAYON-
NAISE
MAYONNAISE
liquid in liquid

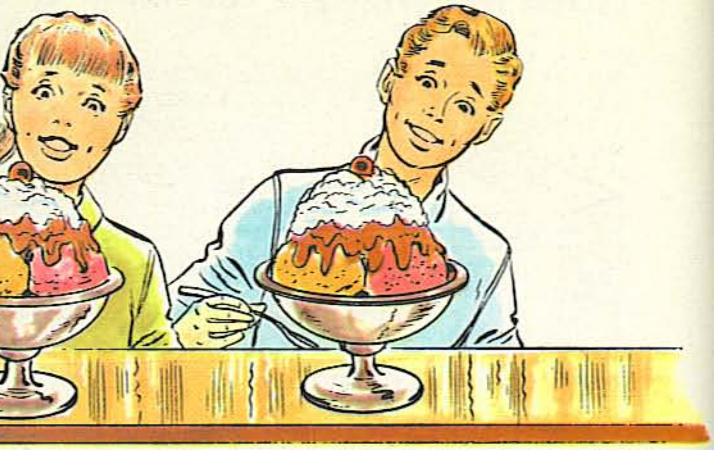

Colloidal Dispersions

WOULD YOU WALK up to a soda fountain and order "a triple, chocolate-flavored colloidal dispersion"? No? Yet that's what you do when you ask for a chocolate sundae. Ice cream is a colloidal dispersion of solids in a liquid; so is chocolate syrup. Whipped cream is a colloidal dispersion of air in a liquid.

It was a Scot, Thomas Graham, who explained colloids, in 1862. He noticed that some solutions passed through parchment paper, others didn't. He discovered that most of those that filtered through were of chemicals that formed crystals — he called them "crystalloids." The others he called "colloids" — from Greek *kollodes*, glue-like.

When a colloid is mixed with water, it does not form a solution but a dispersion. In a solution, the molecules of the dissolved chemical are too small to be seen even with the strongest microscope. In a colloidal dispersion, the much larger particles can be seen in an ultra-microscope — and you can see them as a light effect when you pass a light beam through the dispersion.

Colloidal dispersions can be formed by gases, liquids, and solids. Eight combinations are possible:

IN 1862, A SCOTTISH CHEMIST, THOMAS GRAHAM, PRESENTED HIS IDEAS ABOUT COLLOIDS.

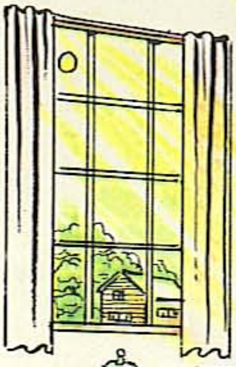
DUSTY AIR
solid in gas

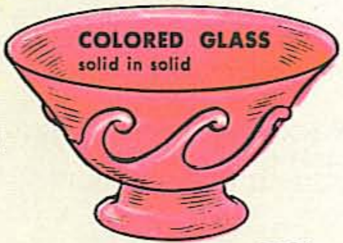
COLORED GLASS
solid in solid

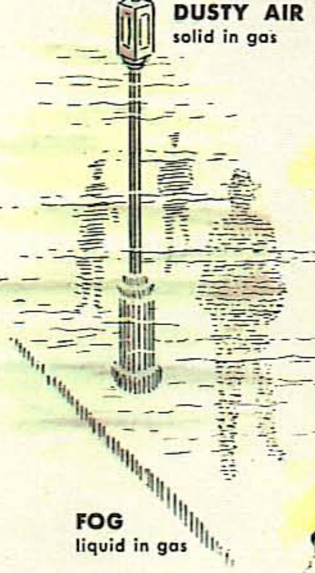
FOG
liquid in gas

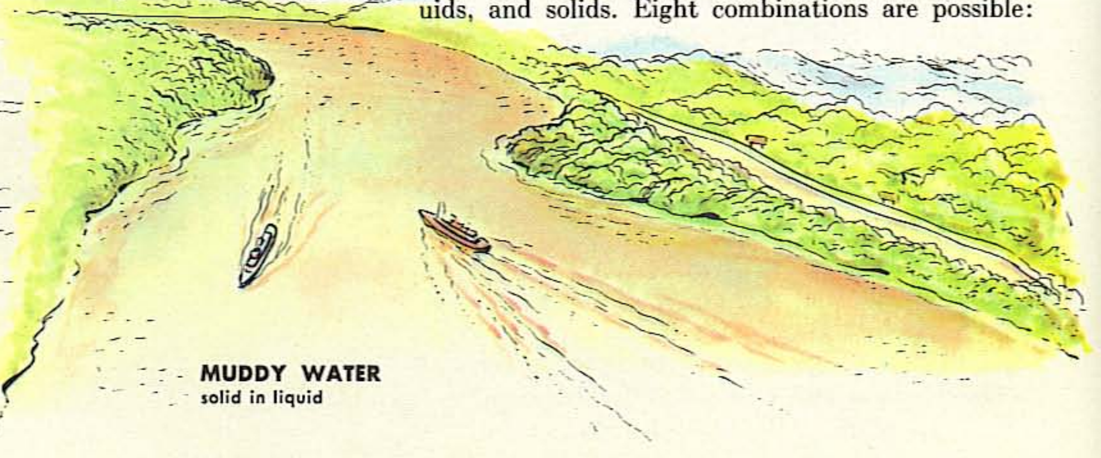
MUDDY WATER
solid in liquid

LIGHT TEST FOR COLLOIDS

THE PARTICLES OF A COLLOIDAL DISPERSION REFLECT LIGHT AND MAKE THE PATH OF A LIGHT BEAM VISIBLE.

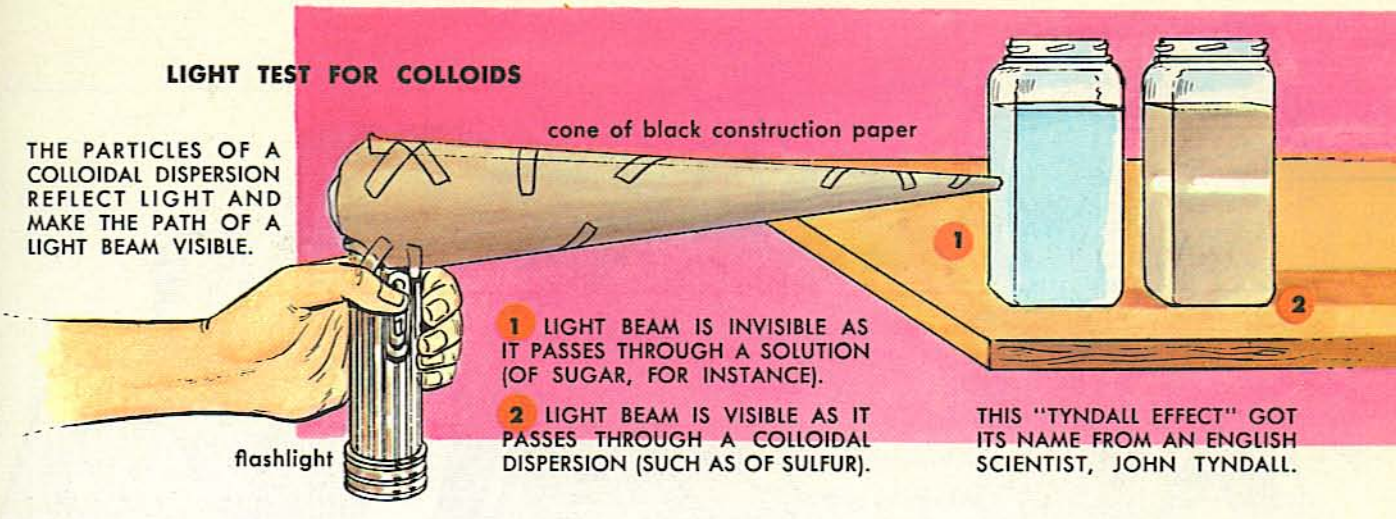
cone of black construction paper
flashlight
1
2

1 LIGHT BEAM IS INVISIBLE AS IT PASSES THROUGH A SOLUTION (OF SUGAR, FOR INSTANCE).

2 LIGHT BEAM IS VISIBLE AS IT PASSES THROUGH A COLLOIDAL DISPERSION (SUCH AS OF SULFUR).

THIS "TYNDALL EFFECT" GOT ITS NAME FROM AN ENGLISH SCIENTIST, JOHN TYNDALL.

gases in liquids and in solids; liquids in gases, in other liquids, and in solids; solids in gases, in liquids, and in other solids. The illustrations show some of these possibilities — you can think of many others.

The colloidal state is important to life. It is the way in which we get most of our food, the way we digest it, and the way the blood carries nourishment throughout our bodies.

IN **PEPTIZATION,** LARGE PARTICLES ARE BROKEN DOWN INTO SMALLER PARTICLES OF COLLOIDAL SIZE.

1 SHAKE UP 1 g STARCH WITH 100 ml COLD WATER. IF LEFT UNDISTURBED, STARCH QUICKLY SETTLES TO BOTTOM.

2 POUR THE MIXTURE OF STARCH AND WATER INTO A CUSTARD CUP. BRING TO A BOIL, THEN COOL. STARCH HAS NOW FORMED A COLLOIDAL DISPERSION.

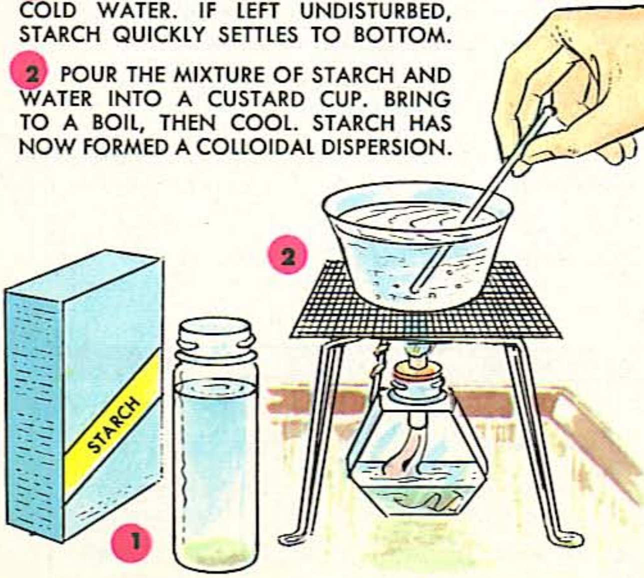

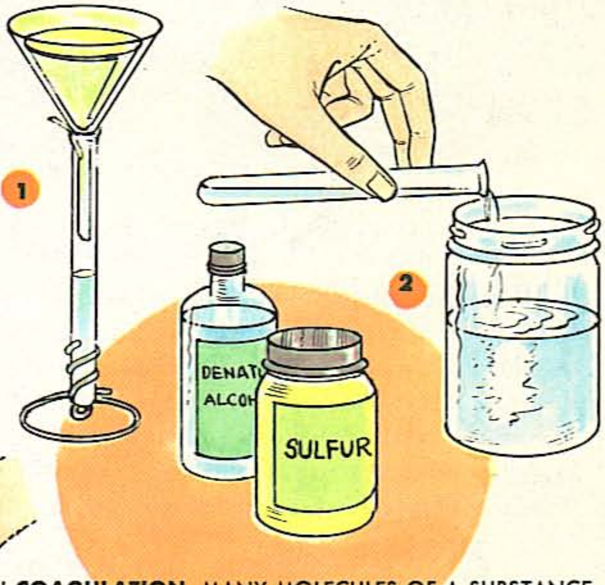

IN **COAGULATION,** MANY MOLECULES OF A SUBSTANCE JOIN TOGETHER INTO PARTICLES OF COLLOIDAL SIZE.

1 SHAKE 1 g FLOWERS OF SULFUR WITH 10 ml DENATURED ALCOHOL. A SMALL AMOUNT OF SULFUR GOES IN SOLUTION. FILTER OUT THE UNDISSOLVED SULFUR.

2 POUR THE ALCOHOLIC SOLUTION OF SULFUR INTO A LARGE AMOUNT OF WATER. YOU WILL SEE A WHITE CLOUD OF FINELY DISPERSED COLLOIDAL SULFUR.

IN **EMULSIFICATION,** ONE LIQUID IS DISPERSED IN ANOTHER. EMULSIONS CAN BE TEMPORARY OR PERMANENT.

1 SHAKE 5 ml KEROSENE AND 5 ml WATER TOGETHER IN A TEST TUBE. LET STAND FOR A SHORT TIME. LIQUIDS SEPARATE. THE EMULSION WAS TEMPORARY.

2 SHAKE 5 ml KEROSENE WITH SOLUTION OF ½ g SOAP IN 5 ml WARM WATER. THEN LET STAND. LIQUIDS DO NOT SEPARATE. THIS IS A PERMANENT EMULSION.

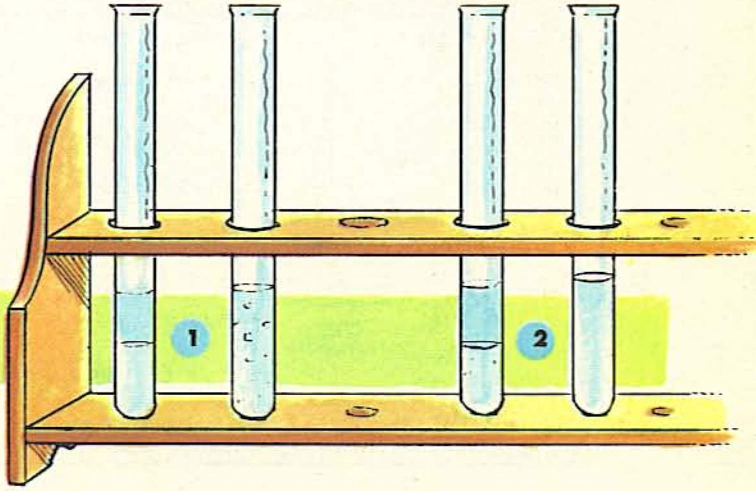

VEGETABLE FIBERS COME FROM PLANTS: COTTON, FLAX (LINEN).

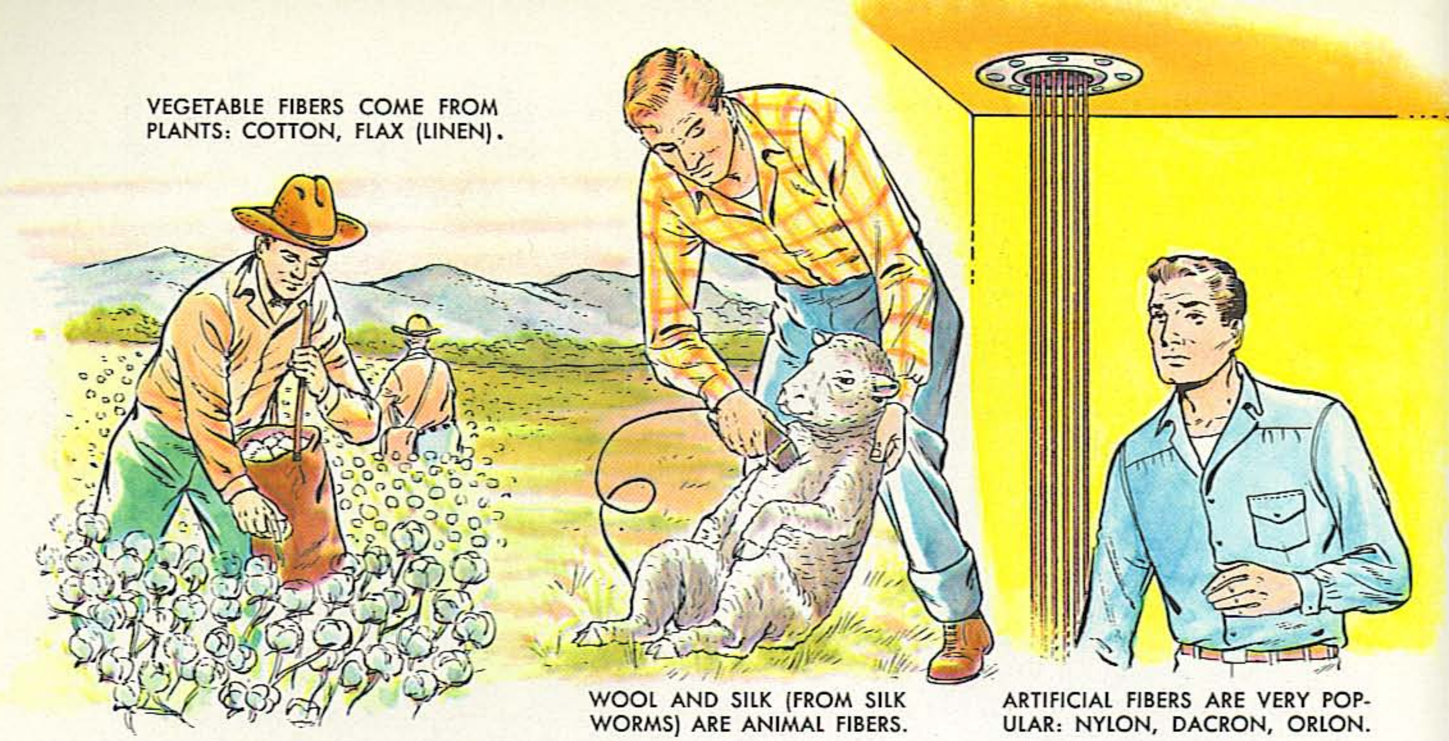

WOOL AND SILK (FROM SILK WORMS) ARE ANIMAL FIBERS.

ARTIFICIAL FIBERS ARE VERY POPULAR: NYLON, DACRON, ORLON.

Natural and Artificial Fibers

It would be tough to get along without fibers in the modern world. Fibers are spun into thread, and the thread is made into cloth for clothing and bedsheets, curtains and towels, and many other things around the house. Fibers also go into such articles as string and rope, rugs and auto tires. Some of these fibers come from the plant and animal worlds, others are manufactured synthetically with coal or petroleum for their starting point.

Fibers belong in different groups of chemical compounds. Animal fibers are proteins; vegetable fibers are cellulose. Artificial fibers such as nylon, Orlon and Dacron are very complex chemical compounds and have enormously long molecules.

BURNING TEST FOR FIBERS

CUT HALF-INCH STRIPS OF DIFFERENT FABRICS. IGNITE EACH STRIP IN TURN. NOTICE HOW FABRIC BURNS, THE SMELL, AND ASH LEFT BEHIND.

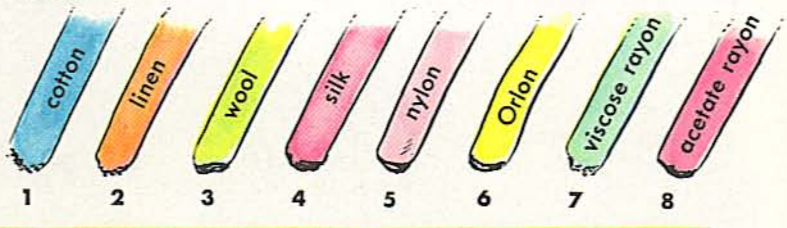

	KIND	FLAME	SMELL	ASH
1	**COTTON**	Rapid, yellow flame	Like burning paper	Small, fine, gray
2	**LINEN**	Fairly fast, yellow flame	Like cotton	Like cotton
3	**WOOL**	Slow, sizzling flame	Like burning hair	Hollow, black bead, easy to crush
4	**SILK**	Small, slow flame	Like wool	Shiny, round bead, easy to crush
5	**NYLON**	Melts; no flame	Like celery	Melts to black bead, hard to crush
6	**ORLON**	Melts and burns	Like broiled fish	Black bead, hard to crush
7	**VISCOSE RAYON**	Rapid, yellow flame	Like cotton	Like cotton
8	**CELLULOSE ACETATE**	Rapid flame with small sparks; melts	Like vinegar	Black bead, hard to crush

CHEMICAL TESTS FOR FIBERS

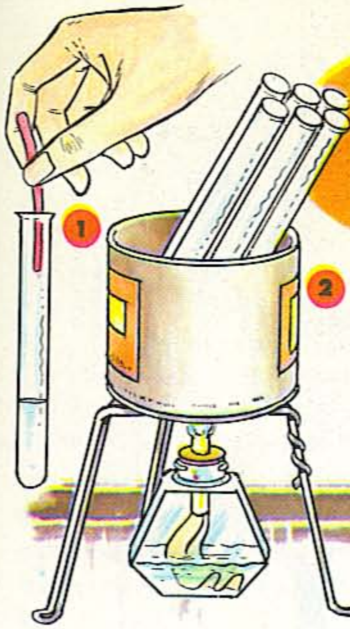

① POUR 5 ml 10% NaOH SOLUTION INTO EACH OF SIX TEST TUBES AND DROP IN STRIPS OF SIX KINDS OF CLOTH. MARK EACH TEST TUBE SO YOU KNOW WHAT IS IN EACH.

② PLACE TEST TUBES IN A CAN OF HOT WATER. BOIL FOR TEN MINUTES.

③ PLACE TEST TUBES IN STAND. NOTE RESULT. WOOL AND SILK HAVE DISSOLVED, THE OTHERS NOT.

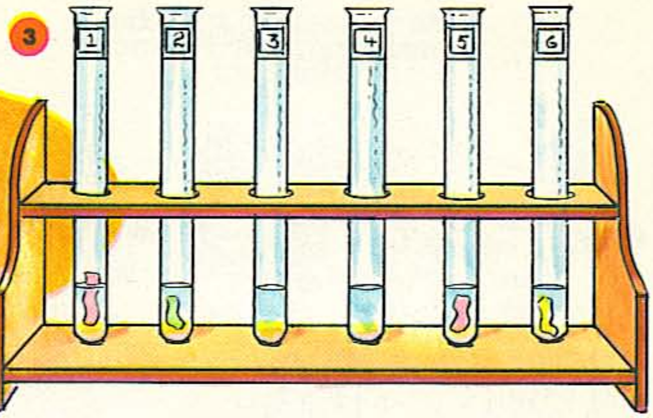

TRY SAME EXPERIMENT WITH STRONG HCl. SILK AND RAYON DISSOLVE. BUT THE WOOL DOES NOT.

MAKING RAYON

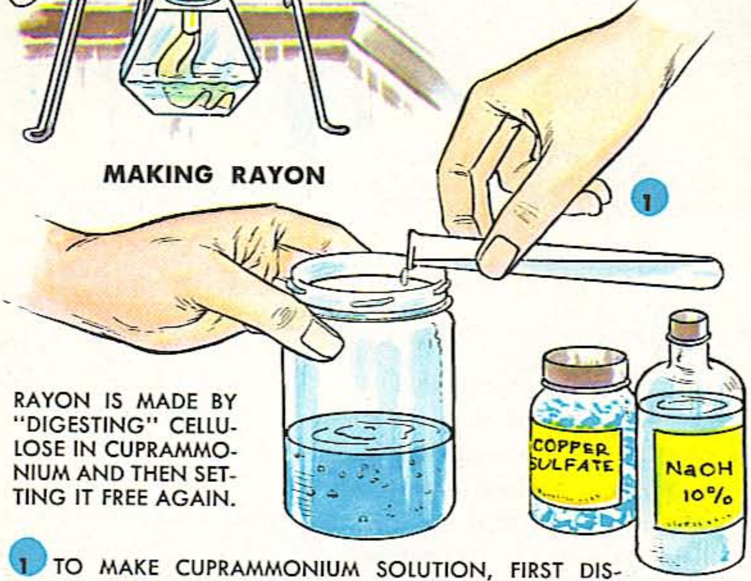

RAYON IS MADE BY "DIGESTING" CELLULOSE IN CUPRAMMONIUM AND THEN SETTING IT FREE AGAIN.

① TO MAKE CUPRAMMONIUM SOLUTION, FIRST DISSOLVE 10 g COPPER SULFATE IN 100 ml WATER IN A PINT JAR. ADD 10% NaOH SOLUTION UNTIL NO MORE LIGHT-BLUE CUPRIC HYDROXIDE FORMS. LET STAND. POUR WATER OFF PRECIPITATE. RE-FILL THE JAR WITH WATER. AGAIN LET STAND. AGAIN POUR WATER OFF PRECIPITATE. REPEAT THIS WASHING PROCESS HALF A DOZEN TIMES.

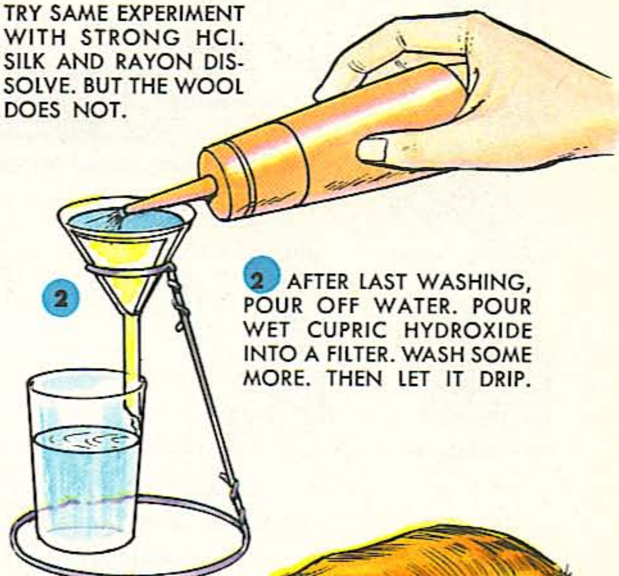

② AFTER LAST WASHING, POUR OFF WATER. POUR WET CUPRIC HYDROXIDE INTO A FILTER. WASH SOME MORE. THEN LET IT DRIP.

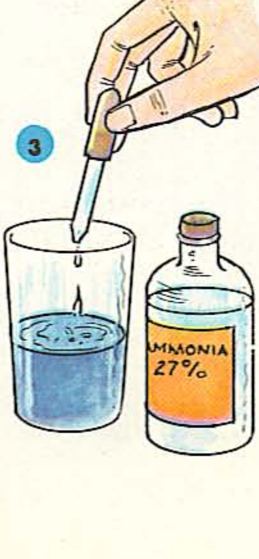

③ SCRAPE THE MOIST CUPRIC HYDROXIDE INTO A GLASS. ADD STRONG AMMONIA (27%, FROM DRUG STORE) BY THE DROP UNTIL ALL $Cu(OH)_2$ HAS DISSOLVED AND HAS BECOME DEEP-BLUE $Cu(NH_3)_4(OH)_2$. THIS SOLUTION IS CALLED "SCHWEITZER'S REAGENT."

④ TEAR UP 3 TO 4 PIECES OF FILTER PAPER. STIR THEM INTO THE BLUE LIQUID. THEY WILL GO INTO SOLUTION.

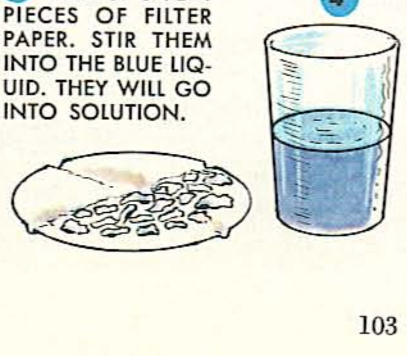

MAKE APPARATUS AS SHOWN BELOW, WITH GLASS TUBE ENDING IN FINE JET TIP. POUR PAPER SOLUTION INTO BOTTLE.

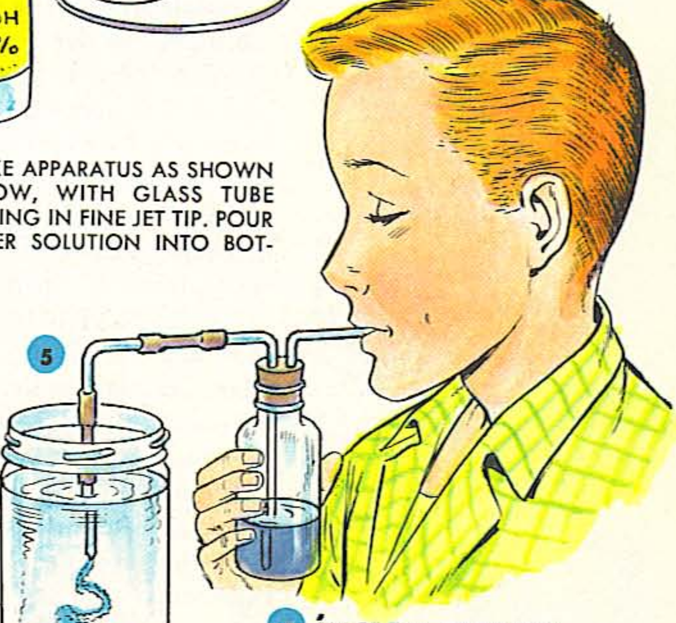

⑤ PLACE THE JET TIP JUST BELOW THE SURFACE OF A MIXTURE OF 10 ml HYDROCHLORIC ACID AND 500 ml WATER. BLOW. AS BLUE LIQUID STREAMS OUT INTO THE DILUTED ACID, IT TURNS INTO WHITISH STRAND OF RAYON THAT SETTLES ON BOTTOM.

THE MAKING OF A TYPICAL THERMOSETTING PLASTIC PHENOLICS

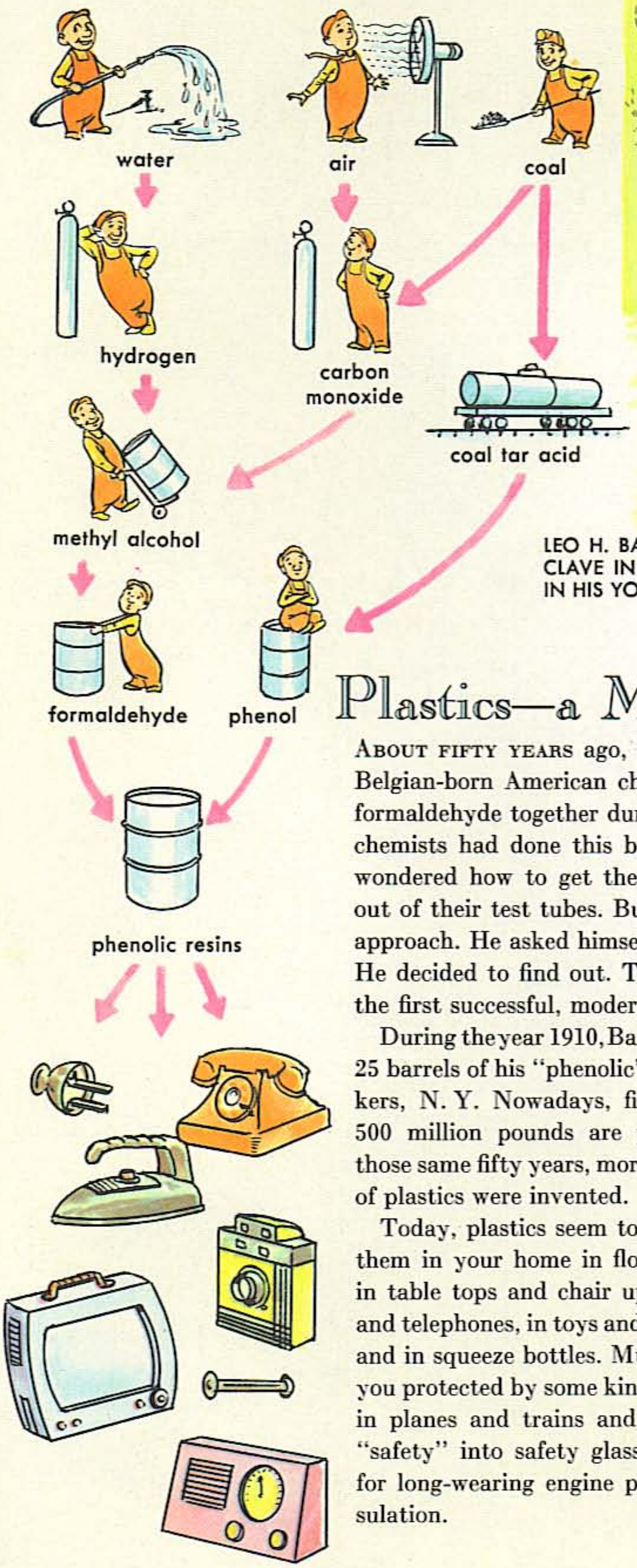

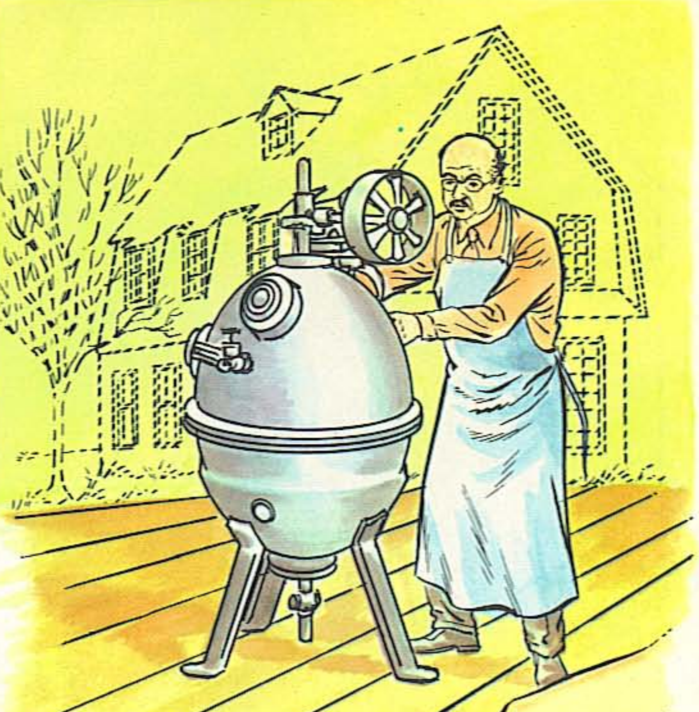

LEO H. BAEKELAND WITH THE AUTOCLAVE IN WHICH HE MADE BAKELITE IN HIS YONKERS, N. Y., LABORATORY.

Plastics—a Modern Giant

About fifty years ago, Dr. Leo H. Baekeland, a Belgian-born American chemist, mixed phenol and formaldehyde together during an experiment. Other chemists had done this before Baekeland and had wondered how to get the messy goo that resulted out of their test tubes. But Baekeland had another approach. He asked himself, "What is it good for?" He decided to find out. The result was Bakelite — the first successful, modern plastic.

During the year 1910, Baekeland produced less than 25 barrels of his "phenolic" plastic in a barn in Yonkers, N. Y. Nowadays, fifty years later, close to 500 million pounds are produced yearly. During those same fifty years, more than a dozen other types of plastics were invented.

Today, plastics seem to be everywhere. You find them in your home in flooring and wall coverings, in table tops and chair upholstery, in TV cabinets and telephones, in toys and games, in rigid containers and in squeeze bottles. Much of your food comes to you protected by some kind of plastic. They are used in planes and trains and cars. A plastic puts the "safety" into safety glass. Other plastics are used for long-wearing engine parts and for electrical insulation.

HEAT TEST

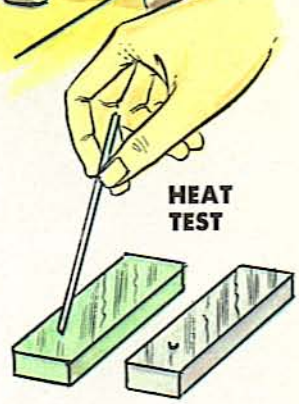

HEAT TIP OF GLASS ROD SLOWLY IN FLAME OF ALCOHOL BURNER. PRESS HOT TIP AGAINST PLASTIC. IT MAKES A DENT IN THERMOPLASTIC, NOT IN THE THERMOSETTING.

BURNING TEST

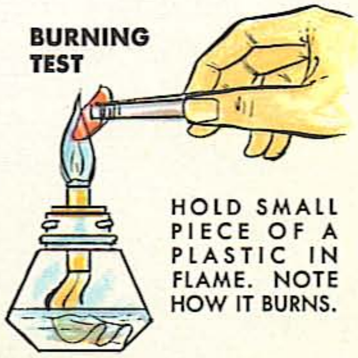

HOLD SMALL PIECE OF A PLASTIC IN FLAME. NOTE HOW IT BURNS.

MOST THERMOSETTING PLASTICS GIVE OFF STRONG ODOR BUT DO NOT BURN. MOST THERMOPLASTICS BURN BUT SOME OF THEM STOP BURNING WHEN REMOVED FROM FLAME.

Plastics are made from a few simple raw materials — some just from water, air, and coal, others with the help of petroleum or natural gas, limestone and salt. The plastics chemist breaks down the comparatively simple molecules of these materials, then builds them up anew into very complex molecules.

Plastics may be divided into two main groups according to their special properties. One group consists of the thermosetting plastics. These can be molded by heat and pressure, but can not be remelted and remolded. They are along the lines of egg white which, once set by heat, stays set. The phenolics and ureas are important thermosetting plastics.

The other group contains the thermoplastics. These are soft when heated, hard when cooled, but can be softened and hardened repeatedly. You can compare them to sulfur and candle wax. The polyethylenes, polystyrenes, vinyls, and acrylics are in the thermoplastics "family."

THE MAKING OF A TYPICAL THERMOPLASTIC VINYL

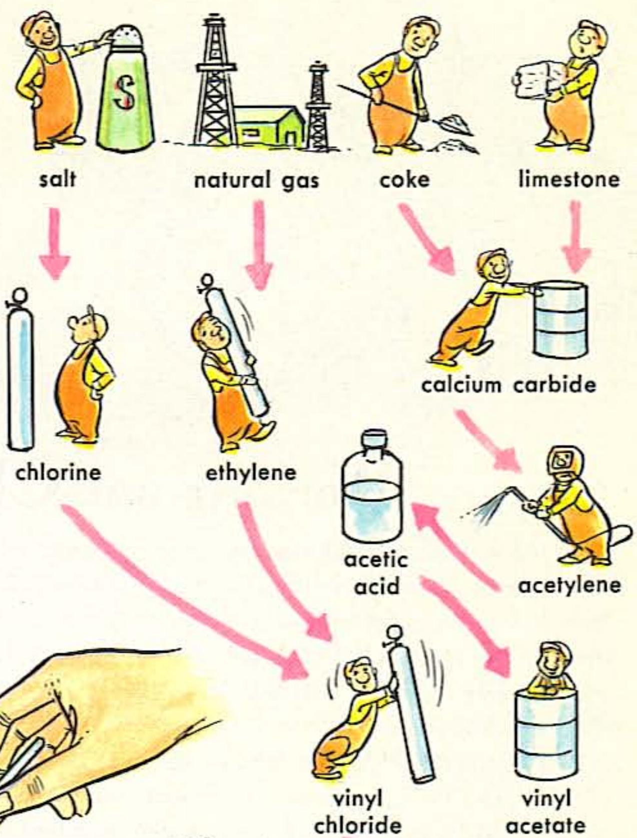

HEAT SHAPING. THERMOPLASTICS BECOME SOFT WHEN HEATED. YOU CAN THEN SHAPE THEM AT WILL.

BRING A POT OF WATER TO A BOIL. DROP IN AN OLD VINYL RECORD. WHEN SOFT, SHAPE IT WITH TWO LONG STICKS. IT BECOMES HARD AGAIN WHEN IT IS REMOVED FROM THE HOT WATER AND COOLED.

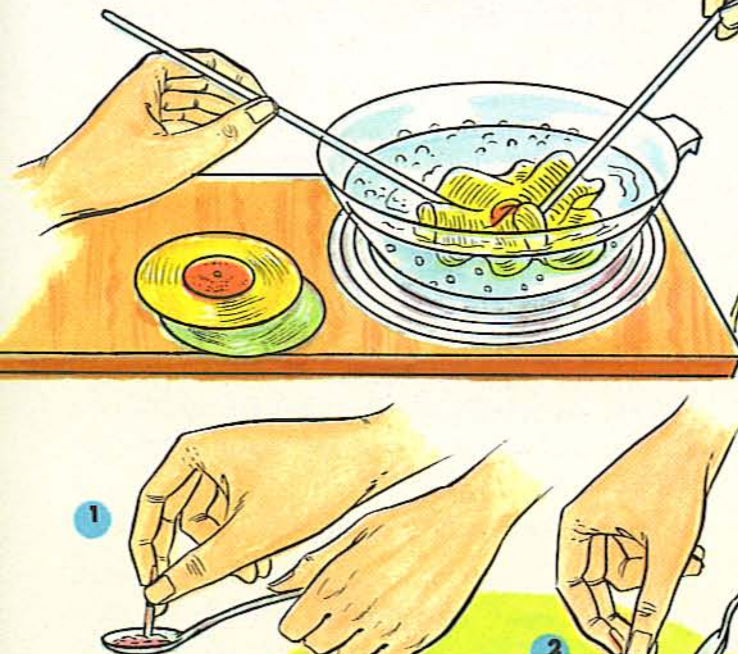

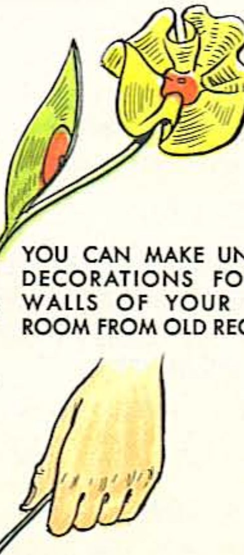

YOU CAN MAKE UNUSUAL DECORATIONS FOR THE WALLS OF YOUR GAME ROOM FROM OLD RECORDS.

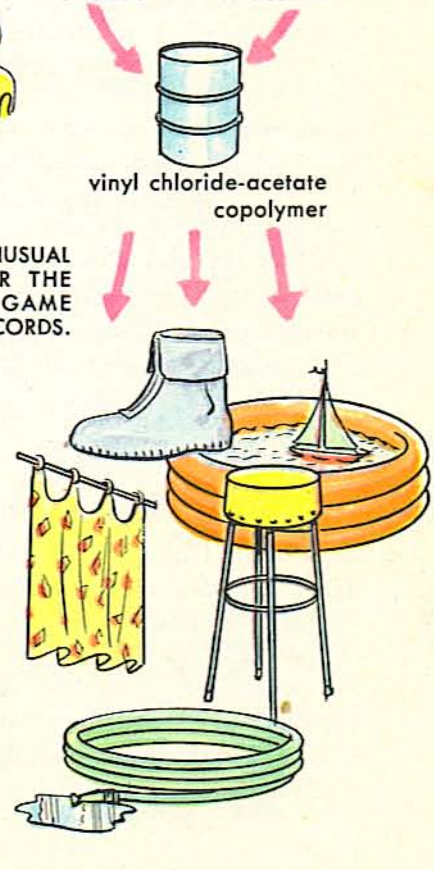

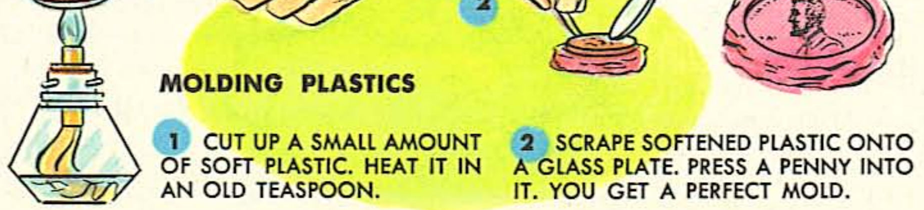

MOLDING PLASTICS

1 CUT UP A SMALL AMOUNT OF SOFT PLASTIC. HEAT IT IN AN OLD TEASPOON.

2 SCRAPE SOFTENED PLASTIC ONTO A GLASS PLATE. PRESS A PENNY INTO IT. YOU GET A PERFECT MOLD.

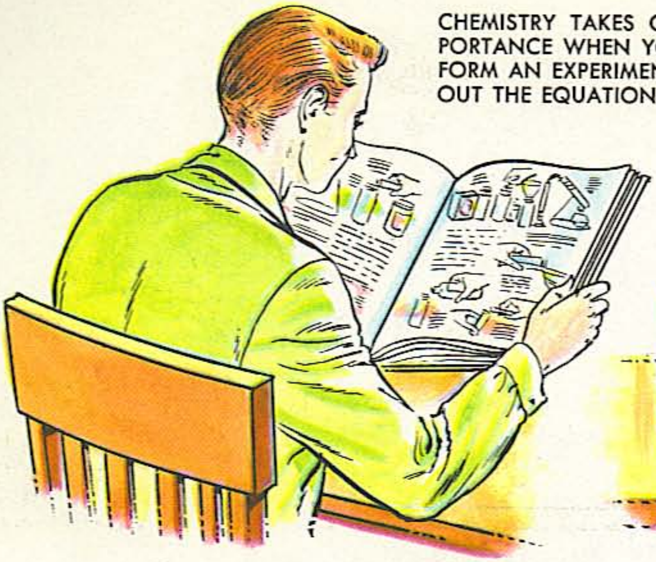

CHEMISTRY TAKES ON A GREATER IMPORTANCE WHEN YOU NOT ONLY PERFORM AN EXPERIMENT BUT ALSO WORK OUT THE EQUATION OF THE REACTION.

Working out Chemical Equations

You have done a great number of experiments by now. You have worked with gases, liquids, and solids. You have precipitated and decanted, filtered and distilled. As you think back over the experiments you will discover that they fall into four main groups of chemical reactions.

The simplest of these reactions is the DIRECT COMBINATION. In this, two or more substances combine to form a single more complex substance, as when iron and sulfur form iron sulfide:

$$Fe + S \rightarrow FeS$$

or when quicklime (calcium oxide) reacts with water to make slaked lime (calcium hydroxide):

$$CaO + H_2O \rightarrow Ca(OH)_2$$

DECOMPOSITION is the opposite of chemical combination. In this, a substance is broken down into simpler substances. This was the case when you separated the two elements found in water:

$$H_2O\ H_2O \rightarrow 2H_2\uparrow + O_2\uparrow$$

or when you made oxygen from hydrogen peroxide:

$$H_2O_2\ H_2O_2 \rightarrow 2H_2O + O_2\uparrow$$

In a SINGLE DISPLACEMENT, one element takes the place of another in a compound, as when you made hydrogen from zinc and hydrochloric acid:

$$Zn + H\,Cl\ H\,Cl \rightarrow H_2\uparrow + ZnCl_2$$

or when you set copper free by dropping a nail in a solution of copper sulfate:

$$Fe + Cu\,SO_4 \rightarrow Cu\downarrow + FeSO_4$$

In a DOUBLE DISPLACEMENT, the two compounds change partners with each other. Think of the time when you precipitated silver chloride from solutions of salt and silver nitrate:

$$Na\,Cl + Ag\,NO_3 \rightarrow AgCl\downarrow + NaNO_3$$

or when you mixed Epsom salt and washing soda:

$$Mg\,SO_4 + Na_2\,CO_3 \rightarrow MgCO_3\downarrow + Na_2SO_4$$

In studying the chemical shorthand above, you notice that, in every instance, there is an equal number of atoms of each element on either side of the arrow that indicates that a reaction takes place. Because of this equal arrangement, these chemical descriptions are called equations.

Many of these equations are scattered throughout this book. Many more are found in advanced chemistry textbooks. But very often, a chemist has to work out an equation from scratch.

Let's say you want to figure out the equation for dissolving aluminum foil in hydrochloric acid. You write out a trial equation:

$$Al + HCl \rightarrow AlCl + H\uparrow$$

But is AlCl correct? Look at the valence chart on page 75. Aluminum has three valence bonds, chlorine only one. One Al atom therefore takes on three Cl atoms, and aluminum chloride must be $AlCl_3$. H isn't right, either. Hydrogen exists in the free state only in molecules containing two atoms (H_2). So you change the equation to this:

$$Al + HCl \rightarrow AlCl_3 + H_2\uparrow$$

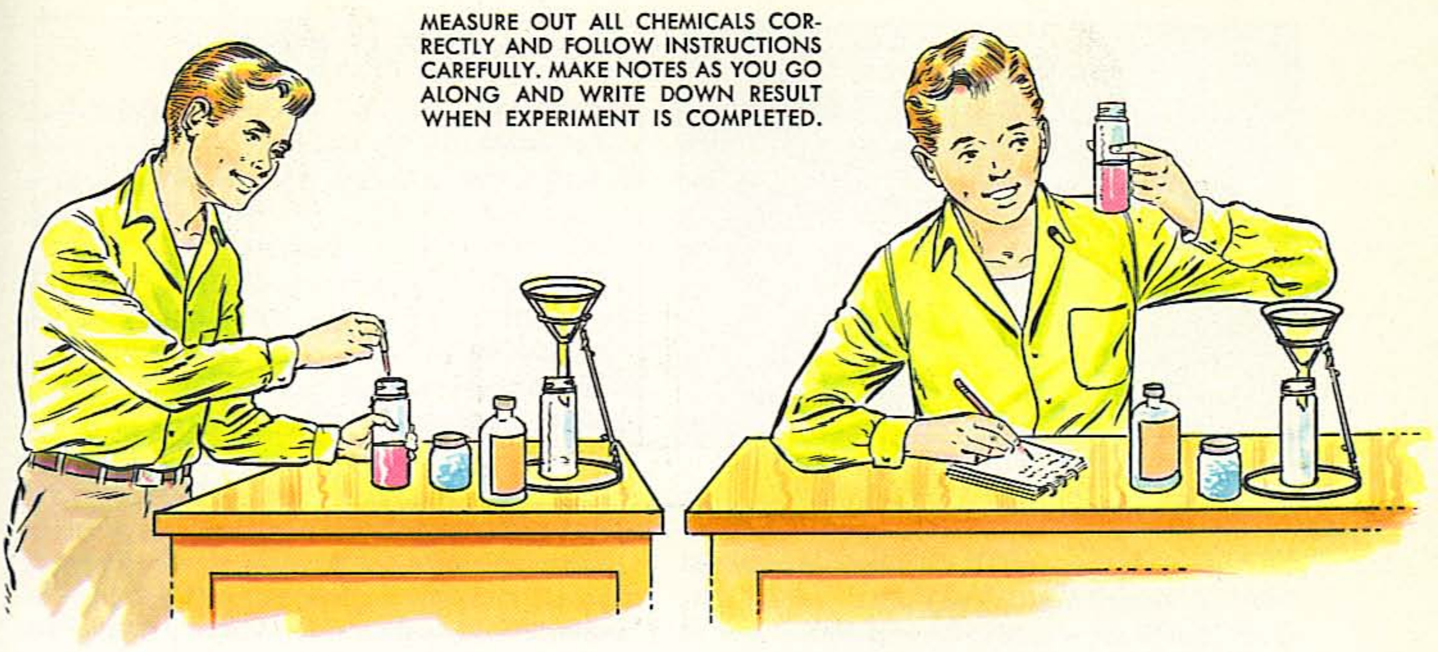

MEASURE OUT ALL CHEMICALS CORRECTLY AND FOLLOW INSTRUCTIONS CAREFULLY. MAKE NOTES AS YOU GO ALONG AND WRITE DOWN RESULT WHEN EXPERIMENT IS COMPLETED.

Now you need an amount of HCl that will give you Cl by the 3's and H by the 2's. 6HCl will do this. So you write in 6HCl and change the rest until the equation balances:

$$2Al + 6HCl \rightarrow 2AlCl_3 + 3H_2 \uparrow$$

Equations tell you what happens — but they tell far more than that.

Take the simple equation:

$$Fe + S \rightarrow FeS$$

APPROXIMATE ATOMIC WEIGHTS FOR CALCULATIONS					
Element	Symbol	Atomic Weight	Element	Symbol	Atomic Weight
ALUMINUM	Al	27	MAGNESIUM	Mg	24
BORON	B	11	MANGANESE	Mn	55
CALCIUM	Ca	40	NITROGEN	N	14
CARBON	C	12	OXYGEN	O	16
CHLORINE	Cl	36	POTASSIUM	K	39
COPPER	Cu	64	SILICON	Si	28
HYDROGEN	H	1	SILVER	Ag	108
IODINE	I	127	SODIUM	Na	23
IRON	Fe	56	SULFUR	S	32
LEAD	Pb	207	ZINC	Zn	65

This not only tells you that iron and sulfur make iron sulfide but also that it takes one iron atom and one sulfur atom to produce one molecule of FeS. Further, by inserting the atomic weights for the two elements from the chart on page 107, the equation tells you how much iron and sulfur are needed and how much iron sulfide you should get:

$$\begin{array}{llll} Fe & + S & \rightarrow & FeS \\ 56 & 32 & & 56 + 32 = 88 \end{array}$$

You can use the atomic weight numerals to indicate numbers of grams or any other unit of weight. By dividing by 16 you get the number of grams you used for experiment on page 22.

Now take a more complicated equation.

Let's say you want to produce magnesium carbonate. The chart of solubilities on page 108 tells you that $MgCO_3$ is insoluble. You should therefore be able to precipitate it from a soluble magnesium salt — the sulfate, for instance — and soluble sodium carbonate:

$$MgSO_4 + Na_2CO_3 \rightarrow MgCO_3 \downarrow + Na_2SO_4$$

Now you need to know how much $MgSO_4$ and how much Na_2CO_3 you need, and how much $MgCO_3$ you will get.

Before you start figuring from the equation above, check the chart on page 108, top right. Here you will discover that each molecule of magnesium sulfate has seven molecules of water of hydration ($7H_2O$) attached to it, and each sodium carbonate molecule,

(CONTINUED ON PAGE 108)

SOLUBILITY OF SALTS AND HYDROXIDES

NITRATES—**SOLUBLE**—WITHOUT EXCEPTIONS.

ACETATES—**SOLUBLE**—WITHOUT EXCEPTIONS.

CHLORIDES—**SOLUBLE**—EXCEPT Ag, Hg (MERCUROUS), AND Pb.

SULFATES—**SOLUBLE**—EXCEPT Pb, Ba, Sr (Ca, Ag AND Hg SLIGHTLY SOLUBLE).

NORMAL CARBONATES, PHOSPHATES, SILICATES, SULFIDES—**INSOLUBLE**—EXCEPT Na, K, NH_4.

HYDROXIDES—**INSOLUBLE**—EXCEPT Na, K, NH_4, Ba. (Ca AND Sr SLIGHTLY SOLUBLE.)

WATER OF HYDRATION
(WATER OF CRYSTALLIZATION)

- $AgNO_3$
- $CaCl_2 \cdot 6H_2O$
- $(CaSO_4)_2 \cdot H_2O$
- $CuSO_4 \cdot 5H_2O$
- $FeCl_3 \cdot 6H_2O$
- $FeCl_2 \cdot 4H_2O$
- $FeSO_4 \cdot 7H_2O$
- $KAl(SO_4)_2 \cdot 12H_2O$
- KNO_3
- $MgSO_4 \cdot 7H_2O$
- $NH_4Al(SO_4)_2 \cdot 12H_2O$
- NH_4Cl
- $Na_2B_4O_7 \cdot 10H_2O$
- $Na_2CO_3 \cdot 10H_2O$
- $NaCl$
- $NaHCO_3$
- $NaHSO_4 \cdot H_2O$
- $NaOH$
- $Na_2SO_4 \cdot 10H_2O$
- $Na_2S_2O_3 \cdot 5H_2O$

Equations—Continued

ten molecules of water ($10H_2O$). These do not enter into the chemical reaction — but you have to include them in the weight of the chemicals.

Write the atomic weight below each element. Then figure the molecular weight of each compound by adding the atomic weights of all the atoms found in the molecule.

This is what you get:

Mg	**S**	**O_4**	**$\bullet\, 7H_2O + Na_2$**		**C**	**O_3**	**$\bullet 10H_2O \rightarrow$**
24	32	16x4	7x18	23x2	12	16x3	10x18
24+32+		64 +	126	46 +	12 +	48 +	180
		246				286	

Mg	**C**	**$O_3 \downarrow +$**	**Na_2**	**S**	**O_4**	**$17H_2O$**
24	12	16x3	23x2	32	16x4	17x18
24+12+		48	46 +	32+	64	306
84				142		306

(When a formula contains subscripts — the small numerals that indicate how many of a kind — be certain to multiply the atomic weight by the number indicated by the subscript. In cases where the formula is preceded by a large number, be sure to multiply the molecular weight by this number.)

Your finished calculation tells you that 246 g (or 24.6 g or 2.46 g) of magnesium sulfate crystals and 286 g (or 28.6 g or 2.86 g) of sodium carbonate crystals will give you 84 g (or 8.4 g or .84 g) magnesium carbonate.

When you get even deeper into chemical mathematics you will be able to figure out the percentage of elements in a compound for which you know the formula, or the formula of a compound when you know the percentage of elements, or the numbers of liters of a gas you prepare in a chemical reaction.

CHEMISTRY AS A HOBBY WILL GIVE YOU MANY HOURS OF ENJOYMENT.

CHEMISTRY AS A SCIENCE TRAINS YOUR REASONING AND OBSERVATION.

What's Ahead in Chemistry?

THE CHEMICAL WONDERS of today are amazing enough — but they are like nothing compared to the wonders the future holds in store for the welfare of all humanity.

FOOD — The fertilizers of tomorrow will greatly increase the crops grown on farms throughout the world. Insect and disease-destroying chemicals will make cattle and poultry healthier and better producers of meat, milk, and eggs. Chemicals unknown today will make it possible to keep food fresh without refrigeration in any climate.

HOMES — The houses of the future will be built of more durable materials than any we have today. Floors and wall covering will last almost indefinitely. New paints will add never-fading colors.

CLOTHING — Many more man-made fibers will be added to those we use today — fibers with longer wear; fabrics that are cool in summer, warm in winter, easy to keep clean.

HEALTH — The miracle drugs of today have wiped out diseases that ranked among our greatest killers just a few years ago. In years to come many more diseases will disappear from the surface of the world under the onslaught of still more effective drugs created in the chemical laboratory.

TRAVEL — Much of the travel of the future will be at supersonic speeds. Planes and rockets will require materials that can stand tremendous heat and new fuels capable of producing enormous energy. Chemistry will provide them.

ATOMIC ENERGY — The force hidden in the atom will be turned into light and heat and power for everyday uses. Chemists of the future, working with their brother-scientists, the physicists, will find new ways of harnessing and using the atoms of numerous elements — some of them unknown to the scientists of today.

Do you want to share in the making of that astonishing and promising future?

If you have enjoyed performing the experiments in this book, figuring out formulas and equations, jotting down observations, you are the kind of person who has the qualifications for making a successful career in chemistry.

If you care to look further into the matter, speak to your science teacher about it and drop a line to one or all three of the organizations mentioned below and ask for their pamphlets on becoming a chemist:

American Chemical Society,
1155 16th Street, N. W., Washington 6, D. C.

American Institute of Chemical Engineers,
25 West 45th Street, New York 36, N. Y.

Manufacturing Chemists' Association,
1625 I Street, N. W., Washington 6, D. C.

But whatever you decide for the future, keep up your interest in chemistry as a hobby. In addition to giving you fun and enjoyment, your chemical hobby will sharpen your powers of observation and reasoning and train your mind for whatever occupation you decide upon for a lifework.

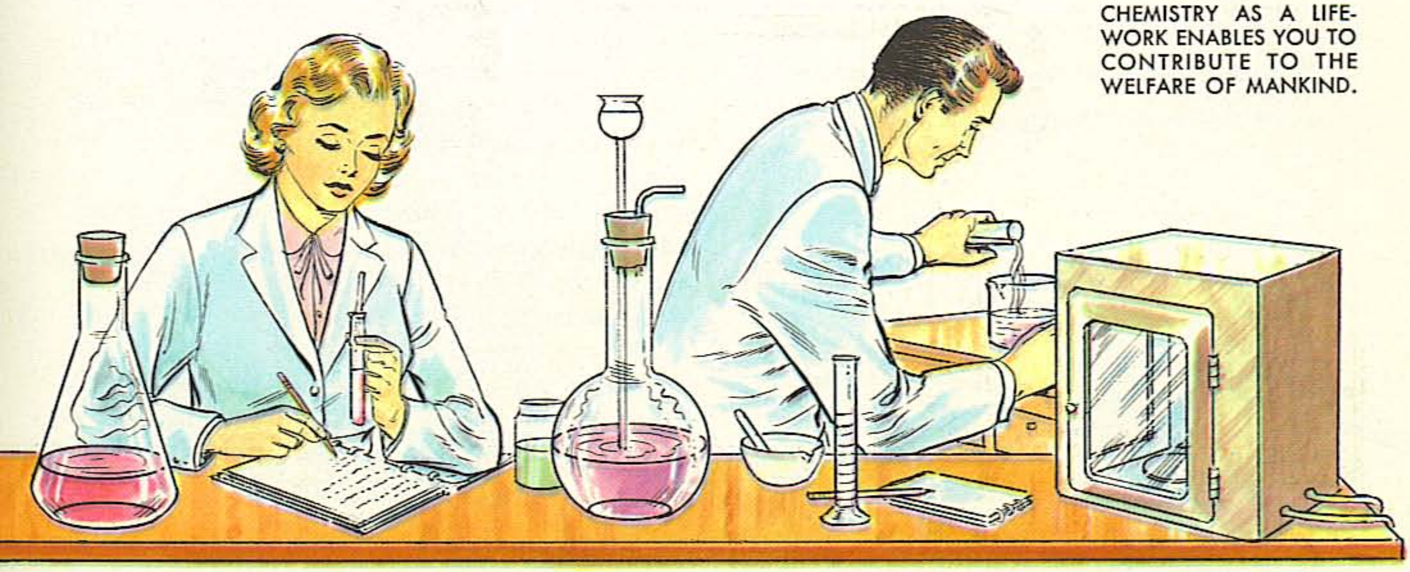

CHEMISTRY AS A LIFEWORK ENABLES YOU TO CONTRIBUTE TO THE WELFARE OF MANKIND.

Where to Get Chemicals and Equipment

A GREAT MANY of the experiments in this book can be performed with equipment found around the house: water glasses, custard cups, jars, bottles, cans, and funnel. For the rest, the following pieces of regular chemical laboratory equipment are needed:

6 test tubes, regular, 150 mm x 16 mm
3 test tubes, Pyrex, 150 mm x 16 mm
1 test tube brush, small
3 wide-mouth bottles, 4 ozs.
6 ft. glass tubing, 6 mm outside diameter
3 ft. rubber tubing, 3⁄16″ inside diameter
2 No. 0 rubber stoppers, one hole
1 No. 5 rubber stopper, one hole
3 No. 5 rubber stoppers, two holes
1 triangular file, 4″
1 glass stirring rod, 5″
1 pkg. filter paper, 12.5 cm, 50 pieces
1 vial litmus paper strips, blue
1 vial litmus paper strips, red

If you can not secure this equipment locally, write to one of the companies below asking for price list or catalog, including cost of catalog where called for. When you receive the answer, mail your order and the correct amount by bank check or money order.

Science Mail Co., 17-33 Murray St., Whitestone 57, N. Y. (Price list free)
Winn Chemical Co., 124 West 23rd St., New York 11, N. Y. (Catalog 25¢)
N. Y. Scientific Supply Co., 28 West 30th St., New York 1, N. Y. (Catalog 50¢)
Home Lab Supply, 511 Homestead Ave., Mount Vernon, N. Y. (Price list free)
Biological Supply Co., 1176 Mt. Hope Ave., Rochester 20, N. Y. (Catalog 25¢)
A. C. Gilbert Co., P. O. Box 1610, New Haven 6, Conn. (Price list free)
Bio-Chemical Products, 30 Somerset St., Belmont, Mass. (Catalog 25¢)
Laboratory Sales, P. O. Box 161, Brighton, Mass. (Catalog 25¢)
The Porter Chemical Co., Hagerstown, Md. (Price list free)
Tracey Scientific Laboratories, P. O. Box 615, Evanston, Ill. (Catalog 25¢)
National Scientific Co., 13 South Park Ave., Lombard, Ill. (Catalog 35¢)
Hagenow Laboratories, Manitowoc, Wis. (Catalog 20¢)

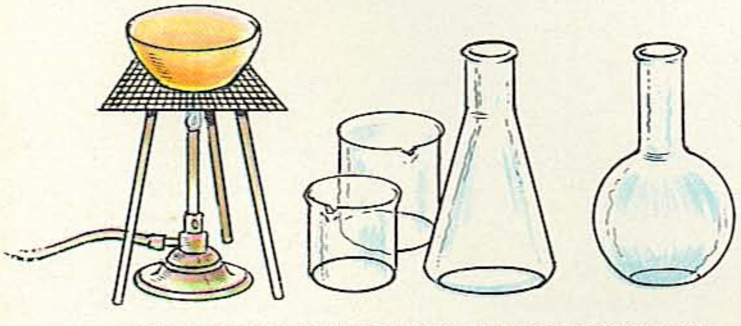

IF YOU DECIDE TO USE REGULAR LABORATORY WARE IN YOUR HOME LAB, GET PRICE LIST FROM SUPPLIER.

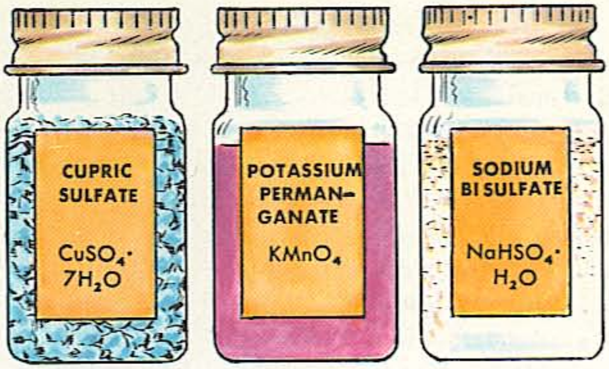

CHEMICALS FOR HOME EXPERIMENTS ARE AVAILABLE IN JARS OF UNIFORM SIZE, ALL PROPERLY LABELED.

WHENEVER YOU NEED a chemical for one of the experiments described in this book, check the list of common chemicals on page 111 to find out where to buy it.

All of these chemicals are, of course, available through chemical supply houses. The trouble is that many of these houses do not sell to individuals but only to schools and established laboratories. Also, the chemicals usually come in a standard amount of ¼ lb. — or even 1 lb. — where, in home experiments, you would need 1 ounce or less. The same is often the case when you buy chemicals in a local store. The minimum-sized packages or jars may be so large that you couldn't possibly use up the contents in a year of experiments. You will probably also have to repack what you actually need into glass jars of suitable size for efficiency and to fit your storage space.

Because of this and the inconvenience of having to shop around, you may find it advantageous to buy your chemicals by the kit, in uniform-sized screw-top glass containers. Such kits are available in the science department of many hobby and model supply stores.

Chem-Kit No. 1 contains the ten chemicals marked 🟥 on the opposite page. Chem-Kit No. 2 contains the ten chemicals marked 🟨. The kits contain sufficient amounts of chemicals to perform each experiment many times over.

You can also make up your own set of chemicals in amounts suitable for home experiments by getting them from one of the companies listed to the left. Be certain to add the cost of the catalog when you write for one and to send the correct amount when you order.

Common Chemicals and Their Formulas

	CHEMICAL NAME	FORMULA	COMMON NAME	WHERE TO BUY
🟦	ACETIC ACID	$CH_3COOH + H_2O$	5% solution: white vinegar	Grocery
🟥	AMMONIUM CHLORIDE	NH_4Cl	sal ammoniac	Drug store
🟦	AMMONIUM HYDROXIDE	$NH_4OH + H_2O$	10% solution: household ammonia 27% solution: strong ammonia	Grocery Drug store
🟦	BORIC ACID	H_3BO_3	boric acid	Drug store
🟥	CALCIUM CARBONATE	$CaCO_3$	chunks: marble, limestone powder: precipitated chalk	Builders' supplies Drug store
🟦	CALCIUM HYDROXIDE	$Ca(OH)_2$	slaked lime, garden lime	Hardware store
🟨	CALCIUM OXIDE	CaO	quicklime	Builders' supplies
🟦	CALCIUM SULFATE	$(CaSO_4) \cdot H_2O$ $CaSO_4 \cdot 2H_2O$	plaster of Paris gypsum	Hardware store Chemical supplies
🟦	CARBON TETRACHLORIDE	CCl_4	carbon tet	Hardware store
🟨	COPPER SULFATE	$CuSO_4 \cdot 5H_2O$	blue vitriol	Drug store
🟨	FERROUS SULFATE	$FeSO_4 \cdot 7H_2O$	iron sulfate, green vitriol, copperas	Drug store
🟦	GLUCOSE	$C_6H_{12}O_6 + H_2O$	solution: corn syrup	Grocery
🟦	HYDROCHLORIC ACID	$HCl + H_2O$	25% solution: muriatic acid	Hardware store
🟦	HYDROGEN PEROXIDE	$H_2O_2 + H_2O$	3% solution: peroxide	Drug store
🟥	IRON, METAL, POWDER	Fe	powdered iron	Chemical supplies
🟨	MAGNESIUM, METAL	Mg	magnesium ribbon	Chemical supplies
🟦	MAGNESIUM SULFATE	$MgSO_4 \cdot 7H_2O$	Epsom salts	Drug store
🟥	MANGANESE DIOXIDE	MnO_2	pyrolusite	Hardware store (flashlight battery)
🟦	NAPHTHALENE	$C_{10}H_8$	moth balls	Hardware store
🟥	PHENOLPHTHALEIN	$C_6H_4COOC(C_6H_4OH)_2$	phenolphthalein	Drug store
🟨	POTASSIUM ALUMINUM SULFATE	$KAl(SO_4)_2 \cdot 12H_2O$	alum, potassium alum	Drug store
🟨	POTASSIUM FERROCYANIDE	$K_4Fe(CN)_6 \cdot 3H_2O$	potassium ferrocyanide	Chemical supplies
🟥	POTASSIUM IODIDE	KI	potassium iodide	Drug store
🟦	POTASSIUM NITRATE	KNO_3	saltpeter, niter	Drug store
🟨	POTASSIUM PERMANGANATE	$KMnO_4$	potassium permanganate	Drug store
🟨	SALICYLIC ACID	$C_6H_4OHCOOH$	salicylic acid	Drug store
🟦	SILVER NITRATE	$AgNO_3$	lunar caustic	Drug store
🟦	SODIUM BICARBONATE	$NaHCO_3$	baking soda, bicarb	Grocery
🟥	SODIUM BISULFATE	$NaHSO_4 \cdot H_2O$	82% of Sani-Flush®	Grocery
🟦	SODIUM CARBONATE	$Na_2CO_3 \cdot 10H_2O$ $Na_2CO_3 \cdot H_2O$	sal soda, crystal washing soda concentrated washing soda	Grocery Grocery
🟦	SODIUM CHLORIDE	$NaCl$	salt, table salt	Grocery
🟦	SODIUM HYDROXIDE	$NaOH$	lye, caustic soda, Drano®	Grocery
🟦	SODIUM HYPOCHLORITE	$NaClO + H_2O$	5% solution: laundry bleach, Clorox®	Grocery
🟨	SODIUM POTASSIUM TARTRATE	$NaKC_4H_4O_6 \cdot 4H_2O$	Rochelle salt	Drug store
🟦	SODIUM SILICATE	$Na_2SiO_3 + H_2O$	solution: water glass	Hardware store
🟨	SODIUM TETRABORATE	$Na_2B_4O_7 \cdot 10H_2O$	borax	Drug store
🟥	SODIUM THIOSULFATE	$Na_2S_2O_3 \cdot 5H_2O$	hypo	Photo store
🟦	SUCROSE	$C_{12}H_{22}O_{11}$	cane sugar	Grocery
🟥	SULFUR	S	powder: flowers of sulfur block: sulfur candle	Drug store Hardware store
🟥	ZINC, METAL	Zn	zinc	Hardware store (flashlight battery)
🟦	ZINC CHLORIDE	$ZnCl_2 + H_2O$	tinners' fluid	Hardware store

Note: Chemicals marked 🟦—many of them liquids—are most easily secured in local stores. Chemicals marked 🟥 are found in Chem-Kit No. 1, chemicals marked 🟨 in Chem-Kit No. 2 (see opposite page).

Index

Acetic acid, 90, 91
Acids, 23, 42, 43; carboxylic, 90; fatty, 92, 93; household items containing, 44; making, 44; test for, 43, 44; traits of, 42
Acrolein, 93
Albumin, 97, 99
Alchemists, 6, 7, 36
Alcohols, 79, 88-89
Alum, 64, 65
Aluminum, 64-65; chloride, 64; foil, 72; hydroxide, 65; sulfate, 64, 65
Ammonia, 32, 33, 43; fountain, 33; making, 33; solubility of, 33; uses of, 32
Ammonium, alum, 64; chloride, 33; cyanate, 78; hydroxide, 33, 43
Apparatus, how to make, 12, 16
Aristotle, 6
Arrhenius, Svante, 40
Atmosphere, 27
Atomic energy, 5
Atomic weights, 38, 39, 107
Atoms, 36, 37, 38, 39

Baekeland, Leo H., 104
Balance, hand, 15
Balloons, 28
Bases, 23; household items containing, 45; test for, 43, 45; traits of, 43
Benzene, 80, 81
Berzelius, Jons, 36
Bessemer, Henry, 69
Borax, 57; bead test, 56
Boric acid, 57; test for, 57
Boron, 56
Boyle, Robert, 6, 7
Bronze, 70

Calcium, 60-61; bicarbonate, 60; carbonate, 45, 60, 61; chloride, 47, 61; hydroxide, 45, 60; oxide, 45, 60; sulfate, 60, 61
Candle, 18-19; contents of, 18
Carbohydrates, 79, 84-87
Carbon, 76; atom, 37; compounds, 23, 76-77, 78, 80-81; forms of, 77; test for, 77
Carbonate, calcium, 45, 60, 61; copper, 71; cupric, 71; ferric, 69; iron, 69; magnesium, 63; manganese, 67; potassium, 59; sodium, 58; zinc, 63
Carbon dioxide, 30-31; cycle of, 30; making, 30, 31; test for, 31; uses of, 30
Carboxylic acids, 79, 90-91
Casein, 98, 99; glue from, 98; making of, 98
Cavendish, Henry, 28
Cellulose, 86
Charles, Jacques, 28
Cheese, 99
Chemical, common names, 111; formulas, 111; where to buy, 110, 111
Chemistry, careers in, 109; future of, 109; importance, 4; what it is, 4
Chemists, 6
Chloride, aluminum, 64; ammonium, 33; calcium, 47, 61; copper, 71; cupric, 71; cuprous, 71; ferric, 68; ferrous, 68; iron, 35, 68; magnesium, 62; manganese, 67; silver, 72; sodium, 41, 58; zinc, 28, 47, 62
Chlorine, 34-35; bleaching with, 35; compounds, 34, 35; making, 35; test for, 34
Chloroform, 89
Coal age, 76; mining, 76
Coagulation, 96, 101
Colloidal dispersion, 100
Colloids, 23, 100-101; light test for, 101
Compounds, 22, 23
Copper, 70-71; carbonate, 71; chloride, 71; hydroxide, 71; replacement of, 62, 71; salicylate, 91; sulfate, 62, 71; sulfide, 53, 71
Copperas, 69
Crystallization, 21, 58
Crystals, 41, 64
Cuprammonium, 103
Cupric salts, 70, 71
Cuprous salts, 70, 71
Curie, Marie, 7, 20
Curie, Pierre, 7, 20

Dalton, John, 36, 37
Davy, Humphry, 7, 59, 60
Decantation, 20
Democritus, 6
Detergents, 95
Dispersions, colloidal, 100-101
Distillation, destructive, 77; of ethanol, 89; of water, 61

Eggs, protein in, 96-97
Electrolysis of water, 25
Elements, 22, 23, 38-39
Empedocles, 6
Emulsification, 101
Equations, chemical, 106-108
Equipment, laboratory, 8; improvised, 9, 11; where to buy, 110
Esters, 79, 92
Ethanol, 88, 89
Evaporation, 21

Faraday, Joseph, 18
Fats, 92-93; extracting, 93; test for, 93
Fehling solution, 85, 87, 98
Ferric salts, 68, 69
Ferrous salts, 68, 69
Fibers, 102-103; tests for, 102, 103
Filtration, 20, 21
Fire extinguisher, 30
Formulas, 74-75, 106; carbon compounds, 80-81; of common chemicals, 111
Fractionating of oil, 82
Frasch, Herman, 50
Fructose, 84

Gasoline, 82
Gelatin, 99
Glass tubes, bending, 13; cutting, 13; glazing, 13
Glycerol, 92, 93
Glossary, 2
Glucose, 84, 85
Graham, Thomas, 100

Hall, Charles, 64, 65
Hydrocarbons, 79, 82-83
Hydrochloric acid, 34, 44
Hydrogen, 28-29; making, 28, 29; safety with, 29; test for, 25, 28
Hydrogen iodide, 49
Hydrogen peroxide, 26, 27, 67
Hydrogen sulfide, 52, 53; in chemical analysis, 53; making, 53
Hydroxide, 43; aluminum, 65; ammonium, 33, 43; calcium, 60, 61; copper, 71; cupric, 71; ferric, 69; ferrous, 69; iron, 69; magnesium, 63; manganese, 66; potassium, 94; sodium, 43, 94; solubility of, 108; zinc, 63
Hypo, 49, 51, 52, 73

Indicators, color table, 44, 45; homemade, 42; types of laboratory, 43
Iodide, potassium, 34, 48, 49, 87
Iodine, 48-49; making, 48, 49; removing, 49; solubility of, 48; test for, 49; tincture of, 48
Iodoform, 89
Iron, 68-69; carbonate, 69; chloride, 35, 68; hydroxide, 69; oxide, 68; salicylate, 91; sulfate, 69; sulfide, 53; tannate, 91

Kalium, 59
Kékulé, August, 80, 81
Kitchen as laboratory, 4

Laboratory, safety, 16; setting up, 10-11; techniques, 16-17
Lavoisier, Antoine, 6
Lime, 43, 60
Lime water, 31
Lye, 43

Magnesium, 62-63; carbonate, 47, 63; hydroxide, 63; sulfate, 41, 47, 62, 63
Manganese, 66-67; carbonate, 67; chloride, 67; dioxide, 25, 66; hydroxide, 66; sulfate, 66, 67; sulfide, 53, 67
Measurements, 14
Mendeleeff, Dmitri, 39
Metalloids, 23
Metals, 23 (see also individual metals); tests for, 53, 56
Methane, 80, 81, 83
Methanol, 88, 89
Metric system, 14
Milk, protein in, 98
Minerals, 54
Mixtures, 22, 23
Moseley, Henry, 39
Muriatic acid, see Hydrochloric acid

Naphthalene, 83
Natrium, 58
Neutralization, 46
Nitrate, potassium, 32, 41, 59; silver, 72
Nitric acid, 42
Nitrogen, 32-33; in atmosphere, 32
Non-metals, 23

Oil, cracking of, 83; crude, 82; fractionating of, 82
Oils, 92-93
Olein, 92
Organic compounds, see Carbon compounds
Orsted, Hans C., 64
Oxide, boric, 57; calcium, 60; magnesium, 62; silver, 72; zinc, 62
Oxygen, 26-27; in atmosphere, 27; making, 27; test for, 25

Paracelsus, 6, 7
Peptization, 101
Periodic table, 38-39
Permanganate of potassium, 66, 67
Petroleum, 82
Phenol, 91
Photography, 73; without a camera, 73
Photosynthesis, 86
Plaster of Paris, 61
Plastics, 104-105; tests for, 104
Potash, 59, 94
Potassium, 58, 59; alum, 64; carbonate, 59; ferrocyanide, 68; flame test for, 59; hydroxide, 94; iodide, 34, 48, 49, 87; nitrate, 32, 41, 59, 61; nitrite, 59; permanganate, 66, 67
Priestley, Joseph, 7, 26
Proteins, 79, 96-99; in egg, 96, 97; in foods, 96, 97; in milk, 98; tests for, 97, 99
Proust, Joseph, 37
Prussian blue, 68
Pyrolusite, 66

Radium, 20, 21
Rayon, 102; making, 103
Rochelle salt, 85
Rubber, 82
Rust, 68

Safety, 16
Sal ammoniac, 43
Salicylic acid, 90, 91
Salt, table, see Sodium chloride
Saltpeter, 32, 41, 59, 61
Salts, 23, 46-47; household items containing, 47; making, 47; names of, 46; solubility of, 108
Scheele, Karl, 7, 26
Schweitzer's reagent, 103
Scientific method, 21
Silicic acid, 55
Silicon, 54-55
Silicones, 54
Silver, 72-73; bromide, 73; chloride, 72; nitrate, 72; oxide, 72; sulfide, 72
Soap, 94-95; how it acts, 94; making, 95; testing, 95
Soda ash, 59, 94
Sodium, 58; acetate, 91; aluminate, 64; bicarbonate, 58; bisulfate, 42, 58; carbonate, 58, 59, 72; chloride, 41, 58; flame test for, 59; hydroxide, 43, 45, 94, 95; hypochlorite, 34; salicylate, 91; silicate, 54, 55; sulfate, 58; sulfide, 53; tetraborate, 57; thiosulfate, 49, 51, 52, 73
Solutions, 20, 23, 40-41; behavior of, 41; conductivity of, 40; making, 41; saturated, 41
Starch, 86-87; making, 87; test for, 87
Stearin, 92
Steel, 69
Stoppers, rubber, 12
Sucrose, 84, 85
Sugars, 84-85; test for, 85
Sulfate, aluminum, 64; calcium, 60, 61; copper, 62, 71; ferric, 69; ferrous, 69; iron, 69; magnesium, 41, 47, 62, 63; manganese, 66, 67; sodium, 58; zinc, 62, 63
Sulfide, antimony, 53; cadmium, 53; copper, 53, 71; cupric, 71; hydrogen, 52, 53; iron, 22, 53; manganese, 53, 67; silver, 72; sodium, 53; zinc, 53, 63
Sulfur, 22, 50-51; casting with, 51; forms of, 50, 51; melting, 51; precipitated, 51; production of, 50
Sulfur dioxide, 52; making, 52
Sulfuric acid, 42
Sulfurous acid, 44, 52
Symbols, alchemists', 2, 6, 36; chemical, 36

Table salt, see Sodium chloride
Tannic acid, 90, 91
Tincture of iodine, 48
Titration, 46
Turpentine, 83

Urea, 78

Valences, 74-75; chart of, 75
Vitriol, green, 69
Vulcanization, 50

Washington Monument, 64
Water, 24-25; as a catalyst, 24; clearing, 65; composition of, 24, 26; distilling, 61; electrolysis of, 25; hardness of, 61, 95; as solvent, 24
Water of hydration, 108
Waterglass, 55
Wöhler, Friedrich, 7, 78

Zinc, 25, 62-63; carbonate, 63; chloride, 28; hydroxide, 63; sulfate, 62; sulfide, 53, 63

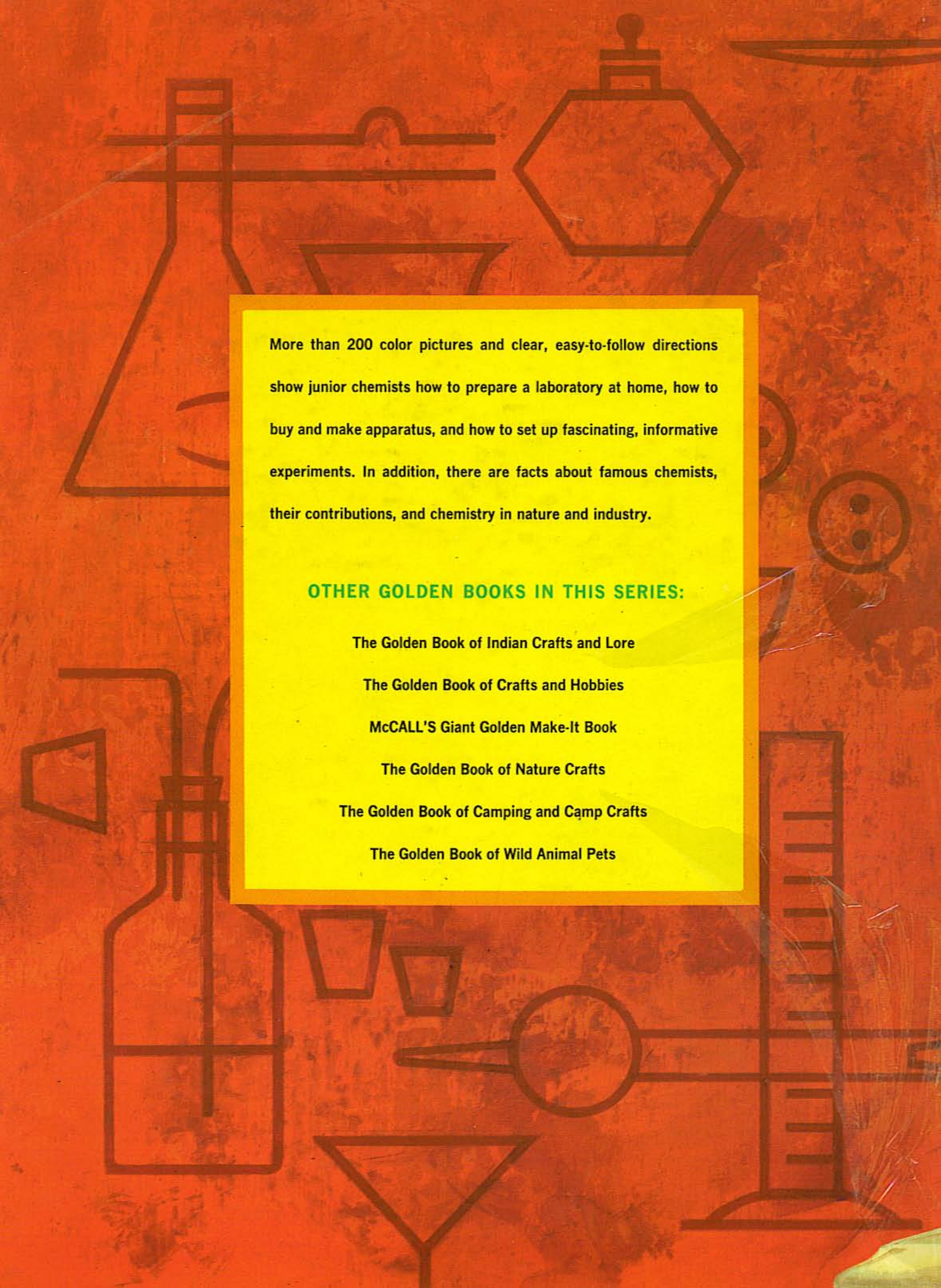

More than 200 color pictures and clear, easy-to-follow directions show junior chemists how to prepare a laboratory at home, how to buy and make apparatus, and how to set up fascinating, informative experiments. In addition, there are facts about famous chemists, their contributions, and chemistry in nature and industry.

OTHER GOLDEN BOOKS IN THIS SERIES:

The Golden Book of Indian Crafts and Lore

The Golden Book of Crafts and Hobbies

McCALL'S Giant Golden Make-It Book

The Golden Book of Nature Crafts

The Golden Book of Camping and Camp Crafts

The Golden Book of Wild Animal Pets

3 CRYSTAL SETS

The simplest type of radio receiver consists of an aerial, a tuned circuit and a detector–plus headphones to listen to the signals received by the tuned circuit. It is generally called a *crystal set* because it works on the same principle as the very earliest radio sets, using a lead-galena crystal and a 'cat's whisker' as a detector. The only real difference is that the crystal and cat's whisker have been replaced by the much more efficient point-contact *diode*.

A simple *tuned circuit* can be made by winding a coil of wire of specific diameter and length (number of turns) to match the capacitance range of a variable capacitor. The combination of coil (inductance) and capacitance can then be adjusted by turning the variable capacitor spindle to be resonant, or tune in to various broadcast frequencies. The signal which is tuned in is then passed to the diode which blocks the RF content, and rectifies and passes on the AF content to the headphones, which transform these signals into audible sounds.

The tuned circuit can be combined with the aerial by winding the coil on a *ferrite rod.* This increases the inductance of the coil (reducing the length of wire required), whilst the rod itself also acts as an aerial to pick up the radio signals. This is the type of aerial employed in most domestic radio receivers. However, with very simple sets working with very weak signals, an external aerial may also be required.

The performance of all simple, low-powered receivers is, in fact, very much dependent on the efficiency of the aerial and tuned circuit, which can be regarded as the critical part of the system. Performance can perhaps be improved by experimenting with different aerial/tuned-circuit systems, and by the addition of further circuitry, as explained in later chapters.

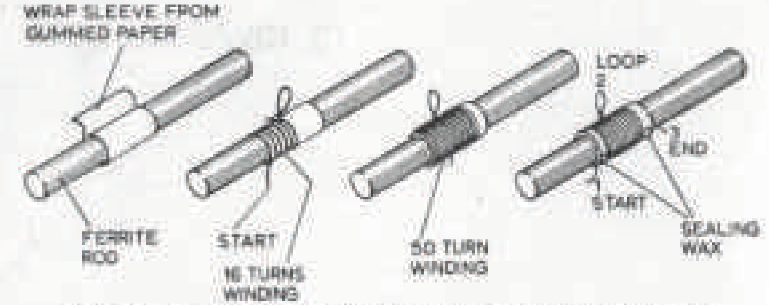

Fig. 3.1 Aerial tuning coil wound from 38-s.w.g. wire on a $\frac{3}{8}$-in. diameter ferrite rod.

Fig. 3.1 details the construction of a suitable coil for a tuned circuit, matched to a ferrite rod of $\frac{3}{8}$ in. diameter and about 4 in. long (see Chapter 4 for alternative coil windings on different rod sizes). If a rod is purchased longer than 4 in., it can be used as it is, or cut down by marking around with a file and then breaking off the surplus length.

Cut seven 1-in. lengths of gumstrip. Moisten one and wrap around the rod gummed side up. Now add about another half a dozen wrappings of similar length over the first, this time with gummed side *down* to form a reasonably rigid tube. Make sure that the paper tube is a *sliding* fit on the ferrite rod and leave to dry thoroughly (preferably removed from the rod so that it cannot become stuck to it).

When the paper tube is quite dry it should be rigid, when the coil windings can be applied. The wire to be used is 38-s.w.g. enamelled copper wire, the number 38 referring to the actual diameter of the wire according to the standard wire gauge (s.w.g.).*

Starting about $\frac{1}{2}$ in. in from one end of the paper tube, wind the wire carefully round the tube, with each turn tight against the one before it, until sixteen full turns have been completed. Then make a loop in the wire, as shown, and carry on winding, with succeeding turns touching, until fifty turns in all have been completed. The two loose ends of the coil (the start and finish) can be secured with a dab of sealing wax whilst the projecting loop can be twisted together (e.g. by putting a pencil through the loop and twisting up). Cut off the loop, leaving about $\frac{1}{2}$ in. protruding from the main coil, bare the wire ends and solder together. This forms point 2 on the coil; the start is point 1, and the end point 3–see Fig. 3.1. It will be easy to remember these

---

* American wire gauge (AWG) is about two sizes smaller than the British standard wire gauge (s.w.g.). For example, 38 s.w.g. is approximately the same as 36 AWG.

without marking since the loop or tapping point (2) comes much closer one end (1) than the other (3).

Cut a panel of *Paxolin sheet* to about the size shown in Fig. 3.2, using a hacksaw.* On this secure a *tag strip*, as shown, and drill a hole to mount a miniature or small-size 500 pF *variable capacitor*.

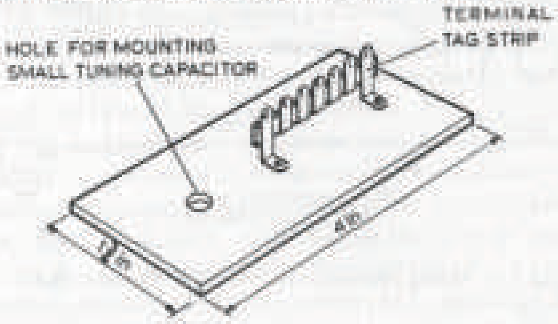

Fig. 3.2 Paxolin panel and tag strip for crystal set circuits.

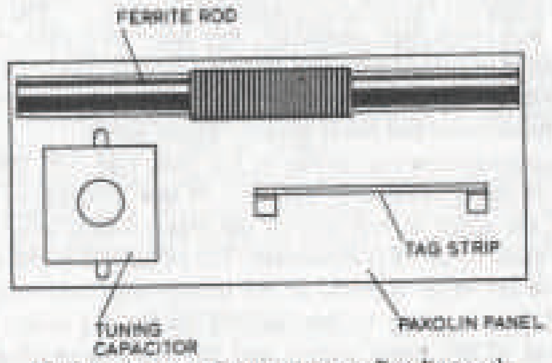

Fig. 3.3 Layout of components on Paxolin panel.

The aerial coil is then mounted on the panel as shown in Fig. 3.3, gluing the coil on to the Paxolin with two or three dabs of sealing wax, or some other suitable adhesive. *Note:* the ferrite rod must be free to slide in the paper tube for 'tuning' adjustments.

Virtually any miniature germanium or silicon diode will be suitable for the detector. Recommended types, which are readily available, are 1N34 and 1N914.

Earphones must be of *high-impedance* type, which need not be expensive to buy, and the higher the impedance the better the reception.

---

* A phenolic or fiberglass sheet may also be used.

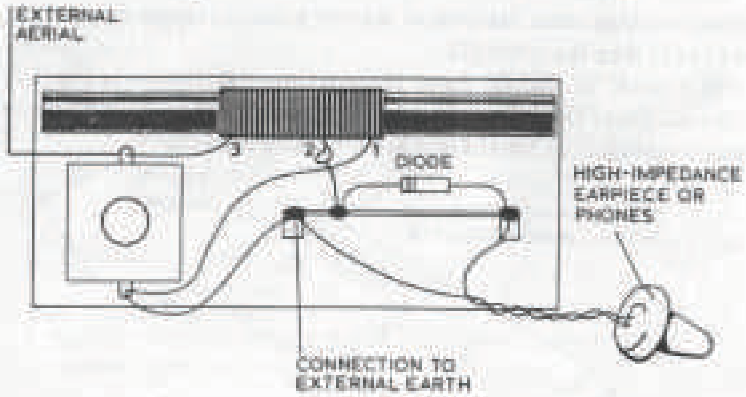
EXTERNAL AERIAL
3
2
1
DIODE
HIGH-IMPEDANCE EARPIECE OR PHONES
CONNECTION TO EXTERNAL EARTH

Fig. 3.4 Wiring connections to complete the crystal set.

Alternatively a deaf-aid type earpiece can be used; this will not give the same volume or quality of reproduction as headphones, but is a less expensive component. This should preferably be of the *high-impedance magnetic* type, with high sensitivity. Any high-impedance earpiece will suffice, but if of crystal type will require a resistor, connected across it, to complete the circuit. This will reduce the amount of current flowing through the earpiece and lower the strength of signal.

Wiring connections are shown in Fig. 3.4. End 3 of the aerial coil (the end of the 50-turn coil) connects to one terminal of the tuning capacitor, the aerial or 'hot' end of the tuned circuit, and the point to which an external aerial is connected. The other end of the coil (end 1) connects to the other terminal of the tuning capacitor, from which an additional wire is taken to the first tag on the tag strip. This is the 'earthy' end of the tuned circuit, and the point to which an external earth is connected. Leave plenty of slack wire between the coil and tuning capacitor.

The other connections are then as follows:

(i) Tapping point of the coil wire bared and connected to the second tag.
(ii) The diode also soldered to this same tag, and to any other free tag.
(iii) Headphone (or earpiece) leads to the 'earthing' tag, and to the 'free' tag to which the above diode has been connected. All connections should be made with *soldered* joints.

The wiring-up can be checked against the *circuit diagram* shown in Fig. 3.5 (ignoring the components shown with broken lines). The set should now be 'working'.

In areas of strong signal strength, no external aerial or earth connections should be necessary. Performance will, however, be improved in any area by attaching an aerial wire (which can be any thin wire, e.g. using the same wire as for the coil winding), of up to 160 feet. The longer the aerial the better the reception, provided it is led away from the receiver to as *high* a point as possible.

An earth connection may further improve aerial performance; by this we mean a connection to some conductor positively in contact with the ground (preferably buried), an excellent example being a metal water-pipe. Thus, if an earth connection is found to be necessary (or you want to try one to see how performance is affected), connect a wire from the 'earthing' tag on the receiver to a convenient water-pipe.

This question of obtaining a *good* aerial and earth is a most important one in areas of poor signal strength. Linking up to a television aerial is often a good plan, since TV aerials are also usually mounted as high as possible. If bare wire is used, it is also important that the upper (free) end of the aerial is not made fast to something which could produce an earth connection (e.g. a damp tree), or at least is suitably insulated from such a support. String is not an efficient insulator; that, too, can conduct when wet.

Quite good results are often obtained by using the springs of a bed as an aerial, in which case an earth connection is usually not necessary. Sometimes, too, when other attempts to yield a good signal strength in

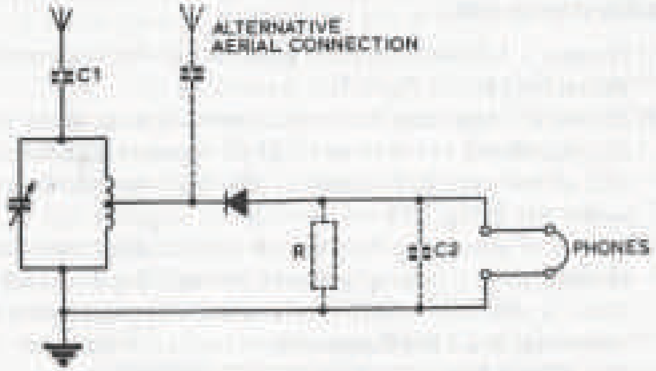

Fig. 3.5 *Circuit 1.* Circuit diagram of basic crystal set.

the aerial have failed, connecting the *aerial* side of the tuning coil to a *good earth* (a water-pipe) can produce better results, the normal earth connection being left off.

Tuning

The receiver is adjusted as follows. Turn the tuning capacitor to fully close the vanes, then open about half a turn on the spindle or knob. (If you are using a trimmer as a tuning capacitor, screw right down and then open half a turn.)

The tuning coil should now be slid up and down the ferrite rod (the coil leads were left a fairly long time to give the necessary freedom of movement) until BBC Radio 3 is heard.*

It may be necessary to slightly alter the adjustment of the variable condenser to tune in to this programme. Also, because of the ferrite-rod aerial the set will be directional, that is, the signal strength received will depend to some extent on the direction in which the aerial rod is pointing, so position the set to pick up the maximum volume.

Having established the best position of the tuning coil on the rod to receive BBC Radio 3, fix permanently with a dab of sealing wax. You should then find it possible to tune in to further stations by altering the setting of the variable capacitor—e.g. typically Radio 4 in about the middle of the capacitor travel and Radio 1 towards the other end.

Any reception you get will almost certainly be very weak and (unless you live close to a broadcast station) you can really feel satisfied if you get any station at all at audible strength. But it is surprising how, sometimes, even quite distant stations can be heard. Also you can often improve the reception and listening strength by quite simple modifications. Try these in order:

(i) Connect a 1,000-pF capacitor across the headphone (earpiece) connections (C2 in Fig. 3.5).

(ii) Instead of connecting the external aerial directly to the tuned circuit, connect one lead to a 220-pF capacitor, and the other end of the capacitor circuit to the 'hot' end of the tuned circuit (C1 in Fig. 3.5).

(iii) Instead of connecting the external aerial to the tuned circuit, connect to the tapping point of the coil (tag to which the diode is also connected). Try a direct connection, and also connecting in a 220-pF capacitor.

---

* Tune to an AM broadcast station near the lower end of the dial—close to 550 kHz.

(iv) Try connecting a 1.2-k resistor (or higher value) across the phone connection (R in Fig. 3.5). You may be using the wrong type of phones or earpiece, which do not provide a proper load or complete the circuit.

If there is a complete lack of response, check for faulty wiring-up. A more likely cause, however, is lack of an external aerial or earth connection in an area where these are strictly necessary for adequate reception; or an inefficient aerial (too short) or poor earth connection (bad electrical contact to a good earth point, or connection to a bad earth point).

Another possible cause of apparent failure may be too much outside noise entering the ear so that it is impossible to detect the very weak radio signal as it is being tuned in. Headphones are better than a single deaf-aid type of earpiece in this respect but, in any case, a really quiet room is virtually essential for initial setting up and tuning adjustments. Also, if your adjustment of the tuning control is too coarse, you may completely miss the setting for the station you are looking for, without realizing it.

Reception will also tend to vary with weather conditions. Some days it may be so poor that what was normally a strong station is hardly heard at all. The simple basic receiver has many limitations but, since it costs very little to construct and nothing at all to operate, this must be regarded as inevitable.

Providing you can hear something—even if too weak a signal to distinguish properly—you can certainly improve the performance of your basic set by further experimentation with tuned circuits (see Chapter 4) and/or the addition of amplification to the circuit. You can also try other types of basic crystal set, as described in the following projects.

Circuit 2 (Fig. 3.6) is identical to *Circuit 1* except that, instead of a diode, a transistor is used as a detector. Only two of the transistor leads are connected—the emitter (*e*) connection to the tapping part of the coil, and the base (*b*) to the 'earthy' end of the circuit. The collector lead of the transistor is ignored (bend it out of the way so that it cannot accidentally short out the other leads).

You can try almost any type of low-cost AF transistor; recommended types are OC42 2N370, 2N2925, 2N5088.

With the addition of two more components, Circuit 2 can be modified to work the transistor both as a detector and an amplifier, to give stronger signals through the headphones. Using the same transistor type

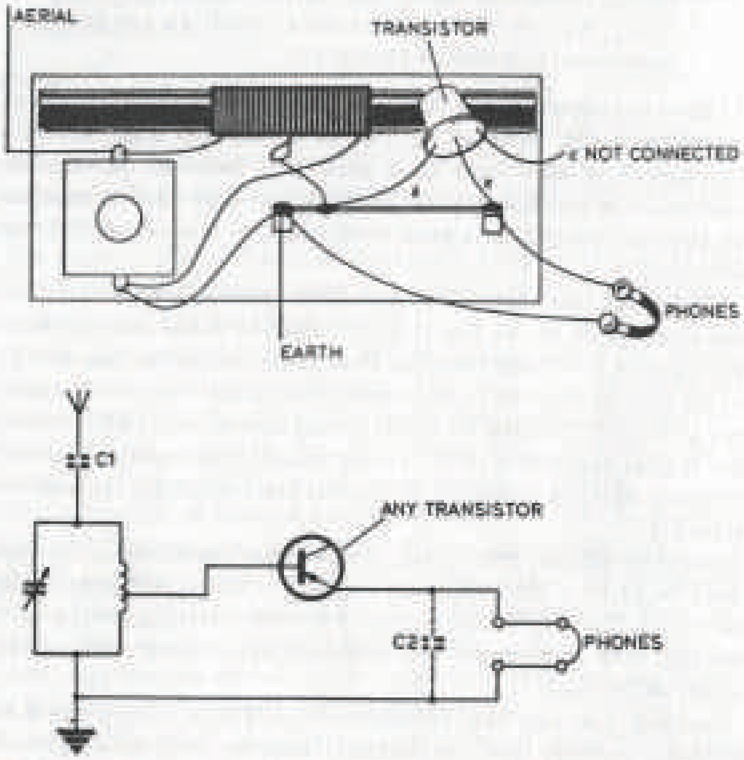
AERIAL
TRANSISTOR
e NOT CONNECTED
b
c
PHONES
EARTH
C1
ANY TRANSISTOR
C2
PHONES

Fig. 3.6 *Circuit 2.* Crystal set using transistor instead of a diode.

as above (or near-equivalent) resistor R should be 15 k and capacitor C3 1 μF or higher.

This time a battery is also required to supply power for the transistor to work as an amplifier. This can be from 1.5 V up to 9 V. Remember the rules for polarity of connection: those shown in Fig. 3.7 are for a p-n-p transistor; an n-p-n transistor would need the battery connected the opposite way round. Battery polarity also affects the connections of capacitor C3 (if an electrolytic or polarized type).

Experiment further by trying the effect of using additional capacitors in the circuit, e.g. C1, 220 pF; C2, 0.001 μF (try other values as well); C4, 0.001 μF (try other values as well).

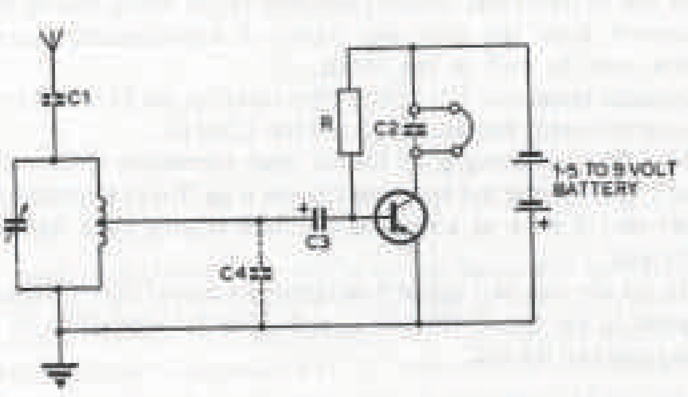
C1
R
C2
1·5 TO 9 VOLT BATTERY
+
+
C3
C4

Fig. 3.7 *Circuit 3.* Crystal set with amplification.

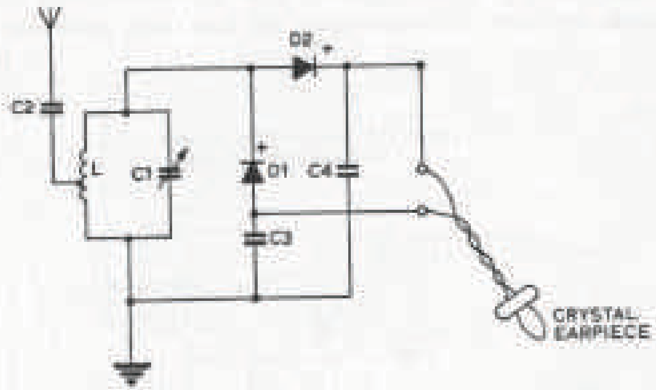
D2
+
C2
L
C1
+
D1
C4
C3
CRYSTAL EARPIECE

Fig. 3.8 *Circuit 4.* Double diode crystal set.

This circuit incorporates 'voltage doubling' to improve the signal volume and should give better performance than a single diode circuit. L and C1 are the usual tuned circuit, but an unusual feature is that the *aerial* is connected to the coil tapping point. Any type of germanium diodes can be used (they should preferably be the same), making sure to connect them the right way round. A high-impedance *crystal* earpiece must be used in this circuit.

Capacitor values are: C2, 220 pF (this capacitor can be omitted–try with and without in the circuit); C3 and C4, 1,000 pF.

The optimum tapping point for the aerial connection to the coil L is best found by trial and error (see Chapter 4 on 'Tuned Circuits'), but the set should work at a nominal one-third tapping point from the 'earthy' end.

This set can also be tried with conventional tuned-circuit coupling– i.e. aerial to the top of the coil L, and diode D2 connection to the tapping point on the coil.

4 MORE ABOUT TUNED CIRCUITS

The basic combination of a coil of certain *inductance* connected in parallel with a *variable capacitor* is the form of tuned circuit used in most radio receivers, from simple crystal sets upwards. The *tuning range* of such a circuit is determined by the respective values of inductance and capacitance used. At the same time, the *resistance* of the coil can have a significant effect on the *selectivity* of the circuit, so that a particular broadcast frequency is tuned into sharply, with other nearby frequencies rejected.

In the simple tuned circuit used for the crystal set of Chapter 3, the aerial is connected to the upper or 'hot' end of the coil, the other end of the coil being earthed; the diode is connected to a tapping point on the coil. Performance may well be improved by adjusting both the aerial-connecting point and the diode-connecting point, as shown in Fig. 4.1.

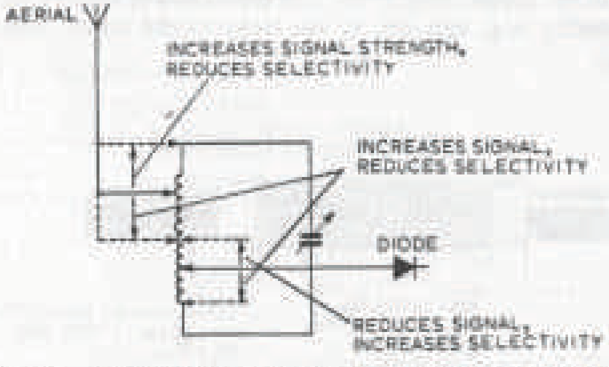
AERIAL
INCREASES SIGNAL STRENGTH, REDUCES SELECTIVITY
INCREASES SIGNAL, REDUCES SELECTIVITY
DIODE
REDUCES SIGNAL, INCREASES SELECTIVITY

Fig. 4.1 The effect of varying the tapping points on the aerial coil.

Connecting the aerial to the 'hot' end should give the loudest signal, but the *selectivity* of the circuit may be poor because of the coil *resistance* involved in the actual circuit to earth. Connecting the aerial to a tapping point lower down the coil (towards the 'earthy' end) should improve selectivity, but reduce signal strength. Experimentation will produce the optimum aerial-connecting point. There will also be an optimum tapping point for the diode.

Finding optimum tapping points for the aerial and diode connections is a bit tricky on a close-wound coil since each tapping point tried has to be bared by scraping off the enamel insulation, with the risk of producing shorted turns. The same *effect* can be provided by removing, or adding, turns at each end of the coil, using the original tapping point for the diode connection–e.g. see Fig. 4.2. Unless the *total* number of turns on the coil remain the same, however, the inductance of the complete coil will be altered, and thus the tuning range of the tuned circuit (see later). Then the *diode* tapping point will be shifted.

For this type of experiment an air-cored coil of relatively large diameter is easier to use, wound on a card or paper tube to which two $\frac{1}{4}$ in.^{2} lengths of hardwood or balsa strip have been cemented. The windings can be bared by scraping or sandpapering over this raised section, and individual wires carefully separated to avoid shorting–Fig. 4.3. Optimum tapping points can be established by pushing bared ends on the aerial- and diode-connecting wires into the raised section of the coil, soldering in place very carefully once they have been found. This should not result in shorting out more than a few turns at these points; which will not greatly modify the coil inductance if allowed for in the initial winding (e.g. add five or six turns to the design number of windings).

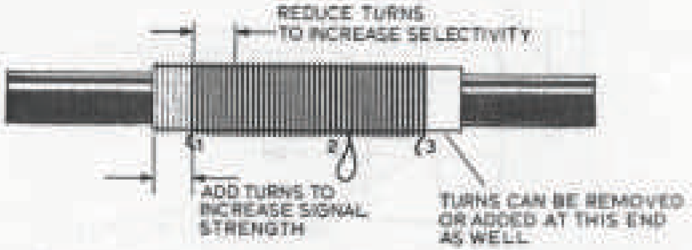
REDUCE TURNS
TO INCREASE SELECTIVITY
1
2
3
ADD TURNS TO INCREASE SIGNAL STRENGTH
TURNS CAN BE REMOVED OR ADDED AT THIS END AS WELL

Fig. 4.2 The effective tapping point on a simple coil can also be varied by adding or removing turns on each end of the winding.

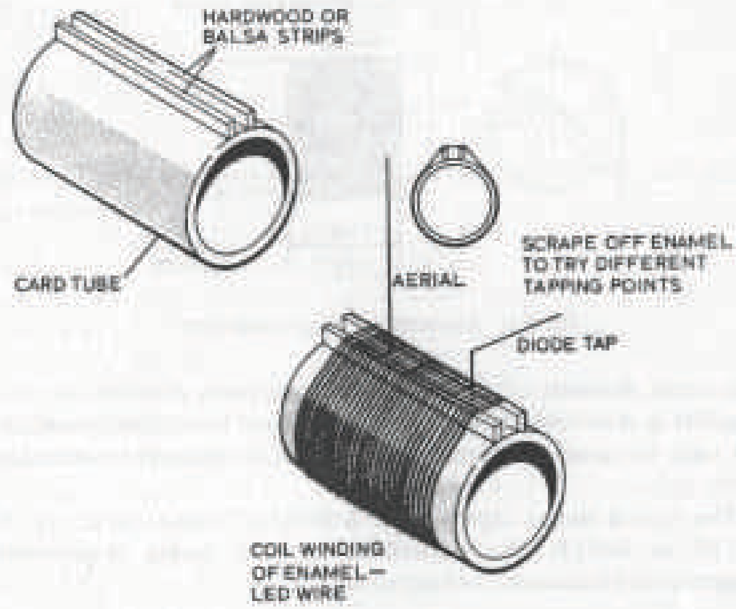
HARDWOOD OR BALSA STRIPS
CARD TUBE
AERIAL
SCRAPE OFF ENAMEL TO TRY DIFFERENT TAPPING POINTS
DIODE TAP
COIL WINDING OF ENAMEL-LED WIRE

Fig. 4.3 Construction of an air-cored aerial coil.

In air-cored coils, the inductance can be calculated with reasonable accuracy from the diameter and number of turns (see also Fig. 4.4):

$$\text{L. inductance, microhenries} = \frac{R^2 \times N^2}{9R + 10l}$$

where R = coil radius in inches,
l = coil length in inches,
N = number of turns.

The more usual form of this formula is:

$$N = \sqrt{\left(\frac{(9R + 10l) \times L}{R^2}\right)}$$

(A worked-out example using this formula is given later.) This is based on the turns of the coil being close-wound, in enamelled copper wire.

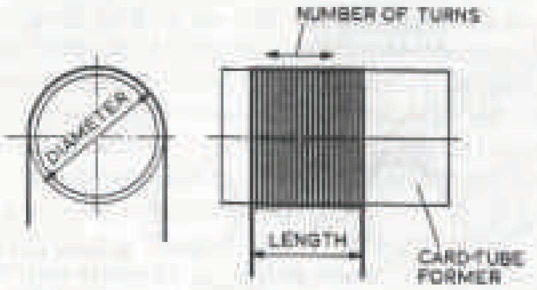

Fig. 4.4 Air-cored coil design parameters.

The actual diameter of the wire is not significant, provided the coil diameter is reasonably large (e.g. greater than 1 inch). The formula is not valid for smaller-diameter air-cored coils, or for coils wound on a ferrite rod.

The typical tuning capacitor has a rating of (about) 50 pF up to 350 pF or 500 pF. The relationship between tuning or *resonant frequency* and inductance and capacitance is:

$$\text{resonant frequency } (f) = \frac{1}{6.28\sqrt{(LC)}} \times 10^6 = \frac{0.16}{\sqrt{(LC)}} \times 10^6$$

where L = inductance in microhenries (μH),
C = capacitance in picofarads (pF).

This is more conveniently written as a solution for the *product* of inductance and capacitance required, i.e.

$$LC = \frac{0.025}{f^2} \times 10^{12}$$

For the medium waveband the range of frequencies to be covered is 500 to 1,500 kHz.* Thus, to encompass this range, the products of LC required are:

at 500 kHz

$$LC = \frac{0.025}{(500)^2} \times 10^{12}$$

$$= 0.001 \times 10^8 \text{ (approx.)}$$

---

* In the U.S.A. the AM broadcast band is from 530 to 1600 kHz.

at 1,500 kHz

$$LC = \frac{0.025}{(1{,}500)^2} \times 10^{12}$$

$$= 0.0001 \times 10^8 \text{ (approx.)}$$

At one end of the capacitor tuning range where C = 50 pF, inductance L required is:

$$\frac{0.001 \times 10^8}{\cancel{50}\ 500}$$

$$= 200\ \mu\text{H}$$

At the other end of the capacitor tuning range where C = 500 pF, inductance L required is:

$$\frac{0.0001 \times 10^8}{\cancel{500}\ 50}$$

$$= 200\ \mu\text{H}$$

Using a 1-in. coil diameter (0.5-in. radius) and a coil length of 1.5 in.

$$N = \sqrt{\left(\frac{(4.5 + 15) \times 200}{0.5^2}\right)}$$

$$= \sqrt{15{,}600}$$

$$= 125$$

In other words, a 125-turn air-cored coil of this size should be about right for covering the medium waveband, matching a 50–500 pF variable capacitor. Tapping point (or diode connection) would nominally be one-third of the coil length from the 'earthy' end.

Air-cored coils can also be designed in this way for long-wave coverage (more turns); or short-wave reception (less turns), which is not likely to be so good because of the low efficiency of the coils. Also, at higher frequencies the resistance of the *capacitor* is also significant, and so simple calculation of coil size is no longer valid.

The efficiency of a tuning coil can be expressed in terms of the magnification of the original signal received, when tuned to resonance, and decreases with increasing coil resistance. The magnification produced in a resonant circuit is referred to as the 'Q'; a high-Q coil is thus desirable for maximum performance and selectivity, and is essentially a low-resistance one. For this reason larger diameter air-cored coils, using thicker wire, are more effective than smaller diameter coils wound from thinner wire (since wire resistance increases with decreasing diameter).

Equally, the inductance of a wound coil can be increased by winding it over a magnetic-iron core, such as a ferrite rod. Much smaller diameter coils can be used, requiring less turns, to produce the inductance required—both reducing the length of wire and its resistance.

Because the actual inductance produced is dependent on the size and characteristics of the core material, windings for ferrite-rod aerials cannot be calculated, and are based on empirical results, or optimum windings established by trial and error.

Common sizes of ferrite rod available are:

$\frac{1}{4}$-in. diameter	length $3\frac{1}{2}$ in.
$\frac{5}{16}$-in. diameter	lengths 4, 5, 6 and 8 in.
$\frac{3}{8}$-in. diameter	lengths 4, 5, 6, 8 and 10 in.

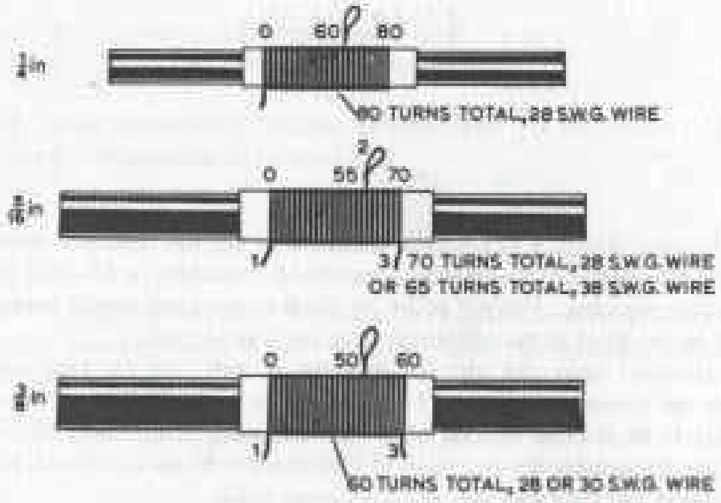
$\frac{1}{4}$in
0
60
80
80 TURNS TOTAL, 28 S.W.G. WIRE
$\frac{5}{16}$in
0
55
2
70
1
3
70 TURNS TOTAL, 28 S.W.G. WIRE
OR 65 TURNS TOTAL, 38 S.W.G. WIRE
$\frac{3}{8}$in
0
50
60
1
3
60 TURNS TOTAL, 28 OR 30 S.W.G. WIRE

Fig. 4.5 Medium-wave aerial coil windings on three sizes of ferrite rod.

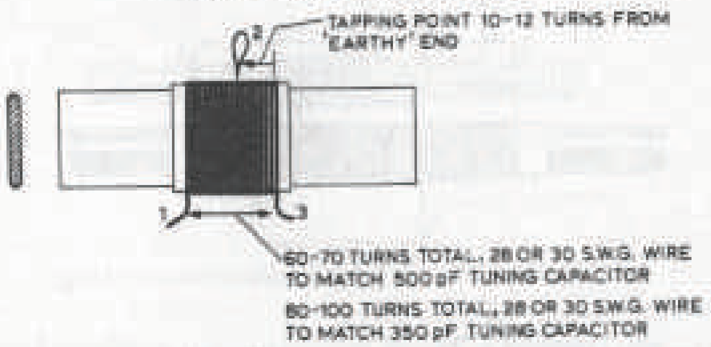
TAPPING POINT 10–12 TURNS FROM 'EARTHY' END
2
1
3
60–70 TURNS TOTAL, 28 OR 30 S.W.G. WIRE
TO MATCH 500 pF TUNING CAPACITOR
80–100 TURNS TOTAL, 28 OR 30 S.W.G. WIRE
TO MATCH 350 pF TUNING CAPACITOR

Fig. 4.6 Medium-wave winding for a ferrite slab.

A long length enables maximum adjustment of coil position (a way of 'trimming' the aerial) and, also, additional coils to be incorporated on the same rod, if the circuit needs them (e.g. for multi-waveband coils). But it is the diameter which really determines the winding specification.

Recommendations for medium-wave coil windings for these three standard ferrite-rod diameters are given in Fig. 4.5, all windings being enamelled copper wire. Normal tapping points are one-third from the 'earthy' end. It is a simple matter to experiment with different windings—e.g. adding more turns to extend the *wavelength* upwards (i.e. lower frequencies), or reducing the number of turns to extend the *wavelength* coverage downwards (i.e. for higher-frequency transmissions).

Ferrite-core material is also available in slab form, which is often easier to accommodate in a small receiver case. A typical ferrite-slab size is $2\frac{3}{4}$ in. long by $\frac{3}{4}$ in. wide, with a thickness of $\frac{1}{8}$ in. or less. A typical winding specification is shown in Fig. 4.6.

Generally, the long waveband should be covered by *doubling* the number of turns used for the medium waveband. *Reducing* the number of turns shifts the coverage out of the medium waveband down into the trawler band (80 m) and 'top-band' (160 m). For short-wave coverage very few turns indeed will be needed—perhaps less than half a dozen—and the optimum number can only be determined by careful experiment.

Separate coils for the various wavebands to be covered can be accommodated on the same ferrite rod and connected to a common tuning capacitor via a waveband switch—Fig. 4.7. This is the usual

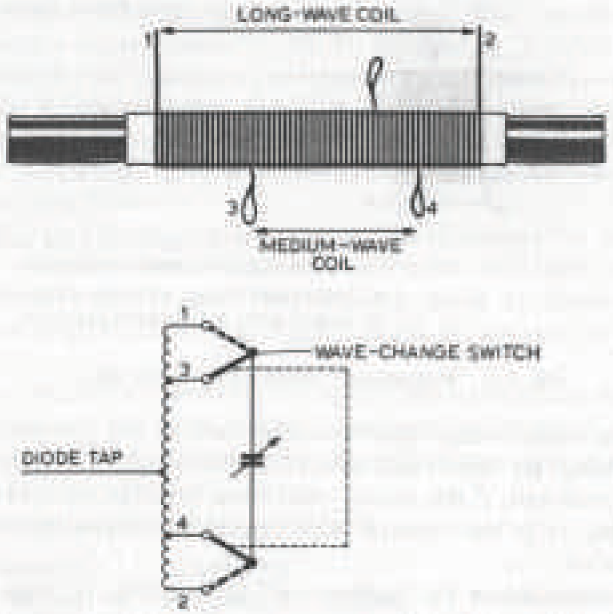
LONG-WAVE COIL
1
2
3
4
MEDIUM-WAVE COIL
1
WAVE-CHANGE SWITCH
3
DIODE TAP
4
2

Fig. 4.7 Long- and medium-wave aerial coil connected to wave-change switch.

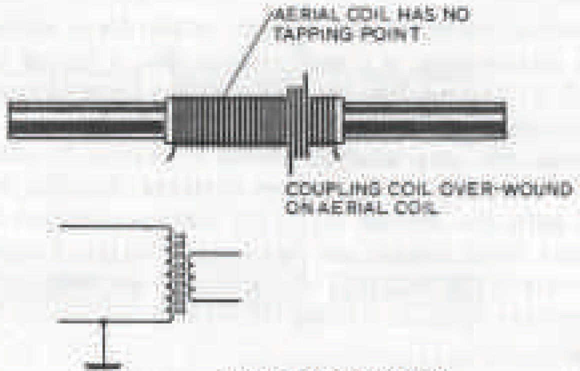
AERIAL COIL HAS NO TAPPING POINT
COUPLING COIL OVER-WOUND ON AERIAL COIL
ELECTRICAL DIAGRAM

Fig. 4.8 Inductively-coupled aerial coil.

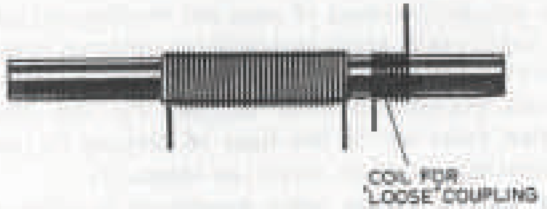
COIL FOR 'LOOSE' COUPLING

Fig. 4.9 'Loose' inductive coupling with separate coupling coil.

procedure with an elaborate receiver circuit and domestic radios, but is seldom satisfactory with the simpler circuits because of the tuned-circuit characteristics (i.e. the need for meticulous adjustment to get optimum results for any single waveband coverage).

All the coil constructions so far described are based on single windings (covering a specific waveband) with a tapping point for the output (diode connection in the case of simple receivers). The alternative form is to provide *inductive coupling* for the output in the form of an overwinding, which can comprise five or more turns of the same gauge wire used for the main winding–Fig. 4.8. The optimum number of turns for the overwinding depends on the degree of coupling required, which in turn is influenced by the characteristics of the following circuit. The position of the coupling coil can also affect the performance, although usually it is best positioned over the 'earthy' end

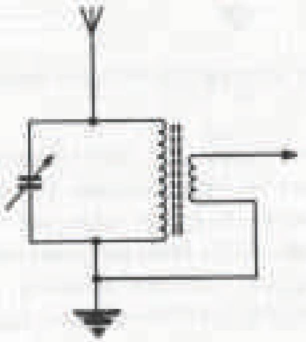

Fig. 4.10 Transformer coupling between tuned circuit and next stage of receiver.

of the main coil. Both number of turns and coupling-coil position are subjects for experimental adjustment to obtain optimum results.

Sometimes a coupling coil is wound on the ferrite rod, separate from the aerial coil, providing a 'loose' coupling–Fig. 4.9. Regenerative receivers often make use of this form of coupling for introducing 'feedback' into the front of the circuit (see Chapter 9).

Inductive coupling is also called *transformer* coupling, since the principle of passing current generated in one coil into another coil is the same. With transformer coupling (Fig. 4.10) the aerial coil has no tapping point, and is an alternative to direct coupling, where the next stage of the circuit is connected directly to a tapping point on the coil.

There is a third method of coupling which may be used between stages in a radio circuit–*capacity coupling*. Here, connection between the stages (e.g. to the tapping point on the aerial coil) is made through a capacitor, which isolates the two stages as far as d.c. flow is concerned (only a.c. will pass through the capacitor)–Fig. 4.11.

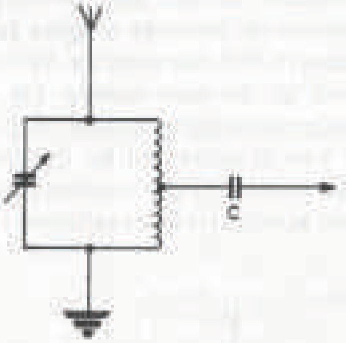

Fig. 4.11 Capacity coupling between tuned circuit and next stage of receiver.

Again the optimum value for the capacitor depends on the characteristics of the circuits being coupled. To couple an aerial coil to the next stage in a receiver, optimum values can vary from 200 pF to 10 μF. In general, though, a high value gives best results.

Direct, transformer, and capacity coupling are alternatives for connecting any two stages in receivers–not just the tuned circuit to the next stage.

5 AMPLIFIERS

It is a characteristic of a transistor that it works as an *amplifier* of signals. In the most widely used mode of connection of transistors for such duties with input and output circuits both connected to the emitter (common-emitter mode), the degree of amplification or current gain is called the 'beta' (β) of the transistor. This is given as h_{FE} (or static forward current ratio) in transistor characteristics data (see also Chapter 7).

A basic transistor amplifier circuit is very simple–and is identical for a p-n-p or n-p-n transistor, except for the battery polarity–see Fig. 5.1. Virtually any low-power AF transistor can be used in this circuit. The *bias resistor* (R) must have a value providing a collector current not exceeding the maximum specified rating for the transistor used, the actual current flowing in the collector circuit also being influenced by the voltage of the battery.

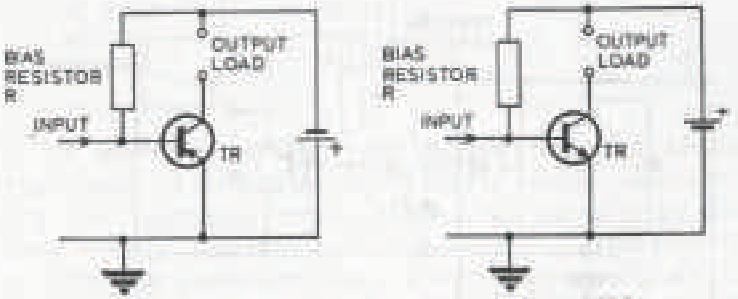

p-n-p TRANSISTOR n-p-n TRANSISTOR

Fig. 5.1 Basic amplifier circuits with simple current biasing via a single resistor.

Knowing the transistor characteristics, a suitable value for R can be calculated as follows:

$$R = \text{gain} \times \frac{\text{battery voltage}}{\text{collector current}}$$

$$= h_{FE} \times \frac{\text{battery voltage}}{I_c}$$

where I_c is equal to, or preferably less than, the specification figure for I_c max.

Assuming that an OC72 transistor is being used, and the battery voltage is 9 V, specified figures are:

$$I_c \text{ max} = 250 \text{ mA}$$
$$h_{FE} = 30 \text{ to } 90$$

Taking the maximum gain, and 2.5 mA as a 'safe' working figure for the collector current:

$$R = 90 \times \frac{9}{0.0025}$$
$$= 324\text{k}\Omega$$

A suitable (preferred-value) resistor would thus be 330k.

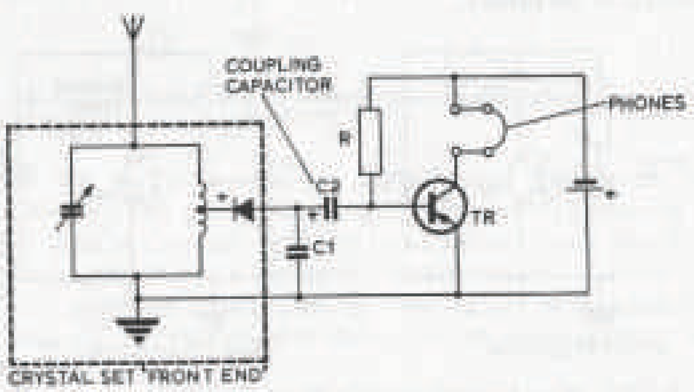

Fig. 5.2 *Circuit 5.* Crystal set with single stage of amplification. Any ~~RF~~ transistor can be used for TR.
AF

Fig. 5.2 shows the complete circuit for a simple AF amplifier of this type, coupled to the 'front end' of a basic crystal set. The detector (output) is coupled to the amplifier via capacitor C2, a suitable value for which would be 10 μF or higher (e.g. 20, 25, 30, 40 or 50 μF); capacitor C1 (0.001 μF) may not be necessary. The output load in the collector circuit is formed by high-impedance phones.

Note that the *polarity* of the battery used to power this circuit is important (as far as the transistor connection is concerned); also the connections to the *diode* and C2 (which will normally need to be an electrolytic type to provide the high-capacity value required). With an n-p-n transistor the battery polarity would be reversed; and also the diode and electrolytic capacitor connections.

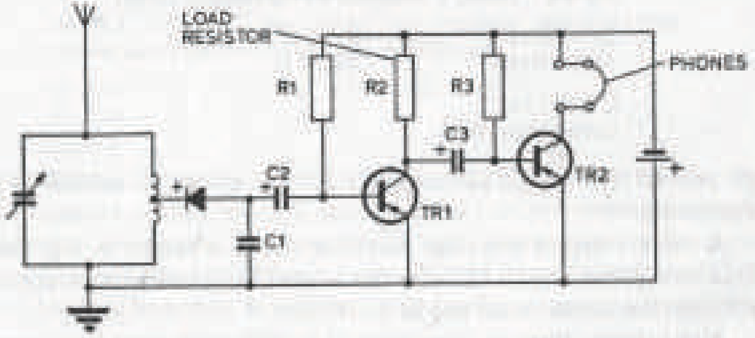

Fig. 5.3 *Circuit 6.* Crystal set with two stages of amplification. Component values matching OC71, OC72 or near equivalent for TR1 and TR2:

R1, 470 k	C1, 0.001 μF
R2, 4.7 k	C2, 8 or 10 μF
R3, 470 k	C3, 8 or 10 μF

Battery voltage 6 to 9 V.

The same type of circuit can be used to provide additional amplification, if required, simply by adding another amplifier stage (Fig. 5.3). The second amplifier stage can be identical to the first or based on a higher-power transistor taking the higher output present and providing even greater gain. The load in the output (collector) of the first amplifier stage is provided by a resistor (R2) (which should be about the same value as the phone resistance, 3.3 to 4.7 k), and the two stages coupled by a capacitor (C3). Value of R3 depends on the second transistor used, and may be anything from 47 Ω to 470 kΩ; capacitors

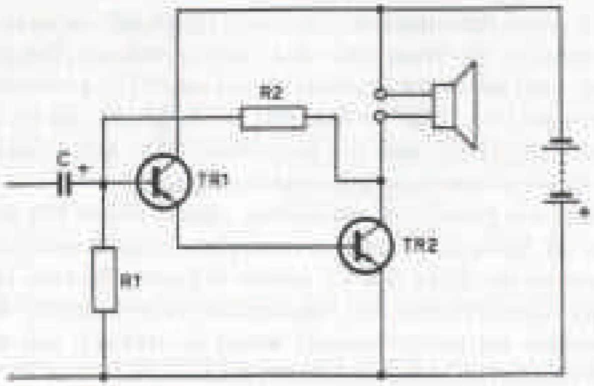

Fig. 5.4 *Circuit 7.* Amplifier for loudspeaker output

TR1, OC72 R1, 6.8 k
TR2, AD140 R2, 33 k
C, 8 or 10 μF
Loudspeaker, 80 Ω

C2 and C3 can be 10 μF or larger (20 μF up to 50 or 100 μF), electrolytic.

A more compact two-stage amplifier circuit is shown in Fig. 5.4. This may prove capable of operating a small 80 Ω loudspeaker direct, although the current drain will be quite high.

Alternatively, three or four stages of amplification using low-power transistors, following a basic 'front end' crystal set, should provide enough power to drive a small loudspeaker at the final output, through a suitable step-down transformer, to provide an impedance match (see Chapter 6).

Simple amplifier circuits of this type have the important limitation that performance of the transistor(s) will tend to vary with temperature. There is also the possibility of 'thermal runaway' developing, which can destroy the transistor(s), because as the external temperature increases the collector current also tends to increase, which in turn causes a further increase in junction temperature, so the effect is cumulative and goes from bad to worse, even to the point of ruining the transistor completely. It is possible to overcome this trouble by arranging for the circuit to be self-biasing or d.c. stabilized so that a constant operating collector current is provided, regardless of transistor type or temperature variations. In other words, the working point of the collector circuit is stabilized.

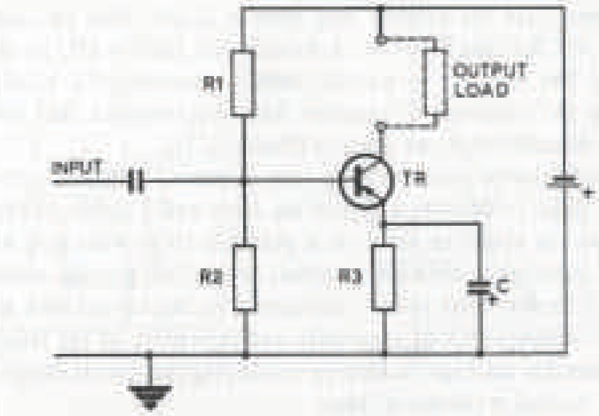

Fig. 5.5 Stabilized bias circuit for transistor amplifier stage. Typical values for low/medium-power transistors

R1, 22 k C, 8 or 10 μF
R2, 10 k
R3, 1 k

A further advantage of a stabilized circuit is that it makes the performance of the amplifier less dependent on the characteristics of individual transistors, which can differ appreciably even for the same type. Capacity coupling between stages should be used since this makes each stage independent as regards working, rather than inter-dependent.

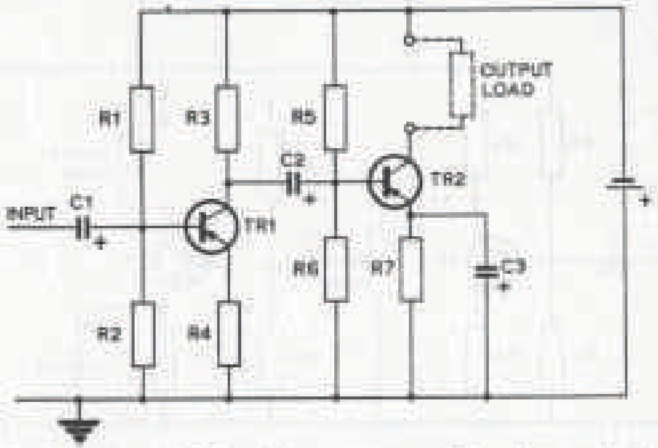

Fig. 5.6 *Circuit 8.* Practical two-stage amplifier using two low/medium-power AF transistors. (See Chapter 7 for determination of matching component values.)

To achieve this the original bias resistor is split into two separate values R1 and R2 (see Fig. 5.5). A further bias resistor (R3) is applied directly to the emitter, in parallel with a capacitor (C) to act as a by-pass for AF currents. (Transistor bias requirements, and data on transistor characteristics, are given in Chapter 7.)

A receiver design incorporating two stages of amplification with stabilized circuits following a crystal set front end is shown in Fig. 5.6. This, in fact, is about as far as it is practical to go with such a basic circuit as, although additional complete stages will provide more gain and greater final output power, deficiency in the circuit will also be aggravated–notably lack of sensitivity and selectivity in the front end. Such deficiencies are best tackled by modifying the circuit design from the simple 'crystal set detector' basis.

The amplifier stage(s) also provides a convenient point to insert a *volume control* into the receiver circuit. This takes the form of a potentiometer which can replace one of the bias resistors, or be placed in series with the flow of the current after the first stage of amplification–Fig. 5.7. This will introduce minimum distortion over the volume control range.

On low-power receivers which give satisfactory reception without an external aerial, a volume control can be an unnecessary refinement. Since the ferrite aerial has directional characteristics, merely turning the set one way and the other will produce changes in signal strength.

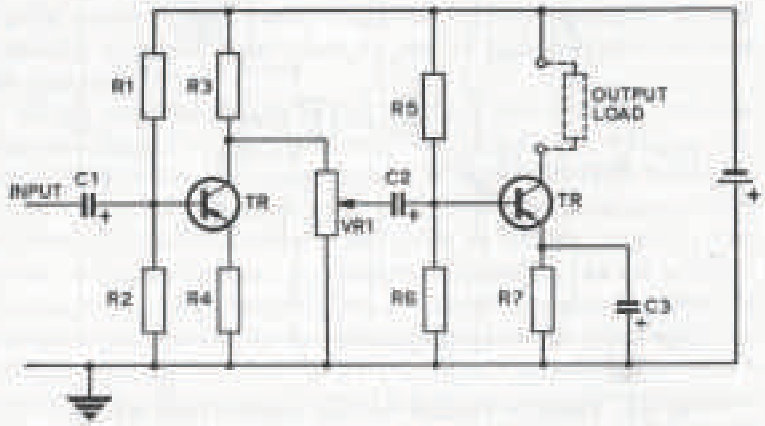

Fig. 5.7 *Circuit 9.* Two-stage amplifier with volume control.

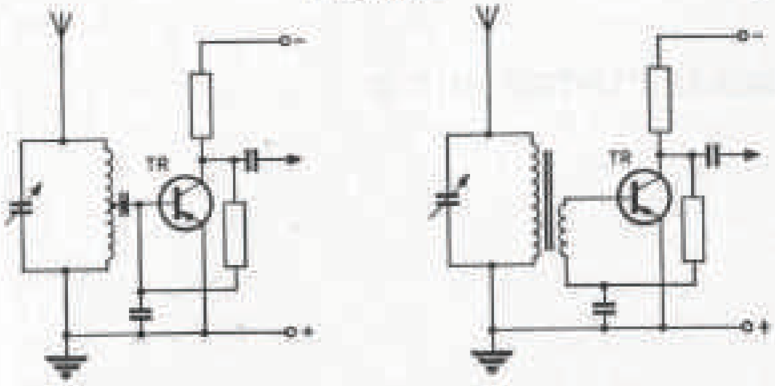

Fig. 5.8 Basic pre-amplifier circuits. The transistor (TR) must be an RF type; p-n-p transistors shown in these diagrams. If n-p-n transistors are used, the battery polarity must be reversed.

Pre-amplifiers

A pre-amplifier is designed to amplify signals between the tuned circuit and detector stages. The main difference is that the transistor(s) selected must be an RF type since it is handling RF signals at this stage. Otherwise the basic circuit is similar to that of an AF amplifier–see Fig. 5.8.

Part II: Diode Test Data Report

Back to Lab..

Test Results

Over a period of several weeks in 2013 I ran I-V characterization measurements on a good number and variety of diodes, crystals and my two tubes to see how very thing stacks up. The following presents the resulting curves, diode photos, and some puzzling questions/conclusions I churned up in the process. These Characteristic Curves carry the voltage out to 2.0 Volts, well beyond any signal strength one might expect to receive with a crystal set. This is important to note because the curves themselves are dominated by the diode series resistance. After each discussion I then include my spreadsheet with the detail DX-level analyses that provide the critical information about diode "n", (Ideality factor) "Is", (Reverse Saturation current) and "Ro", (Zero-bias Diode resistance). These are the data the set designer requires.

I start off with the realization that when one orders diodes, it is by no means certain what one will get. This seems especially true for the supposedly ubiquitous 1N34A. Unless you see the part number actually on the diode, you probably need to test it to know what it really is. So, lets take a look!

Germanium Diodes

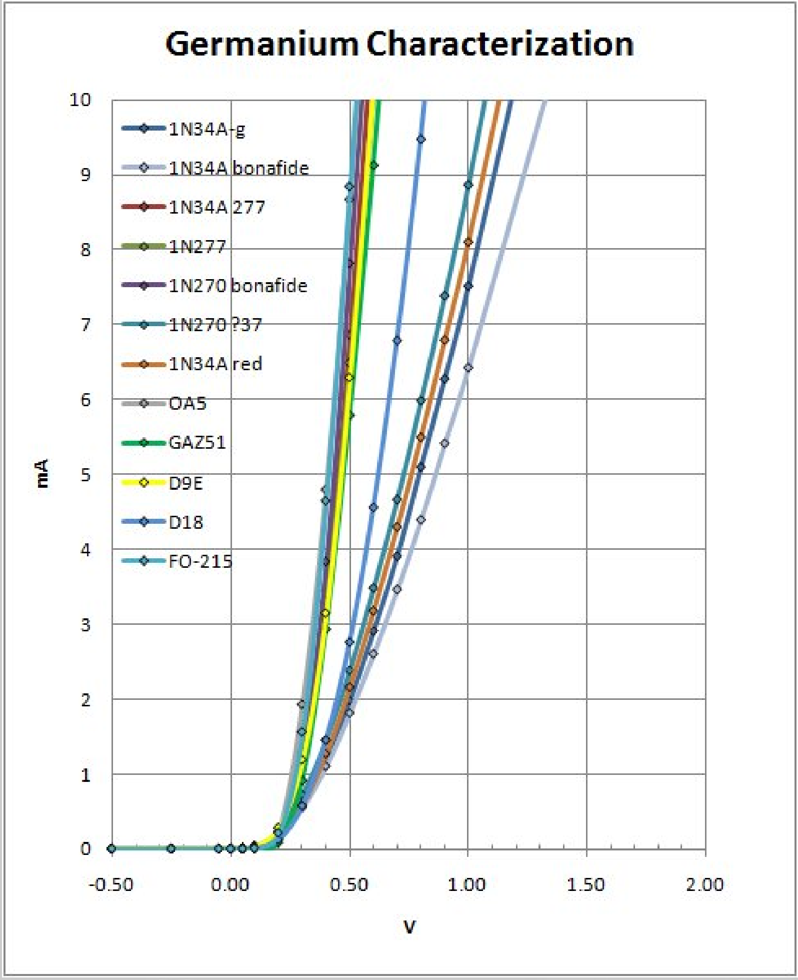

1N34A series

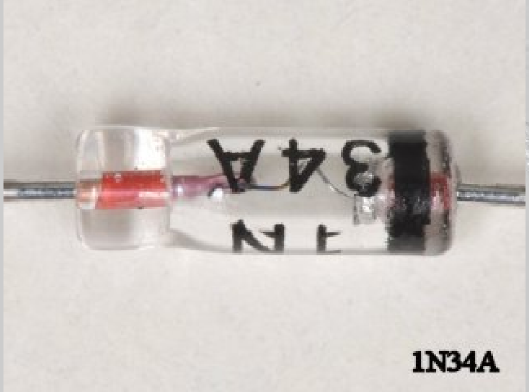

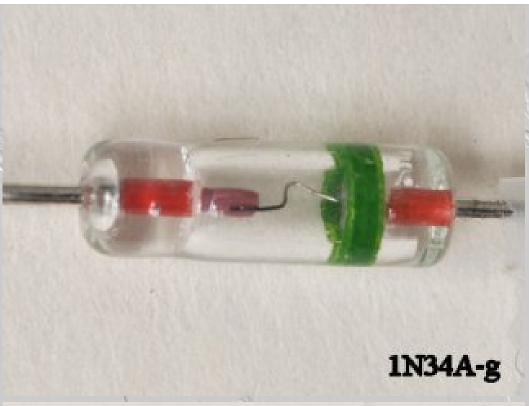

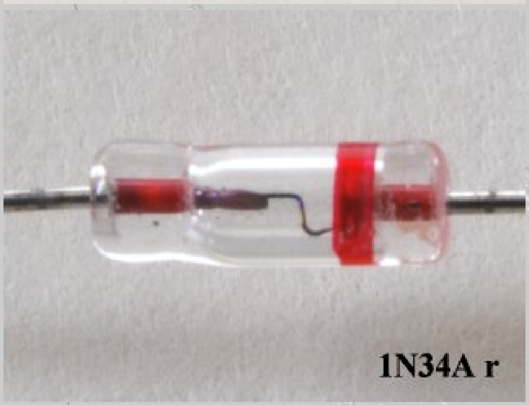

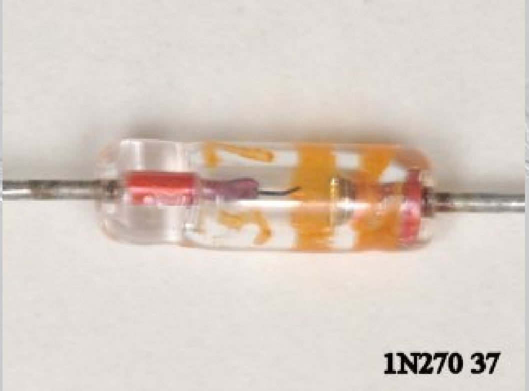

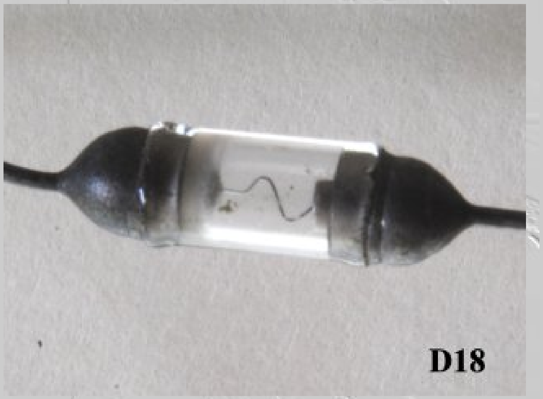

1N270-277 series

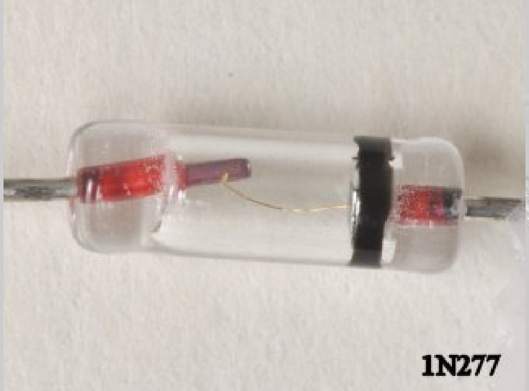

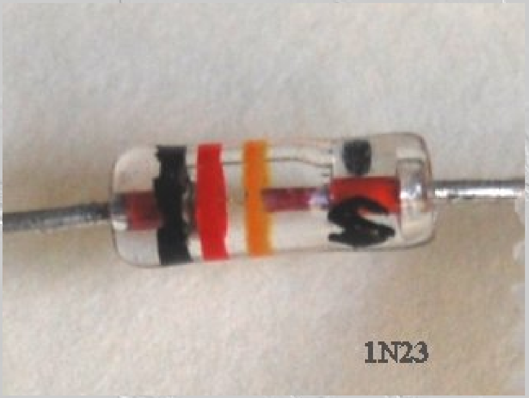

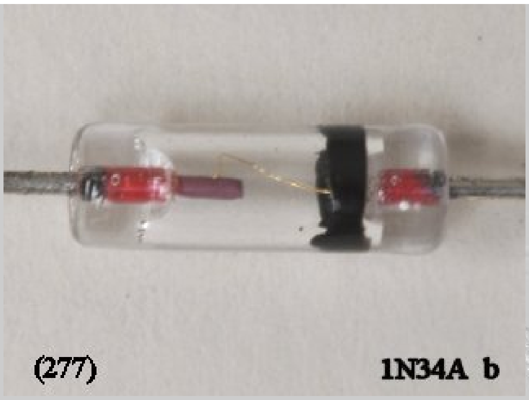
(277) 1N34A b

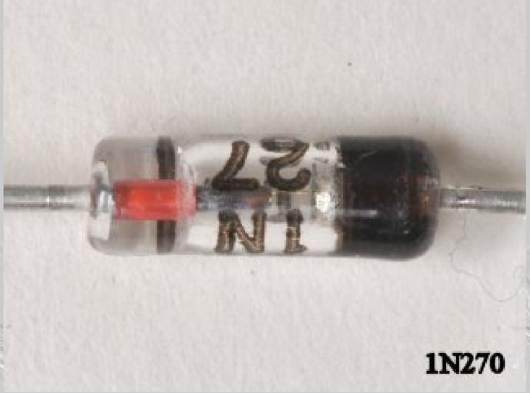
1N270

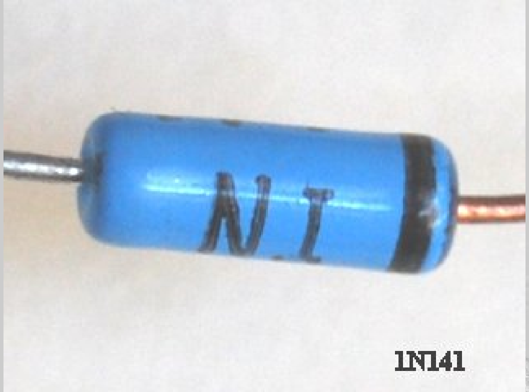
1N141

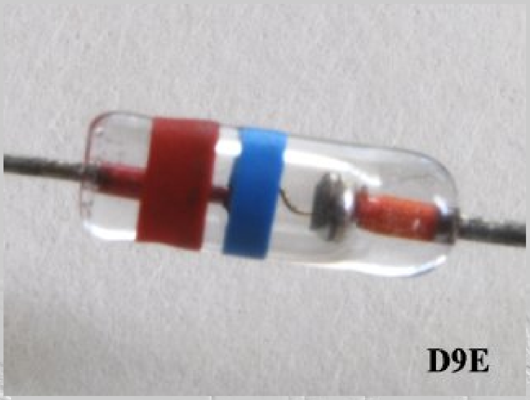
D9E

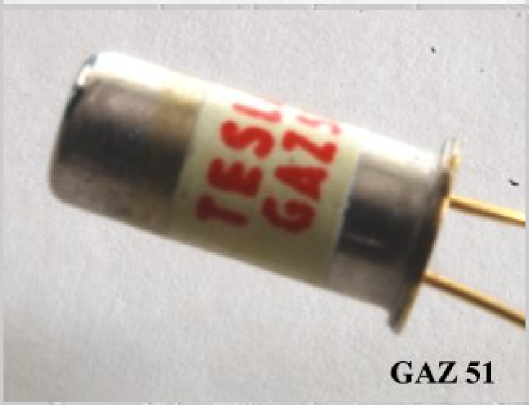
GAZ 51

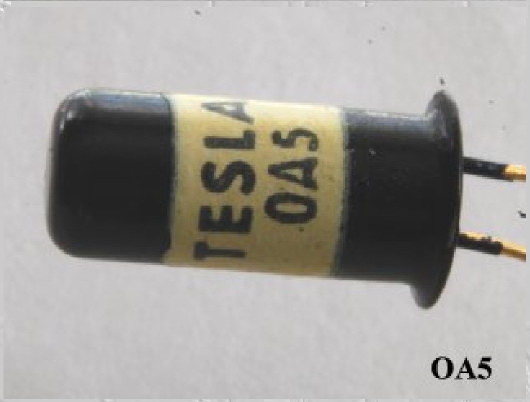
OA5

ITT Diodes, Something special here.

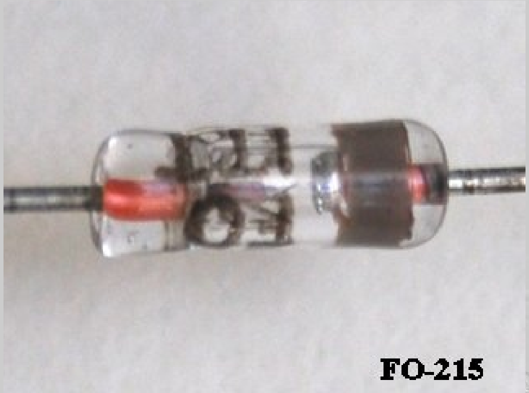
FO-215

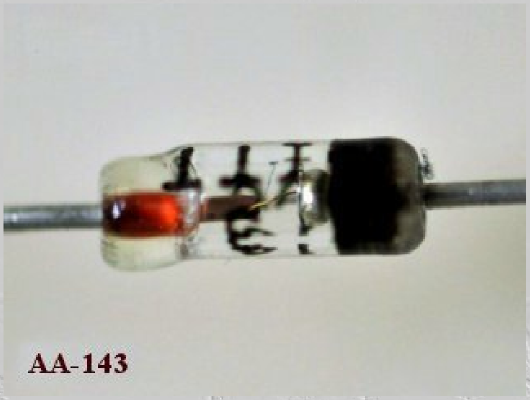
AA-143

One oddball Ge diode, not sure what it is for..

Here we see a set of Germanium diodes and photos of the diodes under test. Immediately you should notice two sets of curves, what I call the "1N34A" set and the "1N270-277" set. In this analysis, the second set appears clearly superior in sensitivity and deservedly so. In the photos it is clear that one bunch of diodes I purchased as 1N34A were in fact better 1N277, the black band and thin gold contact wire as well as common curve unite these. "True" 1N34A's have a more robust contact wire. I also purchased a 1N270 and received pretty orange diodes with the number "37" stamped on it, a fairly robust contact wire and a curve that most closely resembles a 1N34A, go figure. Mercifully, I DID manage to get some bona-fide 1N34A's with the part number on the diode, as well as a single bona-fide ITT 1N270 diode with the part number. These are my base for comparison.
In addition to the standard Germanium Diodes tested above, I also have latched my hands on a few "other" germanium diodes for the fun of testing. These include two vintage Russian types, a D9E (in the 270 class) and a D-18 (transitional between 270 and 1N34A). Additionally, I acquired a few FUZZ diodes popular, I suspect, with the guitar gadget crowd. These are larger packages in metal cases so I cannot see how the internal contact is made other that they are stated to be gold bonded point contact type. Clearly these are highly sensitive diodes in my 270 class, the OA5 looking best of the lot. Interestingly, I recently purchased from good Mr. Peebles a few of his "Holy Grail" ITT FO-215 diodes. My measured characteristic for one of these is a dead lay down on the gold-bonded OA5. From the photos, these are in different packaging with the FO-215 in a traditional glass casing. What makes the FO-215 so great is the fact that its resistance Rd is an interesting 150k ohm or so which matches very well with typical moderate Q tank circuits. Interestingly, I find that an old Russian D18 diode has very similar characteristics to the FO-215 and so must also be placed into the "Holy Grail" class of germanium diode. Take your pick!

Spreadsheet for calculation of Germanium diode n and Is (modified from Mike Tuggle's spreadsheet)

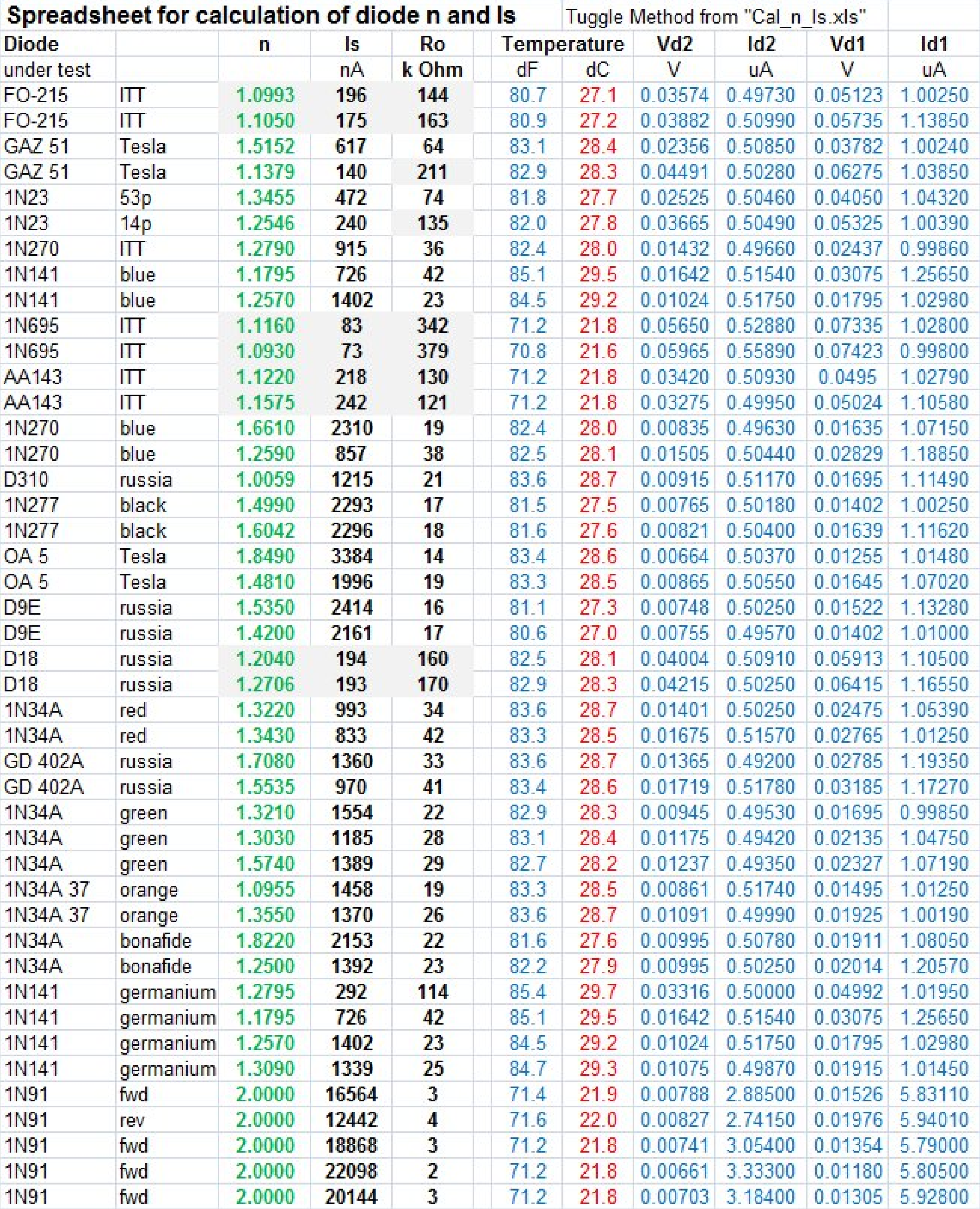

Spreadsheet for calculation of diode n and Is Tuggle Method from "Cal_n_Is.xls"

Diode		n	Is	Ro	Temperature		Vd2	Id2	Vd1	Id1
under test			nA	k Ohm	dF	dC	V	uA	V	uA
FO-215	ITT	1.0993	196	144	80.7	27.1	0.03574	0.49730	0.05123	1.00250
FO-215	ITT	1.1050	175	163	80.9	27.2	0.03882	0.50990	0.05735	1.13850
GAZ 51	Tesla	1.5152	617	64	83.1	28.4	0.02356	0.50850	0.03782	1.00240
GAZ 51	Tesla	1.1379	140	211	82.9	28.3	0.04491	0.50280	0.06275	1.03850
1N23	53p	1.3455	472	74	81.8	27.7	0.02525	0.50460	0.04050	1.04320
1N23	14p	1.2546	240	135	82.0	27.8	0.03665	0.50490	0.05325	1.00390
1N270	ITT	1.2790	915	36	82.4	28.0	0.01432	0.49660	0.02437	0.99860
1N141	blue	1.1795	726	42	85.1	29.5	0.01642	0.51540	0.03075	1.25650
1N141	blue	1.2570	1402	23	84.5	29.2	0.01024	0.51750	0.01795	1.02980
1N695	ITT	1.1160	83	342	71.2	21.8	0.05650	0.52880	0.07335	1.02800
1N695	ITT	1.0930	73	379	70.8	21.6	0.05965	0.55890	0.07423	0.99800
AA143	ITT	1.1220	218	130	71.2	21.8	0.03420	0.50930	0.0495	1.02790
AA143	ITT	1.1575	242	121	71.2	21.8	0.03275	0.49950	0.05024	1.10580
1N270	blue	1.6610	2310	19	82.4	28.0	0.00835	0.49630	0.01635	1.07150
1N270	blue	1.2590	857	38	82.5	28.1	0.01505	0.50440	0.02829	1.18850
D310	russia	1.0059	1215	21	83.6	28.7	0.00915	0.51170	0.01695	1.11490
1N277	black	1.4990	2293	17	81.5	27.5	0.00765	0.50180	0.01402	1.00250
1N277	black	1.6042	2296	18	81.6	27.6	0.00821	0.50400	0.01639	1.11620
OA 5	Tesla	1.8490	3384	14	83.4	28.6	0.00664	0.50370	0.01255	1.01480
OA 5	Tesla	1.4810	1996	19	83.3	28.5	0.00865	0.50550	0.01645	1.07020
D9E	russia	1.5350	2414	16	81.1	27.3	0.00748	0.50250	0.01522	1.13280
D9E	russia	1.4200	2161	17	80.6	27.0	0.00755	0.49570	0.01402	1.01000
D18	russia	1.2040	194	160	82.5	28.1	0.04004	0.50910	0.05913	1.10500
D18	russia	1.2706	193	170	82.9	28.3	0.04215	0.50250	0.06415	1.16550
1N34A	red	1.3220	993	34	83.6	28.7	0.01401	0.50250	0.02475	1.05390
1N34A	red	1.3430	833	42	83.3	28.5	0.01675	0.51570	0.02765	1.01250
GD 402A	russia	1.7080	1360	33	83.6	28.7	0.01365	0.49200	0.02785	1.19350
GD 402A	russia	1.5535	970	41	83.4	28.6	0.01719	0.51780	0.03185	1.17270
1N34A	green	1.3210	1554	22	82.9	28.3	0.00945	0.49530	0.01695	0.99850
1N34A	green	1.3030	1185	28	83.1	28.4	0.01175	0.49420	0.02135	1.04750
1N34A	green	1.5740	1389	29	82.7	28.2	0.01237	0.49350	0.02327	1.07190
1N34A 37	orange	1.0955	1458	19	83.3	28.5	0.00861	0.51740	0.01495	1.01250
1N34A 37	orange	1.3550	1370	26	83.6	28.7	0.01091	0.49990	0.01925	1.00190
1N34A	bonafide	1.8220	2153	22	81.6	27.6	0.00995	0.50780	0.01911	1.08050
1N34A	bonafide	1.2500	1392	23	82.2	27.9	0.00995	0.50250	0.02014	1.20570
1N141	germanium	1.2795	292	114	85.4	29.7	0.03316	0.50000	0.04992	1.01950
1N141	germanium	1.1795	726	42	85.1	29.5	0.01642	0.51540	0.03075	1.25650
1N141	germanium	1.2570	1402	23	84.5	29.2	0.01024	0.51750	0.01795	1.02980
1N141	germanium	1.3090	1339	25	84.7	29.3	0.01075	0.49870	0.01915	1.01450
1N91	fwd	2.0000	16564	3	71.4	21.9	0.00788	2.88500	0.01526	5.83110
1N91	rev	2.0000	12442	4	71.6	22.0	0.00827	2.74150	0.01976	5.94010
1N91	fwd	2.0000	18868	3	71.2	21.8	0.00741	3.05400	0.01354	5.79000
1N91	fwd	2.0000	22098	2	71.2	21.8	0.00661	3.33300	0.01180	5.80500
1N91	fwd	2.0000	20144	3	71.2	21.8	0.00703	3.18400	0.01305	5.92800

The above spreadsheet is based on measurements of the diodes shown above. For the determination of Is and n, I chose at random two examples from my collection of various diodes, (all germanium in this case) and measured via a modified version of the methodology outlined by Ben Tongue and Mike Tuggle. For this work I have chosen to set Id2 about 0.5uA and Id1 about 1.0uA and then read the needed voltage. This is the inverse of the normal method but the justification is that for any diode regardless of forward voltage drop the measurement is made at the same part of the LOG I vs V characteristic. This allows comparison between all diodes. I have included the actual room temperature in the calculations in order to eliminate this variable as a source of doubt or error. All the measured parameters as well as the determined results are presented in order to facilitate repeatability.

A final note on the plots above. You will have noticed that I sub-divide germanium diodes into a "1N34A" class and a "1N270-277" class. This is a result of looking at their I / V characteristics on a plot of 1 - 2V DC versus 0 - 10mA scale. At such a scale the slope of the characteristic for the 1N270 class remains much steeper than that for 1N34 type diodes. As this is not seen as a distinguishing feature at the small-signal scales looked at here, I judge the chief difference between the two types lies in the diode series resistance Rs. This impacts strong-signals but is of no concern for small-signal DX work. Just so you know.

Schottky Diodes

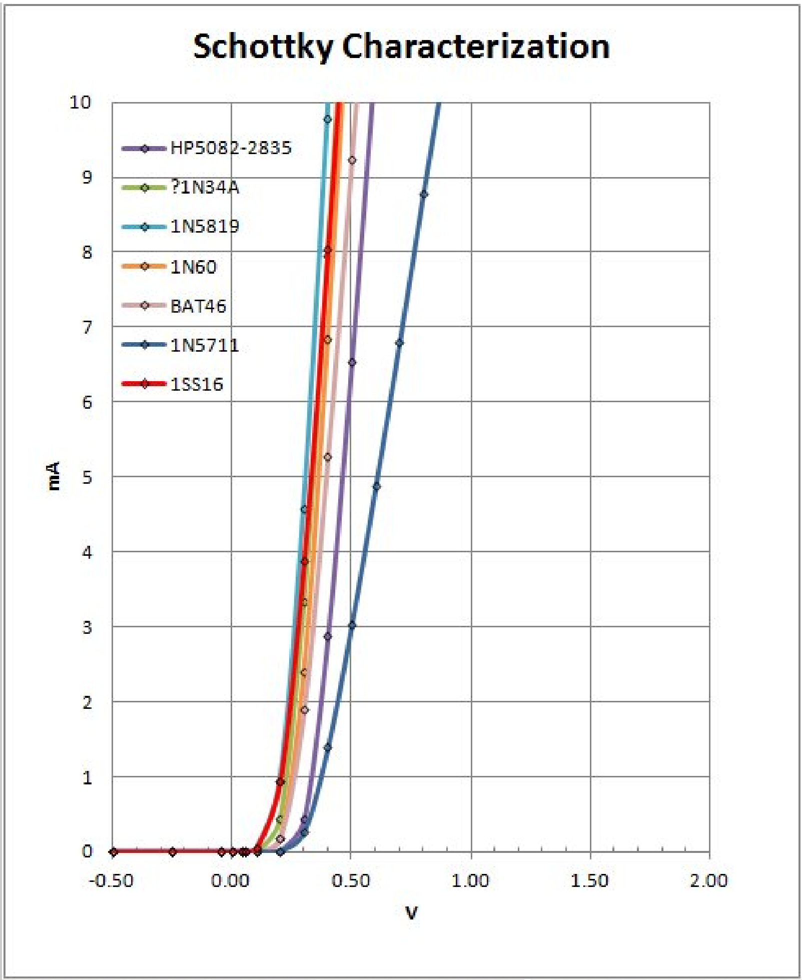
Schottky Characterization
mA
10
9
8
7
6
5
4
3
2
1
0
-0.50
0.00
0.50
1.00
1.50
2.00
V
HP5082-2835
?1N34A
1N5819
1N60
BAT46
1N5711
1SS16

HP5082-2835

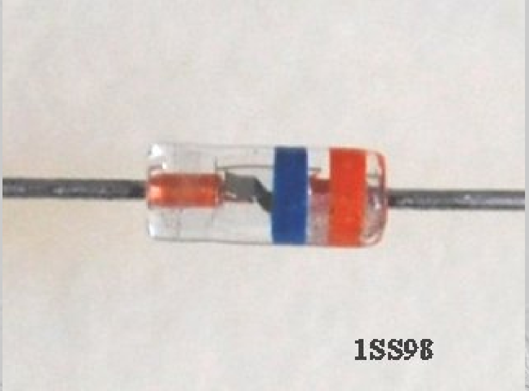
1SS98

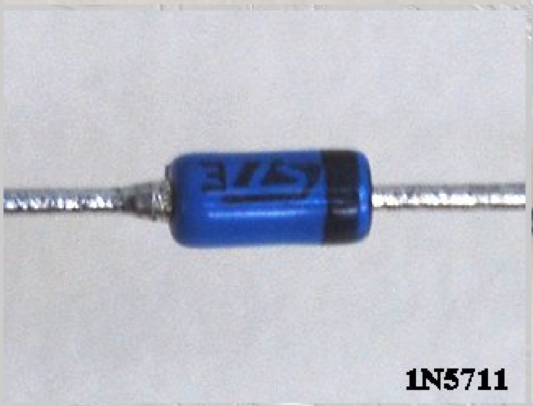
1N5711

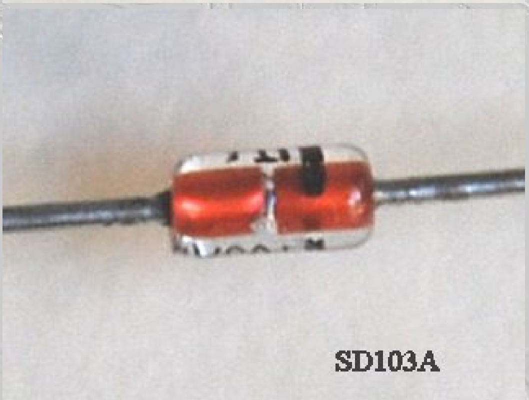
SD103A

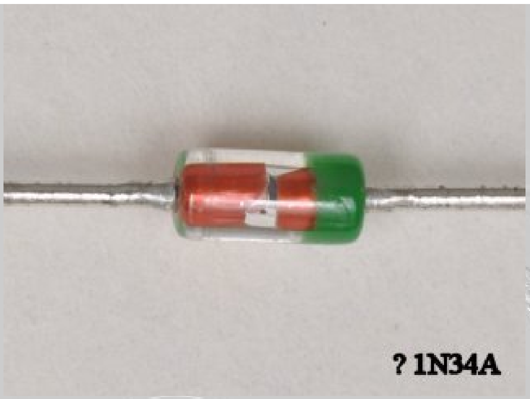

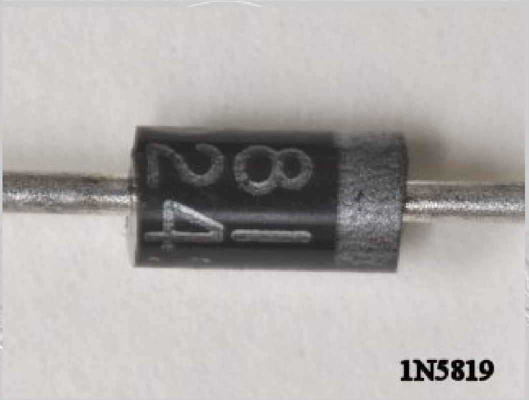

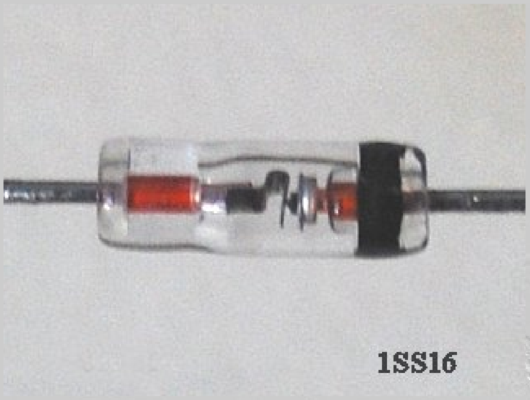

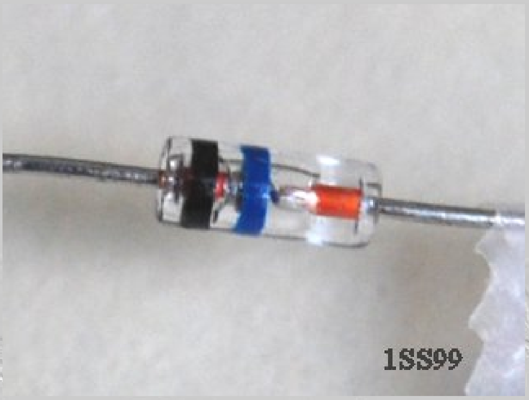

Schottky diodes are very sensitive diodes that work excellently in crystal radios. Their construction and theory are different and I confess to not fully understanding these components. Still, from the characteristic curves, they are excellent! Again, note that you don’t always get what you bargain for. Here I found what was supposed to be 1N34A's to be some sort of Schottky of unknown pedigree. The 1N5819, while having the lowest forward voltage drop, has a rather very high Junction Capacitance and will not perform well. Posts on the RaidoBoard Crystal Radio Forum however highly recommend the 1N5711 for crystal sets although the characteristic curve doesn’t look that fabulous. Many web pages out there on Shottky diodes, I recommend you do your homework. Of all the schottky's, I note that Ben Tongue recommends most highly the HP5083-2835. The high resistance Ro makes them useful for DX sets with very very high Q tanks. Even so the diodes need to be paralleled with up to 4 or 5 diodes to match correctly the Rd with the tank Rp. I have found these somewhat hard to find and expensive, especially when one requires using several in parallel. I recently measured a few 1SS98 diodes and I discover them to have characteristics extremely similar to the HP and I feel they deserve more attention. Sadly, on ebay they seem to be just as difficult to find and expensive. No free ride!

Spreadsheet for calculation of Schottky diode n and Is (modified from Mike Tuggle's spreadsheet)

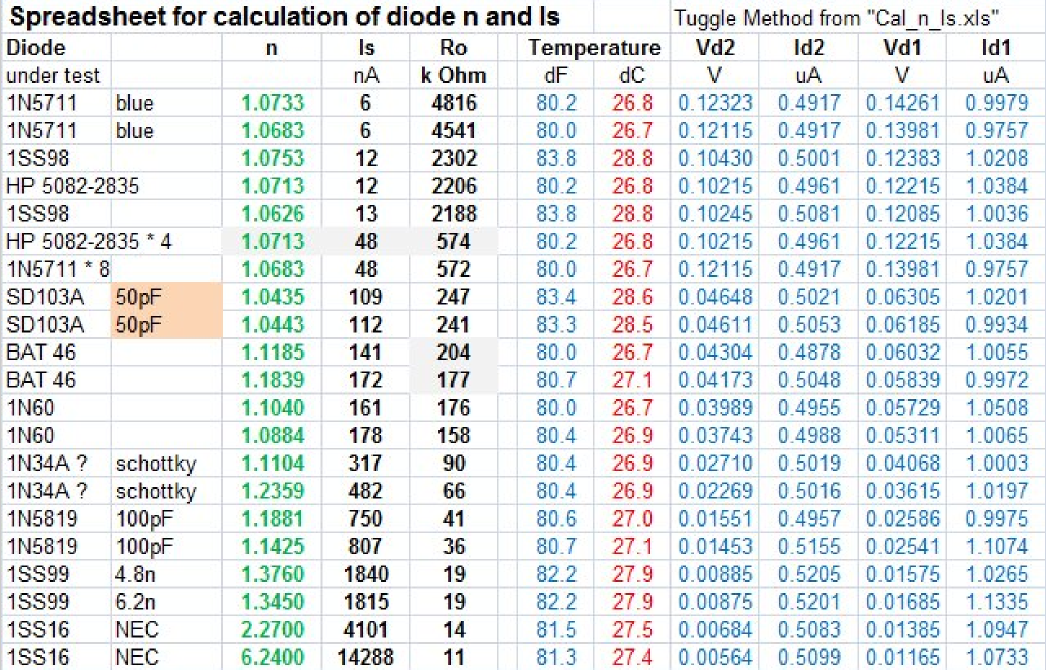

Spreadsheet for calculation of diode n and Is Tuggle Method from "Cal_n_Is.xls"

Diode		n	Is	Ro	Temperature		Vd2	Id2	Vd1	Id1
under test			nA	k Ohm	dF	dC	V	uA	V	uA
1N5711	blue	1.0733	6	4816	80.2	26.8	0.12323	0.4917	0.14261	0.9979
1N5711	blue	1.0683	6	4541	80.0	26.7	0.12115	0.4917	0.13981	0.9757
1SS98		1.0753	12	2302	83.8	28.8	0.10430	0.5001	0.12383	1.0208
HP 5082-2835		1.0713	12	2206	80.2	26.8	0.10215	0.4961	0.12215	1.0384
1SS98		1.0626	13	2188	83.8	28.8	0.10245	0.5081	0.12085	1.0036
HP 5082-2835 * 4		1.0713	48	574	80.2	26.8	0.10215	0.4961	0.12215	1.0384
1N5711 * 8		1.0683	48	572	80.0	26.7	0.12115	0.4917	0.13981	0.9757
SD103A	50pF	1.0435	109	247	83.4	28.6	0.04648	0.5021	0.06305	1.0201
SD103A	50pF	1.0443	112	241	83.3	28.5	0.04611	0.5053	0.06185	0.9934
BAT 46		1.1185	141	204	80.0	26.7	0.04304	0.4878	0.06032	1.0055
BAT 46		1.1839	172	177	80.7	27.1	0.04173	0.5048	0.05839	0.9972
1N60		1.1040	161	176	80.0	26.7	0.03989	0.4955	0.05729	1.0508
1N60		1.0884	178	158	80.4	26.9	0.03743	0.4988	0.05311	1.0065
1N34A ?	schottky	1.1104	317	90	80.4	26.9	0.02710	0.5019	0.04068	1.0003
1N34A ?	schottky	1.2359	482	66	80.4	26.9	0.02269	0.5016	0.03615	1.0197
1N5819	100pF	1.1881	750	41	80.6	27.0	0.01551	0.4957	0.02586	0.9975
1N5819	100pF	1.1425	807	36	80.7	27.1	0.01453	0.5155	0.02541	1.1074
1SS99	4.8n	1.3760	1840	19	82.2	27.9	0.00885	0.5205	0.01575	1.0265
1SS99	6.2n	1.3450	1815	19	82.2	27.9	0.00875	0.5201	0.01685	1.1335
1SS16	NEC	2.2700	4101	14	81.5	27.5	0.00684	0.5083	0.01385	1.0947
1SS16	NEC	6.2400	14288	11	81.3	27.4	0.00564	0.5099	0.01165	1.0733

The above spreadsheet is based on measurements of the diodes shown above. For the determination of Is and n, I chose at random two examples from my collection of various diodes, (all schottky in this case) and measured via a modified version of the methodology outlined by Ben Tongue and Mike Tuggle. For this work I have chosen to set Id2 about 0.5uA and Id1 about 1.0uA and then read the needed voltage. This is the inverse of the normal method but the justification is that for any diode regardless of forward voltage drop the measurement is made at the same part of the LOG I vs V characteristic. This allows comparison between all diodes. I have included the actual room temperature in the calculations in order to eliminate this variable as a source of doubt or error. All the measured parameters as well as the determined results are presented in order to facilitate repeatability.

Note in the above data that both the BAT-46 and the 1N60 diodes appear to have excellent properties for crystal radio work. In fact they have very similar parameters overall. Looking at the photos you should also note that the two types of diode look suspiciously exactly alike. The BAT 46 has it correct markings but the 1N60 is unmarked. You never know what you get!

Silicon Diodes

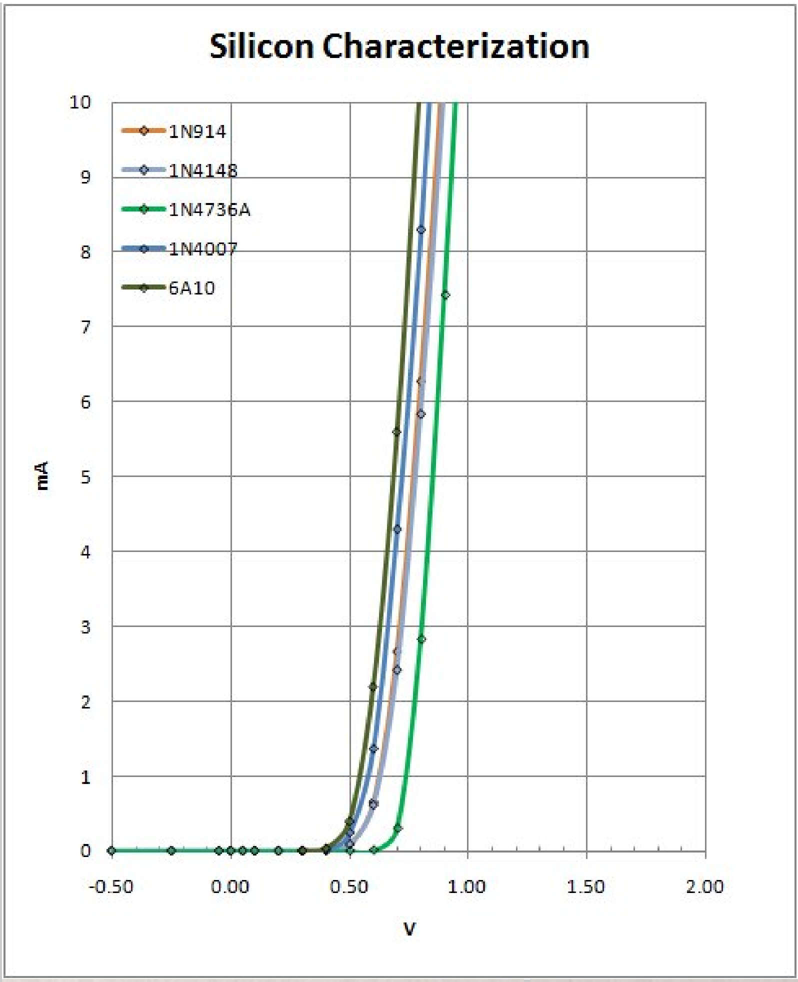

1N914

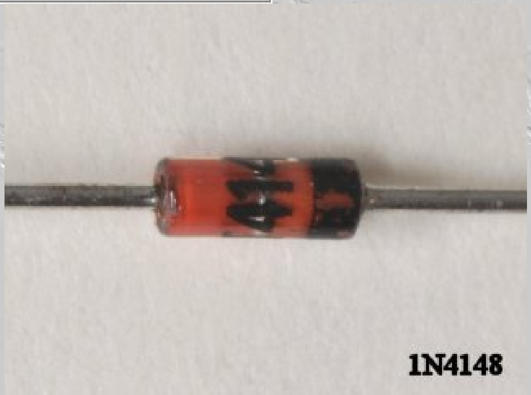
1N4148

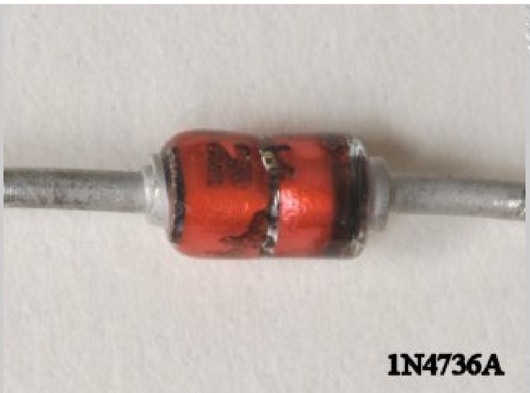

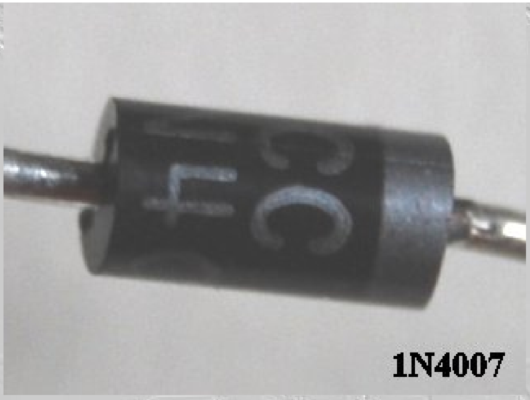

Silicon diodes have good characteristics, but an unacceptably high forward voltage drop making them a very poor choice for crystal radio unless used with bias. The 1N4736A is a Zenner diode.

Spreadsheet for calculation of Silicon diode n and Is (modified from Mike Tuggle's spreadsheet)

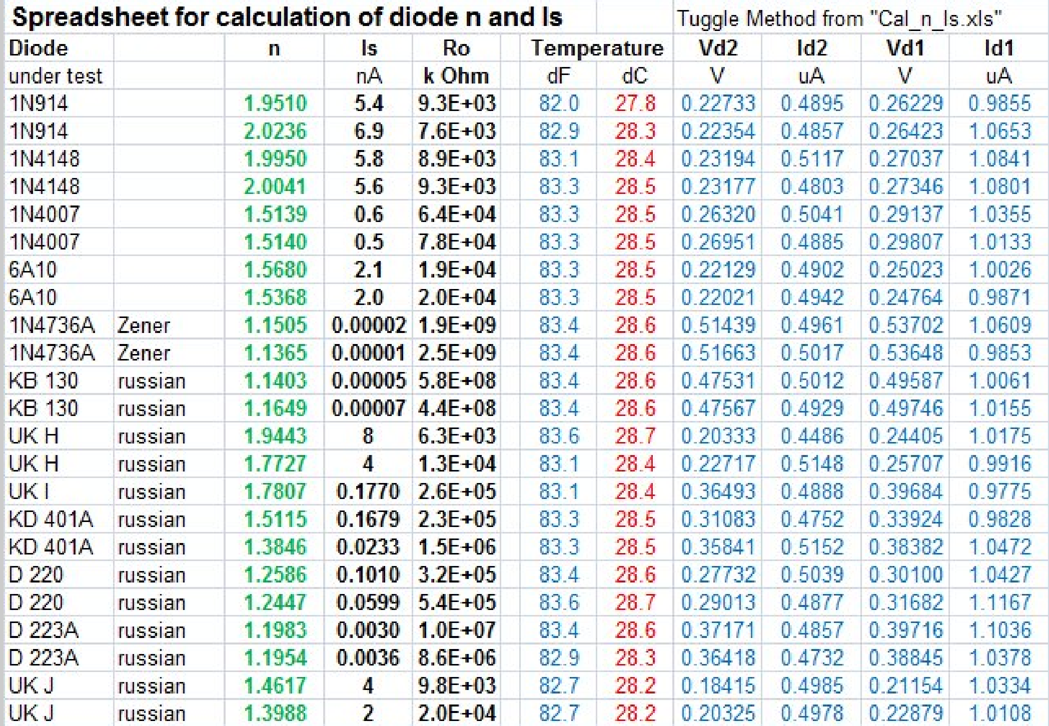

Spreadsheet for calculation of diode n and Is							Tuggle Method from "Cal_n_Is.xls"			
Diode		n	Is	Ro	Temperature		Vd2	Id2	Vd1	Id1
under test			nA	k Ohm	dF	dC	V	uA	V	uA
1N914		1.9510	5.4	9.3E+03	82.0	27.8	0.22733	0.4895	0.26229	0.9855
1N914		2.0236	6.9	7.6E+03	82.9	28.3	0.22354	0.4857	0.26423	1.0653
1N4148		1.9950	5.8	8.9E+03	83.1	28.4	0.23194	0.5117	0.27037	1.0841
1N4148		2.0041	5.6	9.3E+03	83.3	28.5	0.23177	0.4803	0.27346	1.0801
1N4007		1.5139	0.6	6.4E+04	83.3	28.5	0.26320	0.5041	0.29137	1.0355
1N4007		1.5140	0.5	7.8E+04	83.3	28.5	0.26951	0.4885	0.29807	1.0133
6A10		1.5680	2.1	1.9E+04	83.3	28.5	0.22129	0.4902	0.25023	1.0026
6A10		1.5368	2.0	2.0E+04	83.3	28.5	0.22021	0.4942	0.24764	0.9871
1N4736A	Zener	1.1505	0.00002	1.9E+09	83.4	28.6	0.51439	0.4961	0.53702	1.0609
1N4736A	Zener	1.1365	0.00001	2.5E+09	83.4	28.6	0.51663	0.5017	0.53648	0.9853
KB 130	russian	1.1403	0.00005	5.8E+08	83.4	28.6	0.47531	0.5012	0.49587	1.0061
KB 130	russian	1.1649	0.00007	4.4E+08	83.4	28.6	0.47567	0.4929	0.49746	1.0155
UK H	russian	1.9443	8	6.3E+03	83.6	28.7	0.20333	0.4486	0.24405	1.0175
UK H	russian	1.7727	4	1.3E+04	83.1	28.4	0.22717	0.5148	0.25707	0.9916
UK I	russian	1.7807	0.1770	2.6E+05	83.1	28.4	0.36493	0.4888	0.39684	0.9775
KD 401A	russian	1.5115	0.1679	2.3E+05	83.3	28.5	0.31083	0.4752	0.33924	0.9828
KD 401A	russian	1.3846	0.0233	1.5E+06	83.3	28.5	0.35841	0.5152	0.38382	1.0472
D 220	russian	1.2586	0.1010	3.2E+05	83.4	28.6	0.27732	0.5039	0.30100	1.0427
D 220	russian	1.2447	0.0599	5.4E+05	83.6	28.7	0.29013	0.4877	0.31682	1.1167
D 223A	russian	1.1983	0.0030	1.0E+07	83.4	28.6	0.37171	0.4857	0.39716	1.1036
D 223A	russian	1.1954	0.0036	8.6E+06	82.9	28.3	0.36418	0.4732	0.38845	1.0378
UK J	russian	1.4617	4	9.8E+03	82.7	28.2	0.18415	0.4985	0.21154	1.0334
UK J	russian	1.3988	2	2.0E+04	82.7	28.2	0.20325	0.4978	0.22879	1.0108

The above spreadsheet is based on measurements of the diodes shown above. For the determination of Is and n, I chose at random two examples from my collection of various diodes, (all silicon in this case) and measured with a modified methodology to that outlined by Ben Tongue and Mike Tuggle. In this case, with the radically different forward voltage drop of these diodes from Germanium or Schottky diodes, I have kept the Id values constant (about Id2=0.5uA and Id1=1.0uA) and varied Vd. I have included the actual room temperature in the calculations in order to eliminate this variable as a source of doubt or error. All the measured parameters as well as the determined results are presented in order to facilate repeatability.

Light Emitting Diodes

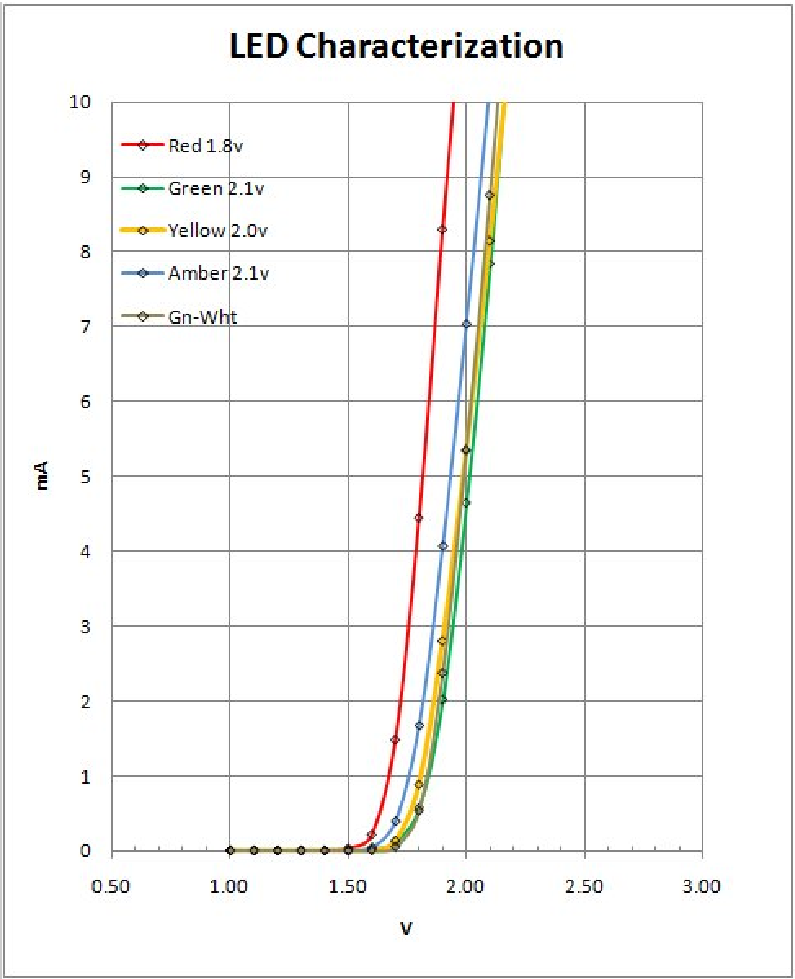
LED Characterization
10
9
8
7
6
5
4
3
2
1
0
mA
Red 1.8v
Green 2.1v
Yellow 2.0v
Amber 2.1v
Gn-Wht
0.50
1.00
1.50
2.00
2.50
3.00
V

Red 1.8v

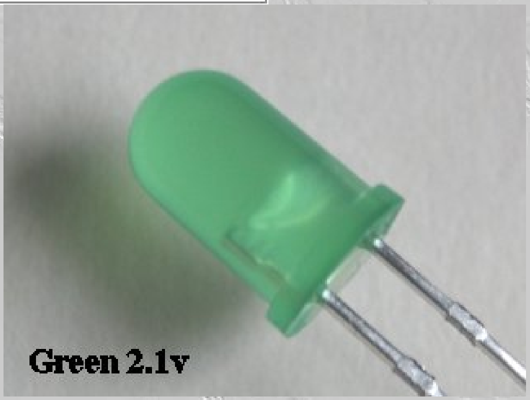
Green 2.1v

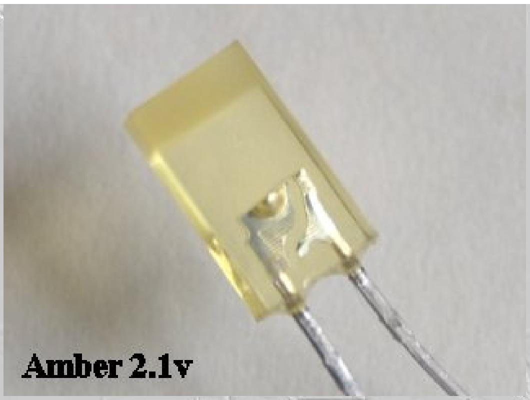

As long as I am measuring various and sundry diodes, I figure I ought to include that most ubiquitous of modern diode, the LED. Found everywhere, these diodes are rapidly becoming the low-energy light source of choice for many lighting applications. I have read occasionally of someone asking whether these ought to be useful for radio applications as well. To this question the answer is generally a resounding "No!". The turn-on voltage is waaay above any reasonable value expected to be delivered by an antenna to a crystal set. Still, this simple answer avoids the actual question, what in fact does the characteristic curve of a LED really look like? Where is the turn-on voltage with respect to the published junction voltage, (assuming you can find that).
To provide just such a look I visited my local electronics store and bought a small handful of LED's, most with the junction voltage listed and took them home to measure. Typical LED junction voltages seem to range from about 1.8v to 2.1v or more. The turn-on voltages look closer to 1.6v-1.7v. Anyone used to working with Carborundum crystals or silicon diodes will be used to biasing their rectifier to get good sensitivity. These LED's once on have a very sharp rise and with a proper bias should work quite well as detector diodes. As a bonus you will get a sweet glow as well. OK, not as cool as the glow of a vacuum diode, but certainly more sensitive!

Spreadsheet for calculation of Light Emitting Diode n and Is (modified from Mike Tuggle's spreadsheet)

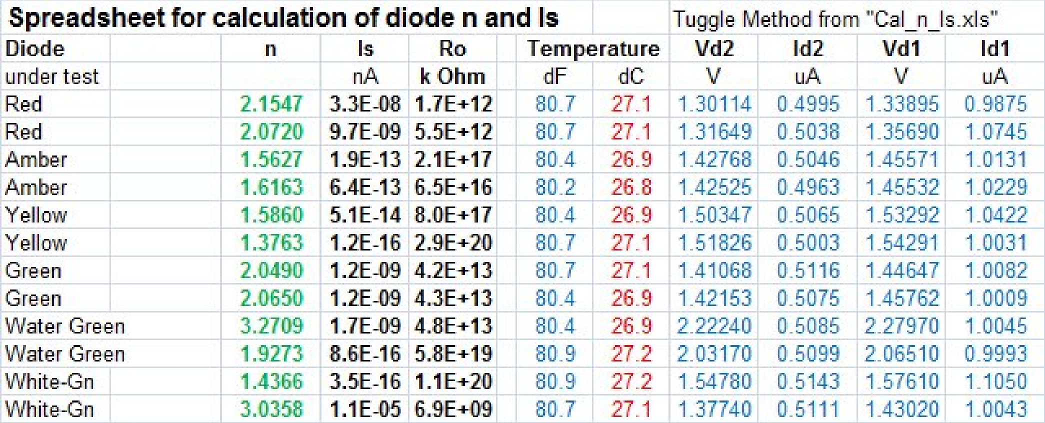

Spreadsheet for calculation of diode n and Is Tuggle Method from "Cal_n_Is.xls"

Diode	n	Is	Ro	Temperature		Vd2	Id2	Vd1	Id1
under test		nA	k Ohm	dF	dC	V	uA	V	uA
Red	2.1547	3.3E-08	1.7E+12	80.7	27.1	1.30114	0.4995	1.33895	0.9875
Red	2.0720	9.7E-09	5.5E+12	80.7	27.1	1.31649	0.5038	1.35690	1.0745
Amber	1.5627	1.9E-13	2.1E+17	80.4	26.9	1.42768	0.5046	1.45571	1.0131
Amber	1.6163	6.4E-13	6.5E+16	80.2	26.8	1.42525	0.4963	1.45532	1.0229
Yellow	1.5860	5.1E-14	8.0E+17	80.4	26.9	1.50347	0.5065	1.53292	1.0422
Yellow	1.3763	1.2E-16	2.9E+20	80.7	27.1	1.51826	0.5003	1.54291	1.0031
Green	2.0490	1.2E-09	4.2E+13	80.7	27.1	1.41068	0.5116	1.44647	1.0082
Green	2.0650	1.2E-09	4.3E+13	80.4	26.9	1.42153	0.5075	1.45762	1.0009
Water Green	3.2709	1.7E-09	4.8E+13	80.4	26.9	2.22240	0.5085	2.27970	1.0045
Water Green	1.9273	8.6E-16	5.8E+19	80.9	27.2	2.03170	0.5099	2.06510	0.9993
White-Gn	1.4366	3.5E-16	1.1E+20	80.9	27.2	1.54780	0.5143	1.57610	1.1050
White-Gn	3.0358	1.1E-05	6.9E+09	80.7	27.1	1.37740	0.5111	1.43020	1.0043

The above spreadsheet is based on measurements of the diodes shown above. For the determination of Is and n, I chose at random two examples from my collection of various diodes, (all led's in this case) and measured with a modified methodology to that outlined by Ben Tongue and Mike Tuggle. In this case, with the radically different forward voltage drop of these diodes from Germanium or Schottky diodes, I have kept the Id values constant (about Id2=0.5uA and Id1=1.0uA) and varied Vd. I have included the actual room temperature in the calculations in order to eliminate this variable as a source of doubt or error. All the measured parameters as well as the determined results are presented in order to facilitate repeatability.

Crystal Galena and Pyrite Diodes

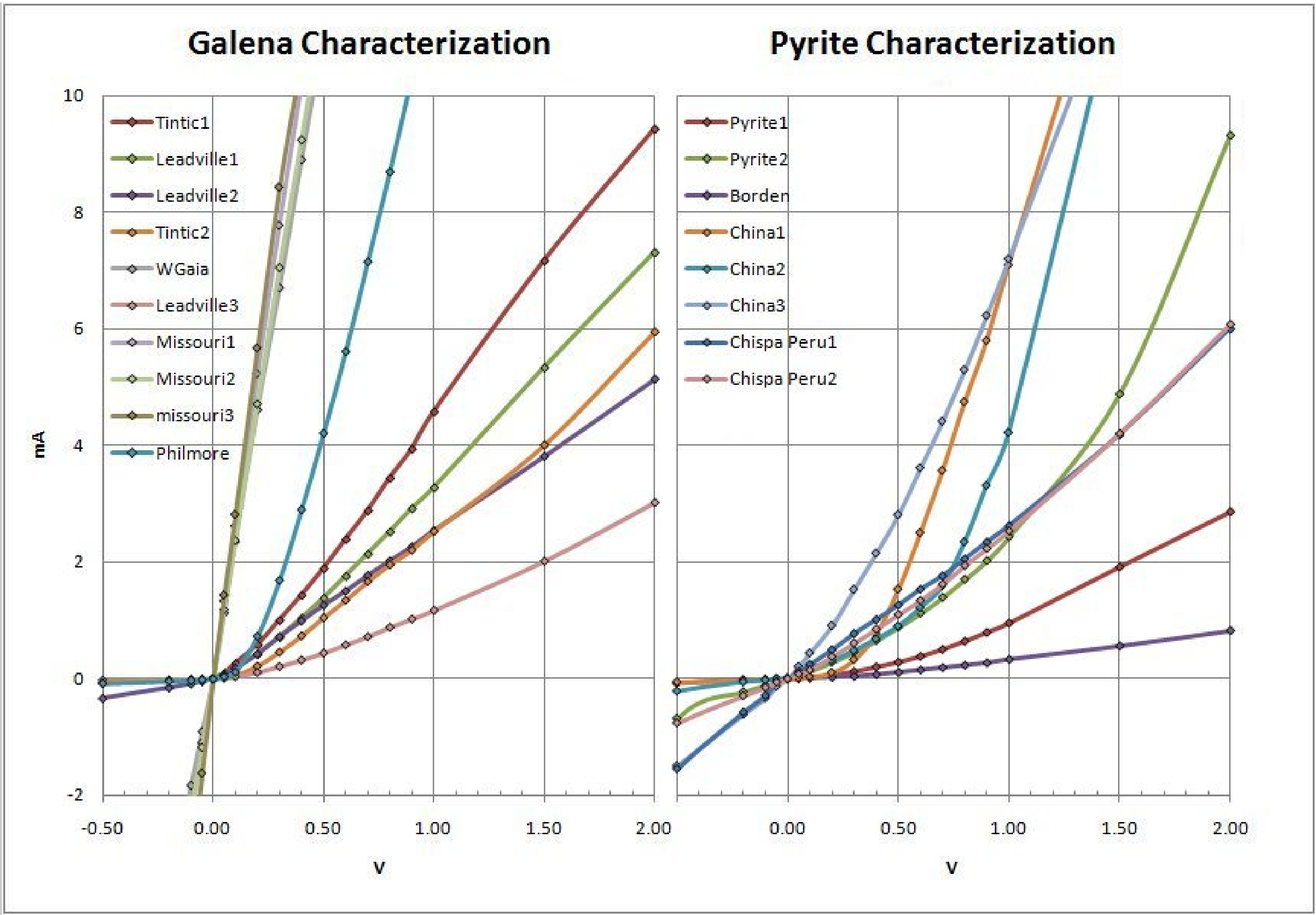

The above curves demonstrate the wide variation in properties and qualities that can be found in natural crystals of galena or pyrite, the two most common and best quality natural stones. In each crystal test I first poked around the crystal for some time to 1) determine a typical sensitivity for the crystal in question and 2) to locate the best hot spot with which to test. This turns out to be a non-trivial exercise on the diode test setup. In a crystal radio one need only listen for the loudest spot. With the test setup one needs test both the forward and reverse current in order to determine if the whisker is on a hot spot or not. Very tedious work! (In retrospect, if I were starting over making a test jig, I would definitely add a DPDT switch to readily change between forward and reverse current measurements. I'd probably toss in a rheostat for fine adjustments as well).
For many crystals there are limited number of possible hot spots, but these may be hot indeed. For most of my "Steel Galena" samples (Tintic Utah, or Leadville Colorado) there are numerous hot spots under virtually every place I touch the probe, but in general the sensitivity is good to so so. These crystals are very kind to work with in terms of finding spots and avoiding frustration. Mirror galena on the other hand may have quality hot spots, but any hot spots at all are rare and frustratingly difficult to locate. Here my Philmore detector crystal shines with an almost ideal "Galena" response. To chase down this rabbit I purchased some lovely mirror galena from Sweetwater Missouri. I broke off a few appropriate-size chunks to pot in woods metal and test. At first I was very excited with the high currents I was seeing at moderate voltages. Figured I had struck gold. When these crystals failed miserably to rectify anything in my radios, I re-measured things in both forward and reverse directions. These crystals obey Ohm's law and act like typical resistors, not suitable for radio work at all.
For my pyrite crystals the work has been especially tedious and frustrating. With one of the crystals one "hot spot" alternated, entirely on its own, between hot and bad while I was making the measurements. I would start over and over, sometimes getting interesting readings then suddenly it would drop to low values and I'd start over, back and forth. I present this data as best as I have measured, and I don't intend to go back! You see at least one of the crystals, my "China 1", (from a lead/silver mine in Hunan) gave a sweet classic-looking curve. More to the point, "ideal" curves for natural minerals are difficult to come by. Most crystals you use will be less than ideal. The good news is that, while listening to your set, poking about for a good spot is far easier than what I have gone through to produce these curves. Your ear will take care of you!

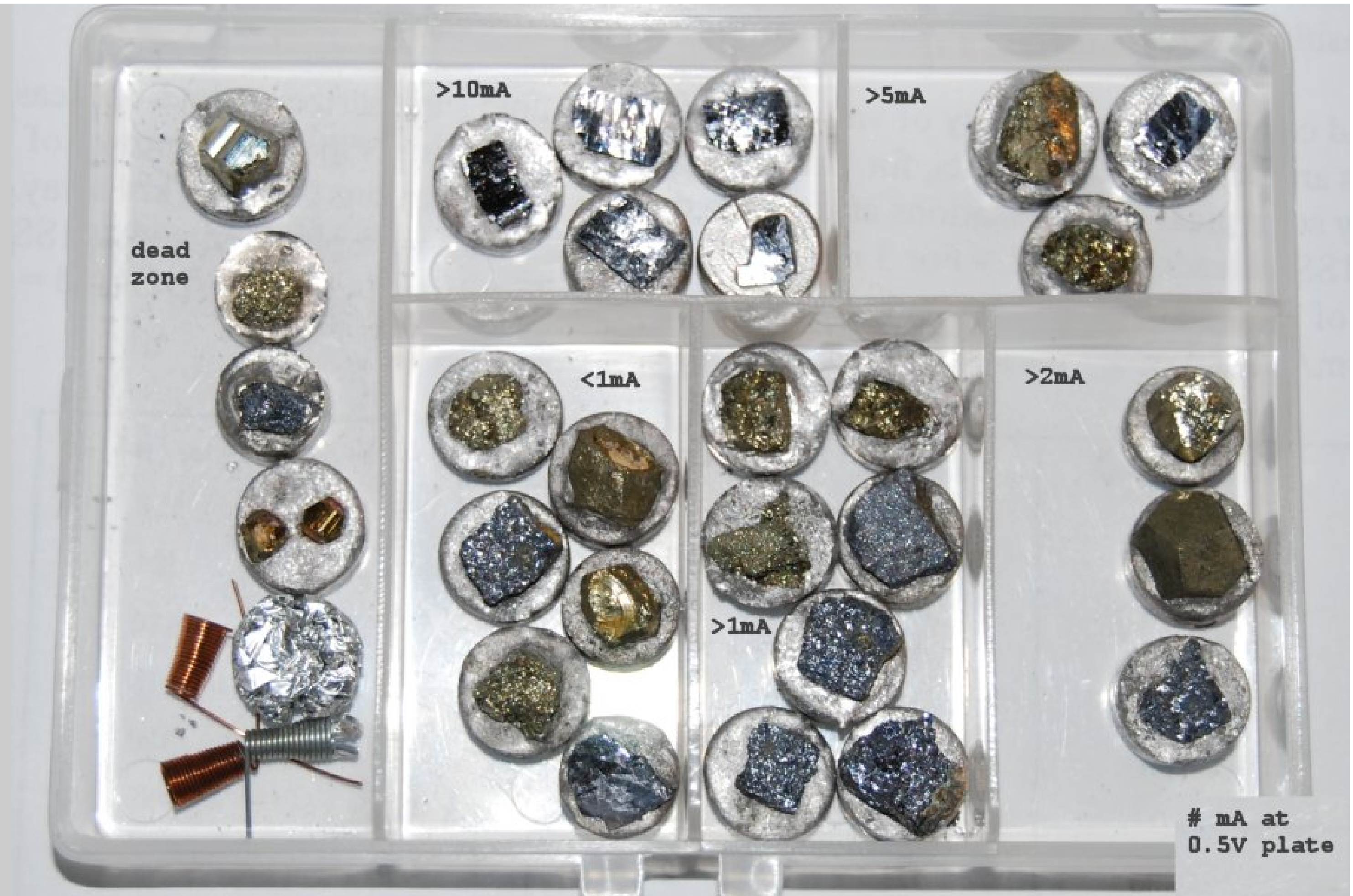

In the photo I indicate groupings based on an easy measure of performance. I note the current in milliamps for each set where the plate voltage is set at 0.5V. The greater the current the better are your chances to get a sensitive crystal, assuming Ohm's Law is not followed! Crystals in the "dead zone" on the left will have their woods metal re-melted for new detector crystals and the bad ones tossed, its tough love for crystals. I find that easily half the potted crystals I make are tossed this way, and only a few can be considered superlative.

My measurements and discussion on vacuum diodes has moved to my section "Diodes III" under the Lab Menu.

Return..

Back to Main....

Sunday, 09-Jul-2017 00:50:33 CDT

Comments and/or Correspondance to: kj-smith@sbcglobal.net

**.::. Sitemap .::. International Pix .::. Western US Pix .::. Pinhole Pix .::. Crystal Radio .::.
.::. Snapshots .::. Thoughts .::. Beer Labels .::. Petrology .::. Life List .::. Ionosphere.::.**

all about
CRYSTAL SETS

by
Charles Green

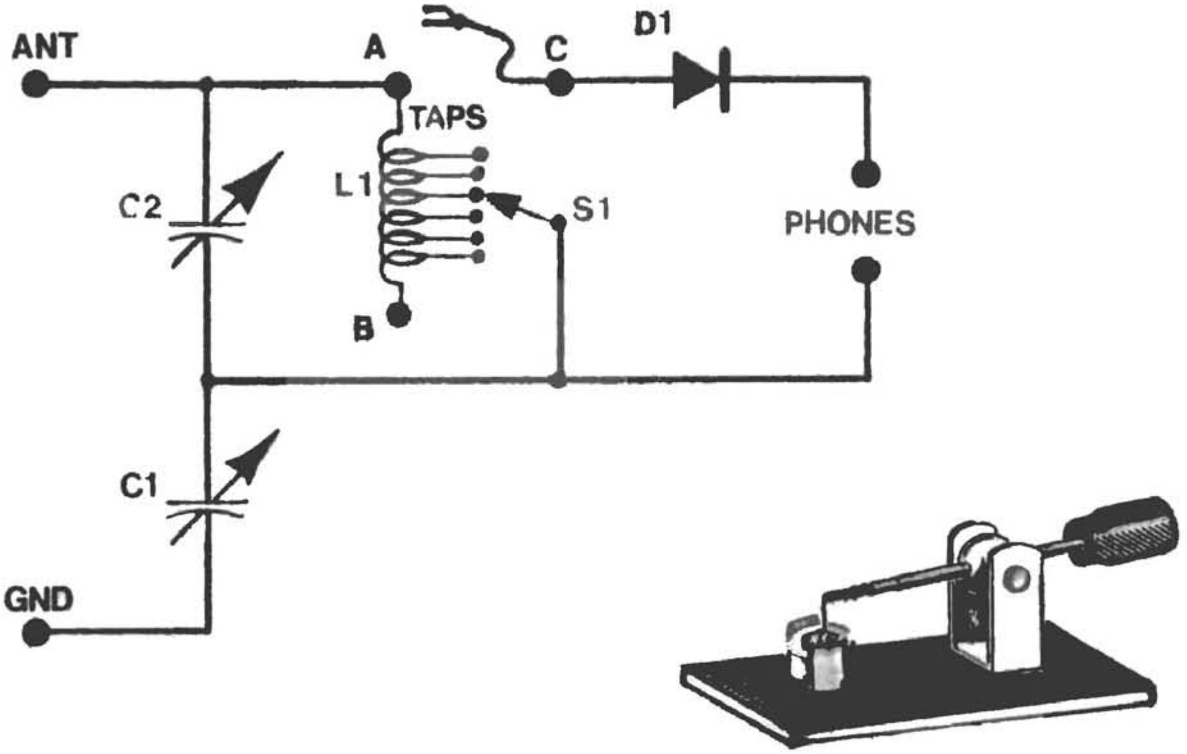
ANT
A
C
D1
TAPS
L1
C2
S1
PHONES
B
C1
GND

How-To-Build Simple Crystal Set Radios

621.3841
G795A
c.1

allabout books no. 101

all about
CRYSTAL SETS

by

Charles Green

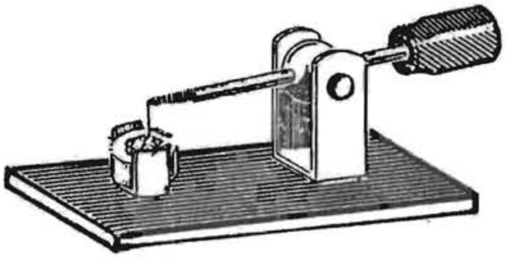

How – To – Build Simple Crystal Set Radios

FIRST EDITION

All rights reserved, including the
right of reproduction in whole or
in part in any form.

Copyright © 1984 by Charles Green

Published by Allabout Books
P.O. Box 22366
San Diego, CA 92192-2366

ISBN 0-930003-00-4

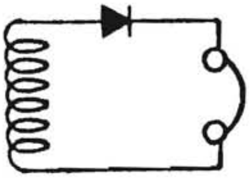

PREFACE

In the beginning of this century, when radio was called "Wireless", most of the early experimenters used a crystal set to receive radio signals.

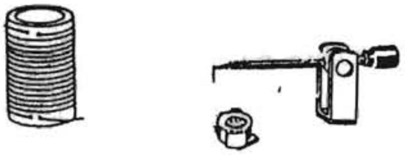

The crystal set is still the simplest type of radio receiver. A catwhisker (made from a thin wire) touching a sensitive spot on a piece of galena or silicon mineral crystal and a tuning coil, form the basic parts of this radio.

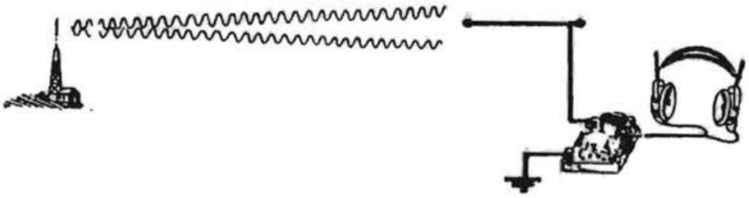

Radio signals from a nearby station are picked up via antenna and ground connections to the crystal set. The galena or silicon crystal detects the signals and the resultant audio is heard with headphones. No battery or electrical power is required.

This book tells you about crystal sets and shows you how to build them.

Charles Green

CONTENTS

1

THE CRYSTAL DETECTOR....................................4
How-it-Works, 4
Detection of Radio Waves, 6
Tuning the Crystal Set, 7
Capacitor, 8
RF Coil, 8
Tuned Circuit, 9
Antenna and Ground, 10
Headphones, 11

2

SIMPLE RECEIVER PROJECTS...............................12
Simple Crystal Set, 12
Slider Tuned Crystal Set, 17
Ferrite Coil Crystal Set, 23
Spiderweb Coil Crystal Set, 27

3

ADVANCED RECEIVER PROJECTS.............................32
Dual-Tuned Crystal Set, 32
Selective Crystal Set, 37
Carborundum Crystal Set, 43

4

COMPONENT PROJECTS.....................................48
Book Tuning Capacitor, 48
Rotary Switch, 53

5

WORKSHOP...54
Soldering, 54
Project Construction Tips, 55
Mounting Minerals, 56

INDEX..57

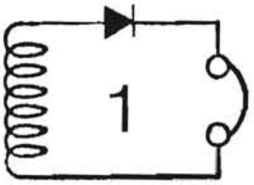

THE CRYSTAL DETECTOR

A crystal detector of radio waves can be either a traditional catwhisker and natural mineral crystal, or a modern semiconductor diode.

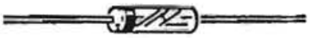

SEMICONDUCTOR DIODE

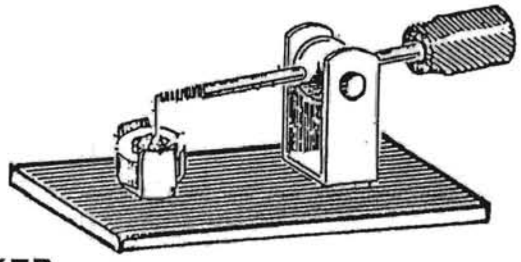

CATWHISKER DETECTOR

Certain natural minerals such as galena (lead sulphide) and silicon have electrically sensitive surface spots. In the beginning of this century, experimenters discovered that electrical current would flow into the sensitive spot on a mineral crystal from a catwhisker, but would not pass readily (high resistance) from the crystal to the catwhisker. This unique feature was used to detect (rectify) radio waves and extract the audio modulation.

HOW-IT-WORKS

There are two general types of crystal detctors;

. Point contact diode
. Junction diode

Both types use semiconductor material such as silicon, germanium, or galena. The basic semiconductor junction diode is made of two portions of the material "doped" in manufacturing by adding controlled amounts of chemicals. The doped portions have opposite electrical properties:

"P" TYPE, with an excess of positive electrical charges (holes).
"N" TYPE, with an excess of negative electrical charges (electrons).

Junction Diode

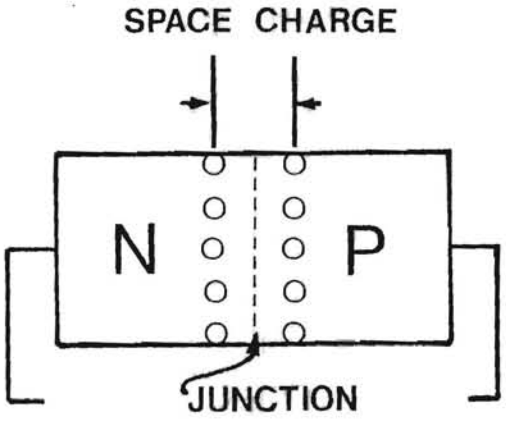

The boundary between the two doped portions is called a PN junction. The area on both sides of the PN junction is called the space charge region. When a positive voltage is applied to the "P" material (and negative voltage to the "N" material), the space charge region becomes narrow and electron current flows through the PN junction.

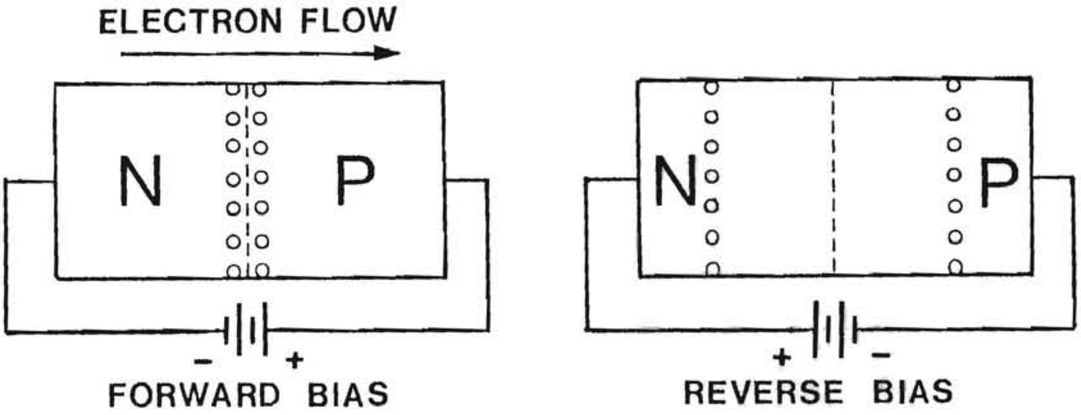

When the voltage polarities are reversed, negative to "P" and positive to "N", the space charge region widens (increasing the effective resistance of the PN junction) and the electron flow is very small (back current). If an ac voltage is applied to the PN junction, it will act as a rectifier (or radio wave detector).

Point Contact Diode

The point contact diode uses a "catwhisker" of thin wire contacting the surface of doped "N" semiconductor material. The pointed end of the wire produces a high intensity electric field in the area of the semiconductor surrounding the wire. When an electric current flows from the wire to the semiconductor material, a small region of "P" material forms around the area of the point contact. Thus a PN junction is formed which operates in the same way as the junction diode.

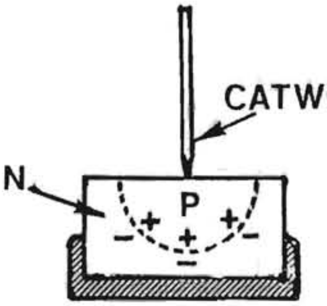

The sensitive spots of galena and silicon crystals used in traditional catwhisker detectors have naturally doped areas of semiconductor material, and operate in the same way as the point contact detector.

DETECTION OF RADIO WAVES

At the transmitting radio station, a microphone converts sound waves into electrical audio signals. The audio signals then modulate the transmitted radio waves.

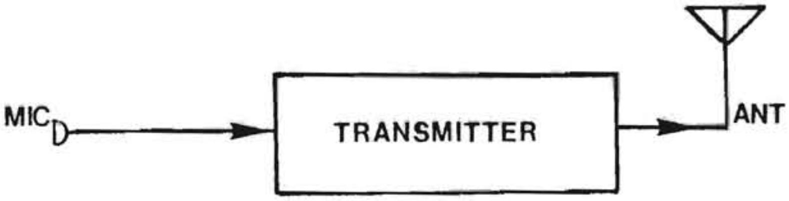

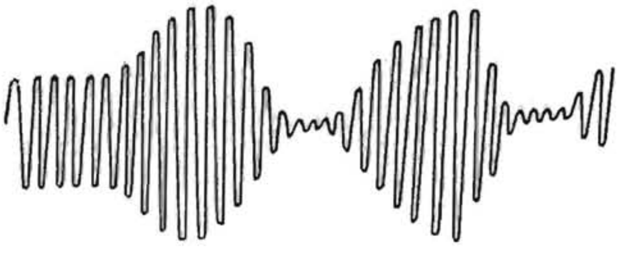

MODULATED RADIO WAVES ⟶

the crystal detector

The transmitted radio waves are intercepted by the receiving antenna and small rf currents are induced in the antenna wire. The rf currents (high frequency ac) flow through the detector to ground. As previously explained, the crystal detector will pass current in only one direction, therefore, pulsating dc voltages (detected audio signals) are developed by the crystal equivalent to the audio modulation of the transmitted radio waves.

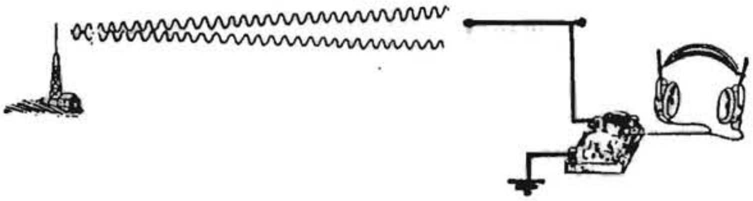

The detected audio signals flow through the headphone coils and affect the intensity of a magnetic field. This causes the steel discs of the headphones to vibrate and produce audible sound waves.

TUNING THE CRYSTAL SET

The antenna intercepts all radio waves in the area, and the signals will be heard jumbled up in the earphones. A nearby radio station may be received with the other stations heard in the background.

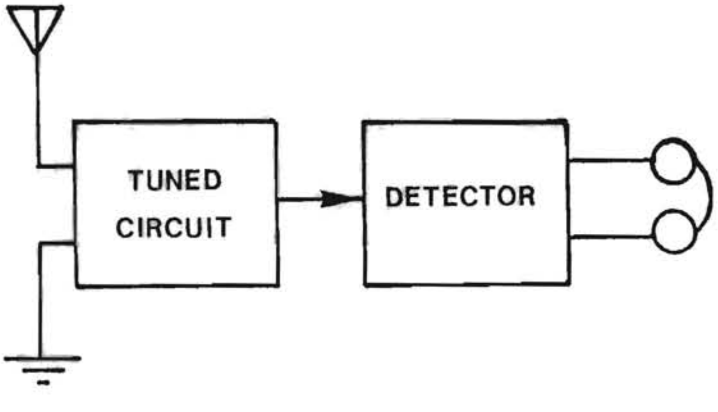

In order to receive a radio station properly, the crystal set must be tuned to the frequency of the transmitter. A tuned circuit composed of an rf coil (inductance) and a capacitor, is connected between the antenna/ground circuit and the detector.

Capacitor

The modern capacitor is evolved from the old Leyden jar. It consists of two or more metal sections separated by a dielectric (air, paper, or plastic). The capacitor stores energy in the form of an electrostatic charge in its dielectric. A variable capacitor (used in a tuned circuit) has movable metal sections that adjust the amount of capacity.

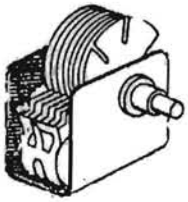

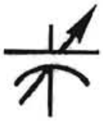

A capacitor will allow pulsating dc current or ac current to flow through it, but will block steady dc current. A capacitor has an oppositon or "reactance" to ac that varies with frequency. The higher the frequency, the lower the reactance of a given value of capacity. Capacity is measured in micro-farads (uF) or micro-micro-farads (picofarads) (uuF or pF).

RF Coil

An rf coil is made of insulated copper wire wound on a plastic or cardboard tube. When electric current flows through the coil, a magnetic field is created. This magnetic field resists any change in electric current flow. This opposition or "reactance" to ac varies with frequency. The higher the frequency, the higher the reactance of a given value of rf coil inductance. Inductance is measured in millihenrys (mH) and microhenrys (uH).

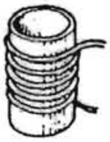

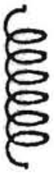

Tuned Circuit

A tuned circuit is composed of a capacitor and an rf coil. As previously described, the reactances of a capacitor and an rf coil change with frequency (but in opposite ways). At a particular frequency, the two reactances will cancel each other out. This condition is called "resonance". The impedance of a parallel tuned circuit is at its highest at resonance. Radio signals at the resonant frequency will be greatly magnified, while signals at other frequencies will be shunted to ground (rejected).

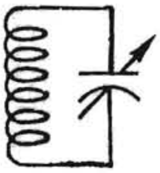

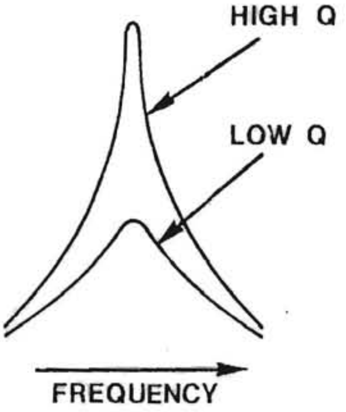

The resonant frequency of a tuned circuit can be changed by adjusting the capacitance (using a variable capacitor),or by varying the inductance (using a slider or taps on the rf coil). A ferrite core can also be used to change the inductance of a coil. The selectivity of a tuned circuit is dependent upon the quality (Q) of the components and the inductance /capacity ratio (L/C ratio). A high-Q tuned circuit is more selective than a low-Q circuit. A tuned circuit with a high L/C ratio (large inductance and small capacity) is more selective than a low L/C ratio circuit.

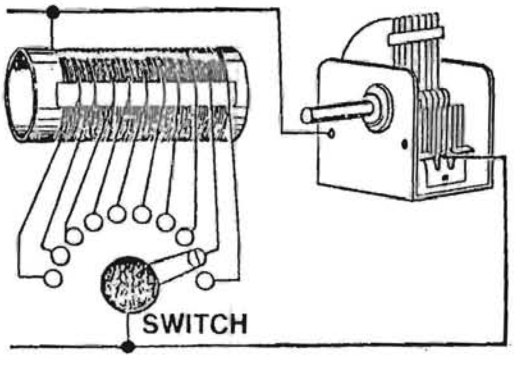

ANTENNA AND GROUND

A simple crystal set depends upon the strength of the radio waves received by the antenna to produce audio in the earphones. If you live in a strong signal area (a nearby radio transmitter), a crystal set can produce loud audio in earphones with only an inside antenna. But in lower signal areas, a good outside antenna with a ground connection is required.

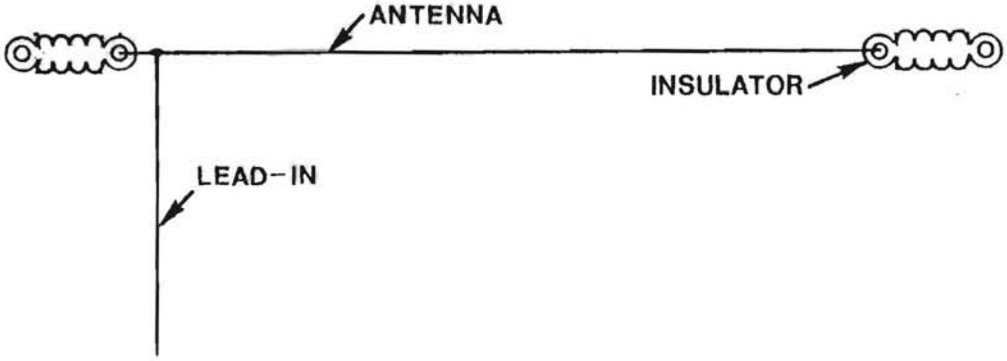

If you are fortunate enough to be able to put up an outside antenna, the basic rule of thumb is the higher, the better reception. Check your local electronics store for an antenna kit. You can construct an antenna using a glass or plastic insulator at each end of a length of copper wire (insulated or bare). Make the antenna as long as possible. It does not have to be in a straight line to work.

CAUTION

ANTENNA INSTALLATION
NEAR ELECTRICAL POWER
LINES IS DANGEROUS.

Make sure that your antenna
is far enough away from any
power line, so that it will
not touch a wire if it falls.

Connect a length of insulated wire between the end of the antenna and your crystal set antenna terminal. This antenna lead-in wire can be fed into your house through a window.

If you can not erect an outside antenna, run a length of insulated wire as long and high as possible through your house or apartment. It will work, but not as well as an outside antenna because of the shielding affect of the house walls.

Connect a ground lead to a cold water pipe. Make sure that it is a metal pipe. Or, you can buy a ground rod from your local electrical supply store and drive it into the ground near your house. Then connect your crystal set ground lead to the rod. For safety install a lightening arrestor between the antenna lead and ground (see the instructions with the device).

HEADPHONES

Headphones (also called earphones, or headset) are required to hear the received audio from a crystal set. The standard headphones are made with dual coils of fine wire connected in series in each phone. Horseshoe magnets are built into each phone case to provide a magnetic field. A disk of thin sheet iron (diaphragm) is mounted near the top of the coils. The phones are connected in series with a set of flexible leads terminated in metal "phone tips".

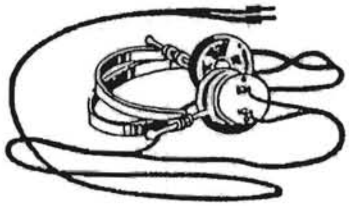

Detected audio signals flow through the coils and affect the intensity of the magnetic field from the horseshoe magnet. This changes the magnetic attraction of the diaphragm, and causes it to vibrate and produce audible sounds.

The total resistance of the series headphone coils is used as a measure of audio sensitivity. A resistance of 2000 ohms is sufficient. More sensitive headphones have a resistance of over 5000 ohms and are used to hear very weak signals.

The 2000 ohm headphones can be obtained at electronic parts mail order houses.

2

SIMPLE RECEIVER PROJECTS

SIMPLE CRYSTAL SET

Look at the schematic diagram. The crystal set uses a tunable ferrite coil L1, and a crystal diode D1. This type of tuning coil is called a "loopstick". The ferrite core is screwed into the coil form and increases the inductance of L1, thus tuning the receiver to a lower frequency.

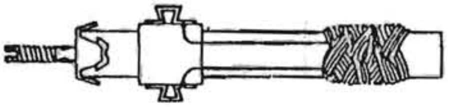

CONSTRUCTION

As shown in the illustration, the crystal set components are mounted on a 5 1/2-inch by 2 1/2-inch by 3/4-inch wood base. Layout and install the crystal set parts in the locations shown. The loopstick L1, is mounted with a metal strip supplied with it. Push the snap-in end of L1 into the large hole of the metal strip. Then bend 1/2-inch of the other end of the strip to form a bracket and mount it on the base with a wood screw. If the metal strip is not available, make one out of sheet metal approximately 2-inches long by 1/2-inch wide. Then drill a hole to fit L1 at one end, and a small hole to fit a wood screw at the other end.

Mount the four fahnestock clips and solder lugs with wood screws in the locations shown in the illustration.

WIRING

Use hookup wire to connect the crystal set components as shown in the schematic and illustration. The loopstick L1 may have three terminals. One of the terminals is a tap, which is not used in this circuit. Make sure that you connect only to the terminals at the start and end of the loopstick coil.

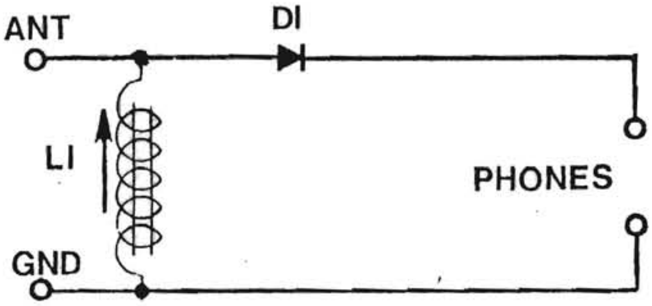
ANT
DI
LI
PHONES
GND

SIMPLE CRYSTAL SET

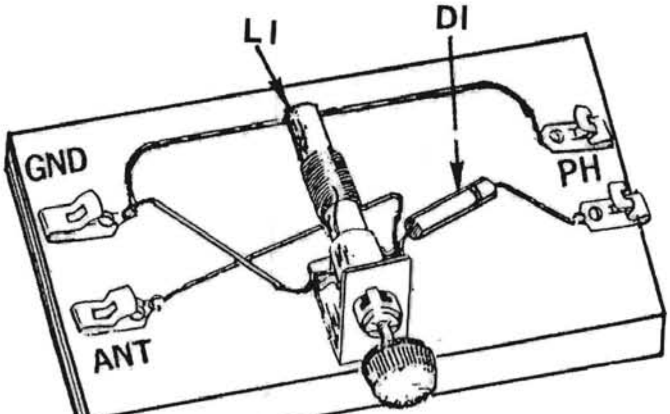
LI
DI
GND
PH
ANT

List of Materials

D1 - Germanium diode
(1N34A or equiv.)
L1 - Loopstick
(see text)
4 - Fahnestock clips
BASE - 5 1/2 X 2 1/2 X
3/4-inch wood section
PHONES - 2000-ohm headphones
MISC. - Solder lugs, wood screws,
knob and bracket to fit L1,
hook-up wire.

Install a small knob on the end of the L1 adjustment screw. Make sure that the screw turns freely with the knob. Position the 1N34A diode as shown before soldering the leads.

Experiment 1

Connect the antenna, ground, and 2000-ohm headphones to the crystal set terminals. Rotate the L1 knob counter-clockwise as far as it will go. Do not turn past this point, or you may break the ferrite rod inside the coil form. The ferrite is now outside of the coil winding. This is the highest frequency point of the tuning range of L1.

Slowly turn the L1 knob clockwise and listen for received radio stations. When the L1 knob is turned fully clockwise, the ferrite will be completly inside the coil (increasing the inductance) and it will be at the lowest frequency point of the tuning range.

Experiment 2

Connect capacitors, with values from 50 mmf to 300 mmf, in parallel with L1. Then tune L1 and note how the frequency range has changed.

AMPLIFIER CONNECTION

The simple Crystal Set (and the other receiver projects) can be connected to an amplifier in place of earphones. This amplifier can be a single channel input of your home stereo.

Parts values are not critical; C1 is approximately .005 mfd, and the volume control R1 can be any value from 1000 ohms to 10,000 ohms. Use shielded wire between R1 and the amplifier input jack.

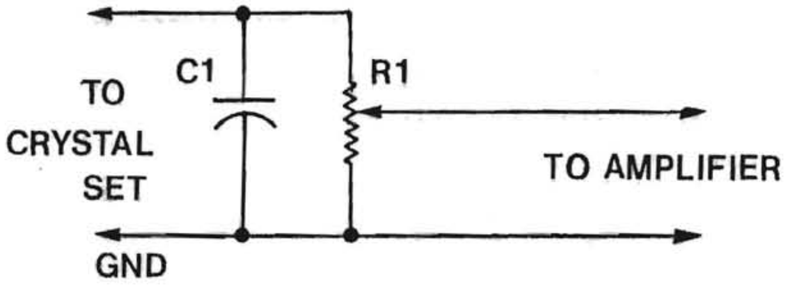
TO
CRYSTAL
SET
GND
C1
R1
TO AMPLIFIER

Amplifier Connection

Schematic Symbols

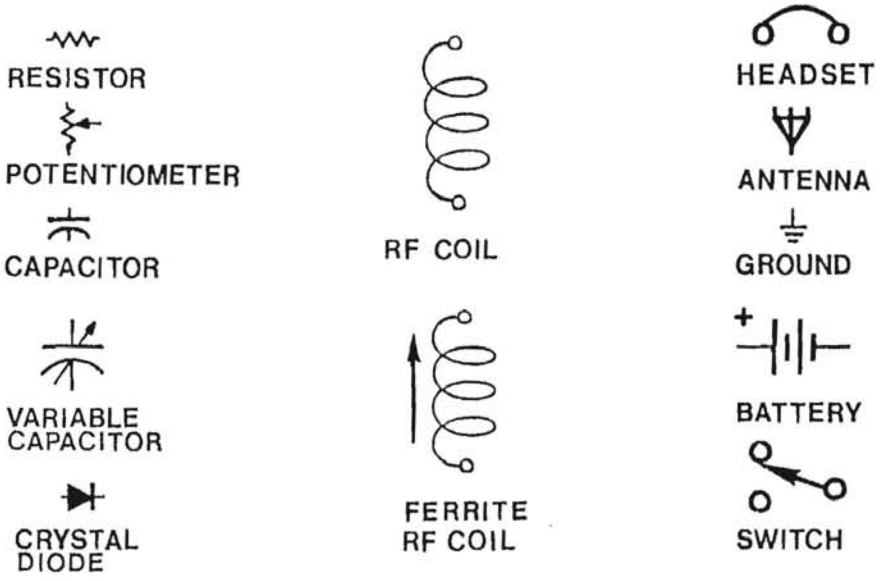
RESISTOR
POTENTIOMETER
CAPACITOR
VARIABLE
CAPACITOR
CRYSTAL
DIODE
RF COIL
FERRITE
RF COIL
HEADSET
ANTENNA
GROUND
+
BATTERY
SWITCH

List of Materials

D1 - Germanium diode, type 1N34A, 1N60, 1N82A, or equiv.
L1 - Slider-tuned coil (see text)
4 - Fahnestock clips (terminals)
BASE - 6 x 3 x 1/2-inch section of wood (see text)
PHONES - 2000 ohms headphones
MISC. - Coil form (see text), No. 24 enameled copper wire, 3/32 inch dia. brass rod, brass ball slider (see text), solder lugs, two small brackets, hook-up wire, wood screws, machine screws and nuts.

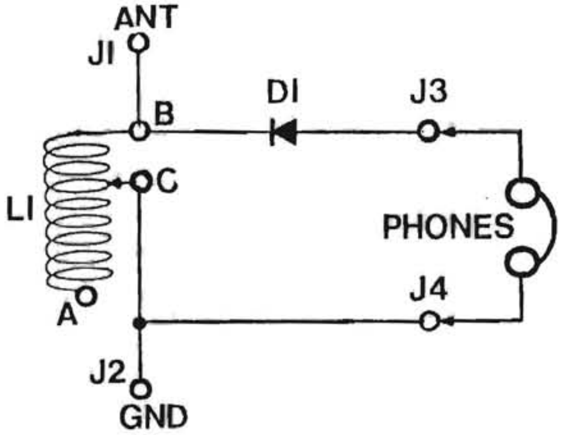
ANT
JI
B
DI
J3
C
LI
PHONES
J4
A
J2
GND

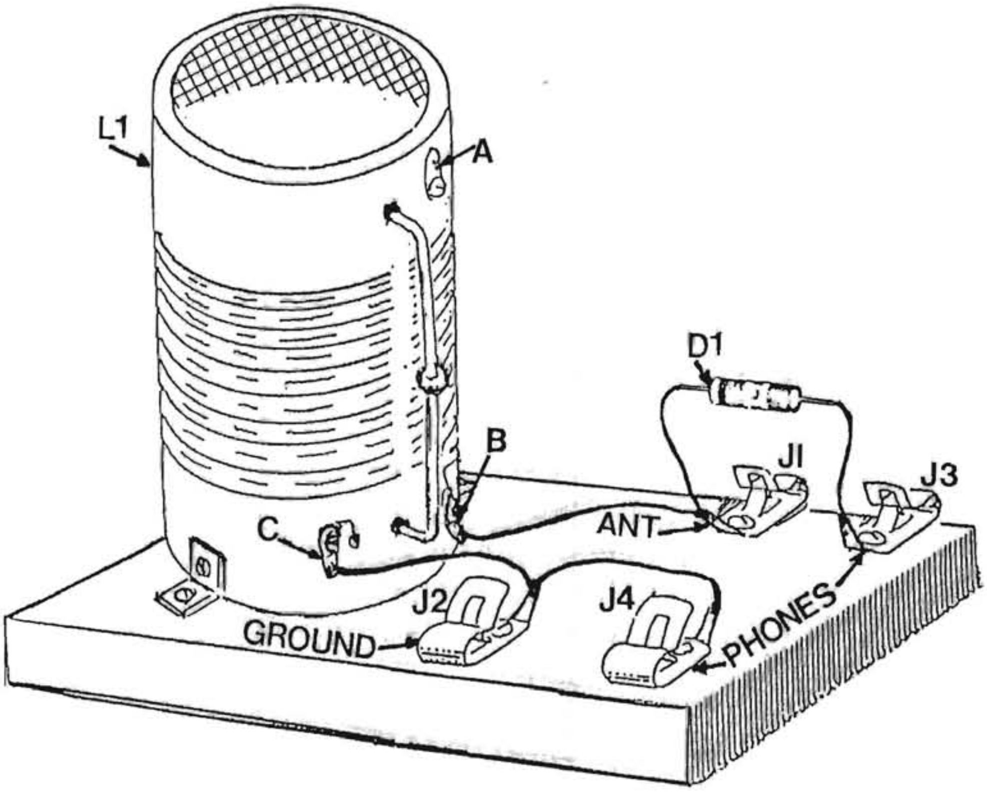
L1
A
D1
B
JI
J3
C
ANT
J2
J4
GROUND
PHONES

SLIDER TUNED CRYSTAL SET

SLIDER TUNED CRYSTAL SET

In the early days of radio, slider tuned coils were used in crystal sets. This method of tuning the inductance uses an adjustable metal contact (slider) that moves over the coil. The slider makes contact with individual coil wires to tune in the radio stations.

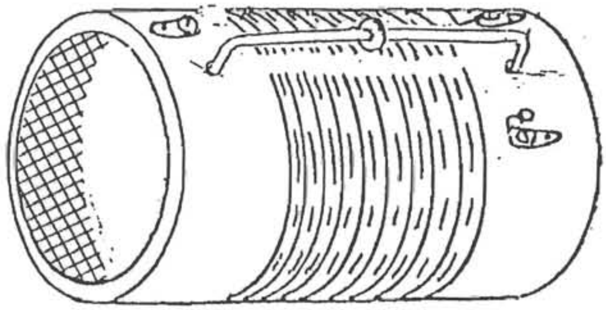

The enamel insulation of the coil wires in the slider path is removed for good electrical contact.

You can experiment with a slider tuned coil by building this crystal set. As shown in the illustration, the coil is mounted upright on a 6-inch by 3 1/2-inch by 1/2-inch wood base. A semiconductor diode is used for a detector.

TUNING COIL CONSTRUCTION

Look at the drawing of the tuning coil L1 construction details on the next page. The tuning coil is wound on a cardboard or plastic mailing tube section, 2 5/8-inches in diameter and 5-inches long. Make sure that you select a rigid, thick-walled tube for a coil form. The tube used in our model has a 3/16-inch wall. The coil form must be rigid to hold the slider ball and rod assembly.

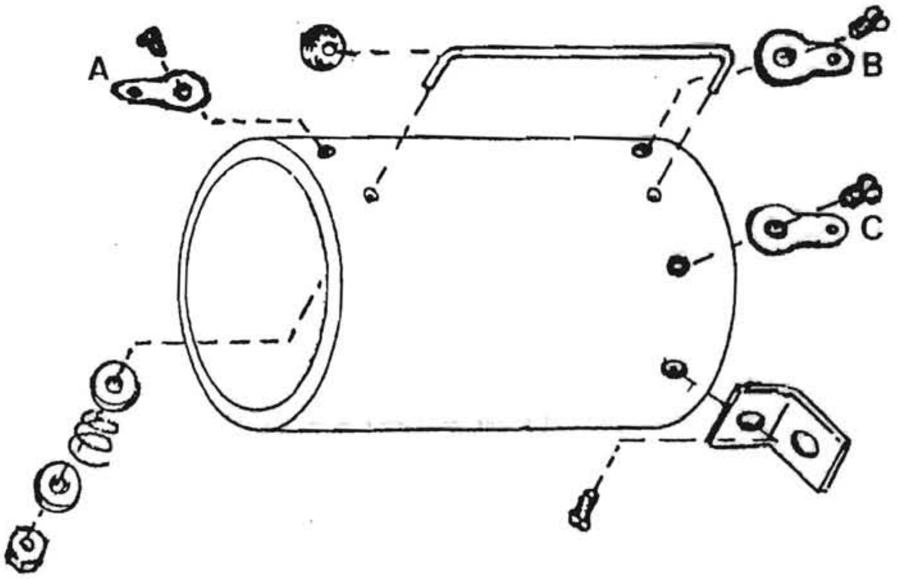

Before winding the coil, drill all the holes in the form as shown in the drawing, and then install the solder lugs. Position the solder lug ends toward the coil form edges.

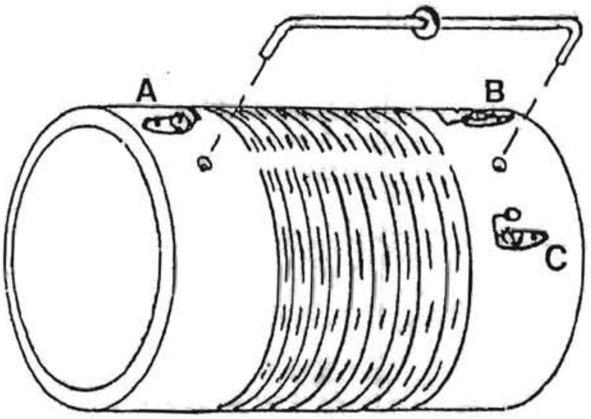

The coil is wound with No. 24 enameled copper wire. Remove the enamel from the wire end with sandpaper and then solder the wire to the lug. Start winding the coil, keeping the turns as tight and close together as possible. Wind the wire to the front of the bottom solder lug B. Remove the enamel and solder the wire end to the lug.

A brass ball approximately 3/8-inches in diameter is used as the moving contact slider. This brass ball can be found at your local electrical supply store. It is a type of ball that is used as a top nut for small table lamps. Drill a hole through the ball so that it will slide over a length of 3/32-inch diameter brass rod. Bend the brass rod at each end (after placing the ball in the center of the rod), to fit through the holes in the coil form as shown in the drawing. Cut the rod 1-inch down from the bend at each end.

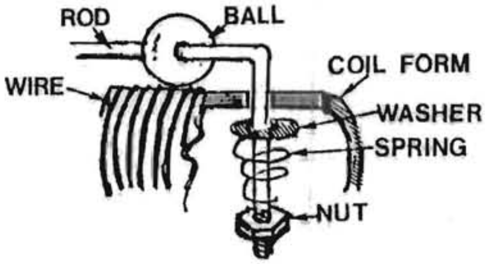

Insert the brass rod with the slider ball into the coil form holes. Then place a small washer and spring on one end of the rod (inside the coil form). Solder a nut on the end of the rod to hold the spring in place. Do the same to the other end of the rod. After soldering, slide the ball over the wires and check that it slides smoothly. Use sandpaper to remove the enamel insulation of the wires in the path of the ball slider. Connect a wire between the soldered nut on the bottom end of the brass rod and the solder lug near the rod.

BASEBOARD

All of the crystal set components are mounted on a wood baseboard as shown in the drawings. Cut the baseboard to size; a 6-inch by 3 1/2-inch section of 1/2-inch thick wood (maple is used in our model). Sand the edges after cutting. It is not necessary to paint or stain the baseboard.

Mount the four terminals and solder lugs as shown in the drawing. Label the four terminals with either rub-on lettering or cemented paper tags. Install two brackets on opposite sides of the bottom of the coil, and mount it on the baseboard with the lugs toward the terminals.

WIRING

Bare No.22 wire was used to wire our model, but any kind of hookup wire can be used. Connect the wires between the solder lugs as shown in the drawing, except do not connect the antenna terminal lug wire to the coil terminal at this time. Connect the crystal diode between the solder lugs as shown.

Experiment 1

This experiment will show how an untuned crystal detector functions.

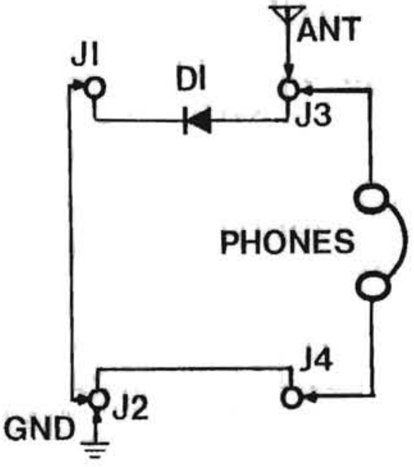

Connect the untuned detector circuit as shown in the drawing, and then hook-up your antenna and ground leads. Connect 2000-ohm headphones to the circuit and listen to the detected radio waves.

You will probably hear a strong local radio station, with other, weaker radio station transmissions all jumbled together.

Experiment 2

Reconnect the baseboard wires to the circuit of the Slider-Tuned Crystal Set as shown in the drawing. Listen to the detected radio waves as you tune the coil by sliding the brass ball over the sanded wire path. You should be able to receive the radio stations much better than the untuned detector circuit of Experiment 1.

LOADING COIL

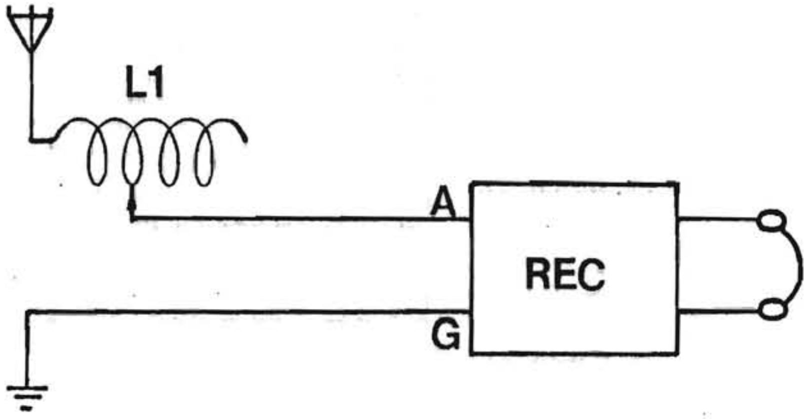

An additional slider tuned coil, L1, can be used as a Loading Coil to increase the electrical length of your antenna. This may allow better reception of the lower broadcast band frequencies. Connect L1 in series with the antenna lead to the receiver, and tune it for best reception.

WAVE TRAP

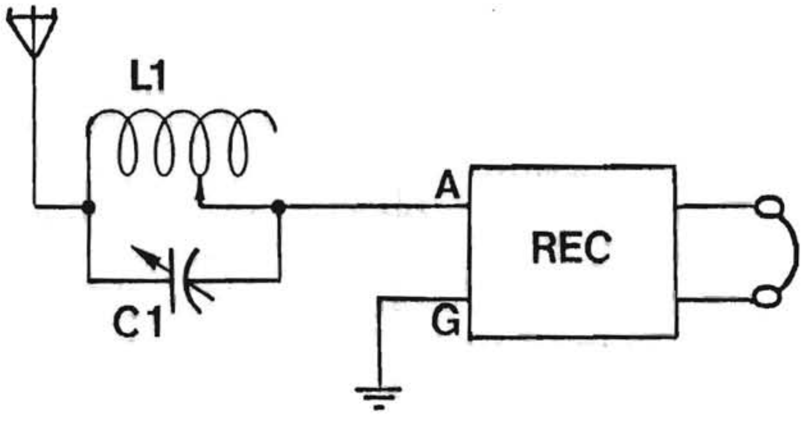

L1 can also be used in a Wave Trap circuit to tune out a strong local signal, and allow easier tuning of weaker stations. C1 is a 365 mmf variable capacitor connected in parallel with L1. Adjust L1 and C1 so that the interfering station is at minimum.

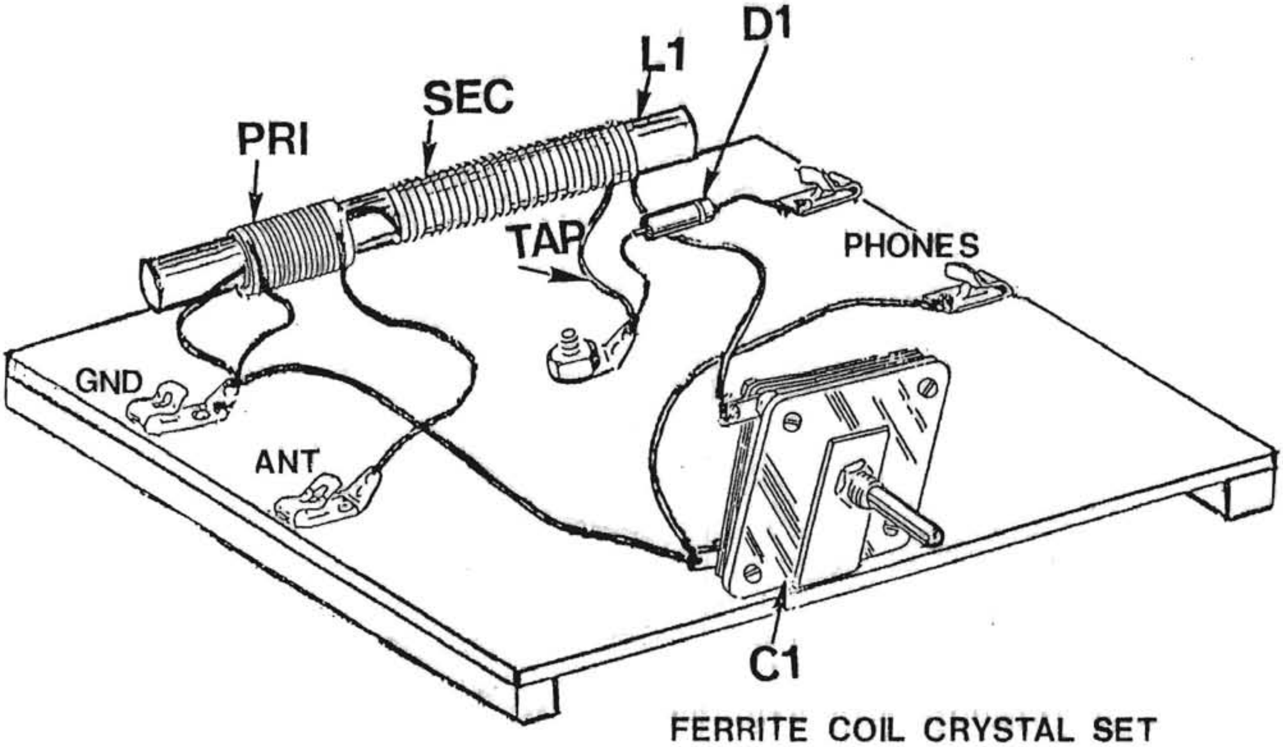
D1
L1
SEC
PRI
TAP
PHONES
GND
ANT
C1

FERRITE COIL CRYSTAL SET

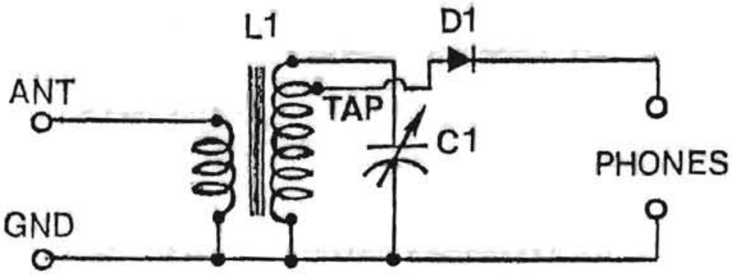
L1
D1
ANT
TAP
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Germanium diode (1N34A or equiv.).
L1 - Ferrite coil (see text).
4 - Fahnestock clips.
2 - wood sections; 6 X 1/2-inch square (see text).
BASE - 8 1/2 X 6-inch sheet plastic (see text).
PHONES - 2000-ohm headphones
MISC. - plastic tube section for L1 primary winding (see text), solder lugs, machine screws and nuts, wood screws, bracket for C1 (see text), knob for C1, hookup wire.

FERRITE COIL CRYSTAL SET

This crystal set has a large ferrite core coil L1 of a type that is normally used for a built-in antenna in portable radios. This type of coil has a very high "Q" and may be more selective than an air-core coil. As you can see in the schematic, a primary coil is wound around the ferrite coil. The original winding is used as the secondary.

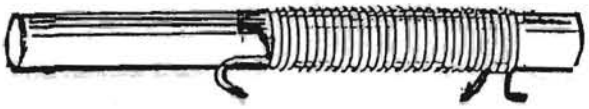

The primary coil (connected to the antenna and ground) can be moved close to, or away from, the secondary (original winding connected to D1). This variable coupling is used to control the selectivity of the crystal set.

PRIMARY COIL

The primary coil is wound on a 1-inch section of plastic tube that can fit loosely over the ferrite rod. A 1/2-inch diameter plastic tube is used on the receiver shown in the illustration. Wind 12-turns of hookup wire evenly over the tube and twist the ends together to prevent movement. Leave 4-inch leads from the primary coil winding. Slide the primary coil over the center of L1.

BASEBOARD

The crystal set components are mounted on a 8 1/2-inch by 6-inch section of sheet plastic. The thickness of the plastic is not critical, but it must be rigid enough to hold the parts without bending. A 6-inch long 1/2-inch square wood section is installed under each end of the baseboard. A section of wood can be used for the baseboard with wood screws to mount the parts, in place of machine screws and nuts.

The model shown uses a plastic variable tuning capacitor, but an air type capacitor can be used for C1 instead. Make a metal bracket approximately 1-inch wide by 1 1/4-inch high and a 1/2-inch right angle mounting foot from sheet aluminum. Drill a hole in the top of the bracket to fit C1 and two holes in the mounting foot to fit small machine screws. Locate the components on the baseboard as shown in the illustration and then drill mounting holes. If you have a cardboard or plastic mounting supplied with ferrite coil L1, you can mount it on the baseboard. If not, L1 can be taped loosely on top of the baseboard. Th ferrite coil L1 in our crystal set was obtained from an old radio. Install the components on the board with solder lugs, machine screws, and nuts.

WIRING

Connect the parts with hookup wire as shown in the illustration. Use the schematic diagram as a guide. Make sure that the ground lead from L1 secondary passes inside the primary coil. Also check that the primary coil slides freely over the L1 winding. Install a knob on the C1 shaft, and mark the terminal designations on the baseboard.

Experiment 1

Hookup the antenna, ground and 2000-ohm headphones to the crystal set terminals. Slide the primary coil over the center of the L1 winding. Tune C1 for radio stations. Then slide the primary coil towards the ground end of L1 while tuning C1. Note the difference in selectivity.

Experiment 2

Change the number of turns of the primary winding and repeat Experiment 1.

Operating Tips

After operating a crystal set for some time, you may notice a difference between day, evening, and night reception of radio stations. This is due to the two ways that radio signals travel between the transmitting antenna and a receiving antenna.

(a) The ground wave; which moves along the surface of the earth.
(b) The sky wave; which travels upward and is reflected by the ionosphere.

The reception of broadcast stations is dependent upon the time of day. During daylight hours, stations are received primarily by ground wave transmission. At night, the broadcast stations will be received via sky wave transmission. Also, some radio stations change to a lower power at night. Reception in the evening may be more difficult as the fading ground wave is interfered with by reflected waves. But later, it may be possible to hear far-away stations transmitted by sky wave.

Your geographical location to radio stations is very important for reception, as a crystal set receiver does not amplifiy signals. It is entirely dependent upon the strength of the radio waves. If you are in a deep valley surrounded by high mountains, reception will be difficult. If you live in a flat plain, with no obstacles to radio transmission, reception may be very good.

An outside antenna is necessary for for best reception. If possible, try erecting an extra antenna at a right angle to the original one, and switch it in to provide a directional effect for increased signal strength. Keeping a log of received stations for a week may be helpful in determining the radio reception conditions in your area.

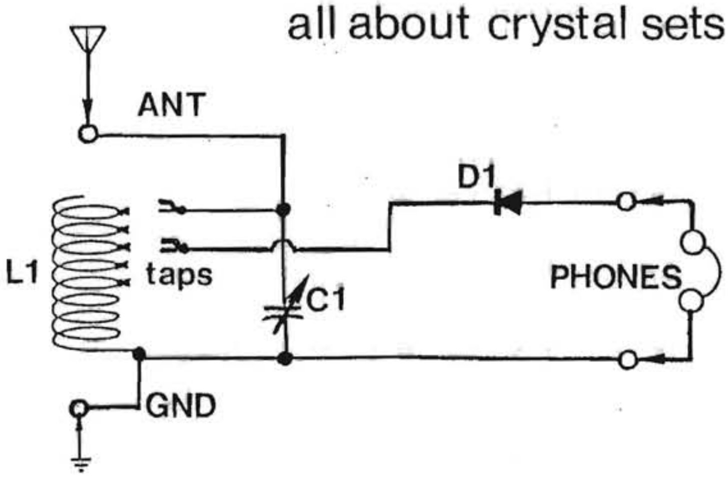
ANT
D1
L1
taps
C1
PHONES
GND

List of Materials

C1 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector (or 1N34A or equiv. diode).
L1 - Spiderweb coil (see text).
4 - Fahnestock clips.
2 - small alligator clips
BASE - 8 X 6 X 1/2-inch section of wood (see text).
MTG. POST - 5 x 3/4 x 3/8-inch wood section (see text).
PHONES - 2000 ohms headphones.
MISC. - Coil form (see text), No. 22 enameled copper wire, solder lugs, hook-up wire, wood screws, machine screws, and nuts.

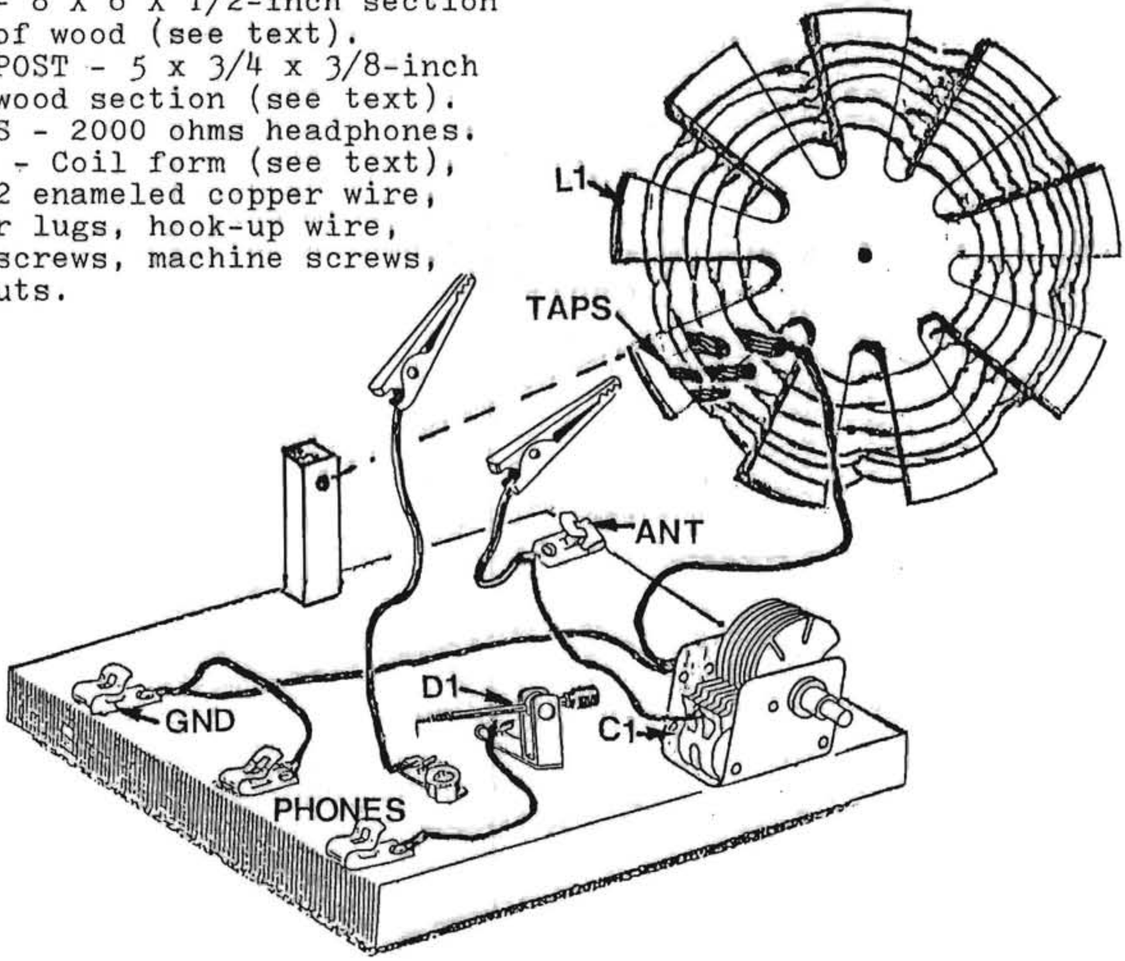
L1
TAPS
ANT
D1
C1
GND
PHONES

SPIDERWEB COIL CRYSTAL SET

SPIDERWEB COIL CRYSTAL SET

The spiderweb coil is a tuning coil that is wound on a flat wheel form that has an uneven amount of spokes. This type of coil was also used as a loop antenna in early radios.

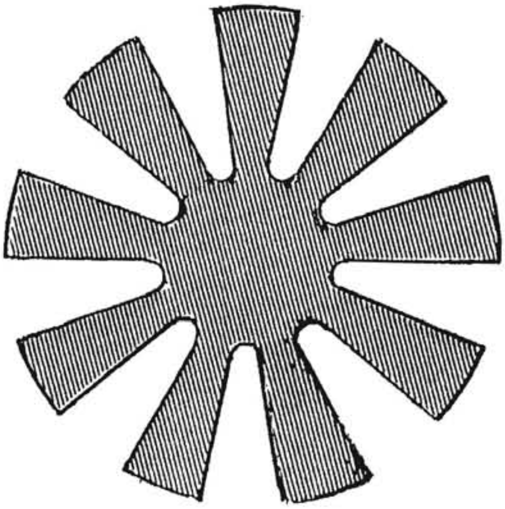

Our receiver uses a tapped spiderweb coil that is tuned by a variable capacitor. As shown in the illustration, the spiderweb coil is mounted at the rear of a 8-inch by 6-inch by 3/4-inch wood base. A catwhisker crystal detector and a variable capacitor are mounted near the front of the wood base.

SPIDERWEB

The spiderweb coil is wound on a 6-inch diameter flat form that has nine vanes. Make a tracing of the spiderweb coil form template on the next page. Then use carbon paper to transfer the tracing to a section of 1/8-inch fibreboard or sheet plastic. Cut the coil form out with a jeweler's saw, and drill a center hole to fit a small wood screw.

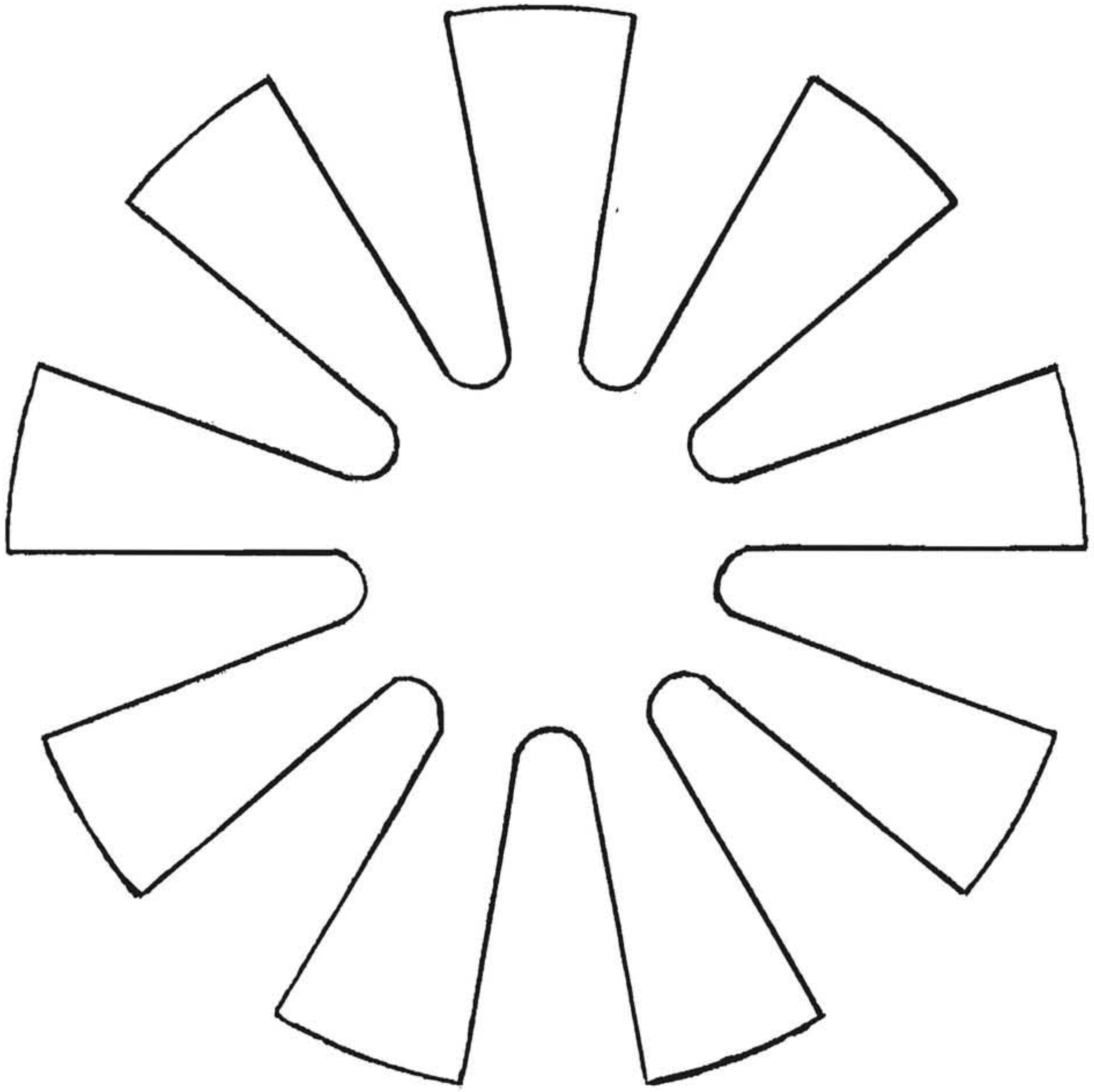

Spiderweb Coil Form Template

Wind the coil with No. 22 enameled copper wire, starting from the center and winding outward. Leave about 5-inches of the starting end of the wire free. It will be used as a connecting lead. Feed the wire behind a vane and over the front of the next vane as you wind. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Wind the wire to about 1/4-inch from the end of a form vane. Then drill a small hole near the vane end, and bend a solder lug through it. Solder the wire end to the lug.

On one of the vanes, count down five wires and then carefully lift it outward. Remove about 1/4-inch of the enamel with a section of sandpaper and then bend the end of a solder lug aound the bared wire. Solder the lug to the wire, then cut off remaining portion of the lug (away from the bentover end). Repeat this every five turns until you have a total of six lug taps on the coil. Stagger the lugs away from each other to keep them from shorting.

BASEBOARD

As shown in the illustration, the Spiderweb Coil Crystal Set parts are mounted on a wood baseboard. Cut the 8-inch by 6-inch baseboard from a section of 1/2-inch thick wood. White pine was used in our model. Sand the surface of the wood after cutting.

Make the spiderweb coil mounting post by cutting out a 5-inch by 3/4-inch by 3/8-inch wood section. Then cut a notch in the center of the rear of the baseboard to fit the wood section. Install the wood section in the notch with small wood screws.

Mount the parts on the baseboard as shown in the illustration with small wood screws. If the mounting holes of the tuning capacitor C1 are on the bottom, drill matching holes in the baseboard and use flathead machine screws to hold the capacitor.

Install solder lugs with the four terminals, and mount the spiderweb coil to the mounting post with a wood screw through a center hole in the coil form. The coil should be positioned with the lugs facing forward (on the left side of the coil form).

WIRING

Connect the crystal set components with hookup wire as shown in the illustration. Use 6-inch lengths of insulated stranded wire to the two small alligator clips.

Experiment 1

Connect antenna, ground, and 2000-ohm headphones to the crystal set. Then connect the D1 alligater clip lead to the antenna terminal. Set the tuning capacitor C1 to minimum capacity (rotor blades all the way out). Connect the antenna clip lead to the tap on L1 near the center of the form.

Adjust the crystal detector D1 catwhisker to a sensitive spot and tune C1 for a radio station. It may be easier to tune for stations if you temporarily connect a crystal diode (1N34A or equiv.) in place of the catwhisker crystal detector.

After a station is heard, you can disconnect the diode and probe for a sensistive spot on the crystal with the catwhisker. Change the antenna clip lead to other L1 taps and continue tuning C1 for radio stations.

Experiment 2

Disconnect the D1 clip lead and reconnect it to a tap on L1 close to the center of the coil. This will load L1 more lightly than if the detector was connected to the top of the coil. Repeat the tuning procedure of Experiment 1 and note the difference in selectivity and volume of the received radio stations. Then tune in a radio station and move the D1 clip lead to other taps on L1. There will be a particular tap point for optimum selectivity and volume for each received station.

STATION LOG

DATE	TIME	DIAL	STATION	NOTES

3 ADVANCED RECEIVER PROJECTS

DUAL TUNED CRYSTAL SET

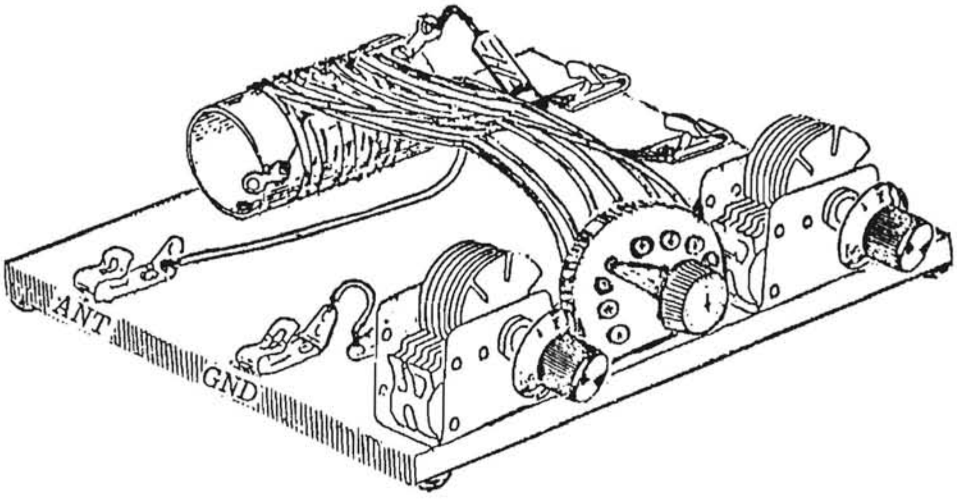

This crystal set circuit uses two variable capacitors C1 and C2. As you can see in the schematic drawing, C1 tunes the ground circuit, and C2 tunes a tapped coil L1. The taps of L1 are connected via the rotary switch S1.

TUNING COIL

Look at the drawing of the coil form construction details on page 34. The tuning coil L1 is wound on a section of 1/8-inch wall plastic tube, 2-inches in diameter, and approximately 4-inches long.

Drill holes in the form and install three solder lugs with machine screws and nuts. Position the solder lug ends towards the coil form ends as shown.

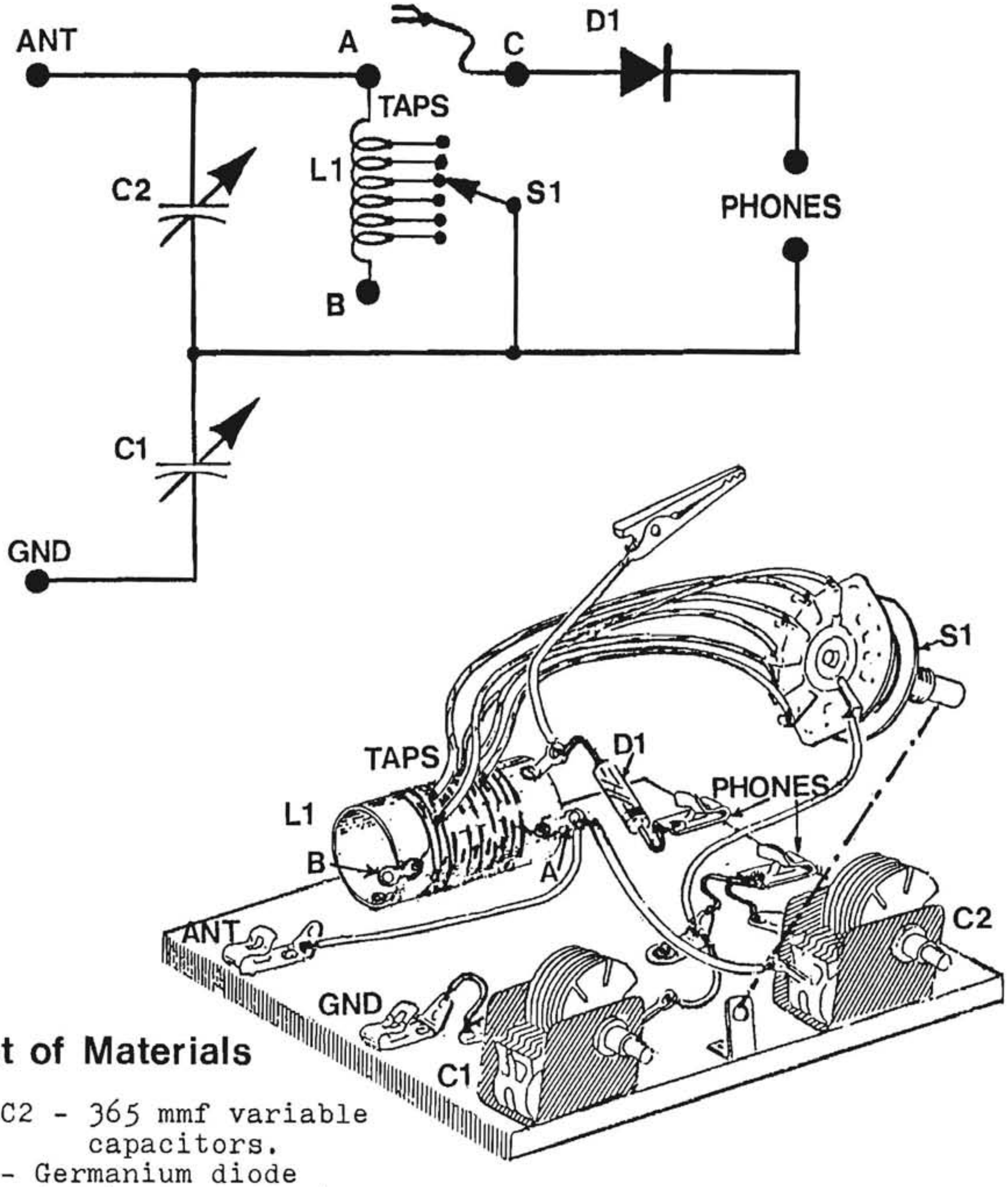
ANT
A
C
D1
TAPS
L1
S1
PHONES
C2
B
C1
GND
S1
TAPS
D1
PHONES
L1
B
A
ANT
C2
GND
C1

List of Materials

C1,C2 - 365 mmf variable
        capacitors.
D1 - Germanium diode
     (1N34A or equiv.).
L1 - Tuning coil (see text).
S1 - 1-pole, 10-position
     rotary switch (see text).
4 - Fahnestock clips.
1 - Small alligator clip.
BASEBOARD - 5 x 8 x 3/8-inch
     wood section (see text).
PHONES - 2000-ohm headphones.
MISC. - Coil form (see text),
No. 24 enameled copper wire,
solder lugs, hook-up wire,
wood screws and nuts, knobs,
bracket for S1.

DUAL TUNED
CRYSTAL SET

The coil L1 is wound with No. 24 enameled copper wire, and is tapped every five turns for a total winding of 75 turns.

Drill a small hole in the coil form approximately 1 1/4-inch from the "A" lug. Remove the enamel from the end of the No. 24 wire and then feed the wire end through the hole. Solder the wire to lug "A".

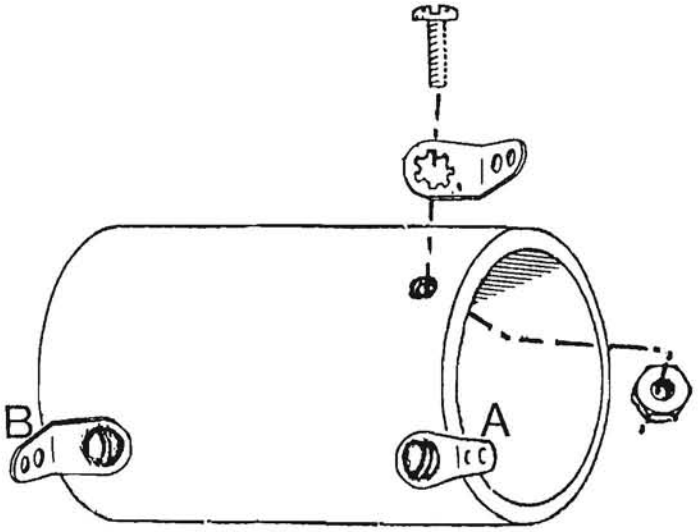

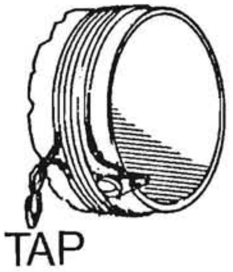

Wind five closely spaced tight turns around the coil form. Then bend back and twist together 1/2-inch of the wire to make a tap. Position the tap outward and then continue winding the coil, keeping the turns as tight and close together as possible. Make a tap every five turns and then drill a small hole in the coil form after the 75th turn. Feed the wire through the hole, remove enamel, and solder the end to lug "B".

BASEBOARD

The Dual-Tuned Crystal Set is built on a wood baseboard as shown in the illustration. Cut the 5-inch by 8-inch baseboard from a section of 1/4-inch thick plywood. Sand the edges and surface of the baseboard to remove any splinters. Drill holes in the baseboard and install the parts with small machine screws and nuts in the locations shown in the illustration. Drill mounting holes at each end of the tuning coil L1 and position it on the rear of the baseboard with the coil taps toward the front. The illustration shows a commercial rotary switch, but a wooden rotary switch can be used for S1 instead. (See page 53 for construction details.)

Install the four fahnestock clips (used as antenna, ground and headphone terminals) with solder lugs. Mount an additional solder lug on the baseboard midway between S1 and L1.

WIRING

Connect the crystal set parts with hookup wire as shown in the pictorial illustration and the schematic diagram. Use the center solder lug as a junction point for leads from C1, C2 and S1 as shown in the illustration. Connect the leads from S1 to the taps on L1 in numerical sequence; every 5 turns for 7 taps. Then connect the last three leads every 10 taps. Insert the lead ends into the taps and hold them with a drop of solder. This will make it easier to move leads to other L1 taps. Use a 5-inch length of stranded hookup wire for the alligator clip lead connected to D1 at solder lug C.

Experiment 1

Connect 2000-ohm headphones, antenna and ground leads, to the crystal set terminals. Then connect the D1 clip lead to the L1 "A" solder lug.

Set S1 to each tap position in turn and tune C1 and C2 for received radio stations. If all of the local stations can not be heard, experiment with different connections to the L1 taps. More turns may have to be added to L1 (with additional five-turn taps) to receive stations at the lower frequency end of the broadcast band.

Experiment 2

Remove the D1 clip lead from "A" and then connect it to the "B" solder lug on L1. Repeat the tuning method of Experiment 1 and note if the crystal set is more selective. This is possible, because L1 is now connected as an rf transformer. You can also connect the D1 clip lead to each of the S1 taps and note if there has been any improvement in reception.

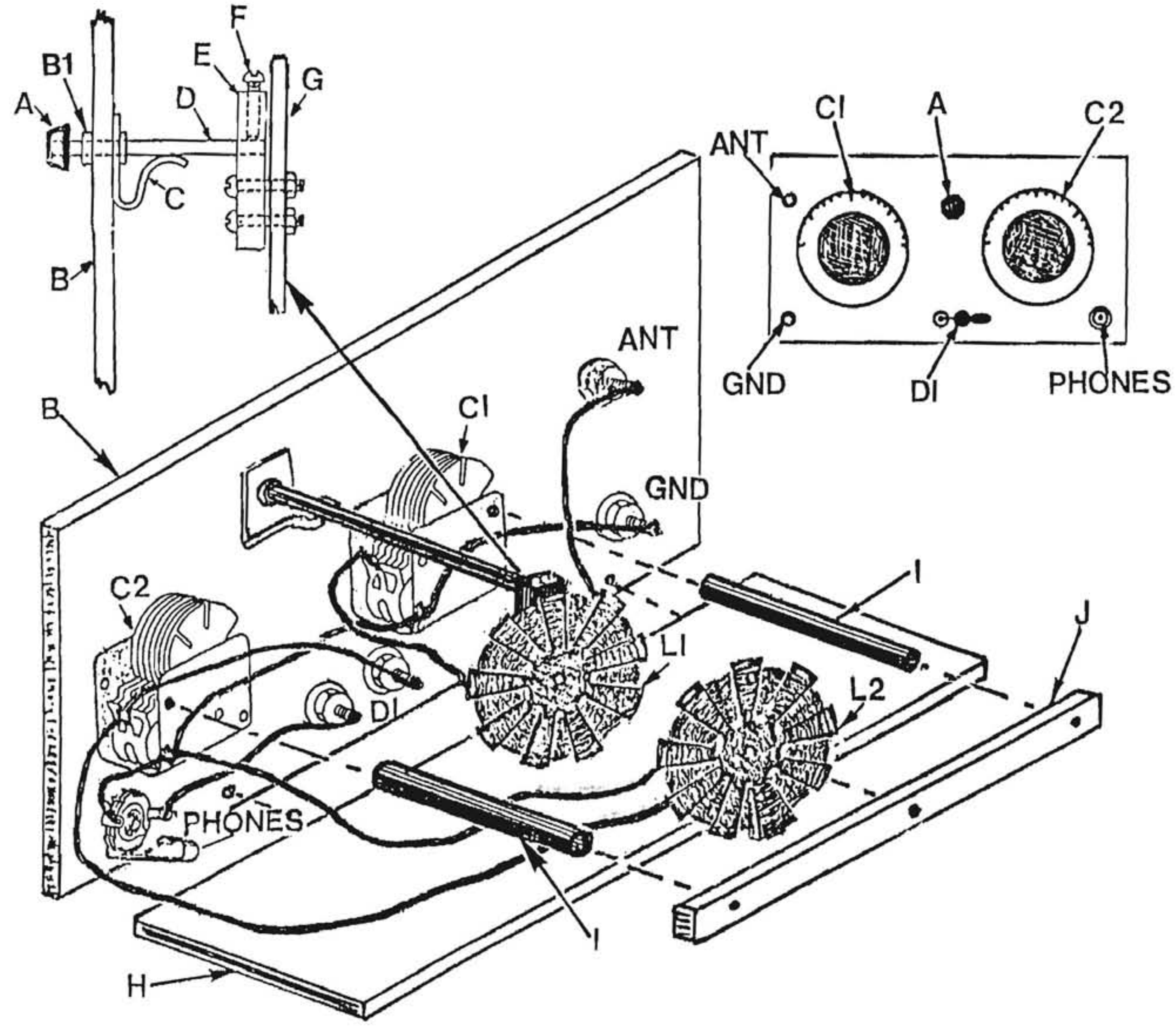
F
B1
E
A
D
G
C
B
ANT
CI
A
C2
GND
DI
PHONES
B
ANT
CI
GND
C2
I
J
LI
L2
DI
PHONES
I
H

SELECTIVE CRYSTAL SET

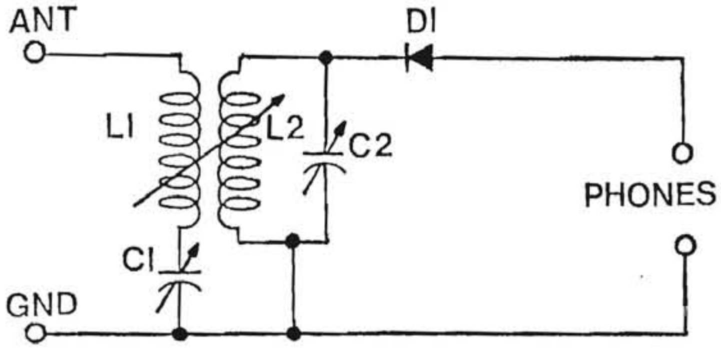
ANT
DI
LI
L2
C2
PHONES
CI
GND

SELECTIVE CRYSTAL SET

This receiver uses two spiderweb coils in a two-circuit tuner for good selectivity and sensitivity. As you can see in the schematic diagram, the antenna-ground rf input circuit is tuned by L1-C1 for maximum signal gain. L1 acts as the primary winding of an rf transformer, with L2-C2 as the tuned secondary. Coupling between L1 and L2 is variable for best selectivity. L2 is connected to a catwhisker crystal detector and the audio signals are fed through the phone jack to 2000-ohm headphones.

As shown in the illustration, the crystal set parts are mounted on the rear of a black plastic panel. This type of layout is similar to the type of construction of the old-time crystal sets.

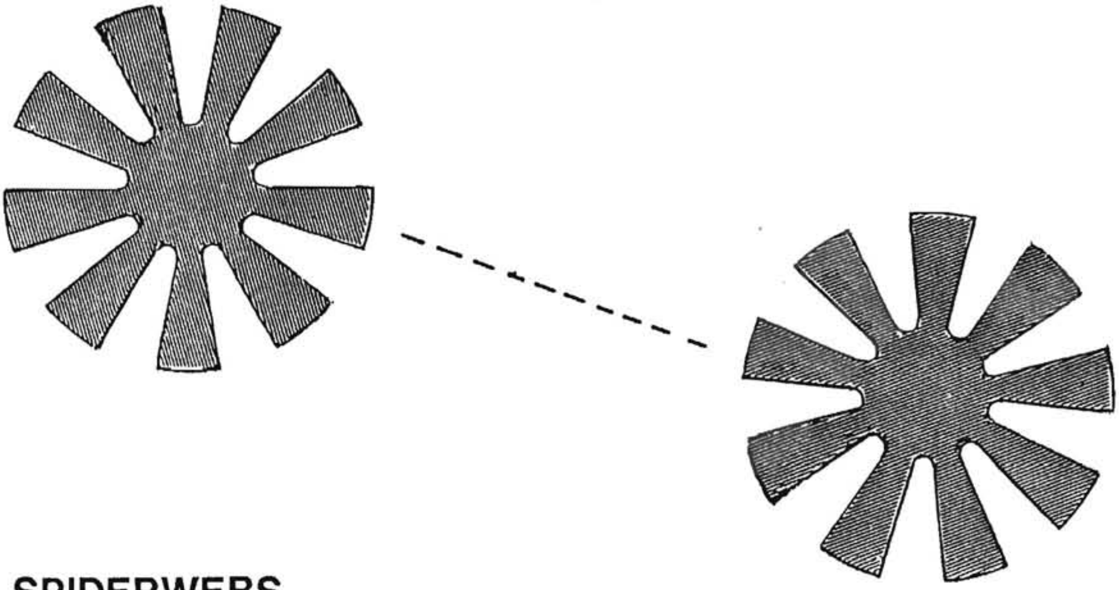

SPIDERWEBS

Both L1 and L2 spiderweb coils are made exactly the same way. L1 and L2 are wound on 17-vane coil forms 3 5/8-inches in diameter. The coil forms in our receiver are made from 1/16-inch sheet plastic of the same type used for printed circuit boards (but without the copper foil). Any kind of rigid sheet plastic should be suitable for the coil forms.

Make a tracing of the spiderweb coil template. Then use carbon paper to transfer the tracing on a section of 1/16-inch thick sheet plastic. Cut out the vanes with a jeweler's saw. Drill two holes as shown, and install the two solder lugs.

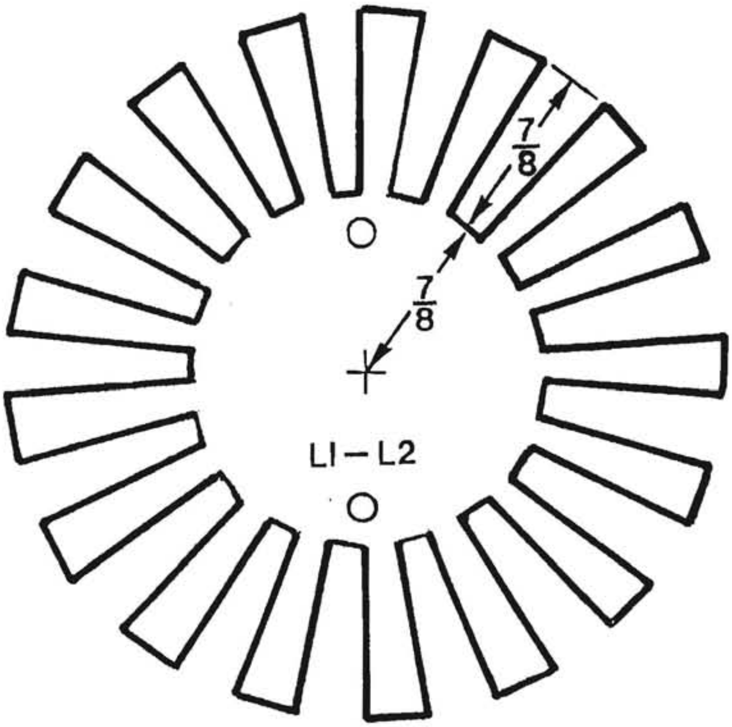

COIL FORM TEMPLATE FOR L1–L2

Wind as much No.22 enameled wire as possible on each coil form. Solder the wire to a solder lug before winding a coil. Then wind the wire over a vane and under the next vane. Wind the wire closely and evenly on the form. It is not necessary to count the turns. Solder the wire end to the remaining solder lug

CONSTRUCTION

The panel used in our receiver is a 7 3/4-inch by 4 1/2-inch section of 1/4-inch thick black acrylic plastic. The size is not critical, but it should be large enough to mount all of the parts as shown in the illustration.

Layout and drill the mounting holes of the components on the panel. Make sure that C1 and C2 are positioned so that the edges of the tuning knob dials are more than 1/2-inch apart. Install the catwhisker crystal detector assembly, antenna and ground terminals, and the phone jack in the locations shown.

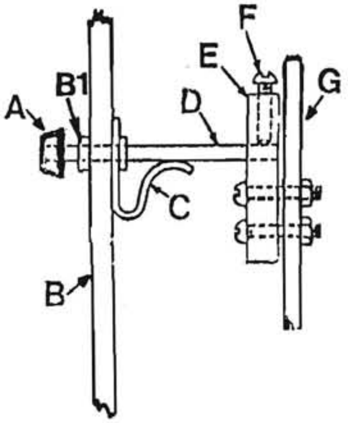

As you can see in the illustration, L1 is installed on the end of a variable coupling metal rod. This 1/4-inch diameter rod is 3-inches long with a flat side. Drill and tap a small hole close to the end of the rod. The rod used in our model came from a discarded volume control. Cut a 2-inch long by 1/2-inch wide by 3/8-inch thick wood section and drill holes to fit the rod, two mounting screws, and a hole for the rod set screw. Locate and mount the wood section in the center of a spiderweb coil L1 so that the rod will be 3/4-inch down from the edge of the top vane.

Hold L1 behind the panel and using the free end of the rod, locate the panel hole. Position L1 midway between C1 and C2, and make sure that the top edge of the coil is below the top of the panel. Drill a hole to fit the rod bushing in the panel hole location. The panel bushing in our model was removed from a discarded volume control.

Drill a hole to fit the bushing in one end of a 2-inch by 1/2-inch section of sheet aluminum. Bend the other end of the section to fit against the flat side of the rod. Install the metal section and the bushing in the panel hole.

Insert the rod into the bushing and make sure that the metal section end fits close to the flat side to prevent rotation of the rod. Install a knob on the rod end and make sure that the rod can be pushed in and out easily. If necessary, put a small amount of grease on the flat side of the rod.

Install the remaining spiderweb coil L2 on a wood section that is mounted on the rear of the panel with two spacers. The spacers should be long enough for L2 to be located 1/4-inch behind L1 with the rod pushed all the way in. The 5 1/2-inch by 1-inch by 1/4-inch wood section in our model is supported with screws installed through 1 1/2-inch metal spacers into tapped holes in the rear of C1 and C2. The screws and bushings can be made longer and mount into panel holes in your receiver.

Install a wood section on the bottom rear of the panel to keep the receiver upright. A 5-inch by 3-inch section of 1/4-inch thick wood is used in our receiver.

SET WIRING

Connect the receiver components with hookup wire as shown in the illustration and schematic diagram. Use flexible stranded wire for connections to L1. Make sure that the L1 leads do not touch any component as the variable coupling rod is moved in and out.

Install knobs on C1, C2, and the variable coupling rod. Then place the crystal in the detector assembly.

OPERATION

Connect antenna and ground leads to the receiver terminals. Plug in a set of 2000-ohm headphones into the panel phone jack. For easier initial operation of the receiver, temporarily connect a germanium diode (1N34A or equivalent) across the D1 connecting lugs. Make sure that the catwhisker is not touching the crystal at this time.

Push the variable coupling rod half way in and tune C1 and C2 for a station. Adjust the variable coupling rod and retune C1-C2 for best reception. Check operation over the entire broadcast band. It may be necessary to add or subtract turns of wire on L1-L2 for complete coverage. After checking operation with the germanium diode, remove it and try using the catwhisker crystal detector.

List of Materials

C1,C2 - 365 mmf variable capacitor.
D1 - Catwhisker crystal detector assembly.
L1,L2 - Spiderweb coils (see text).
2 - Terminals.
1 - Phone jack.
PHONES - 2000-ohm headphones.
MISC.- Plastic sheet for spiderweb coils, solder lugs, machine screws and nuts, wood screws, hookup wire.
CONSTRUCTION MATERIALS (See Text).
A - knob
B - Panel
B1- Bushing
C - Aluminum section
D - Rod
E - Wood section
F - Setscrew
G - L1 coil form
H - Wood section
I - Spacers (2)
J - Wood section

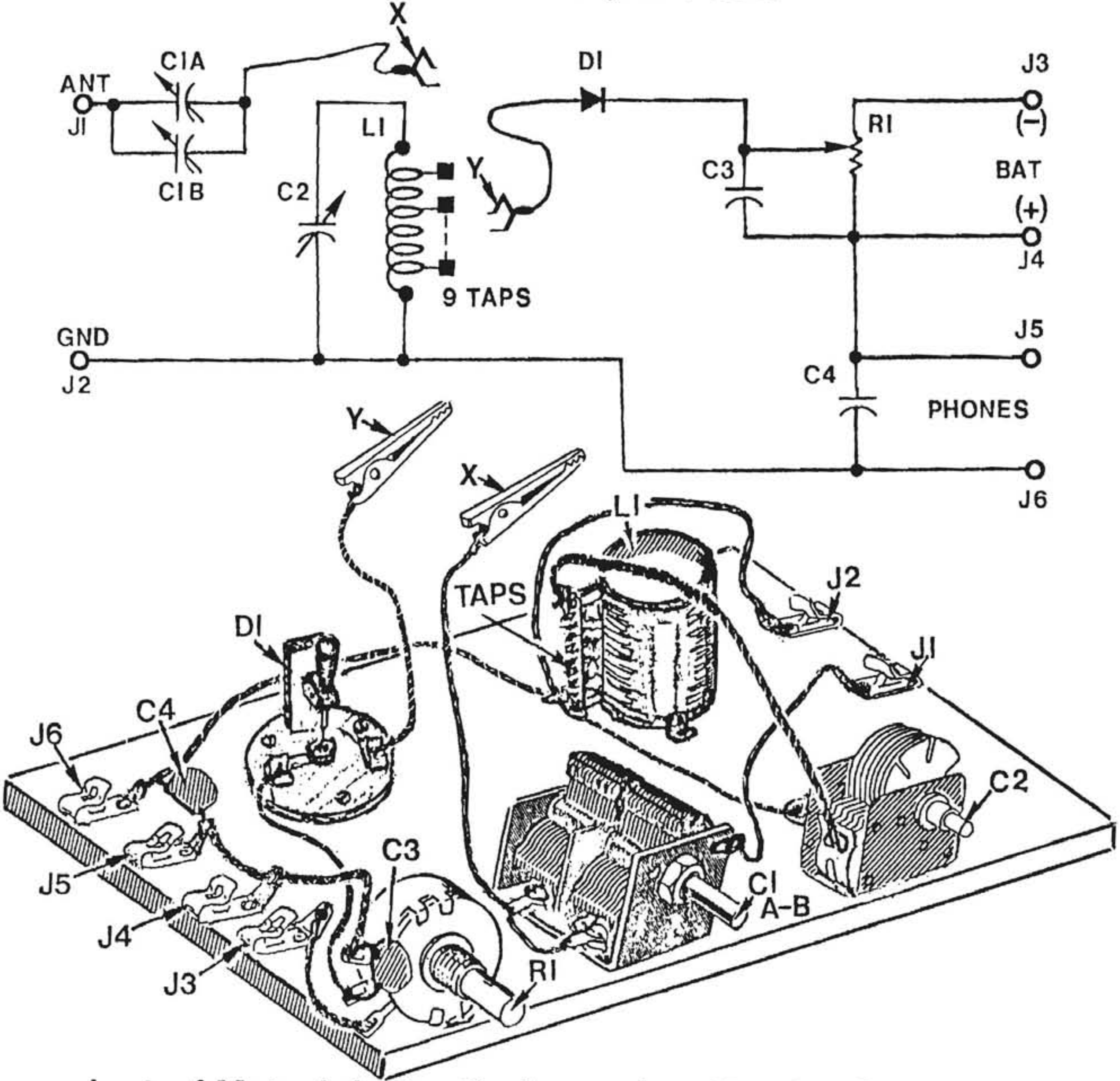
X
C1A
D1
J3
ANT
J1
L1
R1
(−)
C1B
C2
Y
C3
BAT
(+)
J4
9 TAPS
GND
J2
J5
C4
PHONES
J6
Y
X
L1
TAPS
J2
J1
D1
C4
J6
C2
C3
J5
C1
A-B
J4
R1
J3

List of Materials for Carborundum Crystal Set

C1A-C1B - Dual 365 mmf variable capacitor.
C2 - 365 mmf variable capacitor.
C3 - 0.1 mf capacitor.
C4 - 1000 mmf capacitor.
D1 - Carborundum crystal and catwhisker detector assembly (see text).
J1 to J6 - Fahnestock clips.
L1 - Tuning coil (see text).
R1 - 5,000-ohm potentiometer (linear taper).
BAT - 6-volt battery
PHONES - 2000-ohm headphones.
MISC. Coil form (see text), No.24 enameled copper wire, 1/2-inch long spacers, perfboard strip (see text), push-in clips, solder lugs, knobs, hook-up wire, 1N34A diode (or equiv.), wood for base and D1 detector assembly (see text), wood screws, machine screws and nuts, alligator clips X-Y, brackets for L1 and R1.

CARBORUNDUM CRYSTAL SET

Materials other than galena, silicon or germanium have the property of detection. In the beginning of radio, carborundum (the name given to a compound of silicon carbide) was also used as a detector. The required heavy catwhisker pressure made it suitable for the early radio stations on ships, as it did not fall out of adjustment. It would also handle the strong signal energy from nearby spark transmitters without burning out. But what makes this type of crystal detector different, is the necessity for a bias battery.

A crystal detector has a high current flow with voltage applied so that it conducts in the forward direction (catwhisker to crystal), and a very low current flow in the reverse direction. The amount of current flow in the forward direction depends upon the characteristics of the crystal material and the applied forward voltage.

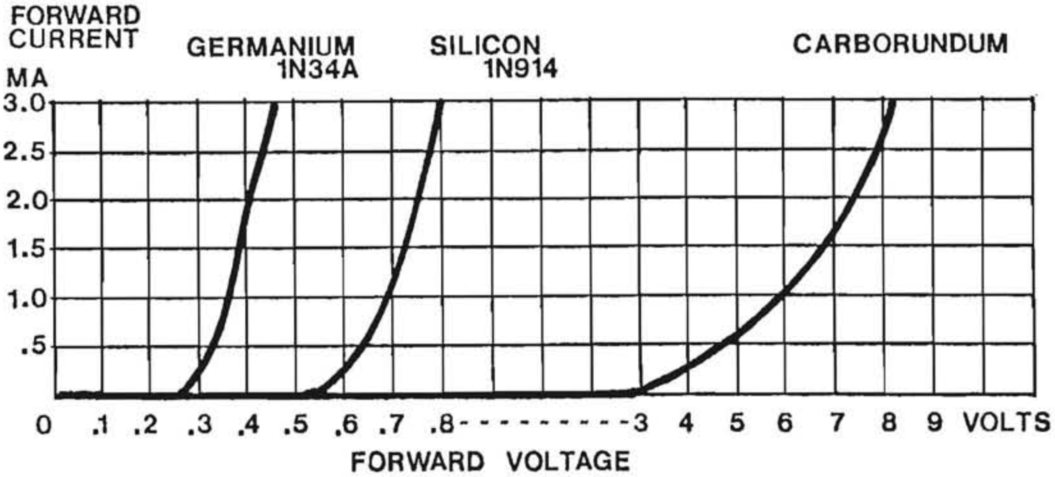

TYPICAL CRYSTAL DIODE CURRENT

The graph compares carborundum with germanium and silicon crystal detectors. Germanium minimum conduction voltage is about 0.3 V. Silicon is 0.6 V, and carborundum is 3 V. The high carborundum minimum conduction voltage is the reason a bias battery is used to move the threshold down so that weak radio signal voltages can be detected.

TUNING COIL

As shown in the illustration, the tuning coil L1 is wound on a cardboard (or plastic) mailing tube section 2-inches in diameter and 2 3/4-inches long.

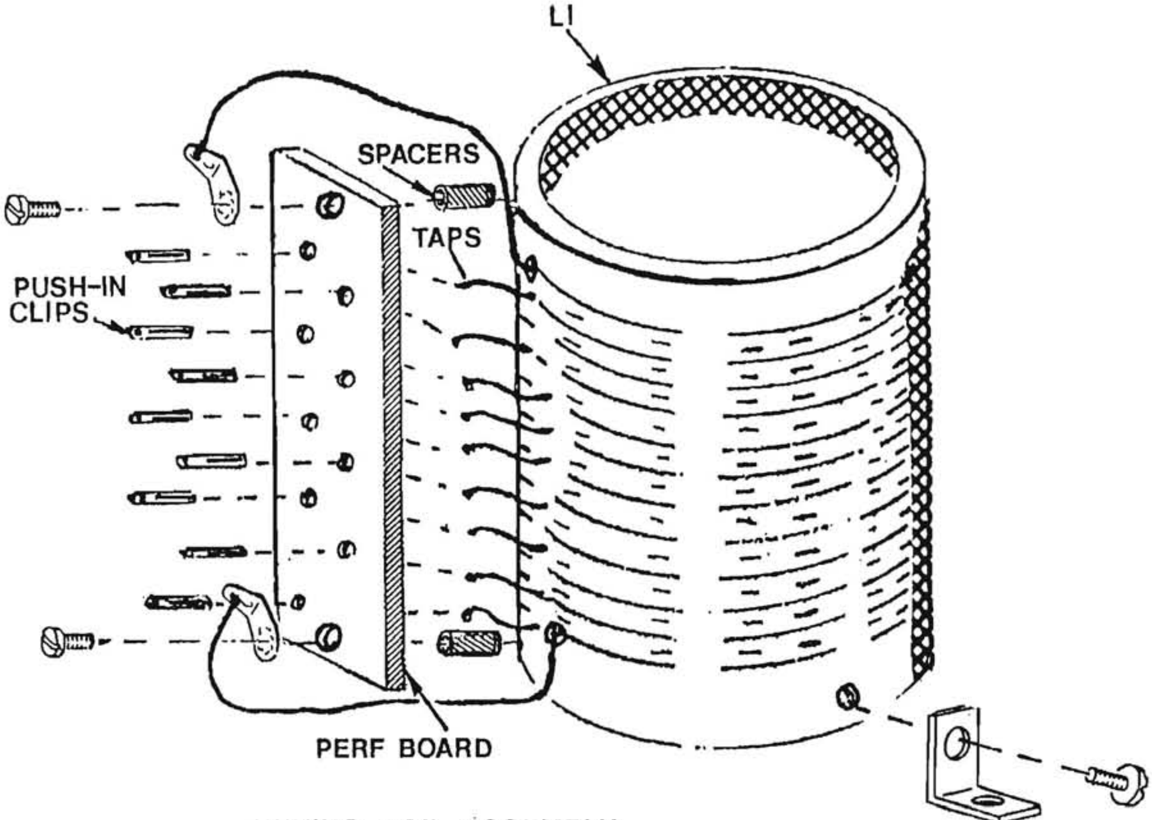

TUNING COIL ASSEMBLY

Locate and drill a hole 1/4-inch from each coil form end for machine screws to fit the two spacers. Also drill holes for two mounting brackets at one end. Then drill a small hole near each spacer hole to feed the wire on the form before you start winding.

Wind the coil with No. 24 enameled copper wire. Tap the coil every 10 turns for a total of 9 taps. An easy way to make the taps is to twist the wire together for a half inch and position the free end out. Make all the coil taps in the area between the two spacer holes. Thread the free end of the wire through the small hole at the end of the form near the spacer hole. Make sure that all of the turns are tightly wound and close together. Use a section of sandpaper to remove the enamel from the tap wire ends. Then tin the ends.

Mount 9 push-in clips in a 1/2-inch by 2 1/4-inch perf board section. Locate and drill holes on the board and install it on the coil form with two solder lugs, 1/2-inch spacers, machine screws, and nuts. Solder the coil taps to the ends of the push-in clips. Connect the coil leads to the solder lugs and mount the two 1/2-inch brackets on the coil form.

DETECTOR

The crystal detector assembly is vertical, (as shown in the illustration) instead of horizontal as in most catwhisker detectors. Begin construction by cutting the 2 1/2-inch diameter base and 2-inch by 4-inch vertical section from 1/4-inch thick wood.

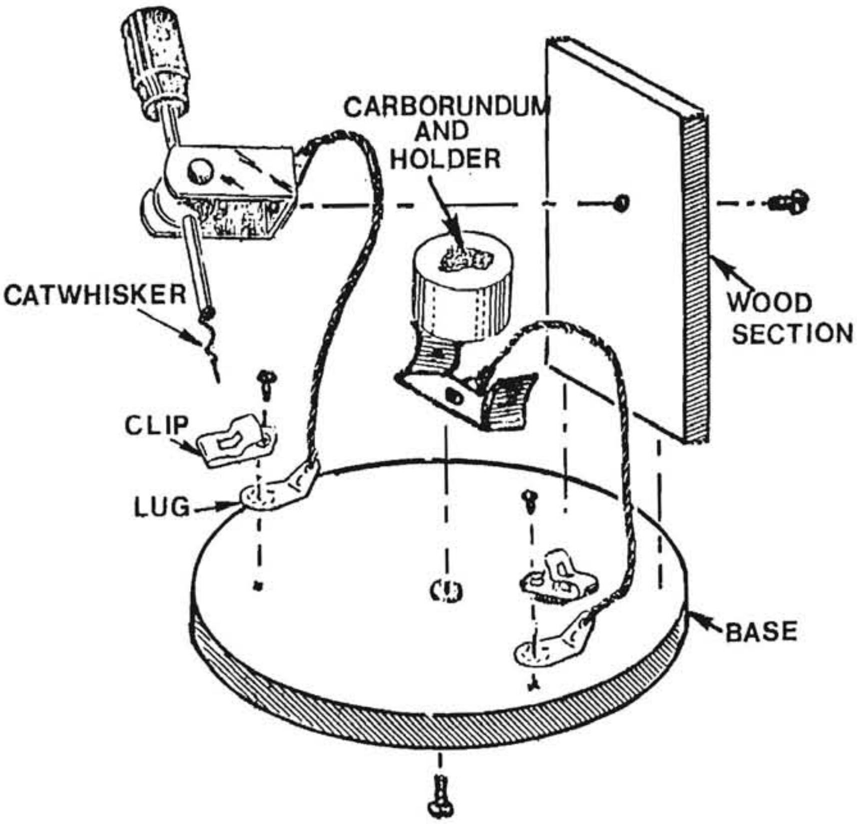

DETECTOR ASSEMBLY

Locate and install the two fahnestock clips on the base with solder lugs and small wood screws. Mount the vertical section on the edge of the base with wood screws (or cement). Locate and install the catwhisker holder with a solder lug, machine screw and nut near the top of the vertical section.

Temporarily position the crystal holder on the base so that the catwhisker will touch the center of the crystal. Locate and mount the crystal holder in this position with a solder lug and machine screw and nut. Connect the detector assembly with hookup wire as shown in the illustration. Solder a short section of coiled steel spring wire to the end of the positioning rod to serve as a catwhisker. Cut the end of the wire diagonally to give it a sharp point.

BASEBOARD

The receiver parts are installed on a 8-inch by 7-inch by 3/4-inch wood base. Wherever poossible, small wood screws are used to hold the components on the base.

Begin construction by laying out and installing the parts in the positions shown in the illustration. It will be necessary to make a metal mounting bracket to fit R1. C1A-B and C2 are mounted with machine screws in the tuning capacitor frames through countersunk holes in the base. Install solder lugs in the rear frame of both tuning capacitors. Also mount solder lugs with all of the terminals. Install knobs on the tuning capacitors and R1.

WIRING

Use hookup wire to connect the crystal set parts as shown in the schematic and illustration. Connect 5-inch leads of flexible stranded insulated wire to each alligator clip. Solder one lead to the stator lugs of C1A-B and connect the other lead to D1. Connect C3 and C4 as shown and keep their leads as short and direct as possible.

Experiment 1

Connect your antenna to J1, ground to J2, and 2000-ohm headphones to J5 and J6. Temporarily connect a germanium diode (1N34A or equiv.) in place of the carborundum crystal detector D1. The diode will make it easier to check out the receiver.

Do not attach the battery
to J3-J4 at this time, or
the diode may be damaged.

Connect **X** C1A-B clip lead to the top tap of L1, and **Y** D1 clip lead to the second tap from the top. Tune in a radio station with C2 and then adjust C1A-B for best volume. Change the position of the clip leads and retune C1A-B and C2 for best reception. Check operation of the receiver over the entire broadcast band.

Experiment 2

Tune the receiver to a radio station and then disconnect the germanium diode. Do not disturb the receiver tuning. Install a carborundum crystal into the D1 assembly. Set R1 to mid range, and connect a 6-volt battery to the receiver; negative lead to J3, positive lead to J4.

Adjust the D1 catwhisker for a sensitive spot on the carborundum crystal. Then set R1 for best volume of received signal. Retune C1A-B and C2 and change the clip leads on the L1 taps as necessary for reception of radio stations.

Note

Carborundum that will work as a detector may be hard to find. Grinding wheels are not suitable as a source, because the carborundum is usually in fine particles mixed with other material. Best bet is to contact collectors of antique radio parts. But you can use this receiver to experiment with minerals for semiconductor operation. See the Workshop Section for information on mounting mineral sections for detector experiments.

4 COMPONENT PROJECTS

BOOK TUNING CAPACITOR

During the roaring 20's,the Crosly Radio Corporation used a type of tuning capacitor that had two plates hinged together on one edge. The dielectric was a sheet of mica between the two plates. This type of tuning capacitor was called a Book Condenser.

The tuning capacitor shown in the illustration is similar to the Crosly Book Condenser. It uses aluminum foil plates cemented to fixed and movable wood sections, and a plastic sheet dielectric (cut from thin transparent plastic notebook protector material).

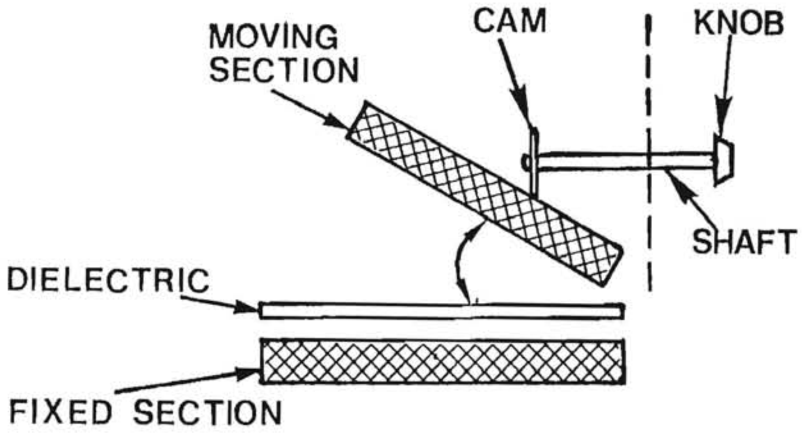

Adjusting the tuning knob causes a metal cam to rotate and control the movement of a spring loaded wood section. This varies the capacity between the two foil plates. As the plates are brought closer together, the capacity increases. As they are separated, the capacity decreases.

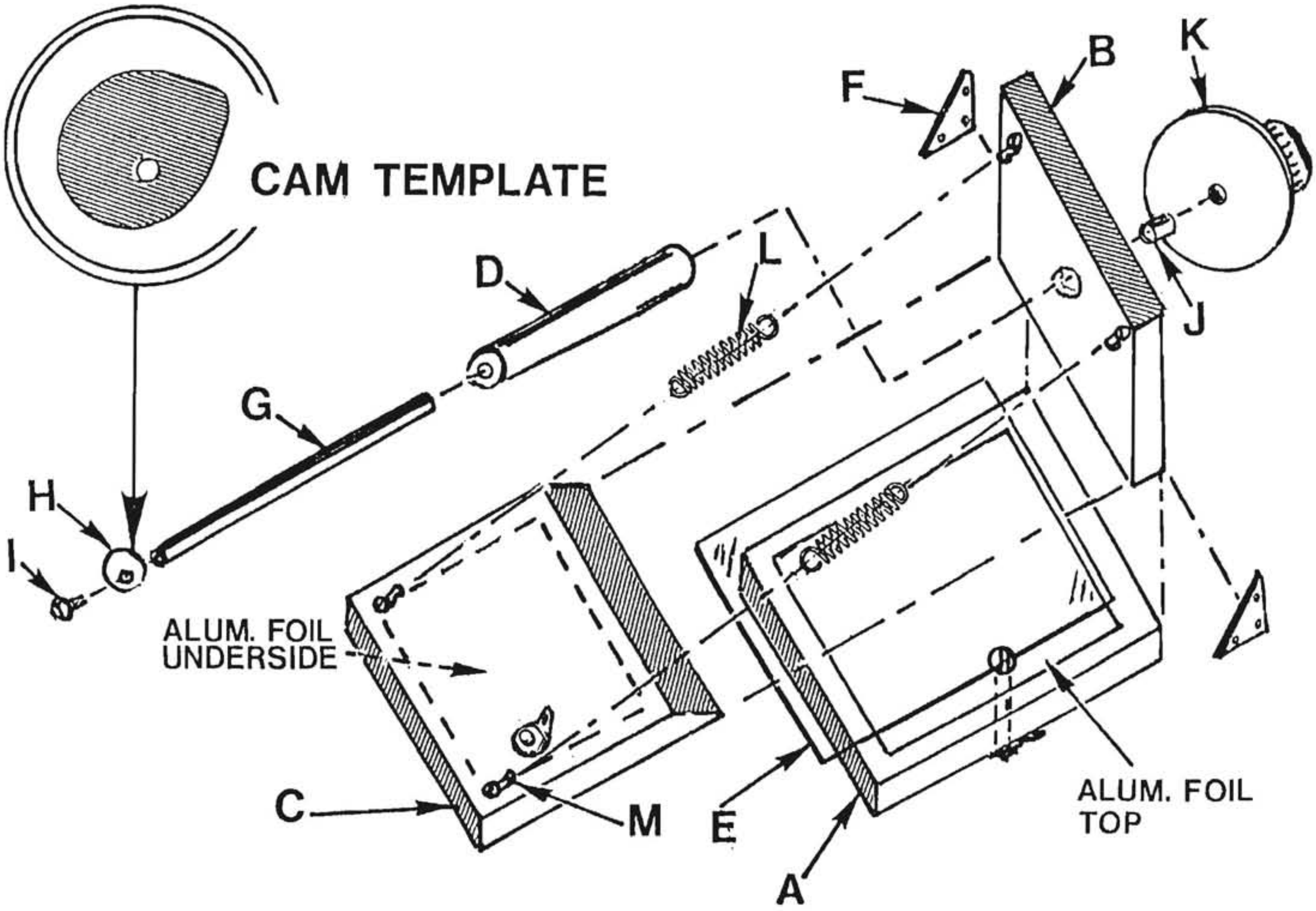

BOOK TUNING CAPACITOR

List of Materials

A - 3 x 3 1/2 X 3/8-inch wood section
B - 3 x 1 3/4 x 3/8-inch wood section
C - 3 x 3 1/8 x 3/8-inch wood section (one end beveled 45 degrees - see text).
D - 3/4 - inch diameter wood dowel, 2 1/4 inches long.
E - 3 1/2 x 3 -inch sheet plastic (See text).
F - 2 triangular 1-inch metal sections
G - 1/4-inch metal rod, 3 -inches long.
H - Metal cam (see text).
I - Machine screw (see text)
J - Bushing 3/8-inch long (to fit rod G).
K - Tuning knob (to fit rod G).
L - 2 springs 1/4-inch diameter by 1 1/4-inch long.
M - 4 solder lugs.
MISC. Small wood screws, nails, aluminum foil, machine screws, nuts, solder lugs and washers.

CONSTRUCTION

Start construction by cutting the wood sections from 3/8-thick wood to the sizes shown in the drawing. Section C has a 45 degree bevel at one end. This acts as a pivot point for the movable section. Drill and countersink a hole to fit a flathead machine screw in the places shown on sections A and C. Do not install the screws at this time.

Cement aluminum foil on the top of section A and on the bottom of section C. Then cut and remove a 1/2-inch strip of foil from the front edge of section A. Cut a 1/8-inch strip from the remaining edges. Also cut away a 1/8-inch strip from all the edges of the foil on section C.

Carefully slit the foil on top of the machine screw hole on section A and press the foil gently inside the hole. Install the flathead machine screw into the hole, making sure that it is in contact with the foil. Fasten the screw with a washer and nut. Make sure that the screw does not turn and rip the foil. Install a solder lug and nut on the free end of the screw. Repeat this operation with section C.

Cut the sheet plastic E to size and place it on top of section A. Then place section B on top of the plastic sheet and install it on the end of section A with wood screws. Install the two metal sections F with small nails.

Saw the metal rod G to size and tap one end to fit a machine screw. Cut the dowel D to size and center drill it to fit the rod G. Cut a flat on one side of the dowel.

Make a tracing of the cam H template and cement it on a section of heavy gauge sheet aluminumm. Cut the cam to size and drill a hole to fit the rod machine screw in the location shown. Install the cam with a lock washer on rod G. Insert the free end of the rod into dowel D. Place section C on top of the plastic sheet E on section A. Then temporarily position dowel D with rod G in the center of section B. Rotate the cam so that its long end touches the top of section C. Mark the placement of the dowel on section B. Drill a hole in section B to fit dowel D. Cement the dowel in place.

Locate and install solder lugs with wood screws on sections A and C as shown in the drawing. Bend over the ends of the solder lugs and install two springs L in the lugs.

Mount tuning knob K and bushing J on the free end of rod G. Adjust the knob and check that section C moves in and out with the cam rotation. changing the capacity between the foil plates.

You may have to experiment with the cam diameter for best operation of section C. File the cam edges in small increments only and recheck the operation often. Use very light springs to keep section C from pressing against the cam and moving it.

Reverse the cam on rod G to alter the direction of capacity (clockwise or counterclockwise). The capacitor can be installed vertically on a receiver project with a small metal bracket mounted on the bottom side of section A.

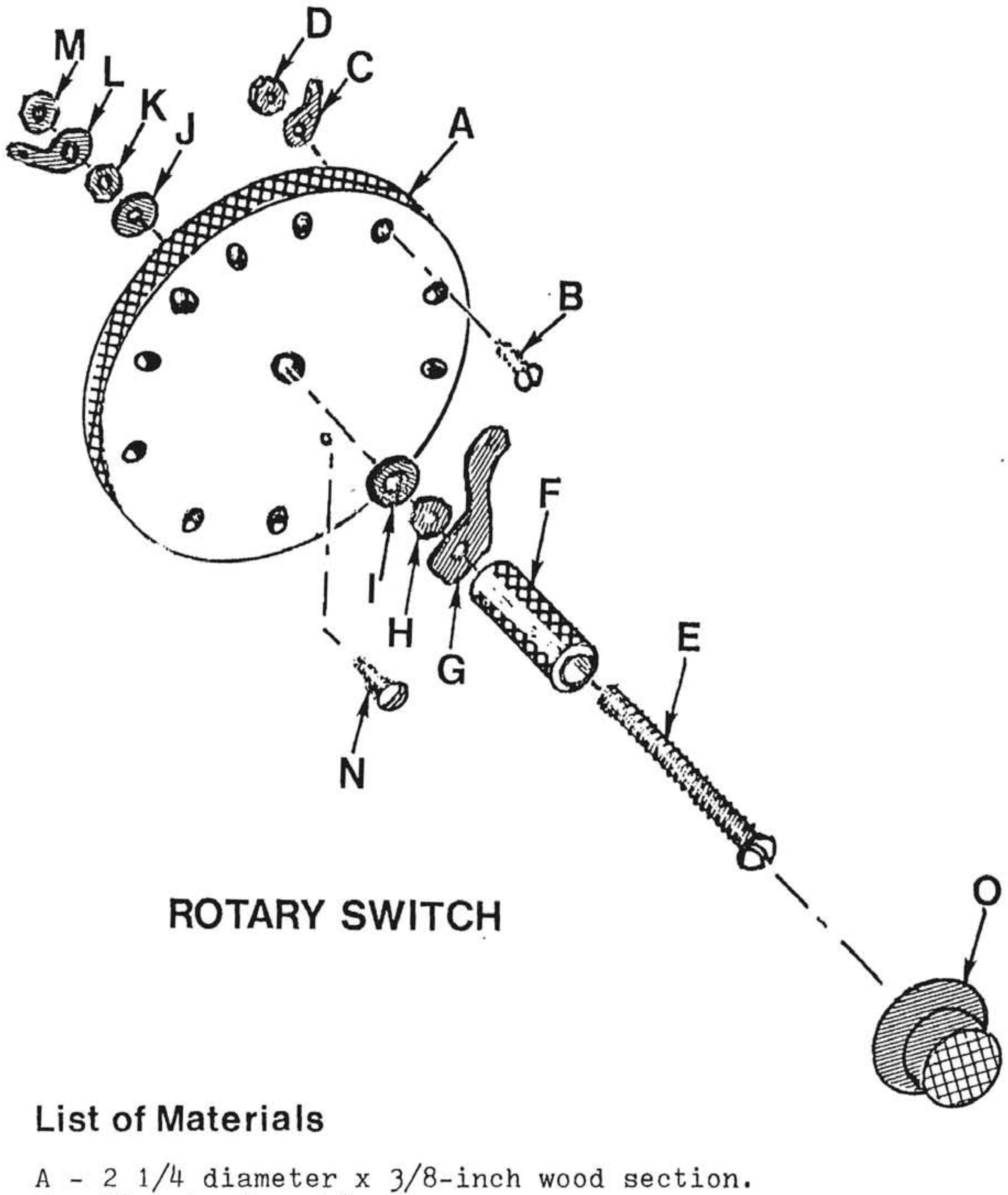

ROTARY SWITCH

List of Materials

A - 2 1/4 diameter x 3/8-inch wood section.
B - Flat head machine screws.
C,L - Solder lugs.
D - Nuts to fit B.
E - Machine screw 1 1/4-inches long.
F - 1/4 diameter x 1/2-inch long bushing (to fit E).
G - Switch arm 1 x 1/4-inch (sheet brass, see text).
H,K,M - Nuts to fit E.
I,J - Washers to fit B.
N - Small woodscrew.

ROTARY SWITCH

In the early days, radio experimenters could not buy parts that are readily available today. They had to build their radio components from wood sections and common hardware.The rotary switch ilustrated is similar to a type constructed for use with tapped coils of crystal sets.

The rotary switch shown uses flat head screws for contact points that are mounted on a wood section. The switch arm is cut from sheet brass and is installed on a machine screw.

CONSTRUCTION

Refer to the drawing of the rotary switch. Begin construction by cutting the 2 1/4-inch diameter section A from 3/8-inch thick wood. Locate, drill,and countersink the holes for the flathead machine screws B. The rotary switch shown has 10 machine screws, but you can space the hole locations for the number of switch contact points desired. Allow a space between the starting and ending contact points for the woodscrew N. This woodscrew is used as a stop for the rotary arm G. Locate and drill a center hole for the machine screw E. Install the screws B in section A with solder lugs C and nuts D. Install the woodscrew N.

Measure and cut the rotary arm G from a section of sheet brass. The arm should be long enough to reach the contacts. Make sure that it is not too wide, or it may short out to a nearby contact. Make a dent or "dimple" on the contact end, and drill a hole for screw E on the other end. Bend the rotary arm as shown in the drawing. Install screw E into bushing F, and arm G, and tighten the assembly with nut H. Place washer I on the screw and install screw E into the center hole of A. loosely fasten the screw with washer J and nut K. Then rotate the bushing F and arm G. Adjust the tightness of nut K as required to allow rotation of G with only a slight "drag". Install solder lug L on screw E with nut M, and bend the lug towards the screw end. Install knob O on bushing F, and recheck operation of the arm G. It should touch all of the contacts B and stop at screw N.

The rotary switch is used as shown in the Dual-Tuned Crystal Set Construction Project. Make sure that extra lead length is allowed for connection to the solder lug L. Bend the lead into a "pig-tail," so that it will not break as the solder lug L is rotated.

5

WORKSHOP

The construction projects in this book can be built with common hand tools found in an average home workshop.

Useful Hand Tools

Safety glasses or face shield
Steel rule
Square
Hack saw
Wood saw
Screwdrivers
Pliers
Wirecutters
Tin snips
Center punch
Files
Hammer
Soldering Iron

Wear safety glasses or face shield while working. Observe common sense safety precautions.

SOLDERING

Make sure that the part to be soldered is clean. Use only rosin core solder. Prepare the soldering iron tip as follows:

File each surface of the tip to remove oxidation. Plug in the iron and let it heat until you can melt solder on the tip (before the copper tip turns blue or bronze color). Melt enough solder until all sides of the tip are coated. Then wipe off excess solder with a damp sponge or cloth.

Make sure that the bare wire lead to be soldered is free of oxide. It is best to tin it first by melting a bit of rosin core solder on the lead end. Crimp the lead in the part hole to keep it from moving during soldering. Hold the iron tip under the part until it is hot enough to melt solder. Apply only enough rosin core solder to cover the surface of the part. Remove the iron. Do not move the part or wire while it is cooling.

PROJECT CONSTRUCTION TIPS

Tape a sheet of graph paper over a project wood section. Mark locations of the parts on the paper. Center punch mounting hole centers through the paper. Use a compass to draw circles the same size as the holes. Then drill holes through the paper.

Lettering on the projects can be done with a labelmaker that prints on strips of adhesive coated plastic. You can also use decals or dry transfer lettering sheets.

Small metal brackets can be made from heavy gauge sheet aluminum. Layout and drill the mounting holes and then cut it to size with a tin snips. Form the bracket by using a heavy mallet and a vise.

Tuning Capacitors, Knobs, and Dials

Old radios are a good source of components. The radios generally use two tuning capacitor sections ganged together. The receivers previously described will operate with one of the capacitor sections (365 mmf). You can cut out cardboard discs and cement them to radio knobs and use them as dials with the tuning capacitors. Mark the dials with dry transfer lettering, or black drawing ink.

Panels and Cabinets

The layout and construction of crystal sets is not critical. They can be fitted into many different types of wood cabinets. You can use almost any type of non-metallic box, but metal boxes are not recommended as sheet metal close to the tuning coil may lower its "Q". Black sheet plastic used for the front panel with white dry transfer lettering will give your receiver project an antique look.

MOUNTING MINERALS FOR CRYSTAL DETECTORS

Minerals (galena, etc) for experimentation as crystal detectors, will work best when mounted in a metallic base. For temporary mounting (while testing), wrap aluminum foil tightly around the sides and bottom of the mineral. Then place it into a sheet metal ring or cup that will fit into the detector assembly.

After testing, the mineral can be permanently mounted in a metal slug. Drill a shallow hole of a diameter equal to the detector assembly crystal in a block of wood. Fill the hole with melted solder. While the solder is still molten, carefully place your mineral section into the center. Hold the mineral in place with tweezers or long nose pliers until the solder cools. Then pry the metal slug out of the wood. Be sure that the top of the mineral section is free of solder.

INDEX

Advanced receiver projects,32
Amplifier connection,14,15
Antenna lead-in,10
Antenna,10

Book tuning capacitor,48
Boundary,5
Bias,5

Capacitor,8
Capacity,8
Carborundum crystal set,43
Carborundum,43
Catwhisker detector,4,6
Catwhisker,6
Component projects,48
Construction tips,55
Crystal detector assembly,45
Crystal detector,4

Detected audio signals,7
Detection,4,6
Diode electron flow,5
Dual spiderweb coils,37
Dual-tuned crystal set,32

Earphones,11
Enamel insulation,17
Electron current flow,5

Ferrite coil crystal set,23
Ferrite coil,12
Ferrite rod,23

Galena,4
Germanium,4
Ground lead,10
Ground wave,25
Ground,10

Hand tools,54
Headphones,7,11
Headset,11

Junction diode,4,5

Knobs and dials,55

Lettering,55
Loading coil,21
Loopstick,12

Modulated radio waves,6
Mounting minerals,56

N type material,4
Natural crystal,4

Operating tips,25

P type material,4
PN junction,5
Panels and cabinets,5
Parallel tuned circuit,9
Point contact diode,4,6
Primary coil,23

Reactance,8
Receiving radio waves,7
Resonant frequency,9
Rf coil,8
Rotary switch project,53

Schematic symbols,15
Secondary coil,23
Selective crystal set,37
Selectivity,9
Semiconductor diode,4
Silicon,4
Simple crystal set,12
Simple receivers,12
Sky wave,25
Slider tuned coil,17
Slider tuned set,17
Soldering,54
Space charge region,5
Spiderweb coil set,27
Station log,31

Tapped coil,32,44
Tapped spiderweb,27
Transmitted waves,6,7
Transmitting station,6
Tuned circuit,7,9
Tuning crystal set,7

Untuned detector,20

Variable capacitor,8
Variable coupling,39

Wave trap,21
Workshop,54

About the Author

Charles Green has been involved in the design and construction of electronic equipment for industry and the U.S. Government for 30 years. Much of his experience has been as a technical writer preparing operation and maintenance manuals. His continuing interest is in constructing radios and he has had over a hundred articles published in hobbyist magazines over the past two decades.

A sun-powered transistor amplifier makes this a—

Space-Age Crystal Set

By D. S. Halacy Jr.

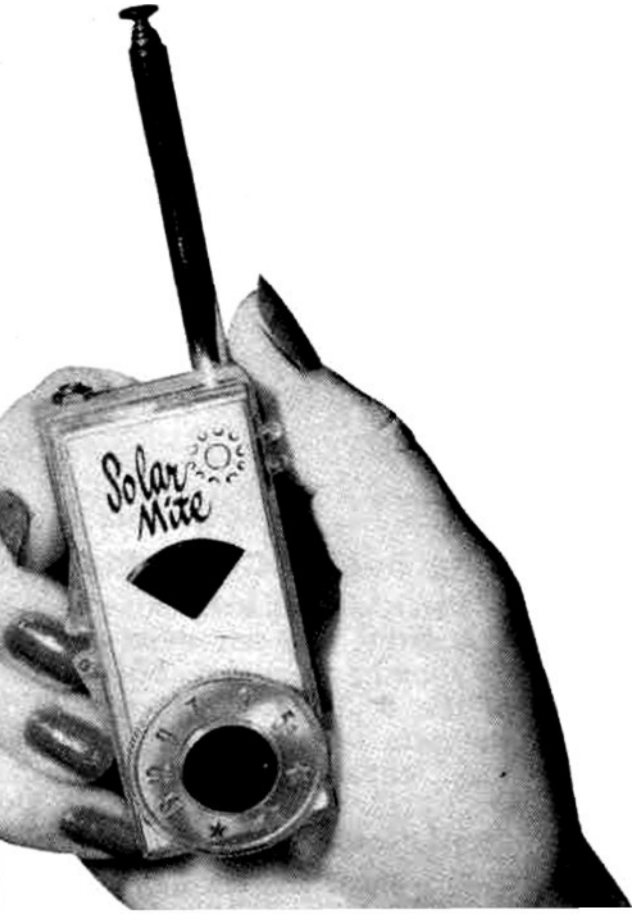

LIGHT-POWERED RADIO plays indoors under artificial light, too. Performance is boosted by connecting antenna to TV aerial or telephone-dial finger stop; a short length of wire, with alligator clips at the ends, helps.

THIS tiny radio puts satellite technology in the palm of your hand. Like radios in artificial moons, it's powered by the sun—or any other light.

Basically, it's a 1960-model crystal set. A rugged sealed-in-glass diode has replaced the shaky catwhisker and galena crystal. The usually faint crystal-set signal is magnified many times by a transistor, in turn powered by a high-efficiency silicon solar cell.

This unit—and other parts—can probably be obtained locally for less than five dollars. All are standard parts, stocked by most retail and mail-order shops. In a pinch you can substitute the less-efficient International Rectifier B2M selenium cell for the quarter-round section of a Hoffman silicon solar battery.

First steps. Best place to begin construction is with the case for the radio—a small plastic box. You may find one in the medicine chest or at the druggist's. Start the necessary holes with a hot ice pick, then file or ream to size Keep try-

Author Halacy is an expert on the. sun. A couple of years ago he wrote Fabulous Fireball, *a book that tells the story of the sun and solar energy. Last fall. Macmillan also published his* Fun with the Sun. *a book describing how to make seven solar-energy projects at home. The present, article is an outgrowth of his research for the latter book.*

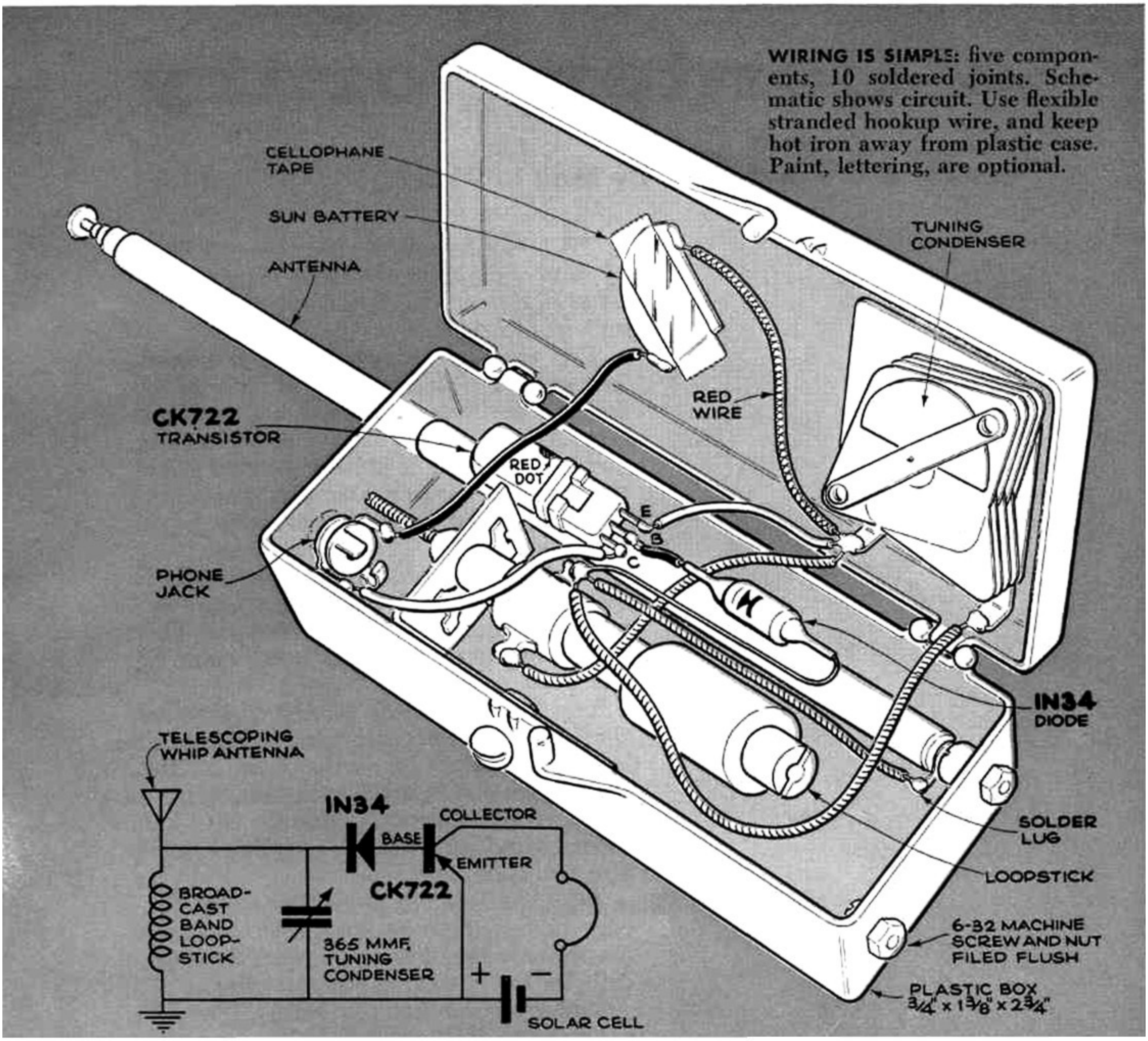
WIRING IS SIMPLE: five components, 10 soldered joints. Schematic shows circuit. Use flexible stranded hookup wire, and keep hot iron away from plastic case. Paint, lettering, are optional.

CELLOPHANE TAPE
SUN BATTERY
ANTENNA
CK722 TRANSISTOR
RED DOT
E
B
C
RED WIRE
TUNING CONDENSER
PHONE JACK
IN34 DIODE
SOLDER LUG
LOOPSTICK
6-32 MACHINE SCREW AND NUT FILED FLUSH
PLASTIC BOX 3/4" x 1 3/8" x 2 3/4"
TELESCOPING WHIP ANTENNA
IN34
BASE
COLLECTOR
EMITTER
CK722
BROAD-CAST BAND LOOP-STICK
365 MMF. TUNING CONDENSER
+
−
SOLAR CELL

ing the part as you enlarge the hole, to insure a snug fit.

The case can be painted or left transparent. The one pictured has a finish of Day-Glo spray paint, applied to the outside of the case. The paint softened the plastic slightly and created a wrinkled finish.

The hinged top is left clear and a piece of painted bond paper cut to fit inside. This paper was lettered, and cut to accept the dial shaft and solar battery. The lettering is pure whim—there's no trademark or commercial name.

After painting, go ahead with the assembly. Set the antenna mounting base through the bottom hole, put on the nut, cut off excess threads, and file smooth. Add a 6-32 screw and nut to the other side if you expect the radio to stand up.

Condenser and battery. Install the tuning condenser in the hinged top by first removing the dial and the small flathead screws. Remove both hex nuts; use one on the outside to hold the condenser in place. Then mount the dial and twist the silver center screw thumb tight. Place the solar battery in position over the opening in the bond paper, and fasten it in place with a strip of cellophane tape.

Bend the loopstick mounting bracket as shown, snap the loopstick in place, and fasten the bracket with a 6-32 screw and

[Continued on page 230]

Space-Age Crystal Set

[Continued from page 171]

nut. Install the phone jack, and the radio is ready for wiring.

Use stranded hookup wire; it can take the flexing imposed by opening and closing the case.

Pitfalls for amateurs. The wiring shown in the drawing is straightforward and follows the circuit schematic. If you're familiar with electronic wiring, you'll probably find some short cuts. If you're a beginner, stick to the diagram. If you run into trouble, don't hesitate to remove the parts, do the soldering outside, and then replace them.

Use care when soldering the diode and transistor—excess heat will damage them. Sidestep part of the risk by soldering to the lugs of a transistor socket, then slipping the transistor in place. You'll have to clip the transistor's leads to make it fit.

Bend the diode's leads to fit before soldering it between the base of the transistor and the loopstick.

Close the case when you've finished soldering, plug in the earphone, and listen. There's no on-off switch: The battery never wears out.

You should pick up stations at their proper position on the dial as indicated by the point of the battery or a dot on the lid. If the calibration is too far off, make a hole in the end of the case in line with the slotted screw on the loopstick. Turning this screw one way or the other will shift the stations enough to align the dial.

Outdoors, you'll find you don't have to point the Solar Mite at the sun—there is plenty of incident light for satisfactory operation. Keep the whip upright for maximum volume.

For stronger signals. Touching the silver center of the tuning dial may give you a stronger signal, since your body acts as a ground. Also try wetting your finger before touching the dial.

While the Solar Mite will play on the whip antenna, for peak performance connect the whip to your TV antenna. If you don't have an outdoor antenna of any kind, clip a lead on the finger stop of a telephone dial. You'll find that the radio works well enough to bring in a strong station even in the dark—acting as a crystal set without the help of its transistor amplifier.

No. 96

CRYSTAL SET CONSTRUCTION

SIMPLE INSTRUCTIONS FOR MAKING NINE CRYSTAL SETS

by B. B. BABANI

There is to-day a marked revival of interest in crystal receivers, and this book has been specially prepared to introduce the home constructor to the design and building of the various types. The text is simple and free from technicalities, so that even the young reader may confidently embark on the construction of any of the receivers shown.

CONTENTS

1. AERIALS.
2. EARTHS.
3. CRYSTAL DETECTORS.
4. COILS.
5. THE BEGINNERS' CRYSTAL RECEIVER.
6. FRAME AERIAL RECEIVER.
7. LONG DISTANCE RECEIVER.
8. HIGH GAIN RECEIVER.
9. HIGH SELECTIVITY RECEIVER.
10. BATTERY AIDED CRYSTAL RECEIVER.
11. MEDIUM AND LONG BAND RECEIVER.
12. ULTRA SENSITIVE ALL WAVE RECEIVER.
13. AN ADVANCED CRYSTAL DIODE RECEIVER.

THERE ARE 12 DIAGRAMS

LONDON: BERNARDS (PUBLISHERS) LIMITED

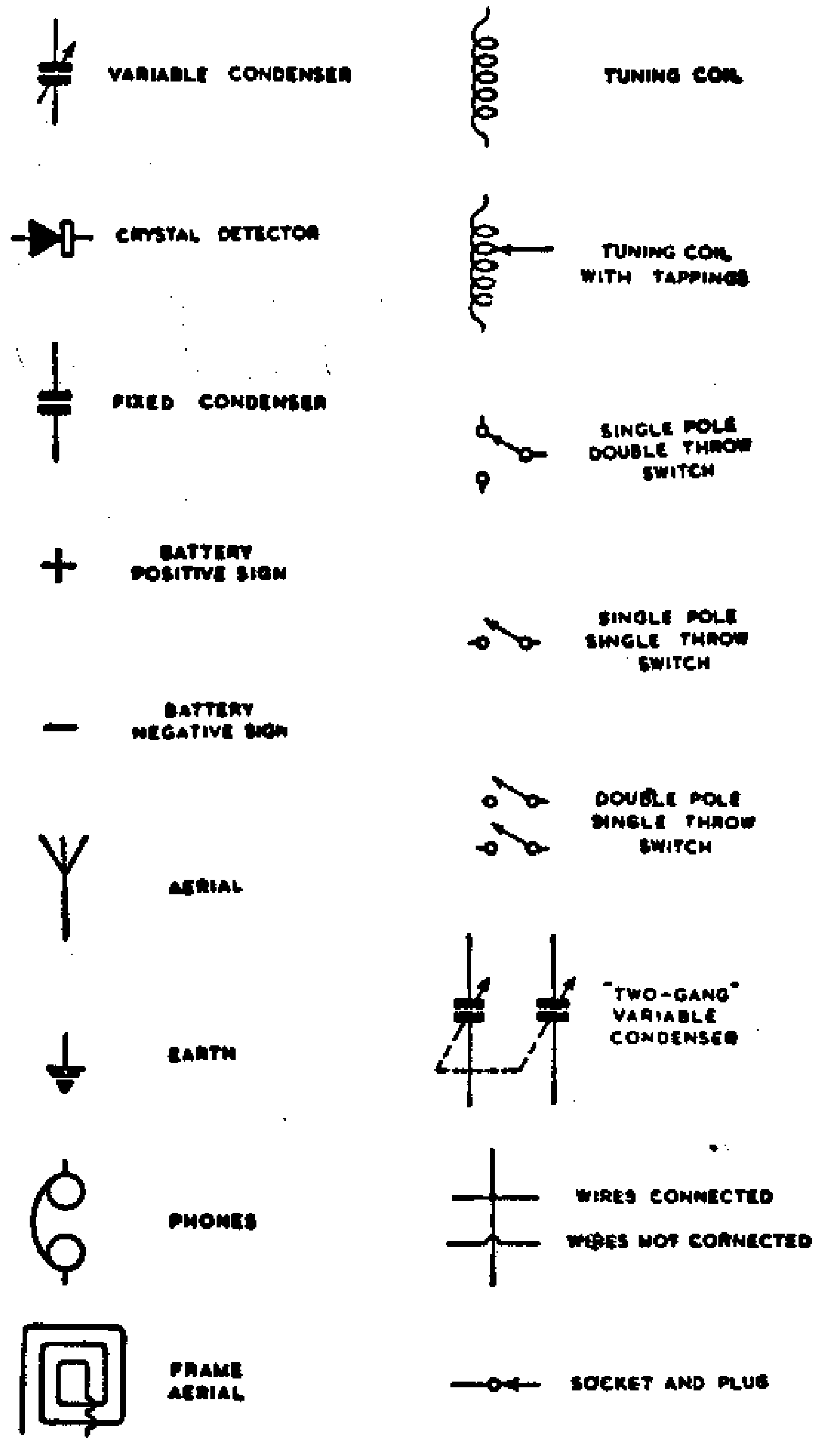
VARIABLE CONDENSER
TUNING COIL
CRYSTAL DETECTOR
TUNING COIL WITH TAPPINGS
FIXED CONDENSER
SINGLE POLE DOUBLE THROW SWITCH
BATTERY POSITIVE SIGN
SINGLE POLE SINGLE THROW SWITCH
BATTERY NEGATIVE SIGN
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
"TWO-GANG" VARIABLE CONDENSER
EARTH
WIRES CONNECTED
WIRES NOT CONNECTED
PHONES
FRAME AERIAL
SOCKET AND PLUG

FIG. I SYMBOLS USED IN THE CIRCUITS DESCRIBED.

1. Aerials.

One of the essentials for satisfactory reception with crystal sets is an efficient aerial. Four different types are suggested and are shown in Fig. 2.

There are several important points to be noted when erecting an aerial for a crystal set.

(1) See that it is suspended as high as possible and is not screened by tall buildings, obstructions, etc.

(2) The aerial itself and the lead-in from the aerial must be adequately insulated. Be sure that the porcelain or plastic insulators do not permit the aerial or lead-in to touch the building, tree, or post from which they are suspended. The lead-in should be brought into the house through an insulated rod.

(3) It is exceedingly important to use insulated copper wire, either single or multi-stranded, for both the aerial and the lead-in.

2. Earths.

An adequate and efficient earth contributes to the good performance of any well-designed crystal set and Fig. 3 shows the best method of obtaining an efficient earth where there is a garden or patch of ground available at a short distance from the receiver.

Obtain a 3-ft. length of ½-in. galvanised iron pipe and at one end drill a hole right through. Insert a bolt with spring washer and locking nut in the hole. Connect the earth lead-in wire to the bolt; the other end of the lead-in to the earth connection on the receiver. Leave both ends of the pipe open. Insert the end opposite the bolt into the ground vertically until only 3-in. are showing above the surface of the surrounding earth.

It is important to see that the surrounding ground is kept damp, if necessary by occasionally pouring a pint or two of water down the open projecting end of the pipe.

Ready-made copper earthing rods with a pointed end for easy hammering into the earth are sometimes obtainable from radio dealers.

Should a garden not be available, an alternative earth is a cold water pipe coming from a rising main. Simply file very lightly the pipe at the back of the tap and wind the copper wire earth lead-in round the filed part and cover the joint with insulating tape.

In no circumstances should a gas pipe be used for an earth.

3. Crystal Detectors.

Most radio dealers can supply very efficient permanent or semi-permanent crystal detectors of proprietary makes. However, there are always available quite a number of ex-Government surplus units which are called crystal valves and, though these were originally designed for a very different use in service equipment, they are easily adaptable and will give extremely efficient results when used in circuits designed for them. To enable the constructor to purchase suitable types for this purpose—and they cost very little—the following list gives the Government reference numbers.

CV 101, CV 102, CV 103, CV 111, CV 112, CV 113, CV 226, CV 241, CV 246, CV 247, CV 253, CV 291, CV 351, CV 361, CV 364, CV 367, CV 727, CV 749, and CV 1785.

One of the satisfactory proprietary makes of this class of crystal detector is the BTH CS7A, which, incidentally, is the commercial equivalent of the CV 253.

For all these Government types, one connection should be made to each end by arranging a holder to fix them into or by carefully soldering the appropriate connecting wires to each end.

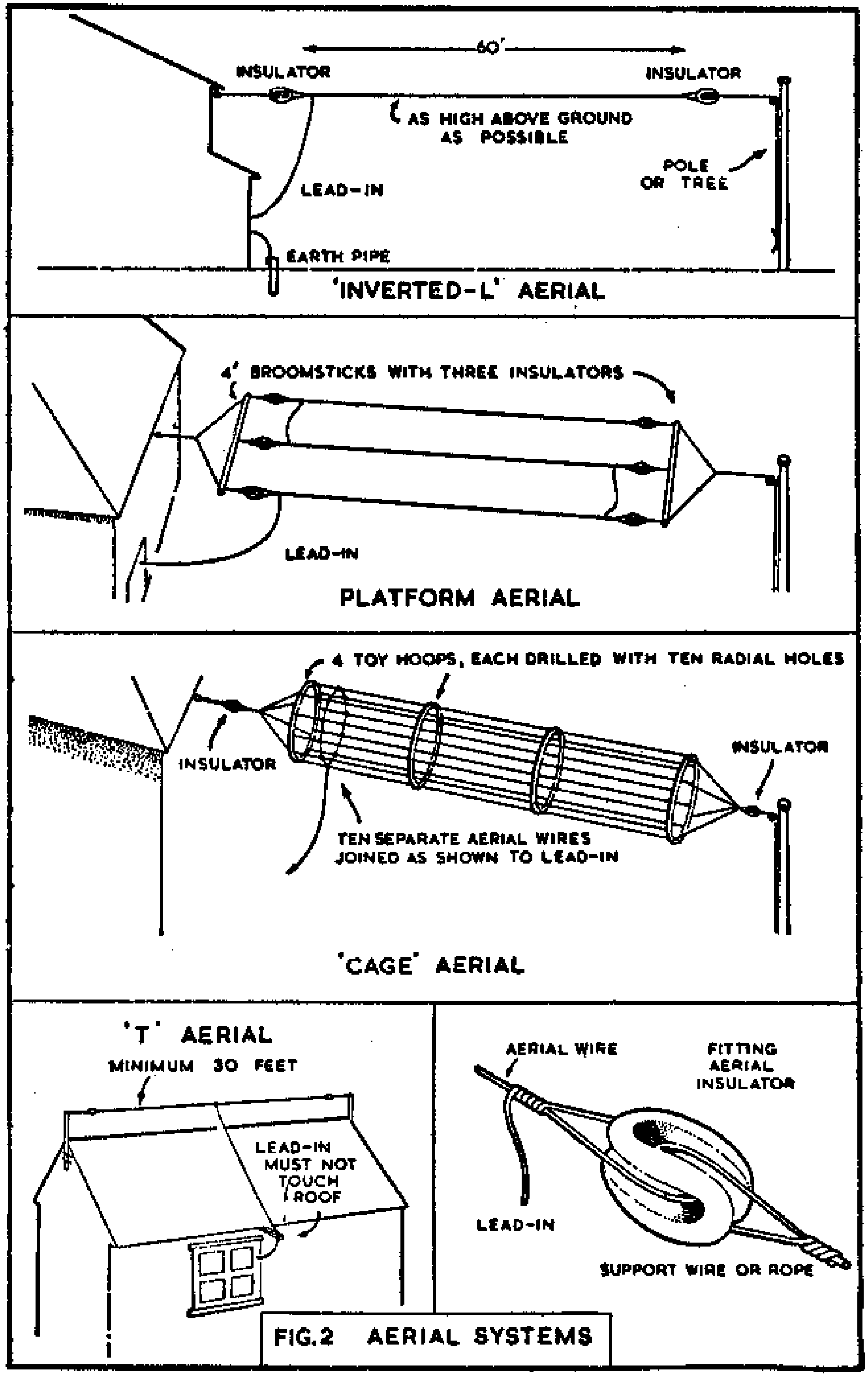
60'
INSULATOR
INSULATOR
AS HIGH ABOVE GROUND AS POSSIBLE
POLE OR TREE
LEAD-IN
EARTH PIPE
'INVERTED-L' AERIAL
4' BROOMSTICKS WITH THREE INSULATORS
LEAD-IN
PLATFORM AERIAL
4 TOY HOOPS, EACH DRILLED WITH TEN RADIAL HOLES
INSULATOR
INSULATOR
TEN SEPARATE AERIAL WIRES JOINED AS SHOWN TO LEAD-IN
'CAGE' AERIAL
'T' AERIAL
MINIMUM 30 FEET
LEAD-IN MUST NOT TOUCH ROOF
AERIAL WIRE
FITTING AERIAL INSULATOR
LEAD-IN
SUPPORT WIRE OR ROPE

FIG.2 AERIAL SYSTEMS

4. Coils.

In Fig. 3 two methods are shown of fixing the coil to the chassis. However, the ingenious constructor may evolve another arrangement.

The important thing to remember is that the coil should be fixed so that it is completely insulated from the metal chassis that may be used.

It is also important to remember that if a metal chassis is used, one must see that all components and wiring, with the exception of earth connections, are adequately insulated from this chassis so as to make sure that no short circuits take place. It is recommended that all crystal sets should be built on a panel, and that both panel and chassis should be of bakelite, paxolin or perspex, as these three materials are efficient insulators in themselves.

5. The Beginners' Crystal Receiver.

The popularity of the crystal receiver among beginners is undoubtedly due to the very modest outlay necessary, and the ease with which such receivers may be built. It is not always appreciated that, to obtain worth while results,

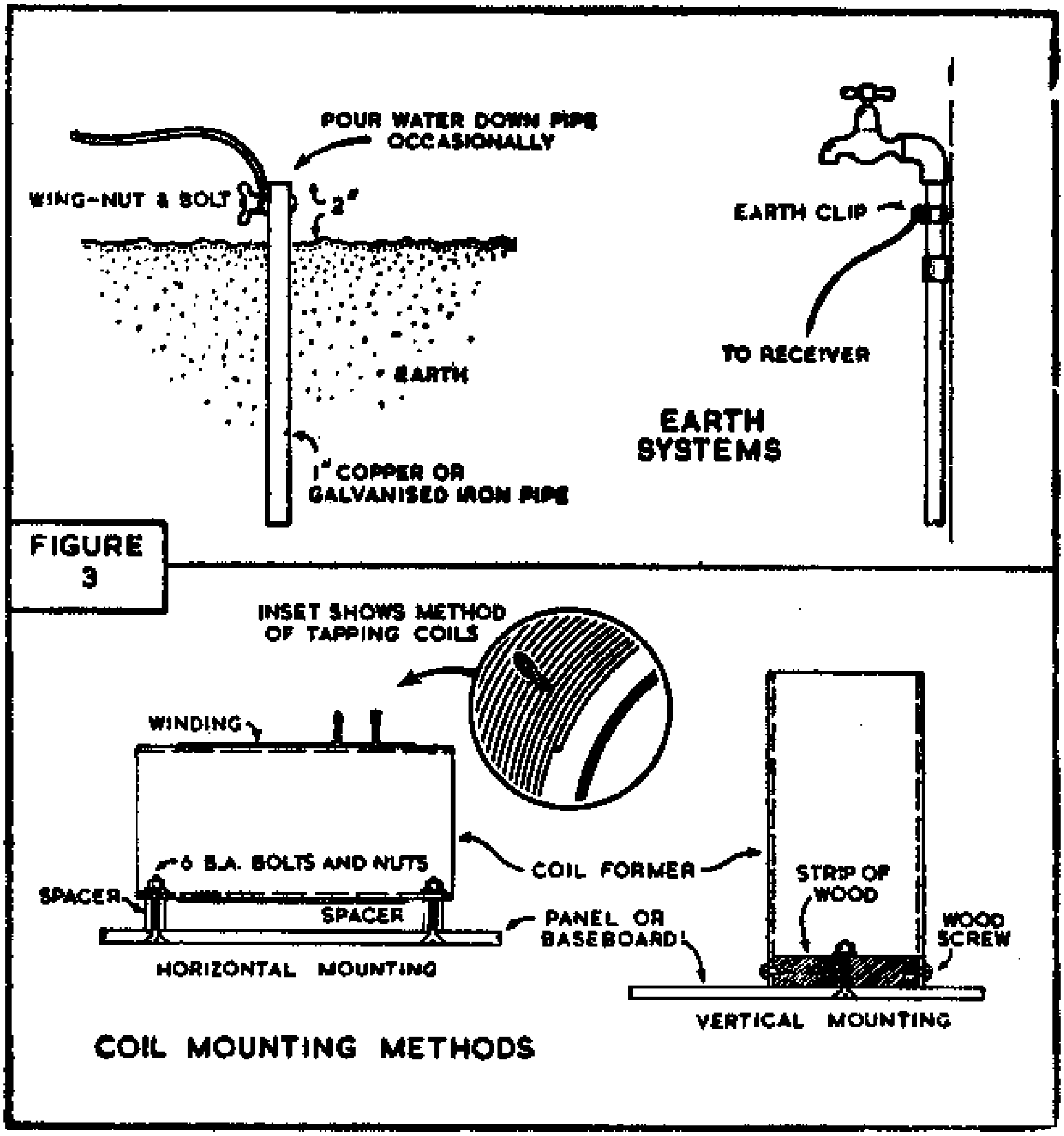
POUR WATER DOWN PIPE OCCASIONALLY
WING-NUT & BOLT
½"
EARTH
1" COPPER OR GALVANISED IRON PIPE
EARTH CLIP
TO RECEIVER
EARTH SYSTEMS
INSET SHOWS METHOD OF TAPPING COILS
WINDING
6 B.A. BOLTS AND NUTS
SPACER
SPACER
COIL FORMER
PANEL OR BASEBOARD
STRIP OF WOOD
WOOD SCREW
HORIZONTAL MOUNTING
VERTICAL MOUNTING
COIL MOUNTING METHODS

FIGURE 3

far more care in avoiding losses is necessary than with a three or four valve receiver. One reason for this is the fact that the crystal receiver can offer no help by way of amplification and it is therefore dependent on an efficient aerial/earth system, low loss design within the receiver itself, and the use of sensitive headphones. The experience gained is well worth the time expended since, when more ambitious receivers are attempted, the constructor will remember the benefits gained by efficient construction and installation with the result that each receiver attempted will be an instrument capable of giving a first class performance.

Losses within the receiver itself may be considerably reduced by efficient coil design. Unfortunately, selectivity is a factor which must not be overlooked; it is this factor which governs the sharpness of the tuning and permits the separation of adjacent powerful stations. The greater the selectivity, the greater the ease with which signals may be separated; but at maximum selectivity the received signals are at minimum strength. In Fig. 4, the coil is tapped at four points, which enables the best point between selectivity and sensitivity to be chosen. The nearer the aerial tap to the earth end of the coil, the greater the selectivity. The pre-set condenser C1 also aids selectivity; at minimum capacity the selectivity is high, and *vice versa.*

Used in conjunction with the coil taps, the constructor should have no difficulty in selecting a condenser setting and tapping point which permit the reception of strong signals and at the same time effect the required separation of powerful stations.

The coil is constructed on a former 3-in. diameter × 2-in., and consists of 50 turns 36 S.W.G. enamelled wire close wound and tapped at every tenth turn. These taps are brought out in the form of loops twisted firmly so that there is no actual break in the winding continuity. After completion the winding may be painted with Durafix to prevent any movement, and the taps and two free ends cleaned with emery cloth to remove the enamel. The layout and wiring is shown in Fig. 4. The headphones should have a resistance of some 4,000 ohms or more and should be of reputable manufacture. Any of the aerials previously described are suitable for use with this receiver.

C2 .0005 variable condenser.
C1 .0002 pre-set condenser.
C3 .001 mica condenser.
1 Crystal detector, semi-permanent pattern.
2 oz. 30 S.W.G. enamelled copper wire.
1 Coil former 3-in. dia. × 2-in.
1 Ebonite or Bakelite panel 6-in. × 5-in. × ⅛-in.
4 terminals, connecting wire, etc.

5. Frame Aerial Receiver for Local Station use on Medium wave bands.

This receiver was specially designed for constructors who are within 5 to 10 miles of a powerful station and who have no facilities for constructing a good outdoor aerial.

The frame itself consists of two ¾-in. × ¾-in. square section wooden rods each 36-in. in length. These are placed together at their centre line to form a cross; they are fastened by a nut and bolt and these in turn pass through an upright piece of wood ½-in. × 1-in. section 2-ft. in length. This upright support is then screwed into a baseboard 8-in. × 8-in. × 1-in. thick. At each of the arms of the 36-in. cross a 4-in. strip of ½-in. square section wood is placed. Reference to Fig. 5B will give the details of the construction of this aerial quite clearly.

It will be noted that approximately 85-ft. of 20 S.W.G. double cotton-covered wire are required for the inductance to cover the medium wave broadcast band. The wire should be wound around the four crossbars, thereby creating

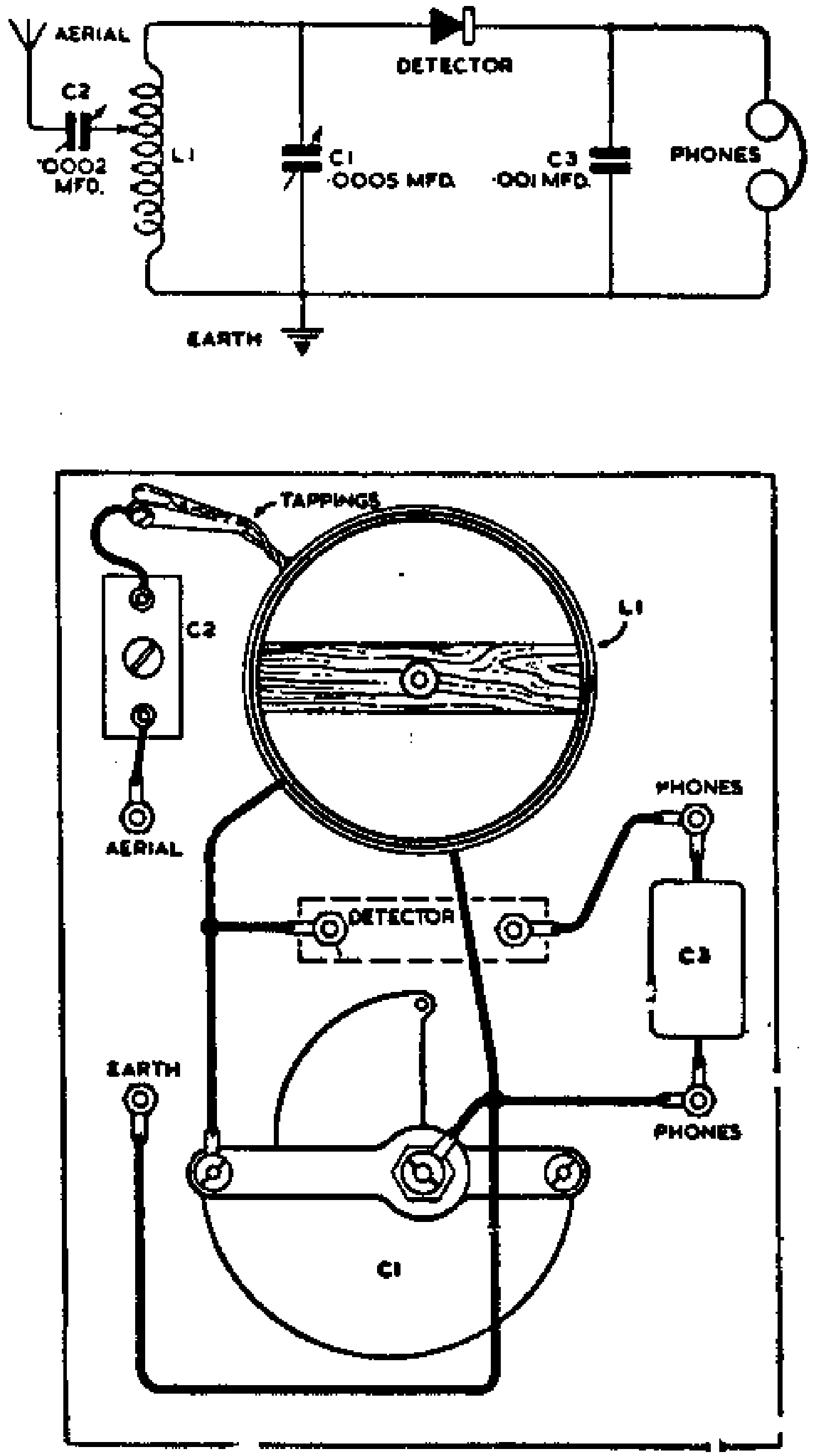
AERIAL
C2
·0002 MFD.
L1
DETECTOR
C1 ·0005 MFD.
C3 ·001 MFD.
PHONES
EARTH
TAPPINGS
L1
C2
AERIAL
PHONES
DETECTOR
C3
EARTH
PHONES
C1

FIG. 4 THE BEGINNER'S CRYSTAL RECEIVER.

a square wire frame, and each turn of the wire should be spaced approximately ⅛-in. The best way to effect this spacing accurately is to file on the 4-in. crossbars a number of small nicks at ⅛-in. intervals. The two ends of the wire should then be brought down on the two terminals which may be screwed into the square baseboard and two leads taken from there to the crystal set itself.

The actual set may be constructed on a small ebonite, paxolin or perspex panel 8-in. × 5-in. which may be screwed to one side of this base supporting piece of wood, thereby making the set virtually self-contained except for the earphones. Fig. 5A shows the circuit of the receiver.

To obtain the best results with this receiver, the aerial should be rotated by hand until the signals become loudest.

1 Semi-permanent crystal detector.
C1 1 .0005 mfd. variable condenser.
C2 1 .0005 mfd. fixed condenser.
1 condenser dial marked 0 to 100 or 0 to 180.
1 pair of high sensitivity earphones.
85-ft. to 90-ft. of 20 S.W.G. double cotton covered wire. (8oz.)
Sufficient connecting wire for wiring purposes.

7. Long Distance Receiver.

This receiver has been specially designed to achieve high sensitivity. With a sufficiently good outdoor aerial effective reception has been obtained up to 150 miles.

The circuit is quite straightforward, as can be seen from Fig. 6 and the only special point is in winding the coil L1.

This coil consists of a 2½-in. diameter former on which are wound a total of 51 turns using 24 S.W.G. double cotton wire. Starting at one end, tap every two turns until there are eight taps; wind a further 15 turns and tap; then a further 10 turns and tap; finally, add 10 turns to end the winding. The wire should be close wound, *i.e.*, with turns touching.

To obtain the best results, the two crocodile clips A and B should be tried on the various coil taps until the loudest reception is obtained. It may be found in use that the loudest reception of any particular station causes another station to be heard at the same time; care should, therefore, be taken to select the appropriate tapping point to separate any two powerful stations that may interfere with one another.

4 oz. 24 S.W.G. DCC copper wire.
Coil former 2½-in. dia. × 3-in.
C1-2 single variable condensers .00035 mfd.
2 Condenser dials marked from 0 to 100 or 0 to 180.
1 Permanent crystal detector.
C3 .001 fixed condenser.
1 pair of high sensitivity earphones (preferably between 4,000 ohms and 8,000 ohms impedance).
2 Crocodile clips.

8. High Gain Receiver.

This receiver has been designed to give good reception up to 50 miles distance from any normal power broadcasting station in the medium band. It is a very selective circuit and will therefore enable the user to separate closely situated stations.

It will be noticed in the circuit (Fig. 7) that the variable condenser C1, has an on/off switch S1 to throw it out of circuit if necessary. It is used to increase the selectivity of the receiver and, in operating, it should be tried with this condenser either in or out of circuit on the various tappings on the coil L1.

L1 consists of 90 turns of 22 S.W.G. double cotton covered wire wound on a 2-in. diameter former and tapped every ten turns. L2 separately wound on a 2-in. former, consists of 90 turns of 22 S.W.G. double cotton covered wire, also tapped as L1. The components required are given on page [illegible]

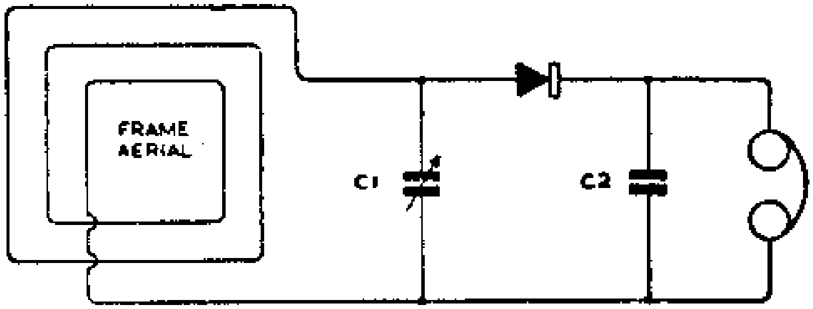
FRAME AERIAL
C1
C2

FIG. 5A FRAME AERIAL CRYSTAL RECEIVER

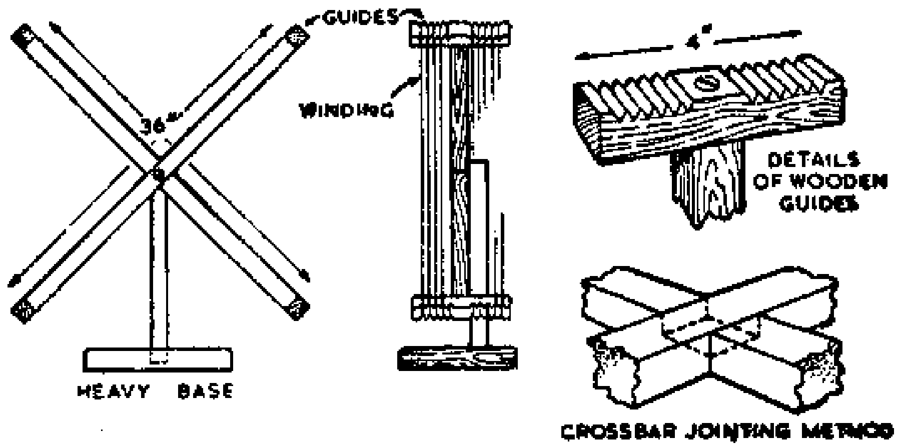
GUIDES
36"
WINDING
HEAVY BASE
4"
DETAILS OF WOODEN GUIDES
CROSSBAR JOINTING METHOD

FIG. 5B. DETAILS OF FRAME AERIAL

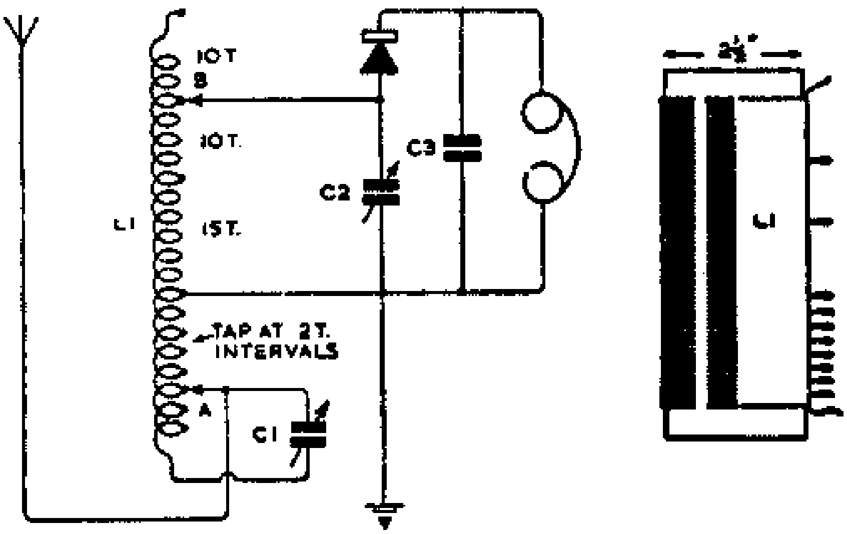
10T
B
10T.
L1
15T.
C2
C3
TAP AT 2T. INTERVALS
A
C1
2½"
L1

FIG. 6 LONG-DISTANCE CRYSTAL RECEIVER

C1-2 Single variable condensers .00035 mfd.
S1 Single pole single throw switch.
C5 .0001 preset condenser.
1 Permanent crystal detector.
C3-4 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
8 oz. 22 S.W.G. DCC.
2 Dials for variable condenser marked 0 to 100 or 0 to 180.

9. High Selectivity Receiver.

This circuit (see Fig. 8) is for use in high saturation strength areas where it is desired to separate powerful stations. The circuit is not complicated and is extremely effective.

It will be noted that C1 and C2, each .00035 mfd. capacity, are 2 gang unit.

The only special point in this set is the design of coils L1, 2 and 3, which are constructed as follows: L1, 60 turns of 22 S.W.G. double cotton covered wire on a 3 × 4 inches former tapped every five turns; L2, 15 turns of 22 S.W.G. double cotton covered wire on a 3 × 6 former; and L3, 85 turns of 22 S.W.G. double cotton covered wire wound on the same former as L2, the two windings being separated by approximately ¼-in. (see Fig. 8).

Note that the aerial lead is brought down to one of the tappings on wire L1. This should be tried on each of the tappings and the one giving the required selectivity should be used.

C1-2 2-gang variable condenser .00035 mfd.
2 3-in. formers for coils L1-2-3.
8 oz. 22 S.W.G. DCC.
C3 .0005 fixed condenser.
1 Semi-permanent crystal detector.

10. Battery-aided Crystal Receiver.

Fig. 9 shows a design which utilises a 9-volt grid bias battery to improve performance.

The charge built up in the .02 mfd. fixed condenser is changed to a positive current when it passes through the coil L1. This, in turn, is superimposed on the positive charge that is commencing to be rectified by the crystal detector and thereby increases the signal strength.

Coil L1 consists of a 4-in. diameter former wound with 85 turns of 24 S.W.G. double cotton covered wire tapped at 5, 25, 45 and 65 turns from the earth end. The earth end and these four taps are brought out to five small sockets and the connection from the earth is taken to a small socket. The plug should be tried in each of the five positions to find which one gives the greatest signal strength.

The 9-volt battery is connected across the .02 mfd. condenser and special care should be taken to see that the battery is connected correctly, *i.e.*, negative pole to earth.

1 Permanent crystal detector.
1 4-in. dia. × 3-in. former.
8 oz. 20 S.W.G. DCC copper wire.
C1 Variable condenser .0005 mfd.
C2 Fixed condenser .005 mfd.
C3 Fixed condenser .02 mfd.
1 pair of high sensitivity headphones.
1 Clix plug.
5 Clix sockets.
1 9-volt grid bias battery.

11. Medium and Long wave band Receiver.

This set has been designed for reception on two wave bands.

Care must be taken in winding the coils L1, 2, 3 and 4. These are wound on a single 3-in. diameter former. L1 consists of 30 turns of 28 S.W.G. DCC tapped at 8, 16 and 24 turns. L2, 60 turns of 32 S.W.G. DCC. L3, 40 turns of 28 S.W.G. DCC, and L4, 80 turns of 32 S.W.G. DCC. All are close wound and are separated from each other as follows: between L1 and L2, 3/16-in.;

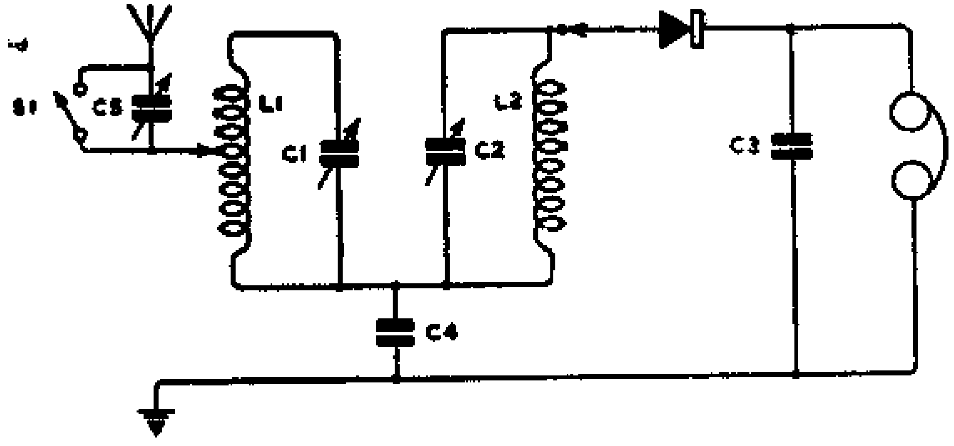
S1
C5
L1
C1
C2
L2
C3
C4

FIG. 7 HIGH-GAIN CRYSTAL RECEIVER

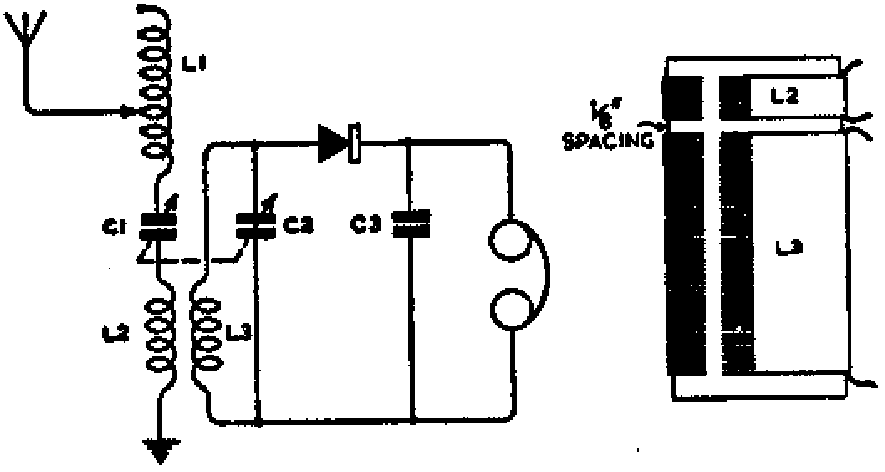
L1
C1
C2
C3
L2
L3
1/8" SPACING
L2
L3

FIG. 8 HIGH SELECTIVITY CRYSTAL RECEIVER

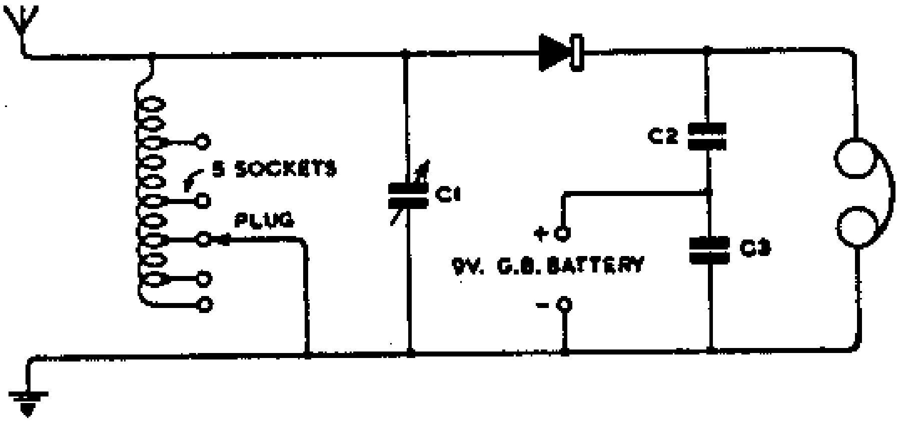
5 SOCKETS
PLUG
C1
C2
C3
+
9V. G.B. BATTERY
-

FIG. 9 BATTERY-AIDED CRYSTAL RECEIVER

between L2 and L3, 1-in.; and between L3 and L4, 3/16-in. See Fig. 10.

C1-2 single variable condensers .0005 mfd.
2 Variable condenser dials marked 0 to 100 or 0 to 180.
1 Clix plug.
3 Clix sockets.
S1 Single pole double throw switch.
C3 .001 mfd. fixed condenser.
1 Pair of high sensitivity earphones.
1 3-in. dia. × 8-in. former.
2oz of both 28 and 32 S.W.G. double cotton covered wire to make these 4 coils.
1 Permanent crystal detector.

In use, the plug should be tried in each of the three sockets to find the point which gives the best results. The switch is used as a wave band change switch.

12. Ultra Sensitive All Wave Receiver.

The receiver shown in Fig. 11 is for long distance reception. It is a novel circuit and, though apparently complicated in construction, it is really very simple to assemble with the minimum of tools and labour.

It will be seen from the circuit diagram that there are two banks of sockets, five black and six red. When using this set the appropriate black and red plugs should be inserted in whichever of the similar coloured sockets give the best results. The coil is quite easy to construct and consists of a primary winding which is inserted at the earth end of the secondary winding. For the medium wave band, the secondary winding L2 consists of 54 turns of 20 S.W.G. enamelled or cotton covered wire. This coil has four taps in addition to the starting point and the end connection, and these taps are taken at six turns, 14 turns, 27 turns and 40 turns from the earth end. The taps are taken out and a connection is made to an appropriate socket, as shown in the diagram.

The coil is wound on a 3-in. diameter former which may be a paper tube, bakelite, paxolin or perspex, and spacing is approximately 17 turns to the inch.

For short wave work down to about 30 metres, L2, the secondary coil, will consist of the same diameter former wound with 15 turns of the same gauge wire, the four taps being taken at 3, 6, 9 and 12 turns from the earth end.

L1, the primary coil, is used for both wave bands and consists of 11 turns of similar gauge wire to that used for L2, wound on a 2-in. × 1-in. former, the winding to be spaced the same as L2. The coil is then inserted at the earth end of whichever secondary coil L2 is being used.

The crystal detector may be one of the semi-permanent types freely available to-day; the old-fashioned cat's whisker pattern; or one of the silicon permanent detectors which are sold by many surplus dealers.

The 'phones should be of a standard type 4,000 ohms impedance although 'phones of other values will probably operate quite satisfactorily with this powerful receiver. It should, however, be stressed that the better the headphones used, the more satisfactory the reception.

Coils for other wave bands may be experimentally designed by the constructor, remembering that four tappings are advisable for the secondary coil L2.

A long aerial is very desirable with this receiver, placed as high as possible. It is recommended that a minimum length of 70-ft. be allowed, to include the aerial and lead-in.

The values of the components are shown in the following components list. The constructor should experience no difficulty in obtaining excellent results from this cleverly designed receiver.

C1 .0005 variable condenser.
C2 .002 fixed condenser.
1 semi-permanent crystal detector.
1 coil former 3-in. dia. × 4½-in.
1 coil former 2-in. dia. × 1-in.
¼ lb. 20 S.W.G. enamelled or DCC copper wire.
6 red wander plug sockets.
6 black wander plug sockets.
1 red and 1 black wander plug.

Headphones, terminals, connecting wire

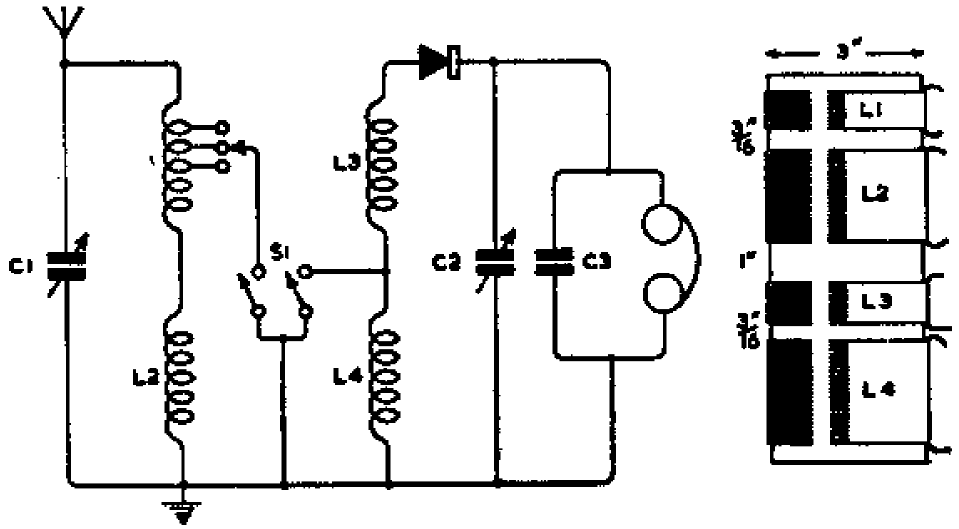
C1
L2
S1
L3
L4
C2
C3
3"
L1
L2
3/16"
1"
L3
3/16"
L4

FIG. 10 MEDIUM AND LONG-WAVE CRYSTAL RECEIVER

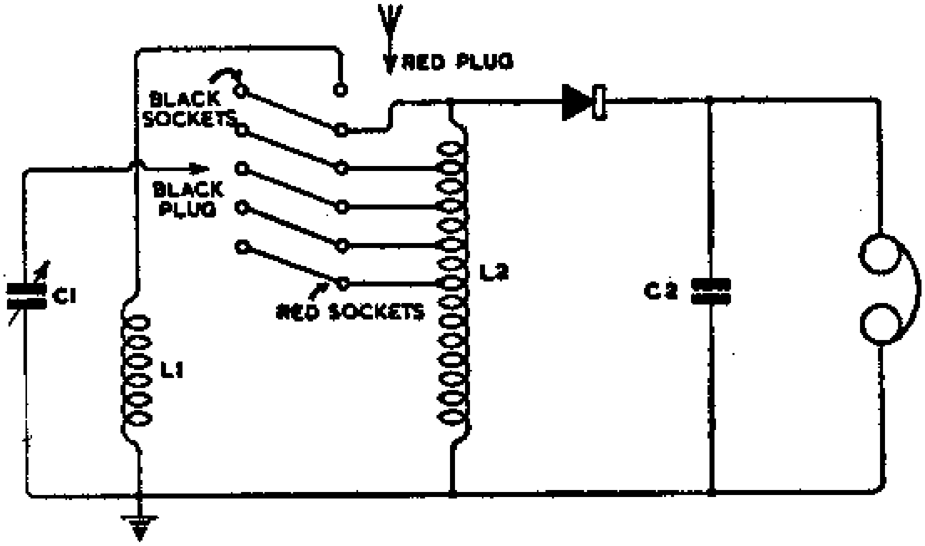
RED PLUG
BLACK SOCKETS
BLACK PLUG
RED SOCKETS
C1
L1
L2
C2

FIG. 11 ULTRA-SENSITIVE ALL-WAVE CRYSTAL RECEIVER

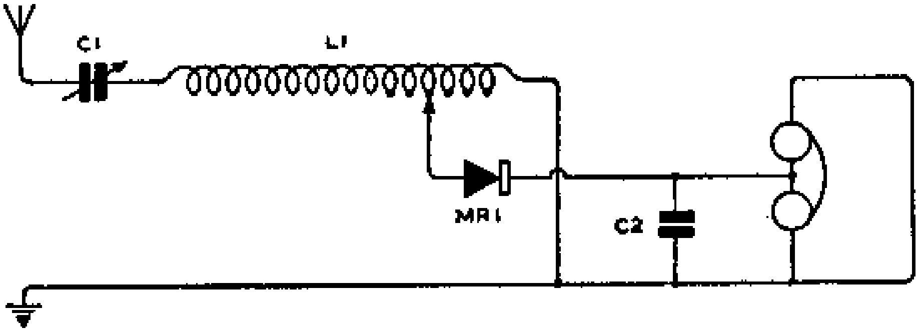

FIG. 12 ADVANCED CRYSTAL DIODE RECEIVER

13. An Advanced Crystal Diode Receiver.

The modern crystal diode of the radar type does not require any adjustment and thus has a distinct advantage over the ordinary crystal detector which needs delicate adjustment—even the semi-permanent type.

Such crystals are now obtainable on both the regular and the surplus markets, and a popular type is the B.T.H. CS7A silicon crystal (also coded as CV 253).

A rather different circuit from the usual crystal set arrangement is needed to suit the characteristics of a crystal diode. The diode must be tapped on to the tuning coil, and it is found that the tuning circuit itself gives best results if a series-tuned acceptor circuit is used. The circuit of a radar crystal receiver is shown in Fig. 12. The type of aerial used with this circuit has a very great bearing on the behaviour of the receiver for, in effect, the tuned circuit, with a series resistance equivalent to the reflected crystal load resistance, is in series with the capacitance of the aerial to earth and the aerial's effective series resistance.

At resonance—when the combination is tuned to any particular signal—the inductance resonates with the capacitances of the tuning condenser and the aerial in series, and the final effect is that the reflected load of the crystal, in series with the coil's R.F. resistance, is paralleled across the aerial's series resistance.

For maximum power transfer, the crystal load resistance must be made equal to the sum of the aerial series resistance and the coil R.F. resistance, and so the method in which the crystal is tapped into the tuning coil, and the exact capacitance required to tune the receiver to any required signal, must depend to a very great extent on the aerial itself.

At the same time the crystal resistance varies with the signal strength, the resistance being high for weak signals and dropping by as much as 50 per cent. and more for strong signals, so that this effect also has a bearing on the correct coil tap. The output impedance of the crystal also varies similarly, affecting the matching of the headphones into the crystal diode, and so for any set of conditions, the receiver requires to be matched up to both the aerial and the signal being received for best results. This would mean a series of coil taps and (theoretically, not practically) a matching transformer between the diode and the headphones; but in practice it will be found that a receiver using standard parts can be built up to give very good results under various conditions.

It has been shown already that the tuning of the receiver depends to a great extent on the characteristics of the aerial, and while a .0005 mfd. variable capacitor is shown in Fig. 12 as the tuner, the constructor must be prepared to experiment with different capacitance values until the required station is tuned in. The range of reception given by the receiver is quite good, if a really long and high aerial, and a good earth connection, are used; but no more than the local station signal can be expected, and the tuning therefore must be adjusted to suit the station frequency.

It must be mentioned that the headphones are shown parallel connected. High resistance headphones of the 4,000 ohm type must be used, and if two of these are connected in parallel rather than in the more usual series method, they will provide a roughly accurate match to the crystal. If more than one pair of headphones are to be connected in, then the pairs of headphones may be left series connected in the usual manner, the sets of headphones being connected in parallel.

The capacitor across the headphone terminals completes the crystal R.F. circuit, and any value between about 0.001 and 0.005 mfd. will serve. The higher capacitance will, of course, by-pass some of the higher audio frequencies, so that if good steady reception is obtained some experiment with this condenser is also worth while.

C1, .0005 mfd. variable condenser.
C2, .005 mfd. fixed condenser.
4oz. 26 S.W.G. enamelled copper wire.
M.R.1 CV 253 Crystal diode.
Coil former 1½-in. dia. × 4½-in.

Coil details are as follows: 150 turns of 26 S.W.G. tapped at 20, 25, 30, 35, 40 and 45 turns from the earth end. At each tapping-point the wire should be twisted up into a loop and the winding then continued without breaking the wire; when the coil is completed and the ends anchored the tapping loops can be bared.

BERNARDS RADIO BOOKS

No.		Price
56.	Radio Aerial Handbook	2/6
57.	Ultra-Shortwave Handbook	2/6
58.	Radio Hints Manual	2/6
64.	Sound Equipment Manual	2/6
68.	Frequency Modulation Receivers' Manual	2/6
73.	Radio Test Equipment Manual	2/6
83.	Radio Instruments and their Construction	2/6
96.	Crystal Set Construction	1/–
99.	One Valve Receivers	1/6
100.	A Comprehensive Radio Valve Guide, Book 1	5/–
103.	"Radiofolder" A. The Master Colour Code Index for Radio and T.V.	1/6
104.	Three Valve Receivers	1/6
107.	Four Valve Circuits	1/6
108.	Five Valve Circuits	2/6
121.	A Comprehensive Radio Valve Guide, Book 2	5/–
123.	"Radiofolder" F. The Beginners' Push-Pull Amplifier	1/6
126.	Boys' Book of Crystal Sets and Simple Circuits	2/6
129.	Universal Gram Motor Speed Indicator	1/–
134.	F.M. Tuner Construction	2/6
135.	All Dry Battery Portable Construction	2/6
138.	How to Make F.M. and T.V. Aerials, Bands 1, 2 and 3	2/6
141.	Radio Servicing for Amateurs	3/6
143.	A Comprehensive Radio Valve Guide, Book 3	5/–
145.	Handbook of AM/FM Circuits and Components	2/–
146.	High Fidelity Loudspeaker Enclosures	5/–
147.	Practical Tape Recording Handbook	5/–
148.	Practical Transistor Receivers, Book 1	5/–
149.	Practical Stereo Handbook	3/6
150.	Practical Radio Inside Out	3/6
151.	Transistor Superhet Receivers	7/6
155.	Portable Transistor Radio and Radiogram	2/6
156.	Transistor Circuits Manual, No. 1	2/6
157.	A Comprehensive Radio Valve Guide, Book 4	5/–
158.	Radio, Television, Industrial Tubes, Semiconductor and Diodes Equivalents Handbook (208 pages)	9/6
159.	Realistic High Fidelity	5/–
160.	Coil Design and Construction Manual	5/–
161.	Radio, Television and Electronics Data Manual	3/6
162.	High Fidelity Stereo Gramophone	5/–
163.	Transistor Circuits Manual, No. 2	2/6
164.	High Fidelity Tape Recorder for the Home Constructor	2/6
165.	Hartley on Hi-Fi, Book 1, Radio Tuners	5/–
166.	Public Address Systems	2/6
167.	Transistor Circuits Manual, No. 3	2/6
168.	Transistor Circuits Manual, No. 4	2/6
169.	Hi-Fi Transistor F.M. Tuner for the Home Constructor	3/6
170.	Transistor Circuits for Radio Controlled Models	7/6
171.	Super Sensitive Transistorised Pocket Set Construction	3/6
172.	International Radio Stations List	2/6
173.	Practical Transistor Audio Amplifiers	3/6
174.	Transistor Subminiature Receivers Handbook	5/–
175.	Transistorised Test Equipment and Servicing Manual	3/6
176.	Transistor Audio Amplifiers Manual	6/–
177.	Modern Transistor Circuits for Beginners	7/6
178.	Comprehensive Radio Valve Guide, Book 5	6/–
179.	Transistor Circuits Manual No. 5	5/–
180.	British Semiconductor Survey	15/–
181.	22 Tested Circuits Using Micro Alloy Transistors	5/–
182.	"At A Glance" Radio Valve & T.V. Tube Equivalents	3/6
183.	How to Receive Foreign T.V. Programmes on your Television Set by simple Modifications	5/–
	Resistor Colour Code Disc Calculator	1/6
	Engineers' Reference Tables	1/6
	International Radio Tube Encyclopaedia—3rd Edition	63/–

THE BOY'S BOOK
of
CRYSTAL SETS

BY

W. J. MAY

* * *

BERNARDS (PUBLISHERS) LTD.
THE GRAMPIANS
WESTERN GATE
LONDON W.6

General Editor

WALTER J. MAY

First Published—March, 1954
Second Impression—June, 1954
Third Impression—April, 1955
Fourth Impression—January, 1956
Fifth Impression—February, 1957
Sixth Impression—August, 1957
Seventh Impression—March, 1958
Eighth Impression—October, 1958
Ninth Impression—April, 1959
Tenth Impression—December, 1959
Eleventh Impression—April, 1960
Twelfth Impression—October, 1960

* * *

Printed by V. Cooper & Partners, Ltd., Flitcroft St., W.C.2
for Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.

PREFACE

The Publishers make no apology for reviving interest in the crystal set. For too long the technical press has immersed itself in an orgy of complex superheterodyne receivers and television equipment, all of which is necessarily expensive. In the 1920's, crystal set construction was a fascinating and inexpensive pastime indulged in by the majority of boys of that day. Today this pleasurable hobby can be even more attractive, the introduction of the germanium crystal diode and high performance coils, has opened up new paths which will capture the interest of boys of all ages. Perhaps most important of all, not only can a receiver capable of impeccable reproduction be produced without any technical knowledge; but the cost is no greater than that of a few visits to the cinema

⁂

CONTENTS

Page

Preface - - - - - 4

Introduction - - - - 5

Receiver No. 1 - - - - 12

Receiver No. 2 - - - - 13

Receiver No. 3 - - - - 14

Receiver No. 4 - - - - 16

Receiver No. 5 - - - - 19

Receiver No. 6 - - - - 20

Receiver No. 7 - - - - 23

Receiver No. 8 - - - - 24

Receiver No. 9 - - - - 26

Receiver No. 10 - - - - 27

Receiver No. 11 - - - - 30

Receiver No. 12 - - - - 31

INTRODUCTION

Before any attempt is made to construct a receiver, it is necessary to examine the problems which surround the crystal set, so that the best can be obtained from any of the designs attempted.

Firstly, it must be understood that the crystal set as it is to-day does **not** provide any amplification. It relies entirely on what is fed into it via the aerial and earth system and gives a very faithful replica of the original transmission.

From this it will be obvious that the aerial and earth system must be as efficient as possible if the final results are to be in any way outstanding. This because these are the only means by which the signals are fed to the receiver.

Secondly it is necessary to understand the nature of the transmitted signal, then it will be easy to understand the working of the set and to appreciate the function of each of the components.

When crystal sets first became popular, very little information was generally available, at least, not in a form that school-boys could understand.

We all built sets of all shapes and sizes, with coil designs that had to be seen to be believed, but very few of us had much idea of how they worked.

I well remember the case of a cousin of mine who, in those days acquired a magnificent variable capacitor or condenser as it was then called, this instrument had a most impressively engraved dial of polished ebonite, brass vanes and nickel plated end plates. Having observed similar dials on several highly priced commercial receivers in the town, he at once ripped out the somewhat tattered coil from the family receiver and replaced it with this device. The profound silence which ensued caused considerable amazement and dismay until a better informed adult explained the mysteries of L and C to him. Readers of this manual, however, will be better informed and there is no risk of failure with any of the designs described provided the instructions are carefully followed.

The Signal

To commence, we will assume that an orchestra is playing in a broadcasting studio. Since the principle of radio transmission is electrical it is necessary to change the sound produced by the orchestra into an electrical equivalent. This is carried out by the microphone, which picks up the sound and changes it into minute electric currents.

As they appear at the output of the microphone they are too small to be of use and accordingly are passed through a high power amplifier. These amplified currents could now be transmitted, but unfortunately, as we shall learn, owing to the inherent nature of the signal in this state, transmission over any useful distance would be impractical.

When the music from the orchestra is transformed into electrical currents they are in the form of alternating currents, usually called A.C., that is they rise to a maximum in one direction, fall to a minimum, rise to a maximum in the opposite direction and then fall to minimum again. This process is repeated over and over again. One complete rise and fall in each direction is called a cycle and is drawn in Fig. 1. Every time a note is struck on a piano, vibrations are sent out which reach the ear enabling you to hear it. These vibrations are also spoken of as cycles, they rise and fall in intensity the same way as an alternating current. The number of cycles radiated by any given note over a period of one second are referred to as its frequency. Middie C on the piano sends out 261 cycles every second and is known as having a frequency of 261. The microphone also " hears " the note and in the case of middie C produces minute A.C. at 261 cycles. This can be drawn as in Fig. 2, the only difference between Fig. 1 and 2 is, that the time factor is given so that the frequency can be identified. The higher the pitch of a note the higher the frequency and the lower the pitch the lower the frequency. On a piano the frequency of the top note is 3515 cycles and that of the lower 27 cycles. Those of you who have listened to an organ in a concert hall will have noticed that when a very deep note was played, it sounded like a growl to the ear, but the vibrating fre-

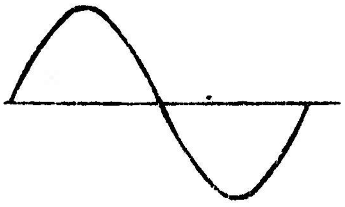

FIG 1

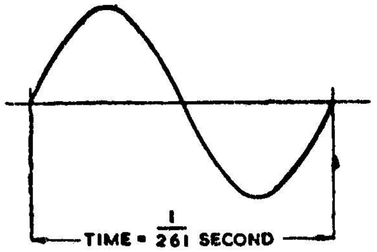

FIG 2

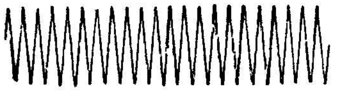

FIG 3

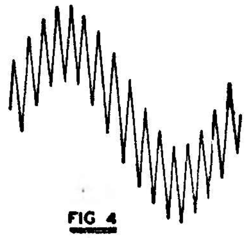

FIG 4

quency could be distinctly felt through the seat. Higher notes have too high a frequency to be observed in this manner.

The range of sounds which can be detected by the human ear are known as audio or low frequencies. From this you will understand that a low frequency amplifier is one which amplifies sound.

So that the transmitter will carry the programme over a useful distance it is necessary to radiate high frequencies. Now, as explained, the programme to be transmitted consists of low frequencies, and to overcome the difficulty, the transmitter generates A.C. of high frequency and combines it with the low frequencies. It will now be understood that the transmitted signal consists essentially of two different parts, a high and a low frequency content.

Fig. 3 gives a representation of the high frequency signal generated by the transmitter. In the case of the London Home Service, the frequency is 908000 cycles.

When referring to a high frequency signal on the medium or long wave-band it is usual to express the frequency in thousands of cycles, thus 908000 cycles becomes 908 kilo-cycles, which in turn may be abbreviated to 908 k/cs. It might at first be thought that by adding the low frequency or L.F. signal to the high frequency or H.F. carrier a form such as in Fig. 4 would result, such a combination is useless, and, so that the original L.F. content can be satisfactorily extracted by the receiver, the L.F. signal must vary the amplitude or output power of the H.F. signal as in Fig. 5.

It is in this form that the signal arrives at the receiving aerial. The aerial in itself is incapable of discriminating between one signal and another, and countless signals will be collected by the aerial at any one time. Many of these are too weak to be of use but the stronger ones must be sorted out since there is no point in receiving several programmes at once.

Fig. 6 shows the basic circuit of the input to a crystal. The coil L possesses a quality known as inductance, and the capacitor C, that of capacitance. If the coil had no capacitance whatever across it, all signals arriving at the aerial would be effectively short circuited to earth. As a matter of interest, it is impossible to obtain this state of affairs since even without any additional capa-

citor any coil **must** contain a certain amount of self capacitance.

By combining a coil and capacitor as in Fig. 6 a peculiar effect is observed, at one particular frequency, the signals are not short circuited to earth, but are developed across the coil. In other words the combined effect of L and C no longer provides a short circuit, **but only at one particular frequency.** If the value of C is altered the effect will be observed at a different frequency; likewise by altering L the frequency at which the effect will take place can be changed. There is a name for this phenomena, the frequency at which it occurs with any given L and C combination is known as the **resonant frequency.**

The values of the coils and capacitors shown in this manual have been carefully chosen so that resonance will be obtained at all frequencies where stations are broadcasting. Broadcasting stations work in bands of frequencies, those of major interest to crystal set constructors are the medium wave-band 1200 k/cs—600 k/cs and the long wave-band 300 kc/s—150 k/cs.

Usually a variable capacitor is used with a fixed inductance to cover one band and an additional coil switched in to increase the inductance to cover the other. In this way the L and C combination can be adjusted to provide resonance at the desired frequency of any given station. In other words you can **select the** station you want by varying C, that is turning the dial of the variable capacitor. This procedure is referred to as tuning. Having selected or tuned the required station it still remains necessary to change the form of the signal back to that of the original transmission. This process is called detection or de-modulation.

Examination of Fig. 5 will show that the signal has been duplicated, in other words, as it rises in one direction it also rises equally in the other. In this form the signal is useless since each half of the signal cancels the other, and if this signal is applied to a pair of headphones silence will result. Obviously some provision must be made to get rid of the unwanted half of the signal, and it is here that the crystal detector must be considered. This device will pass current in one direction only, ignoring any signal in the opposite direction, so that if the signal of Fig. 5 is passed through such a crystal, that of Fig. 7 will result. There is still the H.F. content to be reckoned

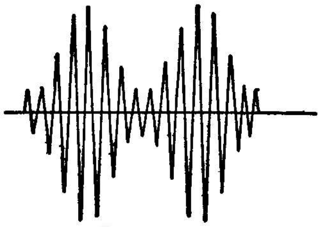

FIG 5

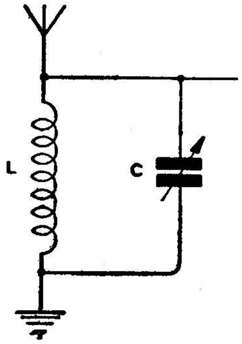

FIG 6

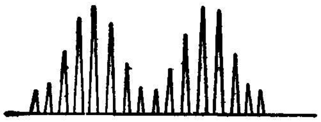

FIG 7

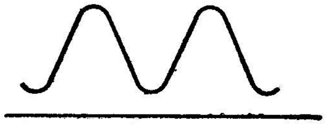

FIG 8

with, fortunately this is easily dealt with, a capacitor connected across the headphone terminals effectively disposes of this, leaving only the audio or L.F. content, as shown in Fig. 8. This audio content, which is a faithful replica of the original transmission is fed to the headphones. These in turn reverse the process of the microphone and transform the electrical currents into sound waves acceptable to the human ear.

Briefly then, your requirements are as follows:—

(1) A good aerial and earth installation, to make the most of the available signals.
(2) A receiver containing:—
  (a) Some form of coil and capacitor (L & C) combination to select or tune in the wanted station.
  (b) A crystal to get rid of the unwanted half of the signal (detection).
  (c) A fixed capacitor across the headphones to get rid of any remaining carrier.
(3) A pair of sensitive high resistance headphones.

Aerials

By this time the intending constructor will be able to appreciate the necessity of a good aerial. It is a point which cannot be over-emphasised. Assuming that you are in the fortunate position of being able to erect an outdoor aerial there are two main considerations, height and length.

One of the best that can be used is the inverted L shown diagramatically in Fig. 9.

It should be erected as high as is practical, every foot counts. The horizontal wire, that is the aerial proper, should have a minimum length of 60′ to which of course the length of the down lead is added. Where it is impossible to erect an aerial with an ideal horizontal length, a compromise must be effected.

Fig. 10 shows a three wire spreader aerial which gives quite a good effective length.

Suitable wire for a receiving aerial will not set any problems, stranded copper about 7/22 gauge is the best. 7/22 means that it consists of 7 strands of 22-gauge wire. This wire may be obtained covered, and for the present purpose is better than the plain or enamelled kind.

Note that insulators are used between the actual aerial wire and its anchoring supports, it is important that these are used, otherwise leakage will occur which will of course spoil its efficiency. Fig. 11 and 12 show how the wire may best be attached to two of the most common types of insulator available. Fig. 11 is of porcelain and is usually referred to as an egg insulator, whereas Fig. 12 shows a more modern (and more expensive) type in glass. If you are using the glass pattern, one is usually sufficient at each end of the aerial, but with the egg type two should be used. No doubt many readers will not be in a position to erect an out-door aerial, and must necessarily be content with an indoor installation.

The next best thing to a good out-door aerial is a replica constructed in a loft. If this form of construction is used, care must be taken when feeding the down lead to avoid close contact with the wall of the house. At the point where the lead feeds under the eaves, a length of rubber tubing can be used to cover the wire. The lead is fed through one of the small spaces left for ventilation purposes. A general idea is given by Fig. 13.

A less elaborate but quite effective aerial can be obtained by using a bed-spring. A length of 7/22 copper insulated wire is connected between the receiver and the spring. The spring should first be cleaned with emery cloth. Remove about 2″ of insulation from the wire and bind it tightly round the prepared spring. The joint may be covered with insulating tape which is obtainable from all electrical stores for a few pence.

Such aerials are quite popular since so many crystal sets are built for bedroom use.

When an aerial is required in the living room, the picture rail can be conveniently used. Insulated screw-eyes are fixed at intervals of about 3 feet along the rail, the wire is firmly anchored to one of them, and stretched right round the room until you arrive back at the starting point. By use of one of the insulated screw-eyes, the wire is secured and the down lead fed to the receiver. The idea is illustrated in Fig. 14.

Soldering

Before leaving the subject of aerials, a few words on soldering will not come amiss. Down leads on outdoor aerials should be soldered, and the same applies to loft types. Apart from aerial leads; earth leads and all the connections in the receiver will require soldering.

For the type of soldering necessary in radio construction an electric iron is the best solution. One of the small types marketed by Adcola or Henleys will be found admirable for the job. These are excellent for actual set construction but are hardly large enough for soldering the down lead to an outdoor type of aerial. Here the heat is dissipated much more quickly and a larger iron is required in order to get the solder to flow.

Ordinary irons which may be heated by a gas flame can be obtained very cheaply from most ironmongers' stores. Assuming the use of such an iron, first heat the iron until the copper bit is giving off a green coloured flame, the iron is now at the correct operating temperature. The bit is now discoloured or oxidised and its tip

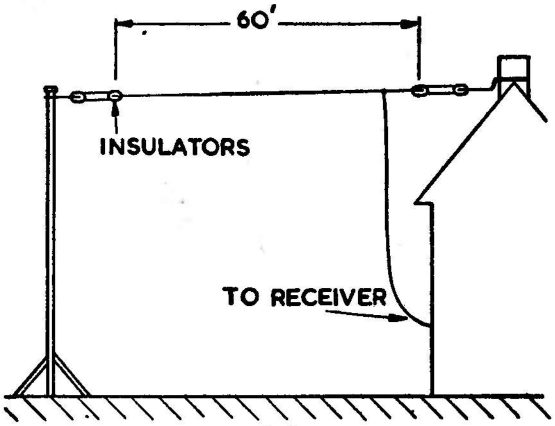
60'
INSULATORS
TO RECEIVER

INVERTED 'L' AERIAL

FIG. 9

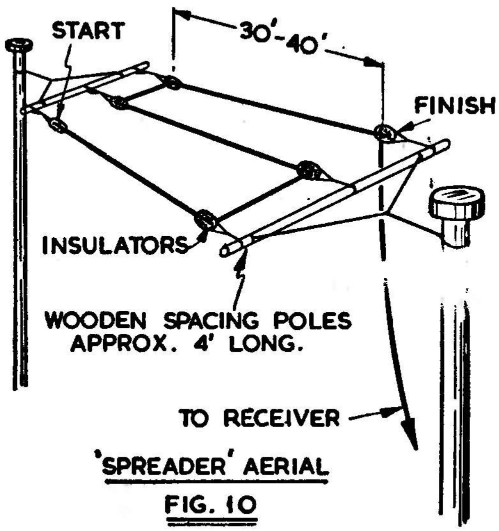
START
30'-40'
FINISH
INSULATORS
WOODEN SPACING POLES
APPROX. 4' LONG.
TO RECEIVER

'SPREADER' AERIAL

FIG. 10

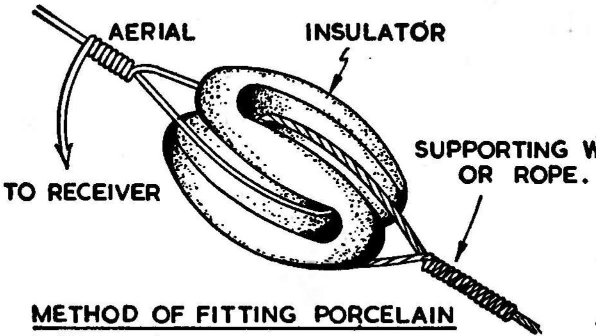
AERIAL
INSULATOR
TO RECEIVER
SUPPORTING WIRE
OR ROPE.

METHOD OF FITTING PORCELAIN INSULATOR

FIG. 11

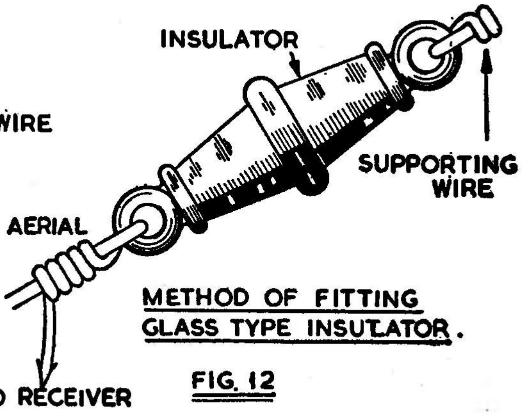
INSULATOR
SUPPORTING
WIRE
AERIAL
TO RECEIVER

METHOD OF FITTING GLASS TYPE INSULATOR.

FIG. 12

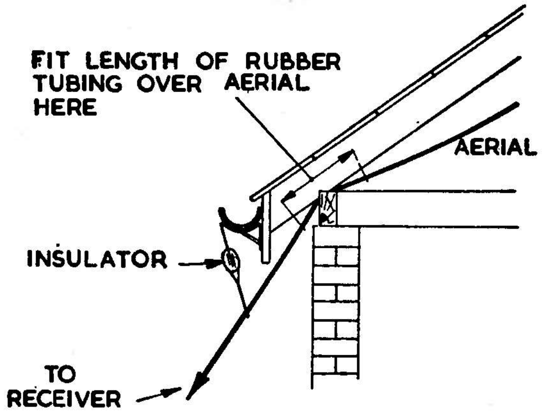
FIT LENGTH OF RUBBER
TUBING OVER AERIAL
HERE
AERIAL
INSULATOR
TO
RECEIVER

LOFT AERIAL.

FIG. 13

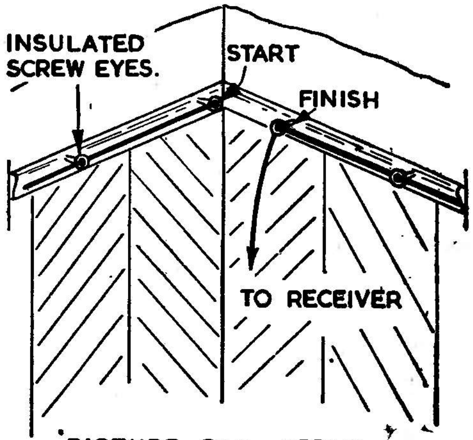
INSULATED
SCREW EYES.
START
FINISH
TO RECEIVER

'PICTURE RAIL AERIAL.'

FIG. 14

should be quickly cleaned with an **old** file which should be kept specially for the purpose. A better idea is to obtain a small block of sal-ammoniac and rub the tip of the hot iron on it. This will clean the tip of the bit perfectly. Next take a length of cored solder such as Ersin Multicore, and melt a little on to the prepared tip faces, now smooth it evenly over the surface with a piece of old rag (be very careful not to burn your fingers) the iron is now "tinned" and ready for use. When heating the iron, be very careful not to let the bit overheat or get red hot, otherwise the tinned surface will be destroyed and the whole process will have to be gone through again.

An electric iron will not overheat, and the tinning will last much longer than with ordinary types, and since the heat does not deteriorate there is less likelihood of making faulty joints.

Having obtained a tinned iron, the process of soldering joints is remarkably simple and anyone with a little patience can acquire the art in a very short time.

To solder two copper wires together: clean the wire with emery cloth, apply a prepared iron and some cored solder to it, the solder will flow evenly over the wire thus tinning it. Repeat the process with the remaining wire and then twist the two together. Now apply the iron to the joint from the underside and the cored solder to the joint on top. Solder will flow evenly over the joint. Remove both iron and solder and allow to set. The solder will harden or set in a few seconds but during this period the joint must not be touched or moved, as otherwise the joint will be "dry" and quite useless mechanically or electrically. On radio components, tags for soldering are already tinned though if they are old, or discoloured it is best to re-tin them. Normally, however, it is only necessary to twist the connecting wire to it and apply the iron and solder as explained.

Remember: **never** apply the solder to the iron and then the iron to the joint, always apply the iron **and** the solder to the joint. It is however a good thing to apply a little solder to the iron tip even when it is perfectly tinned just before making a new joint.

One final "DON'T," you will have noticed that the solder referred to is "cored," that is, it contains resin and other substances through its centre, plain solder as used by electricians and plumbers will not do, as it is the resin or flux

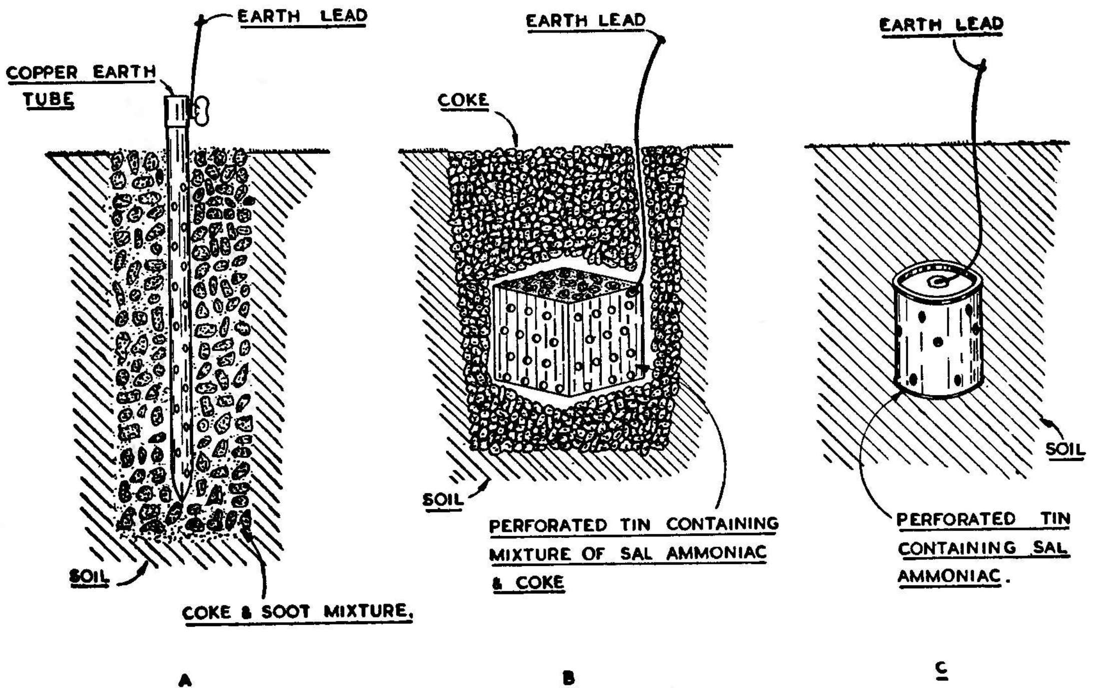
EARTH LEAD
COPPER EARTH TUBE
SOIL
COKE & SOOT MIXTURE.
A
EARTH LEAD
COKE
SOIL
PERFORATED TIN CONTAINING MIXTURE OF SAL AMMONIAC & COKE
B
EARTH LEAD
SOIL
PERFORATED TIN CONTAINING SAL AMMONIAC.
C

FIG. 15.

as it is called which makes the solder flow evenly and permits a good electrical joint.

Equally important do not attempt to use a separate flux or soldering fluid with plain solder or to "help" the cored solder; since these are almost certain to cause eventual corrosion and will destroy any components that have been contaminated.

Earths

The provision of a good earth is just as important as the rest of the installation.

Much disappointment would be avoided if this fact were not lost sight of, to avoid any slip up in this direction I am proposing to outline several well tried and efficient earth systems.

If it can be obtained, an earth rod, specially designed for the purpose provides the basis of a good earth connection. These were very popular some years ago when the majority of receivers were either crystal or battery operated, but in these days of modern mains driven sets they are not often used and consequently not always readily available.

Fig. 15 shows how it is used. First excavate a hole to a depth of three feet and fill with a mixture of soot and coke. Drive in the earth rod, which is a hollow copper tube, perforated. The lead to the receiver is connected at the top, and should be of covered 7/22 gauge, as used for the aerial system. Though not obvious it is important that this lead-in should be in covered wire, otherwise a number of indifferent earth contacts are likely to be made at various points along its length until it reaches the set. This is very undesirable and will spoil the efficiency of the system. It is essential to keep the soil surrounding the rod moist, which is one reason why the tube is hollow, and care must be taken to pour water into the tube at intervals.

A very efficient earth is the percolative type also popular at one time. Due to the chemicals used, it will extract moisture from its surroundings, thus maintaining a permanently moist earth.

It should be installed as in Fig. 15b.

The container is of copper or zinc, anything else will quickly rust away, again a good quantity of coke is used. Fill the container with a mixture of sal-ammoniac and coke, and then bury in coke as illustrated. This earth will not require further attention. It is possible to use powdered calcium chloride instead of sal-ammoniac but unless a chemist can be persuaded to make some up it is better to stick to sal-ammoniac.

If a zinc container is not available and you do not use or cannot easily get coke, Fig. 15c should be used. This makes a better earth than many so-called "earths" that I have come across. Obtain as large a tin as possible, make a number of holes as shown. Solder the lead in the bottom and fill it with sal-ammoniac. Replace the lid and bury in the ground. The tin will eventually rust away, but the replacement cost is negligible.

If it is quite impossible to make direct contact with the ground, a water pipe must be pressed into service. This should be a main pipe feeding straight to ground and not a hot water pipe or one fed from a tank. Scrape the pipe clean and twist the lead in tightly around it, a copper clip is even better. Do not attempt to solder on to the water pipe; since cold water is flowing through the pipe, it is extremely unlikely that your soldering iron will heat up the water supply sufficiently to allow a sound electrical joint, though you may spring a leak.

No attempt should be made to utilise gas pipes; the possibility of causing a fire is certainly very remote but, they make incredibly bad earth connections due to a number of joints made before true ground is reached. These joints are at best only semi-conductors, at least from an electrical standpoint.

Headphones

Since the late war there have been a large number of head-phone sets available on the surplus market. These may be roughly divided into two types, high-impedance and low-impedance. For the crystal sets detailed in this book high-impedance 'phones are required and the low-impedance pattern will not be suitable unless a matching transformer is used. As this is likely to cost more than the rest of the installation including the set, it will be as well to avoid them.

High-impedance types have an impedance of 2000Ω to 4000Ω whereas the low-impedance types are usually 600Ω.

The remaining consideration is weight; often cheap headphones are very heavy and uncomfortable to wear, every endeavour should be made to obtain 'phones as light in weight as possible.

Crystals

All the sets shown in this book have been designed to work with modern germanium crystals rather than the older galena crystal.

These germanium crystals require no adjustment, which, in itself, removes the main objection to this class of receiver. Suitable crystals are available from the following manufacturers and on the surplus market: Mullard, G.E.C., Brimar, Westinghouse and B.T.H. They are of robust construction, some in glass, others in ceramic or plastic, but must not be subjected to heavy knocks, otherwise the contact point may become dislodged.

From the information given, you should now be able to install an excellent aerial and earth system, and have some idea of how the crystal set works, so it is time to pass on to the actual receiver construction.

CONSTRUCTION 1

If you examine the following circuits you will find that each one is different. In most cases the difference lies in the coil design and/or the method by which the crystal and aerial is tapped into it. Each of these circuits has its own particular advantage to suit different conditions and the ideal circuit in some localities is not necessarily the best in others. It is not just a matter of a given circuit giving louder results than another, if it were there would be no point in showing more than one.

The main problem is to obtain adequate selectivity without reducing the volume level.

A receiver is said to be selective when it tunes sharply, a set with poor selectivity allows the stations to spread over the dial and when used near a transmitter will receive the local stations mixed together, which of course is useless.

Consider Fig. 16a, this is a very simple receiver, with no special attempt to provide any great amount of selectivity. In areas where signal

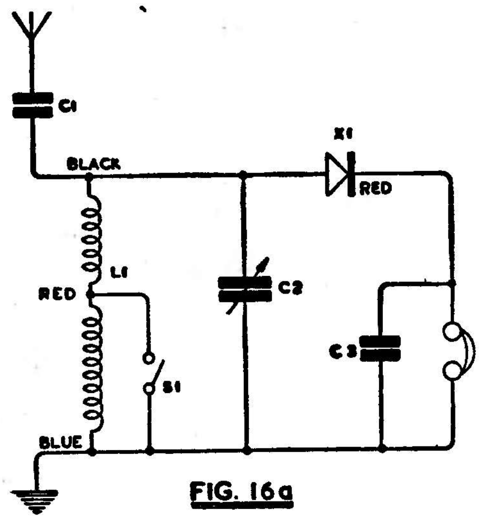

FIG. 16a

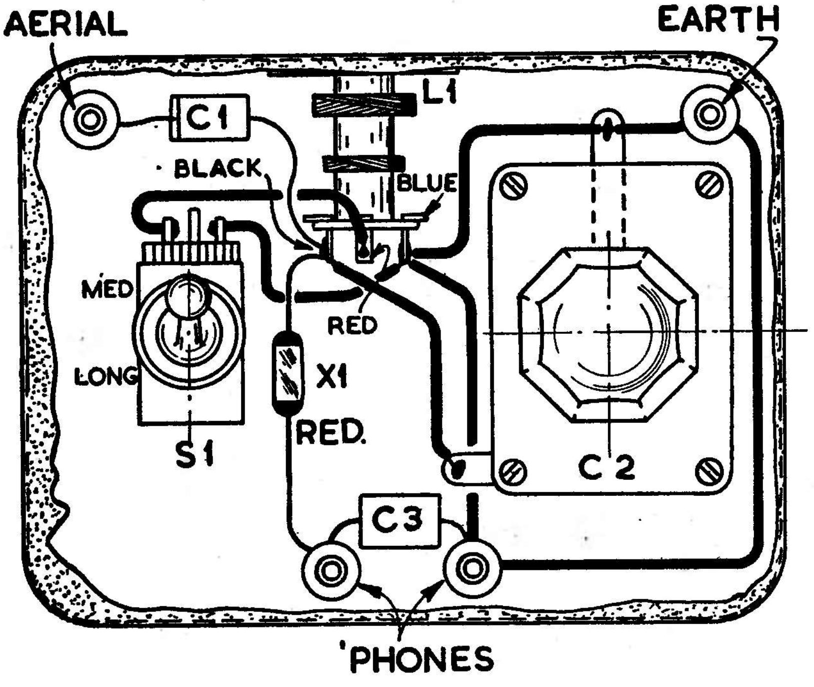

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG 16 b

strength is not high, or a short aerial is used, it will probably be ideal.

It would have been quite easy to increase the selectivity by providing a tap on the coil for the aerial as in Fig. 17a, but unfortunately, as the selectivity is increased overall volume is likely to decrease, so that unless you live close enough to the transmitter to have a large signal available and therefore need the selectivity, the circuit of Fig. 16a will be quite satisfactory.

Capacitor C1 is to prevent the aerial damping the circuit too heavily because this would flatten the tuning unnecessarily, however with small aerials it may be better to take the aerial direct to the black tag on L1. The signals are selected or tuned by L1 and C2, X1 is the crystal and C3 the capacitor across the 'phones to prevent unwanted carrier or R.F. reaching the 'phones.

A practical diagram (Fig. 16b) is provided showing the layout and all the wiring. You will require nuts and bolts to fix the coil, about ½", 4BA size will do, the other parts have locking nuts provided. An old 2-oz. tobacco tin makes a very good container and keeps the size down. Note that the metal box is connected to earth. With the exception of AC/DC receivers the metal work on any receiver or amplifier is connected to earth.

When S1 is open as in the diagram, the set will tune in long-wave stations, but when closed the medium wave-band will be received.

Components List, Fig. 16a

C1	100pF mica capacitor.
C2	500pF tuning capacitor, solid dielectric.
C3	1000pF mica capacitor.
X1	Germanium crystal.
L1	Crystal Set Coil. R.E.P.
S1	Single Pole toggle switch.
4	Insulated wander-plug sockets and plugs.
1	2-oz. tobacco tin (or similar container).

Make sure that the wander-plug sockets are of the **insulated** type, otherwise the metal case will join all the sockets together electrically.

Try to follow the theoretical diagram when wiring, a little practice will soon enable you to wire up a set without a practical diagram, which is a great advantage because often only the theoretical diagram is given when circuits are detailed in the technical press.

Fig 28 on page 34 gives a list of symbols used on the theoretical diagrams, so that you can readily identify the components.

2

Fig. 17a circuit is similar in many respects to that of Fig. 16a. The difference is purely one of selectivity, tuning will certainly be sharper, and even with comparatively inefficient aerials the design will put up a very good performance. There is no reason why the construction should not follow the same lines as the previous receiver but, by enlarging the set a little and using an air spaced tuning capacitor, the efficiency is improved. Observe that the colour coding on the coil, an R.E.P. Dual Range (Blue Box) is different from that of the Crystal set coil used on the previous design, and make sure it is correctly wired in.

C1 is a mica compression capacitor, often referred to as a " trimmer " and because it is adjustable, permits the set to be matched to aerials of varying lengths. Construction is carried out on a square panel of bakelite, perspex or wood, as shown in Fig. 17b. Perspex being clear like glass presents a most attractive finish, provided the set is neatly wired. It must be drilled slowly, however, otherwise the generated heat of the drill will make the hole wander, and spoil the panel.

After the set has been wired and tested a small wooden container can be made to house the completed set.

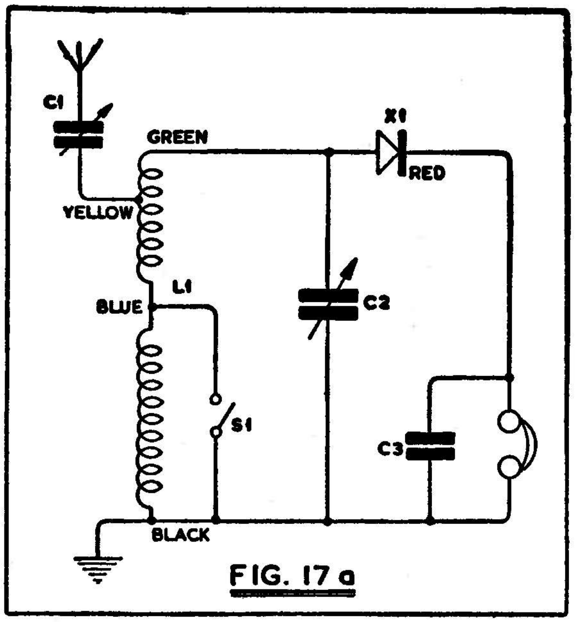

FIG. 17a

When carrying out reception tests, a little experiment is well worth while, to get the best out of the set. Try removing the germanium crystal from the green tag on L1 and connecting it to yellow, at the same time removing the lead from C1 to yellow and connecting it to green. Once the best arrangement has been found the wiring

can be left permanently in that position. Note that two tags on the coil are unused, this is intentional because on this circuit the extra winding connected to these tags is not required. With the switch S1 closed, the coil will cover the medium wave-band, and when open, long wave stations can be received. It is customary to abbreviate the expression medium wave-band to M.W. and long wave-band to L.W.

Components List, Fig. 17a

C1	100pF mica trimmer capacitor.
C2	500pF variable capacitor (air spaced).
C3	1000pF mica capacitor.
L1	Dual Range Coil R.E.P. (Blue Box).
S1	Single Pole toggle switch.
X1	Germanium Crystal.
4	Terminals or wander-plugs and sockets.
	Perspex, bakelite or wood for mounting panel.

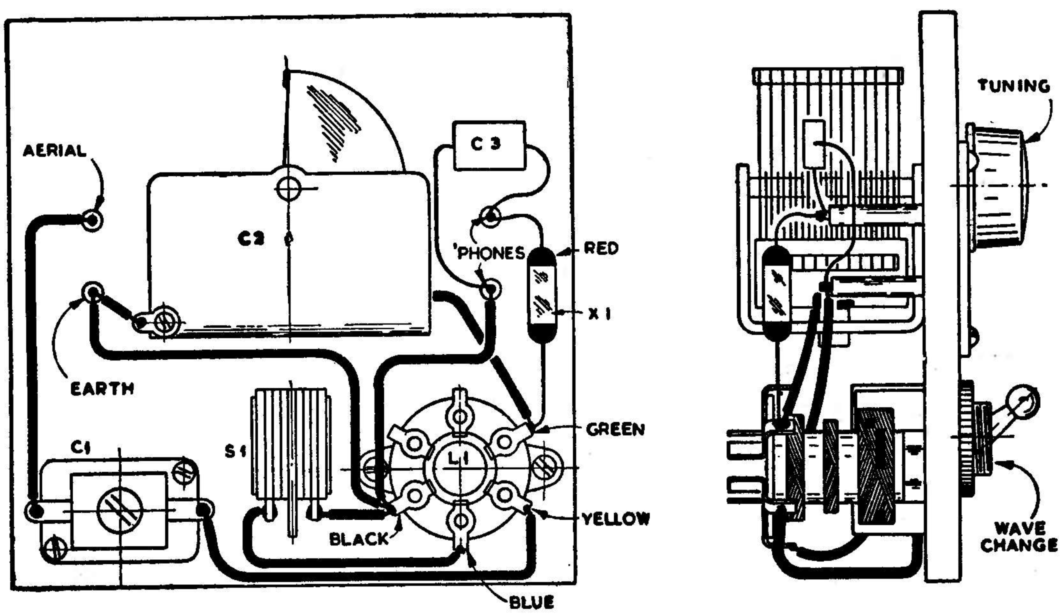

FIG. 17b

3

The design shown in Fig. 18a and 18b is more flexible than either of the preceding sets. That is, it can be varied to suit local conditions to a greater extent than the first two sets. This is made possible by the large number of taps provided on the coil.

You will notice that in this instance the coil is home-constructed on a cardboard former and is much larger than commercially produced coils. Coils wound in this manner are called **solenoids**, the turns are wound on side by side; in this case to a depth of some three inches. The diameter of the coil is made purposely large so as to obtain high efficiency. Modern commercial coils are invariably wave-wound and quite often are litzendraht which is the German for litz wire. This wire is made up of a number of strands of fine copper wire, each strand is enamelled to insulate it from the others, the whole is then silk covered. Litzendraht is more efficient than solid copper wire, and by using this and wave-winding, manufacturers can produce an efficient coil which is also small. Unfortunately wave-wound coils cannot be produced without a complex winding machine and home-constructed coils must take the solenoid form. It would be possible to use litzendraht but it is not easy to obtain by the reel, it is most expensive, and is difficult to handle. This last point is because at termination points, each strand must be cleaned of its enamel before a joint is made, and if one strand is broken, its advantage over plain copper wire is lost. How-

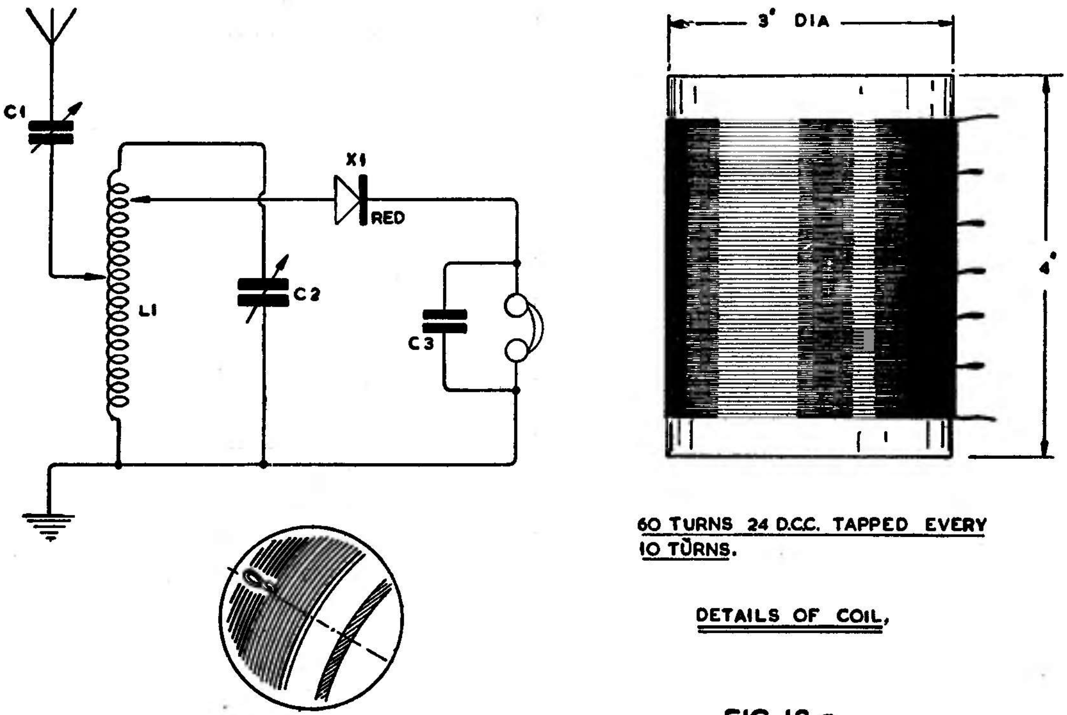

60 TURNS 24 D.C.C. TAPPED EVERY 10 TURNS.

DETAILS OF COIL.

INSET SHOWING METHOD OF TAPPING COIL.

FIG. 18 a

ever, by using a former of reasonably large dimensions an efficient coil can be produced cheaply using ordinary copper wire.

The wire must be firmly secured at the start and finish of the windings. Pierce three small holes about ¼″ apart ½″ from the end of the former. Pass the wire through the first from the outside, return it through the second and pass it back again through the third. Leave some 6″ of wire at the end to make off the connection. It will now be possible to wind the turns on tightly without wire slipping.

Count on ten turns and make a loop 1″ long. The method of preparing loops is shown in Fig. 18a, loops or taps should always be made in this manner, never by baring the wire and soldering a further length of wire to it. Carry on with the winding, making off the taps every ten turns as directed until the coil is complete.

Construction is carried out on a wooden base-board and front panel. Secure the coil to the base-board as shown on Fig. 18b, mount the remaining components and proceed with the wiring.

When testing out, it will be found that the further the aerial is tapped down the coil towards the earth end, the greater the selectivity. A position should be found which permits separation of local stations without excessive loss of sensitivity.

The crystal tap is adjusted for best results and different settings of C1 tried out, when choosing the best position for the aerial tap.

Components List, Fig. 18a

C1	450pF padder.
C2	500pF variable capacitor.
C3	1000pF mica capacitor.
L1	See text and Fig. 18a.
X1	Germanium Crystal.
2	Crocodile Clips (to connect leads to coil taps).
4	Wander-plugs and sockets.

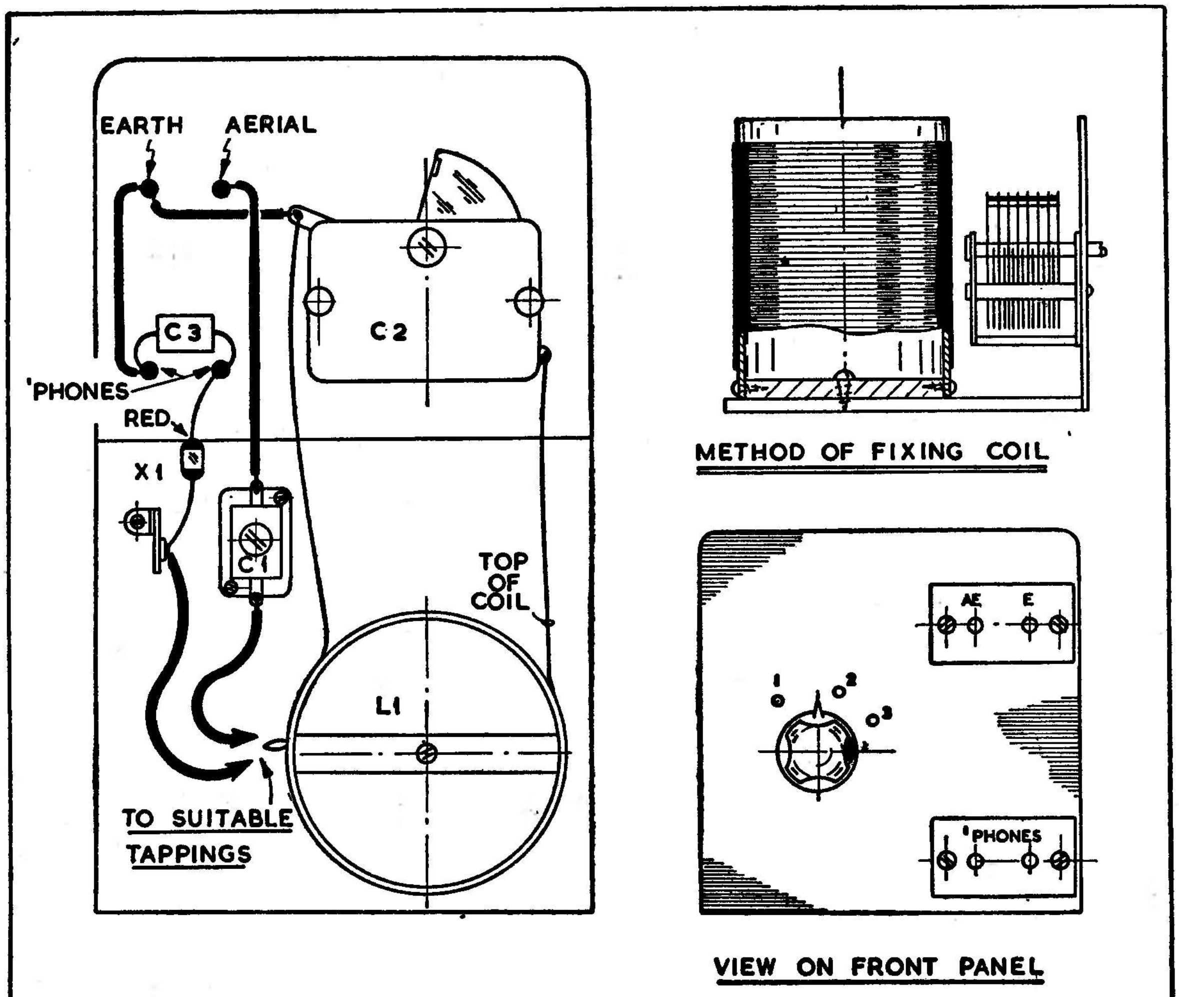

METHOD OF FIXING COIL

VIEW ON FRONT PANEL

FIG. 18 b

4

A most unusual design is reproduced in Fig. 19a.

It first appeared in America about three years ago, and it certainly does offer some advantage over more conventional sets. The tuning circuits are duplicated and two crystals are used, so the circuit may be described as a full-wave receiver. Signals received will certainly be louder than with simple sets, but care must be taken with the coil winding, as with the other receivers using home-made coils, otherwise results will be disappointing. Be very careful to ensure that all three windings are in the same direction, this is very important.

The distance between each winding should be ⅛". After the coil has been wound it is a good plan to warm it before a fire and paint the windings with "Durafix." Heating the coils makes the "Durafix" run freely. This substance sets quite hard and there will be no risk of the windings loosening. A small quantity of enamelled wire will cover the requirements of this coil, a 2 oz. reel will provide more than sufficient.

Before wiring in the germanium crystals, examine them carefully, note that one end is coloured red or in some cases marked with a positive sign thus +. You will notice, the sign is the same as the addition symbol used in arithmetic.

It is essential that both the red or positive ends are connected together, note that this is clearly marked on Fig. 19b. The receiver cannot work if

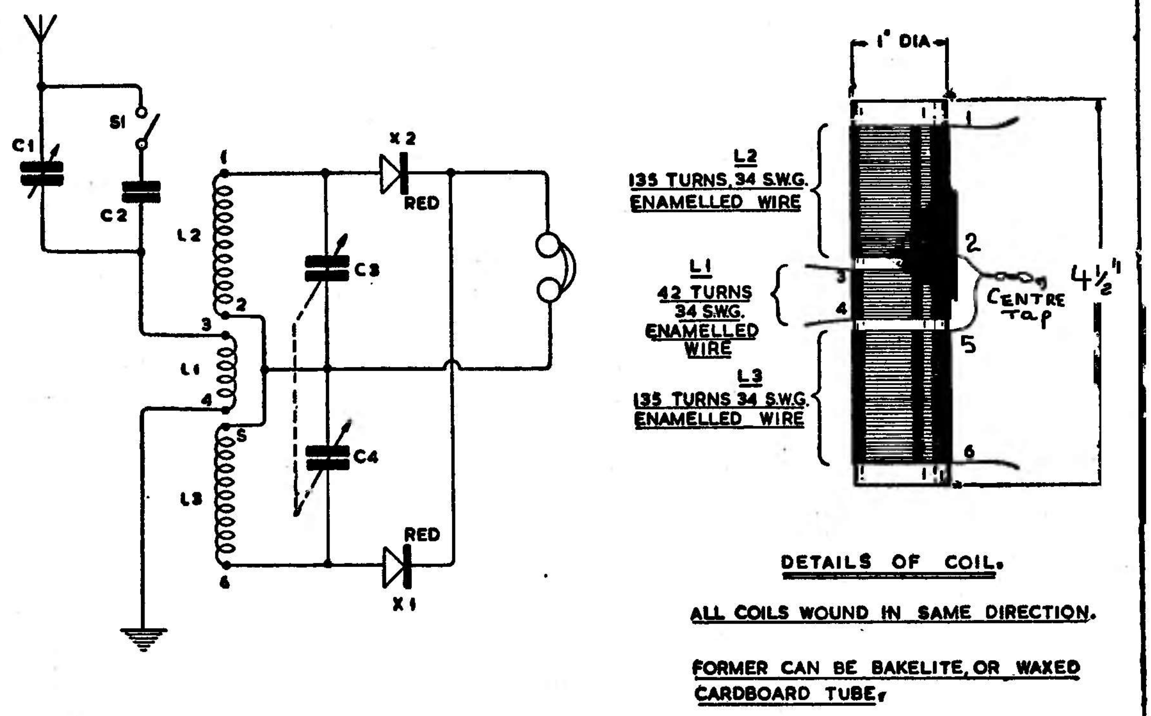

DETAILS OF COIL.

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE, OR WAXED CARDBOARD TUBE.

FIG. 19a

one of the crystals is connected the reverse way round.

A baseboard and panel form of assembly is used for this set which looks quite attractive if housed in a small polished cabinet. Dimensions are not at all critical but Fig. 19b gives a general idea of the layout that should be used. To avoid any unnecessary losses the coil should be mounted on perspex.

Obtain two strips of perspex 5½" long by ½" wide, drill fixing holes at both ends of each strip. It is best to clamp the two together when drilling these holes. Place one strip on the board in position, then put the coil and remaining strip over the first one and screw down. The coil will be clamped neatly and rigidly into position.

Twin socket bakelite strips are used to carry the aerial/earth and 'phone connections, if desired terminals mounted on pieces of bakelite could be used. These socket strips can be purchased from most shops selling components. If feet are not provided, drill two holes and screw into the edge of the base-board. Tag-strips are used to anchor some of the wires from the coil, and the germanium crystals. These also can be purchased for a few pence. At least two tags are necessary on one, and three on the other, though if the strips have more tags than required it is of no consequence. C3-C4, the ganged tuning capacitors have trimmers fitted, this should be stipulated when purchasing, and when first operating the set these small trimming capacitances are set about half-way. The idea of these trimmers is, that they compensate for any difference between the self-capacity of the tuning coils. With the switch S.1 open, that is in the OFF position, turn

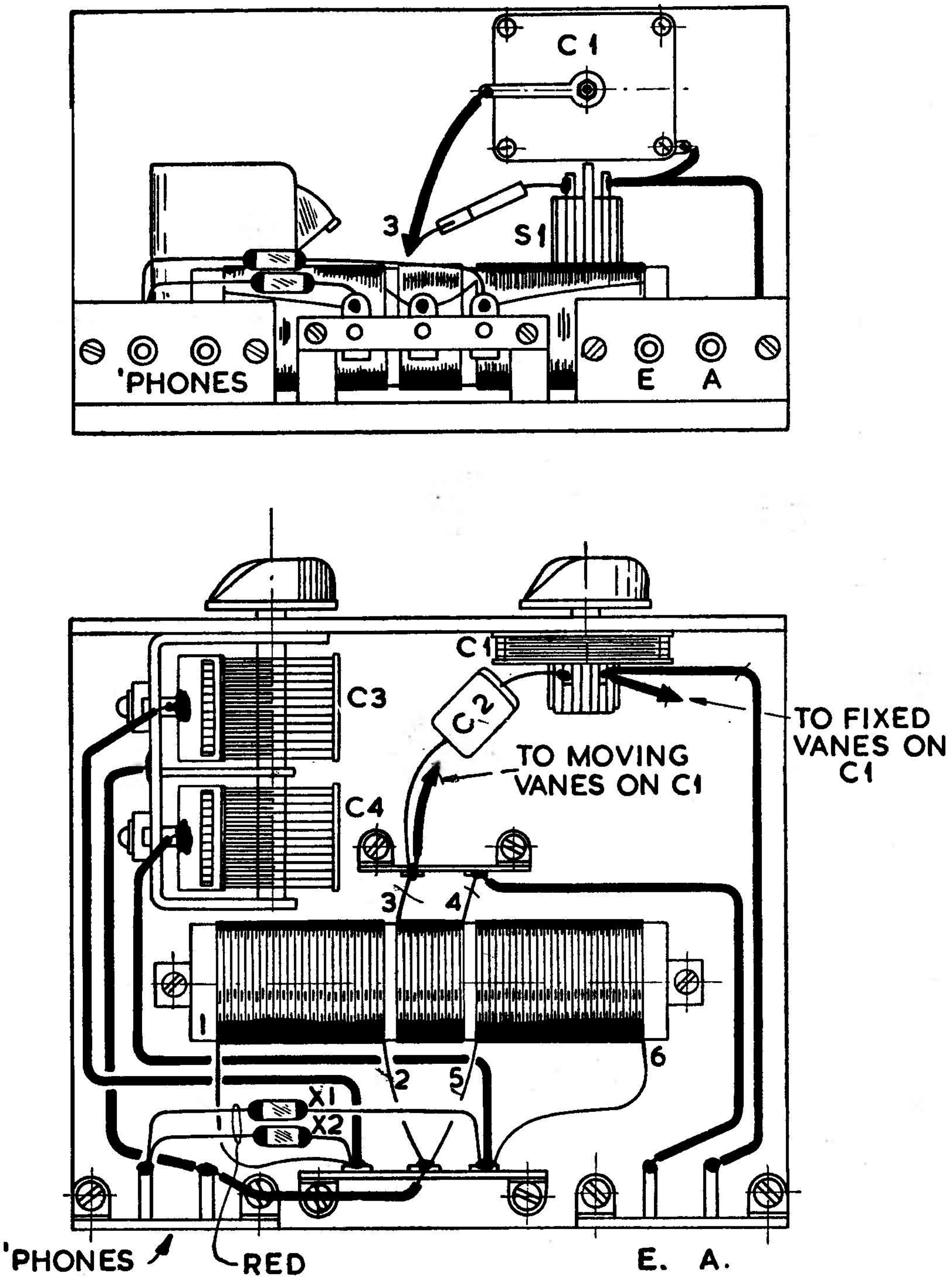
C1
3
S1
'PHONES
E
A
C1
C3
C2
TO FIXED VANES ON C1
TO MOVING VANES ON C1
C4
3
4
1
2
5
6
X1
X2
1
'PHONES
RED
E.
A.

FIG. 19 b

in a station at the high-frequency end of the band, that is with the vanes of C3-C4 towards the disengaged position. Adjust C1 for maximum volume, without interference from other stations. Now adjust the trimmers for loudest headphone strength. It may be necessary to run over the adjustments several times for best results. Once they are correctly set, the trimmers require no further adjustment. The tuning coil covers the M.W. band, 1500 kc/s. to 600 kc/s., as the capacitor C3-C4 is advanced and the vanes start to mesh, the frequency to which the receiver tunes **decreases** so that with the vanes disengaged the set is tuned to 1500 kc/s. approximately and at full mesh, 600 kc/s. Switch S1 is used as follows, for stations **lower** in frequency than 850 kc/s it should be in the ON position with the contact closed, but for stations **higher** in frequency, it is in the OFF position. As an example both the London Home Service and London Light Programme transmitters on the M.W. band operate at a higher frequency than 850 kc/s. If any doubt exists as to the operating frequency of a given station, reference should be made to the Radio Times which quotes both wave-lengths and frequency. If you know the wave-length in metres, it is simple to find the frequency. Divide 300,000 by the wave-length in metres, the dividend equals the frequency in kilocycles, i.e., 300,000 ÷ 300 metres=1,000 kc/s. In the same way, dividing 300,000 by the frequency in kilocycles will produce the wave-length in metres.

To use the receiver, tune in the signal by C3-C4, adjust C1 for maximum volume without allowing stations to overlap.

Components List, Fig. 19a

C1	300pF solid dielectric variable capacitor.
C2	1000pF mica capacitor.
C3-4	500pF twin gang variable capacitor.
L1-2-3	See text and Fig. 19a.
X1-2	Germanium Crystal.
2	Tag-strips.
2	Twin socket strips.
S1	Single Pole toggle switch.

5

Whenever crystal sets are discussed, constructors are apt to think in terms of medium-wave reception. When you consider that at least 98% of published circuits are designed for this band (sometimes with the long-wave band thrown in as an afterthought) it is understandable. However, if you can provide a good outdoor aerial (and an equally good earth) there is a lot of fun to be had listening to the short-wave bands. This receiver is designed specially for short-wave reception.

A metal chassis is used for construction, these can be obtained ready made in aluminium from most good supply houses. It can be quite small and on the original model a 6″ x 4″ was used. Low loss components are used so as to obtain greatest efficiency on the short wave bands. The coil is a commercial product which plugs into a 4-pin base. Best results were obtained on the 3mc/s to 7mc/s band though coils covering other bands are available.

Note that the tuning capacitor C1 is smaller than normally used and has a maximum capacitance of 140pF.

Coil type 706/R covers the 3mc/s to 7mc/s band, but the set is equally satisfactory on the M.W. band, and to cover this coil type 706P should be used. This coil has an iron dustcore which can be adjusted to make the coil cover the required band. The effect of the core is as if turns were being added or removed from an ordinary coil. When dealing with the short waves it is customary to refer to the frequencies in terms of mega-cycles (mc/s) rather than kilo-cycles. One mega-cycle is equivalent to 1000 kilo-cycles, that is 1,000,000 cycles.

Layout and wiring can be clearly followed from Fig. 20b and a theoretical diagram is shown in Fig. 20a. Care should be taken when arranging a mounting for the aerial terminal. The best plan is to cut a 1″ hole in the chassis, mount the aerial terminal on a piece of perspex and screw into the chassis. This avoids any unnecessary losses between aerial and chassis. The earth terminal is screwed directly into the chassis.

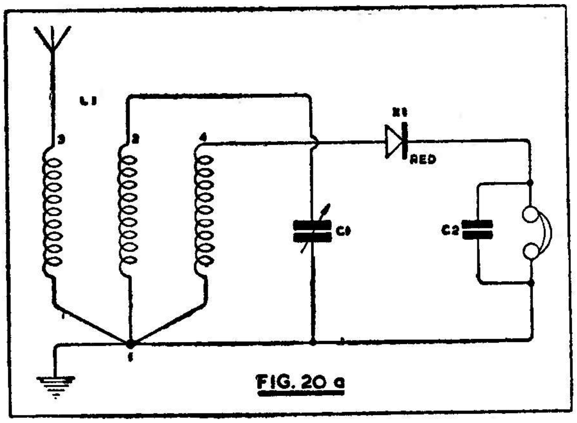

FIG. 20 a

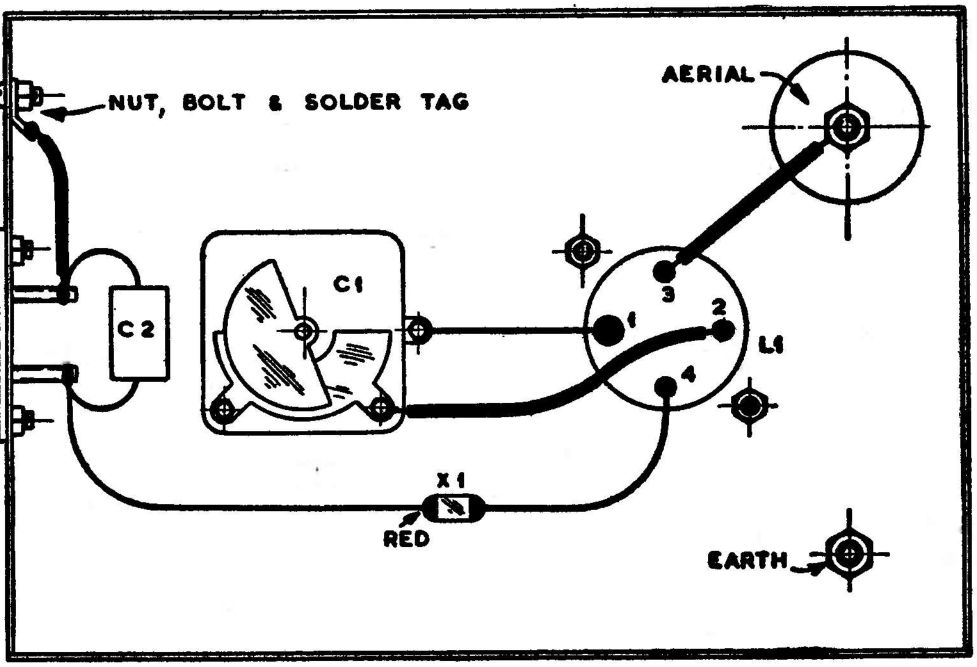

FIG. 20 b

Wiring as shown in Fig. 20a is likely to give the best results, but other arrangements are possible and to get the best out of the set they should be tried.

Refer to Fig. 20a. The aerial is connected to pin 3 on L1, C1 to pin 2 and X1 the crystal to pin 4. If you have a long aerial try this combination—Aerial to pin 4, C1 to pin 2, and the crystal to pin 3. Connections to pin 1 are not altered. In a few cases the following arrangement may prove best. Crystal and C1 to pin 2, aerial to pin 3, pin 4 left free and again pin 1 is unaltered.

Tuning on the short-wave bands is more critical than on the medium waves so tune very slowly over the band and remember signal strength is likely to vary from day to day.

Components List, Fig. 20a

C1	140pF variable capacitor Eddystone 586.
C2	1000pF mica capacitor.
1	Coil holder Eddystone 707.
L1	3m/c—7m/c coil Eddystone 706/R.
1	Engraved tuning dial.
1	6"x4"x2½" chassis.
2	Terminals (Aerial-Earth).
1	Twin socket strip ('phones).
X1	Germanium Crystal.

6

Fig. 21a-b is yet another design from America and appeared in Radio Craft some years ago, designed by Mr. W. J. Spain. The original used a silicon crystal which preceded the development of the germanium. Results are certainly very good though of course on the model built here, a germanium crystal was used. Selectivity can be adjusted to suit all conditions and even at short distances from local stations. Good volume can be obtained without the programmes overlapping. Home made coils are used and for best results a 75′ outdoor aerial is desirable.

Baseboard and panel construction is used, which can conveniently be housed in a small wooden cabinet.

Operation is a little more complicated than some of the other designs, but the results certainly merit the extra trouble taken. As with the other receivers much of the success obtainable is due to the coils and these must be carefully made otherwise results will be disappointing.

Layout and wiring can be followed quite easily from Fig. 21b and requires no special comment.

First make the coils. L1 is a tapped coil, the taps should be made in the same way as shown in Fig. 18b. It consists of 90 turns of 22 D.C.C. (double cotton covered) copper wire tapped at the following number of turns, 5, 10, 15, 25, 30, 40, 50, 60, 70 and 80. The former used has a diameter of 2". L2 is also wound on a 2" diameter former, this coil however is not tapped, it consists of 110 turns of 38 D.C.C. or enamelled copper wire. Great care must be taken when handling this wire, it is quite fine and in the hands of the inexperienced will easily break.

After wiring has been carried out the set can be put into operation.

Selectivity is controlled by S2, in position B tuning is broad but in position A it can be quite sharp.

When searching for a station switch S2 to B. Tune the station and switch S2 to A. The frequency range is controlled by the position of the tap on L1, a good plan is to start with the 5 turn tap on L1.

Components List, Fig. 21a

C1	500pF variable capacitor.
C2	500pF variable capacitor.
C3	1000pF Mica Capacitor.
X1	Germanium Crystal.
L1 & L2	See text.
S1	Single Pole toggle switch.
S2	Single Pole 2-way toggle switch.
2	Twin socket strips (Aerial/Earth and 'Phones).

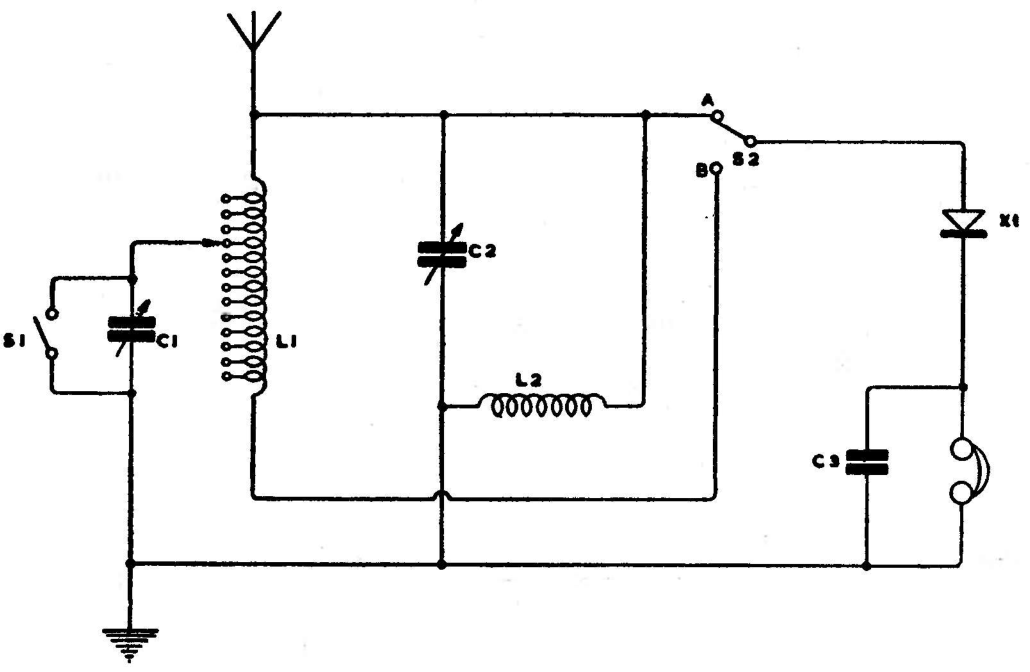

FIG. 21 a

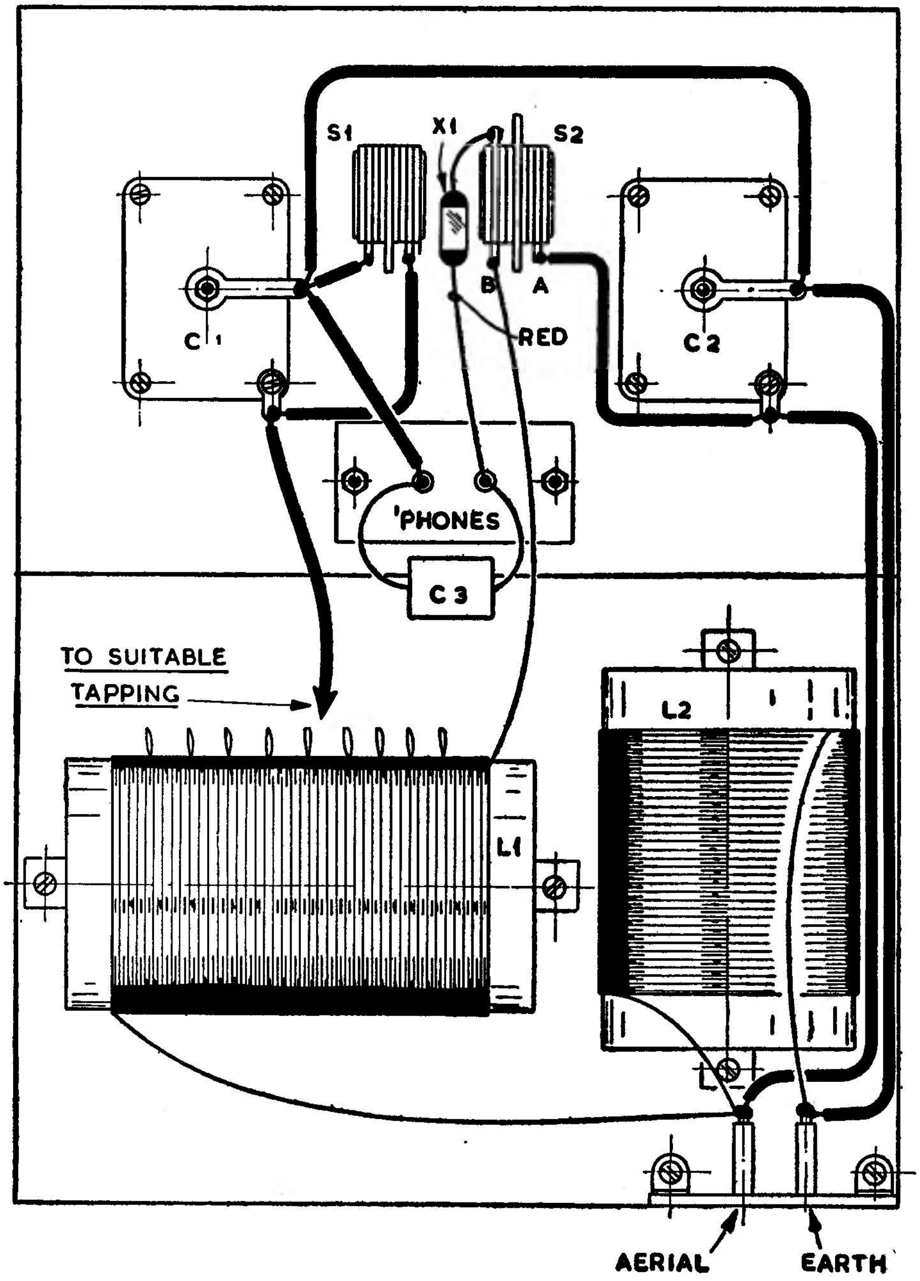
S1
X1
S2
B
A
RED
C 1
C 2
'PHONES
C 3
TO SUITABLE TAPPING
L2
L1
AERIAL
EARTH

FIG. 21 b

Keep S1 open and C1 for tuning, adjusting the tap on L1 for best results. Very powerful signals are best brought in with S1 closed. When using C1 and adjusting the taps on L1, switch S2 is kept in the A position.

It will be found that C2 acts as a fine control on C1. One division of the tuning dial on C1 is roughly equal to a movement of 20 divisions on C2.

Note that the coils are mounted at right angles to one another and that in this instance wood has been used as far as possible, an aluminium or other metal chassis is not suitable.

7

A very useful receiver for both medium and long-wave reception is possible with the circuit of Fig. 22a-b. A home-made coil is used. On the original a 1¼″ diameter former was used but in fact this is not very critical and a 1½″ former could be used. An interesting point is the construction of the long-wave coil section. Solenoid coils as used for the medium wave-band are not efficient if the length is too great relative to the diameter. Now as the long-wave section has 300 turns, if wound as a solenoid the winding length would be very great and a lot of efficiency would be lost. To overcome the difficulty the coil is pile wound. If the whole coil were to be wound in one pile it would still be inefficient because of the capacity formed in the winding, but by dividing it into five sections a coil of reasonable efficiency results. The same remarks concerning winding apply to this coil as to all the others in

Components List, Fig. 22a

C1	300pF Mica Capacitor.
C2	500pF Variable Capacitor.
C3	1000 Mica Capacitor.
X1	Germanium Crystal.
L1-2	See Text.
S1	Single Pole Toggle Switch.

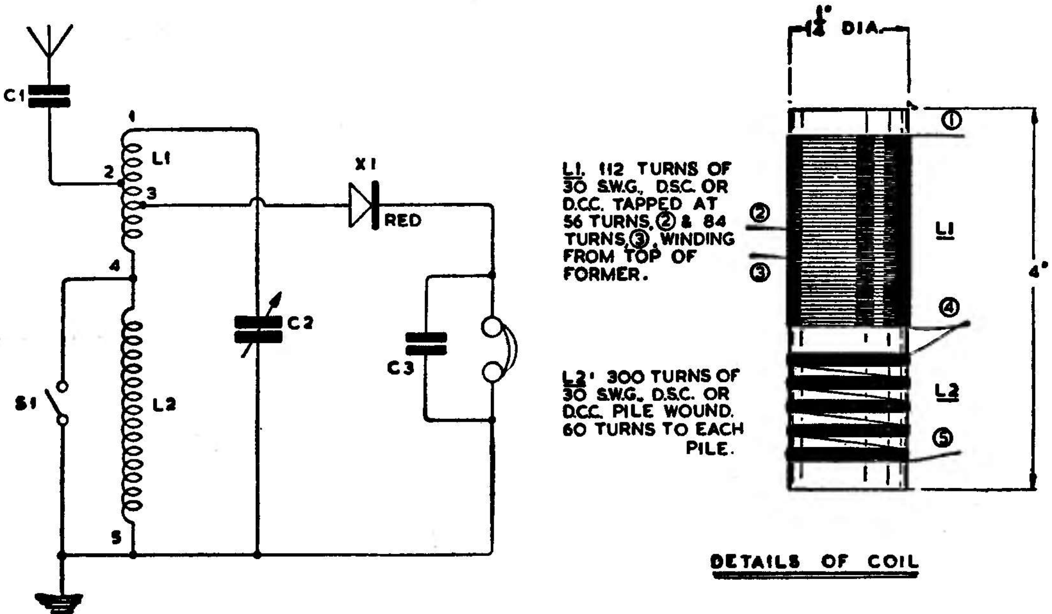
C1
1
2
L1
3
4
S1
L2
5
X1
RED
C2
C3
1¼" DIA.
①
②
③
④
⑤
L1. 112 TURNS OF 30 S.W.G., D.S.C. OR D.C.C. TAPPED AT 56 TURNS,② & 84 TURNS,③, WINDING FROM TOP OF FORMER.
L2. 300 TURNS OF 30 S.W.G. D.S.C. OR D.C.C. PILE WOUND. 60 TURNS TO EACH PILE.
L1
L2
4"
DETAILS OF COIL

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE OR WAXED CARDBOARD TUBE.

FIG. 22a

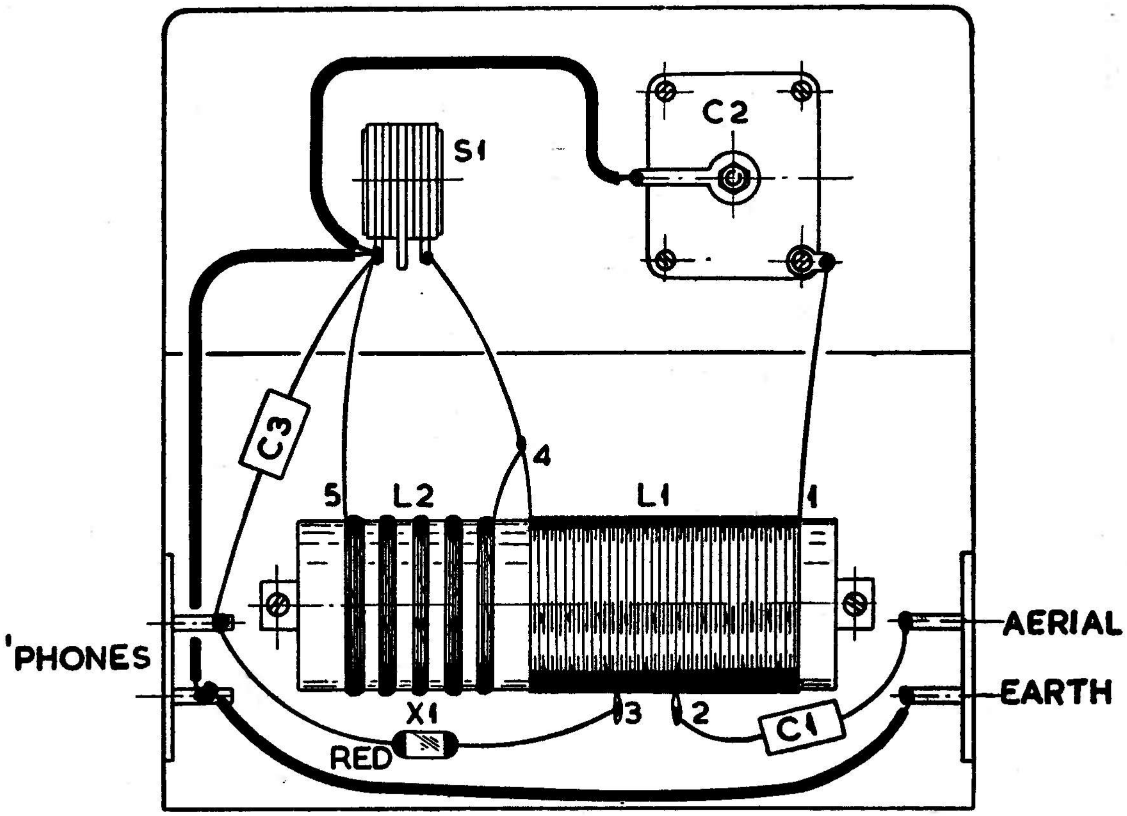

FIG. 22 b

this book, and readers should not, by this time, encounter any difficulties.

The design provides good selectivity with good sensitivity, and provided the set is not required to operate within a few miles of a powerful transmitter, no trouble will be encountered from interference between adjacent stations.

A wooden base-board and front panel is best for this design, and the wiring and layout can be clearly followed from Fig. 22b. Switch S1 is open for long-wave reception and closed for the medium-wave band.

Coil mounting can be on the same principle as for the full-wave design Fig. 19a-b.

8

If you look at all the other circuits shown in this book you will find that in each case the tuning capacitor and tuning coil are in parallel, this is by no means essential and the circuit of Fig. 23a shows a series arrangement that is the tuning capacitor is between the coil and earth instead of being wired up across it. This design also appeared in **Radio Craft** a few years ago.

An interesting point is that the coil is a conversion adapted from another type of component. To make this coil you must obtain an old I.F. transformer as used in superheterodyne receivers.

You will require one designed to work anywhere between 450 and 470 kc/s. Some of the

very early ones worked at 110 kcs/ but these are unsuitable.

The I.F. transformer must also be of the type tuned by an iron dust core at each end.

Dismantle the transformer, remove the parallel capacitors and mounting wires so that you are left with two coil bobbins and their former. Be very careful when un-soldering the wires from the supports because the coils are usually wound in litzendraht and must remain intact. Remove 25% (approximately) of the turns from one coil which we will call the secondary, and completely remove the iron dust slug from this coil, now cut the former in half, and mount as in Fig. 23a. Warm the coils with heat from your soldering iron or from a fire and gently slide the coils into position at the ends of the former as in Fig. 23a. The wax will set again holding the coils in their new position.

Note that the wires from the coils are connected to tags on the mountings. These are pieces of bakelite, and usually the original end pieces used on the I.F. transformer can be used.

A word about soldering litzendraht. Do not attempt to use emery paper. Prepare the end of wire by removing the outer silk covering and dip the end in methylated spirits. Ignite with a match and after about five seconds wipe with a cloth held between finger and thumb. This will quench the flame and will remove all traces of enamel from the wire. It is a good plan to experiment with the odd length of wire removed from the secondary coil. If the coils prove to be constructed of ordinary copper wire, it can of course be prepared with fine emery cloth in the normal manner.

Components List, Fig. 23a

C1	500pF variable capacitor.
X1	Germanium Crystal.
L1	See Text.
4	Plugs and sockets or terminals.

Wiring is straight forward as can be seen from Fig. 23b and requires no explanation.

Operation is as follows, selectivity is controlled by varying the distance between coils with the wing nut, and to an extent by the position of the primary iron dust core. Remember that the aerial and earth installations are as important with this set as with any other crystal receiver. No parallel capacitance appears necessary across the 'phones, the set works quite happily relying on the self-capacitance of the head-phones.

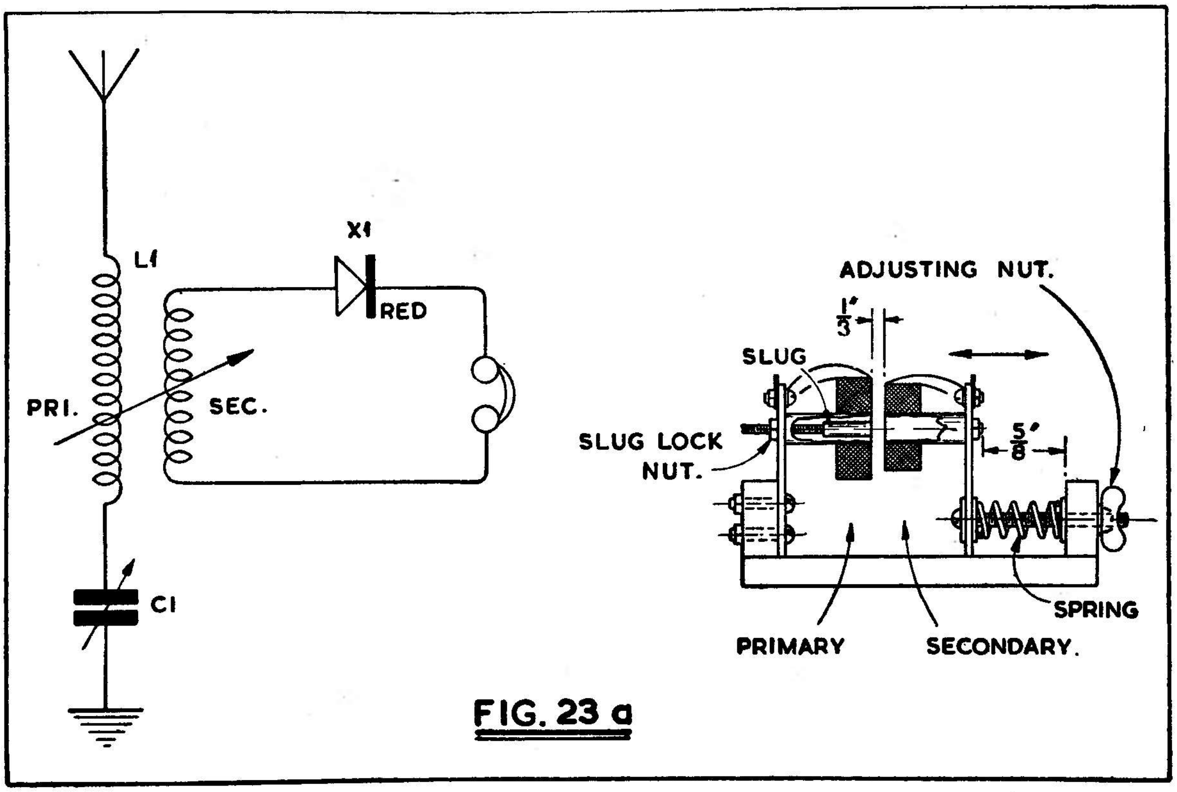

FIG. 23 a

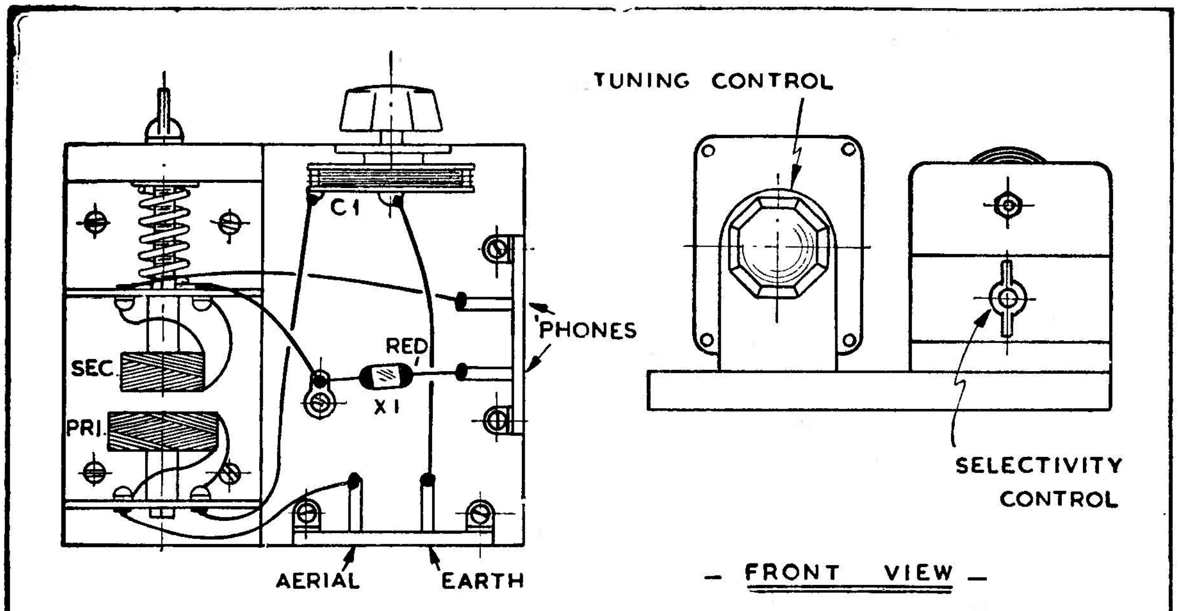
TUNING CONTROL
C1
'PHONES
RED
SEC.
X1
PRI.
SELECTIVITY CONTROL
AERIAL
EARTH
— FRONT VIEW —

FIG. 23 b

9

So as to give a truly varied selection of circuits, some of the sets have been designed around commercial coils whilst others use coils which must be home-constructed.

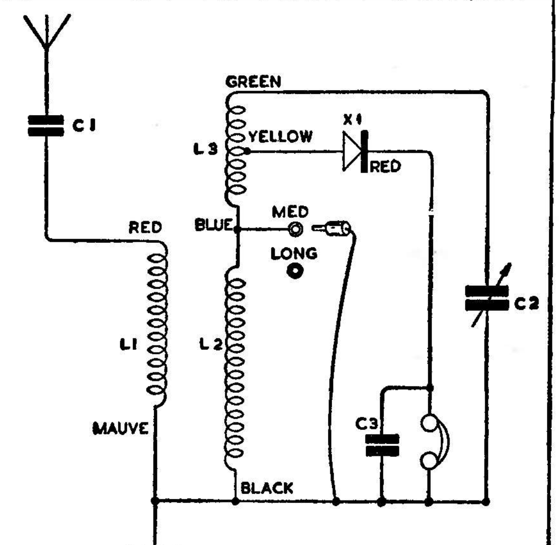
C1
GREEN
X1
L3
YELLOW
RED
RED
BLUE
MED
LONG
C2
L1
L2
MAUVE
C3
BLACK

FIG. 24 a

Most enthusiasts like to build their own coils but when it is a question of something really small combined with efficiency a commercial coil is the obvious choice.

Fig. 24a-b uses a tobacco tin or a similar container just as the design shown in Fig. 16a.

This receiver however is a little more selective and is more suitable than the earlier design, if you are near a transmitter or have a very long aerial. Take great care to connect the coil according to the colours shown on the diagrams. Sockets are used for wave-changing, this cheapens the cost a little and they are just as efficient.

Components List, Fig. 24a

C1	100pF Mica capacitor.
C2	500pF variable capacitor solid dielectric.
C3	1000pF Mica capacitor.
X1	Germanium crystal.
L1-2	R.E.P. dual range coil. Blue box.
6	Insulated wander-plug sockets.
5	Wander Plugs.
2oz.	Tobacco tin or similar container.

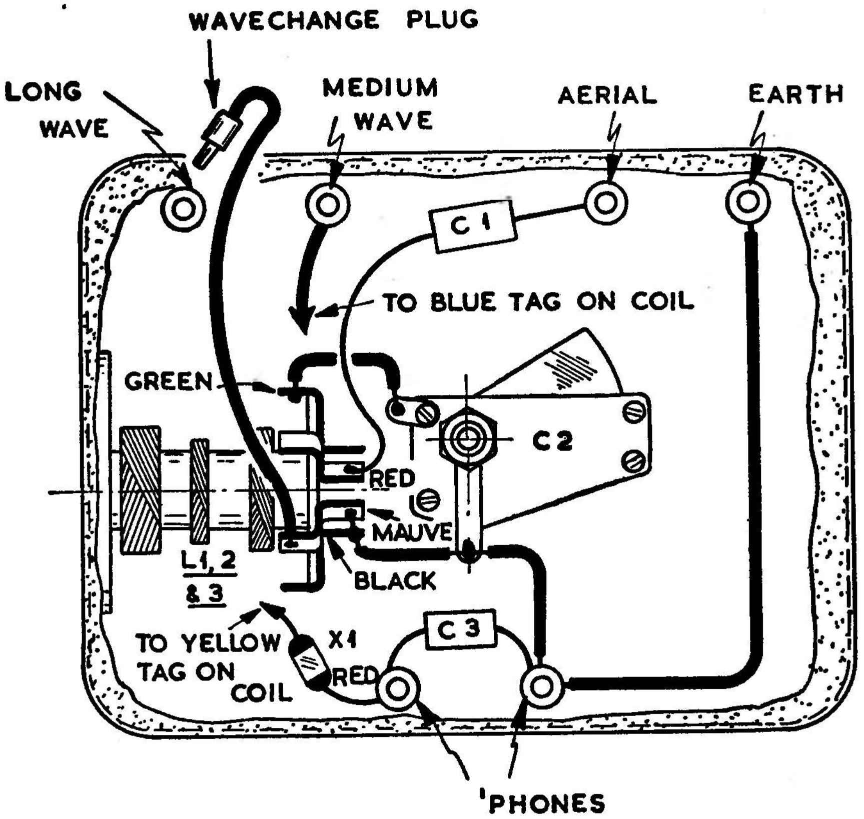

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG. 24 b

10

In certain cases, signals from the local transmitter are too powerful to permit a crystal set with a normal coil to be of much use. I myself live in Hertfordshire a few miles from the London transmitters, and find a more elaborate circuit is necessary. Now in radio circles it is a well-known fact that the greater the number of tuned circuits, the greater the overall selectivity. Look at Fig. 25a, you will see that the usual L and C tuning arrangements have been duplicated. This type of circuit is known as a Band-Pass Filter. There are many kinds of such filters, this particular type is an inductively-coupled filter, note that energy from the first half (L2) is conveyed to L3 by way of two small inductors L1 and L4.

Tuning is quite sharp and the "feel" of the set is quite different from that of single coil designs. A small chassis is ideal for construction. 6"x4"x2½" was used on the original, though it could have been smaller.

Fig. 25b gives the layout and wiring.

Take particular notice that L1-2-3 is mounted horizontally and at right angles to L4-5-6, this is to prevent the coupling from being too "tight" which would ruin the idea of the filter. A good aerial and earth is of course necessary. It may be found on test that the selectivity is too great

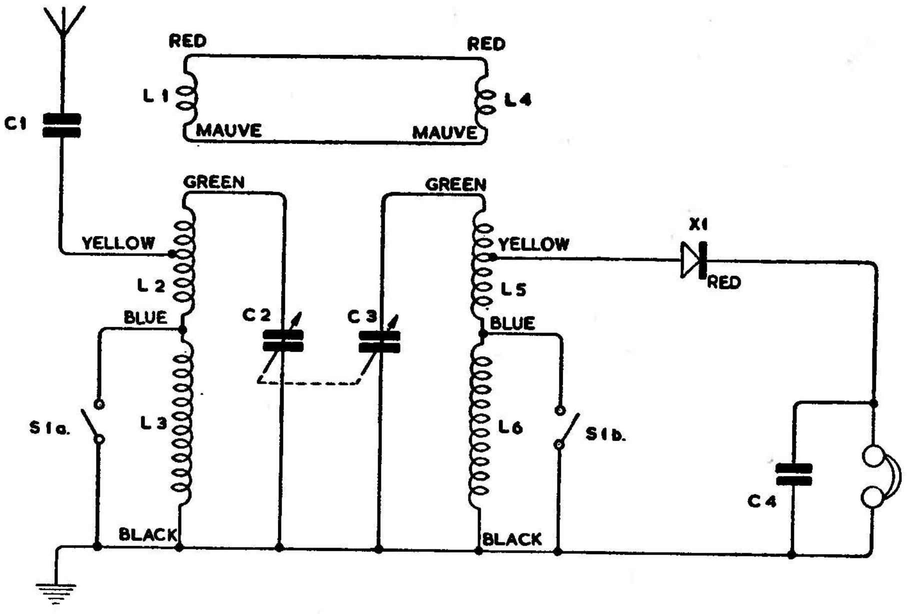

FIG. 25 a.

and that volume has suffered. This can be ingeniously overcome by connecting a very small capacitance between the two Green tags on the coils. The simplest way is to connect it across the fixed vane connections on C2-C3. Values for this extra capacitance are a matter for experiment usually a value between 10pF and 47pF is sufficient. A small 50pF trimmer could be used and adjusted to suit. The ideal setting is to enable powerful adjacent stations to be received just short of overlap so that as much volume as possible is obtained.

There is one point to be observed with the ganged capacitor. It should be of the type fitted with trimmers though of course, there is no reason why you should not fit them yourself. These are to allow for differences between coils to be balanced out, and consist of a small variable capacitance across each main section.

When setting up the receiver, first set each trimmer at half way and tune in a station near the high-frequency end of the band (vanes nearly out) adjust the trimmers for maximum volume. If extra coupling capacity is added as described, try adjusting them further.

Once properly set up, no further adjustment is necessary.

Components List, Fig. 25a

C1	100pF Mica capacitor.
C2-3	2 x 500pF variable capacitor (see text).
C4	1000pF Mica capacitor.
L1-2-3	Dual Range Coil R.E.P. (Blue Box).
L4-5-6	Dual Range Coil R.E.P. (Blue Box).
X1	Germanium Crystal.
S1a-b	2 pole single throw toggle switch.
4	Terminals or plugs and sockets.

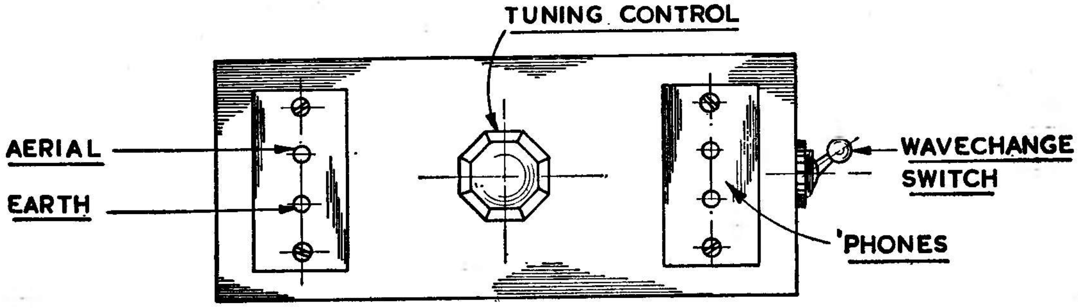
TUNING CONTROL
AERIAL
EARTH
WAVECHANGE SWITCH
'PHONES

FRONT OF COMPLETED RECEIVER.

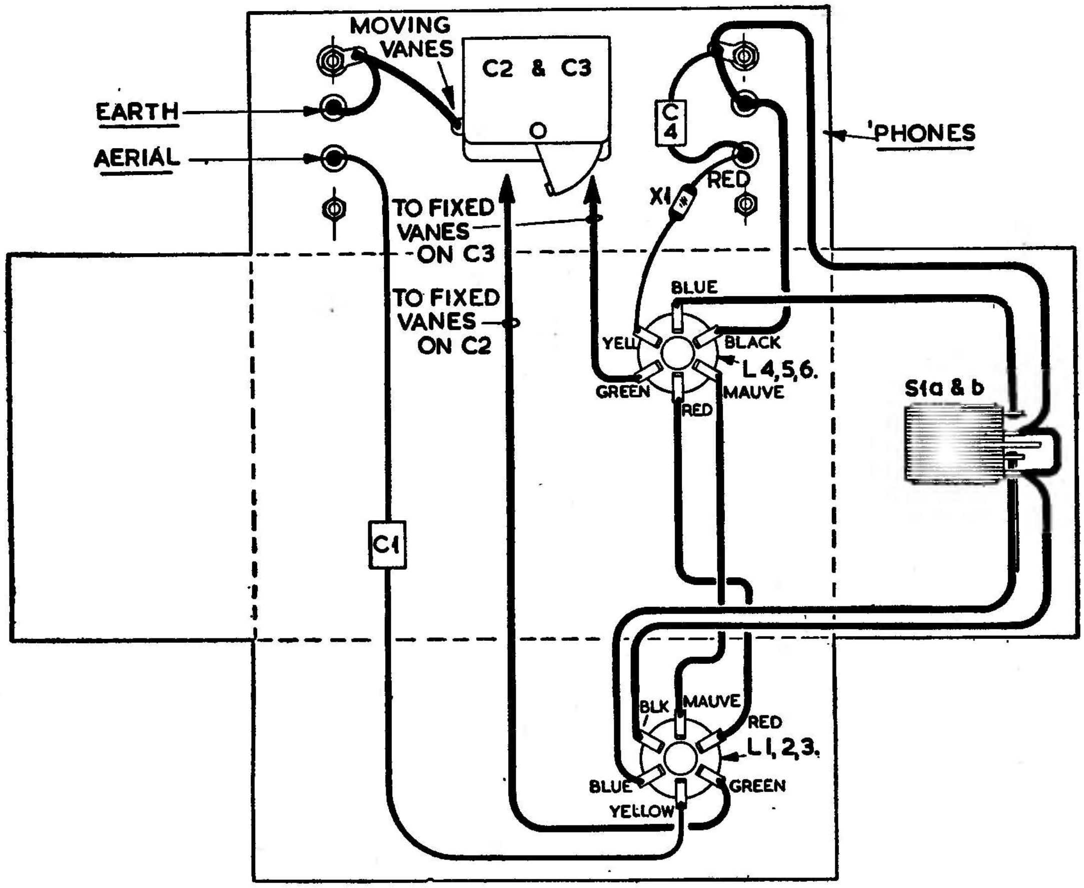
MOVING VANES
C2 & C3
EARTH
AERIAL
C4
'PHONES
RED
X1
TO FIXED VANES ON C3
TO FIXED VANES ON C2
BLUE
YELL
BLACK
L 4,5,6.
GREEN
RED
MAUVE
S1a & b
C1
BLK
MAUVE
RED
L1,2,3.
BLUE
YELLOW
GREEN

EXPLODED UNDER-CHASSIS VIEW

FIG. 25 b

11

Another band-pass circuit is shown in Fig. 26a. This time home-made coils are used. This filter is not inductively coupled but entirely capacitive, relying on C3. As in the previous circuit it must be adjusted to give just sufficient " broadness " to the tuning to provide adequate volume without station overlap. To avoid unwanted coupling a screen is mounted between the coils, this is necessary because due to the size of the coils they would be bound to couple quite tightly without it. Such precautions were not necessary with the circuit of Fig. 25a, as the commercial coils are much smaller physically and the risk of unwanted coupling reduced.

The coils for this receiver are for medium wave reception only, this keeps the size down. Coil winding procedure has already been fully explained, and covers all the requirements of these coils. Base-board and panel construction is used, and for more accurate tuning since the coils are home-made, separate tuning capacitors are recommended. Fig. 26b gives the layout and practical wiring.

Components List, Fig. 26a

C1	100pF Mica capacitor.
C2	500pF variable capacitor.
C3	50pF Mica trimmer.
C4	500pF variable capacitor.
C5	1000pF mica capacitor.
L1-2	See text.
X1	Germanium Crystal.
4	Terminals or plugs and sockets.

For those who do not wish to bother with these a ganged capacitor can be used, but tuning may not be as good as with separate ones. The screen mounted between the coils is of aluminium or copper and should be about 1″ higher than the coils and the same length as the baseboard depth.

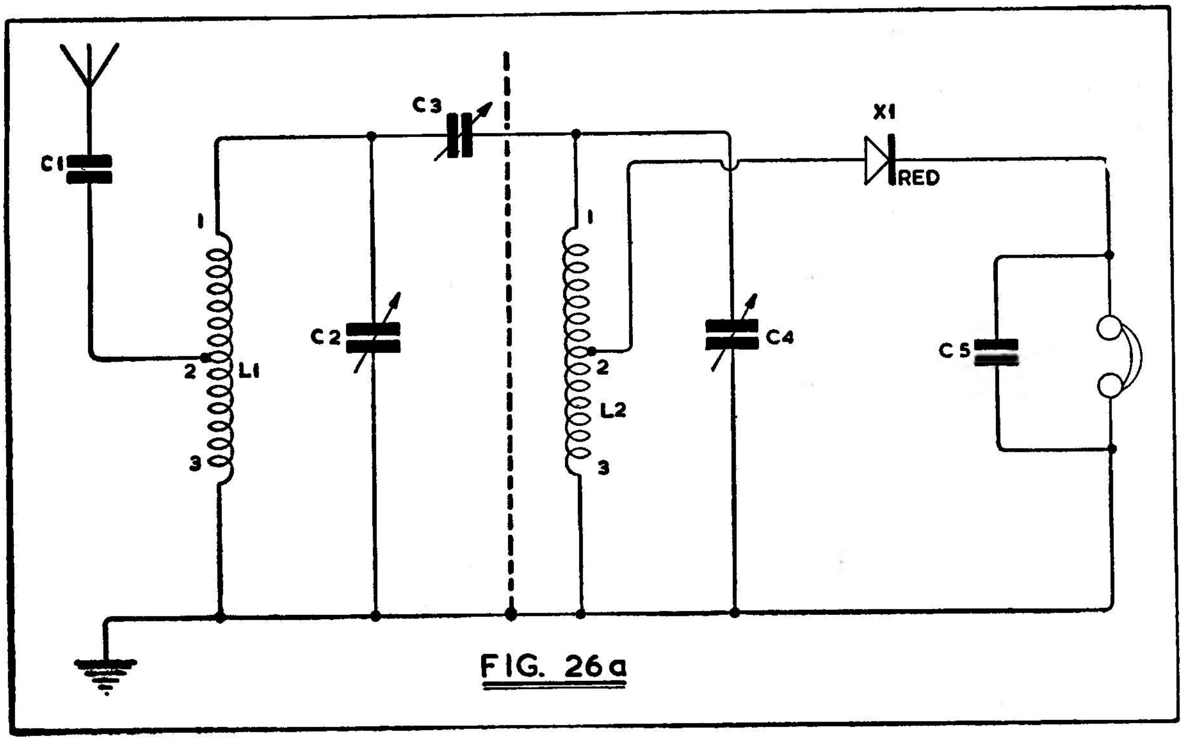

FIG. 26a

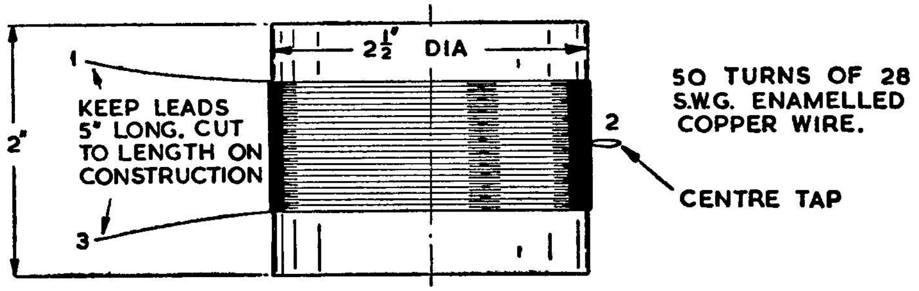
1
2½" DIA
2"
KEEP LEADS
5" LONG. CUT
TO LENGTH ON
CONSTRUCTION
3
2
50 TURNS OF 28
S.W.G. ENAMELLED
COPPER WIRE.
CENTRE TAP

COIL DETAILS.

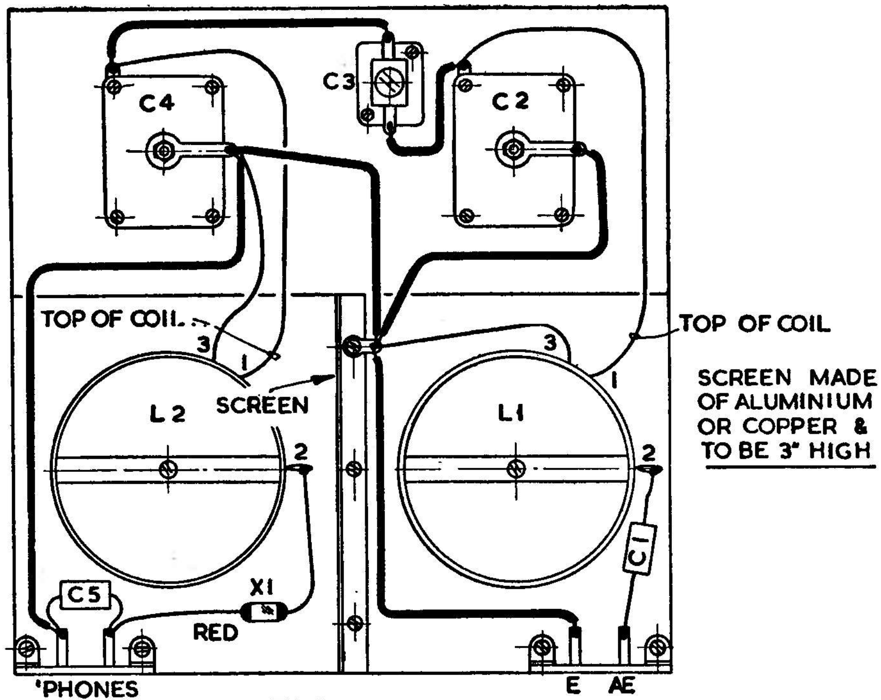
C4
C3
C2
TOP OF COIL
3
1
L 2
SCREEN
2
TOP OF COIL
3
1
L1
2
SCREEN MADE
OF ALUMINIUM
OR COPPER &
TO BE 3" HIGH
C1
C5
X1
RED
'PHONES
E
AE

FIG. 26 b

12

Fig. 27a-b uses a band-pass filter for tuning, and the coils are home-constructed.

With this design a ganged capacitor is quite suitable as the coils tune quite accurately. It is essential that the coupling condenser C3 is **non-inductive,** there will be no difficulty in obtaining a new component of this pattern but it is as well to avoid old components which may not employ this form of construction.

Receivers using this type of filter are used to the best advantage when situated a short distance from a powerful transmitter and if selectivity is not important this particular design is not the most suitable.

Both medium-waves and long-waves are catered for: consider L1-2-3, this is the first half of the filter.

L1 is the aerial coupling coil, with S1 closed,

the medium-waves are received. This coil is inductively coupled to L2 and L3. Again with S2 open L2 and L3 combined cover the long-waves and when S2 is closed the medium waves are covered.

Tuning for this half of the filter is by means of the capacitor C1, which is one half of the ganged capacitor. A screen is erected between the two sets of coils comprising the filter, and coupling is effected by C3, an earth return for the coils is provided by R1.

L4 and L5 operated in a similar manner to L2 and L3, tuning is by means of C2, the remaining section of the ganged capacitor. You will realise that the coils must be accurately wound otherwise tuning will not remain constant between the two sections over the band.

The ganged capacitor should be fitted with trimmer capacitors when purchased so that the two circuits can be balanced.

You will have noticed that this circuit uses three sets of switch contacts. A three-pole two-way wafer switch is the best choice otherwise either two or even three separate switches are necessary. All the necessary coil details are given in the diagrams, the tap on L4 is made in the same way as shown for other coils in the manual. One other component used only on this design, is the resistor R1, its size is not important because the current flowing is so small it can be ignored.

A ¼ watt rating is adequate, its precise value is not critical so that a 20% tolerance component is quite suitable.

Resistors are coded by colours and the 1000Ω type you require will bear the following sequence. At one end will be painted three coloured rings.

Components List, Fig. 27a

C1-2	500pF 2 gang, variable capacitor.
C3	0.05mfd. non-inductive paper capacitor.
C4	1000pF mica capacitor.
R1	1000 ohm resistor.
X1	Germanium Crystal.
S1-2-3	3-pole 2-way wafer switch.
L1-2-3 & L4-5	See text.
4	Terminals (Aerial, Earth and 'Phones). Sheet copper or aluminium for screen.

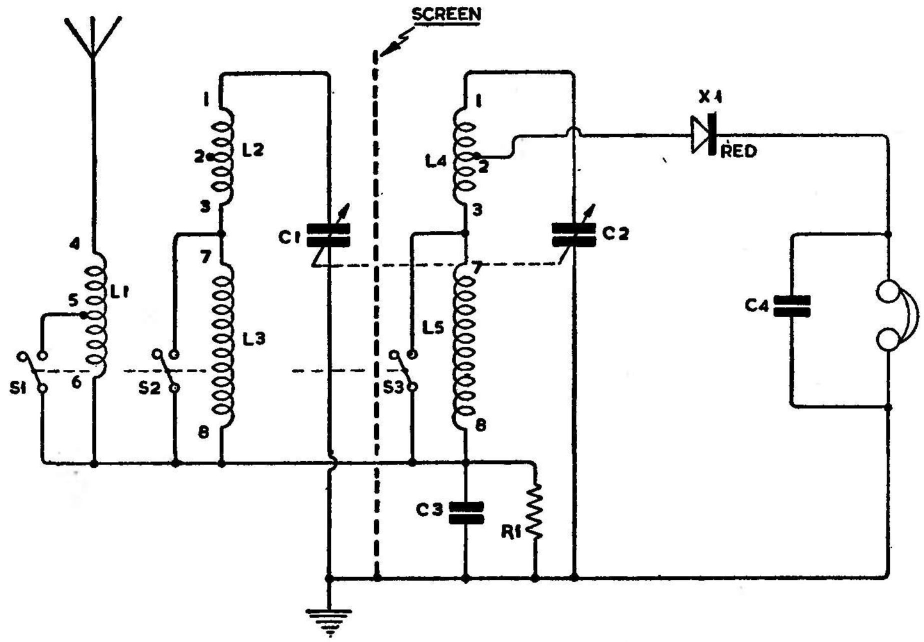

FIG. 27a

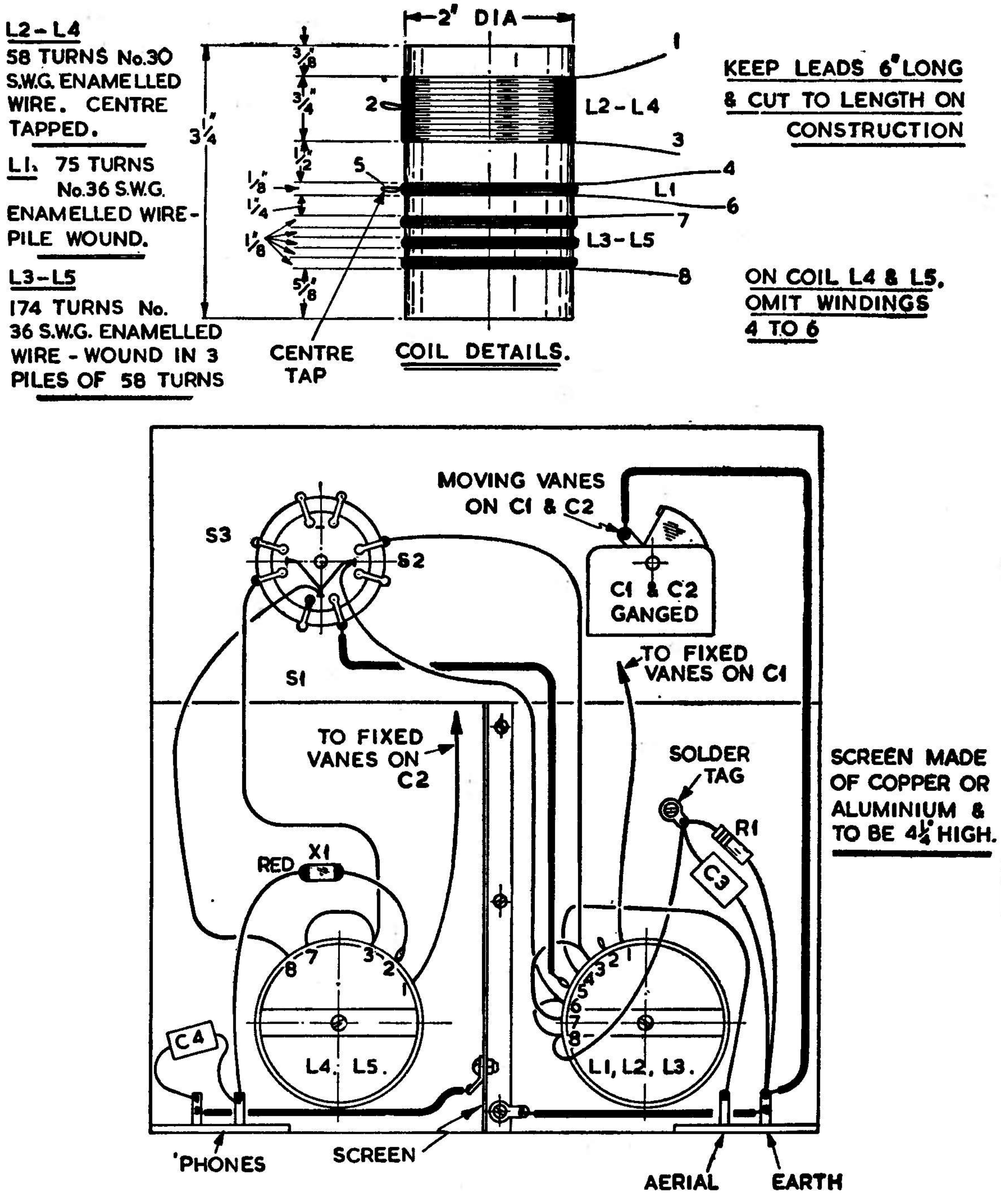
L2-L4
58 TURNS No.30 S.W.G. ENAMELLED WIRE. CENTRE TAPPED.
L1. 75 TURNS No.36 S.W.G. ENAMELLED WIRE-PILE WOUND.
L3-L5
174 TURNS No. 36 S.W.G. ENAMELLED WIRE - WOUND IN 3 PILES OF 58 TURNS
2" DIA
3¼"
⅜
¾"
½"
⅛"
¼"
⅛"
⅝"
1
2
3
4
5
6
7
8
L2-L4
L1
L3-L5
CENTRE TAP
COIL DETAILS.
KEEP LEADS 6" LONG & CUT TO LENGTH ON CONSTRUCTION
ON COIL L4 & L5, OMIT WINDINGS 4 TO 6
MOVING VANES ON C1 & C2
S3
S2
S1
C1 & C2 GANGED
TO FIXED VANES ON C1
TO FIXED VANES ON C2
SOLDER TAG
R1
C3
SCREEN MADE OF COPPER OR ALUMINIUM & TO BE 4½" HIGH.
RED
X1
C4
L4, L5.
L1, L2, L3.
'PHONES
SCREEN
AERIAL
EARTH

FIG. 27 b

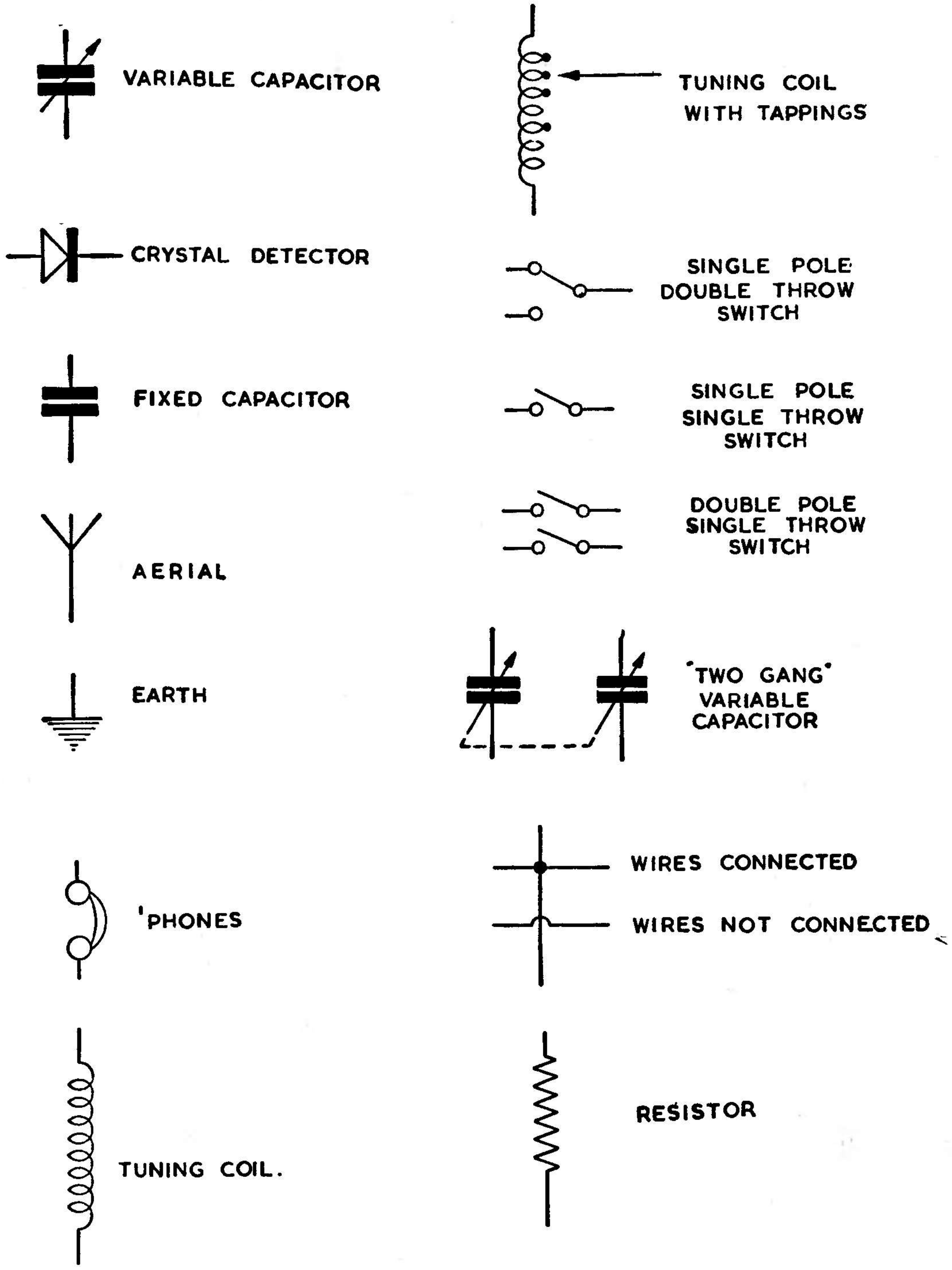
VARIABLE CAPACITOR
TUNING COIL WITH TAPPINGS
CRYSTAL DETECTOR
SINGLE POLE DOUBLE THROW SWITCH
FIXED CAPACITOR
SINGLE POLE SINGLE THROW SWITCH
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
EARTH
'TWO GANG' VARIABLE CAPACITOR
WIRES CONNECTED
WIRES NOT CONNECTED
'PHONES
RESISTOR
TUNING COIL.

SYMBOLS USED IN THE CIRCUITS DESCRIBED.

FIG. 28

The first one is BROWN which represents 1, the first figure of the value, the second BLACK, this indicates the second numeral is nought, and the third colour is RED which tells us that two further ciphers or noughts are added to give the total value. From this you can see, we have 1 plus a nought equalling 10, plus two further noughts which total 1000. Some earlier types of resistor were painted differently, the whole body colour represented the first figure, one tip was coloured to represent the second figure, and a painted dot in the centre gave the number of ciphers to be added.

There are no other special points regarding construction except perhaps that if a long aerial is used, a small 100pF capacitor might be tried in the aerial lead to prevent damping the first tuned circuit.

When setting up the receiver, tune to a station at the high-frequency end of the M.W. band and set the trimmers for loudest volume, no other adjustments are necessary.

CONCLUSION

The twelve receivers described are representative of the best crystal set designs available to-day. Their construction will provide many hours of useful enjoyment, and the results will give lasting pleasure.

A final word about components, if no actual maker is specified, any good class component can be used. Switches, fixed and variable capacitors and crystals, are available from numerous manufacturers all of which are invariably of excellent quality. Where a particular manufacturer's product is called for, the specification should be adhered to. The commercial coils specified are freely available from most supply houses specialising in components for constructors, but in case of difficulty write to Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.6, who will, on receipt of a S.A.E. be pleased to supply the address of your nearest stockist. Coil formers are not always easy to obtain and in this case advice should be sought from Post Radio Supplies, 395, Queensbridge Road, London, E.8, who will be pleased to help in any way possible.

* *

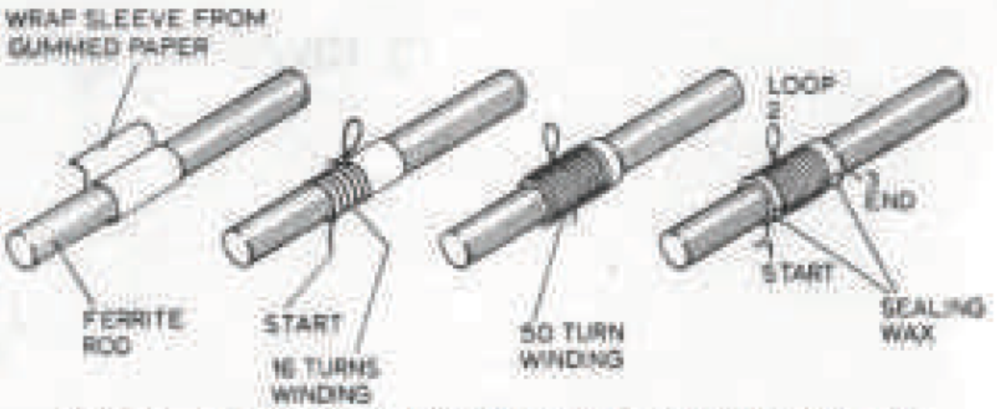

Fig. 3.1 Aerial tuning coil wound from 38-s.w.g. wire on a $\frac{3}{8}$-in. diameter ferrite rod.

Fig. 3.1 details the construction of a suitable coil for a tuned circuit, matched to a ferrite rod of $\frac{3}{8}$ in. diameter and about 4 in. long (see Chapter 4 for alternative coil windings on different rod sizes). If a rod is purchased longer than 4 in., it can be used as it is, or cut down by marking around with a file and then breaking off the surplus length.

Cut seven 1-in. lengths of gumstrip. Moisten one and wrap around the rod gummed side up. Now add about another half a dozen wrappings of similar length over the first, this time with gummed side *down* to form a reasonably rigid tube. Make sure that the paper tube is a *sliding* fit on the ferrite rod and leave to dry thoroughly (preferably removed from the rod so that it cannot become stuck to it).

When the paper tube is quite dry it should be rigid, when the coil windings can be applied. The wire to be used is 38-s.w.g. enamelled copper wire, the number 38 referring to the actual diameter of the wire according to the standard wire gauge (s.w.g.).*

Starting about $\frac{1}{4}$ in. in from one end of the paper tube, wind the wire carefully round the tube, with each turn tight against the one before it, until sixteen full turns have been completed. Then make a loop in the wire, as shown, and carry on winding, with succeeding turns touching, until fifty turns in all have been completed. The two loose ends of the coil (the start and finish) can be secured with a dab of sealing wax whilst the projecting loop can be twisted together (e.g. by putting a pencil through the loop and twisting up). Cut off the loop, leaving about $\frac{1}{2}$ in. protruding from the main coil, bare the wire ends and solder together. This forms point 2 on the coil; the start is point 1, and the end point 3—see Fig. 3.1. It will be easy to remember these without marking since the loop or tapping point (2) comes much closer one end (1) than the other (3).

---

* American wire gauge (AWG) is about two sizes smaller than the British standard wire gauge (s.w.g.). For example, 38 s.w.g. is approximately the same as 36 AWG

Cut a panel of *Paxolin sheet* to about the size shown in Fig. 3.2, using a hacksaw.* On this secure a *tag strip*, as shown, and drill a hole to mount a miniature or small-size 500 pF *variable capacitor*.

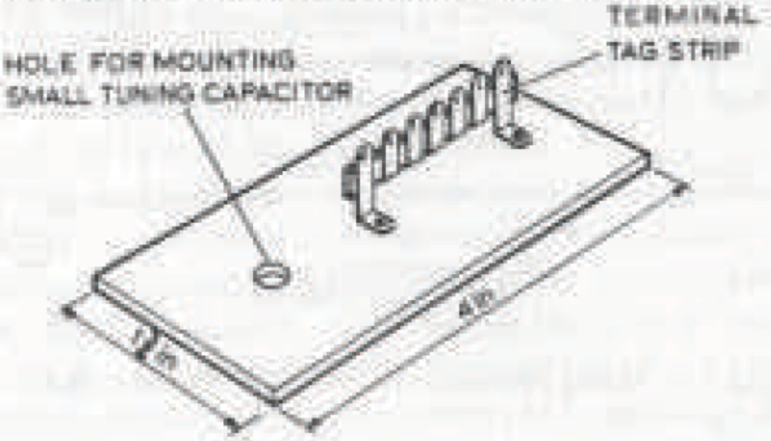

Fig. 3.2 Paxolin panel and tag strip for crystal set circuits.

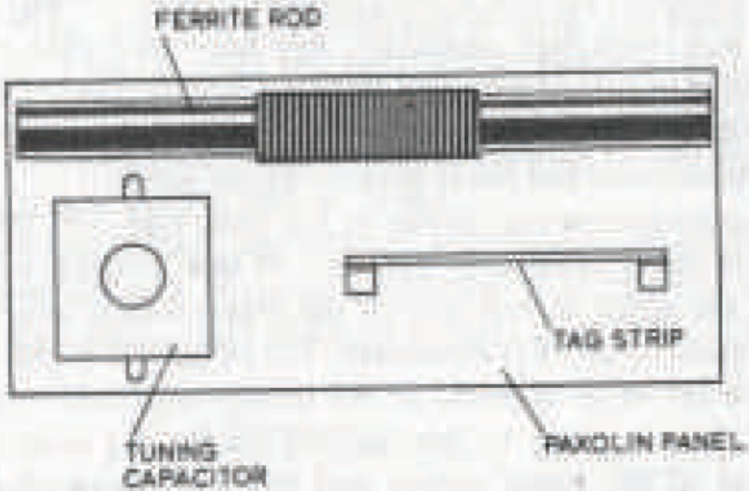

Fig. 3.3 Layout of components on Paxolin panel.

The aerial coil is then mounted on the panel as shown in Fig. 3.3, gluing the coil on to the Paxolin with two or three dabs of sealing wax, or some other suitable adhesive. *Note:* the ferrite rod must be free to slide in the paper tube for 'tuning' adjustments.

Virtually any miniature germanium or silicon diode will be suitable for the detector. Recommended types, which are readily available, are 1N34 and 1N914.

Earphones must be of *high-impedance* type, which need not be expensive to buy, and the higher the impedance the better the reception.

---

* A phenolic or fiberglass sheet may also be used.

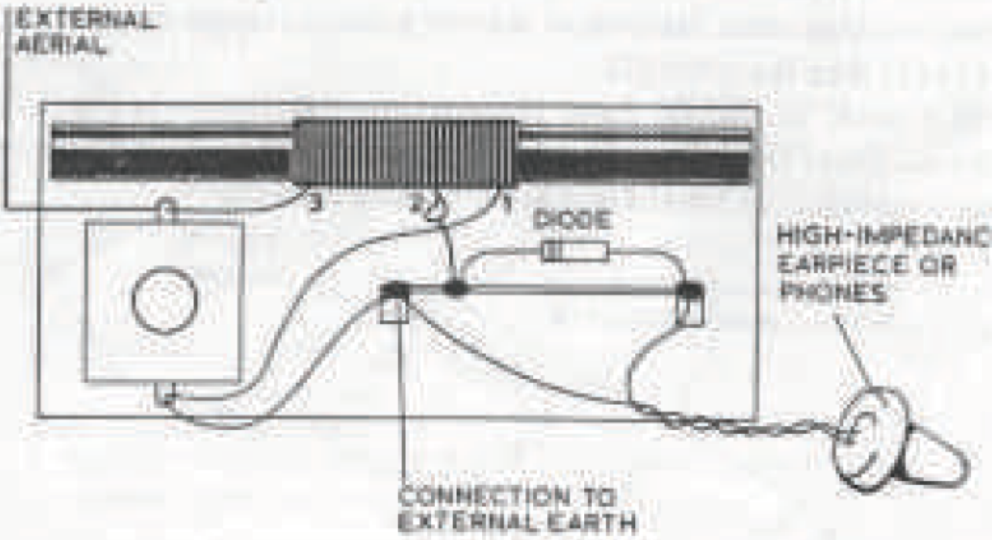
EXTERNAL AERIAL
3
2
1
DIODE
HIGH-IMPEDANCE EARPIECE OR PHONES
CONNECTION TO EXTERNAL EARTH

Fig. 3.4 Wiring connections to complete the crystal set.

Alternatively a deaf-aid type earpiece can be used; this will not give the same volume or quality of reproduction as headphones, but is a less expensive component. This should preferably be of the *high-impedance magnetic* type, with high sensitivity. Any high-impedance earpiece will suffice, but if of crystal type will require a resistor, connected across it, to complete the circuit. This will reduce the amount of current flowing through the earpiece and lower the strength of signal.

Wiring connections are shown in Fig. 3.4. End 3 of the aerial coil (the end of the 50-turn coil) connects to one terminal of the tuning capacitor, the aerial or 'hot' end of the tuned circuit, and the point to which an external aerial is connected. The other end of the coil (end 1) connects to the other terminal of the tuning capacitor, from which an additional wire is taken to the first tag on the tag strip. This is the 'earthy' end of the tuned circuit, and the point to which an external earth is connected. Leave plenty of slack wire between the coil and tuning capacitor.

The other connections are then as follows:

(i) Tapping point of the coil wire bared and connected to the second tag.

(ii) The diode also soldered to this same tag, and to any other free tag.

(iii) Headphone (or earpiece) leads to the 'earthing' tag, and to the 'free' tag to which the above diode has been connected. All connections should be made with *soldered* joints.

The wiring-up can be checked against the *circuit diagram* shown in Fig. 3.5 (ignoring the components shown with broken lines). The set should now be 'working'.

In areas of strong signal strength, no external aerial or earth connections should be necessary. Performance will, however, be improved in any area by attaching an aerial wire (which can be any thin wire, e.g. using the same wire as for the coil winding), of up to 160 feet. The longer the aerial the better the reception, provided it is led away from the receiver to as *high* a point as possible.

An earth connection may further improve aerial performance; by this we mean a connection to some conductor positively in contact with the ground (preferably buried), an excellent example being a metal water-pipe. Thus, if an earth connection is found to be necessary (or you want to try one to see how performance is affected), connect a wire from the 'earthing' tag on the receiver to a convenient water-pipe.

This question of obtaining a *good* aerial and earth is a most important one in areas of poor signal strength. Linking up to a television aerial is often a good plan, since TV aerials are also usually mounted as high as possible. If bare wire is used, it is also important that the upper (free) end of the aerial is not made fast to something which could produce an earth connection (e.g. a damp tree), or at least is suitably insulated from such a support. String is not an efficient insulator; that, too, can conduct when wet.

Quite good results are often obtained by using the springs of a bed as an aerial, in which case an earth connection is usually not necessary. Sometimes, too, when other attempts to yield a good signal strength in

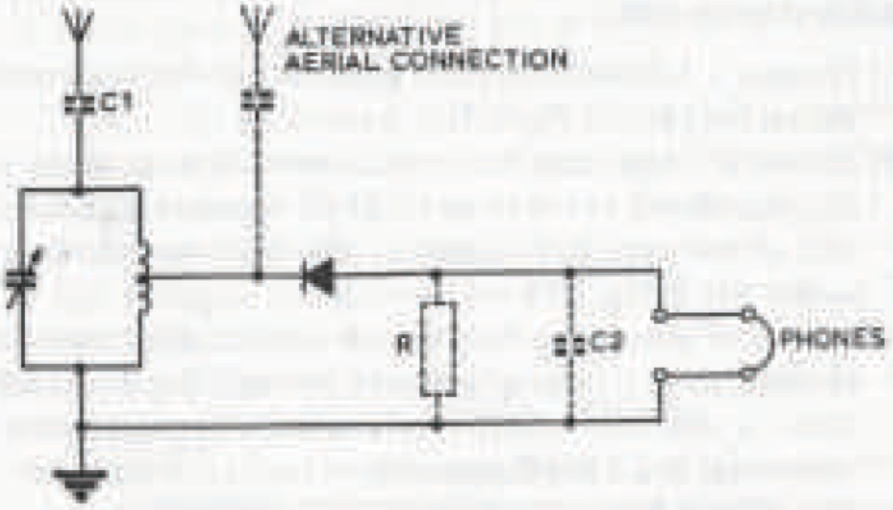

Fig. 3.5 *Circuit I.* Circuit diagram of basic crystal set.

the aerial have failed, connecting the *aerial* side of the tuning coil to a *good earth* (a water-pipe) can produce better results, the normal earth connection being left off.

Tuning

The receiver is adjusted as follows. Turn the tuning capacitor to fully close the vanes, then open about half a turn on the spindle or knob. (If you are using a trimmer as a tuning capacitor, screw right down and then open half a turn.)

The tuning coil should now be slid up and down the ferrite rod (the coil leads were left a fairly long time to give the necessary freedom of movement) until BBC Radio 3 is heard.*

It may be necessary to slightly alter the adjustment of the variable condenser to tune in to this programme. Also, because of the ferrite-rod aerial the set will be directional, that is, the signal strength received will depend to some extent on the direction in which the aerial rod is pointing, so position the set to pick up the maximum volume.

Having established the best position of the tuning coil on the rod to receive BBC Radio 3, fix permanently with a dab of sealing wax. You should then find it possible to tune in to further stations by altering the setting of the variable capacitor—e.g. typically Radio 4 in about the middle of the capacitor travel and Radio 1 towards the other end.

Any reception you get will almost certainly be very weak and (unless you live close to a broadcast station) you can really feel satisfied if you get any station at all at audible strength. But it is surprising how, sometimes, even quite distant stations can be heard. Also you can often improve the reception and listening strength by quite simple modifications. Try these in order:

(i) Connect a 1,000-pF capacitor across the headphone (earpiece) connections (C2 in Fig. 3.5).

(ii) Instead of connecting the external aerial directly to the tuned circuit, connect one lead to a 220-pF capacitor, and the other end of the capacitor circuit to the 'hot' end of the tuned circuit (C1 in Fig. 3.5).

(iii) Instead of connecting the external aerial to the tuned circuit, connect to the tapping point of the coil (tag to which the diode is also connected). Try a direct connection, and also connecting in a 220-pF capacitor.

---

* Tune to an AM broadcast station near the lower end of the dial—close to 550 kHz.

(iv) Try connecting a 1.2-k resistor (or higher value) across the phone connection (R in Fig. 3.5). You may be using the wrong type of phones or earpiece, which do not provide a proper load or complete the circuit.

If there is a complete lack of response, check for faulty wiring-up. A more likely cause, however, is lack of an external aerial or earth connection in an area where these are strictly necessary for adequate reception; or an inefficient aerial (too short) or poor earth connection (bad electrical contact to a good earth point, or connection to a bad earth point).

Another possible cause of apparent failure may be too much outside noise entering the ear so that it is impossible to detect the very weak radio signal as it is being tuned in. Headphones are better than a single deaf-aid type of earpiece in this respect but, in any case, a really quiet room is virtually essential for initial setting up and tuning adjustments. Also, if your adjustment of the tuning control is too coarse, you may completely miss the setting for the station you are looking for, without realizing it.

Reception will also tend to vary with weather conditions. Some days it may be so poor that what was normally a strong station is hardly heard at all. The simple basic receiver has many limitations but, since it costs very little to construct and nothing at all to operate, this must be regarded as inevitable.

Providing you can hear something–even if too weak a signal to distinguish properly–you can certainly improve the performance of your basic set by further experimentation with tuned circuits (see Chapter 4) and/or the addition of amplification to the circuit. You can also try other types of basic crystal set, as described in the following projects.

Circuit 2 (Fig. 3.6) is identical to *Circuit 1* except that, instead of a diode, a transistor is used as a detector. Only two of the transistor leads are connected–the emitter (*e*) connection to the tapping part of the coil, and the base (*b*) to the 'earthy' end of the circuit. The collector lead of the transistor is ignored (bend it out of the way so that it cannot accidentally short out the other leads).

You can try almost any type of low-cost AF transistor; recommended types are OC42 2N370, 2N2925, 2N5088.

With the addition of two more components, Circuit 2 can be modified to work the transistor both as a detector and an amplifier, to give stronger signals through the headphones. Using the same transistor type

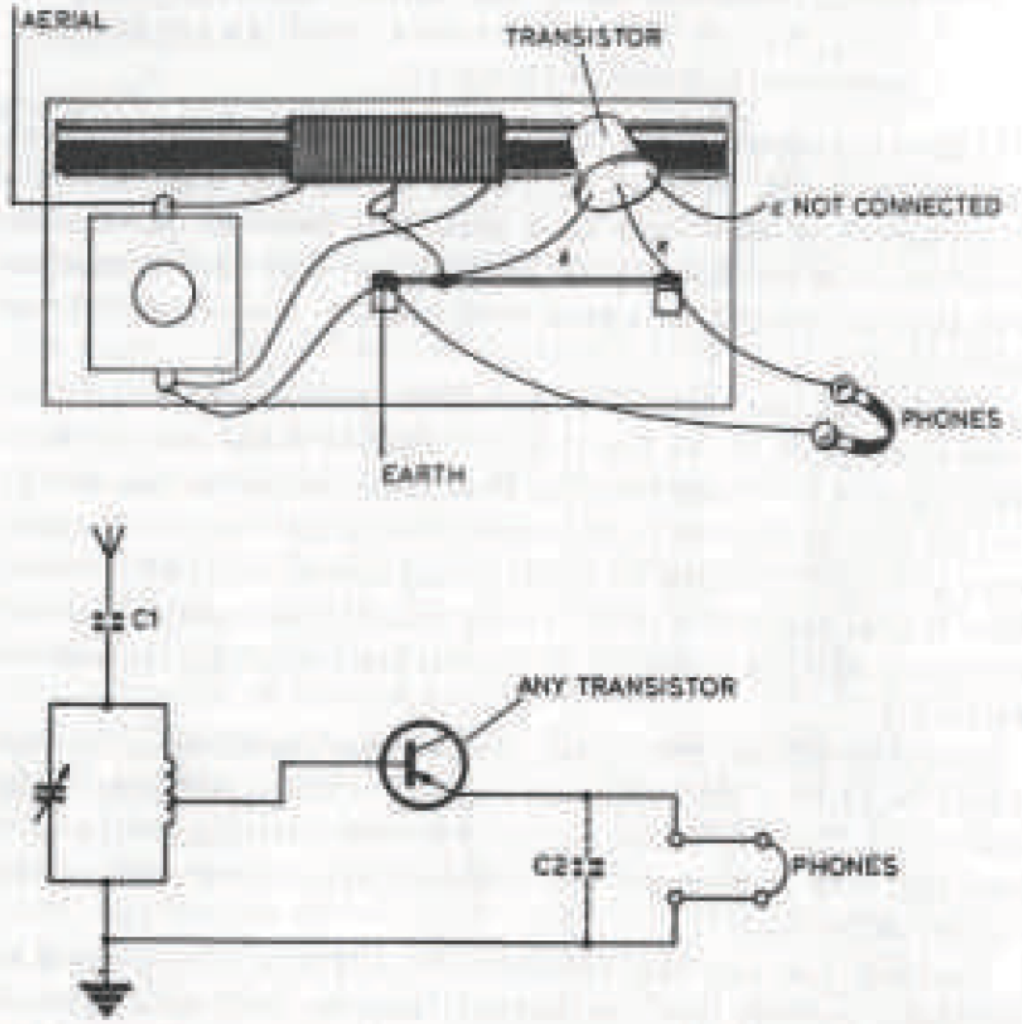

Fig. 3.6 *Circuit 2.* Crystal set using transistor instead of a diode.

as above (or near-equivalent) resistor R should be 15 k and capacitor C3 1 μF or higher.

This time a battery is also required to supply power for the transistor to work as an amplifier. This can be from 1.5 V up to 9 V. Remember the rules for polarity of connection: those shown in Fig. 3.7 are for a p-n-p transistor; an n-p-n transistor would need the battery connected the opposite way round. Battery polarity also affects the connections of capacitor C3 (if an electrolytic or polarized type).

Experiment further by trying the effect of using additional capacitors in the circuit, e.g. C1, 220 pF; C2, 0.001 μF (try other values as well); C4, 0.001 μF (try other values as well).

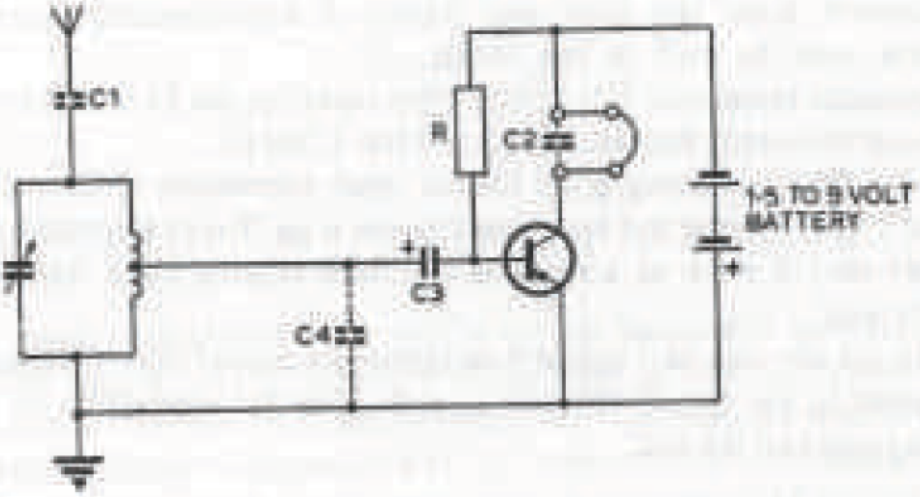
C1
R
C2
1·5 TO 9 VOLT
BATTERY
+
C3
C4
+

Fig. 3.7 *Circuit 3.* Crystal set with amplification.

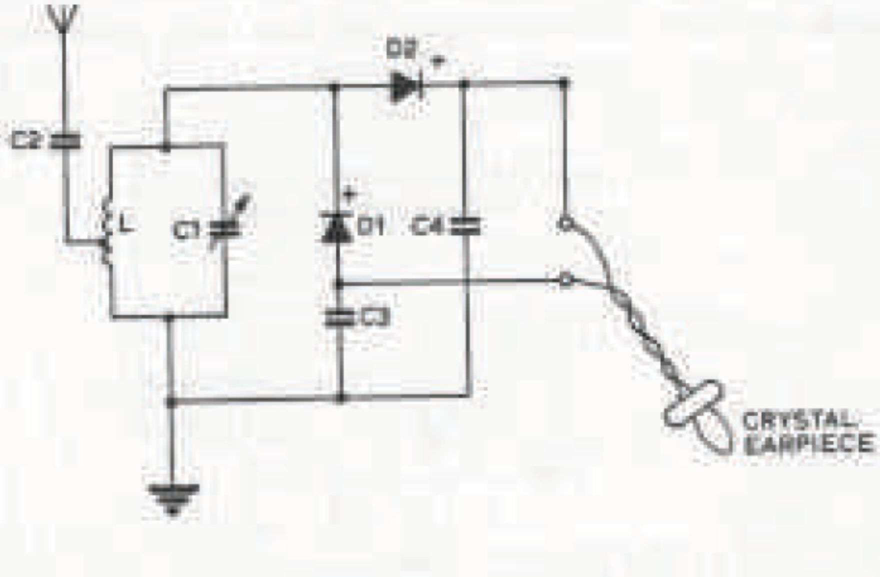
D2
+
C2
L
C1
+
D1
C4
C3
CRYSTAL
EARPIECE

Fig. 3.8 *Circuit 4.* Double diode crystal set.

This circuit incorporates 'voltage doubling' to improve the signal volume and should give better performance than a single diode circuit. L and C1 are the usual tuned circuit, but an unusual feature is that the *aerial* is connected to the coil tapping point. Any type of germanium diodes can be used (they should preferably be the same), making sure to connect them the right way round. A high-impedance *crystal* earpiece must be used in this circuit.

Capacitor values are: C2, 220 pF (this capacitor can be omitted–try with and without in the circuit); C3 and C4, 1,000 pF.

The optimum tapping point for the aerial connection to the coil L is best found by trial and error (see Chapter 4 on 'Tuned Circuits'), but the set should work at a nominal one-third tapping point from the 'earthy' end.

This set can also be tried with conventional tuned-circuit coupling–*i.e.* aerial to the top of the coil L, and diode D2 connection to the tapping point on the coil.

BENJAMIN AND DAVID GOLDENBERG 'S VERY-FINE OLD-TIME CRYSTAL RADIO

Parts List:

Item	Qty.
PVC pipe, 1" outside diameter (3/4” size) x 4" long	1
Aluminum or steel sheet metal, 0.5” wide x 5” long	1
Coil wire (#24 magnet wire)	36 ft
Hook-up wire (#22)	25 ft
Crystal earphone (GC electronics, cat. no. 30-5302)	1
Jack for the earphone plug	1
Crystal diode (1N34A)	1
Capacitor, 470 picofarad	1
#8 sheet metal screws	8
8-32 machine screw, 1" long	1
8-32 nuts	2
flat washers	6
Ring terminal	1
Board, 3/4" thick, approximately 7" x 9"	1

What else you need:

Vise-Grip pliers
Wire cutters
Soldering iron and solder
Drill, with 1/8", 3/16” and 3/8” bits
File or motorized grinder (e.g. Moto-Tool)
Screwdriver
Scotch tape

Instructions

I. CONSTRUCTION OF SPECIAL PARTS.

The following steps involve making a few special parts for the radio. Some of these steps are most easily performed using power tools and some involve soldering. This work should be done by someone old enough and knowledgeable enough to handle these tools safely. Once these parts are made, however, the rest of the project requires only hand tools and can be completed by nearly anyone.

1. Coil Form

The coil form is a piece of PVC pipe, 1” in diameter (outside) and 4” long. (For reasons fully understood only by plumbers, pipe of this diameter is referred to as 3/4”.)

To facilitate mounting the coil to the board, drill two sets of holes, on opposite sides of the form. On one side of the form, drill two holes, 3/16” in diameter and 3/8” from the ends. Directly opposite from these holes, drill two more holes 3/8” in diameter. Screws will be placed through the smaller holes, while the larger ones make it easy to reach the screws with a screwdriver.

2. Slider

Drill two 3/16” holes in the piece of sheet metal, each 1/4” from an end. After the holes are drilled, place a bend in the metal 3/4” from each end, as shown in the figure on page 3. Use a vise or pliers to bend the metal. Use a file to smooth all of the edges and corners of the metal.

CAUTION: Drilling sheet metal can be very dangerous! If the bit catches on the metal, you can have a very sharp edge flying through your hands. Hold the metal firmly with a pair of pliers (preferably Vice-Grip type), not with your hands! Also, place a piece of wood under the metal as you drill. This will help keep the metal from tearing and grabbing the bit as the passes through.

3. Pointed screw for slider

Using either a file or a grinder, grind a point on the 8-32 machine screw. The point doesn’t need to be very sharp; about like the tip of a medium ball-point pen is ideal.

4. Earphone jack with wires.

Cut two pieces of hook-up wire 2” long. Remove 1/4” of insulation form one end of each wire and solder the wires to the terminals of the earphone jack. Remove 1/2” of insulation from the other ends of the wires.

Note: Headphone jacks often have more than two terminals. Make sure that the terminals you solder to are the ones that will be connected to the tip and side of the plug when it is inserted.

5. Ring terminal with two wires.

Cut two pieces of hook-up wire, one 7” long and the other 2”. Remove 1/4” of insulation from one end of each wire. Solder the two wires to the ring terminal. Remove 1/2” of insulation from the other ends of the wires.

6. Board

Tape the layout template (the last page of this document) to a board, approximately 7” x 9”, and drill a 1/8" hole at each position marked with a cross.

II. ASSEMBLY

1. Wind the coil.

The coil is made by wrapping the #24 magnet wire onto the coil form, which is made from a piece of PVC pipe. The magnet wire is the wire that looks like it has no plastic insulation on it. But, it is actually coated with a thin layer of enamel. This coating must be removed from the ends of the wire before making any connections.

The best way to wind the coil is with two people, one to hold the wire taught (the holder) and the other to wind it onto the coil form (the wrapper). First, thread the end of the coil wire from the outside of the form through one of the large holes and out the end of the form. Leave about 4" of wire going out the end. Hold onto the other end of the wire and begin wrapping it around the form, starting right next to the hole. After you have wound a few turns on, push the wire together so it is nice and neat and put a piece of Scotch tape over it. The coil should look like this:

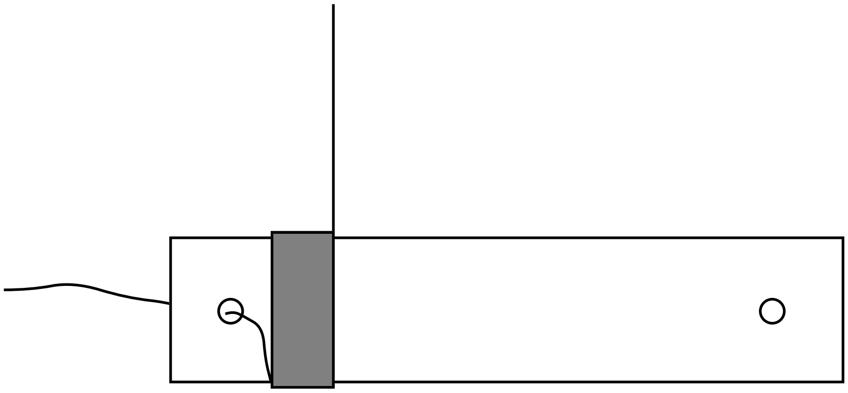

Now, the "holder" should hold the spool of wire loosely while the "wrapper" walks across the room unwinding the wire. Then, the holder should hold the spool very tightly so that the wrapper can wind the wire onto the pipe evenly. Each turn should be wrapped on right next to the previous turn, but shouldn't overlap. As the wire is wrapped on, the wrapper slowly walks across the room keeping the wire tight. When the wrapper reaches the holder, hold the wire that has been wrapped onto the form tightly, while the holder lets the wire on the spool unwrap. The wrapper then walks back across the room unwinding the wire from the spool, and the same procedure is repeated.

When the coil is within about 1/8" of the holes at the other end of the coil, stop winding. Put a piece of Scotch tape on the end of the wrapping, and cut the wire a few inches from the end of the coil. Thread the end of the wire through the large hole in the form.

2. Attach the coil to the board.

Use two of the #8 sheet metal screws to attach the coil to the board, using holes B and C that you drilled. Place the screws through the small holes in the coil form from the inside. You can then use the large holes on the other side of the form to reach the screws with a screwdriver. Make sure the coil is attached snugly to the board.

3. Assemble the slider.

Find the slider, the 8-32 screw and the two 8-32 nuts. Thread one of the nuts onto the screw until it is about 3/8" from the head of the screw. Put the screw through one of the holes in the slider and thread the other nut on, so that the slider looks like this from the side:

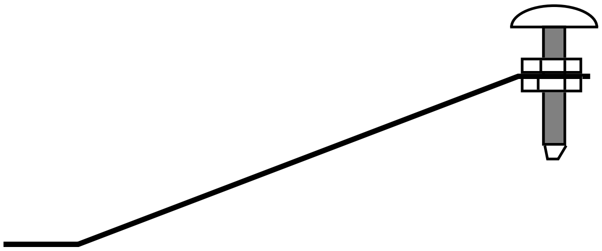

Tighten the second nut against the slider and the first nut with a wrench.

4. Attach the slider to the board.

Find a sheet metal screw, a washer and the ring terminal with two wires attached to it. Put the screw through the ring terminal, the washer and finally the open hole in the slider assembly (from the top of the slider as shown above). Attach the entire assembly to the board with the screw at hole E. Tighten the screw so that it is snug but still allows the slider to move back and forth. The pointed end of the machine screw at the other end of the slider should press firmly against the wire on the coil. If necessary bend the slider so that it makes firm contact with the coil but can still move.

Move the slider back and forth on the coil to scrape the enamel insulation from the coil wire where the pointed end of the machine screw touches it. The screw should be able to make an electrical contact with the coil wire at any position along the length of the coil.

5. Put the remaining screws in the board.

Screw the remaining sheet metal screws into holes A, D, F, G and H, with a washer on each screw. Do not tighten the screws yet.

6. Wire the components of the radio together.

Use the circuit drawing as a guide. For each connection, make sure that the insulation is removed from the end of the wire. Wrap the bare ends of the wires around the screw between the screw head and the washer underneath.

a. Wrap the two wires from the earphone jack around screws F and H. Do not tighten these screws yet.

b. Take the short wire from the slider and wrap the bare end around screw F (which should already be attached to one of the earphone jack wires). Do not tighten the screw yet.

c. Take the long wire from the slider and wrap the bare end around screw A. Do not tighten this screw yet.

d. Find the crystal diode, the small cylinder with two bare wires attached to it. Wrap one of the wires around screw H (attached to the other earphone jack wire), and tighten the screw.

e. Take the other wire from the diode and wrap it around screw G. Do not tighten this screw yet.

f. Find the capacitor, the small disc with two bare wires attached to it. Wrap one of the wires around screw F and tighten the screw.

g. Wrap the other wire from the capacitor around screw G, but do not tighten this screw yet.

h. Cut a piece of hook up wire about 6" long and remove about 1/2" of insulation from each end. Wrap one end around screw G and tighten the screw. Wrap the other end of the wire around screw D, but do not tighten this screw yet.

i. Take the wire from the right hand end of the coil and remove the insulation from about 1/2" from the end of the wire. To remove the insulation, either lightly scrape the wire with a knife or use a small piece of sandpaper. Wrap the bare end of the wire around screw D, but do not tighten the screw yet.

7. Connect the radio to ground.

In order to work, the crystal radio must be wired to a good ground connection. Often, the easiest way to make a ground connection is to use the grounding circuit for the house's electrical system. Cut a piece of hook-up wire long enough to reach from the radio to a nearby electrical outlet. Remove 1/2" of insulation from each end of the wire. Wrap one end around screw A of the radio and tighten the screw. On the electrical outlet, loosen the screw that holds the plate on the box. Wrap the other end of the wire around this screw and tighten it.

CAUTION: Be sure not to let the wire or anything else touch the 120 V socket!

Alternatively, you can use a cold water pipe to connect the radio to ground. Remove a few inches of insulation from the wire from screw A of the radio and wrap the bare end around a cold water pipe.

8. Connect an antenna to the radio.

Use the rest of the hook-up wire to make an antenna. Remove 1/2" of insulation from one end of the wire and wrap it around screw D. Tighten the screw. Unwind the wire from the spool and try to place it as high as possible in the room.

In areas with many powerful radio stations nearby, an inside antenna should be adequate to receive a few stations. If you have trouble hearing anything, it may help to set up a long outside antenna. Radio Shack sells a kit with the necessary wire and other parts (cat. no. 278-758).

9. Try the radio!

Plug the earphone into the jack. Listen to the earphone and slowly move the slider across the coil. Try to find the position where a station is loudest. You should be able to hear a few different stations with the slider at different positions.

10. If the radio doesn't work, check the following:

a. Make sure all of the connections are correct and that the insulation has been removed from the wires where they are wrapped around screws. Also check that there aren't any short circuits between the screws or bare wires.

b. Make sure that you scraped the insulation from the coil wire with the slider so that the slider can make good contact with the coil.

c. Make sure that you have a good ground connection. Try a different electrical outlet or a cold water pipe.

d. Place the antenna higher up or try a longer wire.

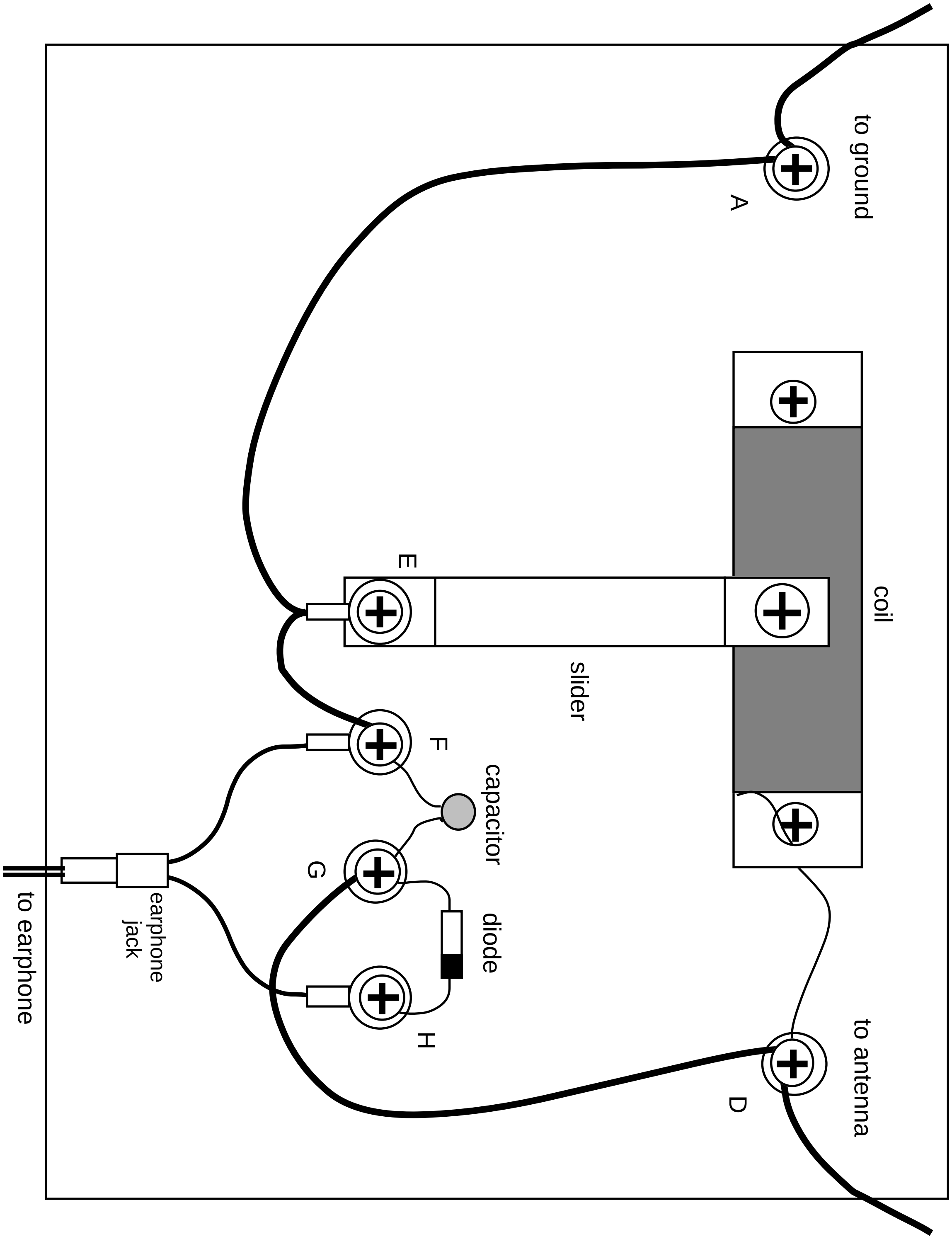
to ground
A
coil
slider
E
F
capacitor
G
diode
H
to antenna
D
earphone
jack
to earphone

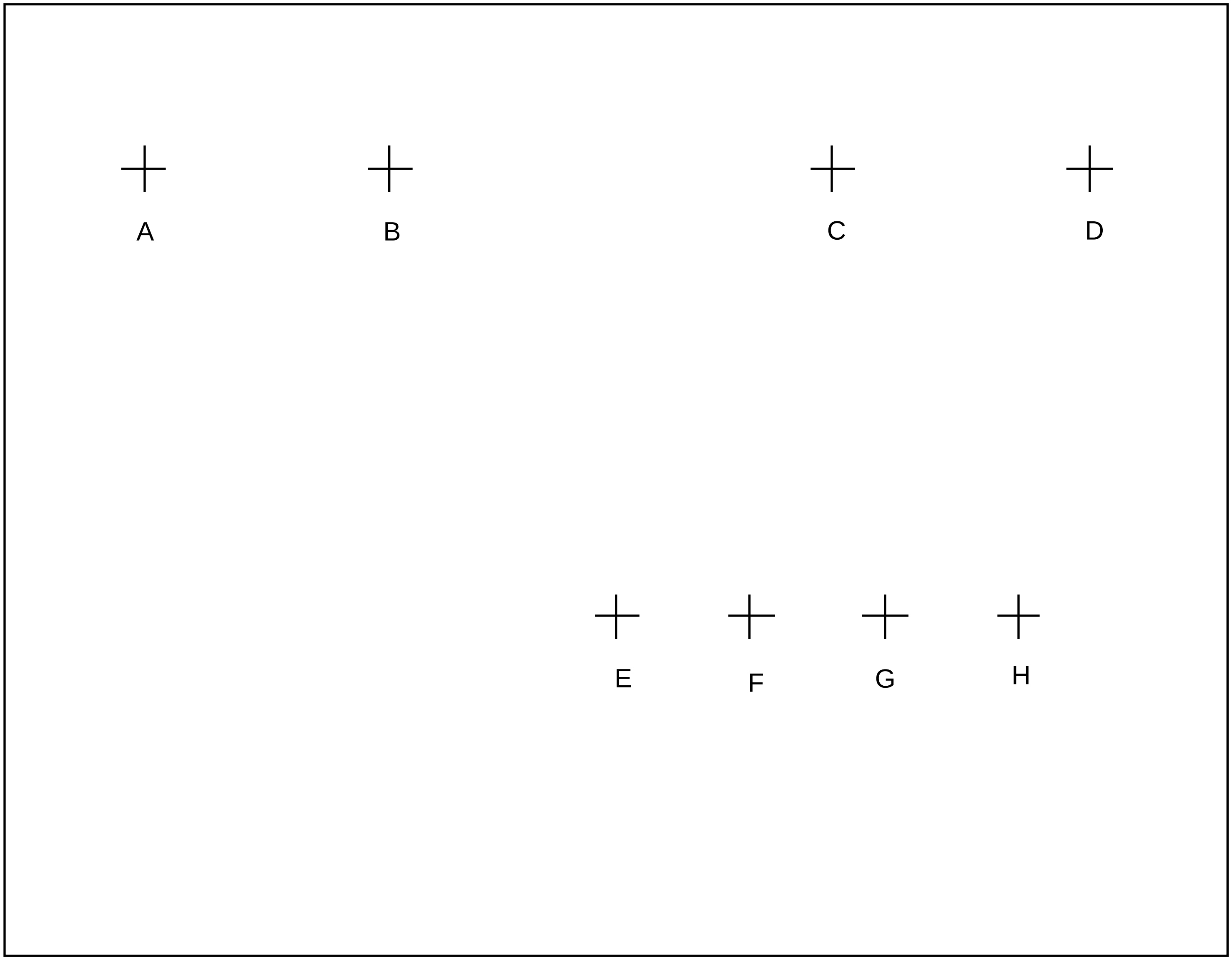
A
B
C
D
E
F
G
H

M0DAD RADIO WEBSITE

Basic Crystal Radio Set

I have been very interested in crystal radios for many years, I can remember building my first set back in my schooldays in the early 1980s. It was the first and still is the most basic of all radio receivers , If you are new to radio and have never seen one of these sets you will be very amazed at how easy it is to build a working radio needing so few components. What also may surprise a newcomer is that it does not need any battery or power to work, The only power it uses is taken from the transmitter of the station you are listening to, In other words it pulls the power from the radio wave.

The biggest craze of building and listening to crystal sets was at its peak in the 1920s. The old style sets used a real mineral crystal for the detector, The most popular being the galena and a fine bit of wire called a cats whisker which you would have to touch the crystal in the correct place to find the best reception where the radio station would come in. During world war II the underground people of Europe used crystal sets to listen to prohibited broadcast stations, Because of there simplicity to build. It is now getting harder to obtain all the original parts like the high impedance headphones, But I will show you how to build the modern day circuit using readily available parts.

My basic crystal radio project is aimed at the absolute beginner or younger generation, to try and interest them in the radio hobby as many amateur radio operators began there interest in radio with building the crystal set. The circuit can be soldered or built using chock block connectors for safety if a young person is doing this project.

CIRCUIT

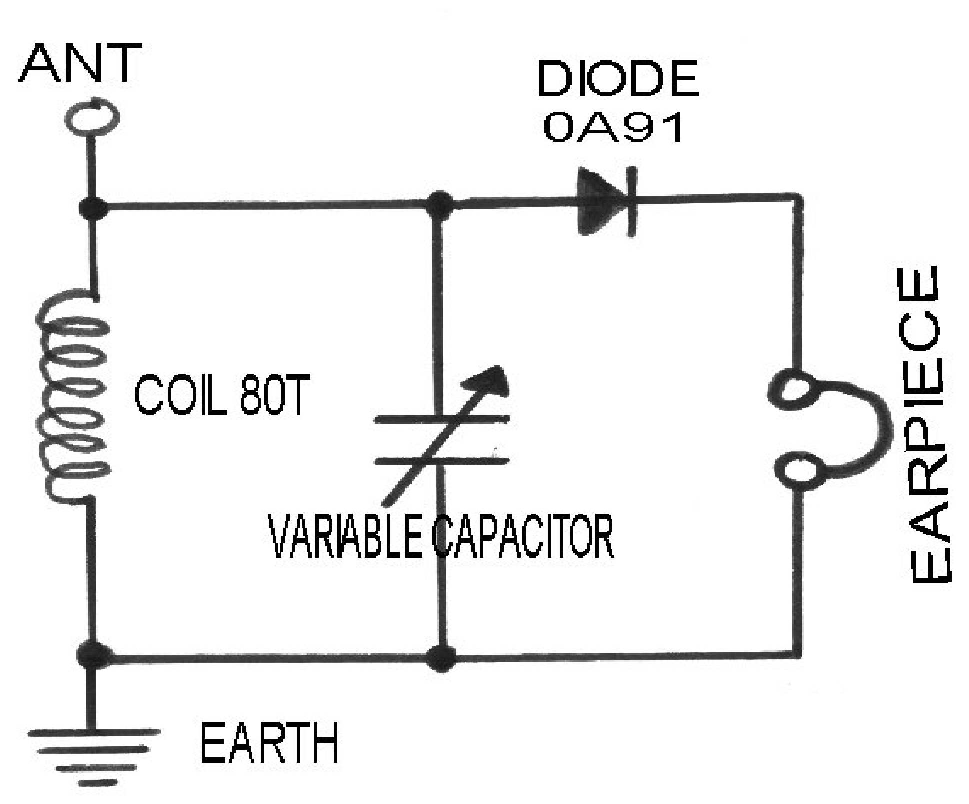
ANT
DIODE
0A91
COIL 80T
VARIABLE CAPACITOR
EARPIECE
EARTH

HOW THE CIRCUIT WORKS

First of all there are many different signals from stations which will be traveling down the long wire antenna, The coil and capacitor together form a tuned circuit, And because the capacitor is a variable type you can change the resonant frequency of the tuned circuit by rotating the tuning knob on the capacitor. Only signals from stations on or near the resonant frequency can pass through the tuned circuit, The other remaining frequencies are bypassed to earth. The passed signals then go through the germanium diode (detector) which rectifies the positive half cycles changing it to DC current and then sending it to the crystal earpiece or high impedance headphones.

BOARD LAYOUT

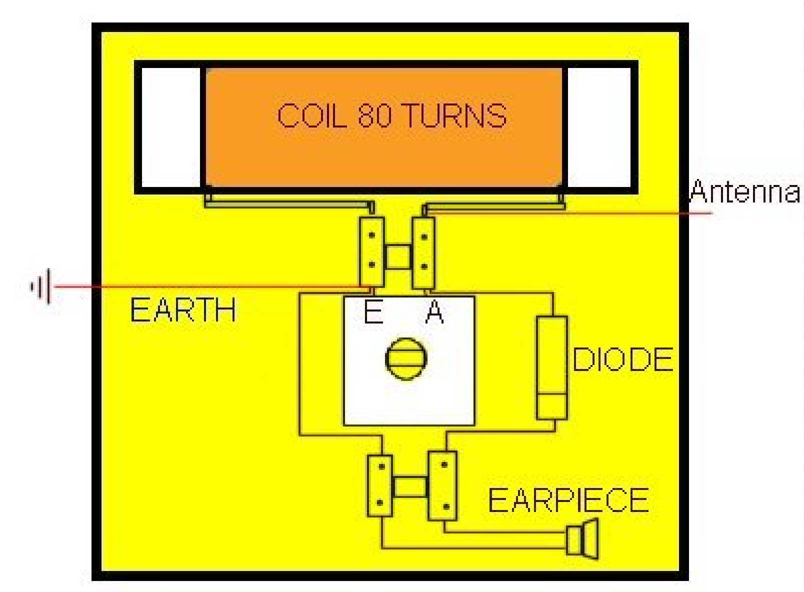
COIL 80 TURNS
Antenna
EARTH
E
A
DIODE
EARPIECE

PARTS

I personally find it a lot more fun to salvage all the parts from an old radio as it will cost you nothing, But just for this project I have listed all the parts that you will need and where to buy them from.

Variable capacitor, available from Maplin Electronics (UK) Code FT78K

Diode: Maplin Electronics 0A91, Code QH72ZP, 0A90 Code QH71N or any germanium point contact diode such as an OA70,OA71, or 1N34.

Chock block connectors Maplin Electronics (UK).

Crystal earplug, avail able from Maplin Electronics (UK) Code LB25C

26 s.w.g. Enameled copper wire.

Coil former: An empty card board toilet roll or an off cut piece of plastic waste pipe.

Base: Any piece of soft scrap wood about 4"x 5"

PVC coated wire for links.

A roll of any kind of wire PVC coated or copper wire for antenna and an earth connection.

CONSTRUCTION

Start the construction with the coil by securing the former (toilet roll, or PVC pipe) and then winding 80 turns of 26 s.w.g enameled copper wire around. When finished secure the windings with sticky tape or glue, Then scrape the wire enamel off both ends of the coil with a knife or sand paper until you can see a shiny tip. The coil can be secured to the wood base with two drawing pins.

Now the most difficult part is over , you can secure a couple of chock block connectors in front of the coil and the two behind the capacitor then wire the circuit as in the board layout, It is straight forward, but make sure you get the capacitor the correct way, the earth and antenna pins are marked e and a on the bottom, There is no connection to the pin marked o. The diode has a glass body with a band on to show the flow of the conducting current. Strip the plastic coating off the end of the crystal earpiece leads and insert them into the two rear chock blocks.

You are now ready to connect the antenna, Just string up the longest length of wire you can, run it from the set through a window and secure it to an insulated pole, tree or fence as high and as long as you possibly can to achieve the best results. If it is impossible for you to have an outdoor wire, String a wire several times around a room ceiling or an attic, It will still receive some stations but at reduced volume.

You will need an earth, run a wire from the set to a cold water pipe, connect it to the pipe or to a water tap with a crocodile clip, Or if you want a more permanent arrangement you can run a wire outside and connect it to an earth rod driven into the ground, This can be a length of copper pipe or brass rod as long and deep into the ground as you can. It is said that you must have a good earth for a crystal set to work, In practice I found it works slightly better but have also achieved good results without any earth.

OPERATING THE SET

This is the best part, When you are finished building the circuit connect the earth and antenna wires, pick up the earpiece and have a listen, You should be able to hear noise, turn the screw on the variable capacitor until you tune into a station, You should be able to hear several stations on the MW band, Depending on your location, length and how high the wire antenna is, Because this is a basic circuit you will soon notice that the set has poor selectivity, This means that sometimes you can hear two or more stations coming in together and may find it difficult to separate the one you want to listen to, That is if you can at all. If you can only receive one or two stations and the screw on the variable capacitor is fully turned, try unwinding about 4 or 5 turns off the coil. One of the most difficult requirements is having the patience, Some stations will come in loud and then may fade out altogether after a wile, If you leave it on the same frequency they can sometimes return as the propagation changes.

FAULT FINDING

If it does not work, Check your wiring with that of the board layout, And make sure all leads are not slack.

You must make sure you have a germanium point contact diode, other diodes will not work.

Make sure you have a proper crystal earpiece or high impedance headphones, Do not be confused with the standard white coloured magnetic earpiece, Unlike the standard 8 or 16 ohm earphones of today crystal earpieces are into the mega ohms, So if you used the wrong type it would be like trying to run your washing machine on a PP3 9 volt battery.

EXPERIMENTING

When I built this crystal set for the web site I hooked a wire to a radiator for an earth, strung a 60 FT length of PVC coated wire through a window down to the rear garden fence, The first station I heard was quite a surprise, Radio Korea, Other stations where Virgin Radio - 1215 KHz, Talk Sport 1071 KHz, Magic 1152 KHz, BBC Radio 5 Live 693 KHz, BBC Radio 4 - 603 KHz, So it had a good range through the medium wave broadcast band.

Do not be afraid to experiment, Due to the simplicity of the set you wont break it, You can try changing the diode around, Try other coils with different number of windings, The more turns should take you lower in frequency, and the smaller number higher, try a ferrite rod as a former or salvage a ready made coil from an old broken radio. Experiment with different lengths of antenna wire. I also find these sets work better with a higher quality air spaced variable capacitor which you can buy from a radio rally or you could try J. Birkett suppliers of electronic components who always stocks them at reasonable prices, TEL NO. +44 (0)1522 520767. The company is UK based and may not export, but I am sure that you will know of local electronics suppliers for such parts.

I hope you have enjoyed reading about the crystal set and if you have built the one from this page e mail me and let me know how it went. If this page has interested you and you would like to build more sets check out my other web pages or web links page.

BACK TO M0DAD HOMEPAGE

A Simple Crystal Radio Set

Whether you are a beginner or a seasoned expert, you would probably agree that there is nothing quite like using this simple design to make a crystal radio and enjoying it, listening to stations both near and far. Add to this the "something for nothing" advantage of not having to power this set with batteries or AC line and you can see why this simple design will be used and enjoyed for a long time to come. Almost like magic... Let' build one!

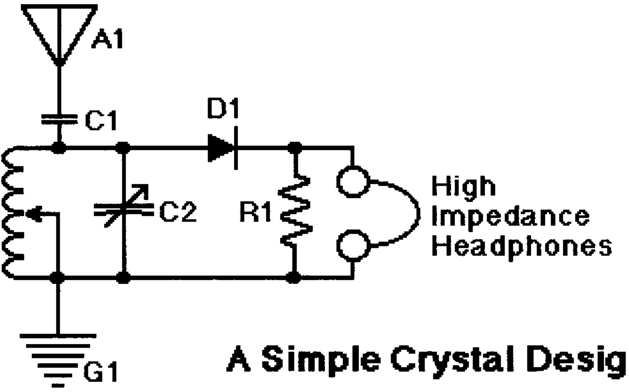
A1
C1
D1
L1
C2
R1
High Impedance Headphones
G1

A Simple Crystal Design

Components:

- **A1- 50-100 foot longwire antenna**
- **C1 - 50 pf ceramic capacitor OSE p/n CD-5**
- **C2 - 365 pf variable capacitor OSE p/n BC14400**
- **D1 - 1N34A diode OSE p/n 1N34A**
- **G1 - Ground (see instructions below)**
- **L1 - 100 turns of #20 AWG enamel coated copper wire OSE p/n MW20, adjustable tap (see instrctions below)**
- **R1 - a 100k ohm resistor OSE p/n RQ100K**
- **Crystal Earphones OSE p/n CE748**
- **Fahnestock clips OSE p/n 2090**

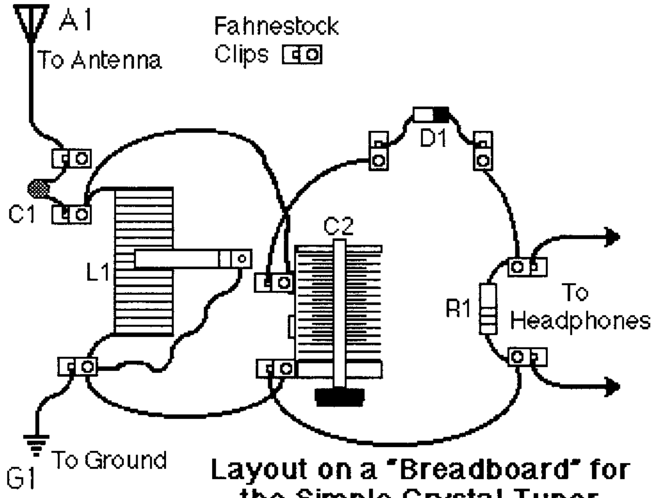
A1
To Antenna
Fahnestock Clips
C1
L1
D1
C2
R1
To Headphones
G1
To Ground

Layout on a "Breadboard" for the Simple Crystal Tuner

AIRCRAFT INSTRUMENTS AND AVIONICS

FOR A&P TECHNICIANS

JEPPESEN

by MAX F. HENDERSON

Table of Contents

About This Book v

About the Author vii

CHAPTER I Basic Instruments 1

Why Study Instruments?; Aircraft Instrument Requirements; Pitot-Static System Instruments; Gyro Instruments; Compass Systems; Electronic Instruments; Computers in Aircraft

CHAPTER II Powerplant Instruments and Logic Gates 55

Liquid Quantity Measuring Systems; Fuel Flow Indicators; Temperature Measuring Systems; Position Indicating Systems; Tachometers; Oil Pressure Indicators; Torquemeters; Engine Pressure Ratio Indicators; Manifold Pressure Gauges; Primary Power Setting Instruments; Vibration Indicators; Logic Circuits and Digital Systems; Takeoff Warning Systems; Angle of Attack Indicators; Stall Warning Systems; Annunciators; FARs for Warning Systems and Annunciators

CHAPTER III Communication and Navigation Systems 103

Radio Fundamentals; Regulations and Standards for Radios; Intercom and Interphone Systems; Communications Radios; Emergency Locator Transmitters (ELTs); Cockpit Voice Recorders and Flight Data Recorders; Navigational Systems; Long Range Navigation Systems; Instrument Landing System (ILS); Microwave Landing System (MLS); Radar Altimeter; Ground Proximity Warning System (GPWS); Weather Radar; Stormscope®; TCAS — Airborne Collision Avoidance System

CHAPTER IV Aircraft Antennas and Autopilots 149

Installation and Inspection of Avionics; Antenna Installations; Autopilots and Flight Directors

Appendix A Glossary 193

Appendix B Abbreviations 197

Index 201

About This Book

This textbook is intended to be used in the instruction of students in an aviation maintenance technician training program. The descriptions, drawings and graphics in this book are for instructional purposes only and should not be used as a technical reference source for specific maintenance tasks on aircraft or aircraft systems or for other operational purposes. Excerpts from Federal Aviation Regulations and other sources have been paraphrased and simplified in order to save space and time.

The author wishes to express his appreciation for permission to use material from the technical publications of the following aviation companies.

Beechcraft Aircraft Corp. (Wichita, Kansas)
Canadair Group, Bombardier Inc. (Montreal, Quebec, Canada)
Cessna Aircraft Co. (Wichita, Kansas)
Comant Industries Inc. (Sante Fe Springs, California)
Dayton-Granger Inc. (Fort Lauderdale, Florida)
Dorne and Margolin Inc. (Bohemia, New York)
Flight Dynamics Inc. (Portland, Oregon)
Piper Aircraft Corp. (Vero Beach, Florida)
Sensor Systems (Chatsworth, California)
Terra Avionics (Albuquerque, New Mexico)
United Technologies — Pratt & Whitney Canada Inc. (Longueuil, Quebec, Canada)

The Canadair drawings which appear throughout this book are the proprietary property of Bombardier Inc, Canadair Group. As such, it is forbidden to copy these drawings without the express written permission of Bombardier Inc., Canadair Group. Readers are cautioned that the drawings are training material only and as such are not subject to revision. They are not to be used in lieu of approved technical manual illustrations for the purposes of carrying out any maintenance procedure or any other activity on any Canadair aircraft or any other aircraft.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

About The Author

Max F. Henderson has been teaching Aviation Maintenance Technology subjects at Embry-Riddle Aeronautical University since 1982. Previous experience includes working as an Electronics Technician in the U.S. Air Force, as a Commercial pilot and A&P mechanic and as a Control Tower Operator. Mr. Henderson holds four FAA certificates:

1. Commercial Pilot Certificate — Ratings for single and multi-engine land, instruments, single-engine seaplanes and gliders.
2. Mechanic Certificate — Airframe and Powerplant ratings
3. Ground Instructor Certificate — Advanced rating
4. Control Tower Operator Certificate

During his years at Embry-Riddle Aeronautical University, Mr. Henderson has earned the following degrees: A.S. Degree in Aircraft Maintenance, B.S. Degree in Professional Aeronautics and a Master's Degree in Aeronautical Science. Mr. Henderson acquired his interest in aviation from his father Floyd B. Henderson whose flying and maintenance experience on aircraft began in 1932. Mr. Henderson's interest in aviation history is evidenced by a collection of books and photographs of early aviation dating back to the World War I era.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

CHAPTER I

Basic Instruments

Chapter one begins with a study of the general requirements for aircraft instruments and their installation. The categories and types of instruments are covered before beginning a study of specific instruments.

The instruments examined in chapter one include pitot-static system instruments, gyro instruments, compass systems and electronic instruments. Reference will be made to Federal Aviation Regulations which apply to these areas, particularly FAR Part 1 Definitions and Abbreviations, FAR Part 23, FAR Part 43 and FAR Part 91.

A. Why Study Instruments?

It is important for aircraft technicians to study aircraft instruments so that they will be able to inspect, install and troubleshoot them properly. There are also occasions when the technician will be running the aircraft's engines or other systems and will have to use the instruments himself. Many instruments are a part of a larger system and it is necessary to understand the interrelationships between the various parts of the system.

All certificated aircraft have instruments, although in the case of hot air balloons and gliders, only a few basic instruments may be required. The number and variety of instruments has increased over the years so that a small single engine airplane of today has more instruments and more sophisticated instruments than airliners had in the 1940s or earlier. Figure 1-1 shows the instrument panel of a Piper Cub from the early 1940s with its six basic instruments. This airplane was a small two place airplane which did not have an electrical system, so all the instruments used mechanical means of operation. The instrument panel shown in figure 1-2 is that of a modern single engine airplane equipped for "blind flying" or IFR flight operations. This airplane has many more instruments and systems that increase the safety of flight and make it a more efficient means of transportation. The most important instruments are placed directly in front of the pilot and the radios are grouped together in the middle for easy access to the controls.

The most common and important types of aircraft instruments and avionics systems will be described in this book along with some FAA requirements for testing and installation. The categories for aircraft instruments and the basic FAA requirements will be covered first.

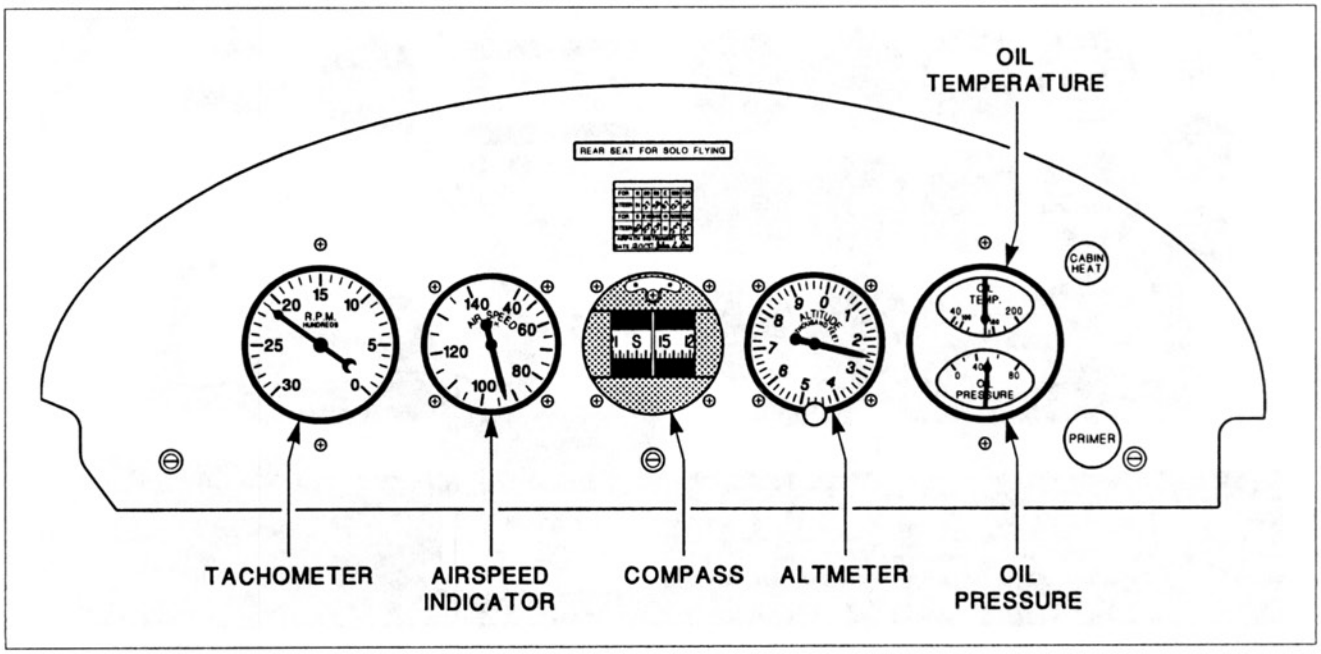

Figure 1-1. The instrument panel of a 1940s Piper Cub. (Courtesy Piper Aircraft Corp.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

B. Aircraft Instrument Requirements

1. Instrument Categories

The instruments found on different types of aircraft have considerable variety, but they can be categorized according to either how they work or what kinds of information they present to the flight crew.

a. Categories According to Application

1. Powerplant Instruments — These give information related to the aircraft's powerplant or powerplants.
2. Flight and Navigation Instruments — These give information such as altitude, speed etc. or information required for navigating the aircraft.
3. Systems Instruments — These concern aircraft systems such as electrical, hydraulic, pressurization, bleed air systems etc.

b. Categories According to Means of Operation

1. Pressure Instruments — These measure the pressure of air, fuel, oil, etc.
2. Mechanical Instruments — These use a mechanical system to obtain and/or transmit information.
3. Gyro Instruments — These use the principles of a gyroscope and are primarily used for IFR flight.
4. Electrical and Electronic Instruments — This group has seen the most change in recent years due to advances in digital technology and other related fields.

2. Instrument Placement and Installation

While there is not a standard placement for all the instruments that might be found on either a small or large aircraft, some of the most important instruments will have a standard layout directly in front of the pilot. This makes it easier for the pilot to scan the important instruments and it makes it easier to transition to a different type of airplane.

Aircraft instruments are manufactured in a number of standard sizes. This mainly applies to the round instruments since some other types come in a wide variety of sizes. The standard sizes for round instruments are:

1. 1″ — Often a vacuum gauge on single-engine airplanes
2. 2″ Flangeless — Many turbine engine powerplant instruments are this type.
3. 2-¼″ — A common size for many different instruments
4. 3-⅛″ — Considered to be a standard full-size instrument

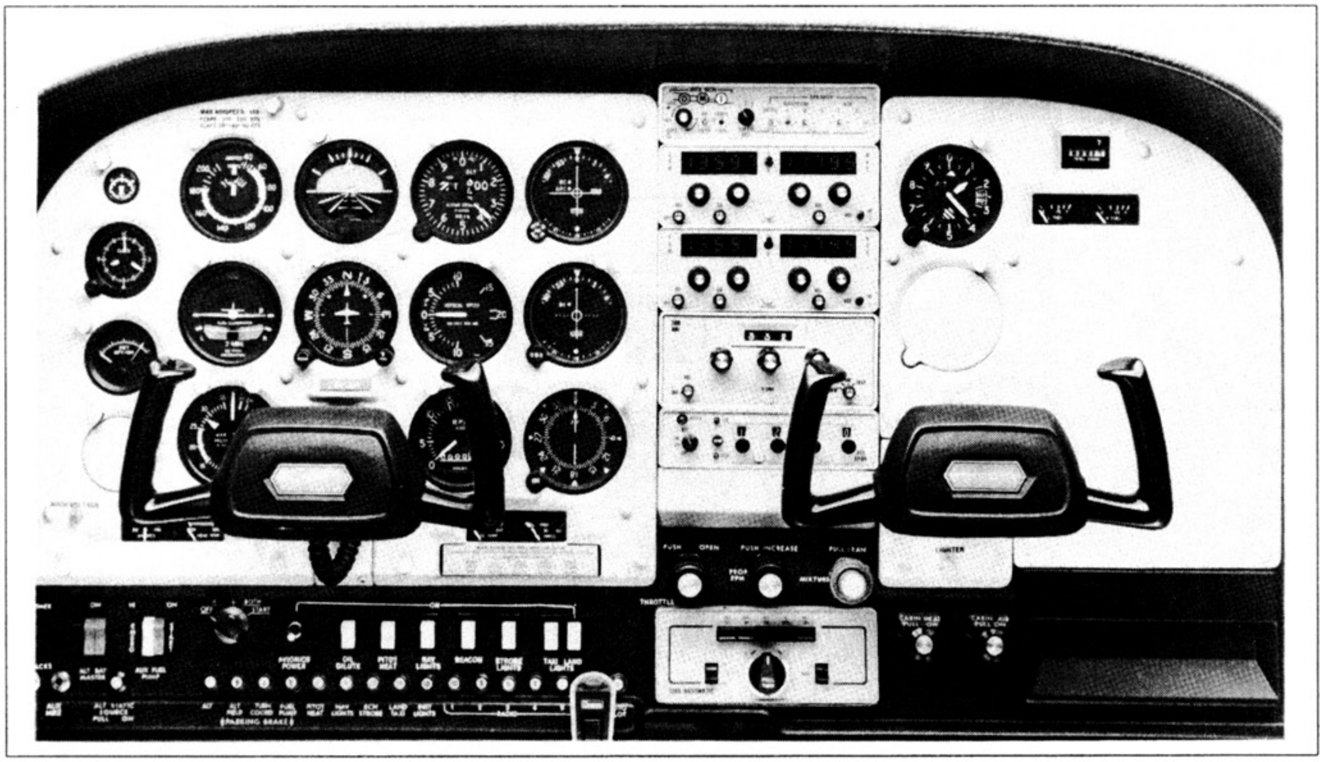
PUSH
OPEN
PUSH INCREASE
PULL LEAN
PROP RPM
MIXTURE
THROTTLE
LIGHTER
CABIN HEAT PULL ON
CABIN AIR PULL ON
AVIONICS POWER
OIL DILUTE
PITOT HEAT
NAV LIGHTS
BEACON
STROBE LIGHTS
TAXI LAND LIGHTS
ALT BAT MASTER
AUX FUEL PUMP
ALT STATIC SOURCE PULL ON
PARKING BRAKE
RADIO

Figure 1-2. *Typical arrangement of instruments and radios in a single engine airplane.*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

5. 4″ — The older style gyro instruments were often this size, but it is seldom used on a modern aircraft.

Examples of the 1″, 2-1/4″ and 3-1/8″ sizes can be seen on the instrument panel in figure 1-2.

There are a number of methods used to install instruments into the aircraft instrument panel. The three most common methods are:

1. Screws — Non-magnetic fine thread machine screws are used, brass screws with a black oxide finish are very common.
2. Circumferential Clamps — These clamps are fastened to the back of the instrument panel and operate somewhat like a hose clamp (figure 1-3).
3. Brackets — small "L" or "U" shaped brackets are installed on studs to hold the instrument in place (figure 1-3).

Installing aircraft instruments is often made difficult by the fact that there isn't much room behind the instrument panel and access is limited. Some aircraft use sliding or hinged panels to improve access to this area.

3. Instrument Markings

Aircraft instruments often utilize colored markings so that safe operating values can be indicated to the pilot. For example, red usually means a maximum or minimum operating limitation for the airplane or engine. These markings are normally on the face of the instrument, inside the cover glass. It is permissible to apply colored markings with paint to the cover glass, but if this is done the marks must not interfere with reading the instrument and a white line must be applied to the cover glass and case to act as a slippage mark.

When applying or inspecting the markings on instruments, a suitable reference source must be used, the acceptable sources are:

1. Approved Aircraft Flight Manual or Pilot's Operating Handbook — This is an FAA-approved document which is a part of the required aircraft equipment.
2. Maintenance Manual or Service Manual.
3. Type Certificate Data Sheet or Specifications.
4. STC, Manufacturers Service Bulletins, and AD Notes — These would indicate a change from the original aircraft requirements.

There are standard meanings for the different colors and markings applied to instruments like powerplant and system instruments, they are:

1. Red Radial Line — This indicates a maximum or minimum operating limitation. Example: on an oil pressure gauge.
2. Red Arc — This indicates a prohibited range of operation. A common example is the red arc on a tachometer because of vibration problems at certain RPMS.
3. Yellow Arc — This indicates a caution range.
4. Green Arc — The normal operating range.
5. Blue Arc or Line — This has a meaning specified by the manufacturer. An example is a blue arc on a manifold pressure gauge for engine operation with a lean mixture.

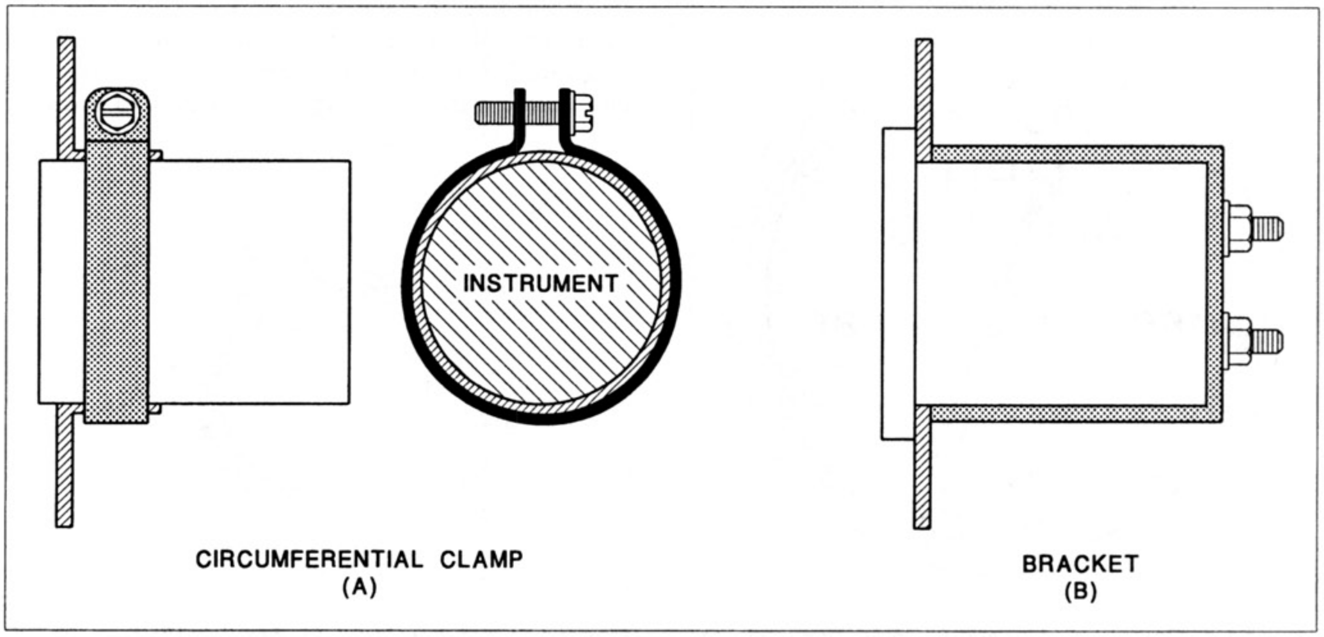

Figure 1-3. Two of the common methods used to install instruments in an aircraft instrument panel.

The colored markings on an airspeed indicator have different and very specific meanings, this will be covered later. The colored markings found on an airspeed indicator are shown in figure 1-4, notice that both arcs and radial lines are used.

4. General Precautions for Installing Instruments

A number of things must be kept in mind when installing instruments, radios and related systems. Instruments are very delicate and must be handled with care. If an instrument must be sent to a repair station or other place for repair, it should be packed in an approved container with any ports or openings plugged and with a desiccant pack to prevent moisture damage if it is sensitive to moisture. When shock mounts are required for a particular installation, you should ensure that the shock mounts are in good condition and that the weight of the equipment does not exceed the weight carrying capability of the shock mounts. Figure 1-5 shows a shock mount unit for aircraft equipment with a bonding jumper installed. Figure 1-6 illustrates the use of a shock mounted sub-panel for the flight instruments on a small airplane.

Many types of instruments can be damaged by static electricity and magnetic fields. Bonding jumpers, shielded wires and static wicks must be properly installed and maintained to prevent these kinds of problems.

Bonding jumpers have a number of functions when installed on an aircraft. Four of these functions are:

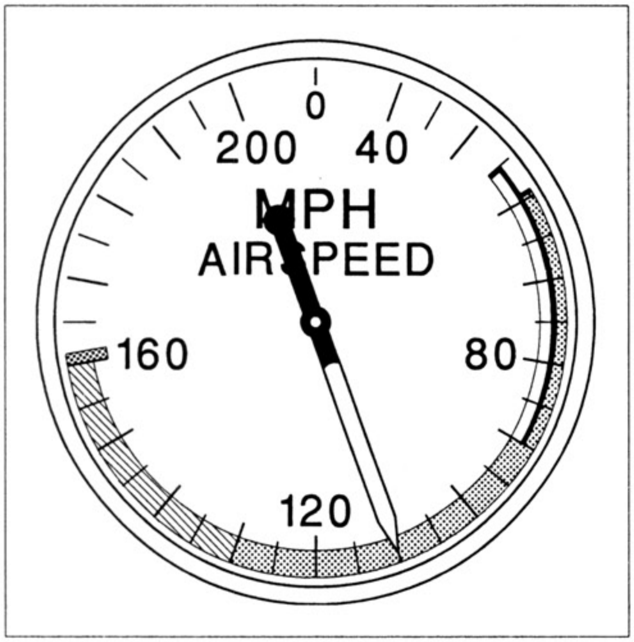

Figure 1-4. Operating limitations markings on an airspeed indicator.

1. A ground return path for aircraft circuits that use single wire type systems.
2. Reducing radio frequency interference in sensitive aircraft systems.
3. Decreasing the possibility of damage due to lightning strikes on control surfaces and other areas.
4. Allowing static charges to move around easily and equalize. This prevents arcing which could create a fire hazard.

A bonding jumper is a small metal braid or metal strap which electrically connects together two parts on the airframe. Some of the recommendations for installing and maintaining bonding jumpers are:

1. Keep them as short as possible but allow for any movement that is necessary as on control surfaces.
2. Clean the contact areas to minimize resistance. This includes removing coatings such as anodizing.
3. Do not solder bonding jumpers. It makes them brittle.
4. Do not paint bonding jumpers. It makes them brittle.
5. Use multiple jumpers on shock mounted electrical equipment. One might break.
6. Ensure that the jumper is compatible with the structural material and hardware to prevent corrosion. Aluminum jumpers are recommended for aluminum aircraft parts and copper jumpers are recommended for stainless steel, cadmium plated steel, brass and bronze.

Some instrument and equipment installations require the use of shielded wire. Don't assume that both ends of the wire shield should be connected to ground. It is sometimes specified that only one end of the wire shielding be attached to ground.

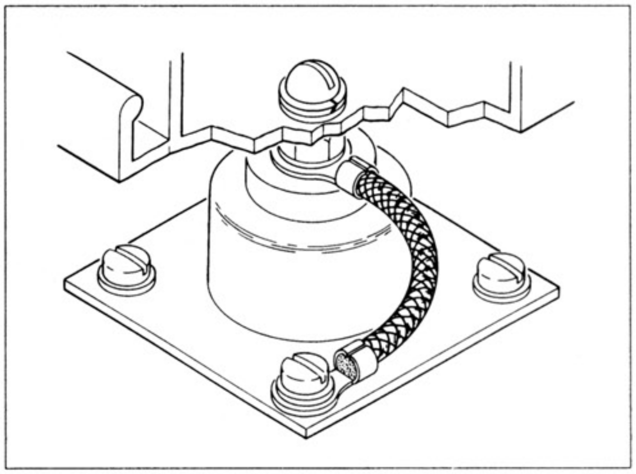

Figure 1-5. Bonding jumpers are installed on shock mounts to ensure proper grounding.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

When installing an instrument that requires lighting, use care to ensure that the light is installed correctly and that it will not short out or cause a problem with another instrument. There are often many small wires behind the instrument panel for the light fixtures and they must be routed and tied carefully. There are four common types of instrument lighting systems found on aircraft:

1. Eyebrow Lights — These are small semi-circular fixtures that fit over the top of a round instrument and look like eyebrows.
2. Post Lights — These are small round lights that install into holes in the instrument panel (figure 1-7).
3. Internal Lights — These are inside the instrument case.
4. Flood Lights — These lights can be aimed at the instrument panel and are shielded so that they don't shine in the pilot's eyes. They are often fitted in addition to one of the other types as a back-up lighting system.

Figure 1-8 shows the fluorescent flood lighting system for a corporate jet airplane.

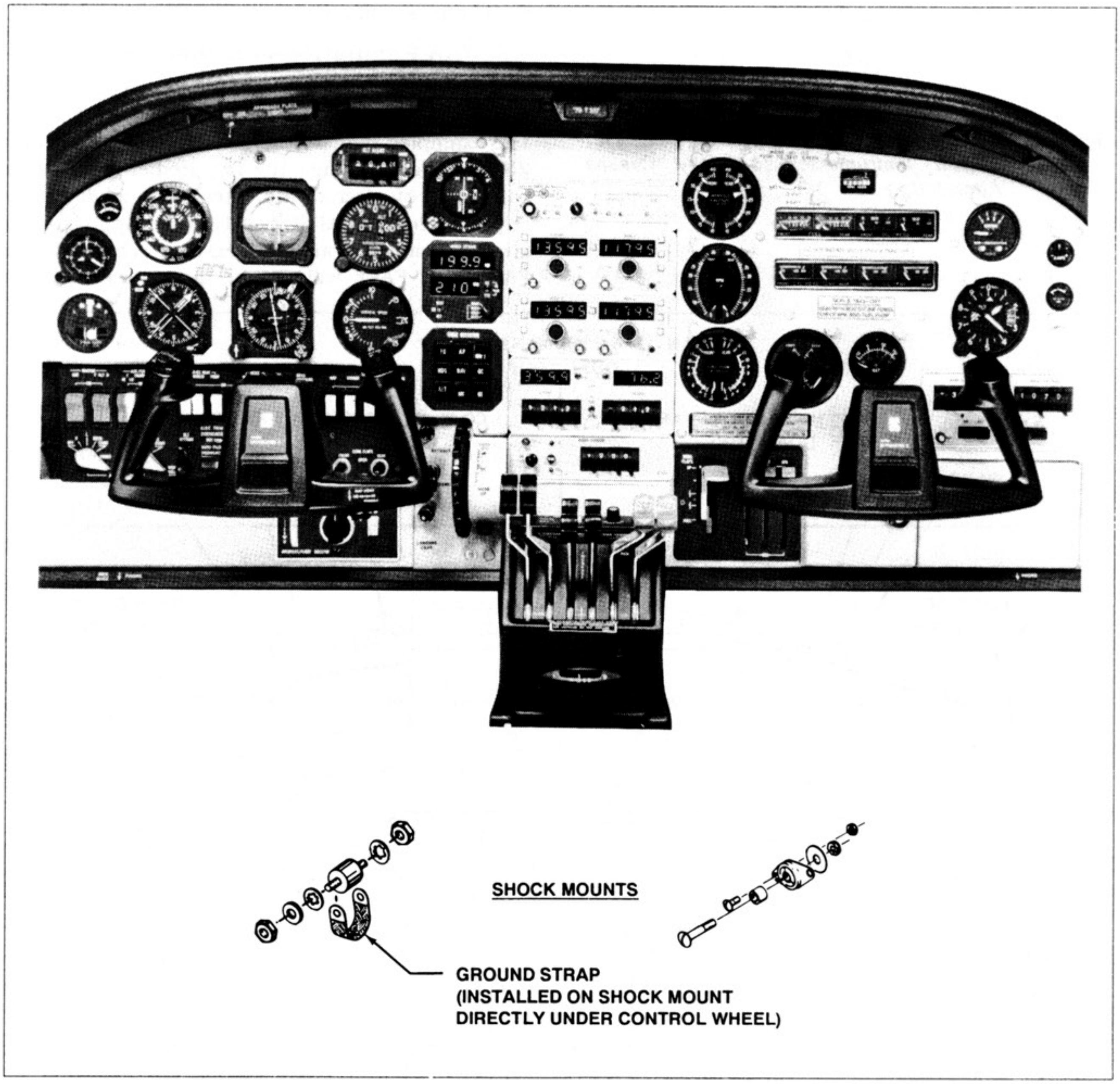

Figure 1-6. The instrument panel for a twin engine airplane showing the shock mounted sub--panel and two kinds of shock mounts. (Courtesy Cessna Aircraft Co.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

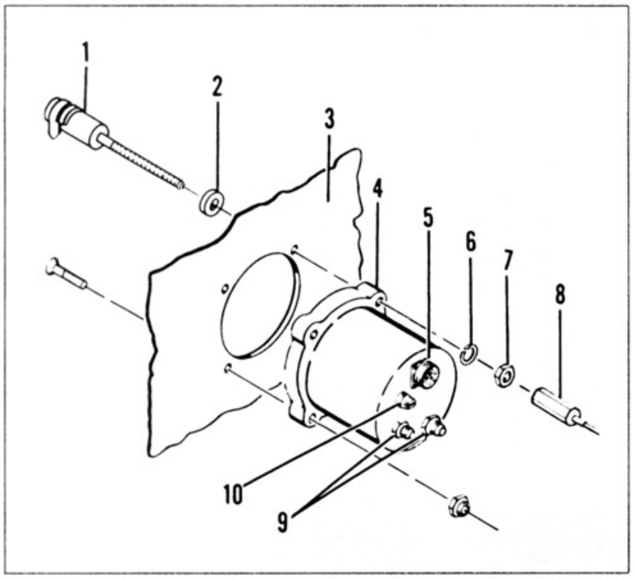
1
2
3
4
5
6
7
8
10
9

Figure 1-7. A post light type of lighting fixture (item 1) for aircraft instruments. (Courtesy Cessna Aircraft Co.)

Precipitation static, also known as P-static, is a build up of static electricity on the aircraft in flight. It can have an adverse effect on the operation of many instruments and radios. P-static is caused by friction between the aircraft structure and particles in the air such as rain, snow, ice and dust particles. It can also be caused by the hot exhaust of a turbine engine as it exits the large metal tailpipe or exhaust pipe. It cannot be prevented but the problems can be reduced by installing good static dischargers on the aircraft. These are normally installed on the trailing edges of main control surfaces and also occasionally on the tips of the wing and horizontal stabilizer.

5. FAA Regulations for Instruments

The FAA has many regulations that concern the installation of instruments in certificated aircraft. The examples that will be given apply to FAR Part 23 airplanes although the requirements for other categories of aircraft are often very similar. The requirements of FAR Part 91 would apply to any aircraft being operated under that section of the FARs. The FARs will not be quoted exactly, but

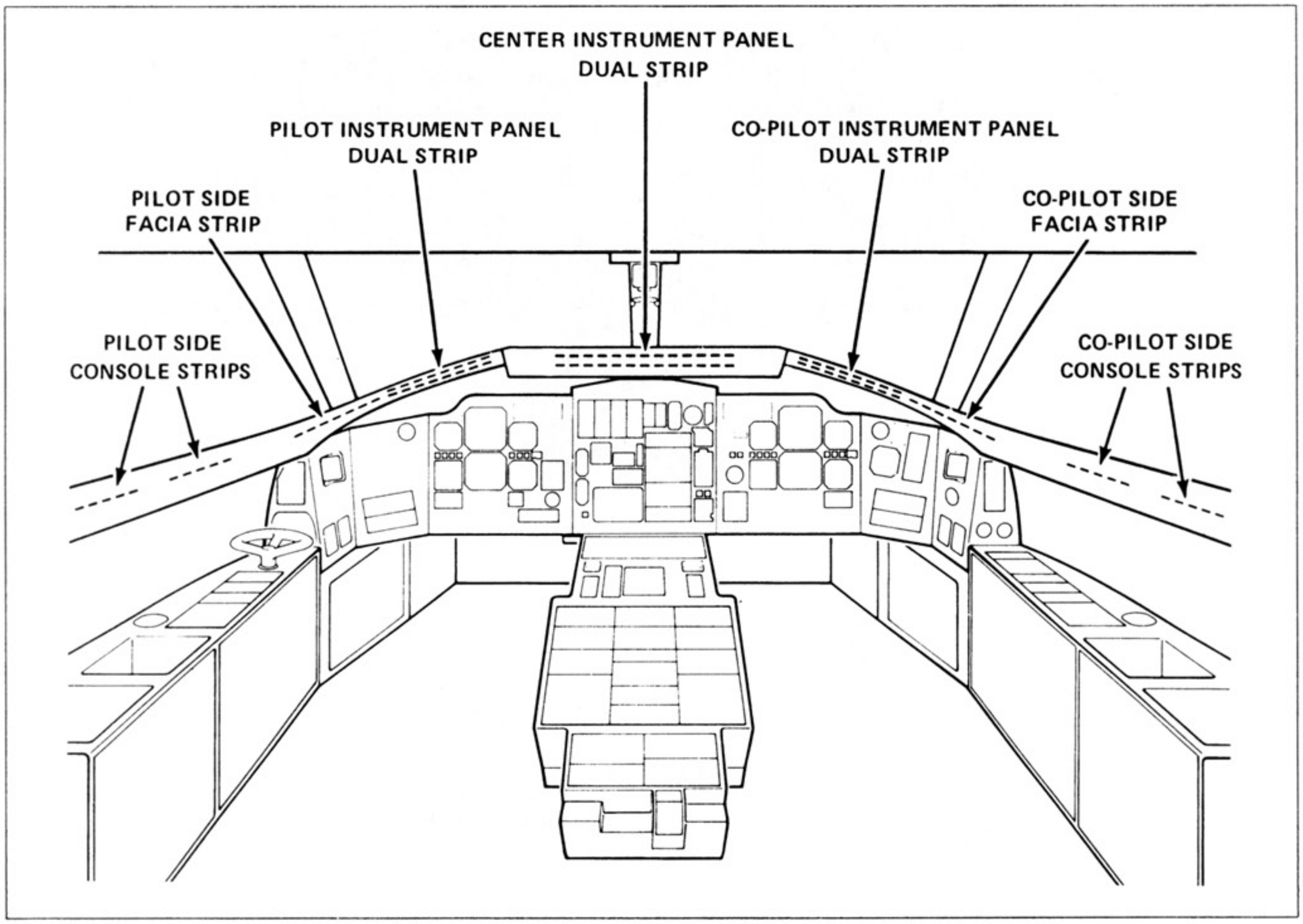
CENTER INSTRUMENT PANEL DUAL STRIP
PILOT INSTRUMENT PANEL DUAL STRIP
CO-PILOT INSTRUMENT PANEL DUAL STRIP
PILOT SIDE FACIA STRIP
CO-PILOT SIDE FACIA STRIP
PILOT SIDE CONSOLE STRIPS
CO-PILOT SIDE CONSOLE STRIPS

Figure 1-8. Fluorescent lighting arrangement for a corporate jet. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

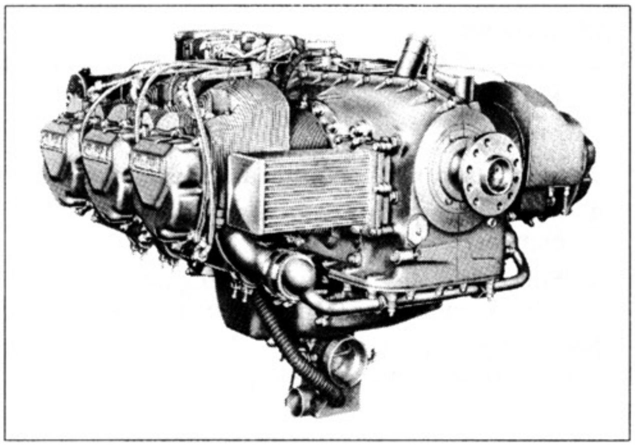

Figure 1-9. An un-supercharged reciprocating aircraft engine.

will be paraphrased in order to simplify the wording and save space.

Since some of the FAR rules for instruments depend on what type of engine the aircraft uses, it is necessary to discuss the different types of powerplants found on certificated aircraft. A reciprocating engine is a piston engine which can either be supercharged or non-supercharged. The most common type of reciprocating engine on modern aircraft is the horizontally opposed type although many radial piston engines are still in service. Figure 1-9 shows an unsupercharged horizontally opposed reciprocating engine. Figure 1-10 shows a radial engine with an internal supercharger. A turbine engine could be one of four types. The turbojet and turbofan engines are similar in that they are both rated in pounds of thrust. The difference is that with the turbofan engine some airflow bypasses the core of the engine and is acted upon only by the fan section as seen in figure 1-11. The turbojet engine illustrated in figure 1-12 does not use the bypass principle since all of the intake air passes through the length of the engine. In the FARs, the use of the term turbojet includes turbofan engines. The other two types of turbine engines are the turbo-propeller and turboshaft engines. These are both usually rated in horsepower because they deliver power to an output shaft (unlike turbojet and turbofan engines). The difference between the two is that the turboprop engine turns a propeller while the turboshaft engine powers the rotor drive gearbox of a helicopter. The reduction gearbox and output shaft of a turboprop engine can be seen on the left in figure 1-13.

FAR Part 23

This FAR covers the Airworthiness Standards for Normal, Utility, Acrobatic and Commuter category airplanes. Certain instrument requirements are a part of these Airworthiness Standards.

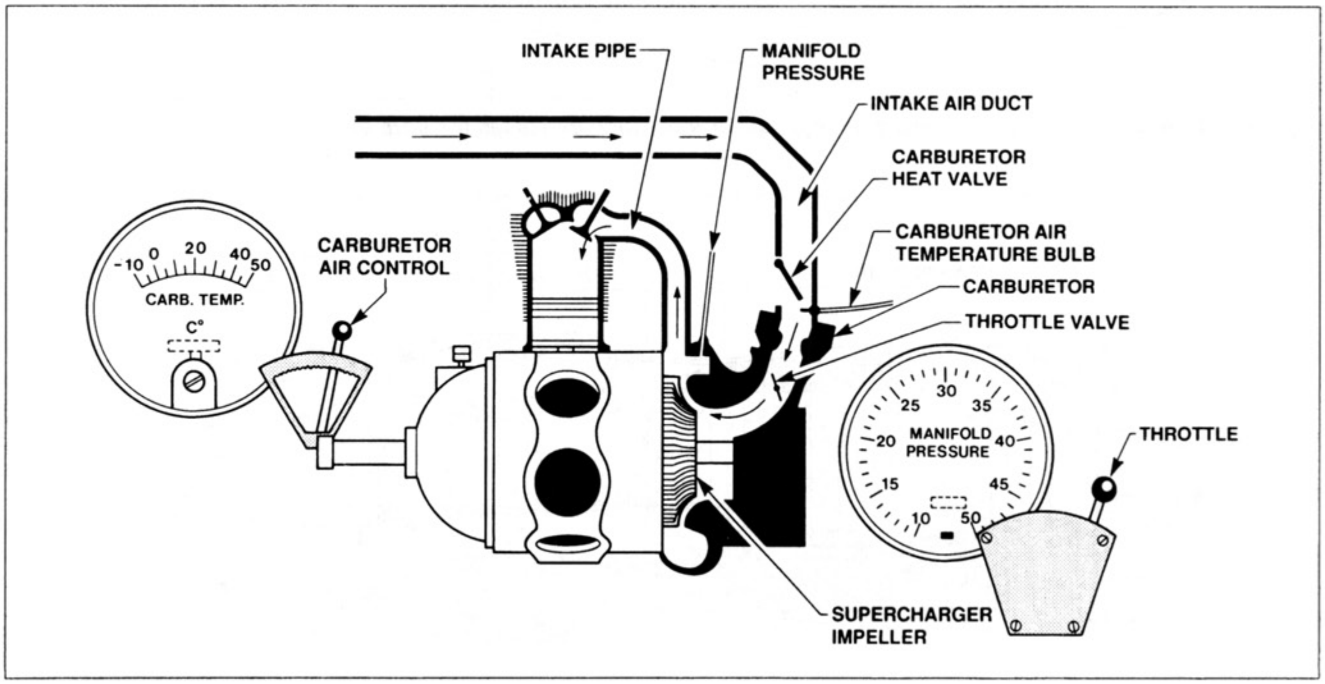
INTAKE PIPE
MANIFOLD PRESSURE
INTAKE AIR DUCT
CARBURETOR HEAT VALVE
CARBURETOR AIR TEMPERATURE BULB
CARBURETOR
THROTTLE VALVE
-10 0 20 40 50
CARB. TEMP.
C°
CARBURETOR AIR CONTROL
10 15 20 25 30 35 40 45 50
MANIFOLD PRESSURE
THROTTLE
SUPERCHARGER IMPELLER

Figure 1-10. A radial piston engine with internal supercharger.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

FAR 23.841

The additional instruments required for pressurized airplanes are specified in this section. Figure 1-14 illustrates these instruments. Pressurized airplanes are required to have instruments to indicate:

1. Cabin differential pressure
2. Cabin altitude
3. Rate of change of cabin altitude (cabin rate of climb)

FAR 23.1301

This FAR states that equipment must be labeled as to its identification, function and operating limitations. The colored markings on an airspeed indicator are examples of these required operating limitations.

FAR 23.1303

Required Flight and Navigation Instruments

1. Airspeed indicator
2. Altimeter
3. Magnetic direction indicator (compass)
4. Free air temperature for airplanes with turbine engines
5. Speed warning for turbine engine aircraft or others when Vmo/Mmo is greater than .8 Vd.

Notice that only the first three would be required for all FAR Part 23 airplanes.

FAR 23.1305

Required Powerplant Instruments

1. Fuel quantity for each tank.

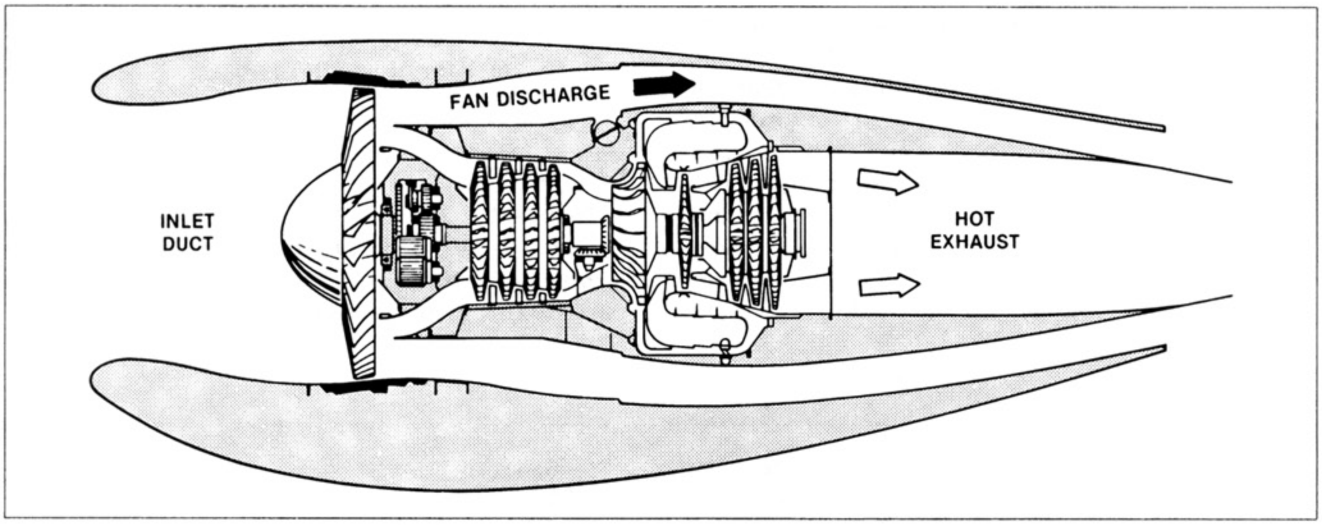

Figure 1-11. A turbofan engine showing the bypass airflow which bypasses the core of the engine.

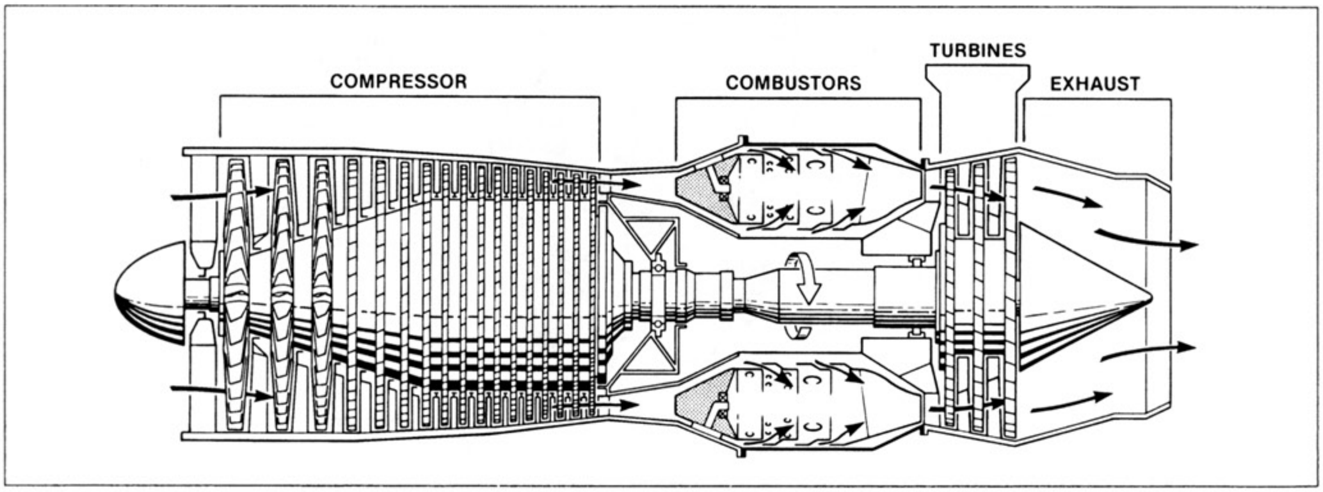

Figure 1-12. A turbojet engine with the major sections identified. This type of engine has no bypass airflow.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

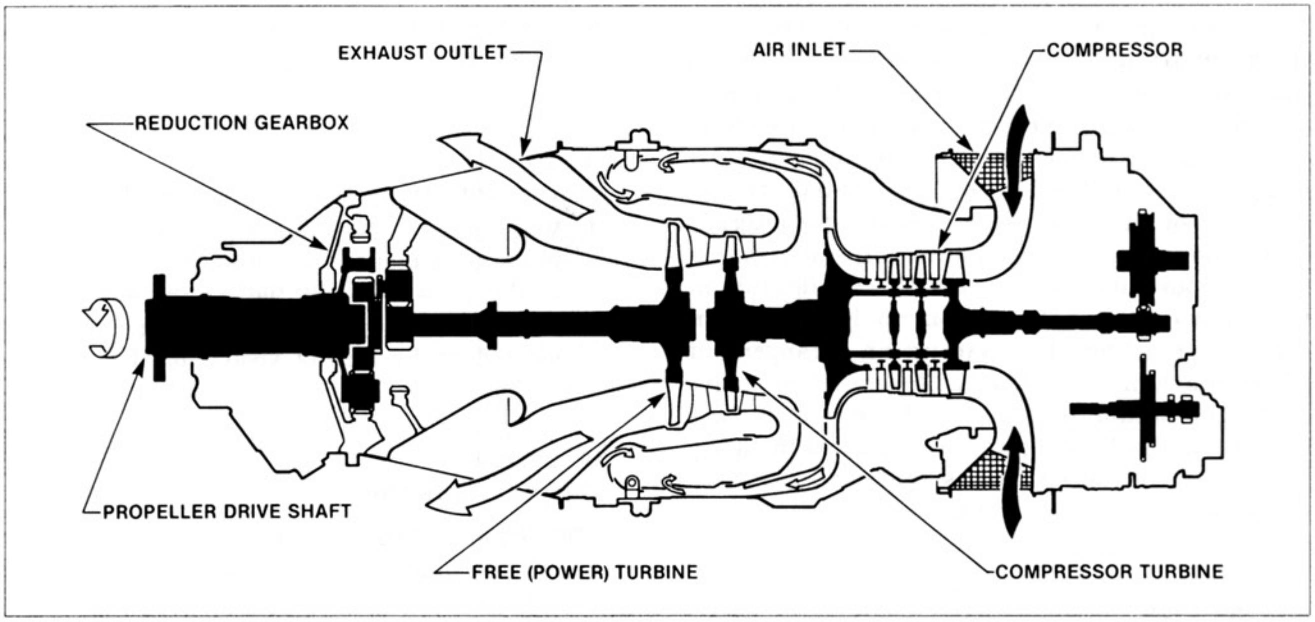

Figure 1-13. A typical turboprop engine with the output shaft on the left. The engine could become a turboshaft engine if it was modified to drive the rotors of a helicopter.

2. Oil pressure for each engine.
3. Oil pressure for each turbosupercharger (only with separate oil system).
4. Oil temperature for each engine.
5. Oil temperature for each turbosupercharger (only with separate oil system).
6. Tachometer for each engine.
7. Cylinder head temperature for:
   a. Air-cooled engines with cowl flaps.
   b. Reciprocating engine commuter category airplanes.
8. Fuel pressure if the engine is pump fed.
9. Manifold pressure for:
   a. Altitude engines.
   b. Reciprocating engine commuter category airplanes.
10. Oil quantity for each oil tank (if separate from engine).

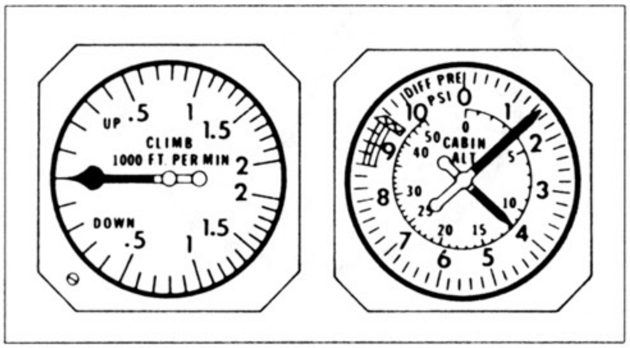

Figure 1-14. The three instruments required for a pressurized aircraft.

11. Gas temperature for turbine engines.
12. Fuel flowmeter for turbine engines.
13. Torquemeter for turbo-propeller engines.
14. EGT and carburetor inlet temperature for turbosupercharger installations if operating limitations can be exceeded.

Figure 1-15 shows a typical tachometer and manifold pressure gauge installation for a twin engine airplane with two pointers in each gauge for the left and right engines.

FAR 23.1321

Instrument Installation

1. Instruments must be plainly visible with minimum deviation of the pilot's position.
2. For multi-engine airplanes, identical instruments must be arranged to prevent confusion.
3. Instrument panel vibration must not damage the instruments.
4. For airplanes more than 6,000 lbs. maximum weight, the following instruments must be installed so that they are centered about the vertical plane of the pilot's vision in this order:
   a. Primary attitude instrument in the center.
   b. Airspeed indicator adjacent and to the left of the attitude instrument.
   c. Altitude instrument adjacent and to the right of the attitude instrument.
   d. Direction of flight instrument adjacent and below the attitude instrument.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The instrument referred to as the primary attitude instrument above would be an artificial horizon on smaller aircraft or an ADI on aircraft with more modern types of instruments. The direction of flight instrument in item 4 is not a magnetic compass. It would be a directional gyro on simple airplanes or an HSI on more sophisticated aircraft. In addition to the four standard instruments mentioned above as part of this "T" configuration, two other instruments are commonly installed in standard positions to make up an arrangement of six instruments. This is illustrated in figure 1-16. The actual appearance of the instruments in the standard configuration is shown in figure 1-17. The gyro instruments in figure 1-17 are the older style instruments.

FAR 23.1322

Warning, Caution and Advisory Lights

There are standard colors specified for certain indicator lights used in airplanes. A fire warning light for example would be red. The colors specified in this FAR are as follows:

RED — Warning — This is used when immediate attention is required.

AMBER — Caution — This is not as serious as a warning.

GREEN — Safe operation — normal operating range.

Any other color of indicator light can be used including white if it differs sufficiently from other colors and its meaning is specified by the manufacturer.

FAR 23.1381

Instrument Lights

Any instrument lights that are installed must make the instruments and controls easily readable and must be shielded so that they don't shine in the pilot's eyes. A cabin dome light is not acceptable as an instrument light.

FAR 23.1541

Markings and Placards

The airplane must have all placards required by the FARs and any additional placards that are required for safe operation if unusual design, handling or operational characteristics are present. In addition, each marking and placard:

1. Must be displayed in a conspicuous place.
2. Must not be easily erased, disfigured or obscured.

FAR 23.1543

Instrument Markings, General Requirements

1. When markings are on the cover glass, there must be a means to maintain the alignment of the glass with the dial (a slippage mark).
2. Each arc and line must be wide enough and located to be clearly visible to the pilot.

FAR 23.1545

Airspeed Indicator

The required markings are:

1. For the never exceed speed Vne, a red radial line.

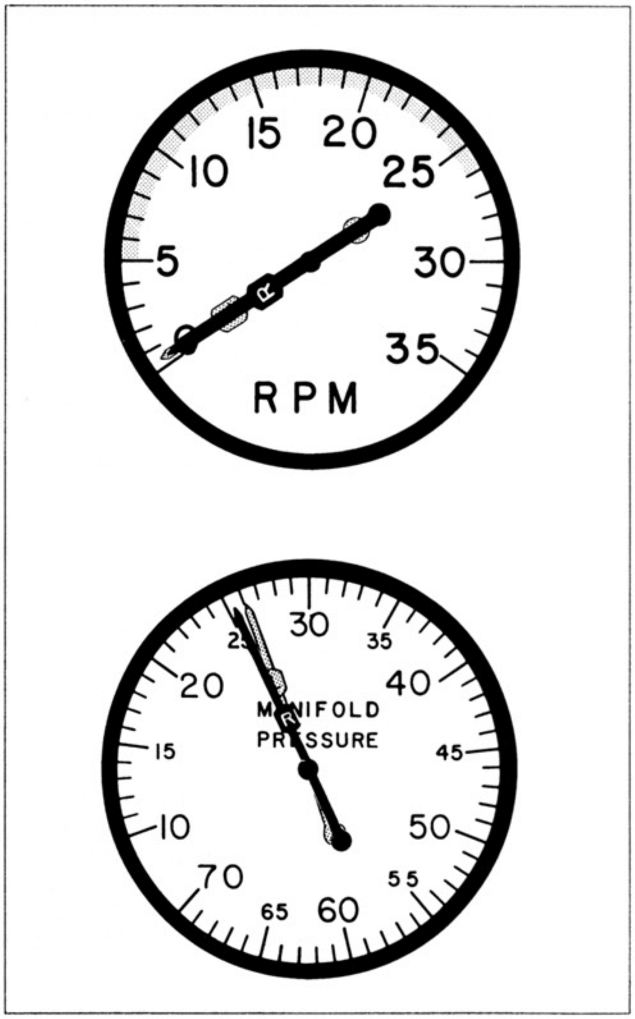

***Figure 1-15.* The manifold pressure gauge is used in conjunction with a tachometer to set power on a supercharged engine.**

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

2. For the caution range, a yellow arc extending from the red radial line in item 1 to the upper limit of the green arc in item 3.
3. For the normal operating range, a green arc with the lower limit at Vs1 (maximum weight, landing gear and flaps retracted) and the upper limit at Vno, maximum structural cruising speed.

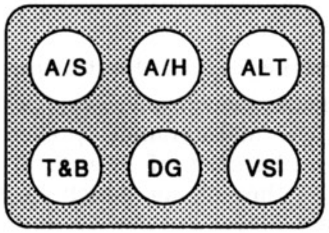

A/S = AIRSPEED INDICATOR
A/H = ARTIFICIAL HORIZON
ALT = ALTIMETER
DG = DIRECTIONAL GYRO
T&B = TURN AND BANK
VSI = VERTICAL SPEED INDICATOR

Figure 1-16. Most modern aircraft have the important flight instruments installed in a standard configuration directly in front of the pilot.

4. For the flap operating range, a white arc with the lower limit at Vso and the upper limit at Vfe.
5. For the one-engine-inoperative best rate of climb speed, Vy, a blue sector extending from the Vy speed at sea level to the Vy speed at 5,000 ft. (or higher altitude as appropriate).
6. For the minimum control speed (one-engine-inoperative) Vmc, a red radial line.

Note: Items 1 through 3 do not apply to an aircraft for which a maximum operating speed, Vmo/Mmo has been established. These aircraft would need a maximum allowable airspeed indicator.

FAR 23.1547

Magnetic Direction Indicator

1. A placard must be installed on or near the magnetic direction indicator to show deviation error corrections.
2. The placard must show the calibration in level flight with the engines running.

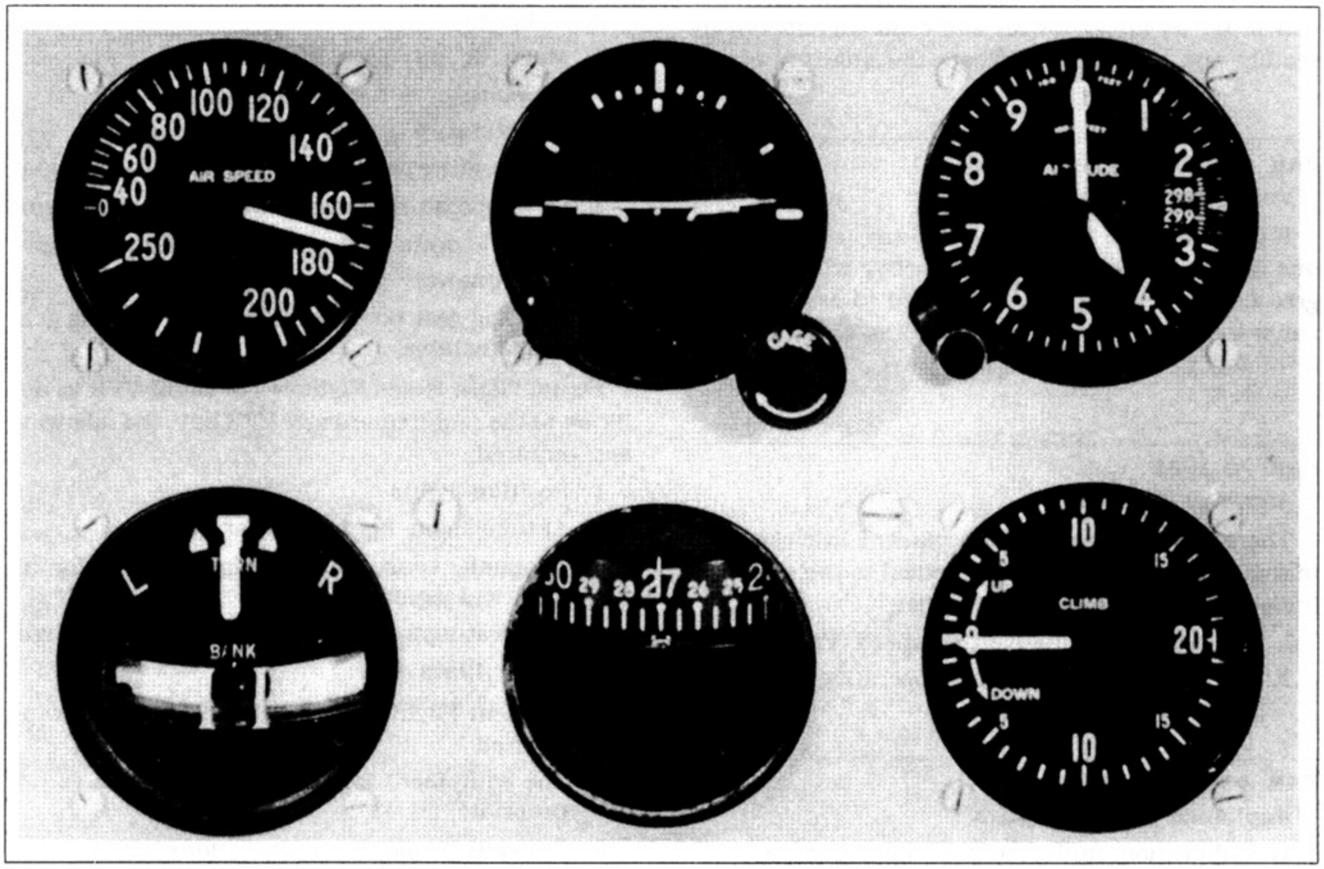

Figure 1-17. The types of instruments that might be found in the standard "T" configuration in a small airplane.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

3. The placard must state if calibration was made with radios on or off.
4. The placard must list the corrections for magnetic headings in increments of no more than 30°.
5. If the operation of electrical equipment can cause a deviation of more than 10°, the placard must identify that equipment.

FAR 23.1549

Powerplant Instruments

The required markings are:

1. Each maximum and minimum safe operating limit must be marked with a red radial line.
2. Each normal operating range must be marked with a green arc.
3. Each takeoff and precautionary range must be marked with a yellow arc.
4. Each engine or propeller range that is restricted because of vibration stresses must be marked with red arcs or red lines.

FAR 23.1551

Oil Quantity Indicator

It must be marked with sufficient increments to readily and accurately indicate the quantity of oil.

FAR 23.1553

Fuel Quantity Indicator

If the unusable fuel supply for any tank exceeds one gallon or 5% of tank capacity, whichever is greater, a red arc must be marked to extend from calibrated zero to the lowest reading obtainable in level flight.

FAR 23.1563

Airspeed Placards

There must be an airspeed placard in clear view of the pilot and as close as practical to the airspeed indicator. This placard must list:

1. The design maneuvering speed Va.
2. The maximum landing gear operating speed Vlo.

FAR 43 Appendix A

Appliance Major Repairs

1. Calibration and repair of instruments
2. Calibration of radio equipment

FAR 43 Appendix D

Scope and Detail of Items to Inspect For 100-Hour and Annual Inspections

(4) Instruments — inspect for poor condition, mounting, marking and where practical for improper operation.

FAR 65.81

General Privileges and Limitations

(a) A certificated mechanic may perform or supervise the maintenance, preventive maintenance or alteration of an aircraft or appliance or a part thereof, for which he is rated but excluding major repairs to and major alterations of propellers and any repair to or alteration of instruments.

FAR 91.9

Powered Civil Aircraft with Standard Category U.S. Airworthiness Certificates; Instrument and Equipment Requirements

Visual Flight Rules Day — For VFR, the following are required:

1. Airspeed indicator
2. Altimeter
3. Magnetic direction indicator
4. Tachometer for each engine
5. Oil pressure gauge for each engine
6. Temperature gauge for each liquid-cooled engine
7. Oil temperature gauge for each air-cooled engine
8. Manifold pressure gauge for each altitude engine
9. Fuel quantity gauge for each tank
10. Landing gear position indicator if landing gear is retractable

Visual Flight Rules Night — For night VFR in addition to the requirements for VFR day, the following are required:

1. Position lights
2. Anti-collision lights (red or white)
3. Adequate source of electrical energy for all electrical equipment
4. If operating for hire, one electric landing light
5. Spare fuses

Instrument Flight Rules — For IFR, the following are required

1. The equipment for VFR day and night as appropriate; PLUS
2. Two-way radio communications equipment
3. Radio navigation equipment

4. Gyroscopic rate-of-turn indicator (except aircraft with three attitude instruments)
5. Slip-skid indicator
6. Sensitive altimeter
7. Clock which displays hours, minutes and seconds
8. Generator or alternator
9. Gyroscopic bank and pitch indicator (artificial horizon)
10. Gyroscopic direction indicator (DG or equivalent)
11. For flight at and above 24,000 ft. MSL, a DME system
12. For Category II operations, additional equipment is required as listed in Appendix A.

A modern digital aircraft clock is shown in figure 1-18. The clock required for IFR does not have to be a digital clock, but it must display hours, minutes and seconds. The older type of aircraft clock was a conventional kind of round clock with three hands.

C. Pitot-Static System Instruments

1. Methods of Pressure Measurement

Pressure instruments are those which obtain readings by measuring the pressure at one or more places in terms of a liquid or gas pressure.

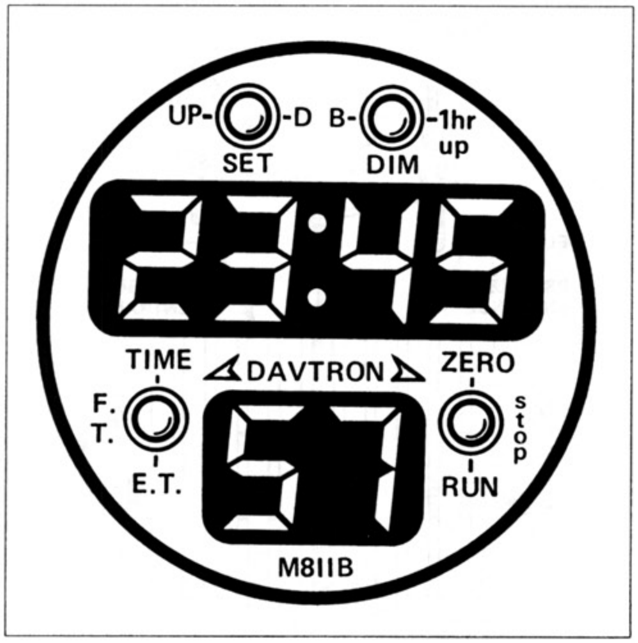

Figure 1-18. A digital aircraft clock able to display hours, minutes and seconds. (Courtesy Canadair Group, Bombardier Inc.

a. Pressure Measuring Sensors

There are three common types of sensors that can be used to mechanically measure a fluid pressure.

1) Bourdon Tube

A bourdon tube is a curved, hollow tube made of a springy metal (often a specially treated brass material) (figure 1-19).

One end of the tube is anchored and the other is free to move and fastened through linkages to the indicator pointer or similar device. As pressure is applied to the inside of the bourdon tube, it will try to straighten out as a result of the imbalance of forces on the walls of the curved tube. This motion is transmitted to the indicator needle or pointer through gears and linkages. The bourdon tube is used to measure relatively high pressure, typically 20 PSI or more.

2) Diaphragm

A diaphragm is a thin, lens shaped hollow metal container. It is used to measure relatively low pressures, as in an altimeter or airspeed indicator. Typically, it is made of a very thin, springy metal (beryllium copper is often used). One side is anchored and the other side is connected through gears and linkages to the pointer. See figure 1-20.

3) Bellows

A bellows is somewhat similar to a diaphragm, but it is longer and has accordion folds to typically allow a greater range of motion. It measures relatively

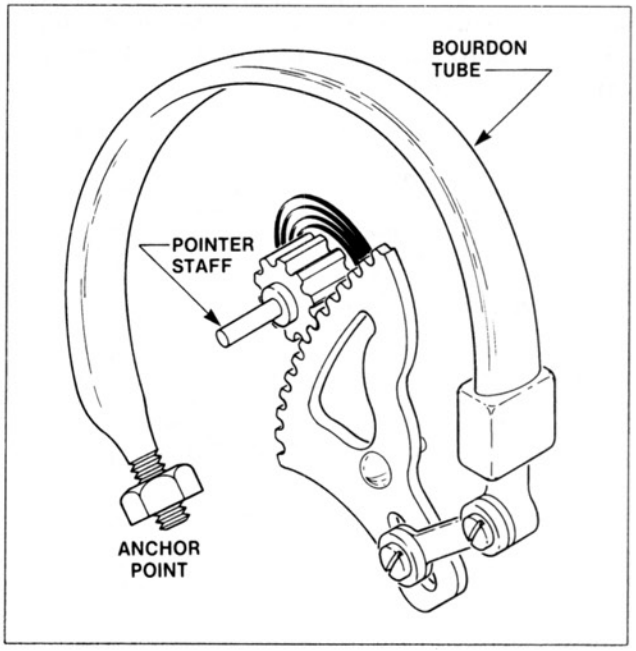

Figure 1-19. The bourdon tube pressure measuring device for instruments.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

low pressures and a common use is to measure differential pressure. It that case the bellows would be divided into two separate chambers with a different pressure source connected to each one as seen in figure 1-21.

b. Units of Measure for Pressure

Many different units of measure are used for pressure. In aviation, a number of units are in common use. Some examples are listed below.

1. Pounds per Square Inch — PSI
2. Inches of Mercury — in. Hg
3. Millibars — Mbar
4. Kilopascals — kPa

c. Types of Pressure Measurement

Pressure is measured compared to some reference value. Standard abbreviations are used to designate what this reference level is for a particular pressure measurement. The three common designations are:

1. Absolute Pressure — This is a pressure compared to a perfect vacuum. No pressure can be below a vacuum, so there is never a negative absolute pressure. Examples of this designation are PSIA and in. Hg absolute.
2. Gauge Pressure — This is a pressure compared to ambient pressure, usually outside atmospheric pressure, as in PSIG, etc.
3. Differential Pressure — This is used to designate two different pressures in the aircraft that are compared to each other, as in PSID.

Examples of these using aircraft instruments are the manifold pressure gauge which measures absolute pressure, the oil pressure gauge measures gauge pressure and the cabin differential pressure gauge is calibrated in PSID.

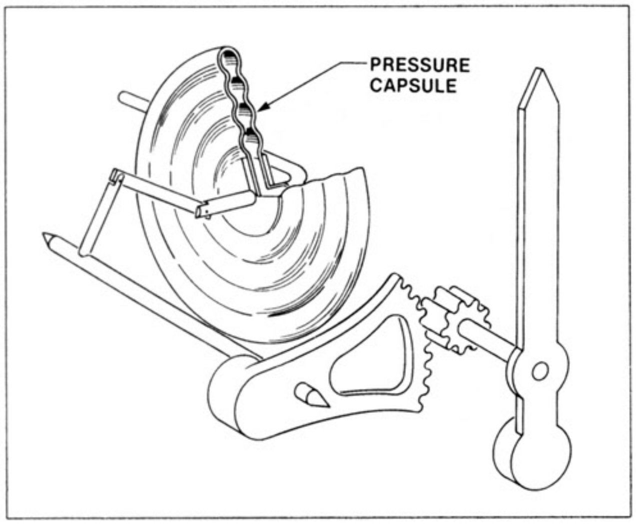

Figure 1-20. The diaphragm or aneroid pressure measuring device.

We will now examine the basic aircraft flight instruments that operate as pressure instruments. These are often referred to as the pitot-static instruments because they utilize pitot pressure and static pressure.

2. Standard Atmosphere Table

Altimeters and other instruments obtain readings by measurements of the pressure at various altitudes. The average or standard pressures and temperatures at different altitudes are contained in a table of the Standard Atmosphere. The values given in this table are accepted by virtually all countries of the world for scientific and aviation purposes (figure 1-22).

3. Altimeters

The altimeter is a basic required instrument for all certificated aircraft. It measures the atmospheric pressure and displays it as altitude in feet. This altitude is called mean sea level (MSL) since it is referenced to the average level of the major oceans. The measurement of altitude is based on the standard atmosphere table and the changes in pressure with altitude changes that it gives.

a. Non-Sensitive and Sensitive Altimeters

1) Non-Sensitive Altimeter

A non-sensitive altimeter has only one pointer and it makes one complete revolution for each 10,000 ft. in altitude. If the pointer was on the number 8,

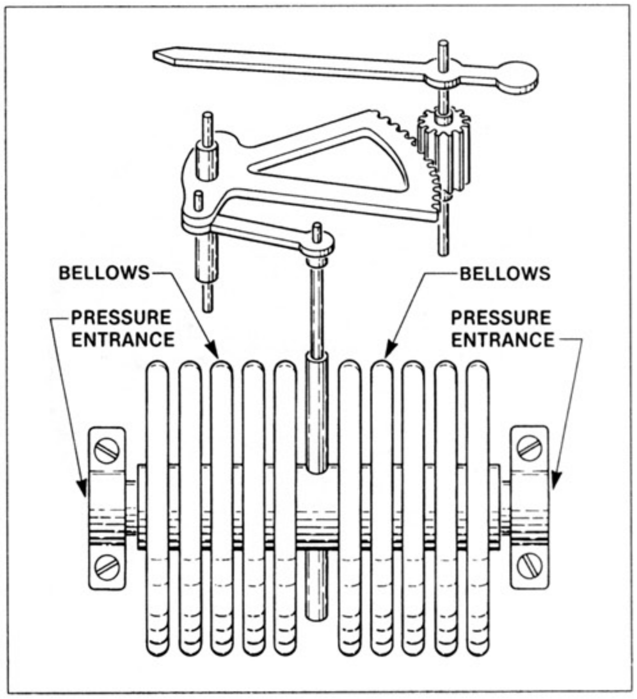

Figure 1-21. An example of a bellows being used to measure a differential pressure.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

that would represent 8,000 ft.; but it could also indicate 18,000 ft. For this reason, this type of altimeter is only used for small, simple aircraft that don't operate at high altitudes. A non-sensitive altimeter is illustrated in figure 1-23.

2) Sensitive Altimeter

A sensitive altimeter is more sensitive and easier to read than the non-sensitive type. A sensitive altimeter may have three separate pointers or a single pointer and a drum readout.

On an altimeter with three pointers, the longest pointer makes one revolution for each 1,000 ft., the second pointer makes one revolution for each 10,000 ft., and the shortest pointer makes one revolution for each 100,000 ft. The three pointer sensitive altimeter is shown in figure 1-24. A major problem with this type of altimeter is that the smallest pointer can be covered up by one of the other two pointers which makes it easy to misread the altimeter. The newer kind of three pointer altimeter

ALTITUDE FT.	DENSITY RATIO σ	PRESSURE RATIO δ	TEMPER-ATURE °F	SPEED OF SOUND α KNOTS
0	1.0000	1.0000	59.00	661.7
1000	0.9711	0.9644	55.43	659.5
2000	0.9428	0.9298	51.87	657.2
3000	0.9151	.08962	48.30	654.9
4000	0.8881	0.8637	44.74	652.6
5000	0.8617	0.8320	41.17	650.3
6000	0.8359	0.8014	37.60	647.9
7000	0.8106	0.7716	34.04	645.6
8000	.07860	0.7428	30.47	643.3
9000	0.7620	0.7148	26.90	640.9
10000	0.7385	0.6877	23.34	638.6
15000	0.6292	0.5643	5.51	626.7
20000	0.5328	0.4595	–12.32	614.6
25000	0.4481	0.3711	–30.15	602.2
30000	0.3741	0.2970	–47.98	589.5
35000	0.3099	0.2353	–65.82	576.6
*36089	0.2971	0.2234	–69.70	573.8
40000	0.2462	0.1851	–69.70	573.8
45000	0.1936	0.1455	–69.70	573.8
50000	0.1522	0.1145	–69.70	573.8
55000	0.1197	0.0900	–69.70	573.8
60000	0.0941	0.0708	–69.70	573.8
65000	0.0740	0.0557	–69.70	573.8
70000	0.0582	0.0438	–69.70	573.8
75000	0.0458	0.0344	–69.70	573.8
80000	0.0360	0.0271	–69.70	573.8
85000	0.0280	0.0213	–64.80	577.4
90000	0.0217	0.0168	–56.57	583.4
95000	0.0169	0.0134	–48.34	589.3
100000	0.0132	0.0107	–40.11	595.2

* GEOPOTENTIAL OF THE TROPOPAUSE

Figure 1-22. A reference table of the Standard Atmosphere.

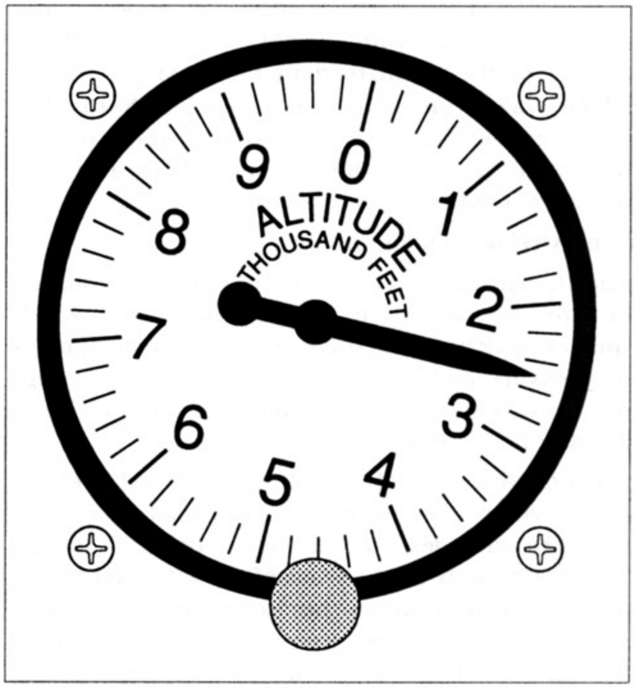

Figure 1-23. A simple non-sensitive aircraft altimeter. (Courtesy Piper Aircraft Corp.)

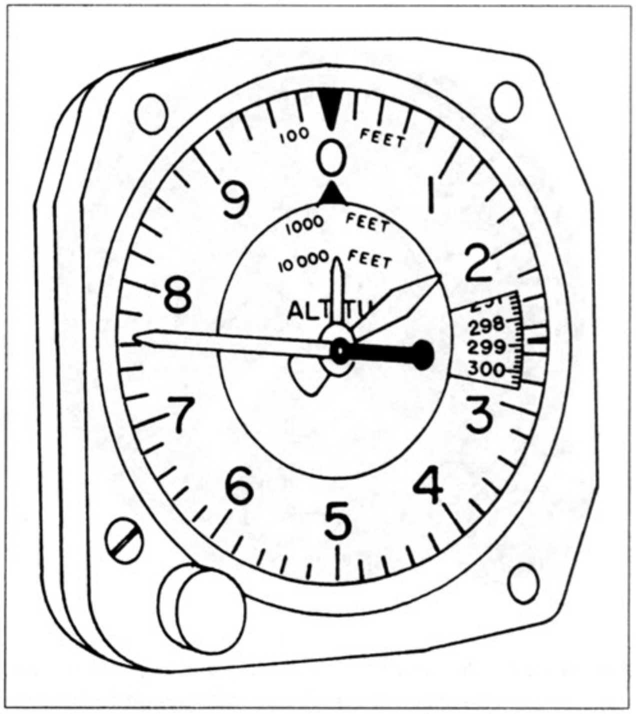

Figure 1-24. The older style three-pointer sensitive altimeter.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

seen in figure 1-25 has been modified to make it easier to read. The pointer that makes one revolution in 100,000 ft. has been extended to the outside of the face with a small triangle on the end so it cannot be covered up. In addition, a small window shows a striped pattern below about 15,000 ft. (at higher altitudes the striped symbol disappears). The sensitive altimeter with one pointer as seen in figure 1-26 uses the drum readout and a single pointer which makes one revolution for each 1,000 ft. The drum readout may have three or five digits displayed, but in any case the combination of the two permits accurate reading of the altimeter.

b. Altimeter Mechanism

The internal mechanism of the altimeter consists of three diaphragms in series. This creates an altimeter which is sensitive to very small changes in pressure and which has a large enough range of measurement to read altitudes of 35,000 to 50,000 ft. or more.

Inside the case of the altimeter is a bimetallic device which gives automatic compensation for temperature changes to ensure accurate readings. There is a knob on the front of the altimeter which operates a mechanism that compensates for nonstandard atmospheric pressure. The standard atmosphere table gives the standard pressure at sea level as 29.92 in. Hg, but the actual atmosphere pressure in any given location seldom matches the values in the table. Weather systems with higher or lower pressure are constantly moving across the surface of the earth.

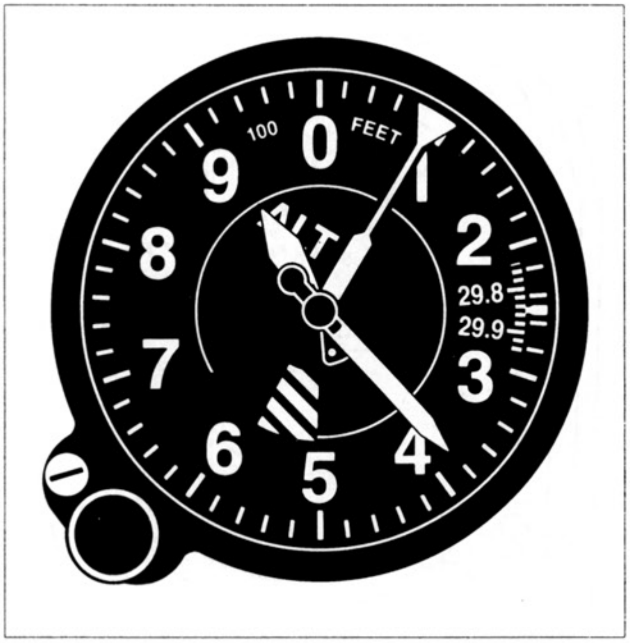

Figure 1-25. The newer style three-pointer sensitive altimeter uses a different pointer and a striped symbol that is visible below about 15,000 ft. as aids to make reading the altimeter easier.

It is due to these variations in pressure that the knob on the front of the altimeter must be adjusted to the current barometric pressure setting for the area in which it is operating. There is a small window on the altimeter which displays the current barometric pressure setting. It is commonly called the Kollsman window. If for example a pilot received an altimeter setting of 30.01 from an air traffic controller, that number would be set into the Kollsman window of the altimeter. When the pilot sets the altimeter to the correct setting before takeoff, the indication on the altimeter will show the height of that airport above sea level. Outside the United Sates, a different unit of measurement is often used for barometric pressure settings. This unit of measurement is millibars (MB) and the altimeters found in many modern aircraft have two Kollsman windows: one for inches of mercury and one for millibars. Figure 1-26 shows the two windows and the barometric pressure setting knob.

c. Altimeter Operation

The diaphragms of an altimeter are sealed at the factory. The case of the altimeter is connected to the static ports on the airplane. The static pressure outside the airplane is conducted to the instrument case by tubing and hose. If the airplane climbs, the reduced pressure in the case causes the diaphragms to expand and move the pointers to indicate a higher altitude.

The altimeter uses three diaphragms in series to increase the range of measurement. The mechanism uses very delicate gearing which is designed in such a way that a very small movement of the diaphragm causes a large motion of the pointer or pointers. Under certain circumstances it has a tendency to stick and some aircraft have a vibrator on the instrument panel or in the altimeter to prevent this sticking.

d. Encoding Altimeter

An altimeter related device which is found on many aircraft is the encoding altimeter system. The purpose of this system is to send information concerning the aircraft's altitude to a radar system on the ground so that it can be displayed on a radar scope. The data is transmitted to the ground using a special aircraft radio called a transponder. An encoding altimeter system (also known as Mode C), supplies the electrical signal to the transponder that contains this altitude data. An aircraft encoding

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

altimeter and transponder are required for three different conditions:

1. Flight operations within a TCA (terminal control area) or Class B airspace.
2. Flight operations within an ARSA (airport radar service area) or Class C airspace.
3. Flight operations above 10,000 ft. MSL and over 2500 ft. AGL.

The connections for the encoding feature of an altimeter and an altimeter vibrator can be seen in figure 1-27.

4. Airspeed Indicators

Airspeed indicators are required on all certificated aircraft except free balloons. This instrument gives the pilot an indication of his speed through the air, it does not measure groundspeed. The airspeed indicator is a pitot-static system instrument that is connected to both the pitot pressure source and the static pressure source. It measures the difference between these two pressures as indicated airspeed.

On aircraft operated in the United States, the unit of measurement for airspeed indicators is nautical miles per hour (knots), statute miles per hour (MPH) or both on the same instrument.

a. Types of Airspeed

There are three types of airspeed associated with the airspeed indicator. This is because it is subject to a number of errors or inaccuracies. The three types of airspeed are:

1. Indicated Airspeed — This is the reading on the instrument
2. Calibrated Airspeed — This is indicated airspeed which has been corrected for position and instrument error. The pitot tube is mounted at a fixed angle to the longitudinal axis of the aircraft and at slow speeds and high angles of attack there will be an error. The pilot determines the correction for calibrated airspeed by consulting a table in the Pilot's Operating Handbook, an example of which is shown in figure 1-28.
3. True Airspeed — This is calibrated airspeed which has been corrected for altitude and temperature effects. At high altitudes the indicated airspeed will be much less than the actual speed through the air which is true airspeed. For example, at 41,000 ft. the indicated airspeed is only about half the true airspeed. At sea level on a standard day with the aircraft in level cruise flight, indicated and true airspeed are usually the same.

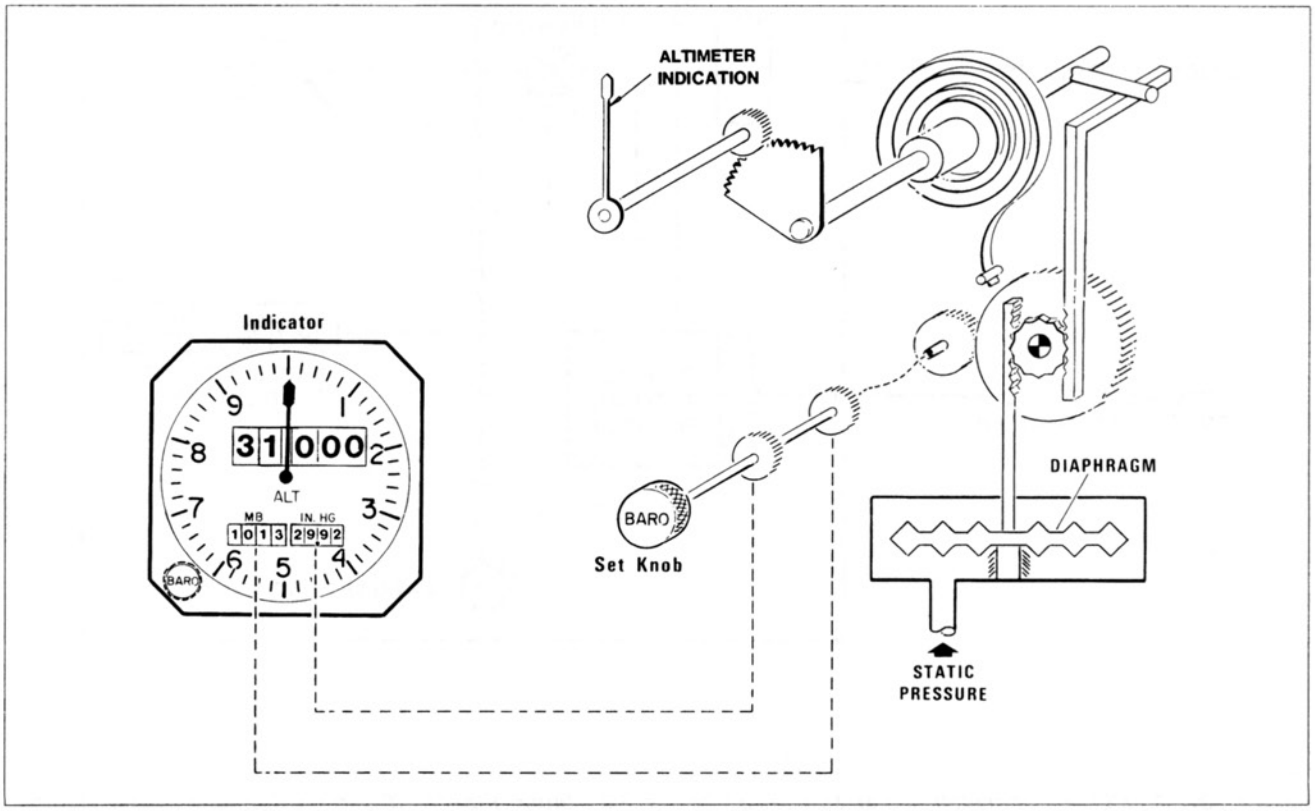

Figure 1-26. A sensitive altimeter with drum readout showing the static connection to the outside of the diaphragm.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

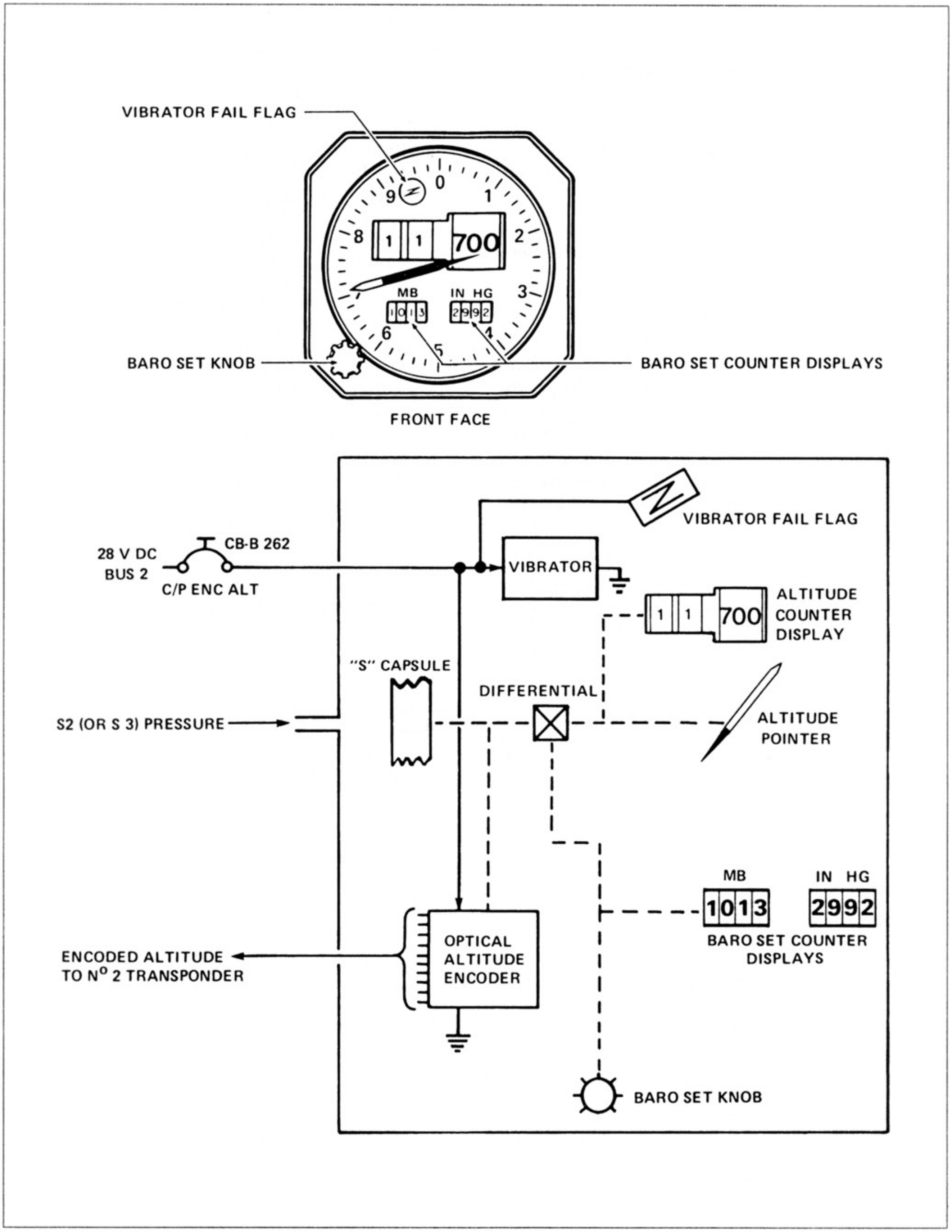
VIBRATOR FAIL FLAG
BARO SET KNOB
BARO SET COUNTER DISPLAYS
FRONT FACE
MB
IN HG
VIBRATOR FAIL FLAG
28 V DC
BUS 2
CB-B 262
C/P ENC ALT
VIBRATOR
ALTITUDE COUNTER DISPLAY
"S" CAPSULE
DIFFERENTIAL
S2 (OR S 3) PRESSURE
ALTITUDE POINTER
MB
IN HG
1013
2992
BARO SET COUNTER DISPLAYS
OPTICAL ALTITUDE ENCODER
ENCODED ALTITUDE TO N° 2 TRANSPONDER
BARO SET KNOB

Figure 1-27. An altimeter installation that includes a vibrator and altitude encoder. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

b. Airspeed Indicator Operation

The airspeed indicator uses a diaphragm that measures the differential pressure between pitot pressure and static pressure. Pitot pressure is connected to the inside of the diaphragm and static pressure to the outside as shown in figure 1-29. A pitot tube has an opening facing forward into the relative wind so that it measures ram pressure effects. The static ports simply measure outside atmospheric pressure. The faster the airplane goes, the greater will be the difference in pressure.

c. Airspeed Indicator Markings

The operating limitations airspeeds that would be found on a general aviation twin-engine airplane are listed below.

FLAPS UP								
KIAS	80	100	120	140	160	180	200	210
KCAS	81	101	120	139	159	177	196	206
FLAPS 10°								
KIAS	70	80	90	100	120	140	160	175
KCAS	72	81	91	100	119	138	157	171
FLAPS 30°								
KIAS	60	70	80	90	100	110	125	- - -
KCAS	64	72	80	89	98	108	123	- - -
Note: For illustration only; not to be used for flight planning								

Figure 1-28. The table of calibrated airspeed corrections for a twin-engine airplane. (Courtesy Cessna Aircraft Co.)

1. Vso — Stall speed with gear and flaps down. This is shown by the bottom of the white arc.
2. Vs1 — Stall speed with gear and flaps up. The bottom of the green arc.
3. Vmc — Minimum control speed with critical engine failed. A red radial line in the lower range of indicated airspeeds.
4. Vfe — Maximum speed with flaps down. Top of white arc.
5. Vno — Maximum structural cruise. Top of green arc or bottom of yellow arc.
6. Vne — Never exceed speed. a red radial line at the highest permitted airspeed, also the top of the yellow arc.
7. Vyse — Best rate of climb speed with one engine inoperative. A blue line or sector.

The aircraft should only be operated in the yellow arc range with caution and in smooth air. Figure 1-30 shows the colored markings for the airspeed indicator on a single-engine airplane. A twin-engine airplane would have the additional markings for Vmc and Vyse. The principle of critical engine and Vmc is illustrated in figure 1-31. The critical engine is the engine whose failure would produce the most adverse effect on the handling characteristics of the airplane. Since the flight controls become more effective at higher speeds, the concept of Vmc involves high engine power and slow speeds such as occurs during takeoff. During initial climb after takeoff, the thrust from the propellers is offset to the right side as shown in figure 1-31 if both

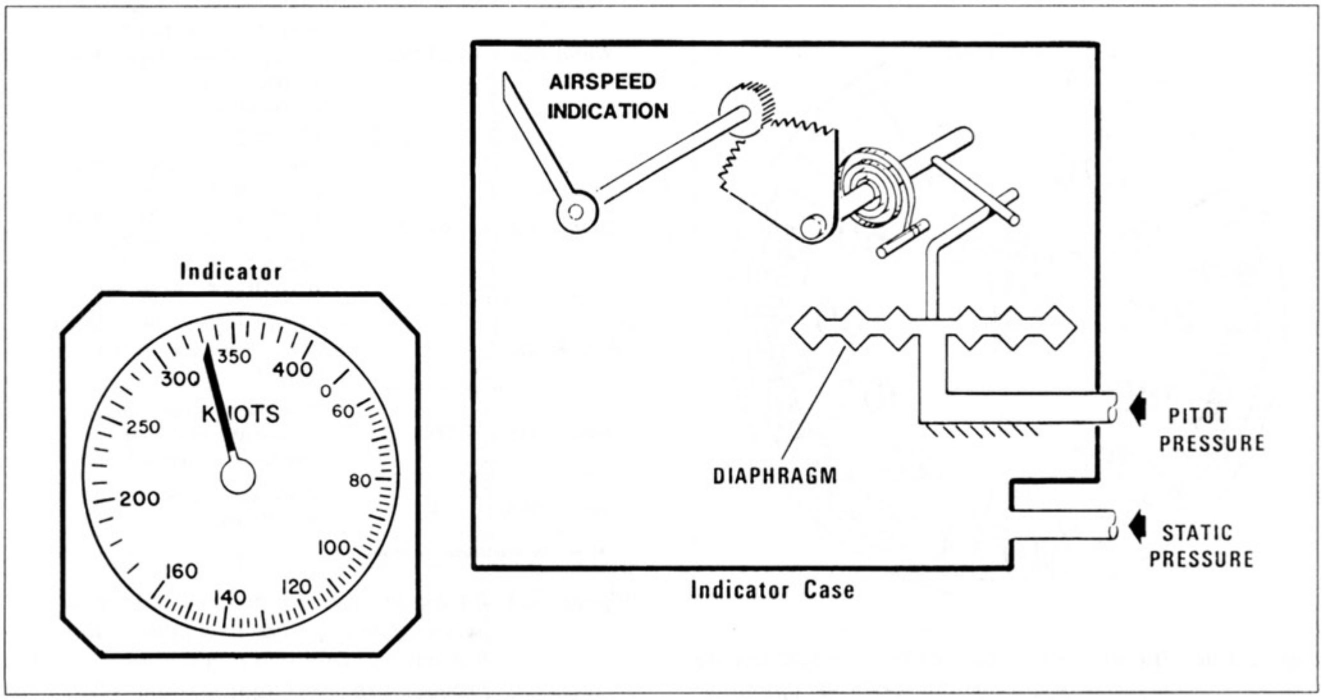

Figure 1-29. The two pressure connections to the diaphragm in an airspeed indicator.

propellers rotate clockwise. In this case, the critical engine is the left engine since the thrust line for the right engine is farther from the longitudinal axis of the airplane. Some twin engine airplanes are designed so that the propellers rotate in opposite directions. In that case, the two engines are equally critical. The Vmc speed would be the same no matter which engine failed. The reason that the Vmc marking is red is that maintaining a speed above Vmc can be very important to prevent loss of control. If an engine fails after takeoff and the airplane is below Vmc speed, the airplane will start to turn and roll and the motion cannot be stopped with full opposite control inputs. This has caused many accidents and many deaths. Vmc is a critical operating limitation that the pilot must keep in mind. The speed for Vyse is obviously higher than Vmc, it is the speed the pilot would use after an engine failure to get the best rate of climb possible.

A table of airspeed indicator operating limitations markings for a small twin engine airplane is shown in figure 1-32.

All of the important operating limitations airspeeds will not be shown on the airspeed indicator. Some of them will be displayed on a placard next to the airspeed indicator like that seen in figure 1-33 and others will only be found in the Pilot's Operating Handbook or Airplane Flight Manual.

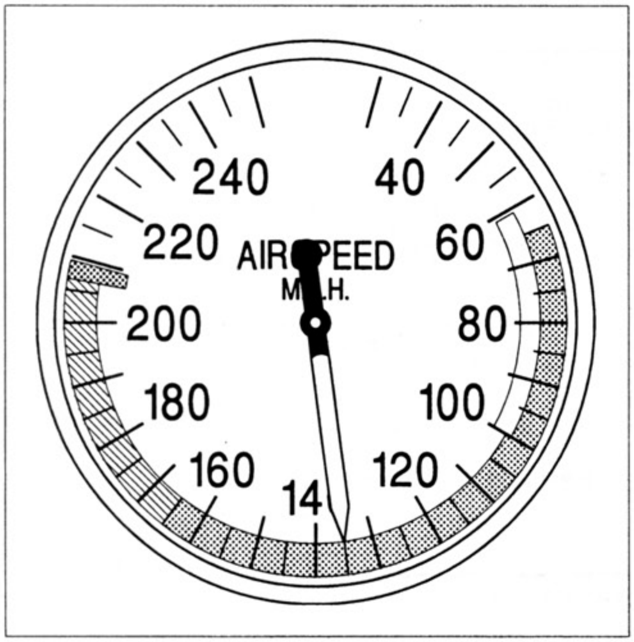

Figure 1-30. The airspeed indicator for a single engine airplane showing the colored operating limitations markings.

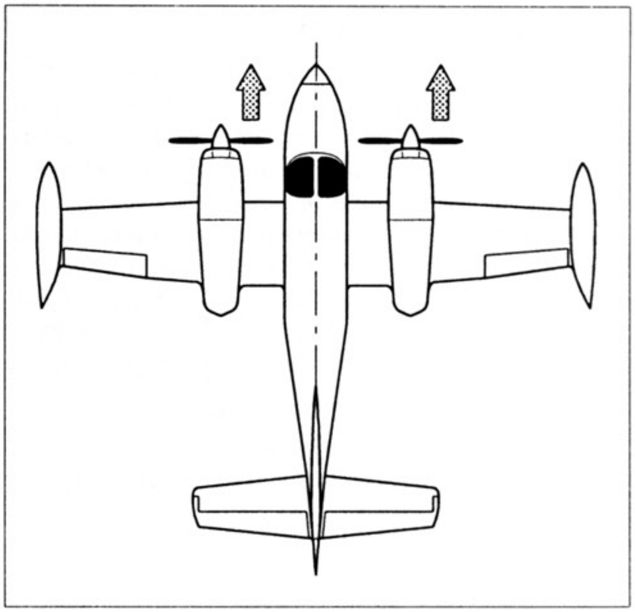

Figure 1-31. An illustration of why the left engine is the critical engine when both propellers turn clockwise. At slow speeds, the thrust lines from the propellers are offset to the right which places the thrust line of the right engine farther from the center line of the aircraft. Vmc is the slowest speed at which control can be maintained when the critical engine fails.

MARKING	KIAS VALUE OR RANGE	SIGNIFICANCE
Red Radial	65	Minimum Control Speed
White Arc	58-125	Full Flap Operating Range. Lower limit is maximum weight V_{S_O} in landing configuration. Upper limit is maximum speed permissible with flaps extended.
Green Arc	66-175	Normal Operating Range. Lower limit is maximum weight V_S at most forward C.G. with flaps retracted. Upper limit is maximum structural cruising speed.
Blue Radial	97	Single Engine Best Rate-of-Climb Speed at Maximum Weight.
Yellow Arc	175-210	Operations must be conducted with caution and only in smooth air.
Red Radial	210	Maximum speed for all operations.

Note: For illustration purposes only; not to be used for flight planning.

Figure 1-32. A table of airspeed indicator markings for a twin engine airplane. This table is found in the Pilot's Operating Handbook *or* Approved Airplane Flight Manual. *(Courtesy Cessna Aircraft Co.)*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

d. Maximum Allowable Airspeed Indicator

On high speed aircraft that operate at high altitudes, it is not practical to use the colored markings on the airspeed indicator because the values change significantly as the aircraft climbs to higher altitudes.

These aircraft would use a different type of airspeed indicator called the maximum allowable airspeed indicator. It is found on turboprop airplanes, bizjets and air carrier jets.

The maximum allowable airspeed indicator has two pointers; one gives the normal indicated airspeed and the second one, which has a striped or checked pattern, is connected to a separate diaphragm and it shows the Vmo speed at the existing altitude. The two pointers can be seen in figure 1-34. On a jet airplane, the maximum allowable airspeed indicator will also usually include a Mach number display as shown in figure 1-34.

e. Machmeter

A machmeter is required for jet airplanes because they will have a maximum safe operating speed expressed in Mach numbers. This is in addition to the maximum safe indicated airspeed. This is necessary because if the safe Mach number is exceeded, the airplane can become difficult or impossible to control. Mach number indicates the ratio of the aircraft's true airspeed compared to the speed of sound. A speed of Mach 0.8 means the aircraft is traveling at 80% of the speed of sound. The speed of sound is about 760 MPH at sea level, but it is affected by temperature. As the temperature decreases, the speed of sound is also decreased. Since temperature decreases with an increase in altitude, the speed of sound is reduced at higher altitudes. At 35,000 ft. the speed of sound would be about 660 MPH. If an airplane climbed from sea level to 35,000 ft. at a constant true airspeed, the Mach number would steadily increase. Most civilian jet airplanes have an Mmo (maximum operating Mach number) of about 0.8 to 0.87. As an airplane approaches its critical Mach number, a shock wave begins to form on the upper surface of the wing as seen in figure 1-35. This will occur at flight speeds below Mach 1 because the air accelerates as it moves over the top of the wing. This shock wave disrupts airflow and causes the center of pressure to shift aft. As aircraft speed gets closer to Mmo, the shock wave gets stronger until a point is reached where control of the aircraft may be lost. A loss of control at high speed and high altitude can and has resulted in the loss of the aircraft—it is an important operating limitation for jet airplanes. The machmeter instrument has a diaphragm like a normal airspeed indicator that is connected to pitot and static pressure. It also has an additional diaphragm which modifies the movement of the pointer to compensate for the effects of altitude and temperature (figure 1-36).

MAX SPEED KIAS

MANEUVERING	**148**
EXTEND GEAR	**175**
RETRACT GEAR	**150**
GEAR DOWN	**210**

Note: For illustration purposes only; not to be used for flight planning.

Figure 1-33. The airspeed limitations placard that would be installed near the airspeed indicator for a twin engine airplane. (Courtesy Cessna Aircraft Co.)

5. Vertical Speed Indicator

The last of the three basic pitot-static system instruments is identified by several names: a vertical speed indicator (VSI), a rate of climb indicator (ROC)

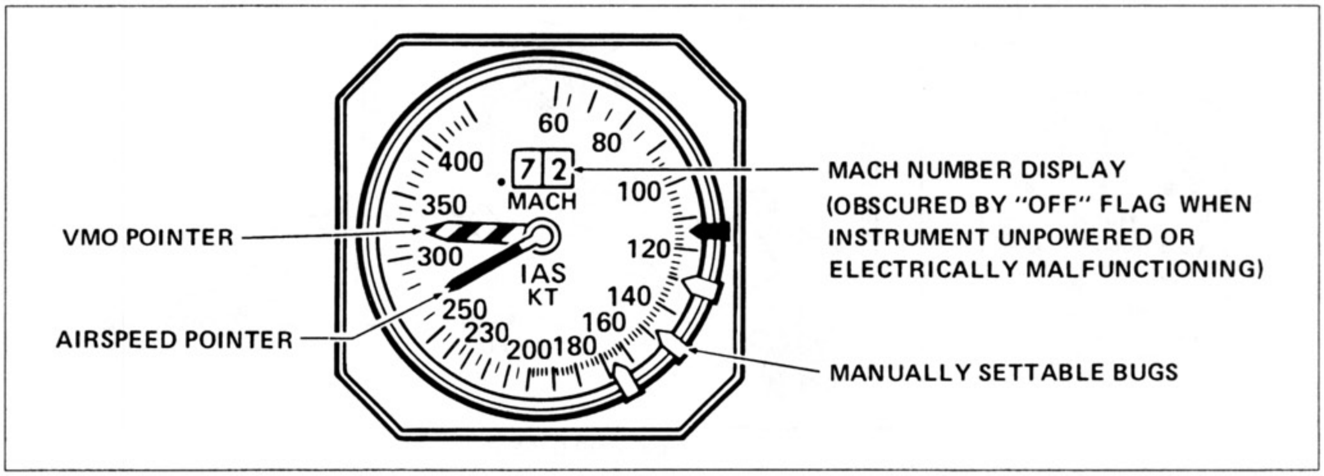

Figure 1-34. The combined maximum allowable airspeed indicator and Machmeter for a corporate jet airplane. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

and a vertical velocity indicator (VVI). This instrument measures the rate of change of static pressure. Since static pressure decreases with an increase in altitude, it can measure the rate of climb or descent of the aircraft. The unit of measurement is feet per minute (FPM) as seen on the instrument face in figure 1-37.

a. Basic Operation of VSI

Static pressure is connected directly to the inside of the diaphragm, but it is connected to the case pressure (outside of the diaphragm) by a small orifice or restrictor opening as indicated in figure 1-38. If the airplane climbs, the pressure will decrease more rapidly inside the diaphragm than outside of it. The diaphragm will be compressed and that will move the pointer to show the rate of change of altitude in feet per minute. The opposite occurs during descents. When the aircraft levels off, the pressures equalize and the pointer returns to zero. The instrument case usually includes an overpressure valve to prevent damage at rates of climb or descent in excess of the maximum reading for that instrument.

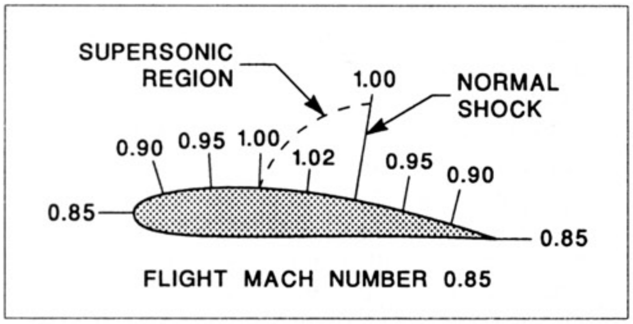

Figure 1-35. A shock wave will form on the upper surface of the wing when the airplane is traveling at less than the speed of sound.

Because of the very small size of the restrictor opening, the pointer tends to react slowly to rapid changes in altitude. This problem is reduced in the instrument called an Instantaneous Vertical Speed Indicator (IVSI).

b. Instantaneous Vertical Speed Indicator

The IVSI has two little cylinders with pistons and springs which can be seen in figure 1-39. They are called accelerometer operated dashpots. When the aircraft pitches up or down suddenly, the pistons will move and force air into or out of the diaphragm to get the pointer moving quickly. In a steady rate climb or descent, they will have no effect on the pointer.

c. Variometer

The variometer is a very sensitive version of the VSI which is used by gliders. The mechanism uses a movable vane in a small air tight metal container.

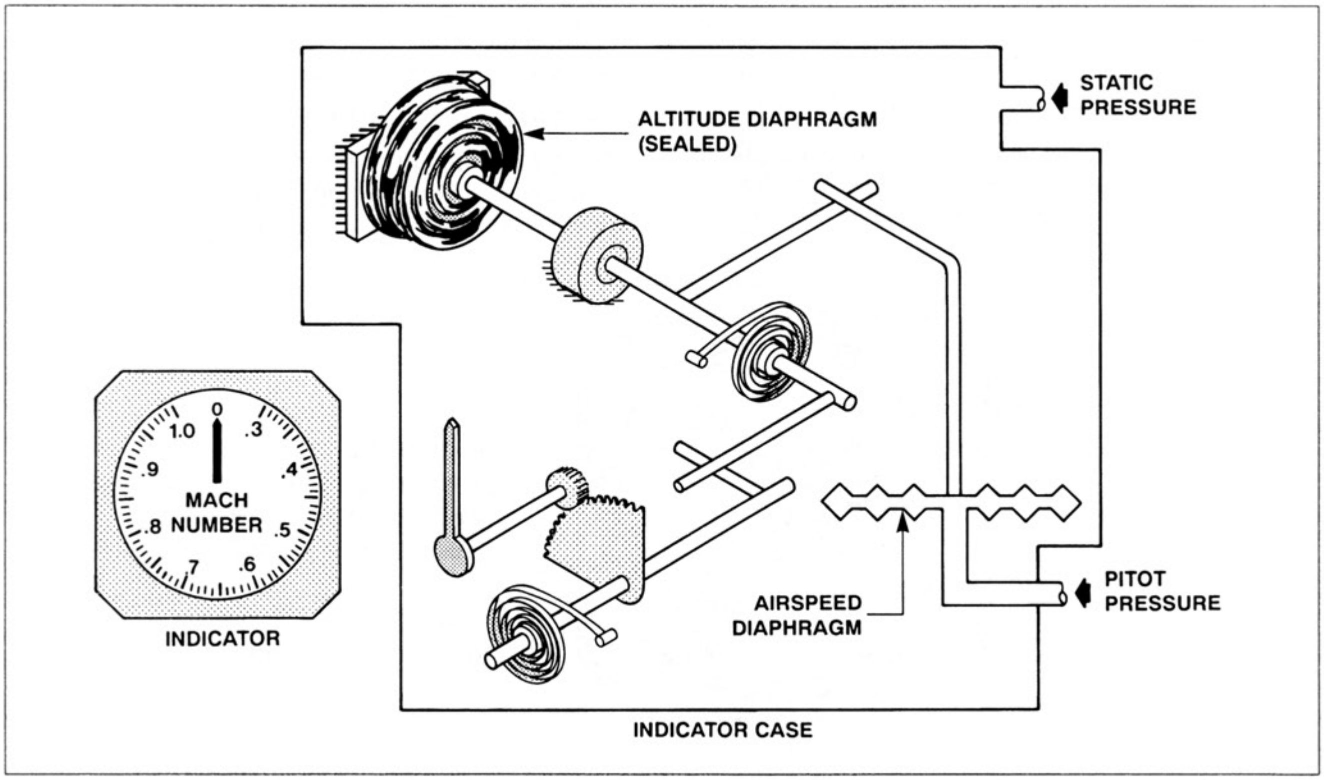

Figure 1-36. The Machmeter has an extra diaphragm to compensate for altitude effects.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The vane divides the container into two chambers. One chamber is connected to the static port and the other is connected to an air bottle. The air storage bottle is used because it will hold more air than a small instrument case and this improves sensitivity. With the variometer, even small changes in rate of climb or descent can be detected.

6. Accelerometer

The accelerometer is an instrument that measures the "G" forces or acceleration forces on an aircraft in flight. The common application is in an aerobatic airplane where the pilot needs to know how much "G" force the airplane is being subjected to in order to prevent overstress of the structure.

The mechanism of the accelerometer consists of a weight which is connected by a cord and pulleys to the shaft that operates the pointer. The internal arrangement of an accelerometer is shown in figure 1-40. The weight is supported by a guide shaft which only allows it to move up and down relative to the guide shaft. A positive G acceleration will cause the weight to move downward and rotate the pointer to show a higher positive G loading. There is a balance spring on the pointer shaft pulley to balance the forces. The instrument is installed in the airplane so that it measures acceleration along the vertical axis of the airplane. The normal at rest indication on the ground or in level flight is +1 Gs. The instrument face of an accelerometer is shown in figure 1-41. The instrument has three pointers connected to the operating mechanism. One pointer gives a readout of the current acceleration force along the vertical axis. The other two pointers have a ratchet device so that they will remain at the highest reading recorded for positive and negative forces. A knob is included on the instrument to reset the two recording pointers.

7. Pitot-Static Systems

a. Pitot and Static Ports

The system of ports and tubing on the aircraft which supplies pitot and static pressure for the instruments is called the pitot-static system. The pitot tube is an open tube which faces forward into the relative wind in flight. It measures the ram pressure of the airstream. On small airplanes, the pitot tube is usually installed below the wing. On other aircraft, it is installed on the nose section of the aircraft.

The static ports are openings at right angles to the relative wind so that they will measure static pressure and not be affected by the speed of the aircraft. The static ports are most often located in pairs along the sides of the fuselage. On some aircraft, the static ports are along the sides or top and bottom surfaces of the pitot tube so that both pressures are measured with the same probe. A pitot tube which includes static ports and electric heaters to prevent icing is shown in figure 1-42.

With the static ports in pairs on opposite sides of the fuselage, any errors caused by sideslip will be eliminated. Aircraft that must operate in adverse weather conditions will require an electrical heating system for the pitot tubes and static ports to prevent icing. Air carrier jets and similar types of aircraft usually employ multiple pitot tubes and static ports for safety. A typical arrangement of this kind with three pitot tubes and three sets of static ports is illustrated in figure 1-43. Notice that the instruments on the left and right sides of the cockpit

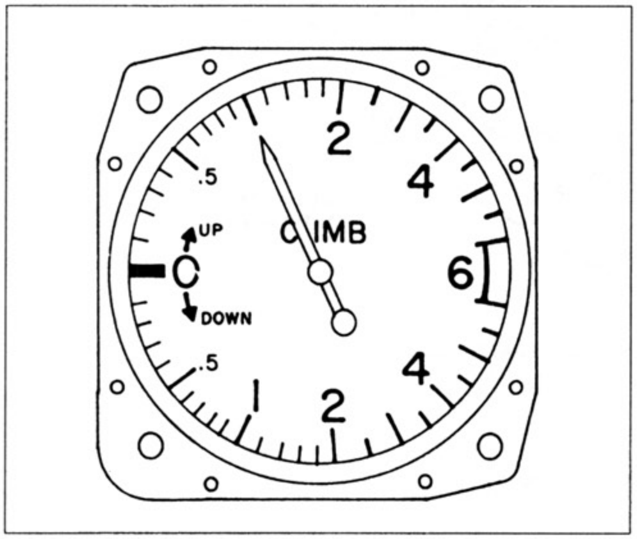

Figure 1-37. The vertical speed indicator has two scales: one for climbs and one for descents.

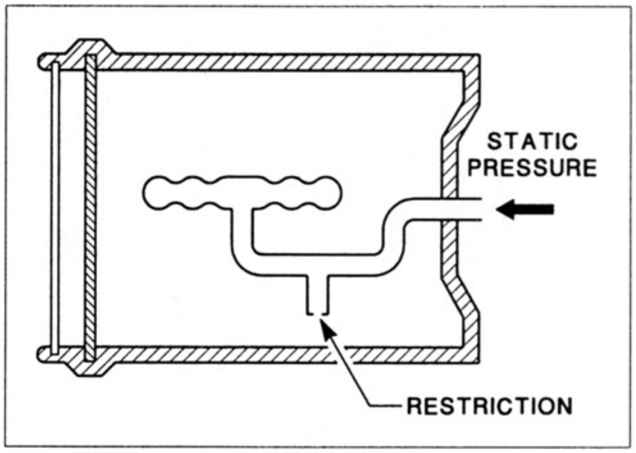

Figure 1-38. The VSI has a restrictor in the connection to the case which causes a difference between diaphragm and case pressure during climbs and descents.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

are operated by totally separate pressure sources. By routinely cross checking the instrument indications from the left and right sides, a faulty indication can be identified.

b. Blockage of Ports

When pitot-static lines or ports become blocked by ice or other factors, it can cause the instruments to give improper readings. If the static pressure is blocked, the altimeter will remain at the current indication and the VSI will continue to read zero even when the aircraft climbs or descends. This problem would usually be noticed by the pilot. The airspeed indicator uses both pitot and static pressure and a blockage in flight could be more difficult to detect.

Let's use the example of an airplane that is flying at 10,000 ft. when the static ports become blocked. The pressure in the static system will be sealed in and won't change when the aircraft climbs or descends. If the airplane climbs at a constant speed, the airspeed indicator will show a decreasing airspeed. If the airplane descends, the airspeed indicator will indicate a higher than actual airspeed. Just the opposite would happen if the pitot tube iced over or was blocked in some other way in flight. A climb would cause the airspeed indicator to read higher than it should and a descent would cause it to read lower than actual airspeed.

The lines and connections in a pitot-static system should be maintained in good condition. Even though they don't have to handle high pressures, the instruments are very sensitive to small changes in pressure so that even very small leaks can cause errors in the instruments.

The tubing and hoses that are used are not very strong and should be inspected carefully for damage. The fittings and connections should be installed with care and torqued to specified values.

c. Altimeter System Tests and Inspections

There are some FAR requirements for testing of altimeter systems. These will be covered next along

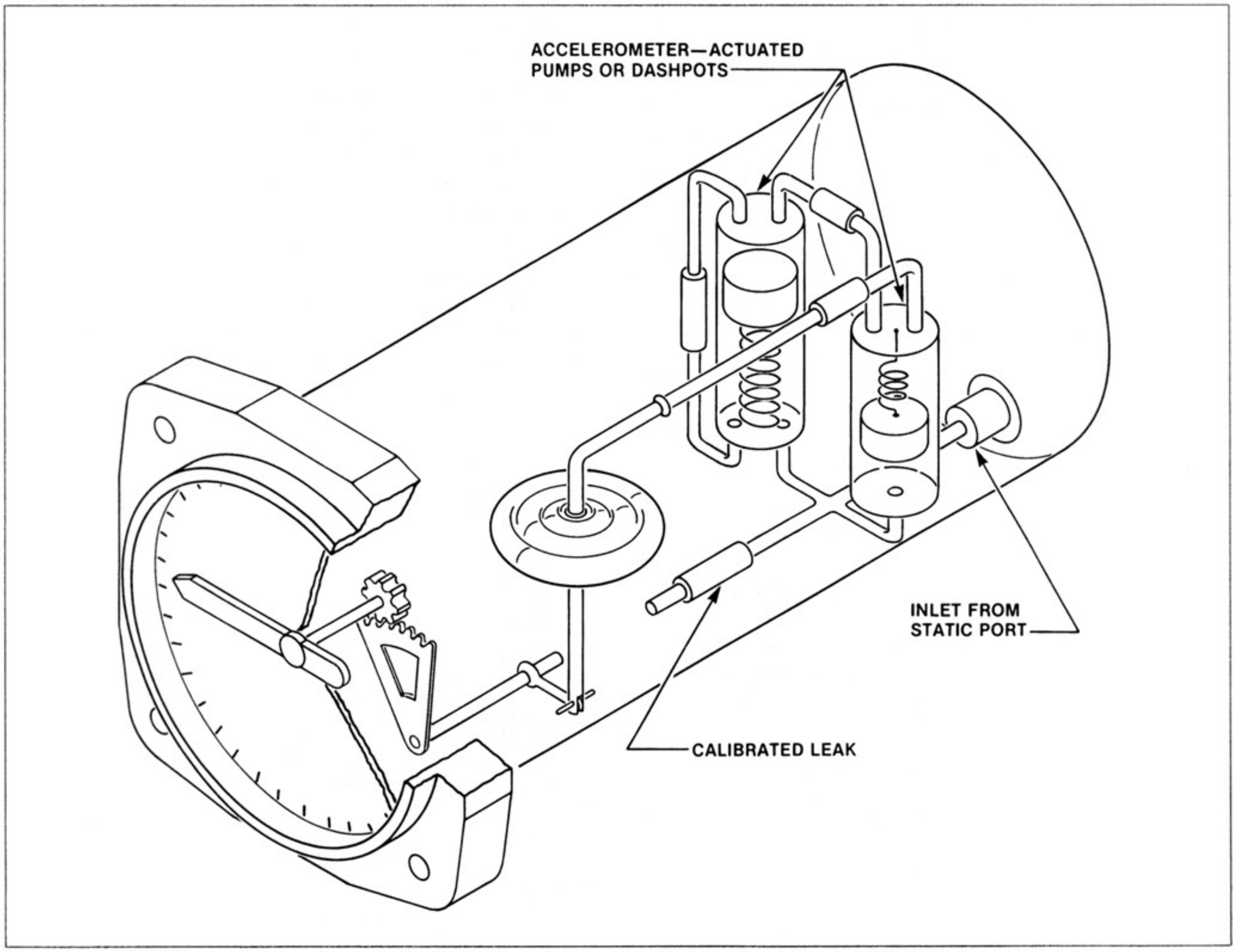

Figure 1-39. The acceleration actuated dashpots in the IVSI reduce the lag in pointer movement.

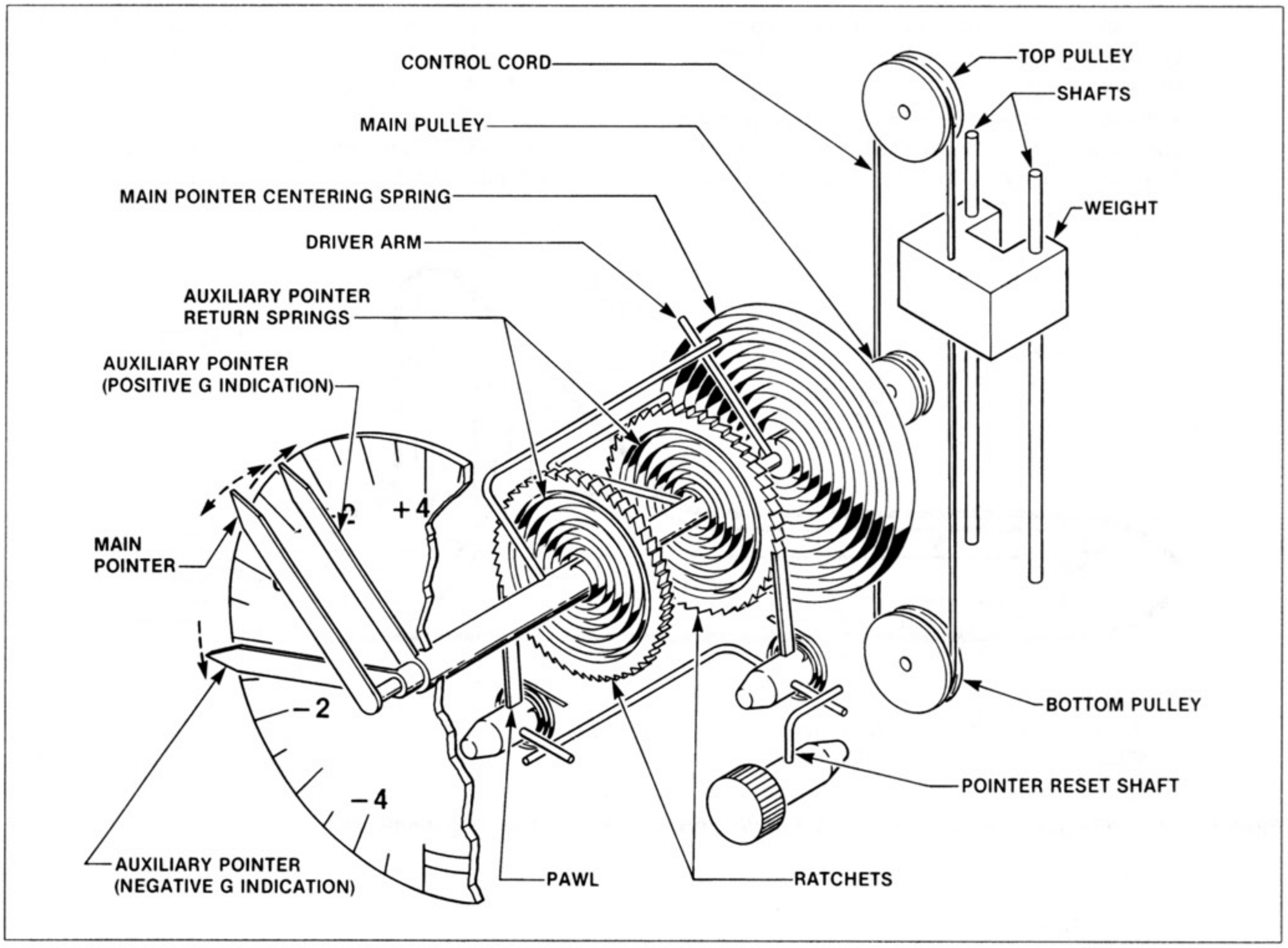
CONTROL CORD
MAIN PULLEY
MAIN POINTER CENTERING SPRING
DRIVER ARM
AUXILIARY POINTER RETURN SPRINGS
AUXILIARY POINTER (POSITIVE G INDICATION)
MAIN POINTER
AUXILIARY POINTER (NEGATIVE G INDICATION)
PAWL
RATCHETS
TOP PULLEY
SHAFTS
WEIGHT
BOTTOM PULLEY
POINTER RESET SHAFT
+2
+4
−2
−4

Figure 1-40. The internal mechanism of a three-pointer accelerometer.

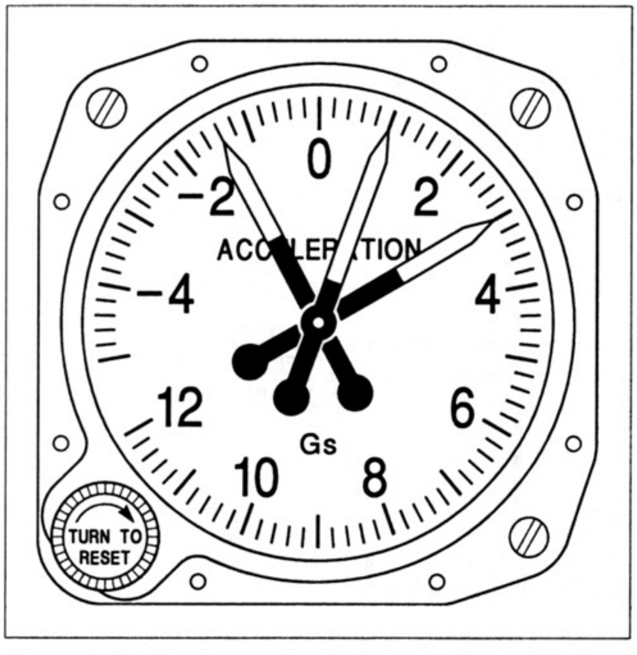
0
2
4
6
8
10
12
−4
−2
ACCELERATION
Gs
TURN TO RESET

Figure 1-41. Accelerometer instrument face with three pointers.

with the procedures for other tests which might not be required at specific intervals but would be conducted whenever a problem was suspected. A portable pitot-static system test set that could be used for these tests is seen in figure 1-44.

FAR 91.411

Altimeter System and Altitude Reporting Equipment Tests and Inspections

a. No person may operate an airplane, or helicopter, in controlled airspace under IFR unless:

   1. Within the preceding 24 calendar months, each static pressure system, each altimeter instrument, and each automatic pressure altitude reporting system has been tested and inspected and found to comply with Appendix E of Part 43 of this chapter;
   2. Except for the use of system drain and alternate static pressure valves, following any opening and closing of a static pressure system, that system has been tested and

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

inspected and found to comply with paragraph (a) Appendices E and F of Part 43 of this chapter; and

3. Following installation or maintenance on the automatic pressure altitude reporting system of the ATC transponder where any error could have been introduced, the entire system has been tested, inspected and found to comply with paragraph (c) Appendix E of part 43 of this chapter.

b. The tests required by paragraph (a.) of this section must be conducted by:

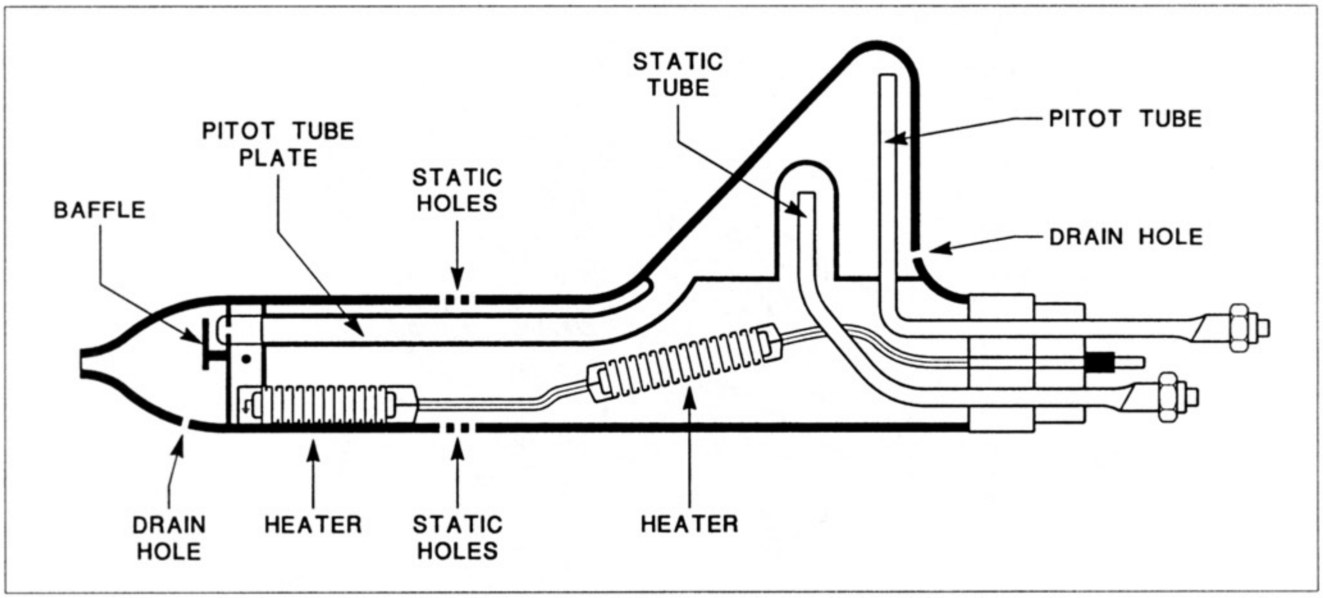
STATIC TUBE
PITOT TUBE PLATE
STATIC HOLES
BAFFLE
PITOT TUBE
DRAIN HOLE
DRAIN HOLE
HEATER
STATIC HOLES
HEATER

Figure 1-42. A pitot tube which also contains static ports and electric heating elements.

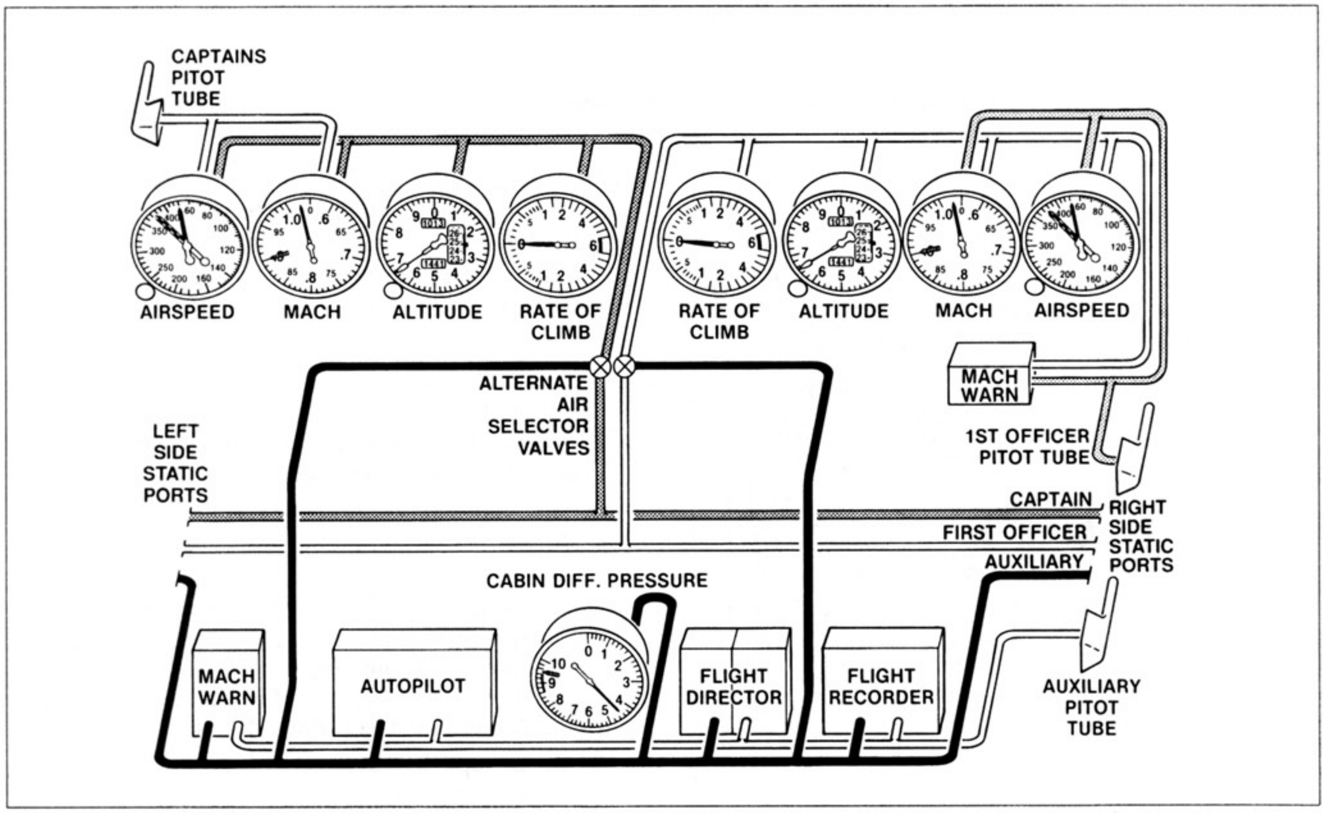
CAPTAINS PITOT TUBE
AIRSPEED
MACH
ALTITUDE
RATE OF CLIMB
RATE OF CLIMB
ALTITUDE
MACH
AIRSPEED
ALTERNATE AIR SELECTOR VALVES
MACH WARN
1ST OFFICER PITOT TUBE
LEFT SIDE STATIC PORTS
CAPTAIN
FIRST OFFICER
AUXILIARY
RIGHT SIDE STATIC PORTS
CABIN DIFF. PRESSURE
MACH WARN
AUTOPILOT
FLIGHT DIRECTOR
FLIGHT RECORDER
AUXILIARY PITOT TUBE

Figure 1-43. The pitot-static system for a large jet airplane showing the instruments and equipment operated by the three pitot tubes and three sets of static ports.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

1. The manufacturer of the airplane or helicopter;
2. A certificated repair station with appropriate rating;
3. A certificated mechanic with an airframe rating (but only for the static pressure system tests and inspections).

FAR 43 Appendix E
Altimeter System Test and Inspection

Each person performing the altimeter system tests and inspections required by FAR 91.411 shall comply with the following:

a. Static pressure system;
   1. Ensure freedom from entrapped moisture and restrictions.
   2. Determine that leakage is within the tolerances established in FAR 23.1325 or 25.1325 whichever is applicable.
   3. Determine that the static port heater, if installed, is operative.
   4. Ensure that no alterations or deformation of the airframe surface have been made that would affect the relationship between air pressure in the static pressure system and true ambient static pressure for any flight condition.

b. Altimeter — omitted here

c. Automatic pressure altitude reporting system — omitted here

---

FAR 23.1325
Static Pressure System

b. If a static pressure system is necessary for the functioning of instruments, systems or devices, it must comply with the provisions of paragraphs (1) through (3) of this section.
   1. The design and installation of the static pressure system must be such that
      i. Positive drainage of moisture is provided
      ii. Chafing of tubing and excessive distortion or restriction in bends is avoided; and
      iii. The materials used are durable, suitable for the purpose and protected against corrosion.
   2. A proof test must be conducted to demonstrate the integrity of the static pressure system in the following manner
      i. Unpressurized Airplanes — Evacuate the static pressure system to a pressure differential of approximately 1 inch of mercury or to a reading of 1,000 ft. above the aircraft elevation at the time of the test. Without additional pumping for a period of one minute, the loss of indicated altitude must not exceed 100 ft. on the altimeter.
      ii. Pressurized Airplanes — Evacuate the static pressure system until a pressure differential equivalent to the maximum cabin pressure differential for which the airplane is type certificated is achieved. Without additional pumping for a period of 1 minute, the loss of indicated altitude must not exceed 2 percent of the equivalent altitude or 100 ft., whichever is greater.
   3. Each static pressure port must be designed and located so that errors will not result when the aircraft encounters icing conditions. An anti-icing means or alternate static ports may be used to show compliance.

---

d. Pitot System Tests

There is no specific test for pitot systems as there is for static systems other than the normal inspections of the entire aircraft. If a problem is reported or suspected with a pitot system, there is a general leak test procedure in AC 43.13-1A, as well as some general guidelines for pitot-static system maintenance. The procedure for leak testing the pitot system is: Apply pressure to the pitot tube to cause the airspeed indicator to show 150 knots. Seal off for 1 minute and the maximum loss of indicated airspeed should not exceed 10 knots.

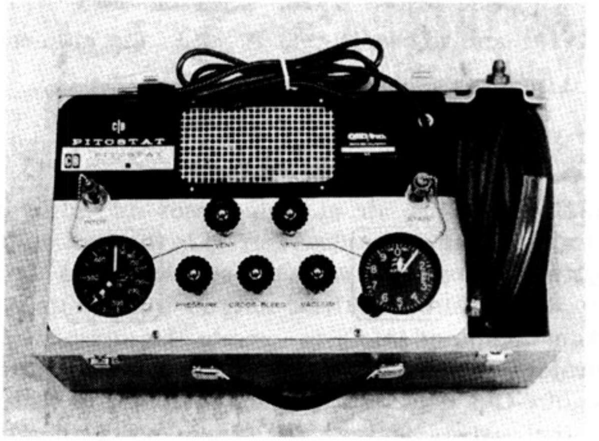

***Figure 1-44.** A portable pitot-static system tester which can be used for leak tests and other maintenance functions.*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

AC 43.13-1A

Precautions in Testing Pitot-Static System

1. Perform all maintenance and inspections before leak testing.
2. Use a system diagram.
3. Check the test unit for leaks before beginning the test.
4. Run full range tests only if you are thoroughly familiar with both the aircraft and the test equipment.
5. Pressure in the pitot system must always be equal to or greater than the pressure in the static system.
6. The rate of change of pressure during testing should not exceed the limits for any installed instrument.
7. After testing make sure that the system is returned to flying condition, such as removing tape from ports and drain holes.

There is an FAR that concerns the altimeter setting which is set by the pilot in the Kollsman window.

FAR 91.121

Altimeter Settings

Each person operating an aircraft shall maintain the altitude of the aircraft by reference to an altimeter that is set:

1. Below 18,000 ft. MSL, to
    - i. The current reported altimeter setting of a station along the route of flight and within 100 nautical miles of the aircraft
    - ii. If there is no station within 100 nm, the nearest appropriate station altimeter setting
    - iii. If the aircraft has no radio, the elevation of the departure airport or the setting available before takeoff shall be used.
2. At and above 18,000 ft. MSL, the altimeter shall be set to 29.92.

---

e. Air Data Computer Systems

Aircraft that operate at high speeds and high altitudes can have significant errors in the pitot-static system instruments with the simple probes used on smaller aircraft. These aircraft will use an Air Data Computer (ADC) to operate the airspeed indicator, altimeter, VSI and any other systems that require this data. The air data computer is placed in the system between the sensor ports and the instruments to automatically apply corrections in order to increase accuracy.

The air data computer has three inputs; pitot pressure from the pitot tube, static pressure from the static ports and total air temperature (TAT) from a special TAT probe. The TAT measurement is needed to correct the instrument indications for friction heating of the air at high speeds. The TAT probe also permits the calculation of SAT (static air temperature) which is used to apply corrections for non-standard temperatures for any flight altitude. The outputs of the air data computer supply a number of cockpit instruments. The three basic pitot-static instruments are operated by the ADC and often several others are added. The common inputs and outputs associated with an air data computer are shown in figure 1-45.

High speed jet airplanes require a machmeter; this could be a separate instrument or included with the airspeed indicator.

Aircraft with an air data computer usually have a digital display on the instrument panel which gives a calculated true airspeed and total air temperature or static air temperature. TAT includes the heating effect of the friction at high speed whereas SAT is just ambient outside air temperature. The temperature indications are especially important for turbine engines which are affected by the temperature of the intake air.

The air data computer system automatically compensates for both temperature effects and the compressibility of the air at higher Mach numbers. This helps to ensure accurate instrument readings throughout a wide range of altitudes and airspeeds. Air data computers are typically found on turboprop airplanes and jet airplanes.

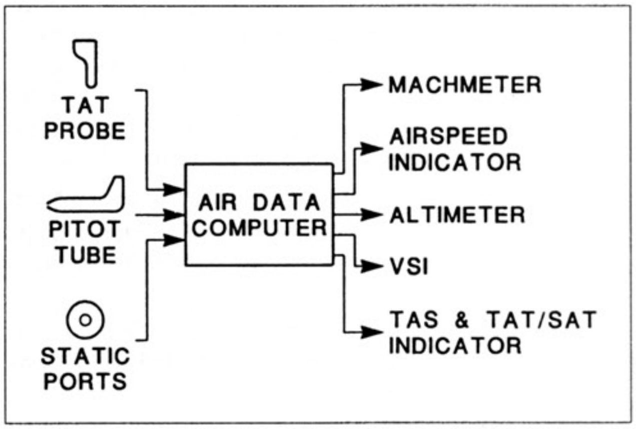

Figure 1-45. An air data computer provides more accurate readings on the pitot-static instruments for high performance aircraft. Total air temperature is measured by a special probe and provided as an input to the ADC.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

D. Gyro Instruments

The instruments know as gyro instruments are required for IFR flight and can also be an aid to accurate flying in VFR conditions. These instruments utilize the principles of a spinning gyroscope to give the pilot information about the aircraft's pitch and roll attitude, heading and rate of turn. A gyroscope is a device which consists of a weighted wheel or rotor which spins at high speed and is held in an arrangement of hinged mounting rings called gimbals (figure 1-46).

The gyro has three axes and one is always the spin axis. Depending on the type of gimbals used, it will be able to move relative to the mounting base around one or both of the remaining axes. If it has 1 degree of freedom, it can move around one axis and if it has two degrees of freedom it can move around both. A gyro with two degrees of freedom is also called a free gyro.

1. Principles of Gyroscopes

There are two main properties of a spinning gyro which are of importance to aircraft use, they are:

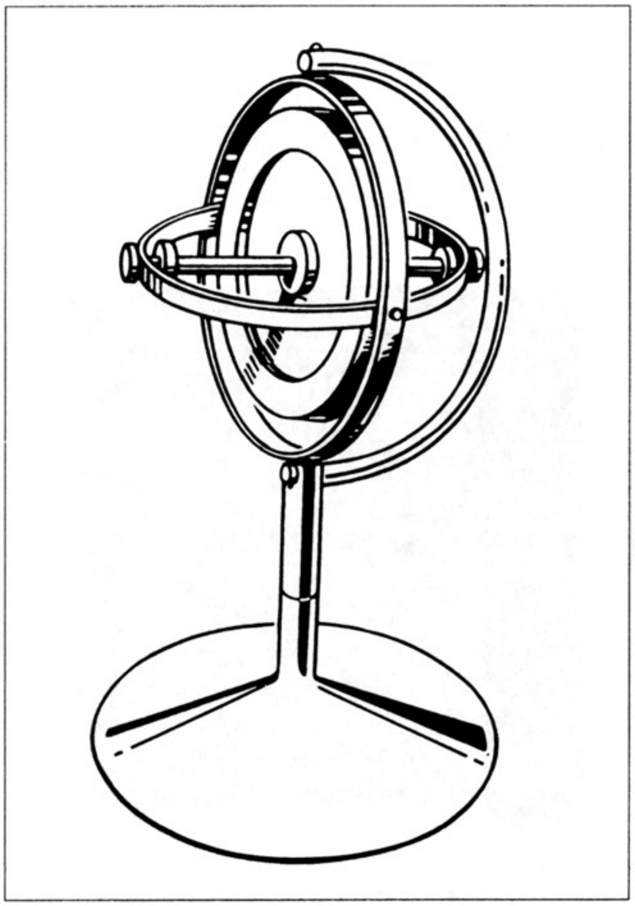

Figure 1-46. A simple gyroscope with both inner and outer gimbals.

1. Rigidity in Space — This means that the gyro rotor will try to maintain its position in space even when its mounting base is tilted and rotated. This is illustrated in figure 1-47.
2. Precession — This effect will cause a gyro, when it is acted upon by an outside force, to tilt or rotate as if the disturbing force was applied to it 90 degrees ahead in the direction of rotation of the rotor (figure 1-48).

Gyros are also subject to other effects such as oscillation. Oscillation is a problem caused by the mass of the gimbals. It can be reduced by making the gimbals lighter. A number of things can be done to make the gyro more stable and more efficient. One way is to concentrate the mass of the rotor on its rim and reduce the mass of the web and shaft which connects it to the bearings. Another way to make it more efficient is to increase the speed of rotation. There are two common methods used to spin the rotor of an aircraft gyro instrument. Pneumatics makes use of a stream of air directed at the rim of the rotor to make it spin at about 8,000 RPM. Electric motors can also be used to spin the rotor and will usually produce a speed of about 24,000 RPM. A gyro can become unstable which is called tumbling. Some aircraft gyro instruments have a caging knob or mechanism which is designed to return the gyro to a stable condition so that it will give correct instrument readings. A caging knob can be seen at the bottom left of figure 1-49. The latest types of gyro instruments are non-tumbling and as long as the instrument is in good condition it will not tumble, even in unusual attitudes. The three common types of aircraft gyro instruments are the directional gyro (DG), the artificial horizon and the turn and bank instrument.

2. Directional Gyro

The directional gyro is the primary heading reference for IFR flight. The magnetic compass is not a good

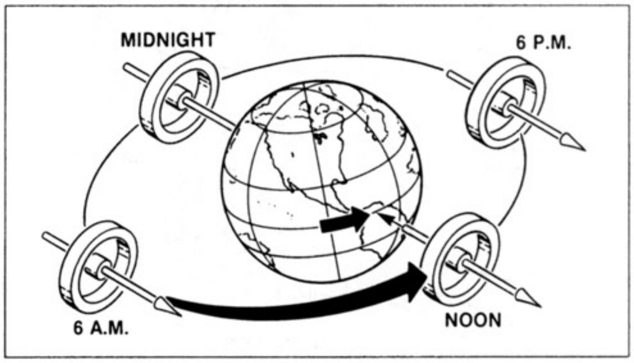

Figure 1-47. The rigidity in space characteristic causes the gyro rotor to try to maintain its orientation in space.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

heading reference for IFR because it tends to oscillate and is not as stable as a DG. This instrument uses a free gyro with a horizontal spin axis. The DG will drift due to precession errors and must be reset every 15 to 20 minutes using the magnetic compass as a reference. Older style DGs had a rectangular window on the face through which the numbers representing the magnetic heading could be read. This older style DG presentation is shown in figure 1-50. Newer style DGs show a full compass card with the indicated magnetic heading under the index mark at the top of the instrument as seen in figure 1-51.

The newer style DG with a full compass card is often called a heading indicator. Another name for a DG type instrument is a gyro compass. The DG uses a free gyro because the spin axis must remain horizontal to give accurate readings. When the airplane banks, for example, the rotor will maintain its horizontal spin axis. On many modern aircraft a more sophisticated instrument replaces the DG, this instrument is the horizontal situation indicator or HSI.

The HSI shown in figure 1-52 is an example of a modern integrated instrument. It has a gyro stabilized compass card like a DG that indicates the aircraft's magnetic heading. Unlike the DG however, this compass card is slaved to a remote compass so it does not have to be reset every 15 to 20 minutes. The HSI is called an integrated instrument because it combines several different types of displays which would normally be found in separate

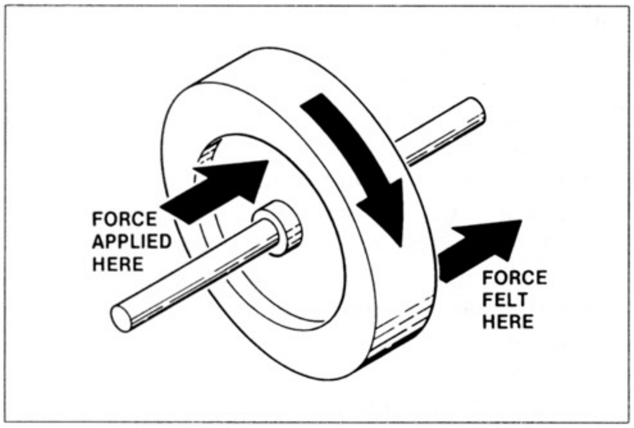

Figure 1-48. Precession causes a gyro rotor to tilt as if a disturbing force was applied 90° ahead in the direction of rotation from the actual point of application.

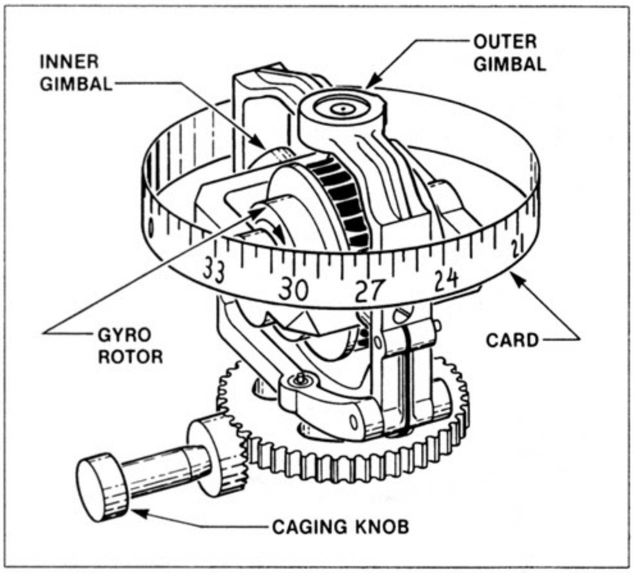

Figure 1-49. The rotor in a directional gyro has a horizontal spin axis and two degrees of freedom.

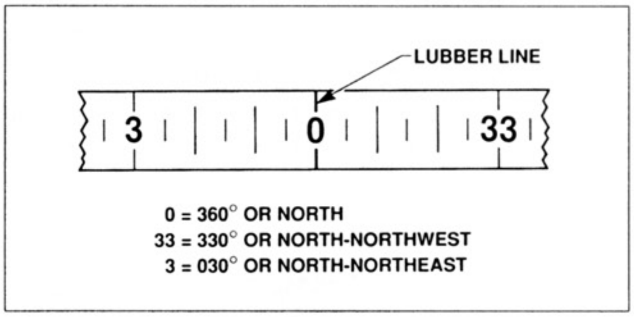

Figure 1-50. The old style DG displayed a small window like a magnetic compass. This type of presentation can be difficult to read.

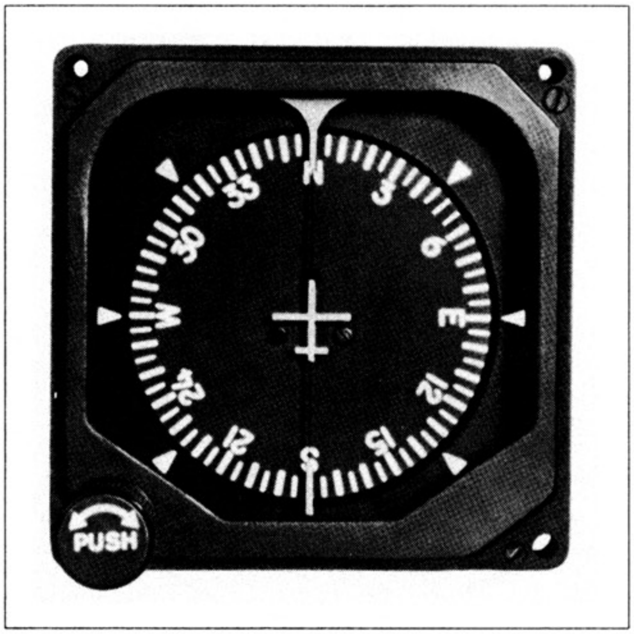

Figure 1-51. The modern DG or heading indicator has a display which shows a complete compass card. This type is easier to use.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

instruments. The wide split bar in the middle is connected to navigational radios and tells the pilot whether to turn right or left to follow the radio navigational signals. The indicators on the right side and at the top are connected to other navigational radios to provide additional information to the pilot.

3. Artificial Horizon

The artificial horizon is the pilot's most important instrument for IFR flying. As the name implies, it replaces the natural earth horizon that a pilot uses in VFR flying to maintain the correct pitch and roll attitude of the aircraft. When an aircraft is flying in the clouds, the pilot must rely on the

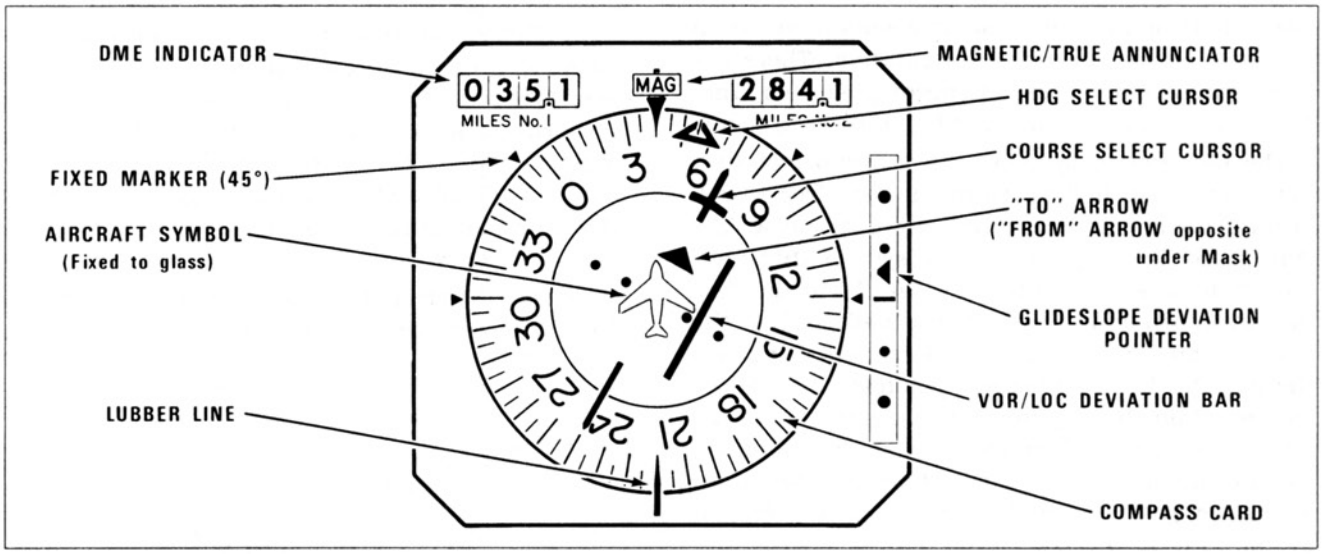
DME INDICATOR
0351
MILES No. 1
MAG
2841
FIXED MARKER (45°)
AIRCRAFT SYMBOL
(Fixed to glass)
LUBBER LINE
MAGNETIC/TRUE ANNUNCIATOR
HDG SELECT CURSOR
COURSE SELECT CURSOR
"TO" ARROW
("FROM" ARROW opposite under Mask)
GLIDESLOPE DEVIATION POINTER
VOR/LOC DEVIATION BAR
COMPASS CARD
0
3
6
9
12
15
18
21
24
27
30
33

Figure 1-52. The horizontal situation indicator is an integrated instrument that displays many additional kinds of information besides gyro stabilized heading data.

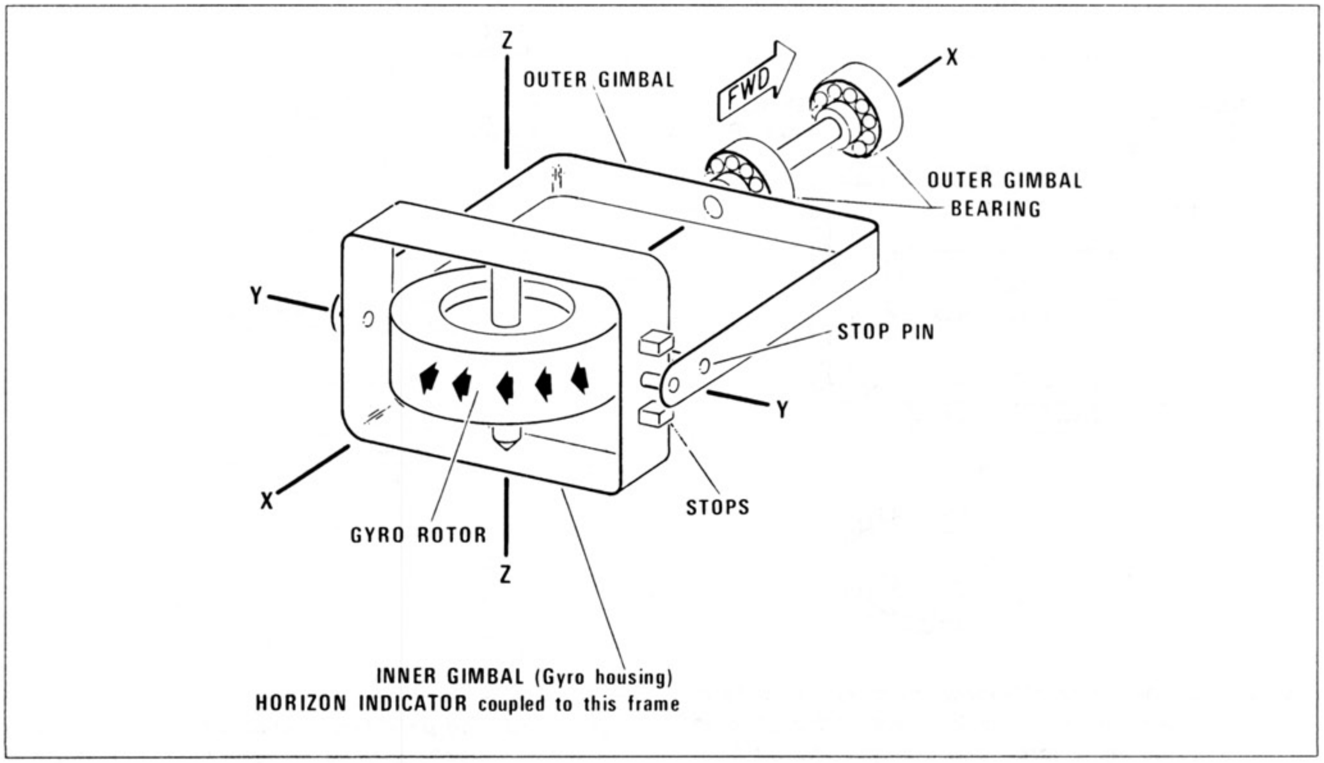
Z
OUTER GIMBAL
FWD
X
OUTER GIMBAL BEARING
Y
STOP PIN
Y
X
STOPS
GYRO ROTOR
Z
INNER GIMBAL (Gyro housing)
HORIZON INDICATOR coupled to this frame

Figure 1-53. The artificial horizon uses a gyro rotor with a vertical spin axis and two degrees of freedom.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

artificial horizon to determine the aircraft attitude and prevent a loss of control. The artificial horizon is also known as a gyro horizon, an attitude indicator, an attitude gyro and a bank and pitch instrument. The arrangement of the spin axis and gimbals for an artificial horizon is illustrated in figure 1-53. The artificial horizon is a free gyro with a vertical spin axis; this allows it to measure the angular displacement of the aircraft in both pitch and roll. The internal mechanism of the artificial horizon must have some means to maintain the spin axis in a vertical orientation. There are two common devices used to accomplish this function, the pendulous vane and ball erector systems.

The ball erector system uses a number of steel balls similar to ball bearings that are free to roll around on a plate mounted above or below the gyro rotor. One type of erector system that uses steel balls is shown in figure 1-54. When the gyro tilts away from the vertical, the balls roll to the low side and this produces a force which pushes the gyro back to the vertical position. This device will maintain the required vertical spin axis of the gyro rotor. The ball erector is usually found on electric motor driven artificial horizons. When an air-driven gyro is used, the pendulous vane erector mechanism is utilized.

The pendulous vanes in this type of erector mechanism are small gravity operated air valves. When the rotor tilts away from the vertical, the vanes move to open the air valves in such a way that streams of air are directed to push the rotor back to the vertical position. The swing of the pendulous vanes to open and close the air valves can be seen in figure 1-55.

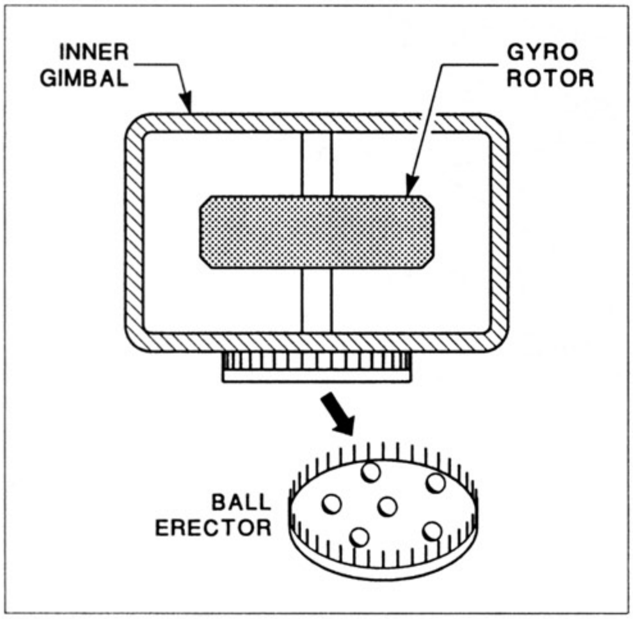

***Figure 1-54.** The ball erector mechanism uses steel balls that move to the low side of the plate to supply a force that corrects the artificial horizon back to a vertical spin axis.*

The artificial horizon has a presentation which shows an airplane symbol with the horizon behind it. It includes index marks to show the angle of bank. Some indications of an artificial horizon for different flight attitudes are illustrated in figure 1-56. The newer types of artificial horizon have a more user friendly presentation on the instrument face. Different colors are used above and below the horizon line to make it easier to read. Converging lines are placed below the horizon line to create perspective and additional markings for pitch attitude are included. These features can be seen on the newer type artificial horizon in figure 1-57.

The small knob on the front of the artificial horizon is used to move the airplane symbol up and down

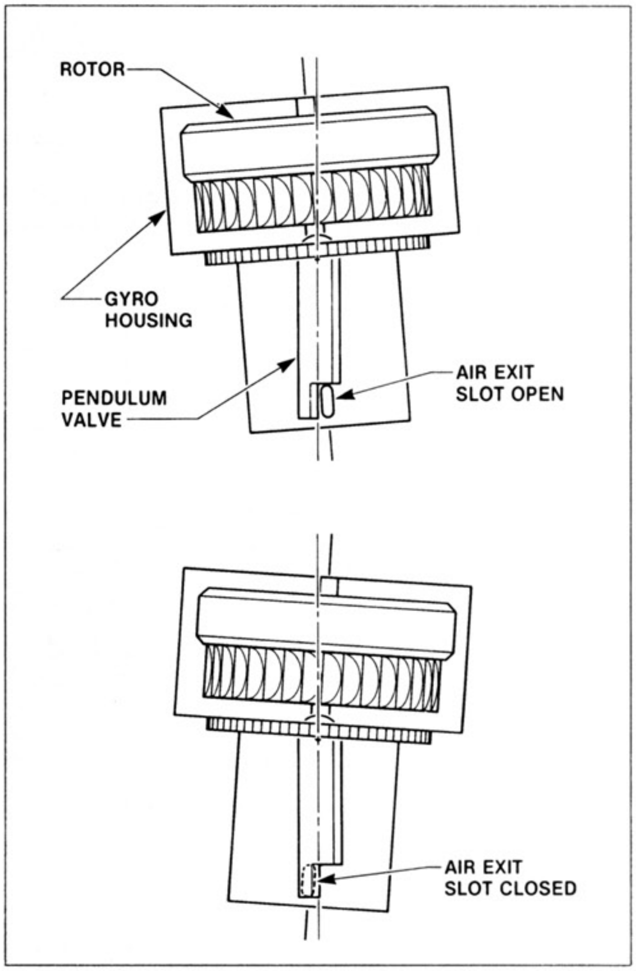

***Figure 1-55.** The pendulous vanes used in the erector mechanism of an air-driven gyro horizon are opened and closed by gravity.*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

to adjust for different aircraft flight attitudes and for tall and short pilots. There also may be a control for quick erect that can be used to stabilize the gyro if it tumbles. There is a newer type of integrated instrument which replaces the artificial horizon on many modern aircraft. This instrument is called the attitude director indicator (ADI). It includes command bars that are operated by a flight director. The flight director is a computer which receives signals from navigational radios and other sources and calculates the correct pitch and roll attitudes to keep the aircraft on course or return it to the desired flight path. The output signals from the flight director computer move the command bars on the ADI and the pilot follows these commands. An example of the use of these command bars on an ADI is found in figure 1-58.

The latest kinds of ADIs use different shapes for the airplane symbol and the command bars as can be seen in figure 1-59. The airplane symbol is a triangle that resembles a delta wing airplane and the command bars are two converging triangular shapes above the airplane symbol. They often use different colors as well to make it easier to distinguish between the airplane symbol and command bar symbol. Some ADIs also include indicators for other aircraft systems in addition to the flight director indicator. The indicator pointer on the right side of the ADI is operated by the aircraft autothrottle system. The pointer on the left side is

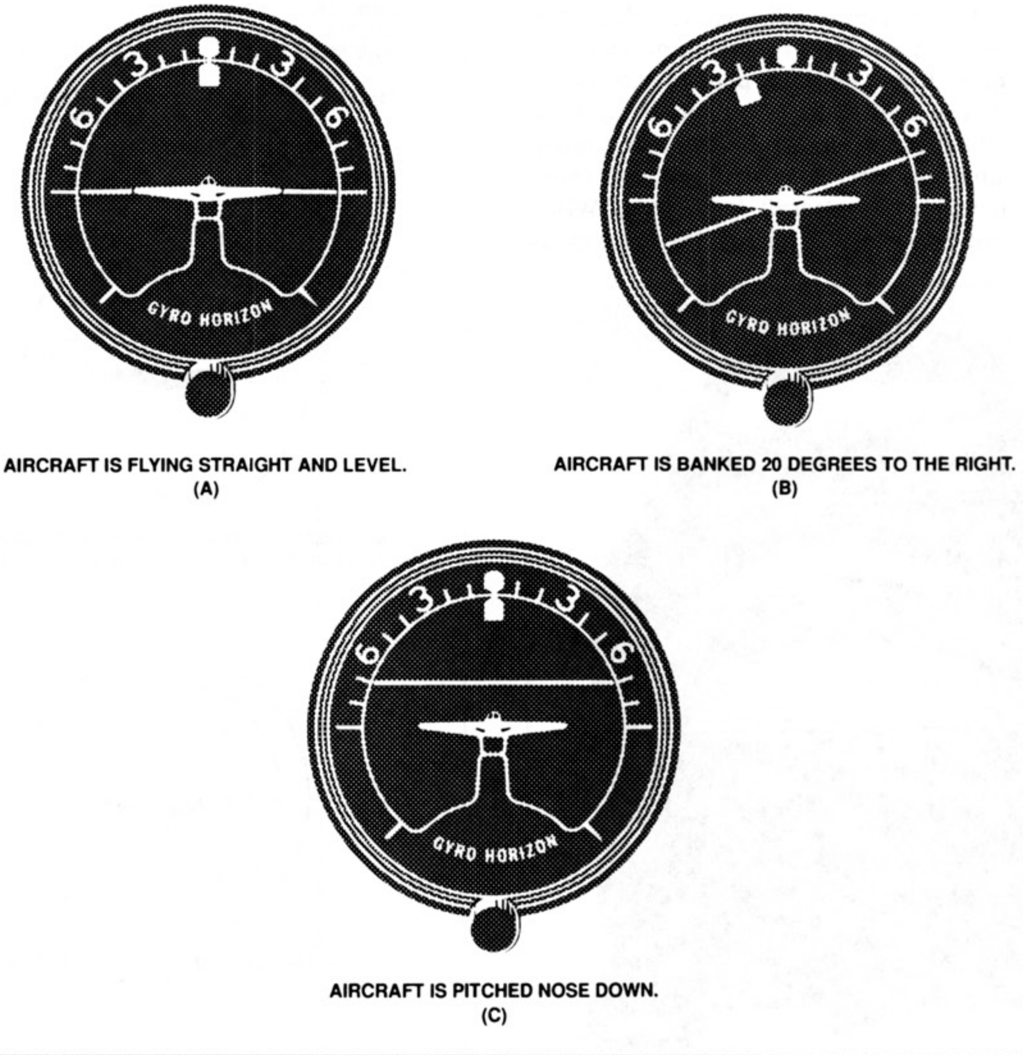

AIRCRAFT IS FLYING STRAIGHT AND LEVEL.
(A)

AIRCRAFT IS BANKED 20 DEGREES TO THE RIGHT.
(B)

AIRCRAFT IS PITCHED NOSE DOWN.
(C)

Figure 1-56. Indications for an artificial horizon instrument.

(A) The airplane is straight and level.

(B) The airplane is in a 20° bank to the right with the nose on the horizon.

(C) The wings are level, but the nose is below the horizon.

operated by a glideslope radio receiver and gives the pilot information needed for an instrument ILS approach. An ADI is an integrated instrument that gives pitch and roll attitude data like an artificial horizon and additional displays of information from radio navigation sources and other aircraft systems like the flight director. On sophisticated aircraft which have backup gyro instruments, a turn and bank is not installed, so the inclinometer is installed at the bottom of the ADI as can be seen in figure 1-59.

4. Turn and Bank

The last of the three basic gyro instruments is the oldest and simplest. It is called the turn and bank and it is really two instruments in one. The gyro part of the instrument measures the rate of turn for the aircraft. The inclinometer or slip-skid indicator is a simple mechanical instrument that consists of a ball in a liquid filled glass tube. This tube is curved and the ball reacts to gravity and centrifugal force. It is used by a pilot to coordinate the use of aileron and rudder control. If the pilot keeps the ball centered, the aircraft is being flown in a coordinated manner. This instrument is especially helpful when the aircraft is turning. When the ball is not centered, it means the aircraft is flying a little sideways. The gyro rotor of the turn and bank is designed to measure the rate of turn of the aircraft. It is the only one of the three basic gyro instruments which is a rate gyro.

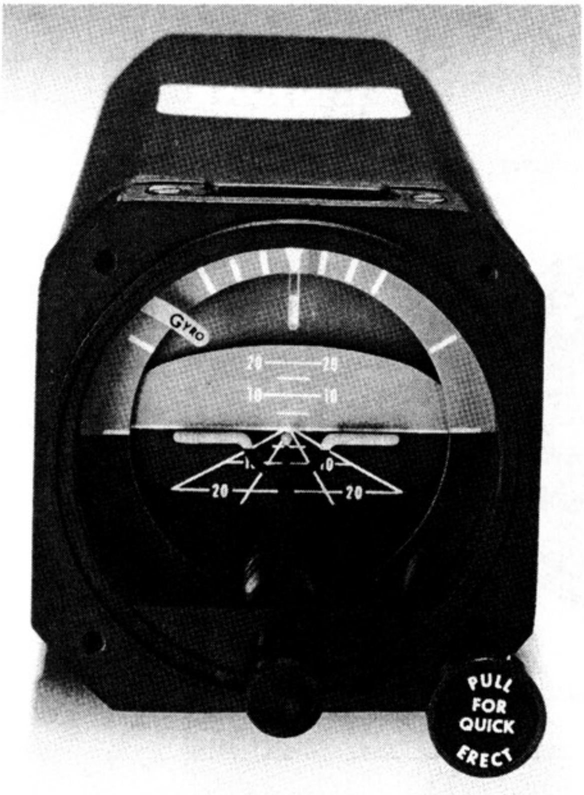

Figure 1-57. The newer type of artificial horizon uses a presentation that is easier to interpret.

The other two basic gyro instruments measure angular displacement about the aircraft's axes. The turn and bank has a gyro with a horizontal spin

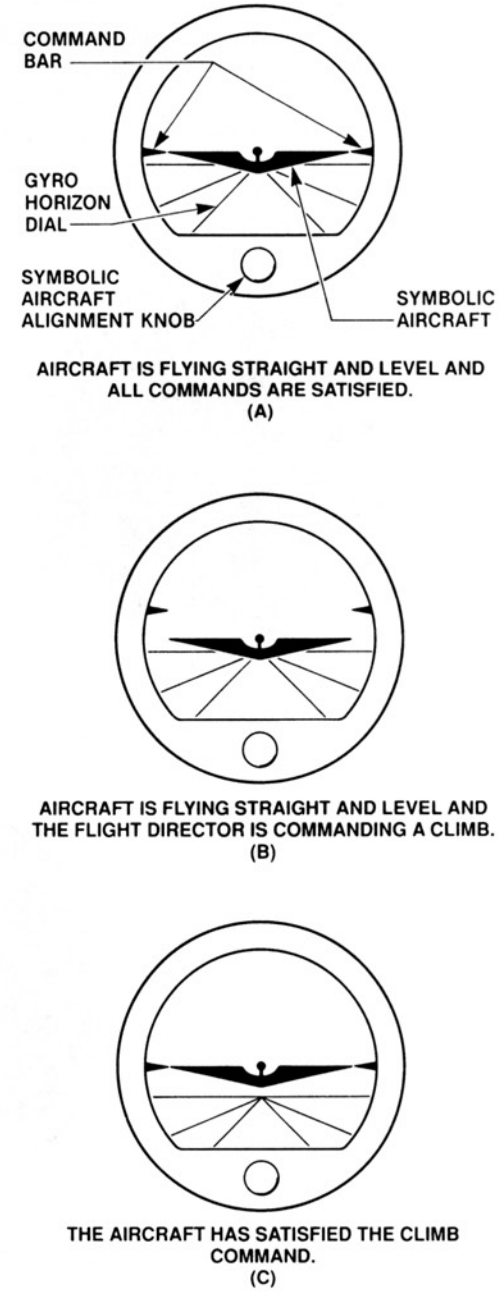

AIRCRAFT IS FLYING STRAIGHT AND LEVEL AND ALL COMMANDS ARE SATISFIED.
(A)

AIRCRAFT IS FLYING STRAIGHT AND LEVEL AND THE FLIGHT DIRECTOR IS COMMANDING A CLIMB.
(B)

THE AIRCRAFT HAS SATISFIED THE CLIMB COMMAND.
(C)

Figure 1-58. The command bars on an attitude director indicator show the pilot the pitch and roll attitude that is needed to satisfy the commands of the flight director.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

axis and one degree of freedom. The feature that makes it a rate gyro are the springs that are connected to the gimbals. These springs oppose the precession force which is caused by the aircraft turning. These features of the turn and bank can be seen in figure 1-60.

When the aircraft turns, the gimbal holding the gyro rotor tilts over against the tension of the spring and moves the pointer to indicate the direction and rate of turn. The turn and bank gives readings based on the concept of a standard rate turn. A standard rate turn is a turning rate of 3° per second. This is also called a 2 minute turn because it would take 2 minutes to turn 360° at this rate. A standard rate turn is not suitable for a high speed aircraft because it would require a steep angle of bank.

These higher speed aircraft would use a½ standard rate turn which is 1-½° per second or a 4 minute turn. Turn and banks are manufactured in both types; 2 minute turn and 4 minute turn, both of which are shown in figure 1-61. The turn and bank indicator is also called a turn and slip indicator and a needle and ball. The face of the instrument shows a needle to indicate turn direction and rate and a ball which is the slip-skid indicator or inclinometer.

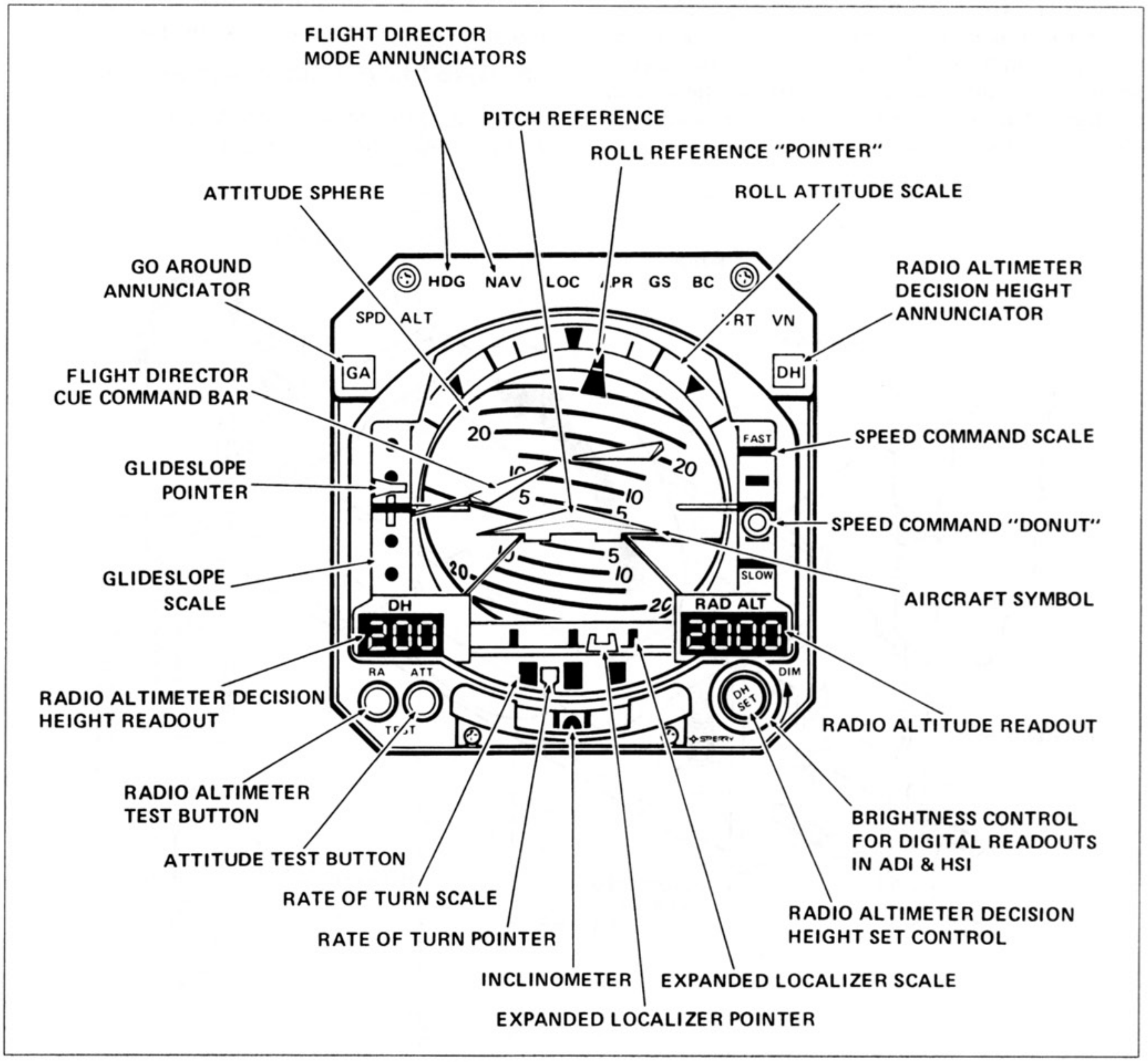
FLIGHT DIRECTOR MODE ANNUNCIATORS
PITCH REFERENCE
ROLL REFERENCE "POINTER"
ATTITUDE SPHERE
ROLL ATTITUDE SCALE
GO AROUND ANNUNCIATOR
RADIO ALTIMETER DECISION HEIGHT ANNUNCIATOR
HDG
NAV
LOC
APR
GS
BC
SPD
ALT
VRT
VN
GA
DH
FLIGHT DIRECTOR CUE COMMAND BAR
20
10
5
FAST
SPEED COMMAND SCALE
GLIDESLOPE POINTER
SPEED COMMAND "DONUT"
SLOW
GLIDESLOPE SCALE
AIRCRAFT SYMBOL
DH
200
RAD ALT
2000
RA
ATT
TEST
DIM
DH SET
SPERRY
RADIO ALTIMETER DECISION HEIGHT READOUT
RADIO ALTITUDE READOUT
RADIO ALTIMETER TEST BUTTON
BRIGHTNESS CONTROL FOR DIGITAL READOUTS IN ADI & HSI
ATTITUDE TEST BUTTON
RATE OF TURN SCALE
RADIO ALTIMETER DECISION HEIGHT SET CONTROL
RATE OF TURN POINTER
INCLINOMETER
EXPANDED LOCALIZER SCALE
EXPANDED LOCALIZER POINTER

Figure 1-59. The newer type of ADI uses different shapes for the airplane symbol and the command bars symbol. It also may include additional displays for other systems. (Courtesy of Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The turn and bank is shown in figure 1-61. The pointer is the rate of turn indicator and the glass tube is the inclinometer.

The index marks on either side of the center position of the pointer on the lower instrument are called dog houses. When the pointer is lined up with a dog house, it indicates a 2-minute turn on the bottom instrument. A 2-minute turn on the upper instrument would be indicated by a one needle width deflection of the turn needle. The turn and bank is considered to be a backup instrumnet for the artificial horizon. If the artificial horizon fails, it is possible to fly the aircraft using the turn and bank in its place.

Another gyro instrument called the turn coordinator is a modified version of the turn and bank. The only significant difference in the internal mechanism is the fact that the tilt axis for the gimbal is changed to a 30° angle from the horizontal as shown in figure 1-62. This causes the gyro rotor to react to rotation around the longitudinal axis as well as the vertical axis. The turn and bank only measures rotation rate about the vertical axis so that it cannot be used accurately to level the wings. The turn coordinator is a better back-up instrument for the artificial horizon for this reason. Since the turn coordinator is not the same as a turn and bank and doesn't give the same kind of information, it has a different appearance so that pilots won't confuse the two instruments.

The turn coordinator as illustrated in figure 1-63 uses a rear view of a small airplane as the indicator. When the wing tip of the airplane symbol is lined up with an index mark, it indicates a standard rate turn for the 2 minute type. The turn coordinator also includes an inclinometer, like the turn and bank.

5. Gyro Instrument Power Sources

Aircraft gyro instruments can be powered by electricity or air. The electric gyros can use 14 or

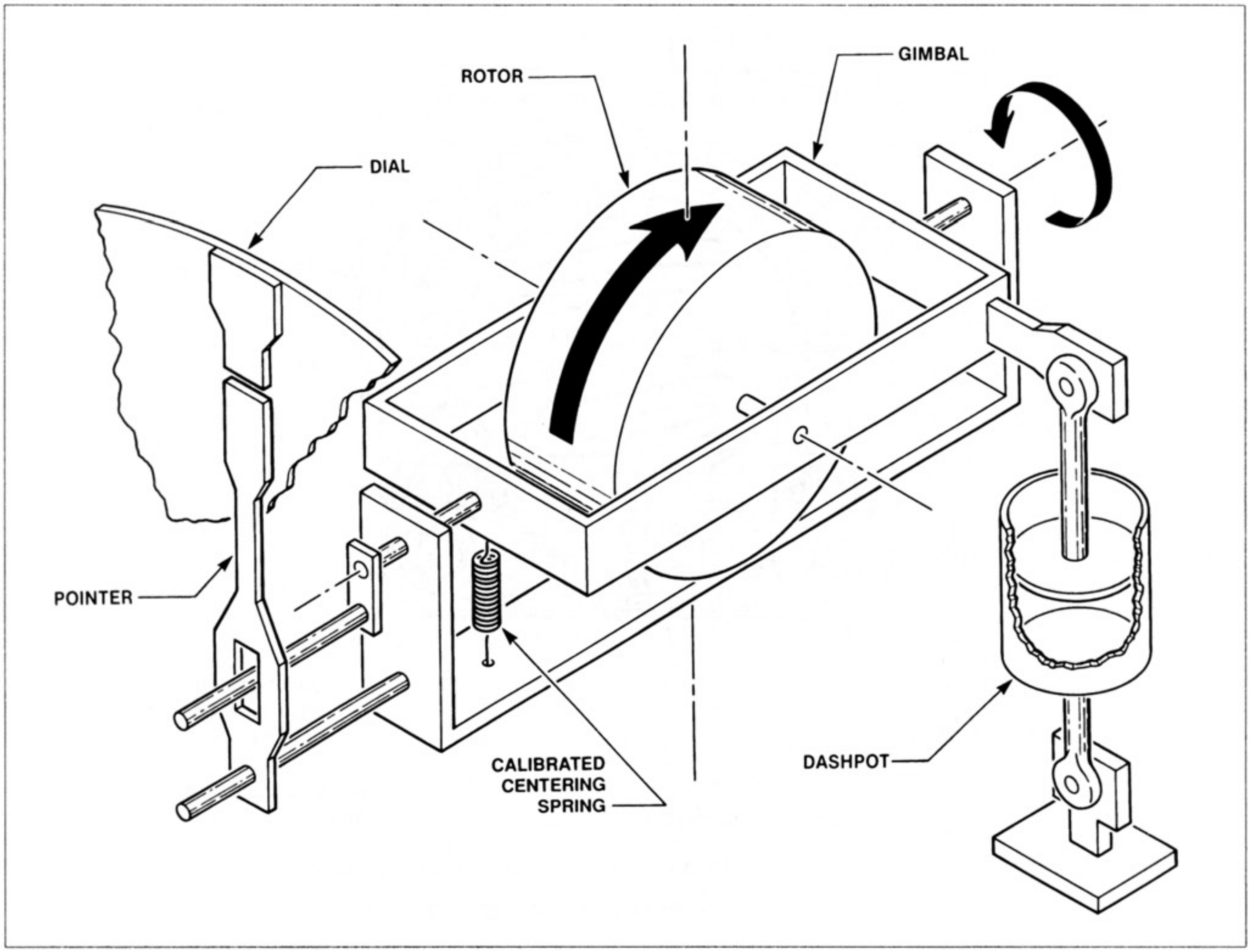

Figure 1-60. The turn-and-bank instrument has a gyro rotor with a horizontal spin axis and one degree of freedom. It also has a centering spring on the gimbal.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

28 volts DC or several different values of AC. The gyros that are air driven can use either an air pump or bleed air from turbine engines. Air-driven gyros can either use suction pressure or positive pressure. Those that use suction pressure are usually called vacuum driven gyros. Some older aircraft used vacuum venturis to power the air-driven gyro instruments. The venturi for gyros is mounted on the fuselage of the aircraft and the airflow caused by the forward motion of the aircraft creates a low pressure or suction in the throat of the venturi. A major problem with using a venturi for IFR flight is that the venturi tends to become blocked with ice under some flight conditions. Another disadvantage of the venturi tube is that the aircraft must maintain a certain minimum airspeed to generate enough vacuum for the gyros. The gyros will not be spun-up and stable during takeoff for example. Examples of 2″ and 4″ venturis

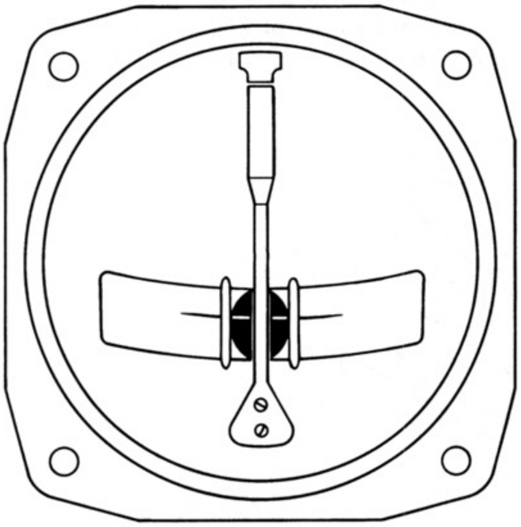

TWO-MINUTE TURN INDICATOR DIAL
(A)

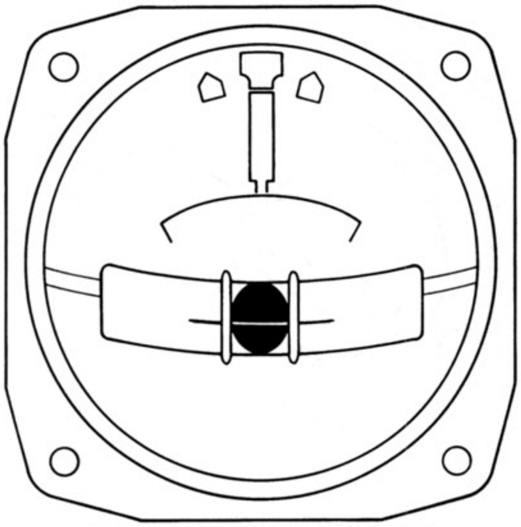

FOUR-MINUTE TURN INDICATOR DIAL
(B)

Figure 1-61. Both 2-minute and 4-minute turn and bank instruments are available.

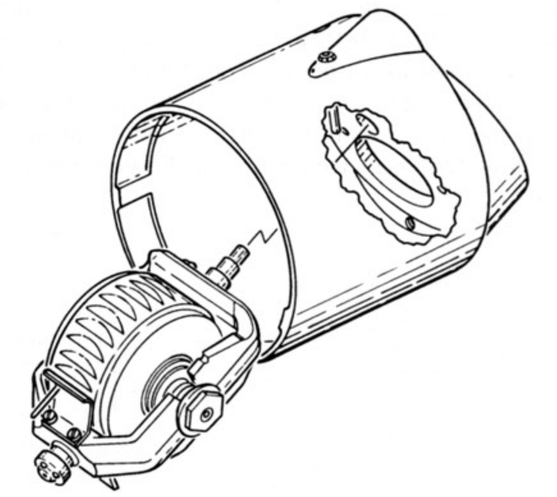

Figure 1-62. By mounting the gimbal at an angle to the horizontal, the turn coordinator senses rotation about both the roll and yaw axes of the aircraft.

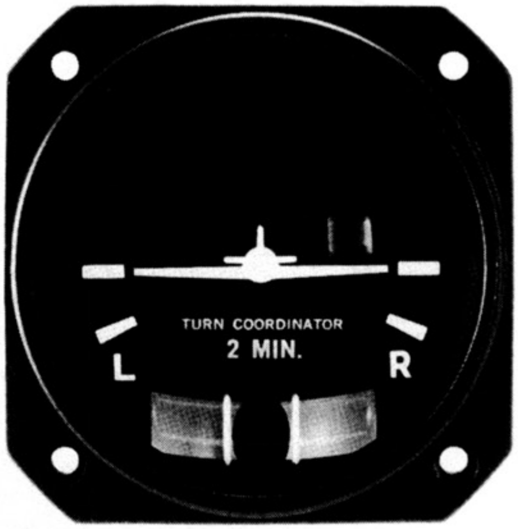

Figure 1-63. The presentation on the face of the turn coordinator is different from that of a turn and bank so that the two instruments will not be confused with each other.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

are found in figure 1-64. The 2″ and 4″ are not physical dimensions. They refer to the amount of suction in inches of mercury that each is designed to provide.

The most common type of air pump used on modern airplanes for the gyro instruments is called a dry air pump. It does not use any oil for sealing or lubrication. It is a vane type pump and the vanes are made of a carbon based material which gradually wears away in service from rubbing against the cylinder walls. Figure 1-65 shows a dry air pump connected to operate as a vacuum pump. Notice that the gyro instruments and gauge are installed in parallel. Figure 1-66 shows the same kind of dry air pump that has been connected to operate as a positive pressure pump. In the vacuum pump system the output of the pump is dumped overboard and the cockpit air is filtered before it flows into the instruments. A filter is required on the regulator of the vacuum system because air is drawn in at that point to regulate the vacuum pressure. An advantage of the positive pressure system is that it is better for aircraft that operate at higher altitudes of 15,000 to 18,000 ft. The positive pressure system requires a filter on the inlet side of the pump and a filter on the outlet side ahead of the instruments. A filter is not required on the regulator in the positive pressure system. There is also a wet pump for air driven gyros which uses engine oil for cooling and lubrication. It can only be used as a vacuum pump and requires an air/oil separator to return oil to the airplane's engine. Figure 1-67 shows the

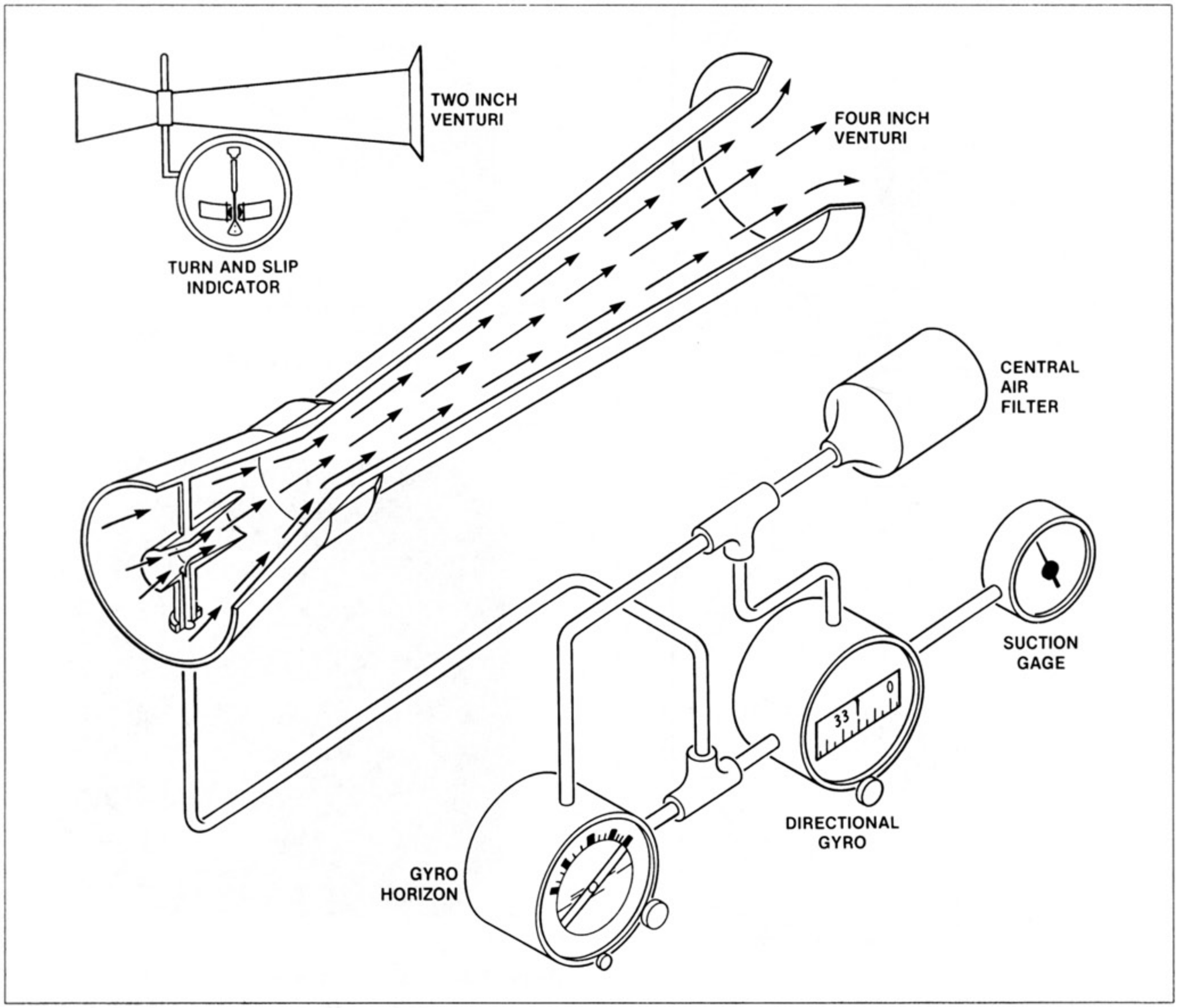

Figure 1-64. A 2 ″ and 4 ″ venturi are available to power air-driven gyro instruments. The ratings apply to the number of inches of Mercury vacuum that are provided, not to physical size.

air/oil separator in a wet pump system as well as a suction reducer that is used to drop the pressure for the turn and slip indicator. An air driven turn and bank or turn and slip requires about 2 inches of mercury while the other two basic types of gyros require 4-5 inches of mercury.

For any kind of air driven gyros, it is very important to change the filters regularly to ensure that only very clean air reaches the gyro instruments. The instruments are very delicate and can wear out rapidly if dust and dirt are allowed to enter with the air supply. A typical air filter for aircraft gyro instrument systems is shown in figure 1-68. The small filter installed on vacuum regulator valves can be seen in figure 1-69. The tubing and hose in an air driven gyro system must be checked to make certain that no restrictions are present which would create higher than normal resistance to the flow of air. The only lubricant approved for vacuum system fittings is usually a silicone spray. Thread lubricants and Teflon® tape should not be used as they might get drawn into the system and cause damage.

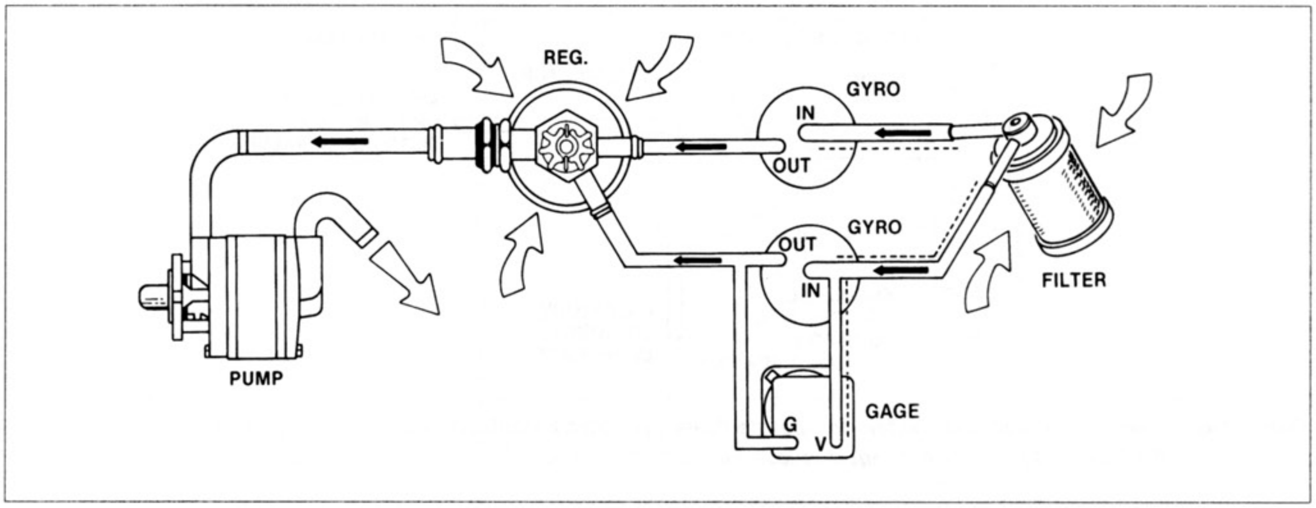

Figure 1-65. A gyro instrument vacuum system that uses a dry air pump.

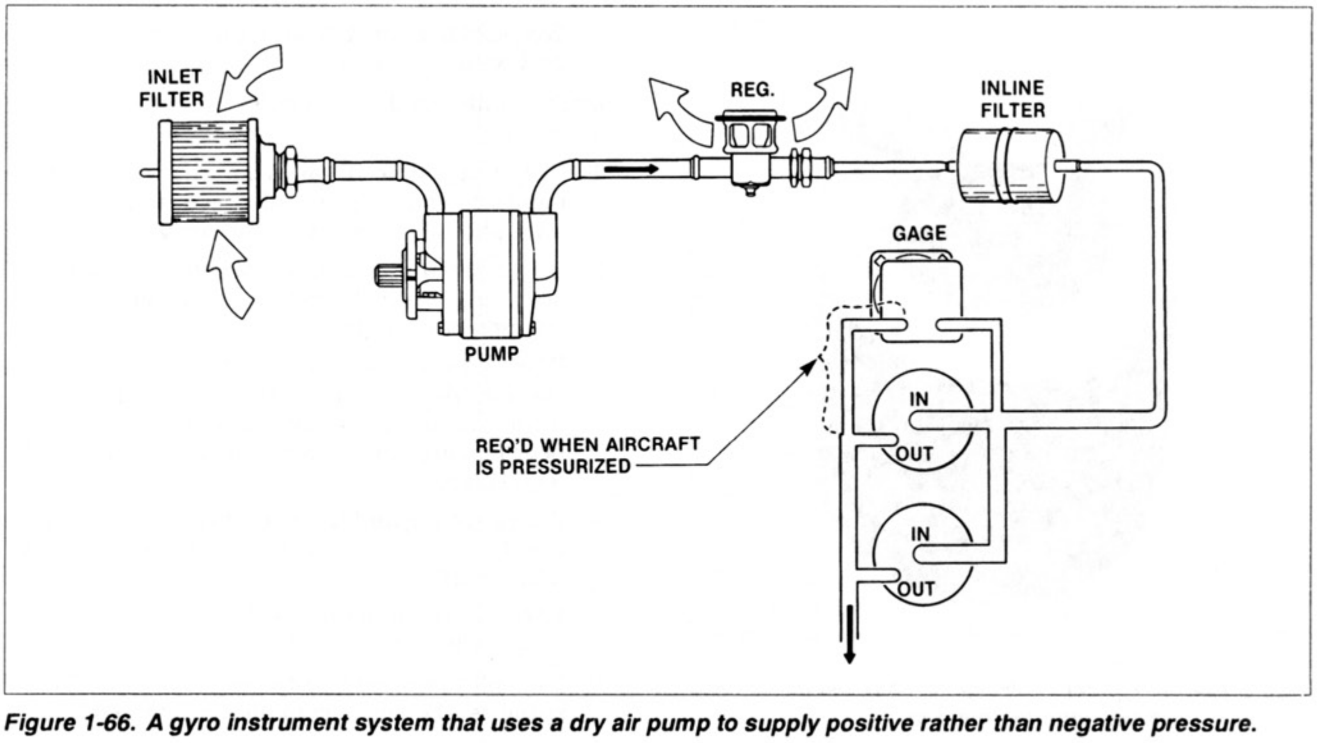

Figure 1-66. A gyro instrument system that uses a dry air pump to supply positive rather than negative pressure.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

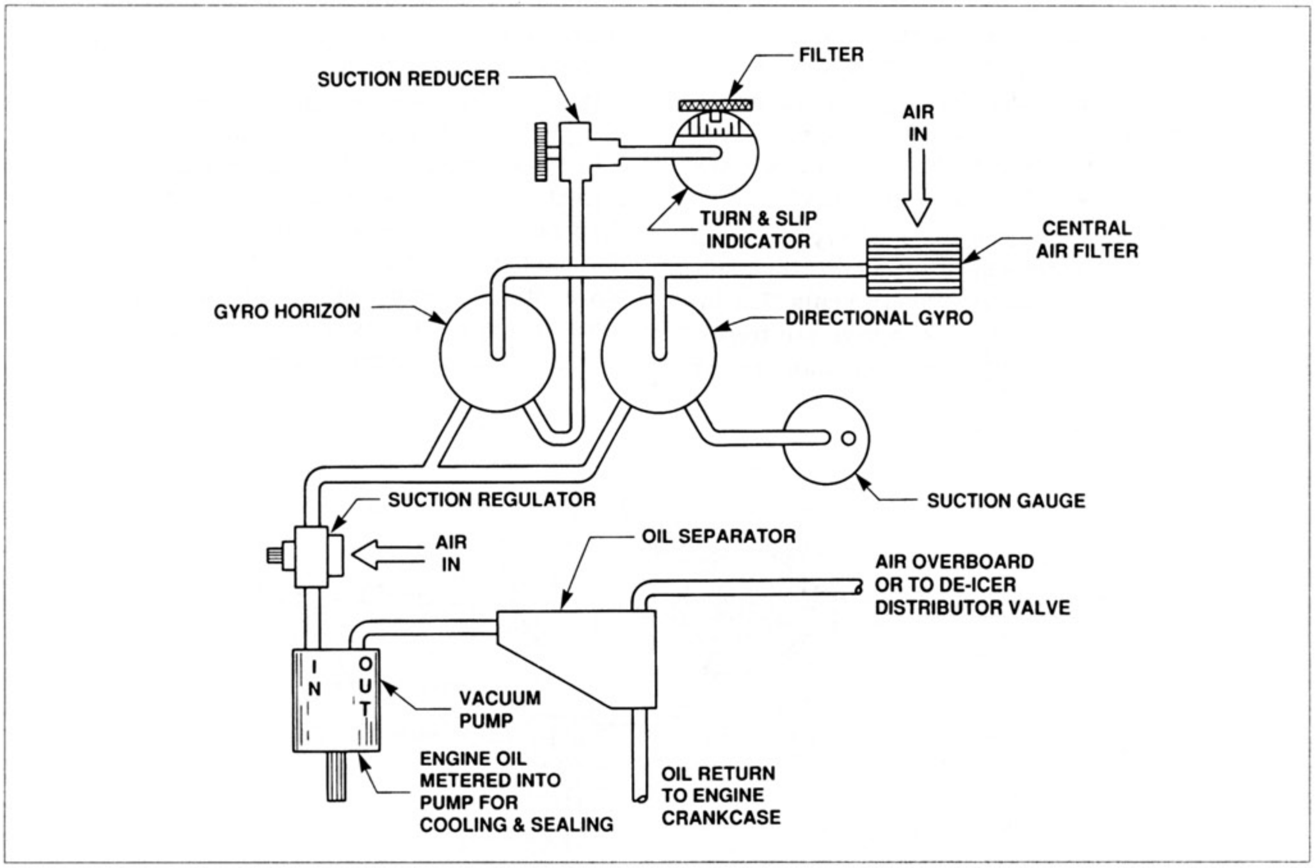
SUCTION REDUCER
FILTER
AIR IN
TURN & SLIP INDICATOR
CENTRAL AIR FILTER
GYRO HORIZON
DIRECTIONAL GYRO
SUCTION GAUGE
SUCTION REGULATOR
AIR IN
OIL SEPARATOR
AIR OVERBOARD OR TO DE-ICER DISTRIBUTOR VALVE
IN
OUT
VACUUM PUMP
ENGINE OIL METERED INTO PUMP FOR COOLING & SEALING
OIL RETURN TO ENGINE CRANKCASE

Figure 1-67. A wet pump vacuum system to operate three gyro instruments. A suction reducer is needed in the line to the turn and slip since it requires less vacuum pressure.

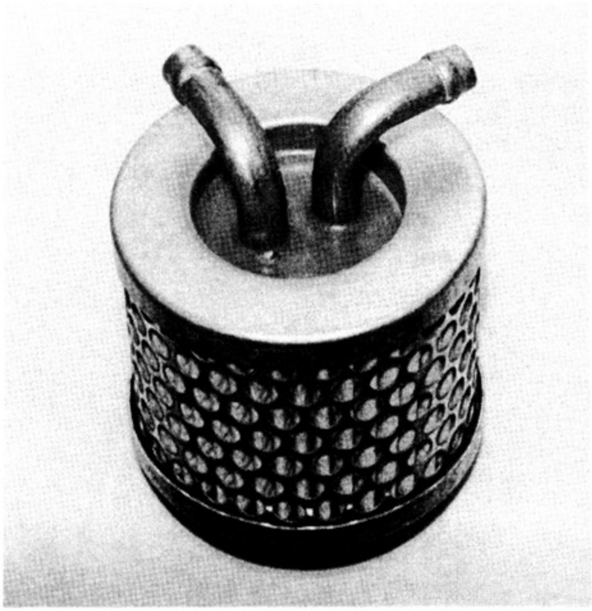

Figure 1-68. A typical replaceable filter used with air-driven gyro instruments.

6. Inspection and Maintenance of Gyro Systems

Some recommended practices for gyro system maintenance are:

1. Check the time it takes for the gyro instruments to come up to full speed and stabilize. This should normally be about 2–4 minutes.
2. Listen for unusual noise when the gyros are spinning. Noise is easier to detect after the engines are stopped.
3. When power to the gyros is removed, measure the run-down time. If there is a shortening of the normal run-down time, it indicates the bearings are getting worn or some other problem exists.
4. Check tubing and hose condition. They should not be worn or restricted. Check for kinks and dents.
5. Fittings should be in good condition and with wide radius bends. Do not over tighten.
6. Use only approved lubricants for fittings. Silicone spray is the most common recommendation.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

7. Route tubing carefully to avoid rubbing and abrasion.
8. If it becomes necessary to blow the lines to remove dirt or moisture, ensure that instruments are completely disconnected. Apply air pressure to instrument end of the lines.

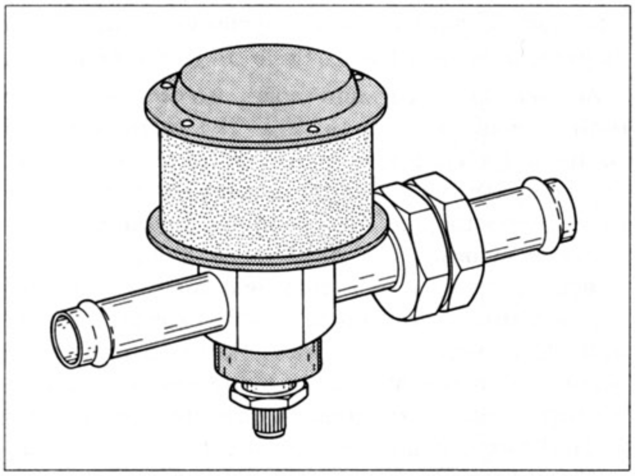

Figure 1-69. A vacuum regulator for gyro instruments includes an air filter.

9. Replace filters at recommended intervals—more often in dusty conditions or if smokers ride in the aircraft.

When installing additional air-driven gyro instruments or if a problem is suspected, the load on the pump should be evaluated. Each gyro instrument requires a certain volume of air which is stated in cubic feet per minute (CFM). Add up the requirement in CFM for all the instruments and ensure that it does not exceed the rated CFM for the pump. You must also evaluate the pressure drop requirements for the instruments and lines. Artificial horizons and directional gyros usually require 4.0–5.0 in. Hg. The turn and bank requires 2.0–2.5 in. Hg. The loss or pressure drop in all the lines and tubing should not exceed 2 in. Hg. If it does, you may have to use larger diameter tubing.

E. Compass Systems

1. Magnetic Compass

The aircraft magnetic direction indicator or compass is a completely independent instrument. It does not require any electrical or tubing connections. It contains a compass card with magnets that line up with

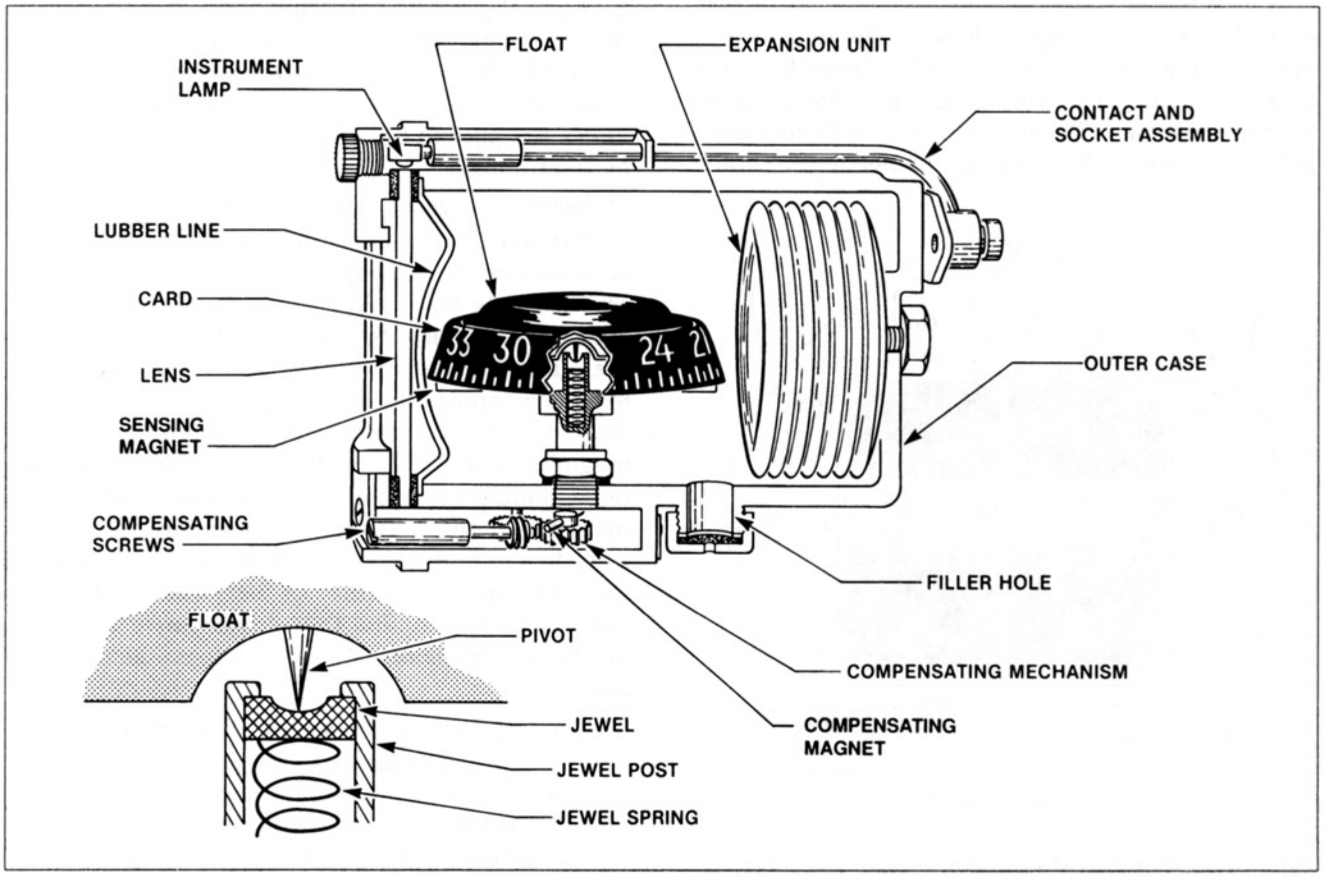
FLOAT
EXPANSION UNIT
INSTRUMENT LAMP
CONTACT AND SOCKET ASSEMBLY
LUBBER LINE
CARD
33 30 24 2
LENS
OUTER CASE
SENSING MAGNET
COMPENSATING SCREWS
FILLER HOLE
FLOAT
PIVOT
COMPENSATING MECHANISM
JEWEL
COMPENSATING MAGNET
JEWEL POST
JEWEL SPRING

Figure 1-70. The parts of a liquid-filled aircraft magnetic compass.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

the magnetic flux lines of the earth. Figure 1-70 shows the following basic parts of a magnetic compass.

1. A compass card or float which is mounted on jeweled bearings. It has numbers and direction markings so that the magnetic heading of the aircraft can be read from the instrument.
2. The case is filled with a light oil (usually refined kerosene) which dampens float motion and lubricates the bearings.
3. A diaphragm or bellows accommodates thermal expansion and contraction.
4. The compensator is two small moveable magnets used to adjust the compass for deviation error.
5. The lubber line is a marker against which readings are taken.

The face of a typical liquid filled magnetic compass is shown in figure 1-71. The indicated magnetic heading is 035°.

2. Compass Errors

The magnetic compass is subject to a number of errors which affect its operation. These include variation, deviation, acceleration error, northerly turning error and oscillation error. Variation error is simply the fact that a magnetic compass will give indications based on the magnetic north pole and not the north geographic pole. The normal grid lines on an aeronautical chart are in true directions based on the geographic poles and the equator. As can be seen in figure 1-72, the north magnetic pole is hundreds of miles from the north geographic pole. In most locations, there will be a difference between true and magnetic directions. This is variation. Figure 1-73 shows how the amount of variation is drawn on aeronautical charts for pilots to use. There are some locations where true and magnetic directions are the same. This would be along the line called the agonic line. Everywhere else the pilot would consult the variation markings on the map and add or subtract the appropriate number of degrees to convert from true to magnetic headings.

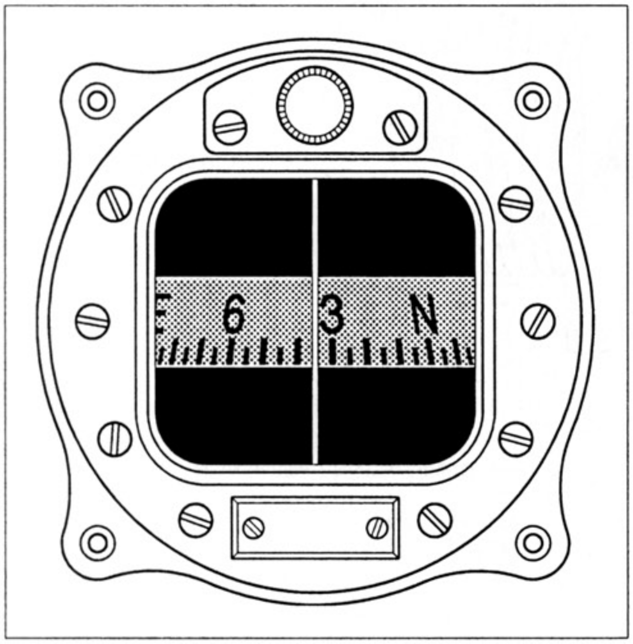

Figure 1-71. The face of a liquid-type magnetic compass.

Acceleration error and north turning error are both a result of compass dip. The earth is round so that at high latitudes in the northern hemisphere, the compass card will tilt downward toward the north magnetic pole. This compass dip causes both of these errors. If an airplane is flying east and it accelerates, the compass will momentarily indicate a turn to the north. If it decelerates, it will indicate a turn to the south. North turning error occurs when the aircraft is flying north or south. If a turn is made from a north heading, the compass will indicate a turn in the opposite direction momentarily and then it will lag behind the actual heading during the turn. Turns from south will cause the compass to lead the actual heading or indicate a higher than actual turning rate. Oscillation error is caused by the very delicate bearings in the compass. In rough air, the compass will oscillate back and forth 40°, 50° or more. The compass may never settle down as long as the turbulence persists. This forces the pilot to have to estimate the actual compass reading.

Deviation error is the most important one for maintenance technicians because they usually perform the checks and adjustments for deviation error. This error is also called magnetic influence error since it is caused by magnetic influences within the aircraft. All aircraft have some steel parts that may have some permanent magnetism. Most aircraft also have electrical circuits that can produce electromagnetic fields. Both of these can affect the magnetic compass and cause errors. The compensator magnets in the compass are used to adjust this error to a minimum. This process is called swinging the compass. It should be performed whenever equipment is installed that could cause a change or when a problem with the compass accuracy is suspected.

Swinging the compass—the basic procedure is:

1. Locate a compass rose on a ramp area which is accurate and can be used as a reference. A compass rose is a circle with magnetic directions indicated as shown in figure 1-74.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

2. Configure the aircraft for the checks by turning on electrical equipment and radios, running the engines and establishing a level attitude.
3. Set the compensators to zero (there are two little screws labeled N-S and E-W)
4. Point the aircraft north on the compass rose and adjust the N-S screw to zero error or as close to zero error as possible.
5. Point the aircraft east and adjust the E-W screw to zero error or as close to zero error as possible.
6. Point the aircraft south and remove half the error.
7. Point the aircraft west and remove half the error.

The process so far has averaged the error for all headings. Now you are ready to record the error.

8. Point the aircraft on all headings every 30°, and record the compass heading for each.
9. Prepare a placard which lists the deviation error at least each 30°. Place it on or near the compass and make a logbook entry.

The compass correction card is used to record the deviation error for the aircraft's compass. An example is shown in figure 1-75.

A newer type of compass is called the vertical card compass. It operates like the other types of

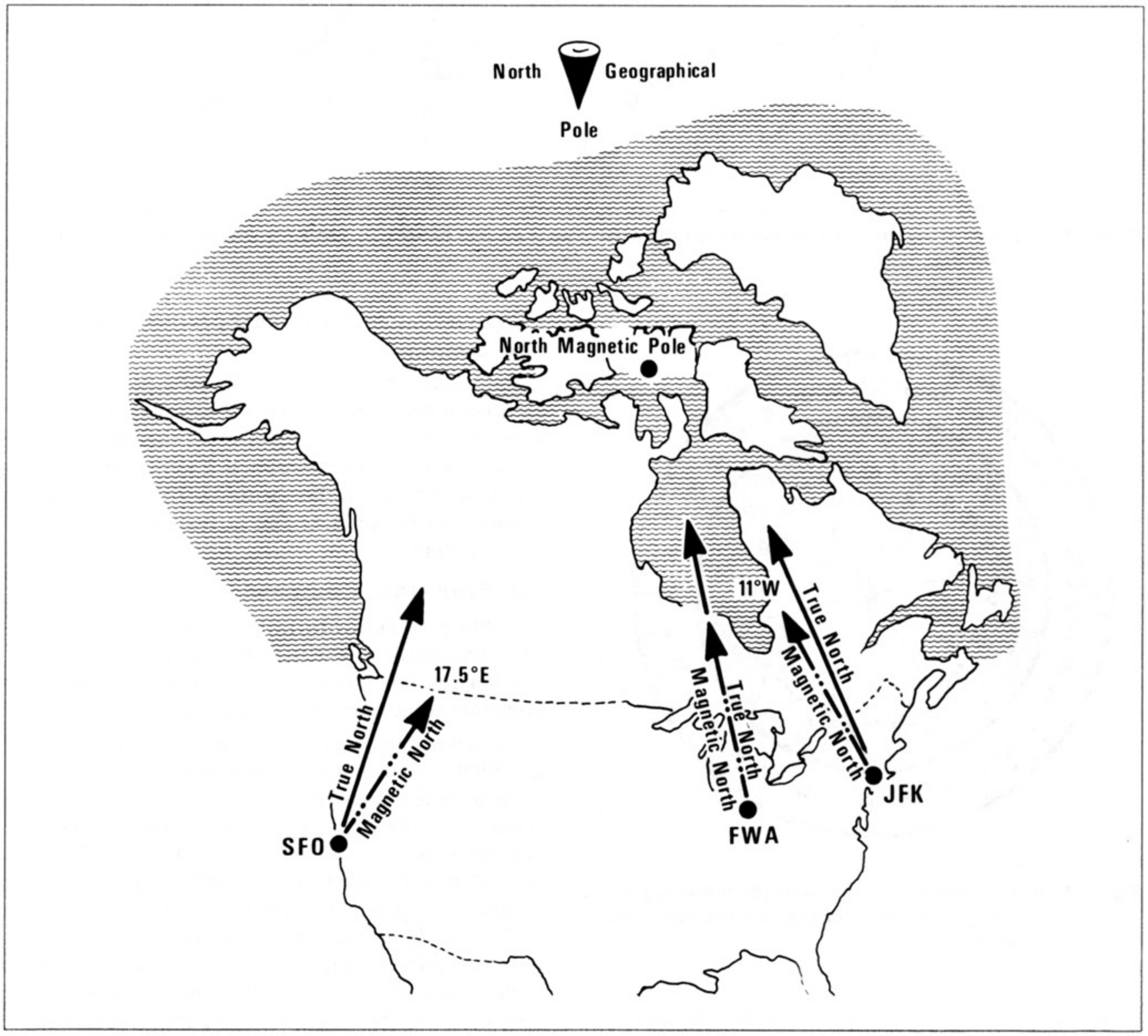

Figure 1-72. Variation error for an aircraft compass is the difference between true headings and magnetic headings. It is caused by the fact that the north geographic pole and the north magnetic pole are not in the same location.

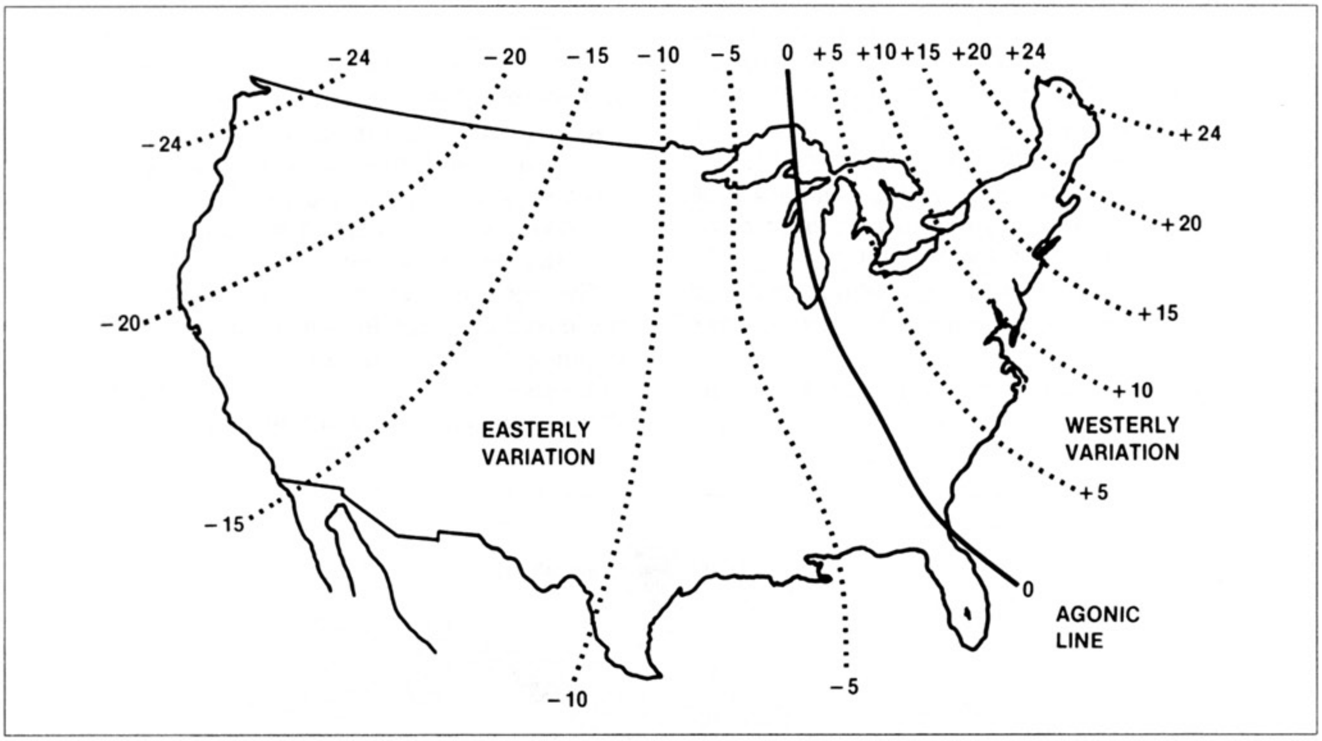
-24
-20
-15
-10
-5
0
+5
+10
+15
+20
+24
EASTERLY VARIATION
WESTERLY VARIATION
AGONIC LINE

Figure 1-73. Lines of variation are drawn on aeronautical charts so that pilots can apply the proper corrections during flight planning.

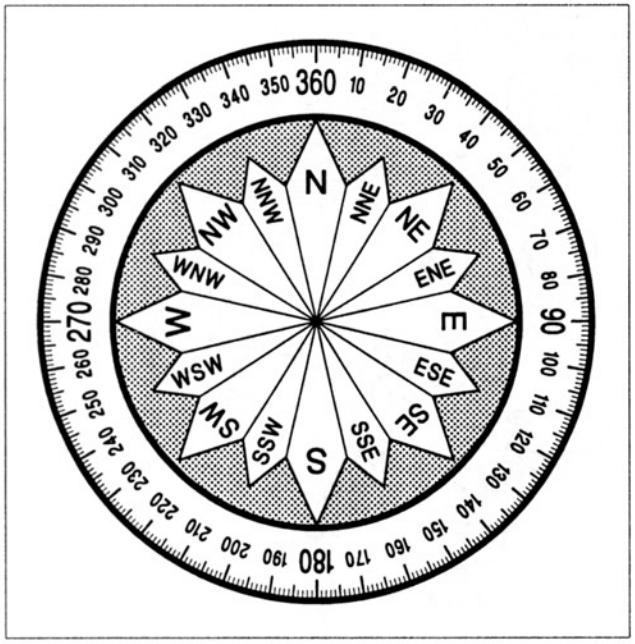
360
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
N
NNE
NE
ENE
E
ESE
SE
SSE
S
SSW
SW
WSW
W
WNW
NW
NNW

Figure 1-74. A compass rose contains the markings and numbers needed for magnetic direction references.

magnetic compass but the presentation on the face of the instrument is a full compass card which is easier to read. It sometimes eliminates the use of oil and employs eddy current damping. See figure 1-76 for an illustration of the appearance of a vertical card compass.

The full compass card presentation of the vertical card compass makes it easier to read. If the pilot wants to turn to a heading of 180°, the presentation makes it easier to determine if it is quicker to turn left or right to reach that heading. The compensator screws can be seen at the bottom of the vertical card compass.

3. Flux Gate Compass

The flux gate compass is a special type of remote mounted compass which is more stable than a standard magnetic compass and usually eliminates the problems of acceleration and north turning errors.

The sensor used with a flux gate compass system is called a flux valve or flux gate. It is a wheel shaped device made of a ferrous material with three spokes and the rim cut into three equal parts. The flux valve sensor can be seen in figure 1-77. The excitation coil is in the center and the pick up coils or output coils are installed with one on each spoke of the flux valve core. The excitation coil is supplied with AC current with a frequency of 400 Hz. It is designed so that when the current flow in the excitation coil is at peak value, the core material is saturated. When the current falls below peak value, the earth's magnetic flux lines cut across the pick up coils and produce an output signal in each one. The excitation coil in effect alternately blocks out the earth's magnetic field and then allows it to move across the output coils. This produces an AC output signal from each of the three output coils. Since the angle of the earth's flux lines to the flux valve changes for each different heading, the relative values of the three output voltages will be different for each different heading. This is illustrated in figure 1-78. An electronic component measures the three output signals and derives the magnetic heading of the aircraft.

FOR		N 0	3	6	E 9	12	15	S 18
STEER	RADIO							
	NO RADIO							
FOR			21	24	W 27	30	33	
STEER	RADIO							
	NO RADIO							

Figure 1-75. The compass correction card is usually made up by the mechanic when he swings the compass to determine the deviation error.

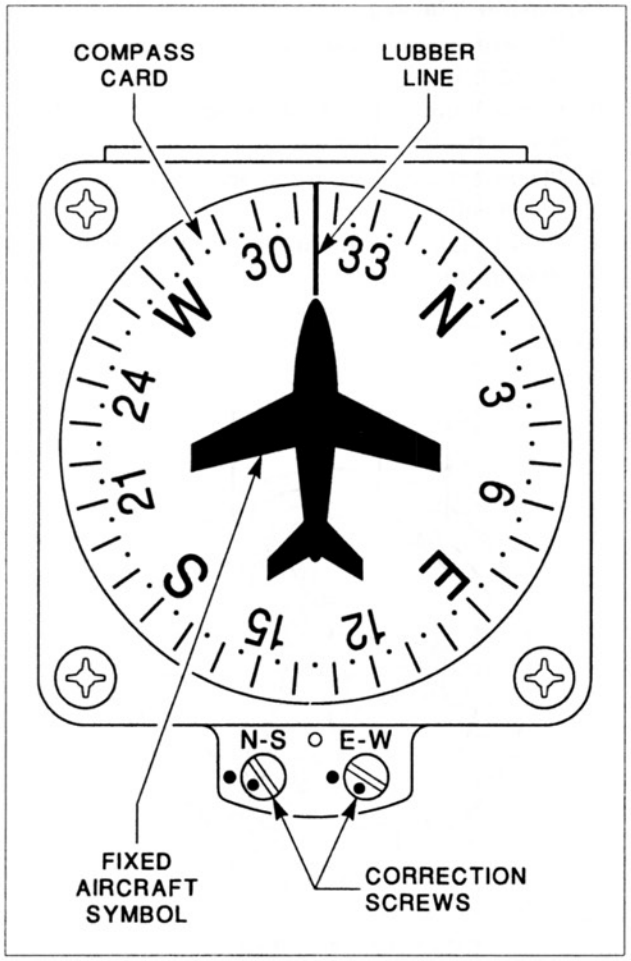

Figure 1-76. The vertical card compass displays a complete compass card and is easier to read than the older type. (Courtesy Canadair Group, Bombardier Inc.)

In order to give accurate readings, the flux gate sensor must normally be maintained in a level, horizontal position with respect to the Earth's surface. This leveling can be accomplished in one of two ways. In the first type of flux gate sensor found on aircraft, the sensor is suspended by a pendulous mechanism so that it can remain level when the aircraft attitude is changed. This type of flux gate has a housing filled with a light oil to dampen the motions of the moving parts. In the second type of flux gate compass system, the flux gate sensor is stabilized by a gyro system to keep it level. The output signals from a flux gate sensor are sent to an electronic unit which amplifies the signals and calculates magnetic heading. The output of this electronic unit is sent to cockpit indicators that require magnetic heading information and sometimes to navigational systems that require heading information. The two common cockpit instruments that receive signals from the flux gate system are the HSI and the RMI (radio magnetic indicator). The HSI and RMI can be seen in figure 1-79 which shows the flux gate compass system for a Challenger

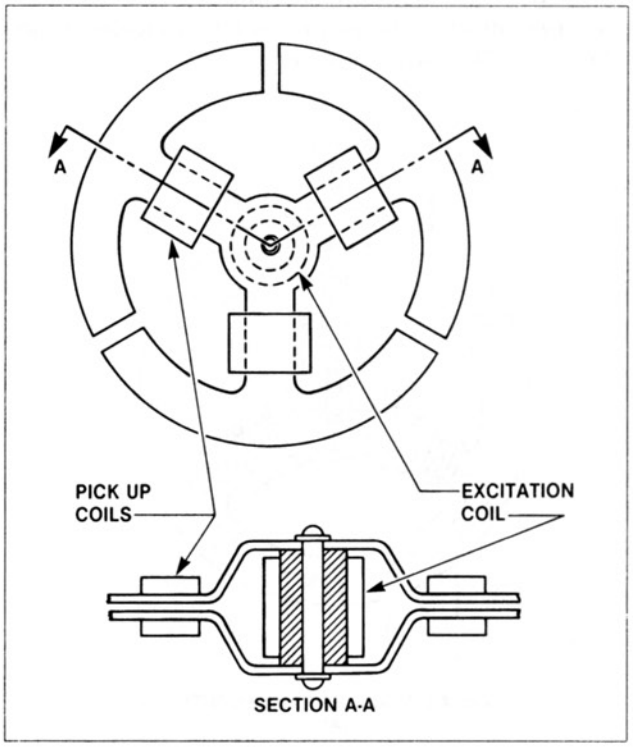

Figure 1-77. The flux valve sensor has an excitation coil in the center and three pick up coils on the spokes or arms.

airplane. The flux valve sensors themselves are normally installed near the wing tips to keep them away from magnetic influences in the aircraft. The location of the vertical card magnetic compass is also shown in figure 1-79.

The HSI and the RMI both have a compass card which indicates the magnetic heading of the aircraft. The heading information comes from a flux gate compass system. The compass cards on both instruments are driven by a remote mounted directional gyro. The DG receives signals from the flux gate compass that automatically reset it to the correct magnetic heading. The remote DG is slaved to the flux gate compass and the compass cards on the instruments are slaved to the remote DG. The connections are shown in figure 1-80 which is a diagram of a flux gate compass system. In this system, the pilot never has to reset the instruments with his magnetic compass in the cockpit unless the flux gate compass system fails.

The face of an RMI is shown in figure 1-81. The compass card in this instrument indicates the magnetic heading of the aircraft as previously described. The RMI can be identified by the two pointers that have a common pivot point in the center of the instrument. These pointers are connected to radio navigation systems so that they point toward the location of the ground transmitter. The selector switches allow each pointer to be connected to an ADF or VOR radio receiver.

4. FARs for Compass Systems

There are a number of FARs that relate directly to compass systems.

FAR 23.1327

Magnetic Direction Indicator

1. Must be installed to prevent influence by airplane vibrations or magnetic fields.
2. Maximum deviation in level flight is 10 degrees on any heading.
3. Magnetic non-stabilized may deviate more than 10 degrees due to electric heated windshield etc. if either a stabilized magnetic direction indicator or DG is installed. Deviation over 10 degrees requires a placard.

FAR 23.1547

Magnetic Direction Indicator
Deviation Placard

1. Placard must be installed on or near the MDI (compass).
2. Placard must list calibration for level flight with engines running.
3. Placard must state if calibration is for radios on or off.
4. Calibration increments must be 30 degrees maximum.

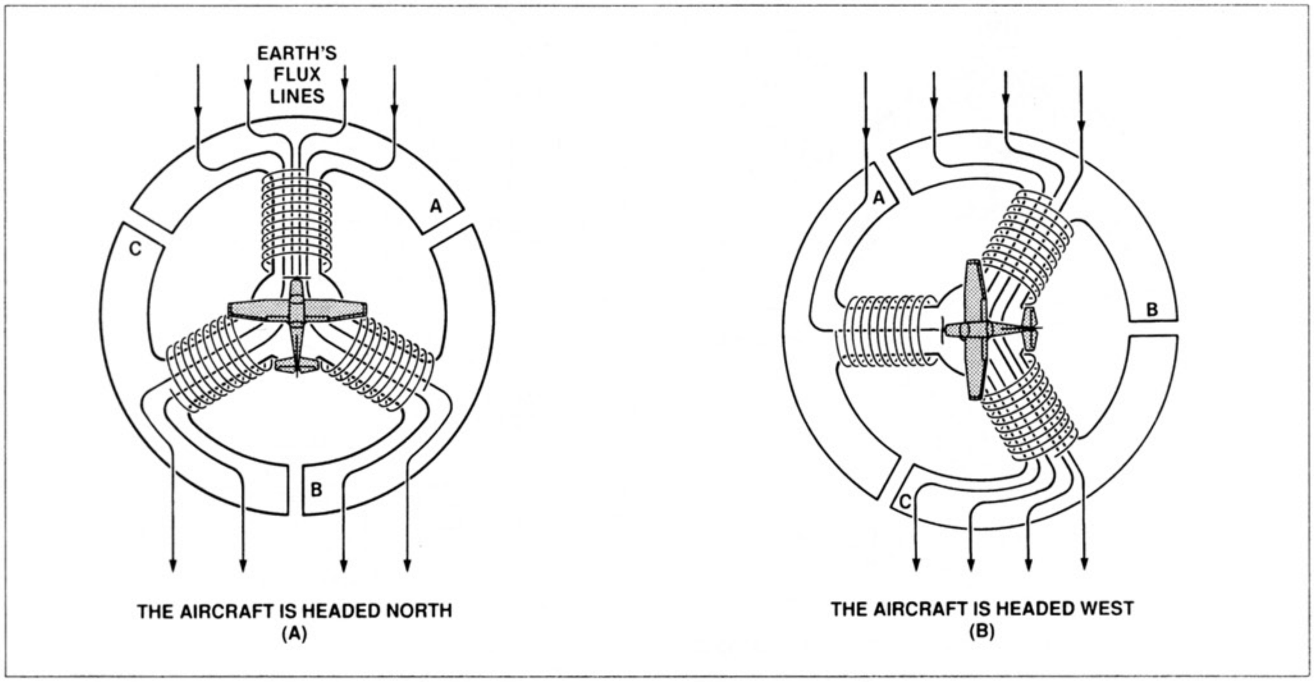

THE AIRCRAFT IS HEADED NORTH
(A)

THE AIRCRAFT IS HEADED WEST
(B)

Figure 1-78. The changing angle of the earth's flux lines to the flux valve produces a different output signal for each different heading of the aircraft.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

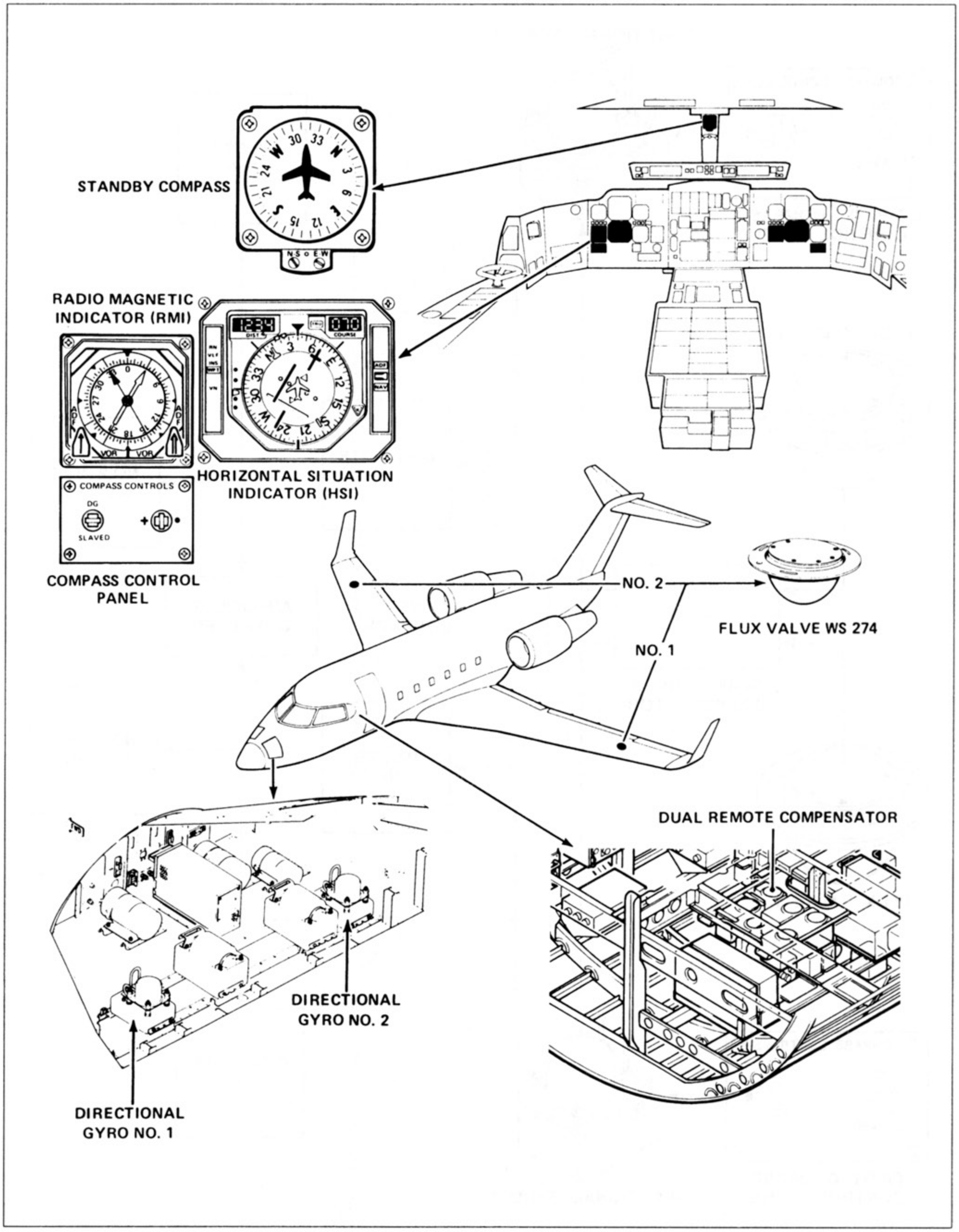
STANDBY COMPASS
NS
EW
RADIO MAGNETIC INDICATOR (RMI)
VOR
VOR
HORIZONTAL SITUATION INDICATOR (HSI)
DIST
COURSE
ADF
NAV
COMPASS CONTROLS
DG
SLAVED
COMPASS CONTROL PANEL
NO. 2
NO. 1
FLUX VALVE WS 274
DUAL REMOTE COMPENSATOR
DIRECTIONAL GYRO NO. 2
DIRECTIONAL GYRO NO. 1

Figure 1-79. The location of the various components of a flux valve compass system are illustrated in this drawing. The standby magnetic compass is also shown. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

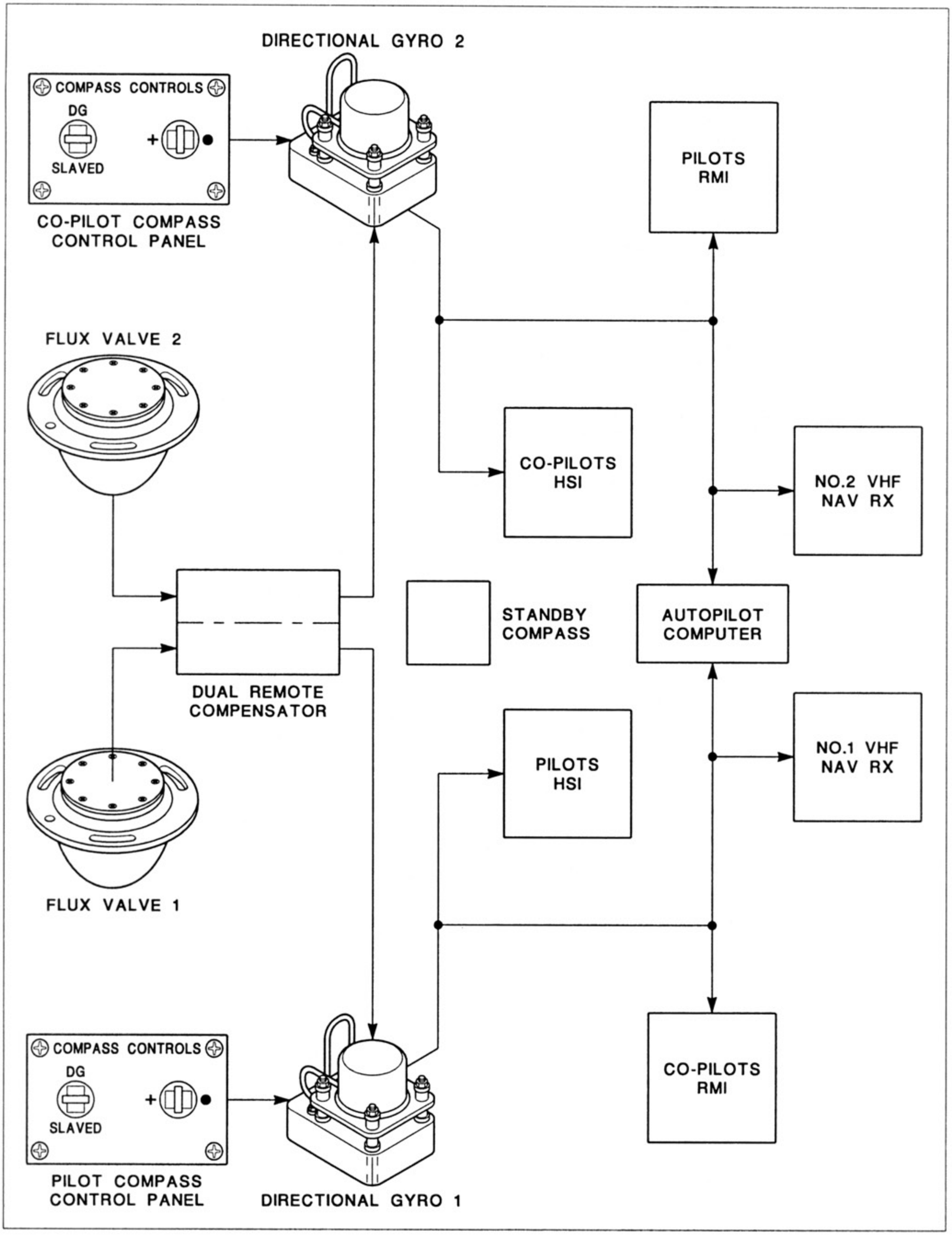
DIRECTIONAL GYRO 2
COMPASS CONTROLS
DG
SLAVED
CO-PILOT COMPASS CONTROL PANEL
PILOTS RMI
FLUX VALVE 2
CO-PILOTS HSI
NO.2 VHF NAV RX
STANDBY COMPASS
AUTOPILOT COMPUTER
DUAL REMOTE COMPENSATOR
PILOTS HSI
NO.1 VHF NAV RX
FLUX VALVE 1
COMPASS CONTROLS
DG
SLAVED
CO-PILOTS RMI
PILOT COMPASS CONTROL PANEL
DIRECTIONAL GYRO 1

Figure 1-80. The compass cards in the aircraft HSI and RMI instruments are operated by both the flux valve compass system and the remote mounted directional gyros. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

5. More than 10 degrees deviation for electric heated windshield etc. must be placarded.

F. Electronic Instruments

1. Basic Principles

The term "electronic instruments" is used to refer to the latest trend in aircraft instruments. This involves the use of CRTs (cathode ray tubes or TV screens) to display aircraft instrument information. Another common term for this system is the "glass cockpit". The use of CRTs permits a greater use of integrated instruments which display numerous types of information on one screen. It also permits greater flexibility because the method of displaying the information and the amount of information on each CRT can be changed in flight. It is also claimed that reliability is increased because complex electro-mechanical instruments are replaced by CRTs that have no moving parts. These CRTs are operated by a special type of computer control called a symbol generator. The latest generation of air carrier jets and bizjets was designed to use the glass cockpit displays. This group includes Boeing 757, 767 and 747-400; McDonnell Douglas MD-11 and Gulfstream G-IVs among others.

Other aircraft have been retrofitted with glass cockpit displays in their latest versions or as an option from the factory.

Some aircraft have only one or two CRTs, while others with a full glass cockpit system will use six or more CRTs. The electronic instruments that make up a full glass cockpit come in three types:

1. Electronic attitude director indicators (EADIs)
2. Electronic horizontal situation indicators (EHSIs)
3. Engine indication and crew alerting system (EICAS)

The appearance of the EADI and EHSI are very similar to the electro-mechanical versions that have been covered previously. The major difference is that the display is more versatile and the pilots can select what types of information they wish

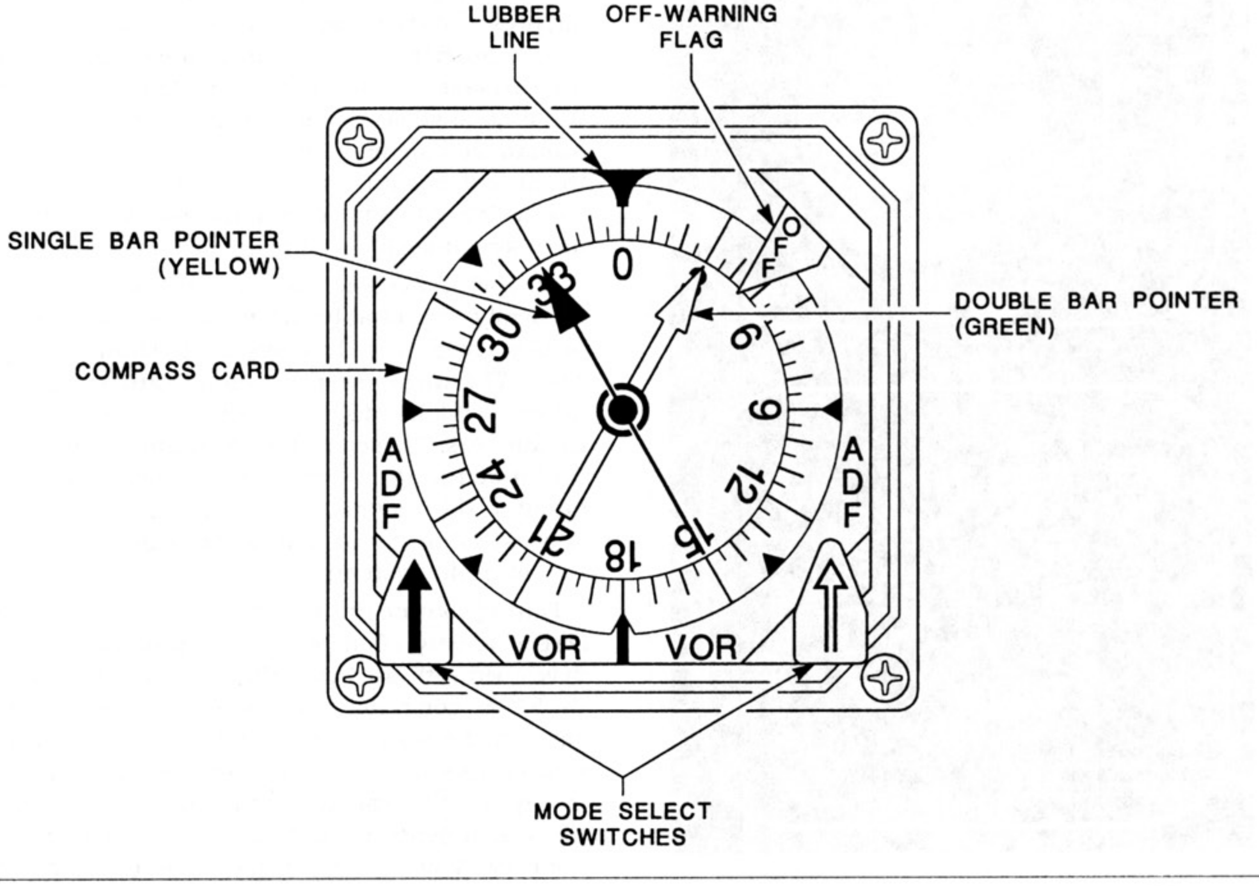

Figure 1-81. The radio magnetic indicator (RMI) has a compass card which indicates the magnetic heading of the aircraft. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

to see and much more information can be presented with the electronic version of the instrument. A typical EADI and EHSI are shown in figure 1-82. The EHSI can be set to a map mode which changes the appearance of the display to that of a map. An EHSI that is showing the map mode is found in figure 1-83. The map mode shows an airplane symbol along with navigational sites, airports and other features on the ground. The map mode presentation shows the aircraft moving across the map in correct relationship to locations on the ground. It is a very user friendly display which shows a large amount of information to the pilot in a way that makes it easier to read than more conventional displays.

2. EADI

Some of the information that can be presented on the EADI other than the basic pitch and roll data includes; radio glideslope data, radio localizer data, radar altimeter data, autopilot status and aircraft indicated airspeed.

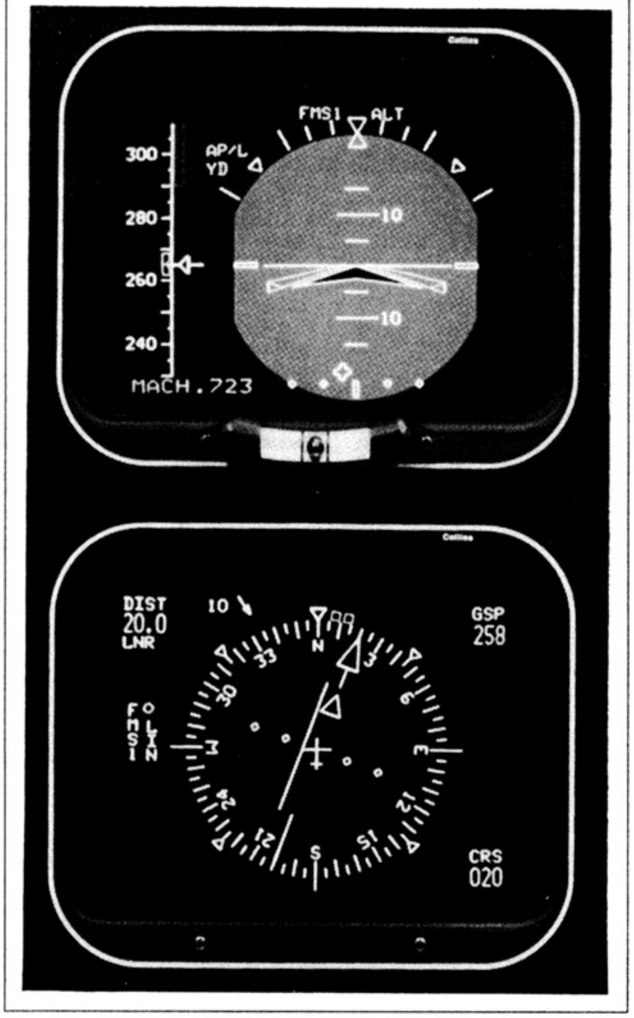

Figure 1-82. Electronic flight instruments: EADI on the top and an EHSI below.

3. EHSI

Information displayed on the EHSI includes: magnetic heading, radio steering commands for VOR/INS, radio glideslope data, DME radio data and weather radar data. The EADI and EHSI are installed directly in front of the pilots to match the standard "T" configuration.

4. EICAS

The EICAS system usually consists of two large CRTs installed in the middle of the instrument panel. The two CRTs may be arranged vertically or horizontally depending on the particular aircraft involved. The EICAS display screens in figure 1-84 are positioned one above the other. The EICAS system has two main types of information that are displayed, as the name implies. The engine indication function displays numerous powerplant instruments in standard columns depending on how many engines the aircraft has. The crew alerting system function consists of many sensors located throughout the aircraft that monitor all the major systems such as engines, electrical, hydraulic, bleed air, pressurization, etc. These sensors are monitored by computer and any faults or abnormal readings are displayed to the flight crew. On many aircraft these sophisticated monitoring systems replace a human crew member, the flight engineer. This allows the aircraft manufacturer to design a large airplane like a 747-400 that only requires two flight crew members instead of three.

An EICAS system that uses two CRTs stacked vertically is probably the most common and will be described here. This system is shown in figure 1-84. The upper screen has a standard presentation which displays the primary engine parameters. These are the most important engine instruments that are used to set power and monitor the engines. Also on the upper screen is a list of alert and status messages concerning the aircraft systems.

During routine cruise flight conditions, the lower screen is very often blank. If a problem suddenly developed with the hydraulic pressure, for example, the EICAS computer would automatically put a message on the upper screen and show the hydraulic system instruments on the lower screen. The basic theory of this system is that normal readings on the instruments do not have to be displayed for the crew. When an abnormal reading occurs, then it will be displayed to the crew. This reduces the workload for the two-man cockpit.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The primary engine parameters on the upper screen are in two identical columns because the airplane has two engines. The instruments displayed are engine pressure ratio (EPR), N1 tachometer and exhaust gas temperature (EGT). During engine starts, the EICAS system will automatically display the secondary engine parameters as shown here. The EICAS screens can also display additional information such as check lists. If an engine flames out during flight, a checklist is automatically displayed which shows the acceptable altitudes and airspeeds for an attempted restart as well as the checklist to accomplish this task. The EICAS system is complex and expensive so it has been installed only on the more sophisticated aircraft. The EADI and EHSI can be found on all classes of aircraft including small single engine airplanes.

5. Heads Up Displays

The glass cockpit instruments were made possible by the rapid advances made in microprocessors and digital computer technology. An even newer item of advanced cockpit displays is the Heads Up Display. The use of a HUD system was developed by the military for combat aircraft. If information about important aircraft systems is displayed in the windshield area, the pilot does not have to shift his attention down to the instrument panel to get this information. The HUD allows the pilot to keep looking out the windshield of the aircraft (head up) and to see the information that is needed projected onto a special screen in the windshield area. A heads up display system for commercial jet airplanes has been developed and is installed in some aircraft at this time. This system is called the Heads Up Guidance System (HGS) and the display is shown in figure 1-85.

The HGS screen itself is a special type of glass plate which the pilot can look through even while information is being projected onto the screen. The images on the screen are focused at infinity so the pilot does not have to refocus his eyes to look at either the world outside the windshield or the

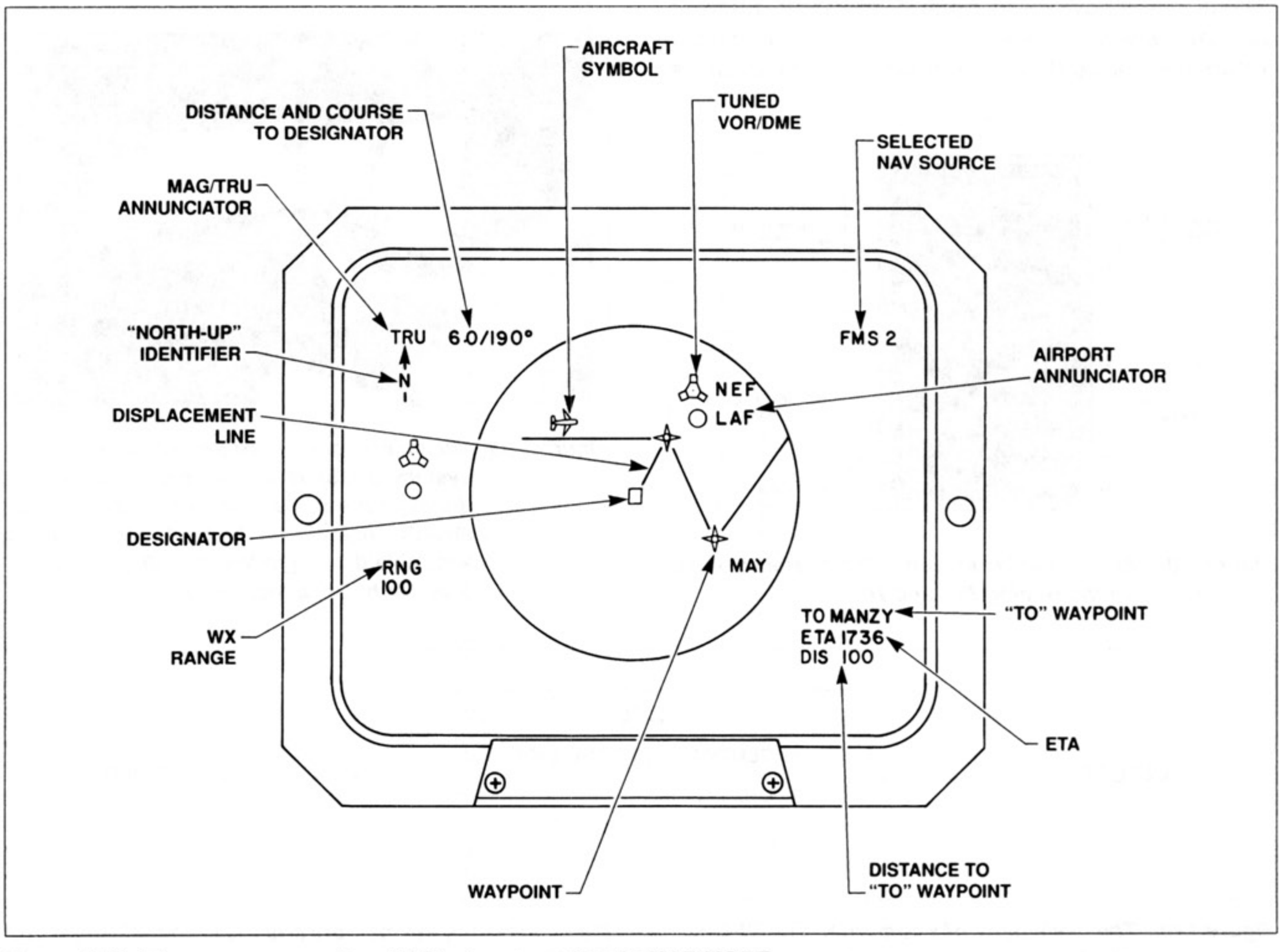
AIRCRAFT SYMBOL
DISTANCE AND COURSE TO DESIGNATOR
TUNED VOR/DME
SELECTED NAV SOURCE
MAG/TRU ANNUNCIATOR
"NORTH-UP" IDENTIFIER
TRU
6.0/190°
N
FMS 2
AIRPORT ANNUNCIATOR
NEF
LAF
DISPLACEMENT LINE
DESIGNATOR
MAY
RNG
100
WX RANGE
TO MANZY
ETA 1736
DIS 100
"TO" WAYPOINT
ETA
WAYPOINT
DISTANCE TO "TO" WAYPOINT

Figure 1-83. The appearance of an EHSI when operating in MAP MODE.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

data on the HGS. The HGS screen is operated by a computer controlled system which has many sensors to display different information. On a civil aircraft the information displayed on the HGS is the same kind of information displayed on an EADI. The first airline to begin using an FAA-approved HGS system was Alaska Airlines. They retrofitted their Boeing 727s with HGS at a cost of about $200,000 for each aircraft. The use of the HGS enables the airline to operate in bad weather conditions that might ground aircraft of other airlines. The FAA has approved this operation because of the elimination of the need for the pilot to switch his attention from the instrument panel to the view out the windshield. The future will no doubt see an increasing use of HGS and other advanced display systems by many other airlines and aircraft operators. Versions of the heads up display designed for use in twin-engine turboprop aircraft are already being developed by several companies.

G. Computers in Aircraft

The rapid advances in computer technology in recent years have been applied to many different aircraft systems such as cockpit displays, autopilots, navigational computers, engine controls etc. The aviation maintenance technician that works on modern aircraft should have a basic understanding of computers in general and their application to aircraft systems. The modern digital computer is made possible by the rapid advances in integrated circuits that have taken place over the last twenty years. A modern microprocessor is in effect a computer on a small chip of silicon. This small and powerful chip makes possible the manufacture of small but powerful computers.

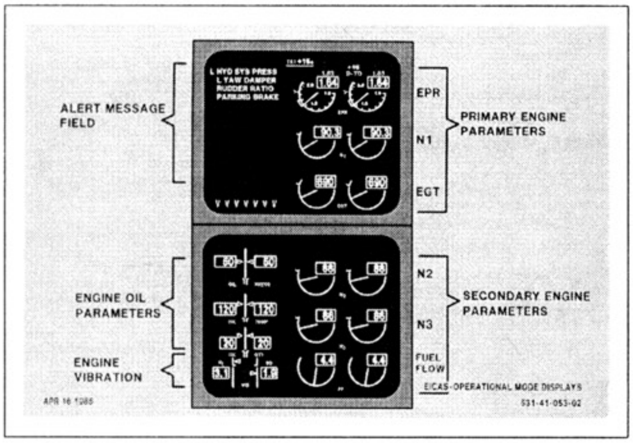
ALERT MESSAGE FIELD
L HYD SYS PRESS
L YAW DAMPER
RUDDER RATIO
PARKING BRAKE
EPR
N1
EGT
PRIMARY ENGINE PARAMETERS
ENGINE OIL PARAMETERS
ENGINE VIBRATION
N2
N3
FUEL FLOW
SECONDARY ENGINE PARAMETERS
EICAS-OPERATIONAL MODE DISPLAYS
APR 16 1985
531-41-053-02

Figure 1-84. EICAS display screens. This system is used on the Boeing 757 and 767.

1. Basic Parts of a Computer

A computer is made up of three basic parts as related to their functions. The basic hardware consists of input devices, output devices and the CPU (central processing unit). Refer to figure 1-86.

Input devices are things like keyboards, mice, scanners etc. Output devices are CRTs, printers, plotters etc. The central processing unit contains the brains of the computer. The CPU can be divided into three different units by their function. The central control unit directs data from one place to another and maintains overall control of the

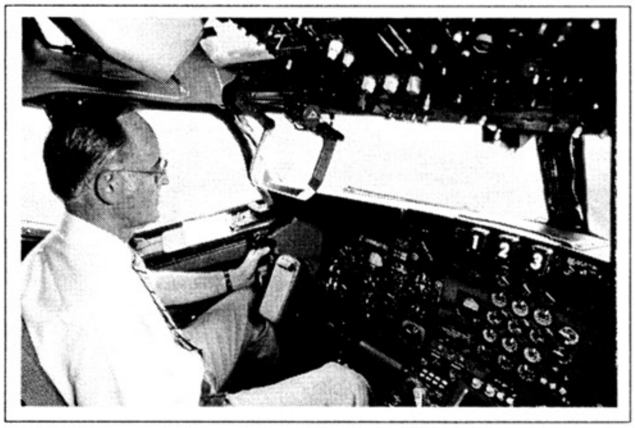

Figure 1-85. The display screen for a heads up guidance system (HGS) is a transparent plate that displays the same kinds of information as an ADI, but it allows the pilot to look out the windshield through the HGS display. (Courtesy Flight Dynamics, Inc.)

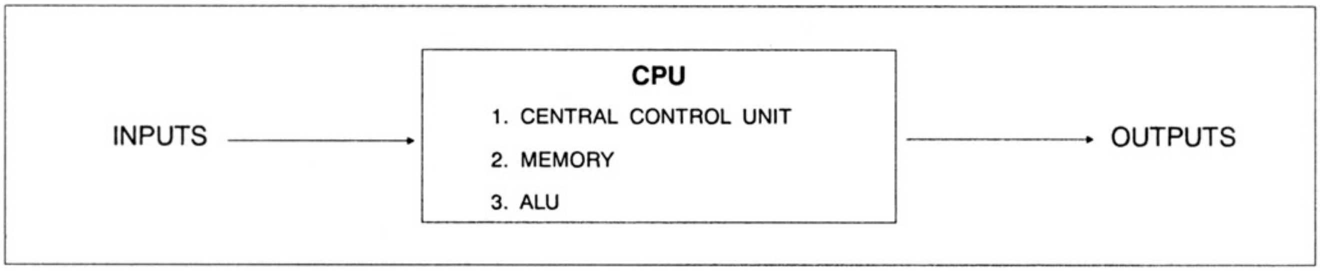
INPUTS
CPU
1. CENTRAL CONTROL UNIT
2. MEMORY
3. ALU
OUTPUTS

Figure 1-86. The basic parts of a computer. The CPU contains the control unit, the memory and the arithmetic logic unit which performs calculations.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

operations. The memory stores information on special computer chips. The ALU is the arithmetic logic unit; it performs the mathematical calculations that are required. The term peripherals is often used in discussing computers. Peripherals are the various input and output devices, examples of which were given above.

The memory of a computer comes in two types that are known as RAM and ROM. The RAM or volatile memory temporarily holds data that is being acted upon by the computer. It is called volatile because it is lost each time the computer is switched off. The operator can change and manipulate the RAM memory with keyboard entries and other actions. The ROM or non-volatile memory is sometimes called hard-wired. The data in the ROM area will not be lost when the computer is switched off and cannot be altered by a simple keyboard entry. An example of ROM is the built-in startup test that most computers have. When the computer is first switched on, it tests itself for errors and checks to see what peripheral devices are connected to it. An example of RAM could be a term paper that you are typing into a computer using a word processor program. If you forget to save the document on a disk, it will be lost when the computer is switched off.

2. Some Applications of Computers in Aircraft

A modern jet airplane may have many different computers that perform a variety of functions. The use of digital systems on aircraft is becoming more and more common because it offers several advantages:

1. Increased reliability
2. Faster response
3. Reduced power consumption
4. Smaller and lighter weight equipment
5. Lower operating cost

Computers have become so common that they are now used in many different aircraft systems such as autopilot, engine controls, navigation, flight planning, etc.

3. BITE Systems

One of the features of the effort to reduce operating cost is the use of BITE (built-in test equipment). The latest types of aircraft electronic equipment and computers have special types of test equipment as a part of the major units. BITE systems often provide three different kinds of tests that can be used to identify and correct faults.

1. Fault Detection — continuous during equipment operation
2. Fault Isolation — faulty equipment can be isolated or bypassed
3. Operational Verification After Defect Repair

The last example is a type of BITE program that maintenance personnel would use most often. After changing a piece of equipment which is thought to be the cause of the problem, a verification test can be conducted to ensure that the system is now operating normally. Running this particular test usually involves just pushing the appropriate button.

To simplify the troubleshooting and repair of modern electronic equipment, it is installed in the aircraft in the form of LRUs. An LRU is a line replaceable unit which means a standard size container which slides easily in and out of a special mounting rack. A typical arrangement for LRUs and equipment installations is shown in figure 1-87. The LRUs use standard types of electrical connectors and mounting attachments; this makes it easy to locate and change one in a short period of time. The BITE systems in an aircraft are designed to identify faulty LRUs so that they can be changed quickly and easily.

5. Digital Data Transmission

The increasing use of computers and sophisticated electronic devices on modern aircraft requires that these devices be able to communicate with each other rapidly and efficiently. This rapid exchange of data is accomplished with digital data transmission using

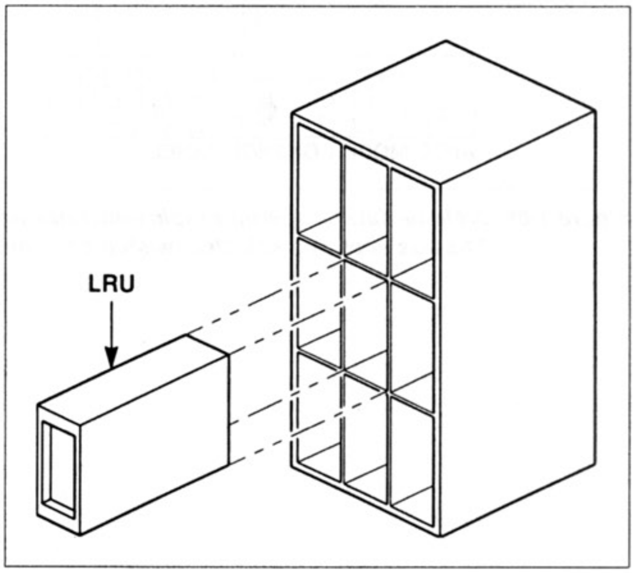

Figure 1-87. Avionics equipment in modern aircraft is installed in special racks that accommodate standard sizes of line replaceable units or LRUs. This makes changing the LRUs a quick and simple procedure.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

digital data busses. Since the various pieces of equipment that use this data bus are manufactured by many different companies, a uniform standard for the method of data transmission is needed. The standard which is used by this type of equipment on modern air carrier jets and bizjets is ARINC 429.

The initials ARINC stand for Aeronautical Radio Incorporated. This organization has been in existence since the 1930s to provide certain services to the airline industry.

The members of ARINC include the major airlines, aircraft manufacturers and equipment manufacturers. They establish many study groups that investigate emerging technology and suggest standards that can be applied to new types of equipment. When these standards are approved, they will be followed by all the members of ARINC. In the case of ARINC 429, this means that computers and similar equipment that utilize digital signals will be compatible with each other. Since the airlines often lead the way in the development of new types of equipment, the manufacturers of equipment for smaller aircraft often use ARINC standards also—even if they are not members of ARINC. Unlike FAA and FCC regulations, ARINC standards are not laws; but anyone who wants to sell airplanes or equipment to the major airlines will comply with these accepted industry standards. ARINC 429 has been used as an example here because it applies to digital information transmission systems used on aircraft. Many of the newer types of equipment described earlier such as EADI, EHSI, EICAS, BITE, etc. will use digital data exchange systems that are designed in accordance with ARINC 429.

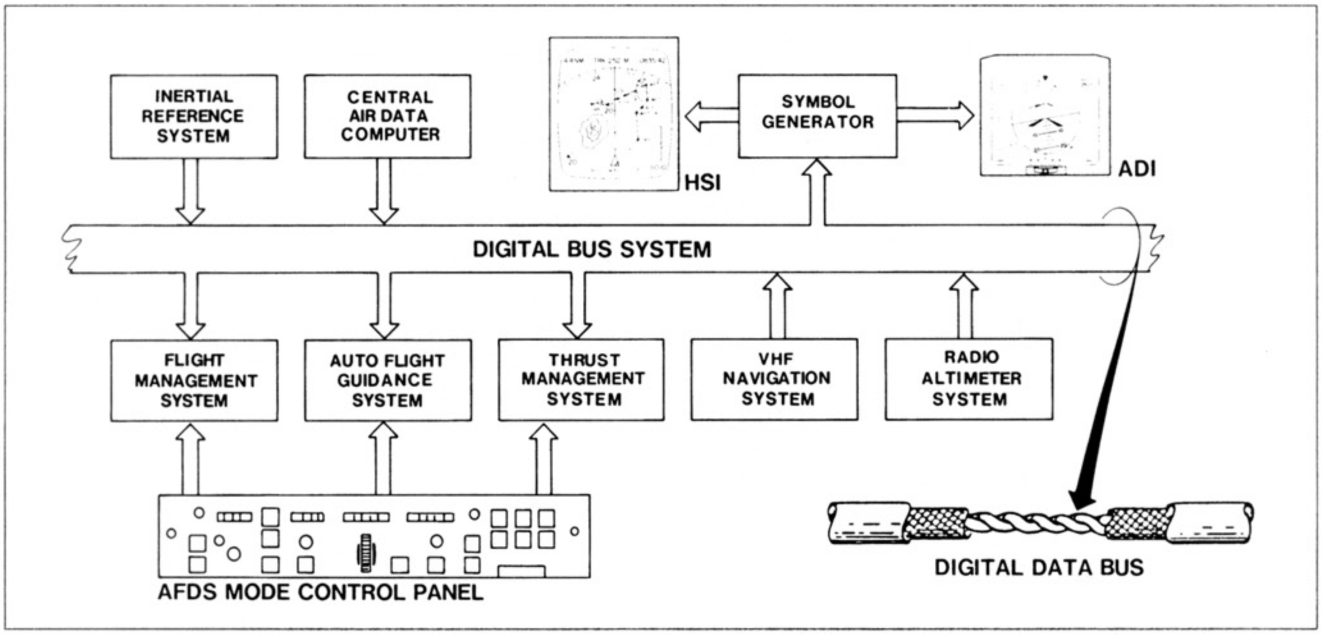
INERTIAL REFERENCE SYSTEM
CENTRAL AIR DATA COMPUTER
HSI
SYMBOL GENERATOR
ADI
DIGITAL BUS SYSTEM
FLIGHT MANAGEMENT SYSTEM
AUTO FLIGHT GUIDANCE SYSTEM
THRUST MANAGEMENT SYSTEM
VHF NAVIGATION SYSTEM
RADIO ALTIMETER SYSTEM
AFDS MODE CONTROL PANEL
DIGITAL DATA BUS

Figure 1-88. A digital data bus permits rapid transmission of data between the various electronic systems on the aircraft. The bus itself is a shielded twisted pair conductor which helps to prevent interference.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

CHAPTER II

Powerplant Instruments and Logic Gates

The information presented in chapter 2 will be in three major topic areas: powerplant instruments, logic gates and binary numbers, and position indicating and annunciator systems.

A. Liquid Quantity Measuring Systems

Depending on the type of aircraft involved, there may be just one or there may be many different liquids carried on the aircraft for which a quantity measurement is required. Most of the examples given here will be fuel quantity systems because they are the most common. It should be noted that for each type of system described, it could be used to measure fuel, oil, water, hydraulic fluid or some other liquid quantity.

Most small, single engine airplanes only have one liquid quantity indicating system in the cockpit and that is for fuel. The simplest types of fuel quantity systems are those that use mechanical systems and require no electrical power to give readings. These will be described first.

1. Sight Glass Gauges

The simplest kind of liquid quantity system is the sight glass gauge. In this system a small glass or plastic tube is connected into the tank so that the level of the liquid in the tube matches that in the tank. Markings on the tube itself or a plate behind it indicate the quantity. A sight glass gauge is shown in figure 2-1.

This type of quantity system has no moving parts, but the tank must be located in or near the cockpit area for it to be practical. It has been used on older aircraft for fuel and hydraulic fluid quantity systems.

2. Float-type Mechanical Gauges

A number of different kinds of mechanical float quantity systems have been used.

A very simple version utilized a float mounted on a metal rod which projected through a hole in the gas cap so that the rod would be visible from the cockpit. The fuel tank was located directly in front of the cockpit in the fuselage so that it could be easily seen. This type is illustrated by figure 2-2.

The float was often made of cork and it had to be coated with a special shellac or varnish so that it would not sink. Two disadvantages of this quantity system are that the rod tends to bounce up and down and there are no index markings at all. A variation of this system was used in many biplanes where the fuel tank was in the center section and the rod stuck down below the tank in a clear tube with an indicator fastened to it. This inverted float system with the indicator below the tank is shown in figure 2-3.

The gauge called the magnetic direct reading is a float-type gauge which uses a gear system to

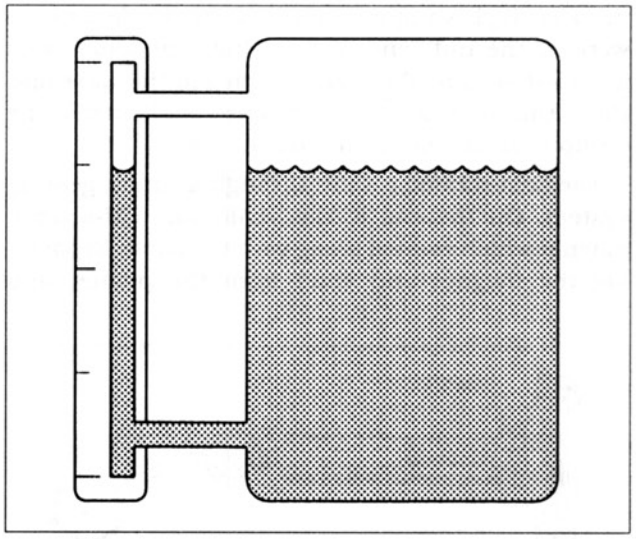

Figure 2-1. A sight glass liquid quantity gauge.

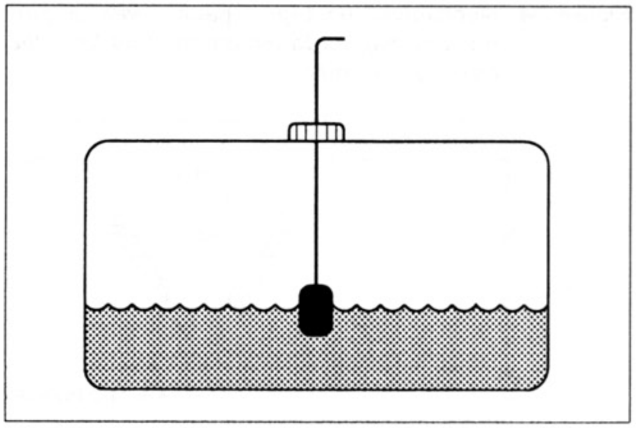

Figure 2-2. A mechanical float-type fuel quantity gauge for a fuselage tank.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

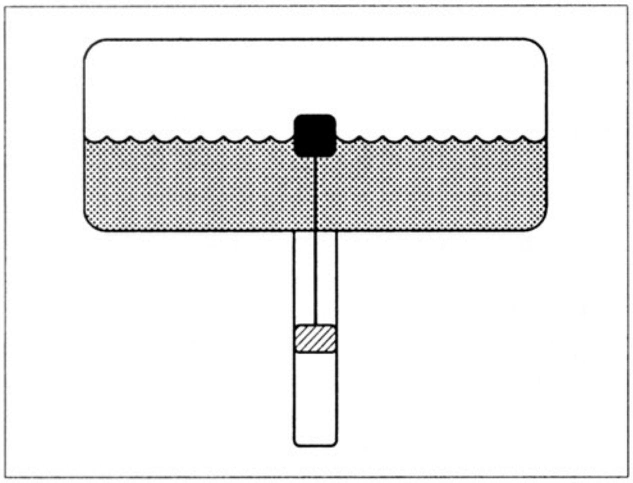

Figure 2-3. A mechanical float fuel quantity gauge for the center section of a biplane.

rotate a pointer in a round gauge and uses a magnetic principle to isolate the glass face cover and pointer from the fuel. This type of gauge was often used on high wing airplanes where the fuel tanks were in the butt end of the wing. The gauge was installed so that the float was inside the tank and the round face of the gauge was visible inside the cockpit. It is shown in figure 2-4.

The float rotated a shaft through a simple gearing system. On the end of the shaft was a U-shaped magnet which rotated along with the shaft. Separating the magnet and shaft from the pointer and the rest of the gauge in the cockpit was a piece of aluminum with sealing gaskets. The magnetic flux traveled through the aluminum and rotated a piece of ferrous metal that in turn rotated the pointer.

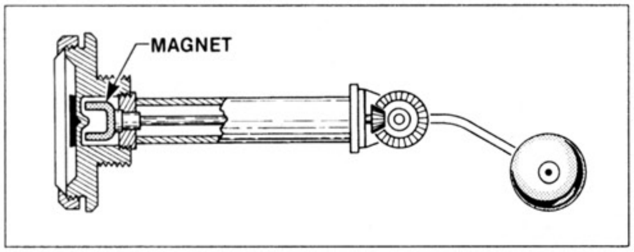

Figure 2-4. Mechanical float-type gauge with a permanent magnet to isolate the fuel from the cover glass and pointer.

3. Resistance Gauges

The type of fuel quantity gauge most common on modern small airplanes is similar to the kind used in cars. It has a float in the tank that moves a variable resistor. The variable resistor alters the current flow in a DC circuit to operate a meter movement that is somewhat similar to those used in voltmeters and ammeters.

The gauge used with the float operated variable resistor is most often the ratiometer type seen in figure 2-5. This gauge uses two opposing magnetic fields so that the pointer reacts to the ratio of current flow in the two sections. In this way, it is less affected by fluctuations in system voltage caused by voltage regulator settings or a weak battery.

4. Underwing Fuel Quantity Indicators

Many large aircraft have two totally different types of fuel quantity measuring systems. One of these operates the cockpit gauges and the other is an underwing fuel quantity system. The underwing system can only be used on the ground and is most often employed by maintenance and service personnel rather than by the flight crew. There are three kinds of underwing fuel quantity systems, but they all share certain features in common. They all utilize a fuel quantity stick of some type which can be extended below the bottom surface of the wing. They measure the fuel quantity in terms of volume and not mass. These underwing quantity systems typically require no electrical power to obtain readings. This last characteristic would be an advantage if it was necessary to take fuel readings while working on the fuel system.

The oldest and simplest type of underwing fuel quantity system is called a drip stick. It uses a

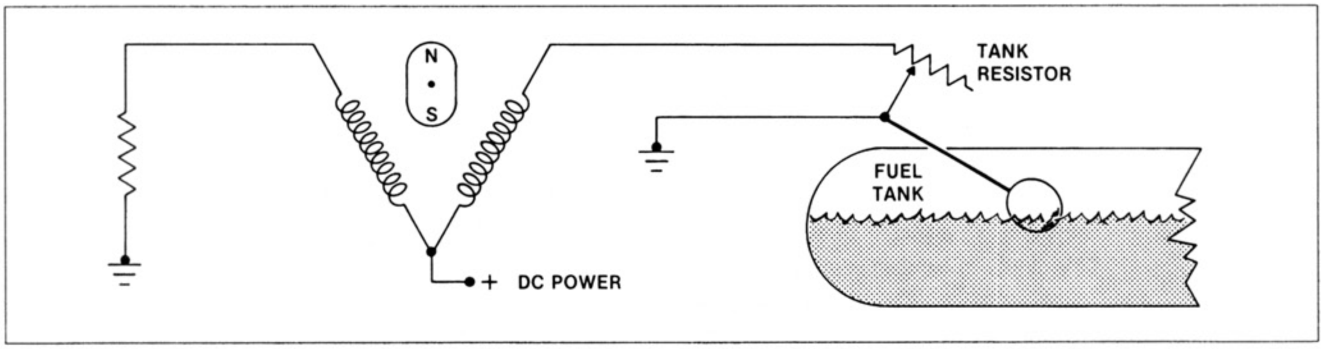

Figure 2-5. Ratiometer fuel quantity system using a float-operated variable resistor.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

hollow tube which extends from the bottom of the wing up inside the fuel tank as illustrated in figure 2-6. The tube is normally stowed by being pushed up inside the tank until the bottom of the tube is flush with the bottom surface of the wing where it is latched in the closed position. In order to obtain a reading, the tube is unlatched and pulled down until the upper end reaches the top of the fuel level. When fuel begins to drip out the bottom of the tube, a reading is taken using the markings on the outside of the tube. This type of fuel quantity system is not used on modern aircraft because of the fire danger when fuel is allowed to drip on the ground or hangar floor.

Another type of underwing fuel quantity stick is the one which uses a clear Lucite® plastic rod. The main features of this device are shown in figure 2-7. The rod is made of clear plastic because it obtains readings by transmitting light along the rod. The principle involved is the refraction of light. Fuel and air have different light refraction characteristics and if a specially shaped quartz tip is installed on the top of the rod, it will produce a particular light pattern when it is positioned at the top of the fuel level. In order to take a reading, the tube is pulled down from the bottom of the wing until the light pattern on the bottom of the rod is focused to a point of light. The reading is then taken using the markings along the length of the rod.

The most common type of underwing fuel quantity stick on modern aircraft is the one which uses a float inside the tank that has a magnet fastened to it. The upper end of the stick has a magnet which will attract the float magnet when they are in alignment. This type of underwing fuel quantity stick is shown in figure 2-8.

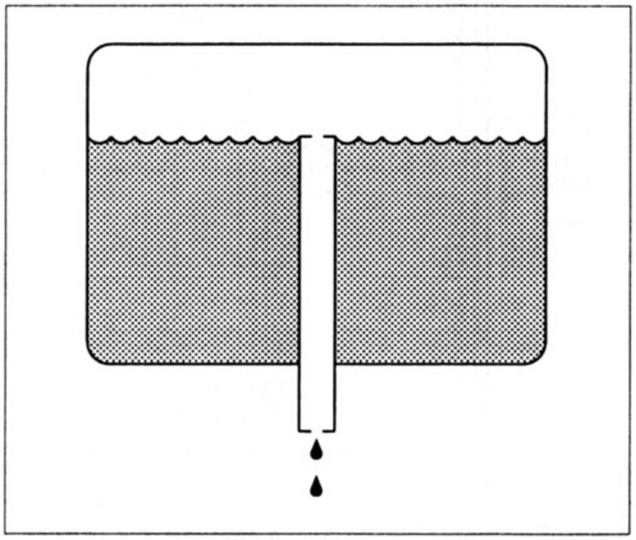

Figure 2-6. Older style drip stick underwing fuel quantity gauge.

A fuel reading is taken by unlatching the stick and pulling it down until the float and the top

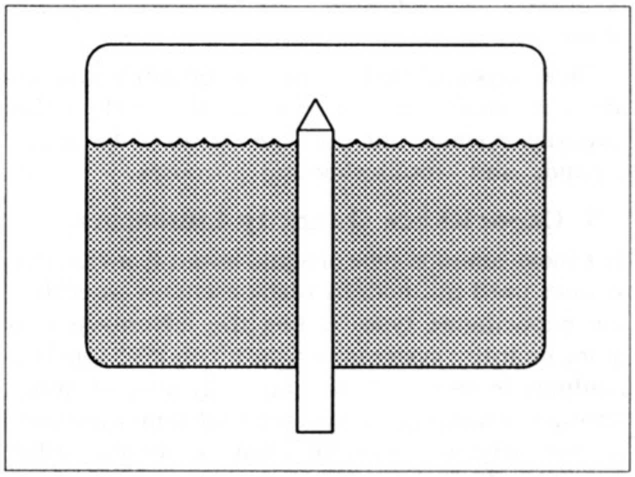

Figure 2-7. An underwing fuel quantity measuring stick that utilizes a clear Lucite rod.

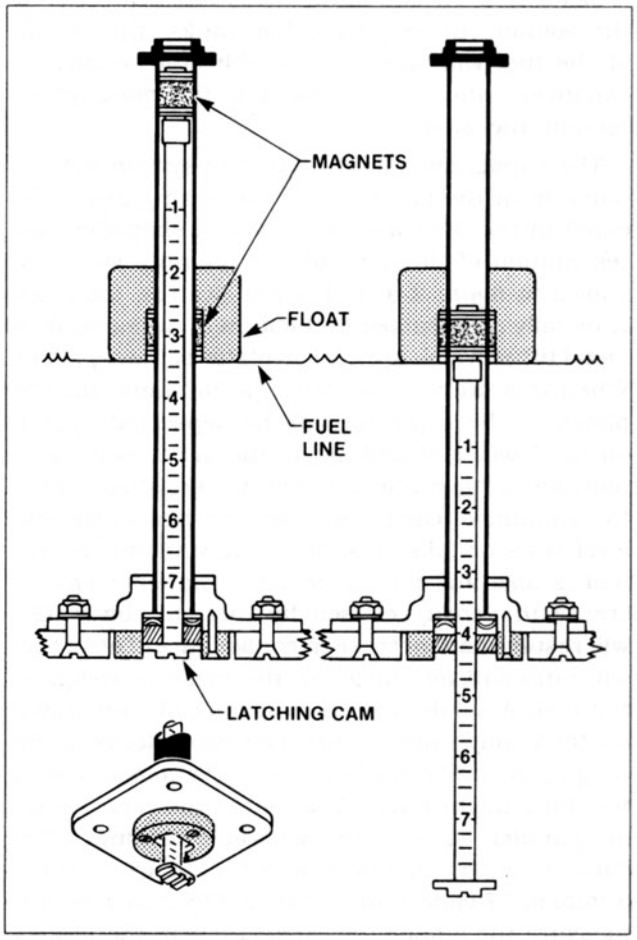

Figure 2-8. Underwing fuel quantity stick which employs a float and permanent magnets.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

of the stick are held in position by magnetic attraction. The reading is then taken from the markings along the tube. The magnetic force is not strong enough to lift the float out of the fuel; so when the stick is pushed up, the magnetic attraction is broken and the stick can be pushed up and stowed.

These types of underwing fuel quantity systems are not usually as accurate as the cockpit fuel quantity system, but they can be used for maintenance and troubleshooting purposes.

5. Capacitance Quantity Indicators

The most common type of liquid quantity measuring system used on modern turbine engine aircraft is the capacitance type. It has the advantage over other quantity systems in that it can give accurate readings in very large or unusually shaped tanks. Another advantage is the fact that liquid quantity is measured in terms of mass or weight rather than in volume. Measuring fuel quantity in mass is especially useful with large turbine engine aircraft because the power produced by the engines is more a factor of the mass of fuel consumed rather than the volume. In very large fuel tanks, the volume of the fuel will vary considerably due to thermal expansion and contraction, but the mass would remain the same.

The capacitance liquid quantity system gets its name from the fact that the measuring probes located in the tank are capacitors. A simplified representation of this type of fuel quantity system is shown in figure 2-9. In the real system, the probe is usually constructed in the form of two concentric metal tubes which are the two plates of the capacitor. When this probe is located in a fuel tank, the two plates of the capacitor will be separated by fuel on the lower end and air on the upper end. Since fuel and air have different dielectric constant values, the amount of capacitance will change as the fuel level rises or falls. The dielectric constant for the fuel is also affected by density. Therefore any increase in density caused by thermal contraction will result in an increase in capacitance. The probes will automatically measure the mass or weight of the fuel. A small, symmetrical tank like an engine oil tank may only require one capacitance probe to give accurate readings. A large, tapered wing fuel tank might have 15 or more probes connected in parallel to ensure accurate readings. The capacitance fuel quantity system of figure 2-10 has a total of 17 capacitance probes. Electronic circuits measure the amount of capacitance in the probes, apply any needed corrections, and send electrical signals to the cockpit gauges to indicate the fuel quantity in pounds. Capacitance fuel quantity systems usually include a totalizer. The totalizer gives a reading of the total fuel on board the aircraft. Some fuel systems will also give the fuel used since takeoff.

B. Fuel Flow Indicators

There are a number of different gauges which might be used for aircraft fuel systems depending on the type and complexity of the particular kind of fuel system used. All powered aircraft will have a fuel quantity system. On small airplanes with gravity flow fuel systems, this would be the only type of fuel system instrument required. Aircraft with pump fed engines will need a fuel pressure gauge in addition to fuel quantity. Aircraft with fuel-injected or large radial piston engines and aircraft with turbine engines will usually have a fuel flow instrument. Some aircraft with turbine engines that operate in cold temperatures will also have a fuel temperature indicator to guard against the danger of ice crystals in the fuel. The various types of fuel flow indicator systems will be described in this section.

1. Fuel-injected Engine Flowmeters

The type of flowmeter commonly installed on aircraft with fuel-injected reciprocating engines is not a true flowmeter at all. The sensor used with this instrument system actually measures pressure not

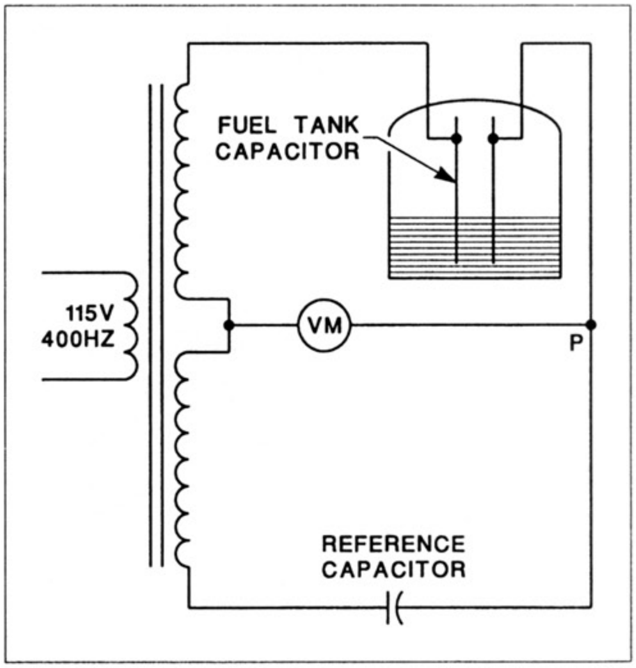

Figure 2-9. ***Simplified circuit to illustrate the principle of operation for a capacitance liquid quantity system.***

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

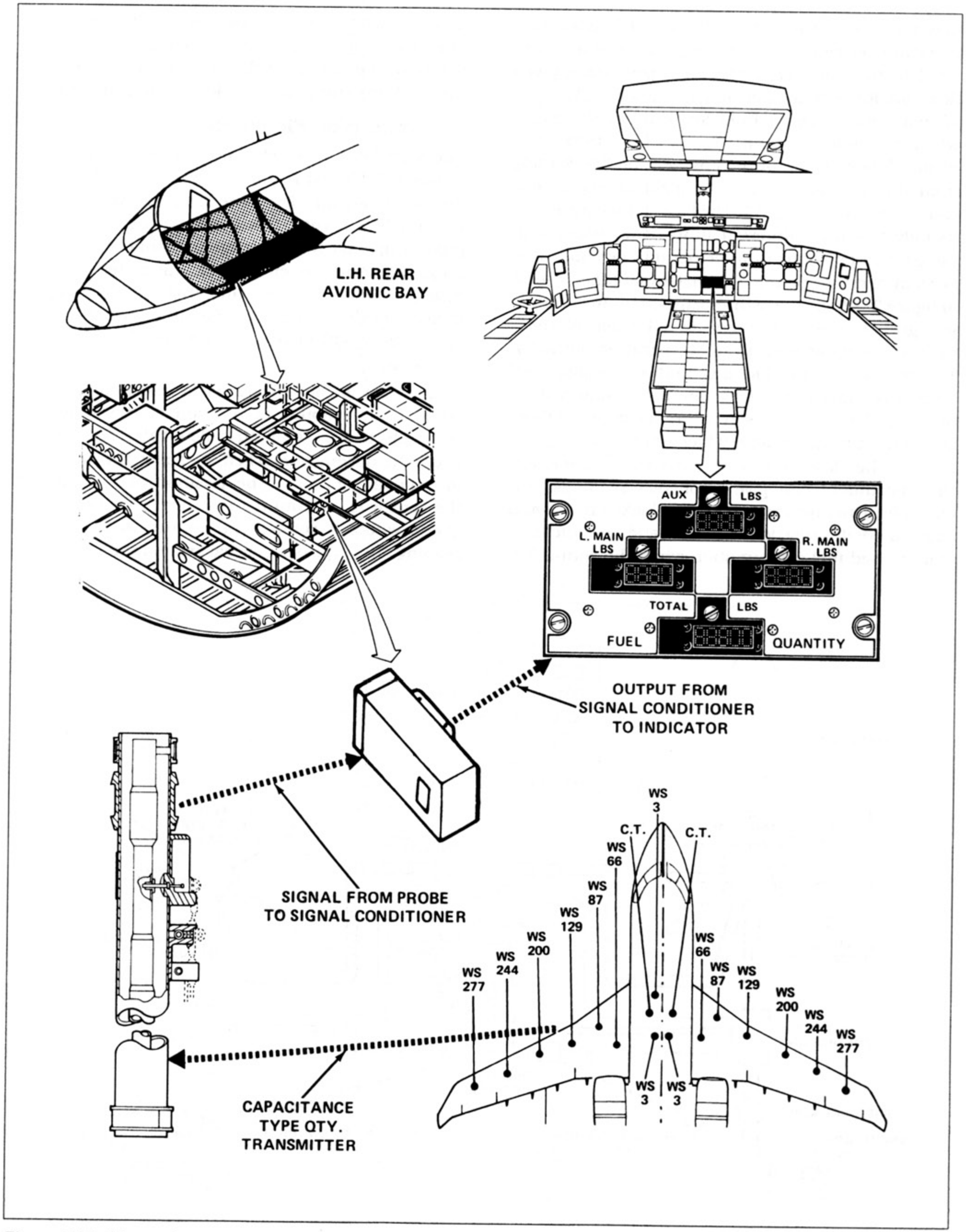
L.H. REAR
AVIONIC BAY
AUX
LBS
L. MAIN
LBS
R. MAIN
LBS
TOTAL
LBS
FUEL
QUANTITY
OUTPUT FROM
SIGNAL CONDITIONER
TO INDICATOR
SIGNAL FROM PROBE
TO SIGNAL CONDITIONER
WS 3
C.T.
C.T.
WS 66
WS 87
WS 129
WS 200
WS 244
WS 277
WS 66
WS 87
WS 129
WS 200
WS 244
WS 277
WS 3
WS 3
CAPACITANCE
TYPE QTY.
TRANSMITTER

Figure 2-10. The major components and their location for a capacitance fuel quantity system on a corporate jet. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

flow rate. Since the injector lines and nozzles have a certain restriction to flow, a given pressure supplied to the injection system will produce a given flow rate for normal operating conditions. This type of instrument uses a Bourdon tube in the gauge which is connected by tubing and hose to the fuel divider block on top of the engine. The line leading from the fuel distribution manifold on the engine can be seen in figure 2-11. The face of the instrument usually has three different units of measurement: PSI, gallons per hour (GPH) or pounds per hour and percent of cruise power. The instrument face shown in figure 2-12 has these three units. This instrument will give accurate readings for all three of these values as long as everything is operating normally. The pressure at the fuel distribution manifold will be proportional to the flow rate if the total restrictions to flow in the system are normal. If there is any fault in the system which causes the restriction to the flow of fuel to increase or decrease, the instrument can give erroneous readings. For example, if an injector nozzle was blocked this would cause a greater restriction to flow and an indication of increased fuel consumption when the actual fuel flow rate would be decreased. A leak in an injector line would decrease the restriction to flow and decrease the indicated flow rate on the gauge but the fuel consumption would actually increase.

2. Vane-type Flowmeters

The vane-type flowmeter uses a sensor like that in figure 2-13 that is installed in the line that feeds fuel to the engine. The vane is mounted on a shaft so that it will rotate through an arc as the fuel pushes against it. The circular chamber that contains the vane has enough clearance between the cylinder walls and the vane that the flow of fuel is not retarded to any significant degree. The vane type sensor will measure the volumetric flow rate of the fuel.

The vane is rotated against a restraining spring so that the amount of rotation of the vane corresponds to the volumetric flow rate. The cockpit gauge is normally marked to show the flow rate in gallons per hour. The position of the vane in the sensor is transmitted electrically to the cockpit gauge where it rotates the pointer to the correct reading. The type of electrical system that transmits

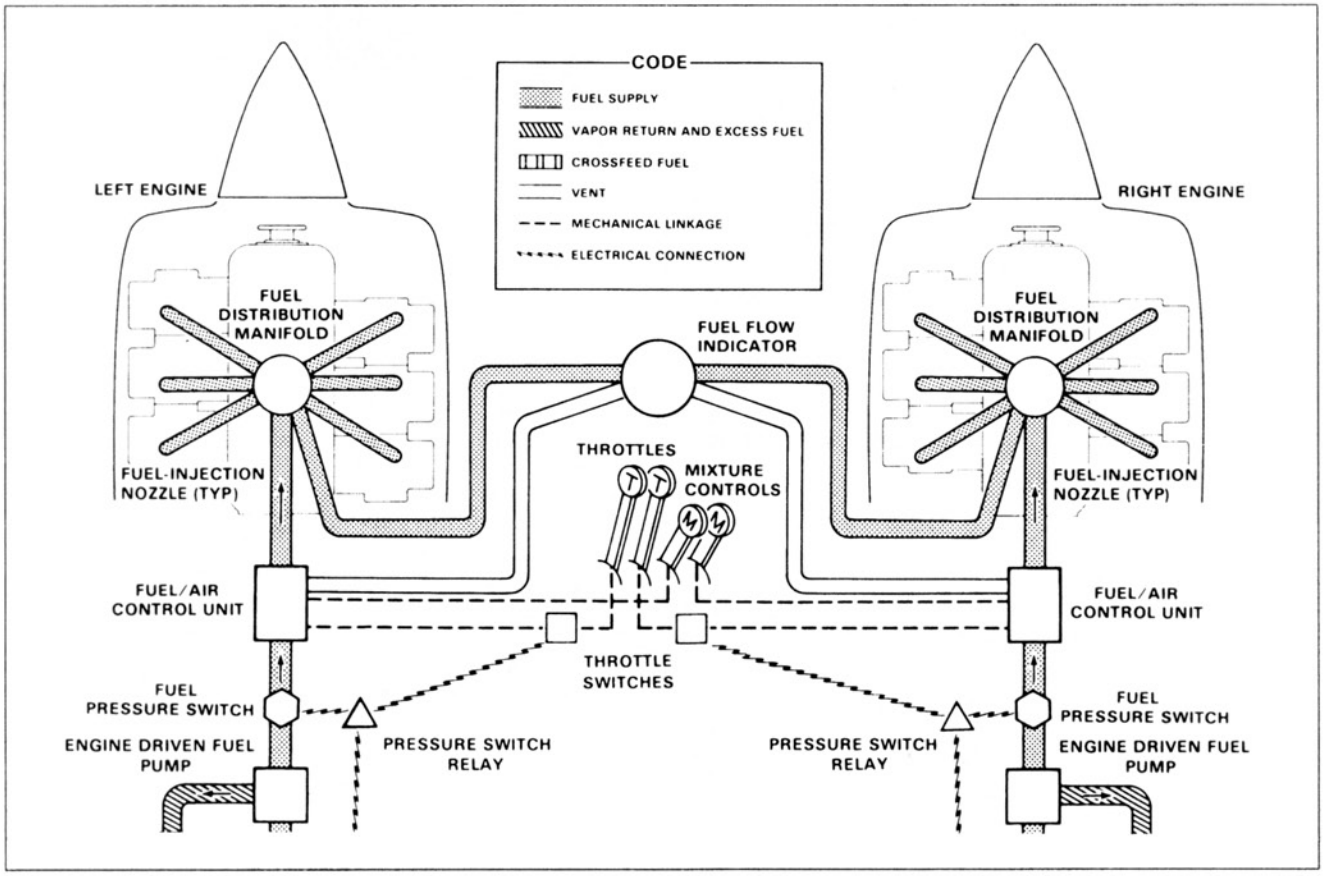
CODE
FUEL SUPPLY
VAPOR RETURN AND EXCESS FUEL
CROSSFEED FUEL
VENT
MECHANICAL LINKAGE
ELECTRICAL CONNECTION
LEFT ENGINE
RIGHT ENGINE
FUEL DISTRIBUTION MANIFOLD
FUEL DISTRIBUTION MANIFOLD
FUEL FLOW INDICATOR
THROTTLES
MIXTURE CONTROLS
T
T
M
M
FUEL-INJECTION NOZZLE (TYP)
FUEL-INJECTION NOZZLE (TYP)
FUEL/AIR CONTROL UNIT
FUEL/AIR CONTROL UNIT
THROTTLE SWITCHES
FUEL PRESSURE SWITCH
FUEL PRESSURE SWITCH
ENGINE DRIVEN FUEL PUMP
ENGINE DRIVEN FUEL PUMP
PRESSURE SWITCH RELAY
PRESSURE SWITCH RELAY

Figure 2-11. The fuel flow indicator system for a fuel-injected reciprocating engine that measures pressure at the fuel distributor manifold. (Courtesy Cessna Aircraft Co.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

this positional information concerning the vane is a type of synchro system. An example of a vane-type flowmeter with a synchro system is shown in figure 2-14. Since these synchro systems are used with many other types of aircraft instruments, they will be described next.

3. Synchro Systems

There are three types of synchro systems and they share the same basic features and are used for similar purposes. A synchro system consists of a transmitter unit and a receiver unit. The two are connected to each other by electrical wiring. The transmitter unit contains an input shaft and it can be connected to anything which will rotate

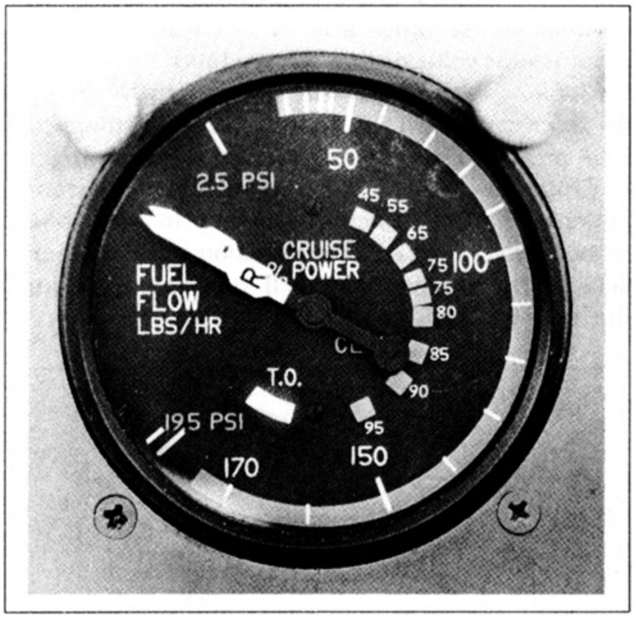
2.5 PSI
50
45
55
65
CRUISE
% POWER
75
100
75
80
FUEL
FLOW
LBS/HR
85
90
T.O.
95
195 PSI
170
150

Figure 2-12. ***Fuel flow indicator with three different measurement units: pounds per hour, PSI and percent of cruise power.***

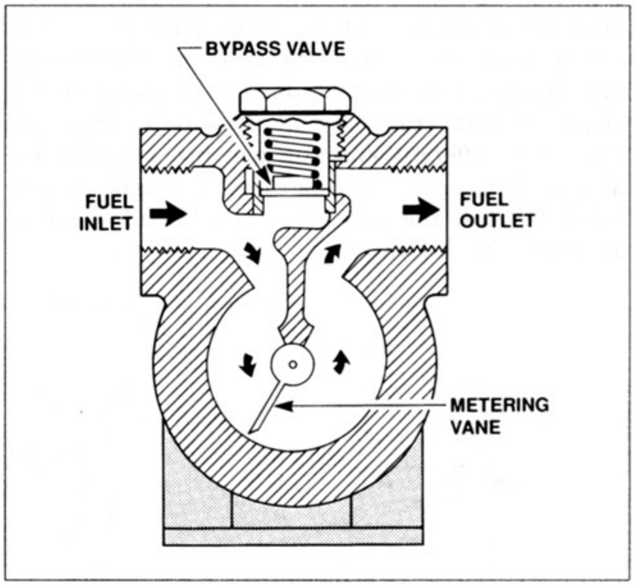
BYPASS VALVE
FUEL
INLET
FUEL
OUTLET
METERING
VANE

Figure 2-13. ***A vane-type flow sensor that measures volumetric flow rate.***

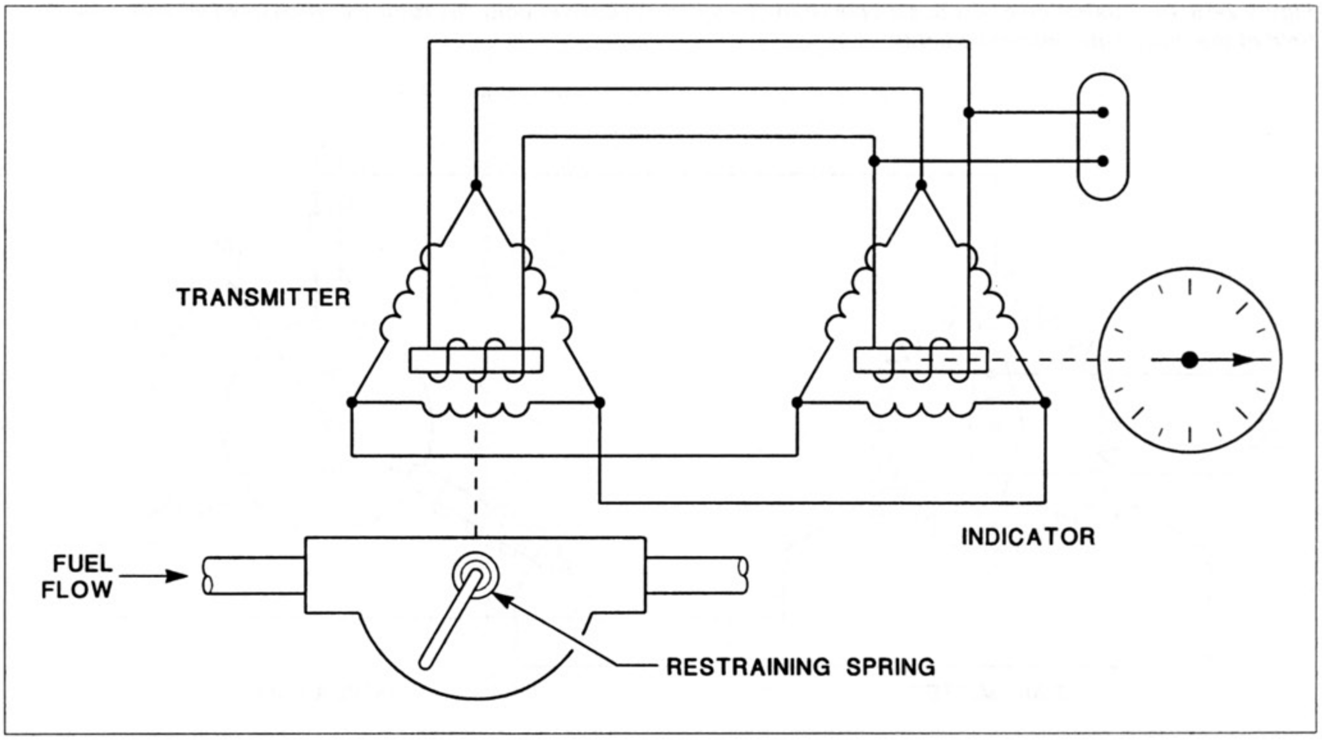
TRANSMITTER
INDICATOR
FUEL
FLOW
RESTRAINING SPRING

Figure 2-14. ***Vane-type flowmeter system for a large airplane which includes a synchro system.***

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

this shaft through an arc. The receiver unit has a shaft which is connected to the pointer in the instrument. If the shaft in the transmitter unit is rotated 20° to the right, the shaft in the receiver will also rotate 20° to the right. The operation of a synchro system causes the receiver unit to move in synchronization with the transmitter unit. A large jet transport may have many different synchro systems for a variety of different instruments. The transmitter unit can be connected to anything which produces a rotation of the shaft through an arc. Figure 2-15 shows how the rotor in the receiver unit will position itself automatically based on the magnetic field created by the three outer magnets. The three different kinds of synchro systems will now be described. They do differ in details of construction, but the basic operation of all three is as described above.

a. DC Selsyn® Synchro

The transmitter unit in the Selsyn synchro is a variable resistor with three sections as seen in figure 2-16. The shaft is connected to the wiper arm. The three sections of the variable resistor are connected by wires to the three coils in the receiver unit. The rotor of the receiver unit is a permanent magnet that is connected to the instrument pointer. The position of the wiper arm in the transmitter determines the voltages that are produced by the three sections of the resistor. The permanent magnet in the receiver unit will line up with the overall magnetic field produced by the three coils surrounding it. Any rotation of the shaft and wiper arm of the transmitter will cause different voltages to be applied to the receiver unit. A new orientation of the magnetic field in the receiver unit will pull the rotor into the correct alignment.

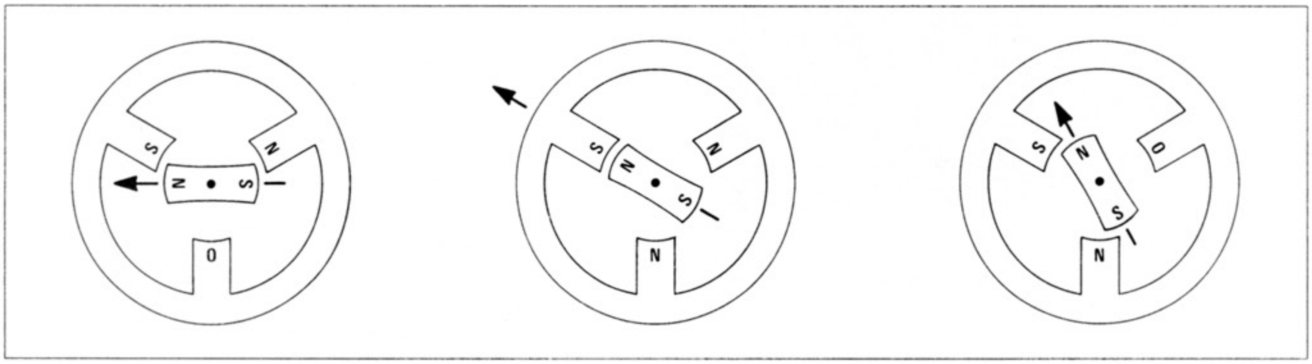

Figure 2-15. An illustration of the basic operation of a synchro receiver unit. The rotor will align itself with the resultant field of the three outer magnetic fields.

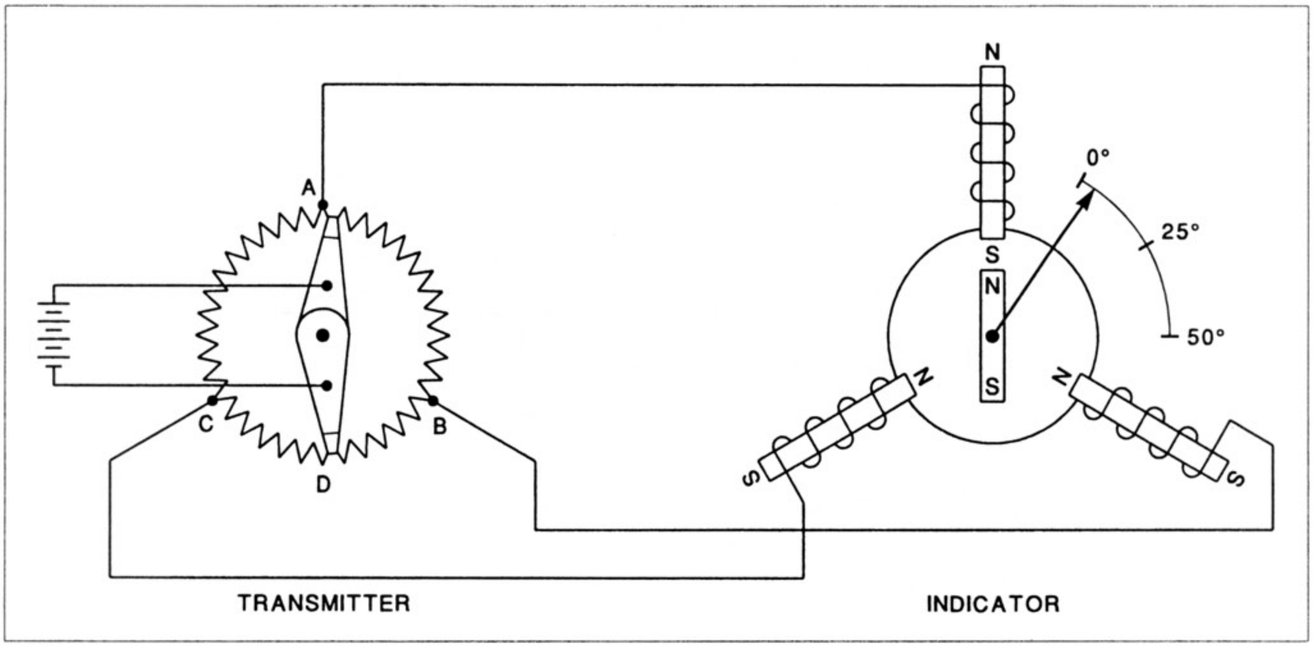

Figure 2-16. The DC-powered Selsyn synchro system.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

b. Magnesyn® Synchro

The Magnesyn synchro system uses AC power, most often 26 volts AC and 400 Hz. The use of AC power eliminates the need for a variable resistor and improves reliability because there are no brushes to wear or get out of adjustment. The construction features of the transmitter and receiver unit are similar, which can be seen in figure 2-17. The rotors are permanent magnets and the three section windings are connected together. The magnetic field produced in the receiver will pull the rotor into a position that corresponds to the rotor position in the transmitter unit.

c. Autosyn® Synchro

The only difference between the Autosyn and Magnesyn synchros is that the Autosyn uses electromagnets instead of permanent magnets for the rotors. Figure 2-18 shows an Autosyn synchro system. It also uses AC power that is most often 26 volts AC at 400 Hz. Many pressure and flow type instruments on modern jet airplanes use a synchro system to transmit the information to the cockpit gauge.

4. Mass Flowmeters

The latest types of turbine engine aircraft use a flowmeter that gives a reading of the mass flow rate in pounds per hour rather than a volumetric reading in gallons per hour. The mass flow rate is a more useful indication for this type of aircraft. Refer to figure 2-19 for a drawing of the mass flowmeter. The mass flowmeter consists of a motor-driven impeller, a turbine and a synchro system to transmit the data to a cockpit gauge. In order to give accurate readings, the impeller must be driven at a constant speed. This is accomplished with an AC synchronous motor or a similar device. As the fuel flows through the impeller, it is given a spin or rotation by the spinning impeller. When the fuel leaves the impeller, it strikes the turbine which is rotated against a restraining spring by the spin energy of the fuel. Because a denser fuel would impart more spin energy to the turbine, the degree of rotation of the turbine is a measure of mass flow rate. The turbine is connected to the transmitter rotor of a synchro system which will cause the pointer on the cockpit gauge to rotate to the proper position to indicate the correct mass flow rate. The sensor for this and other types of flowmeters is installed in the fuel system downstream of the fuel control device so that the flow rate represents the fuel consumption rate for that engine.

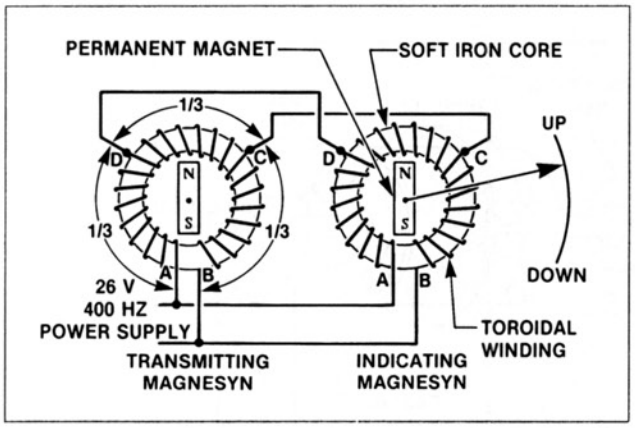

Figure 2-17. The AC-powered Magnesyn synchro system.

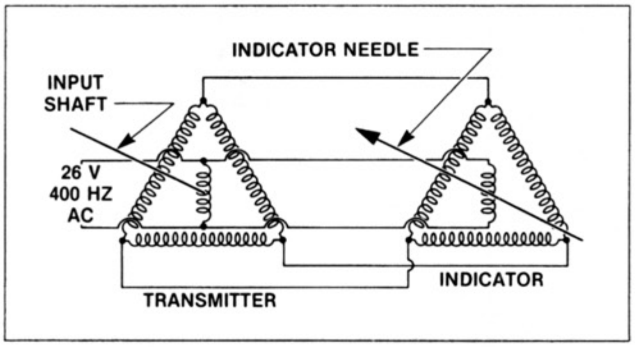

Figure 2-18. The AC-powered Autosyn synchro system.

5. Computerized Fuel System

The computerized fuel system is a volumetric flow measuring system found on some fuel-injected reciprocating engines. The sensor is a small unit which is installed on top of the engine in the fuel line that feeds the fuel splitter or manifold. The sensor or transducer contains a small rotor that has the same density as the fuel to ensure accurate readings.

As the fuel flows past the rim of the rotor, it spins the rotor at a rate which is proportional to the volumetric flow rate of the fuel. The rotor has notches on its rim which interrupt a light beam from a light emitting diode (LED). This light beam falls on a phototransistor which produces an output signal with a frequency that matches the flicker rate of the light beam. The electrical output of the phototransistor is connected to the computer in the cockpit instrument which processes the data and displays information for the pilot. The input signal to the computer has a frequency which is an indication of flow rate, but the computer can calculate and display fuel flow, fuel used and fuel remaining in several different units of measurement. By including a computer to process the signal from the transducer, the computerized fuel system can give the pilot a number of different kinds of useful information. The cockpit indicator is usually a lighted display like that shown in figure 2-20. This is not a CRT but a digital lighted display. There are three common types of lighted digital displays

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

available that use light emitting diodes (LEDs), liquid crystal displays (LCDs), or gas discharge tubes.

6. FARs for Fuel Systems

In chapter 1, the requirements for powerplant instruments were covered and it would be useful to look over that section again to review the information appropriate to the instruments covered in this section. Some additional FARs concerning fuel systems are given here.

FAR 23.993

Fuel Systems

1. Each fuel line must be installed and supported to prevent vibration and to withstand fuel pressure and flight loads.
2. Where relative motion could exist, fuel lines must have provisions for flexibility.
3. Each flexible hose must be approved or shown to be suitable for the particular application.
4. No flexible hose that might be adversely affected by exposure to high temperatures may be used where excessive temperatures will exist during operation or after engine shutdown.

FAR 23.1337

Powerplant Instruments

1. Each line carrying flammable fluids under pressure must:
   a. Have restricting orifices or other safety devices at the source of pressure to prevent the escape of excessive fluid if a line fails; and
   b. Be installed and located so that the escape of fluids would not create a hazard.
2. Each powerplant instrument that utilizes flammable fluids must be installed and located so that the escape of fluid would not create a hazard.
3. Fuel Quantity Indicator — There must be a means to indicate to the flight crew members the quantity of fuel in each tank during flight. An indicator, calibrated in either gallons or pounds, and clearly marked to indicate which scale is used, may be used. In addition:
   a. Each fuel quantity indicator must be calibrated to read zero during level flight when the quantity of fuel remaining in the tank is equal to the unusable fuel supply.
   b. Each exposed sight gauge used as a fuel quantity indicator must be protected against damage.

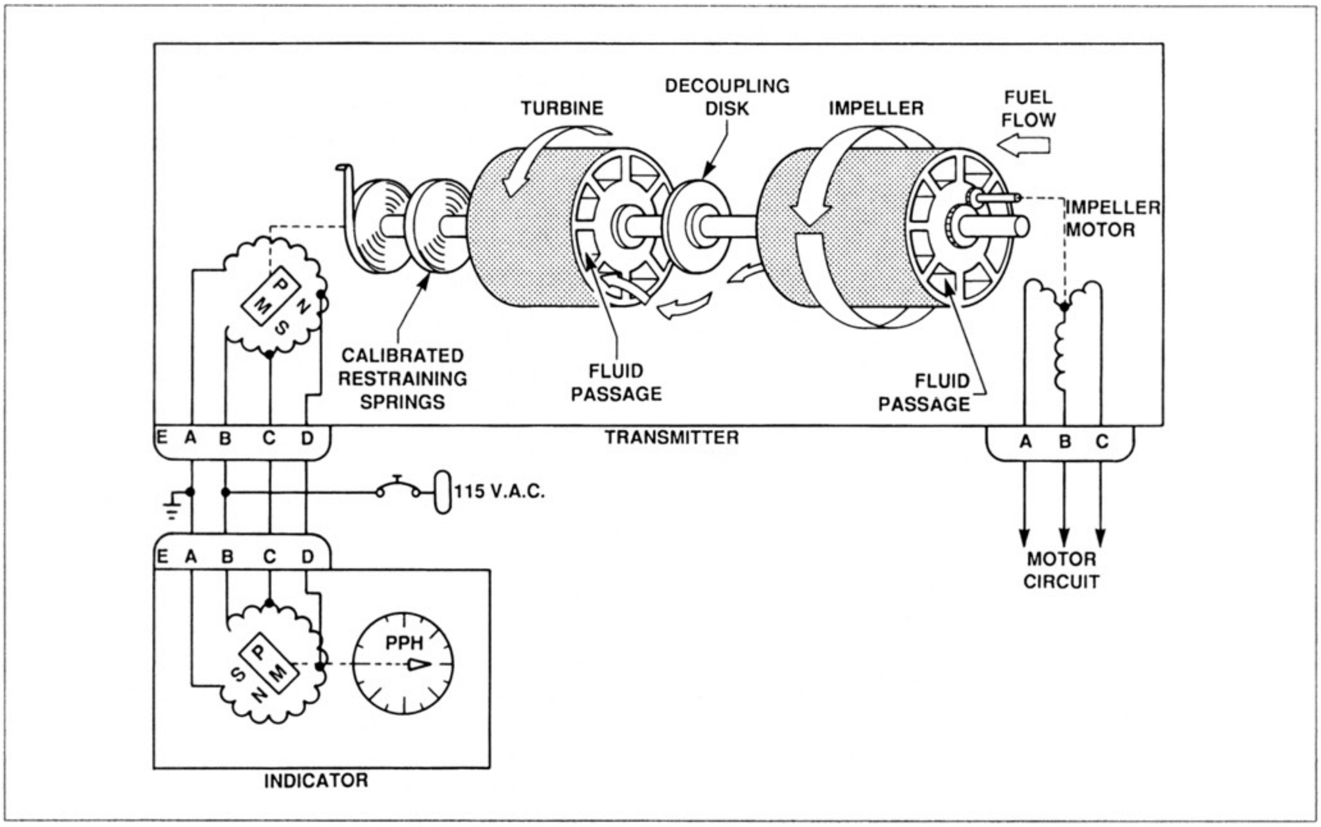

Figure 2-19. A fuel flowmeter system that measures the mass flow rate of fuel for a turbine engine.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

c. Each sight gauge that can collect water and freeze must have a means to allow drainage on the ground.

d. Tanks with interconnected outlets and airspaces may be considered as one tank and need not have separate indicators.

4. Fuel Flowmeter System — Each metering component must have a means to bypass fuel if a malfunction of that component severely restricts fuel flow.

FAR 23.1553

Fuel Quantity Indicator

If the unusable fuel supply for any tank exceeds one gallon, or five percent of tank capacity, whichever is greater, a red arc must be marked on its indicator extending from the calibrated zero reading to the lowest reading obtainable in level flight.

FAR 23.1557

Miscellaneous Markings and Placards

1. Fuel and oil filler openings.
   a. Each fuel filler opening must be marked on or near the filler cover with the word "fuel" and the acceptable fuel grades.
   b. For pressure fueling systems, the maximum permissible fueling and defueling pressures must be indicated.
   c. Oil filler openings must be marked at or near the filler cover with the word "oil".

C. Temperature Measuring Systems

There are a number of common methods of measuring temperatures on an aircraft. All of them will have a limit as to how high a temperature they can be used to measure. The common types of temperature measuring systems will be described, but not all of the possible applications can be included. Aircraft temperature indicators may give readings in degrees Fahrenheit or in degrees Celsius. Most of the temperatures given here will be in degrees Fahrenheit for ease of comparison.

1. Bimetallic

The bimetallic temperature system is limited to measuring temperatures up to a maximum of 140°F. The outside air temperature gauge (OAT) or free air temperature gauge is an example of the bimetallic system. The device that reacts to changes in temperature is a bimetallic sensor that consists of two thin strips of metal joined together. The strips are made of different metals that have different coefficients of expansion. The metals iron and brass are often used.

As the temperature changes, one of the metal strips will expand or contract more than the other causing the device to bend and move the indicator pointer. The basic principle of a bimetallic temperature sensor is shown in figure 2-21. The two metal strips are

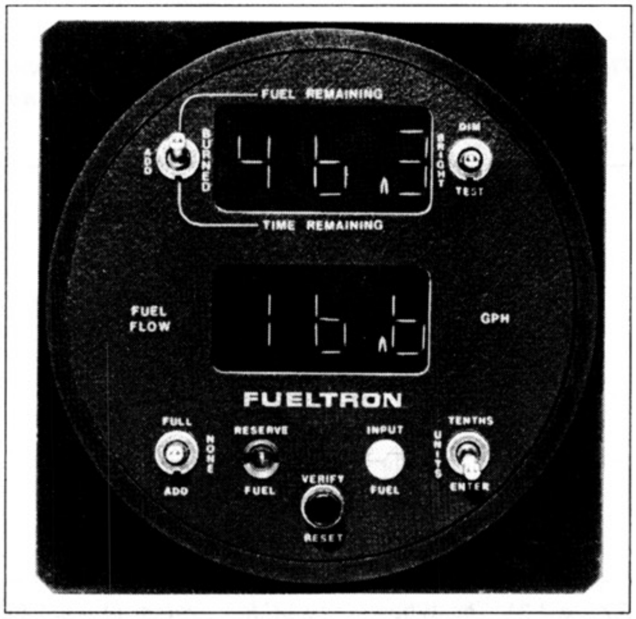

Figure 2-20. The lighted digital display instrument for a computerized fuel flow system.

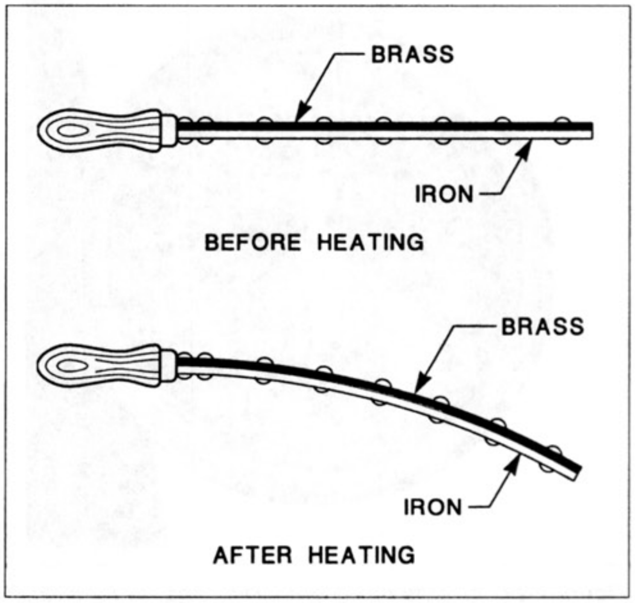

Figure 2-21. An illustration of the basic principle of operation for a bimetallic temperature sensor.

often formed into a spiral shape so that a temperature change will cause a rotating motion of the sensor strips. A typical bimetallic outside air temperature gauge is shown in figure 2-22. This type is often installed through the windshield.

2. Mechanical Bulb

The mechanical bulb utilizes the principle of the increase in pressure of a confined gas with temperature increases to measure temperatures. As shown in figure 2-23, the mechanical bulb system consists of a bourdon tube gauge to measure pressure, a thin-walled bulb which is at the point of measurement and a thin tube (capillary tube) to connect them together.

The system is filled with a chemical such as methyl chloride which will be part liquid and part gas. The system is sealed with the proper amount of the chemical so that the change in pressure with temperature changes will give an accurate reading on the bourdon tube gauge. The mechanical bulb system is found as an oil temperature indicator system on many smaller airplanes that don't have an electrical system. A mechanical bulb is also found on some jet engines where it transmits compressor inlet temperature (CIT) data to the fuel control unit. The maximum temperature for this type of system is about 300°F.

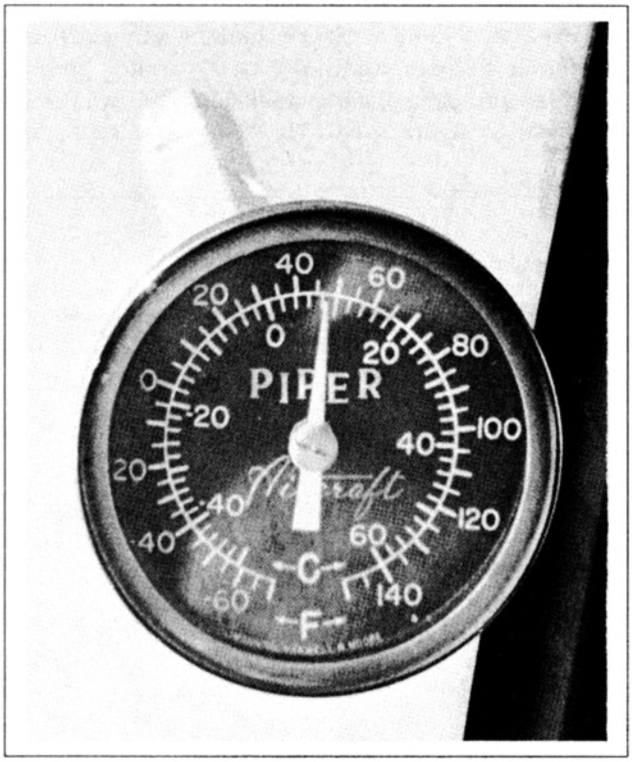

***Figure 2-22.* An outside air temperature gauge for a small airplane. This gauge uses the bimetallic principle.**

3. Wheatstone Bridge

This method of measuring temperature is powered by electricity and limited to about 300°F. The Wheatstone bridge system is illustrated in figure 2-24. The bridge circuit consists of three fixed resistors

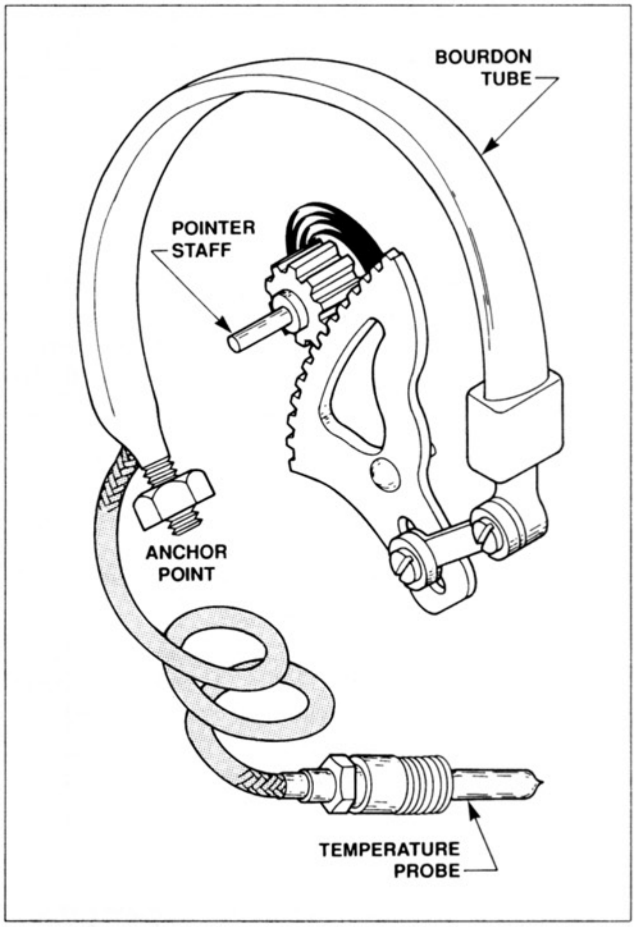

***Figure 2-23.* A mechanical bulb temperature measuring system which measures the vapor pressure of a special chemical.**

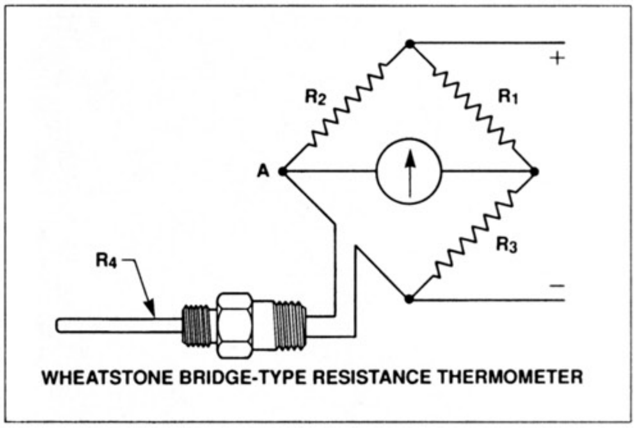

***Figure 2-24.* The Wheatstone bridge system is used to measure temperatures with a variable resistance probe.**

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

and one variable resistor. The variable resistor is the temperature probe which contains a coil of fine nickel wire. As the coil of wire is heated, its resistance increases and alters the current flow in the bridge which moves the needle in the gauge. Electromagnetic attraction and repulsion will move the pointer whenever the current flow through the meter changes. A disadvantage of the Wheatstone bridge is that any added resistance due to bad connections or any fluctuations in the system voltage can cause inaccurate readings. For this reason, it has been largely replaced by the ratiometer.

4. Ratiometer

The ratiometer uses the same kind of electrical power and the same kind of probe that the Wheatstone bridge uses. The difference is in the meter movement that moves the indicator pointer. The ratiometer has two opposing magnetic fields that combine to produce a resultant field that moves the pointer. The resultant field is a ratio of the two opposing fields so that a lower voltage applied to the system will not cause inaccurate readings. The ratiometer can measure temperatures up to 300°F. and is used for oil temperature indicators and other similar requirements. The schematic for a ratiometer is shown in figure 2-25.

5. Thermocouples

When it becomes necessary to measure temperatures of about 500°F or more the thermocouple is most often used. The principle of the thermocouple is shown in figure 2-26. When a junction of two dissimilar metals is heated, it will produce a difference of potential or voltage. The amount of voltage produced is proportional to the temperature. The terms "hot junction" and "cold junction" are used with thermocouples. The hot junction is where the temperature measurement is being taken, while the cold junction is at the opposite end of the wires in the instrument. The voltage output of the thermocouple is a result of the temperature difference between the hot junction and the cold junction. It sometimes is necessary to compensate for any temperature variances at the cold junction in order to obtain accurate readings. Only a few combinations of metals are used for thermocouple hot junctions. The metals must not only withstand the high temperature being measured, but they must produce a usable amount of voltage. The chart in figure 2-27 shows the three commonly used pairs of metals and their voltage output at various temperatures. The actual voltages produced are very low, particularly at lower temperatures. This is why the thermocouple system is not usually used for temperatures below about 400°F. The three types of thermocouples are the iron-constantan, the copper-constantan and the chromel-alumel. Notice that the chromel-alumel can measure much higher temperatures than the other two types. Thermocouple leads are available in standard lengths with specific values of resistance. The length of a thermocouple lead should not be altered in the field.

An application of the thermocouple on piston engines is the cylinder head temperature (CHT) gauge. This gauge is used to monitor the cooling of an air-cooled engine. If only one probe is used, it will be installed in the hottest running cylinder. This would usually be a rear cylinder on a horizontally opposed engine. One type of CHT probe is a gasket that goes under the spark plug, another type fits in a special fitting in the cylinder head. These two types are shown in figures 2-28 and 2-29 respectively. It is better to install CHT probes in all the cylinders. In addition to monitoring all the cylinders, it can be used to troubleshoot some types of engine problems. Another application for the thermocouple is as an exhaust gas temperature (EGT) gauge. All turbine engines have an

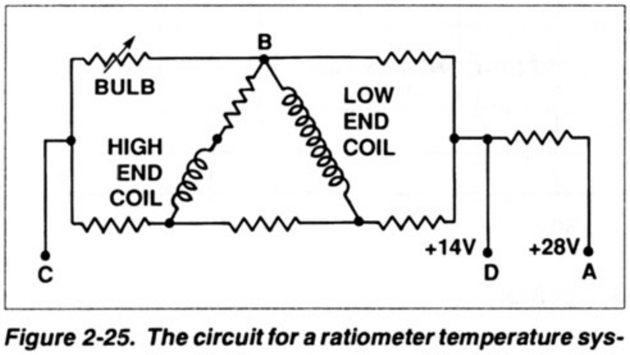

Figure 2-25. The circuit for a ratiometer temperature system that can operate on 14 or 28 volts DC.

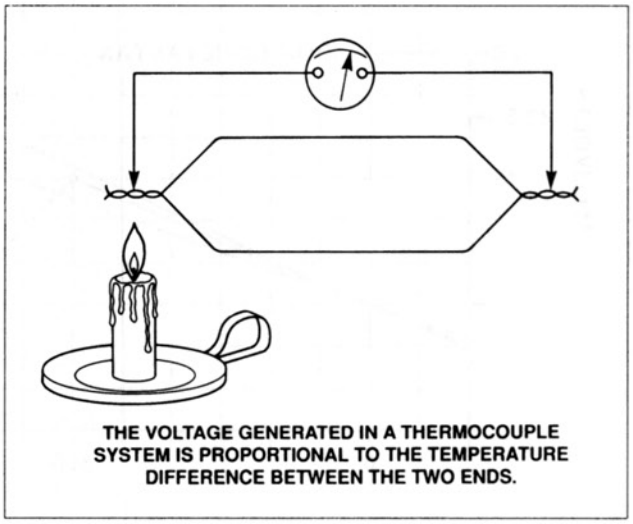
THE VOLTAGE GENERATED IN A THERMOCOUPLE SYSTEM IS PROPORTIONAL TO THE TEMPERATURE DIFFERENCE BETWEEN THE TWO ENDS.

Figure 2-26. An example of a simple thermocouple system showing the hot junction and cold junction.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

EGT gauge or gas temperature gauge because turbine engines can be severely damaged by high temperatures in the turbine sections. A number of terms and abbreviations are used for the gas temperature gauge on turbine engines. Turbine inlet temperature (TIT) refers to a system where the probes are installed ahead of the turbine stages. Inter turbine temperature (ITT) means the probes are located between the different sections of the turbine. Turbine outlet temperature (TOT) and EGT refer to measurements that have the probes installed downstream of all the turbine stages. These locations and abbreviations are shown in figure 2-30. Turbine engines use multiple thermocouple probes that are connected in parallel to give an average temperature as seen in figures 2-31 and 2-32.

Some aircraft use the term measured gas temperature (MGT) for the required gas temperature indicator. An MGT gauge for a turbine engine helicopter is shown in figure 2-33.

On reciprocating engine aircraft, the EGT gauge has a different purpose. It is used to manually lean the fuel-air mixture for better economy. The thermocouple probe is installed in an exhaust pipe and connected to a simple gauge in the cockpit. The actual procedure for leaning the engine will vary from one aircraft to another, the example given here is for purposes of illustration. When ready to set the fuel-air mixture, the pilot watches the EGT gauge as the mixture control knob is pulled back. The temperature will rise as the mixture is leaned because more efficient combustion is taking place. When the EGT reaches a peak, the proper mixture for maximum economy has been reached. If only one probe is installed for the EGT system, the indication is really the average EGT for all the cylinders and they can vary. It is best to have an EGT probe for each cylinder as this gives more information and permits the use of EGT for troubleshooting engine faults. When the EGT gauge is used for leaning, there is usually no redline at all on the gauge. This can be seen on the gauge in figure 2-34 which is an EGT system for a reciprocating engine airplane. However, a turbo-supercharged reciprocating engine will have a redline because the turbo-supercharger can be damaged by high temperatures.

In aircraft schematics, the thermocouple wires are given standard color codes for ease of identification. The standard colors for the wires are:

1. IRON — Black
2. CONSTANTAN — Yellow
3. COPPER — Red
4. CHROMEL — White
5. ALUMEL — Green

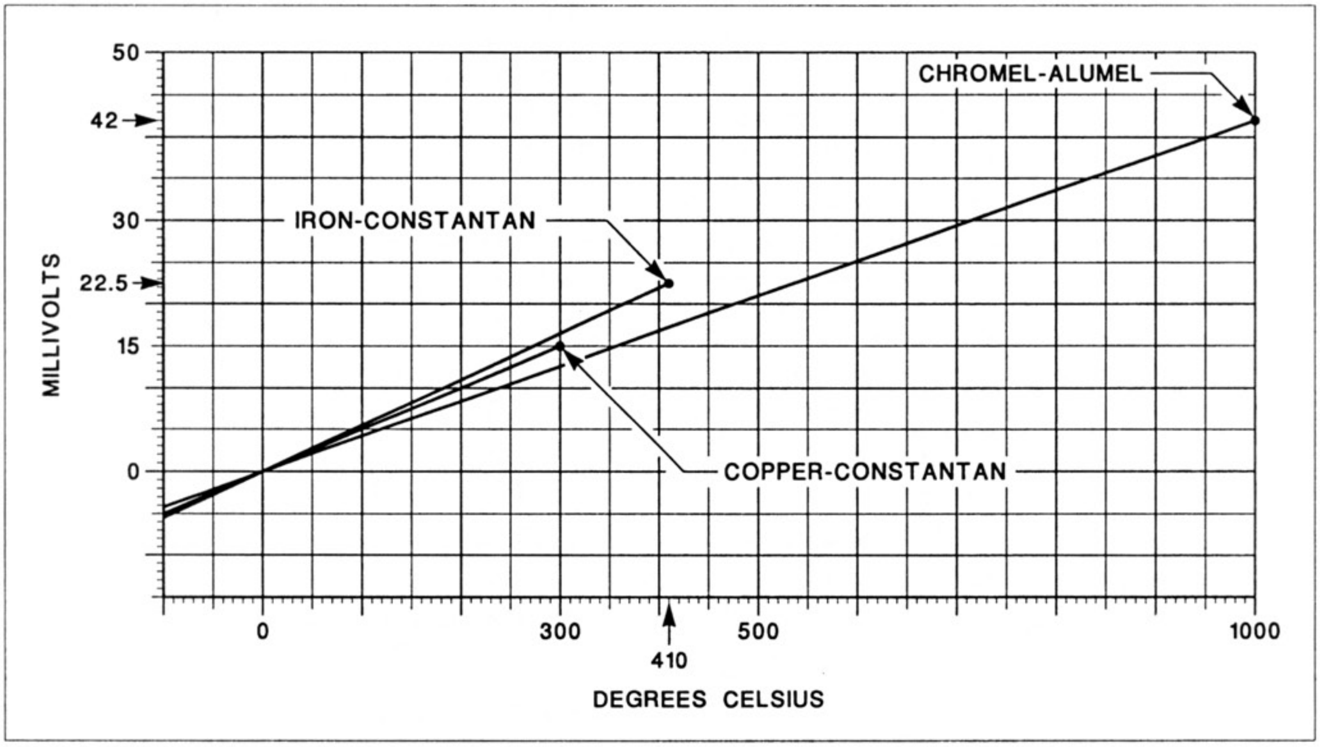

Figure 2-27. Voltage output versus temperature for different combinations of thermocouple metals.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

D. Position Indicating Systems

There are many different components and systems on aircraft that might have a position indicating system. There are a relatively small number of different methods that are used to obtain these indications. The basic operation of common types of position indicating systems will be explained and some specific applications will be discussed.

Some of the different methods used to operate position indicating systems are:

1. Mechanical — Rods, levers, cables, etc.
2. Microswitches — Sometimes called limit switches (figure 2-35).
3. Variable Resistance — Wheatstone bridge and ratiometer.
4. Proximity sensors — Mainly on larger, more modern aircraft.
5. Synchro Systems — Selsyn, Magnesyn and Autosyn.

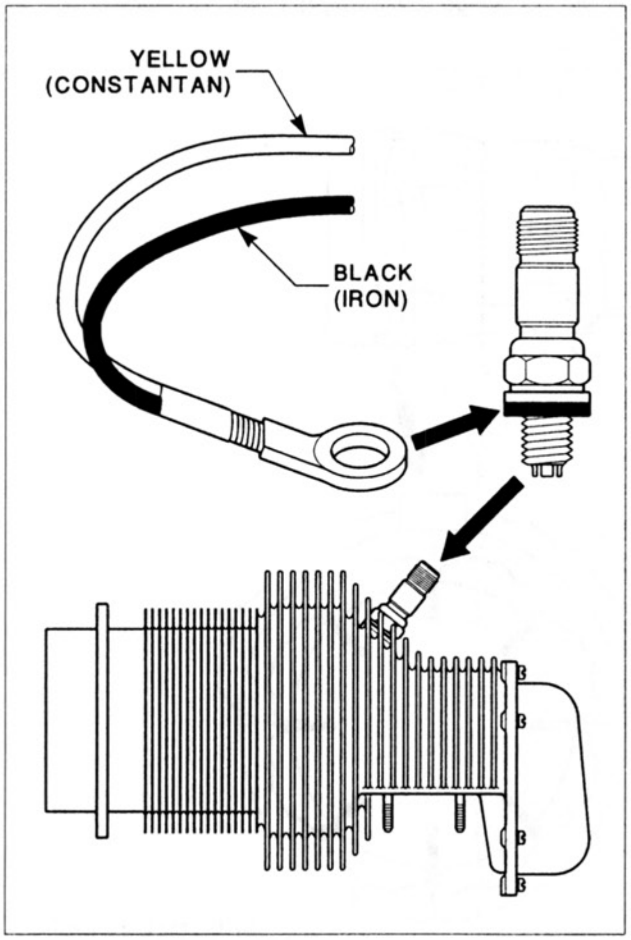

Figure 2-28. A cylinder head temperature (CHT) system for a small reciprocating engine. The thermocouple probe is in the form of a spark plug gasket.

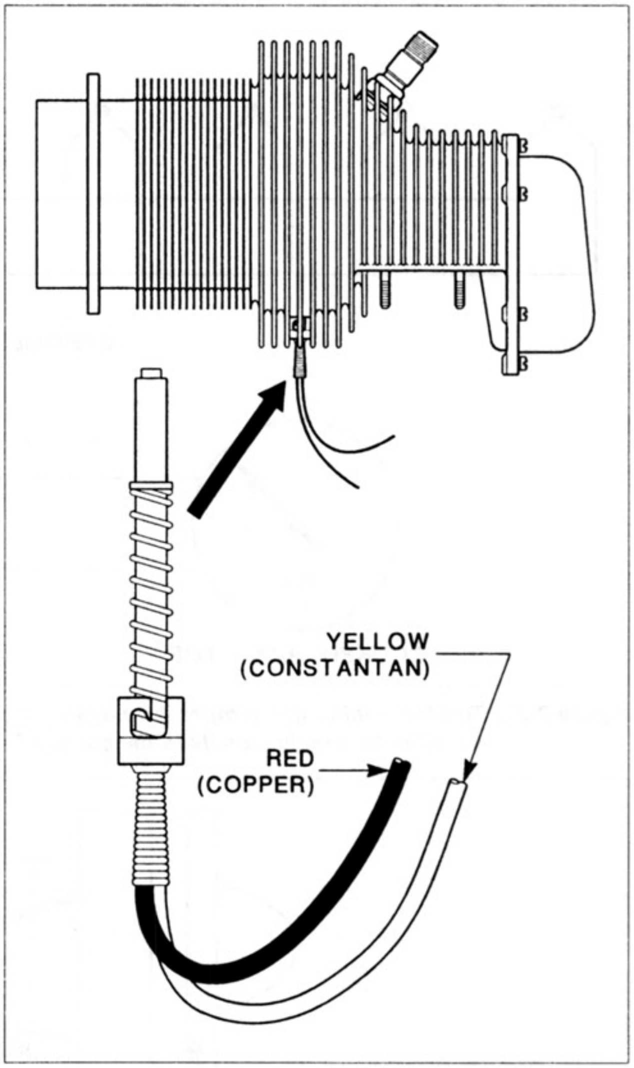

Figure 2-29. A cylinder head temperature system that uses a bayonet-type probe which fits into a special recess in the cylinder head.

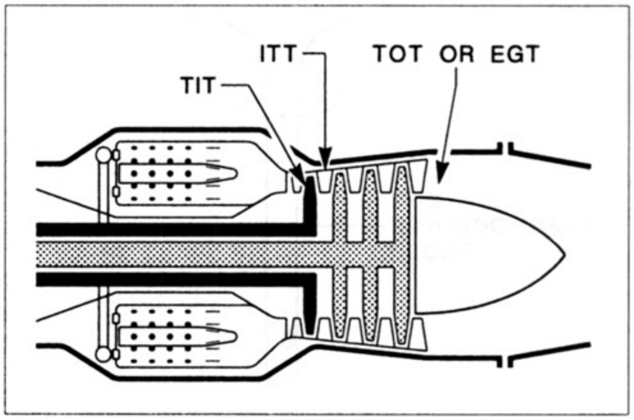

Figure 2-30. The measurement points in the turbine section for turbine inlet temperature, inter-turbine temperature, turbine outlet temperature and exhaust gas temperature.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

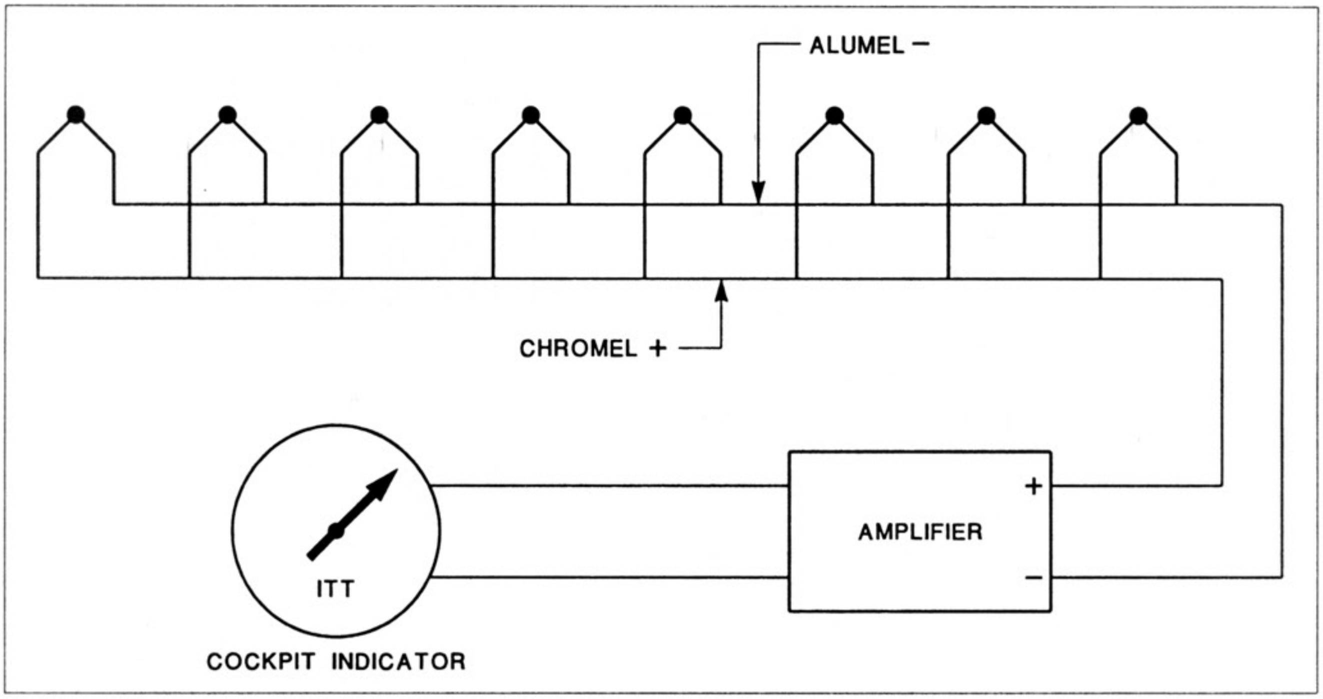
ALUMEL −
CHROMEL +
+
AMPLIFIER
−
ITT
COCKPIT INDICATOR

Figure 2-31. Turbine engine gas temperature systems such as the ITT gauge use multiple probes in parallel and an amplifier to supply signals to the cockpit indicator.

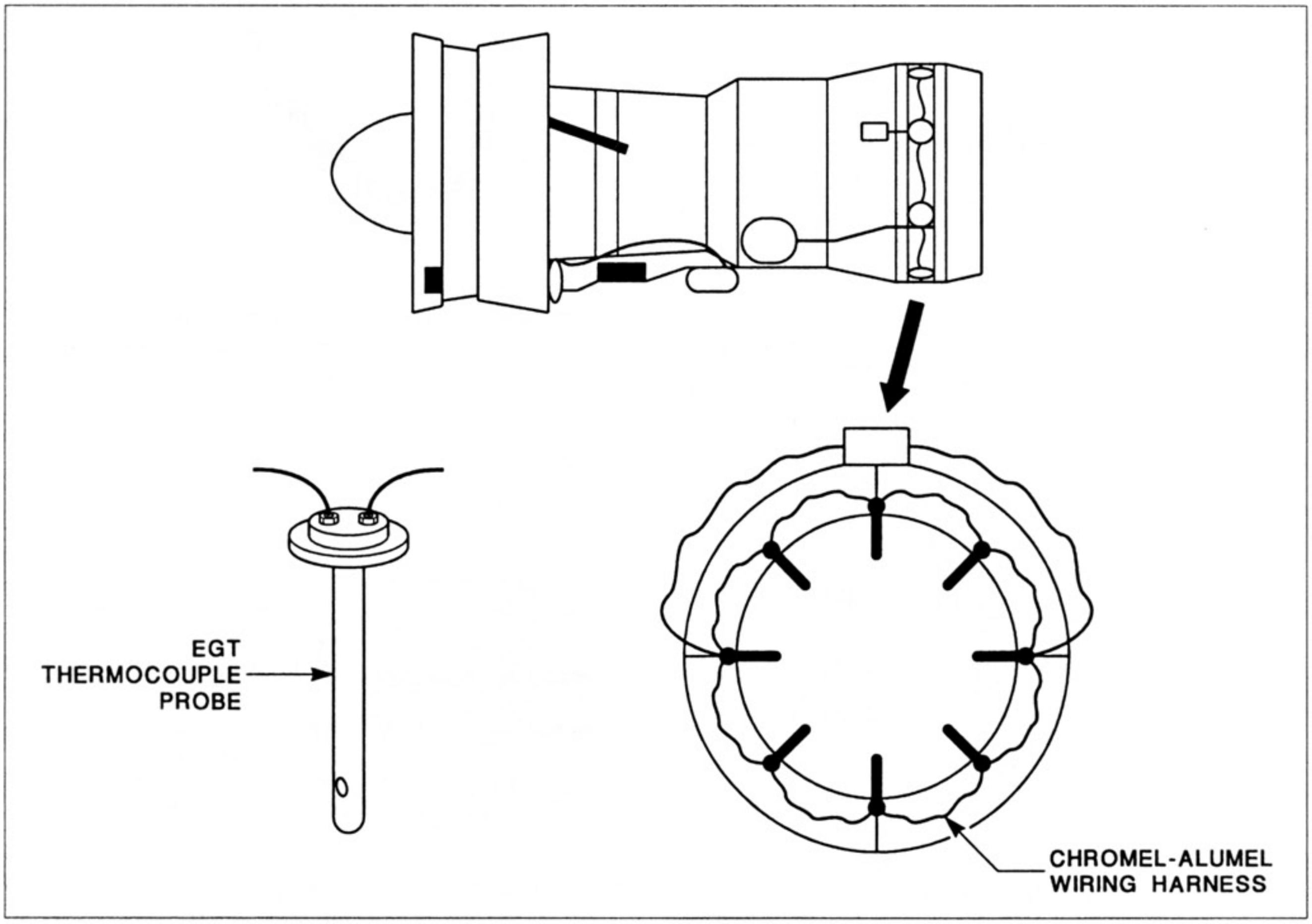
EGT
THERMOCOUPLE
PROBE
CHROMEL-ALUMEL
WIRING HARNESS

Figure 2-32. Turbine engines use multiple EGT probes of the chromel / alumel type to take an average EGT for the exhaust section of the tailpipe.

All the systems and components in an aircraft that might have a position indicating system can be divided into two categories: those that only have two operating positions and those that have many or an infinite number of different operating positions. Some examples follow:

1. Landing Gear — 2 Positions
2. Cabin and Cargo Doors — 2 Positions
3. Thrust Reversers — 2 Positions
4. Ground Spoilers — 2 Positions
5. Trailing Edge Flaps — Many Positions
6. Flight Control Surfaces — Infinite Positions
7. Trim Tabs — Infinite Positions

On smaller aircraft, the systems that only have two operating positions most often use microswitches. Landing gear position indicators are usually lights that are operated by microswitches on the landing gear. This type is shown in figure 2-36. A problem with microswitches is that they are prone to damage from rocks, sand, water, etc. that are thrown up onto the landing gear in service. They also have problems with arcing and burning of the small contact points. For these reasons, most larger aircraft employ proximity sensors instead of microswitches. Figure 2-37 illustrates the operation of one type of proximity sensor. The proximity sensor is a sealed unit that operates in conjunction with a metal target. The proximity sensor produces an electromagnetic field that is distorted when the target piece of metal moves close to it. This change is detected by an electronic circuit to give an indication of gear-up, gear-down etc. The target never touches the proximity sensor and there are no small contacts to cause problems. About the only disadvantages of the proximity sensor systems are that they cost more and since they use electronic circuits, adjustment is more complicated than for a simple microswitch.

On simple aircraft, the systems that have many or an infinite number of operating positions will use mechanical or variable resistance position indicating systems. An example of a mechanical trim tab position indicator is shown in figure 2-38. The wheel that the pilot rotates to move the trim tab contains a spiral groove which moves a small wire pointer to show the position of the trim tab. The ratiometer type variable resistance system is used as a flap position indicator on small airplanes. It works like the ratiometer system already described except that the variable resistor is moved by a part of the flap mechanism.

On the more sophisticated aircraft, the systems that have many or an infinite number of operating positions use synchros. Surface position indicators are usually found on bizjets and air carrier jets to show the flight crew the position of the flight control surfaces. These use a synchro system with the transmitter unit attached to the mechanical linkage of the flight control surface and the receiver unit in the cockpit gauge or indicator. This type of system is shown in figure 2-39.

E. Tachometers

Most types of aircraft have at least one tachometer to indicate the rotational speed of the engine. Aircraft with reciprocating engines have tachometers that indicate the crankshaft RPM. This is true for engines with reduction gearing also; the tachometer gives engine crankshaft RPM, not propeller RPM. Helicopters have a tachometer for the main rotor or rotors to enable the pilot to maintain a safe rotor RPM. Turbine engines use tachometers that give readings in percent of RPM rather than actual revolutions per minute.

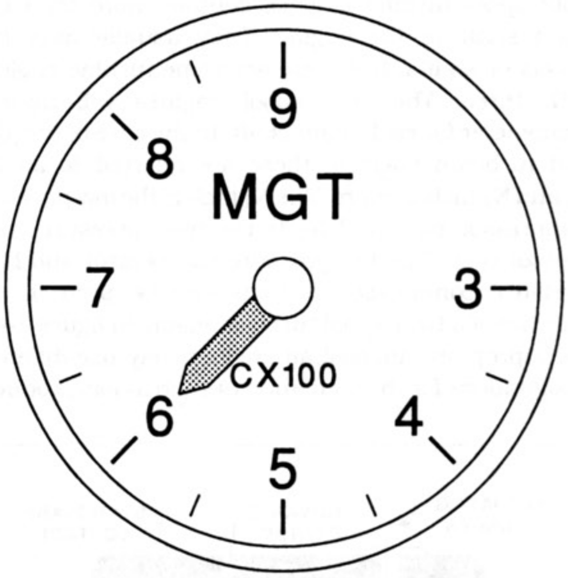

300 - 750 °C CONTINUOUS OPERATION

750 - 780 °C TAKEOFF (5 MINUTES)

780 °C MAXIMUM TAKEOFF

900 °C MAXIMUM FOR STARTING
(12 SECONDS)

Figure 2-33. Some gas temperature gauges are labeled MGT for measured gas temperature. The maximum operating temperatures are often time limited for specific operations.

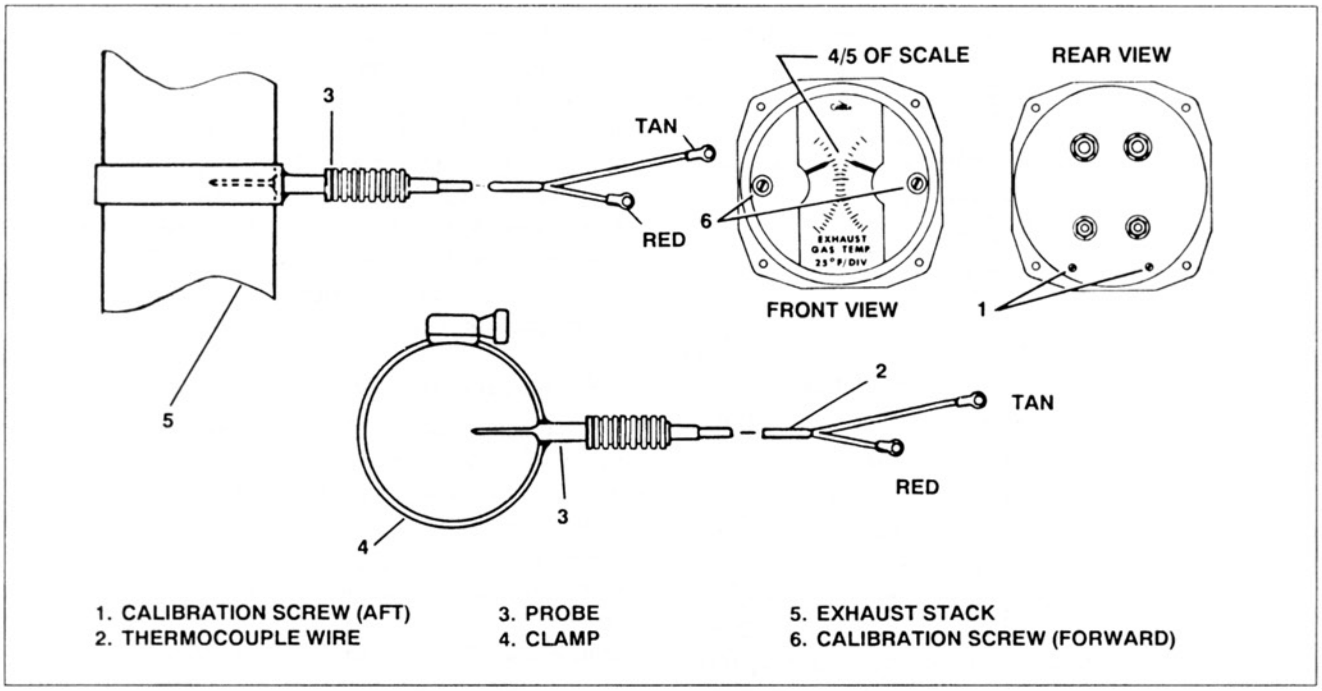

Figure 2-34. The components in an EGT system for a twin-engine airplane with reciprocating engines. This instrument is sometimes referred to as a mixture indicator. (Courtesy Cessna Aircraft Co.)

Split spool turbine engines contain more than one main shaft in the engine. They usually have two spools or shafts, but there are some turbine engines with three. The split spool engines will have a tachometer for each main shaft. In the case of turbojet and turbofan engines, these are referred to as the N_1 and N_2 tachometers. The N_1 tach is the low pressure compressor tach and N_2 is the high pressure compressor tach. The low pressure compressor and high pressure compressor sections can be seen in the drawing of a twin-spool turbojet engine in figure 2-40. Turboprop and turboshaft engines may use different designations for the tachometers such as gas producer tach and power section tach. The gas producer tach for a turbine engine helicopter shown in figure 2-41 shows some time limited permissible readings above 100%. All tachometers have a red radial line to indicate the maximum permissible RPM. On turbine engines this redline is not necessarily at 100%; it could be above or below the 100% reading. The turbine engine manufacturer will establish what engine section RPM is equal to 100%. On one particular model of CF34 turbofan engine, a reading of 100% N_2 is equal to 17,820 RPM. Some reciprocating engine tachometers will have a red arc which denotes a range of engine speeds that is prohibited due to vibration problems at those rotational speeds. A triple tach for a twin-engine helicopter is shown in figure 2-42. It is three tachometers in one to give readings for the power sections of both engines and the main rotor.

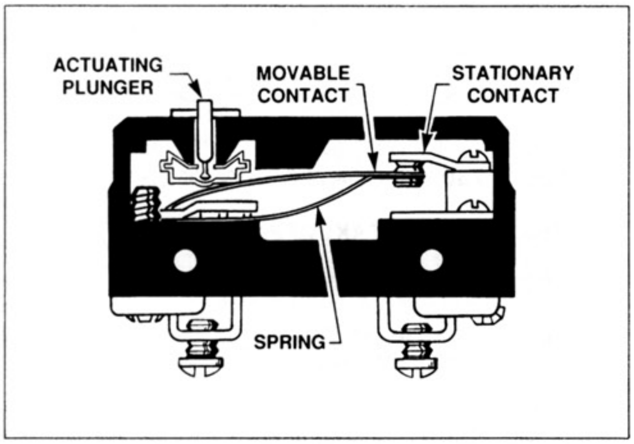

Figure 2-35. A microswitch is designed so that the contact points open and close with a very small motion of the plunger.

1. Mechanical Tachometers

Most small general aviation aircraft use simple mechanical tachometers that utilize a flexible drive cable similar to the speedometer drive cable in a car. This flexible drive cable is connected to a drive gear in the engine accessory section and the other end is connected to the tachometer in the cockpit. Older style tachometers used rotating flyweights to move the pointer in the tachometer instrument as illustrated in figure 2-43. Later mechanical tachometers use a rotating permanent magnet and a drag cup to move the pointer. A tachometer drag cup is shown in figure 2-44. The small permanent

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

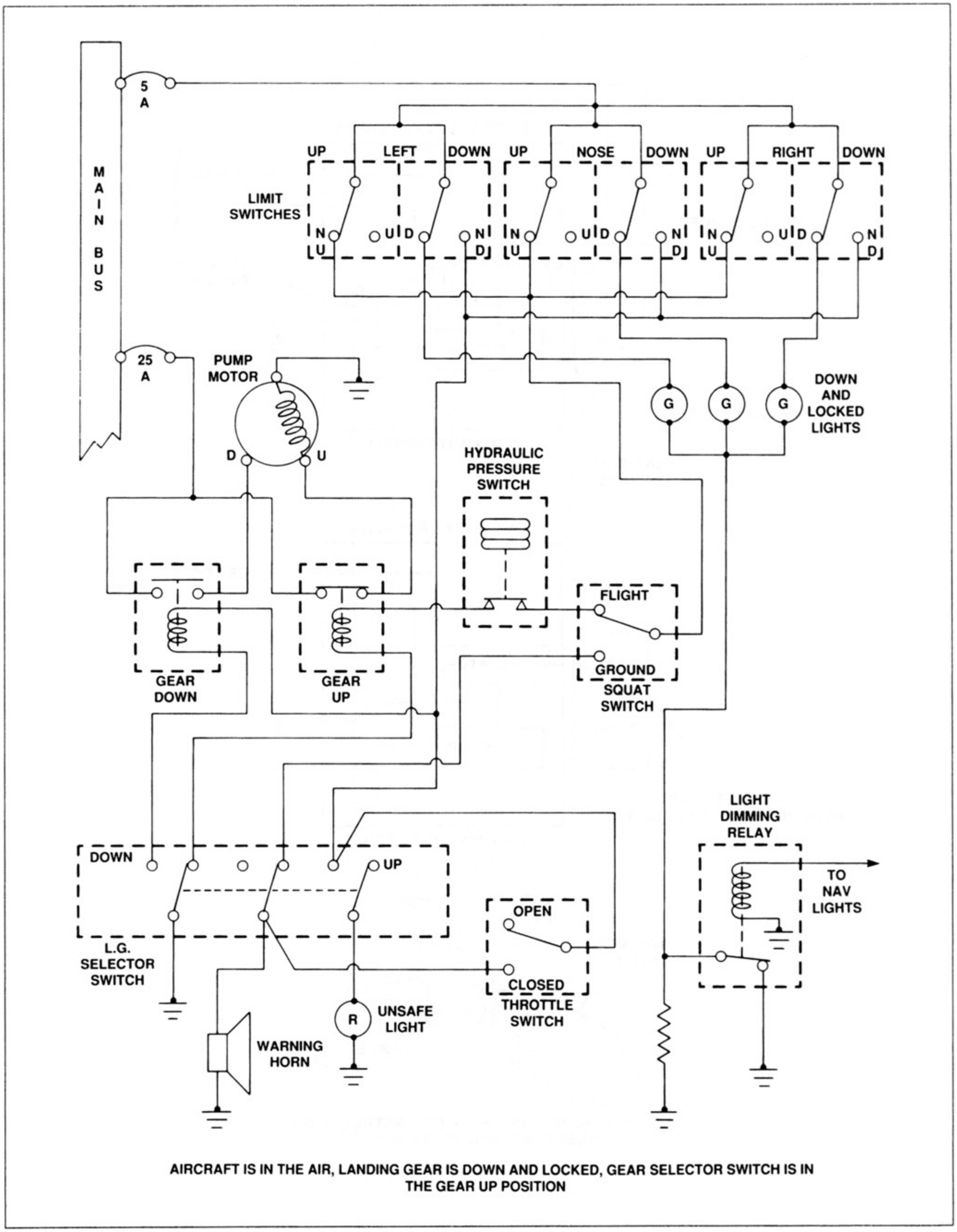
MAIN BUS
5 A
25 A
LIMIT SWITCHES
UP
LEFT
DOWN
UP
NOSE
DOWN
UP
RIGHT
DOWN
N U
U
D
N D
PUMP MOTOR
D
U
DOWN AND LOCKED LIGHTS
G
HYDRAULIC PRESSURE SWITCH
GEAR DOWN
GEAR UP
FLIGHT
GROUND
SQUAT SWITCH
LIGHT DIMMING RELAY
TO NAV LIGHTS
DOWN
UP
L.G. SELECTOR SWITCH
OPEN
CLOSED
THROTTLE SWITCH
R
UNSAFE LIGHT
WARNING HORN
AIRCRAFT IS IN THE AIR, LANDING GEAR IS DOWN AND LOCKED, GEAR SELECTOR SWITCH IS IN THE GEAR UP POSITION

Figure 2-36. The landing gear indicating and warning system for a small airplane. A number of microswitches are utilized.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

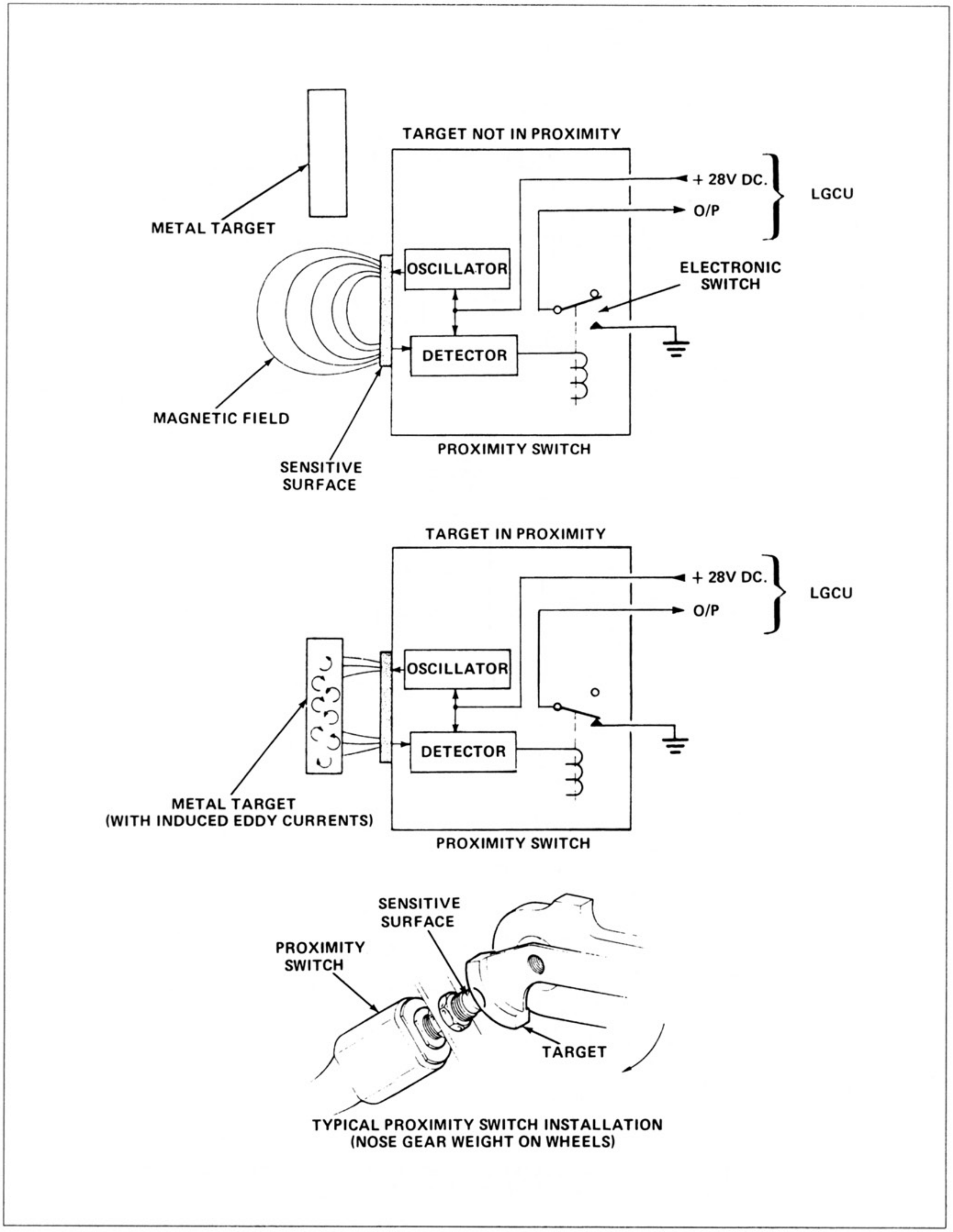
TARGET NOT IN PROXIMITY
METAL TARGET
+ 28V DC.
O/P
LGCU
OSCILLATOR
ELECTRONIC SWITCH
DETECTOR
MAGNETIC FIELD
PROXIMITY SWITCH
SENSITIVE SURFACE
TARGET IN PROXIMITY
+ 28V DC.
O/P
LGCU
OSCILLATOR
DETECTOR
METAL TARGET (WITH INDUCED EDDY CURRENTS)
PROXIMITY SWITCH
SENSITIVE SURFACE
PROXIMITY SWITCH
TARGET
TYPICAL PROXIMITY SWITCH INSTALLATION (NOSE GEAR WEIGHT ON WHEELS)

***Figure 2-37.** The operation of a proximity sensor installed on the landing gear system of a corporate jet airplane. (Courtesy Canadair Group, Bombardier Inc.)*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

magnet is fastened to the end of the drive mechanism so that it produces a rotating magnetic field. Surrounding the magnet is a drag cup made of aluminum. As the magnet rotates, it sets up eddy currents in the aluminum drag cup and the magnetic fields of the eddy currents interact with the rotating field of the permanent magnet. The interaction of the two fields causes a torque force or drag force to be applied to the drag cup which rotates it against spring tension to move the pointer. The main advantage of the drag cup tachometer is that there is no direct mechanical connection between the drive cable and the pointer mechanism. This makes it smoother in operation and less likely to break if some minor binding occurs.

2. Tachometer Generators

This type of tachometer system uses an electrical generator that is mechanically driven by gears at the engine and which transmits electrical energy to the cockpit instrument to give an indication of RPM. There are both AC and DC tach generator systems, but the DC type is not found on modern aircraft. The DC type had a small DC generator on the engine and a cockpit gauge which indicated the voltage output of the generator as RPM. The major disadvantage of this type of tach generator is that any fault which caused a lower than normal voltage would result in an error in the indicated RPM. The AC tach generator eliminates that problem by using an AC generator on the engine and an AC synchronous motor in the indicator. An AC tach generator system is shown in figure 2-45. The primary determining factor in the motor RPM is the frequency of the AC that powers it. A lower voltage caused by loose connections for example would not have much affect on the indicated RPM. Both types of tach generator systems use permanent magnets so that they are totally independent of the electrical systems of the aircraft.

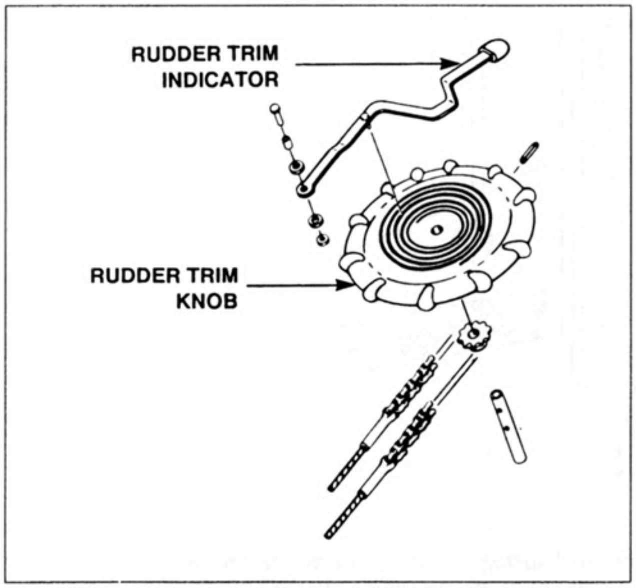

Figure 2-38. A mechanical position indicator for the rudder trim on a twin-engine airplane. (Courtesy Cessna Aircraft Co.)

3. Electronic Tachometers

There are several different types of electronic tachometers used on aircraft. The kind used on some reciprocating engines is operated by a special set of points in the engine magneto. This set of contact points opens and closes like the normal points, but only supplies signals for the tach system.

The points in the magneto are connected by wiring to the cockpit instrument. Since the frequency of opening of the points is proportional to the engine RPM, an electronic circuit measures the frequency at which the points open and close and moves the pointer to indicate the proper RPM of the engine.

Two slightly different kinds of electronic tachometers are found on turbine engines. The first type is often used as a fan speed sensor to measure the RPM of the fan section of a turbofan engine. Figure 2-46 shows this type. It uses a sensor which contains a coil of wire that generates a magnetic field. The sensor is mounted in the shroud around the fan. As each fan blade goes by, it cuts the field of the coil and this is sensed and measured by an electronic circuit. The frequency at which the fan blades cut across the field of the sensor is directly proportional to the fan RPM.

Another type of electronic tach used on turbine engines has a gear driven shaft on the engine which turns a rotor with a permanent magnet embedded in its rim. The sensor contains a coil which is located close to the rotating magnet. Each time the field of the rotating permanent magnet cuts across the coil, it induces a voltage. The frequency of this signal is measured by an electronic circuit and used to position the pointer for the correct RPM indication. This type of tachometer is used in figure 2-47 for the N_2 indication for a large turbofan engine.

F. Oil Pressure Indicators

The oil pressure gauges on small aircraft are usually the direct reading type. The oil pressure line is connected into an oil passage in the engine and transmits that pressure through tubing and hose to the cockpit instrument which contains a bourdon tube to move the pointer. Larger aircraft such as corporate turboprops, bizjets and air carrier jets will use instruments that do not rely on having fluids under pressure in the cockpit area. These aircraft may use a bourdon tube or similar pressure sensor, but it will be installed

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

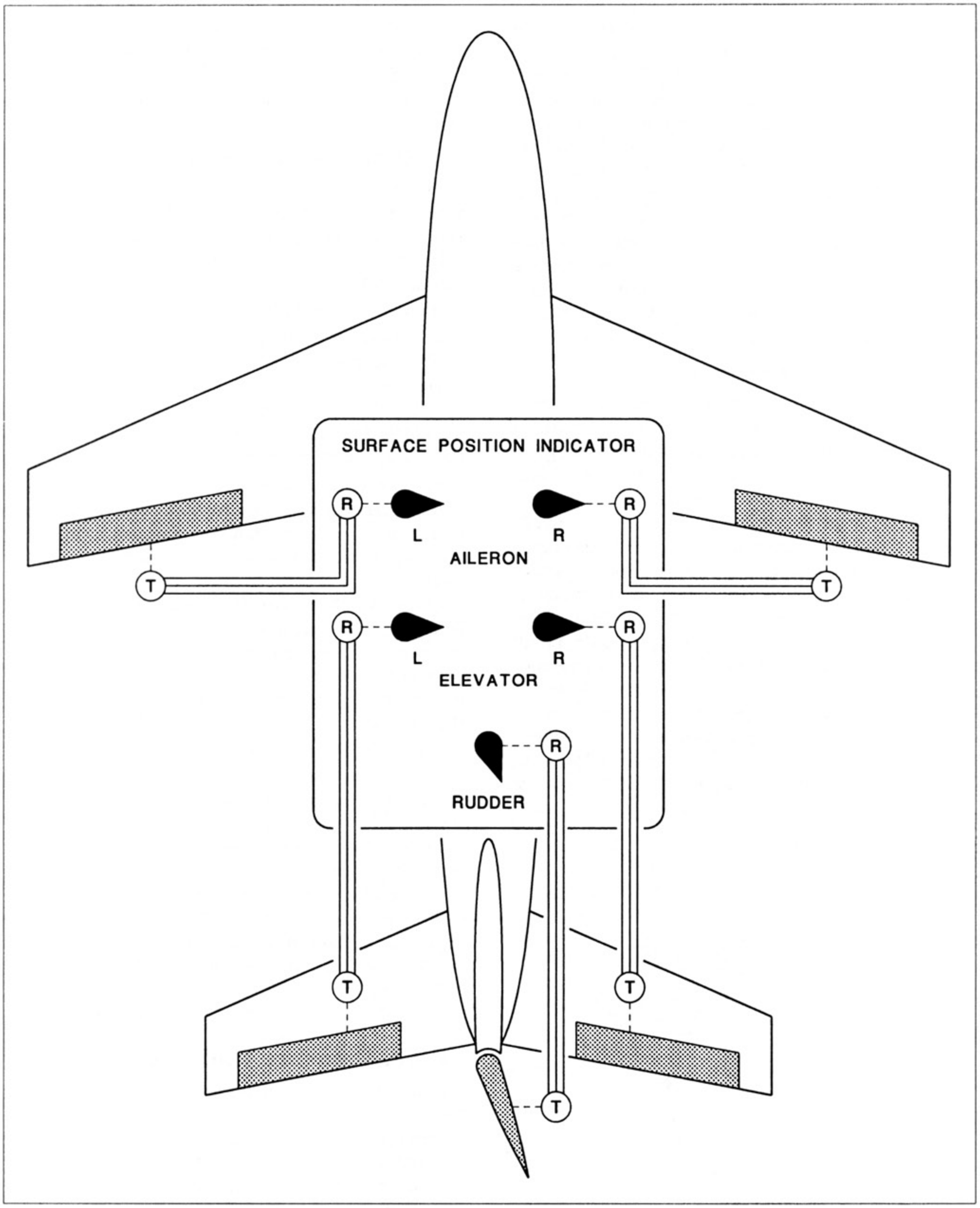
SURFACE POSITION INDICATOR
R
R
L
R
AILERON
T
T
R
R
L
R
ELEVATOR
R
RUDDER
T
T
T

Figure 2-39. A surface position indicator system for the flight control surfaces. Synchro transmitters and receivers are used to transmit the information from the control surfaces to the cockpit indicator.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

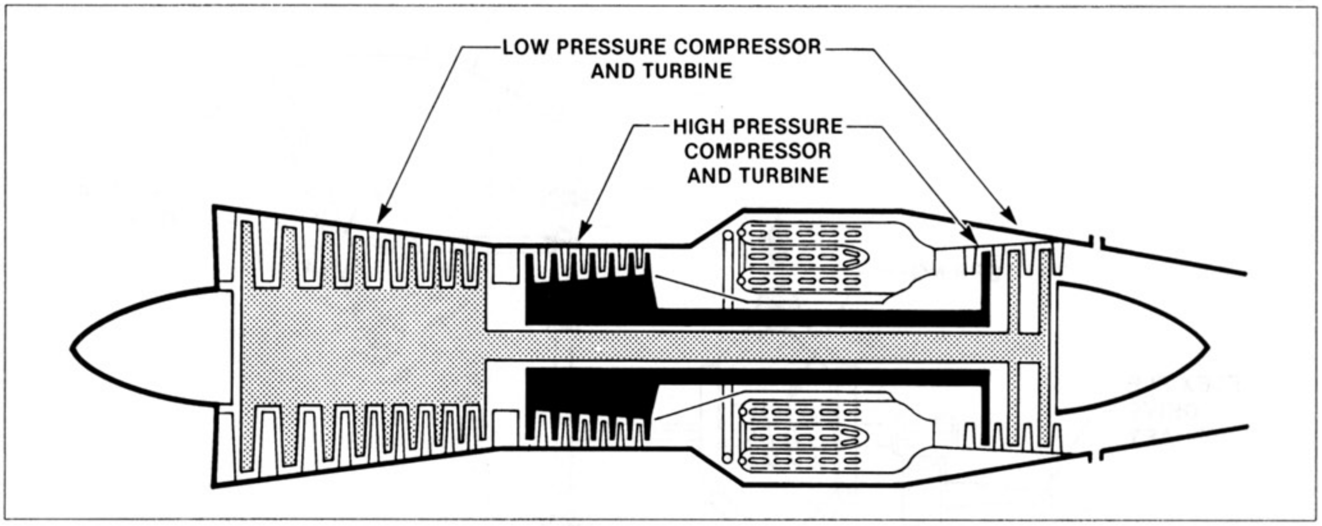
LOW PRESSURE COMPRESSOR AND TURBINE
HIGH PRESSURE COMPRESSOR AND TURBINE

Figure 2-40. Diagram of a twin-spool turbojet engine showing the low pressure compressor and high pressure compressor.

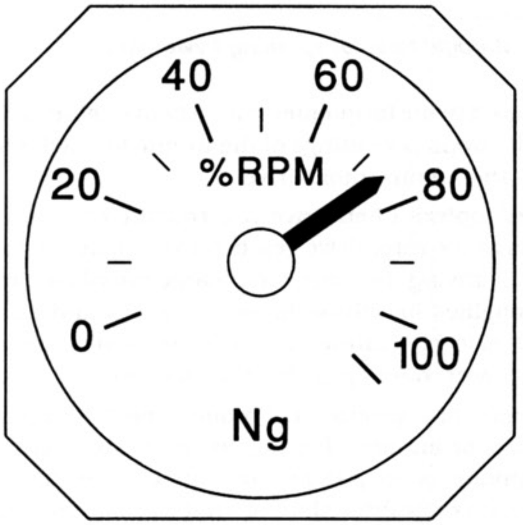
0
20
40
60
80
100
%RPM
Ng

70% - 102% CONTINUOUS OPERATION

102% - 104% TAKEOFF POWER
(5 MINUTES)

104% MAXIMUM TAKEOFF

105% ONE ENGINE INOPERATIVE
(30 MINUTE LIMIT)

106.5% MAXIMUM ONE ENGINE
INOPERATIVE
(2-1/2 MINUTE LIMIT)

Figure 2-41. The label Ng is commonly applied to the gas producer tachometer for turboprop and turboshaft engines. Some maximum values are time limited and some only apply when one engine has failed on a twin engine aircraft.

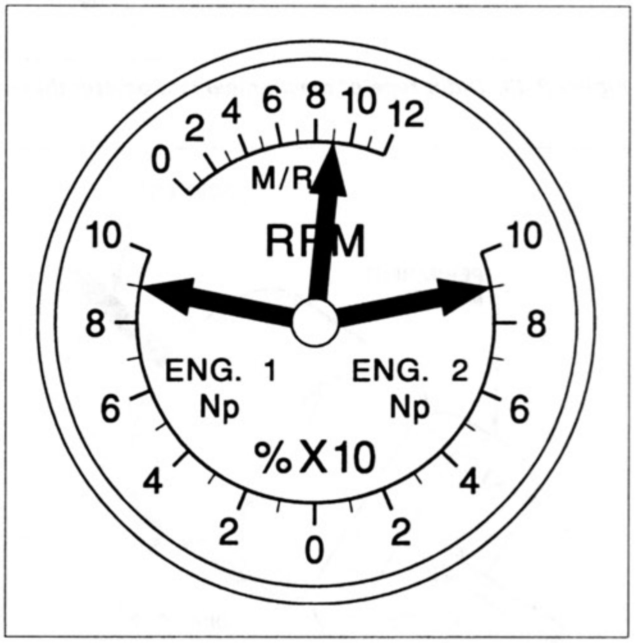
0
2
4
6
8
10
12
M/R
RPM
10
8
6
4
2
0
2
4
6
8
10
ENG. 1
Np
ENG. 2
Np
%X10

Figure 2-42. Twin-engine helicopters often use a triple tachometer that provides indications for both engines and the main rotor. The engine RPM is Np for twin spool engines since the indication is for the power turbine section.

on the engine or wherever the pressure source is located. The information will be transmitted to the cockpit gauges by electrical signals from a synchro transmitter or similar device that is located at the sensor end of the system. Figure 2-48 illustrates an oil pressure system with an Autosyn synchro. The same basic principle of operation would apply to fuel pressure gauges, hydraulic pressure gauges, and similar instruments.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

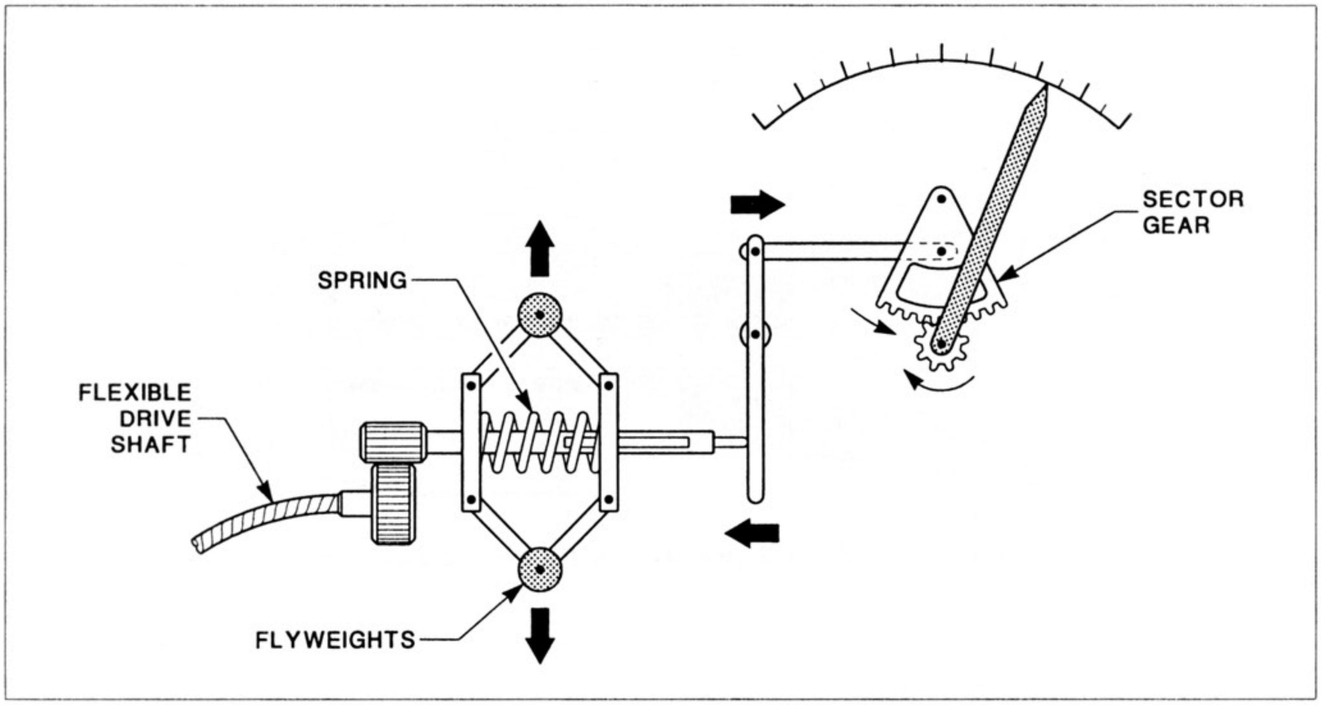

Figure 2-43. Older type of mechanical tachometer that used the centrifugal force of spinning flyweights.

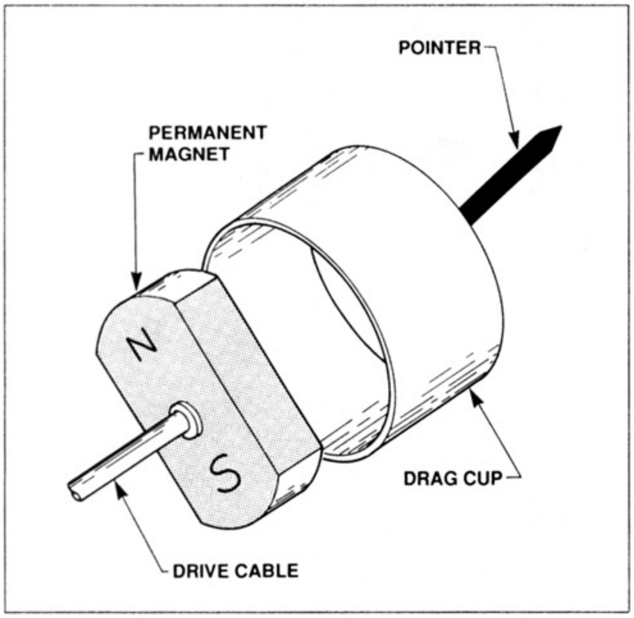

Figure 2-44. The permanent magnet and drag cup of a modern mechanical tachometer.

G. Torquemeters

Torquemeters are used to give an indication of the torque being produced by an engine or the torque being delivered to the main rotor drive of a helicopter. Turboprop airplanes and aircraft with turboshaft engines will have a torquemeter installed because it is the best way to measure the power being produced by these types of turbine engines. Figure 2-49 is a triple torquemeter for a twin engine helicopter. It supplies readings of the torque for each engine and the main rotor drive.

Helicopters often have a torquemeter that is located at the rotor drive gearbox to indicate the torque that is driving the main rotor. Large radial reciprocating engines like those found in DC-6s and Convair 240s also had torquemeters to accurately measure the power developed by the engines.

There are several techniques used to measure torque for an aircraft torquemeter instrument. One technique is to put sensors on a driveshaft, like the main rotor drive shaft of a helicopter. The sensors can be seen at the bottom of the main rotor drive shaft in figure 2-50 which illustrates this principle. The sensors measure the amount of twist in the shaft which is caused by the torque force. The electrical signals from the sensors are processed and used to position the cockpit indicator. Another technique which is used relies on a measurement of torque pressure. The sensor for this type of system is a small oil filled cylinder with a piston in it. The sensor would be located in the reduction gearbox. The reduction gearbox for a turboprop engine is shown in figure 2-51. The sensor is installed in the reduction gearbox so that the torque reaction force is applied to the piston and creates a build-up of pressure that is proportional to the torque force. Figure 2-52 shows another example of this type of torque sensor. The torque pressure is measured and causes the instrument pointer to show the

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

corresponding torque reading. The cockpit indicator for a torquemeter system may use a number of different units of measure. Those units of measure most often seen are horsepower, PSI, foot-pounds and percent.

H. Engine Pressure Ratio Indicators

This type of instrument is used on some kinds of turbojet and turbofan airplane engines. Those engines built by Pratt & Whitney and Rolls Royce use an engine pressure ratio gauge (EPR) as a primary engine instrument. Jet engines built by GE and Garrett usually do not have an EPR gauge, but use the N_1 tachometer in its place. As its name implies, the engine pressure ratio gauge indicates the ratio of two different pressures measured on the engine. The two pressures are most often called Pt_2 and Pt_7. The total inlet pressure at the front of the engine is Pt_2. The total outlet pressure at the aft end of the engine is Pt_7. Figure 2-53 illustrates the location of the probes and the differential pressure transducer. These pressures are called total pressures because the probes measure both static and dynamic pressure. The probes operate like pitot tubes since they measure ram pressure or total pressure in the airstream. The two probes are connected by tubing to a transducer mounted on the engine. The transducer is a differential pressure device that produces an electrical output related to the ratio of Pt_7/Pt_2. The transducer may use a synchro transmitter that is connected to the synchro receiver in the cockpit instrument. The Pt_2 probe in the front of the engine is prone to icing so it includes a heating system to prevent blockage by ice. The transducer and cockpit gauge for a typical EPR system is shown in figure 2-54. Engine pressure ratio gauges often include an index mark which can be set manually by the pilot. When the correct power setting for takeoff has been determined, a knob is used to set the bug or index mark to the correct value on the face of the instrument. During takeoff, the power is set by lining up the EPR gauge pointer with the bug. This makes it easier to set the correct level of engine power for takeoff.

I. Manifold Pressure Gauges

Manifold pressure gauges are only found on certain reciprocating engines where they are required to

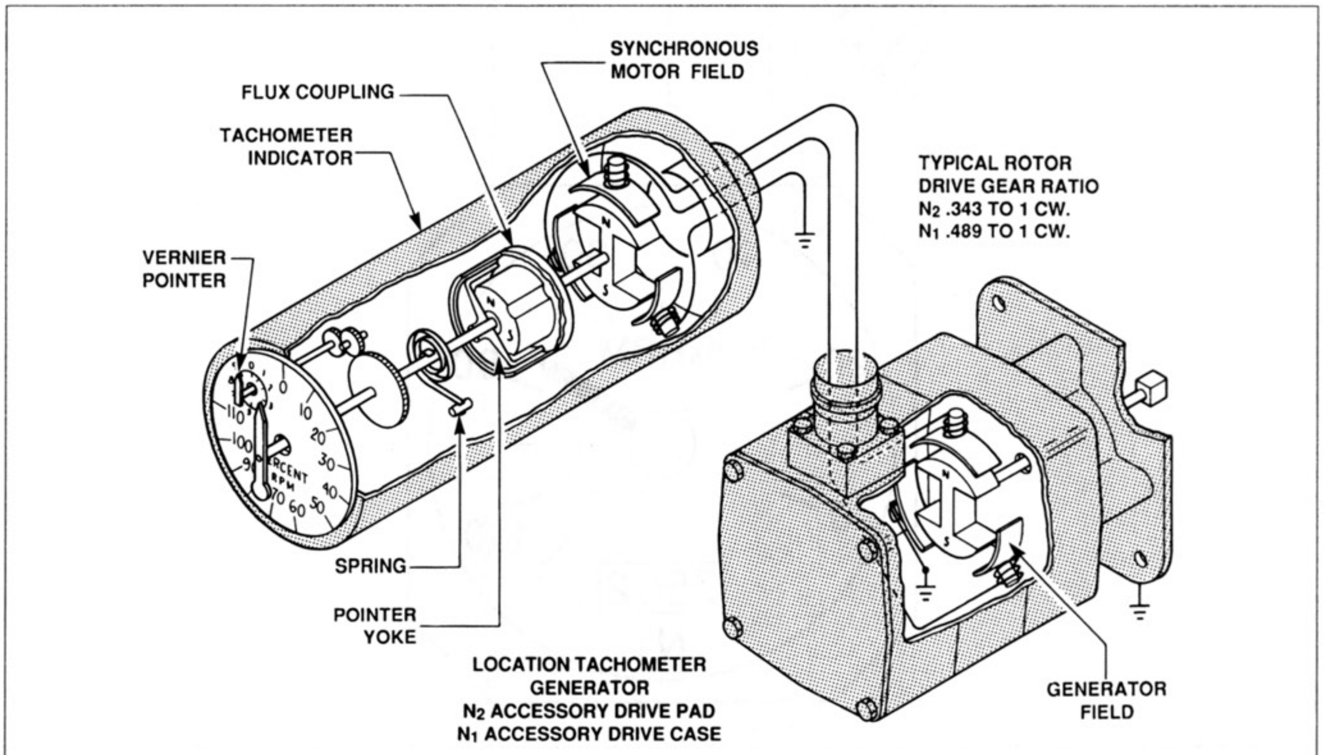
SYNCHRONOUS MOTOR FIELD
FLUX COUPLING
TACHOMETER INDICATOR
VERNIER POINTER
TYPICAL ROTOR DRIVE GEAR RATIO
N_2 .343 TO 1 CW.
N_1 .489 TO 1 CW.
PERCENT RPM
0
10
20
30
40
50
60
70
90
100
110
N
S
SPRING
POINTER YOKE
LOCATION TACHOMETER GENERATOR
N_2 ACCESSORY DRIVE PAD
N_1 ACCESSORY DRIVE CASE
GENERATOR FIELD

THE THREE-PHASE GENERATOR IS DRIVEN BY THE ENGINE TO PRODUCE AC WHOSE FREQUENCY RELATES TO ENGINE RPM. THE INDICATOR HOLDS A SYNCHRONOUS MOTOR WHICH DRIVES A MAGNETIC DRAG TACHOMETER MAGNET.

Figure 2-45. AN AC tach generator system.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

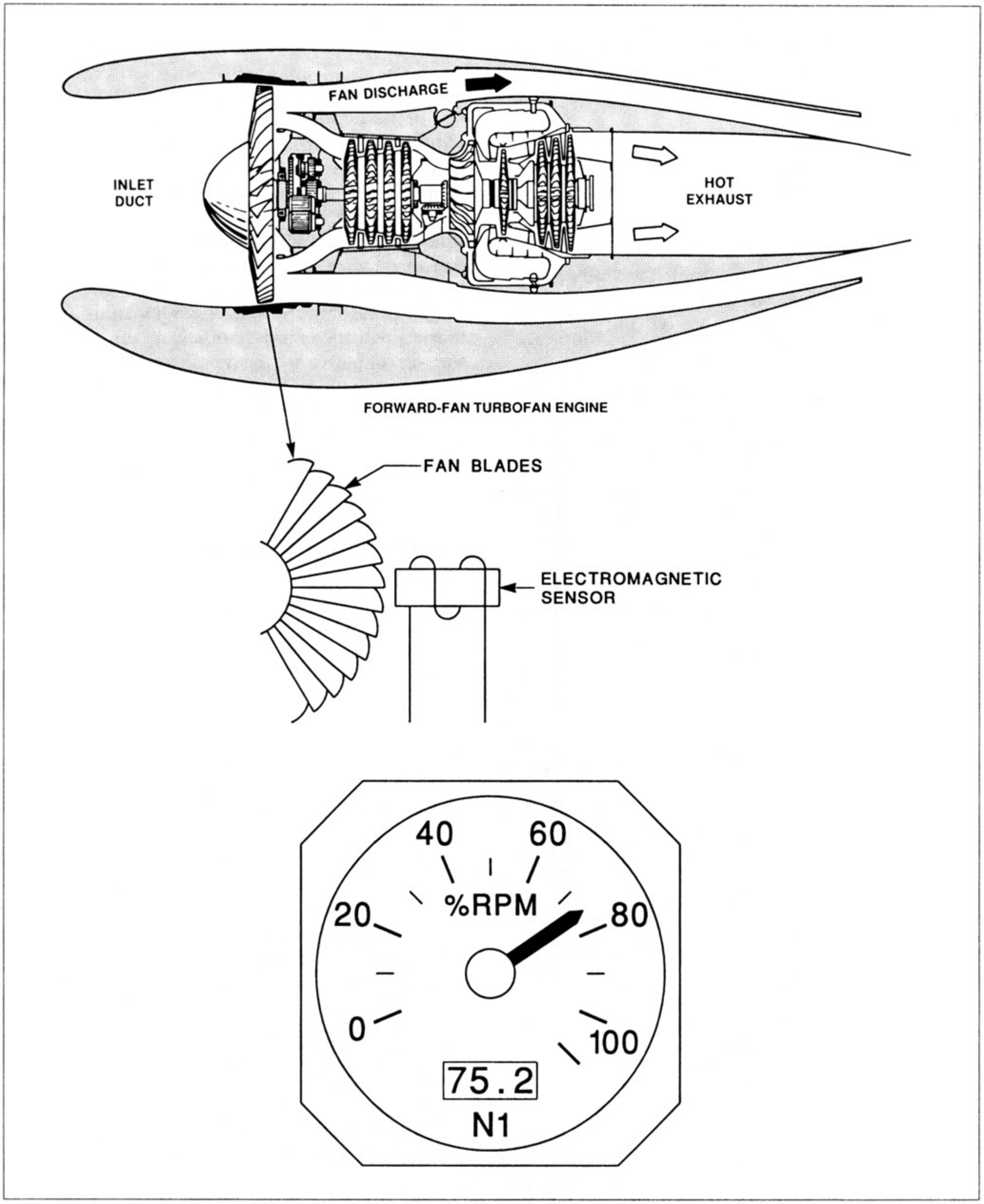
FAN DISCHARGE
INLET DUCT
HOT EXHAUST
FORWARD-FAN TURBOFAN ENGINE
FAN BLADES
ELECTROMAGNETIC SENSOR
40
60
20
%RPM
80
0
100
75.2
N1

Figure 2-46. Electronic tachometer which measures fan speed for the N_1 tachometer. The sensors produce an output signal each time a fan blade cuts through the magnetic field. The frequency of this output signal is measured to provide indications of N_1 RPM.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

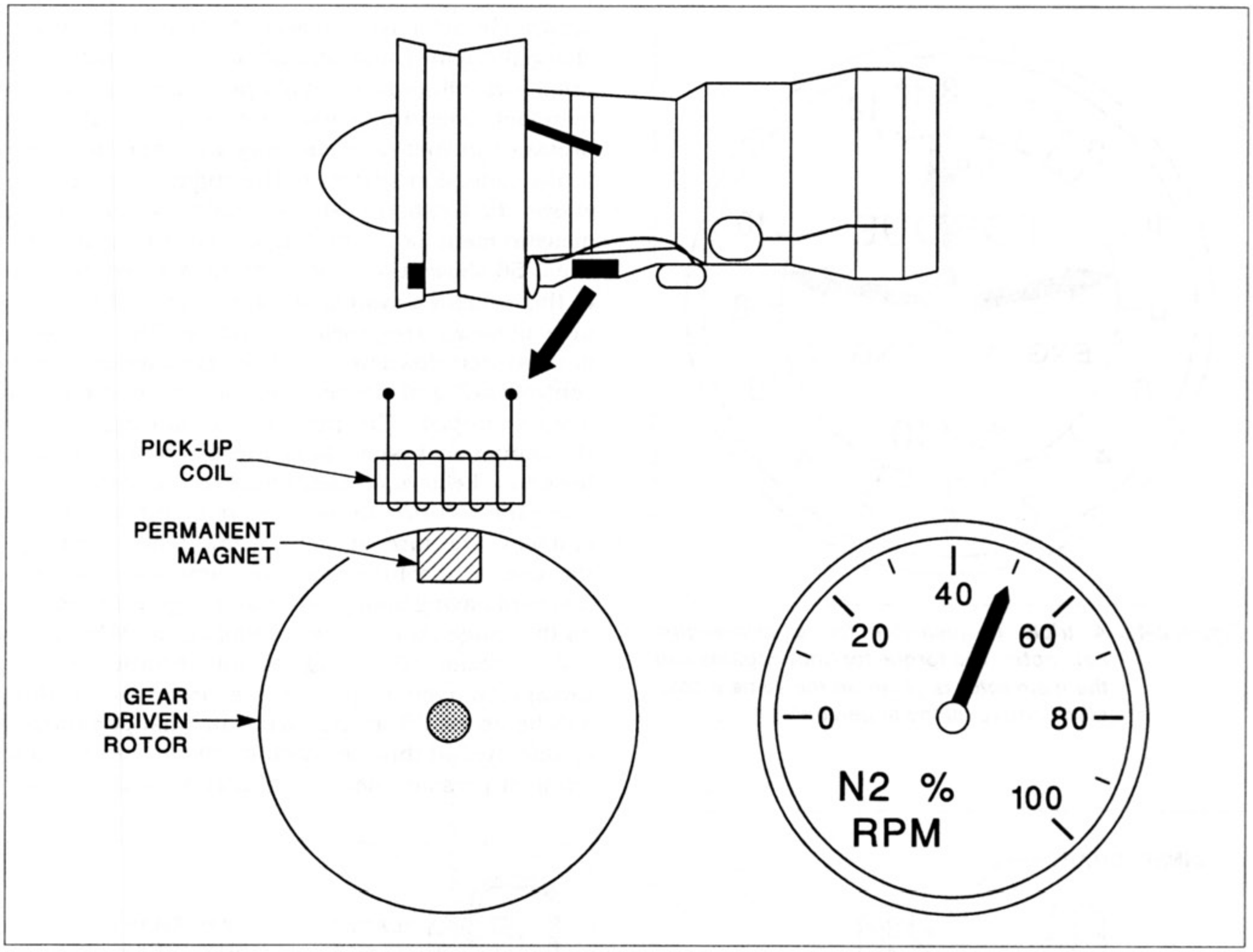
PICK-UP COIL
PERMANENT MAGNET
GEAR DRIVEN ROTOR
0
20
40
60
80
100
N2 %
RPM

Figure 2-47. Tachometer system for the high pressure compressor of a turbofan engine. The gear-driven rotor from the accessory section has a permanent magnet which induces signals in a pick-up coil.

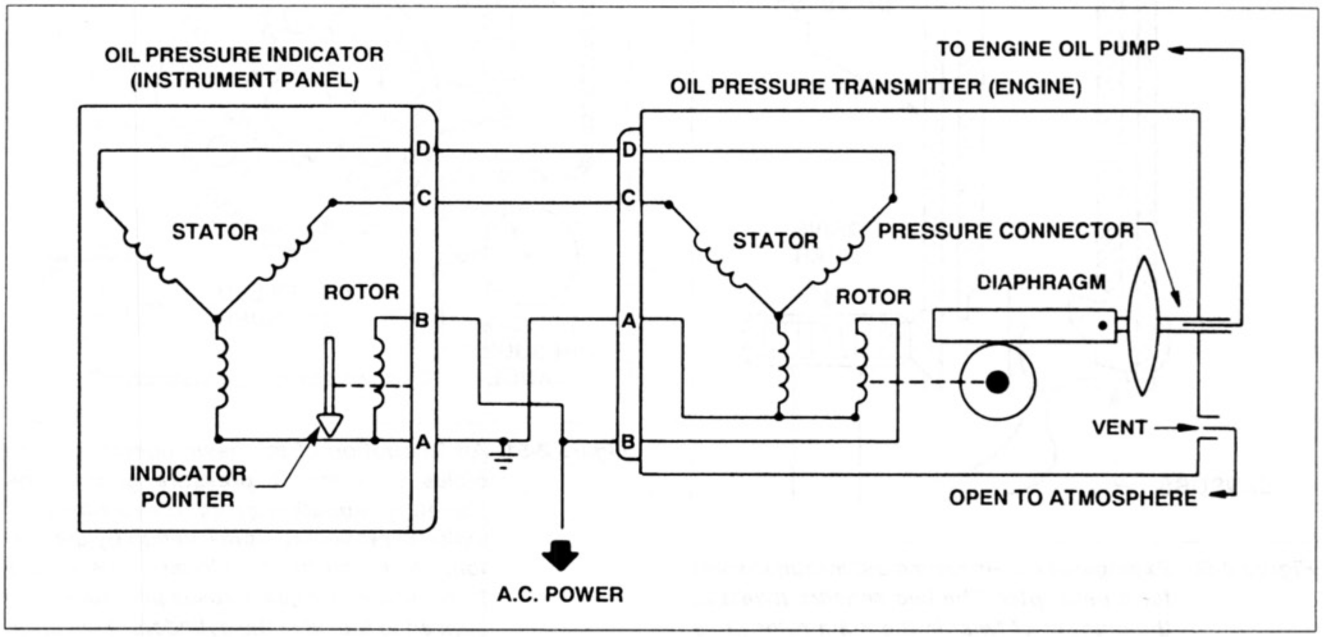
OIL PRESSURE INDICATOR (INSTRUMENT PANEL)
STATOR
ROTOR
D
C
B
A
INDICATOR POINTER
A.C. POWER
TO ENGINE OIL PUMP
OIL PRESSURE TRANSMITTER (ENGINE)
D
C
A
B
STATOR
ROTOR
PRESSURE CONNECTOR
DIAPHRAGM
VENT
OPEN TO ATMOSPHERE

Figure 2-48. An oil pressure indicating system for a large airplane.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

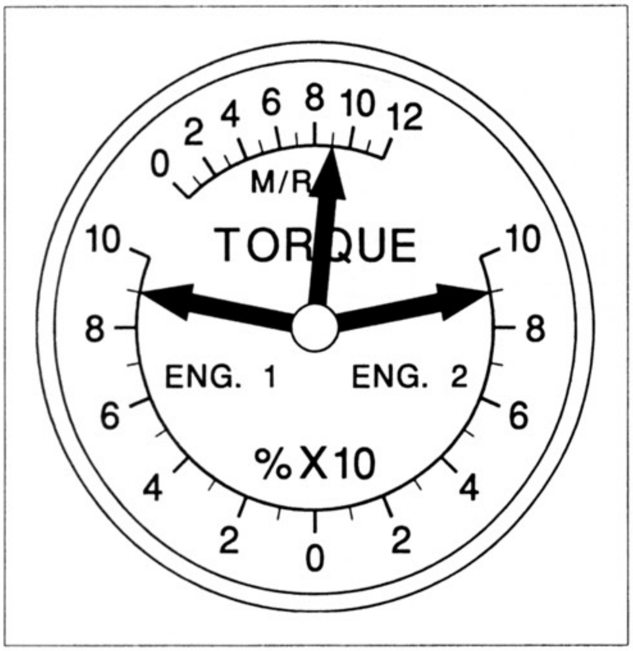

Figure 2-49. ***A triple torquemeter for a twin-engine helicopter. The torque for both engines and the main rotor is given on the same instrument with readings in percent.***

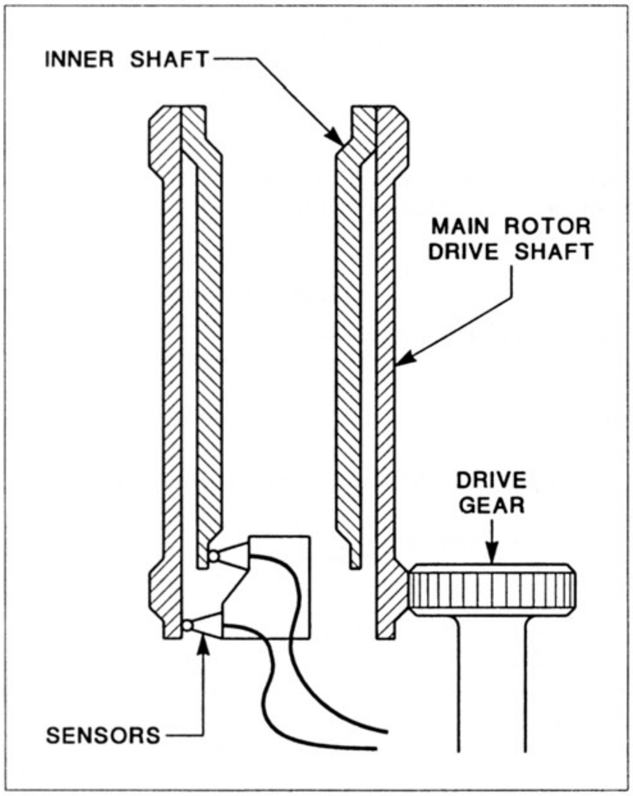

Figure 2-50. ***Example of a main rotor mast torque system for a helicopter. The two sensors measure the amount of twist in the main rotor drive shaft as an indication of torque.***

accurately set engine power. Aircraft with supercharged engines and aircraft with constant-speed propellers will have manifold pressure gauges. The manifold pressure gauge measures the absolute pressure in inches of mercury at a specific point in the induction system of the engine. Figure 2-55 shows the location of the manifold pressure (MAP) measurement for a radial supercharged engine. Figure 2-56 shows the location of the MAP connection in the induction system of a turbocharged horizontally opposed reciprocating engine. The pressure is measured downstream of the carburetor or fuel control unit and downstream of the supercharger if so equipped. The pressure measuring port in the induction system is connected by tubing and hose to a bellows or diaphragm in the instrument. Since the pressure in the induction system is below ambient pressure at idle or low power settings, the use of absolute pressure eliminates the confusion of having both positive and negative numbers on the gauge. At idle, the reading on the MAP gauge will be about 10 in. Hg. At full throttle with an unsupercharged engine at sea level, the reading will be about 28 in. Hg. With an unsupercharged engine the full throttle reading will always be below ambient pressure because of friction and pressure

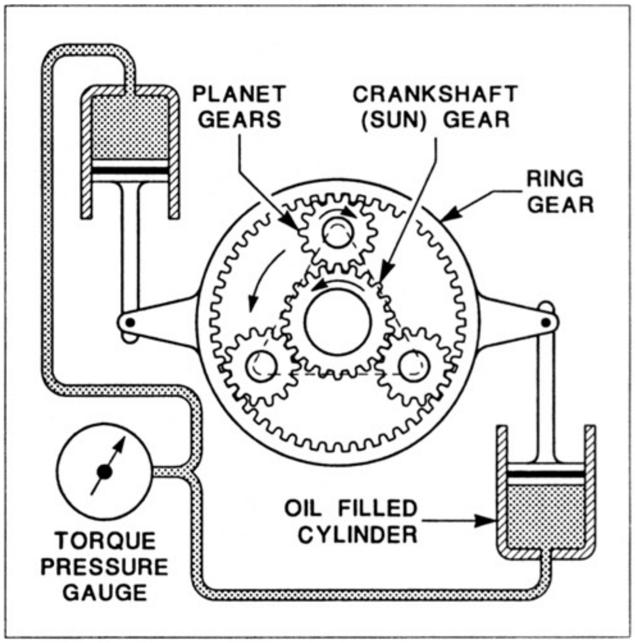

Figure 2-51. ***An illustration of the basic operating principles of a torque pressure gauge. The planetary reduction gears have a ring gear which is prevented from rotating by the pistons in the oil filled cylinders. The torque force on the ring gear causes pressure to be applied to the oil in the cylinders. This pressure is a measurement of torque.***

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

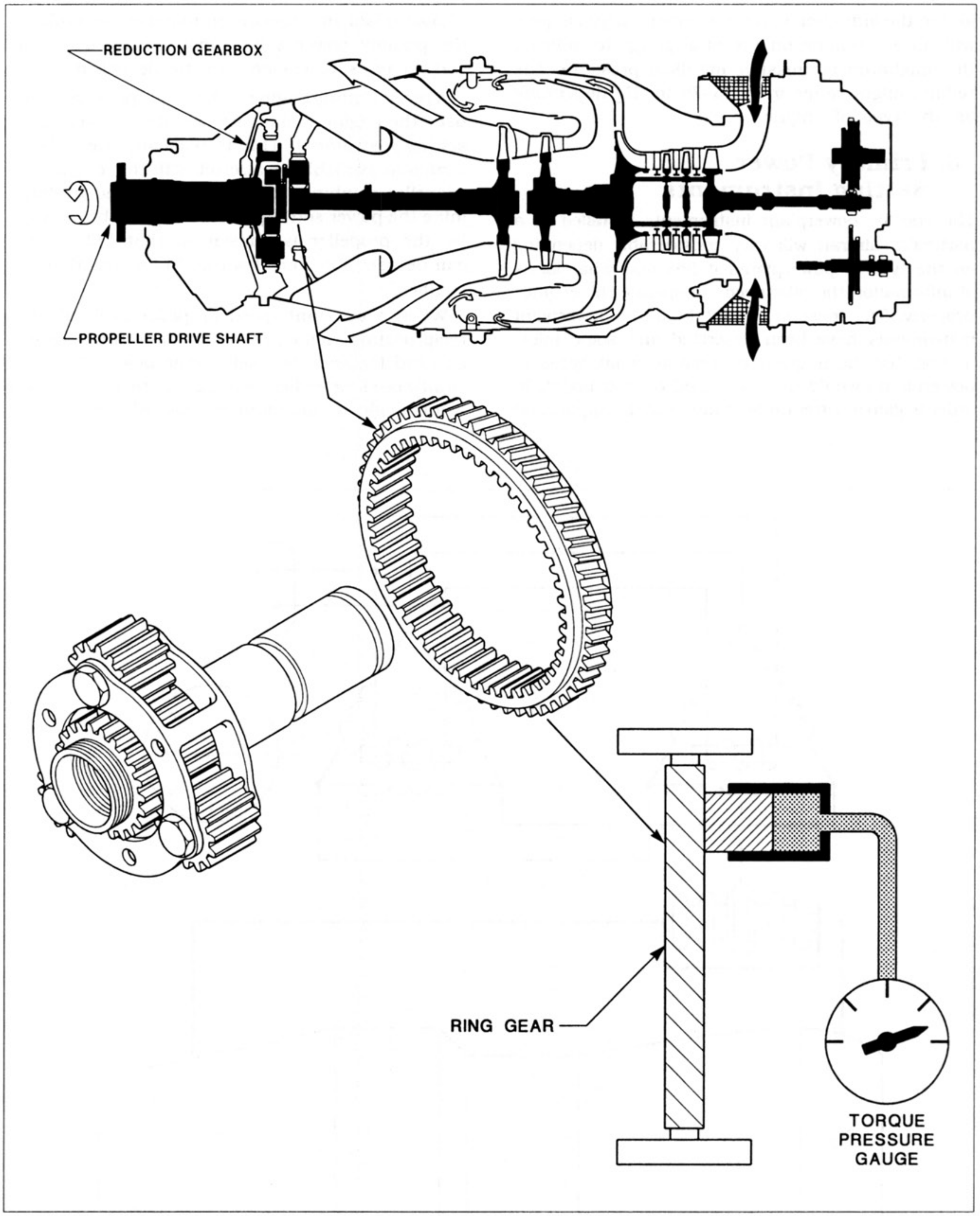
REDUCTION GEARBOX
PROPELLER DRIVE SHAFT
RING GEAR
TORQUE
PRESSURE
GAUGE

Figure 2-52. The torquemeter system for a turboprop engine. The ring gear of the planetary gear set is prevented from rotating by the helical splines on the outer circumference that mesh with splines in the outer housing. Torque force causes the ring gear to more rearward and push against the small piston. The piston applies pressure to the oil in the cylinder and this pressure is a measure of the torque produced.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

loss in the induction system. A supercharged engine will have a redline on the MAP gauge to indicate the maximum permissible manifold pressure. The redline might range from 35–75 in. Hg depending on the type of engine.

J. Primary Power Setting Instruments

The specific powerplant instruments installed in a particular aircraft will vary considerably depending on the type of powerplants it has and what kinds of information the pilot needs to operate the engines properly. The most important kinds of powerplant instruments have been described and some information has been given concerning what types of powerplants would use each kind of instrument. In order to gain a better understanding of the application of powerplant instruments to different powerplants, the primary power setting instruments used with various types of engines will be described.

A reciprocating engine with a fixed-pitch propeller uses the engine tachometer as the primary power setting instrument. This is normally the only instrument available on aircraft with this engine and propeller combination that can be used to determine the power setting. The fixed-pitch blade angle for the propeller is chosen so that full throttle can be used for takeoff without over-speeding the engine.

When a constant-speed propeller is fitted on a reciprocating engine, the tachometer alone cannot be used to accurately set engine power. The constant-speed propeller will automatically vary the blade angle to maintain a selected RPM.

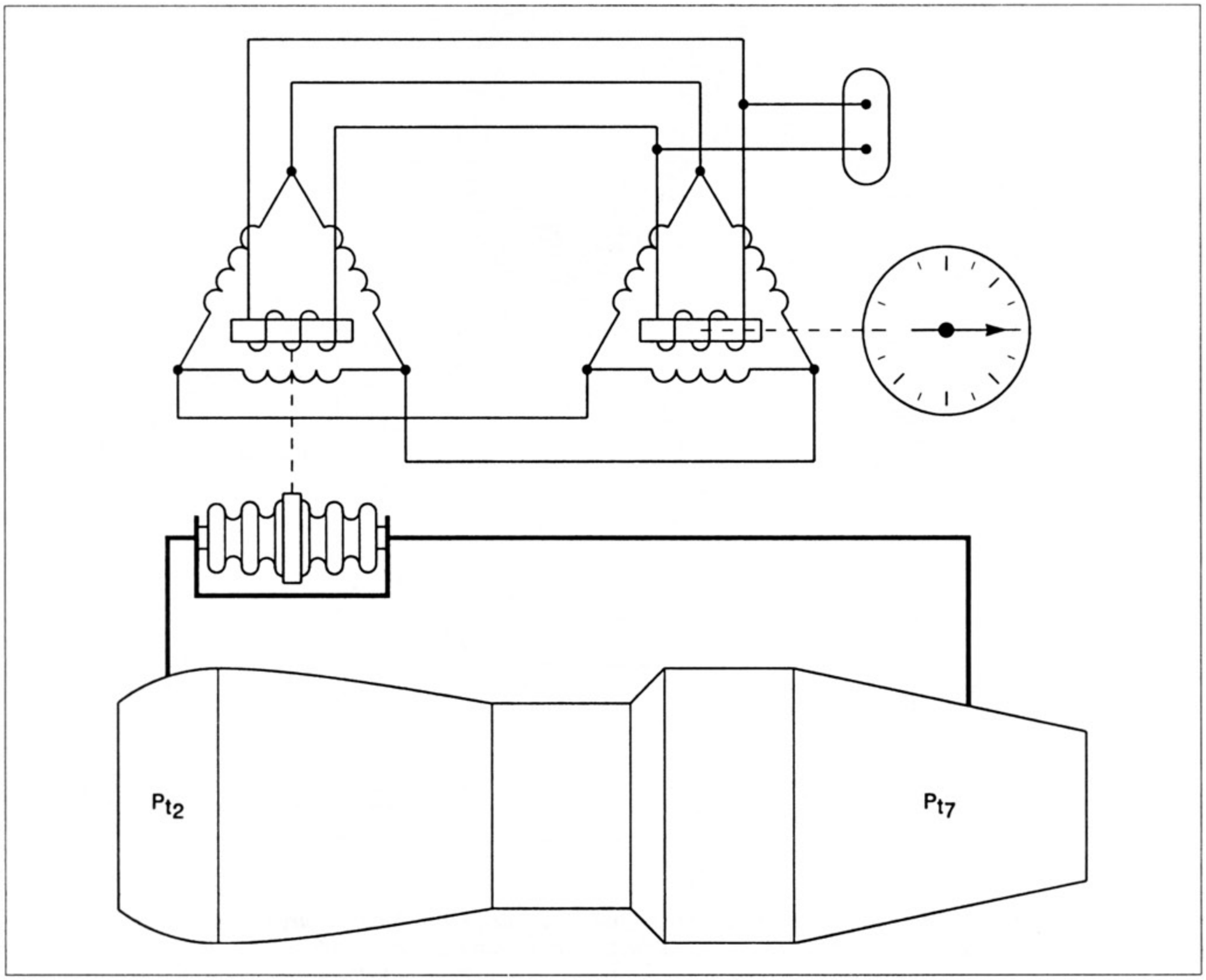

Figure 2-53. The engine pressure ratio system for a jet engine.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Fairly large movements of the throttle lever can be made without affecting the engine RPM, so that the tachometer by itself cannot be used to set engine power. The primary power setting instrument for this type of aircraft is the manifold pressure gauge. A cruise power setting would be made by first setting the desired engine RPM with the propeller control and then adjusting the manifold pressure gauge to the desired power setting. Information is available in the Pilot's Operating Handbook so the pilot can determine what settings will produce a given percent of power for cruise. Common cruise settings would be 55, 65 or 75 percent of maximum engine power. Supercharged reciprocating engines also use the manifold pressure gauge as the primary power setting instrument.

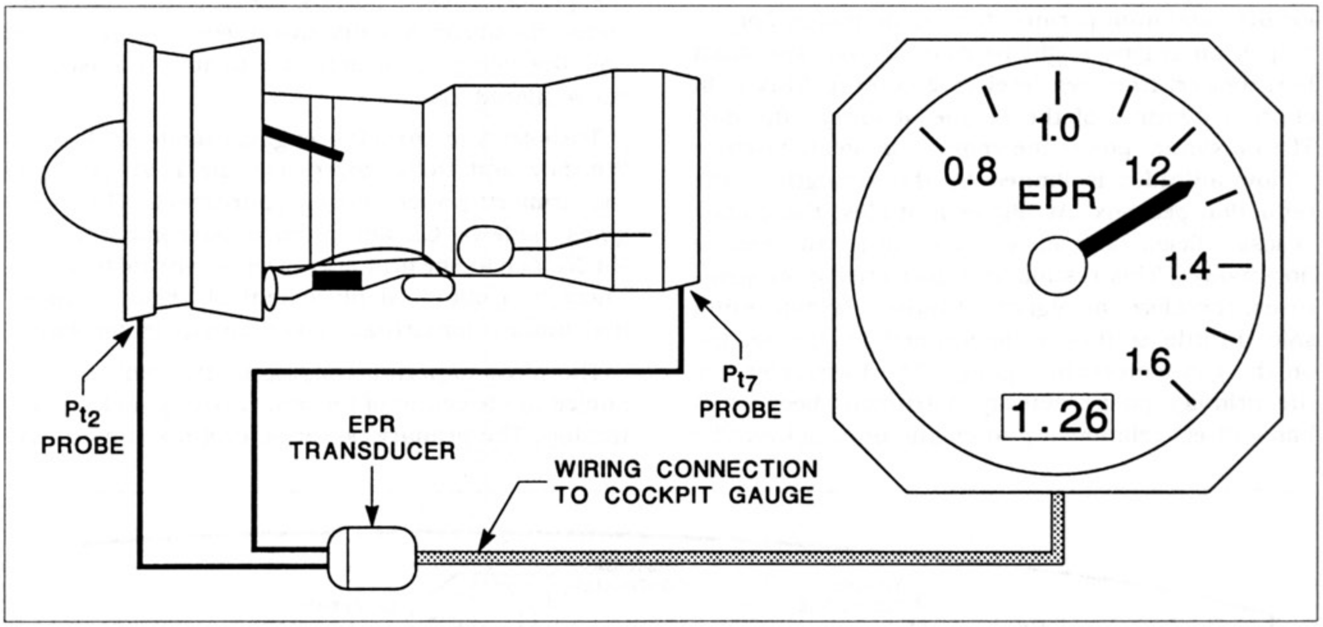
1.0
0.8
1.2
EPR
1.4
1.6
1.26
P_{t_2} PROBE
P_{t_7} PROBE
EPR TRANSDUCER
WIRING CONNECTION TO COCKPIT GAUGE

Figure 2-54. The EPR gauge measures engine pressure ratio which is the ratio of Pt_7 / Pt_2. It provides indications of the thrust being produced by a turbojet or turbofan engine.

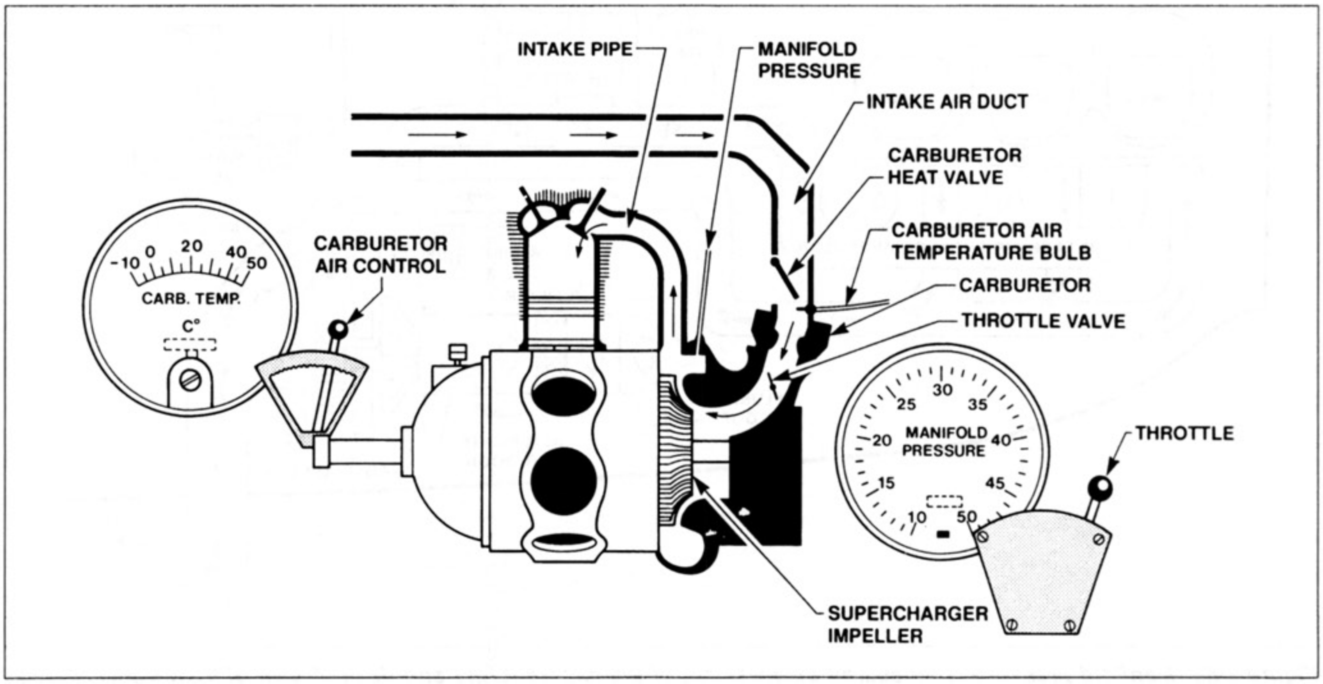
INTAKE PIPE
MANIFOLD PRESSURE
INTAKE AIR DUCT
CARBURETOR HEAT VALVE
CARBURETOR AIR TEMPERATURE BULB
CARBURETOR
THROTTLE VALVE
CARBURETOR AIR CONTROL
-10 0 20 40 50
CARB. TEMP.
C°
10 15 20 25 30 35 40 45 50
MANIFOLD PRESSURE
THROTTLE
SUPERCHARGER IMPELLER

Figure 2-55. The measurement point for manifold pressure is downstream of the carburetor and downstream of the supercharger in this radial engine.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The primary power setting instrument for aircraft with turboshaft and turboprop engines is the torquemeter. Engine RPM is not a good measure of engine power because of the way these engines operate. The gas temperature gauge is a very important instrument for these types of engines. The pilot needs to monitor the gas temperature reading to prevent damage to the engine by excessive heat. Turboshaft and turboprop engines are often given a flat rating for the maximum permissible engine power. For example, an engine might be rated at 575 SHP (shaft horsepower) from sea level to 20,000 ft. This is in effect a derating of the engine at lower altitudes. The maximum power the engine can safely produce at low altitudes is limited by the strength of the reduction gearbox. At higher altitudes, the engine is less efficient and must work harder to produce horsepower. This results in higher engine temperatures; therefore, at higher altitudes, the maximum safe throttle setting is determined by the redline on the gas temperature gauge. The torquemeter is the primary power setting instrument because it has a direct relationship to engine horsepower. The maximum power setting that can be safely used is determined by the strength of the gearbox at low altitudes and the gas temperature at higher altitudes.

The primary power setting instrument for turbojet and turbofan engines depends on who manufactured the engine. Some of these engines use the EPR gauge and others use the N_1 tachometer as the primary power setting instrument. There are some advantages and disadvantages to each method and the choice of which one to use is based on conventional usage.

Turbojet and turbofan engines made by Pratt & Whitney and Rolls Royce use the EPR gauge as the primary power setting instrument. Those engines made by GE and a few other companies use the N_1 tach as the primary power setting instrument. There is a standard placement of primary engine instruments for virtually all air carrier jet airplanes.

The most important engine instruments are installed in the center of the instrument panel, toward the top. The primary power setting instrument will

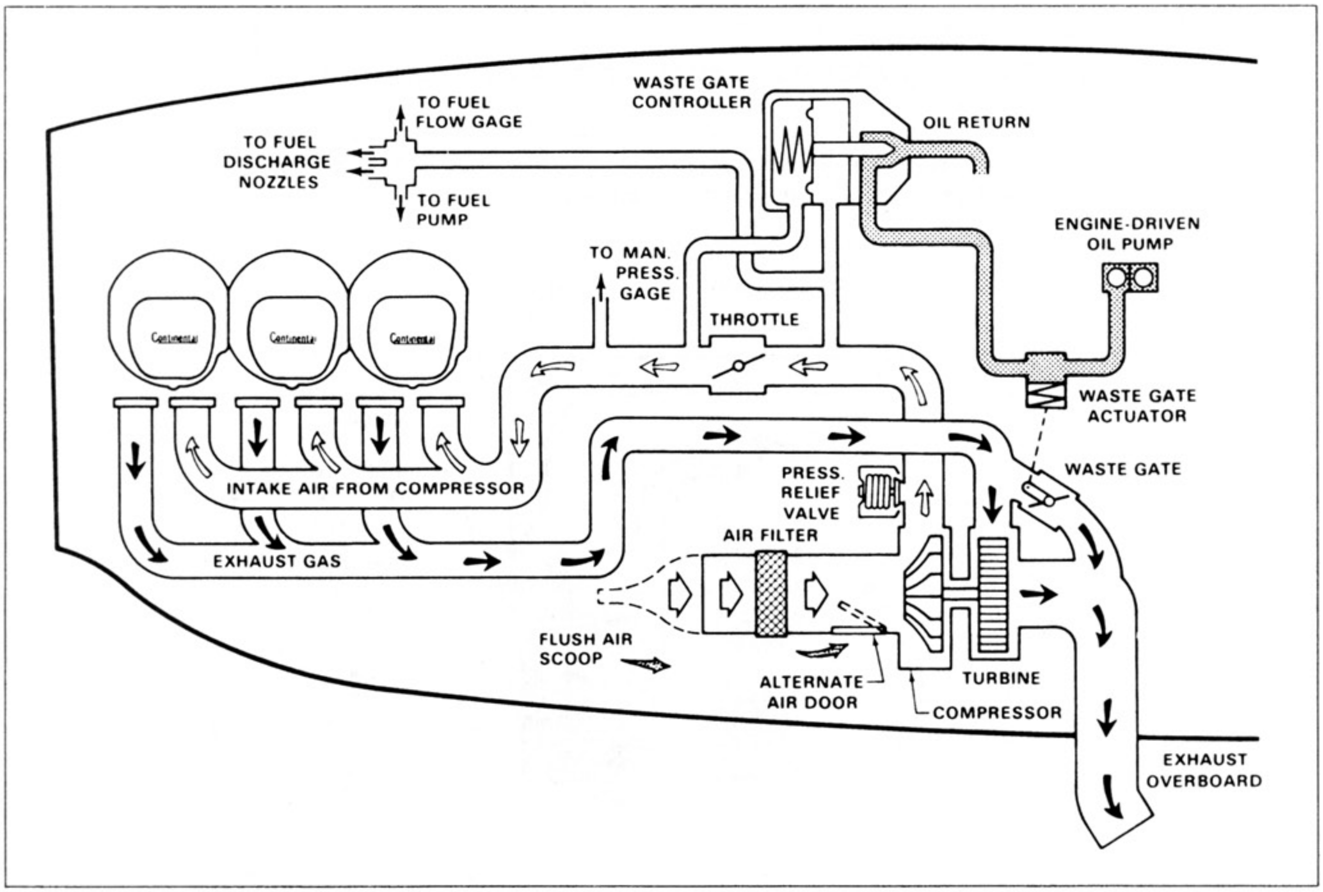
WASTE GATE CONTROLLER
TO FUEL FLOW GAGE
TO FUEL DISCHARGE NOZZLES
TO FUEL PUMP
OIL RETURN
ENGINE-DRIVEN OIL PUMP
TO MAN. PRESS. GAGE
THROTTLE
Continental
Continental
Continental
WASTE GATE ACTUATOR
WASTE GATE
INTAKE AIR FROM COMPRESSOR
PRESS. RELIEF VALVE
AIR FILTER
EXHAUST GAS
FLUSH AIR SCOOP
ALTERNATE AIR DOOR
TURBINE
COMPRESSOR
EXHAUST OVERBOARD

Figure 2-56. Manifold pressure is measured downstream of the supercharger and downstream of the throttle plate in the fuel-air control unit for a turbo-supercharged, fuel-injected reciprocating engine. (Courtesy Cessna Aircraft Co.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

be the one at the top of this stack of engine instruments. The EPR gauge will be at the top of the stack on some air carrier jets, and the N_1 tach on other airplanes. Figure 2-57 shows the primary engine instruments for an air carrier jet with Pratt & Whitney engines. Figure 2-58 shows the primary engine instruments for an air carrier jet with GE engines.

K. Vibration Indicators

Aircraft with turbine engines often have a vibration indicator system which monitors vibration from the engines. Any significant imbalance in a turbine engine can cause serious damage due to the very high rotational speeds of the rotors. Vibration caused by an imbalance or other factors is indicated to the crew in the cockpit so that they can take appropriate action. The vibration indicator system shown in figure 2-59 is typical of the type found on modern jet airplanes. The sensor is a piezoelectric crystal which produces an electrical signal when it is vibrated. This signal is sent to a signal conditioner and then to the cockpit instrument to provide an indication of the amount of vibration at the sensor location on the side of the engine.

L. Logic Circuits and Digital Systems

Logic circuits and microprocessors have made possible some of the very sophisticated electronic instruments and similar systems on modern aircraft. They are used for many different applications from relatively simple switching functions to complex computer systems. Certain basic principles are involved in these advanced systems, and these will be covered first.

1. Binary Numbers

The binary number system and binary codes are the method used by logic gates to transmit and process information. The word digital refers to the use of binary numbers and codes. A simple example of the concept of digital and analog type signals is shown in figure 2-60. The values of voltage and current in the circuits represent the digital and analog systems. In a circuit with a variable resistor, there are an infinite number of different values for current flow. Another example of this would be a simple fuel quantity system that uses a float operated variable resistor. The resistor moves in very small increments so that there are an infinite number of different current flows that can occur in the circuit. The circuit with the switch illustrates the concept of digital values. The switch is either on or off with no in-between settings. This follows the binary or digital signal system because only two different values are used. The two conditions in a binary or digital circuit are called 1 and 0, or high and low, or on and off.

The binary number system is a base 2 number system. The decimal number system that we are more familiar with is a base 10 system. The decimal system uses 10 different digits to make up numbers. The 10 numbers are 0 through 9. By showing some examples of converting numbers from binary to decimal and vice versa, the binary system will be better understood.

The procedure for converting a number from binary to decimal is illustrated in figure 2-61. We start by writing down the binary number as shown by row (A). Then just above each binary digit, the base 2 equivalent for that digit is written as shown by row (B). In row (C), the decimal equivalents are entered starting at the right and working back to the left. Notice that the decimal equivalents start with 1 at the right and are doubled each digit as you move toward the left. Next you look at the binary digits in row (A) and wherever there is a binary 1 you bring down the decimal equivalent as shown in row (D). Where there is a binary zero in row (A), you do not bring down any decimal equivalent since binary zero and decimal zero have the same value. Finally, the decimal equivalents are added up to produce the decimal equivalent of 77.

The procedure for converting a decimal number to binary will be illustrated by converting the decimal number 77 to its binary equivalent. The technique for converting from decimal to binary consists of a series of divisions by the number two. We start by setting up a table with three columns as shown in figure 2-62. The first step is to divide 77 by 2 and enter the quotient and remainder in the proper column. The quotient is then brought down each time and divided by 2 and the values recorded. The process is repeated until the quotient is zero. The digits in the remainder column are read from the bottom up and this will be the binary equivalent.

There are some terms used with binary numbers that refer to how the binary digits are grouped. For example, in the binary number [101 110], there are two sets of three digits each. In digital or binary terminology, we would say that there are three bits in each byte. A bit is a binary digit, while a byte is a group of bits together. As an analogy, we might compare them to letters and words. The previous example would be like a group of words with three letters in each word. This is what is meant by binary codes, how the binary digits are grouped.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

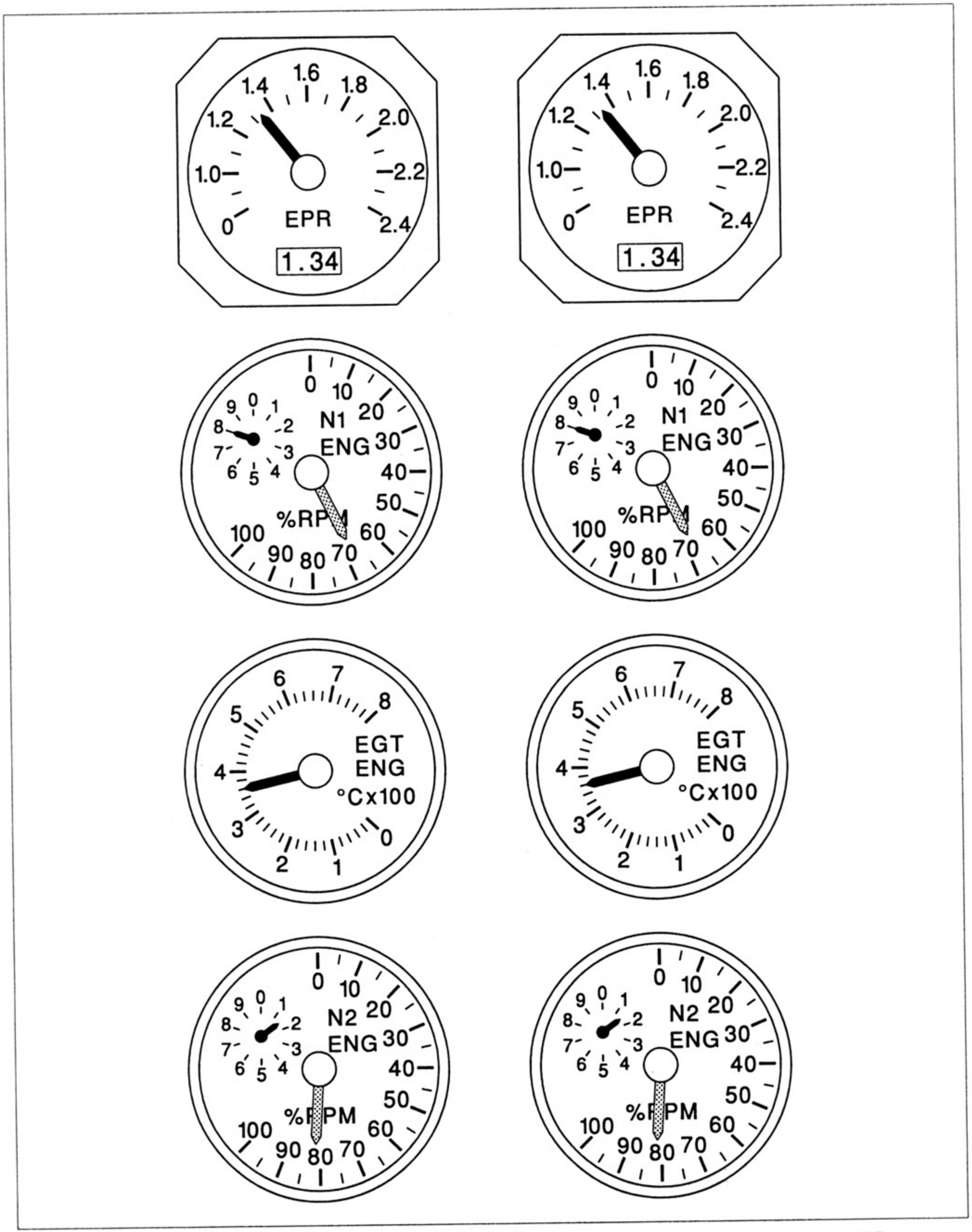
1.4
1.6
1.8
1.2
2.0
1.0
2.2
0
EPR
2.4
1.34
1.4
1.6
1.8
1.2
2.0
1.0
2.2
0
EPR
2.4
1.34
0
10
20
30
40
50
60
70
80
90
100
9
0
1
2
3
4
5
6
7
8
N1
ENG
%RPM
0
10
20
30
40
50
60
70
80
90
100
9
0
1
2
3
4
5
6
7
8
N1
ENG
%RPM
6
7
8
5
4
3
2
1
0
EGT
ENG
°Cx100
6
7
8
5
4
3
2
1
0
EGT
ENG
°Cx100
0
10
20
30
40
50
60
70
80
90
100
9
0
1
2
3
4
5
6
7
8
N2
ENG
%RPM
0
10
20
30
40
50
60
70
80
90
100
9
0
1
2
3
4
5
6
7
8
N2
ENG
%RPM

Figure 2-57. Powerplant instruments for an air carrier jet with Pratt and Whitney engines that utilize an EPR gauge.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

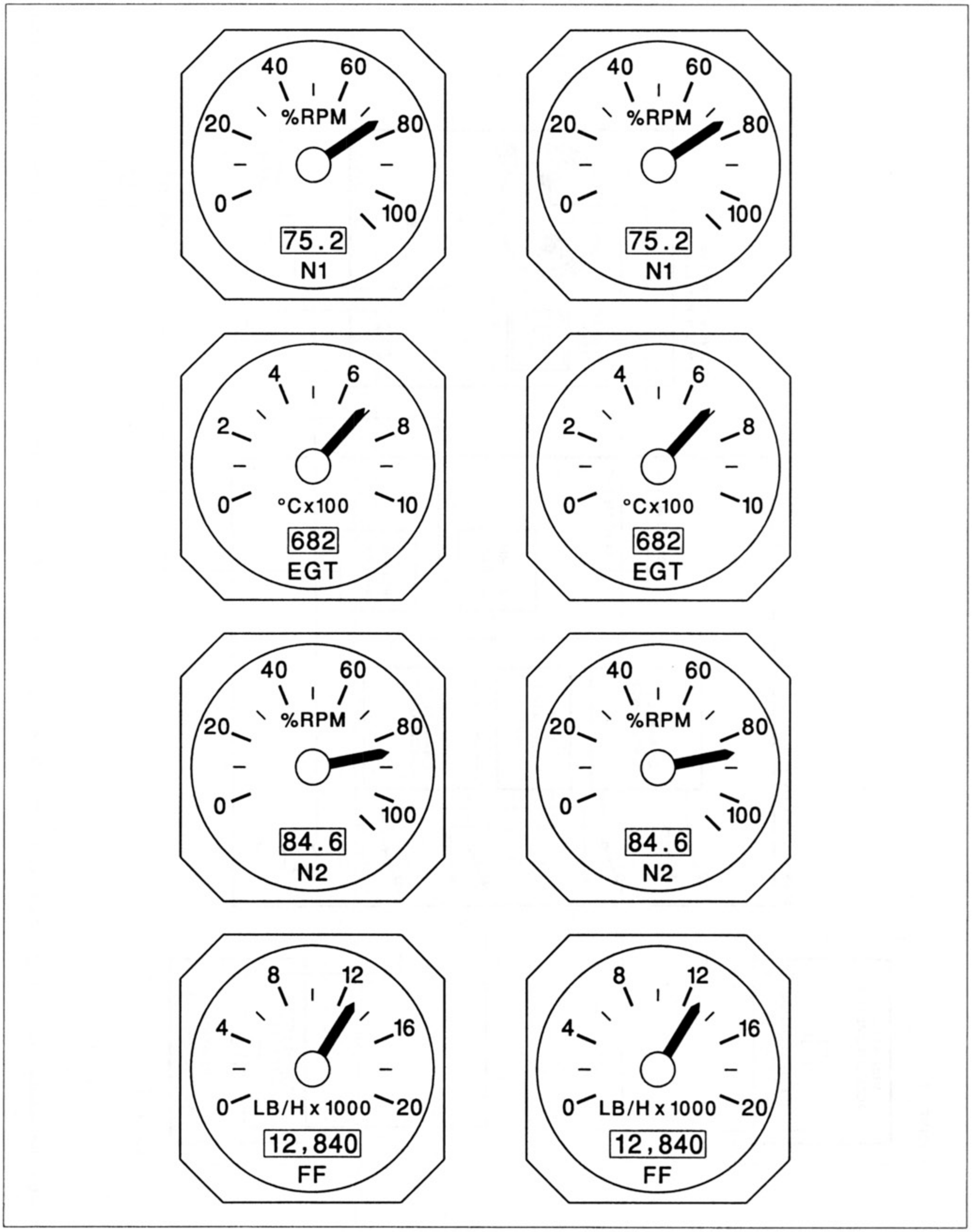
%RPM
0
20
40
60
80
100
75.2
N1
%RPM
0
20
40
60
80
100
75.2
N1
°Cx100
0
2
4
6
8
10
682
EGT
°Cx100
0
2
4
6
8
10
682
EGT
%RPM
0
20
40
60
80
100
84.6
N2
%RPM
0
20
40
60
80
100
84.6
N2
LB/H x 1000
0
4
8
12
16
20
12,840
FF
LB/H x 1000
0
4
8
12
16
20
12,840
FF

Figure 2-58. Powerplant instrument layout for an air carrier jet with G.E. engines and the N_1 tachs at the top of the stack.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

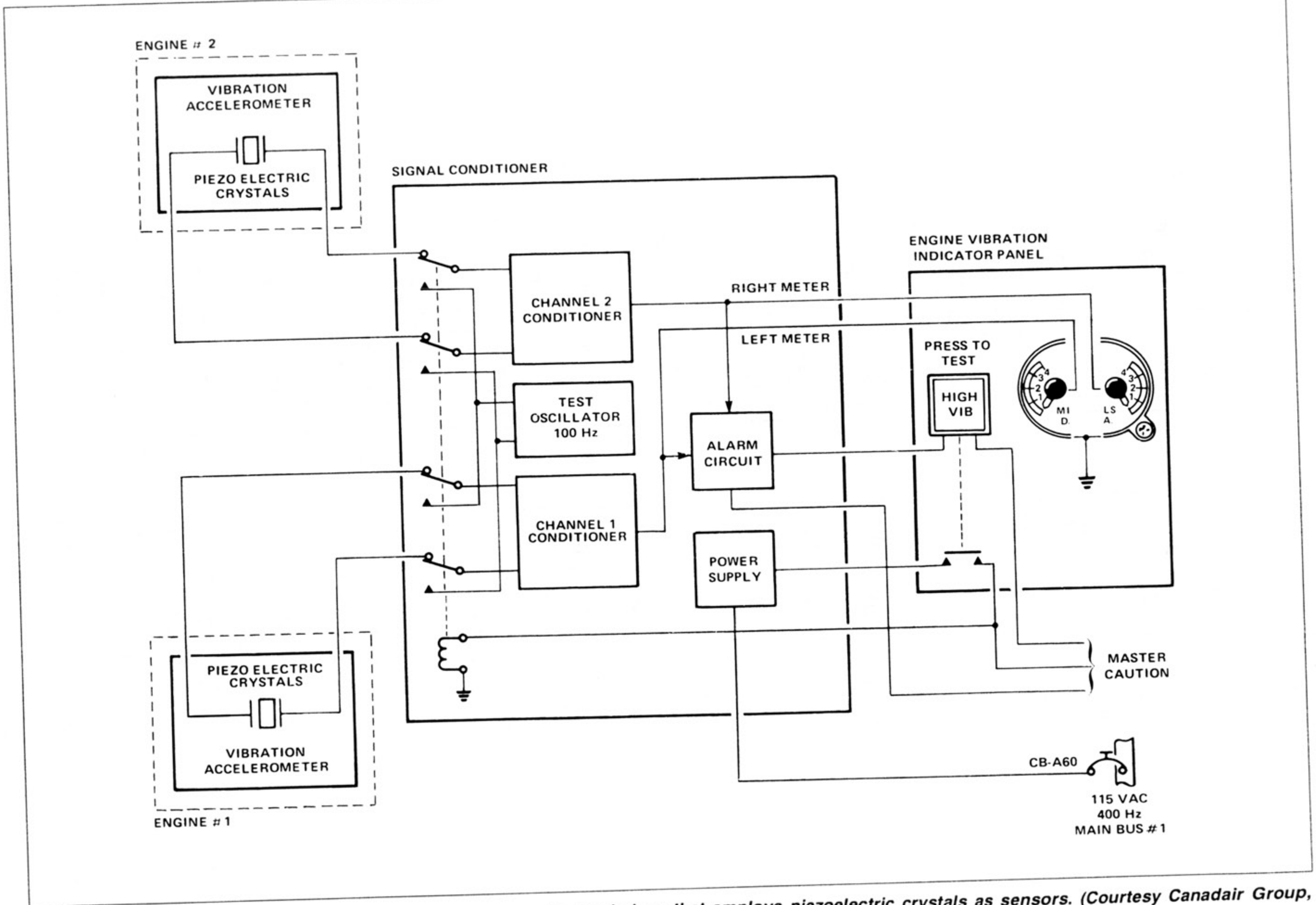
ENGINE # 2
VIBRATION ACCELEROMETER
PIEZO ELECTRIC CRYSTALS
SIGNAL CONDITIONER
CHANNEL 2 CONDITIONER
TEST OSCILLATOR 100 Hz
CHANNEL 1 CONDITIONER
RIGHT METER
LEFT METER
ALARM CIRCUIT
POWER SUPPLY
ENGINE VIBRATION INDICATOR PANEL
PRESS TO TEST
HIGH VIB
4
3
2
1
MI D.
LS A.
MASTER CAUTION
CB-A60
115 VAC
400 Hz
MAIN BUS # 1
PIEZO ELECTRIC CRYSTALS
VIBRATION ACCELEROMETER
ENGINE # 1

Figure 2-59.Engine vibration monitoring system for a corporate jet airplane that employs piezoelectric crystals as sensors. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

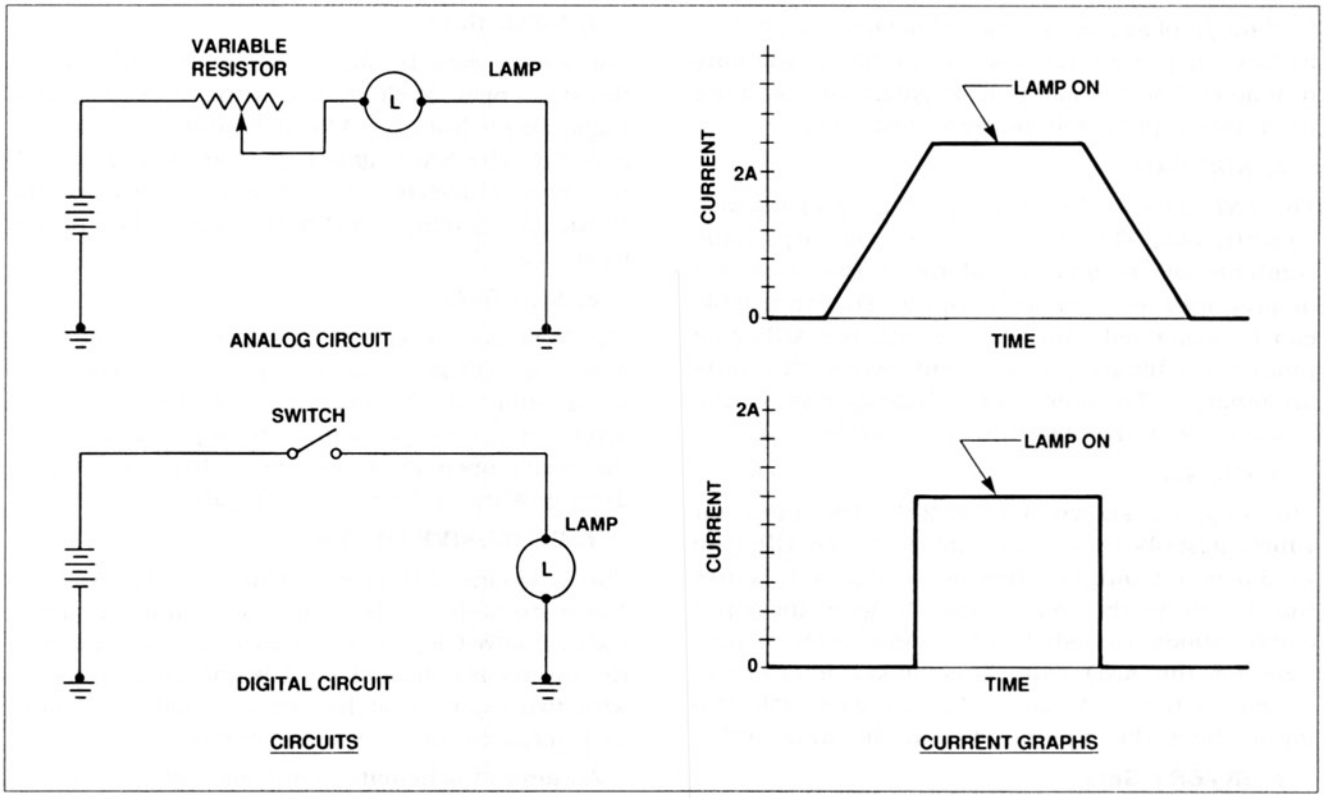

Figure 2-60. An example of the difference between analog and digital. The circuit with the switch represents digital because it has two conditions: on and off. The circuit with the variable resistor is analog because it can have an infinite number of different current flows between maximum and minimum.

(c)	64	32	16	8	4	2	1
(b)	2^6	2^5	2^4	2^3	2^2	2^1	2^0
(a)	1	0	0	1	1	0	1
(d)	64	0 +	0 +	8 +	4 +	0 +	1 = 77

Figure 2-61. An example of the method for converting a number from binary to decimal notation.

Two common types of binary codes that are used with logic circuits are Binary Coded Decimal (BCD) and Octal. An example of BCD coding would be [1001 1101 1110]. The rule for BCD is that there are four bits in each byte. An example of Octal coding is [100 111 101]. This is the same as an earlier example. There are three bits in each byte. The reason that standard coding systems such as these are used is that it is more efficient than having a variable number of bits in each byte. The Octal code gets its name from the fact that there are eight different decimal equivalent numbers that can be encoded in a three bit byte. For example, 000 is equal to decimal 0 and 111 is equal to decimal 7.

2. Logic Gates

Logic gates can be thought of as the most simple and basic building blocks for digital control systems.

division	quotient	remainder
77/2	38	1
38/2	19	0
19/2	9	1
9/2	4	1
4/2	2	0
2/2	1	0
1/2	0	1

Figure 2-62. An example of converting a number from decimal to binary notation.

A logic gate uses binary signals as its inputs and outputs. It is a semi-conductor device which is manufactured using techniques similar to those used to produce diodes and transistors. Since logic gates are found in all types of systems on modern aircraft, it is useful to know something about them so that their function in a circuit can be understood. There are six basic kinds of logic gates that can be identified by their standard schematic symbols. All of them have certain things in common: they may have different numbers of inputs, but each logic gate has just one output. Logic gates can

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

be thought of as special types of electronic switches, in fact, they are often used to perform switching functions. For simplicity, logic gates with no more than two inputs will be described here.

a. AND Gate

The AND gate is shown in figure 2-63 along with its truth table. The truth table shows all the possible combinations of inputs and the output that will be produced for each set of inputs. The truth table can be explained with a statement: the AND gate produces a binary 1 output only when all inputs are binary 1. The name of this logic gate can make it easier to remember the truth table.

b. OR Gate

The OR gate is shown in figure 2-64. The statement which describes the truth table is: the OR gate produces a 1 output when any input is 1. Notice that in the truth table for the OR gate, the input combinations are listed in the same order as they were for the AND gate. This makes it easier to remember the truth tables. All logic gates with two inputs have the inputs listed in the same order.

c. INVERT Gate

The Invert gate is shown in figure 2-65. This is a very simple logic gate, it only has one input and the output is always the opposite of the input. The Invert gate simply inverts any signal that is applied to it. Notice the small open circle on the output side of the Invert gate symbol. This will be used in combination with the basic shapes already covered to identify other kinds of logic gates.

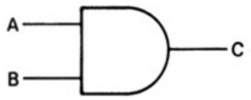

A	B	C
0	0	0
1	0	0
0	1	0
1	1	1

Figure 2-63. The AND gate and its truth table.

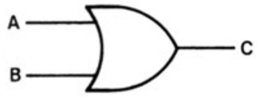

A	B	C
0	0	0
1	0	1
0	1	1
1	1	1

Figure 2-64. The OR gate and its truth table.

d. NAND Gate

The NAND gate is shown in figure 2-66. Notice the small open circle on the output side. This distinguishes it from the AND gate and also tells how it works. The NAND gate is just an AND gate with the ouputs inverted. This can be stated as: the NAND gate produces a 0 output only when all the inputs are 1.

e. NOR Gate

The NOR gate is shown in figure 2-67. The NOR gate is an OR gate with the outputs inverted. The truth table can be described as: the NOR gate produces a 0 output when any input is 1. Notice the small open circle on the output side which distinguishes it from the OR gate.

f. EXCLUSIVE OR Gate

The Exclusive OR gate is shown in figure 2-68. The truth table can be explained by the statement: the Exclusive OR gate produces a 1 output whenever the inputs are dissimilar. It is the only logic gate with two inputs that has equal numbers of ones and zeros in the outputs column.

An aircraft schematic using logic gates is shown in figure 2-75. It is evident that in order to

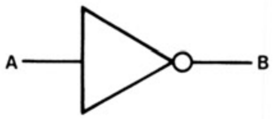

A	B
1	0
0	1

Figure 2-65. The INVERT gate and its truth table.

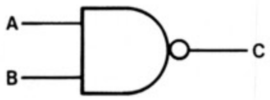

A	B	C
0	0	1
1	0	1
0	1	1
1	1	0

Figure 2-66. The NAND gate and its truth table.

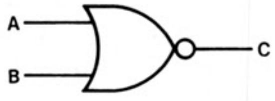

A	B	C
0	0	1
1	0	0
0	1	0
1	1	0

Figure 2-67. The NOR gate and its truth table.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

understand and troubleshoot this aircraft system, the technician should be familiar with logic gates and their truth tables.

g. Amplifier or Buffer

The buffer or amplifier is shown in figure 2-69. It produces no change in the logic state of the input signal. It is used to amplify or increase signal strength. It can also be used to isolate one part of a circuit from another part of the same circuit or from some other unit.

h. Positive and Negative Logic

Logic gates use different values of voltage to represent binary ones and zeros. The terms positive and negative logic refer to whether the voltage value for a binary 1 is more positive or more negative than the voltage value for binary 0. Figure 2-70(A) shows examples of positive logic. The voltage value for binary 1 is more positive (or less negative) than the value for binary 0. Figure 2-70(B) shows examples of negative logic which is the opposite situation.

3. Integrated Circuits

The evolution of electronic circuit technology is shown in figure 2-71(A) and figure 2-71(B). For many years, the vacuum tube represented the state of the art in delicate control of electron flow for radios and early computers. The development of the transistor in the 1960s allowed much smaller and more efficient electronic components to be made.

The transistor was made from a semi-conductor material like silicon or germanium. In the 1970s, the integrated circuit was invented. An integrated circuit is a small chip of semi-conductor material with the equivalent of many transistors on it. It was no longer necessary to manufacture transistors one at a time in individual units. Since the invention of the integrated circuit, the technology has advanced rapidly by crowding more and more transistors onto this small chip of silicon or germanium.

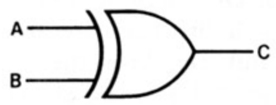

A	B	C
0	0	0
1	0	1
0	1	1
1	1	0

Figure 2-68. The EXCLUSIVE OR gate and its truth table.

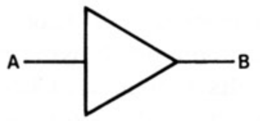

A	B
0	0
1	1

Figure 2-69. The AMPLIFIER and its truth table.

A	Binary 1	+5V	+2V	+10V	0V	–5V
	Binary 0	0V	–2V	+5V	–5V	–10V
B	Binary 1	0V	–2V	+5V	–5V	–10V
	Binary 0	+5V	+2V	+10V	0V	–5V

Figure 2-70.
- *(A) Example of positive logic.*
- *(B) Example of negative logic.*

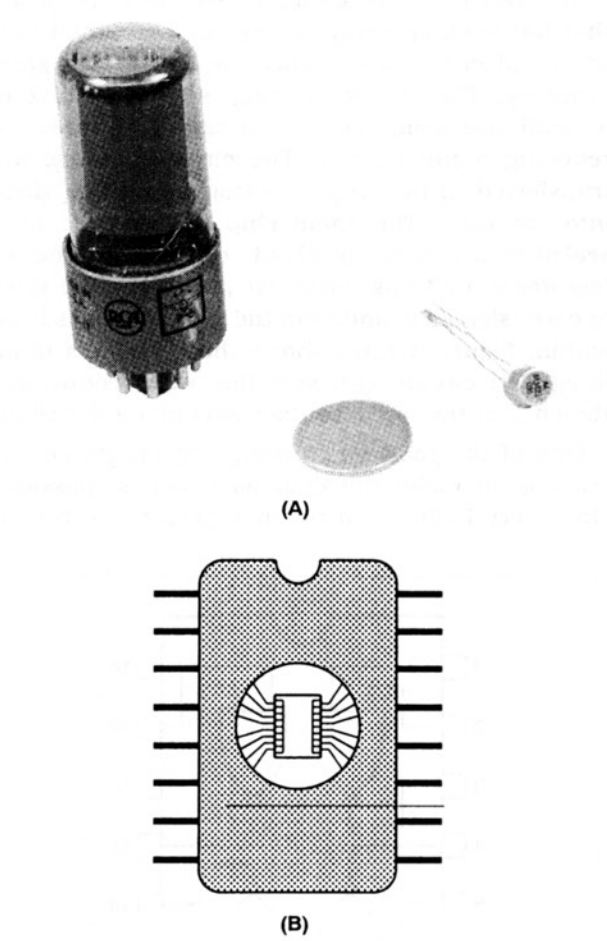

Figure 2-71
- *(A) In the 1960s, transistors began to replace vacuum tubes as a means to provide precise control of electrical signals for amplifiers and other electronic devices including computers.*
- *(B) An integrated circuit contains the equivalent of many transistors on a small chip of silicon. Very fine wires connect the chip to the outer pins.*

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The term microprocessor refers to an integrated circuit that has the equivalent of thousands of transistors on one tiny chip of silicon. The microprocessor has made possible the widespread use of computer controls for cars, radios, and in airplanes as well. The logic gates that we have been discussing are not manufactured one at a time. Logic gates are contained within integrated circuits. Integrated circuits vary considerably in complexity. Simple integrated circuits that contain 6 or 8 logic gates can be purchased for about 25 cents. The most complex integrated circuits are those called microprocessors, which would have the equivalent of thousands of logic gates on one chip. A schematic representation of the logic gates in a simple integrated circuit is shown in figure 2-72.

Integrated circuits are manufactured from silicon that has been specially processed. The circuit that will be placed onto the silicon is drawn with great accuracy. The circuit drawing is then shrunk to a small size using techniques similar to those for reducing a photograph. The circuit tracings are transferred to the chip of silicon by etching them into the chip. The small chip of silicon is then sealed in a housing of plastic or ceramic. The integrated circuit has connector pins along the sides to carry signals to and from the small chip of silicon within. Figure 2-71(B) shows these features of an integrated circuit with very fine wires connecting the chip to the outer contact pins along the sides.

One of the goals when designing integrated circuits is to make the chip as small as possible. On a very high speed microprocessor, the time it takes electrons to move the width of the chip limits the speed of computation.

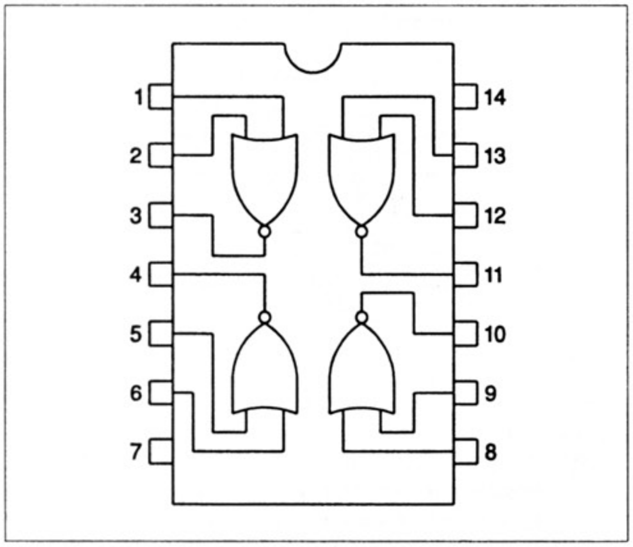

Figure 2-72. An example of a relatively simple integrated circuit (IC) showing logic gate functions and pin connections.

Integrated circuits are divided into categories based on their complexity. The four categories normally referred to are:

1. Small scale integration (SSI)
2. Medium scale integration (MSI)
3. Large scale integration (LSI)
4. Very Large scale integration (VLSI)

If a digital system has many different integrated circuits that are connected to each other, they must be compatible with each other. Integrated circuits that use the same values of voltage for binary ones and zeros and operate at similar speeds are said to be in the same family. Two of the more common families are TTL (transistor-transistor logic) and CMOS (complimentary metal oxide semi-conductor) . The TTL family, for instance, uses positive logic with binary 1 = +5V and binary 0 = 0V.

4. Dip Standards

Integrated circuits (ICs) are manufactured in standard sizes and shapes. This means that ICs made by different companies can be installed in the same way and simplifies replacement. The DIP (dual inline package) standard concerns the numbering of the connector pins and the shape and size of the integrated circuit. A typical DIP integrated circuit is shown in figure 2-72. With the notch at the top, the connector pins are numbered down the left side and then up the right side. The spacing of the pins fits the standard pattern. The total number of pins varies from about 8 to more than 40 depending on the complexity of the integrated circuit. Logic gates can be used for many different purposes. The smaller ones can be used as high speed switches. Adders and subtractors are used in a computer to perform mathematical calculations. Clocks are logic gates that supply a set frequency to synchronize the operation of different units. Latches and flip-flops are used for memory functions. A typical computer contains many integrated circuits that perform a variety of functions such as those mentioned here.

5. ARINC 429 Digital Standards

The latest air carrier jets have many different computers and digital systems on board the aircraft. These many digital systems must have a rapid and efficient way to communicate using binary coded signals. ARINC 429 is the standard for Digital Information Transfer Systems that is used on these airplanes. This standard specifies the use of a dual digital data bus where the various units connected will receive messages on both sections of the bus

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

but only transmit on one of them. The type of message transmission is serial which means different messages are sent in a series one after the other. The messages are separated by time intervals. This industry standard also specifies what coding systems will be used such as binary, binary coded decimal and discrete. The standard length of an ARINC 429 word is 32 bits as shown in figure 2-73. The use of ARINC 429 by the equipment manufacturers permits the airlines and other users to purchase electronic equipment from different sources and know that the different units will be compatible with each other.

6. Digital Aircraft Systems

The trend in the design of modern aircraft and aircraft systems is to make greater use of digital circuits and digital signals because of the advantages they have over older style analog systems. This follows another trend which is the use of more automated monitoring and control systems.

Some of the fundamental concepts involved in the design of electronic systems for current production aircraft are:

1. The use of proximity sensors to replace microswitches.
2. The use of transistors to replace potentiometers, relays etc.
3. Designing equipment to fit in standard size LRUs (line replaceable units).
4. Increased use of BITE (built-in test equipment).
5. Computer controls for navigation, engines, systems instruments, etc.
6. Electronic instruments (CRTs) to replace electro-mechanical instruments and displays.
7. LEDs and LCDs employed for lighted displays that replace older style displays and controls.

The use of computer monitoring systems like EICAS has permitted the elimination of a flight crew member on many aircraft. The flight engineer is replaced by the sophisticated computer monitoring systems and automatic control systems to create what is called the two-man cockpit in aircraft such as the Boeing 747-400 and McDonnell Douglas MD-11.

The use of ARINC 600 standard LRUs permits more efficient installation and maintenance of aircraft equipment. Many of the LRUs contain BITE systems which can be used to troubleshoot problems and identify faulty LRUs for replacement. Maintenance technicians still need to be very familiar with the aircraft systems so that they can do a better job of troubleshooting and repair, and so that the number of false removals of good LRUs can be reduced.

M. Takeoff Warning Systems

A takeoff warning system is designed to sound a warning if the flight crew tries to takeoff in an airplane with the flaps or other important systems in an incorrect position for a safe takeoff. Air carrier jets will have takeoff warning systems while most simpler aircraft will not. The critical items that are monitored by the takeoff warning system are:

1. Pitch trim or stabilizer trim
2. Speed brake
3. Leading edge flaps and slats
4. Trailing edge flaps

A schematic of a typical takeoff warning system is shown in figure 2-74. The two series switches are on the throttles and landing gear. When there is weight on the wheels and the throttles are advanced for takeoff, both of these switches will be

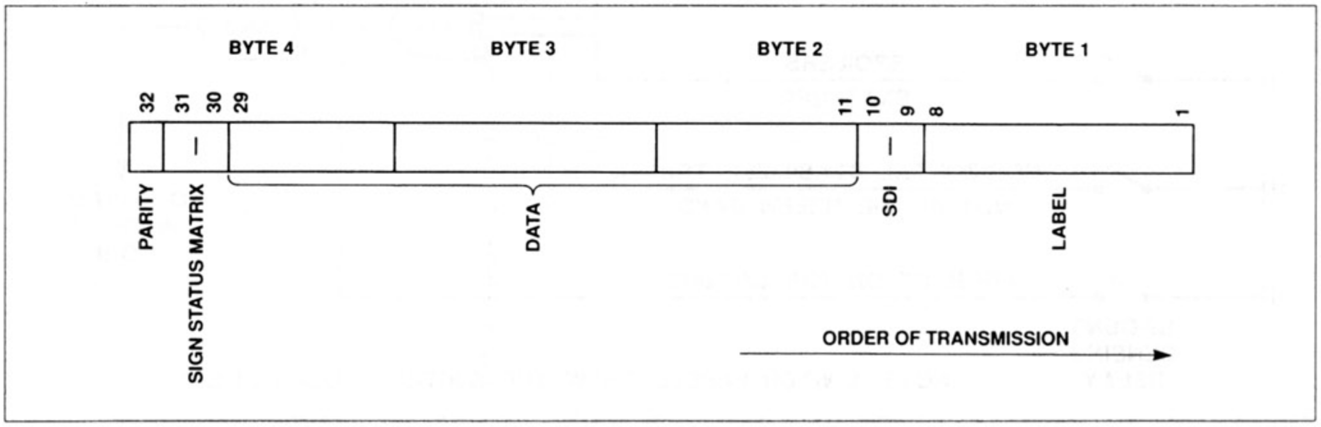

Figure 2-73. ARINC 429 is the industry standard for digital information transmission systems. An ARINC 429 standard word contains 32 bits.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

closed. If any of the parallel switches is closed, a takeoff warning will be initiated.

A more detailed schematic of a takeoff warning system is shown in figure 2-75. Notice the use of logic gate symbols on this relatively simple aircraft system.

There have been aircraft accidents caused by the failure of the takeoff warning system and an attempted takeoff with one of the critical systems set improperly. It is an important system that should be inspected regularly and maintained properly to give warnings when needed.

N. Angle of Attack Indicators

The angle of attack indicator is a system that gives an instrument indication in the cockpit of the angle of attack for the wings of the airplane. Angle of attack is the angle between the chord line of an airfoil and the relative wind. The angle of attack instrument

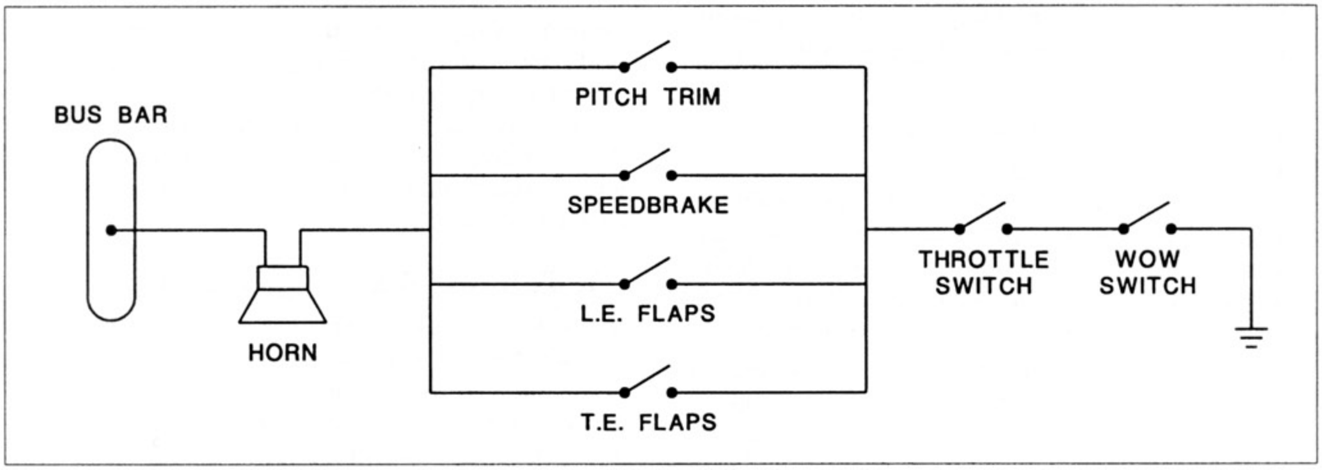
BUS BAR
HORN
PITCH TRIM
SPEEDBRAKE
L.E. FLAPS
T.E. FLAPS
THROTTLE SWITCH
WOW SWITCH

Figure 2-74. A simplified schematic of a takeoff warning system.

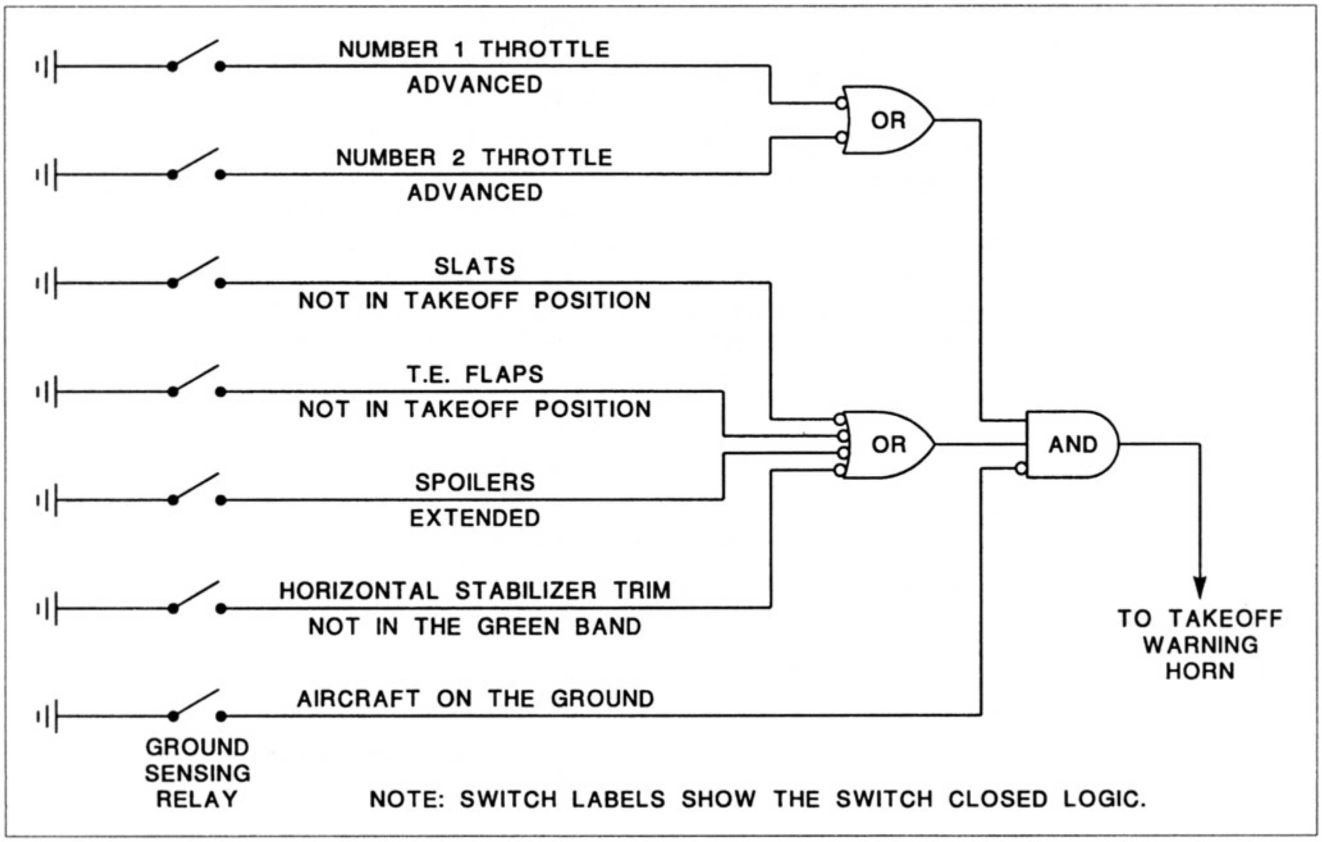
NUMBER 1 THROTTLE
ADVANCED
NUMBER 2 THROTTLE
ADVANCED
SLATS
NOT IN TAKEOFF POSITION
T.E. FLAPS
NOT IN TAKEOFF POSITION
SPOILERS
EXTENDED
HORIZONTAL STABILIZER TRIM
NOT IN THE GREEN BAND
AIRCRAFT ON THE GROUND
GROUND SENSING RELAY
OR
OR
AND
TO TAKEOFF WARNING HORN
NOTE: SWITCH LABELS SHOW THE SWITCH CLOSED LOGIC.

Figure 2-75. The takeoff warning system for an air carrier jet airplane. Logic gate symbols are often found on aircraft schematics.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

can be very useful in critical flight conditions where maximum performance must be employed. During landing and takeoff from short runways and in an encounter with wind shear, the angle of attack instrument can allow more precise control of the airplane. There are two common types of sensors used for AOA systems. The first type is a small vane on a pivot that is installed on the forward fuselage of the airplane. This type can be seen as a part of the stall warning system shown in figure 2-76. Angle of attack sensors can be used with a stall warning system, an AOA gauge, or both. The vane rotates to align itself with the relative wind and this information is transmitted to the cockpit with a synchro system where it moves a pointer on the AOA instrument. The second type uses a sensor which is a tube with two slots that projects out into the airstream, normally from the forward fuselage. This

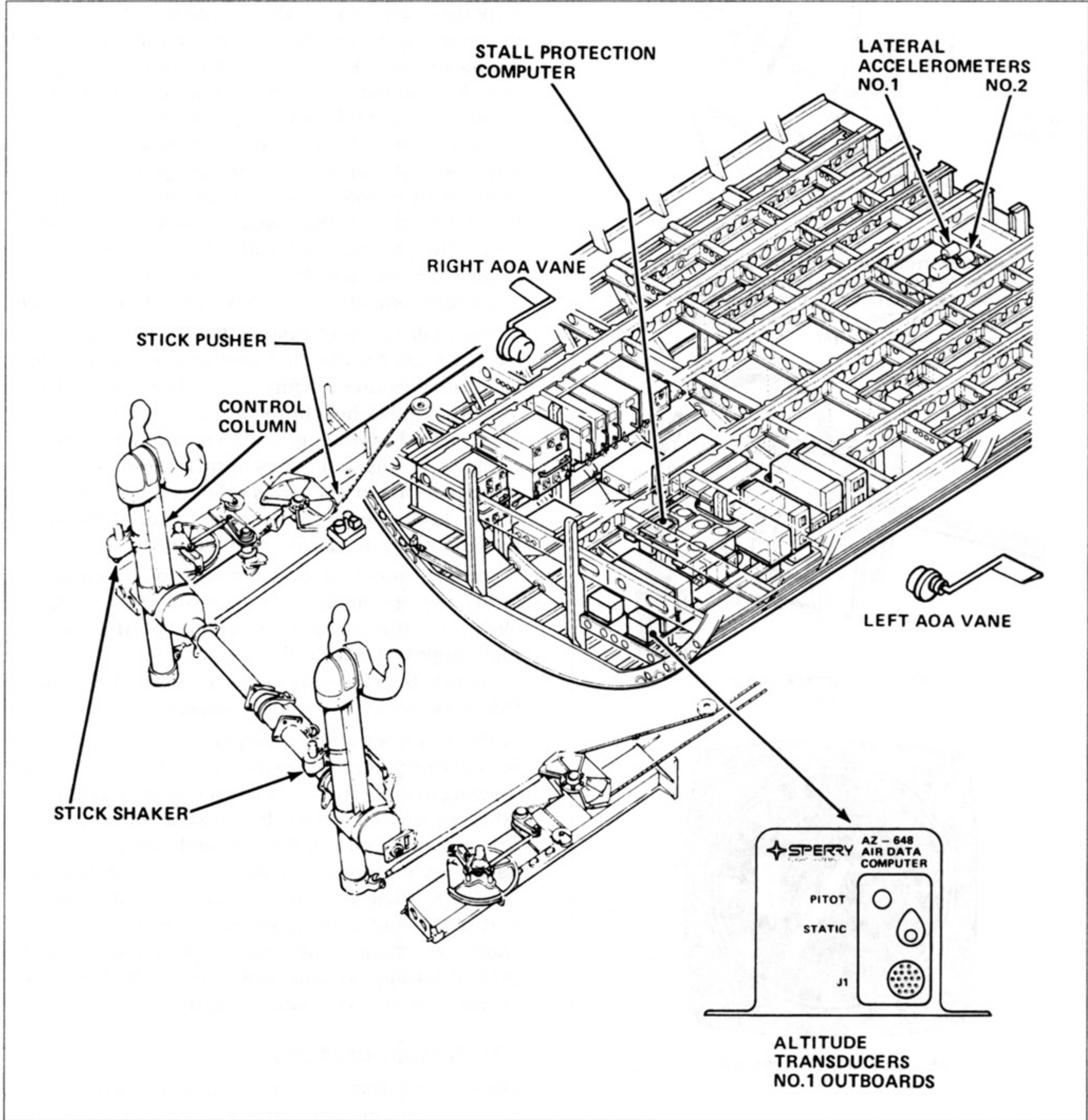
STALL PROTECTION COMPUTER
LATERAL ACCELEROMETERS
NO.1
NO.2
RIGHT AOA VANE
STICK PUSHER
CONTROL COLUMN
LEFT AOA VANE
STICK SHAKER
SPERRY
AZ – 648 AIR DATA COMPUTER
PITOT
STATIC
J1
ALTITUDE TRANSDUCERS NO.1 OUTBOARDS

Figure 2-76. The stall warning system for a corporate jet airplane that employs vane-type angle-of-attack sensors. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

type of probe, along with the AOA indicator, is shown in figure 2-77. The two slots in the probe are connected to two chambers. As the angle of attack changes, the relative pressure in the two slots changes along with the pressures in the air chambers. The difference in air pressure rotates a vane in the unit inside the fuselage which is connected by synchro to the cockpit instrument. The units displayed on the cockpit instrument face can be a percent or decimal numbers that indicate the angle of attack. The reading of 100% or 1.0 would indicate a stall angle of attack, so the pilot can readily determine if a safe margin is being maintained from the stall angle-of-attack.

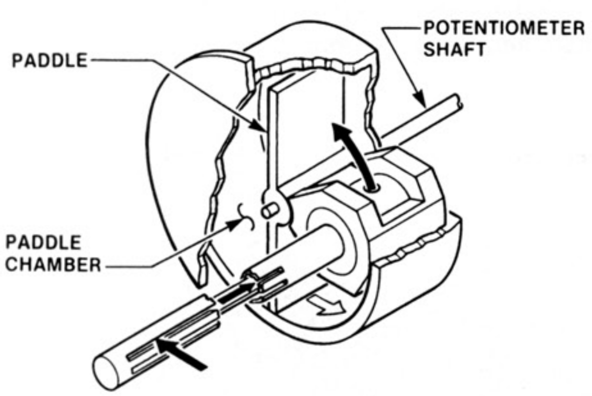

AIRFLOW THROUGH THE SLOTS IN THE PROBE MOVE THE PADDLE WHICH ROTATES THE POTENTIOMETER SHAFT TO CHANGE THE RESISTANCE AS THE ANGLE OF ATTACK CHANGES.
(A)

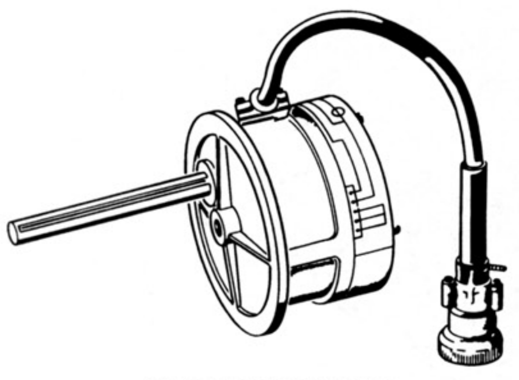

ANGLE OF ATTACK PROBE.
(B)

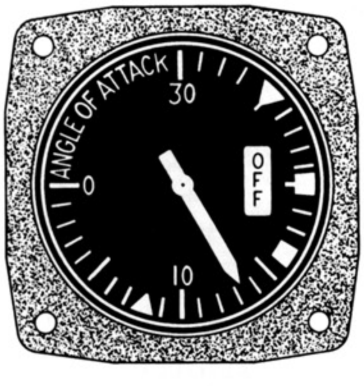

ANGLE OF ATTACK INDICATOR.
(C)

Figure 2-77 An angle-of-attack indicating system that uses the air pressure type of AOA probe.

O. Stall Warning Systems

A stall warning system is required for all modern airplanes. Several of the common types will be described, but they all rely on some measurement of angle of attack to activate the stall warning system. The airspeed of the airplane cannot be used to operate a stall warning system because the airplane can stall at many different airspeeds. An airplane will stall at a higher airspeed in a steep bank than it will in level flight because some of the wing lift is being used to make the airplane turn. The airplane will stall at the same angle of attack in both straight and turning flight, so AOA is a better indicator for stall warning than airspeed.

The stall warning system found on many older airplanes used a vibrating reed and required no outside power source. Figure 2-78 shows this type of system. A small hole in the leading edge of the wing is connected by tubing to a reed and horn near the cockpit. At high angles of attack close to a stall, a low pressure or suction is produced at the opening in the leading edge and this pulls air through the reed to make it vibrate and produce a noise.

On later small airplanes, a vane operated switch is located on the leading edge and this switch is closed by the upward movement of the vane at high angles of attack. This system uses DC power to operate the stall warning horn. Figure 2-79 shows this type of stall warning sensor.

Modern high performance airplanes use an AOA probe as previously described which is connected to a stall warning circuit. These airplanes usually have several other sensors connected to the stall warning circuit or computer. Sensors for flap and slat position are used to give an accurate stall warning for any flight condition. Most jet airplanes have a stick shaker as part of the stall warning system. This device actually shakes or vibrates the control column to warn the pilot of an approaching stall. The stick shaker can be seen at the left side of figure 2-76.

P. Annunciators

Other than simple single-engine airplanes, most aircraft have an annunciator panel which groups together a number of different indicator lights for a variety of aircraft systems and equipment. There

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

are usually three categories of annunciator indications based on how important the information is to the flight crew. Warnings are the most serious and normally activate a red light. For very serious conditions like fires, a sound warning is also given. Caution annunciators are usually amber in color. Status or special purpose annunciators may be white or some other color lights. If the airplane has a large number of annunciators and the panel must be located overhead or some other location which is not in front of the crew, master caution and warning lights are employed. With this system, there is one red light and one amber light in plain view in front of the pilot and/or copilot. When any warning light on the main panel is illuminated, the master warning light also comes on. The crew is alerted to look up at the main annunciator panel to find the source of the warning. The master warning light is then reset so it will be available to give any additional warnings that might be needed. The master caution light works in a similar manner. Some of the typical indications for an annunciator panel for a small twin-engine airplane are shown in figure 2-80

Q. FARs for Warning Systems and Annunciators

Some of the FAA requirements for warning systems and annunciators are given below to familiarize the reader with the kinds of rules that apply to these systems.

FAR 23.207

Stall Warning

A. There must be a clear and distinctive stall warning with flaps and landing gear in any normal position in straight and turning flight.

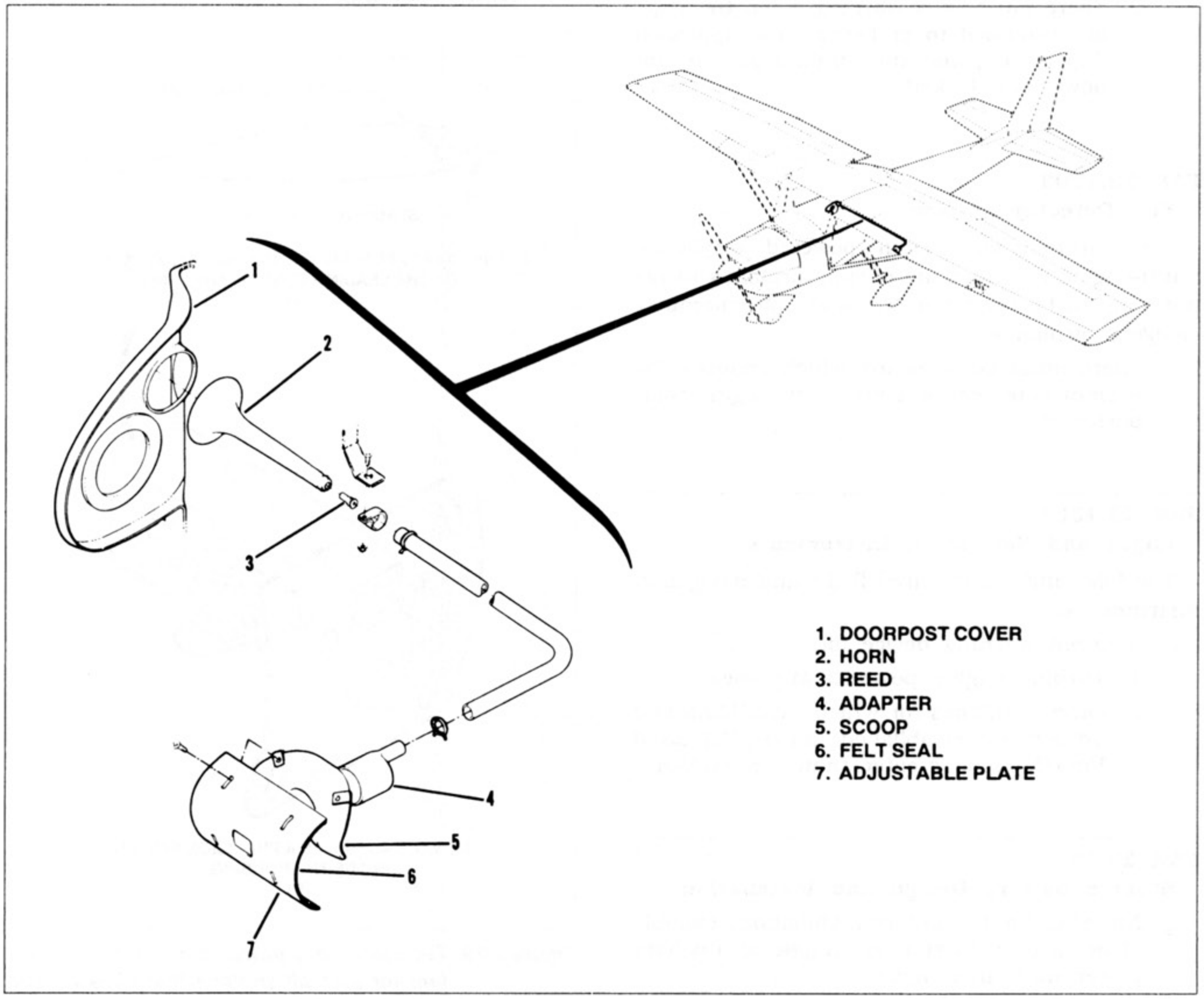

Figure 2-78. The components in a vibrating reed stall warning system. (Courtesy Cessna Aircraft Co.)

B. The stall warning may be inherent buffet or a device. If a device is used, visual indications by themselves that require the pilot's attention to be directed in the cockpit are not acceptable.

C. The stall warning must begin not less than 5 knots above stall speed; but not more than 10 knots or 15% above stall speed, whichever is greater. The stall warning must continue to the stall.

FAR 23.729

Retractable Landing Gear

A. There must be a position indicating system for extended and retracted.

B. Warning system.

  1. There must be a warning if one or more throttles are retarded and the landing gear is not down and locked.
  2. There must be a warning when the flaps are extended to or beyond the approach flap setting and the landing gear is not down and locked.

FAR 23.1203

Fire Detector System

For multi-engine, turbine-powered airplanes, multi-engine reciprocating engine airplanes incorporating turbosuperchargers and all commuter category airplanes:

A. There must be a means which ensures the prompt detection of a fire in an engine compartment.

FAR 23.1303

Flight and Navigation Instruments

The following are required flight and navigation instruments:

E. A speed warning device for:

  1. Turbine engine powered airplanes
  2. Other airplanes for which Vmo/Mmo and Vd/Md are established under FAR 23 if Vmo/Mmo is greater than 0.8 Vd/Md.

FAR 23.1353

Storage Battery Design and Installation

g. Nickel cadmium battery installations capable of being used to start an engine or auxiliary power unit must have:

  1. A system to control the charging rate of the battery automatically so as to prevent battery overheating.
  2. A battery temperature sensing and over-temperature warning system with a means for disconnecting the battery from its charging source in the event of an over-temperature condition.
  3. A battery failure sensing and warning system with a means for disconnecting the battery from its charging source in the event of a battery failure.

FAR 91.219

Altitude Alerting System or Device Turbojet-Powered Civil Airplanes:

1. No person may operate a turbojet powered civil airplane unless it is equipped with an altitude alerting system or device that:

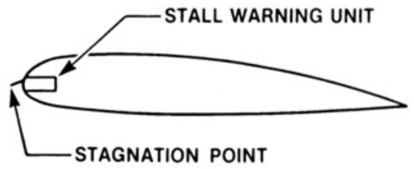

THE VANE IS PLACED AT THE STAGNATION POINT ON THE LEADING EDGE OF THE WING.
(A)

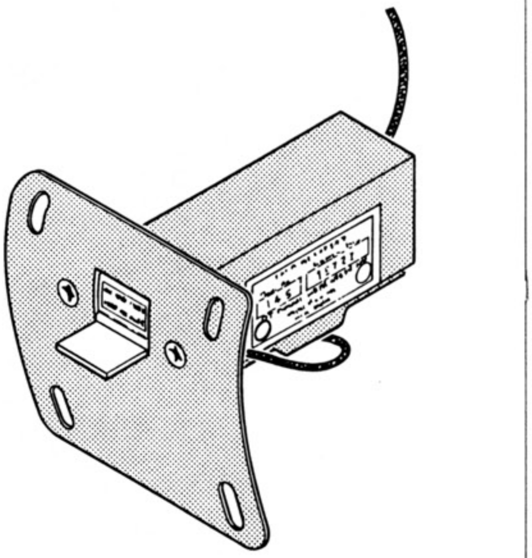

THE VANE ACTUATES A PRECISION SWITCH INSIDE THE HOUSING.
(B)

Figure 2-79. The electrically powered stall warning system for a small airplane that uses a vane operated switch.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

2. Will alert the pilot upon approaching a preselected altitude in either ascent or descent, by a sequence of both aural and visual signals in sufficient time to establish level flight at the preselected altitude.

The control and operation of a typical altitude alerting system are shown in figure 2-81.

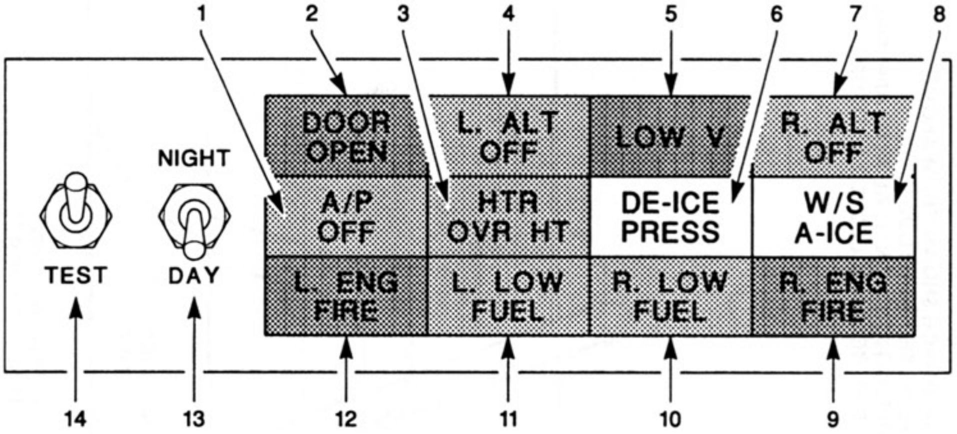

1. AUTOPILOT OFF LIGHT (AMBER) — Indicates the autopilot has disengaged.
2. DOOR OPEN LIGHT (RED) — Indicates that the forward baggage compartment door, the cabin entry doors and/or the emergency exit door are not secured safely for flight.
3. HEATER OVERHEAT LIGHT (AMBER) — Indicates an abnormally high temperature has occurred in the combustion heater and it has been automatically shut off. Once the light illuminates, the heater will not operate until the overheat switch in the right forward nose section (accessible in nosewheel well) has been reset.
4. LEFT ALTERNATOR OFF LIGHT (AMBER) — Indicates the left alternator is not supplying electrical current.
5. LOW VOLTAGE LIGHT (RED) — Indicates electrical system bus voltage is less than 24.5 volts.
6. WING AND STABILIZER DEICE SYSTEM PRESSURE LIGHT (GREEN) — Indicates pressure is being applied to the surface deice boots to inflate them.
7. RIGHT ALTERNATOR OFF LIGHT (AMBER) — Indicates the right alternator is not supplying electrical current.
8. WINDSHIELD ANTI-ICE SYSTEM LIGHT (GREEN) — Indicates that heating elements in the windshield anti-ice system are operating.
9. RIGHT ENGINE FIRE LIGHT (RED) — Indicates an excessive temperature condition or possible fire has occurred in the right engine compartment.
10. RIGHT LOW FUEL LIGHT (AMBER) — Indicates fuel quantity in right main fuel tank is 60 lbs. or less.
11. LEFT LOW FUEL LIGHT (AMBER) — Indicaties fuel quantity in left main fuel tank is 60 lbs. or less.
12. LEFT ENGINE FIRE LIGHT (RED) — Indicates an excessive temperature condition or possible fire has occurred in the left engine compartment.
13. DAY/NIGHT SWITCH — Sets brightness level of annunciator panel indicator lamps for either day or night operation.
14. TEST SWITCH — Tests operation of annunciator panel lamps, landing gear system position indicator lights, and aural warning tones of landing gear, fire detection and stall warning systems. Also, switch can be used to silence an activated engine fire detection warning tone.

Note: For illustration only. Not to be used for operational purposes.

Figure 2-80. (Courtesy Cessna Aircraft Co.)
(A) The annunciator panel for a small twin-engine airplane.
(B) The meanings for the various annunciator lights

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

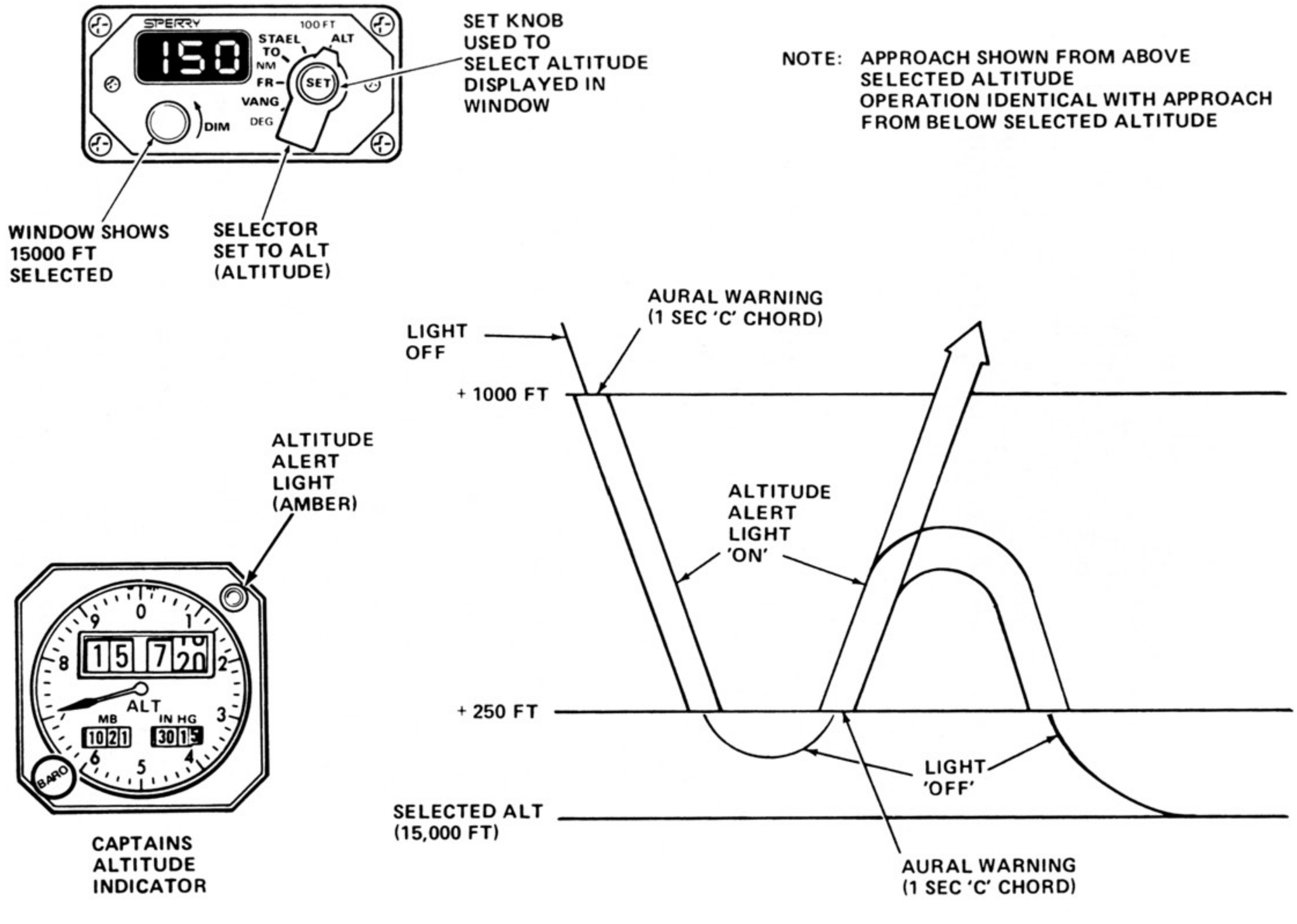
SPERRY
150
100 FT
STAEL
TO
NM
FR
VANG
DEG
ALT
SET
DIM
SET KNOB
USED TO
SELECT ALTITUDE
DISPLAYED IN
WINDOW
NOTE: APPROACH SHOWN FROM ABOVE
SELECTED ALTITUDE
OPERATION IDENTICAL WITH APPROACH
FROM BELOW SELECTED ALTITUDE
WINDOW SHOWS
15000 FT
SELECTED
SELECTOR
SET TO ALT
(ALTITUDE)
AURAL WARNING
(1 SEC 'C' CHORD)
LIGHT
OFF
+ 1000 FT
ALTITUDE
ALERT
LIGHT
(AMBER)
ALTITUDE
ALERT
LIGHT
'ON'
+ 250 FT
LIGHT
'OFF'
SELECTED ALT
(15,000 FT)
AURAL WARNING
(1 SEC 'C' CHORD)
CAPTAINS
ALTITUDE
INDICATOR
15 7 20
ALT
MB
IN HG
1021
3015
BARO

Figure 2-81.The visual and aural warnings associated with a typical altitude alerting system. (Courtesy Canadair Group, Bombardier Inc.)

CHAPTER III

Communication and Navigation Systems

The avionics and radio equipment found in aircraft has seen more change and advancement in recent years than any other part of the aircraft. Most of these changes have resulted from the same technology that has produced personal computers and a world wide telecommunications industry. Although early radios were in use prior to World War I, it is only since the 1940s that extensive use has been made of radios for communication and navigation in all types of aircraft. The basic VHF communications and navigation systems that are used in aviation were developed in the 1940s, but the introduction of newer systems has been increasing dramatically in recent years. This chapter will cover the types of avionics equipment that are already in widespread use and the latest developments that are only beginning to be installed on airplanes. Most of the systems described utilize radio waves, so we will begin with the fundamentals of radio systems.

A. Radio Fundamentals

The basic components found in radios and the operating principles involved will be explained here so that the specific aircraft avionics systems described later can be understood more easily.

1. Types of Aircraft Radio Systems

The first radios installed in airplanes were used for communications and it was only much later that navigational radio systems were developed. Radio systems for other purposes have also been developed, especially in the last 20 to 30 years. Some of the uses for radio in modern aircraft can be categorized as follows.

1. Communications — Both voice and data.
2. Navigation — Many different systems are in use today.
3. ATC Radar — The Air Traffic Control system relies on radar.
4. Weather Avoidance — Used to avoid areas of adverse weather.
5. Approach Aids — A specialized type of navigation to guide an aircraft down to the runway in bad weather.
6. Altitude Measurement — Gives precise altitude above ground level.
7. Airborne Collision Avoidance — Warns the pilot of nearby aircraft.

All of these examples use radio waves, but there are other avionics systems that do not rely on the use of radio waves.

2. Electromagnetic Waves

a. Frequency Bands

The range and diversity of electromagnetic waves, or EM waves, is very broad. The entire spectrum of EM waves includes not only radio waves but visible light, gamma rays, infrared, etc. The range of radio wave frequencies that we will be involved with is illustrated in figure 3-1. The frequencies are divided into bands which, starting at the low end, are: very low frequency (VLF), low frequency (LF), medium frequency (MF), high frequency (HF), very high frequency (VHF), ultra high frequency (UHF), super high frequency (SHF) and extremely high frequency (EHF). Common aircraft radio systems are included on the right side in figure 3-1. Notice that there is an aviation application for all of the bands except EHF. The two common radio broadcast bands are also listed for purposes of comparison.

All EM waves consist of two different invisible energy fields that travel through space. The electric field and the magnetic field are at right angles to each other and to the direction of propagation or travel. Figure 3-2 shows the two fields and the direction of propagation. Radio waves are produced when a radio frequency electrical signal is sent down a conductor to an antenna. The antenna transforms the electrical signal into EM waves which propagate outward from the antenna through space. The EM waves travel through space at the velocity of light, which is 186,284 miles per second or 300,000,000 meters per second. When the EM waves strike a receive antenna, they induce voltage and current to produce an electrical signal that matches the one applied to the transmit antenna, but of course the signal is much weaker. The basic operating frequency of a radio is called the carrier frequency, because this is the signal that carries the data or information that needs to be transmitted from one place to another.

The wavelength of EM waves is often used in describing antennas and other radio components. The

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

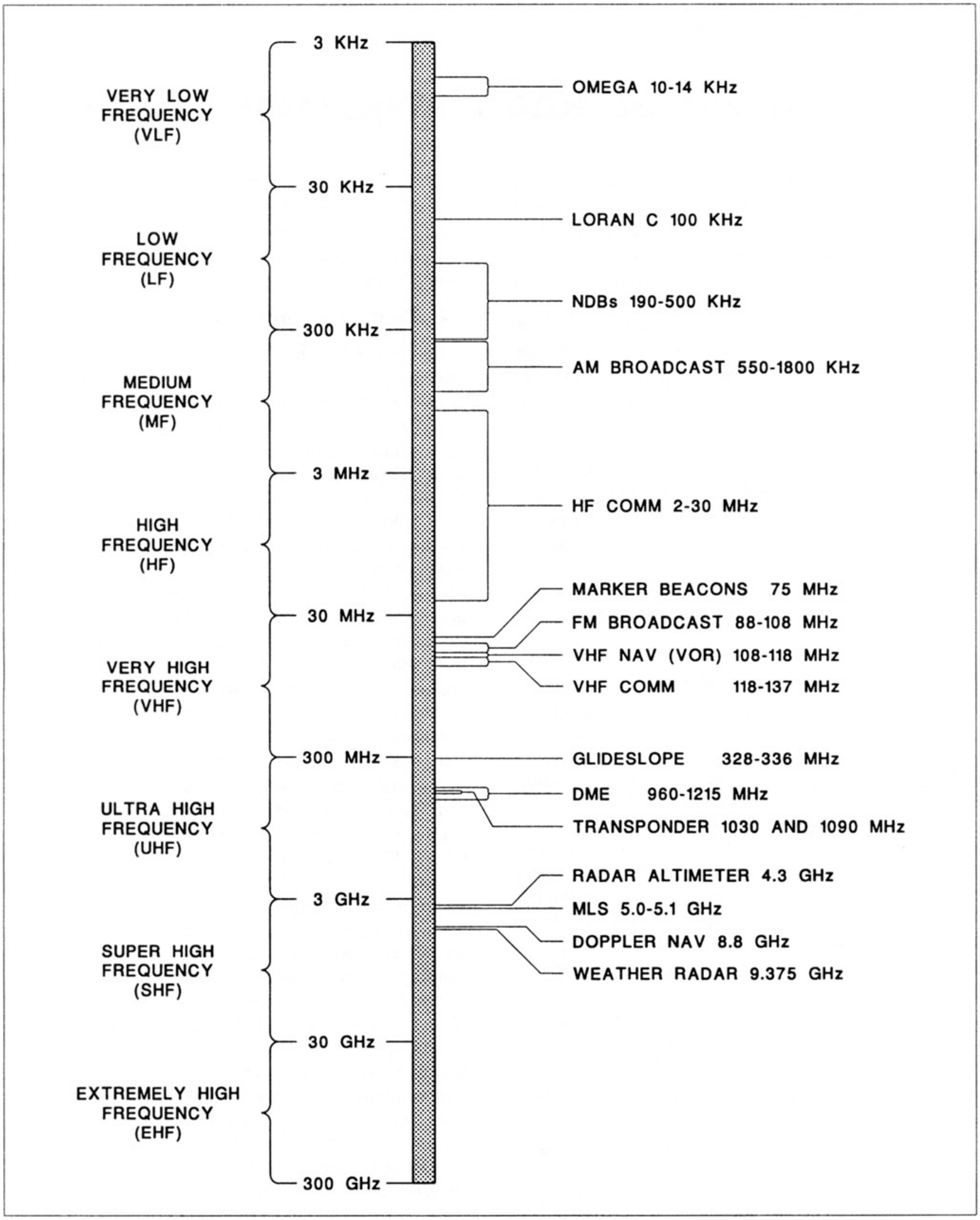
3 KHz
30 KHz
300 KHz
3 MHz
30 MHz
300 MHz
3 GHz
30 GHz
300 GHz
VERY LOW FREQUENCY (VLF)
LOW FREQUENCY (LF)
MEDIUM FREQUENCY (MF)
HIGH FREQUENCY (HF)
VERY HIGH FREQUENCY (VHF)
ULTRA HIGH FREQUENCY (UHF)
SUPER HIGH FREQUENCY (SHF)
EXTREMELY HIGH FREQUENCY (EHF)
OMEGA 10-14 KHz
LORAN C 100 KHz
NDBs 190-500 KHz
AM BROADCAST 550-1800 KHz
HF COMM 2-30 MHz
MARKER BEACONS 75 MHz
FM BROADCAST 88-108 MHz
VHF NAV (VOR) 108-118 MHz
VHF COMM 118-137 MHz
GLIDESLOPE 328-336 MHz
DME 960-1215 MHz
TRANSPONDER 1030 AND 1090 MHz
RADAR ALTIMETER 4.3 GHz
MLS 5.0-5.1 GHz
DOPPLER NAV 8.8 GHz
WEATHER RADAR 9.375 GHz

Figure 3-1. Radio frequency chart showing the operating frequencies of common aircraft systems.

wavelength of an EM wave is the distance from peak to peak for the invisible waves in the electric and magnetic fields. Wavelength is measured in meters and it is inversely proportional to the frequency. The wavelength in meters can be found by dividing the constant 300,000,000 by the frequency in hertz. At higher frequencies, it is easier to use 300 divided by the frequency in MHz to obtain the wavelength.

b. Modulation

If we design a radio transmitter that sends out a steady carrier wave, we would not be able to transmit any data or information. The receiver would only produce a steady hum or nothing at all depending on the design of the receiver. In order to transmit any data or intelligence, it is necessary to vary or alter the carrier wave in some way. The process of changing or varying the carrier wave is called modulation. When the Italian inventor Guglielmo Marconi was developing the first practical radio system in 1896, there was no way to modulate the radio wave in order to transmit voice, so he used a method of switching the transmitter on and off to transmit Morse code signals. This simplest form of modulation is called CW or radio telegraphy since it borrowed the Morse code from the telegraph industry. Figure 3-3 shows a radio carrier wave which uses CW to send Morse code dots and dashes. Within ten years, new inventions permitted voice and music to be transmitted by radio using improved types of modulation. Figure 3-4 illustrates the principles

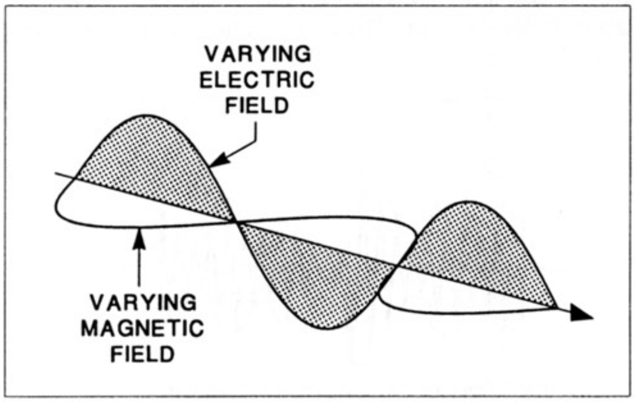

Figure 3-2. An electromagnetic wave with the electric field and magnetic field at right angles to each other and the direction of propagation.

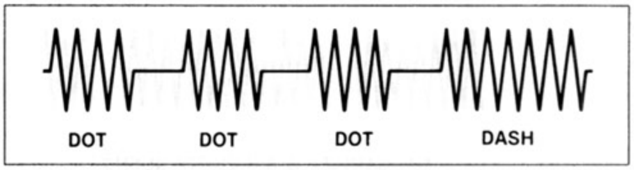

Figure 3-3. The simplest form of transmitting data with radio waves is with Morse code dots and dashes or CW.

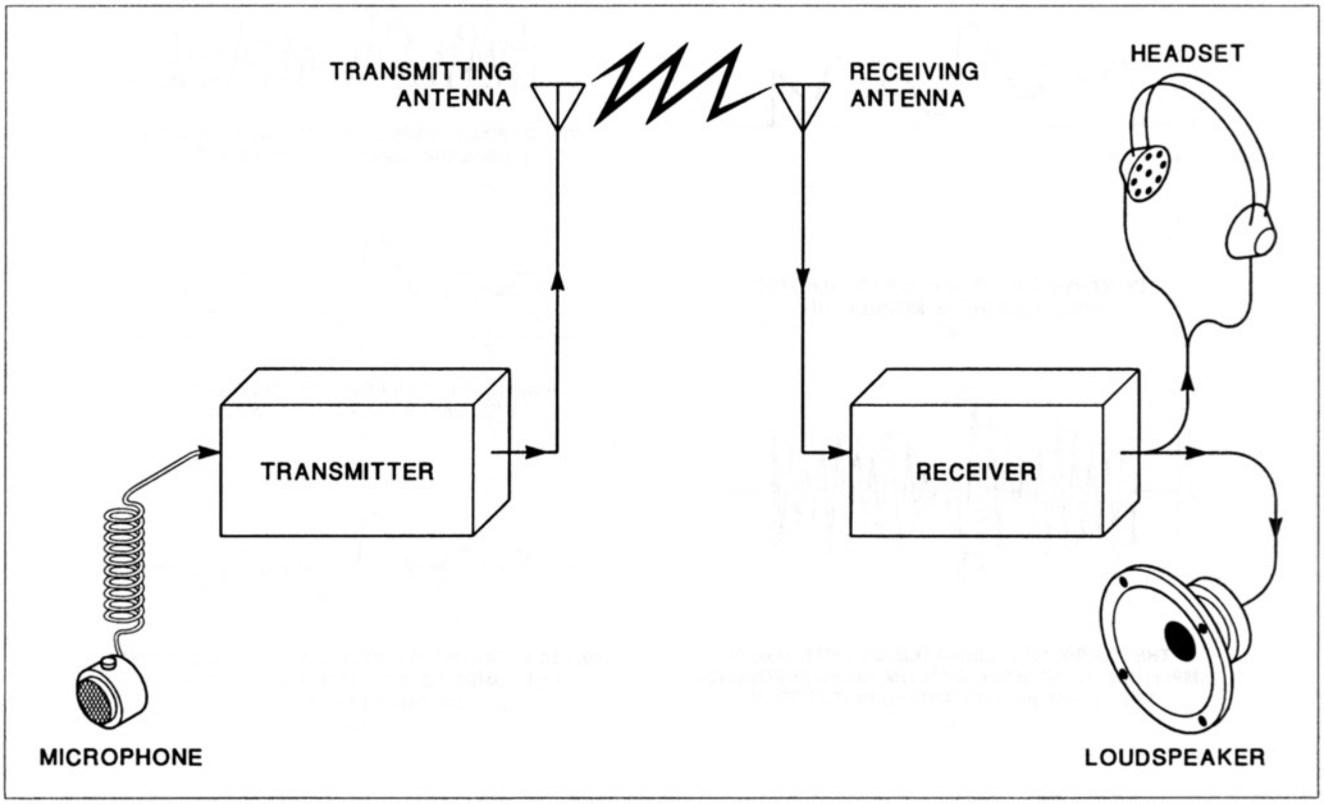

Figure 3-4. Simplified voice radio system.

of a simple voice radio system. A microphone converts sound waves to electrical signals that are combined with the carrier wave using AM or FM modulation. The receiver picks up the modulated carrier wave and then separates out the audio frequency using a demodulator. The audio output of the demodulator drives a speaker or headset to reproduce the sound waves that were picked up by the microphone in the transmitter. Current radio systems use a number of different types of modulation, the two most common being amplitude modulation (AM) and frequency modulation (FM). When amplitude modulation is used, the intensity or amplitude of the carrier wave signal is varied up and down as shown in figure 3-5. The frequency of the carrier wave is changed when frequency modulation is used. Figure 3-6 shows a carrier wave with this type of modulation. The two commercial broadcast radio bands are a good example here since one uses AM and the other FM. One of the advantages of FM is that it is less affected by atmospheric noise from thunderstorms and other disturbances.

c. Audio and Radio Frequencies

Two terms that are often used in discussions about radio systems are audio frequency (AF) and radio frequency (RF). Audio frequencies are those of 20,000 Hz or less. They are called audio because

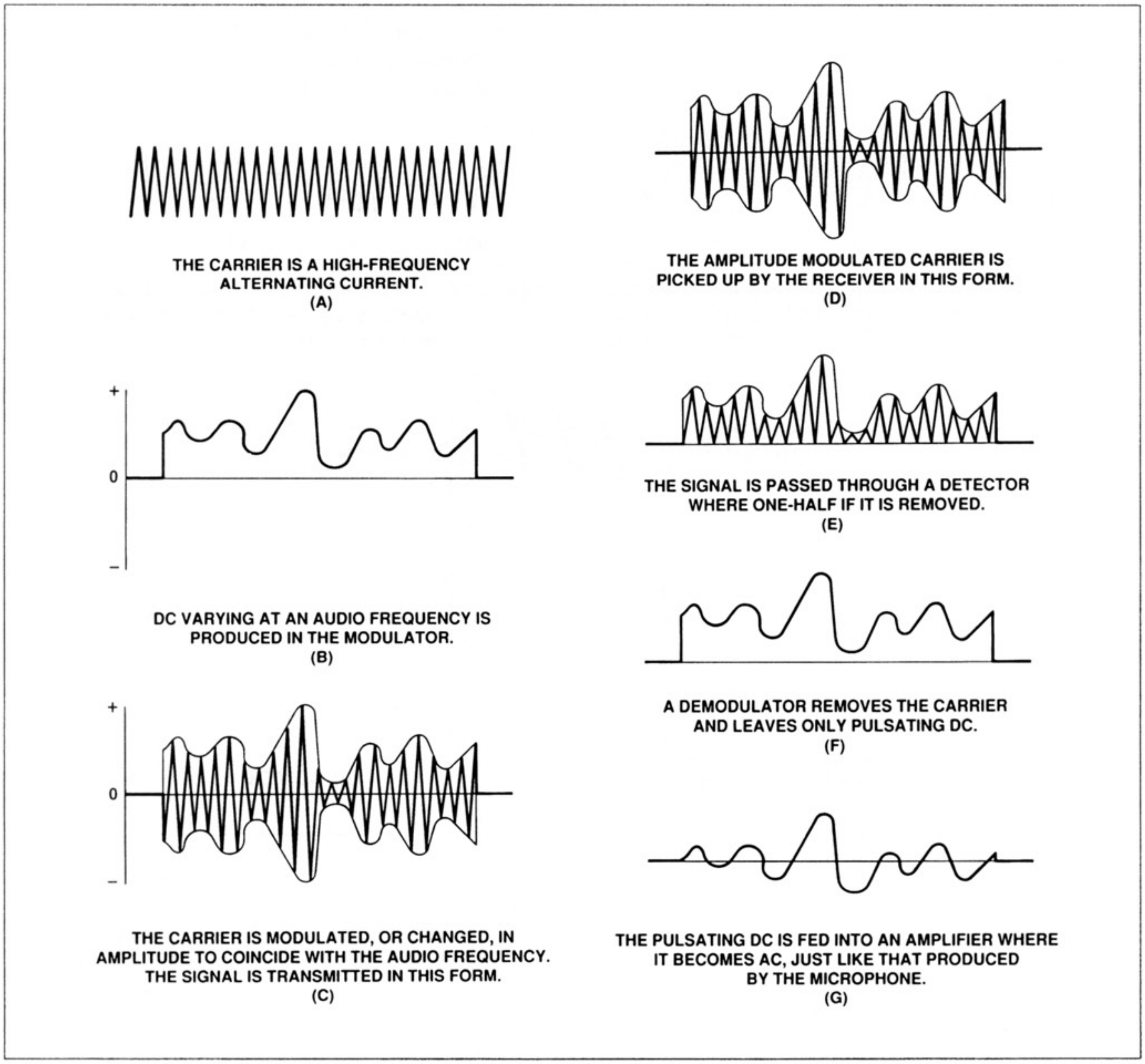

THE CARRIER IS A HIGH-FREQUENCY ALTERNATING CURRENT.
(A)

DC VARYING AT AN AUDIO FREQUENCY IS PRODUCED IN THE MODULATOR.
(B)

THE CARRIER IS MODULATED, OR CHANGED, IN AMPLITUDE TO COINCIDE WITH THE AUDIO FREQUENCY. THE SIGNAL IS TRANSMITTED IN THIS FORM.
(C)

THE AMPLITUDE MODULATED CARRIER IS PICKED UP BY THE RECEIVER IN THIS FORM.
(D)

THE SIGNAL IS PASSED THROUGH A DETECTOR WHERE ONE-HALF IF IT IS REMOVED.
(E)

A DEMODULATOR REMOVES THE CARRIER AND LEAVES ONLY PULSATING DC.
(F)

THE PULSATING DC IS FED INTO AN AMPLIFIER WHERE IT BECOMES AC, JUST LIKE THAT PRODUCED BY THE MICROPHONE.
(G)

Figure 3-5. An example of the use of amplitude modulation (AM).

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

these are approximately the frequencies of sound waves that can be heard by the human ear. Radio frequencies are those above 20,000 Hz. These terms are most often used when discussing the electrical signals that are produced and used by radio receivers and transmitters.

d. Ground, Sky and Space Waves

The behavior of radio waves as they travel through the earth's atmosphere and beyond are classified by the terms ground, space and sky waves. The propagation characteristics of ground waves, sky waves and space waves are illustrated in figure 3-7. Radio waves at frequencies below the HF band (below 3 MHz) are called ground waves because they will follow the curvature of the earth and bend. Radio waves that operate in the HF band from 3 MHz to 30 MHz are called sky waves. They tend to travel in straight lines and will not follow the curvature of the earth. Sky waves will bounce or refract off the ionosphere. The ionosphere is made up of layers of ionized particles from about 60–200 miles high. When sky waves strike a layer of the ionosphere in the right way, they will be refracted so that they will come back to earth hundreds of miles away. This characteristic of sky waves can be used to achieve long range transmission of radio signals. Space waves are the result of transmissions of radio waves at frequencies above 30 MHz or above the HF band. Space waves travel in straight lines, but they will not bounce off the ionosphere. The radio signals used to communicate with orbiting satellites are above 30 MHz. A good example can be given here for those who are familiar with AM and FM radio receivers. The commercial broadcast stations known as AM broadcast in the MF band so that the EM waves behave as ground waves. At night, it is not unusual to pick up an AM station that is hundreds of miles away because the signal has followed the curvature of the earth. Commercial FM stations operate in the VHF band and the reception range is limited to line-of-sight or straight lines because they are space waves. The maximum reception range for these stations is usually about 60 miles because they are blocked by the earth's surface at greater distances.

3. Basic Radio Components

a. Amplifiers

An amplifier is a device that increases the strength of a signal. They are found in both transmitters and receivers. A transmitter must increase the strength of the signal sent to the antenna so that the EM waves will travel a useful distance outward from the antenna. A receiver needs amplifiers because the strength of the signal from the antenna is very low and must be increased to enable the signal to be heard. Up until the 1960s, most amplifiers relied on vacuum tubes to increase the strength of signals. The transistor has replaced the vacuum tube for most amplifier applications. Amplifiers can be categorized as Class A, Class B and Class C. The difference between these is the shape of the output waveform. Figure 3-8 shows the output of a Class A amplifier, it is complete sine waves just like the input. The Class B amplifier has an output which shows only half of each sine wave as shown in figure 3-8. The Class C amplifier

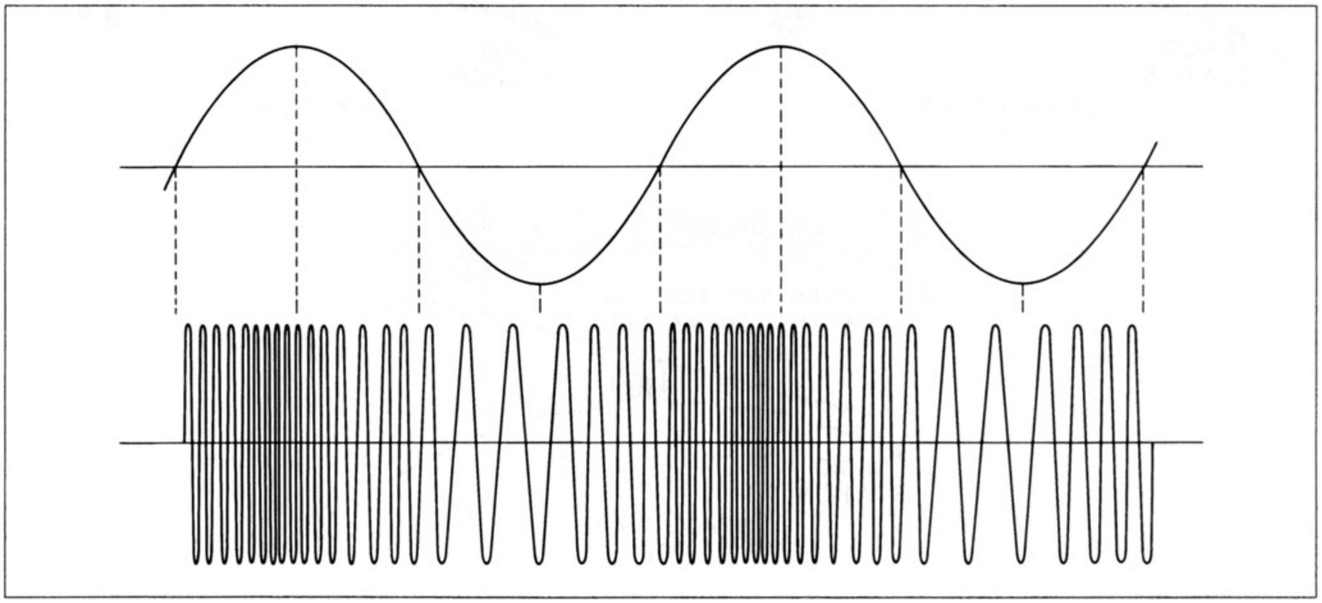

Figure 3-6. An example of the use of frequency modulation (FM).

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

has an output waveform which is less than half of the sine wave as shown at the bottom in figure 3-8. The Class C amplifier is often used as a power amplifier because of its higher efficiency. The output of the Class C amplifier can be sent through an LC circuit or other device to restore the complete sine wave shape.

b. Oscillators

An oscillator is a device that produces the frequencies needed by both receivers and transmitters. A simple oscillator is an LC circuit or tank circuit made up of a capacitor and inductor in parallel. The LC circuit will have a resonant frequency which matches the desired frequency. An LC circuit by itself will not continue to oscillate because of resistance in the components and wires. Figure 3-9 shows an LC tank circuit which can be connected to a battery to produce oscillations. If the switch in figure 3-9 is moved to position A, the capacitor will be charged by the battery. If the switch is then moved to position C, the tank circuit will start to oscillate as energy is transferred rapidly back and forth between the capacitor and inductor. The oscillations will become weaker and die out because of the resistance in the circuit. In order to maintain oscillations, some energy must be fed back into the tank circuit. In earlier radios a vacuum tube was used to supply the needed feedback. A transistor is used instead of a vacuum tube in newer radio designs as shown in figure 3-10. The resonant frequency or oscillation frequency is determined by the values of capacitance and inductance in the tank circuit. The LC circuit will not be stable over a period of time and may drift off the correct frequency. A common technique to stabilize the oscillator and produce a more accurate frequency is to use a crystal as shown in figure 3-11. The piezoelectric effect of the crystal will produce a more accurate and consistent output frequency from the oscillator.

c. Modulators and Demodulators

We will use a voice communications radio as an example of the purpose and function of modulators and demodulators. In the radio transmitter, a device is needed which will combine the AF signal with the RF carrier wave signal before it is sent to the antenna. This is the function of a modulator, it combines the AF and RF signals so that information can be transmitted. The output of the modulator is called modulated RF. The signal produced by the antenna in a receiver will be modulated RF. In order to hear the voice as an output of the

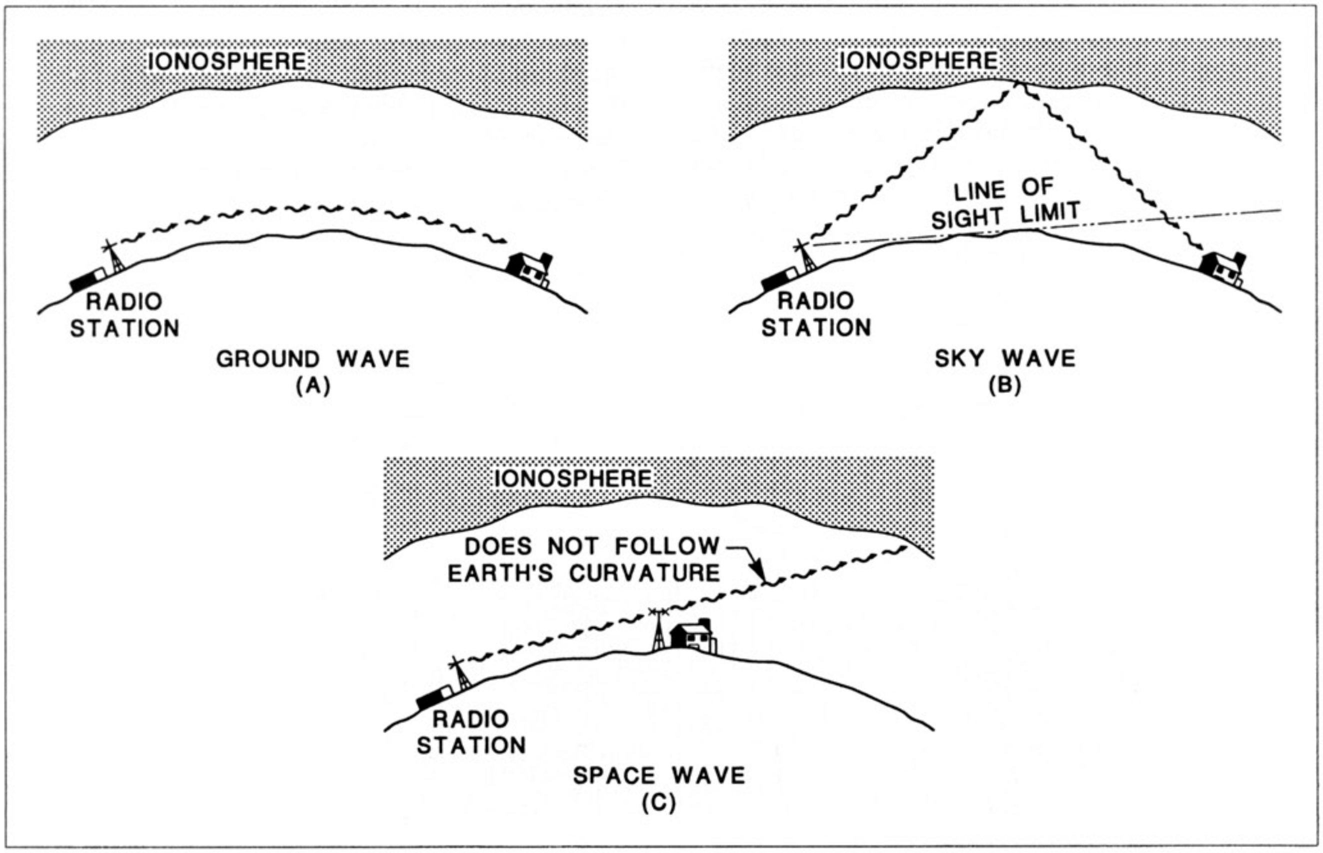

Figure 3-7. The propagation characteristics of ground waves, sky waves and space waves.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

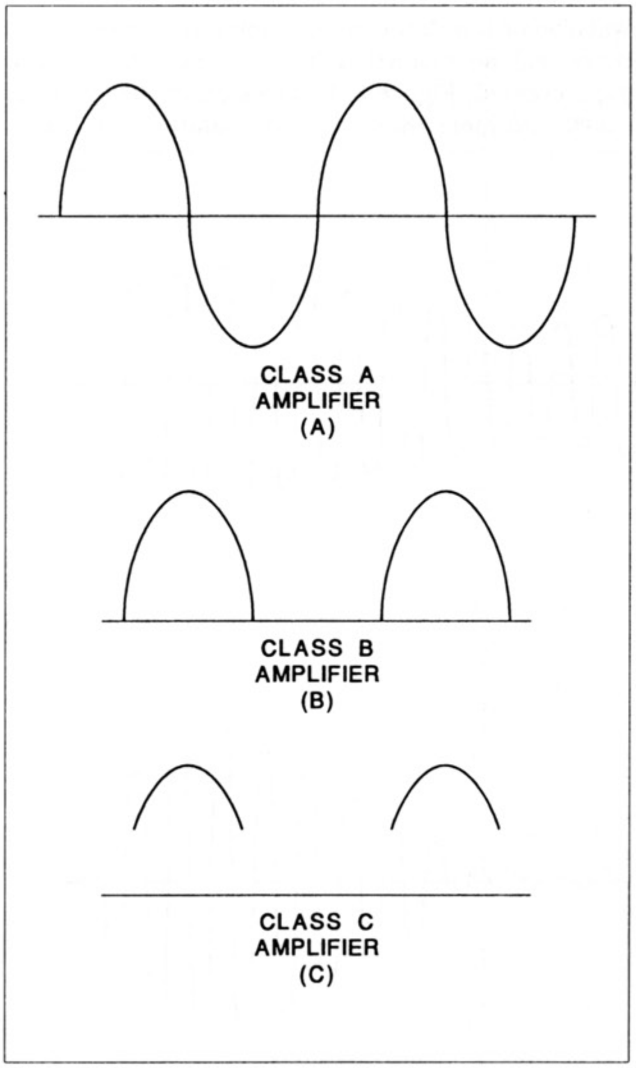

Figure 3-8. An example of the output waveform for Class A, B and C amplifiers.

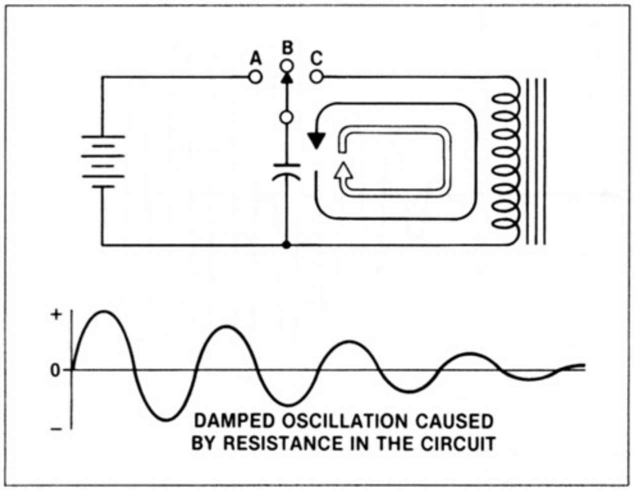
A
B
C
+
0
–
DAMPED OSCILLATION CAUSED
BY RESISTANCE IN THE CIRCUIT

Figure 3-9. A parallel LC tank circuit. Without feedback, the oscillations will be weakened by resistance in the circuit and die out.

receiver, the AF component must be separated out. The demodulator removes the RF component of the modulated RF signal and produces an AF output.

When the AF and RF signals are combined in the modulator, they must have the proper relative

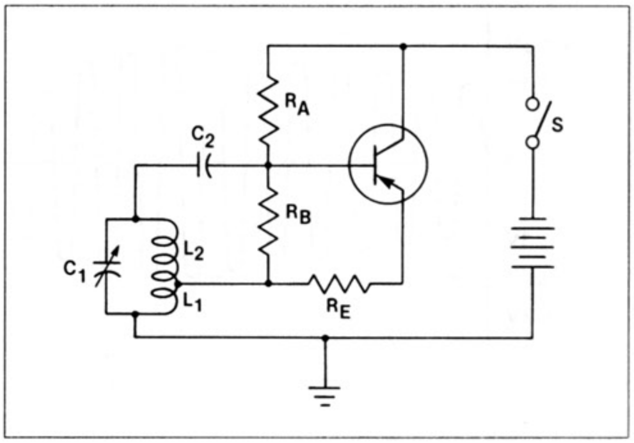
R_A
C_2
S
R_B
L_2
C_1
L_1
R_E

Figure 3-10. An oscillator circuit with feedback supplied by a transistor.

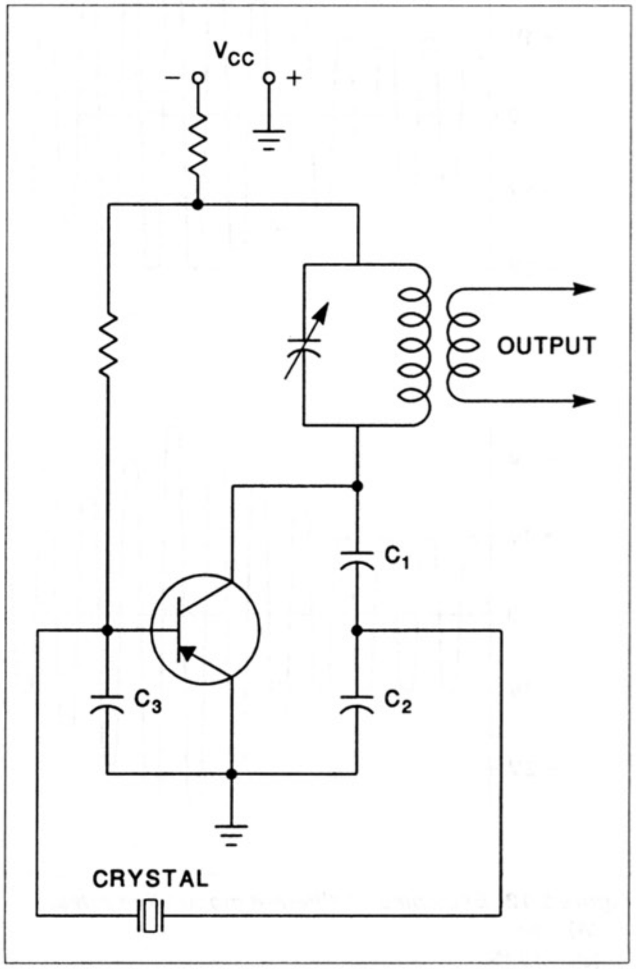
V_CC
–
+
OUTPUT
C_1
C_3
C_2
CRYSTAL

Figure 3-11. A crystal controlled oscillator.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

strengths for maximum efficiency. The amount of modulation is called the modulation rate. If the AF signal is too weak compared to the RF signal, the modulation rate will be low and the efficiency will also be low. If the modulation rate is over 100%, there will be distortion in the signal due to the gaps created. Figure 3-12 gives examples of 50%, 100% and more than 100% modulation rates with

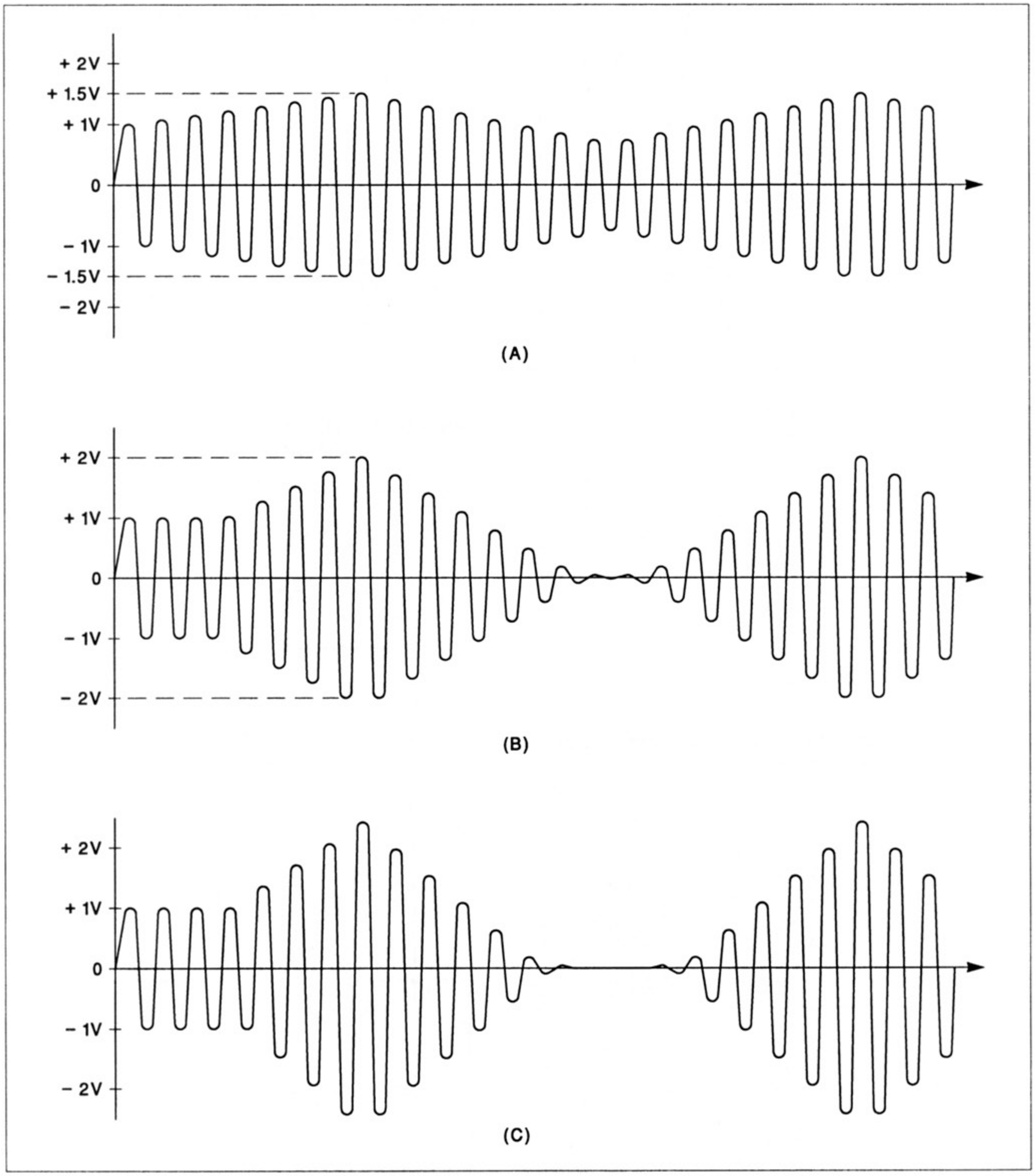

Figure 3-12. Examples of different modulation rates.
(A) 50%
(B) 100%
(C) Over 100%

AM modulation. Most radio transmitters are adjusted to about 90–95% modulation to provide a little margin to prevent distortion. Shouting into a microphone when using a radio can cause overmodulation and should be avoided.

d. Filters

A filter is used in a radio circuit to remove or filter out unwanted frequencies. The signals that are processed by the circuits in a radio often have additional frequencies present that are not needed. If the proper filter is installed, it will filter out the frequency or frequencies that are not wanted. A filter is usually made up of an arrangement of inductors and capacitors as shown in figure 3-13. A low pass filter will remove all frequencies above a certain value and pass the low ones. A high pass filter does the opposite. If a range of frequencies must be blocked, a band reject filter will be used. A bandpass filter will allow a certain band of frequencies to go through and block frequencies either above or below that range.

e. Antennas

An antenna is a device that transforms electrical signals into EM waves in the case of a transmit antenna, or transforms EM waves into electrical signals in the case of a receive antenna. Depending on the particular radio system involved, an antenna may be used for transmit only, receive only, or both. The maintenance, inspection and installation of antennas is usually the responsibility of the airframe technician since they are attached to the structure or skin of the aircraft. Antennas often have general names that describe some of their basic characteristics. Two of the more common types are the Hertz dipole antenna and the Marconi monopole antenna. The Hertz dipole antenna has two metal conductors in a straight line with the connection in the middle. It is called a half-wave antenna because the overall length is equal to one half the wavelength of the EM wave it is designed to be used with. Figure 3-14 is an example of a Hertz dipole antenna. The Marconi antenna is a single metal conductor with a length of ¼ wavelength as illustrated in figure 3-15. In order to work properly, the Marconi antenna must have metal surrounding the mounting base. The metal at the base is needed for efficient operation of the antenna. The necessary metal at the base is called the groundplane or counterpoise. In figure 3-15, the groundplane is the four metal rods at the base of the antenna; the metal skin of an aircraft is used as a groundplane for most aircraft antennas. Most antennas must be installed with the correct

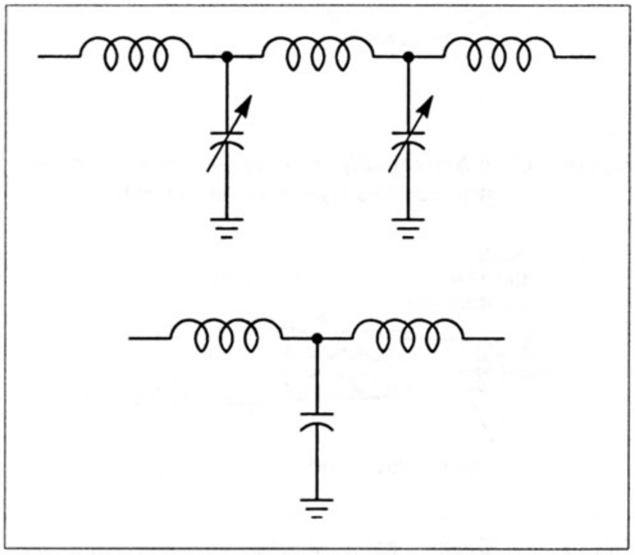

Figure 3-13. Radio frequency filters are combinations of inductors and capacitors.

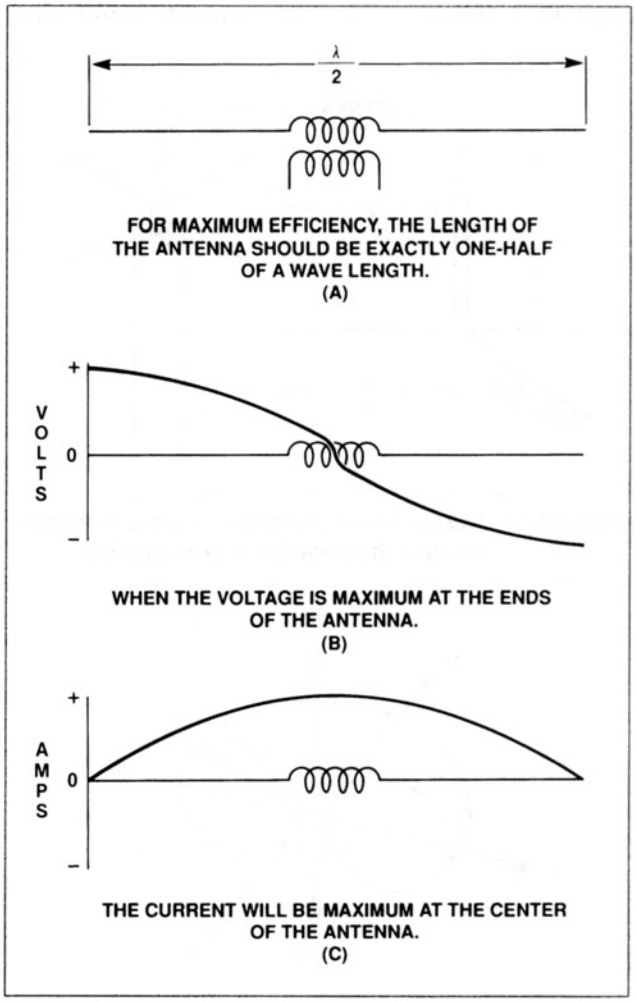
$\frac{\lambda}{2}$
FOR MAXIMUM EFFICIENCY, THE LENGTH OF THE ANTENNA SHOULD BE EXACTLY ONE-HALF OF A WAVE LENGTH.
(A)
+
VOLTS
0
-
WHEN THE VOLTAGE IS MAXIMUM AT THE ENDS OF THE ANTENNA.
(B)
+
AMPS
0
-
THE CURRENT WILL BE MAXIMUM AT THE CENTER OF THE ANTENNA.
(C)

Figure 3-14. The Hertz dipole antenna is a half-wave antenna.

polarization. Polarization refers to the orientation of the electric field relative to the earth. If the electric field is vertical, it has vertical polarization.

The Marconi antenna will produce a vertically polarized radiation pattern as shown in figure 3-16. Horizontal polarization means that the electric field will be parallel to the earth's surface. The horizontally polarized pattern of a Hertz dipole antenna is shown in figure 3-17. Another way to understand polarization is that the polarization of the aircraft antenna should normally match the polarization of the ground based antenna and the EM waves it will utilize. An aircraft VHF communications antenna is an example of a Marconi antenna with vertical polarization. The common example of a Hertz antenna is the VHF navigation antenna found on small airplanes. It is a V-shaped dipole antenna with horizontal polarization.

Antennas must not only be matched to the proper radio, but the conductor that connects the radio and antenna is very critical. A special type of conductor used to connect radios and antennas is called a coaxial cable or coax. Figure 3-18 shows the basic parts of a coaxial cable designed to carry RF signals. It consists of a center conductor covered by a special kind of insulation and an outer conductor around the insulation. Plain wires cannot be used for radio frequency signals because the energy loss would be too great at these frequencies. The antennas and coaxial cables must be maintained in good condition to ensure proper performance from the radio system.

The connection between an antenna and a radio normally requires a coupler in order to give the best transfer of energy between the two of them. Two common types of antenna couplers are the LC circuit and the transformer types. A transformer type of antenna coupler is shown in the antenna connection in figure 3-19. The use of an LC circuit as a coupler between the coax and antenna is shown in figure 3-20. Most aircraft antennas are speed-rated. For example, an antenna rated at 250 mph should not be installed on an aircraft with a higher Vne speed.

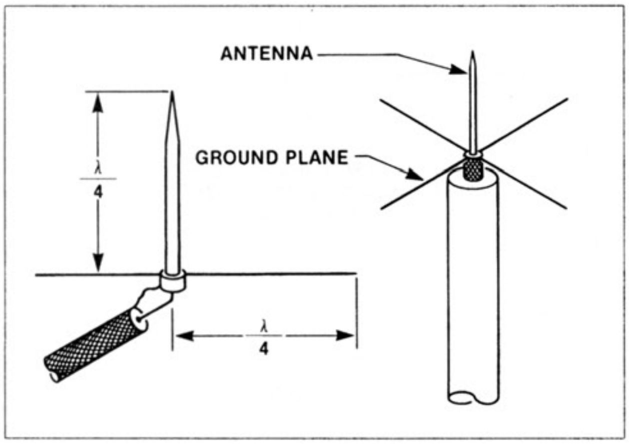

Figure 3-15. The Marconi antenna is a ¼-wave monopole antenna that requires a groundplane.

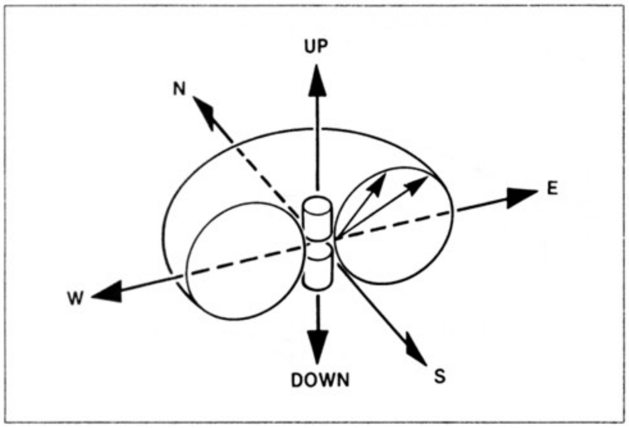

Figure 3-16. A vertically polarized marconi antenna will produce this type of radiation pattern.

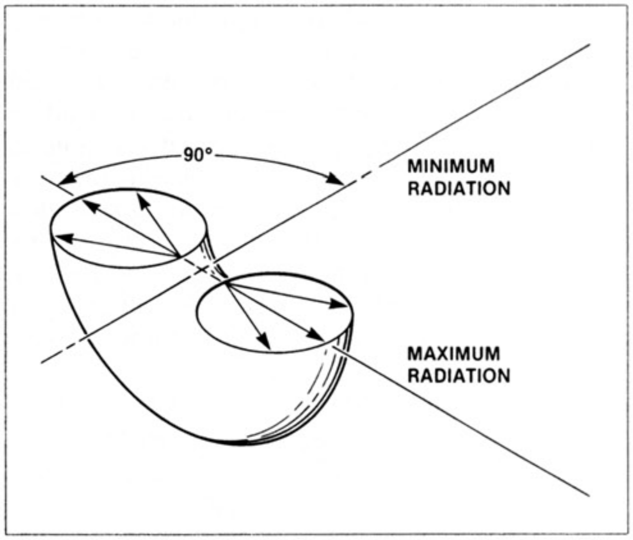

Figure 3-17. A horizontally polarized Hertz antenna will produce this type of radiation pattern.

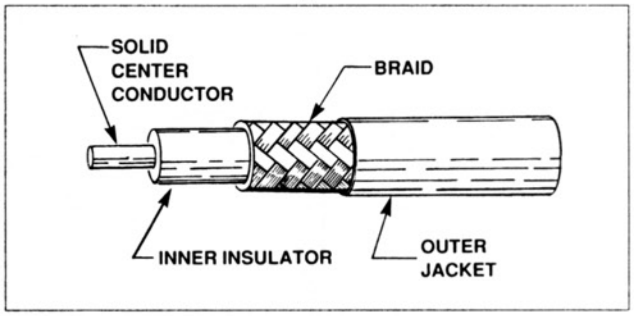

Figure 3-18. Coaxial cable is used to carry radio frequency electrical signals between radios and antennas.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

f. Tuning Circuits

An antenna will intercept many different EM waves of different frequencies so some method must be used to separate out the desired frequency. The tuning circuit performs this function. A simple tuning circuit is shown in figure 3-21, it consists of a variable capacitor and an inductor in parallel. As the tuning knob is rotated on the radio, it moves the variable capacitor until the resonant frequency of the circuit matches the frequency of the desired station. This signal is passed into the radio and the other frequencies are blocked out. A better type of tuner which is found on most modern radios uses a frequency synthesizer which contains a number of crystals that can be combined to match the desired frequency. The basic operation of a frequency synthesizer is shown in figure 3-22. Each crystal has a particular frequency and by using switches the crystals can be combined to produce many additional frequencies. When two frequencies are combined, two new frequencies are created that are equal to the sum and the difference of the two frequencies. By using this technique, hundreds of frequencies can be created using a relatively small number of fixed frequency crystals.

g. Transmitters

The components that have been described will be combined in a block diagram to see how they work together. A voice radio transmitter is shown in figure 3-23. The microphone changes the sound waves of a human voice into AF signals that are amplified and then sent to the modulator. The oscillator in this radio operates at one half the carrier frequency, so its output is amplified and then doubled. The modulator combines the AF and RF signals which are then amplified in the power amp before being sent down the coax to the antenna.

h. Receivers

In the 1920s, a new type of radio receiver was invented that produced better sound quality. It was called the superheterodyne or superhet radio. The only major difference between the superhet and earlier radios was that it reduced the modulated RF signal from the antenna to an AF signal in more than one jump or stage. Since modern radio components are much smaller and more efficient than in the 1920s, virtually all modern radio receivers are superhet. Figure 3-24 shows a block diagram of a VHF superhet aircraft receiver. The RF signal from the antenna is combined with a local oscillator frequency to produce a lower IF frequency. The intermediate frequencies found in a superhet radio are abbreviated as IF. The basic principle of the mixer is that when two different frequencies are combined, two new frequencies are created; the sum and the difference of the two combined frequencies. In this example, the output of the mixer is the difference between the RF frequency and the local oscillator frequency. The IF signal is amplified and then sent to the detector and demodulator. The detector chops off half of each sine wave to produce a varying DC signal from an AC signal. The AF signal is amplified and used to drive the speaker.

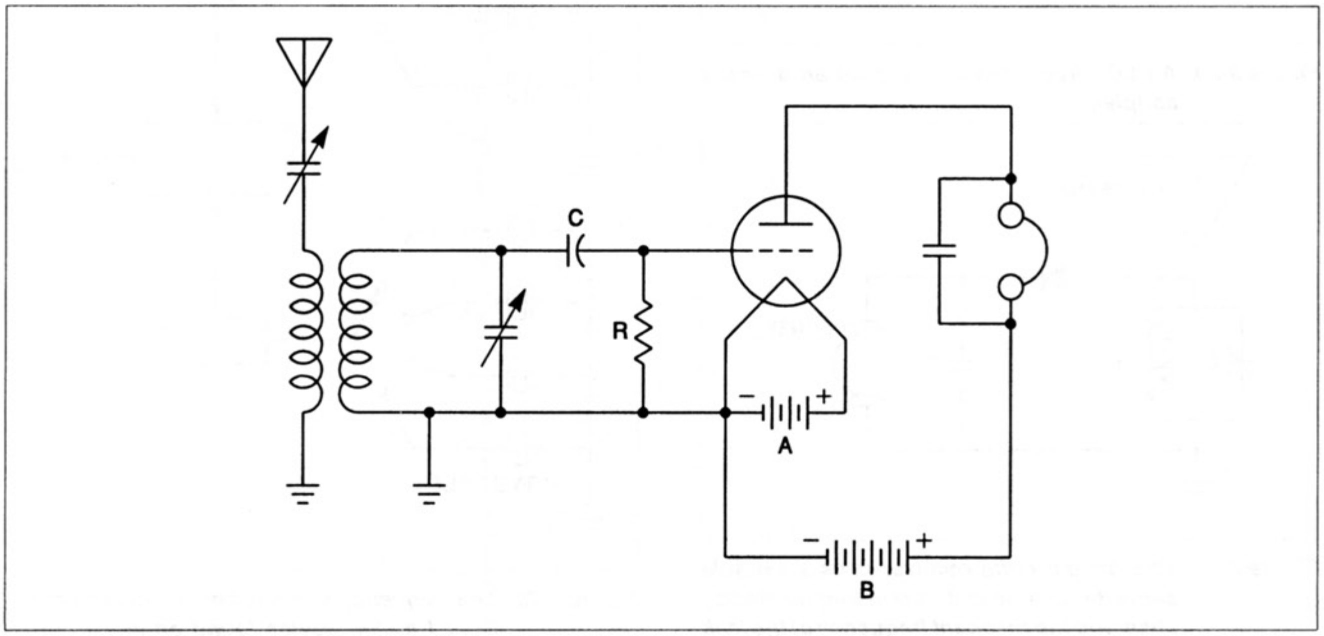

Figure 3-19. An isolation transformer can be used as an antenna coupling device.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

i. Speakers and Microphones

Aircraft radios often supply an audio output for the pilot and voice transmitters require an audio input from a microphone. A speaker is a device that transforms electrical signals into sound waves. A dynamic speaker is shown in figure 3-25.

When the audio frequency signal is applied to the windings in the speaker, it sets up a magnetic field that expands and contracts at an audio rate. This field causes the metal diaphragm to vibrate at a corresponding rate to produce the movement of air that generates sound waves. Dynamic microphones are available which operate in the opposite way. Many newer and more efficient types of microphones are now being manufactured, but they all work by transforming the vibrations of sound waves into varying electrical signals.

j. Audio Control Panels

When an aircraft has more than one radio, an efficient means of switching the microphone and speaker connections from one radio to another is needed. The audio control panel performs this function. A typical audio control panel is shown in figure 3-26. An audio control panel is not a radio because it only uses audio frequencies, but it is associated with the radios in the aircraft. This audio control panel has a row of toggle switches that can be used to connect the audio output of the various radios to the speaker or headphones. It also has a rotary selector switch to connect the microphone audio output to the different radio transmitters and intercom systems available for the aircraft. The audio control panel illustrated also has three lights that are the indicators for the marker beacon system on the aircraft. The marker beacon system will be described later.

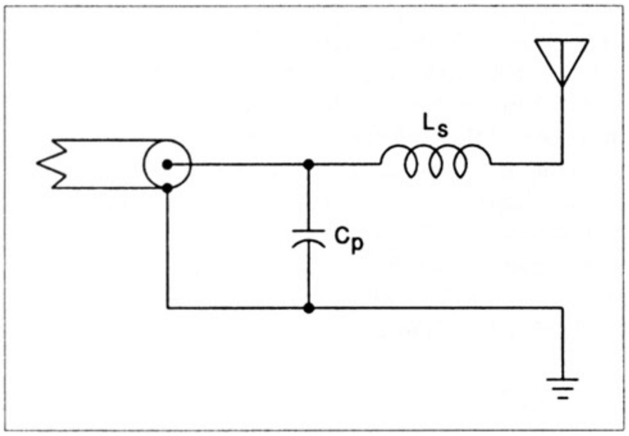

Figure 3-20. An LC circuit can be used as an antenna coupler.

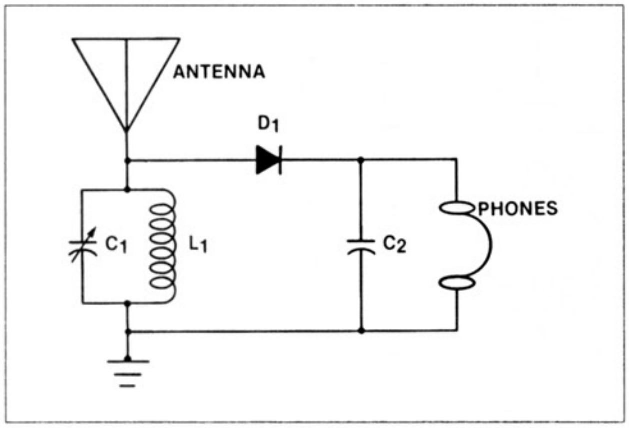

Figure 3-21. This simple radio receiver uses a variable capacitor to tune in different frequencies by changing the resonant frequency of the tank circuit.

B. Regulations and Standards for Radios

Aircraft avionics equipment might have to comply with a number of different regulations and standards depending on the type of equipment and the type of aircraft in which it is installed. Regulations from the FAA and the FCC apply to the manufacture and use of most types of equipment and carry the force of law. FAA standards for equipment are usually in the form of TSO (Technical Standard Order) approvals. FCC rules generally apply to equipment which produces radio waves. An FCC Station License is required for aircraft

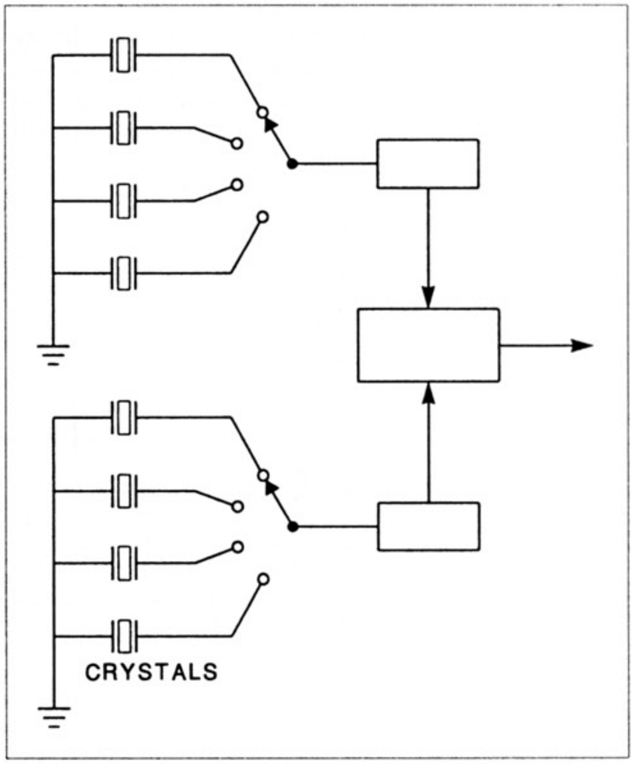

Figure 3-22. The frequency synthesizer is a crystal controlled tuning device found on many modern radios and other electronic units.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

that have radio transmitters other than ELT. Each different type of transmitter must be listed on the license that is displayed in the cockpit. The role of ARINC in established standards has been described earlier. These standards apply to the equipment in air carrier jets and bizjets primarily. In addition to the rules for the equipment itself, there are some FAA Regulations concerning the use of radio equipment in flight. Some of these FAA rules are given here. Others will be covered later when the specific types of equipment to which the rules apply are discussed.

FAR 91.130 (c)

No person may operate an aircraft in an Airport Radar Service Area (ARSA) unless two-way radio communication is established and maintained with

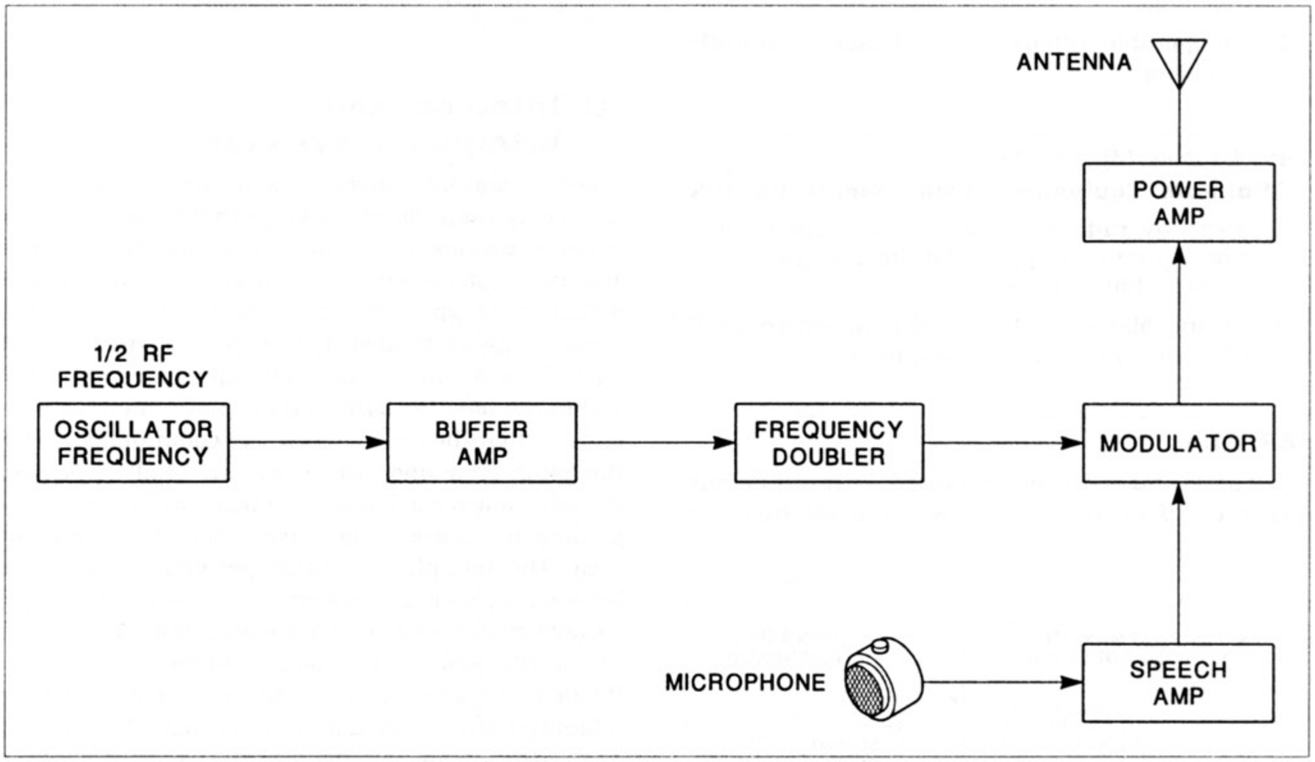
ANTENNA
POWER AMP
1/2 RF FREQUENCY
OSCILLATOR FREQUENCY
BUFFER AMP
FREQUENCY DOUBLER
MODULATOR
MICROPHONE
SPEECH AMP

Figure 3-23. Simplified block diagram of a VHF voice radio transmitter.

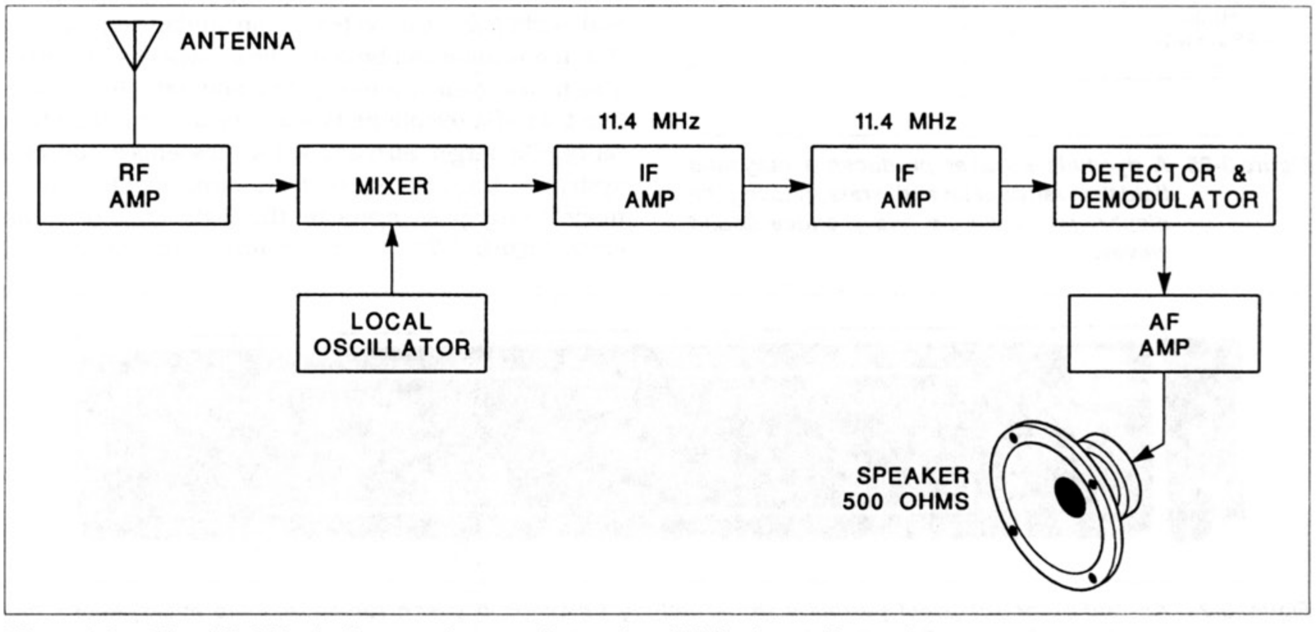
ANTENNA
RF AMP
MIXER
11.4 MHz
IF AMP
11.4 MHz
IF AMP
DETECTOR & DEMODULATOR
LOCAL OSCILLATOR
AF AMP
SPEAKER 500 OHMS

Figure 3-24. Simplified block diagram of a superheterodyne VHF voice radio receiver.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

ATC. A transponder with Mode C automatic reporting of aircraft pressure altitude is also required in all ARSAs.

FAR 91.131 (c) and (d)

No person may operate an aircraft in a Terminal Control Area (TCA) unless it has:

1. A two-way radio with appropriate frequencies available.
2. An operable transponder with Mode C altitude reporting.

FAR 91.205 (d) and (e)

Minimum Equipment Requirements for IFR

1. Two-way radio communications and navigation equipment appropriate to the ground facilities that will be used.
2. At and above 24,000 ft. MSL; approved DME (distance measuring equipment).

FAR 91.511

Large and turbine-powered, multi-engine airplanes, if operating over water more than 30 minutes or 100 nautical miles from the nearest shoreline, must have:

1. Radio communication equipment appropriate to the ground facilities.
2. Two transmitters.
3. Two microphones.
4. Two headsets (or headset and speaker).
5. Two independent receivers.
6. If needed, one HF transceiver.

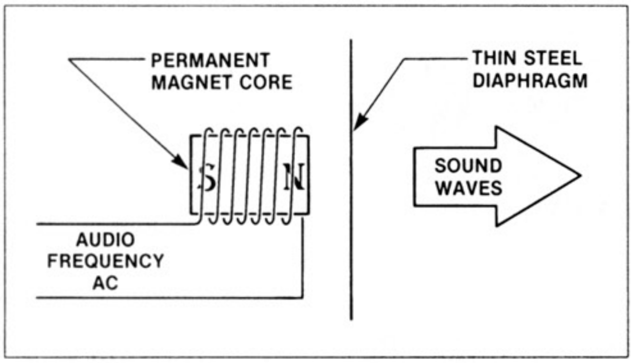

Figure 3-25. A dynamic speaker produces a magnetic field that varies at an audio rate, causing the diaphragm to vibrate and produce sound waves.

C. Intercom and Interphone Systems

Intercom and interphone systems are not radio systems, they use audio signals to permit communication between various points in and around the aircraft. The two systems operate in a similar manner, the difference is who uses the systems and where the phone jacks are located. The intercom system is used for voice communications from one point to another within the aircraft. Large aircraft have intercom systems so that the cockpit crew can communicate with the cabin crew and vice versa. On small airplanes, the intercom is used to communicate within the cockpit area and is needed because of noise in the cockpit area. The interphone system permits conversation between the cockpit and someone outside the aircraft, usually maintenance or service personnel. The operation of intercom and interphone systems is the same. Phone jacks are available at different locations where a handset or headset can be connected. The handset or headset contains a microphone, a small speaker and a push-to-talk switch (PTT). The phone jacks and wiring are connected to an audio amplifier so that the volume can be controlled. Switches are available to select the desired system and a ringing system like that of a telephone is used for alerting the other party. On larger aircraft, a PA (passenger address) system is included so that announcements can be made to the passengers by the flight crew or cabin crew. Figure 3-27 shows the interphone system for

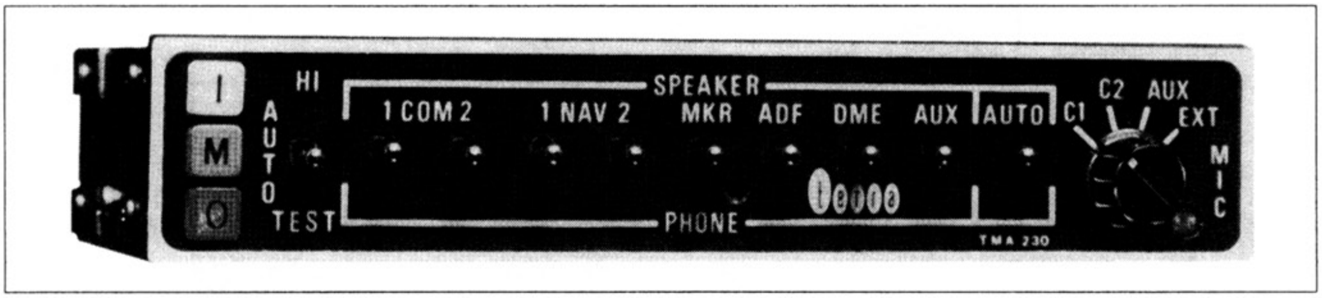

Figure 3-26. An audio control panel performs the switching functions between radios and the microphones and speakers. (Courtesy Terra Corp.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

a bizjet airplane. External interphone jacks are located in the nosewheel area, avionics equipment bay area and in the aft fuselage near the auxiliary power unit (APU). These external jacks permit communication between the cockpit and maintenance personnel at these locations outside the aircraft.

D. Communications Radios

There are a number of different radio communications systems available for aircraft use. They differ primarily in the frequencies used and the type of communication involved. The most important use of communications radios is for Air Traffic Control since the controllers need to be in contact with the pilots to give necessary instructions. The general trend since the 1930s has been the use of higher frequencies and the development of specialized communications for other than ATC purposes.

1. HF Communications

Up until the 1940s, most aircraft radio communications utilized frequencies in the LF, MF and HF bands because suitable equipment was not available to use higher frequencies. Aircraft HF radios operate on frequencies between 2 and 30 MHz. The only modern aircraft that carry HF comm radios are those that operate long distances over water or in the remote regions of the earth. Air carrier jets and bizjets that routinely fly the Atlantic and Pacific oceans will have HF comm radios for ATC purposes.

The HF comm radios have a maximum reception range of about 1,500 to 2,000 miles compared to a maximum of about 250 miles for VHF comm. The reception range of VHF comm radios is restricted to line-of-sight distances as shown in figure 3-28. The probe and flush mount antennas used for HF comm require a special antenna tuning and coupling device. This is automatically repositioned each time a new frequency is selected in order to tune the antenna for that particular frequency. Smaller aircraft with HF comm will use a long wire antenna that usually extends from a wing tip up to the vertical fin. Up to the 1960s, many aircraft used a long wire trailing antenna which extended out the aft fuselage of the airplane. This antenna could be run in and out to select the proper antenna length. It is not suitable for high speed aircraft, so it is little used today. HF comm radios utilize ground and sky waves to achieve their greater reception range. Aircraft HF transmitters produce an output power of 80–200 watts which is much higher than the output power typically found with VHF transmitters. This is necessary to achieve long distance communication. A disadvantage of HF is that it is more affected by atmospheric interference than VHF. Sometimes an aircraft in the middle of the ocean will lose communication because of thunderstorms or other disturbances.

2. VHF Communication

The use of frequencies in the VHF band for aircraft communication was developed in the 1940s. VHF provides much clearer reception and is much less affected by atmospheric conditions. EM waves in the VHF band are space waves so that the reception range is limited to line-of-sight distances. At 1000 ft., the reception range is about 30–40 miles. The maximum reception range using ground based stations is about 250 miles at altitudes above 35,000 ft. Much less power is required for VHF than for HF comm. Aircraft VHF transmitters have an output power of 5–20 watts. The standard radio communications system in the U.S.A. for ATC purposes is VHF. This is also true for most other countries of the world. The International Civil Aviation Organization (ICAO) has designated VHF as the standard radio communication system for ATC purposes over land.

The range of frequencies used for VHF comm is 118–137 MHz using AM modulation. In the 1950s, an aircraft VHF comm radio could tune only 90 channels or different frequencies. Later the separation between channels was reduced to produce 360 channels. Modern VHF comm radios have 720 or 760 channels available. The spacing between channels is now 25 kHz, so that adjacent usable frequencies would be 120.15, 120.175, 120.20 etc. The latest models of aircraft VHF comm radios use lighted displays that employ LEDs (light emitting diodes), LCDs (liquid crystal displays), or gas discharge tubes. Figure 3-29 shows a complete set of radios that might be found on a typical general aviation airplane. The radios use lighted displays for the frequencies and other information that is needed. The use of lighted displays and crystal controlled tuning has eliminated the complex switching systems employed on earlier tuners. The modern radio displays two different frequencies, the one on the left is the active frequency and the one on the right is the standby frequency that is held in memory. This is a very nice feature which allows the pilot to switch the two numbers in the display by simply pushing a transfer button. The transfer buttons can be seen in figure 3-29 on the VHF comm and NAV radios. Technicians should be familiar with the use of aircraft radios for troubleshooting purposes and also because they might have to taxi an airplane which may require the use of the radio. The antennas used with VHF comm are Marconi ¼ wave monopoles that use

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

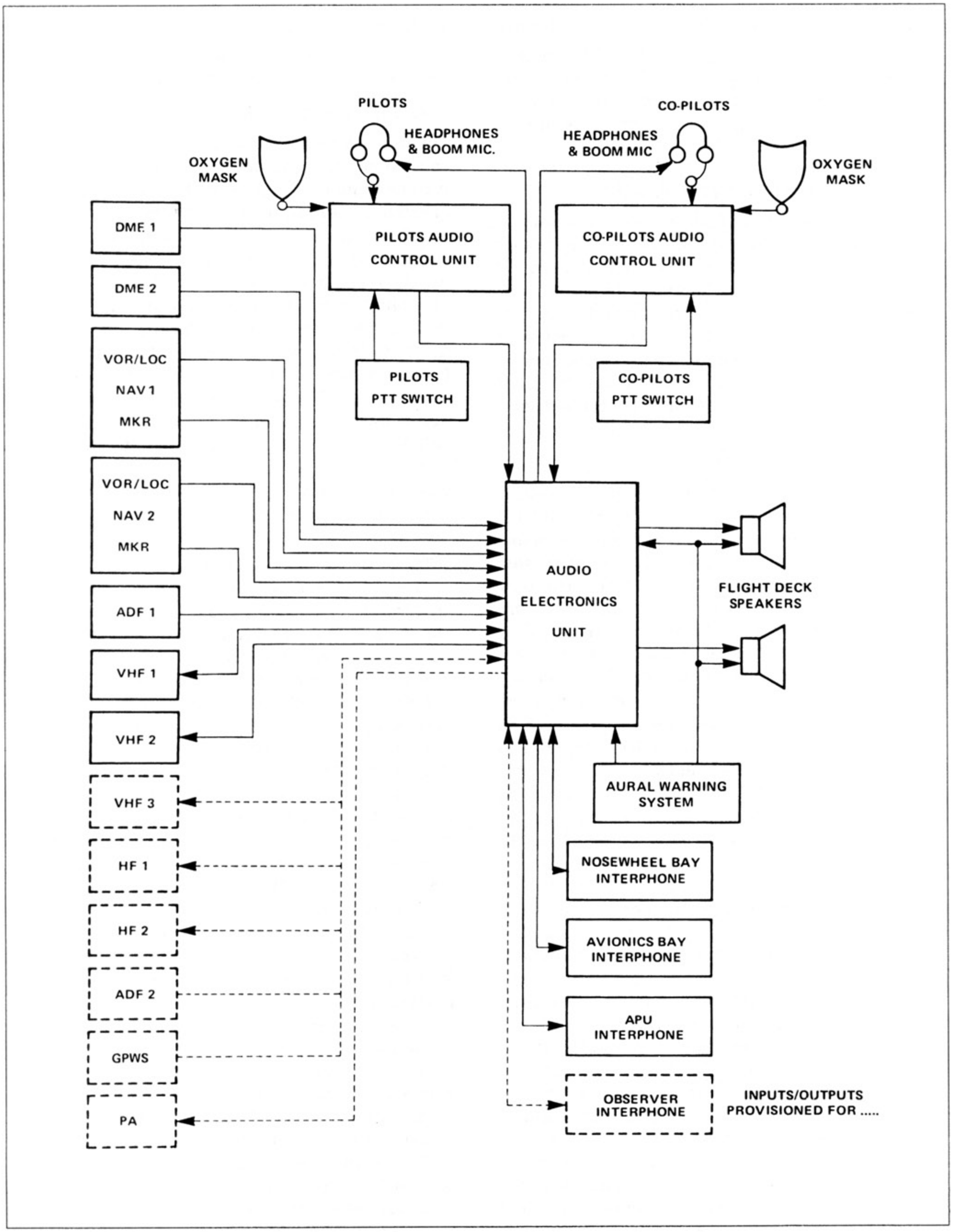
PILOTS
HEADPHONES & BOOM MIC.
CO-PILOTS
HEADPHONES & BOOM MIC
OXYGEN MASK
OXYGEN MASK
DME 1
DME 2
PILOTS AUDIO CONTROL UNIT
CO-PILOTS AUDIO CONTROL UNIT
VOR/LOC NAV 1 MKR
PILOTS PTT SWITCH
CO-PILOTS PTT SWITCH
VOR/LOC NAV 2 MKR
AUDIO ELECTRONICS UNIT
FLIGHT DECK SPEAKERS
ADF 1
VHF 1
VHF 2
VHF 3
HF 1
HF 2
ADF 2
GPWS
PA
AURAL WARNING SYSTEM
NOSEWHEEL BAY INTERPHONE
AVIONICS BAY INTERPHONE
APU INTERPHONE
OBSERVER INTERPHONE
INPUTS/OUTPUTS PROVISIONED FOR .....

Figure 3-27. The intercom and interphone system for a corporate jet. The external interphone jacks are shown in the lower right. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

vertical polarization. There is a separate antenna for each VHF comm radio.

3. Radiotelephone

Aircraft often carry a radiotelephone system which is somewhat similar to the portable cellular phone available for cars. It employs radio signals to permit telephone calls to be made from the aircraft in flight. The frequencies used are 450–500 MHz in the UHF band. The antenna used is a Marconi antenna of a slightly different shape and size compared to a VHF comm antenna.

4. SATCOMM

A very recent development in aircraft systems is Satellite Communications or SATCOMM. A UHF radio is installed in the aircraft to communicate with commercial satellites in orbit overhead. To date, it is being used primarily for telephone calls from bizjets and air carrier jets. It is beginning to be used for datalinks from an aircraft in flight to the airline computer system. This permits monitoring of the progress of the flight and the status of the aircraft systems. In the future, SATCOMM will be used to replace HF comm for communications and ATC purposes for aircraft over the oceans or remote areas. The equipment currently available is very expensive, usually costing hundreds of thousands of dollars. The antenna used with SATCOMM is a special type that must be installed on the top of the aircraft.

5. Selcal

Selcal is an abbreviation of selective calling, a special communications system for air carrier aircraft. Selcal is not a separate radio system, it is a piece of equipment that is connected to the existing comm radios on the aircraft. It is connected to the VHF and HF radios on the aircraft.

The system is used for communications between aircraft in flight and certain airline managers. It is called selective calling because it works somewhat like a telephone system. An example of a SELCAL decoder and the connections to the aircraft's VHF and HF radios is shown in figure 3-30. Each aircraft is assigned a code number which is a part of the SELCAL equipment. When the proper code is received, a tone is heard in the cockpit to tell the crew that someone is calling them. They pick up a handset in the cockpit and talk to the person that has called them. The code consists of four tones that are transmitted to the aircraft in series. Each of the four tones has twelve possible frequencies, so that over 20,000 different codes are available. An example of a SELCAL communication will illustrate the operation of the system. The head dispatcher for the airline is sitting in his office in Chicago and needs to call the flight crew of one of the airline's aircraft to give them a message. According to the schedule, the airplane is somewhere between Boston and Atlanta. The dispatcher picks up his telephone and dials a special access code and the code for that airplane. The signal is sent out over many different ground transmitters and received by hundreds of airplanes in flight. The phone will ring only in the cockpit of the airplane he is calling and he will pass on the message when they answer. The SELCAL system is a great help to the airline when they must reroute a flight or pass along important information to the crew in flight.

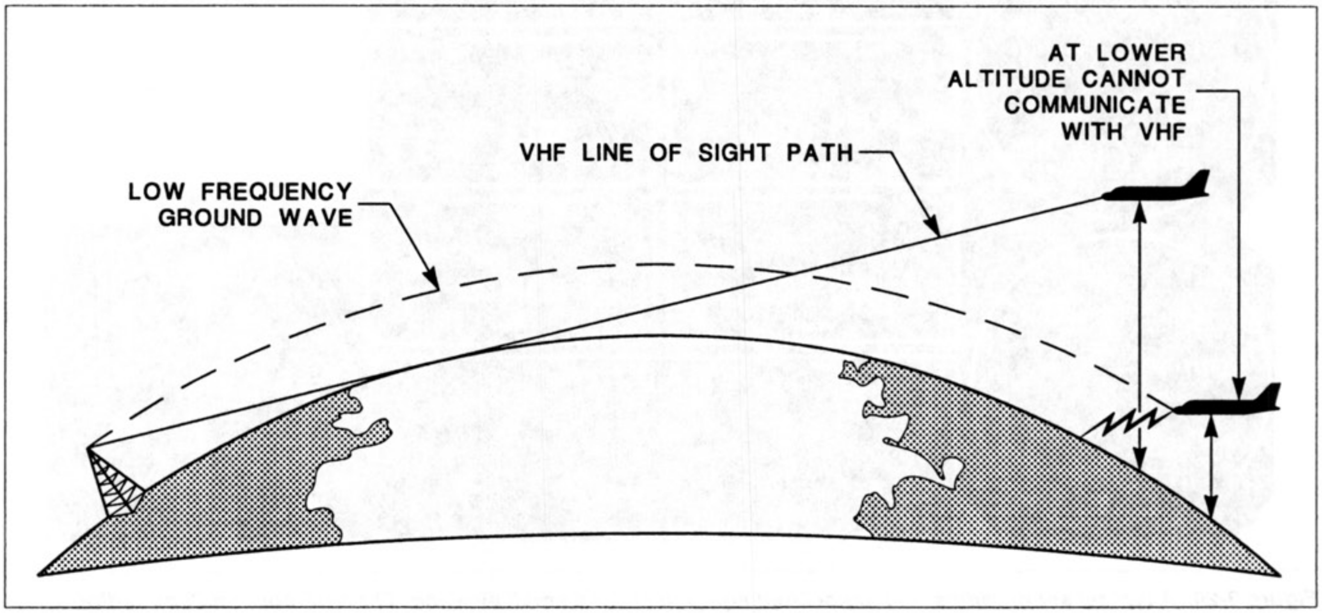

Figure 3-28. Example of the line-of-sight restriction that applies to VHF and other space wave transmissions.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

E. Emergency Locator Transmitters (ELTs)

The ELT is a self-contained transmitter that is designed to help locate an airplane after a crash. A typical ELT with its antenna and coaxial cable is shown in figure 3-31. It is required on most small airplanes, but is not required on air carrier jets and bizjets. The ELT is battery powered and is automatically turned on by crash forces. It will transmit a special swept tone for 48 hours on two different emergency frequencies. The two frequencies are 121.5 MHz and 243.0 MHz; 121.5 is the civilian emergency frequency and 243.0 is the military emergency frequency. The transmitter is activated by an acceleration operated switch when a rapid deceleration force is applied along the longitudinal axis of the aircraft. The ELT must be installed as far aft as possible but in front of the tail surfaces since this area has been shown to remain intact in most airplane crashes. The batteries in the ELT must be replaced or recharged at specific intervals as required by the FARs. There are times when an aircraft technician may need to test an ELT so he should be familiar with the procedure. If possible, the ELT should be tested with the antenna disconnected or shielded to prevent the transmission of emergency signals into the air. If this cannot be done, it is still permissible to test the ELT, but only during the first five minutes of any hour and for three audio sweeps maximum. A VHF comm radio is turned on and tuned to 121.5 MHz. The ELT is then switched on manually until the signal is heard on the receiver and then switched off again.

FAR 91.207

Emergency Locator Transmitters

No person may operate a U.S. registered civil airplane unless it meets the applicable requirements listed below for ELTs.

1. Each emergency locator transmitter must be in operable condition and meet the requirements of TSO-C91 or TSO-C91A and it must be installed as far aft as practicable.
2. Batteries used in the ELT must be replaced or recharged as appropriate:
   a. When the transmitter has been in use for more than one cumulative hour; or
   b. When 50% of the useful life has expired.

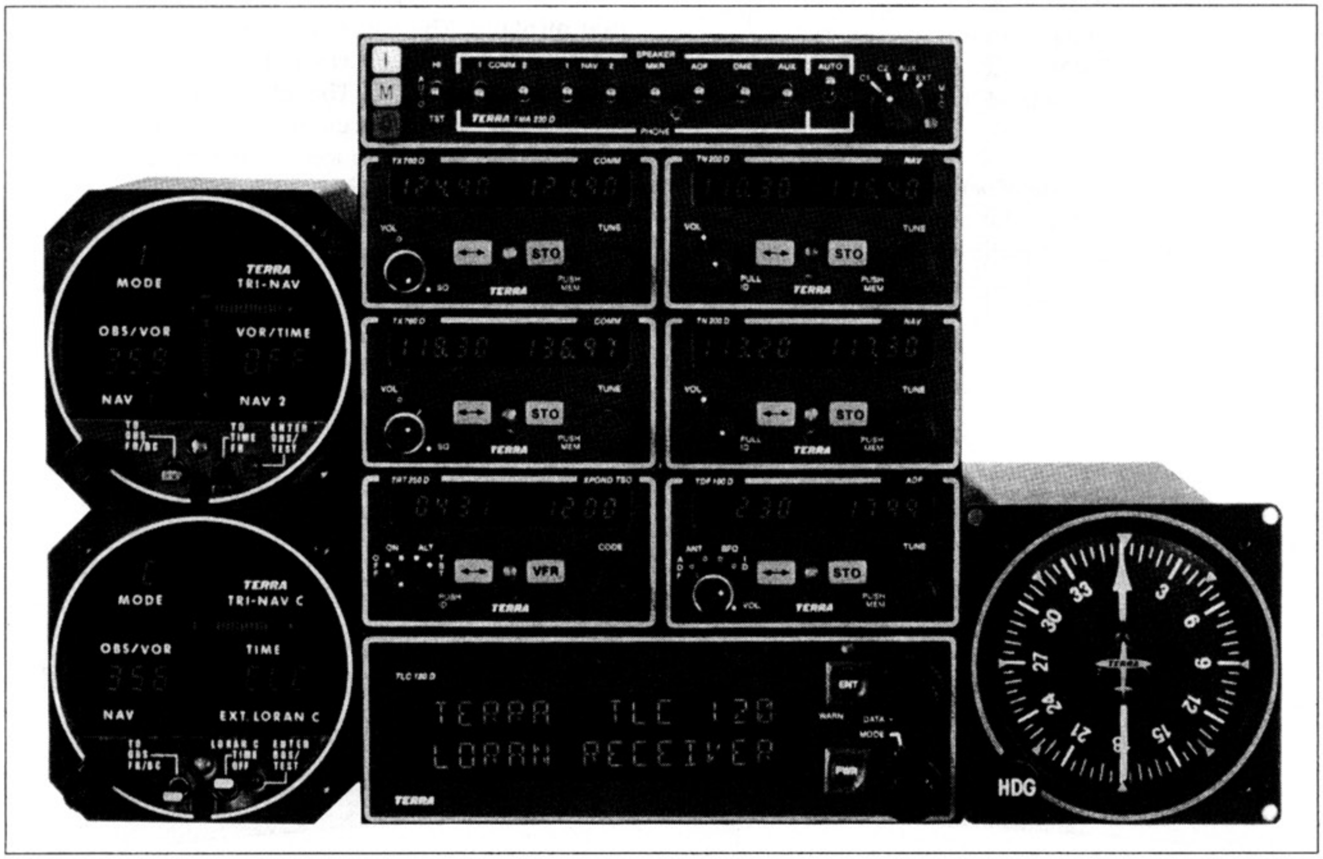
SPEAKER
COMM
NAV
MKR
ADF
DME
AUX
AUTO
PHONE
TERRA TMA 350 D
TX 760 D
COMM
TN 200 D
NAV
VOL
TUNE
STO
PUSH MEM
TERRA
MODE
TERRA TRI-NAV
OBS/VOR
VOR/TIME
NAV
NAV 2
TRT 250 D
XPOND TSO
CODE
VFR
TDF 100 D
ADF
TERRA TRI-NAV C
TIME
EXT. LORAN C
TLC 120 D
TERRA TLC 120
LORAN RECEIVER
ENT
WARN
DATA
MODE
PWR
HDG

Figure 3-29. A typical set of radios and associated equipment for a small airplane. The VHF com and nav radios show both active and standby frequencies. (Courtesy Terra Corp.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

3. The expiration date for the replacement or recharge of the battery must be legibly marked on the outside of the transmitter and entered in the aircraft maintenance record.
4. The requirements for ELT do not apply to:
   a. A newly acquired aircraft that must be ferried to a place where the ELT will be installed.
   b. An aircraft with an inoperative ELT that must be ferried to a place for ELT repair.
   c. Turbojet powered aircraft.
   d. Scheduled air carrier flights.
   e. Training flights conducted entirely within 50 nautical miles of the airport of operations.
   f. Design and test flights.
   g. Delivery flights of new aircraft.
   h. Aircraft engaged in aerial application of chemicals for agricultural purposes.
   i. Research and development aircraft.
   j. Exhibition and air racing aircraft.

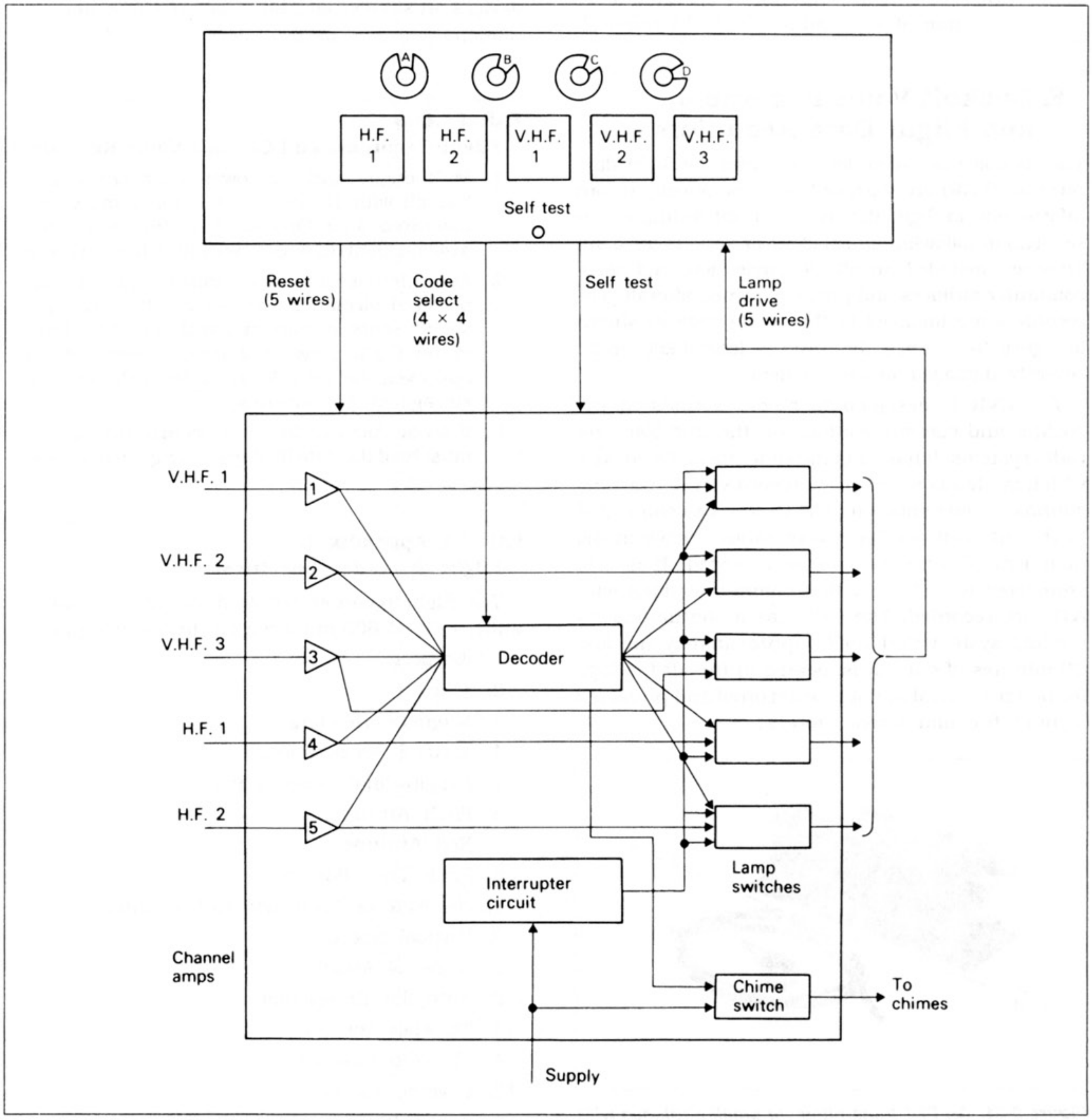
A
B
C
D
H.F. 1
H.F. 2
V.H.F. 1
V.H.F. 2
V.H.F. 3
Self test
Reset (5 wires)
Code select (4 × 4 wires)
Self test
Lamp drive (5 wires)
V.H.F. 1
V.H.F. 2
V.H.F. 3
H.F. 1
H.F. 2
1
2
3
4
5
Decoder
Lamp switches
Interrupter circuit
Channel amps
Chime switch
To chimes
Supply

Figure 3-30. A typical SELCAL decoder unit showing the connections to the VHF and HF com radios.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

k. Aircraft equipped to carry only one person.

l. An aircraft during any period in which the ELT has been temporarily removed for inspection, repair, modification or replacement, subject to the following:
    1. A maintenance record entry must be made that includes the date of removal, the serial number and the reason for removal.
    2. A placard must be placed in view of the pilot which states "ELT not installed".
    3. The aircraft must not be operated more than 90 days after initial ELT removal.

F. Cockpit Voice Recorders and Flight Data Recorders

The cockpit voice recorder (CVR) and the flight data recorder (FDR) are designed to automatically record information in flight that can be used during an investigation following an accident or serious incident. They are installed on all air carrier jets and some commuter airliners and privately owned aircraft. The recorders are installed in the aft fuselage as shown in figure 3-32 since this area is least likely to be severely damaged in an accident.

The CVR is designed to record sounds in the cockpit and communications on the intercom and radio systems. It has a hot microphone in the cockpit which is always activated to record voices, warning sounds, engine noise etc. The CVR is also connected to the intercom so that conversations between the members of the crew can be recorded. It is also connected to radios so that communications with ATC are recorded. The CVR has a continuous recording system that holds approximately the last 30 minutes of data. It is located in the aft fuselage for better survival and it is waterproof and protected against fire and impact forces.

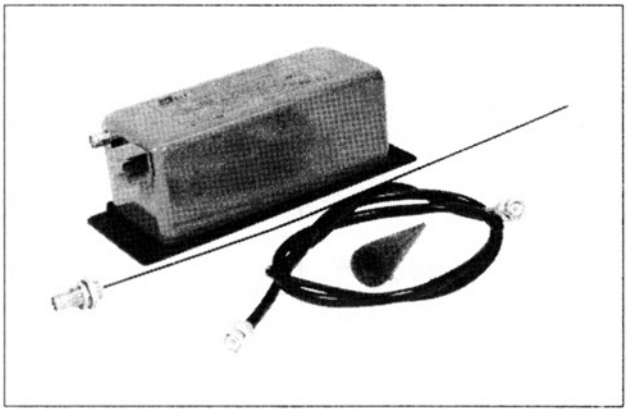

Figure 3-31. An ELT transmitter for small airplanes with the antenna and coaxial cable.

The flight data recorder has many more inputs than the cockpit voice recorder. It has a recording time of 8 hours on smaller aircraft and about 24 hours on larger aircraft. The CVR and FDR are located in the same area of the aft fuselage and have similar protection from water, fire etc. Some of the typical types of data that are recorded on the FDR are listed below.

Air carrier jets have been required to carry CVRs and FDRs for some years, but recently new regulations have gone into effect that require these devices on smaller aircraft. Some of these new regulations are summarized here.

FAR 91.609
Flight Recorders and Cockpit Voice Recorders

1. Multi-engine turbine powered airplanes or rotorcraft with 10 passenger seats or more manufactured after October 11, 1991 must have a digital flight data recorder with 8 hours storage.
2. After October 11, 1991, multi-engine turbine powered airplanes and rotorcraft with 6 passenger seats or more and with a required minimum flight crew of 2 pilots must have an approved cockpit voice recorder with minimum storage of 15 minutes.
3. If an accident or incident occurs, the operator must hold the data 90 days or longer if requested.

FAR 91 Appendix E
Flight Recorder Specifications

The flight recorder required for certain aircraft under FAR 91.609 must record the following items:

1. Airspeed.
2. Altitude.
3. Magnetic Heading.
4. Vertical Acceleration.
5. Longitudinal Acceleration.
6. Pitch Attitude.
7. Roll Attitude.
8. Pitch Trim Position.
9. N_1, EPR or Prop RPM and Torque.
10. Vertical Speed.
11. Angle of Attack.
12. Autopilot Engagement.
13. TE Flap Position.
14. LE Flap Position.
15. Reverse Thrust.
16. Spoiler/Speedbrake Position.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

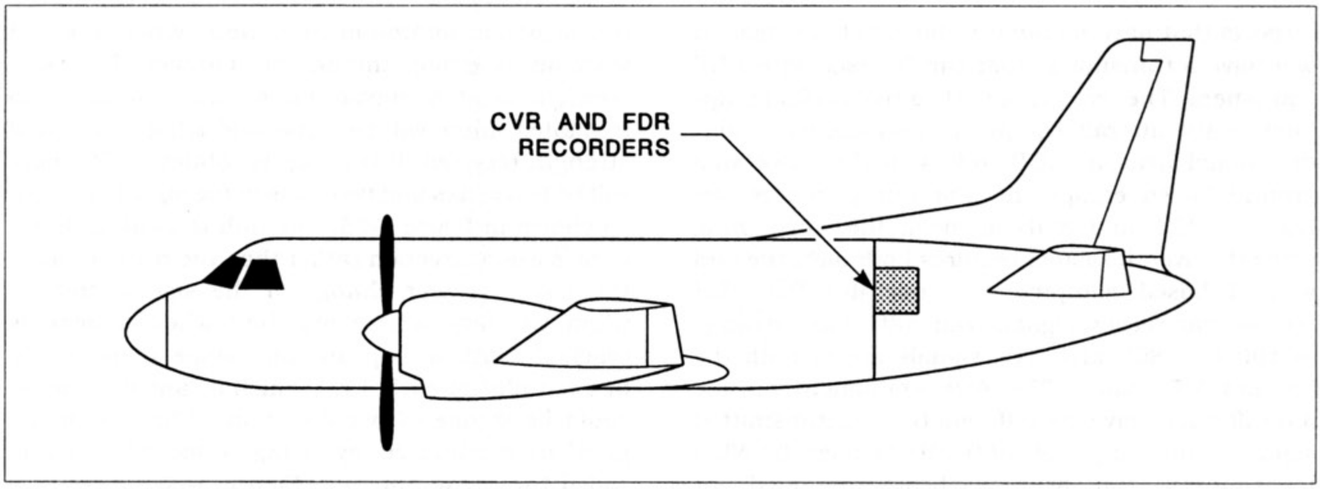

Figure 3-32. Large commuter aircraft and air carrier jets have a Cockpit Voice Recorder and Flight Data Recorder installed in the aft section of the fuselage.

G. Navigational Systems

There is a much wider variety of navigational systems available to aircraft than communications systems. There is a wide range of capabilities for the various systems and some of them are very specialized. Some of them have been around for 50 years and others are much newer. We will start with a brief description of a NAV system that is now obsolete. A basic understanding of this older system will make clearer the greater capabilities of more modern NAV systems.

1. Four-course Radio Range

The four-course radio range was the first radio navigation system developed in the U.S. to guide aircraft in poor weather conditions. Before its invention, the standard navigation system consisted of powerful light beacons that the pilot followed during either day or night flying. The four-course radio range gets its name from the fact that only four pathways or courses were usable with this system. The ground stations transmitted signals in the LF band, so the signals could be very difficult to use when atmospheric conditions caused interference. The antenna arrangement on the ground transmitted four different directional radio beams as shown in figure 3-33. The signals transmitted north and south of the station were modulated with the Morse code letter "N". The signals transmitted east and west were modulated with the letter "A". If the pilot was flying in any one of these sectors and listening to the signals, the Morse code keying for the letter A or N, as appropriate, would be heard. If the aircraft was exactly along a line where the signals merged, a steady hum or tone would be heard in the headphones. The pilot navigated by listening to the sounds and aligning the airplane along the centerline of one of the four courses. As long as the aircraft was on course, the pilot would hear the steady tone in the headphones. Since the radio signals were affected by atmospheric noise, the pilot often had to fly for hours on end listening to noise and static and trying to pick out the correct signals. The system was not very easy to use and the last one in the U.S. was taken off the air in the 1970s.

2. Automatic Direction Finder (ADF)

The ADF system has been in use since the 1930s and even though it is not as accurate or easy to use as the more common VOR system, it is still widely used because it is inexpensive. Many smaller

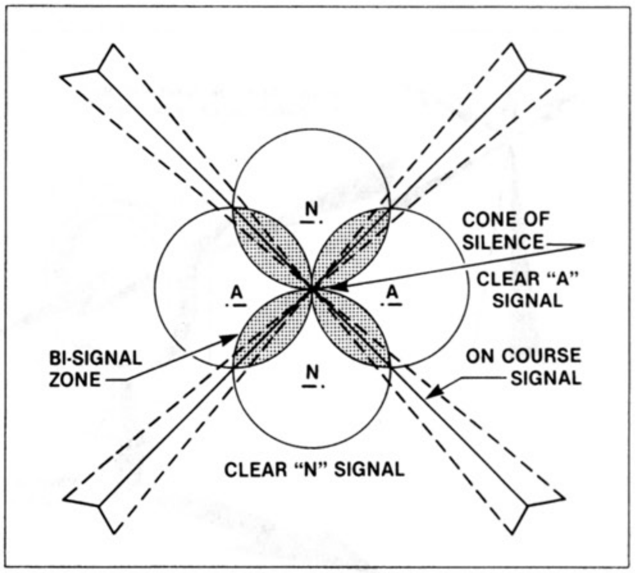

Figure 3-33. The first radio navigation system was the 4 course radio range which operated in the LF band.

airports that have no other radio aids to navigation will have a transmitter that can be used with ADF equipment. The term automatic direction finder applies to the aircraft equipment and the term non-directional beacon (NDB) refers to the associated ground based equipment. For our purposes, we can use ADF and NDB to mean the same thing since the overall system requires both airborne and ground based equipment to operate. The ADF receiver can receive signals transmitted in the range of 190 to 1,800 kHz. The signals are in both the LF and MF bands. The ADF equipment on the aircraft can receive two different types of transmitted signals. The range 190–500 kHz is used by NDB transmitters that are specifically designed for aircraft use. The range of 550–1,800 kHz is the band used by commercial AM broadcast stations. The broadcast stations can be used for navigation, but are not as good as the NDBs because the pilot does not always know exactly where the transmitter site is located. The NDB locations are shown on aeronautical charts so the location of the transmit antenna can be determined more accurately. The signals transmitted from the ground sites are omnidirectional. The ADF equipment determines station direction through the use of a directional antenna. The directional antenna is called a loop antenna. The older versions were actually shaped like a loop as shown in figure 3-34. The strength of the output signal from the antenna depends on the angle between the plane of the loop and the direction of travel of the EM wave. When the EM wave is at right angles to the plane of the loop, the signal is minimum or a null. When the EM wave and the loop antenna are parallel, the signal strength is at a maximum. As the loop antenna is rotated, there will be a rise and fall in the signal strength received. If the loop is rotated 360°, there will be two peaks and two nulls in the signal strength as shown in figure 3-35. The null is used to determine station direction rather than the peak because there is a greater change in the signal strength when the null is reached than when a peak is reached. With a loop antenna alone there would be two nulls for each 360°, which means the station could be in one of two directions. This ambiguity problem is removed by using a second antenna called the sense antenna. Figure 3-36 shows how the signals from both antennas are combined to determine the direction of the transmit station. Older loop antennas were rotated by electric motors. The newer types of loop antenna do not rotate themselves, they use an electronic system to cause rotation of the signal. In either type, the principle of operation is the same. The simple type of cockpit indicator used with an ADF is called a radio compass indicator or ADF indicator. This type of simple ADF indicator can be seen in figure 3-37. It has a compass rose with degree markings and a pointer which points in the direction of the transmitter site. When the pointer is straight up, it shows that the station is directly in front of the aircraft. If the pilot keeps the pointer at the top of the instrument, it will guide him to the location of the transmitter. On newer types of ADF equipment, the cockpit indicator is called a radio magnetic indicator (RMI).

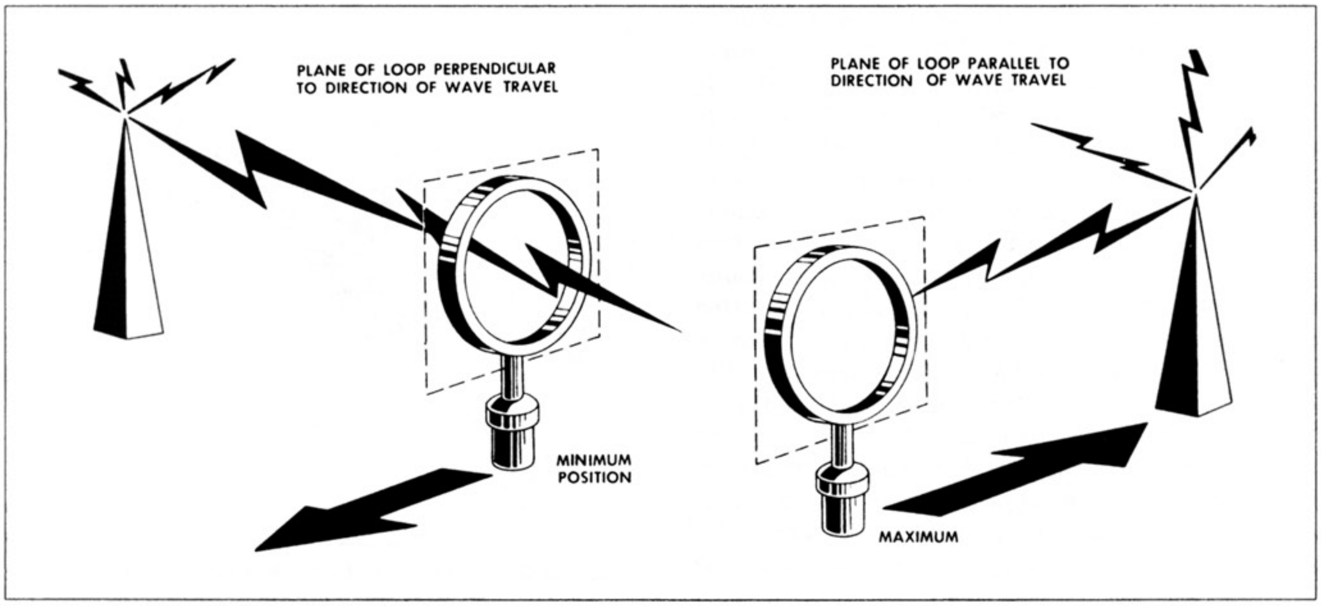
PLANE OF LOOP PERPENDICULAR TO DIRECTION OF WAVE TRAVEL
PLANE OF LOOP PARALLEL TO DIRECTION OF WAVE TRAVEL
MINIMUM POSITION
MAXIMUM

Figure 3-34. Example of the directional characteristics of a loop antenna.

The appearance of this RMI is illustrated in figure 3-38. The RMI has two pointers that can be operated by two ADF signals, two VOR signals, or one of each. It also includes a compass card which is slaved to a remote DG and flux gate compass as described in chapter 1. The primary use of ADF is for approaches to airports that don't have any better radio aids to navigation available.

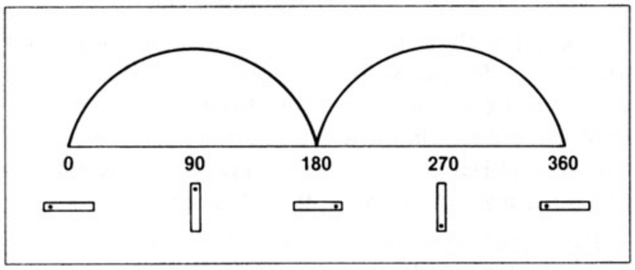

Figure 3-35. If a loop antenna is rotated 360°, two peaks and two nulls will occur.

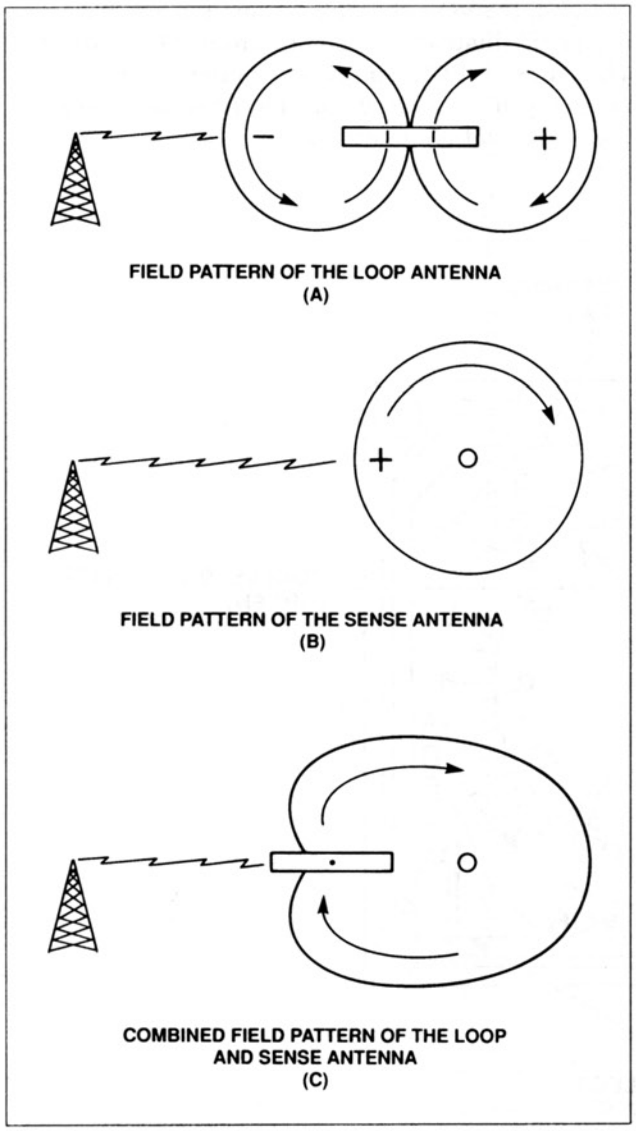
−
+
FIELD PATTERN OF THE LOOP ANTENNA
(A)
+
FIELD PATTERN OF THE SENSE ANTENNA
(B)
COMBINED FIELD PATTERN OF THE LOOP AND SENSE ANTENNA
(C)

Figure 3-36. When the signals from the loop and sense antennas are combined, the ambiguity problem is eliminated.

3. Very High Frequency Omnirange (VOR)

The VOR system is the standard IFR radio navigation system for cross-country flying in the U.S. and most of the rest of the world. The VOR system was developed to overcome some of the problems with the old four course radio range. The major advantages of VOR over the older system are:

1. An infinite number of radials or courses are available, not just four.
2. Since it operates in the VHF band, the VOR is much less affected by thunderstorms and atmospheric conditions.
3. The VOR is much more accurate.
4. The VOR is much easier to use, the pilot follows an indicator needle instead of listening to Morse code signals for navigation.

The first VOR was installed in Indianapolis in 1939 and by the 1950s, coverage was almost complete over the entire U.S. Since the reception range of VOR is limited by the same line-of-sight considerations that apply to VHF comm, about 1,500 VOR ground sites are required for nationwide coverage. The range of frequencies used by VORs is 108–118 MHz. The ground sites transmit two kinds of signals, a reference signal and a rotating signal. The reference signal uses FM and the rotating signal uses AM. The two signals are aligned so that they will be in phase when the receiver is straight north of the ground site. The phase angle

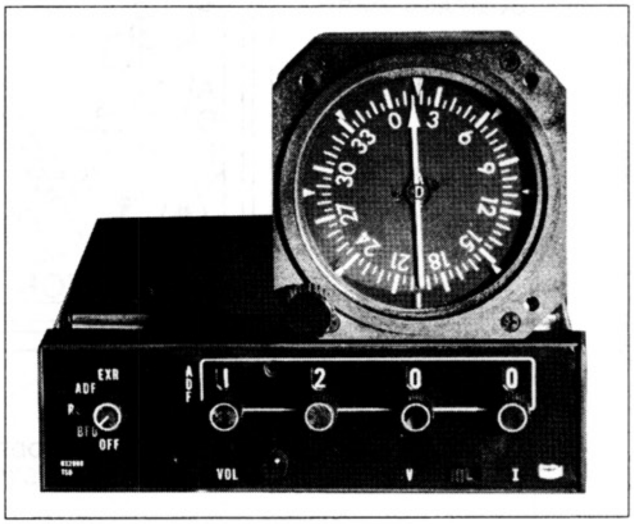

Figure 3-37. An ADF radio receiver and its indicator instrument.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

for east is 90° and for south it is 180°. No matter where the aircraft is in relation to the ground site, the direction can be determined by measuring the phase angle between the two signals. The cockpit indicator for VOR is shown in figure 3-39. This indicator has three different parts: the course deviation indicator (CDI), the omni bearing selector (OBS) and the TO-FROM indicator.

1. OBS — The omni bearing selector is a knob that the pilot rotates to select the desired radial from the station.
2. CDI — The course deviation indicator is the vertical needle or pointer. When it moves to the left of center, the pilot must turn left to get back on the desired radial set with the OBS.
3. TO-FROM Indicator — The to-from indicator is a small window where one of three words is displayed: TO, FROM or OFF. The word TO means that the pilot is flying toward the station. The word FROM means the pilot is flying away from it and the word OFF means that usable signals are not being received or the indication is changing between the TO and FROM conditions.

The procedure for tuning in a station with the VOR is similar to that used for tuning an ADF. The pilot locates the station on an aeronautical chart or other reference and determines the frequency.

The pilot then tunes in the desired frequency and listens for the Morse code identifier. Both VORs and ADFs transmit a two or three letter identifier in Morse code. When the station has been identified, the pilot is ready to use it for navigation by means of the appropriate cockpit indicator instrument.

The latest types of VOR indicators use light bars instead of a needle for the CDI. A VOR indicator that uses light bars can be seen in figure 3-29 on the left side. On aircraft equipped with an HSI or EHSI, the VOR steering commands are displayed on the HSI instead of the more simple VOR indicator. The aircraft VOR equipment must be tested for accuracy if it is used for IFR (this is covered in the FAR section later on).

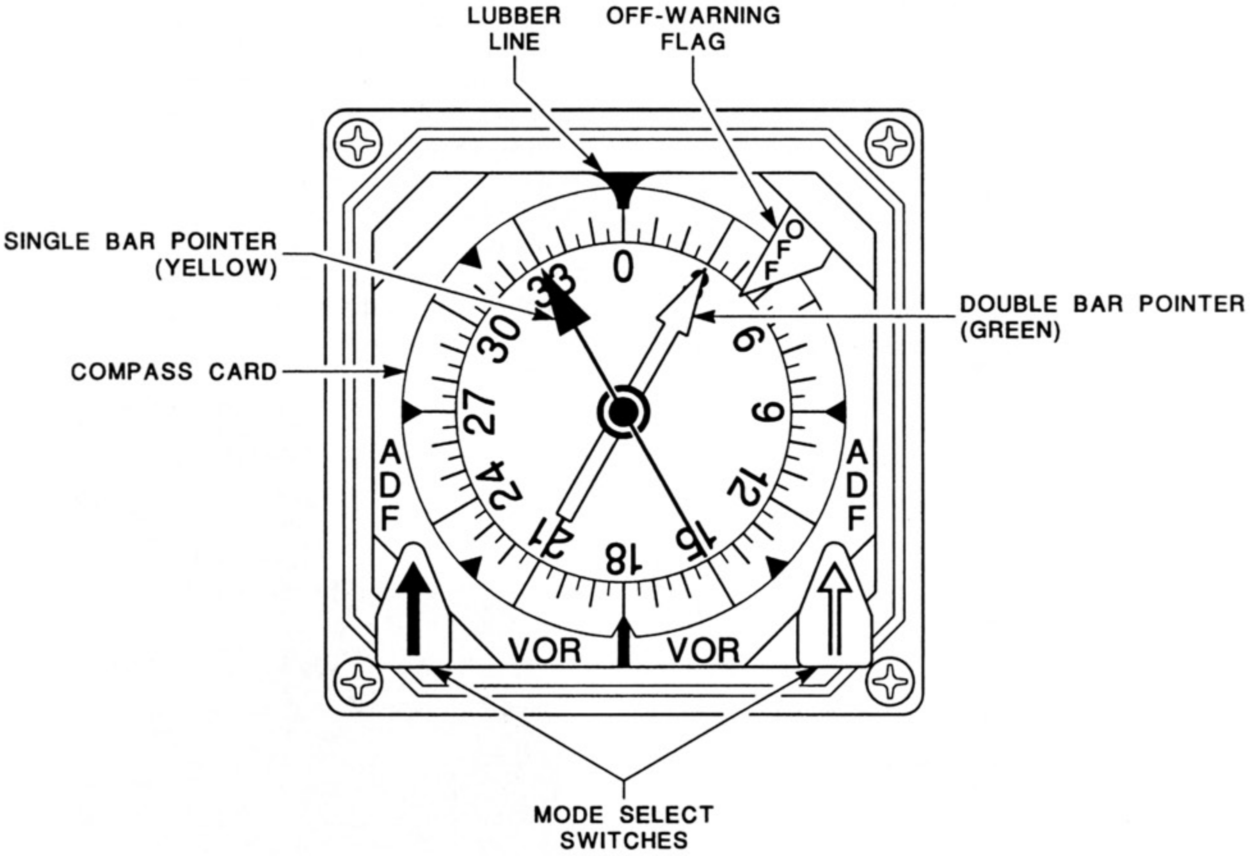

Figure 3-38. The RMI can display either ADF or VOR radio navigational information. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

4. Distance Measuring Equipment (DME)

The military services have their own radio navigation system which operates on principles similar to those of VOR. The system is known as TACAN (Tactical Air Navigation) and it uses signals in the UHF band. An additional feature of TACAN that is not a part of the VOR system is the use of distance measuring equipment as an integral part of TACAN. The DME portion of TACAN is used by civilian aircraft to augment the information available from the VOR. The VOR and TACAN transmitters are usually located at the same ground sites and referred to as VORTACs. The frequencies utilized by DME are in the range of 960–1,215 MHz. The basic operation of DME is illustrated in figure 3-40. The airplane DME transmitter sends out a pulse signal in all directions. This is referred to as the interrogation. When a DME ground station receives a valid interrogation from an aircraft, it sends back a reply after a fixed delay of 50 microseconds. The aircraft DME equipment measures the travel time for the signals to be sent and received back, and calculates the distance in nautical miles. The distance information is displayed on an indicator in the cockpit for the pilots. The distance measurement given by DME is a slant range distance so some error will result from the altitude of the aircraft. The amount of difference between slant range distance and horizontal or map distance is normally small and the error can be ignored. If the aircraft is at a high altitude and almost directly over the DME ground site, the error will be at its greatest. For example, if the aircraft is directly over the DME site and 18,000 ft. above it, the DME will indicate 3 nm (1 nm = 6,080 ft.). With the use of microprocessors, a modern DME can give other indications in addition to distance. If the DME distance is known, then groundspeed and time to station can be found through mathematical calculations. The pilot can select which readout is needed; distance in nautical miles, groundspeed in knots or time-to-station in minutes. Due to the fact that most DME ground sites are located in the same place as a VOR, the two radios are tuned at the same time. When the pilot selects the proper frequency for the VOR that is being used, the DME equipment is tuned automatically to the proper DME channel.

5. RNAV — Area Navigation

The RNAV equipment in aircraft contains a computer that processes the signals received from VOR and DME ground sites. The main advantage of RNAV is that it permits random direct routes of flight. The use of conventional VOR navigation along airways requires that the aircraft be flown directly from one VOR site to the next. Since the VOR sites seldom line up directly along the desired flight path, the aircraft ends up flying a zigzag course to get from one place to another over long distances. The use of RNAV equipment permits the aircraft to fly directly to the destination without having to fly straight to and from each of the VOR sites. This more direct routing is illustrated in figure 3-41. The RNAV computer processes signals from VOR and DME transmitters and displays steering information to the pilot to guide the aircraft along a direct route of flight.

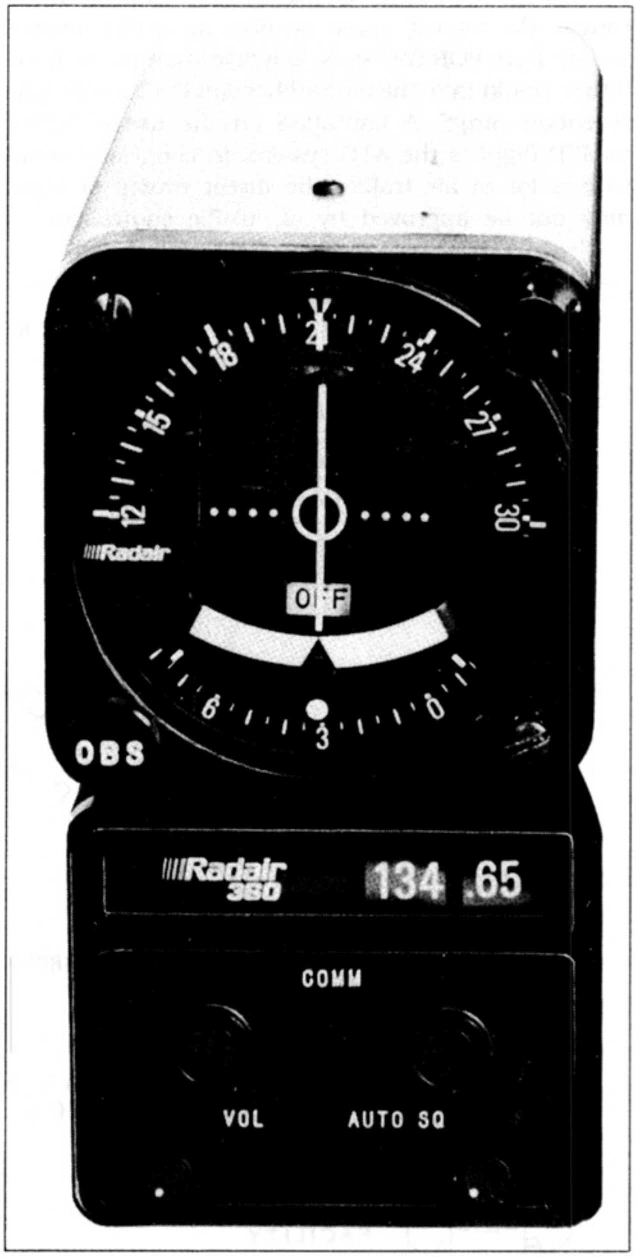

Figure 3-39. The VOR indicator contains the CDI needle, the TO-FROM indicator and the OBS knob to select the desired radial.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Before takeoff the pilot will program the computer with the desired waypoints that establish the desired route of flight. A waypoint is established as a direction and distance from a VOR and DME site. For example, the waypoint OMN 240/25 would indicate a point that is 25 nautical miles southwest of the OMN transmitter site. The pilot programs the RNAV computer by designating a number of waypoints along the desired flight path. During flight, the RNAV computer performs the calculations needed to display guidance commands using a CDI or HSI that will guide the aircraft from one waypoint to the next. Even though the RNAV equipment is designed to permit direct routes, the aircraft must be able to receive usable signals from VORTAC sites. Waypoints cannot be used if they would take the aircraft beyond the line-of-sight reception range. A limitation on the use of RNAV for IFR flight is the ATC system. In congested areas with a lot of air traffic, the direct routes of flight may not be approved by air traffic controllers.

6. Transponders

The transponder equipment found on aircraft is designed to make it easier for air traffic controllers to identify specific aircraft so that they can prevent mid-air collisions and provide guidance to the aircraft. The transponder is a device which is related to radar, so we will begin with a short history of the use of radar to identify aircraft. The use of radar to locate aircraft in flight dates back to the 1930s. The principle used is called primary radar or echo location radar. The radar transmitter sends out a brief pulse of EM waves which travel outward at the speed of light and bounce off the metal parts of an airplane. The reflected energy or echo is received back at the radar site where it produces a spot of light on the radar scope. The problem with this primary radar is that all the blips on the radar scope look the same. During World War II, a system was developed to make it easier to distinguish the friendly aircraft from

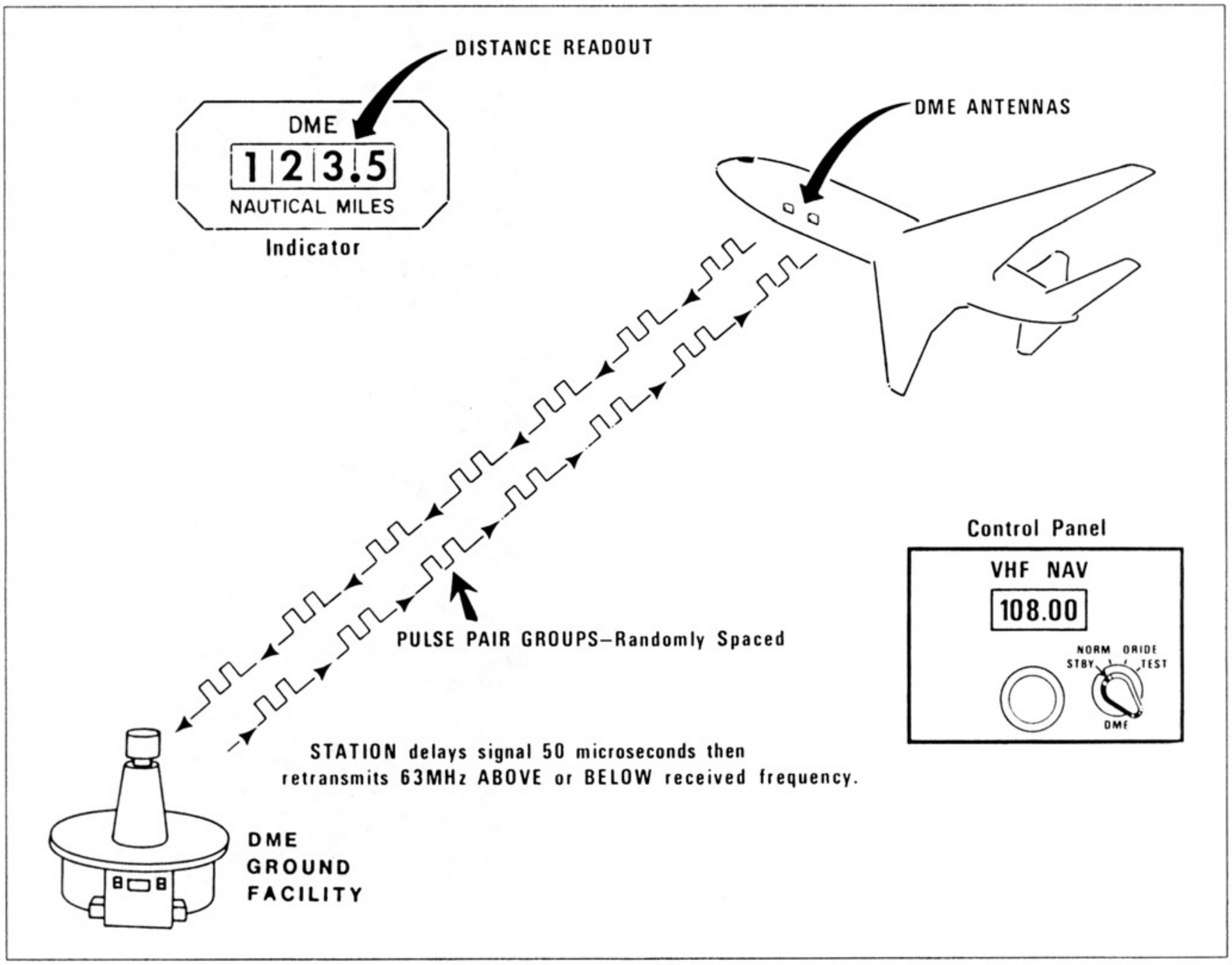
DISTANCE READOUT
DME
1 2 3.5
NAUTICAL MILES
Indicator
DME ANTENNAS
PULSE PAIR GROUPS—Randomly Spaced
Control Panel
VHF NAV
108.00
NORM
ORIDE
STBY
TEST
DME
STATION delays signal 50 microseconds then
retransmits 63MHz ABOVE or BELOW received frequency.
DME
GROUND
FACILITY

Figure 3-40. The operation of the DME radio system using pulse signals.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

the enemy aircraft. The system was known as IFF (identification friend or foe) and it is still called that by the military. This type of radar is also called secondary radar. A small radar frequency receiver and transmitter unit is installed in each airplane. When the radar pulse from the ground site strikes the aircraft, the IFF equipment sends a coded signal back to the ground site. The basic operation of the primary and secondary radar systems is shown in figure 3-42. The coded signal received at the radar site from the aircraft permits it to be identified. In the years since World War II, both primary and secondary radar have been adapted for ATC purposes. Transponder is the name of the secondary radar equipment installed on aircraft.

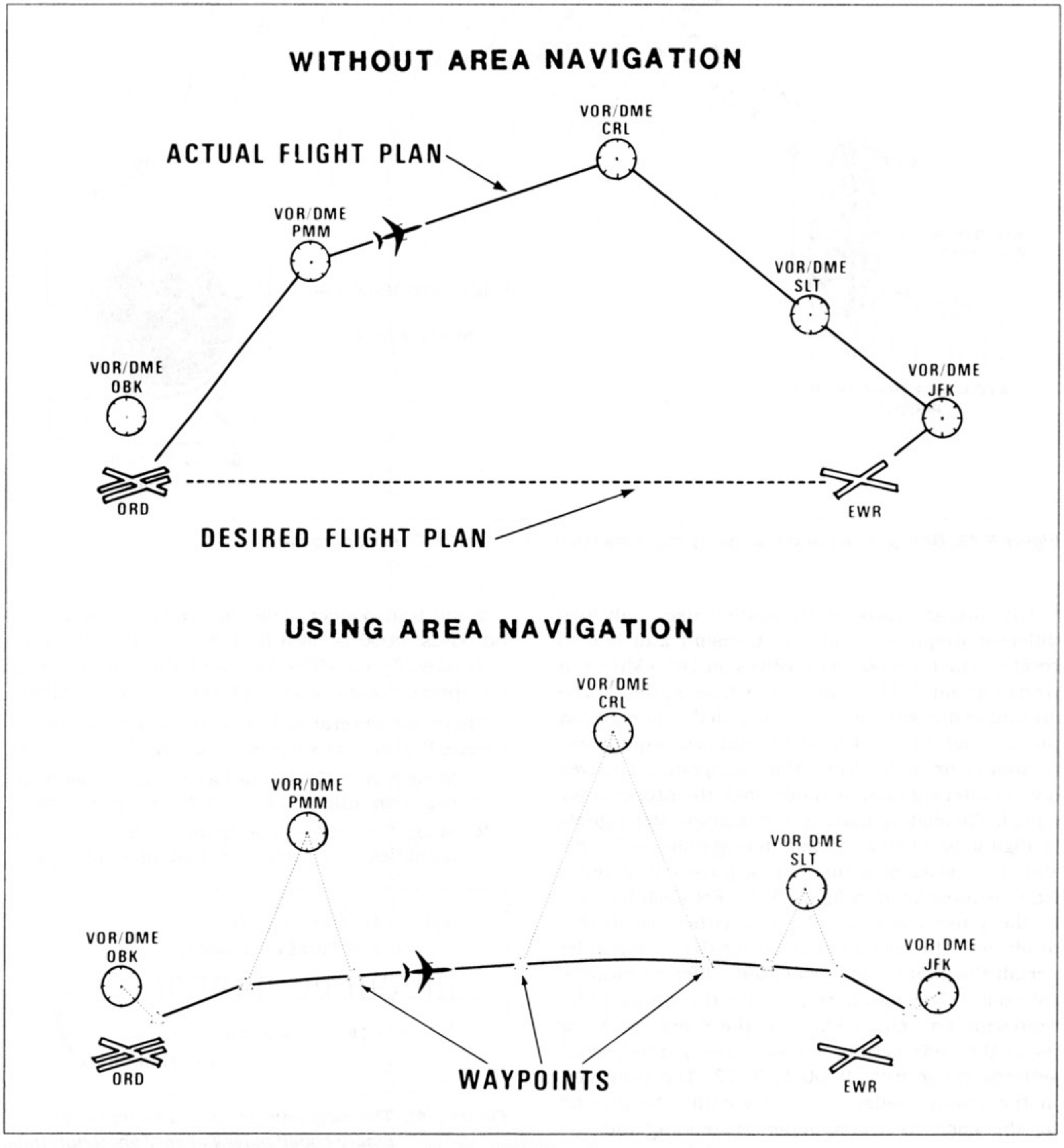
WITHOUT AREA NAVIGATION
ACTUAL FLIGHT PLAN
VOR/DME CRL
VOR/DME PMM
VOR/DME SLT
VOR/DME OBK
VOR/DME JFK
ORD
EWR
DESIRED FLIGHT PLAN
USING AREA NAVIGATION
VOR/DME CRL
VOR/DME PMM
VOR DME SLT
VOR/DME OBK
VOR DME JFK
ORD
WAYPOINTS
EWR

Figure 3-41. The use of RNAV equipment permits direct flights using way points.

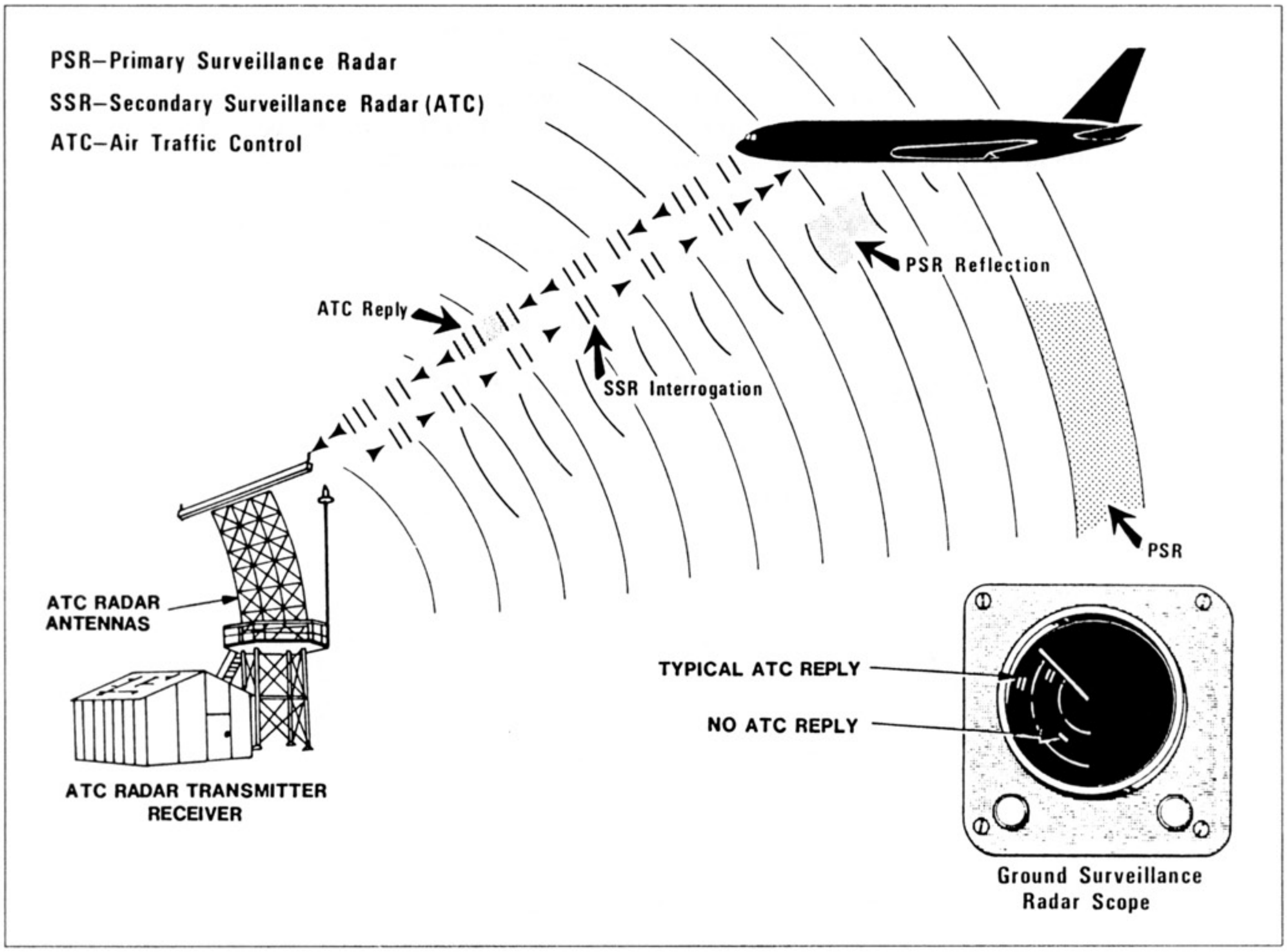
PSR–Primary Surveillance Radar
SSR–Secondary Surveillance Radar (ATC)
ATC–Air Traffic Control
PSR Reflection
ATC Reply
SSR Interrogation
PSR
ATC RADAR ANTENNAS
TYPICAL ATC REPLY
NO ATC REPLY
ATC RADAR TRANSMITTER RECEIVER
Ground Surveillance Radar Scope

Figure 3-42. Both primary and secondary radar are used for Air Traffic Control purposes.

The aircraft transponder system uses only two different frequencies, one to transmit and one to receive. The transponder receives on 1030 MHz and transmits on 1090 MHz (it is a UHF system). The ground radar site sends out a coded interrogation pulse which in effect asks the airborne equipment to answer or reply. When the transponder receives a valid interrogation, it sends back the proper reply signal. The coding used in the transponder signals is digital or binary. Each interrogation and reply signal consists of a number of pulses in a pulse train as illustrated in figure 3-43. For each location in the pulse train, a pulse can either be present or absent. The cockpit controls for the transponder permit the pilot to set one of 4096 different numerical codes. The numbers set into the transponder represent an octal coding so there are no 8s or 9s in the code setting window. The possible code settings range from 0000 to 7777. The computer in the ground radar site can identify the aircraft by the code its transponder is sending out.

Some transponder codes are reserved for special purposes; 0000 is used by the military, 1200 is for VFR aircraft and 7500,7600 and 7700 are reserved for specific types of aircraft emergency situations.

There are several different operating modes associated with transponder equipment. They are:

1. Mode 3/A — This is the basic transponder mode that can utilize one of 4096 different codes.
2. Mode C — This mode includes the above capabilities but adds a coded message giving

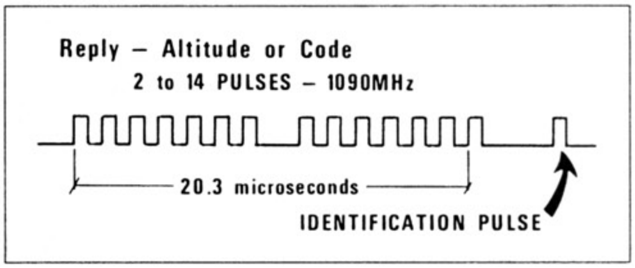
Reply – Altitude or Code
2 to 14 PULSES – 1090MHz
20.3 microseconds
IDENTIFICATION PULSE

Figure 3-43. The reply from an aircraft transponder uses binary coded pulses of very short duration.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

the aircraft's pressure altitude using an altitude encoder.

3. Mode S — This is the latest development in transponders and is not yet fully operational, it will have the capability of sending additional messages such as ATC instructions or weather reports that can be viewed on a CRT or printed on paper in the cockpit. Mode S also increases the number of different identification codes for the aircraft to over one million.

7. FARs for VOR and Transponder

The FAA regulations which apply to testing and operation of VOR and transponder equipment are summarized here.

FAR 91.171

VOR Equipment Check for IFR Operations

(a) No person may operate a civil aircraft under IFR using the VOR system of radio navigation unless the VOR equipment of that aircraft:
  1. Is maintained, checked and inspected under an approved procedure; or
  2. Has been operationally checked within the preceding 30 days and was found to be within the limits for bearing error set forth below.

(b) The check must use one of the following:
  1. An approved FAA or Repair station ground test signal — ±4°.
  2. Designated VOR checkpoint on the airport surface — ±4°.
  3. Designated airborne checkpoint — ±6°.
  4. An airborne check using a VOR radial and prominent ground point that can be seen from the air as established by the person doing the check — ±6°.
  5. If two separate VOR receivers are installed, they can be checked against each other — ±4°.

(c) Maintenance record entry
  1. Each person performing one of the above checks shall enter the date, place and bearing error in the aircraft log or other record and sign it.
  2. If a test signal from a repair station is used, the repair station certificate holder must enter the bearing transmitted and date in the aircraft log or other record.

FAR 91.215

ATC Transponder and Altitude Reporting Equipment and Use

(a) TSO requirements
  1. All aircraft transponders must meet the appropriate requirements of TSO-C74b, TSO-C74c or TSO-C112.
  2. After July 1, 1991, all initial installations of ATC transponders in aircraft must meet the requirements of TSO-C112 (Mode S).

Note: Due to development delays this requirement has been dropped.

(b) Airspace requirements — all aircraft operating in the following airspace must have a 4096 code transponder with Mode C altitude reporting or a Mode S transponder.
  1. Terminal Control Areas — TCAs
  2. Airport Radar Service Areas — ARSAs
  3. In all controlled airspace above 10,000 ft. MSL and over 2,500 AGL.

FAR 91.413

ATC Transponder Tests and Inspections

(a) No person may use an ATC transponder that is required by the rules of this Chapter unless it has been tested and inspected within the preceding 24 calendar months and meets the requirement of Part 43 Appendix F.

(b) Following any installation or maintenance which could have introduced errors, the integrated system must be tested in accordance with paragraph (c) of Appendix E, Part 43.

(c) The above tests and inspections must be conducted by:
  1. An appropriately rated repair station; or
  2. A holder of a continuous airworthiness program as provided in Part 121, 127, or 135; or
  3. The manufacturer of the aircraft, if the transponder was installed by that manufacturer.

H. Long Range Navigation Systems

The navigation systems in this section all have usable ranges that exceed those of the VOR, ADF and other systems already covered. Some of these long range navigation systems do not rely on ground based transmitters and some do not use radio signals at all. All of them have been developed since the 1960s so they are relatively more modern than most of the systems previously described. The long range navigation systems use geographical coordinates to establish aircraft position and waypoints along a desired flight path. As shown in figure 3-44, geographical coordinates are based on the grid system of lines of latitude and longitude. Longitude

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

is expressed either east or west of the prime meridian that runs through Greenwich, England. The maximum longitude is 180° which is the International Date Line exactly half way around the earth from the prime meridian. Latitude is measured north and south of the equator with a maximum value of 90° at the poles. Since 1° is equal to approximately 60 nm, for accurate positions each degree is divided into 60 minutes and each minute into 60 seconds. An example of a lat/long position is 29°, 10 minutes, 51 seconds North latitude by 81°, 3 minutes, 22 seconds West longitude. This coordinate position locates the airport at Daytona Beach, Florida.

1. Loran C

The first Loran system was developed in the 1940s by the U.S. Navy for use by ships. Modifications were made to produce Loran A, B, C and D. Loran C is the only one that has any large scale use by aircraft. All the early transmitter sites were located along coastlines since it was a system for ships. Starting about 30 years ago, pilots of privately owned airplanes began modifying Loran C units from boats and ships for aircraft use. The system has been improved and newer equipment is now available that make Loran C a very useful radio navigation system. Within the last several years new ground sites have been installed

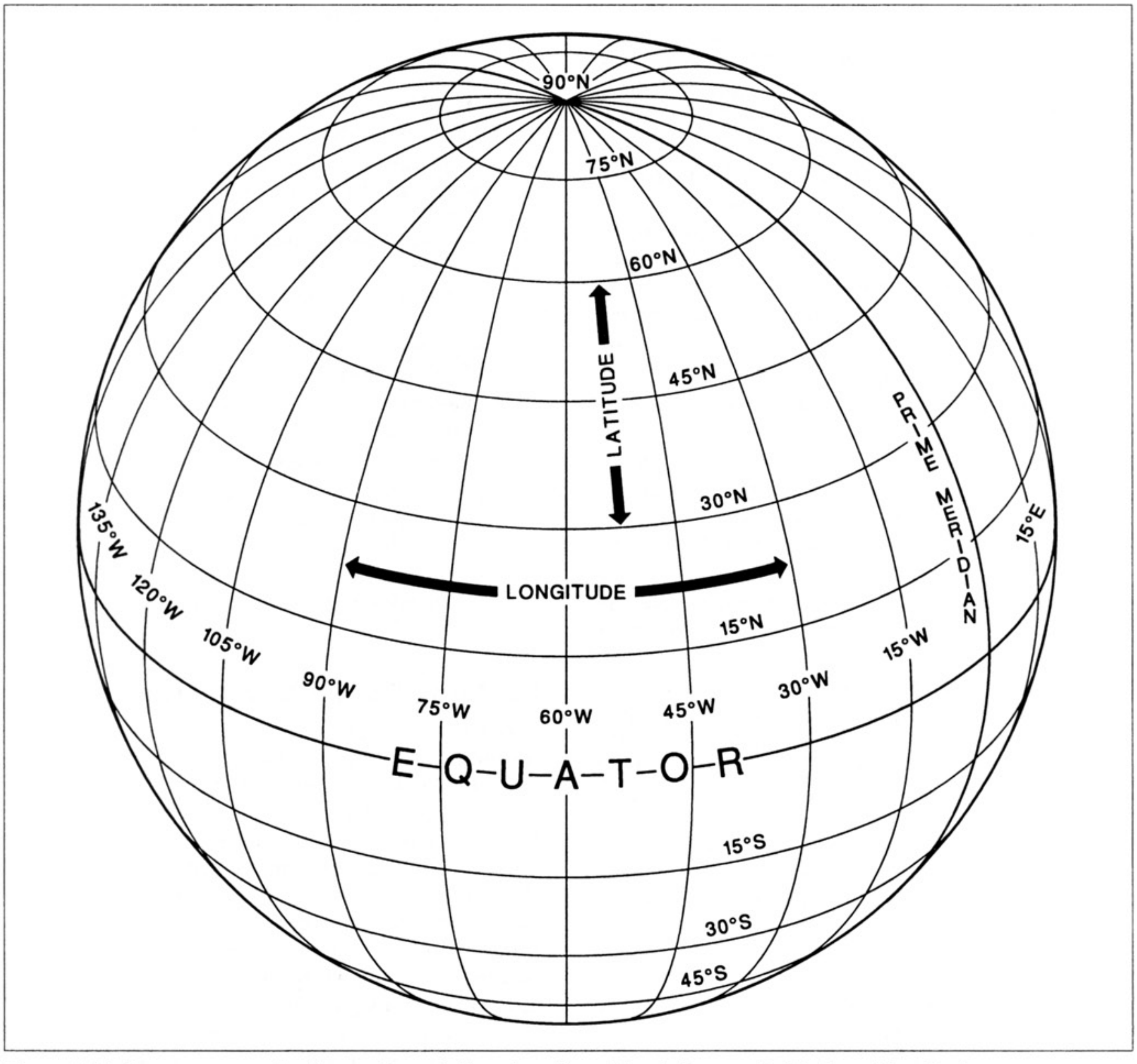

Figure 3-44. Long range navigation systems define aircraft position in terms of latitude and longitude.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

in the western U.S. to give nationwide coverage. The primary aircraft users of Loran C continue to be general aviation aircraft.

The Loran C system uses ground transmitter sites that all transmit on the same frequency—100 kHz. The EM waves produced are ground waves. The operation of this system involves transmitter chains in which each chain consists of one master station and two or more slave stations. For purposes of explanation, we will consider a chain with two slave stations like that shown in figure 3-45. Each transmitter uses a tower about 1,000 ft. tall and has an output power of approximately 4,000,000 watts.

The transmissions are sequenced so that the master transmits first and then the slaves transmit. The location of the master and slaves are hundreds of miles apart. The signals received by the aircraft will have a time separation that is determined by the aircraft location relative to the transmitter sites. A computer in the Loran C receiver performs the calculations that determine location. The location determination has an accuracy on the order of 400–1,000 ft. in most cases. The Loran C equipment in the aircraft does not have to be tuned since all signals are received on 100 kHz. Each chain can be identified by the time delay between transmission pulses and the information is stored in the computer memory of the aircraft receiver. The reception range is 1,000 miles or more. An advantage of this system is that signals can be received at any altitude, even with the aircraft on the ground. The use of microprocessors has made the modern Loran C unit a very powerful and versatile navigation

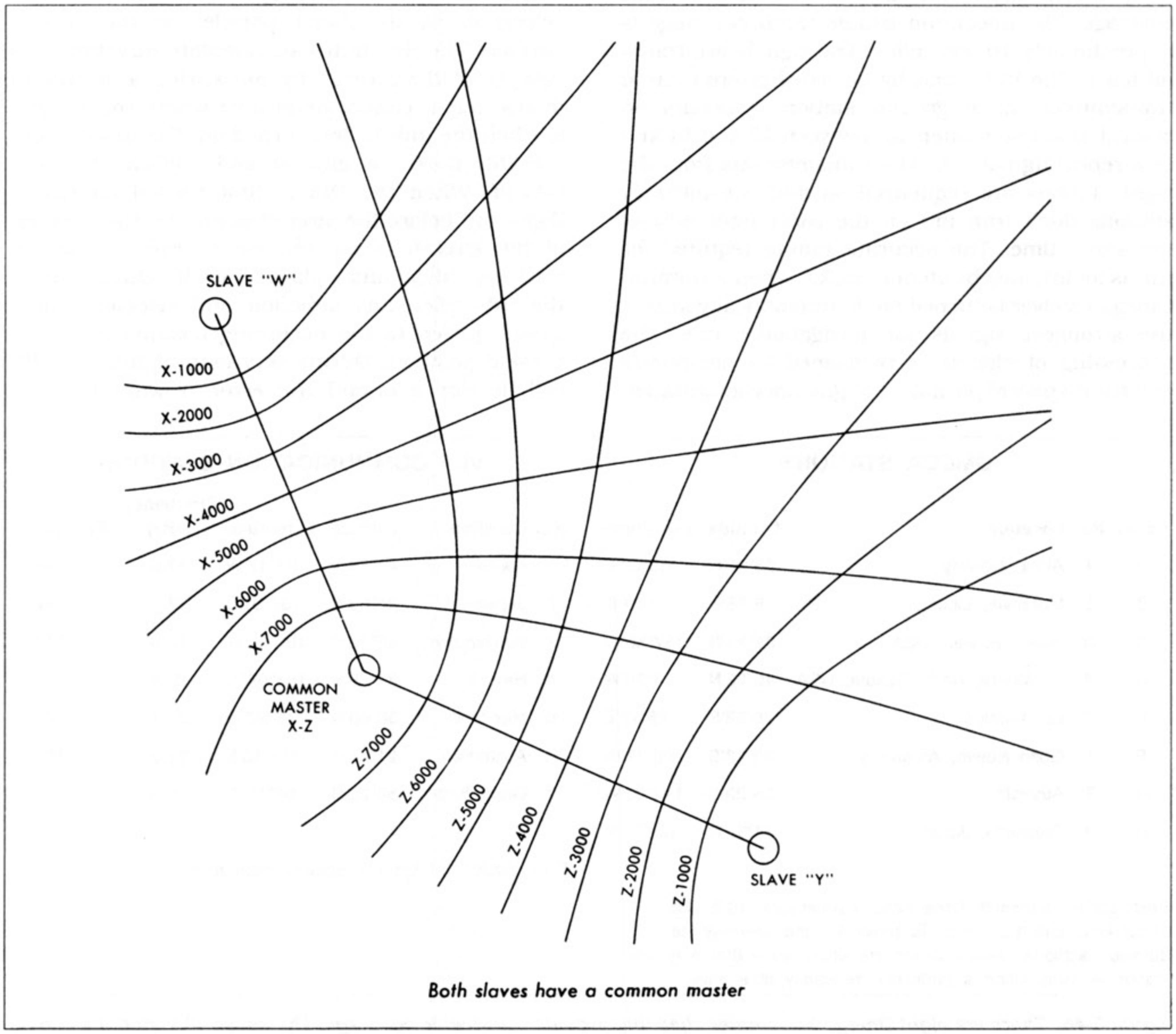

Figure 3-45. A Loran C chain consists of one master station and two or more slave stations.

system. A Loran C receiver contains a large amount of memory which stores the location of every airport and VOR in the U.S. The memory often includes extra information for each airport such as radio frequencies available, runway direction and length, fuel available etc. The memory can also store pilot designated waypoints to enable direct routes to be flown. The output of the Loran C includes a CDI type indicator which can be used to guide the aircraft along the desired route of flight. A Loran C receiver and its digital lighted display can be seen at the bottom of the radio stack in figure 3-29.

2. Omega

The Omega radio navigation system was developed by the U.S. Navy for use by ships and aircraft. There are only eight Omega transmitter sites scattered around the earth, but they provide worldwide coverage. The maximum usable reception range is approximately 10,000 miles. The signals are transmitted in the VLF band by powerful ground based transmitters. Each ground station transmits on several different frequencies between 10 and 14 kHz in a repeating pattern. The transmissions from the eight stations are sequenced so that two different stations don't transmit on the same frequency at the same time. The accurate timing required for this is maintained by atomic clocks. When an aircraft Omega receiver is turned on, it automatically selects the strongest signals for navigational use. The processing of signals is performed by computers and the display of position and guidance information is similar to that used with Loran C. The U.S. Navy also operates a VLF communications system that utilizes seven sites around the world. Many aircraft navigation receivers can pick up both Omega and VLF comm signals. The Omega signals are more accurate for navigation and are used as the primary source, the VLF communications station signals are used as a back-up or secondary means of navigation. The locations for the eight Omega stations and the seven VLF stations are shown in figure 3-46.

3. Inertial Navigation System — INS

The inertial navigation system or INS is a long range NAV system that does not rely on the reception of radio waves. The system is totally self-contained within the aircraft. The key to the operation of INS is the very accurate measurement of acceleration forces. The accelerometer sensors measure acceleration in directions parallel to the earth's surface. The INS unit can calculate direction and velocity of the aircraft by measuring acceleration forces, but it cannot determine where the aircraft is when the unit is first turned on. For this reason, the INS must be aligned and calibrated before takeoff. When the INS is first turned on before flight and before the aircraft is moved, the position of the aircraft is entered on a keyboard so the unit can align and calibrate itself. During flight the INS calculates direction and velocity, which when applied to the beginning position gives the present position. During very long flights, the INS will develop a cumulative error. Toward the end

OMEGA STATIONS

Letter	No.	Location	Latitude	Longitude
A	1	Aldra, Norway	66°25'N	13°08'E
B	2	Monrovia, Liberia	6°18'N	10°40'W
C	3	Haiku, Hawaii, USA	21°24'N	157°50'W
D	4	La Moure, North Dakota, USA	46°22'N	98°20'W
E	5	La Reunion	20°58'S	55°17'E
F	6	Golfo Nuevo, Argentina	43°03'S	65°11'W
G	7	Australia	38°29'S	146°56'W
H	8	Tsushima, Japan	34°37'N	129°27'E

Each station transmits three basic frequencies: 10.2 kHz, 11.33 kHz, and 13.6 kHz. To prevent signal interference between stations, transmisisons are timed such that only one station is transmitting a particular frequency at a time.

VLF COMMUNICATION STATIONS

No.	Location	Latitude	Longitude	Frequency (kHz)	Pwr (KW)
1	Maine	44°39'N	67°17'W	17.8	1026
2	Japan	34°58'N	137°01'E	17.4	48
3	Washington	48°12'N	121°55'W	18.6	124
4	Hawaii	21°26'N	158°09'W	23.4	588
5	Maryland	38°60'N	76°27'W	21.4	588
6	Australia	21°49'S	114°10'E	22.3	989
7	Great Britain	52°22'N	01°11'W	16.0	40

Each station transmits a specific frequency.

Figure 3-46. There are eight Omega transmitter sites that provide world-wide coverage. The seven VLF stations can be used as back-ups.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

of a long flight the error might be 20 miles. To eliminate this cumulative error, modern INS units can be updated using radio signals received from ground stations where they are available. The simple type of INS sensors are small weights which react to acceleration forces by movement about a hinge point as shown in figure 3-47. In order to give accurate readings, the accelerometers must be mounted on a gyro stabilized platform so that they only measure horizontal forces. This stable platform and the control panel for an INS system is shown in figure 3-48(A) and figure 3-48(B). In effect, the accelerometers measure north-south and east-west accelerations in order to determine aircraft position.

The newest type of inertial navigation system does not use conventional spinning gyroscopes and does not need a gyro stabilized platform. A modern Inertial Reference Unit (IRU) for an INS system contains three accelerometers and three ring laser gyros (RLGs). The three accelerometers measure acceleration forces along the aircraft's three axes; vertical, lateral and longitudinal. The RLG is a device with no moving parts that replaces a conventional gyro with a spinning rotor. As shown in figure 3-49, the laser gyro uses a triangular housing and two different laser beams. The mirrors at the corners direct the two laser beams in opposite directions around the triangular course. Sensitive detectors measure the Doppler frequency shift that occurs when the unit is rotated. Three of these RLGs are needed to measure rotation around the three axes of the aircraft. A computer processes the signals from the three accelerometers and the three laser gyro sensors to determine aircraft heading, position and groundspeed.

This modern type of IRU is referred to as a strapdown system because it does not require a gyro stabilized platform like that shown in figure 3-48(A). The corrections that are needed are calculated by the computer. Like any inertial navigation system, the strapdown INS must be given the geographical coordinates for present position during the alignment before takeoff.

Inertial navigation systems can be programmed with complete routes of flight and can be coupled to the aircraft autopilot to provide steering commands.

4. Doppler Navigation

The Doppler navigation system does not rely on the reception of radio signals from ground based transmitter sites, but it does use radio waves.

The Doppler system uses radar beams that are projected downward and received back at the aircraft after they have bounced off the surface of the earth. The frequency commonly used is 8.8 GHz. The arrangement of the radar beams is shown in figure 3-50. Notice that two beams are projected forward and two to the rear. When the aircraft is in motion, the frequency of the received signals will be shifted upward or downward compared to the frequency that was transmitted. The change in frequency of a wave when there is relative motion between the source of the wave and the observer is the Doppler effect. Figure 3-51 illustrates the Doppler effect with sound waves where the observer hears a change in pitch or frequency of the sound waves as the ambulance goes by his position. The Doppler effect is the same for both sound waves and radio waves. That is where this system gets its name. If the aircraft is traveling forward over the surface of the earth and not drifting right or left, the frequency of the two forward beams is shifted upward equally and the frequency of the two rearward beams is shifted downward equally. If the aircraft is drifting, there will be a difference in the received frequency between the right side and left side beams. By measuring all four beams, the groundspeed and side drift of the aircraft can be calculated accurately. This information can be used to make a continuous determination of position. Like INS, Doppler NAV systems must be aligned before takeoff and will develop cumulative errors in flight. Doppler can be updated using available ground based radio signals as is done with INS. Doppler navigation units were common in the past, but have now been largely replaced by INS and other newer long range navigation systems.

5. Satellite Navigation

The latest development in long range navigation systems is the use of satellites in earth orbit. Two

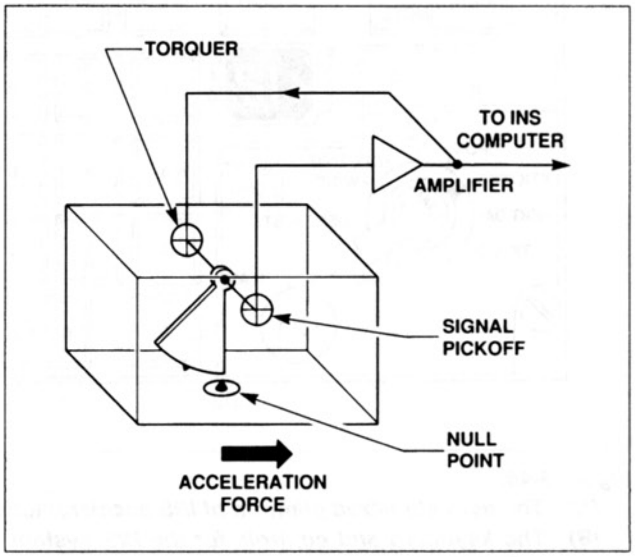

Figure 3-47. One simple type of INS accelerometer.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

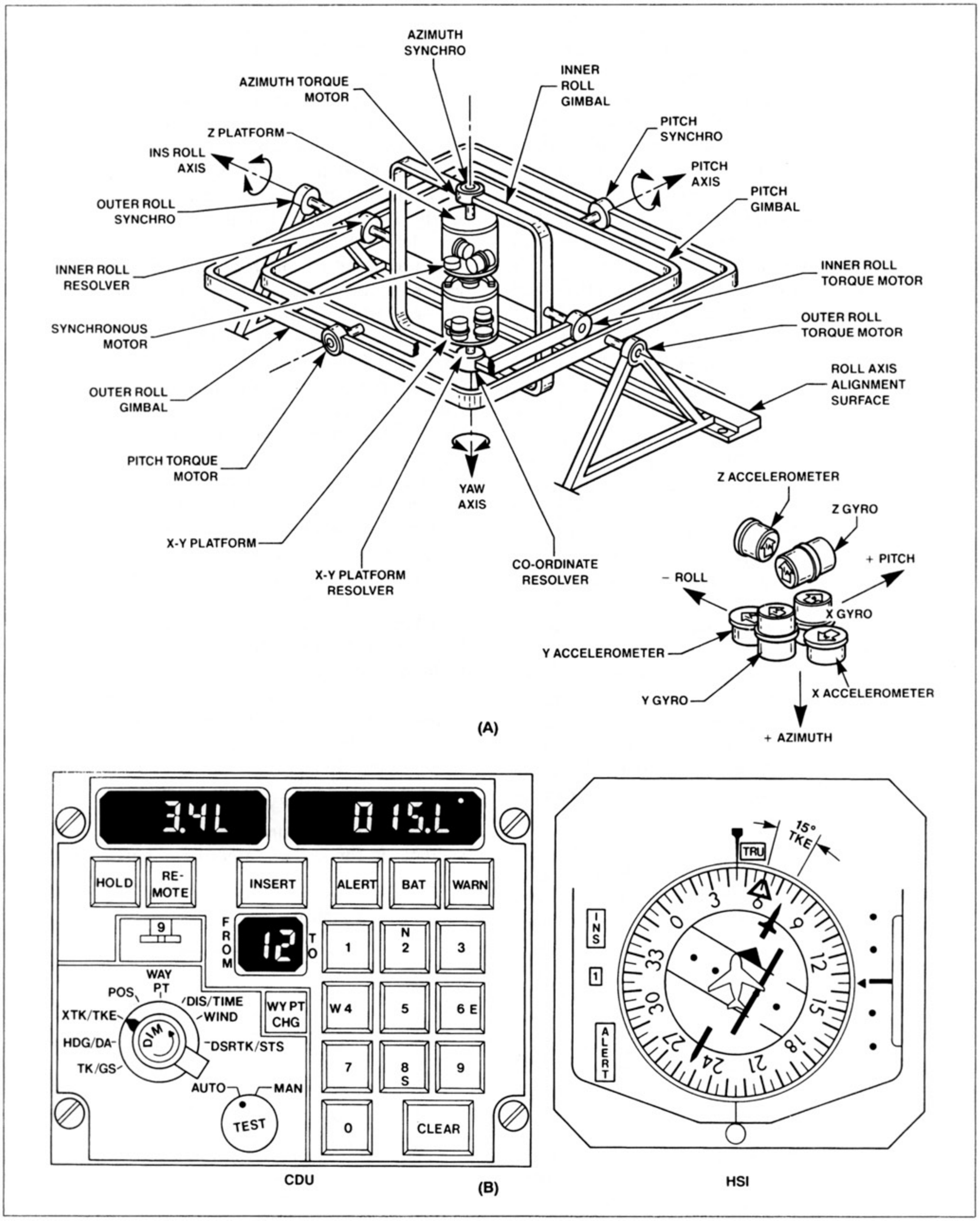
AZIMUTH SYNCHRO
AZIMUTH TORQUE MOTOR
INNER ROLL GIMBAL
Z PLATFORM
PITCH SYNCHRO
INS ROLL AXIS
PITCH AXIS
PITCH GIMBAL
OUTER ROLL SYNCHRO
INNER ROLL RESOLVER
INNER ROLL TORQUE MOTOR
OUTER ROLL TORQUE MOTOR
SYNCHRONOUS MOTOR
ROLL AXIS ALIGNMENT SURFACE
OUTER ROLL GIMBAL
PITCH TORQUE MOTOR
YAW AXIS
Z ACCELEROMETER
Z GYRO
X-Y PLATFORM
X-Y PLATFORM RESOLVER
CO-ORDINATE RESOLVER
– ROLL
+ PITCH
X GYRO
Y ACCELEROMETER
Y GYRO
X ACCELEROMETER
+ AZIMUTH
(A)
3.4L
015.L
HOLD
RE-MOTE
INSERT
ALERT
BAT
WARN
9
FROM
12
TO
1
N 2
3
WAY PT
POS
/DIS/TIME
XTK/TKE
WIND
HDG/DA
DSRTK/STS
TK/GS
DIM
WY PT CHG
W 4
5
6 E
7
8 S
9
AUTO
MAN
TEST
0
CLEAR
CDU
15° TKE
TRU
INS
1
ALERT
0
3
6
9
12
15
18
21
24
27
30
33
HSI
(B)

Figure 3-48.

(A) The gyro stabilized platform of INS accelerometers.

(B) The keyboard and controls for the INS system and the HSI which can be used to display INS navigational information.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

systems began development in the 1970s, one by the USSR and one by the U.S. The U.S. system is called GPS (global positioning system) and the Russian system is called GLONASS.

Both systems are still incomplete although some satellites are in orbit now and can be used. The two systems are very similar in terms of the frequencies and principles of operation. The GPS system was designed to include 24 satellites, 21 primary and 3 spares. Due to delays caused by the Space Shuttle Challenger disaster and other factors only 18 satellites were in place as of the middle of 1992. The GLONASS system was also designed to use a total of 24 satellites, but only about 15 were in place by the middle of 1992. The satellites of both systems orbit at an altitude of about 10,000 nm with a period of 11 to 12 hours. The position and altitude of the satellites is known with great precision. The aircraft with a satellite navigation system communicates with the satellites using frequencies in the 1.6 GHz range. For accurate navigation, the aircraft must be able to communicate with at least four different satellites as seen in figure 3-52. Currently the coverage in most areas is over 90%. Since the altitude of all the satellites is known, this system can provide altitude as well as position information. The accuracy is on the order of 80 ft. which makes the system potentially more accurate than the other systems so far described. The GPS system was developed and is operated by the Department of Defense. A possible limitation on the use of GPS for IFR is the difficulty in monitoring the accuracy of the satellite signals. Satellite navigation systems can be purchased now, but they are primarily used for VFR navigation or as a secondary system. The use of this system as a sole source for IFR navigation has not yet been approved. It has been predicted that within a few years satellite navigation may become the dominant long range NAV system and may also be used for precision approach guidance to airport runways.

I. Instrument Landing System (ILS)

An instrument approach procedure is a method used to guide an aircraft to an airport runway for

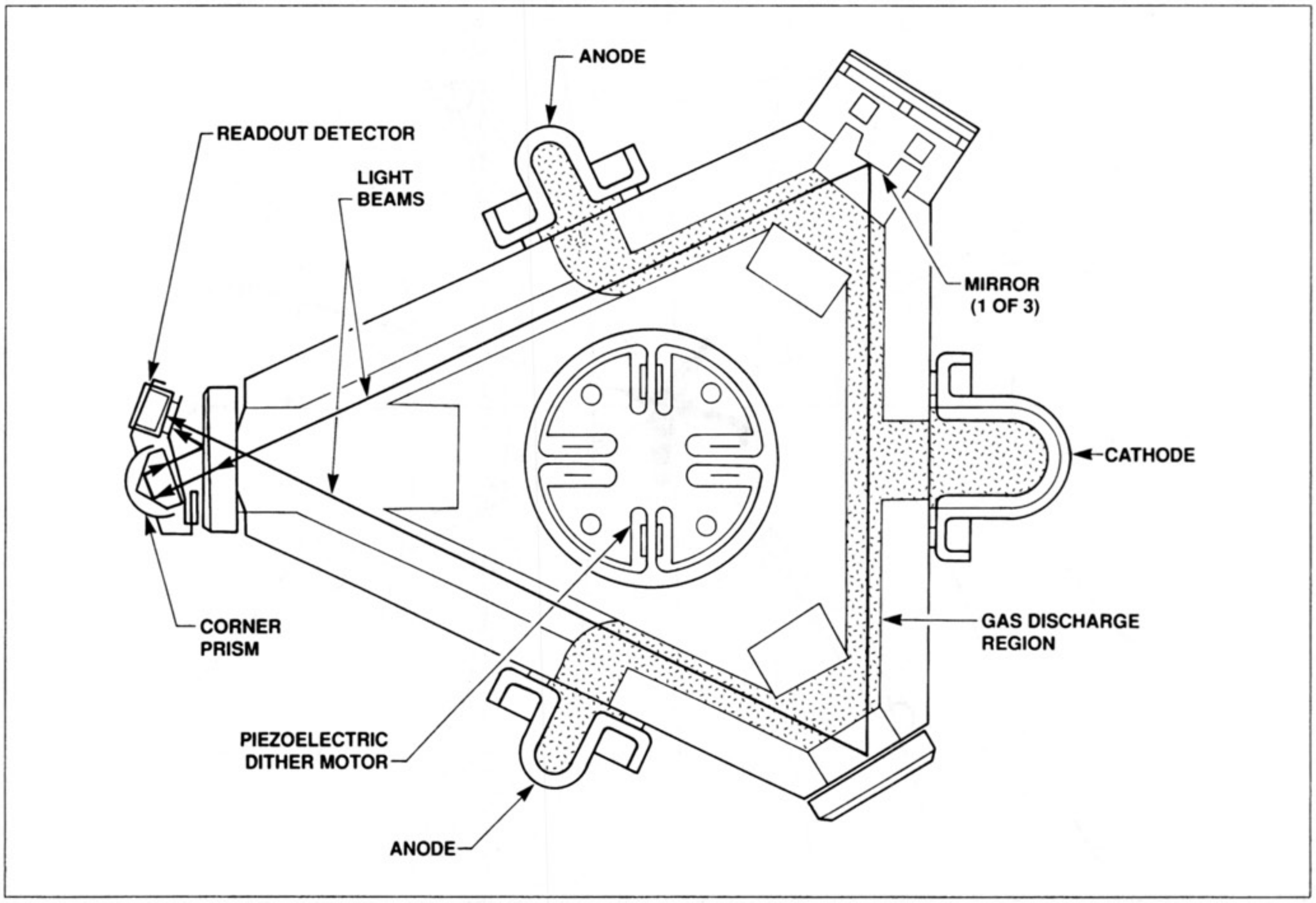

Figure 3-49. A laser beam IRU (inertial reference unit) which uses laser beams to replace conventional gyros for an INS navigation system.

landing in bad weather conditions. The procedures must be FAA-approved and are published in special books for pilots to use. There are two basic types of instrument approach procedures: precision approaches and non-precision approaches. The difference is that precision approaches give the pilot vertical or descent guidance while non-precision approaches do not. Signals from VORs and NDBs can be used for non-precision approaches. The standard type of precision approach system used in the U.S. and most of the world for civilian aircraft is the ILS. The military still uses a form of precision approach radar, but that kind of system is no longer operated by the FAA. Instrument approaches have weather minimums which specify the minimum ceiling (cloud height) and visibility needed to successfully complete the approach. The weather minimums for a Category I standard ILS are 200 ft. ceiling and ½ mile visibility. The minimums for a non-precision approach would be about twice as much.

The ground equipment needed for an ILS system has four parts.

1. Localizer — A radio beam for lateral guidance.
2. Glideslope — A radio beam for vertical guidance.
3. Marker Beacons — Radio signals that give distance to the runway data.
4. Runway and approach lights.

The lights will not be discussed further because they require no equipment on board the aircraft.

The first three parts will be described in order. The localizer layout is shown in figure 3-53 with a view from above to show the localizer beam.

The signals transmitted by the localizer are on frequencies between 108 and 112 MHz. A dual beam is transmitted outward from the far end of the runway. The right half of the signal is modulated at 150 Hz and the left half is modulated at 90 Hz. The aircraft receiver measures the relative strength of the 90 and

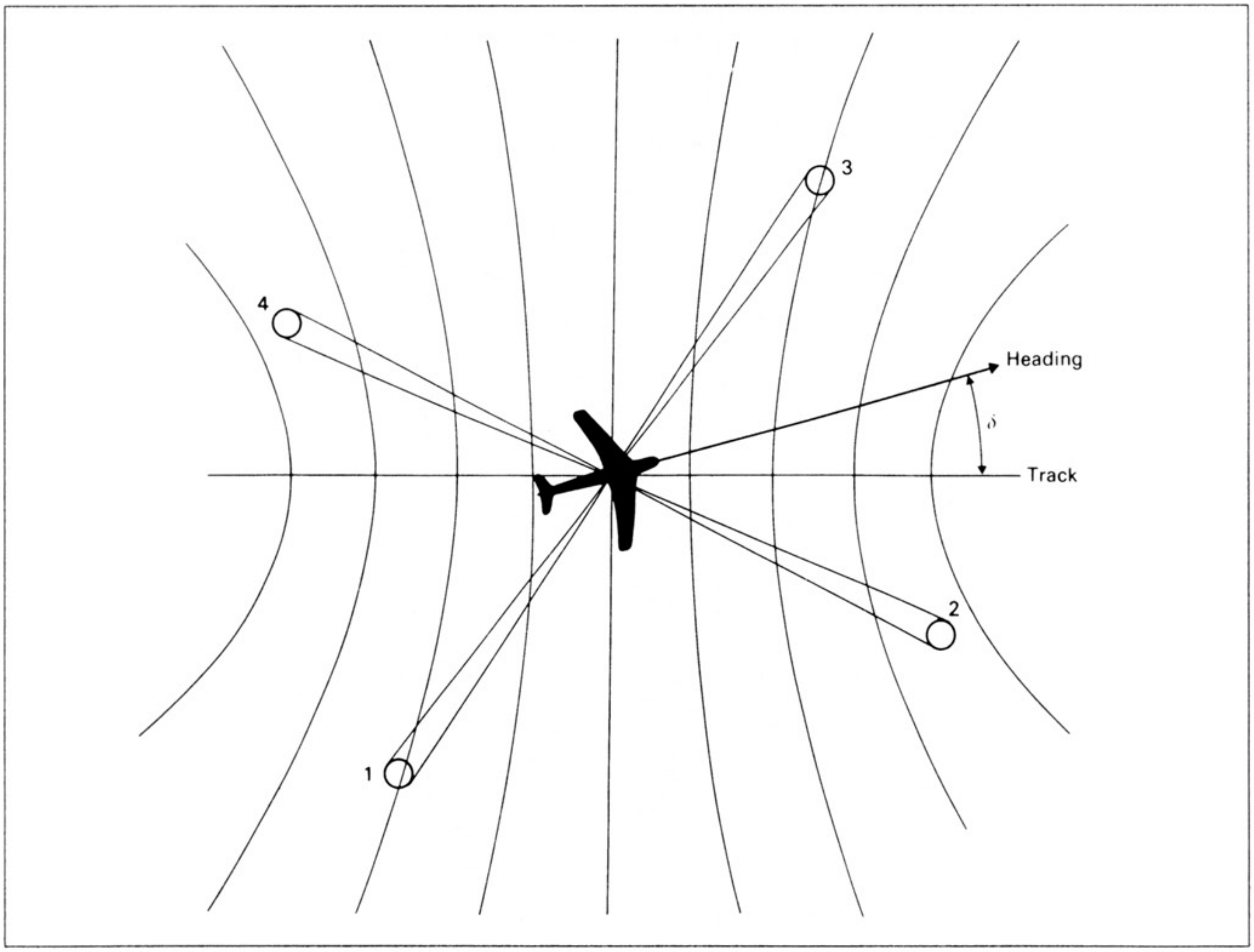

Figure 3-50. The orientation of the four radio beams projected from the bottom of the aircraft by a Doppler navigation system.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

150 Hz signals that are received. When they are equal, the aircraft is lined up with the centerline of the runway. The cockpit indicator is a vertical needle just like a VOR indicator, in fact it is often the same one. If the needle swings to the left, the pilot must turn left to get back on course.

The glideslope uses a principle like that of the localizer, but it transmits on frequencies of 328 to 336 MHz in the UHF band. A side view of the glideslope and runway is shown in figure 3-54.

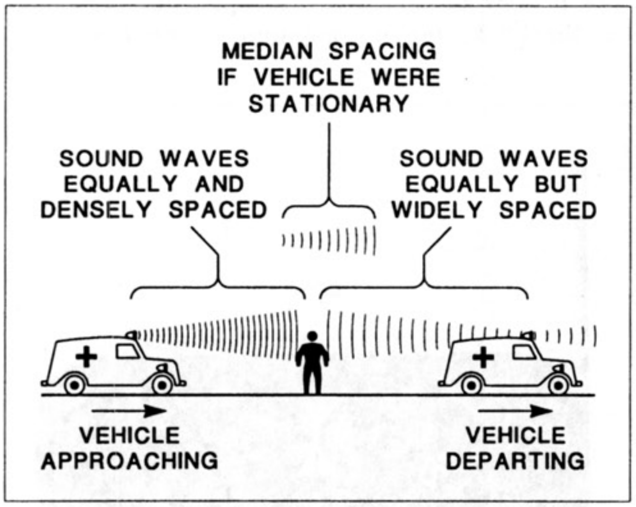

Figure 3-51. Example of the Doppler principle applied to sound waves rather than radio waves.

The glideslope signal uses 90 Hz modulation above the glidepath and 150 Hz modulation below the glidepath. The center of the glidepath would produce equal parts of 90 and 150 Hz signal in the receiver. The cockpit indicator for glideslope is a horizontal needle as shown in figure 3-55 which shows a simple ILS indicator. A needle deflected upward means the pilot must fly up or decrease his rate of descent. The actual glidepath angle used in an ILS system is about 2½–3°. This angle permits both large and small aircraft to use the ILS. The glideslope and localizer frequencies for an ILS are paired together in set combinations. The glideslope receiver is usually slaved to the localizer receiver so that when a localizer frequency is tuned in, the correct glideslope frequency is automatically set in the glideslope receiver.

The marker beacons are low powered transmitters that transmit a cone shaped pattern straight up into the air. When the aircraft flies directly over the marker beacon site, an indication is given in the cockpit to show the pilot the distance to the approach end of the runway. This is illustrated in figure 3-56 which gives the approximate distances for the outer, middle and inner markers. All marker

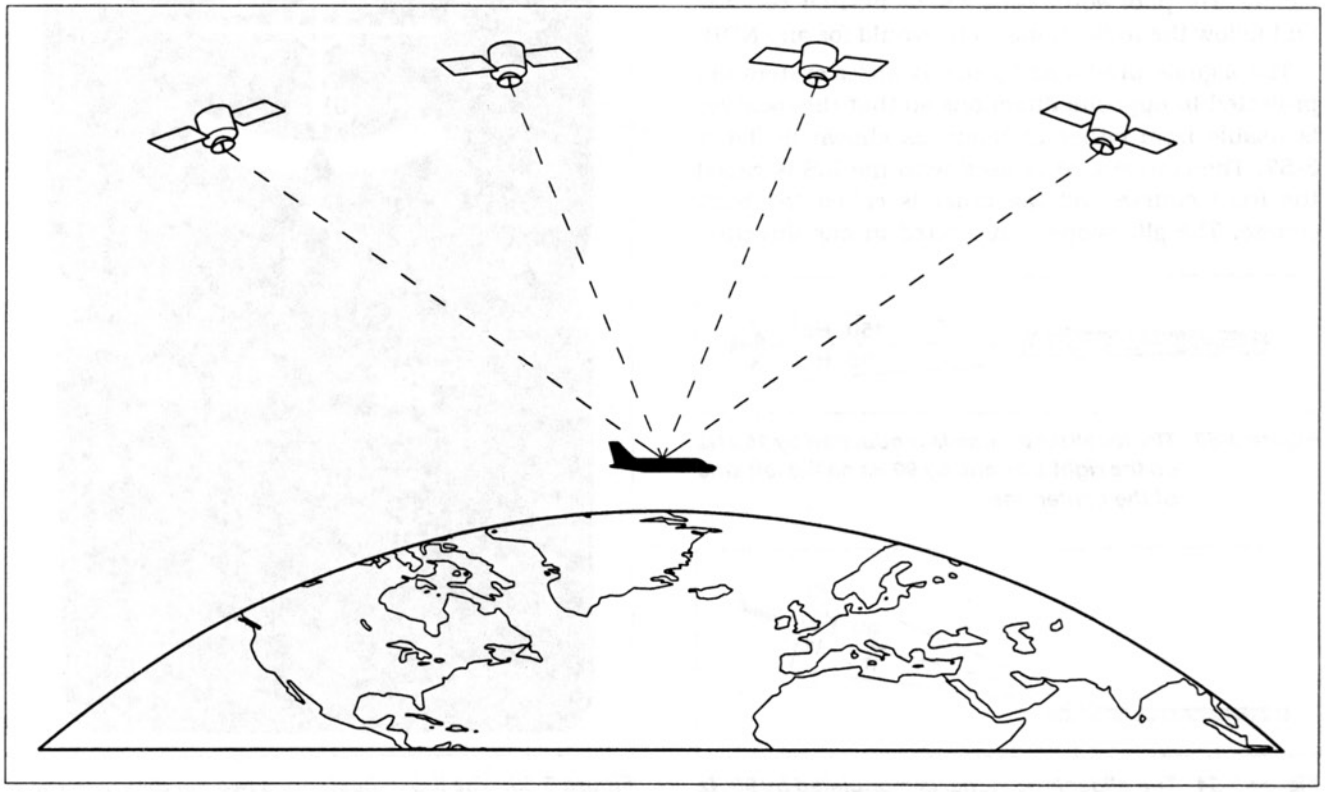

Figure 3-52. When using satellite navigation, the aircraft usually needs to communicate with four satellites for accurate information.

beacons transmit on 75 MHz so different modulations must be used to identify the inner, middle and outer markers. The aircraft receiver does not need to be tuned and in fact is often turned on automatically with the electrical system. The outer marker is modulated with a frequency of 400 Hz and a Morse code sequence of dashes. It also causes a blue light to illuminate in the cockpit so it can be identified by sound or with the blue light or both.

The middle marker is modulated with a frequency of 1300 Hz and a sequence of Morse code dots and dashes.

The amber indicator light comes on over the middle marker. The inner marker is not used with all ILS systems. It uses a modulation of 3000 Hz and a sequence of Morse code dots. The white light comes on over the inner marker. The three marker beacon indicator lights can be seen on the left side of the audio control panel in figure 3-26.

Some ILS systems place an NDB type transmitter at the outer or middle marker locations. These make it easier for the pilot to navigate to the proper location to begin the approach. These are examples of what are called transitional navigational aids. When an NDB transmitter is associated with a marker beacon location, it is called a compass locator. The pilot would tune it in on his ADF receiver and follow the indications as he would for any NDB.

The signals produced by the localizer system are projected in opposite directions so that the localizer is usable from either direction as shown in figure 3-57. The course that is used with the ILS is called the front course and the other is called the back course. The glideslope is projected in one direction only. If an airport has an ILS for runway 9, the full ILS is only available for landings to the east on runway 9 (090°). The back course of the localizer would be available for landings on runway 27 (the opposite end of the same runway). The back course approach is a localizer only approach which will have higher minimums than the ILS because it does not have the glideslope. The sensing of the indicator needle is backwards on the back course approach. The pilot would have to use opposite corrections on the back course compared to the ILS.

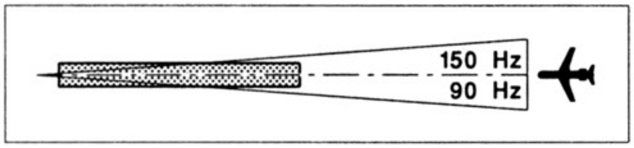

Figure 3-53. The localizer course is modulated by 150 Hz on the right side and by 90 Hz on the left side of the center line.

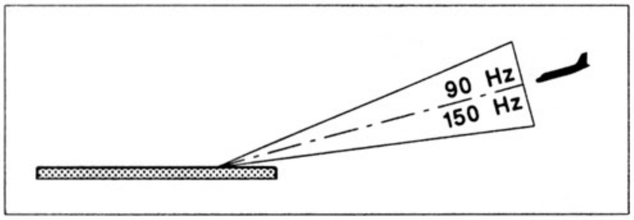

Figure 3-54. The glideslope signal is modulated by 90 Hz above and 150 Hz below the middle of the glidepath.

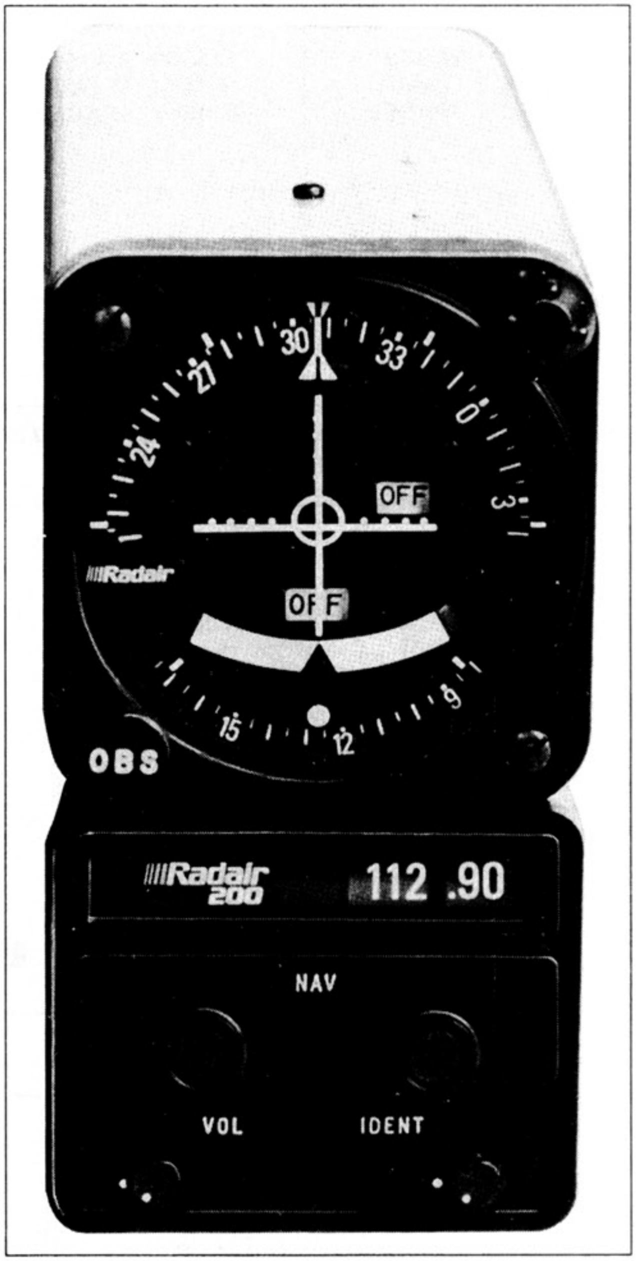

Figure 3-55. The ILS indicator has two needles: a vertical needle for localizer and a horizontal needle for glideslope.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

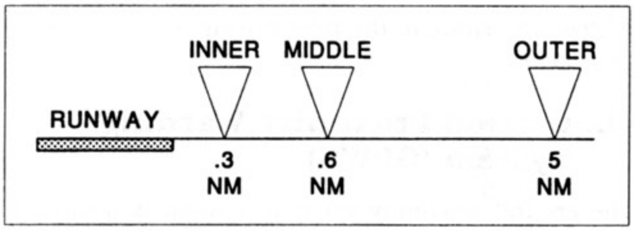

Figure 3-56. The marker beacon transmitters send signals upward which provide indications in the cockpit of distance to the runway.

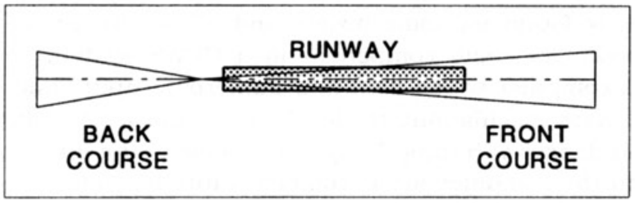

Figure 3-57. The localizer transmits signals that provide guidance on both the front course and the back course. Only the front course is used for an ILS.

J. Microwave Landing System (MLS)

The microwave landing system is a recently developed precision approach system that is in limited use at this time with about eight systems in service in the U.S. The MLS was designed to overcome some of the problems with ILS. A major problem with ILS is that not enough frequencies are available to install the system in all the places it might be needed. ILS is affected by some bending of the beams by obstructions and can only have one glidepath angle that all aircraft must use. The MLS transmitters operate at frequencies between 5.0 and 5.1 GHz. Many new frequencies are available in this range and the signals do not suffer the same kind of interference or bending that affects ILS signals. The principle of operation of the MLS is called a time referenced scanning beam system. Two beams are used: one that scans side to side and one that scans up and down. The aircraft receiver measures the time difference between reception of the TO and FRO scans for the two beams in order to determine lateral and vertical position. The scanning beams used by the MLS are illustrated in figure 3-58. The cockpit display works in a way similar to the one used with ILS. Because of delays in development, only a few MLS approach systems are in use. One of the first was installed at a small airport in Colorado that is served by a commuter airline. The airline operates STOL (short takeoff and landing) aircraft and the glideslope that is used is steeper than normal. This allows a steep approach which gives greater terrain clearance in mountainous areas. The glidepath angle for MLS is determined by the processing of the signals by the aircraft equipment. It is not determined by the installation of the ground antennas. This means that with MLS a different glidepath angle could be used by aircraft with different flight characteristics. It will probably be many years before MLS has replaced ILS to any great extent, in fact, the FAA just recently ordered new ILS equipment for installation at a number of U.S. airports. The use of MLS requires different receivers and antennas than ILS, but both types of equipment may be found on some aircraft.

K. Radar Altimeter

The radio altimeter or radar altimeter is a system which measures the aircraft's height above ground level (AGL) with an accuracy of about 5 ft. A conventional altimeter is not that accurate and it measures MSL not AGL altitude. The usable range for a radar altimeter extends up to 2,500 ft., but it is mainly used during instrument approaches in bad weather. The basic Category I ILS minimums are 200 ft. ceiling and ½ mile visibility. There are other categories of ILS with lower minimums. A

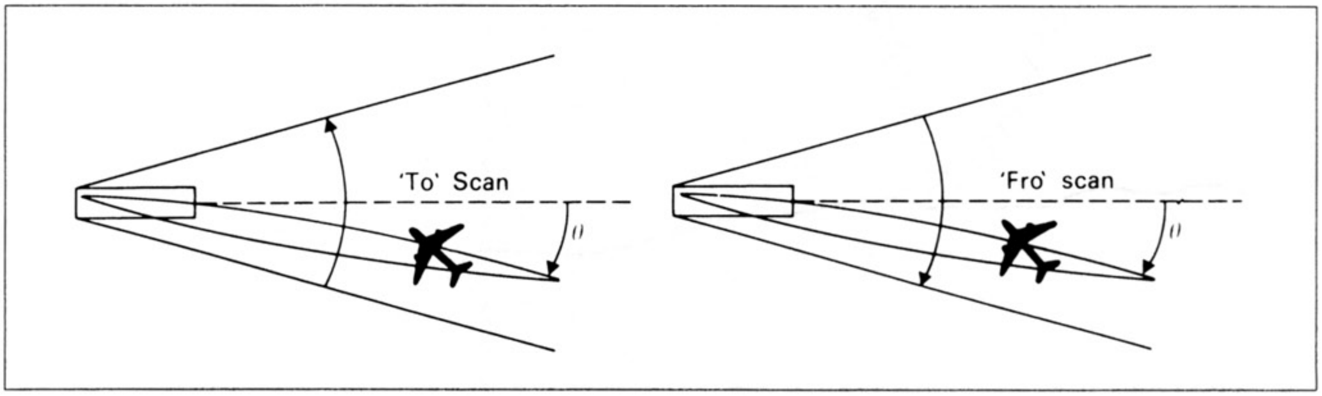

Figure 3-58. The MLS system has two scanning beams. The lateral beam that scans side to side is shown here.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Category II ILS has minimums of 100 ft. ceiling and ¼ mile visibility. The minimums for Category III are even lower. When the aircraft is flying below 200 ft. in bad weather, a better method of measuring altitude than a barometric altimeter is needed for safety. This is the main use for a radar altimeter, as a precise way to measure AGL altitude during IFR approaches. The radar altimeter uses antennas that are installed on the belly of the aircraft. The transmitter sends out radio waves at 4.3 GHz which strike the earth and bounce back to the receive antenna. By measuring the travel time for the radio waves, an accurate calculation can be made of AGL altitude. The basic parts of a radio altimeter or radar altimeter system and one type of cockpit indicator are shown in figure 3-59. Another type of cockpit indicator for a radar altimeter that uses a round display is shown in figure 3-60. This instrument has a bug that can be set at 200 ft. or some other altitude to give a warning to the pilot during an instrument approach.

L. Ground Proximity Warning System (GPWS)

The ground proximity warning system is designed to provide warnings to the flight crew when the aircraft is in danger of striking the ground due to excessive descent rate or rising terrain. This equipment is required on all air carrier jets and it is found on some bizjets and other aircraft as well. The main component in a GPWS system is a computer which monitors numerous inputs and makes calculations to determine if the aircraft is in danger of hitting the ground. Some of the inputs to the computer are barometric altitude, radar altitude, rate of climb or descent, flap position and landing gear position. The GPWS is one of the few

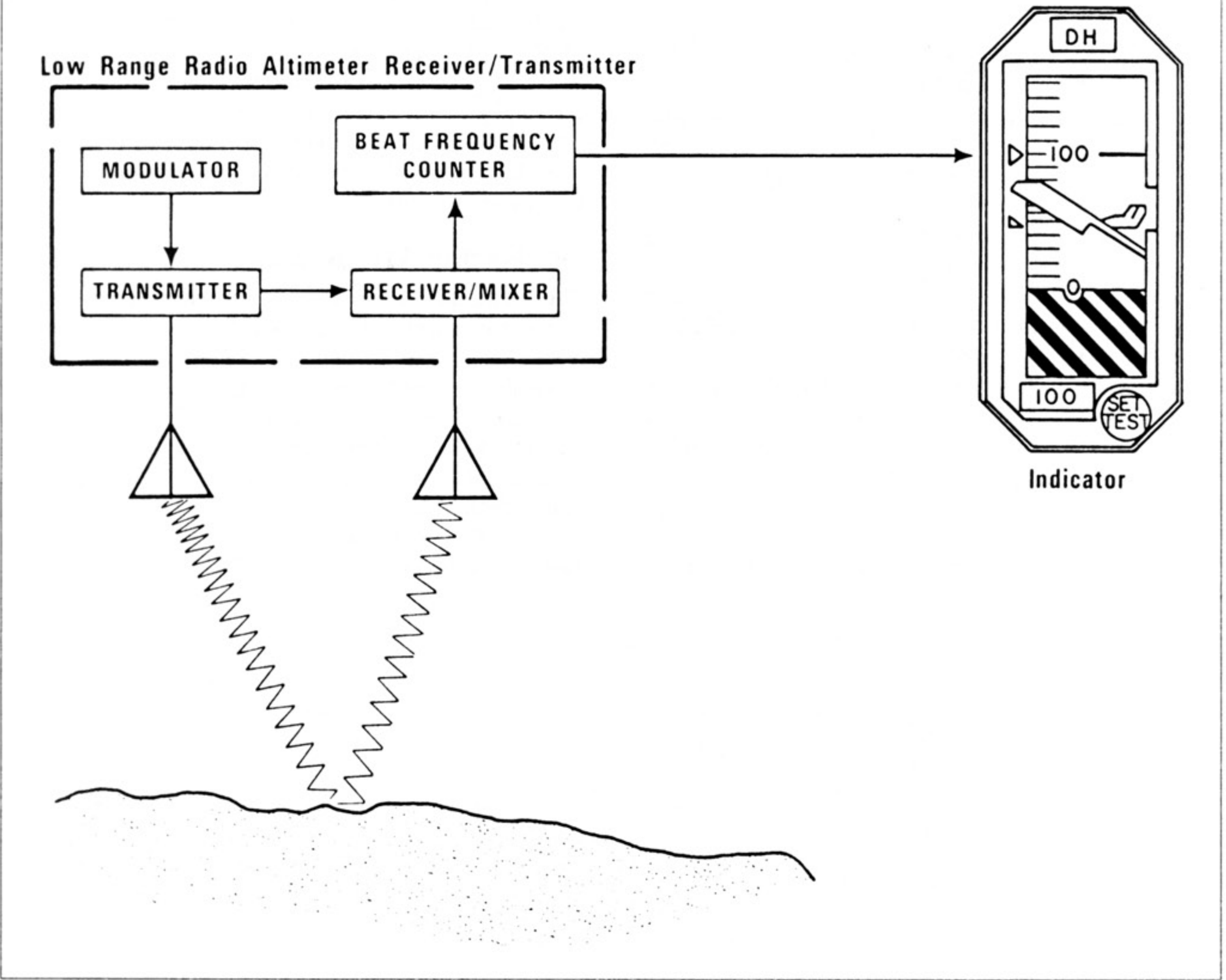

Figure 3-59. The aircraft radar altimeter system is also called a radio altimeter.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

systems on a civil aircraft that gives a spoken voice command to the flight crew. When it is determined that a warning must be given, a recorded voice on tape is activated which tells the flight crew to "PULL UP, PULL UP" or a similar type of message.

M. Weather Radar

A radar weather unit is another piece of equipment which is required for all air carrier jets and is common on many other types of aircraft. Aircraft weather radar is a pulse radar that typically operates at 9.375 GHz. The radar antenna is installed on the front of the aircraft where it sends out brief pulses of radar frequency EM waves in order to locate and avoid thunderstorms. There must be something present in a thunderstorm which will reflect the radar pulse. Clouds are invisible to radar, but ice, hail and especially rain will reflect the energy back to the aircraft radar antenna. The strength of the return is affected by the size of the raindrops and the rainfall intensity. Color radars use different colors for different intensity levels. Green, yellow and red are often used with red indicating the highest intensity of rainfall.

Radar signals can also be reflected from the ground and the radar system can be used to locate surface features on the earth below. This mapping feature is especially effective when used to pick up well defined coastlines.

Weather radar is called a pulse radar system because it transmits very brief pulses of energy. This is necessary in order to use the same antenna for transmit and receive and to produce a usable maximum range. The transmitter sends out a pulse that has a duration of about one microsecond. Then the antenna is switched to the receiver for a period of about 2,500 microseconds. The receiver must be connected long enough for the pulse to travel out to the maximum range and back again. The use of pulse radar also makes the system more efficient since the transmit energy is concentrated in brief pulses which permits much higher values of peak power than would otherwise be possible. The major components and their location for a weather radar system on a business jet airplane are shown in figure 3-61.

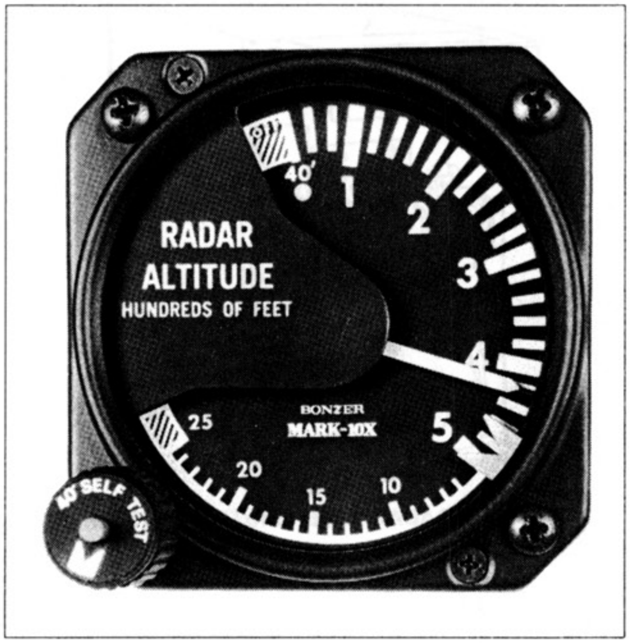

Figure 3-60. Some radar altimeters use a round indicator instrument.

The major components of a weather radar system and their functions will now be listed:

1. Antenna — The antenna is a parabolic reflector or a newer and more efficient flat plate antenna. It does not rotate 360° like ATC radar, but scans side to side through an arc of approximately 120°.
2. Radar screen or display — This shows the returns picked up on the radar, usually in 3 or 4 different colors.
3. Cavity magnetron — This is a special device used to produce the radar frequency EM waves for the radar system (figure 3-62).
4. Synchronizer — The antenna and the screen must be synchronized in order to show the correct location of the returns.
5. Duplexer — This unit rapidly switches the antenna between the transmitter and receiver.
6. Stabilization System — The antenna unit needs to be gyro stabilized so that pitch and roll attitudes of the aircraft will not cause incorrect display presentations.
7. Waveguide — because of the power and frequency of the EM waves, coaxial cables cannot be used to connect the antenna with the R/T unit. A hollow tube called waveguide is used for this purpose (figure 3-63).

The weather radar has a maximum range from 200–300 miles on a typical installation. The controls allow the pilot to select different ranges and different settings so that the best indications of thunderstorms can be produced depending on the conditions encountered.

A tilt control is included so that the antenna can be tiled up and down to gauge the vertical extent of the storm cell. The radar antenna is protected by a plastic or fiberglass radome which must be carefully maintained to prevent adverse effects on radar performance. The radome often

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

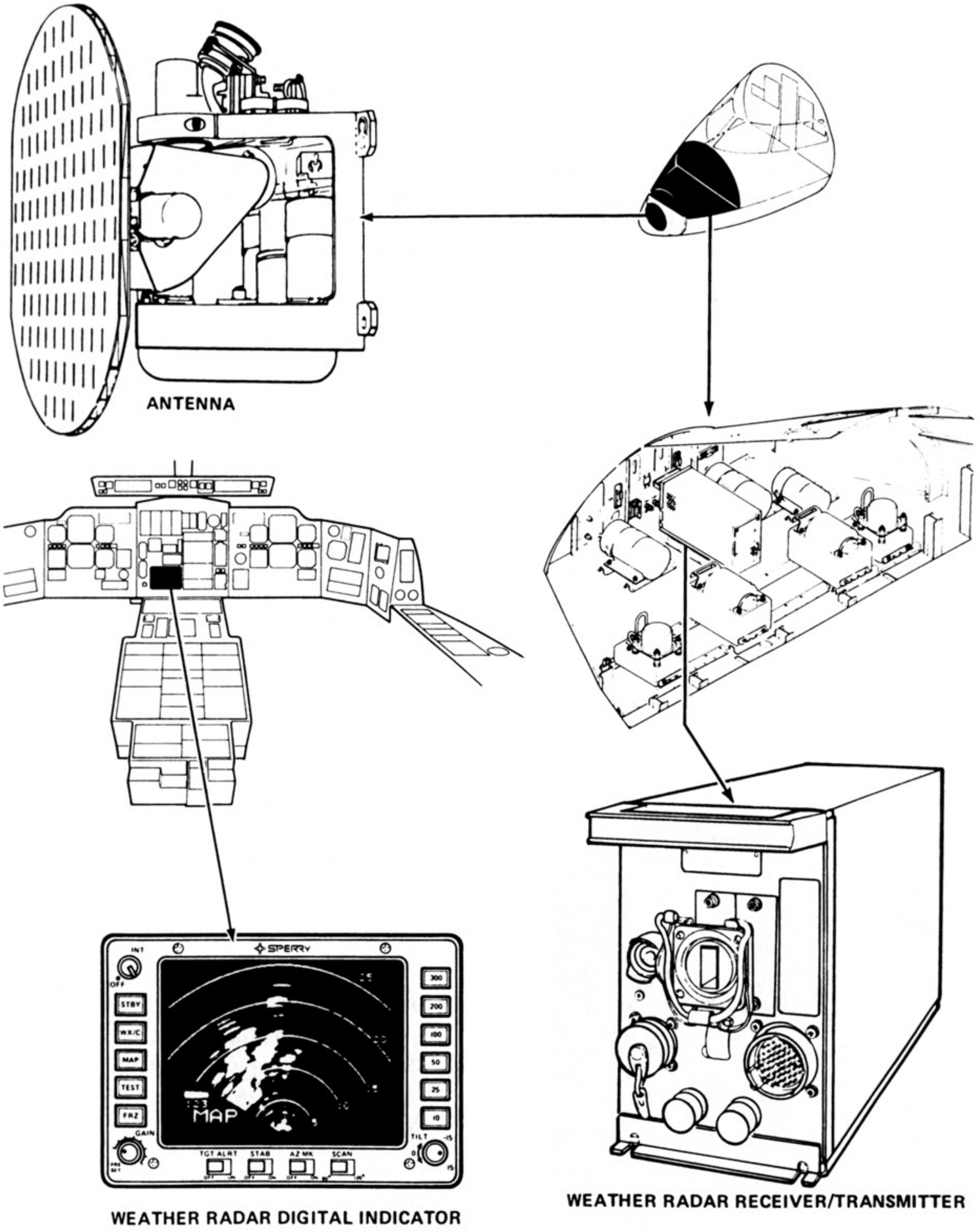
ANTENNA
SPERRY
INT
OFF
STBY
WX/C
MAP
TEST
FRZ
GAIN
300
200
100
50
25
10
TILT
TGT ALRT
STAB
AZ MK
SCAN
MAP
WEATHER RADAR RECEIVER/TRANSMITTER
WEATHER RADAR DIGITAL INDICATOR

Figure 3-61. The major components of a weather radar system. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

has conducting strips fastened on the outside to conduct static charges and lightning strikes away from the radome. A typical arrangement of lightning diverter strips is shown in figure 3-64. The radome should only be painted with approved types of paint which will not interfere with the radar frequency signals that must pass through the radome.

Personal safety is very important when working on aircraft with radar systems. Some of the components in the receiver/transmitter unit can hold very high voltages and should be worked on only by personnel that are familiar with the necessary safety precautions. The emissions from the radar antenna can be very hazardous to human beings. The radar should never be turned on while on the ground unless special precautions are taken. The manufacturers maintenance instructions usually include some information on the MPEL. The maximum permissible exposure level gives safe distances from aircraft radar antennas. The best procedure is to never walk in front of an aircraft when the radar might be turned on.

N. Stormscope®

The Stormscope is a weather avoidance system that uses completely different methods to locate thunderstorms than a radar system. The Stormscope is designed to receive the radio frequency EM waves produced by lightning discharges. It uses a directional antenna system similar to that used by the ADF equipment. In fact in some cases it is possible to connect the Stormscope to the aircraft ADF antennas with special couplers. The direction of the lightning is determined using the directional antenna and the relative intensity of the discharge is measured. The intensity is used as pseudo range on the display. It is not actual range like that obtained from weather radar, but it does give useful information to the pilot. The display instrument in the cockpit is normally a small round LCD display that shows a light dot for each lightning strike that is detected. From the patterns on the display the pilot can determine where the worst areas are located and avoid them. Figure 3-65 shows the appearance of a typical Stormscope display instrument. The purpose of all weather detection systems is avoidance. A very strong thunderstorm cell has the capability of tearing apart even the strongest of aircraft.

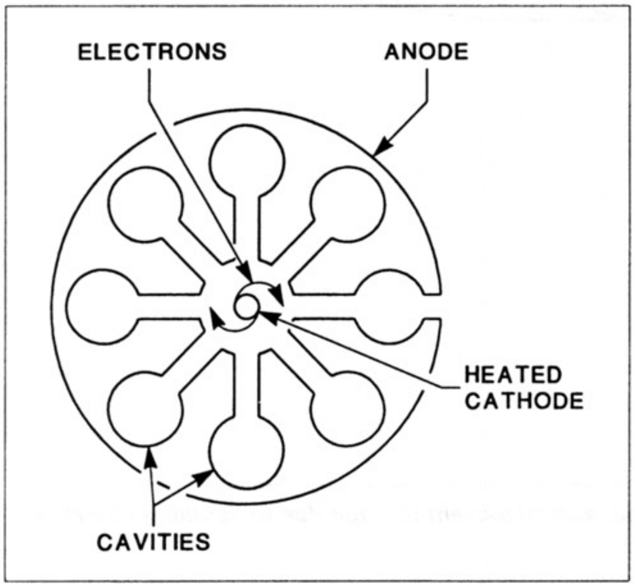

Figure 3-62. A cavity magnetron produces the powerful SHF band EM waves for a weather radar transmitter.

Since the Stormscope and the weather radar react to different aspects of thunderstorms, the best weather avoidance system would be to have both installed in the aircraft. Many corporate aircraft in fact do have both systems installed.

O. TCAS — Airborne Collision Avoidance System

The full meaning of the abbreviation TCAS is traffic alert and collision avoidance system. The prevention

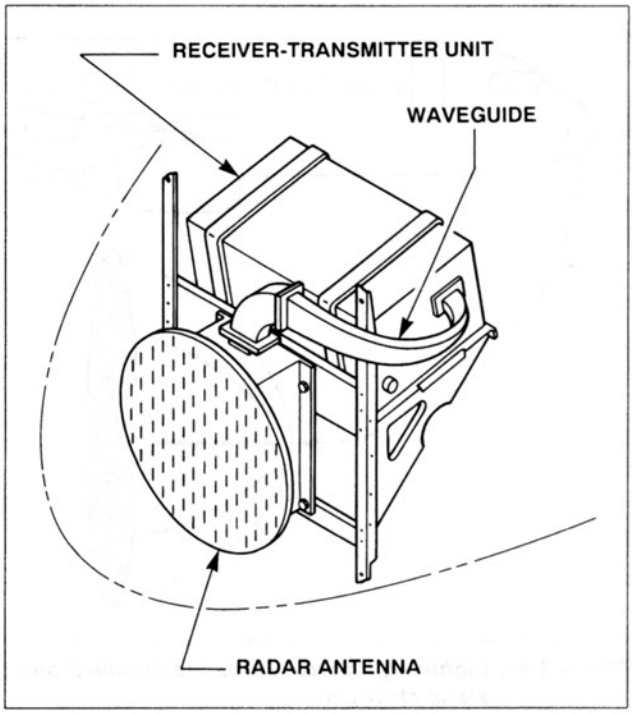

Figure 3-63. Waveguide is used to carry the radar frequency energy between the R/T unit and the antenna in a weather radar system. (Courtesy Piper Aircraft Co.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

of collisions between aircraft is the primary responsibility of the ATC system. This TCAS system was developed because both the FAA and the airlines were interested in having a system which would display collision threats to the pilots. This type of equipment is being installed at the present time in air carrier jets. About half of them have had the equipment installed already. The TCAS equipment uses some of the same equipment and principles as the transponders discussed earlier. The TCAS equipment on the aircraft includes a computer, a display screen in the cockpit and a directional antenna system. The unit sends out interrogations in all directions around the TCAS aircraft. Any transponder equipped aircraft within range will send back a reply and the TCAS calculates direction, range and altitude of the other aircraft. Of course it can only determine altitude if the other aircraft is Mode C equipped. If the other aircraft does not have a transponder, it will not be detected at all. Each aircraft that has been detected within a certain range will be displayed as a lighted symbol on the display screen. If the other aircraft gets closer and creates a threat, the symbol will change color and shape. If the TCAS equipment determines that a sufficient danger level is present, it will display an avoidance maneuver command to the pilots. The avoidance maneuver will be in the vertical plane only, the present equipment is not able to suggest turns as avoidance maneuvers. The pilot will be told on the display to climb or descend at a certain rate to avoid the threat aircraft. The general appearance of the cockpit indicator for a TCAS is shown in figure 3-66. The aircraft symbols are different shapes and colors with an altitude and climb or descent arrow next to the symbol. The position of the TCAS equipped aircraft is shown by the airplane symbol at the center of the range circle. The installation of TCAS in an aircraft normally uses a Mode S transponder and a special type of directional antenna just for the TCAS equipment.

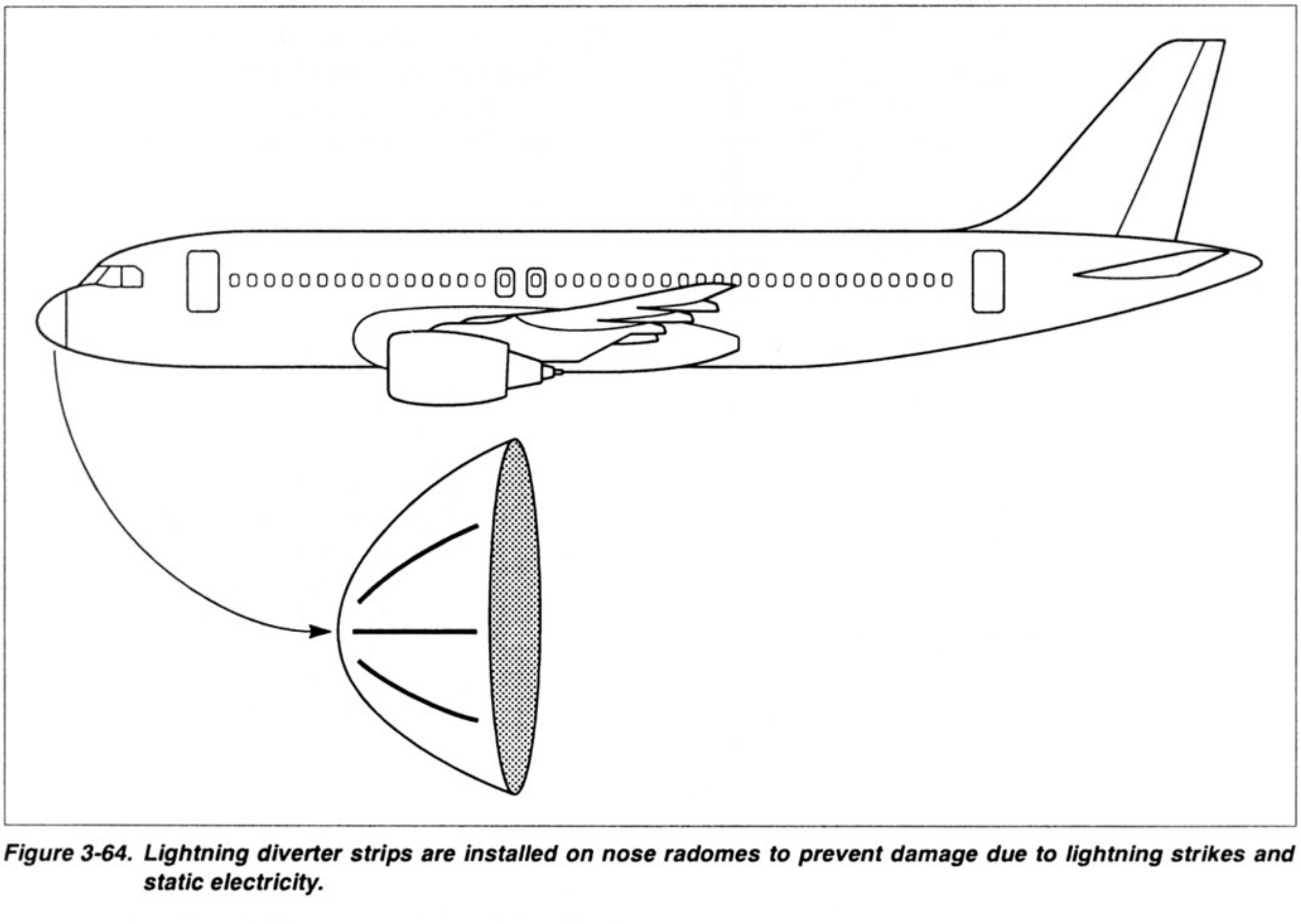

Figure 3-64. Lightning diverter strips are installed on nose radomes to prevent damage due to lightning strikes and static electricity.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

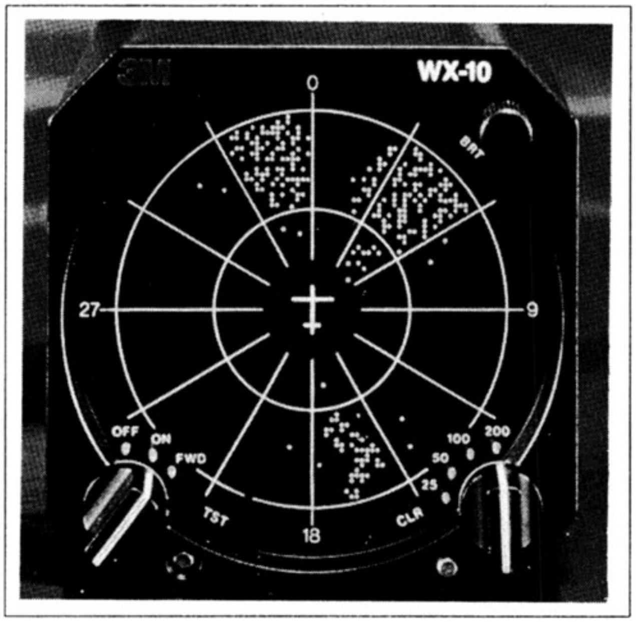
3M
WX-10
0
BRT
27
9
OFF
ON
FWD
TST
18
CLR
25
50
100
200

Figure 3-65. The type of display normally used with a Stormscope.

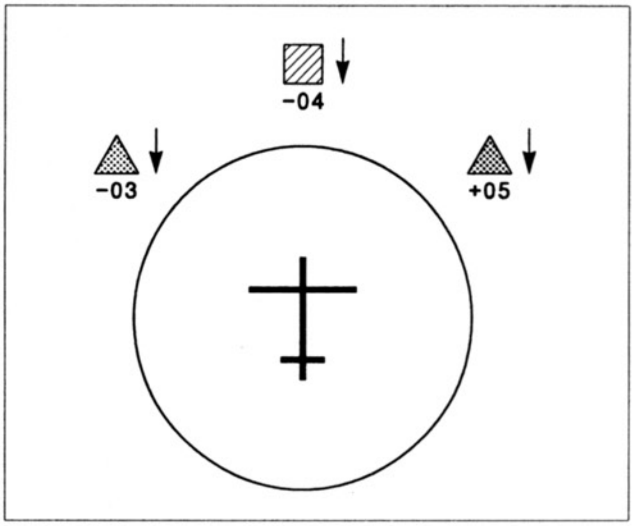

Figure 3-66. Simplified example of a TCAS display. The symbols for threat aircraft use different shapes and colors.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

CHAPTER IV

Aircraft Antennas and Autopilots

This chapter contains three major topic areas: recommendations for installing and inspecting avionics equipment, additional information concerning aircraft antennas and aircraft autopilot systems.

A. Installation and Inspection of Avionics

There are a number of factors which make the installation of avionics equipment more critical than the installation of other types of equipment. Radios and avionics are very sensitive to electromagnetic interference which can be created by nearby wiring and other electrically operated devices. The installation and maintenance of good bonding jumper connections is important to ensure proper operation of avionics units. Avionics equipment is easily damaged by excessive heat which requires that provisions be made for adequate air circulation. The most common cause for failures in avionics equipment is probably overheating. Figure 4-1 shows the avionics cooling arrangement for a small airplane. The tubing and plenums in this type of system must be maintained properly to ensure that adequate cooling airflow is available to the equipment. Other types of avionics cooling systems use electric motor driven fans, these must be maintained in good condition to prevent overheating. A thorough knowledge of these and other important considerations is necessary to ensure proper performance from installed avionics equipment in aircraft. The repair of avionics equipment is accomplished by appropriately rated FAA repair stations, but the installation, inspection and routine maintenance on these units is often performed by A&P technicians. Some of the important considerations when installing and inspecting avionics equipment will be discussed next.

1. Cleaning of Electronic Equipment

Cleaning of electronic equipment is important in order to remove accumulations of dust, dirt and lint that can block cooling holes and cause overheating. Dirt and lint which collects on open terminal strips and other electrical connections can absorb moisture and cause short circuits. Open terminal strips like those shown in figure 4-2 should be cleaned regularly to prevent the accumulation of dirt and lint that can cause these types of problems.

Electrical connections should be kept clean and free of corrosion and oxidation which can add unwanted resistance. When a mild abrasive is needed to remove corrosion and oxidation on terminal strips and mating surfaces, emery cloth is recommended.

Commutators and slip rings should be cleaned with crocus cloth or very fine sandpaper.

Older electronic equipment made extensive use of rotary selector switches and similar devices with many sets of contacts. Spray cans of a special cleaning solvent are available for cleaning the contacts of these devices. Before using a spray can contact cleaner, you should ensure that it is compatible with any plastic or non-metallic parts of these switches.

2. Routing Wires

Wiring of all types should be routed above lines that carry fluids and clamped securely to the aircraft structure. The wires should be routed so as to prevent abrasive damage from control cables, mechanical linkages and other moving parts in the aircraft. Frequent clamps and ties should be used to prevent excessive wire movement due to in-flight vibration and other factors. The proper installation of clamps to prevent excessive wire movement is illustrated in figure 4-3. Where wires terminate at pieces of equipment, enough slack should be left after the last clamp to allow for shock mount motions. If wires are clamped tightly too close to the termination point, normal aircraft motions and vibration will put bending loads on the wire connectors and cause premature failure in the wires.

Another factor to consider when routing wiring for electronic equipment is electromagnetic interference (EMI). Antenna leads and other wiring sensitive to EMI should be routed away from the wires for inverters, power supplies, strobe lights, motors and other components that are known to cause interference. When troubleshooting a noise or interference problem in aircraft radios and sensitive electronic equipment, it is often necessary to reroute wires away from the source of the EMI once it has been located.

3. Switches and Circuit Breakers

The radios on aircraft are usually connected to an avionics master switch. This switch is separate from

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

the normal master switch as shown in figure 4-4. This is a useful feature which makes it easier for the pilot to ensure that all the radios are turned off when starting and stopping the engine. The radios should be turned off when the engine is started and stopped in order to prevent damage caused by surge currents and spikes of high voltage that can occur during engine starts and engine shut down. When installing switches in aircraft circuits, the rating of the switch must be adequate to handle both the type and amount of current and voltage for the circuit. Figure 4-5 shows a typical

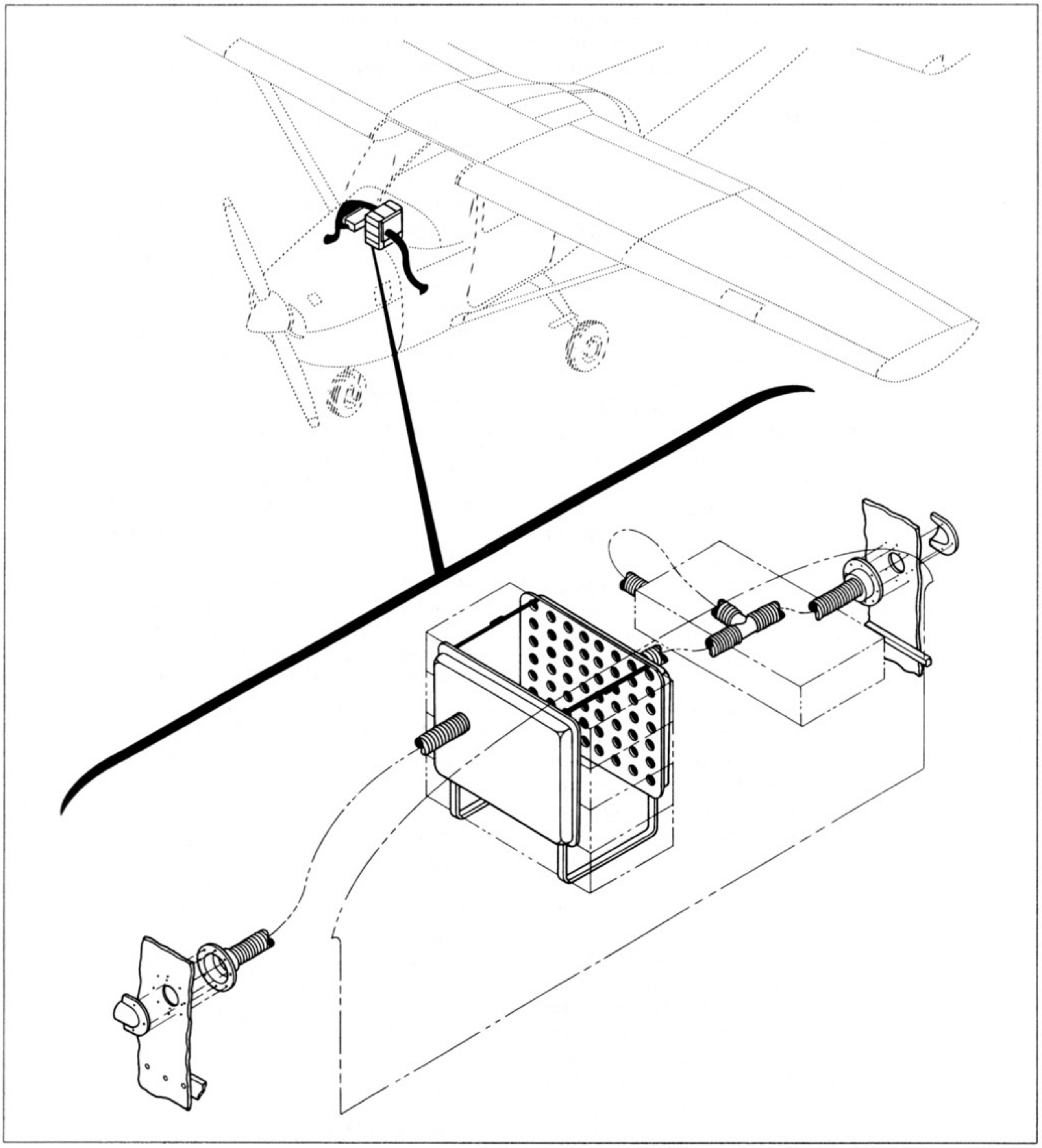

Figure 4-1. Avionics cooling installation for a small airplane that uses vents on the sides of the fuselage to move air across the radios.

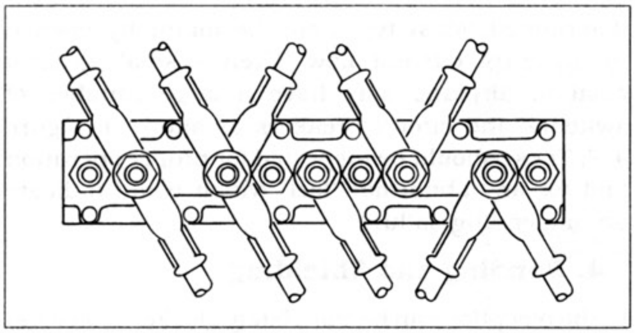

Figure 4-2. Open terminal strips should be kept clean and free of corrosion. Check for metal objects that could fall across the terminals and cause shorts.

aircraft toggle switch. Whether the circuit is AC or DC makes a significant difference in the proper selection of switches.

For example, there is a common aircraft quality switch that is rated for 10 amps at 125–250 volts AC. The same switch is rated at 0.3 amps when it is used in a DC circuit up to 125 volts DC. If this switch was installed in a 10 amp DC circuit the points would quickly burn and fail. The reason for this difference in ratings is that the current in an AC circuit drops to zero twice each cycle, this greatly reduces the problem of arcing as the points in the switch are opening. The proper rating of a switch in terms of both current flow and AC

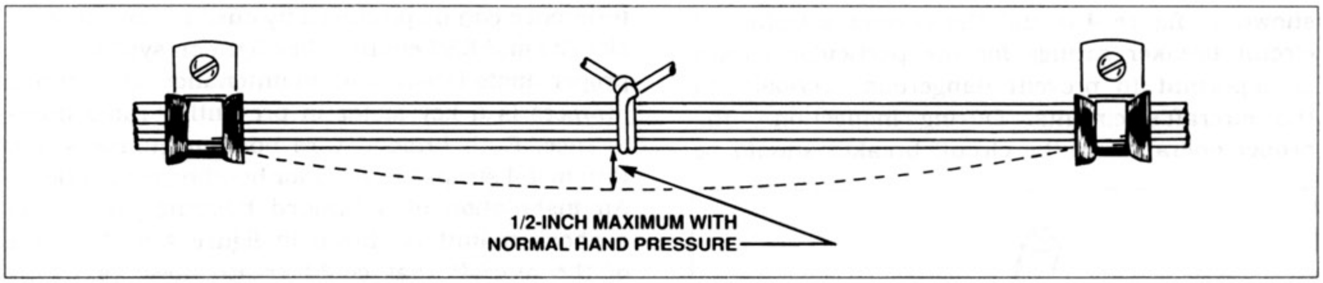

Figure 4-3. Wires and antenna leads should be supported with proper clamps and ties.

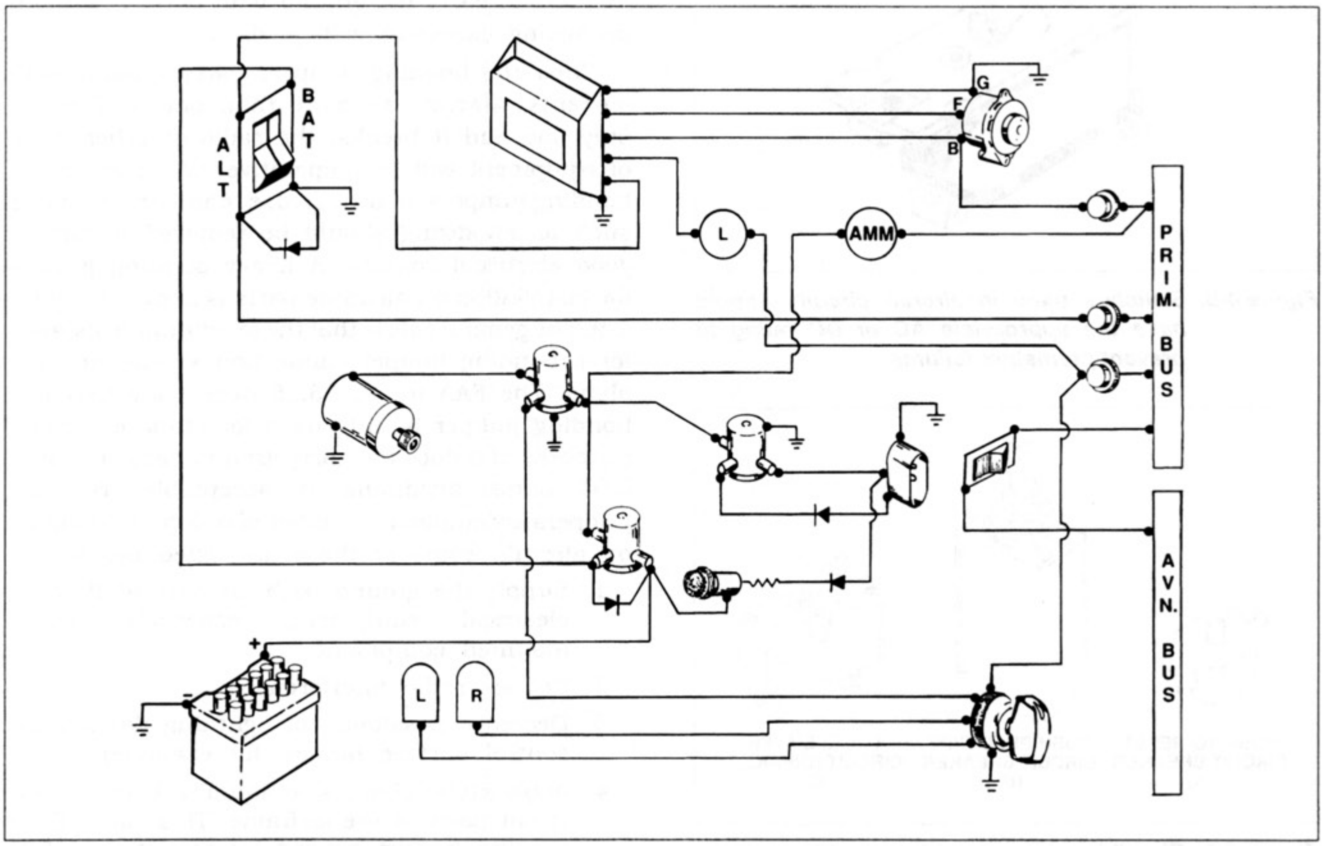

Figure 4-4. An avionics master switch supplies electrical power to the avionics bus. It should be off when starting and stopping the engine.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

versus DC rating is very important to ensure adequate performance and service life. The condition of switches can be checked during an inspection by operating the switch and checking the "feel" during operation. Many switches have over-center mechanisms and other devices that produce a distinct feel to the switch. When the switch is getting worn and ready to fail, it often starts to feel sloppy in operation.

Circuit breakers for aircraft circuits should be the "trip free" type. This means that the circuit breaker cannot be overridden by holding it in the engaged position. It will open the circuit regardless of the position of the control toggle or push button. Various types of circuit breakers are available as shown in figure 4-6 and the correct selection of circuit breaker ratings for the particular circuit is important to prevent dangerous overloads in the aircraft's circuits. During inspections, the proper operation of the circuit breaker should be determined. Most types can be manually opened to interrupt current flow. Even a small general aviation airplane may have a large number of switches and circuit breakers as shown in figure 4-7. These should be inspected for proper operation and for any abnormal "feel" which could indicate an impending failure.

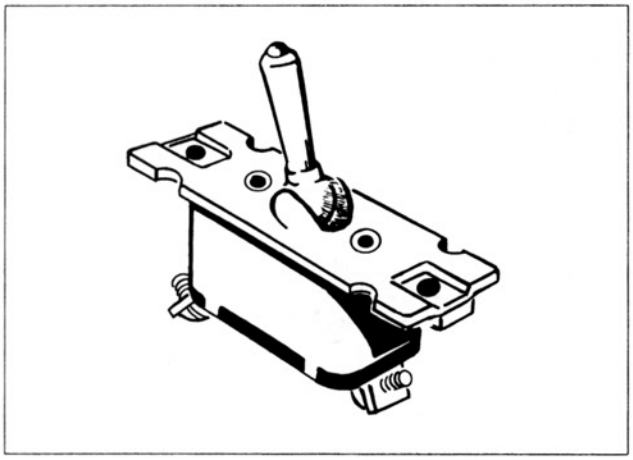

Figure 4-5. Switches used in aircraft circuits should have the appropriate AC or DC rating to prevent premature failures.

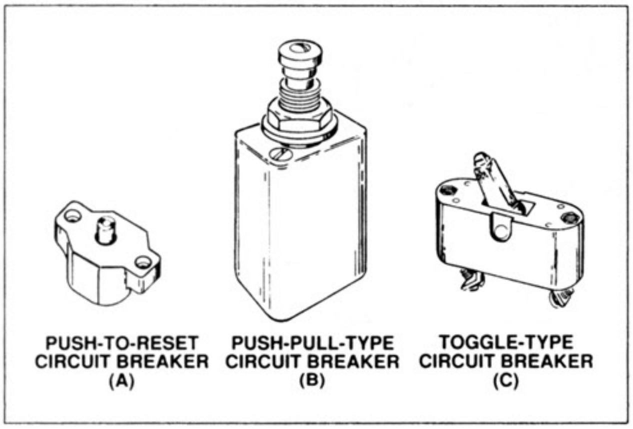

PUSH-TO-RESET CIRCUIT BREAKER
(A)
PUSH-PULL-TYPE CIRCUIT BREAKER
(B)
TOGGLE-TYPE CIRCUIT BREAKER
(C)

Figure 4-6. Circuit breakers should be the trip-free type and inspected regularly for proper operation.

4. Bonding and Shielding

Radio reception can be completely blocked or severely interfered with by improper bonding and shielding in the aircraft. The source of the noise interference that affects aircraft radios is both inside and outside the aircraft. Outside interference comes from precipitation static and thunderstorms. Inside interference can be produced by current flow in other circuits and EMI emitters like ignition systems. The proper installation and maintenance of bonding jumpers is a key factor in preventing radio interference. Both braided wire bonding jumpers and thin metal straps are used for bonding connections. An installation of a braided bonding jumper on a shock mount is shown in figure 4-8. All parts of the aircraft that could create noise problems should be bonded. Electrical equipment that is shock mounted should have adequate bonding jumpers to carry the ground path current without producing excessive voltage drop.

When the bonding jumpers carry ground path currents, always use more than one. If there is only one and it breaks, the radio or other piece of equipment will be inoperative. When attaching bonding jumpers all dirt, grease, paint and coatings such as anodizing should be removed to ensure good electrical contact. A heavy bonding jumper for installation on airframe parts is shown in figure 4-9. The general rule is that the maximum resistance for a bonding jumper connection should be .003 ohms. The FAA in AC 65.15 does state that if a bonding jumper is only used for static electricity purposes and does not carry ground path currents, 0.01 ohms maximum is acceptable. Bonding jumpers accomplish a number of different functions on aircraft, some of these are listed below.

1. Supply the ground path for current flow for electrical equipment, especially shock mounted equipment.
2. Reduce radio interference.
3. Decrease possibility of lightning damage (at control surface hinges, for example).
4. Allow static charges to equalize between different parts of the airframe. This can reduce the fire hazard caused by arcing near fuel tank vents, etc.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

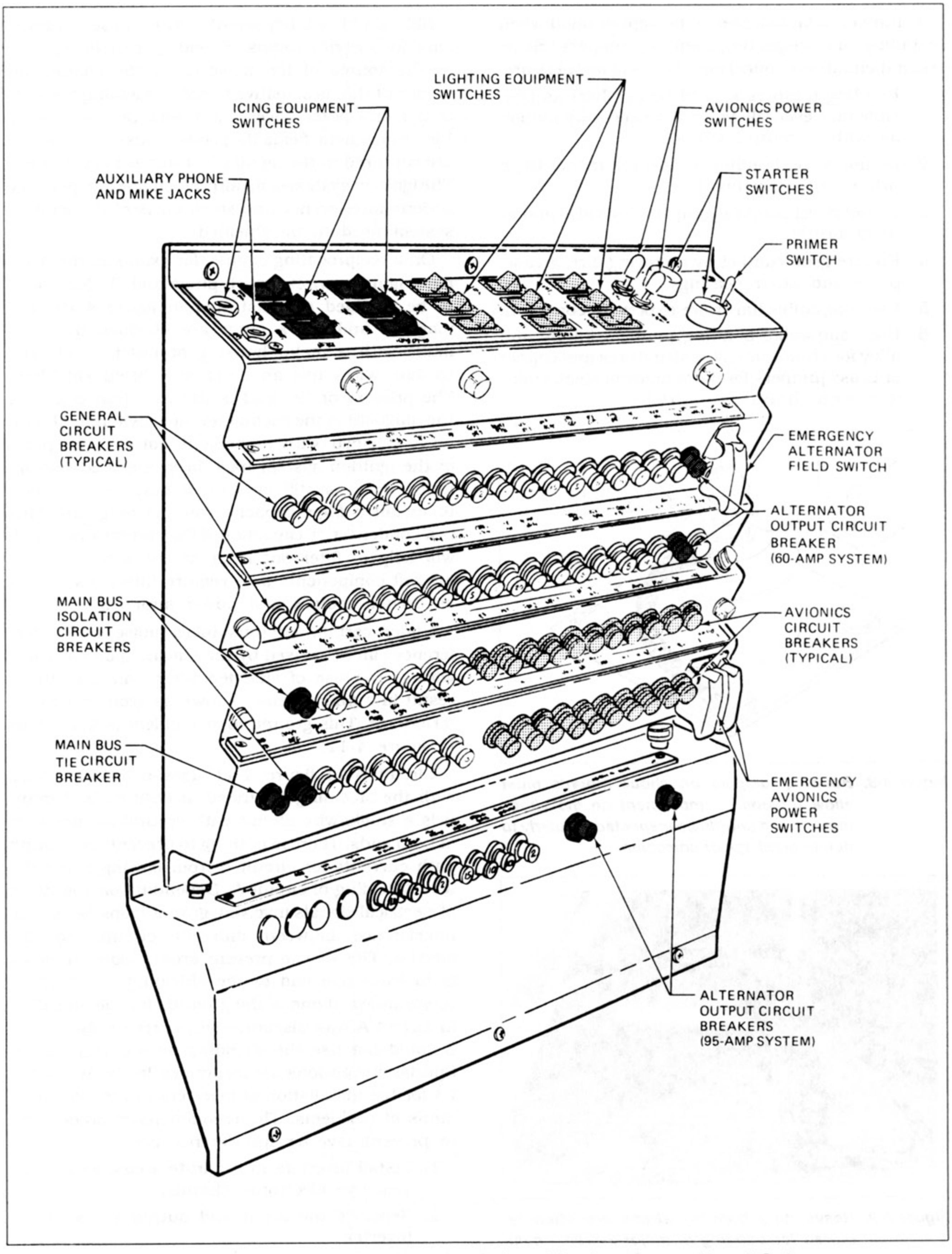
LIGHTING EQUIPMENT SWITCHES
ICING EQUIPMENT SWITCHES
AVIONICS POWER SWITCHES
AUXILIARY PHONE AND MIKE JACKS
STARTER SWITCHES
PRIMER SWITCH
GENERAL CIRCUIT BREAKERS (TYPICAL)
EMERGENCY ALTERNATOR FIELD SWITCH
ALTERNATOR OUTPUT CIRCUIT BREAKER (60-AMP SYSTEM)
MAIN BUS ISOLATION CIRCUIT BREAKERS
AVIONICS CIRCUIT BREAKERS (TYPICAL)
MAIN BUS TIE CIRCUIT BREAKER
EMERGENCY AVIONICS POWER SWITCHES
ALTERNATOR OUTPUT CIRCUIT BREAKERS (95-AMP SYSTEM)

Figure 4-7. Typical circuit breaker panel for a twin-engine airplane. (Courtesy Cessna Aircraft Co.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

A number of factors should be kept in mind when installing and inspecting bonding jumpers. Some recommendations concerning bonding jumpers are:

1. Bonding jumpers should be as short as possible (however, allow for any necessary motion as with a control surface).
2. Do not solder bonding jumpers. It makes them brittle and they break.
3. Do not paint bonding jumpers. This also makes them brittle.
4. Ensure good contact by removing dirt, grease, paint and other coatings.
5. Use compatible hardware to prevent corrosion.
6. Use compatible bonding jumpers (aluminum alloy for aluminum alloy structures and copper or brass jumpers for parts made of steel, stainless steel, brass or bronze).

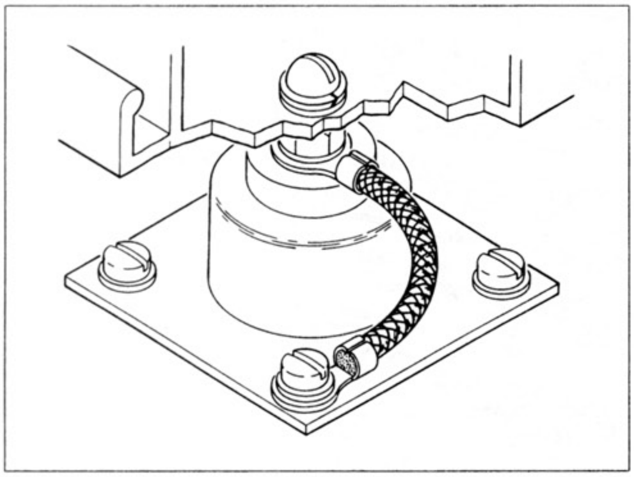

Figure 4-8. Bonding jumpers on shock mounts must allow freedom of movement on the shock mounts and should be inspected regularly to detect breakage or corrosion.

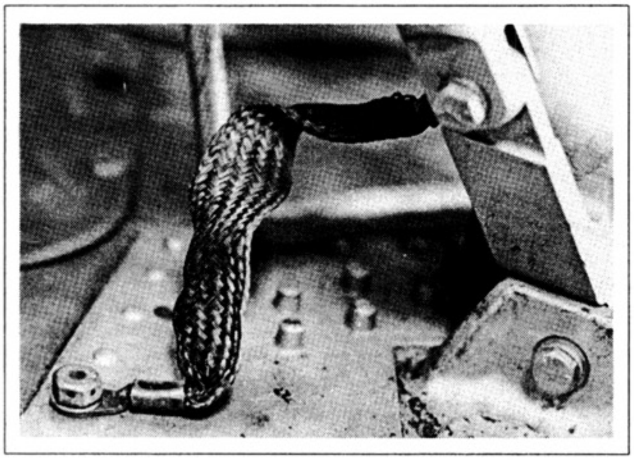

Figure 4-9. Heavy duty bonding straps are often required for bonding of major airframe components.

Shielding is an important part of noise suppression for aircraft radios. Shielding can be applied at the source of the noise or at the component or circuit that is sensitive to EMI. Shielding consists of a metal outer cover for a wire or component. Electromagnetic fields that could cause interference are captured in the metal cover and sent to ground. The ignition system of an aircraft engine can produce serious interference and so all parts of the ignition system need to be shielded.

On a reciprocating engine, for example, the magneto, ignition wires, spark plugs and "P" lead need to be shielded as illustrated in figure 4-10. The magneto and spark plugs are shielded by being made with a metal housing or outer cover. The ignition wires use an outer wire braid shielding. The primary or "P" lead is the wire that connects the magneto to the cockpit ignition switch. It should be a shielded wire to prevent noise. If all parts of the ignition system have been shielded and ignition noise is still present, it may be necessary to install a filter capacitor on the magneto. This is a condenser or capacitor of the correct size which will help to filter out noise at the source. Other aircraft components may require filters also, such as certain motors and power supplies.

Under certain circumstances noise and interference can be caused by the shielding on electrical wiring. The use of shielded wires can sometimes result in a phenomenon known as ground-loop interference. This ground-loop problem is illustrated in figure 4-11.

Circuit A in figure 4-11 uses a shielded wire with the shielding grounded at both ends. Circuit B is a single-wire circuit with ground connections at both ends. There is nothing to prevent the ground path currents for circuit B from flowing along the wire shielding of circuit A. Depending on the types of electrical signals involved, groundloops can cause interference between different circuits in the aircraft. The way to prevent groundloop problems is to leave one end of the shielding "floating" or ungrounded. If one of the grounds for the shielding in circuit A was disconnected, currents for circuit B could not use the shielding as a current path. Special precautions are recommended in AC 43.13-2A for the installation of inverters to prevent these kinds of problems. The recommended procedures to prevent inverter interference are:

1. Install inverters in separate areas, away from sensitive electronic circuits.
2. Separate the input and output wires of the inverter.
3. Properly bond the inverter case to the airframe.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

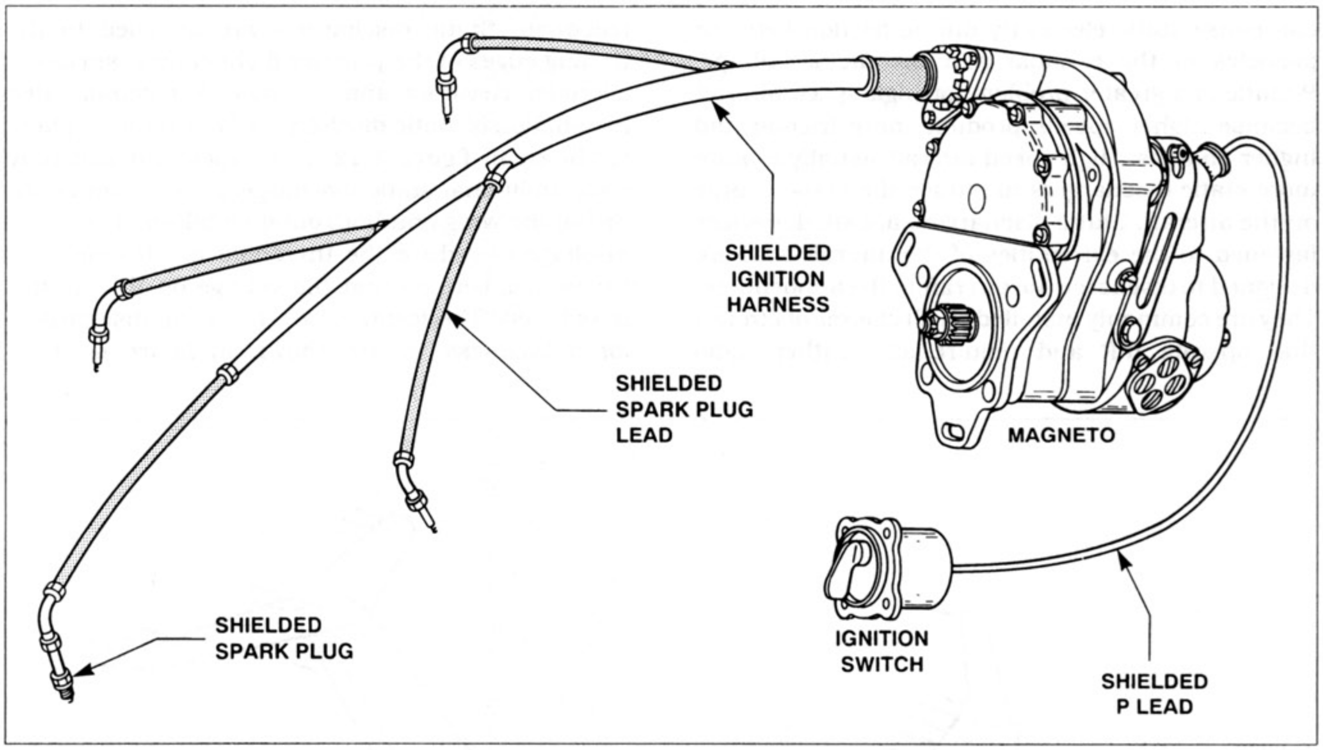
SHIELDED
IGNITION
HARNESS
SHIELDED
SPARK PLUG
LEAD
MAGNETO
SHIELDED
SPARK PLUG
IGNITION
SWITCH
SHIELDED
P LEAD

Figure 4-10. All parts of the ignition system should be shielded to prevent radio noise.

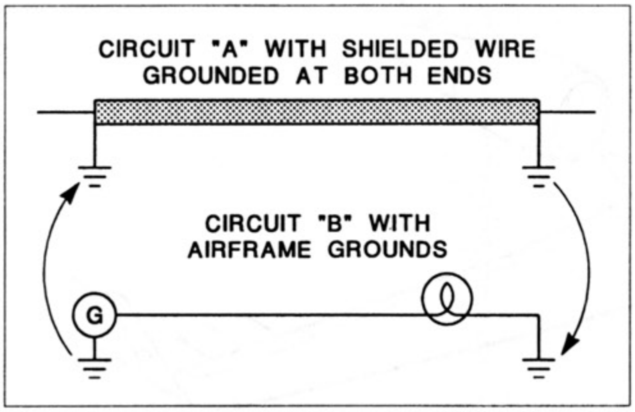
CIRCUIT "A" WITH SHIELDED WIRE
GROUNDED AT BOTH ENDS
CIRCUIT "B" WITH
AIRFRAME GROUNDS
G

Figure 4-11. Example of how "ground loop" interference can occur.

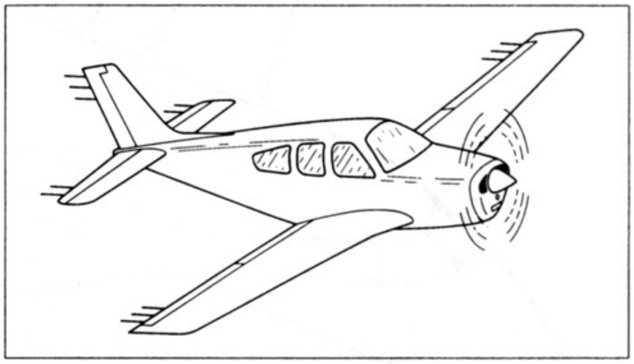

Figure 4-12. Static wicks are installed on the trailing edges of the flight control surfaces to help remove static charges in flight in order to prevent noise in the radios.

4. Use shielded wires for inverter output wires and ground the shielding at the inverter end only.

A number of items of aircraft equipment can create special interference problems, examples include inverters, motors, strobe lights, rotating beacon lights etc. Sometimes trial and error is necessary to eliminate noise and interference problems. The use of shielded wires and physical separation are basic techniques that can be used to prevent or eliminate noise and interference between different aircraft systems and equipment.

5. Static Dischargers

A common cause of noise in aircraft radios and related equipment is P-static interference. Precipitation or P-static noise is caused by static electricity that builds up on an aircraft in flight. The static electricity is produced by friction and can build up to 80,000 volts or more under certain conditions. Friction between the metal skin of the aircraft and particles in the atmosphere is a common cause of P-static. Flying through rain, snow, ice or even dust particles can result in a static charge on the airframe. The exhaust stream of a turbine engine can cause static electricity due to friction between particles in the exhaust and the metal tailpipe. P-static is a greater problem for high speed aircraft because higher speeds produce more friction and higher charges. High speed aircraft usually require more static dischargers to reduce the static charge on the aircraft. Static dischargers are small devices fastened to the extremities of the aircraft that are designed to discharge the aircraft to the atmosphere. They are commonly installed on all classes of aircraft that operate IFR and require all weather radio reception. Static dischargers are fastened to the trailing edges of the primary flight control surfaces: ailerons, elevators and rudders. A recommended installation of static dischargers for a small airplane is shown in figure 4-12. High speed aircraft may have additional static dischargers on the outboard tips of the wing and horizontal stabilizer. The static dischargers reduce the threshold for discharge to the atmosphere so that the voltage on the aircraft is reduced. The locations of the static dischargers for a business jet are shown in figure 4-13.

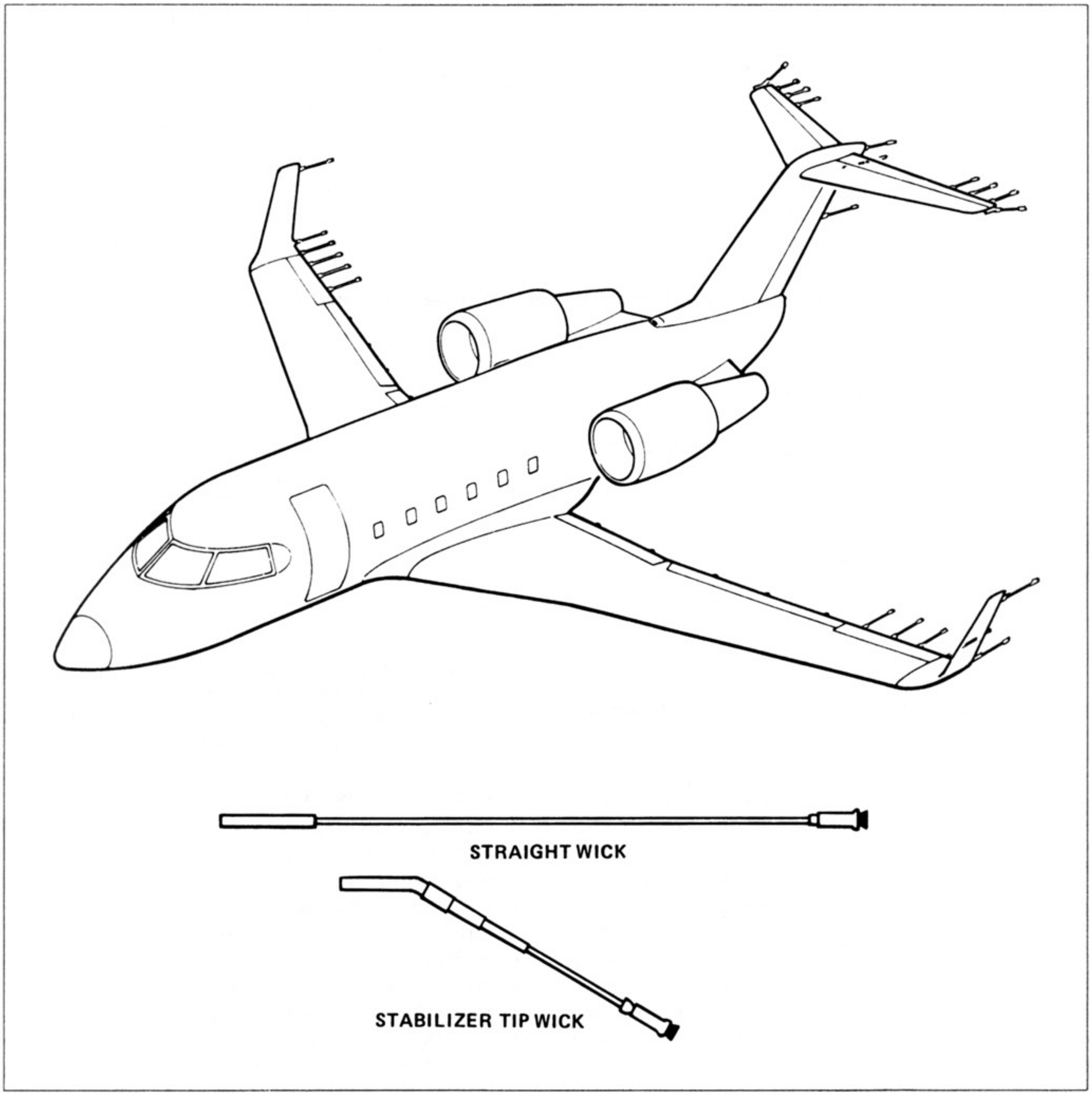

Figure 4-13. The installation of static dischargers for a corporate jet airplane. (Courtesy Canadair Group, Bombardier Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Corona is a term associated with P-static charges on aircraft. Corona refers to the glow that is sometimes visible on the extremities of the aircraft when static electricity is discharging to the atmosphere. St. Elmo's Fire is an older term that means the same thing. The installation of static dischargers will eliminate or reduce corona by controlling the discharges to the atmosphere.

There are three basic types of static dischargers used on aircraft: static wicks, wire braid dischargers and null field dischargers. The static wick is also called a flexible static discharger and is found on low speed aircraft. It consists of a plastic tube or outer covering with a fabric braid inside. The inner braid can be cotton, nylon, or some other material. The inner braid extends beyond the plastic covering where it is fanned out to produce the discharge points. The FAA recommends that one inch of the inner braid should extend beyond the outer cover.

When they become worn, they can be retrimmed to this dimension until they become too short and must be replaced. The inner braid of a static discharger is designed to have some built-in resistance to control the discharge current and further reduce noise. The wire braid static dischargers are also called the semi-flexible type. This type is simply a piece of wire braid made of stainless steel wires as shown in figure 4-15. The wire braid does not have any built-in resistance, so this kind is not as effective as the other two kinds of static dischargers. Jet airplanes normally use the null field discharger which is more rugged than the others for high speed aircraft use. The null field discharger consists of a rigid shaft made of fiberglass or composite materials with very sharp metal points at the aft end. The metal points are sometimes made of tungsten for longer life. Static wicks and null field dischargers are illustrated in figure 4-14. As shown in figure 4-14, the metal points of the null field dischargers are at right angles to the direction of flight. This feature helps to further reduce noise compared to the other kinds of static dischargers.

Static dischargers should be maintained properly to ensure that they will perform their intended function. The attachment to the aircraft must be tight and with good electrical contact. Any corrosion or looseness at the attachment point can create noise in the radios. Damaged or badly worn static dischargers should be replaced with new ones of the approved type. The noise produced by P-static affects the frequency bands of HF and below more than the higher frequency bands. If the pilot complains

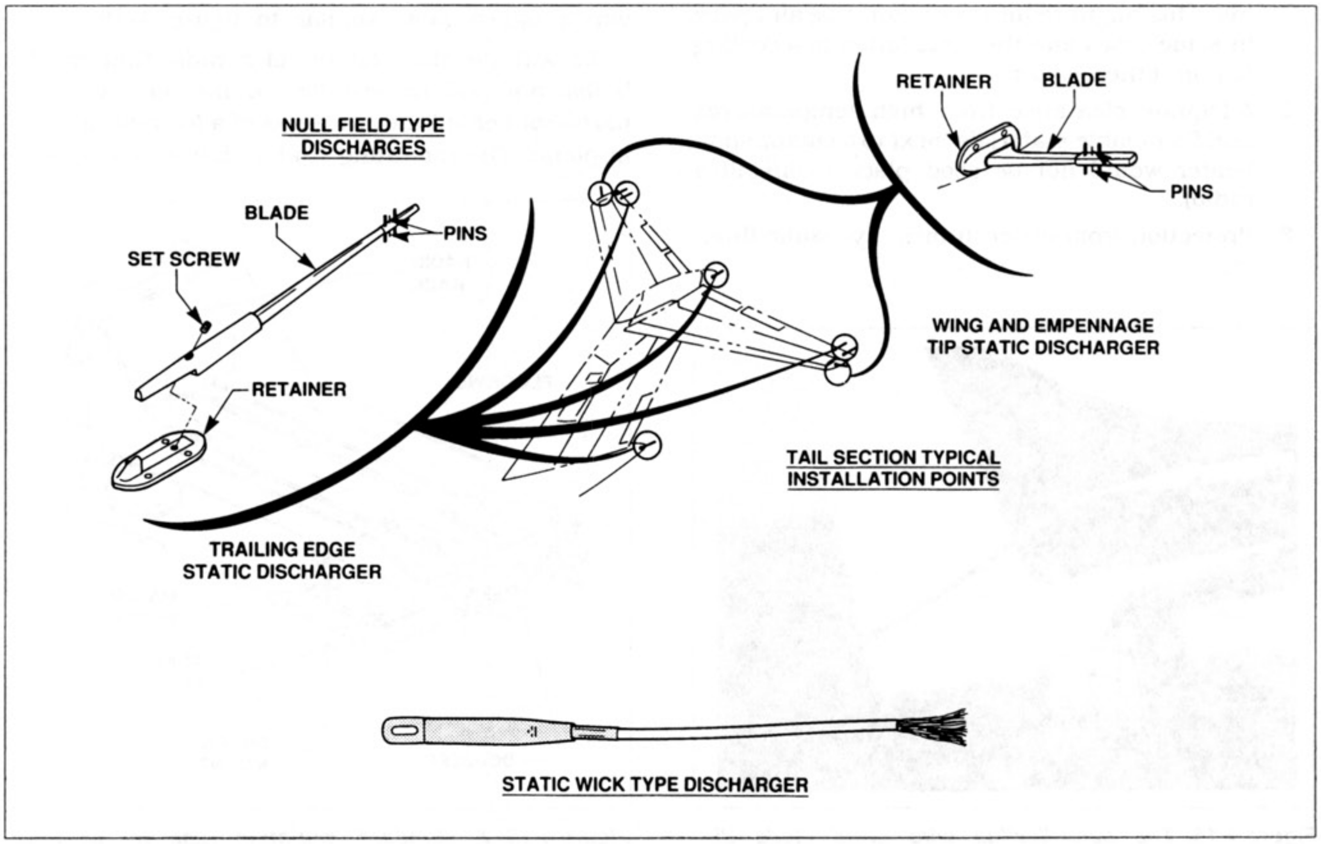
NULL FIELD TYPE DISCHARGES
BLADE
SET SCREW
PINS
RETAINER
TRAILING EDGE STATIC DISCHARGER
RETAINER
BLADE
PINS
WING AND EMPENNAGE TIP STATIC DISCHARGER
TAIL SECTION TYPICAL INSTALLATION POINTS
STATIC WICK TYPE DISCHARGER

Figure 4-14. The Null Field and Static Wick types of static dischargers.

of noise on a radio system that operates at HF or below, the static dischargers should be inspected to determine if the noise is P-static related.

6. Installation Methods

The installation of electronic equipment and radio equipment follows some of the same basic practices that are used for other equipment, but special procedures may be required to prevent interference or other problems that especially affect these types of aircraft systems. The specific instructions of the manufacturer should always be followed when available. Some general recommendations from AC 43.13-1A and -2A will be described here along with some precautions that should be observed for all types of installations.

a. General Precautions

If a standard location and mounting rack is available from the aircraft manufacturer, it should be used to install items of equipment. A standard type of shock mounted installation for electronic equipment is shown in figure 4-16. If this is not available, the installer will have to determine the best location and means of mounting for the equipment. Some of the factors that should be considered when making this type of determination are:

1. Sufficient air circulation to prevent overheating. This might require a certain free air space in some cases and the installation of a cooling fan in other cases.
2. Adequate clearance from high temperatures and flammable materials (next to a combustion heater would not be good place to install a radio).
3. Protection from water, fumes, hydraulic fluid, etc.
4. Protection from damage by baggage or seat deflection.
5. Sufficient clearance to prevent rubbing or striking aircraft structures, control cables, movable parts, etc.
6. Preventing interference and noise. Separate sensitive electronic equipment from inverters, power supplies, strobe lights, motors, etc.
7. If shock mounts will be used, ensure that the equipment does not exceed the weight carrying capability of the shock mounts and install adequate bonding jumpers or straps.

b. Static Loads

Whenever it is necessary for the installer to fabricate a mounting for aircraft equipment, the strength of the mounting should be verified with a load test. An example of a fabricated mounting for aircraft equipment is illustrated in figure 4-17. The equipment installed in aircraft must be able to withstand the acceleration forces or "G" loads that are experienced in flight. In a steep turn, for example, the additional "G" load is felt by the equipment in the aircraft as well as by the wings and other structures.

A simple example of how a static load test might be performed will be explained here. The load factors for the test can be obtained from AC 43.13-2A which has a table similar to figure 4-18.

We will use the example of a radio that weighs 5 lbs. and will be installed in the baggage compartment behind the rear seats of a Normal category airplane. The mounting that is fabricated to hold

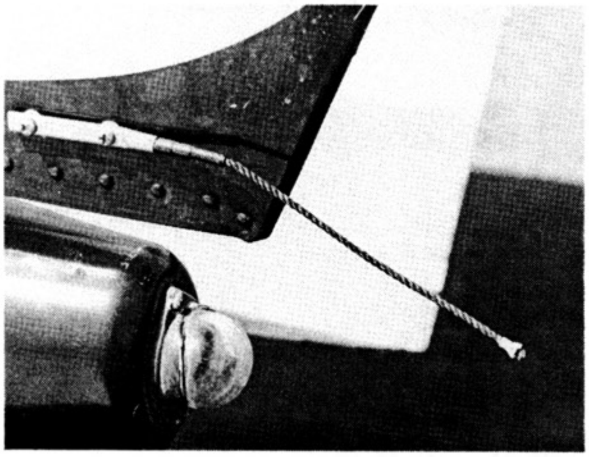

Figure 4-15. The semi-flexible wire braid static dischargers do not have any built-in resistance.

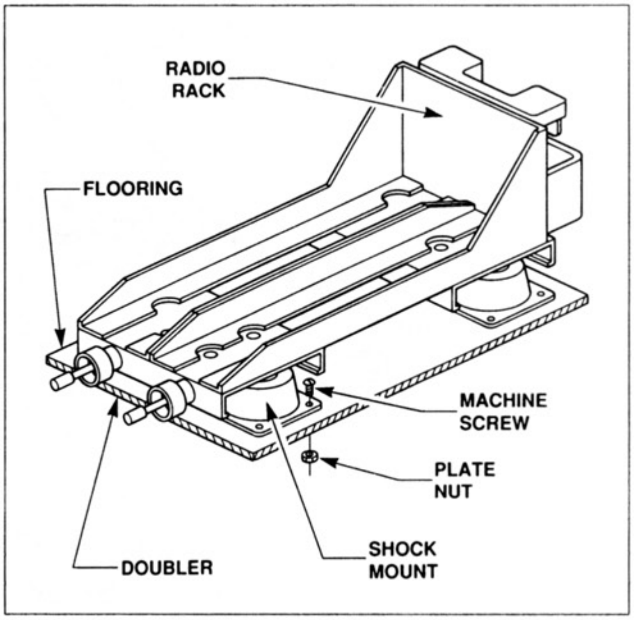

Figure 4-16. A standard mounting rack for avionics equipment that includes shock mounts.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

the radio would be tested by applying loads equal to the weight of the equipment multiplied by the appropriate load factor. The sideward test load would be 7.5 lbs., upward load 15 lbs., forward load 45 lbs. and the downward load 33 lbs. If a location was chosen that was in the nose section of the aircraft and forward of all occupants, a forward test load of 2.0 Gs or 10 lbs. would have been sufficient. The mountings for aircraft equipment must be able to withstand the appropriate level of acceleration forces or load factors that might be experienced in flight. Standard industry practices for rivets, bolts, screws, etc. would be followed to ensure that adequate levels of strength are provided in the fasteners and in mounting brackets and similar parts. The FAA in AC 43.13-2A recommends the use of machine screws and anchor nuts for the removable fasteners to hold aircraft radios in place. Where possible, existing plate nuts should be used or new ones installed. If that is not practical, then machine screws and self-locking nuts can be used.

When radios or other equipment are installed in an instrument panel and the item will extend some distance behind the instrument panel, a brace or support should be installed to the side or back of the equipment to minimize the load on the instrument panel itself. An example of a rear brace for an item of equipment installed in an instrument panel is shown in figure 4-19.

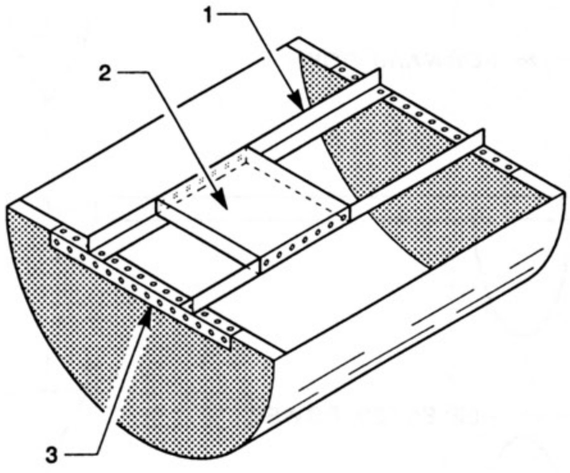

1–BULB ANGLE

2–STIFFENING FLANGE OR ANGLE AT ENDS OF PLATFORM

3–REINFORCEMENT ANGLE FOR BULKHEAD

Figure 4-17. A fabricated mounting unit for avionics equipment.

DIRECTION OF FORCE APPLIED	CERTIFICATION CATEGORY OF AIRCRAFT		
	NORMAL/ UTILITY	ACROBATIC	ROTORCRAFT
SIDEWARD	1.5 Gs	1.5 Gs	2.0 Gs
UPWARD	3.0 Gs	4.5 Gs	1.5 Gs
FORWARD*	9.0 Gs	9.0 Gs	4.0 Gs
DOWNWARD	6.6 Gs	9.0 Gs	4.0 Gs

*When equipment mounting is located externally to one side, or forward of occupants, a forward load factor of 2.0 g is sufficient.

Figure 4-18. Static test load factors that could be used for testing equipment mountings and attachments.

B. Antenna Installations

The antennas found in aircraft radio installations are critical to the proper operation of the radio system. Antennas must be carefully installed and maintained in order to provide the efficiency that is needed for good radio reception and transmission. There are many factors that can affect the efficiency of aircraft antennas. An A&P technician should be familiar with the basic factors that affect the proper operation of an antenna. The inspection and maintenance of aircraft antennas is part of an A&P technicians responsibility in most cases because the antennas are fastened to the skin or other structure of the aircraft. Some of the key concepts that affect antenna operation will be covered in this section.

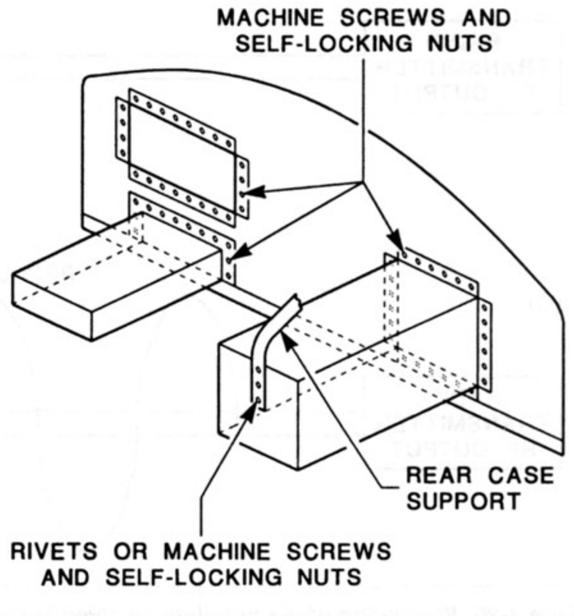

Figure 4-19. Example of the installation of a rear brace or support for radio equipment installed in an instrument panel.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

1. Standing Wave Ratio

The standing wave ratio is a measure of the efficiency of an antenna installation. The standing wave ratio (SWR) is also referred to as voltage standing wave ratio (VSWR). In order to demonstrate the principle of VSWR, we will consider what would happen if just one sine wave was sent down a transmission line from a radio transmitter. Figure 4-20(A) shows this sine wave traveling from left to right. If the transmission line had an infinite length, the sine wave signal would eventually be reduced to zero by line resistance. In an actual installation, the transmission line is of a limited length and terminates at the antenna. The purpose of the antenna is to transform the sine wave signal into radio waves, but this can never be accomplished with 100 percent efficiency. The result of this less than perfect efficiency is that some of the energy is reflected back toward the transmitter from the antenna end of the transmission line. This is illustrated by figure 4-20(B) which shows some energy being reflected back toward the transmitter and moving from right to left. The output of the transmitter is not just one sine wave at a time, but a continuous series of sine waves. The reflected waves will combine with the transmitter output waves to produce standing waves on the transmission line as indicated by figure 4-20(C). A calculation based on the relationship between forward power and reflected

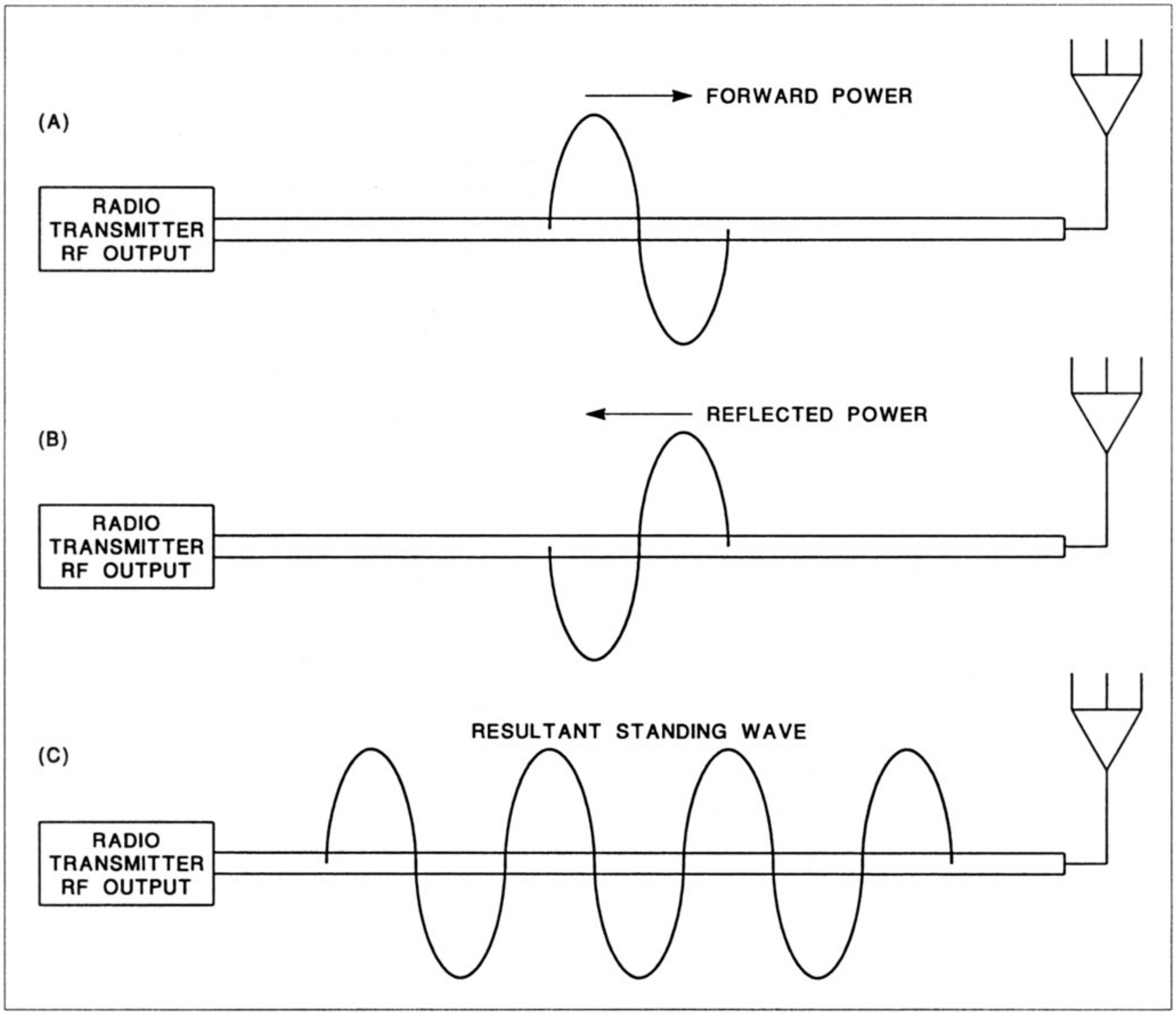
(A)
FORWARD POWER
RADIO TRANSMITTER RF OUTPUT
(B)
REFLECTED POWER
RADIO TRANSMITTER RF OUTPUT
(C)
RESULTANT STANDING WAVE
RADIO TRANSMITTER RF OUTPUT

Figure 4-20. Illustration of the principle involved in the standing wave ratio for an antenna installation.
(A) Forward power.
(B) Reflected power.
(C) Resultant Standing Wave.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

power gives the voltage standing wave ratio, which is a measure of efficiency. With a perfect antenna installation, there would be zero reflected power and the VSWR would be 1:1 or simply 1. In an actual aircraft antenna system, the lowest VSWR is the most efficient. Typical values for VSWR of aircraft antennas are in the range of 1.1 to 5.0 for the various types of antennas. Manufacturer's catalogs usually list the VSWR for antennas so that the relative efficiency of different types can be compared when selecting an antenna. The listing of VSWR in the specifications for an aircraft antenna can be seen in figure 4-21.

The example given here was for the VSWR of a transmit antenna, but the manufacturer's data also lists the VSWR for receive only antennas. If the transmission line or coaxial cable that connects the antenna is in good condition and properly suited to the installation, the VSWR is affected by the antenna itself. However, if there is a fault in the coaxial cable the VSWR will go up significantly, which reduces efficiency. Special types of wattmeters and VSWR meters can be used to measure the VSWR of an aircraft antenna installation for troubleshooting purposes.

2. Coaxial Cables and Connectors

Coaxial cables are required for the antenna connections on most aircraft radios because of the RF frequencies that are used. A coaxial cable is shown in figure 4-22. The proper installation and maintenance of coaxial cables is very important since large losses can occur if a fault is present. Coaxial cables should be rejected if they have become dented or if kinks are found. Any distortion or crushing which causes the cable to be oval in shape or flattened are also cause for rejection. If abrasion or rubbing has exposed or damaged the wire braid, the cable should be replaced. Coaxial cable should be supported by clamps about every 2 ft. to help prevent damage. A good rule of thumb for coaxial cable bend radius is to use a minimum bend radius of 10 times the cable diameter. This will help to reduce the possibility of kinks from sharp bends. Special types of end connectors are used with coaxial cable and they come in a number of different styles. Some can be removed and reused and other types are crimped or swaged on and cannot be reused. When installing and removing coaxial cable connections, care should be used to prevent damage to the connectors. If corrosion is found on connectors, it is usually better to replace them rather than try to clean them. Even small amounts of corrosion or corrosion pits can cause a signal loss. Figure 4-23 shows a reusable coaxial cable connector. When installing this type, the wire braid should be carefully spread out over the braid clamp and breakage of the wires should be avoided. The connector should be assembled carefully to provide tight connections with good electrical contact and to avoid distorting the coaxial cable or the connector itself. If it is necessary to solder a connector pin onto the center conductor, only an approved electrical solder should be used—never use acid core solder or acid flux on electrical connectors. An acceptable solder is 60/40 rosin core solder. Great care must be used in soldering to prevent excessive heat damage to the coaxial cable insulation materials.

Some antenna cables are matched to the radio and antenna and should not be shortened or spliced.

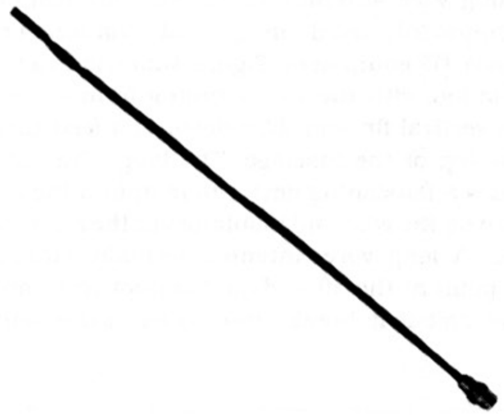

SPECIFICATIONS

V.S.W.R.	2.0:1
IMPEDANCE	50 ohms
POWER	40 watts
WEIGHT	0.2 lbs.
CONSTRUCTION	Whip
HEIGHT	14.0 in.
ELEMENT	Open

Figure 4-21. Antenna manufacturers catalogs usually list the V.S.W.R. for each antenna. (Courtesy of Dayton-Granger, Inc.)

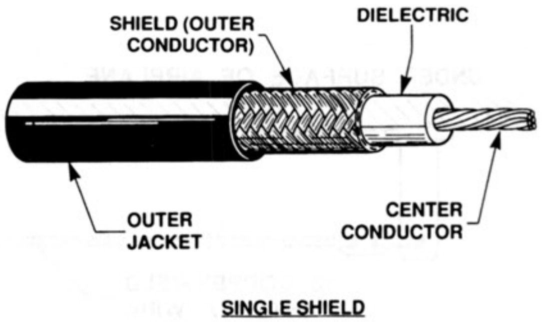

Figure 4-22. The parts of a coaxial cable for antennas.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

This is true for some ADF antenna leads, for example. On other installations the antenna coax should be kept as short as possible and routed as directly as possible to reduce line loss.

The specific antenna or radio manufacturers installation instructions should be followed carefully in this area since there are many different procedures that may apply depending on the specific installation.

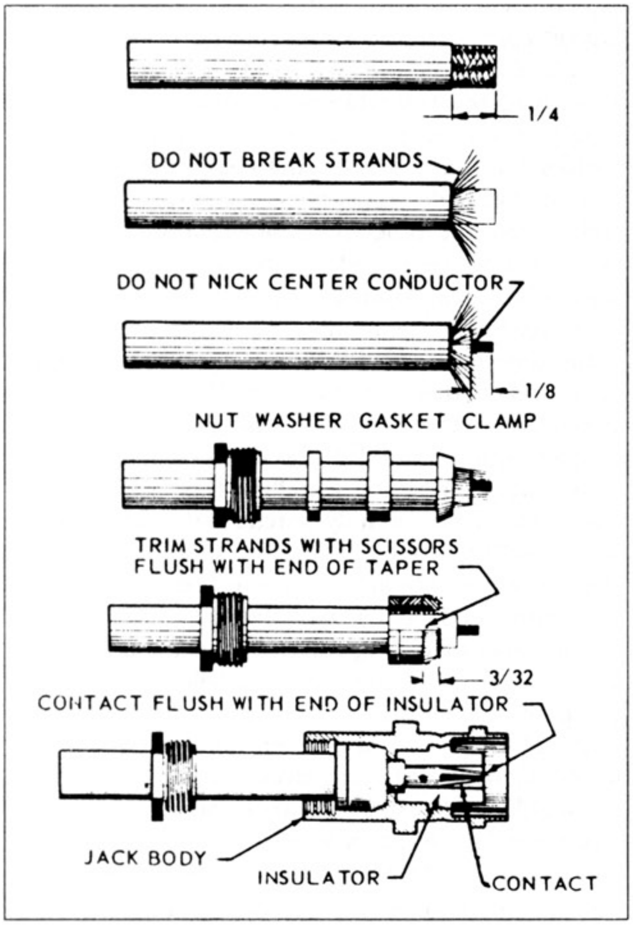
1/4
DO NOT BREAK STRANDS
DO NOT NICK CENTER CONDUCTOR
1/8
NUT WASHER GASKET CLAMP
TRIM STRANDS WITH SCISSORS FLUSH WITH END OF TAPER
3/32
CONTACT FLUSH WITH END OF INSULATOR
JACK BODY
INSULATOR
CONTACT

Figure 4-23. Installation procedure for a reusable coaxial cable connector.

3. Wire Antennas

A wire antenna is a length of wire that is supported by masts and attachments above or below the aircraft fuselage. They are found most often on smaller aircraft and older aircraft. Jet airplanes seldom use wire antennas because of the vibration and increased chance of damage at high speeds. The type of wire used is most often a copper coated steel wire that is a solid single strand. Wire with an outer covering of insulating material is superior to non-insulated wire in reducing noise caused by P-static.

A type of wire antenna that is seldom used today is the trailing wire antenna. The trailing wire antenna was a roll of wire on a drum in the aft fuselage which could be extended out the back of the aircraft in flight. It was very common in the 1930s and 1940s for HF communications radios. The advantage was that 200 ft. or more of wire could be extended out the back of the airplane for better radio performance. The disadvantages were the added complexity and weight of the mechanism to extend and retract the antenna. It is not suitable for high speed aircraft and is rarely used on modern aircraft.

The wire type marker beacon antenna is shown in figure 4-24. This type may still be found on small airplanes. It is about 4 ft. long and fastened to standoff and support masts on the bottom of the aircraft. The minimum fuselage separation should be 6 inches.

A long wire antenna for HF communications is still commonly used on general aviation aircraft that have HF equipment. Figure 4-25 shows a typical installation with the wire running from a wing tip to the vertical fin and then down to a feed-through on the top of the fuselage. The long wire antenna includes a tensioning device to maintain the proper tension on the wire and insulators at the appropriate points. A long wire antenna normally employs a weak point at the aft end so it cannot wrap around the aircraft if it breaks due to excessive tension.

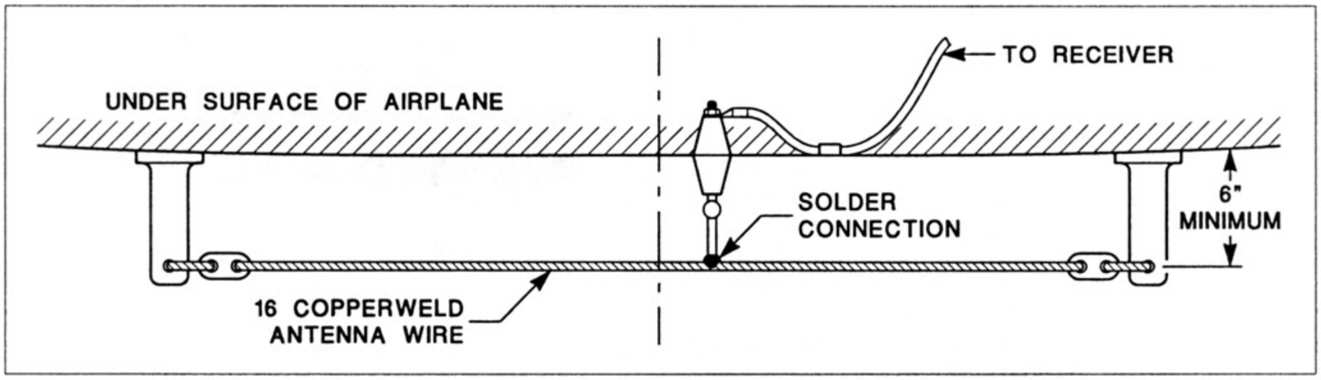
TO RECEIVER
UNDER SURFACE OF AIRPLANE
SOLDER CONNECTION
6" MINIMUM
16 COPPERWELD ANTENNA WIRE

Figure 4-24. A wire-type marker beacon antenna.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The most common use of wire antennas on modern aircraft is as an ADF sense antenna. These will be described in the next section.

4. ADF Antennas

All aircraft ADF receivers require two antennas, the loop antenna and the sense antenna. The loop antenna is the directional antenna and the sense antenna is needed to eliminate the ambiguity caused by the two nulls in the reception pattern. Air carrier jets have an ADF antenna that combines the loop and sense antennas in one housing that is a low profile or flush mount and it is installed on the top or bottom of the fuselage. The ADF antennas on general aviation aircraft come in a greater variety and are most often separate loop and sense antennas.

The loop antenna that is rotated by an electric motor is still used, but is being replaced by the type that rotates the signal rather than the antenna itself. A motor driven loop antenna for installation inside a housing is shown in figure 4-26. The newer non-rotating types are usually contained in a teardrop shaped streamlined housing that installs on the top or bottom of the aircraft as shown in figure 4-27. The sense antennas used with the dual antenna installations are either the wire type or whip type. The whip type sense antenna is a metal rod about 4 ft. long and installed on the top or bottom. It is still found on some helicopters were there isn't enough room for a long wire sense antenna, this is shown in figure 4-28. The long wire sense antenna is about 15–20 ft. long and most often installed using the vertical fin as the aft anchor point to gain more fuselage clearance. The recommended minimum clearance from the fuselage is 12". A top installation for a sense wire is shown in figure 4-29. The sense wire can be installed on the bottom as shown in figure 4-30 if adequate ground and fuselage clearance can be obtained. Like the long wire HF antenna, the ADF sense wire will use masts, tension units and weak links as part of the installation.

In order to give accurate navigational information, ADF antennas must be installed and calibrated correctly. The loop antenna normally needs to be installed close to the electrical center of the sense antenna to give accurate indications of station passage. This relationship is illustrated in figure 4-31.

Both the loop and sense antennas can be installed on the top or the bottom, but they must have the

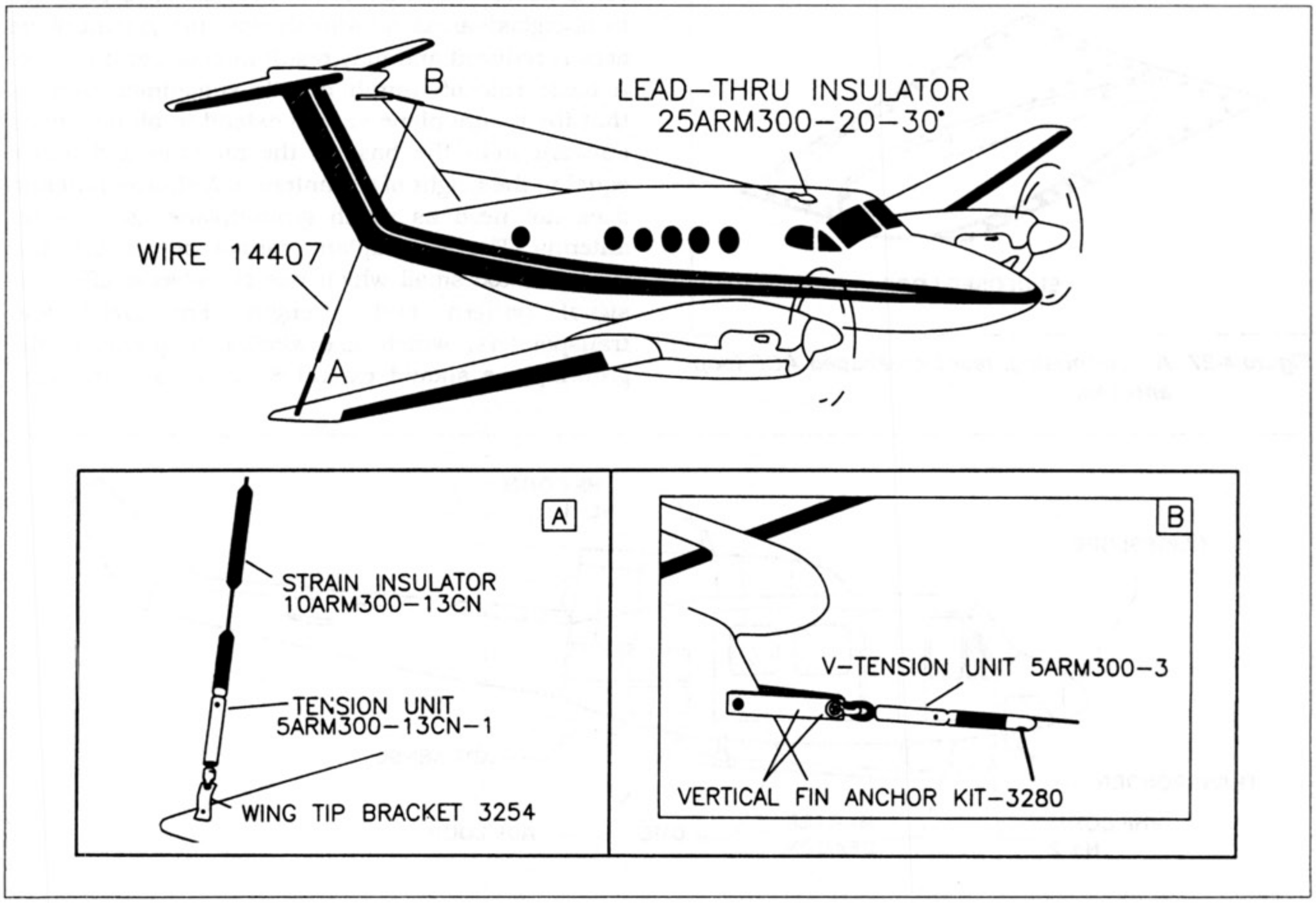

Figure 4-25. A long wire HF comm antenna installation. (Courtesy Dayton-Granger Inc)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

correct relationship to each other for accurate readings to be obtained. Since the ADF antenna system is highly directional, it must be calibrated to give the correct indications of ground station direction. This is called the check for quadrantal error or the calibration check. Whenever an antenna is installed or any change is made which could affect the accuracy of the ADF, a check for quadrantal error should be performed. The checks can be made on the ground, but should always be confirmed with a flight check. To perform the ground check a nearby NDB of known location is tuned in and the bearing is checked and adjusted at least every 45° as the aircraft is turned on the ground. The flight check involves locating geographical points on the ground with known bearings from the NDB and flying the aircraft over those locations to confirm the accuracy of the ADF bearing information. This flight check should be performed at low altitude to reduce errors in established the aircraft position accurately.

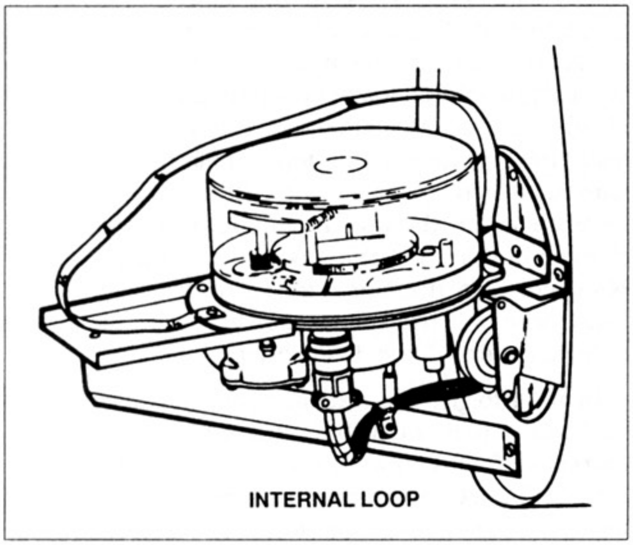

Figure 4-26. A motor-driven ADF loop antenna for internal installations.

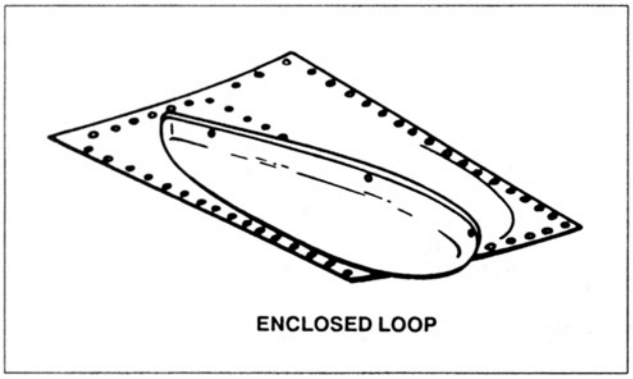

Figure 4-27. A non-rotating, teardrop-shaped ADF loop antenna.

5. Groundplane Considerations

When a 1/4 wave, Marconi-type antenna is installed on an aircraft, an adequate groundplane or counterpoise is required for proper operation. The aircraft systems that use 1/4 wave antennas are VHF communications, ATC transponder, DME and UHF radiotelephone. When these antennas are installed on metal skinned aircraft, the metal skin supplies the groundplane. If the antenna is installed too close to fiberglass areas or windshields, the groundplane area is reduced and may result in poor performance. A basic rule of thumb that is sometimes used is that the groundplane should extend in all directions outward from the base of the antenna a distance equal to the height of the antenna. A shorter antenna does not need as much groundplane as a longer antenna. The groundplane cannot be too big, but it can be too small which has an adverse effect on signal pattern and strength. For DME and transponders, which use similar frequencies, the groundplane should extend 8–12″ in all directions

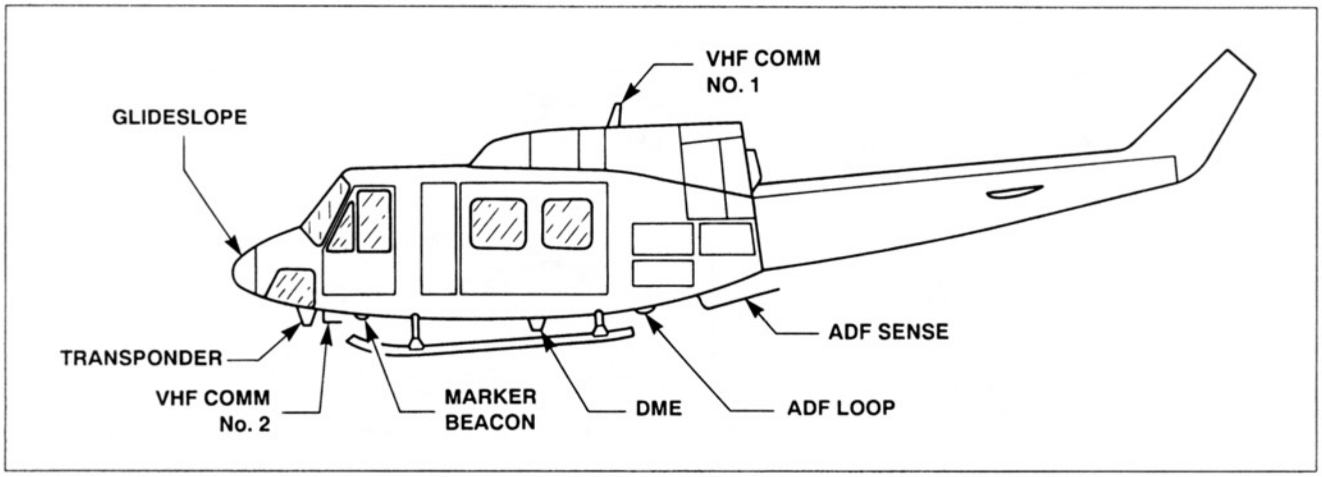

Figure 4-28. Helicopter antenna installations are difficult because of limited skin area and limited ground clearance.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

from the antenna base. For VHF communications antennas a groundplane that extends 24" in all directions is desirable. These sizes would give a groundplane that is a little larger than if the length of the antenna was used as the dimension. It is not always possible to supply a large enough groundplane when installing antennas on aircraft with limited metal skin area such as small helicopters, but the groundplane area should always be considered and provided for to the extent possible.

If it is necessary to install these types of antennas on aircraft with non-metal skin, a groundplane must be provided by the installer. This usually means installing metal foil strips or wire mesh fastened on the inside of the aircraft covering. The same rules would apply as to desirable lengths. An example of the use of a foil strip groundplane is seen in figure 4-32.

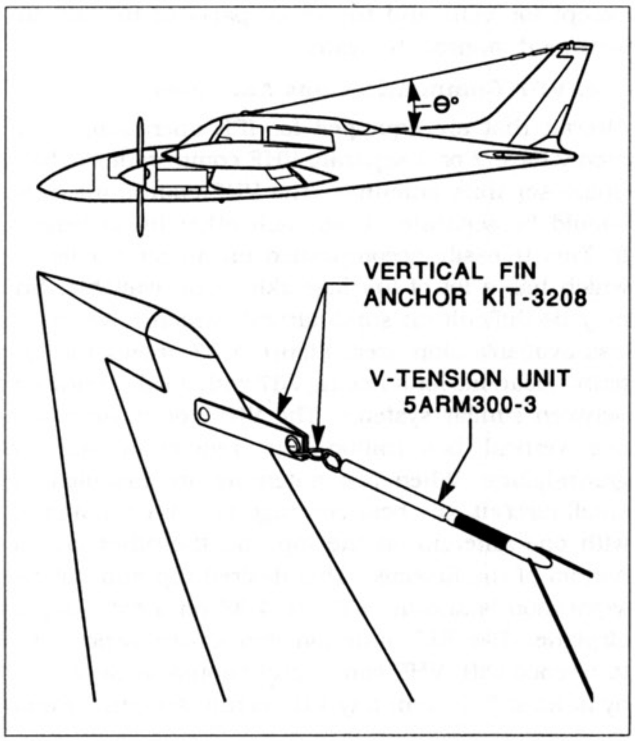
θ°
VERTICAL FIN ANCHOR KIT-3208
V-TENSION UNIT 5ARM300-3

Figure 4-29. Top-mounted ADF sense wire antenna. (Courtesy Dayton-Granger Inc.)

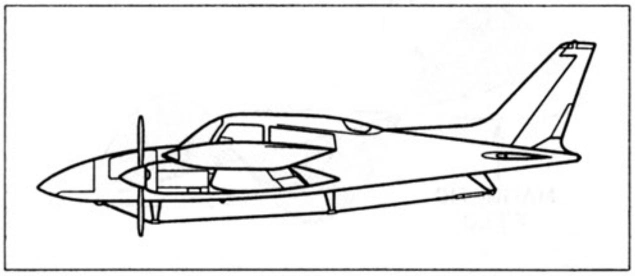

Figure 4-30. Bottom-mounted ADF sense wire antenna. (Courtesy Dayton-Granger Inc.)

When installing 1/4 wave antennas, it is recommended that all grease, dirt and paint be removed from the skin area under the base of the antenna. Some avionics experts recommend that a gasket not be used so that the base of the antenna contacts the skin of the aircraft. Whether or not a gasket is used, the skin should be cleaned and stripped and a sealant applied around the base of the antenna after installation.

The installation of antennas to the skin of aircraft requires that some additional reinforcement be given

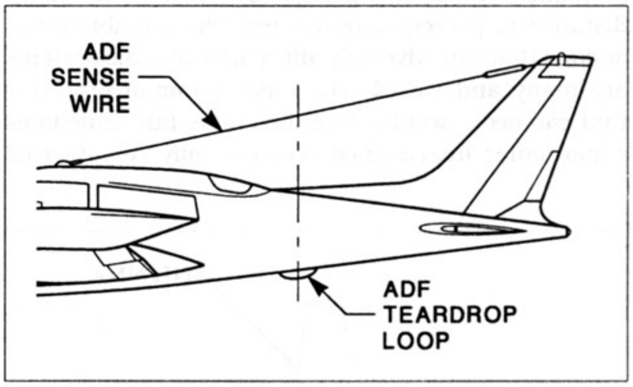
ADF SENSE WIRE
ADF TEARDROP LOOP

Figure 4-31. The ADF loop antenna should normally be installed near the electrical center of the sense wire antenna.

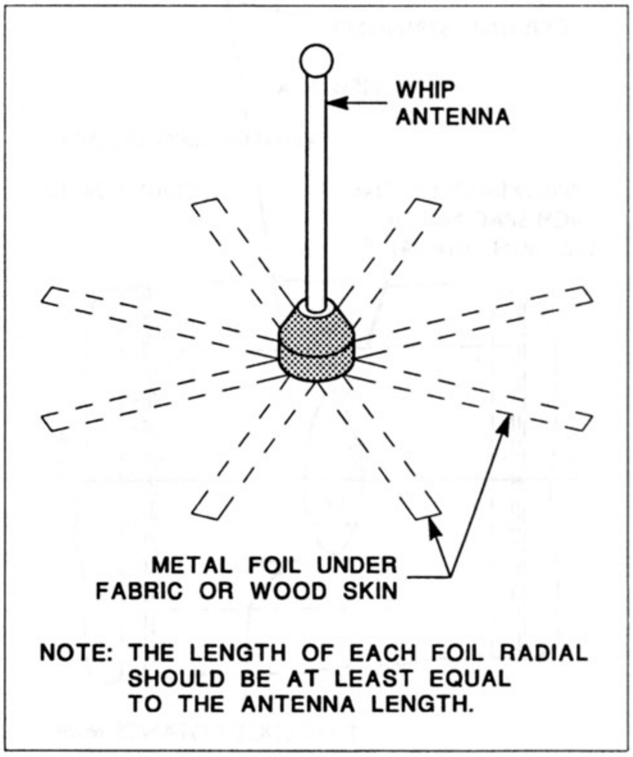
WHIP ANTENNA
METAL FOIL UNDER FABRIC OR WOOD SKIN
NOTE: THE LENGTH OF EACH FOIL RADIAL SHOULD BE AT LEAST EQUAL TO THE ANTENNA LENGTH.

Figure 4-32. When installing Marconi antennas on an aircraft with non-metal skin, a groundplane must be provided.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

to preserve the strength of the aircraft structure. The use of a doubler as shown in figure 4-33 will reinforce the aircraft structure and provide the additional support needed for antenna drag loads.

6. Reducing Antenna Interference

A very important factor in the proper performance of aircraft antennas is the prevention of interference between one system and another. Interference can also occur between a radio system antenna and other components of the aircraft. A basic consideration is that antennas for systems that operate on similar frequencies must be separated by a certain minimum distance to prevent interference. The possible interactions that can adversely affect aircraft radio systems are many and varied. The more common problems that can occur will be described here, but sometimes a particular interference problem may require trial and error to eliminate the cause of the antenna interaction.

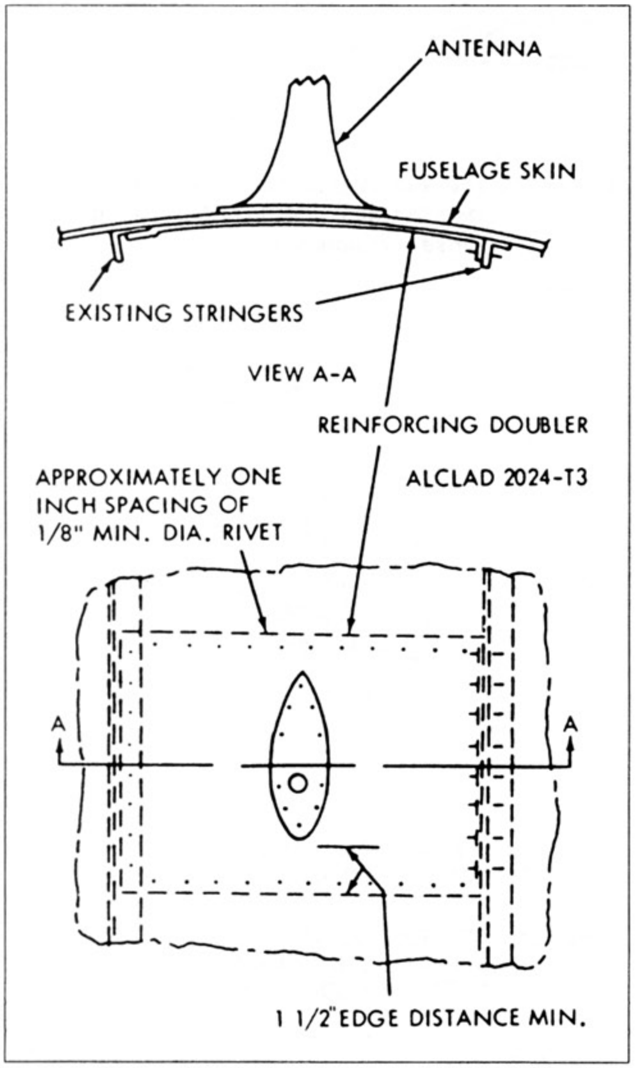

Figure 4-33. A reinforcing doubler should be installed inside the skin at the base of the antenna.

The important factors that affect mutual interference are frequency and wavelength, polarization and type of modulation. The operating frequencies for the various radio systems are listed in the frequency chart in chapter 3. The polarization of radio waves is based on the orientation of the electric field relative to the earth's surface. The field orientations for vertical and horizontal polarization can be seen in figures 4-34 and 4-35. The antenna installed on the aircraft needs to have the proper polarization relative to the ground based antenna for optimum performance—particularly at frequencies above HF. Figure 4-36 gives the polarization for the various types of aircraft radio systems.

From the information in figure 4-36, it can be seen that all the systems use vertical polarization except for VOR and the three parts of the ILS instrument approach system.

a. VHF Communications Antennas

Aircraft that are equipped for IFR operations commonly have 2 or 3 separate VHF comm radios which utilize separate antennas. The VHF comm antennas should be separated from each other by at least 5 ft. This is easily accomplished on an air carrier jet which has a lot of fuselage skin area available, but may be difficult on small aircraft which have much less available skin area. Figure 4-37 shows the antenna locations for a Boeing 767 with good separation between similar systems. The VHF comm antennas use vertical polarization and require a suitable groundplane. When two antennas are installed on small aircraft, the best coverage is usually obtained with one antenna on the top and the other on the bottom of the fuselage. This desired top and bottom separation is shown in figure 4-38 on a twin-engine airplane. The ELT antenna can cause serious interference with VHF comm and should be separated by at least 5 ft. from any VHF comm antenna. Radio

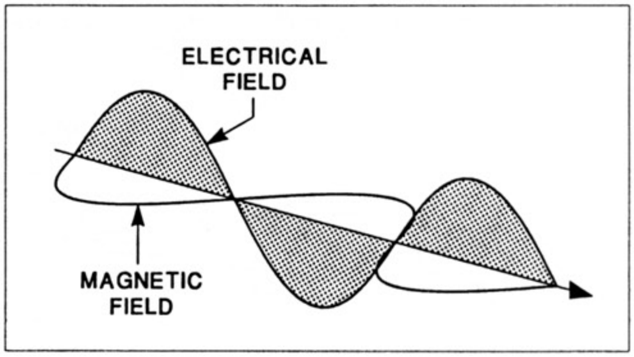

Figure 4-34. When an EM wave has vertical polarization, the electric field is in the vertical plane.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

interference can be caused by parts of the aircraft as well as by other antennas. The vertical fin of an aircraft can cause significant signal blockage to any VHF comm antenna that is installed too close. A top mounted VHF comm antenna that is installed closer than 5 ft. to the vertical fin will result in blockage and poor radio reception and transmission to the rear of the aircraft. The VHF comm antenna is a 1/4 wave Marconi antenna which must have an adequate groundplane or counterpoise for proper operation. A common mistake is the installation of a VHF comm too far forward on the upper fuselage. If it is less than 24″ from the top of the windshield, the signal pattern can be distorted by the lack of groundplane in the forward direction.

b. DME and Transponder Antennas

These two antennas are treated as equals because they use similar frequencies, polarization and modulation. The antennas used for these two systems are 1/4 wave Marconi antennas with vertical polarization and they both transmit and receive. Since the wavelength is shorter at higher frequencies, the minimum separation distance is less than that for VHF comm antennas. The DME and transponder antennas should be separated from each other by at least 2 ft. and an adequate groundplane must be provided around the base of the antenna. These antennas are normally installed on the bottom of the aircraft to prevent signal blockage by the fuselage. A top mounted antenna may be used on a narrow portion of the aircraft that will not cause significant blockage. The top of the tail boom on a helicopter can be an acceptable location.

c. VOR and Localizer

VOR antennas are most often installed on the vertical fin of the aircraft. This gives good reception characteristics from all directions on most aircraft. On small aircraft, the VOR antenna is sometimes mounted on the top of the fuselage. If the VOR antenna is mounted too far forward, a propeller modulation problem can occur. When signals are being received from the front of the aircraft, the radio wave is chopped by the propeller blades. At certain RPMS, this can cause serious propeller modulation interference. The cure for this involves changing propeller RPM or relocating the antenna. Small aircraft often use the same antenna for both VOR and localizer reception. This is practical because the two systems operate on similar frequencies. When the localizer is being used for an instrument approach, the signals are always received from the front of the aircraft. On a large aircraft, it is not possible to use the tail mounted VOR antenna for localizer reception because of fuselage blockage. These aircraft will use a separate localizer antenna or antennas that are mounted in the nose section inside the radome for the weather radar.

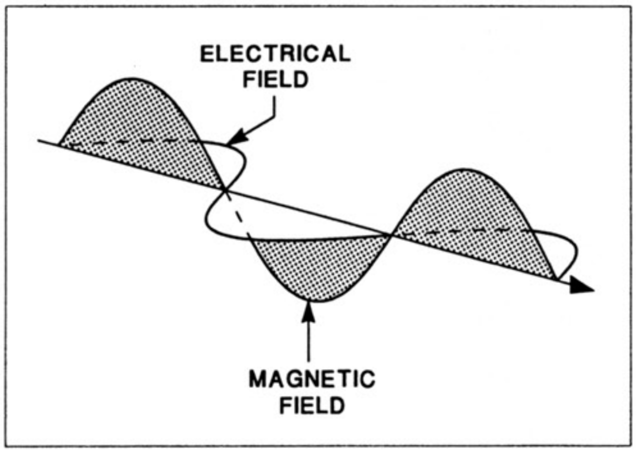

Figure 4-35. When an EM wave has horizontal polarization, the electric field is in the horizontal plane.

The location of the VOR and localizer antennas usually provides sufficient separation that interference from other antennas is not a problem. If a VHF comm or other antenna is mounted closer than 5 ft. from the VOR, it can cause some interference depending on the type of VHF comm antenna used.

d. Glideslope Antennas

Like the localizer, the signals from the ground transmitters for the glideslope are always received from the front of the aircraft. Some small aircraft use the VOR antenna to receive glideslope signals as well as localizer signals. The glideslope operates on frequencies that are the third harmonic of VOR frequencies. This means that the glideslope frequencies are three times the frequencies for VOR. A special antenna coupler is used so that the VOR antenna can supply two separate VOR and localizer receivers and also supply signals for the glideslope receiver.

RADIO SYSTEM	RECEIVE, TRANSMIT OR BOTH	POLARIZATION
LORAN C	RECEIVE	VERTICAL
ADF	RECEIVE	VERTICAL
VHF COM	BOTH	VERTICAL
DME & TRANSPONDER	BOTH	VERTICAL
ELT	TRANSMIT	VERTICAL
VOR & LOCALIZER	RECEIVE	HORIZONTAL
MARKER BEACONS	RECEIVE	HORIZONTAL
GLIDESLOPE	RECEIVE	HORIZONTAL

Figure 4-36. The polarization for various types of aircraft radio systems.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

The same fuselage blockage problems occur on large aircraft for both localizer and glideslope reception. The glideslope antenna or antennas for air carrier jets are installed inside the radome on the nose of the aircraft. Aircraft that do not have a nose radome can utilize a separate glideslope antenna that is mounted on the forward fuselage on either the top or bottom. Blockage of signals by the fuselage or other parts of the aircraft is the primary consideration in locating localizer and glideslope antennas. Interference from other antennas is not as great a problem with these systems as it is for some other radio systems.

e. Loran C and Omega

Loran C and Omega system antennas are receive only antennas and they operate at frequencies that are widely separated from those of most other aircraft radios. The major sources of interference for these radio systems are P-static noise and noise from aircraft electrical systems. The Loran C and Omega antennas can be mounted on the top or the bottom of the aircraft. The best location for these types of antennas is based on preventing interference from aircraft motors, generators, power supplies and similar systems. The proper installation and maintenance of bonding jumpers and static dischargers is critical to ensure good performance from these lower frequency radio systems.

f. ADF Antennas

The primary consideration in locating ADF antennas is to obtain the proper relationship between the loop and sense antenna to ensure accurate indications of station direction. The ADF antennas can be installed with both loop and sense antennas on the top of the fuselage, both on the bottom or one on the top and one on the bottom.

The most common installation on small aircraft is with a wire sense antenna on the top and the loop antenna on the bottom of the aircraft. In any

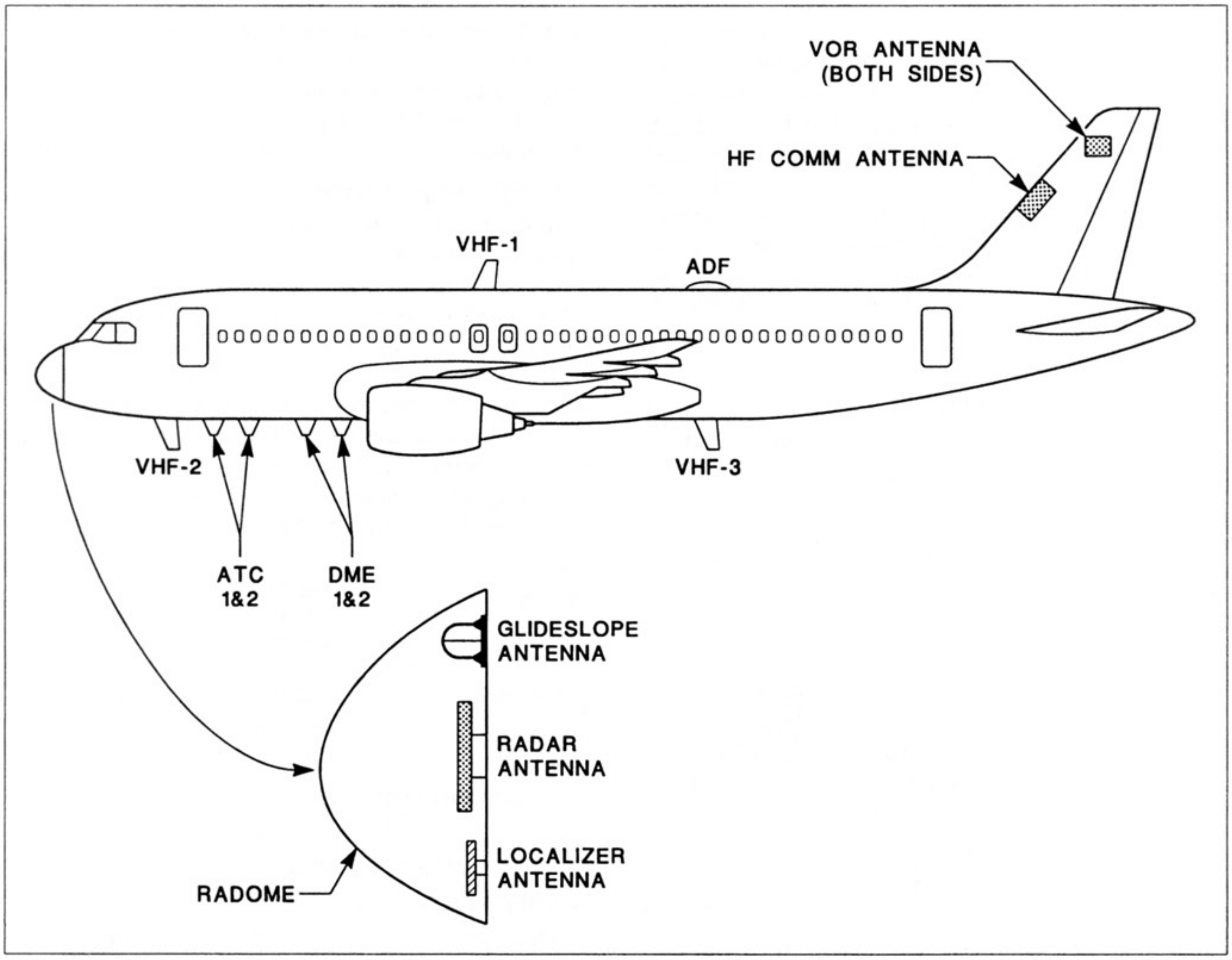

Figure 4-37. Antenna installations on modern air carrier jets often include localizer and glideslope antennas inside the radome and flush mount VOR and HF comm antennas in the vertical fin.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

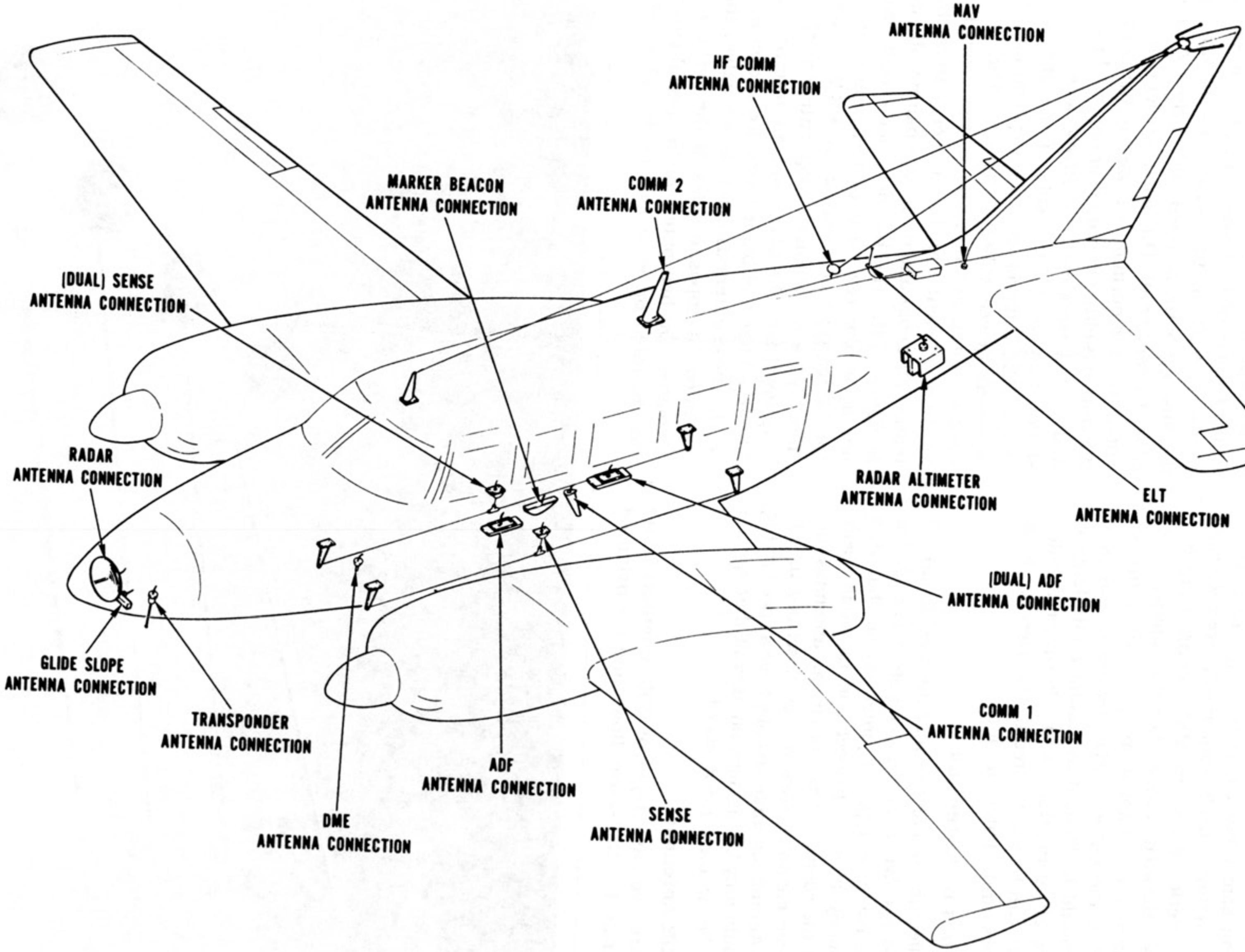
NAV ANTENNA CONNECTION
HF COMM ANTENNA CONNECTION
MARKER BEACON ANTENNA CONNECTION
COMM 2 ANTENNA CONNECTION
(DUAL) SENSE ANTENNA CONNECTION
RADAR ANTENNA CONNECTION
RADAR ALTIMETER ANTENNA CONNECTION
ELT ANTENNA CONNECTION
(DUAL) ADF ANTENNA CONNECTION
GLIDE SLOPE ANTENNA CONNECTION
TRANSPONDER ANTENNA CONNECTION
COMM 1 ANTENNA CONNECTION
ADF ANTENNA CONNECTION
DME ANTENNA CONNECTION
SENSE ANTENNA CONNECTION

Figure 4-38. ***The antenna installations for a twin-engine airplane. (Courtesy Piper Aircraft Corp.)***

case, the loop antenna must be located in the electrical center of the sense antenna for accurate readings. The ADF antenna system is a directional antenna system and interference from parts of the aircraft can sometimes cause bearing errors. This is one reason that a check of quadrantal error should always be performed when ADF antennas are installed or relocated. Proper bonding jumper and static discharger installations are important to prevent P-static noise in ADF receivers. ADF antennas should be located to minimize interference from aircraft generators and alternators. Filter capacitors can be used to reduce interference from alternators and similar devices.

7. Types of Antennas

Many different types of antennas are used in aircraft radio systems. Aviation technicians should be familiar with the common types of antennas so that they can properly identify, inspect and maintain them. Some of the common types of aircraft antennas and their basic characteristics will be described in this section. Aircraft antennas usually have a speed rating and should only be installed on aircraft that operate at and below their rated speed.

a. VOR Antennas

There are two basic types of VOR antennas found on aircraft: the half-wave dipole and the balanced loop types. The half-wave dipole antenna is a "V" shaped antenna that has a figure eight-shaped reception pattern. This kind of antenna is shown in figure 4-39. The antenna has two metal rods in the shape of the letter "V" or a fiberglass covered element made of thin sheet metal. It is installed on the aircraft on the vertical fin or on top of the fuselage with the open end of the "V" pointed either forward or aft. The figure 8 reception pattern works well for normal VOR airway flying because the station is either in front of or behind the aircraft. It does not work well for RNAV when the VOR station may be off the side of the aircraft. The dipole VOR antenna requires a special impedance matching device called a "balun". The balun is located at the antenna end of the coaxial cable for more efficient transfer of energy from the antenna to the coax and receiver. A balun is illustrated in figure 4-40. The balanced loop VOR antenna has a circular reception pattern and is therefore the better type of antenna for RNAV. There are three types of balanced loop antenna: the open loop towel bar, the blade and the internal mount. The towel bar and blade types are shown in figure 4-41. These antennas come in two halves that are mounted on opposite sides of the vertical fin on airplanes. On helicopters or in special cases they are mounted on each side of the aft fuselage or

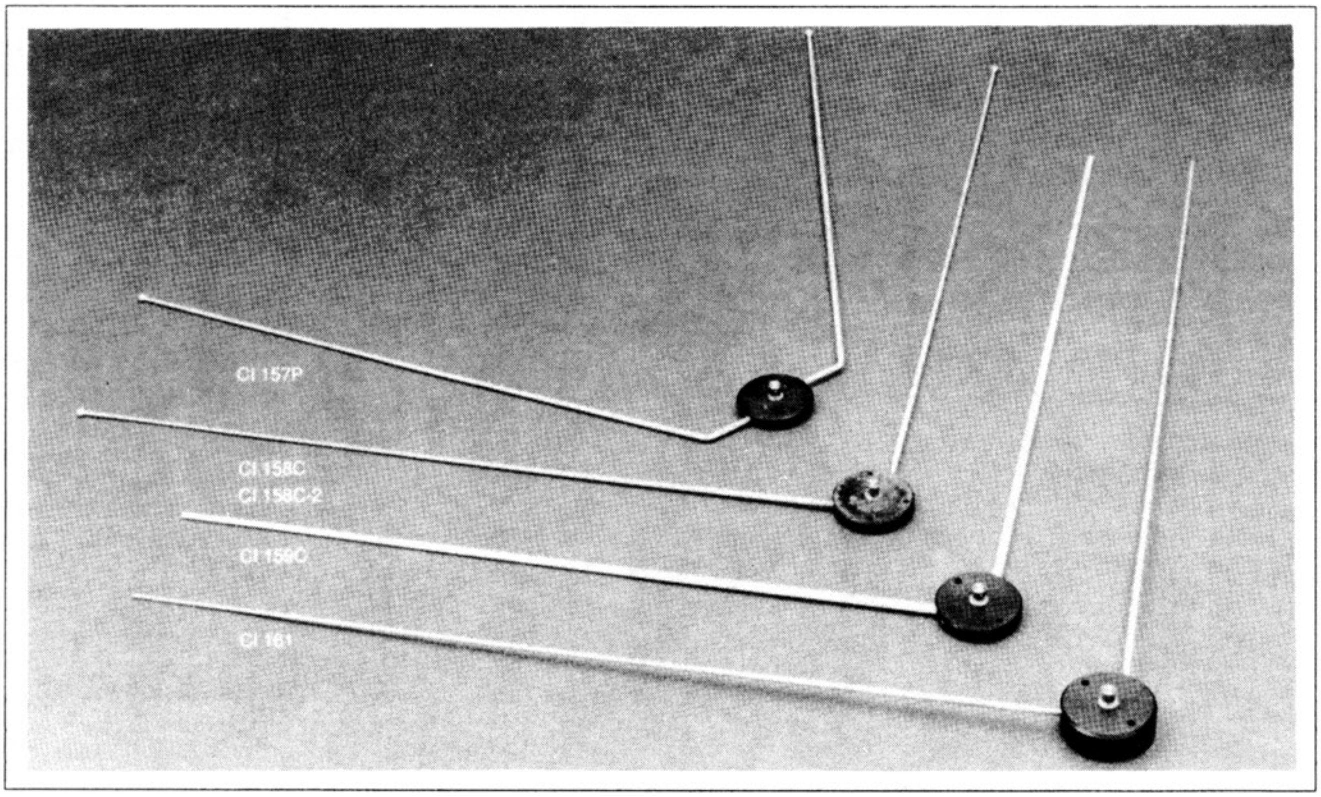

Figure 4-39. Hertz dipole "V" type antennas for VOR reception. (Courtesy Comant Industries Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

tail boom. The blade-type, balanced loop VOR antenna has a higher speed rating than the towel bar or V-type and is used on bizjets and similar aircraft. Air carrier jets use a VOR antenna that is mounted inside the vertical fin with non-metallic flush covers on each side. This kind of antenna is shown in figure 4-37.

b. Localizer

Small airplanes usually do not have a separate localizer antenna, the VOR antenna is used to receive localizer signals. On air carrier jets and similar aircraft, the large fuselage can cause blockage of the localizer signals so a separate localizer antenna is installed. A type of separate localizer antenna is seen in figure 4-42. This antenna is installed inside the radome on the nose section of the aircraft.

c. Glideslope

The signals from glideslope transmitters can be received on a VOR antenna because they operate at a frequency that is approximately the third harmonic of the VOR frequency. Single-engine airplanes commonly use a signal splitter or coupler to supply the glideslope receiver from the VOR antenna. Other general aviation airplanes often use a V-shaped glideslope antenna like that shown in figure 4-43 to receive glideslope signals. This antenna looks a lot like a V-shaped VOR antenna but it is only about 1/3 the size because of the shorter wavelength of glideslope signals. When a separate glideslope antenna is installed on the aircraft, it needs to be located on the front of the aircraft to prevent blockage. The loop type glideslope antenna in figure 4-44 can be

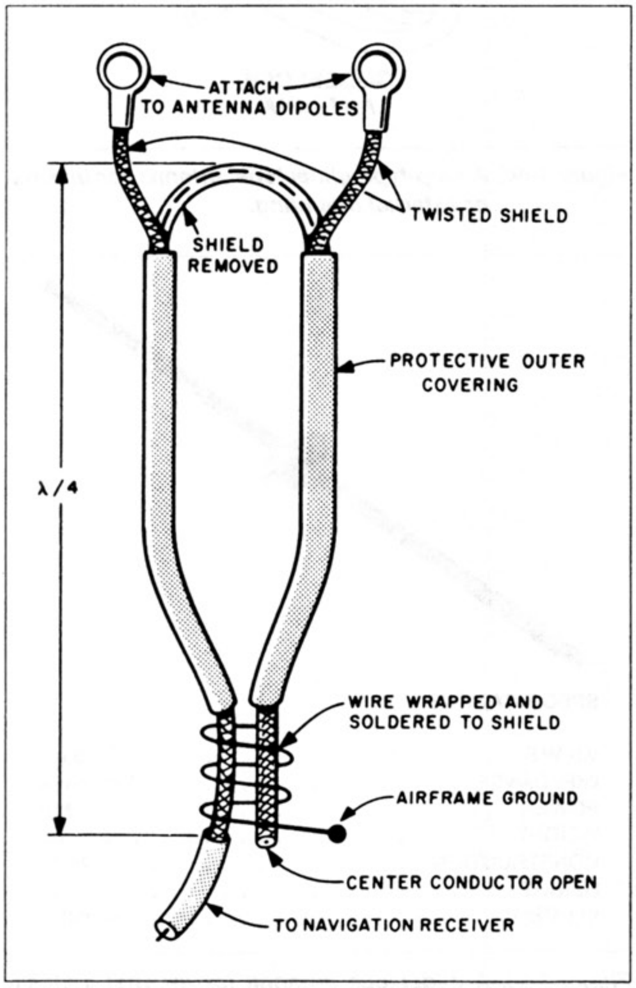
ATTACH
TO ANTENNA DIPOLES
TWISTED SHIELD
SHIELD REMOVED
PROTECTIVE OUTER COVERING
λ/4
WIRE WRAPPED AND SOLDERED TO SHIELD
AIRFRAME GROUND
CENTER CONDUCTOR OPEN
TO NAVIGATION RECEIVER

Figure 4-40. Balun for a VOR antenna.

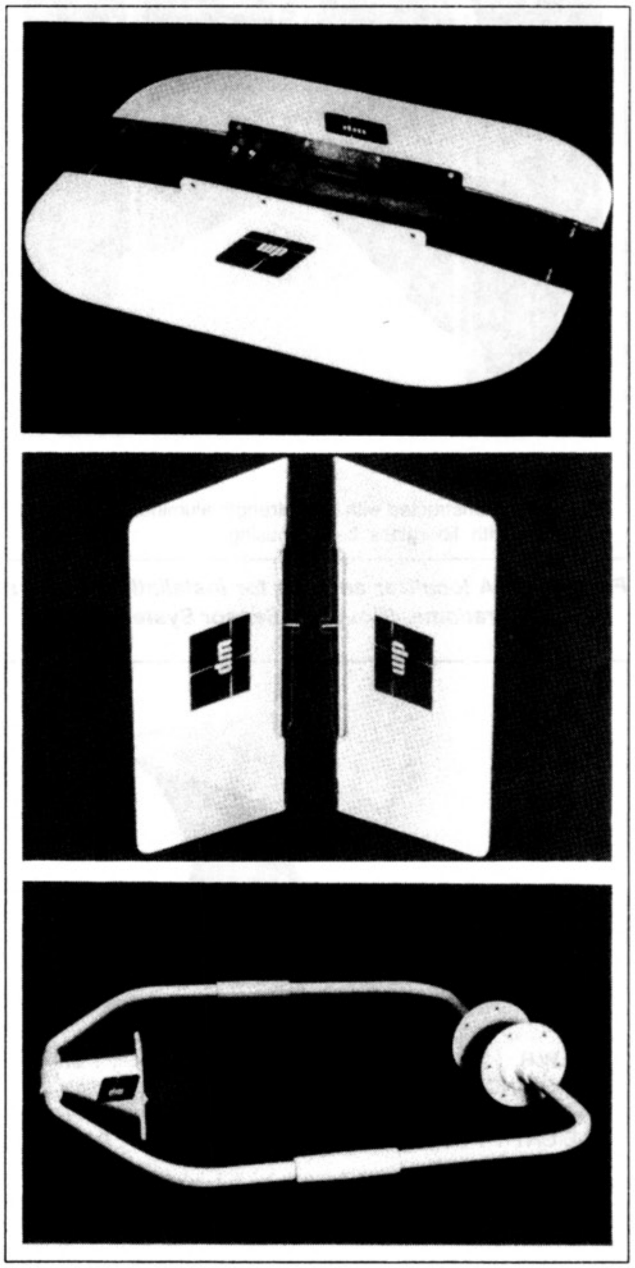

Figure 4-41. Balanced loop antennas for VOR reception. (Courtesy Dorne & Margolin Inc.)

installed either externally or internally on the forward part of an aircraft. The dipole glideslope antenna in figure 4-45 is designed to be installed inside a radome as it is not a streamlined design.

d. Marker Beacon

The older style wire-type marker beacon antenna has been previously described under the heading of wire antennas. All marker beacon antennas need to be installed on the bottom of the aircraft because the signals are received when the aircraft is directly over the transmitter site. Another type of marker beacon antenna found on smaller aircraft is the sled type. This is a bent metal rod which is about 3-½ to 4 ft. long and uses a sliding clip for the lead-in connection. When the antenna is installed on the aircraft, the clip can be loosened and moved to tune the antenna. A newer type of marker beacon antenna is the boat type antenna that is illustrated in figure 4-46. This antenna is smaller and more streamlined than the wire or sled type antennas.

Air carrier jets most often use a flush mounted marker beacon antenna that is installed in the belly of the airplane.

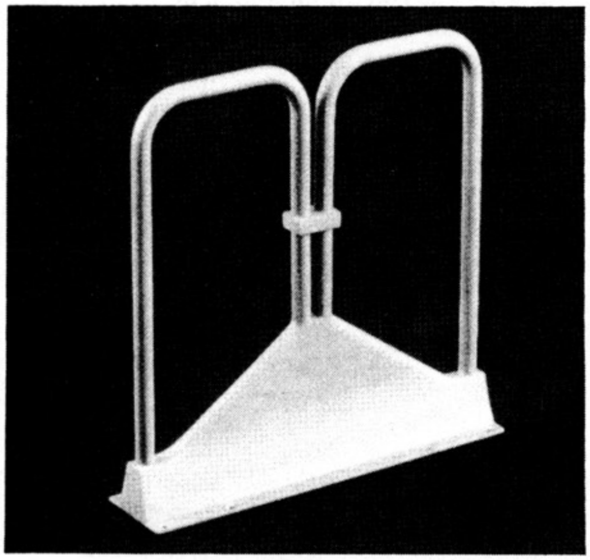

DESCRIPTION

S65-147-2: Constructed with high-strength aluminum tubing and extrusion, with fiberglass base housing.

Figure 4-42.A localizer antenna for installation inside a radome. (Courtesy Sensor Systems)

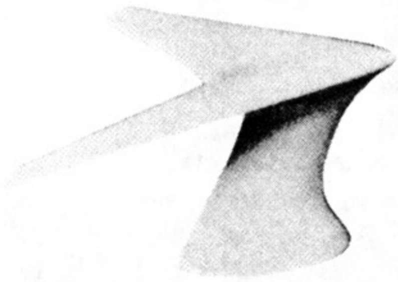

SPECIFICATIONS

V.S.W.R.	5.0:1
IMPEDANCE	50 ohms
POWER	N/A
WEIGHT	0.2 lbs.
CONSTRUCTION	Fiberglass
HEIGHT	3.4 in.
ELEMENT	Grounded

Figure 4-43. A "V"-type glideslope antenna. (Courtesy Dayton-Granger Inc.)

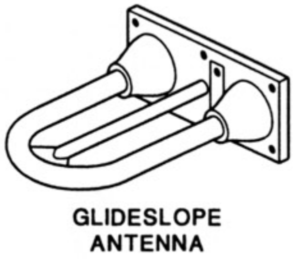

Figure 4-44. A loop-type glideslope antenna for internal or external mounting.

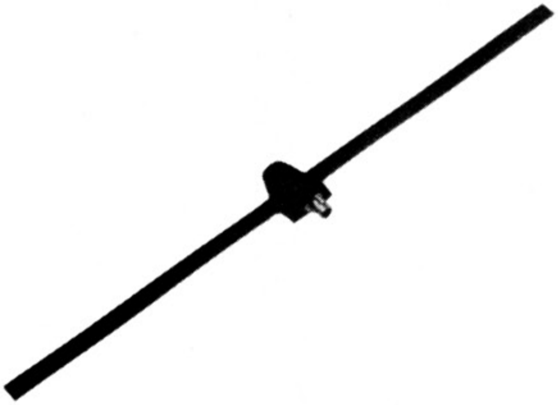

SPECIFICATIONS

V.S.W.R.	3.0:1
IMPEDANCE	50 ohms
POWER	N/A
WEIGHT	0.1 lbs.
CONSTRUCTION	Metal
HEIGHT	15.3 in.
ELEMENT	Grounded

Figure 4-45. A glideslope antenna for internal installation. (Courtesy Dayton-Granger Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

e. HF Communication

The trailing wire and long wire HF antennas found on older aircraft and slow speed aircraft have already been covered. Older air carrier jets used a probe-type HF antenna similar to the vertical fin mounted antenna shown in figure 4-47. This antenna includes a special coupler/tuner that retunes the antenna each time the frequency is changed on the HF radio. This kind of antenna can be mounted on the vertical fin as shown or on a wing tip. The later model air carrier jets use a flush mounted HF comm antenna that is installed inside the vertical fin as seen in figure 4-37. This antenna also requires a special tuning device that is installed at the antenna connection point.

f. VHF Communication

The VHF comm radios on aircraft use a separate antenna for each radio. These antennas are ¼ wave, monopole antennas that can be mounted on the top or bottom of the aircraft. Lower speed aircraft use the thin whip type antennas while higher speed aircraft employ blade type antennas that create less drag. The antenna may either be straight or bent, the bent antennas having the advantages of less drag and less height for belly mountings. A variety of VHF comm antennas is shown in figure 4-48 of both whip and blade types. Some blade-type VHF comm antennas have a stainless steel leading edge to prevent damage, this feature can be seen on the antenna in figure 4-49.

g. DME/Transponder

The same type of antenna can be used for either DME or transponder systems on aircraft. This is practical because they operate at similar frequencies and have similar characteristics. These antennas are almost always installed on the bottom of the aircraft, but they can be located on the top of a narrow tail boom or other location that does not cause serious blockage. The two common types are the spike and blade antennas as illustrated in figure 4-50. The spike is a short metal rod with a ball on the end. This type is cheaper and easier to install, but it is more easily damaged and creates more vibration and drag. The blade type is the most common type on modern aircraft. This antenna can be distinguished from the VHF comm blade because it is much smaller, about 2–4″ long. These antennas are all ¼ wave monopoles with vertical polarization so an adequate groundplane must be provided during installation.

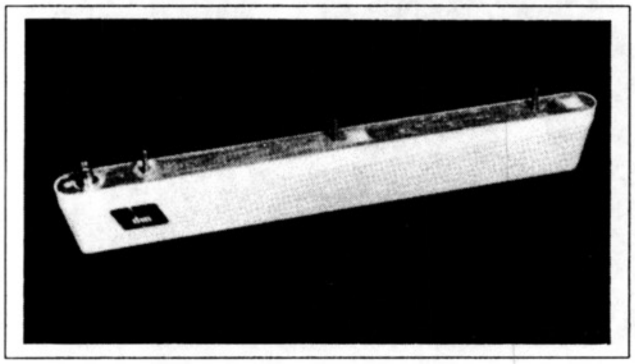

Figure 4-46. A boat-type marker beacon antenna. (Courtesy Dorne & Margolin Inc.)

h. ELT Antennas

Figure 4-51 shows the common type of ELT antenna, it is a thin metal rod that is located close to the ELT itself. The antenna is a Marconi ¼ wave antenna that requires a groundplane. It should normally be installed as close as possible to the ELT because of the low output power of ELT transmitters. A blade type of ELT antenna is also available for higher speed aircraft.

i. Satellite Navigation

The signals from GPS and GLONASS satellites are received from above the aircraft so the antenna needs to be installed on the upper surfaces of the aircraft. A typical GPS antenna is shown in figure 4-52. This small, round antenna creates very low drag and yet has a VSWR of 2:1 which provides good signal reception for the GPS/GLONASS navigation system.

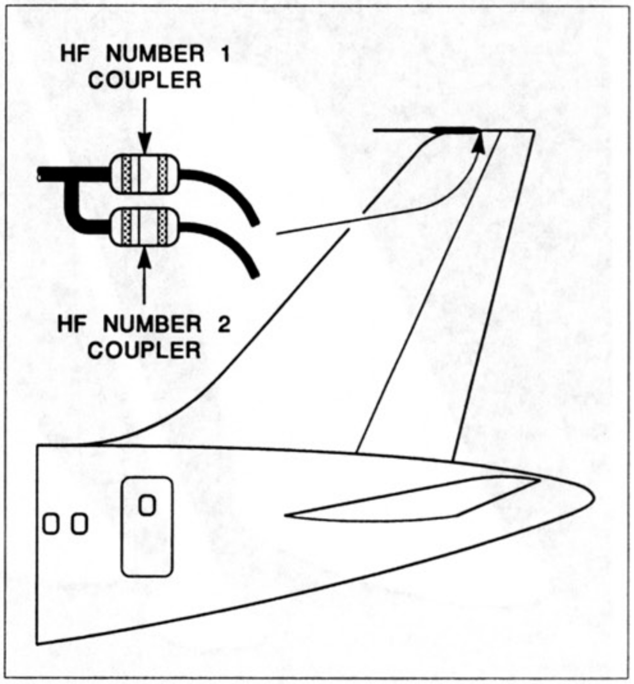

Figure 4-47. Some jet transports have an HF probe-type antenna installed in the vertical fin. Two antenna coupling and tuning devices are also installed in the fin to retune the antenna when different frequencies are selected.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

j. Satellite Communications

The SATCOM antenna, like the SATNAV antenna, must be installed on the top of the aircraft to prevent signal blockage. A variety of different designs are produced for this kind of antenna. The antenna in figure 4-53 is just one of the kinds of antennas being produced for satellite communications systems for aircraft.

k. Loran C

An ADF antenna can be used to receive Loran C navigational signals by utilizing a special antenna coupler. Specific antennas for Loran C are now being produced and they often bear a resemblance to VHF comm antennas as indicated in figure 4-54. These antennas can be installed on either the top or the bottom of the aircraft and still provide good reception because of the frequencies involved. These antennas often include a special anti-static coating to reduce P-static noise in the radio.

l. Omega

Aircraft antennas designed to receive Omega/VLF signals are available in two basic types: the "E" field and "H" field types. The antenna shown in figure 4-55 is the "E" field kind. These antennas can be installed on either upper or lower surfaces of the aircraft. The most important consideration when choosing a location is to reduce noise interference from aircraft systems. A "skin noise map" is often required which consists of measuring the VLF noise on various parts of the aircraft to find the best antenna location. The lowest noise is usually found on the aft underbelly of most aircraft.

m. MLS

The MLS receive antenna seen in figure 4-56 is a low profile, vertically polarized antenna designed to receive the MLS signals that operate on frequencies of 5.03 to 5.09 GHz. This kind of antenna should be located on the nose section of the aircraft for

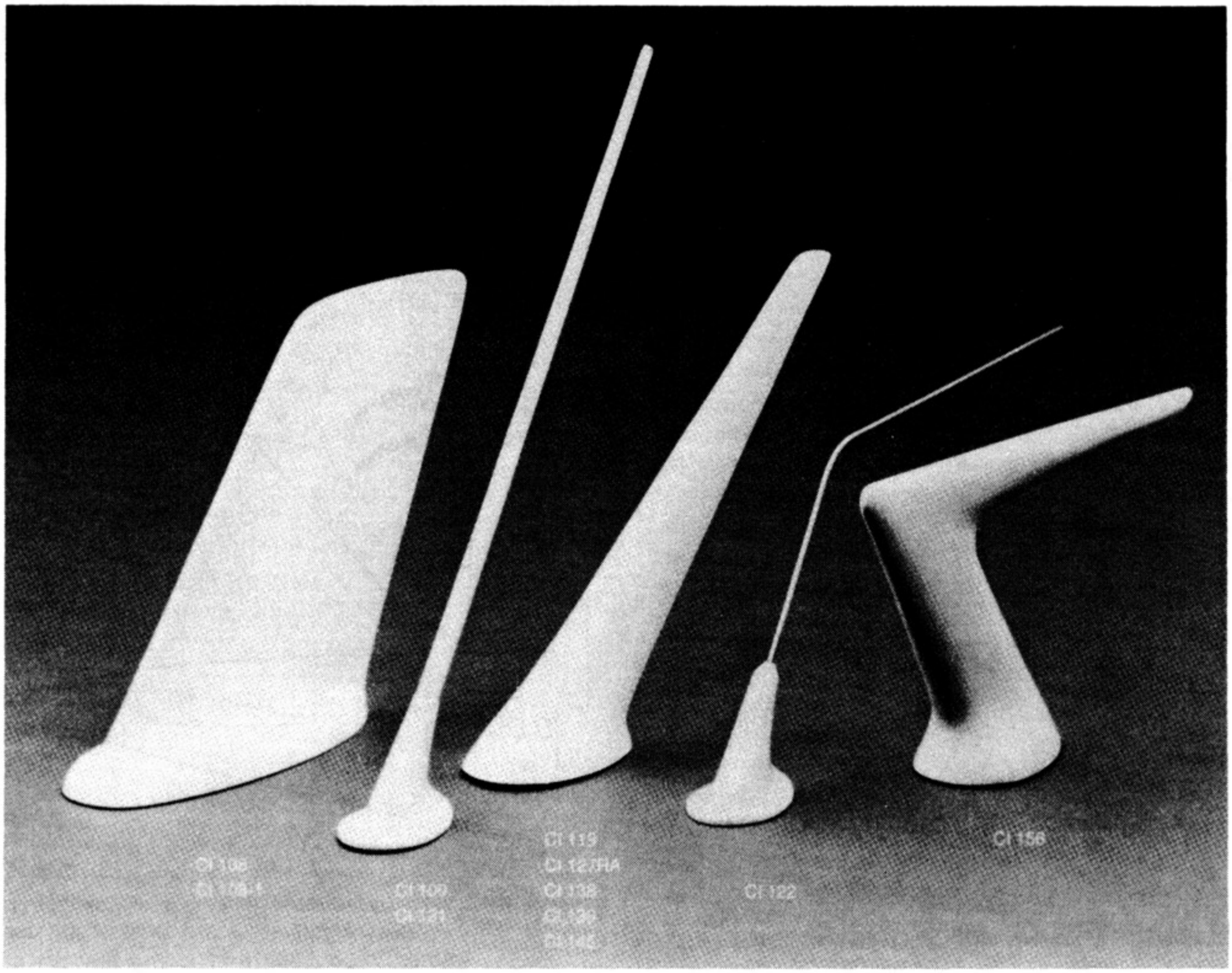

Figure 4-48. Various VHF comm antennas. (Courtesy Comant Industries Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

best reception and minimum blockage. Some MLS systems require two antennas to be installed on the aircraft for proper signal reception.

n. TCAS

The Traffic Alert and Collision Avoidance system found on air carrier jets requires a special type of directional antenna like that seen in figure 4-57. This TCAS I antenna is normally located on the top of the fuselage and has three connector ports for connection to the aircraft's TCAS I equipment.

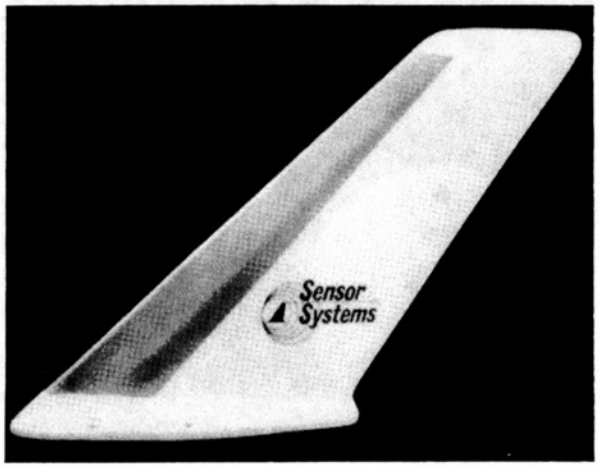

DESCRIPTION

S65-8282: This broadband fixed tuned antenna operates in the frequency range of 116-156 MHz.

Figure 4-49. A blade-type VHF comm antenna with a stainless steel guard on the leading edge. (Courtesy Sensor Systems)

o. Radiotelephone

Radiotelephone antennas come in a wide variety of shapes and sizes. These UHF antennas are normally installed on the bottom of the aircraft since they operate in conjunction with ground based line-of-sight radio waves. A number of different kinds of radiotelephone antennas are shown in figure 4-58. A major consideration when installing this type of antenna is preventing noise that can be caused by loose joints and poorly bonded surfaces on the aircraft.

C. Autopilots and Flight Directors

The FAA classifies autopilots as aircraft instruments so A&P technicians cannot repair or alter autopilots. There are many tasks related to autopilots that might be performed by aircraft technicians such as installation, inspection, troubleshooting etc. An autopilot is an expensive and complicated device. It often has various components located in many different areas of the aircraft and many interconnections. The autopilot is connected to the flight control system of the aircraft and autopilot malfunctions can be very serious indeed.

An autopilot system must always be approved by the FAA for the specific make and model of aircraft in which it will be installed. A type of autopilot may be approved for a number of different

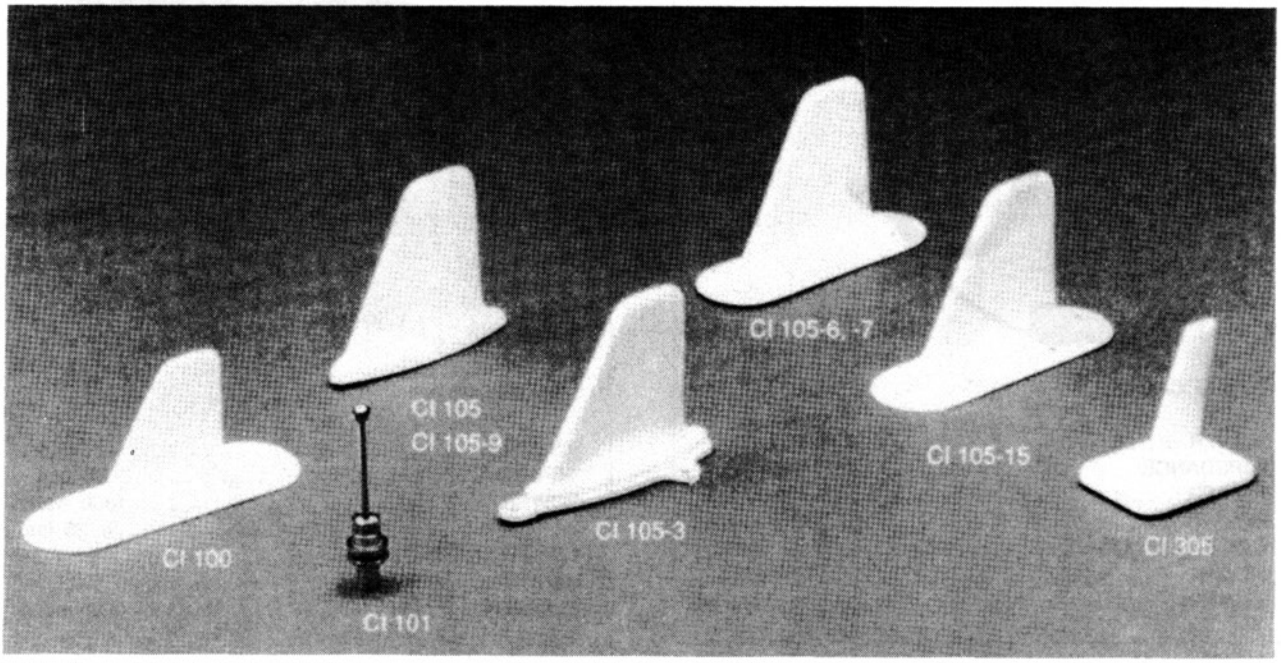

Figure 4-50. Typical antennas used for DME and transponder. (Courtesy Comant Industries Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

aircraft, but different torque settings and adjustments may have to be made for each application. The maintenance instructions that apply to the specific autopilot installation should always be followed as there are many differences in adjustments and testing for the various aircraft installations. The basic principles of operation for aircraft autopilots will be described here along with some specific examples of aircraft autopilot installations.

1. Types of Autopilots

Autopilot systems are categorized according to the number of aircraft axes of rotation they control and according to their complexity. The autopilot utilizes the same control surfaces that the human pilot does. The three control axes of an airplane are shown in figure 4-59. The rudder controls aircraft rotation about or around the vertical or yaw axis. The elevators control rotation about the lateral or pitch axis. The ailerons control aircraft rotation about the longitudinal or roll axis. Autopilots can be described as single-axis, two-axis or three-axis types. The single-axis autopilot usually operates the ailerons only and is often referred to as a wing leveler. The two-axis autopilot controls the ailerons and elevator to provide additional control of the aircraft. A three-axis autopilot operates all three types of control surfaces: ailerons, elevator and rudder. There is a very large difference in the capabilities of a three-axis autopilot found on a small general aviation airplane and the three-axis autopilot found on air carrier jets and similar

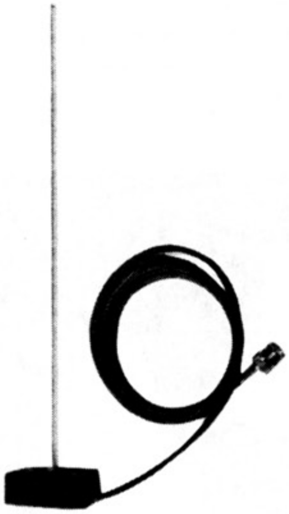

SPECIFICATIONS

V.S.W.R.	2.0:1
IMPEDANCE	50 ohms
POWER	40 watts
WEIGHT	0.3 lbs.
CONSTRUCTION	Whip
HEIGHT	18.3 in.
ELEMENT	Open

Figure 4-51. A whip-type ELT antenna. (Courtesy Dayton-Granger Inc.)

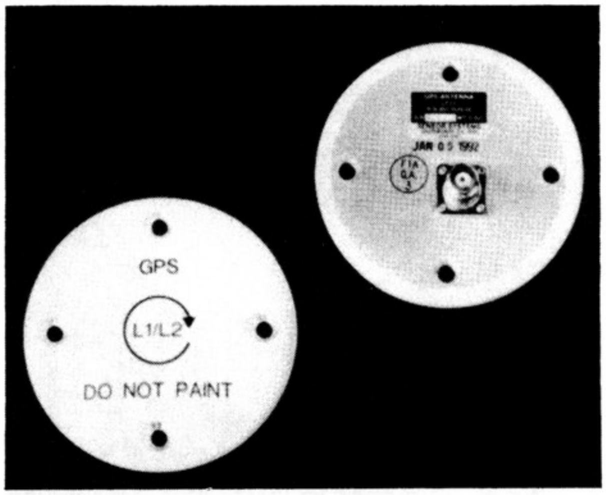

DESCRIPTION

S67-1575-14: Dual band L1/L2 GPS Antenna provides coverage at 1227.6 MHz and 1575.42 MHz with a VSWR of 2.0:1.

Figure 4-52. GPS antenna for satellite nav. (Courtesy Sensor Systems)

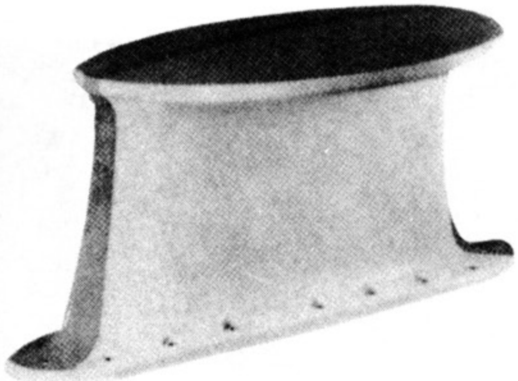

SPECIFICATIONS

V.S.W.R.	2.5:1
IMPEDANCE	50 ohms
POWER	1300 watts
WEIGHT	26 lbs.
CONSTRUCTION	Fiberglass
HEIGHT	10.5 in.
ELEMENT	Grounded

Figure 4-53. An antenna for satellite communications. (Courtesy Dayton-Granger Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

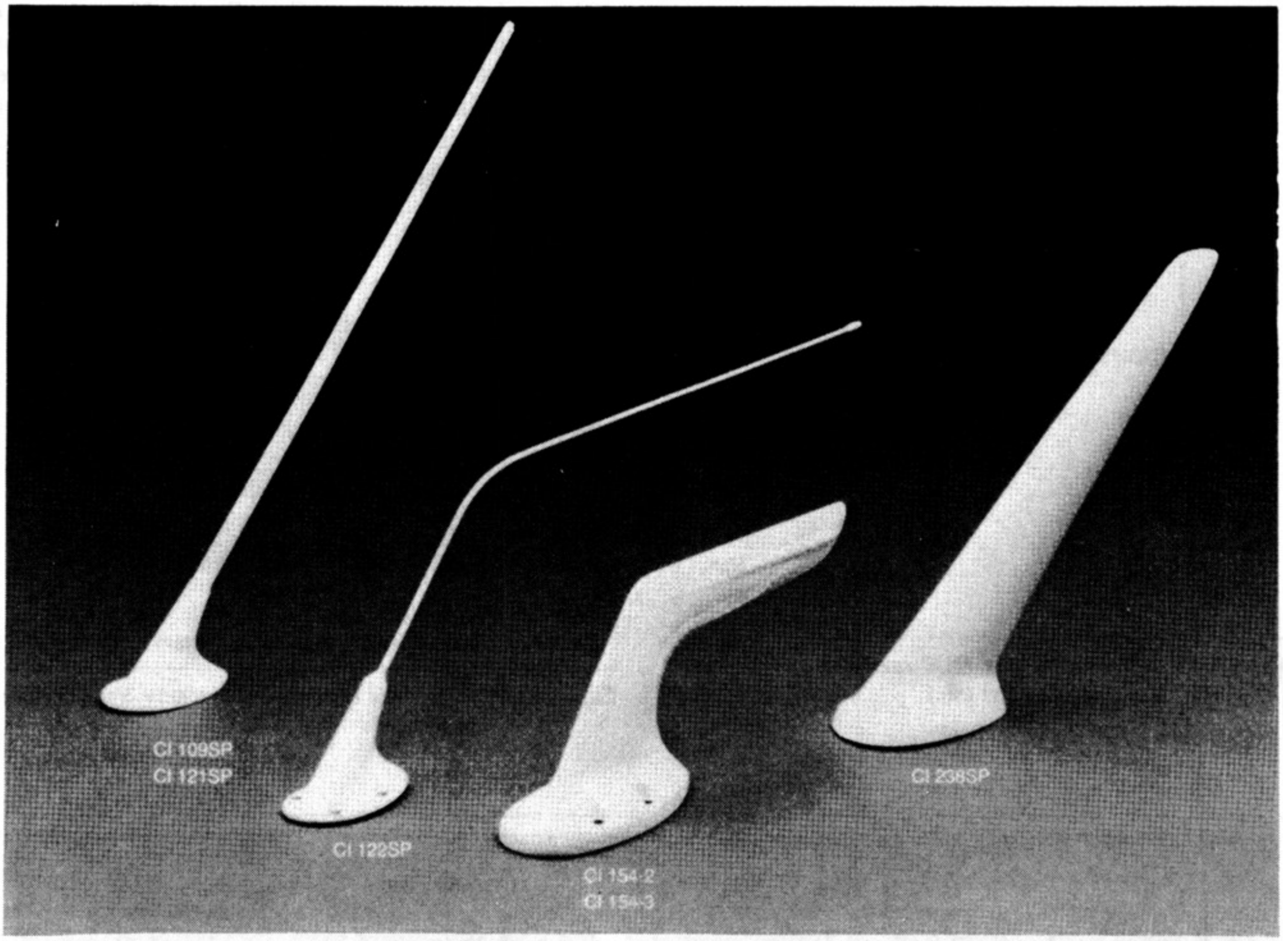

Figure 4-54. Antennas for Loran C nav receivers. (Courtesy Comant Industries Inc.)

aircraft. For this reason two other categories of autopilot will be added to the three already mentioned. Two common abbreviations for these advanced autopilots will be used to distinguish them from the other types.

The term "Automatic Flight Control System" (AFCS) generally represents the state-of-the-art that was reached a few years ago. The autopilot in the Lockheed L-1011 is an example of an AFCS.

This is a three axis autopilot that can control the aircraft during climbs, descents, cruise flight and during instrument approaches. It also has an auto-throttle system which will automatically control engine power or thrust. Some AFCS autopilots have auto-land capability where the autopilot can actually land the airplane on the runway. These types of autopilots require many back-up systems and high levels of redundancy. The AFCS includes a flight director function which will be explained later.

The latest types of autopilots are referred to as Flight Management Systems (FMS). These include

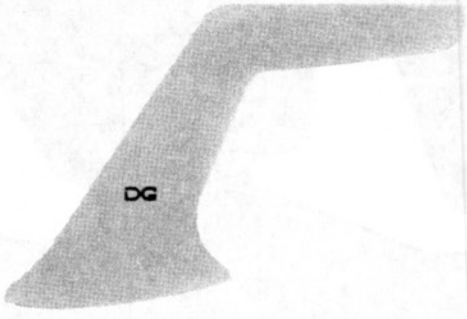

SPECIFICATIONS

V.S.W.R.	N/A
IMPEDANCE	N/A
POWER	N/A
WEIGHT	1.4 lbs.
CONSTRUCTION	Fiberglass
HEIGHT	8.2 in.
ELEMENT	Open

Figure 4-55. An "E" field Omega antenna. (Courtesy Dayton-Granger Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

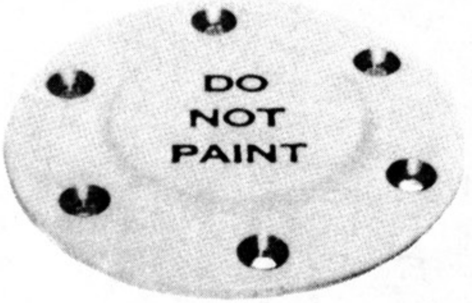
DO
NOT
PAINT

SPECIFICATIONS

V.S.W.R.	2.0:1
IMPEDANCE	50 ohms
POWER	N/A
WEIGHT	0.1 lbs.
CONSTRUCTION	Fiberglass
HEIGHT	0.3 in.
ELEMENT	N/A
APPROVALS	TSO-C104

Figure 4-56. An antenna for MLS reception. (Courtesy Dayton-Granger Inc.)

additional computers called Flight Management Computers that permit an entire flight from just after takeoff to landing to be programmed in the computers and automatically controlled. The Flight

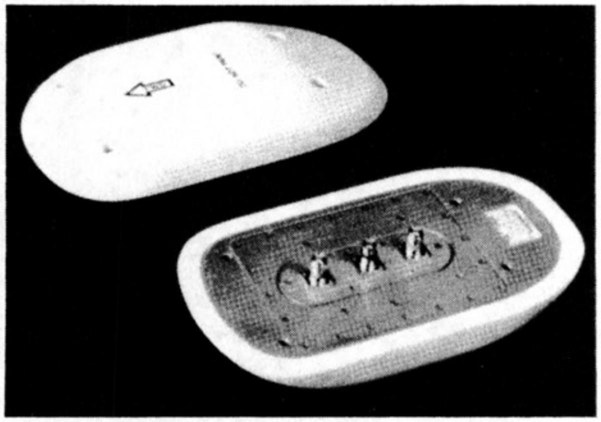

DESCRIPTION

S72-1744: Traffic Collision Avoidance System I.

Figure 4-57. A TCAS I antenna. (Courtesy Sensor Systems)

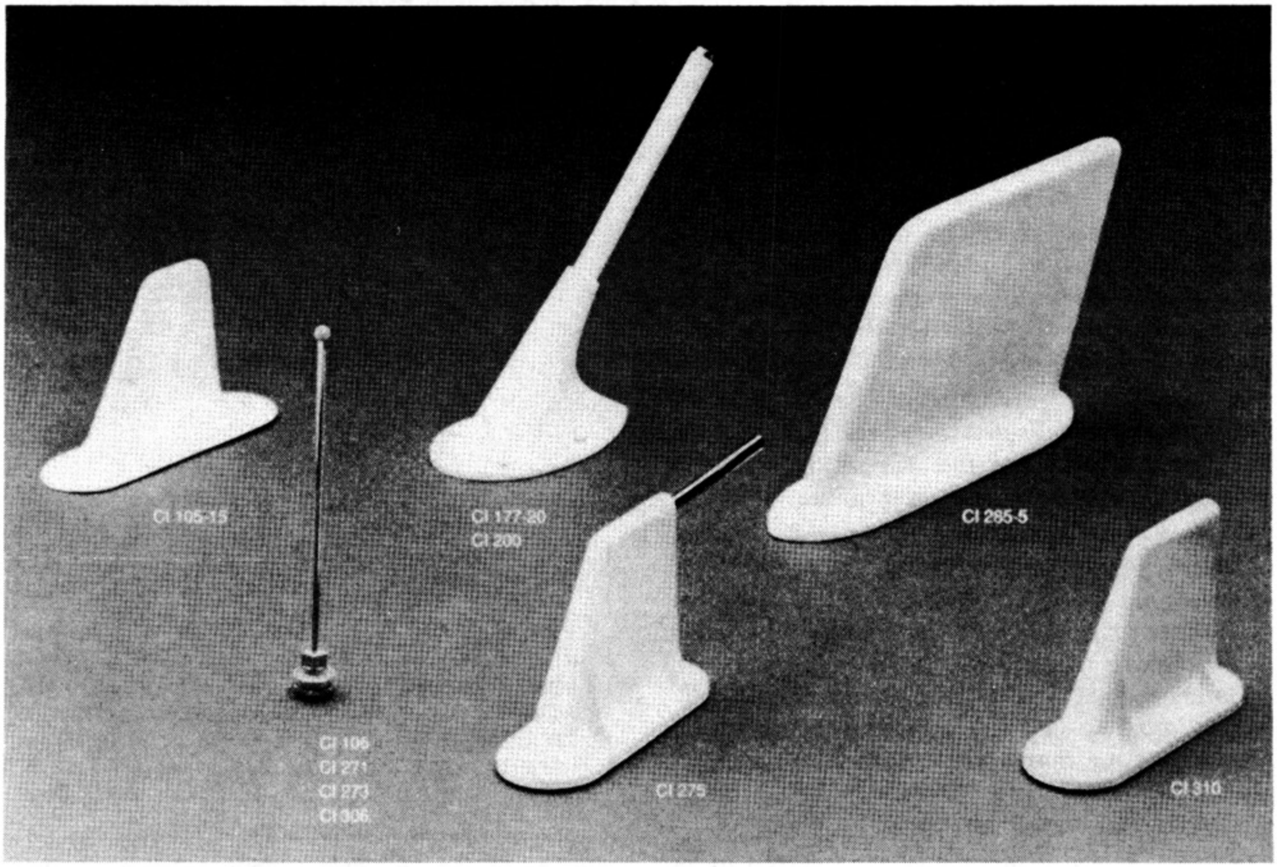
CI 105-15
CI 177-20
CI 200
CI 285-5
CI 106
CI 271
CI 273
CI 306
CI 275
CI 310

Figure 4-58. A variety of radiotelephone antennas. (Courtesy Comant Industries Inc.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Management Computer can be thought of as a master computer which controls the autopilot and auto-throttle computers. The computers can store in their memory many different routes and flight profiles and they can be used to provide a maximum economy in fuel consumption or other desired controlling factor. The standard Boeing 767 autopilot systems will be used later as an example of the capabilities of an FMS installation.

2. Basic Autopilot Operation

The FAA states in AC 65-15A that the purpose of an automatic pilot system is primarily to reduce the work, strain and fatigue of controlling the aircraft during long flights. The capabilities of a modern autopilot go way beyond simply controlling the aircraft during cruise operations. A sophisticated autopilot system can land the airplane in weather conditions that are so bad that the human pilot could not legally land the airplane. We would have to say that the statement is true for simple autopilot systems, but is obsolete or outdated in describing a sophisticated modern autopilot. In this section, the basic parts and operation of simple autopilots will be described. Figure 4-60 shows the basic parts of the rudder control channel of an autopilot. The aileron and elevator channels would work in a similar fashion. The basic parts and their functions are:

1. Sensors or Gyros — These detect a change in aircraft attitude using gyros or similar sensing devices.
2. Amplifier or Computer — This component processes the signals from the sensors and sends signals to the servos to correct the attitude.
3. Servos — The servos receive the signals from the computer and supply the physical force necessary to move the flight control surface.

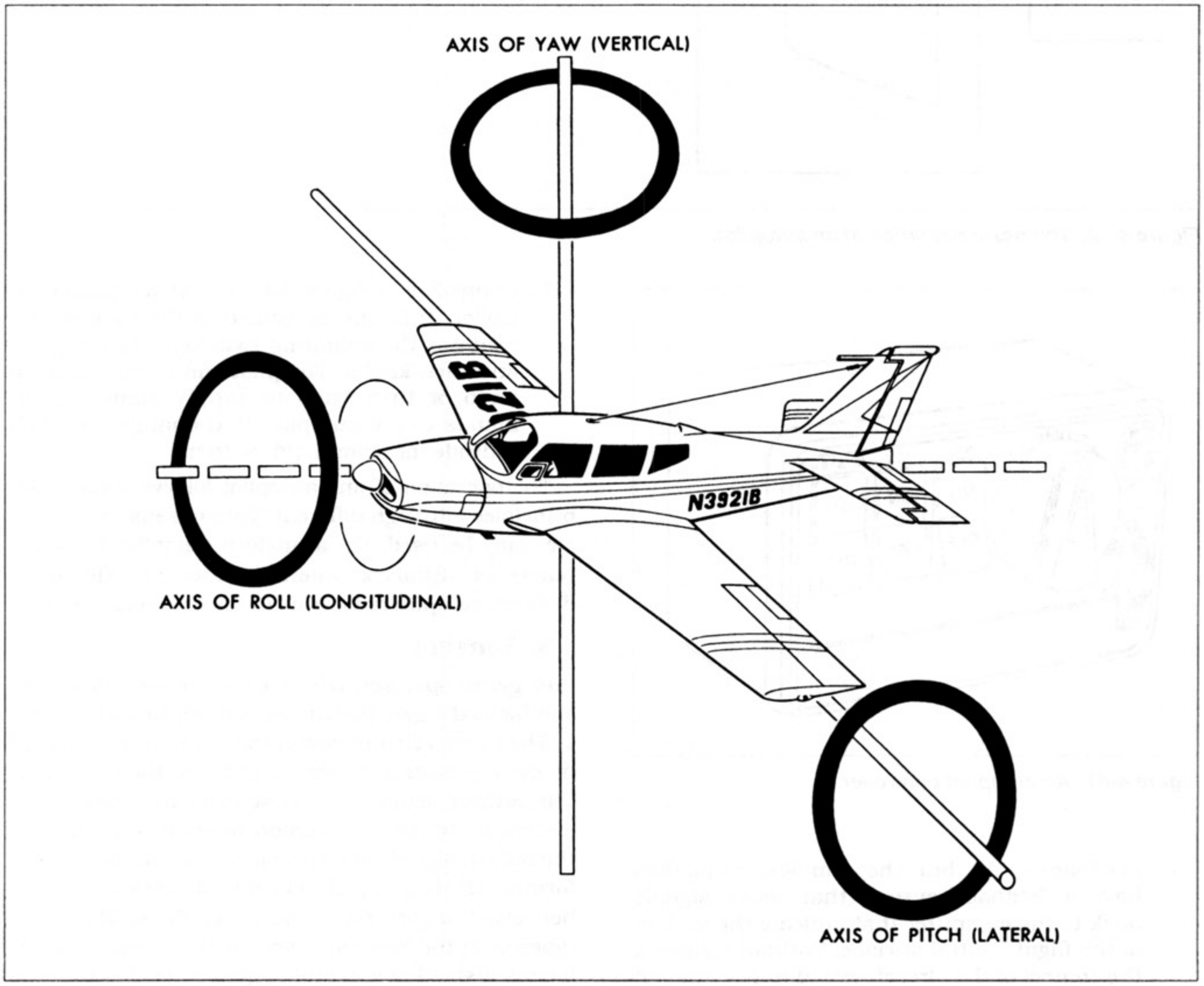

Figure 4-59. The control axes for an airplane.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

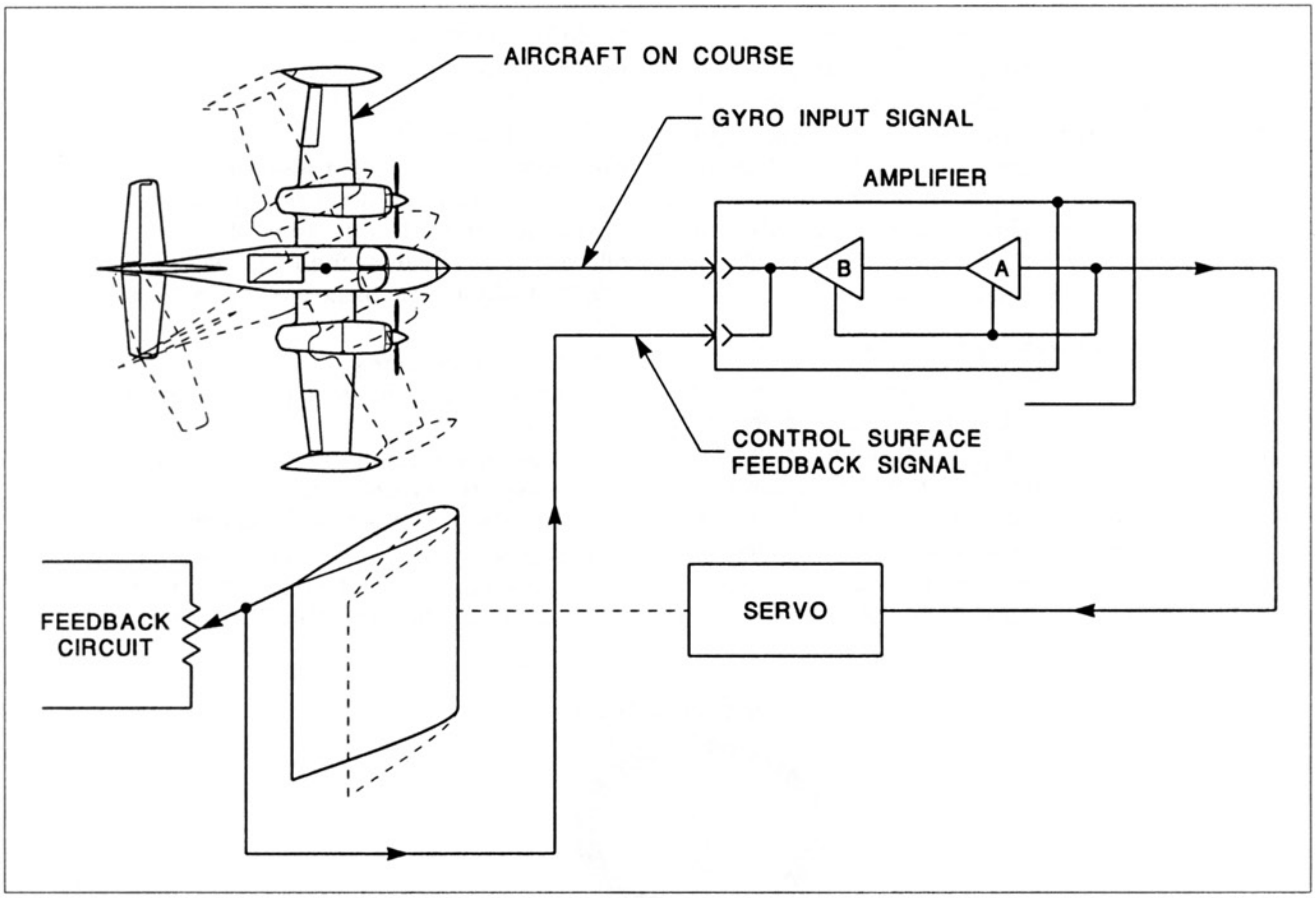

Figure 4-60. The basic operation of an autopilot.

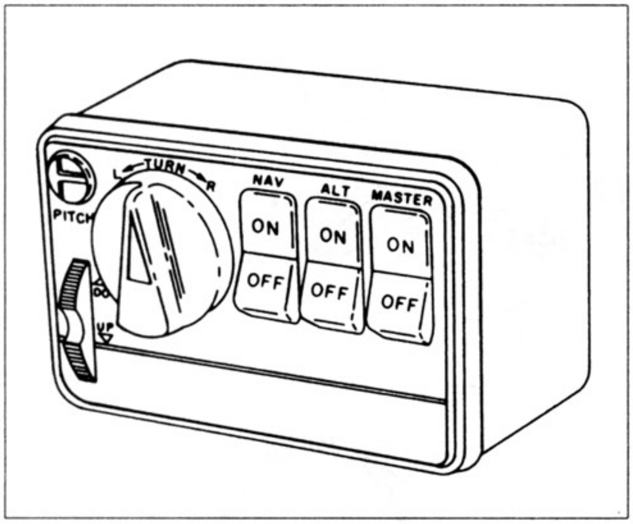

Figure 4-61. An autopilot controller.

4. Feedback — All but the simplest autopilots have a feedback system that sends signals back to the computer that indicate the motion of the flight control surface. Without feedback the control of the aircraft would not be smooth and precise.
5. Controller — Figure 4-61 shows a typical controller. This unit is located in the cockpit and contains the actuating switches and the pitch and turn knobs. The pilot can move the pitch knob or turn knob to supply manual commands to the autopilot that change the pitch attitude or command a turn.

The operation of any autopilot follows these basic principles although different types of sensors, servos, etc. may be used. On a modern autopilot the computers are digital computers and there are often three different computers for each of the three control axes.

3. Sensors

The gyroscopic sensors used with autopilots are similar to the gyro instruments described in chapter 1. The pitch, roll and yaw of the aircraft are detected by gyro sensors that send signals to the computer. The output signals of the sensors are most often electrical signals. A common method of producing the output signals is a special type of variable transformer called an EI pick-off which detects the motion between the gyro rotor and its gimbals. The latest types of autopilots use a sensor that employs laser beams instead of a spinning gyro rotor. Figure 4-62 shows one of these laser sensors that are called

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

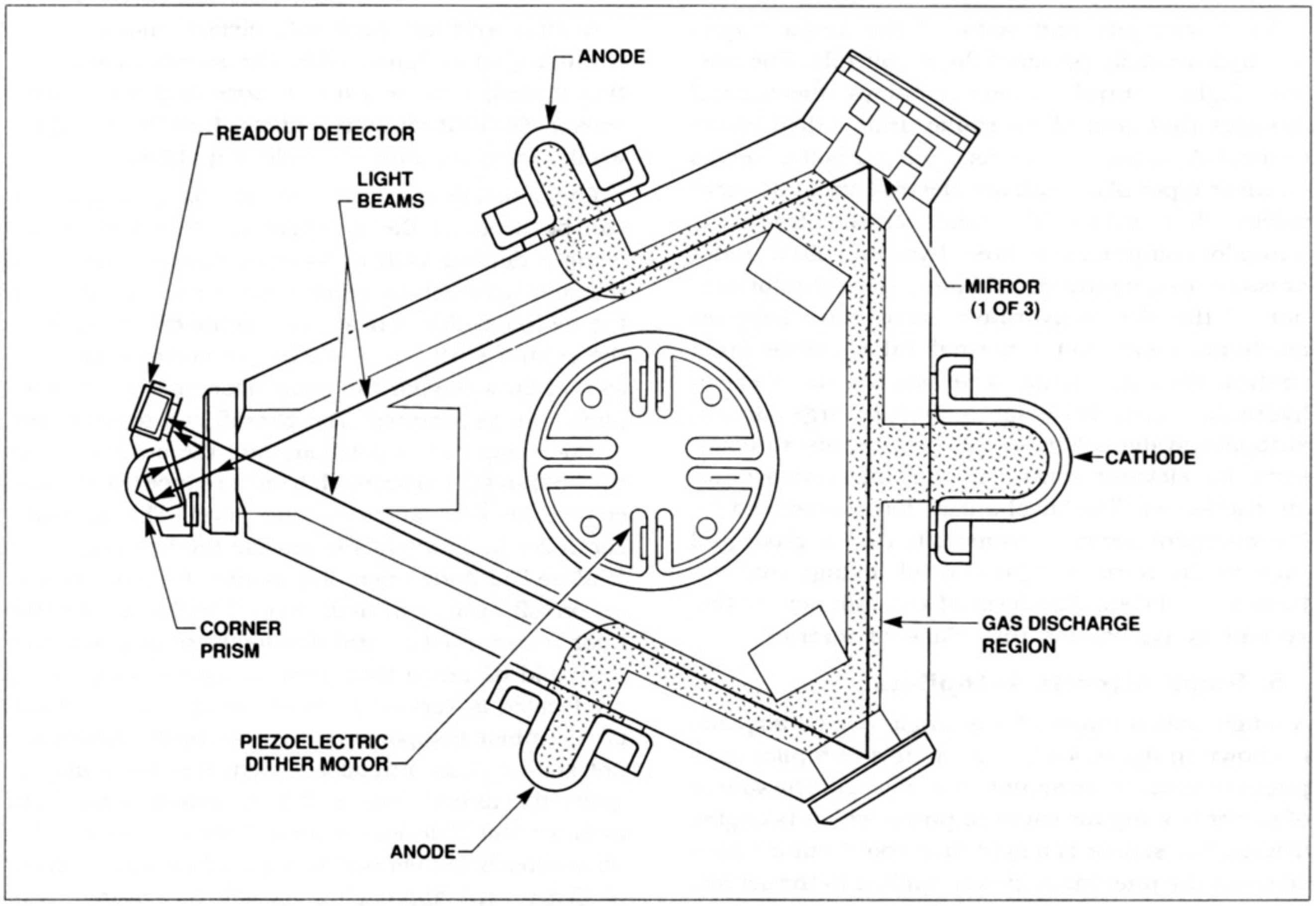

Figure 4-62. Ring laser gyro sensor for an autopilot.

ring laser gyros or RLGs. The RLG has two laser beams that travel in opposite directions around a triangular course. Sensitive detectors measure the Doppler shift or frequency change whenever the unit is rotated. One of these is needed for each axis that must be measured for the autopilot. These RLGs are much more expensive than an actual gyro, but they eliminate the moving parts that cause a conventional gyro to gradually wear out.

4. Servos

The servos supply the force needed to move the flight control surfaces of the aircraft. There are four basic kinds which will be described here. Some simple autopilots found on small airplanes use vacuum sources like those used to operate gyro instruments. The vacuum is directed to pneumatic servos that are connected mechanically to the normal flight control system. As seen in figure 4-63, the pneumatic servo is an air tight housing which contains a movable diaphragm. When vacuum is applied to the servo, the diaphragm is displaced which pulls on the bridle cable that is connected to the main control cable by a bridle clamp. Two of these servos would be needed for each control axis.

Servos that utilize electric motors are shown in Figures 4-64 and 4-65.

The servo shown in figure 4-64 uses a reversible DC motor and reduction gearing to supply the force to move the control surface in both directions. The servo in figure 4-65 has an electric motor that runs continuously and uses magnetic clutches to engage the mechanism and apply torque to the capstan and control cable. This type has the advantage that the inertia forces in starting and stopping the motor are eliminated. It can be engaged and disengaged more rapidly and precisely.

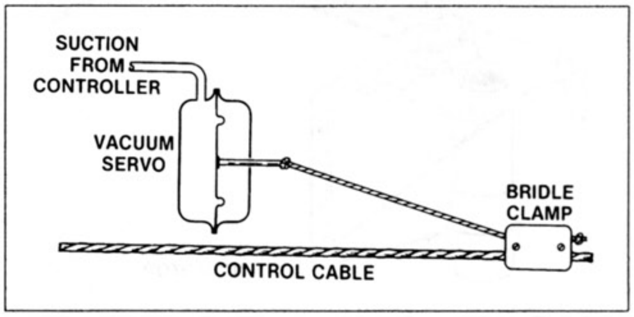

Figure 4-63. Pneumatic servo for a small aircraft autopilot.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Air carrier jets and some of the larger bizjets use hydraulically powered flight controls. The normal flight control system employs mechanical linkages that control hydraulic units called Power Control Actuators or PCAs. The autopilot servos on these types of aircraft are electro-hydraulic servo valves that utilize electrical signals from the autopilot computers to direct hydraulic fluid under pressure to a hydraulic actuator. The actuator portion of the electro-hydraulic servo valve supplies mechanical force to the normal linkage of the flight control system. Figure 4-66 shows the electro-hydraulic servo valve for a typical large aircraft autopilot system. Figure 4-67 shows an autopilot servo for elevator control in the tail section of an air carrier jet. The mechanical force produced by the autopilot servo is transmitted by a push-pull tube to the normal flight control linkage that activates the PCAs. The level of redundancy in this system is typical for this class of aircraft.

5. Small Aircraft Autopilots

A single-axis autopilot for a single engine airplane is shown in figure 4-68. This simple autopilot uses pneumatic servos to actuate the ailerons. The source of power is a dry air vacuum pump which is engine driven. The sensor is a gyro turn coordinator which controls the pneumatic power applied to the servos. Some of the torque settings and rigging instructions for the autopilot can be seen in this drawing. This is the type of autopilot which is often called a wing leveler since it controls only the aileron control surfaces.

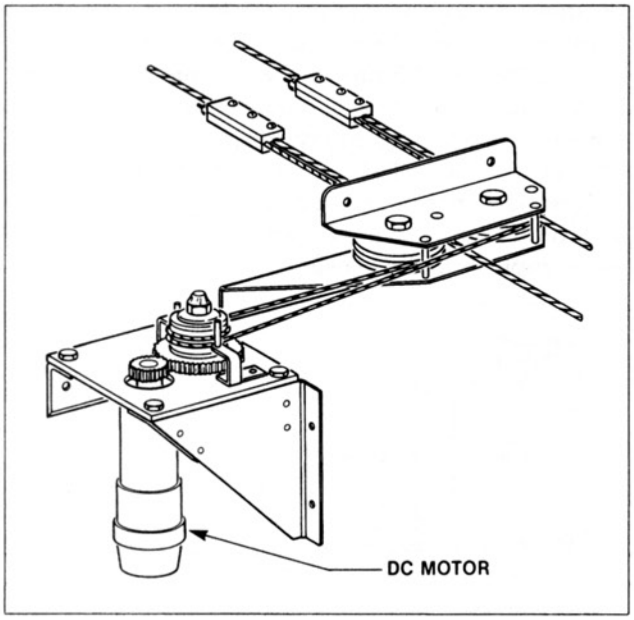

Figure 4-64. Autopilot servo with reversible DC motor, reduction gears and bridle cables.

A three axis autopilot with electric motor servos is illustrated in figure 4-69. The sensors used with this system include gyro sensors and an altitude sensor. The altitude sensor shows that this autopilot would have an altitude hold capability.

Radio signals from the aircraft's navigation radios can be used by the autopilot to steer the aircraft along a desired VOR or localizer course. The pitch, roll and yaw servos receive electrical signals from the computer that activate the electric motors to move the control surfaces. A pitch trim servo is included so that the autopilot can apply nose up or nose down pitch trim as required. The aircraft can operate with a wide range of CG positions and the autopilot, like the human pilot, uses pitch trim to reduce the elevator control force to an acceptable level. The autopilot controller has switches to engage the heading, radio NAV and altitude operating modes. It also contains an on/off switch, a pitch control indicator and the knobs for manual control of autopilot pitch and turns. It should be noted that these autopilot components are located in various parts of the aircraft and some of the minor components such as bridle cables are not shown. This autopilot system has the ability to guide the aircraft on an ILS approach using both localizer and glideslope signals. This feature is called an approach coupler and is required for certain types of instrument approaches as will be covered later in the section on FARs.

6. Flight Management System (FMS)

The Boeing 767 will be used as an example of a flight management system or FMS. This system has the capability of automatically controlling the airplane from just after takeoff (above 400 ft. AGL)

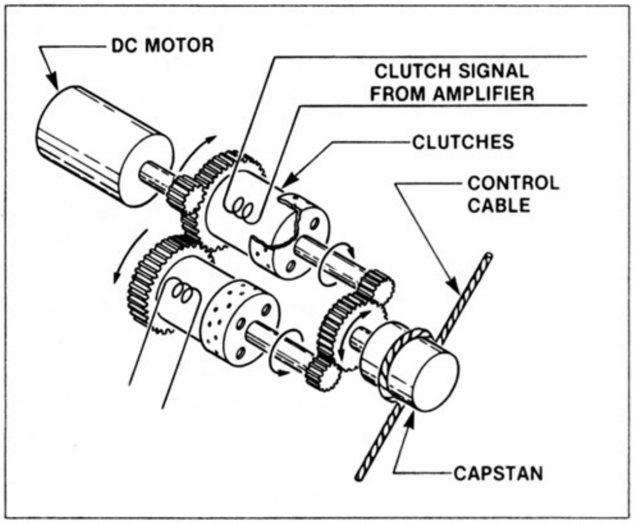

Figure 4-65. Autopilot servo with a motor that runs continuously and is engaged by magnetic clutches.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

through roll out on the runway after landing at the destination airport. The human pilot must take over to turn off the runway and taxi to the gate. This does not mean that all flights will use all these capabilities, but the autopilot and flight director will be used for some portion of each flight under normal circumstances.

a. Flight Management Computers

The Flight Management Computers (FMCs) provide a number of advanced features and functions which were not found on earlier autopilot systems. Some of the functions of the Flight Management Computer are:

1. Flight Planning — The entire flight can be programmed into the computer using a cockpit keyboard.
2. Performance Management — The system can provide optimum profiles for climb, cruise, descent and holding patterns. A minimum cost flight can be flown automatically by using optimum climb settings, cruise settings etc.
3. Navigation Calculations — The FMC can calculate great circle routes, climb and descent profiles etc.
4. Auto Tune of VOR and DME — The FMC can automatically tune the radios to the correct station frequencies.
5. Autothrottle Speed Commands — These are displayed on the EADI as FAST/SLOW indications.

The FMC is in effect a master computer which integrates the functions of the laser sensors, Flight Control Computers, Thrust Management Computers, Air Data Computers, navigation sensors and EICAS computers. The autopilot sensors are called Inertial Reference Units (IRUs) and they are the same in basic operation as the Ring Laser Gyros (RLGs) previously described.

b. Flight Control Computers

The flight control computers are the autopilot computers and there are three of them. A block diagram of the connections to the three Flight Control Computers is shown in figure 4-70. The three computers

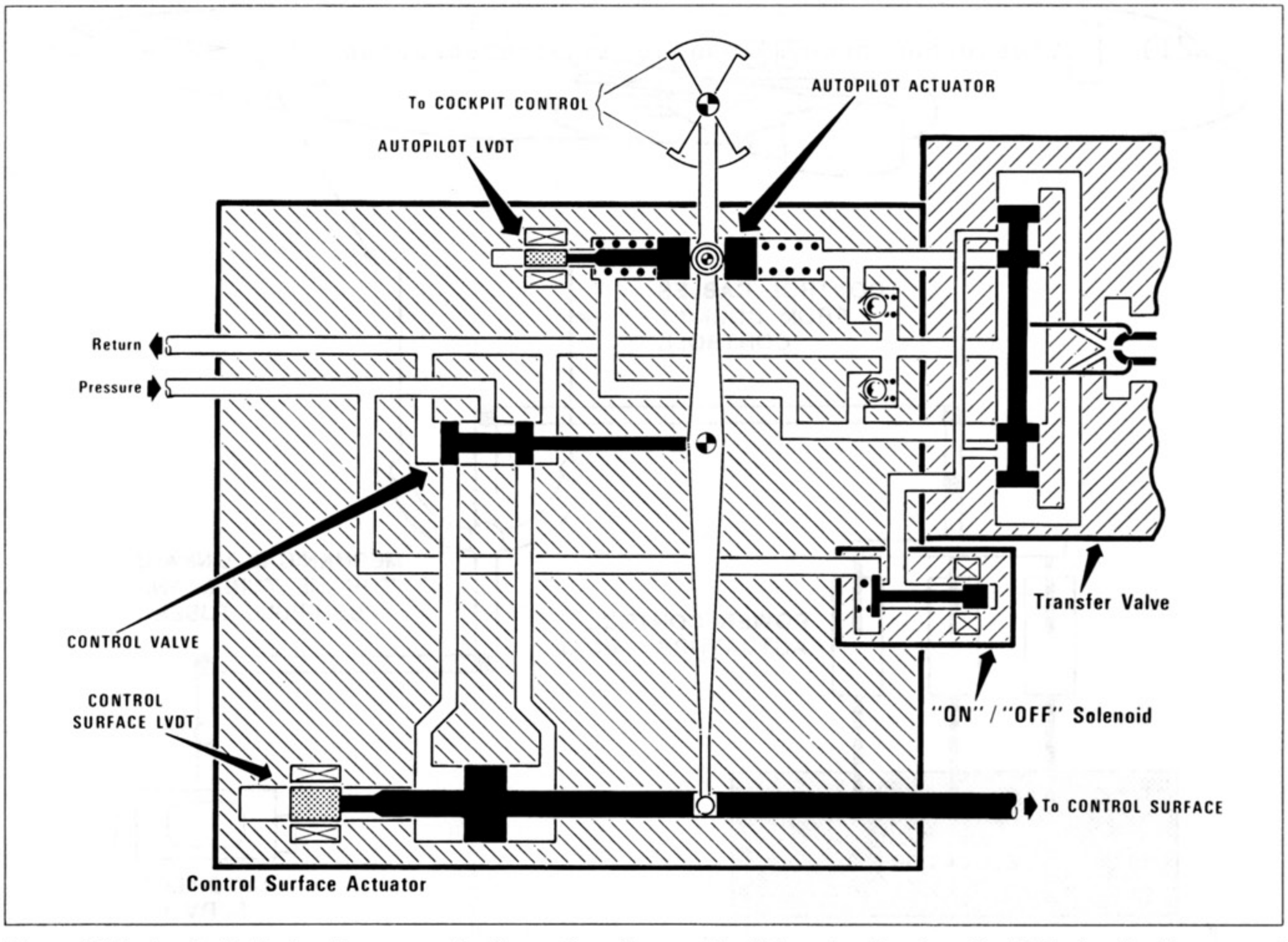

Figure 4-66. An electrohydraulic servo valve is employed in an autopilot system for aircraft with hydraulically powered flight controls. Linear variable differential transformers (LVDTs) provide feedback signals of the movement of the mechanical parts of the system.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

are independent so that a failure in one will not affect the other two. The computers are modern digital computers that are more compact and faster than earlier types of computers.

c. Thrust Management Computer (TMC)

The purpose of the TMC is to automatically set the proper thrust level for the engines. A diagram of the autothrottle system is shown in figure 4-71. The output servo moves the throttle linkage to set the level of engine power calculated by the TMC. The system includes sensors on the engines which monitor the important engine operating parameters. The monitoring of engine parameters is used to prevent exceeding any engine operating limitation for RPM, EPR, EGT, etc. The autothrottle system can be used to maintain a given climb rate, indicated airspeed, Mach number or descent rate. Since the 767 has autoland capabilities, the autothrottle system will automatically close the throttles just prior to landing so that a smooth touchdown can be made. The TMC system also provides a minimum speed protection which will maintain a safe margin above stall speed for the particular flight configuration. The autopilot system and the autothrottle system can be engaged separately or together using the controls on the flight control panel.

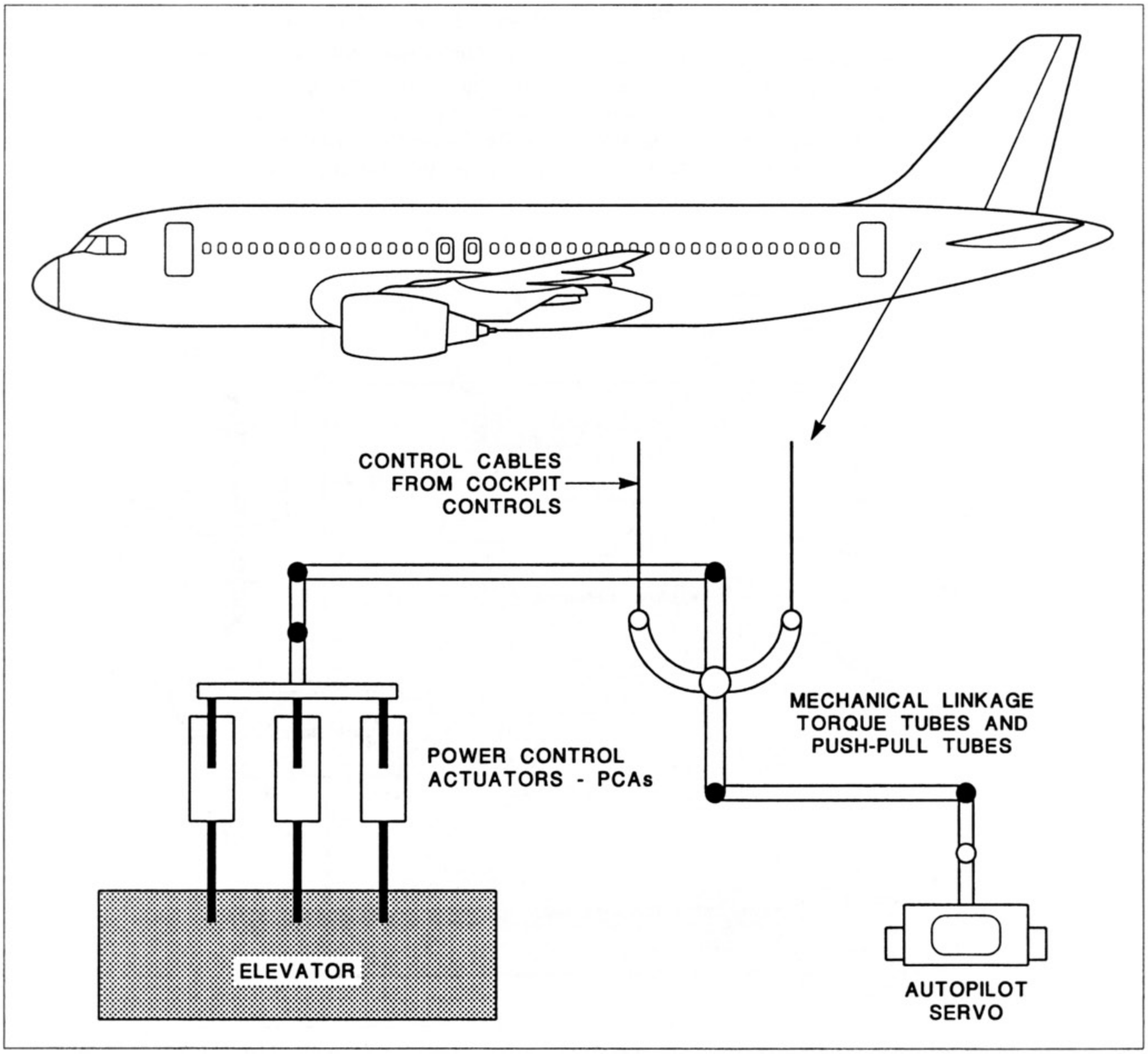

Figure 4-67. The autopilot servos on a large jet airplane provide mechanical force to move the normal control linkage and activate the hydraulic PCAs that move the flight control surfaces.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

NOTE

TORQUE HOSE MOUNTING NUTS (3) TO 12-14 LB. INCHES AND CABLE CLAMP (6) TO 70-90 LB. INCHES WHEN INSTALLING

THE OPTIONAL GYRO SYSTEM AND THE WING LEVELER SYSTEM OBTAIN VACUUM FROM THE SAME VACUUM SOURCE

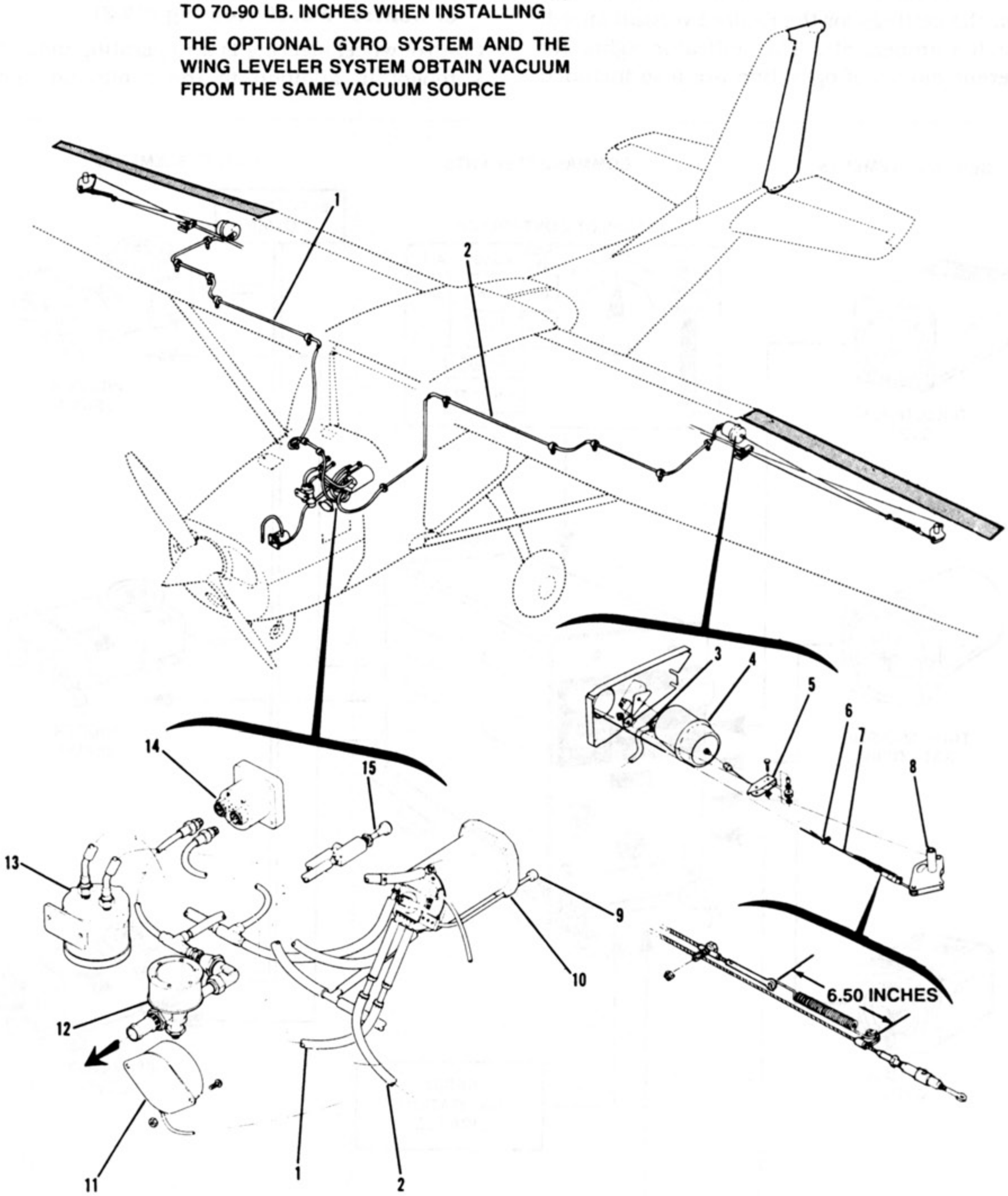

1. RIGHT AILERON VACUUM HOSE
2. LEFT AILERON VACUUM HOSE
3. NUT
4. SERVO
5. BRACKET
6. CLAMP
7. DIRECT CABLE
8. BELLCRANK
9. ROLL-TRIM KNOB
10. TURN COORDINATOR
11. INVERTER
12. VACUUM RELIEF VALVE
13. FILTER
14. SUCTION GAGE
15. ON-OFF CONTROL

Note: For illustration only. Not to be used for maintenance purposes.

Figure 4-68. A single-axis autopilot with pneumatic servos. (Courtesy Cessna Aircraft Corp.)

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

d. Flight Control Panel

The flight control panel contains the switches to activate the various functions of the autopilot and to adjust the settings for the desired vertical speed, IAS, Mach number, etc. The indicator lights for the different modes of operation are also included in the flight control panel. This panel is located in the glareshield above the center instrument panel and it is illustrated in figure 4-72.

e. Control Wheel Steering (CWS)

Control wheel steering is an operating mode for the autopilot in addition to the command operating

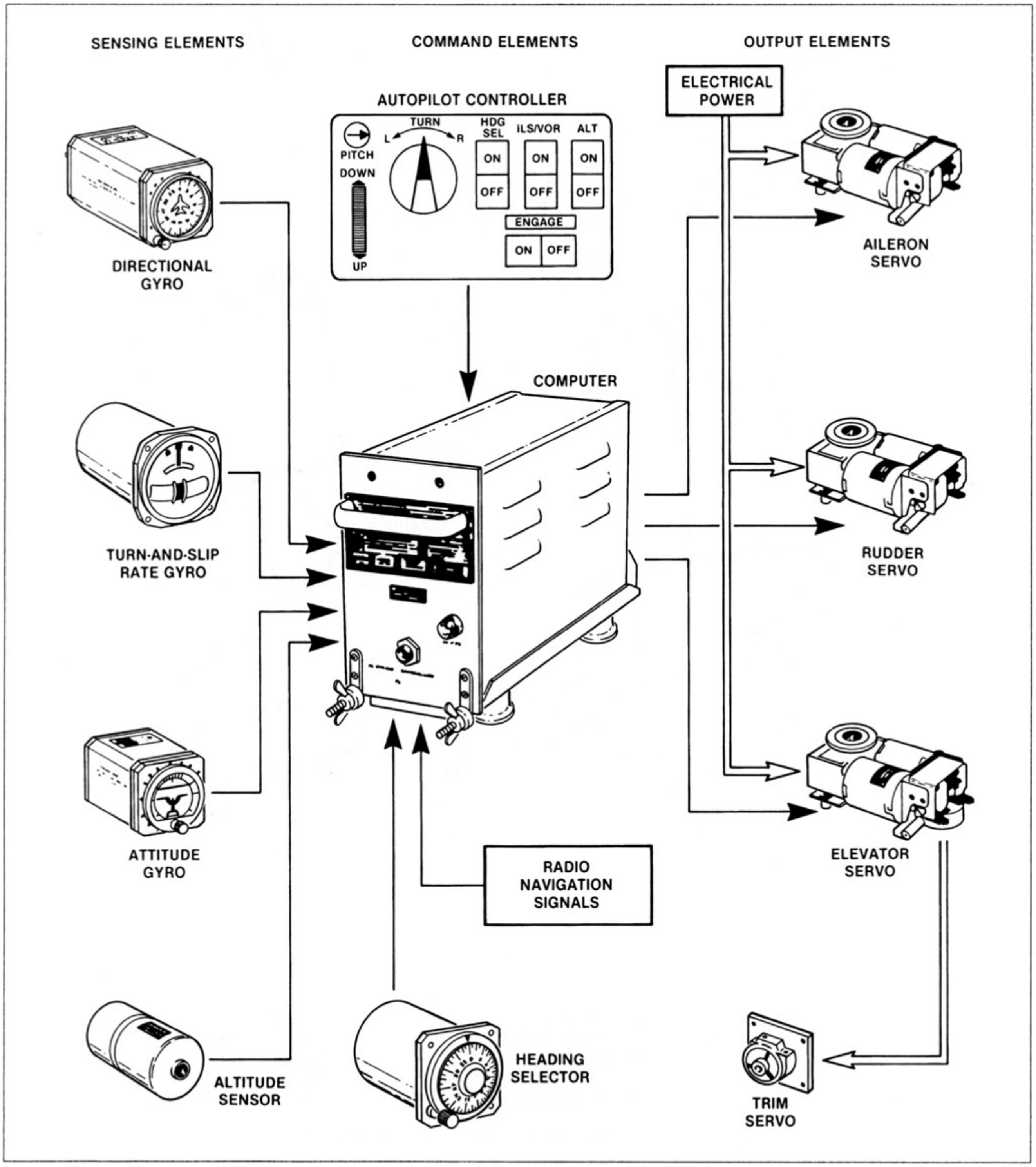
SENSING ELEMENTS
COMMAND ELEMENTS
OUTPUT ELEMENTS
AUTOPILOT CONTROLLER
PITCH
TURN
L
R
HDG SEL
ILS/VOR
ALT
DOWN
UP
ON
OFF
ON
OFF
ON
OFF
ENGAGE
ON
OFF
ELECTRICAL POWER
DIRECTIONAL GYRO
AILERON SERVO
COMPUTER
TURN-AND-SLIP RATE GYRO
RUDDER SERVO
ATTITUDE GYRO
ELEVATOR SERVO
RADIO NAVIGATION SIGNALS
ALTITUDE SENSOR
HEADING SELECTOR
TRIM SERVO

Figure 4-69. Diagram of a 3-axis autopilot that can be coupled to radio navigation receivers.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

mode. The command mode is the normal autopilot mode where the pilot does not touch the controls because the autopilot is flying the airplane. In the CWS mode, the controls are moved by the pilot as in normal flight and the force that is applied to the controls is measured and used as an input signal to the autopilot computers. In effect, the human pilot is flying the airplane, but the autopilot is helping to move the control surfaces. Figure 4-73 shows the connections between the force transducer and the flight control computer. The operation of a typical force transducer is illustrated by figure 4-74. The three electrical windings and the armature above them make up a special type of variable transformer. The AC input signal is applied to the center winding and the outer windings produce the output signal. The housing of the force transducer is flexible so that its length will change based on the force applied to it. When the housing changes in length, it causes relative motion between the armature and the coils. This motion alters the magnetic coupling and therefore produces a change in the output signal.

f. Flight Director

A flight director is a system that uses some of the basic components of an autopilot, but not all

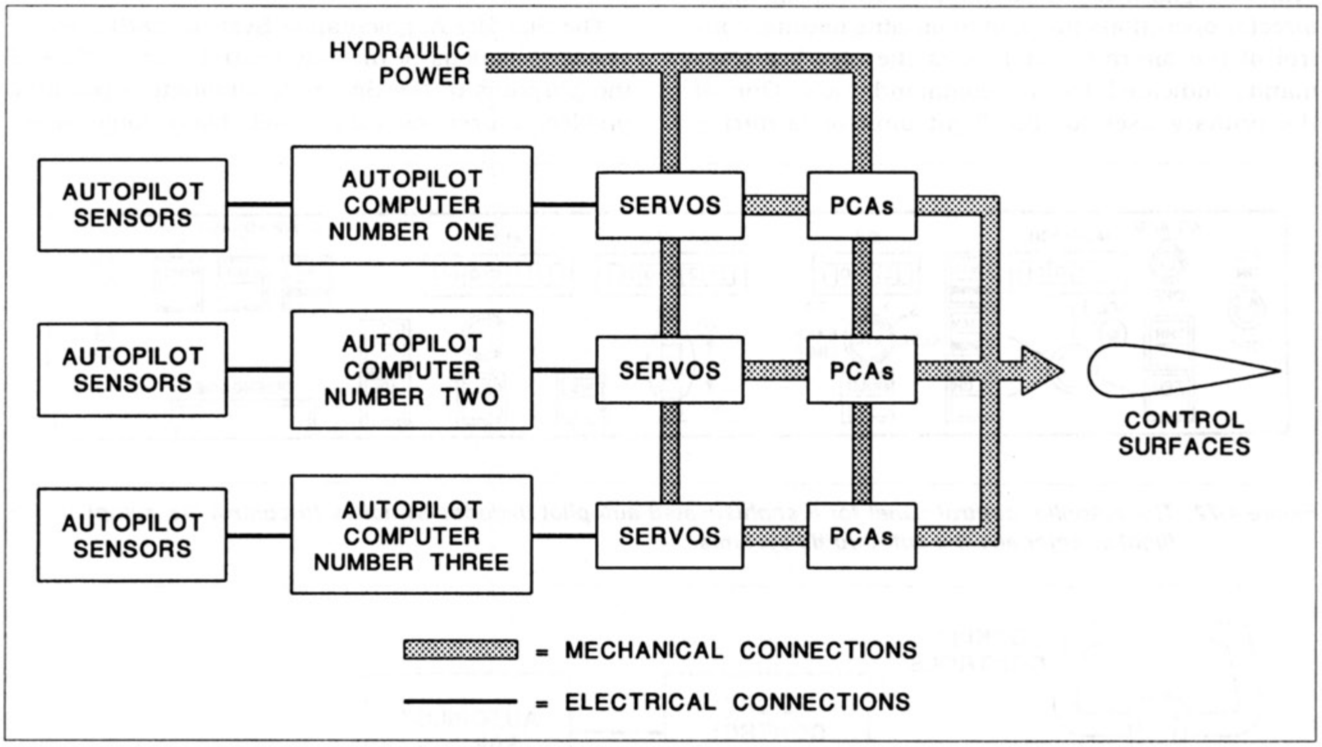
HYDRAULIC POWER
AUTOPILOT SENSORS
AUTOPILOT COMPUTER NUMBER ONE
SERVOS
PCAs
AUTOPILOT SENSORS
AUTOPILOT COMPUTER NUMBER TWO
SERVOS
PCAs
AUTOPILOT SENSORS
AUTOPILOT COMPUTER NUMBER THREE
SERVOS
PCAs
CONTROL SURFACES
= MECHANICAL CONNECTIONS
= ELECTRICAL CONNECTIONS

Figure 4-70. Block diagram of a three-channel autopilot for a large aircraft with hydraulically powered flight controls.

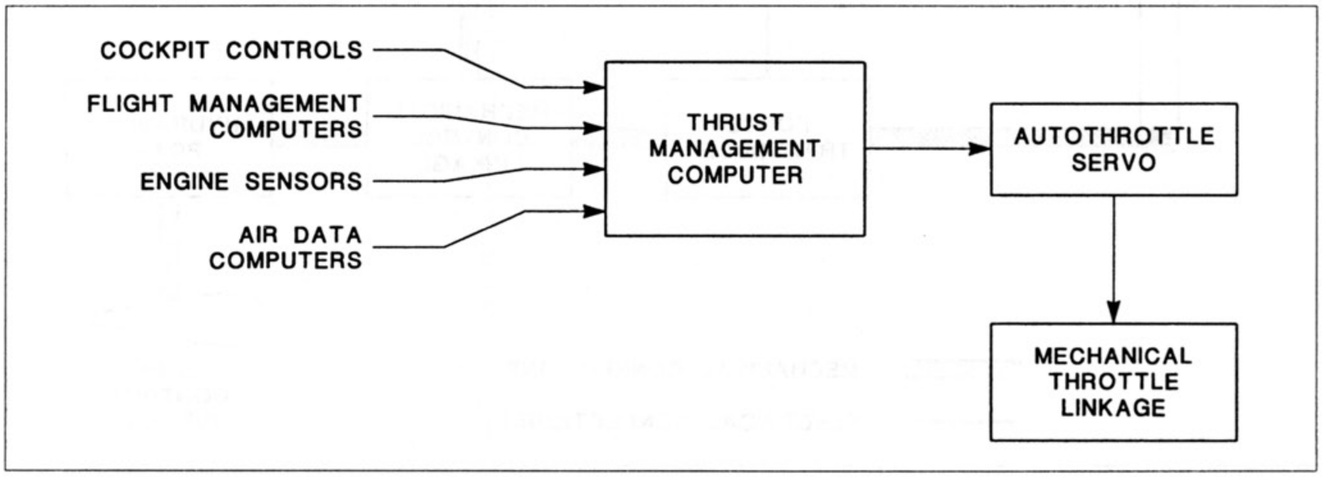
COCKPIT CONTROLS
FLIGHT MANAGEMENT COMPUTERS
ENGINE SENSORS
AIR DATA COMPUTERS
THRUST MANAGEMENT COMPUTER
AUTOTHROTTLE SERVO
MECHANICAL THROTTLE LINKAGE

Figure 4-71. Block diagram of an autothrottle system with a thrust management computer.

of them. A flight director uses sensors and computers, but it does not have servos. The flight director computer uses the signals from sensors to calculate a correction which is then displayed as a command for the pilot to follow. The commands from the flight director are displayed to the pilot on the EADI by the command bars. The operation of the command bars on the EADI is shown in figure 4-75. On the left, the command bar symbol is above the airplane symbol. The indication is that the pilot needs to raise the nose to satisfy this flight director command. On the right, the nose of the airplane has been raised so that the airplane symbol aligns with the command bar. During flight director operations the pilot maintains manual control of the aircraft, but follows the steering commands indicated by the command bars. One of the primary uses for the flight director is during an instrument approach. By using the flight director the pilot can more accurately fly the airplane on an ILS approach because the computer is making rapid calculations to predict the optimum heading and attitude for the approach. Corrections for wind drift are automatic, all the pilot has to do is follow the flight director commands.

Another condition when the flight director is helpful is in setting the proper takeoff pitch attitude.

g. Additional Features

Some of the additional features of the Boeing 767 Autopilot and Flight Director System that are typical for this class of aircraft will be described briefly.

The Stability Augmentation System (SAS) involves certain functions of the yaw control system. One of the purposes of the SAS is to eliminate a potential problem known as Dutch Roll. Many large swept

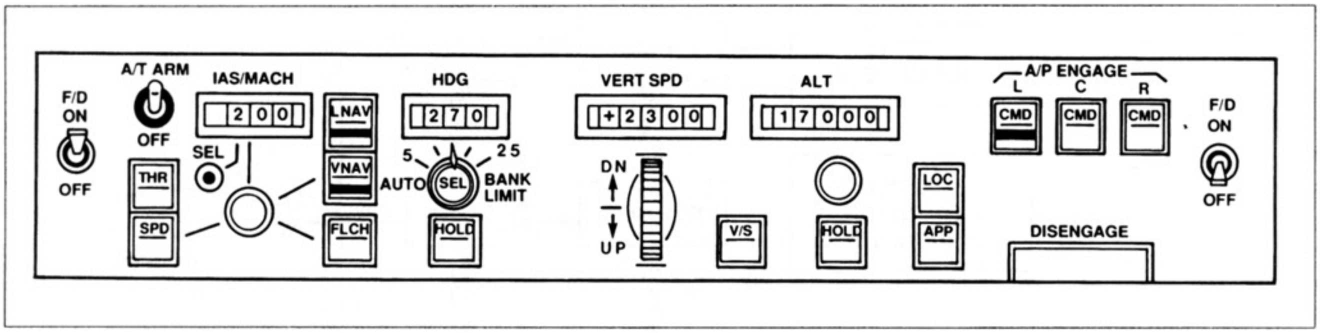
F/D
ON
OFF
A/T ARM
OFF
IAS/MACH
200
SEL
THR
SPD
LNAV
VNAV
FLCH
HDG
270
5
25
AUTO
SEL
BANK
LIMIT
HOLD
VERT SPD
+2300
DN
UP
V/S
ALT
17000
HOLD
LOC
APP
A/P ENGAGE
L
C
R
CMD
CMD
CMD
DISENGAGE
F/D
ON
OFF

Figure 4-72. The autopilot control panel for a sophisticated autopilot includes switches to control the autopilot, the flight director and the autothrottle systems.

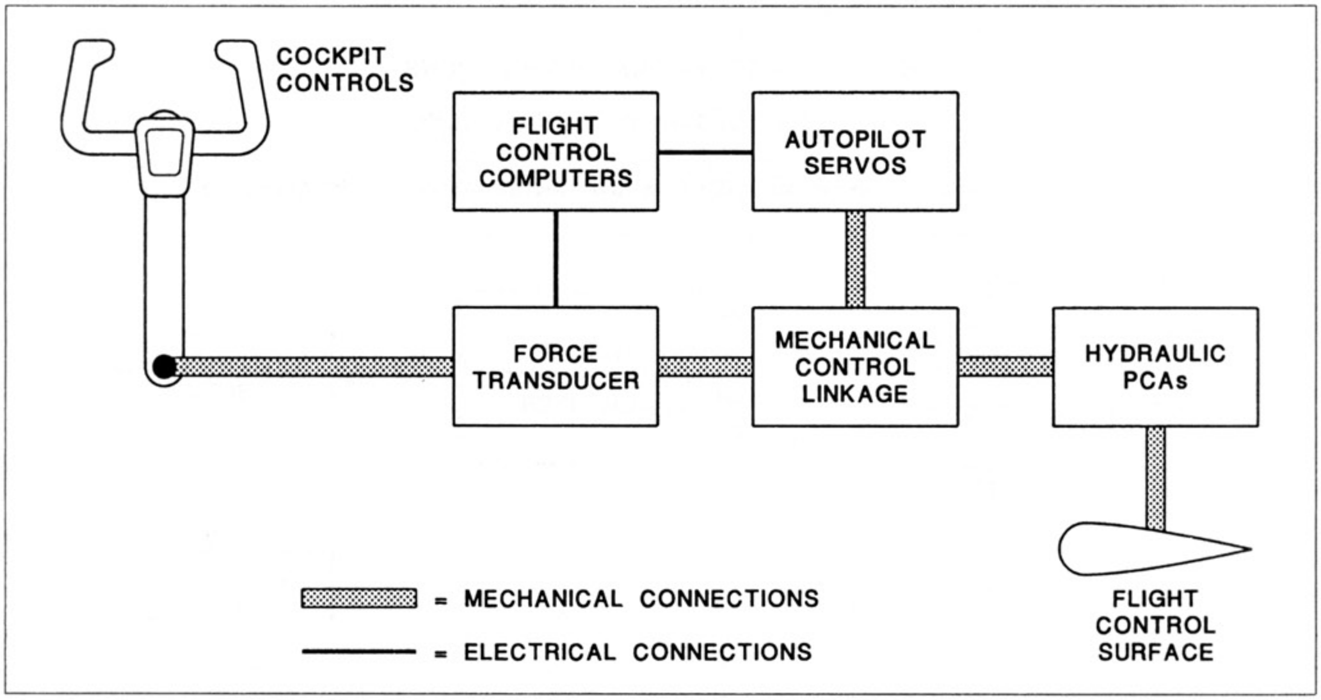
COCKPIT
CONTROLS
FLIGHT
CONTROL
COMPUTERS
AUTOPILOT
SERVOS
FORCE
TRANSDUCER
MECHANICAL
CONTROL
LINKAGE
HYDRAULIC
PCAs
FLIGHT
CONTROL
SURFACE
▨ = MECHANICAL CONNECTIONS
—— = ELECTRICAL CONNECTIONS

Figure 4-73. Location of the force transducers and servos in the control system of an air carrier jet airplane.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

wing airplanes display a peculiar type of instability at high altitudes under certain flight conditions. This can result in a continuous pitching and rolling motion known as Dutch Roll. The SAS will automatically make rapid and precise rudder movements to reduce Dutch Roll motions of the airplane. The system is also referred to as the yaw damper and it can be engaged separately from the rest of the autopilot.

The runway alignment feature of the Boeing 767 is a part of the autoland system. It will automatically align the longitudinal axis of the airplane with the runway prior to touchdown. This feature is important during crosswind landings because it is designed to prevent the airplane from landing at a crab angle to the runway. This is illustrated in figure 4-76. The runway alignment feature is limited to control surface deflections of 25° for the rudder and 2° for ailerons. A very strong crosswind at 90° to the runway could not be completely counteracted. This same feature will also provide

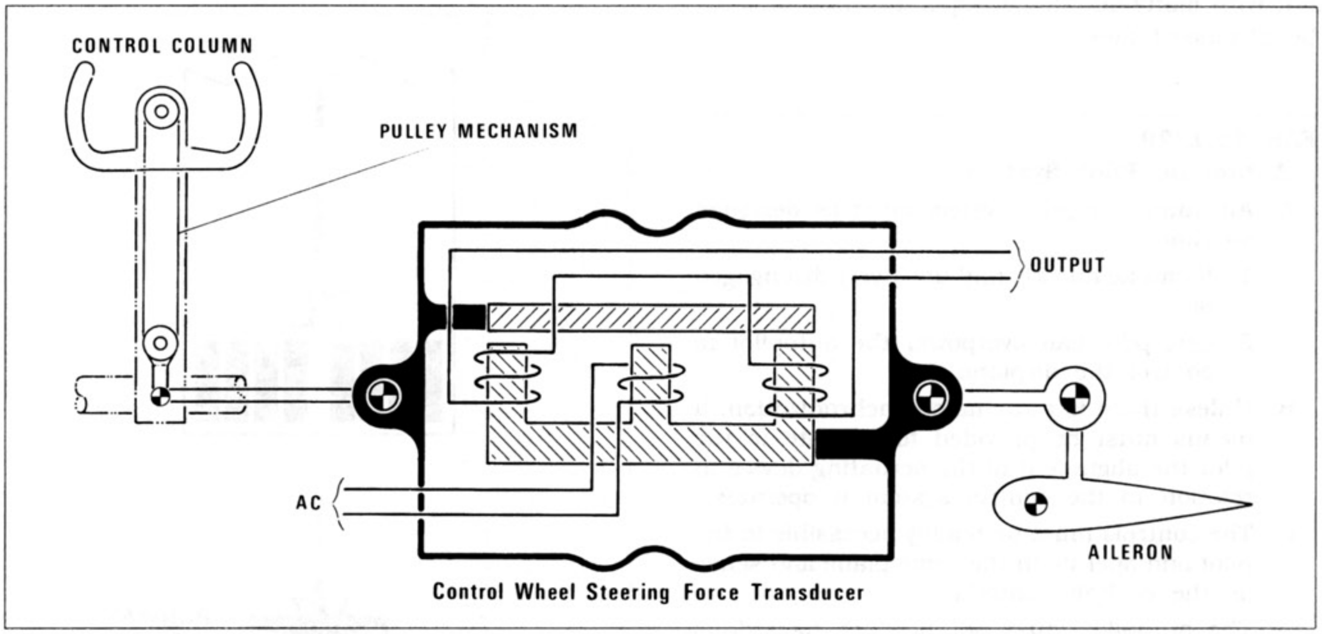
CONTROL COLUMN
PULLEY MECHANISM
OUTPUT
AC
AILERON
Control Wheel Steering Force Transducer

Figure 4-74. Basic operation of one type of force transducer.

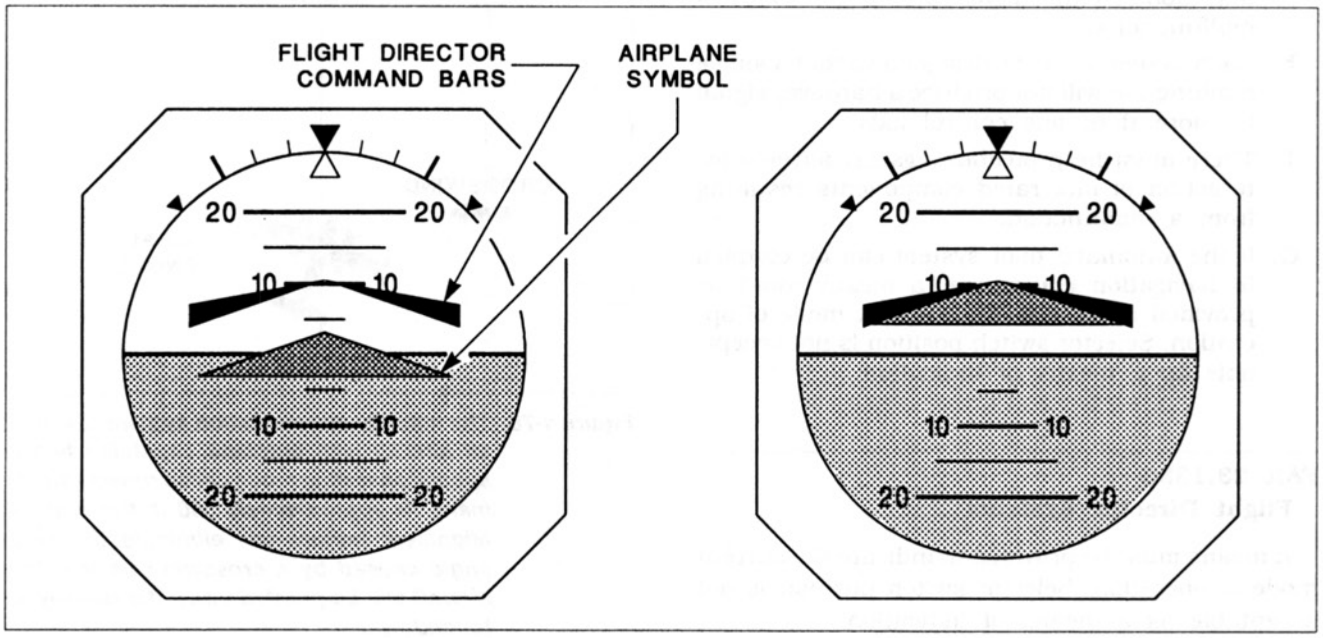
FLIGHT DIRECTOR COMMAND BARS
AIRPLANE SYMBOL
20
20
10
10
10
10
20
20
20
20
10
10
10
10
20
20

Figure 4-75. The command bars in an ADI are positioned by signals from the flight director computer. The pilot follows the commands by aligning the airplane symbol with the command bars.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

corrections if an engine fails during the landing approach. It would supply the control corrections to counteract the asymmetrical thrust situation.

7. FARs for Autopilots

Some of the Federal Aviation Regulations that apply to autopilots and related systems will be given here. Most of the references are to FAR Part 23 and FAR Part 91 which apply mainly to small airplanes. The requirements for transport category aircraft autopilots include some of these same requirements, but with many additional requirements that won't be discussed here.

FAR 23.1329

Automatic Pilot System

A. An automatic pilot system must be designed so that:
    1. It can be quickly and positively disengaged or
    2. One pilot can overpower the autopilot to control the airplane.

B. Unless there is automatic synchronization, a means must be provided to indicate to the pilot the alignment of the actuating device in relation to the control system it operates.

C. The controls must be readily accessible to the pilot and operate in the same plane and sense as the cockpit controls.

D. The autopilot must not produce hazardous loads on the airplane or produce hazardous deviations in the flight path in the event of malfunctions.

E. Each system must be designed so that a single malfunction will not produce a hardover signal in more than one control axis.

F. There must be protection against adverse interaction of integrated components resulting from a malfunction.

G. If the automatic pilot system can be coupled to navigation equipment, a means must be provided to indicate the current mode of operation. Selector switch position is not acceptable as a means of indication.

FAR 23.1335

Flight Director Systems

A means must be provided to indicate the current mode of operation. Selector switch position is not acceptable as a means of indication.

FAR 91 Appendix A

Category II Operations: Required Instruments and Equipment

Refer to figure 4-77 for an example of the weather minimums associated with Category II and Category III ILS approaches.

2A9 An automatic pilot approach coupler or a flight director system is required for Category II ILS operations.

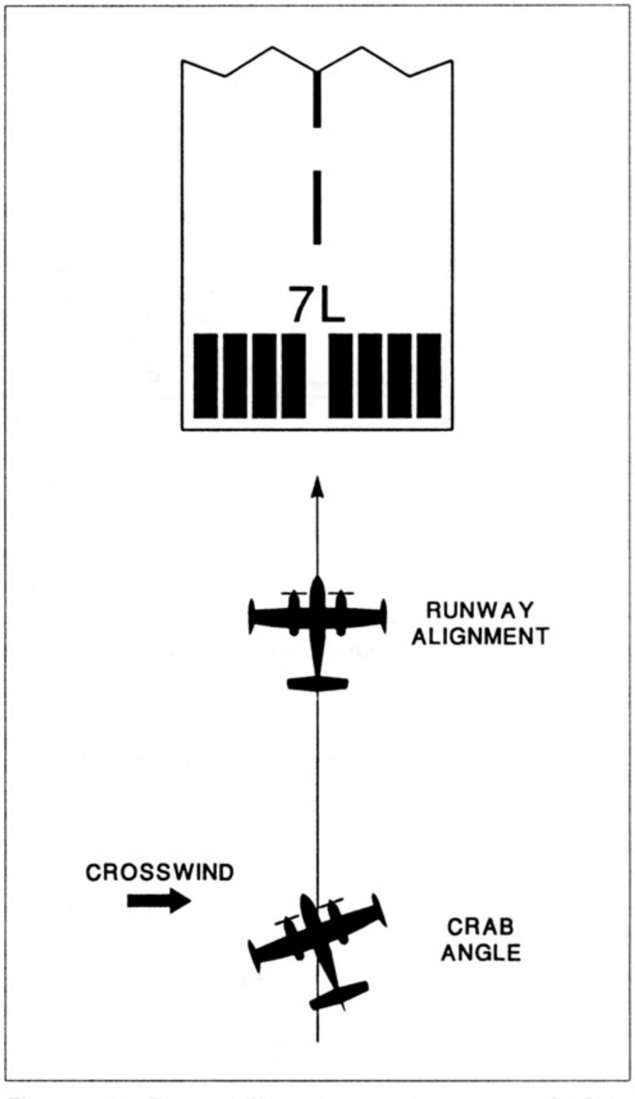

***Figure 4-76.* The stability augmentation system (SAS) in an aircraft with autoland capability has a special operating mode called runway alignment. At approximately 500 ft. the runway alignment feature will eliminate the crab angle caused by a crosswind so that the aircraft will be pointed down the runway at touchdown.**

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

2B1 For Category III ILS operations additional equipment requirements are a radio altimeter and autothrottle system.

FAR 135.105

Autopilot in Place of Second Pilot

The certificate holder may use an autopilot in place of a second pilot if:

A. The autopilot and its operation are approved.

B. The Administrator issues an approval amendment.

C. The autopilot is a 3-axis type.

D. The certificate holder shows that operations can be conducted safely.

E. Exceptions: A second pilot must be used:

  1. For IFR operations.
  2. For Category II approaches.
  3. If required by the aircraft manufacturer.

---

8. Autopilot Maintenance

The information in this section is not meant to relate to any particular aircraft autopilot system. The procedures are general and could be applied to most autopilots as appropriate. Maintenance of autopilots consists of visual inspections, replacement of components, cleaning, lubrication, troubleshooting and operational checkouts of the system.

An operational check of the autopilot should be performed whenever an autopilot is installed, when components are replaced and whenever a malfunction is suspected. Many things can be operationally checked on the ground, but some situations may require a test flight with an airborne checkout of the autopilot. Some general procedures for a ground checkout of an autopilot are as follows.

With the autopilot disengaged, manipulate the flight controls to see if they function smoothly and without excessive drag or interference from autopilot components. The alignment of the autopilot to the aircraft should be checked. This normally involves checking such things as cable tension, torque settings, dimensional adjustments etc.

ILS APPROACH CATEGORY	RUNWAY VISUAL RANGE (RVR)	DECISION HEIGHT (ALTITUDE)
Category I	2400 feet	200 feet
Category II	1200 feet	100 feet
Category IIIA	700 feet	None
Category IIIB	150 feet	None
Category IIIC	None	None

Figure 4-77. The airport weather minimums for the various ILS approach categories.

Before engaging the autopilot for an operational check, allow the gyros to come up to speed. This normally requires from 2–4 minutes. After engaging the autopilot, the following checks should be made.

1. Rotate the turn knob on the controller to the left. The rudder pedals and control column should move in the proper directions to indicate a left turn. The motion should be smooth and without excessive binding, jerking or hesitation.
2. Rotate the turn knob to the right and watch for the proper operation and motion of the controls.
3. Rotate the pitch knob up and down and watch for the correct motion of the control column aft and forward.
4. If the autopilot has automatic pitch trim, check the proper motion of the trim control as the control column moves fore and aft. When the control column moves back, the system should apply nose up trim and vice versa.
5. With the autopilot engaged, try to overpower it by grasping the controls and applying force. It should be possible to overpower the autopilot if it is adjusted properly.
6. Check all of the controls and switches for proper actuation and correct indications.
7. It may be desirable to taxi the aircraft in order to check out some of the operating modes. If you engage the heading hold mode and make a taxi turn to the right, the controls should show motion commanding a turn to the left.
8. Check the autopilot disconnect switches to ensure that the autopilot disconnects rapidly and positively. There may be several ways to disconnect the autopilot; check them all.
9. If the aircraft has a flight director, check for proper indications by the command bars in the ADI or EADI. Check the autopilot mode indicators in the ADI or EADI if so equipped.

If the aircraft has both an autopilot and flight director, they can be checked against each other as an aid in troubleshooting. The autopilot and flight director share some components while others are only used by the autopilot. This can be used to help locate the source of the problem when malfunctions are suspected. A ground checkout can often help to locate the source of a problem by comparing the indications of the flight director and autopilot. If the flight director is commanding an incorrect control movement and the autopilot is moving the controls in the same incorrect direction, then the fault is most likely a component they share such as the sensors or computer. If the flight director shows a correct command for nose up pitch, but the autopilot does not move the controls to agree with this command then the problem is likely not in the sensors or computer.

Some complaints about autopilot malfunctions are the result of faults in components other than the autopilot itself. If a pilot reports that the autopilot will not track a VOR radial, the problem could be a wiring fault in the wires that carry radio signals to the autopilot rather than a problem with the autopilot itself. Conditions such as rigging problems or binding of the main control cables in the aircraft itself can adversely affect the operation of the autopilot. Because the interactions and interconnections associated with autopilots can be very complex, a good system schematic and a thorough knowledge of the autopilot are necessary for efficient troubleshooting and maintenance of autopilot systems.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

APPENDIX A — Glossary

Accelerometer — A device or instrument which measures acceleration forces. The unit of measure is most often "Gs" based on the acceleration of gravity.

Aircraft — A machine that operates in the air. Examples include airplanes, gliders, balloons, airships and rotorcraft.

Airplane — An engine-driven, fixed-wing aircraft that is heavier than air and supported in flight by wings.

Altitude engine — A reciprocating aircraft engine that employs a supercharger to maintain sea level power at higher altitudes.

Annunciator — The indicating lights that are used to alert crew members to operating conditions that they need to be aware of. As in warning, caution and status indicator lights.

Appliance — Any instrument, mechanism, equipment, part, apparatus, appurtenance or accessory, including communications equipment, that is used in operating or controlling an aircraft in flight and is not part of an airframe, engine or propeller.

Artificial horizon — An aircraft instrument used to indicate pitch and roll attitudes. A gyro instrument used for IFR and also known as a gyro horizon, attitude gyro, bank and pitch indicator etc.

Autosyn — A synchro system that transmits positional information electrically from one place to another. The Autosyn type uses AC power and electromagnet rotors in both the sending and receiving units.

Avionics — Aircraft electronic equipment. Most often refers to aircraft radios and similar components.

Balun — A special type of impedance matching device used with V-shaped aircraft VOR antennas.

Bellows — A mechanical device for measuring pressure. It is a thin metal container with accordion shaped folds on the sides. Often separated into two chambers for use as a differential pressure measuring sensor.

Bonding jumper — A metal conductor used to electrically connect two parts of an aircraft. Most often a wire braid or thin metal strap.

Bourdon tube — A curved, hollow metal tube used as a pressure measuring sensor. Used for relatively higher pressures than a bellows or diaphragm.

Calibrated airspeed — Indicated airspeed that has been corrected for airspeed system errors and errors caused by the location and alignment of the pitot and static ports or sensors.

Category II operation — An ILS instrument approach using the specified procedures and meeting the requirements for operations in weather conditions below those required for a Category I approach.

Category III operation — An ILS instrument approach using the additional procedures and requirements for operations in weather conditions below those required for a Category II approach.

Class B Airspace — This is the same as a TCA (termimal control area), the airspace around a major airport where special equipment and flight restrictions apply.

Class C Airspace — this is the same as an ARSA (airport radar service area), the airspace around a busy airport where special equipment and flight restrictions apply.

Coaxial cable — A special conductor having a center conductor, a layer of dielectric insulation and an outer conductor. Designed to carry radio frequency electrical signals as in an antenna lead.

Compass dip — The tilting of the float in a magnetic compass due to the curvature of the earth. Compass dip becomes greater as the distance from the equator increases. Magnetic compass acceleration error and north turning error are a result of compass dip.

Compensator — A device to adjust or compensate for some error in an aircraft instrument or radio system. The compensator magnets in an aircraft compass installation.

Critical engine — The engine whose failure would most adversely affect the peformance or handling qualities of an aircraft.

Dashpot — A small acceleration operated air pump used in an aircraft IVSI to decrease the lag in the indication.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Deviation — An error in a magnetic compass caused by magnetic fields in the aircraft. This error is minimized and recorded in a procedure called swinging the compass.

Diaphragm — A small lens shaped metal container that is used as a sensor to measure pressures. Used in altimeters and airspeed indicators because of its sensitivity.

Doppler effect — The apparent change in frequency of a sound wave or radio wave when there is relative motion between the observer and the source of the waves. Used in the Doppler navigation system by bouncing radar waves off the surface of the earth. The Doppler principle is also used on the latest types of weather radar systems on aircraft.

Drip stick — A fuel quantity measuring system that uses a tube or stick that is extended below the wing of the airplane. It is a mechanical measuring system that requires no outside power.

Encoding altimeter — An altimeter or sensor that supplies electrical outputs related to aircraft altitude. The output is supplied to the aircraft transponder which transmits the data to a ground station. Also referred to as Mode C.

Erector mechanism — A device or mechanism in an artificial horizon that maintains the vertical spin axis of the gyro to ensure accurate readings.

Flight level — A level of constant atmospheric pressure related to a reference level of 29.92 in. Hg. Flight levels are indicated in hundreds of feet as in FL310, FL330, FL350 etc. Aircraft must use flight levels rather than MSL altitudes at and above 18,000 ft.

Flux valve — The sensor for a flux gate compass system. The flux valve produces an electrical output from the earth's magnetic field to operate remote compass systems and other devices.

Gyroscope — A small wheel or rotor that is spun at high speed to operate aircraft intruments. Also used to stabilize certain components such as radar antennas and INS accelerometers.

Heading indicator — A gyro instrument that indicates the magnetic heading of the aircraft. Also called the DG, gyro compass, heading gyro etc.

IFR conditions — Weather conditions below the minimum for VFR flight.

Inclinometer — An aircraft instrument that indicates slips and skids. Also called the slip-skid indicator. Usually a small curved glass tube with a ball in it.

Indicated airspeed — The speed of an aircraft as shown on its pitot-static airspeed indicator calibrated to reflect standard atmospheric adiabatic compressible flow at sea level and uncorrected for airspeed system errors.

Instrument — A device using an internal mechanism to show visually or aurally the altitude, attitude, or operation of an aircraft or aircraft part. It includes electronic devices for automatically controlling an aircraft in flight (autopilots).

Knots — A velocity in units of nautical miles per hour. To convert from knots to MPH, multiply by 1.151.

Kollsman window — The small window on the face of an altimeter that indicates the barometric pressure for which the altimeter has been set with the setting knob. Unit of measure is inches of mercury or millibars.

Large aircraft — Aircraft of more than 12,500 lbs. maximum certificated takeoff weight.

Logic gate — A semi-conductor device that produces an output from one or more inputs. It uses digital signals and is a basic component in many aircraft electronic systems.

Loran C — A radio navigation system that uses signals transmitted from ground stations on a frequency of 100 KHz. Most often used by general aviation aircraft.

Lubber line — The line on a magnetic compass that is used to take readings. The numbers on the compass card are read against the lubber line.

Mach number — The ratio of true airspeed to the speed of sound for the specified flight condition and altitude.

Magnesyn — A synchro system used on aircraft. The Magnesyn system uses AC power and permanent magnet rotors in the sending and receiving units.

Manifold pressure — The absolute pressure measured at the appropriate point in the induction system of a reciprocating aircraft engine and expressed in inches of Mercury.

Octal — A binary code system used with integrated circuits and logic gates. The octal code uses three bit bytes.

Omega — A radio navigation system that employs ground stations that transmit on frequencies from 10-14 KHz. It provides world wide coverage. VLF ground sites are also used by aircraft Omega systems.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Piezoelectric — A device that employs a crystal sensor that produces electrical output signals when it is squeezed or vibrated. An example is the piezoelectric sensor for an engine vibration indicator.

Polarization — The orientation of the electric field portion of an electro-magnetic radio wave. Vertical polarization would mean that the electric field is vertical relative to the earth's surface.

Precession — The displacement of a gyro rotor by an outside force. The gyro will precess as if the force was applied 90° ahead in the direction of rotation from the actual point of application of the disturbing force.

Pressure Altitude — The altitude shown on an aircraft altimeter when 29.92 is set into the barometric pressure setting window. This represents the altitude in a Standard Atmosphere.

Proximity sensor — A sensor or transducer used to replace microswitches in an aircraft position indicating system. It is an electronic device with no moving parts and is considered to be more reliable than microswitches.

Quadrantal error — The error in indication for an aircraft ADF system. A check of quadrantal error should be made whenever an antenna is installed or other maintenance is performed which could affect the accuracy of the ADF system.

Radar — Radio detection and ranging. Used in aviation for ATC purposes, weather avoidance systems, navigation and precision altitude measurement.

Ratiometer — An electrical circuit used to operate an aircraft instrument. It is called a ratiometer because the pointer is positoned by the ratio of the field strength of two electromagnetic fields.

Rigidity — The characteristic of a gyro rotor that causes it to try to maintain its spin axis fixed in space.

Sea level engine — A reciprocating aircraft engine having a rated takeoff power that is producible only at sea level. An unsupercharged engine.

Selsyn — A synchro system used on aircraft. The Selsyn uses DC power with a variable resistor in the sending unit and a permanent magnet rotor and three section coil in the receiving unit.

Slip-skid indicator — The same as inclinometer. See above.

Small aircraft — Aircraft of 12,500 lbs. or less maximum certificated takeoff weight.

Thermocouple — A device which uses two different metals to produce a DC output at the cold junction when the hot junction is heated. Commonly used for CHT, EGT and other relatively high temperature measurements in an aircraft.

Torquemeter — An instrument system that measures torque delivered to a shaft, usually by the aircraft engine. Common on large recip engines, turboprop engines and turboshaft engines.

True Airspeed — True airspeed is calibrated airspeed that has been corrected for altitude and temperature effects. An airspeed indicator is designed to be accurate for the standard pressure and temperature at sea level. At higher altitudes, indicated or calibrated airspeed is less than true airspeed.

Variation — The apparent error in the indication of a magnetic compass caused by the fact that the north geographic pole and the north magnetic pole of the earth are not in the same location.

Venturi — A tube with curved inner walls that produces a reduction in pressure in accordance with Bernoulli's Principle. Used in aircraft to produce suction to operate gyros and as a jet pump.

Waveguide — A hollow tube used as a conductor for radar frequency EM waves. Usually rectangular in cross-section and found in aircraft weather radar systems.

Wheatstone bridge — A bridge circuit with three fixed resistors and a variable resistor. It is used to operate a meter movement that rotates a pointer in an aircraft instrument for temperature measurements.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

APPENDIX B — Abbreviations

ADC — *Air data computer*: A computer which processes inputs from pitot tubes, static ports and TAT probes and provides outputs to the typical pitot-static instruments as well as other aircraft systems.

ADF — *Automatic direction finder*: A radio navigation system using signals in the LF and MF bands.

ADI — *Attitude director indicator*: An instrument that combines pitch and roll data with the command bars of a flight director.

AF — *Audio frequency*: 20,000 Hz and below.

AFCS — *Automatic flight control system*: Usually refers to more advanced autopilots that include features such as auto-throttle and stored flight plan routes.

AFM — *Aircraft flight manual*: An FAA-approved manual that lists operating requirements. Useful to mechanics as well as to pilots.

AGL — *Altitude above ground level*.

AM — *Amplitude modulated*: This refers to a radio wave that has been modulated in such a way that the amplitude of the signal varies up and down to match the modulating signal.

AOA — *Angle-of-attack*: The angle between the chord line of an airfoil and the relative wind. Also AOA instrument presentation.

ARINC — *Aeronautical Radio Incorporated*: This is an organization made up of airlines and manufacturers which establishes standards for aircraft equipment. For example, ARINC 429 for digital data.

ARSA — *Airport radar service area*: An area of airspace around an airport with radar air traffic control (ATC) in which special restrictions are placed on aircraft flight operations.

ATC — *Air traffic control*.

BCD — *Binary coded decimal*: A binary code system that uses four bit bytes.

BIT — Each individual binary digit or number in a digital word or message.

BYTE — A group of binary bits which are treated together as in a binary word with 32 bits. In this example, one byte = 32 bits.

CADC — *Central air data computer*: Same as ADC.

CAS — *Calibrated airspeed*.

CAT II — Category II ILS instrument approaches.

CAT III — Category III ILS instrument approaches.

CFM — *Cubic feet per minute*: A measure of flow rate used with air-operated gyros.

CHT — *Cylinder head temperature*: An instrument found on many aircraft with air-cooled reciprocating engines, usually a thermocouple system.

CIT — *Compressor inlet temperature*: Also called Tt_2, it refers to the temperature of the air entering the inlet of a turbine engine.

CPU — *Central processing unit*: One of the major components of a computer, the CPU contains the ALU, control and memory functions.

CRT — *Cathode ray tube*: This is the display for electronic aircraft instruments. It looks like a TV screen, but is specially designed to be readable in the bright conditions of the cockpit.

CVR — *Cockpit voice recorder*: Records cockpit sounds and conversations from radio and intercom systems.

DG — *Directional gyro*: An instrument which gives information concerning aircraft rotation about the vertical axis. Usually referenced to magnetic headings.

DME — *Distance measuring equipment*: A two-way radio system for determining aircraft distance in nautical miles from a ground site.

EADI — *Electronic attitude director indicator instrument*: A display which combines pitch and roll data along with indications from the flight director in the form of command bar movements. Other data such as radio navigation displays are also included. A CRT instrument.

EGT — *Exhaust gas temperature*: An instrument which displays the temperature of the engine exhaust gasses. Found on both reciprocating and turbine engines.

EHSI — *Electronic horizontal situation indicator*: A display which combines gyro stabilized magnetic heading information along with radio navigational information using a deviation bar indicator. A CRT instrument.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

EICAS — *Engine indication and crew alerting system*: An electronic instrument system that provides indications for powerplant and aircraft system instruments, and provides alert, caution and status messages for the crew. EICAS typically uses two CRTs.

ELT — *Emergency locator transmitter*: A small self-contained radio transmitter for crash location purposes found on most small aircraft.

EM WAVES — *Electro-magnetic waves*: Most often used to mean radio waves.

EPR — *Engine pressure ratio*: An instrument which indicates the power being produced by certain turbojet and turbofan engines. EPR is the ratio of total outlet pressure divided by total inlet pressure.

FCC — *Federal Communications Commission*: This government agency establishes rules for many types of electronic equipment including the radio equipment on aircraft.

FDR — *Flight data recorder*: A system which records many different operating parameters such as altitude, airspeed, engine power, G loadings, flap settings, etc. Used for accident investigation.

FM — *Frequency modulation*: A radio carrier wave uses FM when the carrier wave frequency is varied up and down by the modulating signal.

FMS — *Flight management system*: A sophisticated autopilot system that includes advanced features for managing virtually the entire flight. Uses an FMC (flight management computer).

FPM — *Feet per minute*: The standard unit of measurement for aircraft rate of climb indicators and similar devices.

GPH — *Gallons per hour*: A standard unit of measurement for fuel flow or fuel consumption for aircraft. Usually used for reciprocating engines.

GPS — *Global positioning system*: A satellite navigation system being developed for the military, but available for use by civilian aircraft.

GPWS — *Ground proximity warning system*: Designed to give a warning to the flight crew to avoid ground impact due to excessive rates of descent, rising terrain, etc.

HSI — *Horizontal situation indicator*: An integrated aircraft instrument which displays magnetic heading, radio navigation steering information and sometimes additional information. It replaces the simpler DG instrument.

IAS — *Indicated airspeed.*

IC — *Integrated circuit*: A semi-conductor device that incorporates a number of logic gates in one compact unit.

ICAO — *International Civil Aviation Organization*.

IFR — *Instrument flight rules*: Aircraft must operate IFR if weather conditions are below the minimums for visual reference flying.

ILS — *Instrument landing system*: A precision approach using radio guidance signals to guide an aircraft to a landing runway.

INS — *Inertial navigation system*: A navigational system that uses very accurate measurements of acceleration to calculate aircraft position, course and speed.

IRU — *Inertial reference unit*: Most often refers to the laser device which is the sensor for INS and other aircraft systems.

ITT — *Inter-turbine temperature*: This refers to a gas temperature measurement on a turbine engine where the probes are located in between two different sections of the turbine.

IVSI — *Instantaneous vertical speed indicator*: An instrument that eliminates the lag of a conventional VSI through the use of acceleration operated dashpots.

KIAS — *Knots indicated airspeed*: We also find KCAS and KTAS.

LMM — Compass locator transmitter at the middle marker.

LCD — *Liquid crystal display*: A common type of device used in aircraft instruments and radios that have a lighted display.

LED — *Light emitting diode*: A common type of device used in lighted displays on radios and other equipment.

LOM — Compass locator transmitter at the outer marker.

LRU — *Line replaceable unit*: Modern aircraft have most electronic equipment installed in the form of LRUs which are standard size boxes that contain the equipment and which make replacement and maintenance simpler and more efficient.

MLS — *Microwave landing system*: A new type of precision approach aid that may eventually replace ILS. In limited use at this time.

MM — *ILS middle marker*.

MSL — Altitude in terms of mean sea level.

N_1 — A tachometer indication of the low pressure compressor speed in a turbine engine.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

N_2 — A tachometer indication of the high pressure compressor speed in a turbine engine.

N_g — A tachometer indication of the gas producer speed in a turbine engine. Usually turboprop or turboshaft.

N_p — A tachometer indication of the power section speed in a turbine engine. Usually turboprop or turboshaft.

N_r — A tachometer indication of the rotor speed for a helicopter. Usually the main rotor.

NDB — *Non-directional radio beacon*: The ground based radio transmitter that sends the signals which are received by the aircraft ADF radio navigation receiver.

OM — *ILS outer marker*.

POH — *Pilot's operating handbook*: This is an FAA-approved document which gives operating information for that particular aircraft. Used by both pilots and mechanics. Same as AFM.

PPH — *Pounds per hour*: A unit of measurement for aircraft fuel flow and fuel consumption. Most often used with turbine engines.

PSIA — *Pounds per square inch absolute*: A measurement of pressure compared to a perfect vacuum.

PSID — *Pounds per square inch differential*: A measurement of the differential pressure between two pressures measured at different points.

PSIG — *Pounds per square inch gauge*: A measurement of pressure compared to ambient conditions, usually ambient atmospheric pressure.

P-static — *Precipitation static*: The static electricity charge on an aircraft produced by friction with ice, snow, rain, sand, dust etc. It can cause noise in the radios and other problems.

RAM — *Random access memory*: The memory in a computer that can be affected by operator input and is lost when the computer is turned off.

RF — *Radio frequency*: Refers to frequencies above audio frequencies, or those frequencies above 20,000 Hz (20 KHz).

RLG — *Ring laser gyro*: A laser beam device that can be used to replace spinning gyroscopes to operate aircraft instruments and other aircraft systems.

RNAV — *Area navigation*: The use of a computer to process signals from VOR and DME transmitters, permitting random direct routes to be flown using waypoints designated as a VOR radial and distance such as OMN 243/24.

ROC — *Rate of climb*: An aircraft instrument that gives readings in FPM of the aircraft rate of climb or descent.

ROM — *Read only memory*: Sometimes called hard-wired, this is the memory in a computer which cannot be changed by the operator and is not lost when the computer is turned off.

RVR — *Runway visual range*: A measured visibility along a runway, stated in feet. RVR 2400 = 1/2 mile visibility.

SAS — *Stability augmentation system*: This system is often associated with an autopilot and it is designed to provide additional stability to the aircraft for certain flight conditions. On swept wing jets, for example, it helps to reduce Dutch roll.

SELCAL — *Selective calling*: A communications system which allows the person on the ground to dial a code to contact a specific airplane in flight. Used by the airlines to contact their aircraft for operational reasons.

SWR — *Standing wave ratio*: A measure of the efficiency of an antenna, it is based on forward power and reflected power measurements.

TACAN — *Tactical air navigation*: a radio navigation system in the UHF band designed primarily for military aircraft.

TAS — *True Airspeed*.

TCA — *Terminal control area*: An area of airspace around a busy airport where special restrictions are placed on aircraft flight operations.

TCAS — *Traffic alert and collision avoidance system:* This system is installed in some larger aircraft where it gives warnings to the pilots to prevent mid-air collisions. The TCAS uses transponder principles of operation.

TIT — *Turbine inlet temperature*: Refers to the measurement of gas temperature for a turbine engine where the probes are located just downstream of the combustion chambers or just in front of the first turbine stage. Can also refer to a turbosupercharger temperature.

TOT — *Turbine outlet temperature*: Refers to the measurement of gas temperature for a turbine engine where the probes are located downstream of all the turbine sections. Also called EGT.

V-speeds — Designated airspeeds related to a specific aircraft certification requirement or operating airspeed.

V — Velocity.

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Va — Design maneuvering speed.

Vd — Design dive speed.

Vfe — Maximum speed with the flaps extended.

Vle — Maximum speed with the landing gear extended.

Vlo — Maximum landing gear operating speed.

Vmc — Minimum control speed: The lowest speed at which directional control of the aircraft can be maintained with the critical engine failed and the remaining engines at maximum continuous power.

Vmo/Mmo — Maximum operating speed: In terms of airspeed and in terms of Mach number.

Vne — Never exceed speed: the maximum permissable speed under any circumstances.

Vno — Maximum structural cruising speed.

Vs$_1$ — The stalling speed or minimum steady flight speed with the landing gear and flaps retracted.

Vso — The stalling speed or minimum steady flight speed with the landing gear and flaps extended.

Vx — The speed for best angle of climb.

Vy — The speed for best rate of climb.

Vyse — The speed for best rate of climb, single engine operations with one engine inoperative.

VFR — *Visual flight rules*: for VFR the pilot must be able to control the aircraft by visual outside references.

VHF — *Very high frequency*.

VOR — *VHF omnidirectional radio range*: a radio navigation system.

VORTAC — A combined VOR and TACAN transmitter site.

VSI — *Vertical speed indicator*: an aircraft instrument that indicates aircraft rate of change of altitude in FPM.

VSWR — *Voltage standing wave ratio*: Same as SWR (see above).

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

INDEX

A

- Absolute pressure 14
- Acceleration error 42
- Accelerometer 23
- Accelerometer operated dashpots 22
- Air data computer systems 28
- Airspeed
  - calibrated 17
  - indicated 17
  - true 17
- Airspeed indicator markings 19
- Airspeed placards 12
- Altimeter
  - encoding 16
  - non-sensitive 14
  - radar 141
  - sensitive 15
- Altimeter system tests and inspections 24, 25
- Altitude alerting system 100
- Amplifiers 107
- Amplitude Modulation (AM) 106
- Angle of attack indicators 96
- Annunciators 98
- Antenna couplers 112
- Antenna interference 166
- Antenna separation, VHF comm 166
- Antennas, types of 170
  - DME/transponder 173
  - ELT 173
  - glideslope 171
  - GPS 173
  - HF comm 173
  - localizer 171
  - Loran C 174
  - marker beacon 172
  - MLS 174
  - Omega 174
  - Radiotelephone 175
  - SATCOM 174
  - TCAS I 175
  - VHF comm 166, 173
  - VOR 170
- ARINC 429 54
- ARINC 429 digital standards 94
- Artificial horizon 31
- Attitude Director Indicator (ADI) 33
- Audio control panels 114
- Audio Frequency (AF) 106
- Automatic Direction Finder (ADF) 123
- Automatic flight control system 177
- Autopilot maintenance 191
- Autopilot sensors 180
- Autopilots, types of 176
- Autosyn synchro 63
- Autothrottle 184
- Avionics master switch 149

B

- Bellows 13
- Bimetallic temperature system 65
- Binary Coded Decimal (BCD) 91
- Binary numbers 87
- BITE systems 53
- Bonding jumpers 4, 152
- Bourdon tube 13
- Bridle cable 181

C

- Caging knob 29
- Calibrated airspeed 17
- Capacitance quantity indicators 58
- Cavity magnetron 143
- CDI 126
- CMOS 94
- Coaxial cables 161
- Cockpit voice recorders 122
- Command bars 33
- Compass dip 42
- Compass error
  - acceleration 42
  - deviation 42
  - north turning 42
  - variation 42
- Compass, flux gate 44
- Computerized fuel system 63
- Computers
  - basic parts of 52
  - in aircraft 52
- Control Wheel Steering (CWS) 186
- Corona 157
- Counterpoise 111
- Critical engine 19
- Crystal 108
- Cylinder Head Temperature (CHT) gauge 67

D

- Deviation error 42
- Diaphragm 13
- Differential pressure 14
- Dip standards 94
- Directional gyro 29
- Distance Measuring Equipment (DME) 127
- Doppler navigation 135
- Drag cup 73, 74, 75
- Drip stick 56
- Dutch roll 188

E

- EADI 50
- EHSI 50
- EICAS 50
- Electromagnetic waves 103
  - wavelength of 103
- Electronic equipment, cleaning 149
- Electronic instruments 49
- Electronic tachometers 75
- Emergency Locator Transmitters (ELTs) 120
- Engine pressure ratio indicators 79
- Erector mechanism 32
- Exhaust gas temperature (EGT) gauge 67

F

- FAR 91.411 Altimeter system tests and inspections 25
- Filters 111
- Flight data recorders 122
- Flight director 187
- Flight management computers 183
- Flight management system (FMS) 177, 182
- Float-type mechanical gauge 55
- Flux gate compass 44
- Four-course radio range 123

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Frequency modulation (FM) 106
Frequency synthesizer 113
Fuel-injected engine flowmeters 58

G

Gauge
  float-type mechanical 55
  magnetic direct reading 55
  sight glass 55
Gauge pressure 14
Gimbals 29
Glass cockpit 49
Glipeslope 139
GPS (global positioning system) ... 137
Ground proximity warning
  system 142
Ground waves 107
Ground-loop interference 154
Groundplane 111
  considerations 164

H

Heads up displays 51
Heads up Guidance System (HGS). 51
Hertz dipole antenna 111
HF comm radio 117
Horizontal polarization 112
Horizontal Situation Indicator
  (HSI) 30

I

IFF 129
Inclinometer 34
Indicated airspeed 17
Inertial Navigation System (INS) ... 134
Inertial Reference Unit (IRU) 135
Instantaneous vertical speed
  indicator 22
Instrument and equipment
  requirements 12
Instrument categories 2
Instrument flight rules 12
Instrument Landing System (ILS) 137
Instrument lighting systems 5
Instruments
  instrument flight rules 12
  methods to install 3
  visual flight rules day 12
  visual flight rules night 12

Integrated circuits 93
Intercom systems 116
Interphone systems 116
Inverters 154
ITT 68

K

Kollsman window 16

L

Latitude 131
Line replaceable unit 53
Localizer 138
Logic gates 91
Long wire sense antenna 163
Longitude 131
Loop antenna 124
Loran C 132
LRU 53, 95

M

Mach number 21
Machmeter 21
Magnesyn synchro 63
Magnetic compass 41
Magnetic directing reading gauge... 55
Maintenance of gyro systems 40
Manifold pressure gauges 79
Marconi antenna 111
Marker beacons 139
Mass flowmeters 63
Master caution lights 99
Master warning lights 99
Maximum allowable airspeed
  indicator 21
Mechanical bulb temperature
  gauge 66
Mechanical tachometers 72
Microphones 114
Microwave landing system (MLS) . 141
Mode 3/A 130
Mode C 130
Mode S 131
Modulation rate 110
Modulators and demodulators 108

N

North turning error 42
Null field discharger 157

O

OBS 126
Octal 91
Oil pressure indicators 75
Omega 134
Oscillators 108

P

P-static 6, 155
Pitot tube 23
Polarization
  horizontal 112, 166
  vertical 112, 166
Position indicating systems 69
Positive and negative logic 93
Powerplants, types of 7
Precession 29
Pressure
  absolute 14
  differential 14
  gauge 14
Primary power setting instruments 84
Proximity sensor 71

Q

Quadrantal error 164

R

Radar altimeter 141
Radio
  HF comm 117
  VHF comm 117
Radio Frequency (RF) 106
Radio frequency chart 104
Radiotelephone 119
Ratiometer tempereature gauge 67
Required instruments
  flight and navigation 8
  powerplant 8
Rigidity in space 29
RLG 181
RMI (radio magnetic indicator) 45
RNAV 127
Runway alignment 189

S

SATCOMM 119
Satellite navigation 135

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

Selcal 119
Selsyn synchro 62
Sense antenna 124
Sensors, autopilot 180
Servos, autopilot 181
Shielding...................................... 154
Shock mounts.................................. 4
Sight glass gauge 55
Silicon.. 93
Sky waves 107
Slip-skid indicator.......................... 34
Space waves................................. 107
Speakers 114
Speed of sound 21
St. Elmo's Fire 157
Stability Augmentation System
    (SAS) 188
Stall warning systems 98
Standard "T" configuration............. 11
Standard Atmosphere table............ 15
Standard rate turn......................... 35
Standard sizes for round
    instruments 2
Standing wave ratio 160
Static discharagers 155
Static loads 158
Static ports 23
Static wicks 157
Stick shaker................................... 98
Stormscope 145
Superheterodyne 113
Swinging the compass.................... 42
Synchro systems 61

T

Tachometer generators................... 75
Takeoff warning systems................ 95
TCAS.. 145
Temperature measuring systems.... 65
    bimetallic 65
    mechanical bulb......................... 66
    ratiometer 67
    thermocouples 67
    Wheatstone bridge...................... 66
Thermocouples 67
Thrust Management Computer
    (TMC) 184
TIT 68
TO-FROM indicator 126
Torquemeters 78
TOT.. 68
Total Air Temperature (TAT) 28
Transponders................................ 128
Trip free circuit breakers.............. 152
True airspeed 17
TTL 94
Turn and bank 34
Turn coordinator 36

V

Vacuum pump 38
Vane-type flowmeters 60
Variation error................................ 42
Variometer 22
Venturi for gyros 37
Vertical card compass 43
Vertical polarization 112
Vertical speed indicator 21
Very high frequency omnirange
    (VOR)....................................... 125
VHF comm radio 117
Vibration indicators........................ 87
Visual flight rules day 12
Visual flight rules night 12
Vmc 19
VSWR... 160

W

Waveguide 143
Weather radar 143
Wet pump for air driven gyros 38
Wheatstone bridge temperature
    gauge 66
Wire antennas 162

Aircraft Technical Books, LLC
(970) 726-5111
http://www.ACTechBooks.com

ABOUT THE BOOK

Aircraft Instruments and Avionics is intended to be used in the instruction of students in an aviation maintenance technician training program.

This textbook includes:

Basic Instruments: Why study instruments, aircraft instrument requirements, pitot-static system instruments, gyro instruments, compass systems, electronic instruments, and computers in aircraft

Powerplant Instruments and Logic Gates: Liquid quantity measuring systems, fuel flow indicators, temperature measuring systems, position indicating systems, tachometers, oil pressure indicators, torquemeters, engine pressure ratio indicators, manifold pressure gauges, primary power setting instruments, vibration indicators, logic circuits and digital systems, takeoff warning systems, angle of attack indicators, stall warning systems, and annunciators

Communication and Navigation Systems: Radio fundamentals, regulations and standards for radios, intercom and interphone systems, communications radios, emergency locator transmitters (ELTs), cockpit voice recorders, flight data recorders, navigation systems, instrument landing system (ILS), radar altimeter, ground proximity warning system (GPWS), weather radar, Stormscope®, and airborne collision avoidance system (TCAS)

Aircraft Antennas and Autopilots: Installation and inspection of avionics, antenna installations, autopilots, and flight directors

ABOUT THE AUTHOR

Max Henderson taught Aviation Maintenance Technology subjects at Embry-Riddle Aeronautical University. Prior to that he was an Electronics Technician in the U.S. Air Force. He holds a commercial pilot as well as a mechanic certificate with an airframe and powerplant rating. He has also worked as a control tower operator.

JS312666-001

Jeppesen Sanderson Inc.
www.jeppesen.com
55 Inverness Drive East
Englewood, Colorado 80112-5498

ISBN 0-89100-422-X
9 780891 004226
90000

ELECTRONIC WARFARE FUNDAMENTALS

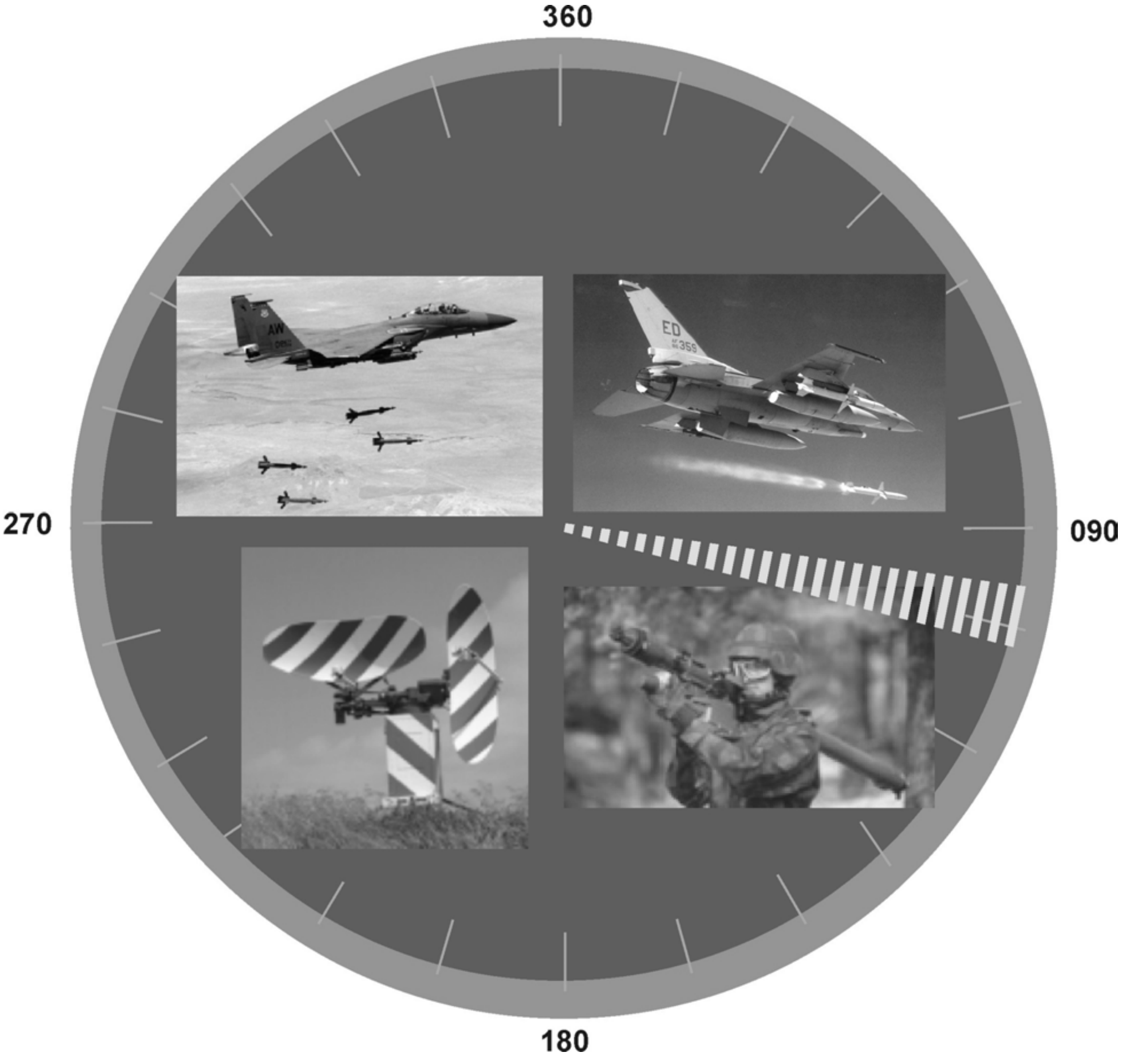
360
270
090
180

NOVEMBER 2000

PREFACE

Electronic Warfare Fundamentals **is a student supplementary text and reference book that provides the foundation for understanding the basic concepts underlying electronic warfare (EW). This text uses a practical building-block approach to facilitate student comprehension of the essential subject matter associated with the combat applications of EW.**

Since radar and infrared (IR) weapons systems present the greatest threat to air operations on today's battlefield, this text emphasizes radar and IR theory and countermeasures. Although command and control (C^2) systems play a vital role in modern warfare, these systems are not a direct threat to the aircrew and hence are not discussed in this book. This text does address the specific types of radar systems most likely to be associated with a modern integrated air defense system (IADS).

To introduce the reader to EW, ***Electronic Warfare Fundamentals*** **begins with a brief history of radar, an overview of radar capabilities, and a brief introduction to the threat systems associated with a typical IADS. The two subsequent chapters introduce the theory and characteristics of radio frequency (RF) energy as it relates to radar operations. These are followed by radar signal characteristics, radar system components, and radar target discrimination capabilities. The book continues with a discussion of antenna types and scans, target tracking, and missile guidance techniques.**

The next step in the building-block approach is a detailed description of countermeasures designed to defeat radar systems. The text presents the theory and employment considerations for both noise and deception jamming techniques and their impact on radar systems. This is followed by a chapter on decoys, both for defeating an IADS as well as for self-protection. Then, the next chapter discusses chaff characteristics, employment, and impact on specific radar systems.

The following two chapters are dedicated to the IR threat. The first covers IR theory, IR target detection and tracking, and advanced IR missile flare rejection techniques. The second chapter presents IR countermeasures to include flare employment, maneuvers, and missile warning equipment.

Electronic Warfare Fundamentals **then addresses an important aspect of EW, specifically electronic protection (EP). This section includes a description of the most common radar EP techniques designed to counter noise jamming, deception jamming, and chaff employment. The book concludes with an overview of the basic components and limitations of a typical radar warning receiver (RWR), current geolocation techniques, and it finally discusses the basic components of a self-protection jamming system.**

In addition to the textual material, the book also contains a detailed glossary of EW-related terms and acronyms for quick reference. A list of references is provided to acknowledge the source material used in the preparation of this text. These sources also provide the interested student with a supplementary reading list to learn more about a specific topic.

For information or comments on this document, please contact:

Det 8, ACC TRSS
4349 Duffer Drive, Ste 437
Nellis AFB NV 89191-7007

DSN: 682-4897
DSN FAX: 682-7391
Comm: (702) 652-4897
Comm FAX: (702) 652-7391

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION TO RADAR 1-1
1. INTRODUCTION 1-1
2. HISTORY 1-1
3. TARGET DISCRIMINANTS 1-2
4. INTRODUCTION TO INTEGRATED AIR DEFENSE SYSTEMS (IADS)........ 1-6
5. SUMMARY........ 1-10

CHAPTER 2. CHARACTERISTICS OF RF RADIATION........ 2-1
1. INTRODUCTION........ 2-1
2. FREQUENCY........ 2-1
3. WAVELENGTH........ 2-2
4. POLARIZATION 2-2
5. DOPPLER EFFECT........ 2-5
6. ELECTROMAGNETIC SPECTRUM 2-7
7. RF PROPAGATION........ 2-9
8. SUMMARY........ 2-14

CHAPTER 3. RADAR SIGNAL CHARACTERISTICS........ 3-1
1. INTRODUCTION........ 3-1
2. PULSE WIDTH (PW) 3-1
3. PULSE RECURRENCE TIME (PRT)........ 3-1
4. PULSE REPETITION FREQUENCY (PRF)........ 3-2
5. RADAR RECEIVER CHARACTERISTICS........ 3-3
6. DUTY CYCLE 3-4
7. PEAK POWER........ 3-4
8. AVERAGE POWER........ 3-5
9. MODULATION........ 3-6
10. SUMMARY........ 3-8

CHAPTER 4. RADAR SYSTEM COMPONENTS 4-1
1. INTRODUCTION........ 4-1
2. PULSE RADAR SYSTEM........ 4-1
3. CONTINUOUS WAVE (CW) RADAR 4-7
4. PULSE DOPPLER RADAR 4-9
5. MONOPULSE RADAR 4-10
6. SUMMARY........ 4-13

CHAPTER 5. RADAR PRINCIPLES 5-1

1. INTRODUCTION 5-1
2. RADAR RANGE 5-1
3. AZIMUTH DETERMINATION 5-7
4. ELEVATION DETERMINATION 5-11
5. RADAR RESOLUTION CELL 5-12
6. PULSE DOPPLER VELOCITY DETERMINATION 5-14
7. BASIC RADAR EQUATION 5-18
8. SUMMARY 5-21

CHAPTER 6. ANTENNA CHARACTERISTICS AND SCANS 6-1

1. INTRODUCTION 6-1
2. PARABOLIC ANTENNA 6-1
3. CASSEGRAIN ANTENNA 6-3
4. PHASED ARRAY ANTENNA 6-5
5. ANTENNA GAIN 6-9
6. POWER DENSITY 6-11
7. CIRCULAR SCAN 6-12
8. LINEAR SCAN 6-16
9. UNIDIRECTIONAL SCAN 6-16
10. BIDIRECTIONAL SCAN 6-17
11. CONICAL SCAN 6-18
12. TRACK-WHILE-SCAN 6-20
13. IMPACT OF TERRAIN ON RADAR SCANS 6-23
14. SUMMARY 6-26

CHAPTER 7. TARGET TRACKING 7-1

1. INTRODUCTION 7-1
2. RANGE TRACKING 7-2
3. CONICAL SCAN 7-4
4. TRACK-WHILE-SCAN (TWS) 7-6
5. LOBE-ON-RECEIVE-ONLY (LORO) 7-11
6. MONOPULSE RADAR 7-12
7. CONTINUOUS WAVE (CW) RADARS 7-16
8. PULSE DOPPLER RADAR 7-20
9. SUMMARY 7-24

CHAPTER 8. RADAR MISSILE GUIDANCE TECHNIQUES 8-1

1. INTRODUCTION 8-1
2. COMMAND GUIDANCE 8-3
3. SEMI-ACTIVE GUIDANCE 8-6
4. ACTIVE GUIDANCE 8-9
5. SEEKER-AIDED GROUND GUIDANCE/TRACK-VIA-MISSILE GUIDANCE 8-11

6. ANTIAIRCRAFT ARTILLERY (AAA) 8-12
7. SUMMARY 8-15

CHAPTER 9. INTRODUCTION TO RADAR JAMMING 9-1
1. INTRODUCTION 9-1
2. RADAR JAMMING TYPES 9-1
3. RADAR JAMMING EMPLOYMENT OPTIONS 9-3
4. FUNDAMENTALS OF RADAR JAMMING 9-7
5. SUMMARY 9-15

CHAPTER 10. RADAR NOISE JAMMING 10-1
1. INTRODUCTION 10-1
2. RADAR NOISE JAMMING EFFECTIVENESS 10-1
3. RADAR NOISE JAMMING GENERATION 10-4
4. BARRAGE JAMMING 10-5
5. SPOT JAMMING 10-6
6. SWEPT-SPOT JAMMING 10-7
7. COVER PULSE JAMMING 10-8
8. MODULATED NOISE JAMMING 10-10
9. SUMMARY 10-11

CHAPTER 11. DECEPTION JAMMING 11-1
1. INTRODUCTION 11-1
2. FALSE TARGET JAMMING 11-3
3. RANGE DECEPTION JAMMING 11-5
4. ANGLE DECEPTION JAMMING 11-8
5. VELOCITY DECEPTION JAMMING 11-11
6. MONOPULSE DECEPTION JAMMING 11-17
7. TERRAIN BOUNCE 11-23
8. SUMMARY 11-24

CHAPTER 12. DECOYS 12-1
1. INTRODUCTION 12-1
2. SATURATION DECOYS 12-1
3. TOWED DECOYS 12-3
4. EXPENDABLE ACTIVE DECOYS 12-6
5. SUMMARY 12-7

CHAPTER 13. CHAFF EMPLOYMENT 13-1
1. INTRODUCTION 13-1
2. CHAFF CHARACTERISTICS 13-2
3. CHAFF OPERATIONAL EMPLOYMENT 13-9
4. SUMMARY 13-16

CHAPTER 14. IR FUNDAMENTALS 14-1
1. INTRODUCTION 14-1
2. BASIC IR THEORY 14-1
3. IR SIGNATURE SOURCES 14-5
4. IR SEEKER CHARACTERISTICS 14-7
5. IR SEEKER TYPES 14-9
6. IR MISSILE FLARE REJECTION 14-12
7. SUMMARY 14-18

CHAPTER 15. IR COUNTERMEASURES 15-1
1. INTRODUCTION 15-1
2. FLARE CHARACTERISTICS 15-1
3. ADVANCED FLARES 15-3
4. FLARE EMPLOYMENT 15-4
5. IRCM TACTICS 15-5
6. MISSILE APPROACH WARNING SYSTEMS (MAWS) 15-6
7. SUMMARY 15-8

CHAPTER 16. RADAR ELECTRONIC PROTECTION (EP) TECHNIQUES 16-1
1. INTRODUCTION 16-1
2. RADAR RECEIVER PROTECTION 16-1
3. JAMMING SIGNAL AVOIDANCE 16-4
4. JAMMING SIGNAL EXPLOITATION 16-5
5. OVERPOWERING THE JAMMING SIGNAL 16-6
6. PULSE DURATION DISCRIMINATION 16-7
7. ANGLE DISCRIMINATION 16-7
8. BANDWIDTH DISCRIMINATION 16-8
9. DOPPLER DISCRIMINATION 16-9
10. TIME DISCRIMINATION 16-9
11. SUMMARY 16-10

CHAPTER 17. RADAR WARNING RECEIVER (RWR) BASIC OPERATIONS AND GEOLOCATION TECHNIQUES 17-1
1. INTRODUCTION 17-1
2. RWR ANTENNAS 17-1
3. RWR RECEIVER/AMPLIFIERS 17-3
4. SIGNAL PROCESSOR 17-6
5. EMITTER IDENTIFICATION (EID) TABLES 17-8
6. RWR SCOPE DISPLAY 17-8
7. RWR AUDIO 17-9
8. RWR INTERFACE CONTROL UNIT (ICU) 17-10
9. RWR LIMITATIONS 17-10
10. THREAT GEOLOCATION TECHNIQUES 17-15
11. SUMMARY 17-20

CHAPTER 18. SELF-PROTECTION JAMMING SYSTEM OPERATIONS.............. 18-1
1. INTRODUCTION.. 18-1
2. RECEIVE ANTENNAS... 18-2
3. RECEIVER SECTION.. 18-2
4. SYSTEM PROCESSOR.. 18-3
5. JAMMING TECHNIQUES GENERATOR... 18-4
6. TRANSMIT ANTENNAS.. 18-4
7. C-9492 CONTROL INDICATOR UNIT.. 18-5
8. JAMMING POD CONSIDERATIONS... 18-5
9. SUMMARY... 18-7

ANNEX A. GLOSSARY...A-1

ANNEX B. EC-RELATED ABBREVIATIONS AND ACRONYMS............................B-1

LIST OF REFERENCES

LIST OF FIGURES

Figure 1-1. Radar System....................1-1
Figure 1-2. Early Continuous Wave Radar Experiments....................1-1
Figure 1-3. Pulse Radar Operation....................1-2
Figure 1-4. Radar Advantages and Disadvantages....................1-3
Figure 1-5. Radar Target Discriminants....................1-3
Figure 1-6. Angle Discrimination....................1-4
Figure 1-7. Velocity Discrimination....................1-5
Figure 1-8. Basic Radar Operation....................1-5
Figure 1-9. Early Warning Radar System....................1-6
Figure 1-10. GCI Radar....................1-7
Figure 1-11. Acquisition Radar and TTR....................1-8
Figure 1-12. TTR with Acquisition....................1-8
Figure 1-13. MiG-29....................1-9
Figure 1-14. IR System....................1-9
Figure 1-15. A Typical IADS....................1-10
Figure 2-1. Radio Frequency....................2-1
Figure 2-2. Radar Signal Wavelength....................2-2
Figure 2-3. Radar Signal Vertical Polarization....................2-3
Figure 2-4. Circular Polarization....................2-4
Figure 2-5. Impact of Polarization....................2-4
Figure 2-6. Zero Doppler Effect – Stationary Target....................2-6
Figure 2-7. Doppler Effect – Closing Target....................2-6
Figure 2-8. Doppler Effect – Opening Target....................2-6
Figure 2-9. Electromagnetic Spectrum....................2-7
Figure 2-10. Impact of Refraction on RF Propagation....................2-10
Figure 2-11. Impact of Refraction on Target Elevation Determination....................2-10
Figure 2-12. Superrefracting Surface Duct....................2-12
Figure 2-13. RF Atmospheric Attenuation....................2-14
Figure 3-1. Typical Radar Pulse....................3-1
Figure 3-2. Constant PRF Radar Pulse....................3-2
Figure 3-3. PRF Stagger....................3-2
Figure 3-4. PRF Jitter....................3-3
Figure 3-5. Basic Radar Pulse....................3-3
Figure 3-6. Peak Power and Average Power....................3-5
Figure 3-7. Radar Power and Energy....................3-6
Figure 3-8. Amplitude Modulation....................3-6
Figure 3-9. Frequency Modulation....................3-7
Figure 3-10. Pulse Modulation....................3-7
Figure 4-1. Pulse Radar Block Diagram....................4-1
Figure 4-2. Transmitter Group....................4-2
Figure 4-3. Receiver Group....................4-3
Figure 4-4. Master Timer/Synchronizer....................4-4
Figure 4-5. Parabolic Antenna....................4-4

Figure 4-6. A Scope Display.. 4-5
Figure 4-7. B Scope Display.. 4-5
Figure 4-8. RHI Scope Display 4-6
Figure 4-9. PPI Scope Display.. 4-6
Figure 4-10. Continuous Wave Radar 4-7
Figure 4-11. Basic CW Doppler Radar.. 4-8
Figure 4-12. CW Doppler Radar Display 4-8
Figure 4-13. CW and Pulse Doppler Radar Comparison 4-9
Figure 4-14. Basic Pulse Doppler Radar Diagram.. 4-10
Figure 4-15. Monopulse Radar.. 4-11
Figure 4-16. Monopulse Magic T's.. 4-11
Figure 4-17. Magic T Output Signals 4-12
Figure 4-18. Monopulse Radar Track 4-12
Figure 4-19. Monopulse Radar Track Logic.. 4-13
Figure 5-1. Basic Radar Pulse 5-1
Figure 5-2. Radar Mile.. 5-2
Figure 5-3. Second Time Around Echo 5-3
Figure 5-4. Radar Pulse 5-4
Figure 5-5. Radar Pulse at T1.. 5-5
Figure 5-6. Radar Pulse at T2.. 5-5
Figure 5-7. Radar Pulse at T3.. 5-6
Figure 5-8. Radar Range Resolution 5-6
Figure 5-9. Radar Beamwidth 5-7
Figure 5-10. Azimuth Determination.. 5-8
Figure 5-11. Antenna Scan.. 5-8
Figure 5-12. Horizontal Beamwidth Comparison 5-9
Figure 5-13. Horizontal Beamwidth and Azimuth Resolution 5-10
Figure 5-14. Azimuth Resolution 5-10
Figure 5-15. Vertical Beamwidth and Elevation Resolution 5-11
Figure 5-16. Elevation Resolution 5-12
Figure 5-17. Radar Resolution Cell.. 5-13
Figure 5-18. Radar Resolution Cell Dimensions 5-13
Figure 5-19. Pulse Modulation 5-14
Figure 5-20. Harmonics of a Pulse Modulated Waveform 5-15
Figure 5-21. Harmonic Content.. 5-15
Figure 5-22. Spectral Line Frequencies 5-16
Figure 5-23. Selective Clutter Canceling.. 5-16
Figure 5-24. PRF and Spectral Lines.. 5-17
Figure 5-25. Pulse Doppler Filters.. 5-18
Figure 6-1. Parabolic Antenna 6-1
Figure 6-2. Parabolic Cylinder Antenna 6-2
Figure 6-3. Height-Finder Parabolic Antenna 6-3
Figure 6-4. Multiple-Feed Parabolic Antenna 6-3
Figure 6-5. Cassegrain Antenna 6-4
Figure 6-6. Flat Plate Cassegrain Antenna 6-4
Figure 6-7. Phased Array Antenna 6-5

Figure 6-8. Phase Relationships 6-6
Figure 6-9. Broadside Array......6-7
Figure 6-10. End-Fire Array......6-7
Figure 6-11. Phased Array Antenna Scan......6-8
Figure 6-12. Planar Array Antenna......6-9
Figure 6-13. Radiation Pattern for a Parabolic Antenna......6-11
Figure 6-14. Radar Antenna Pattern......6-12
Figure 6-15. Circular Scan Radar 6-13
Figure 6-16. PPI Scope Display 6-14
Figure 6-17. Resolution Cell 6-14
Figure 6-18. V-Beam Radar......6-15
Figure 6-19. Stacked Beam Radar......6-15
Figure 6-20. Unidirectional Linear Scan 6-16
Figure 6-21. Helical Scan 6-17
Figure 6-22. Raster Scan......6-17
Figure 6-23. Conical Scan......6-18
Figure 6-24. Conical Scan Tracking Area 6-18
Figure 6-25. Palmer-Raster Scan......6-19
Figure 6-26. Palmer-Helical Scan 6-20
Figure 6-27. Conventional TWS Radar......6-21
Figure 6-28. Phased Array Antenna 6-21
Figure 6-29. Phased Array TWS 6-22
Figure 6-30. Phased Array TWS Radar Display......6-23
Figure 6-31. The Radar Horizon......6-23
Figure 6-32. Direct Terrain Masking......6-25
Figure 6-33. Indirect Terrain Masking 6-26
Figure 7-1. Split-Gate Range Tracker......7-2
Figure 7-2. Leading-Edge Range Tracker 7-3
Figure 7-3. Conical Scan Positions......7-4
Figure 7-4. Conical Scan Modulation......7-5
Figure 7-5. Conical Scan Tracking Errors 7-5
Figure 7-6. Conical Scan Tracking......7-6
Figure 7-7. Conventional TWS Radar Beams 7-7
Figure 7-8. Conventional TWS Azimuth Tracking......7-8
Figure 7-9. Phased Array TWS 7-9
Figure 7-10. Air-to-Air TWS Display 7-10
Figure 7-11. Planar/Phased Array TWS Tracking Loop......7-10
Figure 7-12. LORO Mode......7-11
Figure 7-13. Monopulse Magic T 7-13
Figure 7-14. Magic T Output Signals......7-13
Figure 7-15. Monopulse Tracking Loops......7-14
Figure 7-16. Monopulse Elevation Tracking Error 7-15
Figure 7-17. Monopulse Elevation Track 7-15
Figure 7-18. Monopulse Azimuth Track......7-16
Figure 7-19. CW Radar 7-16
Figure 7-20. FM CW Radar Range Determination (No Relative Motion)......7-18

Figure 7-21. FM CW Radar Range Determination (Relative Motion) 7-18
Figure 7-22. Comparison of CW Radar and Pulse Doppler Radar 7-20
Figure 7-23. Range Ambiguity 7-21
Figure 7-24. Pulse Doppler Speed Gate 7-22
Figure 7-25. Range Ambiguity 7-23
Figure 7-26. Range Eclipsing 7-23
Figure 7-27. Resolving Range Ambiguity and Eclipsing 7-24
Figure 8-1. Initial Boost Phase 8-1
Figure 8-2. Mid-Course Guidance Phase 8-2
Figure 8-3. Terminal Guidance Phase 8-3
Figure 8-4. Rectified Flight Profile 8-4
Figure 8-5. Command Guidance: Three-Point Pursuit 8-5
Figure 8-6. Semi-Active Guided Missile 8-6
Figure 8-7. Semi-Active Guidance (Mid-Course) 8-7
Figure 8-8. Semi-Active Guidance (Terminal Phase) 8-8
Figure 8-9. Semi-Active Guidance CW Reference Signal 8-9
Figure 8-10. Active Guidance Mid-Course Intercept Phase 8-9
Figure 8-11. Active Guidance Phase 8-10
Figure 8-12. SAGG/Track-Via-Missile Guidance 8-11
Figure 8-13. Optically Aimed AAA 8-13
Figure 8-14. Radar-Directed AAA 8-14
Figure 8-15. Sector/Barrage AAA 8-15
Figure 9-1. Noise Jamming 9-1
Figure 9-2. Deception Jamming 9-2
Figure 9-3. Stand-Off Jamming 9-3
Figure 9-4. False Target Jamming 9-4
Figure 9-5. Escort Jamming 9-5
Figure 9-6. Escort Jamming Alignment 9-5
Figure 9-7. Self-Protection Jamming 9-6
Figure 9-8. Jamming Frequency Error 9-7
Figure 9-9. Correct Jamming Frequency Matching 9-8
Figure 9-10. S/N Ratio and Target Detection 9-11
Figure 9-11. Receiver False Alarm Threshold 9-11
Figure 9-12. S/N Ratio Greater Than One 9-12
Figure 9-13. S/N Ratio Less Than One 9-12
Figure 9-14. J/S Ratio Greater Than One 9-14
Figure 9-15. J/S Ratio Less Than One 9-14
Figure 9-16. Burnthrough Range 9-15
Figure 10-1. Noise Level in a Typical Radar Receiver Output 10-1
Figure 10-2. Impact of Noise Jamming 10-2
Figure 10-3. Power Density – Narrow Bandwidth 10-2
Figure 10-4. Power Density – Wide Bandwidth 10-3
Figure 10-5. Frequency Modulated (FM) Noise Jamming System 10-5
Figure 10-6. Barrage Jamming 10-6
Figure 10-7. Spot Jamming 10-6
Figure 10-8. Spot Jamming Effectiveness 10-7

Figure 10-9. Swept-Spot Jamming 10-8
Figure 10-10. Cover Pulse Jamming 10-8
Figure 10-11. Range Gate Pull-Off Cover Pulse 10-9
Figure 10-12. Velocity Gate Pull-Off Cover Pulse 10-10
Figure 10-13. Conical Scan Modulated Jamming 10-10
Figure 10-14. TWS Modulated Jamming 10-11
Figure 11-1. Deception Jamming System 11-2
Figure 11-2. False Target Generation 11-3
Figure 11-3. Range Gate Tracking 11-5
Figure 11-4. Range Gate Jamming Cover Pulse 11-6
Figure 11-5. Range Gate Pull-Off 11-7
Figure 11-6. Inverse Amplitude Modulation Jamming 11-8
Figure 11-7. Inverse Gain Jamming 11-9
Figure 11-8. Swept Square Wave Jamming 11-10
Figure 11-9. Velocity Tracking Gate 11-11
Figure 11-10. Velocity Gate Capture 11-12
Figure 11-11. Velocity Gate Pull-Off 11-13
Figure 11-12. Doppler Noise Jamming 11-14
Figure 11-13. Impact of Doppler Noise Jamming 11-14
Figure 11-14. Narrowband Doppler Noise 11-15
Figure 11-15. Velocity Bin Masking 11-16
Figure 11-16. Impact of Doppler False Target Jamming 11-17
Figure 11-17. Monopulse Radar Receiver 11-18
Figure 11-18. Filter Skirt Jamming Pulse 11-19
Figure 11-19. Ineffective Filter Skirt Jamming 11-19
Figure 11-20. Monopulse Image Frequency 11-20
Figure 11-21. Monopulse Image Jamming 11-21
Figure 11-22. Cross-Polarization Antenna Pattern 11-21
Figure 11-23. Cross-Eye Jamming 11-23
Figure 11-24. Terrain Bounce 11-24
Figure 12-1. Tactical Air Launched Decoy 12-2
Figure 12-2. Miniature Air Launched Decoy (MALD) 12-3
Figure 12-3. Towed Decoy 12-4
Figure 12-4. AN/ALE-50 System 12-5
Figure 12-5. Generic Expendable 12-6
Figure 13-1. “WINDOW” – The First Operational Employment of Chaff 13-1
Figure 13-2. Aircraft Radar Cross Section (RCS) 13-2
Figure 13-3. RR-170 Chaff Cartridge RCS 13-3
Figure 13-4. Threat Radar Aspect and Chaff RCS 13-4
Figure 13-5. Impact of Multiple Chaff Cartridge Employment on Chaff RCS 13-4
Figure 13-6. Chaff Length and Frequency Coverage 13-5
Figure 13-7. RR-170 Chaff Cartridge 13-6
Figure 13-8. Impact of Turbulence and Chaff Bloom Rate 13-7
Figure 13-9. Chaff Bloom Rate and Radar Resolution Cell 13-7
Figure 13-10. Impact of Chaff Polarization 13-9
Figure 13-11. Area Saturation Tactics 13-10

Figure 13-12. Chaff Corridor Tactics ... 13-11
Figure 13-13. Self-Protection Chaff Tactics ... 13-12
Figure 13-14. Self-Protection Chaff Effect on a TWS Radar ... 13-13
Figure 13-15. Self-Protection Chaff Impact on a Conical Scan Radar ... 13-14
Figure 13-16. Self-Protection Chaff Impact on a Monopulse Radar ... 13-15
Figure 13-17. Self-Protection Chaff Impact on a Doppler Radar ... 13-16
Figure 14-1. The IR Threat ... 14-1
Figure 14-2. IR Frequency Band ... 14-2
Figure 14-3. Atmospheric IR Transmission ... 14-3
Figure 14-4. IR Missile Seeker ... 14-4
Figure 14-5. Aircraft IR Intensity ... 14-4
Figure 14-6. IR Signature Versus Aspect Angle ... 14-5
Figure 14-7. Target Signature Sources ... 14-6
Figure 14-8. IR Detection Materials ... 14-7
Figure 14-9. IR Missile Seeker ... 14-7
Figure 14-10. Spin Scan Reticle ... 14-8
Figure 14-11. Missile Field of View ... 14-9
Figure 14-12. IR Missile Reticle ... 14-10
Figure 14-13. Conical Scan IR Seeker ... 14-11
Figure 14-14. Impact of IR Missile Field of View ... 14-13
Figure 14-15. IR Missile – Narrow Field of View ... 14-13
Figure 14-16. Two-Color IR “Trigger” ... 14-15
Figure 14-17. Seeker Push-Ahead “Response” Technique ... 14-16
Figure 14-18. Seeker Push-Pull “Response” Technique ... 14-17
Figure 14-19. Sector Attenuation “Response” Technique ... 14-18
Figure 15-1. The MJU-7 Flare ... 15-1
Figure 15-2. Impact of Flare Rise Time ... 15-2
Figure 15-3. Kinematic Flare ... 15-3
Figure 15-4. Flare Visual Signature ... 15-4
Figure 15-5. Initial Maneuver and Flare Employment ... 15-5
Figure 15-6. IRCM ... 15-6
Figure 15-7. Missile Alert Warning System (MAWS) ... 15-7
Figure 17-1. Radar Warning Receiver (RWR) Antennas ... 17-2
Figure 17-2. RWR Antenna Sensitivity ... 17-2
Figure 17-3. RWR Antenna Polarization ... 17-3
Figure 17-4. Crystal Video RWR Receiver ... 17-4
Figure 17-5. Superheterodyne RWR Receiver ... 17-5
Figure 17-6. RWR System ... 17-6
Figure 17-7. Signal Processor Signal Priority ... 17-7
Figure 17-8. Sample EID Table ... 17-8
Figure 17-9. RWR Scope Azimuth Positioning ... 17-9
Figure 17-10. RWR Interface Control Unit ... 17-10
Figure 17-11. Signal Frequency Spectrum ... 17-11
Figure 17-12. RWR Threat Azimuth Position Determination ... 17-12
Figure 17-13. RWR Azimuth Error ... 17-13
Figure 17-14. RWR Multiple Threat Symbols ... 17-14

Figure 17-15. Triangulation 17-16
Figure 17-16. Interferometer Components 17-17
Figure 17-17. Interferometer Array 17-17
Figure 17-18. Time of Arrival Measurements 17-18
Figure 17-19. Time Difference of Arrival Measurements 17-19
Figure 18-1. ALQ-184 Pod 18-1
Figure 18-2. ALQ-184 Receive Antenna Location 18-2
Figure 18-3. ALQ-184 Simplified Block Diagram.. 18-3
Figure 18-4. ALQ-184 Transmit Antenna Location 18-4
Figure 18-5. C-9492 Control Indicator Unit 18-5
Figure 18-6. F-16 ECM CMS 18-6

LIST OF EQUATIONS

Equation 1-1. Basic Radar Range Equation 1-4
Equation 2-1. Basic Wavelength Equation 2-2
Equation 2-2. Index of Refraction 2-11
Equation 2-3. Refractivity of RF Waves (Normal Radar Frequencies)................ 2-11
Equation 3-1. Duty Cycle 3-4
Equation 3-2. Energy Per Pulse 3-5
Equation 5-1. Basic Radar Range Determination Equation 5-2
Equation 5-2. Radar Mile 5-2
Equation 5-3. Maximum Unambiguous Range 5-4
Equation 5-4. Range Resolution 5-7
Equation 5-5. Azimuth Resolution 5-11
Equation 5-6. Altitude Resolution... 5-12
Equation 5-7. Power Density From an Antenna 5-18
Equation 5-8. Power Density at the Target 5-19
Equation 5-9. Power Density at the Antenna 5-19
Equation 5-10. Signal Power Density 5-20
Equation 5-11. Basic Radar Equation 5-20
Equation 6-1. Directive Gain 6-9
Equation 6-2. Power Gain... 6-10
Equation 6-3. Power Density for an Isotropic Antenna 6-10
Equation 6-4. Power Density for a Practical Antenna.. 6-10
Equation 6-5. Power Gain of a Practical Antenna 6-11
Equation 6-6. Basic Radar Horizon Equation 6-24
Equation 6-7. Simplified Radar Horizon Equation.. 6-24
Equation 6-8. Target LOS Distance 6-24
Equation 9-1. Signal Power Density 9-9
Equation 9-2. Thermal Noise.. 9-9
Equation 9-3. Signal-to-Noise Ratio 9-10
Equation 9-4. Jamming-to-Signal Ratio 9-13

LIST OF TABLES

Table 2-1. Antenna Polarization Loss 2-5
Table 2-2. Commonly Used Prefixes for Scientific Notation 2-8
Table 2-3. Radar Frequency Band Designations 2-9
Table 7-1. Pulse Doppler Tracking Capabilities 7-21
Table 17-1. RWR Frequency Band Designators 17-3

CHAPTER 1. INTRODUCTION TO RADAR

1. INTRODUCTION

The word "RADAR" is an acronym for RAdio Detection And Ranging. As it was originally conceived, radio waves were used to detect the presence of a target and to determine its distance or range (Figure 1-1).

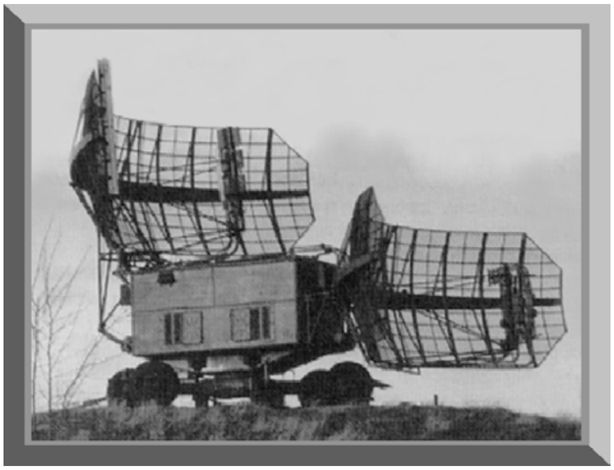

Figure 1-1. Radar System

2. HISTORY

The reflection of radio waves by objects was first noted more than a century ago. In 1903, the reflection of radio waves was employed in Germany to demonstrate detection of ships at sea (Figure 1-2). In 1922, Marconi presented the same idea in Britain but received little official interest. These early experiments used continuous wave, or CW, transmissions and relied on the reflection of a transmitted wave from a target to indicate the presence of a target. CW transmissions can detect the presence of an object and, if the radio wave is formed into a narrow beam, can also provide azimuth information. CW transmissions cannot provide range.

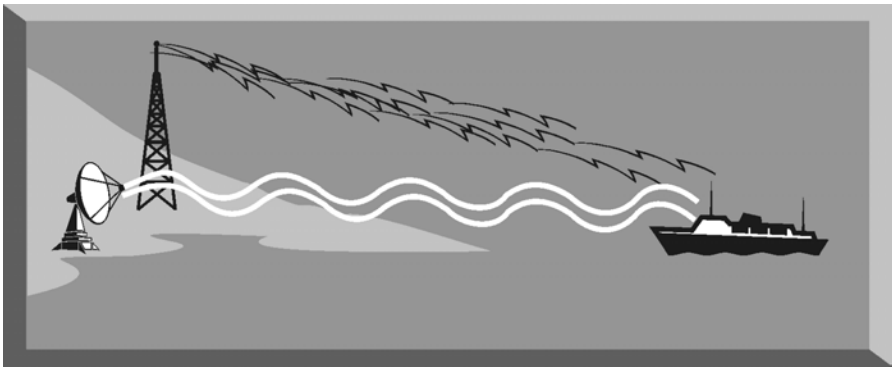

Figure 1-2. Early Continuous Wave Radar Experiments

a. The lack of range information was a serious limitation but was finally overcome by modulating the radio wave transmissions to send out a train of short pulses (Figure 1-3). The time between pulse transmission and an echo return to the receiver provides a direct measurement of range. Practical development of the pulse radar began in the 1930s, principally in the United States, Britain, and Germany. Due to deteriorating relations with Germany and the threat of invasion, the British intensified their efforts to develop a pulse radar in 1935. These efforts culminated in the development and deployment of multiple radar stations which formed the Chain Home System. These radar installations provided critical information to British pilots on the size and location of German bomber formations during the Battle of Britain.

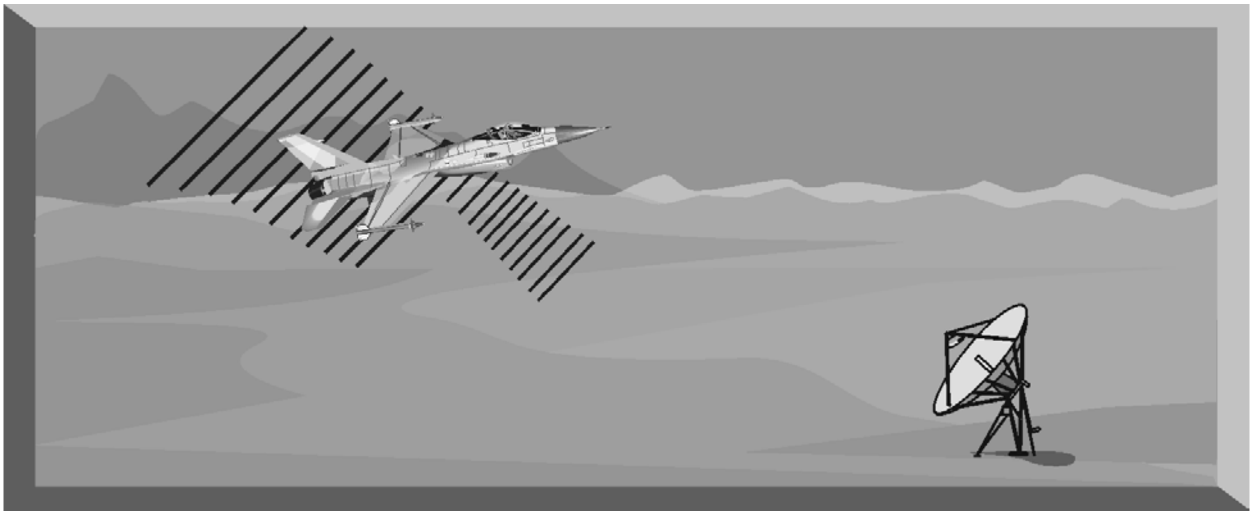

Figure 1-3. Pulse Radar Operation

b. The Chain Home System is considered the first integrated air defense system, or IADS. Radar development, with its obvious military and civilian applications, has continued unabated to the present day.

3. TARGET DISCRIMINANTS

The widespread military and civilian application of radar is based on its inherent advantages over the human eye (Figure 1-4). Radar can “see” farther than the human eye and more accurately assess the range or distance of an object. Radar works well in all-weather conditions and is relatively immune to smoke, haze, and clouds. What's more, radar works 24 hours a day because it can transmit its own energy and does not have to rely on sunshine or ambient radiation. There are some disadvantages of radar when compared to the human eye. First, radar does not have the resolution that the human eye has. While radar can detect the presence of an airplane, the human eye can discern, in great detail, the shape, size, color, and even markings. This can be a serious limitation if positive identification is required prior to engagement.

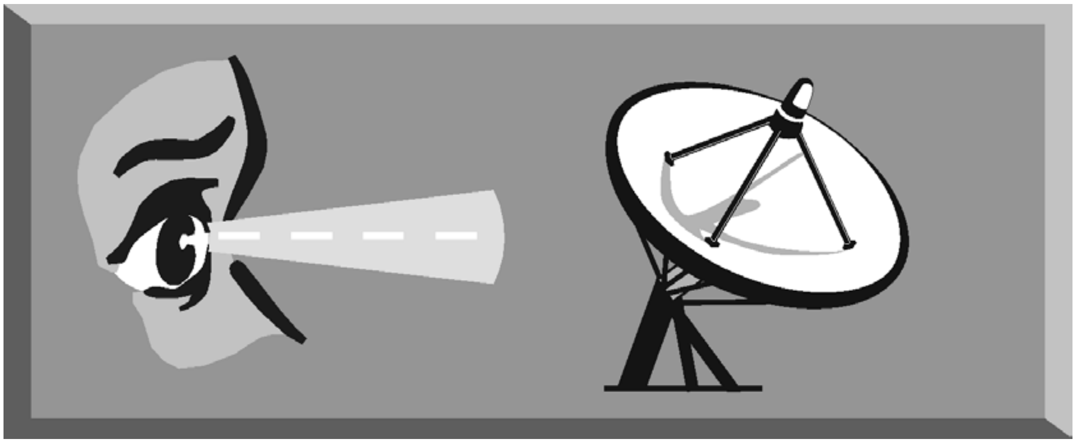

Figure 1-4. Radar Advantages and Disadvantages

Second, the human eye is not bothered by undesirable reflections, called clutter, the way radar sometimes is. Although metal is the best reflector of radio frequency (RF) energy, nearly any material will reflect some RF energy. Mountains, trees, buildings, rain, birds, and chaff all reflect RF energy. Radar systems must use target discriminants to isolate the desired target return from the clutter. These target discriminants include range, velocity, and angle (Figure 1-5).

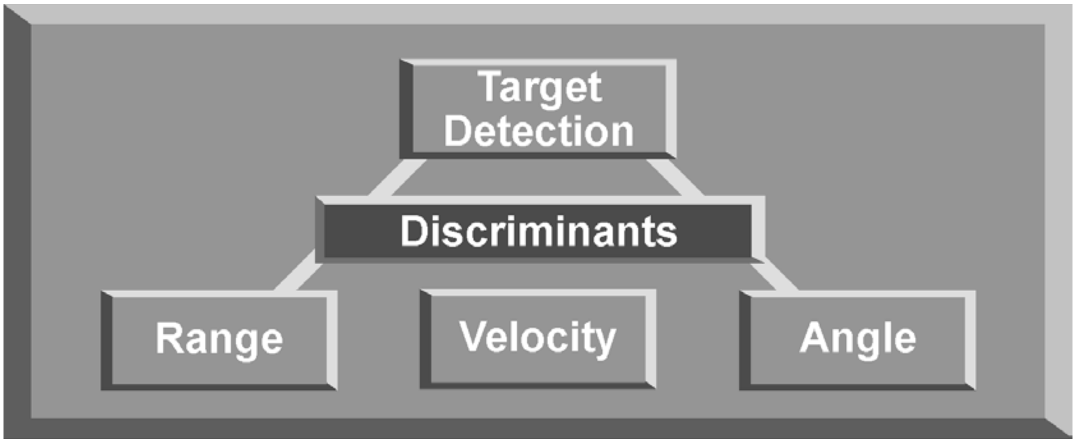

Figure 1-5. Radar Target Discriminants

a. The first target discriminant is range. The time an RF wave takes to go to, and return from, a target allows measurement of the range to that target. We know that RF energy travels at the speed of light, or “c” which is 3 x 10^8 meters per second. Target range can be determined by using the basic radar range equation (Equation 1-1), target range equals measured time, multiplied by the speed of light (“c”), divided by 2.

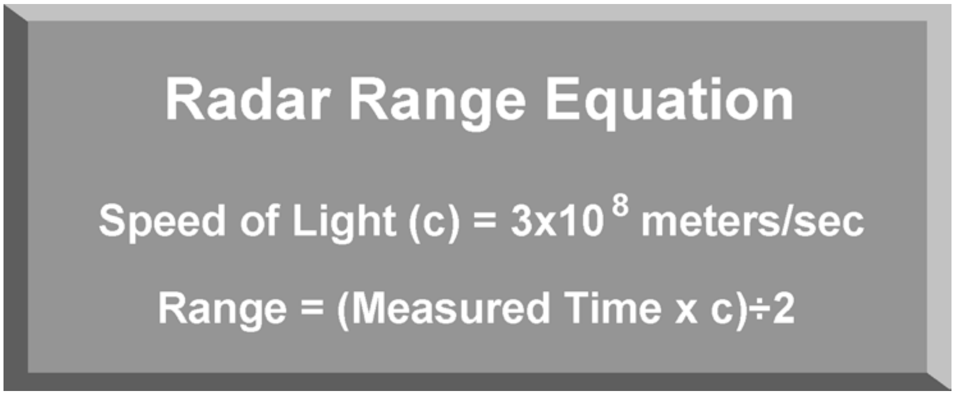

Radar Range Equation

Speed of Light (c) = 3×10^8 meters/sec

Range = (Measured Time x c)÷2

Equation 1-1. Basic Radar Range Equation

b. Target angle discrimination is another critical capability of radar systems. In order for a radar system to detect a target, the antenna must be pointed at the target during the transmission and reception of RF energy. The ability of a radar system to accurately determine angle is a function of the horizontal beamwidth of the antenna. If the radar sweep is referenced to true North, the angle of a radar return can be measured relative to true North (Figure 1-6).

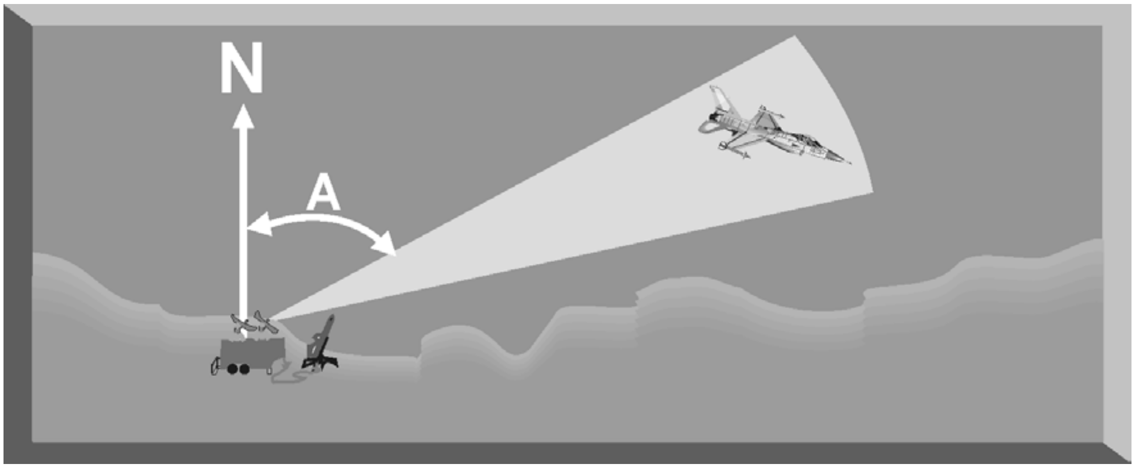

Figure 1-6. Angle Discrimination

c. Velocity discrimination is a specific capability of CW and pulse Doppler radar systems. The transmitters of CW radars send out continuous RF at a specific frequency (Figure 1-7). The reflected signal frequency is changed, or shifted, by a specific amount by a moving target. This frequency shift, called the Doppler effect, allows the measurement of the velocity of that target relative to the radar. The receiver measures this frequency difference which equates to a specific radial velocity. Pulse Doppler radars can measure the Doppler effect while still obtaining the range.

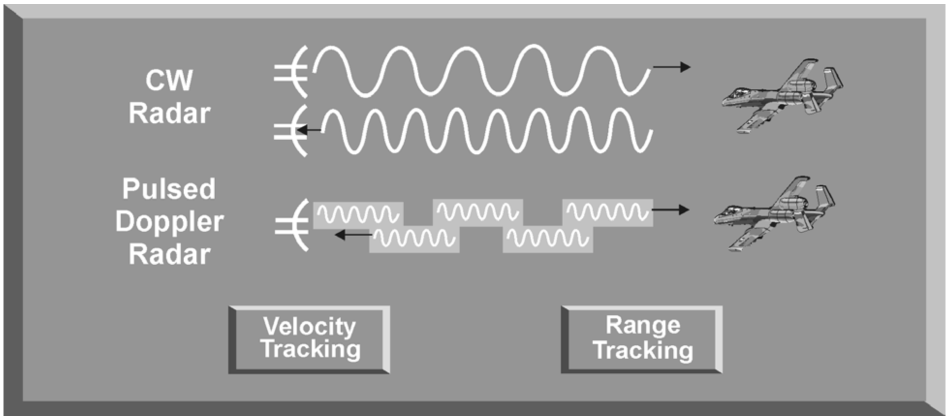
CW
Radar
Pulsed
Doppler
Radar
Velocity
Tracking
Range
Tracking

Figure 1-7. Velocity Discrimination

d. A basic pulse radar system consists of a transmitter, antenna, receiver, and a master timer (Figure 1-8). The transmitter sends electromagnetic energy (RF) to the antenna. This energy is radiated through the atmosphere. When this RF energy is interrupted by any object, such as a plane, ship, or the earth, a portion of the RF energy is reflected back to the antenna and processed by the receiver.

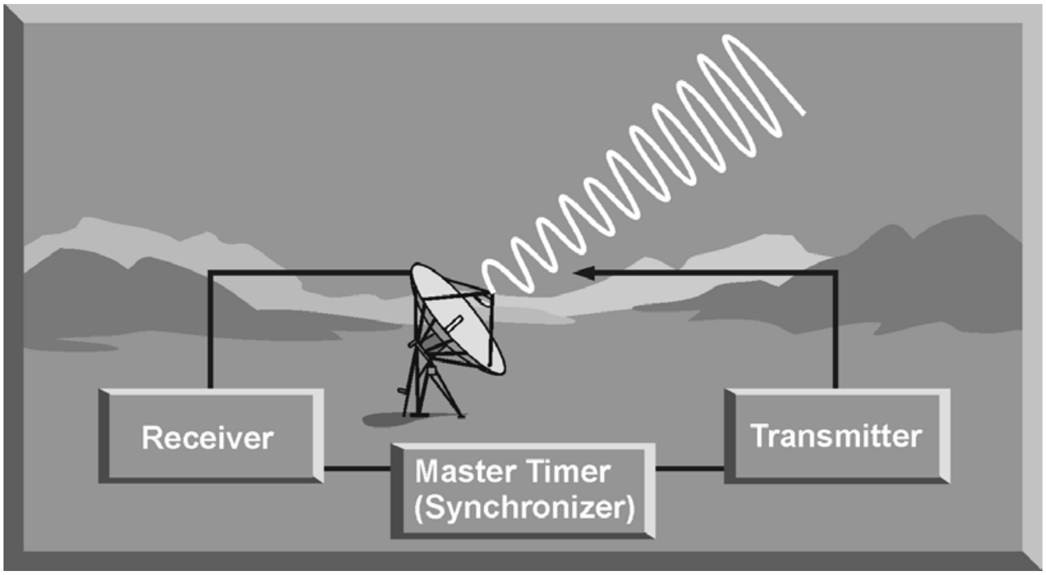
Receiver
Master Timer
(Synchronizer)
Transmitter

Figure 1-8. Basic Radar Operation

The reflection is called an echo, and the object interrupting the RF energy is called a target. The presence of an echo indicates target detection. If the target detected is the desired target, the echo is referred to as a target signal. If the echo is from undesired targets, such as the earth, the echo is referred to as clutter. This is especially true when the undesirable echoes make the detection of the desired target difficult. The capability of the antenna to focus the RF energy

affects the angular discrimination of the radar. The ability of the master timer to determine the time between RF transmission and target echo reception impacts the range determination capability of the radar. The ability of the receiver to analyze the Doppler frequency shift in the target echo determines the velocity discrimination capability of the radar and the ability of the radar to reject clutter.

4. INTRODUCTION TO INTEGRATED AIR DEFENSE SYSTEMS (IADS)

Radar systems have the inherent capability to determine accurate range, azimuth, and/or velocity information on airborne targets. Radar systems can provide this information in nearly all types of weather, day or night, and at distances that far exceed the capabilities of the human eye. Military commanders have taken advantage of these capabilities by employing radar systems to provide air defense for high-value targets. The primary missions of radar systems employed for air defense are attack warning and threat engagement.

a. Radar systems specifically designed to provide attack warning are called early warning (EW) radars (Figure 1-9). These radars are characterized by high power output, large antennas, and low frequencies. These same characteristics limit the accuracy of the target parameters available from early warning radars. The long-range detection of aircraft and the earliest possible attack warning capabilities of early warning radars provide the first line of defense for the air defense system.

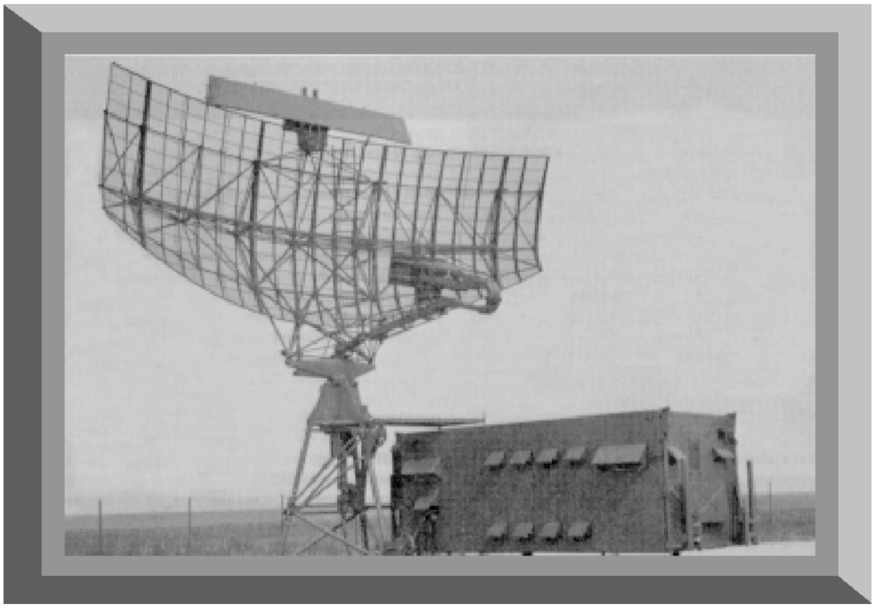

Figure 1-9. Early Warning Radar System

b. Radar systems designed to provide target engagement information include ground control intercept (GCI) radars, acquisition radars, target tracking radars (TTRs), and airborne interceptor (AI) radars.

(1) GCI radars are designed to provide sufficiently accurate target aircraft range, azimuth, and altitude information to vector AI assets to intercept and destroy attacking aircraft (Figure 1-10). To provide this data, early warning radars can be deployed along with specialized height finder radars. This combination of radar systems is commonly referred to as a GCI site. Newer GCI radar systems, employing phased array antennas and Doppler processing, can provide the required 3-dimensional target information. Any radar system, or combination of radar systems, that can determine 3-dimensional target data, and is equipped with the communication equipment to pass this information to AI assets, can act as a GCI site. GCI radar systems can be used to supplement early warning radar systems to provide critical attack warning.

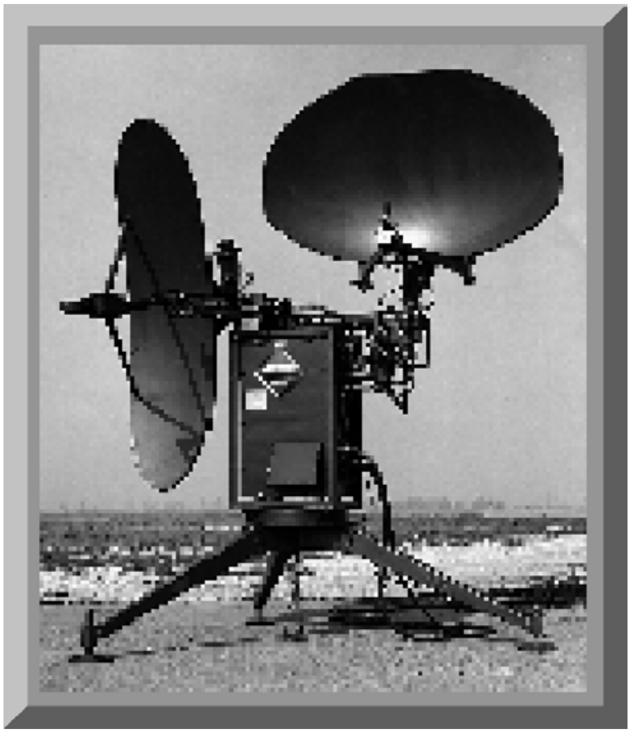

Figure 1-10. GCI Radar

(2) Acquisition radar systems are designed to act as GCI radars for ground based TTRs. Acquisition radar systems generally have shorter range capability than early warning radars and operate at higher frequencies. These radar systems provide accurate target range and azimuth data to TTRs to facilitate target engagement. Acquisition radars can be a distinct radar system (Figure 1-11) or be incorporated as part of the TTR (Figure 1-12).

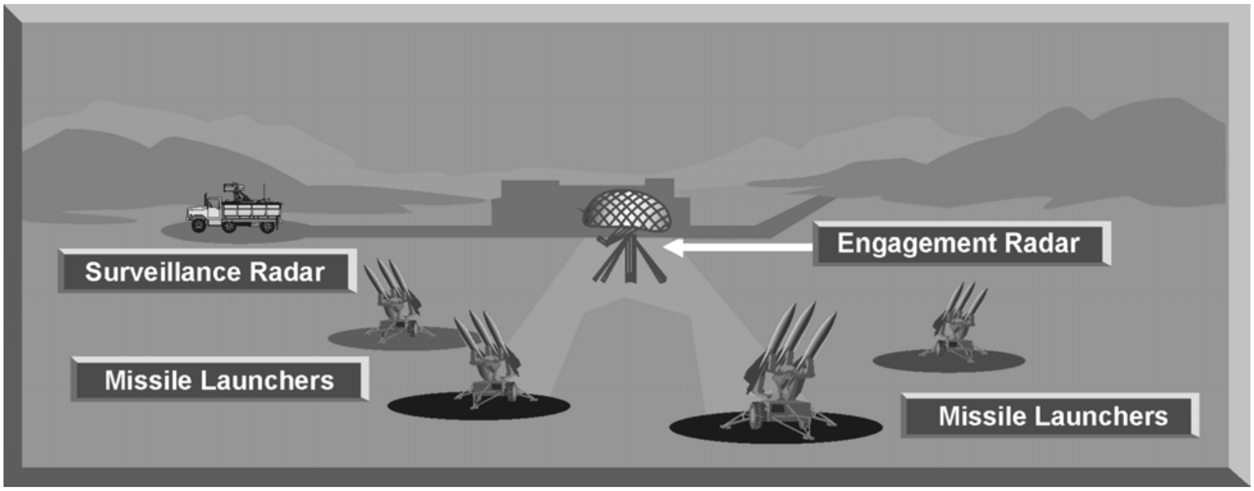
Engagement Radar
Surveillance Radar
Missile Launchers
Missile Launchers

Figure 1-11. Acquisition Radar and TTR

Figure 1-12. TTR with Acquisition

(3) The primary role of TTRs, in support of an air defense system, is to provide continuous and accurate target parameters to a fire control computer. The fire control computer uses this data to guide missiles or aim antiaircraft artillery (AAA) to destroy attacking aircraft. TTRs employ various tracking techniques to continuously update target parameters. TTRs generally employ high frequencies, narrow beamwidths, and computer signal processing to enhance the accuracy of target parameters provided to the fire control computer.

(4) AI radar systems are TTRs employed by fighter aircraft to engage and destroy airborne targets (Figure 1-13). These radar systems are characterized by high frequency, sophisticated computer processing, and accurate target tracking capability. They are designed to allow the AI asset to employ air-to-air missiles and guns/cannons. TTRs and AI radars constitute the highest radar threat associated with an air defense system.

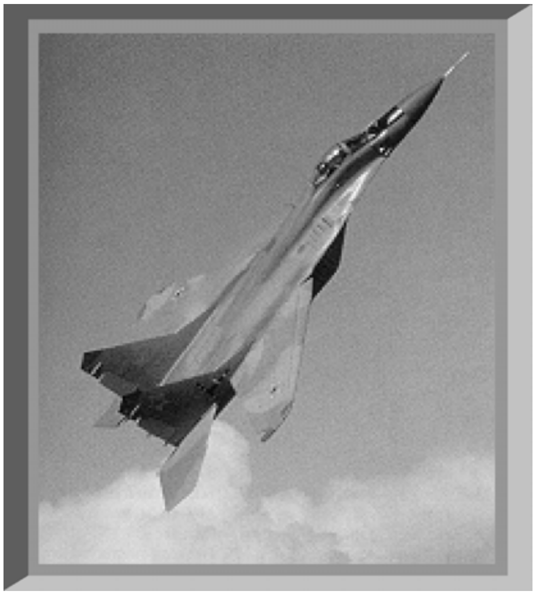

Figure 1-13. MiG-29

(5) Another growing lethal threat associated with an air defense system is infrared (IR) missiles. IR missile systems can be man-portable (Figure 1-14), mounted on vehicles, or employed by AI assets. These missile systems guide on the distinctive IR signature of aircraft. The recent proliferation and enhanced performance of IR systems has increased the contribution of these systems to air defense.

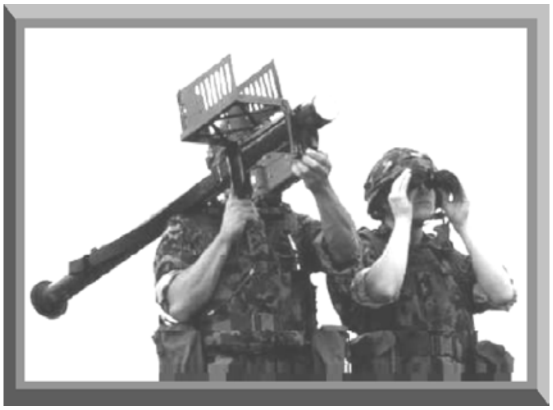

Figure 1-14. IR System

c. All these radar systems can be deployed to provide air defense for a particular country or geographical area. When the employment of these radar systems is integrated by a command and control (C^2) structure, this constitutes an IADS (Figure 1-15). The C^2 structure allows the military commander to take advantage of the threat warning provided by early warning radars. Based on this threat warning, the military commander can allocate specific assets (GCI and AI assets, or acquisition radars and TTRs) to engage airborne targets. This allocation decision is based on the capabilities of these systems and the tactical situation. This allocation process enables the military commander to maximize the capabilities of his forces to engage and destroy attacking aircraft.

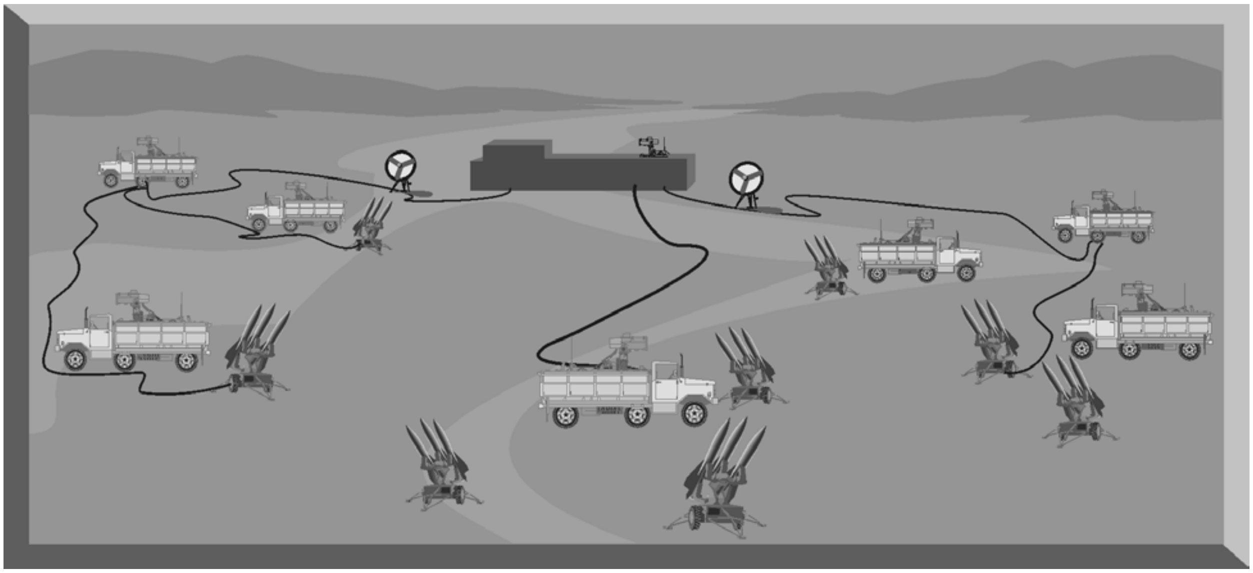

Figure 1-15. A Typical IADS

5. SUMMARY

Radar systems are the cornerstones of a modern IADS. Radar and IR threat systems operate at frequencies that span most of the electromagnetic spectrum. Each system has unique capabilities and operating characteristics that enable it to accomplish assigned tasks in support of the IADS. In order to effectively employ offensive air power on the modern battlefield, the systems that support the IADS must be negated. A basic knowledge of how radar and IR systems operate, their capabilities, limitations, and the available countermeasures is the key to defeating these systems. The purpose of this book is to provide this information.

CHAPTER 2. CHARACTERISTICS OF RF RADIATION

1. INTRODUCTION

In order for a radar system to determine range, azimuth, elevation, or velocity data, it must transmit and receive electromagnetic radiation. This electromagnetic radiation is referred to as radio frequency (RF) radiation. RF transmissions have specific characteristics that determine the capabilities and limitations of a radar system to provide these target discriminants, based on an analysis of the characteristics of the target return. The frequency of transmitted RF energy affects the ability of a radar system to analyze target return, based on time, to determine target range. RF frequency also affects the ability of the transmitting antenna to focus RF energy into a narrow beam to provide azimuth and elevation information. The wavelength and frequency of the transmitted RF energy impact the propagation of the radar signal through the atmosphere. The polarization of the RF signal affects the amount of clutter the radar must contend with. The ability of a radar system to use the Doppler effect in analyzing the radar return impacts the velocity discrimination capability of the radar. These characteristics of RF radiation will be discussed in this chapter.

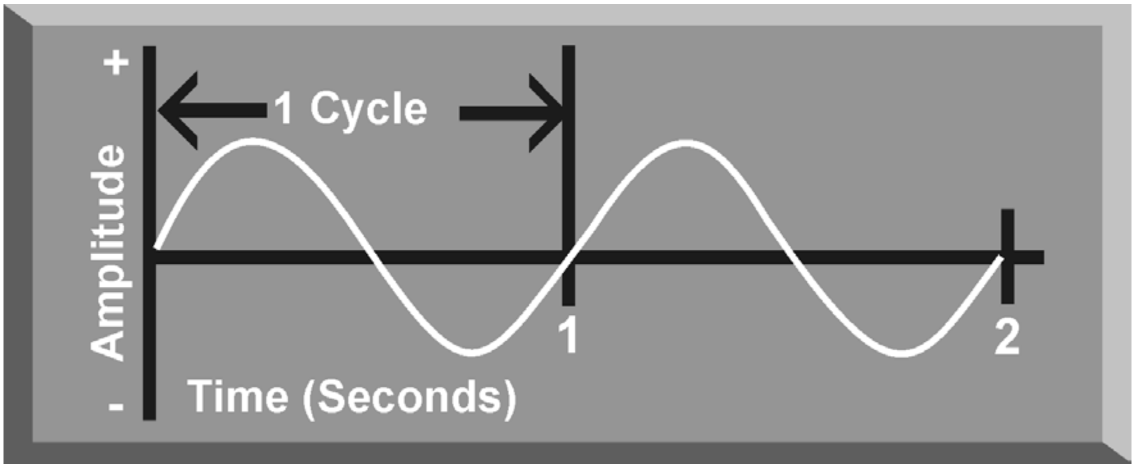

Figure 2-1. Radio Frequency

2. FREQUENCY

The output signal from a typical radio or radar system has several important characteristics that affect the capabilities and limitations of radio or radar systems. The first characteristic considered is usually RF. The frequency of the transmitted signal is the number of times per second the RF energy completes one cycle. The RF signal depicted in Figure 2-1 has a frequency of one cycle per second. The basic unit of measurement is the hertz (Hz). One hertz equals one cycle per second. Most radars have an RF in the millions of hertz, or megahertz (MHz).

3. WAVELENGTH

A characteristic of any RF signal is wavelength. Wavelength is a measure of the physical distance between peaks of a sine wave propagated in space (Figure 2-2).

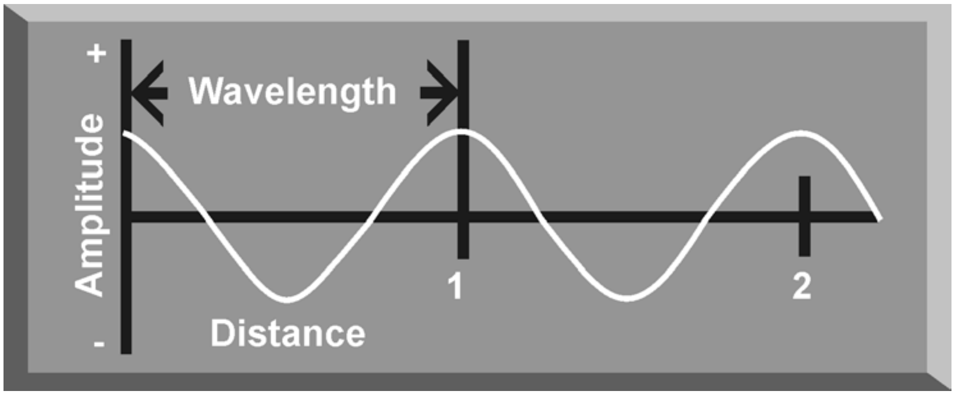

Figure 2-2. Radar Signal Wavelength

Though wavelength is measured in meters, most radar signals have wavelengths measured in centimeters or millimeters. The wavelength of a radar signal can be computed using the equation shown in Equation 2-1. The relationship between wavelength and frequency is inverse: the higher the frequency, the shorter the wavelength. In early radio and radar terminology, wavelength was used instead of frequency to describe operating characteristics of a system. Today, wavelength is used to describe systems operating at very high frequencies, such as millimeter wave, and for describing infrared (IR) systems.

$$\text{Wavelength } (\lambda) = \frac{c \quad \text{(Speed of Light)}}{f \quad \text{(Radar Frequency)}}$$

Equation 2-1. Basic Wavelength Equation

4. POLARIZATION

Another characteristic of a radio frequency wave is polarization. Polarization is determined by the radar antenna and refers to the orientation of the RF wave as it travels through space. There are two types of polarization: linear and circular.

a. Traveling electromagnetic energy has two components: an electrostatic field and a magnetic field. These two fields are always perpendicular to each other and perpendicular to the direction of travel. The polarization of the wave is defined in terms of the orientation to the electrostatic field. Many radar antennas are linearly polarized, either vertically or horizontally. The signal depicted in Figure 2-3 is vertically polarized.

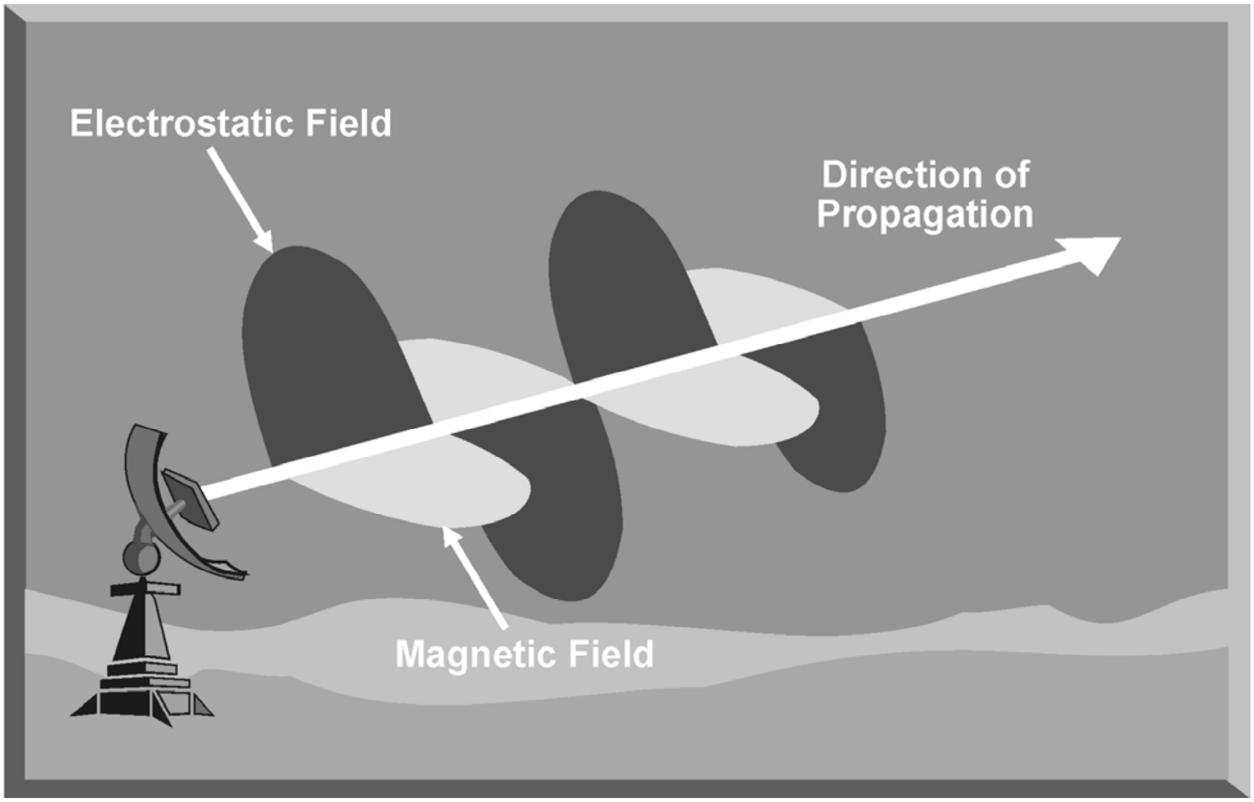

Figure 2-3. Radar Signal Vertical Polarization

b. Some radars use circular polarization to improve target detection in rain. Circular polarization can be right-hand, or left-hand orientation. For circular polarization, the direction of the electrostatic field varies with time and traces a circular locus about a fixed plane perpendicular to the direction of propagation. For a right-hand circular polarized signal, the electrostatic vector appears to rotate in a clockwise direction. For a left-hand circular polarized signal, the rotation is counterclockwise. Circular polarization can be visualized by pointing the thumb of either hand in the direction of propagation and curling the fingers in the direction of electrostatic field rotation (Figure 2-4).

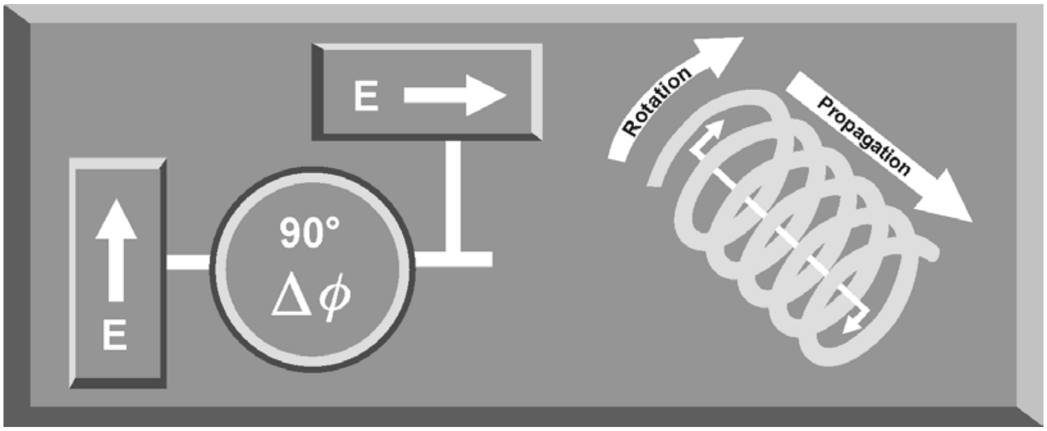
E
E
90°
$\Delta\phi$
Rotation
Propagation

Figure 2-4. Circular Polarization

c. The impact of polarization on receivers and transmitters is fairly straightforward. If an antenna is designed to receive a particular polarization, it will have difficulty receiving a signal with an opposite polarization. This situation is defined as cross polarization (Figure 2-5).

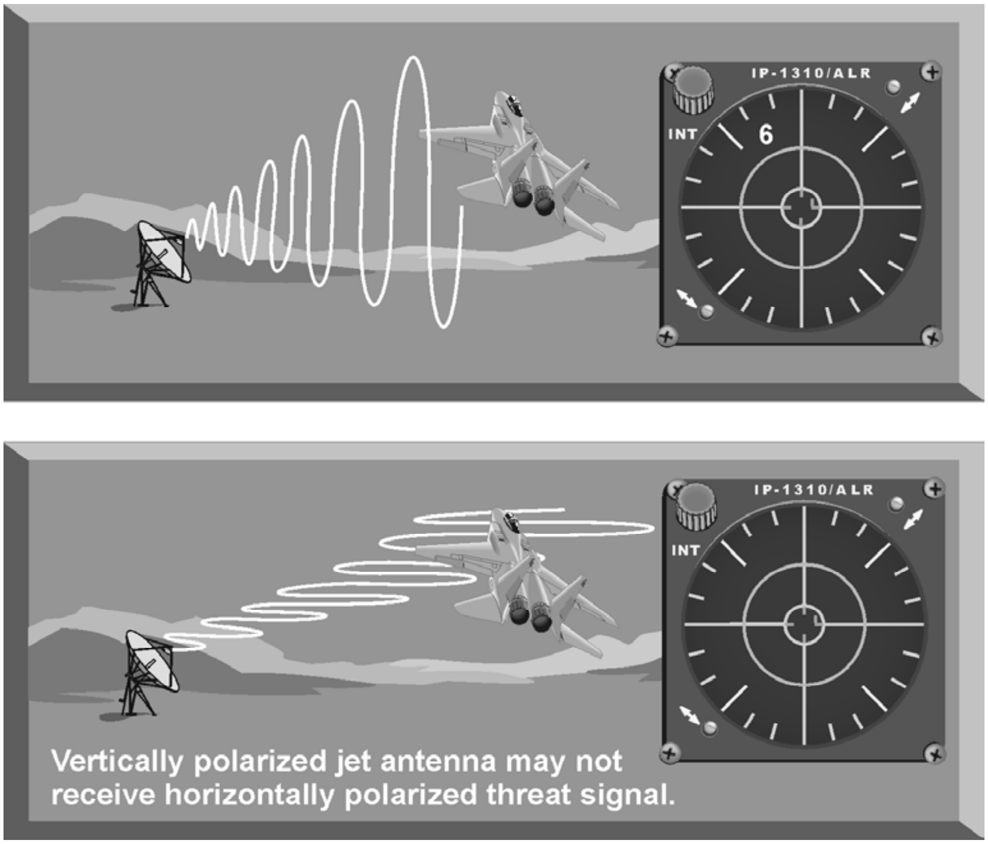
IP-1310/ALR
INT
6
IP-1310/ALR
INT
Vertically polarized jet antenna may not receive horizontally polarized threat signal.

Figure 2-5. Impact of Polarization

The impact of cross polarization on electronic combat can be dramatic. If a radar warning receiver antenna is polarized to receive vertically polarized signals, a threat system employing a horizontally polarized radar signal may not be detected, or may be displayed on the scope well after the threat has acquired the aircraft. In addition, if the jamming antenna on an electronic attack (EA) system is also vertically polarized, it may not be able to jam this system. Fortunately, this potentially lethal situation rarely occurs, but future threat systems may take advantage of this situation. Table 2-1 details the impact of polarization on selected transmit and receive antenna combinations.

Table 2-1. Antenna Polarization Loss

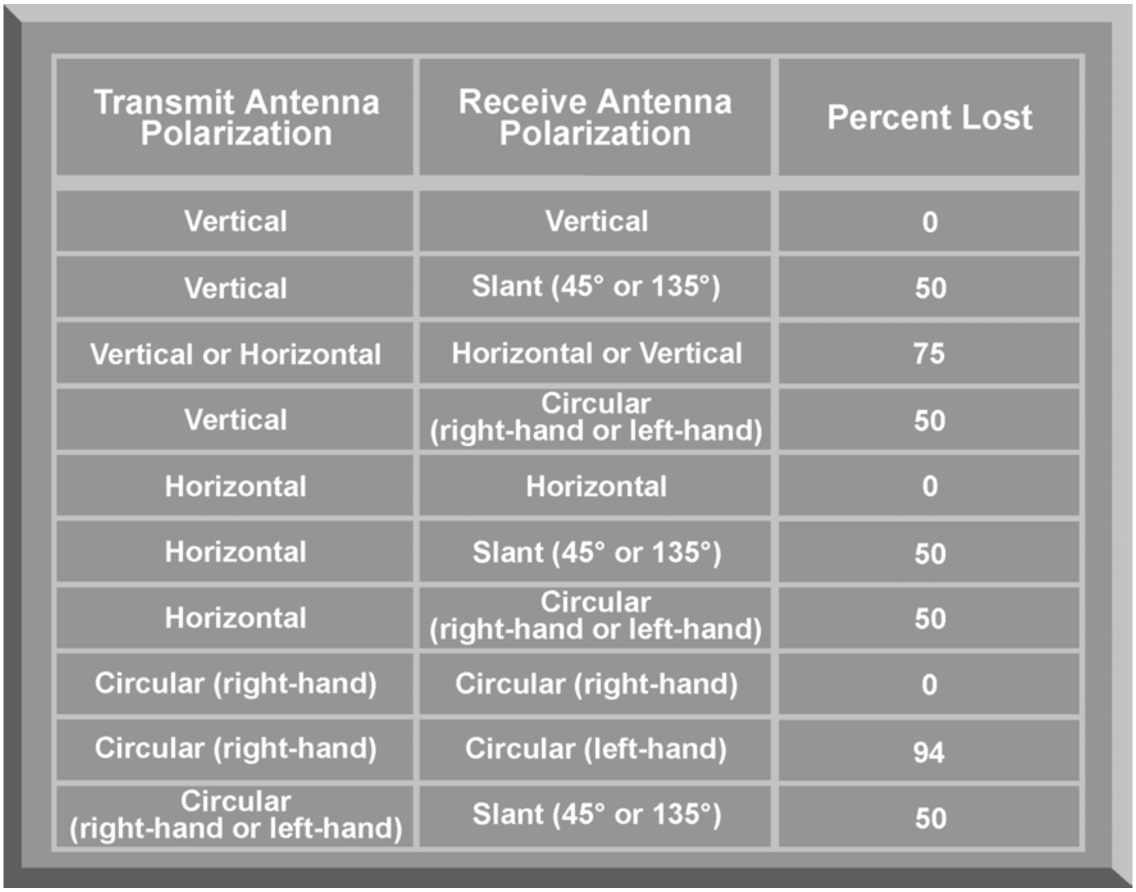

Transmit Antenna Polarization	Receive Antenna Polarization	Percent Lost
Vertical	Vertical	0
Vertical	Slant (45° or 135°)	50
Vertical or Horizontal	Horizontal or Vertical	75
Vertical	Circular (right-hand or left-hand)	50
Horizontal	Horizontal	0
Horizontal	Slant (45° or 135°)	50
Horizontal	Circular (right-hand or left-hand)	50
Circular (right-hand)	Circular (right-hand)	0
Circular (right-hand)	Circular (left-hand)	94
Circular (right-hand or left-hand)	Slant (45° or 135°)	50

5. DOPPLER EFFECT

The “Doppler effect” takes advantage of the fact that the frequency of RF waves will be changed or shifted when reflected from a target moving relative to the radar. The shifted frequency of the returning RF wave depends on the movement of the aircraft in relation to the radar. In Figures 2-6, 2-7, and 2-8, f_o is the transmitted frequency of the radar, and f_t is the frequency of the reflected RF

wave from the target. For a stationary target, the frequency of the reflected signal will equal the frequency of the transmitted signal (Figure 2-6).

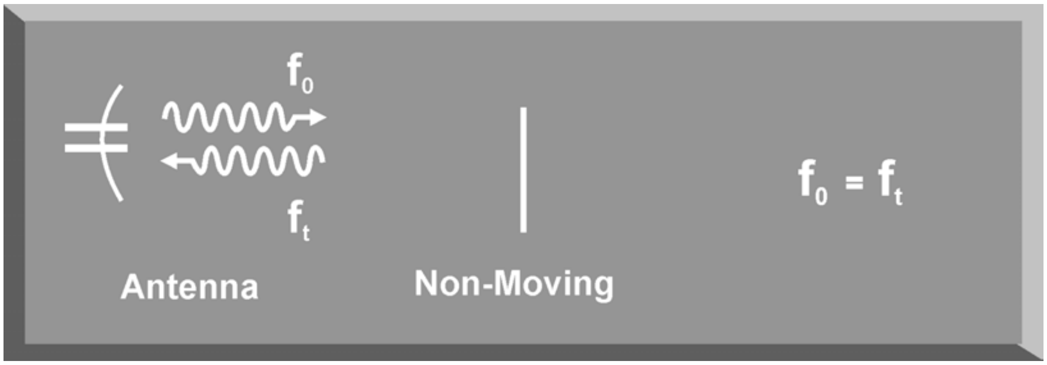
f_0
f_t
Antenna
Non-Moving
$f_0 = f_t$

Figure 2-6. Zero Doppler Effect – Stationary Target

a. For a target moving toward the radar, the frequency of the reflected signal will be higher than the transmitted signal (Figure 2-7).

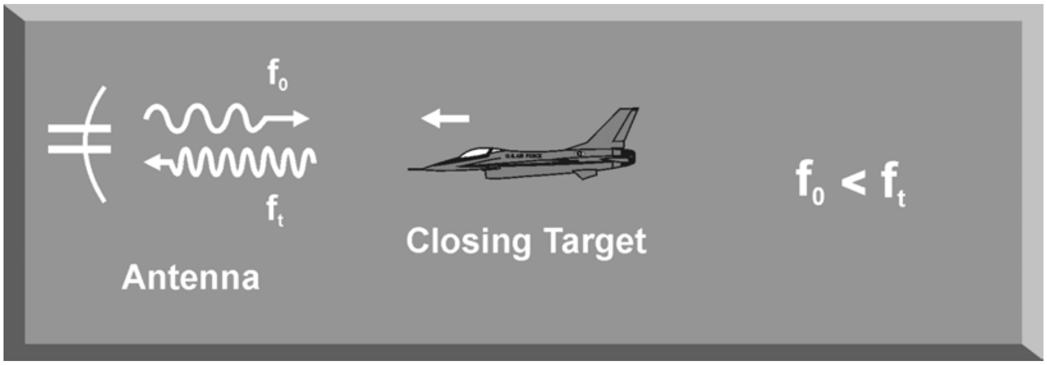
f_0
f_t
Antenna
Closing Target
$f_0 < f_t$

Figure 2-7. Doppler Effect – Closing Target

b. The reflected frequency for a target moving away from the radar will be lower than the transmitted frequency (Figure 2-8).

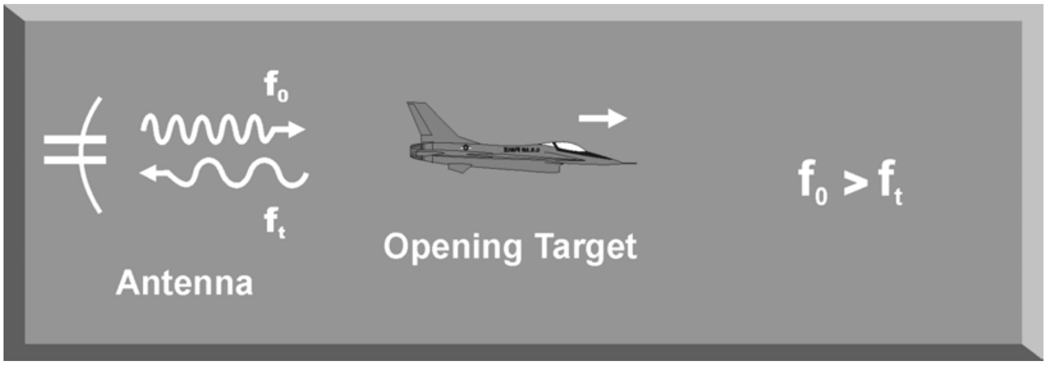
f_0
f_t
Antenna
Opening Target
$f_0 > f_t$

Figure 2-8. Doppler Effect – Opening Target

6. ELECTROMAGNETIC SPECTRUM

The portion of the electromagnetic spectrum (Figure 2-9) that today's electronic combat systems must deal with starts with radio waves and encompasses microwaves, infrared, and a small portion of the ultraviolet region. Communications systems generally operate in the HF, UHF, and VHF regions. Some satellite communications operate in the SHF region. Radars operate in the microwave region, normally from 0.2 – 200 gigahertz. Infrared systems operate in the region just below visible light.

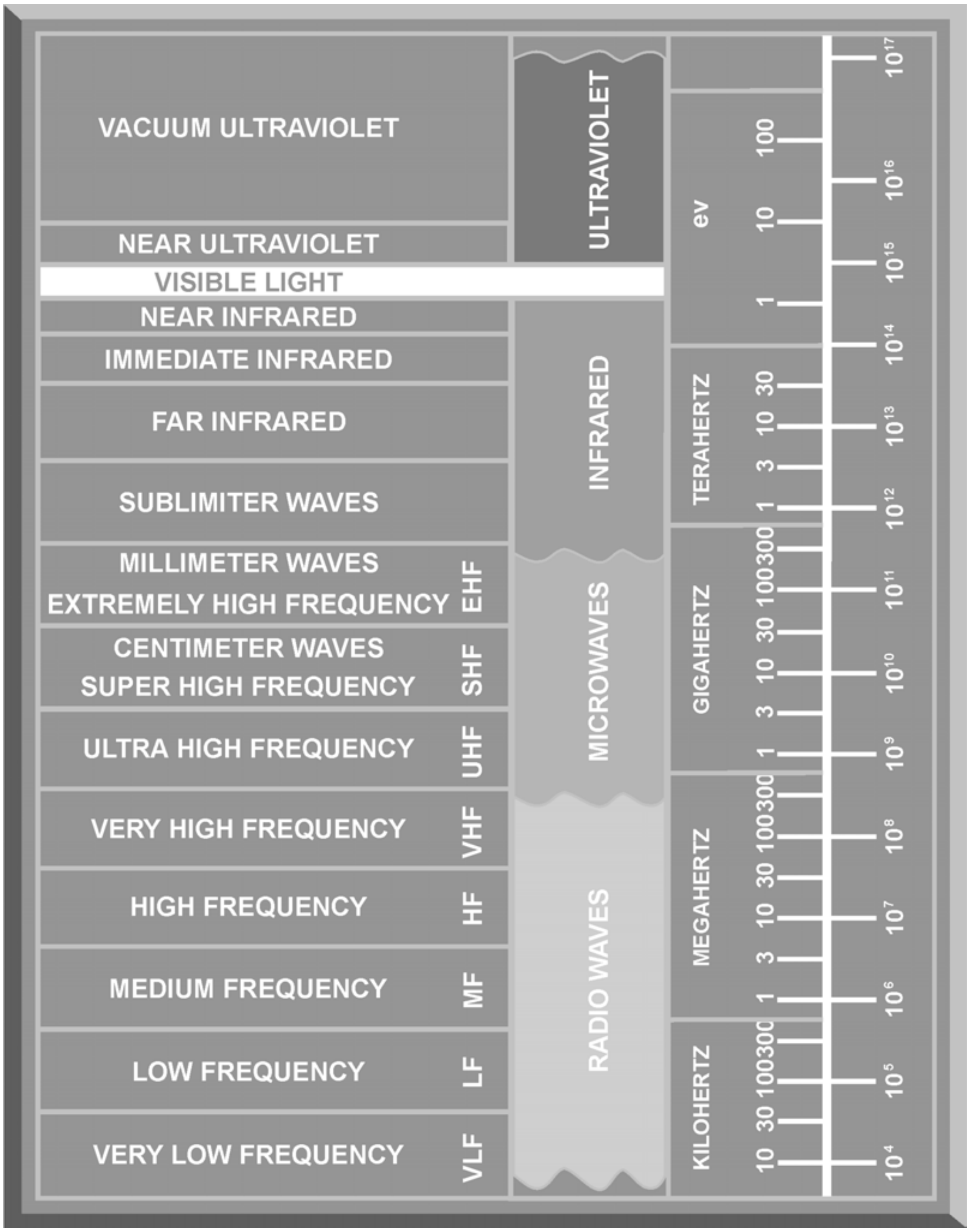
VACUUM ULTRAVIOLET
NEAR ULTRAVIOLET
VISIBLE LIGHT
NEAR INFRARED
IMMEDIATE INFRARED
FAR INFRARED
SUBLIMITER WAVES
MILLIMETER WAVES
EXTREMELY HIGH FREQUENCY
EHF
CENTIMETER WAVES
SUPER HIGH FREQUENCY
SHF
ULTRA HIGH FREQUENCY
UHF
VERY HIGH FREQUENCY
VHF
HIGH FREQUENCY
HF
MEDIUM FREQUENCY
MF
LOW FREQUENCY
LF
VERY LOW FREQUENCY
VLF
ULTRAVIOLET
INFRARED
MICROWAVES
RADIO WAVES
ev
1
10
100
TERAHERTZ
1
3
10
30
GIGAHERTZ
1
3
10
30
100
300
MEGAHERTZ
1
3
10
30
100
300
KILOHERTZ
10
30
100
300
10^{17}
10^{16}
10^{15}
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}

Figure 2-9. Electromagnetic Spectrum

a. Table 2-2 lists prefixes commonly used to indicate scientific notation and their abbreviations. These prefixes are often used when dealing with extremely high frequency radar signals or RF waves with extremely small wavelengths. For example, when dealing with extremely high frequency RF, it is simpler to use gigahertz, or GHz. One gigahertz is one thousand megahertz (MHz). For example, a radar operating at 3150 MHz is also operating at an RF of 3.150 GHz.

Table 2-2. Commonly Used Prefixes for Scientific Notation

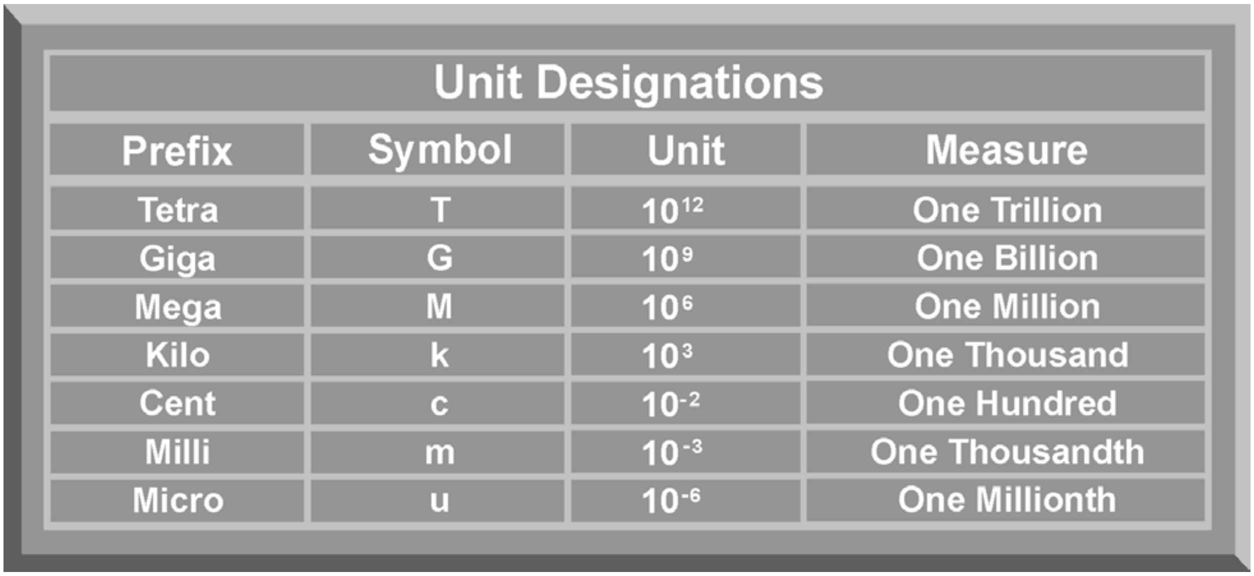

Unit Designations			
Prefix	Symbol	Unit	Measure
Tetra	T	10^{12}	One Trillion
Giga	G	10^{9}	One Billion
Mega	M	10^{6}	One Million
Kilo	k	10^{3}	One Thousand
Cent	c	10^{-2}	One Hundred
Milli	m	10^{-3}	One Thousandth
Micro	u	10^{-6}	One Millionth

b. Frequency bands are often used when discussing radars and electronic combat systems. Radar designers use a frequency band designation system entirely different from the one used in electronic combat. Table 2-3 depicts both frequency band designation systems and the corresponding frequency ranges.

Table 2-3. Radar Frequency Band Designations

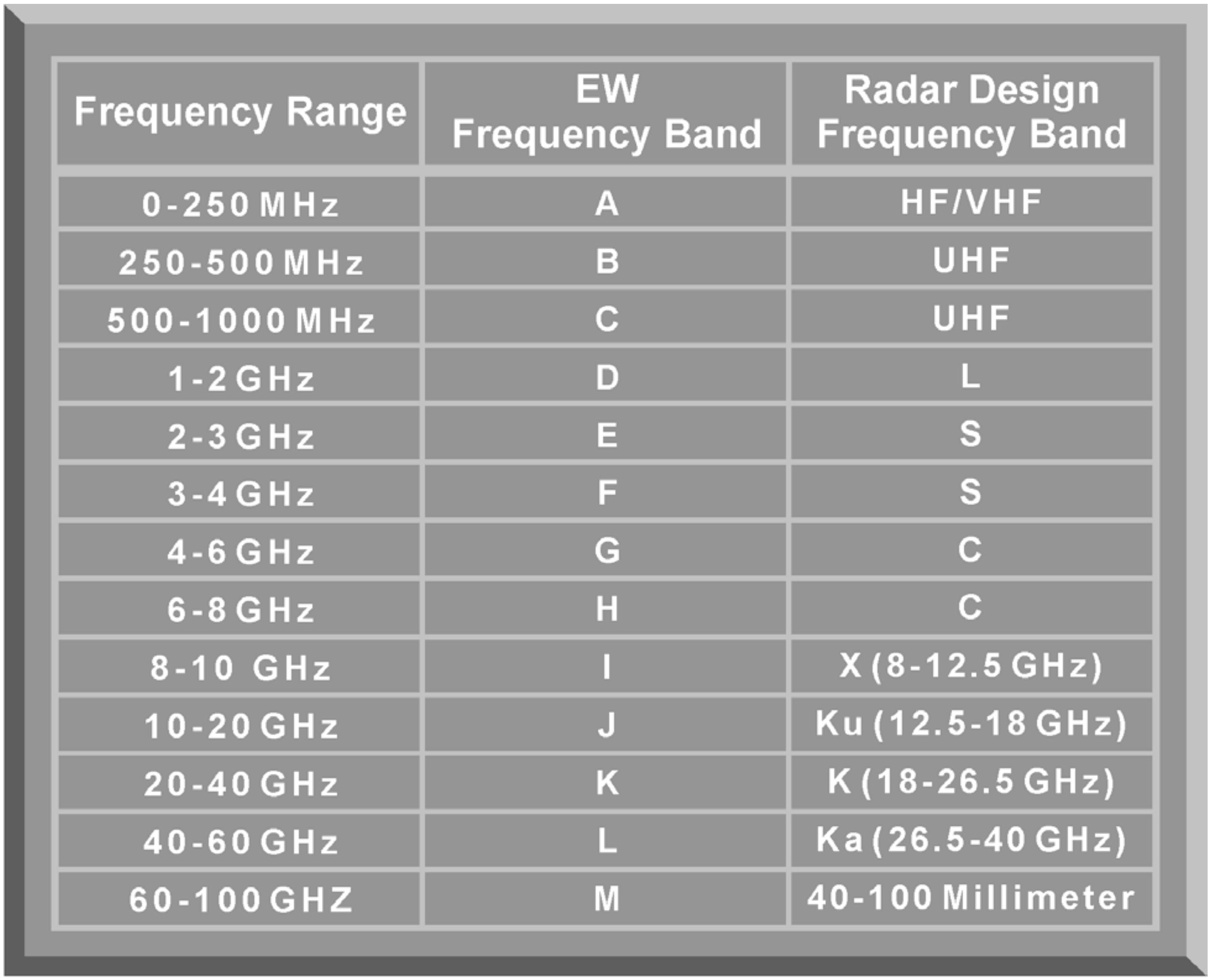

Frequency Range	EW Frequency Band	Radar Design Frequency Band
0-250 MHz	A	HF/VHF
250-500 MHz	B	UHF
500-1000 MHz	C	UHF
1-2 GHz	D	L
2-3 GHz	E	S
3-4 GHz	F	S
4-6 GHz	G	C
6-8 GHz	H	C
8-10 GHz	I	X (8-12.5 GHz)
10-20 GHz	J	Ku (12.5-18 GHz)
20-40 GHz	K	K (18-26.5 GHz)
40-60 GHz	L	Ka (26.5-40 GHz)
60-100 GHZ	M	40-100 Millimeter

7. RF PROPAGATION

Propagation characteristics of RF energy are profoundly affected by the earth's surface and atmospheric conditions. Any analysis of radar performance must take into account the propagation phenomena associated with RF radiation in a "real world" environment. The most important propagation phenomena include refraction, anomalous propagation (ducting), and attenuation.

a. In a vacuum, RF waves travel in a straight line. However, RF waves propagating within the earth's atmosphere do not travel in a straight line. The earth's atmosphere bends, or refracts, RF waves. One impact of the atmospheric refraction of RF waves is an increase in the line of sight (LOS) of the radar. This increase in radar LOS effectively extends the range of the radar system (Figure 2-10).

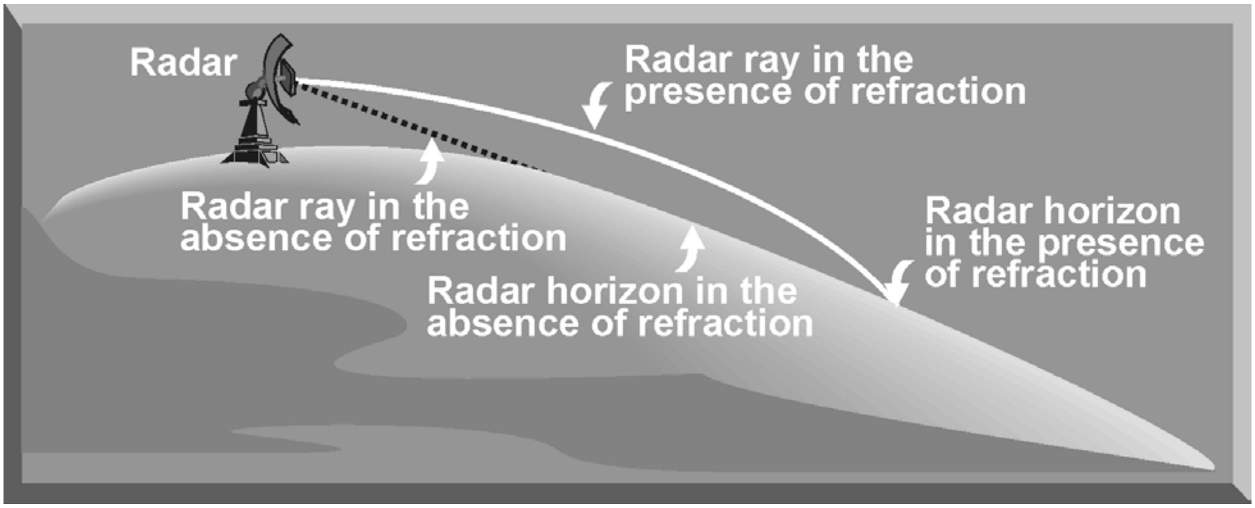
Radar
Radar ray in the presence of refraction
Radar ray in the absence of refraction
Radar horizon in the presence of refraction
Radar horizon in the absence of refraction

Figure 2-10. Impact of Refraction on RF Propagation

Atmospheric refraction of RF energy can also induce elevation measurement errors in radar systems (Figure 2-11).

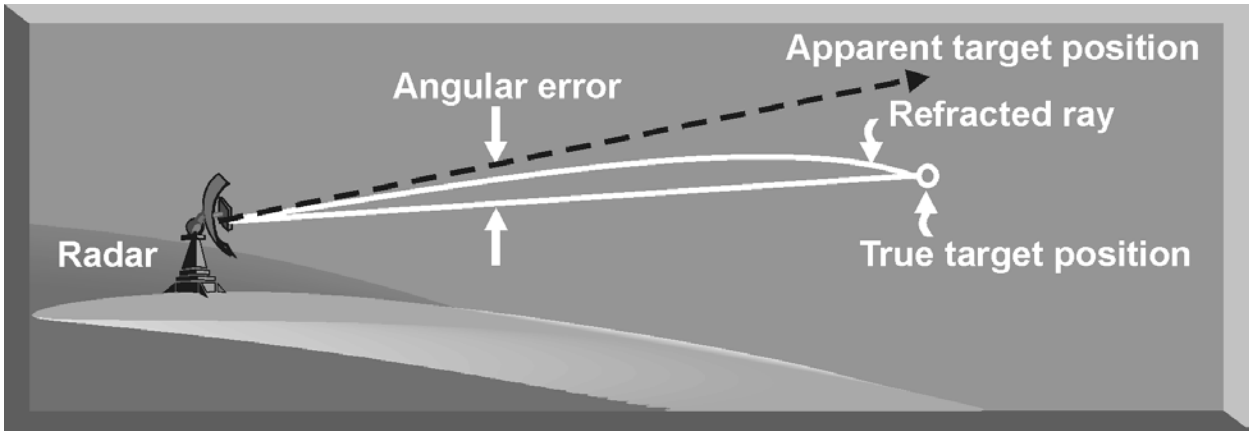
Apparent target position
Angular error
Refracted ray
Radar
True target position

Figure 2-11. Impact of Refraction on Target Elevation Determination

(1) The refraction of RF waves in the atmosphere is caused by the variation in the velocity of propagation with altitude. The index of refraction (n) is used to describe this velocity variation and is defined by Equation 2-2.

$$\text{Index of Refraction } (n) = \frac{\text{Velocity of Propagation in a Vacuum}}{\text{Velocity of Propagation in the Atmosphere}}$$

Equation 2-2. Index of Refraction

(2) The term refractivity (N) is used for predicting the impact of refraction on RF wave propagation. Refractivity is a “scaled up” expression for the index of refraction and is used by radar designers to calculate the impact of refraction on actual radar systems. At normal radar operating frequencies, the refractivity for air containing water vapor can be computed using Equation 2-3.

$$\text{Refractivity } (N) = (n-1) \times 10^{6} = \frac{77.6\,p}{T} + \frac{3.75 \times 10^{5} e}{T^{2}}$$

n = index of refraction

p = barometric pressure

e = partial pressure of water vapor

T = absolute temperature (degrees K)

Equation 2-3. Refractivity of RF Waves (Normal Radar Frequencies)

(3) As altitude increases, the barometric pressure and water vapor content decrease rapidly. At the same time, the absolute temperature decreases slowly based on the standard lapse rate. Using Equation 2-3, it can be seen that the refractivity of the atmosphere decreases with increasing altitude. This decrease in refractivity means that the velocity of RF waves increases with altitude. The result is a downward bending, or refraction, of RF waves as depicted in Figure 2-10. RF wave refraction primarily affects ground-based radar systems at low antenna elevation angles, especially at or near the horizon. For most radar applications, refraction is not a factor at elevation angles above 5 degrees.

b. The term anomalous, or nonstandard, propagation is used to describe atmospheric conditions that extend the propagation of RF waves and increase radar range. The most common anomalous propagation phenomena is called superrefraction, or ducting.

(1) A superrefracting duct is formed when the refractivity of the atmosphere (Equation 2-2) rapidly decreases with altitude. Based on Equation 2-2, this occurs when the temperature increases with altitude and/or the water vapor content decreases with altitude. An increase in temperature with altitude is called a temperature inversion. To produce a duct, the temperature inversion must be very pronounced.

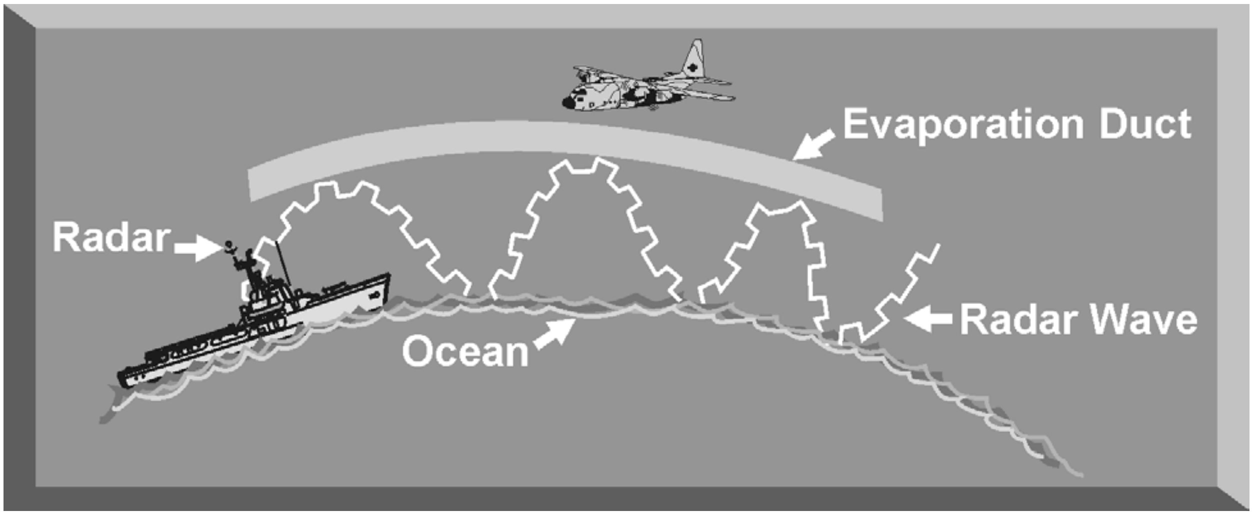

Figure 2-12. Superrefracting Surface Duct

(2) A superrefracting duct acts like a wave guide which traps the RF wave (Figure 2-12). This channels the radar signal and reduces attenuation. In order for an RF wave to propagate within a duct, the angle of the radar signal, in relation to the duct, should be less than one degree. The RF waves trapped by the duct take advantage of the decrease in refractivity and travel much further than normal. This can greatly extend the range of a radar system.

(3) The extension of radar range inside a duct can result in a reduction of radar coverage outside the duct. The area of reduced radar coverage because of ducting is called a radar hole. Due to radar holes, the extended radar range caused by ducting may result in a decrease in radar coverage along other paths of propagation. These holes can seriously degrade the effectiveness of early warning radar systems. For example, a radar system is taking advantage of a duct formed at the surface to extend low altitude radar range (Figure 2-12). Airborne targets flying just above the duct would normally be detected, but because of ducting, these targets may be missed.

(4) Water vapor content is a significant factor in producing ducts. Consequently, most ducts are formed over water and in warm climates. Any atmospheric phenomenon that results in a pronounced increase in temperature and/or a decrease in water vapor content as altitude increases can generate a superrefracting duct, of which there are three types. A superrefracting duct which is formed just above the surface of the earth is referred to as a surface duct. A surface duct formed just above the surface of the ocean is called an evaporation duct. A duct which is formed well above the surface of the earth is known as an elevated duct.

(a) Surface ducts formed over land are usually a result of the nighttime radiation of heat from the earth. Duct formation is especially prevalent during the

summer months when the ground is moist. As the earth loses heat, a temperature inversion is created at the surface coupled with a sharp decrease in the moisture content. These conditions are favorable to the formation of a surface duct. A superrefracting duct can also be produced by the diverging downdraft under a thunderstorm. The cool air that is dispersed creates a local temperature inversion while the water vapor content decreases due to rain. Surface ducts formed in conjunction with thunderstorms are difficult to predict and normally persist for a short period of time.

(b) A superrefracting surface duct that lies just above the surface of the ocean is a result of evaporated water, thus the term evaporation duct. The air in contact with the ocean is saturated with water vapor, while the air several feet above the ocean contains a much lower level. This rapid decrease in water vapor pressure with an increase in altitude creates an evaporation duct. An evaporation duct exists over the ocean almost all the time. The height of this duct varies from 20 to 100 feet based on the season, time of day, and wind speed. One positive aspect of an evaporation duct is the extended range available to a shipborne radar system with a properly aligned antenna. This extended range coverage against surface ships and low altitude aircraft is a definite advantage of ducted propagation.

(c) An elevated duct is generally formed by a temperature inversion in the upper atmosphere. To take maximum advantage of the increased radar range inside an elevated duct, both the radar and the target should be inside the elevated duct. In addition, radar systems operating below an elevated duct may also experience enhanced range performance.

(4) The presence of surface ducts and elevated ducts, especially over land, are extremely difficult to predict and may persist for very short periods of time. The atmospheric conditions favorable to duct formation are difficult to predict using conventional weather forecasting techniques.

c. The attenuation of RF energy in a clear atmosphere is due to the presence of oxygen and water vapor. Attenuation results when a portion of the RF energy strikes these molecules and is absorbed as heat. Figure 2-13 details the RF attenuation loss due to atmospheric gasses based on the frequency of the RF energy.

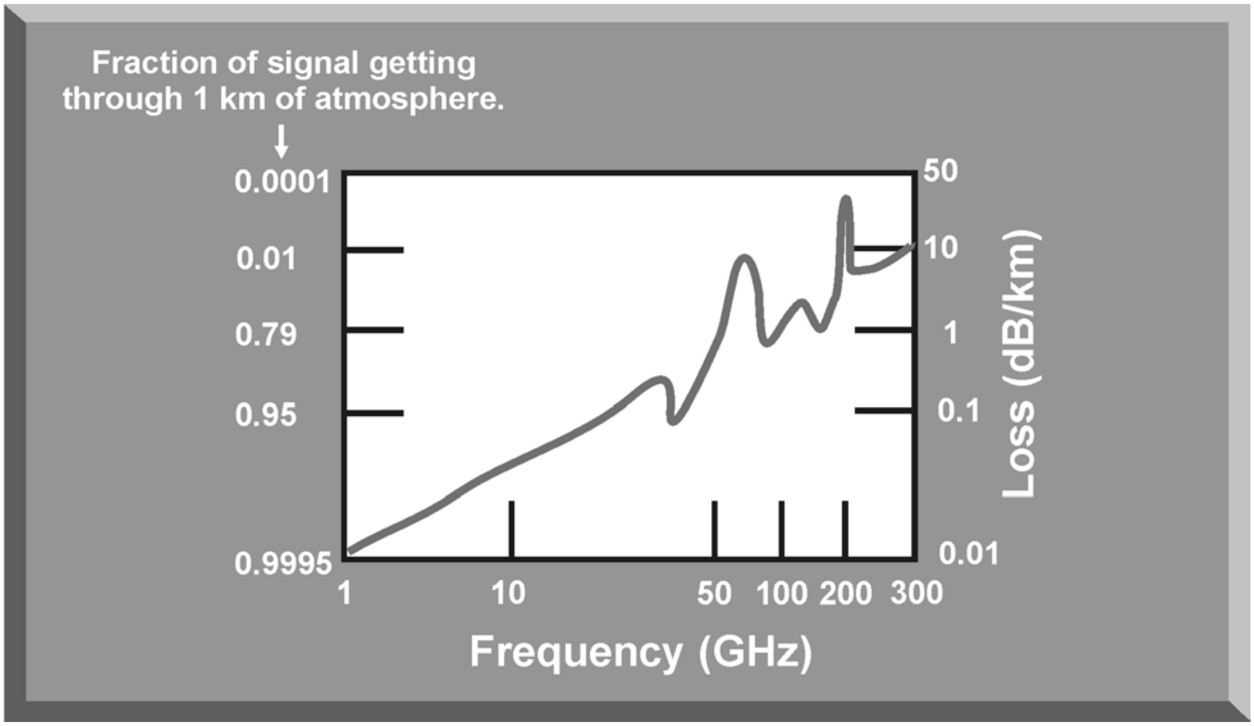

Figure 2-13. RF Atmospheric Attenuation

(1) At frequencies below 1 GHz, the effect of atmospheric attenuation is negligible. Above 10 GHz, atmospheric attenuation increases dramatically. This dramatic signal loss impacts the maximum detection range of radars operating in the millimeter wavelength band.

(2) RF energy attenuation decreases as altitude increases. The RF attenuation experienced by an air-to-air radar will depend on the altitude of the target as well as target range. For a ground-based radar, RF attenuation will decrease as antenna elevation increases.

8. SUMMARY

Since all radar operations depend on the transmission and reception of RF energy, a basic knowledge of RF frequency, wavelength, and polarization provides the basis for understanding the more complex radar characteristics. Since most modern radar systems employ some form of Doppler signal processing, the concept of the Doppler effect is fundamental to understanding modern radar operation. The concepts of refraction, anomalous propagation (ducting), and atmospheric attenuation are key to understanding how RF waves propagate in the atmosphere. The topics in this chapter provide a foundation for understanding radar and jamming system operation.

CHAPTER 3. RADAR SIGNAL CHARACTERISTICS

1. INTRODUCTION

Every radar produces a radio frequency (RF) signal with specific characteristics that differentiate it from all other signals and define its capabilities and limitations. Pulse width (pulse duration), pulse recurrence time (pulse repetition interval), pulse repetition frequency, and power are all radar signal characteristics determined by the radar transmitter. Listening time, rest time, and recovery time are radar receiver characteristics. An understanding of the terms used to describe these characteristics is critical to understanding radar operation.

2. PULSE WIDTH (PW)

Figure 3-1 depicts the output from a typical pulse radar. PW, sometimes called pulse duration (PD), is the time that the transmitter is sending out RF energy. PW is measured in microseconds. It has an impact on range resolution capability, that is, how accurately the radar can discriminate between two targets based on range.

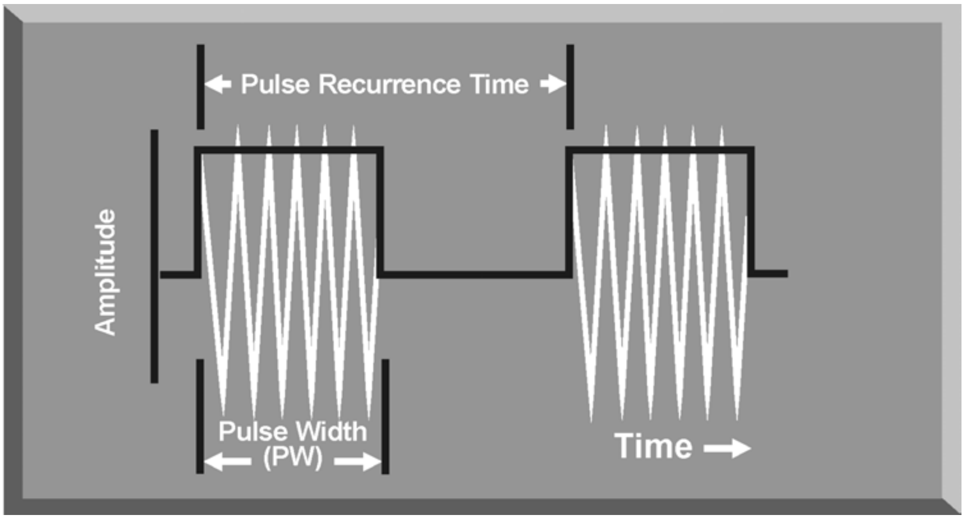

Figure 3-1. Typical Radar Pulse

3. PULSE RECURRENCE TIME (PRT)

Pulse recurrence time is also known as pulse repetition time. PRT is the time required for a complete transmission cycle. This is the time from the beginning of one pulse of RF energy to the beginning of the next. PRT is measured in microseconds. PRT is the same as pulse repetition interval (PRI), which is used in radar warning receivers and other electronic warfare support (ES) assets to discriminate between radar systems. It also affects maximum radar range.

4. PULSE REPETITION FREQUENCY (PRF)

One of the most important characteristics of a pulse radar signal is pulse repetition frequency. PRF is the rate at which pulses or pulse groups are transmitted. Generally, PRF is the number of pulses generated per second and is expressed in hertz (Hz). PRF and PRI are related in that PRI is the inverse of PRF. A word of caution—do not confuse the operating frequency of the radar, which is measured in Hz, with the pulse repetition frequency, which is also measured in Hz. They are entirely different characteristics of a pulsed radar signal.

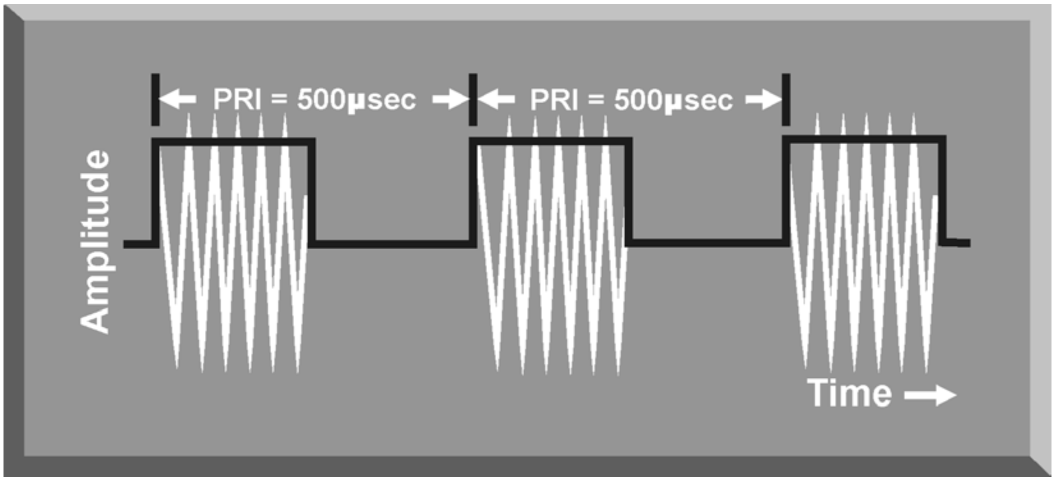

Figure 3-2. Constant PRF Radar Pulse

a. A pulse radar operating at an unvarying PRF is called a constant PRF radar (Figure 3-2). Pulse radar systems can employ PRF stagger or PRF jitter as an electronic protection (EP) technique against repeater or synchronous jammers. The time between each pulse is the PRI.

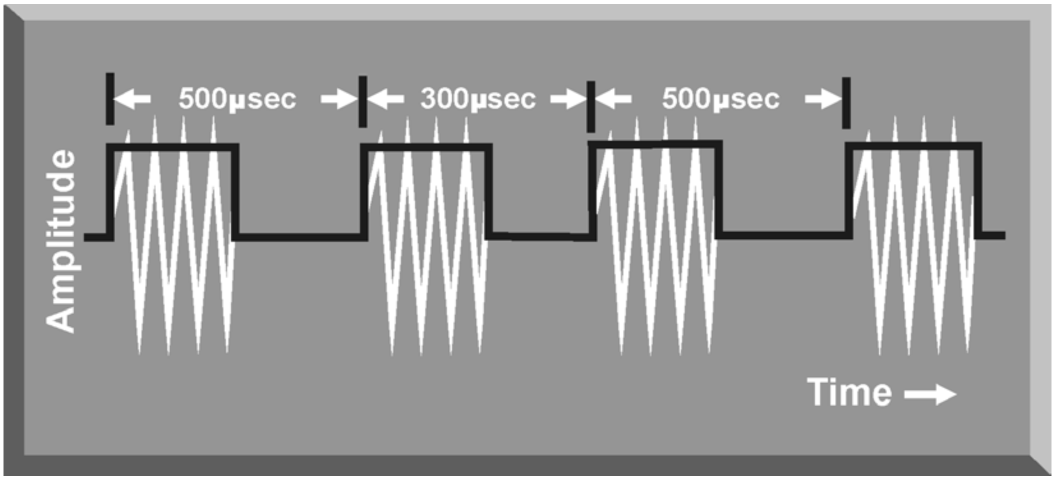

Figure 3-3. PRF Stagger

b. PRF stagger is accomplished by assuring that no adjacent PRIs are equal. The number of different PRIs generated is called the "position" of the stagger. Two-position stagger would have two PRI values, for example, 300 microseconds and 500 microseconds (Figure 3-3).

c. PRF jitter may be considered a random stagger. It is also an EP technique to counter synchronous jammers. PRF jitter has no repeating pattern of PRI values (Figure 3-4).

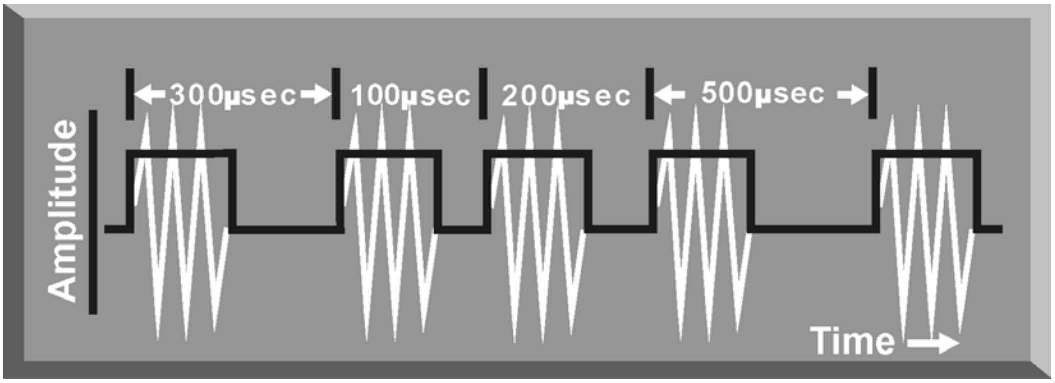
300µsec
100µsec
200µsec
500µsec
Amplitude
Time

Figure 3-4. PRF Jitter

5. RADAR RECEIVER CHARACTERISTICS

Pulse repetition frequency, pulse recurrence time, and pulse width are determined by the transmitter. The pulse radar signal characteristics that relate to receiver operation are rest time, recovery time (RT), and listening time (LT) (Figure 3-5).

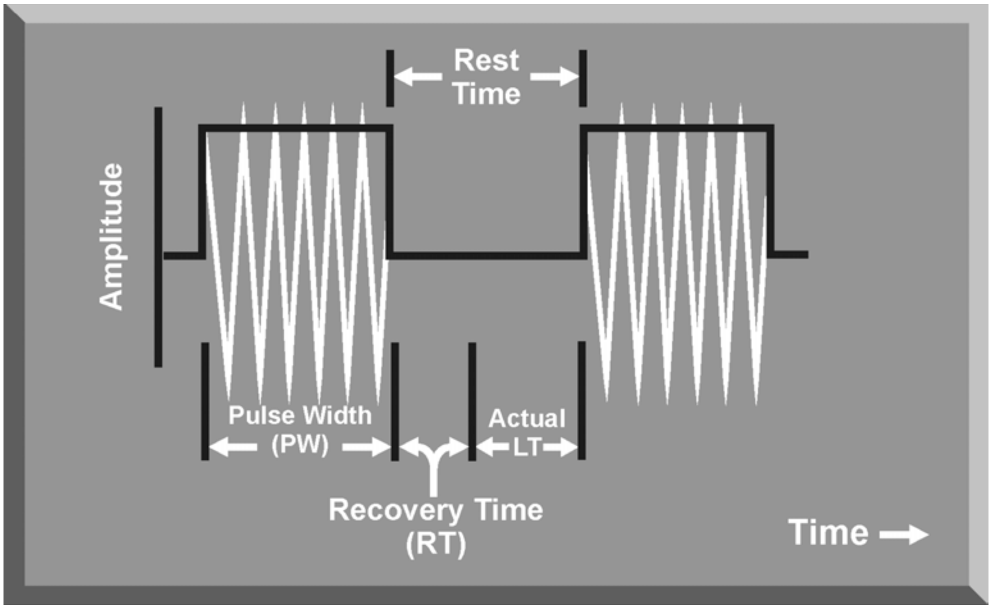
Rest Time
Amplitude
Pulse Width (PW)
Actual LT
Recovery Time (RT)
Time

Figure 3-5. Basic Radar Pulse

a. Rest time is the time between the end of one transmitted pulse and the beginning of the next. It represents the total time that the radar is not transmitting. Rest time is measured in microseconds.

b. Recovery time (RT) is the time immediately following transmission time during which the receiver is unable to process returning radar energy. RT is determined by the amount of isolation between the transmitter and receiver and the efficiency of the duplexer. A part of the high power transmitter output spills over into the receiver and saturates this system. The time required for the receiver to recover from this condition is RT.

c. Listening time (LT) is the time the receiver can process target returns. Listening time is measured from the end of the recovery time to the beginning of the next pulse, or PRT minus (PW + RT). Listening time is measured in microseconds.

6. DUTY CYCLE

Duty cycle is the ratio of the time the transmitter operates to the time it could operate during a given transmission cycle. The duty cycle of a radar can be computed by dividing the PW by the PRT, or by multiplying the PW times the PRF. Duty cycle has no units (Equation 3-1). CW radars have a duty cycle of 100%, while early warning radars may have a duty cycle of around 1%.

$$\text{Duty Cycle} = \frac{PW}{PRT} \quad \text{or} \quad \text{Duty Cycle} = PW \times PRF$$

Equation 3-1. Duty Cycle

7. PEAK POWER

The power output of a radar is normally expressed in terms of peak power or average power (Figure 3-6). Peak power is the amplitude, or power, of an individual radar pulse. It is simply the power, measured in watts or megawatts, that is radiated when the transmitter is on. The power a radar transmits is normally used to determine the maximum detection range of that radar. However, it is the energy in a radar pulse that determines maximum radar detection range. Since power is the rate of flow of energy, the energy in a radar pulse is equal to the peak power multiplied by the time the radar is transmitting, or pulse width (Equation 3-2).

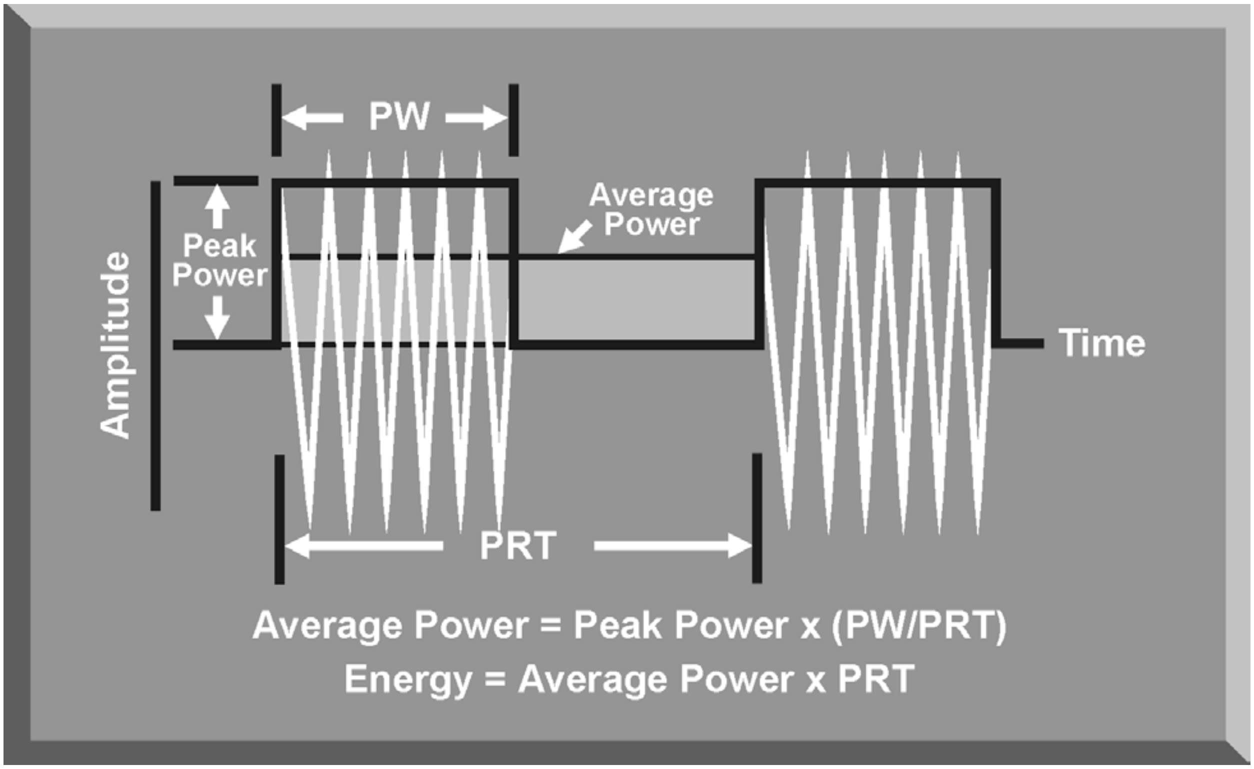
PW
Average Power
Peak Power
Amplitude
Time
PRT
Average Power = Peak Power x (PW/PRT)
Energy = Average Power x PRT

Figure 3-6. Peak Power and Average Power

Energy Per Pulse = Peak Power x Pulse Width

Equation 3-2. Energy Per Pulse

8. AVERAGE POWER

Average power is the power distributed over the pulse recurrence time. It can be computed using the formula in Figure 3-6. The energy transmitted by average power can be computed by multiplying average power by PRT. Since the energy in a set of pulses determines detection range, average power or energy provides a better measure of the detection range of a radar than does peak power. Average power can be increased by increasing the PRF, by increasing the pulse width, or by increasing peak power (Figure 3-7).

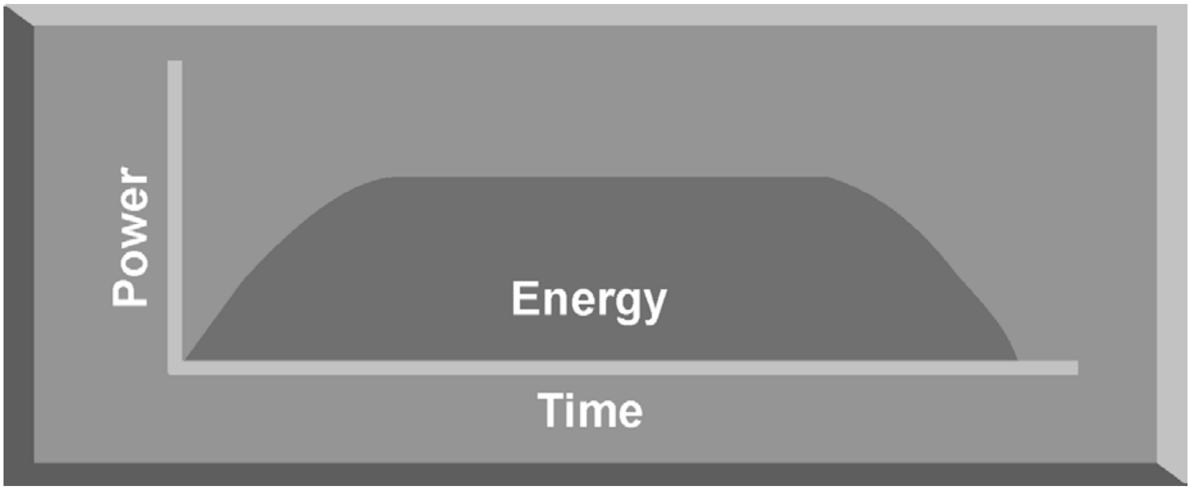

Figure 3-7. Radar Power and Energy

9. MODULATION

The characteristics of an RF signal must be changed in order to transmit information on the signal. This process is called modulation. Modulation is accomplished by combining a basic RF signal, called a carrier wave, with a modulating signal that contains the desired information. The resulting waveform is then used to transmit the desired information.

a. One basic modulation technique is amplitude modulation (AM). The carrier wave is combined with a modulating signal containing information of varying amplitude. Waveforms produced have the same frequency as the carrier wave but with a varying amplitude based on the information from the modulating signal. AM is used extensively in communications and broadcast radio transmissions (Figure 3-8).

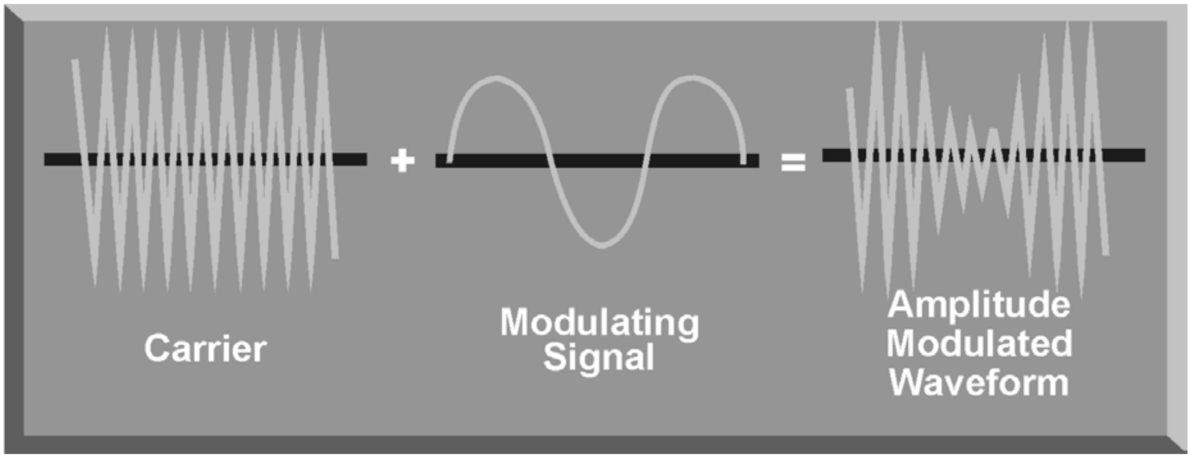

Figure 3-8. Amplitude Modulation

b. Frequency modulation (FM) is another means of impressing information on a carrier wave. Frequency modulation is accomplished by combining the carrier wave with a modulating signal containing information of varying frequency. The waveform produced has the same amplitude as the carrier wave, but the frequency varies based on the information from the modulating signal (Figure 3-9). FM is used extensively in communications and commercial radio. FM is also used with continuous wave (CW) radars to make them more resistant to jamming and to add range determination capability.

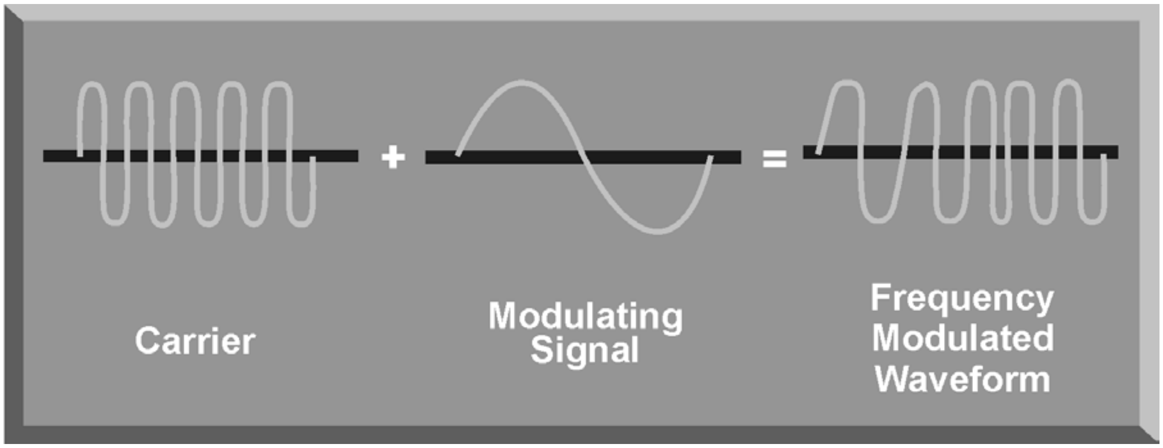

Figure 3-9. Frequency Modulation

c. A type of amplitude modulation known as pulse modulation (PM) is used in pulse radars to produce the short, powerful bursts of RF energy. PM combines the carrier wave with a rectangular pulse that acts like a switch. PM turns the transmitter on, leaves it on for a predetermined time, and then turns it off. The result is a waveform that produces radar pulses that can be used to measure range and angle to the target (Figure 3-10).

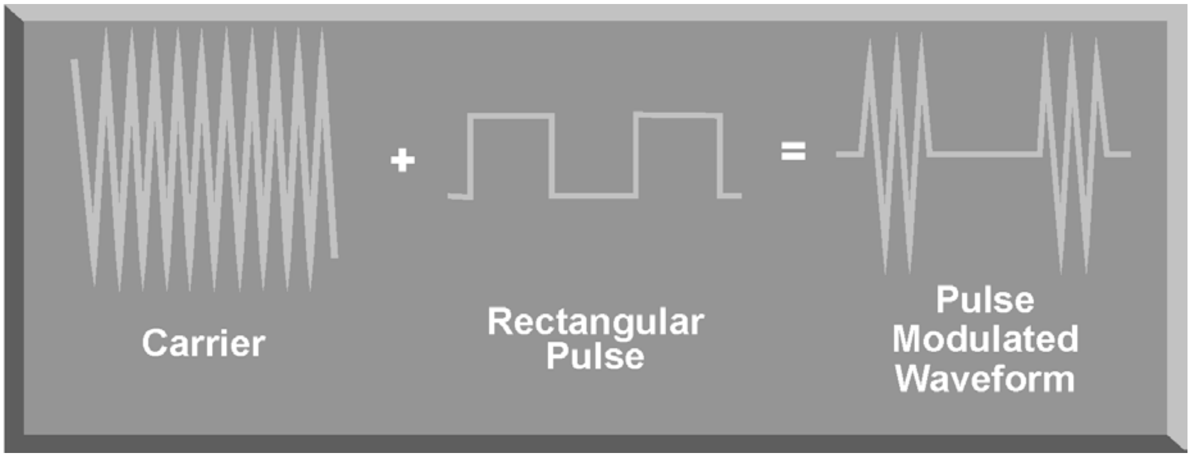

Figure 3-10. Pulse Modulation

10. SUMMARY

The radar signal characteristics of PW, PRI, PRF, and power determine the maximum range and the range resolution capability of a specific radar. When combined with the frequency of the carrier wave of the radar signal, these parameters provide a unique signature to identify a specific radar signal. Modulation is the method used to put information on an RF carrier wave. The primary modulation techniques used in radar signal generation include amplitude, frequency, and pulse modulation. The radar signal characteristics of PRF, PRI, power, and modulation are the keys to understanding radar operation and jamming techniques.

CHAPTER 4. RADAR SYSTEM COMPONENTS

1. INTRODUCTION

The individual components of a radar determine the capabilities and limitations of a particular radar system. The characteristics of these components also determine the countermeasures that will be effective against a specific radar system. This chapter will discuss the components of a basic pulse radar, a continuous wave (CW) radar, a pulse Doppler radar, and a monopulse radar.

2. PULSE RADAR SYSTEM

The most common type of radar design is the pulse radar system. The name describes a process of transmitting discrete bursts of RF energy at the frequency of the radar system. The time that pulses are transmitted determines the pulse repetition frequency (PRF) of the radar system. A pulse radar system can figure out range and azimuth. Range is determined by the time that it takes a pulse to go to a target and return. Target azimuth is determined by the relative position, or antenna orientation, when the pulse strikes the target. Figure 4-1 is a basic block diagram of a simple pulse radar system.

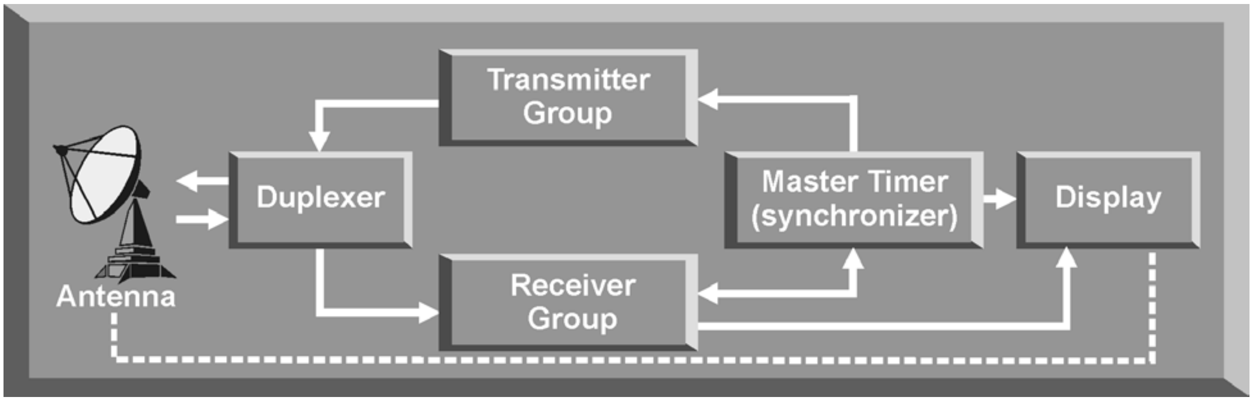

Figure 4-1. Pulse Radar Block Diagram

a. The purpose of the transmitter is to deliver a series of high-energy bursts of radio frequency (RF) energy to the antenna. The transmitter group of a modern pulse radar normally consists of a pulse generator or waveform generator, a modulator, and some kind of power amplifier (Figure 4-2).

(1) The purpose of the waveform generator is to generate the proper waveform or pulse, normally at a low power level, before delivery to the modulator. It is much easier to generate complex waveforms at a lower power level. These complex waveforms are required for coherent systems employing digital moving target indicator techniques and for pulse Doppler radar operations.

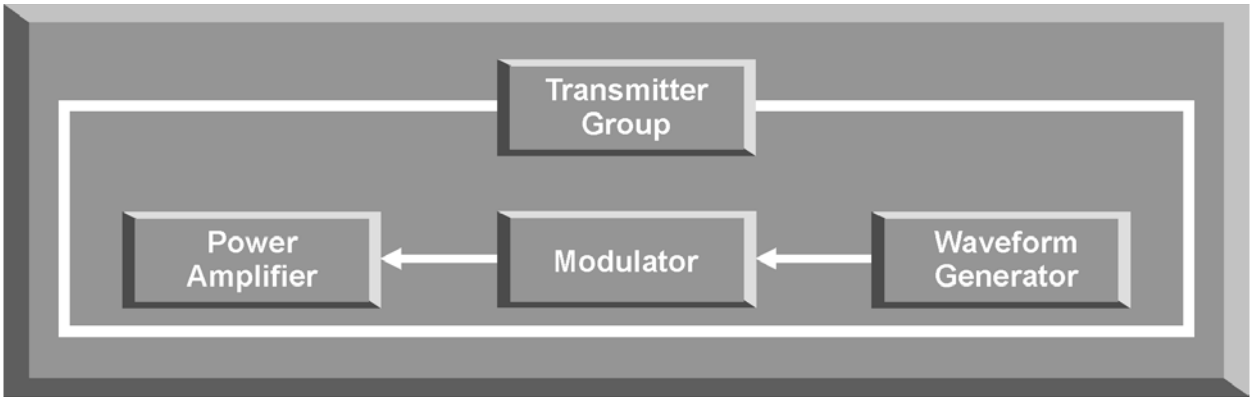

Figure 4-2. Transmitter Group

(2) The modulator is a major portion of the transmitter. The modulator provides an extremely powerful, very short pulse of direct current (DC) voltage to the power amplifier. This is similar to the ignition system of an automobile but with very stringent requirements. The modulator has an energy storage device and a switch. Between pulses, during the resting time of the transmitter, energy is accumulated and stored in the storage device. When keyed by the master timer, all this energy is switched to the power amplifier as a pulse. The waveform of this pulse is determined by the waveform generator.

(3) The power amplifier for a modern radar is normally a klystron, traveling wave tube, cross field amplifier, or solid state amplifier. Most common pulse radars use a klystron power amplifier. No matter what power amplifier is used, the purpose of the transmitter group is to produce a series of pulses at the correct amplitude, at the proper interval, with the exact waveform, and at the operating frequency of the radar.

b. A duplexer is required when both the transmitter and receiver use the same antenna. The duplexer acts as a rapid switch to protect the sensitive receiver from damage when the high-power transmitter is on. When the transmitter is off, the duplexer directs the weak target signals to the receiver. The duplexer's main purpose is to minimize power loss and maximize isolation. Power lost in the duplexer during transmission reduces the maximum detection range of the radar. Isolation refers to the amount of transmitter power that "bleeds through" the duplexer to the receiver during transmission. This "bleed through" must be extremely small to avoid receiver saturation or damage.

c. The capabilities of the receiver group are critical to radar performance (Figure 4-3). The ability of the radar receiver to detect the presence of the target return and extract the required information is limited primarily by noise. Noise can enter the receiver through the antenna along with the target return. This type of noise is called external noise. Noise generated within the receiver is called thermal noise. Radar noise can never be completely eliminated. Minimizing noise

is the most important consideration in the design of the sensitive receivers used with modern radars. In addition, relative immunity to noise makes a radar system more resistant to jamming.

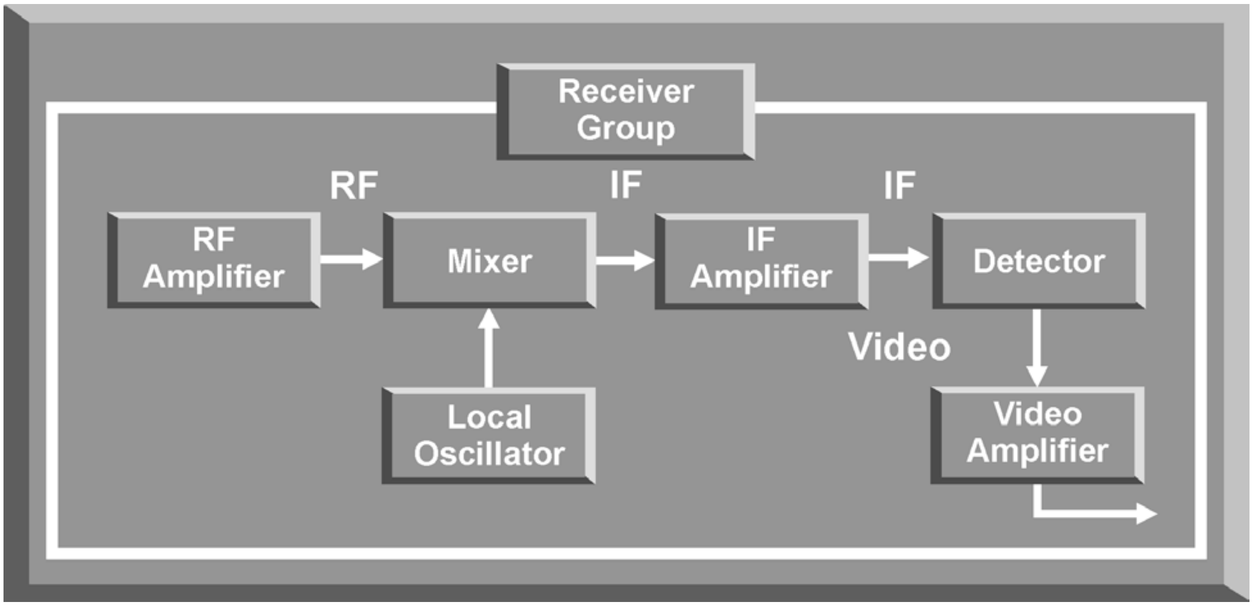

Figure 4-3. Receiver Group

(1) The most common pulse radar receiver is the superheterodyne receiver. A superheterodyne receiver consists of an RF amplifier, a mixer and local intermediate frequency (IF) amplifier, a detector, and a video amplifier.

(2) Radar target returns enter the receiver group via the antenna and duplexer. Since these signals are normally very low power, the RF amplifier boosts the signal gain and filters out as much external noise as possible. The capability of the RF amplifier to minimize noise determines the receiver sensitivity. The boosted RF signal is sent to the mixer where it is converted to a lower IF. This is accomplished by mixing the RF signal with the signal from the local oscillator to produce an IF that is easier to process. The IF amplifier increases the IF signal level and includes a matched filter. The matched filter maximizes the signal-to-noise ratio which enhances detection of the target return. The detector, which is usually a crystal diode, extracts the video modulation from the IF or converts the IF to a video signal.

d. The brain of a basic pulse radar is the master timer, or synchronizer, which coordinates the operation of the various parts of the radar (Figure 4-4). Exact timing within the radar is necessary to get accurate range. The master timer is an oscillator that triggers the transmitter to initiate transmission of a pulse. Simultaneously, the master timer sends a signal to initialize the display to ensure that range and azimuth information is accurately displayed.

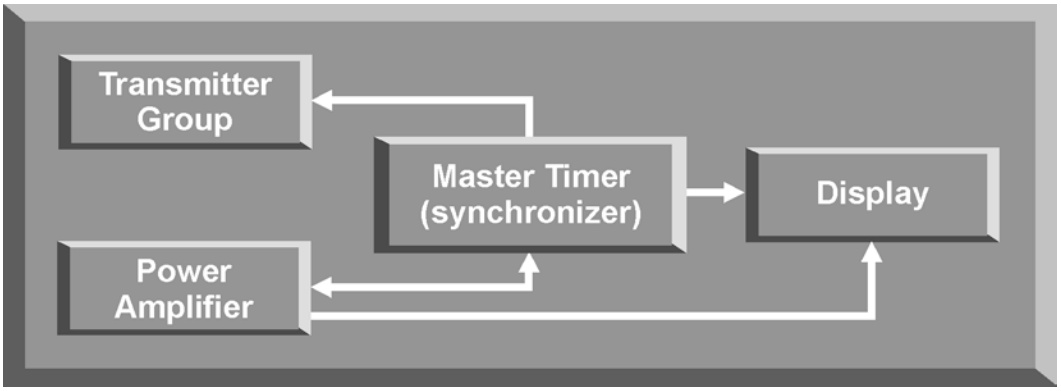

Figure 4-4. Master Timer/Synchronizer

e. The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction (Figure 4-5). During reception, or listening time, the function of the antenna is to collect the returning radar energy contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern.

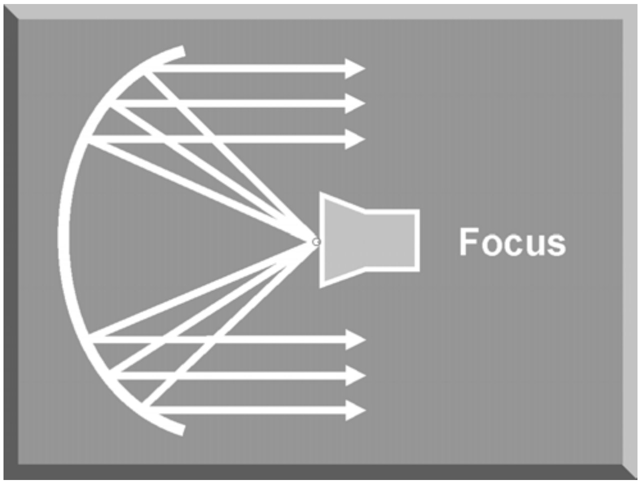

Figure 4-5. Parabolic Antenna

f. The purpose of the radar display is to take the information derived from a radar target in the receiver group and present it to the operator in a usable format. There are many different types of scope displays depending on the purpose of the radar and how the radar information is to be used. There are four basic types of radar displays: the A scope, B scope, range height indicator (RHI) scope, and plan position indicator (PPI) scope.

(1) The A scope is used to display target range or velocity (Figure 4-6). Threat systems using A scope displays include air interceptors (AIs) with range-only radar, surface-to-air missiles (SAMs), and radar-directed antiaircraft artillery (AAA) systems. SAM and AAA systems may use the A scope for range or velocity information, and other radar displays for azimuth and elevation data. The A scope displays range or velocity in relation to amplitude. The operator must distinguish the target return from other returns, including ground return and noise.

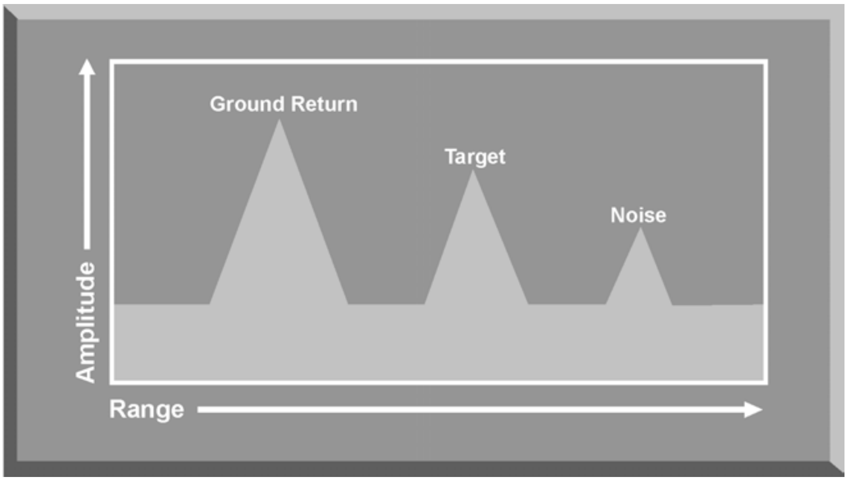

Figure 4-6. A Scope Display

(2) The B scope is used to display target range and azimuth (Figure 4-7). Threat systems using B scope displays include AI and SAM systems. The position of the target return to the right or left of the centerline of the screen shows the azimuth of the target. The position of the target return in relation to the bottom of the display, or zero range, shows target range.

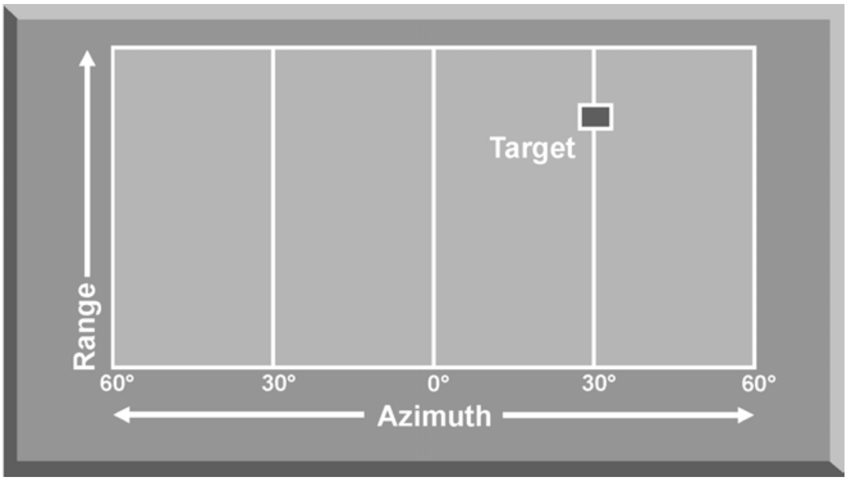

Figure 4-7. B Scope Display

(3) The RHI scope is used to display range and elevation (Figure 4-8). The RHI scope is used with height finder radars, and a modified RHI scope is used for ground-controlled approach (GCA) radars. The sweep trace of the display produces a fan-shaped display with the vertex at the lower left of the scope. The antenna sweeps up and down and is synchronized with the display.

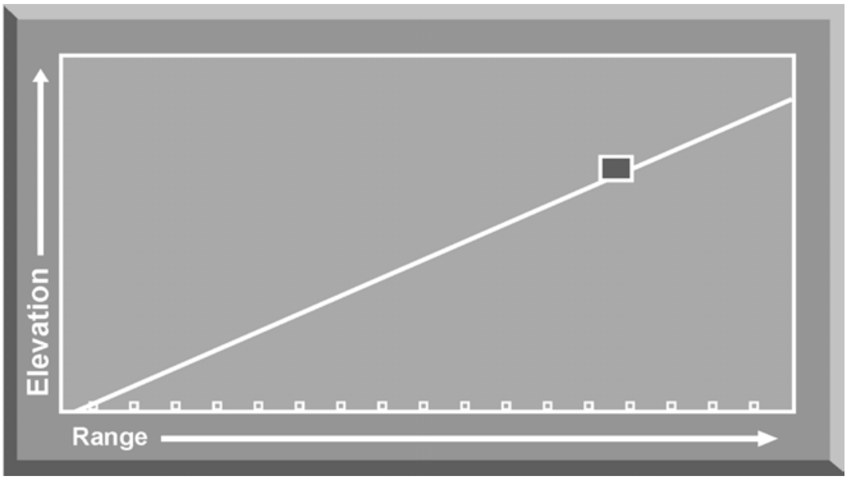

Figure 4-8. RHI Scope Display

(4) The PPI display is probably the best known radar display (Figure 4-9). The display represents a map picture of the area scanned by the radar beam, usually 360 degrees. The PPI display is used by early warning, acquisition, ground-controlled intercept (GCI), and SAM radar systems. The target return's angular position shows target azimuth, while distance from the center of the display shows range.

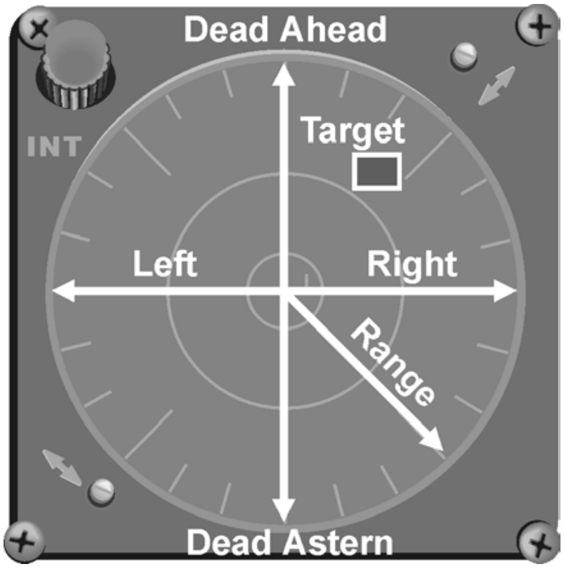

Figure 4-9. PPI Scope Display

3. CONTINUOUS WAVE (CW) RADAR

A continuous radar transmission from the antenna requires that classic CW radars have two antennas, one for transmission and one for reception (Figure 4-10). Since a continuous transmission results in a continuous echo signal, it is impossible to tell what part of the echo is associated with any particular part of the transmission. This makes conventional range determination (based on timing) impossible. However, the simple application of the Doppler principle provides a means for a CW radar to track a target. The Doppler principle deals with the fact that a radar return from a moving target will be shifted in frequency by an amount proportional to its radial velocity relative to the radar site. Using the difference in frequency from the transmitted signal to the received signal, a CW radar can separate the target return from clutter, based on a change in frequency. This type of radar is called a CW Doppler radar.

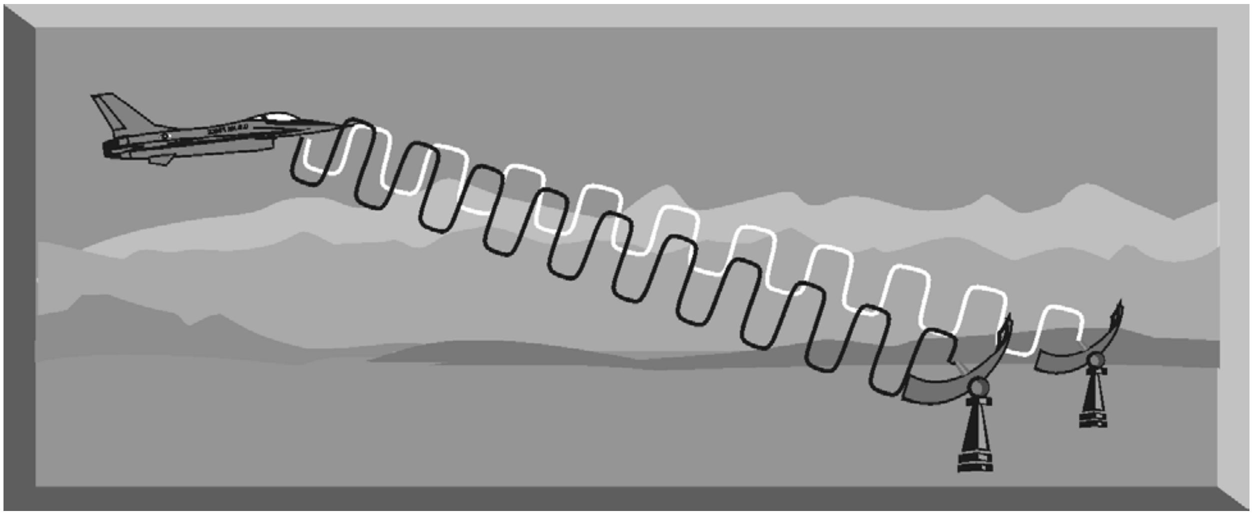

Figure 4-10. Continuous Wave Radar

a. Figure 4-11 depicts a basic CW Doppler radar. In a simple CW Doppler radar, the transmitter transmits a continuous signal at the radar's operating frequency. This signal is reflected by a moving target and travels back to the receiving antenna. The frequency of the reflected signal (f_d) is the frequency change due to the Doppler effect. This target frequency is passed to the detector. The transmitted frequency (f_o) is also fed to the detector as a reference. The detector notes the difference between the transmitted and received frequencies and passes this frequency to the Doppler filters. The Doppler filters only allow Doppler frequencies within a certain range to pass through. A filter is required for each Doppler frequency. The number of Doppler filters determines the number of targets that the radar can resolve in velocity.

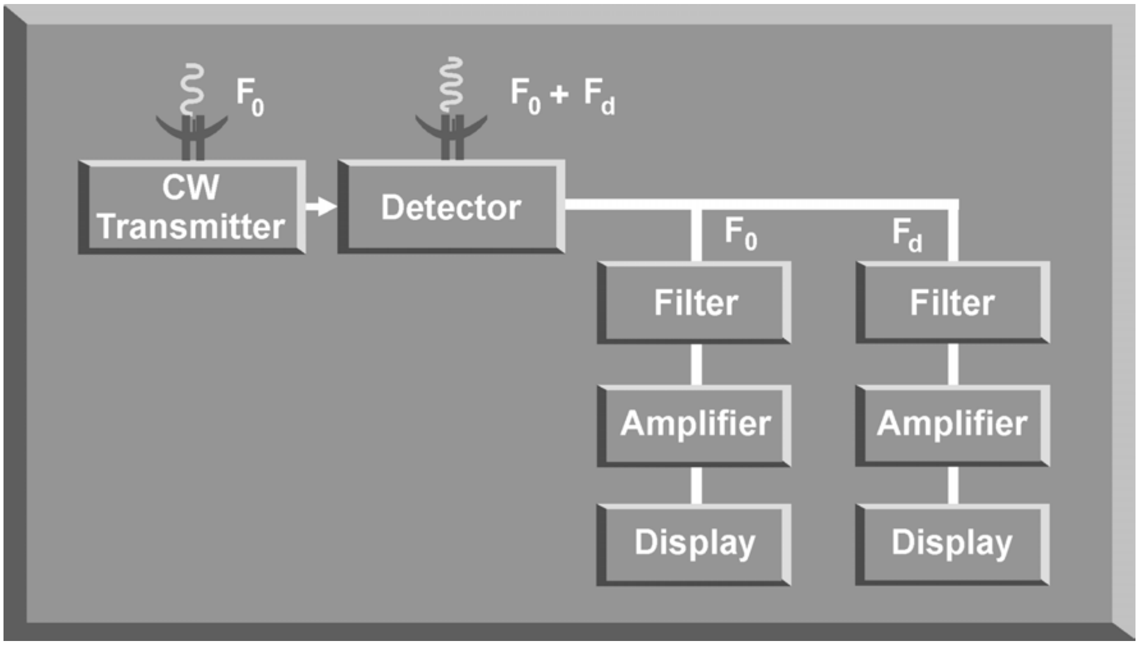
F_0
$F_0 + F_d$
CW Transmitter
Detector
F_0
F_d
Filter
Filter
Amplifier
Amplifier
Display
Display

Figure 4-11. Basic CW Doppler Radar

b. The output of each Doppler filter is amplified and passed to its own display. The display is normally an A scope as shown in Figure 4-12.

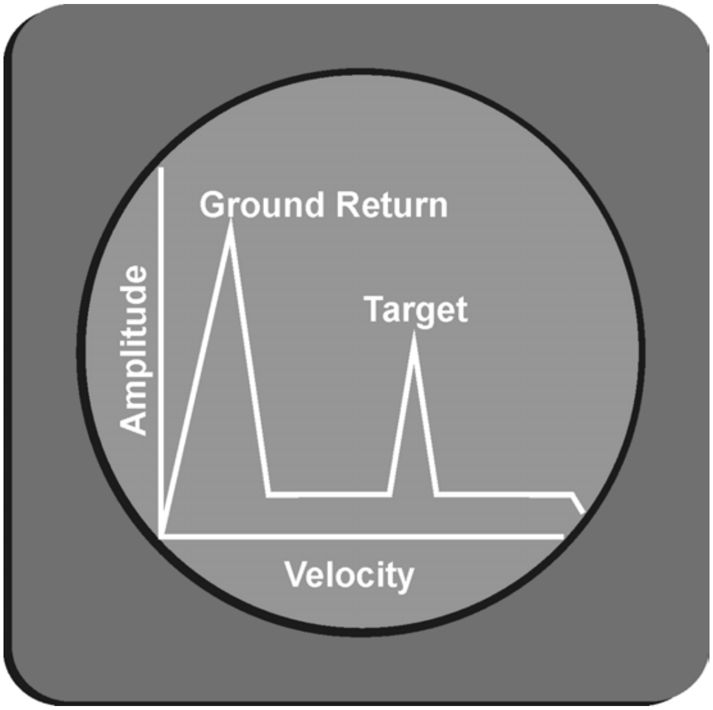
Ground Return
Target
Amplitude
Velocity

Figure 4-12. CW Doppler Radar Display

4. PULSE DOPPLER RADAR

Pulse Doppler radars combine the advantages of both pulse and Doppler radar systems (Figure 4-13). Because the signal is pulsed, the radar can determine range, azimuth, and elevation, similar to a conventional pulsed radar. A pulse Doppler radar can also compute overtake, or rate of closure, relative to the radar system on a pulse-to-pulse basis. Pulse Doppler radars also use multiple PRFs to eliminate target eclipsing and for range determination in medium PRF. The beauty of a pulse Doppler radar is that it eliminates ground clutter and provides range, azimuth, and velocity resolution.

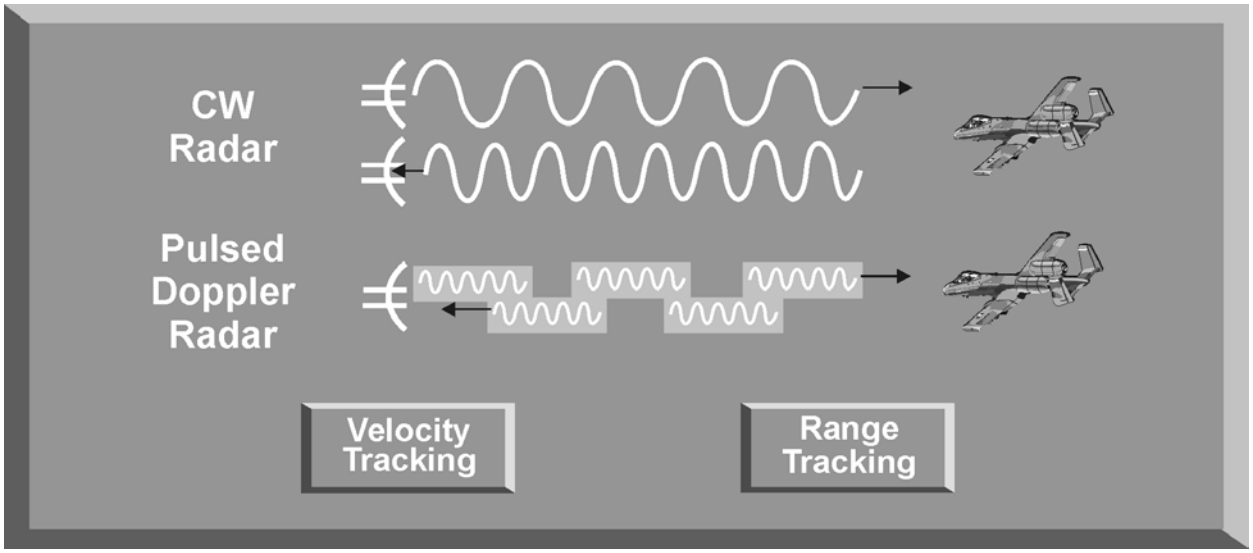

Figure 4-13. CW and Pulse Doppler Radar Comparison

a. A pulse Doppler radar transmits a box, or pulse, of coherent RF energy at the operating frequency of the radar. The frequency inside these boxes reacts the same way as the continuous waves of a CW radar. However, since the RF waves are pulsed, range determination can be accomplished by measuring the time it takes for the reflected pulse to return from the target. Velocity determination and tracking are accomplished by capturing and quantifying the Doppler shift of the frequencies in each box or pulse.

b. The basic block diagram of a coherent pulse Doppler radar (Figure 4-14) is similar to a pulse radar except for the addition of an exciter, a radar computer, and a digital signal processor. The exciter generates a continuous stable low power signal at the desired frequency and phase for the transmitter. It also sends this signal as a reference to the receiver. The digital signal processor performs the adding and subtracting functions required to find, track, and sort targets with respect to velocity and range. The radar computer performs all routine functions of the radar such as changing modes and accounting for aircraft flight parameters.

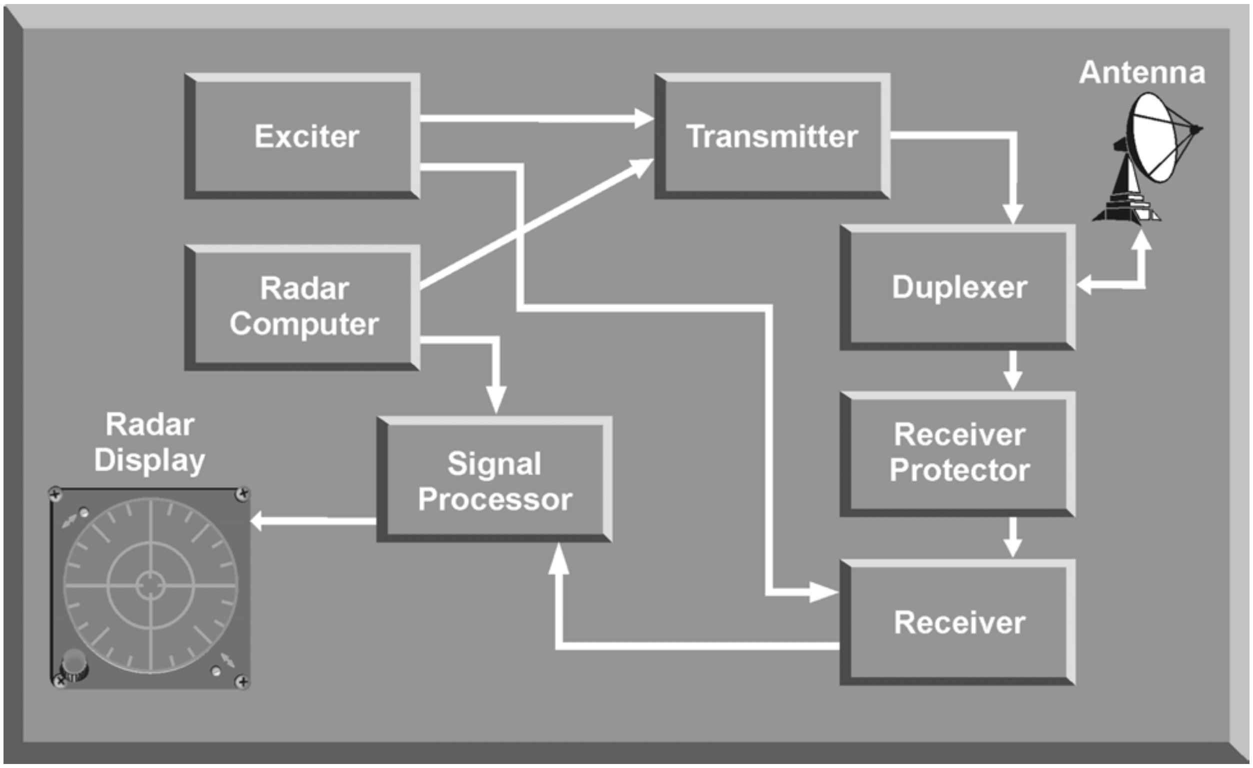

Figure 4-14. Basic Pulse Doppler Radar Diagram

5. MONOPULSE RADAR

The primary functions of monopulse radars include target tracking and weapon guidance. Monopulse radars were developed to overcome the limitations and jamming susceptibility of scanning radar systems. A monopulse radar typically receives returned radar energy in two, three, or four separate receivers, or channels, each looking at a different area. By comparing the returns from each receiver, track errors can be determined. For example, in Figure 4-15, this radar compares the amplitude, or signal strength, in each channel. Obviously, an aircraft in area "A" will produce a stronger return in receiver "A" since that receiver is focused in that direction. Since this type of system uses signal strength to determine position, typical noise jamming will only highlight an aircraft" position. Other coherent radars can measure phase differences verses the amplitude differences used in the last example. If a signal is directly ahead of the antenna, each quadrant will receive the returned signal at exactly the same time and therefore the same exact phase. However, if a return is off-center, the returned signal will strike part of the antenna slightly ahead of the opposite side and each quadrant will measure a different phase. This phase difference is used to determine azimuth error.

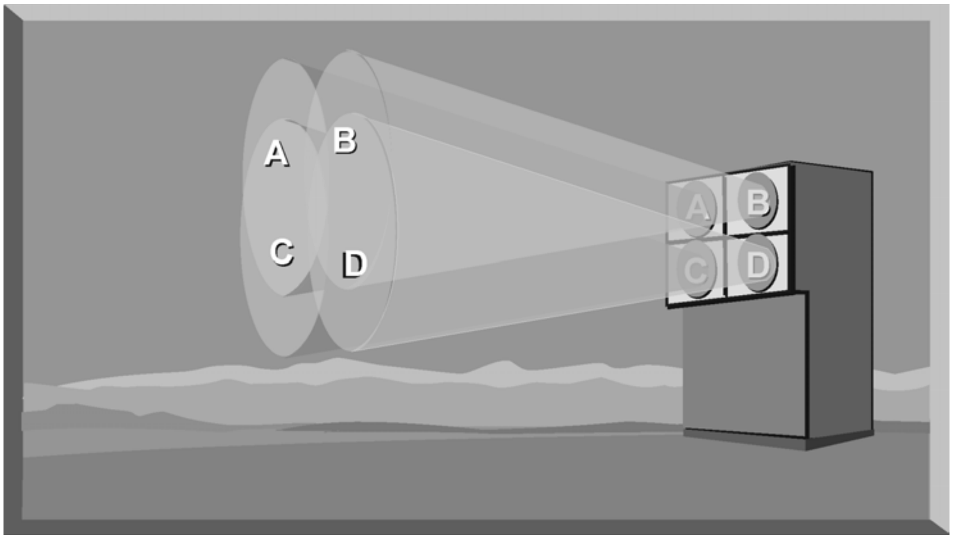

Figure 4-15. Monopulse Radar

a. Monopulse radars operate by comparing the amplitude or the phase of the received signal in each of the transmitted beams. A complex set of comparator circuits, called magic T's, do the addition and subtraction of the received signals based on the orientation of the magnetic field, or H-plane, and the electrostatic field, or E-plane (Figure 4-16). The received signal from antenna A enters the magic T in the H-plane arm while the signal from antenna B enters the magic T in the E-plane arm.

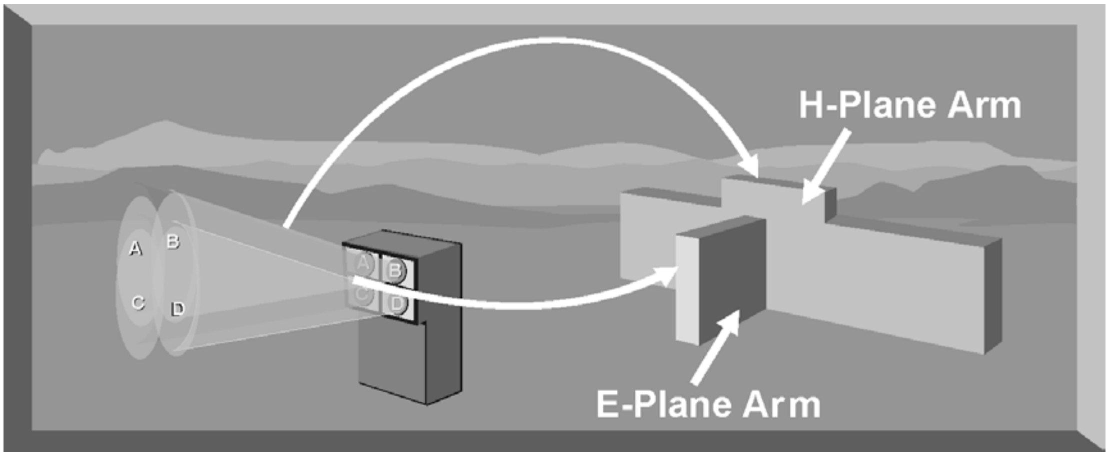

Figure 4-16. Monopulse Magic T’s

b. Figure 4-17 depicts the output of a magic T. Output is the sum and difference of the two input signals. These sum and difference values, in amplitude or phase, are used to generate azimuth and elevation error signals and to compute range.

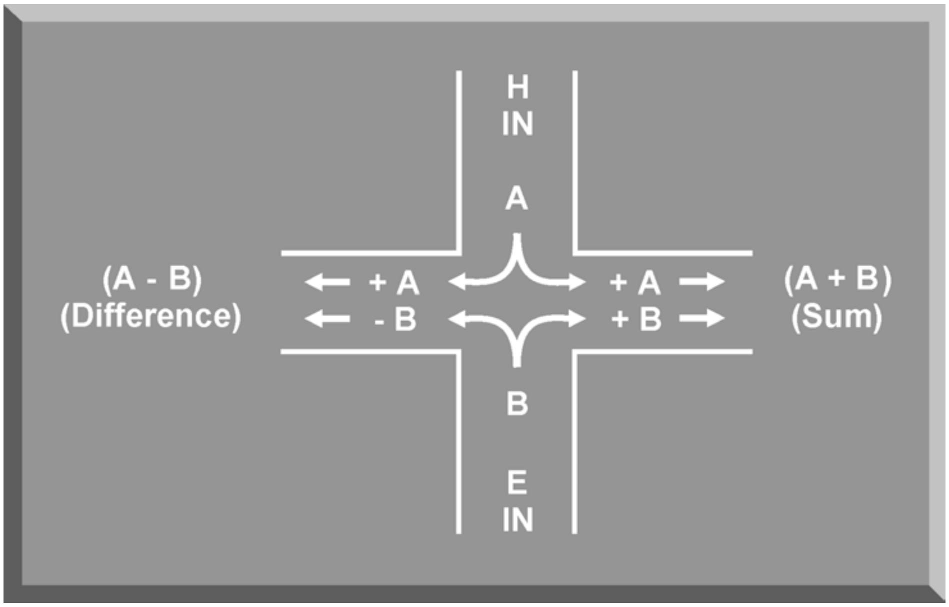
H
IN
A
(A - B)
(Difference)
+ A
- B
+ A
+ B
(A + B)
(Sum)
B
E
IN

Figure 4-17. Magic T Output Signals

c. These sum and differences computed by the magic T's are sent as signals to the monopulse tracking servos to generate both azimuth and elevation corrections to keep the monopulse antenna centered on the target (Figure 4-18). Target range is computed by calculating the difference of signal transmission time with the reception time of the target echo. Range information is displayed to the operator for range tracking.

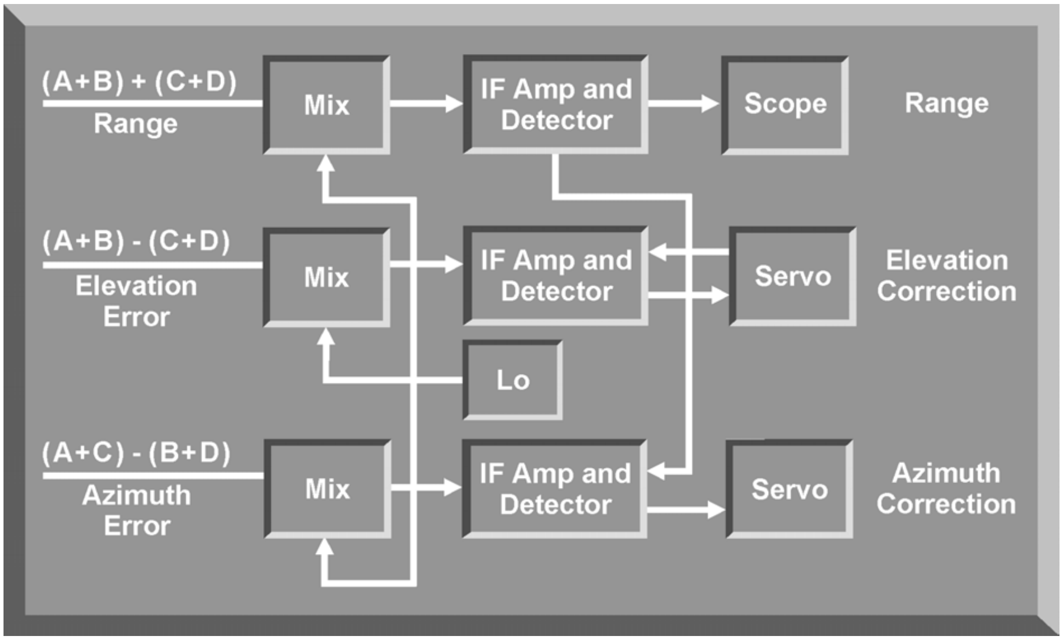
(A+B) + (C+D)
Range
Mix
IF Amp and Detector
Scope
Range
(A+B) - (C+D)
Elevation Error
Mix
IF Amp and Detector
Servo
Elevation Correction
Lo
(A+C) - (B+D)
Azimuth Error
Mix
IF Amp and Detector
Servo
Azimuth Correction

Figure 4-18. Monopulse Radar Track

d. If the signal in antenna A equals the signal in antenna B, equals the signal in antenna C, and equals the signal in antenna D, the target is in the center of the four radar beams (Figure 4-19). This is the condition the monopulse tracker attempts to maintain when tracking a target.

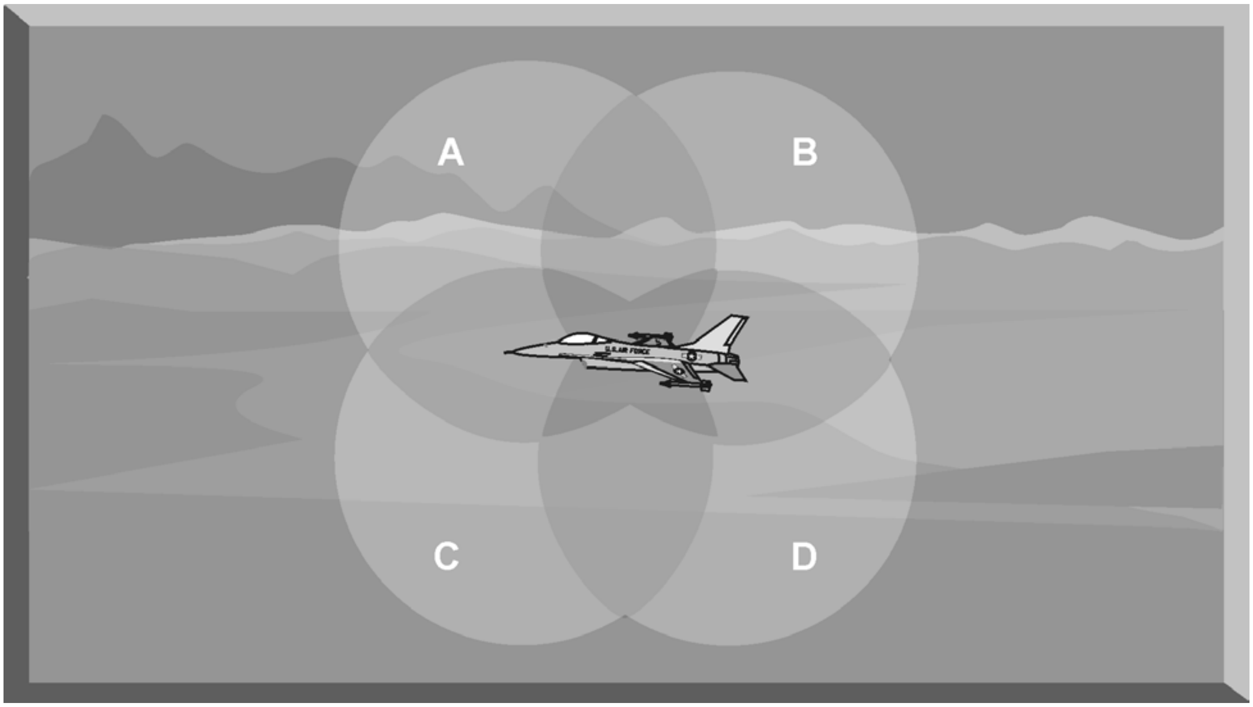

Figure 4-19. Monopulse Radar Track Logic

6. SUMMARY

This chapter has discussed the basic components of a simple pulse radar, a continuous wave radar, a pulse Doppler radar, and a monopulse radar. The characteristics of these components determine the capabilities and limitations of a particular radar system.

CHAPTER 5. RADAR PRINCIPLES

1. INTRODUCTION

The primary purpose of radar systems is to determine the range, azimuth, elevation, or velocity of a target. The ability of a radar system to determine and resolve these important target parameters depends on the characteristics of the transmitted radar signal. This chapter explains the relationship of radar frequency (RF), pulse repetition frequency (PRF), pulse width (PW), and beamwidth to target detection and resolution.

2. RADAR RANGE

A basic pulse radar system consists of four fundamental elements: the transmitter, the receiver, the antenna, and the synchronizer, or master timer.

a. The transmitter, through the antenna, sends out a pulse of RF energy at a designated frequency. The presence of a target is revealed when the RF energy bounces off the target, returns to the radar antenna, and goes into the receiver (Figure 5-1). The master timer measures the time between the transmission of a pulse and the arrival of a target echo.

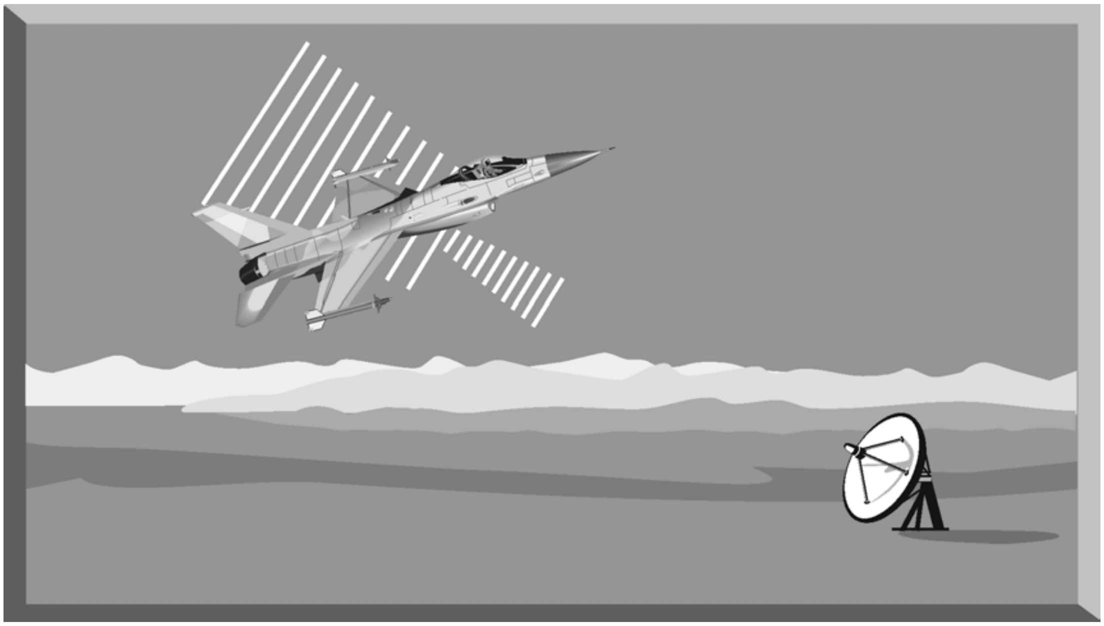

Figure 5-1. Basic Radar Pulse

(1) RF energy travels at the speed of light (c) which is 3 x 10^8 meters per second. Target range can be computed by using the basic radar range determination equation (Equation 5-1).

$$\text{Target Range} = \frac{\text{Measured Time} \times \text{Speed of Light (c)}}{2}$$

Equation 5-1. Basic Radar Range Determination Equation

(2) Another useful measurement is the radar mile, which is the round trip time for an RF wave to travel to and from a target one nautical mile away (Figure 5-2). Solving the radar range equation for time results in Equation 5-2. Substituting the appropriate values into the equation and solving for time gives measured time of 12.4 microseconds for a one nautical mile (1853 meters) round trip.

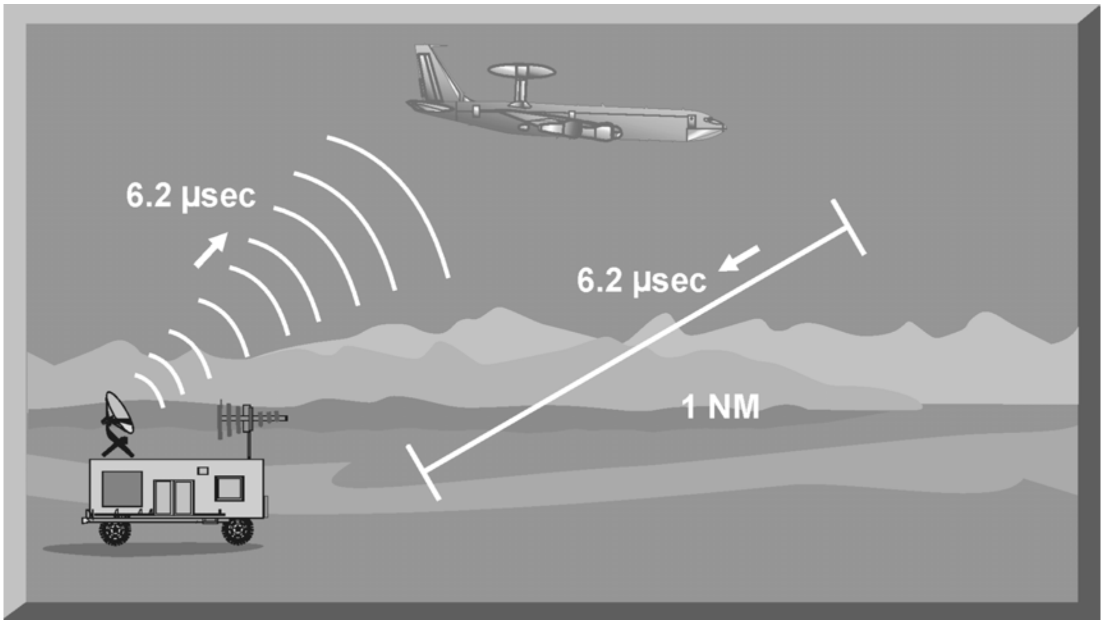

Figure 5-2. Radar Mile

$$\text{Measured Time} = \frac{\text{Target Range} \times 2}{\text{Speed of Light (c)}} = \frac{1853 \text{ meters} \times 2}{3 \times 10^{8} \text{ meters/second}} = 12.4\,\mu\text{sec}$$

Equation 5-2. Radar Mile

(3) A limitation on radar detection range is the concept of a second time around echo. A second time around echo occurs when a target echo associated with a particular radar pulse arrives at the antenna after another radar pulse has been transmitted. The radar master timer always assumes the target echo is associated with the last pulse transmitted. This makes the target echo ambiguous in range. Figure 5-3 depicts a radar signal with a pulse recurrence time (PRT) of 248 microseconds. Radar pulse A takes 372 microseconds to travel to the target and return. Using the range determination equation, actual target range is 30 nautical miles (nm). However, before the target echo returns to the antenna, radar pulse B is transmitted. The master timer associates the target echo of pulse A with radar pulse B, and calculates a target range of 10 nm. This ambiguous and false range is displayed to the operator.

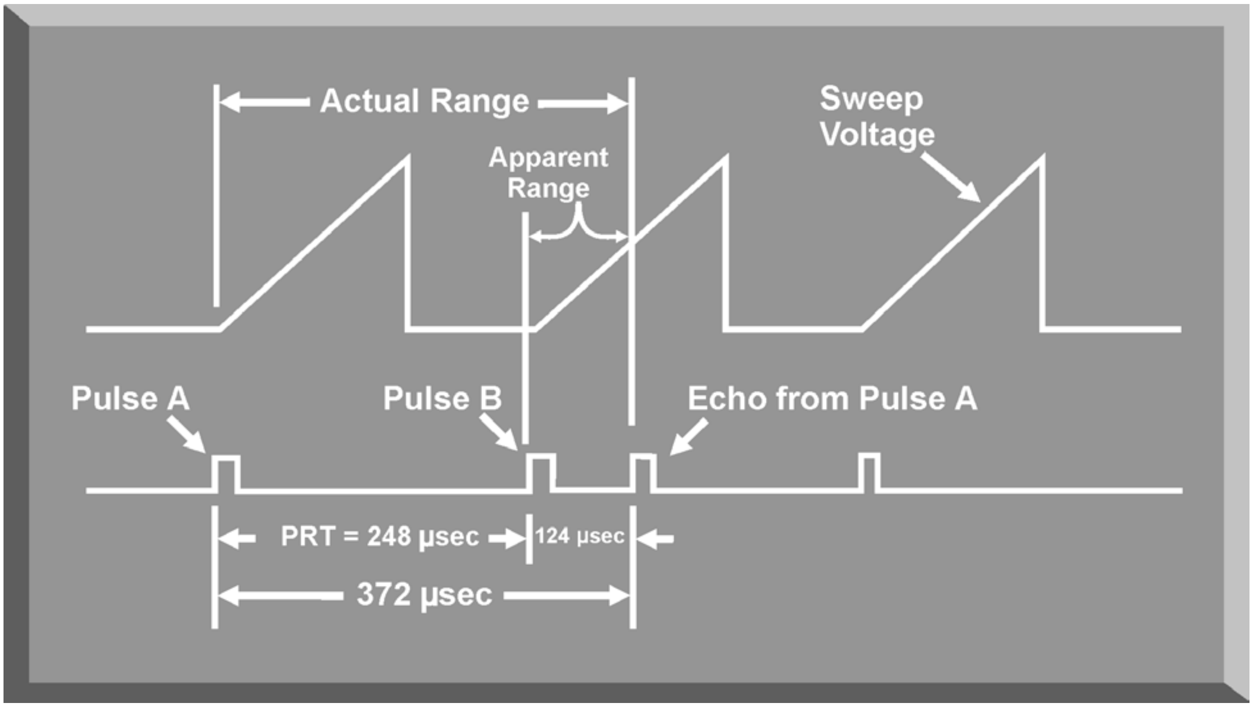

Figure 5-3. Second Time Around Echo

(4) Range ambiguities caused by second time around echoes limit the maximum unambiguous range of a radar system. This important capability can be calculated by using Equation 5-3. An analysis of this equation shows that a radar system designed for long-range detection should transmit a radar signal with a large PRT. In addition, as the PRF of a radar signal increases, the PRT decreases, and the maximum unambiguous range decreases.

$$\text{Maximum Unambiguous Range} = \frac{\text{PRT } (\mu\text{secs})}{12.4\ (\mu\text{secs/nm})}$$

Equation 5-3. Maximum Unambiguous Range

b. A critical aspect of range determination is range resolution. Range resolution is the ability of a radar to separate two targets that are close together in range and are at approximately the same azimuth (Figure 5-4). The range resolution capability is determined by pulse width. Pulse width is the time that the radar is transmitting RF energy. Pulse width is measured in microseconds.

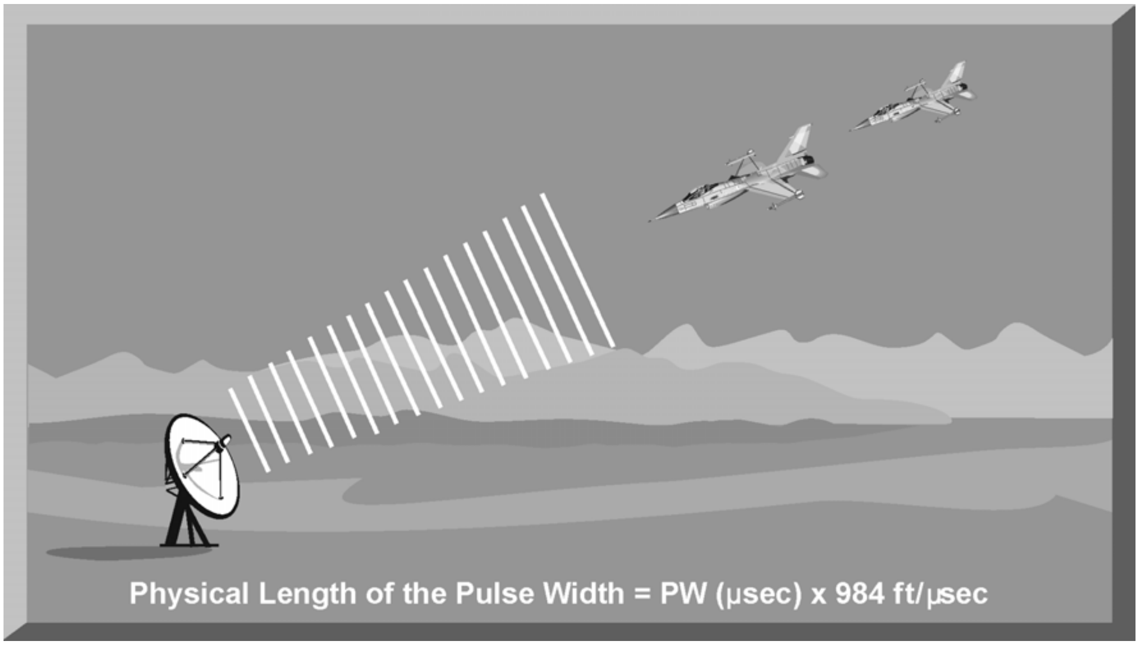
Physical Length of the Pulse Width = PW (µsec) x 984 ft/µsec

Figure 5-4. Radar Pulse

(1) A radar pulse in free space occupies a physical distance equal to the pulse width multiplied by the speed of light, which is about 984 feet per microsecond. If two targets are closer together than one-half of this physical distance, the radar cannot resolve the returns in range, and only one target will be displayed.

(2) To illustrate range resolution, consider Figure 5-5, in which two aircraft are separated by a distance of one-half the pulse width or less. At T1, the leading edge of the radar pulse hits the lead aircraft.

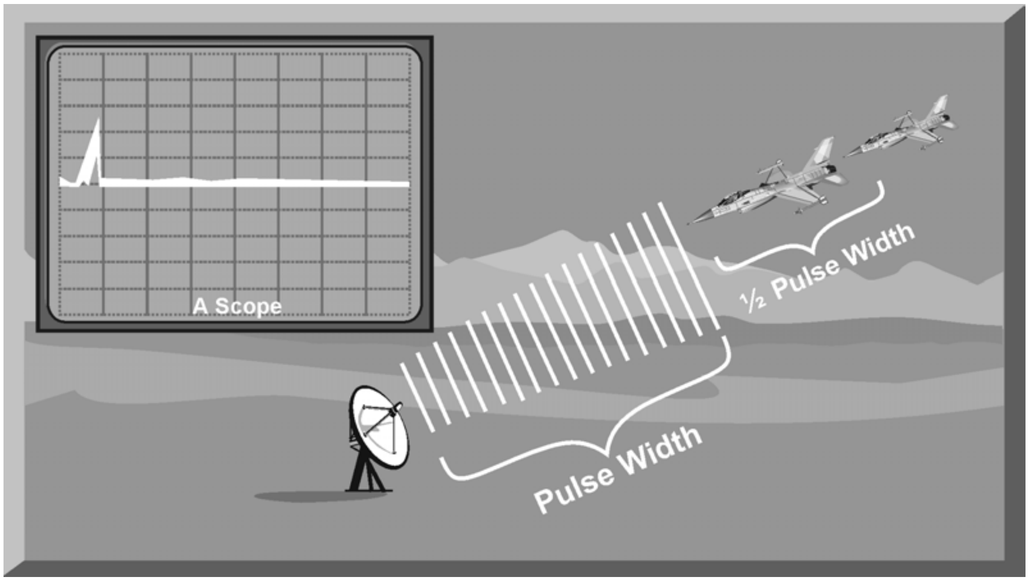

Figure 5-5. Radar Pulse at T1

(3) In Figure 5-6, at T2 the leading edge of this same pulse hits the trailing aircraft. Since the trailing aircraft is less than one-half the pulse width from the lead aircraft, the return echo from the lead aircraft is received by the antenna before the entire pulse has left the trailing aircraft.

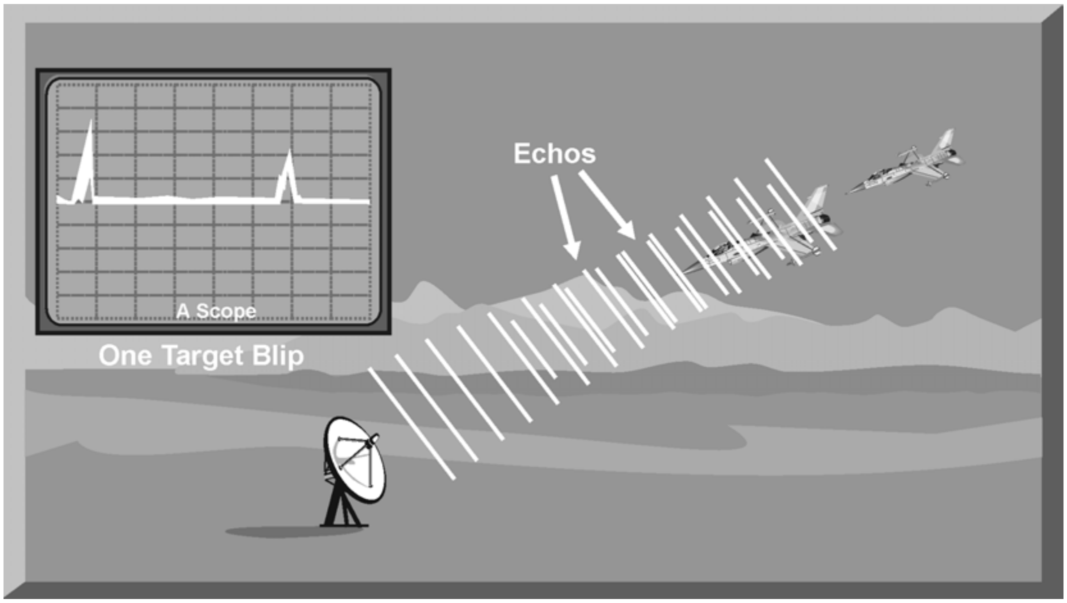

Figure 5-6. Radar Pulse at T2

(4) T3 depicts the merging target echoes at the radar (Figure 5-7). The radar would display only one target in this situation, such as the one shown on the radar scope in Figure 5-6.

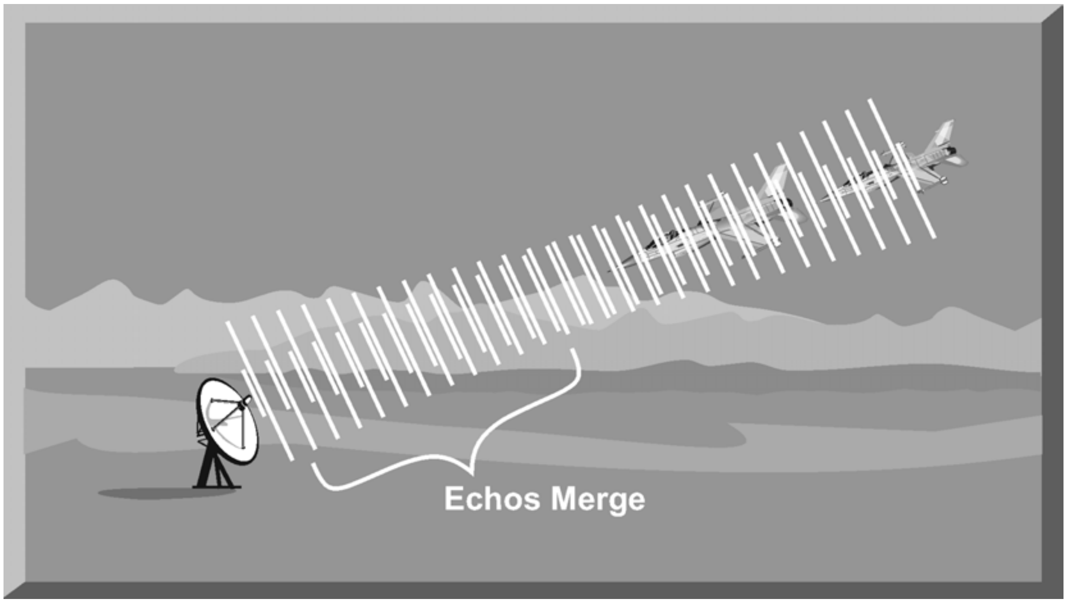

Figure 5-7. Radar Pulse at T3

(5) Two targets separated by more than one-half the pulse width, as in Figure 5-8, will be displayed as two targets. In this case, the transmitted pulse is completely past the lead aircraft before the return echo from the trailing aircraft reaches the antenna.

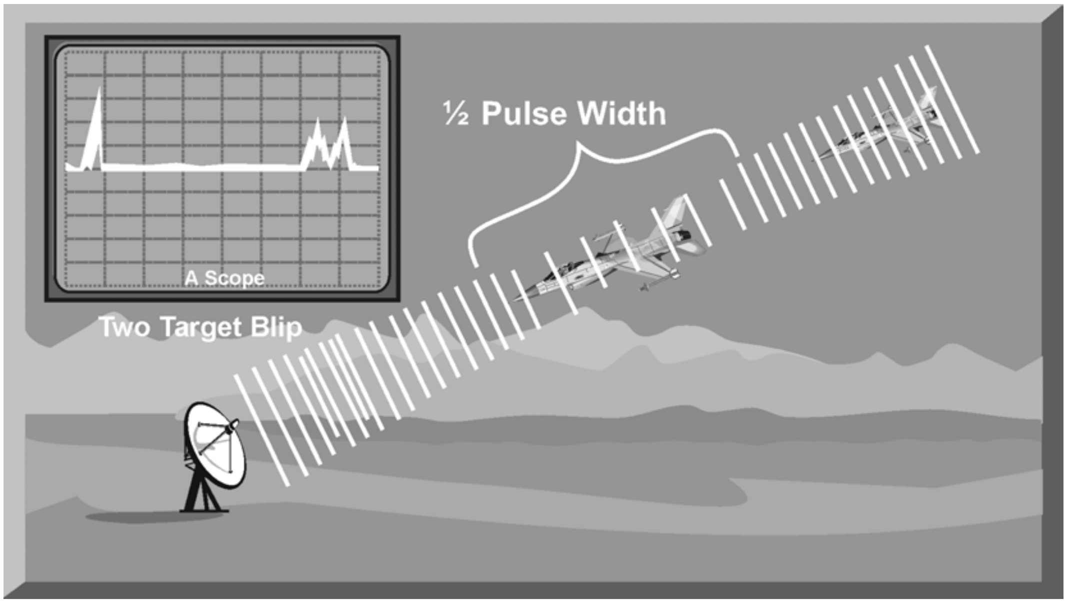

Figure 5-8. Radar Range Resolution

(6) The range resolution of the radar is usually expressed in feet and can be computed using Equation 5-4. It is the minimum separation required between two targets in order for the radar to display them separately on the radar scope.

$$\text{Range Resolution} = \frac{\text{Pulse Width} \times 984\ \text{feet}}{2}$$

Equation 5-4. Range Resolution

3. AZIMUTH DETERMINATION

The beamwidth of a radar system is the horizontal and vertical thickness of the radar beam (Figure 5-9). Beamwidth depends on antenna design and is normally measured in degrees from the center of the beam to the point at which the power drops off by half. This half-power point is -3 dB in power drop-off. Beamwidth governs the azimuth and elevation accuracy and resolution capability of a radar system in the same way that pulse width governs radar range accuracy and resolution.

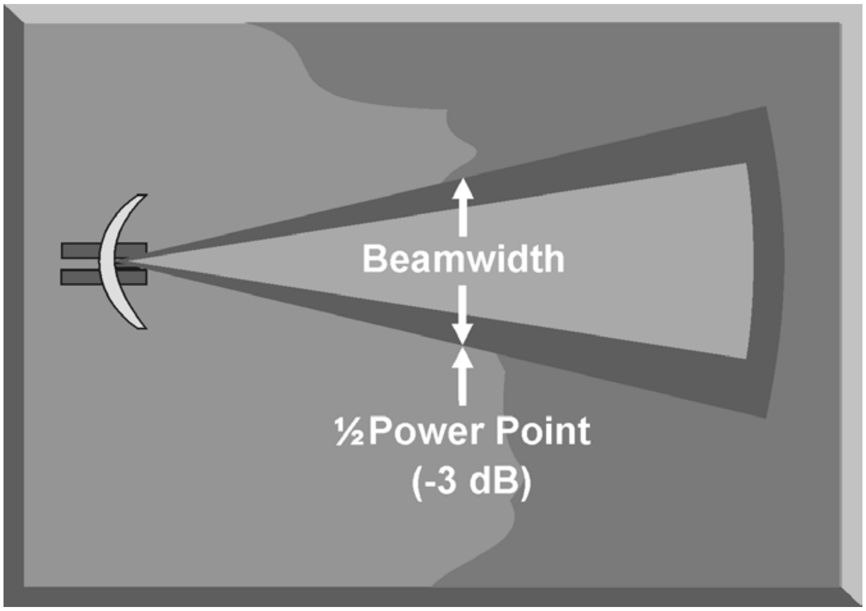

Figure 5-9. Radar Beamwidth

a. In order for a radar system to figure out target azimuth, the antenna must be aligned with a point of reference and pointed at the target during the transmission and reception of several pulses of radar energy. If the antenna is

referenced to true North, the azimuth of the target can be measured relative to true North (Figure 5-10). Azimuth determination is based on the position of the antenna when the target is being illuminated.

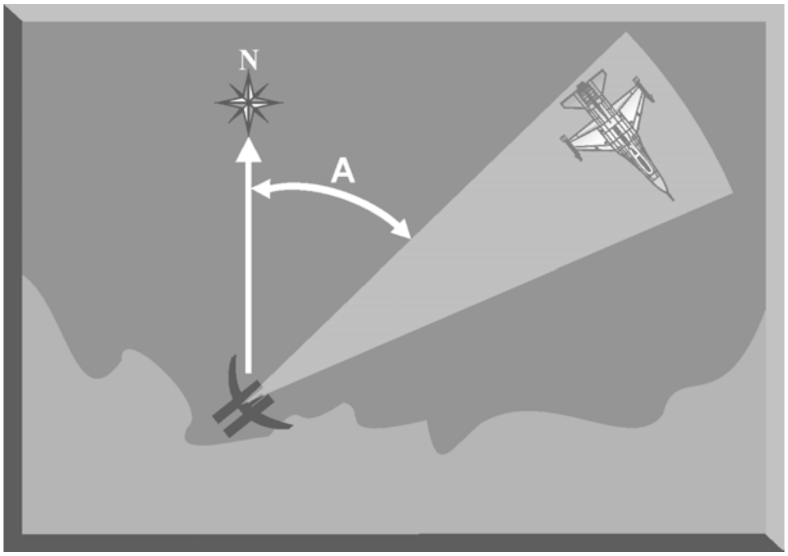

Figure 5-10. Azimuth Determination

(1) To provide accurate azimuth determination over a large area, many radars employ a narrow beam and scan the antenna in a predictable pattern. The most common scan pattern is a 360° circular scan at a constant rate (Figure 5-11). The plan position indicator (PPI) radar scope display is normally associated with this scan pattern. As the radar beam sweeps, a target is detected and displayed. The position of the antenna, when the target is displayed, shows the relative azimuth.

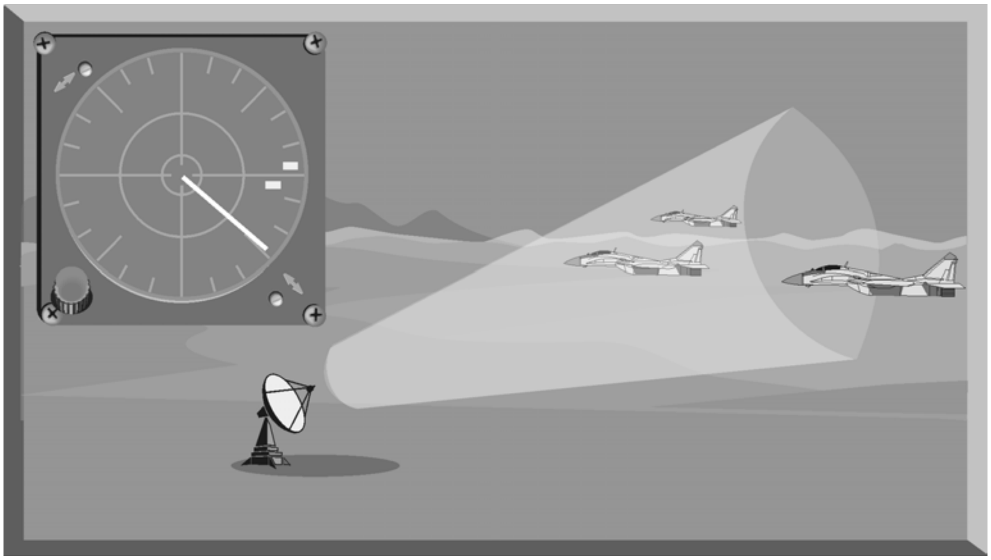

Figure 5-11. Antenna Scan

(2) The azimuth accuracy of a radar system is determined by the horizontal beamwidth (HBW). In Figure 5-12, radar system A has a horizontal beamwidth of 10°. As the beam sweeps, the target is illuminated for as long as it is in the beam. This means that the target covers 10° in azimuth on the PPI scope. Radar system B has a beamwidth of 1°. A target displayed on the PPI scope will cover 1° in azimuth. The narrower the horizontal beamwidth, the better the azimuth accuracy.

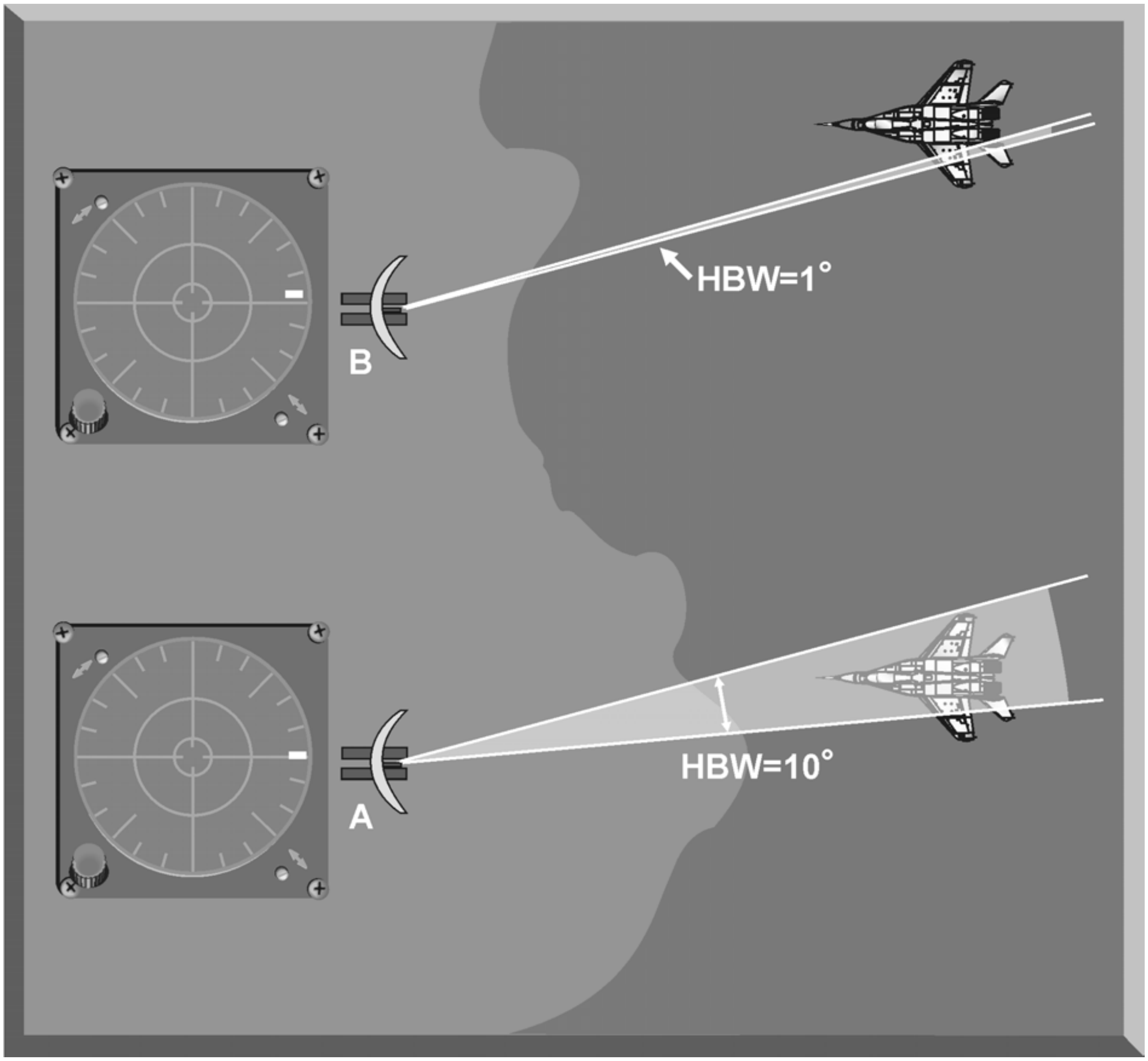

Figure 5-12. Horizontal Beamwidth Comparison

b. Azimuth resolution is the ability of a radar to display two targets flying at approximately the same range with little angular separation, such as two fighters flying line-abreast tactical formation. The azimuth resolution capability is usually expressed in nautical miles and corresponds to the minimum azimuth separation

required between two targets for separate display. Azimuth resolution depends on the horizontal beamwidth of the radar. The radar system in Figure 5-13 has a horizontal beamwidth of 10°. The two targets are so close in azimuth that the return echoes are blended into one return.

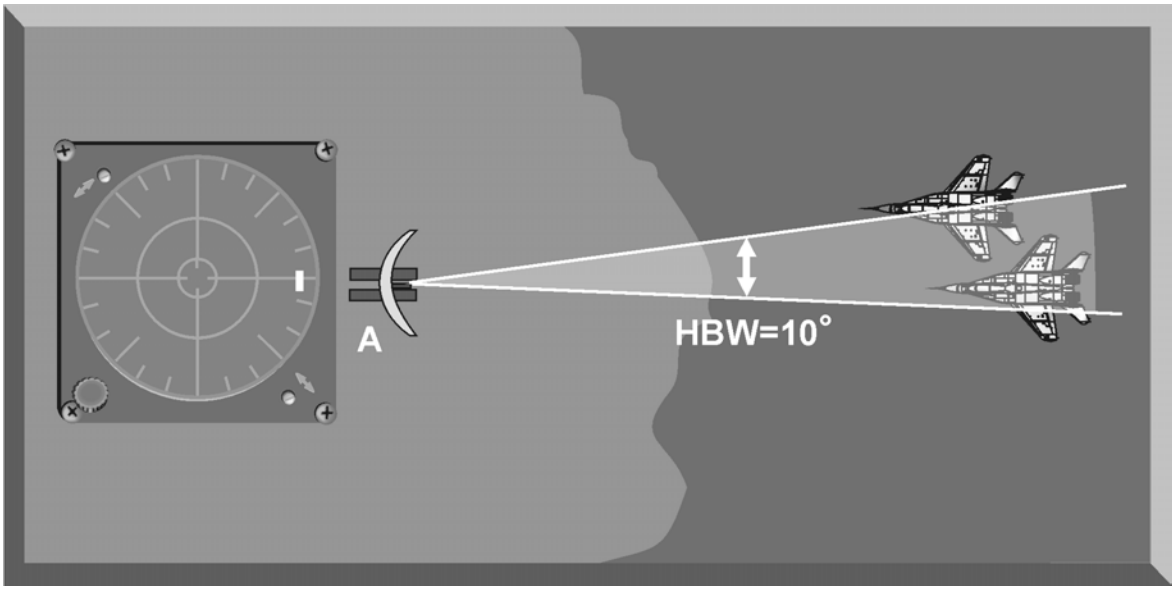

Figure 5-13. Horizontal Beamwidth and Azimuth Resolution

(1) The radar system in Figure 5-14 has a horizontal beamwidth of 1°. The radar beam not only hits the targets, but passes between them without causing a return. This allows the radars to display two distinct radar returns. A small horizontal beamwidth improves azimuth resolution.

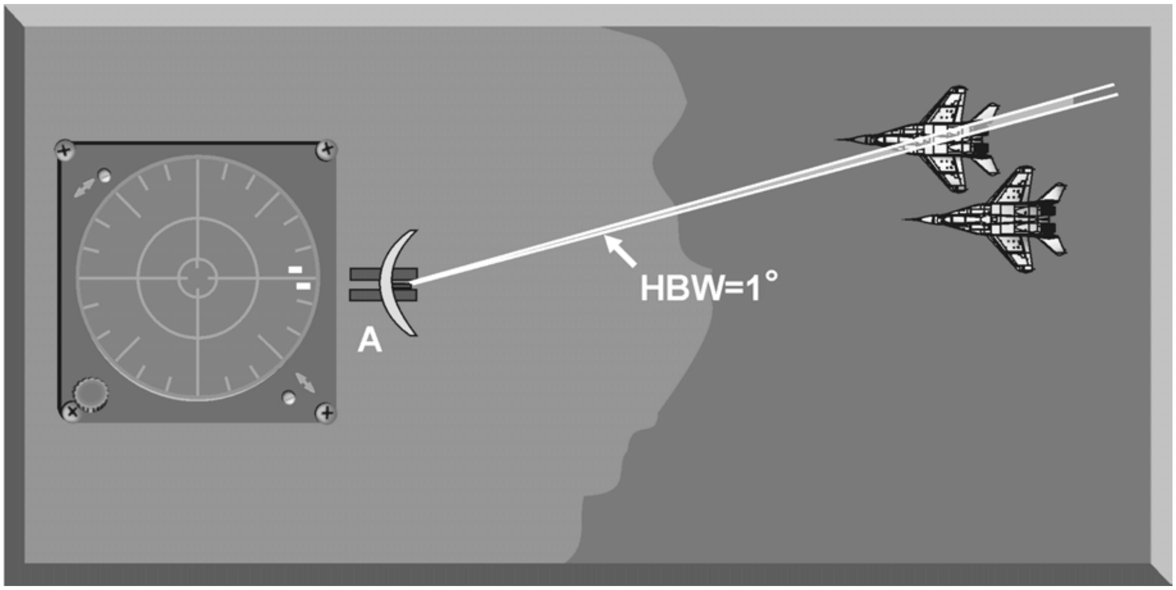

Figure 5-14. Azimuth Resolution

(2) Azimuth resolution, in nautical miles, can be computed using Equation 5-5. Notice that this equation is the "60 to 1 rule" used for navigation. A 1° beamwidth will yield a one-mile-wide cell at 60 nautical miles.

$$\text{Azimuth Resolution} = \frac{\text{Horizontal Beamwidth x Range}}{60}$$

Equation 5-5. Azimuth Resolution

4. ELEVATION DETERMINATION

Since a radar beam is three-dimensional, the vertical beamwidth is the primary factor in determining altitude resolution capability. Altitude resolution is the ability of a radar to display two targets flying at approximately the same range and azimuth with little altitude separation, such as two fighters flying a vertical stack formation. The altitude resolution capability is usually expressed in feet and corresponds to the minimum altitude separation required between two targets for separate display. The radar system in Figure 5-15 has a vertical beamwidth of 10°. The two targets are so close in altitude that the return echoes depicted on the range height indicator (RHI) are blended into one.

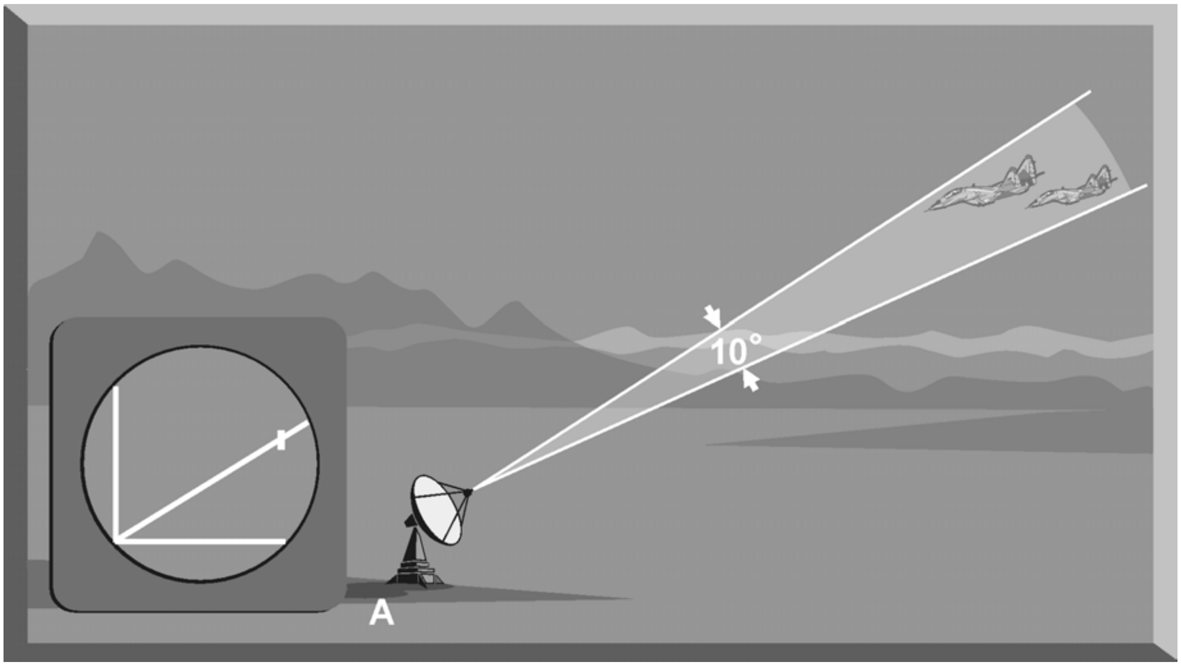

Figure 5-15. Vertical Beamwidth and Elevation Resolution

a. The radar system depicted in Figure 5-16 has a vertical beamwidth of 1°. This small beam not only hits the targets, but passes between them without causing a return. This allows the radar to display two distinct targets.

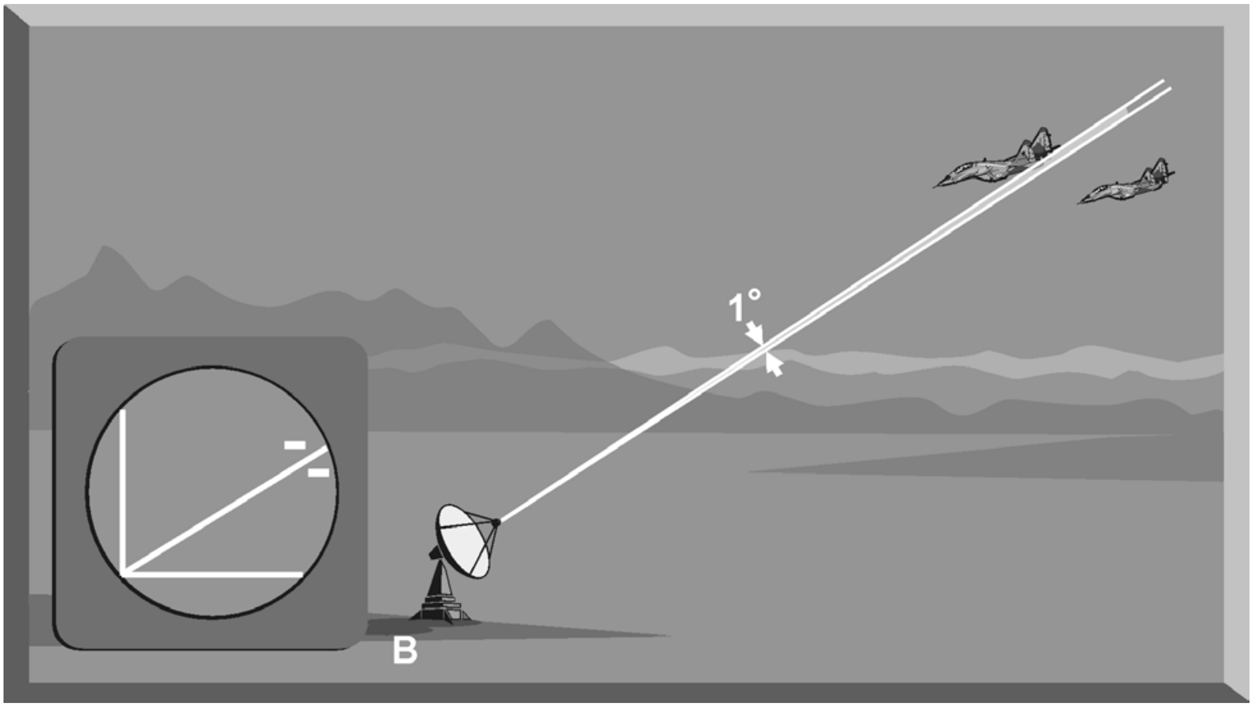

Figure 5-16. Elevation Resolution

b. Altitude/elevation resolution, in thousands of feet, can be computed using Equation 5-6.

$$\text{Altitude Resolution} = \frac{\text{Vertical Beamwidth} \times \text{Range}}{60}$$

Equation 5-6. Altitude Resolution

5. RADAR RESOLUTION CELL

A radar's pulse width, horizontal beamwidth, and vertical beamwidth form a three-dimensional resolution cell (RC) (Figure 5-17). A resolution cell is the smallest volume of airspace in which a radar cannot determine the presence of more than one target. The resolution cell of a radar is a measure of how well the radar can resolve targets in range, azimuth, and altitude. The horizontal and vertical

dimensions of a resolution cell vary with range. The closer to the radar, the smaller the resolution cell.

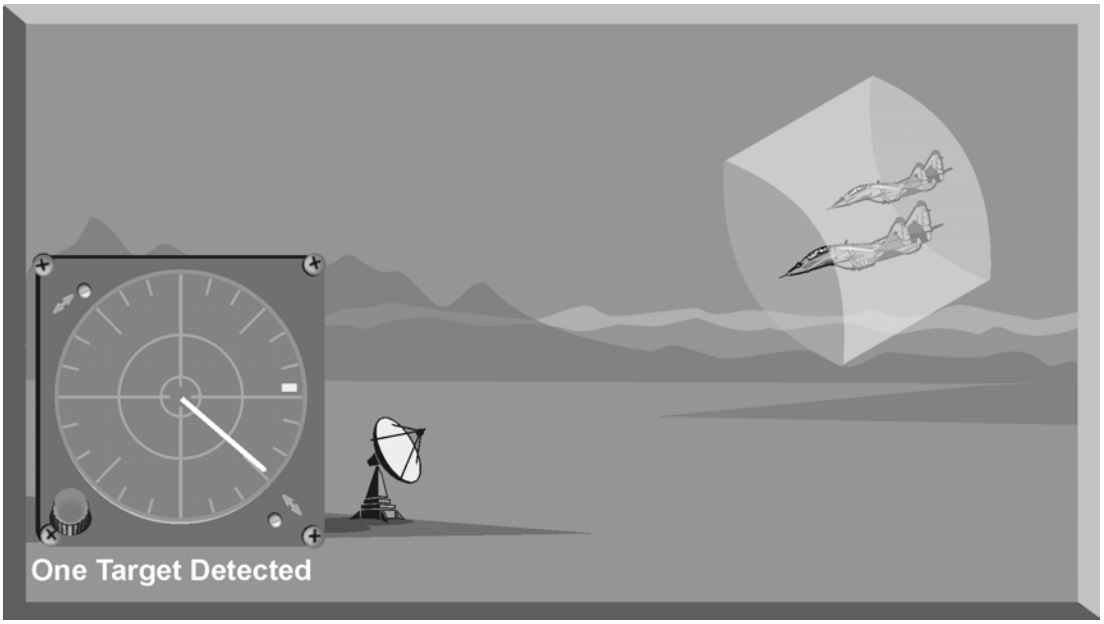

Figure 5-17. Radar Resolution Cell

a. The physical dimensions of a radar's resolution cell can be computed. For a radar with a pulse width of 1 microsecond, a horizontal beamwidth of 1°, and a vertical beamwidth of 10°, the formulas for range resolution, azimuth resolution, and altitude resolution can be used to compute the dimensions of the resolution cell. In the example in Figure 5-18, at a target range of 10 nm, the physical dimensions of the radar's resolution cell are 492 feet in range, by 1000 feet in azimuth, and 10,000 feet in altitude. These figures can be confirmed by using Equations 5-4, 5-5, and 5-6.

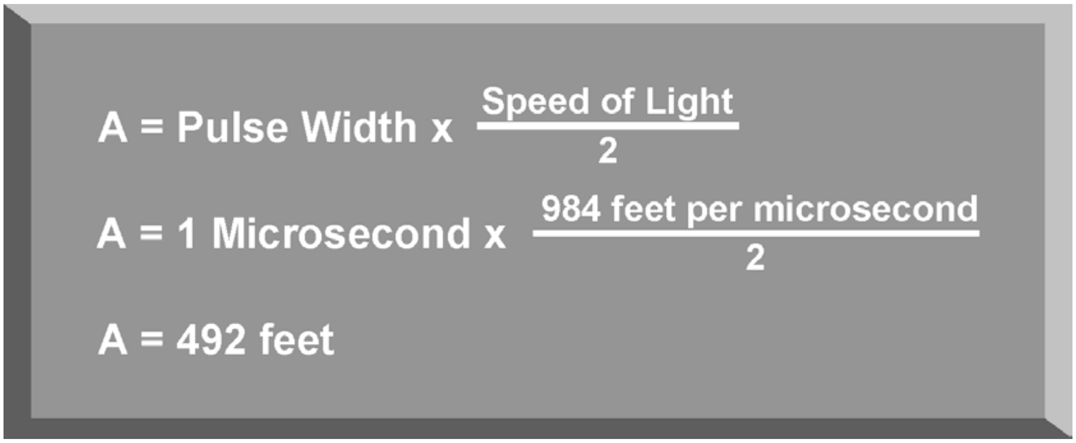

$$A = \text{Pulse Width} \times \frac{\text{Speed of Light}}{2}$$

$$A = 1 \text{ Microsecond} \times \frac{984 \text{ feet per microsecond}}{2}$$

$$A = 492 \text{ feet}$$

Figure 5-18. Radar Resolution Cell Dimensions

b. Based on these computations, two, or more, aircraft flying a trail formation closer than 492 feet would be displayed as a single target. Two, or more, aircraft flying line abreast closer than 1000 feet would be displayed as a single target. Two, or more, aircraft flying a vertical stack closer than 10,000 feet would be displayed as a single target. This also shows that the shorter the pulse width, the better the range resolution capability of a radar system. The narrower the horizontal beamwidth, the better the azimuth resolution capability. The narrower the vertical beamwidth, the better the altitude resolution capability.

c. Another type of resolution is velocity resolution. For a Doppler radar aircraft flying within the conventional resolution cell described above can be distinguished as separate targets if they have enough speed differential. Paragraph 6 below will describe how this is carried out.

6. PULSE DOPPLER VELOCITY DETERMINATION

To fully understand how a pulse Doppler radar determines target velocity, it is necessary to know more about the pulsed waveform. To generate a pulse modulated wave, a continuous carrier sine wave, like the output from a CW radar, is combined with a rectangular wave, like that of a pulse radar, to produce the pulse modulated waveform. Figure 5-19 depicts pulse modulation.

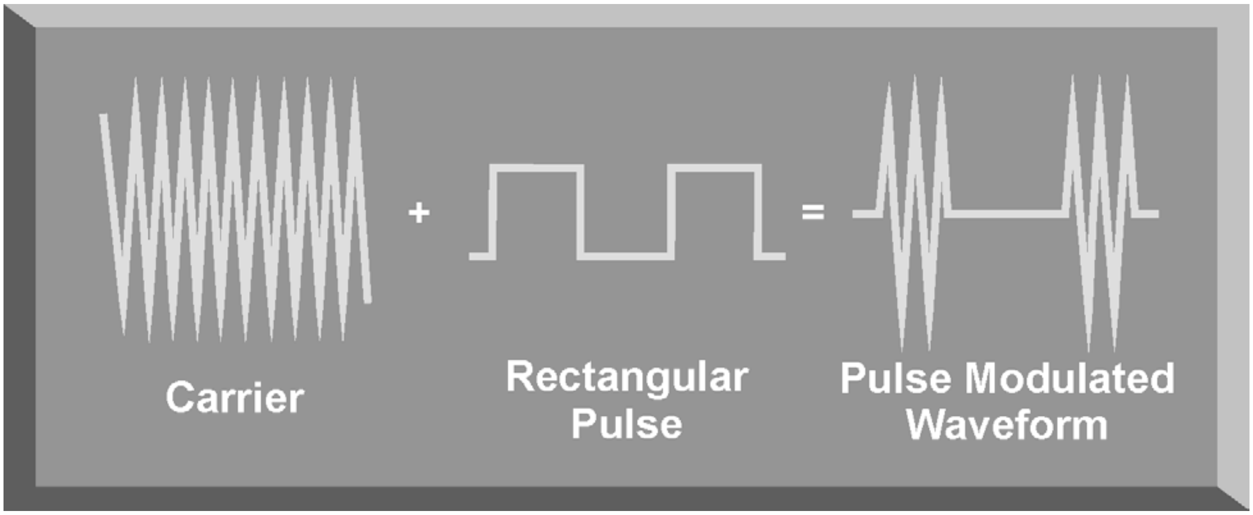

Figure 5-19. Pulse Modulation

a. Mathematically, any waveform other than a sine wave is composed of many different pure sine waves added in the proper amplitude and phase relationships (Figure 5-20). In a pulsed modulated waveform, the sine waves correspond to the fundamental frequency, which is the PRF, and the sum of all harmonics in the proper amplitude and phase. The frequency of the harmonic is the basic frequency plus or minus a multiple of the PRF.

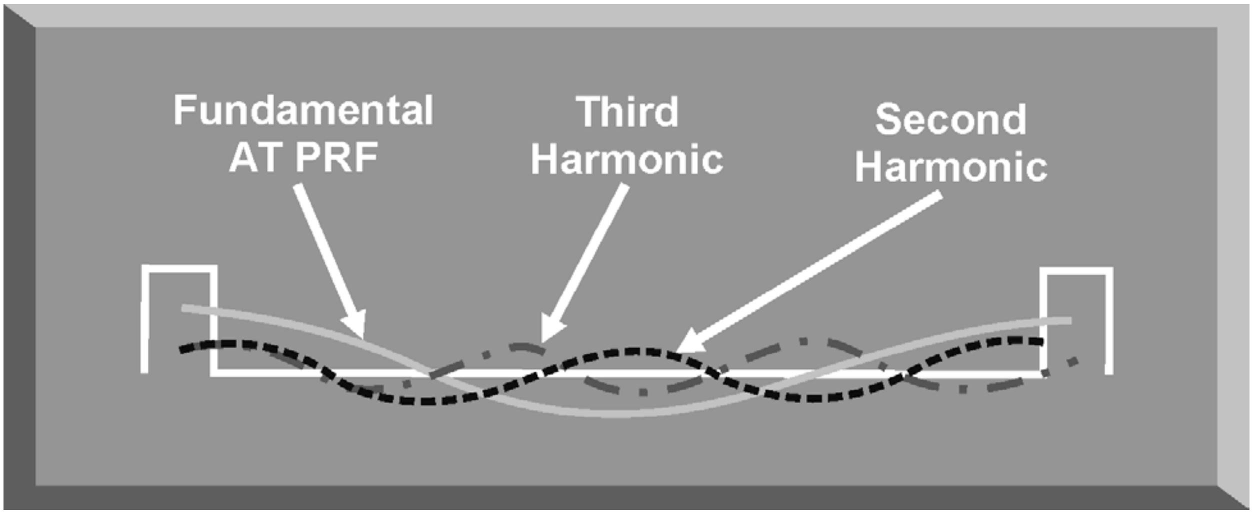
Fundamental
AT PRF
Third
Harmonic
Second
Harmonic

Figure 5-20. Harmonics of a Pulse Modulated Waveform

(1) Figure 5-21 is a plot of the harmonic content of a pulse modulated waveform operating at a carrier frequency of 2800 megahertz (MHz) with a PRF of 1 MHz. Note the loops of frequencies on either side of the carrier frequency. These are the additions and subtractions of all the frequencies in the rectangular pulse to the carrier frequency. The important thing to remember is that there are many frequencies present, and a pulse Doppler radar must deal with a crowded frequency spectrum. This becomes even more important when one considers the fact that every frequency present will experience a Doppler shift when it is reflected by a moving target. The individual frequencies shown are called spectral lines.

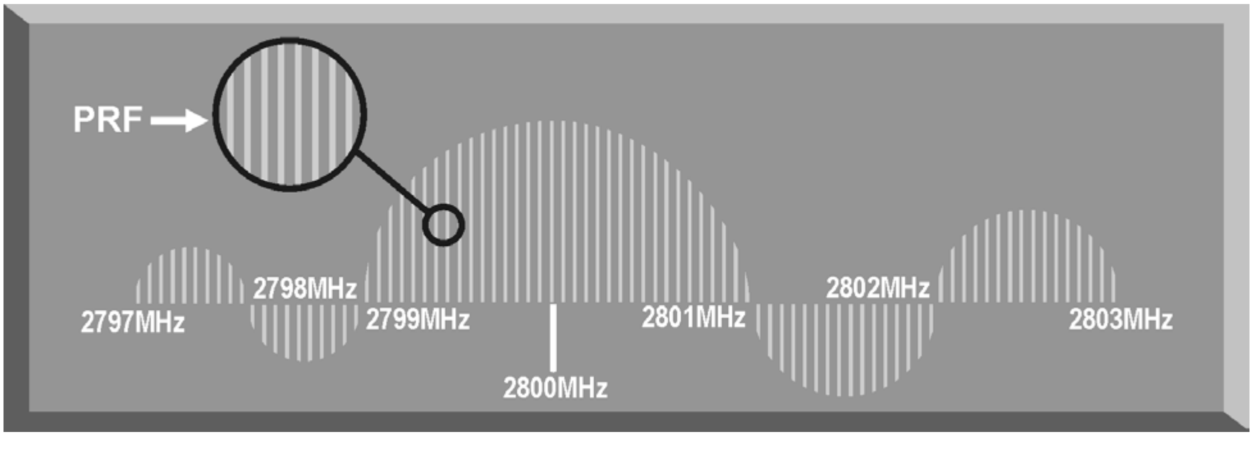
PRF
2797MHz
2798MHz
2799MHz
2800MHz
2801MHz
2802MHz
2803MHz

Figure 5-21. Harmonic Content

(2) For a pulse Doppler radar to accurately measure velocity, it must compare the frequency change, or Doppler shift, between the carrier frequency

and the frequency returning from the target. It is a difficult task for the radar to differentiate between the returning carrier and all the harmonic frequencies (Figure 5-22).

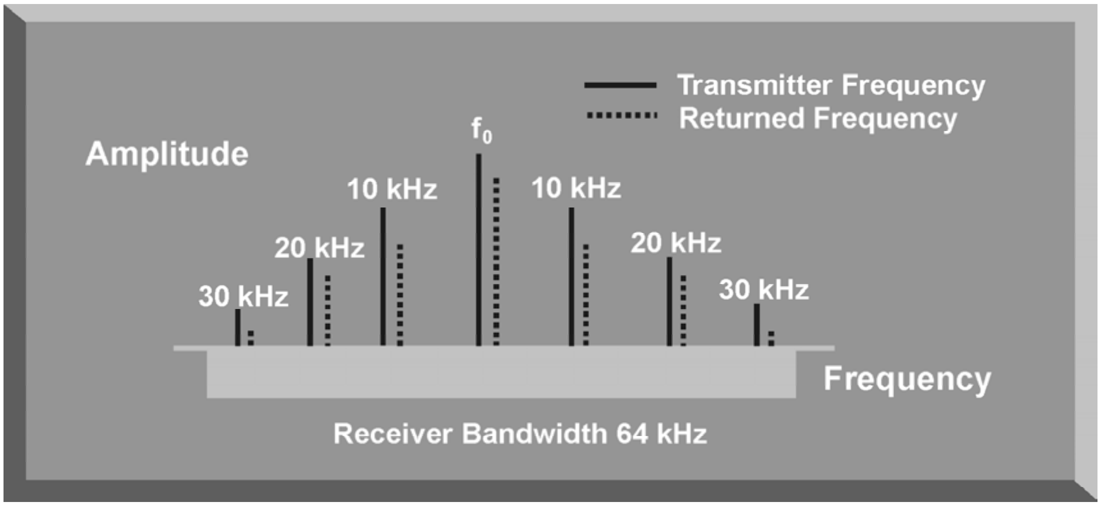
Transmitter Frequency
Returned Frequency
Amplitude
f0
10 kHz
10 kHz
20 kHz
20 kHz
30 kHz
30 kHz
Frequency
Receiver Bandwidth 64 kHz

Figure 5-22. Spectral Line Frequencies

b. The radar differentiates between the returning carrier frequencies and all other harmonic frequencies by using clutter cancellers, or filters, at the known harmonic frequencies (Figure 5-23). The radar cannot process frequencies cancelled by these filters. The filters create "blind speeds" for the radar. The closer together the spectral lines, the more "blind speeds" the radar will have.

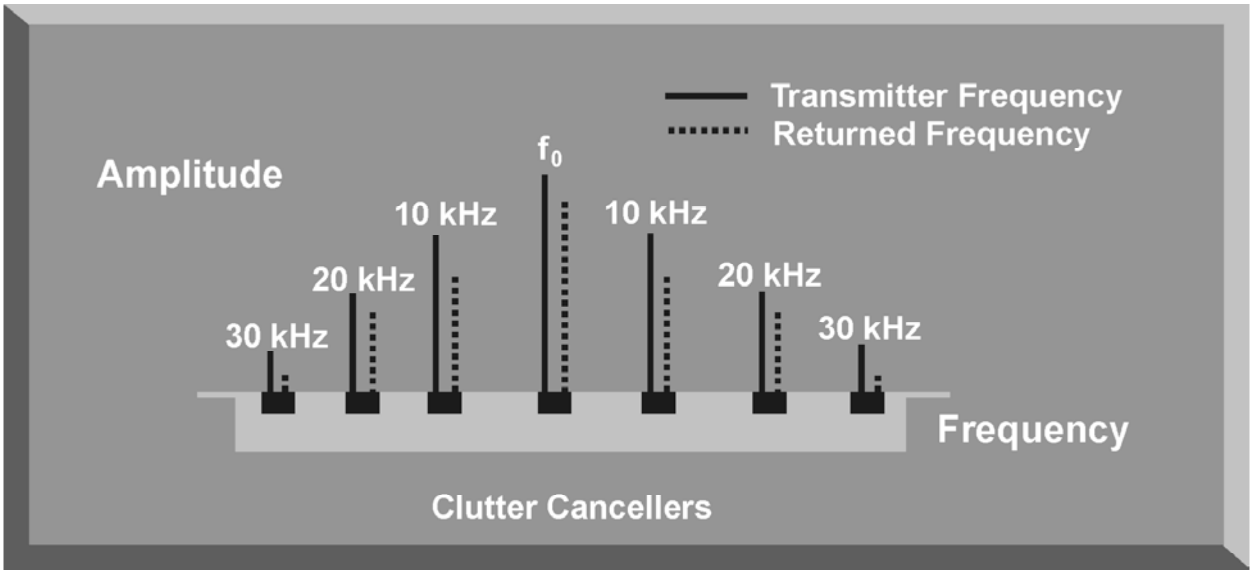
Transmitter Frequency
Returned Frequency
Amplitude
f0
10 kHz
10 kHz
20 kHz
20 kHz
30 kHz
30 kHz
Frequency
Clutter Cancellers

Figure 5-23. Selective Clutter Canceling

(1) Since the position of the harmonics in relation to the carrier frequency is based on PRF, the number of blind speeds can be reduced by changing the PRF of the radar. The higher the PRF, the wider the spacing of the spectral lines and the fewer blind speeds due to selective clutter canceling. However, a high PRF increases the problem of range ambiguities. Most modern pulse Doppler radars employ a medium and high PRF mode. Medium PRF equates to fewer range ambiguities but more blind speeds. High PRF has fewer blind speeds but more range ambiguities (Figure 5-24).

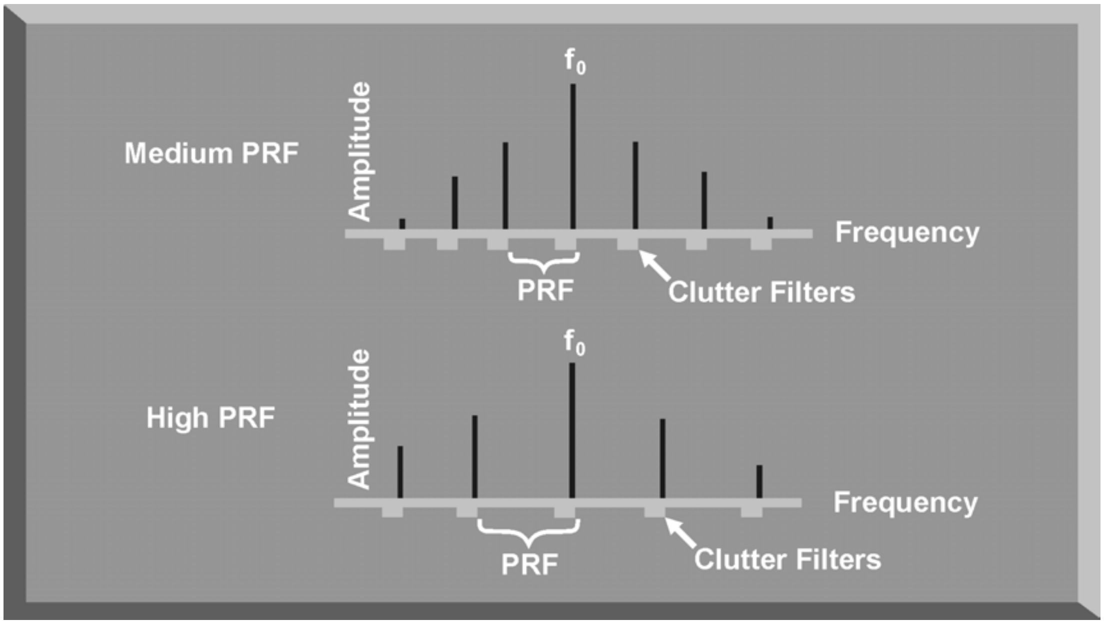

Figure 5-24. PRF and Spectral Lines

(2) To separate the returning target frequency shifts from all other frequencies in the returning waveform, the pulse Doppler radar employs filters to cancel the known harmonic frequency shifts. In addition, the radar cancels out all returns with no frequency shift, which equates to canceling all returns with no movement relative to the radar. However, if the radar has too many clutter filters, this creates multiple blind speeds, and targets will be missed. In Figure 5-25, the detection filters allow target frequencies to be processed, and clutter filters cancel unwanted frequency shifts. Target 1 will be detected, but Target 2 will be canceled.

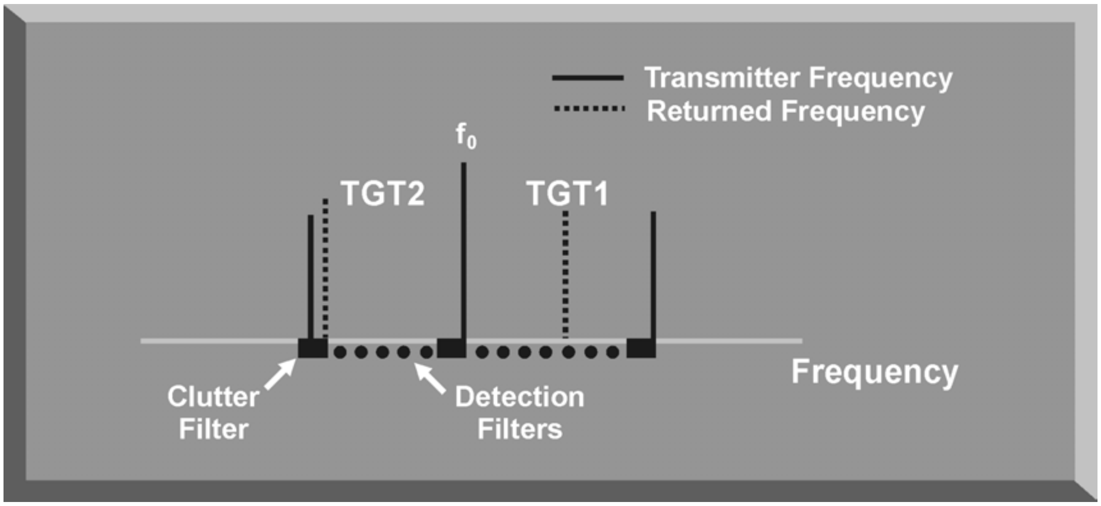

Figure 5-25. Pulse Doppler Filters

7. BASIC RADAR EQUATION

The basic radar equation relates the range of a radar system to the characteristics of the transmitter, receiver, antenna, and the target. The radar equation provides a means not only to figure out the maximum range of a particular radar system, but it can be used to understand the factors that affect radar operation. In this section, the simple forms of the radar equation are developed, starting with the power density of the transmitting antenna to the power received by the receiving antenna.

a. Power density is the power of a radio wave per unit of area normal to the direction of propagation. The power density generated by a practical antenna can be expressed in Equation 5-7.

$$\text{Power Density from Antenna} = \frac{P_T \times G}{4\pi r^2}$$

P_T = transmitted power
G = antenna gain
r = radius of the antenna

Equation 5-7. Power Density From an Antenna

b. As the radar beam propagates through space, it arrives at a target at some range (R) from the antenna. As the radar beam travels through space, the wavefront of the beam expands to a very large cross-sectional area, especially in relation to the target dimensions. The power density of the radar beam, across this wide area, at the target, is detailed in Equation 5-8.

$$\text{Power Density at Target} = \frac{P_T \times G}{4\pi R^2}$$

P_T = transmitted power
G = antenna gain
R = range to the target

Equation 5-8. Power Density at the Target

c. Since the cross-sectional area of the radar beam is so large, only a small portion of the total power in the beam can be reflected toward the antenna. The rest of the radar energy continues through space and is dissipated, absorbed, or reflected by other targets. The small portion of the radar beam that hits the target is reradiated in various directions. The measure of the amount of incident power intercepted by the target and reradiated back in the direction of the antenna depends on the radar cross section (RCS) of the target. Equation 5-9 details the power density of the target echo signal reflected back to the radar antenna.

$$\text{Power Density at Antenna} = \frac{P_T \times G}{4\pi R^2} \times \frac{\sigma}{4\pi R^2}$$

P_T = transmitted power
G = antenna gain
σ = RCS
R = range to the target

Equation 5-9. Power Density at the Antenna

d. As the target echo reaches the antenna, part of the echo is captured by the antenna based on the effective aperture (A_e). Equation 5-10 details the actual signal power received by the radar system. This is one form of the basic radar equation and is the signal strength of a radar return from a specific target at range (R) from the radar.

$$\text{Signal Power Density (S)} = \frac{P_T G \sigma A_e}{(4\pi)^2 R^4}$$

P_T = transmitted power

G = antenna gain

σ = RCS

R = range to the target

A_e = antenna aperture area

Equation 5-10. Signal Power Density

e. A detailed analysis of this equation is not required to draw some basic conclusions about the factors affecting the detection of an aircraft. If any factor in the numerator, such as transmitted power, is increased by a factor of three, the signal received by the radar will increase by only 30 percent. This clearly shows why radar system operation is characterized by the transmission of megawatts of power and the reception of microwatts of returning power. In addition, this equation shows that the most critical factor in determining radar detection is target range.

f. The maximum radar range (R_{MAX}) occurs when the signal power density received just equals the minimum detectable signal (S_{MIN}) for the receiver. Solving Equation 5-11 for range, and substituting S_{MIN}, yields the basic radar equation for R_{MAX} for a specific target. This is another form of the basic radar equation.

$$R_{MAX} = \frac{\left[P_T G \sigma A_e\right]^{1/4}}{\left[(4\pi)^2 S_{MIN}\right]}$$

P_T = transmitted power

G = antenna gain

σ = RCS

A_e = antenna aperture area

R = range to the target

Equation 5-11. Basic Radar Equation

g. It is important to note that the basic radar equations do not consider such factors as meteorological conditions, changes in aircraft RCS, the impact of clutter on gain, or operator abilities. The radar equation does explain why a radar system designed for long-range detection should transmit a very high power signal, concentrated into a narrow beam, collected by a large antenna, and processed by a very sensitive receiver.

8. SUMMARY

This chapter has discussed the methods employed by radar systems to determine target range, azimuth, elevation, and velocity. The relationship between pulse width and range resolution, beamwidth and azimuth/elevation resolution, and PRF and velocity resolution have been explained. In addition, an explanation of the complex radar equation has been presented. The capabilities and limitations of a specific radar system to determine these critical target parameters is the key to understanding the countermeasures designed to defeat this system.

CHAPTER 6. ANTENNA CHARACTERISTICS AND SCANS

1. INTRODUCTION

The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction. During reception, or listening time, the function of the antenna is to collect the returning radar energy, contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern. The primary antenna types in use today fall into three categories: parabolic, Cassegrain, or phased array antennas. Additionally, the method radar antennas employ to sample the environment is a critical design feature of the radar system. The scan type selected for a particular radar system often decides the employment of that radar in an integrated air defense system (IADS). The process the radar antenna uses to search airspace for targets is called scanning or sweeping. This chapter discusses circular, unidirectional, bidirectional, helical, raster, Palmer, and conical scans, and track-while-scan (TWS) radar systems.

2. PARABOLIC ANTENNA

One of the most widely used radar antennas is the parabolic reflector (Figure 6-1). The parabola-shaped antenna is illuminated by a source of radar energy, from the

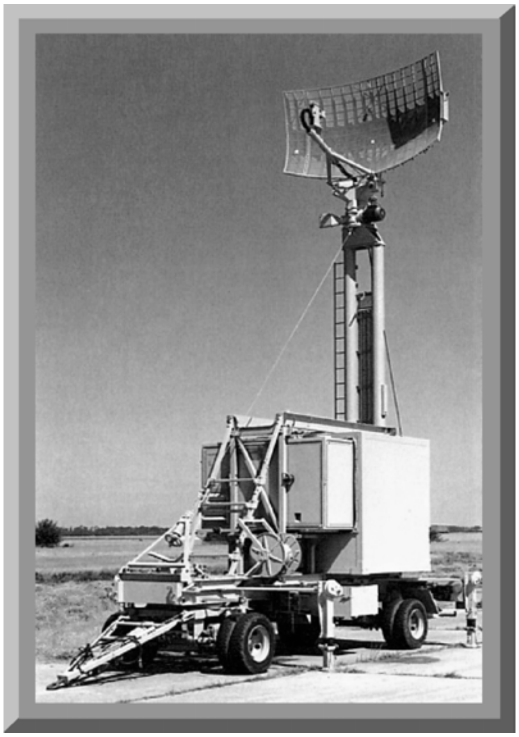

Figure 6-1. Parabolic Antenna

transmitter, called the feed. The feed is placed at the focus of the parabola, and the radar energy is directed at the reflector surface. Because a point source of energy, located at the focus, is converted into a wavefront of uniform phase, the parabola is well suited for radar antenna applications. By changing the size and shape of the parabolic reflecting surface, a variety of radar beam shapes can be transmitted.

a. The antenna depicted in Figure 6-1 generates a nearly symmetrical pencil beam that can be used for target tracking.

b. Elongating the horizontal dimensions of the parabolic antenna creates a radar antenna called the parabolic cylinder antenna (Figure 6-2). The pattern of this antenna is a vertical fan-shaped beam. Combining this antenna pattern with a circular scan technique creates a radar system well suited for long-range search and target acquisition.

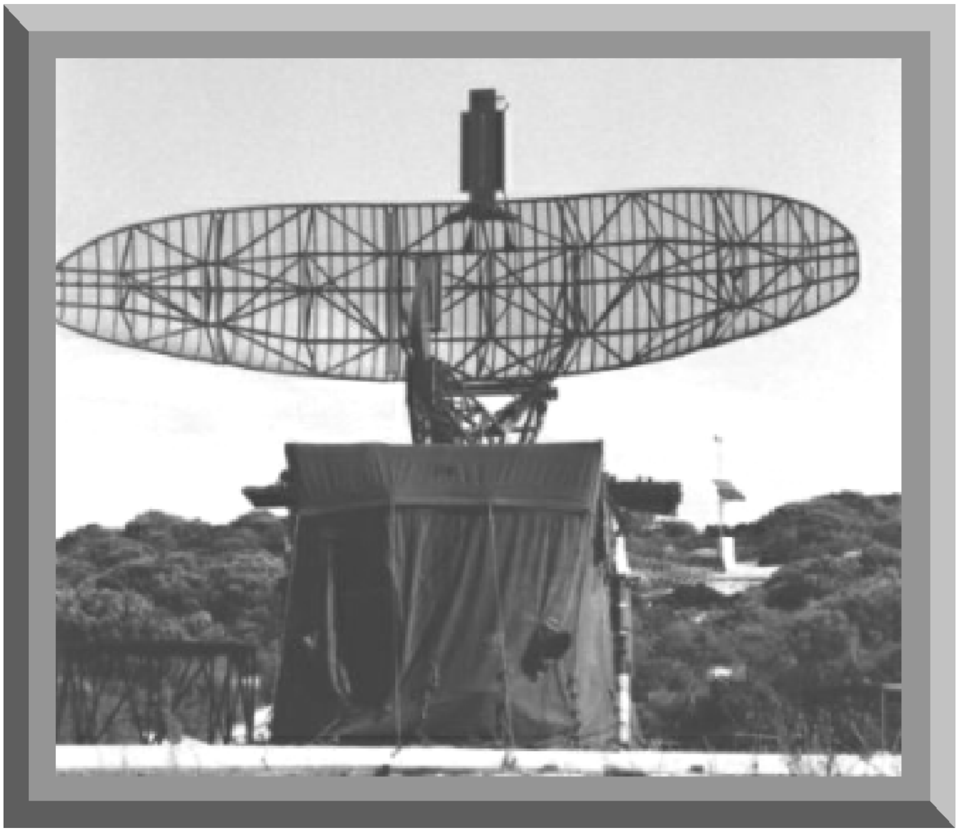

Figure 6-2. Parabolic Cylinder Antenna

c. Elongating the vertical dimensions of the parabola creates a radar antenna that generates a horizontal fan-shaped beam with a small vertical dimension (Figure 6-3). This type of antenna is generally used in height-finding radar systems.

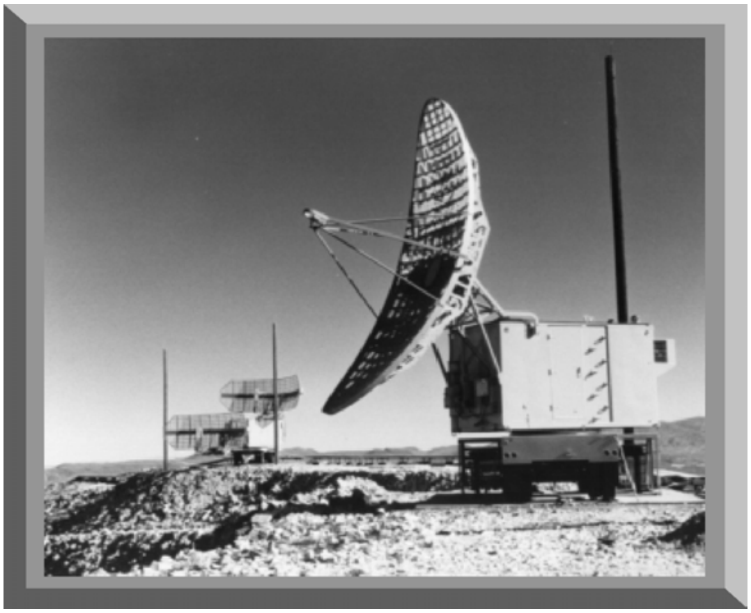

Figure 6-3. Height-Finder Parabolic Antenna

d. Another variation of the basic parabolic antenna includes using an array of multiple feeds instead of a single feed (Figure 6-4). This type of parabolic antenna can produce multiple radar beams, either symmetrical or asymmetrical, depending on the angle and spacing of the individual feeds.

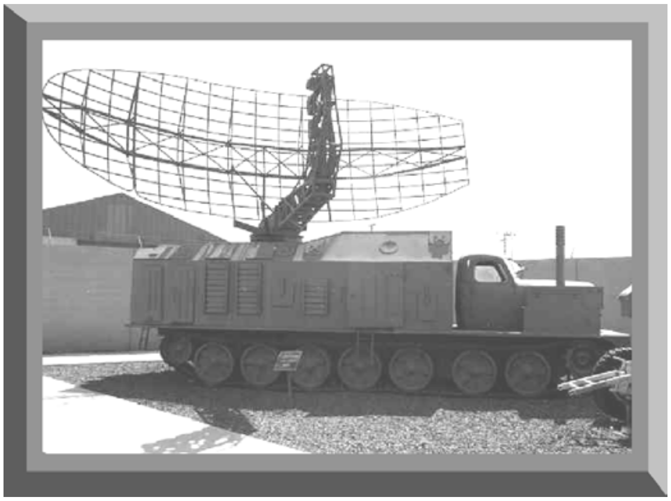

Figure 6-4. Multiple-Feed Parabolic Antenna

3. CASSEGRAIN ANTENNA

A Cassegrain antenna uses a two-reflector system to generate and focus a radar beam (Figure 6-5). The primary reflector uses a parabolic contour, and the secondary reflector, or subreflector, has a hyperbolic contour. The antenna feed

is located at one of the two foci of the hyperbola. Radar energy from the transmitter is reflected from the subreflector to the primary reflector to focus the radar beam. Radar energy returning from a target is collected by the primary reflector and reflected as a convergent beam to the subreflector. The radar energy is rereflected by the subreflector, converging at the position of the antenna feed. The larger the subreflector, the closer it can be to the primary reflector. This reduces the axial dimensions of the radar but increases aperture blockage due to the subreflector. A small subreflector reduces aperture blockage, but it must be positioned at a greater distance from the primary reflector.

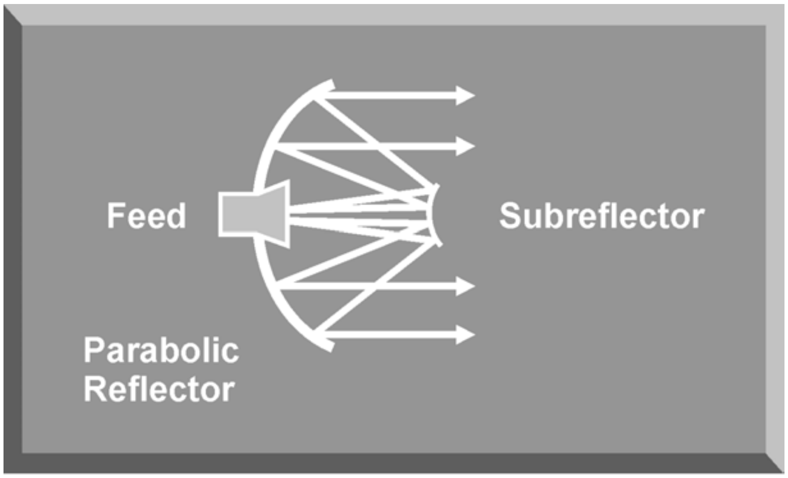
Feed
Subreflector
Parabolic
Reflector

Figure 6-5. Cassegrain Antenna

a. To reduce the aperture blockage by the subreflector and to provide a method to rapidly scan the radar beam, the flat plate Cassegrain antenna was developed. This type of Cassegrain antenna is depicted in Figure 6-6.

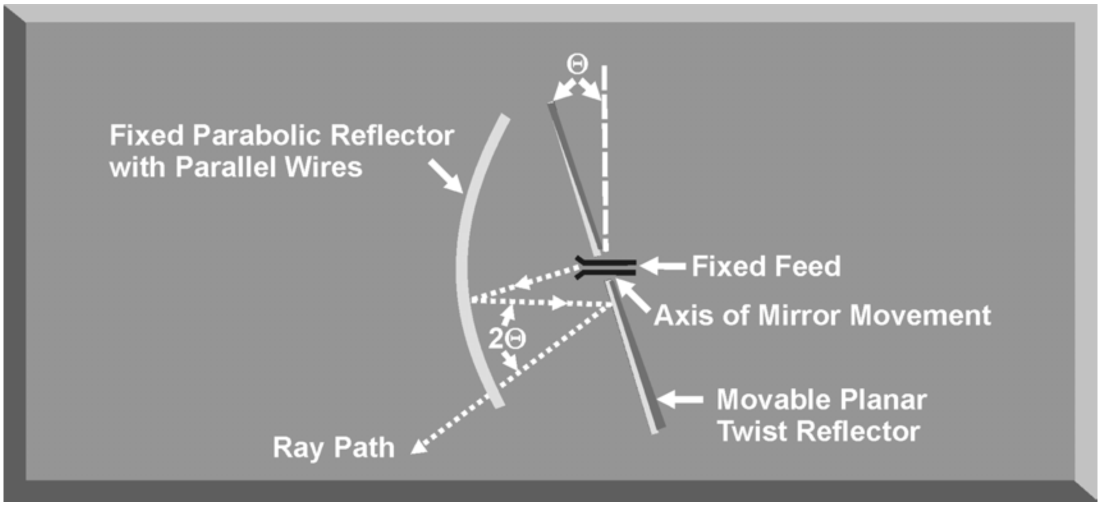
Θ
Fixed Parabolic Reflector
with Parallel Wires
Fixed Feed
Axis of Mirror Movement
2Θ
Movable Planar
Twist Reflector
Ray Path

Figure 6-6. Flat Plate Cassegrain Antenna

The fixed parabolic reflector is made up of parallel wires spaced less than a half wavelength apart and supported by a low-loss dielectric material. This makes the fixed parabolic reflector polarization sensitive. It will completely reflect one type of linear polarization and be transparent to the orthogonal polarization. The fixed antenna feed, in the middle of the moveable mirror, transmits a radar signal polarized to be reflected by the parabolic reflector. The moveable mirror is constructed as a twist reflector that changes the polarization of the radar signal by 90°. The signal from the feed is reflected by the parabolic reflector to the mirror, which rotates the polarization 90°. This rotation makes the transmitted signal transparent to the parabolic reflector, and the signal passes through with minimal attenuation. The radar beam can be scanned over a wide area by rotating the moveable mirror. A deflection of the mirror by the angle θ results in the beam scanning through an angle of 2θ.

b. The geometry of the Cassegrain antenna is especially well suited for monopulse tracking radar applications. Unlike the parabolic antenna, the complex feed assembly required for a monopulse radar can be placed behind the reflector to avoid aperture blocking.

4. PHASED ARRAY ANTENNA

The phased array radar is a product of the application of computer and digital technologies to the field of radar design. A phased array antenna is a complex arrangement of many individual transmitting and receiving elements in a particular pattern. A phased array antenna can, in effect, radiate more than one beam from the antenna by using a computer to rapidly and independently control groups of these individual elements. Multiple beams and computer processing of radar returns give the phased array radar the ability to track-while-scanning and engage multiple targets simultaneously. Figure 6-7 is a view of the phased array radar antenna.

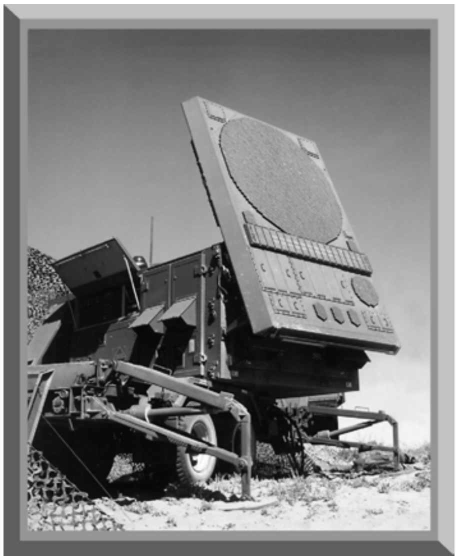

Figure 6-7. Phased Array Antenna

a. A phased array radar uses the principle of radar phase to control the individual transmitting and receiving elements. When two transmitted frequencies are in-phase, their amplitudes add together, and the radiated energy is doubled. When two transmitted frequencies are out-of-phase, they cancel each other. Phased array radars use this principle to control the shape of the transmitted radar beam (Figure 6-8).

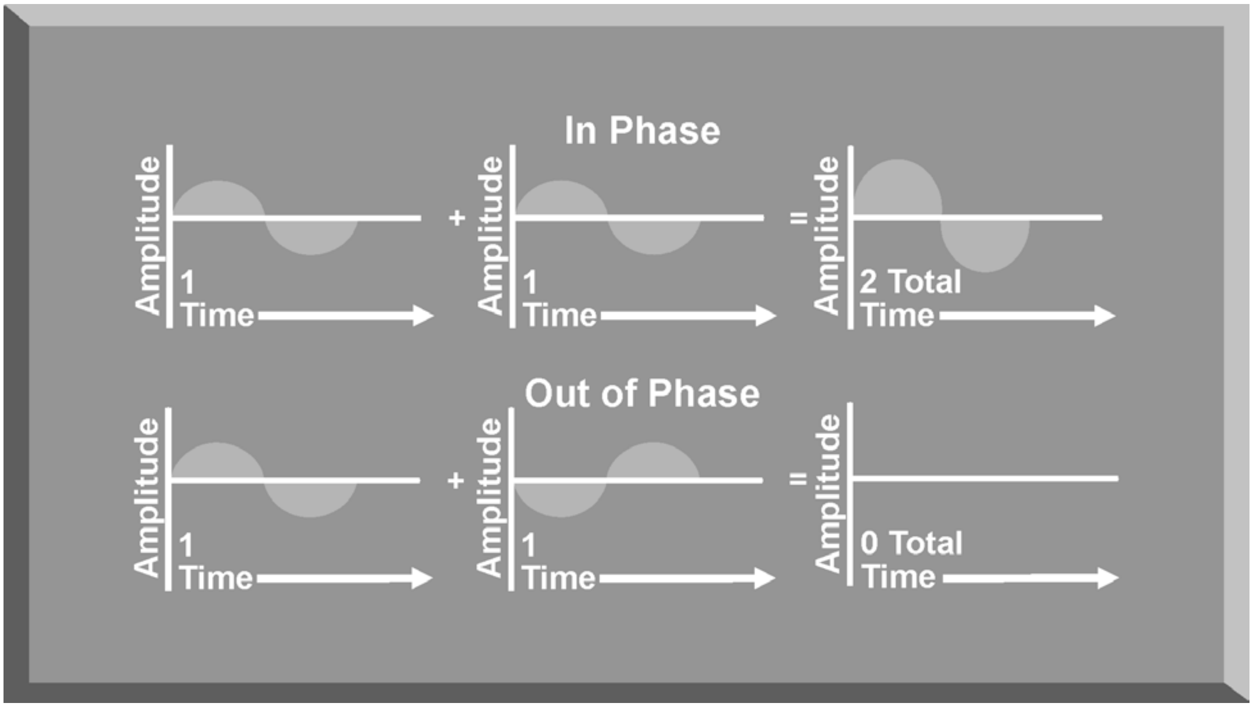

Figure 6-8. Phase Relationships

b. Phase relationships and antenna element spacing determine the orientation of the transmitted beam. In Figure 6-9, antenna elements A and B are separated by one-half wavelength and are radiating in-phase, that is, when one is at the positive peak, the other is also at a positive peak. Since the elements are one-half wavelength apart, when the positive peak radiated by A reaches B, B will be radiating a negative peak. As the peaks propagate along the X axis, they will cancel each other out. The total radiated power along that axis will be zero. Along the Y axis, however, the positive peaks from A will add to the positive peaks from B, causing the total radiation along this axis to be at its maximum value. This type of array is called a "broadside array" because most of the radiation is in the direction that is broadside to the line of the antenna array.

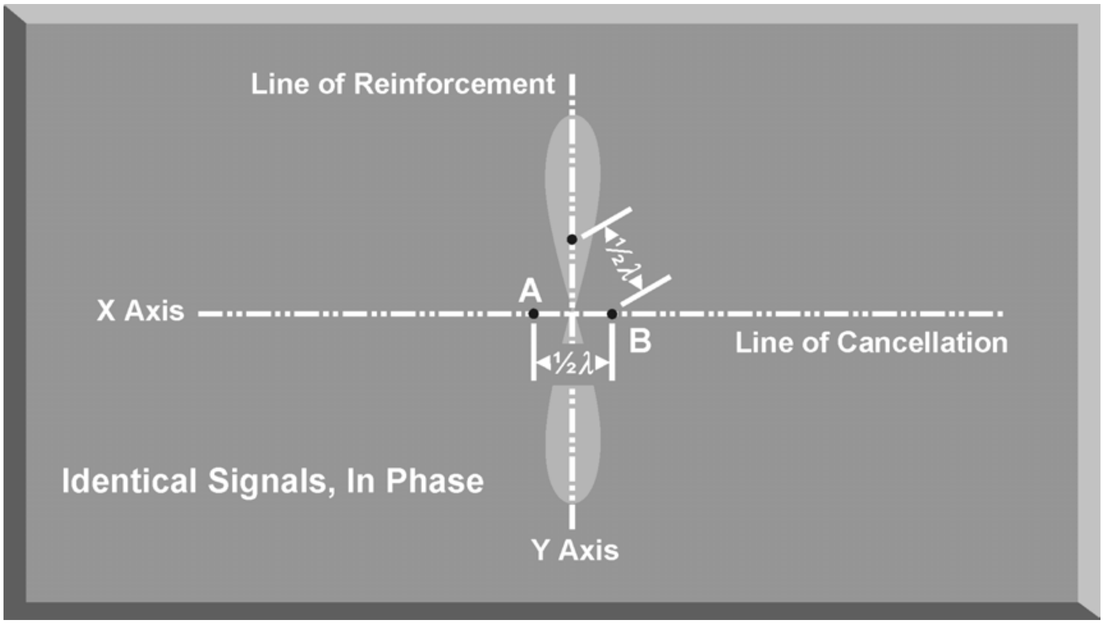
Line of Reinforcement
½λ
A
X Axis
B
Line of Cancellation
½λ
Identical Signals, In Phase
Y Axis

Figure 6-9. Broadside Array

c. If the same antenna elements are fed out-of-phase, the principal direction of radiation will be along the axis of the antenna elements. In Figure 6-10, when the positive peak from A arrives at B, B is now positive also. These energies interact to strengthen the energy being radiated from the ends of the array. Meanwhile, when the positive peak from A, radiating along the Y axis, meets the negative peak from B, they are canceled. This type of array is called an "end-fire array."

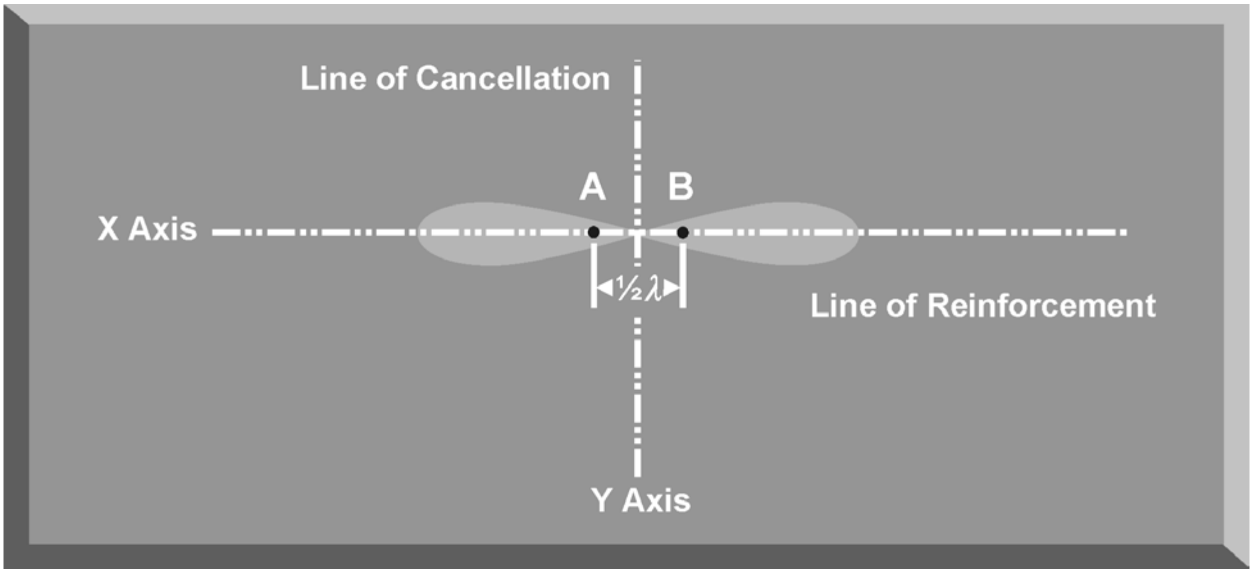
Line of Cancellation
A
B
X Axis
½λ
Line of Reinforcement
Y Axis

Figure 6-10. End-Fire Array

d. The computer controlling the phase of the signal delivered to each transmitting and receiving element of a phased array antenna controls the direction and shape of the radiated beam (Figure 6-11). By shifting the phase of the signals between 0° and 180°, the beam sweeps. This is the basic means of producing an antenna scan. In addition, the amplitude, or power, of the signal applied to each element can be varied to control the sidelobes. This alters the shape of the beam which affects the range capability and angular resolution of the radar.

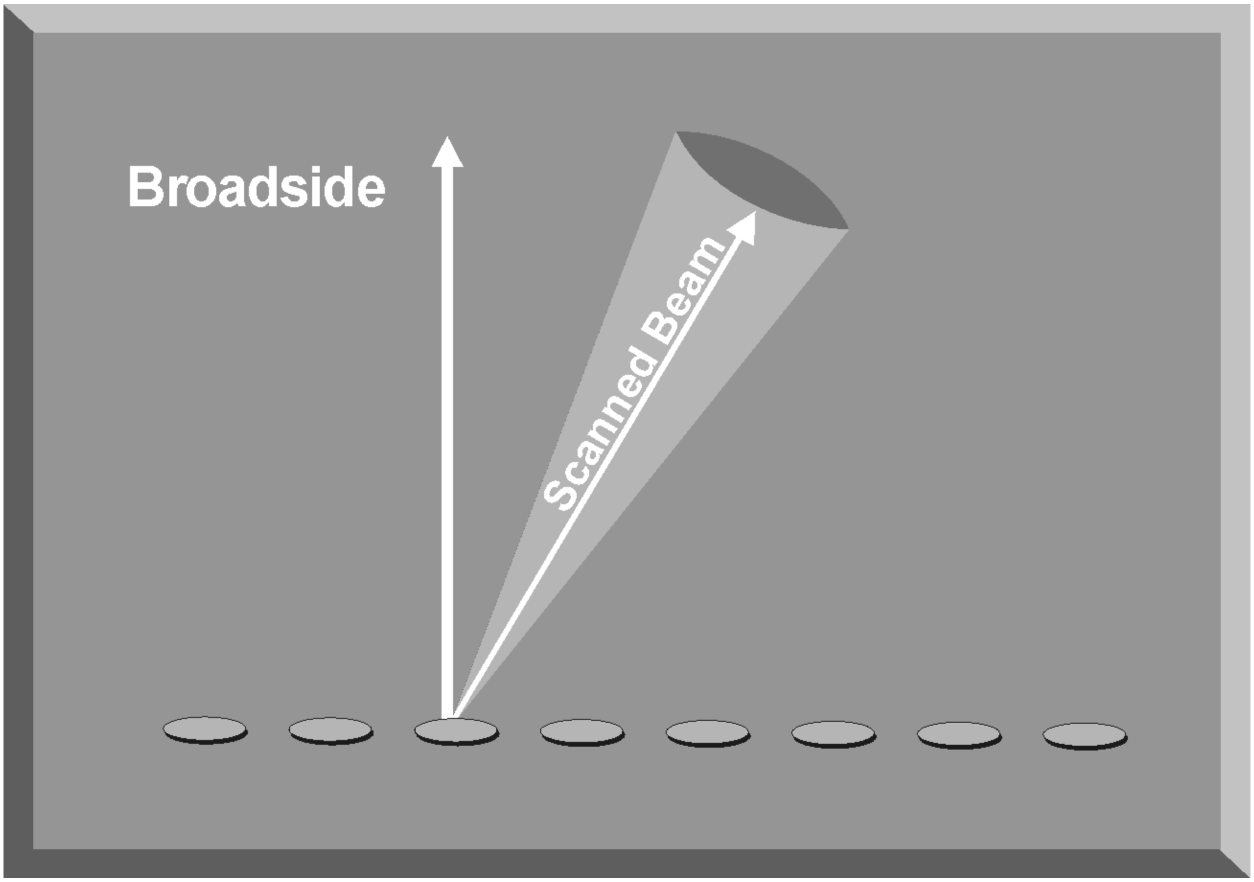

Figure 6-11. Phased Array Antenna Scan

e. Figure 6-12 depicts a variation of the phased array antenna, known as a planar array antenna. A planar array antenna uses transmit and receive elements in a linear array, but, unlike the phased array radar, the elements are smaller and are placed on a movable flat plate. The ability to simultaneously track several targets is one advantage of this type of radar.

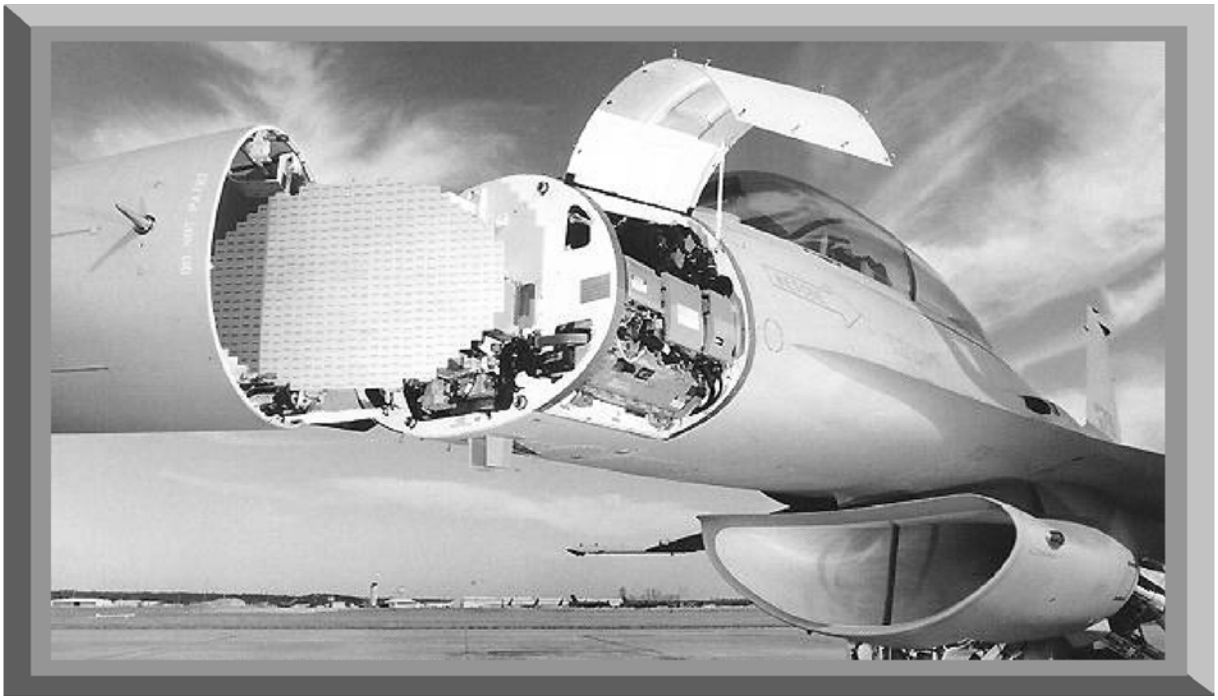

Figure 6-12. Planar Array Antenna

5. ANTENNA GAIN

The most important characteristic of any type of antenna is antenna gain. Antenna gain is a measure of the ability of an antenna to concentrate energy in the desired direction. Antenna gain should not be confused with receiver gain, which is designed to control the sensitivity of the receiver section of a radar system. There are two types of antenna gain: directive and power.

a. The directive gain of a transmitting antenna is the measure of signal intensity radiated in a particular direction. Directive gain is dependent on the shape of the radiation pattern of a specific radar antenna. The directive gain does not take into account the dissipative losses of the antenna. Directive gain is computed using Equation 6-1.

$$G_D \text{ (Directive Gain)} = \frac{\text{Maximum Radiation Intensity (Desired Direction)}}{\text{Average Radiation Intensity}}$$

Equation 6-1. Directive Gain

b. The power gain does include the antenna dissipative losses and is computed using Equation 6-2.

$$\text{G (Power Gain)} = \frac{\text{Maximum Radiation Intensity (Practical Antenna)}}{\text{Radiation Intensity of an Isotropic Antenna}}$$

Equation 6-2. Power Gain

c. The term isotropic antenna describes a theoretical spherical antenna that radiates with equal intensity in all directions. This results in a spherical radiation pattern. The power density for any point on an isotropic antenna is the radiation intensity and can be calculated by dividing the total power transmitted (P_T) by the total surface area of the sphere, as shown in Equation 6-3.

$$\text{Power Density (Isotropic Antenna)} = \frac{P_T\ \text{(Watts)}}{4\pi r^2\ \text{(Centimeters}^2\text{)}}$$

Equation 6-3. Power Density for an Isotropic Antenna

d. The radiation pattern of an isotropic, or spherical, antenna would provide neither azimuth or elevation resolution and would be unusable for radar applications. To provide azimuth and elevation resolution, a practical antenna must focus the radar energy. The power density of a practical antenna differs from the isotropic antenna only in terms of antenna gain (G). Solving Equation 6-3 for the power density of a practical antenna yields Equation 6-4.

$$\text{Power Density (Practical Antenna)} = \frac{P_T\, G}{4\pi r^2}$$

Equation 6-4. Power Density for a Practical Antenna

e. The actual power gain (G) of a practical antenna can be calculated by using Equation 6-5.

$$G = \frac{4\pi A_e}{\lambda}$$

A_e = effective area of aperture

λ = wavelength of the radar

NOTE: The effective area of aperture (A_e) is the effective antenna area presented to the returning radar echo.

Equation 6-5. Power Gain of a Practical Antenna

6. POWER DENSITY

The power density and gain of an antenna are a function of the antenna pattern of a radar system. Figures 6-13 and 6-14 illustrate the antenna pattern of a typical parabolic antenna. Most of the power density of the radar is concentrated in the main beam. However, since the radar is not a perfect reflector, some radar energy is transmitted in the sidelobes. In addition, there is spillover radiation due to the energy radiated by the feed that is not intercepted by the reflector. Finally, the radar has a back lobe caused by diffraction effects of the reflector and direct signal leakage. Sidelobes and backlobes are all undesirable radiations that adversely affect the maximum radar range and increases the vulnerability of the radar to certain jamming techniques.

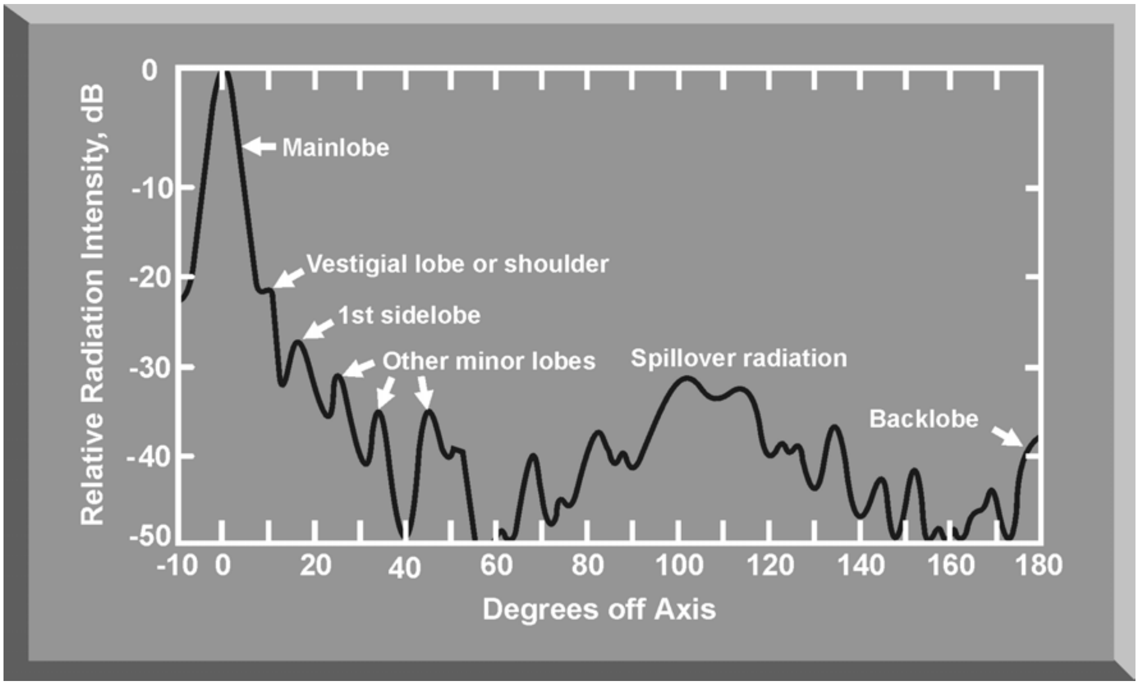

Figure 6-13. Radiation Pattern for a Parabolic Antenna

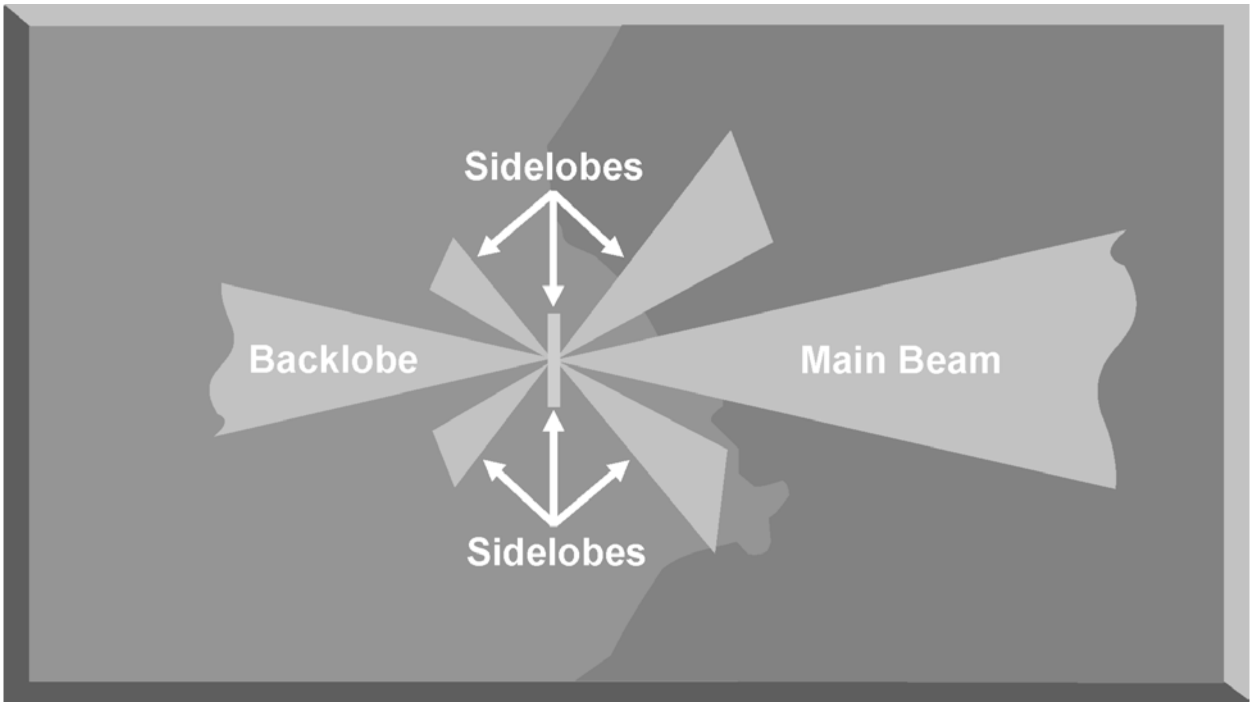

Figure 6-14. Radar Antenna Pattern

a. All radars have a primary main beam, which is where the radar has the most power and where target detection usually occurs. The dimensions of this main beam are highly dependent on the design of the antenna.

b. Besides the main beam, all radars have what is called a backlobe. This lobe is directly opposite to the location of the main beam. The sensitivity and signal strength associated with the backlobe is significantly less than that in the main beam.

c. Sidelobes add another dimension to the radar pattern. As with the backlobe, sidelobes do not have the signal strength or sensitivity associated with the main beam. Normally, the sensitivity associated with the sidelobes is 40-50 decibels (dBs) less than the main beam. The radar signal weakness in the backlobe and sidelobes of the main beam make these areas of the radar signal vulnerable to jamming. It is much easier to introduce jamming into these areas because of the reduced jamming-to-signal ratio needed to be effective. It is difficult for jamming to be effective in the main beam because the radar signal is very powerful in that region.

7. CIRCULAR SCAN

A circular scanning radar uses an antenna system that continuously scans 360° in azimuth (Figure 6-15). The time required for the antenna to sweep one complete 360° cycle is called the scan rate. Scan duration is the number of "hits per scan,"

or the number of pulses, reflected by a target as the radar beam crosses it during one full scan. Most pulse radars require 15 to 20 hits per scan to obtain sufficient information to display a target. The factors that determine the number of hits per scan the radar receives include pulse repetition frequency (PRF), antenna beamwidth, and scan duration.

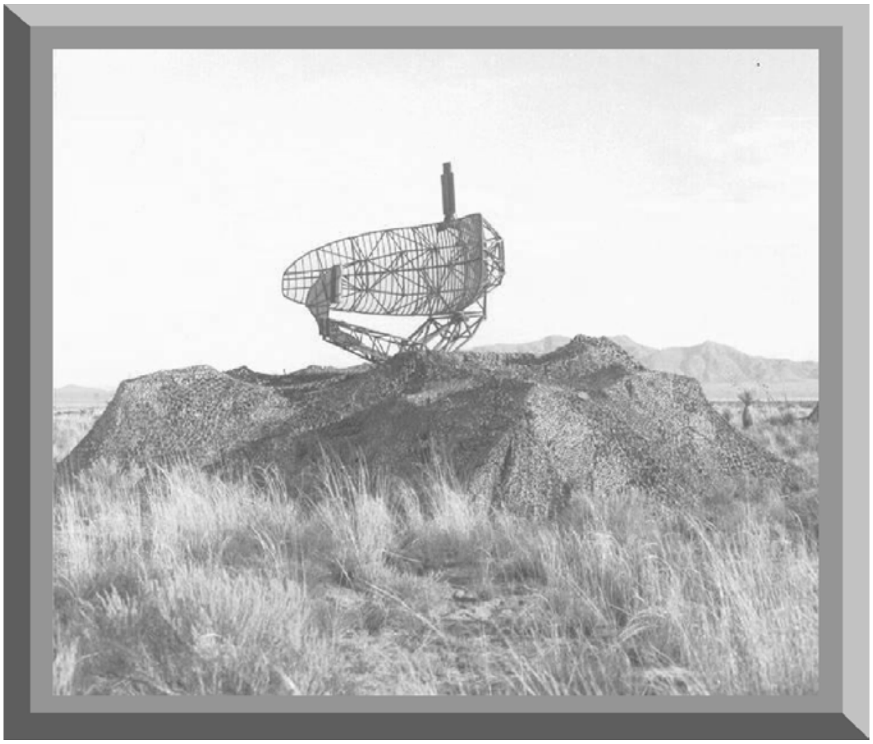

Figure 6-15. Circular Scan Radar

a. Circular scan radars provide accurate target range and azimuth information. This makes these radars ideal for the roles of early warning and initial target acquisition. To accomplish these missions, the antenna generates a fan beam that has a large vertical beamwidth and a small horizontal beamwidth. Since elevation information will normally be provided by height finder radars, the size of the vertical beamwidth is not a limitation. This antenna scan allows the radar to scan large volumes of airspace for early target detection. Since early detection is the primary goal of early warning radars, accurate altitude and azimuth resolution are secondary considerations.

b. Circular scan radars designed for early warning transmit a radar signal with a low PRF. A low PRF allows sufficient time for the radar pulse to travel long distances, and return, before another pulse is transmitted. This gives the radar system a long, unambiguous range capability. Circular scan radars with low PRFs generally use long pulse widths in order to increase their average power and long-range detection capability. The scan durations of early warning radars are relatively long to provide the required "hits per scan" for long-range target

detection. The plan position indicator (PPI) scope display is normally used with a circular scan radar (Figure 6-16).

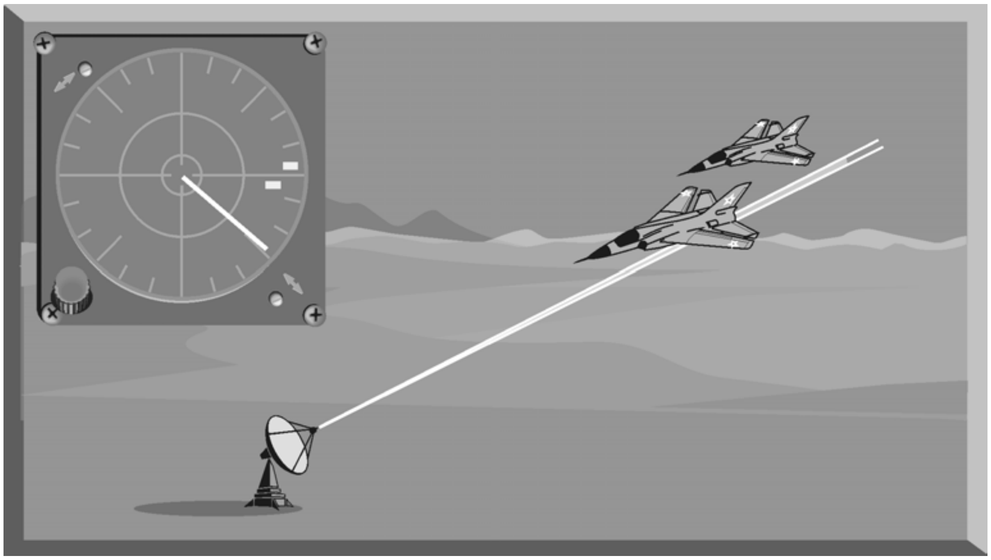

Figure 6-16. PPI Scope Display

c. In order to provide coverage for a large volume of airspace, the beamwidth associated with a circular scan radar is relatively wide. This wide beamwidth, coupled with the long pulse width and low PRF, gives the circular scan radar a large resolution cell, especially at long ranges (Figure 6-17). This limitation can be exploited to mask force size and composition. However, as range decreases, the dimensions of the resolution cell decrease, and a circular scan radar will begin to break out target formations.

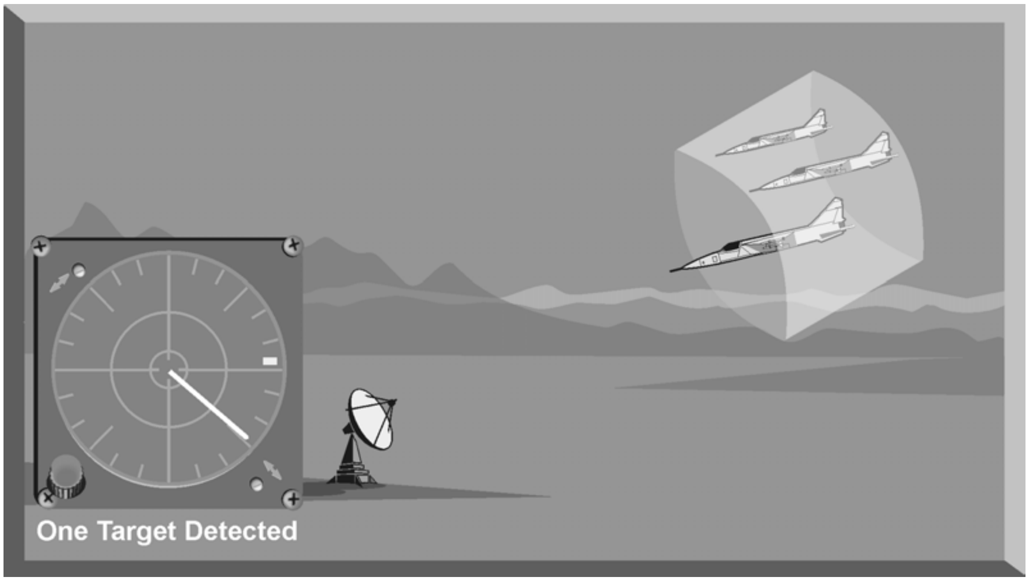

Figure 6-17. Resolution Cell

d. Circular scan radars provide range and azimuth information for both early warning and acquisition roles. Modified circular scan radars that can also provide elevation information may be used for ground control intercept (GCI) roles. Two modified circular scan radars that determine range, azimuth, and elevation are the V-beam and the stacked beam.

(1) The V-beam radar transmits two fan-shaped beams that are swept together (Figure 6-18). A vertical beam provides range and azimuth information. A second beam, rotated at some convenient angle, provides a measure of the altitude of the target.

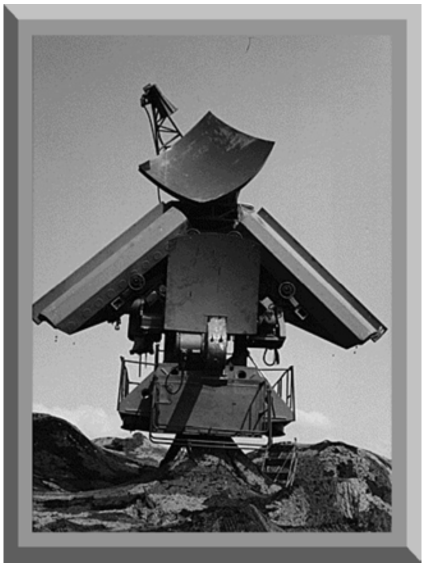

Figure 6-18. V-Beam Radar

(2) A stacked beam radar (Figure 6-19) employs a vertical stack of fixed elevation "pencil" beams which rotate 360°. Elevation information is obtained by noting which beam contains the target return. Range and azimuth information is determined in the same manner as in an early warning radar.

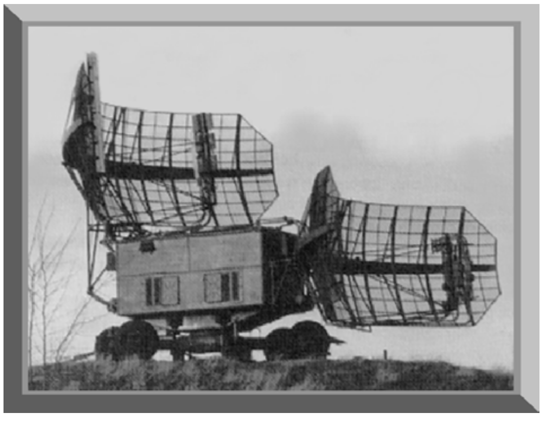

Figure 6-19. Stacked Beam Radar

8. LINEAR SCAN

Linear scan is a method used by some radar systems to sweep a narrow radar beam in a set pattern to cover a large volume of airspace. Linear scans can be oriented in a vertical direction for height finder radars or in a horizontal direction, or raster, for acquisition and target tracking radars. A unidirectional linear radar scans in a single direction then begins its sweep all over again (Figure 6-20). Generally, linear scans offer excellent single-axis coverage, and the narrow beam offers enhanced azimuth and elevation resolution.

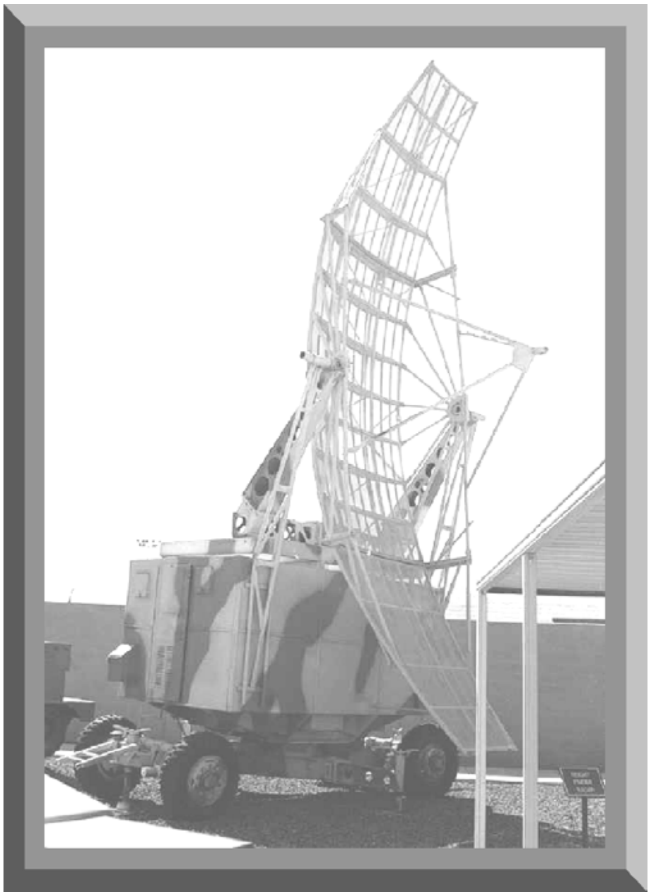

Figure 6-20. Unidirectional Linear Scan

9. UNIDIRECTIONAL SCAN

A helical scan is a unidirectional scan pattern that allows a “pencil” beam to search a 360° pattern. In Figure 6-21, the antenna sweeps a 360° sector in a clockwise direction. After each complete revolution, the antenna elevation is increased. This scan pattern is repeated for a specified number of revolutions, in this case, three, 360° sweeps. At the end of the scan pattern, the antenna elevation is reset to the initial elevation and the scan is repeated. A helical scan pattern is commonly used as a target acquisition mode for radar systems with narrow vertical and horizontal beamwidths.

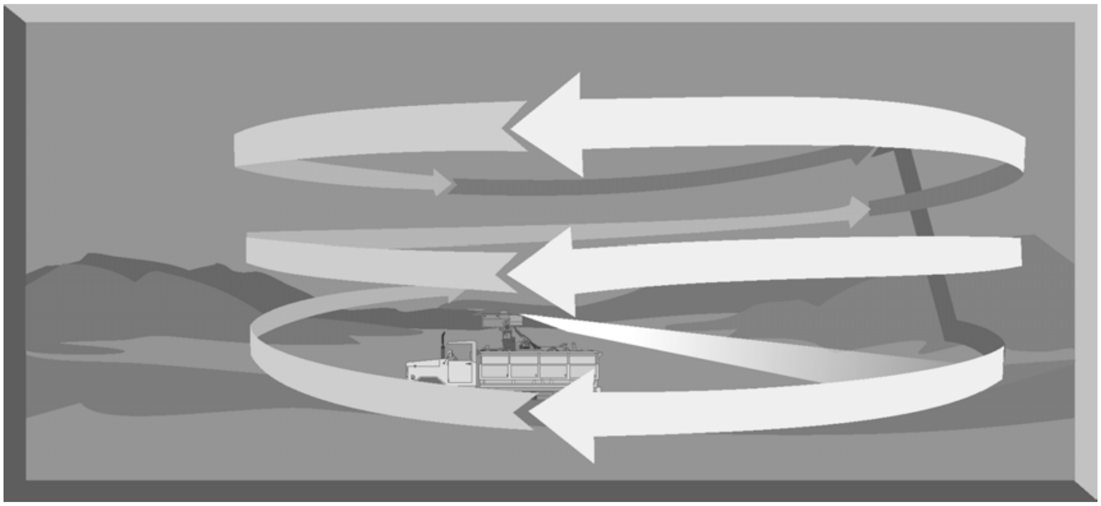

Figure 6-21. Helical Scan

10. BIDIRECTIONAL SCAN

A bidirectional linear scan, such as a raster scan, sweeps both horizontally and vertically (Figure 6-22). A raster scan uses a thin beam to cover a rectangular area by horizontally sweeping the area. The angle of elevation is incrementally stepped up or down with each horizontal sweep of the desired sector. After the sector has been covered, the angle of elevation is reset to the original value and the process is repeated. The number of raster bars is set by the number of horizontal sweeps in the basic raster pattern. Figure 6-22 shows a four-bar raster scan, which is normally associated with an airborne interceptor (AI) radar.

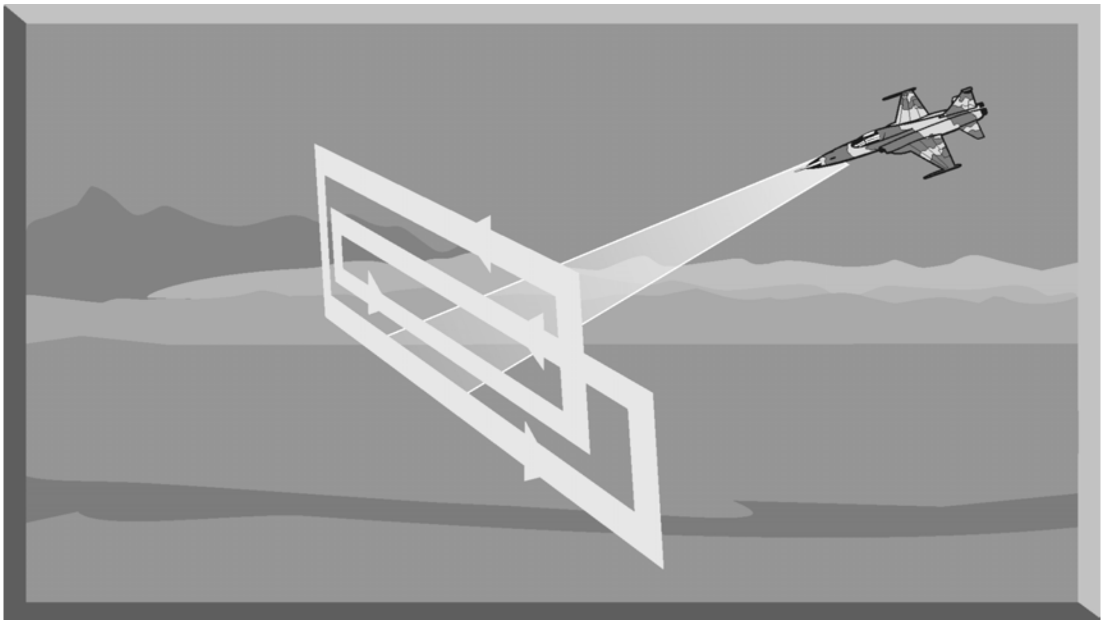

Figure 6-22. Raster Scan

11. CONICAL SCAN

A conical scan, or conscan, radar is generally used for precision target tracking. A conical scan radar (Figure 6-23) employs a pencil beam of radar energy that is continuously rotated around the target. This circular rotation of a pencil beam generates a cone-shaped scan pattern with the apex of the cone located at the antenna. Thus, the name conical scan.

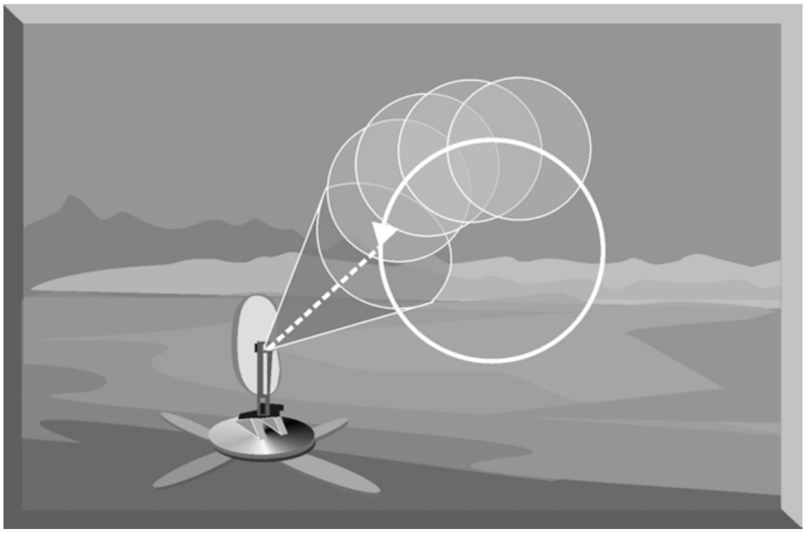

Figure 6-23. Conical Scan

a. As the pencil beam rotates, the circular scan patterns overlap in the center. This creates a central tracking area that has a much smaller effective beamwidth than the rotating pencil beam (Figure 6-24). This results in a very precise tracking solution.

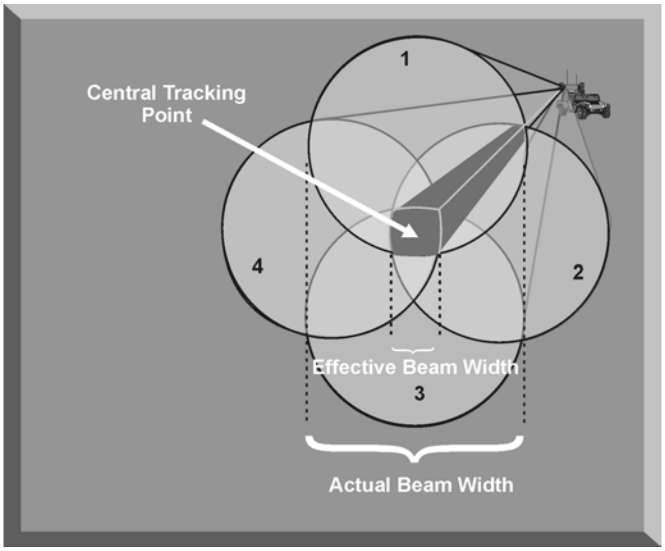

Figure 6-24. Conical Scan Tracking Area

b. Since conical scan radars are designed for precision target tracking, these radars normally operate at high frequencies, high PRF, narrow pulse widths, and narrow beamwidths. The rotation rate of the pencil beam can exceed 1,800 revolutions per minute. This means that both azimuth and elevation data can be updated about 30 times per second.

c. The combination of conical scan and raster scan is called a Palmer-raster scan (Figure 6-25). A Palmer-raster scan uses a thin beam, employing a conical scan searching pattern, for a specific sector of airspace. With each sweep of the sector, the angle of elevation is incrementally stepped up or down. After the vertical sector has been covered, the angle of elevation is set at the original elevation and the process is repeated. The number of bars is determined by the number of vertical search scans.

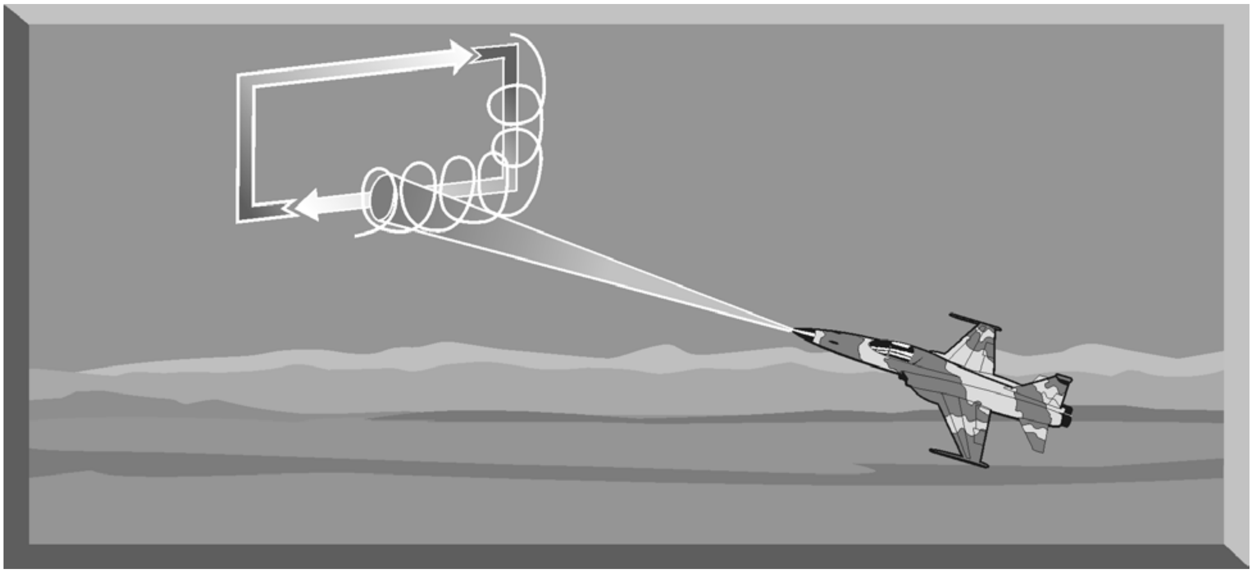

Figure 6-25. Palmer-Raster Scan

d. The combination of a conical scan and a circular scan is called a Palmer scan. Palmer scans incorporate a circular scanning antenna to search the entire horizon while simultaneously performing a conical scan. If the radar antenna is also performing a unidirectional altitude search in conjunction with this scan, it is employing a Palmer-helical scan (Figure 6-26).

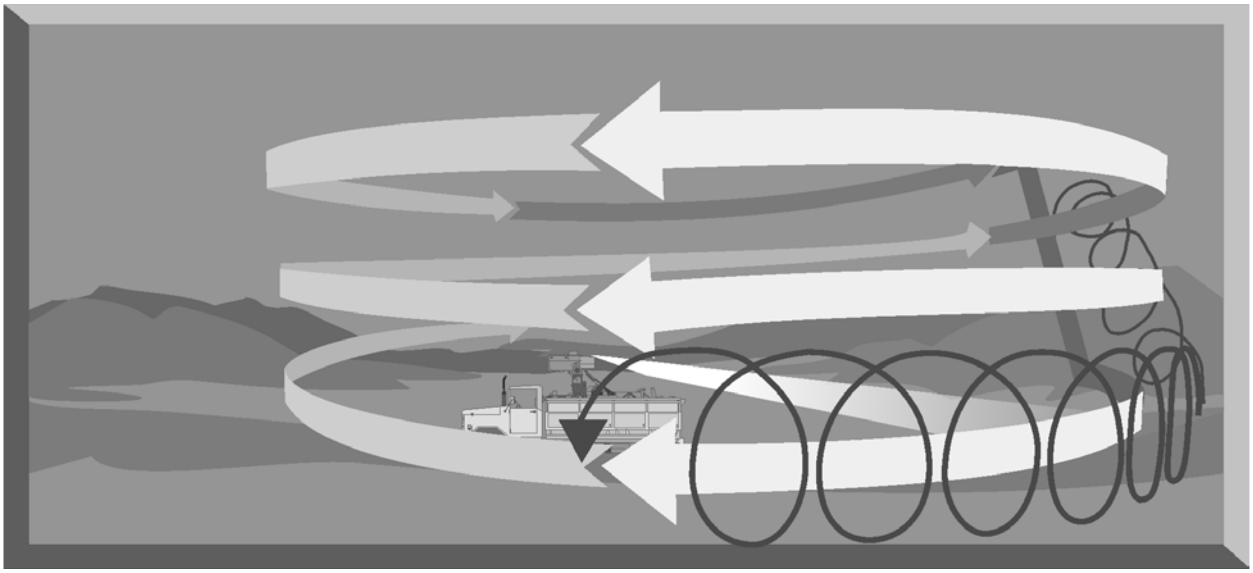

Figure 6-26. Palmer-Helical Scan

12. TRACK-WHILE-SCAN

A track-while-scan (TWS) system uses a technique that allows a radar to track one or more targets while scanning for others. Radar systems with a TWS capability must be able to generate two or more distinct radar beams.

a. A conventional TWS radar employs two antennas that work with each other to perform the scan function (Figure 6-27). Each antenna produces a separate unidirectional beam. Each beam is transmitted at a different frequency. The vertical antenna generates a beam employing a vertical sector scan similar to a height finder radar except the beamwidth is narrower and it scans at a higher rate. The horizontal antenna generates an identical beam employing a horizontal sector scan at a different frequency. The track function is accomplished in the area where the two beams pass through each other. A target that is within this center area is tracked, and positional information on range, elevation, and azimuth is updated each time the beams sweep through the area.

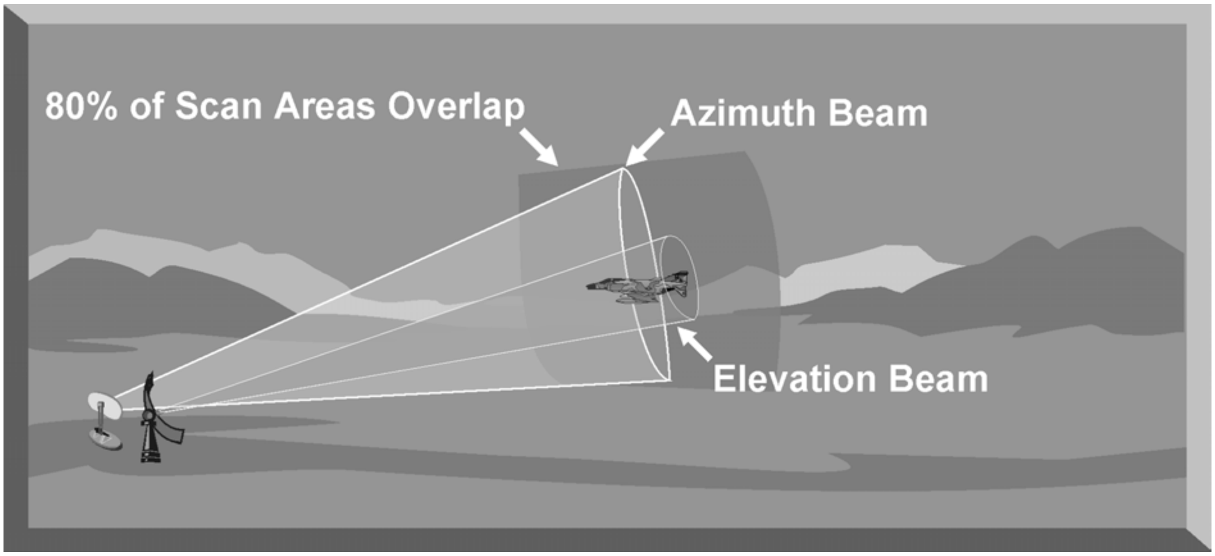

Figure 6-27. Conventional TWS Radar

b. The phased array radar is a product of the application of computer and digital technologies to the field of radar design. A phased array is a complex arrangement of many individual transmitting and receiving elements in a particular pattern (Figure 6-28). Common arrays include linear, planar, curved, and conformal, with linear being the most common. By using a computer to rapidly and independently control groups of these individual elements, a phased array antenna can, in effect, radiate more than one beam from the antenna. Multiple beams and computer processing of radar returns give the phased array radar the ability to perform the TWS function. The most common employment of the TWS capability of the phased array radar is in the air-to-air arena.

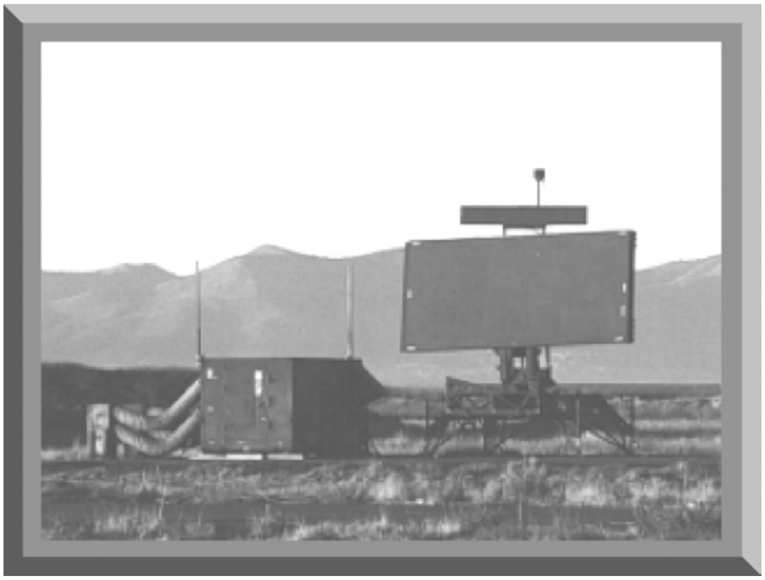

Figure 6-28. Phased Array Antenna

(1) The number of individual transmitting and receiving elements is limited by the size of the radar antenna. The number of targets a phased array radar can track is limited by the number of independent beams the antenna can generate. Many phased array radars, especially air-to-air radars, do not track and scan simultaneously, but rapidly switch between the two modes to overcome this limitation (Figure 6-29).

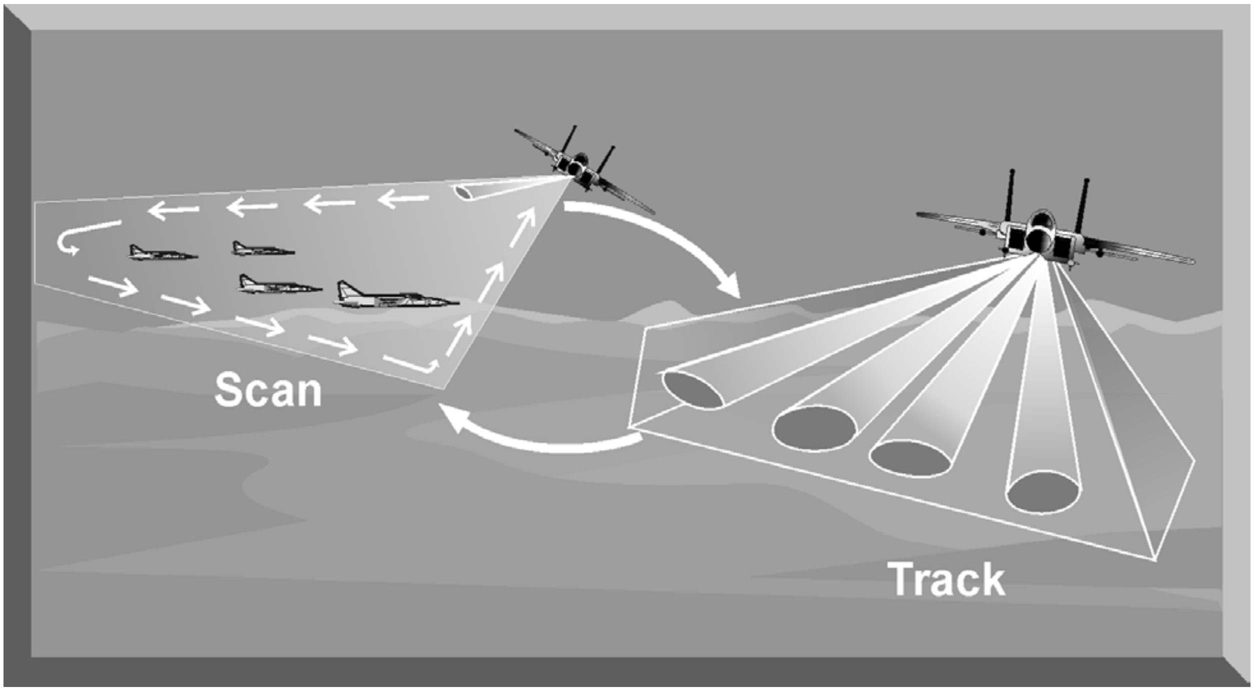

Figure 6-29. Phased Array TWS

(2) Modern TWS radars employ computer signal processing and complex computer algorithms to simplify the problem of target correlation (Figure 6-30). An air-to-air radar typically uses a raster scan to search a volume of airspace. In the search mode, the radar simply presents all targets detected in this airspace to the pilot on his radar display. In the TWS mode, the radar employs computer processing to figure out target correlation and update target information. This is done automatically, and the results are presented on the display.

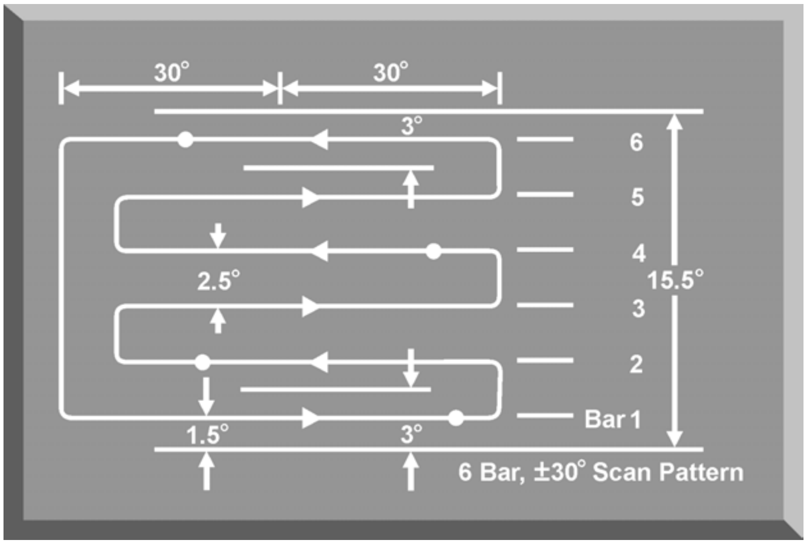

Figure 6-30. Phased Array TWS Radar Display

13. IMPACT OF TERRAIN ON RADAR SCANS

No matter what type of scan a radar system employs, terrain can limit radar line of sight (LOS) and target detection. The concepts of radar horizon, direct terrain masking, and indirect terrain masking are important factors in radar target detection and mission planning.

a. RF waves traveling in the atmosphere are bent, or refracted, and do not travel in a straight line. However, the degree of refraction depends on atmospheric conditions which vary significantly and are difficult to accurately quantify and predict. For these reasons, most radar computations are based on the assumption that RF waves travel in a straight line. The concept of the radar horizon is based on this assumption.

(1) The radar horizon shown in Figure 6-31 is the maximum range a radar system can detect a target due to the curvature of the earth. The distance (d) to the horizon for a radar antenna at a height (h) can be computed using Equation 6-6.

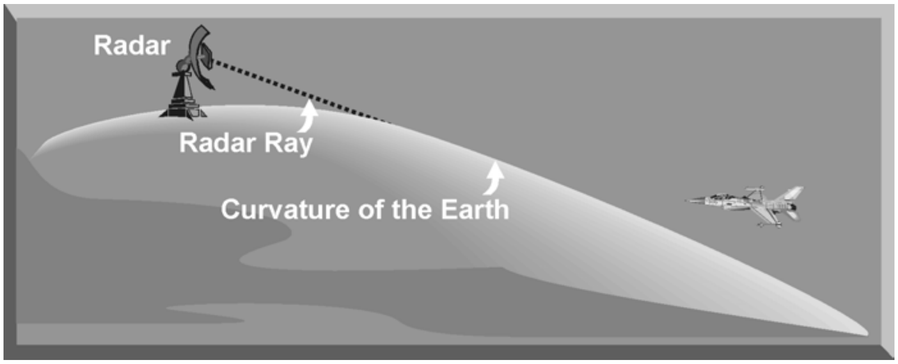

Figure 6-31. The Radar Horizon

$$\text{Distance to Horizon } (d) = \sqrt{2kah}$$

k = constant (4/3)
a = radius of earth
h = height of radar antenna

Equation 6-6. Basic Radar Horizon Equation

(2) In Equation 6-6, the constant, k, is commonly used to compensate for the assumption of straight line propagation of RF waves. Using the assumption that the height of the radar antenna is small compared to the radius of the earth, distance is measured in nautical miles (nm), and height is measured in feet, then Equation 6-6 reduces to Equation 6-7.

$$\text{Distance to Horizon } (d) = 1.06\sqrt{h}$$

Equation 6-7. Simplified Radar Horizon Equation

(3) Another application of Equation 6-7 is in calculating the range at which a radar antenna will achieve LOS with a low-altitude target. To compute this distance (D), Equation 6-8 can be used.

$$\text{Target Detection Distance } (D) = 1.06\left(\sqrt{h} + \sqrt{\text{target altitude}}\right)$$

Equation 6-8. Target LOS Distance

(4) To illustrate the use of Equations 6-7 and 6-8, consider the example of a radar antenna located at a height of 25 feet and a target aircraft flying at an altitude of 100 feet. From Equation 6-7, the radar horizon for this system would be 5.3 nm. From Equation 6-8, the radar antenna will have LOS with the target aircraft at 16 nm.

b. The previous discussions of the radar horizon assumes the radar is operating over water or level terrain. Radar operations over rough terrain can present other radar LOS limitations. As illustrated in Figure 6-32, prominent terrain features can limit radar detection. In this illustration, a mountain acts as a radar horizon and limits target LOS in one sector of the radar scan. This situation is called direct terrain masking. Placing prominent terrain features between the aircraft and threat radar systems effectively negates these systems and is an integral part of threat avoidance during combat mission planning.

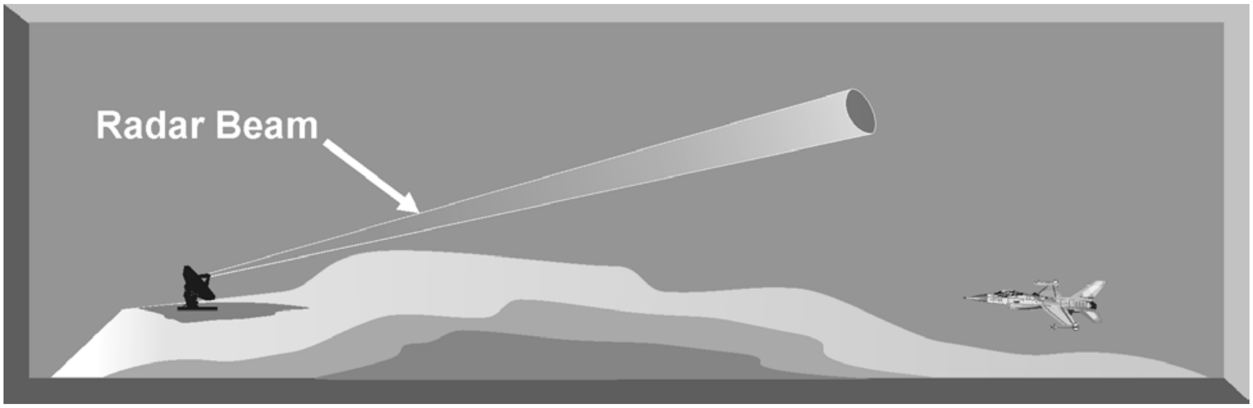

Figure 6-32. Direct Terrain Masking

c. Figure 6-33 depicts another impact of terrain on target detection/indirect terrain masking. When both the aircraft and a prominent terrain feature are illuminated by a radar beam, a pulse radar system may not be able to differentiate the target return from the return generated by the terrain. Indirect terrain masking is most effective when the aircraft is flying abeam the radar site. Pulse Doppler radars and radar systems employing moving target indicator (MTI) circuits may be able to negate the effectiveness of indirect terrain masking. However, indirect terrain masking is another important consideration for threat avoidance during combat mission planning.

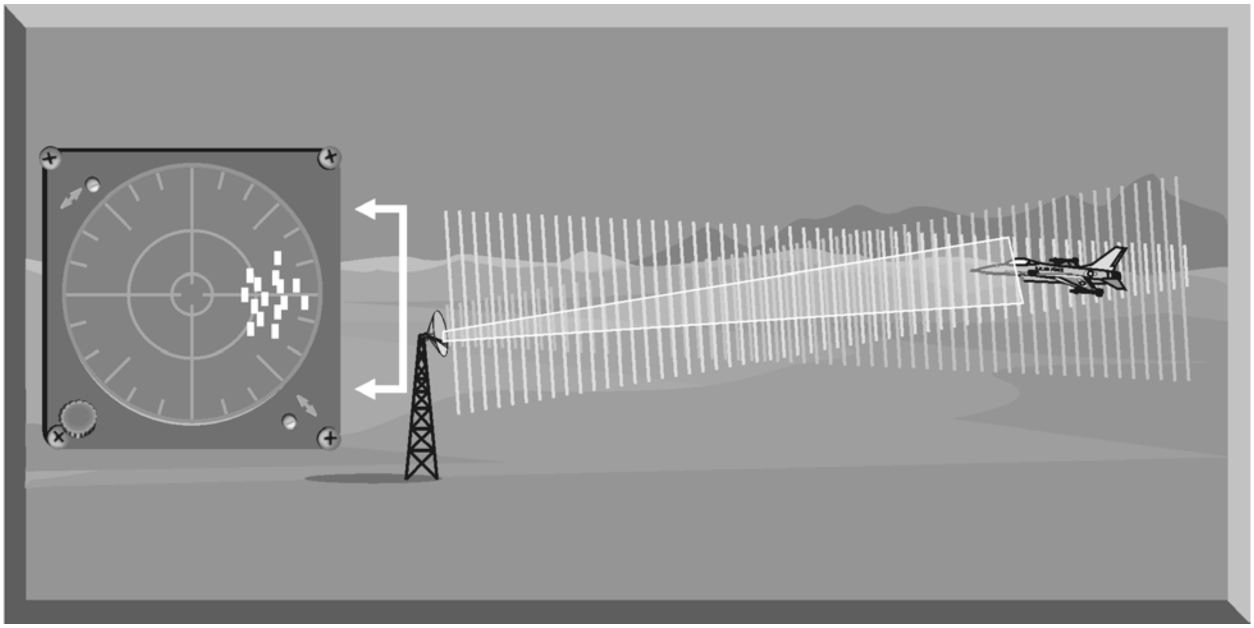

Figure 6-33. Indirect Terrain Masking

14. SUMMARY

This chapter has introduced basic antenna characteristics and how these characteristics influence their employment. The concepts of antenna gain and power density were also explained. The remainder of the chapter discussed the different types of radar scans. A radar's scan pattern is designed to enhance target detection and facilitate target tracking. The radar horizon, direct terrain masking, and indirect terrain masking are LOS limitations to all radar scans. The scan pattern chosen for a specific radar system determines the ability of that radar to provide accurate range, azimuth, elevation, or velocity of the detected target.

CHAPTER 7. TARGET TRACKING

1. INTRODUCTION

A target tracking radar (TTR) is designed to provide all the necessary information to guide a missile or aim a gun to destroy an aircraft. Once a target has been detected, either by a dedicated search radar or by using an acquisition mode, the TTR is designed to provide accurate target range, azimuth, elevation, or velocity information to a fire control computer.

a. A typical TTR has individual tracking loops to track a target in range, azimuth, elevation, or velocity. The antenna of the TTR is pointed at a single target, and the radar initiates acquisition and target track. TTRs normally employ automatic trackers to continuously measure target data. The range tracking loop employs an early gate/late gate range tracker to maintain automatic range tracking. The azimuth and elevation tracking loops generate error signals to position the antenna and maintain constant target illumination. The velocity tracking loop found on pulse Doppler and CW radars is used to reject clutter and generate accurate target radial velocity information. All this critical information is passed to a fire control computer for weapons employment.

b. The fire control computer is programmed with critical information on the capability of the weapon to be employed. For a missile, the fire control computer is programmed with the aerodynamic and range capabilities of the missile. For antiaircraft artillery (AAA), the fire control computer is programmed with the ballistics for the gun, rate of fire, and tracking rate. The fire control computer uses the precise target information from the TTR and the programmed weapon's parameters to compute a firing solution. Once a firing solution has been computed, the fire control computer either fires the weapon automatically or alerts the operator, who fires the weapon. For missile employment, the fire control computer may continue to provide missile guidance and fusing commands until missile impact or initiation of an active missile guidance mode. For AAA engagement, the fire control computer computes the required lead angle, aims the guns, and initiates firing.

c. To provide the required azimuth and elevation resolution, most TTRs use a high frequency to provide narrow antenna beamwidths for accurate target tracking. High frequency operation also allows the radar to transmit wide bandwidths. To provide the required range resolution, most TTRs employ narrow pulse widths and high pulse repetition frequencies (PRFs) to rapidly update target information. In this chapter, the target tracking techniques of conical scan, track-while-scan, lobe-on-receive-only, monopulse, Doppler radars, and pulse Doppler radars will be discussed.

2. RANGE TRACKING

In most TTR applications, the target is continuously tracked in range, azimuth, and elevation. Range tracking can be accomplished by an operator who watches an "A" scope presentation and manually positions a handwheel to maintain a marker over the desired target return. The setting of the handwheel is a measure of target range and is converted to a voltage used by the fire control computer. As target speeds and maneuvers increase, the operator may have extreme difficulty maintaining manual target range tracking. To avoid this situation, most TTRs employ an automatic range tracking loop. All pulse TTRs, which includes conical scan, track-while-scan, monopulse, and pulse Doppler radars, employ either a split gate or leading-edge automatic range tracking system. In a TTR, automatic range tracking serves two essential functions: (1) it provides the critical value of target range, and (2) it provides a target acceptance range gate that excludes clutter and interference from other returns. Since radar range is normally the first target discriminator used to initiate automatic target tracking, the second function is essential to the proper operation of the other tracking loops.

a. A range gate circuit is simply an electronic switch that is turned on for a period of time after a pulse has been transmitted. The time delay for switch activation corresponds to a specific range. Any target return that appears inside this range gate is automatically tracked. The most common type of automatic range tracking is accomplished by a split-gate tracker. Two range gates are generated as shown in Figure 7-1. The automatic range tracking loop attempts to keep the amount of energy from the target return in the early gate and late gate equal. The range tracking error is computed by subtracting the output of the late gate from the output of the early gate. The amount of the range tracking error signal is the difference between the center of the pulse and the center of the range gate. The sign of the error signal determines the direction in which the gates must be repositioned to continue to track the target.

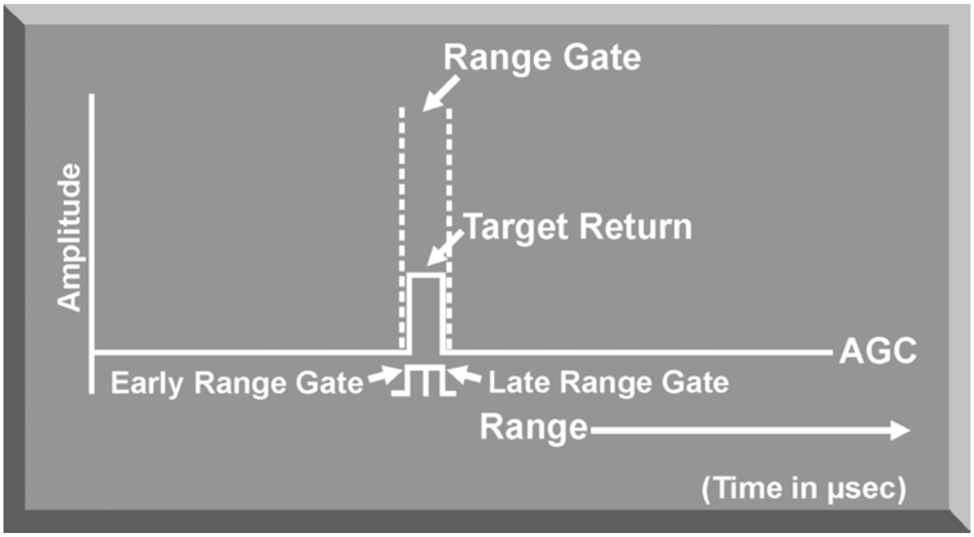

Figure 7-1. Split-Gate Range Tracker

b. Leading-edge range tracking is an electronic protection (EP) technique used to defeat range-gate-pull-off (RGPO) jamming. Figure 7-2 illustrates the application of leading-edge tracking. The leading-edge tracker obtains all range data from the leading edge of the target return. All RGPO cover pulse jamming tends to lag the target return by some increment of time (see (a) in Figure 7-2). By differentiating the entire return with respect to time, the target return can be separated from the jamming pulse (see (b) in Figure 7-2). Employing a split-gate tracker electronically positioned at the initial pan, or leading edge, of the returning pulse, the range tracking loop can track the target return and ignore any jamming signals. The range tracking loop then uses split-gate tracking logic to determine the magnitude and direction of range tracking errors and reposition the range gate.

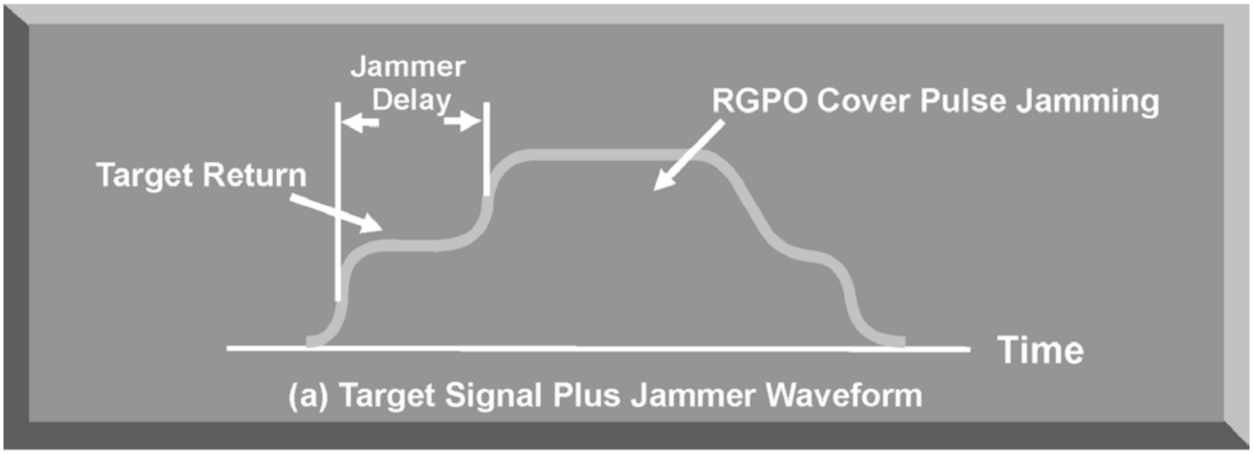

(a) Target Signal Plus Jammer Waveform

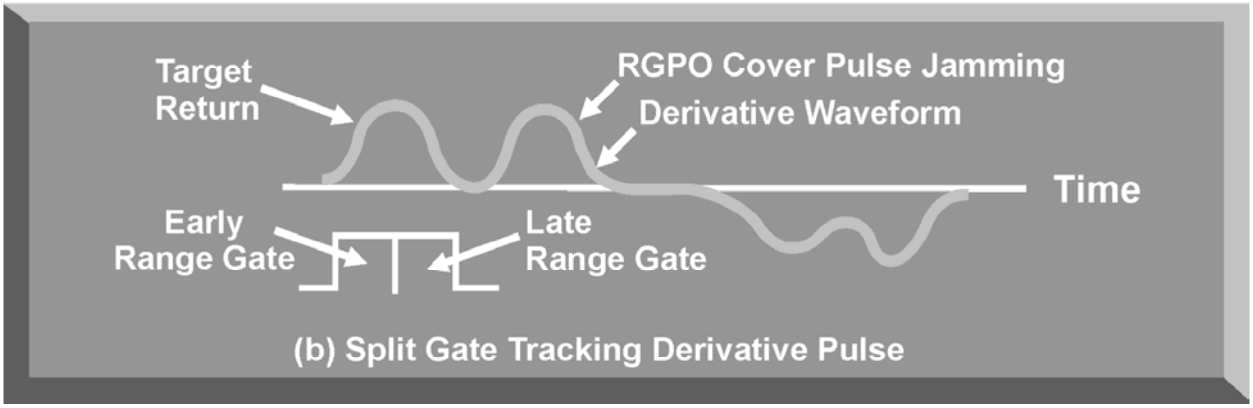

(b) Split Gate Tracking Derivative Pulse

Figure 7-2. Leading-Edge Range Tracker

c. The width of the tracking gate is an important radar design consideration. The range gate should be sufficiently narrow to effectively isolate the target from other returns at different ranges. It should be wide enough to allow sufficient energy from the target echo to be displayed. The width of the range tracking gate is normally equal to the pulse width of the radar.

d. Nearly all range tracking gates employ some form of automatic gain control (AGC). AGC is designed to limit target clutter and glint. It is also designed to avoid excessive false alarms.

3. CONICAL SCAN

A conical scan tracking system is a special form of sequential lobing. Sequential lobing implies that the radar antenna beam is sequentially moved between beam positions around the target to develop angle-error data. For a conical scan radar to generate azimuth and elevation tracking data, the beam must be switched between at least four beam positions as shown in Figure 7-3.

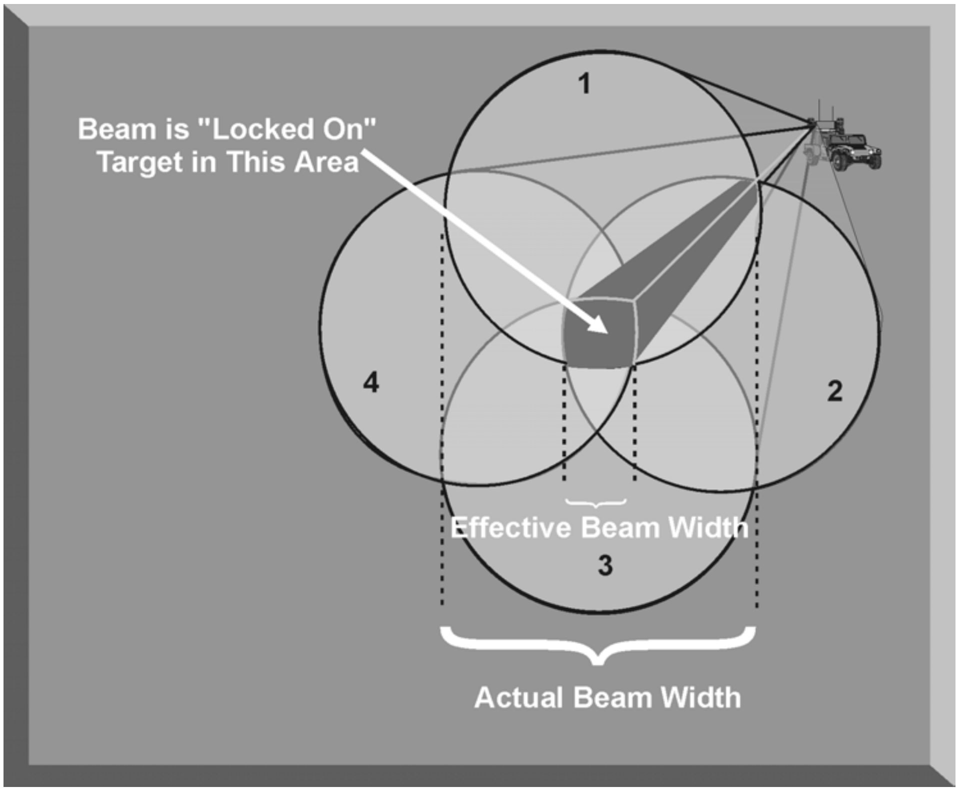

Figure 7-3. Conical Scan Positions

a. One of the simplest conical scan antennas is a parabola with an offset rear feed that rotates, or nutates, to maintain the plane of signal polarization. The radar beam is rotated at a fixed frequency around the target. The angle between the axis of rotation (normally the axis of the antenna) and the axis of the antenna beam is called the squint angle.

b. A conical scan radar first tracks the target aircraft in range. For azimuth and elevation tracking, the target return is modulated at a frequency equal to the rotation frequency of the radar beam. This results in a target signal output that resembles a sine wave (Figure 7-4). The azimuth and elevation tracking loops

drive servo motors to position the antenna to keep the energy level in each of the four positions equal. The amount of the modulated signal determines how far the target is off the antenna boresight while the phase of the modulation (positive or negative) determines the direction.

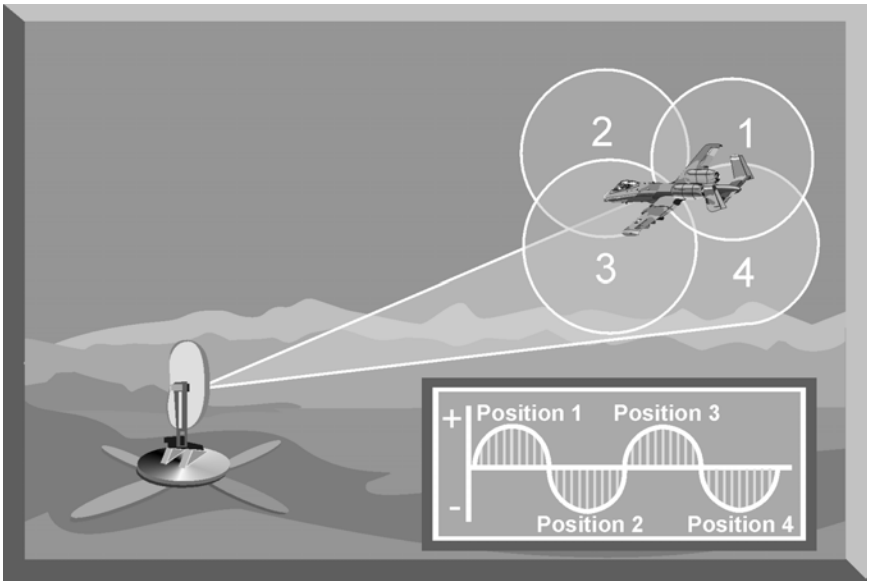
2
1
3
4
+
-
Position 1
Position 3
Position 2
Position 4

Figure 7-4. Conical Scan Modulation

c. In Figure 7-5, most of the target energy is in position 1, with a small amount of energy in position 4. The output from the elevation tracking loop is positive and drives the antenna servos upward. The output from the azimuth tracking loop moves the antenna to the right.

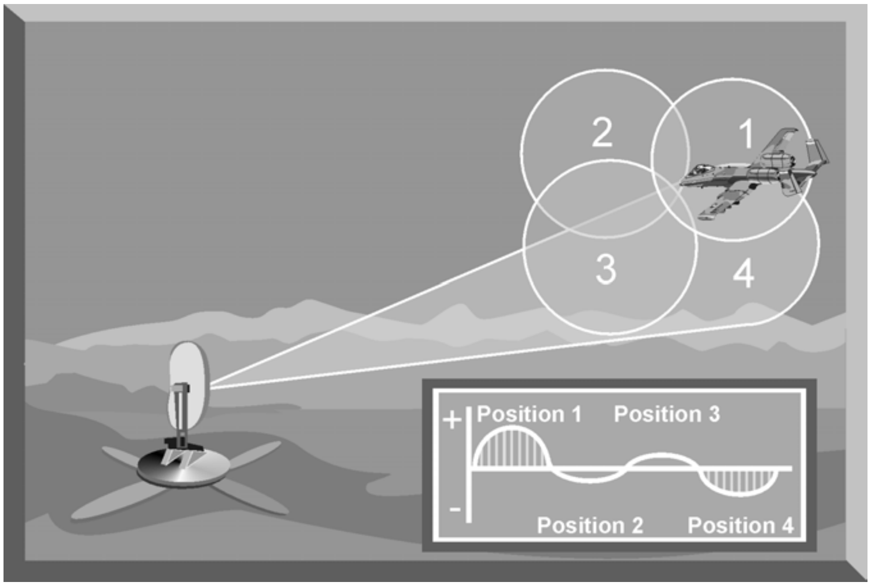
2
1
3
4
+
-
Position 1
Position 3
Position 2
Position 4

Figure 7-5. Conical Scan Tracking Errors

d. Once a balance of target energy in each scan position is achieved, the target is in the central tracking area (Figure 7-6). The azimuth and elevation tracking circuits continue to drive the antenna servos to maintain this energy balance which keeps the radar beam on the target.

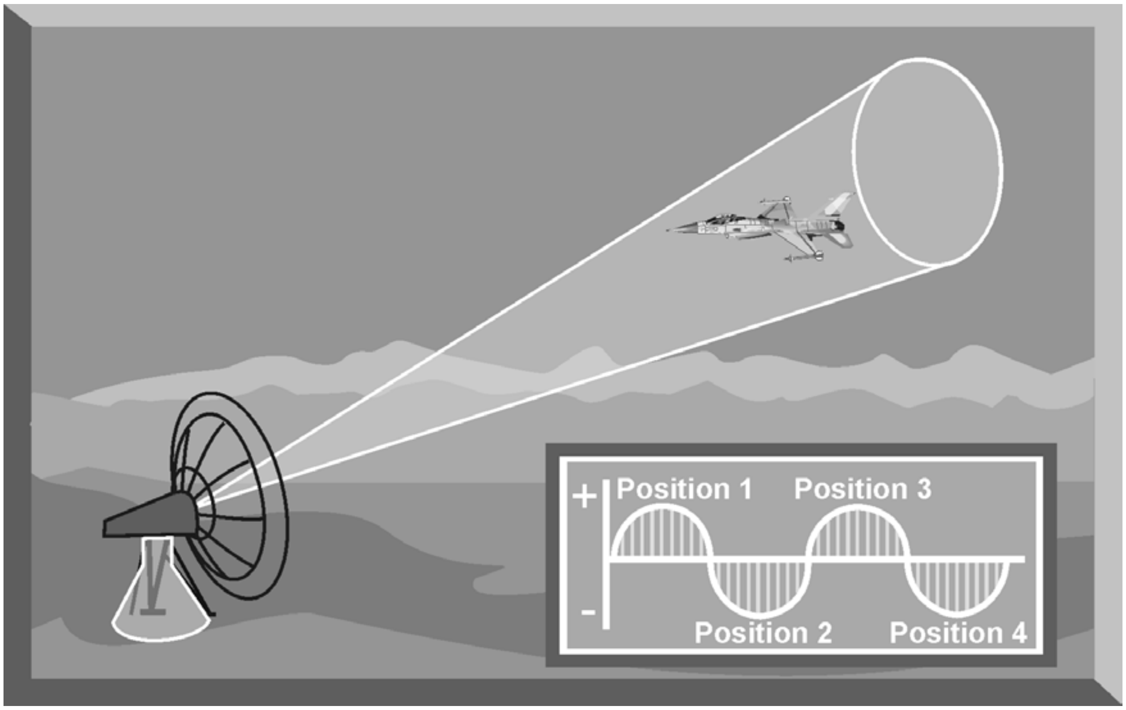

Figure 7-6. Conical Scan Tracking

e. The primary advantage of a conical scan radar is the small beamwidth which provides extremely accurate target tracking information. The primary disadvantages of conical scan include the following: (1) the narrow beamwidth makes target acquisition a problem. Even using a Palmer-helical scan, it may take considerable time to find and initiate track on a target; (2) conical scan radars are vulnerable to inverse gain modulation jamming based on the scanning frequency of the rotating beam; (3) a conical scan radar must analyze many radar return pulses to generate a tracking solution.

4. TRACK-WHILE-SCAN (TWS)

TWS is a combined search and tracking mode that sacrifices the continuous target observation capability of the dedicated tracker in return for the ability to monitor a finite sector of airspace. This is accomplished while maintaining tracks on multiple targets moving through the covered airspace. There are two types of radar systems capable of TWS operation: conventional and phased array.

a. Conventional track-while-scan threat radars use two separate antennas to generate two separate beams (Figure 7-7). These beams operate at two different

frequencies and are sectored so they overlap the same region of space. This overlap area provides a tracking area for a single target. One beam is sectored in the vertical plane to give range and elevation. The other beam is sectored in the horizontal plane to provide range and azimuth. Each beam scans its sector at a rate of 5 to 50 times per second. This provides a rapid update on target range, azimuth, and elevation.

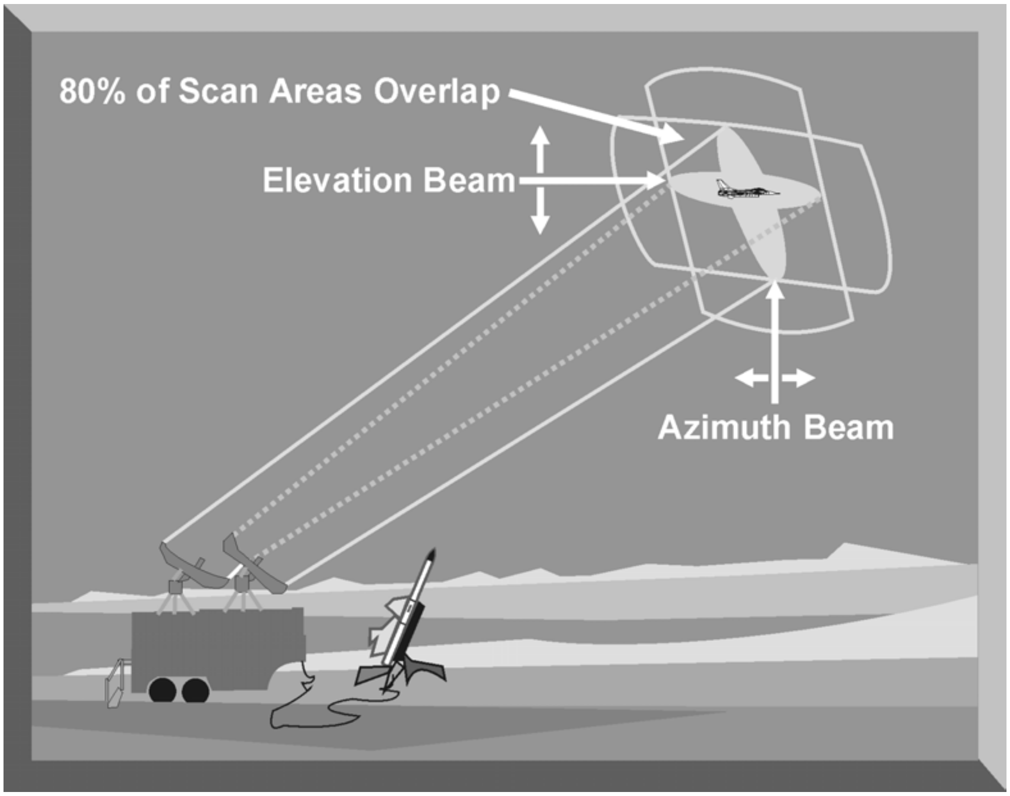

Figure 7-7. Conventional TWS Radar Beams

(1) The two TWS antennas generate their beams using an electro-mechanical principle. Each antenna provides inputs to its own display and provides angle and range information for all targets in the coverage of the radar. The display from the elevation beam is calibrated in range and elevation, while the display from the azimuth beam is calibrated in azimuth and range. Operators position a cursor over the returns on these displays using range as the primary parameter. Once a target has been designated for engagement, the radar automatically attempts to keep the tracking axis of the radar beams centered on the target.

(2) Once the target is designated by the operator, the range gate is enabled and tracks the target using a split-gate tracker. The azimuth and elevation tracking loops receive information only from targets inside the range gate. As the beams scan across the target, a burst of pulse returns is received that have an

amplitude envelope corresponding to the beam pattern. The azimuth beam pattern is shown in Figure 7-8.

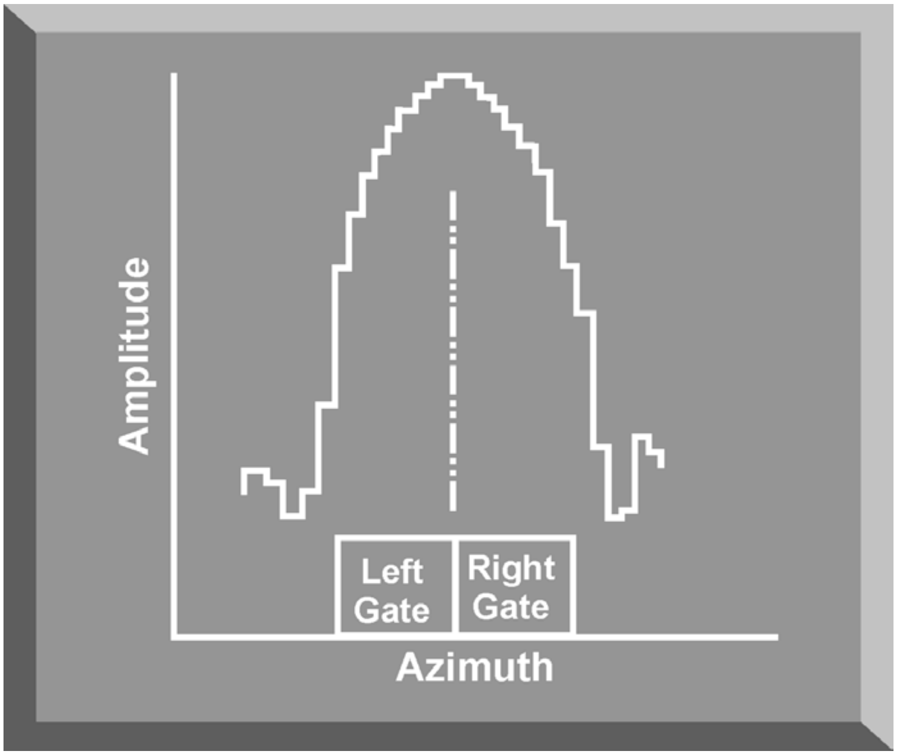

Figure 7-8. Conventional TWS Azimuth Tracking

(a) The azimuth tracker is typically a split-gate tracker, identical in concept to a split-gate range tracker. However, range delay time is replaced by azimuth scan time. The azimuth tracker uses a left gate and right gate. Each gate integrates its share of the target return to generate a voltage/time value. When the azimuth gate is centered on the target, the areas are equal and the error signal (right gate minus left gate) is zero. The azimuth tracking loop sends signals to the antenna servos to keep the target centered in the scan area.

(b) Elevation tracking is accomplished in the same manner by using an up gate and a down gate. The elevation tracking loop also sends signals to the antenna servos to keep the target centered in the scan area.

(3) Once the target is designated and the radar is automatically keeping the radar return centered in the tracking area, target range, azimuth, and elevation information is sent to a fire control computer. The radar continues to provide information on other targets in the scan area. The fire control computer indicates the firing solution has been achieved for the designated target, and a missile is launched. The radar tracks the target and the missile and provides in-flight corrections to the missile right up to the moment of missile impact. These corrections are based on both target and missile azimuth, range, and elevation

information. Information is passed to the missile from a dedicated antenna on the radar to special antennas on the missile. Commands from the radar to the missile are called uplink guidance commands. Information from the missile back to the radar and fire control computer is called downlink information.

(4) The advantages of a conventional TWS radar include the following: (1) TWS radars have the ability to maintain radar contact with all targets in the sector scan area while maintaining target track on a single target, and (2) the rapid sector scan rate provides a rapid update on target parameters. The primary disadvantages of a conventional TWS radar include: (1) a large resolution cell due to the wide azimuth and elevation beams, and (2) vulnerability to modulation jamming based on the scan rate of the independent beams.

b. Many modern radars employing a planar or phased array antenna system have a TWS mode. The radar does not really track and scan simultaneously, but rapidly switches between search and track (Figure 7-9).

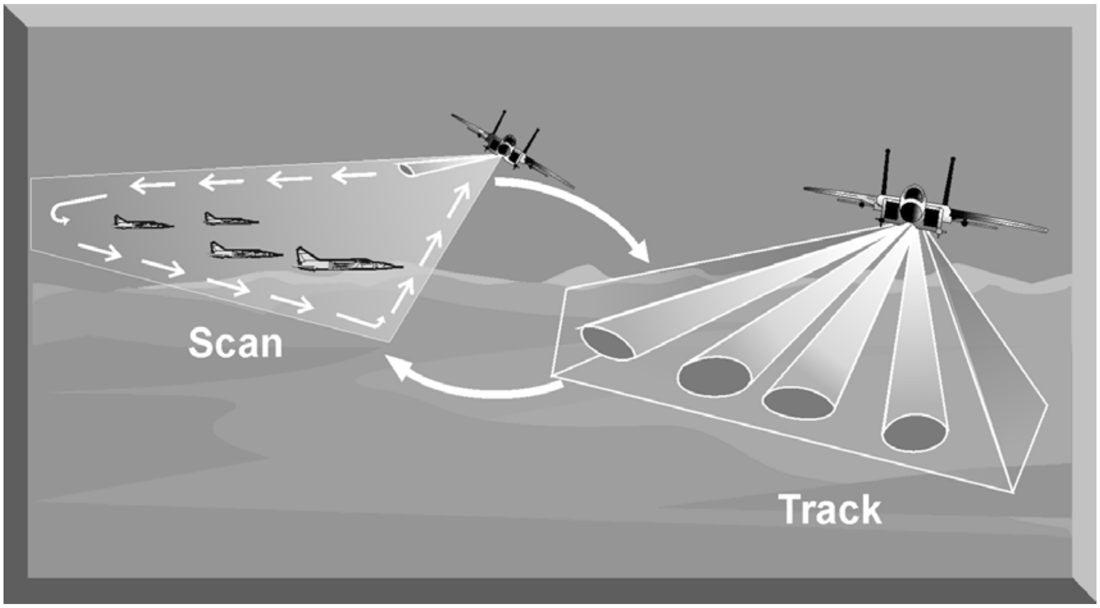

Figure 7-9. Phased Array TWS

(1) The most common air-to-air radar system uses a planar array antenna. In the scan mode, the radar antenna generates a pencil beam and uses a raster scan to detect targets in the search area. Targets detected are presented to the pilot on the aircraft's radar display (Figure 7-10).

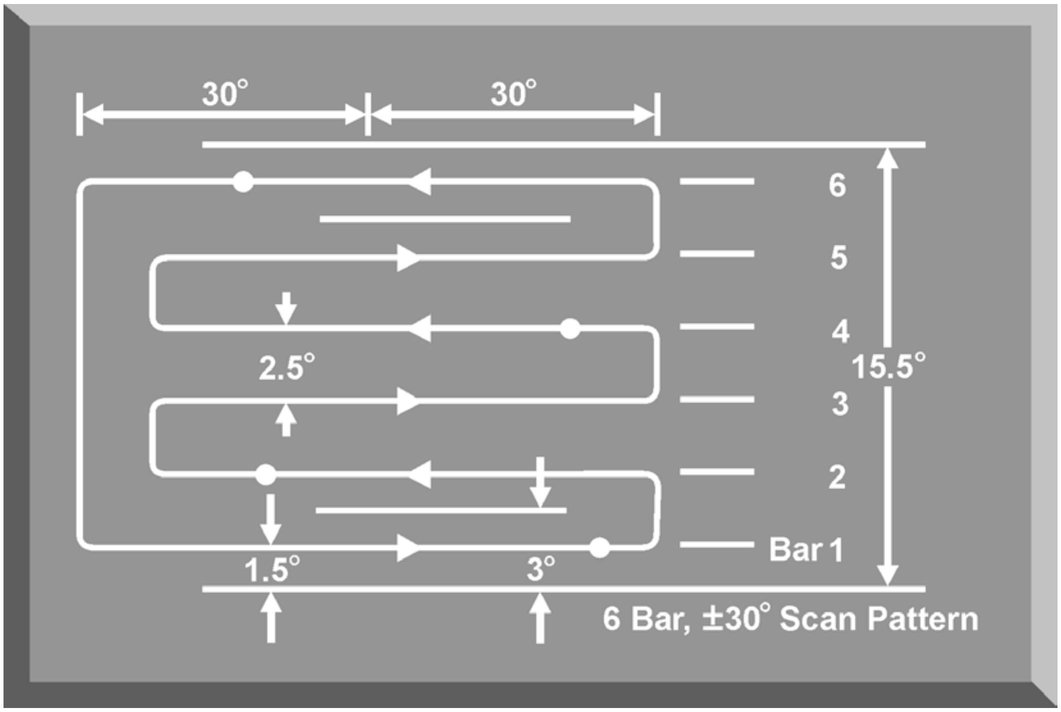
30°
30°
6
5
4
3
2
Bar 1
15.5°
2.5°
1.5°
3°
6 Bar, ±30° Scan Pattern

Figure 7-10. Air-to-Air TWS Display

(a) In the track mode, the antenna generates multiple beams to illuminate individual targets. The radar typically uses monopulse or pulse Doppler techniques to update target range, azimuth, elevation or velocity. These tracking techniques will be covered later in this chapter. The radar initiates a track file on each detected target that contains all current data on the target and an estimate of future target position.

(b) As the radar switches between track and scan modes, target parameters are updated in the tracking loop (Figure 7-11). The new target information is compared to the predicted information in the measurement data processing cell. If the two sets of data agree within certain limits, target position and information are updated. This process is called gating.

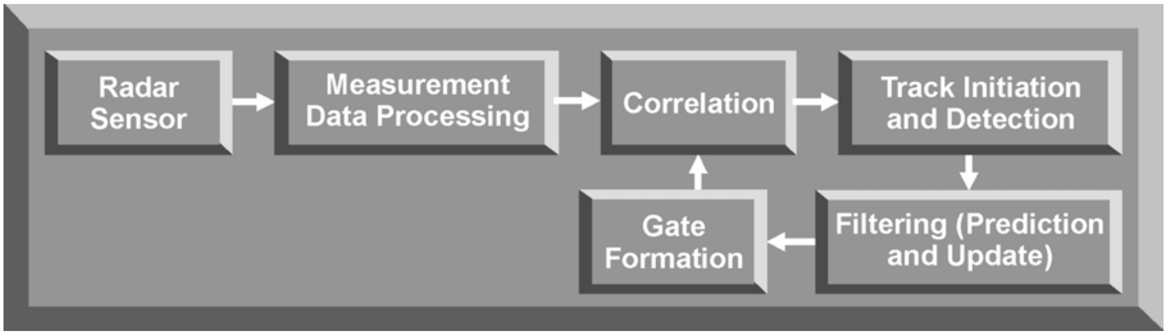
Radar Sensor
Measurement Data Processing
Correlation
Track Initiation and Detection
Filtering (Prediction and Update)
Gate Formation

Figure 7-11. Planar/Phased Array TWS Tracking Loop

(c) If the updated target information does not correspond to the predicted values, the information is sent to the correlation processor. The correlation processor attempts to resolve the conflict based on further refinement of target data. If the correlation processor cannot assign the target parameters to an existing track file, a new track file is generated and displayed.

(2) The obvious advantage of a planar/phased array TWS radar is that it can search a large volume of airspace while tracking individual targets. The number of targets that can be tracked is limited by the number of beams the radar can generate. Planar/phased array radars have increased peak and average power when compared to pulse radar systems. Since the radar beam of a planar/phased array radar is electronically controlled and can rapidly change beams and scans, it is resistant to many jamming techniques. The primary disadvantages of a planar/phased array TWS radar include its complexity, cost, and reliance on computer processing.

5. LOBE-ON-RECEIVE-ONLY (LORO)

LORO is a mode of radar operation developed as an EP feature for a track-while-scan radar. LORO can be employed by any radar that has the capability to passively track a target. In a LORO mode, the radar transmits a continuous signal from a set of illuminating antennas. This continuous signal hits the target, and the return echo is received by a different set of receive antennas (Figure 7-12). The receive antennas are passive and generate azimuth and elevation tracking signals by electronically scanning the reflected signal. The tracking signals are sent to the antenna servos to keep the illuminating antennas pointed at the target and centered in the receive antenna tracking area. The range tracking circuit uses the time delay between the transmission and reception of the illuminating antenna signals. A split-gate tracker is used to provide range tracking.

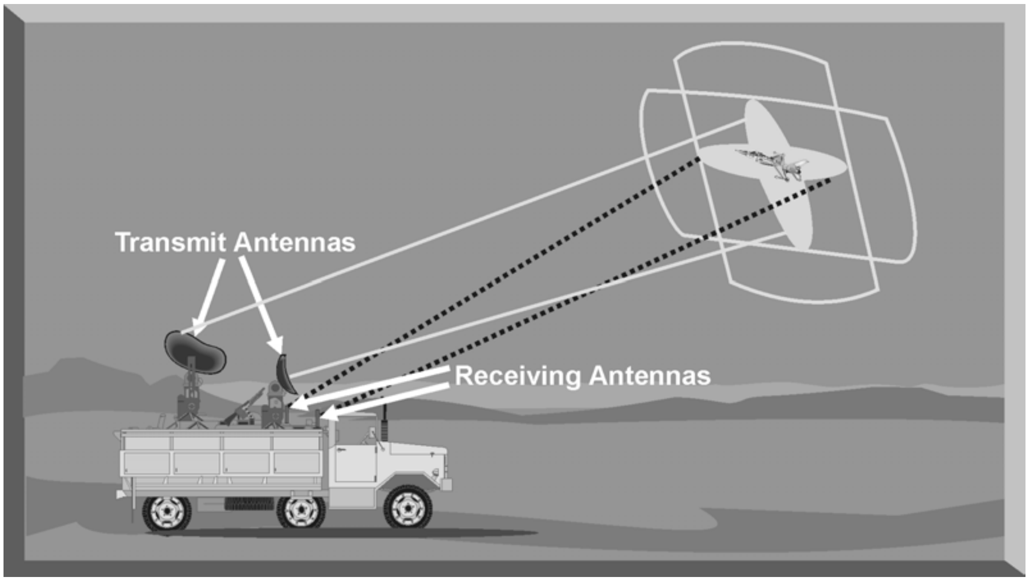

Figure 7-12. LORO Mode

a. The illuminating antennas used in the LORO mode have very narrow beamwidths and transmit at a high power level. This reduces the effectiveness of noise jamming techniques against a radar employing a LORO mode. In addition, the continuous signal from the illuminating antennas negate the effectiveness of most angle deception jamming techniques designed to defeat TWS radars. These specialized jamming techniques exploit the scan rate of TWS antennas. In the LORO mode, the illuminating antennas do not have a scan rate. The limited effectiveness of both noise and deception jamming techniques is the major advantage of the LORO mode.

b. The LORO mode also provides a track-on-jam (TOJ) capability to exploit noise jamming techniques. In a TOJ mode, the receive antennas passively track any detected noise jamming signals. The radar assumes that the most intense jamming signal is the target. The receive antennas process the strongest jamming signal as if it were a target echo from the transmit antenna signal. The receive antennas generate azimuth and elevation tracking signals to keep the jamming signal centered in the tracking area. The TOJ mode does not provide target range.

6. MONOPULSE RADAR

Monopulse radars are among the most complex radar systems. From a single pulse, a monopulse radar can derive all the data needed to update a target's position. It does this by comparing the relationship of two or more radar beams that are transmitted together from the same antenna but received separately. By comparing the phase or amplitude of the energy in these returned beams, target azimuth and elevation can be found. The speed that a monopulse radar updates the target's position, coupled with its azimuth/elevation accuracy and resistance to jamming, make this a popular choice amongst many newer TTRs.

a. The Magic T circuit allows monopulse radars to gather and process information from a single pulse that is transmitted and received using separate antennas. Figure 7-13 depicts a four-beam monopulse radar system. The Magic T is a sophisticated wave guide that can separate multiple signals by their phase relationships. This allows the radar tracking computer to compare the signal amplitude from the reflected pulses in several distinct ways. As the reflected energy enters the Magic T, it is separated by phase. The energy in the "H" arm will be in-phase and will exit from ports 1 and 2. The received energies entering the wave guide in the "E" arm exit at the number 1 port. This energy is exactly 180° out-of-phase with energy entering the H arm. This ensures there is no transfer of energy between the E and H plane arms. A typical monopulse radar would have eight Magic T's.

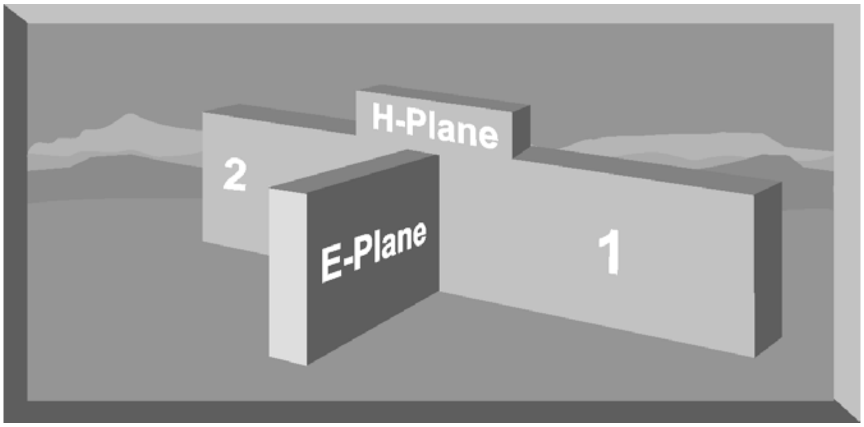

Figure 7-13. Monopulse Magic T

b. The output of a Magic T is the sum and difference of the two signals. These sum and difference values in amplitude or phase are used to generate azimuth and elevation error signals as well as to compute range. Monopulse radars may split the incoming signal as depicted in Figure 7-14. Upper antennas receive the A and B signals. Lower antennas receive the C and D signals. The various combinations of signals are processed and compared by simple addition and subtraction of the signal characteristics. From these steps, azimuth, range and elevation data are computed.

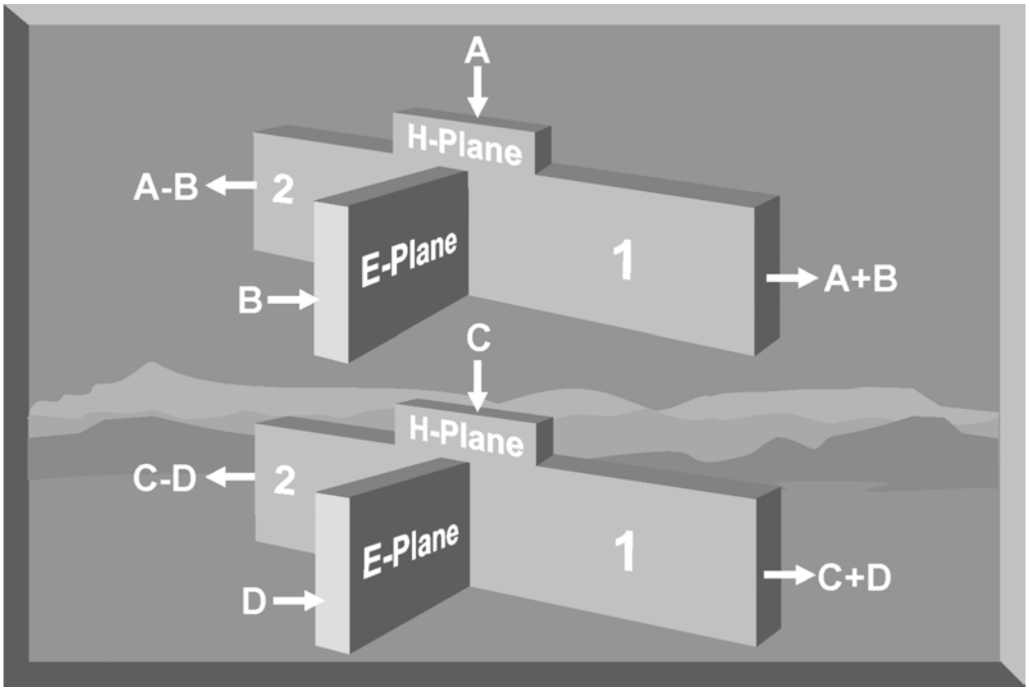

Figure 7-14. Magic T Output Signals

(1) The top equation in Figure 7-15 is used to compute target range. Target range is derived by adding the signal from the A scan to the signal from the B scan. This value is then added to the sum of the C and D scan signals. The output of these combinations is then passed to the range circuit which figures out the range of the target and displays it to the operator. Monopulse range tracking is accomplished by using either a leading-edge or split-gate tracking loop.

(2) Target elevation tracking error is derived using the middle equation from Figure 7-15. The signal from the A scan is added to the signal from the B scan. The signals from the C and D scans are also added. This time, however, the sum of A+B is subtracted from the sum of C+D. This value is then passed to the elevation circuit. Elevation signals are sent to the operator display and the servo mechanism, which corrects to the updated position of the target.

(3) The bottom equation from Figure 7-15 is used to compute the azimuth error. The signal from the A scan is added to the signal from the C scan. The B and D scans are also added together. The sums of these values are subtracted from each other. This difference equals the tracking error in azimuth. The radar system will then position the search beam to even the energy level between the two pairs of sums. When this occurs, the azimuth tracking error is zero.

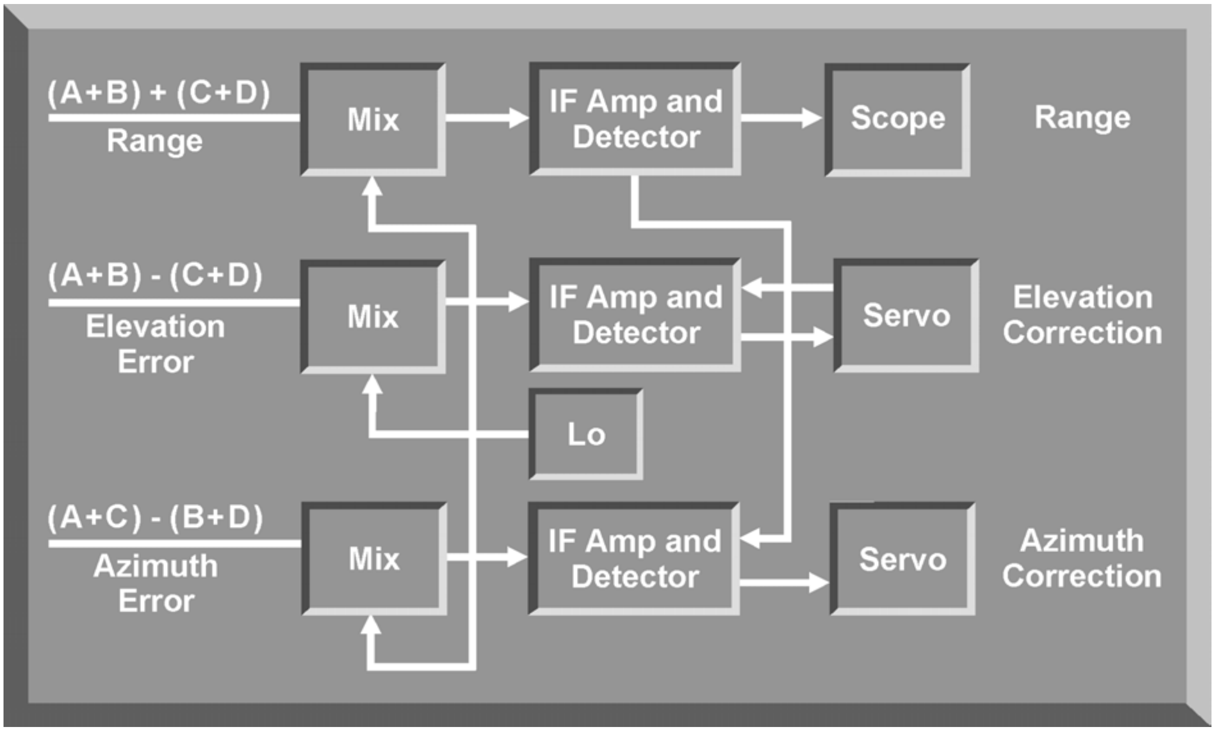
(A+B) + (C+D)
Range
Mix
IF Amp and Detector
Scope
Range
(A+B) - (C+D)
Elevation Error
Mix
IF Amp and Detector
Servo
Elevation Correction
Lo
(A+C) - (B+D)
Azimuth Error
Mix
IF Amp and Detector
Servo
Azimuth Correction

Figure 7-15. Monopulse Tracking Loops

c. A further illustration of the idea of signal combinations can be seen by referring to the F-16 in Figure 7-16. All the energy is received in the B scan area.

The A scan signal is added to the B scan signal. The signals from the C and D scans are also added. However, the sum of A+B is now subtracted from the sum of C+D. In this case, the values from the A scan and the C scan are zero. This total value of (A+B) - (C+D) is then passed to the elevation circuit.

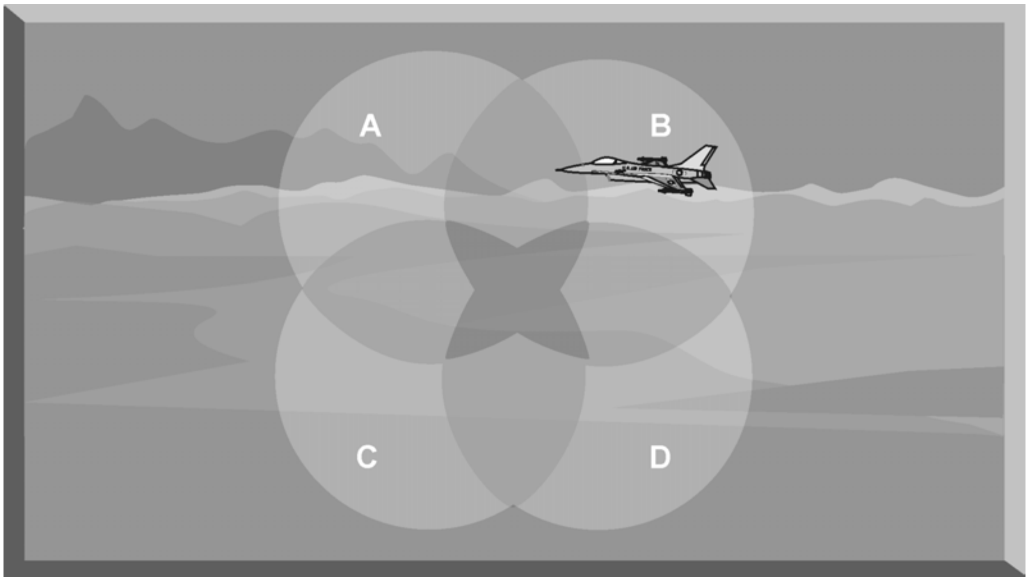

Figure 7-16. Monopulse Elevation Tracking Error

(1) The comparison in Figure 7-17 shows that all the energy is in the B scan. The radar will reposition the scan vertically to balance the energy between the B and D scans. When the energy level is balanced, the elevation error is zero.

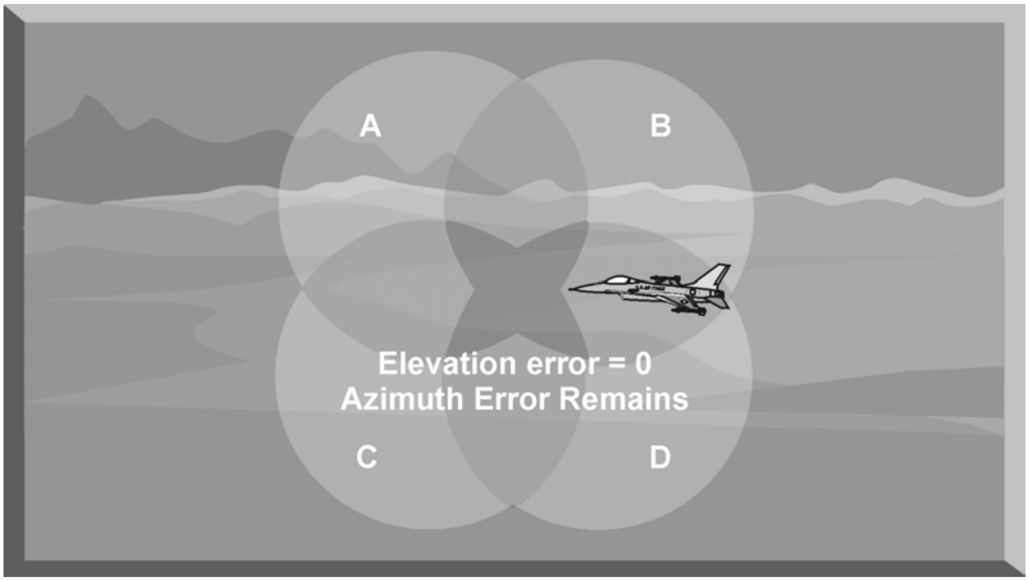

Figure 7-17. Monopulse Elevation Track

(2) Using the azimuth error equation from Figure 7-15, the azimuth tracking loop computes the azimuth error and repositions the antenna to equalize the received energy in all the beams. The monopulse radar has now established a tracking solution (Figure 7-18). All these computations are done instantaneously on a pulse-to-pulse basis.

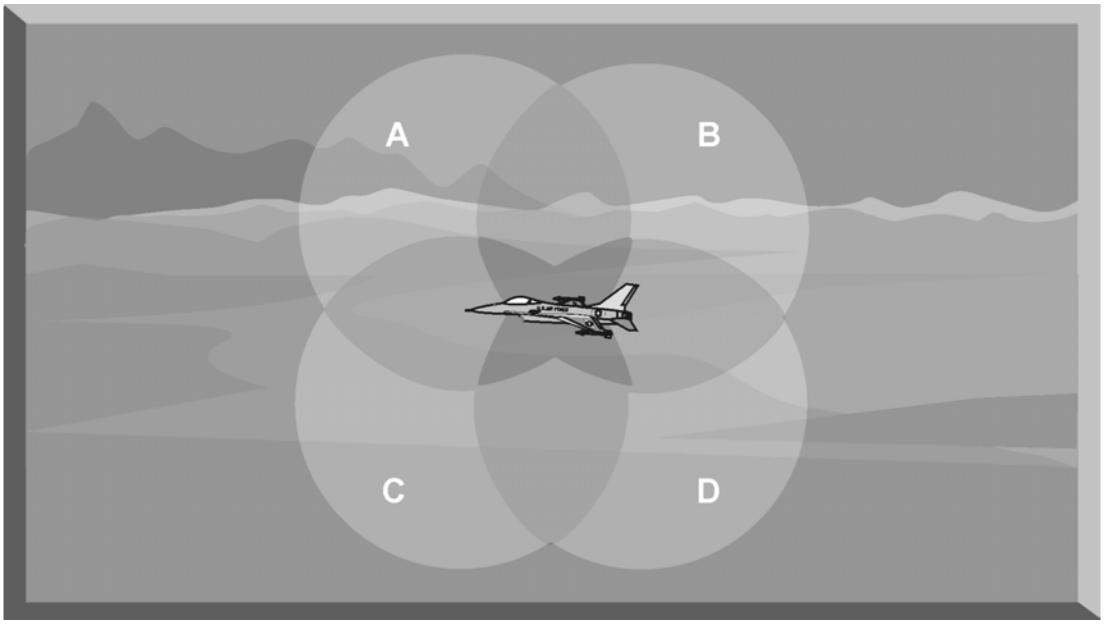

Figure 7-18. Monopulse Azimuth Track

7. CONTINUOUS WAVE (CW) RADARS

CW radar was one of the earliest forms of radar systems. Unlike pulse radar systems, CW radars emit a continuous beam of RF energy with no interruptions in the transmissions to detect returning echoes. A continuous radar transmission from the antenna requires that a classic CW radar have two antennas, one for transmission and one for reception (Figure 7-19).

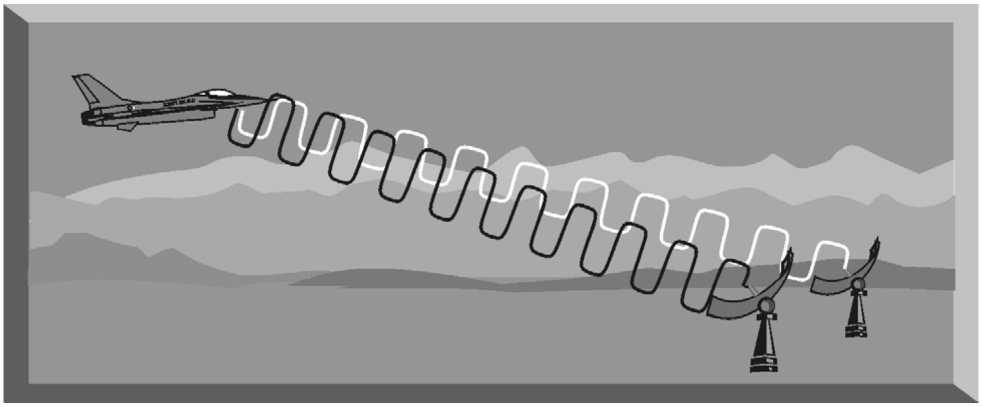

Figure 7-19. CW Radar

a. Since a continuous transmission results in a continuous echo signal, it is impossible to tell what part of the echo is associated with any particular part of the transmission. This makes range typical determination impossible. The azimuth and elevation tracking capability of a CW radar is based on the antenna position when the target is illuminated. The simple application of the Doppler principle provides a means for a CW radar to track a target in velocity and reject clutter. The Doppler principle deals with the fact that a radar return from a moving target will be shifted in frequency by an amount proportional to its radial velocity compared with the radar site. Using the difference in frequency from the transmitted signal to the received signal, a CW radar can separate the target return from clutter based on a change in frequency. This type of radar is called a CW Doppler radar.

b. The most serious disadvantage of a simple CW Doppler radar is that it does not provide any range information on the target. One method of obtaining range information from a CW radar uses frequency modulation (FM). The modulation can be sinusoidal, sawtooth, triangular, or any shape, as long as the rate of frequency change is known. The transmitter emits a continuous signal, but the frequency is changed in a known pattern. When the echo returns from the target, it is then compared to the frequency being transmitted. This frequency difference is directly proportional to the range of the target.

(1) Figure 7-20 shows how a FM CW radar measures range. Using a triangular wave for modulation, a plot of transmitted frequency over time would look like the solid line. It is important to note that this is not a depiction of the transmitted wave but a plot of how the frequency of the wave varies with time. For a target, without any relative motion, the frequency returning to the receiver is depicted by the dotted line. The target echo frequency lags the transmitted frequency by time (t). There is also a frequency difference between the transmitted and received signals. Range to the target may be computed by measuring this frequency difference and dividing by the rate of change of the transmitted frequency. The result is time. Dividing this time by 12.4 microseconds per radar mile yields range to the target. The frequency difference is constant for a target, without any relative motion, except for the brief intervals when the change in frequency goes from a positive to a negative slope. These “ditches” are negligible and can be disregarded when calculating range. The average amplitude is equal to target range.

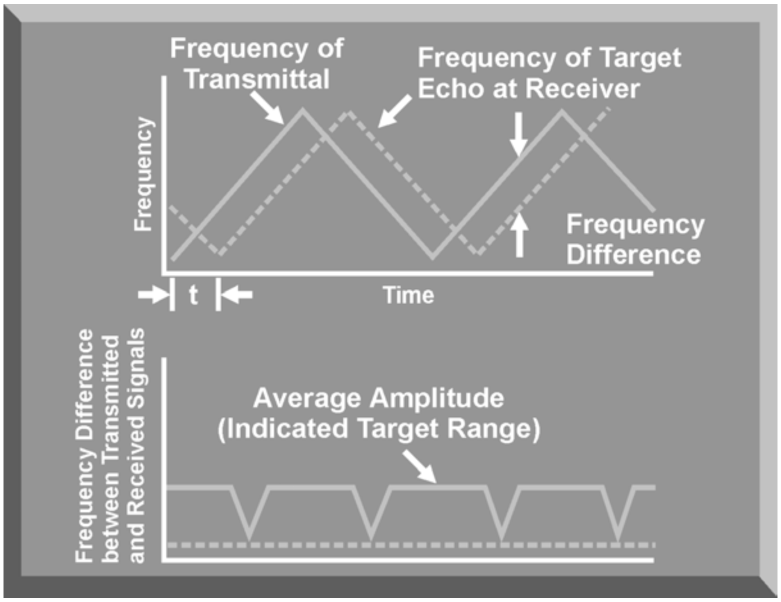
Frequency of Transmittal
Frequency of Target Echo at Receiver
Frequency
Frequency Difference
t
Time
Frequency Difference between Transmitted and Received Signals
Average Amplitude (Indicated Target Range)

Figure 7-20. FM CW Radar Range Determination (No Relative Motion)

(2) For a target moving toward the radar, the FM CW radar measures both target range and velocity. In Figure 7-21, the frequency of the return signal will be increased as depicted by the dotted line. Remember, this is a graph of frequency versus time, not a depiction of the radar wave. For a moving target, this results in a varying frequency difference. A plot of the frequency difference over time provides target range by averaging the difference, while target velocity is found by comparing the two frequency differences.

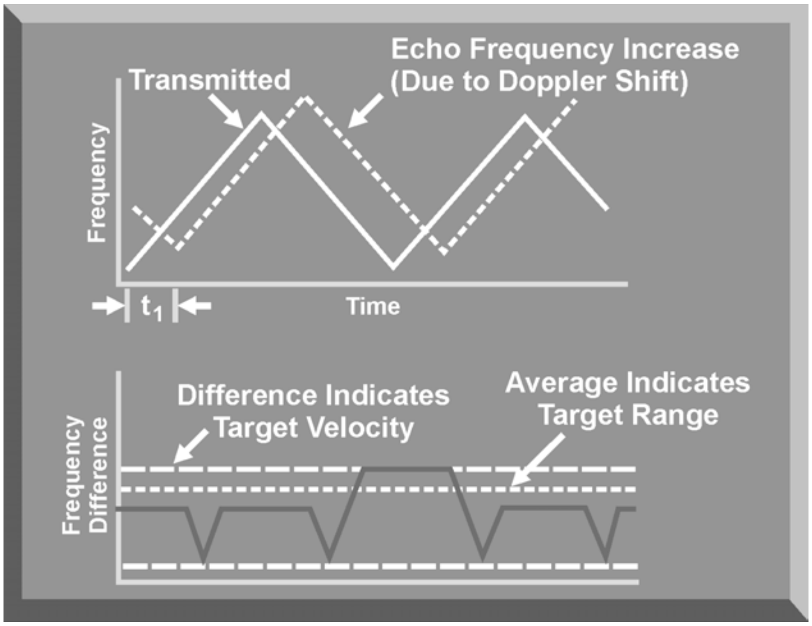
Transmitted
Echo Frequency Increase (Due to Doppler Shift)
Frequency
t1
Time
Frequency Difference
Difference Indicates Target Velocity
Average Indicates Target Range

Figure 7-21. FM CW Radar Range Determination (Relative Motion)

c. The primary advantage of the FM CW Doppler radar is its ability to combine the clutter rejection features of a simple CW Doppler radar with the capability to detect range. The widest application of FM CW Doppler radars is in radar altimeters for aircraft. In addition, the HAWK missile system uses an FM CW Doppler acquisition radar and a CW target illuminator.

d. Another method used for clutter rejection in a pulse radar system is to employ special circuits, or Doppler processing, to identify and reject clutter. These special circuits are added to the receiver section of a pulse radar and are called moving target indicators (MTIs). There are two types of MTIs, non-coherent and coherent.

(1) The earliest form of MTI was called non-coherent MTI or “area” MTI. Non-coherent MTI radars do not process Doppler frequencies. The returns from one scan are subtracted from returns from the next scan. All targets that move at least one resolution cell in the time between scans are displayed. All stationary objects, including fixed clutter, are cancelled and not displayed. In this type of MTI, clutter cancellation is based on the size and movement of the return. Due to changes in the clutter cross section, instabilities in radar operations, variations such as rain or clouds, and noise from the transmitter, clutter cancellation is never complete. In another form of non-coherent MTI, the radar returns from moving targets are compared to the returns from fixed targets, and the fixed targets are cancelled. These non-coherent MTIs are simple, but they do not provide the clutter rejection available from coherent MTI radars.

(2) A coherent MTI uses the fact that Doppler shifts appear to a pulse radar as phase shifts on the received target pulse. Coherent MTI uses sophisticated circuitry, including stable local oscillators (STALOs) and coherent local oscillators (COHOs) to capture and process these phase shifts. Further processing of these phase shifts yields velocities for each return. Those velocities associated with stationary targets are rejected and only moving targets are displayed. Coherent MTIs have a major problem called “blind speeds.” Blind speeds occur for all target Doppler frequencies that are the exact PRF, or any multiple of the PRF, of the radar signal. When a target is moving at a velocity that produces this Doppler frequency, its return is cancelled along with fixed returns.

(3) There are three techniques to limit “blind speeds.” The first technique is to use a PRF stagger. By staggering the PRF, the blind speed associated with one PRF will be covered by the other PRF. The second method is called the delay line and canceler. This involves delaying each pulse so it can be compared to the next pulse before processing. This method enhances Doppler frequency comparison and rejects clutter more effectively. The third way is to use range gates and Doppler filters. A range gate is simply a switch that opens for a time corresponding to the time a radar return would arrive for a target at a specific range. For example, a range gate for all targets between 10 and 11 nautical miles would open 124 microseconds after the transmitted pulse (12.4 microseconds per mile) and close 12.4 microseconds later. A target return at this range would trip

this gate and be processed by Doppler filters to find velocity. Fixed targets would trip the range gate and be eliminated by the Doppler filters. This is a sophisticated technique that is also used by pulse Doppler radars.

8. PULSE DOPPLER RADAR

Pulse Doppler radars combine the advantages of both pulse and Doppler radar systems. Because the signal is pulsed, the radar can find range, azimuth, and elevation, similar to a conventional pulsed radar. A pulse Doppler radar can also compute overtake, or rate of closure, compared with the radar system on a pulse-to-pulse basis. A pulse Doppler radar operates much like an MTI, and the terms are sometimes used interchangeably. However, an MTI uses Doppler frequency shifts only to reject clutter while a pulse Doppler radar uses Doppler frequency shifts to reject clutter and to track targets in velocity. A pulse Doppler radar transmits a box or pulse of RF energy at the operating frequency of the radar (Figure 7-22). The frequency inside these boxes reacts the same way as the continuous waves of a CW radar, but since the RF waves are pulsed, range determination can be accomplished by measuring the time it takes for the reflected pulse to return from the target. Velocity determination and tracking are accomplished by capturing and quantifying the Doppler shift of the frequencies in each pulse.

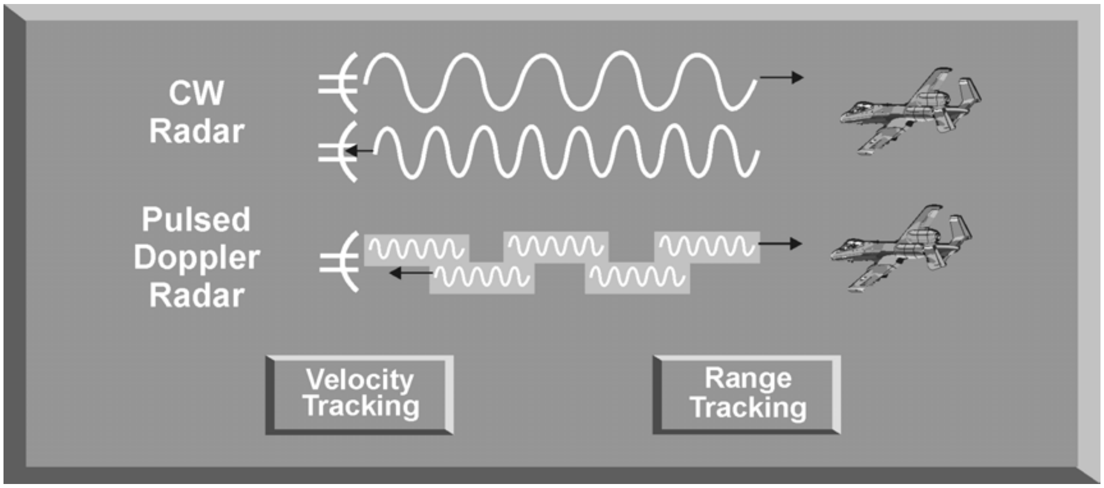

Figure 7-22. Comparison of CW Radar and Pulse Doppler Radar

a. A pulse Doppler radar tracks a single target in azimuth and elevation by employing either conical scan (sequential lobing) or monopulse tracking. Angle and elevation tracking employing these techniques is covered in Sections 3 and 6 of this chapter. The error signals generated by these techniques are sent to the antenna servos to keep the target return centered in the antenna beam.

b. Range tracking of a single target in a pulse Doppler radar is normally accomplished by a split-gate or leading-edge range tracking loop. Some pulse

Doppler radars employ an FM technique to provide range information during high PRF operations.

c. For velocity tracking, each range gate has a complete set of Doppler filter banks as depicted in Figure 7-23. Each pulse of RF energy is composed of many frequencies. To separate the returning target frequency from all other frequencies in the returning waveform, the pulse Doppler radar employs filters to cancel the unwanted frequencies. In addition, it cancels out all returns with no frequency shift, which equates to canceling all returns with no movement relative to the radar.

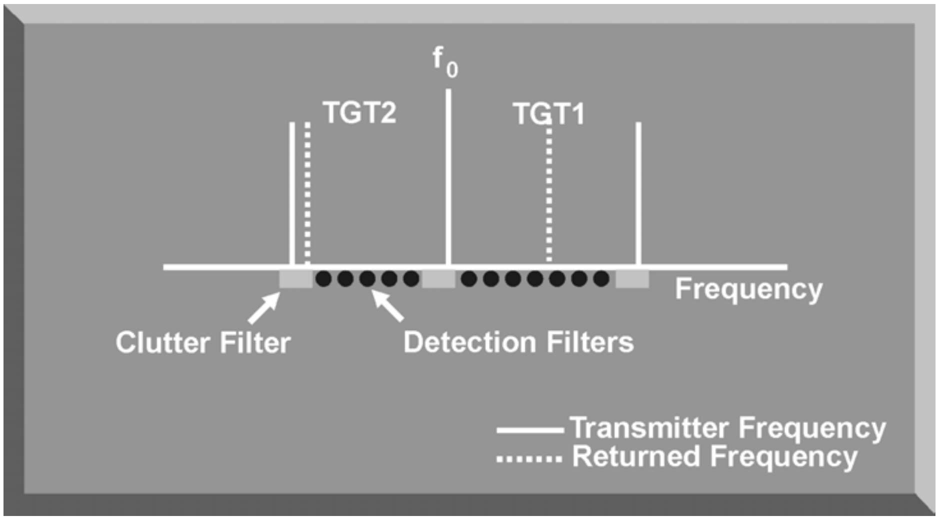

Figure 7-23. Range Ambiguity

d. The ability of a pulse Doppler radar to accomplish range, azimuth, and velocity tracking is dependent on the PRF of the radar. Table 7-1 summarizes these capabilities based on PRF. Low PRF tracking in velocity is extremely difficult due to the spacing of the spectral lines. Low PRF gives accurate range and azimuth as well as long, unambiguous range. Medium PRF tracking in range, azimuth, and velocity is easy for the radar to handle. High PRF provides excellent velocity resolution, but range ambiguities become a problem.

Table 7-1. Pulse Doppler Tracking Capabilities

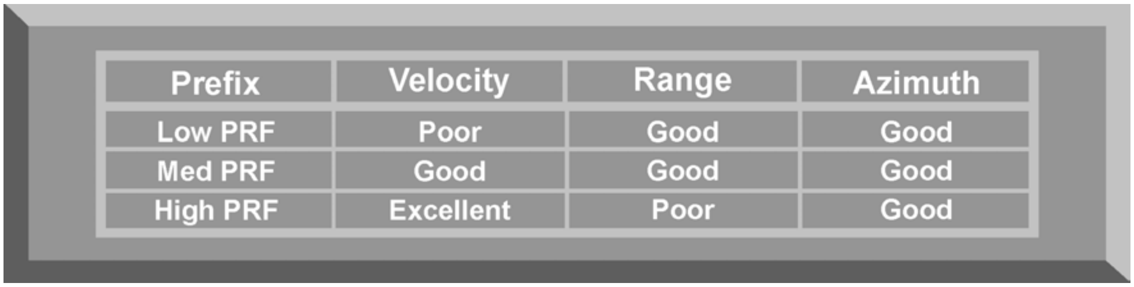

Prefix	Velocity	Range	Azimuth
Low PRF	Poor	Good	Good
Med PRF	Good	Good	Good
High PRF	Excellent	Poor	Good

e. Weaknesses of pulse Doppler radars include velocity blind speeds, range ambiguity, and range eclipsing.

(1) The primary operator exploitable weakness of a pulse Doppler radar takes advantage of the pulse Doppler radar's biggest strength. The pulse Doppler is designed to eliminate ground returns so that the attacker is able to track an aircraft that used to be able to hide in ground clutter. To remove the ground clutter and avoid tracking unwanted targets like cars on a road, a filter is designed in the radar to eliminate all targets with a low velocity relative to the radar. The key to breaking track of a pulse Doppler radar is to place the aircraft in a speed less than the speed gate relative to the radar (Figure 7-24), commonly referred to as the Doppler notch.

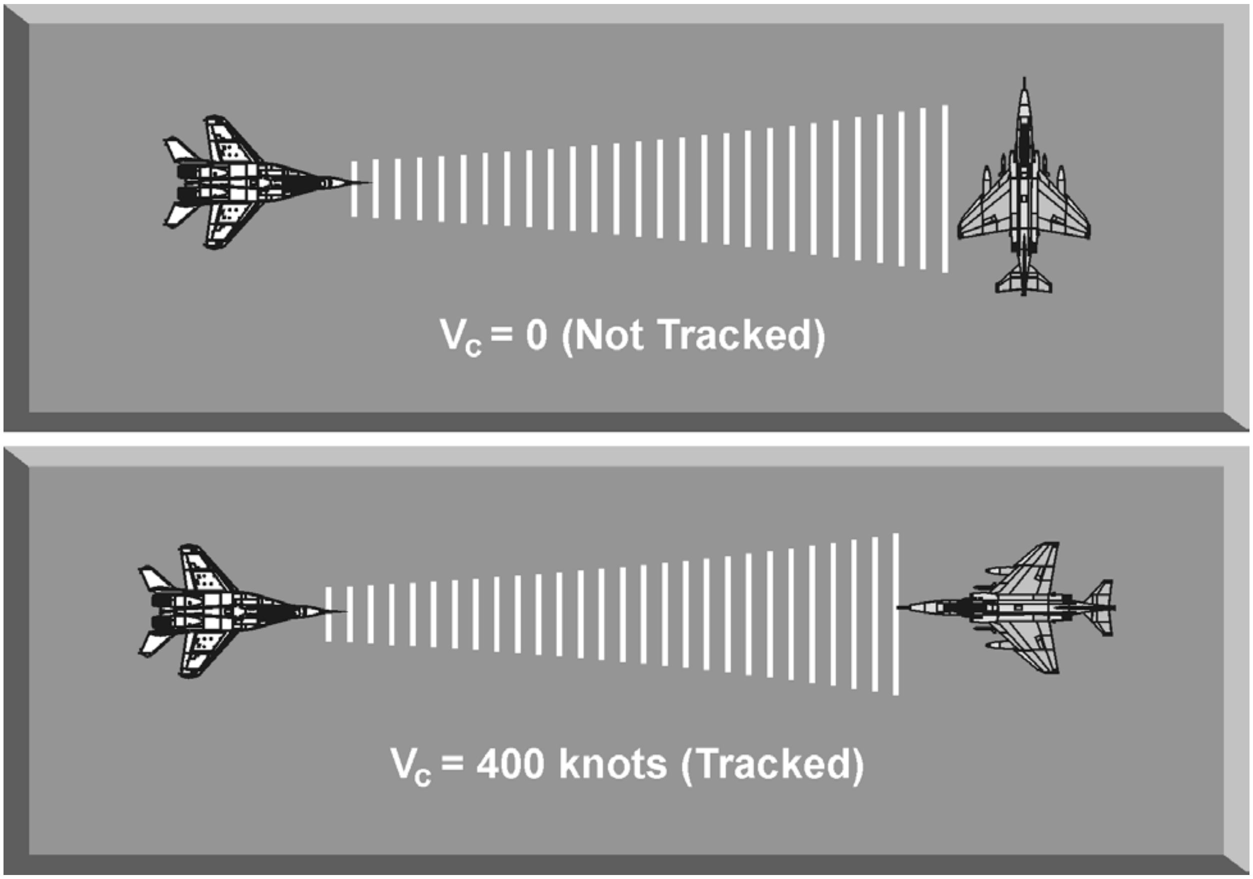

Figure 7-24. Pulse Doppler Speed Gate

(2) Range ambiguity occurs primarily with long-range targets when the return comes back to the radar after another pulse has already been transmitted (Figure 7-25). The radar will see this return and base its range calculations on the transmission time of the immediately preceding pulse, instead of the pulse that generated the return. The result will be an incorrect range calculation.

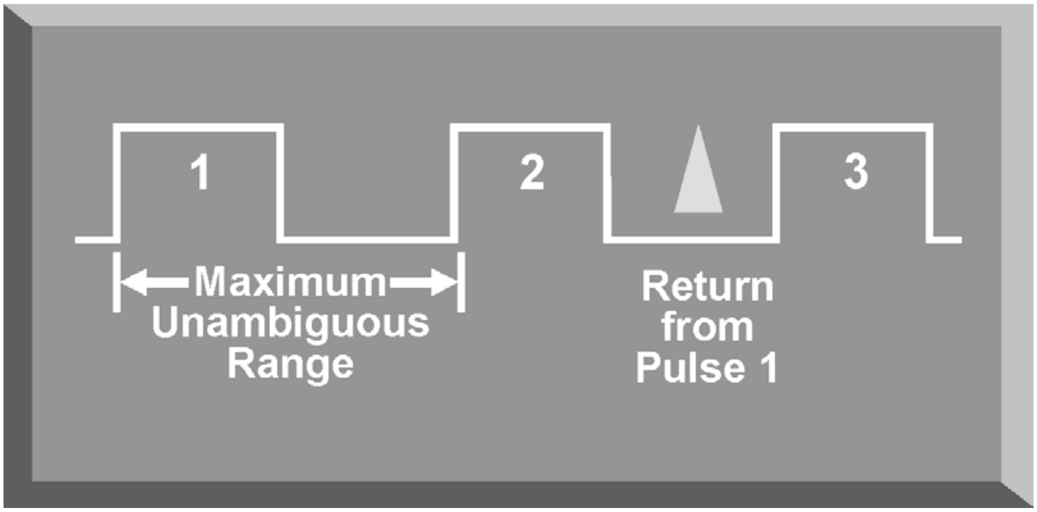

Figure 7-25. Range Ambiguity

(3) Range eclipsing occurs when a target return comes back to the radar antenna while a pulse is being transmitted (Figure 7-26). Since the radar cannot receive while transmitting, the return will not be displayed.

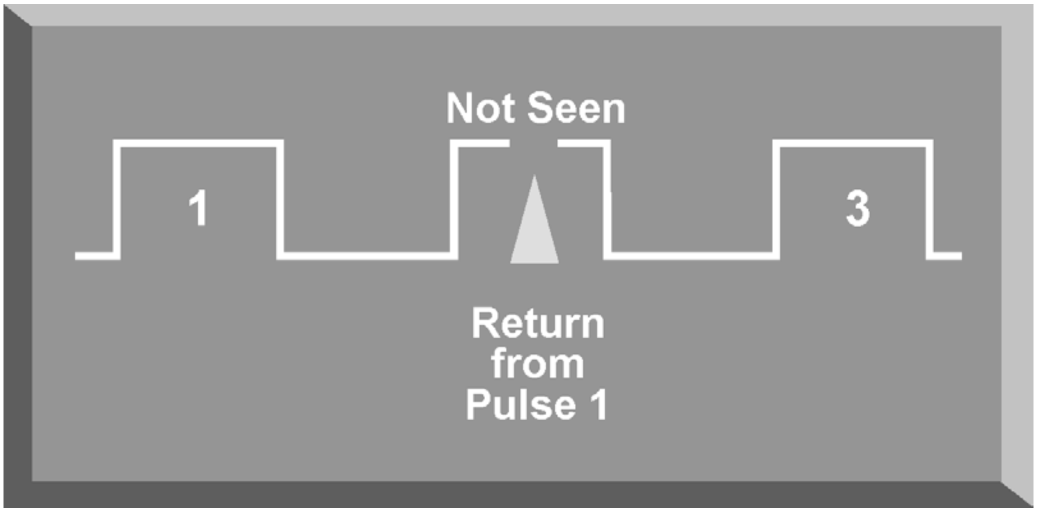

Figure 7-26. Range Eclipsing

(4) To solve the problems of range ambiguity and eclipsing, a pulse Doppler radar employs different PRFs and computer logic. In Figure 7-27, at PRF 1, the return from target 1 is eclipsed, and target 2 is ambiguous. By changing the PRF slightly, these range problems can be resolved. Notice that at PRF 2, neither target is eclipsed, and at PRF 3, target 2 is eclipsed. The computer logic needs to be extremely advanced to compensate for these problems. These simplistic examples show the complex problem of using multiple PRFs.

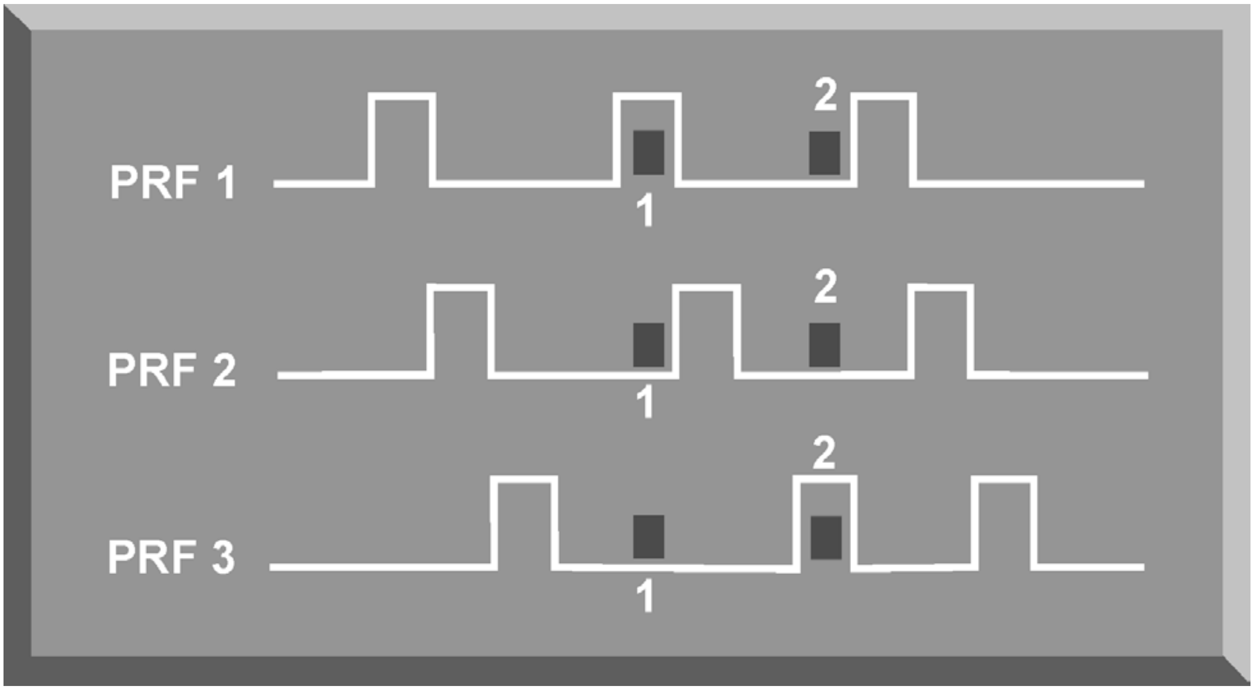

Figure 7-27. Resolving Range Ambiguity and Eclipsing

9. SUMMARY

This chapter has discussed the range, azimuth, elevation, and velocity tracking techniques employed by conical scan, TWS, LORO, monopulse, CW Doppler, and pulse Doppler radar systems. The method used by a specific radar system to track a target determines the type of jamming technique required to counter this system. A basic understanding of the target tracking techniques will enable you to understand the basic jamming techniques employed to defeat these threats.

CHAPTER 8. RADAR MISSILE GUIDANCE TECHNIQUES

1. INTRODUCTION

Once a target has been designated, acquired, and tracked by a radar system, the final stage in target engagement is to guide a missile or projectile to destroy the target. There are three basic requirements for successful missile guidance: (1) precise target tracking by a target tracking radar (TTR) to provide target parameters (range, azimuth, elevation, velocity, etc.), (2) a method to track the position of the missile compared with the target, and (3) a fire control computer to generate missile guidance commands based on target and missile position. The missile guidance techniques employed by modern air-to-air and surface-to-air missile (SAM) systems will be covered in this chapter. In addition, the target engagement techniques employed by antiaircraft artillery (AAA) systems will also be discussed. There are three distinct phases in any missile intercept: boost, mid-course, and terminal.

a. Nearly all missiles are unguided during the initial boost phase (Figure 8-1). During the boost phase, the missile electrical and hydraulic systems are activated and are coming up to operating parameters. The missile is gathering speed and normally will be in an unguided mode of flight.

Figure 8-1. Initial Boost Phase

b. During the mid-course phase, the missile is actively being guided to the target using some type of guidance signal (Figure 8-2). Guidance signals deflect

the control vanes of the missile to change its direction. These vanes change the roll, pitch, and yaw, in some combination, to control the missile flight path. Normally a gas grain generator powers a small hydraulic pump that deflects the control vanes in response to guidance signals. Each missile carries a limited supply of hydraulic fluid for maneuvering. The fluid is expended through vents with every control surface activation. The limited quantity of hydraulic fluid can be a significant factor during a long-range missile intercept.

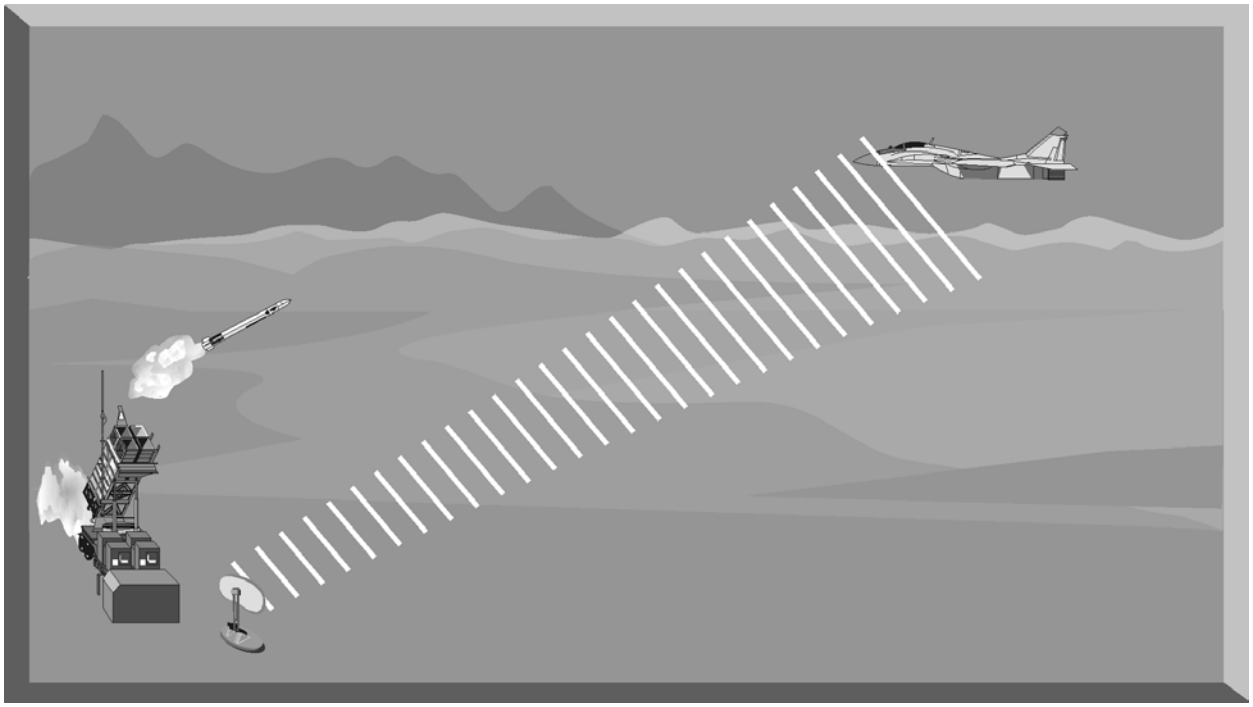

Figure 8-2. Mid-Course Guidance Phase

c. The final phase of an intercept is the terminal phase (Figure 8-3). During this phase, the missile attempts to pass close enough to the target to detonate the fuse while the target is within the lethal radius of the warhead. Modern missiles employ both a contact fuse and some type of proximity fuse. Proximity fuses range from command detonation for command-guided missiles, fractional Doppler gates for semi-active guided missiles, to active laser fuses for IR-guided missiles.

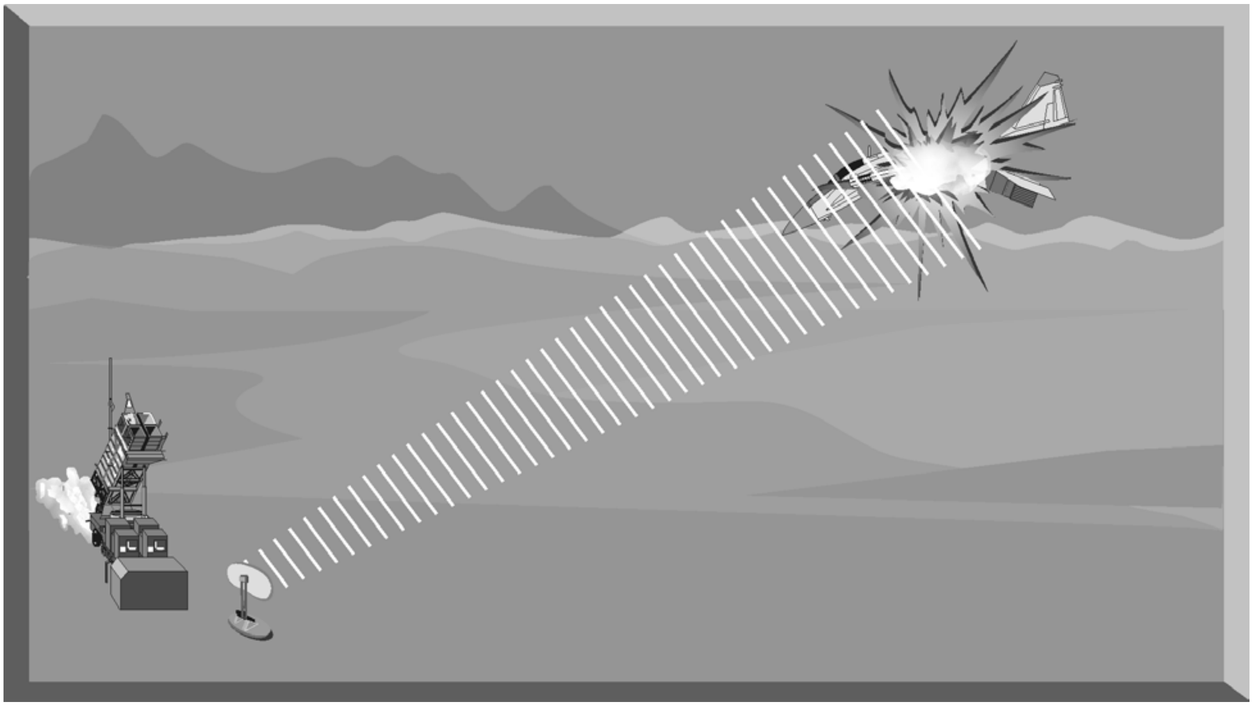

Figure 8-3. Terminal Guidance Phase

2. COMMAND GUIDANCE

Command guidance uses a fire control computer to constantly send course correction commands to the missile throughout its flight. These commands are a series of electrical missile guidance pulses called doublets or triplets. These pulses provide steering commands to the missile by varying the spacing between each guidance pulse. Each pulse, or pulse combination, relays some roll, pitch, and yaw command to the missile. These inputs are constantly corrected for the spatial relationship between the missile and the target's present position and rate of motion. Guidance commands are passed to the missile by specialized antennas on the TTR and an antenna installed on the missile, called a missile beacon. The beacon is a special radio receiver and transmitter that is attached to the rear of the missile. It acts like a transponder in that the TTR tracks and receives guidance commands. The guidance frequency may be widely separated from the target tracking radar frequency to minimize interference. This beacon is usually masked until missile booster separation. This results in the missile being launched unguided for the first 2-3 seconds. This type of delay is one of the reasons that all command-guided missile systems have a minimum launch range. Command guidance is used by the SA-2, SA-3, SA-4, and SA-8.

a. Command-guided missiles will generally fly a rectified (full or half) or three-point pursuit geometry during the mid-course portion of the intercept (Figure 8-4). However, a command-guided missile may transition to a pure pursuit geometry during the terminal phase of the intercept. Rectified geometry involves the prediction of where the target and the missile will be at some point in the future.

The target's direction and rate of movement is tracked and predicted. The missile is then launched, pulls lead on the target, and is guided to the point in the sky where the intercept is predicted to take place. This profile requires the constant update of both the target and missile positions.

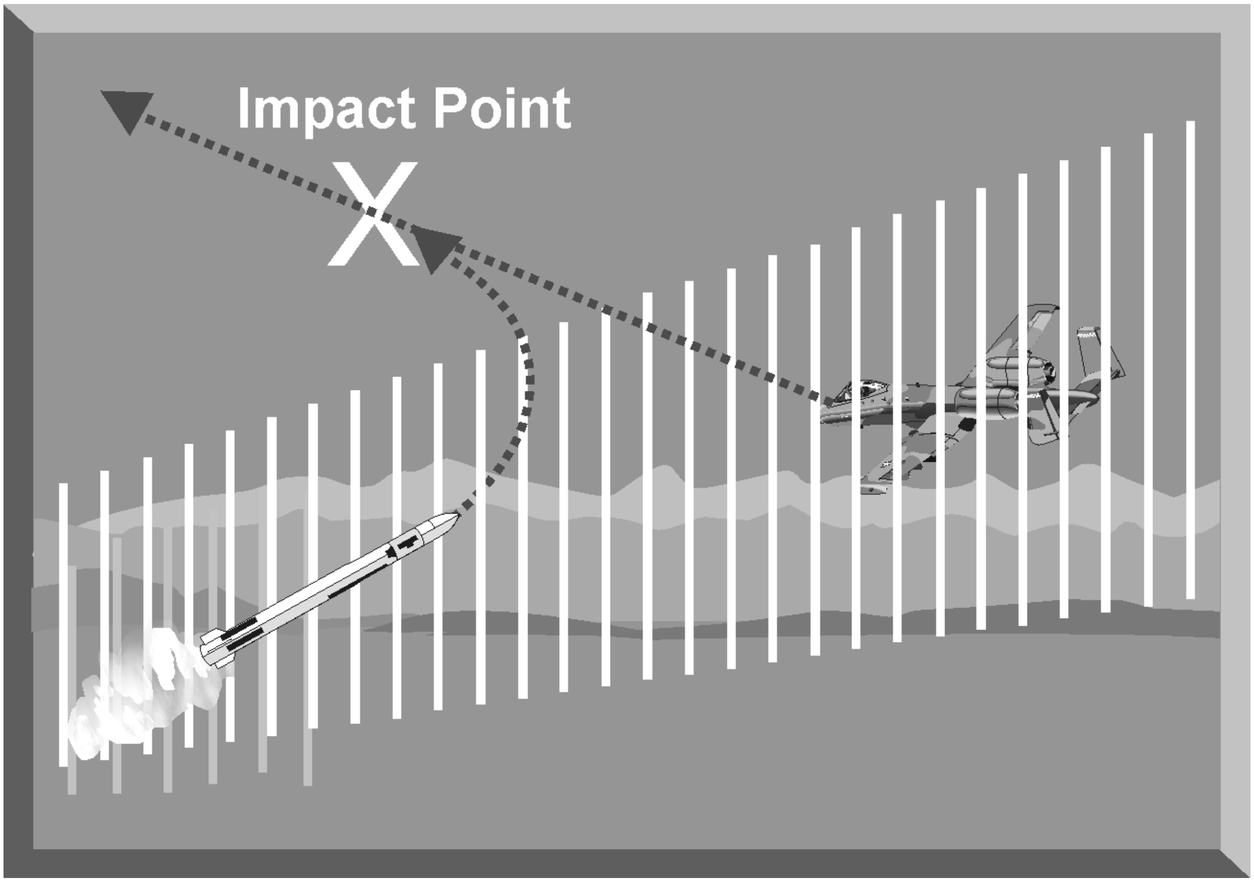

Figure 8-4. Rectified Flight Profile

b. Three-point pursuit geometry is often used when there is incomplete range tracking data on the target. In this case, it will be impossible to predict exactly where the target will be at some point in the future. In this profile, the target tracking radar constantly tracks the target. The missile location will be updated by the missile beacon. The fire control computer will direct the missile to fly directly down the tracking radar beam toward the target. In this geometry, the missile may start out on a direct intercept course and, depending on the target's direction and rate of movement, transition to a pure pursuit intercept. The three points in three-point missile geometry are depicted in Figure 8-5. Point one is the target tracking radar, point two is the missile itself, and point three is the target. By keeping all three points always in a line, the missile will intercept the target at some point, although the range of the target is unknown.

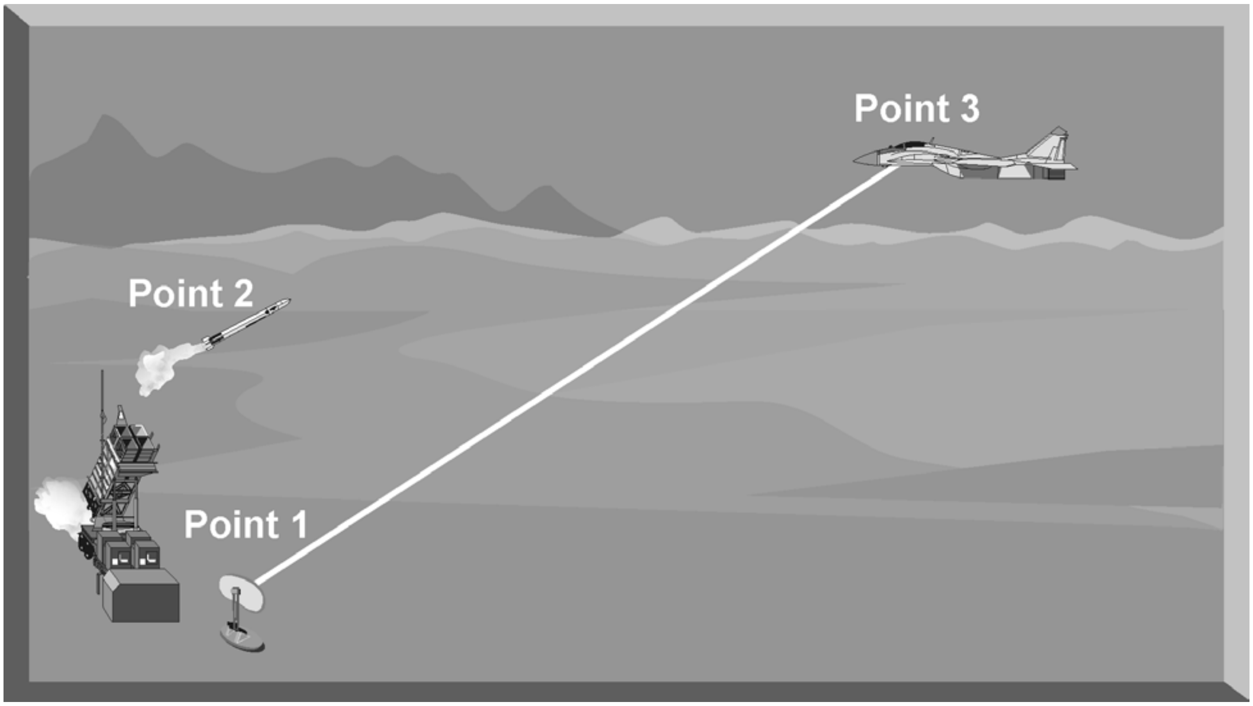

Figure 8-5. Command Guidance: Three-Point Pursuit

c. Command guidance techniques have many advantages. First, command-guided missiles can adjust their flight geometry throughout an intercept profile. Second, the missiles are uncomplicated since they do not carry onboard computers or target tracking equipment. The fire control computer associated with the TTR accomplishes all intercept calculations. Third, the primary intercept profile, a full- or half-rectified intercept, is the fastest and most fuel-efficient intercept. Fourth, command guidance is difficult to jam since the missile beacon antenna is at the rear of the missile and can be relatively high-powered. And finally, an intercept is possible even without accurate range information by using the three point intercept profile.

d. Command guidance has several disadvantages. First, the use of a missile beacon delays the capture of the missile by the tracking radar. This can cause a large dead zone which equates to a larger minimum engagement range. Second, the accuracy of the intercept geometry is only as good as the tracking information provided by the target tracking radar. Jamming, interference, or loss of signal will adversely affect the intercept accuracy. In addition, normal radar characteristics could produce sufficient errors to cause the missile to miss the target, especially at longer ranges. Third, with insufficient range information, the three-point intercept profile is very slow and could result in the missile running out of energy before it gets to the target. Fourth, command guidance is reactive. The fire control computer constantly updates the intercept geometry based on target maneuvering. This results in missile maneuvering lagging target maneuvers.

3. SEMI-ACTIVE GUIDANCE

Semi-active guidance is significantly different from command guidance, but only after launch. The first requirement is still for the target tracking radar to maintain a solid target track, with the tracking data being supplied to the fire control computer. The fire control computer then directs a target illumination antenna to point at the target and illuminates it with CW energy. The missile then passively homes on the reflected CW energy.

a. The missile used by a threat system that uses CW homing is vastly different from the missile being guided by a command guidance signal. The missile that homes on CW energy must be equipped with a seeker section composed of an antenna and an internal receiver. The seeker section processes and computes the necessary course corrections as it flies toward the target. It can do this by knowing the zero boresight line of the antenna within the missile (Figure 8-6). As the reflected CW energy is received by the seeker, there is normally some deviation from the zero reference position. The onboard computer then directs the control surfaces to change the flight path to reduce the reference errors in the antenna to zero, if possible. When the error between the antenna position and the boresight position is zero, the missile is pointed directly at the target.

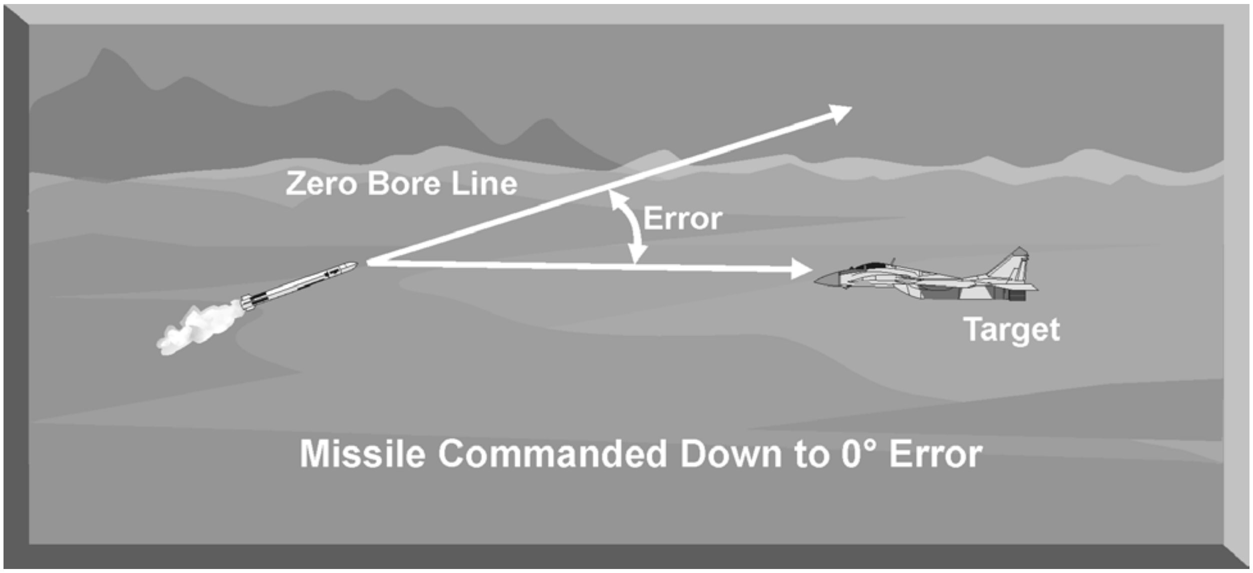
Zero Bore Line
Error
Target
Missile Commanded Down to 0° Error

Figure 8-6. Semi-Active Guided Missile

b. Missile systems that use semi-active guidance normally use velocity as the primary target discriminator during the intercept. The missile seeker locks onto a reference Doppler signal provided by the fire control computer before launch. This Doppler signal establishes a tracking gate around the velocity of the target. After the missile is launched, it initially compares the reference Doppler to the target Doppler signal.

c. The mid-course phase for a semi-active missile is also different from that of a command-guided missile. A semi-active guided missile follows the reflected CW energy during the mid-course phase of the intercept and normally attempts to fly a lead pursuit profile to the target (Figure 8-7). If the target maneuvers, however, the missile may transition to a pure pursuit flight path. Unlike a command-guided missile, a semi-active guided missile does not use a missile beacon. The fire control computer does not need to know where the missile is to compute course corrections since all that is necessary is to illuminate the target with the CW illuminator. This also means that the missile can begin to track and guide when it is launched and locked on to the reference Doppler gate. Semi-active guidance is the primary mode of guidance for many surface-to-air missiles, and almost all radar-guided air-to-air missiles.

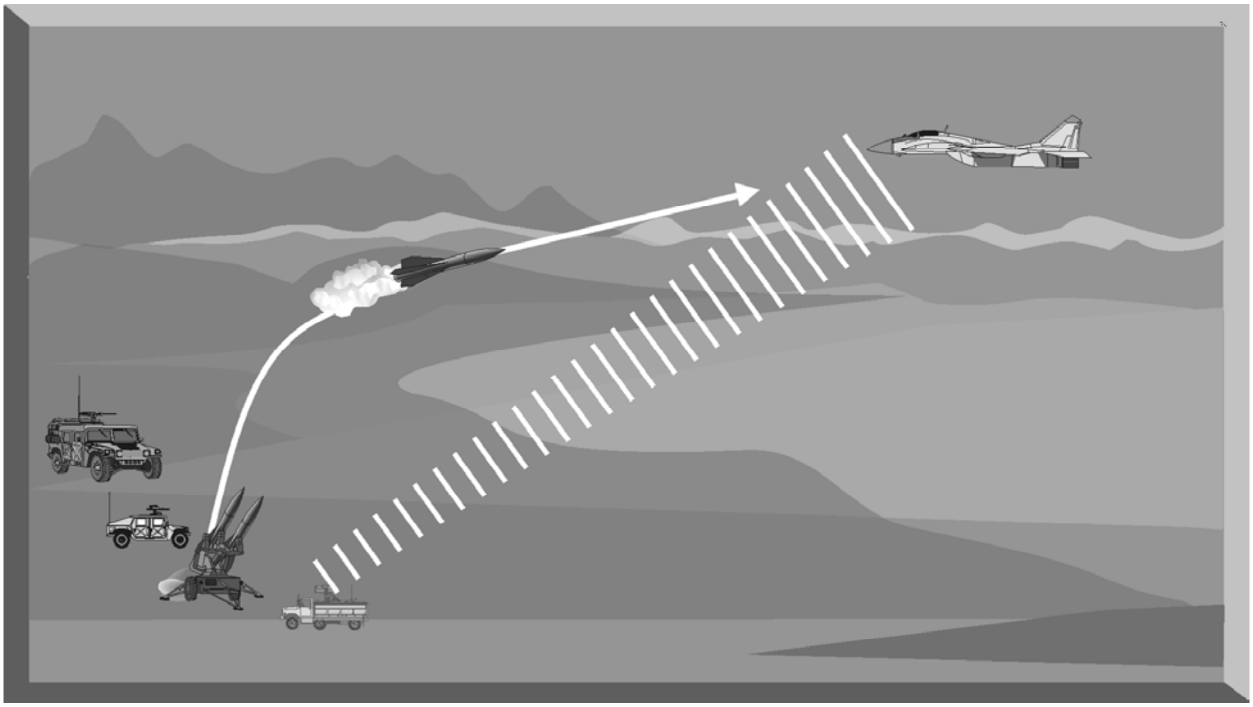

Figure 8-7. Semi-Active Guidance (Mid-Course)

d. As the missile enters the terminal phase of the intercept, there is no change in the guidance mode used by a CW homing missile. The missile may complete the terminal phase of the intercept geometry by going to a pure pursuit flight path, if necessary (Figure 8-8). The missile continues to home in on the reflected CW signal until it passes close enough for the fuse to function.

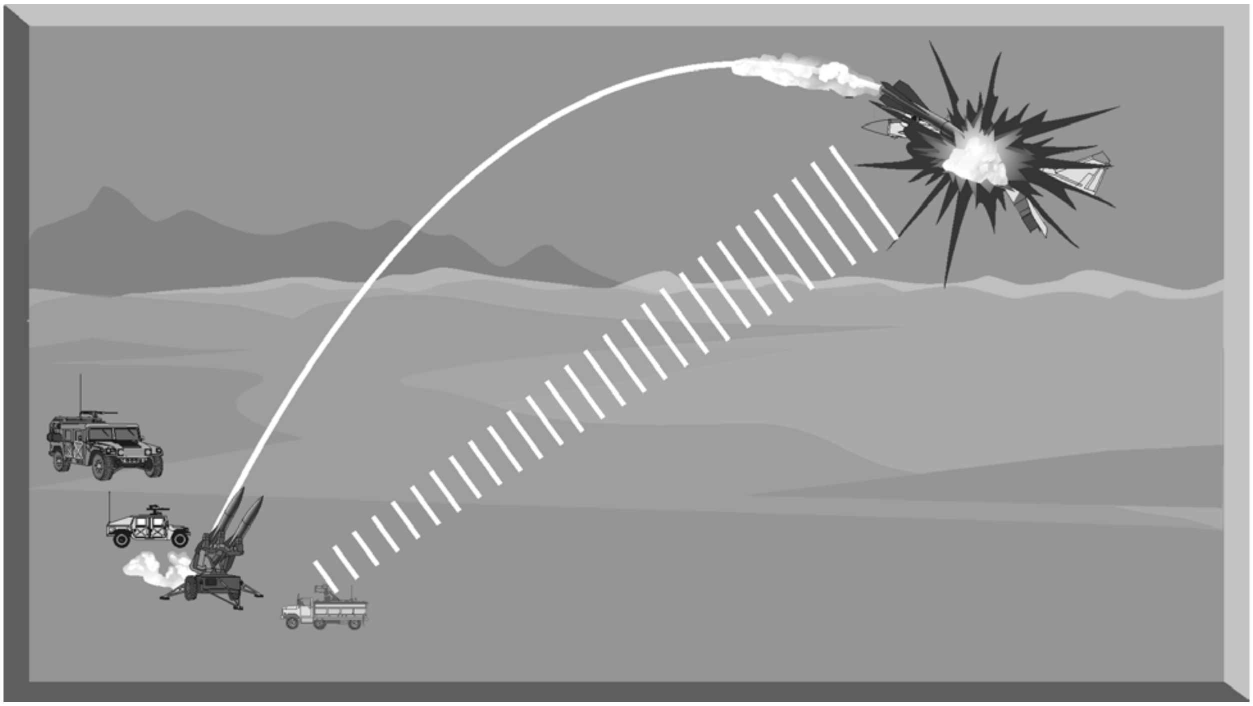

Figure 8-8. Semi-Active Guidance (Terminal Phase)

e. Semi-active missile guidance has many advantages. First, a semi-active guided missile is resistant to electronic jamming that may be used to deny range information. Second, a semi-active missile can be guided almost immediately after launch. This gives it a very small minimum range since it can maneuver almost as soon as it clears the launch rail. Third, it computes its own course corrections as necessary. This allows for a much quicker reaction to target maneuvers compared to a command-guided missile. Fourth, during a long-range intercept, a CW missile can be more accurate than a command-guided missile. This is accomplished by taking the inherent long-range radar tracking errors out of the equation. The target tracking radar only has to keep the target illuminated so that it can point the CW antenna at the target.

f. Although semi-active missile guidance is generally considered an excellent guidance technique, it does have some disadvantages. First, a semi-active guided missile normally requires reference Doppler values to be entered into the missile computer before launch. Without this reference, a semi-active missile cannot be launched (Figure 8-9). Second, a semi-active homing missile must maintain a lock onto the target Doppler. The use of chaff and beam maneuvers, which result in a near zero target Doppler, may cause a missile or radar to break lock. Third, if a break-lock occurs, a CW homing missile normally cannot regain target track and complete the intercept.

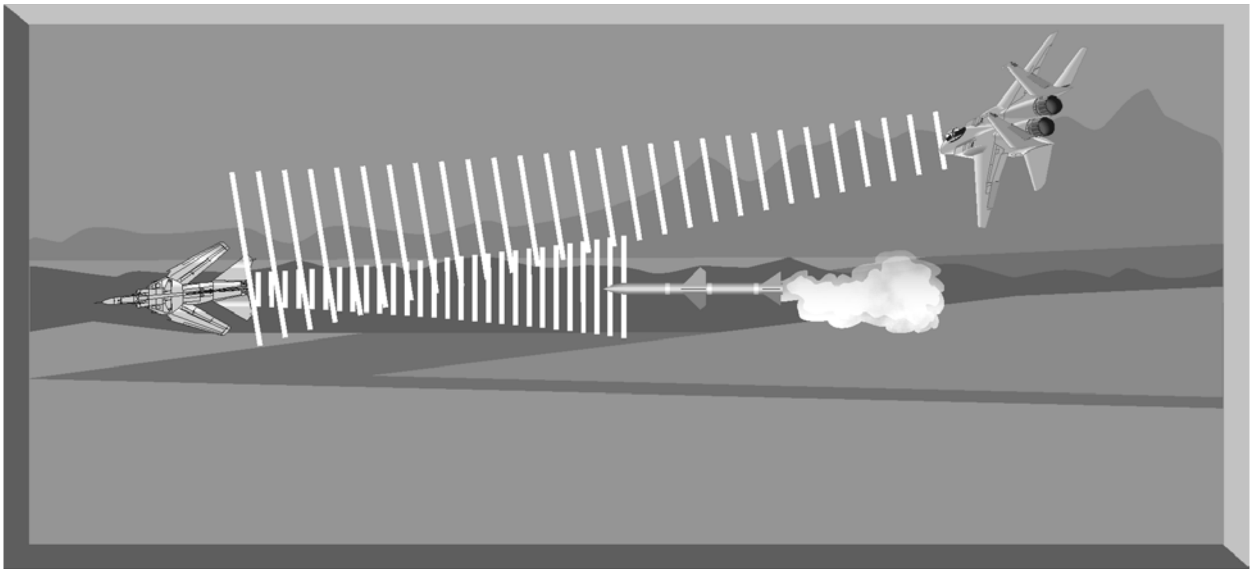

Figure 8-9. Semi-Active Guidance CW Reference Signal

4. ACTIVE GUIDANCE

Active guidance is an improvement that has been included in several new long-range missiles such as the AIM-54 Phoenix and the AIM-120 AMRAAM. This specialized guidance mode is only active during the terminal phase of flight. The mid-course phase usually employs semi-active or command guidance (Figure 8-10).

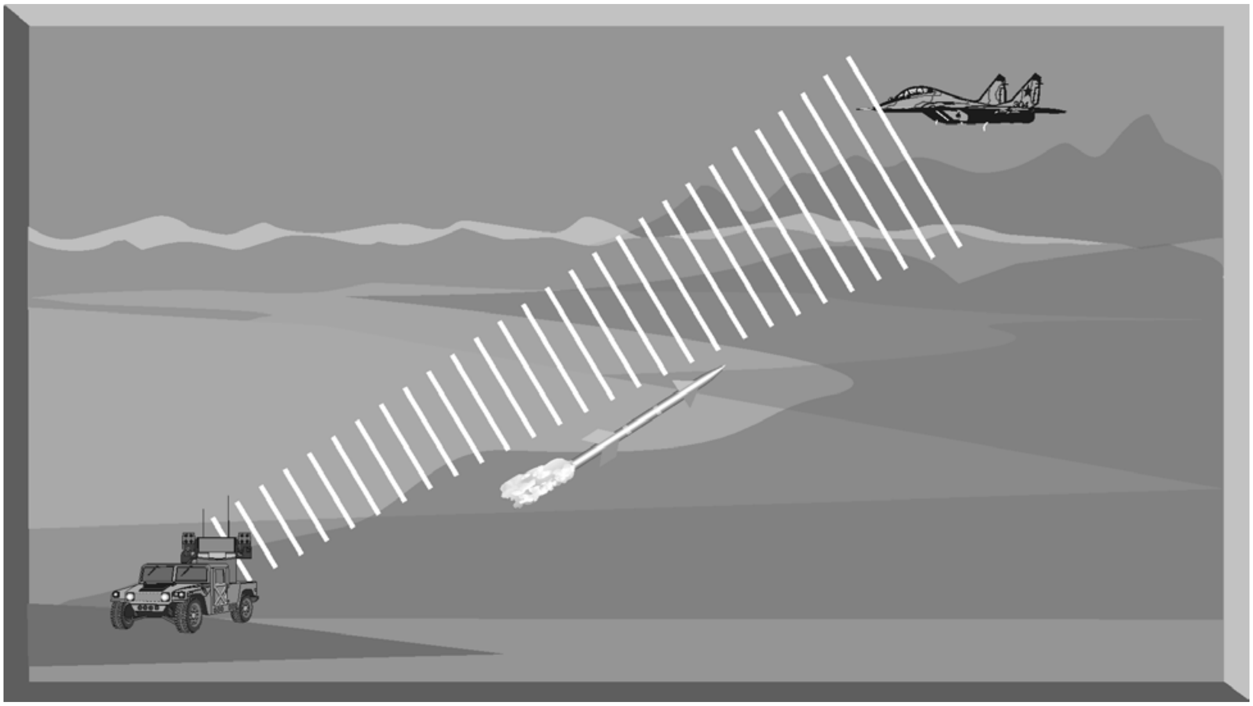

Figure 8-10. Active Guidance Mid-Course Intercept Phase

a. The range at which the missile goes "active" is dependent on the intercept geometry. High-aspect angle intercepts allow the activation of active guidance sooner than beam or tail-aspect intercepts. Missiles that employ active guidance carry a complete miniature radar system and fire control computer within the missile. As the missile nears the target, its internal radar system turns on and locks onto the target. The internal fire control computer directs control inputs to complete the intercept (Figure 8-11).

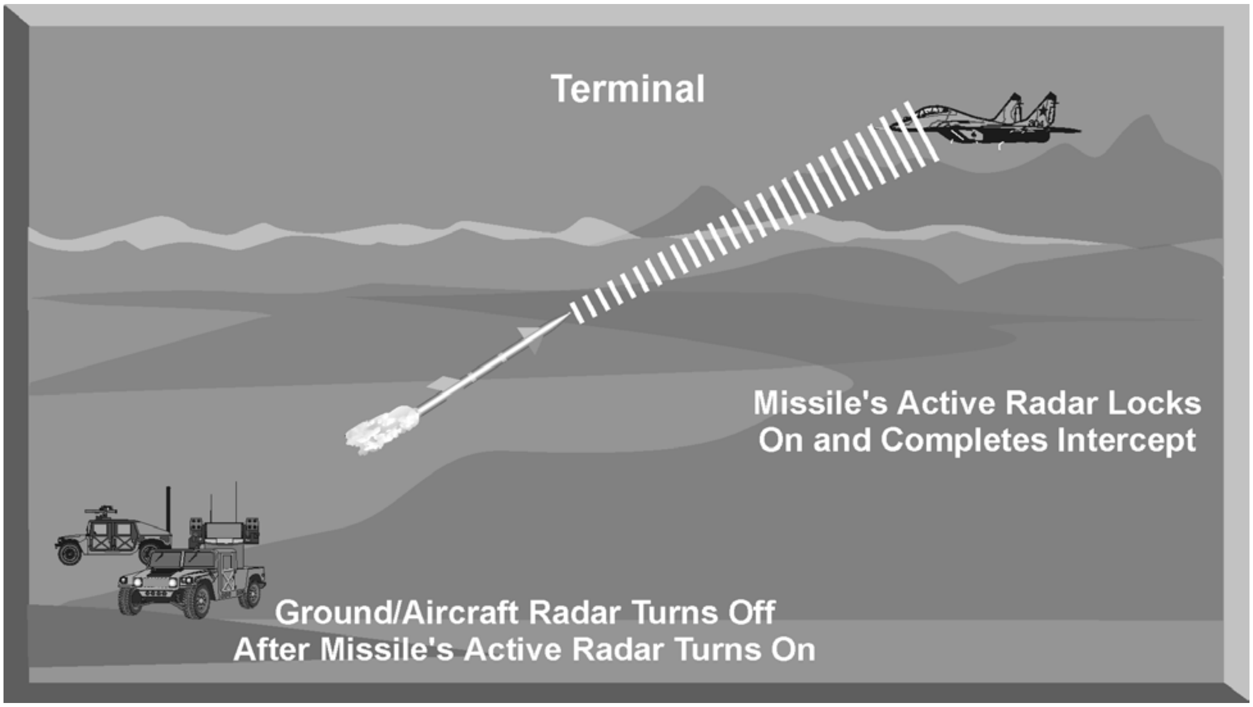
Terminal
Missile's Active Radar Locks On and Completes Intercept
Ground/Aircraft Radar Turns Off After Missile's Active Radar Turns On

Figure 8-11. Active Guidance Phase

b. Active-guided missiles have many advantages. First, active-guided missiles are very accurate at long ranges. This is because they do not rely on the target tracking radar once their internal radar takes over the intercept. Second, an active missile is extremely difficult to jam. It uses a narrow beam and its relative power is constantly increasing as it nears the target. Third, an active-guided missile is a fire-and-forget weapon. Command or semi-active missile guidance requires the target tracking radar to maintain lock-on until the intercept is completed. In an air-to-air engagement, this means the interceptor is predictable until the missile hits the target, and vulnerable to an enemy missile attack. An interceptor with an active missile, however, may launch the missile and, once it goes "active," can then turn around or maneuver defensively.

c. Active-guided missiles have a few disadvantages as well. First, the active homing missile is a complex missile integrating both command and active guidance modes. Second, the missile may still be susceptible to electronic jamming during the mid-course phase of flight. Remember, during the mid-course

phase, the missile relies on command or semi-active guidance. Jamming the target tracking radar may affect the missile's ability to "see" the target near the terminal phase.

5. SEEKER-AIDED GROUND GUIDANCE/TRACK-VIA-MISSILE GUIDANCE

In seeker-aided ground guidance (SAGG) and track-via-missile (TVM) guidance, the target is illuminated by the ground-based radar and the missile receives reflected energy from the target. Unlike conventional semi-active homing, the missile does not generate its own guidance commands. Instead, the missile transmits raw engagement data to the ground-based fire control system (FCS) in order to generate uplink guidance commands. TVM is similar to SAGG; however, additional processing is done on-board the missile prior to transmitting the engagement data to the ground-based FCS.

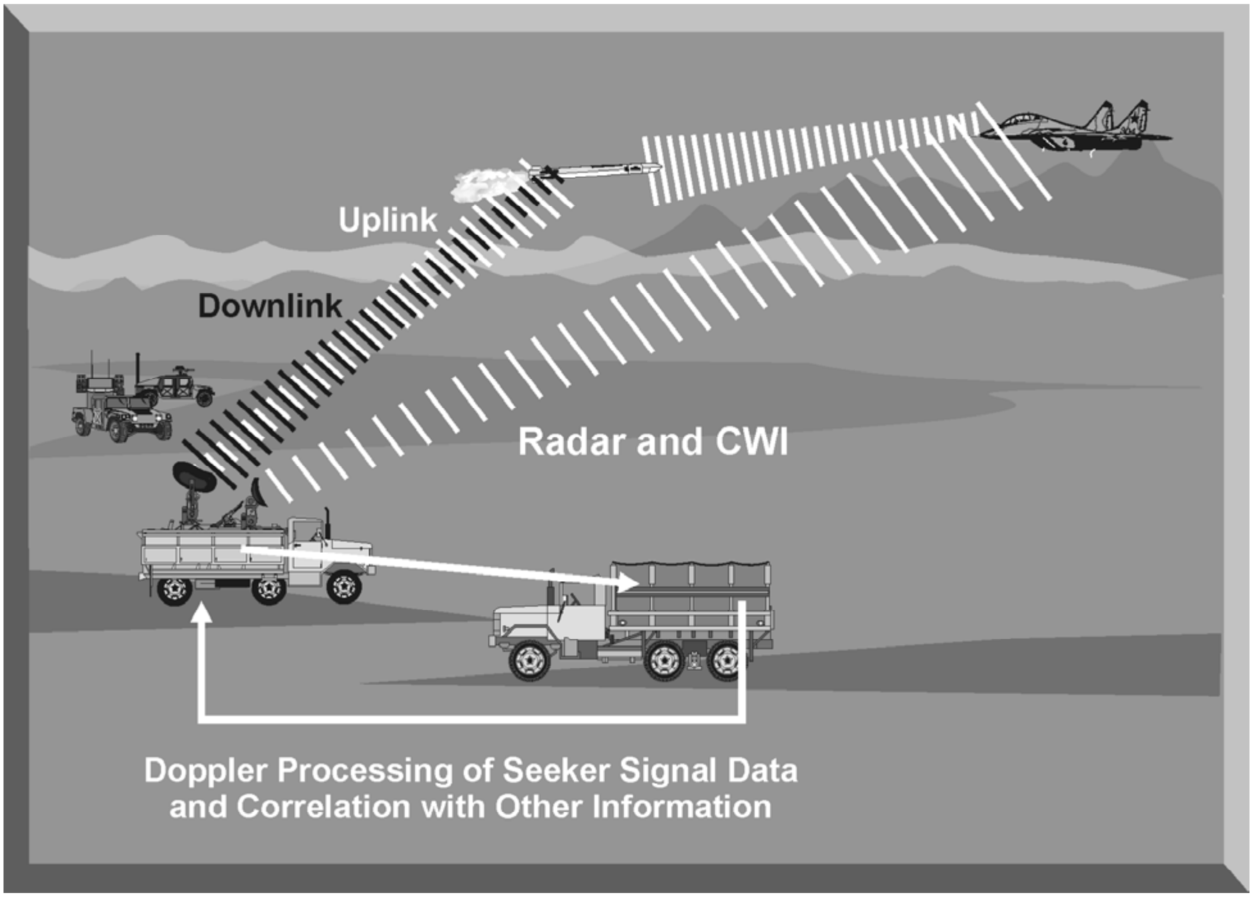
Uplink
Downlink
Radar and CWI
Doppler Processing of Seeker Signal Data
and Correlation with Other Information

Figure 8-12. SAGG/Track-Via-Missile Guidance

a. Track-via-missile and seeker-aided ground guidance are two relatively new missile guidance techniques with similar advantages. First, they are extremely accurate at long ranges where the inherent radar tracking errors may be large

enough to cause a miss. Second, they can respond very quickly to any actions taken by the target since the missile seeker can track these changes and transmit the new position to the TTR fire control computer. Third, TVM and SAGG can be used with a large and capable fire control computer since most computations are accomplished by the TTR. Fourth, the integration of a phased array radar and the powerful TTR fire control computer allows the missile system to engage multiple targets. The Patriot missile battery, for example, can track and engage at least four targets simultaneously.

b. The major disadvantage of track-via-missile and seeker-aided ground guidance is that they are the most complex forms of missile guidance. They require the use of sophisticated computers to combine radar tracking data and data received from the missile. This required hardware is expensive and demands greater maintenance and logistical support. In addition, the missile itself needs to be large enough to store the appropriate hardware for computations and data transfer.

6. ANTIAIRCRAFT ARTILLERY (AAA)

The classic role of AAA is point defense. AAA systems provide close-in defense for high-value targets. AAA systems are deployed to defend cities, airfields, bridges, industrial centers, lines of communications, command and control centers, infantry/tank units, and SAM sites. There are two types of AAA systems: towed and mobile. Towed AAA is normally deployed in fixed sites around key targets. Mobile AAA systems are deployed to provide air defense for army units and to protect mobile SAM sites. The effectiveness of AAA systems, towed or mobile, depends on the ability of the system to predict an aircraft's future position to fire its unguided ballistic projectile to intercept the aircraft and destroy it. To accomplish this objective, AAA systems employ two primary tactics, aimed fire and sector/barrage fire.

a. Aimed AAA fire requires very accurate aircraft position information and an accurate prediction of future position. For aimed AAA fire, this information can be derived by using an optical sighting system on the gun or by employing a radar system coupled with a fire control computer. Smaller caliber AAA guns generally rely on optical target acquisition and firing (Figure 8-13). The high rate of fire, short range, and short projectile time of flight (TOF) simplifies the prediction and aiming problem for these systems. Smaller caliber AAA can also use tracer ammunition to help the gunner in correcting his optical firing solution. Larger caliber AAA systems, with slow rates of fire, long range, and long projectile TOF, generally use a TTR and a fire control computer to solve the problems associated with aimed fire (Figure 8-14).

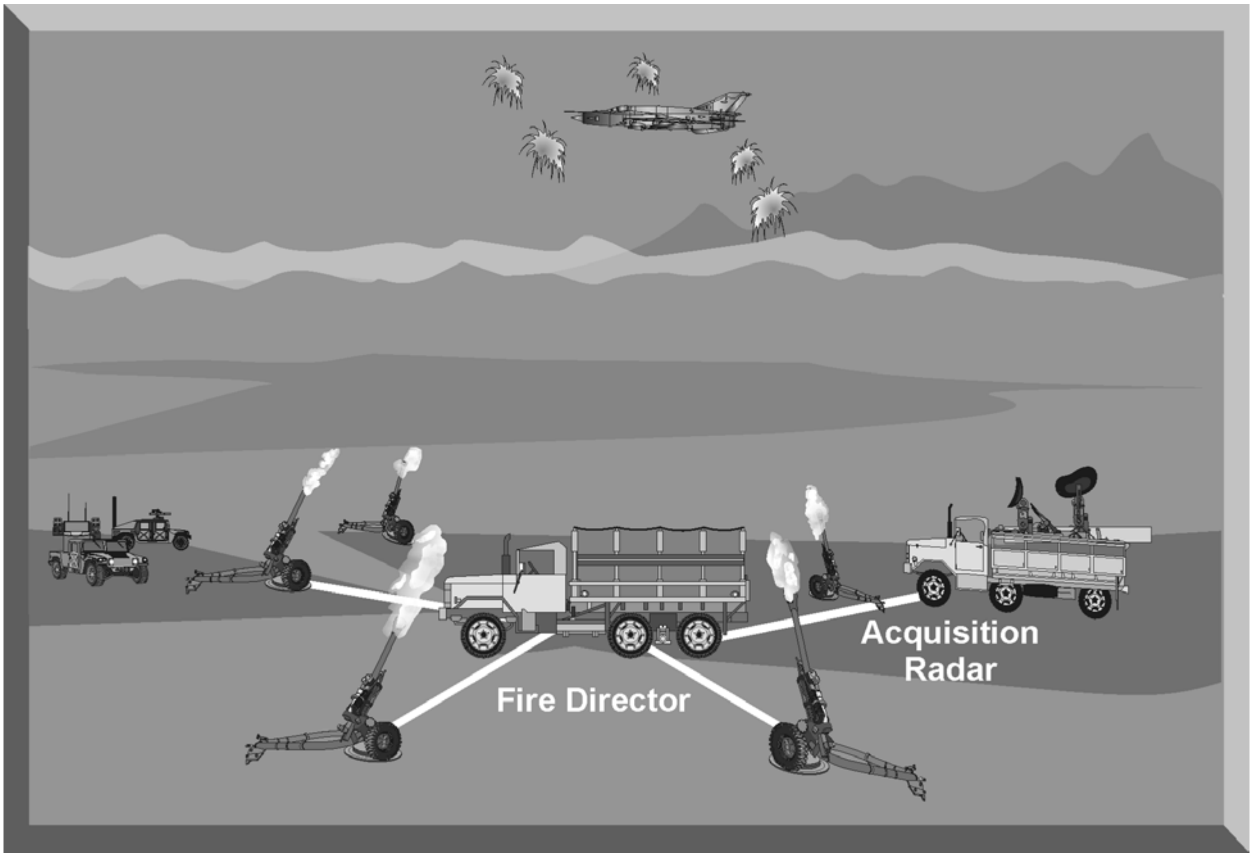

Figure 8-13. Optically Aimed AAA

(1) The typical engagement sequence for an aimed AAA engagement employing a TTR and fire control computer begins with initial target data from an acquisition radar. The guns and TTR are pointed toward the target. The TTR initiates search and lock-on to the target. The TTR associated with large caliber AAA is usually a conical scan radar to provide accurate target positioning information. Target information is fed into the fire control computer which calculates the aim point, points the guns, and initiates firing. The fire control computer uses the target kinematic data, gun ballistics, wind, air density, and projectile dispersion pattern to compute the required aim point. All these computations are based on the assumption that the target will continue on the same heading, at the same altitude, and at the same airspeed during the projectile TOF.

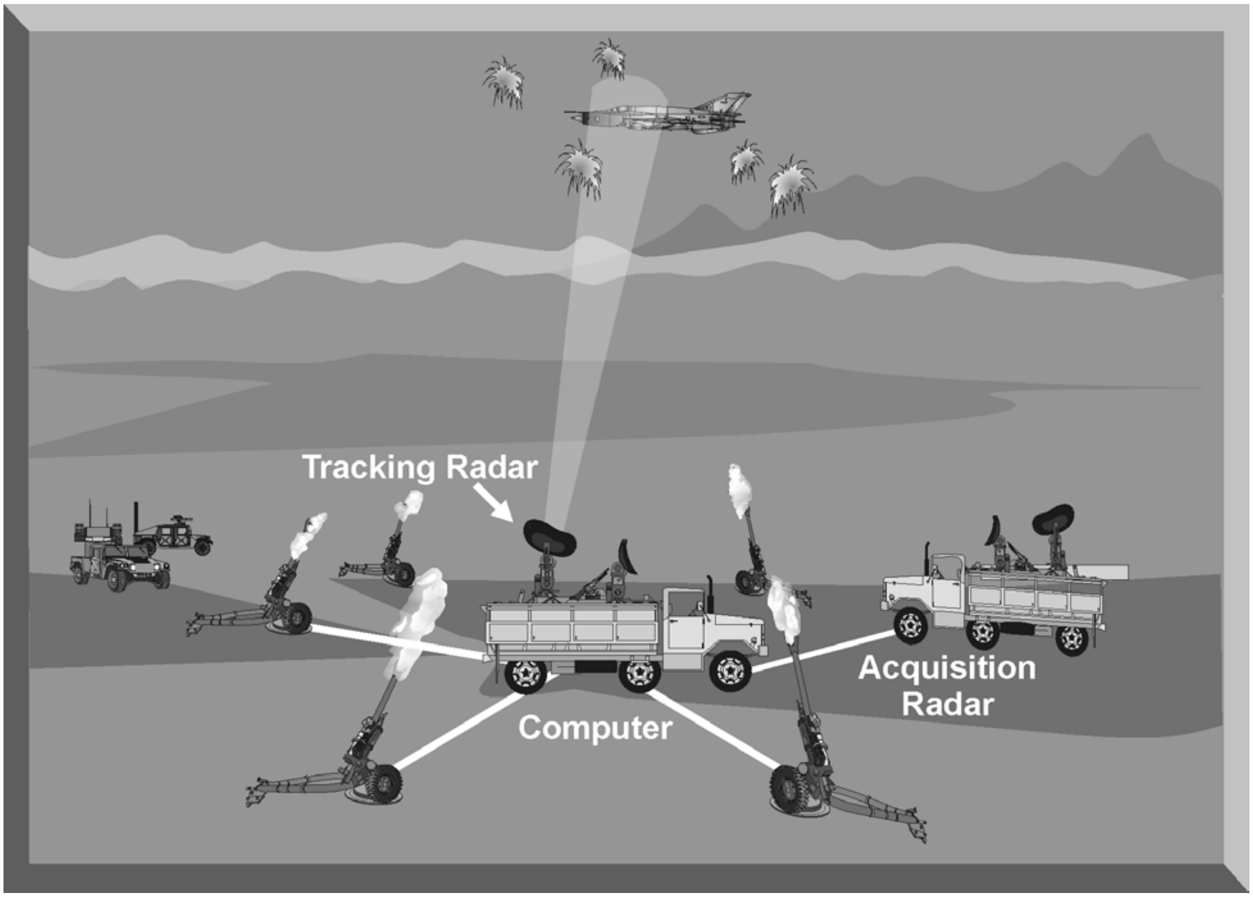

Figure 8-14. Radar-Directed AAA

(2) The typical engagement sequence for an aimed AAA engagement, employing optical target tracking begins with initial target information from an acquisition radar to the fire director. The fire director gives gross aiming commands to the individual guns. The gunners then visually search for the target and use the on-carriage gun sights to predict the required lead angle and initiate firing.

b. Sector or barrage fire tactics are employed when the aircraft cannot be accurately tracked (Figure 8-15). Acquisition information suggests an aircraft will traverse a volume of airspace or a specific sector. The fire director instructs the gunners to fire randomly into this sector in an effort to hit the aircraft with the barrage of AAA fire, or have the aircraft fly into a "curtain" of AAA fire. This tactic is especially effective for point defense for a fixed target. Attacking aircraft may have to fly a predictable flight path during weapons delivery. Sector/barrage fire can be directed to cover the expected attack directions and altitudes.

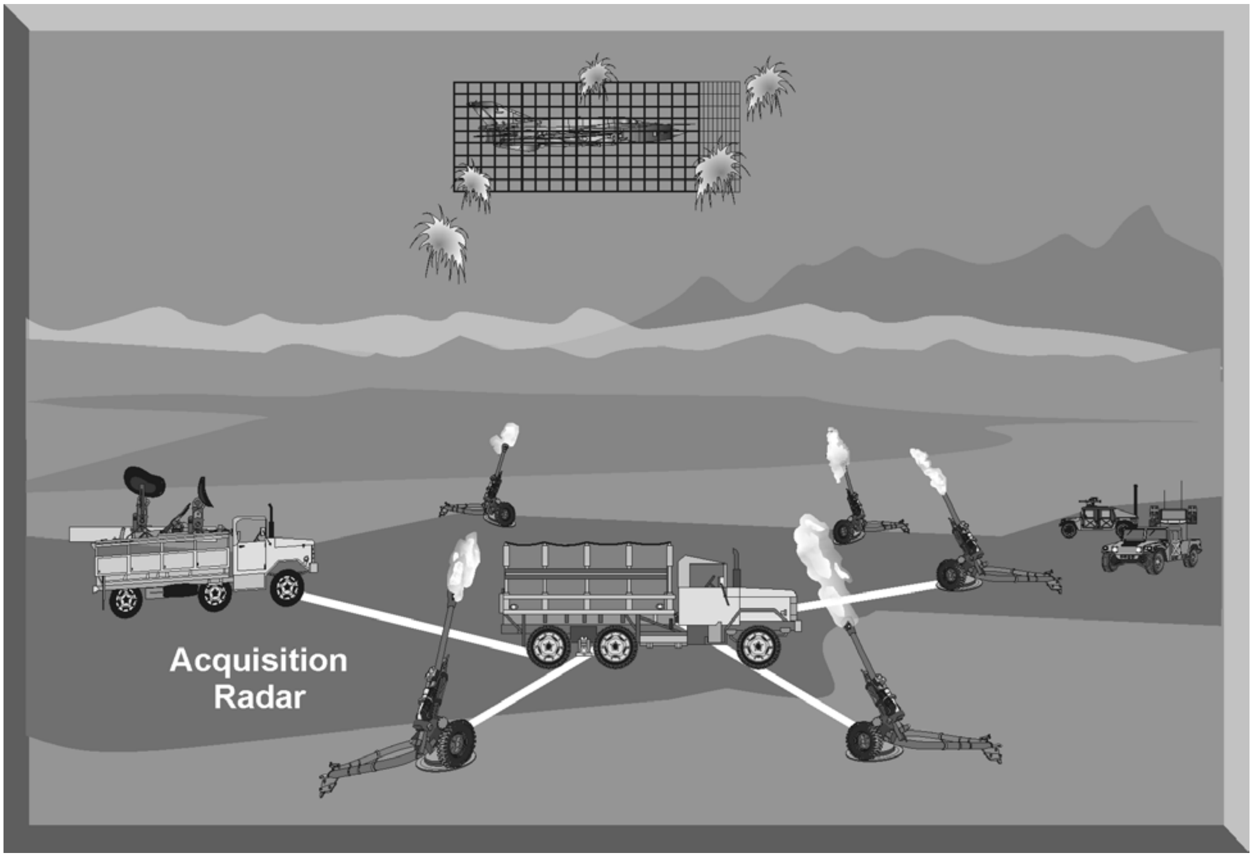

Figure 8-15. Sector/Barrage AAA

7. SUMMARY

This chapter has discussed the most common missile guidance techniques and AAA firing modes used by modern threat systems. A familiarity with the guidance technique employed by specific threat systems is the key to understanding the jamming techniques, chaff/flare employment settings, and tactical maneuvers designed to counter these systems.

CHAPTER 9. INTRODUCTION TO RADAR JAMMING

1. INTRODUCTION

Radar jamming is the intentional radiation or reradiation of radio frequency (RF) signals to interfere with the operation of a radar by saturating its receiver with false targets or false target information. Radar jamming is one principal component of electronic combat (EC). Specifically, it is the electronic attack (EA) component of electronic warfare (EW). Radar jamming is designed to counter the radar systems that play a vital role in support of an enemy integrated air defense system (IADS). The primary purpose of radar jamming is to create confusion and deny critical information to negate the effectiveness of enemy radar systems. This chapter will introduce the two types of radar jamming, the three radar jamming employment options, and discuss the fundamental principles that determine the effectiveness of radar jamming.

2. RADAR JAMMING TYPES

There are two types of radar jamming: noise and deception.

a. Noise jamming is produced by modulating a RF carrier wave with noise, or random amplitude changes, and transmitting that wave at the victim's radar frequency. It relies on high power levels to saturate the radar receiver and deny range and, occasionally, azimuth and elevation information to the victim radar (Figure 9-1). Noise jamming takes advantage of the extreme sensitivity of the radar receiver and the transmission pattern of the radar antenna to deny critical information to the victim radar.

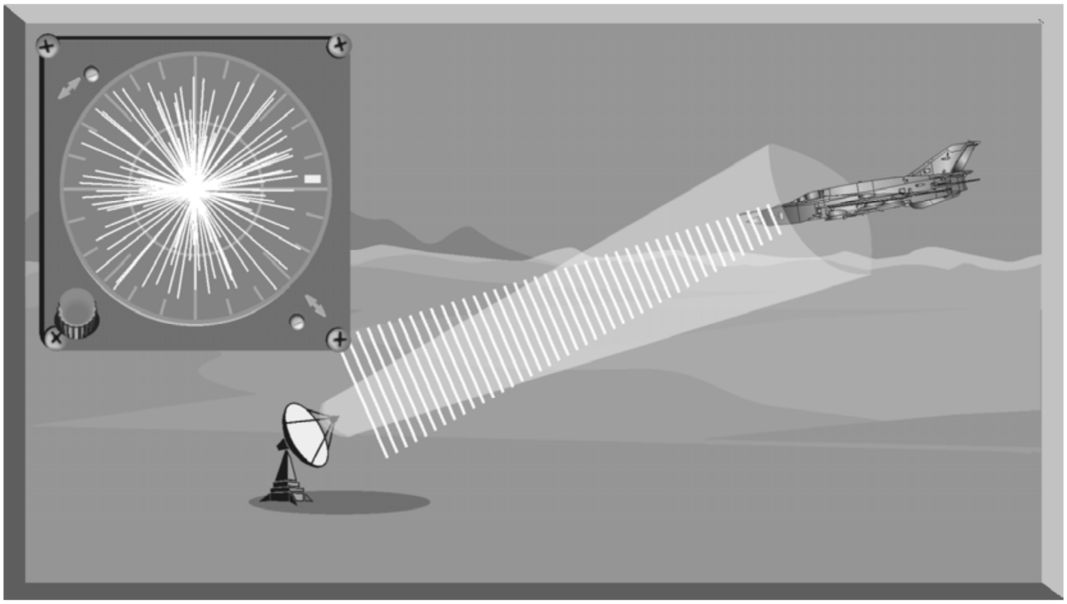

Figure 9-1. Noise Jamming

b. Deception jamming uses complex receiving and transmitting circuits to process and retransmit jamming pulses that appear as a real target to the victim radar. A deception jammer receives the signal from the victim radar and alters the signal to provide false range, azimuth, or velocity information. The altered signal is then retransmitted (Figure 9-2). The victim radar processes this signal, which disrupts the victim radar and confuses the radar operator. To be effective, deception jamming must match not only the victim radar's operating frequency, but all the other operating characteristics, including pulse repetition frequency (PRF), pulse repetition interval (PRI), pulse width, and scan rate.

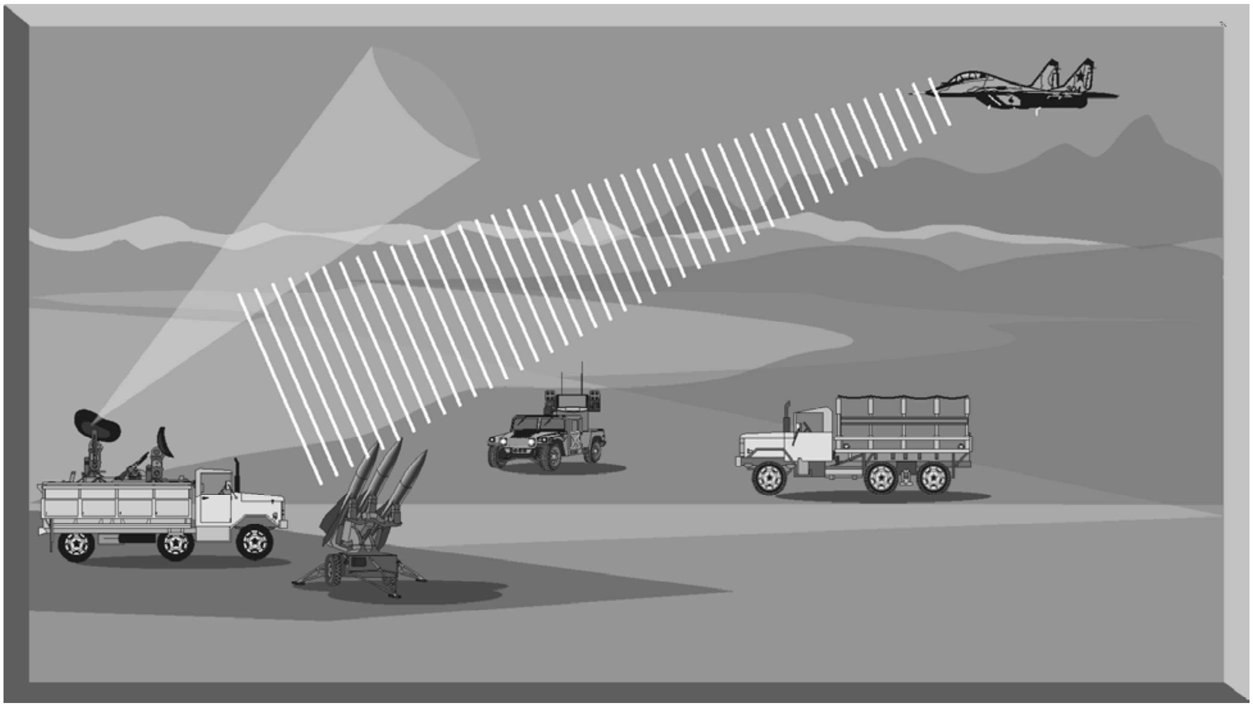

Figure 9-2. Deception Jamming

c. Both noise and deception jamming effectiveness are heavily dependent on another component of EW, specifically, electronic warfare support (ES). ES assets, either airborne or ground-based, provide the threat system specific radar parametric data and update this critical information based on observed threat system operations. This data provides the foundation for developing noise and deception jamming techniques. Intelligence and engineering assessment of this data are used to identify specific threat system weaknesses that can be exploited with the optimum noise, deception, or combination of jamming techniques. This information is then programmed into jamming systems to counter specific threats.

3. RADAR JAMMING EMPLOYMENT OPTIONS

There are currently two primary employment options for both noise and deception jamming techniques. These options are: (1) support jamming, and (2) self-protection jamming. Support jamming can be broken down further into stand-off jamming (SOJ), and escort jamming.

a. To counter early warning, ground control intercept (GCI), and acquisition radars associated with an enemy IADS, noise and deception jamming techniques are employed by specialized support jamming aircraft. The goal of support jamming is to create confusion and delays within the command and control structure of the IADS. Deny, delay or degrade the enemy's ability to engage friendly forces. Support jamming operations can be focused against a national level IADS through the use of a stand-off jamming (SOJ) profile (Figure 9-3) or against a target area threat array using an escort jamming profile.

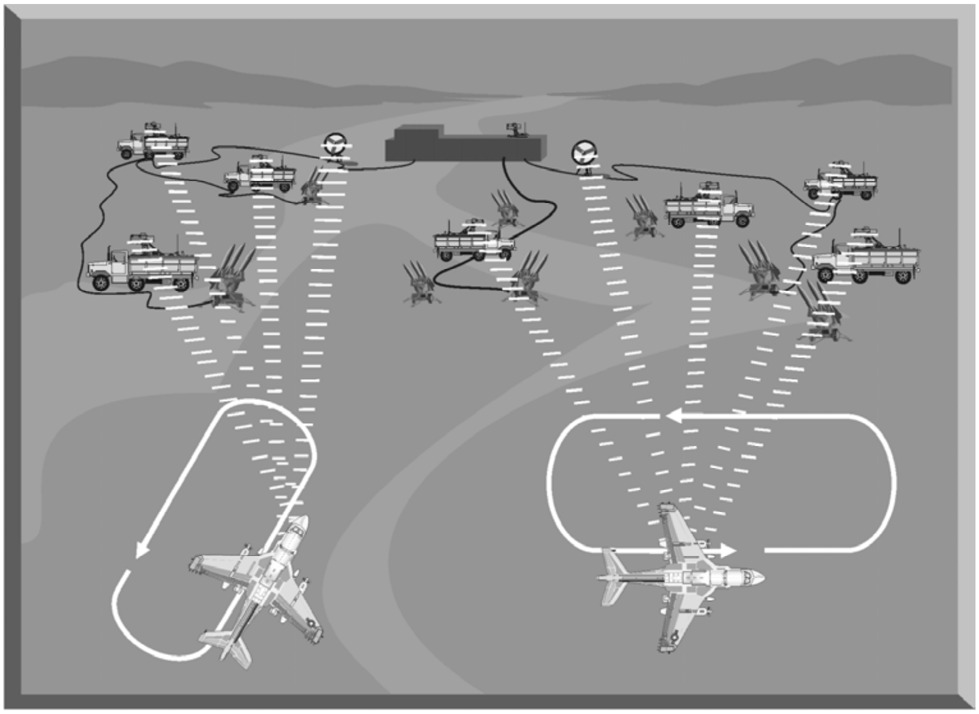

Figure 9-3. Stand-Off Jamming

(1) From an orbit area outside the surface-to-air missile (SAM) engagement zone, SOJ aircraft employ specialized jamming techniques to deny the enemy information about the attack package. SOJ aircraft employ specialized noise jamming techniques to generate jamming strobes on the victim radar display. This effectively denies range and azimuth information on aircraft ingressing and egressing the area covered by the noise jamming strobes. Intensity of the strobes is based on the power in the jamming. The area covered is based on the amount of jamming that can be injected into the main beam and sidelobes of the victim

radar. The effectiveness of SOJ noise jamming is determined by the power the jammer can generate relative to the power the victim radar can generate. This is called the jamming-to-signal (J/S) ratio.

(2) SOJ aircraft can also employ a deception technique to generate false targets to confuse the radar operator and mask the presence of real targets (Figure 9-4). In this specialized technique, the deception jammer must tune to the frequency, PRF, and scan rate of the victim radar. The jammer then transmits multiple jamming pulses that the victim radar receiver processes like real target returns. With enough power, the deception jammer can generate multiple false azimuth targets by injecting jamming pulses into the sidelobes of the victim radar. False moving targets and false range targets are generated by varying the time delay of the jamming pulses based on the PRF and scan rate of the victim radar.

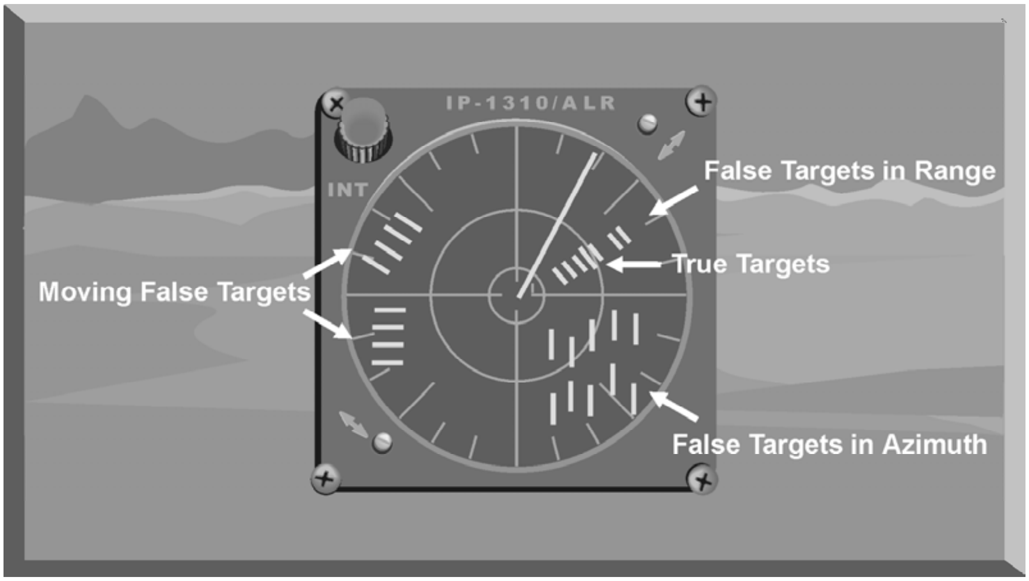

Figure 9-4. False Target Jamming

(3) Escort jamming is a specific tactic used by the EA-6B Prowler. The EA-6B is employed as an integral part of the attack package and is normally positioned behind and above the attack package (Figure 9-5).

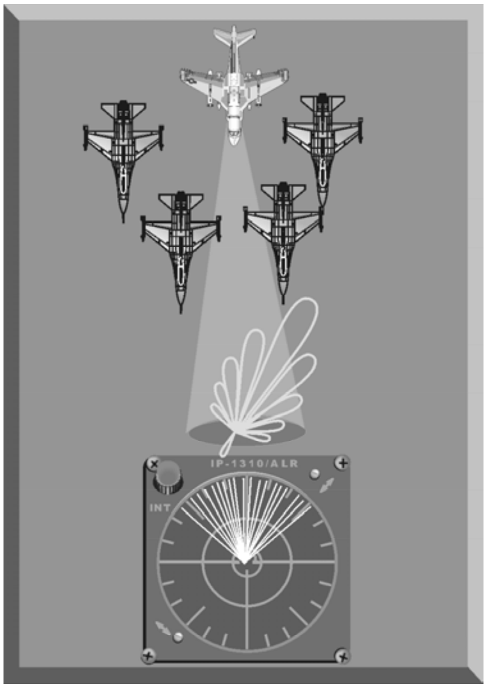

Figure 9-5. Escort Jamming

Using noise jamming, the EA-6B attempts to deny range and azimuth information to the victim radar by injecting high power signals into the main radar beam and sidelobes. To be effective, the EA-6B must be properly positioned in relation to the ingressing or egressing attack package (Figure 9-6).

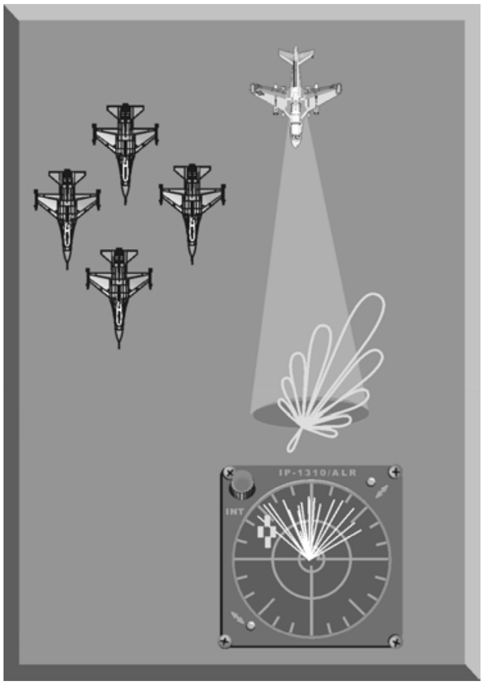

Figure 9-6. Escort Jamming Alignment

b. Self-protection radar jamming targets the radar systems that support jamming cannot negate. Self-protection jamming systems are part of a self-protection suite that includes a self-protection jamming pod, a chaff/flare dispenser, and on some aircraft, a towed decoy system. The overall purpose of these systems is individual aircraft survivability. These systems are designed to counter the individual SAM, AAA, and AI assets associated with the enemy IADS. They employ deception jamming techniques against the target tracking radars (TTRs) associated with these threats. They are designed to break the radar track or generate sufficient tracking errors to cause the missile or bullet to miss the aircraft.

(1) Self-protection radar jamming systems usually employ deception jamming techniques based on several factors. First, effective deception jamming techniques generally require less power than noise jamming techniques. Second, less power means less weight and space, which are very important considerations for modern tactical aircraft. Finally, deception jammers can be designed to jam multiple threats, which is a critical requirement for operations in a dense threat environment (Figure 9-7).

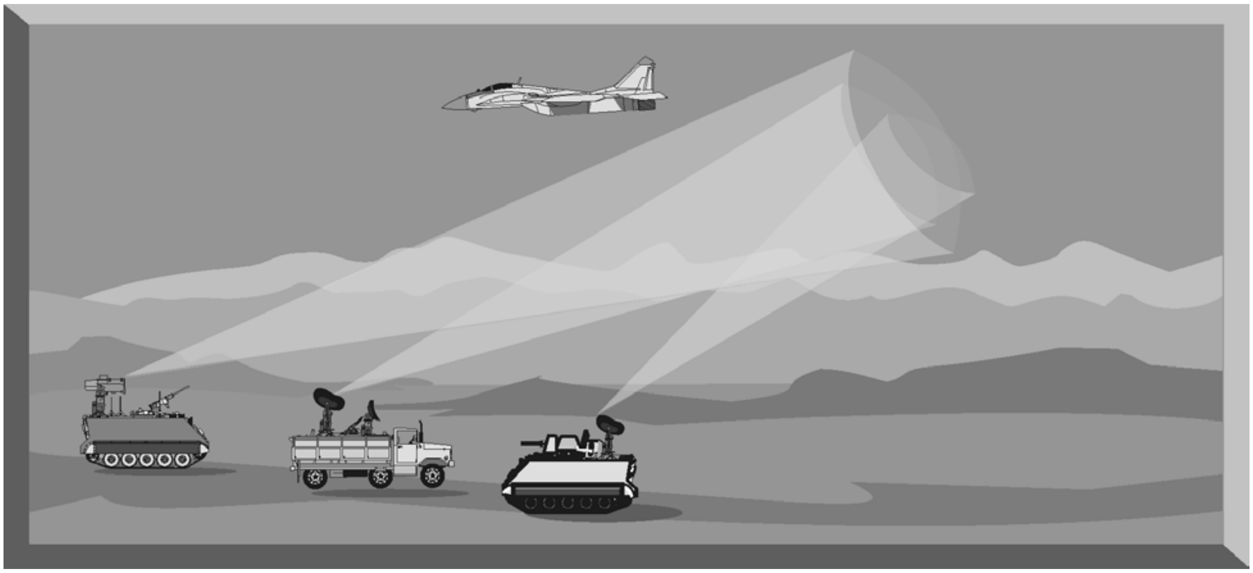

Figure 9-7. Self-Protection Jamming

(2) Despite the advantages of deception jamming techniques for self-protection jamming, there are some limitations that must be considered. First, deception jammers are complex electronic systems that must receive a victim radar's signal, memorize all its characteristics, modify the signal, and retransmit this modified signal at a high power level. Second, to be effective, deception jammers must be programmed with all the signal parameters (frequency, PRF, PRI, pulse width, scan rate, etc.) of the victim radar. Finally, because many deception techniques can be effective against specific threats, selecting optimum techniques to employ against these threats must be based on identified threat

system limitations. Identifying these specific threat systems limitations may be difficult.

4. FUNDAMENTALS OF RADAR JAMMING

There are some fundamental principles that apply to all types of jamming and to all jamming employment options. These principles are based on the characteristics of the jamming system and the characteristics of the victim radar. They include frequency matching, continuous interference, signal-to-noise ratio, jamming-to-signal ratio, and burnthrough range.

a. Based on the data provided by ES systems and intelligence evaluations, radar jamming systems must transmit signals at the frequency of the victim radar This applies to both noise and deception jamming. If a jamming signal does not match the transmitter frequency, the jamming signal is not received and displayed on the scope (Figure 9-8). When a jamming signal matches the transmitter frequency, the jamming signal is received and masks the target display (Figure 9-9).

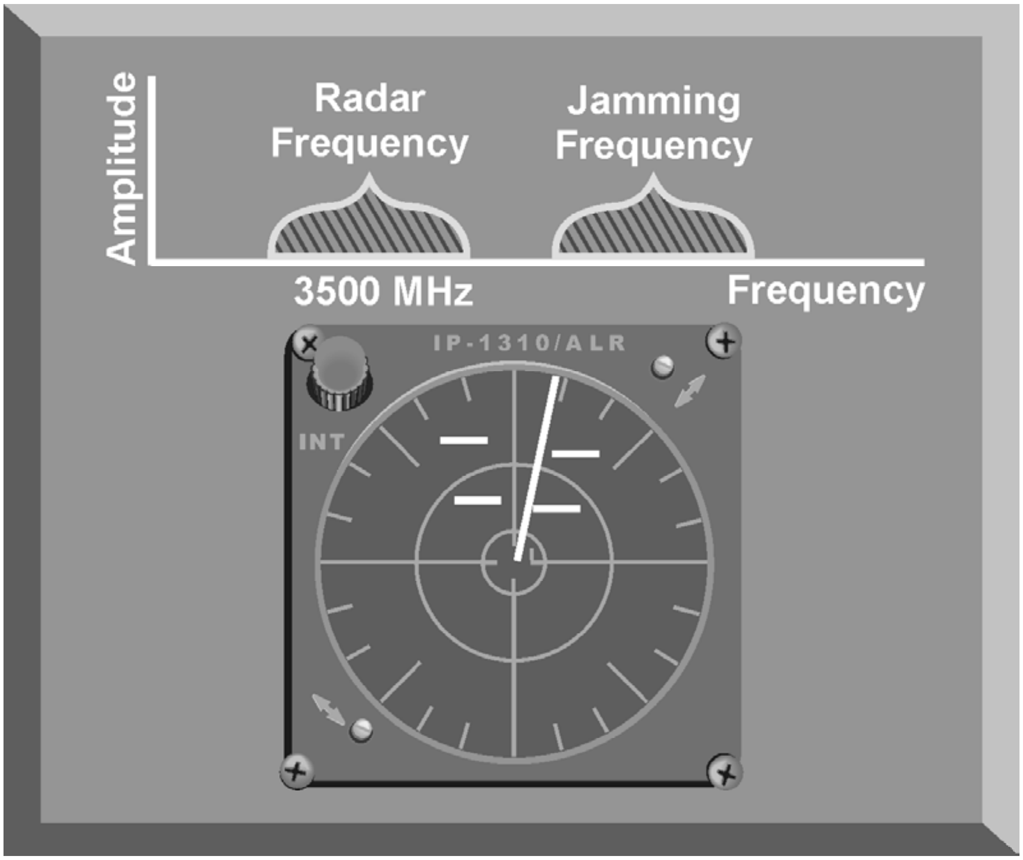

Figure 9-8. Jamming Frequency Error

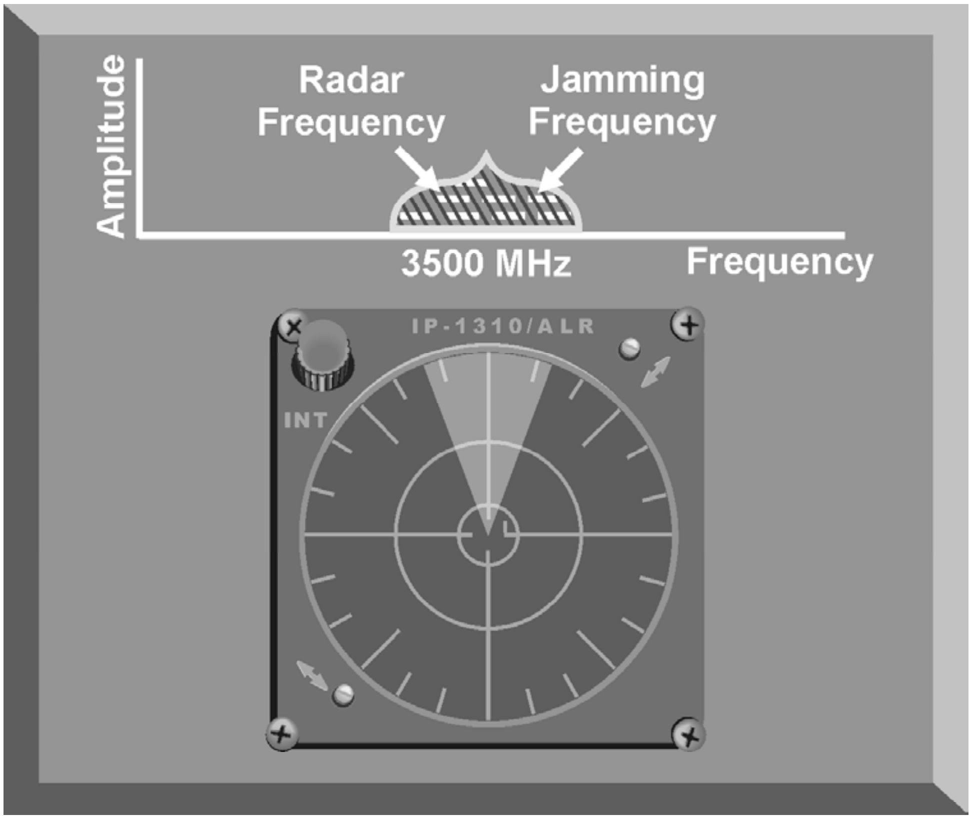

Figure 9-9. Correct Jamming Frequency Matching

b. For maximum effectiveness, a jamming transmitter should produce continuous interference. In much the same way intermittent static on a radio receiver does not completely block out a signal, intermittent jamming on a radar scope may not completely mask the target. An experienced radar operator or advanced automatic tracker can "read through" intermittent jamming and derive sufficient target information to negate jamming effectiveness. While true for noise jamming techniques, continuous interference also applies to deception techniques, especially when target reacquisition is considered.

c. The signal-to-noise (S/N) ratio is a measure of the ability of the victim radar to detect targets. It is also an indication of the vulnerability of the radar to certain jamming techniques, especially noise jamming.

(1) From the discussion of the basic radar equation in Chapter 5 (Equation 5-10), Equation 9-1 is the signal power density of a target return at the radar receiver. The signal power density of the target return is so weak that it requires very strong amplification before processing and display. Besides the signal power from the target, some level of thermal noise is also generated and amplified along with the target signal. For an "ideal" (no noise) amplifier, Equation 9-2 is used to compute the level of thermal noise generated by the amplifier.

$$\text{Signal Power Density} = \frac{P_T G \sigma A_e}{(4\pi)^2 R^4}$$

P_T = transmitted power
G = antenna gain
σ = target radar cross section (RCS)
A_e = antenna aperture area
R = range to the target

Equation 9-1. Signal Power Density

$$\text{Thermal Noise (N)} = KTBF$$

K = Boltzman's Constant
(1.38 x 10^{-23} watts/Hz degrees K)
T = standard temperature (290° K)
B = radar receiver equivalent bandwidth
F = radar receiver noise figure
(one for "ideal" receiver)

Equation 9-2. Thermal Noise

Note: The instantaneous bandwidth of a receiver is the frequency range over which the receiver can simultaneously amplify two or more signals to within a specified gain.

(2) The radar receiver amplifies both target signal and thermal noise. The output of the radar receiver will contain the target signal and the noise amplified across the bandwidth of the receiver. Separating the desired target signal from the undesired noise signal is one of the major problems confronting radar designers.

(3) Equation 9-3 is derived by dividing Equation 9-1 by Equation 9-2. Many factors in this equation fluctuate and must be estimated using statistical calculations. For example, target RCS fluctuates based on the changing angle of the antenna beam and corresponding changes in the reflected signal. Effective antenna aperture is also a statistical phenomenon based on the fluctuations in target RCS. The thermal noise generated by a receiver is also a fluctuating factor and must be treated statistically. This means that the S/N ratio is a statistical factor associated with a probability of target detection and a probability of a false alarm. A false alarm occurs when the radar operator or automatic tracking circuit designates a fluctuation in noise level as a target. The higher the S/N ratio, the

higher the probability of target detection with a corresponding reduction in the probability of a false alarm.

$$\text{Signal-to-Noise Ratio} = \frac{P_T G \sigma A_e \left(\frac{1}{(4\pi)^2 R^4} \right)}{KTBF}$$

P_T = transmitted power
G = antenna gain
σ = target radar cross section (RCS)
A_e = antenna aperture area
R = range to the target

K = Boltzman's Constant (1.38×10^{-23} watts/Hz degrees K)
T = standard temperature (290° K)
B = radar receiver equivalent bandwidth
F = radar receiver noise figure (one for "ideal" receiver)

Equation 9-3. Signal-to-Noise Ratio

(4) An analysis of Equation 9-3 suggests that any action that increases the power in the target signal (for example, increasing transmitted power, increasing antenna gain/aperture area, or decreasing target range) will improve the S/N ratio and improve the probability of target detection. It would also appear that decreasing the bandwidth of the radar receiver will increase the S/N ratio and enhance the probability of target detection. However, if the effective bandwidth of the receiver is reduced, this may eliminate a significant portion of the radar signal spectrum and decrease the probability of target detection.

(5) The S/N ratio is also an indication of the range at which a target will be detected. A plot of the receiver output of a typical radar is shown in Figure 9-10. The weak target signal at an extended range is just above the receiver noise level. The target at closer range is easily detected above the noise level. A radar operator or automatic target detector could mistake the very weak target return as a fluctuation in the receiver noise level. This could result in a missed detection. The lack of discrimination between noise and target returns because of a poor S/N ratio can also result in designating fluctuations in the noise level as actual target signals, known as false alarms.

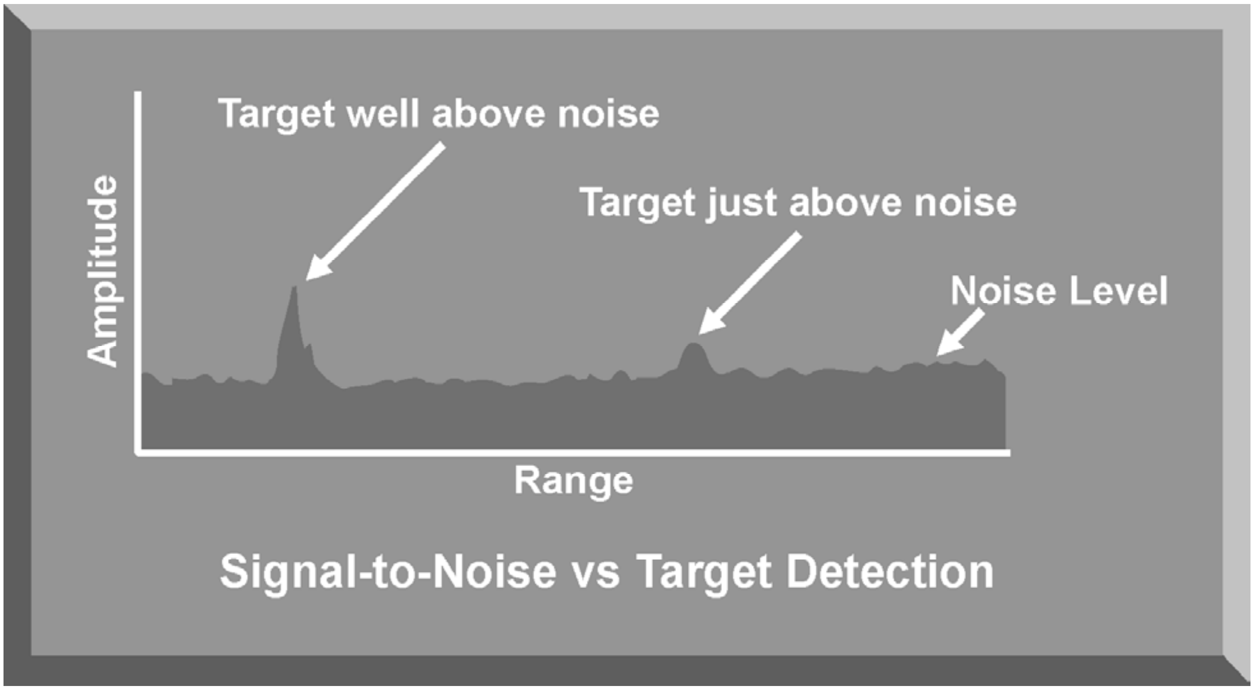

Figure 9-10. S/N Ratio and Target Detection

(6) To preclude, or minimize false alarms, the radar receiver may be equipped with electronic circuits to establish a false alarm threshold. If the signal strength of a radar return is below this threshold level, it will not be detected or displayed (Figure 9-11). This false alarm threshold also influences the probability of target detection. With the threshold set too high, many detected targets will not be displayed. Additionally, if the false alarm threshold is raised automatically in relation to the amplitude of the receiver noise, the radar receiver is more vulnerable to noise jamming.

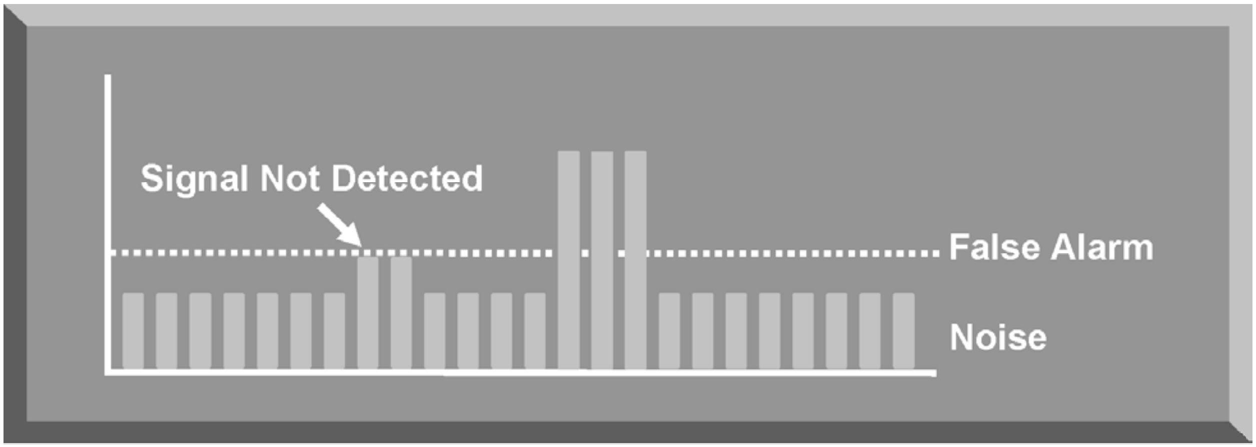

Figure 9-11. Receiver False Alarm Threshold

(7) For any target return to be detected by the radar, the S/N ratio must be greater than one. If the S/N ratio is less than one, the target will not be detected above the receiver noise level. The purpose of noise jamming is to raise the level of noise in the radar receiver to reduce the S/N ratio to less than one. This masks the presence of the true target return. If a false alarm threshold is used, noise jamming raises this threshold to further complicate target detection. Figure 9-12 depicts a S/N ratio greater than one. Figure 9-13 depicts a S/N ratio of less than one due to noise jamming.

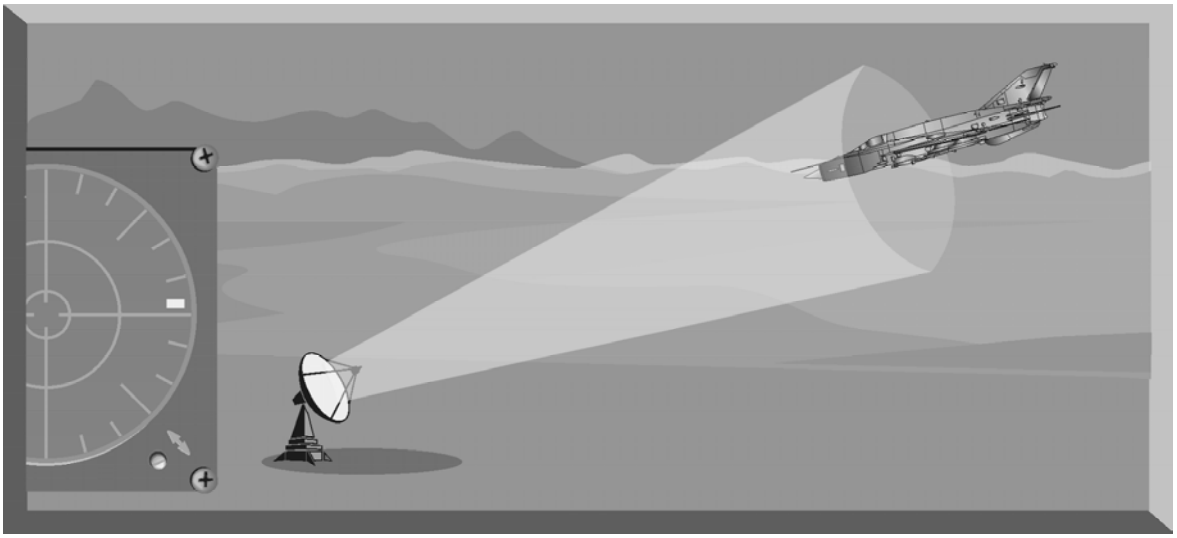

Figure 9-12. S/N Ratio Greater Than One

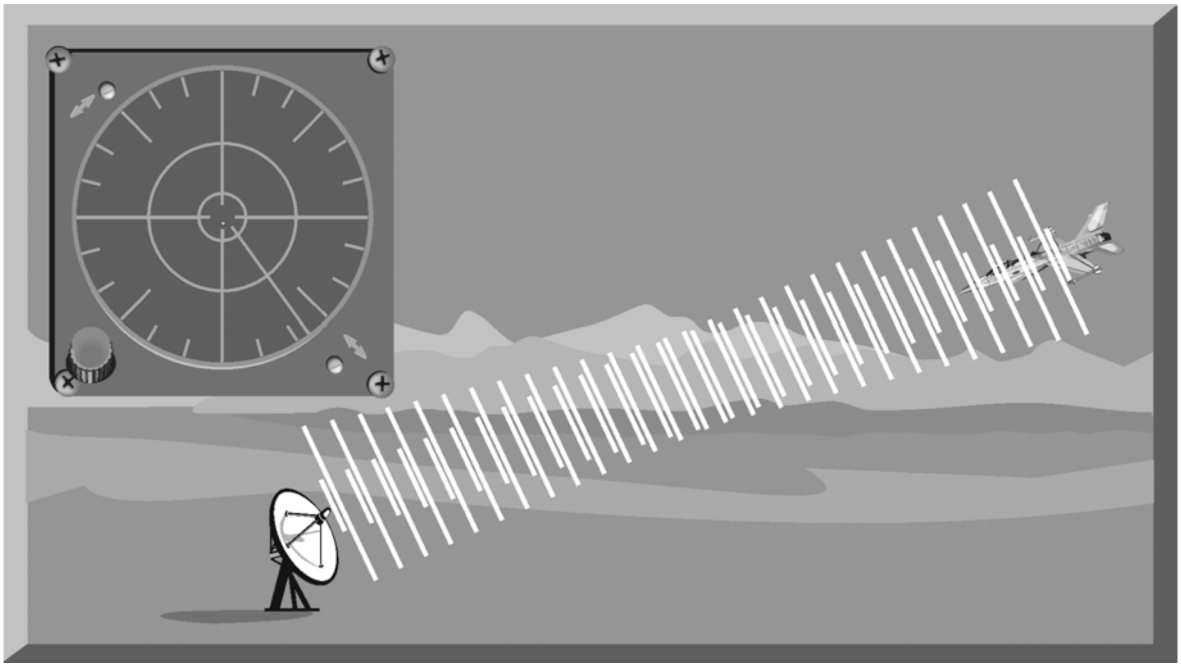

Figure 9-13. S/N Ratio Less Than One

d. The jamming-to-signal (J/S) ratio is a fundamental measure of jamming effectiveness. The J/S ratio compares the power in the jamming signal with the power in the radar return. Equation 9-4 is an expression of the J/S ratio. It is important to note that the J/S ratio should be measured at the output of the radar receiver. This will allow consideration of the receiver signal processing gain applied to the jamming signal.

$$\text{Jamming-to-Signal Ratio} = \frac{P_J G_J}{P_T G_T} \times \frac{4\pi R^2}{\sigma}$$

P_J = jamming power transmitted
G_J = jamming antenna gain
P_T = peak power transmitted by the radar
G_T = radar antenna gain
R = range from jammer to radar
σ = aircraft RCS

Equation 9-4. Jamming-to-Signal Ratio

(1) The most critical factor in both the S/N and the J/S ratios is range. The S/N ratio is calculated based on R to the fourth power. This equates to a signal traveling from the radar to the target, and back to the radar receiver. The J/S ratio is calculated using R to the second power. This factor reflects the "one way" transmission of the jamming pulse from the jammer to the victim radar's receiver.

(2) For a jamming signal to be effective, the J/S ratio must be greater than one. In general, threat radars, especially ground-based radars, transmit much more power than does an airborne jamming system. However, this power must travel twice as far as the airborne jamming signal. At long ranges, a low power jamming system can generate a J/S ratio much greater than one. In Figure 9-14, the jamming pulse completely masks the target return. As the jamming system approaches the target, the distance the radar pulse travels decreases with a corresponding increase of power in the radar return. This reduces the J/S ratio to a value less than one and the radar "sees" the target. This is called the burnthrough range (Figure 9-15).

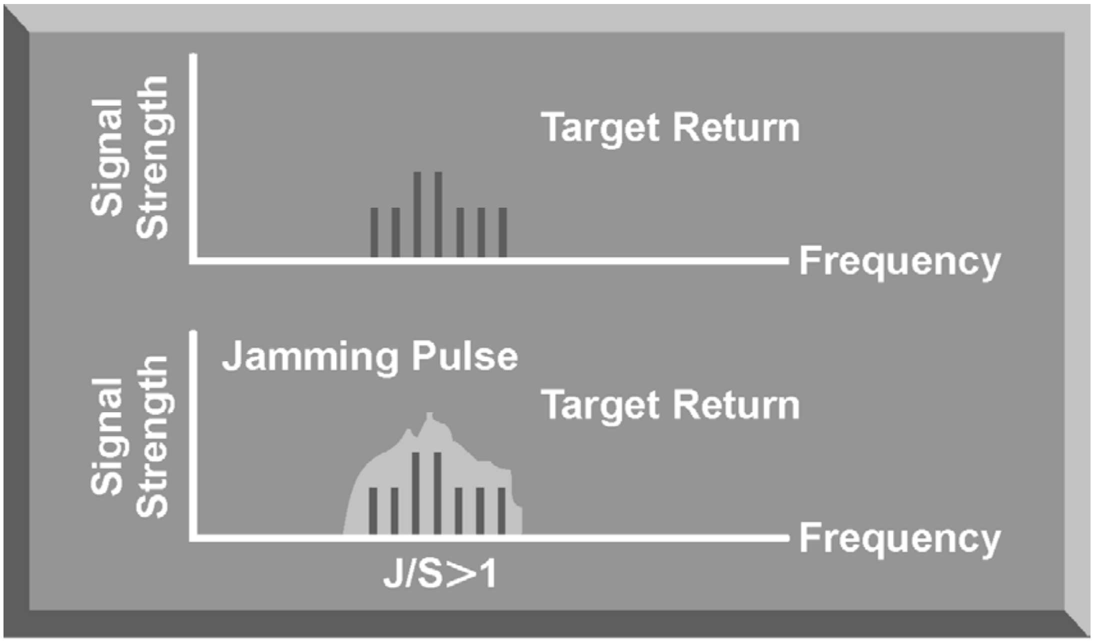
Signal Strength
Target Return
Frequency
Signal Strength
Jamming Pulse
Target Return
Frequency
J/S>1

Figure 9-14. J/S Ratio Greater Than One

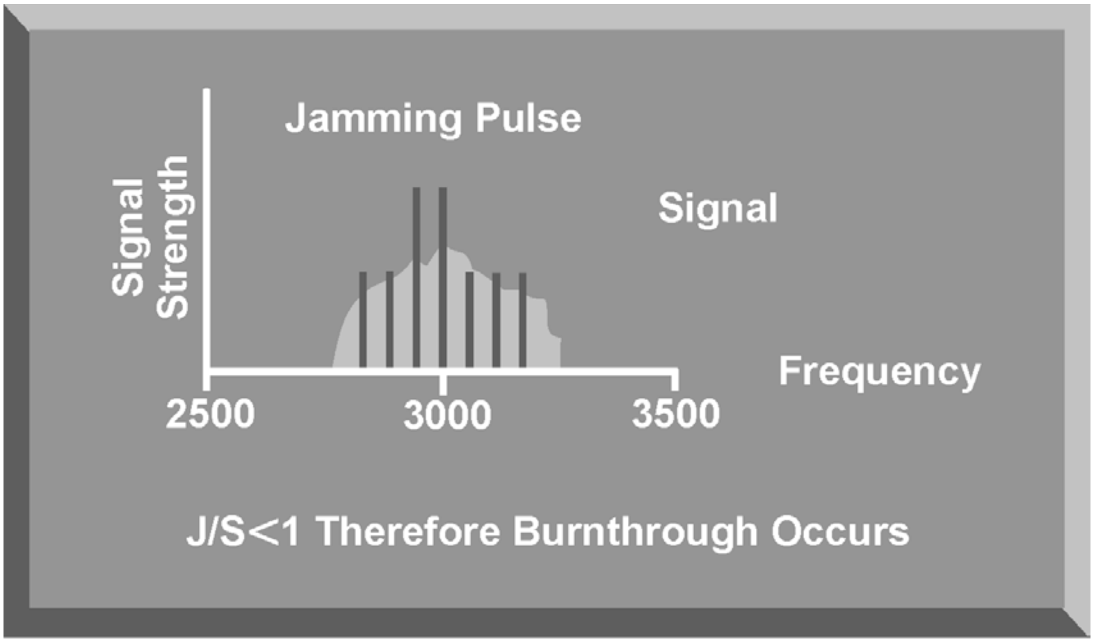
Jamming Pulse
Signal Strength
Signal
Frequency
2500
3000
3500
J/S<1 Therefore Burnthrough Occurs

Figure 9-15. J/S Ratio Less Than One

e. Burnthrough occurs when the power in the reflected target signal exceeds the power in the jamming signal. Even when an optimum and continuous jamming technique is transmitting on the exact frequency of the victim radar, the jamming starts to lose effectiveness as it nears the radar. For a particular radar

jamming technique, burnthrough range depends on the detection capability of the victim radar, expressed as the S/N ratio, and the capability of the aircraft's jamming system, expressed as the J/S ratio. The idea of burnthrough range explains why a jamming technique, especially noise jamming, loses its effectiveness as the aircraft approaches the radar. When plotting the jamming and signal power versus range (Figure 9-16), these two values intersect at the point where the J/S ratio is one. At closer ranges, the jamming pulse is no longer masking the aircraft, and the aircraft can be detected. Burnthrough range is the point where the radar can see through the jamming.

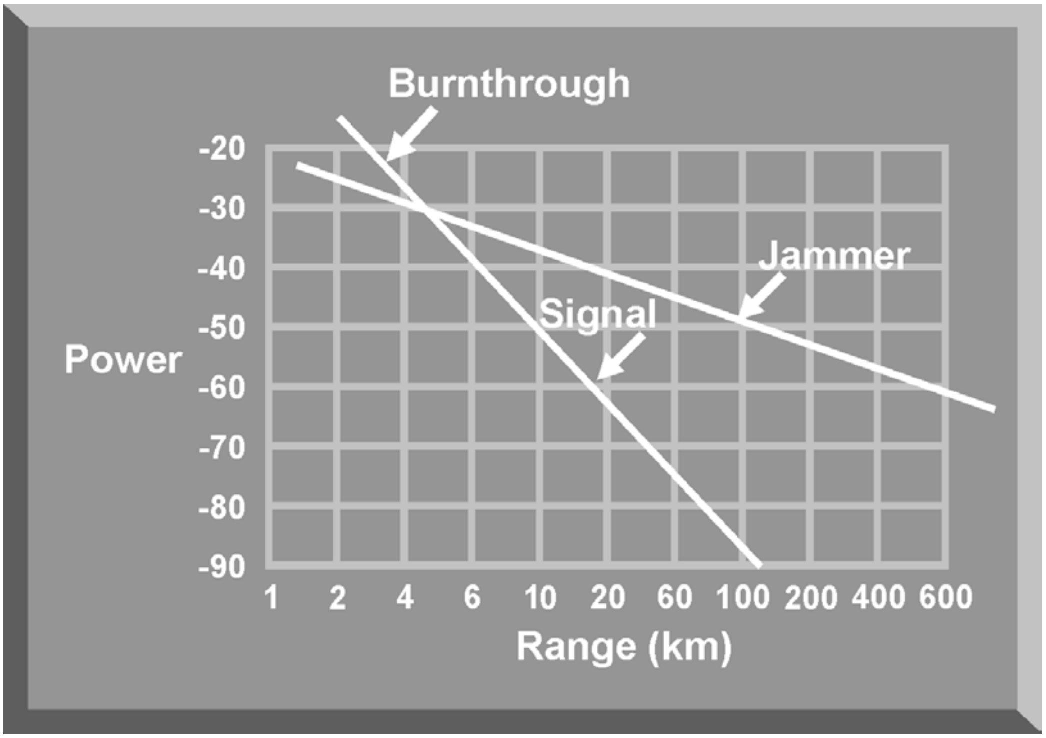

Figure 9-16. Burnthrough Range

5. SUMMARY

The purpose of radar jamming is to confuse or deny critical data to the radar systems that play a vital role in supporting the mission of an integrated air defense system. Two types of radar jamming, noise and deception, can be employed in a support-jamming role, or in a self-protection role for individual aircraft. The effectiveness of a jamming technique depends on the ability of the jamming system to generate a jamming signal that replicates the parameters of the victim radar, especially its frequency. The signal-to-noise ratio of the victim radar determines the vulnerability of the radar receiver to jamming while the jamming-to-signal ratio is an indication of the ability of the jamming system to effectively jam the victim radar. These basic radar jamming concepts are fundamental to understanding the impact of specific jamming techniques on radar systems.

CHAPTER 10. RADAR NOISE JAMMING

1. INTRODUCTION

A radar noise jamming system is designed to generate a disturbance in a radar receiver to delay or deny target detection. Since thermal noise is always present in the radar receiver, noise jamming attempts to mask the presence of targets by substantially adding to this noise level. Radar noise jamming can be employed by support jamming assets or as a self-protection jamming technique. Radar noise jamming usually employs high-power jamming signals tuned to the frequency of the victim radar. This chapter will discuss the factors that determine the effectiveness of radar noise jamming, radar noise jamming generation, and the most common noise jamming techniques. These noise jamming techniques include barrage, spot, swept spot, cover pulse, and modulated noise jamming.

2. RADAR NOISE JAMMING EFFECTIVENESS

The effectiveness of radar noise jamming depends on numerous factors. These factors include the jamming-to-signal (J/S) ratio, power density, the quality of the noise signal, and the polarization of the transmitted jamming signal (Figure 10-1).

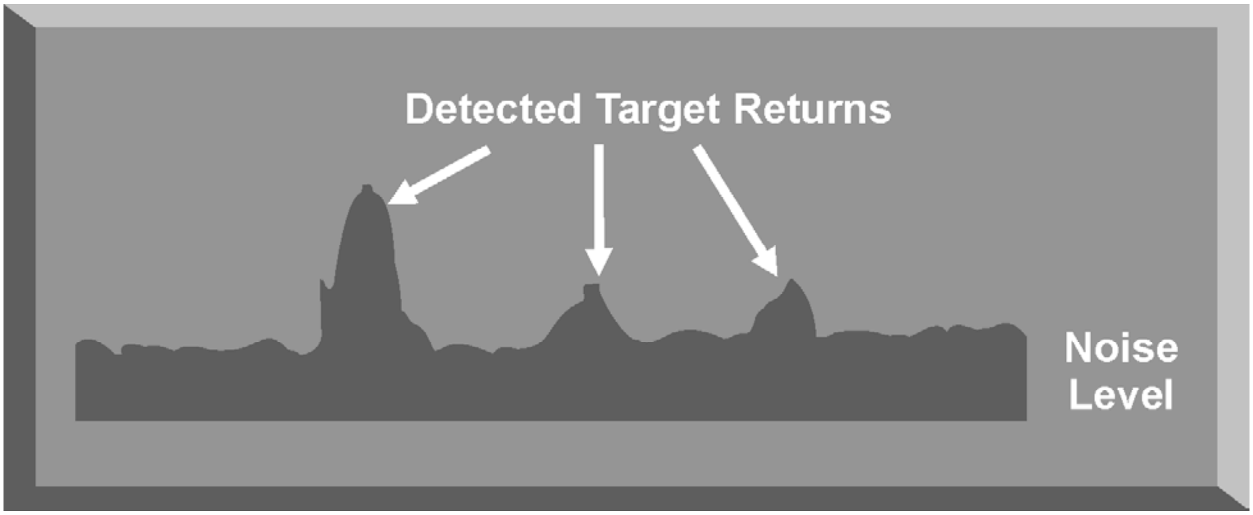

Figure 10-1. Noise Level in a Typical Radar Receiver Output

a. One of the most important factors that impacts the effectiveness of radar noise jamming is the J/S ratio (Figure 10-2). As discussed in Chapter 9, the power output of the noise jammer must be greater than the power in the target return, as measured at the output of the radar receiver. To achieve this level of jamming power, radar noise jammers usually generate high-power jamming signals. These high-power jamming signals can be introduced into the victim radar's main beam

to deny range information and into the victim radar's sidelobes to deny azimuth information.

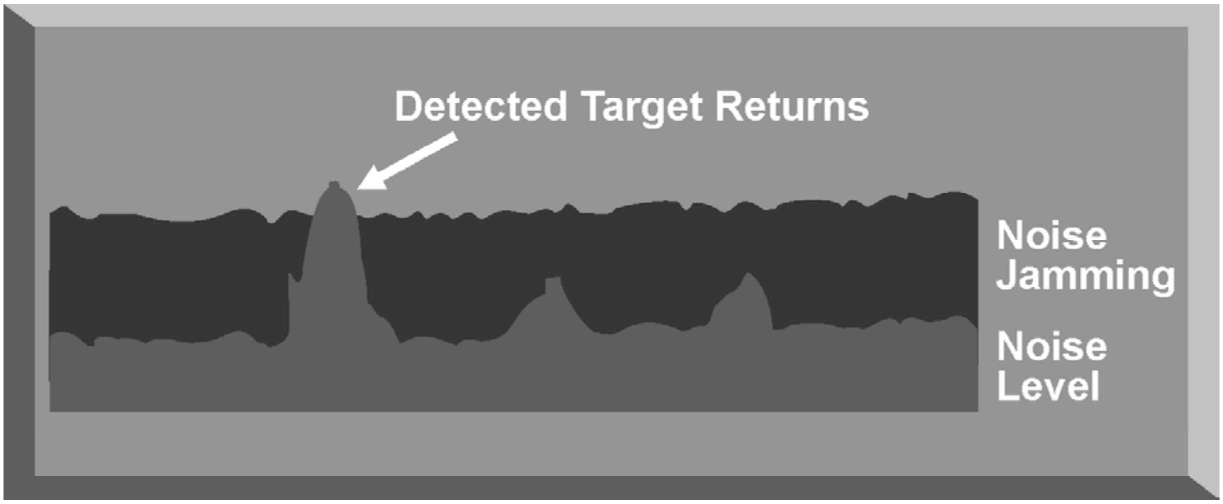
Detected Target Returns
Noise Jamming
Noise Level

Figure 10-2. Impact of Noise Jamming

b. Another factor which impacts the effectiveness of radar noise jamming is the power density. The power density of the noise jamming signal has a direct relation to the J/S ratio.

(1) If the noise jamming signal is centered on the frequency and bandwidth of the victim radar, the jamming signal has a high power density. The ability of a noise jammer to concentrate the jamming signal depends on the ability of the jammer to identify the exact frequency and bandwidth of the victim radar (Figure 10-3).

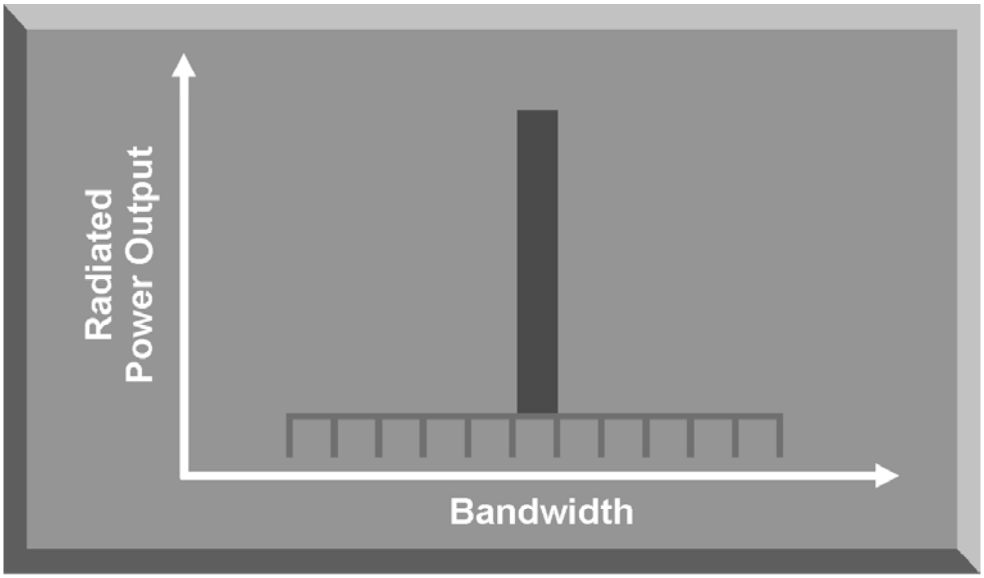
Radiated Power Output
Bandwidth

Figure 10-3. Power Density – Narrow Bandwidth

(2) If the generated noise jamming signal has to cover a wide bandwidth or frequency range, the power density at any one frequency is reduced (Figure 10-4). Radar systems that are frequency agile or that employ a wide bandwidth can reduce, or negate, the effectiveness of noise jamming by reducing the power density of the jamming signal.

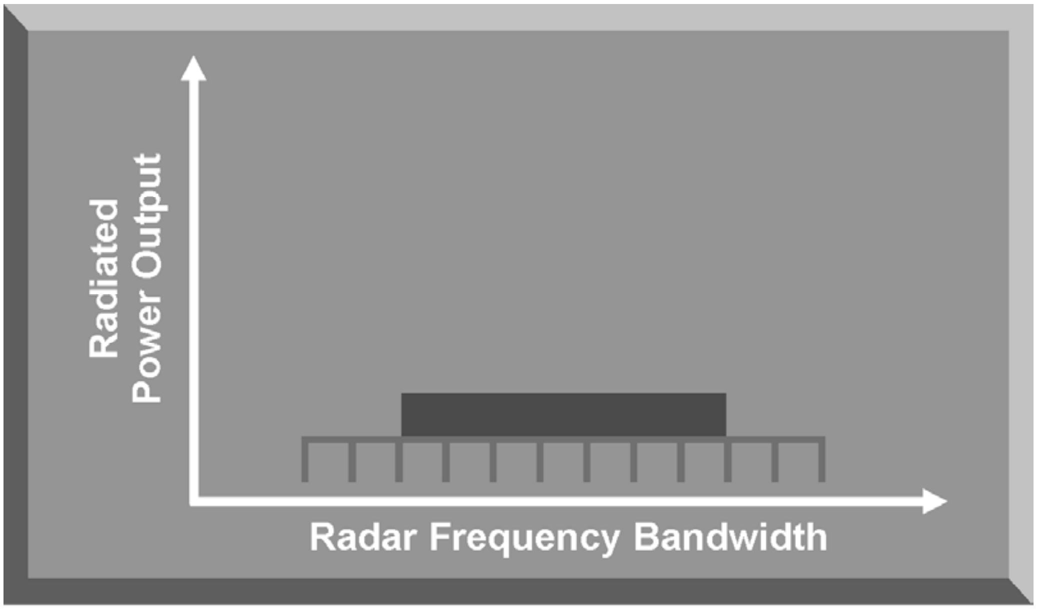

Figure 10-4. Power Density – Wide Bandwidth

c. The quality of the noise jamming also determines its effectiveness. To effectively jam a radar receiver with noise, the jamming signal must emulate the thermal noise generated by the receiver. This ensures that the radar operator or automatic detection circuit cannot distinguish between the noise jamming and normal thermal noise. Thermal noise is referred to as white noise and has a uniform spectrum. All of the frequencies in the bandwidth of the receiver have the same spectrum and an amplitude that varies based on Gaussian distribution. A Gaussian distribution is simply a bell-shaped distribution of amplitudes. In order to be effective, the jamming signal should exactly match the characteristics of the thermal noise signal of the victim radar receiver.

d. Polarization of the noise jamming signal is another significant factor that impacts its effectiveness. As discussed in Chapter 2, if the polarization of the jamming signal does not match the antenna polarization of the victim radar, there is a significant power loss in the jamming signal. Noise jamming systems designed to counter multiple threat radars, with various polarizations, generally use a transmitting antenna with a 45° slant or use circular polarization. Most threat systems are horizontally or vertically polarized. This results in a 50% reduction in effective radiated power (ERP) for most threat systems. A more serious power loss, nearly 100%, in ERP occurs when the jamming antenna is

orthogonally polarized with the victim antenna. The polarization of the noise jamming signal impacts the J/S ratio and the power density.

3. RADAR NOISE JAMMING GENERATION

Noise jamming is produced by modulating an RF carrier wave with random amplitude or frequency changes, called noise, and retransmitting that wave at the victim radar's frequency. Since noise from numerous sources is always present and displayed on a radar scope, noise jamming adds to the problem of target detection. Reflected radar pulses from target aircraft are extremely weak. To detect these pulses, a radar receiver must be very sensitive and be able to amplify the weak target returns. Noise jamming takes advantage of this radar characteristic to delay or deny target detection.

a. The simplest method of generating a high-power Gaussian noise jamming signal is to employ a highly amplified diode to generate a noise signal at the frequency of the victim radar. This signal is filtered and directly amplified to the maximum power that can be generated by the transmitter. This method is called direct noise amplification (DINA). The DINA method of noise generation has a serious limitation. The maximum power available from linear wideband power amplification is extremely limited. Employing any other form of power amplification would alter the Gaussian distribution of the jamming signal. This method of generating radar noise jamming was used extensively during WW II.

b. Modern noise jamming systems generate noise jamming signals by frequency modulating a carrier wave at the frequency of the victim radar. FM noise jammers employ a receiving antenna to intercept the victim's radar signal. The antenna passes the victim radar signal to the receiver for identification. The receiver also tunes the jamming signal generator to the correct frequency. The receiver uses an automatic frequency control (AFC) circuit to tune the voltage-controlled oscillator (VCO) to the frequency of the victim radar. A noise signal is generated by the jamming signal generator and added to the tuning voltage of the VCO to get an FM jamming signal. This signal is sent to a traveling wave tube (TWT) power transmitter. The TWT is normally operated in a saturated mode which produces a high-power jamming signal that covers a wider bandwidth than the victim radar. This reduces the power density of the signal, but the high power levels available from the TWT amplification of an FM signal compensate for this loss. The signal is sent to the transmitting antenna and directed toward the victim radar.

c. Figure 10-5 highlights an important feature of a modern radar noise jamming system: a look-through capability. A look-through mode allows the receiver to periodically sample the signal environment. The objective of the look-through mode is to allow the jammer to update victim radar parameters and change the jamming signal to respond to changes in the signal environment. This greatly enhances the effectiveness of noise jamming systems. One method used to provide a look-through capability is to isolate the transmit and receive

antennas to allow continuous operation of the receiver to update signal parameters. Another method is to switch off the jammer for a brief period to allow the receiver to sample the signal environment. Since this latter look-through method eliminates the jamming signal, the amount of time the jammer is switched off must be kept to a minimum.

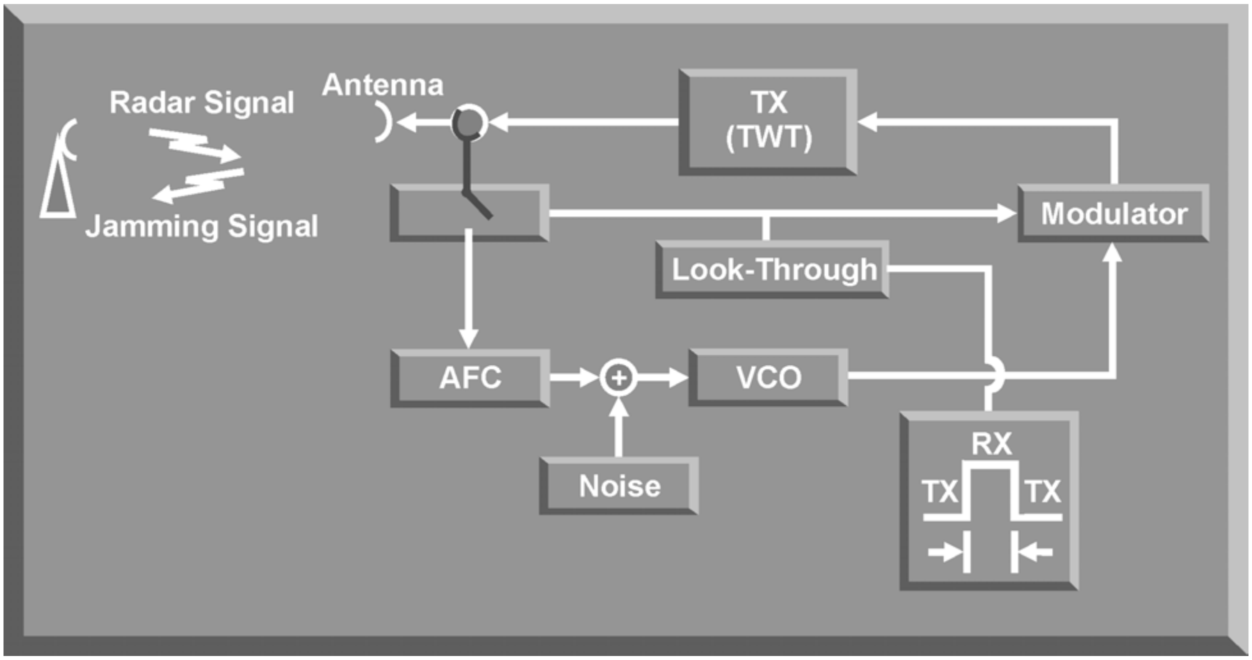

Figure 10-5. Frequency Modulated (FM) Noise Jamming System

4. BARRAGE JAMMING

An important aspect of jamming power is power density. Noise jamming depends on power density for its effectiveness. Power density is a function of the frequency range, or bandwidth, of the jamming signal. If a jammer covers a narrow frequency range, it can concentrate energy in a narrow band. If a jammer covers a wide frequency range, the energy is spread over that entire range. Since the jammer has fixed radiated power, this lowers the effective jamming power at a given frequency. Barrage jamming is a jamming technique where high power is sacrificed for the continuous coverage of several radar frequencies (Figure 10-6). The jamming signal is spread over a wide frequency range, which lowers the ERP at any one frequency. This type of jamming is useful against frequency-agile radars, against a radar system that uses multiple beams, or against multiple radar systems operating in a specific frequency range. By spreading the jamming over a wide frequency range, there is some level of jamming no matter what frequency the radar uses. Barrage jamming was used extensively during World War II. Advantages of barrage jamming are its simplicity and ability to cover a wide portion of the electromagnetic spectrum. The primary disadvantage is the low power density, especially when a high J/S ratio is needed against modern radars.

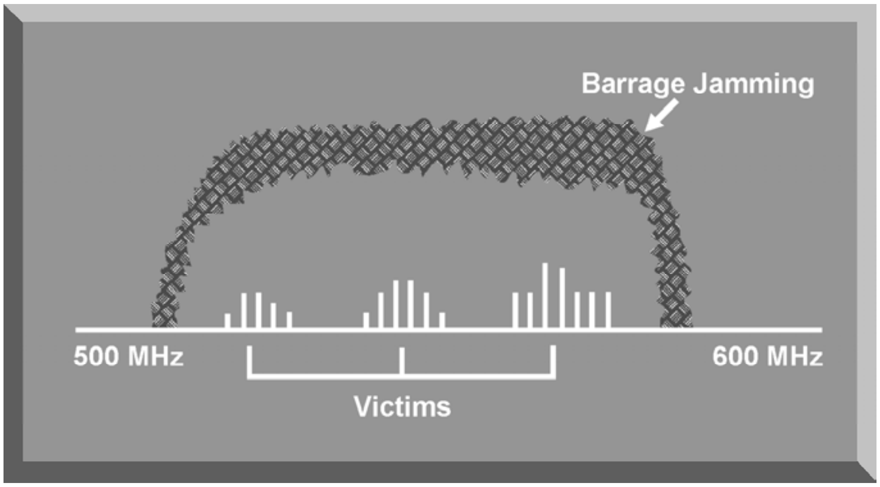

Figure 10-6. Barrage Jamming

5. SPOT JAMMING

One way to take advantage of the noise jammer's simplicity, but raise the jamming signal power, is to use a spot jammer. The earliest spot jammers were very narrow band jammers covering a bandwidth of 10 megahertz or less (Figure 10-7). This narrow band spot jammer was tuned to the anticipated frequency of the target radar. When it is necessary to jam a number of radars at different frequencies, more than one jammer is used. One problem that developed was of carrying the required number of spot jammers to counter a modern IADS. Also, radars that change their operating frequency, or are frequency-agile, defeat the spot jammer. Today, intercept panoramic receivers work with spot jammers to determine the frequency of the victim radar. A look-through capability is included in the system so that the target radar signal can be monitored to assess jamming effectiveness. The jamming signal can be adjusted for any changes in the operating frequency of the radar.

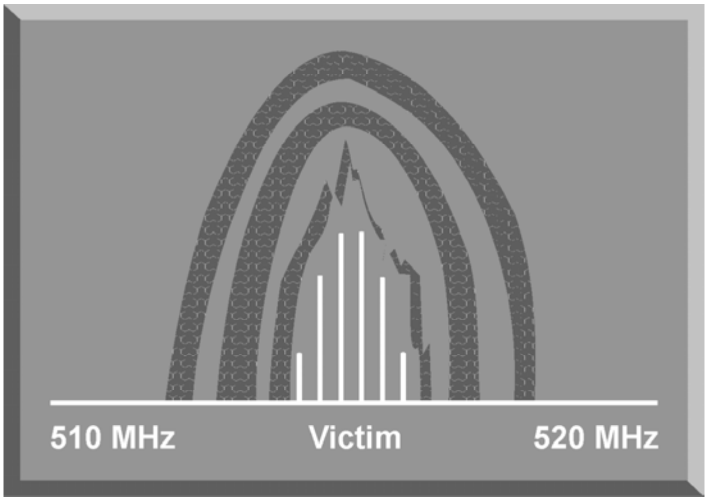

Figure 10-7. Spot Jamming

a. The primary advantage of spot jamming is its power density. Radar or communications receivers can be countered at longer ranges than when using a barrage jammer of equal output power (Figure 10-8).

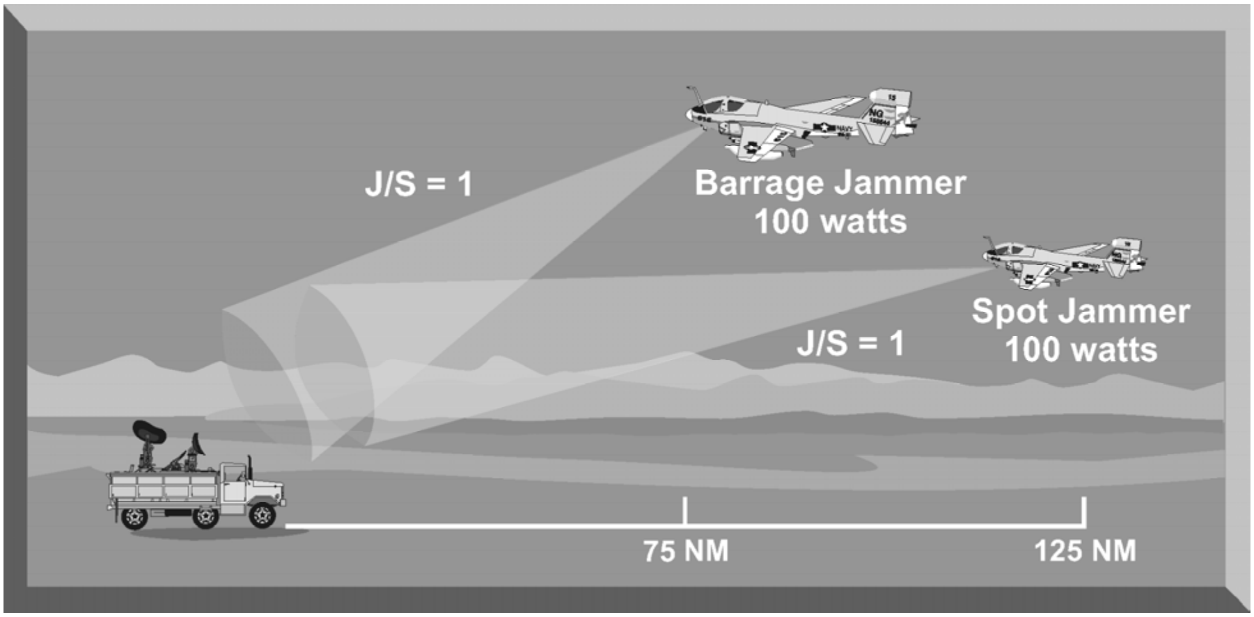

Figure 10-8. Spot Jamming Effectiveness

b. A disadvantage of the spot jammer is its coverage of a narrow band of the frequency spectrum. An operator or computer in the receiver must constantly monitor and tune the jamming signal to the target radar's frequency. The complexity of this process increases when jamming frequency-agile radars that can change frequencies with every pulse.

6. SWEPT-SPOT JAMMING

When high power density is required over a large bandwidth, one solution is to take spot jamming and sweep it across a wide frequency range (Figure 10-9). This preserves the high power density but allows the jamming to cover a large bandwidth. The jamming spot is swept across a broad frequency range at varying speeds. With this technique, a number of radar systems can be covered. Because of their high jamming power, swept-spot jammers are able to cover a number of radars operating in a broad frequency range. However, jamming is not continuous. Fast swept-spot jamming can approximate continuous jamming by causing a phenomenon known as “ringing.” Fast sweeping spot noise is like a burst of energy which sets up vibrations within the receiver section. When these vibrations last until the next burst of energy is received, this is known as ringing. Three factors determine swept-spot jamming effectiveness. The first is the power in the spot. The next is the bandwidth, or frequency range, the spot covers. The last is the sweep rate.

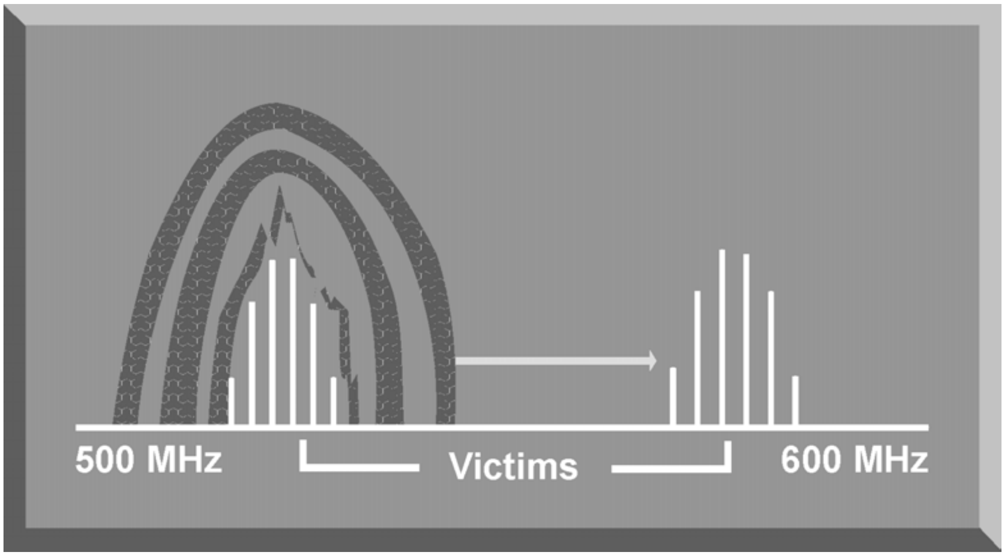

Figure 10-9. Swept-Spot Jamming

7. COVER PULSE JAMMING

Cover pulse jamming is a modification of swept-spot jamming. This is a "smart noise" technique that is responsive for a short period of time (Figure 10-10). A repeater jammer acts as a transponder. It receives several radar pulses and determines the PRF of the victim radar. It then uses this data to predict when the next radar pulse should arrive. Using an oscillator that is gated for a period of time based on predicted pulse arrival time, a noise-modulated signal is amplified and transmitted. This process works against a radar with a steady PRF, and allows a low-powered repeater to respond to a number of threats by time-sharing.

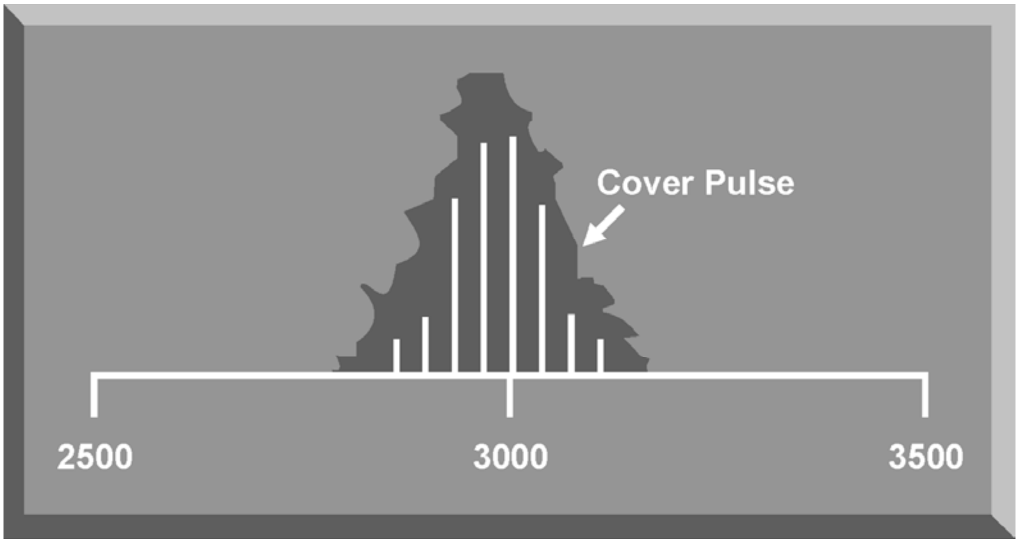

Figure 10-10. Cover Pulse Jamming

a. Cover pulse jamming is used to initiate a range gate pull-off (RGPO) deception jamming technique. The deception jammer transmits a noise jamming signal, or cover pulse, that is much stronger than the target return. The cover pulse raises the automatic gain inside the range gate, and the range tracking loop initiates tracking on the cover pulse. The deception jammer then increases the time delay in the jamming pulse and moves the range tracking gate away from the real target (Figure 10-11).

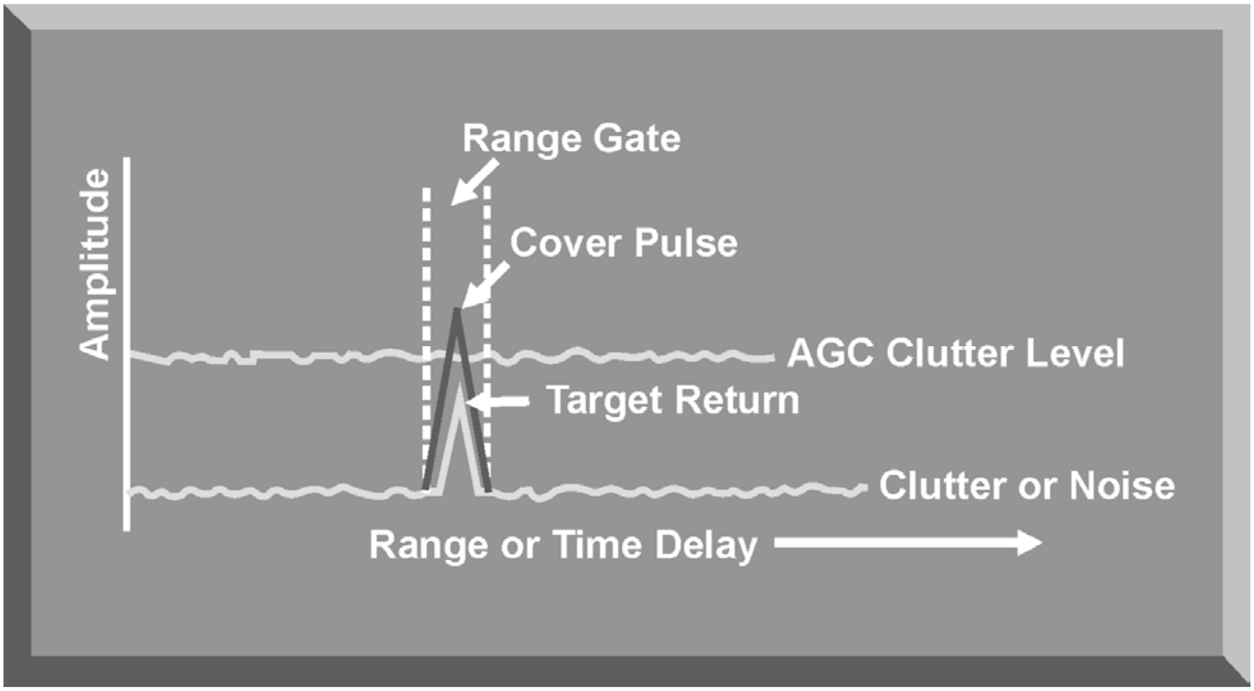

Figure 10-11. Range Gate Pull-Off Cover Pulse

b. A form of cover pulse jamming is also used to initiate a velocity gate pull-off (VGPO) technique against continuous wave and pulse Doppler radars. The cover pulse, in this case, is a strong jamming signal with the same frequency shift as the aircraft return. This cover pulse steals the velocity tracking gate and sets up the velocity tracking loop to steal the velocity tracking gate based on false target Doppler shifts (Figure 10-12).

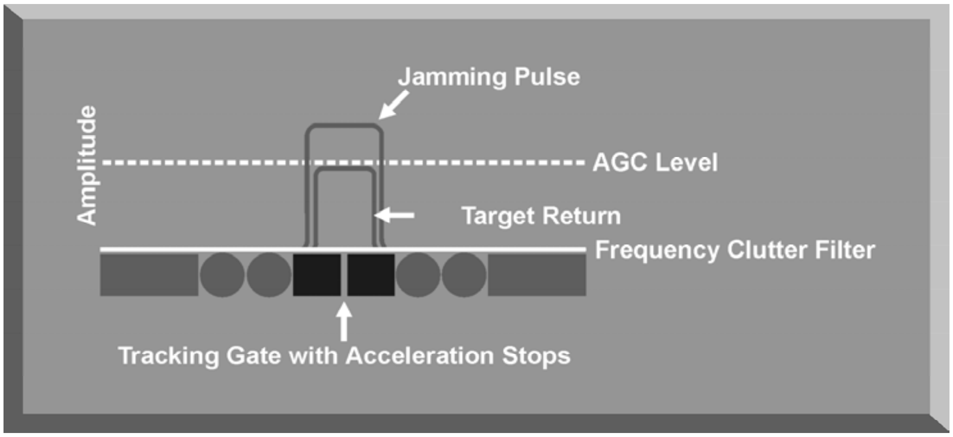
Jamming Pulse
Amplitude
AGC Level
Target Return
Frequency Clutter Filter
Tracking Gate with Acceleration Stops

Figure 10-12. Velocity Gate Pull-Off Cover Pulse

8. MODULATED NOISE JAMMING

Modulated jammers are special hybrid jammers which employ noise jamming that is either amplitude or frequency modulated. The purpose of this modulated noise is to defeat target tracking radars (TTRs) rather than deny range information. Modulated noise jamming has proven effective against conical scan and track-while-scan (TWS) TTRs.

a. Modulated jamming alters the noise jamming signal at a frequency that is related to the scan rate of the target radar. If modulated jamming is used against a conical scan radar, a sine wave signal is used (Figure 10-13). The frequency of the sine wave is slightly higher than the scan rate of the victim radar. The amplitude difference results in a constantly varying phase between the radar and the jamming signal. This phase differential produces false targets with a strong signal amplitude everywhere the signals reinforce each other. This causes the conical scan radar to track the false returns and lose the real target return. For this technique to work, the scan rate of the intended victim radar must be known.

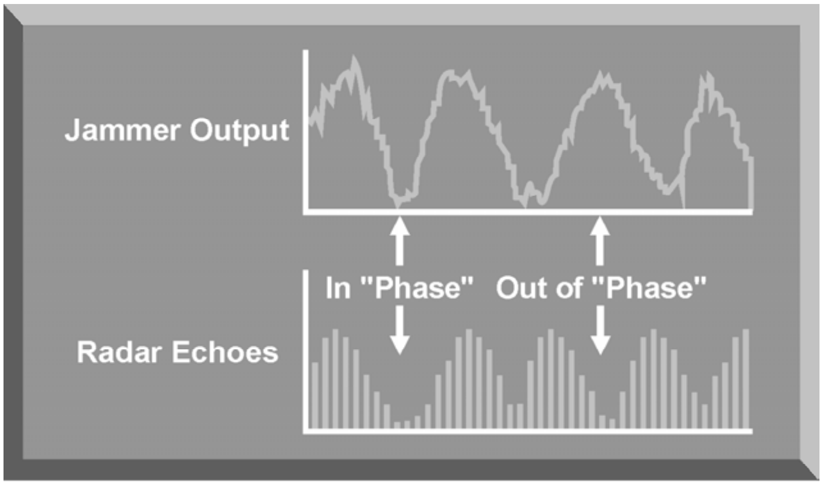
Jammer Output
In "Phase"
Out of "Phase"
Radar Echoes

Figure 10-13. Conical Scan Modulated Jamming

b. Against a TWS radar, a rectangular waveform is used to modulate the noise signal. The PRF of the modulation is set at some harmonic of the TWS rate. This synchronization results in a number of jamming strobes on the radar scope. Each jamming strobe is at a different azimuth or elevation depending on which radar beam is being jammed. The number of jamming strobes depends directly on the harmonic used to modulate the signal. In Figure 10-14, a modulating signal frequency that is four times the scan rate of the radar will produce four jamming strobes on the scope. If the jamming is slightly out of tune with the scan rate, the jamming strobes will appear to roll across the radar scope.

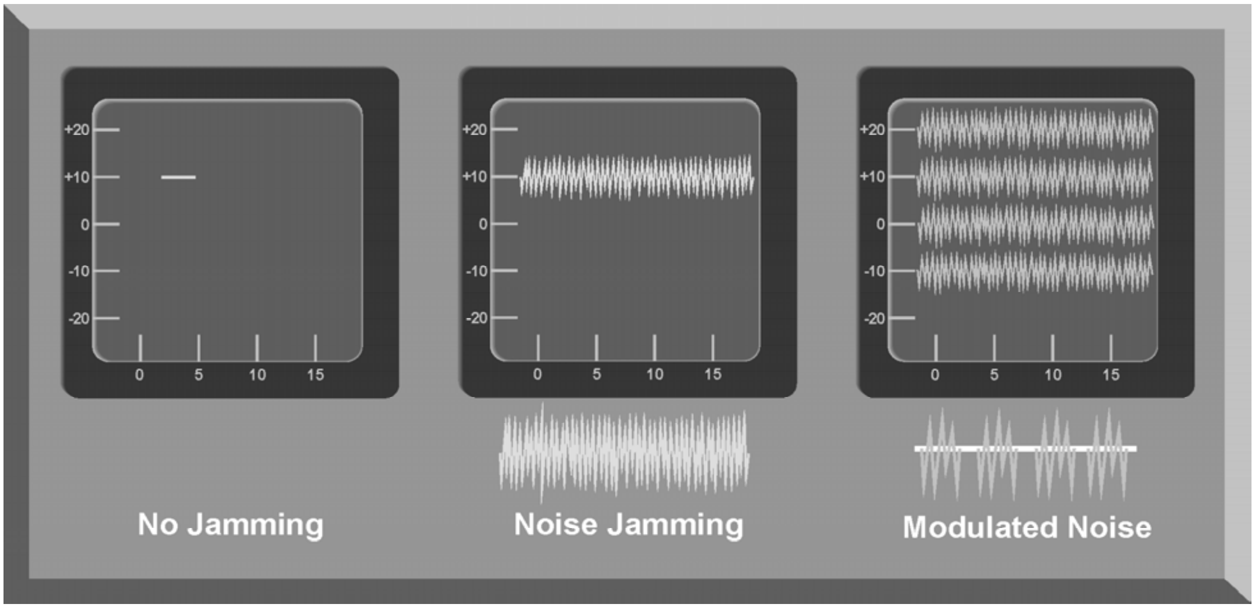

Figure 10-14. TWS Modulated Jamming

9. SUMMARY

Radar noise jamming is employed to deny target acquisition and target tracking data to a victim radar. This is accomplished by injecting amplitude or frequency modulated noise jamming signals into the victim radar's receiver. The radar noise jamming techniques discussed in this chapter included barrage, spot, swept-spot, cover pulse, and modulated jamming. The effectiveness of these noise jamming techniques depends on the power density of the jamming signal compared to the power in the radar return, or the J/S ratio. Radar noise jammers are generally simple, high-power systems which can be effectively employed in a support or self-protection role. Radar noise jamming can be employed in conjunction with deception jamming techniques to maximize the impact of jamming on victim radars.

CHAPTER 11. DECEPTION JAMMING

1. INTRODUCTION

Deception jamming systems are designed to inject false information into a victim radar to deny critical information on target azimuth, range, velocity, or a combination of these parameters. To be effective, a deception jammer receives the victim radar signal, modifies this signal, and retransmits this altered signal back to the victim radar. Because these systems retransmit, or repeat, a replica of the victim's radar signal, deception jammers are known as repeater jammers. The retransmitted signal must match all victim radar signal characteristics including frequency, pulse repetition frequency (PRF), pulse repetition interval (PRI), pulse width, and scan rate. However, the deception jammer does not have to replicate the power of the victim radar system.

a. A deception jammer requires significantly less power than a noise jamming system. The deception jammer gains this advantage by using a waveform that is identical to the waveform the radar's receiver is specifically designed to process. Therefore, the deception jammer can match its operating cycle to the operating cycle of the victim radar instead of using the 100% duty cycle required of a noise jammer. To be effective, a deception jammer's power requirements are dictated by the average power of a radar rather than the peak power required for a noise jammer. In addition, since the jammer waveform looks identical to the radar's waveform, it is processed like a real return. The jamming signal is amplified by the victim radar receiver, which increases its effectiveness. The reduced power required for effective deception jamming is particularly significant when designing and building self-protection jamming systems for tactical aircraft that penetrate a dense threat environment. Deception jamming systems can be smaller, lighter, and can jam more than one threat simultaneously. These characteristics give deception jammers a great advantage over noise jamming systems.

b. Although deception jammers require less power, they are much more complex than noise jammers (Figure 11-1). Memory is the most critical element of any deception jammer. The memory element must store the signal characteristics of the victim radar and pass these parameters to the control circuitry for processing. This must be done almost instantaneously for every signal that will be jammed. Any delay in the memory loop diminishes the effectiveness of the deception technique. Using digital RF memory (DRFM) reduces the time delay and enhances deception jammer effectiveness. Deception jamming employed in a self-protection role is designed to counter lethal radar systems. To be effective, deception jamming systems must be programmed with detailed and exact signal parameters for each lethal threat.

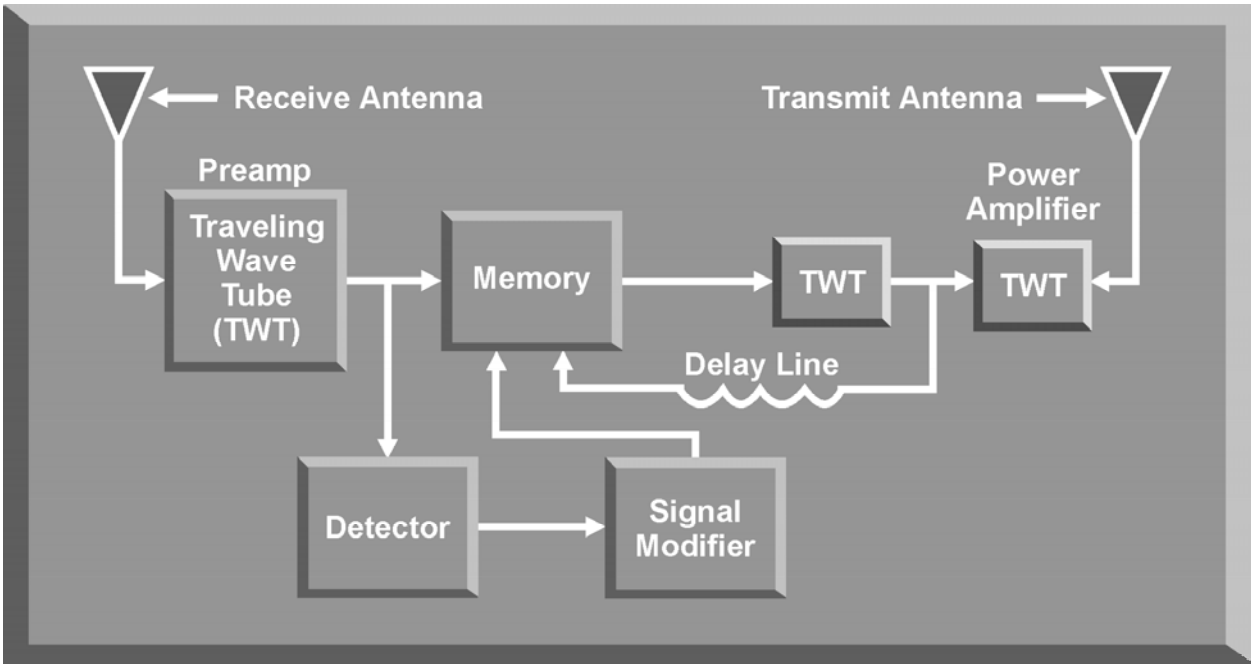

Figure 11-1. Deception Jamming System

c. The requirement for exact signal parameters increases the burden on electronic warfare support (ES) systems to provide and update threat information on operating frequency, PRF, PRI, power pulse width, scan rate, and other unique signal characteristics. An electronic intelligence (ELINT) architecture is required to collect, update, and provide changes to deception jamming systems. In addition, intelligence and engineering information on exactly how a specific threat system acquires, tracks and engages a target is essential in identifying system weaknesses. Once a weakness has been identified, an effective deception jamming technique can be developed and programmed into a deception jammer. For example, if a particular radar system relies primarily on Doppler tracking, a Doppler deception technique will greatly reduce its effectiveness. Threat system exploitation is the best source of detailed information on threat system capabilities and vulnerabilities. Effective deception jamming requires much more intelligence support than does noise jamming.

d. Most self-protection jamming techniques employ some form of deception against a target tracking radar (TTR). The purpose of a TTR is to continuously update target range, azimuth, and velocity. Target parameters are fed to a fire control computer that computes a future impact point for a weapon based on these parameters and the characteristics of the weapon being employed. The fire control computer is constantly updating this predicted impact point based on changes in target parameters. Deception jamming is designed to take advantage of any weaknesses in either target tracking or impact point calculation to maximize the miss distance of the weapon or to prevent automatic tracking. This chapter will discuss the most commonly employed deception jamming

techniques, including false target jamming, range deception jamming, angle deception jamming, velocity deception jamming, and monopulse jamming.

2. FALSE TARGET JAMMING

False target jamming is an effective jamming technique employed against acquisition, early warning, and ground control intercept (GCI) radars. The purpose of this type of jamming is to confuse the enemy radar operator by generating many false target returns on the victim radar scope. When false target deception jamming is successfully employed, the radar operator cannot distinguish between false targets and real targets (Figure 11-2).

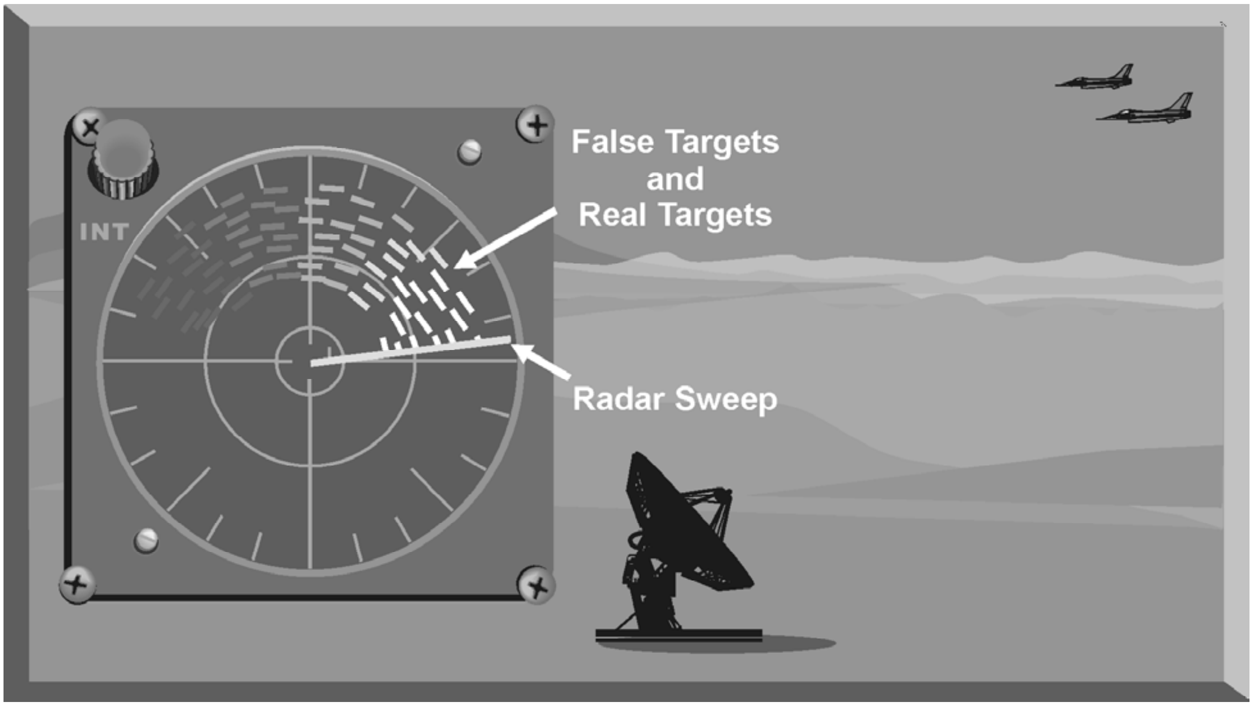

Figure 11-2. False Target Generation

a. To generate false targets, the deception jammer must tune to the frequency, PRF, and scan rate of the victim radar. The jamming pulse must appear on the radar scope exactly like a radar return from an aircraft. Multiple false targets greater in range than the jammer are generated by delaying the transmission of a jamming pulse until after the victim radar pulse has been received. False targets closer in range are generated by anticipating the arrival of a radar pulse and transmitting a jamming pulse before the victim radar pulse hits the aircraft. If the victim radar employs a jittered PRF, only targets greater in range can be generated.

b. To generate different azimuth false targets, the deception jammer synchronizes its transmitted pulse with the victim radar's sidelobes. Due to their

reduced power, when compared to the main beam, sidelobes are difficult to detect and analyze. The receiver in the deception jammer must be sensitive enough to detect these sidelobes and not be saturated by the power in the main radar beam. A false target deception jammer must inject a jamming pulse that looks like a target return into these sidelobes. To penetrate the radar sidelobes requires a lot of power. However, the power must be judiciously used. If a powerful jamming pulse is injected into the main beam, the false targets will be easy to detect. Most false target jammers vary the power in the jamming pulse inversely with the power in the received signal, on a pulse-by-pulse basis. This means the repeater jamming signal is at minimum power when the main beam of the victim radar is on the aircraft and at maximum power when the sidelobes are being jammed. To effectively generate false azimuth targets, the jammer must have a receiver with a wide dynamic range to detect both the main beam and the sidelobes. In addition, the jamming system must be able to generate high power that can be effectively controlled by the receiver.

c. To generate moving false targets, the deception jammer must synchronize with the main beam and the sidelobes in frequency, pulse width and PRF. Amplitude modulated jamming signals, with variable time delays, are transmitted into the sidelobes of the victim radar. The variable time delay provides a false target that changes range, either toward or away from the radar, depending on the time delay. The amplitude modulation provides false azimuth targets that appear to be moving.

d. The effectiveness of false target generation is based on the credibility of the generated false radar returns. If the victim radar can easily distinguish between false returns and target returns, the technique is a failure. The false returns must look identical to an aircraft return. The radar return on the victim radar scope should have the same intensity, depth, and width as a target return.

(1) Power determines the false target intensity when it is displayed on the victim radar scope. Varying jammer output power inversely with received power ensures that each false target has nearly the same intensity as a true target return. The depth, or thickness, of the false target depends on the pulse width of the victim radar. By matching the pulse width of the jamming pulse with the pulse width of the victim radar, the jammer can generate false targets with the same depth as a real target return.

(2) The width of the false target depends on the antenna pattern of the victim radar. This can pose a problem for false-target deception jammers. Because the jamming pulse is transmitted the entire time the radar beam is on the jammer, the width of a false target will tend to be greater than a real target return. Aircraft radar return varies with main beam cross-section. To correct this problem, most false target deception jammers use random modulation in the power of the transmitted pulses. This will vary the width of the false targets and make them look more like the variable returns of actual targets.

3. RANGE DECEPTION JAMMING

Although a specific TTR can track multiple targets and direct multiple weapons, the tracking circuit must select a single target return and track it while ignoring all other returns. Target selection is done by using gate bins. The range gate is used as the primary gate for target selection. A range gate is an electronic switch that is turned on for a period of microseconds based on a certain range or time delay after a pulse is transmitted (Figure 11-3).

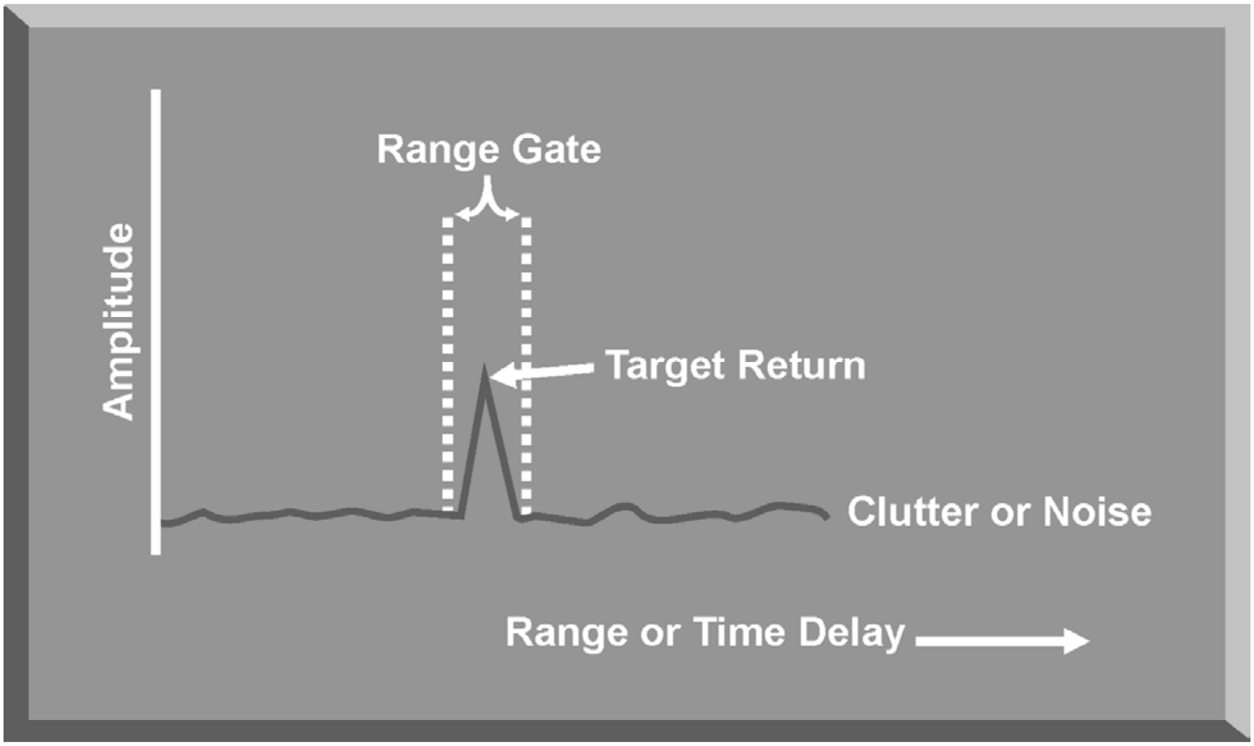

Figure 11-3. Range Gate Tracking

a. Range deception jamming exploits any inherent weakness in a TTR's automatic range gate tracking circuits. When a TTR's range gate locks on to an aircraft, the range deception jammer detects the radar signal. The range deception jammer then amplifies and retransmits a signal much stronger than the radar return. This retransmitted signal, called a cover pulse, is displayed in the range gate with the target signal (Figure 11-4).

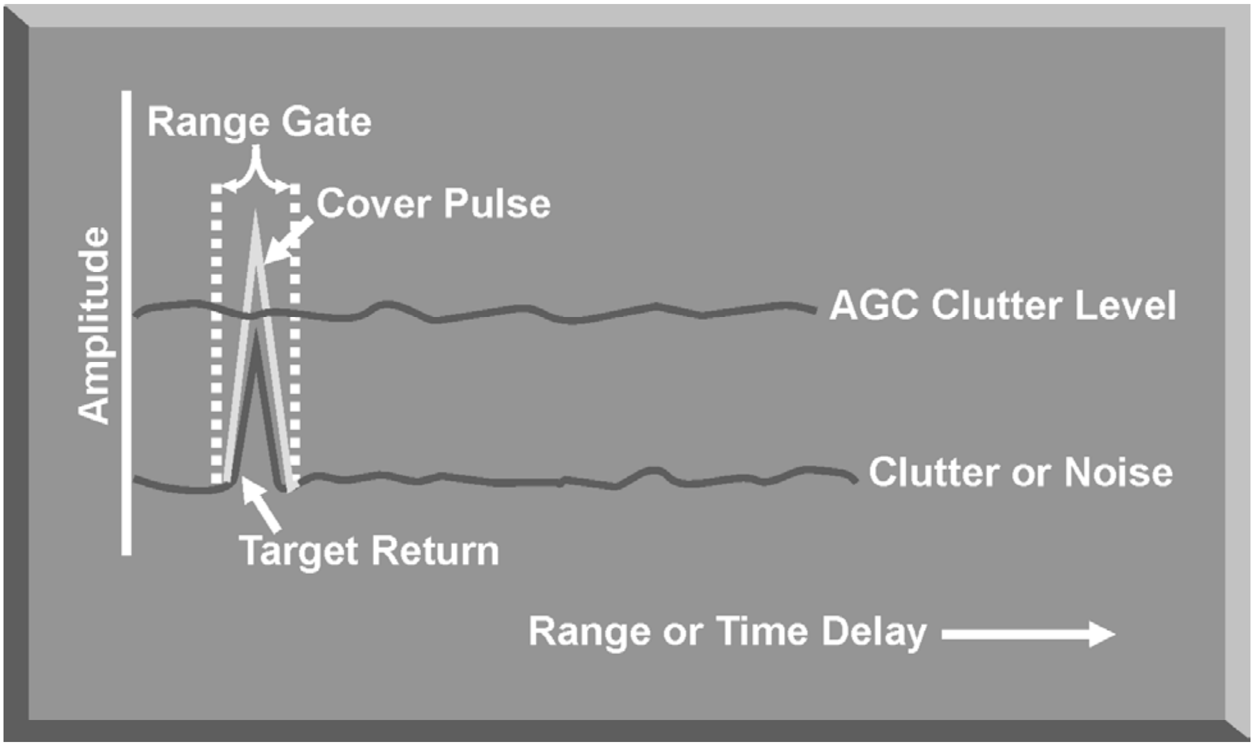

Figure 11-4. Range Gate Jamming Cover Pulse

b. The automatic gain control (AGC) circuit lowers the gain in the range tracking gate to control the amplitude of the cover pulse in the range gate. Reduced gain causes the real target return to be lost, and the range gate only tracks the jamming signal. This is known as range gate capture.

c. Once the range gate is captured by the cover pulse, a technique called range gate pull-off (RGPO) is employed (Figure 11-5). The deception jammer memorizes the radar signal and introduces a series of time delays before retransmitting. By increasing these time delays, the range gate will detect an increase in range and automatically move off to a false range. Once the range gate has moved well away from the real target, the range deception jammer shuts down, and the radar range gate is left with no target to track. The range gate breaks lock and the TTR must again go through the process of search, acquisition, and lock-on to re-engage the target.

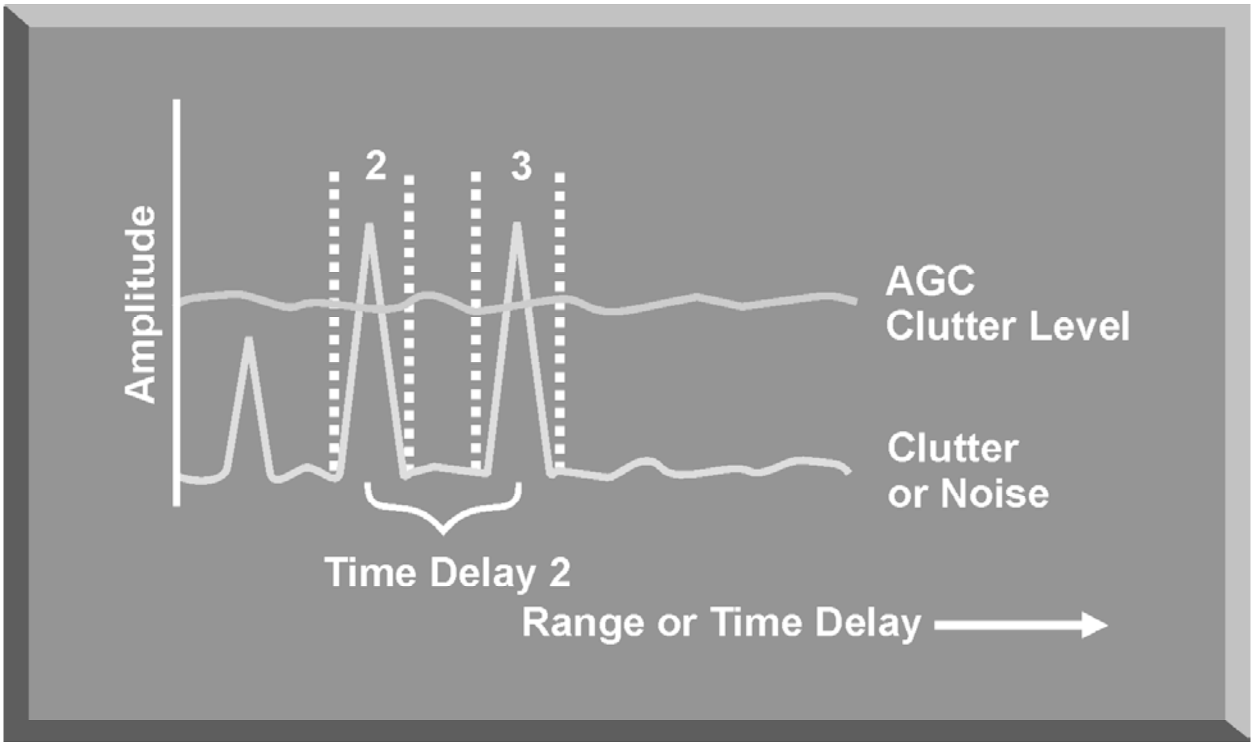

Figure 11-5. Range Gate Pull-Off

d. There are several advantages of range deception jamming, especially when used as a self-protection technique. It can generate sufficient errors to deny range information and is effective against most automatic range tracking systems. This technique does not require a large amount of power, just enough to cover the radar return of the aircraft. If the time delays are not exaggerated, an operator may not detect the loss of range lock-on until after a missile has been fired. The insidious nature of range deception jamming may generate enough miss distance to save the aircraft and pilot.

e. There are disadvantages to using range deception jamming. First, it can be defeated by a trained radar operator. If the operator detects a problem with the automatic range tracking circuit, the system can be switched to manual range tracking mode to defeat RGPO. Also, if the threat system is still able to track the aircraft's azimuth and elevation, range information may not be required to complete target engagement. To maximize range deception jamming effectiveness, it should be employed in conjunction with azimuth and elevation jamming. Finally, this type of range deception jamming is not effective against a leading-edge range tracking system. A leading-edge tracker will not see the delayed cover pulse. As the cover pulse moves off the target, AGC circuits reset the gain to continue tracking the real target. The only way to defeat a leading-edge range tracker is with a deceptive jammer that anticipates the next radar pulse and sends a jamming cover pulse before it reaches the aircraft. This jamming technique can also be defeated by randomly varying the radar PRF.

4. ANGLE DECEPTION JAMMING

Angle deception jamming is designed to exploit weaknesses in the angle tracking loop of the victim radar. The specific technique depends on the tracking method used to derive azimuth and elevation information. Inverse amplitude modulation jamming is the main angle deception technique used against TWS radars. For conical scan radars, scan rate modulation and inverse gain jamming are used. Swept square wave (SSW) jamming is used against LORO tracking radars. Monopulse angle deception jamming will be covered separately.

a. The azimuth and elevation tracking loop for a TWS radar is based on target signal amplitude modulation. The inverse amplitude modulation jammer generates a signal with modulation exactly opposite the expected return. To accomplish this, the angle deception jammer must receive the radar signals from the tracking beams. The jammer responds with a signal of the same frequency, PRF, and scan rate synchronized to the inverse of the radar antenna pattern (Figure 11-6). This induces an error in the angle tracking gate that, over a series of scans, causes the radar to lose target angle tracking.

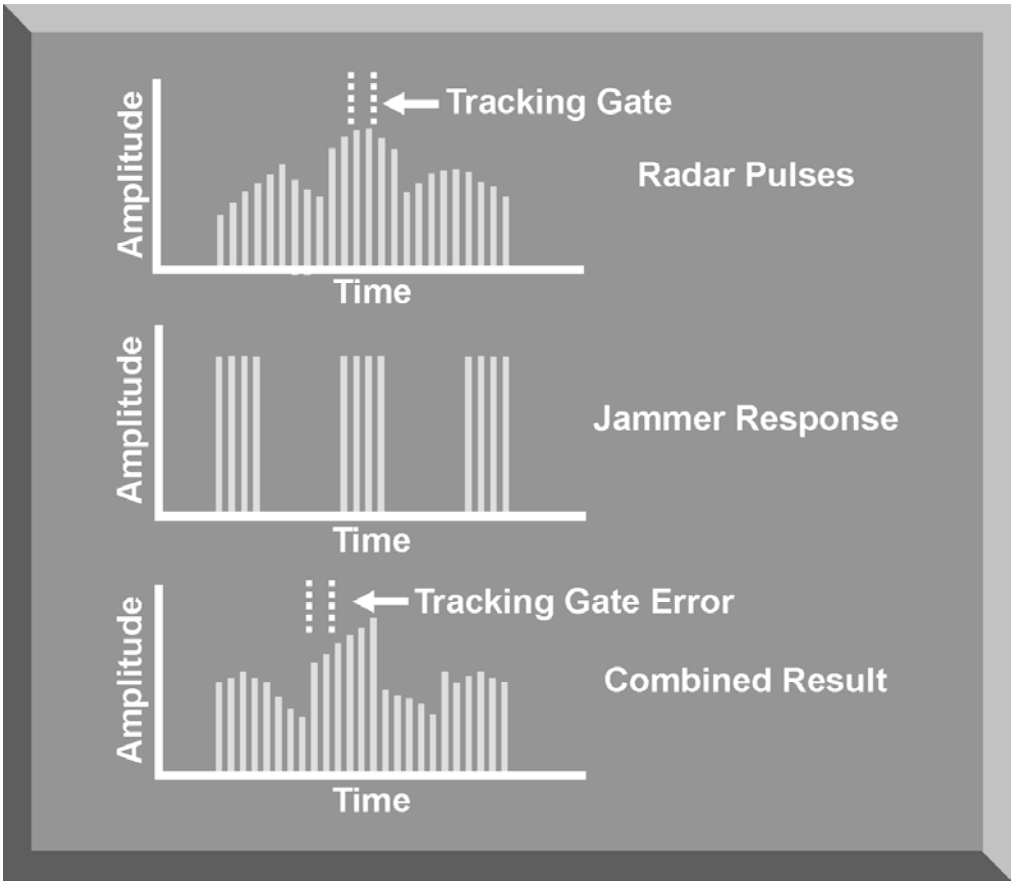

Figure 11-6. Inverse Amplitude Modulation Jamming

b. Inverse gain jamming is also effective against conical scan radars. Since conical scan radars use the phase of the target returns to generate error signals, an inverse gain deception jammer attempts to alter the phase by inducing fake signals into the antennas. In addition, by altering the amplitude of the signal, the jammer induces large errors into the tracking loop. To accomplish this, the jammer must determine the frequency, PRF, and scan rate of the victim radar. It then transmits signals that change the phase and amplitude of the target signal, resulting in a signal 180 degrees out of phase with the actual target (Figure 11-7). This 180-degree error rapidly drives the antenna off the target and causes break-lock.

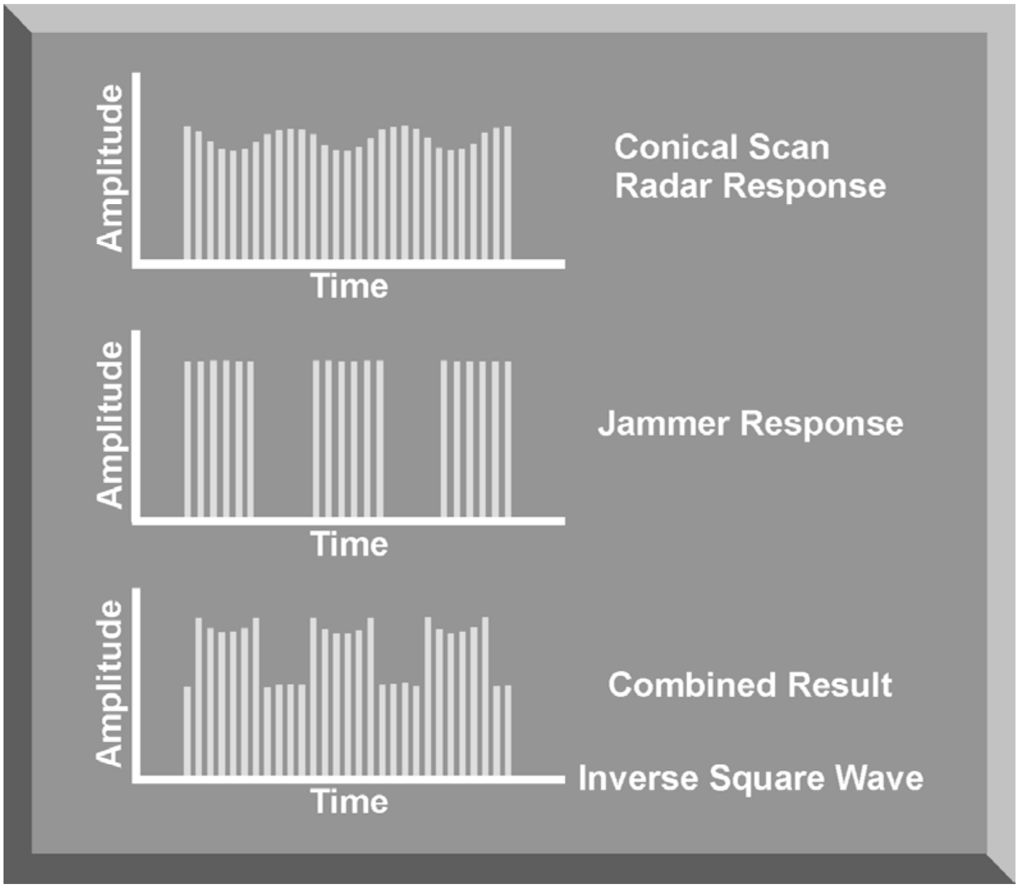

Figure 11-7. Inverse Gain Jamming

c. Scan rate modulation is also used against conical scan radars. This angle deception technique modulates the jamming pulse at or near the victim radar nutation frequency. As the modulation approaches the radar's nutation frequency, large error signals appear in the radar servo tracking loops, producing random gyrations in the antenna system, causing break-lock. This technique is most effective if the modulation jamming is slowly swept in frequency until it matches the nutation rate.

d. Both inverse scan and scan rate modulation jamming require very little power and have proven extremely effective against TWS and conical scan radars. To be effective, however, the angle deception jammer must find the precise scan rate of the victim radar. The jammer must concentrate on one signal at a time, limiting the number of threat systems that can be jammed simultaneously. In a dense threat environment, this can be a severe limitation.

e. The effectiveness of inverse gain and scan rate modulation jamming led radar designers to employ antennas that scan only during the receiving function of the radar system. Generally, this is accomplished by using two antennas. The transmitting antenna illuminates the target. Receiving antennas scan to produce the amplitude modulation of the reflected signal for effective angle tracking. This technique is called Lobe-On-Receive-Only (LORO). Since the transmitting antenna does not nutate, or scan, angle deception jammers cannot detect the modulation required to generate effective inverse gain modulation. Swept square wave (SSW) jamming is the angle deception technique developed to counter LORO angle tracking.

f. SSW jamming continuously varies the frequency of amplitude modulation on the jamming pulse over an expected range of nutation or scanning frequencies. This range is established by either electronic intelligence (ELINT) data on a particular system, or by exploitation. The dotted line in Figure 11-8 shows a threat's nutation or scan frequency. As the frequency of the modulated jamming pulse approaches the threat scan frequency, it induces errors in the angle tracking loop of the victim radar. The longer the SSW jamming stays near the scan frequency, the greater the induced errors. It is important that the sweep rate of the modulating jamming be slow enough to maximize its impact on the victim radar.

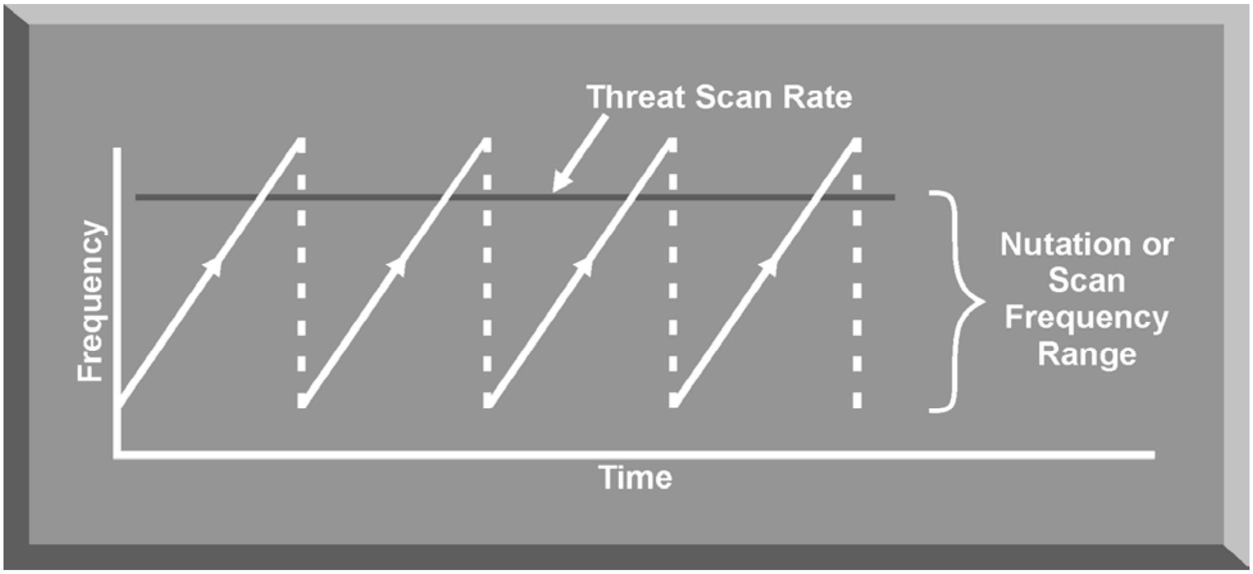

Figure 11-8. Swept Square Wave Jamming

5. VELOCITY DECEPTION JAMMING

Pulse Doppler and continuous wave (CW) radars track targets based on velocity or Doppler-shifted frequency (Figure 11-9). The objective of velocity deception jamming is to deny velocity tracking information and generate false velocity targets. The primary techniques include velocity gate pull-off (VGPO), Doppler noise, narrowband Doppler noise, and Doppler false targets.

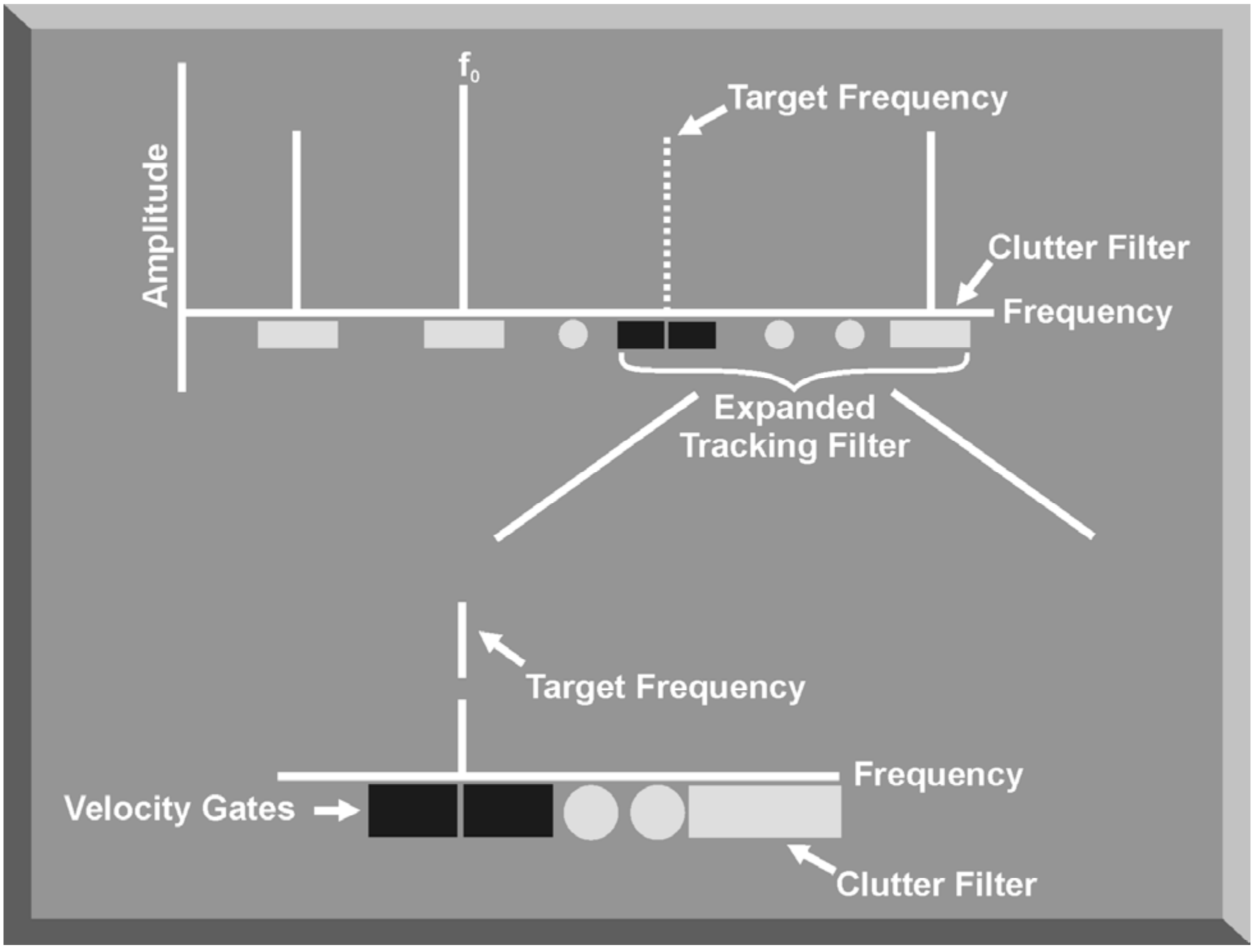
f_0
Target Frequency
Amplitude
Clutter Filter
Frequency
Expanded
Tracking Filter
Target Frequency
Frequency
Velocity Gates
Clutter Filter

Figure 11-9. Velocity Tracking Gate

a. Velocity gate pull-off counters pulse Doppler or CW radars by stealing the velocity gate of their automatic tracking loop. The objective of VGPO is to capture the Doppler velocity tracking gate by transmitting an intense false Doppler signal. Then the frequency of the false signal is changed to move the tracking gate away from the true target Doppler. This is analogous to the RGPO technique used against the range gate tracking loop.

(1) To accomplish an effective VGPO technique, the jammer receives the CW or pulse Doppler signal. It then retransmits a CW or pulse Doppler signal that is higher in power than the return from the aircraft, but at approximately the same

Doppler frequency (Figure 11-10). It is important that the frequency of this initial jamming pulse appears within the same velocity tracking filters as the target return or the victim radar will disregard it. The frequency band of the Doppler tracking filters is an important piece of intelligence information. The velocity tracking gates are quite narrow, roughly 50 to 250 MHz. Once the jamming pulse appears in the tracking gate, the automatic gain control circuit gains out the target return, and the jamming pulse has captured the velocity gate.

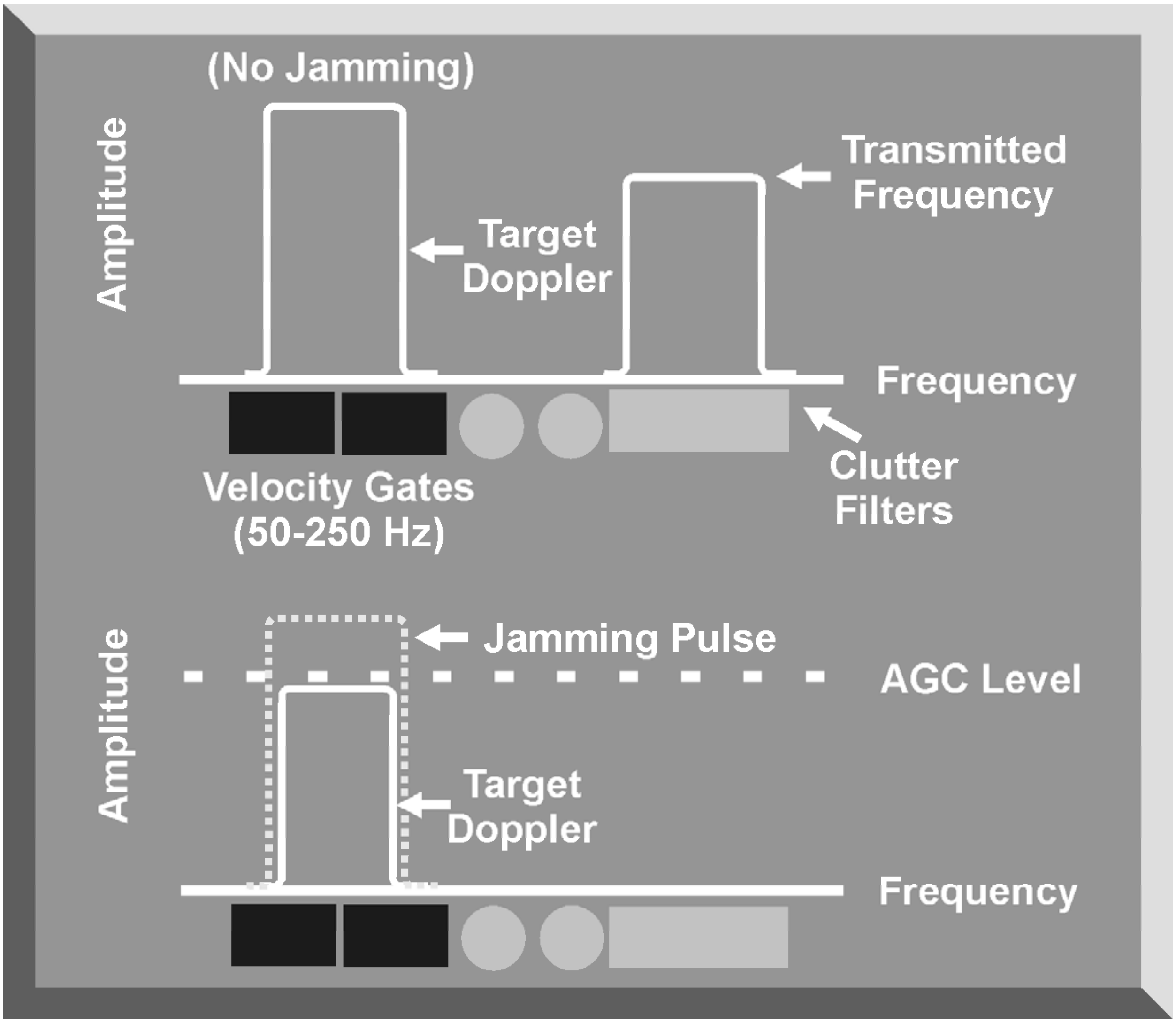
(No Jamming)
Amplitude
Transmitted Frequency
Target Doppler
Frequency
Clutter Filters
Velocity Gates (50-250 Hz)
Amplitude
Jamming Pulse
AGC Level
Target Doppler
Frequency

Figure 11-10. Velocity Gate Capture

(2) Once the jamming pulse has captured the tracking gate, the deception jammer slowly changes the Doppler frequency (Figure 11-11). This frequency shift is accomplished by several methods. The most common method uses frequency modulation (FM) within the jammer's traveling wave tube (TWT). By varying the TWT voltage, the Doppler frequency of the jamming pulse is changed

linearly, and the radar tracking gates follow the jamming pulse. By using FM, the jamming pulse can be moved in either a positive or negative direction, depending on the slope of the voltage. By slowly changing the frequency of the modulation, the jamming pulse pulls the tracking gates off the target. When the maximum offset has been achieved, nominally 5 to 50 kHz, the FM is "snapped back" to a minimum value, and the process is repeated to preclude target reacquisition.

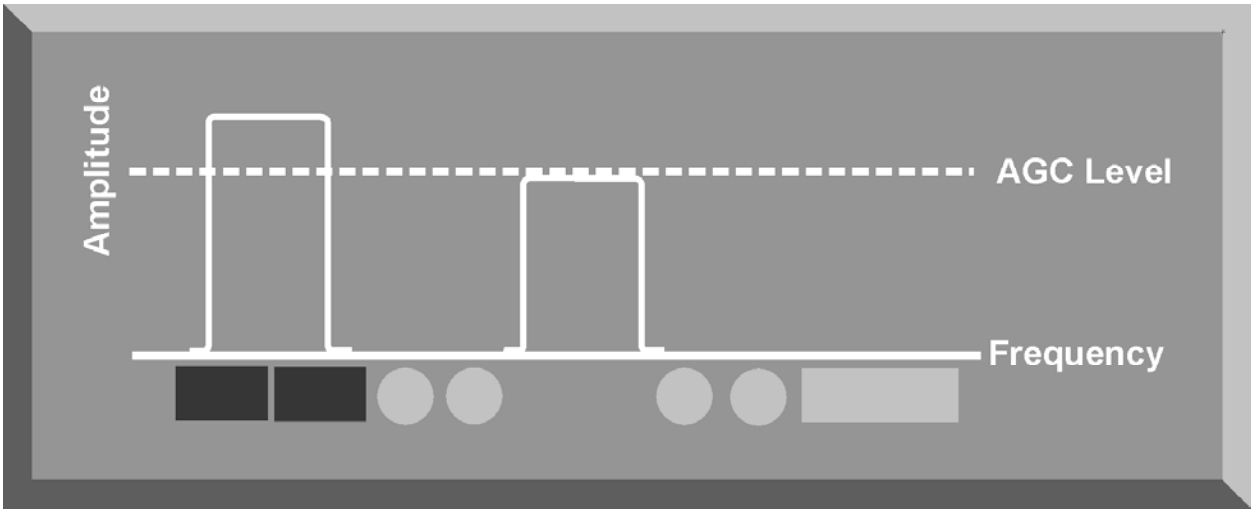

Figure 11-11. Velocity Gate Pull-Off

(3) The rate of change of frequency offset in a VGPO pulse is an extremely critical parameter. Many CW and pulse Doppler radars employ acceleration stops as part of the tracking gates. By differentiating the velocity outputs of the velocity tracking gates with respect to time, the velocity tracker computes target acceleration. Acceleration stops detect and reject unusually large changes in target acceleration. If the VGPO technique changes the frequency of the jamming pulse too rapidly, the tracking loop, with acceleration stops, will reject the jamming pulse and stay on the target. This means that an effective VGPO technique may take from one to ten seconds.

b. Doppler noise differs from most noise techniques in that it is a repeater technique. The jamming system must receive the pulse Doppler radar signal in order to generate an appropriate jamming pulse. Also, noise jamming output is done on a pulse-by-pulse basis and only lasts as long as the pulse duration, or pulse width, of the victim radar signal (Figure 11-12). The Doppler noise jammer receives each pulse and applies a random frequency shift, either positive or negative, to each pulse.

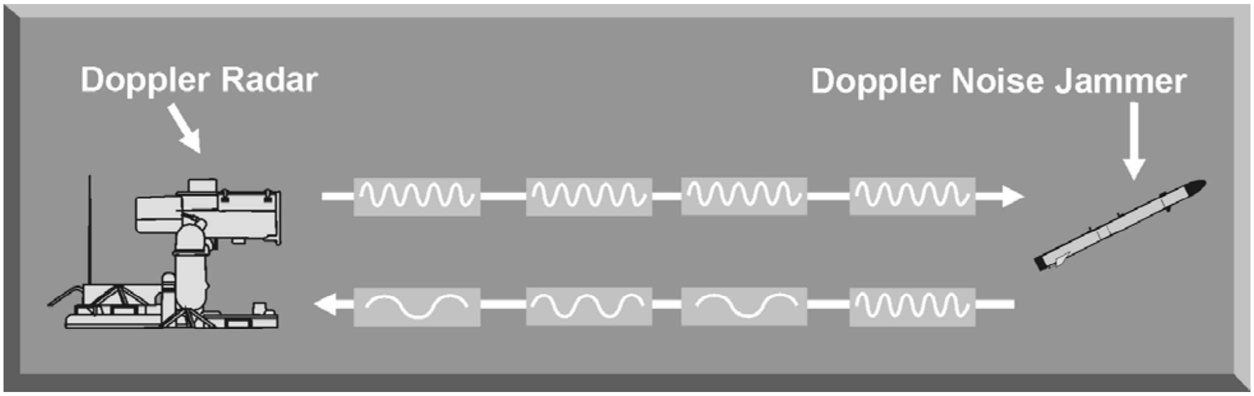
Doppler Radar
Doppler Noise Jammer

Figure 11-12. Doppler Noise Jamming

(1) When Doppler noise jamming pulses are processed by the signal processor, and the Doppler frequencies are sent to the velocity tracking gate, there are so many different velocities that the tracking gate cannot distinguish the target from the jamming. The random distribution of target velocities effectively masks the true target Doppler velocity. If the velocity tracking loop is not saturated, multiple false targets traveling at different speeds will be displayed.

(2) When a technique called Doppler noise blinking is employed, it interferes with the angle and velocity tracking within most semi-active radar missiles. Doppler noise blinking is accomplished by rapidly transmitting bursts of Doppler noise jamming (Figure 11-13).

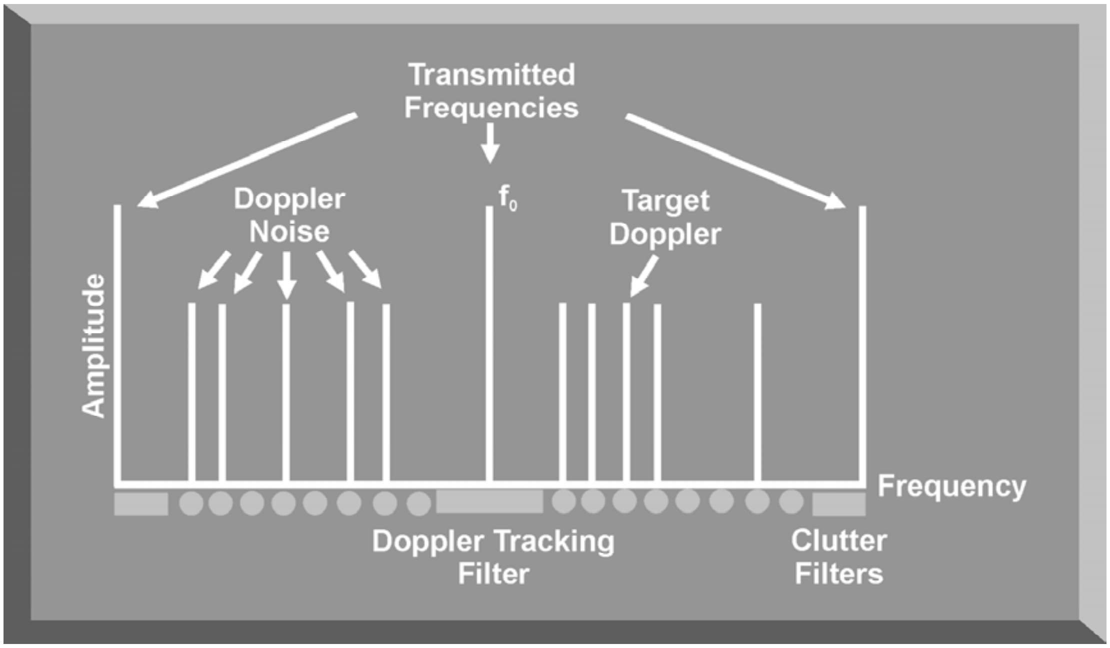
Transmitted Frequencies
Doppler Noise
f0
Target Doppler
Amplitude
Frequency
Doppler Tracking Filter
Clutter Filters

Figure 11-13. Impact of Doppler Noise Jamming

(3) Doppler noise jamming is effective against most pulse Doppler radars and the semi-active missiles employed with these radars. One disadvantage, however, is that it is only effective against the velocity tracking loop. If range tracking is still available to the radar, Doppler noise may highlight the jamming aircraft. Another disadvantage is that Doppler noise requires a sophisticated jammer able to receive the victim radar pulse, generate random positive and negative frequency modulations on this pulse, and retransmit the jamming pulses at the PRF and pulse width of the victim radar. This requires an extremely fast signal processing capability and detailed intelligence information on the victim radar.

c. Narrowband Doppler noise is also a repeater technique. The jamming system receives the pulse Doppler radar signal and generates a noise jamming signal on a pulse-by-pulse basis (Figure 11-14). Narrowband Doppler noise requires detailed information on the frequency coverage of an individual velocity tracking filter, or velocity bin, employed by the victim radar. Once this frequency range is known, the jammer receives each pulse from the victim radar and transmits jamming pulses with a higher and lower frequency shift based on the real target Doppler. These frequency shifts are always within the frequency range of the velocity bin.

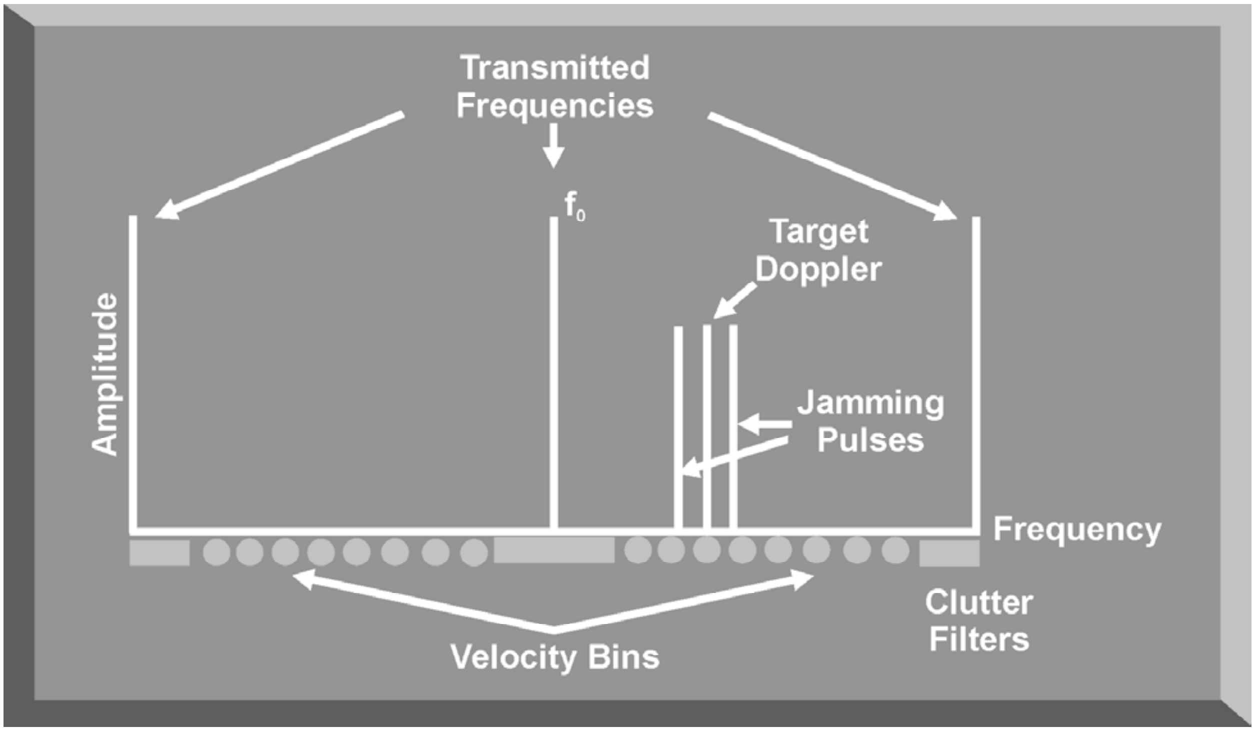

Figure 11-14. Narrowband Doppler Noise

(1) When these pulses are processed by the signal processor and the Doppler signals are sent to the velocity tracking gates, the particular bin that

contains the target Doppler also contains several other targets generated by the jammer. The victim radar signal processor attempts to distinguish the target Doppler from the jamming pulses. It raises the gain in the velocity tracking bins, thinking that the signal with the highest amplitude is the target. But, as the signal gain is increased, the target is "gained out" with the jamming signals and no target is displayed. This is called velocity bin masking and can completely deny target information to a pulse Doppler radar (Figure 11-15).

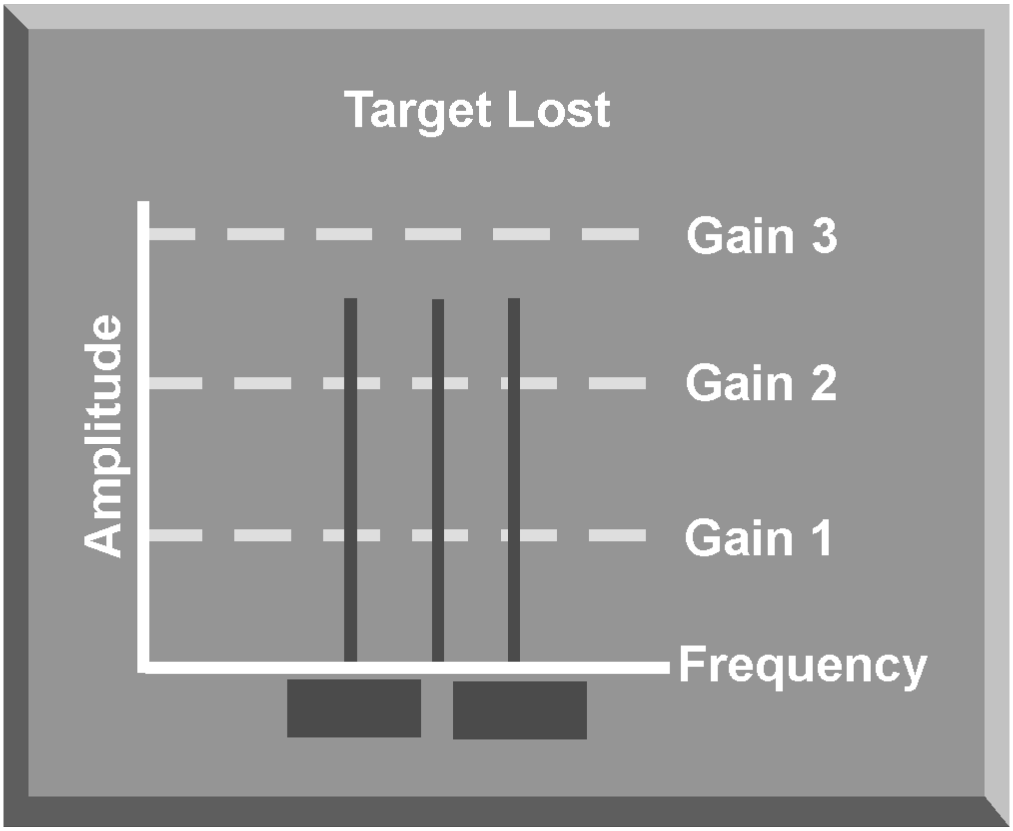

Figure 11-15. Velocity Bin Masking

(2) The advantage of narrowband Doppler noise is that it completely masks an aircraft's velocity from a pulse Doppler radar. The disadvantages include the following: When the victim radar can range-track an aircraft, narrowband Doppler noise highlights the aircraft's presence. To be effective, narrowband Doppler noise requires knowledge of the frequency range of the victim radar's velocity tracking bins, or filters. This detailed information may be available only through threat system exploitation. Finally, sophisticated signal processing and jamming systems are required to receive and transmit in the very narrow frequency band of the velocity bin.

d. Doppler false target jamming is normally used with narrowband Doppler noise or other deception techniques. Its purpose is to initially confuse the radar signal processor with multiple targets and then force the radar signal processor

to raise its gain levels in the velocity tracking loop. The Doppler false target jammer receives each pulse of the victim radar and applies a random frequency shift to a selected number of these pulses (Figure 11-16).

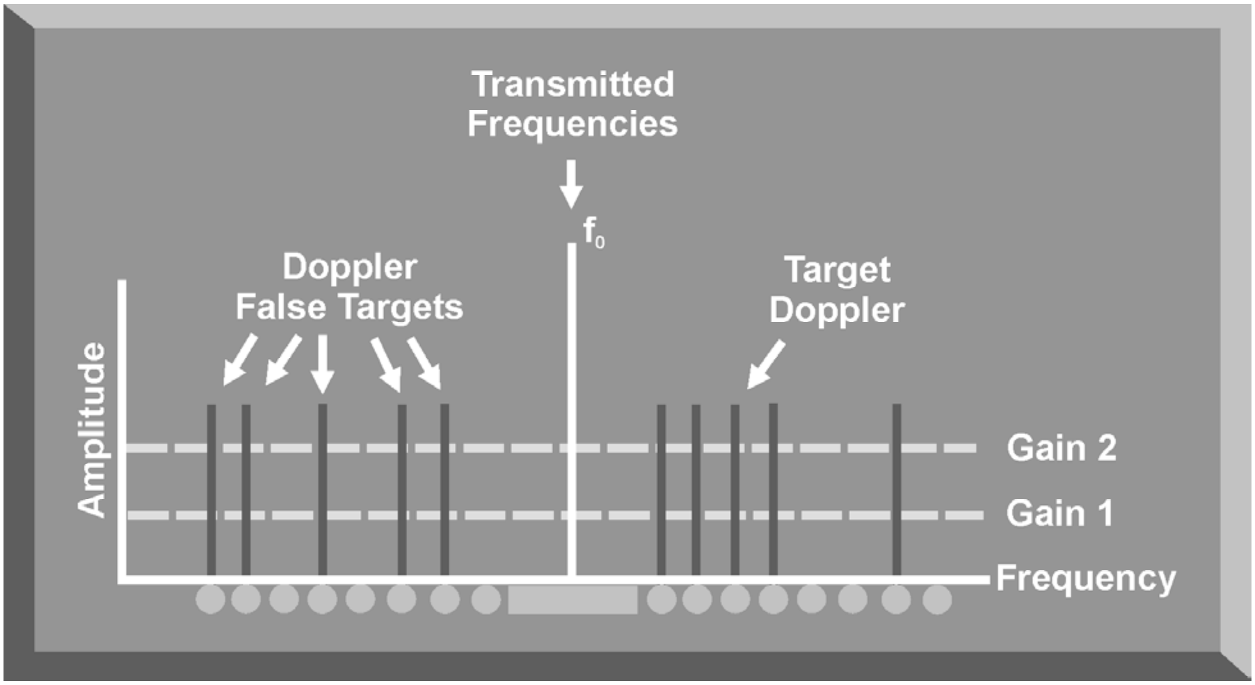

Figure 11-16. Impact of Doppler False Target Jamming

(1) The selected pulses are processed by the signal processor, and multiple Doppler frequencies are sent to the velocity tracking gate. In an attempt to distinguish the target from the jamming pulses, the signal processor increases the gain in each tracking filter, assuming the target Doppler has a higher amplitude than the jamming pulses. This increase in gain sets up the velocity tracking loop for a narrowband Doppler noise technique that will cause the real target to be lost among the generated false targets.

(2) The advantage of Doppler false target jamming is that it can initially confuse the radar signal processor and the radar operator as to the velocity of the real target. It also sets up the radar for narrowband Doppler noise technique and increases its effectiveness. The disadvantage is that the signal processor or the radar operator will eventually be able to distinguish the real target from the false targets based on its velocity. This jamming technique is much more effective when used in conjunction with other Doppler jamming techniques.

6. MONOPULSE DECEPTION JAMMING

The ability of monopulse tracking radars to obtain azimuth, range, and elevation information on a pulse-by-pulse basis make them extremely difficult to jam

(Figure 11-17). Amplitude modulation jamming used against conical scan or TWS radars, such as inverse scan and swept square wave, highlights a target, making monopulse tracking easier. Frequency modulation techniques, such as RGPO and VGPO, are equally ineffective. They serve as a beacon that aids the monopulse radar's target tracking ability. The monopulse radar may be able to track the jammer with more accuracy than tracking actual radar returns because target glint effects are absent from the jamming pulse. Monopulse angle jamming techniques can be divided into two main categories, system-specific and universal. Examples of system-specific jamming techniques include skirt frequency jamming, image jamming, and cross-polarization jamming. These techniques attempt to exploit weaknesses in the design and operation of specific monopulse radars. Cross-eye jamming, a universal technique, attempts to exploit all monopulse radar systems.

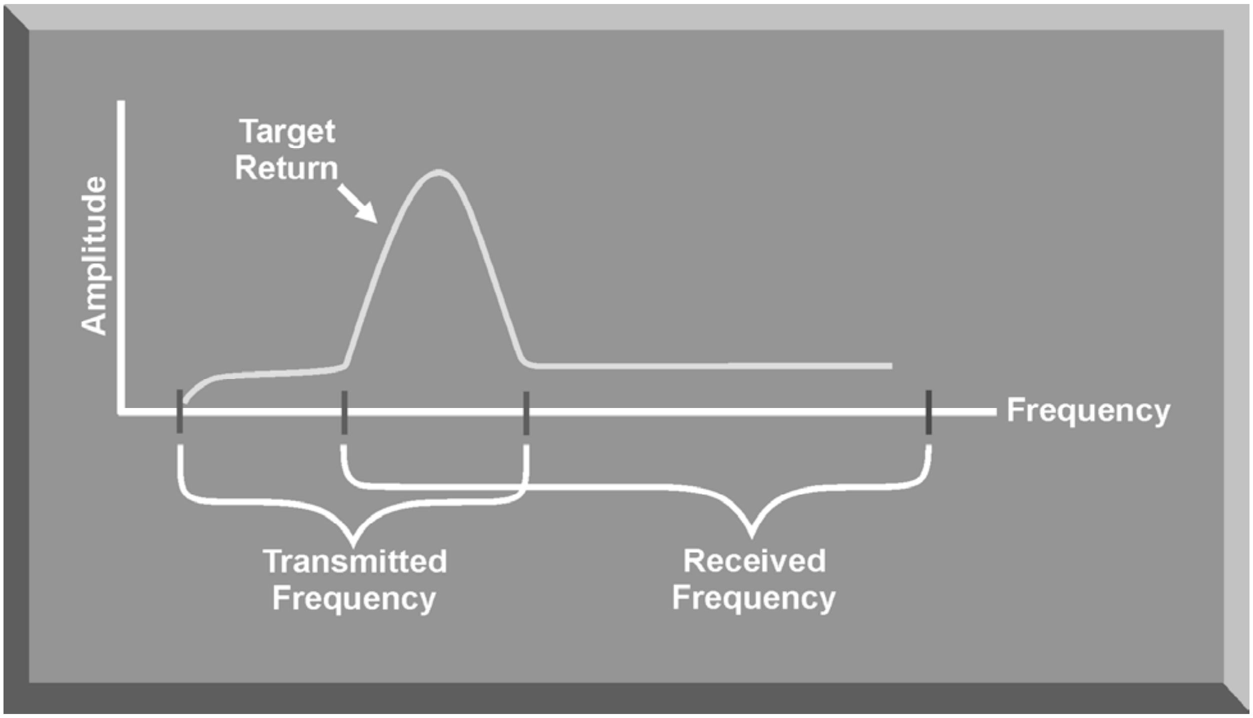

Figure 11-17. Monopulse Radar Receiver

a. Skirt frequency jamming, or filter skirt jamming, is designed to counter the monopulse receiver. Skirt frequency jamming is based on the fact that the intermediate frequency (IF) filter of the monopulse receiver must be correctly tuned to the transmitting frequency of the monopulse radar. It these two components are not exactly tuned, the target signal may be presented on the edge, or skirt, of the receiver IF filter. This offers an opportunity to inject a jamming signal into this skirt (Figure 11-18).

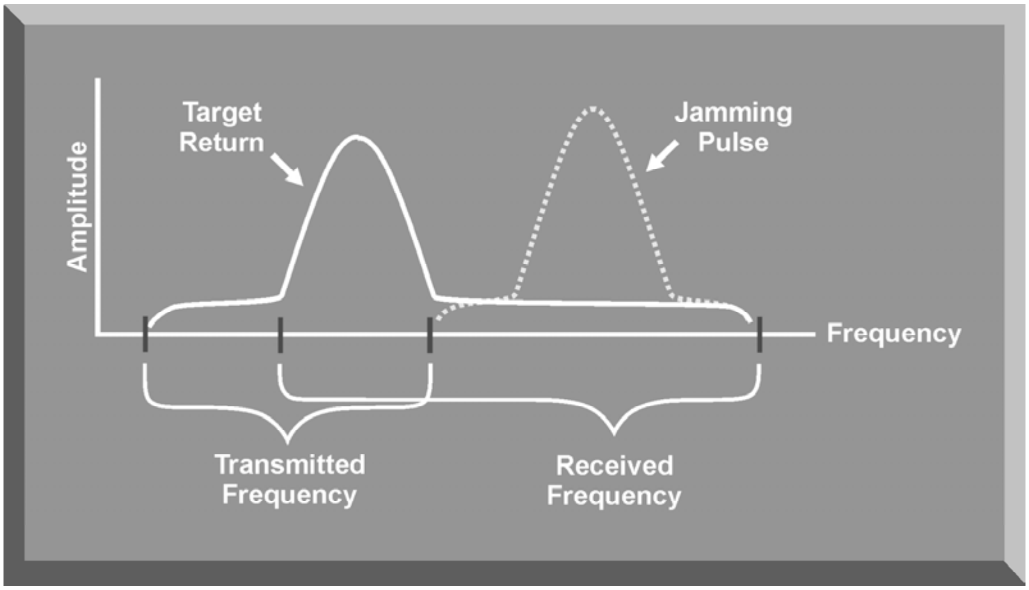
Target
Return
Jamming
Pulse
Amplitude
Frequency
Transmitted
Frequency
Received
Frequency

Figure 11-18. Filter Skirt Jamming Pulse

(1) Filter skirt jamming attempts to take advantage of this frequency imbalance by transmitting a jamming pulse tuned slightly off the radar transmitted frequency and in the middle of the receiver IF filter. This jamming pulse will generate a false error signal and drive the antenna away from the true target return.

(2) A well designed and maintained monopulse system does not have a frequency imbalance. The transmitter and IF filter frequencies will be identical. Jamming signals that are even slightly out of this narrow frequency range will not affect the monopulse tracking capability of the radar (Figure 11-19).

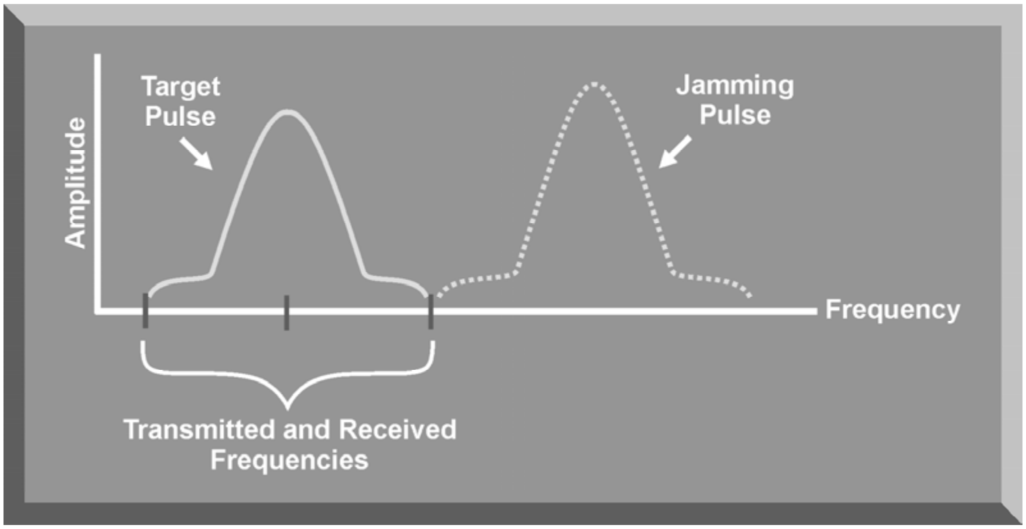
Target
Pulse
Jamming
Pulse
Amplitude
Frequency
Transmitted and Received
Frequencies

Figure 11-19. Ineffective Filter Skirt Jamming

(3) Effective filter skirt jamming requires extensive knowledge of the internal operation of the IF filter. This information can normally be obtained only by system exploitation. Variances from radar to radar and frequency imbalance exists from one radar IF filter to another. This creates a high degree of uncertainty in the effectiveness of this technique.

b. Image jamming exploits another potential weakness in the monopulse receiver (Figure 11-20). Some monopulse receivers have a wide-open front end with no preselection before the mixer. If the jammer transmits a pulse at the intermediate, or image, frequency, but out of phase with this frequency, the phase of the target tracking signal will be reversed and the antenna will be driven away from the target (Figure 11-21). Effective image jamming requires detailed information on the operation of the monopulse receiver. Of particular importance is the image, or intermediate, frequency and whether the local oscillation frequency is above or below the transmitted frequency. This may require exploitation of the monopulse threat system. In addition, a well-designed monopulse system has preselection in the front end and will reject signals that are out of phase with the transmitted frequencies. This capability renders image jamming ineffective.

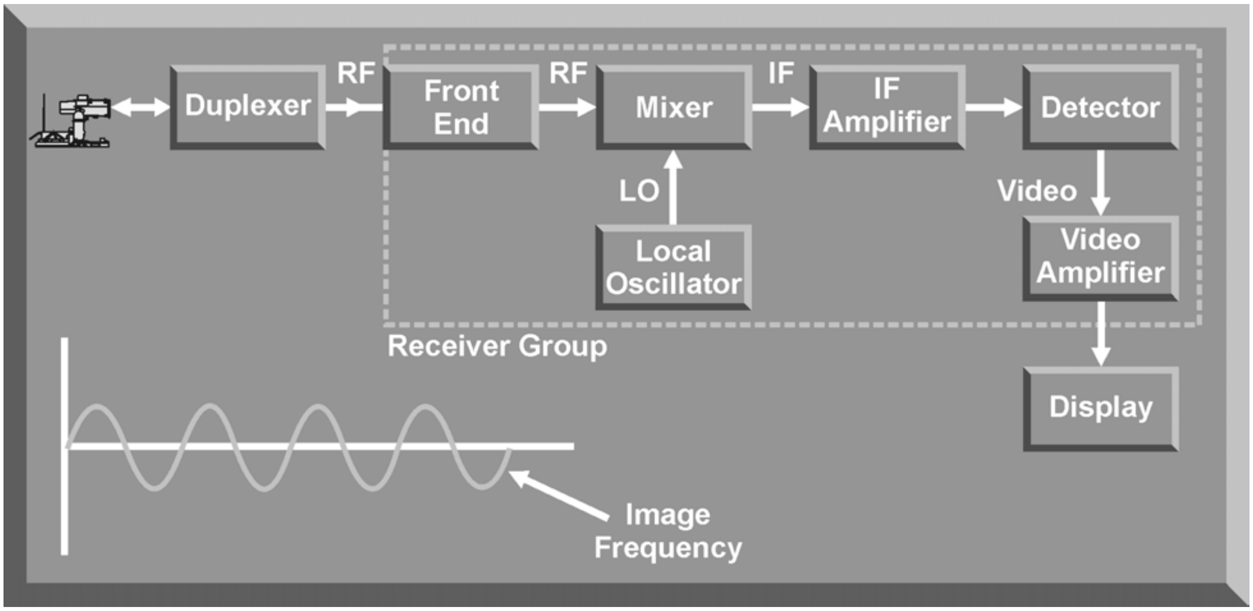

Figure 11-20. Monopulse Image Frequency

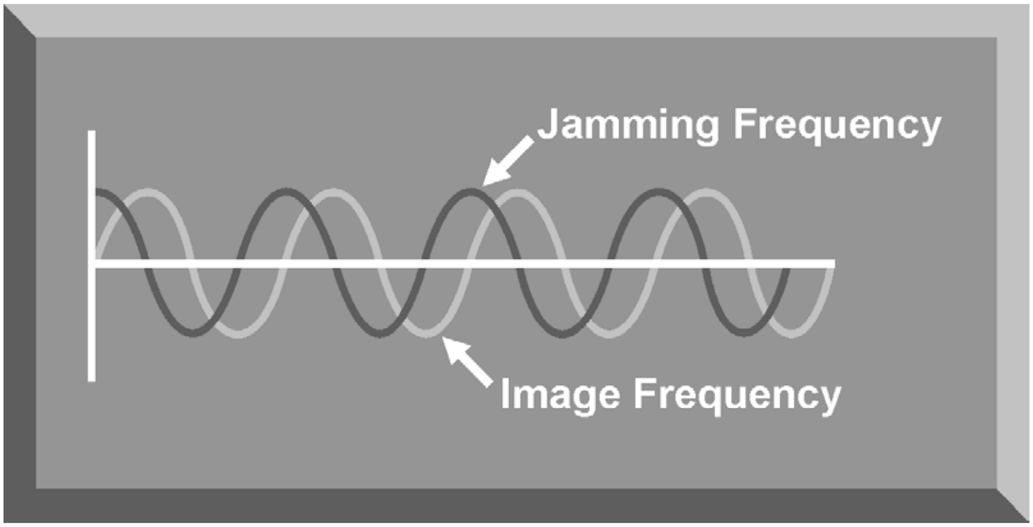
Jamming Frequency
Image Frequency

Figure 11-21. Monopulse Image Jamming

c. Cross-polarization jamming exploits the difference in the monopulse antenna pattern for a jamming pulse that is polarized orthogonal to the design polarization. The antenna pattern for a two-channel monopulse radar using sigma and delta beams shows the tracking point to be between the two beams (Figure 11-22). This is true if the radar is using its design polarization. However, the radar antenna also has a receiving pattern for a signal that is cross-polarized with the design frequency. For a cross-polarized signal, the tracking point is shifted one beamwidth to the right. This shift in the tracking point results in a target tracking signal that is 180° out of phase with the real signal. To be effective, a jamming signal polarized orthogonally to the design frequency of the radar would have to be 25 to 30 decibels, or about 1000 times, stronger than the radar signal.

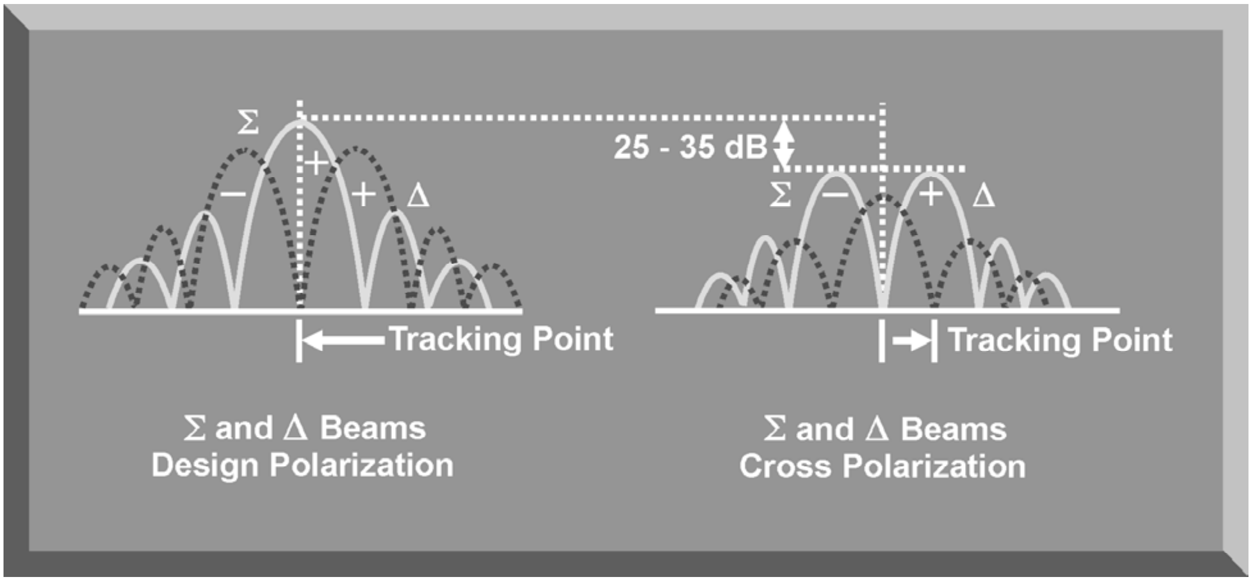
Σ
+
−
+
Δ
25 - 35 dB
Σ
−
+
Δ
Tracking Point
Tracking Point
Σ and Δ Beams
Design Polarization
Σ and Δ Beams
Cross Polarization

Figure 11-22. Cross-Polarization Antenna Pattern

(1) A cross-polarized jammer must receive and measure the polarization of the victim monopulse radar. The jammer then transmits a very high power jamming signal at the same frequency, but orthogonally polarized, to the victim radar. As a rule, the jamming signal must be 25 to 30 dBs stronger than the target return to exploit the tracking errors in the cross-polarized antenna pattern. Additionally, it must be as purely orthogonal to the design polarization as possible. Any jamming signal component that is not purely orthogonal will highlight the target and require more jamming power to cover the target return.

(2) A cross-polarized jammer must be able to generate a powerful jamming pulse that is polarized orthogonal to the victim radar. A cross-polarized jammer that generates the power and purity of polarization required to defeat monopulse angle tracking poses extreme technological challenges.

d. Cross-eye jamming is a complex technique that attempts to distort the wavefront of the beams in a monopulse radar and induce angle tracking errors. It exploits two basic assumptions of monopulse tracking logic in comparing target returns on a pulse-by-pulse basis. The first assumption is that a target return will always be a normal radar pulse echo. The second assumption is that any shift in amplitude or phase in a target return is due to the tracking antenna not pointing directly at a target. This condition generates an error signal and the antenna tries to null, but the amplitude or phase shifts.

(1) Cross-eye jamming attacks the two assumptions through a process of receiving and transmitting jamming pulses from different antennas separated as far apart as possible. In Figure 11-23, the phase front of a monopulse signal is received by the number 1 receive antenna, amplified by the repeater, and transmitted by the number 2 transmit antenna. The same phase front then hits receive antenna number 2, is shifted 180°, amplified by the repeater, and transmitted by the number 1 transmit antenna. These two out-of-phase signals must be matched in amplitude and must exceed the amplitude of the target return.

(2) When these jamming signals arrive at the victim radar, the tracking loop attempts to null out the amplitude and phase differences. With two widely spaced jamming sources at different phases, the antenna never achieves a null position or tracking solution. The distance between antenna pairs is an important parameter that determines the effectiveness of cross-eye jamming. The wider the spacing between antenna pairs, the more distortion in the victim's wave front near the true radar return. Most fighter aircraft do not provide sufficient spacing between the antennas to maximize effectiveness. Effectiveness is also lost when the aircraft is abeam or going away from the radar. To further complicate matters, when the radar is directly in front of the aircraft, the jamming pulses must have a power at least 20 dBs above the target return. Cross-eye jamming can also be defeated with a leading-edge tracker that rejects jamming signals arriving at the antenna behind the target return.

(3) Countering monopulse angle tracking is the greatest challenge for self-protection jamming systems. Skirt jamming and image jamming have had limited success. Cross-polarization and cross-eye jamming techniques require complex and sophisticated circuitry and much power.

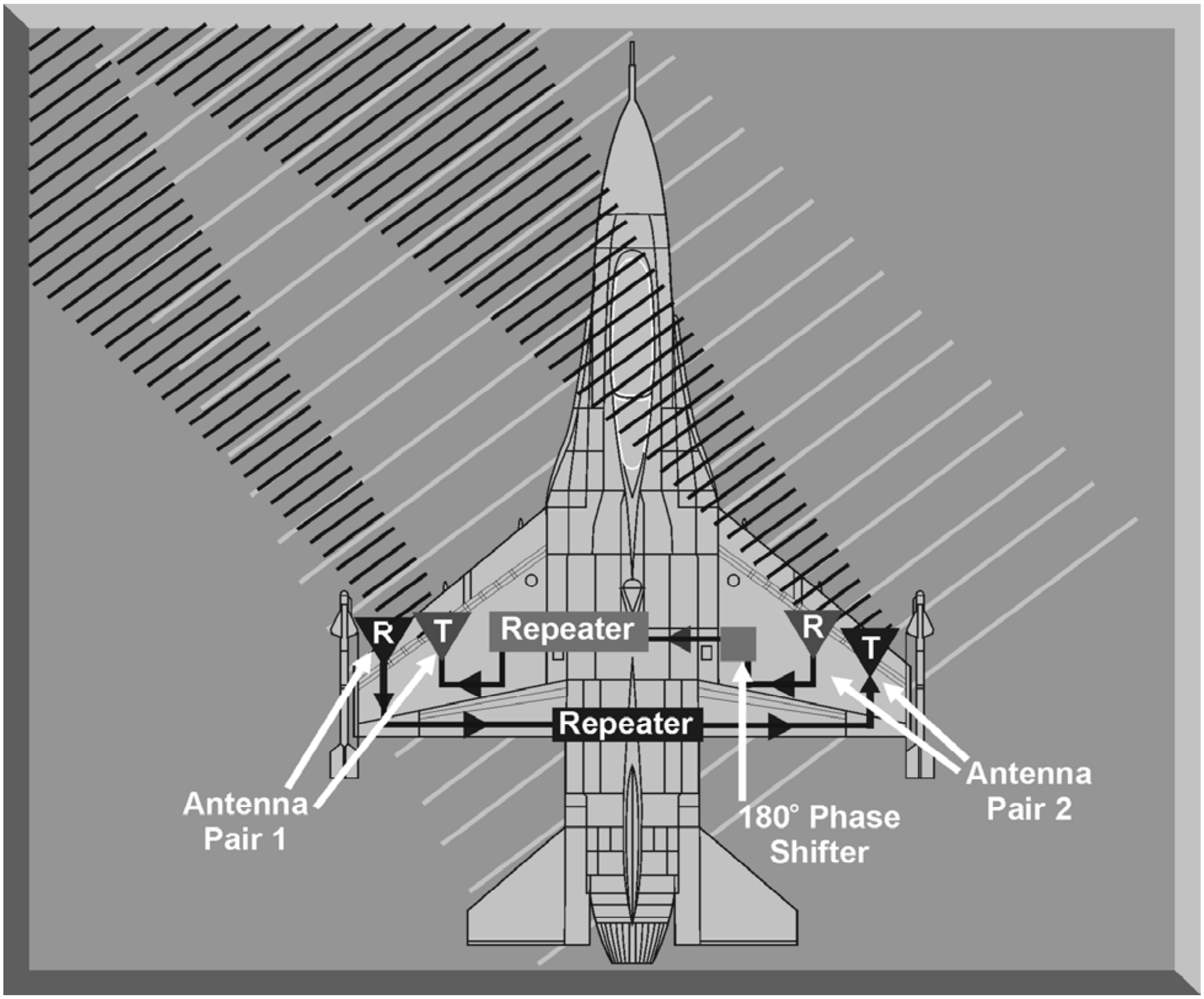

Figure 11-23. Cross-Eye Jamming

7. TERRAIN BOUNCE

Terrain bounce is a jamming technique used primarily at low altitude. It is used to counter semi-active, air-to-air missiles and monopulse tracking radars. The technique involves a repeater jammer that receives the radar or missile guidance signal. The jammer amplifies and directs this signal to illuminate the terrain directly in front of the aircraft. The missile or radar tracks the reflected energy from the spot on the ground instead of the aircraft (Figure 11-24).

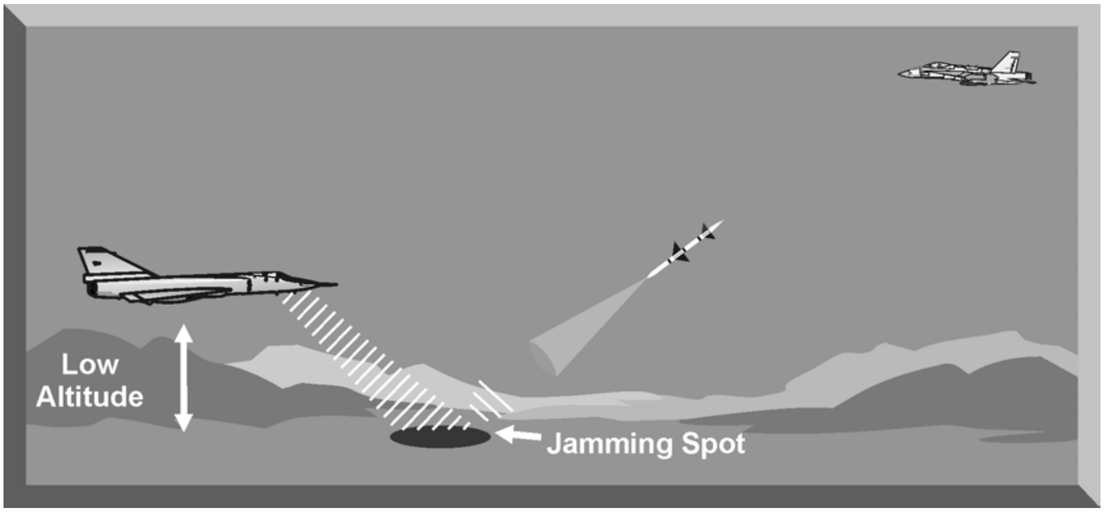

Figure 11-24. Terrain Bounce

a. To be effective, the terrain bounce jamming antennas should have a narrow elevation beamwidth and a broad azimuth beamwidth. This transmission pattern maximizes the energy directed toward the ground and minimizes the energy transmitted toward the missile or radar. To overcome signal losses associated with uncertain terrain propagation, the jamming system should also generate high jamming power. This ensures the energy reflected from the terrain is higher than the energy in the aircraft return. The terrain bounce jamming antennas should have very low sidelobes to preclude activation of any home-on-jam (HOJ) missile capability. For an air-to-air missile, the terrain bounce technique should be activated at long range. This will initially put the aircraft and the jamming spot in the same resolution cell. As the range decreases, the missile will be decoyed by the higher power in the jamming spot.

b. Some problems associated with terrain bounce jamming include the uncertainty of the signal scattering parameters of the various terrain features and the possible changes in signal polarization caused by terrain propagation. In addition, terrain bounce jamming can place maneuvering restrictions and maximum altitude limitations on the aircraft.

8. SUMMARY

There are several deception jamming techniques that can be employed to counter threat radar systems. The effectiveness of these techniques can be enhanced when they are employed in combination. For example, the effectiveness of an RGPO technique is enhanced when an angle deception technique is also employed. Determining the most effective deception technique, or combination of techniques, can present a challenge to intelligence and engineering analysts. However, when employed with maneuvers and chaff, deception techniques can mean the difference between success and failure on the modern battlefield.

CHAPTER 12. DECOYS

1. INTRODUCTION

A decoy is a device designed to look to an enemy radar more like an aircraft than the actual aircraft itself. Decoys do three primary missions: they saturate the enemy's integrated air defense system (IADS), coerce the enemy into exposing his forces prematurely, and defeat tracking by enemy radar. This chapter will discuss saturation decoys, towed decoys, and expendable active decoys. Chaff and flare systems will be discussed in separate chapters.

2. SATURATION DECOYS

A saturation decoy is usually an expendable vehicle designed to emulate a penetrating aircraft. Its mission is to deceive and saturate an enemy's IADS. Employing multiple saturation decoys can force an IADS to devote critical resources to engage these false targets. This depletes enemy assets available to engage penetrating aircraft. In addition, ground or air launched saturation decoys can be used to stimulate the IADS, to collect intelligence data, or to initiate attacks by suppression of enemy air defense (SEAD) assets. The three main characteristics of saturation decoys are their electronic signature, their flight program, and their mission type.

a. Saturation decoys must present an electronic signature, or radar return, that is indistinguishable from the aircraft they are protecting. Decoys can do this by either passive or active measures, or use a combination of both. A passive decoy is essentially a flying radar reflector. The size, shape, and materials used in the decoy are optimized to ensure that the proper amount of radar energy is returned to the enemy radars. Active decoys employ radar repeater systems to receive the enemy radar signal, amplify it and send back a radar return of the proper size to confuse the enemy. Reflecting or transmitting the proper size radar return is critical for both passive and active decoys. A return that is too large or too small will allow the enemy radar operator to differentiate between decoys and aircraft, causing the decoys to be ignored.

b. To continue deceiving an enemy IADS, a decoy must do more than provide the proper-sized radar return. Possessing flight characteristics similar to the aircraft it is protecting increases the probability that the decoy will effectively deceive an IADS for a sustained period of time. Modern decoys can either be powered with rockets, miniature engines, or simply glide for very long distances based upon the altitude and airspeed of the jet that releases them. Additionally, their flight paths can be pre-programmed into an onboard autopilot, allowing the decoy to fly an independent ground track, thus increasing their appearance as attack aircraft worth tracking.

c. Saturation decoys carry out two of the three decoy missions. Launched in significant numbers, they can saturate or overburden an IADS. Meanwhile, their realistic electronic image and preprogrammed flight paths entice the enemy to turn on radars and show his forces.

(1) Saturation decoys launched in coordination with an attacking strike package force the enemy to take time to process meaningless tracks and tie up critical assets. In this role, decoys primarily work against the early warning network of the enemy IADS by presenting the IADS with numerous targets to sort and track. Resources committed to tracking decoys may not be available to track actual aircraft. Additionally, if an enemy knows that decoys are present, he may not commit any resources against targets for fear they are just decoys.

(2) Time of radiation or “emission control” is a critical factor for acquisition and target tracking radars. To be effective and survive on the battlefield, ground threat radars radiate as little as possible; too much time radiating allows ELINT collectors to find their location and either direct aircraft to avoid them or call in an attack upon them. Therefore, when a decoy can get a radar to emit, the radar is now essentially compromised and can be avoided or attacked. Getting the enemy's radars to emit is called “stimulating the IADS,” which is generally a precursor to any threat suppression mission.

(3) An extremely successful example of using decoys to stimulate the IADS was carried out in the Bekaa Valley in 1982. The Israelis opened the conflict by launching saturation decoys to successfully simulate an attack. While the Syrians reloaded, Israeli fighters attacked, destroying 17 of 19 Syrian SA-6s in the beginning of the battle. With the ground threat neutralized, the Israeli Air Force went on to destroy 85 Syrian fighters in the pure air-to-air conflict that resulted.

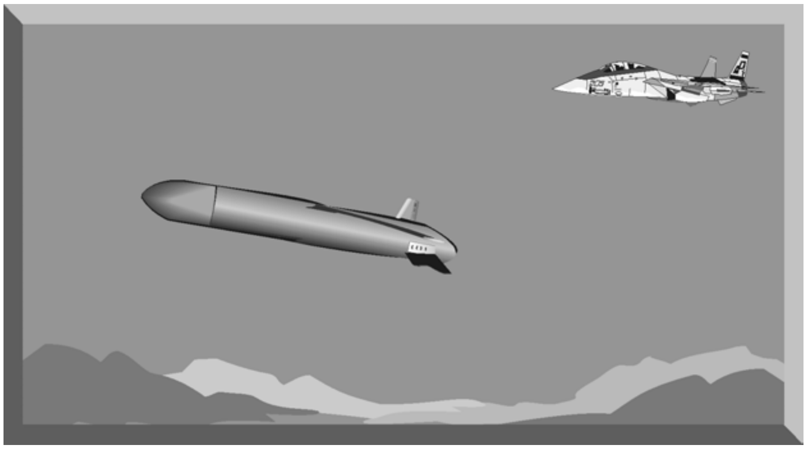

Figure 12-1. Tactical Air Launched Decoy

d. Two examples of saturation decoys are the Navy's Tactical Air Launched Decoy (TALD) in Figure 12-1 and the Air Force's proposed Miniature Air Launched

Decoy (MALD) in Figure 12-2. Both of these decoys can work actively or passively and both have pre-programmable flight paths. The TALD is an unpowered decoy normally launched from an F-14 Tomcat. The Air Force's MALD is a smaller jet-powered decoy also designed to be used by fighter aircraft. The MALD is 90 inches long, 6 inches in diameter, and has 25-inch wings that are foldable–essentially it is the size of an air-to-air missile. Because of its small size, the MALD can be carried into the target area before it is launched. Once launched it uses its speed, independent flight path, and electronically manipulated radar signature to make acquisition radars and target tracking radars mistake it for one of the attacking aircraft.

Figure 12-2. Miniature Air Launched Decoy (MALD)

3. TOWED DECOYS

A towed decoy is a small jammer that is physically attached to the aircraft (Figure 12-3). Unlike the saturation decoys that work against the IADS, the towed decoys are for individual aircraft survival. Towed decoys are designed to defeat enemy missiles in the final stages of an engagement; therefore, towed decoys, as well as other expendables, are known as endgame countermeasures. While towed decoys are primarily designed to provide sufficient miss distance between an attacking semi-active radar missile and the protected aircraft, they may also be effective against pulse Doppler radars and monopulse radars.

a. To be effective, the towed decoy must turn on within the threat radar's resolution cell after the radar is tracking the protected target. To successfully decoy the missile, the towed decoy must return radar signals with sufficient power to simulate a radar cross section (RCS) significantly larger than that of the

protected target. There are currently two generations of towed decoys on the market. Their primary difference lies in the connection each has with the aircraft towing it.

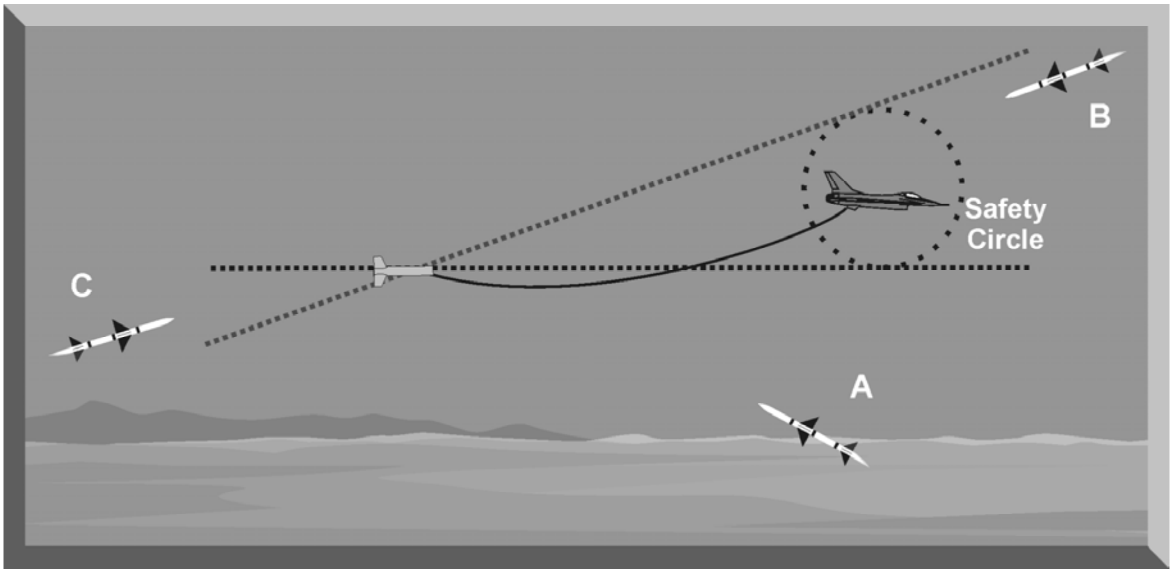

Figure 12-3. Towed Decoy

(1) The first generation of towed decoys contains a simple repeater jammer that enhances any signal it receives in the proper frequency range. The enhanced signal is stronger than the aircraft signal so the missile is lured towards the decoy. These decoys are stand-alone units that contain all the electronics, processors, receivers and transmitters within them. The only tie to the aircraft is for power and status. One of the big advantages of these simple repeater devices is that they do not require the exact frequency of the enemy radar systems to be effective, they will enhance any signal coming at them. An area of concern with the use of towed decoys is possible conflict between the onboard jamming system and the towed decoy. The onboard system could overpower the decoy, causing the attacking missile to ignore the decoy and track the aircraft.

(2) The second generation of decoys is tethered to the aircraft via fiber optic cable. Through this cable travels the different jamming modulations to be used by the decoy. These fiber optic towed decoys (FOTD) only contain the transmitters; the remaining items are in the jet or the pod. This system allows for more complex jamming through the decoy, including cooperative jamming between the aircraft and the decoy.

b. The separation required between the decoy and the aircraft is a primary consideration in developing a towed decoy system. The towed decoy should be positioned far enough behind the aircraft to preclude warhead fragments from missiles guiding on the decoy from also impacting the aircraft. Missile A in Figure

12-3 depicts a situation where the missile will detonate well outside of the aircraft's safety circle. From a pilot perspective, any restrictions on aircraft maneuvering imposed by a towed decoy are very important. The number of decoys that can be carried and the time required for decoy deployment are also important employment considerations.

c. Achieving 360° coverage is a primary limitation of a towed decoy system. When an aircraft equipped with a towed decoy is abeam a threat radar, the radar may be able to discriminate between the aircraft and the decoy. This is a function of the resolution cell of the radar. In addition, missiles approaching from a high-aspect angle, and above the aircraft (Figure 12-3 - Missile B), may fuse on the aircraft while guiding to the decoy. Missiles approaching from a low-aspect angle (Figure 12-3 - Missile C) may not fuse on the decoy and subsequently acquire and fuse on the aircraft. Finally, if the decoy is destroyed or lost, the time required to deploy a replacement decoy is critical, especially if the aircraft is engaged by multiple missiles.

d. An example of a fielded towed decoy system is the AN/ALE-50 (Figure 12-4). This first generation towed decoy system is found on Air Force F-16 and B-1 aircraft, and there is a version that is integrated into the ALQ-184 pod.

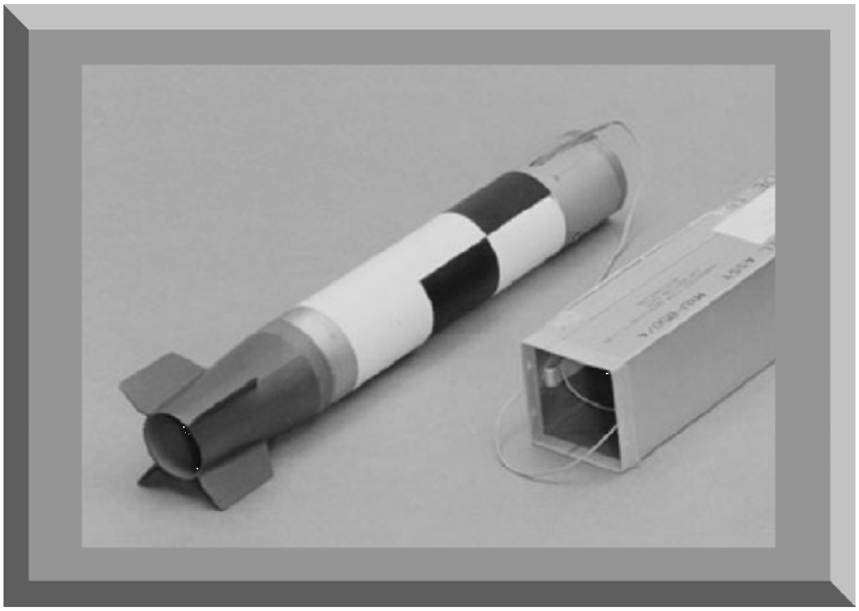

Figure 12-4. AN/ALE-50 System

(1) The system consists of a launch controller subsystem and towed decoys. The launch controller houses the decoy before it is launched, provides power to the decoy, and provides for the monitoring of the electronics. The decoy body is a factory sealed, self-contained unit with everything except for power. Power comes through the tether from the host aircraft; the decoy sends its operating status back through the tether to the aircraft.

(2) ALE-50 decoy use is cleared throughout the flight regime of the F-16 and B-1. Decoys can be deployed without being turned on, but once a decoy has been deployed, it cannot be reeled back in and must be severed before landing. Procedures are in place to reduce the chance that the onboard jamming system will negate the decoy.

(3) The ALE-50 towed decoy is a wideband RF repeater that provides self-protection EA by receiving, electronically amplifying, and retransmitting enemy radar RF signals. Upon receiving a threat radar signal, this simple repeater amplifies the signal and retransmits it. This provides the radar with two signals, one reflected from the aircraft and a stronger one from the decoy. With the signal from the decoy being the more attractive, the radar or missile guides towards the decoy. During combat operations over Kosovo, ALE-50 decoys were credited with saves for both F-16s and B-1s.

4. EXPENDABLE ACTIVE DECOYS

Expendable active decoys are designed to lure the tracking gates of an enemy's radar away from the aircraft. They are endgame countermeasures like towed decoys, but they differ in that expendable decoys free-fall or glide to the ground as opposed to being towed behind the aircraft.

a. Expendable decoys are small, active jamming systems designed to be expended by existing aircraft chaff and flare dispensers, such as the AN/ALE-40 or the AN/ALE-47. Expendable decoys can employ noise or deception jamming with noise jamming being the most common. Deception jamming techniques can be employed to enhance effectiveness against pulse Doppler radars. There are two challenges associated with expendable jammers: the amount of the time the jammer is effective and the packaging (Figure 12-5).

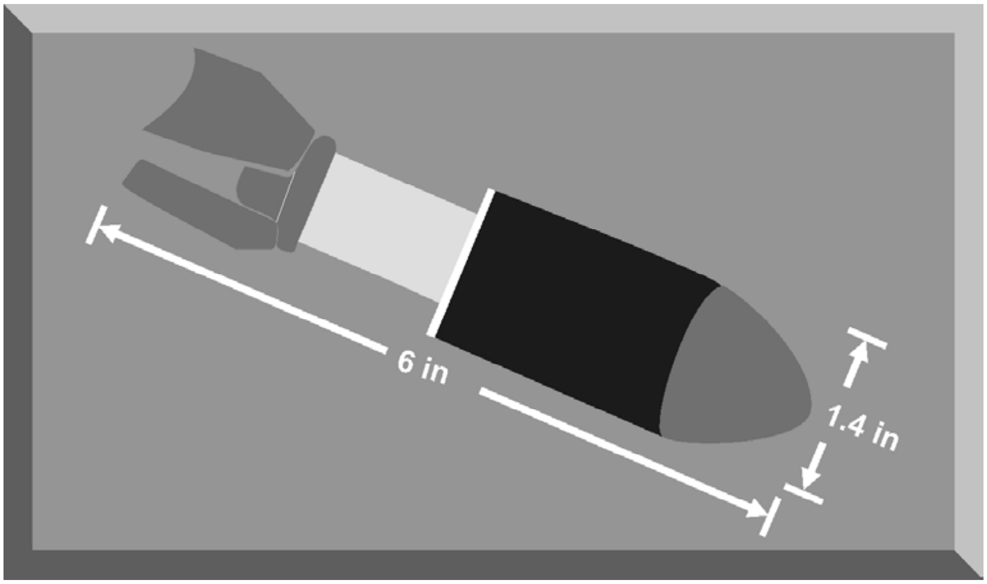

Figure 12-5. Generic Expendable

b. Expendable decoys are designed to provide protection for the dispensing aircraft for a specific period. The dispensing altitude and rate of fall determine this period of effective coverage. Expendable decoys can employ small parachutes of aerodynamic design to slow the rate of fall and increase the time of effective coverage. If the period of coverage is too short, multiple expendable decoys must be employed. This places a premium on timely employment and expendables management.

c. The primary components of an expendable decoy are the transmit and receive antennas, techniques generator, amplifier, and power supply. The transmit and receive antennas should be isolated and capable of high gain, wide bandwidth, and should use compatible polarization with the victim radar. The techniques generator must recognize the victim radar signal and generate the appropriate jamming response. The amplifier must be capable of generating a high power jamming signal over a wide frequency range. To meet these requirements, sophisticated computer and miniaturization techniques are used, and the components packaged to all fit in the aircraft dispenser. These factors impact the cost of expendable decoys and may limit the availability of these assets.

d. The Generic Expendable, RTE-1489, commonly called the GEN-X decoy is a fielded expendable active decoy. The decoy is sized to fit into a 1.4 x 1.4 x 5.8 inch cartridge and take advantage of new microwave/millimeter-wave integrated circuit (MMIC) technology. The GEN-X is programmable and features a broadband antenna and wide frequency coverage. After ejection, the decoy extends three small fins for stability. Its battery ignites to provide power, the receiver locks on to the threat radar signal, and a deception signal is generated and transmitted.

5. SUMMARY

Decoys simply provide the enemy with more targets to process. In the case of saturation decoys, this forces the enemy to commit resources against false targets, or show his defenses. For towed decoys and expendable active decoys, it makes the missile or tracking radar separate a real target from more electronically attractive decoys.

CHAPTER 13. CHAFF EMPLOYMENT

1. INTRODUCTION

Chaff was first used during World War II when the Royal Air Force, under the code name "WINDOW," dropped bales of metallic foil during a night bombing raid in July 1943 (Figure 13-1). The bales of foil were thrown from each bomber as it approached the target. The disruption of German AAA fire control and ground control intercept (GCI) radars rendered these systems almost totally ineffective. Based on this early success, chaff employment became a standard bomber tactic for the rest of the war.

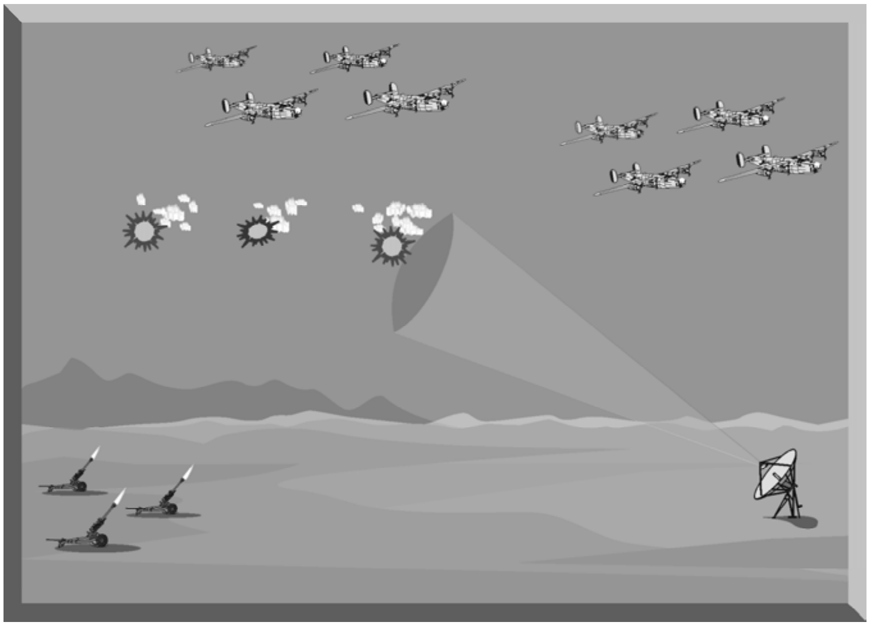

Figure 13-1. "WINDOW" – The First Operational Employment of Chaff

a. Chaff is one of the most widely used and effective expendable electronic attack (EA) devices. It is a form of volumetric radar clutter consisting of multiple metalized radar reflectors designed to interfere with and confuse radar operation. It is dispensed into the atmosphere to deny radar acquisition, generate false targets, and to deny or disrupt radar tracking. Chaff is designed to be dispensed from an aircraft and function for a limited period.

b. Even with the development and deployment of advanced radar threat systems, chaff continues to be an extremely effective EA device. Experience gained during the Vietnam conflict, the 1973 Yom Kippur War, and DESERT STORM clearly shows that chaff effectiveness against radar threats is still a factor with which the enemy must contend. This is especially true when chaff is employed with self-protection jamming and aircraft maneuvers.

c. Chaff screening and self-protection are the two basic chaff employment tactics. Chaff screening tactics, including area saturation and chaff corridor employment, are designed to confuse and deny acquisition information to the early warning, GCI, and acquisition radars supporting surface-to-air missile (SAM) systems. Self-protection tactics are designed to counter acquisition and target tracking radars (TTRs). When used with jamming and maneuvers, chaff can cause TTRs to break lock or generate survivable miss distances if a SAM is fired at the aircraft.

2. CHAFF CHARACTERISTICS

To understand how chaff affects radar systems, it is important to understand its characteristics. The most important chaff characteristics are radar cross section (RCS), frequency coverage, bloom rate, Doppler content, polarization, and persistence.

a. RCS is a measure of the net reradiated energy from a target to the illuminating radar. The RCS of an aircraft varies based on the size, shape, type of skin surface, configuration, and aspect to the illuminating radar. Figure 13-2 shows the effect of aspect on aircraft RCS. The RCS is greatest when the aircraft aspect is 90°, or abeam the radar. The lowest RCS occurs near the 30-70° and 110-150° of aspect. Since the aircraft RCS also varies based on frequency, the victim radar's frequency is a key factor. To be effective, chaff must be dispensed in large enough quantities to create an RCS greater than the aircraft RCS.

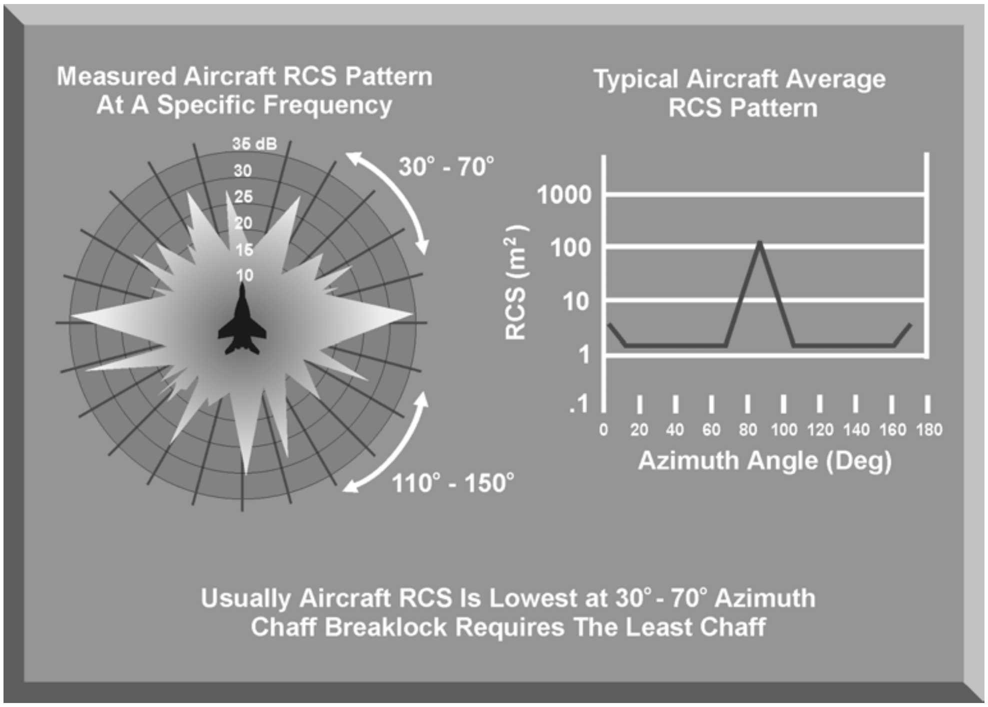

Figure 13-2. Aircraft Radar Cross Section (RCS)

(1) The RCS of a chaff bundle depends on the frequency of the victim radar and the dispensing aircraft's relative position, or aspect. Figure 13-3 shows the RCS of a single RR-170 chaff cartridge based on frequency. It shows that the largest RCS occurs at about 3 GHz. However, for the spectrum between 2-18 GHz, which includes most SAM TTRs, the RCS of the RR-170 cartridge is over 50 square meters. Since the typical fighter aircraft RCS varies between 1 and 10 square meters, depending upon frequency and aspect, the RR-170 chaff cartridge should provide a sufficient RCS to mask the aircraft RCS.

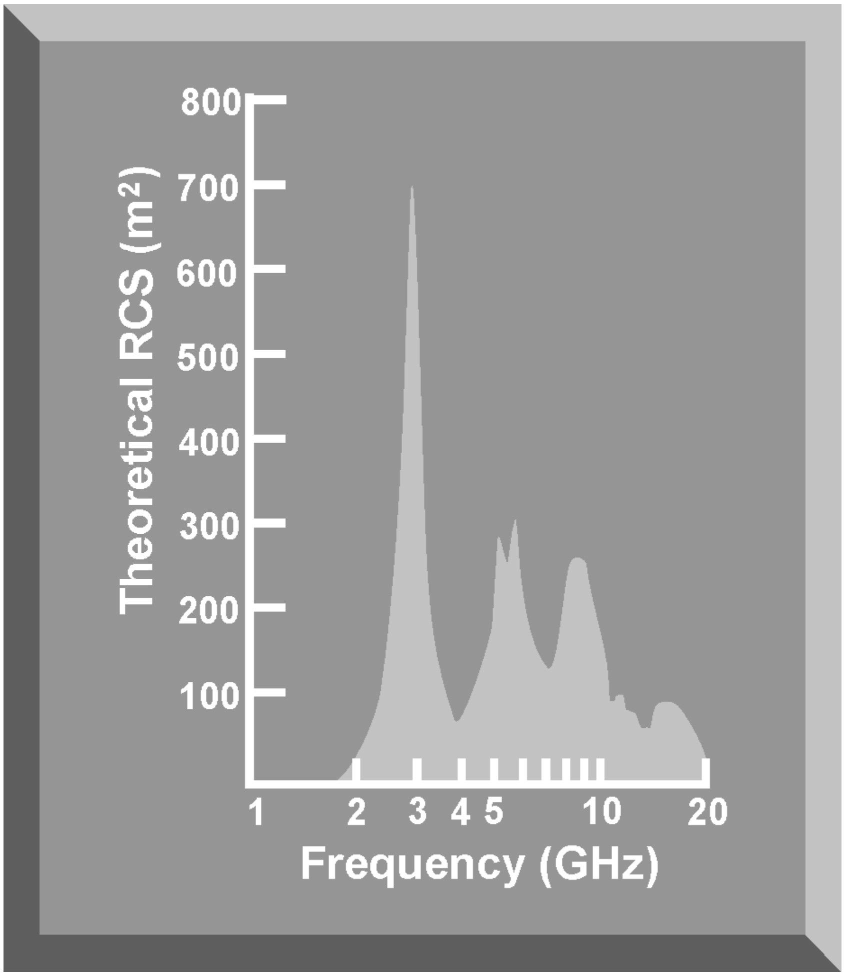

Figure 13-3. RR-170 Chaff Cartridge RCS

(2) The angular relationship, or aspect, between the aircraft and chaff bundle affects the chaff RCS presented to the victim radar. Chaff RCS is greatest when the chaff bundle and the aircraft are abeam the threat radar. It is smallest when the threat radar is off the nose or tail of the aircraft. Aspect is important when developing self-protection maneuvering and chaff dispensing tactics against threat radars (Figure 13-4).

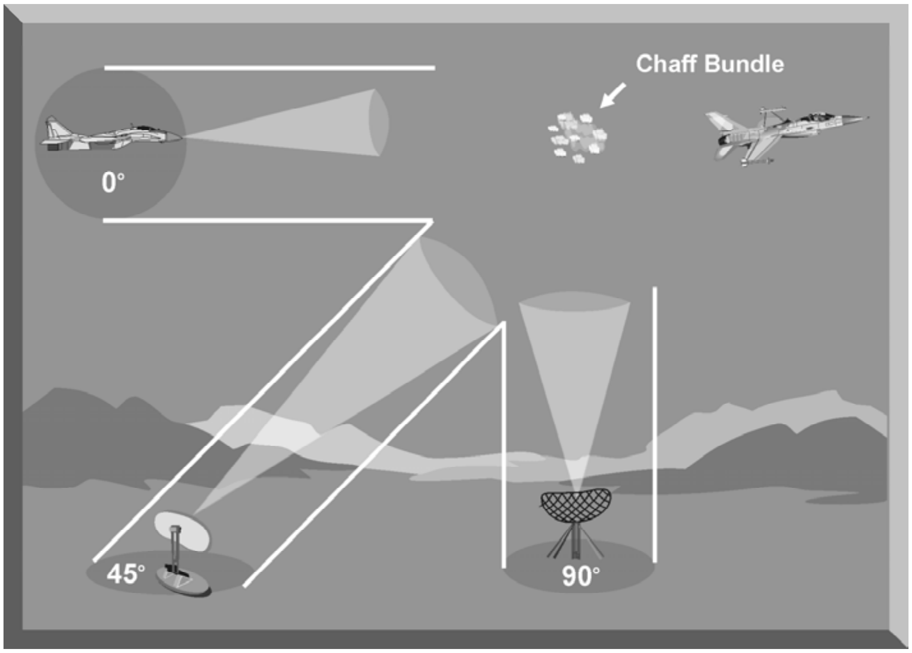
Chaff Bundle
0°
45°
90°

Figure 13-4. Threat Radar Aspect and Chaff RCS

(3) Dispensing multiple chaff bundles simultaneously does not necessarily increase chaff RCS. Multiple bundles increase the density of the chaff but do not directly enhance self-protection capabilities (Figure 13-5). This is an important consideration when developing chaff dispenser rates to counter threats.

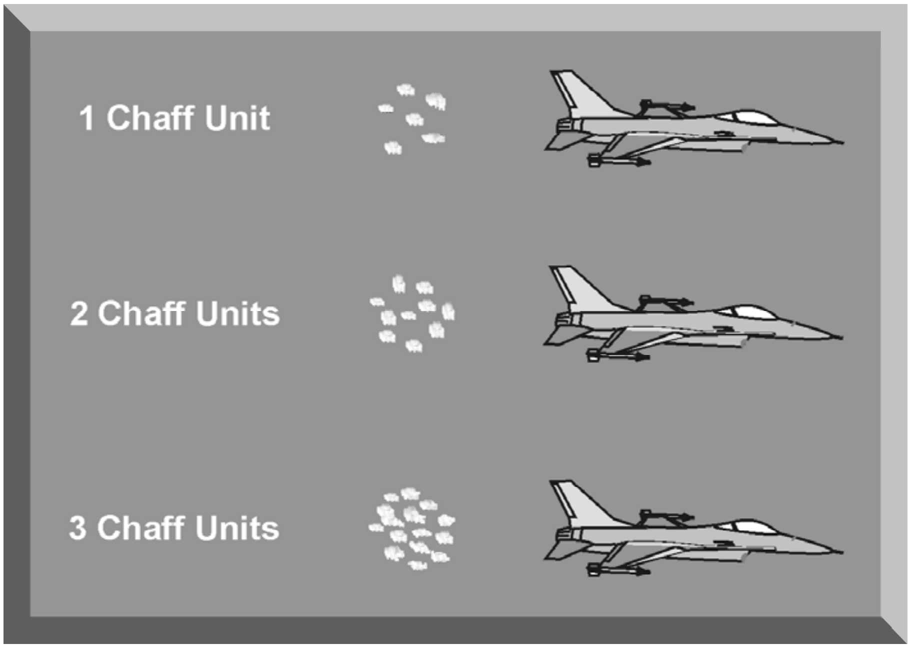
1 Chaff Unit
2 Chaff Units
3 Chaff Units

Figure 13-5. Impact of Multiple Chaff Cartridge Employment on Chaff RCS

b. Each strip of chaff is a dipole reflector that reradiates the electromagnetic energy received from an emitting radar and creates a radar echo. The optimum size is cut to about one-half the wavelength of the victim radar's RF. Since a single cut length is restricted in effectiveness to a narrow range of frequencies, different lengths are normally packaged together to provide coverage over a wide range of frequencies (Figure 13-6).

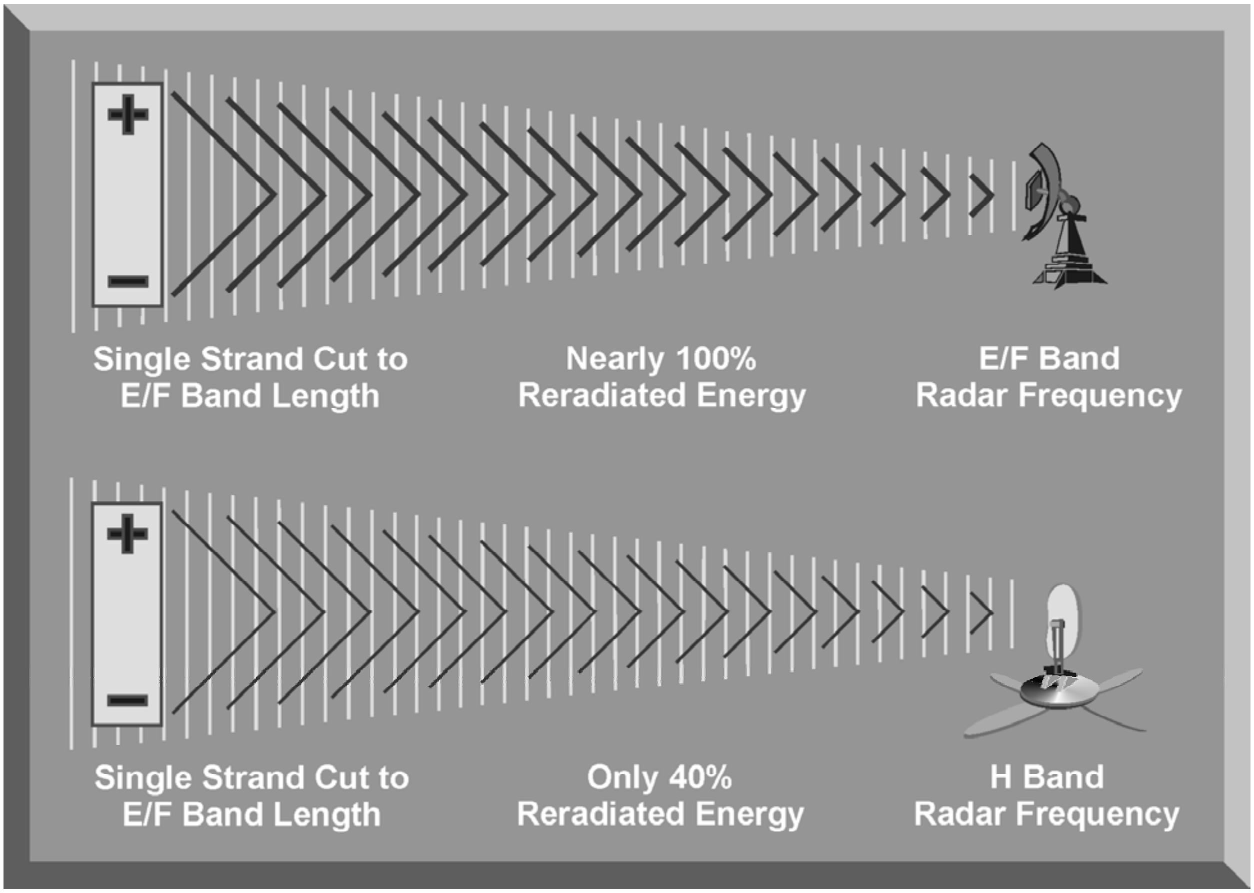
+
−
Single Strand Cut to E/F Band Length
Nearly 100% Reradiated Energy
E/F Band Radar Frequency
+
−
Single Strand Cut to E/F Band Length
Only 40% Reradiated Energy
H Band Radar Frequency

Figure 13-6. Chaff Length and Frequency Coverage

(1) Considerable research and development has reduced the size and increased the effectiveness of self-protection chaff. There are various chaff sizes, shapes, and materials. Most chaff carried on fighter aircraft are made of small aluminum strips, coated strips of nylon, or fiberglass. These strips are cut to various lengths and compressed into bundles that are small and light enough to allow the aircraft to carry and dispense multiple chaff bundles. These cuts of chaff are packaged into chaff cartridges and inserted into a dispenser on the aircraft. An explosive squib assembly ejects the cartridges from the dispenser and disperses the chaff (Figure 13-7).

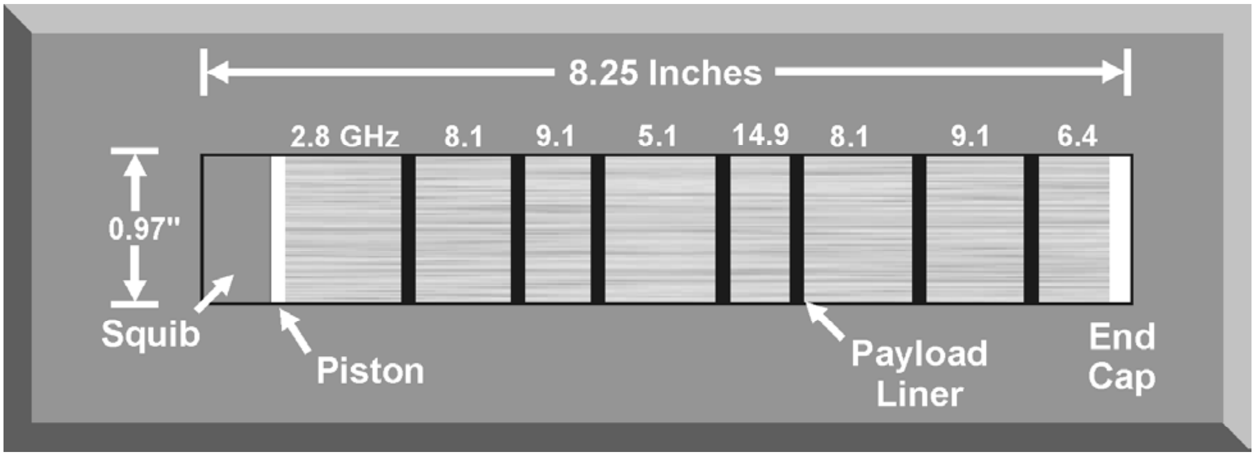

Figure 13-7. RR-170 Chaff Cartridge

(2) To provide as many dipoles as possible and present the maximum radar cross section, each chaff bundle has numerous chaff cuts to match a predetermined range of frequencies. Each chaff cartridge contains almost 3 million dipoles packaged in an eight inch by one-inch cartridge. The dipole frequencies cover the frequency range where most SAM TTRs and air-to-air radars operate (2 - 18 GHZ).

c. Bloom rate, the rate at which chaff will scatter, is also a very important characteristic of self-protection chaff. Self-protection chaff effectiveness is based on the relationship of bloom rate, chaff RCS, aircraft RCS, and the resolution cell of the threat radar system. The ability of chaff to effectively defeat a target tracking radar is directly related to the chaff dispense rate, which determines the chaff RCS, which should be larger than the aircraft's RCS. The chaff bundles must also bloom within the resolution cell of the radar.

(1) Chaff bloom rate is dependent on aerodynamic factors associated with the chaff type, the location of the dispenser on the aircraft, and the aircraft wake or turbulence. Heavy or dense chaff falls faster and blooms slower than lighter and less dense chaff. The location of the chaff dispenser on the aircraft affects the airflow in which the chaff will be dispensed. The ideal position for the dispenser is in the area where there is the most turbulence from the aircraft. Turbulence behind the aircraft is probably the most important factor affecting bloom rate. The more turbulent the airflow, the greater the bloom rate (Figure 13-8). Maneuvering the aircraft while dispensing chaff also enhances the chaff bloom rate.

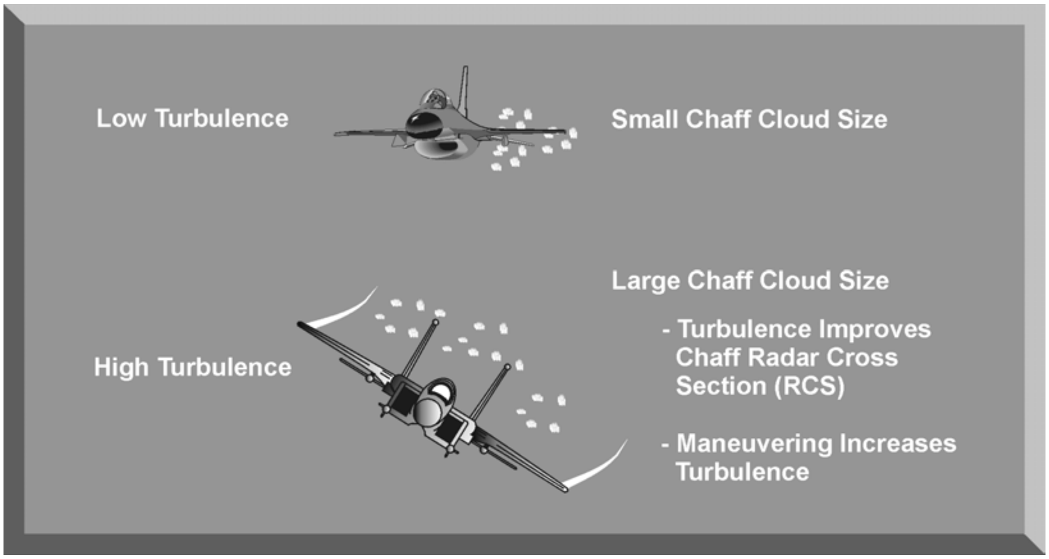
Low Turbulence
Small Chaff Cloud Size
Large Chaff Cloud Size
- Turbulence Improves Chaff Radar Cross Section (RCS)
High Turbulence
- Maneuvering Increases Turbulence

Figure 13-8. Impact of Turbulence and Chaff Bloom Rate

(2) To ensure that the victim radar is decoyed or that it transfers automatic tracking to the chaff, the chaff must bloom within the radar resolution cell. This resolution cell is a three-dimensional spheroid with dimensions based on the pulse width, horizontal beamwidth, vertical beamwidth, and the range of the aircraft (Figure 13-9). There are some rules of thumb that can be used when considering the bloom rate of chaff and the resolution cell of a particular radar. The shorter the pulse width of a radar, the faster the chaff has to bloom to be effective. The narrower the horizontal and vertical beamwidths, the faster the chaff has to bloom to be effective.

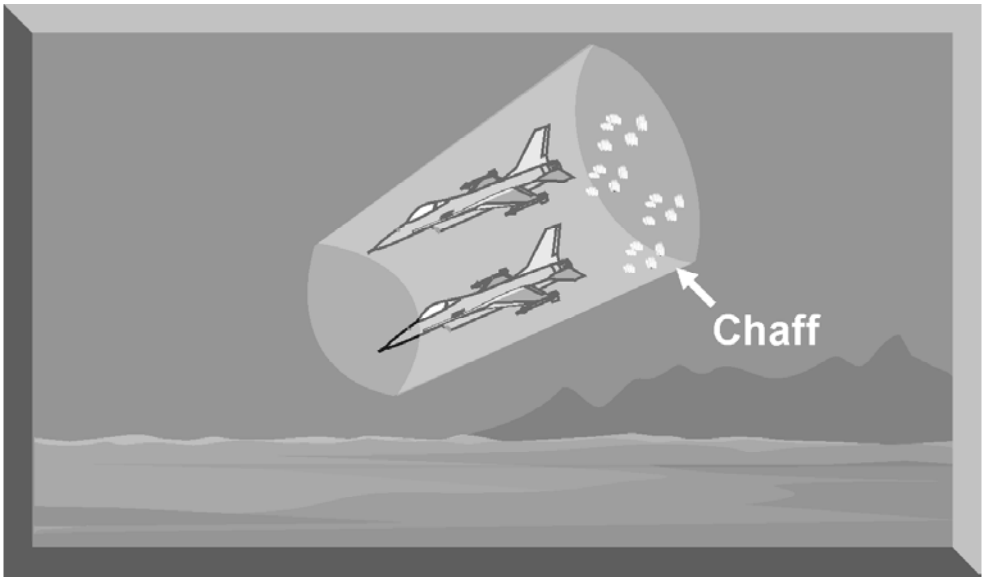

Figure 13-9. Chaff Bloom Rate and Radar Resolution Cell

d. Against Doppler radars, self-protection chaff is most effective when dispensed at or near the beam, relative to the threat radar. When chaff is dispensed in the airstream, the drag on an individual dipole is so great compared to its mass that it slows to the velocity of the surrounding air mass almost instantly. Since the relative velocity of the chaff, in relation to the radar, is zero, radar systems employing Doppler processing and tracking will not display the chaff. Doppler processing radars will continue to track the aircraft unless it also has a relative velocity of zero. This occurs when the aircraft is abeam the radar. Chaff corridor and area saturation tactics against Doppler tracking radars will have limited effectiveness.

e. Chaff persistence and polarization are two additional characteristics that are important employment considerations for area saturation or chaff corridor operations. These individual chaff element characteristics are directly related to the combined effects of aerodynamic, atmospheric, and gravitational influences.

(1) Chaff persistence is the length of time the chaff is at an effective altitude to screen ingressing aircraft during area saturation or chaff corridor operations. The time span depends on the fall rate of the chaff and varies according to the density of the dipoles. The prevailing atmospheric conditions, such as wind and temperature also affect chaff persistence. Generally, the longer cuts used for lower frequency radars fall faster than the shorter cuts used for higher frequency radars. Each type has its own rate of fall based on these conditions. The rate of fall is a critical mission planning consideration for determining the amount of time between chaff corridor or area saturation initiation and the arrival of the aircraft being screened. If the chaff is employed too early, it may not be at the correct altitude or may have dispersed to the point that it is not effective to screen ingressing aircraft.

(2) Each chaff strand is a polarized dipole with positive and negative ends. The orientation of these strands determine their polarity (Figure 13-10). Chaff cuts with the positive and negative ends oriented vertically are vertically polarized. Chaff cuts with the positive and negative ends oriented horizontally are horizontally polarized. Since chaff strands are initially buffeted by turbulence and airstream vortices, the dipole orientation and polarization, changes rapidly and randomly. Eventually, the strands separate into two groups; one descending horizontally, and one descending vertically. Since the vertically oriented strands tend to fall faster, the lower part of the chaff cloud tends to become more vertically polarized, while the upper portion is horizontally polarized. A threat radar that uses vertical polarization will receive minimal affects from the upper (horizontally polarized) portion of the chaff cloud. If the aircraft being screened are flying within this portion of the chaff cloud, they may be detected and engaged. This is another mission planning consideration for chaff area saturation or chaff corridor operations.

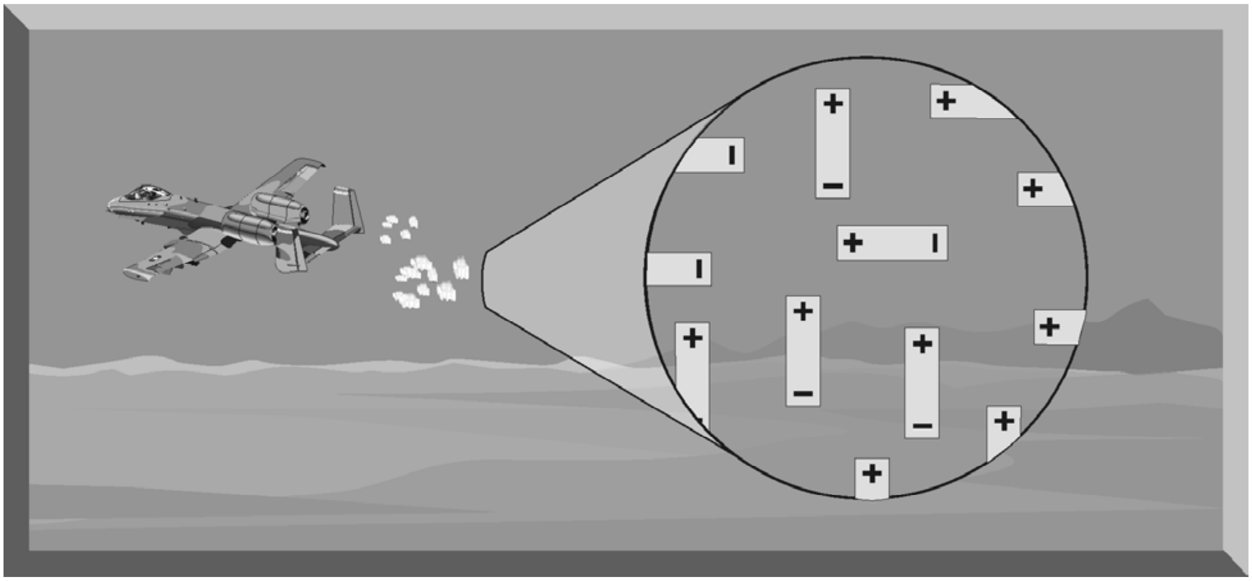

Figure 13-10. Impact of Chaff Polarization

3. CHAFF OPERATIONAL EMPLOYMENT

The two primary chaff employment tactics are force screening and self-protection. Force screening tactics include area saturation and corridor operations. Self-protection tactics include the reactive employment of chaff to negate a potentially lethal engagement. Different chaff dispensing techniques are used for each employment tactic and are important planning considerations for all chaff employment tactics. This section will discuss area saturation, corridor operations, and self-protection chaff employment.

a. The objective of area saturation operations is to present multiple false targets in a specific area in order to saturate radar systems and confuse the enemy integrated air defense system (IADS). Area saturation can be accomplished by fighter aircraft or drones equipped with chaff pods employing random chaff dispensing techniques. The chaff dispenser is set to release random bursts of chaff along the ingress and egress route of the attack package. Chaff pods may be supplemented with chaff bombs containing special fuses that provide false targets at varying altitudes. Attack aircraft can also contribute to area saturation by randomly dispensing self-protection chaff as they ingress and egress. However, this tactic can deplete the number of chaff bundles an attack aircraft may need to defeat a potentially lethal radar system encountered at a later time in the mission.

(1) The chaff cuts must provide frequency coverage for the threat radar systems. Also the RCS of each chaff burst should be large enough to present a realistic target to the victim radars. Multiple false targets created by chaff area saturation may confuse threat system operators and encourage them to expend missiles on false chaff targets.

(2) Saturation also masks the number of attacking aircraft (Figure 13-11). When used with false target deception jamming, area saturation can greatly enhance mission success. However, the technique is resource-intensive since aircraft employing chaff pods and chaff bombs cannot attack targets. These aircraft are vulnerable to attack and should be supported by standoff jamming. Area saturation tactics may have limited success against Doppler processing radars.

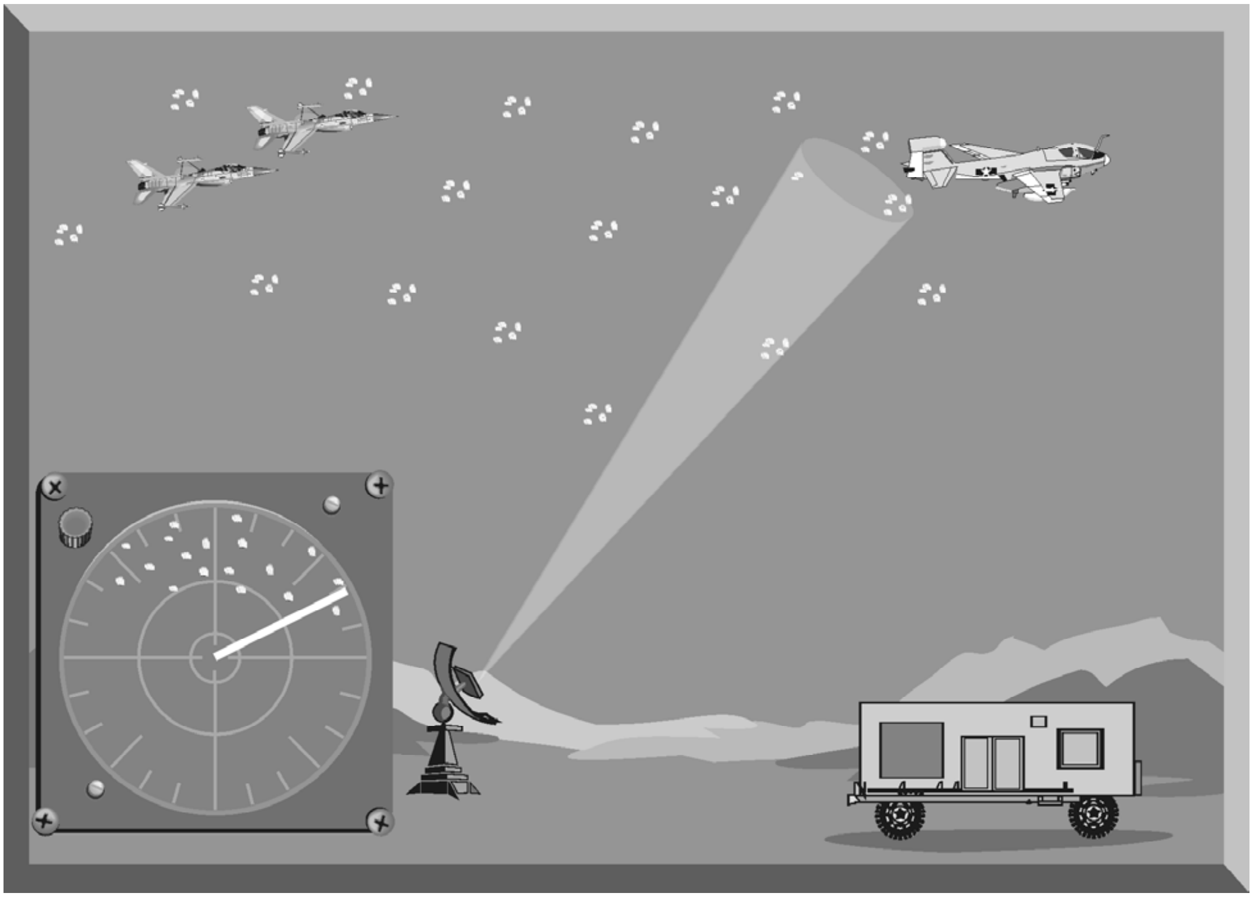

Figure 13-11. Area Saturation Tactics

b. The objective of chaff corridor operations is to screen the ingress and egress of an attack package by dispensing large quantities of chaff in a continuous "ribbon." Fighter aircraft, or drones equipped with chaff pods such as the ALE-38, employ a stream chaff dispensing technique to "lay" the chaff corridor. The pods are set to provide a continuous line of chaff dense enough to hide ingressing and egressing aircraft. The chaff cuts should provide frequency coverage for the radar systems that must be countered. Timing for the chaff aircraft in relation to the attack package must consider the fall rate and persistency of the chaff to ensure that the chaff corridor covers the required altitude for a time sufficient to allow the attack package to ingress and egress. An effective chaff corridor completely denies a radar's ability to distinguish between

the chaff and the attack aircraft. To do this, the radar cross section, or RCS, of the chaff within the resolution cell of the radar must exceed the RCS of the aircraft. This condition must be met throughout the length of the chaff corridor. When this condition is met, the chaff corridor will appear as a continuous return on the victim radar scope, and the attack package cannot be detected (Figure 13-12).

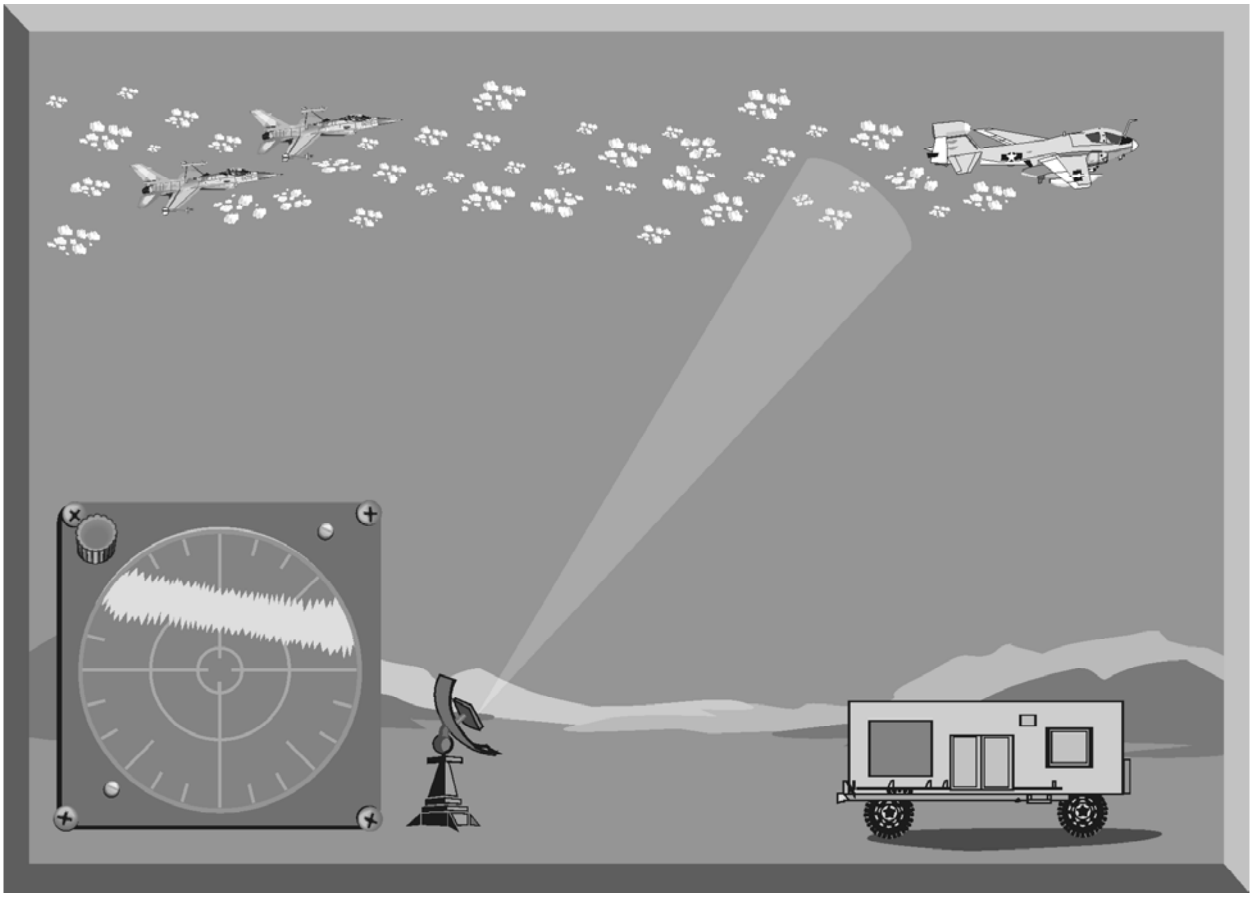

Figure 13-12. Chaff Corridor Tactics

(1) One advantage of a chaff corridor is that it can screen ingressing and egressing aircraft from pulse radar systems. However, chaff corridors are resource-intensive. Aircraft "laying" the corridor cannot strike critical targets. The chaff aircraft are also vulnerable to attack. Therefore, standoff jamming and self-protection jamming systems should be employed to provide some screening and protection for the chaff dispensing aircraft. Finally, chaff corridors may not be effective against radars with Doppler processing.

(2) To be effective, chaff corridor operations require detailed planning. Electronic combat (EC) planners must first determine that a chaff corridor is the most effective way to screen the attack force. This decision is based on the vulnerability of the attack aircraft to the anticipated threat radar systems and the availability of chaff assets. Once the decision is made to employ a chaff corridor,

planners must select the location, determine the length of the chaff corridor, select the ingress and egress altitudes, and establish the timing for the chaff aircraft and the attack package. Once the location of the chaff corridor is determined, planners must assess the threat radar systems that must be countered. The specific operating frequencies of the threat radars will determine the cuts of chaff that must be dispensed. The resolution cells of the threat radars will determine the density of chaff required. The length of the chaff corridor and the chaff density will determine the number of chaff aircraft required to seed the chaff corridor. The chaff fall rate and the atmospheric conditions impact the timing between the chaff aircraft and the attack package, and the altitude that the chaff dispensing aircraft must fly.

c. Self-protection chaff tactics are based on the use of chaff dispensers that use burst chaff dispensing techniques to defeat a TTR. Burst chaff dispensing, employed during the final phase of an engagement by air-to-air or surface-to-air weapons, can generate tracking errors or a radar break-lock. Burst chaff effectiveness is greatly enhanced when accompanied by jamming and evasive maneuvers (Figure 13-13).

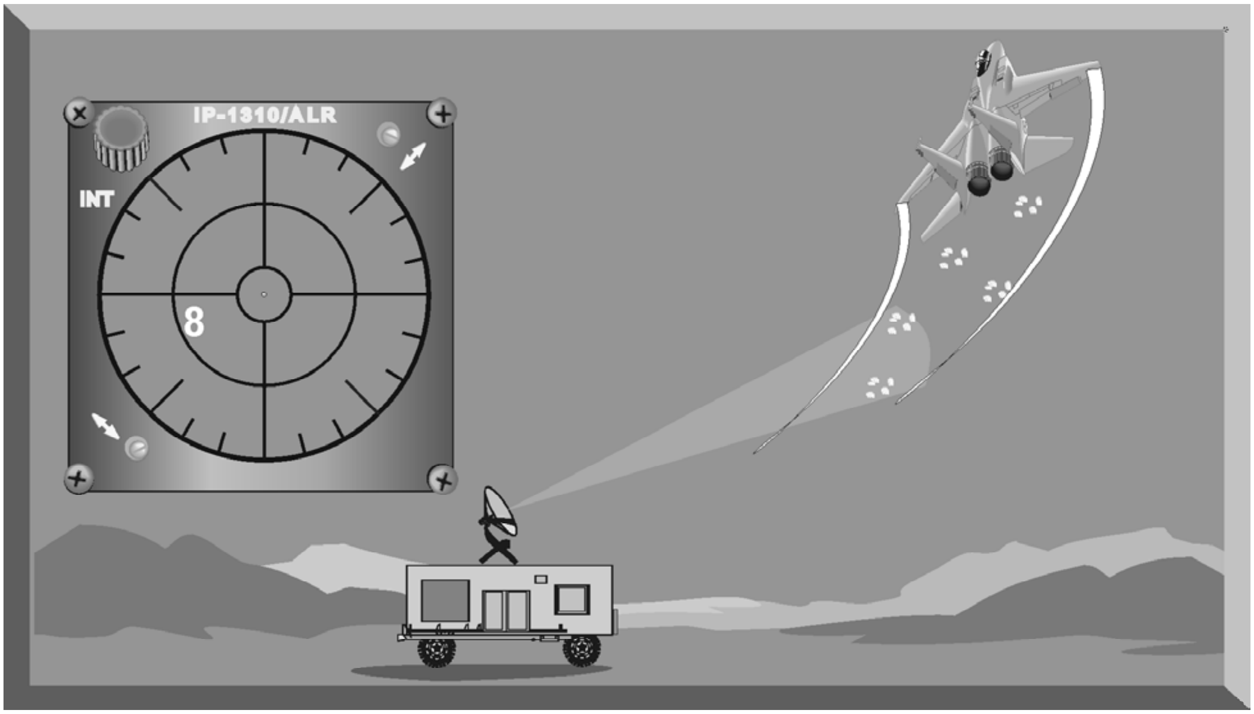

Figure 13-13. Self-Protection Chaff Tactics

(1) Self-protection chaff has proven effective against all pulse radar threat systems when employed with maneuvers and jamming. This is especially true for TTRs operating in an automatic tracking mode.

(a) Chaff employed against a track-while-scan (TWS) radar is designed to put multiple targets, with an RCS greater than the aircraft, in the resolution cell of the horizontal and vertical radar beams (Figure 13-14). Since the tracking loop tracks the largest return, the TWS radar will automatically switch to the chaff. After dispensing chaff, the pilot can maneuver vertically or horizontally to move the aircraft out of the resolution cell.

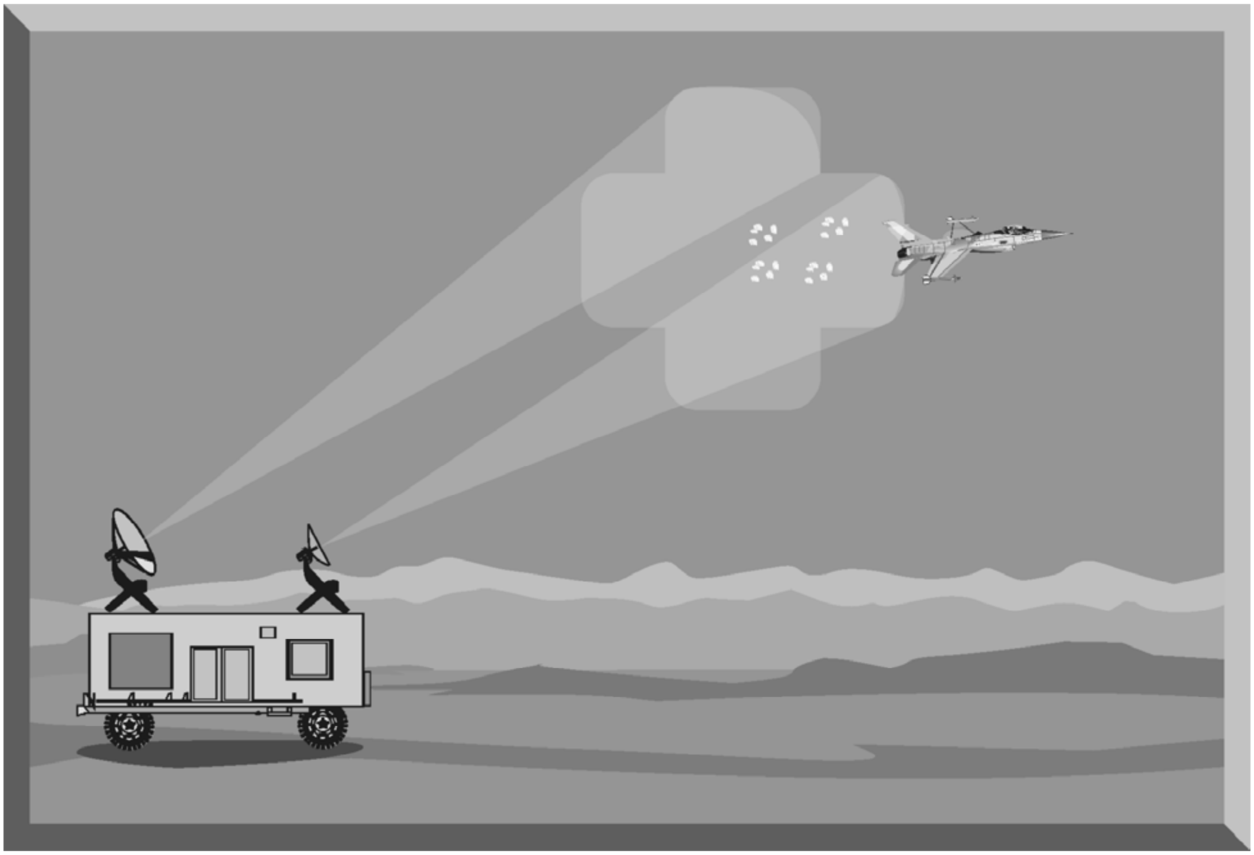

Figure 13-14. Self-Protection Chaff Effect on a TWS Radar

(b) Against a conical scan radar, chaff puts multiple, large RCS targets within the separate scans of the radar (Figure 13-15). These multiple targets generate error signals in the tracking loop and drive the separate scans off the aircraft return. As the conical scan radar tracking loop attempts to resolve these error signals, it will eventually lock on to the chaff. Maneuvering outside the overlapping scan area enhances chaff effectiveness and facilitates the transfer of radar lock-on to the chaff.

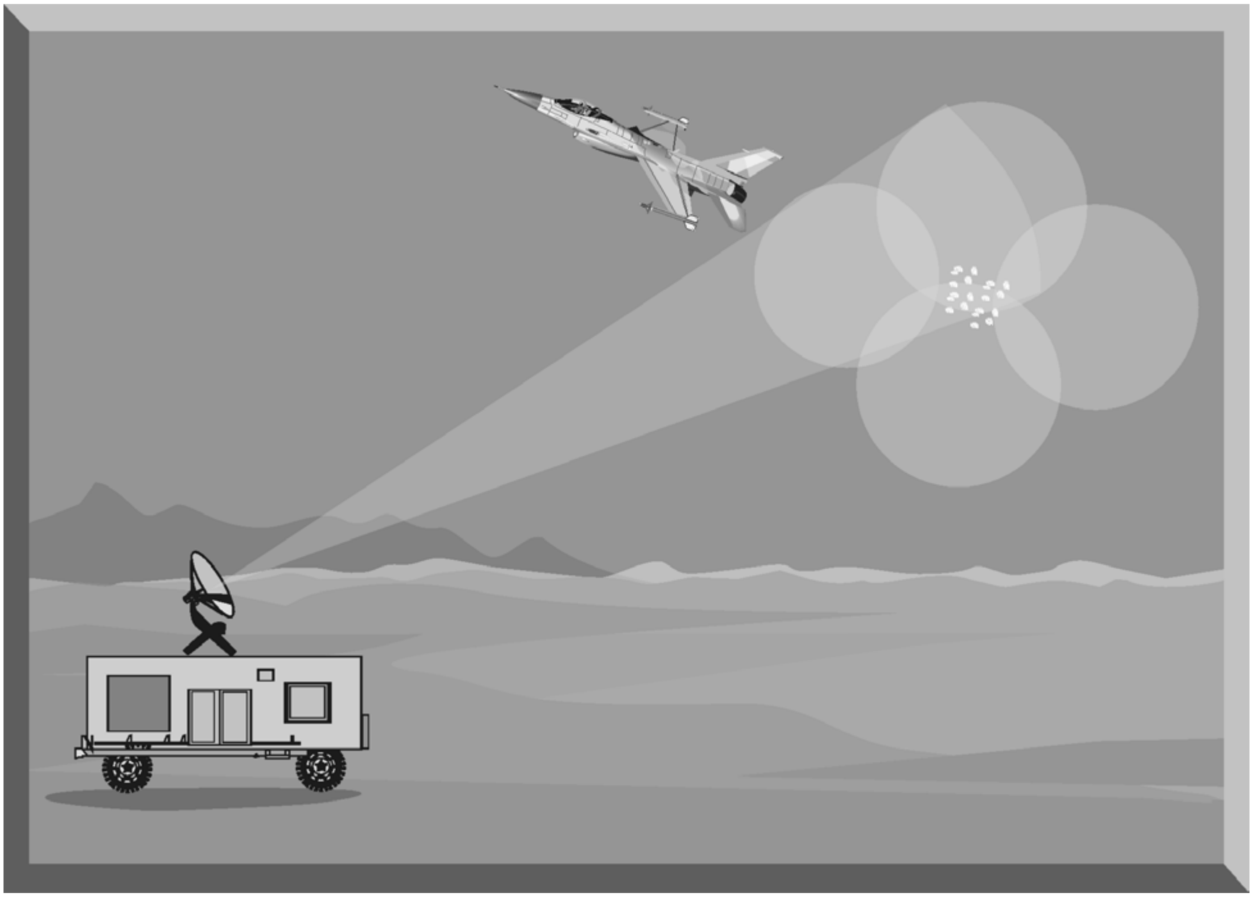

Figure 13-15. Self-Protection Chaff Impact on a Conical Scan Radar

(c) Chaff employed against a monopulse radar is designed to put multiple targets in at least two of the tracking beams (Figure 13-16). This generates errors in the azimuth, elevation, and range tracking circuits. Multiple chaff targets continue to generate azimuth and elevation errors that can eventually generate a break-lock condition, as the radar transfers lock-on to the chaff. Chaff is most effective against monopulse radars when employed on the beam in order to create the maximum angular tracking error.

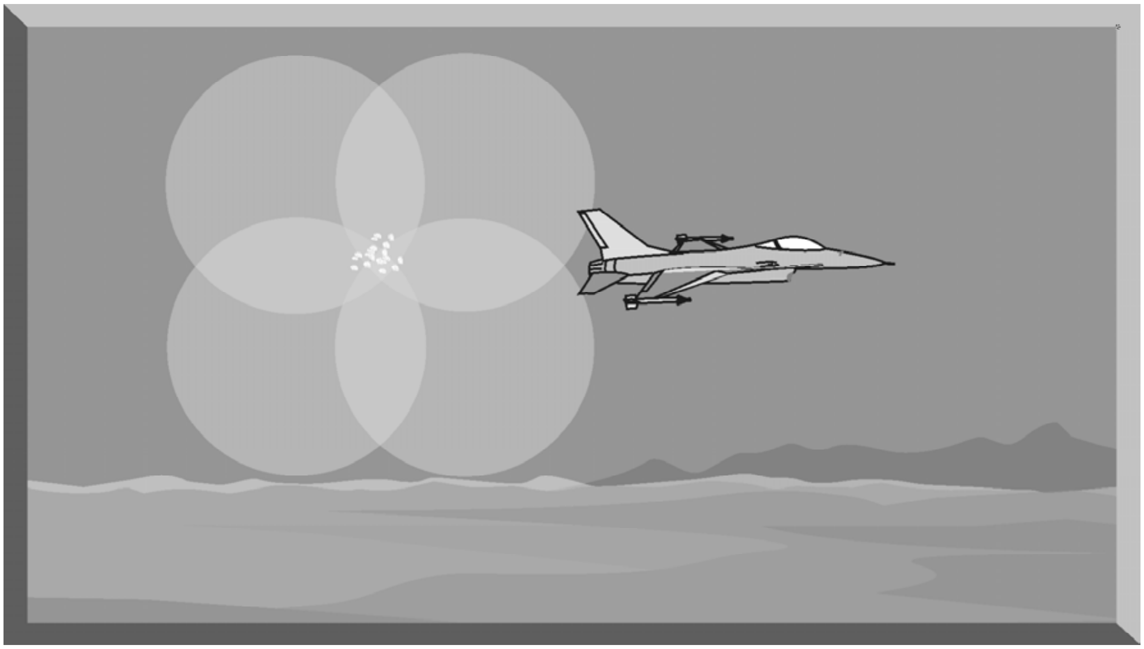

Figure 13-16. Self-Protection Chaff Impact on a Monopulse Radar

(d) Modern radars may employ some form of Doppler filtering to negate the effectiveness of chaff and other sources of clutter. Pulse Doppler and continuous wave radar systems track targets based on target velocity relative to the radar. Radars employing a moving target indicator (MTI) use relative target velocity to distinguish between targets and clutter. Chaff slows to near zero relative velocity almost immediately after dispensing. For self-protection chaff to be effective, the aircraft velocity relative to the radar site must also be near zero This occurs when the aircraft's aspect to the radar is 90°, or on the beam. By maneuvering to a beam aspect against a Doppler radar, the pilot is exploiting the "notch" where radar cannot discriminate targets based on Doppler frequency shift (Figure 13-17).

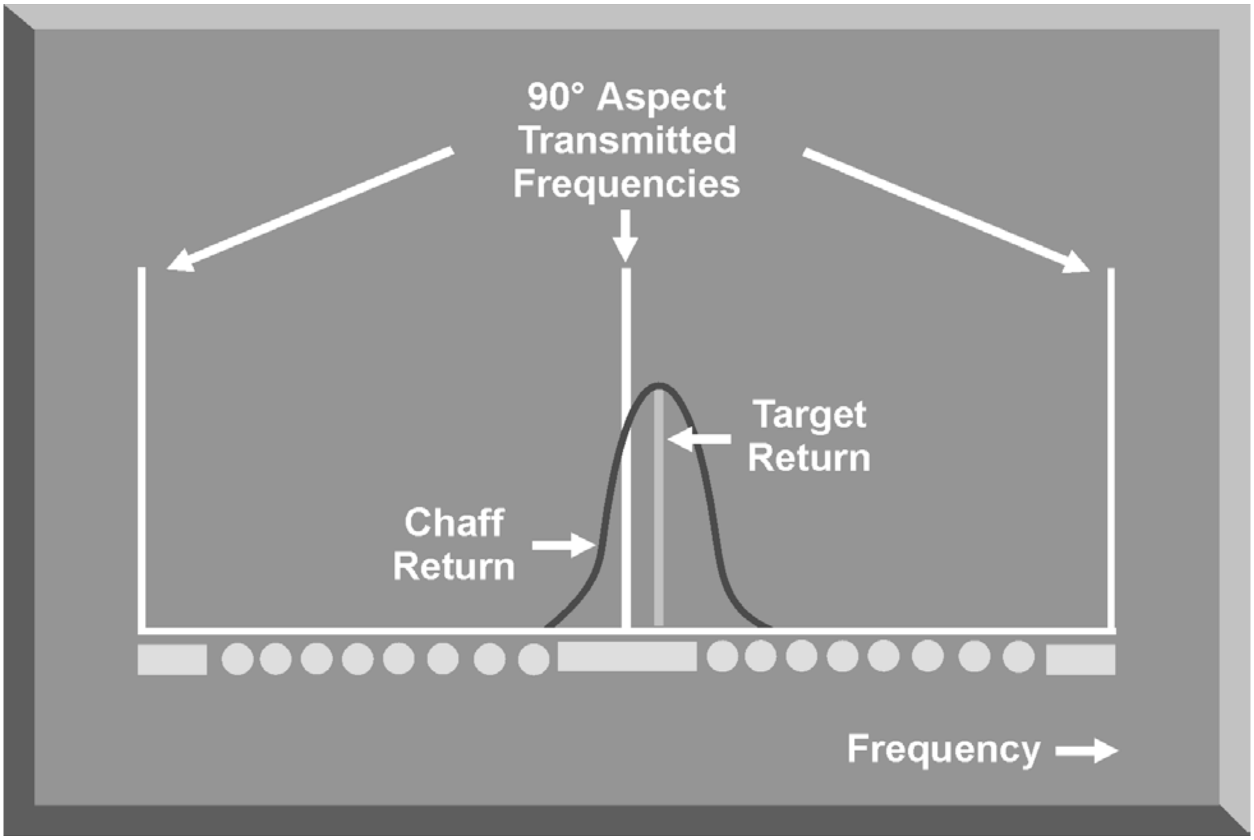
90° Aspect Transmitted Frequencies
Target Return
Chaff Return
Frequency

Figure 13-17. Self-Protection Chaff Impact on a Doppler Radar

4. SUMMARY

Chaff is one of the oldest and most effective pulse radar countermeasures. The fundamental characteristics of chaff (RCS, frequency coverage, bloom rate, Doppler content, polarization, and persistence) determine the effectiveness of chaff employment. The primary chaff employment tactics of force screening and self-protection are designed to maximize the impact of chaff on threat radar systems. Self-protection chaff, together with jamming and maneuvers, is often the "last line of defense" against lethal radar threat systems.

CHAPTER 14. IR FUNDAMENTALS

1. INTRODUCTION

Since their introduction in the 1950s, infrared (IR) missiles have been an increasing threat from both ground-based and airborne systems. The range, reliability, and effectiveness of IR missiles have been continuously improved by advanced detector materials and computer technology. Since IR missiles are passive, they are relatively simple and inexpensive to produce. These characteristics have contributed to the proliferation of IR missiles in the combat arena. Nearly every aircraft flying in either the air-to-air or air-to-surface role now carries an all-aspect IR missile. Additionally, every infantry unit down to the platoon level is equipped with shoulder-fired IR missiles (Figure 14-1). This chapter will cover basic IR theory, IR missile detection, IR seekers, and conclude with a section on IR flare rejection.

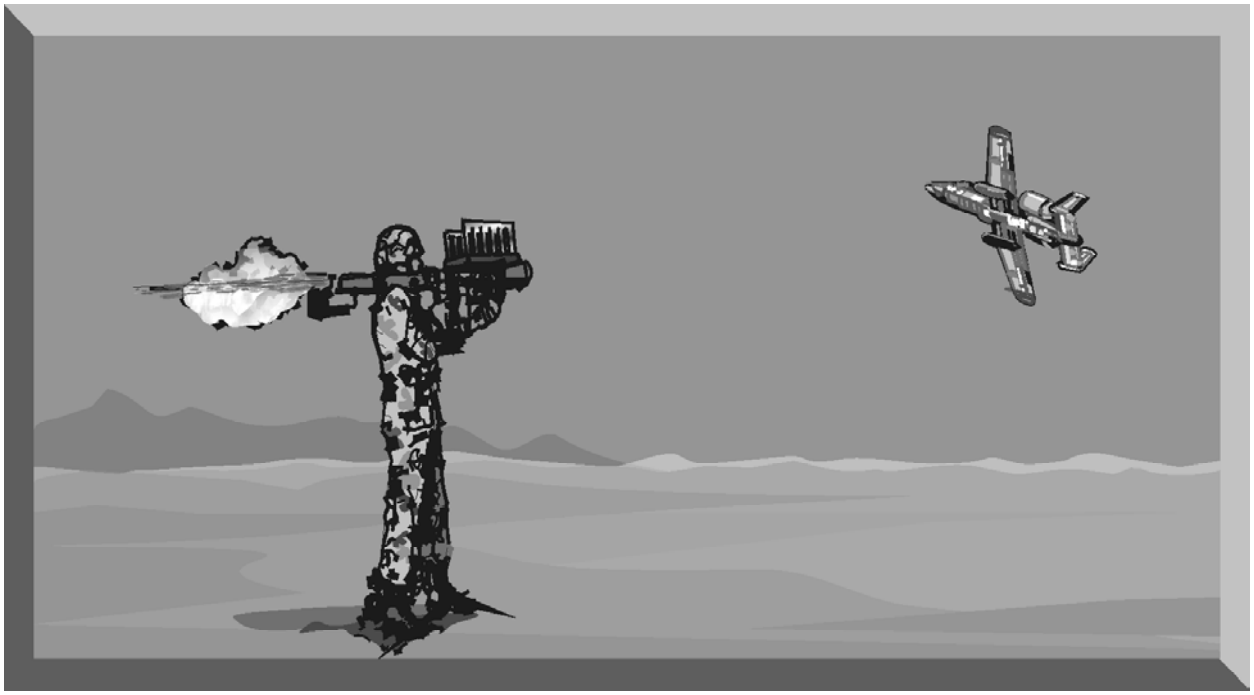

Figure 14-1. The IR Threat

2. BASIC IR THEORY

Because of its location in the frequency spectrum, IR radiation exhibits some of the characteristics and limitations of microwaves and visible light.

a. All warm objects emit IR energy. The object's temperature dictates the characteristics of this radiation. As the temperature of the material increases, the radiant intensity increases and shifts to shorter and shorter wavelengths or

higher and higher frequencies. The frequency band of IR radiation falls between the upper limit of microwaves and the lower limit of visible light. When discussing IR radiation, it is more convenient to refer to wavelength instead of frequency (Figure 14-2). The wavelength of the highest frequency IR is 0.72 x 10^{-6} meters. A unit of measure called the micron (μ), is one millionth, or 10^{-6}, of a meter and is used to designate IR wavelengths. IR energy falls in the electromagnetic spectrum between the wavelengths of 1000 and 0.72 microns, while visible light occupies the spectrum from 0.72 to 0.39 microns.

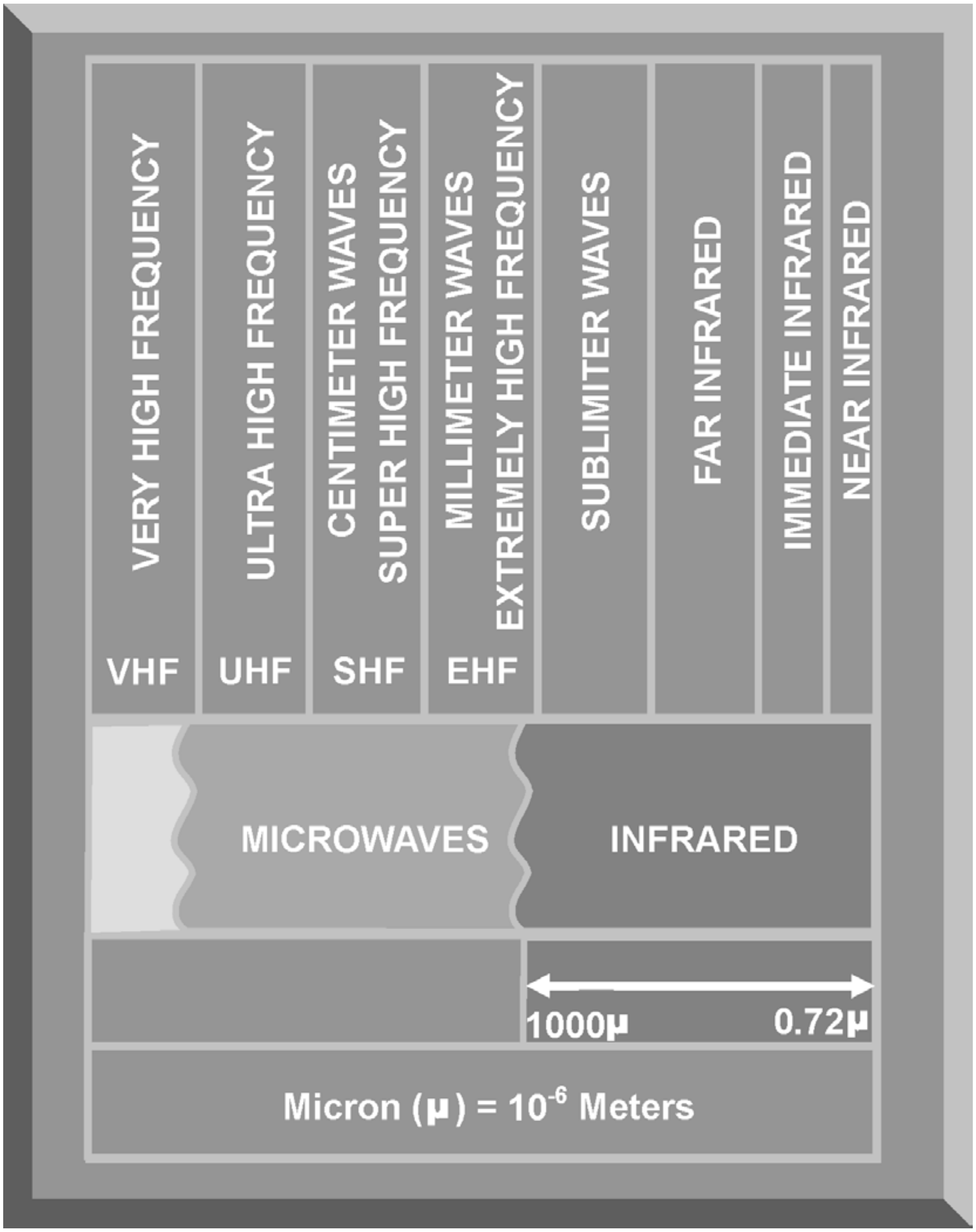
VERY HIGH FREQUENCY
VHF
ULTRA HIGH FREQUENCY
UHF
CENTIMETER WAVES
SUPER HIGH FREQUENCY
SHF
MILLIMETER WAVES
EXTREMELY HIGH FREQUENCY
EHF
SUBLIMITER WAVES
FAR INFRARED
IMMEDIATE INFRARED
NEAR INFRARED
MICROWAVES
INFRARED
1000μ
0.72μ
Micron (μ) = 10^-6 Meters

Figure 14-2. IR Frequency Band

b. As the IR energy travels through the atmosphere, certain wavelengths are absorbed or attenuated. The greatest IR attenuator is atmospheric water vapor, which varies as the weather conditions vary, with negligible absorption at altitudes above 30,000 feet. Another significant attenuator is carbon dioxide. The percentage of carbon dioxide in the atmosphere is practically constant up to a height of 30 miles. Carbon dioxide absorption is predictable and occurs only in the IR region. Scattering is another form of atmospheric attenuation and is caused by dust particles and water droplets. Scattering is also largely dependent on weather conditions and cannot be predicted. Most of the scattering occurs at lower altitudes and at the shorter wavelengths. Other atmospheric elements cause little or no attenuation of IR energy. Figure 14-3 shows atmospheric IR transmission at sea level. There is a relatively large window of IR transmission in the region from one to five microns. This is the region where the aircraft engine heat signature is at its maximum intensity and where most IR missiles are designed to operate.

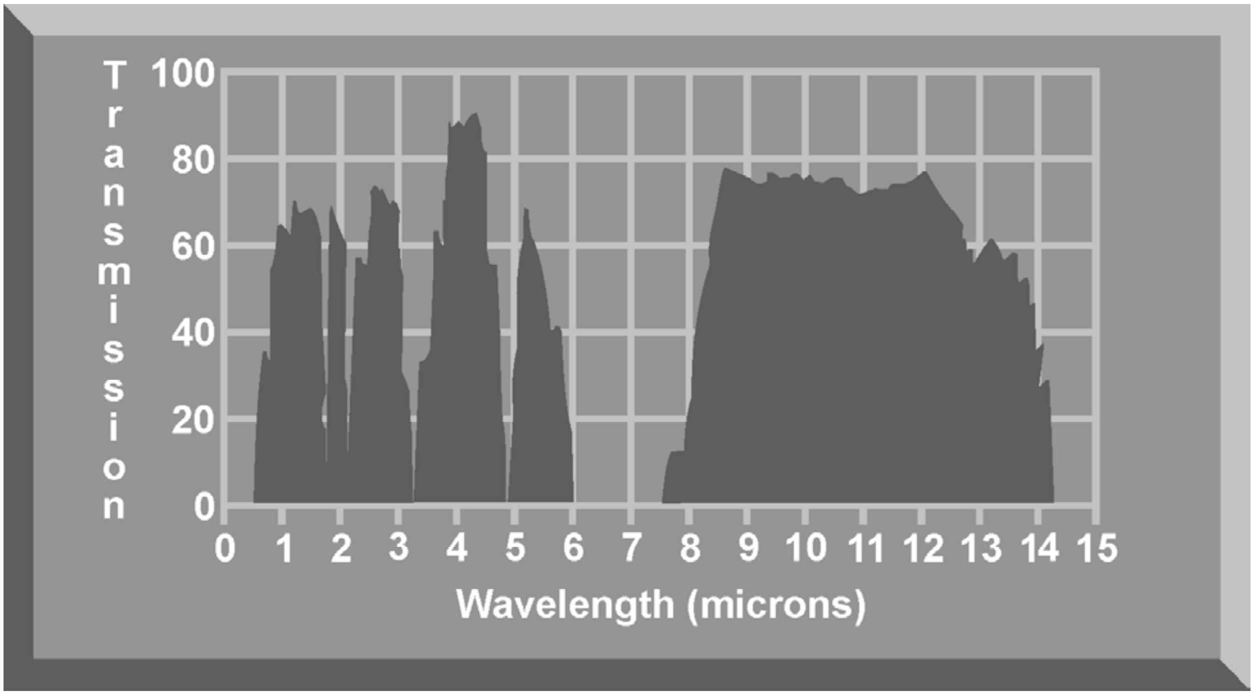

Figure 14-3. Atmospheric IR Transmission

c. IR energy exhibits some of the transmission characteristics of both RF energy and visible light. As with visible light, it can be optically focused by lenses and mirrors. This characteristic is used in the IR missile detector elements (Figure 14-4). As with RF energy, the intensity of IR radiation diminishes inversely with the square of the distance between the source and the receiver.

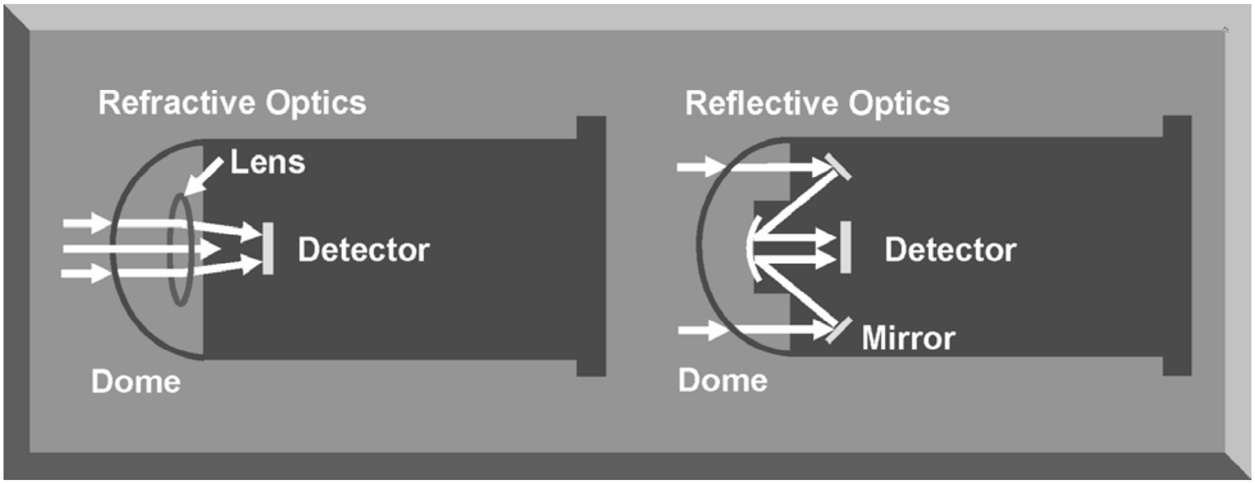
Refractive Optics
Lens
Detector
Dome
Reflective Optics
Detector
Mirror
Dome

Figure 14-4. IR Missile Seeker

d. If the intensity of the aircraft's IR signature, while in the military power setting, is plotted in relation to wavelength, it reaches a peak at approximately 3 microns (Figure 14-5). In afterburner, the aircraft's IR intensity reaches a peak at approximately 1.5 microns. Since IR missiles are designed to detect and track the aircraft's IR signature, most IR missiles operate in the region of 1 to 5 microns. To be effective, the IR intensity of a flare must also fall within this micron region.

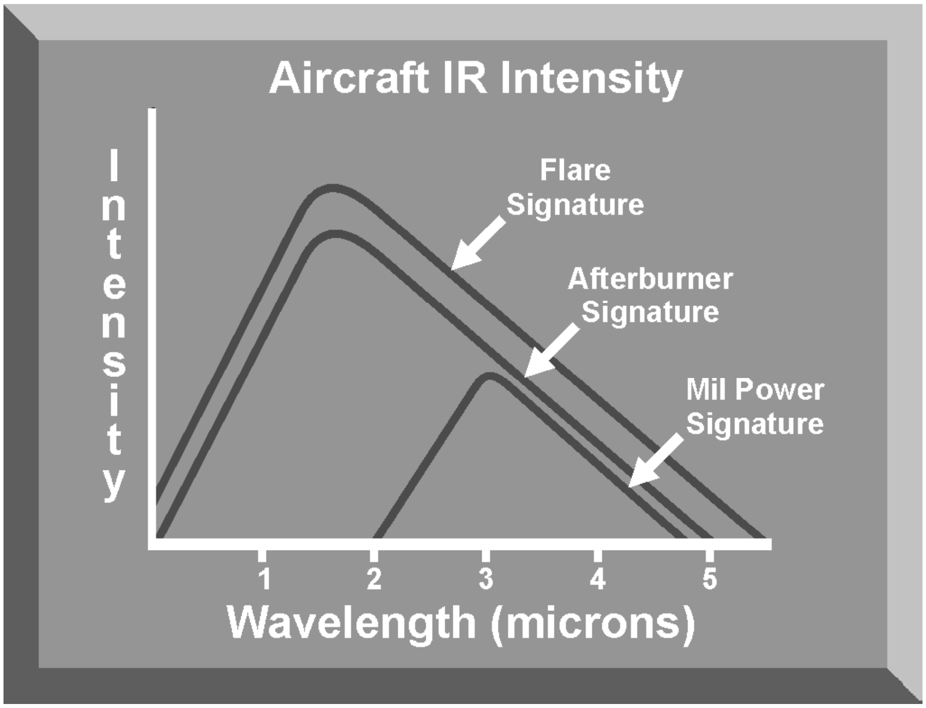

Figure 14-5. Aircraft IR Intensity

e. An aircraft's relative IR signature is based on aspect angle, airspeed, altitude, and afterburner status. The minimum relative IR intensity is at the nose and maximum at the tail. As airspeed increases, the relative IR intensity increases due to the heat generated by friction and increased engine temperatures. As altitude increases, the relative IR signature increases at all aspect angles, due to the reduction in atmospheric attenuation (Figure 14-6).

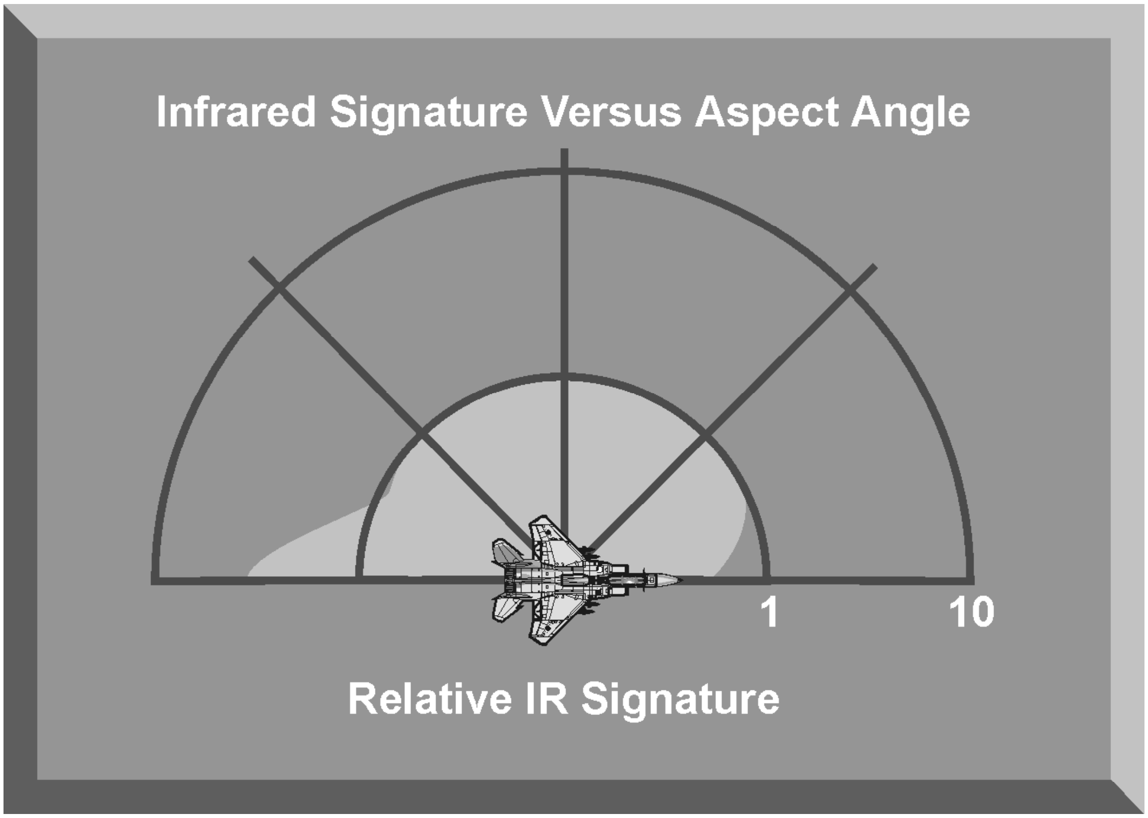

Figure 14-6. IR Signature Versus Aspect Angle

3. IR SIGNATURE SOURCES

IR guidance is based on the fact that every object with a temperature above zero degrees Kelvin emits IR radiation. The temperature of the object dictates the characteristics of this radiation. As the temperature of the object increases, the radiant intensity increases and shifts to higher and higher frequencies and correspondingly shorter and shorter wavelengths. The F-16 at Figure 14-7 demonstrates the different wavelengths found at different areas of the aircraft. Emissivity in Figure 14-7 is a relative measure of the IR energy emitted when a surface is directly viewed.

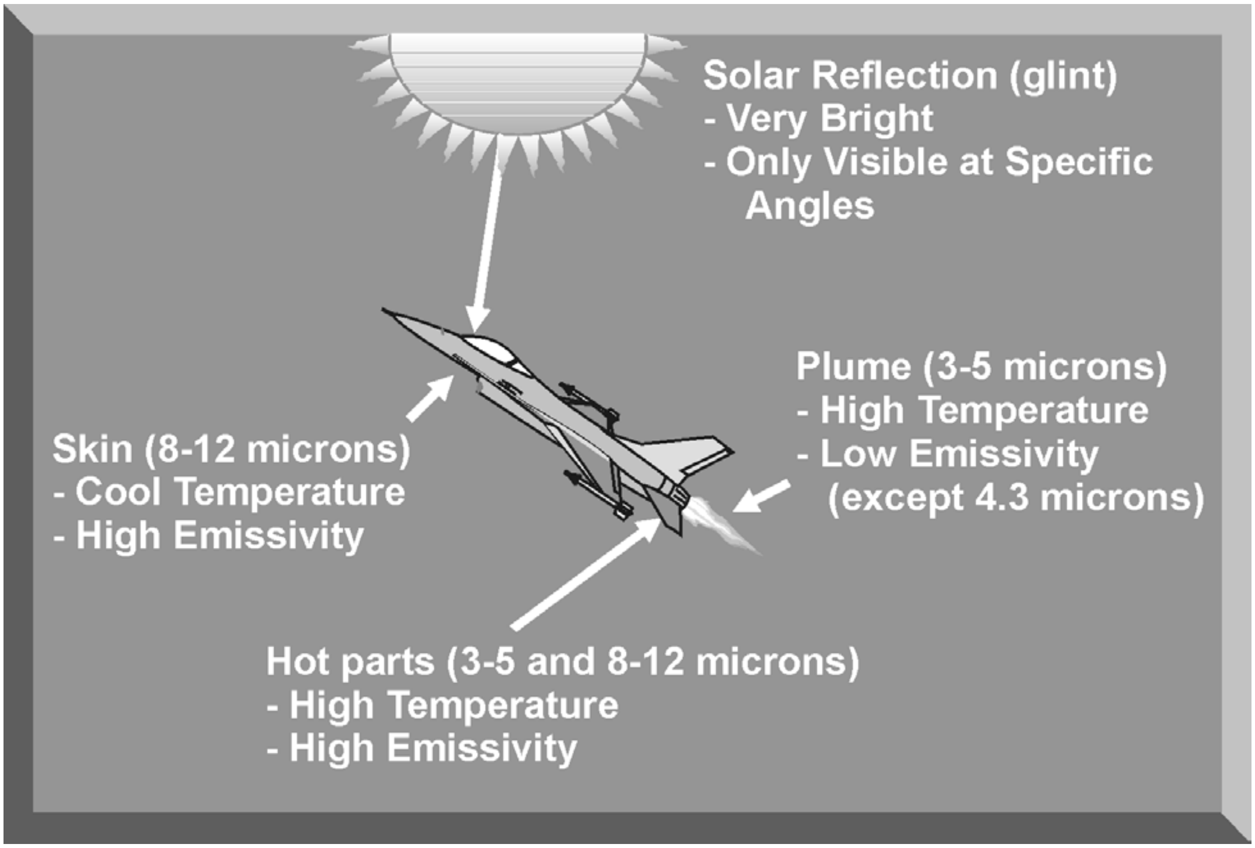
Solar Reflection (glint)
- Very Bright
- Only Visible at Specific Angles
Plume (3-5 microns)
- High Temperature
- Low Emissivity (except 4.3 microns)
Skin (8-12 microns)
- Cool Temperature
- High Emissivity
Hot parts (3-5 and 8-12 microns)
- High Temperature
- High Emissivity

Figure 14-7. Target Signature Sources

a. The main aircraft signature sources are the plume, the engine hot parts, and the skin. The skin is large in area, but usually fairly cool. Thus the best detection of the aircraft skin is usually in the long wave IR band (8-12 microns).

b. The engine hot parts offer excellent detection when one is looking at the right angle. They have high temperatures and high emissivity. However, depending on the viewing angle, they may have a low perceived area.

c. The plume has high temperatures and a high perceived area. This large perceived area allows near all-aspect detection. Unfortunately, it has a relatively low emissivity except near 4.2 microns. It is the strategy of the new breed of all-aspect missiles to detect the plume in the mid-wave IR, around 4.2 microns.

d. Figure 14-8 shows some of the common IR detector materials. Note that the cooled detectors are sensitive to longer wavelength (lower energy) photons, as discussed earlier.

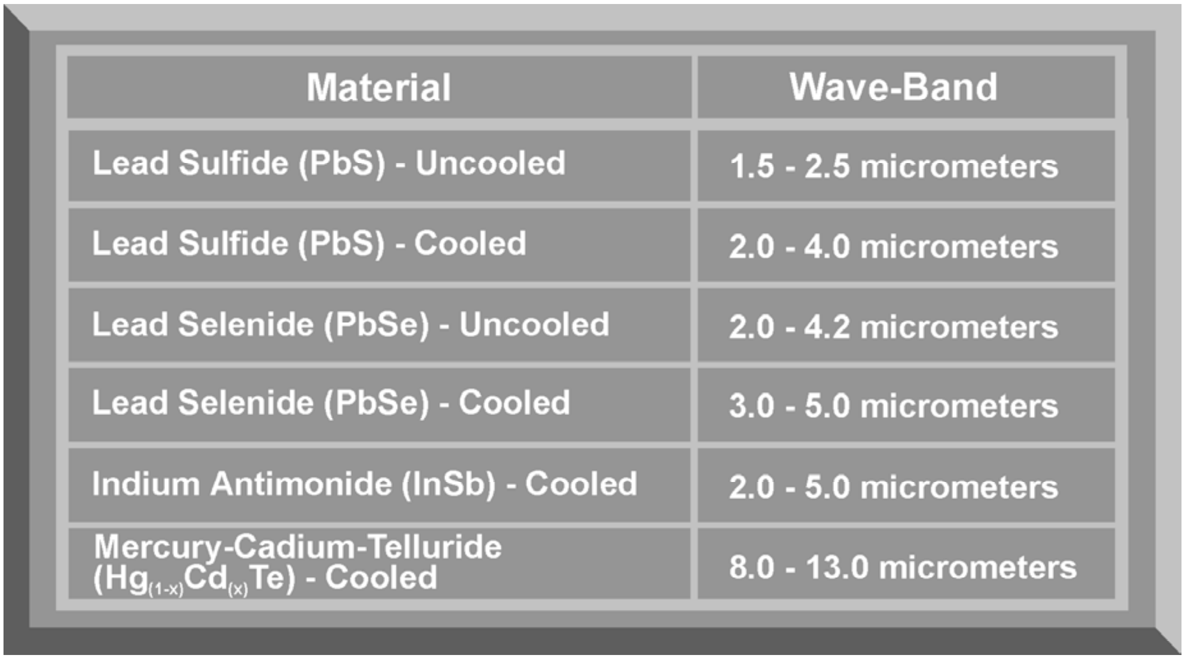

Material	Wave-Band
Lead Sulfide (PbS) - Uncooled	1.5 - 2.5 micrometers
Lead Sulfide (PbS) - Cooled	2.0 - 4.0 micrometers
Lead Selenide (PbSe) - Uncooled	2.0 - 4.2 micrometers
Lead Selenide (PbSe) - Cooled	3.0 - 5.0 micrometers
Indium Antimonide (InSb) - Cooled	2.0 - 5.0 micrometers
Mercury-Cadium-Telluride ($Hg_{(1-x)}Cd_{(x)}Te$) - Cooled	8.0 - 13.0 micrometers

Figure 14-8. IR Detection Materials

4. IR SEEKER CHARACTERISTICS

Guidance units are designed to detect and home in on the IR radiation of the aircraft. The job of the seeker is to view the scene and output the estimated target position. Figure 14-9 shows the hardware that makes up the seeker of a generic IR missile system.

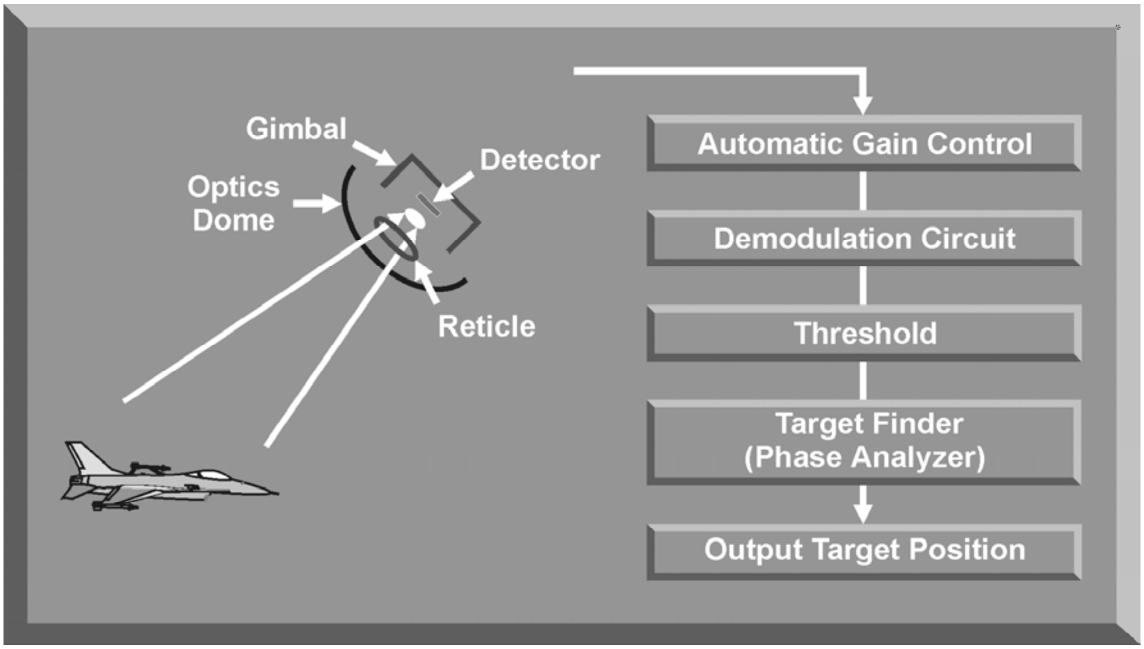
Gimbal
Optics Dome
Detector
Reticle
Automatic Gain Control
Demodulation Circuit
Threshold
Target Finder (Phase Analyzer)
Output Target Position

Figure 14-9. IR Missile Seeker

a. The dome acts as an IR transparent protective cover. Most IR missiles use the hemispherical dome because that shape does not alter the path of the incoming light rays. The optics focus the scene through the reticle onto the detector. The reticle pattern enables the detector output to "code" the position of the target.

b. The internal workings of the missile are similar to those found in a radar tracking system. There is an automatic gain control (AGC) to adjust the levels of the detector output so that it is not too large or too small. The demodulation circuit "decodes" the detector signal. The threshold circuit cancels signals that are below the specified threshold value, similar to the clutter rejection function found in radar systems. Finally, the phase analyzer reads the target position from the signal and sends the result to the autopilot. The autopilot then adjusts the missile's flight path to track the target.

c. IR systems employ filters and detectors to filter out unwanted IR radiation from the sun, sunlit clouds, smoke, the earth, and other background radiation sources (Figure 14-10). The detection unit is coupled with a guidance system to generate commands to the missile control vanes to keep the target centered in the field of view.

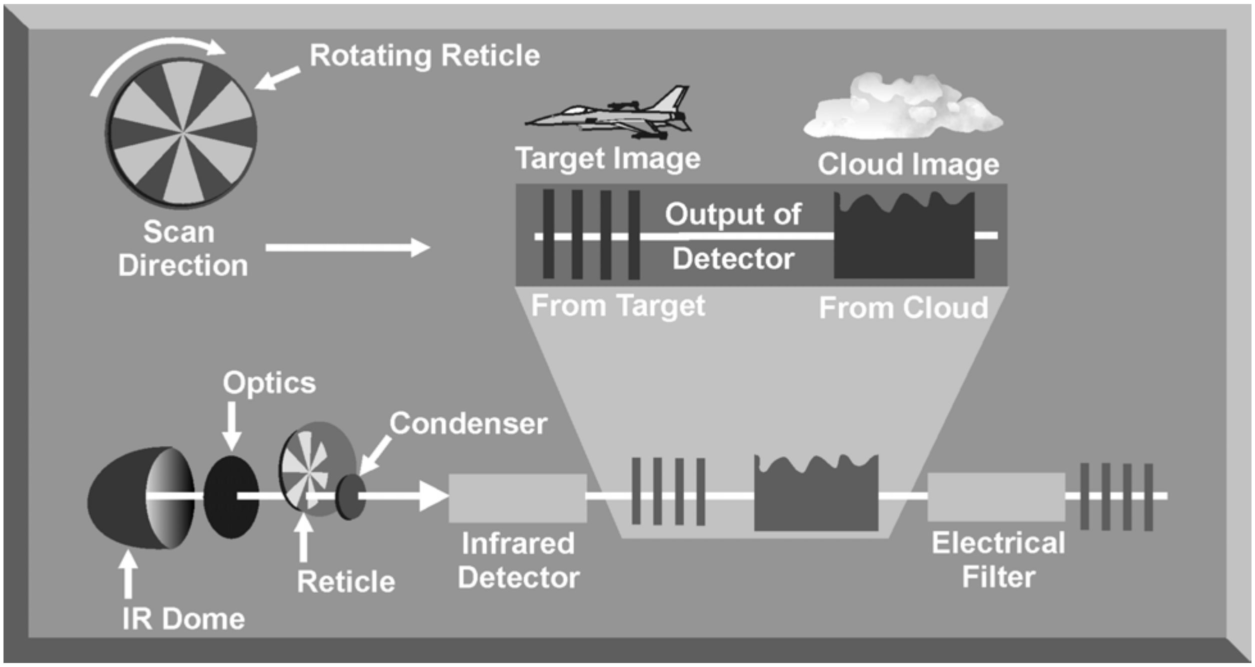

Figure 14-10. Spin Scan Reticle

d. A defining characteristic of an IR missile is its field of view (FOV) and field of regard (FOR) (Figure 14-11). The angular size of the image in degrees is called the field of view. To provide the greatest possibility of collecting IR radiation from the target, the receiver must have the greatest possible FOV. This in turn may

create problems. A large FOV increases the possibility that the receiver may not be able to distinguish the target from other sources of IR radiation. To avoid this problem, a relatively small FOV is scanned through a wider area. This wider area is referred to as the FOR. The gimbal enables the FOV to be scanned, so that an entire FOR can be searched.

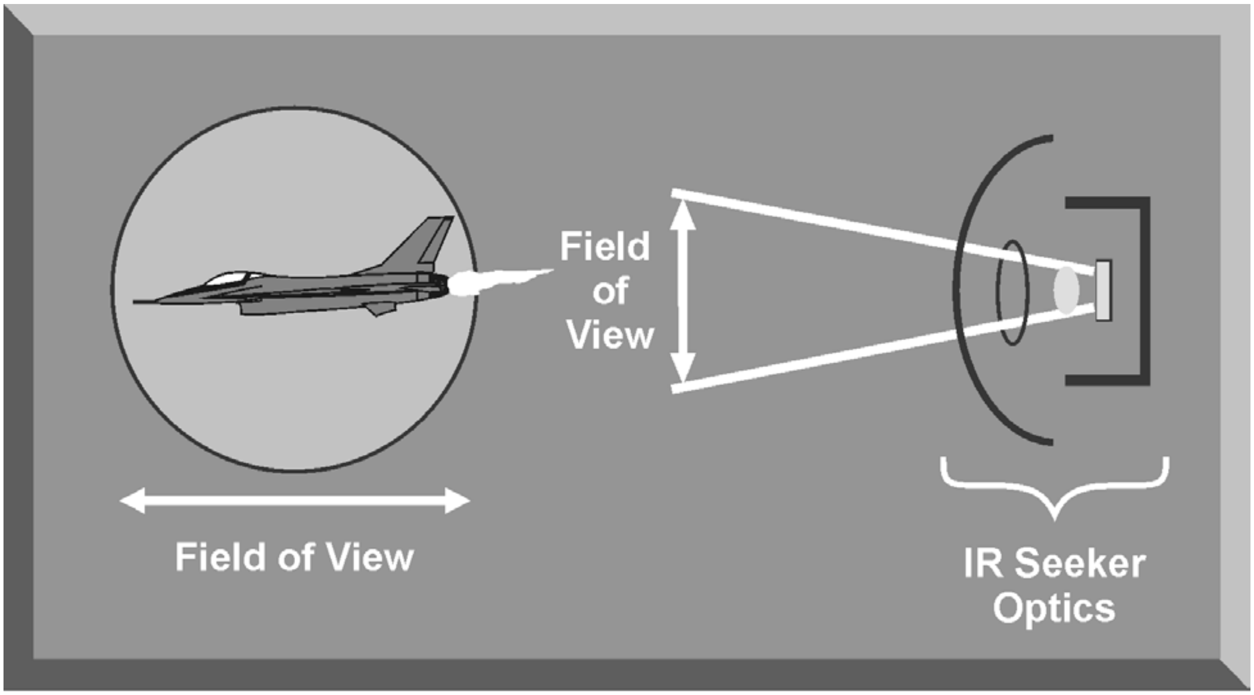

Figure 14-11. Missile Field of View

5. IR SEEKER TYPES

Through the years advances in seeker technology have resulted in significant changes to IR missile engagement tactics. This section will discuss spinning reticle, conical scan, cooled, and imaging IR seekers.

a. First generation IR missiles, like the SA-7, use a spinning reticle as the means to track the target. Due to their relatively low cost and ease of use, IR missiles of the first generation can still be encountered. The spinning reticle is inserted in the seeker just before the IR radiation reaches the detector. The reticle is a thin plate of optical material which has a transparent and opaque pattern on it. As the reticle is rotated, the IR energy is chopped at a rate determined by the reticle pattern. This system produces error signals when the target is not exactly centered in the field of view. Figure 14-12 is an example of a reticle pattern that can provide both azimuth and elevation information. If the target is located in the upper half of the pattern, the IR intensity on the detector is constant as the reticle rotates. As the pie-shaped half of the disc rotates over the target, the IR energy is pulsed and the amplitude of the pulses is an indication of relative elevation angle.

When the target moves to the right or left, the pulsing starts and stops at different times, indicating target azimuth. Center spun spin-scan seekers, also called center null reticles, are relatively insensitive when the target is in the center of the seeker scan where there is no tracking error. This is because the point target tends to bleed energy into all the spokes at once, eliminating the pulsed signal output of the detector. Once the target falls off the center of the reticle, the seeker generates an error signal that initiates guidance commands to recenter the target. This is the reason early IR missiles flew an undulating path toward the target.

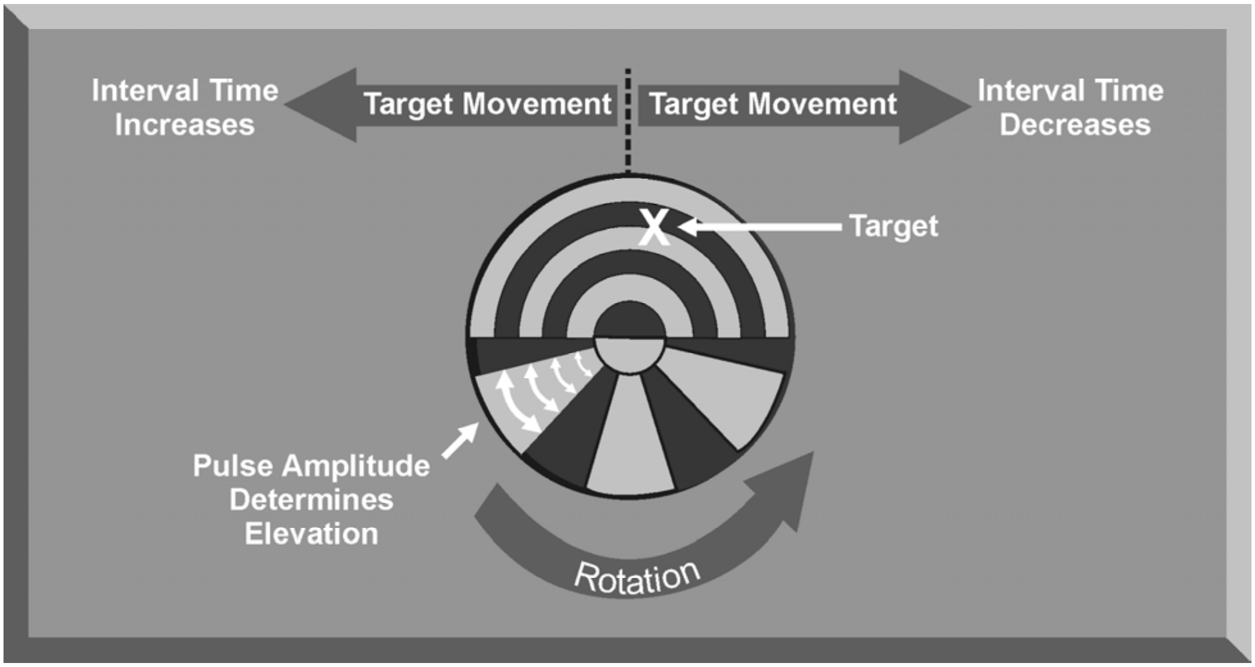

Figure 14-12. IR Missile Reticle

b. IR conical-scan (con-scan) seekers were developed to solve some of the problems with spin-scan seekers; notably the lack of error response when the target is near the center of the seeker field of view (FOV) (Figure 14-13). In a typical con-scan seeker, the reticle is fixed and does not spin. Instead, a secondary mirror is tilted and spun. This causes the target image to be scanned in a circular path around the outer edge of the reticle. When the target is centered in the seeker scan, the detector generates a pulsed output similar to that of the spin-scan seeker. However, as the target leaves the center, the output of the detector is a frequency modulated (FM) sine wave. The frequency of the modulation is directly proportional to the amount of target displacement from the center of the seeker scan. Con-scan optics are usually designed to spin the target very close to the edge of the reticle. This generates the greatest amount of FM modulation for a given target tracking error and gives the con-scan IR missile a more sensitive and tighter tracking loop. The center of the reticle is only used for acquisition.

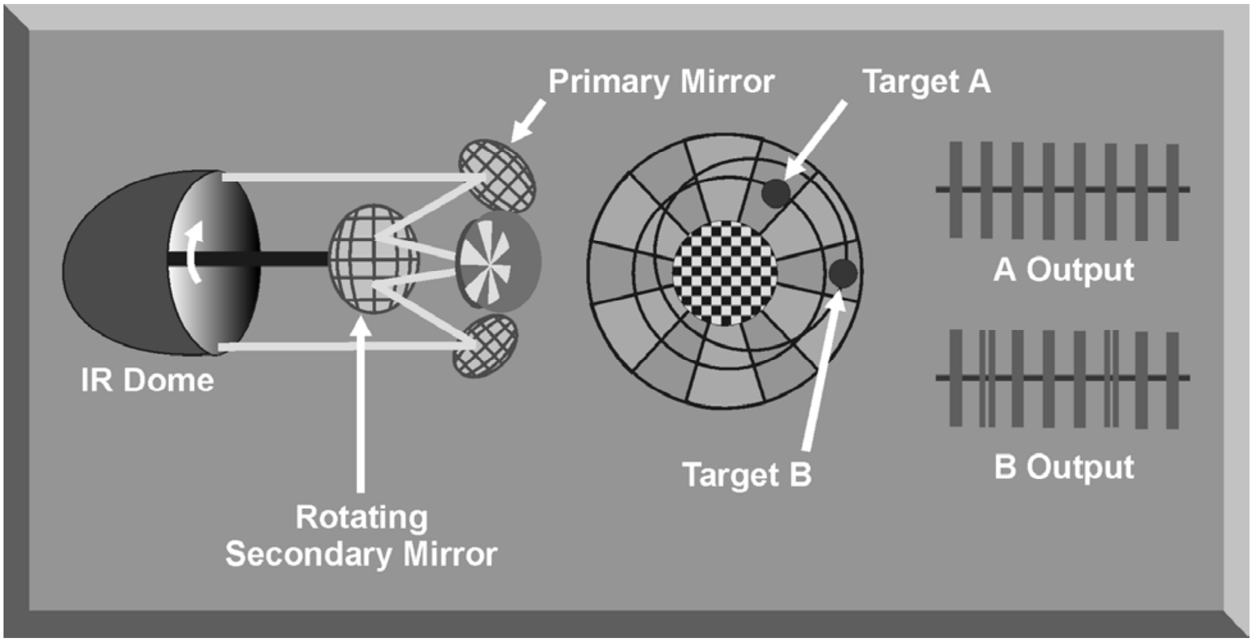

Figure 14-13. Conical Scan IR Seeker

c. A significant improvement to IR seekers resulted from the cooling of the detector with an inert gas such as argon. Older IR missiles, using uncooled lead-sulfide detectors, have a peak sensitivity in the 2 micron region. This limits these missiles, the SA-7 for example, to stern attacks because the missiles can only discriminate the IR signature of the engine turbine from background IR energy. By cooling the detector with an inert gas, like argon, the detectors of newer IR missiles can track longer wavelength IR radiation associated with airframe friction. Using newer detector materials like indium antimonide (InSb), require cooling to have increased target detection range and all-aspect tracking capability.

d. Imaging IR is the most recent advancement in IR seeker technology. The technology for these seekers is similar to that found in the AGM-65 Maverick missile. Imaging IR seekers are harder to decoy with flares than older seekers, and they are resistant to pulsed light jamming. Imaging detection involves creating an IR picture of the scene in one of two ways, scanning or staring. A scanning system uses one detector (or a mix of detectors and mirrors) which moves relative to the scene until the entire scene is scanned. This is an easy system to fabricate, but it can be noisy because the detector can't stay very long at each position, and it does not have a lot of time to measure the signal. A staring system uses many detectors, each of which detects a small portion of the scene. Each detector can “dwell” on its part of the scene for the entire frame time. However, such systems, also called focal plane arrays, are difficult to fabricate in a way such that each detector has the same sensitivity. One of the prime advantages with using imaging IR seekers is that they can be programmed

to track a particular IR shape or scene, significantly reducing the effectiveness of decoy flares.

6. IR MISSILE FLARE REJECTION

Flares are the primary countermeasure used to defeat the IR missile. Advanced IR missiles use different techniques to overcome the use of self-protection flares. There are two important characteristics of infrared (IR) missiles that influence the effectiveness of self-protection flares. The first is the ability of the IR missile seeker to discriminate between the IR signature of the aircraft and the IR signature of background interference, especially clouds. The second is the flare rejection capability built into the missile seeker and the missile guidance section.

a. The major source of IR background interference is sunlit clouds. Space filtering using a spinning reticle as described in the previous section is the method most widely used to suppress background interference and improve discrimination. IR energy from the target is gathered by the optical system and focused on the spinning reticle. The reticle chops the signal into a series of pulses that are focused onto the IR detector element. The signal output of a point source of IR energy, such as a target aircraft, will be a series of pulses. The signal output of a cloud, which covers several segments of the reticle, will be a single large pulse. An electrical filter eliminates the single large cloud pulse, and passes the multiple pulses generated by the target to the missile guidance section.

(1) IR missiles using spin-scan seekers have very little ability to reject flares. Flares provide the missile seeker with a hotter target than that of the aircraft, causing the spin-scan seeker to track the flare. The typical flare burns with a peak energy emission in the 2 micron range. Since the flare energy emission is greater than that of the target aircraft, the missile seeker transfers lock to the flare.

(2) Due to the scan pattern, con-scan seekers have some inherent resistance to flares. As described in the previous section, the reticle in a con-scan seeker is fixed and does not spin. Instead, a secondary mirror is tilted and spun. This causes the target image to be scanned in a circular path around the outer edge of the reticle. Because flares tend to drop away from the aircraft, they will drop off the con-scan reticle much faster than for a spin-scan reticle; therefore, flare resistance is built-in.

(3) Reducing the missile FOV is another method to help an IR missile discriminate the target IR signature from background IR (Figure 14-14). Limiting the FOV of the missile also makes the missile more resistant to flares. To decoy an IR missile seeker, the flare must create a heat source hotter than the aircraft and within the missile FOV. To decoy an IR missile with a narrow FOV, the flare must reach peak intensity almost immediately after ejection (Figure 14-15).

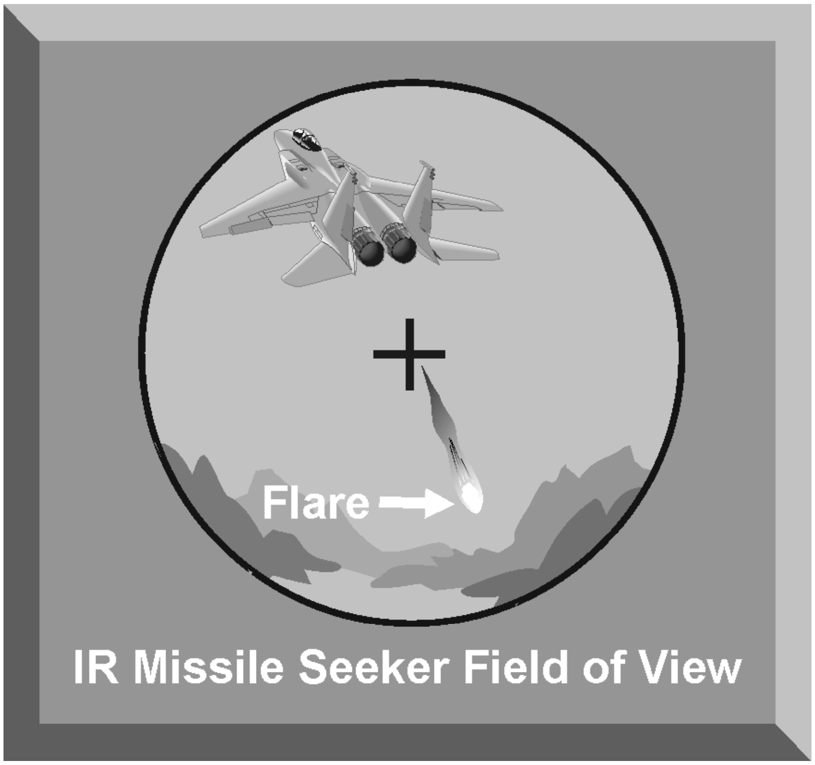
Flare
IR Missile Seeker Field of View

Figure 14-14. Impact of IR Missile Field of View

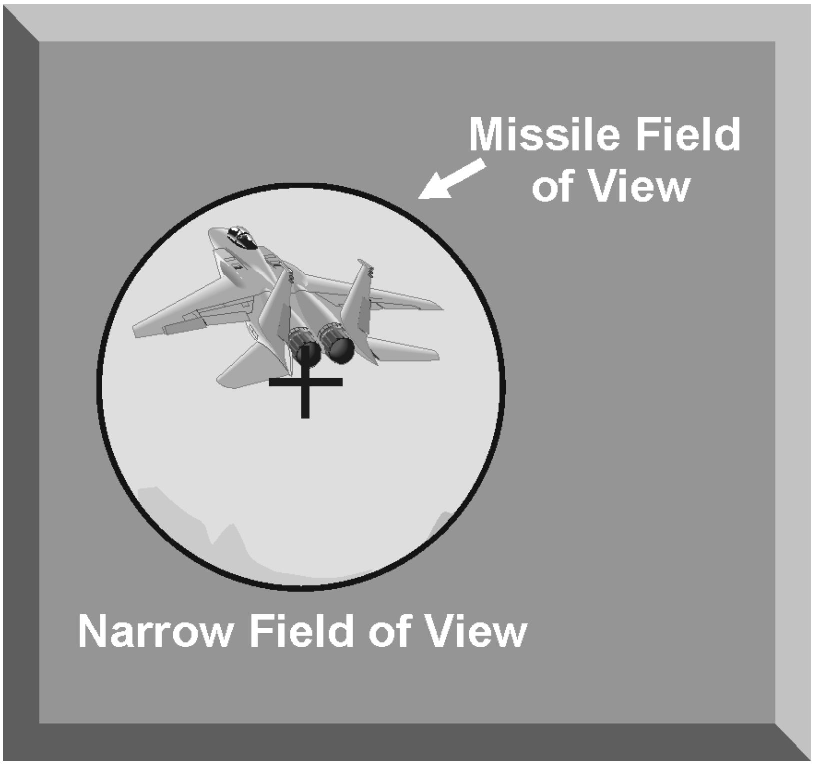
Missile Field of View
Narrow Field of View

Figure 14-15. IR Missile – Narrow Field of View

(4) The detector also influences the ability of an IR missile to discriminate target IR from background IR. Newer IR seekers can track longer wavelength IR radiation. The seeker's ability to track through a wider frequency range forces flares to decoy that wider frequency range to be successful. Additionally, the cooled seeker's ability to track from all angles could cause geometry problems, the missile may never see a flare ejected from the bottom rear of an aircraft.

b. IR missile designers have combined computer signal processing and modern IR missile seekers to detect the presence of flares and reject them as targets. This flare rejection capability, or IR counter-countermeasure (IRCCM), allows newer IR missiles to track the aircraft IR signature while rejecting multiple flares. Flare rejection is based on two computer functions. The first is called the "trigger," which detects the flare in the seeker field of view. The "trigger" function activates the "response" computer processing, the action the seeker takes to reject the flare. Both the "trigger" and "response" must operate to successfully reject flares. Each advanced IR missile employs different computer techniques for both the "trigger" and "response." A flare technique developed to counter one IR missile may not work against another IR missile that uses different flare rejection computer techniques.

(1) There are several "trigger" techniques that can be used by IR missiles to detect flares. Advanced IR missiles may employ one or more of these techniques to detect the presence of a flare in the seeker FOV. These "trigger" techniques include rise time, two-color, kinematic, and spatial.

(a) An IR missile using a rise time "trigger" monitors the IR energy level of the target. A sharp rise in the received IR energy within a specified time limit indicates a flare in the IR seeker FOV. When the missile detects this rapid rise in IR energy, the rise-time "trigger" triggers a flare "response." The "response" is switched off when the received IR energy drops to its original level. The thresholds for the rise in IR energy and the response time are set to preclude activating the rise time "trigger" when the aircraft selects afterburner. An IR missile using the rise time "trigger" can be decoyed by multiple flares with slow rise times.

(b) IR missiles using a two-color "trigger" to detect flares sample the energy level in two different wavelength bands. In Figure 14-16, a non-afterburning target would have more IR intensity in band B than in band A. The typical flare produces more IR intensity in band A than band B. A sudden increase in band A intensity compared to band B intensity indicates a flare in the seeker FOV. The two-color "trigger" would then trigger a flare "response." Advanced IR missiles using a two-color "trigger" can employ different detectors composed of different materials to monitor the intensity in two bands. A lead sulfide detector could be used for band A and indium antimonide for band B. IR missiles using a single detector can employ a reticle with different bandpass filters to monitor IR intensity in both bands. To track the target, IR missiles employing a two-color "trigger" may use either band, or use data from both

bands of a dual-channel tracker. An IR missile using the two-color "trigger" can be decoyed by multiple flares that provide equal IR intensity in each band.

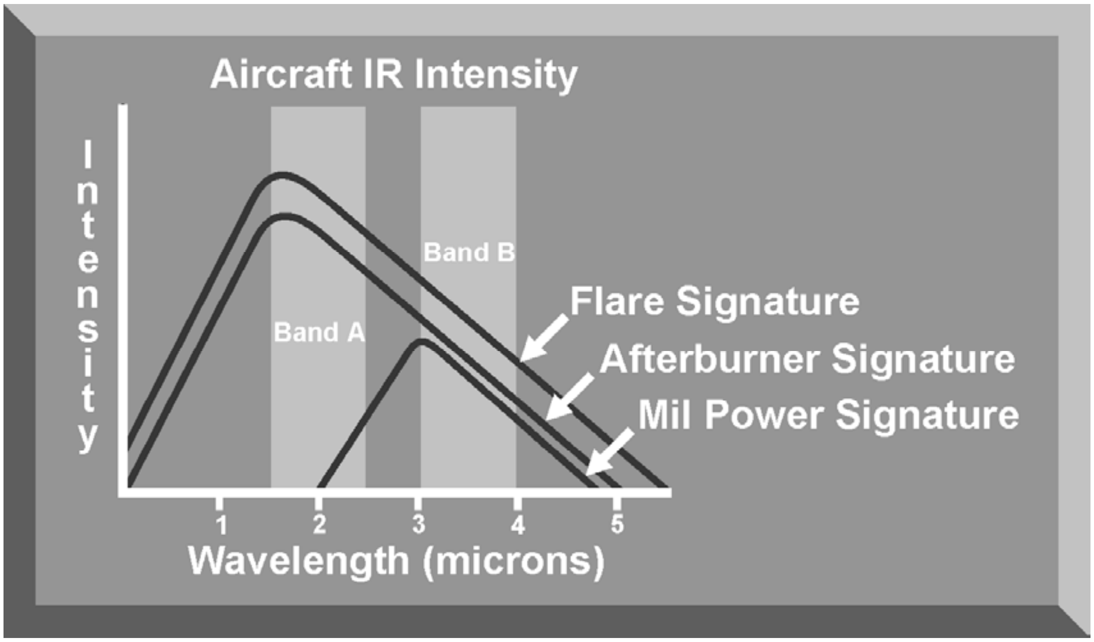

Figure 14-16. Two-Color IR "Trigger"

(c) A kinematic "trigger" takes advantage of the fact that flares separate very quickly from the dispensing aircraft due to aerodynamic drag. In a beam aspect engagement, an IR seeker that transfers track from the target to the flare will have a dramatic change in the line-of-sight rate due to the rapid separation of the flare from the aircraft. An IR missile employing a kinematic "trigger" detects this change and initiates the "response." IR missiles employing a kinematic "trigger" may have difficulty in flare detection in a head-on or stern engagement due to the small line-of-sight change between target and flare. Multiple flares dispensed at very short intervals will probably decoy an IR missile employing a kinematic "trigger".

(d) The spatial "trigger" operates like the kinematic "trigger" in that it uses the rapid separation of the flare from the aircraft to trigger the "response." An IR missile employing a spatial "trigger" uses the seeker to detect a flare. When the flare separates to the rear of the aircraft, the seeker will see the target on the edge of the FOV corresponding to the direction of target movement. The flare will be on the opposite side of the FOV. Once two hot objects on opposite sides of the FOV are distinguishable, the spatial "trigger" triggers the flare "response." IR missiles employing a spatial "trigger" can be decoyed by dispensing multiple flares at very short intervals.

(2) The seeker's "response" to the "trigger" is to reject the flare or limit its effect on target track. As long as the flare remains in the seeker FOV, the missile

is tracking the target in a degraded mode. Most IR seekers have a FOV of less than 2.5°. At long ranges, the flare will remain in the FOV for a relatively long period of time. At close range, the flare will be in the missile FOV for a relatively short period of time. Several different "response" techniques may be used, either alone or in combinations, to defeat a flare. These "response" techniques include simple memory, seeker push-ahead, seeker push-pull, and sector attenuation.

(a) When the simple memory "response" is initiated, the missile continues the maneuver it was performing just before the "trigger." This "response" assumes the flare will separate to the rear of the target. The missile rejects the seeker track data and maintains its motion relative to the target, waiting for the flare to leave the seeker FOV. The missile will continue to reject track data until the flare leaves the FOV or until the "trigger" times out. When the "trigger" times out, the "response" is discarded and the seeker operates in the normal track mode. If the "trigger" times out while a flare remains in the FOV, the seeker will usually transfer lock to the flare.

(b) The seeker push-ahead "response" causes the seeker gimbals to drive the seeker forward in the direction the target is moving (Figure 14-17). Pushing the seeker forward causes the flare to depart the FOV faster than with simple memory, minimizing the amount of time the missile is not tracking the target. The greater the amount of forward movement (called "bias"), the faster the flare will depart the FOV. If the amount of forward bias applied is too great, the seeker may be pushed forward of the target. This could cause both the target and the flare to depart the FOV and the missile would have to reacquire the target.

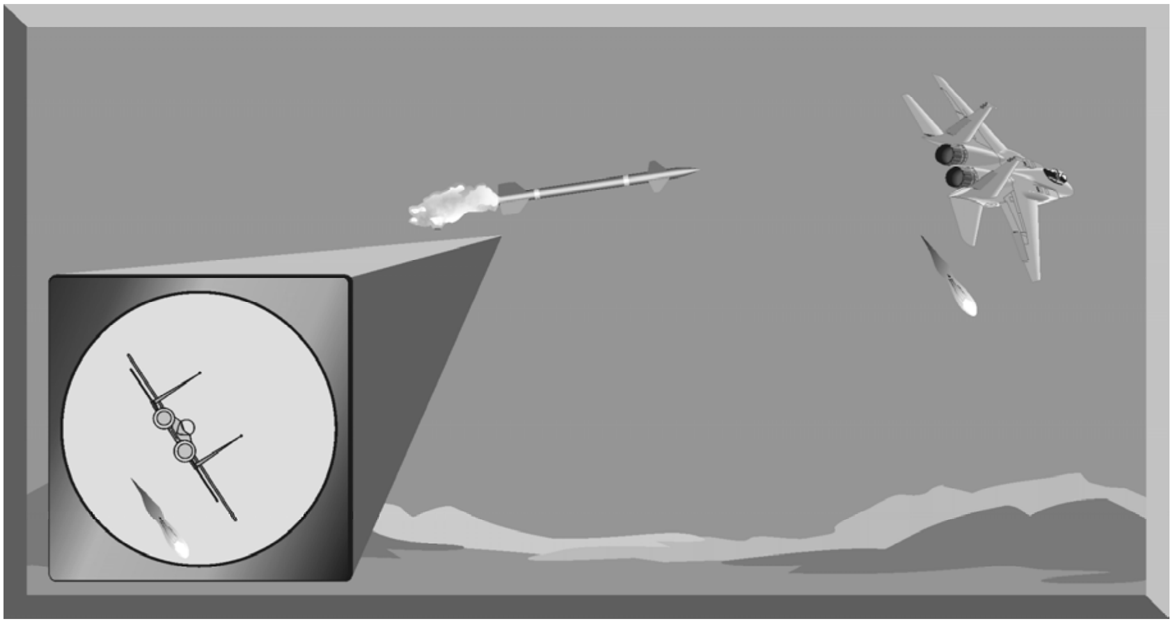

Figure 14-17. Seeker Push-Ahead "Response" Technique

(c) The seeker push-pull "response" assumes flares will have a higher intensity IR signature than the target. The "response" is initiated when the target and flare are on opposite sides of the seeker FOV (Figure 14-18). This corresponds to a spatial "trigger" condition. The received energy will rise and fall as the energy of the target and flare is scanned across the detector. When the flare energy is at a peak, the seeker gimbals drive the seeker away from the flare. When the lesser energy from the target is detected, the seeker's gimbals pull the seeker in the direction of the target. As a result, the seeker is moved away from the flare and toward the coolest IR source in the FOV, the target aircraft.

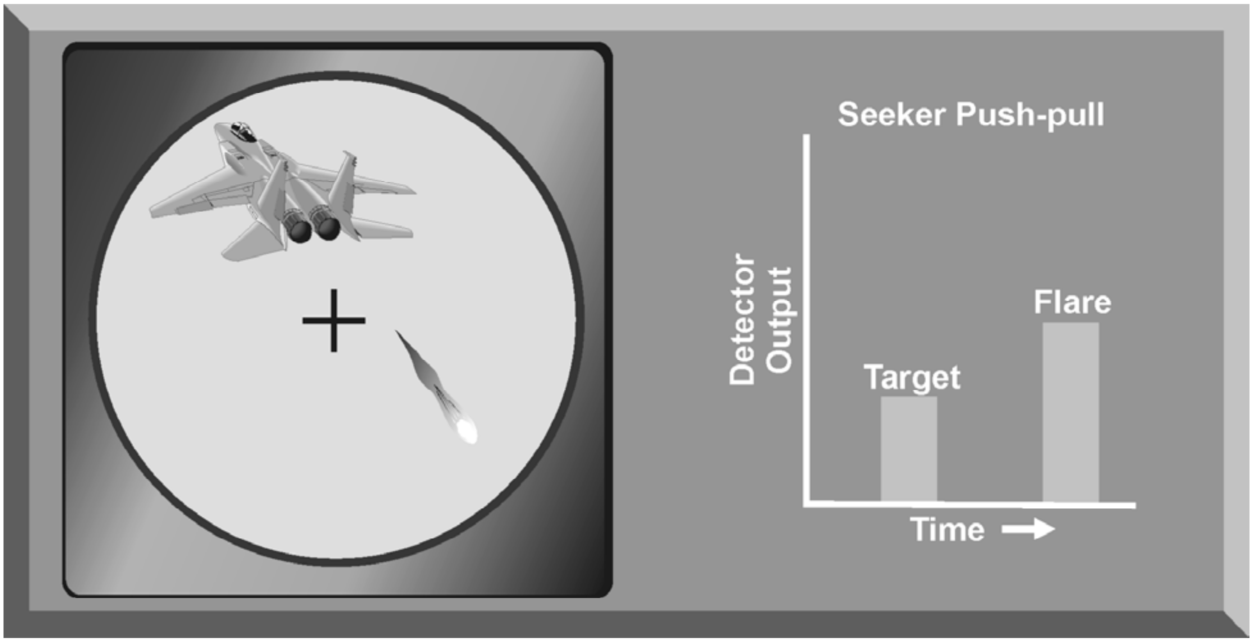

Figure 14-18. Seeker Push-Pull "Response" Technique

(d) The sector attenuation "response" is initiated by placing an attenuation filter across part of the seeker FOV (Figure 14-19). This filter reduces the seeker sensitivity in that part of the FOV. If the target being tracked is in the center of the FOV, then placing an attenuator in the quadrant below and to the rear of the target should reduce any energy received from a flare. If the attenuated flare energy is below that of the target energy, the seeker will continue to track the target.

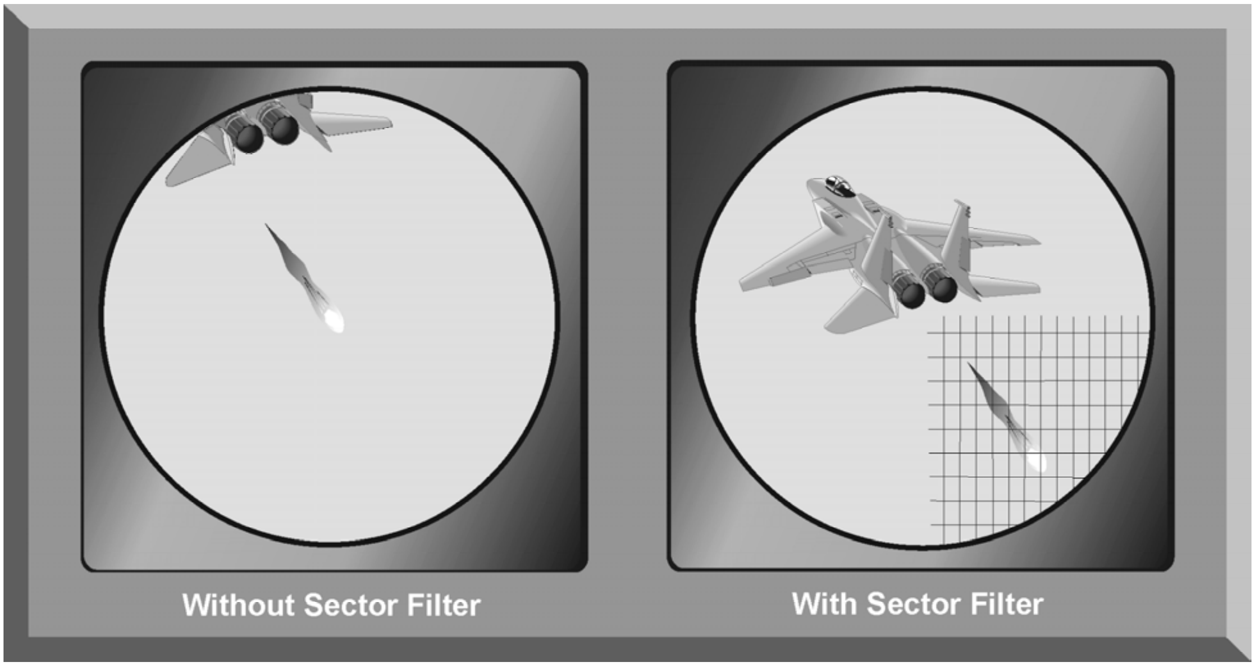

Figure 14-19. Sector Attenuation “Response” Technique

7. SUMMARY

This chapter provided some background on IR theory and IR detection, then went into some of the basics about the different types of IR missile seekers. Finally factors that impact the ability of an IR missile to reduce the effectiveness of self-protection flares were discussed. This ability depends on the discrimination capabilities of the missile seeker, the type of detector, the missile FOV, and the missile flare rejection capabilities. Modern IR missiles that employ sophisticated flare rejection techniques and advanced missile seeker technologies present a growing, and potentially lethal, threat.

CHAPTER 15. IR COUNTERMEASURES

1. INTRODUCTION

Defeating IR missiles used to be an afterthought placed well behind surviving the radar missile threat. Now with the proliferation of advanced IR missiles, defeating the IR threat is becoming more important and more difficult. The primary countermeasure used to defeat IR missiles is the self-protection flare. This chapter will discuss flare characteristics and employment considerations. One of the main difficulties in defeating an IR missile is knowing that one has been launched. This chapter will also discuss some of the methods currently available to detect an IR missile attack.

2. FLARE CHARACTERISTICS

Self-protection flares were developed to counter threat systems operating in the IR spectrum. Self-protection chaff and flare dispensers, such as the ALE-40, ALE-45, or the ALE-47, are designed to allow the pilot to dispense flare cartridges when engaged by an IR threat. These flare cartridges are pyrotechnic and pyrophoric devices designed to produce an IR source that is more attractive than the IR signature of the aircraft. To decoy an IR missile seeker, the flare must create a heat source more attractive than the aircraft, within the missile field of view. The most important flare characteristics that determine the ability of a flare to decoy an IR missile are IR wavelength matching, flare rise time, and flare burn time. The MJU-7 flare cartridge will be used as an example of a typical flare cartridge.

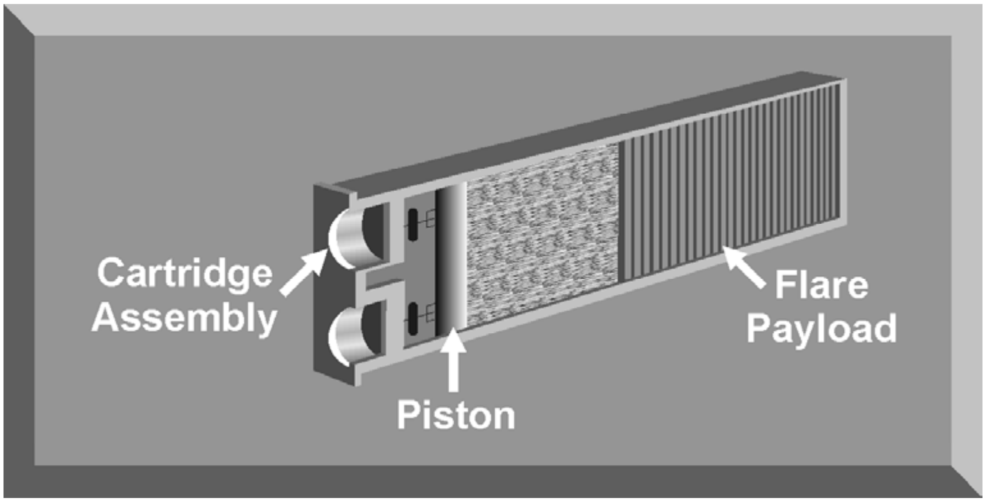

Figure 15-1. The MJU-7 Flare

a. The MJU-7 flare cartridge (Figure 15-1) is an example of a flare cartridge used in both the ALE-40 and the ALE-47 dispensers. The flare grain is composed of magnesium and tetraflouroethylene, or C_2F_4, which burns at 2000 to 2200°K. As the flare burns, it emits IR energy of different wavelengths from the luminous zone that emulates the aircraft IR signature. The burning flare also produces a large quantity of white smoke, which may highlight the position of the dispensing aircraft.

b. A flare must reach peak intensity shortly after ejection or it will not be effective in decoying the IR missile seeker. Flare rise time is the time required for the flare to reach peak intensity.

(1) To counter a short-range IR missile with a narrow field of view, a flare must reach peak intensity quickly. On the other hand, some advanced IR missiles now look for a rapid rise in IR energy as a trigger to know when a flare is trying to decoy it (Figure 15-2).

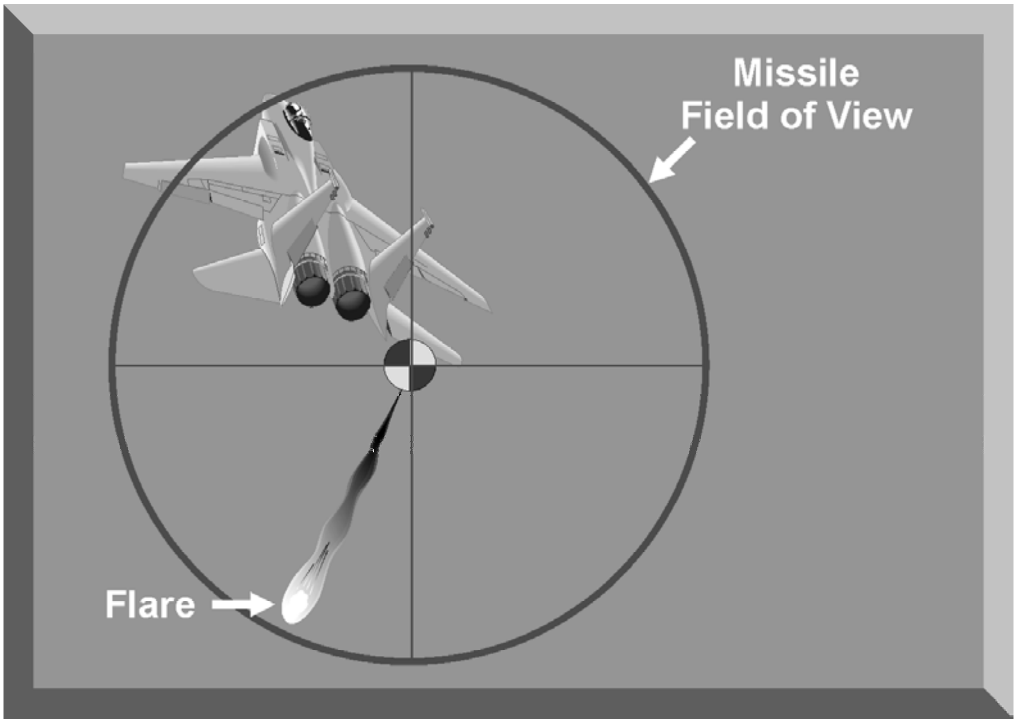

Figure 15-2. Impact of Flare Rise Time

(2) Flare rise time varies dramatically with altitude. The flare burns longer at high altitude, but it takes much longer to reach peak intensity. This increase in flare rise time at high altitude can impact the effectiveness of flares to decoy and defeat IR missiles particularly in the air-to-air combat environment.

c. Flare burn time is the time span that the flare burns and determines how far the IR seeker will be pulled off the target. The longer the burn time, the longer the

IR seeker is pulled off the target aircraft. The flare should burn long enough to ensure that the aircraft is no longer in the missile field of view. Flare burn time, like flare rise time, varies with altitude. A flare will burn longer, but at lower intensity, at higher altitudes. A longer flare burn time increases the probability the IR missile will be decoyed by a single flare.

3. ADVANCED FLARES

To counter the advances in IR missile seeker technology the Air Force and Navy formed the Advanced Strategic and Tactical Expendables program to develop and field advanced IR decoys. Two products to come out of this program are the kinematic flare, MJU-47, and the covert flare, MJU-50/51.

a. Kinematic Flares. A significant characteristic of conventional flares is that upon ejection they rapidly slow down and separate from the aircraft. As described in Chapter 14, modern IR missiles exploit this rapid separation between the aircraft and the flare. The rapid separation triggers a flare rejection response in the missile seeker causing it to ignore the flare and continue tracking the aircraft. The kinematic, or thrusted, flare delays the missile response by propelling itself in the direction of the aircraft and negating or delaying the missile's flare rejection trigger. The MJU-47 is the same size as the currently fielded MJU-10 flares. The flare's decoying pyrotechnics are vectored out the end of the flare to provide a means of propulsion (Figure 15-3).

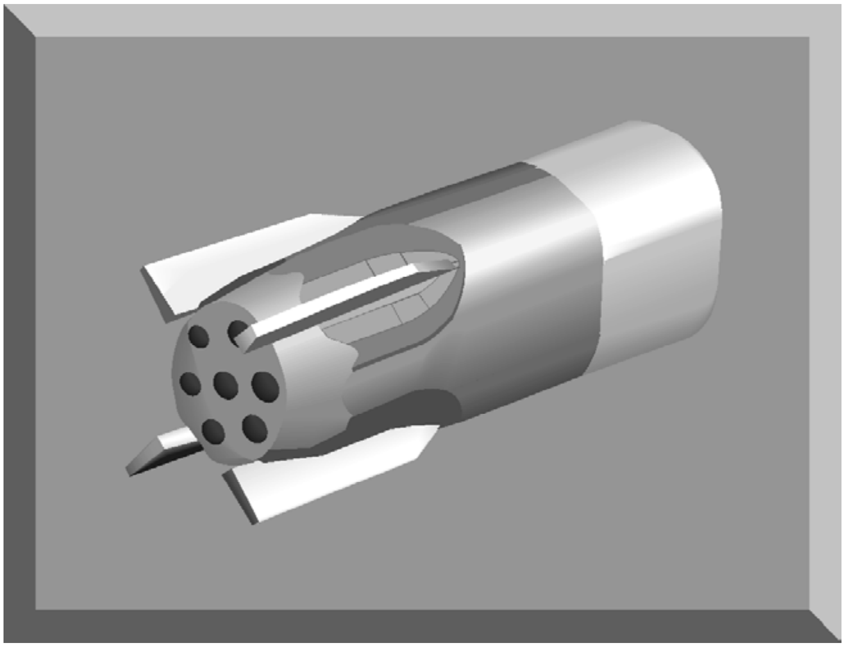

Figure 15-3. Kinematic Flare

b. Covert Flares. Though an effective tactic at defeating MANPADS, pre-emptively dropping flares in a target area carries a significant risk. The risk is that enemy air defense systems that may not have known the aircraft's location will surely see the flares, day or night, and any chance for surprise is lost. Figure 15-4 shows the visual signature of conventional flares. The danger of highlighting oneself becomes even more likely as aircraft become fitted with chaff and flare systems that automatically dispense expendables based upon inputs from the RWR or a missile attack warning system (MAWS). Automatic systems tend to error on the side of caution which means expendables will be dropped if there is any ambiguity. Visually covert flares, named the MJU-50 and MJU-51, remove this problem by not leaving a visual signature such as smoke or flame. These covert flares are made of material that oxidizes, pyrophoric instead of pyrotechnic, when released in the air. The MJU-50 is the size of the small M-206 flare while the MJU-51 is the size of the MJU-7 flare.

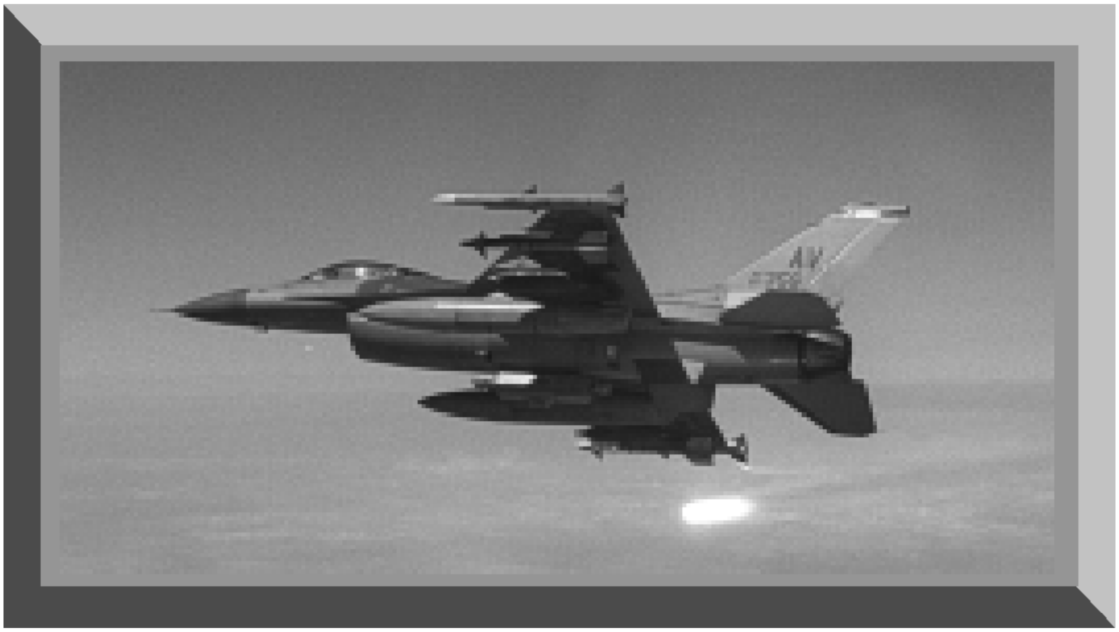

Figure 15-4. Flare Visual Signature

4. FLARE EMPLOYMENT

The purpose of employing a self-protection flare cartridge is to decoy the seeker head of an IR missile. This is accomplished by presenting the IR missile with a second heat source with an IR signature that exceeds the aircraft signature. The flare or IR source must appear in the field of view of the IR missile at the same time as the aircraft. As the flare separates from the aircraft, the IR missile seeker tracks the most intense IR signature, which ideally is the flare, and is decoyed away from the aircraft.

a. It is important to perform maneuvers along with flare employment to effectively defeat IR missiles. Maneuvers compound the IR missile's tracking problems and increase the distance between the aircraft and flare. Maneuvers that put the aircraft outside the IR missile field of view increase the ability of the flare to decoy the missile. When engaged by an IR threat, one tactic is to immediately dispense a flare and maneuver to put the missile on the beam. This increases the line-of-sight rate the missile guidance system must process and gives maximum separation between the aircraft and the flare (Figure 15-5).

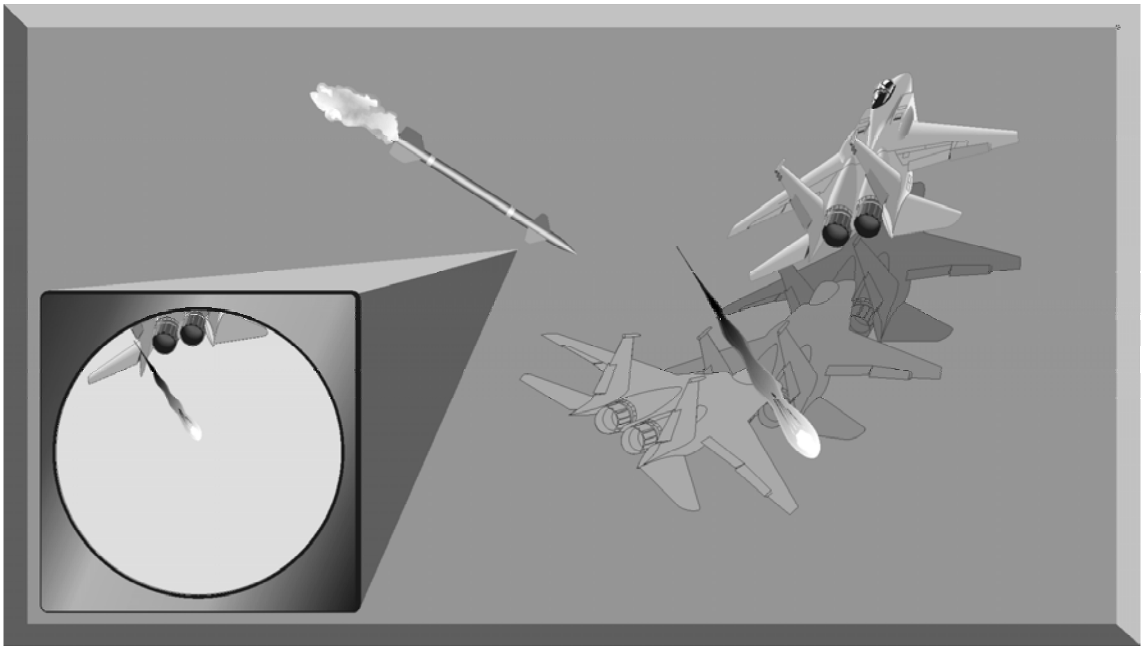

Figure 15-5. Initial Maneuver and Flare Employment

b. With the IR threat on the beam, the pilot has the best chance for achieving a "tally-ho" on the missile to determine range and keep the missile on the beam. Visually acquiring the missile increases the chances of surviving the encounter. Modern chaff and flare dispensers, such as the ALE-47 and ALQ-213, can be programmed to dispense flares in a sequence optimized to defeat specific IR missiles. Programs are selected by the pilot based upon the most likely threat. Classified tactics manuals provide aircraft specific maneuvers and flare dispensing programs to defeat the IR threat.

5. IRCM TACTICS

Besides maneuvers and flare employment, there are other IR countermeasures that can reduce the effectiveness of IR threat systems. The first is to reduce the intensity of the heat signature of the aircraft by reducing the power setting. The engines produce the largest IR signature and are the only source of IR radiation that the pilot can influence. By reducing the power setting, the IR signature is

reduced and the required flare intensity is thereby reduced. Use of afterburner should be minimized for exactly the same reason; the required flare intensity necessary to cover an aircraft in afterburner is difficult to obtain. The pilot should not reduce the power setting below the minimum required to maintain sufficient maneuvering airspeed.

a. Another effective IRCM is to use the sun. Maneuvering into the sun masks the aircraft's IR signature from most IR threat systems. When attacking a target defended by IR systems, an attack axis that places the sun behind the attacker may limit the effectiveness of these systems.

b. Clouds or smoke can also confuse IR threat systems (Figure 15-6). The water vapor making up clouds diffuses the IR energy making it difficult for an IR system to get a point to track, similar to how sun's light is diffused on a cloudy day. The particles found in smoke can have a similar effect on IR energy. The final IRCM is to reduce the IR signature of the aircraft by careful design of the engines and exhaust system. For example, the F-117 was specifically designed to provide the lowest possible IR signature.

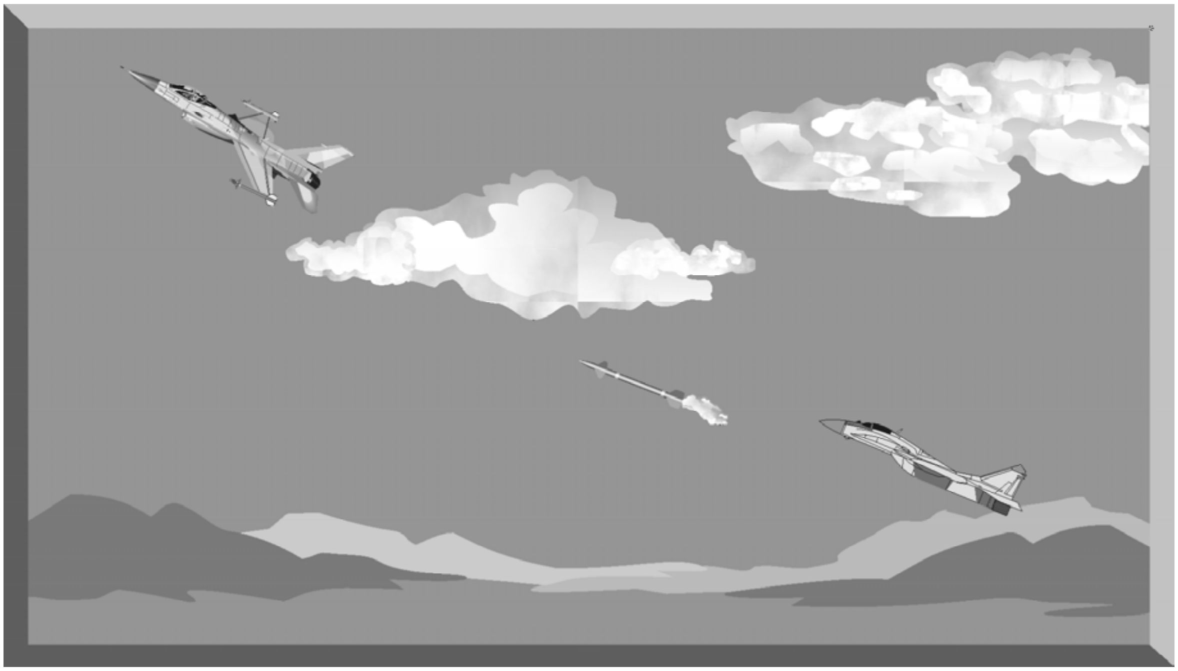

Figure 15-6. IRCM

6. MISSILE APPROACH WARNING SYSTEMS (MAWS)

One of the most important factors in defeating an IR missile attack is knowing that an attack is in progress. Since IR threat systems are generally passive, the radar warning receiver (RWR) will provide no attack warning unless the threat uses some detectable radar energy for acquisition prior to launching an IR

missile. MAWS are designed to provide the crew some warning of an ongoing missile attack. These systems have been widely deployed on larger aircraft. Advances in technology allowing them to be smaller and promising lower false alarm rates have increased the possibility that fighter size aircraft may be outfitted with this equipment.

a. The two primary requirements for MAWS are timeliness and reliability. An IR missile engagement is an extremely short event providing very little time for the targeted aircraft to respond with maneuvers and flares. Secondly, there must be a very low false alarm rate. In earlier systems the sun, flares, or the wingman's aircraft have been known to cause false alarms. Aside from being annoying, a false alarm can become a tactical problem if the MAWS is tied directly into the countermeasures dispenser set allowing for automatic dispensing of chaff and flares when the MAWS senses an attack. False alarms in this type of scenario would only serve to highlight a previously untargeted aircraft.

b. MAWS detect incoming missiles either actively or passively. The active MAWS use a pulse Doppler radar to detect and track the in-flight approach of an attacking missile. Pulse Doppler radar is used for this purpose because of its ability to use relative velocity to pull targets out of clutter. An incoming missile will have a high velocity relative to the surrounding background. Range and time to impact are computed automatically, updated continuously, and provided to the crew to assist in defensive maneuvers (Figure 15-7). The down side to active MAWS is the requirement to continuously radiate for long periods of time to ensure protection. This contradicts the common tactic of keeping radar transmissions to a minimum to avoid being tracked via passive detection measures. This situation has led to the latest generation of MAWS being mostly of the passive type.

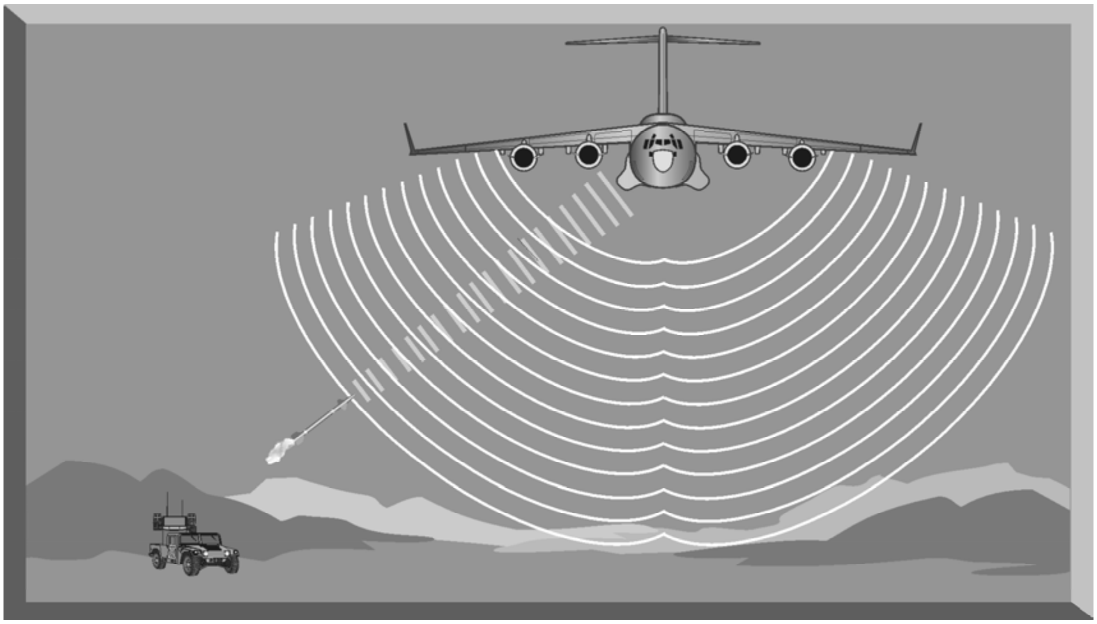

Figure 15-7. Missile Alert Warning System (MAWS)

c. The passive MAWS operate in either the ultraviolet, or the infrared frequency spectrum. These systems are tuned to look for the IR signature of a missile's rocket motor, then pass a warning and the position of the inbound threat to the pilot. To obtain all-aspect missile warning, multiple MAWS sensors must be positioned on the aircraft in a similar fashion to RWR antennas. All-aspect MAWS face the same challenges that RWR systems face: false multiple targets due to maneuvers and accurate threat position reporting, to name a few.

7. SUMMARY

The proliferation of IR threat systems has elevated the importance of IRCM to survival on the modern battlefield. The most effective IRCM is still the employment of flares in conjunction with maneuvers. The specific flare characteristics of IR spectrum coverage, rise time, and burn time are critical factors in determining flare effectiveness. Other IRCM tactics are designed to enhance the effectiveness of flares and take advantage of IR missile limitations. The effectiveness of all IRCM tactics depends on some type of attack warning. The new generation of MAWS should provide some measure of warning to the crews and in some cases automatically defeat the IR missile.

CHAPTER 16. RADAR ELECTRONIC PROTECTION (EP) TECHNIQUES

1. INTRODUCTION

Electronic warfare (EW) is defined as military action involving the use of electromagnetic and directed energy to control the electromagnetic spectrum or to attack the enemy. Nearly every military action, from command and control of an entire integrated air defense system (IADS) to precision guidance of an individual weapon, depends on effective use of the electromagnetic spectrum. Radar systems have become a vital element of nearly every military operation. Since these systems operate across the entire electromagnetic spectrum, much of the EW effort is concerned with countering radar systems. All of the jamming techniques discussed in Chapters 10 and 11 and the chaff employment options discussed in Chapter 13 are specifically designed to counter radar systems. These actions are classified as electronic attack (EA), which is a part of EW.

a. EW is somewhat like a chess game—a series of moves and countermoves within the electromagnetic spectrum. As we develop jamming techniques to counter radar systems, our adversaries develop counter-countermeasures to negate the effectiveness of these techniques. In response, we develop newer techniques and our adversaries respond with new modifications to their radar systems. This series of moves and countermoves can continue for decades. The development and application of radar counter-countermeasures are classified as electronic protection (EP), also a part of EW.

b. The continuing battle to control the electromagnetic spectrum for unrestricted radar employment has resulted in over 150 radar EP techniques. These techniques are designed to negate the effectiveness of electronic jamming and chaff on radar systems. These radar EP techniques can be incorporated into the design of a radar system or added to an existing radar system in response to a jamming technique. It is beyond the scope of this text to discuss all the radar EP techniques in use today. This chapter will discuss the most common EP techniques. They have been organized by function of the technique within the radar. These functions include radar receiver protection, jamming avoidance, jamming signal exploitation, overpowering the jamming signal, pulse duration discrimination, angle discrimination, bandwidth discrimination, Doppler discrimination, and time discrimination.

2. RADAR RECEIVER PROTECTION

The following are some of the most common radar counter-countermeasures designed to prevent receiver overload or saturation.

a. Sensitivity time control (STC) is used to counter close-in chaff or close-in clutter. Receiver gain is set at normal for long ranges and reduced for close-in ranges. One problem with using STC is that close-in targets may be missed if STC is improperly adjusted.

b. Automatic gain control (AGC) is used to counter chaff, clutter, and most types of transmitted jamming. AGC senses the signal level of a receiver's output and develops a back-bias, producing a constant output level. This technique has a slow response time compared to fast AGC and instantaneous AGC, both of which are employed instead of AGC. Also, it cannot maintain correct IF output levels for different intensity signals that are close in range because the bias voltage has a long buildup and decay time.

c. Fast automatic gain control (FAGC) is also employed against chaff, clutter, and most types of transmitted jamming. FAGC works by sensing the signal level of receiver output and develops a back-bias, tending to hold output constant. Response time is within milliseconds, permitting fast response and recovery as the antenna traverses the jammer's bearing. There are several precautions to note when using FAGC. First, targets may be suppressed and lost without the operator knowing that jamming is present. Second, a strong pulse or echo may cause ensuing weak targets to be lost. Lastly, FAGC has difficulty getting an accurate bearing on the jamming source.

d. Instantaneous automatic gain control (IAGC) is another technique to counter chaff, clutter, and most types of transmitted jamming. IAGC senses the signal level of each echo or jamming pulse and develops a back-bias that holds the stage output constant. Gain control response time is within milliseconds and extends the dynamic range of the receiver. There are several precautions to note when using IAGC. First, it is not effective against signals whose "in band" time is less than the IAGC response time. Also, with continuous duty cycle jammers, targets may be lost without the operator knowing that jamming is present. Finally, it is difficult to get an accurate bearing on the jamming source.

e. Automatic noise leveling (ANL) counters noise jamming and modulated or unmodulated constant wave jamming. ANL samples receiver noise content at the end of each PRF and sets the gain accordingly for the next pulse interval. Continuous jamming reduces gain to keep output the same as the original noise level. ANL also follows the scanning rate of the antenna so that receiver noise output is constant as the antenna rotates. When using ANL, targets may be suppressed and lost without the operator knowing that jamming is present. Also, receiver gain is unstable when pulses or swept jamming enter the sampling gate intermittently.

f. The logarithmic receiver (LOG) counters most types of transmitted jamming by amplifying and demodulating large dynamic-range signals in logarithmic amplifiers. This produces "amplitude compression" of the strong

signals. However, when using LOG, output is nearly constant so the operator cannot easily tell when jamming is present.

g. The logarithmic receiver with fast time constant (LOG-FTC) counters narrowband jamming, chaff, and clutter. This technique amplifies and demodulates large dynamic-range signals in logarithmic amplifiers, producing "amplitude compression" of the strong signals. Video is coupled through FTC circuits to eliminate rectified carrier and low frequency sideband products. There are problems with using LOG-FTC. First, the receiver output is nearly constant, so the operator cannot always tell when jamming is present. Second, LOG-FTC is not effective against wideband, or fast-swept, short pulse jamming. Lastly, LOG-FTC causes a broadening of displayed jam sector. as well as degrading bearing accuracy on the jam source.

h. Dicke-fix (DF) counters wideband and fast-swept jamming and is similar in employment to wideband limiting (WBL). DF amplifies without ringing, clips down all pulses to a common level, then amplifies the narrowband echo signal more than the wideband jamming. Noise level is held constant, independent of jamming intensity. There are precautions to note when using DF. Jamming that enters the wideband limiter can capture limiters, causing poor receiver sensitivity. Targets may be suppressed without the operator knowing that jamming is present. Also, resolution and target detection range are reduced, even in a non-jamming environment. Finally, DF is ineffective against extremely fast-swept spot jamming.

i. WBL is used to counter wideband jamming and fast-swept jamming. WBL amplifies without ringing, clips down all pulses to a common level, then amplifies the narrowband echo signal more than the wideband jamming. Noise level is held constant, independent of jamming intensity. However, jamming that enters the wideband limiter can capture limiters and cause poor receiver sensitivity. Resolution and target detection range is reduced, even in a clear environment. Targets may be suppressed without the operator knowing that jamming is present. Finally, WBL is ineffective against fast-swept spot jamming.

j. Adaptive video processing (AVP) counters chaff corridors, weather, sea clutter, and most types of transmitted jamming. AVP combines the adaptive threshold, beam-to-beam correlation, and wide-pulse blanking in frequency-scanning three-dimensional radars to avoid collapsing undesired returns on the PPI display. However, when using AVP, there is a decreased probability of detection in some multiple target situations. Also, targets may be suppressed without the operator knowing that jamming is present. Finally, AVP passes all point targets.

3. JAMMING SIGNAL AVOIDANCE

The following are some EP techniques used to avoid jamming signals.

a. Frequency agility (FA) counters narrowband jamming and some types of repeater and deception jamming. FA enables the radar to make rapid changes of transmitter and receiver operating frequency, sometimes on a pulse-to-pulse basis. Manual frequency changes may cause mutual interference with other radars and services.

b. Frequency diversity works against narrowband jamming and some types of repeaters and transponders. It is a multiple-radar coordination procedure in which radars are assigned operating frequencies that are separated to reduce mutual interference and their susceptibility to a single jammer. It is important to note that other radars may be operating at the same allocated operating frequency.

c. Polarization diversity is used to counter chaff, weather, and transmitted jamming. Polarization diversity attenuates jamming input to a radar receiver by using antenna polarization different from jammer polarization, and usually involves separate radars of different polarization. There are two precautions when using polarization diversity: (1) ground clutter worsens on vertical polarization, and (2) close coordination is necessary if separate radars are used; for example, one horizontally polarized search radar and one vertically polarized search radar.

d. Circular polarization (CP) works against chaff, weather, and transmitted jamming. CP attenuates jamming input to a radar receiver by using antenna polarization different from jammer polarization, and usually involves separate radars of different polarization. CP also improves target detection in rain clutter.

e. Conical-scan-on-receive-only (COSRO) is employed against inverse conical scan jamming to deny a jammer the ability to sense and upset scan angle tracking information. A constantly transmitted illumination beam is received and scanned to derive target angle information. However, the jammer can still degrade angle tracking if it can approximate the received signal scan rate.

f. Speedgate tracking is used against all types of transmitted jamming. The technique provides a very narrow bandpass having a center frequency related to Doppler shift. Only jamming within the restricted band is effective. It has the advantages of accurate target Doppler discrimination and good target tracking at low target levels. However, the speedgate can be stolen by gate stealers and some types of swept jamming.

g. Leading-edge track (LET) is used to counter an incoming target dropping chaff by allowing target tracking on the leading edge of the target. Trailing edge track (TET) is used to counter a receding target dropping chaff.

h. Track coast is used to counter chaff, clutter, multiple targets, range gate stealers, jam fades, and blinking jamming by placing tracking radar in a rate-aided coast condition. The system "estimates" target position to avoid interrupting the fire control solution. A lock-on or return to acquisition mode terminates the track coast condition. Track coast requires adequate storage of rate-aided information, and no true tracking information will be developed while track coast is operating.

i. Guard gates work against chaff, clutter, multiple targets, range gate stealers, jam fades, and blinking jamming. Guard gates provide automatic detection of a foreign signal and "estimates" target position to avoid interrupting fire control solutions. Like track coast, guard gates require an adequate store of rate-aided information with no true tracking information developed.

4. JAMMING SIGNAL EXPLOITATION

The following are some EP techniques that use the jamming signal for target acquisition and engagement.

a. Passive angle tracking (PAT) counters most types of transmitted jamming by allowing the radar to acquire and angle-track the source of jamming signals. There are some problems with this technique. Blinking jamming can cause severe instability, and the range of the jammer is unavailable until the target reaches burnthrough range.

b. Home-on-jamming (HOJ) counters most types of transmitted jamming by allowing the missile or radar to use the jamming signals, locate the source, and home on it. However, blinking jamming can cause severe instability, and the range of the jammer is unavailable until burnthrough.

c. Jamming signals produce recognizable sounds that help in their detection and identification. Aural recognition allows an operator to listen to the Doppler frequency associated with a moving target. It is used to counter most types of jamming.

d. The local oscillator off technique counters continuously transmitted jamming. No receiver output occurs unless a target echo signal and a jamming signal are present. Limitations of this technique include: targets only display in an area where jamming is present; and, if the antenna rotates away from the jammer, or if jamming is turned off, no targets are displayed on the radar scope.

e. The jamming strobe indicator counters any transmitted jamming with high-duty-cycle modulation. The indicator is a variable marker strobe on the radar display that moves in range proportional to jamming strength. The indicator traces an antenna lobe pattern on the display, showing the azimuth of the jamming source. There are some problems with the jamming strobe indicator. First, it interrupts normal video in some radars. Second, inverse or sidelobe

jamming can cause erroneous strobes. Finally, the jam strobe does not react to unmodulated CW or low-duty-cycle jamming for some radar systems.

f. The jamming indicator lamp, located on the operator console, is used on radars with automatic noise leveling (ANL) to counter continuous transmitted jamming. The lamp alerts the operator to the presence of jamming, and the ANL is manually shut off. This action allows the operator to determine jammer bearing.

g. Clean strobe generation (CSG) counters any transmitted jamming by using the sidelobe blanking circuits of a radar. An azimuth strobe appears when the jamming level in the main antenna exceeds the jamming level in the sidelobe auxiliary antennas. The operator is alerted to the presence and bearing of a jamming source, even with a constant false alarm radar (CFAR) receiver.

h. Jamming attenuation (JAM ATTEN) counters both clutter and any type of jamming. Receiver gain is reduced to avoid receiver saturation by inserting an attenuator pad that enables the operator to recognize presence, type, and bearing of a jamming source. When using JAM ATTEN, however, the reduced gain may cause loss of targets, even in non-jammed sectors. Also, any improvement in signal-to-jam ratio is not possible.

i. Receiver manual IF gain (MAN GAIN or IF GAIN) also counters clutter and jamming. Receiver gain is reduced to avoid jamming saturation by manually reducing stage gain, allowing the operator to identify jammer presence, type, and bearing. When using IF GAIN, the reduced gain may cause a loss of targets, even in non-jammed sectors. Also, any improvement in signal-to-jam ratio is not possible.

5. OVERPOWERING THE JAMMING SIGNAL

Following are some EP techniques a radar system can employ to overpower jamming and reduce the jamming-to-signal (J/S) ratio to less than one.

a. Burnthrough counters most types of transmitted jamming. Energy in the target pulse is raised by increasing the peak power, that is, the PRF or pulse width, or by increasing the time the radar illuminates the target by reducing the scan rate or scan angle. Some radars have modes in which the radar concentrates its power in narrow azimuth and elevation sectors about the suspected target position. However, burnthrough can degrade general radar performance by overloading the receiver if a large radar cross section target is detected. High power may impact radar operation in clutter or dense chaff environments.

b. Narrowband long pulse (NBLP or NLP) counters most types of transmitted jamming by using a high-energy long pulse. The signal uses a narrowband receiver for reception, and increases detection range for targets in jamming and

in the clear. Simultaneously, it reduces resolution, which causes poor radar performance in chaff and clutter.

6. PULSE DURATION DISCRIMINATION

The following are some radar counter-countermeasures that use pulse duration to discriminate between radar and jamming signals.

a. The fast time constant (FTC) is used to counter chaff, clutter, and narrowband jamming. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation jamming. FTC passes normal radar pulse lengths with little attenuation, but causes some loss of receiver sensitivity.

b. Pulse width discrimination (PWD), clutter eliminate (CE), and wide pulse blanking (WPB) are designed to counter chaff, most types of jamming, EMI, and some types of deception jammers. A video coincidence gate, involving a delay line matched to the expected signal duration, senses if a return is the proper pulse width. PWD, CE, and WPB provide an enabling path for qualified signals. However, weak signals may be lost in the signal processing.

c. Pulse expansion-compression (PC) is used to counter most types of noise jamming and some types of deception jamming. An expanded pulse is coded for transmission. This expanded pulse is transmitted and decoded on return. Echo responses are then compressed in a decoding process. This expansion/compression is equivalent to NLP, which provides longer detection ranges, and wideband short pulse, which provides increased resolution. Using PC is not without problems. Unwanted residues may cause loss of weak targets. Additionally, range error proportional to the Doppler shift, or radial velocity, affects the accuracy of the PC.

7. ANGLE DISCRIMINATION

The following techniques use angle discrimination to distinguish between radar returns and jamming signals.

a. Sidelobe blanking (SLB) and sidelobe cancellation (SLC) are types of sidelobe suppression (SLS) used to counter sidelobe response to chaff, clutter, transmitted jamming, sidelobe jamming, and deception jamming. An auxiliary antenna approximates the pattern and gain of sidelobes of the main antenna and produces a signal for comparison with the signal received in the main antenna. If the signal in the auxiliary antenna is greater, the signal in the main antenna channel is blanked. This permits bearings to be obtained on a jamming source and rejects sidelobe jamming. SLB is useful only for determining the bearing to the jamming source.

b. Antenna manual positioning, antenna traverse and elevation angle offset, antenna jog, and antenna slow scan are EP techniques used to counter main

beam and sidelobe jamming and deception. These techniques are designed to increase the antenna scans across the jammed sector to increase the blip-scan ratio. These techniques increase the number of pulses integrated, as well as the operators' sorting capability.

8. BANDWIDTH DISCRIMINATION

The following are EP techniques that use bandwidth to distinguish between radar jamming and target returns.

a. Dicke-fix (DF) counters wideband and fast-swept jamming. A wideband limiter amplifies without ringing and clips all pulses down to a common level. Then an amplifier increases narrowband target echo signals more than the wideband jamming. There are some problems associated with DF. Any jamming entering the wideband limiter can "capture" the limiters and cause poor receiver sensitivity. Targets may be suppressed without the operator knowing that jamming is present. Resolution and target range is reduced, even in a clear environment. Finally, DF is ineffective, even harmful, when the jamming bandwidth is near the bandwidth of the desired echo signal.

b. Transmitter pulse lengthening (TPL) counters wideband and fast-swept jamming. TPL concentrates power into a narrow band about the carrier frequency by lengthening the transmitting pulse. While this allows use of a narrowband receiver, it impairs resolution, causing poor chaff and clutter performance.

c. Transmitter pulse shaping (TPS) counters wideband and fast-swept jamming. The sideband range is limited by shaping the transmitted pulse. This allows use of a narrowband receiver, but impairs resolution, causing poor performance in chaff and clutter.

d. Narrowband pulse limiting (NBLP or NLP) is a form of transmitter pulse lengthening that counters wideband jamming and fast-swept jamming. NBLP concentrates power into a narrow band in the carrier frequency by lengthening the transmitting pulse. This allows use of a narrowband receiver, but impairs resolution, causing poor chaff and clutter performance.

e. The fast time constant (FTC) is used to counter chaff, clutter, and narrowband jamming. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation of jamming. FTC passes normal radar pulse lengths with little attenuation, but causes some loss of receiver sensitivity.

f. High video pass (HVP) is used to counter chaff, clutter, and narrowband jamming. It is similar to FTC. A video circuit provides low frequency attenuation to reject carrier and low frequency modulation jamming. HVP passes only the leading edge of the received pulses. HVP can cause some loss of receiver sensitivity.

g. The wideband short pulse (WSP) counters chaff, clutter, and narrowband jamming by transmitting a short pulse and using a wideband receiver for reception. Echo resolution and accuracy are improved, and performance against narrowband jamming is enhanced. However, the maximum detection range is decreased and system vulnerability to wideband jamming is increased.

h. Narrowband limiting (NBL) counters chaff, clutter and narrowband jamming. A narrowband filter positioned in the front of the amplifier section allows only the target signal bandwidth to enter the limiter, reducing wideband and off-frequency jamming. The limiter clips all signals and noise to a common level. This technique is useful only when followed by pulse compression or other "decode" techniques. NBL is not effective against wideband jammers capable of causing "ringing" of the NBL bandpass filters. Targets may be suppressed and lost without the operator knowing that jamming is present. Finally, target detection and resolution are poor.

9. DOPPLER DISCRIMINATION

EP techniques that use Doppler frequency discrimination between radar and jamming signals to negate jamming effectiveness include the following:

a. Moving target indication (MTI) is used to counter chaff and clutter. The phase of returned target echoes is compared on a pulse-to-pulse basis. Those with no phase change (no change in radial velocity) are cancelled using a delay-line canceler. Sensitivity using MTI is poor for weak targets, even in the clear. Also, it is blind to targets that have a Doppler frequency that is equal to a multiple of the radar PRF, unless PRF stagger is used. Finally, limited dynamic range does not allow full cancellation of strong clutter echoes.

b. Compensated coherent MTI, also known as compensated COHO MTI, counters chaff and clutter by comparing the phase of returned target echoes on a pulse-to-pulse basis. Those pulses with no phase change, that is, no change in radial velocity, are cancelled. Corrections to the coherent oscillator are applied to compensate for motion of the platform and radar antenna. Sensitivity is poor for weak targets, even in the clear, and it is blind to targets that have a Doppler frequency equal to, or a multiple of, the radar PRF, unless PRF stagger is used.

10. TIME DISCRIMINATION

The following EP techniques use time discrimination between radar and jamming signals to negate jamming effectiveness.

a. Video integration (VINT) and integrate-multiply (INT-MULT) counter any form of transmitted jamming not synchronous with radar PRF. The video continuously circulates through a delay line, delaying signals exactly one pulse recurrence time (PRT), then combines them with signals from the next PRT. Synchronous target signals add together to increase video output, but noise and

random pulses are suppressed. Unless MTI and FTC are employed, VINT and INT-MULT will enhance chaff, clutter, and jamming along with target returns. Also, feedback control, or loop gain, must be carefully adjusted for optimum results.

b. PRF stagger and jitter are EP techniques designed to counter quasi-synchronous jamming, EMI, MTI blind-speeds, "second trip" echoes, and repeater jammers simulating "closer-than-real" targets. The transmitter pulse interval is varied to break up synchronous patterns. Received signals must be "de-staggered" for use with MTI or integration. When using any of these techniques, video de-stagger balance must be accurate or "double video" occurs. These techniques are not effective against exact synchronous deception jamming.

c. Pulse-to-pulse correlation (PPC) is used to counter slow-swept blinking or pulsed jamming not synchronized with radar PRF, and some types of deception jammers. To be displayed, target video must exceed a threshold voltage for two successive pulses. The technique avoids displaying non-synchronous jamming and EMI, but is not effective against synchronous jamming, and may cause weak targets to be missed.

d. Beam-to-beam correlation (BBC) is used to counter slow-swept blinking or pulse jamming not synchronous with radar PRF and some types of repeaters and transponders. It is used in three-dimensional frequency scanning or frequency agile radars. To be displayed, the target return echo signal must exceed a threshold value in two adjacent antenna beams. BBC is not effective against synchronous jamming. It may also cause weak targets to be missed.

e. Single beam blanking (SBB) is used to counter slow-swept, blinking, pulsed jamming, and narrowband jamming. It is used in three-dimensional frequency scanning or frequency agile radars to avoid displaying vertical beams containing jamming. A blanking pulse is generated for vertical beams containing jamming so that they are not displayed. The technique is not used on RHI video. Also, it can cause a loss of targets at the jammed elevation angle on the PPI display.

11. SUMMARY

This chapter has discussed some of the most widely employed EP techniques designed to counter radar jamming. The capabilities of the individual radar operator were not discussed. However, the radar operator is as important as the EP techniques designed for the radar system. Many of the most effective EP techniques are designed to ease operator interpretation of the radar display. In the chess game of EW, the capabilities of individual radar operators can be as important as the sophisticated EP techniques in determining the final outcome.

CHAPTER 17. RADAR WARNING RECEIVER (RWR) BASIC OPERATIONS AND GEOLOCATION TECHNIQUES

1. INTRODUCTION

Radar surveillance and radar-directed weapons represent the biggest threat to aircraft survival on the modern battlefield. The first step in countering these threat systems is to provide the pilot or crew with timely information on the signal environment. The radar warning receiver (RWR) is designed to provide this vital information to the pilot. The RWR system is an example of an electronic warfare support (ES) system. The primary purpose of an RWR system is to provide a depiction of the electronic order of battle (EOB) that can have an immediate impact on aircraft survival. Though the RWR system is complex, the basic operations of the various components are straightforward. A step above RWR systems is threat geolocation. While an RWR provides the EOB for a single aircraft, threat geolocation systems can provide accurate threat location data for numerous aircraft over an entire region. Threat location data is used for aircraft threat avoidance and, more common today, the preemptive attacking of enemy radar sites. This chapter will discuss the functions of the various components of a RWR system including the antennas, receiver/amplifiers, signal processor, emitter identification (EID) tables, RWR scope, RWR audio, and limitations to RWR systems. This chapter will then go on to discuss three of the methods used to geolocate radar threat systems.

2. RWR ANTENNAS

Antennas are designed to receive radar pulses from threat radar systems. Factors that impact the operation of the RWR antennas include location, pattern, sensitivity, and polarization.

a. The physical location of the RWR antennas on the aircraft can affect its ability to detect a radar signal. Antennas are arranged to cover a predetermined area of horizontal and vertical space around the aircraft (Figure 17-1).

b. The antennas and their patterns play an essential part in displaying the spatial relationship of a threat radar to the aircraft. The antenna patterns are the areas, or “footprints,” that the antennas are specifically designed to cover. These footprints are directly affected by the relative position of the antennas to the threat systems. This is because the signal processor measures and compares signal strength from all the aircraft antennas to compute threat signal location relative to the aircraft. This relative location is then presented on the RWR scope display. Aircraft movement and maneuvering shifts these relative positions during flight and can distort the true threat position on the RWR scope. Precise position determination is not possible with most RWRs.

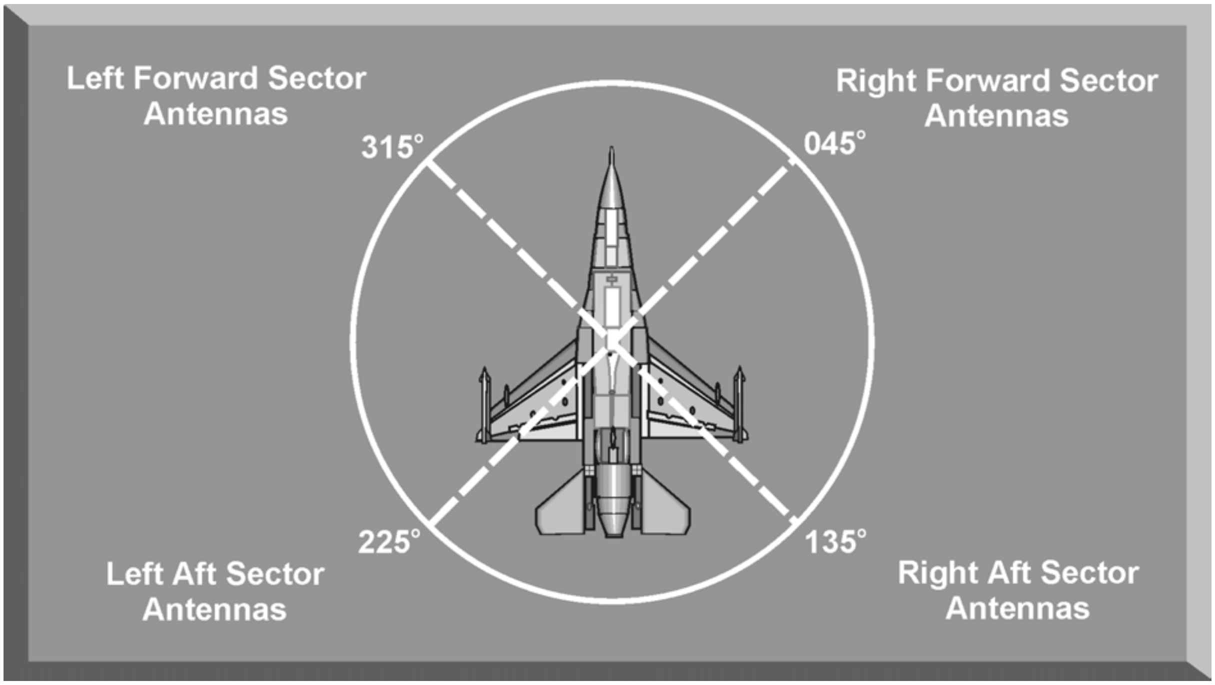
Left Forward Sector Antennas
Right Forward Sector Antennas
315°
045°
225°
135°
Left Aft Sector Antennas
Right Aft Sector Antennas

Figure 17-1. Radar Warning Receiver (RWR) Antennas

c. The sensitivity of an RWR antenna directly affects its ability to detect a radar signal. The more sensitive the antenna, the further it can detect a signal. The sensitivity of a system and its ability to intercept a radar signal is usually expressed in decibels relative to milliwatts or dBm units. A 10 dBm change in sensitivity can result in a 25 nm range difference in target detection. In general, sensitivity levels of -50 to -60 dBms are required to detect signals at long ranges (Figure 17-2).

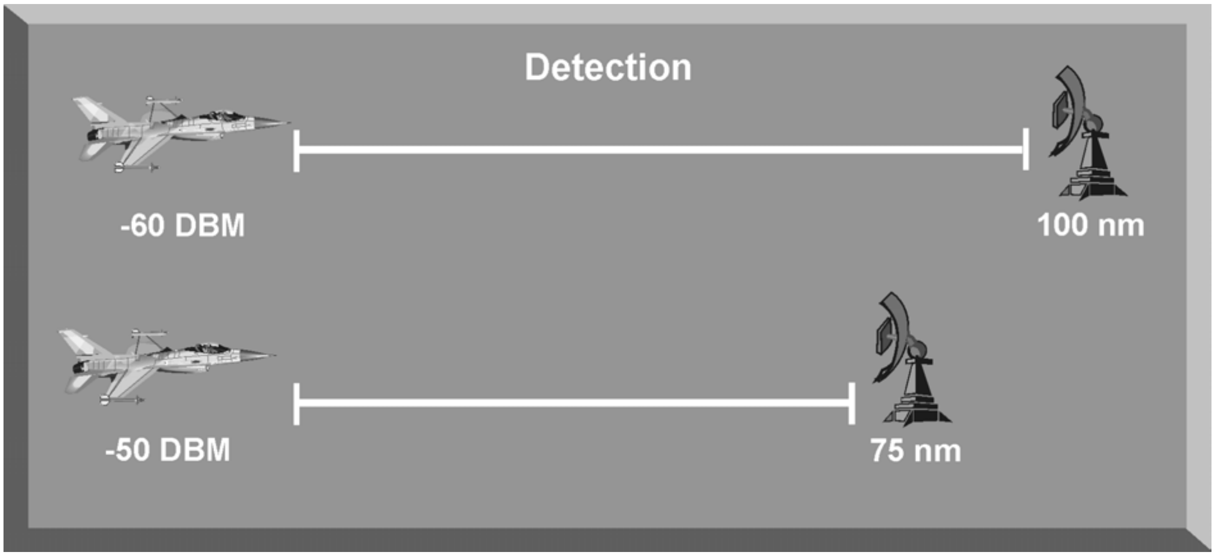
Detection
-60 DBM
100 nm
-50 DBM
75 nm

Figure 17-2. RWR Antenna Sensitivity

d. Another factor affecting antenna detection range is the polarization of the antennas. If the polarization of the RWR antenna and the threat system antenna are mismatched, or cross-polarized, initial detection of a threat signal could be delayed until the aircraft is within the lethal range of a threat system. In this situation, the aircraft could be engaged with minimal warning (Figure 17-3).

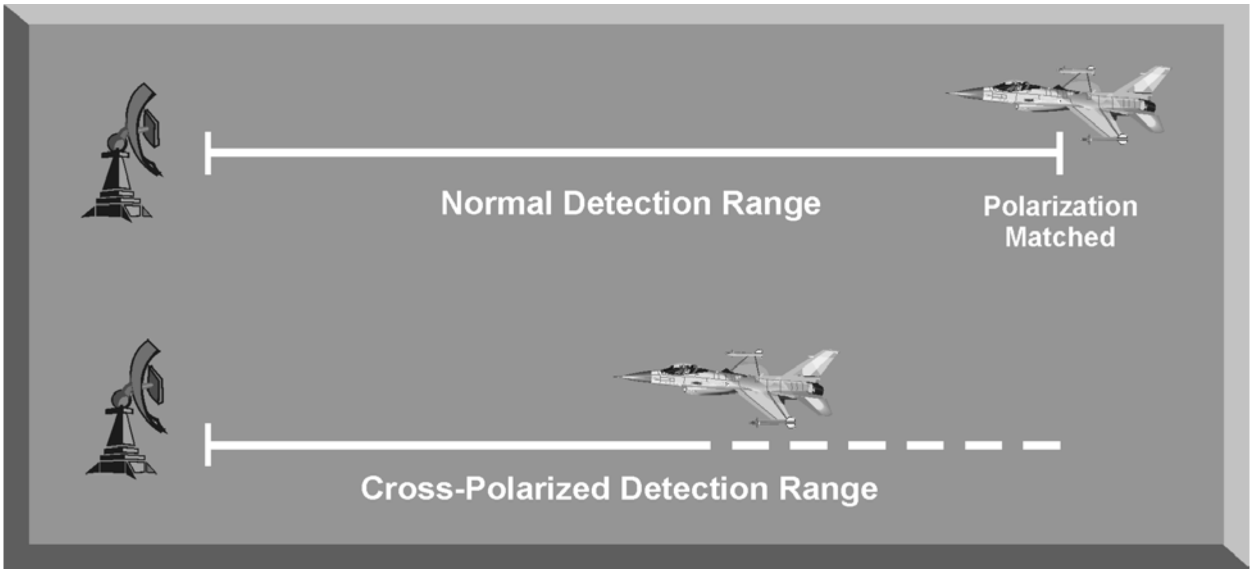

Figure 17-3. RWR Antenna Polarization

3. RWR RECEIVER/AMPLIFIERS

The RWR receiver/amplifier section processes the radar signals from the antennas. Most RWR systems use frequency bands to differentiate signals. Nominal band designations are summarized in Table 17-1. There are two types of receivers currently used in RWR systems: the crystal video receiver and the superheterodyne receiver.

Table 17-1. RWR Frequency Band Designators

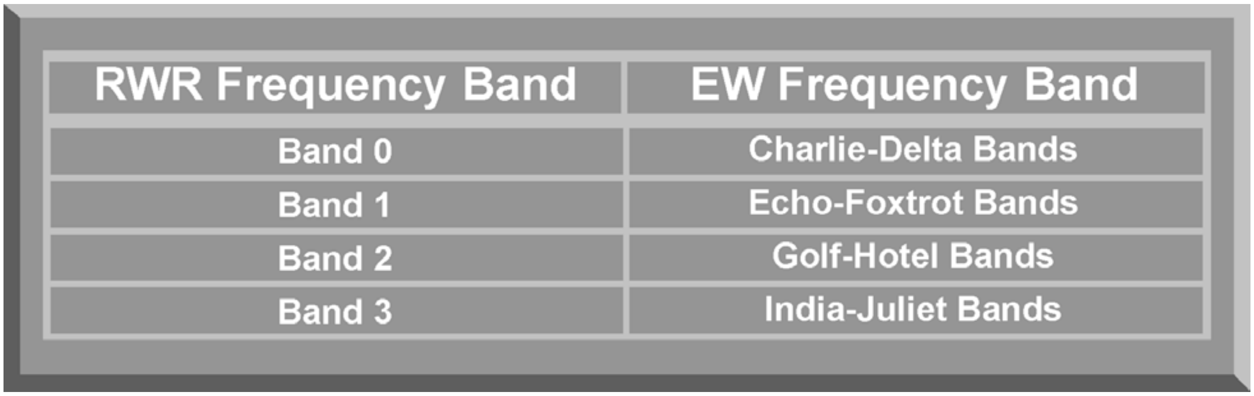

RWR Frequency Band	EW Frequency Band
Band 0	Charlie-Delta Bands
Band 1	Echo-Foxtrot Bands
Band 2	Golf-Hotel Bands
Band 3	India-Juliet Bands

a. A crystal video receiver (CVR) is the simplest type of microwave receiver. It is used primarily for detection of pulse radar signals in the 2 to 18 GHz band. A CVR used in a radar warning receiver incorporates crystal detectors for each designated frequency band. A pulse radar signal is detected by the antenna and passed to the multiplexer. The multiplexer divides the received radar signals by frequency band and sends the signals to the appropriate band channel. The RF amplifier boosts the radar signal and passes it to the crystal detector (Figure 17-4).

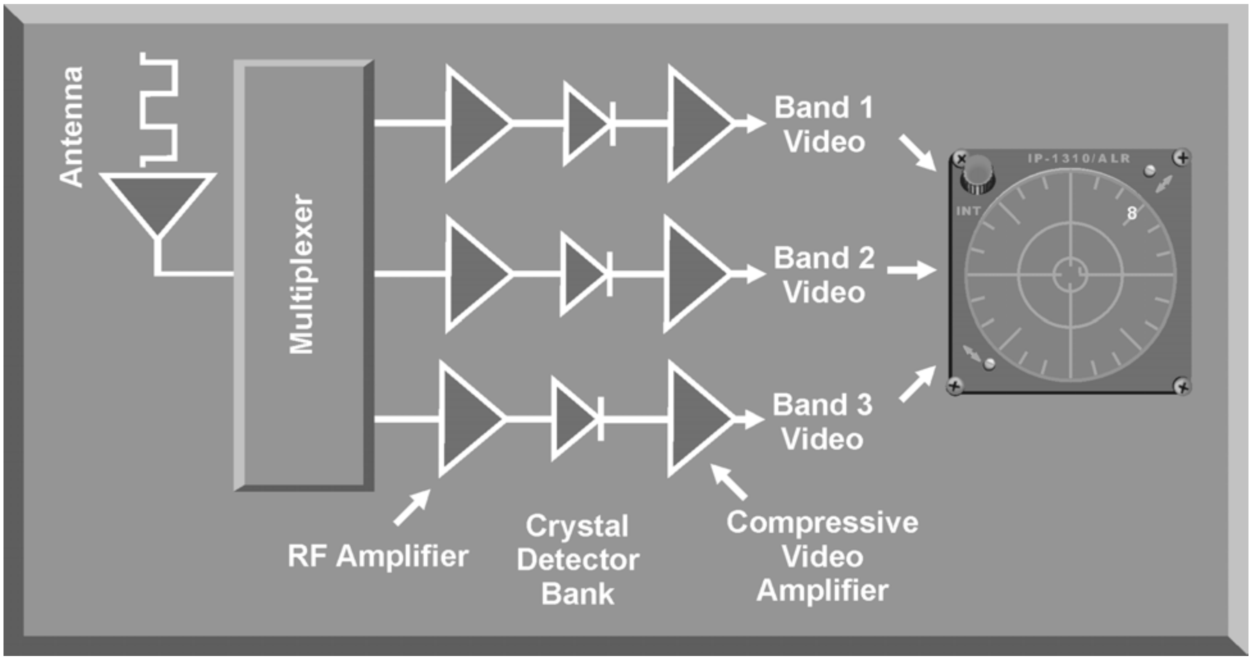

Figure 17-4. Crystal Video RWR Receiver

(1) The crystal detector is an RF diode, which converts the RF signal into a video signal. The voltage level of the output video signal is dependent only on the amplitude of the input signal and not on the frequency or phase. The sensitivity of a CVR is limited by the sensitivity of these crystal detectors. The sensitivity of crystal detectors currently available is generally adequate to detect main beam radiation from most threat radars. The video output of the crystal detectors is amplified by a high-grain compressive video amplifier and sent to the RWR scope for display.

(2) A CVR is extremely fast, sensitive, and covers a wide frequency range. These characteristics coupled with low cost and small size make CVRs ideal for use in radar warning receivers. The primary disadvantage of a CVR is that it is indiscriminate in reception and can be saturated in a dense signal environment. Multiple signals in the same band can cause amplitude distortion which can mask key threat signals.

b. A superheterodyne RWR receiver uses a pre-selector filter, mixer, and a local oscillator to translate the received signal to a lower intermediate frequency (IF). This lower IF allows the receiver to amplify and filter the received signal to provide greater sensitivity and frequency selectivity than a CVR (Figure 17-5).

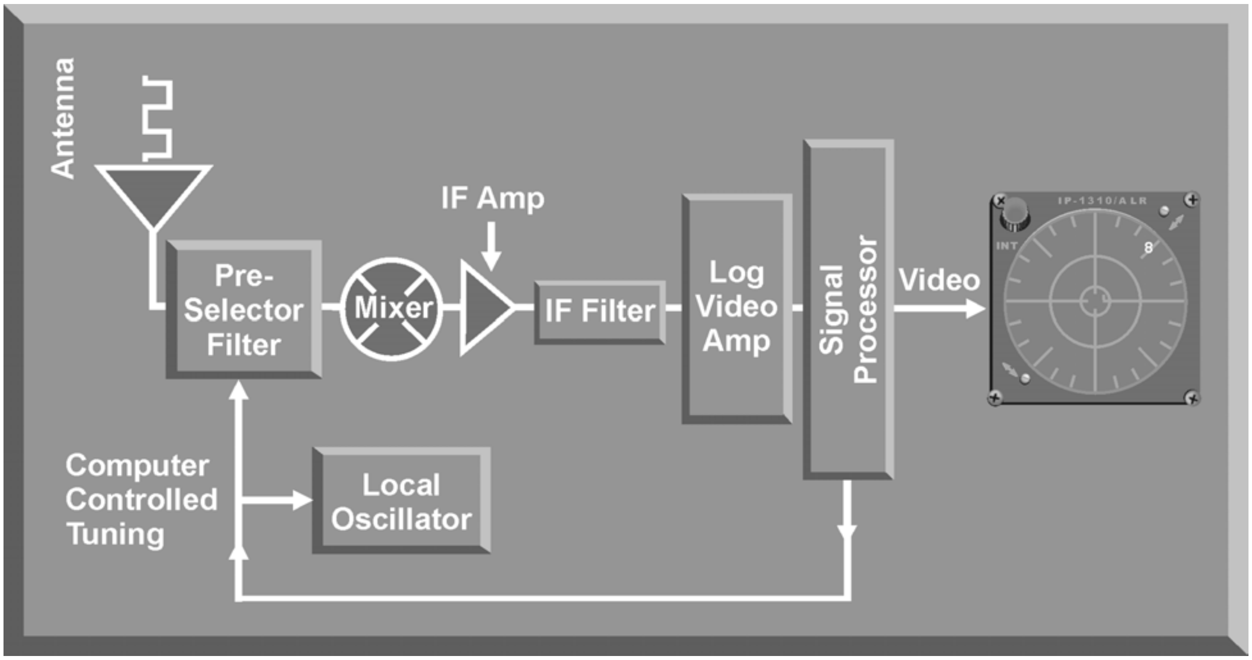

Figure 17-5. Superheterodyne RWR Receiver

(1) Superheterodyne RWR receivers use special scanning techniques controlled by the signal processor to tune the pre-selector filter and the local oscillator to rapidly scan selected threat system frequency bands. If the receiver detects activity in any of these bands, the scan stops to allow the processor to analyze the detected signals. The signal is combined with the local oscillator signal to lower the frequency to the IF. This signal is amplified, filtered, and amplified again before it reaches the signal processor. The signal processor classifies the threat and displays the proper threat symbol to the pilot. This entire process is accomplished in a matter of microseconds.

(2) The scanning superheterodyne receiver has important features that make it effective for RWR system application. It has excellent sensitivity and selectivity. It also has good frequency resolution. These features give the superheterodyne receiver a very low false alarm rate. The major disadvantage of the scanning superheterodyne receiver is its limited capability to receive signals from threat systems employing scanning antenna patterns. This limitation can be overcome with specific computer-controlled tuning to look for these threat signals.

4. SIGNAL PROCESSOR

The signal processor is the heart of the radar warning receiver. The signal processor is also known as the digital processor or analysis processor in different RWR systems. Its primary functions are to process numerous complex radar signals and identify, among the thousands of similar signals, those generated by lethal threat systems. The signal processor accomplishes this task continuously over the duration of the mission and displays the identified threat system to the aircrew almost instantly.

a. Signal processing begins when RF energy strikes the receive antennas on the aircraft. The received signals are then boosted in strength by intermediate amplifiers or antenna receivers. These amplified signals are then sent to the signal processor, or digital processor, where they are assigned a track file for reference to other signal characteristics. Data in these files is compared to those in the emitter identification (EID) table to process the signal for identification. Once identification is complete, a video and, if necessary, an audio signal, is sent to the cockpit display. The audio and video signals alert the aircrew to the electronic environment around the aircraft. This whole process takes less than a microsecond (Figure 17-6).

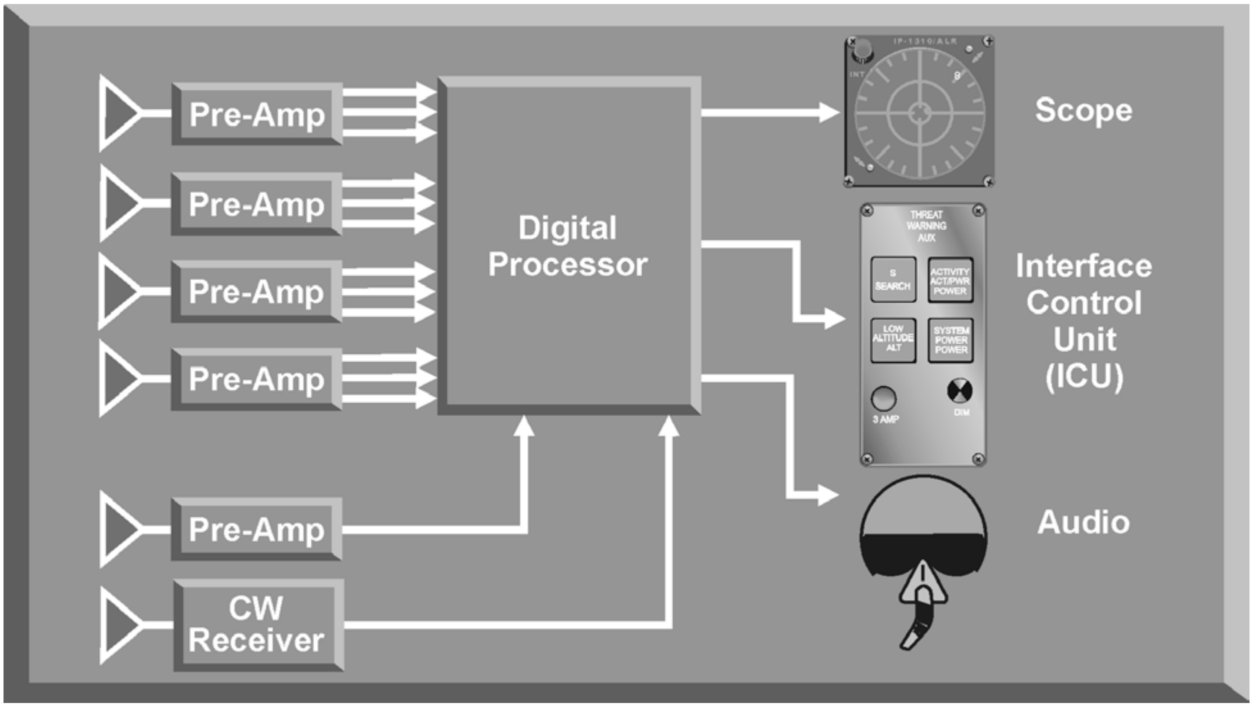
Pre-Amp
Pre-Amp
Pre-Amp
Pre-Amp
Pre-Amp
CW Receiver
Digital Processor
IP-1310/ALR
INT
8
Scope
THREAT WARNING AUX
S SEARCH
ACTIVITY ACT/PWR POWER
LOW ALTITUDE ALT
SYSTEM POWER POWER
3 AMP
DIM
Interface Control Unit (ICU)
Audio

Figure 17-6. RWR System

b. Since many signals may be present, the amplifier detectors boost the signal strength, and also tag each signal by certain characteristics such as its time of arrival, direction of arrival, and/or frequency. These signals, along with their respective tags, are sent to the signal processor for further processing and

identification. The signal processor then makes a track file for each signal it receives from the amplifier detector.

c. The signal processor classifies each received signal and corresponding track file by its unique radar signal characteristics. Identifying characteristics used by a signal processor can include radio frequency, pulse width, pulse repetition frequency, EP techniques, and more. Characteristics of one signal may be identical to characteristics from different signals, while certain other characteristics can be as unique as a human fingerprint. The signal processor uses these primary characteristics to identify specific signals. When the primary characteristics of two or more signals are similar, the signal processor uses additional signal characteristics to resolve any confusion between two or more signals.

d. The signal processor ranks the track files based on priorities determined from tests it conducts on the signal characteristics and the threat priorities contained in the EID tables. It then quickly processes signals belonging to lethal threats before it processes signals belonging to non-lethal threats. For example, three signals enter the processor together and separate track files are established for each signal (Figure 17-7). A test on the first characteristic discriminant, frequency, will delay the further processing of Signal 3, since no lethal threat systems operate at a frequency less than 2000 megahertz. A further test on the remaining prioritized signals may eliminate Signal 2 as a threat system signal, leaving more processor time for the identification of the threat system which generated Signal 1. These tests do not stop the processor from attempting to identify all received signals. The signal processor merely delays the identifying sequence until all high priority signals have been processed.

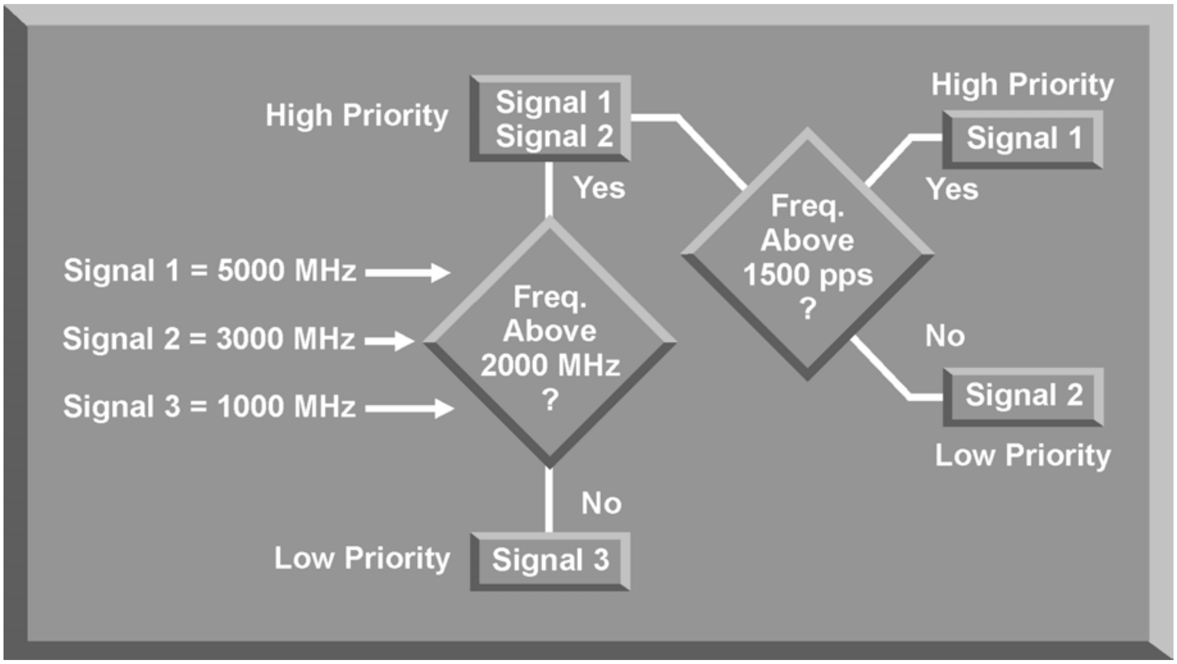

Figure 17-7. Signal Processor Signal Priority

5. EMITTER IDENTIFICATION (EID) TABLES

The signal characteristics in each track file are filled with processed data, and are constantly updated based on the time of arrival and location of the received signals. In addition, the track files are constantly compared to the EID table installed in the signal processor's computer memory. The EID table is a predefined table of radar characteristics associated with known radar systems (Figure 17-8). It is created from information gathered from electronic warfare support (ES) assets and intelligence sources. This table can be changed and updated as necessary to reflect the most current radar characteristics available for the anticipated threats in the planned theater of conflict. Each RWR system has unique procedures to reprogram the signal processor and update the EID tables. Emergency reprogramming actions, such as would be taken if a new threat appears that is not part of the current EID, are called a Pacer Ware.

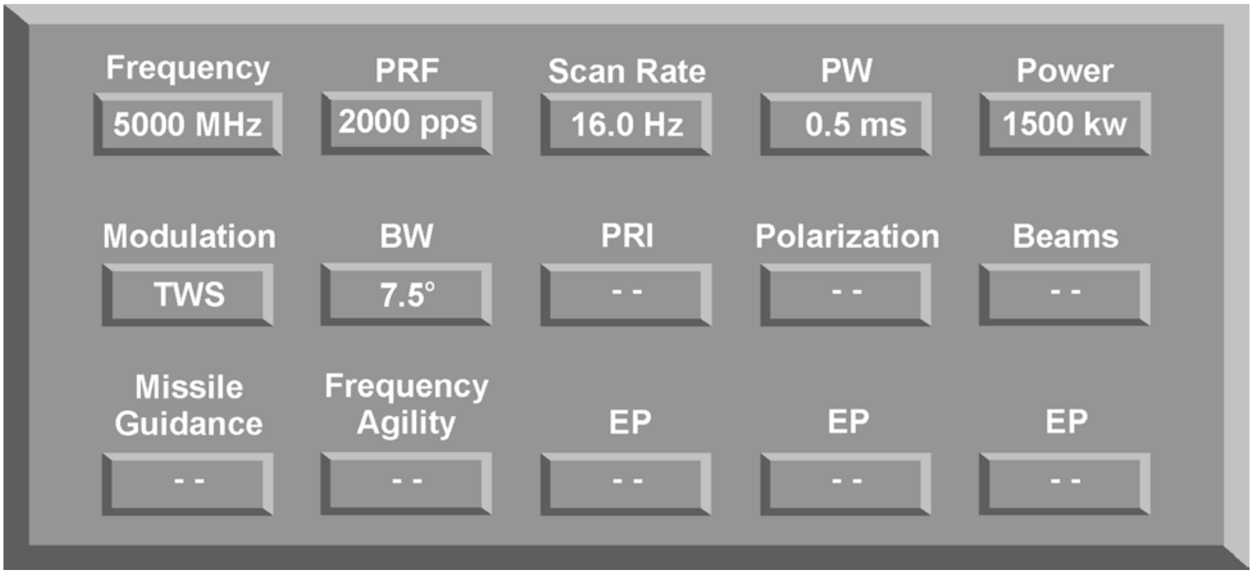

Frequency	PRF	Scan Rate	PW	Power
5000 MHz	2000 pps	16.0 Hz	0.5 ms	1500 kw
Modulation	BW	PRI	Polarization	Beams
TWS	7.5°	- -	- -	- -
Missile Guidance	Frequency Agility	EP	EP	EP
- -	- -	- -	- -	- -

Figure 17-8. Sample EID Table

6. RWR SCOPE DISPLAY

The signal processor continually compares signal characteristics in the track files with the data in the EID tables. Once the signal processor has determined that enough of the signal characteristics match the information in the EID tables, it generates and positions a video symbol on the RWR scope. The video symbol represents a specific threat, and each threat system has its own unique symbol. In addition, an audio tone is generated to alert the pilot. The signal processor also generates symbols and audio associated with specific threat system actions, including search, track, and missile launch. The position of the threat symbol on the RWR scope always represents the relative position of the threat in relation to the aircraft which is the center of the RWR scope. The signal processor compares the received signal strength in the different antennas to determine the proper location of the threat symbol. Figure 17-9 depicts a situation where the two

forward antennas receive equal signal strength therefore the signal processor places the symbol at the 12 o'clock position.

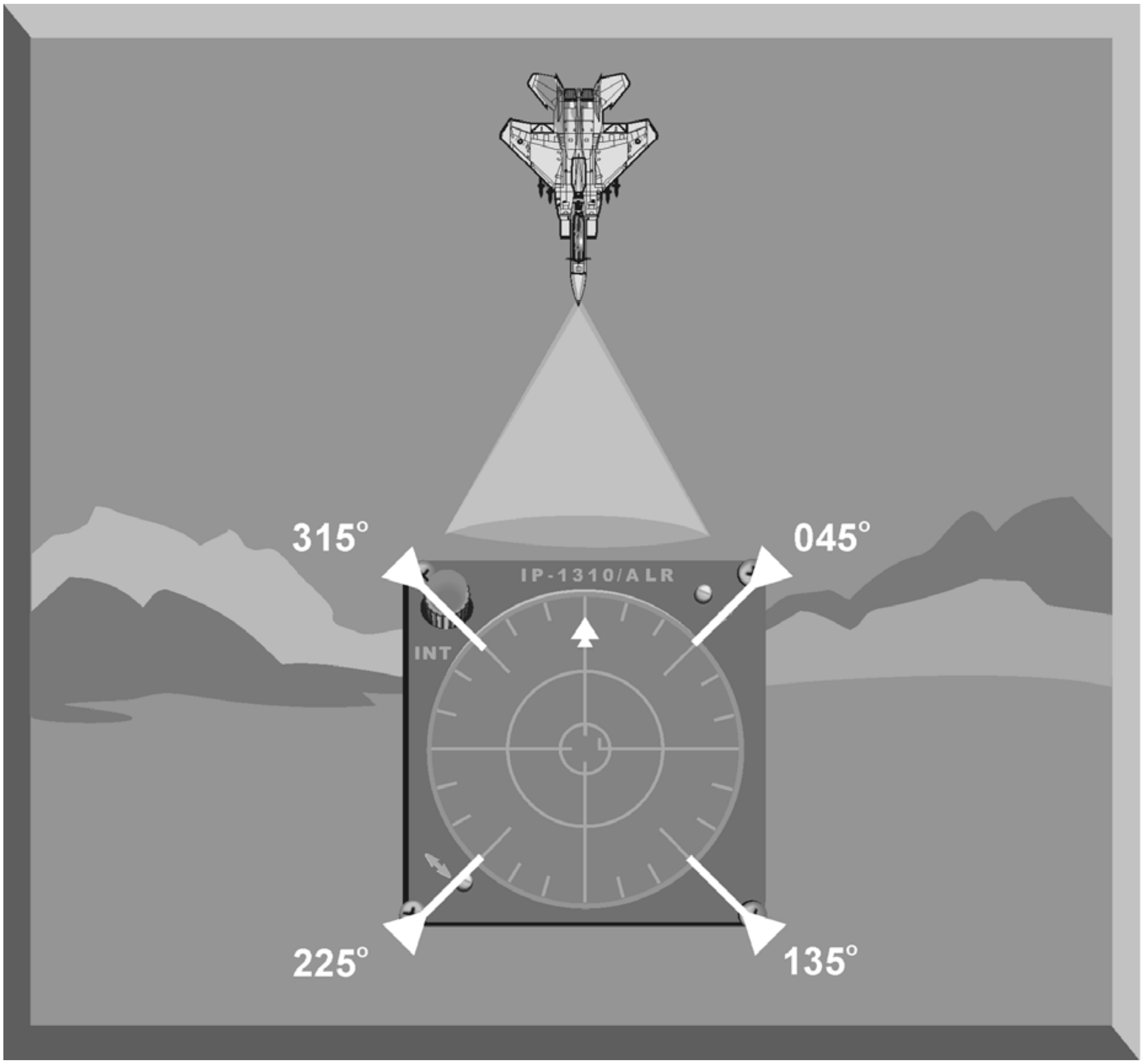

Figure 17-9. RWR Scope Azimuth Positioning

7. RWR AUDIO

In addition to generating threat symbols for each identified threat, the signal processor also generates threat audio. Threat audio first alerts the aircrew to the detection of a threat system. This RWR audio is generally referred to as "new guy" alert audio. The signal processor can also present constant audio from a selected threat. The aircrew controls this function through the interface control unit. The constant audio provided by an RWR system can be either "real" or synthetic. "Real" audio is normally based on the actual pulse repetition frequency (PRF) of the threat system radar whether the signal processor has identified it or not. Synthetic audio is based on the classification of the threat (SAM, AI, etc.) as

determined by the signal processor. The signal processor also generates a launch warning audio when the signal characteristics of the threat indicate a missile launch condition exists.

8. RWR INTERFACE CONTROL UNIT (ICU)

Every RWR system has some type of an ICU which provides the aircrew interface with the signal processor (Figure 17-10). The buttons on the ICU control specific functions of the signal processor. The ICU allows the aircrew to optimize the RWR system based on mission tactics. This optimization includes selecting appropriate priority lists based on ingress and egress tactics, controlling threat audio presentation, and determining the number and types of threats displayed. In addition, the ICU provides an additional visual indication of missile launch. All system test functions are controlled by the ICU to allow the aircrew to monitor the status of the RWR system.

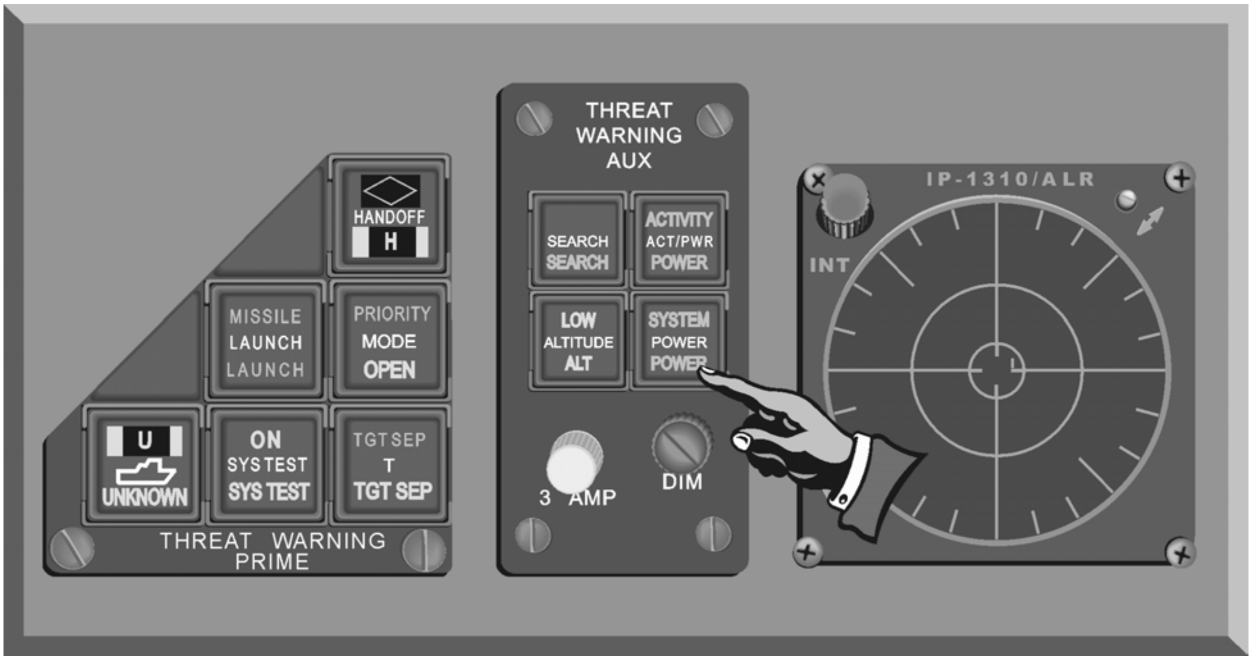
HANDOFF
H
MISSILE
LAUNCH
LAUNCH
PRIORITY
MODE
OPEN
U
UNKNOWN
ON
SYS TEST
SYS TEST
TGT SEP
T
TGT SEP
THREAT WARNING
PRIME
THREAT
WARNING
AUX
SEARCH
SEARCH
ACTIVITY
ACT/PWR
POWER
LOW
ALTITUDE
ALT
SYSTEM
POWER
POWER
3 AMP
DIM
IP-1310/ALR
INT

Figure 17-10. RWR Interface Control Unit

9. RWR LIMITATIONS

There are several limitations associated with all RWR systems. The most important limitations include ambiguities, impact of maneuvering, and electromagnetic interference (EMI).

a. The sheer number and diversity of radar systems associated with an enemy IADS greatly compound the problem of threat identification and warning for RWRs. Adding to this problem is the fact that many different threat systems use operating modes that are parametrically similar. When an RWR processes a

radar signal that has the same characteristics of a signal from a different system, an ambiguity may occur. An RWR ambiguity is defined as the display of more than one symbol for a specific threat signal. RWR ambiguities may occur from both enemy and friendly radar systems.

(1) Figure 17-11 depicts a number of friendly and threat signals that operate between 8000 and 10,000 megahertz. On any given combat mission, it is quite possible that the RWR will receive signals from one or more of these threat systems at the same time. If frequency is the only signal characteristic available for processing, the RWR will not be able to determine which system the signal represents. Since threat systems operating in this frequency range are potentially lethal systems, the signal processor will attempt to match the frequency with a threat system from the EID table.

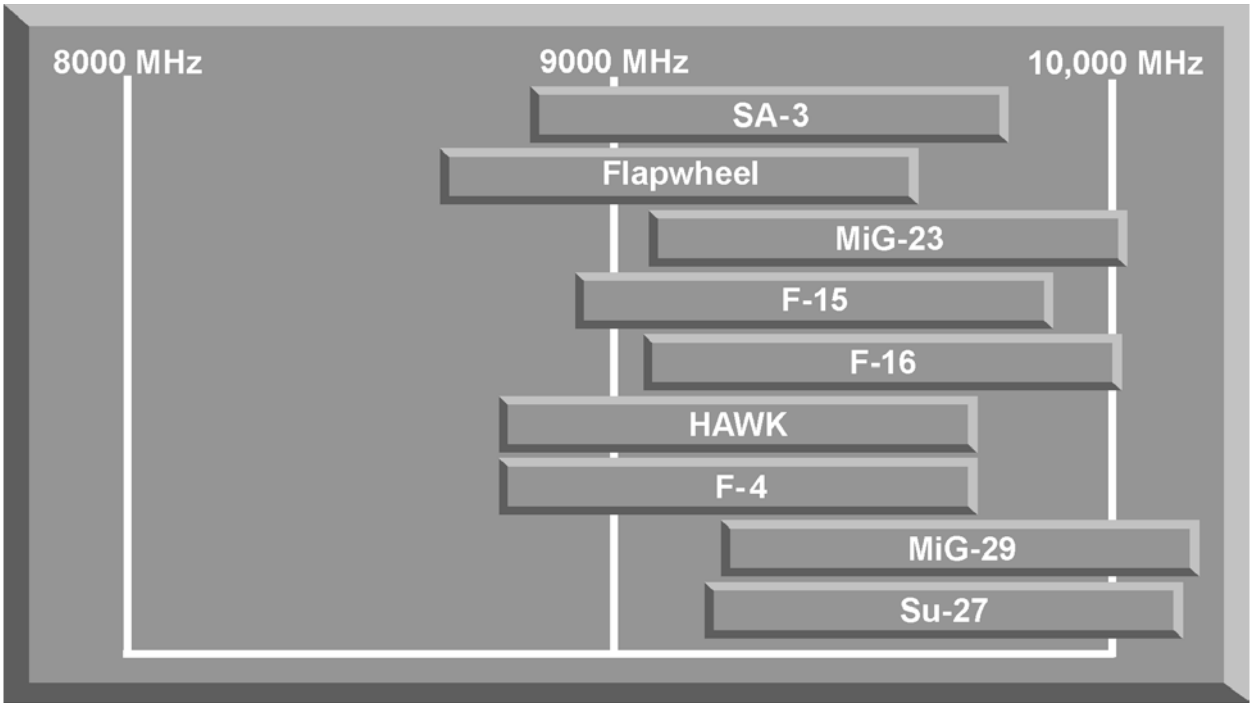

Figure 17-11. Signal Frequency Spectrum

(2) Matching partially processed signals to threat systems in the EID table may result in the wrong threat symbol being displayed, or numerous symbols being displayed on top of each other, making it difficult for the pilot to distinguish the exact threat. Incorrect threat symbology, or numerous combined symbols, are called RWR ambiguities.

b. RWRs are designed to provide accurate threat positioning information when the aircraft is flying straight and level. Most RWRs will also provide accurate threat positioning information when the aircraft is maneuvering up to certain limits of bank angle and turn rate. If aircraft maneuvering exceeds these

limits, RWR threat positioning data becomes unreliable. The two RWR limitations associated with aggressive maneuvering are inaccurate threat azimuth and multiple threat symbols.

(1) In Figure 17-12, the right forward and right aft RWR antennas detect a threat signal from a TTR. The left forward and left aft antennas are shielded by the aircraft and do not detect the signal. The signal processor determines the threat position, using the azimuth positioning algorithm, and displays the threat symbol at the 2 o'clock position.

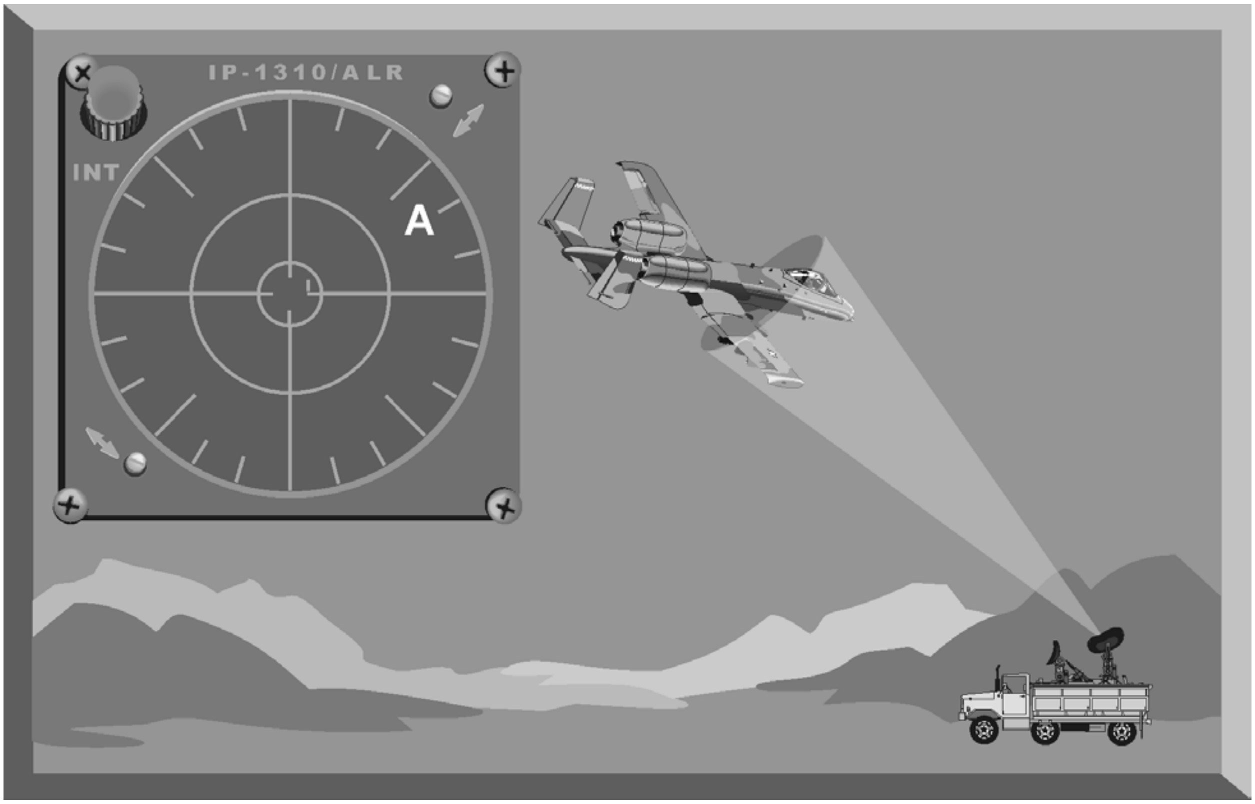

Figure 17-12. RWR Threat Azimuth Position Determination

(2) When the pilot maneuvers aggressively to put the threat symbol on the beam, he exceeds the RWR maneuvering limitations (Figure 17-13). Now all four RWR antennas detect the TTR signal. The signal strength in the right and left forward antennas is nearly equal. Based on this information, the signal processor displays the RWR symbol at the one o'clock position while the threat is actually at the three o'clock position.

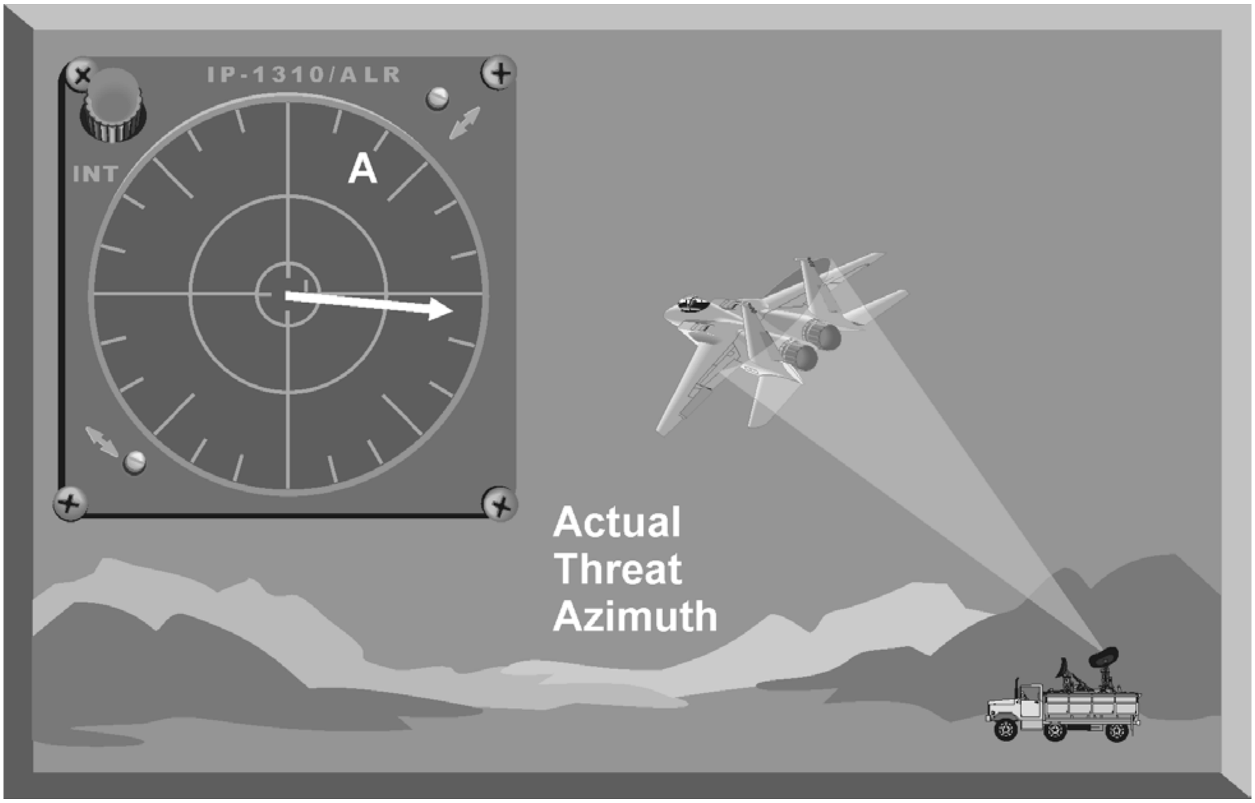

Figure 17-13. RWR Azimuth Error

(3) During aggressive maneuvering, the RWR signal processor may generate multiple symbols for a single threat emitter. In Figure 17-14, as the aircraft maneuvers, the relative azimuth of the threat position changes rapidly causing signal strength detected by each antenna to also change rapidly. The signal processor interprets these changing signals as new and different threat systems with the same signal characteristics. The number of “false” threat symbols displayed for a single threat is determined by the processing speed of the signal processor and a parameter called the symbol “age-out” time. Symbol age-out is the time, normally in seconds, that the RWR will continue to display a threat symbol after the signal processor has determined that the threat is no longer transmitting. The symbol age-out time is set for each threat in the EID tables.

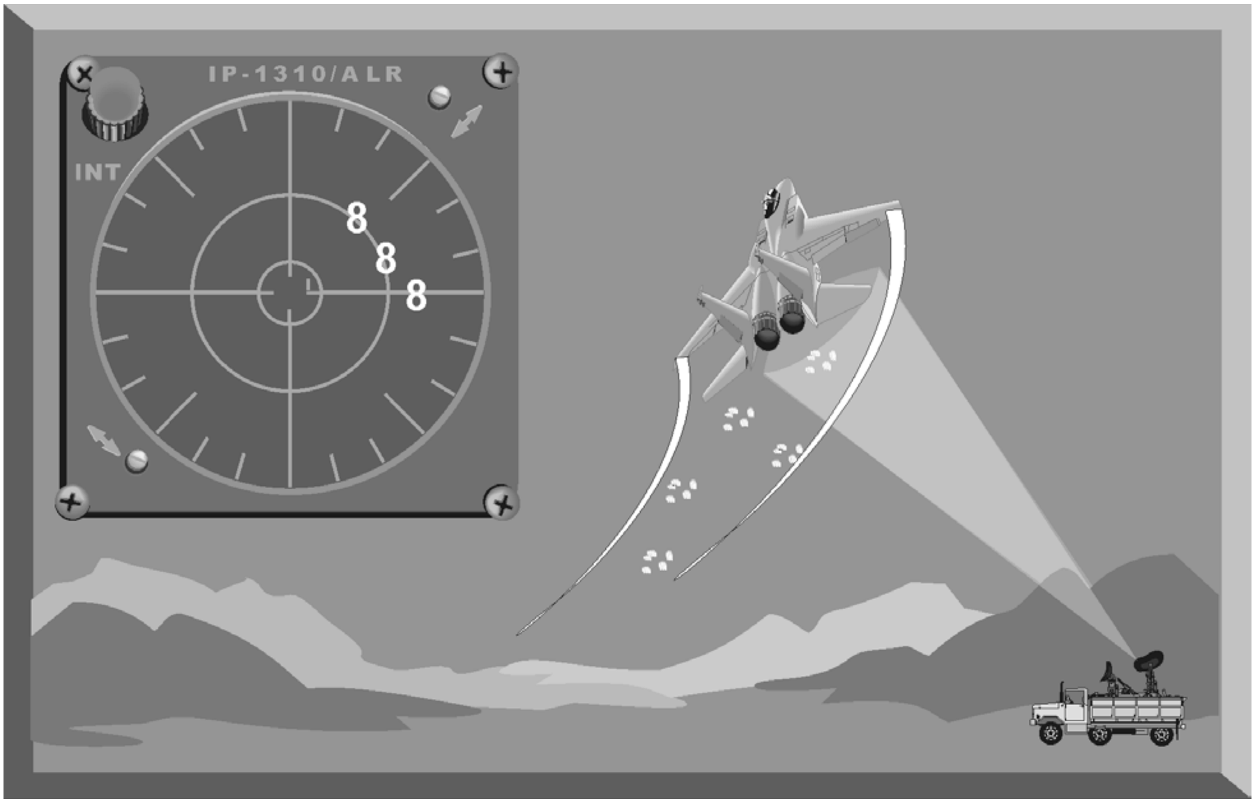

Figure 17-14. RWR Multiple Threat Symbols

c. Electromagnetic interference (EMI) is defined as any electromagnetic disturbance that interrupts, obstructs, or otherwise degrades or limits the effective performance of electronic systems. EMI can be induced intentionally, by way of jamming, or unintentionally as a result of spurious emissions and modulations. Certain RWR characteristics make them susceptible to EMI. Modern RWR systems are designed to receive and process signals in a wide frequency range, nominally 0.5 to 18 GHz, where most threats operate. This broad frequency coverage combined with the sensitive antennas make RWR systems susceptible to EMI. The primary source of EMI that impacts RWR operation is noise and deception jamming designed to counter enemy threat systems.

(1) High power noise jamming, such as that provided by a stand-off jamming aircraft, causes the RWR to raise the receiver threshold in the frequency band of the jamming. This effectively reduces the sensitivity of the RWR receiver and could delay the display of threat signals in that band. The reduction in sensitivity depends on the power that the jammer is transmitting, the beamwidth of the jamming beam, and the distance from the jammer to the aircraft. High power jamming may also generate multiple threat symbols on the RWR scope at the approximate azimuth to the jammer's position relative to the aircraft. Additionally, jamming from a wingman's self-protection system can generate multiple threat symbols and reduced sensitivity.

(2) A limitation of all RWR systems, related to EMI, is the problem of inaccurately identifying threat radar systems as friendly radar systems. This occurs when the parameters of a friendly radar are similar to the parameters of a threat radar system. The RWR will either display a threat radar symbol or an ambiguity associated with a threat radar. These RWR misidentifications are especially prevalent for AI radar systems.

(3) The impact of EMI on the operation of an RWR system depends on the signal environment. The pilot has little control over the number and diversity of friendly and enemy signals the RWR system must process. EMI is an unfortunate consequence of the reliance of modern military forces on operations in the electromagnetic spectrum. Aircrews should be keenly aware that EMI will impact RWR operation and be must be familiar with common RWR displays of EMI.

10. THREAT GEOLOCATION TECHNIQUES

The purpose of threat geolocation is to put a defined position, normally coordinates, on a threat radar. This information can be used to simply warn other aircraft about the threat or, if the coordinates are accurate enough, to allow for targeting and attack of the threat. Until recently, threat geolocation could only be performed by strategic assets and specialized tactical aircraft, specifically the F-4G Wild Weasel. Due to inherent time delays, data provided through strategic channels often did not apply to mobile threat systems. The mobile systems would relocate making the data obsolete. This section will discuss three techniques used to geolocate, also known as direction finding (DF), emitting radars that can be used by tactical assets to rapidly locate radar threat systems. The three are triangulation, interferometry, and time of arrival. All three techniques are heavily dependent upon the receiving aircraft's ability to accurately determine its present position, and the advent of GPS receivers has made this significantly easier.

a. Triangulation is the most basic form of DF available. It involves taking direction measurements from more than one source. The intersection of the azimuth measurements, called “lines of bearing”, is the likely location of the emitter (Figure 17-15). To be effective, the participating aircraft must have accurate data of their current positions when getting the lines of bearing.

(1) Triangulation can be carried out by multiple aircraft equipped with receiver equipment or by one aircraft over a period of time. The advantage of having multiple aircraft providing azimuth measurements is the increased angle-off and the speed of interception. In triangulation the best azimuth cuts are those that approach 90° angles. The speed of interception comes into play because threat emitters, knowing that DF operations are underway, attempt to transmit for as little time as possible. The disadvantage of multiple platforms is the communication required to ensure that all the platforms are measuring the same radar. In a dense radar environment this can be a difficult task.

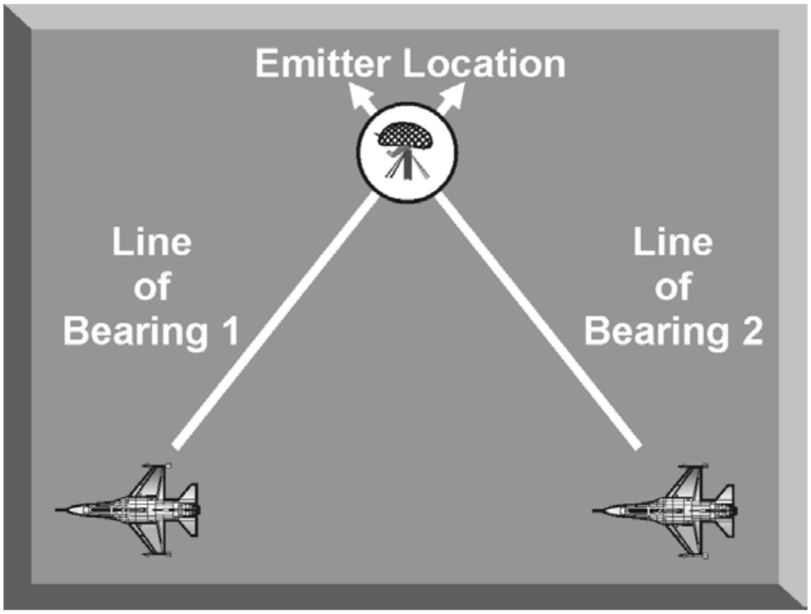

Figure 17-15. Triangulation

(2) Single aircraft triangulation eliminates the problem of signal coordination with other aircraft, but it also requires that the aircraft transit some distance to get multiple azimuth measurements. The accuracy of single ship DF operations is a function of the quality of the receiver equipment, distance away from the targeted radar, speed of the aircraft, and the amount of time that the targeted radar radiates.

b. The second technique is called interferometry. This technique is also known as phase interferometry, or phase difference of arrival. These systems operate by comparing the phase of a radar wave as it impacts two or more DF antennas; this phase difference is then used to compute an angle of arrival (AOA). For aircraft, the desire is to have these DF antennas on different parts of the same aircraft. Multiple AOA measurements are then used to provide the range and position of the threat.

(1) The key elements in an interferometer system are two antennas in fixed locations with matched receivers, a phase comparator, and a processor (Figure 17-16). An intermediate frequency output from each receiver is passed to a phase comparator, which measures the relative phase of the two signals. This relative phase position is passed to a processor, which calculates the AOA relative to the orientation of the two antennas (called the baseline). In most systems, the processor also accepts information about the orientation of the baseline (relative to true North or local horizontal) to determine the true azimuth or elevation angle to the emitter.

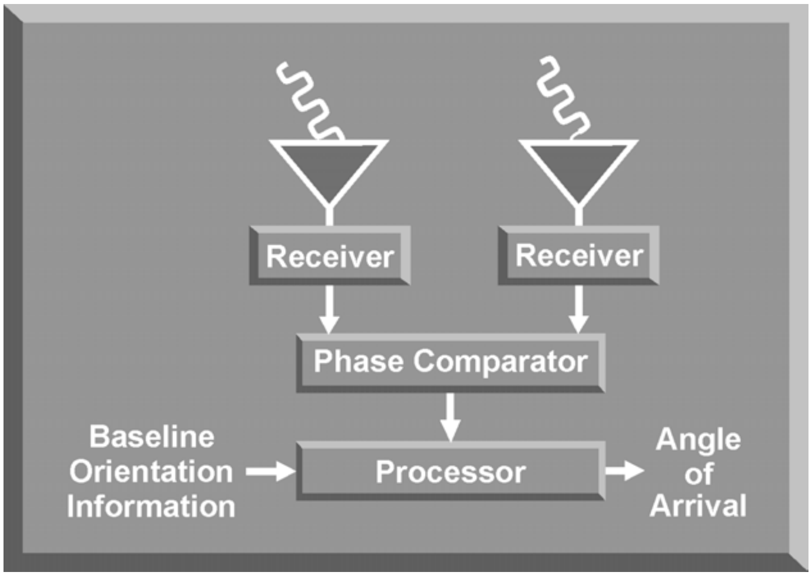
Receiver
Receiver
Phase Comparator
Baseline Orientation Information
Processor
Angle of Arrival

Figure 17-16. Interferometer Components

(2) To obtain rapid emitter locations some interferometer systems use multiple receive antennas in an array setup. This allows for simultaneous azimuth and elevation measurements to rapidly locate the emitter. An array allows for the mixing of long and short baselines in different patterns by selecting different pairs of antennas (Figure 17-17). The terms long baseline and short baseline are often used to designate the distance between the antenna elements in an interferometer system. Long baselines have the advantage of providing a quick and accurate location of the emitter, but they can suffer from ambiguities resulting from different wavefronts hitting the different antenna elements. In addition to the array depicted in Figure 17-17, a long baseline system could be created by using the existing RWR antennas on an aircraft, and supplementing these with a small short baseline system to compensate for the ambiguities.

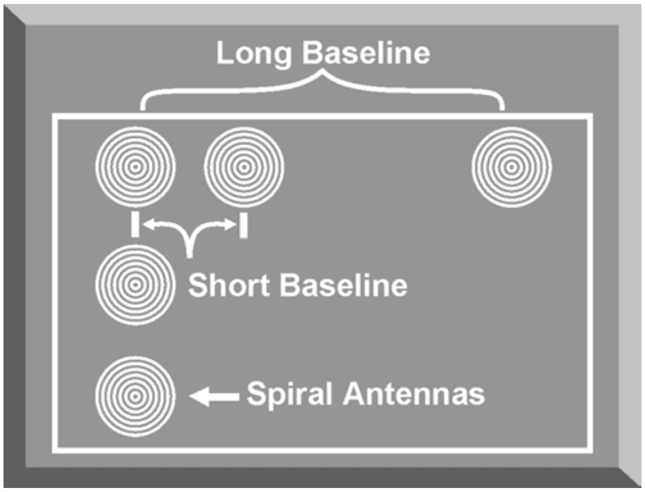
Long Baseline
Short Baseline
Spiral Antennas

Figure 17-17. Interferometer Array

c. Time of arrival (TOA) and time difference of arrival (TDOA) techniques are the final type of location techniques to be covered in this section. Both techniques are based around the fact that radar signals travel at approximately the speed of light. Both techniques, using the speed of light as a constant, solve for the distance that the emitter is away from the receiver using the equation distance equals rate multiplied by time. Because there is not any directional information, the equation represents the radius of a circle around the receiving antenna; multiple distance measurements taken from multiple receivers are then overlayed. The intersection of the circles is the position of the emitter. If only two receivers are used, a simple DF technique can solve the ambiguity of which intersection represents the emitter (Figure 17-18).

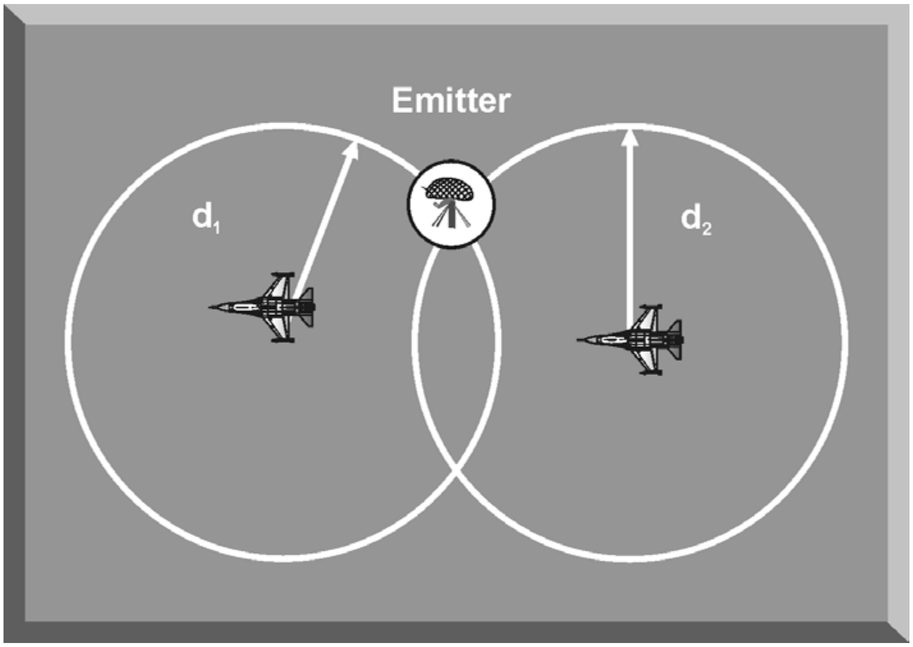

Figure 17-18. Time of Arrival Measurements

(1) TOA positioning is calculated by taking the time that a signal leaves a radar, measuring its arrival time at the receiver and mathematically solving for distance using the techniques described above. One of the primary challenges involved with TOA positioning is determining when the measured signal was transmitted, this requires either a very cooperative enemy or a radar signal with some type of exploitable time reference. Another challenge involves insuring that multiple receivers are timing the same signal. This is especially difficult in a threat intensive environment.

(2) TDOA is used when it is not possible to determine when a signal was transmitted. TDOA uses most of the same principles as TOA except that it must compensate for not knowing the time the signal was transmitted. It does this by using an extra receiver a known distance from the first receiver to generate a distance curve. The arrival of the signal is precisely measured at the two

receivers, the theory is that a signal that arrives at the two receivers at times t1 and t2 had to originate from a point on a curve defined by that difference in arrival time. For example, if the signal arrives at the two points at exactly the same time, then the transmitter must be equal distance from the two receivers. In this case the curve is actually a line of possible locations equal distance from the receivers and can easily be drawn. For situations where the signal arrives at different times at the receivers, a hyperbolic curve instead of a straight line denotes all the possible locations of the transmitter. Figure 17-19 shows an example of when the signal arrives at a different time, the constant value is the time difference multiplied by the speed of light yielding a distance. To solve for the emitter's location using TDOA another antenna receiver is required that is not in line with the first two receivers to generate an independent curve that will cross the first curve at the transmitter location.

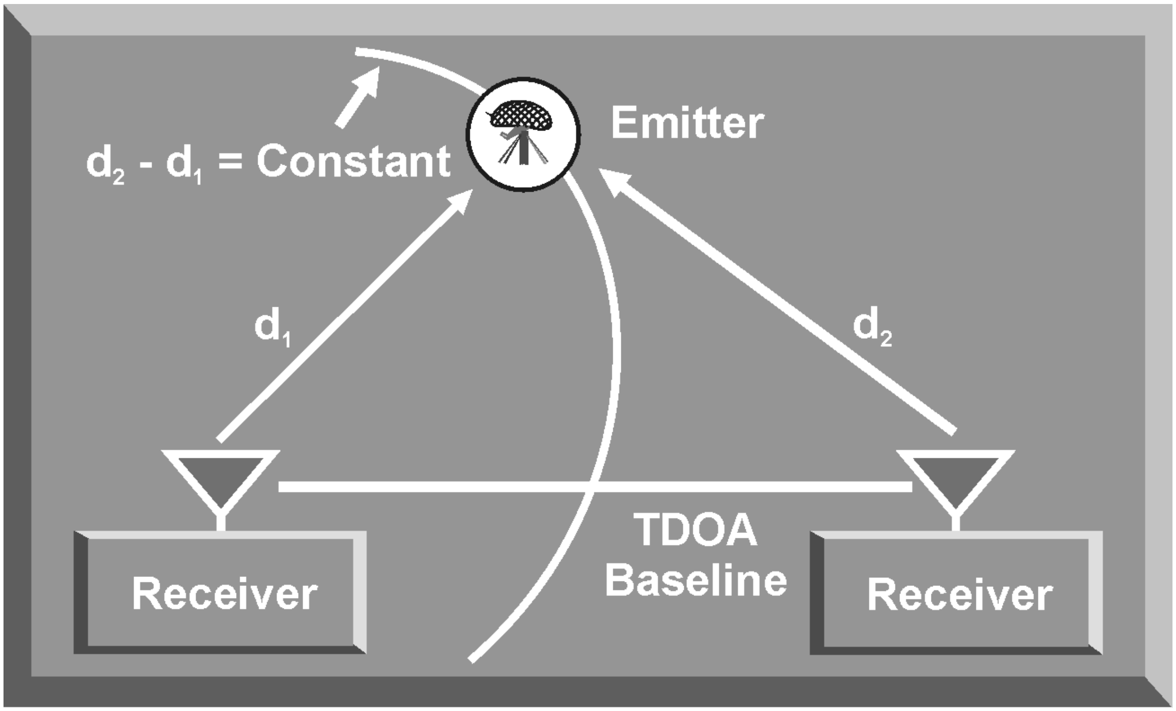

Figure 17-19. Time Difference of Arrival Measurements

(3) Timing techniques require very good timing accuracy in the receiving systems. If the receivers are far apart, such as different aircraft, then a separate time measurement is required at the source before sending to a processor. As with triangulation, one of the major challenges of multiple aircraft measurements is the coordination to ensure that the same signal is being looked at by all aircraft involved.

(4) Timing techniques are affected by the type of radar that they are trying to locate. Pure CW radars are not practical for a timing technique because there

is not a pulse to time. Pulsed signals, on the other hand, are much more susceptible to this type of location technique.

11. SUMMARY

An RWR system is designed to provide a picture of the electronic order of battle (EOB) operating in the vicinity of an aircraft. The signal processor is the heart of the RWR system and controls the function of all other system components. The signal processor, using inputs from the antennas and the receiver/amplifiers, compares the signal parameters with the parameters in the EID tables. The identified signals are displayed on the RWR scope with the appropriate audio. The ICU provides the aircrew interface with the signal processor to allow the aircrew to customize RWR operation for each combat mission. Modern RWR systems have some limitations that can effect aircrew survival. These limitations include ambiguous threat displays, maneuvering limitations, and EMI. Despite these limitations, an operational RWR system is one of the keys to survival on today's electronic battlefield. A product of advances on the electronic battlefield is rapid threat geolocation. New threat geolocation techniques are designed to provide advanced threat information to allow pilots to avoid, suppress, or destroy the mobile threat. The three most common techniques are basic triangulation using lines of bearing, interferometry using phase difference of arrival, and time difference of arrival using time differences to determine distance curves.

CHAPTER 18. SELF-PROTECTION JAMMING SYSTEM OPERATIONS

1. INTRODUCTION

Self-protection jamming systems are designed to counter surface-to-air missile (SAM), airborne interceptor (AI), and antiaircraft artillery (AAA) acquisition and target tracking radars. Self-protection jamming systems generate noise and deception jamming techniques to either deny threat system automatic tracking capability or generate sufficient tracking errors to prevent a successful engagement.

a. To counter the diverse array of threats and their associated frequencies in an integrated air defense system (IADS), a self-protection jamming system must be able to simultaneously jam multiple signals operating in a wide frequency range. The system must also be able to generate sufficient power to mask the radar return of the aircraft. Since many modern self-protection jamming systems are carried internally or externally on aircraft, their size, shape, and weight must be carefully controlled to minimize adverse effects on aircraft performance and handling. These design requirements may require a trade-off between additional power or capability and aircraft compatibility.

b. A self-protection jamming system must have receive antennas to receive radar signals, a system processor to identify received signals, a jamming techniques generator to produce an optimum jamming technique for the identified threat, and transmit antennas to transmit the required jamming techniques. These system components will be discussed in this chapter.

c. There are numerous internal and externally mounted self-protection jamming systems in use today. To simplify the discussion, the ALQ-184 pod will be used as an example of a modern, self-protection jamming system (Figure 18-1). The functions and operations of the ALQ-184 components are representative of currently deployed self-protection systems.

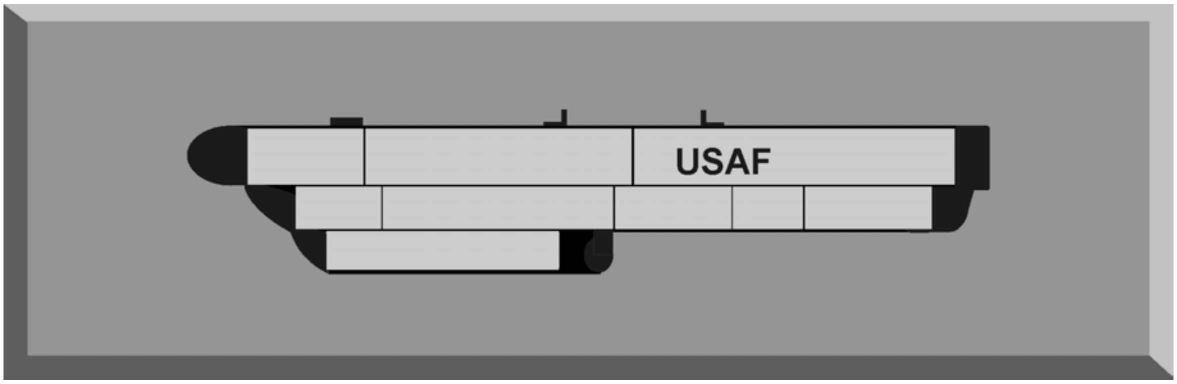

Figure 18-1. ALQ-184 Pod

2. RECEIVE ANTENNAS

The ALQ-184 pod has two sets of receive antenna assemblies located on the front and rear of the lower pod "gondola" (Figure 18-2). Each receive antenna assembly set consists of a low-band antenna and a mid/high-band antenna. The low-band antenna's signals are combined to form the input for the low-band receiver. The mid/high-band antennas provide signals to the separate mid/high-band portion of the pod.

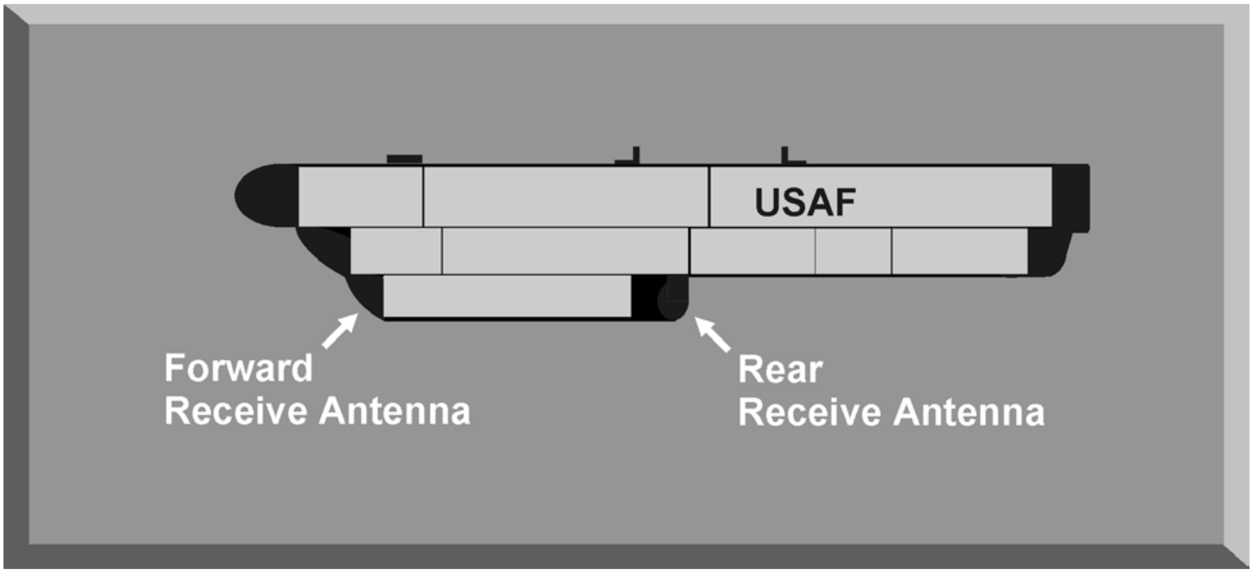

Figure 18-2. ALQ-184 Receive Antenna Location

a. The low-band receive antennas are circularly polarized spiral arrays. The mid/high-band receive antenna covers eight sub-bands via an eight element array that corresponds to the eight element transmit array. The mid/high-band antenna is also circularly polarized.

b. Each signal received via the receive antenna generates a corresponding transmit signal through a crystal video receiver. The angle of arrival (AOA) of each signal is determined by comparing the output of the crystal video receiver for each antenna.

3. RECEIVER SECTION

The forward and aft receive antennas detect radar signals and send them to the receiver section. The receiver section contains an AOA receiver and a multiplexer that separates frequencies. The AOA receiver determines the AOA of a signal based on the output signal level of each antenna element. The AOA data is passed to the system processor to control the angle of transmission of the jamming signal. The multiplexer separates all received threat signals into eight frequency sub-bands consisting of five mid-band and three high-band frequency

ranges. The multiplexer categorizes the received radar signals by sub-band and azimuth and then sends them to the system processor and the exciter.

4. SYSTEM PROCESSOR

The system processor is the "brain" of the ALQ-184 pod. It receives the threat signals from the receiver section that have already been categorized by sub-band and distinguished by AOA. Each received signal and its corresponding AOA is counted. When the signal count exceeds a pre-set threshold, the system processor validates the signal and identifies it using the emitter identification (EID) tables. Based on this threat identification, the system processor directs the techniques generator to initiate a jamming program through the exciter. At the same time, the system processor, through a signal switch control, directs the transmitter section to use either the forward or aft transmit antenna array, and specifies the transmit angle for the jamming program based on AOA.

a. The system processor also controls the low-band portion of the pod (Figure 18-3). The forward or aft low-band antenna receives and combines the threat signal and sends it to the multiplexer. The multiplexer channelizes this signal into one of two low sub-bands. The signal then passes to the modulator where the voltage controlled oscillators (VCOs) generate a jamming technique as directed by the system processor techniques generator. This ensures the jamming technique is at the proper frequency and modulation to provide maximum effect against the threat system. The generated jamming signal passes to a second multiplexer and then to a solid state amplifier for amplification. The amplified jamming signal is then sent to the two low-band traveling wave tubes (TWTs) and transmitted from the forward and aft low-band transmit antennas.

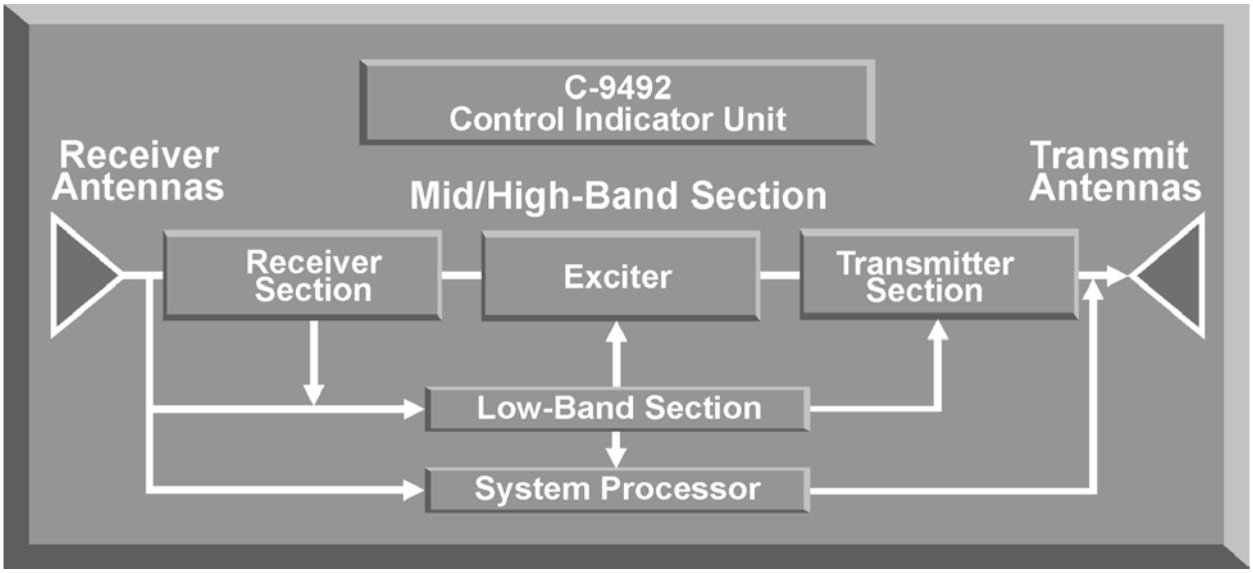

Figure 18-3. ALQ-184 Simplified Block Diagram

b. The system processor periodically performs a background built-in-test (B-BIT), to check the status and calibration of the ALQ-184. The system processor software is updated or reprogrammed as required from a memory loader verifier (MLV). The system processor uses two major software programs. The operational flight program (OFP) manages the receiver and transmitter functions, signal processing, and techniques generation. The mission data generator (MDG) contains the threat EID tables and the jamming techniques matrix for these threats.

5. JAMMING TECHNIQUES GENERATOR

The exciter section contains the VCOs and the keyed oscillators, which generate the jamming waveforms that are directed by the system processor. The exciter takes the threat signal from the receiver and modifies it with deception modulation or generates a noise program based on the techniques generator in the system processor. The selected jamming technique is then sent to the transmitter section.

6. TRANSMIT ANTENNAS

The transmitter section contains antenna switching circuits, a signal forming network, and TWTs. The antenna switching circuits are controlled by the system processor to ensure the transmitted jamming pulse is radiated at the proper azimuth. The signal-forming network controls the phase of each jamming signal. There are sixteen TWTs in the transmitter section, one for each front and aft antenna. The TWTs amplify the jamming signal for transmission from the appropriate transmit antenna.

a. The ALQ-184 has two sets of transmit antenna assemblies. One set is located on the front and one on the rear of the main pod assembly (Figure 18-4).

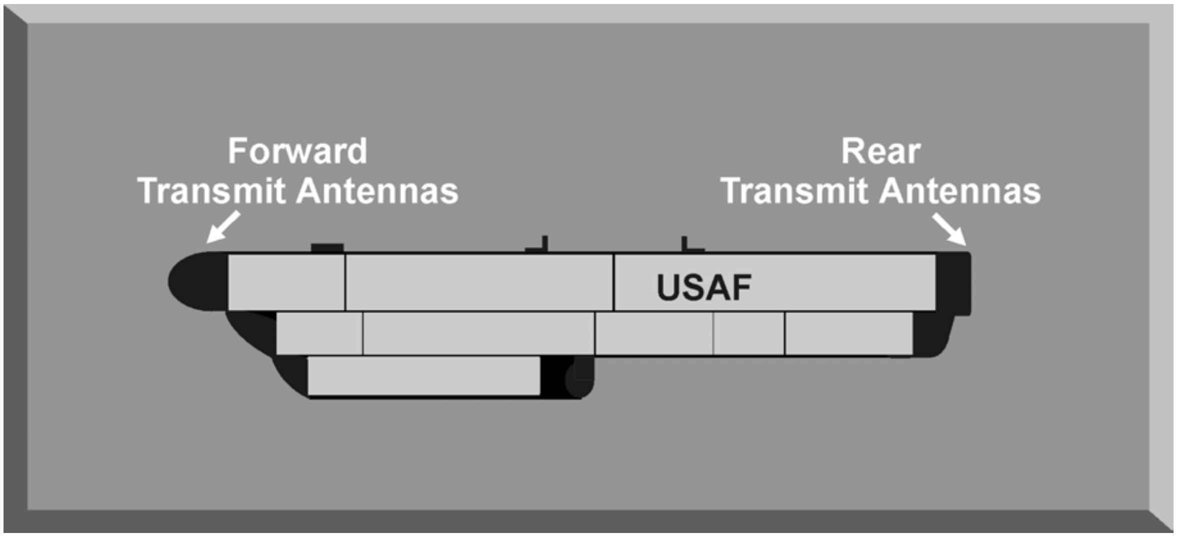

Figure 18-4. ALQ-184 Transmit Antenna Location

Each set has a low-band and mid/high-band antenna. The transmit antennas have opposite polarization from the receive antennas. This enables the pod to transmit and receive at the same time while preventing pod "ringing." Ringing occurs when the pod receives its own jamming signal and generates a jamming program to counter its own signal. This condition can highlight the aircraft to enemy radar and seriously degrade pod effectiveness.

b. The low-band transmit horn antennas are circularly polarized. The mid/high-band transmit antennas have the same eight-element array as the receive antennas. The transmit antennas generate a directional jamming signal for pulsed radar signals based on the AOA determined by the receive antennas. Detailed information on the antenna jamming pattern is contained in the ALQ-184 ECM Handbook.

7. C-9492 CONTROL INDICATOR UNIT

The C-9492 Control Indicator Unit (CIU) provides control of all functions of the ALQ-184. It allows the aircrew to control system power, select jamming techniques, and monitor system status (Figure 18-5).

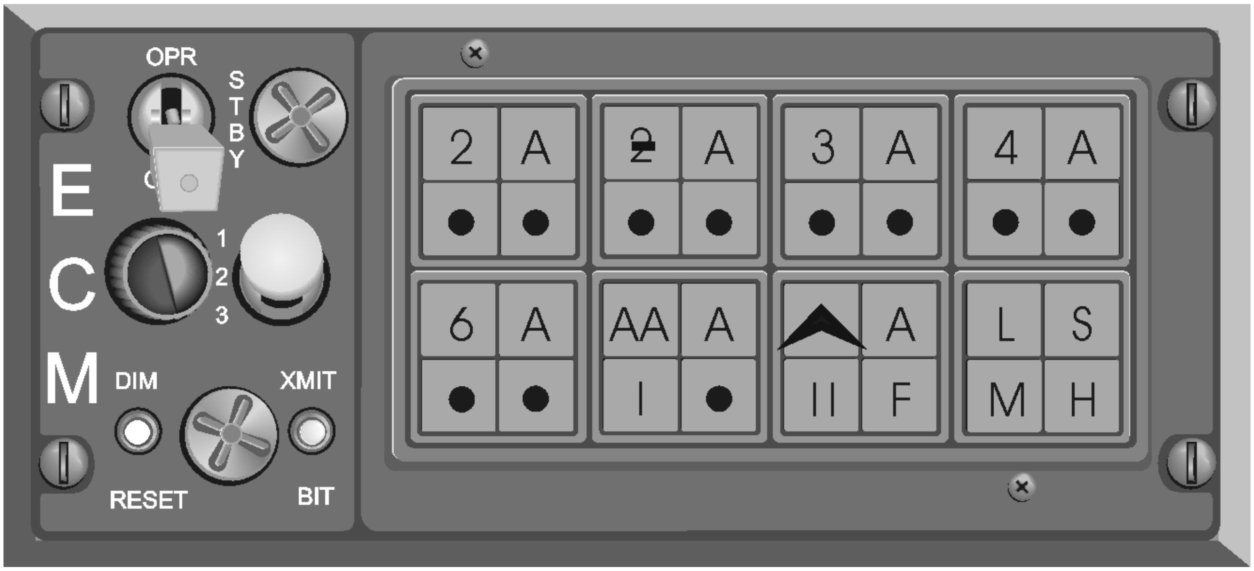

Figure 18-5. C-9492 Control Indicator Unit

8. JAMMING POD CONSIDERATIONS

Unfortunately, jamming pods do not make the aircraft completely invisible from enemy threat systems. When used in conjunction with the RWR, expendables, and threat countertactics they do drastically increase the probability of surviving a threat engagement. This section contains some items of consideration to ensure that your self-protection jamming is being properly utilized. These considerations include highlighting, burnthrough, and reprogramming.

a. If used improperly a self-protection pod can actually highlight the aircraft. This happens when a pod transmits a jamming program constantly or when another type of radar, not targeted for jamming but in the same frequency range, sees the jamming. Highlighting is normally associated with noise programs that transmit constantly. Most pods today avoid this by transmitting only when a threat requires it. Additionally, some aircraft have pod inhibit switches that allow the pod to remain in a standby condition until the pilot switches the pod to operate. What makes this system effective is that the switch is either on the stick or throttle allowing the pilot to rapidly go to operate without looking inside the cockpit (Figure 18-6).

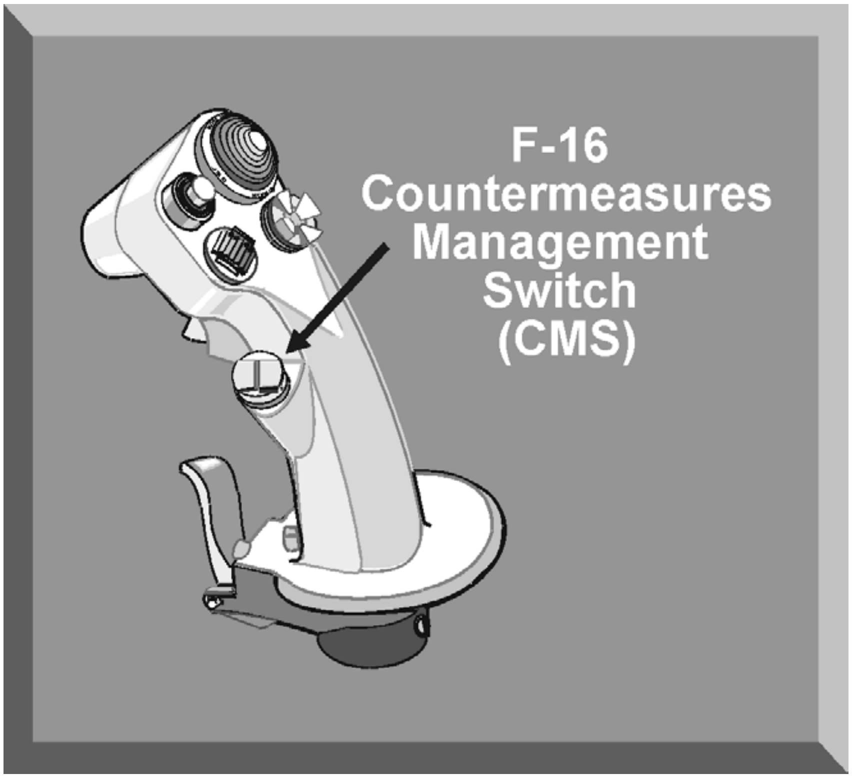

Figure 18-6. F-16 ECM CMS

b. Burnthrough is described as the range that the aircraft's radar return is stronger than its jamming signal. Basically, the jet has gotten so close to the threat radar that the threat radar is overpowering the jamming pod. The range that this occurs varies for different radars, but it should be planned for particularly when there are threats in the target area. Burnthrough is also a consideration for support jamming systems; there is a range at which the support jammer's signal will be overpowered and the protected aircraft can be seen.

c. Many self-protection jamming systems are what is called software-reprogrammable. This means that jamming programs can be rapidly changed to adjust for changes in enemy radar systems. These updates can be made to

systems while they are still on the jet, and can take as little as 15 minutes using an MLV. The process of changing electronic warfare systems is known as a Pacer Ware change. An exercise to check the ability to make changes is known as a Serene Byte exercise. The pilot of the aircraft carrying the self-protection system has one responsibility in this process: to ensure that his jamming system is carrying the most current software program, usually referred to as a software version. While Pacer Ware changes are not common during peacetime operations, they do happen quickly during combat operations often to fine tune the jamming systems for the particular theater of combat.

9. SUMMARY

The effective employment of self-protection jamming is one of the factors that can mean the difference between success and failure on the modern battlefield. When employed in concert with chaff and maneuvers, self-protection jamming can be extremely effective in negating potentially lethal attacks. Despite their limitations, self-protection jamming systems are an important part of the "bag of tricks" aircrews can employ to put bombs on target.

ANNEX A
GLOSSARY

Absorption - Dissipation of energy of electromagnetic waves, sound, and light waves into other forms of energy because of interaction with matter. Absorption characteristics of specific materials are used as blankets, coatings, or structural and surface materials for aircraft to reduce effective radar cross-sections.

Acoustic jamming - The deliberate radiation of mechanical or electro-acoustic signals with the objectives of obliterating or obscuring signals which the enemy is attempting to receive, and of deterring enemy weapon systems.

Acquire –

1. When applied to acquisition radars: the process of detecting the presence and location of a target in sufficient detail to permit identification.

2. When applied to tracking radars: the process of positioning a radar beam so that a target is in that beam to permit the effective employment of weapons.

Acquisition radar - A radar set that detects an approaching target and feeds approximate position data to a fire control or missile guidance radar that then takes over the function of tracking the target.

Active array radar - A phased array radar in which each radiating element contains a transmitter and receiver front end, as opposed to a single transmitter/receiver serving all phased array elements. Advantages attributed to active array radars include efficient use of prime power, no waveguide losses, very wide bandwidth, and extreme reliability due to the lack of single point of failure.

Active homing guidance - A system of homing guidance in which both the source for illuminating the target, and the receiver for detecting the energy reflected from the target as the result of illumination, are carried within the missile.

Aerosols - Solid particles dispersed in the atmosphere that have resonant size particles with a high index of refraction. The particles both scatter and absorb visual and laser-directed energy so as to lessen the effect of weapon systems directed by these techniques.

Airborne early warning and control - Air surveillance and control provided by airborne early warning vehicles that are equipped with search and height finding radar and communications equipment for controlling weapons.

Airborne interceptor (AI) - A manned aircraft used for identification and/or engagement of airborne objects. (An AI may or may not be equipped with radar to assist in the interception.)

Airborne Warning and Control System (AWACS) - An aircraft suitably equipped to provide control, surveillance, and communications capability for strategic defense and/or tactical air operations.

Air defense - All defensive measures designed to destroy attacking enemy aircraft or missiles in the earth's envelope of atmosphere, or to nullify or reduce the effectiveness of such attack.

Air surveillance radar (ASR) - A radar displaying range and azimuth that is normally employed in a terminal area as an aid to approach and departure control.

Air-to-air missile (AAM) - A missile launched from an airborne carrier at a target above the surface.

Air-to-surface missile (ASM) - A missile launched from an airborne carrier to impact on a surface target.

Alternating current (AC) - An electric current that reverses its direction at regularly recurring intervals, the frequency being determined by the frequency of the alternator supplying the current.

AM-CW jamming - Jamming in which a carrier wave is modulated at a constant recurring rate. The recurrence rate of amplitude modulation is variable, and reflects a noticeable change in the radar scope. Because of the spiraling or chaining effect produced, AM-CW jamming is easily identified.

Amplifier - An electronic circuit usually used to obtain amplification of voltage, current, or power.

Amplitude shift keying - A method of impressing a digital signal upon a carrier signal by causing the carrier amplitude to take different values corresponding to the different values of the digital signal.

Amplitude modulation (AM) - A method of impressing a message upon a carrier signal by causing the carrier amplitude to vary proportionally to the message waveform.

Angle jamming - A deception jamming technique used to deny azimuth and elevation information to a TTR by transmitting a jamming pulse similar to the radar pulse, but with modulation information out of phase with the returning target azimuth modulation information.

Angle tracking - Accomplished through the use of pulses to determine angular definition of a target.

Antenna - A device used for transmitting or receiving RF energy. The function of the antenna during transmission is to concentrate the radar energy from the transmitter into a shaped beam that points in the desired direction. During reception, or listening time, the function of the antenna is to collect the returning radar energy, contained in the echo signals, and deliver these signals to the receiver. Radar antennas are characterized by directive beams that are usually scanned in a recognizable pattern. The primary radar antenna types in use today fall into three categories: parabolic, Cassegrain, and phased array antennas.

Antenna cross talk - A measure of undesired power transfer through space from one antenna to another. Ratio of power received by one antenna to power transmitted by the other usually expressed in decibels.

Antenna gain - See Gain

Antenna polarization - See Polarization

Antenna sidelobes - See Sidelobes

Antiaircraft artillery (AAA) - Guns used to shoot unguided projectiles at airborne aircraft. Usually used in the air defense system.

Anti-clutter circuits (in radar) - Circuits that attenuate undesired reflections to permit detection of targets otherwise obscured by such reflections.

Antiradiation missile (ARM) - A missile that homes passively on a radiation source.

Area defense - The concept of locating defense units to intercept enemy attacks, remote from, and without reference to, individual vital installations, industrial complexes, or population centers.

Asynchronous pulsed jamming - An effective form of pulsed jamming where the jammer nearly matches the PRF of the radar then transmits multiples of the PRF. It is more effective if the jammer pulse width is greater than that of the radar. Asynchronous pulsed jamming is similar to synchronous jamming except that the target lines tend to curve inward or outward and appear fuzzy in the jammed sector of the radar scope.

Attenuation - The decrease in amplitude of a signal during its transmission from one point to another. Attenuation increases as distance increases. The higher the frequency of the propagating signal, the higher the rate of attenuation.

Automatic frequency control (AFC) - Circuits in a receiver that automatically correct the local oscillator frequency to prevent receiver drift in tuned frequency.

Automatic gain control (AGC) –

1. A feature involving special circuitry designed to maintain the output of a radio, radar, or television receiver essentially constant, or to prevent its exceeding certain limits, despite variations in the strength of the incoming signal. In a radio receiver, in particular, though something of a misnomer, also known as automatic volume control.

2. A self-acting compensating device that maintains the output of a transmission system constant with narrow limits, even in the face of wide variations in the attenuation of the system.

3. A radar circuit that prevents saturation of the radar receiver by long blocks of receiver signals, or by a carrier modulated at low frequency.

Automatic search jamming - An intercept receiver and jamming transmitting system that automatically searches for and jams enemy signals of specific radiation characteristics.

Automatic tracking - Tracking in which a system employs some mechanism, e.g., servo or computer, to automatically follow some characteristics of the signal.

Azimuth resolution - The ability of a radar to distinguish two targets in close azimuth proximity and distance.

Backlobe - The portion of the radiation pattern of an antenna that is oriented 180° in relation to the main beam. The antenna backlobe is a result of diffraction effects of the reflector and direct leakage through the reflector surface.

Backward wave oscillator (BWO) - A special traveling wave tube in which oscillatory currents are produced by using an oscillatory electromagnetic field to bunch the electrons as they flow from cathode to anode.

Ballistic missile - Any missile that does not rely upon aerodynamic surfaces to produce lift and consequently follows a ballistic trajectory when thrust is terminated.

Bandpass filter - A filter that allows a select range of frequencies to pass while attenuating all frequencies outside the range.

Bandwidth - The range of frequencies within which performance, with respect to some characteristics, falls with specific limits (i.e., the width of frequency of a barrage noise package).

Barrage noise jamming - Noise jamming spread in frequency to deny the use of multiple radar frequencies to effectively deny range information. Although this is attractive because it enables one jammer to simultaneously jam several radars of different frequencies, it does have the inherent problem that the wider the jamming spread, the less jamming power available per radar.

Beam rider - A missile guided by an electronic beam.

Beam-to-beam correlation (BBC) - Used by frequency scan radars to reject pulse jamming and jamming at swept frequencies. Correlation is made from two adjacent beams (pulses). The receiver rejects those targets (signals) that do not occur at the same place in two adjacent beams.

Beamwidth - The width of a radar beam measured between lines of half-power points on the polar pattern of the antenna. This width is measured at the 3 dB points.

Beat frequency oscillator (BFO) - Any oscillator whose output is intended to be mixed with another signal to produce a sum or difference beat frequency. Used particularly in reception of CW transmissions.

Bistatic radar - A radar where the transmitting and receiving antennas are separated by a considerable distance. Bistatic operation provides several advantages for its user. The covert positioning of the receivers poses problems for a potential attacking force since ELINT techniques locate the transmitter not the receiver. The proper placement of jamming assets is difficult, since the receiving sites are unknown. In addition, if a stand-off jammer is directed at the transmitter, its effectiveness in the direction of the covert receiver is diminished.

Jammers not in the same beam as the wanted targets will be attenuated by the receiver's sidelobe protection and these targets will be more readily detected.

Blanking –

1. The cutting off of the electronic beam in a cathode-ray tube when the picture is not being formed. The beam is blanked while the spot is returning to the starting position (normally right to left).

2. Process of making a channel or device noneffective for a desired interval.

Blinking - A jamming technique employed by two aircraft separated by a short distance and within the same azimuth resolution to appear as one target to a tracking radar. The two aircraft alternately spot jam, causing the radar system to oscillate from one plane to another, making an accurate solution of fire control problem impossible. However, keep in mind that too high a blinking frequency can cause the tracker to average the data while too low a frequency will cause a missile to home-in on one of the jammers.

Broad pulse jamming - Transmission of broad pulses for control system jamming when little is known about the command pulse group. For example, a broad pulse might cover a whole group of command pulses, thus jamming that command.

Burnthrough range - The ability of a radar to see through jamming. Usually described as the point when the radar's target return is stronger than the jamming signal.

Burst chaff - The formation of a reflective volume of chaff from an individual bundle.

Buzzer - Code name for airborne jamming.

Capture - Where the jammer takes control over the guidance signal by active jamming.

Capture effect - The tendency of a receiver to suppress the weaker of two signals within its bandpass.

Capture of AGC - Domination of the radar's automatic gain control (AGC) level by a strong transmitted jamming signal.

Carrier frequency - Frequency of an unmodulated radio wave emanated from a radio, radar, or other type of transmitter.

Carrier wave - Electromagnetic radiation used to “carry” information from one point to another.

Centroid homing - When applied to antiradiation missiles, the effect on a missile that has two or more radiation sources in its field of view causing it to home on the centroid of the power from the radiating sources.

Chaff - Ribbon-like pieces of metallic materials or metalized plastic that are dispensed by aircraft to mask or screen other aircraft or to cause a tracking radar to break lock. The foil materials are generally cut into small pieces for which the size is dependent upon the radar interrogation frequency (approximately 1/2 wavelength of the victim radar frequency). Being 1/2 wavelength long, chaff acts as a resonant dipole and reflects much of the energy back to the radar.

Chaff corridor - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a “ribbon” or “stream” of returns many miles long on radar scopes. The penetrating strike force can then use the resulting chaff corridor to mask its penetration.

Chirp - A pulse compression technique characterized by linear frequency modulation on pulse (LFMOP).

Chirp radar - Radar in which a swept-frequency signal is transmitted, received from a target, and then compressed in time to give a final narrow pulse called the chirp signal. It has high immunity to jamming and an inherent rejection of random noise signals.

Circularly polarized jamming - The techniques of radiating jamming energy in both planes of polarization. With this method, there is a 3-dB loss of effective power in either plane, but the enemy cannot cross-polarize his antenna to escape jamming.

Circular scan - The pattern generated by an antenna that is continuously rotating in one direction.

Clipped noise modulation - A clipping action is performed to increase the. bandwidth of the jamming signal. Results in more energy in the sidebands, correspondingly less energy in the carrier, and an increase in the ratio of average power to peak power.

Clutter - Unwanted signals, echoes, or images on the face of a scope that interferes with the observation of desired signals. Also called noise. This tends to mask the true target from detection or cause a tracking radar to break lock.

Clutter elimination - The clutter eliminator circuit discriminates against any target echo that exceeds three times the transmitted pulse width, and will not display it on the indicator. It is normally employed on the lower beams of a high frequency radar. This will allow targets above a preset signal strength to be presented, while the clutter (land) will be eliminated.

Clutter gating - A technique that provides switching between MTI and normal videos. It results in the normal video being displayed in regions with no clutter and the MTI video being switched in only for the clutter areas. Clutter gating is achieved automatically by the PW discrimination or the use of storage tubes. It also can be achieved by a manually operated range azimuth gate. The clutter gate vastly increases the effectiveness of noncoherent MTI against chaff.

Coherent MTI (in radar MTI) - A system in which the target echo is selected based on its Doppler frequency when compared to a local reference frequency maintained by a coherent oscillator.

Coherent noise jamming - Noise-like jamming signal that is repetitive and in synch with the PRI of the radar.

Coherent repeater jammer - A jammer that uses the phase information of the receiver radar signal in creating false targets thus transmitting a signal that is acceptable to the receiver processor of a victim coherent radar.

Command and control warfare (C2W) - The integrated use of operations security (OPSEC), military deception, psychological operations (PSYOP), electronic warfare (EW), and physical destruction, mutually supported by intelligence to deny information, influence, degrade, or destroy adversary C^2 capabilities while protecting friendly C^2 capabilities.

Command, control, communications, and computer systems (C4) - The process of, and means for, the exercise of authority and direction by a properly designated commander over assigned forces in the accomplishment of the commander's mission.

Command guidance - A guidance system in which intelligence transmitted to the missile from an offboard source causes the missile to traverse a directed flight path.

Communications intelligence (COMINT) - Intelligence derived from the interception of enemy communications signals.

Communications security (COMSEC) - The protection resulting from all measures designed to deny unauthorized persons information of value that might be derived from the possession and study of telecommunications, or to mislead unauthorized persons in their interpretation of the results of such possession and study. COMSEC includes: 1. Cryptosecurity; 2. transmission security; 3. emission security; and 4. physical security of communications security material and information.

1. **Cryptosecurity** - The component of communications security that results from the provision of technically sound cryptosystems and their proper use.

2. **Transmission security** - The component of communications security from which all measures designed to protect transmissions from interception and exploitation by means other than cryptoanalysis.

3. **Emission security** - The component of communications security that results from all measures taken to deny unauthorized persons information of value that might be derived from intercept and analysis of compromising emanations from crypto equipment and telecommunications systems.

4. **Physical security** - The component of communications security that results from all physical measures necessary to safeguard classified equipment, material, and documents from access thereto or observation thereof by being within a friendly power.

Complex pulse - A pulse train in which there is more than one pulse width and/or more than one pulse repetition interval.

Conformal antenna - An antenna which conforms to a surface whose shape is determined by considerations other than electromagnetic, for example, an antenna shaped to aerodynamically fit the side of an aircraft.

Conical scan (CONSCAN) - A type of scanning in which the axis of the RF beam is tilted away from the axis of the reflector and rotated about it, thus generating a cone.

Constant false alarm rate (CFAR) receiver - A radar receiver with automatic detection circuits designed to produce a constant number of erroneous target detections independent of noise level at the receiver input. CFAR techniques are intended to prevent receiver saturation and overload, to present clear video information to the display, and a constant noise level to an automatic detector. A device that accomplishes these objectives may respond to the signal-to-noise ratio, for example, rather than the absolute signal level above a fixed threshold. CFAR does not usually permit the detection of a target if the target is weaker than the jamming, but it does attempt to remove the confusing jamming effects.

Continuous wave (CW) - A constant, single-frequency, unmodulated carrier wave that is transmitted and then reflected. A good system for determining velocity.

Correlation detection (modulation systems) - Detection based on the averaged product of the received signal and locally generated function possessing some known characteristics of the transmitted wave. Note the following caveats:

1. The averaged product can be formed, for example, by multiplying and integrating, or by using a matched filter whose impulse response, when reversed in time, is the locally generated function.

2. Strictly, the above definition applies to detection based on cross-correlation. The term correlation detection may also apply to detection involving autocorrelation, in which case the locally generated function is merely a delayed form of the received signal.

Countdown - A technique for forcing the radar AGC to continuously change value and oscillate. The jammer rapidly changes the duty cycle of the deception pulses.

Countdown blink - Self-screening AGC deception using a gated repeater or noise source; the period of which is short compared to the victim's AGC time constant.

Cover pulse - A jammer covers the radar return with an AM pulse usually much wider than the radar pulse. Since tracking circuits are looking for largest return, they will transfer to the cover pulse, thereby denying range information.

Cross-eye - A jamming technique used to produce angular errors in monopulse and other passive lobing radars. Jammer is a two-source interferometer that causes the phase front of the signal reaching the radar to be highly distorted. With such a technique, it is difficult for the radar to determine the points from which the transmissions are originating. Requires a high jam-to-signal ratio or the skin echo will show up in the pattern nulls.

Cross-gated CFAR - A CFAR technique employed to achieve the fast switching required for an optimum combination of normal and MTI modes. Here, the MTI video signals are used to “gate” on the normal video when the MTI indicates a target in clutter. CFAR action is achieved by the wideband as in the zero-crossing and Dicke fix CFARs.

Cross polarization - or “Cross Pole,” is a monopulse jamming technique where a cross-polarized signal is transmitted to give erroneous angle data to the radar. The component of the jamming signal with the same polarization as the radar must be very small.

Cross polarization jamming - A technique whereby the received signal is retransmitted cross-polarized and at a much higher power than the skin return-effective against some monopulse radars by generating erroneous angle information.

CW jamming - The transmission of constant-amplitude, constant-frequency, unmodulated jamming signals to change the signal-to-noise ratio of a radar receiver.

Data link - A communications link which permits automatic transmission of information in digital form.

Deception - Those measures designed to mislead the enemy by manipulation, distortion, or falsification of evidence to induce him to react in a manner prejudicial to his interests. (See Electronic Deception, or Manipulative Deception.)

Deception jamming - Any means of jamming consisting of false signals that have similar characteristics to the victim radar thereby deceiving the operator into erroneous conclusions.

Decibel (dB) - A dimensionless unit for expressing the ratio of two values, the number of decibels being 10 times the logarithm to the base 10 of a power ratio, or 20 times the logarithm to the base 10 of a voltage of current ratio. A power increase by 3 dB indicates a doubling of the original power.

(dBm) - Same as dBw except the reference level is one milliwatt instead of one watt.

(dBw) - Unit used to describe the ratio of the power at any point in a transmission system to a referenced level of one watt. The ratio expresses decibels above and below the reference level of one watt.

Defense suppression - A term applied to weapons systems that are intended to eliminate or degrade enemy detection, acquisition, or tracking equipment.

Delayed opening chaff - Chaff that blooms at a specific elapsed time after it is dispensed.

Detector balanced bias - Controlling circuit used in radar systems for anti-clutter purposes.

Dicke fix - A technique specifically designed to protect the receiver from ringing caused by noise, fast-sweep, or narrow-pulse jamming. The basic configuration

consists of a broadband limiting IF amplifier, followed by an IF amplifier of optimum bandwidth. The limit level is preset at approximately the peak amplitude of receiver noise. The bandwidth may vary from 10 to 20 MHz, depending on the jamming environment. The device provides excellent discrimination against fast-sweep jamming (10 - 500 MHz), usually something about 20 to 40 dB, without appreciable loss in sensitivity. However, strong CW jamming will seriously degrade the performance of the Dicke fix because the CW signal captures the radar signal in the limiter.

Dicke fix CFAR - Constant False Alarm Rate.

Dicke fix MT CFAR - An MTI CFAR technique similar to Dicke fix CFAR. Limiting and narrowbanding follow wideband amplification, phase detection, and cancellation so as not to impair the MTI performance.

Digital radio frequency memory (DRFM) - A computer-controlled digital device used in radar jamming systems. DRFM provides an extremely fast and accurate storage capability for victim radar signal parameters. Jamming systems employing DRFM can rapidly and accurately generate coherent jamming based on the memorized victim radar signal.

DINA - See Direct Amplified Noise.

Diode - An electron device having two electrodes, a cathode, and an anode.

Diplex - Two transmitters operating alternately on approximately the same RF and using a common antenna. The normal procedure is to pulse each transmitter at ½ the desired PRF, 180° out-of-phase. The advantage is that higher peak power per transmitter is possible because each transmitter is operating at ½ the normal duty cycle.

Dipole antenna - A straight, center-fed, ½ wavelength antenna. Horizontally polarized, it produces a figure-eight radiation pattern with maximum radiation at right angles to the plane of the antenna.

Direction finding (DF) - A procedure for obtaining bearings of radio frequency emitters with the use of a highly directional antenna and a display unit on an intercept receiver or ancillary equipment.

Direct amplified noise (DINA) - DINA without a carrier frequency is used for increasing (saturating) the radar receiver's noise level. This is called video stealing. The biggest danger from this type of jamming is that the radar operator may not realize he is being jammed when automatic gain control (AGC) or automatic noise leveling (ANL) are employed. This is a barrage type of jamming

with a bandwidth normally more than 50 MHz. DINA appears as a bright, high intensity wedge when the manual gain is employed. Thus, all useful radar scope information is lost.

Doppler (effect) - Continuous wave (CW) Doppler radar modules are sensors that measure the shift in frequency created when an object moves. A transmitter emits energy at a specific frequency which, when reflected, can indicate both speed and direction of the target. When objects move closer to the Doppler source, they increase in shift (positive value), and when they move further away, they decrease in shift (negative value).

Doppler radar - A radar system that measures the velocity of a moving object by the apparent shift in carrier frequency of the returned signal as it approaches or recedes.

Downlink - The signal from a transponder beacon located on a surface-to-air missile (SAM) used to provide a traceable radar return for missile guidance.

Downlink jamming (DLJ) - Some command guidance missiles carry a beacon (downlink) which is used by the parent radar to track the missile. If this beacon reply can be hidden from the parent tracking radar, the missile guidance solution can be defeated. Hence, downlink (beacon) jamming is intended to screen the missile beacon signal from the parent radar's view.

Ducting - The bending of radar rays due to atmospheric conditions. Ducting can either extend radar coverage beyond normal line of sight or it can deny the radar picture above a duct. Ducting is also called Anomalous Propagation.

Dummy antenna - A device that has the necessary impedance characteristics of an antenna and the necessary power-handling capabilities, but does not radiate or receive radio waves. Note: In receiver practice, that portion of the impedance not included in the signal generator is often called a dummy antenna.

Dummy load (radio transmission) - A dissipative but essentially nonradiating substitute device having impedance characteristics simulating those of the substituted device. This allows power to be applied to the radar unit without radiating into free space.

Duplex - In radar, a condition of operation when two identical and interchangeable equipments are provided—one in an active state, and the other immediately available for operation.

Duplexer - A switching device used in radar to permit alternate use of the same antenna for both transmitting and receiving.

Duty cycle - The ratio of the time the transmitter is actually on versus the time it could be on in a given transmission cycle. Mathematically, it can be expressed as:

$$\text{Duty Cycle} = \frac{PD}{PRT} \quad \text{or} \quad \text{Duty Cycle} = PD \times PRF$$

Dynamic range –

1. The difference, in decibels, between the overload level and the minimum acceptable signal level in a system or transducer. Note: The minimum acceptable signal level of a system or transducer is ordinarily fixed by one or more of the following: noise level, low-level distortion, interference, or resolution level.

2. Ratio of the specified maximum signal level capability of a system or component to its noise or resolution level, usually expressed in decibels.

Early warning radar - A radar set or system used near the periphery of a defended area to provide early notification of hostile aircraft approaching the area.

EA pod - A jamming system that is designed to be carried externally on an aircraft.

Effective radiated power (ERP) - Input power to antenna time multiplied by the gain of the antenna, expressed in watts.

Electromagnetic interference (EMI) - Any electromagnetic disturbance that interrupts, obstructs, or otherwise degrades or limits the effective performance of electronic systems. EMI can be induced intentionally, by way of jamming, or unintentionally because of spurious emissions and modulations.

Electromagnetic pulse (EMP) - The generation and radiation in a transmission medium of a very narrow and very high-amplitude pulse of electromagnetic noise. The term is associated with the high-level pulse because of a nuclear detonation and with an intentionally generated narrow, high-amplitude pulse for EA applications. In nuclear detonations, the EMP signal consists of a continuous spectrum with most of its energy distributed throughout the low frequency band of 3 to 30 kHz.

Electromagnetic radiation - Radiation made up of oscillating electric and magnetic fields and propagated with the speed of light. Includes gamma radiation, x-rays, ultraviolet, visible and infrared radiation, plus radar and radio waves.

Electromagnetic spectrum - The total range of frequencies (or wavelengths) over which any form of electromagnetic radiation occurs.

Electromagnetic test environment (EMTE) - A range complex of radars, such as those located at Eglin AFB, FL, operating in different frequency bands and modes to provide a very flexible test facility for evaluating aircraft antenna patterns, reflectivity measurements, infrared, reconnaissance, airborne interceptors, and electronic warfare devices and techniques.

Electronic attack (EA) - The use of electromagnetic energy, directed energy, or antiradiation weapons to attack personnel, facilities, or equipment with the intent of degrading, neutralizing, or destroying enemy combat capability. Action taken to reduce the enemy’s effective use of the electromagnetic spectrum. EA is a division of electronic warfare (EW).

Electronic combat (EC) - Action taken in support of military operations against the enemy's electromagnetic capabilities. EC is task-oriented and includes electronic warfare (EW), command and control warfare (C2W), and suppression of enemy air defenses (SEAD).

Electronic protection (EP) - Active and passive means taken to protect personnel, facilities, and equipment from any effects of friendly or enemy employment of electronic warfare that degrade, neutralize or destroy friendly combat capability. EP is a division of electronic warfare (EW).

Electromagnetic deception - The deliberate radiation, reradiation, alteration, absorption, or reflection of electromagnetic radiations in a manner intended to mislead an enemy in the interpretation of, or use of, information received by his electronic systems. There are two categories of electronic deception:

1. **Manipulative deception** - The alteration or simulation of friendly electromagnetic radiation to accomplish deception.

2. **Imitative deception** - The introduction of radiations into enemy channels that imitate his own emissions.

Electronic intelligence (ELINT) - The intelligence information product of activities engaged in the collection and processing for subsequent intelligence purposes of foreign, noncommunications, electromagnetic radiations emanating from other than nuclear detonations or radioactive sources.

Electronic intelligence parameter limit list (EPL) - A compilation of identified signals with assigned ELINT notations.

Electronic jammers –

1. **Expendable** - A transmitter designed for special use such as being dropped behind enemy lines.

2. **Repeater** - A receiver-transmitter device that, when triggered by enemy radar impulses, returns synchronized false signals to the enemy equipment. The returned impulses are spaced and timed to produce false echoes or bearing errors in the enemy equipment. See Expendable and Repeater Jammers.

Electronic jamming - The deliberate radiation, reradiation, or reflection of electromagnetic energy with the object of impairing the use of electronic devices, equipment, or systems.

Electronic order of battle - A listing of all the electronic radiating equipment of a military force giving location, type function, and other pertinent data.

Electronic reconnaissance - Specific reconnaissance directed toward the collection of electromagnetic radiations. Examples:

COMINT	Communications Intelligence
ELINT	Electronic Intelligence
OPINT	Optical Intelligence
RINT	Radiated Intelligence
SIGINT	Signal Intelligence

Electronic warfare (EW) - Military action involving the use of electromagnetic energy and directed energy to control the electromagnetic spectrum. EW has three divisions: electronic attack (EA), electronic protection (EP), and electronic warfare support (ES).

Electronic warfare support (ES) - Actions taken to search for, intercept, identify, and locate sources of intentional radiated electromagnetic energy for the purpose of immediate threat recognition. Surveillance of the electromagnetic spectrum that directly supports an operational commander's electromagnetic information needs. ES is a division of EW.

Electro-optics (EO) - The interaction between optics and electronics leading to the transformation of electrical energy into light, or vice versa, with the use of an optical device.

Electro-optic counter-countermeasures (EOCCM) - Actions taken to ensure the effective friendly use of the electro-optic spectrum despite the enemy's use of countermeasures in that spectrum.

Emission control (EMCON) –

1. The management of electromagnetic radiations to counter an enemy's capability to detect, identify, or locate friendly emitters for exploitation by hostile action.

2. Controlling the radiation of an active system to minimize detection by enemy sensors.

Endgame - The period of military engagement 3-5 seconds before missile impact.

Endgame countermeasures (EGCM) - Actions taken to defeat a tracking missile. This includes expendables, decoys, and maneuvers.

Essential elements of information (EEI) - The critical items of information regarding the enemy and his environment needed by the commander by a particular time to relate with other available information and intelligence to be able to reach a logical decision.

Expendable jammer - A nonrecoverable jammer. Early expendables were limited to chaff and flare deployments; however, various radiating jamming systems exist that use noise or repeater techniques. These are dispensed by aircraft or other delivery systems and are designed to disrupt or deceive a victim radar for a short period of time.

Extremely high frequency (EHF) - Frequencies in the range of 30 to 300 GHz.

False target - Radiated bundle of electromagnetic energy that is displaced in time from the echo that creates a response in the receiver where no reflecting surface exists.

False target generator - Device for generating electromagnetic energy of the correct frequency of the receiver that is displaced in time from the reflected energy of the target.

Fast automatic gain control (FAGC) - An AGC scheme characterized by a response time that is long with respect to a PW and short with respect to the target. An ultra fast FAGC will reduce the CW capture effect on the Dicke fix.

Fast time constant circuit (in radar) –

1. A circuit with short time constant used to emphasize signals of short duration to produce discrimination against extended clutter, long-pulse jamming, or noise.

2. A resistance-capacitance differentiating network with a time constant about equal to the transmitted pulse width and employed in the video portion of the receiver to provide the discrimination against jamming with low modulating frequencies. FTC is effective against CW, long jamming pulses, swept jamming, and clutter. It is also fairly effective against chaff corridors. Based on time, the radar receiver will allow only pulses equal to its own pulse width to pass and be presented on the indicators as targets. Because of this, only the leading edges of long pulses will be displayed.

Fence –

1. Line or network of early warning radars.

2. Concentric steel fence erected around a ground radar transmitting antenna to serve as an artificial horizon and suppress ground clutter that would otherwise drown out weak signals returning at a low angle from the target.

Ferret - An aircraft, ship, or vehicle especially equipped for the detection, location, recording, and analyzing of electromagnetic radiations.

Field of view (FOV) - The maximum solid angle visible by an optical or electro-optic system.

Fire control radar - Specialized radar systems used to locate and track airborne and surface targets, compute an optimum weapons firing point, and control the firing and sometimes guidance of its weapons.

FM-by-noise modulation - A method of frequency modulating with effective jamming method against AM and fix-tuned FM receivers. Not very effective against continuously tunable PFM receivers; careful tuning can defeat a great portion of the jamming signal. For this reason, FM-by-noise is not considered optimum as a type of modulation for jamming FM receivers.

FM jamming - Technique consisting of a constant amplitude RF signal that is varied in frequency around a center frequency to produce a signal over a band of frequencies.

Forward edge of the battle area (FEBA) - The foremost limits of a series of areas in which ground combat units are deployed, excluding the areas in which the covering or screening forces are operating, designated to coordinate fire support, the positioning of forces, or the maneuver of units.

Frequency agility - A radar's ability to change frequency within its operating band, usually on a pulse-to-pulse basis. This is an EP technique employed to avoid spot jamming and to force the jammer to go into a less effective barrage mode.

Frequency band designations –

1. As decided upon by the Atlantic City Radio Convention of 1947, and later modified by Comite Consultatif International Radio (CCIR) Recommendation No. 142 in 1953:

Band	RF Range	Band	RF Range
VLF	0-30 kHz	VHF	30-300 MHz
LF	30-300 kHz	UHF	300-3000 MHz
MF	300-3000 kHz	SHF	3-30 GHz
HF	3-30 MHz	EHF	30-300 GHz

2. Electronic warfare uses the following designations:

A	0-250 MHz	H	6-8 GHz
B	250-500 MHz	I	8-10 GHz
C	500-1000 MHz	J	10-20 GHz
D	1000-2000 MHz	K	20-40 GHz
E	2-3 GHz	L	40-60 GHz
F	3-4 GHz	M	60-100 GHz
G	4-6 GHz		

Frequency bandpass - The number of hertz where maximum output is obtained between two limits usually defined and bounded by lower and upper half power (3 dB) points.

Frequency hopping - An anti-jamming technique used by a radar system. The carrier frequency of the pulsed transmissions are periodically or continuously shifted within limits on each successive pulse.

Frequency diversity - Method of transmission and/or reception using several frequencies simultaneously to minimize the effects of selective fading, deliberate jamming, or interference.

Frequency modulated CW (FM-CW) jamming - FM of a CW signal to produce a wide band of signals when the exact operating frequency of the victim is not known within the CW limits.

Frequency modulation (FM) - A method of impressing a message upon a carrier signal by causing the carrier frequency to vary proportionally to the message waveform.

Frequency shift keying (FSK) - A form of FM where the carrier is mined code. In multiple FSK, the carrier is shifted to more than two frequencies. FSK is used principally with teletype communications.

Frequency spectrum - The entire range of frequencies of electromagnetic radiation.

G - Acceleration due to gravity (32.2 ft/sec^2).

Gain (manual) - The receiver gain control allows the operator to vary the receiver sensitivity. It is not designed as an AJ feature; however, when properly employed it may greatly reduce the effects of jamming. The radar detection capability is also reduced by an equal amount.

Gain (transmission gain) - The increase in signal power in transmission from one point to another under static conditions. Note: Power gain is usually expressed in decibels.

Glint (in radar) –

1. The random component of target location error caused by variations in the phase front of the target signal (as contrasted with scintillation error). GLINT may affect angle, range of Doppler measurements, and may have peak values corresponding to locations beyond the true target extent in the measured coordinate.

2. Electronic countermeasures that use the scintillating, or flashing, effect of shuttered or rotating reflectors to degrade tracking or seeking functions of an enemy weapons system.

Ground controlled intercept (GCI) - Vectoring an interceptor aircraft to an airborne target by means of information relayed from a ground-based radar site that observes both the interceptor and target.

Guidance system (missile) - A system that evaluates flight information, correlates it with target data, determines the desired flight path of the missile, and communicates the necessary commands to the missile flight control system.

Guided missile - An unmanned vehicle moving above the surface of the earth, whose trajectory of flight path is capable of being altered by an external or internal mechanism.

Height finder - A radar used to detect the angular elevation, slant range and height of objects in the vertical sight plane. An air defense ground radar used specifically to accurately determine aircraft altitude for tracking and ground controlled intercepts.

Hertz (Hz) - The unit of frequency, equal to one cycle of variation per second. It supersedes the unit cycle per second (cps).

High frequency (HF) - Frequencies from 3000 - 30,000 kHz.

Home-on-jam (HOJ) - A missile mode of operation in which a jamming signal is used to develop steering information for the missile to home in on the jamming source.

Homing guidance - A system by which a missile steers itself toward a target by means of a self-contained mechanism which is activated by some distinguishing characteristics of the target.

Identification, friend or foe (IFF) - A system using radar transmission to which equipment carried by friendly forces automatically responds, for example, by emitting pulses, thereby distinguishing themselves from enemy forces. It is the primary method or determining the friendly or unfriendly character of aircraft and ships by other aircraft and ships, and by ground forces employing radar detection equipment and associated identification, friend or foe units.

Image frequency - An undesired input frequency capable of producing the selected frequency by the same process. NOTE: An image frequency is a frequency which differs from, but has a certain symmetrical relationship to, that which a superheterodyne receiver is tuned. Consequently, the image frequency can be mistakenly accepted and processed as a true frequency by the receiver.

Image jamming - Jamming at the image frequency of the radar receiver. Barrage jamming is made most effective by generating energy at both the normal operating and imaging frequency of the radar. Image jamming inverts the phase of the response and is thereby useful as an eagle deception technique. Not effective if the radar uses image rejection.

Imitative deception - The introduction of radiations into enemy channels which imitates their own emissions.

Imitative jamming - The jamming technique of transmitting a signal identical to the original guidance signal.

Infrared (IR) - That portion of the frequency spectrum lying between the upper end of the millimeter wave region and the lower (red) end of the visible spectrum. In wavelength, the IR lies between 0.78 and 300 microns; in frequency, it lies between one and 400 terahertz (THz).

Infrared counter-countermeasures (IRCCM) - Actions taken to effectively employ our own infrared radiation equipment and systems in spite of the enemy's actions to counter their use.

Infrared countermeasures (IRCM) –

1. Countermeasures used specifically against enemy threats operating in the infrared spectrum.

2. Actions taken to prevent or reduce the effectiveness of enemy equipment and tactics employing infrared radiation.

Instantaneous automatic gain control (IAGC) –

1. That portion of a system that automatically adjusts the gain of an amplifier with varying input pulse amplitudes. The adjustment is sufficiently fast to operate while a pulse is passing through the amplifier.

2. In radar, a quick-acting automatic gain control that responds to variations of mean clutter level, or jamming over different range or angular regions, avoiding receiver saturation (also known as back-bias). IAGC automatically adjusts the gain of the radar receiver so that strong signals do not block adjoining weak signals. IGAC is not quick enough to block short pulse jamming, extended clutter, or chaff.

Instantaneous frequency CFAR (IFCFAR) - A technique similar to the IF DICKE FIX CFAR, except for having a phase detector instead of a frequency discriminator. The primary use of this technique, combined with rapid random frequency changes, is to make chaff signals appear noise-like so they can be handled conventionally.

Instantaneous inverse gain - Pulse-by-pulse amplification, modulation, and reradiation of the victim's radar pulse to obscure angle data.

Intercept point - A computed point in space toward which an interceptor is vectored to complete an interception.

Interference (electronic) - An electrical or electromagnetic disturbance that causes undesirable responses on electronic equipment. Electrical interference refers specifically to interference caused by the operation of electrical apparatus that is not designed to radiate electromagnetic energy.

Interferometer - As pertains to radar, a receiving system which determines the angle of arrival of a wave by phase comparison of the signals received at separate antennas or separate points on the same antenna.

Interleaved jamming - Time sharing of a jammer's output around many threats usually to cover pulse jamming.

Intermediate frequency (IF) –

1. A fixed frequency to which all carrier waves are converted in a superheterodyne receiver.

2. A frequency to which a signaling wave is shifted locally as an intermediate step during transmission or reception.

3. A frequency resulting from the combination of the received signal and that of the local oscillator in a superheterodyne receiver.

Intermediate frequency jamming - Form of CW jamming that is accomplished by transmitting two CW signals separated by a frequency equal to the center frequency of the radar receiver IF amplifier.

Interrogator - A device used to transmit pulse-coded challenges to an IFF transponder and then evaluates the pulse-coded reply for identification purposes.

Intrapulse modulation repeater - A classified deception jamming technique.

Intrusion –

1. **The entry of a nonfriendly aircraft or system into friendly air space.**

2. **The intentional interference in a communication system by which the intruder attempts to confuse, delay, or cause error by the selective introduction of additional data.**

Inverse conscan - One method of confusing a radar operator or fire control radar system is to provide erroneous target bearings. This is accomplished by first sensing the radar antenna scan rate and then modulating repeater amplifier gain so that the weapons systems will fire at some bearing other than the true target bearing. The angle deception technique is used to break lock on CONSCAN radars.

Inverse gain - Amplification, inverse modulation, and reradiation of a radar's pulse train at the rotation rate of the radar scan. Deceives a conical scanning radar in angle.

Inverse-gain repeater jammer - A form of repeater in which the jammer creates false targets by varying the output power inverse with the strength of the received radar signal.

Jaff - Expression for the combination of electronic and chaff jamming. An ECM tactic involving the use of jammers to illuminate chaff corridors or chaff bursts to produce false targets.

Jam attenuator control - Used to prevent receiver saturation from any strong signals, including electronic jamming, chaff, or clutter. It also permits the determination of the bearing and elevation of jammers that would cause wide sectors of the scope to be obscured. The control reduces all signals equally, jamming as well as targets. This control should be used in the jammed sector only and not for an entire antenna revolution. Targets in the jammed sector would only be seen if they were stronger than the jamming. However, the jammed sector may be reduced enough in size to allow the operator to determine either bearing or elevation.

Jammer - A device used to deprive, limit, or degrade the use of communications or radar systems. Radio frequency jammers include noise, discrete frequency repeater, and deceptive equipment.

Jamming-to-signal (J/S) ratio - The relative power ratio of jamming to the radar return signal at the radar antenna. The inverse of the signal-to-jamming ratio.

Jam strobe - Also called JAVA (jamming amplitude versus azimuth). A circuit that generates a marker on the PPI to indicate signal strength as a function of bearing. It does this by sampling the jamming intensity once each repetition period. Besides showing the direction of the jammer, it also indicates the severity of main beam and sidelobe jamming.

Jet engine modulation (JEM) - Modulation present in the radar returns received from a jet aircraft, caused by the rotation of the fan or turbine blades of the aircraft's engines.

Jittered pulse repetition frequency (jittered PRF) - The PRF is rapidly varied at a random rate so that false targets appear to jitter or appear fuzzy on the scope. An alternative to jittered PRF is to change the PRF momentarily. This causes the false targets to change their position on the scope. It provides a discrimination against repeater-type jammers.

Klystron - A very stable microwave amplifier that provides high gain at good efficiency. This is accomplished by velocity modulating (accelerating a beam of electronics flowing from its cathode to its anode.

Laser target designation - The use of a laser to direct a light beam onto the target so that appropriate sensors can track or home on the reflected energy.

Leading-edge tracker - A tracking radar that obtains its data from the leading edge of the echo pulse from the target.

Light amplification by stimulated emission of radiation (LASER) - A process of generating coherent light. The process uses a natural molecular (and atomic) phenomenon whereby molecules absorb incident electromagnetic energy at specific frequencies. It then stores this energy for short but usable periods, then releases the stored energy as light at particular frequencies, and in an extremely narrow frequency band.

Lobe - One of the three-dimensional sections of the radiation pattern of a directional antenna bounded by 1-2 cones of nulls.

Lobe-on-receive-only (LORO) - Mode of operation consisting of transmitting on one antenna system and receiving the reflected energy on another system (TWS, conical, or monopulse).

Local oscillator off - A simple expedient of shutting off the local oscillator during barrage jamming. The barrage jammer will not be seen unless there is a beating signal such as a target; therefore, targets will appear. Since targets will not

appear in directions where no jamming arises, either an automatic azimuth switch or an additional receiver display is required.

Logarithmic fast time constant log (LOG-FTCL) - A device consisting of a logarithmic IF amplifier followed by an FTC circuit. The LOG-FTC combination is very effective in removing variations in the video output noise level caused by spot noise, wideband noise, and slow sweep noise modulated AM jamming.

Logarithmic receiver - A receiver whose response approximates the logarithm of the strength of the incoming signal. A special type of receiver having a large dynamic range of automatic gain control that gives considerable protection against receiver saturation by strong jamming on interference signals. Useful against weather, clutter, chaff, and spot jamming.

Look-down, shoot-down - Refers to an air interceptor (AI) equipped with a pulse Doppler radar, or a radar that has a moving target indicator (MTI) feature, that can detect and lock-on to a target within ground return clutter enabling the AI to track and shoot the target.

Look-through –

1. **When jamming, a technique by which the jamming emission is interrupted irregularly for extremely short periods to allow monitoring of the victim signal during jamming operations.**

2. **When being jammed, the technique of observing or monitoring a desired signal during interruptions in the jamming signals.**

Low frequency (LF) - Frequencies from 30 - 300 kHz.

Low power spread spectrum radar - A low power, high duty cycle radar whose spectrum is spread 100 MHz or more. Since this radar has a broad output spectrum and a high duty cycle, neither time nor frequency can be effectively used to resolve these signals. This leaves direction as the prime method of resolution. The spectrum of these radars is spread over the bandwidth by any of the pseudo random noise modulating techniques commonly used in communications. Techniques such as bi-phase modulation, quaternary phase modulation, chirp, random frequency jumping, etc., may be used to spread either a CW signal or a very high duty cycle signal. Such signals have a very good range resolution—approximately equal to the reciprocal of the bandwidth.

Magnetron - A radar microwave device whose operation is based on the motion of electrons (AC) under the influence of combined electric and magnetic fields.

Mainlobe - The lobe of a transmitting or receiving antenna centered on the directivity axis of the antenna.

Manipulative deception - The alteration or simulation of friendly electromagnetic radiations to accomplish the deception.

Meaconing, interference, jamming, and intrusion (MIJI) - An acronym of four component parts to determine the intent and technique of electromagnetic radiation.

Meaconing - a system of receiving radio beacon signals and rebroadcasting them on the same frequency to confuse navigation. Interference - unintentional electromagnetic radiation that may cause interference with electronic equipment.

Jamming - the deliberate radiation, reradiation, or reflection of electromagnetic energy with the intent of impairing the use of electronic devices, equipment, or systems being used by the enemy.

Intrusion - the intentional insertion of electromagnetic energy into transmission paths in any manner with the objective of deceiving operators or causing confusion.

Medium frequency (MF) - Frequencies from 300 to 3,000 kHz.

Micron - A unit of length equal to a micrometer (10^{-6} meters).

Microwave amplification by stimulated emission or radiation (MASER) - A low-noise, radio-frequency amplifier. The emission of energy stored in a molecular or atomic system by a microwave power supply is stimulated by the input signal.

Microwave communications - Line-of-sight communications, the frequency of which is higher than 300 MHz.

Millimeter waves - Frequencies (30 GHz to 300 GHz) in the millimeter portion of the electromagnetic spectrum.

Miss distance - The distance measured between the closest paths of a target and interceptor (i.e., aircraft and missile). One objective of self-protection jamming systems is to increase the miss distance to avoid destruction if missile launch cannot be prevented.

Missile approach warning system (MAWS) - A system used to detect and provide warning of approaching missiles. MAWS may be partitioned into active MAWS and passive MAWS.

1. Active missile approach warning system (AMAWS) - Generally employs pulse Doppler radar as its sensor. This radar is able to discern a moving target in stationary or slow-moving background clutter.

2. Passive missile approach warning system (PMAWS) - An ultraviolet (UV) or infrared-based detector system with the ability to detect and distinguish threat missiles from surrounding clutter and non-lethal missiles.

Modulated barrage jamming - A technique that varies the amplitude of the output power of a barrage jammer with sinusoidal or complex modulation.

Modulated CW jamming - A CW carrier waveform that has been modulated with another signal (such as noise, low, medium, or high frequencies), and is transmitted for the purpose of deception. May be AM, FM, pulse modulated (PM), etc.

Modulated PRF (MPRF) - The deliberate modulation of the interpulse spacing in a pulse train. See PRF Jitter and PRF Stagger.

Modulation - The variation of amplitude, frequency, or phase of an electromagnetic wave by impressing another wave on it.

Modulator - A device (such as an electron tube) for modulating a carrier wave or signal for the transmission of intelligence of some sort.

Monopinch - Anti-jam application of the monopulse technique where the error signal is used to provide discrimination against jamming signals.

Monopulse - A method of pulse generation that allows the simultaneous determining of azimuth, elevation and range, and/or speed from a single pulse.

Monopulse radar - A radar using a receiving antenna system having two or more partially overlapping lobes in the radiation patterns. Sum and difference channels in the receiver compare the amplitudes or the phases of the antenna outputs to determine the angle of arrival of the received signal relative to the antenna boresight. A well-designed monopulse tracking system will achieve a more accurate track under conventional jamming techniques than on the skin return. Certain monopulse trackers are susceptible to angular jamming techniques such as skirt and image jamming. Techniques such as “CROSS EYE” are designed to

attack all monopulse tracking systems. Monopulse deception is a major area of advanced R&D with no clear "best technique" yet in sight.

Moving target indicator (MTI) - A radar presentation that shows only targets that are in motion. Signals from stationary targets are subtracted out of the return signal by the output of a suitable memory circuit.

MTI CFAR - A technique that provides CFAR capability in an MTI receiver. The cancellation of ground clutter is not impaired during radar jamming.

Multiband radar - Radar that simultaneously operates on more than one frequency band through a common antenna. This technique allows for many sophisticated forms of video processing and requires that a jammer must jam all channels simultaneously.

Multiplex - Simultaneous transmission of two or more signals on a common carrier wave. The three types of multiplex are called time division, frequency division, and phase division.

Multitarget generator jamming - A technique where a generator takes the radar's PRF, scan rate, and antenna lobe pattern and computes when to transmit false targets at the radar's frequency. The targets appear as true targets, but normally only about 20 percent of the targets are programmed to give a logical course and speed. The many targets saturate the tracking operator's capability by the sudden appearance of multiple targets, and/or many targets either stationary or on illogical courses and speeds, or maneuvers such as 90 degree turns at high speeds but with no displacement. Also, targets may appear in back and sidelobe positions.

Music - In air intercept, a term meaning electronic jamming.

Narrowband (NB-1) - Narrows the receiver frequency making it more selective. It limits the target signals and the jamming signal to a set level of amplitude and reduces the level of a jamming signal if the jammer is not tuned to the radar's exact frequency.

Narrow pulse jamming - Jamming where the jammer pulse width is less than the radar's pulse width. Technique may be required by interleaved jamming.

Noise –

1. Any unwanted disturbance within a dynamic electrical or mechanical system, such as undesired electromagnetic radiation, and any transmission channel or device.

2. Uncontrolled random disturbances that arise in a guided missile system because of various physical phenomena.

Noise jamming - Direct (straight) AM or FM noise on a carrier frequency that has a highly variable bandwidth for the purpose of increasing (saturating) the radar receiver's noise level.

Nuclear effects - The electromagnetic phenomena resulting from a nuclear explosion. The phenomena are listed as follows:

1. **Argus phenomena (trapped electrons)**. The trapping in the earth's magnetic field of electrons produced by a nuclear burst.

2. **Blackout (radio frequency interference)**. An effect that is the result of ionization produced by a nuclear explosion in or above the atmosphere. This ionization can cause interference (blackout) by attenuating, reflecting, cluttering, and scattering radar electromagnetic pulses in a high-intensity burst of electromagnetic radiation, predominantly in the radio frequency range of the spectrum.

3. **Optical phenomena**. Intense radiations covering all parts of the optical spectrum are produced by the interactions between the atmosphere and the nuclear radiation and fission products resulting from a nuclear detonation. The resulting auroras and airglows are created as an optical background that can affect reconnaissance, tracking, warning and homing systems, and personnel.

4. **Transient radiation effects on electronics**. Nuclear radiation impinging on electronic systems or components can substantially alter the operation and output of these systems. The word transient refers to the type of environment and not to the duration of the effect since the effect may be either transient or permanent.

Nulls - Points of no radiation in an antenna radiation pattern.

Obscuration - Effects produced by masking-type jammers. Denial of either range or angle is achieved by submerging data interference caused by noise or noise-like signals.

Off-frequency spot jamming - A type of spot jamming in which the jammer frequency is off the radar operating frequency but still within the radar receiver bandpass.

Optical countermeasures - Applications of electronic countermeasures in the visible light portion of the electromagnetic spectrum. Actions taken to prevent or reduce an enemy's effective use of the visible spectrum.

Oscillator - Electronic circuit or device capable of converting direct current (DC) into alternating current (AC) at a frequency determined by the inductive and the capacitive constants of the oscillator.

Over-the-horizon radar - A radar system that makes use of the ionosphere to extend its range of detection beyond line-of-sight. Over-the-horizon radars may be either forward scatter or backscatter systems.

Palmer scan - Conical scan superimposed on another type of scan pattern- usually a spiral pattern.

Passive angle tracking (PAT) - A target may be tracked “passively” if that target emits electromagnetic radiation; i.e., jamming radio, radar signal of sufficient duration that a DF bearing may be obtained. The emission from the target is DF-ed in azimuth and/or elevation. No range information will be available unless cross DF techniques are used by two or more passively tracking sites.

Passive detection and tracking - By combining azimuth data on jamming strobes from several stations, intersections are obtained which indicate the position of the jammers. The number of ghosts can be reduced by increasing the number of friendly stations and obtaining elevation angles of strobes when available.

Passive electronic countermeasures - Electronic countermeasures based on the reflection, absorption or modification of the enemy's electromagnetic energy. This distinction between active and passive countermeasures is not currently used, but is based on the presence or absence of an electronic transmitter.

Passive homing guidance - A system of homing guidance in which the receiver in the missile uses radiations only from the target.

Peak power - Maximum power output during transmission time. Expressed in watts or megawatts.

Penetration aids - Techniques and/or devices employed by aerospace systems to increase the probability of weapon system penetration of any enemy defense.

Examples are: Low altitude profiles, trajectory adjustments, reduced radar cross-sections of attack vehicles, improved vehicle hardness to effects of defense engagements, terrain avoidance radar, bomber defense missiles, decoys, chaff, electronic countermeasures, etc. Penetration aids are used by an offensive system to penetrate enemy defenses more effectively. Also called PENAIDS.

Phase modulation - A method of impressing a message upon a carrier signal by causing the carrier phase to vary proportionally to the waveform.

Phase shift keying - A method of impressing a digital signal upon a carrier signal by causing the carrier phase to take different values corresponding to the different values of the digital signal.

Phased array radar - Radar using many antenna elements that are fed out-of-phase to each other. The direction of the beam can be changed as rapidly as the phase relationships (usually less than 20 µsec). Thus, the antenna remains stationary while the beam is moved electronically. The use of many antenna elements allows for very rapid and high directivity of the beam(s) with a large peak and/or average power.

Point defense - The defense of specified geographical areas, cities, and vital installations. One distinguishing feature of point defense missiles is that their guidance information is received from radars located near the launching sites.

Polarization - The direction of an electrical field is considered the direction of polarization. When a half-wave dipole antenna is horizontally oriented, the emitted wave is horizontally polarized. A vertical polarized wave is emitted when the antenna is erected vertically.

Polarization diversity - The variation of polarization (such as horizontal, vertical, circular, or elliptical for radar use) either simultaneously or singularly.

Power management - Generally classified methods to optimize all related EW activities and facilities-usually in reference to a coordinate, optimized EW suite.

PPI scope - A radar display yielding range and azimuth information via an intensity modulated display and a circular sweep of a radial line. The radar is located at the center of the display.

PRF jitter - PRF rapidly varied at a random rate so that false targets appear to jitter, or appear fuzzy, on the scope. An alternative to jittered PRF is to momentarily change the PRF. This causes the false targets to change their position on the scope.

PRF stagger - The technique of switching PRF or PRI to different values on a pulse-to-pulse basis such that the various intervals follow a regular pattern. This is useful in compensating for “blind speeds” in pulsed MTI radars. Interpulse intervals differ but follow a regular pattern.

Pseudo noise - A modulation technique resulting in low signal selectability and decreased vulnerability to jamming.

Pulse compression - A scheme whereby a specifically modulated, medium power long pulse is stretched and transmitted. The returned pulse is compressed in the receiver demodulation process to obtain the advantage of narrow pulse operation. Long pulses provide the advantage of long range detection and short (compressed) pulses provide the advantage of better resolution and accuracy. This technique uses matched filter techniques for discriminating against signals that do not correspond to the transmitted code. It permits an increase in average transmitted power (without an increase in peak power) with no loss in range resolution. Pulse compression is a special form of pulse coding and correlation.

Pulse deception jamming - Any of the many false target techniques used to deceive a pulse radar, as opposed to obscurative noise techniques.

Pulse discriminator - Device that responds only to a pulse that has a particular characteristic, such as duration or period.

Pulse Doppler radar - A highly complex radar system that employs a very high pulse repetition frequency (usually 10,000 PPS or higher) to reduce “blind speeds” and measure the Doppler frequency shift to resolve target velocity. Pulse Doppler is applied principally to radar systems requiring the detection of moving targets in a ground clutter environment. It uses pulse modulation to achieve higher peak power, greater range, less susceptibility to unfriendly detection, and enhanced range resolution.

Pulse duration - The time in microseconds that the radar set is transmitting RF energy. Generally, the greater the pulse duration, the higher the average power, but the poorer the range resolution. Also known as pulse width. More technically, it is the time interval, measured at the half-amplitude points, from the leading edge to the trailing edge of a pulse.

Pulse group - In complex pulse trains, two or more pulses that are recognizably distinct from the others.

Pulse interference suppression and blanking (PISAB) - An EP automatic interference blanking device that will blank all video signals not synchronous with the radar PRF. PISAB does not require any trigger and operates on both normal and MTI modes. It is effective against random pulse signals.

Pulse jitter - Random variation of interpulse interval.

Pulse modulation (PM) - A special case of amplitude modulation in which the carrier wave is varied at a pulsed rate. This series of pulses is generally for transmitting data. The result is a short, powerful burst of electromagnetic radiation that can be used for measuring the distance from a radar set to a target.

Pulse modulated jamming - Use of jamming pulses of various widths and repetition rates.

Pulse position modulation - Modulation by variation of the interval that elapses between the pulse to be modulated in a group of pulses and a synchronizing pulse, usually the first pulse of the group.

Pulse repetition frequency (PRF) - The rate at which pulses or pulse groups are transmitted from a radar set. Generally, PRF is the number of pulses generated per second and is expressed in hertz (Hz).

Pulse repetition interval (PRI) - The interval of time between two transmitted radar pulses, usually expressed in microseconds. PRI is the inverse of PRF.

Pulse recurrence time (PRT) - Time elapsing between the start of one transmitted pulse and the start of the next. It is the reciprocal of PRF.

Pulse recurrence time (PRT) agility - Ability of the radar set to vary the number of pulses per set.

Pulse width (PW) - See Pulse Duration.

Pulse width discrimination (PWD) - An EP technique used to discriminate against received pulses that do not have the same duration as the radar transmitted pulse. PWD is used in eliminating the effects of pulse type interference when the interference pulses are not the same length as the real radar pulse. Since this circuit generates a blanking gate that shuts off the receiver whenever a pulse of improper length is sensed, loss of valid target data can result. PWD offers good discrimination against long-pulse jamming and jamming signals employing low frequency noise modulation. It affords little or no discrimination against short pulses and HF noise modulations.

Pulse width discriminator - A device that measures the pulse length of video signals and passes only those pulses whose time duration tails into some predetermined design tolerances. A pulse width discriminator will generally provide some gain against barrage jamming, similar to that of a matched filter in the video.

“Rabbits” - Interference from another radar on or near the frequency of the receiving radar. Shows on the indicators as interference at the rate of the relative PRF of the interfering radar.

Radar - From Radio Detection And Ranging. A device used to detect a distant target and determine and display its relative direction (azimuth) and its relative distance (range).

Radar absorbent material (RAM) - Material used as a radar camouflage device to reduce the echo area of an object.

Radar beacon - A receiver-transmitter combination that sends out a coded signal when triggered by the proper type of pulse enabling determination of range and bearing information by the interrogating station or aircraft.

Radar cross section - The equivalent area intercepted by a radiated signal and, if scattered uniformly in all directions, produces an echo at the radar receiver equal to that of the target. Typical radar cross sections of aircraft vary from one to over 1,000 square meters. The RCS of ships may exceed 10,000 square meters.

Radar definition - The accuracy with which a radar obtains target information such as range, azimuth, or elevation.

Radar homing - Homing on the source of a radar beam.

Radar homing and warning (RHAW) - Typically consists of an airborne, wideband video receiver designed to intercept, identify, and display the direction to pulse-type emitters.

Radar resolution - A measure of a radar's ability to separate targets that are close together in some aspect of range, azimuth, or elevation into individual returns.

Radar warning receiver (RWR) - A receiver onboard an aircraft that analyzes the hostile radar environment and determines radar threat by type, frequency, relative bearing, and relative distance. The threat is displayed to the aircrew by means of display lights, video symbols, and aural tones.

Radio frequency (RF) - Electromagnetic energy radiated at some frequency.

Radio frequency interference - An unintentional interfering signal capable of being propagated into electronic equipment, usually derived from sources outside the system.

Railing –

1. Pertains to radar pulse jamming at high recurrence rates (50 to 150 kHz). It results in an image on a radar indicator resembling fence railing.

2. The name given to that pattern produced on an “A” scope by CW modulated with a high frequency signal. Railings appear as a series of vertical lines resembling target echoes along the baseline.

Random modulation CW jamming - The use of a nonperiodic function to an AM, FM, or AM/FM CW carrier. The effects produced on a radar are similar to those produced by DINA.

Random noise - Electromagnetic energy having no particular modulation or pattern. May be generated by either natural atmospheric phenomena or by electromagnetic radiation devices.

Random pulse jamming - The technique by which a pulse transmission system is pulsed irregularly by random noise signals.

Range - The distance from one object to another.

Range gate - A gate voltage used to select radar echoes from a very short range interval.

Range gate capture - A jamming technique using a spoofer radar transmitter to produce a false target echo that can make a target tracking radar move off the real target and follow the false one.

Range gate pull-in (RGPI) - Same as range gate pull-off (see next text entry) except that the deceptive pulse is transmitted before the radar pulse is received. This is accomplished by digital storage of the pulse repetition period which must be extremely stable.

Range gate pull-off (RGPO) - A deception technique used against pulse tracking radars using range gates. The jammer initially repeats the skin echo with minimum time delay at a high power to capture the AGC circuitry. The delay is progressively increased, forcing the tracking gates to be pulled away ("walked off") from the target echo. Frequency memory loops (FML) or transponders provide the variable delay.

Range gate tracker - A radar system that tracks targets in range by measuring the elapsed time from the transmitted pulse to the echo return.

Range tracking - Pulse radars measure the time difference between radar pulse transmission and echo reception. The range gate is positioned at a range where the target is expected. The receiver is blanked off except during the period where the range gate is positioned. Range tracking may occur at the leading edge of the return pulse or between ON and OFF gates.

Recovery time (RT) - The time that a radar requires to "get ready" to receive after a pulse is sent out. Time is required to get ready because the high-powered pulse tends to fill the sensitive radar receiver with RF energy, which prevents target returns from being seen. A short "damping out" period occurs after the pulse width during which time the RF energy dissipates, allowing weak target echoes to once again be detected.

Rectifier - A device (such as a vacuum tube) for converting alternating current (AC) into direct current (DC).

Repeater - A receiver-transmitter combination that amplifies the received signal and retransmits it.

Repeater jammer - Equipment used to confuse or deceive the enemy by causing his equipment to present false information. This is done by a system that intercepts and reradiates a signal on the frequency of the enemy equipment. The reradiated signal is modified to present erroneous data on azimuth, range, number of targets, etc.

Repeater jamming - Interception and reradiation of a signal with the reradiation of a signal being modified to give erroneous azimuth, range, velocity, acceleration, or number of targets.

Resolution - The ability of a system to distinguish between two adjacent objects and to display them separately.

Resolution cell - The minimum volume in space in which a radar can discriminate between targets. It is determined numerically-for a conventional radar-by the width of the beam in angle, the pulse width in range, and the velocity gate width in speed.

Response noise jammer - A repeater jammer that transmits spot noise over the received radar frequency even in the case of radar frequency agility.

Ringing - The undesired oscillation or triggering of an electrical device by its own transmitter.

Rope - An element of chaff consisting of a long roll of metallic foil or wire that is designed for broad low frequency response. (See Chaff.)

Rope chaff - Chaff that contains one or more rope elements. (See Chaff.)

Rotating polarization - The result of a rotating feed. This should not be confused with circular polarization where the electric field rotates about the axis of propagation at the radar frequency.

SAM - Surface-to-air missile.

Saturating signal - In radar, a signal of an amplitude greater than the dynamic range of the receiving system.

Sawtooth modulated jamming - Electronic countermeasure technique when a high-level jamming signal is transmitted, thus causing large AGC voltages to be developed at the radar receiver that, in turn, causes the target return and receiver noise to completely disappear.

Scan - The process of directing a beam of RF energy successively over a given region, or the corresponding process in reception.

Scan interval - The time interval from the peak of one mainlobe in a scan pattern to the peak of the next mainlobe.

Scan period - The time period of basic scan types (except conical and lobe switching) or the period of the lowest repetitive cycle of complex scan combinations. The basic unit of measurement is degrees/mils per second or seconds per cycle.

Scan rate modulation - Modulation of the deception jamming signal with one or more frequencies that are related to the scan rate of the radar.

Scan type - The path made in space by a point on the radar beam, for example, circular, helical, conical, spiral, or sector.

Search –

1. A term applied to that phase of radar operation when the lobe, or beam of radiated energy, is directed in such a way to search for targets in the area.

2. A systematic examination of space to locate and identify targets of interest.

Sector scan - A scan in which the antenna sweeps back and forth through a selected angle.

Selective identification feature (SIF) - A capability which, when added to the basic IFF system, provides the means to transmit, receive, and display selected coded replies.

Self-protection jamming - Jamming to protect the vehicle upon which the jammer is deployed.

Semiactive radar homing - Semiactive homing guidance combines principles from both the beam rider and the active radar homing missile. Track on the target is established by the AI's radar; the missile is launched when the target comes within its effective range. During missile flight, the AI maintains track on the target. Radar returns from the target are received by the missile. Guidance commands are generated within the missile from the radar returns.

Sensitivity time control (STC) - Programmed variation of the gain (sensitivity) of a radar receiver as a function of time within each pulse repetition interval or observation time. STC prevents overloading of the receiver by strong echoes from targets or clutter at close ranges. STC reduces the gain of the radar receiver for detection of close-in targets. It is particularly effective in removing close-in clutter and strong nearby signals. STC also refers to a circuit that reduces the gain of the radar receiver immediately following the transmission of the radar pulse so that the receiver is not saturated by strong radar returns from nearby objects.

Serrodyne - A method of “pulling off” the velocity gate of a Doppler radar by using a voltage-controlled phase shifter, usually a traveling wave tube (TWT). This introduces a frequency shift from zero to some maximum value, pulling the Doppler tracking gate away from the skin echo. The phase shift is usually accomplished by modulating the TWT's helix voltages.

Sidelobe - Part of the beam from an antenna, other than the mainlobe. Sidelobe gain is usually less than mainlobe gain. Given that the mainlobe radiates most of the power at zero degrees azimuth, sidelobes inherently radiate significant power in the direction of +20°, 90°, and 150° relative to the mainlobe.

Sidelobe blanking (SLB) - A device that employs an auxiliary wide angle antenna and receiver to sense whether a received pulse originates in the sidelobe region of the main antenna and to gate it from the output signal if it does. This technique uses an omnidirectional antenna and compares relative signal strength between the omni and the radar antenna. The omnichannel (plus receiver) has slightly more gain than the sidelobes of the normal channel, but less gain than the main beam. Any signal that is greater in the omnichannel must have been received from a sidelobe and is blanked. This technique is effective in removing spoofed signals with a duty cycle up to 50%.

Sidelobe canceller (SLC) - A device that employs one or more auxiliary antennas and receivers to allow linear subtraction of interfering signals from the desired output if they are sensed to originate in the sidelobe of the main antenna. This technique employs the same antenna and receiver configurations of the SLB, except that a gain matching and canceling process takes place. Extraneous signals entering the sidelobes of the main antenna are canceled while the targets remain. This type of system exhibits cancellation roughly 20 dB against a single noise jammer. With multiple jammers at various azimuths, the performance of this device rapidly deteriorates.

Sidelobe jamming - Jamming through a sidelobe of the receiving antenna in an attempt to obliterate the desired signal received through the mainlobe of the receiving antenna at fixed points.

Sidelobe suppression - The suppression of that portion of the beam from a radar antenna other than the mainlobe.

Sidewinder - A solid-propellant, air-to-air missile with nonnuclear warhead and an infrared guidance system. Designated as the AIM-9 missile.

Signal intelligence (SIGINT) - Intelligence derived from the interception of enemy communications and noncommunication signals. A generic term that includes both COMINT and ELINT.

Signal-to-jamming ratio (S/J) - The ratio of the signal power to the jamming power or intentional interference at some point in the system. This ratio is often expressed in decibels.

Signal-to-noise ratio (S/N) - Ratio of the power of the signal to the power of the noise.

Signature - The set of parameters that describes the characteristics of a radar target or an RF emitter and distinguishes one emitter from another. Signature

parameters include the RF of the carrier, the modulation characteristics (typically the pulse modulation code), and the scan pattern.

Sine-wave modulated jamming - Jamming signal produced by modulating a CW signal with one or more sine waves.

Single beam blanking (SBB) - Used by phased array radars as an alternative method of Beam to Beam Correlation (BBC). It is effective, to some degree, against many multiple target generators and swept frequency jammers. Because of overlapping beamwidths, a target signal will appear in more than one beam as the beams are scanned past a true target. When jamming signals are transmitted along one beam, that beam is blanked by the radar receiver.

Skirt jamming - Jamming that places the signal somewhat off the radar center frequency but within the IF skirts. The technique will degrade the tracking accuracy of some monopulse radars; however, good radar design makes the technique ineffective.

Smart jamming - Selective jamming of threat radars with an optimized modulation signal at a correct time. Power management techniques are used to control smart jamming systems.

Smith modulation - Deceptive technique that operates on the servo loop of the victim's radar. Two RF carriers are transmitted with a few cycles difference.

Spiral scan - A pencil beam scan which moves around a central axis describing the surface of a cone in an ever-increasing cone size.

Spot jamming - Narrow frequency band jamming concentrated against radar at a particular frequency. The jamming bandwidth is comparable to the radar bandpass. Used to deny range and sometimes angle information.

Spread spectrum - Use of broader frequency bandwidths than normally required to transmit information, in order to gain advantages in interference rejection (anti-jam), message privacy/security, low probability of intercept (LPI), deny frequency resolution, multiplexing of more than one message in the same bandwidth, or high resolution range measurement.

Spurious radiation - Emissions from a radio transmitter at frequencies outside its assigned or intended emission frequency. Spurious emission includes harmonic emission, parasitic emission, and intermodulation products, but excludes emissions in the immediate vicinity of the necessary band that are a result of the modulation process for the transmission of information.

Spurious response –

1. Any response, other than the desired response, of an electric transducer or device.

2. A term used in electronic warfare to describe the undesirable signal images in the intercept receiver resulting from the mixing of the intercepted signal with harmonics of the local oscillators in the receiver.

Staggered PRF - A technique of using more than one PRF to reduce the blind speeds associated with MTI radars. PRF switching occurs on a pulse-to-pulse basis.

Stand-off jammer (SOJ) - A powerful jammer that remains at a safe range while providing jamming coverage (masking) for the attacking elements. The Navy EA-6B is an example of a SOJ aircraft.

Stand-off jamming (SOJ) - A jamming aircraft that orbits outside the maximum range of the SAM threat. As the attack package penetrates, the jamming aircraft directs jamming against all significant radars in the area.

Stream - Dispensing of chaff (solid/random interval/bursts).

Stream chaff - Operational technique of dropping large quantities of chaff for a continuous period of time. This results in a "ribbon" or "stream" of returns many miles in length on victim radar scopes. The penetrating attack package can then use the resulting chaff corridor to mask their penetration.

Super high frequency (SHF) - Frequencies from 3 to 30 GHz.

Support jamming - A tactic by which aircraft carrying electronic jamming equipment orbit at a safe distance from the enemy threat defenses or fly escort with the strike force for the primary purpose of screening them from the threat radars.

Suppression of enemy air defenses (SEAD) - That activity which neutralizes, destroys, or temporarily degrades enemy air defense systems in a specific area by using physical attack, deception, and/or electronic warfare.

Surface-to-air missile (SAM) - A missile launched from a surface launcher at a target above the surface.

Sweep jammer - Electronic jammer that sweeps a narrow band of electronic energy over a broad bandwidth.

Sweep lock-on jammer - A transmitter in which a narrowband jamming signal can be tuned over a broad frequency band and the signal locked on a particular frequency.

Swept audio - Jamming technique usually employed against conical scan-on-receive-only (COSRO) radars. The received pulses are amplified and retransmitted by the target and amplitude modulated at a frequency close to the suspected receiver antenna scan frequency.

Swept jamming - An EA technique of barrage jamming in which a CW carrier or noise source is swept over a selected bandwidth.

Swept-spot jamming - A jamming technique in which an oscillator is swept over a specific range of frequencies in the band of interest in order to be assured of exciting a receiver tuned to any frequency in that band.

Synchronized-pulse jamming - The technique of attempting to insert jamming pulses into a receiver each time the receiver gate opens.

Synchronous-pulsed jamming - A jamming technique that matches exactly the PRF of the victim's radar; then transmits multiples of the PRF. It is most effective if the jammer also matches the PW of the radar. Synchronous-pulsed jamming is easily recognized since the spacing between successive target lines is equal and each target line is the same in depth from the center outward. The width of the jammed sector is dependent upon the range of the jammer from the radar.

Synthetic aperture radar (SAR) - A high-resolution ground mapping technique in which advantage is taken of the forward motion of a coherent pulsed radar to synthesize the equivalent of a very long sidelooking array antenna from the radar returns received over a period of up to several seconds or more.

Target acquisition - The detection, identification, and location of a target in sufficient detail to permit the effective employment of weapons.

Terminal guidance –

1. The guidance applied to a guided missile between mid-course and arrival in the vicinity of the target.

2. Electronic, mechanical, visual, or other assistance given to aircraft pilots to facilitate arrival at, operation within or over, landing upon or departure from an air landing or air drop facility.

Terminal threat - The weapon systems, generally near a target, used to directly engage an aircraft in order to destroy it.

Terrain-avoidance radar - An airborne radar that provides a display of terrain ahead of low-flying aircraft to permit horizontal avoidance of obstacles.

Terrain-following radar (TFR) - An airborne radar that provides a display of terrain ahead of low-flying aircraft to permit manual control, or signals for automatic control to maintain constant altitude above the ground.

Thermal crossover - The natural phenomenon which normally occurs twice daily when temperature conditions are such that there is a loss of contrast between two adjacent objects on infrared imagery.

Threshold - The minimum value of a signal that can be detected by a system or sensor under consideration.

Time-of-arrival (TOA) - A method of locating a distant pulse emitter by measuring the difference in the time-of-arrival of its pulses at three separate locations. This method is also called Inverse LORAN.

Track –

1. A series of related contacts displayed on a plotting board.

2. To display or record the successive positions of a moving object.

3. To lock onto a point of radiation and obtain guidance from it.

4. To keep a gun properly aimed, or to continuously point a target-locating instrument at a moving target.

5. The actual path of an aircraft above, or a ship on, the surface of the earth. The course is the path that is planned; the track is the path that is taken.

Tracking - The continuous monitoring of range, velocity, or position of a target in space from a reference position. This is accomplished via radar and/or optical means.

Tracking radar - A radar that measures the range, azimuth, elevation, and/or velocity of the target and provides data that may be used by the fire control computer to determine the target path and predict its future position.

Track-on-jam - A method of passive target tracking using the jamming signal emitted by the target.

Track-while-scan (TWS) radar - Although it is not really a tracking radar in the true sense of the word, it does provide complete and accurate position information for missile guidance by using two separate beams produced by two separate antennas on two different frequencies. The system uses electronic computer techniques whereby raw data are used to track an assigned target, compute target velocity, and predict its future position while maintaining normal sector scan.

Transceiver - A combined radio transmitter and receiver in which some circuits other than those of the power supply are common to both transmitter and receiver, and not providing for simultaneous transmission and reception.

Transponder - A transmitter-receiver capable of accepting the electronic challenge of an interrogator and automatically transmitting an appropriate reply.

Traveling wave tube (TWT) - An electron tube in which a beam of electrons interacts continuously with a guided electromagnetic wave to produce amplification at microwave frequencies. A TWT capable of providing high amplification (60 dB) in frequencies over several octaves without adjustment. TWTs are classified by:

1. **Power Output:**

    Low (less than one watt)
    High (more than 10 watts)
    Intermediate (one to 10 watts)

2. **Noise Characteristics:**

   Low (less than 20 dB)

3. **Operating Mode:**

   Pulse
   CW
   Dual Mode (pulse or CW)

4. **Method of Focusing:**

   PPM (periodic permanent magnet)
   Solenoid (single electromagnet)
   Coupled Cavity

5. **Terms associated with TWT operation:**

   a. **Gain Compression:** Change in amplification a device provides as it operates near saturation.

   b. **Overdrive:** TWT is operating beyond the point of saturation.

   c. **Saturation Gain:** Ratio of output power to input power when the device is being driven at maximum output.

   d. **Serrodyne:** Linear translation of the phase of a signal on the helix of a TWT by a linear sawtooth waveform that enables the TWT to operate as a single sideband frequency translator.

   e. **Small Signal Gain:** Ratio of output power to input power when the TWT is operating linearly.

Two-signal jamming (also called straddle jamming) - Jamming whereby two signals are transmitted on two RF frequencies slightly separated. Effective against certain types of radar where receiver bandwidth is narrow enough to defeat noise jamming.

Ultra high frequency (UHF) - Frequencies from 300 to 3,000 MHz.

Unit prefixes - Prefixes used to indicate scientific units:

Multiple	Prefix	Symbol
10^{12}	tera	T
10^{9}	giga	G
10^{6}	mega	M
10^{3}	kilo	k
10^{2}	hecto	h
10^{1}	deka	da
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	µ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

Unmodulated CW jamming - The transmission of a high power carrier frequency that causes an overload effect to occur in the radar receiver. Used against bandpass radars that have a limited tuning capability. Unmodulated CW jamming can be identified by a blackening of the scope background (no video present) in a wedge-shaped sector, or by a solid brightening of a wedge or sector, normally exceeding one bandwidth.

Uplink - The missile guidance signal that passes command guidance intelligence from the site to the missile.

Velocity gate pull-off (VGPO) - Method of capturing the velocity gate of a Doppler radar and moving it away from the skin echo. Similar to the RGPO, but used against Doppler radar systems. The target Doppler frequency, which is amplified and retransmitted, is shifted in frequency to provide an apparent rate change or Doppler shift.

Very high frequency (VHF) - Frequencies from 30 to 300 MHz.

Very low frequency (VLF) - Frequencies from 3 to 30 kHz.

Video frequency –

1. A band of frequencies extending from approximately 100 Hz to several MHz.

2. The frequency of the voltage resulting from television scanning. Range from zero to 4 MHz or more.

Warning receiver - A receiver with the primary function of warning the user that his unit is being illuminated by an electromagnetic signal of interest.

Wideband constant false alarm rate receiver (WB-1) - Used against individual or combinations of rapidly swept FM-CW, noise, or CW jamming. This mode has nonlinear limiting and gives poor resolution of overlapping targets.

Window –

1. Strips of frequency-cut metal foil, wire, or bars that may be dropped from aircraft or missiles, or expelled from shells or rockets as a radar countermeasure. A confusion reflector.

2. A passive radar deception or confusion device; usually consisting of some metallic structure, designed in size and shape to effectively reflect impinging signals, to simulate a true target.

3. British name for chaff.

Wobbulation - A periodic and usually slow variation of a parameter (such as frequency, amplitude, or period) about a central value. A low frequency modulation of a jamming carrier that appears on a radar display as a "wobbling" target. This effect is an undesired result of an angle deception or time-varying barrage technique.

Wooden round - An ordnance round (shell, missile, etc.) requiring no maintenance or preparation time prior to loading for firing.

ANNEX B
EC-RELATED ABBREVIATIONS AND ACRONYMS

A

AAA	**Antiaircraft artillery**
AAM	**Air-to-air missile**
AAS	**ARM alarm sensor**
AC	**Alternating current**
ACET	**Automatic cancellation of extended targets**
ACQ	**Acquisition**
AC&W	**Aircraft control and warning**
ADA	**Air defense artillery**
ADF	**Automatic direction finding**
ADL	**Automatic data link**
AEW	**Airborne early warning**
AF	**Air Force/audio frequency**
AFB	**Air Force Base**
AFC	**Automatic frequency control**
AFRES	**Air Force Reserves**
AGC	**Automatic gain control**
AGI	**Auxiliary general intelligence (ship)**
AGL	**Above ground level**
AGM	**Air-to-ground missile**
AI	**Airborne interceptor**
AJ	**Anti-jamming**
AM	**Amplitude modulation**
AM-CW	**Amplitude modulated continuous wave**
AMOP	**Amplitude modulation on pulse**
AMP	**Amplifier**
AMRAAM	**Advanced medium range air-to-air missile**
AMTI	**Airborne moving target indicator**
ANG	**Air National Guard**
ANL	**Automatic noise limiting**
AOA	**Angle of arrival/angle of attack**
AOJ	**Angle-on-jam**
ARM	**Antiradiation missile**
ASK	**Amplitude shift keying**
ASM	**Air-to-surface missile**
ASR	**Air surveillance radar**
ASW	**Anti-submarine warfare**
ATC	**Air traffic control**
ATD	**Automatic target detector**
ATR	**Antenna, transmit receive**

ATV	**Automatic threshold variation**
AVNL	**Automatic video noise leveling**
AVP	**Adaptive video processing**
AWACS	**Airborne warning and control system**
AZ	**Azimuth**

B

BBC	**Beam-to-beam correlation**
BCC	**Battery control center**
BFO	**Beat frequency oscillator**
BIT	**Built-in test**
BMEWS	**Ballistic Missile Early Warning System**
BN	**Barrage noise**
BSE	**Boresight error**
BTE	**Battery terminal equipment**
BW	**Bandwidth**
BWO	**Backward wave oscillator**

C

C	**Degree centigrade**
C^2	**Command and control**
C^3	**Command, control, and communications**
C^2W	**Command and control warfare**
CAP	**Combat air patrol**
CAS	**Close air support**
CE	**Clutter eliminator**
CEP	**Circular error probable**
CFAR	**Constant false alarm rate**
CIJ	**Close-in jamming**
CIU	**Control indicator unit**
CM	**Countermeasures**
COHO	**Coherent oscillator**
COMINT	**Communications intelligence**
COMSEC	**Communication security**
CONSCAN	**Conical scan**
COSRO	**Conical scan-on-receive-only**
CPAC	**Coded pulse anti-clutter**
CP	**Circular polarization**
CRT	**Cathode ray tube**
CSG	**Clean strobe generator**
CSLC	**Coherent sidelobe canceler**
CV	**Coincident video**
CVR	**Crystal video receiver**

CW	**Continuous wave**
CWAR	**Continuous wave acquisition radar**
CWI	**Continuous wave illuminator**

D

dB	**Decibel**
dBm	**Decibels relative to 1 milliwatt**
dBw	**Decibels relative to 1 watt**
DBB	**Detector back bias**
DBF	**Digital beam forming**
DC	**Direct current**
DET	**Detector**
DF	**Direction finder/Dicke fix**
DINA	**Direct amplified noise**
DIFM	**Digital instantaneous frequency measurement**
DME	**Distance measuring equipment**
DOA	**Direction of arrival/dead on arrival**
DOF	**Degrees of freedom**
DPI	**Detected pulse interference**
DR	**Dynamic range/dead reckoning**
DRFM	**Digital radio frequency memory**
DSB	**Double sideband**
DTM	**Data transfer module**

E

EA	**Electronic attack**
EC	**Electronic combat**
ECCM	**Electronic counter-countermeasures**
ECM	**Electronic countermeasures**
ECP	**Electronic combat pilot**
EDAC	**Error detection and correction**
EHF	**Extremely high frequency**
EID	**Emitter identification (table)**
EL	**Elevation**
ELINT	**Electronic intelligence**
ELSCAN	**Elevation scan**
EMCON	**Emission control**
EMI	**Electromagnetic interference**
EMP	**Electromagnetic pulse**
EO	**Electro-optical**
EOB	**Electronic order of battle**
EP	**Electronic protection**
ER	**Electronic reconnaissance**

ERP	**Effective radiated power**
ES	**Electronic warfare support**
EVIL	**Evaluated vs. integrated log**
EW	**Electronic warfare**
EWO	**Electronic warfare officer**
EWWS	**Electronic warfare warning set**

F

F	**Degrees Fahrenheit**
FA	**Frequency agility**
FC	**Fire control**
FAAR	**Forward area alerting radar**
FAGC	**Fast automatic gain control**
FCR	**Fire control radar**
FCS	**Fire control system**
FEBA	**Forward edge of the battle area**
FFO	**Fixed frequency oscillator**
FFT	**Fast Fourier transform**
FLIR	**Forward looking infrared**
FM	**Frequency modulation**
FM-CW	**Frequency modulation continuous wave pulsed**
FMOP	**Frequency modulation on pulse**
FOJ	**Fuse-on-jam**
FOR	**Field of regard**
FOTD	**Fiber optic towed decoy**
FOV	**Field of view**
FRESCAN	**Frequency scanning**
FS	**Frequency sector**
FSK	**Frequency shift keying**
ft	**Feet**
FTC	**Fast time constant**

G

G	**Antenna gain/acceleration due to gravity**
GAS	**Ground aided seeker**
GCA	**Ground controlled approach**
GCI	**Ground controlled intercept**
GEN	**Generator**
GHz	**Gigahertz**

H

HARM	**High-speed antiradiation missile**
HAWK	**Homing all-the-way killer**
HBW	**Horizontal beam width**
HF	**High frequency/height finder**
HFDF	**High frequency direction finding**
HIPIR	**High-power illuminator radar**
HOJ	**Home-on-jam**
HOT	**Home-on-target**
HQ	**Headquarters**
HUMINT	**Human intelligence**
HVP	**High video pass**
HVPS	**High voltage power supply**
Hz	**Hertz**

I

I	**Improved**
IADS	**Integrated air defense system**
IAGC	**Instantaneous automatic gain control**
ICD	**Imitative communications deception**
ICW	**Interrupted continual wave**
IF	**Intermediate frequency**
IFC	**Instantaneous frequency correlator**
IFF	**Identification, friend or foe**
IPM	**Instantaneous position memory**
IR	**Infrared**
IRCM	**Infrared countermeasures**

J

JAT	**Jam angle track**
JAVA	**Jamming amplitude versus azimuth**
JCS	**Joint Chiefs of Staff**
JLT	**Jammer look-through**
J/S	**Jamming-to-signal (ratio)**
JTIDS	**Joint Tactical Information Distribution System**

K

K	**Degrees Kelvin**
kHz	**Kilohertz**
km	**Kilometer**

KO	**Keyed oscillation**
kV	**Kilovolts**
kW	**Kilowatts**

L

LASER	**Light amplification by stimulated emission of radiation**
LET	**Leading-edge tracker**
LF	**Low frequency**
LFMOP	**Linear frequency modulation on pulse (chirp)**
LIN	**Linear**
LIN-LOG	**Linear-logarithmic amplifier**
LLLTV	**Low light level television**
LO	**Local oscillator**
LOG	**Logarithmic**
LOG-FTC	**Logarithmic receiver with fast time constant**
LORO	**Lobe-on-receive-only**
LOS	**Line of sight**
LPI	**Low probability of intercept**
LT	**Listening time**
LVA	**Log video amplifier**

M

M	**Meter**
MALD	**Miniature air launched decoy**
MASER	**Microwave amplification by stimulated emission of radiation**
MANPADS	**Manportable air defense system**
MAWS	**Missile approach warning system**
MBB	**Main beam blanking**
MCG	**Midcourse guidance**
MCD	**Manipulative communications deception**
MDT	**Mission data type**
MDF	**Mission data file**
MDG	**Mission data generator**
MF	**Medium frequency**
MFAR	**Multifunction array radar**
MG	**Missile guidance**
MGC	**Manual gain control**
MHz	**Megahertz**
MIC	**Microwave integrated circuit**
MIJI	**Meaconing, Intrusion, Jamming, and Interference**
MLC	**Mainlobe cancellation**
MLV	**Memory loader verifier**
mm	**Millimeter**

MMIC	**Monolithic microwave integrated circuit**
MMW	**Millimeter wave**
MOE	**Measure of Effectiveness**
MOPA	**Master oscillator power amplifier**
msec	**Millisecond**
MTD	**Moving target detector**
MTI	**Moving target indicator**
MUX	**Multiplexer**
MW	**Megawatts**

N

NBFM	**Narrowband frequency modulation**
NBL	**Narrowband limiting**
NBLP	**Narrowband long pulse**
nm	**Nautical mile**
NLP	**Narrowband long pulse**
NOTAM	**Notice to airmen**
NSPS	**Nonsynchronous pulse suppression**

O

OFP	**Operational flight program**
OOB	**Order of battle**
OOK	**On-off keying**
OOK CW	**Interrupted CW**
OPDEC	**Operational deception**
OPSEC	**Operational security**
OTH	**Over-the-horizon**

P

PAM	**Pulse amplitude modulation**
PAR	**Pulse acquisition radar/precision approach radar**
PAT	**Passive angle track**
PC	**Pulse compression**
PCM	**Pulse code modulation**
PD	**Pulse Doppler/pulse duration (same as PW)**
PDM	**Pulse duration modulation**
PENAIDS	**Penetration aids**
PFM	**Pulse frequency modulation**
PFN	**Pulse-forming network**
PHOTOINT	**Photographic intelligence**
PIE	**Pulse interference elimination**

PISAB	**Pulse interference suppression and blanking**
PLD	**Pulse length discrimination (same as PWD)**
PM	**Phase modulation**
PN	**Pseudo noise**
PPC	**Pulse-to-pulse correlation**
PPI	**Plan position indicator**
PPM	**Pulse position modulation**
PPS	**Pulses per second**
PRF	**Pulse repetition frequency**
PRI	**Pulse repetition interval**
PRT	**Pulse recurrence time**
PSK	**Phase shift keying**
PW	**Pulse width**
PWD	**Pulse width discrimination**
PWM	**Pulse width modulation**

R

RADAR	**Radio detection and ranging**
RAGM	**Range angle gate memory**
RAM	**Radar absorbing material/repeater amplitude modulation**
RANRAP	**Random range program**
RC	**Resolution cell**
RCS	**Radar cross section**
RDF	**Radio direction finding**
RF	**Radio frequency**
RFD	**Radio frequency discriminator**
RFI	**Radio frequency interference**
RGM	**Range gate memory**
RGPI	**Range gate pull in**
RGPO	**Range gate pull off**
RHAW	**Radar homing and warning**
RHI	**Range height indicator**
ROB	**Radar order of battle**
ROR	**Range only radar**
R/P	**Receiver/processor**
RPD	**Random pulse discrimination**
RPJ	**Random pulse jamming**
RPM	**Revolutions per minute**
RPS	**Random pulse suppression**
RPV	**Remotely piloted vehicle**
RSBN	**Radio short-range beacon for navigation**
RT	**Recovery time**
R/T	**Receiver/transmitter**
RWR	**Radar warning receiver**

S

SAM	**Surface-to-air missile**
SAR	**Synthetic aperture radar**
SASE	**Semi-automatic support equipment**
SATCOM	**Satellite communications**
SBB	**Single beam blanking**
SCV	**Sub-clutter visibility**
SEAD	**Suppression of enemy air defenses**
sec	**Seconds**
SHF	**Super high frequency**
SIF	**Selective identification feature**
SIGINT	**Signals intelligence**
S/J	**Signal-to-jamming (ratio)**
SLAR	**Side-looking airborne radar**
SLB	**Sidelobe blanking**
SLC	**Sidelobe cancellation**
SLS	**Sidelobe suppression**
S/N	**Signal-to-noise (ratio)**
SOJ	**Stand-off jamming**
SORO	**Scan-on-receive-only**
SRM	**Swept rectangular modulation**
SS	**Swept-spot**
SSB	**Single sideband**
SSN	**Swept-spot noise/social security number**
SRW	**Swept rectangular wave**
SSW	**Swept square wave**
STALO	**Stable local oscillator**
STC	**Sensitivity time control**
SWC	**Scan with compensation**

T

TA	**Target acquisition**
TACAN	**Tactical air navigation**
TALD	**Tactical air launched decoy**
TBM	**Track break modulation**
TDD	**Target detection device**
TDOA	**Time difference of arrival**
TEL	**Transporter-erector-launcher**
TELAR	**Transporter-erector-launcher and radar**
TET	**Trailing edge tracker**
TEWS	**Tactical electronic warfare system**
TFR	**Terrain following radar**
TGT	**Target**
TJS	**Tactical jamming system**

TOA	**Time of arrival**
TOF	**Time of flight**
TOJ	**Track-on-jam**
TPL	**Transmitter pulse lengthening**
TPS	**Transmitter pulse shaping**
TTR	**Target tracking radar**
TV	**Television**
TVM	**Track-via-missile**
TWS	**Track-while-scan**
TWT	**Traveling wave tube**

U

UHF	**Ultra high frequency**
µsec	**Microseconds**

V

VCO	**Voltage controlled oscillator**
VDL	**Video data link**
VGPO	**Velocity gate pull off**
VHF	**Very high frequency**
VHSIC	**Very high speed integrate circuit**
VINT	**Video integration**
VLF	**Very low frequency**
VSRW	**VGPO with swept rectangular wave**
VSWR	**Voltage standing-wave ratio**

W

W	**Watts**
WBFM	**Wideband frequency modulation**
WBL	**Wideband limiting**
WPB	**Wide pulse blanking**
WSP	**Wideband short pulse**

LIST OF REFERENCES

1. **Adamy, D., “Seduction Decoys”, *Journal of Electronic Defense*, Vol. 20, No 7, pp. 58-59, July, 1997.**

2. **Adamy, D., “EW 101”, *Journal of Electronic Defense*, Vol. 21, No 1, pp. 14-19, January, 1998.**

3. **Bar-Shalom, Y., *Multitarget-Multisensor Tracking: Advanced Applications*, Artec House, Inc., Norwood, MA, 1990.**

4. **Chrzanowshi, E. J., *Active Radar Electronic Countermeasures*, Artec House, Inc., Norwood, MA, 1990.**

5. **Det 8, Training Support Squadron, *Radar Concepts for Electronic Warfare*, January 1997.**

6. **Deyerle, C. M., “Advanced Infrared Missile Counter-Countermeasures”, *Journal of Electronic Defense*, Vol. 17, No 1, pp. 47-70, January, 1994.**

7. **Lothes, R. N. et al., “Radar Vulnerability to Jamming”, Artec House, Inc., Norwood, MA, 1990.**

8. **Maksimov, M. V. et al., *Radar Anti-jamming Techniques*, Artec House, Inc., Norwood, MA, 1979.**

9. **Neri, F., *Introduction to Electronic Defense Systems*, Artec House, Inc., Norwood, MA, 1990.**

10. **Schleher, D. C., *Introduction to Electronic Warfare*, Artec House, Inc., Norwood, MA, 1986.**

11. **Skolnik, M. L., *Introduction to Radar Systems*, McGraw-Hill, Inc., New York, NY, 1980.**

12. **Stimson, G. W., *Introduction to Airborne Radar*, 2nd Edition, SciTech Publishing, Inc., Mendham, NJ, 1998.**

13. **USAF Air Warfare Center, *Radar, ECM and ECCM Fundamentals*, January 1992.**

14. Wiegand, R. J., *Radar Electronic Countermeasures System Design*, Artec House, Inc., Norwood, MA, 1991.

15. Wiley, R. G., *Electronic Intelligence: The Interception of Radar Signals*, Artec House, Inc., Norwood, MA, 1985.

16. 4513 ATTG/INW, *Advanced Radar Principles for Electronic Combat*, 15 April 1991.

17. 4513 ATTG/INW, *Basic Radar for Electronic Warfare*, 1 April 1990.

18. 4513 ATTG/INW, *Principles of Chaff and Flares*, 15 April 1991.

FAA-H-8083-15A

With editorial change

Instrument Flying Handbook

UP VERTICAL SPEE
THOUSAND FT PE
.5 DOWN
AIRSPEED
KNOTS
4000
4200
4100
3900
100
90
80
103KT
270°
VOR 1

U.S. Department
of Transportation
**FEDERAL AVIATION
ADMINISTRATION**

Instrument Flying Handbook

2008

U.S. Department of Transportation
FEDERAL AVIATION ADMINISTRATION
Flight Standards Service

Preface

This Instrument Flying Handbook is designed for use by instrument flight instructors and pilots preparing for instrument rating tests. Instructors may find this handbook a valuable training aid as it includes basic reference material for knowledge testing and instrument flight training. Other Federal Aviation Administration (FAA) publications should be consulted for more detailed information on related topics.

This handbook conforms to pilot training and certification concepts established by the FAA. There are different ways of teaching, as well as performing, flight procedures and maneuvers and many variations in the explanations of aerodynamic theories and principles. This handbook adopts selected methods and concepts for instrument flying. The discussion and explanations reflect the most commonly used practices and principles. Occasionally the word "must" or similar language is used where the desired action is deemed critical. The use of such language is not intended to add to, interpret, or relieve a duty imposed by Title 14 of the Code of Federal Regulations (14 CFR).

All of the aeronautical knowledge and skills required to operate in instrument meteorological conditions (IMC) are detailed. Chapters are dedicated to human and aerodynamic factors affecting instrument flight, the flight instruments, attitude instrument flying for airplanes, basic flight maneuvers used in IMC, attitude instrument flying for helicopters, navigation systems, the National Airspace System (NAS), the air traffic control (ATC) system, instrument flight rules (IFR) flight procedures, and IFR emergencies. Clearance shorthand and an integrated instrument lesson guide are also included.

This handbook supersedes FAA-H-8081-15A, Instrument Flying Handbook, dated 2007.

This handbook may be purchased from the Superintendent of Documents, United States Government Printing Office (GPO), Washington, DC 20402-9325, or from GPO's web site.

http://bookstore.gpo.gov

This handbook is also available for download, in PDF format, from the Regulatory Support Division's (AFS-600) web site.

http://www.faa.gov/about/office_org/headquarters_offices/avs/offices/afs/afs600

This handbook is published by the United States Department of Transportation, Federal Aviation Administration, Airman Testing Standards Branch, AFS-630, P.O. Box 25082, Oklahoma City, OK 73125.

Comments regarding this publication should be sent, in email form, to the following address.

AFS630comments@faa.gov

Record of Changes

The following editorial changes were made in 2008 that supersede the 2007 publication of FAA-H-8081-15A, Instrument Flying Handbook:

Page 2-2, corrections made to Figure 2-1. The Airfoil.

Page 2-3, correction to paragraph concerning induced drag.

Page 2-7, corrections made to Figure 2-8. Thrust and Power Required Curves.

Page 2-11, correction to paragraph concerning the coordination of rudder and aileron controls.

Page 2-12, correction to Figure 2-15. Adverse Yaw.

Page 3-4, correction to Figure 3-3. Sensitive Altimeter Components.

Page 3-5, correction to Figure 3-6. The loss of altitude experienced when flying into an area where the air is colder (more dense) than standard.

Page 4-2, correction to Figure 4-1. Control Instruments.

Page 4-3, correction to Figure 4-2. Performance Instruments.

Page 4-18, correction to Figure 4-25. Control Instruments.

Page 6-11, correction to Figure 6-12. Flight instrument indications in a stabilized constant-airspeed climb.

Page 6-12, correction to Figure 6-13. Flight Instrument Indications in a Stabilized Constant-Rate Climb.

Page 7-7, correction to Figure 7-7. ADF Tracking Inbound.

Page 7-9, corrections made to Figure 7-8. ADF Interception and Tracking Outbound.

Page 7-13, correction made to Figure 7-15. CDI Interpretation.

Page 7-15, correction made to Figure 7-16. Course Interception (VOR).

Page 7-18, correction made to Figure 7-18. Using DME and RMI to Maintain an Arc.

Page 7-27, correction made to paragraph concerning GPS Components.

Page 7-41, corrections made to Figure 7-37. Precision and Nonprecision ALS Configuration.

Page 10-12, correction made to paragraph concerning a parallel procedure.

Acknowledgments

This handbook was produced by the Federal Aviation Administration (FAA) with the assistance of Safety Research Corporation of America, LLC. The FAA wishes to acknowledge the following contributors:

The laboratory of Dale Purves, M.D. and Mr. Al Seckel in providing imagery (found in Chapter 1) for visual illusions from the book, *The Great Book of Optical Illusions*, Firefly Books, 2004

Sikorsky Aircraft Corporation and Robinson Helicopter Company for imagery provided in Chapter 9

Garmin Ltd. for providing flight system information and multiple display systems to include integrated flight, GPS and communication systems; information and hardware used with WAAS, LAAS; and information concerning encountering emergencies with high-technology systems

Universal Avionics System Corporation for providing background information of the Flight Management System and an overview on Vision–1 and Traffic Alert and Collision Avoidance systems (TCAS)

Meggitt/S-Tec for providing detailed autopilot information regarding installation and use

Cessna Aircraft Company in providing instrument panel layout support and information on the use of onboard systems

Kearfott Guidance and Navigation Corporation in providing background information on the Ring-LASAR gyroscope and its history

Honeywell International Inc., for Terrain Awareness Systems (TAWS) and various communication and radio systems sold under the Bendix-King name

Chelton Flight Systems and Century Flight Systems, Inc., for providing autopilot information relating to Highway in the Sky (Chelton) and HSI displays (Century)

Avidyne Corporation for providing displays with alert systems developed and sold by Ryan International, L3 Communications, and Tectronics.

Additional appreciation is extended to the Aircraft Owners and Pilots Association (AOPA), the AOPA Air Safety Foundation, and the National Business Aviation Association (NBAA) for their technical support and input.

Introduction

Is an Instrument Rating Necessary?

The answer to this question depends entirely upon individual needs. Pilots may not need an instrument rating if they fly in familiar uncongested areas, stay continually alert to weather developments, and accept an alternative to their original plan. However, some cross-country destinations may take a pilot to unfamiliar airports and/or through high activity areas in marginal visual or instrument meteorological conditions (IMC). Under these conditions, an instrument rating may be an alternative to rerouting, rescheduling, or canceling a flight. Many accidents are the result of pilots who lack the necessary skills or equipment to fly in marginal visual meteorological conditions (VMC) or IMC and attempt flight without outside references.

Pilots originally flew aircraft strictly by sight, sound, and feel while comparing the aircraft's attitude to the natural horizon. As aircraft performance increased, pilots required more inflight information to enhance the safe operation of their aircraft. This information has ranged from a string tied to a wing strut, to development of sophisticated electronic flight information systems (EFIS) and flight management systems (FMS). Interpretation of the instruments and aircraft control have advanced from the "one, two, three" or "needle, ball, and airspeed" system to the use of "attitude instrument flying" techniques.

Navigation began by using ground references with dead reckoning and has led to the development of electronic navigation systems. These include the automatic direction finder (ADF), very-high frequency omnidirectional range (VOR), distance measuring equipment (DME), tactical air navigation (TACAN), long range navigation (LORAN), global positioning system (GPS), instrument landing system (ILS), microwave landing system (MLS), and inertial navigation system (INS).

Perhaps you want an instrument rating for the same basic reason you learned to fly in the first place—because you like flying. Maintaining and extending your proficiency, once you have the rating, means less reliance on chance and more on skill and knowledge. Earn the rating—not because you might need it sometime, but because it represents achievement and provides training you will use continually and build upon as long as you fly. But most importantly it means greater safety in flying.

Instrument Rating Requirements

A private or commercial pilot must have an instrument rating and meet the appropriate currency requirements if that pilot operates an aircraft using an instrument flight rules (IFR) flight plan in conditions less than the minimums prescribed for visual flight rules (VFR), or in any flight in Class A airspace.

You will need to carefully review the aeronautical knowledge and experience requirements for the instrument rating as outlined in Title 14 of the Code of Federal Regulations (14 CFR) part 61. After completing the Federal Aviation Administration (FAA) Knowledge Test issued for the instrument rating, and all the experience requirements have been satisfied, you are eligible to take the practical test. The regulations specify minimum total and pilot-in-command time requirements. This minimum applies to all applicants regardless of ability or previous aviation experience.

Training for the Instrument Rating

A person who wishes to add the instrument rating to his or her pilot certificate must first make commitments of time, money, and quality of training. There are many combinations of training methods available. Independent studies may be adequate preparation to pass the required FAA Knowledge Test for the instrument rating. Occasional periods of ground and flight instruction may provide the skills necessary to pass the required test. Or, individuals may choose a training facility that provides comprehensive aviation education and the training necessary to ensure the pilot will pass all the required tests and operate safely in the National Airspace System (NAS). The aeronautical knowledge may be administered by educational institutions, aviation-oriented schools, correspondence courses, and appropriately rated instructors. Each person must decide for themselves which training program best meets his or her needs and at the same time maintain a high quality of training. Interested persons should make inquiries regarding the available training at nearby airports, training facilities, in aviation publications, and through the FAA Flight Standards District Office (FSDO).

Although the regulations specify minimum requirements, the amount of instructional time needed is determined not by the regulation, but by the individual's ability to achieve a satisfactory level of proficiency. A professional pilot with diversified flying experience may easily attain a satisfactory level of proficiency in the minimum time required by regulation. Your own time requirements will depend upon a variety of factors, including previous flying experience, rate of learning, basic ability, frequency of flight training, type of aircraft flown, quality of ground school training, and quality of flight instruction, to name a few. The total instructional time you will need, the scheduling of such time, is up to the individual most qualified to judge your proficiency—the instructor who supervises your progress and endorses your record of flight training.

You can accelerate and enrich much of your training by informal study. An increasing number of visual aids and programmed instrument courses is available. The best course is one that includes a well-integrated flight and ground school curriculum. The sequential nature of the learning process requires that each element of knowledge and skill be learned and applied in the right manner at the right time.

Part of your instrument training may utilize a flight simulator, flight training device, or a personal computer-based aviation training device (PCATD). This ground-based flight training equipment is a valuable tool for developing your instrument cross-check and learning procedures, such as intercepting and tracking, holding patterns, and instrument approaches. Once these concepts are fully understood, you can then continue with inflight training and refine these techniques for full transference of your new knowledge and skills.

Holding the instrument rating does not necessarily make you a competent all-weather pilot. The rating certifies only that you have complied with the minimum experience requirements, that you can plan and execute a flight under IFR, that you can execute basic instrument maneuvers, and that you have shown acceptable skill and judgment in performing these activities. Your instrument rating permits you to fly into instrument weather conditions with no previous instrument weather experience. Your instrument rating is issued on the assumption that you have the good judgment to avoid situations beyond your capabilities. The instrument training program you undertake should help you to develop not only essential flying skills but also the judgment necessary to use the skills within your own limits.

Regardless of the method of training selected, the curriculum in Appendix B, Instrument Training Lesson Guide, provides guidance as to the minimum training required for the addition of an instrument rating to a private or commercial pilot certificate.

Maintaining the Instrument Rating

Once you hold the instrument rating, you may not act as pilot-in-command under IFR or in weather conditions less than the minimums prescribed for VFR, unless you meet the recent flight experience requirements outlined in 14 CFR part 61. These procedures must be accomplished within the preceding 6 months and include six instrument approaches, holding procedures, and intercepting and tracking courses through the use of navigation systems. If you do not meet the experience requirements during these 6 months, you have another 6 months to meet these minimums. If the requirements are still not met, you must pass an instrument proficiency check, which is an inflight evaluation by a qualified instrument flight instructor using tasks outlined in the instrument rating practical test standards (PTS).

The instrument currency requirements must be accomplished under actual or simulated instrument conditions. You may log instrument flight time during the time for which you control the aircraft solely by reference to the instruments. This can be accomplished by wearing a view-limiting device, such as a hood, flying an approved flight-training device, or flying in actual IMC.

It takes only one harrowing experience to clarify the distinction between minimum practical knowledge and a thorough understanding of how to apply the procedures and techniques used in instrument flight. Your instrument training is never complete; it is adequate when you have absorbed every foreseeable detail of knowledge and skill to ensure a solution will be available if and when you need it.

Table of Contents

Preface iii
Acknowledgements..v
Introduction..vii
Is an Instrument Rating Necessary? vii
Instrument Rating Requirements................................vii
Training for the Instrument Rating vii
Maintaining the Instrument Rating............................viii
Table of Contents ix

Chapter 1
Human Factors **1-1**

Introduction...1-1
Sensory Systems for Orientation 1-2
Eyes 1-2
Vision Under Dim and Bright Illumination............1-3
Ears..1-4
Nerves..1-5
Illusions Leading to Spatial Disorientation 1-5
Vestibular Illusions..1-5
The Leans 1-5
Coriolis Illusion 1-6
Graveyard Spiral..1-6
Somatogravic Illusion..1-6
Inversion Illusion 1-6
Elevator Illusion...1-6
Visual Illusions..1-7
False Horizon 1-7
Autokinesis 1-7
Postural Considerations...1-7
Demonstration of Spatial Disorientation 1-7
Climbing While Accelerating....................................1-8
Climbing While Turning..1-8
Diving While Turning..1-8
Tilting to Right or Left..1-8
Reversal of Motion..1-8
Diving or Rolling Beyond the Vertical Plane.............1-8
Coping with Spatial Disorientation...............................1-8
Optical Illusions..1-9
Runway Width Illusion...1-9
Runway and Terrain Slopes Illusion..........................1-9
Featureless Terrain Illusion..1-9
Water Refraction...1-9
Haze...1-9
Fog...1-9
Ground Lighting Illusions 1-9
How To Prevent Landing Errors Due To Optical Illusions..1-9
Physiological and Psychological Factors.....................1-11
Stress 1-11
Medical Factors...1-12
Alcohol 1-12
Fatigue 1-12
Acute Fatigue 1-12
Chronic Fatigue..1-13
IMSAFE Checklist 1-13
Hazard Identification 1-13
Situation 1 1-13
Situation 2 1-13
Risk Analysis...1-13
Crew Resource Management (CRM) and Single-Pilot Resource Management (SRM)....................................1-14
Situational Awareness...1-14
Flight Deck Resource Management............................1-14
Human Resources..1-14
Equipment 1-14
Information Workload 1-14
Task Management 1-15
Aeronautical Decision-Making (ADM).......................1-15
The Decision-Making Process.................................1-16
Defining the Problem 1-16
Choosing a Course of Action 1-16
Implementing the Decision and Evaluating the Outcome 1-16
Improper Decision-Making Outcomes 1-16
Models for Practicing ADM 1-17
Perceive, Process, Perform 1-17
The DECIDE Model..1-17
Hazardous Attitudes and Antidotes 1-18

Chapter 2
Aerodynamic Factors 2-1
Introduction 2-1
  The Wing 2-2
Review of Basic Aerodynamics 2-2
  The Four Forces 2-2
    Lift 2-2
    Weight 2-3
    Thrust 2-3
    Drag 2-3
  Newton’s First Law, the Law of Inertia 2-4
  Newton’s Second Law, the Law of Momentum 2-4
  Newton’s Third Law, the Law of Reaction 2-4
Atmosphere 2-4
  Layers of the Atmosphere 2-5
  International Standard Atmosphere (ISA) 2-5
    Pressure Altitude 2-5
    Density Altitude 2-5
Lift 2-6
  Pitch/Power Relationship 2-6
Drag Curves 2-6
  Regions of Command 2-7
    Control Characteristics 2-7
  Speed Stability 2-7
    Normal Command 2-7
    Reversed Command 2-8
Trim 2-8
Slow-Speed Flight 2-8
  Small Airplanes 2-9
  Large Airplanes 2-9
Climbs 2-10
  Acceleration in Cruise Flight 2-10
Turns 2-10
  Rate of Turn 2-10
  Radius of Turn 2-11
  Coordination of Rudder and Aileron Controls 2-11
Load Factor 2-11
Icing 2-12
Types of Icing 2-13
  Structural Icing 2-13
  Induction Icing 2-13
  Clear Ice 2-13
  Rime Ice 2-14
  Mixed Ice 2-14
  General Effects of Icing on Airfoils 2-14
    Piper PA-34-200T (Des Moines, Iowa) 2-15
  Tailplane Stall Symptoms 2-16
  Propeller Icing 2-16
  Effects of Icing on Critical Aircraft Systems 2-16
    Flight Instruments 2-16
    Stall Warning Systems 2-16
    Windshields 2-16
  Antenna Icing 2-17
Summary 2-17

Chapter 3
Flight Instruments 3-1
Introduction 3-1
Pitot/Static Systems 3-2
  Static Pressure 3-2
  Blockage Considerations 3-2
    Indications of Pitot Tube Blockage 3-3
    Indications from Static Port Blockage 3-3
    Effects of Flight Conditions 3-3
Pitot/Static Instruments 3-3
  Sensitive Altimeter 3-3
    Principle of Operation 3-3
    Altimeter Errors 3-4
    Cold Weather Altimeter Errors 3-5
  ICAO Cold Temperature Error Table 3-5
    Nonstandard Pressure on an Altimeter 3-6
    Altimeter Enhancements (Encoding) 3-7
    Reduced Vertical Separation Minimum (RVSM) 3-7
  Vertical Speed Indicator (VSI) 3-8
Dynamic Pressure Type Instruments 3-8
  Airspeed Indicator (ASI) 3-8
    Types of Airspeed 3-9
    Airspeed Color Codes 3-10
Magnetism 3-10
  The Basic Aviation Magnetic Compass 3-11
    Magnetic Compass Overview 3-11
    Magnetic Compass Induced Errors 3-12
  The Vertical Card Magnetic Compass 3-14
  The Flux Gate Compass System 3-14
  Remote Indicating Compass 3-15
Gyroscopic Systems 3-16
  Power Sources 3-16
    Pneumatic Systems 3-16
    Vacuum Pump Systems 3-17
    Electrical Systems 3-18
Gyroscopic Instruments 3-18
  Attitude Indicators 3-18
  Heading Indicators 3-19
  Turn Indicators 3-20
    Turn-and-Slip Indicator 3-20
    Turn Coordinator 3-21
Flight Support Systems 3-22
  Attitude and Heading Reference System (AHRS) 3-22
  Air Data Computer (ADC) 3-22
Analog Pictorial Displays 3-22
  Horizontal Situation Indicator (HSI) 3-22

  - Attitude Direction Indicator (ADI) 3-23
  - Flight Director System (FDS) 3-23
  - Integrated Flight Control System 3-24
    - Autopilot Systems...3-24
- Flight Management Systems (FMS)............................3-25
  - Electronic Flight Instrument Systems 3-27
- Primary Flight Display (PFD).......................................3-27
  - Synthetic Vision 3-27
  - Multi-Function Display (MFD)................................3-28
- Advanced Technology Systems....................................3-28
  - Automatic Dependent Surveillance—Broadcast (ADS-B) 3-28
- Safety Systems 3-30
  - Radio Altimeters..3-30
  - Traffic Advisory Systems 3-31
    - Traffic Information System 3-31
    - Traffic Alert Systems 3-31
    - Traffic Avoidance Systems 3-31
    - Terrain Alerting Systems 3-34
- Required Navigation Instrument System Inspection ...3-34
  - Systems Preflight Procedures...................................3-34
  - Before Engine Start 3-36
  - After Engine Start...3-37
  - Taxiing and Takeoff 3-37
  - Engine Shut Down 3-37

Chapter 4, Section I
Airplane Attitude Instrument Flying Using Analog Instrumentation 4-1

- Introduction..4-1
- Learning Methods 4-2
  - Attitude Instrument Flying Using the Control and Performance Method 4-2
    - Control Instruments 4-2
    - Performance Instruments 4-2
    - Navigation Instruments 4-2
    - Procedural Steps in Using Control and Performance 4-2
    - Aircraft Control During Instrument Flight 4-3
  - Attitude Instrument Flying Using the Primary and Supporting Method...4-4
    - Pitch Control 4-4
    - Bank Control...4-7
    - Power Control 4-8
    - Trim Control...4-8
    - Airplane Trim...4-8
    - Helicopter Trim 4-10
  - Example of Primary and Support Instruments 4-10
- Fundamental Skills...4-10
  - Instrument Cross-Check..4-10
    - Common Cross-Check Errors.............................4-11
  - Instrument Interpretation..4-13

Chapter 4, Section II
Airplane Attitude Instrument Flying Using an Electronic Flight Display 4-15

- Introduction..4-15
- Learning Methods 4-16
  - Control and Performance Method 4-18
    - Control Instruments 4-18
    - Performance Instruments 4-19
    - Navigation Instruments.......................................4-19
  - The Four-Step Process Used to Change Attitude 4-20
    - Establish...4-20
    - Trim 4-20
    - Cross-Check...4-20
    - Adjust...4-20
  - Applying the Four-Step Process 4-20
    - Pitch Control 4-20
    - Bank Control..4-20
    - Power Control 4-21
  - Attitude Instrument Flying—Primary and Supporting Method..4-21
    - Pitch Control 4-21
    - Straight-and-Level Flight 4-22
    - Primary Pitch 4-22
    - Primary Bank 4-23
    - Primary Yaw 4-23
    - Primary Power 4-23
- Fundamental Skills of Attitude Instrument Flying.......4-23
  - Instrument Cross-Check..4-24
- Scanning Techniques..4-24
  - Selected Radial Cross-Check 4-24
    - Starting the Scan 4-24
    - Trend Indicators 4-26
- Common Errors..4-28
  - Fixation..4-28
  - Omission...4-28
  - Emphasis 4-28

Chapter 5, Section I
Airplane Basic Flight Maneuvers Using Analog Instrumentation 5-1

- Introduction..5-1
- Straight-and-Level Flight...5-2
  - Pitch Control...5-2
    - Attitude Indicator 5-2
    - Altimeter 5-3
    - Vertical Speed Indicator (VSI)...............................5-4

Airspeed Indicator (ASI) 5-6
Bank Control 5-6
Attitude Indicator 5-6
Heading Indicator 5-7
Turn Coordinator 5-7
Turn-and-Slip Indicator (Needle and Ball) 5-8
Power Control 5-8
Power Settings 5-9
Airspeed Changes in Straight-and-Level Flight ... 5-11
Trim Technique 5-12
Common Errors in Straight-and-Level Flight 5-12
Pitch 5-12
Heading 5-13
Power 5-13
Trim 5-13
Straight Climbs and Descents 5-14
Climbs 5-14
Entry 5-14
Leveling Off 5-16
Descents 5-16
Entry 5-17
Leveling Off 5-17
Common Errors in Straight Climbs and Descents ... 5-17
Turns 5-19
Standard Rate Turns 5-19
Turns to Predetermined Headings 5-20
Timed Turns 5-21
Compass Turns 5-21
Steep Turns 5-22
Climbing and Descending Turns 5-24
Change of Airspeed During Turns 5-24
Common Errors in Turns 5-25
Pitch 5-25
Bank 5-25
Power 5-26
Trim 5-26
Errors During Compass Turns 5-26
Approach to Stall 5-26
Unusual Attitudes and Recoveries 5-26
Recognizing Unusual Attitudes 5-27
Recovery from Unusual Attitudes 5-27
Nose-High Attitudes 5-27
Nose-Low Attitudes 5-28
Common Errors in Unusual Attitudes 5-28
Instrument Takeoff 5-29
Common Errors in Instrument Takeoffs 5-29
Basic Instrument Flight Patterns 5-30
Racetrack Pattern 5-30
Procedure Turn 5-30
Standard 45° Procedure Turn 5-30
80/260 Procedure Turn 5-31
Teardrop Patterns 5-31
Circling Approach Patterns 5-32
Pattern I 5-32
Pattern II 5-32

Chapter 5, Section II
Airplane Basic Flight Maneuvers
Using an Electronic Flight Display 5-33

Introduction 5-33
Straight-and-Level Flight 5-34
Pitch Control 5-34
Attitude Indicator 5-34
Altimeter 5-36
Partial Panel Flight 5-36
VSI Tape 5-36
Airspeed Indicator (ASI) 5-37
Bank Control 5-37
Attitude Indicator 5-37
Horizontal Situation Indicator (HSI) 5-38
Heading Indicator 5-38
Turn Rate Indicator 5-38
Slip/Skid Indicator 5-39
Power Control 5-39
Power Settings 5-39
Airspeed Changes in Straight-and-Level Flight ... 5-40
Trim Technique 5-43
Common Errors in Straight-and-Level Flight 5-43
Pitch 5-43
Heading 5-44
Power 5-45
Trim 5-45
Straight Climbs and Descents 5-46
Entry 5-46
Constant Airspeed Climb From Cruise Airspeed ... 5-46
Constant Airspeed Climb from Established Airspeed ... 5-47
Constant Rate Climbs 5-47
Leveling Off 5-48
Descents 5-49
Entry 5-49
Leveling Off 5-50
Common Errors in Straight Climbs and Descents ... 5-50
Turns 5-51
Standard Rate Turns 5-51
Establishing A Standard Rate Turn 5-51
Common Errors 5-51
Turns to Predetermined Headings 5-52

Timed Turns 5-53
Compass Turns 5-53
Steep Turns 5-53
Unusual Attitude Recovery Protection 5-55
Common Errors Leading to Unusual Attitudes 5-58
Instrument Takeoff 5-60
Common Errors in Instrument Takeoffs 5-61
Basic Instrument Flight Patterns 5-61

Chapter 6

Helicopter Attitude Instrument Flying 6-1
Introduction 6-1
Flight Instruments 6-2
Instrument Flight 6-2
Instrument Cross-Check 6-2
Instrument Interpretation 6-3
Aircraft Control 6-3
Straight-and-Level Flight 6-3
Pitch Control 6-3
Attitude Indicator 6-3
Altimeter 6-4
Vertical Speed Indicator (VSI) 6-5
Airspeed Indicator 6-5
Bank Control 6-5
Attitude Indicator 6-5
Heading Indicator 6-6
Turn Indicator 6-7
Common Errors During Straight-and-Level Flight 6-7
Power Control During Straight-and-Level Flight 6-7
Common Errors During Airspeed Changes 6-10
Straight Climbs (Constant Airspeed and Constant Rate) 6-10
Entry 6-10
Level Off 6-12
Straight Descents (Constant Airspeed and Constant Rate) 6-12
Entry 6-12
Level Off 6-13
Common Errors During Straight Climbs and Descents 6-13
Turns 6-13
Turn to a Predetermined Heading 6-13
Timed Turns 6-13
Change of Airspeed in Turns 6-14
Compass Turns 6-15
30° Bank Turn 6-15
Climbing and Descending Turns 6-15
Common Errors During Turns 6-15
Unusual Attitudes 6-16
Common Errors During Unusual Attitude Recoveries 6-16
Emergencies 6-16
Autorotations 6-17
Common Errors During Autorotations 6-17
Servo Failure 6-17
Instrument Takeoff 6-17
Common Errors During Instrument Takeoffs 6-18
Changing Technology 6-18

Chapter 7

Navigation Systems 7-1
Introduction 7-1
Basic Radio Principles 7-2
How Radio Waves Propagate 7-2
Ground Wave 7-2
Sky Wave 7-2
Space Wave 7-2
Disturbances to Radio Wave Reception 7-3
Traditional Navigation Systems 7-3
Nondirectional Radio Beacon (NDB) 7-3
NDB Components 7-3
ADF Components 7-3
Function of ADF 7-4
Operational Errors of ADF 7-8
Very High Frequency Omnidirectional Range (VOR) 7-8
VOR Components 7-10
Function of VOR 7-12
VOR Operational Errors 7-14
VOR Accuracy 7-16
VOR Receiver Accuracy Check 7-16
VOR Test Facility (VOT) 7-16
Certified Checkpoints 7-16
Distance Measuring Equipment (DME) 7-16
DME Components 7-17
Function of DME 7-17
DME Arc 7-17
Intercepting Lead Radials 7-19
DME Errors 7-19
Area Navigation (RNAV) 7-19
VOR/DME RNAV 7-23
VOR/DME RNAV Components 7-23
Function of VOR/DME RNAV 7-23
VOR/DME RNAV Errors 7-24
Long Range Navigation (LORAN) 7-24
LORAN Components 7-25
Function of LORAN 7-26
LORAN Errors 7-26
Advanced Technologies 7-26
Global Navigation Satellite System (GNSS) 7-26

Global Positioning System (GPS) 7-27
GPS Components 7-27
Function of GPS 7-28
GPS Substitution 7-28
GPS Substitution for ADF or DME 7-29
To Determine Aircraft Position Over a DME Fix: 7-29
To Fly a DME Arc: 7-29
To Navigate TO or FROM an NDB/Compass Locator: 7-29
To Determine Aircraft Position Over an NDB/Compass Locator: 7-29
To Determine Aircraft Position Over a Fix Made up of an NDB/Compass Locator Bearing Crossing a VOR/LOC Course: 7-30
To Hold Over an NDB/Compass Locator: 7-30
IFR Flight Using GPS 7-30
GPS Instrument Approaches 7-31
Departures and Instrument Departure Procedures (DPs) 7-33
GPS Errors 7-33
System Status 7-33
GPS Familiarization 7-34
Differential Global Positioning Systems (DGPS) 7-34
Wide Area Augmentation System (WAAS) 7-34
General Requirements 7-34
Instrument Approach Capabilities 7-36
Local Area Augmentation System (LAAS) 7-36
Inertial Navigation System (INS) 7-36
INS Components 7-37
INS Errors 7-37
Instrument Approach Systems 7-37
Instrument Landing Systems (ILS) 7-37
ILS Components 7-39
Approach Lighting Systems (ALS) 7-40
ILS Airborne Components 7-42
ILS Function 7-42
ILS Errors 7-44
Marker Beacons 7-44
Operational Errors 7-45
Simplified Directional Facility (SDF) 7-45
Localizer Type Directional Aid (LDA) 7-45
Microwave Landing System (MLS) 7-45
Approach Azimuth Guidance 7-45
Required Navigation Performance 7-46
Flight Management Systems (FMS) 7-48
Function of FMS 7-48
Head-Up Display (HUD) 7-49
Radar Navigation (Ground Based) 7-49
Functions of Radar Navigation 7-49
Airport Surface Detection Equipment 7-50
Radar Limitations 7-50

Chapter 8
The National Airspace System 8-1
Introduction 8-1
Airspace Classification 8-2
Special Use Airspace 8-2
Federal Airways 8-4
Other Routing 8-5
IFR En Route Charts 8-6
Airport Information 8-6
Charted IFR Altitudes 8-6
Navigation Features 8-7
Types of NAVAIDs 8-7
Identifying Intersections 8-7
Other Route Information 8-10
Weather Information and Communication Features 8-10
New Technologies 8-10
Terminal Procedures Publications 8-12
Departure Procedures (DPs) 8-12
Standard Terminal Arrival Routes (STARs) 8-12
Instrument Approach Procedure (IAP) Charts 8-12
Margin Identification 8-12
The Pilot Briefing 8-16
The Plan View 8-16
Terminal Arrival Area (TAA) 8-18
Course Reversal Elements in Plan View and Profile View 8-20
Procedure Turns 8-20
Holding in Lieu of Procedure Turn 8-20
Teardrop Procedure 8-21
The Profile View 8-21
Landing Minimums 8-23
Airport Sketch /Airport Diagram 8-27
Inoperative Components 8-27
RNAV Instrument Approach Charts 8-32

Chapter 9
The Air Traffic Control System 9-1
Introduction 9-1
Communication Equipment 9-2
Navigation/Communication (NAV/COM) Equipment 9-2
Radar and Transponders 9-3
Mode C (Altitude Reporting) 9-3
Communication Procedures 9-4
Communication Facilities 9-4

  - Automated Flight Service Stations (AFSS) 9-4
  - ATC Towers 9-5
  - Terminal Radar Approach Control (TRACON) 9-6
  - Tower En Route Control (TEC) 9-7
  - Air Route Traffic Control Center (ARTCC) 9-7
  - Center Approach/Departure Control 9-7
- ATC Inflight Weather Avoidance Assistance 9-11
  - ATC Radar Weather Displays 9-11
  - Weather Avoidance Assistance 9-11
- Approach Control Facility 9-12
- Approach Control Advances 9-12
  - Precision Runway Monitor (PRM) 9-12
    - Precision Runway Monitor (PRM) Radar 9-12
    - PRM Benefits 9-13
- Control Sequence 9-13
  - Letters of Agreement (LOA) 9-14

Chapter 10

IFR Flight 10-1

- Introduction 10-1
- Sources of Flight Planning Information 10-2
  - Aeronautical Information Manual (AIM) 10-2
  - Airport/Facility Directory (A/FD) 10-2
  - Notices to Airmen Publication (NTAP) 10-2
  - POH/AFM 10-2
- IFR Flight Plan 10-2
  - Filing in Flight 10-2
  - Cancelling IFR Flight Plans 10-3
- Clearances 10-3
  - Examples 10-3
  - Clearance Separations 10-4
- Departure Procedures (DPs) 10-5
  - Obstacle Departure Procedures (ODP) 10-5
  - Standard Instrument Departures 10-5
  - Radar Controlled Departures 10-5
  - Departures From Airports Without an Operating Control Tower 10-7
- En Route Procedures 10-7
  - ATC Reports 10-7
  - Position Reports 10-7
  - Additional Reports 10-7
  - Planning the Descent and Approach 10-8
  - Standard Terminal Arrival Routes (STARs) 10-9
  - Substitutes for Inoperative or Unusable Components 10-9
- Holding Procedures 10-9
  - Standard Holding Pattern (No Wind) 10-9
  - Standard Holding Pattern (With Wind) 10-9
  - Holding Instructions 10-9
  - Standard Entry Procedures 10-11
  - Time Factors 10-12
  - DME Holding 10-12
- Approaches 10-12
  - Compliance With Published Standard Instrument Approach Procedures 10-12
  - Instrument Approaches to Civil Airports 10-13
    - Approach to Airport Without an Operating Control Tower 10-14
    - Approach to Airport With an Operating Tower, With No Approach Control 10-14
    - Approach to an Airport With an Operating Tower, With an Approach Control 10-14
  - Radar Approaches 10-17
  - Radar Monitoring of Instrument Approaches 10-18
  - Timed Approaches From a Holding Fix 10-18
  - Approaches to Parallel Runways 10-20
  - Side-Step Maneuver 10-20
  - Circling Approaches 10-20
  - IAP Minimums 10-21
  - Missed Approaches 10-21
  - Landing 10-22
- Instrument Weather Flying 10-22
  - Flying Experience 10-22
    - Recency of Experience 10-22
    - Airborne Equipment and Ground Facilities 10-22
  - Weather Conditions 10-22
    - Turbulence 10-23
    - Structural Icing 10-24
    - Fog 10-24
    - Volcanic Ash 10-24
    - Thunderstorms 10-25
    - Wind Shear 10-25
  - VFR-On-Top 10-26
  - VFR Over-The-Top 10-27
- Conducting an IFR Flight 10-27
  - Preflight 10-27
  - Departure 10-31
  - En Route 10-32
  - Arrival 10-33

Chapter 11

Emergency Operations 11-1

- Introduction 11-1
- Unforecast Adverse Weather 11-2
  - Inadvertent Thunderstorm Encounter 11-2
  - Inadvertent Icing Encounter 11-2
  - Precipitation Static 11-3
- Aircraft System Malfunctions 11-3
  - Electronic Flight Display Malfunction 11-4
  - Alternator/Generator Failure 11-4
  - Techniques for Electrical Usage 11-5

Master Battery Switch 11-5
Operating on the Main Battery 11-5
Loss of Alternator/Generator for Electronic Flight Instrumentation..11-5
Techniques for Electrical Usage 11-6
Standby Battery 11-6
Operating on the Main Battery 11-6
Analog Instrument Failure 11-6
Pneumatic System Failure..11-7
Pitot/Static System Failure...11-7
Communication/Navigation System Malfunction 11-8
GPS Nearest Airport Function.....................................11-9
Nearest Airports Using the PFD...............................11-9
Additional Information for a Specific Airport......11-9
Nearest Airports Using the MFD 11-10
Navigating the MFD Page Groups......................11-10
Nearest Airport Page Group 11-10
Nearest Airports Page Soft Keys 11-10
Situational Awareness..11-11
Traffic Avoidance 11-12

Appendix A
Clearance Shorthand A-1

Appendix B
Instrument Training Lesson Guide.................. B-1

Glossary G-1

Index I-1

Semicircular canals
Tubular ducts containing endolymph
Utricle
Saccule
Cochlea
Ampullae
Pilot capabilitie
Margin of Safety
Effort
Task requirements
Time
Preflight
Takeoff
Cruise
Approach & Landing

Chapter 1

Human Factors

Introduction

Human factors is a broad field that examines the interaction between people, machines, and the environment for the purpose of improving performance and reducing errors. As aircraft became more reliable and less prone to mechanical failure, the percentage of accidents related to human factors increased. Some aspect of human factors now accounts for over 80 percent of all accidents. Pilots who have a good understanding of human factors are better equipped to plan and execute a safe and uneventful flight.

Flying in instrument meteorological conditions (IMC) can result in sensations that are misleading to the body's sensory system. A safe pilot needs to understand these sensations and effectively counteract them. Instrument flying requires a pilot to make decisions using all available resources.

The elements of human factors covered in this chapter include sensory systems used for orientation, illusions in flight, physiological and psychological factors, medical factors, aeronautical decision-making, and crew resource management (CRM).

Sensory Systems for Orientation

Orientation is the awareness of the position of the aircraft and of oneself in relation to a specific reference point. Disorientation is the lack of orientation, and spatial disorientation specifically refers to the lack of orientation with regard to position in space and to other objects.

Orientation is maintained through the body's sensory organs in three areas: visual, vestibular, and postural. The eyes maintain visual orientation. The motion sensing system in the inner ear maintains vestibular orientation. The nerves in the skin, joints, and muscles of the body maintain postural orientation. When healthy human beings are in their natural environment, these three systems work well. When the human body is subjected to the forces of flight, these senses can provide misleading information. It is this misleading information that causes pilots to become disoriented.

Eyes

Of all the senses, vision is most important in providing information to maintain safe flight. Even though the human eye is optimized for day vision, it is also capable of vision in very low light environments. During the day, the eye uses receptors called cones, while at night, vision is facilitated by the use of rods. Both of these provide a level of vision optimized for the lighting conditions that they were intended. That is, cones are ineffective at night and rods are ineffective during the day.

Rods, which contain rhodopsin (called visual purple), are especially sensitive to light and increased light washes out the rhodopsin compromising the night vision. Hence, when strong light is momentarily introduced at night, vision may be totally ineffective as the rods take time to become effective again in darkness. Smoking, alcohol, oxygen deprivation, and age affect vision, especially at night. It should be noted that at night, oxygen deprivation such as one caused from a climb to a high altitude causes a significant reduction in vision. A return back to the lower altitude will not restore a pilot's vision in the same transitory period used at the climb altitude.

The eye also has two blind spots. The day blind spot is the location on the light sensitive retina where the optic nerve fiber bundle (which carries messages from the eye to the brain) passes through. This location has no light receptors, and a message cannot be created there to be sent to the brain. The night blind spot is due to a concentration of cones in an area surrounding the fovea on the retina. Because there are no rods in this area, direct vision on an object at night will disappear. As a result, off-center viewing and scanning at night is best for both obstacle avoidance and to maximize situational awareness. [See the Pilot's Handbook of Aeronautical Knowledge and the Aeronautical Information Manual (AIM) for detailed reading.]

The brain also processes visual information based upon color, relationship of colors, and vision from objects around us. *Figure 1-1* demonstrates the visual processing of information. The brain assigns color based on many items to include an object's surroundings. In the figure below, the orange square on the shaded side of the cube is actually the same color as the brown square in the center of the cube's top face.

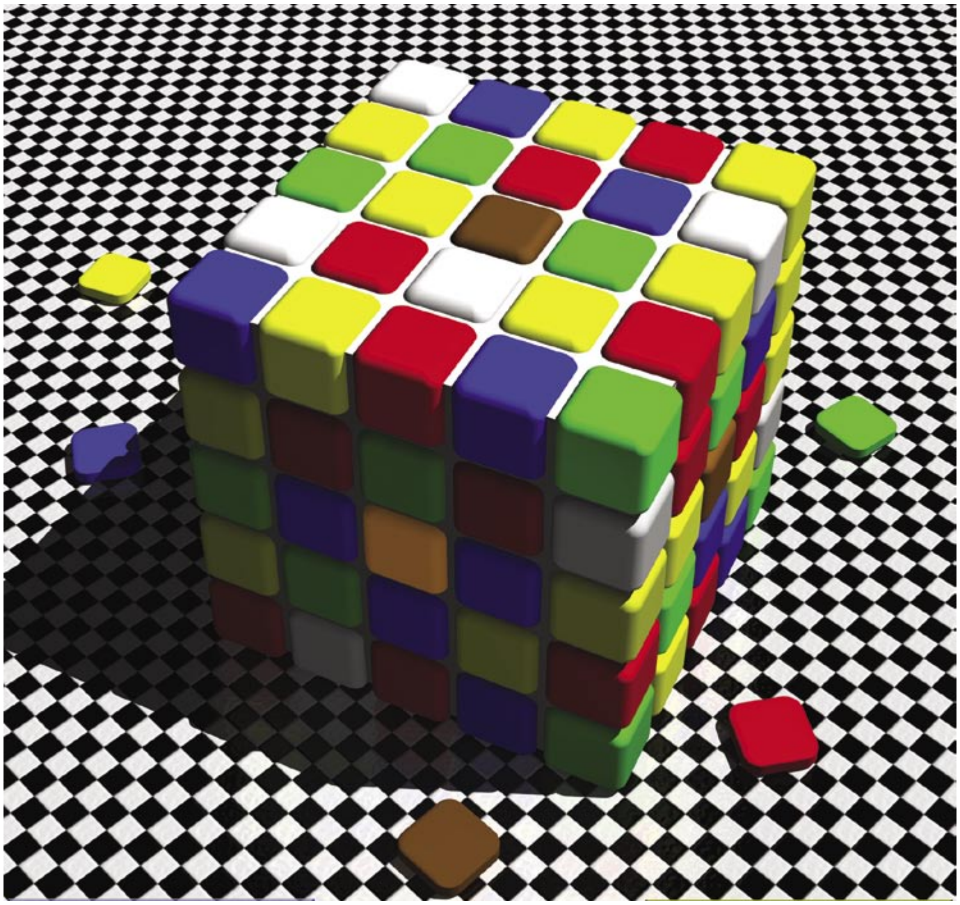

Figure 1-1. *Rubik's Cube Graphic.*

Isolating the orange square from surrounding influences will reveal that it is actually brown. The application to a real environment is evident when processing visual information that is influenced by surroundings. The ability to pick out an airport in varied terrain or another aircraft in a light haze are examples of problems with interpretation that make vigilance all the more necessary.

Figure 1-2 illustrates problems with perception. Both tables are the same lengths. Objects are easily misinterpreted in size to include both length and width. Being accustomed to a 75-foot-wide runway on flat terrain is most likely going to influence a pilot's perception of a wider runway on uneven terrain simply because of the inherent processing experience.

Vision Under Dim and Bright Illumination

Under conditions of dim illumination, aeronautical charts and aircraft instruments can become unreadable unless adequate flight deck lighting is available. In darkness, vision becomes more sensitive to light. This process is called dark adaptation. Although exposure to total darkness for at least 30 minutes is required for complete dark adaptation, a pilot can achieve a moderate degree of dark adaptation within 20 minutes under dim red flight deck lighting.

Red light distorts colors (filters the red spectrum), especially on aeronautical charts, and makes it very difficult for the eyes to focus on objects inside the aircraft. Pilots should use it only where optimum outside night vision capability is necessary. White flight deck lighting (dim lighting) should be available when needed for map and instrument reading, especially under IMC conditions.

Since any degree of dark adaptation is lost within a few seconds of viewing a bright light, pilots should close one eye when using a light to preserve some degree of night vision. During night flights in the vicinity of lightning, flight deck lights should be turned up to help prevent loss of night vision due to the bright flashes. Dark adaptation is also impaired by exposure to cabin pressure altitudes above 5,000 feet, carbon monoxide inhaled through smoking, deficiency of Vitamin A in the diet, and by prolonged exposure to bright sunlight.

During flight in visual meteorological conditions (VMC), the eyes are the major orientation source and usually provide accurate and reliable information. Visual cues usually prevail over false sensations from other sensory systems. When these visual cues are taken away, as they are in IMC, false sensations can cause the pilot to quickly become disoriented.

An effective way to counter these false sensations is to recognize the problem, disregard the false sensations, rely on the flight instruments, and use the eyes to determine the aircraft attitude. The pilot must have an understanding of the problem and the skill to control the aircraft using only instrument indications.

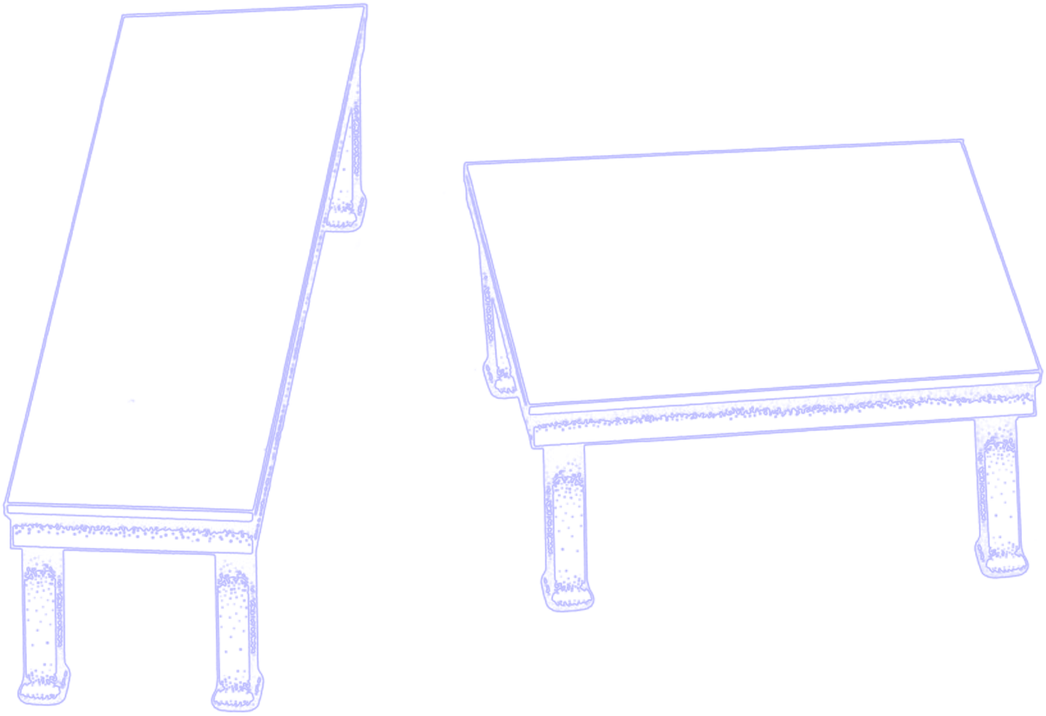

Figure 1-2. *Shepard's Tables.*

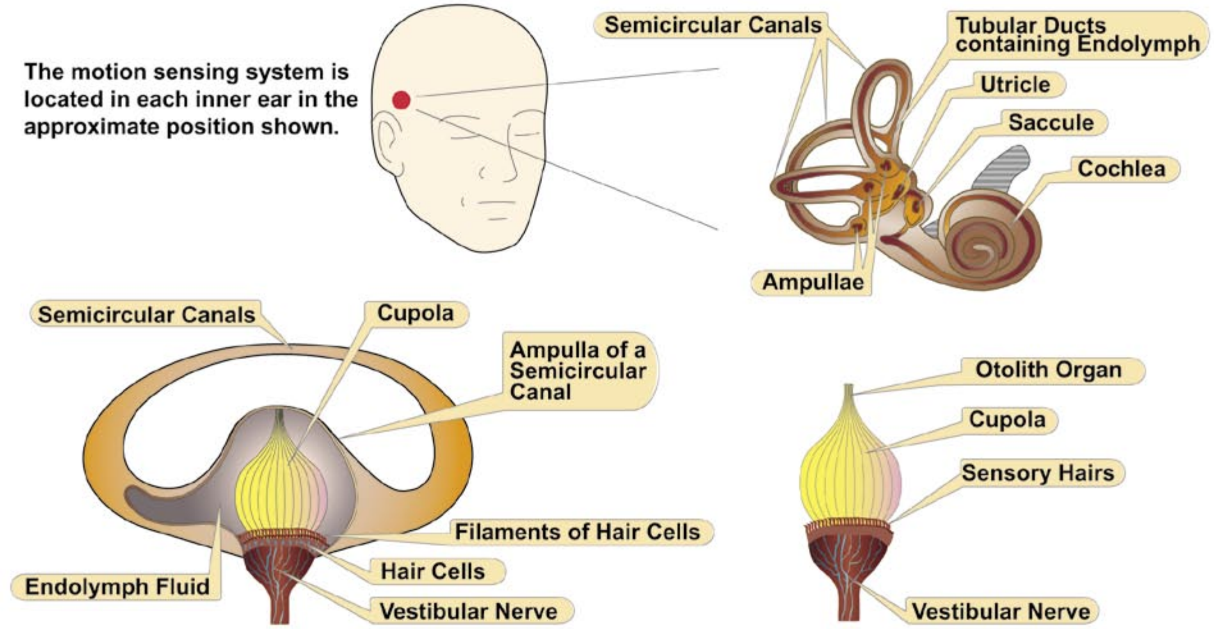
The motion sensing system is located in each inner ear in the approximate position shown.
Semicircular Canals
Tubular Ducts containing Endolymph
Utricle
Saccule
Cochlea
Ampullae
Semicircular Canals
Cupola
Ampulla of a Semicircular Canal
Filaments of Hair Cells
Hair Cells
Endolymph Fluid
Vestibular Nerve
Otolith Organ
Cupola
Sensory Hairs
Vestibular Nerve

Figure 1-3. *Inner Ear Orientation.*

Ears

The inner ear has two major parts concerned with orientation, the semicircular canals and the otolith organs. *[Figure 1-3]* The semicircular canals detect angular acceleration of the body while the otolith organs detect linear acceleration and gravity. The semicircular canals consist of three tubes at right angles to each other, each located on one of three axes: pitch, roll, or yaw as illustrated in *Figure 1-4*. Each canal is filled with a fluid called endolymph fluid. In the center of the canal is the cupola, a gelatinous structure that rests upon sensory hairs located at the end of the vestibular nerves. It is the movement of these hairs within the fluid which causes sensations of motion.

Because of the friction between the fluid and the canal, it may take about 15–20 seconds for the fluid in the ear canal to reach the same speed as the canal's motion.

To illustrate what happens during a turn, visualize the aircraft in straight and level flight. With no acceleration of the aircraft, the hair cells are upright and the body senses that no turn has occurred. Therefore, the position of the hair cells and the actual sensation correspond.

Placing the aircraft into a turn puts the semicircular canal and its fluid into motion, with the fluid within the semicircular canal lagging behind the accelerated canal walls.*[Figure 1-5]* This lag creates a relative movement of the fluid within the canal. The canal wall and the cupula move in the opposite direction from the motion of the fluid.

The brain interprets the movement of the hairs to be a turn in the same direction as the canal wall. The body correctly senses that a turn is being made. If the turn continues at a constant rate for several seconds or longer, the motion of the fluid in

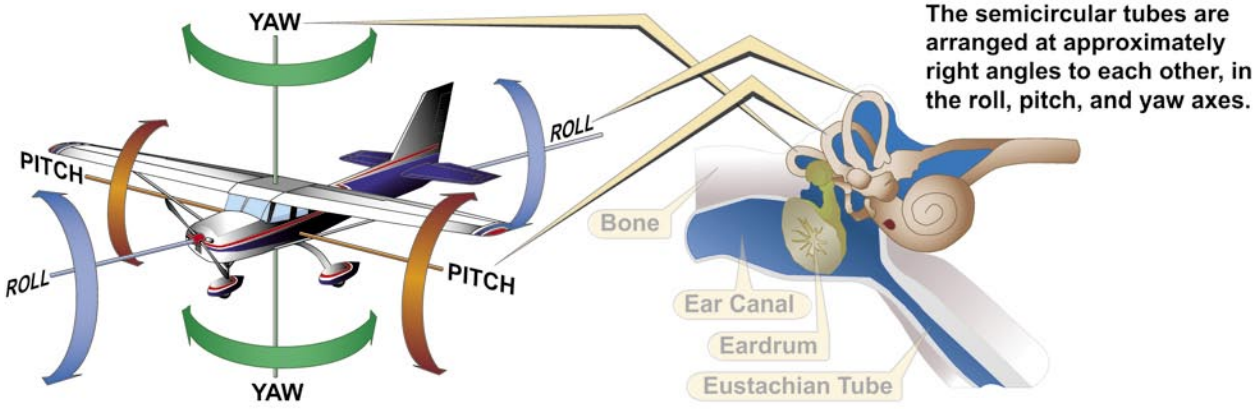
YAW
ROLL
PITCH
ROLL
PITCH
YAW
The semicircular tubes are arranged at approximately right angles to each other, in the roll, pitch, and yaw axes.
Bone
Ear Canal
Eardrum
Eustachian Tube

Figure 1-4. *Angular Acceleration and the Semicircular Tubes.*

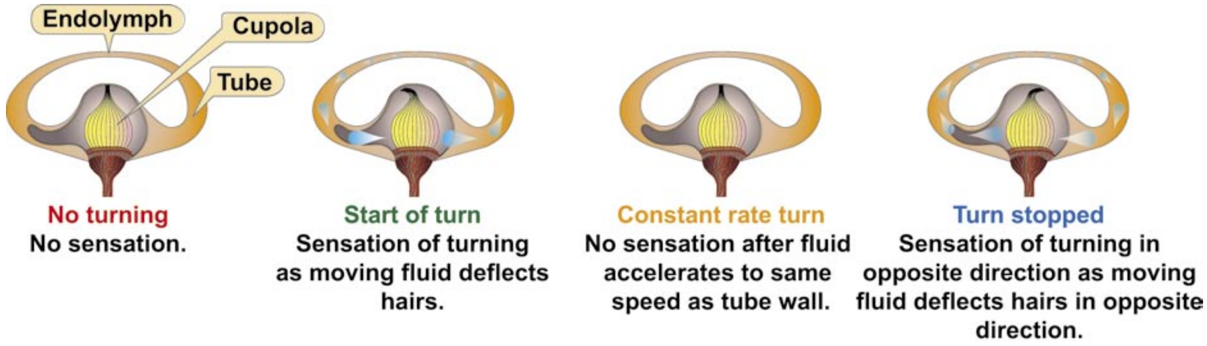
Endolymph
Cupola
Tube
No turning
No sensation.
Start of turn
Sensation of turning as moving fluid deflects hairs.
Constant rate turn
No sensation after fluid accelerates to same speed as tube wall.
Turn stopped
Sensation of turning in opposite direction as moving fluid deflects hairs in opposite direction.

Figure 1-5. *Angular Acceleration.*

the canals catches up with the canal walls. The hairs are no longer bent, and the brain receives the false impression that turning has stopped. Thus, the position of the hair cells and the resulting sensation during a prolonged, constant turn in either direction will result in the false sensation of no turn.

When the aircraft returns to straight-and-level flight, the fluid in the canal moves briefly in the opposite direction. This sends a signal to the brain that is falsely interpreted as movement in the opposite direction. In an attempt to correct the falsely perceived turn, the pilot may reenter the turn placing the aircraft in an out of control situation.

The otolith organs detect linear acceleration and gravity in a similar way. Instead of being filled with a fluid, a gelatinous membrane containing chalk-like crystals covers the sensory hairs. When the pilot tilts his or her head, the weight of these crystals causes this membrane to shift due to gravity and the sensory hairs detect this shift. The brain orients this new position to what it perceives as vertical. Acceleration and deceleration also cause the membrane to shift in a similar manner. Forward acceleration gives the illusion of the head tilting backward. *[Figure 1-6]* As a result, during takeoff and while accelerating, the pilot may sense a steeper than normal climb resulting in a tendency to nose-down.

Nerves

Nerves in the body's skin, muscles, and joints constantly send signals to the brain, which signals the body's relation to gravity. These signals tell the pilot his or her current position. Acceleration will be felt as the pilot is pushed back into the seat. Forces created in turns can lead to false sensations of the true direction of gravity, and may give the pilot a false sense of which way is up.

Uncoordinated turns, especially climbing turns, can cause misleading signals to be sent to the brain. Skids and slips give the sensation of banking or tilting. Turbulence can create motions that confuse the brain as well. Pilots need to be aware that fatigue or illness can exacerbate these sensations and ultimately lead to subtle incapacitation.

Illusions Leading to Spatial Disorientation

The sensory system responsible for most of the illusions leading to spatial disorientation is the vestibular system. Visual illusions can also cause spatial disorientation.

Vestibular Illusions

The Leans

A condition called the leans can result when a banked attitude, to the left for example, may be entered too slowly to set in motion the fluid in the "roll" semicircular tubes. *[Figure 1-5]* An abrupt correction of this attitude sets the fluid in motion, creating the illusion of a banked attitude to the right. The disoriented pilot may make the error of rolling the aircraft into the original left banked attitude, or if level flight is maintained, will feel compelled to lean in the perceived vertical plane until this illusion subsides.

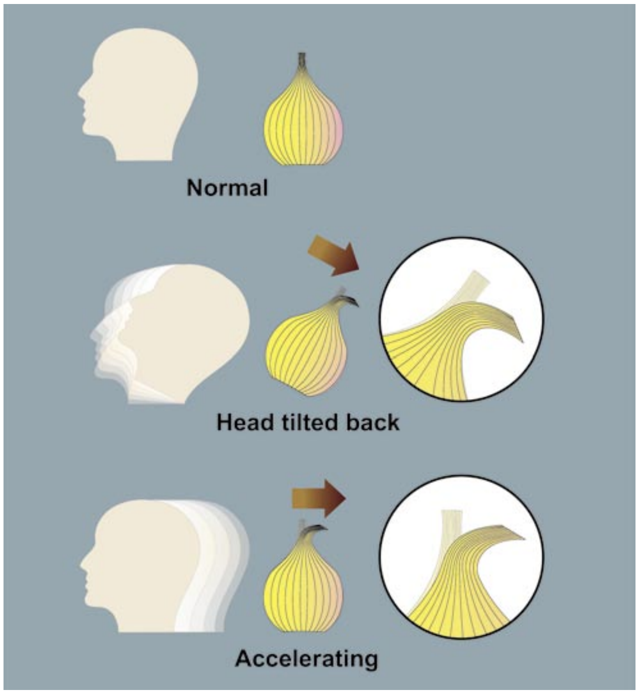
Normal
Head tilted back
Accelerating

Figure 1-6. *Linear Acceleration.*

Coriolis Illusion

The coriolis illusion occurs when a pilot has been in a turn long enough for the fluid in the ear canal to move at the same speed as the canal. A movement of the head in a different plane, such as looking at something in a different part of the flight deck, may set the fluid moving and create the illusion of turning or accelerating on an entirely different axis. This action causes the pilot to think the aircraft is doing a maneuver that it is not. The disoriented pilot may maneuver the aircraft into a dangerous attitude in an attempt to correct the aircraft's perceived attitude.

For this reason, it is important that pilots develop an instrument cross-check or scan that involves minimal head movement. Take care when retrieving charts and other objects in the flight deck—if something is dropped, retrieve it with minimal head movement and be alert for the coriolis illusion.

Graveyard Spiral

As in other illusions, a pilot in a prolonged coordinated, constant rate turn, will have the illusion of not turning. During the recovery to level flight, the pilot will experience the sensation of turning in the opposite direction. The disoriented pilot may return the aircraft to its original turn. Because an aircraft tends to lose altitude in turns unless the pilot compensates for the loss in lift, the pilot may notice a loss of altitude. The absence of any sensation of turning creates the illusion of being in a level descent. The pilot may pull back on the controls in an attempt to climb or stop the descent. This action tightens the spiral and increases the loss of altitude; hence, this illusion is referred to as a graveyard spiral. *[Figure 1-7]* At some point, this could lead to a loss of control by the pilot.

Somatogravic Illusion

A rapid acceleration, such as experienced during takeoff, stimulates the otolith organs in the same way as tilting the head backwards. This action creates the somatogravic illusion of being in a nose-up attitude, especially in situations without good visual references. The disoriented pilot may push the aircraft into a nose-low or dive attitude. A rapid deceleration by quick reduction of the throttle(s) can have the opposite effect, with the disoriented pilot pulling the aircraft into a nose-up or stall attitude.

Inversion Illusion

An abrupt change from climb to straight-and-level flight can stimulate the otolith organs enough to create the illusion of tumbling backwards, or inversion illusion. The disoriented pilot may push the aircraft abruptly into a nose-low attitude, possibly intensifying this illusion.

Elevator Illusion

An abrupt upward vertical acceleration, as can occur in an updraft, can stimulate the otolith organs to create the illusion of being in a climb. This is called elevator illusion. The disoriented pilot may push the aircraft into a nose-low attitude. An abrupt downward vertical acceleration, usually

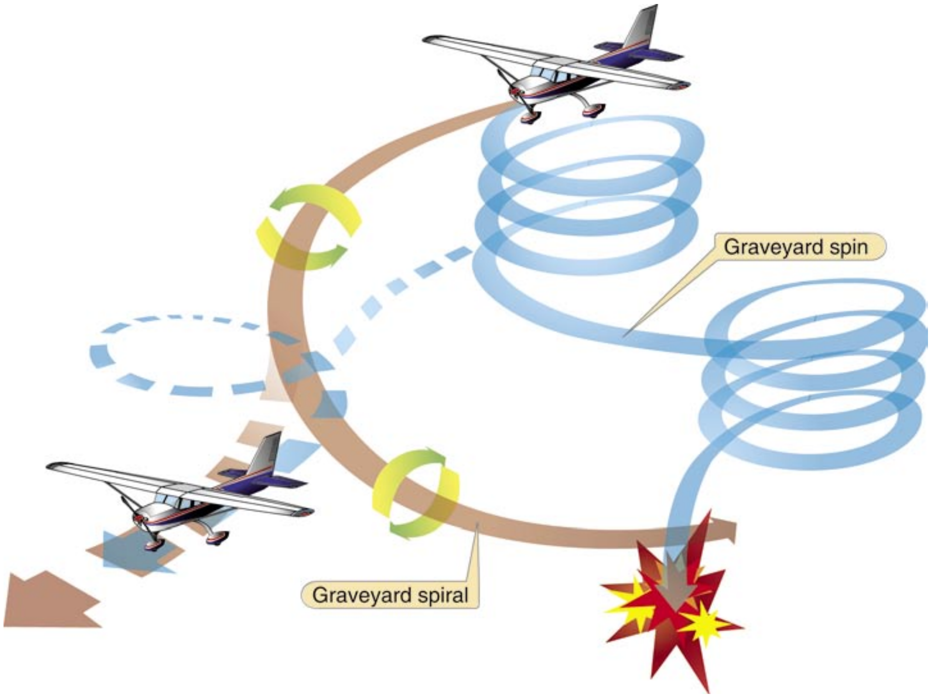

Figure 1-7. *Graveyard Spiral.*

in a downdraft, has the opposite effect, with the disoriented pilot pulling the aircraft into a nose-up attitude.

Visual Illusions

Visual illusions are especially hazardous because pilots rely on their eyes for correct information. Two illusions that lead to spatial disorientation, false horizon and autokinesis, are concerned with only the visual system.

False Horizon

A sloping cloud formation, an obscured horizon, an aurora borealis, a dark scene spread with ground lights and stars, and certain geometric patterns of ground lights can provide inaccurate visual information, or false horizon, for aligning the aircraft correctly with the actual horizon. The disoriented pilot may place the aircraft in a dangerous attitude.

Autokinesis

In the dark, a stationary light will appear to move about when stared at for many seconds. The disoriented pilot could lose control of the aircraft in attempting to align it with the false movements of this light, called autokinesis.

Postural Considerations

The postural system sends signals from the skin, joints, and muscles to the brain that are interpreted in relation to the Earth's gravitational pull. These signals determine posture. Inputs from each movement update the body's position to the brain on a constant basis. "Seat of the pants" flying is largely dependent upon these signals. Used in conjunction with visual and vestibular clues, these sensations can be fairly reliable. However, because of the forces acting upon the body in certain flight situations, many false sensations can occur due to acceleration forces overpowering gravity. *[Figure 1-8]* These situations include uncoordinated turns, climbing turns, and turbulence.

Demonstration of Spatial Disorientation

There are a number of controlled aircraft maneuvers a pilot can perform to experiment with spatial disorientation. While each maneuver will normally create a specific illusion, any false sensation is an effective demonstration of disorientation. Thus, even if there is no sensation during any of these maneuvers, the absence of sensation is still an effective demonstration in that it shows the inability to detect bank or roll. There are several objectives in demonstrating these various maneuvers.

1. They teach pilots to understand the susceptibility of the human system to spatial disorientation.
2. They demonstrate that judgments of aircraft attitude based on bodily sensations are frequently false.
3. They help lessen the occurrence and degree of disorientation through a better understanding of the relationship between aircraft motion, head movements, and resulting disorientation.
4. They help instill a greater confidence in relying on flight instruments for assessing true aircraft attitude.

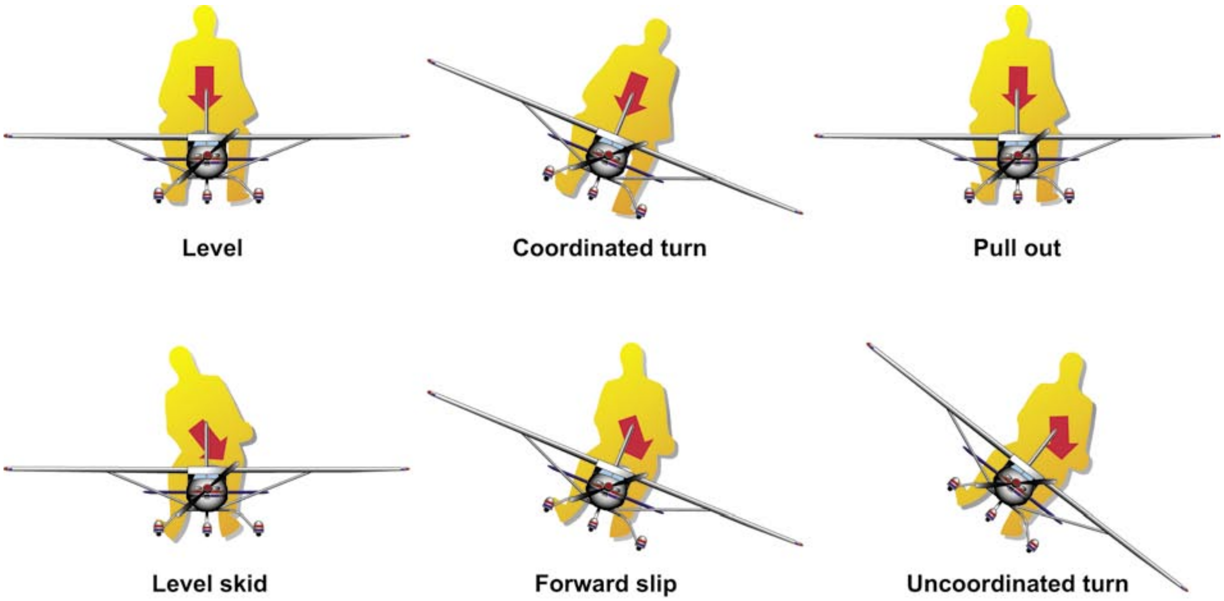

Level **Coordinated turn** **Pull out**

Level skid **Forward slip** **Uncoordinated turn**

Skid, slip, and uncoordinated turns feel alike.
Pilots feel they are being forced sideways in their seat.

Figure 1-8. *Sensations From Centrifugal Force.*

A pilot should not attempt any of these maneuvers at low altitudes, or in the absence of an instructor pilot or an appropriate safety pilot.

Climbing While Accelerating

With the pilot's eyes closed, the instructor pilot maintains approach airspeed in a straight-and-level attitude for several seconds, and then accelerates while maintaining straight-and-level attitude. The usual illusion during this maneuver, without visual references, will be that the aircraft is climbing.

Climbing While Turning

With the pilot's eyes still closed and the aircraft in a straight-and-level attitude, the instructor pilot now executes, with a relatively slow entry, a well-coordinated turn of about 1.5 positive G (approximately 50° bank) for 90°. While in the turn, without outside visual references and under the effect of the slight positive G, the usual illusion produced is that of a climb. Upon sensing the climb, the pilot should immediately open the eyes and see that a slowly established, coordinated turn produces the same feeling as a climb.

Diving While Turning

Repeating the previous procedure, with the exception that the pilot's eyes should be kept closed until recovery from the turn is approximately one-half completed can create this sensation. With the eyes closed, the usual illusion will be that the aircraft is diving.

Tilting to Right or Left

While in a straight-and-level attitude, with the pilot's eyes closed, the instructor pilot executes a moderate or slight skid to the left with wings level. This creates the illusion of the body being tilted to the right.

Reversal of Motion

This illusion can be demonstrated in any of the three planes of motion. While straight and level, with the pilot's eyes closed, the instructor pilot smoothly and positively rolls the aircraft to approximately a 45° bank attitude while maintaining heading and pitch attitude. This creates the illusion of a strong sense of rotation in the opposite direction. After this illusion is noted, the pilot should open his or her eyes and observe that the aircraft is in a banked attitude.

Diving or Rolling Beyond the Vertical Plane

This maneuver may produce extreme disorientation. While in straight-and-level flight, the pilot should sit normally, either with eyes closed or gaze lowered to the floor. The instructor pilot starts a positive, coordinated roll toward a 30° or 40° angle of bank. As this is in progress, the pilot tilts his or her head forward, looks to the right or left, then immediately returns his or her head to an upright position. The instructor pilot should time the maneuver so the roll is stopped as the pilot returns his or her head upright. An intense disorientation is usually produced by this maneuver, and the pilot experiences the sensation of falling downward into the direction of the roll.

In the descriptions of these maneuvers, the instructor pilot is doing the flying, but having the pilot do the flying can also be a very effective demonstration. The pilot should close his or her eyes and tilt the head to one side. The instructor pilot tells the pilot what control inputs to perform. The pilot then attempts to establish the correct attitude or control input with eyes closed and head tilted. While it is clear the pilot has no idea of the actual attitude, he or she will react to what the senses are saying. After a short time, the pilot will become disoriented and the instructor pilot then tells the pilot to look up and recover. The benefit of this exercise is the pilot experiences the disorientation while flying the aircraft.

Coping with Spatial Disorientation

To prevent illusions and their potentially disastrous consequences, pilots can:

1. Understand the causes of these illusions and remain constantly alert for them. Take the opportunity to understand and then experience spatial disorientation illusions in a device such as a Barany chair, a Vertigon, or a Virtual Reality Spatial Disorientation Demonstrator.
2. Always obtain and understand preflight weather briefings.
3. Before flying in marginal visibility (less than 3 miles) or where a visible horizon is not evident such as flight over open water during the night, obtain training and maintain proficiency in airplane control by reference to instruments.
4. Do not continue flight into adverse weather conditions or into dusk or darkness unless proficient in the use of flight instruments. If intending to fly at night, maintain night-flight currency and proficiency. Include cross-country and local operations at various airfields.
5. Ensure that when outside visual references are used, they are reliable, fixed points on the Earth's surface.
6. Avoid sudden head movement, particularly during takeoffs, turns, and approaches to landing.
7. Be physically tuned for flight into reduced visibility. That is, ensure proper rest, adequate diet, and, if flying at night, allow for night adaptation. Remember that illness, medication, alcohol, fatigue, sleep loss, and mild hypoxia are likely to increase susceptibility to spatial disorientation.

8. Most importantly, become proficient in the use of flight instruments and rely upon them. Trust the instruments and disregard your sensory perceptions.

The sensations that lead to illusions during instrument flight conditions are normal perceptions experienced by pilots. These undesirable sensations cannot be completely prevented, but through training and awareness, pilots can ignore or suppress them by developing absolute reliance on the flight instruments. As pilots gain proficiency in instrument flying, they become less susceptible to these illusions and their effects.

Optical Illusions

Of the senses, vision is the most important for safe flight. However, various terrain features and atmospheric conditions can create optical illusions. These illusions are primarily associated with landing. Since pilots must transition from reliance on instruments to visual cues outside the flight deck for landing at the end of an instrument approach, it is imperative they be aware of the potential problems associated with these illusions, and take appropriate corrective action. The major illusions leading to landing errors are described below.

Runway Width Illusion

A narrower-than-usual runway can create an illusion the aircraft is at a higher altitude than it actually is, especially when runway length-to-width relationships are comparable. *[Figure 1-9A]* The pilot who does not recognize this illusion will fly a lower approach, with the risk of striking objects along the approach path or landing short. A wider-than-usual runway can have the opposite effect, with the risk of leveling out high and landing hard, or overshooting the runway.

Runway and Terrain Slopes Illusion

An upsloping runway, upsloping terrain, or both, can create an illusion the aircraft is at a higher altitude than it actually is. *[Figure 1-9B]* The pilot who does not recognize this illusion will fly a lower approach. Downsloping runways and downsloping approach terrain can have the opposite effect.

Featureless Terrain Illusion

An absence of surrounding ground features, as in an overwater approach, over darkened areas, or terrain made featureless by snow, can create an illusion the aircraft is at a higher altitude than it actually is. This illusion, sometimes referred to as the "black hole approach," causes pilots to fly a lower approach than is desired.

Water Refraction

Rain on the windscreen can create an illusion of being at a higher altitude due to the horizon appearing lower than it is. This can result in the pilot flying a lower approach.

Haze

Atmospheric haze can create an illusion of being at a greater distance and height from the runway. As a result, the pilot will have a tendency to be low on the approach. Conversely, extremely clear air (clear bright conditions of a high attitude airport) can give the pilot the illusion of being closer than he or she actually is, resulting in a high approach, which may result in an overshoot or go around. The diffusion of light due to water particles on the windshield can adversely affect depth perception. The lights and terrain features normally used to gauge height during landing become less effective for the pilot.

Fog

Flying into fog can create an illusion of pitching up. Pilots who do not recognize this illusion will often steepen the approach quite abruptly.

Ground Lighting Illusions

Lights along a straight path, such as a road or lights on moving trains, can be mistaken for runway and approach lights. Bright runway and approach lighting systems, especially where few lights illuminate the surrounding terrain, may create the illusion of less distance to the runway. The pilot who does not recognize this illusion will often fly a higher approach.

How To Prevent Landing Errors Due to Optical Illusions

To prevent these illusions and their potentially hazardous consequences, pilots can:

1. Anticipate the possibility of visual illusions during approaches to unfamiliar airports, particularly at night or in adverse weather conditions. Consult airport diagrams and the Airport/Facility Directory (A/FD) for information on runway slope, terrain, and lighting.
2. Make frequent reference to the altimeter, especially during all approaches, day and night.
3. If possible, conduct aerial visual inspection of unfamiliar airports before landing.

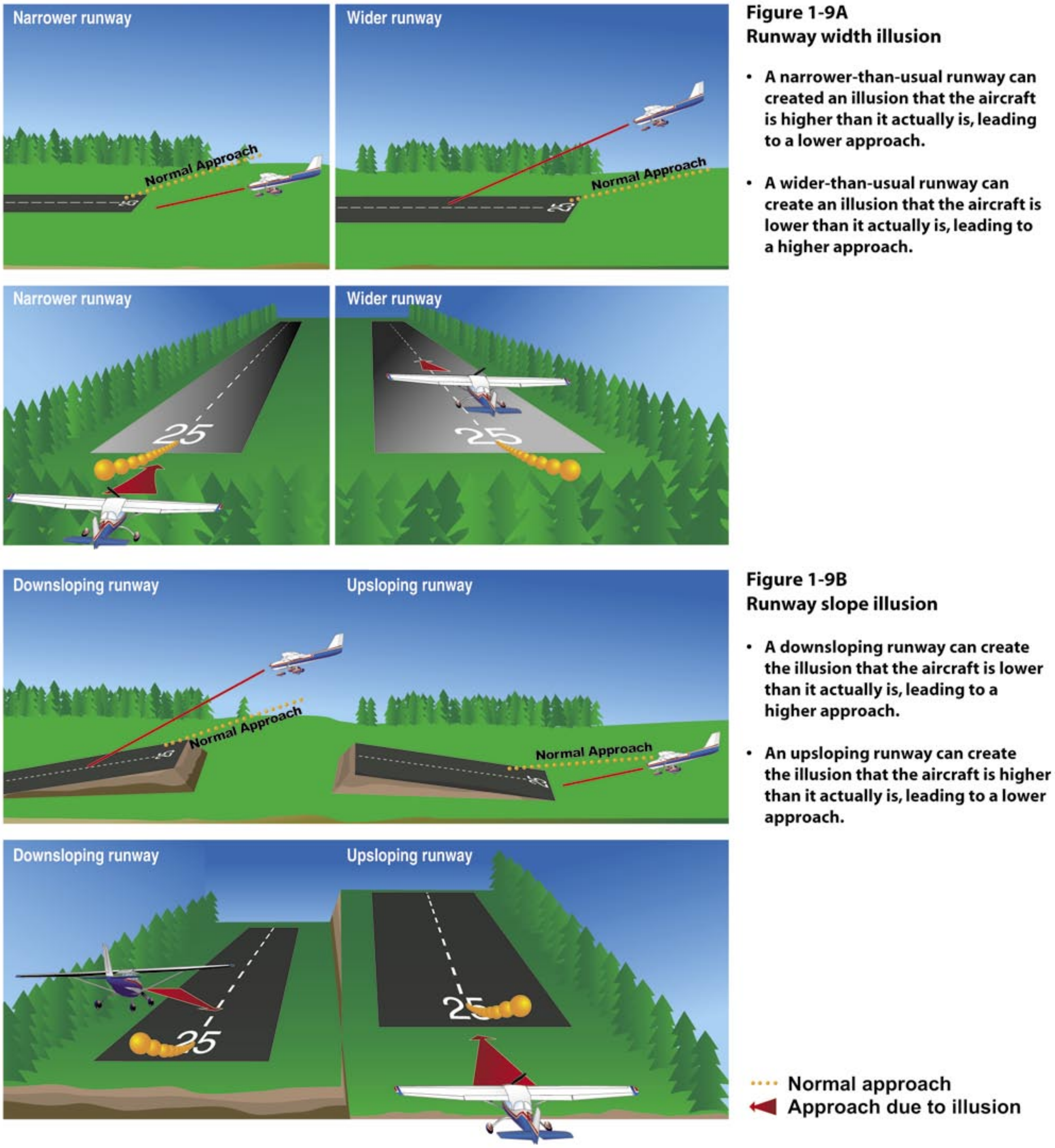
Narrower runway
Wider runway
Normal Approach
Normal Approach
Narrower runway
Wider runway
25
25
Downsloping runway
Upsloping runway
Normal Approach
Normal Approach
Downsloping runway
Upsloping runway
25
25
···· Normal approach
◀ Approach due to illusion

Figure 1-9A
Runway width illusion

- **A narrower-than-usual runway can created an illusion that the aircraft is higher than it actually is, leading to a lower approach.**
- **A wider-than-usual runway can create an illusion that the aircraft is lower than it actually is, leading to a higher approach.**

Figure 1-9B
Runway slope illusion

- **A downsloping runway can create the illusion that the aircraft is lower than it actually is, leading to a higher approach.**
- **An upsloping runway can create the illusion that the aircraft is higher than it actually is, leading to a lower approach.**

Figure 1-9. *Runway Width and Slope Illusions.*

4. Use Visual Approach Slope Indicator (VASI) or Precision Approach Path Indicator (PAPI) systems for a visual reference, or an electronic glide slope, whenever they are available.
5. Utilize the visual descent point (VDP) found on many nonprecision instrument approach procedure charts.
6. Recognize that the chances of being involved in an approach accident increase when some emergency or other activity distracts from usual procedures.
7. Maintain optimum proficiency in landing procedures.

Physiological and Psychological Factors

Physiological or psychological factors can affect a pilot and compromise the safety of a flight. These factors are stress, medical, alcohol, and fatigue. Any of these factors, individually or in combination, significantly degrade the pilot's decision-making or flying abilities.

Stress

Stress is the body's response to demands placed upon it. These demands can be either pleasant or unpleasant in nature. The causes of stress for a pilot can range from unexpected weather or mechanical problems while in flight, to personal issues unrelated to flying. Stress is an inevitable and necessary part of life; it adds motivation to life and heightens an individual's response to meet any challenge. The effects of stress are cumulative and there is a limit to a person's adaptive nature. This limit, called the stress tolerance level (or channel capacity), is based on the ability to cope with the situation.

At first, some amount of stress can be desirable and can actually improve performance. However, higher stress levels, particularly over long periods of time, can adversely affect performance. Performance will generally increase with the onset of stress, but will peak and then begin to fall off rapidly as stress levels exceed the ability to cope. *[Figure 1-10]*

At this point, a pilot's performance begins to decline and judgment deteriorates. Complex or unfamiliar tasks require higher levels of performance than simple or overlearned tasks. Complex or unfamiliar tasks are also more subject to the adverse effects of increasing stress than simple or familiar tasks. *[Figure 1-10]*

The indicators of excessive stress often show as three types of symptoms: (1) emotional, (2) physical, and (3) behavioral. Emotional symptoms may surface as over-compensation, denial, suspicion, paranoia, agitation, restlessness, or defensiveness. Physical stress can result in acute fatigue while behavioral degradation will be manifested as sensitivity to criticism, tendency to be argumentative, arrogance, and hostility. Pilots need to learn to recognize the symptoms of stress as they begin to occur.

There are many techniques available that can help reduce stress in life or help people cope with it better. Not all of the following ideas may be a solution, but some of them should be effective.

1. Become knowledgeable about stress.
2. Take a realistic self-assessment. (See the Pilot's Handbook of Aeronautical Knowledge).
3. Take a systematic approach to problem solving.
4. Develop a lifestyle that will buffer against the effects of stress.
5. Practice behavior management techniques.
6. Establish and maintain a strong support network.

Good flight deck stress management begins with good life stress management. Many of the stress coping techniques practiced for life stress management are not usually practical in flight. Rather, pilots must condition themselves to relax and think rationally when stress appears. The following checklist outlines some methods of flight deck stress management.

1. Avoid situations that distract from flying the aircraft.
2. Reduce flight deck workload to reduce stress levels. This will create a proper environment in which to make good decisions. Typically, flying involves higher stress levels during takeoff and landing phases. Between the two generally lies a period of low activity resulting in a lower stress level. Transitioning from the cruise phase to the landing phase is generally accompanied by a significant workload that, if not properly accommodated, will increase stress significantly. Proper planning and prioritization of flight deck duties are key to avoiding events that affect the pilot's capacity to maintain situational awareness.

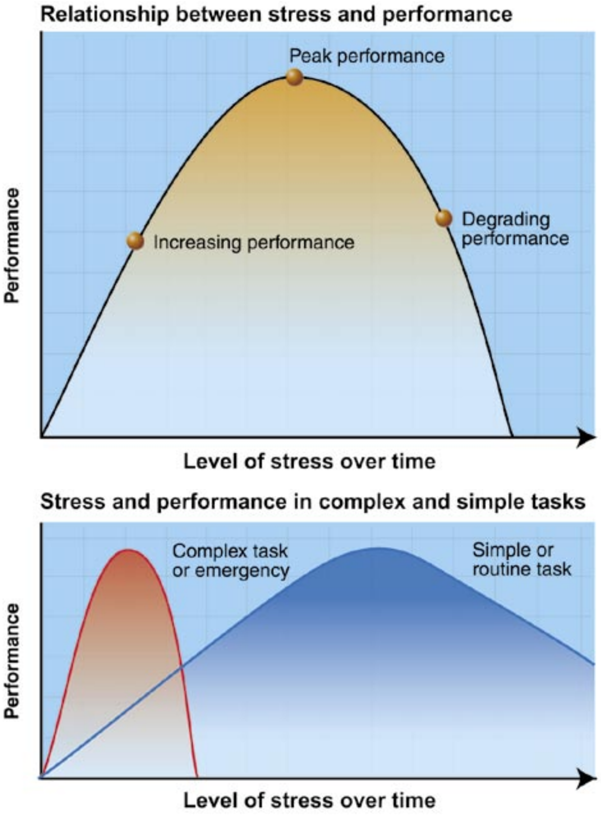

Figure 1-10. *Stress and Performance.*

3. If a problem occurs, remain calm. If time is not a pressing factor, follow the analytical approach to decision-making: think for a moment, weigh the alternatives, select and take an appropriate course of action, and then evaluate its effects.

   If an emergency situation occurs, remain calm and use the aeronautical decision-making (ADM) process to resolve the emergency. This process relies on the pilot's training and experience to accurately and automatically respond to an emergency situation. Constant training in handling emergency procedures will help reduce pilot stress when an emergency occurs.

4. Become thoroughly familiar with the aircraft, its operation, and emergency procedures. Also, maintain flight proficiency to build confidence.
5. Know and respect personal limits. Studies have suggested that highly experienced pilots have taken more chances when flying into potential icing conditions than low time or inexperienced pilots. Very low time pilots without experience may analyze and interpret the likelihood for "potential" flight into icing without the benefit of life experience, thereby making decisions closely aligned with the compilation of their training and recent academic knowledge. Highly experienced pilots may evaluate the current situation based upon the empirical information (sometimes diluted with time) coupled with their vast experience. This may lead to a level of greater acceptability of the situation because their experience has illustrated successful navigation of this problem before. Therefore, the automatic decision may be in error because not all salient facts are evaluated.
6. Do not allow small mistakes to be distractions during flight; rather, review and analyze them after landing.
7. If flying adds stress, either stop flying or seek professional help to manage stress within acceptable limits.

Medical Factors

A "go/no-go" decision based on a pilot's medical factors is made before each flight. The pilot should not only preflight check the aircraft, but also himself or herself before every flight. A pilot should ask, "Can I pass my medical examination right now?" If the answer is not an absolute "yes," *do not fly*. This is especially true for pilots embarking on flights in IMC. Instrument flying is much more demanding than flying in VMC, and peak performance is critical for the safety of flight.

Pilot performance can be seriously degraded by both prescribed and over-the-counter medications, as well as by the medical conditions for which they are taken. Many medications, such as tranquilizers, sedatives, strong pain relievers, and cough suppressants, have primary effects that impair judgment, memory, alertness, coordination, vision, and the ability to make calculations. Others, such as antihistamines, blood pressure drugs, muscle relaxants, and agents to control diarrhea and motion sickness, have side effects that impair the same critical functions. Any medication that depresses the nervous system, such as a sedative, tranquilizer, or antihistamine, makes a pilot much more susceptible to hypoxia.

Title 14 of the Code of Federal Regulations (14 CFR) prohibits pilots from performing crewmember duties while using any medication that affects the faculties in any way contrary to safety. The safest rule is not to fly as a crewmember while taking any medication, unless approved to do so by the Federal Aviation Administration (FAA). If there is any doubt regarding the effects of any medication, consult an Aviation Medical Examiner (AME) before flying.

Alcohol

14 CFR part 91 prohibits pilots from performing crewmember duties within 8 hours after drinking any alcoholic beverage or while under the influence. Extensive research has provided a number of facts about the hazards of alcohol consumption and flying. As little as one ounce of liquor, one bottle of beer, or four ounces of wine can impair flying skills and render a pilot much more susceptible to disorientation and hypoxia. Even after the body completely metabolizes a moderate amount of alcohol, a pilot can still be impaired for many hours. There is simply no way of increasing the metabolism of alcohol or alleviating a hangover.

Fatigue

Fatigue is one of the most treacherous hazards to flight safety, as it may not be apparent to a pilot until serious errors are made. Fatigue can be either acute (short-term) or chronic (long-term).

Acute Fatigue

A normal occurrence of everyday living, acute fatigue is the tiredness felt after long periods of physical and mental strain, including strenuous muscular effort, immobility, heavy mental workload, strong emotional pressure, monotony, and lack of sleep. Adequate rest, regular exercise, and proper nutrition prevent acute fatigue.

Indications of fatigue are generally subtle and hard to recognize because the individual being assessed is generally the person making the assessment, as in single pilot operations. Therefore, the pilot must look at small errors that occur to provide an indication of becoming fatigued. These include:

- Misplacing items during the preflight;
- Leaving material (pencils, charts) in the planning area;
- Missing radio calls;
- Answering calls improperly (read-backs); and
- Improper tuning of frequencies.

Chronic Fatigue

Chronic fatigue occurs when there is not enough time for a full recovery from repeated episodes of acute fatigue. Chronic fatigue's underlying cause is generally not "rest-related" and may have deeper points of origin. Therefore, rest alone may not resolve chronic fatigue.

Chronic fatigue is a combination of both physiological problems and psychological issues. Psychological problems such as financial, home life, or job related stresses cause a lack of qualified rest that is only resolved by mitigating the underpinning problems. Without resolution, performance continues to fall off, judgment becomes impaired, and unwarranted risks are taken. Recovery from chronic fatigue requires a prolonged and deliberate solution. In either case, unless adequate precautions are taken, personal performance could be impaired and adversely affect pilot judgment and decision-making.

IMSAFE Checklist

The following checklist, IMSAFE, is intended for a pilot's personal preflight use. A quick check of the items on this list will help a pilot make a good self-evaluation prior to any flight. If the answer to any of the checklist questions is yes, then the pilot should consider not flying.

Illness
Do I have any symptoms?

Medication
Have I been taking prescription or over-the-counter drugs?

Stress
Am I under psychological pressure from the job? Do I have money, health, or family problems?

Alcohol
Have I been drinking within 8 hours? Within 24 hours?

Fatigue
Am I tired and not adequately rested?

Eating
Have I eaten enough of the proper foods to keep adequately nourished during the entire flight?

Hazard Identification

In order to identify a hazard, it would be useful to define what a hazard is. The FAA System Safety course defines a hazard as: "a present condition, event, object, or circumstance that could lead or contribute to an unplanned or undesired event." Put simply, a hazard is a source of danger. Potential hazards may be identified from a number of internal and external sources. These may be based upon several concurrent factors that provide an indication and ultimate identification of a hazard. Consider the following situations:

Situation 1

The pilot has just taken off and is entering the clouds. Suddenly, there is an explosive sound. The sudden noise is disturbing and occurs as the pilot is given a new heading, a climb restriction, and the frequency for the departure control.

Situation 2

The pilot took off late in a rented aircraft (first time flying this model), and is now in night conditions due to the delay, and flying on an instrument flight rules (IFR) flight plan in IMC conditions. The radios do not seem to work well and develop static. They seem to be getting weaker. As the pilot proceeds, the rotating beacon stops flashing/rotating, and the lights become dimmer. As the situation progresses, the pilot is unaware of the problem because the generator warning light, (on the lower left of the panel) is obscured by the chart on the pilot's lap.

Both situations above represent hazards that must be dealt with differently and a level of risk must be associated with each depending on various factors affecting the flight.

Risk Analysis

Risk is defined as the future impact of a hazard that is not eliminated or controlled. It is the possibility of loss or injury. Risk analysis is the process whereby hazards are characterized by their likelihood and severity. Risk analysis evaluates the hazards to determine the outcomes and how abrupt that outcome will occur. The analysis applied will be qualitative to the degree that time allows resulting in either an analytical or automatic approach in the decision-making process.

In the first situation, the decision may be automatic: fly the airplane to a safe landing. Since automatic decision-making is based upon education and experience, an inexperienced pilot may react improperly to the situation which results in an inadequate action. To mitigate improper decision-making, immediate action items from emergency procedures should be learned. Training, education, and mentorship are all key factors in honing automatic decision-making skills.

In the second situation, if the pilot has a flashlight onboard, it can be used for illumination, although its light may degrade night vision. After changing the appropriate transponder code, and making calls in the blind, awareness of present location becomes imperative, especially if the pilot must execute a controlled descent to VMC conditions. Proper preflight planning conducted before departure and constant awareness of location provide an element of both comfort (reduces stress) and information from which the pilot can draw credible information.

In both cases, the outcomes can be successful through systems understanding, emergency procedures training, and correctly analyzing the risks associated with each course of action.

Crew Resource Management (CRM) and Single-Pilot Resource Management (SRM)

Crew resource management (CRM) and single-pilot resource management (SRM) is the ability for the crew or pilot to manage all resources effectively to ensure the outcome of the flight is successful. In general aviation, SRM will be most often used and its focus is on the single-pilot operation. SRM integrates the following:

- Situational Awareness
- Flight Deck Resource Management
- Task Management
- Aeronautical Decision-making (ADM) and Risk Management

SRM recognizes the need to seek proper information from these sources to make a valid decision. For instance, the pilot may have to request assistance from others and be assertive to resolve situations. Pilots should understand the need to seek information from other sources until they have the proper information to make the best decision. Once a pilot has gathered all pertinent information and made the appropriate decision, the pilot needs to perform an assessment of the action taken.

Situational Awareness

Situational awareness is the accurate perception of operational and environmental factors that affect the flight. It is a logical analysis based upon the machine, external support, environment, and the pilot. It is knowing what is going on.

Flight Deck Resource Management

CRM is the effective use of all available resources: human, equipment, and information. It focuses on communication skills, teamwork, task allocation, and decision-making. While CRM often concentrates on pilots who operate in crew environments, the elements and concepts also apply to single-pilot operations.

Human Resources

Human resources include everyone routinely working with the pilot to ensure flight safety. These people include, but are not limited to: weather briefers, flight line personnel, maintenance personnel, crew members, pilots, and air traffic personnel. Pilots need to effectively communicate with these people. This is accomplished by using the key components of the communication process: inquiry, advocacy, and assertion.

Pilots must recognize the need to seek enough information from these resources to make a valid decision. After the necessary information has been gathered, the pilot's decision must be passed on to those concerned, such as air traffic controllers, crew members, and passengers. The pilot may have to request assistance from others and be assertive to safely resolve some situations.

Equipment

Equipment in many of today's aircraft includes automated flight and navigation systems. These automatic systems, while providing relief from many routine flight deck tasks, present a different set of problems for pilots. The automation intended to reduce pilot workload essentially removes the pilot from the process of managing the aircraft, thereby reducing situational awareness and leading to complacency. Information from these systems needs to be continually monitored to ensure proper situational awareness. Pilots should be thoroughly familiar with the operation of and information provided by all systems used. It is essential that pilots be aware not only of equipment capabilities, but also equipment limitations in order to manage those systems effectively and safely.

Information Workload

Information workloads and automated systems, such as autopilots, need to be properly managed to ensure a safe flight. The pilot flying in IMC is faced with many tasks, each with a different level of importance to the outcome of the flight. For example, a pilot preparing to execute an instrument approach to an airport needs to review the approach chart, prepare the aircraft for the approach and landing, complete checklists, obtain information from Automatic Terminal Information Service (ATIS) or air traffic control (ATC), and set the navigation radios and equipment.

The pilot who effectively manages his or her workload will complete as many of these tasks as early as possible to preclude the possibility of becoming overloaded by last minute changes and communication priorities in the later, more critical stages of the approach. *Figure 1-11* shows the margin of safety is at the minimum level during this stage of the approach. Routine tasks delayed until the last minute can contribute to the pilot becoming overloaded and stressed, resulting in erosion of performance.

By planning ahead, a pilot can effectively reduce workload during critical phases of flight. If a pilot enters the final phases of the instrument approach unprepared, the pilot should recognize the situation, abandon the approach, and try it again after becoming better prepared. Effective resource management includes recognizing hazardous situations and attitudes, decision-making to promote good judgment and headwork, and managing the situation to ensure the safe outcome of the IFR flight.

Task Management

Pilots have a limited capacity for information. Once information flow exceeds the pilot's ability to mentally process the information any additional information will become unattended or displace other tasks and information already being processed. This is termed channel capacity and once reached only two alternatives exist: shed the unimportant tasks or perform all tasks at a less than optimal level. Like an electrical circuit being overloaded, either the consumption must be reduced or a circuit failure is experienced.

The pilot who effectively manages the tasks and properly prioritizes them will have a successful flight. For example, do not become distracted and fixate on an instrument light failure. This unnecessary focus displaces capability and prevents the pilot's ability to appreciate tasks of greater importance. By planning ahead, a pilot can effectively reduce workload during critical phases of a flight.

Aeronautical Decision-Making (ADM)

Flying safely requires the effective integration of three separate sets of skills. Most obvious are the basic stick-and-rudder skills needed to control the airplane. Next, are skills related to proficient operation of aircraft systems, and last, but not least, are ADM skills.

ADM is a systematic approach to the mental process used by pilots to consistently determine the best course of action in response to a given set of circumstances. The importance of learning effective ADM skills cannot be overemphasized. While progress is continually being made in the advancement of pilot training methods, airplane equipment and systems, and services for pilots, accidents still occur. Despite all the changes in technology to improve flight safety, one factor remains the same—the human factor. While the FAA strives to eliminate errors through training and safety programs, one fact remains: humans make errors. It is estimated that approximately 80 percent of all aviation accidents are human factors related.

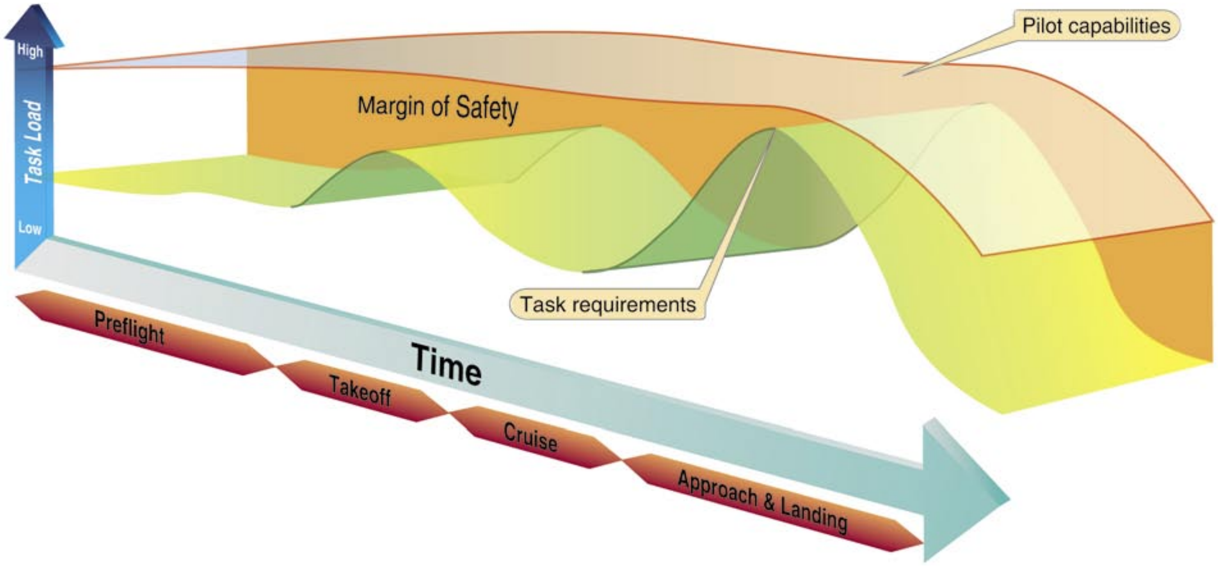

Figure 1-11. *The Margin of Safety.*

The ADM process addresses all aspects of decision making in the flight deck and identifies the steps involved in good decision making. While the ADM process will not eliminate errors, it will help the pilot recognize errors, and in turn enable the pilot to manage the error to minimize its effects. These steps are:

1. Identifying personal attitudes hazardous to safe flight;
2. Learning behavior modification techniques;
3. Learning how to recognize and cope with stress;
4. Developing risk assessment skills;
5. Using all resources; and
6. Evaluating the effectiveness of one's ADM skills.

Historically, the term "pilot error" has been used to describe the causes of these accidents. Pilot error means that an action or decision made by the pilot was the cause, or a contributing factor that led to the accident. This definition also includes the pilot's failure to make a decision or take action. From a broader perspective, the phrase "human factors related" more aptly describes these accidents since it is usually not a single decision that leads to an accident, but a chain of events triggered by a number of factors.

The poor judgment chain, sometimes referred to as the "error chain," is a term used to describe this concept of contributing factors in a human factors related accident. Breaking one link in the chain normally is all that is necessary to change the outcome of the sequence of events.

The Decision-Making Process

An understanding of the decision-making process provides a pilot with a foundation for developing ADM skills. Some situations, such as engine failures, require a pilot to respond immediately using established procedures with a little time for detailed analysis. This is termed automatic decision-making and is based upon training, experience, and recognition. Traditionally, pilots have been well trained to react to emergencies, but are not as well prepared to make decisions requiring a more reflective response where greater analysis is required. Typically during a flight, there is time to examine any changes that occur, gather information, and assess risk before reaching a decision. The steps leading to this conclusion constitute the decision-making process.

Defining the Problem

Problem definition is the first step in the decision-making process. Defining the problem begins with recognizing that a change has occurred or that an expected change did not occur. A problem is perceived first by the senses, then is distinguished through insight and experience. One critical error that can be made during the decision-making process is incorrectly defining the problem. For example, a low oil pressure reading could indicate that the engine is about to fail and an emergency landing should be planned, or it could mean that the oil pressure sensor has failed. The actions to be taken in each of these circumstances would be significantly different. One requires an immediate decision based upon training, experience, and evaluation of the situation; whereas the latter decision is based upon an analysis. It should be noted that the same indication could result in two different actions depending upon other influences.

Choosing a Course of Action

After the problem has been identified, the pilot must evaluate the need to react to it and determine the actions that may be taken to resolve the situation in the time available. The expected outcome of each possible action should be considered and the risks assessed before deciding on a response to the situation.

Implementing the Decision and Evaluating the Outcome

Although a decision may be reached and a course of action implemented, the decision-making process is not complete. It is important to think ahead and determine how the decision could affect other phases of flight. As the flight progresses, the pilot must continue to evaluate the outcome of the decision to ensure that it is producing the desired result.

Improper Decision-Making Outcomes

Pilots sometimes get in trouble not because of deficient basic skills or system knowledge, but rather because of faulty decision-making skills. Although aeronautical decisions may appear to be simple or routine, each individual decision in aviation often defines the options available for the next decision the pilot must make and the options, good or bad, they provide. Therefore, a poor decision early on in a flight can compromise the safety of the flight at a later time necessitating decisions that must be more accurate and decisive. Conversely, good decision-making early on in an emergency provide greater latitude for options later on.

FAA Advisory Circular (AC) 60-22, defines ADM as a systematic approach to the mental process of evaluating a given set of circumstances and determining the best course of action. ADM thus builds upon the foundation of conventional decision-making, but enhances the process to decrease the probability of pilot error. Specifically, ADM provides a structure to help the pilot use all resources to develop comprehensive situational awareness.

Models for Practicing ADM

Two models for practicing ADM are presented below.

Perceive, Process, Perform

The Perceive–Process–Perform (3P) model for ADM offers a simple, practical, and systematic approach that can be used during all phases of flight. *[Figure 1-12]* To use it, the pilot will:

- Perceive the given set of circumstances for a flight;
- Process by evaluating their impact on flight safety; and
- Perform by implementing the best course of action.

P
P
P
Aeronautical
Decision-
Making

Figure 1-12. *The 3P Model for Aeronautical Decision-Making.*

In the first step, the goal is to develop situational awareness by perceiving hazards, which are present events, objects, or circumstances that could contribute to an undesired future event. In this step, the pilot will systematically identify and list hazards associated with all aspects of the flight: pilot, aircraft, environment, and external pressures. It is important to consider how individual hazards might combine. Consider, for example, the hazard that arises when a new instrument pilot with no experience in actual instrument conditions wants to make a cross-country flight to an airport with low ceilings in order to attend an important business meeting.

In the second step, the goal is to process this information to determine whether the identified hazards constitute risk, which is defined as the future impact of a hazard that is not controlled or eliminated. The degree of risk posed by a given hazard can be measured in terms of exposure (number of people or resources affected), severity (extent of possible loss), and probability (the likelihood that a hazard will cause a loss). If the hazard is low ceilings, for example, the level of risk depends on a number of other factors, such as pilot training and experience, aircraft equipment and fuel capacity, and others.

In the third step, the goal is to perform by taking action to eliminate hazards or mitigate risk, and then continuously evaluate the outcome of this action. With the example of low ceilings at destination, for instance, the pilot can perform good ADM by selecting a suitable alternate, knowing where to find good weather, and carrying sufficient fuel to reach it. This course of action would mitigate the risk. The pilot also has the option to eliminate it entirely by waiting for better weather.

Once the pilot has completed the 3P decision process and selected a course of action, the process begins anew because now the set of circumstances brought about by the course of action requires analysis. The decision-making process is a continuous loop of perceiving, processing and performing.

The DECIDE Model

Another structured approach to ADM is the DECIDE model, which is a six-step process intended to provide a logical way of approaching decision-making. As in the 3P model, the elements of the DECIDE model represent a continuous loop process to assist a pilot in the decision-making required when faced with a situational change that requires judgment. *[Figure 1-13C]* The model is primarily focused on the intellectual component, but can have an impact on the motivational component of judgment as well. If a pilot continually uses the DECIDE Model in all decision-making, it becomes natural and results in better decisions being made under all types of situations. The steps in this approach are listed in *Figure 1-13C*.

In conventional decision-making, the need for a decision is triggered by recognition that something has changed or an expected change did not occur. Recognition of the change, or lack of change, is a vital step in any decision making process. Not noticing change in a situation can lead directly to a mishap. *[Figure 1-13A]* The change indicates that an appropriate response or action is necessary in order to modify the situation (or, at least, one of the elements that comprise it) and bring about a desired new situation. Therefore, situational awareness is the key to successful and safe decision making. At this point in the process, the pilot is faced with a need to evaluate the entire range of possible responses to the detected change and to determine the best course of action.

Figure 1-13B illustrates how the ADM process expands conventional decision-making, shows the interactions of the

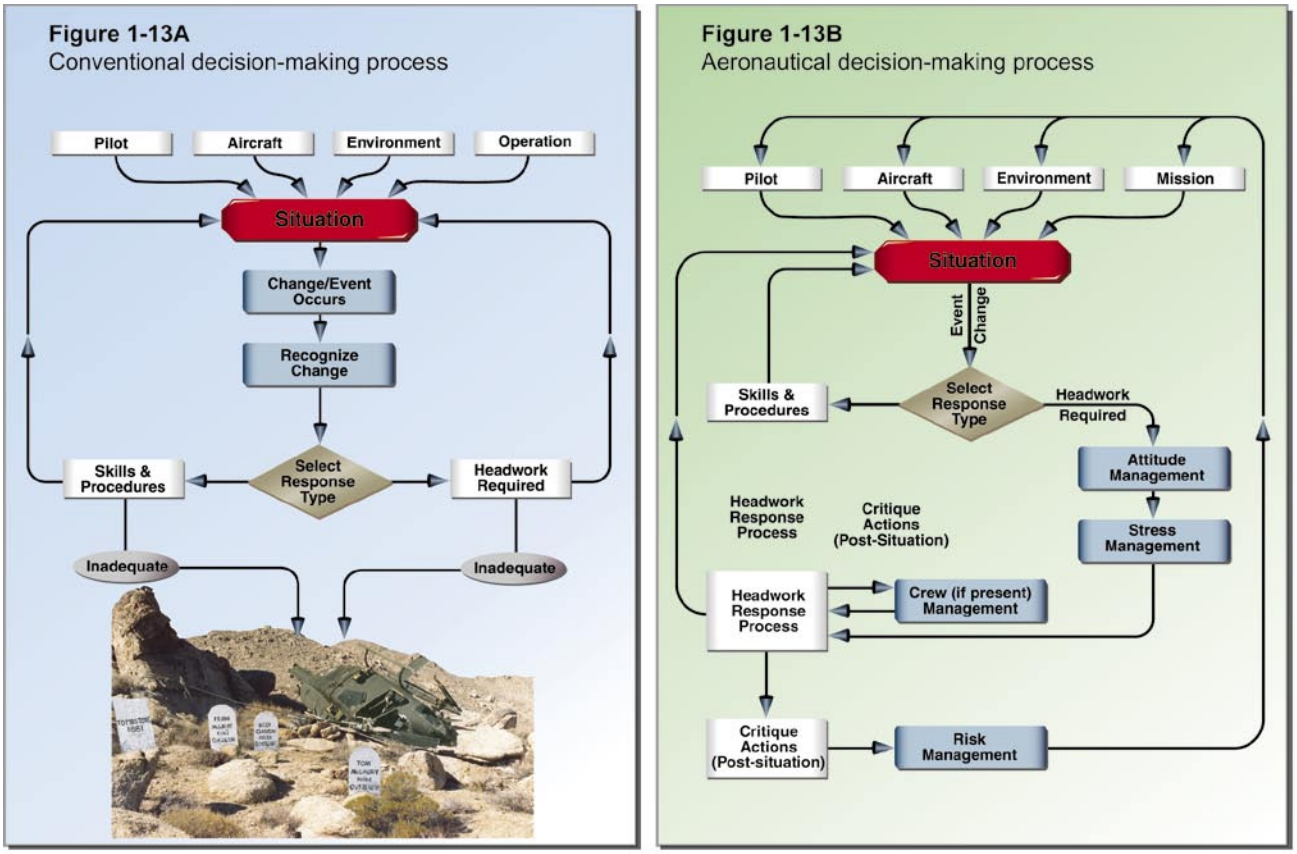

Figure 1-13A
Conventional decision-making process

Pilot
Aircraft
Environment
Operation
Situation
Change/Event Occurs
Recognize Change
Select Response Type
Skills & Procedures
Headwork Required
Inadequate
Inadequate

Figure 1-13B
Aeronautical decision-making process

Pilot
Aircraft
Environment
Mission
Situation
Event Change
Select Response Type
Skills & Procedures
Headwork Required
Attitude Management
Stress Management
Headwork Response Process
Critique Actions (Post-Situation)
Headwork Response Process
Crew (if present) Management
Critique Actions (Post-situation)
Risk Management

Figure 1-13C
The DECIDE Model

1. **Detect.** The decision maker detects the fact that change has occurred.
2. **Estimate.** The decision maker estimates the need to counter or react to the change.
3. **Choose.** The decision maker chooses a desirable outcome (in terms of success) for the flight.
4. **Identify.** The decision maker identifies actions which could successfully control the change.
5. **Do.** The decision maker takes the necessary action.
6. **Evaluate.** The decision maker evaluates the effect(s) of his/her action countering the change.

Figure 1-13. *Decision-Making.*

ADM steps, and how these steps can produce a safe outcome. Starting with the recognition of change, and following with an assessment of alternatives, a decision to act or not act is made, and the results are monitored. Pilots can use ADM to enhance their conventional decision-making process because it:

1. Increases their awareness of the importance of attitude in decision-making;
2. Teaches the ability to search for and establish relevance of information; and
3. Increases their motivation to choose and execute actions that ensure safety in the situational timeframe.

Hazardous Attitudes and Antidotes

Hazardous attitudes, which contribute to poor pilot judgment, can be effectively counteracted by redirecting that hazardous attitude so that correct action can be taken. Recognition of hazardous thoughts is the first step toward neutralizing them. After recognizing a thought as hazardous, the pilot should label it as hazardous, then state the corresponding antidote. Antidotes should be memorized for each of the hazardous attitudes so they automatically come to mind when needed. Each hazardous attitude along with its appropriate antidote is shown in *Figure 1-14*.

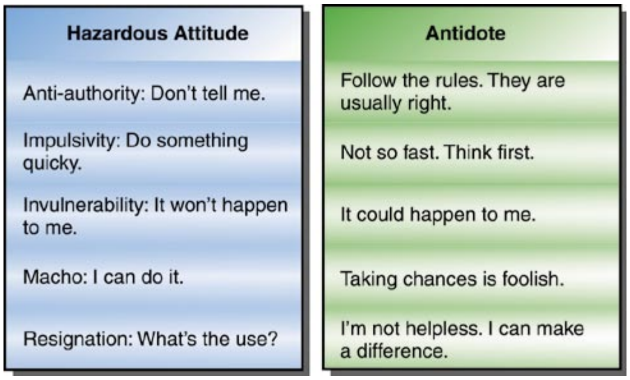

Hazardous Attitude	Antidote
Anti-authority: Don't tell me.	Follow the rules. They are usually right.
Impulsivity: Do something quicky.	Not so fast. Think first.
Invulnerability: It won't happen to me.	It could happen to me.
Macho: I can do it.	Taking chances is foolish.
Resignation: What's the use?	I'm not helpless. I can make a difference.

Figure 1-14. *The Five Antidotes to Hazardous Attitudes.*

Research has identified five hazardous attitudes that can affect a pilot's judgment, as well as antidotes for each of these five attitudes. ADM addresses the following:

1. Anti-authority ("Don't tell me!"). This attitude is found in pilots who do not like anyone telling them what to do. They may be resentful of having someone tell them what to do or may regard rules, regulations, and procedures as silly or unnecessary. However, there is always the prerogative to question authority if it is perceived to be in error.
2. Impulsivity ("Do something quickly!"). This attitude is found in pilots who frequently feel the need to do something—anything—immediately. They do not stop to think about what they are about to do, they do not select the best course of action, and they do the first thing that comes to mind.
3. Invulnerability ("It won't happen to me!"). Many pilots feel that accidents happen to others, but never to them. They know accidents can happen, and they know that anyone can be affected. They never really feel or believe that they will be personally involved. Pilots who think this way are more likely to take chances and increase risk.
4. Macho ("I can do it!"). Pilots who are always trying to prove that they are better than anyone else are thinking, "I can do it—I'll show them." Pilots with this type of attitude will try to prove themselves by taking risks in order to impress others. This pattern is characteristic in both men and women.
5. Resignation ("What's the use?"). These pilots do not see themselves as being able to make a great deal of difference in what happens to them. When things go well, these pilots are apt to think it is due to good luck. When things go badly, they may feel that someone is out to get them, or attribute it to bad luck. The pilot will leave the action to others, for better or worse. Sometimes, they will even go along with unreasonable requests just to be a "nice guy."

Chapter 2

Aerodynamic Factors

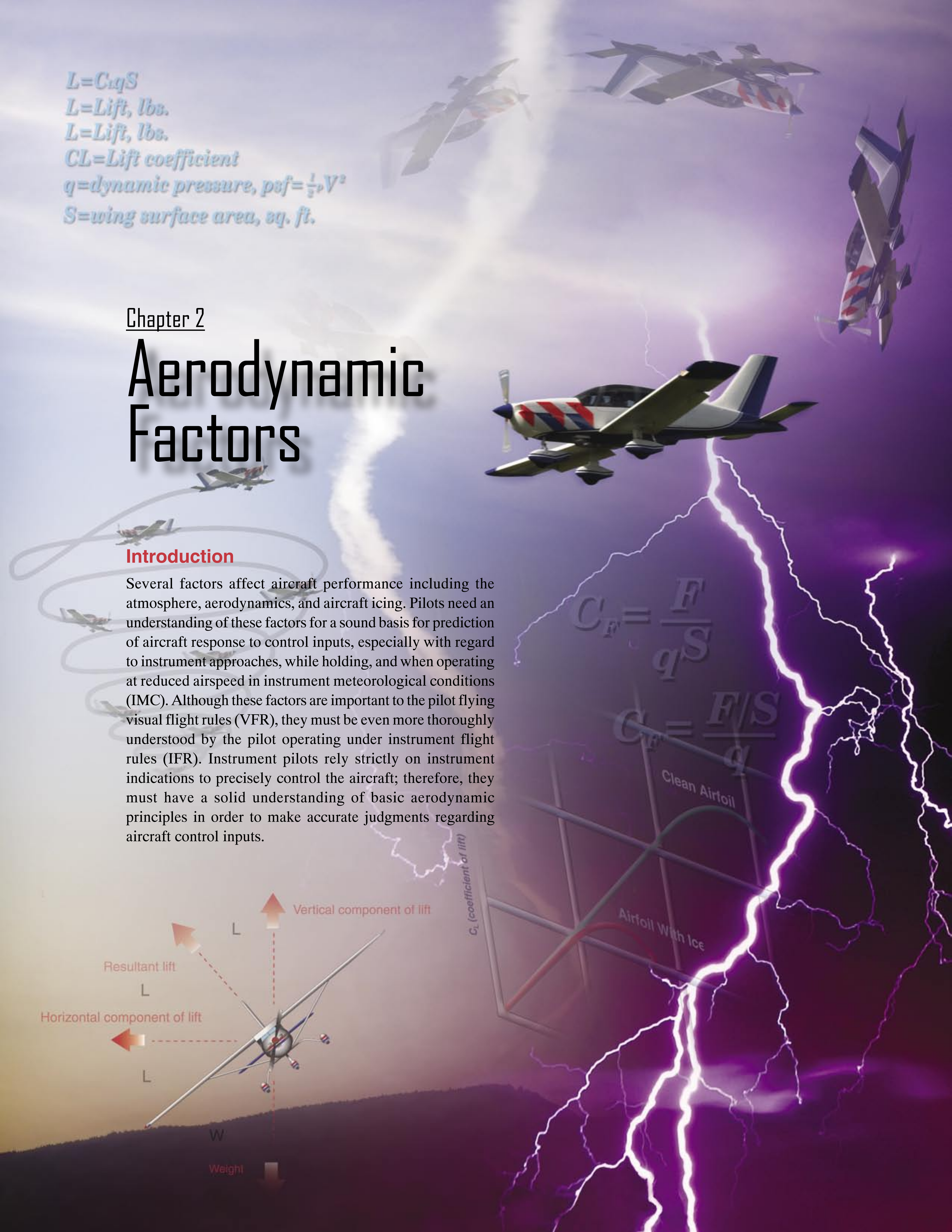
L=C_LqS
L=Lift, lbs.
L=Lift, lbs.
CL=Lift coefficient
q=dynamic pressure, psf=½ρV²
S=wing surface area, sq. ft.
C_F = F/qS
C_F = (F/S)/q
Clean Airfoil
Airfoil With Ice
C_L (coefficient of lift)
Vertical component of lift
L
Resultant lift
L
Horizontal component of lift
L
W
Weight

Introduction

Several factors affect aircraft performance including the atmosphere, aerodynamics, and aircraft icing. Pilots need an understanding of these factors for a sound basis for prediction of aircraft response to control inputs, especially with regard to instrument approaches, while holding, and when operating at reduced airspeed in instrument meteorological conditions (IMC). Although these factors are important to the pilot flying visual flight rules (VFR), they must be even more thoroughly understood by the pilot operating under instrument flight rules (IFR). Instrument pilots rely strictly on instrument indications to precisely control the aircraft; therefore, they must have a solid understanding of basic aerodynamic principles in order to make accurate judgments regarding aircraft control inputs.

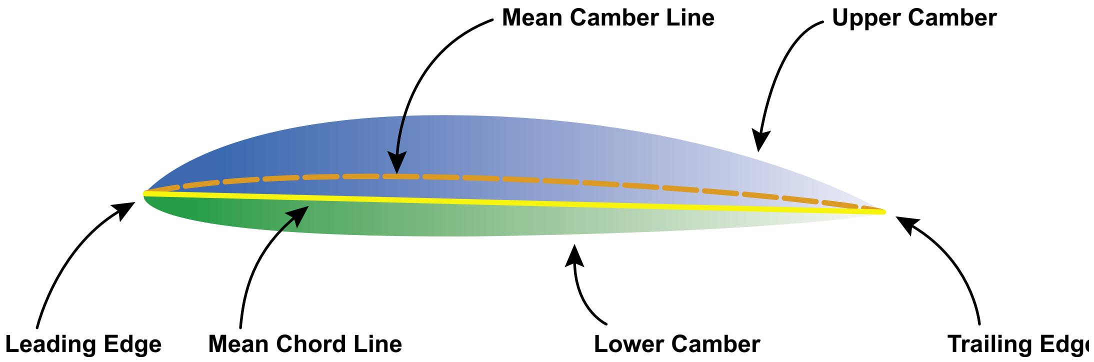

Figure 2-1. *The Airfoil.*

The Wing

To understand aerodynamic forces, a pilot needs to understand basic terminology associated with airfoils. *Figure 2-1* illustrates a typical airfoil.

The chord line is the straight line intersecting the leading and trailing edges of the airfoil, and the term chord refers to the chord line longitudinal length (length as viewed from the side).

The mean camber is a line located halfway between the upper and lower surfaces. Viewing the wing edgewise, the mean camber connects with the chord line at each end. The mean camber is important because it assists in determining aerodynamic qualities of an airfoil. The measurement of the maximum camber; inclusive of both the displacement of the mean camber line and its linear measurement from the end of the chord line, provide properties useful in evaluating airfoils.

Review of Basic Aerodynamics

The instrument pilot must understand the relationship and differences between several factors that affect the performance of an aircraft in flight. Also, it is crucial to understand how the aircraft reacts to various control and power changes, because the environment in which instrument pilots fly has inherent hazards not found in visual flying. The basis for this understanding is found in the four forces acting on an aircraft and Newton's Three Laws of Motion.

Relative Wind is the direction of the airflow with respect to an airfoil.

Angle of Attack is the acute angle measured between the relative wind, or flight path and the chord of the airfoil. *[Figure 2-2]*

Flight path is the course or track along which the aircraft is flying or is intended to be flown.

The Four Forces

The four basic forces *[Figure 2-3]* acting upon an aircraft in flight are lift, weight, thrust, and drag.

Lift

Lift is a component of the total aerodynamic force on an airfoil and acts perpendicular to the relative wind. Relative wind is the direction of the airflow with respect to an airfoil. This force acts straight up from the average (called mean) center of pressure (CP), which is called the center of lift. It should be noted that it is a point along the chord line of an airfoil through which all aerodynamic forces are considered to act. The magnitude of lift varies proportionately with speed, air density, shape and size of the airfoil, and angle of attack. During straight-and-level flight, lift and weight are equal.

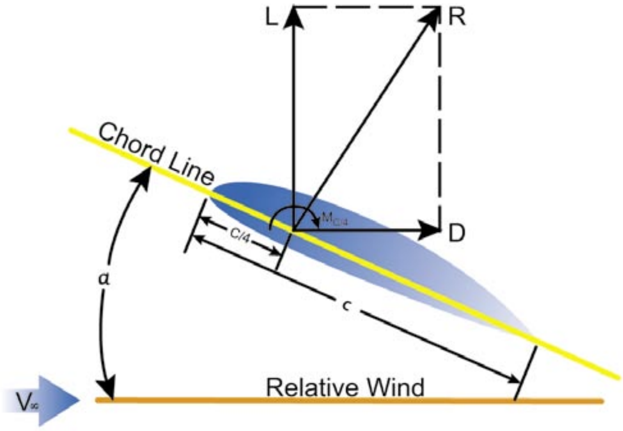

Figure 2-2. *Angle of Attack and Relative Wind.*

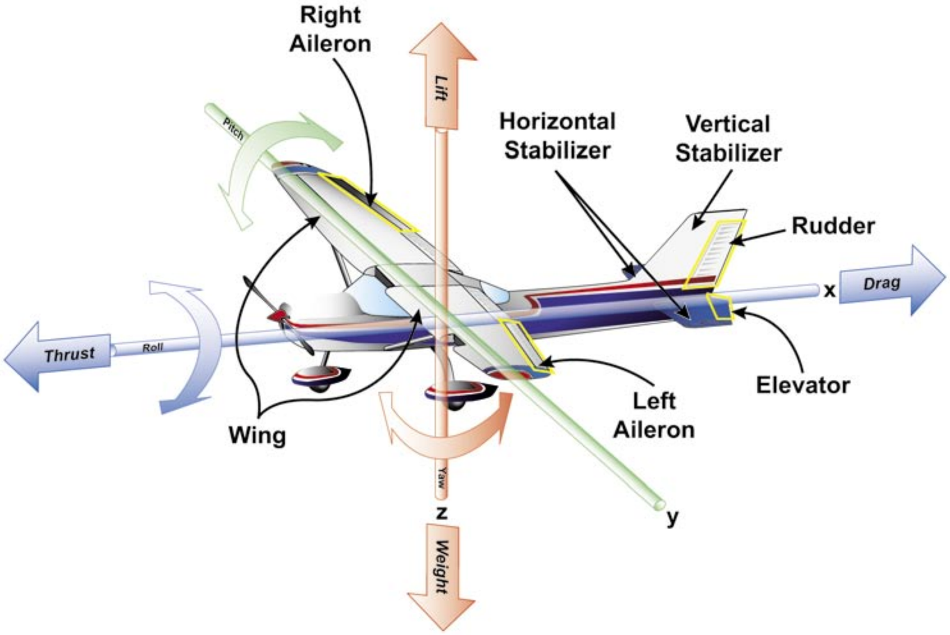

Figure 2-3. *The Four Forces and Three Axes of Rotation.*

Weight

Weight is the force exerted by an aircraft from the pull of gravity. It acts on an aircraft through its center of gravity (CG) and is straight down. This should not be confused with the center of lift, which can be significantly different from the CG. As an aircraft is descending, weight is greater than lift.

Thrust

Thrust is a force that drives an aircraft through the air and can be measured in thrust and/or horsepower. It is a component that is parallel to the center of thrust and overcomes drag providing the aircraft with its forward speed component.

Drag

Drag is the net aerodynamic force parallel to the relative wind and is generally a sum of two components: induced drag and parasite drag.

Induced drag

Induced drag is caused from the creation of lift and increases with angle of attack. Therefore, if the wing is not producing lift, induced drag is zero. Conversely, induced drag decreases with airspeed.

Parasite drag

Parasite drag is all drag not caused from the production of lift. Parasite drag is created by displacement of air by the aircraft, turbulence generated by the airfoil, and the hindrance of airflow as it passes over the surface of the aircraft or components. All of these forces create drag not from the production of lift but the movement of an object through an air mass. Parasite drag increases with speed and includes skin friction drag, interference drag, and form drag.

- Skin Friction Drag

Covering the entire "wetted" surface of the aircraft is a thin layer of air called a boundary layer. The air molecules on the surface have zero velocity in relation to the surface; however, the layer just above moves over the stagnant molecules below because it is pulled along by a third layer close to the free stream of air. The velocities of the layers increase as the distance from the surface increases until free stream velocity is reached, but all are affected by the free stream. The distance (total) between the skin surface and where free stream velocity is reached is called the boundary layer. At subsonic levels the cumulative layers are about the thickness of a playing card, yet their motion sliding over one another creates a drag force. This force retards motion due to the viscosity of the air and is called skin friction drag. Because skin friction drag is related to a large surface area its affect on smaller aircraft is small versus large transport aircraft where skin friction drag may be considerable.

- Interference Drag

Interference drag is generated by the collision of airstreams creating eddy currents, turbulence, or restrictions to smooth flow. For instance, the airflow around a fuselage and around the wing meet at some point, usually near the wing's root. These airflows interfere with each other causing a greater drag than the individual values. This is often the case when external items are placed on an aircraft. That is, the drag of each item individually, added to that of the aircraft, are less than that of the two items when allowed to interfere with one another.

- Form Drag

Form drag is the drag created because of the shape of a component or the aircraft. If one were to place a circular disk in an air stream, the pressure on both the top and bottom would be equal. However, the airflow starts to break down as the air flows around the back of the disk. This creates turbulence and hence a lower pressure results. Because the total pressure is affected by this reduced pressure, it creates a drag. Newer aircraft are generally made with consideration to this by fairing parts along the fuselage (teardrop) so that turbulence and form drag is reduced.

Total lift must overcome the total weight of the aircraft, which is comprised of the actual weight and the tail-down force used to control the aircraft's pitch attitude. Thrust must overcome total drag in order to provide forward speed with which to produce lift. Understanding how the aircraft's relationship between these elements and the environment provide proper interpretation of the aircraft's instruments.

Newton's First Law, the Law of Inertia

Newton's First Law of Motion is the Law of Inertia. It states that a body at rest will remain at rest, and a body in motion will remain in motion, at the same speed and in the same direction until affected by an outside force. The force with which a body offers resistance to change is called the force of inertia. Two outside forces are always present on an aircraft in flight: gravity and drag. The pilot uses pitch and thrust controls to counter or change these forces to maintain the desired flight path. If a pilot reduces power while in straight-and-level flight, the aircraft will slow due to drag. However, as the aircraft slows there is a reduction of lift, which causes the aircraft to begin a descent due to gravity. *[Figure 2-4]*

Newton's Second Law, the Law of Momentum

Newton's Second Law of Motion is the Law of Momentum, which states that a body will accelerate in the same direction as the force acting upon that body, and the acceleration will be directly proportional to the net force and inversely proportional to the mass of the body. Acceleration refers either to an increase or decrease in velocity, although deceleration is commonly used to indicate a decrease. This law governs the aircraft's ability to change flight path and speed, which are controlled by attitude (both pitch and bank) and thrust inputs. Speeding up, slowing down, entering climbs or descents, and turning are examples of accelerations that the pilot controls in everyday flight. *[Figure 2-5]*

Newton's Third Law, the Law of Reaction

Newton's Third Law of Motion is the Law of Reaction, which states that for every action there is an equal and opposite reaction. As shown in *Figure 2-6*, the action of the jet engine's thrust or the pull of the propeller lead to the reaction of the aircraft's forward motion. This law is also responsible for a portion of the lift that is produced by a wing, from the downward deflection of the airflow around it. This downward force of the relative wind results in an equal but opposite (upward) lifting force created by the airflow over the wing. *[Figure 2-6]*

Atmosphere

The atmosphere is the envelope of air which surrounds the Earth. A given volume of dry air contains about 78 percent nitrogen, 21 percent oxygen, and about 1 percent other gases such as argon, carbon dioxide, and others to a lesser degree. Although seemingly light, air does have weight and a one square inch column of the atmosphere at sea level weighs approximately 14.7 pounds. About one-half of the air by weight is within the first 18,000 feet. The remainder of the air is spread over a vertical distance in excess of 1,000 miles.

Air density is a result of the relationship between temperature and pressure. Air density is inversely related to temperature and directly related to pressure. For a constant pressure to be

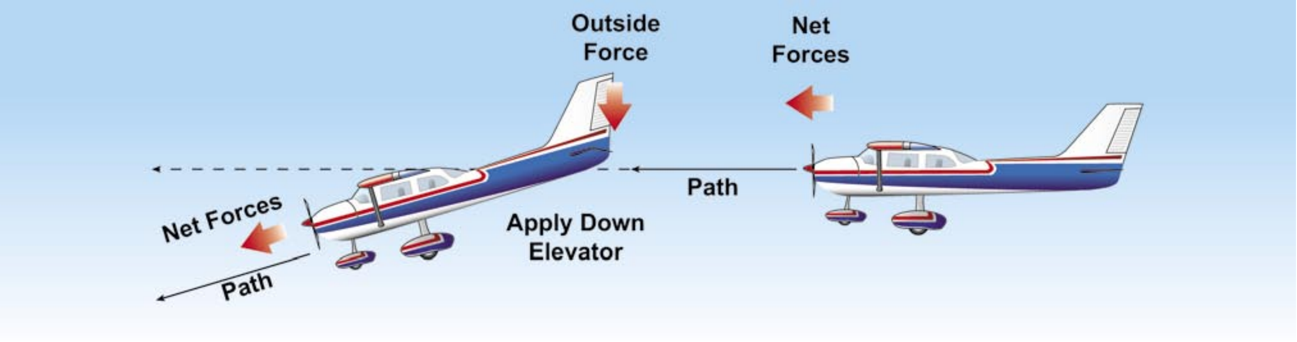

Figure 2-4. *Newton's First Law of Motion: the Law of Inertia.*

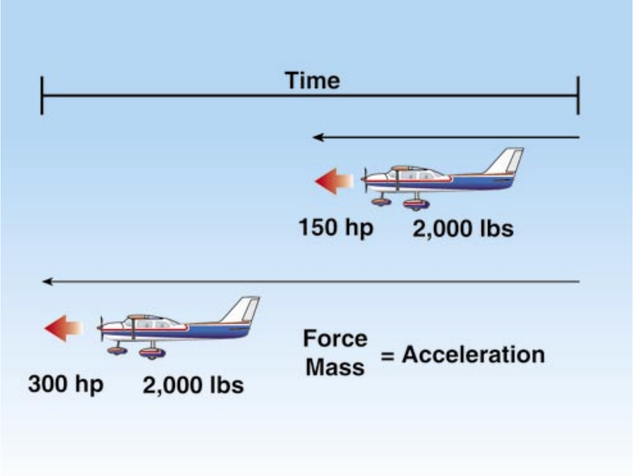

Figure 2-5. *Newton's Second Law of Motion: the Law of Momentum.*

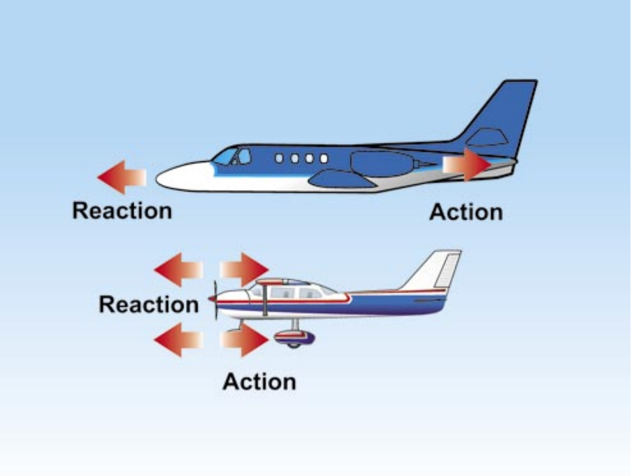

Figure 2-6. *Newton's Third Law of Motion: the Law of Reaction.*

maintained as temperature increases, density must decrease, and vice versa. For a constant temperature to be maintained as pressure increases, density must increase, and vice versa. These relationships provide a basis for understanding instrument indications and aircraft performance.

Layers of the Atmosphere

There are several layers to the atmosphere with the troposphere being closest to the Earth's surface extending to about 60,000 feet at the equator. Following is the stratosphere, mesosphere, ionosphere, thermosphere, and finally the exosphere. The tropopause is the thin layer between the troposphere and the stratosphere. It varies in both thickness and altitude but is generally defined where the standard lapse (generally accepted at 2° C per 1,000 feet) decreases significantly (usually down to 1° C or less).

International Standard Atmosphere (ISA)

The International Civil Aviation Organization (ICAO) established the ICAO Standard Atmosphere as a way of creating an international standard for reference and performance computations. Instrument indications and aircraft performance specifications are derived using this standard as a reference. Because the standard atmosphere is a derived set of conditions that rarely exist in reality, pilots need to understand how deviations from the standard affect both instrument indications and aircraft performance.

In the standard atmosphere, sea level pressure is 29.92" inches of mercury (Hg) and the temperature is 15° C (59° F). The standard lapse rate for pressure is approximately a 1" Hg decrease per 1,000 feet increase in altitude. The standard lapse rate for temperature is a 2° C (3.6° F) decrease per 1,000 feet increase, up to the top of the stratosphere. Since all aircraft performance is compared and evaluated in the environment of the standard atmosphere, all aircraft performance instrumentation is calibrated for the standard atmosphere. Because the actual operating conditions rarely, if ever, fit the standard atmosphere, certain corrections must apply to the instrumentation and aircraft performance. For instance, at 10,000 ISA predicts that the air pressure should be 19.92" Hg (29.92" - 10" Hg = 19.92") and the outside temperature at -5°C (15° C - 20° C). If the temperature or the pressure is different than the International Standard Atmosphere (ISA) prediction an adjustment must be made to performance predictions and various instrument indications.

Pressure Altitude

There are two measurements of the atmosphere that affect performance and instrument calibrations: pressure altitude and density altitude. Pressure altitude is the height above the standard datum pressure (SDP) (29.92" Hg, sea level under ISA) and is used for standardizing altitudes for flight levels (FL). Generally, flight levels are at or above 18,000 feet (FL 180), providing the pressure is at or above 29.92"Hg. For calculations involving aircraft performance when the altimeter is set for 29.92" Hg, the altitude indicated is the pressure altitude.

Density Altitude

Density altitude is pressure altitude corrected for nonstandard temperatures, and is used for determining aerodynamic performance in the nonstandard atmosphere. Density altitude increases as the density decreases. Since density varies directly with pressure, and inversely with temperature, a wide range of temperatures may exist with a given pressure altitude, which allows the density to vary. However, a known density occurs for any one temperature and pressure altitude combination. The density of the air has a significant effect on aircraft and engine performance. Regardless of the actual altitude above sea level an aircraft is operating at, its performance will be as though it were operating at an altitude equal to the existing density altitude.

If a chart is not available the density altitude can be estimated by adding 120 feet for every degree Celsius above the ISA. For example, at 3,000 feet pressure altitude (PA), the ISA prediction is 9° C (15° C - [lapse rate of 2° C per 1,000 feet x 3 = 6° C]). However, if the actual temperature is 20° C (11° C more than that predicted by ISA) then the difference of 11° C is multiplied by 120 feet equaling 1,320. Adding this figure to the original 3,000 feet provides a density altitude of 4,320 feet (3,000 feet + 1,320 feet).

Lift

Lift always acts in a direction perpendicular to the relative wind and to the lateral axis of the aircraft. The fact that lift is referenced to the wing, not to the Earth's surface, is the source of many errors in learning flight control. Lift is not always "up." Its direction relative to the Earth's surface changes as the pilot maneuvers the aircraft.

The magnitude of the force of lift is directly proportional to the density of the air, the area of the wings, and the airspeed. It also depends upon the type of wing and the angle of attack. Lift increases with an increase in angle of attack up to the stalling angle, at which point it decreases with any further increase in angle of attack. In conventional aircraft, lift is therefore controlled by varying the angle of attack and speed.

Pitch/Power Relationship

An examination of *Figure 2-7* illustrates the relationship between pitch and power while controlling flight path and airspeed. In order to maintain a constant lift, as airspeed is reduced, pitch must be increased. The pilot controls pitch through the elevators, which control the angle of attack. When back pressure is applied on the elevator control, the tail lowers and the nose rises, thus increasing the wing's angle of attack and lift. Under most conditions the elevator is placing downward pressure on the tail. This pressure requires energy that is taken from aircraft performance (speed). Therefore, when the CG is closer to the aft portion of the aircraft the elevator downward forces are less. This results in less energy used for downward forces, in turn resulting in more energy applied to aircraft performance.

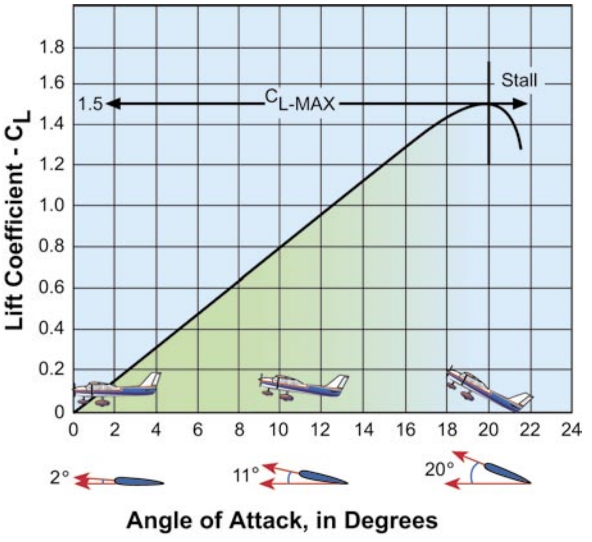

Figure 2-7. *Relationship of Lift to Angle of Attack.*

Thrust is controlled by using the throttle to establish or maintain desired airspeeds. The most precise method of controlling flight path is to use pitch control while simultaneously using power (thrust) to control airspeed. In order to maintain a constant lift, a change in pitch requires a change in power, and vice versa.

If the pilot wants the aircraft to accelerate while maintaining altitude, thrust must be increased to overcome drag. As the aircraft speeds up, lift is increased. To prevent gaining altitude, the pitch angle must be lowered to reduce the angle of attack and maintain altitude. To decelerate while maintaining altitude, thrust must be decreased to less than the value of drag. As the aircraft slows down, lift is reduced. To prevent losing altitude, the pitch angle must be increased in order to increase the angle of attack and maintain altitude.

Drag Curves

When induced drag and parasite drag are plotted on a graph, the total drag on the aircraft appears in the form of a "drag curve." Graph A of *Figure 2-8* shows a curve based on thrust versus drag, which is primarily used for jet aircraft. Graph B of *Figure 2-8* is based on power versus drag, and it is used for propeller-driven aircraft. This chapter focuses on power versus drag charts for propeller-driven aircraft.

Understanding the drag curve can provide valuable insight into the various performance parameters and limitations of the aircraft. Because power must equal drag to maintain a steady airspeed, the curve can be either a drag curve or a power required curve. The power required curve represents the amount of power needed to overcome drag in order to maintain a steady speed in level flight.

The propellers used on most reciprocating engines achieve peak propeller efficiencies in the range of 80 to 88 percent. As airspeed increases, the propeller efficiency increases until it reaches its maximum. Any airspeed above this maximum point causes a reduction in propeller efficiency. An engine that produces 160 horsepower will have only about 80 percent of that power converted into available horsepower,

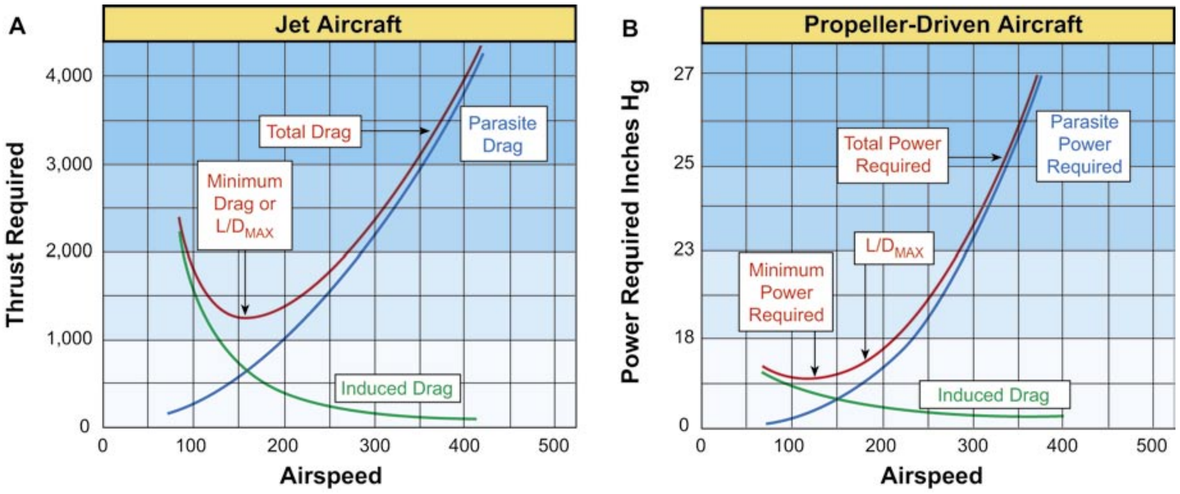

Figure 2-8. *Thrust and Power Required Curves.*

approximately 128 horsepower. The remainder is lost energy. This is the reason the thrust and power available curves change with speed.

Regions of Command

The drag curve also illustrates the two regions of command: the region of normal command, and the region of reversed command. The term "region of command" refers to the relationship between speed and the power required to maintain or change that speed. "Command" refers to the input the pilot must give in terms of power or thrust to maintain a new speed once reached.

The "region of normal command" occurs where power must be added to increase speed. This region exists at speeds higher than the minimum drag point primarily as a result of parasite drag. The "region of reversed command" occurs where additional power is needed to maintain a slower airspeed. This region exists at speeds slower than the minimum drag point (L/D$_{MAX}$ on the thrust required curve, *Figure 2-8*) and is primarily due to induced drag. *Figure 2-9* shows how one power setting can yield two speeds, points 1 and 2. This is because at point 1 there is high induced drag and low parasite drag, while at point 2 there is high parasite drag and low induced drag.

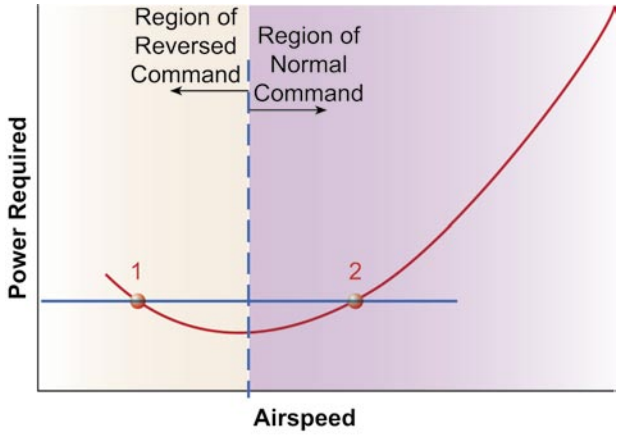

Figure 2-9. *Regions of Command.*

Control Characteristics

Most flying is conducted in the region of normal command: for example, cruise, climb, and maneuvers. The region of reversed command may be encountered in the slow-speed phases of flight during takeoff and landing; however, for most general aviation aircraft, this region is very small and is below normal approach speeds.

Flight in the region of normal command is characterized by a relatively strong tendency of the aircraft to maintain the trim speed. Flight in the region of reversed command is characterized by a relatively weak tendency of the aircraft to maintain the trim speed. In fact, it is likely the aircraft exhibits no inherent tendency to maintain the trim speed in this area. For this reason, the pilot must give particular attention to precise control of airspeed when operating in the slow-speed phases of the region of reversed command.

Operation in the region of reversed command does not imply that great control difficulty and dangerous conditions exist. However, it does amplify errors of basic flying technique—making proper flying technique and precise control of the aircraft very important.

Speed Stability

Normal Command

The characteristics of flight in the region of normal command are illustrated at point A on the curve in *Figure 2-10*. If the aircraft is established in steady, level flight at point A, lift is equal to weight, and the power available is set equal to the power required. If the airspeed is increased with no changes

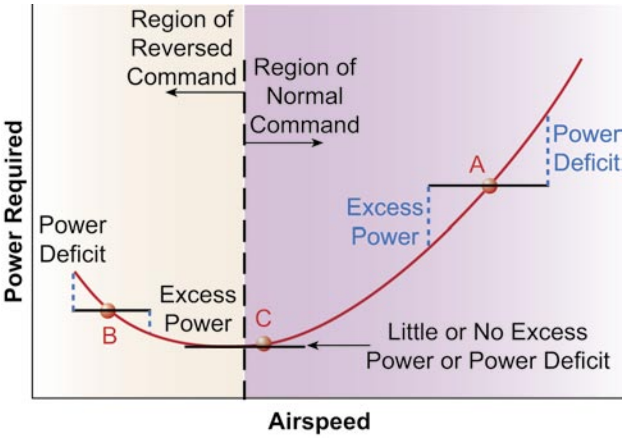

Figure 2-10. *Region of Speed Stability.*

to the power setting, a power deficiency exists. The aircraft has a natural tendency to return to the initial speed to balance power and drag. If the airspeed is reduced with no changes to the power setting, an excess of power exists. The aircraft has a natural tendency to speed up to regain the balance between power and drag. Keeping the aircraft in proper trim enhances this natural tendency. The static longitudinal stability of the aircraft tends to return the aircraft to the original trimmed condition.

An aircraft flying in steady, level flight at point C is in equilibrium. *[Figure 2-10]* If the speed were increased or decreased slightly, the aircraft would tend to remain at that speed. This is because the curve is relatively flat and a slight change in speed does not produce any significant excess or deficiency in power. It has the characteristic of neutral stability, i.e., the aircraft's tendency is to remain at the new speed.

Reversed Command

The characteristics of flight in the region of reversed command are illustrated at point B on the curve in *Figure 2-10*. If the aircraft is established in steady, level flight at point B, lift is equal to weight, and the power available is set equal to the power required. When the airspeed is increased greater than point B, an excess of power exists. This causes the aircraft to accelerate to an even higher speed. When the aircraft is slowed to some airspeed lower than point B, a deficiency of power exists. The natural tendency of the aircraft is to continue to slow to an even lower airspeed.

This tendency toward instability happens because the variation of excess power to either side of point B magnifies the original change in speed. Although the static longitudinal stability of the aircraft tries to maintain the original trimmed condition, this instability is more of an influence because of the increased induced drag due to the higher angles of attack in slow-speed flight.

Trim

The term trim refers to employing adjustable aerodynamic devices on the aircraft to adjust forces so the pilot does not have to manually hold pressure on the controls. One means is to employ trim tabs. A trim tab is a small, adjustable hinged surface, located on the trailing edge of the elevator, aileron, or rudder control surfaces. (Some aircraft use adjustable stabilizers instead of trim tabs for pitch trim.) Trimming is accomplished by deflecting the tab in the direction opposite to that in which the primary control surface must be held. The force of the airflow striking the tab causes the main control surface to be deflected to a position that corrects the unbalanced condition of the aircraft.

Because the trim tabs use airflow to function, trim is a function of speed. Any change in speed results in the need to re-trim the aircraft. An aircraft properly trimmed in pitch seeks to return to the original speed before the change. It is very important for instrument pilots to keep the aircraft in constant trim. This reduces the pilot's workload significantly, allowing attention to other duties without compromising aircraft control.

Slow-Speed Flight

Anytime an aircraft is flying near the stalling speed or the region of reversed command, such as in final approach for a normal landing, the initial part of a go around, or maneuvering in slow flight, it is operating in what is called slow-speed flight. If the aircraft weighs 4,000 pounds, the lift produced by the aircraft must be 4,000 pounds. When lift is less than 4,000 pounds, the aircraft is no longer able to sustain level flight, and consequently descends. During intentional descents, this is an important factor and is used in the total control of the aircraft.

However, because lift is required during low speed flight and is characterized by high angles of attack, flaps or other high lift devices are needed to either change the camber of the airfoil, or delay the boundary level separation. Plain and split flaps *[Figure 2-11]* are most commonly used to change the camber of an airfoil. It should be noted that with the application of flaps, the aircraft will stall at a lower angle of attack. The basic wing stalls at 18° without flaps but with the application of the flaps extended (to $C_{L\text{-}MAX}$ position) the new angle of attack at which point the aircraft will stall is 15°. However, the value of lift (flaps extended to the $C_{L\text{-}MAX}$ position) produces more lift than lift at 18° on the basic wing.

Delaying the boundary layer separation is another way to increase $C_{L\text{-}MAX}$. Several methods are employed (such as suction and use of a blowing boundary layer control), but the most common device used on general aviation light aircraft is the vortex generator. Small strips of metal placed along the wing (usually in front of the control surfaces) create turbulence. The turbulence in turn mixes high energy air from outside the boundary layer with boundary layer air. The effect is similar to other boundary layer devices. *[Figure 2-12]*

Small Airplanes

Most small airplanes maintain a speed well in excess of 1.3 times V_{SO} on an instrument approach. An airplane with a stall speed of 50 knots (V_{SO}) has a normal approach speed of 65 knots. However, this same airplane may maintain 90 knots (1.8 V_{SO}) while on the final segment of an instrument approach. The landing gear will most likely be extended at the beginning of the descent to the minimum descent altitude, or upon intercepting the glide slope of the instrument landing system. The pilot may also select an intermediate flap setting for this phase of the approach. The airplane at this speed has good positive speed stability, as represented by point A on *Figure 2-10*. Flying in this regime permits the pilot to make slight pitch changes without changing power settings, and accept minor speed changes knowing that when the pitch is returned to the initial setting, the speed returns to the original setting. This reduces the pilot's workload.

Aircraft are usually slowed to a normal landing speed when on the final approach just prior to landing. When slowed to 65 knots, (1.3 V_{SO}), the airplane will be close to point C. *[Figure 2-10]* At this point, precise control of the pitch and power becomes more crucial for maintaining the correct speed. Pitch and power coordination is necessary because the speed stability is relatively neutral since the speed tends to remain at the new value and not return to the original setting. In addition to the need for more precise airspeed control, the pilot normally changes the aircraft's configuration by extending landing flaps. This configuration change means the pilot must be alert to unwanted pitch changes at a low altitude. If allowed to slow several knots, the airplane could enter the region of reversed command. At this point, the airplane could develop an unsafe sink rate and continue to lose speed unless the pilot takes a prompt corrective action. Proper pitch and power coordination is critical in this region due to speed instability and the tendency of increased divergence from the desired speed.

Large Airplanes

Pilots of larger airplanes with higher stall speeds may find the speed they maintain on the instrument approach is near 1.3 V_{SO}, putting them near point C *[Figure 2-10]* the entire time the airplane is on the final approach segment. In this case, precise speed control is necessary throughout the approach. It may be necessary to temporarily select excessive, or deficient thrust in relation to the target thrust setting in order to quickly correct for airspeed deviations.

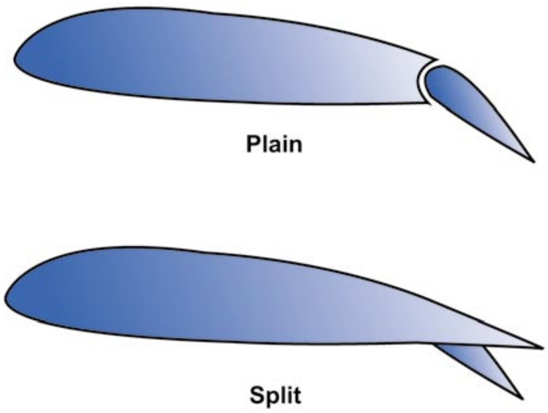

Figure 2-11. *Various Types of Flaps.*

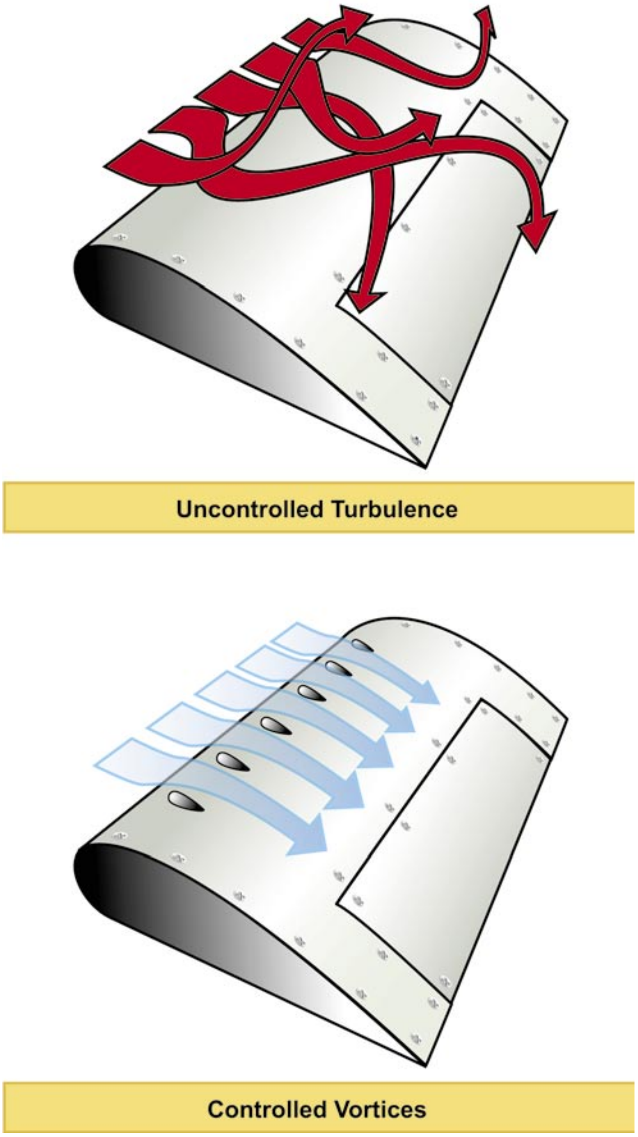

Figure 2-12. *Vortex Generators.*

For example, a pilot is on an instrument approach at 1.3 V_{SO}, a speed near L/D_{MAX}, and knows that a certain power setting maintains that speed. The airplane slows several knots below the desired speed because of a slight reduction in the power setting. The pilot increases the power slightly, and the airplane begins to accelerate, but at a slow rate. Because the airplane is still in the "flat part" of the drag curve, this slight increase in power will not cause a rapid return to the desired speed. The pilot may need to increase the power higher than normally needed to maintain the new speed, allow the airplane to accelerate, then reduce the power to the setting that maintains the desired speed.

Climbs

The ability for an aircraft to climb depends upon an excess power or thrust over what it takes to maintain equilibrium. Excess power is the available power over and above that required to maintain horizontal flight at a given speed. Although the terms power and thrust are sometimes used interchangeably (erroneously implying they are synonymous), distinguishing between the two is important when considering climb performance. Work is the product of a force moving through a distance and is usually independent of time. Power implies work rate or units of work per unit of time, and as such is a function of the speed at which the force is developed. Thrust, also a function of work, means the force which imparts a change in the velocity of a mass.

During take off, the aircraft does not stall even though it may be in a climb near the stall speed. The reason is that excess power (used to produce thrust) is used during this flight regime. Therefore, it is important if an engine fails after take off, to compensate the loss of thrust with pitch and airspeed.

For a given weight of the aircraft, the angle of climb depends on the difference between thrust and drag, or the excess thrust. When the excess thrust is zero, the inclination of the flight path is zero, and the aircraft is in steady, level flight. When thrust is greater than drag, the excess thrust allows a climb angle depending on the amount of excess thrust. When thrust is less than drag, the deficiency of thrust induces an angle of descent.

Acceleration in Cruise Flight

Aircraft accelerate in level flight because of an excess of power over what is required to maintain a steady speed. This is the same excess power used to climb. Upon reaching the desired altitude with pitch being lowered to maintain that altitude, the excess power now accelerates the aircraft to its cruise speed. However, reducing power too soon after level off results in a longer period of time to accelerate.

Turns

Like any moving object, an aircraft requires a sideward force to make it turn. In a normal turn, this force is supplied by banking the aircraft in order to exert lift inward, as well as upward. The force of lift is separated into two components at right angles to each other. *[Figure 2-13]* The upward acting lift together with the opposing weight becomes the vertical lift component. The horizontally acting lift and its opposing centrifugal force are the horizontal lift component, or centripetal force. This horizontal lift component is the sideward force that causes an aircraft to turn. The equal and opposite reaction to this sideward force is centrifugal force, which is merely an apparent force as a result of inertia.

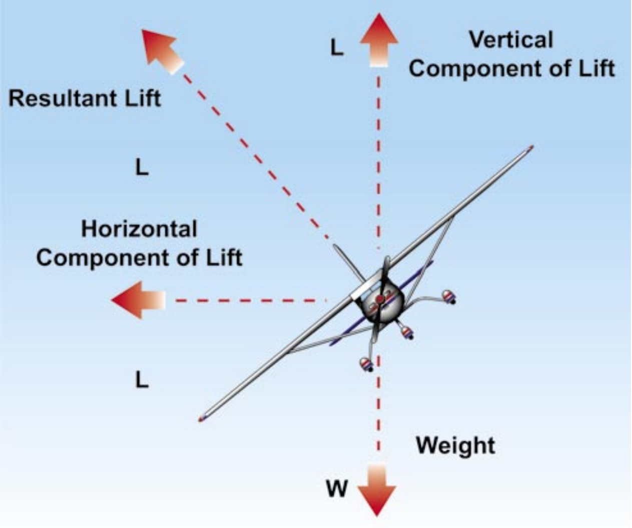

Figure 2-13. *Forces In a Turn.*

The relationship between the aircraft's speed and bank angle to the rate and radius of turns is important for instrument pilots to understand. The pilot can use this knowledge to properly estimate bank angles needed for certain rates of turn, or to determine how much to lead when intercepting a course.

Rate of Turn

The rate of turn, normally measured in degrees per second, is based upon a set bank angle at a set speed. If either one of these elements changes, the rate of turn changes. If the aircraft increases its speed without changing the bank angle, the rate of turn decreases. Likewise, if the speed decreases without changing the bank angle, the rate of turn increases.

Changing the bank angle without changing speed also causes the rate of turn to change. Increasing the bank angle without changing speed increases the rate of turn, while decreasing the bank angle reduces the rate of turn.

The standard rate of turn, 3° per second, is used as the main reference for bank angle. Therefore, the pilot must understand how the angle of bank varies with speed changes, such as slowing down for holding or an instrument approach. *Figure 2-14* shows the turn relationship with reference to a constant bank angle or a constant airspeed, and the effects on rate of turn and radius of turn. A rule of thumb for determining the standard rate turn is to divide the airspeed by ten and add 7. An aircraft with an airspeed of 90 knots takes a bank angle of 16° to maintain a standard rate turn (90 divided by 10 plus 7 equals 16°).

Radius of Turn

The radius of turn varies with changes in either speed or bank. If the speed is increased without changing the bank angle, the radius of turn increases, and vice versa. If the speed is constant, increasing the bank angle reduces the radius of turn, while decreasing the bank angle increases the radius of turn. This means that intercepting a course at a higher speed requires more distance, and therefore, requires a longer lead. If the speed is slowed considerably in preparation for holding or an approach, a shorter lead is needed than that required for cruise flight.

Coordination of Rudder and Aileron Controls

Any time ailerons are used, adverse yaw is produced. Adverse yaw is caused when the ailerons are deflected as a roll motion (as in turn) is initiated. In a right turn, the right aileron is deflected upward while the left is deflected downward. Lift is increased on the left side and reduced on the right, resulting in a bank to the right. However, as a result of producing lift on the left, induced drag is also increased on the left side. The drag causes the left wing to slow down, in turn causing the nose of the aircraft to initially move (left) in the direction opposite of the turn. Correcting for this yaw with rudder, when entering and exiting turns, is necessary for precise control of the airplane when flying on instruments. The pilot can tell if the turn is coordinated by checking the ball in the turn-and-slip indicator or the turn coordinator. *[Figure 2-15]*

As the aircraft banks to enter a turn, a portion of the wing's vertical lift becomes the horizontal component; therefore, without an increase in back pressure, the aircraft loses altitude during the turn. The loss of vertical lift can be offset by increasing the pitch in one-half bar width increments. Trim may be used to relieve the control pressures; however, if used, it has to be removed once the turn is complete.

In a slipping turn, the aircraft is not turning at the rate appropriate to the bank being used, and the aircraft falls to the inside of the turn. The aircraft is banked too much for the rate of turn, so the horizontal lift component is greater than the centrifugal force. A skidding turn results from excess of centrifugal force over the horizontal lift component, pulling the aircraft toward the outside of the turn. The rate of turn is too great for the angle of bank, so the horizontal lift component is less than the centrifugal force.

The ball instrument indicates the quality of the turn, and should be centered when the wings are banked. If the ball is off of center on the side toward the turn, the aircraft is slipping and rudder pressure should be added on that side to increase the rate of turn or the bank angle should be reduced. If the ball is off of center on the side away from the turn, the aircraft is skidding and rudder pressure toward the turn should be relaxed or the bank angle should be increased. If the aircraft is properly rigged, the ball should be in the center when the wings are level; use rudder and/or aileron trim if available.

The increase in induced drag (caused by the increase in angle of attack necessary to maintain altitude) results in a minor loss of airspeed if the power setting is not changed.

Load Factor

Any force applied to an aircraft to deflect its flight from a straight line produces a stress on its structure; the amount of this force is termed load factor. A load factor is the ratio of

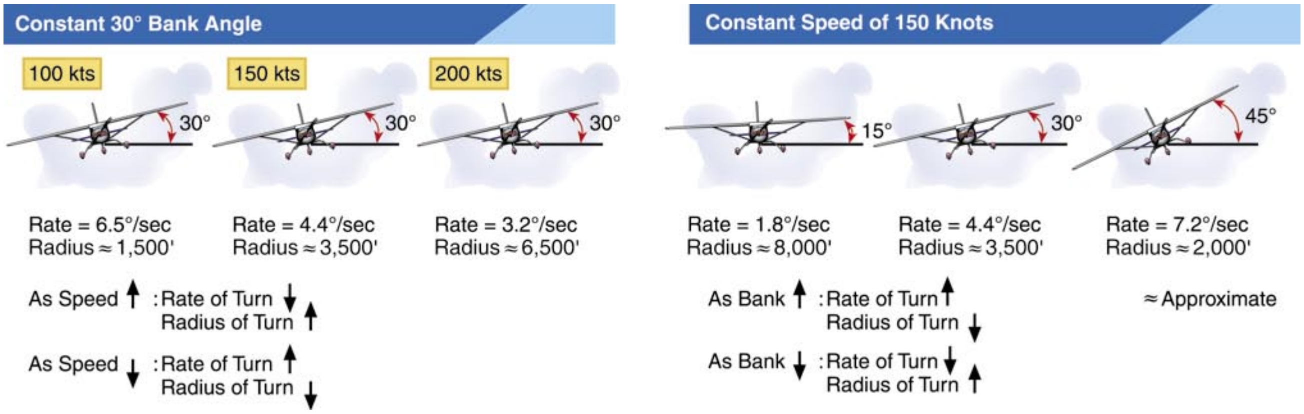

Figure 2-14. *Turns.*

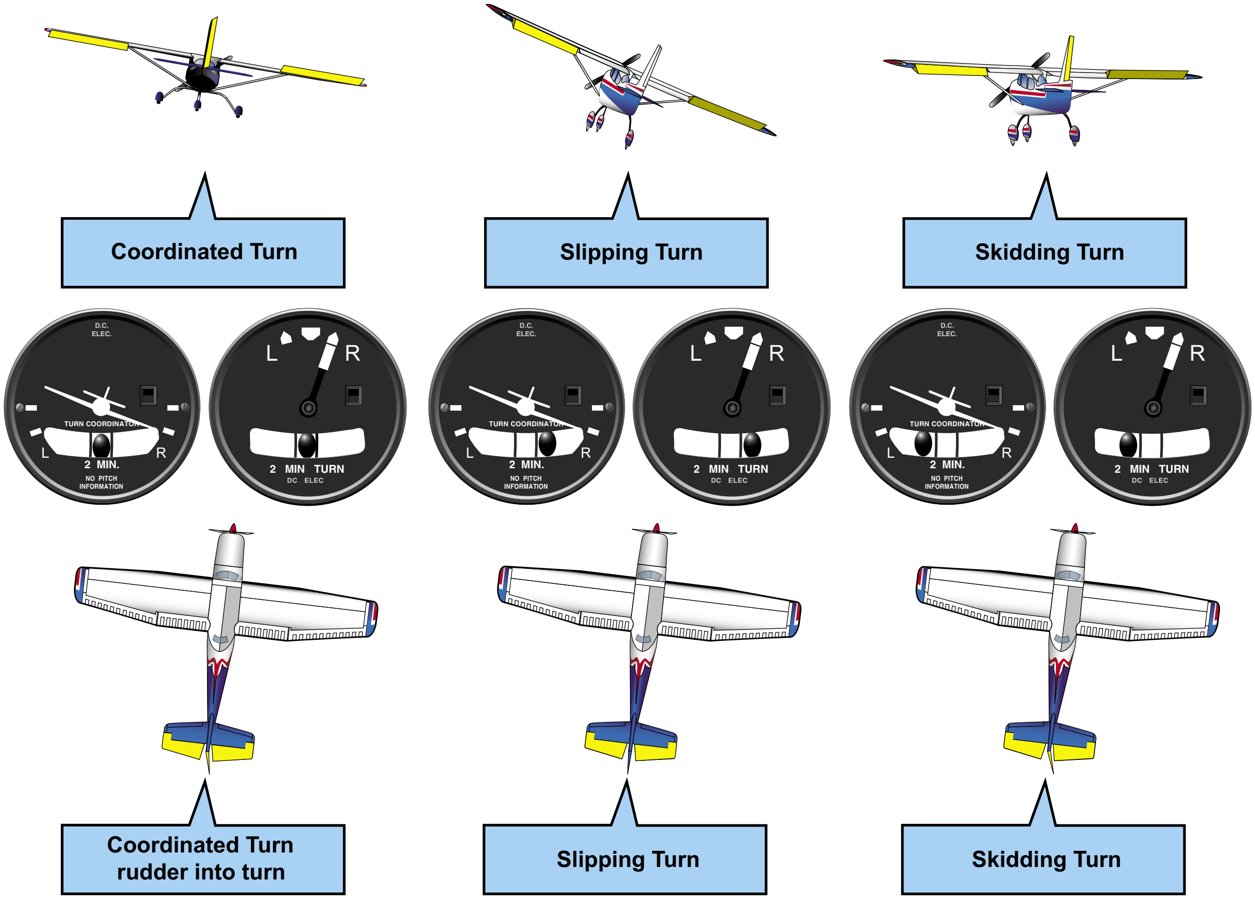
Coordinated Turn
Slipping Turn
Skidding Turn
D.C. ELEC.
TURN COORDINATOR
L
R
2 MIN.
NO PITCH INFORMATION
L
R
2 MIN TURN
DC ELEC
D.C. ELEC.
TURN COORDINATOR
L
R
2 MIN.
NO PITCH INFORMATION
L
R
2 MIN TURN
DC ELEC
D.C. ELEC.
TURN COORDINATOR
L
R
2 MIN.
NO PITCH INFORMATION
L
R
2 MIN TURN
DC ELEC
Coordinated Turn rudder into turn
Slipping Turn
Skidding Turn

Note the slight differences in rudder placement.

Figure 2-15. *Adverse Yaw.*

the aerodynamic force on the aircraft to the gross weight of the aircraft (e.g., lift/weight). For example, a load factor of 3 means the total load on an aircraft's structure is three times its gross weight. When designing an aircraft, it is necessary to determine the highest load factors that can be expected in normal operation under various operational situations. These "highest" load factors are called "limit load factors."

Aircraft are placed in various categories, i.e., normal, utility, and acrobatic, depending upon the load factors they are designed to take. For reasons of safety, the aircraft must be designed to withstand certain maximum load factors without any structural damage.

The specified load may be expected in terms of aerodynamic forces, as in turns. In level flight in undisturbed air, the wings are supporting not only the weight of the aircraft, but centrifugal force as well. As the bank steepens, the horizontal lift component increases, centrifugal force increases, and the load factor increases. If the load factor becomes so great that an increase in angle of attack cannot provide enough lift to support the load, the wing stalls. Since the stalling speed increases directly with the square root of the load factor, the pilot should be aware of the flight conditions during which the load factor can become critical. Steep turns at slow airspeed, structural ice accumulation, and vertical gusts in turbulent air can increase the load factor to a critical level.

Icing

One of the greatest hazards to flight is aircraft icing. The instrument pilot must be aware of the conditions conducive to aircraft icing. These conditions include the types of icing, the effects of icing on aircraft control and performance, effects of icing on aircraft systems, and the use and limitations of aircraft deice and anti-ice equipment. Coping with the hazards of icing begins with preflight planning to determine where icing may occur during a flight and ensuring the aircraft is free of ice and frost prior to takeoff. This attention to detail extends to managing deice and anti-ice systems properly during the flight, because weather conditions may change rapidly, and the pilot must be able to recognize when a change of flight plan is required.

Types of Icing

Structural Icing

Structural icing refers to the accumulation of ice on the exterior of the aircraft. Ice forms on aircraft structures and surfaces when super-cooled droplets impinge on them and freeze. Small and/or narrow objects are the best collectors of droplets and ice up most rapidly. This is why a small protuberance within sight of the pilot can be used as an "ice evidence probe." It is generally one of the first parts of the airplane on which an appreciable amount of ice forms. An aircraft's tailplane is a better collector than its wings, because the tailplane presents a thinner surface to the airstream.

Induction Icing

Ice in the induction system can reduce the amount of air available for combustion. The most common example of reciprocating engine induction icing is carburetor ice. Most pilots are familiar with this phenomenon, which occurs when moist air passes through a carburetor venturi and is cooled. As a result of this process, ice may form on the venturi walls and throttle plate, restricting airflow to the engine. This may occur at temperatures between 20° F (-7° C) and 70° F (21° C). The problem is remedied by applying carburetor heat, which uses the engine's own exhaust as a heat source to melt the ice or prevent its formation. On the other hand, fuel-injected aircraft engines usually are less vulnerable to icing but still can be affected if the engine's air source becomes blocked with ice. Manufacturers provide an alternate air source that may be selected in case the normal system malfunctions.

In turbojet aircraft, air that is drawn into the engines creates an area of reduced pressure at the inlet, which lowers the temperature below that of the surrounding air. In marginal icing conditions (i.e., conditions where icing is possible), this reduction in temperature may be sufficient to cause ice to form on the engine inlet, disrupting the airflow into the engine. Another hazard occurs when ice breaks off and is ingested into a running engine, which can cause damage to fan blades, engine compressor stall, or combustor flameout. When anti-icing systems are used, runback water also can refreeze on unprotected surfaces of the inlet and, if excessive, reduce airflow into the engine or distort the airflow pattern in such a manner as to cause compressor or fan blades to vibrate, possibly damaging the engine. Another problem in turbine engines is the icing of engine probes used to set power levels (for example, engine inlet temperature or engine pressure ratio (EPR) probes), which can lead to erroneous readings of engine instrumentation operational difficulties or total power loss.

The type of ice that forms can be classified as clear, rime, or mixed, based on the structure and appearance of the ice. The type of ice that forms varies depending on the atmospheric and flight conditions in which it forms. Significant structural icing on an aircraft can cause serious aircraft control and performance problems.

Clear Ice

A glossy, transparent ice formed by the relatively slow freezing of super cooled water is referred to as clear ice. *[Figure 2-16]* The terms "clear" and "glaze" have been used for essentially the same type of ice accretion. This type of ice is denser, harder, and sometimes more transparent than rime ice. With larger accretions, clear ice may form "horns." *[Figure 2-17]* Temperatures close to the freezing point, large amounts of liquid water, high aircraft velocities, and large droplets are conducive to the formation of clear ice.

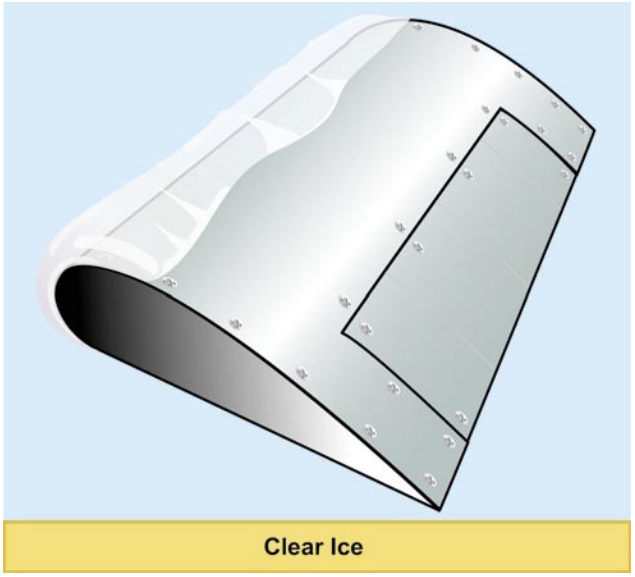

Figure 2-16. *Clear Ice.*

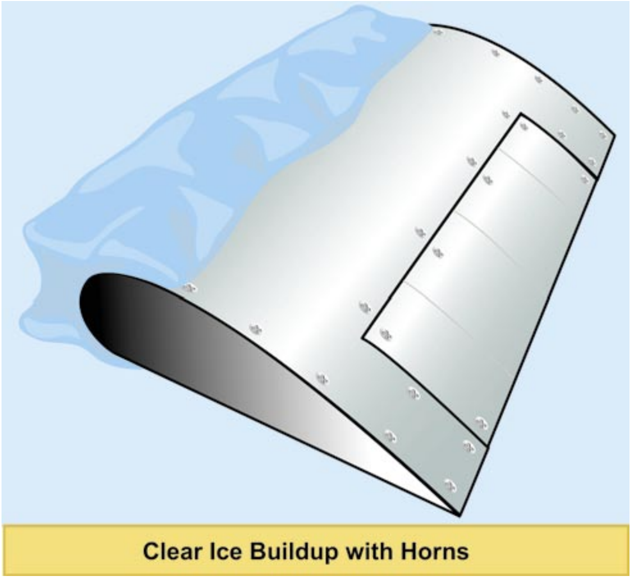
Clear Ice Buildup with Horns

Figure 2-17. *Clear Ice Buildup.*

Rime Ice

A rough, milky, opaque ice formed by the instantaneous or very rapid freezing of super cooled droplets as they strike the aircraft is known as rime ice. *[Figure 2-18]* The rapid freezing results in the formation of air pockets in the ice, giving it an opaque appearance and making it porous and brittle. For larger accretions, rime ice may form a streamlined extension of the wing. Low temperatures, lesser amounts of liquid water, low velocities, and small droplets are conducive to the formation of rime ice.

Figure 2-18. *Rime Ice.*

Mixed Ice

Mixed ice is a combination of clear and rime ice formed on the same surface. It is the shape and roughness of the ice that is most important from an aerodynamic point of view.

General Effects of Icing on Airfoils

The most hazardous aspect of structural icing is its aerodynamic effects. *[Figure 2-19]* Ice alters the shape of an airfoil, reducing the maximum coefficient of lift and angle of attack at which the aircraft stalls. Note that at very low angles of attack, there may be little or no effect of the ice on the coefficient of lift. Therefore, when cruising at a low angle of attack, ice on the wing may have little effect on the lift. However, note that the ice significantly reduces the $C_{L\text{-}MAX}$, and the angle of attack at which it occurs (the stall angle) is much lower. Thus, when slowing down and increasing the angle of attack for approach, the pilot may find that ice on the wing, which had little effect on lift in cruise now, causes stall to occur at a lower angle of attack and higher speed. Even a thin layer of ice at the leading edge of a wing, especially if it is rough, can have a significant effect in increasing stall speed. For large ice shapes, especially those with horns, the lift may also be reduced at a lower angle of attack. The accumulation of ice affects the coefficient of drag of the airfoil. *[Figure 2-19]* Note that the effect is significant even at very small angles of attack.

A significant reduction in $C_{L\text{-}MAX}$ and a reduction in the angle of attack where stall occurs can result from a relatively small ice accretion. A reduction of $C_{L\text{-}MAX}$ by 30 percent is not unusual, and a large horn ice accretion can result in reductions of 40 percent to 50 percent. Drag tends to increase steadily as ice accretes. An airfoil drag increase of 100 percent is not unusual, and for large horn ice accretions, the increase can be 200 percent or even higher.

Ice on an airfoil can have other effects not depicted in these curves. Even before airfoil stall, there can be changes in the pressure over the airfoil that may affect a control surface at the trailing edge. Furthermore, on takeoff, approach, and landing, the wings of many aircraft are multi-element airfoils with three or more elements. Ice may affect the different

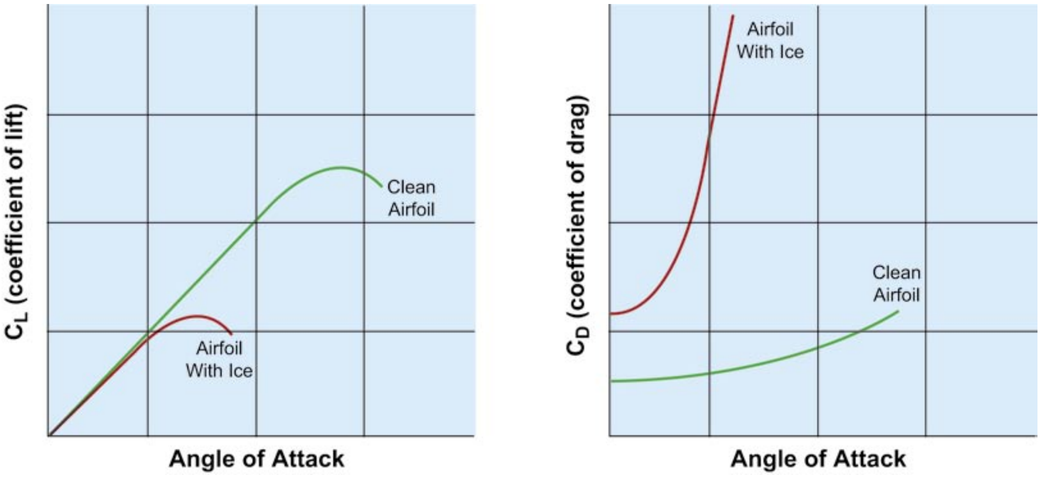

Figure 2-19. *Aerodynamic Effects of Icing.*

elements in different ways. Ice may also affect the way in which the air streams interact over the elements.

Ice can partially block or limit control surfaces, which limits or makes control movements ineffective. Also, if the extra weight caused by ice accumulation is too great, the aircraft may not be able to become airborne and, if in flight, the aircraft may not be able to maintain altitude. Therefore any accumulation of ice or frost should be removed before attempting flight.

Another hazard of structural icing is the possible uncommanded and uncontrolled roll phenomenon, referred to as roll upset, associated with severe in-flight icing. Pilots flying aircraft certificated for flight in known icing conditions should be aware that severe icing is a condition outside of the aircraft's certification icing envelope. Roll upset may be caused by airflow separation (aerodynamic stall), which induces self-deflection of the ailerons and loss of or degraded roll handling characteristics *[Figure 2-20]*. These phenomena can result from severe icing conditions without the usual symptoms of ice accumulation or a perceived aerodynamic stall.

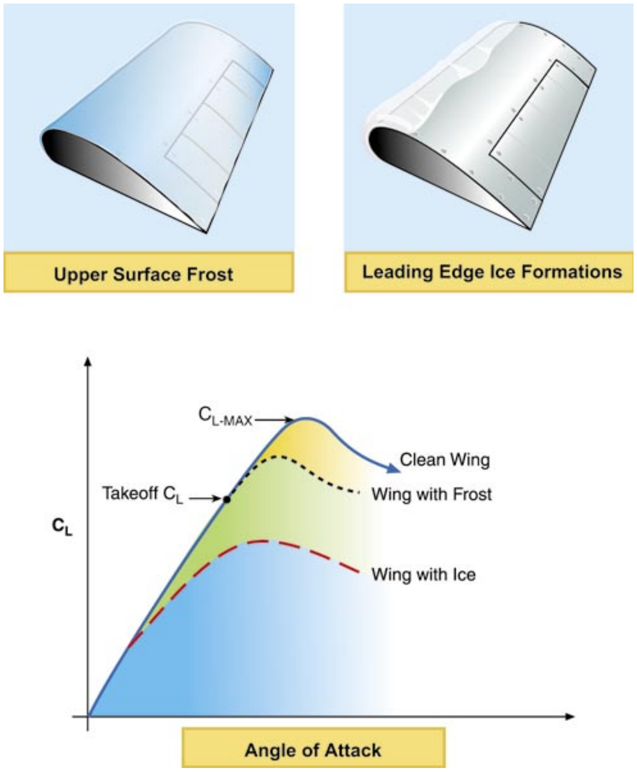

Figure 2-20. *Effect of Ice and Frost on Lift.*

Most aircraft have a nose-down pitching moment from the wings because the CG is ahead of the CP. It is the role of the tailplane to counteract this moment by providing a downward force. *[Figure 2-21]* The result of this configuration is that actions which move the wing away from stall, such as deployment of flaps or increasing speed, may increase the negative angle of attack of the tail. With ice on the tailplane, it may stall after full or partial deployment of flaps. *[Figure 2-22]*

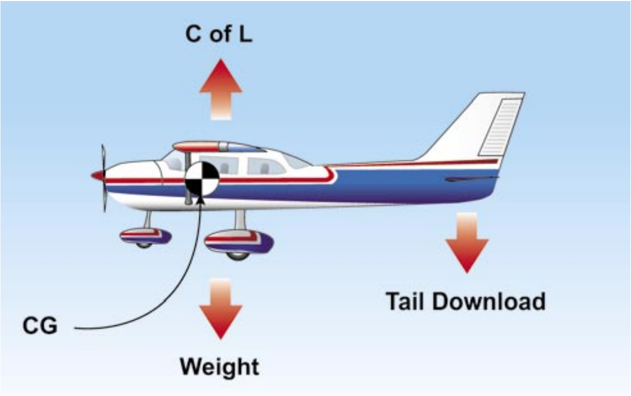

Figure 2-21. *Downward Force on the Tailplane.*

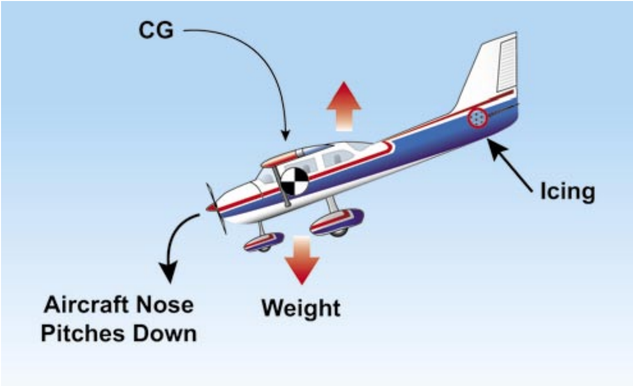

Figure 2-22. *Ice on the Tailplane.*

Since the tailplane is ordinarily thinner than the wing, it is a more efficient collector of ice. On most aircraft the tailplane is not visible to the pilot, who therefore cannot observe how well it has been cleared of ice by any deicing system. Thus, it is important that the pilot be alert to the possibility of tailplane stall, particularly on approach and landing.

Piper PA-34-200T (Des Moines, Iowa)

The pilot of this flight, which took place on January 9, 1996, said that upon crossing the runway threshold and lowering the flaps 25°, "the airplane pitched down." The pilot "immediately released the flaps and added power, but the airplane was basically uncontrollable at this point." The pilot reduced power and lowered the flaps before striking the runway on its centerline and sliding 1,000 feet before coming to a stop. The accident resulted in serious injury to the pilot, the sole occupant.

Examination of the wreckage revealed heavy impact damage to the airplane's forward fuselage, engines, and wings. Approximately one-half inch of rime ice was observed adhering to the leading edges of the left and right horizontal stabilizers and along the leading edge of the vertical stabilizer.

The National Transportation Safety Board (NTSB) determined the probable cause of the accident was the pilot's failure to use the airplane's deicing system, which resulted in an accumulation of empennage ice and a tailplane stall. Factors relating to this accident were the icing conditions and the pilot's intentional flight into those known conditions.

Tailplane Stall Symptoms

Any of the following symptoms, occurring singly or in combination, may be a warning of tailplane icing:

- Elevator control pulsing, oscillations, or vibrations;
- Abnormal nose-down trim change;
- Any other unusual or abnormal pitch anomalies (possibly resulting in pilot induced oscillations);
- Reduction or loss of elevator effectiveness;
- Sudden change in elevator force (control would move nose-down if unrestrained); and
- Sudden uncommanded nose-down pitch.

If any of the above symptoms occur, the pilot should:

- Immediately retract the flaps to the previous setting and apply appropriate nose-up elevator pressure;
- Increase airspeed appropriately for the reduced flap extension setting;
- Apply sufficient power for aircraft configuration and conditions. (High engine power settings may adversely impact response to tailplane stall conditions at high airspeed in some aircraft designs. Observe the manufacturer's recommendations regarding power settings.);
- Make nose-down pitch changes slowly, even in gusting conditions, if circumstances allow; and
- If a pneumatic deicing system is used, operate the system several times in an attempt to clear the tailplane of ice.

Once a tailplane stall is encountered, the stall condition tends to worsen with increased airspeed and possibly may worsen with increased power settings at the same flap setting. Airspeed, at any flap setting, in excess of the airplane manufacturer's recommendations, accompanied by uncleared ice contaminating the tailplane, may result in a tailplane stall and uncommanded pitch down from which recovery may not be possible. A tailplane stall may occur at speeds less than the maximum flap extended speed (VFE).

Propeller Icing

Ice buildup on propeller blades reduces thrust for the same aerodynamic reasons that wings tend to lose lift and increase drag when ice accumulates on them. The greatest quantity of ice normally collects on the spinner and inner radius of the propeller. Propeller areas on which ice may accumulate and be ingested into the engine normally are anti-iced rather than deiced to reduce the probability of ice being shed into the engine.

Effects of Icing on Critical Aircraft Systems

In addition to the hazards of structural and induction icing, the pilot must be aware of other aircraft systems susceptible to icing. The effects of icing do not produce the performance loss of structural icing or the power loss of induction icing but can present serious problems to the instrument pilot. Examples of such systems are flight instruments, stall warning systems, and windshields.

Flight Instruments

Various aircraft instruments including the airspeed indicator, altimeter, and rate-of-climb indicator utilize pressures sensed by pitot tubes and static ports for normal operation. When covered by ice these instruments display incorrect information thereby presenting serious hazard to instrument flight. Detailed information on the operation of these instruments and the specific effects of icing is presented in Chapter 3, Flight Instruments.

Stall Warning Systems

Stall warning systems provide essential information to pilots. These systems range from a sophisticated stall warning vane to a simple stall warning switch. Icing affects these systems in several ways resulting in possible loss of stall warning to the pilot. The loss of these systems can exacerbate an already hazardous situation. Even when an aircraft's stall warning system remains operational during icing conditions, it may be ineffective because the wing stalls at a lower angle of attack due to ice on the airfoil.

Windshields

Accumulation of ice on flight deck windows can severely restrict the pilot's visibility outside of the aircraft. Aircraft equipped for flight into known icing conditions typically have some form of windshield anti-icing to enable the pilot to see outside the aircraft in case icing is encountered in flight. One system consists of an electrically heated plate installed onto the airplane's windshield to give the pilot a narrow band of clear visibility. Another system uses a bar at the lower end of the windshield to spray deicing fluid onto it and prevent ice from forming. On high performance aircraft that require complex windshields to protect against bird strikes and withstand pressurization loads, the heating element often is a layer of conductive film or thin wire strands through which electric current is run to heat the windshield and prevent ice from forming.

Antenna Icing

Because of their small size and shape, antennas that do not lay flush with the aircraft's skin tend to accumulate ice rapidly. Furthermore, they often are devoid of internal anti-icing or deicing capability for protection. During flight in icing conditions, ice accumulations on an antenna may cause it to begin to vibrate or cause radio signals to become distorted and it may cause damage to the antenna. If a frozen antenna breaks off, it can damage other areas of the aircraft in addition to causing a communication or navigation system failure.

Summary

Ice-contaminated aircraft have been involved in many accidents. Takeoff accidents have usually been due to failure to deice or anti-ice critical surfaces properly on the ground. Proper deicing and anti-icing procedures are addressed in two other pilot guides, Advisory Circular (AC) 120-58, Pilot Guide: Large Aircraft Ground Deicing and AC 135-17, Pilot Guide: Small Aircraft Ground Deicing.

The pilot of an aircraft, which is not certificated or equipped for flight in icing conditions, should avoid all icing conditions. The aforementioned guides provide direction on how to do this, and on how to exit icing conditions promptly and safely should they be inadvertently encountered.

The pilot of an aircraft, which is certificated for flight in icing conditions can safely operate in the conditions for which the aircraft was evaluated during the certification process but should never become complacent about icing. Even short encounters with small amounts of rough icing can be very hazardous. The pilot should be familiar with all information in the Aircraft Flight Manual (AFM) or Pilot's Operating Handbook (POH) concerning flight in icing conditions and follow it carefully. Of particular importance are proper operation of ice protection systems and any airspeed minimums to be observed during or after flight in icing conditions. There are some icing conditions for which no aircraft is evaluated in the certification process, such as super-cooled large drops (SLD). These subfreezing water droplets, with diameters greater than 50 microns, occur within or below clouds and sustained flight in these conditions can be very hazardous. The pilot should be familiar with any information in the AFM or POH relating to these conditions, including aircraft-specific cues for recognizing these hazardous conditions within clouds.

The information in this chapter is an overview of the hazards of aircraft icing. For more detailed information refer to AC 91-74, Pilot Guide: Flight in Icing Conditions, AC 91-51A, Effect of Icing on Aircraft Control and Airplane Deice and Anti-Ice Systems, AC 20-73A, Aircraft Ice Protection and AC 23.143-1, Ice Contaminated Tailplane Stall (ICTS).

HDG
NAV
N
3
6
E
12
15
S
21
W
30
33

Chapter 3

Flight Instruments

Introduction

Aircraft became a practical means of transportation when accurate flight instruments freed the pilot from the necessity of maintaining visual contact with the ground. Flight instruments are crucial to conducting safe flight operations and it is important that the pilot have a basic understanding of their operation. The basic flight instruments required for operation under visual flight rules (VFR) are airspeed indicator (ASI), altimeter, and magnetic direction indicator. In addition to these, operation under instrument flight rules (IFR) requires a gyroscopic rate-of-turn indicator, slip-skid indicator, sensitive altimeter adjustable for barometric pressure, clock displaying hours, minutes, and seconds with a sweep-second pointer or digital presentation, gyroscopic pitch-and-bank indicator (artificial horizon), and gyroscopic direction indicator (directional gyro or equivalent).

Aircraft that are flown in instrument meteorological conditions (IMC) are equipped with instruments that provide attitude and direction reference, as well as navigation instruments that allow precision flight from takeoff to landing with limited or no outside visual reference.

The instruments discussed in this chapter are those required by Title 14 of the Code of Federal Regulations (14 CFR) part 91, and are organized into three groups: pitot-static instruments, compass systems, and gyroscopic instruments. The chapter concludes with a discussion of how to preflight these systems for IFR flight. This chapter addresses additional avionics systems such as Electronic Flight Information Systems (EFIS), Ground Proximity Warning System (GPWS), Terrain Awareness and Warning System (TAWS), Traffic Alert and Collision Avoidance System (TCAS), Head Up Display (HUD), etc., that are increasingly being incorporated into general aviation aircraft.

Pitot/Static Systems

Pitot pressure, or impact air pressure, is sensed through an open-end tube pointed directly into the relative wind flowing around the aircraft. The pitot tube connects to pressure operated flight instruments such as the ASI.

Static Pressure

Other instruments depend upon accurate sampling of the ambient still air atmospheric pressure to determine the height and speed of movement of the aircraft through the air, both horizontally and vertically. This pressure, called static pressure, is sampled at one or more locations outside the aircraft. The pressure of the static air is sensed at a flush port where the air is not disturbed. On some aircraft, air is sampled by static ports on the side of the electrically heated pitot-static head. *[Figure 3-1]* Other aircraft pick up the static pressure through flush ports on the side of the fuselage or the vertical fin. These ports are in locations proven by flight tests to be in undisturbed air, and they are normally paired, one on either side of the aircraft. This dual location prevents lateral movement of the aircraft from giving erroneous static pressure indications. The areas around the static ports may be heated with electric heater elements to prevent ice forming over the port and blocking the entry of the static air.

Three basic pressure-operated instruments are found in most aircraft instrument panels. These are the sensitive altimeter, ASI, and vertical speed indicator (VSI). All three receive pressures sensed by the aircraft pitot-static system. The static ports supply pressure to the ASI, altimeter, and VSI.

Blockage Considerations

The pitot tube is particularly sensitive to blockage especially by icing. Even light icing can block the entry hole of the pitot tube where ram air enters the system. This affects the ASI and is the reason most airplanes are equipped with a pitot heating system.

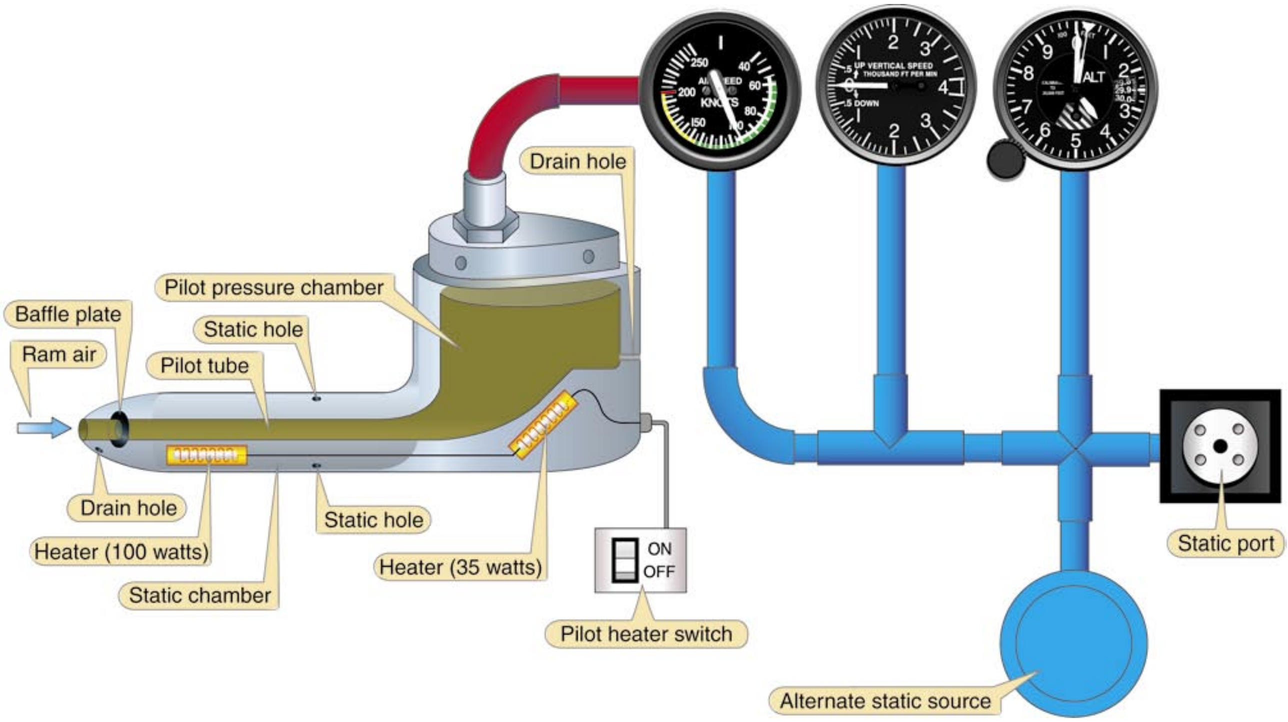

Figure 3-1. *A Typical Electrically Heated Pitot-Static Head.*

Indications of Pitot Tube Blockage

If the pitot tube becomes blocked, the ASI displays inaccurate speeds. At the altitude where the pitot tube becomes blocked, the ASI remains at the existing airspeed and doesn't reflect actual changes in speed.

- At altitudes above where the pitot tube became blocked, the ASI displays a higher-than-actual airspeed increasing steadily as altitude increases.
- At lower altitudes, the ASI displays a lower-than-actual airspeed decreasing steadily as altitude decreases.

Indications from Static Port Blockage

Many aircraft also have a heating system to protect the static ports to ensure the entire pitot-static system is clear of ice. If the static ports become blocked, the ASI would still function but could produce inaccurate indications. At the altitude where the blockage occurs, airspeed indications would be normal.

- At altitudes above which the static ports became blocked, the ASI displays a lower-than-actual airspeed continually decreasing as altitude is increased.
- At lower altitudes, the ASI displays a higher-than-actual airspeed increasing steadily as altitude decreases.

The trapped pressure in the static system causes the altimeter to remain at the altitude where the blockage occurred. The VSI remains at zero. On some aircraft, an alternate static air source valve is used for emergencies. *[Figure 3-2]* If the alternate source is vented inside the airplane, where static pressure is usually lower than outside static pressure, selection of the alternate source may result in the following erroneous instrument indications:

1. Altimeter reads higher than normal,
2. Indicated airspeed (IAS) reads greater than normal, and
3. VSI momentarily shows a climb. Consult the Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM) to determine the amount of error.

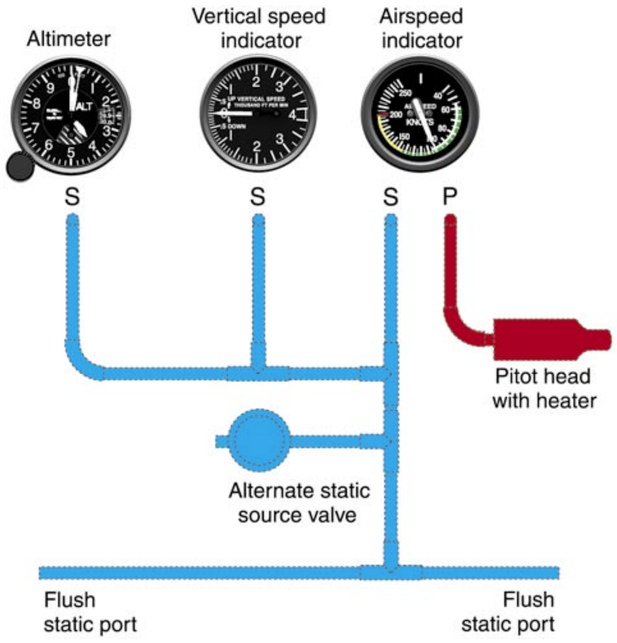

Figure 3-2. *A Typical Pitot-Static System.*

Effects of Flight Conditions

The static ports are located in a position where the air at their surface is as undisturbed as possible. But under some flight conditions, particularly at a high angle of attack with the landing gear and flaps down, the air around the static port may be disturbed to the extent that it can cause an error in the indication of the altimeter and ASI. Because of the importance of accuracy in these instruments, part of the certification tests for an aircraft is a check of position error in the static system.

The POH/AFM contains any corrections that must be applied to the airspeed for the various configurations of flaps and landing gear.

Pitot/Static Instruments

Sensitive Altimeter

A sensitive altimeter is an aneroid barometer that measures the absolute pressure of the ambient air and displays it in terms of feet or meters above a selected pressure level.

Principle of Operation

The sensitive element in a sensitive altimeter is a stack of evacuated, corrugated bronze aneroid capsules. *[Figure 3-3]* The air pressure acting on these aneroids tries to compress them against their natural springiness, which tries to expand them. The result is that their thickness changes as the air pressure changes. Stacking several aneroids increases the dimension change as the pressure varies over the usable range of the instrument.

Below 10,000 feet, a striped segment is visible. Above this altitude, a mask begins to cover it, and above 15,000 feet, all of the stripes are covered. *[Figure 3-4]*

Another configuration of the altimeter is the drum-type. *[Figure 3-5]* These instruments have only one pointer that makes one revolution for every 1,000 feet. Each number represents 100 feet and each mark represents 20 feet. A drum, marked in thousands of feet, is geared to the mechanism that drives the pointer. To read this type of altimeter, first look at

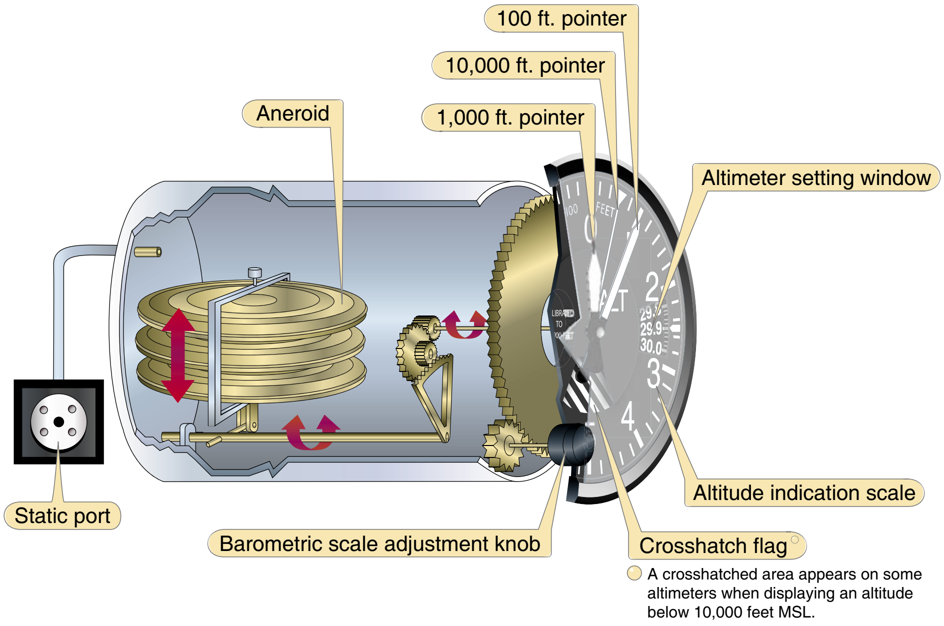
100 ft. pointer
10,000 ft. pointer
1,000 ft. pointer
Aneroid
Altimeter setting window
Static port
Barometric scale adjustment knob
Altitude indication scale
Crosshatch flag
A crosshatched area appears on some altimeters when displaying an altitude below 10,000 feet MSL.

Figure 3-3. *Sensitive Altimeter Components.*

the drum to get the thousands of feet, and then at the pointer to get the feet and hundreds of feet.

A sensitive altimeter is one with an adjustable barometric scale allowing the pilot to set the reference pressure from which the altitude is measured. This scale is visible in a small window called the Kollsman window. A knob on the instrument adjusts the scale. The range of the scale is from 28.00" to 31.00" inches of mercury (Hg), or 948 to 1,050 millibars.

Rotating the knob changes both the barometric scale and the altimeter pointers in such a way that a change in the barometric scale of 1" Hg changes the pointer indication by 1,000 feet. This is the standard pressure lapse rate below 5,000 feet. When the barometric scale is adjusted to 29.92" Hg or 1,013.2 millibars, the pointers indicate the pressure altitude. The pilot displays indicate altitude by adjusting the barometric scale to the local altimeter setting. The altimeter then indicates the height above the existing sea level pressure.

Altimeter Errors

A sensitive altimeter is designed to indicate standard changes from standard conditions, but most flying involves errors caused by nonstandard conditions and the pilot must be able to modify the indications to correct for these errors. There are two types of errors: mechanical and inherent.

Mechanical

A preflight check to determine the condition of an altimeter consists of setting the barometric scale to the local altimeter setting. The altimeter should indicate the surveyed elevation

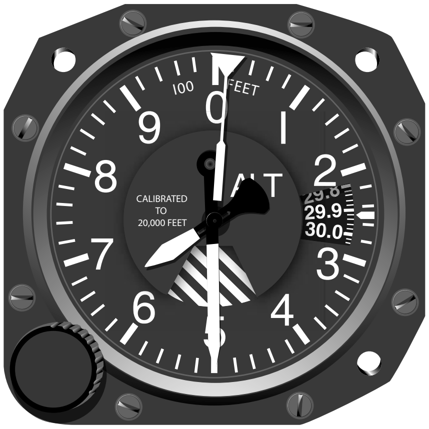
100
FEET
ALT
CALIBRATED TO 20,000 FEET
29.8
29.9
30.0

Figure 3-4. *Three-Pointer Altimeter.*

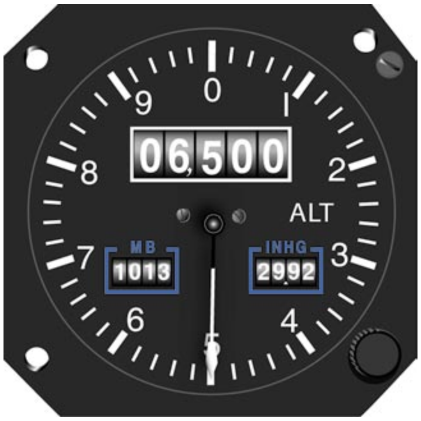
06500
ALT
MB
1013
INHG
2992

Figure 3-5. *Drum-Type Altimeter.*

of the airport. If the indication is off by more than 75 feet from the surveyed elevation, the instrument should be referred to a certificated instrument repair station for recalibration. Differences between ambient temperature and/or pressure causes an erroneous indication on the altimeter.

Inherent Altimeter Error

When the aircraft is flying in air that is warmer than standard, the air is less dense and the pressure levels are farther apart. When the aircraft is flying at an indicated altitude of 5,000 feet, the pressure level for that altitude is higher than it would be in air at standard temperature, and the aircraft is higher than it would be if the air were cooler. If the air is colder than standard, it is denser and the pressure levels are closer together. When the aircraft is flying at an indicated altitude of 5,000 feet, its true altitude is lower than it would be if the air were warmer. *[Figure 3-6]*

Cold Weather Altimeter Errors

A correctly calibrated pressure altimeter indicates true altitude above mean sea level (MSL) when operating within the International Standard Atmosphere (ISA) parameters of pressure and temperature. Nonstandard pressure conditions are corrected by applying the correct local area altimeter setting.

Temperature errors from ISA result in true altitude being higher than indicated altitude whenever the temperature is warmer than ISA and true altitude being lower than indicated altitude whenever the temperature is colder than ISA. True altitude variance under conditions of colder than ISA temperatures poses the risk of inadequate obstacle clearance. Under extremely cold conditions, pilots may need to add an appropriate temperature correction determined from the chart in *Figure 3-7* to charted IFR altitudes to ensure terrain and obstacle clearance with the following restrictions:

- Altitudes specifically assigned by Air Traffic Control (ATC), such as "maintain 5,000 feet" shall not be corrected. Assigned altitudes may be rejected if the pilot decides that low temperatures pose a risk of inadequate terrain or obstacle clearance.
- If temperature corrections are applied to charted IFR altitudes (such as procedure turn altitudes, final approach fix crossing altitudes, etc.), the pilot must advise ATC of the applied correction.

ICAO Cold Temperature Error Table

The cold temperature induced altimeter error may be significant when considering obstacle clearances when temperatures are well below standard. Pilots may wish to increase their minimum terrain clearance altitudes with a corresponding increase in ceiling from the normal minimum when flying in extreme cold temperature conditions. Higher altitudes may need to be selected when flying at low terrain clearances. Most flight management systems (FMS) with air data computers implement a capability to compensate for cold temperature errors. Pilots flying with these systems should ensure they are aware of the conditions under which the system will automatically compensate. If compensation is applied by the FMS or manually, ATC must be informed that the aircraft is not flying the assigned altitude. Otherwise, vertical separation from other aircraft may be reduced creating a potentially hazardous situation. The table in *Figure 3-7*, derived from International Civil Aviation

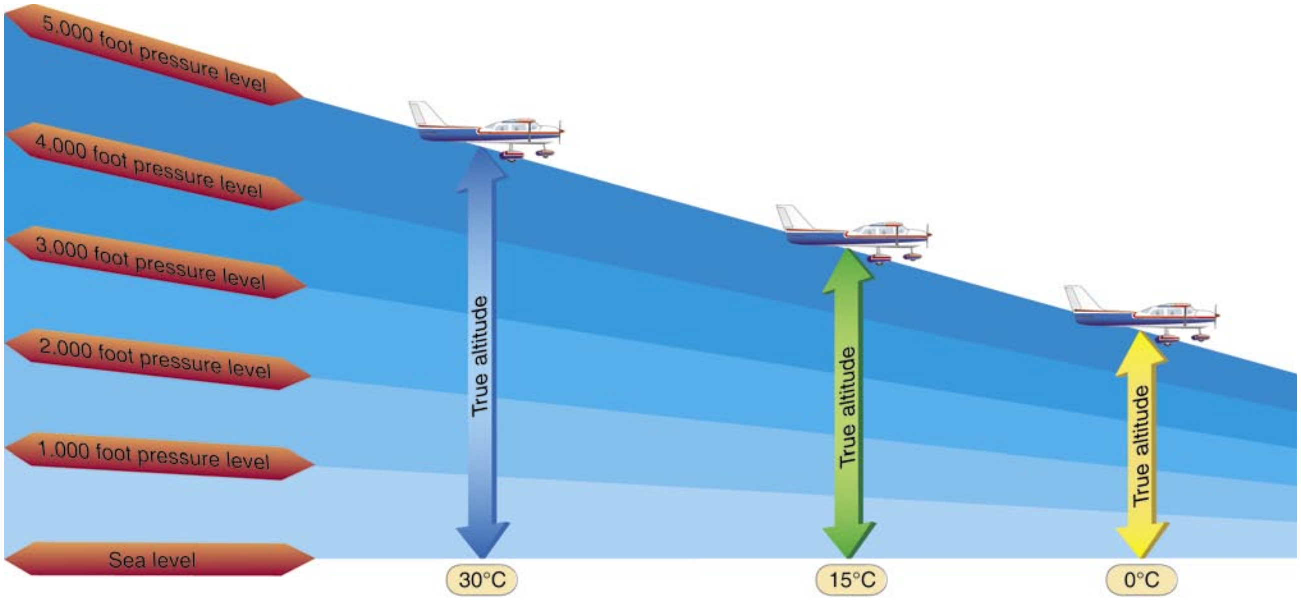
5,000 foot pressure level
4,000 foot pressure level
3,000 foot pressure level
2,000 foot pressure level
1,000 foot pressure level
Sea level
True altitude
True altitude
True altitude
30°C
15°C
0°C

Figure 3-6. *The loss of altitude experienced when flying into an area where the air is colder (more dense) than standard.*

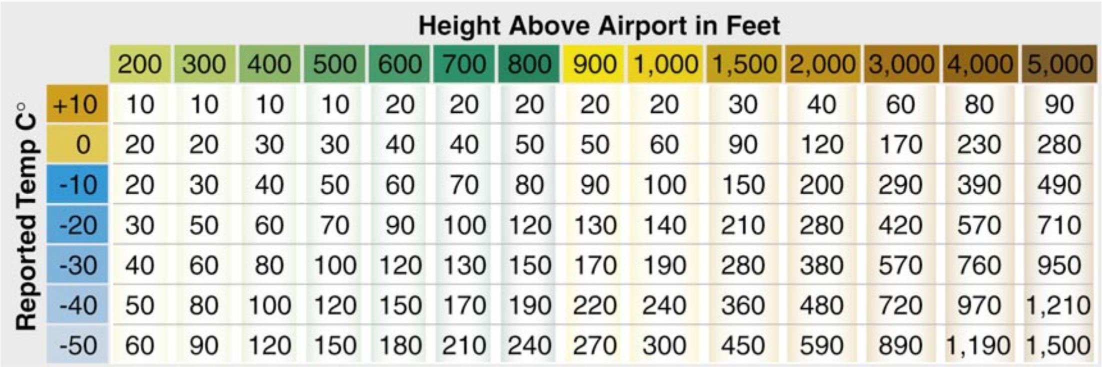

Reported Temp C°	Height Above Airport in Feet													
	200	300	400	500	600	700	800	900	1,000	1,500	2,000	3,000	4,000	5,000
+10	10	10	10	10	20	20	20	20	20	30	40	60	80	90
0	20	20	30	30	40	40	50	50	60	90	120	170	230	280
-10	20	30	40	50	60	70	80	90	100	150	200	290	390	490
-20	30	50	60	70	90	100	120	130	140	210	280	420	570	710
-30	40	60	80	100	120	130	150	170	190	280	380	570	760	950
-40	50	80	100	120	150	170	190	220	240	360	480	720	970	1,210
-50	60	90	120	150	180	210	240	270	300	450	590	890	1,190	1,500

Figure 3-7. *ICAO Cold Temperature Error Table.*

Organization (ICAO) standard formulas, shows how much error can exist when the temperature is extremely cold. To use the table, find the reported temperature in the left column, and then read across the top row to the height above the airport/reporting station. Subtract the airport elevation from the altitude of the final approach fix (FAF). The intersection of the column and row is the amount of possible error.

Example: The reported temperature is -10° Celsius and the FAF is 500 feet above the airport elevation. The reported current altimeter setting may place the aircraft as much as 50 feet below the altitude indicated by the altimeter.

When using the cold temperature error table, the altitude error is proportional to both the height above the reporting station elevation and the temperature at the reporting station. For IFR approach procedures, the reporting station elevation is assumed to be airport elevation. It is important to understand that corrections are based upon the temperature at the reporting station, not the temperature observed at the aircraft's current altitude and height above the reporting station and not the charted IFR altitude.

To see how corrections are applied, note the following example:

Airport Elevation	496 feet
Airport Temperature	- 50° C

A charted IFR approach to the airport provides the following data:

Minimum Procedure Turn Altitude	1,800 feet
Minimum FAF Crossing Altitude	1,200 feet
Straight-in Minimum Descent Altitude	800 feet
Circling MDA	1,000 feet

The Minimum Procedure Turn Altitude of 1,800 feet will be used as an example to demonstrate determination of the appropriate temperature correction. Typically, altitude values are rounded up to the nearest 100-foot level. The charted procedure turn altitude of 1,800 feet minus the airport elevation of 500 feet equals 1,300 feet. The altitude difference of 1,300 feet falls between the correction chart elevations of 1,000 feet and 1,500 feet. At the station temperature of -50°C, the correction falls between 300 feet and 450 feet. Dividing the difference in compensation values by the difference in altitude above the airport gives the error value per foot.

In this case, 150 feet divided by 500 feet = 0.33 feet for each additional foot of altitude above 1,000 feet. This provides a correction of 300 feet for the first 1,000 feet and an additional value of 0.33 times 300 feet, or 99 feet, which is rounded to 100 feet. 300 feet + 100 feet = total temperature correction of 400 feet. For the given conditions, correcting the charted value of 1,800 feet above MSL (equal to a height above the reporting station of 1,300 feet) requires the addition of 400 feet. Thus, when flying at an indicated altitude of 2,200 feet, the aircraft is actually flying a true altitude of 1,800 feet.

Minimum Procedure Turn Altitude
1,800 feet charted = 2,200 feet corrected
Minimum FAF Crossing Altitude
1,200 feet charted = 1,500 feet corrected
Straight-in MDA
800 feet charted = 900 feet corrected
Circling MDA
1,000 feet charted = 1,200 feet corrected

Nonstandard Pressure on an Altimeter

Maintaining a current altimeter setting is critical because the atmosphere pressure is not constant. That is, in one location the pressure might be higher than the pressure just a short distance away. Take an aircraft whose altimeter setting is set to 29.92" of local pressure. As the aircraft moves to an area of lower pressure (Point A to B in *Figure 3-8*) and the pilot fails to readjust the altimeter setting (essentially calibrating it to local pressure), then as the pressure decreases, the true altitude will be lower. Adjusting the altimeter settings compensates for this. When the altimeter shows an indicated altitude of 5,000 feet, the true altitude at Point A (the height above mean sea level) is only 3,500 feet at Point B. The fact that the altitude indication is not always true lends itself to the memory aid, “When flying from hot to cold or from a high to a low, look out below.” *[Figure 3-8]*

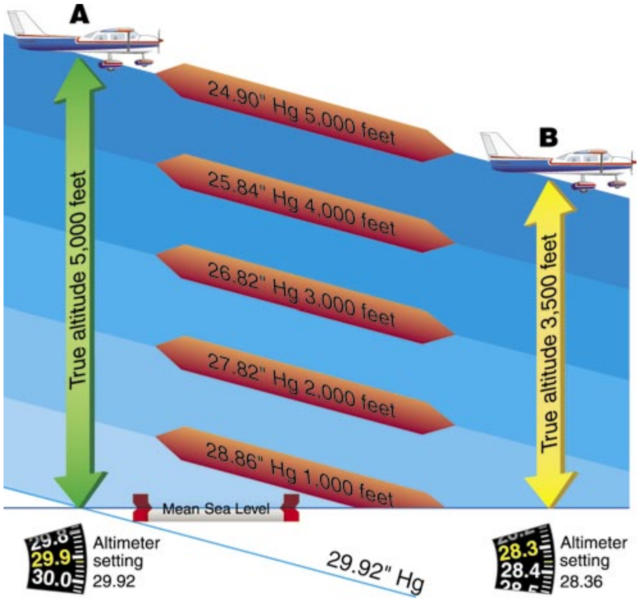

Figure 3-8. *Effects of Nonstandard Pressure on an Altimeter of an Aircraft Flown into Air of Lower Than Standard Pressure (Air is Less Dense).*

Altimeter Enhancements (Encoding)

It is not sufficient in the airspace system for only the pilot to have an indication of the aircraft’s altitude; the air traffic controller on the ground must also know the altitude of the aircraft. To provide this information, the aircraft is typically equipped with an encoding altimeter.

When the ATC transponder is set to Mode C, the encoding altimeter supplies the transponder with a series of pulses identifying the flight level (in increments of 100 feet) at which the aircraft is flying. This series of pulses is transmitted to the ground radar where they appear on the controller’s scope as an alphanumeric display around the return for the aircraft. The transponder allows the ground controller to identify the aircraft and determine the pressure altitude at which it is flying.

A computer inside the encoding altimeter measures the pressure referenced from 29.92" Hg and delivers this data to the transponder. When the pilot adjusts the barometric scale to the local altimeter setting, the data sent to the transponder is not affected. This is to ensure that all Mode C aircraft are transmitting data referenced to a common pressure level. ATC equipment adjusts the displayed altitudes to compensate for local pressure differences allowing display of targets at correct altitudes. 14 CFR part 91 requires the altitude transmitted by the transponder to be within 125 feet of the altitude indicated on the instrument used to maintain flight altitude.

Reduced Vertical Separation Minimum (RVSM)

Below 31,000 feet, a 1,000 foot separation is the minimum required between usable flight levels. Flight levels (FLs) generally start at 18,000 feet where the local pressure is 29.92" Hg or greater. All aircraft 18,000 feet and above use a standard altimeter setting of 29.92" Hg, and the altitudes are in reference to a standard hence termed FL. Between FL 180 and FL 290, the minimum altitude separation is 1,000 feet between aircraft. However, for flight above FL 290 (primarily due to aircraft equipage and reporting capability; potential error) ATC applied the requirement of 2,000 feet of separation. FL 290, an altitude appropriate for an eastbound aircraft, would be followed by FL 310 for a westbound aircraft, and so on to FL 410, or seven FLs available for flight. With 1,000-foot separation, or a reduction of the vertical separation between FL 290 and FL 410, an additional six FLs become available. This results in normal flight level and direction management being maintained from FL 180 through FL 410. Hence the name is Reduced Vertical Separation Minimum (RVSM). Because it is applied domestically, it is called United States Domestic Reduced Vertical Separation Minimum, or DRVSM.

However, there is a cost to participate in the DRVSM program which relates to both aircraft equipage and pilot training. For example, altimetry error must be reduced significantly and operators using RVSM must receive authorization from the appropriate civil aviation authority. RVSM aircraft must meet required altitude-keeping performance standards. Additionally, operators must operate in accordance with RVSM policies/procedures applicable to the airspace where they are flying.

The aircraft must be equipped with at least one automatic altitude control—

- Within a tolerance band of ±65 feet about an acquired altitude when the aircraft is operated in straight-and-level flight.
- Within a tolerance band of ±130 feet under no turbulent, conditions for aircraft for which application for type certification occurred on or before April 9, 1997 that are equipped with an automatic altitude control system with flight management/performance system inputs.

That aircraft must be equipped with an altitude alert system that signals an alert when the altitude displayed to the flight crew deviates from the selected altitude by more than (in most cases) 200 feet. For each condition in the full RVSM flight envelope, the largest combined absolute value for residual static source error plus the avionics error may not exceed 200 feet. Aircraft with TCAS must have compatibility with RVSM Operations. *Figure 3-9* illustrates the increase in aircraft permitted between FL 180 and FL 410. Most noteworthy, however, is the economization that aircraft can take advantage of by the higher FLs being available to more aircraft.

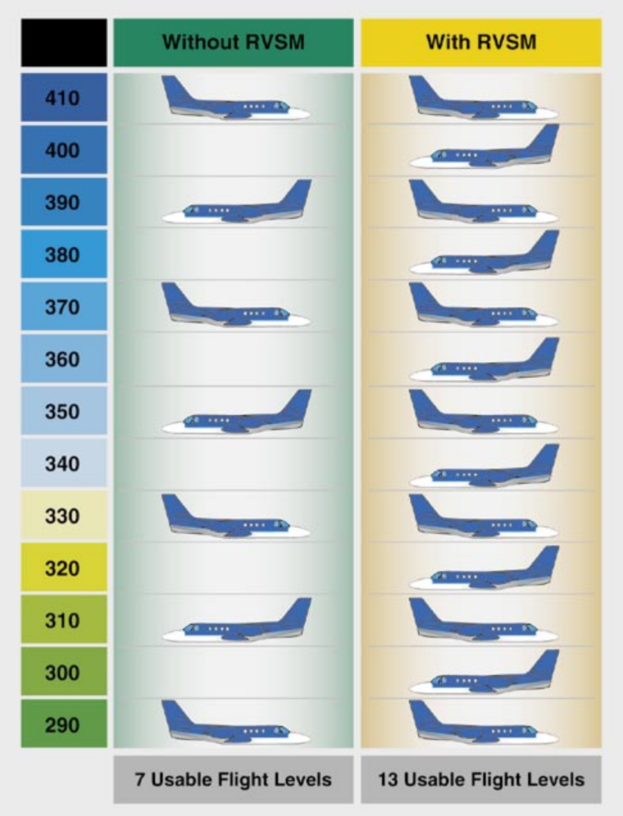

Figure 3-9. *Increase in Aircraft Permitted Between FL 180 and FL 410.*

Vertical Speed Indicator (VSI)

The VSI in *Figure 3-10* is also called a vertical velocity indicator (VVI), and was formerly known as a rate-of-climb indicator. It is a rate-of-pressure change instrument that gives an indication of any deviation from a constant pressure level.

Inside the instrument case is an aneroid very much like the one in an ASI. Both the inside of this aneroid and the inside of the instrument case are vented to the static system, but the case is vented through a calibrated orifice that causes the pressure inside the case to change more slowly than the pressure inside the aneroid. As the aircraft ascends, the static pressure becomes lower. The pressure inside the case compresses the aneroid, moving the pointer upward, showing a climb and indicating the rate of ascent in number of feet per minute (fpm).

When the aircraft levels off, the pressure no longer changes. The pressure inside the case becomes equal to that inside the aneroid, and the pointer returns to its horizontal, or zero, position. When the aircraft descends, the static pressure increases. The aneroid expands, moving the pointer downward, indicating a descent.

The pointer indication in a VSI lags a few seconds behind the actual change in pressure. However, it is more sensitive than an altimeter and is useful in alerting the pilot of an upward or downward trend, thereby helping maintain a constant altitude.

Some of the more complex VSIs, called instantaneous vertical speed indicators (IVSI), have two accelerometer-actuated air pumps that sense an upward or downward pitch of the aircraft and instantaneously create a pressure differential. By the time the pressure caused by the pitch acceleration dissipates, the altitude pressure change is effective.

Dynamic Pressure Type Instruments

Airspeed Indicator (ASI)

An ASI is a differential pressure gauge that measures the dynamic pressure of the air through which the aircraft is flying. Dynamic pressure is the difference in the ambient static air pressure and the total, or ram, pressure caused by the motion of the aircraft through the air. These two pressures are taken from the pitot-static system.

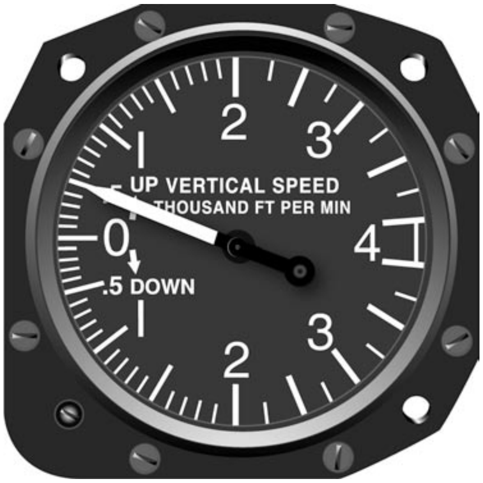

Figure 3-10. *Rate of Climb or Descent in Thousands of Feet Per Minute.*

The mechanism of the ASI in *Figure 3-11* consists of a thin, corrugated phosphor bronze aneroid, or diaphragm, that receives its pressure from the pitot tube. The instrument case is sealed and connected to the static ports. As the pitot pressure increases or the static pressure decreases, the diaphragm expands. This dimensional change is measured by a rocking shaft and a set of gears that drives a pointer across the instrument dial. Most ASIs are calibrated in knots, or nautical miles per hour; some instruments show statute miles per hour, and some instruments show both.

Types of Airspeed

Just as there are several types of altitude, there are multiple types of airspeed: Indicated Airspeed (IAS), Calibrated Airspeed (CAS), Equivalent Airspeed (EAS), and True Airspeed (TAS).

Indicated Airspeed (IAS)

IAS is shown on the dial of the instrument, uncorrected for instrument or system errors.

Calibrated Airspeed (CAS)

CAS is the speed at which the aircraft is moving through the air, which is found by correcting IAS for instrument and position errors. The POH/AFM has a chart or graph to correct IAS for these errors and provide the correct CAS for the various flap and landing gear configurations.

Equivalent Airspeed (EAS)

EAS is CAS corrected for compression of the air inside the pitot tube. EAS is the same as CAS in standard atmosphere at sea level. As the airspeed and pressure altitude increase, the CAS becomes higher than it should be, and a correction for compression must be subtracted from the CAS.

True Airspeed (TAS)

TAS is CAS corrected for nonstandard pressure and temperature. TAS and CAS are the same in standard atmosphere at sea level. Under nonstandard conditions, TAS is found by applying a correction for pressure altitude and temperature to the CAS.

Some aircraft are equipped with true ASIs that have a temperature-compensated aneroid bellows inside the instrument case. This bellows modifies the movement of the rocking shaft inside the instrument case so the pointer shows the actual TAS.

The TAS indicator provides both true and IAS. These instruments have the conventional airspeed mechanism, with an added subdial visible through cutouts in the regular dial. A knob on the instrument allows the pilot to rotate the subdial and align an indication of the outside air temperature with the pressure altitude being flown. This alignment causes the instrument pointer to indicate the TAS on the subdial. *[Figure 3-12]*

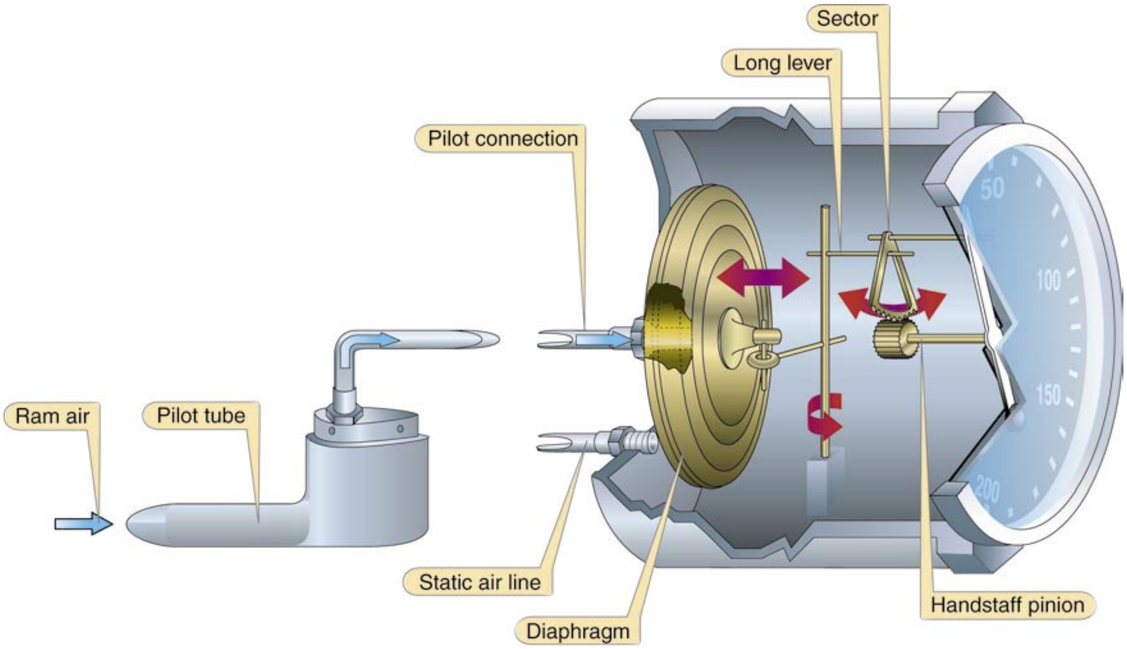

Figure 3-11. *Mechanism of an Airspeed Indicator.*

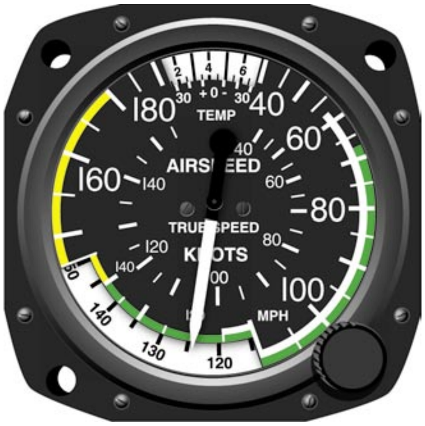

Figure 3-12. *A true airspeed indicator allows the pilot to correct IAS for nonstandard temperature and pressure.*

Mach Number

As an aircraft approaches the speed of sound, the air flowing over certain areas of its surface speeds up until it reaches the speed of sound, and shock waves form. The IAS at which these conditions occur changes with temperature. Therefore, in this case, airspeed is not entirely adequate to warn the pilot of the impending problems. Mach number is more useful. Mach number is the ratio of the TAS of the aircraft to the speed of sound in the same atmospheric conditions. An aircraft flying at the speed of sound is flying at Mach 1.0. Some older mechanical Machmeters not driven from an air data computer use an altitude aneroid inside the instrument that converts pitot-static pressure into Mach number. These systems assume that the temperature at any altitude is standard; therefore, the indicated Mach number is inaccurate whenever the temperature deviates from standard. These systems are called indicated Machmeters. Modern electronic Machmeters use information from an air data computer system to correct for temperature errors. These systems display true Mach number.

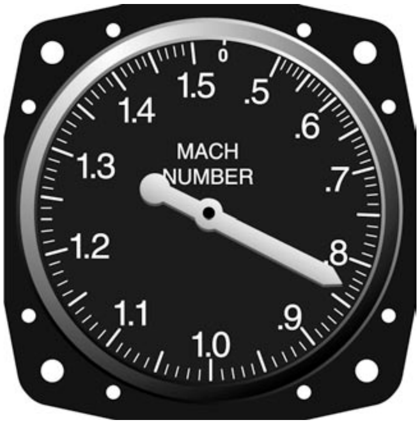

Figure 3-13. *A Machmeter shows the ratio of the speed of sound to the TAS the aircraft is flying.*

Most high-speed aircraft are limited to a maximum Mach number at which they can fly. This is shown on a Machmeter as a decimal fraction. *[Figure 3-13]* For example, if the Machmeter indicates .83 and the aircraft is flying at 30,000 feet where the speed of sound under standard conditions is 589.5 knots, the airspeed is 489.3 knots. The speed of sound varies with the air temperature. If the aircraft were flying at Mach .83 at 10,000 feet where the air is much warmer, its airspeed would be 530 knots.

Maximum Allowable Airspeed

Some aircraft that fly at high subsonic speeds are equipped with maximum allowable ASIs like the one in *Figure 3-14*. This instrument looks much like a standard air-speed indicator, calibrated in knots, but has an additional pointer colored red, checkered, or striped. The maximum airspeed pointer is actuated by an aneroid, or altimeter mechanism, that moves it to a lower value as air density decreases. By keeping the airspeed pointer at a lower value than the maximum pointer, the pilot avoids the onset of transonic shock waves.

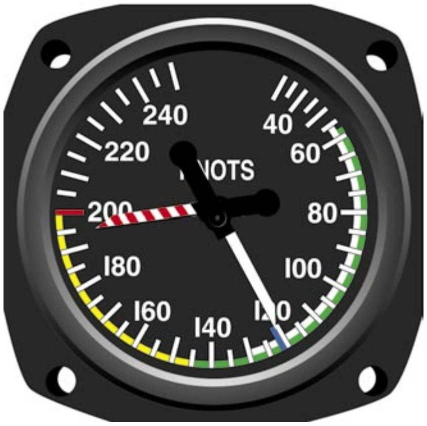

Figure 3-14. *A maximum allowable airspeed indicator has a movable pointer that indicates the never-exceed speed, which changes with altitude to avoid the onset of transonic shock waves.*

Airspeed Color Codes

The dial of an ASI is color coded to alert the pilot, at a glance, of the significance of the speed at which the aircraft is flying. These colors and their associated airspeeds are shown in *Figure 3-15*.

Magnetism

The Earth is a huge magnet, spinning in space, surrounded by a magnetic field made up of invisible lines of flux. These lines leave the surface at the magnetic north pole and reenter at the magnetic South Pole.

Lines of magnetic flux have two important characteristics: any magnet that is free to rotate will align with them, and

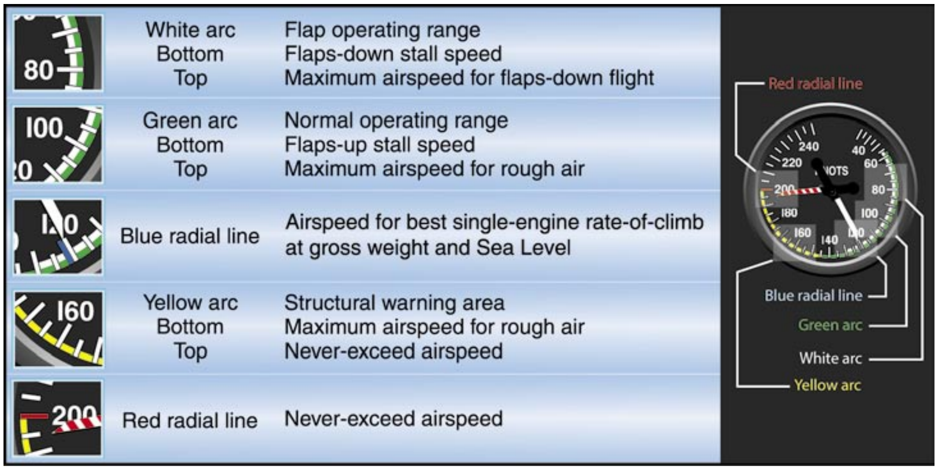

White arc	Flap operating range
Bottom	Flaps-down stall speed
Top	Maximum airspeed for flaps-down flight
Green arc	Normal operating range
Bottom	Flaps-up stall speed
Top	Maximum airspeed for rough air
Blue radial line	Airspeed for best single-engine rate-of-climb at gross weight and Sea Level
Yellow arc	Structural warning area
Bottom	Maximum airspeed for rough air
Top	Never-exceed airspeed
Red radial line	Never-exceed airspeed

Figure 3-15. *Color Codes for an Airspeed Indicator.*

an electrical current is induced into any conductor that cuts across them. Most direction indicators installed in aircraft make use of one of these two characteristics.

The Basic Aviation Magnetic Compass

One of the oldest and simplest instruments for indicating direction is the magnetic compass. It is also one of the basic instruments required by 14 CFR part 91 for both VFR and IFR flight.

Magnetic Compass Overview

A magnet is a piece of material, usually a metal containing iron, which attracts and holds lines of magnetic flux. Regardless of size, every magnet has two poles: a north pole and a south pole. When one magnet is placed in the field of another, the unlike poles attract each other and like poles repel.

An aircraft magnetic compass, such as the one in *Figure 3-16*, has two small magnets attached to a metal float sealed inside a bowl of clear compass fluid similar to kerosene. A graduated scale, called a card, is wrapped around the float and viewed through a glass window with a lubber line across it. The card is marked with letters representing the cardinal directions, north, east, south, and west, and a number for each 30° between these letters. The final "0" is omitted from these directions; for example, 3 = 30°, 6 = 60°, and 33 = 330°. There are long and short graduation marks between the letters and numbers, with each long mark representing 10° and each short mark representing 5°.

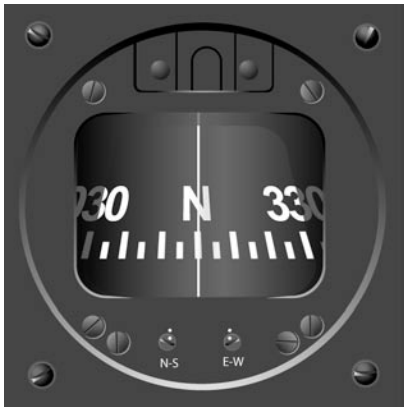

Figure 3-16. *A Magnetic Compass. The vertical line is called the lubber line.*

Magnetic Compass Construction

The float and card assembly has a hardened steel pivot in its center that rides inside a special, spring-loaded, hard-glass jewel cup. The buoyancy of the float takes most of the weight off the pivot, and the fluid damps the oscillation of the float and card. This jewel-and-pivot type mounting allows the float freedom to rotate and tilt up to approximately 18° angle of bank. At steeper bank angles, the compass indications are erratic and unpredictable.

The compass housing is entirely full of compass fluid. To prevent damage or leakage when the fluid expands and contracts with temperature changes, the rear of the compass case is sealed with a flexible diaphragm, or with a metal bellows in some compasses.

Magnetic Compass Theory of Operations

The magnets align with the Earth's magnetic field and the pilot reads the direction on the scale opposite the lubber line. Note that in *Figure 3-16*, the pilot sees the compass card from its backside. When the pilot is flying north as the compass shows, east is to the pilot's right, but on the card "33", which represents 330° (west of north), is to the right of north. The reason for this apparent backward graduation is that the card remains stationary, and the compass housing and the pilot turn around it, always viewing the card from its backside.

A compensator assembly mounted on the top or bottom of the compass allows an aviation maintenance technician (AMT) to create a magnetic field inside the compass housing that cancels the influence of local outside magnetic fields. This is done to correct for deviation error. The compensator assembly has two shafts whose ends have screwdriver slots accessible from the front of the compass. Each shaft rotates one or two small compensating magnets. The end of one shaft is marked E-W, and its magnets affect the compass when the aircraft is pointed east or west. The other shaft is marked N-S and its magnets affect the compass when the aircraft is pointed north or south.

Magnetic Compass Induced Errors

The magnetic compass is the simplest instrument in the panel, but it is subject to a number of errors that must be considered.

Variation

The Earth rotates about its geographic axis; maps and charts are drawn using meridians of longitude that pass through the geographic poles. Directions measured from the geographic poles are called true directions. The north magnetic pole to which the magnetic compass points is not collocated with the geographic north pole, but is some 1,300 miles away; directions measured from the magnetic poles are called magnetic directions. In aerial navigation, the difference between true and magnetic directions is called variation. This same angular difference in surveying and land navigation is called declination.

Figure 3-17 shows the isogonic lines that identify the number of degrees of variation in their area. The line that passes near Chicago is called the agonic line. Anywhere along this line the two poles are aligned, and there is no variation. East of this line, the magnetic pole is to the west of the geographic pole and a correction must be applied to a compass indication to get a true direction.

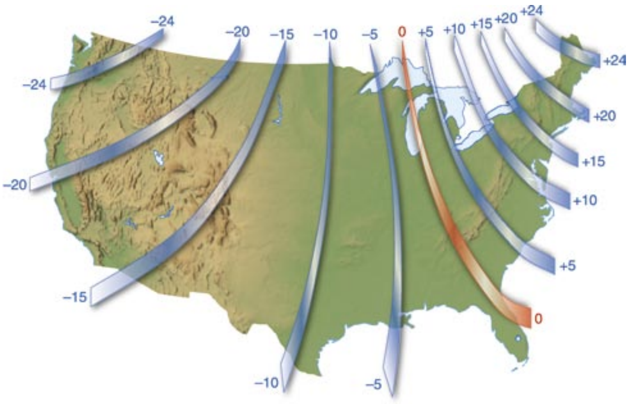

Figure 3-17. *Isogonic lines are lines of equal variation.*

Flying in the Washington, D.C. area, for example, the variation is 10° west. If the pilot wants to fly a true course of south (180°), the variation must be added to this resulting in a magnetic course to fly of 190°. Flying in the Los Angeles, CA area, the variation is 14° east. To fly a true course of 180° there, the pilot would have to subtract the variation and fly a magnetic course of 166°. The variation error does not change with the heading of the aircraft; it is the same anywhere along the isogonic line.

Deviation

The magnets in a compass align with any magnetic field. Local magnetic fields in an aircraft caused by electrical current flowing in the structure, in nearby wiring or any magnetized part of the structure, conflict with the Earth's magnetic field and cause a compass error called deviation.

Deviation, unlike variation, is different on each heading, but it is not affected by the geographic location. Variation error cannot be reduced or changed, but deviation error can be minimized when a pilot or AMT performs the maintenance task known as "swinging the compass."

Most airports have a compass rose, which is a series of lines marked out on a taxiway or ramp at some location where there is no magnetic interference. Lines, oriented to magnetic north, are painted every 30°, as shown in *Figure 3-18*.

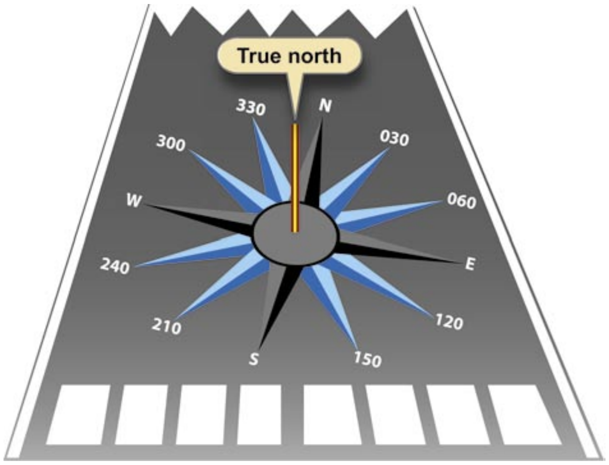

Figure 3-18. *Utilization of a Compass Rose Aids Compensation for Deviation Errors.*

The pilot or AMT aligns the aircraft on each magnetic heading and adjusts the compensating magnets to minimize the difference between the compass indication and the actual magnetic heading of the aircraft. Any error that cannot be removed is recorded on a compass correction card, like the one in *Figure 3-19*, and placed in a cardholder near the compass. If the pilot wants to fly a magnetic heading of 120° and the

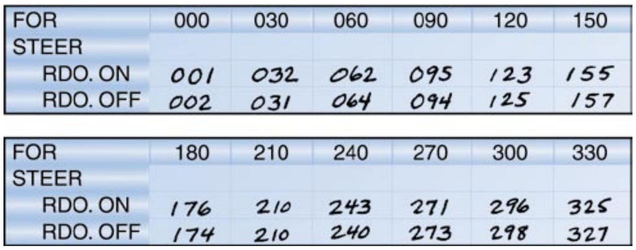

FOR	000	030	060	090	120	150
STEER						
RDO. ON	001	032	062	095	123	155
RDO. OFF	002	031	064	094	125	157

FOR	180	210	240	270	300	330
STEER						
RDO. ON	176	210	243	271	296	325
RDO. OFF	174	210	240	273	298	327

Figure 3-19. *A compass correction card shows the deviation correction for any heading.*

aircraft is operating with the radios on, the pilot should fly a compass heading of 123°.

The corrections for variation and deviation must be applied in the correct sequence and is shown below starting from the true course desired.

Step 1: Determine the Magnetic Course
True Course (180°) ± Variation (+10°) = Magnetic Course (190°)

The Magnetic Course (190°) is steered if there is no deviation error to be applied. The compass card must now be considered for the compass course of 190°.

Step 2: Determine the Compass Course
Magnetic Course (190°, from step 1) ± Deviation (-2°, from correction card) = Compass Course (188°)

NOTE: Intermediate magnetic courses between those listed on the compass card need to be interpreted. Therefore, to steer a true course of 180°, the pilot would follow a compass course of 188°.

To find the true course that is being flown when the compass course is known:
Compass Course ± Deviation = Magnetic Course ± Variation = True Course

Dip Errors

The lines of magnetic flux are considered to leave the Earth at the magnetic north pole and enter at the magnetic South Pole. At both locations the lines are perpendicular to the Earth's surface. At the magnetic equator, which is halfway between the poles, the lines are parallel with the surface. The magnets in a compass align with this field, and near the poles they dip, or tilt, the float and card. The float is balanced with a small dip-compensating weight, so it stays relatively level when operating in the middle latitudes of the northern hemisphere. This dip along with this weight causes two very noticeable errors: northerly turning error and acceleration error.

The pull of the vertical component of the Earth's magnetic field causes northerly turning error, which is apparent on a heading of north or south. When an aircraft flying on a heading of north makes a turn toward east, the aircraft banks to the right, and the compass card tilts to the right. The vertical component of the Earth's magnetic field pulls the north-seeking end of the magnet to the right, and the float rotates, causing the card to rotate toward west, the direction opposite the direction the turn is being made. *[Figure 3-20]*

If the turn is made from north to west, the aircraft banks to the left and the compass card tilts down on the left side. The magnetic field pulls on the end of the magnet that causes the card to rotate toward east. This indication is again opposite to the direction

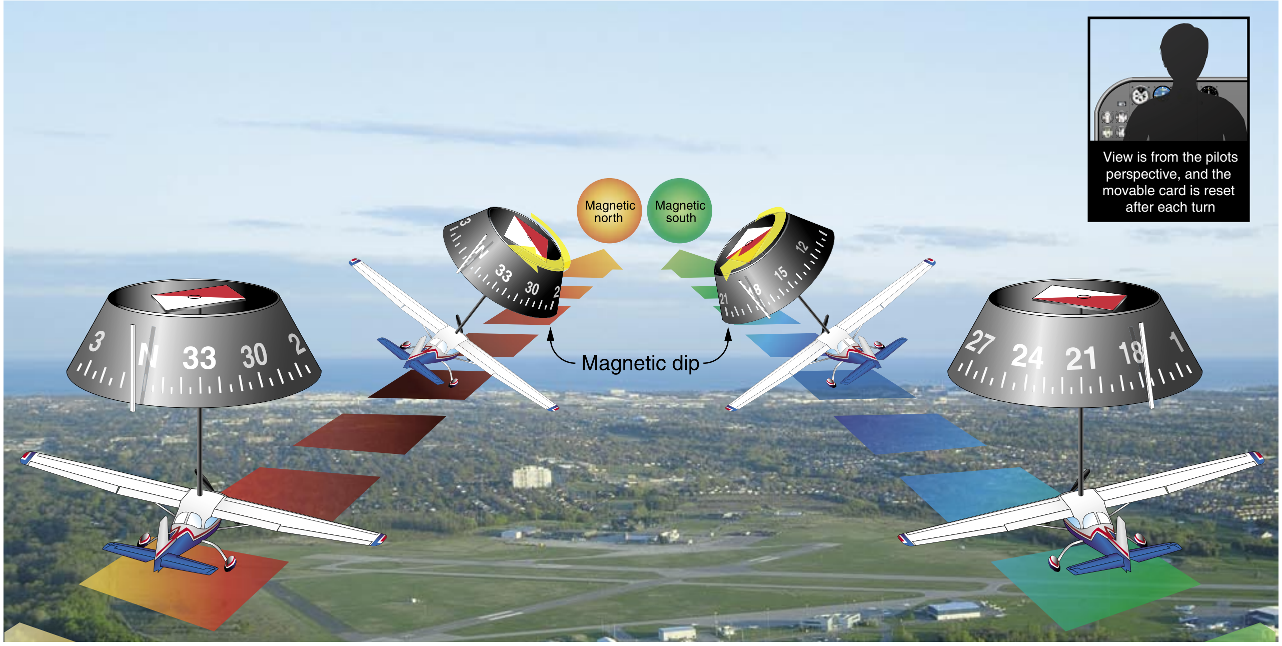
View is from the pilots perspective, and the movable card is reset after each turn
Magnetic north
Magnetic south
Magnetic dip
3
N
33
30
2
27
24
21
18
1
12
15

Figure 3-20. *Northerly Turning Error.*

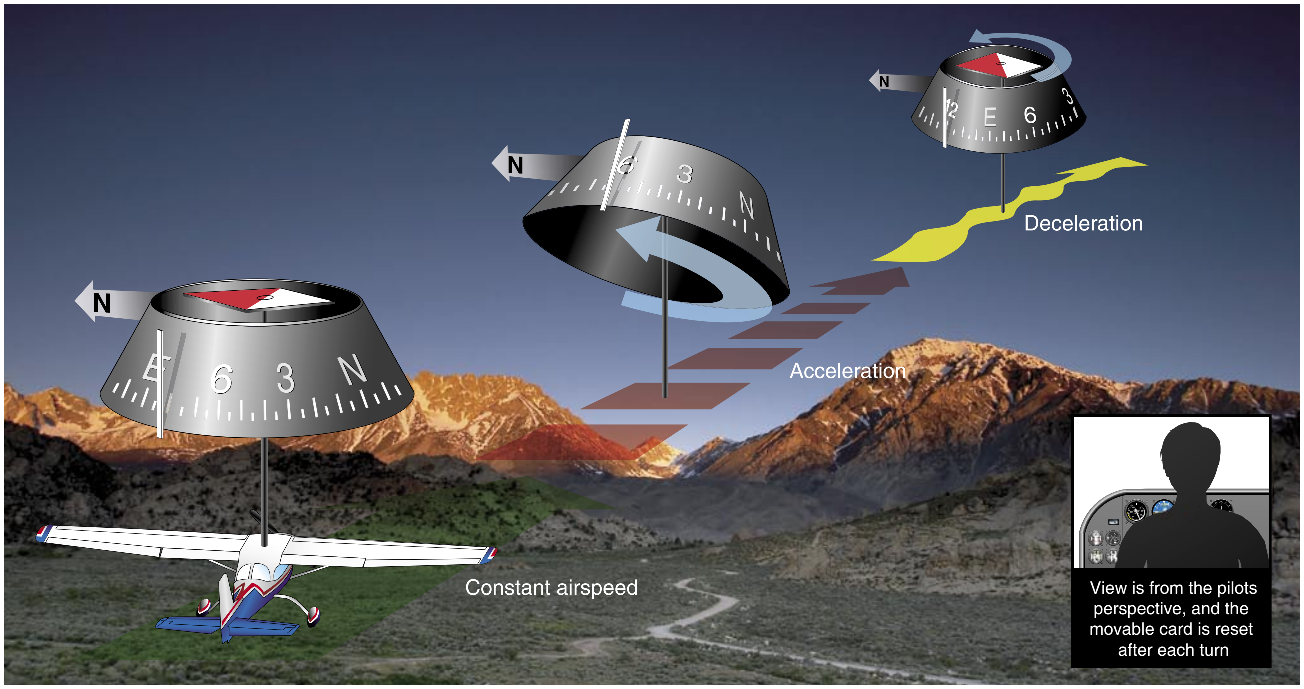
N
N
N
E 6 3 N
6 3 N
2 E 6 3
Deceleration
Acceleration
Constant airspeed
View is from the pilots perspective, and the movable card is reset after each turn

Figure 3-21. *The Effects of Acceleration Error.*

the turn is being made. The rule for this error is: when starting a turn from a northerly heading, the compass indication lags behind the turn.

When an aircraft is flying on a heading of south and begins a turn toward east, the Earth's magnetic field pulls on the end of the magnet that rotates the card toward east, the same direction the turn is being made. If the turn is made from south toward west, the magnetic pull starts the card rotating toward west—again, in the same direction the turn is being made. The rule for this error is: When starting a turn from a southerly heading, the compass indication leads the turn.

In acceleration error, the dip-correction weight causes the end of the float and card marked N (the south-seeking end) to be heavier than the opposite end. When the aircraft is flying at a constant speed on a heading of east or west, the float and card is level. The effects of magnetic dip and the weight are approximately equal. If the aircraft accelerates on a heading of east *[Figure 3-21]*, the inertia of the weight holds its end of the float back and the card rotates toward north. As soon as the speed of the aircraft stabilizes, the card swings back to its east indication. If, while flying on this easterly heading, the aircraft decelerates, the inertia causes the weight to move ahead and the card rotates toward south until the speed again stabilizes.

When flying on a heading of west, the same things happen. Inertia from acceleration causes the weight to lag, and the card rotates toward north. When the aircraft decelerates on a heading of west, inertia causes the weight to move ahead and the card rotates toward south.

Oscillation Error

Oscillation is a combination of all of the other errors, and it results in the compass card swinging back and forth around the heading being flown. When setting the gyroscopic heading indicator to agree with the magnetic compass, use the average indication between the swings.

The Vertical Card Magnetic Compass

The floating magnet type of compass not only has all the errors just described, but also lends itself to confused reading. It is easy to begin a turn in the wrong direction because its card appears backward. East is on what the pilot would expect to be the west side. The vertical card magnetic compass eliminates some of the errors and confusion. The dial of this compass is graduated with letters representing the cardinal directions, numbers every 30°, and marks every 5°. The dial is rotated by a set of gears from the shaft-mounted magnet, and the nose of the symbolic airplane on the instrument glass represents the lubber line for reading the heading of the aircraft from the dial. Eddy currents induced into an aluminum-damping cup damp oscillation of the magnet. *[Figure 3-22]*

The Flux Gate Compass System

As mentioned earlier, the lines of flux in the Earth's magnetic field have two basic characteristics: a magnet aligns with these lines, and an electrical current is induced, or generated, in any wire crossed by them.

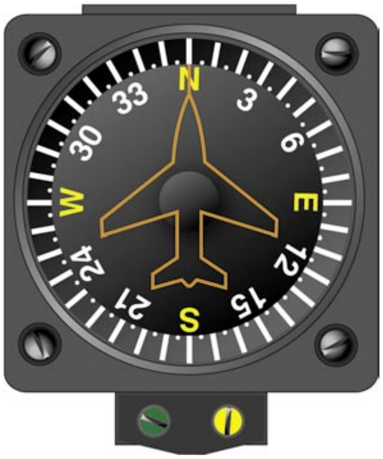

Figure 3-22. *Vertical Card Magnetic Compass.*

The flux gate compass that drives slaved gyros uses the characteristic of current induction. The flux valve is a small, segmented ring, like the one in *Figure 3-23*, made of soft iron that readily accepts lines of magnetic flux. An electrical coil is wound around each of the three legs to accept the current induced in this ring by the Earth's magnetic field. A coil wound around the iron spacer in the center of the frame has 400-Hz alternating current (A.C.) flowing through it. During the times when this current reaches its peak, twice during each cycle, there is so much magnetism produced by this coil that the frame cannot accept the lines of flux from the Earth's field.

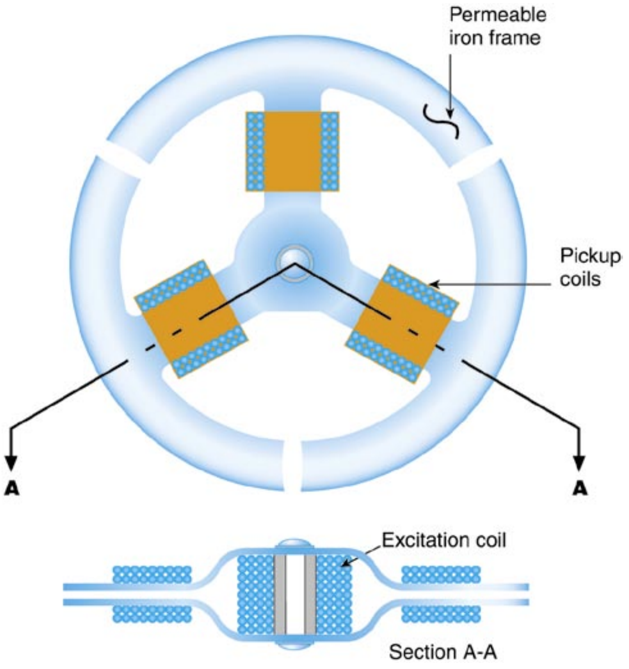

Figure 3-23. *The soft iron frame of the flux valve accepts the flux from the Earth's magnetic field each time the current in the center coil reverses. This flux causes current to flow in the three pickup coils.*

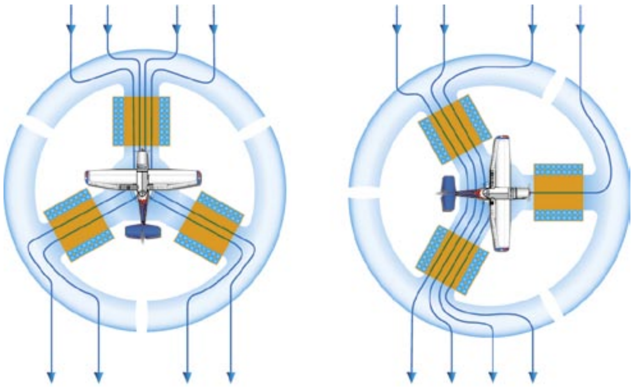

Figure 3-24. *The current in each of the three pickup coils changes with the heading of the aircraft.*

But as the current reverses between the peaks, it demagnetizes the frame so it can accept the flux from the Earth's field. As this flux cuts across the windings in the three coils, it causes current to flow in them. These three coils are connected in such a way that the current flowing in them changes as the heading of the aircraft changes. *[Figure 3-24]*

The three coils are connected to three similar but smaller coils in a synchro inside the instrument case. The synchro rotates the dial of a radio magnetic indicator (RMI) or a horizontal situation indicator (HSI).

Remote Indicating Compass

Remote indicating compasses were developed to compensate for the errors and limitations of the older type of heading indicators. The two panel-mounted components of a typical system are the pictorial navigation indicator and the slaving control and compensator unit. *[Figure 3-25]* The pictorial navigation indicator is commonly referred to as a HSI.

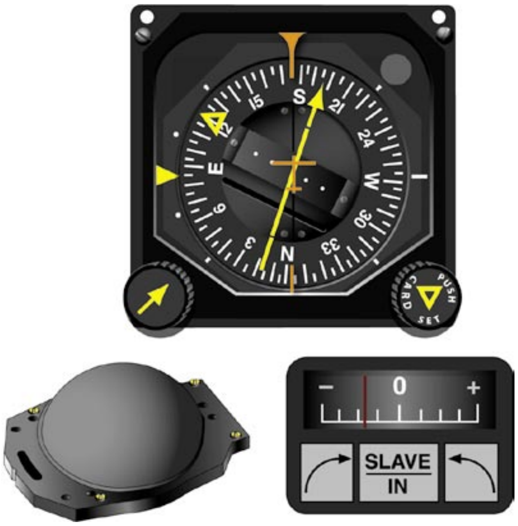

Figure 3-25. *Pictorial Navigation Indicator (HSI Top), Slaving Control and Compensator Unit.*

The slaving control and compensator unit has a pushbutton that provides a means of selecting either the "slaved gyro" or "free gyro" mode. This unit also has a slaving meter and two manual heading-drive buttons. The slaving meter indicates the difference between the displayed heading and the magnetic heading. A right deflection indicates a clockwise error of the compass card; a left deflection indicates a counterclockwise error. Whenever the aircraft is in a turn and the card rotates, the slaving meter shows a full deflection to one side or the other. When the system is in "free gyro" mode, the compass card may be adjusted by depressing the appropriate heading-drive button.

A separate unit, the magnetic slaving transmitter is mounted remotely; usually in a wingtip to eliminate the possibility of magnetic interference. It contains the flux valve, which is the direction-sensing device of the system. A concentration of lines of magnetic force, after being amplified, becomes a signal relayed to the heading indicator unit, which is also remotely mounted. This signal operates a torque motor in the heading indicator unit that processes the gyro unit until it is aligned with the transmitter signal. The magnetic slaving transmitter is connected electrically to the HSI.

There are a number of designs of the remote indicating compass; therefore, only the basic features of the system are covered here. Instrument pilots must become familiar with the characteristics of the equipment in their aircraft.

As instrument panels become more crowded and the pilot's available scan time is reduced by a heavier flight deck workload, instrument manufacturers have worked toward combining instruments. One good example of this is the RMI in *Figure 3-26*. The compass card is driven by signals from the flux valve, and the two pointers are driven by an automatic direction finder (ADF) and a very high frequency omnidirectional range (VOR).

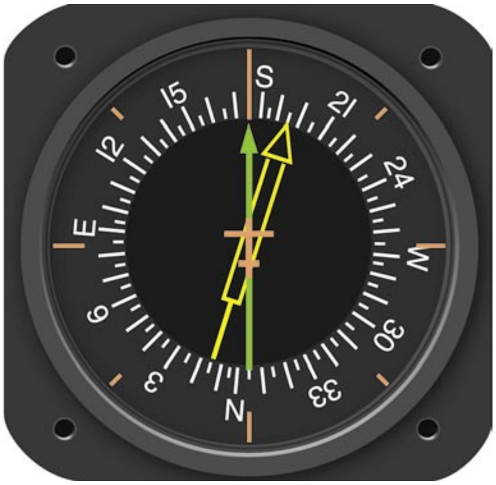

Figure 3-26. *Driven by signals from a flux valve, the compass card in this RMI indicates the heading of the aircraft opposite the upper center index mark. The green pointer is driven by the ADF.*

Gyroscopic Systems

Flight without reference to a visible horizon can be safely accomplished by the use of gyroscopic instrument systems and the two characteristics of gyroscopes, which are rigidity and precession. These systems include attitude, heading, and rate instruments, along with their power sources. These instruments include a gyroscope (or gyro) that is a small wheel with its weight concentrated around its periphery. When this wheel is spun at high speed, it becomes rigid and resists tilting or turning in any direction other than around its spin axis.

Attitude and heading instruments operate on the principle of rigidity. For these instruments, the gyro remains rigid in its case and the aircraft rotates about it. Rate indicators, such as turn indicators and turn coordinators, operate on the principle of precession. In this case, the gyro processes (or rolls over) proportionate to the rate the aircraft rotates about one or more of its axes.

Power Sources

Aircraft and instrument manufacturers have designed redundancy in the flight instruments so that any single failure will not deprive the pilot of the ability to safely conclude the flight. Gyroscopic instruments are crucial for instrument flight; therefore, they are powered by separate electrical or pneumatic sources.

Pneumatic Systems

Pneumatic gyros are driven by a jet of air impinging on buckets cut into the periphery of the wheel. On many aircraft this stream of air is obtained by evacuating the instrument case with a vacuum source and allowing filtered air to flow into the case through a nozzle to spin the wheel.

Venturi Tube Systems

Aircraft that do not have a pneumatic pump to evacuate the instrument case can use venturi tubes mounted on the outside of the aircraft, similar to the system shown in *Figure 3-27*. Air flowing through the venturi tube speeds up in the narrowest part and, according to Bernoulli's principle, the pressure drops. This location is connected to the instrument case by a piece of tubing. The two attitude instruments operate on approximately 4" Hg of suction; the turn-and-slip indicator needs only 2" Hg, so a pressure-reducing needle valve is used to decrease the suction. Air flows into the instruments through filters built into the instrument cases. In this system, ice can clog the venturi tube and stop the instruments when they are most needed.

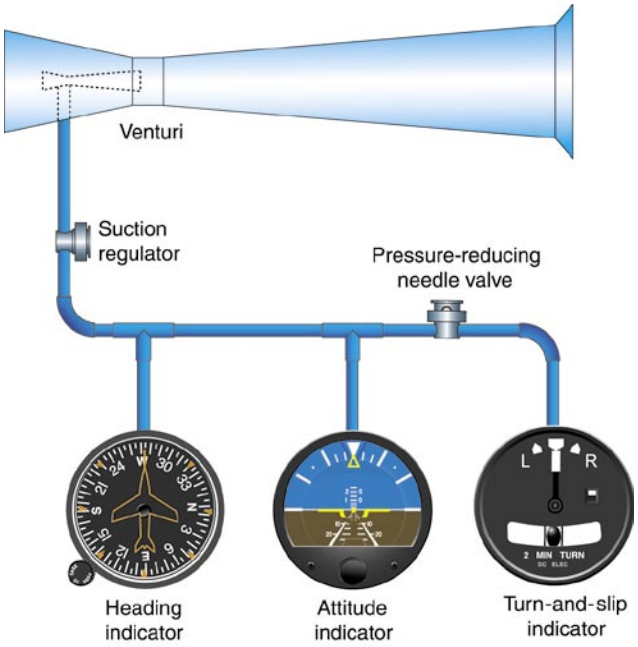
Venturi
Suction regulator
Pressure-reducing needle valve
L
R
2 MIN TURN
DC ELEC
Heading indicator
Attitude indicator
Turn-and-slip indicator

Figure 3-27. *A venturi tube system that provides necessary vacuum to operate key instruments.*

Vacuum Pump Systems

Wet-Type Vacuum Pump

Steel-vane air pumps have been used for many years to evacuate the instrument cases. The vanes in these pumps are lubricated by a small amount of engine oil metered into the pump and discharged with the air. In some aircraft the discharge air is used to inflate rubber deicer boots on the wing and empennage leading edges. To keep the oil from deteriorating the rubber boots, it must be removed with an oil separator like the one in *Figure 3-28*.

The vacuum pump moves a greater volume of air than is needed to supply the instruments with the suction needed, so a suction-relief valve is installed in the inlet side of the pump. This spring-loaded valve draws in just enough air to maintain the required low pressure inside the instruments, as is shown on the suction gauge in the instrument panel. Filtered air enters the instrument cases from a central air filter. As long as aircraft fly at relatively low altitudes, enough air is drawn into the instrument cases to spin the gyros at a sufficiently high speed.

Dry Air Vacuum Pump

As flight altitudes increase, the air is less dense and more air must be forced through the instruments. Air pumps that do not mix oil with the discharge air are used in high flying aircraft.

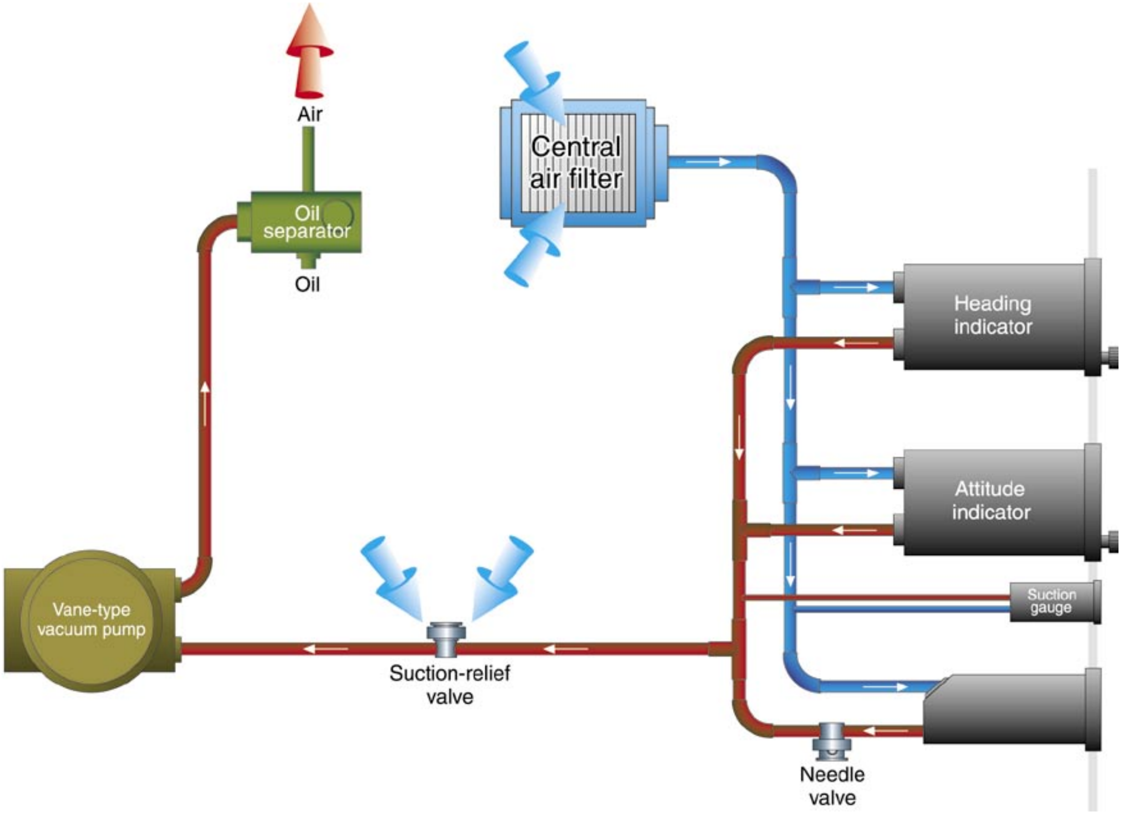
Air
Oil separator
Oil
Central air filter
Heading indicator
Attitude indicator
Suction gauge
Vane-type vacuum pump
Suction-relief valve
Needle valve

Figure 3-28. *Single-engine instrument vacuum system using a steel-vane wet-type vacuum pump.*

Steel vanes sliding in a steel housing need to be lubricated, but vanes made of a special formulation of carbon sliding inside carbon housing provide their own lubrication in a microscopic amount as they wear.

Pressure Indicating Systems

Figure 3-29 is a diagram of the instrument pneumatic system of a twin-engine general aviation airplane. Two dry air pumps are used with filters in their inlet to filter out any contaminants that could damage the fragile carbon vanes in the pump. The discharge air from the pump flows through a regulator, where excess air is bled off to maintain the pressure in the system at the desired level. The regulated air then flows through inline filters to remove any contamination that could have been picked up from the pump, and from there into a manifold check valve. If either engine should become inoperative or either pump should fail, the check valve isolates the inoperative system and the instruments are driven by air from the operating system. After the air passes through the instruments and drives the gyros, it is exhausted from the case. The gyro pressure gauge measures the pressure drop across the instruments.

Electrical Systems

Many general aviation aircraft that use pneumatic attitude indicators use electric rate indicators and/or the reverse. Some instruments identify their power source on their dial, but it is extremely important that pilots consult the POH/AFM to determine the power source of all instruments to know what action to take in the event of an instrument failure. Direct current (D.C.) electrical instruments are available in 14- or 28-volt models, depending upon the electrical system in the aircraft. A.C. is used to operate some attitude gyros and autopilots. Aircraft with only D.C. electrical systems can use A.C. instruments via installation of a solid-state D.C. to A.C. inverter, which changes 14 or 28 volts D.C. into three-phase 115-volt, 400-Hz A.C.

Gyroscopic Instruments

Attitude Indicators

The first attitude instrument (AI) was originally referred to as an artificial horizon, later as a gyro horizon; now it is more properly called an attitude indicator. Its operating mechanism is a small brass wheel with a vertical spin axis, spun at a high speed by either a stream of air impinging on buckets cut into its periphery, or by an electric motor. The gyro is mounted in a double gimbal, which allows the aircraft to pitch and roll about the gyro as it remains fixed in space.

A horizon disk is attached to the gimbals so it remains in the same plane as the gyro, and the aircraft pitches and rolls about it. On early instruments, this was just a bar that

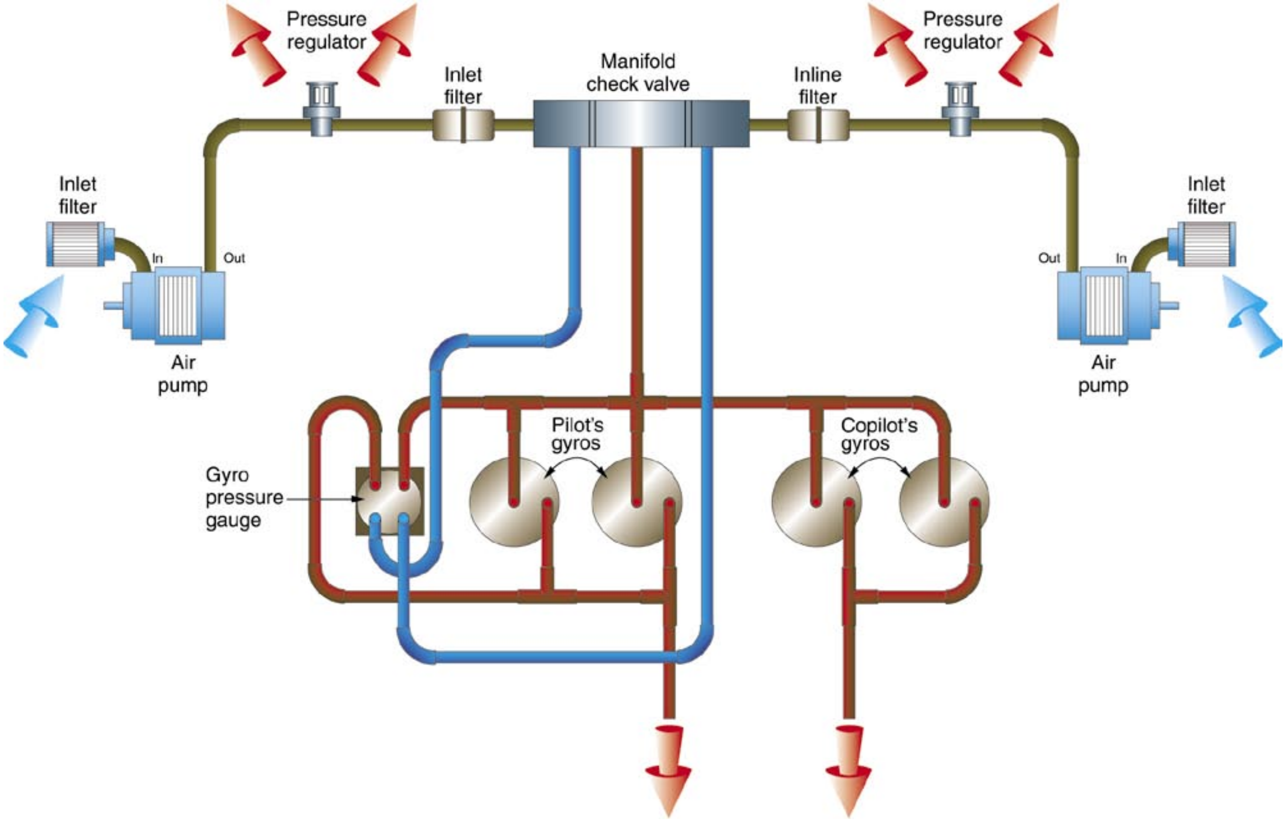

Figure 3-29. *Twin-Engine Instrument Pressure System Using a Carbon-Vane Dry-Type Air Pump.*

represented the horizon, but now it is a disc with a line representing the horizon and both pitch marks and bank-angle lines. The top half of the instrument dial and horizon disc is blue, representing the sky; and the bottom half is brown, representing the ground. A bank index at the top of the instrument shows the angle of bank marked on the banking scale with lines that represent 10°, 20°, 30°, 45°, and 60°. *[Figure 3-30]*

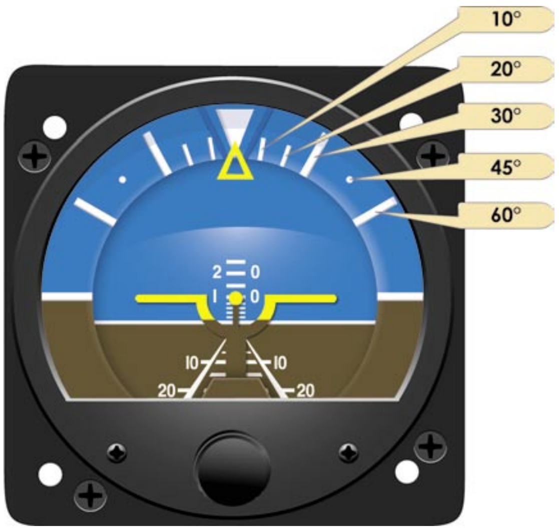

Figure 3-30. *The dial of this attitude indicator has reference lines to show pitch and roll.*

A small symbolic aircraft is mounted in the instrument case so it appears to be flying relative to the horizon. A knob at the bottom center of the instrument case raises or lowers the aircraft to compensate for pitch trim changes as the airspeed changes. The width of the wings of the symbolic aircraft and the dot in the center of the wings represent a pitch change of approximately 2°.

For an AI to function properly, the gyro must remain vertically upright while the aircraft rolls and pitches around it. The bearings in these instruments have a minimum of friction; however, even this small amount places a restraint on the gyro producing precession and causing the gyro to tilt. To minimize this tilting, an erection mechanism inside the instrument case applies a force any time the gyro tilts from its vertical position. This force acts in such a way to return the spinning wheel to its upright position.

The older artificial horizons were limited in the amount of pitch or roll they could tolerate, normally about 60° in pitch and 100° in roll. After either of these limits was exceeded, the gyro housing contacted the gimbals, applying such a precessing force that the gyro tumbled. Because of this limitation, these instruments had a caging mechanism that locked the gyro in its vertical position during any maneuvers that exceeded the instrument limits. Newer instruments do not have these restrictive tumble limits; therefore, they do not have a caging mechanism.

When an aircraft engine is first started and pneumatic or electric power is supplied to the instruments, the gyro is not erect. A self-erecting mechanism inside the instrument actuated by the force of gravity applies a precessing force, causing the gyro to rise to its vertical position. This erection can take as long as 5 minutes, but is normally done within 2 to 3 minutes.

Attitude indicators are free from most errors, but depending upon the speed with which the erection system functions, there may be a slight nose-up indication during a rapid acceleration and a nose-down indication during a rapid deceleration. There is also a possibility of a small bank angle and pitch error after a 180° turn. These inherent errors are small and correct themselves within a minute or so after returning to straight-and-level flight.

Heading Indicators

A magnetic compass is a dependable instrument used as a backup instrument. Although very reliable, it has so many inherent errors that it has been supplemented with gyroscopic heading indicators.

The gyro in a heading indicator is mounted in a double gimbal, as in an attitude indicator, but its spin axis is horizontal permitting sensing of rotation about the vertical axis of the aircraft. Gyro heading indicators, with the exception of slaved gyro indicators, are not north seeking, therefore they must be manually set to the appropriate heading by referring to a magnetic compass. Rigidity causes them to maintain this heading indication, without the oscillation and other errors inherent in a magnetic compass.

Older directional gyros use a drum-like card marked in the same way as the magnetic compass card. The gyro and the card remain rigid inside the case with the pilot viewing the card from the back. This creates the possibility the pilot might start a turn in the wrong direction similar to using a magnetic compass. A knob on the front of the instrument, below the dial, can be pushed in to engage the gimbals. This locks the gimbals allowing the pilot to rotate the gyro and card until the number opposite the lubber line agrees with the magnetic compass. When the knob is pulled out, the gyro remains rigid and the aircraft is free to turn around the card.

Directional gyros are almost all air-driven by evacuating the case and allowing filtered air to flow into the case and out through a nozzle, blowing against buckets cut in the periphery of the wheel. The Earth constantly rotates at 15° per hour while the gyro is maintaining a position relative to space, thus causing an apparent drift in the displayed heading of 15° per hour. When using these instruments, it is standard practice to compare the heading indicated on the directional gyro with the magnetic compass at least every 15 minutes and to reset the heading as necessary to agree with the magnetic compass.

Heading indicators like the one in *Figure 3-31* work on the same principle as the older horizontal card indicators, except that the gyro drives a vertical dial that looks much like the dial of a vertical card magnetic compass. The heading of the aircraft is shown against the nose of the symbolic aircraft on the instrument glass, which serves as the lubber line. A knob in the front of the instrument may be pushed in and turned to rotate the gyro and dial. The knob is spring loaded so it disengages from the gimbals as soon as it is released. This instrument should be checked about every 15 minutes to see if it agrees with the magnetic compass.

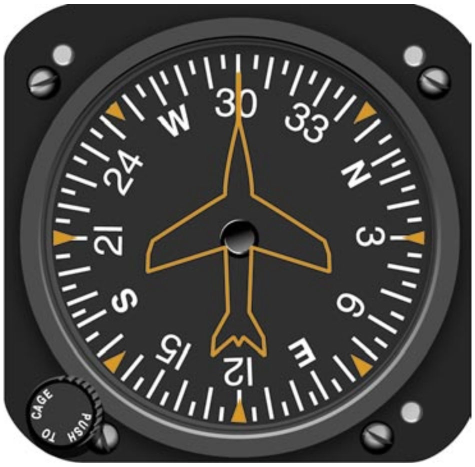

Figure 3-31. *The heading indicator is not north seeking, but must be set periodically (about every 15 minutes) to agree with the magnetic compass.*

Turn Indicators

Attitude and heading indicators function on the principle of rigidity, but rate instruments such as the turn-and-slip indicator operate on precession. Precession is the characteristic of a gyroscope that causes an applied force to produce a movement, not at the point of application, but at a point 90° from the point of application in the direction of rotation. *[Figure 3-32]*

Turn-and-Slip Indicator

The first gyroscopic aircraft instrument was the turn indicator in the needle and ball, or turn-and-bank indicator, which has more recently been called a turn-and-slip indicator. *[Figure 3-33]*

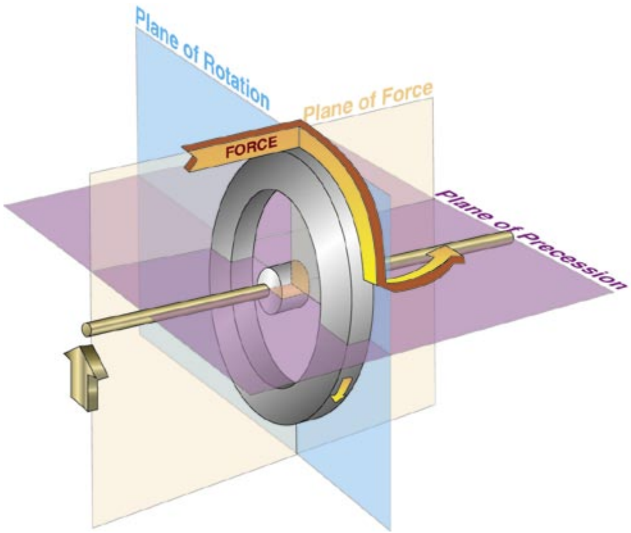

Figure 3-32. *Precession causes a force applied to a spinning wheel to be felt 90° from the point of application in the direction of rotation.*

The inclinometer in the instrument is a black glass ball sealed inside a curved glass tube that is partially filled with a liquid for damping. This ball measures the relative strength of the force of gravity and the force of inertia caused by a turn. When the aircraft is flying straight-and-level, there is no inertia acting on the ball, and it remains in the center of the tube between two wires. In a turn made with a bank angle that is too steep, the force of gravity is greater than the inertia and the ball rolls down to the inside of the turn. If the turn is made with too shallow a bank angle, the inertia is greater than gravity and the ball rolls upward to the outside of the turn.

The inclinometer does not indicate the amount of bank, nor does it indicate slip; it only indicates the relationship between the angle of bank and the rate of yaw.

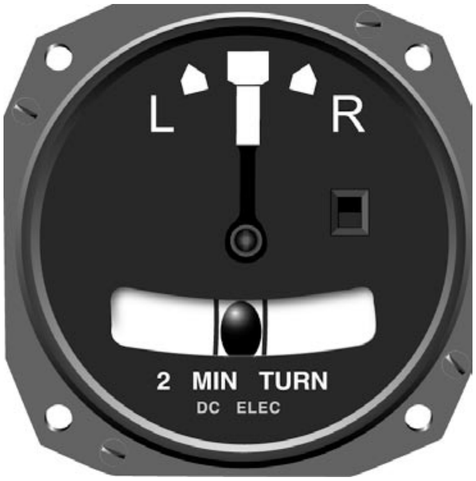

Figure 3-33. *Turn-and-Slip Indicator.*

The turn indicator is a small gyro spun either by air or by an electric motor. The gyro is mounted in a single gimbal with its spin axis parallel to the lateral axis of the aircraft and the axis of the gimbal parallel with the longitudinal axis. *[Figure 3-34]*

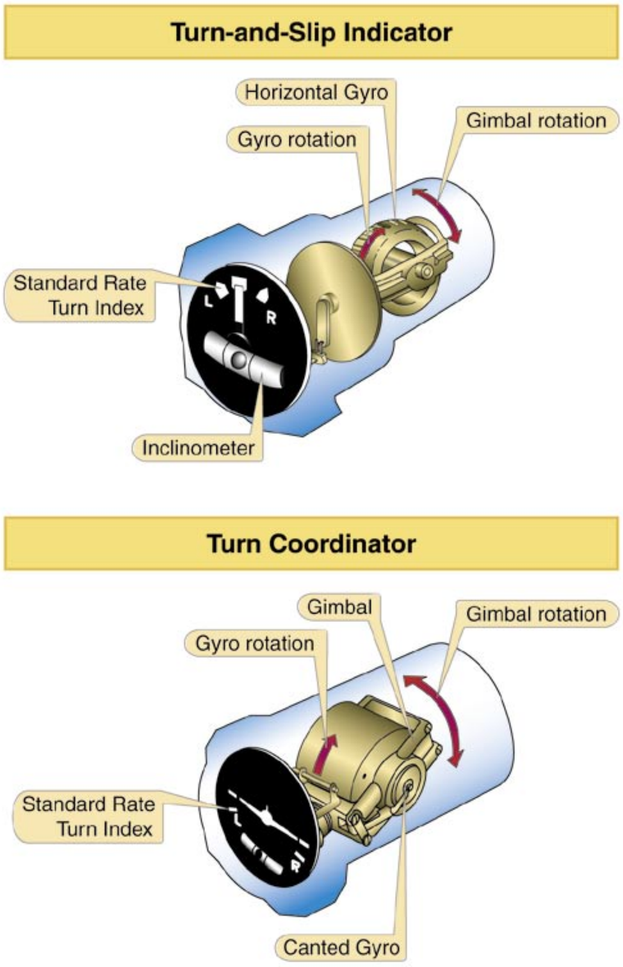

Figure 3-34. *The rate gyro in both turn-and-slip indicator and turn coordinator.*

When the aircraft yaws, or rotates about its vertical axis, it produces a force in the horizontal plane that, due to precession, causes the gyro and its gimbal to rotate about the gimbal's axis. It is restrained in this rotation plane by a calibration spring; it rolls over just enough to cause the pointer to deflect until it aligns with one of the doghouse-shaped marks on the dial, when the aircraft is making a standard rate turn.

The dial of these instruments is marked "2 MIN TURN." Some turn-and-slip indicators used in faster aircraft are marked "4 MIN TURN." In either instrument, a standard rate turn is being made whenever the needle aligns with a doghouse. A standard rate turn is 3° per second. In a 2 minute instrument, if the needle is one needle width either side of the center alignment mark, the turn is 3° per second and the turn takes 2 minutes to execute a 360° turn. In a 4 minute instrument, the same turn takes two widths deflection of the needle to achieve 3° per second.

Turn Coordinator

The major limitation of the older turn-and-slip indicator is that it senses rotation only about the vertical axis of the aircraft. It tells nothing of the rotation around the longitudinal axis, which in normal flight occurs before the aircraft begins to turn.

A turn coordinator operates on precession, the same as the turn indicator, but its gimbals frame is angled upward about 30° from the longitudinal axis of the aircraft. *[Figure 3-34]* This allows it to sense both roll and yaw. Therefore during a turn, the indicator first shows the rate of banking and once stabilized, the turn rate. Some turn coordinator gyros are dual-powered and can be driven by either air or electricity.

Rather than using a needle as an indicator, the gimbal moves a dial that is the rear view of a symbolic aircraft. The bezel of the instrument is marked to show wings-level flight and bank angles for a standard rate turn. *[Figure 3-35]*

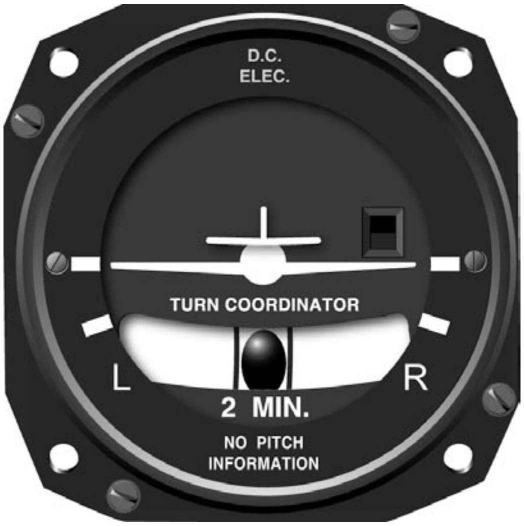

Figure 3-35. *A turn coordinator senses rotation about both roll and yaw axes.*

The inclinometer, similar to the one in a turn-and-slip indicator, is called a coordination ball, which shows the relationship between the bank angle and the rate of yaw. The turn is coordinated when the ball is in the center, between the marks. The aircraft is skidding when the ball rolls toward the outside of the turn and is slipping when it moves toward the inside of the turn. A turn coordinator does not sense pitch. This is indicated on some instruments by placing the words "NO PITCH INFORMATION" on the dial.

Flight Support Systems

Attitude and Heading Reference System (AHRS)

As aircraft displays have transitioned to new technology, the sensors that feed them have also undergone significant change. Traditional gyroscopic flight instruments have been replaced by Attitude and Heading Reference Systems (AHRS) improving reliability and thereby reducing cost and maintenance.

The function of an AHRS is the same as gyroscopic systems; that is, to determine which way is level and which way is north. By knowing the initial heading the AHRS can determine both the attitude and magnetic heading of the aircraft.

The genesis of this system was initiated by the development of the ring-LASAR gyroscope developed by Kearfott located in Little Falls, New Jersey. *[Figure 3-36]* Their development of the Ring-LASAR gyroscope in the 1960s/1970s was in support of Department of Defense (DOD) programs to include cruise missile technology. With the precision of these gyroscopes, it became readily apparent that they could be leveraged for multiple tasks and functions. Gyroscopic miniaturization has become so common that solid-state gyroscopes are found in products from robotics to toys.

Because the AHRS system replaces separate gyroscopes, such as those associated with an attitude indicator, magnetic heading indicator and turn indicator these individual systems are no longer needed. As with many systems today, AHRS itself had matured with time. Early AHRS systems used expensive inertial sensors and flux valves. However, today the AHRS for aviation and general aviation in particular are small solid-state systems integrating a variety of technology such as low cost inertial sensors, rate gyros, and magnetometers, and have capability for satellite signal reception.

Air Data Computer (ADC)

An Air Data Computer (ADC) *[Figure 3-37]* is an aircraft computer that receives and processes pitot pressure, static pressure, and temperature to calculate very precise altitude, IAS, TAS, and air temperature. The ADC outputs this information in a digital format that can be used by a variety of aircraft systems including an EFIS. Modern ADCs are small solid-state units. Increasingly, aircraft systems such as autopilots, pressurization, and FMS utilize ADC information for normal operations. NOTE: In most modern general aviation systems, both the AHRS and ADC are integrated within the electronic displays themselves thereby reducing the number of units, reducing weight, and providing simplification for installation resulting in reduced costs.

Analog Pictorial Displays

Horizontal Situation Indicator (HSI)

The HSI is a direction indicator that uses the output from a flux valve to drive the dial, which acts as the compass card. This instrument, shown in *Figure 3-38*, combines the magnetic compass with navigation signals and a glide slope. This gives the pilot an indication of the location of the aircraft with relationship to the chosen course.

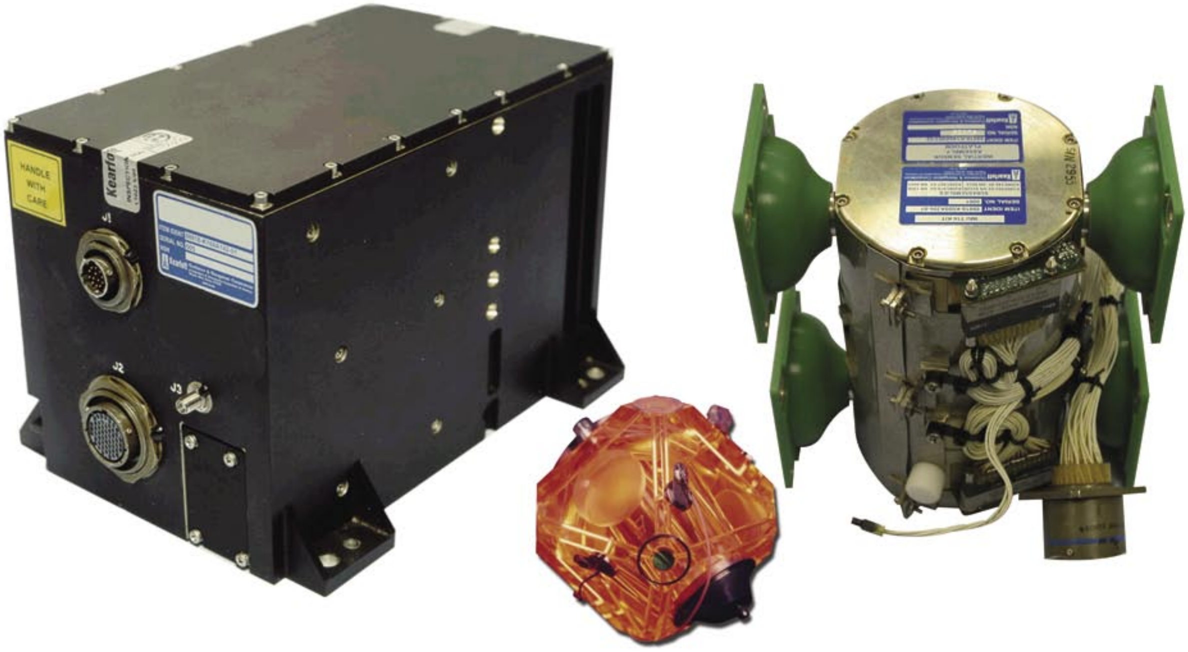

Figure 3-36. *The Kearfott Attitude Heading Reference System (AHRS) on the left incorporates a Monolithic Ring Laser Gyro (MRLG) (center), which is housed in an Inertial Sensor Assembly (ISA) on the right.*

Collins
PITOT
STATIC
WARNING — FAA REGULATIONS REQUIRE QUALIFIED PERSONNEL TO WORK ON STATIC SYSTEM AND A PNEUMATIC TEST IF SYSTEM IS OPENED.
CAUTION — DO NOT BOTTOM TAPERED END OF ADJUSTABLE FITTINGS IN PORTS. USE ONLY MS28778-4 AND -6 GASKETS.
ADC-850
ALERT
VALID
TEST

Figure 3-37. *Air Data Computer (Collins).*

In *Figure 3-38*, the aircraft heading displayed on the rotating azimuth card under the upper lubber line is North or 360°. The course-indicating arrowhead shown is set to 020; the tail indicates the reciprocal, 200°. The course deviation bar operates with a VOR/Localizer (VOR/LOC) navigation receiver to indicate left or right deviations from the course selected with the course-indicating arrow, operating in the same manner that the angular movement of a conventional VOR/LOC needle indicates deviation from course.

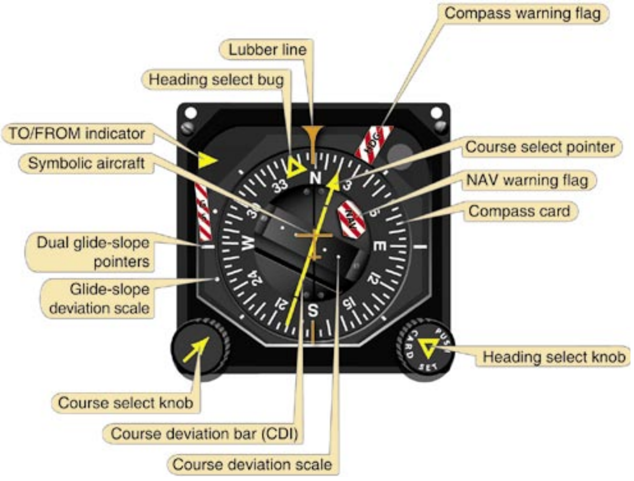
Compass warning flag
Lubber line
Heading select bug
TO/FROM indicator
Symbolic aircraft
Course select pointer
NAV warning flag
Compass card
Dual glide-slope pointers
Glide-slope deviation scale
Heading select knob
Course select knob
Course deviation bar (CDI)
Course deviation scale
N
33
30
W
24
21
S
15
12
E
6
3
NAV
HDG
PUSH CARD SET

Figure 3-38. *Horizontal Situation Indicator (HSI).*

The desired course is selected by rotating the course-indicating arrow in relation to the azimuth card by means of the course select knob. This gives the pilot a pictorial presentation: the fixed aircraft symbol and course deviation bar display the aircraft relative to the selected course, as though the pilot were above the aircraft looking down. The TO/FROM indicator is a triangular pointer. When the indicator points to the head of the course arrow, it shows that the course selected, if properly intercepted and flown, takes the aircraft to the selected facility. When the indicator points to the tail of the course arrow, it shows that the course selected, if properly intercepted and flown, takes the aircraft directly away from the selected facility.

The glide slope deviation pointer indicates the relation of the aircraft to the glide slope. When the pointer is below the center position, the aircraft is above the glide slope, and an increased rate of descent is required. In most installations, the azimuth card is a remote indicating compass driven by a fluxgate; however, in few installations where a fluxgate is not installed, or in emergency operation, the heading must be checked against the magnetic compass occasionally and reset with the course select knob.

Attitude Direction Indicator (ADI)

Advances in attitude instrumentation combine the gyro horizon with other instruments such as the HSI, thereby reducing the number of separate instruments to which the pilot must devote attention. The attitude direction indicator (ADI) is an example of such technological advancement. A flight director incorporates the ADI within its system, which is further explained below (Flight Director System). However, an ADI need not have command cues; however, it is normally equipped with this feature.

Flight Director System (FDS)

A Flight Director System (FDS) combines many instruments into one display that provides an easily interpreted understanding of the aircraft's flight path. The computed solution furnishes the steering commands necessary to obtain and hold a desired path.

Major components of an FDS include an ADI, also called a Flight Director Indicator (FDI), an HSI, a mode selector, and a flight director computer. It should be noted that a flight director in use does not infer the aircraft is being manipulated by the autopilot (coupled), but is providing steering commands that the pilot (or the autopilot, if coupled) follows.

Typical flight directors use one of two display systems for steerage. The first is a set of command bars, one horizontal and one vertical. The command bars in this configuration are maintained in a centered position (much like a centered glide slope). The second uses a miniature aircraft aligned to a command cue.

A flight director displays steerage commands to the pilot on the ADI. As previously mentioned, the flight director receives its signals from one of various sources and provides that to the ADI for steerage commands. The mode controller provides signals through the ADI to drive the steering bars, e.g., the pilot flies the aircraft to place the delta symbol in the V of the steering bars. "Command" indicators tell the pilot in which direction and how much to change aircraft attitude to achieve the desired result.

The computed command indications relieve the pilot of many of the mental calculations required for instrument flight. The yellow cue in the ADI *[Figure 3-39]* provides all steering commands to the pilot. It is driven by a computer that receives information from the navigation systems, the ADC, AHRS, and other sources of data. The computer processes this information, providing the pilot with a single cue to follow. Following the cue provides the pilot with the necessary three-dimensional flight trajectory to maintain the desired path.

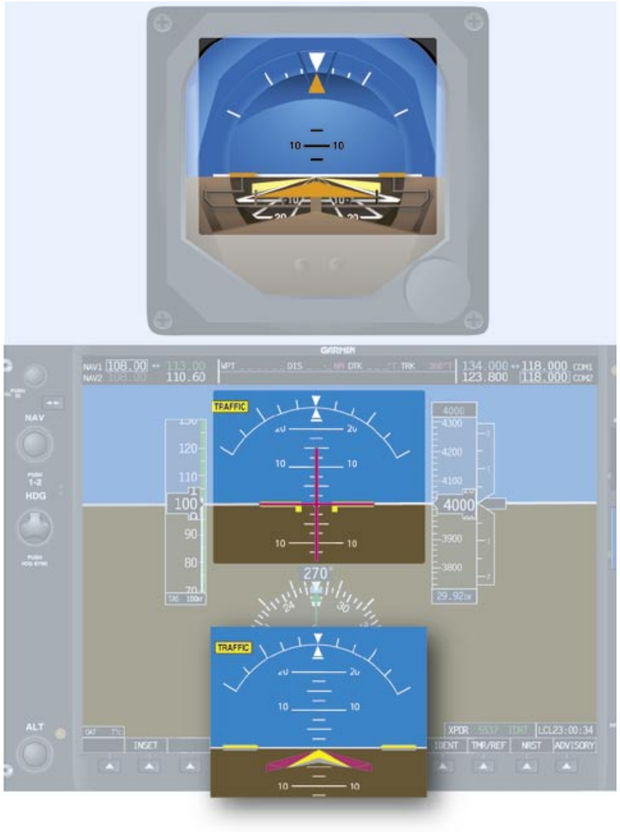

Figure 3-39. *A Typical Cue That a Pilot Would Follow.*

One of the first widely used flight directors was developed by Sperry and was called the Sperry Three Axis Attitude Reference System (STARS). Developed in the 1960s, it was commonly found on both commercial and business aircraft alike. STARS (with a modification) and successive flight directors were integrated with the autopilots and aircraft providing a fully integrated flight system.

The flight director/autopilot system described below is typical of installations in many general aviation aircraft.

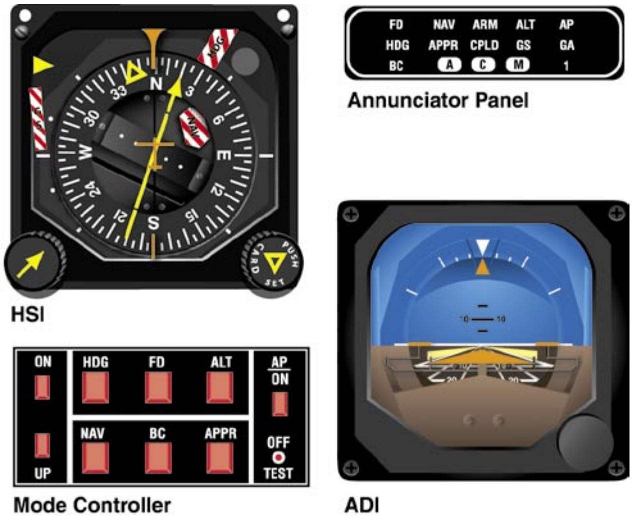

Figure 3-40. *Components of a Typical Flight Director System.*

The components of a typical flight director include the mode controller, ADI, HSI, and annunciator panel. These units are illustrated in *Figure 3-40*.

The pilot may choose from among many modes including the HDG (heading) mode, the VOR/LOC (localizer tracking) mode, or the AUTO Approach (APP) or G/S (automatic capture and tracking of instrument landing system (ILS) localizers and glide path) mode. The auto mode has a fully automatic pitch selection computer that takes into account aircraft performance and wind conditions, and operates once the pilot has reached the ILS glide slope. More sophisticated systems allow more flight director modes.

Integrated Flight Control System

The integrated flight control system integrates and merges various systems into a system operated and controlled by one principal component. *Figure 3-41* illustrates key components of the flight control system that was developed from the onset as a fully integrated system comprised of the airframe, autopilot, and flight director system. This trend of complete integration, once seen only in large commercial aircraft, are now becoming common in the general aviation field.

Autopilot Systems

An autopilot is a mechanical means to control an aircraft using electrical, hydraulic, or digital systems. Autopilots can control three axes of the aircraft: roll, pitch, and yaw. Most autopilots in general aviation control roll and pitch.

Autopilots also function using different methods. The first is position based. That is, the attitude gyro senses the degree of difference from a position such as wings level, a change in pitch, or a heading change.

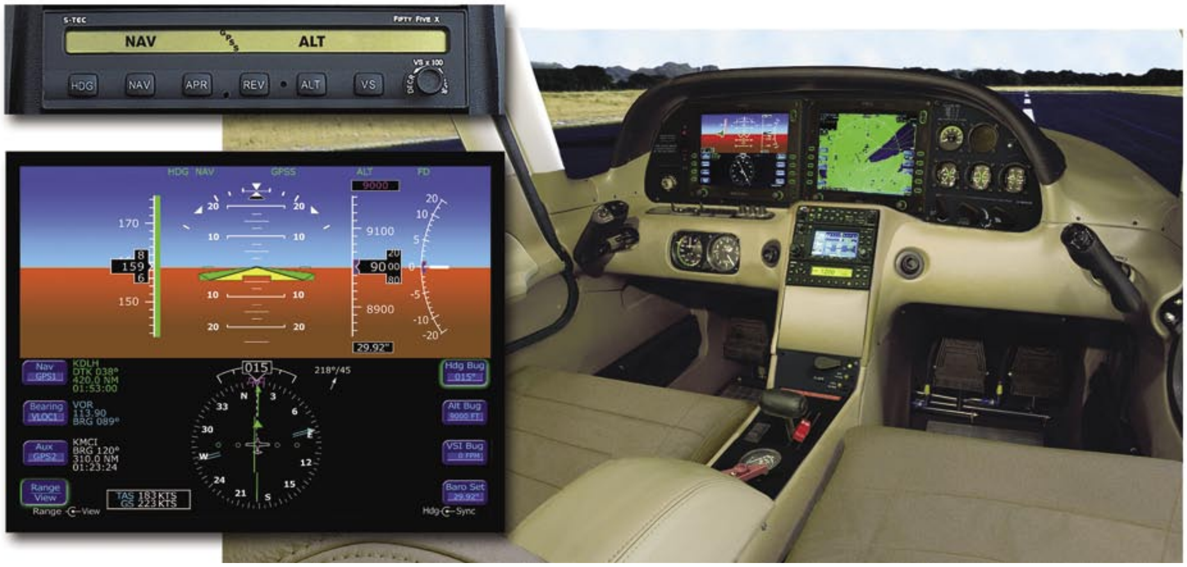

Figure 3-41. *The S-TEC/Meggit Corporation Integrated Autopilot Installed in the Cirrus.*

Determining whether a design is position based and/or rate based lies primarily within the type of sensors used. In order for an autopilot to possess the capability of controlling an aircraft's attitude (i.e., roll and pitch), that system must be provided with constant information on the actual attitude of that aircraft. This is accomplished by the use of several different types of gyroscopic sensors. Some sensors are designed to indicate the aircraft's attitude in the form of position in relation to the horizon, while others indicate rate (position change over time).

Rate-based systems use the turn-and-bank sensor for the autopilot system. The autopilot uses rate information on two of the aircraft's three axes: movement about the vertical axis (heading change or yaw) and about the longitudinal axis (roll). This combined information from a single sensor is made possible by the 30° offset in the gyro's axis to the longitudinal axis.

Other systems use a combination of both position and rate-based information to benefit from the attributes of both systems while newer autopilots are digital. *Figure 3-42* illustrates an autopilot by Century.

Figure 3-43 is a diagram layout of a rate-based autopilot by S-Tec, which permits the purchaser to add modular capability form basic wing leveling to increased capability.

Flight Management Systems (FMS)

In the mid-1970s, visionaries in the avionics industry such as Hubert Naimer of Universal, and followed by others such as Ed King, Jr., were looking to advance the technology of aircraft navigation. As early as 1976, Naimer had a vision of a "Master Navigation System" that would accept inputs from a variety of different types of sensors on an aircraft and automatically provide guidance throughout all phases of flight.

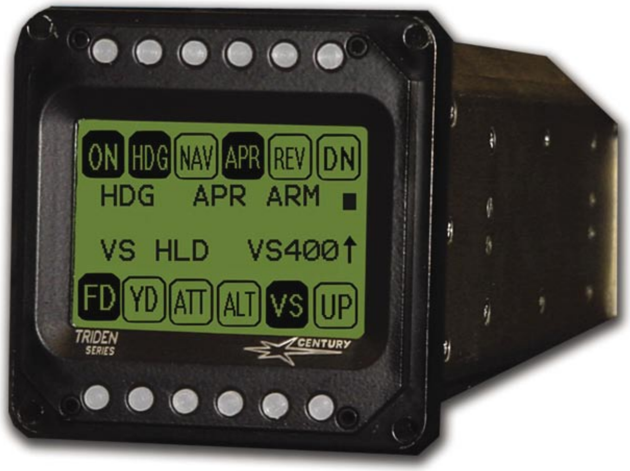

Figure 3-42. *An Autopilot by Century.*

At that time aircraft navigated over relatively short distances with radio systems, principally VOR or ADF. For long-range flight inertial navigation systems (INS), Omega, Doppler, and Loran were in common use. Short-range radio systems usually did not provide area navigation capability. Long-range systems were only capable of en route point-to-point navigation between manually entered waypoints described as longitude and latitude coordinates, with typical systems containing a limited number of waypoints.

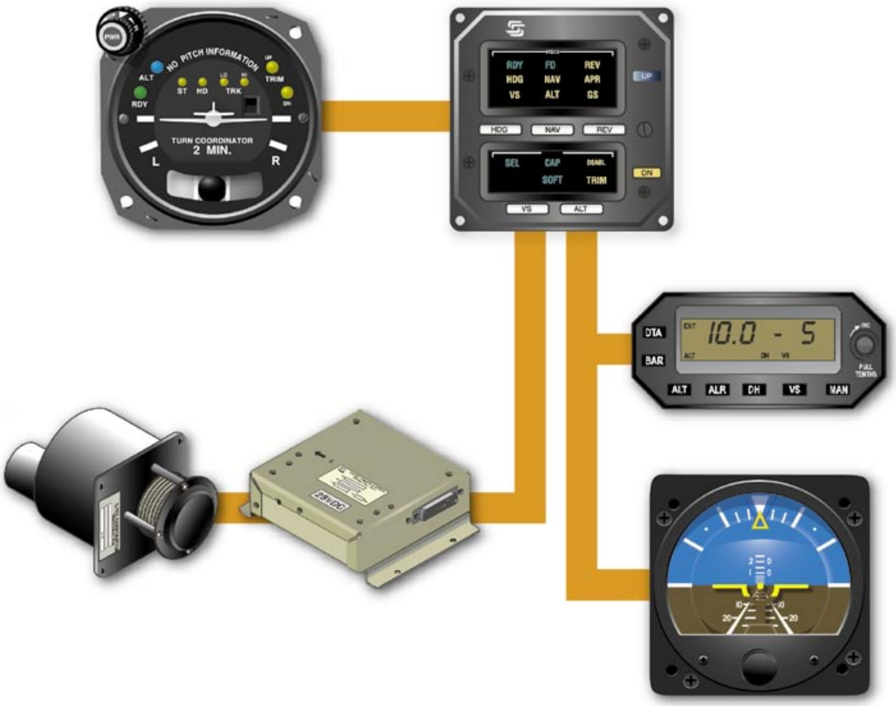
PWR
NO PITCH INFORMATION
ALT
RDY
ST
HD
TRK
TRIM
DN
TURN COORDINATOR
2 MIN.
L
R
RDY FD REV
HDG NAV APR
VS ALT GS
HDG
NAV
REV
UP
SEL CAP DSABL
SOFT TRIM
DN
VS
ALT
DTA
BAR
10.0 - 5
ALT
ALR
DH
VS
MAN

Figure 3-43. *A Diagram Layout of an Autopilot by S-Tec.*

The laborious process of manually entering cryptic latitude and longitude data for each flight waypoint created high crew workloads and frequently resulted in incorrect data entry. The requirement of a separate control panel for each long-range system consumed precious flight deck space and increased the complexity of interfacing the systems with display instruments, flight directors, and autopilots.

The concept employed a master computer interfaced with all of the navigation sensors on the aircraft. A common control display unit (CDU) interfaced with the master computer would provide the pilot with a single control point for all navigation systems, thereby reducing the number of required flight deck panels. Management of the various individual sensors would be transferred from the pilot to the new computer.

Since navigation sensors rarely agree exactly about position, Naimer believed that blending all available sensor position data through a highly sophisticated, mathematical filtering system would produce a more accurate aircraft position. He called the process output the "Best Computed Position." By using all available sensors to keep track of position, the system could readily provide area navigation capability. The master computer, not the individual sensors, would be integrated into the airplane, greatly reducing wiring complexity.

To solve the problems of manual waypoint entry, a pre-loaded database of global navigation information would be readily accessible by the pilot through the CDU. Using such a system a pilot could quickly and accurately construct a flight plan consisting of dozens of waypoints, avoiding the tedious typing of data and the error potential of latitude/longitude coordinates. Rather than simply navigating point-to-point, the master system would be able to maneuver the aircraft, permitting use of the system for terminal procedures including departures, arrivals, and approaches. The system would be able to automate any aspect of manual pilot navigation of the aircraft. When the first system, called the UNS-1, was released by Universal in 1982, it was called a flight management system (FMS). *[Figure 3-44]*

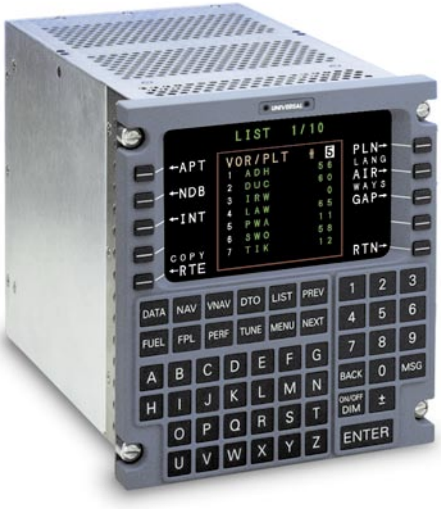
UNIVERSAL
LIST 1/10
←APT
←NDB
←INT
COPY
←RTE
VOR/PLT
PLN→
LANG
AIR→
WAYS
GAP→
RTN→
DATA
NAV
VNAV
DTO
LIST
PREV
FUEL
FPL
PERF
TUNE
MENU
NEXT
BACK
MSG
ON/OFF DIM
ENTER

Figure 3-44. *A Control Display Unit (CDU) Used to Control the Flight Management System.*

An FMS uses an electronic database of worldwide navigational data including navigation aids, airways and intersections, Standard Instrument Departures (SIDs), Standard Terminal Arrival Routes (STARs), and Instrument Approach Procedures (IAPs) together with pilot input through a CDU to create a flight plan. The FMS provides outputs to several aircraft systems including desired track, bearing and distance to the active waypoint, lateral course deviation and related data to the flight guidance system for the HSI displays, and roll steering command for the autopilot/flight director system. This allows outputs from the FMS to command the airplane where to go and when and how to turn. To support adaptation to numerous aircraft types, an FMS is usually capable of receiving and outputting both analog and digital data and discrete information. Currently, electronic navigation databases are updated every 28 days.

The introduction of the Global Positioning System (GPS) has provided extremely precise position at low cost, making GPS the dominant FMS navigation sensor today. Currently, typical FMS installations require that air data and heading information be available electronically from the aircraft. This limits FMS usage in smaller aircraft, but emerging technologies allow this data from increasingly smaller and less costly systems.

Some systems interface with a dedicated Distance Measuring Equipment (DME) receiver channel under the control of the FMS to provide an additional sensor. In these systems, the FMS determines which DME sites should be interrogated for distance information using aircraft position and the navigation database to locate appropriate DME sites. The FMS then compensates aircraft altitude and station altitude with the aid of the database to determine the precise distance to the station. With the distances from a number of sites the FMS can compute a position nearly as accurately as GPS.

Aimer visualized three-dimensional aircraft control with an FMS. Modern systems provide Vertical Navigation (VNAV) as well as Lateral Navigation (LNAV) allowing the pilot to create a vertical flight profile synchronous with the lateral flight plan. Unlike early systems, such as Inertial Reference Systems (IRS) that were only suitable for en route navigation, the modern FMS can guide an aircraft during instrument approaches.

Today, an FMS provides not only real-time navigation capability but typically interfaces with other aircraft systems providing fuel management, control of cabin briefing and display systems, display of uplinked text and graphic weather data and air/ground data link communications.

Electronic Flight Instrument Systems

Modern technology has introduced into aviation a new method of displaying flight instruments, such as electronic flight instrument systems, integrated flight deck displays, and others. For the purpose of the practical test standards, any flight instrument display that utilizes LCD or picture tube like displays is referred to as "electronic flight instrument display" and/or a glass flight deck. In general aviation there is typically a primary flight display (PFD) and a multi-function display (MFD). Although both displays are in many cases identical, the PFD provides the pilot instrumentation necessary for flight to include altitude, airspeed, vertical velocity, attitude, heading and trim and trend information.

Glass flight decks (a term coined to describe electronic flight instrument systems) are becoming more widespread as cost falls and dependability continually increases. These systems provide many advantages such as being lighter, more reliable, no moving parts to wear out, consuming less power, and replacing numerous mechanical indicators with a single glass display. Because the versatility offered by glass displays is much greater than that offered by analog displays, the use of such systems will only increase with time until analog systems are eclipsed.

Primary Flight Display (PFD)

PFDs provide increased situational awareness to the pilot by replacing the traditional six instruments used for instrument flight with an easy-to-scan display that provides the horizon, airspeed, altitude, vertical speed, trend, trim, rate of turn among other key relevant indications. Examples of PFDs are illustrated in *Figure 3-45*.

Synthetic Vision

Synthetic vision provides a realistic depiction of the aircraft in relation to terrain and flight path. Systems such as those produced by Chelton Flight Systems, Universal Flight Systems, and others provide for depictions of terrain and course. *Figure 3-46* is an example of the Chelton Flight System providing both 3-dimensional situational awareness and a synthetic highway in the sky, representing the desired flight path. Synthetic vision is used as a PFD, but provides guidance in a more normal, outside reference format.

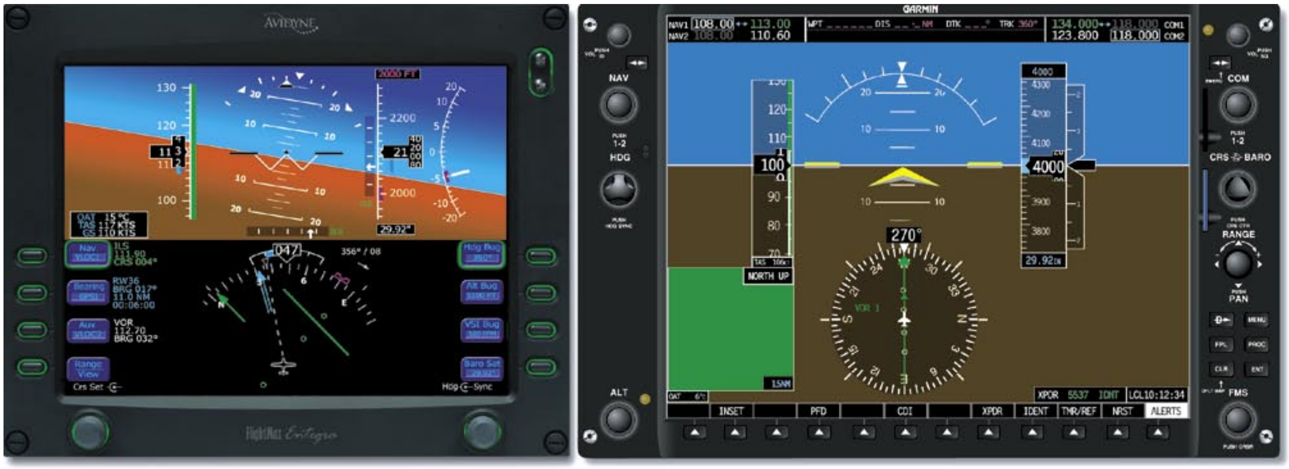

Figure 3-45. *Two Primary Flight Displays (Avidyne on the Left and Garmin on the Right).*

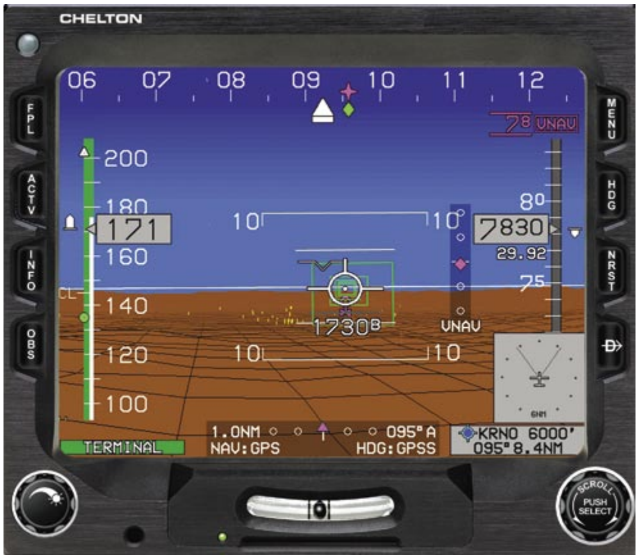

Figure 3-46. *The benefits of realistic visualization imagery, as illustrated by Synthetic Vision manufactured by Chelton Flight Systems. The system provides the pilot a realistic, real-time, three-dimensional depiction of the aircraft and its relation to terrain around it.*

Multi-Function Display (MFD)

In addition to a PFD directly in front of the pilot, an MFD that provides the display of information in addition to primary flight information is used within the flight deck. *[Figure 3-47]* Information such as a moving map, approach charts, Terrain Awareness Warning System, and weather depiction can all be illustrated on the MFD. For additional redundancy both the PFD and MFD can display all critical information that the other normally presents thereby providing redundancy (using a reversionary mode) not normally found in general aviation flight decks.

Advanced Technology Systems

Automatic Dependent Surveillance—Broadcast (ADS-B)

Although standards for Automatic Dependent Surveillance (Broadcast) (ADS-B) are still under continuing development, the concept is simple: aircraft broadcast a message on a regular basis, which includes their position (such as latitude, longitude and altitude), velocity, and possibly other information. Other aircraft or systems can receive this information for use in a wide variety of applications. The key to ADS-B is GPS, which provides three-dimensional position of the aircraft.

As an simplified example, consider air-traffic radar. The radar measures the range and bearing of an aircraft. The bearing is measured by the position of the rotating radar antenna when it receives a reply to its interrogation from the aircraft, and the range by the time it takes for the radar to receive the reply.

An ADS-B based system, on the other hand, would listen for position reports broadcast by the aircraft. *[Figure 3-48]* These position reports are based on satellite navigation systems. These transmissions include the transmitting aircraft's position, which the receiving aircraft processes into usable pilot information. The accuracy of the system is now determined by the accuracy of the navigation system, not measurement errors. Furthermore the accuracy is unaffected by the range to the aircraft as in the case of radar. With radar, detecting aircraft speed changes require tracking the data and changes can only be detected over a period of several position updates. With ADS-B, speed changes are broadcast almost instantaneously and received by properly equipped aircraft.

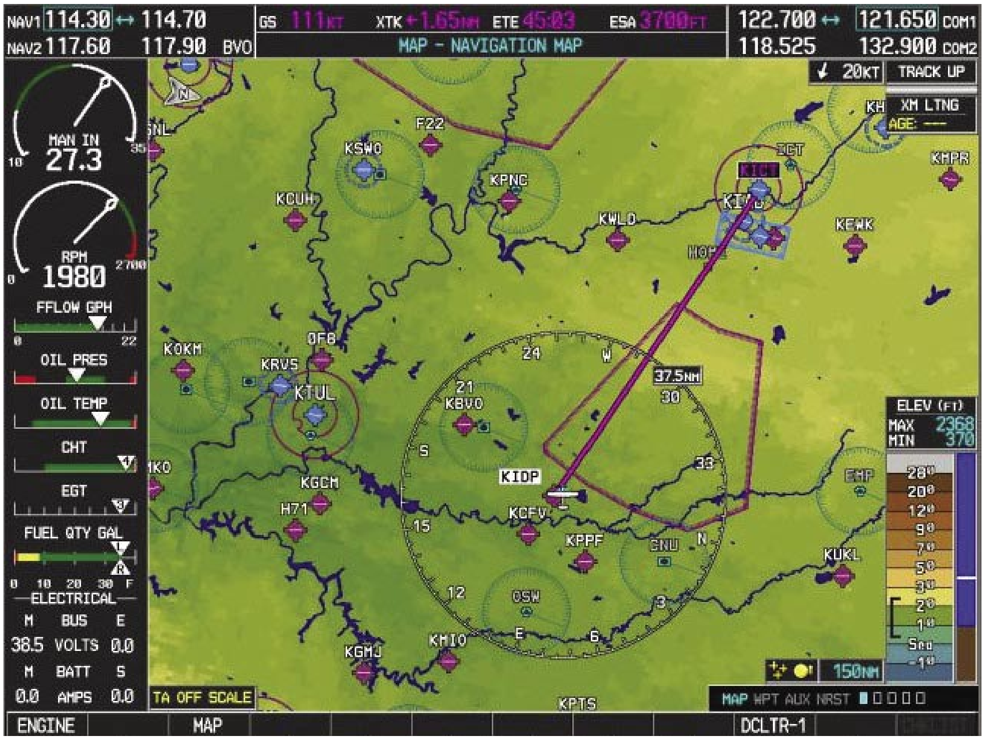
NAV1 114.30 ↔ 114.70
NAV2 117.60 117.90 BVO
GS 111KT XTK 1.65NM ETE 45:03 ESA 3700FT
MAP - NAVIGATION MAP
122.700 ↔ 121.650 COM1
118.525 132.900 COM2
20KT TRACK UP
XM LTNG
AGE: ---
MAN IN
27.3
RPM
1980
FFLOW GPH
OIL PRES
OIL TEMP
CHT
EGT
FUEL QTY GAL
ELECTRICAL
M BUS E
38.5 VOLTS 0.0
M BATT S
0.0 AMPS 0.0
KSWO
F22
KCUH
KPNC
KWLD
KICT
KEWK
KMPR
KOKM
KRVS
0F8
KTUL
KBVO
KIDP
37.5NM
KGCM
H71
KCFV
KPPF
KUKL
KMIO
KGMJ
KPTS
ELEV (FT)
MAX 2368
MIN 370
150NM
TA OFF SCALE
MAP WPT AUX NRST
ENGINE
MAP
DCLTR-1

Figure 3-47. *Example of a Multi-Function Display (MFD).*

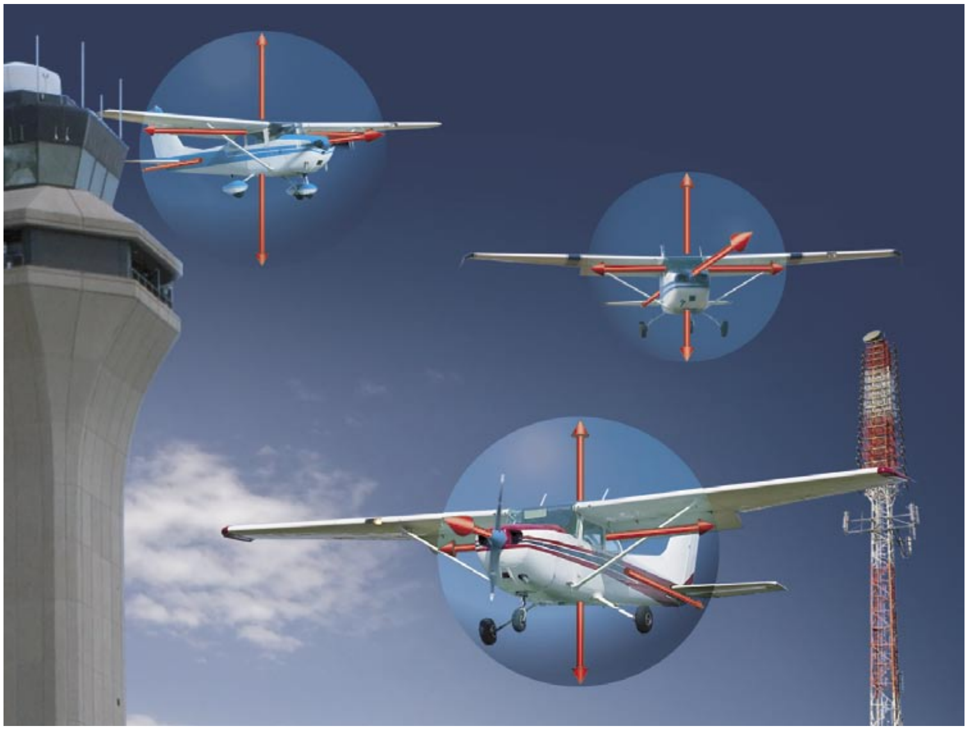

Figure 3-48. *Aircraft equipped with Automatic Dependent Surveillance—Broadcast (ADS-B) continuously broadcast their identification, altitude, direction, and vertical trend. The transmitted signal carries significant information for other aircraft and ground stations alike. Other ADS-equipped aircraft receive this information and process it in a variety of ways. It is possible that in a saturated environment (assuming all aircraft are ADS equipped), the systems can project tracks for their respective aircraft and retransmit to other aircraft their projected tracks, thereby enhancing collision avoidance. At one time, there was an Automatic Dependent Surveillance—Addressed (ADS-A) and that is explained in the Pilot's Handbook of Aeronautical Knowledge.*

Additionally, other information can be obtained by properly equipped aircraft to include notices to airmen (NOTAM), weather, etc. *[Figures 3-49* and *3-50]* At the present time, ADS-B is predominantly available along the east coast of the United States where it is matured.

Safety Systems

Radio Altimeters

A radio altimeter, commonly referred to as a radar altimeter, is a system used for accurately measuring and displaying the height above the terrain directly beneath the aircraft. It sends a signal to the ground and processes the timed information.

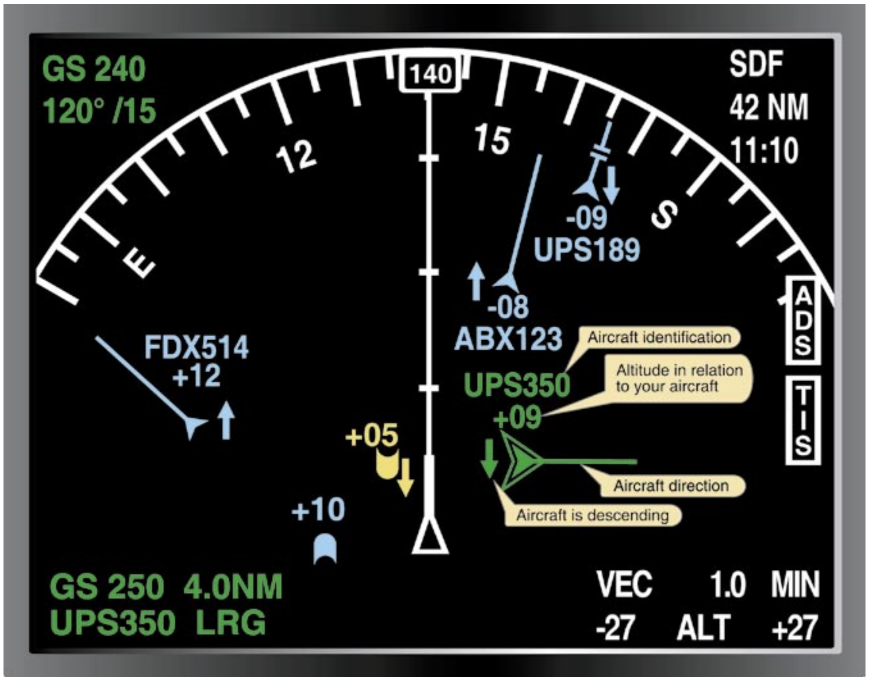
GS 240
120° /15
140
12
15
E
S
SDF
42 NM
11:10
-09
UPS189
-08
ABX123
Aircraft identification
Altitude in relation to your aircraft
UPS350
+09
Aircraft direction
Aircraft is descending
FDX514
+12
+05
+10
ADS
TIS
GS 250 4.0NM
UPS350 LRG
VEC 1.0 MIN
-27 ALT +27

Figure 3-49. *An aircraft equipped with ADS will receive identification, altitude in hundreds of feet (above or below using + or -), direction of the traffic, and aircraft descent or climb using an up or down arrow. The yellow target is an illustration of how a non-ADS equipped aircraft would appear on an ADS-equipped aircraft's display.*

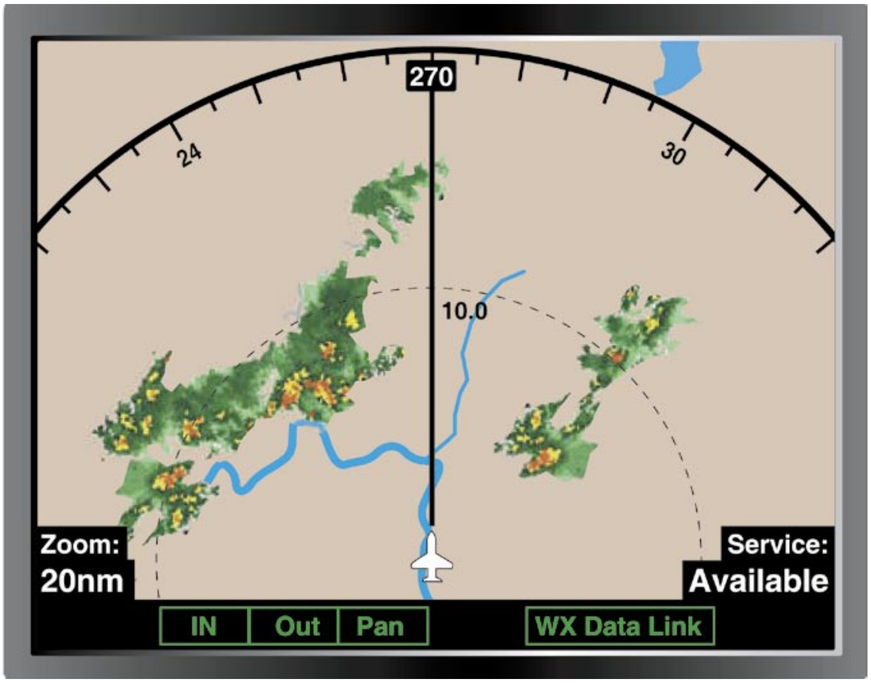
270
24
30
10.0
Zoom:
20nm
Service:
Available
IN
Out
Pan
WX Data Link

Figure 3-50. *An aircraft equipped with ADS has the ability to upload and display weather.*

Its primary application is to provide accurate absolute altitude information to the pilot during approach and landing. In advanced aircraft today, the radar altimeter also provides its information to other onboard systems such as the autopilot and flight directors while they are in the glide slope capture mode below 200-300 feet above ground level (AGL).

A typical system consists of a receiver-transmitter (RT) unit, antenna(s) for receiving and transmitting the signal, and an indicator. *[Figure 3-51]* Category II and III precision approach procedures require the use of a radar altimeter and specify the exact minimum height above the terrain as a decision height (DH) or radio altitude (RA).

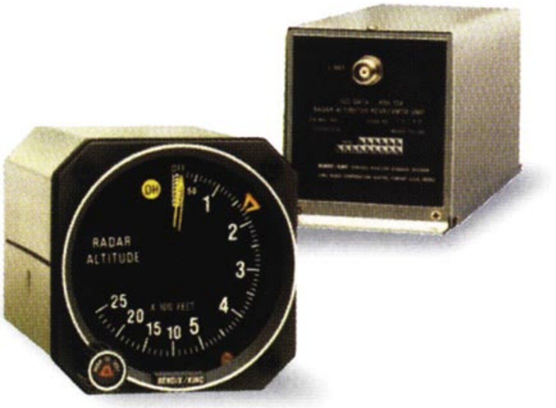

Figure 3-51. *Components of a Radar Altimeter.*

Traffic Advisory Systems

Traffic Information System

The Traffic Information Service (TIS) is a ground-based service providing information to the flight deck via data link using the S-mode transponder and altitude encoder. TIS improves the safety and efficiency of "see and avoid" flight through an automatic display that informs the pilot of nearby traffic. The display can show location, direction, altitude and the climb/descent trend of other transponder-equipped aircraft. TIS provides estimated position, altitude, altitude trend, and ground track information for up to several aircraft simultaneously within about 7 NM horizontally, 3,500 feet above and 3,500 feet below the aircraft. *[Figure 3-52]* This data can be displayed on a variety of MFDs. *[Figure 3-53]*

Figure 3-54 displays the pictorial concept of the traffic information system. Noteworthy is the requirement to have Mode S and that the ground air traffic station processes the Mode S signal.

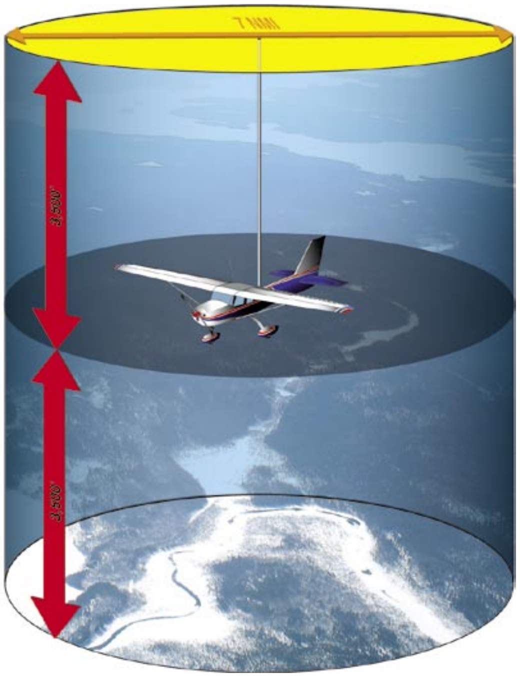

Figure 3-52. *Coverage Provided by a Traffic Information System.*

Traffic Alert Systems

Traffic alert systems receive transponder information from nearby aircraft to help determine their relative position to the equipped aircraft. They provide three-dimensional location of other aircraft *[Figures 3-55, 3-56,* and *3-57]* and are cost effective alternatives to TCAS equipage for smaller aircraft.

Traffic Avoidance Systems

Traffic Alert and Collision Avoidance System (TCAS)

The TCAS is an airborne system developed by the FAA that operates independently from the ground-based ATC system. TCAS was designed to increase flight deck awareness of proximate aircraft and to serve as a "last line of defense" for the prevention of mid-air collisions.

There are two levels of TCAS systems. TCAS I was developed to accommodate the general aviation (GA) community and the regional airlines. This system issues traffic advisories (TAs) to assist pilots in visual acquisition of intruder aircraft. TCAS I provides approximate bearing and relative altitude of aircraft with a selectable range. It provides the pilot with traffic advisory (TA) alerting him or her to potentially conflicting traffic. The pilot then visually acquires the traffic and takes appropriate action for collision avoidance.

TCAS II is a more sophisticated system which provides the same information of TCAS I. It also analyzes the projected flight path of approaching aircraft and issues resolution advisories (RAs) to the pilot to resolve potential mid-air collisions. Additionally, if communicating with another TCAS II equipped aircraft, the two systems coordinate the resolution alerts provided to their respective flight crews. *[Figure 3-58]*

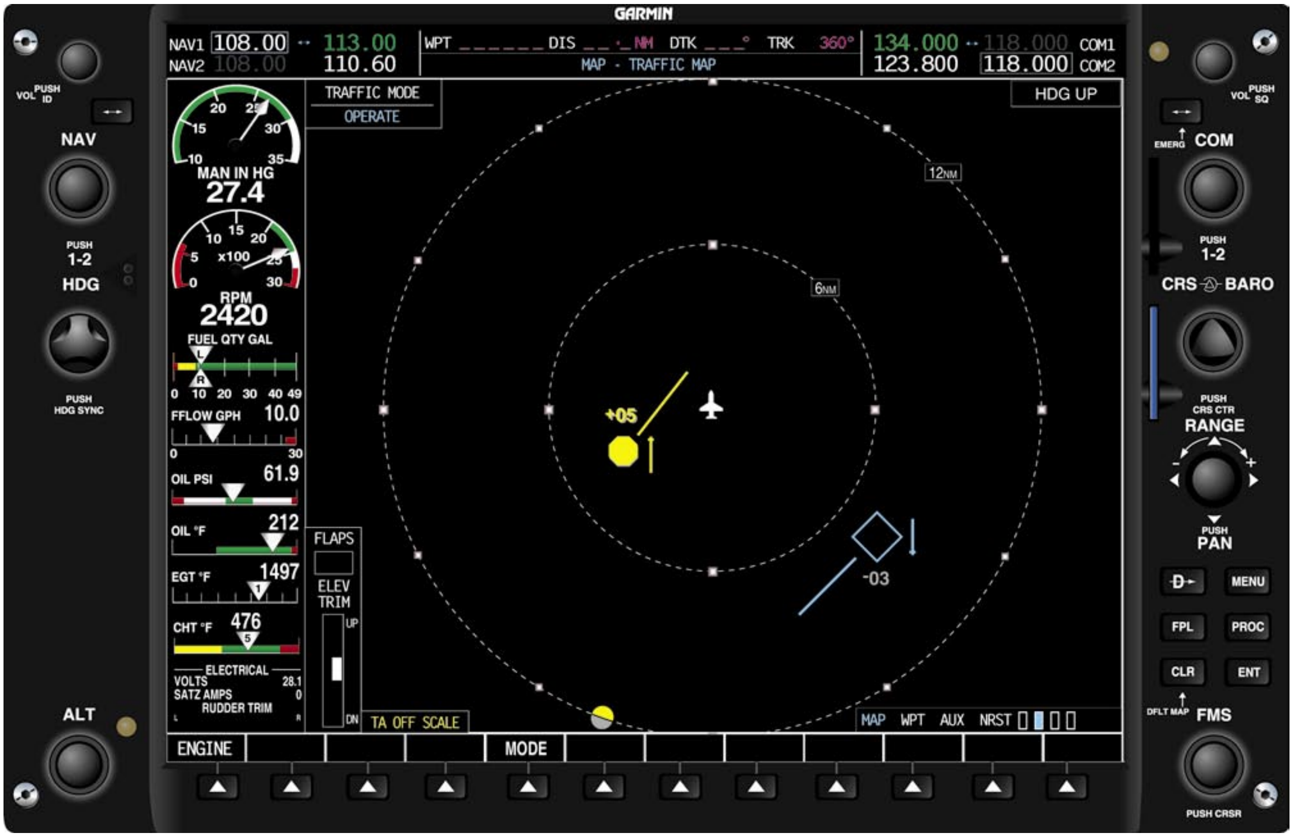
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK ° TRK 360°
MAP - TRAFFIC MAP
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
TRAFFIC MODE
OPERATE
HDG UP
12NM
6NM
+05
-03
MAN IN HG
27.4
RPM
2420
FUEL QTY GAL
FFLOW GPH 10.0
OIL PSI 61.9
OIL °F 212
EGT °F 1497
CHT °F 476
ELECTRICAL
VOLTS 28.1
SATZ AMPS 0
RUDDER TRIM
FLAPS
ELEV TRIM
UP
DN
TA OFF SCALE
MAP WPT AUX NRST
ENGINE
MODE
NAV
HDG
ALT
PUSH HDG SYNC
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR

Figure 3-53. *Multi-Function Display (MFD).*

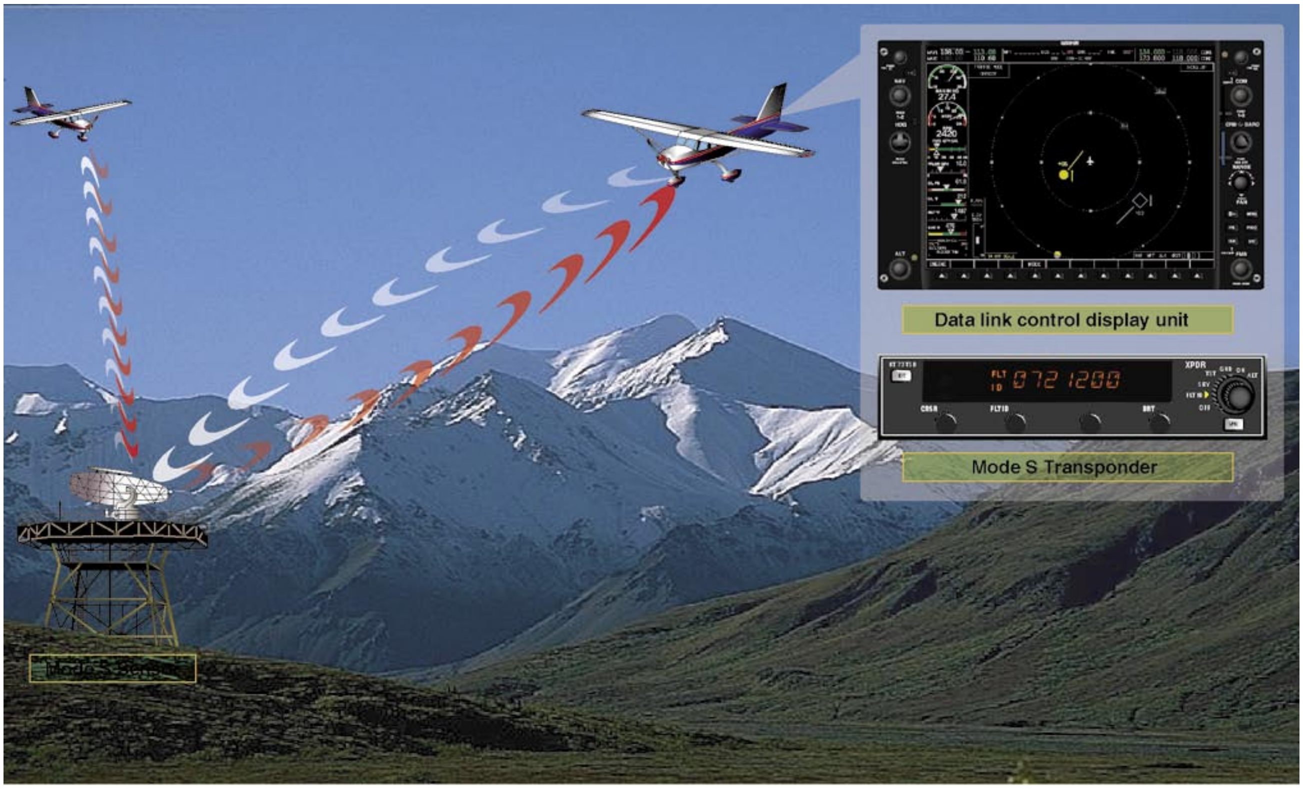
Data link control display unit
Mode S Transponder
Mode S Sensor

Figure 3-54. *Concept of the Traffic Information System.*

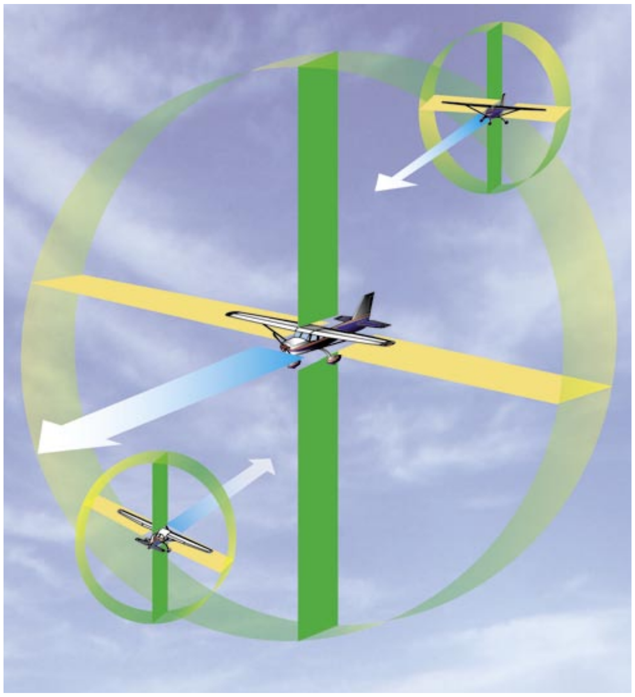

Figure 3-55. *Theory of a Typical Alert System.*

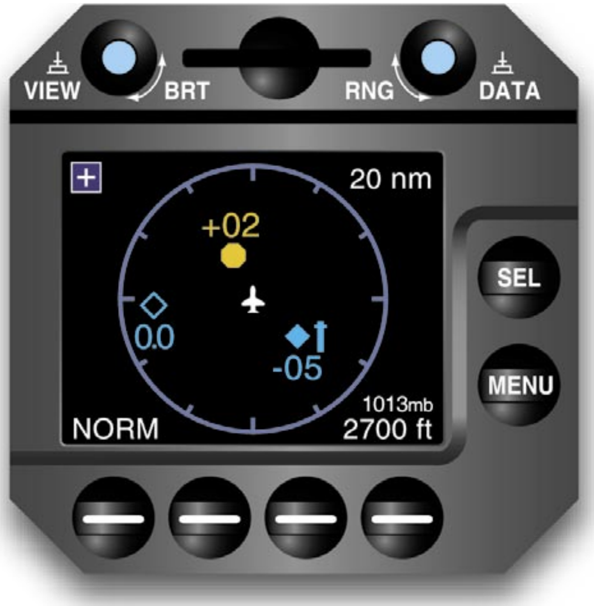
VIEW
BRT
RNG
DATA
20 nm
+02
0.0
-05
NORM
1013mb
2700 ft
SEL
MENU

Figure 3-56. *A Skywatch System.*

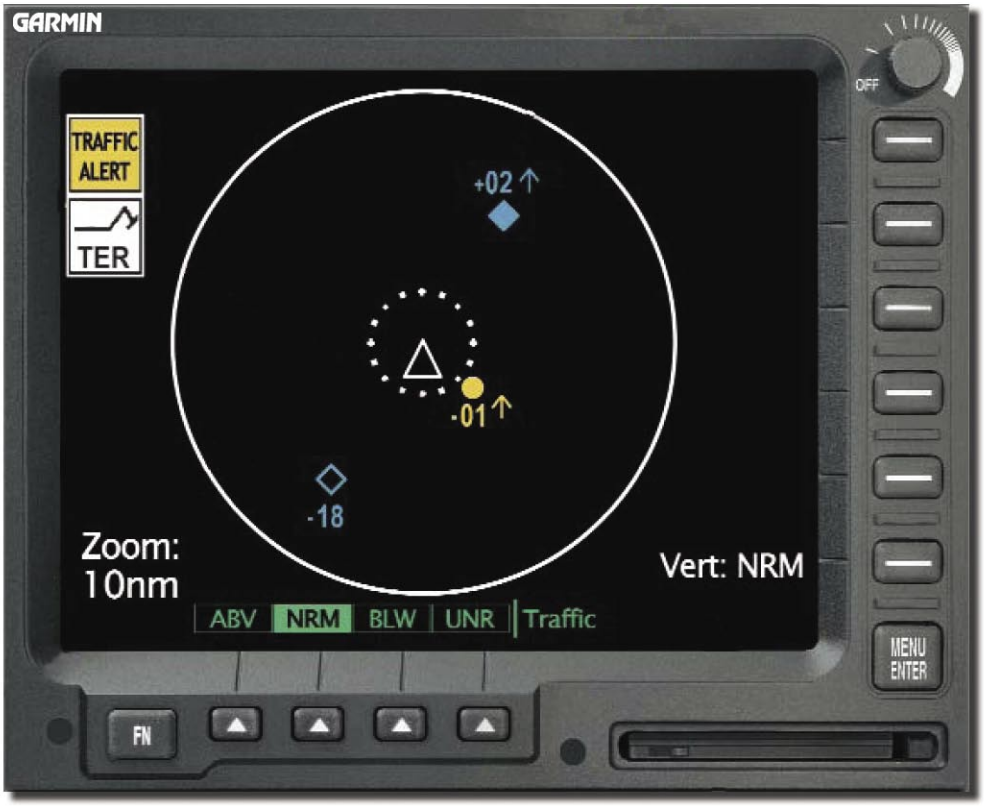
GARMIN
OFF
TRAFFIC ALERT
TER
+02
-01
-18
Zoom: 10nm
Vert: NRM
ABV
NRM
BLW
UNR
Traffic
MENU ENTER
FN

Figure 3-57. *Alert System by Avidyne (Ryan).*

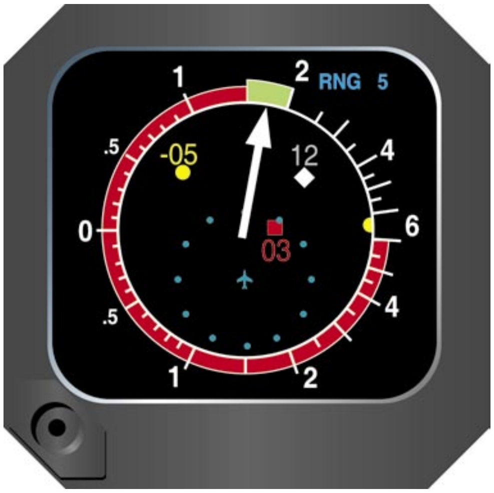

Figure 3-58. *An example of a resolution advisory being provided the pilot. In this case, the pilot is requested to climb, with 1,200 feet being the appropriate rate of ascent to avoid traffic conflict. This visual indication plus the aural warning provide the pilot with excellent traffic awareness that augments see and avoid practices.*

Terrain Alerting Systems

Ground Proximity Warning System (GPWS)

An early application of technology to reduce CFIT was the GPWS. In airline use since the early 1970s, GPWS uses the radio altimeter, speed, and barometric altitude to determine the aircraft's position relative to the ground. The system uses this information in determining aircraft clearance above the Earth and provides limited predictability about aircraft position relative to rising terrain. It does this based upon algorithms within the system and developed by the manufacturer for different airplanes or helicopters. However, in mountainous areas the system is unable to provide predictive information due to the unusual slope encountered.

This inability to provide predictive information was evidenced in 1999 when a DH-7 crashed in South America. The crew had a GPWS onboard, but the sudden rise of the terrain rendered it ineffective; the crew continued unintentionally into a mountain with steep terrain. Another incident involved Secretary of Commerce Brown who, along with all on board, was lost when the crew flew over rapidly rising terrain where the GPWS capability is offset by terrain gradient. However, the GPWS is tied into and considers landing gear status, flap position, and ILS glide slope deviation to detect unsafe aircraft operation with respect to terrain, excessive descent rate, excessive closure rate to terrain, unsafe terrain clearance while not in a landing configuration, excessive deviation below an ILS glide slope. It also provides advisory callouts.

Generally, the GPWS is tied into the hot bus bar of the electrical system to prevent inadvertent switch off. This was demonstrated in an accident involving a large four-engine turboprop airplane. While on final for landing with the landing gear inadvertently up, the crew failed to heed the GPWS warning as the aircraft crossed a large berm close to the threshold. In fact, the crew attempted without success to shut the system down and attributed the signal to a malfunction. Only after the mishap did the crew realize the importance of the GPWS warning.

Terrain Awareness and Warning System (TAWS)

A TAWS uses GPS positioning and a database of terrain and obstructions to provide true predictability of the upcoming terrain and obstacles. The warnings it provides pilots are both aural and visual, instructing the pilot to take specific action. Because TAWS relies on GPS and a database of terrain/obstacle information, predictability is based upon aircraft location and projected location. The system is time based and therefore compensates for the performance of the aircraft and its speed. *[Figure 3-59]*

Head-Up Display (HUD)

The HUD is a display system that provides a projection of navigation and air data (airspeed in relation to approach reference speed, altitude, left/right and up/down glide slope) on a transparent screen between the pilot and the windshield. The concept of a HUD is to diminish the shift between looking at the instrument panel and outside. Virtually any information desired can be displayed on the HUD if it is available in the aircraft's flight computer. The display for the HUD can be projected on a separate panel near the windscreen or as shown in *Figure 3-60* on an eye piece. Other information may be displayed, including a runway target in relation to the nose of the aircraft, which allows the pilot to see the information necessary to make the approach while also being able to see out the windshield.

Required Navigation Instrument System Inspection

Systems Preflight Procedures

Inspecting the instrument system requires a relatively small part of the total time required for preflight activities, but its importance cannot be overemphasized. Before any flight involving aircraft control by instrument reference, the pilot should check all instruments and their sources of power for proper operation. NOTE: The following procedures are appropriate for conventional aircraft instrument systems. Aircraft equipped with electronic instrument systems utilize different procedures.

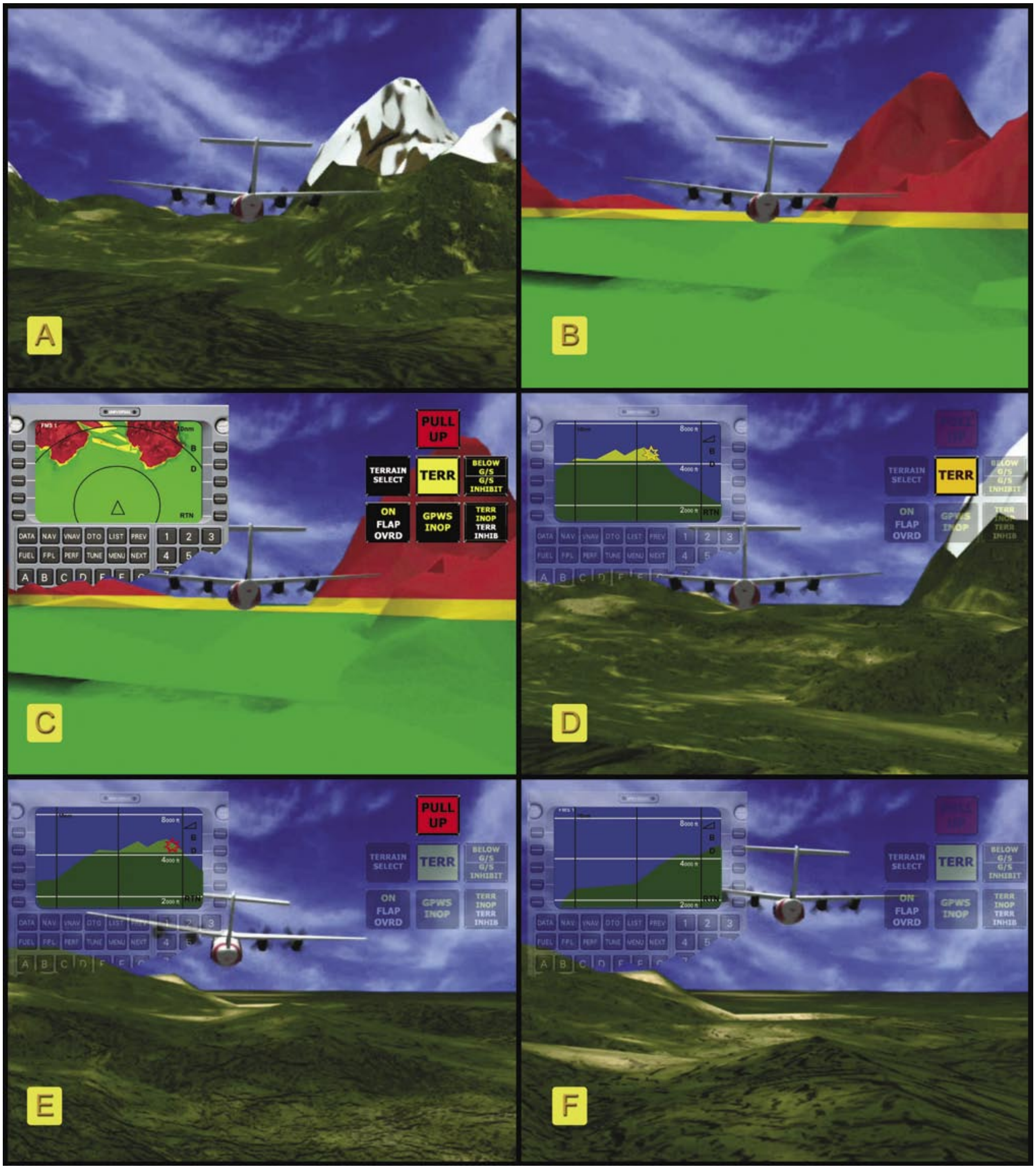
A
B
PULL UP
TERRAIN SELECT
TERR
BELOW G/S G/S INHIBIT
ON FLAP OVRD
GPWS INOP
TERR INOP TERR INHIB
DATA NAV VNAV DTO LIST PREV
FUEL FPL PERF TUNE MENU NEXT
1 2 3 4 5
RTN
C
8000 ft
4000 ft
2000 ft
D
E
F

Figure 3-59. *A six-frame sequence illustrating the manner in which TAWS operates. A TAWS installation is aircraft specific and provides warnings and cautions based upon time to potential impact with terrain rather than distance. The TAWS is illustrated in an upper left window while aircrew view is provided out of the windscreen.* A *illustrates the aircraft in relation to the outside terrain while* B *and* C *illustrate the manner in which the TAWS system displays the terrain.* D *is providing a caution of terrain to be traversed, while* E *provides an illustration of a warning with an aural and textural advisory (red) to pull up.* E *also illustrates a pilot taking appropriate action (climb in this case) while* F *illustrates that a hazard is no longer a factor.*

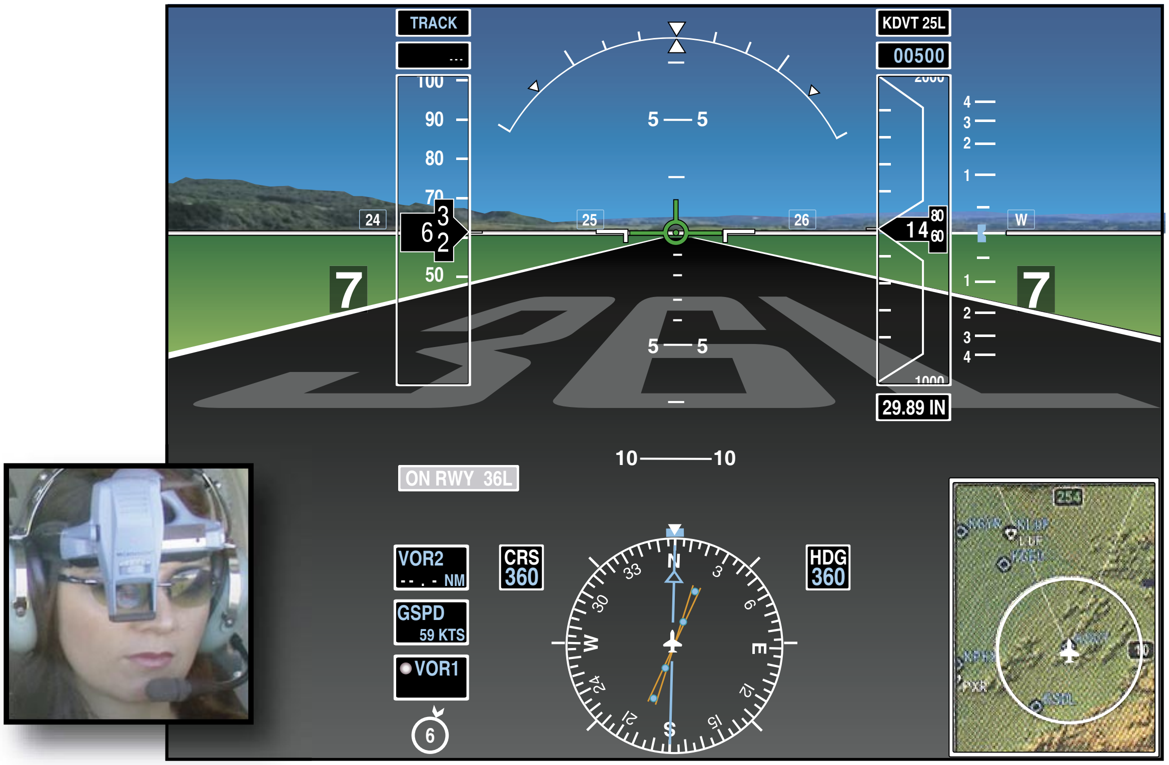

Figure 3-60. *A Head-Up Display.*

Before Engine Start

1. Walk-around inspection: Check the condition of all antennas and check the pitot tube for the presence of any obstructions and remove the cover. Check the static ports to be sure they are free from dirt and obstructions, and ensure there is nothing on the structure near the ports that would disturb the air flowing over them.
2. Aircraft records: Confirm that the altimeter and static system have been checked and found within approved limits within the past 24 calendar months. Check the replacement date for the emergency locator transmitter (ELT) batteries noted in the maintenance record, and be sure they have been replaced within this time interval.
3. Preflight paperwork: Check the Airport/Facility Directory (A/FD) and all Notices to Airmen (NOTAMs) for the condition and frequencies of all the navigation aid (NAVAIDs) that are used on the flight. Handbooks, en route charts, approach charts, computer and flight log should be appropriate for the departure, en route, destination, and alternate airports.
4. Radio equipment: Switches off.
5. Suction gauge: Proper markings as applicable if electronic flight instrumentation is installed.
6. ASI: Proper reading, as applicable. If electronic flight instrumentation is installed, check emergency instrument.
7. Attitude indicator: Uncaged, if applicable. If electronic flight instrumentation is installed, check emergency system to include its battery as appropriate.
8. Altimeter: Set the current altimeter setting and ensure that the pointers indicate the elevation of the airport.
9. VSI: Zero indication, as applicable (if electronic flight instrumentation is installed).
10. Heading indicator: Uncaged, if applicable.
11. Turn coordinator: If applicable, miniature aircraft level, ball approximately centered (level terrain).
12. Magnetic compass: Full of fluid and the correction card is in place and current.
13. Clock: Set to the correct time and running.
14. Engine instruments: Proper markings and readings, as applicable if electronic flight instrumentation is installed.
15. Deicing and anti-icing equipment: Check availability and fluid quantity.
16. Alternate static-source valve: Be sure it can be opened if needed, and that it is fully closed.

17. Pitot tube heater: Check by watching the ammeter when it is turned on, or by using the method specified in the POH/AFM.

After Engine Start

1. When the master switch is turned on, listen to the gyros as they spin up. Any hesitation or unusual noises should be investigated before flight.
2. Suction gauge or electrical indicators: Check the source of power for the gyro instruments. The suction developed should be appropriate for the instruments in that particular aircraft. If the gyros are electrically driven, check the generators and inverters for proper operation.
3. Magnetic compass: Check the card for freedom of movement and confirm the bowl is full of fluid. Determine compass accuracy by comparing the indicated heading against a known heading (runway heading) while the airplane is stopped or taxiing straight. Remote indicating compasses should also be checked against known headings. Note the compass card correction for the takeoff runway heading.
4. Heading indicator: Allow 5 minutes after starting engines for the gyro to spin up. Before taxiing, or while taxiing straight, set the heading indicator to correspond with the magnetic compass heading. A slaved gyrocompass should be checked for slaving action and its indications compared with those of the magnetic compass. If an electronic flight instrument system is installed, consult the flight manual for proper procedures.
5. Attitude indicator: Allow the same time as noted above for gyros to spin up. If the horizon bar erects to the horizontal position and remains at the correct position for the attitude of the airplane, or if it begins to vibrate after this attitude is reached and then slowly stops vibrating altogether, the instrument is operating properly. If an electronic flight instrument system is installed, consult the flight manual for proper procedures.
6. Altimeter: With the altimeter set to the current reported altimeter setting, note any variation between the known field elevation and the altimeter indication. If the indication is not within 75 feet of field elevation, the accuracy of the altimeter is questionable and the problem should be referred to a repair station for evaluation and possible correction. Because the elevation of the ramp or hangar area might differ significantly from field elevation, recheck when in the run-up area if the error exceeds 75 feet. When no altimeter setting is available, set the altimeter to the published field elevation during the preflight instrument check.
7. VSI: The instrument should read zero. If it does not, tap the panel gently. If an electronic flight instrument system is installed, consult the flight manual for proper procedures.
8. Engine instruments: Check for proper readings.
9. Radio equipment: Check for proper operation and set as desired.
10. Deicing and anti-icing equipment: Check operation.

Taxiing and Takeoff

1. Turn coordinator: During taxi turns, check the miniature aircraft for proper turn indications. The ball or slip/skid should move freely. The ball or slip/skid indicator should move opposite to the direction of turns. The turn instrument should indicate the direction of the turn. While taxiing straight, the miniature aircraft (as appropriate) should be level.
2. Heading indicator: Before takeoff, recheck the heading indicator. If the magnetic compass and deviation card are accurate, the heading indicator should show the known taxiway or runway direction when the airplane is aligned with them (within 5°).
3. Attitude indicator: If the horizon bar fails to remain in the horizontal position during straight taxiing, or tips in excess of 5° during taxi turns, the instrument is unreliable. Adjust the miniature aircraft with reference to the horizon bar for the particular airplane while on the ground. For some tricycle-gear airplanes, a slightly nose-low attitude on the ground gives a level flight attitude at normal cruising speed.

Engine Shut Down

When shutting down the engine, note any abnormal instrument indications.

FEET
100
ALT
CALIBRATED
29.8
29.9
30.0
UP VERTICAL SPEED
THOUSAND FT PER MIN
DOWN
L
R
OFF
2 MIN TURN
DC ELEC
STBY PWR
TEST
PULL TO CAGE

Chapter 4, Section I

Airplane Attitude Instrument Flying

Using Analog Instrumentation

Introduction

Attitude instrument flying is defined as the control of an aircraft's spatial position by using instruments rather than outside visual references. Today's aircraft come equipped with analog and/or digital instruments. Analog instrument systems are mechanical and operate with numbers representing directly measurable quantities, such as a watch with a sweep second hand. In contrast, digital instrument systems are electronic and operate with numbers expressed in digits. Although more manufacturers are providing aircraft with digital instrumentation, analog instruments remain more prevalent. This section acquaints the pilot with the use of analog flight instruments.

Any flight, regardless of the aircraft used or route flown, consists of basic maneuvers. In visual flight, aircraft attitude is controlled by using certain reference points on the aircraft with relation to the natural horizon. In instrument flight, the aircraft attitude is controlled by reference to the flight instruments. Proper interpretation of the flight instruments provides essentially the same information that outside references do in visual flight. Once the role of each instrument in establishing and maintaining a desired aircraft attitude is learned, a pilot is better equipped to control the aircraft in emergency situations involving failure of one or more key instruments.

Learning Methods

The two basic methods used for learning attitude instrument flying are "control and performance" and "primary and supporting." Both methods utilize the same instruments and responses for attitude control. They differ in their reliance on the attitude indicator and interpretation of other instruments.

Attitude Instrument Flying Using the Control and Performance Method

Aircraft performance is achieved by controlling the aircraft attitude and power. Aircraft attitude is the relationship of both the aircraft's pitch and roll axes in relation to the Earth's horizon. An aircraft is flown in instrument flight by controlling the attitude and power, as necessary, to produce both controlled and stabilized flight without reference to a visible horizon. This overall process is known as the control and performance method of attitude instrument flying. Starting with basic instrument maneuvers, this process can be applied through the use of control, performance, and navigation instruments, resulting in a smooth flight, from takeoff to landing.

Control Instruments

The control instruments display immediate attitude and power indications and are calibrated to permit those respective adjustments in precise increments. In this discussion, the term "power" is used in place of the more technically correct term "thrust or drag relationship." Control is determined by reference to the attitude and power indicators. Power indicators vary with aircraft and may include manifold pressure, tachometers, fuel flow, etc. *[Figure 4-1]*

Performance Instruments

The performance instruments indicate the aircraft's actual performance. Performance is determined by reference to the altimeter, airspeed or vertical speed indicator (VSI). *[Figure 4-2]*

Navigation Instruments

The navigation instruments indicate the position of the aircraft in relation to a selected navigation facility or fix. This group of instruments includes various types of course indicators, range indicators, glide-slope indicators, and bearing pointers. *[Figure 4-3]* Newer aircraft with more technologically advanced instrumentation provide blended information, giving the pilot more accurate positional information.

Procedural Steps in Using Control and Performance

1. Establish an attitude and power setting on the control instruments that results in the desired performance. Known or computed attitude changes and approximated power settings helps to reduce the pilot's workload.
2. Trim (fine tune the control forces) until control pressures are neutralized. Trimming for hands-off flight is essential for smooth, precise aircraft control.

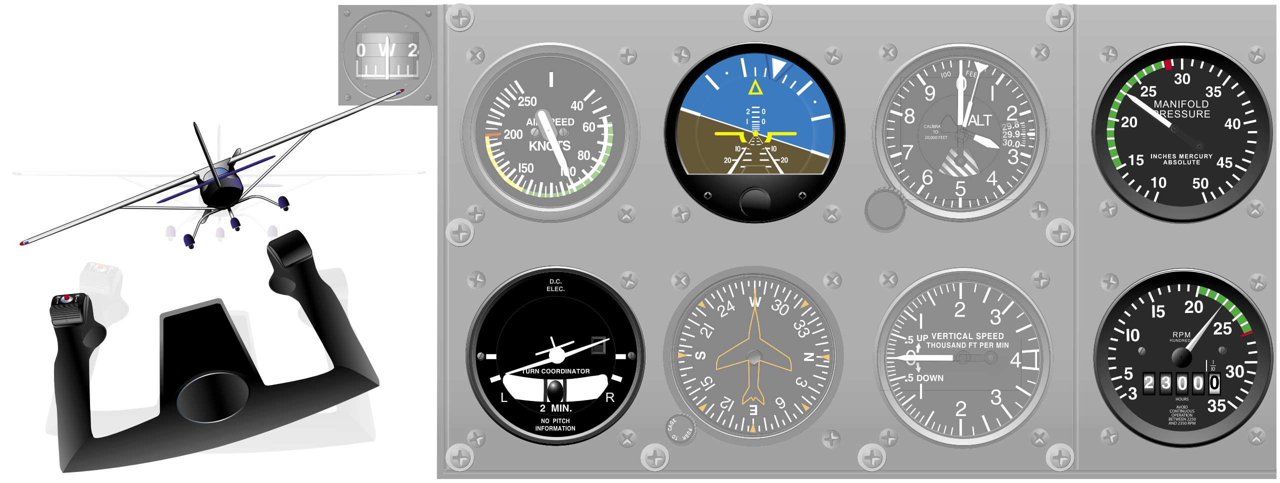

Figure 4-1. *Control Instruments.*

It allows a pilot to attend to other flight deck duties with minimum deviation from the desired attitude.

3. Cross-check the performance instruments to determine if the established attitude or power setting is providing the desired performance. The cross-check involves both seeing and interpreting. If a deviation is noted, determine the magnitude and direction of adjustment required to achieve the desired performance.

4. Adjust the attitude and/or power setting on the control instruments as necessary.

Aircraft Control During Instrument Flight

Attitude Control

Proper control of aircraft attitude is the result of proper use of the attitude indicator, knowledge of when to change the

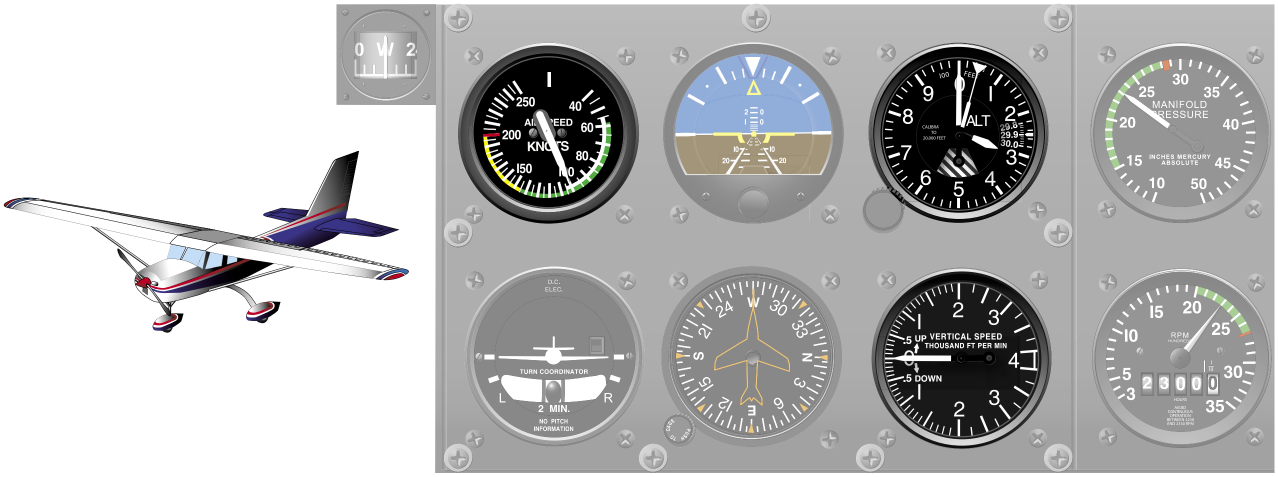
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
D.C. ELEC.
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
VERTICAL SPEED
THOUSAND FT PER MIN
UP
DOWN
RPM
HOURS

Figure 4-2. *Performance Instruments.*

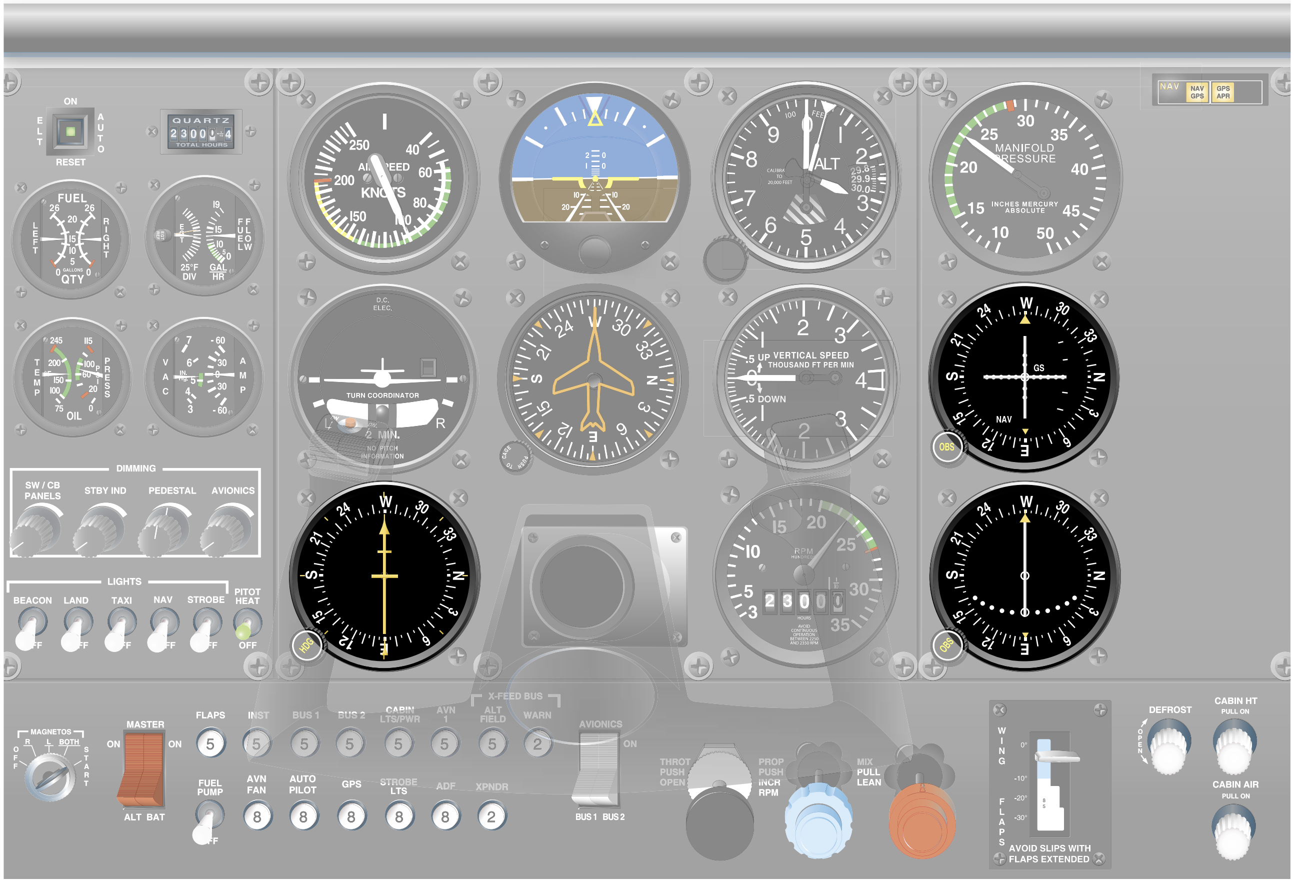
ELT
ON
AUTO
RESET
QUARTZ
TOTAL HOURS
FUEL
LEFT
RIGHT
QTY
GAL HR
FUEL FLOW
TEMP
PRESS
OIL
VAC
AMP
DIMMING
SW / CB PANELS
STBY IND
PEDESTAL
AVIONICS
LIGHTS
BEACON
LAND
TAXI
NAV
STROBE
PITOT HEAT
OFF
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
NAV
GPS
APR
D.C. ELEC.
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
VERTICAL SPEED
THOUSAND FT PER MIN
UP
DOWN
GS
OBS
HDG
RPM
MAGNETOS
OFF
R
L
BOTH
START
MASTER
ALT
BAT
FLAPS
INST
BUS 1
BUS 2
CABIN LTS/PWR
AVN 1
X-FEED BUS
ALT FIELD
WARN
AVIONICS
ON
BUS 1
BUS 2
FUEL PUMP
AVN FAN
AUTO PILOT
GPS
STROBE LTS
ADF
XPNDR
THROT PUSH OPEN
PROP PUSH INCR RPM
MIX PULL LEAN
WING FLAPS
AVOID SLIPS WITH FLAPS EXTENDED
DEFROST
OPEN
CABIN HT
PULL ON
CABIN AIR
PULL ON

Figure 4-3. *Flight Panel Instrumentation.*

attitude, and then smoothly changing the attitude a precise amount. The attitude reference provides an immediate, direct, and corresponding indication of any change in aircraft pitch or bank attitude.

Pitch Control

Changing the "pitch attitude" of the miniature aircraft or fuselage dot by precise amounts in relation to the horizon makes pitch changes. These changes are measured in degrees or fractions thereof, or bar widths depending upon the type of attitude reference. The amount of deviation from the desired performance determines the magnitude of the correction.

Bank Control

Bank changes are made by changing the "bank attitude" or bank pointers by precise amounts in relation to the bank scale. The bank scale is normally graduated at 0°, 10°, 20°, 30°, 60°, and 90° and is located at the top or bottom of the attitude reference. Normally, use a bank angle that approximates the degrees to turn, not to exceed 30°.

Power Control

Proper power control results from the ability to smoothly establish or maintain desired airspeeds in coordination with attitude changes. Power changes are made by throttle adjustments and reference to the power indicators. Power indicators are not affected by such factors as turbulence, improper trim, or inadvertent control pressures. Therefore, in most aircraft little attention is required to ensure the power setting remains constant.

Experience in an aircraft teaches a pilot approximately how far to move the throttle to change the power a given amount. Power changes are made primarily by throttle movement, followed by an indicator cross-check to establish a more precise setting. The key is to avoid fixating on the indicators while setting the power. Knowledge of approximate power settings for various flight configurations helps the pilot avoid overcontrolling power.

Attitude Instrument Flying Using the Primary and Supporting Method

Another basic method for teaching attitude instrument flying classifies the instruments as they relate to control function as well as aircraft performance. All maneuvers involve some degree of motion about the lateral (pitch), longitudinal (bank/roll), and vertical (yaw) axes. Attitude control is stressed in this handbook in terms of pitch control, bank control, power control, and trim control. Instruments are grouped as they relate to control function and aircraft performance as pitch control, bank control, power control, and trim.

Pitch Control

Pitch control is controlling the rotation of the aircraft about the lateral axis by movement of the elevators. After interpreting the pitch attitude from the proper flight instruments, exert control pressures to effect the desired pitch attitude with reference to the horizon. These instruments include the attitude indicator, altimeter, VSI, and airspeed indicator. *[Figure 4-4]* The attitude indicator displays a direct indication of the aircraft's pitch attitude while the other pitch attitude control instruments indirectly indicate the pitch attitude of the aircraft.

Attitude Indicator

The pitch attitude control of an aircraft controls the angular relationship between the longitudinal axis of the aircraft and the actual horizon. The attitude indicator gives a direct and immediate indication of the pitch attitude of the aircraft. The aircraft controls are used to position the miniature aircraft in relation to the horizon bar or horizon line for any pitch attitude required. *[Figure 4-5]*

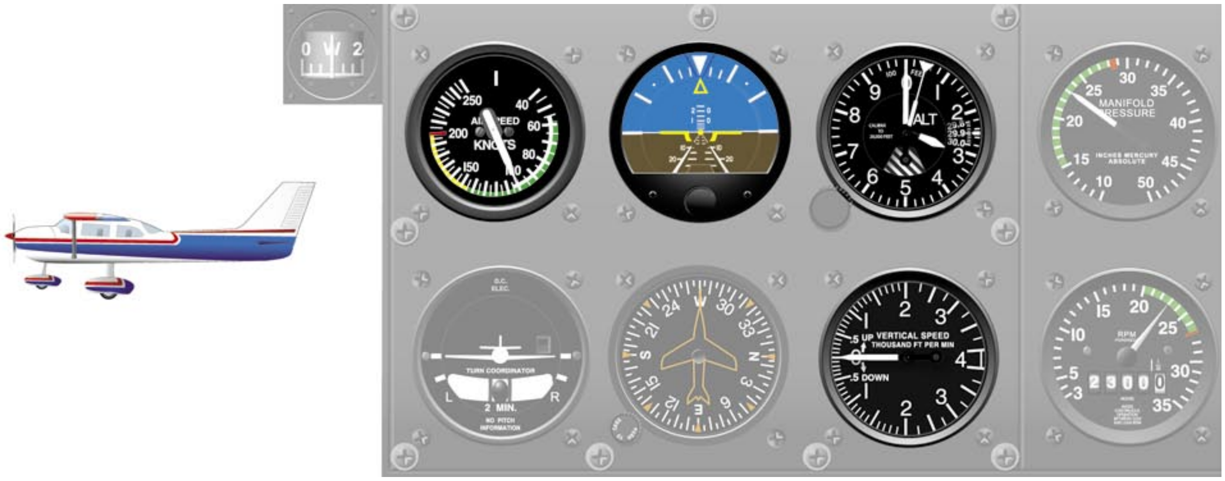

Figure 4-4. *Pitch Instruments.*

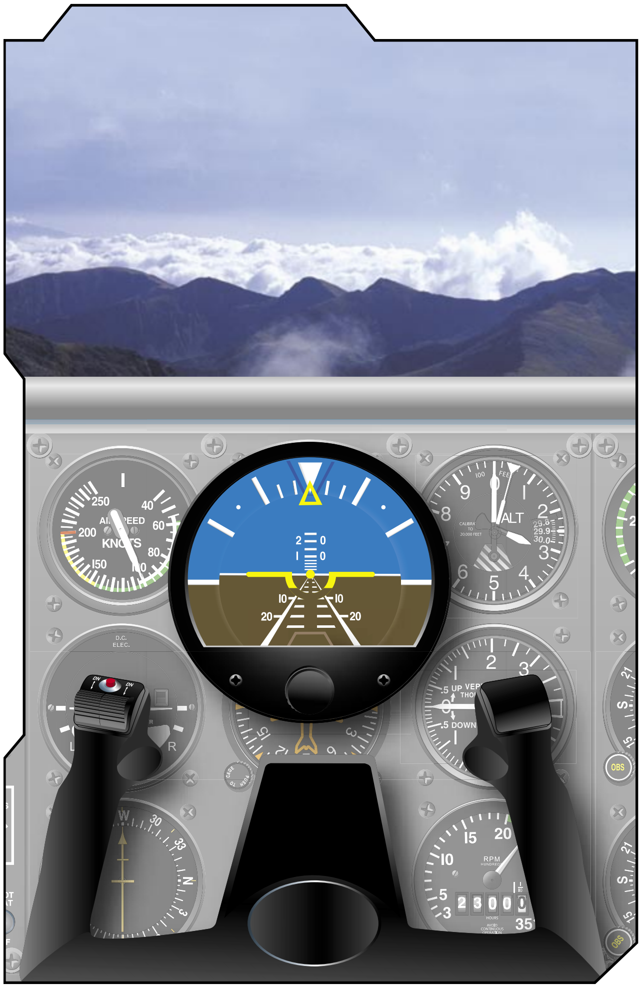

Figure 4-5. *Attitude Indicator.*

The miniature aircraft should be placed in the proper position in relation to the horizon bar or horizon line before takeoff. The aircraft operator's manual explains this position. As soon as practicable in level flight and at desired cruise airspeed, the miniature aircraft should be moved to a position that aligns its wings in front of the horizon bar or horizon line. This adjustment can be made anytime varying loads or other conditions indicate a need. Otherwise, the position of the miniature aircraft should not be changed for flight at other than cruise speed. This is to make sure that the attitude indicator displays a true picture of pitch attitude in all maneuvers.

When using the attitude indicator in applying pitch attitude corrections, control pressure should be extremely light. Movement of the horizon bar above or below the miniature aircraft of the attitude indicator in an airplane should not exceed one-half the bar width. *[Figure 4-6]* If further change is required, an additional correction of not more than one-half horizon bar wide normally counteracts any deviation from normal flight.

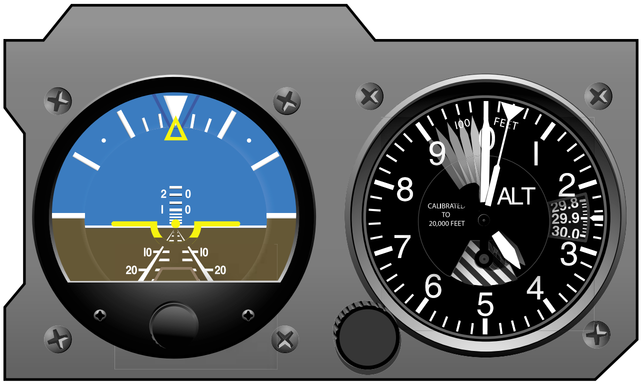

Figure 4-6. *Pitch Correction Using the Attitude Indicator.*

Altimeter

If the aircraft is maintaining level flight, the altimeter needles maintain a constant indication of altitude. If the altimeter indicates a loss of altitude, the pitch attitude must be adjusted upward to stop the descent. If the altimeter indicates a gain in altitude, the pitch attitude must be adjusted downward to stop the climb. *[Figure 4-7]* The altimeter can also indicate the pitch attitude in a climb or descent by how rapidly the needles move. A minor adjustment in pitch attitude may be made to control the rate at which altitude is gained or lost. Pitch attitude is used only to correct small altitude changes caused by external forces, such as turbulence or up and down drafts.

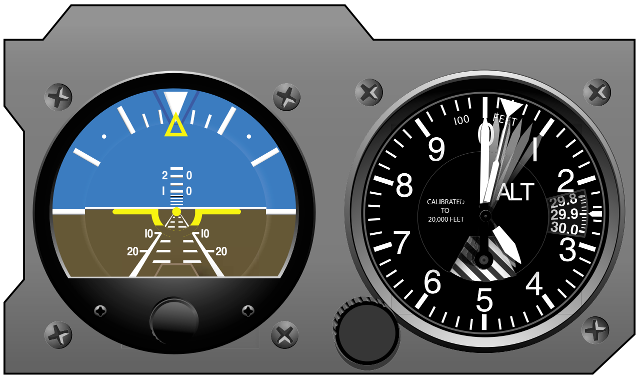

Figure 4-7. *Pitch Correction Using the Altimeter.*

Vertical Speed Indicator (VSI)

In flight at a constant altitude, the VSI (sometimes referred to as vertical velocity indicator or rate-of-climb indicator) remains at zero. If the needle moves above zero, the pitch attitude must be adjusted downward to stop the climb and return to level flight. Prompt adjustments to the changes in the indications of the VSI can prevent any significant change in altitude. *[Figure 4-8]* Turbulent air causes the needle to fluctuate near zero. In such conditions, the average of the fluctuations should be considered as the correct reading. Reference to the altimeter helps in turbulent air because it is not as sensitive as the VSI.

Vertical speed is represented in feet per minute (fpm). *[Figure 4-8]* The face of the instrument is graduated with numbers such as 1, 2, 3, etc. These represent thousands of feet up or down in a minute. For instance, if the pointer is aligned with .5 (1/2 of a thousand, or 500 fpm) the aircraft will climb 500 feet in one minute. The instrument is divided into two regions, one for climbing (up) and one for descending (down).

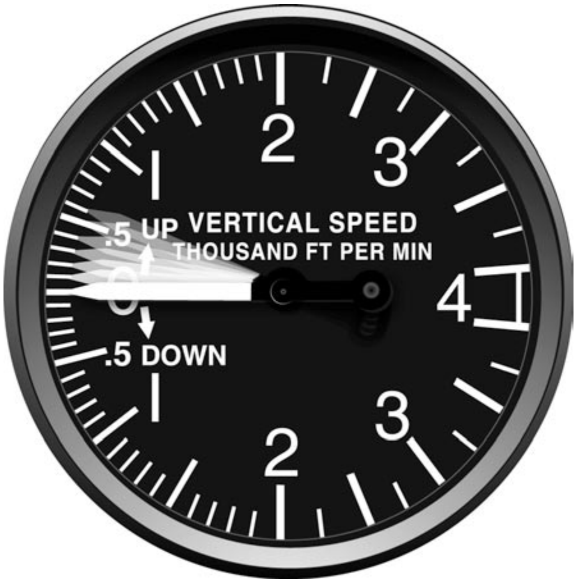

Figure 4-8. *Vertical Speed Indicator.*

During turbulence, it is not uncommon to see large fluctuations on the VSI. It is important to remember that small corrections should be employed to avoid further exacerbating a potentially divergent situation.

Overcorrecting causes the aircraft to overshoot the desired altitude; however, corrections should not be so small that the return to altitude is unnecessarily prolonged. As a guide, the pitch attitude should produce a rate of change on the VSI about twice the size of the altitude deviation. For example, if the aircraft is 100 feet off the desired altitude, a 200 fpm rate of correction would be used.

During climbs or descents, the VSI is used to change the altitude at a desired rate. Pitch attitude and power adjustments are made to maintain the desired rate of climb or descent on the VSI.

When pressure is applied to the controls and the VSI shows an excess of 200 fpm from that desired, overcontrolling is indicated. For example, if attempting to regain lost altitude at the rate of 500 fpm, a reading of more than 700 fpm would indicate overcontrolling. Initial movement of the needle indicates the trend of vertical movement. The time for the VSI to reach its maximum point of deflection after a correction is called lag. The lag is proportional to speed and magnitude of pitch change. In an airplane, overcontrolling may be reduced by relaxing pressure on the controls, allowing the pitch attitude to neutralize. In some helicopters with servo-assisted controls, no control pressures are apparent. In this case, overcontrolling can be reduced by reference to the attitude indicator.

Some aircraft are equipped with an instantaneous vertical speed indicator (IVSI). The letters "IVSI" appear on the face of the indicator. This instrument assists in interpretation by instantaneously indicating the rate of climb or descent at a given moment with little or no lag as displayed in a VSI.

Occasionally, the VSI is slightly out of calibration and indicates a gradual climb or descent when the aircraft is in level flight. If readjustments cannot be accomplished, the error in the indicator should be considered when the instrument is used for pitch control. For example, an improperly set VSI may indicate a descent of 100 fpm when the aircraft is in level flight. Any deviation from this reading would indicate a change in pitch attitude.

Airspeed Indicator

The airspeed indicator gives an indirect reading of the pitch attitude. With a constant power setting and a constant altitude, the aircraft is in level flight and airspeed remains constant. If the airspeed increases, the pitch attitude has lowered and should be raised. *[Figure 4-9]* If the airspeed decreases, the pitch attitude has moved higher and should be lowered. *[Figure 4-10]* A rapid change in airspeed indicates a large change in pitch; a slow change in airspeed indicates a small change in pitch. Although the airspeed indicator is used as a pitch instrument, it may be used in level flight for power control. Changes in pitch are reflected immediately by a change in airspeed. There is very little lag in the airspeed indicator.

Pitch Attitude Instrument Cross-Check

The altimeter is an important instrument for indicating pitch attitude in level flight except when used in conditions of exceptionally strong vertical currents, such as thunderstorms. With proper power settings, any of the pitch attitude instruments can be used to hold reasonably level flight attitude. However, only the altimeter gives the exact altitude information. Regardless of which pitch attitude control instrument indicates a need for a pitch attitude adjustment, the attitude indicator, if available, should be used to make the

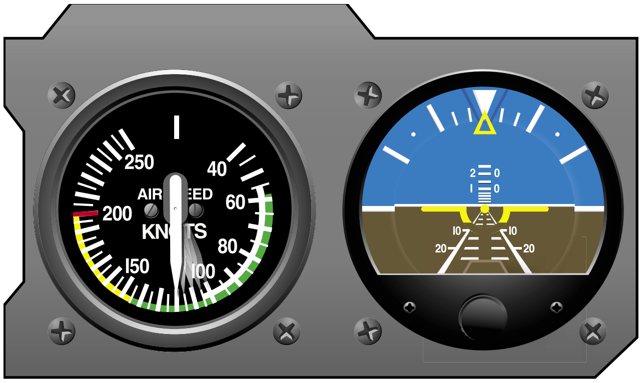
250
40
AIRSPEED
60
200
KNOTS
80
150
100
20
10
20

Figure 4-9. *Pitch attitude has lowered.*

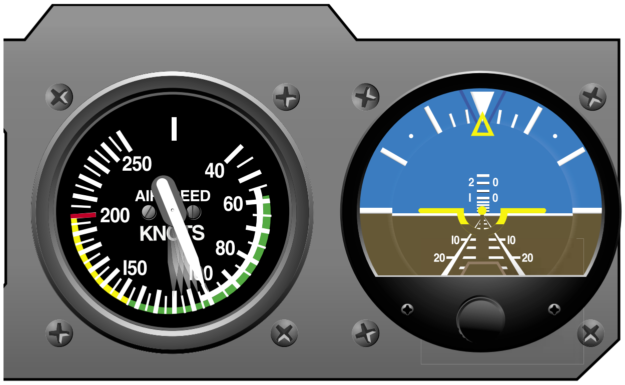
250
40
AIRSPEED
60
200
KNOTS
80
150
100
20
10
20

Figure 4-10. *Pitch attitude has moved higher.*

adjustment. Common errors in pitch attitude control are:

- Overcontrolling,
- Improperly using power, and
- Failing to adequately cross-check the pitch attitude instruments and take corrective action when pitch attitude change is needed

Bank Control

Bank control is controlling the angle made by the wing and the horizon. After interpreting the bank attitude from the appropriate instruments, exert the necessary pressures to move the ailerons and roll the aircraft about the longitudinal axis. As illustrated in *Figure 4-11*, these instruments include:

Attitude Indicator

As previously discussed, the attitude indicator is the only instrument that portrays both instantly and directly the actual flight attitude and is the basic attitude reference.

Heading Indicator

The heading indicator supplies the pertinent bank and heading information and is considered a primary instrument for bank.

Magnetic Compass

The magnetic compass provides heading information and is considered a bank instrument when used with the heading indicator. Care should be exercised when using the magnetic compass as it is affected by acceleration, deceleration in flight caused by turbulence, climbing, descending, power changes, and airspeed adjustments. Additionally, the magnetic compass indication will lead and lag in its reading depending upon the direction of turn. As a result, acceptance of its indication should be considered with other instruments that indicate turn information. These include the already mentioned attitude and heading indicators as well as the turn-and-slip indicator and turn coordinator.

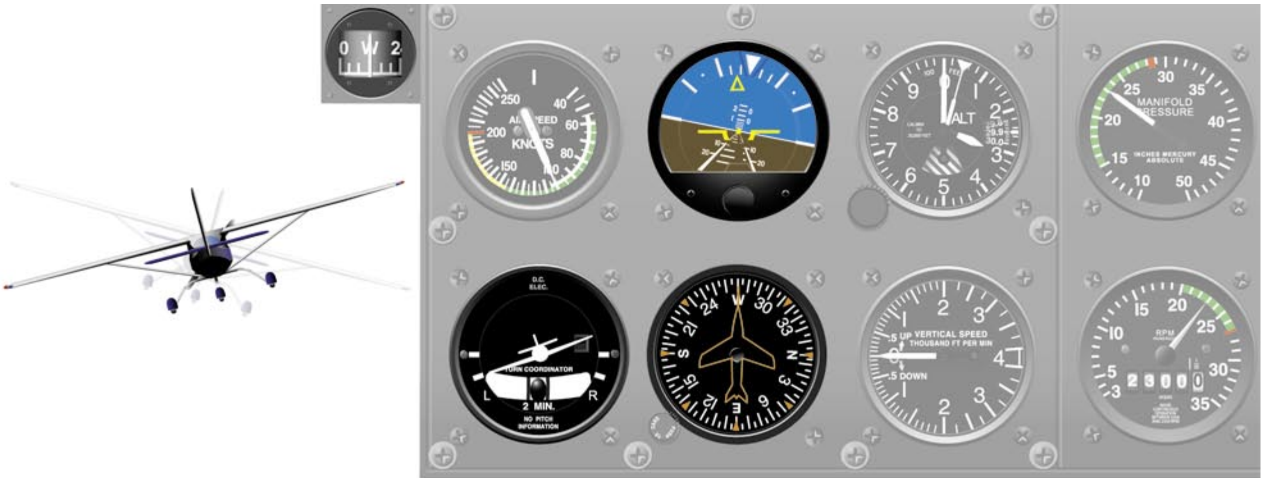
W
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
L
R
D.C. ELEC.
VERTICAL SPEED
THOUSAND FT PER MIN
UP
DOWN
RPM
2 3 0 0 0

Figure 4-11. *Bank Instruments.*

Turn Coordinator/Turn-and-Slip Indicator

Both of these instruments provide turn information. *[Figure 4-12]* The turn coordinator provides both bank rate and then turn rate once stabilized. The turn-and-slip indicator provides only turn rate.

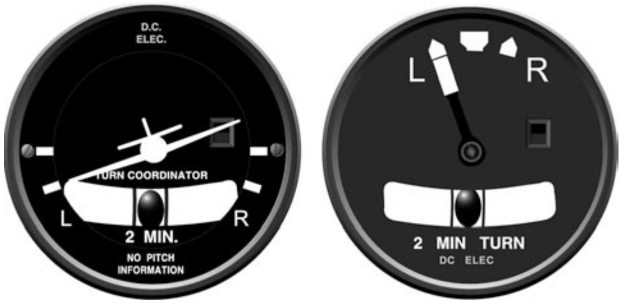

Figure 4-12. *Turn Coordinator and Turn-and-Slip Indicator.*

Power Control

A power change to adjust airspeed may cause movement around some or all of the aircraft axes. The amount and direction of movement depends on how much or how rapidly the power is changed, whether single-engine or multiengine airplane or helicopter. The effect on pitch attitude and airspeed caused by power changes during level flight is illustrated in *Figures 4-13* and *4-14*. During or immediately after adjusting the power control(s), the power instruments should be cross-checked to see if the power adjustment is as desired. Whether or not the need for a power adjustment is indicated by another instrument(s), adjustment is made by cross-checking the power instruments. Aircraft are powered by a variety of power plants, each power plant having certain instruments that indicate the amount of power being applied to operate the aircraft. During instrument flight, these instruments must be used to make the required power adjustments.

As illustrated in *Figure 4-15,* power indicator instruments include:

Airspeed Indicator

The airspeed indicator provides an indication of power best observed initially in level flight where the aircraft is in balance and trim. If in level flight the airspeed is increasing, it can generally be assumed that the power has increased, necessitating the need to adjust power or re-trim the aircraft.

Engine Instruments

Engine instruments, such as the manifold pressure (MP) indicator, provide an indication of aircraft performance for a given setting under stable conditions. If the power conditions are changed, as reflected in the respective engine instrument readings, there is an affect upon the aircraft performance, either an increase or decrease of airspeed. When the propeller rotational speed (revolutions per minute (RPM) as viewed on a tachometer) is increased or decreased on fixed-pitch propellers, the performance of the aircraft reflects a gain or loss of airspeed as well.

Trim Control

Proper trim technique is essential for smooth and accurate instrument flying and utilizes instrumentation illustrated in *Figure 4-16.* The aircraft should be properly trimmed while executing a maneuver. The degree of flying skill, which ultimately develops, depends largely upon how well the aviator learns to keep the aircraft trimmed.

Airplane Trim

An airplane is correctly trimmed when it is maintaining a desired attitude with all control pressures neutralized. By relieving all control pressures, it is much easier to maintain the aircraft at a certain attitude. This allows more time to devote to the navigation instruments and additional flight deck duties.

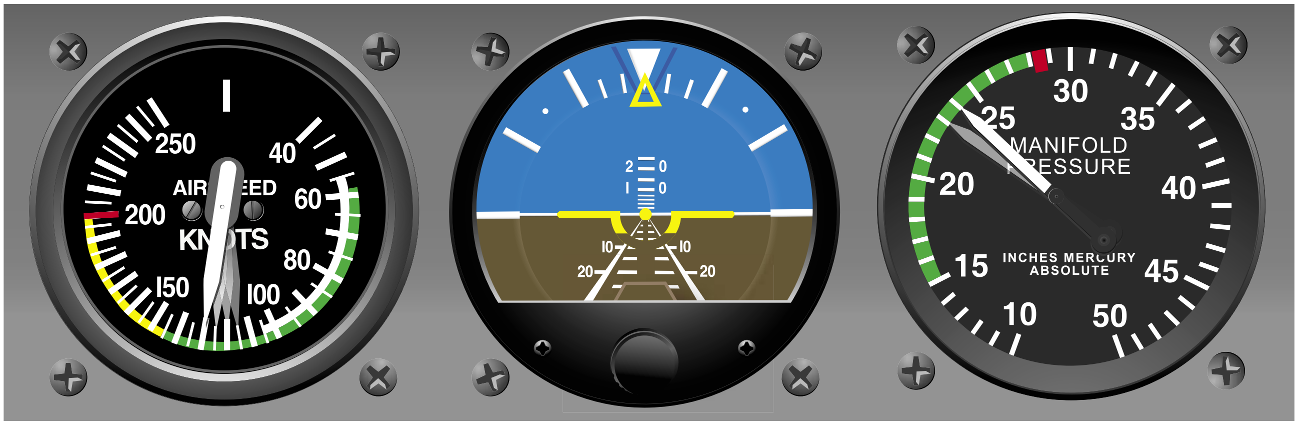

Figure 4-13. *An Increase in Power Inscreasing Airpseed Accordingly in Level Flight.*

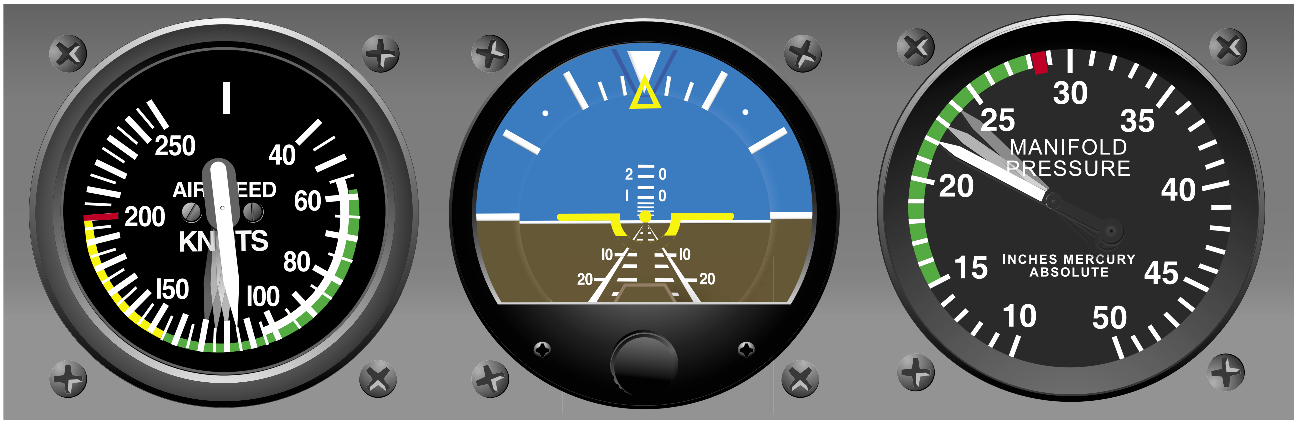
250
40
60
200
AIRSPEED
KNOTS
80
150
100
2
1
10
20
30
25
35
MANIFOLD PRESSURE
20
40
15
INCHES MERCURY ABSOLUTE
45
10
50

Figure 4-14. *Pitch Control and Power Adjustment Required To Bring Aircraft to Level Flight.*

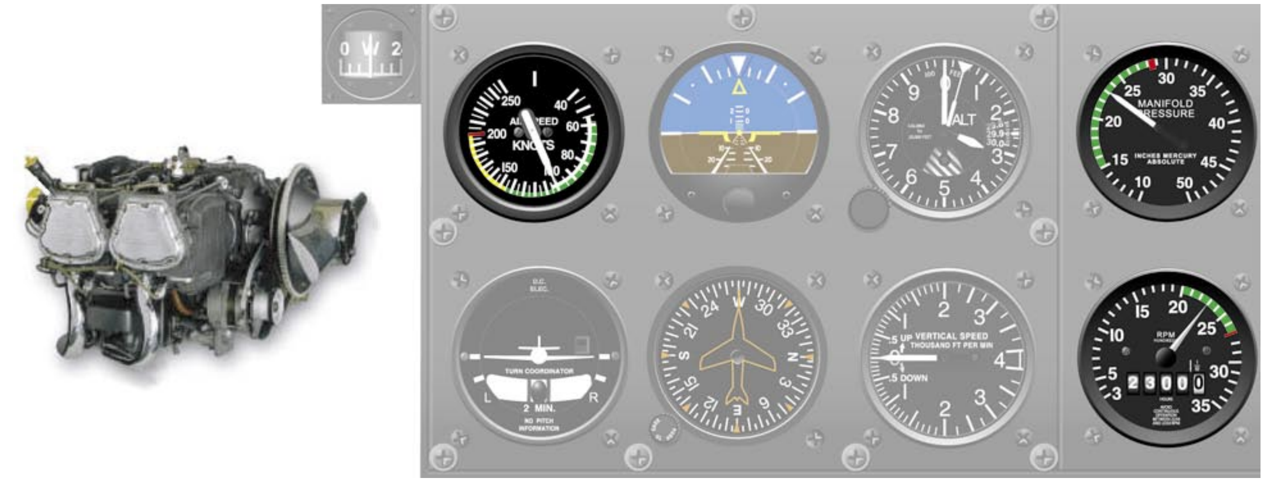
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
L
R
D.C. ELEC.
VERTICAL SPEED
THOUSAND FT PER MIN
UP
DOWN
RPM
2 3 0 0 0

Figure 4-15. *Power Instruments.*

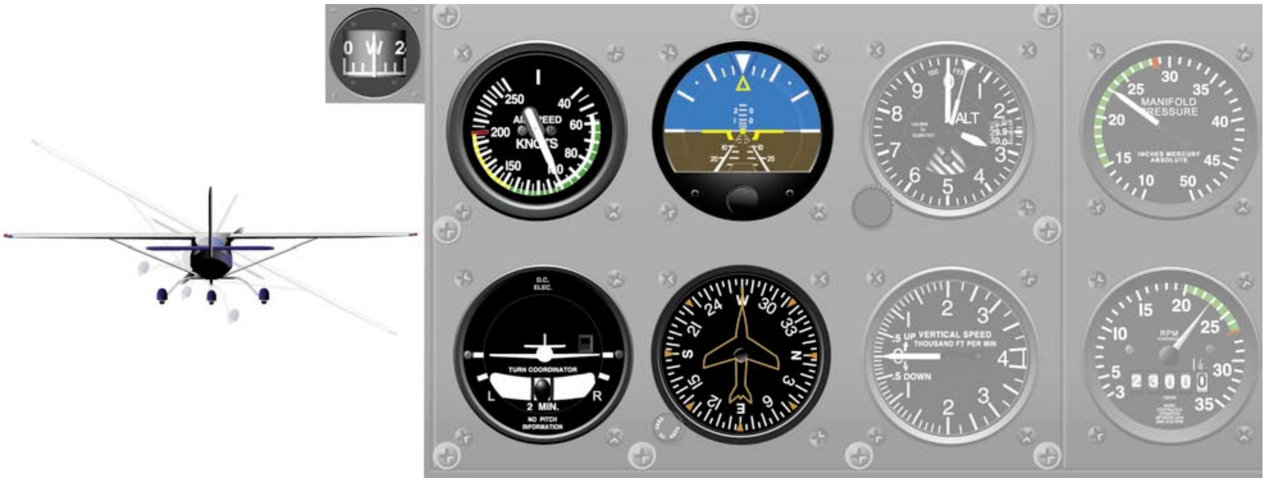
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
L
R
D.C. ELEC.
VERTICAL SPEED
THOUSAND FT PER MIN
UP
DOWN
RPM
2 3 0 0 0

Figure 4-16. *Trim Instruments.*

An aircraft is placed in trim by:

- Applying control pressure(s) to establish a desired attitude. Then, the trim is adjusted so that the aircraft maintains that attitude when flight controls are released. The aircraft is trimmed for coordinated flight by centering the ball of the turn-and-slip indicator.
- Moving the rudder trim in the direction where the ball is displaced from center. Aileron trim may then be adjusted to maintain a wings-level attitude.
- Using balanced power or thrust when possible to aid in maintaining coordinated flight. Changes in attitude, power, or configuration may require trim adjustments. Use of trim alone to establish a change in aircraft attitude usually results in erratic aircraft control. Smooth and precise attitude changes are best attained by a combination of control pressures and subsequent trim adjustments. The trim controls are aids to smooth aircraft control.

Helicopter Trim

A helicopter is placed in trim by continually cross-checking the instruments and performing the following:

- Using the cyclic centering button. If the helicopter is so equipped, this relieves all possible cyclic pressures.
- Using the pedal adjustment to center the ball of the turn indicator. Pedal trim is required during all power changes and is used to relieve all control pressures held after a desired attitude has been attained.

An improperly trimmed helicopter requires constant control pressures, produces tension, distracts attention from cross-checking, and contributes to abrupt and erratic attitude control. The pressures felt on the controls should be only those applied while controlling the helicopter.

Adjust the pitch attitude, as airspeed changes, to maintain desired attitude for the maneuver being executed. The bank must be adjusted to maintain a desired rate of turn, and the pedals must be used to maintain coordinated flight. Trim must be adjusted as control pressures indicate a change is needed.

Example of Primary and Support Instruments

Straight-and-level flight at a constant airspeed means that an exact altitude is to be maintained with zero bank (constant heading). The primary pitch, bank, and power instruments used to maintain this flight condition are:

- Altimeter—supplies the most pertinent altitude information and is primary for pitch.
- Heading Indicator—supplies the most pertinent bank or heading information and is primary for bank.
- Airspeed Indicator—supplies the most pertinent information concerning performance in level flight in terms of power output and is primary for power.

Although the attitude indicator is the basic attitude reference, the concept of primary and supporting instruments does not devalue any particular flight instrument, when available, in establishing and maintaining pitch-and-bank attitudes. It is the only instrument that instantly and directly portrays the actual flight attitude. It should always be used, when available, in establishing and maintaining pitch-and-bank attitudes. The specific use of primary and supporting instruments during basic instrument maneuvers is presented in more detail in Chapter 5, Airplane Basic Flight Maneuvers.

Fundamental Skills

During attitude instrument training, two fundamental flight skills must be developed. They are instrument cross-check and instrument interpretation, both resulting in positive aircraft control. Although these skills are learned separately and in deliberate sequence, a measure of proficiency in precision flying is the ability to integrate these skills into unified, smooth, positive control responses to maintain any prescribed flight path.

Instrument Cross-Check

The first fundamental skill is cross-checking (also called "scanning" or "instrument coverage"). Cross-checking is the continuous and logical observation of instruments for attitude and performance information. In attitude instrument flying, the pilot maintains an attitude by reference to instruments, producing the desired result in performance. Observing and interpreting two or more instruments to determine attitude and performance of an aircraft is called cross-checking. Although no specific method of cross-checking is recommended, those instruments that give the best information for controlling the aircraft in any given maneuver should be used. The important instruments are the ones that give the most pertinent information for any particular phase of the maneuver. These are usually the instruments that should be held at a constant indication. The remaining instruments should help maintain the important instruments at the desired indications, which is also true in using the emergency panel.

Cross-checking is mandatory in instrument flying. In visual flight, a level attitude can be maintained by outside references. However, even then the altimeter must be checked to determine if altitude is being maintained. Due to human error, instrument error, and airplane performance differences in various atmospheric and loading conditions, it is impossible to establish an attitude and have performance remain constant for a long period of time. These variables make it necessary for the pilot to constantly check the instruments and make appropriate changes in airplane attitude using cross-checking of instruments. Examples of cross-checking are explained in the following paragraphs.

Selected Radial Cross-Check

When the selected radial cross-check is used, a pilot spends 80 to 90 percent of flight time looking at the attitude indicator, taking only quick glances at the other flight instruments (for this discussion, the five instruments surrounding the attitude indicator are called the flight instruments). With this method, the pilot's eyes never travel directly between the flight instruments but move by way of the attitude indicator. The maneuver being performed determines which instruments to look at in the pattern. *[Figure 4-17]*

Inverted-V Cross-Check

In the inverted-V cross-check, the pilot scans from the attitude indicator down to the turn coordinator, up to the attitude indicator, down to the VSI, and back up to the attitude indicator. *[Figure 4-18]*

Rectangular Cross-Check

In the rectangular cross-check, the pilot scans across the top three instruments (airspeed indicator, attitude indicator, and altimeter) and then drops down to scan the bottom three instruments (VSI, heading indicator, and turn instrument). This scan follows a rectangular path (clockwise or counterclockwise rotation is a personal choice). *[Figure 4-19]*

This cross-checking method gives equal weight to the information from each instrument, regardless of its importance to the maneuver being performed. However, this method lengthens the time it takes to return to an instrument critical to the successful completion of the maneuver.

Common Cross-Check Errors

A beginner might cross-check rapidly, looking at the instruments without knowing exactly what to look for. With increasing experience in basic instrument maneuvers and familiarity with the instrument indications associated with them, a pilot learns what to look for, when to look for it, and what response to make. As proficiency increases, a pilot cross-checks primarily from habit, suiting scanning rate and sequence to the demands of the flight situation. Failure to maintain basic instrument proficiency through practice can result in many of the following common scanning errors, both during training and at any subsequent time.

Fixation, or staring at a single instrument, usually occurs for a reason, but has poor results. For example, a pilot may stare at the altimeter reading 200 feet below the assigned altitude, and wonder how the needle got there. While fixated on the

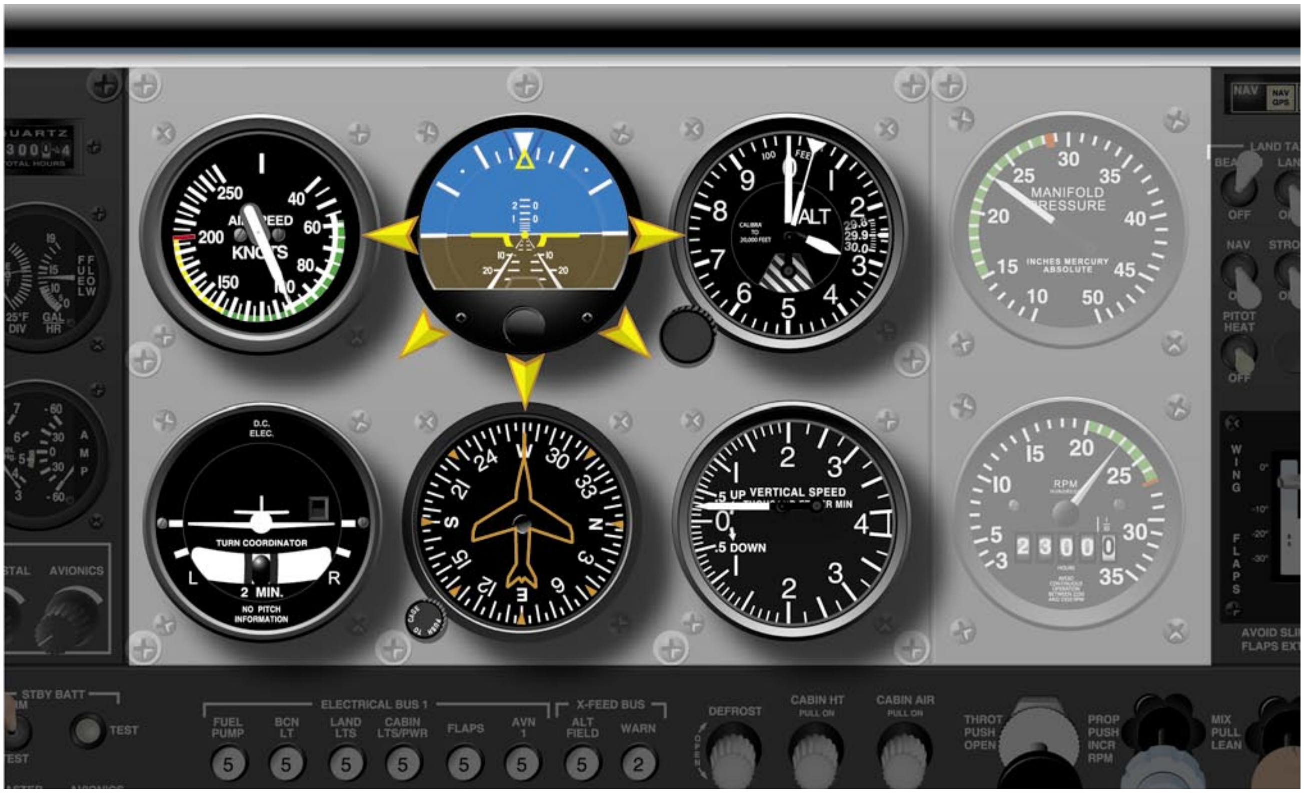
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
D.C. ELEC.
TURN COORDINATOR
L
R
2 MIN.
NO PITCH INFORMATION
VERTICAL SPEED
UP
DOWN
RPM
STBY BATT
TEST
ELECTRICAL BUS 1
X-FEED BUS
FUEL PUMP
BCN LT
LAND LTS
CABIN LTS/PWR
FLAPS
AVN 1
ALT FIELD
WARN
DEFROST
CABIN HT
CABIN AIR
THROT PUSH OPEN
PROP PUSH INCR RPM
MIX PULL LEAN
NAV
PITOT HEAT
AVIONICS
FLAPS

Figure 4-17. *Radial Cross-Check.*

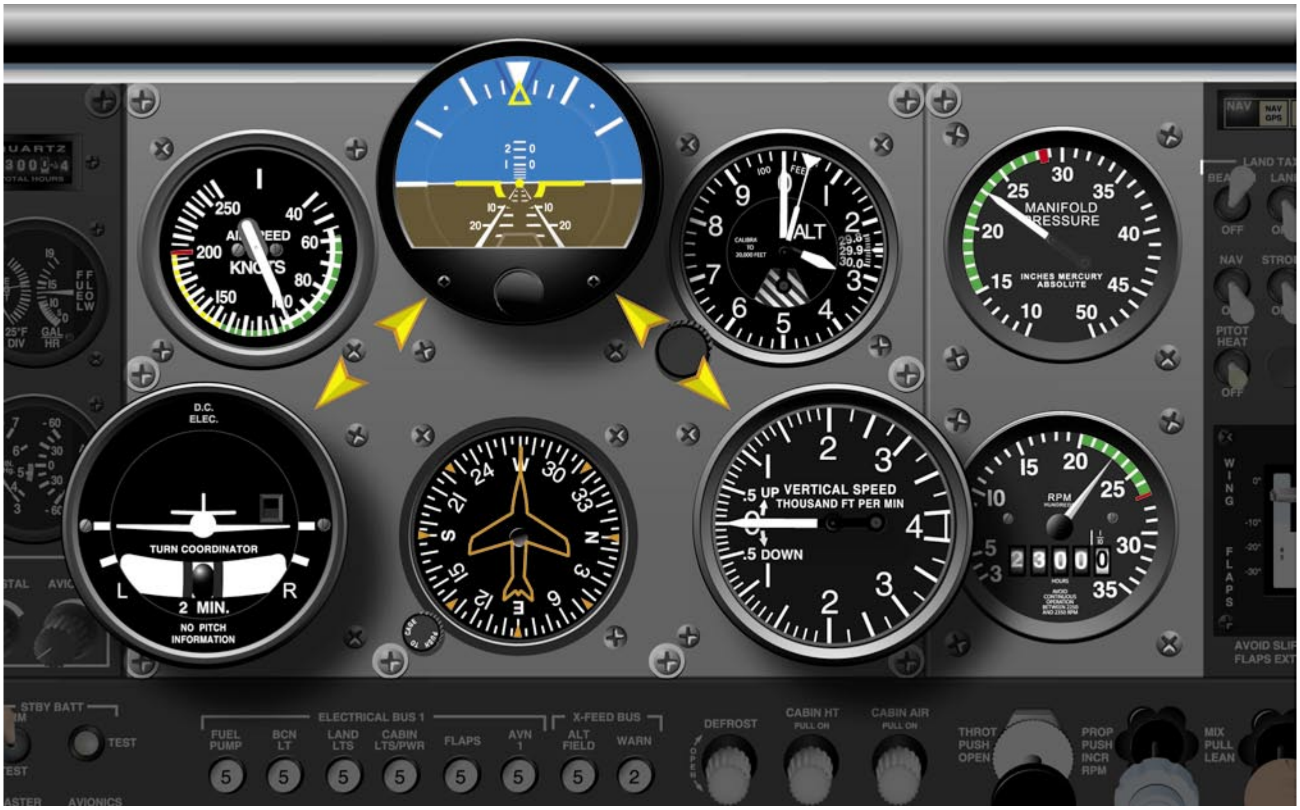

Figure 4-18. *Inverted-V Cross-Check.*

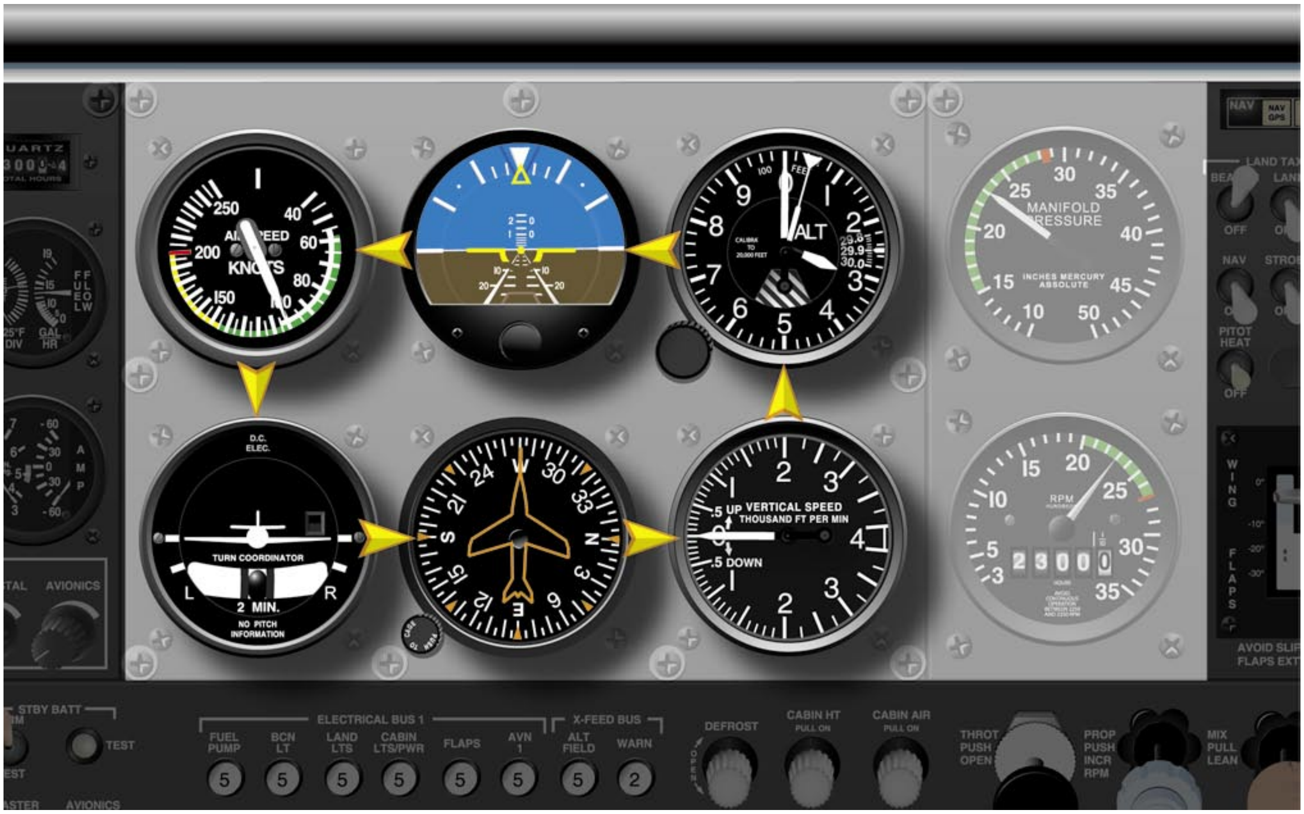

Figure 4-19. *Rectangular Cross-Check.*

instrument, increasing tension may be unconsciously exerted on the controls, which leads to an unnoticed heading change that leads to more errors. Another common fixation is likely when initiating an attitude change. For example, a shallow bank is established for a 90° turn and, instead of maintaining a cross-check of other pertinent instruments, the pilot stares at the heading indicator throughout the turn. Since the aircraft is turning, there is no need to recheck the heading indicator for approximately 25 seconds after turn entry. The problem here may not be entirely due to cross-check error. It may be related to difficulties with instrument interpretation. Uncertainty about reading the heading indicator (interpretation) or uncertainty because of inconsistency in rolling out of turns (control) may cause the fixation.

Omission of an instrument from a cross-check is another likely fault. It may be caused by failure to anticipate significant instrument indications following attitude changes. For example, in a roll-out from a 180° steep turn, straight-and-level flight is established with reference only to the attitude indicator, and the pilot neglects to check the heading indicator for constant heading information. Because of precession error, the attitude indicator temporarily shows a slight error, correctable by quick reference to the other flight instruments.

Emphasis on a single instrument, instead of on the combination of instruments necessary for attitude information, is an understandable fault during the initial stages of training. It is a natural tendency to rely on the instrument that is most readily understood, even when it provides erroneous or inadequate information. Reliance on a single instrument is poor technique. For example, a pilot can maintain reasonably close altitude control with the attitude indicator, but cannot hold altitude with precision without including the altimeter in the cross-check.

Instrument Interpretation

The second fundamental skill, instrument interpretation, requires more thorough study and analysis. It begins by understanding each instrument's construction and operating principles. Then, this knowledge must be applied to the performance of the aircraft being flown, the particular maneuvers to be executed, the cross-check and control techniques applicable to that aircraft, and the flight conditions.

For example, a pilot uses full power in a small airplane for a 5-minute climb from near sea level, and the attitude indicator shows the miniature aircraft two bar widths (twice the thickness of the miniature aircraft wings) above the artificial horizon. *[Figure 4-20]* The airplane is climbing at 500 fpm as shown on the VSI, and at airspeed of 90 knots, as shown on the airspeed indicator. With the power available in this particular airplane and the attitude selected by the pilot, the performance is shown on the instruments. Now, set up the identical picture on the attitude indicator in a jet airplane. With the same airplane attitude as shown in the first example, the VSI in the jet reads 2,000 fpm and the airspeed indicator reads 250 knots.

As the performance capabilities of the aircraft are learned, a pilot interprets the instrument indications appropriately

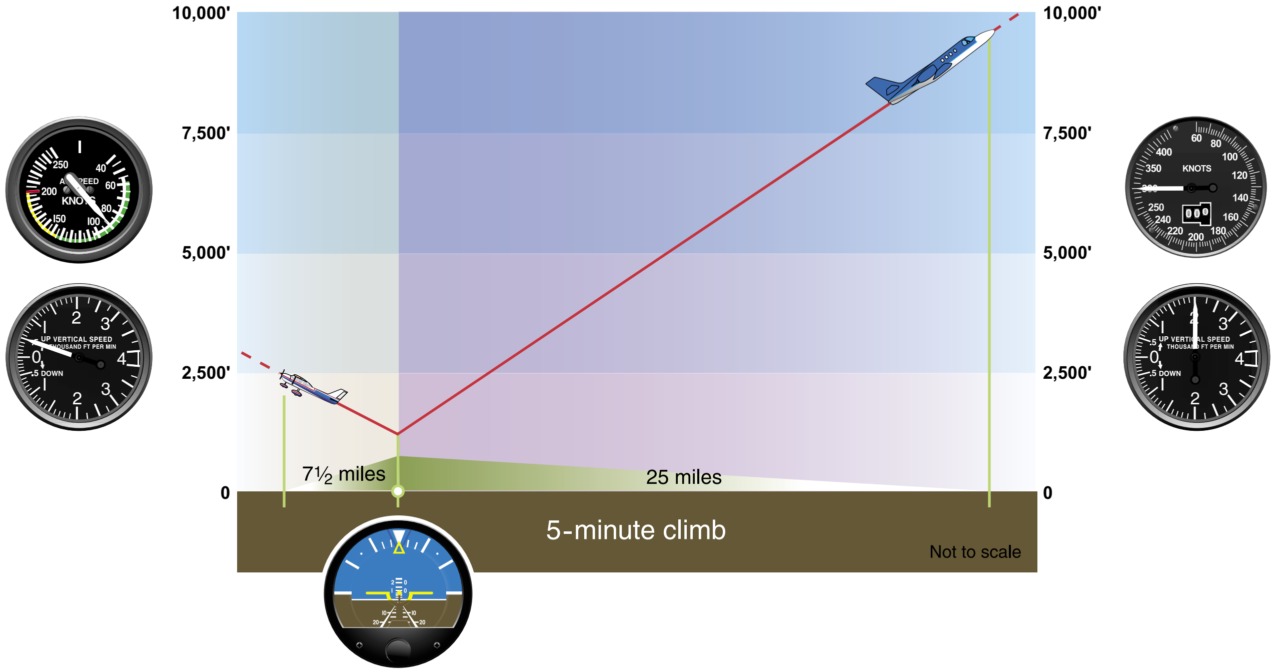

Figure 4-20. *Power and Attitude Equal Performance.*

in terms of the attitude of the aircraft. If the pitch attitude is to be determined, the airspeed indicator, altimeter, VSI, and attitude indicator provide the necessary information. If the bank attitude is to be determined, the heading indicator, turn coordinator, and attitude indicator must be interpreted. For each maneuver, learn what performance to expect and the combination of instruments to be interpreted in order to control aircraft attitude during the maneuver. It is the two fundamental flight skills, instrument cross-check and instrument interpretation, that provide the smooth and seamless control necessary for basic instrument flight as discussed at the beginning of the chapter.

NAV1 108.00
NAV2 108.00
113.00
110.60
WPT
DIS
NM
DTK
TRK 360°
GARMIN
134.000
123.800
118.000 COM1
118.000 COM2
MAN IN HG
15.0
RPM
x100
2200
FUEL QTY GAL
GPH
13.7
46
200
1652
338
RON TRIM
120
110
100
90
80
70
TAS 106KT
OAT 7°C
4000
4300
4200
3900
3800
29.92IN
270°
VOR 1
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
XPDR 5537 IDNT LCL10:12:34
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS

Chapter 4, Section II

Airplane Attitude Instrument Flying

Using an Electronic Flight Display

Introduction

Attitude instrument flying is defined as the control of an aircraft's spatial position by using instruments rather than outside visual references. As noted in Section I, today's aircraft come equipped with analog and/or digital instruments. Section II acquaints the pilot with the use of digital instruments known as an electronic flight display (EFD).

The improvements in avionics coupled with the introduction of EFDs to general aviation aircraft offer today's pilot an unprecedented array of accurate instrumentation to use in the support of instrument flying.

Until recently, most general aviation aircraft were equipped with individual instruments utilized collectively to safely maneuver the aircraft by instrument reference alone. With the release of the electronic flight display system, the conventional instruments have been replaced by multiple liquid crystal display (LCD) screens. The first screen is installed in front of the left seat pilot position and is referred to as the primary flight display (PFD). *[Figure 4-21]* The second screen is positioned in approximately the center of the instrument panel and is referred to as the multi-function display (MFD). *[Figure 4-22]* The pilot can use the MFD to display navigation information (moving maps), aircraft systems information (engine monitoring), or should the need arise, a PFD. *[Figure 4-23]* With just these two screens, aircraft designers have been able to de-clutter instrument panels while increasing safety. This has been accomplished through the utilization of solid-state instruments which have a failure rate far lower than those of conventional analog instrumentation.

However, in the event of electrical failure, the pilot still has emergency instruments as a backup. These instruments either do not require electrical power, or as in the case of many attitude indicators, they are battery equipped. *[Figure 4-24]*

Pilots flying under visual flight rules (VFR) maneuver their aircraft by reference to the natural horizon, utilizing specific reference points on the aircraft. In order to operate the aircraft in other than VFR weather, with no visual reference to the natural horizon, pilots need to develop additional skills. These skills come from the ability to maneuver the aircraft by reference to flight instruments alone. These flight instruments replicate all the same key elements that a VFR pilot utilizes during a normal flight. The natural horizon is replicated on the attitude indicator by the artificial horizon.

Understanding how each flight instrument operates and what role it plays in controlling the attitude of the aircraft is fundamental in learning attitude instrument flying. When the pilot understands how all the instruments are used in establishing and maintaining a desired aircraft attitude, the pilot is better prepared to control the aircraft should one or more key instruments fail or if the pilot should enter instrument flight conditions.

Learning Methods

There are two basic methods utilized for learning attitude instrument flying. They are “control and performance” and “primary and supporting.” These methods rely on the same flight instruments and require the pilot to make the same adjustments to the flight and power controls to control aircraft attitude. The main difference between the two methods is the importance that is placed on the attitude indicator and the interpretation of the other flight instruments.

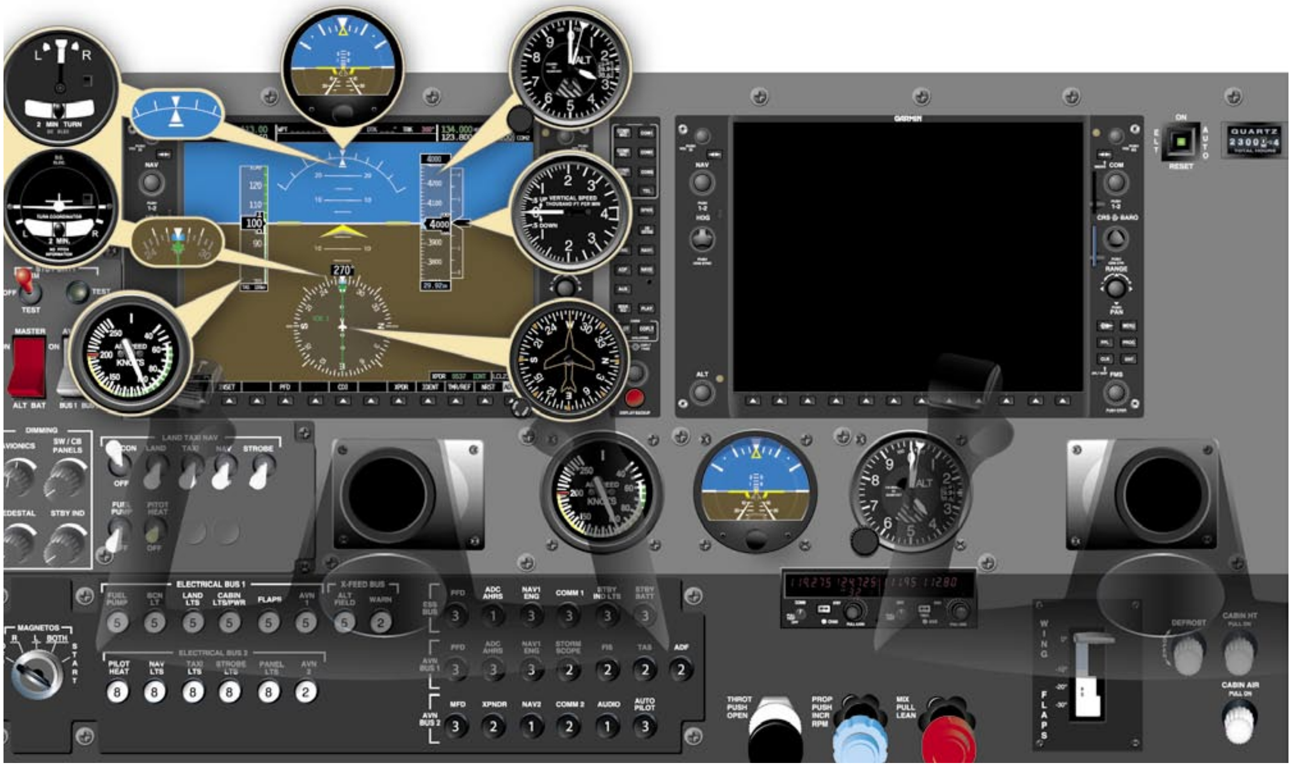

Figure 4-21. *Primary Flight Display (PFD) and Analog Counterparts.*

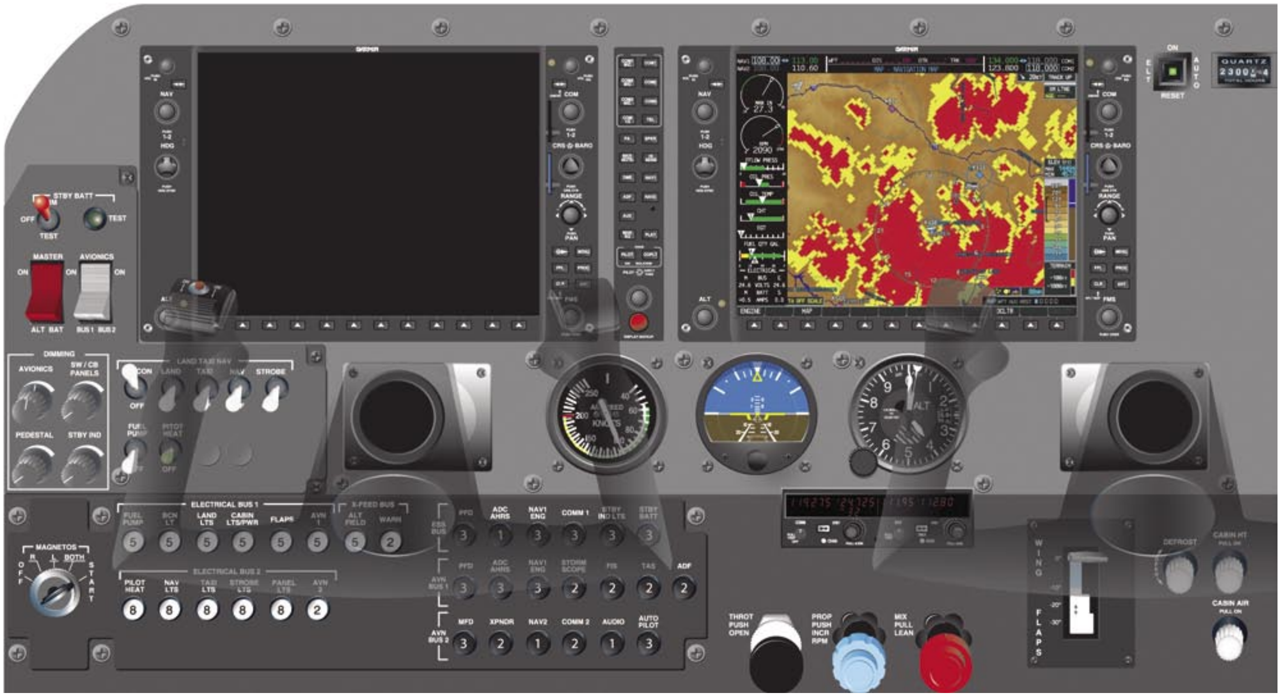
STBY BATT
OFF
TEST
MASTER
AVIONICS
ON
ALT BAT
BUS 1 BUS 2
DIMMING
AVIONICS
SW / CB PANELS
PEDESTAL
STBY IND
BCN
LAND
TAXI
NAV
STROBE
FUEL PUMP
PITOT HEAT
ELECTRICAL BUS 1
ELECTRICAL BUS 2
X-FEED BUS
MAGNETOS
R
L
BOTH
START
THROT PUSH OPEN
PROP PUSH INCR RPM
MIX PULL LEAN
WING FLAPS
DEFROST
CABIN HT
CABIN AIR
ELT
RESET
QUARTZ

Figure 4-22. *Multifunction Display (MFD).*

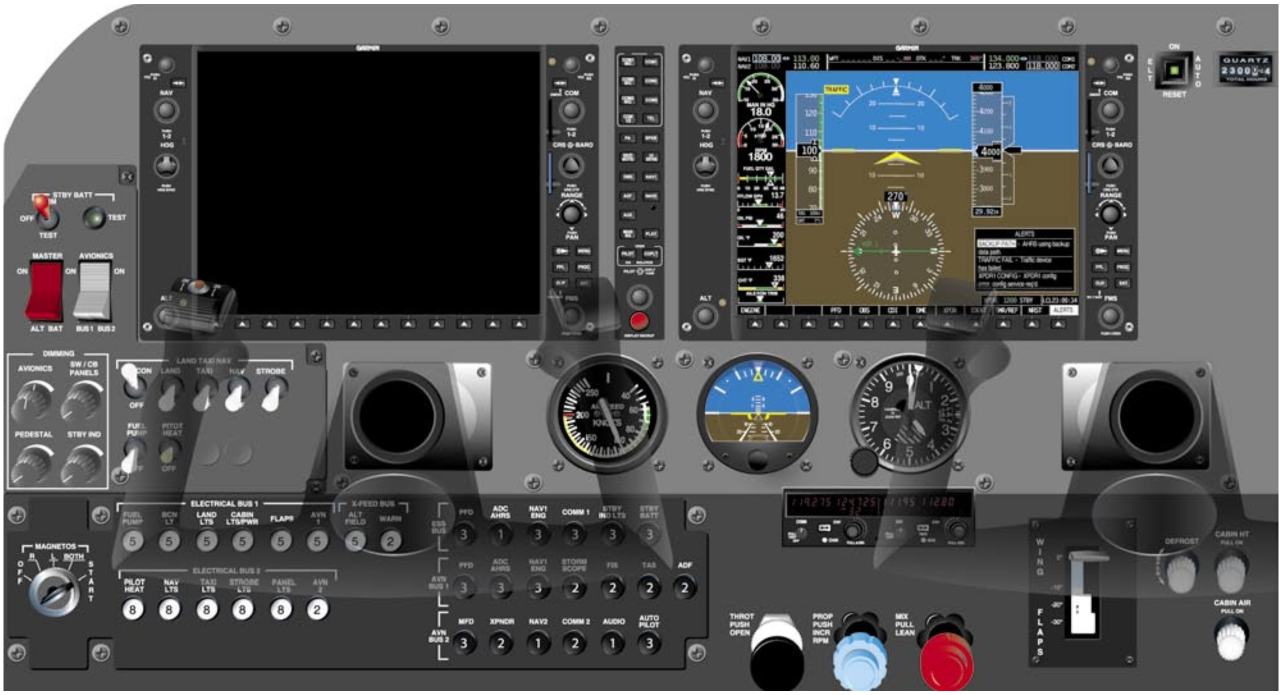
STBY BATT
OFF
TEST
MASTER
AVIONICS
ON
ALT BAT
BUS 1 BUS 2
DIMMING
AVIONICS
SW / CB PANELS
PEDESTAL
STBY IND
BCN
LAND
TAXI
NAV
STROBE
FUEL PUMP
PITOT HEAT
ELECTRICAL BUS 1
ELECTRICAL BUS 2
X-FEED BUS
MAGNETOS
R
L
BOTH
START
THROT PUSH OPEN
PROP PUSH INCR RPM
MIX PULL LEAN
WING FLAPS
DEFROST
CABIN HT
CABIN AIR
ELT
RESET
QUARTZ
ALERTS

Figure 4-23. *Reversionary Displays.*

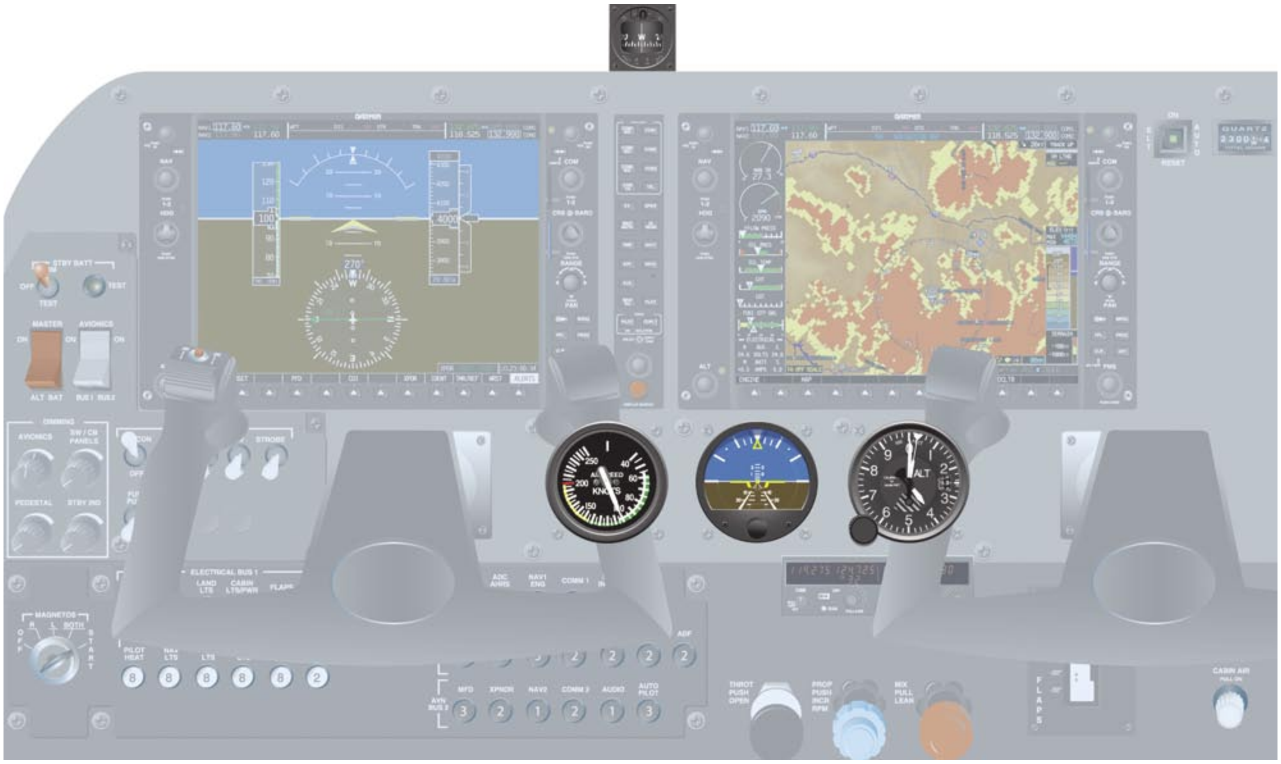

Figure 4-24. *Emergency Back-up of the Airspeed Indicator, Attitude Indicator, and Altitude Indicator.*

Control and Performance Method

Aircraft performance is accomplished by controlling the aircraft attitude and power output. Aircraft attitude is the relationship of its longitudinal and lateral axes to the Earth's horizon. When flying in instrument flight conditions, the pilot controls the attitude of the aircraft by referencing the flight instruments and manipulating the power output of the engine to achieve the performance desired. This method can be used to achieve a specific performance level enabling a pilot to perform any basic instrument maneuver.

The instrumentation can be broken up into three different categories: control, performance, and navigation.

Control Instruments

The control instruments depict immediate attitude and power changes. The instrument for attitude display is the attitude indicator. Power changes are directly reflected on the manifold pressure gauge and the tachometer. *[Figure 4-25]* All three of these instruments can reflect small adjustments, allowing for precise control of aircraft attitude.

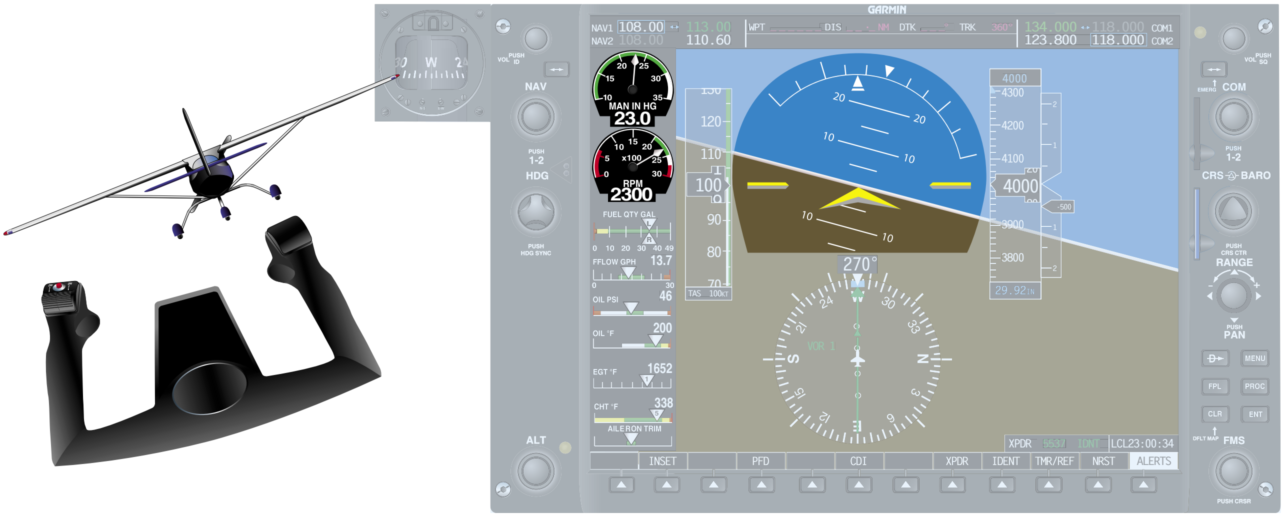
GARMIN
MAN IN HG
23.0
RPM
2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
270°
4000
29.92IN
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 4-25. *Control Instruments.*

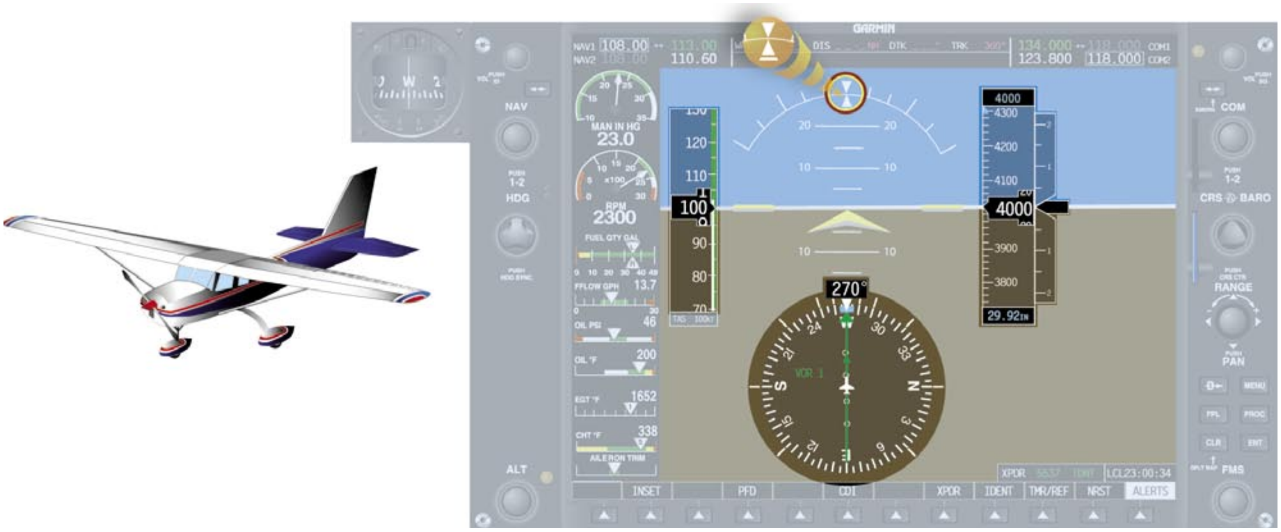
GARMIN
NAV1 108.00 113.00
NAV2 108.00 110.60
DIS
DTK
TRK
134.000 118.000 COM1
123.800 118.000 COM2
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS
MAN IN HG
23.0
RPM
2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
100
4000
270°
29.92IN
VOR 1
TAS 100KT
XPDR
LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 4-26. *Performance Instruments.*

In addition, the configuration of the power indicators installed in each aircraft may vary to include the following types of power indicators: tachometers, manifold pressure indicator, engine pressure ratio indicator, fuel flow gauges, etc.

The control instruments do not indicate how fast the aircraft is flying or at what altitude it is flying. In order to determine these variables and others, a pilot needs to refer to the performance instruments.

Performance Instruments

The performance instruments directly reflect the performance the aircraft is achieving. The speed of the aircraft can be referenced on the airspeed indicator. The altitude can be referenced on the altimeter. The aircraft's climb performance can be determined by referencing the vertical speed indicator (VSI). *[Figure 4-26]* Other performance instruments available are the heading indicator, angle of attack indicator, and the slip/skid indicator.

The performance instruments will most directly reflect a change in acceleration, which is defined as change in velocity or direction. Therefore, these instruments indicate if the aircraft is changing airspeed, altitude, or heading, which are horizontal, vertical, or lateral vectors.

Navigation Instruments

The navigation instruments are comprised of global positioning system (GPS) displays and indicators, very high frequency omnidirectional range/nondirectional radio beacon (VOR/NDB) indicators, moving map displays, localizer, and glide slope (GS) indicators. *[Figure 4-27]* The instruments indicate the position of the aircraft relative to a selected navigation facility or fix. Navigation instruments allow the pilot to maneuver the aircraft along a predetermined path of ground-based or spaced-based navigation signals without reference to any external visual cues. The navigation instruments can support both lateral and visual inputs.

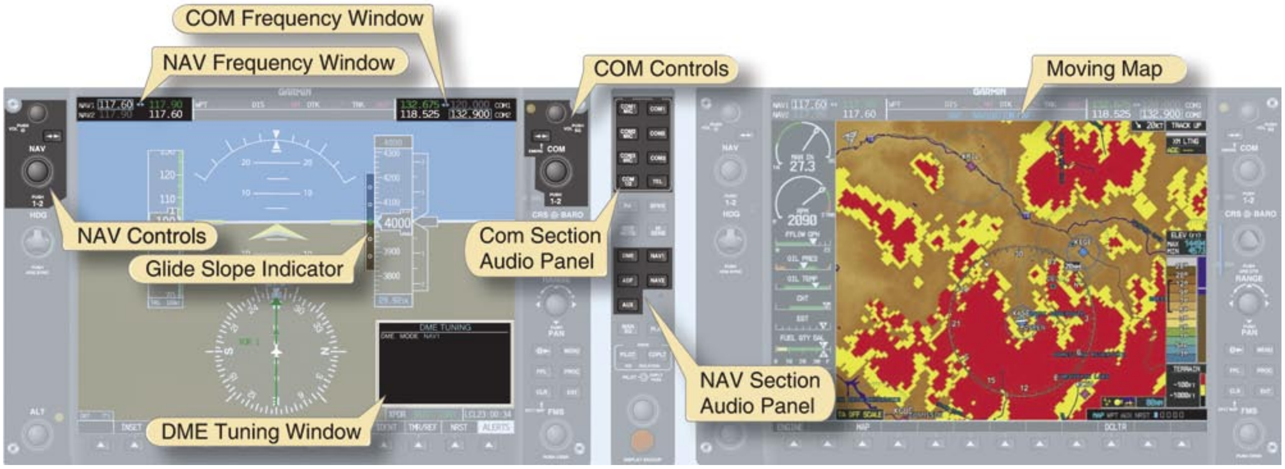
COM Frequency Window
NAV Frequency Window
COM Controls
Moving Map
NAV Controls
Glide Slope Indicator
Com Section
Audio Panel
NAV Section
Audio Panel
DME Tuning Window

Figure 4-27. *Navigation Instruments.*

The Four-Step Process Used to Change Attitude

In order to change the attitude of the aircraft, the pilot must make the proper changes to the pitch, bank, or power settings of the aircraft. Four steps (establish, trim, cross-check, and adjust) have been developed in order to aid in the process.

Establish

Any time the attitude of the aircraft requires changing, the pilot must adjust the pitch and/or bank in conjunction with power to establish the desired performance. The changes in pitch and bank require the pilot to reference the attitude indicator in order to make precise changes. Power changes should be verified on the tachometer, manifold pressure gauge, etc. To ease the workload, the pilot should become familiar with the approximate pitch and power changes necessary to establish a specified attitude.

Trim

Another important step in attitude instrument flying is trimming the aircraft. Trim is utilized to eliminate the need to apply force to the control yoke in order to maintain the desired attitude. When the aircraft is trimmed appropriately, the pilot is able to relax pressure on the control yoke and momentarily divert attention to another task at hand without deviating from the desired attitude. Trimming the aircraft is very important, and poor trim is one of the most common errors instructors note in instrument students.

Cross-Check

Once the initial attitude changes have been made, the pilot should verify the performance of the aircraft. Cross-checking the control and performance instruments requires the pilot to visually scan the instruments as well as interpret the indications. All the instruments must be utilized collectively in order to develop a full understanding of the aircraft attitude. During the cross-check, the pilot needs to determine the magnitude of any deviations and determine how much of a change is required. All changes are then made based on the control instrument indications.

Adjust

The final step in the process is adjusting for any deviations that have been noted during the cross-check. Adjustments should be made in small increments. The attitude indicator and the power instruments are graduated in small increments to allow for precise changes to be made. The pitch should be made in reference to bar widths on the miniature airplane. The bank angle can be changed in reference to the roll scale and the power can be adjusted in reference to the tachometer, manifold pressure gauge, etc.

By utilizing these four steps, pilots can better manage the attitude of their aircraft. One common error associated with this process is making a larger than necessary change when a deviation is noted. Pilots need to become familiar with the aircraft and learn how great a change in attitude is needed to produce the desired performance.

Applying the Four-Step Process

In attitude instrument flight, the four-step process is used to control pitch attitude, bank attitude, and power application of the aircraft. The EFD displays indications precisely enough that a pilot can apply control more accurately.

Pitch Control

The pitch control is indicated on the attitude indicator which spans the full width of the PFD. Due to the increased size of the display, minute changes in pitch can be made and corrected for. The pitch scale on the attitude indicator is graduated in 5-degree increments which allow the pilot to make correction with precision to approximately 1/2 degree. The miniature airplane utilized to represent the aircraft in conventional attitude indicators is replaced in glass panel displays by a yellow chevron. *[Figure 4-28]* Representing the nose of the aircraft, the point of the chevron affords the pilot a much more precise indication of the degree of pitch and allows the pilot to make small, precise changes should the desired aircraft performance change. When the desired performance is not being achieved, precise pitch changes should be made by referencing the point of the yellow chevron.

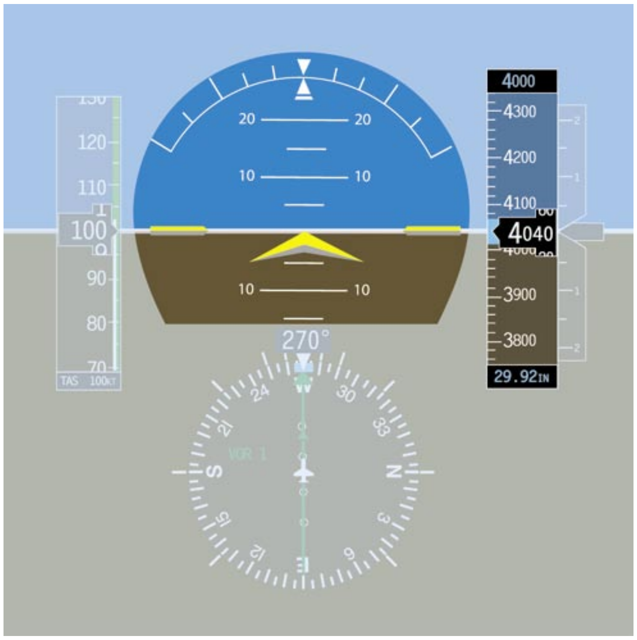

Figure 4-28. *The chevron's relationship to the horizon line indicates the pitch of the aircraft.*

Bank Control

Precise bank control can be developed utilizing the roll pointer in conjunction with the roll index displayed on the attitude indicator. The roll index is sectioned by hash marks at 0°, 10°, 20°, 30°, 45°, 60° and the horizon line which depicts 90° of bank. *[Figure 4-29]* The addition of the 45° hash mark is an improvement over conventional attitude indicators. In addition to the roll index, the instrument pilot utilizes the turn rate indicator to maintain the aircraft in a standard rate turn (3° per second). Most instrument maneuvers can be done comfortably, safely, and efficiently by utilizing a standard rate turn.

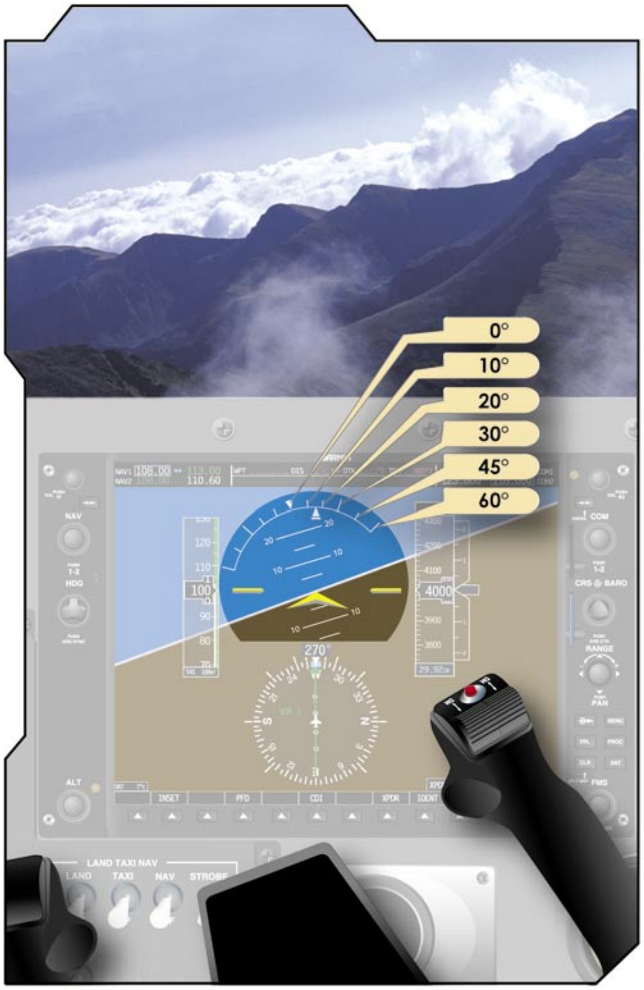

Figure 4-29. *Bank Control Index Lines.*

Power Control

The power instruments indicate how much power is being generated by the engine. They are not affected by turbulence, improper trim, or control pressures. All changes in power should be made with reference to power instruments and cross-checked on performance instruments.

Power control needs to be learned from the beginning of flight training. Attitude instrument flying demands increased precision when it comes to power control. As experience increases, pilots begin to know approximately how much change in throttle position is required to produce the desired change in airspeed. Different aircraft demand differing amounts of throttle change to produce specific performance. It is imperative that the pilot make the specific changes on the power instruments and allow the performance to stabilize. Avoid the tendency to overcontrol.

One common error encountered with glass panel displays is associated with the precision of the digital readouts. This precision causes pilots to focus too much attention on establishing the exact power setting.

Control and power instruments are the foundation for precise attitude instrument flying. The keys to attitude instrument flying are establishing the desired aircraft attitude on the attitude indicator and selecting the desired engine output on the power instruments. Cross-checking is the vital ingredient in maintaining precise attitude instrument flight.

Attitude Instrument Flying—Primary and Supporting Method

The second method for performing attitude instrument flight is a direct extension of the control/power method. By utilizing the primary and supporting flight instruments in conjunction with the control and power instruments, the pilot can precisely maintain aircraft attitude. This method utilizes the same instruments as the control/power method; however, it focuses more on the instruments that depict the most accurate indication for the aspect of the aircraft attitude being controlled. The four key elements (pitch, bank, roll, and trim) are discussed in detail.

Similar to the control/power method, all changes to aircraft attitude need to be made using the attitude indicator and the power instruments (tachometer, manifold pressure gauge, etc.). The following explains how each component of the aircraft attitude is monitored for performance.

Pitch Control

The pitch of the aircraft refers to the angle between the longitudinal axis of the aircraft and the natural horizon. When flying in instrument meteorological conditions, the natural horizon is unavailable for reference, and an artificial horizon is utilized in its place. *[Figure 4-30]* The only instrument capable of depicting the aircraft attitude is the attitude indicator displayed on the PFD. The attitude and heading reference system (AHRS) is the engine that drives the attitude display. The AHRS unit is capable of precisely tracking minute changes in the pitch, bank, and yaw axes, thereby making the PFD very accurate and reliable. The AHRS unit determines the angle between the aircraft's longitudinal axis and the horizon line on initialization. There is no need

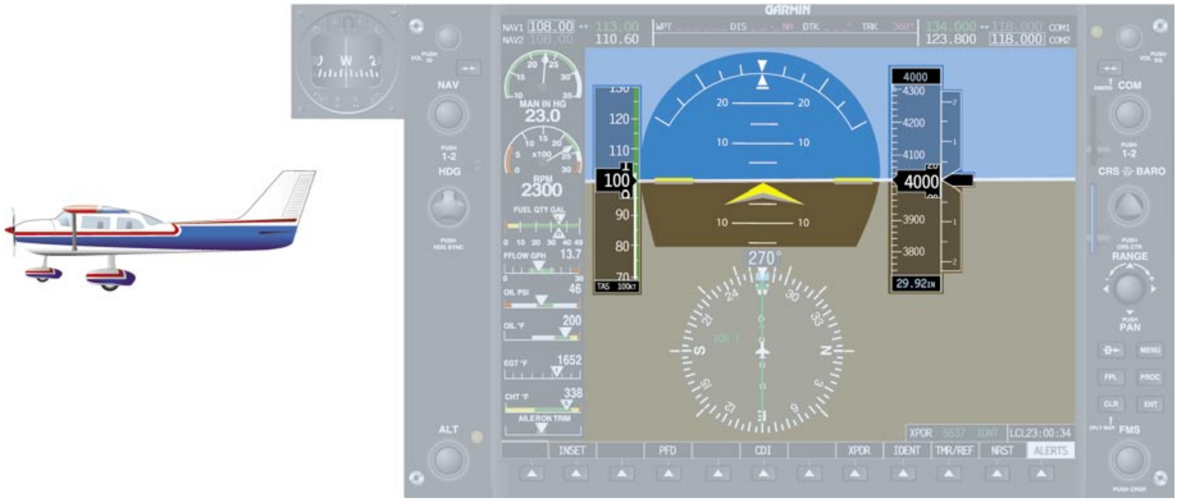

Figure 4-30. *Pitch of the Aircraft.*

or means for the pilot to adjust the position of the yellow chevron which represents the nose of the aircraft.

Straight-and-Level Flight

In straight-and-level flight, the pilot maintains a constant altitude, airspeed and, for the most part, heading for extended periods of time. To achieve this, three primary instruments need to be referenced in order to maintain these three variables.

Primary Pitch

When the pilot is maintaining a constant altitude, the primary instrument for pitch is the altimeter. As long as the aircraft maintains a constant airspeed and pitch attitude, the altitude should remain constant.

Two factors that cause the altitude to deviate are turbulence and momentary distractions. When a deviation occurs, a change in the pitch needs to be made on the attitude indicator. Small deviations require small corrections while large deviations require larger corrections. Pilots should avoid making large corrections that result in rapid attitude changes, for this may lead to spatial disorientation. Smooth, timely corrections should be made to bring the aircraft back to the desired attitude.

Pay close attention to indications on the PFD. An increase in pitch of 2.5° produces a climb rate of 450 feet per minute (fpm). Small deviations do not require large attitude changes.

A rule of thumb for correcting altitude deviations is to establish a change rate of twice the altitude deviation, not to exceed 500 fpm. For example, if the aircraft is off altitude by 40 feet, 2 x 40 = 80 feet, so a descent of approximately 100 fpm allows the aircraft to return to the desired altitude in a controlled, timely fashion.

In addition to the primary instrument, there are also supporting instruments that assist the pilot in cross-checking the pitch attitude. The supporting instruments indicate trend, but they do not indicate precise attitude indications. Three instruments (vertical speed, airspeed, and altitude trend tape) indicate when the pitch attitude has changed and that the altitude is changing. *[Figure 4-31]* When the altitude is constant, the VSI and altitude trend tape are not shown on the PFD. When these two trend indicators are displayed, the pilot is made aware that the pitch attitude of the aircraft has changed and may need adjustment.

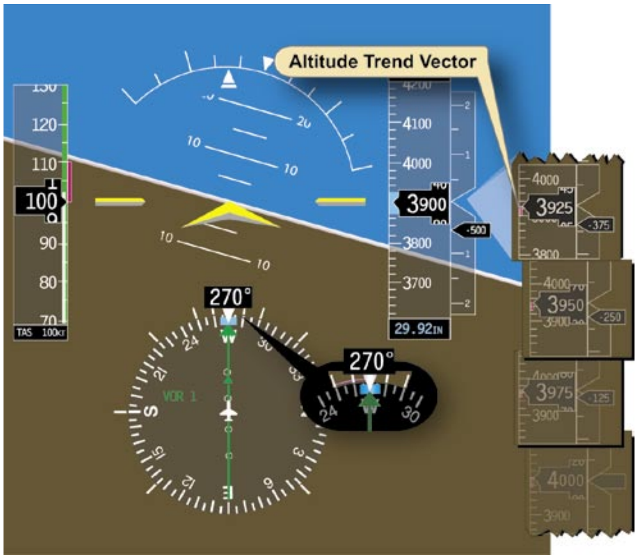

Figure 4-31. *Supporting Instruments.*

The instrument cross-check necessitates utilizing these supporting instruments to better manage altitude control. The VSI and trend tape provide the pilot with information regarding the direction and rate of altitude deviations. The pilot is thus able to make correction to the pitch attitude before a large deviation in altitude occurs. The airspeed indicator depicts an increase if the pitch attitude is lowered. Conversely, when the pitch attitude increases, the pilot should note a decrease in the airspeed.

Primary Bank

When flying in instrument meteorological conditions, pilots maintain preplanned or assigned headings. With this in mind, the primary instrument for bank angle is the heading indicator. Heading changes are displayed instantaneously. The heading indicator is the only instrument that displays the current magnetic heading, provided that it is matched to the magnetic compass with all deviation adjustments accounted for. *[Figure 4-32]*

There are supporting instruments associated with bank as well. The turn rate trend indicator shows the pilot when the aircraft is changing heading. The magnetic compass is also useful for maintaining a heading; however, it is influenced by several errors in various phases of flight.

Primary Yaw

The slip/skid indicator is the primary instrument for yaw. It is the only instrument that can indicate if the aircraft is moving through the air with the longitudinal axis of the aircraft aligned with the relative wind.

Primary Power

The primary power instrument for straight-and-level flight is the airspeed indicator. The main focus of power is to maintain a desired airspeed during level flight. No other instrument delivers instantaneous indication.

Learning the primary and supporting instruments for each variable is the key to successfully mastering attitude instrument flying. At no point does the primary and supporting method devalue the importance of the attitude indicator or the power instruments. All instruments (control, performance, primary, and supporting) must be utilized collectively.

Fundamental Skills of Attitude Instrument Flying

When first learning attitude instrument flying, it is very important that two major skills be mastered. Instrument cross-check and instrument interpretation comprise the foundation for safely maneuvering the aircraft by reference to instruments alone. Without mastering both skills, the pilot will not be able to maintain precise control of aircraft attitude.

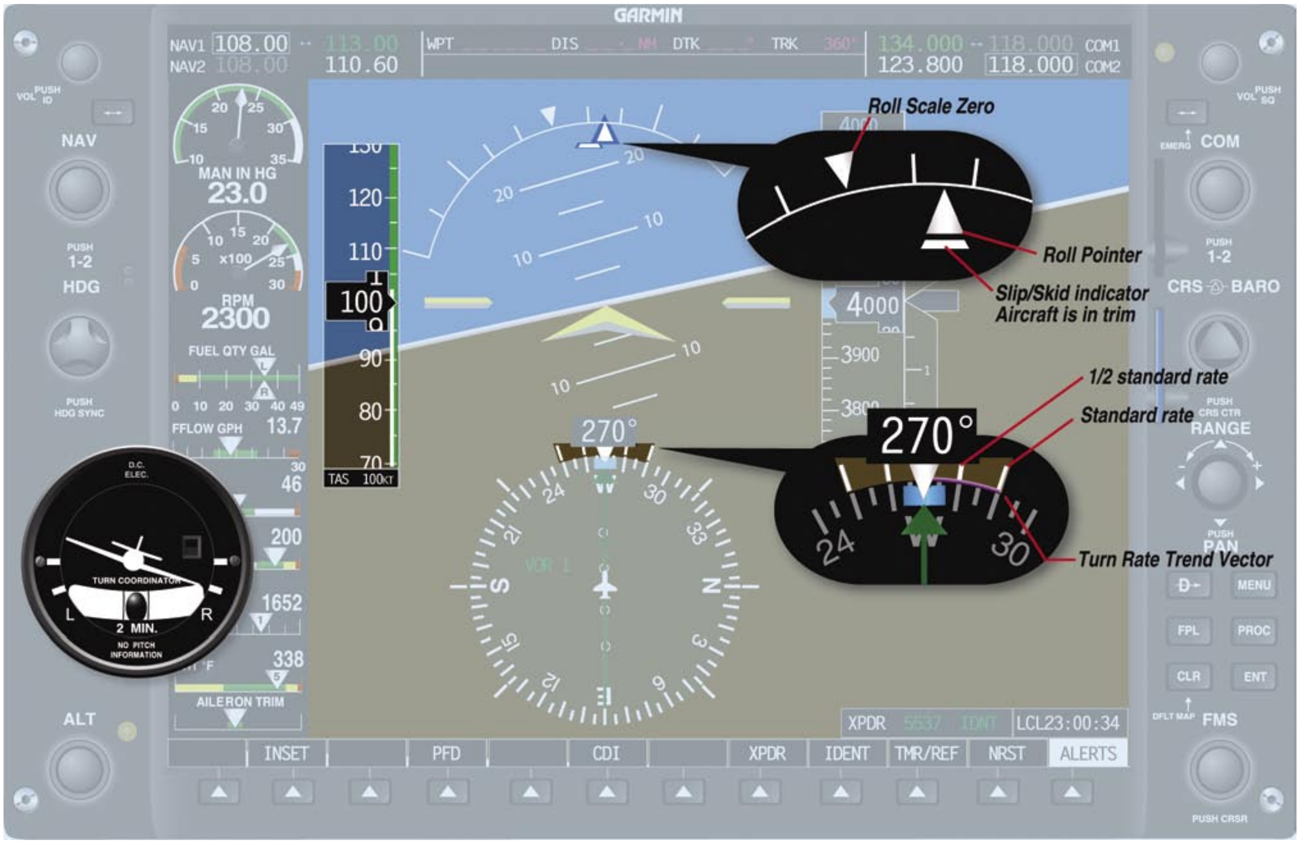
GARMIN
Roll Scale Zero
Roll Pointer
Slip/Skid indicator
Aircraft is in trim
1/2 standard rate
Standard rate
Turn Rate Trend Vector
270°
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
NAV
HDG
ALT
COM
CRS BARO
RANGE
FMS
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 4-32. *Primary Bank.*

Instrument Cross-Check

The first fundamental skill is cross-checking (also call "scanning"). Cross-checking is the continuous observation of the indications on the control and performance instruments. It is imperative that the new instrument pilot learn to observe and interpret the various indications in order to control the attitude and performance of the aircraft. Due to the configuration of some glass panel displays such as the Garmin G1000, one or more of the performance instruments may be located on an MFD installed to the right of the pilot's direct forward line of sight. *[Figure 4-33]*

How a pilot gathers the necessary information to control the aircraft varies by individual pilot. No specific method of cross-checking (scanning) is recommended; the pilot must learn to determine which instruments give the most pertinent information for any particular phase of a maneuver. With practice, the pilot is able to observe the primary instruments quickly and cross-check with the supporting instruments in order to maintain the desired attitude. At no time during instrument flying should the pilot stop cross-checking the instrumentation.

Scanning Techniques

Since most of the primary and supporting aircraft attitude information is displayed on the PFD, standard scanning techniques can be utilized. It is important to remember to include the stand-by flight instruments as well as the engine indications in the scan. Due to the size of the attitude instrument display, scanning techniques have been simplified because the attitude indicator is never out of peripheral view.

Selected Radial Cross-Check

The radial scan is designed so that your eyes remain on the attitude indicator 80–90 percent of the time. The remainder of the time is spent transitioning from the attitude indicator to the various other flight instruments. *[Figure 4-34]*

The radial scan pattern works well for scanning the PFD. The close proximity of the instrument tape displays necessitates very little eye movement in order to focus in on the desired instrument. While the eyes move in any direction, the extended artificial horizon line allows the pilot to keep the pitch attitude in his or her peripheral vision. This extended horizon line greatly reduces the tendency to fixate on one instrument and completely ignore all others. Because of the size of the attitude display, some portion of the attitude indicator is always visible while viewing another instrument display on the PFD.

Starting the Scan

Start the scan in the center of the PFD on the yellow chevron. Note the pitch attitude and then transition the eyes upward to the slip/skid indicator. Ensure that the aircraft is coordinated

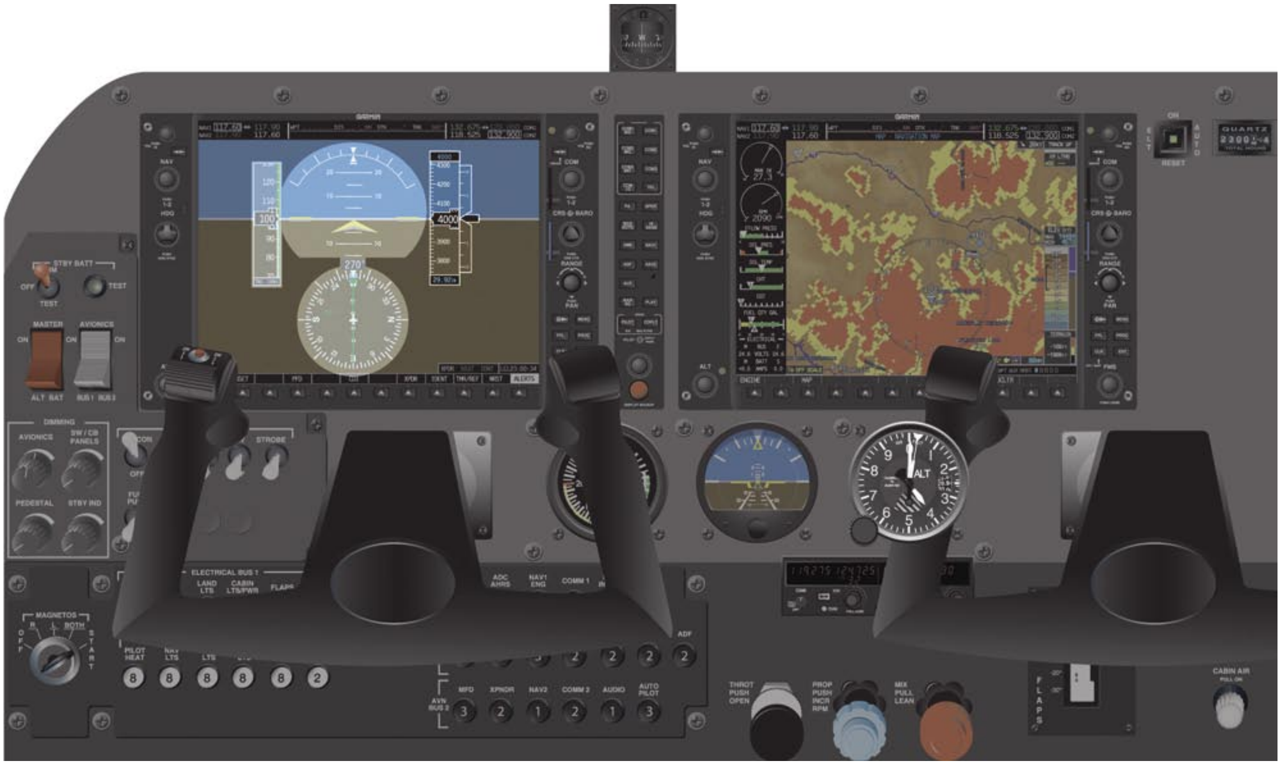

Figure 4-33. *Note that the altitude and vertical speed tapes are slightly to the right of the pilot's direct forward line of sight.*

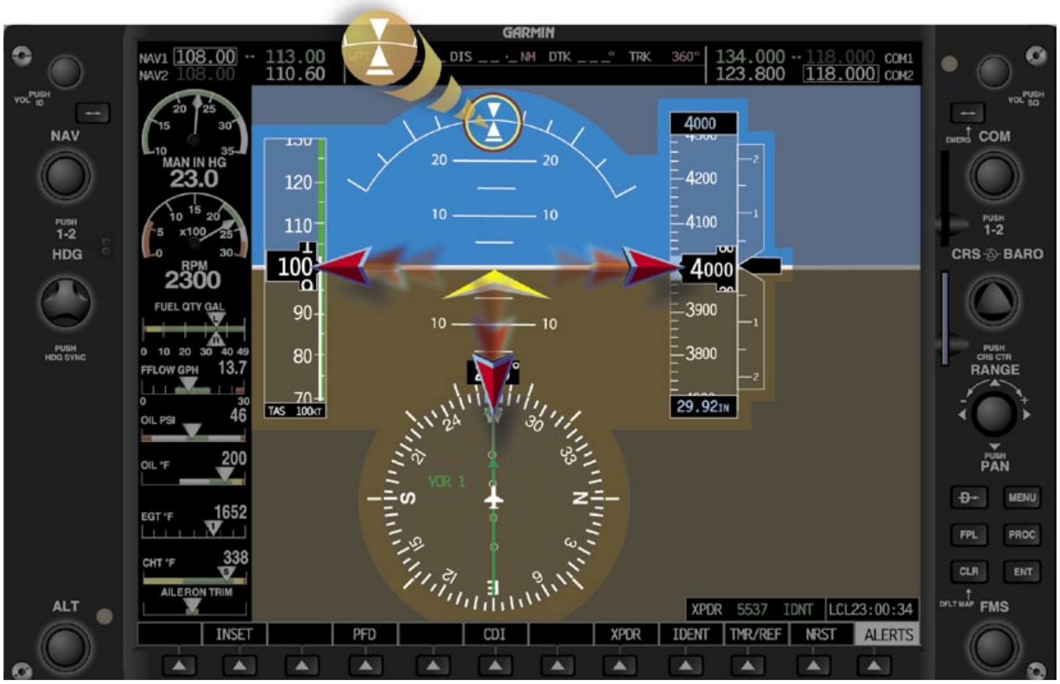
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
DIS _ _ ._ NM DTK _ _ _° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
MAN IN HG 23.0
RPM 2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
120
110
100
90
80
TAS 100KT
4000
4200
4100
4000
3900
3800
29.92IN
VOR 1
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS

Figure 4-34. *Selected Radial Cross-Check.*

by aligning the split triangle symbol. The top of the split triangle is referred to as the roll pointer. The lower portion of the split triangle is the slip/skid indicator. If the lower portion of the triangle is off to one side, step on the rudder pedal on the same side to offset it. *[Figure 4-35 NOTE: The aircraft is not changing heading. There is no trend vector on the turn rate indicator.]*

While scanning that region, check the roll pointer and assure that the desired degree of roll is being indicated on the bank scale. The roll index and the bank scale remain stationary at the top of the attitude indicator. The index is marked with angles of 10°, 20°, 30°, 45° and 60° in both directions. If the desired bank angle is not indicated, make the appropriate

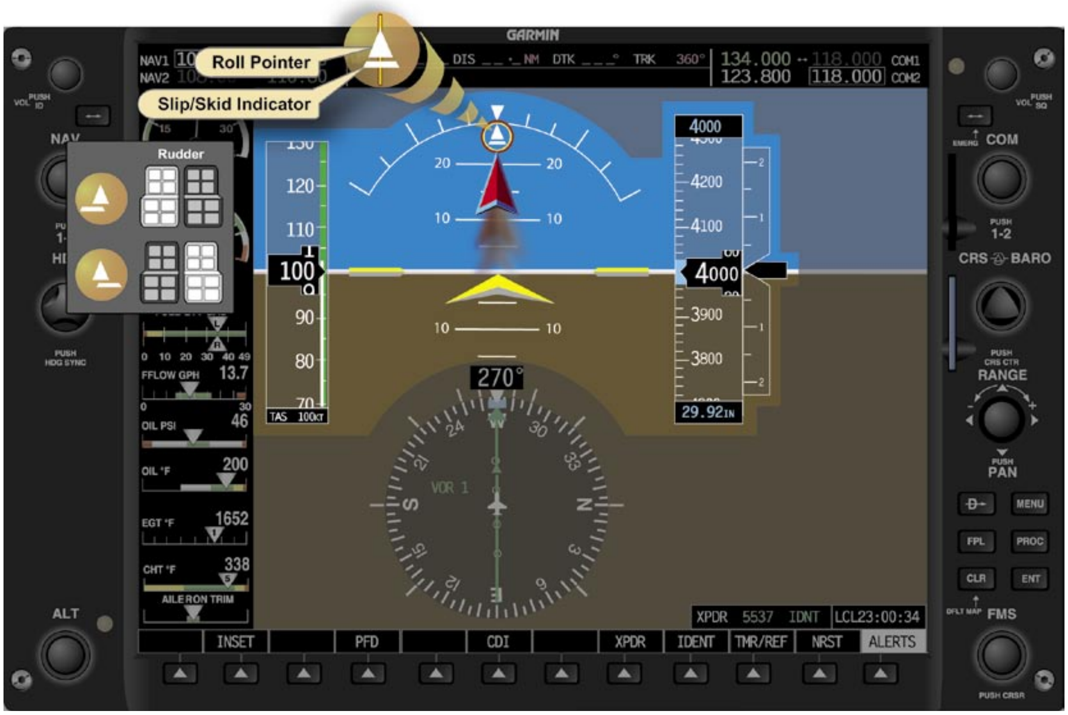
Roll Pointer

Slip/Skid Indicator

Rudder
GARMIN
DIS _ _ ._ NM DTK _ _ _° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
270°
4000
29.92IN
VOR 1
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 4-35. *Roll Pointer and Slip/Skid Indicator.*

aileron corrections. Verify the bank angle is correct and continue scanning back to the yellow chevron.

Scan left to the airspeed tape and verify that the airspeed is as desired, then return back to the center of the display. Scan right to the altimeter tape. Verify that the desired altitude is being maintained. If it is not, make the appropriate pitch change and verify the result. Once the desired altitude has been verified, return to the center of the display. Transition down to the heading indicator to verify the desired heading. When the heading has been confirmed, scan to the center of the display.

It is also important to include the engine indications in the scan. Individualized scan methods may require adjustment if engine indications are presented on a separate MFD. A modified radial scan can be performed to incorporate these instruments into the scan pattern. Another critical component to include in the scan is the moving map display located on the MFD. To aid in situational awareness and facilitate a more centralized scan, a smaller inset map can be displayed in the lower left corner of the PFD screen.

Trend Indicators

One improvement the glass panel displays brought to the general aviation industry is the trend vector. Trend vectors are magenta lines that appear on the airspeed and altitude tapes as well as on the turn rate indicator. These magenta lines indicate what the associated airspeed, altitude, or heading will be in 6 seconds *[Figure 4-36]* if the current rate is maintained. The trend vector is not displayed if there is no change to the associated tape and the value remains constant *[Figure 4-37]* or if there is a failure in some portion of the system that would preclude the vector from being determined.

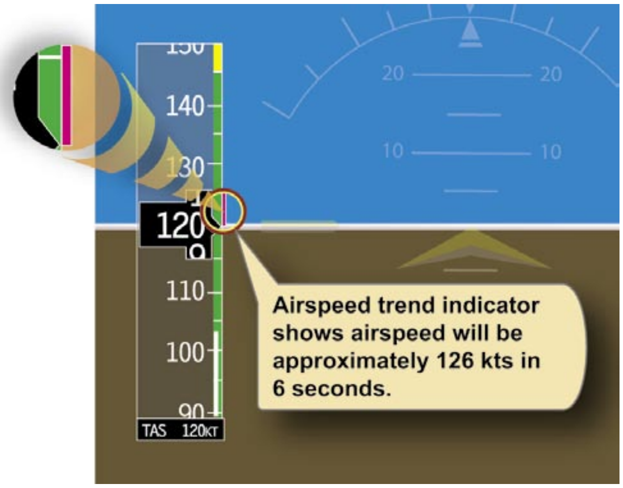

Airspeed trend indicator shows airspeed will be approximately 126 kts in 6 seconds.

Figure 4-36. *Airspeed Trend Indicators.*

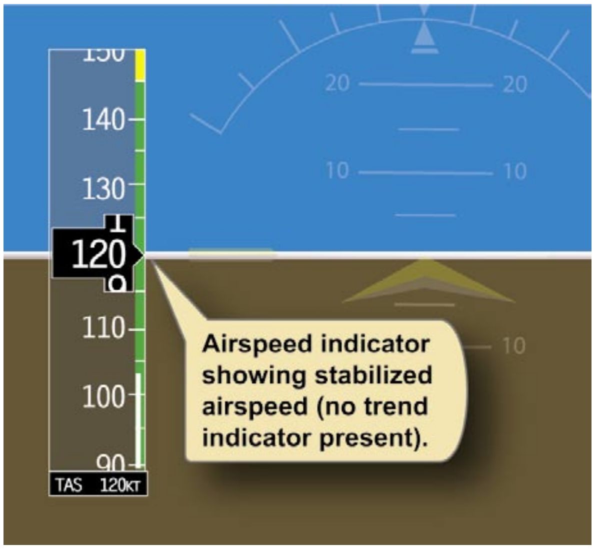

Airspeed indicator showing stabilized airspeed (no trend indicator present).

Figure 4-37. *Airspeed Indicators With No Trend Present.*

Trend vectors are a very good source of information for the new instrument flight rules (IFR) pilot. Pilots who utilize good scanning techniques can pick up subtle deviations from desired parameters and make small correction to the desired attitude. As soon as a trend is indicated on the PFD, a conscientious pilot can adjust to regain the desired attitude. *[Figure 4-38]*

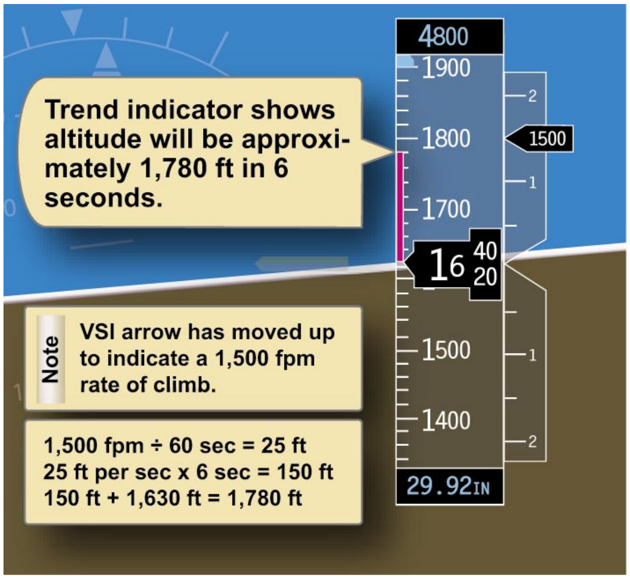

Trend indicator shows altitude will be approximately 1,780 ft in 6 seconds.

Note

VSI arrow has moved up to indicate a 1,500 fpm rate of climb.

1,500 fpm ÷ 60 sec = 25 ft
25 ft per sec x 6 sec = 150 ft
150 ft + 1,630 ft = 1,780 ft

Figure 4-38. *Altimeter Trend Indicators.*

Another advancement in attitude instrument flying is the turn rate trend indicator. As in the cases of airspeed, altitude, and vertical speed trend indicators, the turn rate trend indicator depicts what the aircraft's heading will be in 6 seconds. While examining the top of the heading indicator, notice two white lines on the exterior of the compass rose. *[Figure 4-39]* These

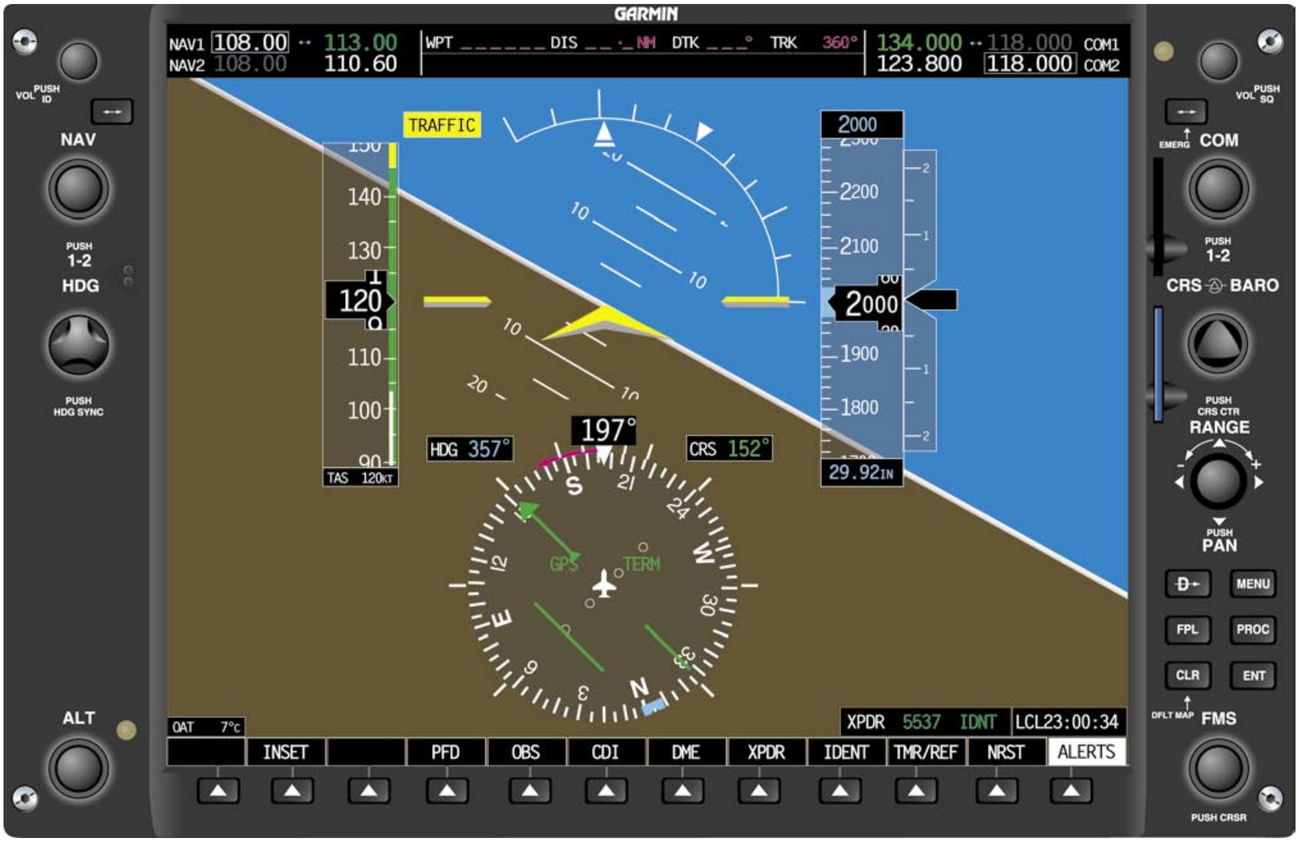
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK ° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
TRAFFIC
140
130
120
110
100
TAS 120KT
2000
2200
2100
1900
1800
29.92IN
HDG 357°
197°
CRS 152°
GPS
TERM
OAT 7°C
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
OBS
CDI
DME
XPDR
IDENT
TMR/REF
NRST
ALERTS
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR

Figure 4-39. *Horizontal Situation Indicator (HSI) Trend Indicator Elongates Proportionally With the Rate of Turn.*

two tick marks located on both sides of the top of the heading indicator show half-standard rate turns as well as standard rate turns.

In *Figure 4-40*, when the aircraft begins its turn to the left, the magenta trend indicator elongates proportionally with the rate of turn. To initiate a half-standard rate turn, position the indicator on the first tick mark. A standard rate turn would be indicated by the trend indicator extending to the second tick mark. A turn rate in excess of standard rate would be indicated by the trend indicator extending past the second tick mark. This trend indicator shows what the aircraft's heading will be in 6 seconds, but is limited to indicate no more than 24° in front of the aircraft, or 4° per second. When the aircraft exceeds a turning rate of 25° in 6 seconds, the trend indicator has an arrowhead attached to it.

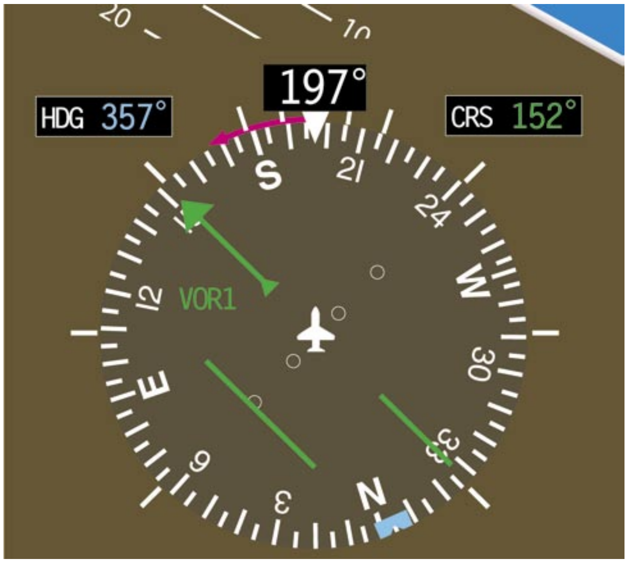
HDG 357°
197°
CRS 152°
S
21
24
W
30
33
N
3
6
E
12
VOR1

Figure 4-40. *HSI Indicator (enlargement).*

Trend indicators are very useful when leveling off at a specific altitude, when rolling out on a heading, or when stabilizing airspeed. One method of determining when to start to level off from a climb or descent is to start leveling at 10 percent of the vertical speed rate prior to the desired altitude.

As the aircraft approaches the desired altitude, adjust the pitch attitude to keep the trend indicator aligned with the target altitude. As the target approaches, the trend indicator gradually shrinks until altitude stabilizes. Trend indicators should be used as a supplement, not as a primary means of determining pitch change.

Common Errors

Fixation

Fixation, or staring at one instrument, is a common error observed in pilots first learning to utilize trend indicators. The pilot may initially fixate on the trend indicator and make adjustments with reference to that alone. Trend indicators are not the only tools to aid the pilot in maintaining the desired power or attitude; they should be used in conjunction with the primary and supporting instruments in order to better manage the flight. With the introduction of airspeed tapes, the pilot can monitor airspeed to within one knot. Fixation can lead to attempting to keep the airspeed to an unnecessarily tight tolerance. There is no need to hold airspeed to within one knot; the Instrument Rating Practical Test Standards (PTS) allows greater latitude.

Omission

Another common error associated with attitude instrument flying is omission of an instrument from the cross-check. Due to the high reliability of the PFD and associated components, pilots tend to omit the stand-by instruments as well as the magnetic compass from their scans. An additional reason for the omission is the position of the stand-by instruments. Pilots should continue to monitor the stand-by instruments in order to detect failures within those systems. One of the most commonly omitted instruments from the scan is the slip/skid indicator.

Emphasis

In initial training, placing emphasis on a single instrument is very common and can become a habit if not corrected. When the importance of a single instrument is elevated above another, the pilot begins to rely solely on that instrument for guidance. When rolling out of a 180° turn, the attitude indicator, heading indicator, slip/skid indicator, and altimeter need to be referenced. If a pilot omits the slip/skid indicator, coordination is sacrificed.

Chapter 5, Section I

Airplane Basic Flight Maneuvers

Using Analog Instrumentation

Introduction

Instrument flying techniques differ according to aircraft type, class, performance capability, and instrumentation. Therefore, the procedures and techniques that follow need to be modified to suit individual aircraft. Recommended procedures, performance data, operating limitations, and flight characteristics of a particular aircraft are available in the Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM) for study before practicing the flight maneuvers.

The flight maneuvers discussed here in Chapter 5-I assume the use of a single-engine, propeller-driven small airplane with retractable gear and flaps and a panel with instruments representative of those discussed earlier in Chapter 3, Flight Instruments. With the exception of the instrument takeoff, all of the maneuvers can be performed on "partial panel," with the attitude gyro and heading indicator covered or inoperative.

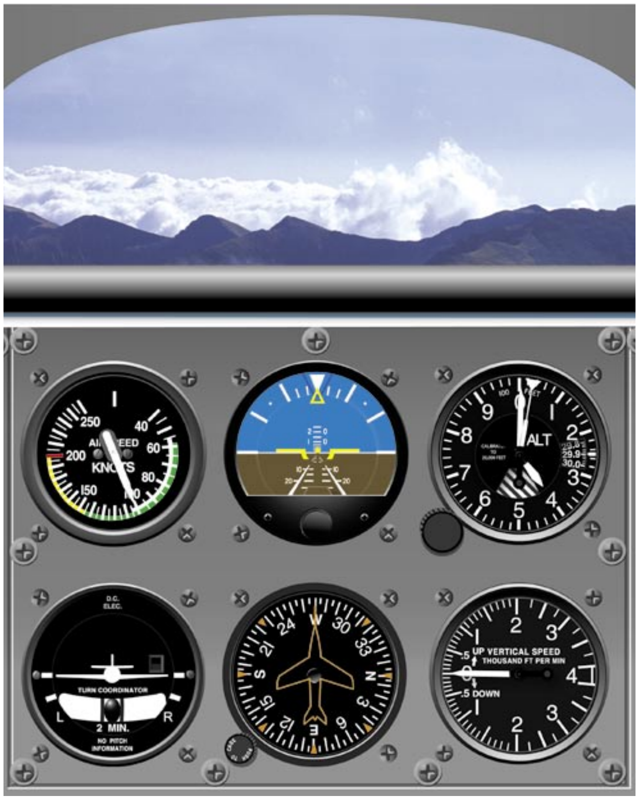
AIRSPEED KNOTS
ALT
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
D.C. ELEC.
L
R
UP VERTICAL SPEED THOUSAND FT PER MIN
.5 DOWN

Figure 5-1. *Pitch Attitude and Airspeed in Level Flight, Slow Cruise Speed.*

Straight-and-Level Flight

Pitch Control

The pitch attitude of an airplane is the angle between the longitudinal axis of the airplane and the actual horizon. In level flight, the pitch attitude varies with airspeed and load. For training purposes, the latter factor can normally be disregarded in small airplanes. At a constant airspeed, there is only one specific pitch attitude for level flight. At slow cruise speeds, the level flight attitude is nose high with indications as in *Figure 5-1*; at fast cruise speeds, the level-flight attitude is nose low. *[Figure 5-2] Figure 5-3* shows the indications for the attitude at normal cruise speeds. The instruments used to determine the pitch attitude of the aircraft are the attitude indicator, the altimeter, the vertical speed indicator (VSI), and the airspeed indicator (ASI).

Attitude Indicator

The attitude indicator gives the direct indication of pitch attitude. The desired pitch attitude is gained by using the elevator control to raise or lower the miniature aircraft in relation to the horizon bar. This corresponds to the way pitch attitude is adjusted in visual flight by raising or lowering the nose of the airplane in relation to the natural horizon. However, unless the airspeed is constant, and until the level flight attitude for that airspeed has been identified and established, there is no way to know whether level flight as

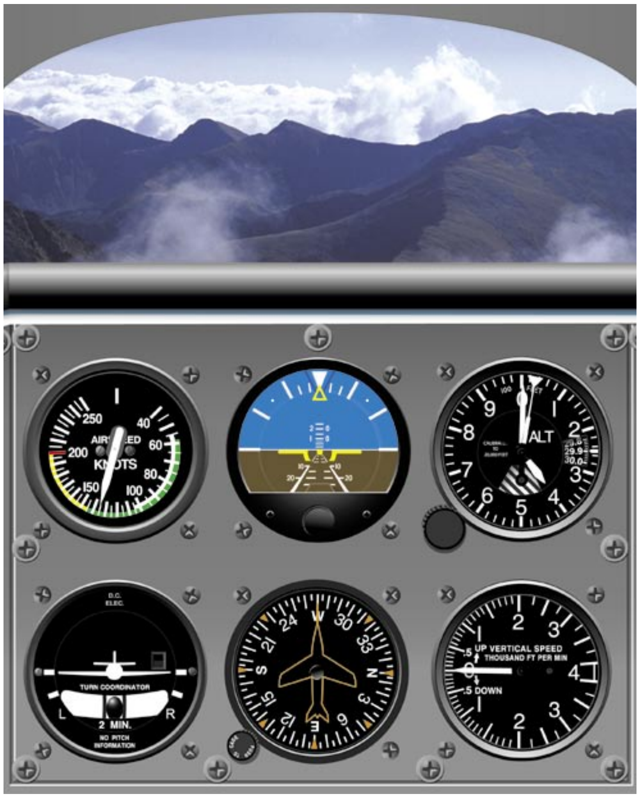
AIRSPEED KNOTS
ALT
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
D.C. ELEC.
L
R
UP VERTICAL SPEED THOUSAND FT PER MIN
.5 DOWN

Figure 5-2. *Pitch Attitude and Airspeed in Level Flight, Fast Cruise Speed.*

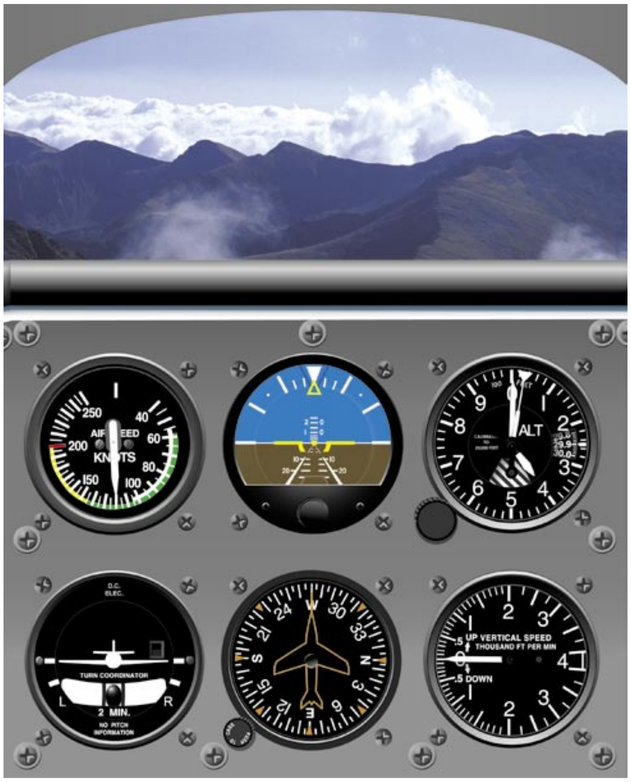
AIRSPEED KNOTS
ALT
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
D.C. ELEC.
L
R
UP VERTICAL SPEED THOUSAND FT PER MIN
.5 DOWN

Figure 5-3. *Pitch Attitude and Airspeed in Level Flight, Normal Cruise Speed.*

indicated on the attitude indicator is resulting in level flight as shown on the altimeter, VSI, and ASI. If the miniature aircraft of the attitude indicator is properly adjusted on the ground before takeoff, it shows approximately level flight at normal cruise speed when the pilot completes the level off from a climb. If further adjustment of the miniature aircraft is necessary, the other pitch instruments must be used to maintain level flight while the adjustment is made.

To practice pitch control for level flight using only the attitude indicator, use the following exercise. Restrict the displacement of the horizon bar to a one-half bar width, a bar width up or down, then a one-and-one-half bar width. One-half, one, and one-and-one-half bar width nose-high attitudes are shown in *Figures 5-4, 5-5,* and *5-6*.

An instructor pilot can demonstrate these normal pitch corrections and compare the indications on the attitude indicator with the airplane's position to the natural horizon.

Pitch attitude changes for corrections to level flight by reference to instruments are much smaller than those commonly used for visual flight. With the airplane correctly trimmed for level flight, the elevator displacement and the control pressures necessary to effect these standard pitch changes are usually very slight. The following are a few helpful hints to help determine how much elevator control pressure is required.

First, a tight grip on the controls makes it difficult to feel control pressure changes. Relaxing and learning to control the aircraft usually takes considerable conscious effort during the early stages of instrument training.

Second, make smooth and small pitch changes with positive pressure. With practice, a pilot can make these small pitch corrections up or down, "freezing" (holding constant) the one-half, full, and one-and-one-half bar widths on the attitude indicator.

Third, with the airplane properly trimmed for level flight, momentarily release all pressure on the elevator control when becoming aware of tenseness. This is a reminder that the airplane is stable; except under turbulent conditions, it will maintain level flight if left alone. Even when no control change is called for, it will be difficult to resist the impulse to move the controls. This may be one of the most difficult initial training problems in instrument flight.

Altimeter

At constant power, any deviation from level flight (except in turbulent air) is the result of a pitch change. Therefore, the altimeter gives an indirect indication of the pitch attitude in level flight, assuming constant power. Since the altitude

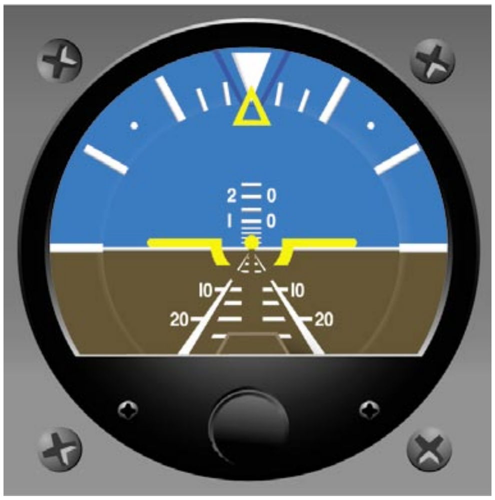

Figure 5-4. *Pitch Correction for Level Flight, One-Half Bar Width.*

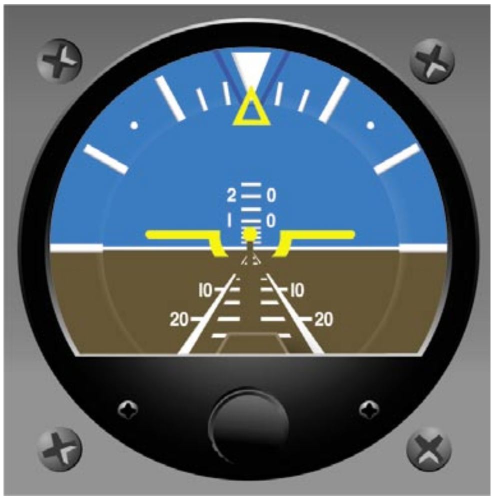

Figure 5-5. *Pitch Correction for Level Flight, One Bar Width.*

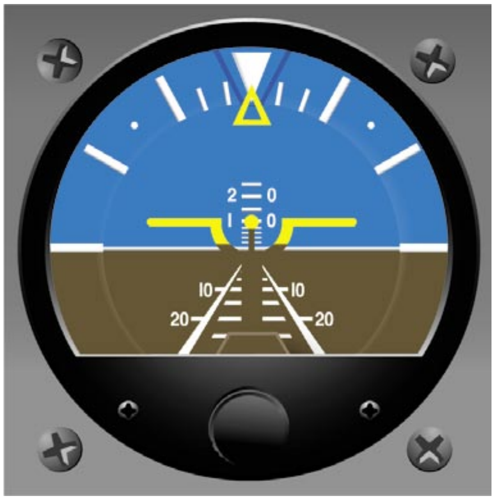

Figure 5-6. *Pitch Correction for Level Flight, One-and-One-Half Bar Width.*

should remain constant when the airplane is in level flight, any deviation from the desired altitude signals the need for a pitch change. If the aircraft is gaining altitude, the nose must be lowered. *[Figures 5-7* and *5-8]*

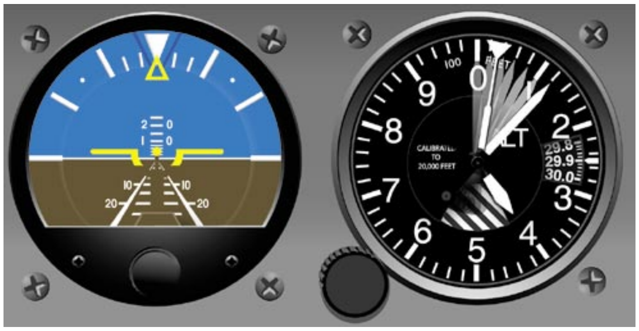

Figure 5-7. *Using the Altimeter for Pitch Interpretation, a High Altitude Means a Nose-High Pitch Attitude.*

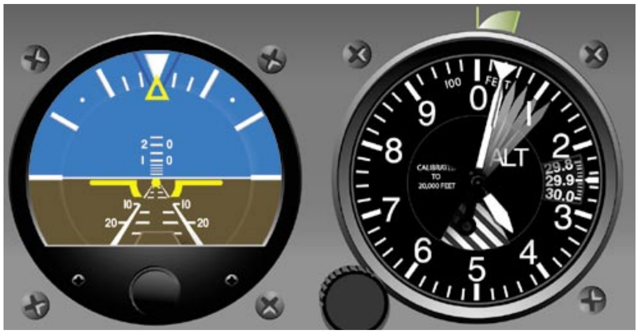

Figure 5-8. *Pitch Correction Following Altitude Increase—Lower Nose to Correct Altitude Error.*

The rate of movement of the altimeter needle is as important as its direction of movement in maintaining level flight without the use of the attitude indicator. An excessive pitch deviation from level flight results in a relatively rapid change of altitude; a slight pitch deviation causes a slow change. Thus, if the altimeter needle moves rapidly clockwise, assume a considerable nose-high deviation from level flight attitude. Conversely, if the needle moves slowly counterclockwise to indicate a slightly nose-low attitude, assume that the pitch correction necessary to regain the desired altitude is small. As the altimeter is added to the attitude indicator in a cross-check, a pilot will learn to recognize the rate of movement of the altimeter needle for a given pitch change as shown on the attitude indicator.

To practice precision control of pitch in an airplane without an attitude indicator, make small pitch changes by visual reference to the natural horizon, and note the rate of movement of the altimeter. Note what amount of pitch change gives the slowest steady rate of change on the altimeter. Then practice small pitch corrections by accurately interpreting and controlling the rate of needle movement.

An instructor pilot can demonstrate an excessive nose-down deviation (indicated by rapid movement of the altimeter needle) and then, as an example, show the result of improper corrective technique. The normal impulse is to make a large pitch correction in a hurry, but this inevitably leads to overcontrolling. The needle slows down, then reverses direction, and finally indicates an excessive nose-high deviation. The result is tension on the controls, erratic control response, and increasingly extreme control movements. The correct technique, which is slower and smoother, will return the airplane to the desired attitude more quickly, with positive control and no confusion.

When a pitch error is detected, corrective action should be taken promptly, but with light control pressures and two distinct changes of attitude: (1) a change of attitude to stop the needle movement and (2) a change of attitude to return to the desired altitude.

When the altimeter indicates an altitude deviation, apply just enough elevator pressure to decrease the rate of needle movement. If it slows down abruptly, ease off some of the pressure until the needle continues to move, but ease off slowly. Slow needle movement means the airplane attitude is close to level flight. Add slightly more corrective pressure to stop the direction of needle movement. At this point level flight is achieved; a reversal of needle movement means the aircraft has passed through it. Relax control pressures carefully, continuing to cross-check since changing airspeed will cause changes in the effectiveness of a given control pressure. Next, adjust the pitch attitude with elevator pressure for the rate of change of altimeter needle movement that is correlated with normal pitch corrections, and return to the desired altitude.

As a rule of thumb, for errors of less than 100 feet, use a half bar width correction. *[Figures 5-9* and *5-10]* For errors in excess of 100 feet, use an initial full bar width correction. *[Figures 5-11* and *5-12]* Practice predetermined altitude changes using the altimeter alone, then in combination with the attitude indicator.

Vertical Speed Indicator (VSI)

The VSI, like the altimeter, gives an indirect indication of pitch attitude and is both a trend and a rate instrument. As a trend instrument, it shows immediately the initial vertical movement of the airplane, which disregarding turbulence can be considered a reflection of pitch change. To maintain level flight, use the VSI in conjunction with the altimeter and attitude indicator. Note any positive or negative trend of the needle from zero and apply a very light corrective elevator

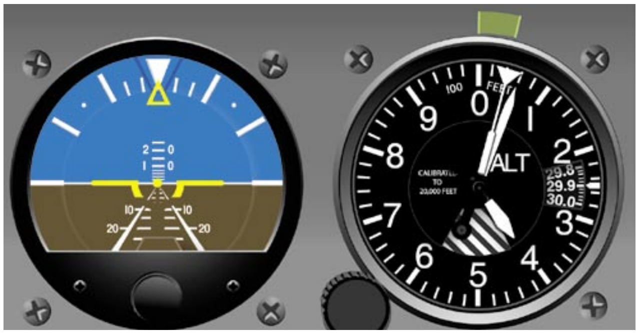

Figure 5-9. *Altitude Error, Less Than 100 Feet.*

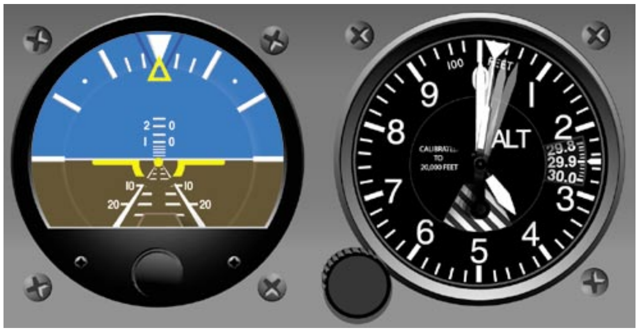

Figure 5-10. *Pitch Correction, Less Than 100 Feet—One-Half Bar Low to Correct Altitude Error.*

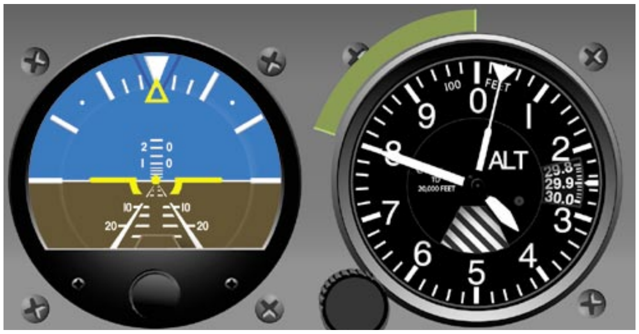

Figure 5-11. *Altitude Error, Greater Than 100 Feet.*

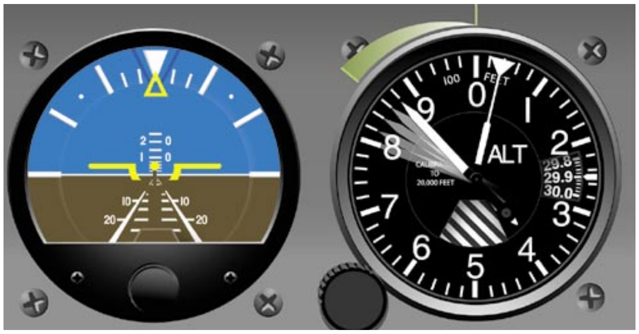

Figure 5-12. *Pitch Correction, Greater Than 100 Feet—One Bar Correction Initially.*

pressure. As the needle returns to zero, relax the corrective pressure. If control pressures have been smooth and light, the needle reacts immediately and slowly, and the altimeter shows little or no change of altitude. As a rate instrument, the VSI requires consideration of lag characteristics.

Lag refers to the delay involved before the needle attains a stable indication following a pitch change. Lag is directly proportional to the speed and magnitude of a pitch change. If a slow, smooth pitch change is initiated, the needle moves with minimum lag to a point of deflection corresponding to the extent of the pitch change, and then stabilizes as the aerodynamic forces are balanced in the climb or descent. A large and abrupt pitch change produces erratic needle movement, a reverse indication, and introduces greater time delay (lag) before the needle stabilizes. Pilots are cautioned not to chase the needle when flight through turbulent conditions produces erratic needle movements. The apparent lag in airspeed indications with pitch changes varies greatly among different airplanes and is due to the time required for the airplane to accelerate or decelerate when the pitch attitude is changed. There is no appreciable lag due to the construction or operation of the instrument. Small pitch changes, smoothly executed, result in an immediate change of airspeed.

When using the VSI as a rate instrument and combining it with the altimeter and attitude indicator to maintain level flight, a pilot should know that the amount the altimeter needle moves from the desired altitude governs the rate which should be used to return to that altitude. A rule of thumb is to make an attitude change that will result in a vertical-speed rate approximately double the error in altitude. For example, if altitude is off by 100 feet, the rate of return to the desired altitude should be approximately 200 feet per minute (fpm). If it is off by more than 100 feet, the correction should be correspondingly greater, but should never exceed the optimum rate of climb or descent for the airplane at a given airspeed and configuration.

A deviation of more than 200 fpm from the desired rate of return is considered overcontrolling. For example, if attempting to change altitude by 200 feet, a rate in excess of 400 fpm indicates overcontrolling.

When returning to an altitude, the VSI is the primary pitch instrument. Occasionally, the VSI is slightly out of calibration and may indicate a climb or descent when the airplane is in level flight. If the instrument cannot be adjusted, take the error into consideration when using it for pitch control. For example, if the needle indicates a descent of 200 fpm while in level flight, use this indication as the zero position.

Airspeed Indicator (ASI)

The ASI presents an indirect indication of the pitch attitude. In non-turbulent conditions with a constant power setting and pitch attitude, airspeed remains constant. *[Figure 5-13]* As the pitch attitude lowers, airspeed increases, and the nose should be raised. *[Figure 5-14]* As the pitch attitude rises, airspeed decreases, and the nose should be lowered. *[Figure 5-15]* A rapid change in airspeed indicates a large pitch change, and a slow change of airspeed indicates a small pitch change.

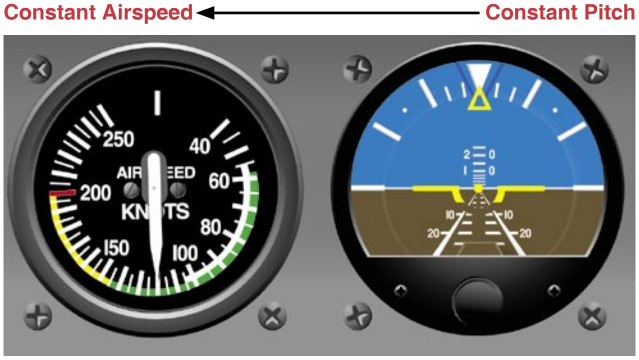

Figure 5-13. *Constant Power Plus Constant Pitch Equals Constant Speed.*

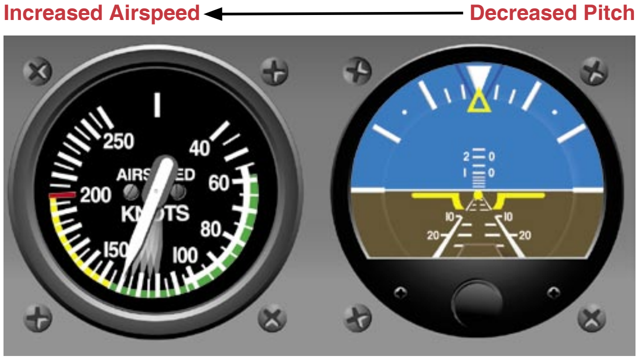

Figure 5-14. *Constant Power Plus Decreased Pitch Equals Increased Airspeed.*

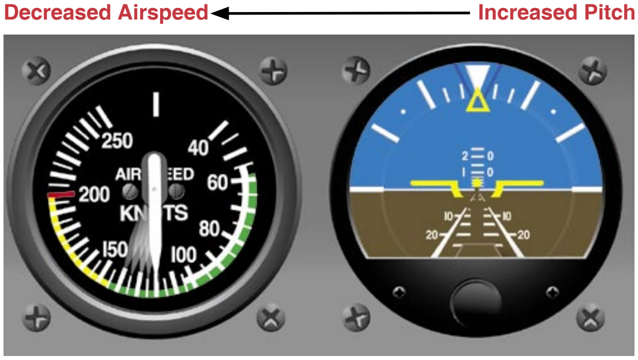

Figure 5-15. *Constant Power Plus Increased Pitch Equals Decreased Airspeed.*

Pitch control in level flight is a question of cross-check and interpretation of the instrument panel for the instrument information that enables a pilot to visualize and control pitch attitude. Regardless of individual differences in cross-check technique, all pilots should use the instruments that give the best information for controlling the airplane in any given maneuver. Pilots should also check the other instruments to aid in maintaining the primary instruments at the desired indication.

As noted previously, the primary instrument is the one that gives the most pertinent information for a particular maneuver. It is usually the one that should be held at a constant indication. Which instrument is primary for pitch control in level flight, for example? This question should be considered in the context of specific airplane, weather conditions, pilot experience, operational conditions, and other factors. Attitude changes must be detected and interpreted instantly for immediate control action in high-performance airplanes. On the other hand, a reasonably proficient instrument pilot in a slower airplane may rely more on the altimeter for primary pitch information, especially if it is determined that too much reliance on the attitude indicator fails to provide the necessary precise attitude information. Whether the pilot decides to regard the altimeter or the attitude indicator as primary depends on which approach will best help control the attitude. In this handbook, the altimeter is normally considered as the primary pitch instrument during level flight.

Bank Control

The bank attitude of an airplane is the angle between the airplane's wings and the natural horizon. To maintain a straight-and-level flight path, the wings of the airplane are kept level with the horizon (assuming the airplane is in coordinated flight). The instruments used for bank control are the attitude indicator, the heading indicator, and the turn coordinator. *Figure 5-16* illustrates coordinated flight. The aircraft is banked left with the attitude indicator and turn coordinator indicating the bank. The heading indicator indicates a left turn by apparent clockwise rotation of the compass card behind the airplane silhouette.

Attitude Indicator

The attitude indicator shows any change in bank attitude directly and instantly and is, therefore, a direct indicator. On the standard attitude indicator, the angle of bank is shown pictorially by the relationship of the miniature aircraft to the artificial horizon bar, and by the alignment of the pointer with the banking scale at the top of the instrument. On the face of the standard three-inch instrument, small angles of bank can be difficult to detect by reference to the miniature aircraft, especially if leaning to one side or changing a seating position

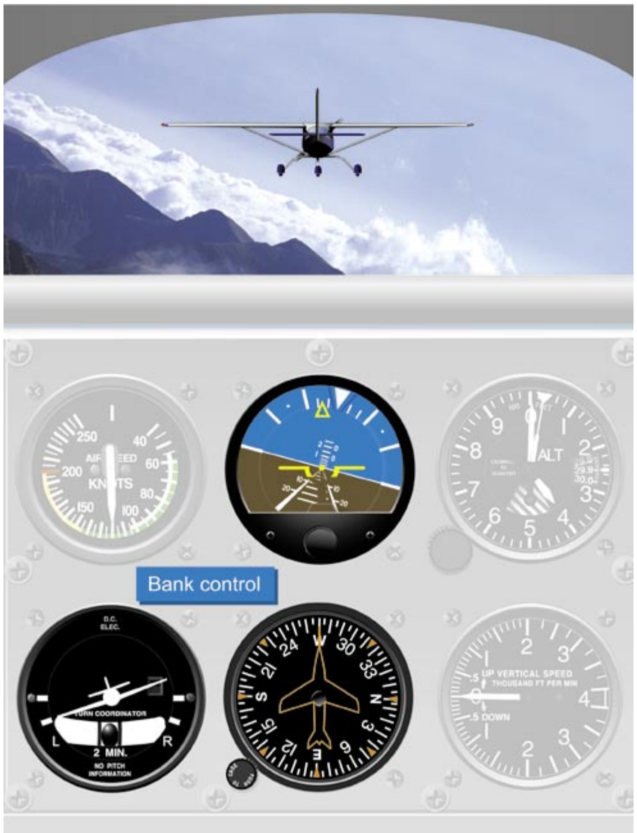

Figure 5-16. *Instruments Used for Bank Control.*

slightly. The position of the scale pointer is a good check against the apparent miniature aircraft position. Disregarding precession error, small deviations from straight coordinated flight can be readily detected on the scale pointer. The banking index may be graduated as shown in *Figure 5-17*, or it may be graduated in 30° increments.

The instrument depicted in *Figure 5-17* has a scale pointer that moves in the same direction of bank shown by the miniature aircraft. In this case, the aircraft is in a left 15° bank. Precession errors in this instrument are common and predictable, but the obvious advantage of the attitude indicator is an immediate indication of both pitch attitude and bank attitude in a single glance. Even with the precession errors associated with many attitude indicators, the quick attitude presentation requires less visual effort and time for positive control than other flight instruments.

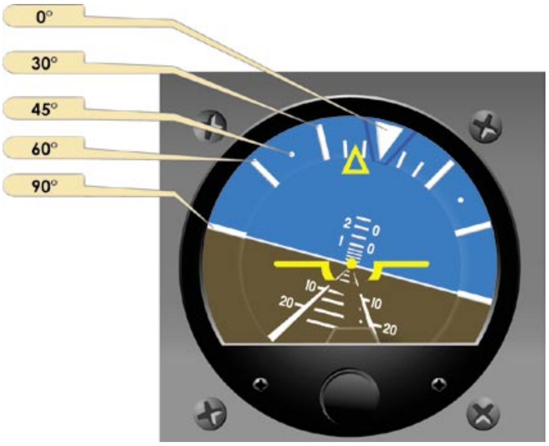

Figure 5-17. *Bank Interpretation with the Attitude Indicator.*

Heading Indicator

The bank attitude of an aircraft in coordinated flight is shown indirectly on the heading indicator, since banking results in a turn and change in heading. Assuming the same airspeed in both instances, a rapid movement of the heading indicator (azimuth card in a directional gyro) indicates a large angle of bank, whereas slow movement reflects a small angle of bank. Note the rate of movement of the heading indicator and compare it to the attitude indicator's degrees of bank. The attitude indicator's precession error makes a precise check of heading information necessary in order to maintain straight flight.

When deviations from straight flight are noted on the heading indicator, correct to the desired heading using a bank angle no greater than the number of degrees to be turned. In any case, limit bank corrections to a bank angle no greater than that required for a standard rate turn. Use of larger bank angles requires a very high level of proficiency, and normally results in overcontrolling and erratic bank control.

Turn Coordinator

The miniature aircraft of the turn coordinator gives an indirect indication of the bank attitude of the airplane. When the miniature aircraft is level, the airplane is in straight flight. When the miniature airplane is aligned with one of the alignment marks and the aircraft is rolling to the left or right the indication represents the roll rate, with the alignment marks indicating a roll of 3° per second in the direction of the miniature aircraft. This can be seen in level flight when a bank is introduced either to the left or the right. The turn coordinator's indicator will indicate the rolling motion although there is no turn being made. Conversely, a pedal input to the right or left causes the aircraft to turn momentarily about its vertical axis (with no rolling motion) with an indication of turn on the turn coordinator. After the turn becomes stabilized and the aircraft is no longer rolling, the turn coordinator displays the rate of turn with the alignment marks equaling a turn of 3° per second. The turn coordinator is able to display both roll and turn parameters because its electrically powered gyroscope is canted at an angle. As a result, the turn-and-slip indicator provides both roll and turn indications. Autopilots in general aviation today use this instrument in determining both roll and turn information. After the completion of a turn, return to straight flight is accomplished by coordinated aileron and rudder pressure to level the miniature aircraft. Include the miniature aircraft in the cross-check and correct for even the smallest deviations from the desired position. When this instrument is used to maintain straight flight, control pressures must be applied very lightly and smoothly.

The ball of the turn coordinator is actually a separate instrument, conveniently located under the miniature aircraft because the two instruments are used together. The ball instrument indicates the quality of the turn. If the ball is off center, the airplane is slipping or skidding. That is, if the coordinator's miniature airplane is tilted right and the ball is displaced to the right, the aircraft is in a skid. *[Figure 5-18]* If however, the miniature airplane is tilted to the right with the ball off-center to the left, the aircraft is in a slip. *[Figure 5-19]* If the wings are level and the airplane is properly trimmed, the ball will remain in the center, and the airplane will be in straight flight. If the ball is not centered, the airplane is improperly trimmed.

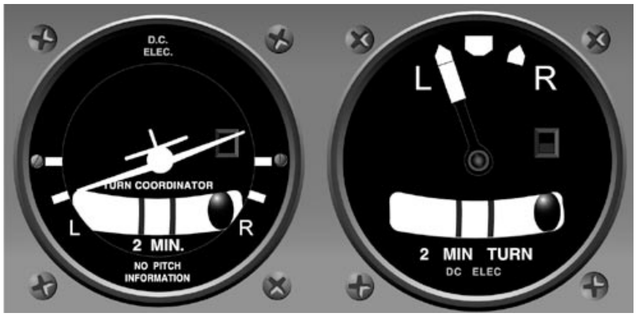

Figure 5-18. *Skid Indication.*

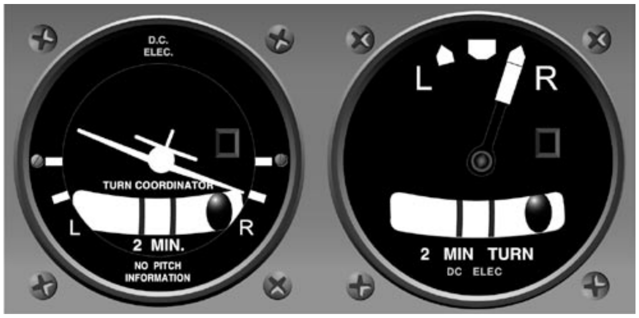

Figure 5-19. *Slip Indication.*

To maintain straight-and-level flight with proper trim, note the direction of ball displacement. If the ball is to the left of center and the left wing is low, apply left rudder pressure to center the ball and correct the slip. At the same time apply right aileron pressure as necessary to level the wings, cross-checking the heading indicator and attitude indicator while centering the ball. If the wings are level and the ball is displaced from the center, the airplane is skidding. Note the direction of ball displacement, and use the same corrective technique as for an indicated slip. Center the ball (left ball/left rudder, right ball/right rudder), use aileron as necessary for bank control, and retrim.

To trim the airplane using only the turn coordinator, use aileron pressure to level the miniature aircraft and rudder pressure to center the ball. Hold these indications with control pressures, gradually releasing them while applying rudder trim sufficient to relieve all rudder pressure. Apply aileron trim, if available, to relieve aileron pressure. With a full instrument panel, maintain a wings level attitude by reference to all available instruments while trimming the airplane.

Turn-and-Slip Indicator (Needle and Ball)

Unlike the turn coordinator that provides three indications (roll, turn, and trim), the turn-and-slip indicator provides two: turn-rate and trim. Although the turn-and-slip indicator needle provides an indication of turn only, it provides an indirect indication of aircraft attitude when used with roll indicators such as a heading indicator or magnetic compass. As with the turn coordinator (after stabilizing from a roll), when the turn-and-slip indicator's needle is aligned with the alignment marks the aircraft is in a standard turn of 3° per second or 360° in 2 minutes.

The ball of the turn-and-bank indicator provides important trim in the same manner that the ball in the turn coordinator does. *Figures 5-18* and *5-19* provide a comparison of the two instruments.

Power Control

Power produces thrust which, with the appropriate angle of attack of the wing, overcomes the forces of gravity, drag, and inertia to determine airplane performance.

Power control must be related to its effect on altitude and airspeed, since any change in power setting results in a change in the airspeed or the altitude of the airplane. At any given airspeed, the power setting determines whether the airplane is in level flight, in a climb, or in a descent. If the power is increased in straight-and-level flight and the airspeed held constant, the airplane climbs. If power is decreased while the airspeed is held constant, the airplane descends. On the other hand, if altitude is held constant, the power applied will determine the airspeed.

The relationship between altitude and airspeed determines the need for a change in pitch or power. If the airspeed is not the desired value, always check the altimeter before deciding that a power change is necessary. Think of altitude and airspeed as interchangeable; altitude can be traded for airspeed by lowering the nose, or convert airspeed to altitude by raising the nose. If altitude is higher than desired and airspeed is

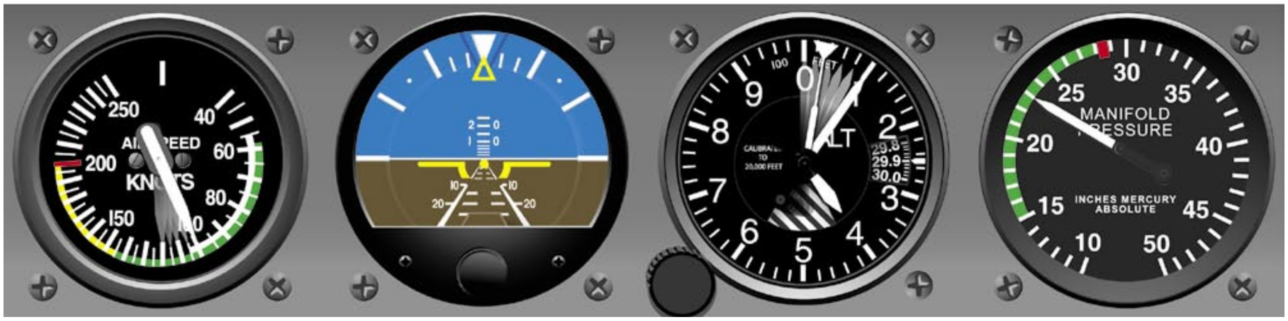
250
40
AIRSPEED
60
200
KNOTS
80
150
100
2
1
0
0
10
20
10
20
100
FEET
9
0
8
ALT
2
CALIBRATED TO 20,000 FEET
29.8
29.9
30.0
7
3
6
5
4
30
25
35
MANIFOLD PRESSURE
20
40
15
INCHES MERCURY ABSOLUTE
45
10
50

Figure 5-20. *Airspeed Low and Altitude High—Lower Pitch.*

low, or vice versa, a change in pitch alone may return the airplane to the desired altitude and airspeed. *[Figure 5-20]* If both airspeed and altitude are high or if both are low, then a change in both pitch and power is necessary in order to return to the desired airspeed and altitude. *[Figure 5-21]*

For changes in airspeed in straight-and-level flight, pitch, bank, and power must be coordinated in order to maintain constant altitude and heading. When power is changed to vary airspeed in straight-and-level flight, a single-engine, propeller-driven airplane tends to change attitude around all axes of movement. Therefore, to maintain constant altitude and heading, apply various control pressures in proportion to the change in power. When power is added to increase airspeed, the pitch instruments indicate a climb unless forward elevator control pressure is applied as the airspeed changes. With an increase in power, the airplane tends to yaw and roll to the left unless counteracting aileron and rudder pressures are applied. Keeping ahead of these changes requires increasing cross-check speed, which varies with the type of airplane and its torque characteristics, the extent of power and speed change involved.

Power Settings

Power control and airspeed changes are much easier when approximate power settings necessary to maintain various airspeeds in straight-and-level flight are known in advance. However, to change airspeed by any appreciable amount, the common procedure is to underpower or overpower on initial power changes to accelerate the rate of airspeed change. (For small speed changes, or in airplanes that decelerate or accelerate rapidly, overpowering or underpowering is not necessary.)

Consider the example of an airplane that requires 23" mercury (Hg) of manifold pressure to maintain a normal cruising airspeed of 120 knots, and 18" Hg of manifold pressure to maintain an airspeed of 100 knots. The reduction in airspeed from 120 knots to 100 knots while maintaining straight-and-level flight is discussed below and illustrated in *Figures 5-22, 5-23,* and *5-24*.

Instrument indications, prior to the power reduction, are shown in *Figure 5-22*. The basic attitude is established and maintained on the attitude indicator. The specific pitch, bank, and power control requirements are detected on these primary instruments:

Altimeter—Primary Pitch
Heading Indicator—Primary Bank
Airspeed Indicator—Primary Power

Supporting pitch-and-bank instruments are shown in *Figure 5-23*. Note that the supporting power instrument is the manifold pressure gauge (or tachometer if the propeller is fixed pitch). However, when a smooth power reduction to approximately 15" Hg (underpower) is made, the manifold pressure gauge becomes the primary power instrument. *[Figure 5-23]* With practice, power setting can be changed with only a brief glance at the power instrument, by sensing

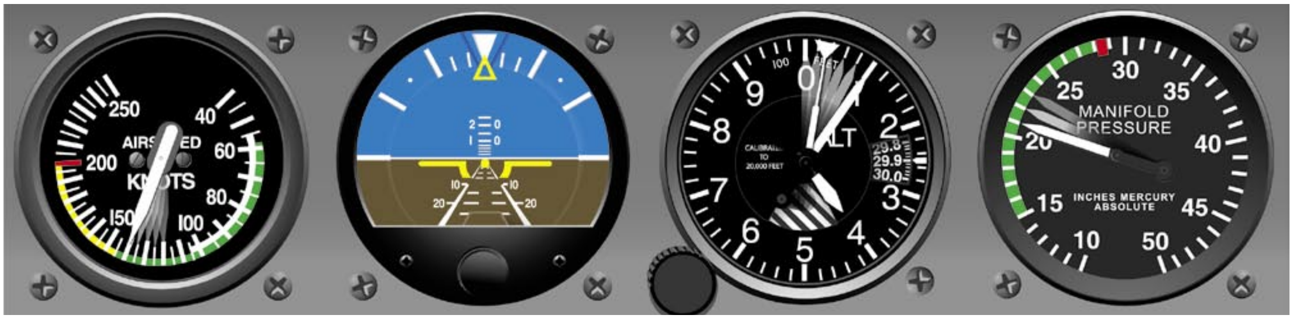
250
40
AIRSPEED
60
200
KNOTS
80
150
100
2
1
0
0
10
20
10
20
100
FEET
9
0
8
ALT
2
CALIBRATED TO 20,000 FEET
29.8
29.9
30.0
7
3
6
5
4
30
25
35
MANIFOLD PRESSURE
20
40
15
INCHES MERCURY ABSOLUTE
45
10
50

Figure 5-21. *Airspeed and Altitude High—Lower Pitch and Reduce Power.*

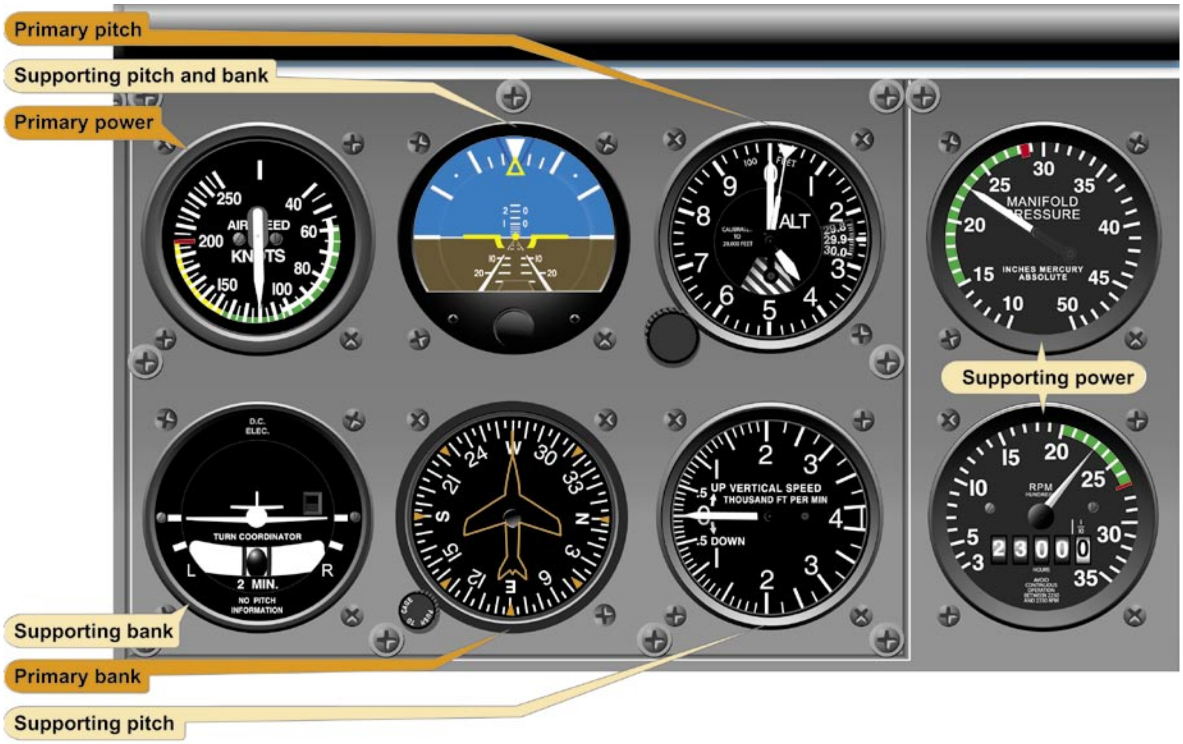
Primary pitch
Supporting pitch and bank
Primary power
Supporting power
Supporting bank
Primary bank
Supporting pitch

Figure 5-22. *Straight-and-Level Flight (Normal Cruising Speed).*

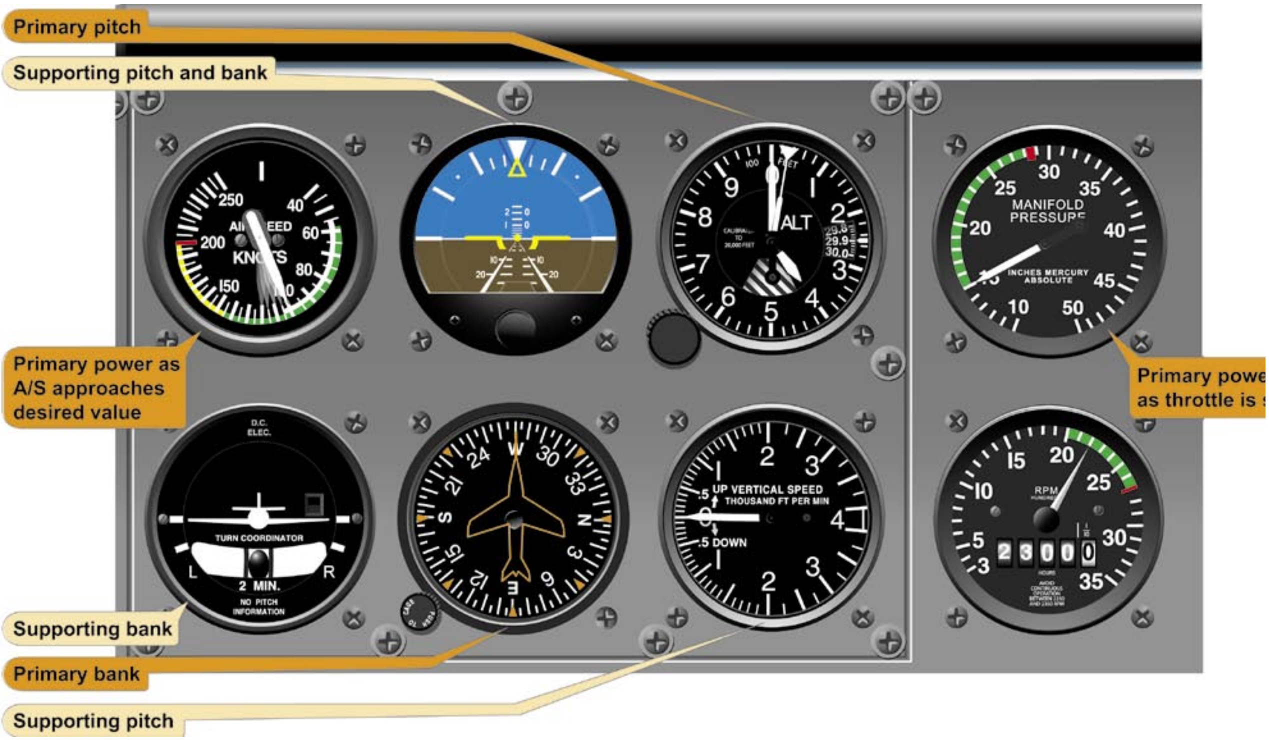
Primary pitch
Supporting pitch and bank
Primary power as A/S approaches desired value
Primary powe as throttle is
Supporting bank
Primary bank
Supporting pitch

Figure 5-23. *Straight-and-Level Flight (Airspeed Decreasing).*

the movement of the throttle, the change in sound, and the changes in the feel of control pressures.

As thrust decreases, increase the speed of the cross-check and be ready to apply left rudder, back-elevator, and aileron control pressure the instant the pitch-and-bank instruments show a deviation from altitude and heading. As proficiency is obtained, a pilot learns to cross-check, interpret, and control the changes with no deviation of heading and altitude. Assuming smooth air and ideal control technique, as airspeed decreases, a proportionate increase in airplane pitch attitude is required to maintain altitude. Similarly, effective torque control means counteracting yaw with rudder pressure.

As the power is reduced, the altimeter is primary for pitch, the heading indicator is primary for bank, and the manifold pressure gauge is momentarily primary for power (at 15" Hg in this example). Control pressures should be trimmed off as the airplane decelerates. As the airspeed approaches the desired airspeed of 100 knots, the manifold pressure is adjusted to approximately 18" Hg and becomes the supporting power instrument. The ASI again becomes primary for power. *[Figure 5-24]*

Airspeed Changes in Straight-and-Level Flight

Practice of airspeed changes in straight-and-level flight provides an excellent means of developing increased proficiency in all three basic instrument skills, and brings out some common errors to be expected during training in straight-and-level flight. Having learned to control the airplane in a clean configuration (minimum drag conditions), increase proficiency in cross-check and control by practicing speed changes while extending or retracting the flaps and landing gear. While practicing, be sure to comply with the airspeed limitations specified in the POH/AFM for gear and flap operation.

Sudden and exaggerated attitude changes may be necessary in order to maintain straight-and-level flight as the landing gear is extended and the flaps are lowered in some airplanes. The nose tends to pitch down with gear extension, and when flaps are lowered, lift increases momentarily (at partial flap settings) followed by a marked increase in drag as the flaps near maximum extension.

Control technique varies according to the lift and drag characteristics of each airplane. Accordingly, knowledge of the power settings and trim changes associated with different combinations of airspeed, gear and flap configurations will reduce instrument cross-check and interpretation problems.

For example, assume that in straight-and-level flight instruments indicate 120 knots with power at 23" Hg/2,300 revolutions per minute (rpm), gear and flaps up. After reduction in airspeed, with gear and flaps fully extended, straight-and-level flight at the same altitude requires 25" Hg manifold pressure/2,500 rpm. Maximum gear extension speed is 115 knots; maximum flap extension speed is 105

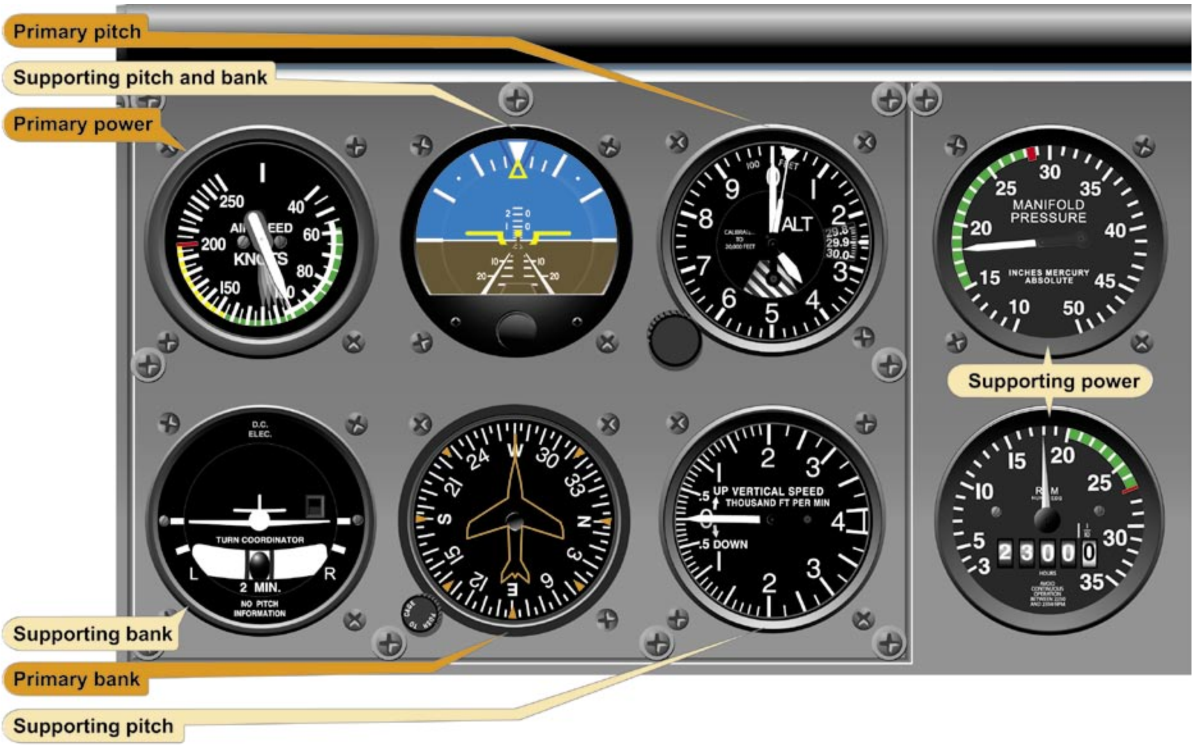

Figure 5-24. *Straight-and-Level Flight (Reduced Airspeed Stabilized).*

knots. Airspeed reduction to 95 knots, gear and flaps down, can be made in the following manner:

1. Maintain rpm at 2,500, since a high power setting will be used in full drag configuration.
2. Reduce manifold pressure to 10" Hg. As the airspeed decreases, increase cross-check speed.
3. Make trim adjustments for an increased angle of attack and decrease in torque.
4. Lower the gear at 115 knots. The nose may tend to pitch down and the rate of deceleration increases. Increase pitch attitude to maintain constant altitude, and trim off some of the back-elevator pressures. If full flaps are lowered at 105 knots, cross-check, interpretation, and control must be very rapid. A simpler technique is to stabilize attitude with gear down before lowering the flaps.
5. Since 18" Hg manifold pressure will hold level flight at 100 knots with the gear down, increase power smoothly to that setting until the ASI shows approximately 105 knots, and retrim. The attitude indicator now shows approximately two-and-a-half bar width nose-high in straight-and-level flight.
6. Actuate the flap control and simultaneously increase power to the predetermined setting (25" Hg) for the desired airspeed, and trim off the pressures necessary to hold constant altitude and heading. The attitude indicator now shows a bar width nose-low in straight-and-level flight at 95 knots.

Proficiency in straight-and-level flight is attained when a pilot can consistently maintain constant altitude and heading with smooth pitch, bank, power, and trim control during the pronounced changes in aircraft attitude.

Trim Technique

Proper trim technique is essential for smooth and precise aircraft control during all phases of flight. By relieving all control pressures, it is much easier to hold a given attitude constant, and devote more attention to other flight deck duties.

An aircraft is trimmed by applying control pressures to establish a desired attitude, then adjusting the trim so the aircraft will maintain that attitude when the flight controls are released. Trim the aircraft for coordinated flight by centering the ball of the turn-and-slip indicator, by using rudder trim in the direction the ball is displaced from the center. Differential power control on multiengine aircraft is an additional factor affecting coordinated flight. Use balanced power or thrust, when possible, to aid in maintaining coordinated flight.

Changes in attitude, power, or configuration will require a trim adjustment in most cases. Using trim alone to establish a change in aircraft attitude invariably leads to erratic aircraft control. Smooth and precise attitude changes are best attained by a combination of control pressures and trim adjustments. Therefore, when used correctly, trim adjustment is an aid to smooth aircraft control.

Common Errors in Straight-and-Level Flight

Pitch

Pitch errors usually result from the following faults:

1. Improper adjustment of the attitude indicator's miniature aircraft to the wings level attitude. Following the initial level off from a climb, check the attitude indicator and make any necessary adjustment in the miniature aircraft for level flight indication at normal cruise airspeed.
2. Insufficient cross-check and interpretation of pitch instruments. For example, the airspeed indication is low. The pilot, believing a nose-high attitude exists, applies forward pressure without noting that a low power setting is the cause of the airspeed discrepancy. Increase cross-check speed to include all relevant instrument indications before making a control input.
3. Uncaging the attitude indicator (if caging feature is present) when the airplane is not in level flight. The altimeter and heading indicator must be stabilized with airspeed indication at normal cruise before pulling out the caging knob, to obtain correct indications in straight-and-level flight at normal cruise airspeed.
4. Failure to interpret the attitude indicator in terms of the existing airspeed.
5. Late pitch corrections. Pilots commonly like to leave well enough alone. When the altimeter indicates a 20 foot error, there is a reluctance to correct it, perhaps because of fear of overcontrolling. If overcontrolling is the anticipated error, practice small corrections and find the cause of overcontrolling. If any deviation is tolerated, errors will increase.
6. Chasing the vertical speed indications. This tendency can be corrected by proper cross-check of other pitch instruments, as well as by increasing overall understanding of instrument characteristics.
7. Using excessive pitch corrections for the altimeter evaluation. Rushing a pitch correction by making a large pitch change usually aggravates the existing error, saving neither time nor effort.

8. Failure to maintain established pitch corrections, a common error associated with cross-check and trim errors. For example, having established a pitch change to correct an altitude error, there is a tendency to slow down the cross-check, waiting for the airplane to stabilize in the new pitch attitude. To maintain the attitude, continue to cross-check and trim off the pressures.
9. Fixations during cross-check. After initiating a heading correction, for example, there is a tendency to become preoccupied with bank control and miss errors in pitch attitude. Likewise, during an airspeed change, unnecessary gazing at the power instrument is common. A small error in power setting is of less consequence than large altitude and heading errors. The airplane will not decelerate any faster by staring at the manifold pressure gauge.

Heading

Heading errors usually result from the following faults:

1. Failure to cross-check the heading indicator, especially during changes in power or pitch attitude.
2. Misinterpretation of changes in heading, with resulting corrections in the wrong direction.
3. Failure to note and remember a preselected heading.
4. Failure to observe the rate of heading change and its relation to bank attitude.
5. Overcontrolling in response to heading changes, especially during changes in power settings.
6. Anticipating heading changes with premature application of rudder control.
7. Failure to correct small heading deviations. Unless zero error in heading is the goal, a pilot will tolerate larger and larger deviations. Correction of a 1° error takes a lot less time and concentration than correction of a 20° error.
8. Correcting with improper bank attitude. If correcting a 10° heading error with 20° of bank, the airplane will roll past the desired heading before the bank is established, requiring another correction in the opposite direction. Do not multiply existing errors with errors in corrective technique.
9. Failure to note the cause of a previous heading error and thus repeating the same error. For example, the airplane is out of trim, with a left wing low tendency. Repeated corrections for a slight left turn are made, yet trim is ignored.
10. Failure to set the heading indicator properly or failure to uncage it.

Power

Power errors usually result from the following faults:

1. Failure to know the power settings and pitch attitudes appropriate to various airspeeds and airplane configurations.
2. Abrupt use of throttle.
3. Failure to lead the airspeed when making power changes. For example, during airspeed reduction in level flight, especially with gear and flaps extended, adjust the throttle to maintain the slower speed before the airspeed actually reaches the desired speed. Otherwise, the airplane will decelerate to a speed lower than that desired, resulting in additional power adjustments. The amount of lead depends upon how fast the airplane responds to power changes.
4. Fixation on airspeed or manifold pressure instruments during airspeed changes, resulting in erratic control of both airspeed and power.

Trim

Trim errors usually result from the following faults:

1. Improper adjustment of seat or rudder pedals for comfortable position of legs and feet. Tension in the ankles makes it difficult to relax rudder pressures.
2. Confusion about the operation of trim devices, which differ among various airplane types. Some trim wheels are aligned appropriately with the airplane's axes; others are not. Some rotate in a direction contrary to what is expected.
3. Faulty sequence in trim technique. Trim should be used not as a substitute for control with the wheel (stick) and rudders, but to relieve pressures already held to stabilize attitude. As proficiency is gained, little conscious effort will be required to trim off the pressures as they occur.
4. Excessive trim control. This induces control pressures that must be held until the airplane is trimmed properly. Use trim frequently and in small amounts.
5. Failure to understand the cause of trim changes. Lack of understanding the basic aerodynamics related to basic instrument skills will cause a pilot to continually lag behind the airplane.

Straight Climbs and Descents

Climbs

For a given power setting and load condition, there is only one attitude that will give the most efficient rate of climb. The airspeed and climb power setting that will determine this climb attitude are given in the performance data found in the POH/AFM. Details of the technique for entering a climb vary according to airspeed on entry and the type of climb (constant airspeed or constant rate) desired. (Heading and trim control are maintained as discussed in Straight-and-Level Flight.)

Entry

To enter a constant-airspeed climb from cruising airspeed, raise the miniature aircraft to the approximate nose-high indication for the predetermined climb speed. The attitude will vary according to the type of airplane. Apply light back-elevator pressure to initiate and maintain the climb attitude. The pressures will vary as the airplane decelerates. Power may be advanced to the climb power setting simultaneously with the pitch change, or after the pitch change is established and the airspeed approaches climb speed. If the transition from level flight to climb is smooth, the VSI will show an immediate trend upward, continue to move slowly, and then stop at a rate appropriate to the stabilized airspeed and attitude. (Primary and supporting instruments for the climb entry are shown in *Figure 5-25*.)

Once the airplane stabilizes at a constant airspeed and attitude, the ASI is primary for pitch and the heading indicator remains primary for bank. *[Figure 5-26]* Monitor the tachometer or manifold pressure gauge as the primary power instrument to ensure the proper climb power setting is being maintained. If the climb attitude is correct for the power setting selected, the airspeed will stabilize at the desired speed. If the airspeed is low or high, make an appropriately small pitch correction.

To enter a constant airspeed climb, first complete the airspeed reduction from cruise airspeed to climb speed in straight-and-level flight. The climb entry is then identical to entry from cruising airspeed, except that power must be increased simultaneously to the climb setting as the pitch attitude is increased. Climb entries on partial panel are more easily and accurately controlled if entering the maneuver from climbing speed.

The technique for entering a constant rate climb is very similar to that used for entry to a constant-airspeed climb from climb airspeed. As the power is increased to the approximate setting for the desired rate, simultaneously raise the miniature aircraft to the climbing attitude for the desired airspeed and rate of climb. As the power is increased, the ASI is primary for pitch control until the vertical speed approaches the desired value. As the vertical speed needle stabilizes, it becomes primary for pitch control and the ASI becomes primary for power control. *[Figure 5-27]*

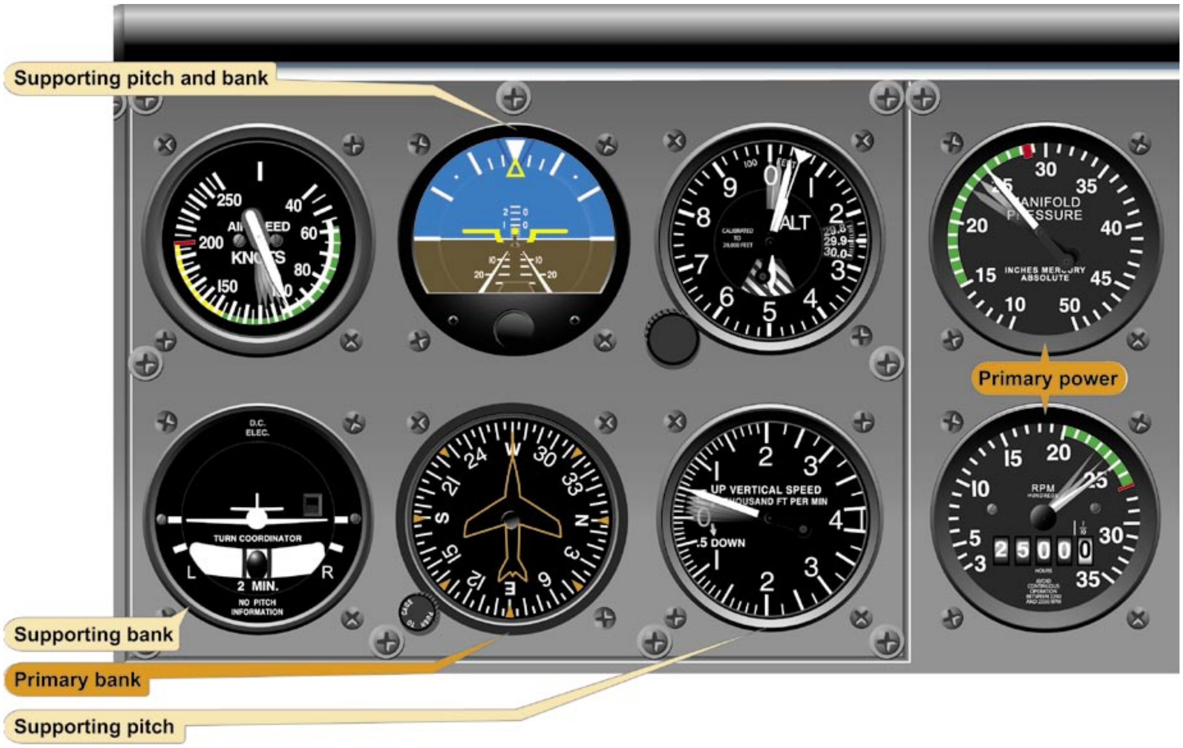

Figure 5-25. *Climb Entry for Constant Airspeed Climb.*

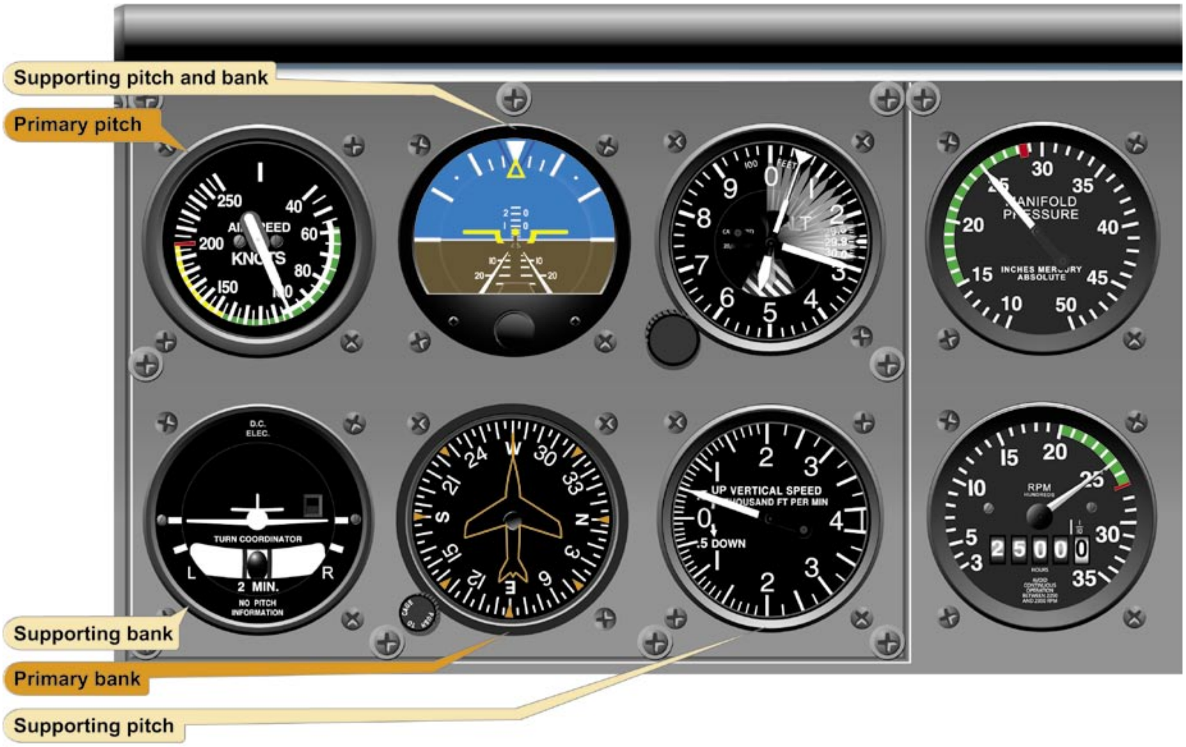
Supporting pitch and bank
Primary pitch
Supporting bank
Primary bank
Supporting pitch

Figure 5-26. *Stabilized Climb at Constant Airspeed.*

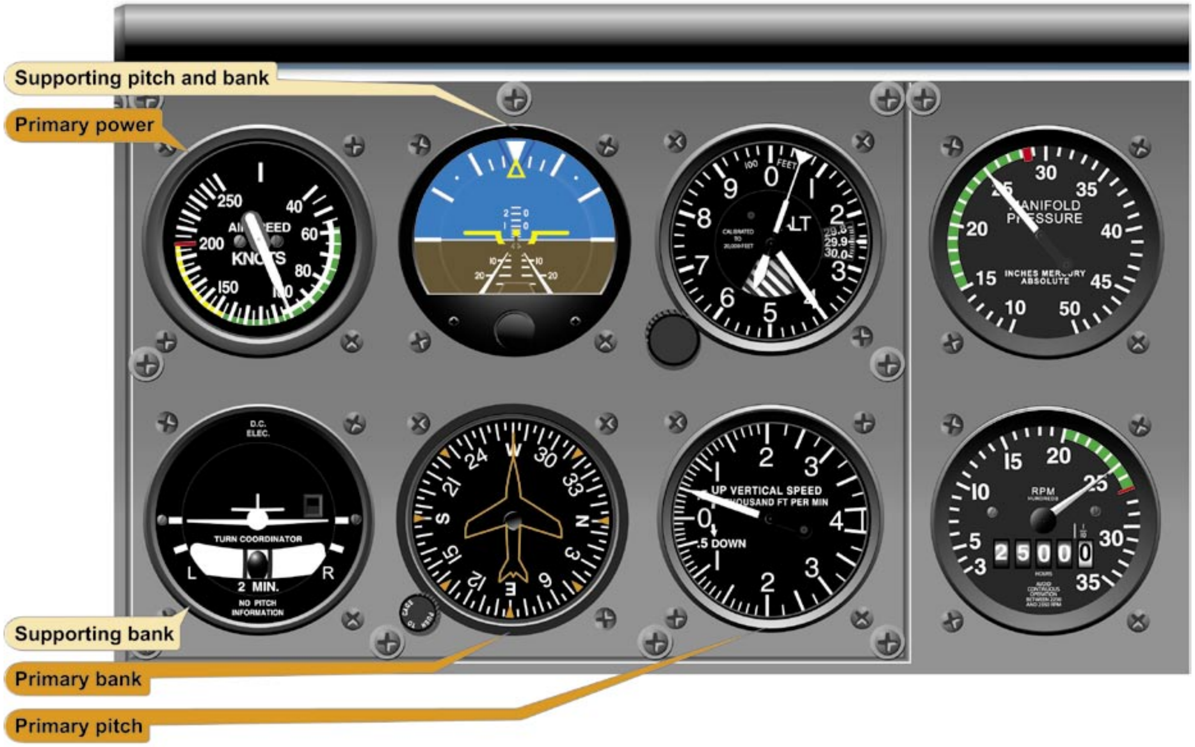
Supporting pitch and bank
Primary power
Supporting bank
Primary bank
Primary pitch

Figure 5-27. *Stabilized Climb at Constant Rate.*

Pitch and power corrections must be promptly and closely coordinated. For example, if the vertical speed is correct, but the airspeed is low, add power. As the power is increased, the miniature aircraft must be lowered slightly to maintain constant vertical speed. If the vertical speed is high and the airspeed is low, lower the miniature aircraft slightly and note the increase in airspeed to determine whether or not a power change is also necessary. *[Figure 5-28]* Familiarity with the approximate power settings helps to keep pitch and power corrections at a minimum.

Leveling Off

To level off from a climb and maintain an altitude, it is necessary to start the level off before reaching the desired altitude. The amount of lead varies with rate of climb and pilot technique. If the airplane is climbing at 1,000 fpm, it will continue to climb at a decreasing rate throughout the transition to level flight. An effective practice is to lead the altitude by 10 percent of the vertical speed shown (500 fpm/ 50-foot lead, 1,000 fpm/100-foot lead).

To level off at cruising airspeed, apply smooth, steady forward-elevator pressure toward level flight attitude for the speed desired. As the attitude indicator shows the pitch change, the vertical speed needle will move slowly toward zero, the altimeter needle will move more slowly, and the airspeed will show acceleration. *[Figure 5-29]* When the altimeter, attitude indicator, and VSI show level flight, constant changes in pitch and torque control will have to be made as the airspeed increases. As the airspeed approaches cruising speed, reduce power to the cruise setting. The amount of lead depends upon the rate of acceleration of the airplane.

To level off at climbing airspeed, lower the nose to the pitch attitude appropriate to that airspeed in level flight. Power is simultaneously reduced to the setting for that airspeed as the pitch attitude is lowered. If power reduction is at a rate proportionate to the pitch change, airspeed will remain constant.

Descents

A descent can be made at a variety of airspeeds and attitudes by reducing power, adding drag, and lowering the nose to a predetermined attitude. The airspeed will eventually stabilize at a constant value. Meanwhile, the only flight instrument providing a positive attitude reference, is the attitude indicator. Without the attitude indicator (such as during a partial panel descent), the ASI, the altimeter, and the VSI will show varying rates of change until the airplane decelerates to a constant airspeed at a constant attitude. During the transition, changes in control pressure and trim, as well as cross-check and interpretation, must be accurate to maintain positive control.

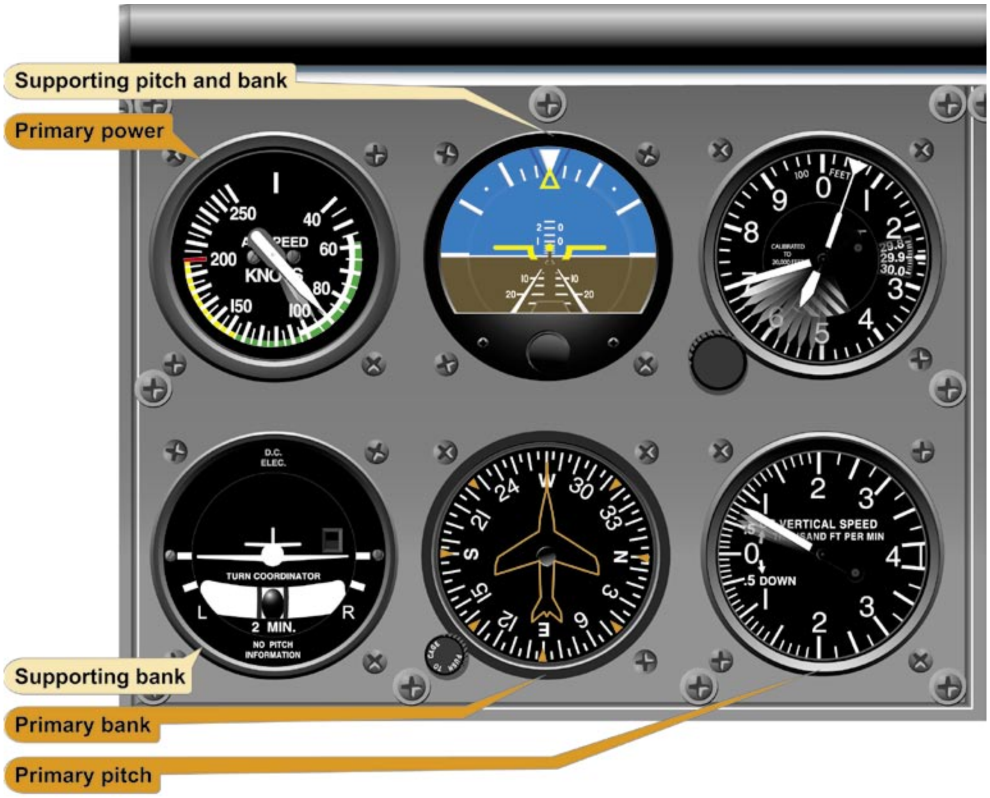

Figure 5-28. *Airspeed Low and Vertical Speed High—Reduce Pitch.*

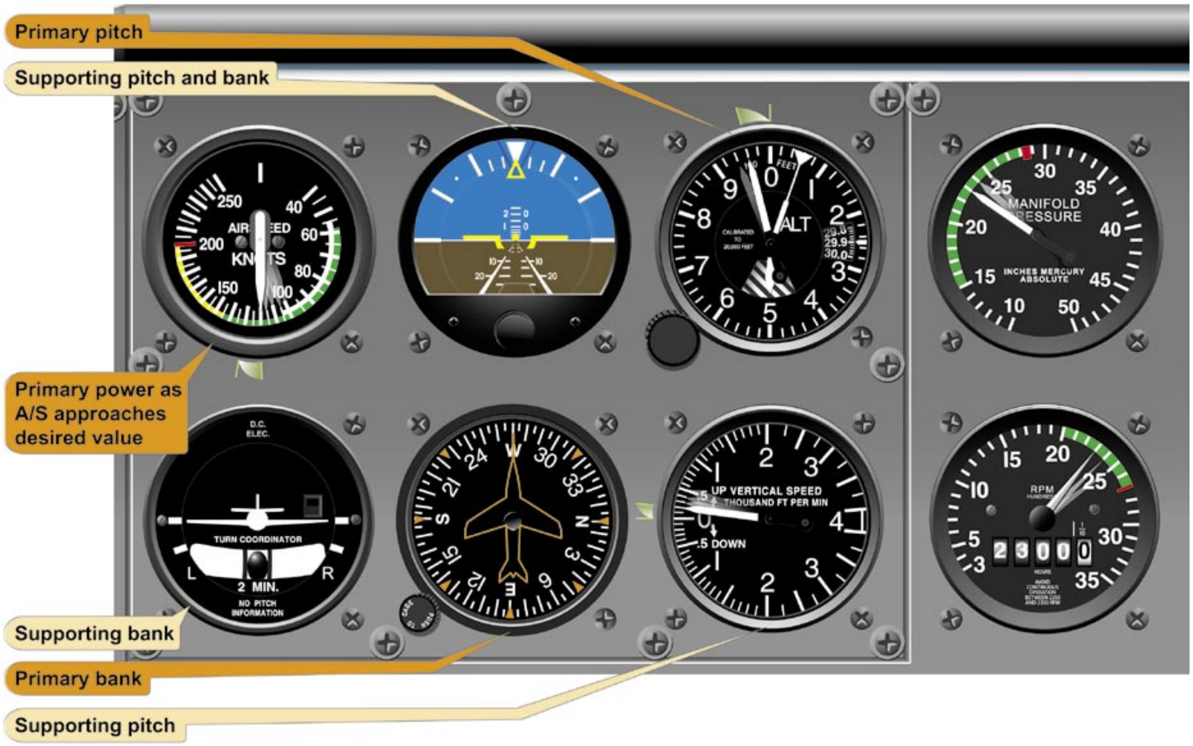
Primary pitch
Supporting pitch and bank
Primary power as A/S approaches desired value
Supporting bank
Primary bank
Supporting pitch
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
TURN COORDINATOR
2 MIN.
NO PITCH INFORMATION
D.C. ELEC.
UP VERTICAL SPEED
THOUSAND FT PER MIN
.5 DOWN
RPM
HUNDREDS

Figure 5-29. *Level Off at Cruising Speed.*

Entry

The following method for entering descents is effective with or without an attitude indicator. First, reduce airspeed to a selected descent airspeed while maintaining straight-and-level flight, then make a further reduction in power (to a predetermined setting). As the power is adjusted, simultaneously lower the nose to maintain constant airspeed, and trim off control pressures.

During a constant airspeed descent, any deviation from the desired airspeed calls for a pitch adjustment. For a constant rate descent, the entry is the same, but the VSI is primary for pitch control (after it stabilizes near the desired rate), and the ASI is primary for power control. Pitch and power must be closely coordinated when corrections are made, as they are in climbs. *[Figure 5-30]*

Leveling Off

The level off from a descent must be started before reaching the desired altitude. The amount of lead depends upon the rate of descent and control technique. With too little lead, the airplane will tend to overshoot the selected altitude unless technique is rapid. Assuming a 500 fpm rate of descent, lead the altitude by 100–150 feet for a level off at an airspeed higher than descending speed. At the lead point, add power to the appropriate level flight cruise setting. *[Figure 5-31]* Since the nose will tend to rise as the airspeed increases, hold forward elevator pressure to maintain the vertical speed at the descending rate until approximately 50 feet above the altitude, and then smoothly adjust the pitch attitude to the level flight attitude for the airspeed selected.

To level off from a descent at descent airspeed, lead the desired altitude by approximately 50 feet, simultaneously adjusting the pitch attitude to level flight and adding power to a setting that will hold the airspeed constant. *[Figure 5-32]* Trim off the control pressures and continue with the normal straight-and-level flight cross-check.

Common Errors in Straight Climbs and Descents

Common errors result from the following faults:

1. Overcontrolling pitch on climb entry. Until the pitch attitudes related to specific power settings used in climbs and descents are known, larger than necessary pitch adjustments are made. One of the most difficult habits to acquire during instrument training is to restrain the impulse to disturb a flight attitude until the result is known. Overcome the inclination to make a large control movement for a pitch change, and learn to apply small control pressures smoothly, cross-checking rapidly for the results of the change, and continuing with the pressures as instruments show the desired results. Small pitch changes can be easily controlled, stopped, and corrected; large changes are more difficult to control.

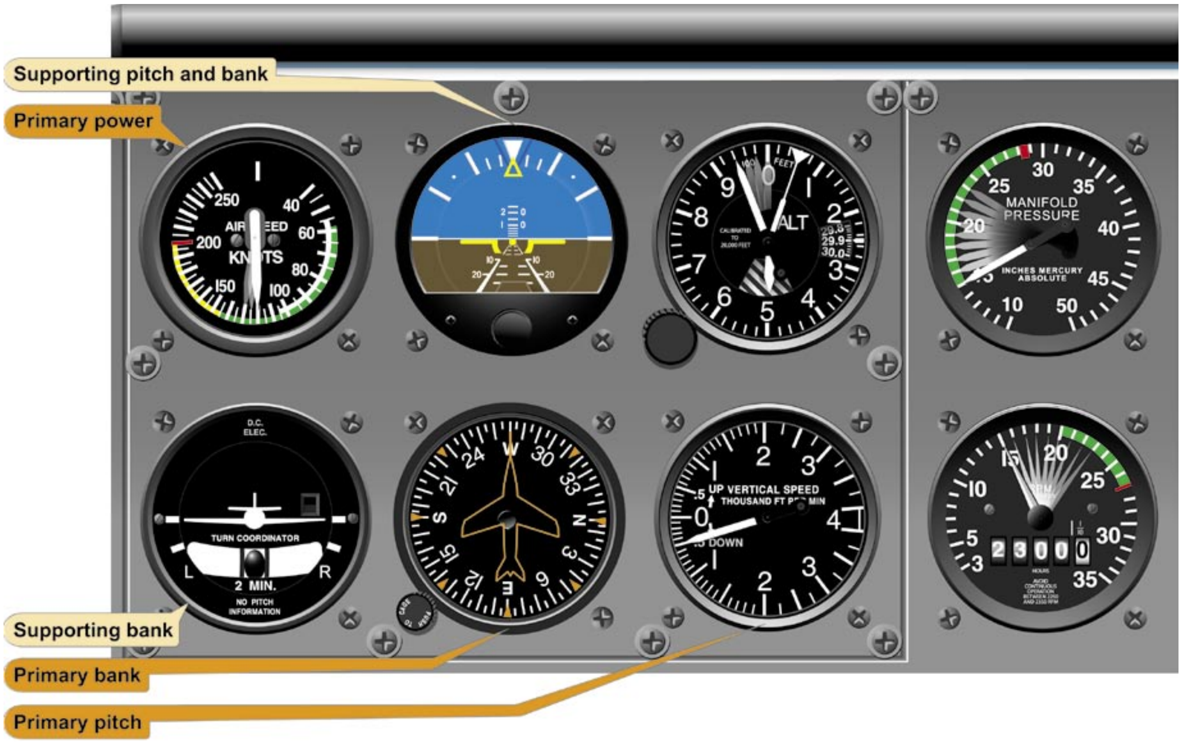
Supporting pitch and bank
Primary power
Supporting bank
Primary bank
Primary pitch

Figure 5-30. *Constant Airspeed Descent, Airspeed High—Reduce Power.*

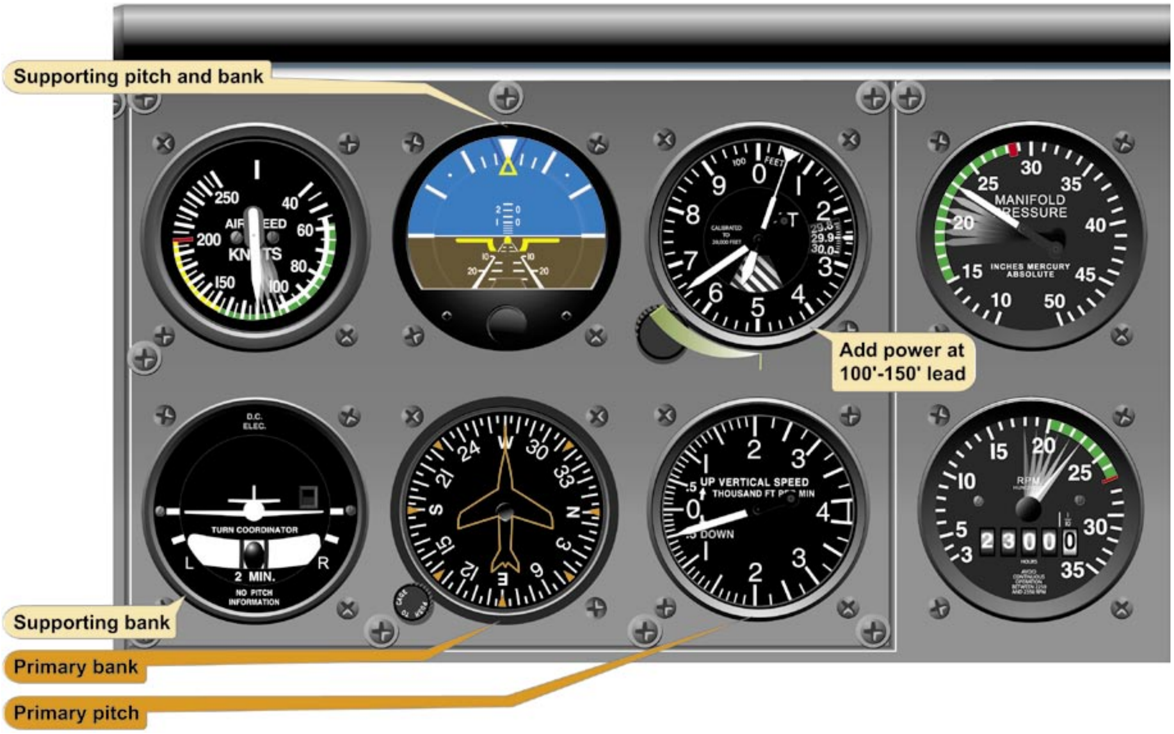
Supporting pitch and bank
Add power at
100'-150' lead
Supporting bank
Primary bank
Primary pitch

Figure 5-31. *Level Off Airspeed Higher Than Descent Airspeed.*

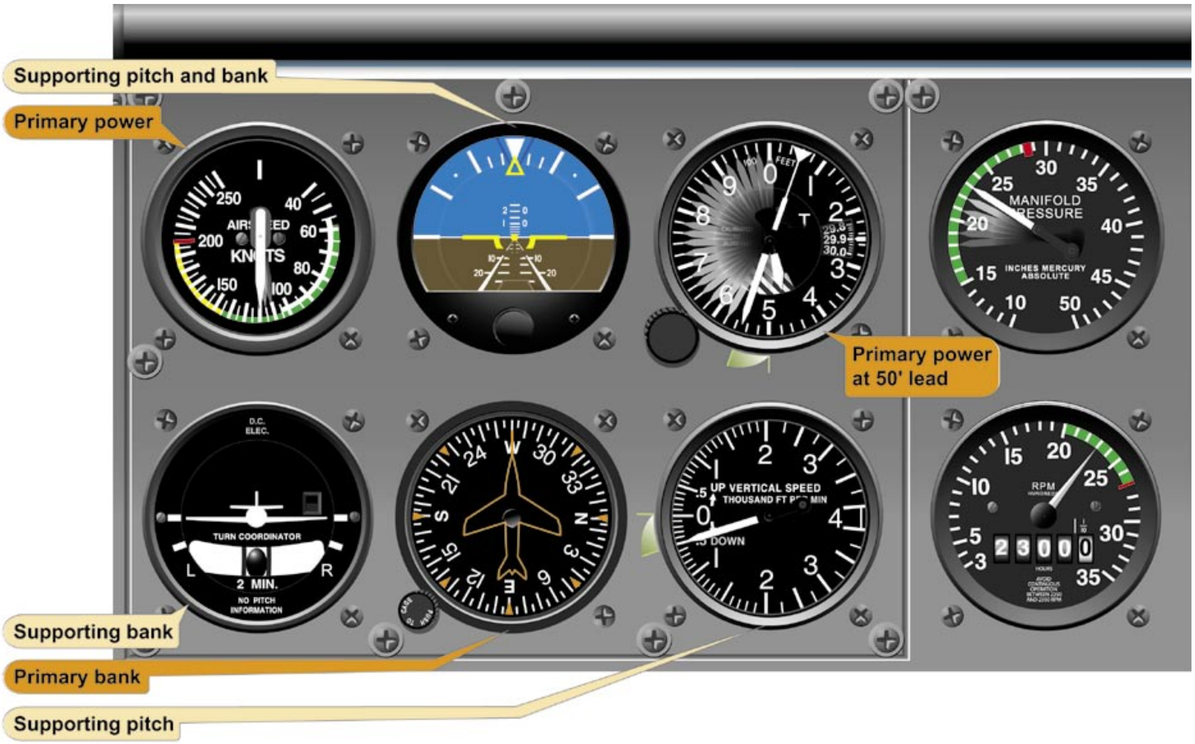

Figure 5-32. *Level Off at Descent Airspeed.*

2. Failure to vary the rate of cross-check during speed, power, or attitude changes or climb or descent entries.
3. Failure to maintain a new pitch attitude. For example, raising the nose to the correct climb attitude, and as the airspeed decreases, either overcontrol and further increase the pitch attitude, or allow the nose to lower. As control pressures change with airspeed changes, cross-check must be increased and pressures readjusted.
4. Failure to trim off pressures. Unless the airplane is trimmed, there will be difficulty in determining whether control pressure changes are induced by aerodynamic changes or by the pilot's own movements.
5. Failure to learn and use proper power settings.
6. Failure to cross-check both airspeed and vertical speed before making pitch or power adjustments.
7. Improper pitch and power coordination on slow-speed level offs, due to slow cross-check of airspeed and altimeter indications.
8. Failure to cross-check the VSI against the other pitch control instruments, resulting in chasing the vertical speed.
9. Failure to note the rate of climb or descent to determine the lead for level offs, resulting in overshooting or undershooting the desired altitude.
10. Ballooning (allowing the nose to pitch up) on level offs from descents, resulting from failure to maintain descending attitude with forward-elevator pressure as power is increased to the level flight cruise setting.
11. Failure to recognize the approaching straight-and-level flight indications as level off is completed. Maintain an accelerated cross-check until positively established in straight-and-level flight.

Turns

Standard Rate Turns

A standard rate turn is one in which the pilot will do a complete 360° circle in two minutes, or 3° per second. A standard rate turn, although always 3° per second, will require higher angles of bank as airspeed increases. To enter a standard rate level turn, apply coordinated aileron and rudder pressures in the desired direction of turn. Pilots commonly roll into turns at a much too rapid rate. During initial training in turns, base control pressures on the rate of cross-check and interpretation. Maneuvering an airplane faster than the capability to keep up with the changes in instrument indications only creates the need to make corrections.

A rule of thumb to determine the approximate angle of bank required for a standard rate turn is to use 15 percent of the true airspeed. A simple way to determine this amount is to divide the airspeed by 10 and add one-half the result. For example, at 100 knots, approximately 15° of bank is required (100 ÷ 10 = 10 + 5 = 15); at 120 knots, approximately 18° of bank is needed for a standard rate turn.

On the roll-in, use the attitude indicator to establish the approximate angle of bank, and then check the turn coordinator's miniature aircraft for a standard rate turn indication or the aircraft's turn-and-bank indicator. Maintain the bank for this rate of turn, using the turn coordinator's miniature aircraft as the primary bank reference and the attitude indicator as the supporting bank instrument. *[Figure 5-33]* Note the exact angle of bank shown on the banking scale of the attitude indicator when the turn coordinator indicates a standard rate turn.

During the roll-in, check the altimeter, VSI, and attitude indicator for the necessary pitch adjustments as the vertical lift component decreases with an increase in bank. If constant airspeed is to be maintained, the ASI becomes primary for power, and the throttle must be adjusted as drag increases. As the bank is established, trim off the pressures applied during pitch and power changes.

To recover to straight-and-level flight, apply coordinated aileron and rudder pressures opposite to the direction of the turn. Strive for the same rate of roll-out used to roll into the turn; fewer problems will be encountered in estimating the lead necessary for roll-out on exact headings, especially on partial panel maneuvers. Upon initiation of the turn recovery, the attitude indicator becomes the primary bank instrument. When the airplane is approximately level, the heading indicator is the primary bank instrument as in straight-and-level flight. Pitch, power, and trim adjustments are made as changes in vertical lift component and airspeed occur. The ball should be checked throughout the turn, especially if control pressures are held rather than trimmed off.

Some airplanes are very stable during turns, requiring only slight trim adjustments that permit hands-off flight while the airplane remains in the established attitude. Other airplanes require constant, rapid cross-check and control during turns to correct overbanking tendencies. Due to the interrelationship of pitch, bank, and airspeed deviations during turns, cross-check must be fast in order to prevent an accumulation of errors.

Turns to Predetermined Headings

As long as an airplane is in a coordinated bank, it continues to turn. Thus, the roll-out to a desired heading must be started before the heading is reached. The amount of lead varies with the relationship between the rate of turn, angle of bank, and rate of recovery. For small heading changes, use a bank angle that does not exceed the number of degrees to be turned. Lead the desired heading by one-half the number of degrees of bank used. For example, if a 10° bank is used during a change in heading, start the roll-out 5° before reaching the desired heading. For larger changes in heading, the amount

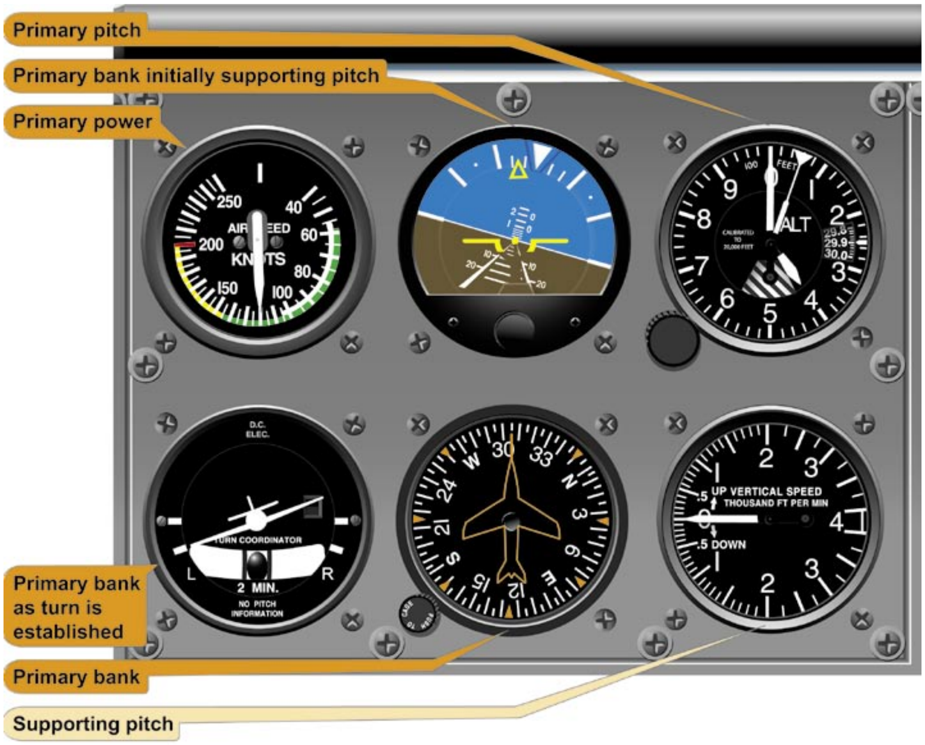

Figure 5-33. *Standard Rate Turn, Constant Airspeed.*

of lead varies since the angle of bank for a standard rate turn varies with the true airspeed.

Practice with a lead of one-half the angle of bank until the precise lead a given technique requires is determined. If rates of roll-in and roll-out are consistent, the precise amount of lead suitable to a particular roll-out technique can be determined.

Timed Turns

A timed turn is a turn in which the clock and the turn coordinator are used to change heading by a specific number of degrees in a given time. For example, in a standard rate turn (3° per second), an airplane turns 45° in 15 seconds; in a half standard rate turn, the airplane turns 45° in 30 seconds.

Prior to performing timed turns, the turn coordinator should be calibrated to determine the accuracy of its indications. *[Figure 5-34]* Establish a standard rate turn as indicated by the turn coordinator, and as the sweep-second hand of the clock passes a cardinal point (12, 3, 6, 9), check the heading on the heading indicator. While holding the indicated rate of turn constant, note the indicated heading changes at 10 second intervals. If the airplane turns more than or less than 30° in that interval, a respectively larger or smaller deflection of the miniature aircraft of the turn coordinator is necessary to produce a standard rate turn. After calibrating the turn coordinator during turns in each direction, note the corrected deflections, if any, and apply them during all timed turns.

The same cross-check and control technique is used in making a timed turn that is used to execute turns to predetermined headings, except the clock is substituted for the heading indicator. The miniature aircraft of the turn coordinator is primary for bank control, the altimeter is primary for pitch control, and the ASI is primary for power control. Start the roll-in when the clock's second hand passes a cardinal point, hold the turn at the calibrated standard rate indication (or half-standard rate for small heading changes), and begin the roll-out when the computed number of seconds has elapsed. If the rates of roll-in and roll-out are the same, the time taken during entry and recovery does not need to be considered in the time computation.

Practice timed turns with a full instrument panel and check the heading indicator for the accuracy of turns. If the turns are executed without the gyro heading indicator, use the magnetic compass at the completion of the turn to check turn accuracy, taking compass deviation errors into consideration.

Compass Turns

In most small airplanes, the magnetic compass is the only direction-indicating instrument independent of other airplane instruments and power sources. Because of its operating characteristics, called compass errors, pilots are prone to use it only as a reference for setting the heading indicator, but knowledge of magnetic compass characteristics permits full use of the instrument to turn the airplane to correct and maintain headings.

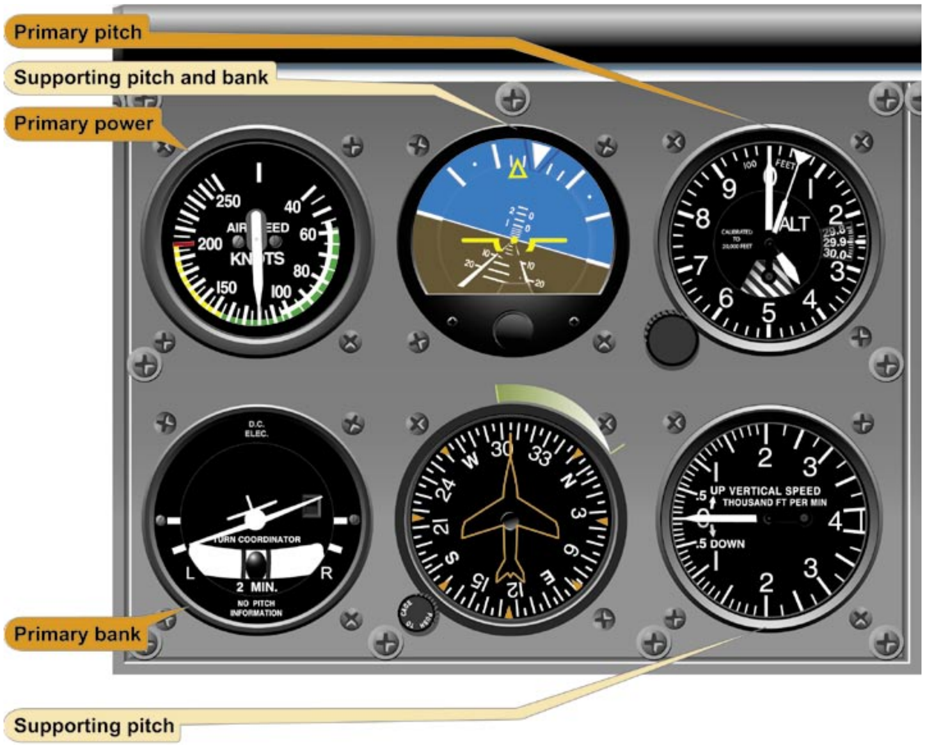

Figure 5-34. *Turn Coordinator Calibration.*

Remember the following points when making turns to magnetic compass headings or when using the magnetic compass as a reference for setting the heading indicator:

1. If on a north heading and a turn is started to the east or west, the compass indication lags, or indicates a turn in the opposite direction.
2. If on a south heading and a turn is started toward the east or west, the compass indication precedes the turn, indicating a greater amount of turn than is actually occurring.
3. When on an east or west heading, the compass indicates correctly when starting a turn in either direction.
4. If on an east or west heading, acceleration results in a north turn indication; deceleration results in a south turn indication.
5. When maintaining a north or south heading, no error results from diving, climbing, or changing airspeed.

With an angle of bank between 15° and 18°, the amount of lead or lag to be used when turning to northerly or southerly headings varies with, and is approximately equal to, the latitude of the locality over which the turn is being made. When turning to a heading of north, the lead for roll-out must include the number of degrees of change of latitude, plus the lead normally used in recovery from turns. During a turn to a south heading, maintain the turn until the compass passes south the number of degrees of latitude, minus normal roll-out lead. *[Figure 5-35]*

For example, when turning from an easterly direction to north, where the latitude is 30°, start the roll-out when the compass reads 37° (30° plus one-half the 15° angle of bank, or whatever amount is appropriate for the rate of roll-out). When turning from an easterly direction to south, start the roll-out when the magnetic compass reads 203° (180° plus 30° minus one-half the angle of bank). When making similar turns from a westerly direction, the appropriate points at which to begin the roll-out would be 323° for a turn to north, and 157° for a turn to south.

When turning to a heading of east or west from a northerly direction, start the roll-out approximately 10° to 12° before the east or west indication is reached. When turning to an east or west heading from a southerly direction, start the rollout approximately 5° before the east or west indication is reached. When turning to other headings, the lead or lag must be interpolated.

Abrupt changes in attitude or airspeed and the resulting erratic movements of the compass card make accurate interpretations of the instrument very difficult. Proficiency in compass turns depends on knowledge of compass characteristics, smooth control technique, and accurate bank-and-pitch control.

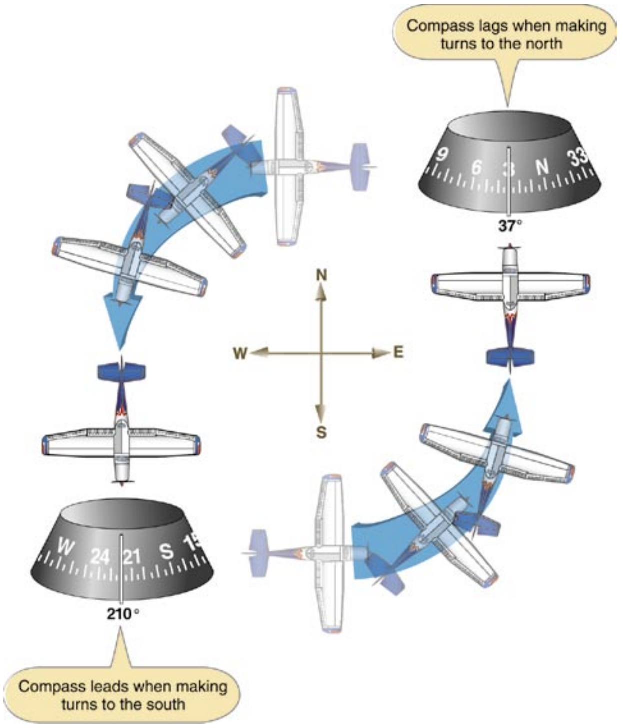
Compass lags when making turns to the north
9 6 3 N 33
37°
N
W
E
S
W 24 21 S 15
210°
Compass leads when making turns to the south

Figure 5-35. *North and South Turn Error.*

Steep Turns

For purposes of instrument flight training in conventional airplanes, any turn greater than a standard rate is considered steep. *[Figure 5-36]* The exact angle of bank at which a normal turn becomes steep is unimportant. What is important is learning to control the airplane with bank attitudes in excess of those normally used on instruments. Practicing steep turns will not only increase proficiency in the basic instrument flying skills, but also enable smooth, quick, and confident reactions to unexpected abnormal flight attitudes under instrument flight conditions.

Pronounced changes occur in the effects of aerodynamic forces on aircraft control at progressively greater bank attitudes. Skill in cross-check, interpretation, and control is increasingly necessary in proportion to the amount of these changes, though the techniques for entering, maintaining, and recovering from the turn are the same in principle for steep turns as for shallower turns.

Enter a steep turn in the same way as a shallower turn, but prepare to cross-check rapidly as the turn steepens. Because of the greatly reduced vertical lift component, pitch control is usually the most difficult aspect of this maneuver. Unless immediately noted and corrected with a pitch increase, the loss of vertical lift results in rapid movement of the altimeter, vertical speed, and airspeed needles. The faster the rate of bank change, the more suddenly the lift changes occur. If a cross-check is fast enough to note the immediate need for pitch changes, smooth, steady back elevator pressure will maintain constant altitude. However, overbanking to excessively steep angles without adjusting pitch as the bank changes occur, requires increasingly stronger elevator pressure. The loss of vertical lift and increase in wing loading finally reach a point at which further application of back-elevator pressure tightens the turn without raising the nose.

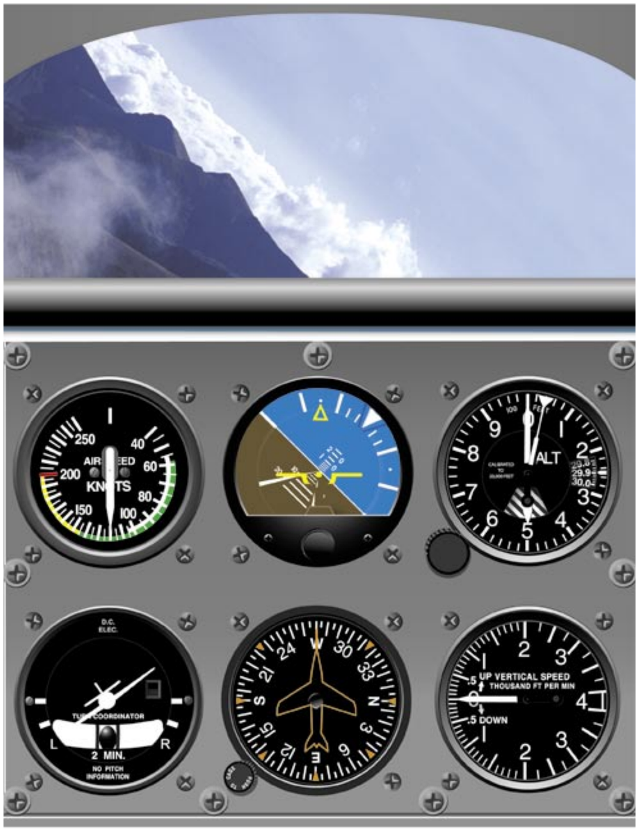
AIRSPEED
KNOTS
ALT
TURN COORDINATOR
D.C. ELEC.
2 MIN.
NO PITCH INFORMATION
L
R
UP VERTICAL SPEED THOUSAND FT PER MIN
DOWN

Figure 5-36. *Steep Left Turn.*

How does a pilot recognize overbanking and low pitch attitude? What should a pilot do to correct them? If a rapid downward movement of the altimeter needle or vertical speed needle, together with an increase in airspeed, is observed despite application of back elevator pressure, the airplane is in a diving spiral. *[Figure 5-37]* Immediately shallow the bank with smooth and coordinated aileron and rudder pressures, hold or slightly relax elevator pressure, and increase the cross-check of the attitude indicator, altimeter, and VSI. Reduce power if the airspeed increase is rapid. When the vertical speed trends upward, the altimeter needle will move slower as the vertical lift increases. When the elevator is effective in raising the nose, hold the bank attitude shown on the attitude indicator and adjust elevator control pressures smoothly for the nose-high attitude appropriate to the bank maintained. If pitch control is consistently late on entries to steep turns, rollout immediately to straight-and-level flight and analyze possible errors. Practice shallower turns initially and learn the attitude changes and control responses required, then increase the banks as a quicker and more accurate cross-check and control techniques are developed.

The power necessary to maintain constant airspeed increases as the bank and drag increase. With practice, the power

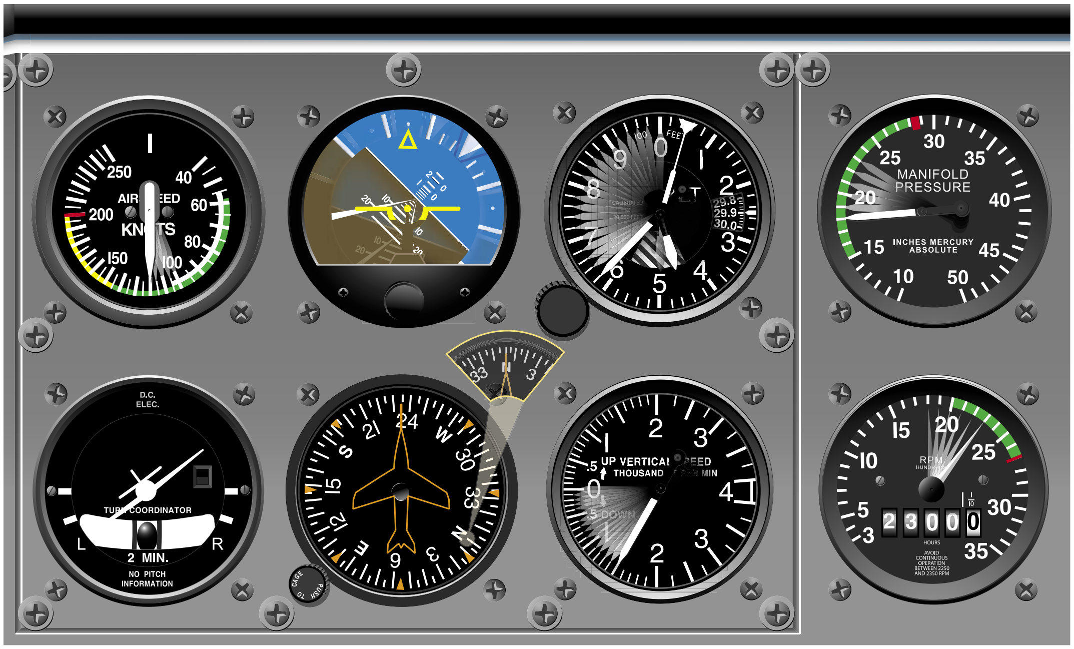
AIRSPEED
KNOTS
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
TURN COORDINATOR
D.C. ELEC.
2 MIN.
NO PITCH INFORMATION
L
R
UP VERTICAL SPEED THOUSAND FT PER MIN
DOWN
RPM
2300

Figure 5-37. *Diving Spiral.*

settings appropriate to specific bank attitudes are learned, and adjustments can be made without undue attention to airspeed and power instruments. During training in steep turns, as in any other maneuver, attend to the most important tasks first. Keep the pitch attitude relatively constant, and more time can be devoted to cross-check and instrument interpretation.

During recovery from steep turns to straight-and-level flight, elevator and power control must be coordinated with bank control in proportion to the changes in aerodynamic forces. Back elevator pressures must be released and power decreased. The common errors associated with steep turns are the same as those discussed later in this section. Remember, errors are more exaggerated, more difficult to correct, and more difficult to analyze unless rates of entry and recovery are consistent with the level of proficiency in the three basic instrument flying skills.

Climbing and Descending Turns

To execute climbing and descending turns, combine the technique used in straight climbs and descents with the various turn techniques. The aerodynamic factors affecting lift and power control must be considered in determining power settings, and the rate of cross-check and interpretation must be increased to enable control of bank as well as pitch changes.

Change of Airspeed During Turns

Changing airspeed during turns is an effective maneuver for increasing proficiency in all three basic instrument skills. Since the maneuver involves simultaneous changes in all components of control, proper execution requires rapid cross-check and interpretation as well as smooth control. Proficiency in the maneuver will also contribute to confidence in the instruments during attitude and power changes involved in more complex maneuvers. Pitch and power control techniques are the same as those used during changes in airspeed in straight-and-level flight.

The angle of bank necessary for a given rate of turn is proportional to the true airspeed. Since the turns are executed at a standard rate, the angle of bank must be varied in direct proportion to the airspeed change in order to maintain a constant rate of turn. During a reduction of airspeed, decrease the angle of bank and increase the pitch attitude to maintain altitude and a standard rate turn.

The altimeter and turn coordinator indications should remain constant throughout the turn. The altimeter is primary for pitch control and the miniature aircraft of the turn coordinator is primary for bank control. The manifold pressure gauge (or tachometer) is primary for power control while the airspeed is changing. As the airspeed approaches the new indication, the ASI becomes primary for power control.

Two methods of changing airspeed in turns may be used. In the first method, airspeed is changed after the turn is established. *[Figure 5-38]* In the second method, the airspeed change is initiated simultaneously with the turn entry. The first method is easier, but regardless of the method used, the rate of cross-check must be increased as power is reduced. As the airplane decelerates, check the altimeter and VSI for necessary pitch changes and the bank instruments for required bank changes. If the miniature aircraft of the turn coordinator indicates a deviation from the desired deflection, adjust the bank. Adjust

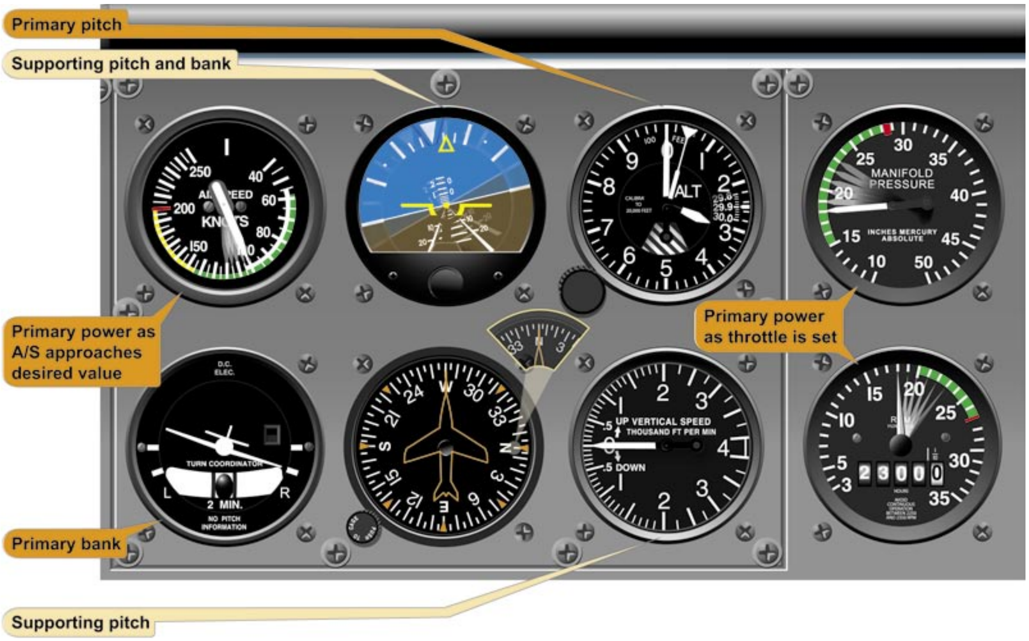

Figure 5-38. *Change of Airspeed During Turn.*

pitch attitude to maintain altitude. When approaching the desired airspeed, pitch attitude becomes primary for power control and the manifold pressure gauge (or tachometer) is adjusted to maintain the desired airspeed. Trim is important throughout the maneuver to relieve control pressures.

Until control technique is very smooth, frequent cross-check of the attitude indicator is essential to prevent overcontrolling and to provide approximate bank angles appropriate to the changing airspeeds.

Common Errors in Turns

Pitch

Pitch errors result from the following faults:

1. Preoccupation with bank control during turn entry and recovery. If 5 seconds are required to roll into a turn, check the pitch instruments as bank pressures are initiated. If bank control pressure and rate of bank change are consistent, a sense of the time required for an attitude change will be developed. During the interval, check pitch, power, and trim—as well as bank—controlling the total attitude instead of one factor at a time.
2. Failure to understand or remember the need for changing the pitch attitude as the vertical lift component changes, resulting in consistent loss of altitude during entries.
3. Changing the pitch attitude before it is necessary. This fault is very likely if a cross-check is slow and rate of entry too rapid. The error occurs during the turn entry due to a mechanical and premature application of back-elevator control pressure.
4. Overcontrolling the pitch changes. This fault commonly occurs with the previous error.
5. Failure to properly adjust the pitch attitude as the vertical lift component increases during the roll-out, resulting in consistent gain in altitude on recovery to headings.
6. Failure to trim during turn entry and following turn recovery (if turn is prolonged).
7. Failure to maintain straight-and-level cross-check after roll-out. This error commonly follows a perfectly executed turn.
8. Erratic rates of bank change on entry and recovery, resulting from failure to cross-check the pitch instruments with a consistent technique appropriate to the changes in lift.

Bank

Bank and heading errors result from the following faults:

1. Overcontrolling, resulting in overbanking upon turn entry, overshooting and undershooting headings, as well as aggravated pitch, airspeed, and trim errors.
2. Fixation on a single bank instrument. On a 90° change of heading, for example, leave the heading indicator out of the cross-check for approximately 20 seconds after establishing a standard rate turn, since at 3° per second the turn will not approach the lead point until that time has elapsed. Make the cross-check selective, checking only what needs to be checked at the appropriate time.
3. Failure to check for precession of the horizon bar following recovery from a turn. If the heading indicator shows a change in heading when the attitude indicator shows level flight, the airplane is turning. If the ball is centered, the attitude gyro has processed; if the ball is not centered, the airplane may be in a slipping or skidding turn. Center the ball with rudder pressure, check the attitude indicator and heading indicator, stop the heading change if it continues, and retrim.
4. Failure to use the proper degree of bank for the amount of heading change desired. Rolling into a 20° bank for a heading change of 10° will normally overshoot the heading. Use the bank attitude appropriate to the amount of heading change desired.
5. Failure to remember the heading to which the aircraft is being turned. This fault is likely when rushing the maneuver.
6. Turning in the wrong direction, due to misreading or misinterpreting the heading indicator, or to confusion regarding the location of points on the compass. Turn in the shortest direction to reach a given heading, unless there is a specific reason to turn the long way around. Study the compass rose and visualize at least the positions of the eight major points around the azimuth. A number of methods can be used to make quick computations for heading changes. For example, to turn from a heading of 305° to a heading of 110°, would a pilot turn right or left for the shortest way around? Subtracting 200 from 305 and adding 20, gives 125° as the reciprocal of 305°; therefore, execute the turn to the right. Likewise, to figure the reciprocal of a heading less than 180°, add 200 and subtract 20. Computations are done more quickly using multiples of 100s and 10s than by adding or subtracting 180° from the actual heading; therefore, the method suggested above may save time and confusion.

7. Failure to check the ball of the turn coordinator when interpreting the instrument for bank information. If the roll rate is reduced to zero, the miniature aircraft of the turn coordinator indicates only direction and rate of turn. Unless the ball is centered, do not assume the turn is resulting from a banked attitude.

Power

Power and airspeed errors result from the following faults:

1. Failure to cross-check the ASI as pitch changes are made.
2. Erratic use of power control. This may be due to improper throttle friction control, inaccurate throttle settings, chasing the airspeed readings, abrupt or overcontrolled pitch-and-bank changes, or failure to recheck the airspeed to note the effect of a power adjustment.
3. Poor coordination of throttle control with pitch-and-bank changes, associated with slow cross-check or failure to understand the aerodynamic factors related to turns.

Trim

Trim errors result from the following faults:

1. Failure to recognize the need for a trim change due to slow cross-check and interpretation. For example, a turn entry at a rate too rapid for a cross-check leads to confusion in cross-check and interpretation, with resulting tension on the controls.
2. Failure to understand the relationship between trim and attitude/power changes.
3. Chasing the vertical speed needle. Overcontrolling leads to tension and prevents sensing the pressures to be trimmed off.
4. Failure to trim following power changes.

Errors During Compass Turns

In addition to the faults discussed above, the following errors connected with compass turns should be noted:

1. Faulty understanding or computation of lead and lag.
2. Fixation on the compass during the roll-out. Until the airplane is in straight-and-level unaccelerated flight, it is unnecessary to read the indicated heading. Accordingly, after the roll-out, cross-check for straight-and-level flight before checking the accuracy of the turn.

Approach to Stall

Practicing approach to stall recoveries in various airplane configurations should build confidence in a pilot's ability to control the airplane in unexpected situations. Approach to stall should be practiced from straight flight and from shallow banks. The objective is to practice recognition and recovery from the approach to a stall.

Prior to stall recovery practice, select a safe altitude above the terrain, an area free of conflicting air traffic, appropriate weather, and the availability of radar traffic advisory service.

Approaches to stalls are accomplished in the following configurations:

1. Takeoff configuration—should begin from level flight near liftoff speed. Power should be applied while simultaneously increasing the angle of attack to induce an indication of a stall.
2. Clean configuration—should begin from a reduced airspeed, such as pattern airspeed, in level flight. Power should be applied while simultaneously increasing the angle of attack to induce an indication of a stall.
3. Approach or landing configuration—should be initiated at the appropriate approach or landing airspeed. The angle of attack should be smoothly increased to induce an indication of a stall.

Recoveries should be prompt in response to a stall warning device or an aerodynamic indication by smoothly reducing the angle of attack and applying maximum power, or as recommended by the POH/AFM. The recovery should be completed without an excessive loss of altitude, and on a predetermined heading, altitude, and airspeed.

Unusual Attitudes and Recoveries

An unusual attitude is an airplane attitude not normally required for instrument flight. Unusual attitudes may result from a number of conditions, such as turbulence, disorientation, instrument failure, confusion, preoccupation with flight deck duties, carelessness in cross-checking, errors in instrument interpretation, or lack of proficiency in aircraft control. Since unusual attitudes are not intentional maneuvers during instrument flight, except in training, they are often unexpected, and the reaction of an inexperienced or inadequately trained pilot to an unexpected abnormal flight attitude is usually instinctive rather than intelligent and deliberate. This individual reacts with abrupt muscular effort, which is purposeless and even hazardous in turbulent conditions, at excessive speeds, or at low altitudes. However, with practice, the techniques for rapid and safe recovery from unusual attitudes can be mastered.

When an unusual attitude is noted during the cross-check, the immediate problem is not how the airplane got there, but what it is doing and how to get it back to straight-and-level flight as quickly as possible.

Recognizing Unusual Attitudes

As a general rule, any time an instrument rate of movement or indication other than those associated with the basic instrument flight maneuvers is noted, assume an unusual attitude and increase the speed of cross-check to confirm the attitude, instrument error, or instrument malfunction.

Nose-high attitudes are shown by the rate and direction of movement of the altimeter needle, vertical speed needle, and airspeed needle, as well as the immediately recognizable indication of the attitude indicator (except in extreme attitudes). *[Figure 5-39]* Nose-low attitudes are shown by the same instruments, but in the opposite direction. *[Figure 5-40]*

Recovery from Unusual Attitudes

In moderate unusual attitudes, the pilot can normally reorient by establishing a level flight indication on the attitude indicator. However, the pilot should not depend on this instrument if the attitude indicator is the spillable type, because its upset limits may have been exceeded or it may have become inoperative due to mechanical malfunction. If it is the nonspillable-type instrument and is operating properly, errors up to 5° of pitch-and-bank may result and its indications are very difficult to interpret in extreme attitudes. As soon as the unusual attitude is detected, the recommended recovery procedures stated in the POH/AFM should be initiated. If there are no recommended procedures stated in the POH/AFM, the recovery should be initiated by reference to the ASI, altimeter, VSI, and turn coordinator.

Nose-High Attitudes

If the airspeed is decreasing, or below the desired airspeed, increase power (as necessary in proportion to the observed deceleration), apply forward elevator pressure to lower the nose and prevent a stall, and correct the bank by applying coordinated aileron and rudder pressure to level the miniature aircraft and center the ball of the turn coordinator. The corrective control applications are made almost simultaneously, but in the sequence given above. A level pitch attitude is indicated by the reversal and stabilization

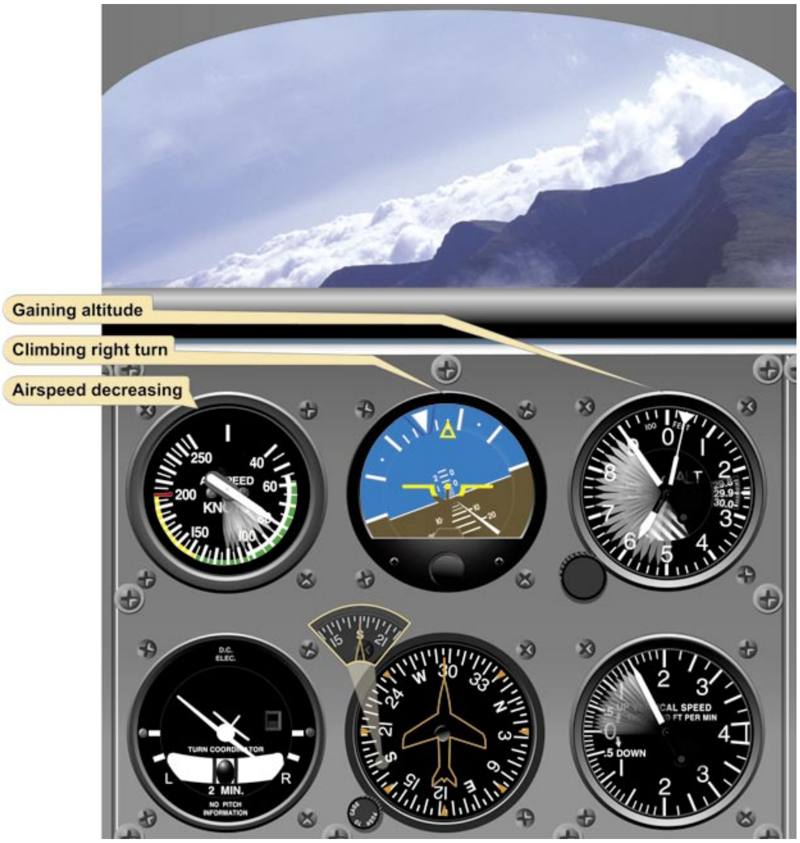

Figure 5-39. *Unusual Attitude—Nose-High.*

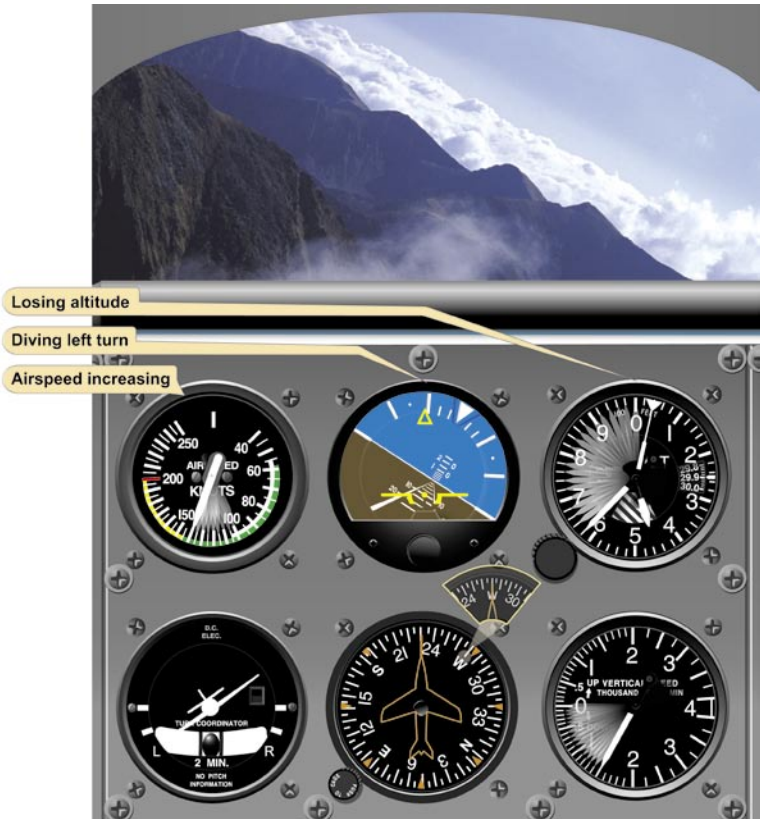

Figure 5-40. *Unusual Attitude—Nose-Low.*

of the ASI and altimeter needles. Straight coordinated flight is indicated by the level miniature aircraft and centered ball of the turn coordinator.

Nose-Low Attitudes

If the airspeed is increasing, or is above the desired airspeed, reduce power to prevent excessive airspeed and loss of altitude. Correct the bank attitude with coordinated aileron and rudder pressure to straight flight by referring to the turn coordinator. Raise the nose to level flight attitude by applying smooth back elevator pressure. All components of control should be changed simultaneously for a smooth, proficient recovery. However, during initial training a positive, confident recovery should be made by the numbers, in the sequence given above. A very important point to remember is that the instinctive reaction to a nose-down attitude is to pull back on the elevator control.

After initial control has been applied, continue with a fast cross-check for possible overcontrolling, since the necessary initial control pressures may be large. As the rate of movement of altimeter and ASI needles decreases, the attitude is approaching level flight. When the needles stop and reverse direction, the aircraft is passing through level flight. As the indications of the ASI, altimeter, and turn coordinator stabilize, incorporate the attitude indicator into the cross-check.

The attitude indicator and turn coordinator should be checked to determine bank attitude and then corrective aileron and rudder pressures should be applied. The ball should be centered. If it is not, skidding and slipping sensations can easily aggravate disorientation and retard recovery. If entering the unusual attitude from an assigned altitude (either by an instructor or by air traffic control (ATC) if operating under instrument flight rules (IFR)), return to the original altitude after stabilizing in straight-and-level flight.

Common Errors in Unusual Attitudes

Common errors associated with unusual attitudes include the following faults:

1. Failure to keep the airplane properly trimmed. A flight deck interruption when holding pressures can easily lead to inadvertent entry into unusual attitudes.
2. Disorganized flight deck. Hunting for charts, logs, computers, etc., can seriously distract attention from the instruments.
3. Slow cross-check and fixations. The impulse is to stop and stare when noting an instrument discrepancy unless a pilot has trained enough to develop the skill required for immediate recognition.

4. Attempting to recover by sensory sensations other than sight. The discussion of disorientation in Chapter 1, Human Factors, indicates the importance of trusting the instruments.
5. Failure to practice basic instrument skills. All of the errors noted in connection with basic instrument skills are aggravated during unusual attitude recoveries until the elementary skills have been mastered.

Instrument Takeoff

Competency in instrument takeoffs will provide the proficiency and confidence necessary for use of flight instruments during departures under conditions of low visibility, rain, low ceilings, or disorientation at night. A sudden rapid transition from "visual" to "instrument" flight can result in serious disorientation and control problems.

Instrument takeoff techniques vary with different types of airplanes, but the method described below is applicable whether the airplane is single- or multiengine; tricycle gear or conventional gear.

Align the airplane with the centerline of the runway with the nosewheel or tailwheel straight. Lock the tailwheel, if so equipped, and hold the brakes firmly to avoid creeping while preparing for takeoff. Set the heading indicator with the nose index on the 5° mark nearest the published runway heading to allow instant detection of slight changes in heading during the takeoff. Make certain that the instrument is uncaged (if it has a caging feature) by rotating the knob after uncaging and checking for constant heading indication. If using an electric heading indicator with a rotatable needle, rotate the needle so that it points to the nose position, under the top index. Advance the throttle to an rpm that will provide partial rudder control. Release the brakes, advancing the power smoothly to takeoff setting.

During the takeoff roll, hold the heading constant on the heading indicator by using the rudder. In multiengine, propeller-driven airplanes, also use differential throttle to maintain direction. The use of brakes should be avoided, except as a last resort, as it usually results in overcontrolling and extending the takeoff roll. Once the brakes are released, any deviation in heading must be corrected instantly.

As the airplane accelerates, cross-check both heading indicator and ASI rapidly. The attitude indicator may precess to a slight nose-up attitude. As flying speed is approached (approximately 15–25 knots below takeoff speed), smoothly apply elevator control for the desired takeoff attitude on the attitude indicator. This is approximately a two bar width climb indication for most small airplanes.

Continue with a rapid cross-check of heading indicator and attitude indicator as the airplane leaves the ground. Do not pull it off; let it fly off while holding the selected attitude constant. Maintain pitch-and-bank control by referencing the attitude indicator, and make coordinated corrections in heading when indicated on the heading indicator. Cross-check the altimeter and VSI for a positive rate of climb (steady clockwise rotation of the altimeter needle, and the VSI showing a stable rate of climb appropriate to the airplane).

When the altimeter shows a safe altitude (approximately 100 feet), raise the landing gear and flaps, maintaining attitude by referencing the attitude indicator. Because of control pressure changes during gear and flap operation, overcontrolling is likely unless the pilot notes pitch indications accurately and quickly. Trim off control pressures necessary to hold the stable climb attitude. Check the altimeter, VSI, and airspeed for a smooth acceleration to the predetermined climb speed (altimeter and airspeed increasing, vertical speed stable). At climb speed, reduce power to climb setting (unless full power is recommended for climb by the POH/AFM and trim).

Throughout the instrument takeoff, cross-check and interpretation must be rapid, and control positive and smooth. During liftoff, gear and flap retraction, power reduction, and the changing control reactions demand rapid cross-check, adjustment of control pressures, and accurate trim changes.

Common Errors in Instrument Takeoffs

Common errors during the instrument takeoff include the following:

1. Failure to perform an adequate flight deck check before the takeoff. Pilots have attempted instrument takeoffs with inoperative airspeed indicators (pitot tube obstructed), gyros caged, controls locked, and numerous other oversights due to haste or carelessness.
2. Improper alignment on the runway. This may result from improper brake application, allowing the airplane to creep after alignment, or from alignment with the nosewheel or tailwheel cocked. In any case, the result is a built-in directional control problem as the takeoff starts.
3. Improper application of power. Abrupt application of power complicates directional control. Add power with a smooth, uninterrupted motion.
4. Improper use of brakes. Incorrect seat or rudder pedal adjustment, with feet in an uncomfortable position, frequently cause inadvertent application of brakes and excessive heading changes.

5. Overcontrolling rudder pedals. This fault may be caused by late recognition of heading changes, tension on the controls, misinterpretation of the heading indicator (and correcting in the wrong direction), failure to appreciate changing effectiveness of rudder control as the aircraft accelerates, and other factors. If heading changes are observed and corrected instantly with small movement of the rudder pedals, swerving tendencies can be reduced.
6. Failure to maintain attitude after becoming airborne. If the pilot reacts to seat-of-the-pants sensations when the airplane lifts off, pitch control is guesswork. The pilot may either allow excessive pitch or apply excessive forward elevator pressure, depending on the reaction to trim changes.
7. Inadequate cross-check. Fixations are likely during trim changes, attitude changes, gear and flap retractions, and power changes. Once an instrument or a control input is applied, continue the cross-check and note the effect during the next cross-check sequence.
8. Inadequate interpretation of instruments. Failure to understand instrument indications immediately indicates that further study of the maneuver is necessary.

Basic Instrument Flight Patterns

Flight patterns are basic maneuvers, flown by sole reference to the instruments rather than outside visual clues, for the purpose of practicing basic attitude flying. The patterns simulate maneuvers encountered on instrument flights such as holding patterns, procedure turns, and approaches. After attaining a reasonable degree of proficiency in basic maneuvers, apply these skills to the various combinations of individual maneuvers. The following practice flight patterns are directly applicable to operational instrument flying.

Racetrack Pattern

1. Time 3 minutes straight-and-level flight from A to B. *[Figure 5-41]* During this interval, reduce airspeed to the holding speed appropriate for the aircraft.
2. Start a 180° standard rate turn to the right at B. Roll-out at C on the reciprocal of the heading originally used at A.
3. Time a 1 minute straight-and-level flight from C to D.
4. Start a 180° standard rate turn to the right at D, rolling-out on the original heading.
5. Fly 1 minute on the original heading, adjusting the outbound leg so that the inbound segment is 1 minute.

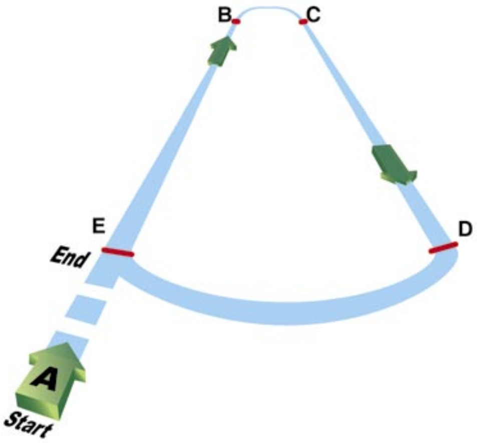

Figure 5-41. *Racetrack Pattern (Entire Pattern in Level Flight).*

NOTE*:* This pattern is an exercise combining use of the clock with basic maneuvers.

Procedure Turn

A procedure turn is a maneuver that facilitates:

- A reversal in flight direction.
- A descent from an initial approach fix or assigned altitude to a permissible altitude (usually the procedure turn altitude).
- An interception of the inbound course at a sufficient distance allowing the aircraft to become aligned with the final approach.

Procedure turn types include the 45° turn, the 80/260 turn, and the teardrop turn. All of these turns are normally conducted no more than 10 nautical miles (NM) from the primary airport. The procedure turn altitude generally provides a minimum of 1,000' obstacle clearance in the procedure turn area (not necessarily within the 10 NM arc around the primary airport). Turns may have to be increased or decreased but should not exceed 30° of a bank angle.

Standard 45° Procedure Turn

1. Start timing at point A (usually identified on approach procedures by a fix). For example, fly outbound on a heading of 360° for a given time (2 minutes, in this example). *[Figure 5-42]*
2. After flying outbound for 2 minutes (point B), turn left 45° to a heading of 315° using a standard rate turn. After roll-out and stabilizing, fly this new heading of 315° for 40 seconds and the aircraft will be at the approximate position of C.

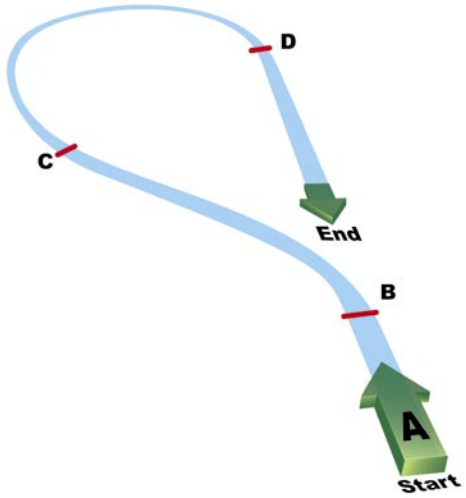

Figure 5-42. *Standard Procedure Turn (Entire Pattern in Level Flight).*

3. At point C, turn 225° right (using a standard rate turn) which will provide a heading of 180°. The timing is such that in a no wind environment, the pilot will be aligned with the final approach course of 180° at D. Wind conditions, however must be considered during the execution of the procedure turn. Compensating for wind may result in changes to outbound time, procedure turn heading and/or time and minor changes in the inbound turn.

80/260 Procedure Turn

1. Start timing at point A (usually identified on approach procedures by a fix). For example, fly outbound on a heading of 360° for 2 minutes. *[Figure 5-43]*

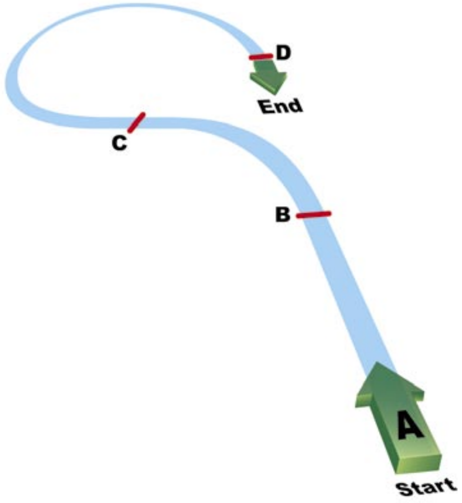

Figure 5-43. *80/260 Procedure Turn (Entire Pattern in Level Flight).*

2. At B, enter a left standard rate turn of 80° to a heading of 280°.
3. At the completion of the 80° turn to 280° (Point C), immediately turn right 260°, rolling-out on a heading of 180° (Point D) and also the reciprocal of the entry heading.

Teardrop Patterns

There are three typical teardrop procedure turns. A 30°, 20°, and a 10° teardrop pattern. The below steps indicate actions for all three starting on a heading of 360°. *[Figure 5-44]*

1. At point B (after stabilizing on the outbound course) turn left:
   - 30° to a heading of 330° and time for 1 minute
   - 20° to a heading of 340° and time for 2 minutes
   - 10° to a heading of 350° and time for 3 minutes
2. After the appropriate time above (Point C), make a standard rate turn to the right for:
   - 30° teardrop—210° to the final course heading of 180° (Point D)
   - 20° teardrop—200° to the final course heading of 180° (Point D)
   - 10° teardrop—190° to the final course heading of 180° (Point D)

By using the different teardrop patterns, a pilot is afforded the ability to manage time more efficiently. For instance, a 10° pattern for 3 minutes provides about three times the distance

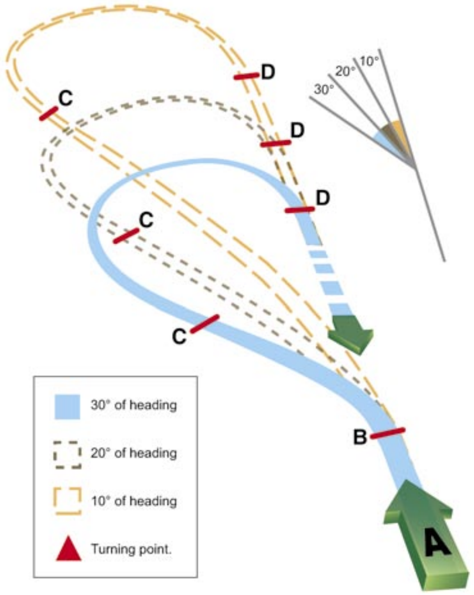

Figure 5-44. *Teardrop Pattern (Entire Pattern in Level Flight).*

(and time) than a 30° pattern. Pattern selection should be based upon an individual assessment of the procedure turn requirements to include wind, complexity, the individual preparedness, etc.

Circling Approach Patterns

Pattern I

1. At A, start timing for 2 minutes from A to B; reduce airspeed to approach speed. *[Figure 5-45]*
2. At B, make a standard rate turn to the left for 45°.
3. At the completion of the turn, time for 45 seconds to C.
4. At C, turn to the original heading; fly 1 minute to D, lowering the landing gear and flaps.
5. At D, turn right 180°, rolling-out at E on the reciprocal of the entry heading.
6. At E, enter a 500 fpm rate descent. At the end of a 500 foot descent, enter a straight constant-airspeed climb, retracting gear and flaps.

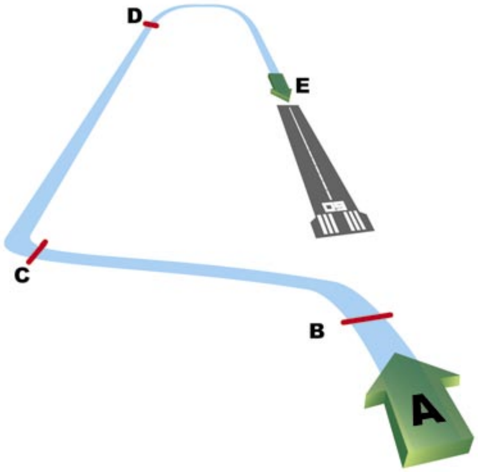

Figure 5-45. *Circling Approach Pattern I (Imaginary Runway).*

Pattern II

Steps:

1. At A, start timing for 2 minutes from A to B; reduce airspeed to approach speed. *[Figure 5-46]*
2. At B, make a standard rate turn to the left for 45°.
3. At the completion of the turn, time for 1 minute to C.
4. At C, turn right for 180° to D; fly for 1-1/2 minutes to E, lowering the landing gear and flaps.
5. At E, turn right for 180°, rolling-out at F.
6. At F, enter a 500 fpm rate descent. At the end of a 500 foot descent, enter a straight constant-airspeed climb, retracting gear and flaps.

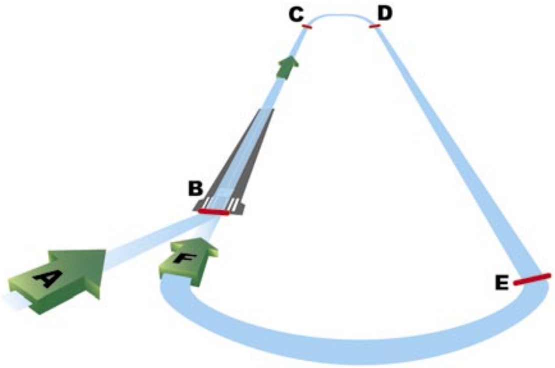

Figure 5-46. *Circling Approach Pattern II (Imaginary Runway).*

270°
VOR 1
140
130
120
110
100
90
GARMIN
NAV1 108.00 113.00
NAV2 108.00 110.60
134.000 COM1
123.800 118.000 COM2
4000
4300
4200
4100
3900
3800
29.92IN
TAS 106KT
NORTH UP
OAT 6°C
XPDR 5537 IDNT LCL10:12:34
INSET
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
STBY BATT
OFF
TEST
MASTER
AVIONICS
ON
ALT BAT
BUS 1 BUS 2
DIMMING
AVIONICS
SW / CB PANELS
LAND TAXI NAV
LAND
TAXI

Chapter 5, Section II

Airplane Basic Flight Maneuvers

Using an Electronic Flight Display

Introduction

The previous chapters have laid the foundation for instrument flying. The pilot's ability to use and interpret the information displayed and apply corrective action is required to maneuver the aircraft and maintain safe flight. A pilot must recognize that each aircraft make and model flown may require a different technique. Aircraft weight, speed, and configuration changes require the pilot to vary his or her technique in order to perform successful attitude instrument flying. A pilot must become familiar with all sections of the Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM) prior to performing any flight maneuver.

Chapter 5–II describes basic attitude instrument flight maneuvers and explains how to perform each one by interpreting the indications presented on the electronic flight display (EFD). In addition to normal flight maneuvers, "partial panel" flight will be addressed. With the exception of the instrument takeoff, all flight maneuvers can be performed on "partial panel" with the Attitude Heading Reference System (AHRS) unit simulated or rendered inoperative.

Straight-and-Level Flight

Pitch Control

The pitch attitude of an airplane is the angle between the longitudinal axis of the airplane and the actual horizon. In level flight, the pitch attitude varies with airspeed and load. For training purposes, the latter factor can normally be disregarded in small airplanes. At a constant airspeed, there is only one specific pitch attitude for level flight. At slow cruise speeds, the level flight attitude is nose-high with indications as in *Figure 5-47*; at fast cruise speeds, the level flight attitude is nose-low. *[Figure 5-48] Figure 5-49* shows the indications for the attitude at normal cruise speeds.

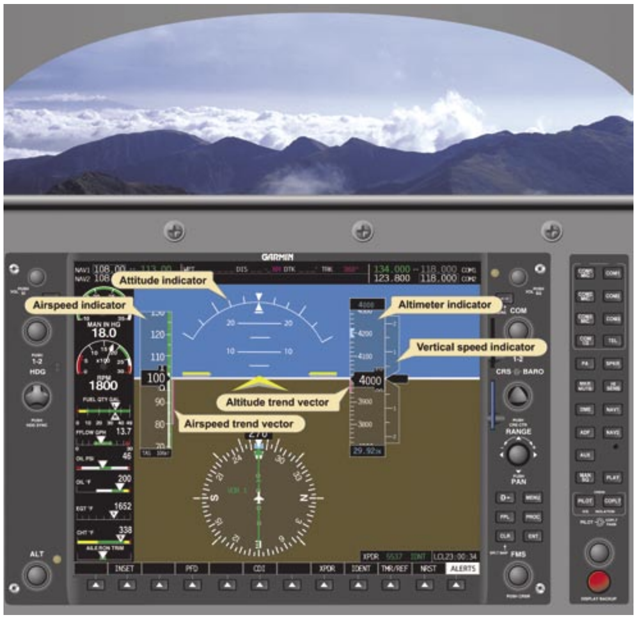

Figure 5-47. *Pitch Attitude and Airspeed in Level Flight, Slow Cruise Speed.*

The instruments that directly or indirectly indicate pitch on the Primary Flight Display (PFD) are the attitude indicator, altimeter, vertical speed indicator (VSI), airspeed indicator (ASI), and both airspeed and altitude trend indicators.

Attitude Indicator

The attitude indicator gives the pilot a direct indication of the pitch attitude. The increased size of the attitude display on the EFD system greatly increases situational awareness for the pilot. Most attitude indicators span the entire width of the PFD screen.

The aircraft pitch attitude is controlled by changing the deflection of the elevator. As the pilot pulls back on the control yoke causing the elevator to rise, the yellow chevron will begin to show a displacement up from the artificial horizon line. This is caused by the AHRS unit sensing the changing angle between the longitudinal plane of the earth and the longitudinal axis of the aircraft.

The attitude indicator displayed on the PFD screen is a representation of outside visual cues. Rather than rely on the natural horizon visible during visual flight rules (VFR) flight, the pilot must rely on the artificial horizon of the PFD screen.

During normal cruise airspeed, the point of the yellow chevron (aircraft symbol) will be positioned on the artificial horizon. Unlike conventional attitude indicators, the EFD attitude indicator does not allow for manipulating the position of the chevron in relationship to the artificial horizon. The position is fixed and therefore will always display the pitch angle as calculated by the AHRS unit.

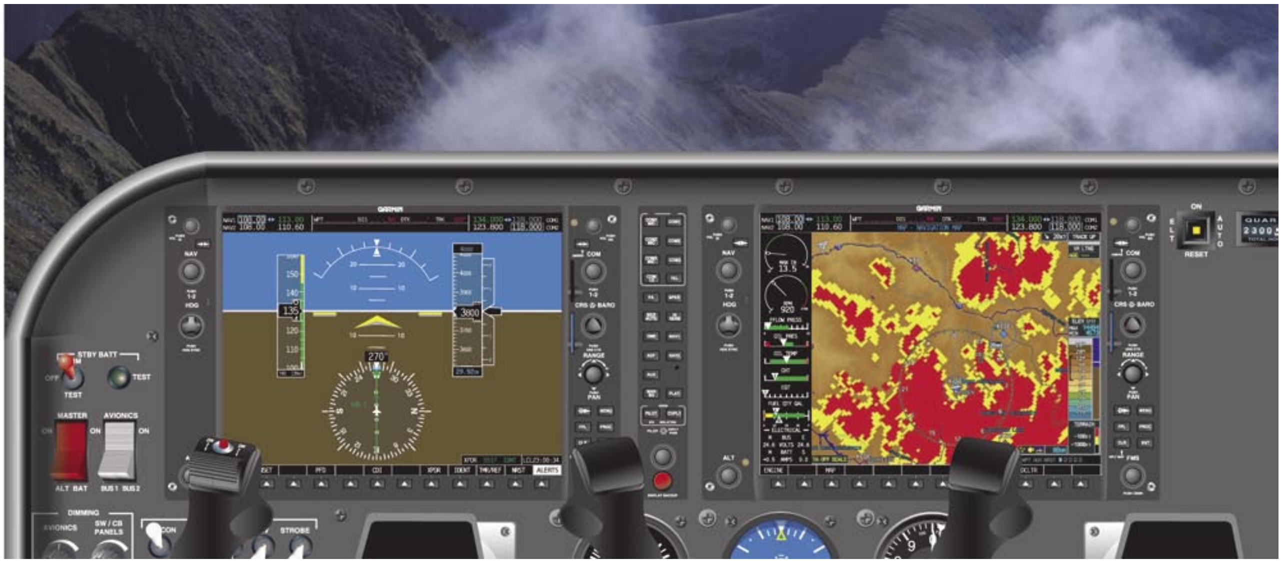

Figure 5-48. *Pitch Attitude Decreasing and Airspeed Increasing—Indicates Need to Increase Pitch.*

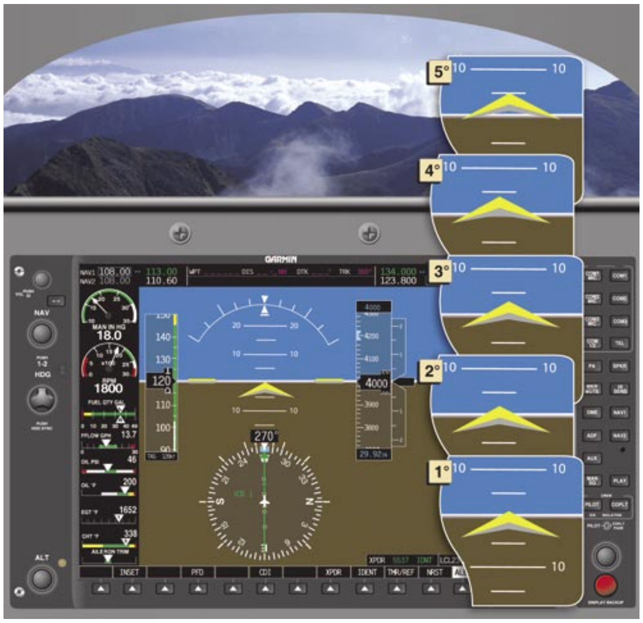

Figure 5-49. *Various Pitch Attitudes (Right), Aircraft Shown in Level Flight.*

The attitude indicator only shows pitch attitude and does not indicate altitude. A pilot should not attempt to maintain level flight using the attitude indicator alone. It is important for the pilot to understand how small displacements both up and down can affect the altitude of the aircraft. To achieve this, the pilot should practice increasing the pitch attitude incrementally to become familiar with how each degree of pitch changes the altitude. *[Figures 5-50* and *5-51]* In both cases, the aircraft will slow and gain altitude.

The full height of the chevron is approximately 5° and provides an accurate reference for pitch adjustment. It is imperative that the pilot make the desired changes to pitch by referencing the attitude indicator and then trimming off any excess control pressures. Relieving these pressures will allow for a more stabilized flight and will reduce pilot work load. Once the aircraft is trimmed for level flight, the pilot must smoothly and precisely manipulate the elevator control forces in order to change the pitch attitude.

To master the ability to smoothly control the elevator, a pilot must develop a very light touch on the control yoke. The thumb and two fingers are normally sufficient to move the control yoke. The pilot should avoid griping the yoke with a full fist. When a pilot grips the yoke with a full fist, there is a tendency to apply excess pressures, thus changing the aircraft attitude.

Practice making smooth, small pitch changes both up and down until precise corrections can be made. With practice a pilot will be able to make pitch changes in 1° increments, smoothly controlling the attitude of the aircraft.

The last step in mastering elevator control is trim. Trimming the aircraft to relieve any control pressures is essential for smooth attitude instrument flight. To accomplish this, momentarily release the control yoke. Note which way the aircraft pitch attitude wants to move. Grasp the control yoke again and then reapply the pressure to return the attitude to the previous position. Apply trim in the direction of the control pressure. Small applications of trim will make large changes in the pitch attitude. Be patient and make multiple changes to trim, if necessary.

Once the aircraft is in trim, relax on the control yoke as much as practicable. When pressure is held on the yoke, unconscious pressures are applied to the elevator and ailerons which displaces the aircraft from its desired flight path. If the aircraft is in trim, in calm, non-turbulent air, a pilot should be able to release the control yoke and maintain level flight for extended periods of time. This is one of the hardest skills to learn prior to successfully flying in instrument meteorological conditions (IMC).

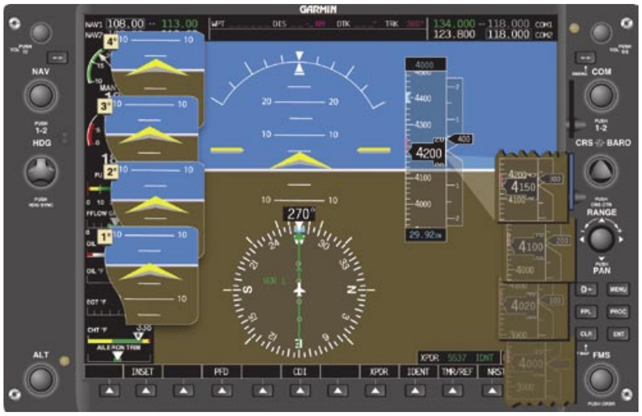

Figure 5-50. *Pitch Indications for Various Attitudes (1° through 5°).*

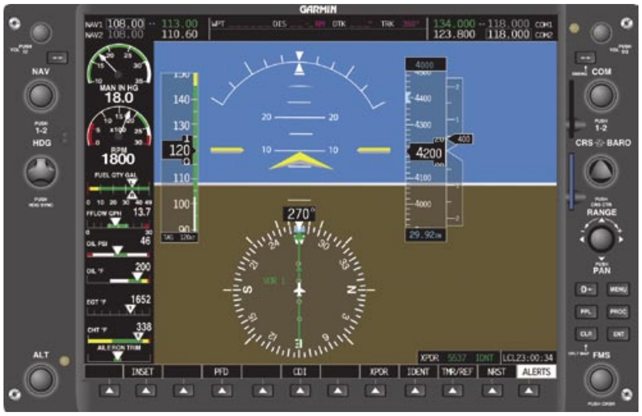

Figure 5-51. *Pitch Illustrated at 10°.*

Altimeter

At constant power, any deviation from level flight (except in turbulent air) must be the result of a pitch change. If the power is constant, the altimeter gives an indirect indication of the pitch attitude in level flight. Since the altitude should remain constant when the airplane is in level flight, any deviation from the desired altitude signals the need for a pitch change. For example, if the aircraft is gaining altitude, the nose must be lowered.

In the PFD, as the pitch starts to change, the altitude trend indicator on the altitude tape will begin to show a change in the direction of displacement. The rate at which the trend indicator grows and the altimeter numbers change aids the pilot in determining how much of a pitch change is necessary to stop the trend.

As a pilot becomes familiar with a specific aircraft's instruments, he or she learns to correlate pitch changes, altimeter tapes, and altitude trend indicators. By adding the altitude tape display and the altitude trend indicator into the scan along with the attitude indicator, a pilot starts to develop the instrument cross-check.

Partial Panel Flight

One important skill to practice is partial panel flight by referencing the altimeter as the primary pitch indicator. Practice controlling the pitch by referencing the altitude tape and trend indicator alone without the use of the attitude indicator. Pilots need to learn to make corrections to altitude deviations by referencing the rate of change of the altitude tape and trend indicator. When operating in IMC and in a partial panel configuration, the pilot should avoid abrupt changes to the control yoke. Reacting abruptly to altitude changes can lead to large pitch changes and thus a larger divergence from the initial altitude.

When a pilot is controlling pitch by the altitude tape and altitude trend indicators alone, it is possible to overcontrol the aircraft by making a larger than necessary pitch correction. Overcontrolling will cause the pilot to move from a nose-high attitude to a nose-low attitude and vice versa. Small changes to pitch are required to insure prompt corrective actions are taken to return the aircraft to its original altitude with less confusion.

When an altitude deviation occurs, two actions need to be accomplished. First, make a smooth control input to stop the needle movement. Once the altitude tape has stopped moving, make a change to the pitch attitude to start back to the entry altitude.

During instrument flight with limited instrumentation, it is imperative that only small and precise control inputs are made. Once a needle movement is indicated denoting a deviation in altitude, the pilot needs to make small control inputs to stop the deviation. Rapid control movements will only compound the deviation by causing an oscillation effect. This type of oscillation can quickly cause the pilot to become disoriented and begin to fixate on the altitude. Fixation on the altimeter can lead to a loss of directional control as well as airspeed control.

As a general rule of thumb, for altitude deviations less than 100 feet, utilize a pitch change of 1°, which equates to 1/5 of the thickness of the chevron. Small incremental pitch changes will allow the performance to be evaluated and will eliminate overcontrolling of the aircraft.

Instrumentation needs to be utilized collectively, but failures will occur which leave the pilot with only limited instrumentation. That is why partial panel flying training is important. If the pilot understands how to utilize each instrument independently, no significant change is encountered in carrying out the flight when other instruments fail.

VSI Tape

The VSI tape provides for an indirect indication of pitch attitude and gives the pilot a more immediate indication of a pending altitude deviation. In addition to trend information, the vertical speed also gives a rate indication. By using the VSI tape in conjunction with the altitude trend tape, a pilot will have a better understanding of how much of a correction needs to be made. With practice, the pilot will learn the performance of a particular aircraft and know how much pitch change is required in order to correct for a specific rate indication.

Unlike older analog VSIs, new glass panel displays have instantaneous VSIs. Older units had a lag designed into the system that was utilized to indicate rate information. The new glass panel displays utilize a digital air data computer that does not indicate a lag. Altitude changes are shown immediately and can be corrected for quickly.

The VSI tape should be used to assist in determining what pitch changes are necessary to return to the desired altitude. A good rule of thumb is to use a vertical speed rate of change that is double the altitude deviation. However, at no time should the rate of change be more than the optimum rate of climb or descent for the specific aircraft being flown. For example, if the altitude is off by 200 feet from the desired altitude, then a 400 feet per minute (fpm) rate of change would be sufficient to get the aircraft back to the original altitude. If the altitude has changed by 700 feet, then doubling that would necessitate a 1,400 fpm change. Most aircraft are not capable of that, so restrict changes to no more than optimum climb and descent. An optimum rate of change would vary between 500 and 1,000 fpm.

One error the instrument pilot encounters is overcontrolling. Overcontrolling occurs when a deviation of more than 200 fpm is indicated over the optimum rate of change. For example, an altitude deviation of 200 feet is indicated on the altimeter, a vertical speed rate of 400 feet should be indicated on the gauge. If the vertical speed rate showed 600 fpm (200 more than optimum), the pilot would be overcontrolling the aircraft.

When returning to altitude, the primary pitch instrument is the VSI tape. If any deviation from the desired vertical speed is indicated, make the appropriate pitch change using the attitude indicator.

As the aircraft approaches the target altitude, the vertical speed rate can be slowed in order to capture the altitude in a more stabilized fashion. Normally within 10 percent of the rate of climb or descent from the target altitude, begin to slow the vertical speed rate in order to level off at the target altitude. This will allow the pilot to level at the desired altitude without rapid control inputs or experiencing discomfort due to G-load.

Airspeed Indicator (ASI)

The ASI presents an indirect indication of the pitch attitude. At a constant power setting and pitch attitude, airspeed remains constant. As the pitch attitude lowers, airspeed increases, and the nose should be raised.

As the pitch attitude is increased, the nose of the aircraft will raise, which will result in an increase in the angle of attack as well as an increase in induced drag. The increased drag will begin to slow the momentum of the aircraft which will be indicated on the ASI. The airspeed trend indicator will show a trend as to where the airspeed will be in 6 seconds. Conversely, if the nose of the aircraft should begin to fall, the angle of attack as well as induced drag will decrease.

There is a lag associated with the ASI when using it as a pitch instrument. It is not a lag associated with the construction of the ASI, but a lag associated with momentum change. Depending on the rate of momentum change, the ASI may not indicate a pitch change in a timely fashion. If the ASI is being used as the sole reference for pitch change, it may not allow for a prompt correction. However, if smooth pitch changes are executed, modern glass panel displays are capable of indicating 1 knot changes in airspeed and also capable of projecting airspeed trends.

When flying by reference to flight instruments alone, it is imperative that all of the flight instruments be cross-checked for pitch control. By cross-checking all pitch related instruments, the pilot can better visualize the aircraft attitude at all times.

As previously stated, the primary instrument for pitch is the instrument that gives the pilot the most pertinent information for a specific parameter. When in level flight and maintaining a constant altitude, what instrument shows a direct indication of altitude? The only instrument that is capable of showing altitude is the altimeter. The other instruments are supporting instruments that are capable of showing a trend away from altitude, but do not directly indicate an altitude.

The supporting instruments forewarn of an impending altitude deviation. With an efficient cross-check, a proficient pilot will be better able to maintain altitude.

Bank Control

This discussion assumes the aircraft is being flown in coordinated flight which means the longitudinal axis of the aircraft is aligned with the relative wind. On the PFD, the attitude indicator shows if the wings are level. The turn rate indicator, slip/skid indicator, and the heading indicator also indicate whether or not the aircraft is maintaining a straight (zero bank) flight path.

Attitude Indicator

The attitude indicator is the only instrument on the PFD that has the capability of displaying the precise bank angle of the aircraft. This is made possible by the display of the roll scale depicted as part of the attitude indicator.

Figure 5-52 identifies the components that make up the attitude indicator display. Note that the top of the display is blue, representing sky, the bottom is brown, depicting dirt, and the white line separating them is the horizon. The lines parallel to the horizon line are the pitch scale which is marked in 5° increments and labeled every 10°. The pitch scale always remains parallel to the horizon.

The curved line in the blue area is the roll scale. The triangle on the top of the scale is the zero index. The hash marks on the scale represent the degree of bank. *[Figure 5-53]* The roll scale always remains in the same position relative to the horizon line.

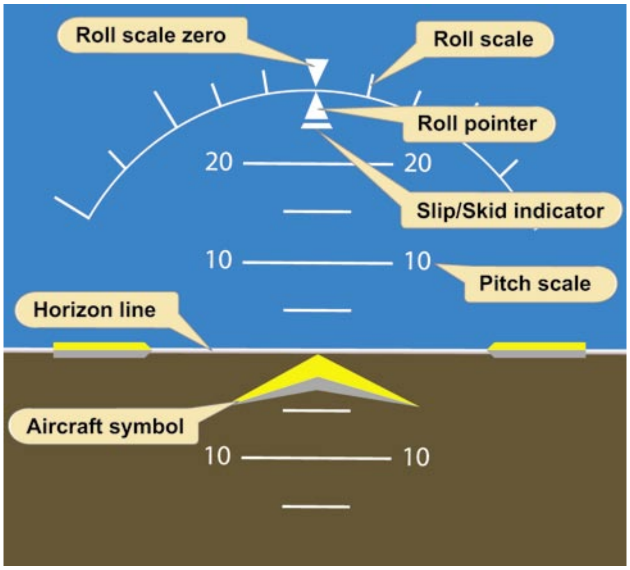

Figure 5-52. *Attitude Indicator.*

The roll pointer indicates the direction and degree of bank. *[Figure 5-53]* The roll pointer is aligned with the aircraft symbol. The roll pointer indicates the angle of the lateral axis of the aircraft compared to the natural horizon. The slip/skid indicator will show if the longitudinal axis of the aircraft is aligned with the relative wind, which is coordinated flight. With the roll index and the slip/skid indicator aligned, any deflection, either right or left of the roll index will cause the aircraft to turn in that direction. With the small graduations on the roll scale, it is easy to determine the bank angle within approximately 1°. In coordinated flight, if the roll index is aligned with the roll pointer, the aircraft is achieving straight flight.

An advantage of EFDs is the elimination of the precession error. Precession error in analog gauges is caused by forces being applied to a spinning gyro. With the new solid state instruments, precession error has been eliminated.

Since the attitude indicator is capable of showing precise pitch and bank angles, the only time that the attitude indicator will be a primary instrument is when attempting to fly at a specific bank angle or pitch angle. Other times, the attitude instrument can be thought of as a control instrument.

Horizontal Situation Indicator (HSI)

The HSI is a rotating 360° compass card that indicates magnetic heading. The HSI is the only instrument that is capable of showing exact headings. The magnetic compass can be used as a backup instrument in case of an HSI failure; however, due to erratic, unstable movements, it is more likely to be used a supporting instrument.

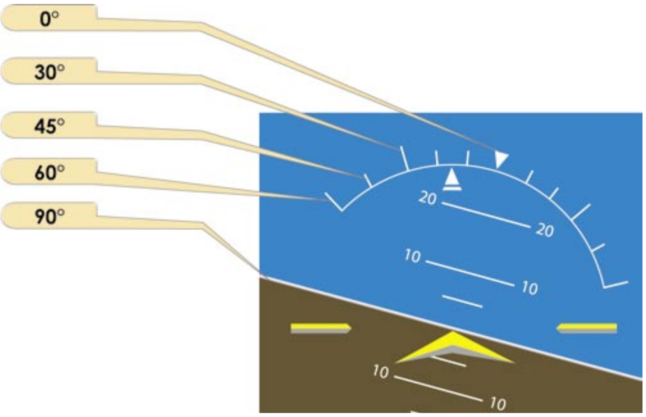

Figure 5-53. *Attitude Indicator Showing a 15° left bank.*

In order for the pilot to achieve the desired rate of change, it is important for him or her to understand the relationship between the rate at which the HSI changes heading displays and the amount of bank angle required to meet that rate of change. A very small rate of heading change means the bank angle is small, and it will take more time to deviate from the desired straight flight path. A larger rate of heading change means a greater bank angle will happen at a faster rate.

Heading Indicator

The heading indicator is the large black box with a white number that indicates the magnetic heading of the aircraft. *[Figure 5-54]* The aircraft heading is displayed to the nearest degree. When this number begins to change, the pilot should be aware that straight flight is no longer being achieved.

Turn Rate Indicator

The turn rate indicator gives an indirect indication of bank. It is a magenta trend indicator capable of displaying half-standard as well as standard rate turns to both the left and right. *[Figure 5-54]* The turn indicator is capable of indicating turns up to 4° per second by extending the magenta line outward from the standard rate mark. If the rate of turn has

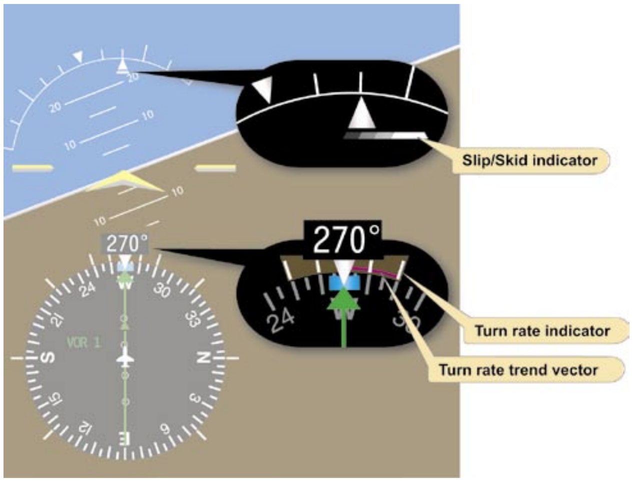

Figure 5-54. *Slip/Skid and Turn Rate Indicator.*

exceeded 4° per second, the magenta line can not precisely indicate where the heading will be in the next 6 seconds, the magenta line freezes and an arrowhead will be displayed. This alerts the pilot to the fact that the normal range of operation has been exceeded.

Slip/Skid Indicator

The slip/skid indicator is the small portion of the lower segmented triangle displayed on the attitude indicator. This instrument depicts whether the aircraft's longitudinal axis is aligned with the relative wind. *[Figure 5-54]*

The pilot must always remember to cross-check the roll index to the roll pointer when attempting to maintain straight flight. Any time the heading remains constant and the roll pointer and the roll index are not aligned, the aircraft is in uncoordinated flight. To make a correction, the pilot should apply rudder pressure to bring the aircraft back to coordinated flight.

Power Control

Power produces thrust which, with the appropriate angle of attack of the wing, overcomes the forces of gravity, drag, and inertia to determine airplane performance.

Power control must be related to its effect on altitude and airspeed, since any change in power setting results in a change in the airspeed or the altitude of the airplane. At any given airspeed, the power setting determines whether the airplane is in level flight, in a climb, or in a descent. If the power is increased in straight-and-level flight and the airspeed held constant, the airplane will climb; if power is decreased while the airspeed is held constant, the airplane will descend. On the other hand, if altitude is held constant, the power applied will determine the airspeed.

The relationship between altitude and airspeed determines the need for a change in pitch or power. If the airspeed is off the desired value, always check the altimeter before deciding that a power change is necessary. Think of altitude and airspeed as interchangeable; altitude can be traded for airspeed by lowering the nose, or convert airspeed to altitude by raising the nose. If altitude is higher than desired and airspeed is low, or vice versa, a change in pitch alone may return the airplane to the desired altitude and airspeed. *[Figure 5-55]* If both airspeed and altitude are high or if both are low, then a change in both pitch and power is necessary in order to return to the desired airspeed and altitude. *[Figure 5-56]*

For changes in airspeed in straight-and-level flight, pitch, bank, and power must be coordinated in order to maintain constant altitude and heading. When power is changed to vary airspeed in straight-and-level flight, a single-engine, propeller-driven airplane tends to change attitude around all axes of movement. Therefore, to maintain constant altitude and heading, apply various control pressures in proportion to the change in power. When power is added to increase airspeed, the pitch instruments indicate a climb unless forward-elevator control pressure is applied as the airspeed changes. With an increase in power, the airplane tends to yaw and roll to the left unless counteracting aileron and rudder pressures are applied. Keeping ahead of these changes requires increasing cross-check speed, which varies with the type of airplane and its torque characteristics, the extent of power and speed change involved.

Power Settings

Power control and airspeed changes are much easier when approximate power settings necessary to maintain various airspeeds in straight-and-level flight are known in advance. However, to change airspeed by any appreciable amount, the common procedure is to underpower or overpower on initial power changes to accelerate the rate of airspeed change. (For small speed changes, or in airplanes that decelerate or accelerate rapidly, overpowering or underpowering is not necessary.)

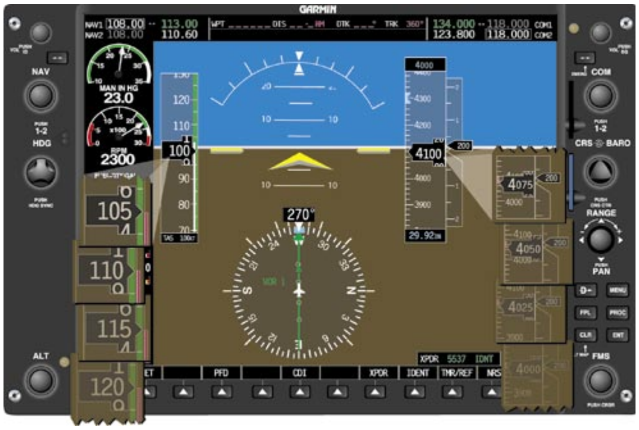

Figure 5-55. *An aircraft decreasing in airspeed while gaining altitude. In this case, the pilot has decreased pitch.*

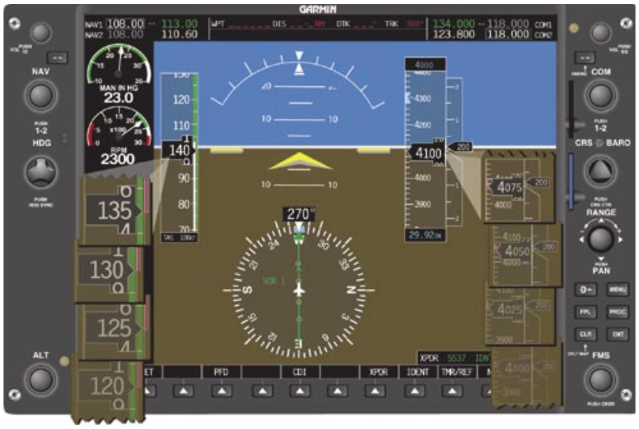

Figure 5-56. *Figure shows both an increase in speed and altitude where pitch adjustment alone is insufficient. In this situation, a reduction of power is also necessary.*

Consider the example of an airplane that requires 23" of manifold pressure (Hg) to maintain a normal cruising airspeed of 120 knots, and 18" Hg to maintain an airspeed of 100 knots. The reduction in airspeed from 120 knots to 100 knots while maintaining straight-and-level flight is discussed below and illustrated in *Figures 5-57, 5-58,* and *5-59*.

Instrument indications, prior to the power reduction, are shown in *Figure 5-57*. The basic attitude is established and maintained on the attitude indicator. The specific pitch, bank, and power control requirements are detected on these primary instruments:

Altimeter—Primary Pitch
Heading Indicator—Primary Bank
Airspeed Indicator—Primary Power

Supporting pitch and bank instruments are shown in *Figure 5-57*. Note that the supporting power instrument is the manifold pressure gauge (or tachometer if the propeller is fixed pitch). However, when a smooth power reduction to approximately 15" Hg (underpower) is made, the manifold pressure gauge becomes the primary power instrument. *[Figure 5-58]* With practice, power setting can be changed with only a brief glance at the power instrument, by sensing the movement of the throttle, the change in sound, and the changes in the feel of control pressures.

As the thrust decreases, increase the speed of the cross-check and be ready to apply left rudder, back-elevator, and aileron control pressure the instant the pitch-and-bank instruments show a deviation from altitude and heading. As proficiency is obtained, a pilot will learn to cross-check, interpret, and control the changes with no deviation of heading and altitude. Assuming smooth air and ideal control technique, as airspeed decreases, a proportionate increase in airplane pitch attitude is required to maintain altitude. Similarly, effective torque control means counteracting yaw with rudder pressure.

As the power is reduced, the altimeter is primary for pitch, the heading indicator is primary for bank, and the manifold pressure gauge is momentarily primary for power (at 15" Hg in *Figure 5-58*). Control pressures should be trimmed off as the airplane decelerates. As the airspeed approaches the desired airspeed of 100 knots, the manifold pressure is adjusted to approximately 18" Hg and becomes the supporting power instrument. The ASI again becomes primary for power. *[Figure 5-58]*

Airspeed Changes in Straight-and-Level Flight

Practice of airspeed changes in straight-and-level flight provides an excellent means of developing increased proficiency in all three basic instrument skills, and brings out some common errors to be expected during training in straight-and-level flight. Having learned to control the airplane in a clean configuration (minimum drag conditions), increase proficiency in cross-check and control by practicing speed changes while extending or retracting the flaps and landing gear. While practicing, be sure to comply with the airspeed limitations specified in the POH/AFM for gear and flap operation.

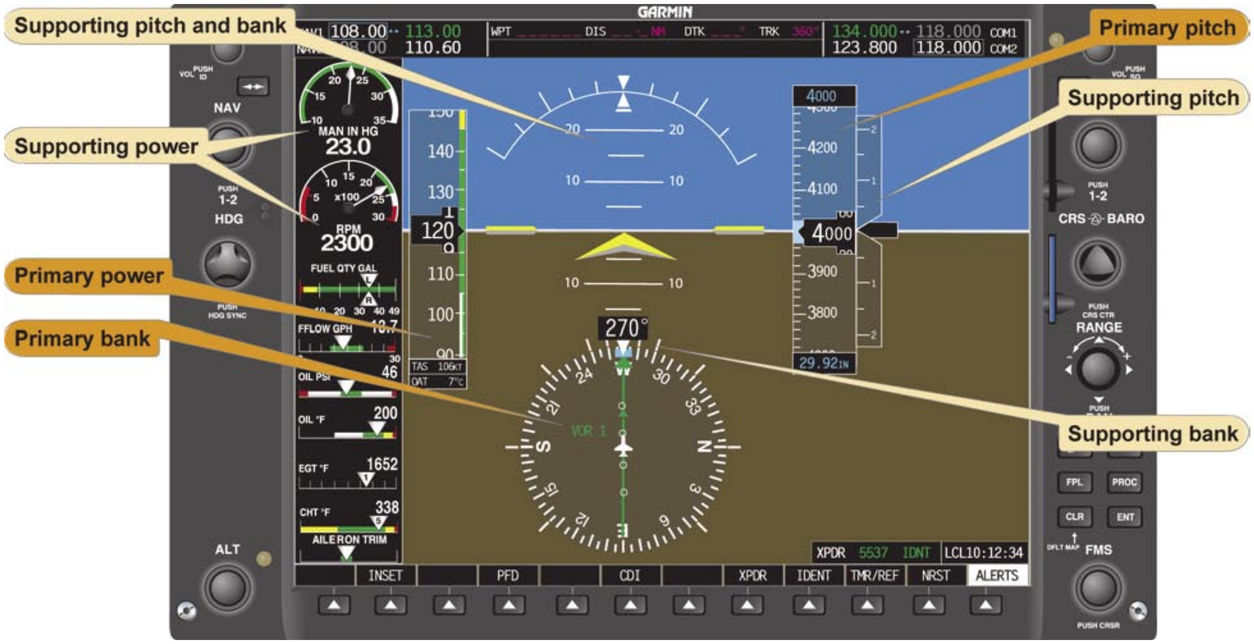

Figure 5-57. *Straight-and-Level Flight (Normal Cruising Speed).*

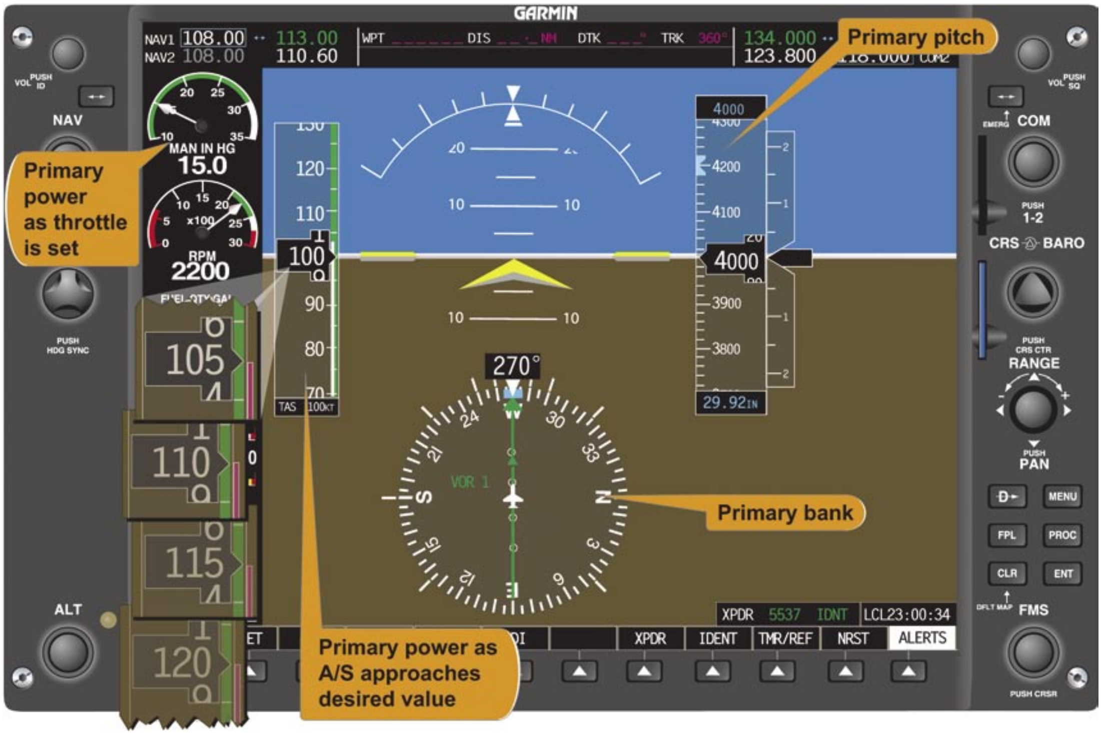
GARMIN
Primary pitch
Primary power as throttle is set
MAN IN HG 15.0
RPM 2200
100
4000
270°
Primary bank
Primary power as A/S approaches desired value

Figure 5-58. *Straight-and-Level Flight (Airspeed Decreasing).*

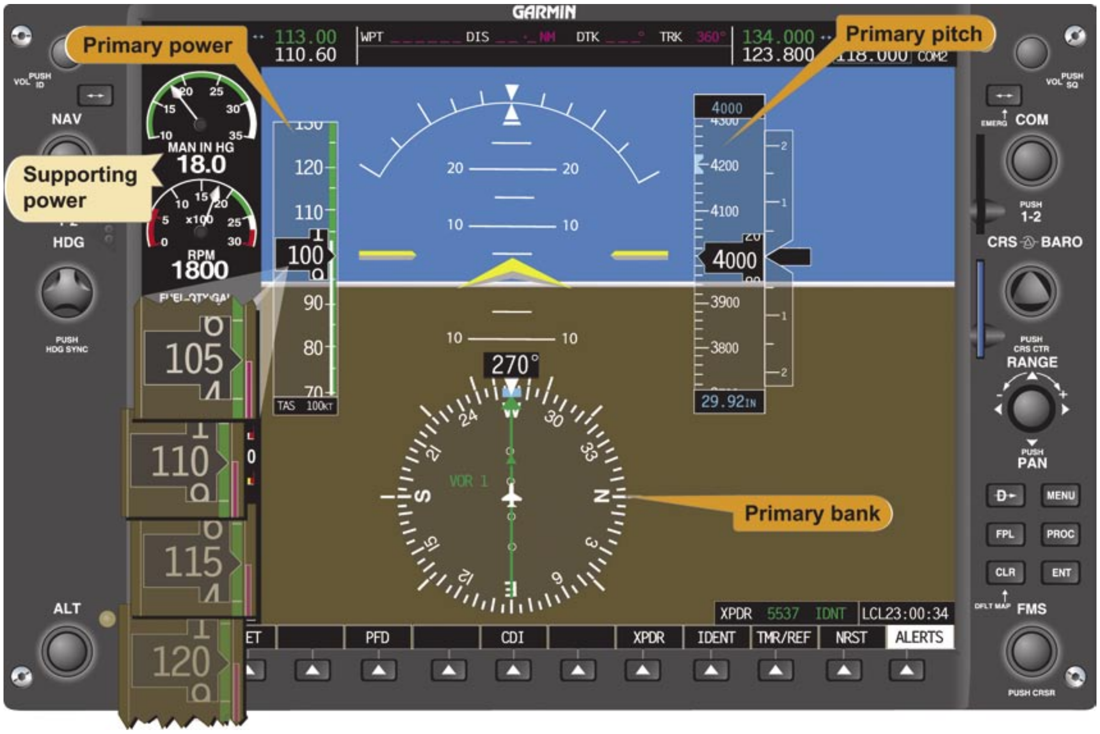
GARMIN
Primary power
Primary pitch
Supporting power
MAN IN HG 18.0
RPM 1800
100
4000
270°
Primary bank

Figure 5-59. *Straight-and-Level Flight (Reduced Airspeed Stabilized).*

Sudden and exaggerated attitude changes may be necessary in order to maintain straight-and-level flight as the landing gear is extended and the flaps are lowered in some airplanes. The nose tends to pitch down with gear extension, and when flaps are lowered, lift increases momentarily (at partial flap settings) followed by a marked increase in drag as the flaps near maximum extension.

Control technique varies according to the lift and drag characteristics of each airplane. Accordingly, knowledge of the power settings and trim changes associated with different combinations of airspeed, gear, and flap configurations will reduce instrument cross-check and interpretation problems. *[Figure 5-60]*

For example, assume that in straight-and-level flight instruments indicate 120 knots with power at 23" Hg manifold pressure/2,300 revolutions per minute (rpm), gear and flaps up. After reduction in airspeed, with gear and flaps fully extended, straight-and-level flight at the same altitude requires 25" Hg manifold pressure/2,500 rpm. Maximum gear extension speed is 115 knots; maximum flap extension speed is 105 knots. Airspeed reduction to 95 knots, gear and flaps down, can be made in the following manner:

1. Maintain rpm at 2,500, since a high power setting will be used in full drag configuration.
2. Reduce manifold pressure to 10" Hg. As the airspeed decreases, increase cross-check speed.
3. Make trim adjustments for an increased angle of attack and decrease in torque.
4. Lower the gear at 115 knots. The nose may tend to pitch down and the rate of deceleration increases. Increase pitch attitude to maintain constant altitude, and trim off some of the back-elevator pressures. If full flaps are lowered at 105 knots, cross-check, interpretation, and control must be very rapid. A simpler technique is to stabilize attitude with gear down before lowering the flaps.
5. Since 18" Hg manifold pressure will hold level flight at 100 knots with the gear down, increase power smoothly to that setting as the ASI shows approximately 105 knots, and retrim. The attitude indicator now shows approximately two-and-a-half bar width nose-high in straight-and-level flight.
6. Actuate the flap control and simultaneously increase power to the predetermined setting (25" Hg) for the desired airspeed, and trim off the pressures necessary to hold constant altitude and heading. The attitude indicator now shows a bar width nose-low in straight-and-level flight at 95 knots.

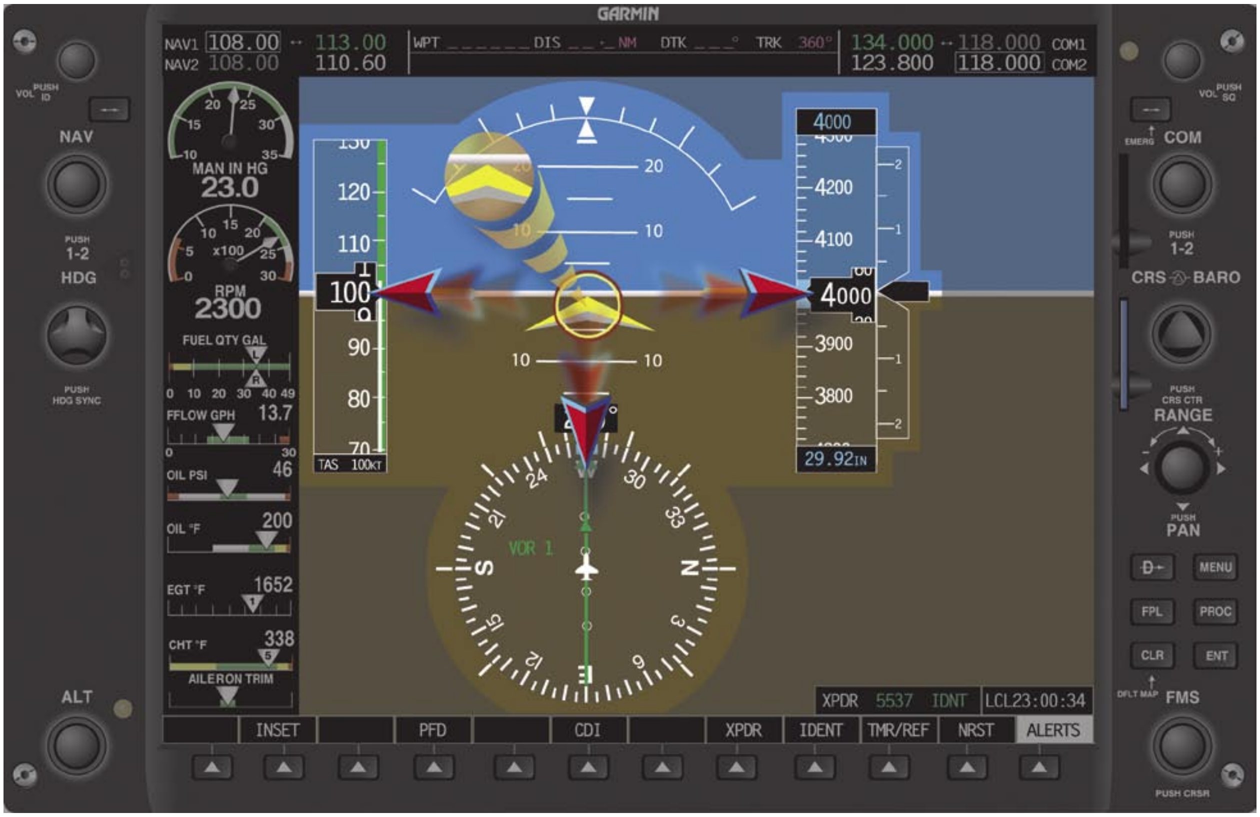
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
MAN IN HG 23.0
RPM 2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
120
110
100
90
80
TAS 100KT
4000
4200
4100
3900
3800
29.92IN
VOR 1
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
NAV
HDG
PUSH HDG SYNC
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS
PUSH CRSR

Figure 5-60. *Cross-check Supporting Instruments.*

Trim Technique

Trim control is one of the most important flight habits to cultivate. Trimming refers to relieving any control pressures that need to be applied by the pilot to the control surfaces to maintain a desired flight attitude. The desired result is for the pilot to be able to take his or her hands off the control surfaces and have the aircraft remain in the current attitude. Once the aircraft is trimmed for hands-off flight, the pilot is able to devote more time to monitoring the flight instruments and other aircraft systems.

In order to trim the aircraft, apply pressure to the control surface that needs trimming and roll the trim wheel in the direction pressure is being held. Relax the pressure that is being applied to the control surface and monitor the primary instrument for that attitude. If the desired performance is achieved, fly hands off. If additional trimming is required, redo the trimming steps.

An aircraft is trimmed for a specific airspeed, not pitch attitude or altitude. Any time an aircraft changes airspeed there is a need to re-trim. For example, an aircraft is flying at 100 knots straight-and-level. An increase of 50 rpm will cause the airspeed to increase. As the airspeed increases, additional lift will be generated and the aircraft will climb. Once the additional thrust has stabilized at some higher altitude, the airspeed will again stabilize at 100 knots.

This demonstrates how trim is associated with airspeed and not altitude. If the initial altitude is to be maintained, forward pressure would need to be applied to the control wheel while the trim wheel needs to be rolled forward to eliminate any control pressures. Rolling forward on the trim wheel is equal to increasing for a trimmed airspeed. Any time the airspeed is changed, re-trimming will be required. Trimming can be accomplished during any transitional period; however, prior to final trimming, the airspeed must be held constant. If the airspeed is allowed to change, the trim will not be adjusted properly and the altitude will vary until the airspeed for which the aircraft is trimmed is achieved.

Common Errors in Straight-and-Level Flight

Pitch

Pitch errors usually result from the following errors:

1. Improper adjustment of the yellow chevron (aircraft symbol) on the attitude indicator.

   Corrective Action: Once the aircraft has leveled off and the airspeed has stabilized, make small corrections to the pitch attitude to achieve the desired performance. Cross-check the supporting instruments for validation.

2. Insufficient cross-check and interpretation of pitch instruments. *[Figure 5-61]*

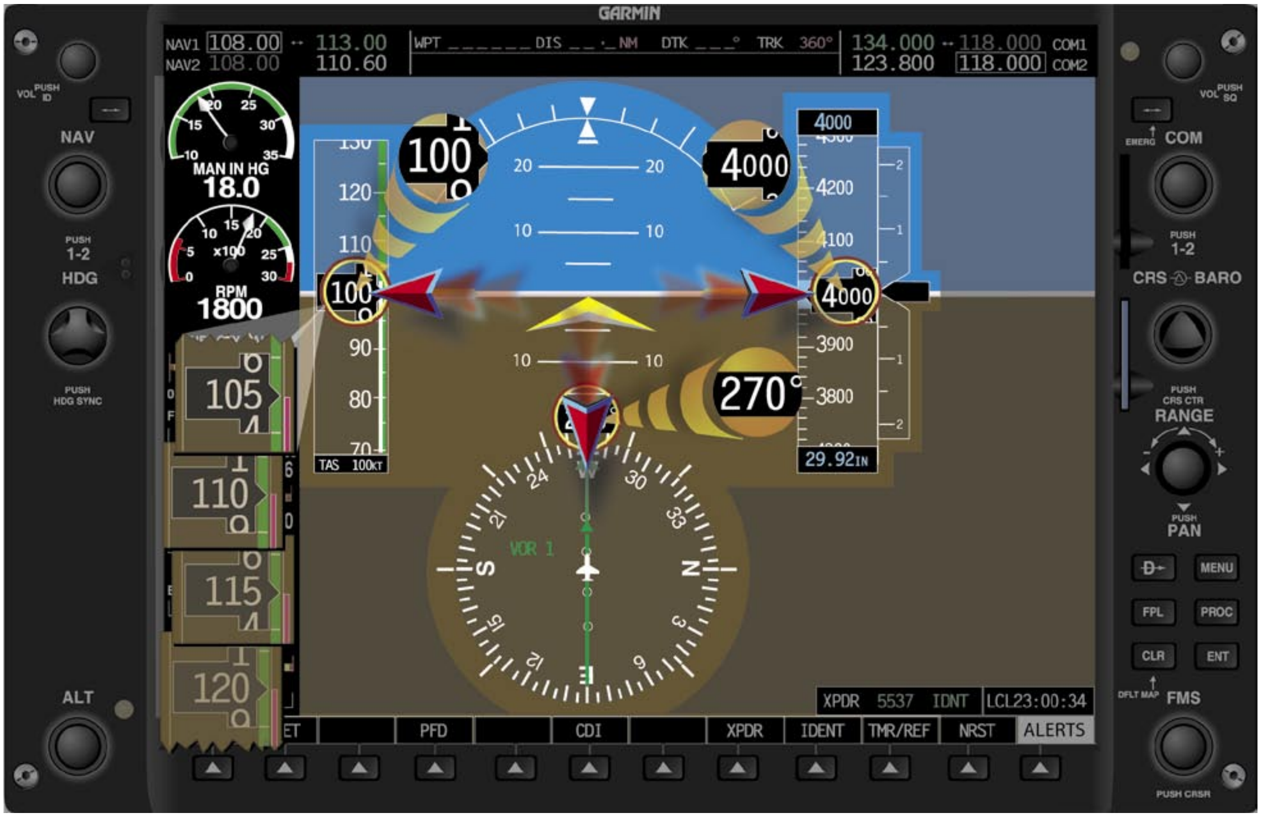

Figure 5-61. *Insufficient cross-check. The problem is power and not nose-high. In this case, the pilot decreased pitch inappropriately.*

Example: The airspeed indication is low. The pilot, believing a nose-high pitch attitude exists, applies forward pressure without noting that a low power setting is the cause of the airspeed discrepancy.

Corrective Action: Increase the rate of cross-check of all the supporting flight instruments. Airspeed and altitude should be stabilized before making a control input.

3. Acceptance of deviations.

   Example: A pilot has an altitude range of ±100 feet according to the practical test standards for straight-and level-flight. When the pilot notices that the altitude has deviated by 60 feet, no correction is made because the altitude is holding steady and is within the standards.

   Corrective Action: The pilot should cross-check the instruments and, when a deviation is noted, prompt corrective actions should be taken in order to bring the aircraft back to the desired altitude. Deviations from altitude should be expected but not accepted.

4. Overcontrolling—Excessive Pitch Changes.

   Example: A pilot notices a deviation in altitude. In an attempt to quickly return to altitude, the pilot makes a large pitch change. The large pitch change destabilizes the attitude and compounds the error.

   Corrective Action: Small, smooth corrections should be made in order to recover to the desired altitude (0.5° to 2° depending on the severity of the deviation). Instrument flying is comprised of small corrections to maintain the aircraft attitude. When flying in IMC, a pilot should avoid making large attitude changes in order to avoid loss of aircraft control and spatial disorientation.

5. Failure to Maintain Pitch Corrections.

   Pitch changes need to be made promptly and held for validation. Many times pilots will make corrections and allow the pitch attitude to change due to not trimming the aircraft. It is imperative that any time a pitch change is made; the trim is readjusted in order to eliminate any control pressures that are being held. A rapid cross-check will aid in avoiding any deviations from the desired pitch attitude.

   Example: A pilot notices a deviation in altitude. A change in the pitch attitude is accomplished but no adjustment to the trim is made. Distractions cause the pilot to slow the cross-check and an inadvertent reduction in the pressure to the control column commences. The pitch attitude then changes, thus complicating recovery to the desired altitude.

   Corrective Action: The pilot should initiate a pitch change and then immediately trim the aircraft to relieve any control pressures. A rapid cross-check should be established in order to validate the desired performance is being achieved.

6. Fixation During Cross-Check.

   Devoting an unequal amount of time to one instrument either for interpretation or assigning too much importance to an instrument. Equal amounts of time should be spent during the cross-check to avoid an unnoticed deviation in one of the aircraft attitudes.

   Example: A pilot makes a correction to the pitch attitude and then devotes all of the attention to the altimeter to determine if the pitch correction is valid. During this time, no attention is paid to the heading indicator which shows a turn to the left. *[Figure 5-62]*

   Corrective Action: The pilot should monitor all instrumentation during the cross-check. Do not fixate on one instrument waiting for validation. Continue to scan all instruments to avoid allowing the aircraft to begin a deviation in another attitude.

Heading

Heading errors usually result from but are not limited to the following errors:

1. Failure to cross-check the heading indicator, especially during changes in power or pitch attitude.
2. Misinterpretation of changes in heading, with resulting corrections in the wrong direction.
3. Failure to note and remember a preselected heading.
4. Failure to observe the rate of heading change and its relation to bank attitude.
5. Overcontrolling in response to heading changes, especially during changes in power settings.
6. Anticipating heading changes with premature application of rudder pressure.
7. Failure to correct small heading deviations. Unless zero error in heading is the goal, a pilot will tolerate larger and larger deviations. Correction of a 1° error takes far less time and concentration than correction of a 20° error.
8. Correcting with improper bank attitude. If correcting a 10° heading error with a 20° bank correction, the aircraft will roll past the desired heading before the bank is established, requiring another correction in the opposite direction. Do not multiply existing errors with errors in corrective technique.

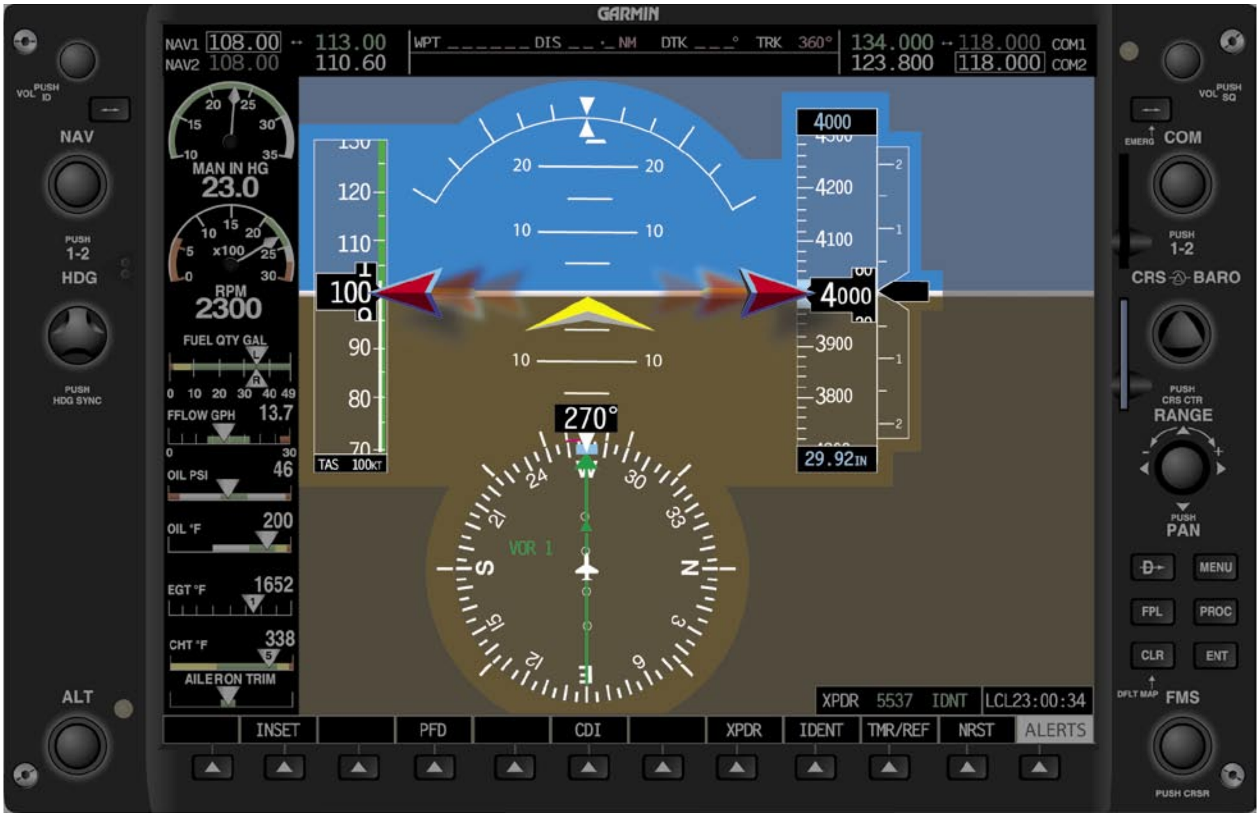
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT ______ DIS __._ NM DTK ___° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
MAN IN HG
23.0
RPM
2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
120
110
100
90
80
TAS 100KT
4000
4200
4100
4000
3900
3800
29.92IN
270°
VOR 1
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP FMS
PUSH CRSR

Figure 5-62. *The pilot has fixated on pitch and altitude, leaving bank indications unattended. Note the trend line to the left.*

9. Failure to note the cause of a previous heading error and thus repeating the same error. For example, the airplane is out of trim, with a left wing low tendency. Repeated corrections for a slight left turn are made, yet trim is ignored.

Power

Power errors usually result from but are not limited to the following errors:

1. Failure to become familiar with the aircraft's specific power settings and pitch attitudes.
2. Abrupt use of throttle.
3. Failure to lead the airspeed when making power changes, climbs or descents.

   Example: When leveling off from a descent, increase the power in order to avoid the airspeed from bleeding off due to the decrease in momentum of the aircraft. If the pilot waits to bring in the power until after the aircraft is established in the level pitch attitude, the aircraft will have already decreased below the speed desired which will require additional adjustment in the power setting.
4. Fixation on airspeed tape or manifold pressure indications during airspeed changes, resulting in erratic control of airspeed, power, as well as pitch and bank attitudes.

Trim

Trim errors usually result from the following faults:

1. Improper adjustment of seat or rudder pedals for comfortable position of legs and feet. Tension in the ankles makes it difficult to relax rudder pressures.
2. Confusion about the operation of trim devices, which differ among various airplane types. Some trim wheels are aligned appropriately with the airplane's axes; others are not. Some rotate in a direction contrary to expectations.
3. Failure to understand the principles of trim and that the aircraft is being trimmed for airspeed, not a pitch attitude.
4. Faulty sequence in trim techniques. Trim should be utilized to relieve control pressures, not to change pitch attitudes. The proper trim technique has the pilot holding the control wheel first and then trimming to relieve any control pressures. Continuous trim changes will be required as the power setting is changed. Utilize the trim continuously, but in small amounts.

Straight Climbs and Descents

Each aircraft will have a specific pitch attitude and airspeed that corresponds to the most efficient climb rate for a specified weight. The POH/AFM contains the speeds that will produce the desired climb. These numbers are based on maximum gross weight. Pilots must be familiar with how the speeds will vary with weight so they can compensate during flight.

Entry

Constant Airspeed Climb From Cruise Airspeed

To enter a constant airspeed climb from cruise airspeed, slowly and smoothly apply aft elevator pressure in order to raise the yellow chevron (aircraft symbol) until the tip points to the desired degree of pitch. *[Figure 5-63]* Hold the aft control pressure and smoothly increase the power to the climb power setting. This increase in power may be initiated either prior to initiating the pitch change or after having established the desired pitch setting. Consult the POH/AFM for specific climb power settings if anything other than a full power climb is desired. Pitch attitudes will vary depending on the type of aircraft being flown. As airspeed decreases, control forces will need to be increased in order to compensate for the additional elevator deflection required to maintain attitude. Utilize trim to eliminate any control pressures. By effectively using trim, the pilot will be better able to maintain the desired pitch without constant attention. The pilot is thus able to devote more time to maintaining an effective scan of all instrumentation.

The VSI should be utilized to monitor the performance of the aircraft. With a smooth pitch transition, the VSI tape should begin to show an immediate trend upward and stabilize on a

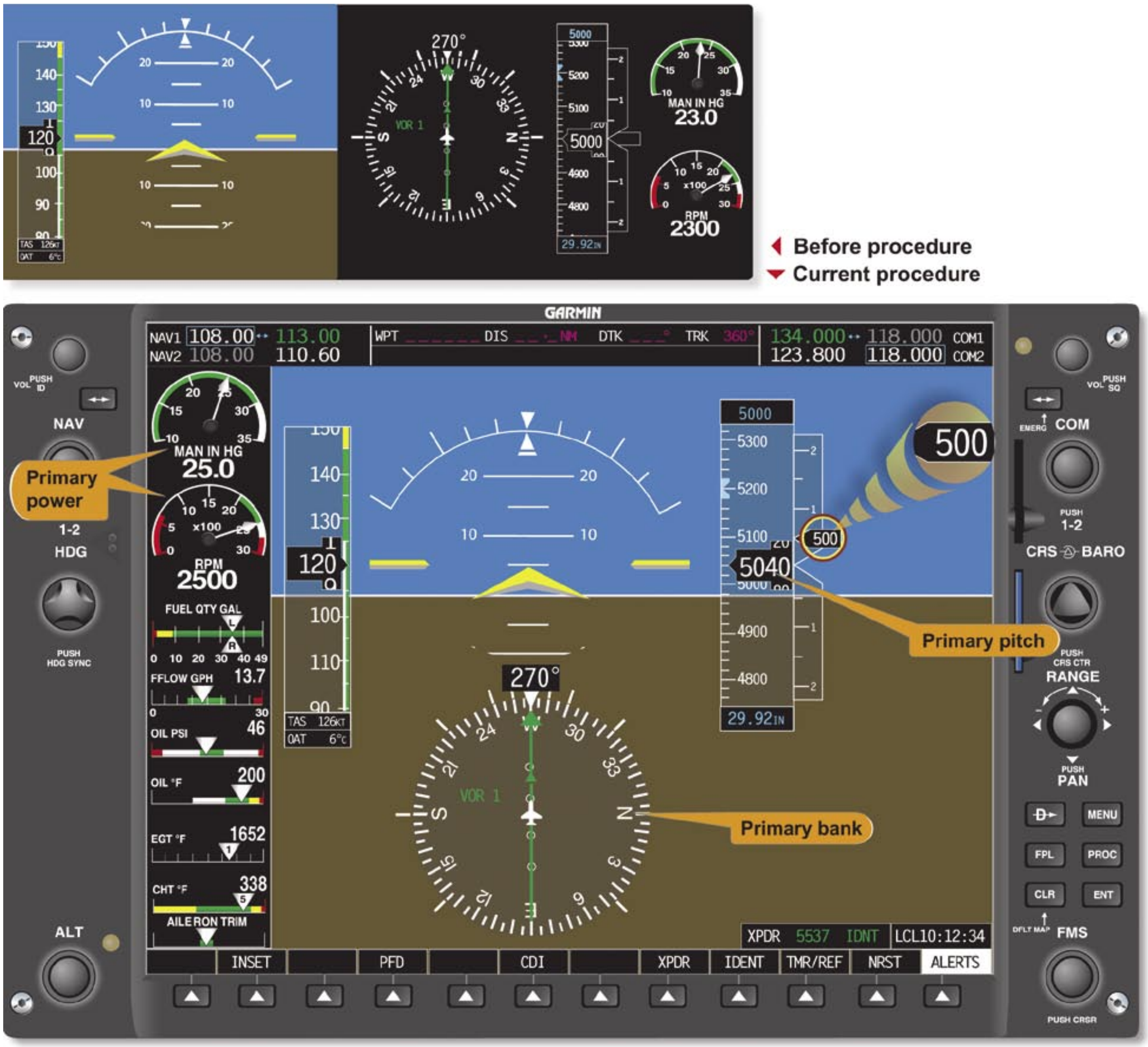

Figure 5-63. *Constant Airspeed Climb From Cruise Airspeed.*

rate of climb equivalent to the pitch and power setting being utilized. Depending on current weight and atmospheric conditions, this rate will be different. This will require the pilot to be knowledgeable of how weight and atmospheric conditions affect aircraft performance.

Once the aircraft is stabilized at a constant airspeed and pitch attitude, the primary flight instrument for pitch will be the ASI and the primary bank instrument will be the heading indicator. The primary power instrument will be the tachometer or the manifold pressure gauge depending on the aircraft type. If the pitch attitude is correct, the airspeed should slowly decrease to the desired speed. If there is any variation in airspeed, make small pitch changes until the aircraft is stabilized at the desired speed. Any change in airspeed will require a trim adjustment.

Constant Airspeed Climb from Established Airspeed

In order to enter a constant airspeed climb, first complete the airspeed reduction from cruise airspeed to climb airspeed. Maintain straight-and-level flight as the airspeed is reduced. The entry to the climb is similar to the entry from cruise airspeed with the exception that the power must be increased when the pitch attitude is raised. *[Figure 5-64]* Power added after the pitch change will show a decrease in airspeed due to the increased drag encountered. Power added prior to a pitch change will cause the airspeed to increase due to the excess thrust.

Constant Rate Climbs

Constant rate climbs are very similar to the constant airspeed climbs in the way the entry is made. As power is added,

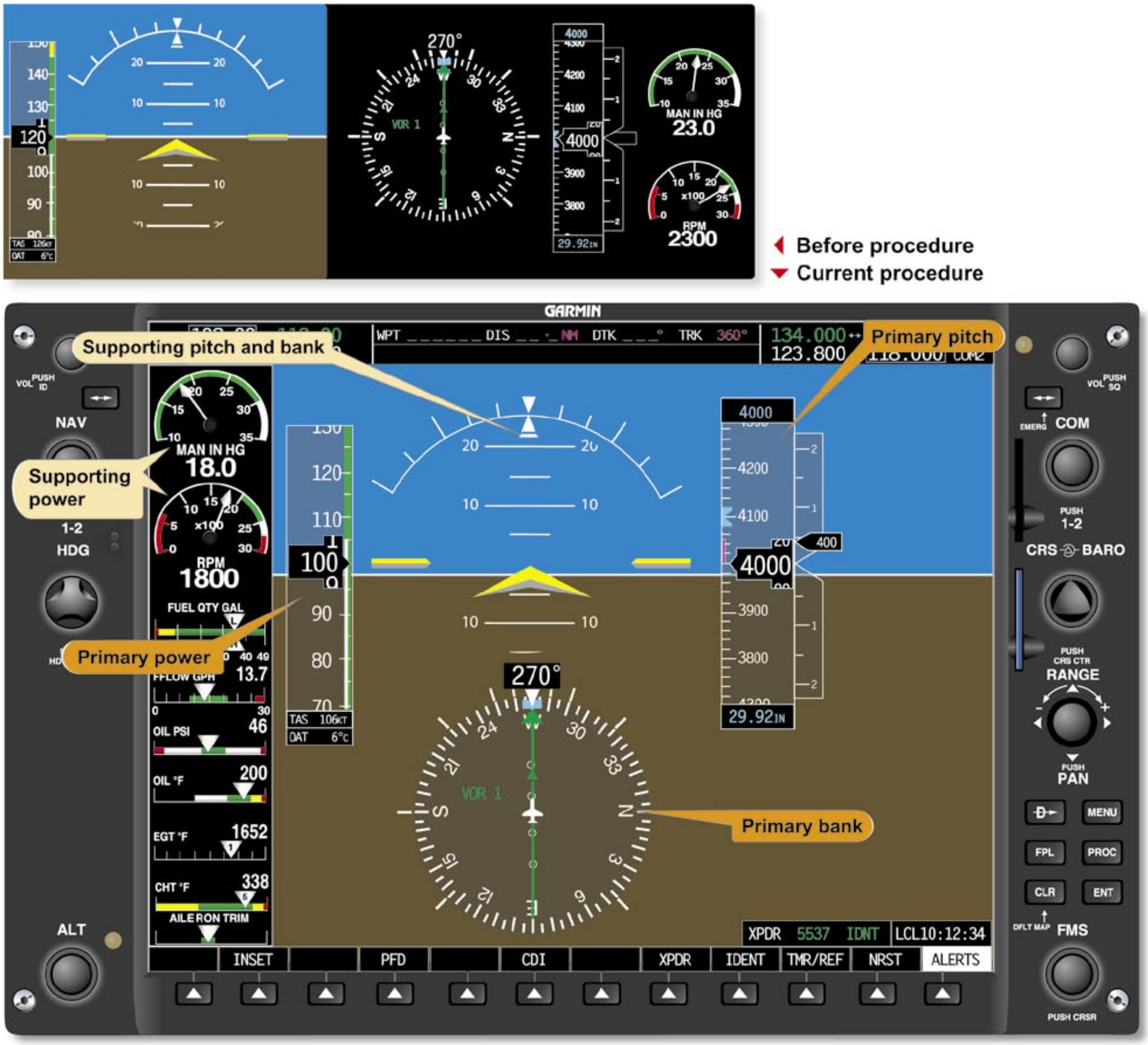

Figure 5-64. *Constant-Airspeed Climb From Established Airspeed.*

smoothly apply elevator pressure to raise the yellow chevron to the desired pitch attitude that equates to the desired vertical speed rate. The primary instrument for pitch during the initial portion of the maneuver is the ASI until the vertical speed rate stabilizes and then the VSI tape becomes primary. The ASI then becomes the primary instrument for power. If any deviation from the desired vertical speed is noted, small pitch changes will be required in order to achieve the desired vertical speed. *[Figure 5-65]*

When making changes to compensate for deviations in performance, pitch, and power, pilot inputs need to be coordinated to maintain a stable flight attitude. For instance, if the vertical speed is lower than desired but the airspeed is correct, an increase in pitch will momentarily increase the vertical speed. However, the increased drag will quickly start to degrade the airspeed if no increase in power is made. A change to any one variable will mandate a coordinated change in the other.

Conversely, if the airspeed is low and the pitch is high, a reduction in the pitch attitude alone may solve the problem. Lower the nose of the aircraft very slightly to see if a power reduction is necessary. Being familiar with the pitch and power settings for the aircraft aids in achieving precise attitude instrument flying.

Leveling Off

Leveling off from a climb requires a reduction in the pitch prior to reaching the desired altitude. If no change in pitch is made until reaching the desired altitude, the momentum of the aircraft causes the aircraft to continue past the desired altitude throughout the transition to a level pitch attitude. The amount of lead to be applied depends on the vertical speed rate. A higher vertical speed requires a larger lead for level off. A good rule of thumb to utilize is to lead the level off by 10 percent of the vertical speed rate (1,000 fpm ÷ 10 = 100 feet lead).

To level off at the desired altitude, refer to the attitude display and apply smooth forward elevator pressure toward the desired level pitch attitude while monitoring the VSI and altimeter tapes. The rates should start to slow and airspeed should begin to increase. Maintain the climb power setting until the airspeed approaches the desired cruise airspeed. Continue to monitor the altimeter to maintain the desired altitude as the airspeed increases. Prior to reaching the cruise airspeed, the power must be reduced to avoid overshooting the desired speed. The amount of lead time that is required depends on the speed at which the aircraft accelerates. Utilization of the airspeed trend indicator can assist by showing how quickly the aircraft will arrive at the desired speed.

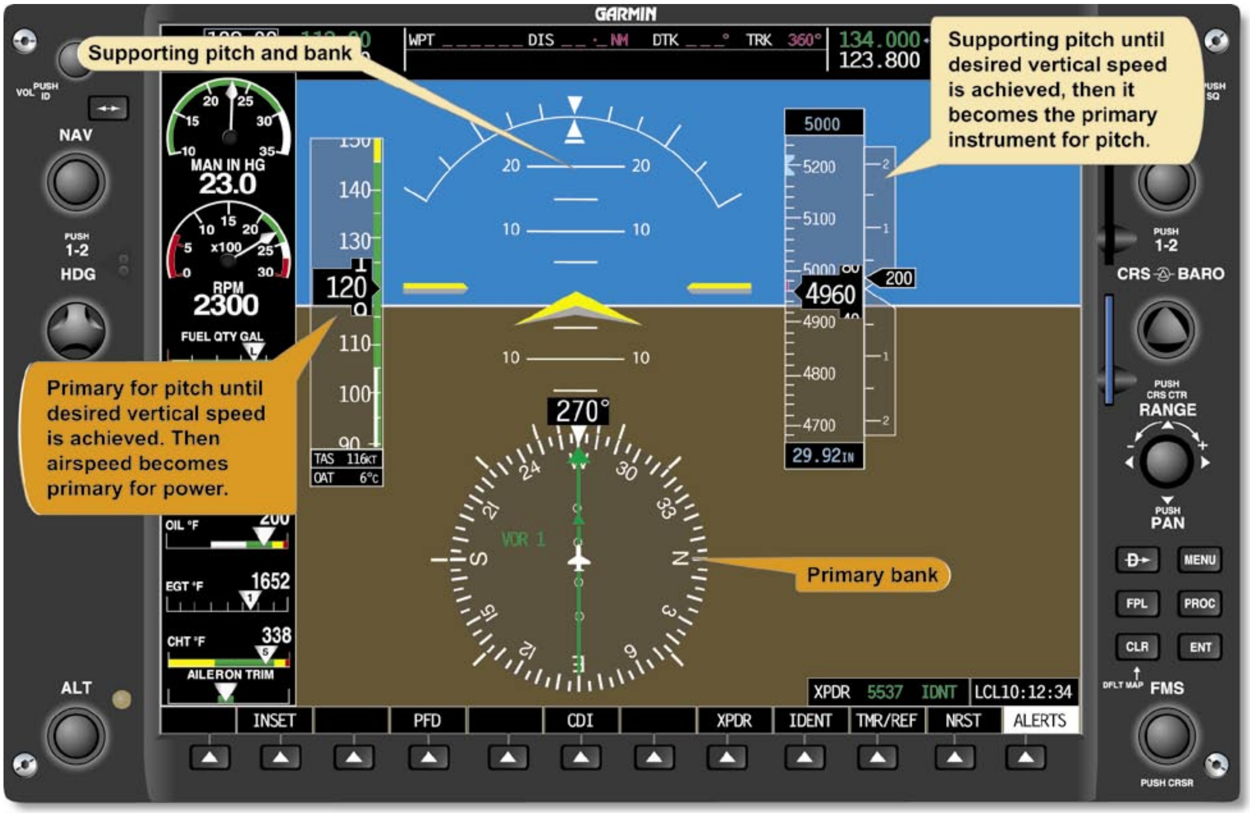
GARMIN
Supporting pitch and bank
Supporting pitch until desired vertical speed is achieved, then it becomes the primary instrument for pitch.
Primary for pitch until desired vertical speed is achieved. Then airspeed becomes primary for power.
Primary bank
134.000
123.800
MAN IN HG
23.0
RPM
2300
120
4960
5000
200
270°
29.92IN
TAS 116KT
OAT 6°C
NAV
HDG
ALT
CRS BARO
RANGE
PAN
FMS
MENU
FPL
PROC
CLR
ENT
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
XPDR 5537 IDNT LCL10:12:34

Figure 5-65. *Constant Rate Climbs.*

To level off at climbing airspeed, lower the nose to the appropriate pitch attitude for level flight with a simultaneous reduction in power to a setting that will maintain the desired speed. With a coordinated reduction in pitch and power there should be no change in the airspeed.

Descents

Descending flight can be accomplished at various airspeeds and pitch attitudes by reducing power, lowering the nose to a pitch attitude lower than the level flight attitude, or adding drag. Once any of these changes have been made, the airspeed will eventually stabilize. During this transitional phase, the only instrument that will display an accurate indication of pitch is the attitude indicator. Without the use of the attitude indicator (such as in partial panel flight), the ASI tape, the VSI tape, and the altimeter tape will show changing values until the aircraft stabilizes at a constant airspeed and constant rate of descent. The altimeter tape continues to show a descent. Hold pitch constant and allow the aircraft to stabilize. During any change in attitude or airspeed, continuous application of trim is required to eliminate any control pressures that need to be applied to the control yoke. An increase in the scan rate during the transition is important since changes are being made to the aircraft flight path and speed. *[Figure 5-66]*

Entry

Descents can be accomplished with a constant rate, constant airspeed or a combination. The following method can accomplish any of these with or without an attitude indicator. Reduce the power to allow the aircraft to decelerate to the desired airspeed while maintaining straight-and-level flight. As the aircraft approaches the desired airspeed, reduce the

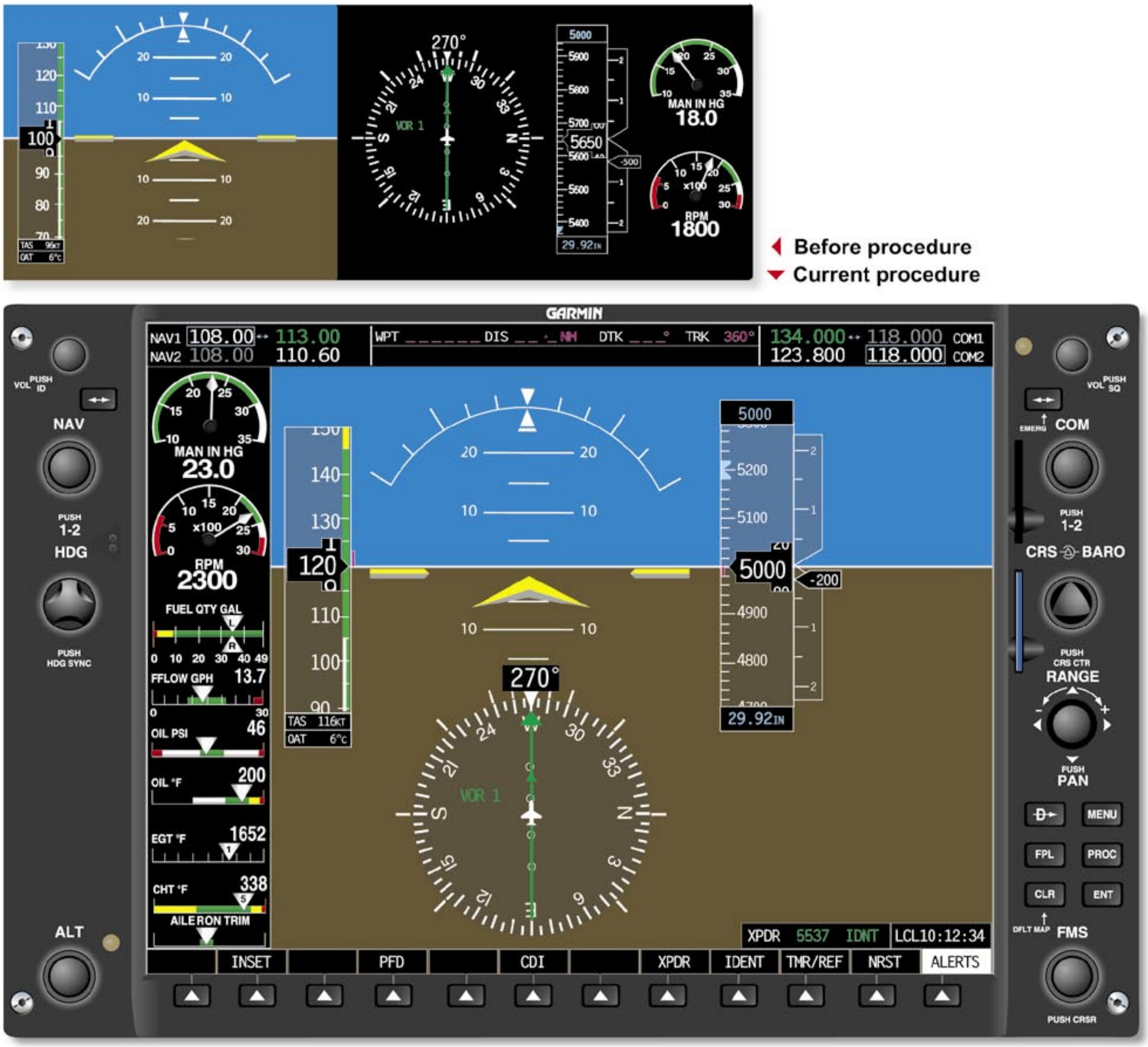
◀ Before procedure
▼ Current procedure
GARMIN
MAN IN HG
18.0
RPM
1800
5650
-500
29.92IN
270°
VOR 1
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT
DIS
DTK
TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
MAN IN HG
23.0
RPM
2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
120
5000
-200
TAS 116KT
OAT 6°C
XPDR 5537 IDNT
LCL10:12:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS

Figure 5-66. *The top image illustrates a reduction of power and descending at 500 fpm to an altitude of 5,000 feet. The bottom image illustrates an increase in power and the initiation of leveling off.*

power to a predetermined value. The airspeed continues to decrease below the desired airspeed unless a simultaneous reduction in pitch is performed. The primary instrument for pitch is the ASI tape. If any deviation from the desired speed is noted, make small pitch corrections by referencing the attitude indicator and validate the changes made with the airspeed tape. Utilize the airspeed trend indicator to judge if the airspeed will be increasing and at what rate. Remember to trim off any control pressures.

The entry procedure for a constant rate descent is the same except the primary instrument for pitch is the VSI tape. The primary instrument for power will be the ASI. When performing a constant rate descent while maintaining a specific airspeed, coordinated use of pitch and power will be required. Any change in pitch directly affects the airspeed. Conversely, any change in airspeed will have a direct impact on vertical speed as long as the pitch is being held constant.

Leveling Off

When leveling off from a descent with the intention of returning to cruise airspeed, first start by increasing the power to cruise prior to increasing the pitch back toward the level flight attitude. A technique used to determine how soon to start the level off is to lead the level off by an altitude corresponding to 10 percent of the rate of descent. For example, if the aircraft is descending at 1,000 fpm, start the level off 100 feet above the level off altitude. If the pitch attitude change is started late, there is a tendency to overshoot the desired altitude unless the pitch change is made with a rapid movement. Avoid making any rapid changes that could lead to control issues or spatial disorientation. Once in level pitch attitude, allow the aircraft to accelerate to the desired speed. Monitor the performance on the airspeed and altitude tapes. Make adjustments to the power in order to correct any deviations in the airspeed. Verify that the aircraft is maintaining level flight by cross-checking the altimeter tape. If deviations are noticed, make an appropriate smooth pitch change in order to arrive back at desired altitude. Any change in pitch requires a smooth coordinated change to the power setting. Monitor the airspeed in order to maintain the desired cruise airspeed.

To level off at a constant airspeed, the pilot must again determine when to start to increase the pitch attitude toward the level attitude. If pitch is the only item that is changing, airspeed varies due to the increase in drag as the aircraft's pitch increases. A smooth coordinated increase in power will need to be made to a predetermined value in order to maintain speed. Trim the aircraft to relieve any control pressure that may have to be applied.

Common Errors in Straight Climbs and Descents

Climbing and descending errors usually result from but are not limited to the following errors:

1. Overcontrolling pitch on beginning the climb. Aircraft familiarization is the key to achieving precise attitude instrument flying. Until the pilot becomes familiar with the pitch attitudes associated with specific airspeeds, the pilot must make corrections to the initial pitch settings. Changes do not produce instantaneous and stabilized results; patience must be maintained while the new speeds and vertical speed rates stabilize. Avoid the temptations to make a change and then rush into making another change until the first one is validated. Small changes will produce more expeditious results and allow for a more stabilized flight path. Large changes to pitch and power are more difficult to control and can further complicate the recovery process.
2. Failure to increase the rate of instrument cross-check. Any time a pitch or power change is made, an increase in the rate a pilot cross-checks the instrument is required. A slow cross-check can lead to deviations in other flight attitudes.
3. Failure to maintain new pitch attitudes. Once a pitch change is made to correct for a deviation, that pitch attitude must be maintained until the change is validated. Utilize trim to assist in maintaining the new pitch attitude. If the pitch is allowed to change, it is impossible to validate whether the initial pitch change was sufficient to correct the deviation. The continuous changing of the pitch attitude delays the recovery process.
4. Failure to utilize effective trim techniques. If control pressures have to be held by the pilot, validation of the initial correction will be impossible if the pitch is allowed to vary. Pilots have the tendency to either apply or relax additional control pressures when manually holding pitch attitudes. Trim allows the pilot to fly without holding pressure on the control yoke.
5. Failure to learn and utilize proper power settings. Any time a pilot is not familiar with an aircraft's specific pitch and power settings, or does not utilize them, a change in flight paths will take longer. Learn pitch and power settings in order to expedite changing the flight path.
6. Failure to cross-check both airspeed and vertical speed prior to making adjustments to pitch and or power. It is possible that a change in one may correct a deviation in the other.

7. Uncoordinated use of pitch and power during level offs. During level offs, both pitch and power settings need to be made in unison in order to achieve the desired results. If pitch is increased before adding power, additional drag will be generated thereby reducing airspeed below the desired value.
8. Failure to utilize supporting pitch instruments which will lead to chasing the VSI. Always utilize the attitude indicator as the control instrument on which to change the pitch.
9. Failure to determine a proper lead time for level off from a climb or descent. Waiting too long can lead to overshooting the altitude.
10. Ballooning—Failure to maintain forward control pressure during level off as power is increased. Additional lift is generated causing the nose of the aircraft to pitch up.

Turns

Standard Rate Turns

The previous sections have addressed flying straight-and-level as well as climbs and descents. However, attitude instrument flying is not accomplished solely by flying in a straight line. At some point, the aircraft will need to be turned to maneuver along victor airways, global positioning system (GPS) courses, and instrument approaches. The key to instrument flying is smooth, controlled changes to pitch and bank. Instrument flying should be a slow but deliberate process that takes the pilot from departure airport to destination airport without any radical flight maneuvers.

A turn to specific heading should be made at standard rate. Standard rate is defined as a turning rate of 3° per second which will yield a complete 360° turn in 2 minutes. A turning rate of 3° per second will allow for a timely heading change, as well as allowing the pilot sufficient time to cross-check the flight instruments and avoid drastic changes to the aerodynamic forces being exerted on the aircraft. At no time should the aircraft be maneuvered faster than the pilot is comfortable cross-checking the flight instruments. Most autopilots are programmed to turn at standard rate.

Establishing A Standard Rate Turn

In order to initiate a standard rate turn, approximate the bank angle and then establish that bank angle on the attitude indicator. A rule of thumb to determine the approximate angle of bank is to use 15 percent of the true airspeed. A simple way to determine this amount is to divide the airspeed by 10 and add one-half the result. For example, at 100 knots, approximately 15° of bank is required (100/10 = 10 + 5 = 15); at 120 knots, approximately 18° of bank is needed for a standard-rate turn. Cross-check the turn rate indicator, located on the HSI, to determine if that bank angle is sufficient to deliver a standard rate turn. Slight modifications may need to be made to the bank angle in order to achieve the desired performance. The primary bank instrument in this case is the turn rate indicator since the goal is to achieve a standard rate turn. The turn rate indicator is the only instrument that can specifically indicate a standard rate turn. The attitude indicator is used only to establish a bank angle (control instrument) but can be utilized as a supporting instrument by cross-checking the bank angle to determine if the bank is greater or less than what was calculated.

As the aircraft rolls into the bank, the vertical component of lift will begin to decrease. *[Figure 5-67]* As this happens, additional lift must be generated to maintain level flight. Apply aft control pressure on the yoke sufficient to stop any altitude loss trend. With the increase in lift that needs to be generated, additional induced drag will also be generated. This additional drag will cause the aircraft to start to decelerate. To counteract this, apply additional thrust by adding power to the power lever. Once altitude and airspeed is being maintained, utilize the trim wheel to eliminate any control forces that need to be held on the control column.

When rolling out from a standard rate turn, the pilot needs to utilize coordinated aileron and rudder and roll-out to a wings level attitude utilizing smooth control inputs. The roll-out rate should be the same as the roll-in rate in order to estimate the lead necessary to arrive at the desired heading without over- or undershooting.

During the transition from the turn back to straight flight, the attitude indicator becomes the primary instrument for bank. Once the wings are level, the heading indicator becomes the primary instrument for bank. As bank decreases, the vertical component increases if the pitch attitude is not decreased sufficiently to maintain level flight. An aggressive cross-check keeps the altimeter stationary if forward control pressure is applied to the control column. As the bank angle is decreased, the pitch attitude should be decreased accordingly in order to arrive at the level pitch attitude when the aircraft reaches zero bank. Remember to utilize the trim wheel to eliminate any excess control forces that would otherwise need to be held.

Common Errors

1. One common error associated with standard rate turns is due to pilot inability to hold the appropriate bank angle that equates to a standard rate. The primary bank instrument during the turn is the turn rate indicator; however the bank angle varies slightly. With an

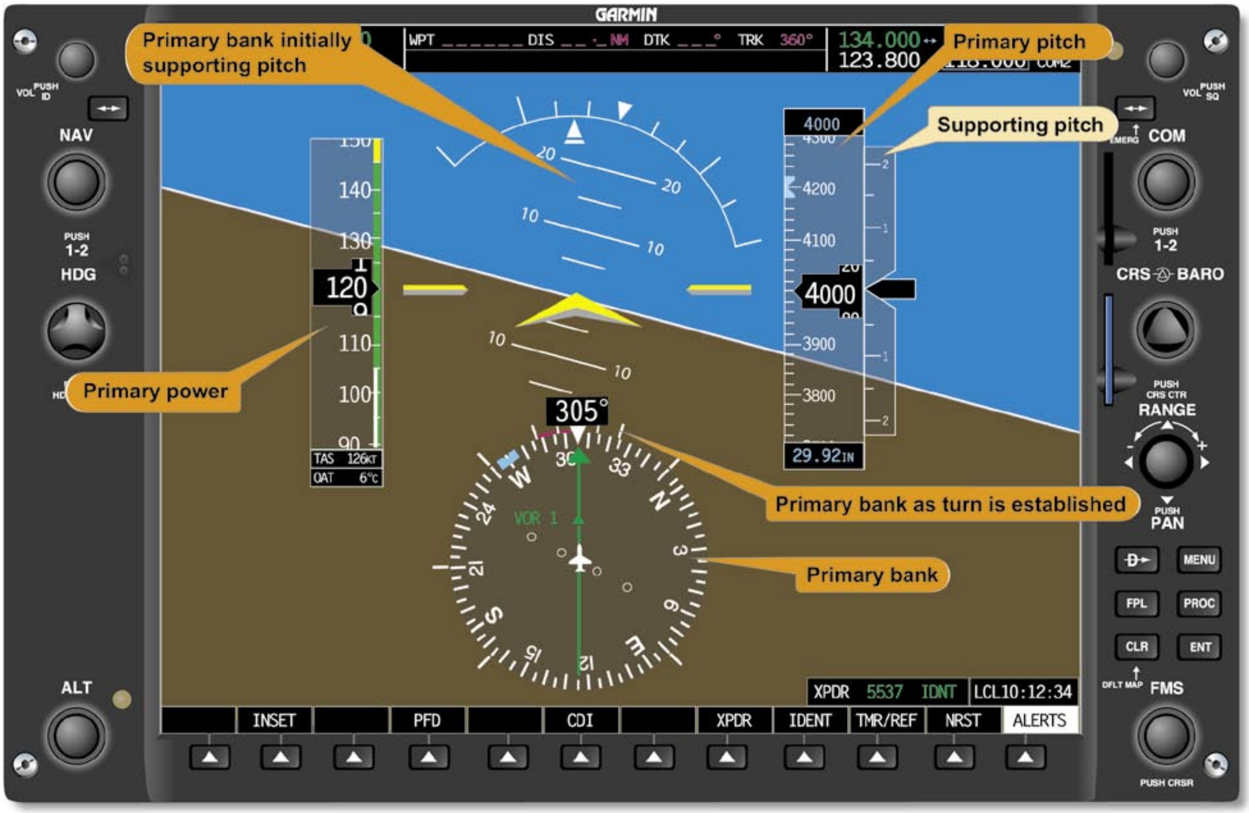
GARMIN
Primary bank initially supporting pitch
Primary pitch
Supporting pitch
Primary power
Primary bank as turn is established
Primary bank
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS
134.000
123.800
4000
120
305°
29.92IN
XPDR 5537 IDNT LCL10:12:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 5-67. *Standard Rate Turn—Constant Airspeed.*

aggressive cross-check, a pilot should be able to minimize errors arising from over- or underbanking.

2. Another error normally encountered during standard rate turns is inefficient or lack of adequate cross-checking. Pilots need to establish an aggressive cross-check in order to detect and eliminate all deviations from altitude, airspeed, and bank angle during a maneuver.
3. Fixation is a major error associated with attitude instrument flying in general. Pilots training for their instrument rating tend to focus on what they perceive to be the most important task at hand and abandon their cross-check by applying all of their attention to the turn rate indicator. A modified radial scan works well to provide the pilot with adequate scanning of all instrumentation during the maneuver.

Turns to Predetermined Headings

Turning the aircraft is one of the most basic maneuvers that a pilot learns during initial flight training. Learning to control the aircraft, maintaining coordination, and smoothly rolling out on a desired heading are all keys to proficient attitude instrument flying.

EFDs allow the pilot to better utilize all instrumentation during all phases of attitude instrument flying by consolidating all traditional instrumentation onto the PFD. The increased size of the attitude indicator, which stretches the entire width of the PFD, allows the pilot to maintain better pitch control while the introduction of the turn rate indicator positioned directly on the compass rose aids the pilot in determining when to begin a roll-out for the desired heading.

When determining what bank angle to utilize when making a heading change, a general rule states that for a small heading change, do not use a bank angle that is greater than the total number of degrees of change needed. For instance, if a heading change of 20° is needed, a bank angle of not more than 20° is required. Another rule of thumb that better defines the bank angle is half the total number of degrees of heading change required, but never greater than standard rate. The exact bank angle that equates to a standard rate turn varies due to true airspeed.

With this in mind and the angle of bank calculated, the next step is determining when to start the roll-out process. For example:

An aircraft begins a turn from a heading of 030° to a heading of 120°. With the given airspeed, a standard rate turn has yielded a 15° bank. The pilot wants to begin a smooth coordinated roll-out to the desired heading when the heading indicator displays approximately 112°. The necessary calculations are:

15° bank (standard rate) ÷ 2 = 7.5°
120° – 7.5° = 112.5°

By utilizing this technique the pilot is better able to judge if any modifications need to be made to the amount of lead once the amount of over- or undershooting is established.

Timed Turns

Timed turns to headings are performed in the same fashion with an EFD as with an analog equipped aircraft. The instrumentation used to perform this maneuver is the turn rate indicator as well as the clock. The purpose of this maneuver is to allow the pilot to gain proficiency in scanning as well as to further develop the pilot's ability to control the aircraft without standard instrumentation.

Timed turns become essential when controlling the aircraft with a loss of the heading indicator. This may become necessary due to a loss of the AHRS unit or the magnetometer. In any case, the magnetic compass will still be available for navigation. The reason for timed turns instead of magnetic compass turns is the simplicity of the maneuver. Magnetic compass turns require the pilot to take into account various errors associated with the compass; timed turns do not.

Prior to initiating a turn, determine if the standard rate indication on the turn rate indicator will actually deliver a 3° per second turn. To accomplish this, a calibration must be made. Establish a turn in either direction at the indicated standard rate. Start the digital timer as the compass rolls past a cardinal heading. Stop the timer once the compass card rolls through another cardinal heading. Roll wings level and compute the rate of turn. If the turn rate indicator is calibrated and indicating correctly, 90° of heading change should take 30 seconds. If the time taken to change heading by 90° is more or less than 30 seconds, then a deflection above or below the standard rate line needs to be made to compensate for the difference. Once the calibration has been completed in one direction, proceed to the opposite direction. When both directions have been calibrated, apply the calibrated calculations to all timed turns.

In order to accomplish a timed turn, the amount of heading change needs to be established. For a change in heading from 120° to a heading of 360°, the pilot calculates the difference and divides that number by 3. In this case, 120° divided by 3° per second equals 40 seconds. This means that it would take 40 seconds for an aircraft to change heading 120° if that aircraft were held in a perfect standard rate turn. Timing for the maneuver should start as the aircraft begins rolling into the standard rate turn. Monitor all flight instruments during this maneuver. The primary pitch instrument is the altimeter. The primary power instrument is the ASI and the primary bank instrument is the turn rate indicator.

Once the calculated time expires, start a smooth coordinated roll-out. As long as the pilot utilizes the same rate of roll-in as roll-out, the time it takes for both will not need to be included in the calculations. With practice the pilot should level the wings on the desired heading. If any deviation has occurred, make small corrections to establish the correct heading.

Compass Turns

The magnetic compass is the only instrument that requires no other source of power for operation. In the event of an AHRS or magnetometer failure, the magnetic compass is the instrument the pilot uses to determine aircraft heading. For a more detailed explanation on the use of the magnetic compass, see page 5-21.

Steep Turns

For the purpose of instrument flight training, a steep turn is defined as any turn in excess of standard rate. A standard rate turn is defined as 3° per second. The bank angle that equates to a turn rate of 3° per second varies according to airspeed. As airspeed increases, the bank angle must be increased. The exact bank angle that equates to a standard rate turn is unimportant. Normal standard rate turn bank angles range from 10° to 20°. The goal of training in steep turn maneuvers is pilot proficiency in controlling the aircraft with excessive bank angles.

Training in excessive bank angles will challenge the pilot in honing cross-checking skills and improve altitude control throughout a wider range of flight attitudes. Although the current instrument flight check practical test standards (PTS) do not call for a demonstration of steep turns on the certification check flight, this does not eliminate the need for the instrument pilot-in-training to demonstrate proficiency to an instructor.

Training in steep turns teaches the pilot to recognize and to adapt to rapidly changing aerodynamic forces that necessitate an increase in the rate of cross-checking all flight instruments. The procedures for entering, maintaining and exiting a steep turn are the same as for shallower turns. Proficiency in instrument cross-check and interpretation is increased due to the higher aerodynamic forces and increased speed at which the forces are changing.

Performing the Maneuver

To enter a steep turn to the left, roll into a coordinated 45° bank turn to the left. An advantage that glass panel displays have over analog instrumentation is a 45° bank indication on the roll scale. This additional index on the roll scale allows the pilot to precisely roll into the desired bank angle instead of having to approximate it as is necessary with analog instrumentation. *[Figure 5-68]*

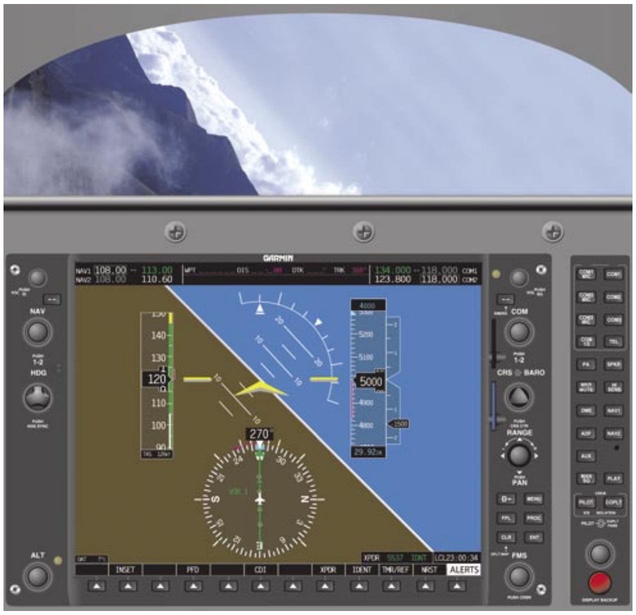

Figure 5-68. *Steep Left Turn.*

As soon as the bank angle increases from level flight, the vertical component of lift begins to decrease. If the vertical component of lift is allowed to continue to decrease, a pronounced loss of altitude is indicated on the altimeter along with the VSI tape, as well as the altitude trend indicator. Additionally, the airspeed will begin to increase due to the lowered pitch attitude. It is very important to have a comprehensive scan developed prior to training in steep turns. Utilization of all of the trend indicators, as well the VSI, altimeter, and ASI, is essential in learning to fly steep turns by reference to instruments alone.

In order to avoid a loss of altitude, the pilot begins to slowly increase back pressure on the control yoke in order to increase the pitch attitude. The pitch change required is usually no more than 3° to 5°, depending on the type of aircraft. As the pilot increases back pressure, the angle of attack increases, thus increasing the vertical component of lift. When a deviation in altitude is indicated, proper control force corrections need to be made. During initial training of steep turns, pilots have a tendency to overbank. Over banking is when the bank angle exceeds 50°. As the outboard wing begins to travel faster through the air it will begin to generate a greater and greater differential in lift compared to the inboard wing. As the bank angle continues to progress more and more steeply past 45°, the two components of lift (vertical and horizontal) become inversely proportionate.

Once the angle has exceeded 45°, the horizontal component of lift is now the greater force. If altitude should continue to decrease and the pilot only applies back yoke pressure, the aircraft's turn radius begins to tighten due to the increased horizontal force. If aft control pressure continues to increase, there will come a point where the loss of the vertical component of lift and aerodynamic wing loading prohibits the nose of the aircraft from being raised. Any increase in pitch only tightens the turning radius.

The key to successfully performing a steep turn by reference to instruments alone is the thorough understanding of the aerodynamics involved, as well as a quick and reliable cross-check. The pilot should utilize the trim to avoid holding control forces for any period of time. With time and practice, a flight instructor can demonstrate how to successfully fly steep turns with and without the use of trim. Once the aircraft is trimmed for the maneuver, accomplishing the maneuver will be virtually a hands-off effort. This allows additional time for cross-checking and interpreting the instruments.

It is imperative when correcting for a deviation in altitude, that the pilot modify the bank angle ±5° in order to vary the vertical component of lift, not just adjust back pressure. These two actions should be accomplished simultaneously.

During the recovery from steep turns to straight-and-level flight, aft control forces must be varied with the power control to arrive back at entry altitude, heading and airspeed.

Steps:

1. Perform clearing turns.
2. Roll left into a 45° bank turn and immediately begin to increase the pitch attitude by approximately 3° to 5°.
3. As the bank rolls past 30°, increase power to maintain the entry airspeed.
4. Apply trim to eliminate any aft control wheel forces.
5. Begin rolling out of the steep turn approximately 20° prior to the desired heading.
6. Apply forward control pressure and place the pitch attitude in the level cruise pitch attitude.
7. Reduce power to the entry power setting to maintain the desired airspeed.
8. Re-trim the aircraft as soon as practical or continue into a right hand steep turn and continue from step 3.

9. Once the maneuver is complete, establish cruise flight and accomplish all appropriate checklist items.

Unusual Attitude Recovery Protection

Unusual attitudes are some of the most hazardous situations for a pilot to be in. Without proper recovery training on instrument interpretation and aircraft control, a pilot can quickly aggravate an abnormal flight attitude into a potentially fatal accident.

Analog gauges require the pilot to scan between instruments to deduce the aircraft attitude. Individually, these gauges lack the necessary information needed for a successful recovery.

EFDs have additional features to aid in recognition and recovery from unusual flight attitudes. The PFD displays all the flight instruments on one screen. Each instrument is superimposed over a full-screen representation of the attitude indicator. With this configuration, the pilot no longer needs to transition from one instrument to another.

The new unusual attitude recovery protection allows the pilot to be able to quickly determine the aircraft's attitude and make a safe, proper and prompt recovery. Situational awareness is increased by the introduction of the large full-width artificial horizon depicted on the PFD. This now allows for the attitude indicator to be in view during all portions of the scan.

One problem with analog gauges is that the attitude indicator displays a complete blue or brown segment when the pitch attitude is increased toward 90° nose-up or nose-down.

With the EFDs, the attitude indicator is designed to retain a portion of both sky and land representation at all times. This improvement allows the pilot to always know the quickest way to return to the horizon. Situational awareness is greatly increased.

NOTE: The horizon line starts moving downward at approximately 47° pitch up. From this point on, the brown segment will remain visible to show the pilot the quickest way to return to the level pitch attitude. *[Figure 5-69]*

NOTE: The horizon line starts moving upward at approximately 27° pitch down. From this point on, the blue segment will remain visible to show the pilot the quickest way to return to the level pitch attitude. *[Figure 5-70]*

It is imperative to understand that the white line on the attitude indicator is the horizon line. The break between the blue and brown symbols is only a reference and should not be thought of as the artificial horizon.

Another important advancement is the development of the unusual attitude recovery protection that is built into the PFD software and made capable by the AHRS. In the case of a nose-high unusual attitude, the unusual attitude recovery protection displays red chevrons which point back to the horizon line.

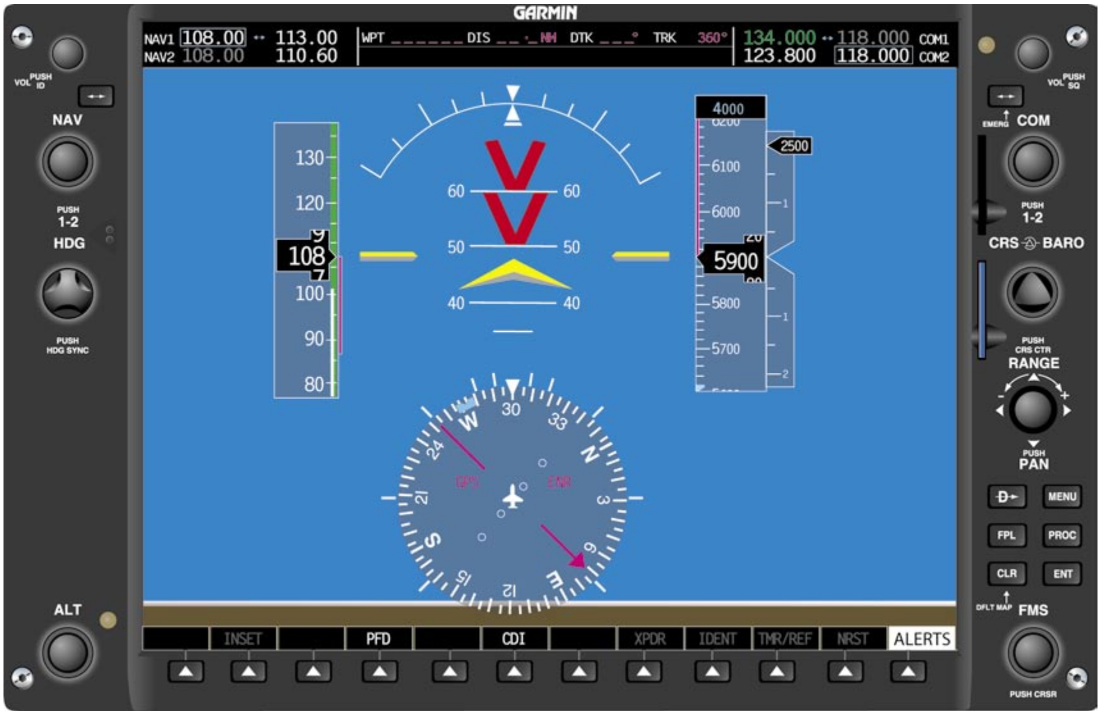
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
130
120
108
100
90
80
60
50
40
4000
2500
6100
6000
5900
5800
5700
30
33
N
3
6
E
12
15
S
21
24
W
GPS
ENR
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR

Figure 5-69. *Unusual Attitude Recovery Protection. Note the brown horizon line is visible at the bottom.*

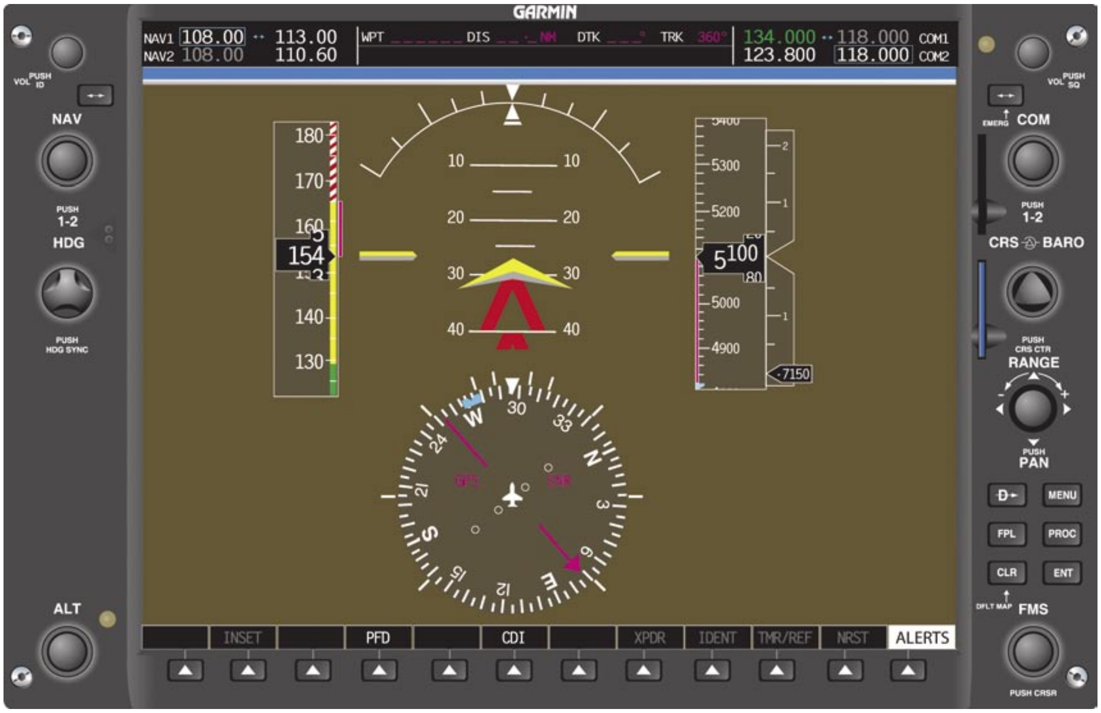

Figure 5-70. *Horizon line starts moving upward at 27°. Note that the blue sky remains visible at 17° nose-down.*

These chevrons are positioned at 50° up on the attitude indicator. The chevrons appear when the aircraft approaches a nose-high attitude of 30°. The software automatically de-clutters the PFD leaving only airspeed, heading, attitude, altimeter, VSI tape, and the trend vectors. The de-cluttered information reappears when the pitch attitude falls below 25°.

For nose-low unusual attitudes, the chevrons are displayed when the pitch exceeds 15° nose-down. If the pitch continues to decrease, the unusual attitude recovery protection de-clutters the screen at 20° nose-down. The de-cluttered information reappears when the pitch increases above 15°.

Additionally, there are bank limits that trigger the unusual attitude protection. If the aircraft's bank increases beyond 60°, a continuation of the roll index occurs to indicate the shortest direction to roll the wings back to level. At 65°, the PFD de-clutters. All information reappears when the bank decreases below 60°.

In *Figure 5-71*, the aircraft has rolled past 60°. Observe the white line that continues from the end of the bank index. This line appears to indicate the shortest distance back to wings level.

When experiencing a failure of the AHRS unit, all unusual attitude protection is lost. The failure of the AHRS results in the loss of all heading and attitude indications on the PFD. In addition, all modes of the autopilot, except for roll and altitude hold, are lost.

The following picture series represents how important this technology is in increasing situational awareness, and how critical it is in improving safety.

Figure 5-72 shows the unusual attitude protection with valid AHRS and air data computer (ADC) inputs. The bright red chevrons pointing down to the horizon indicate a nose-high unusual attitude that can be easily recognized and corrected.

NOTE: The red chevrons point back to the level pitch attitude. The trend indicators show where the airspeed and altitude will be in 6 seconds. The trend indicator on the heading indicator shows which direction the aircraft is turning. The slip/skid indicator clearly shows if the aircraft is coordinated. This information helps the pilot determine which type of unusual attitude the aircraft has taken.

Now look at *Figure 5-73*. The display shows the same airspeed as the picture above; however, the AHRS unit has failed. The altimeter and the VSI tape are the only clear indications that the aircraft is in a nose-high attitude. The one key instrument that is no longer present is the slip/skid indicator. There is not a standby turn coordinator installed in the aircraft for the pilot to reference.

The magnetic compass indicates a heading is being maintained; however, it is not as useful as a turn coordinator or slip/skid indicator.

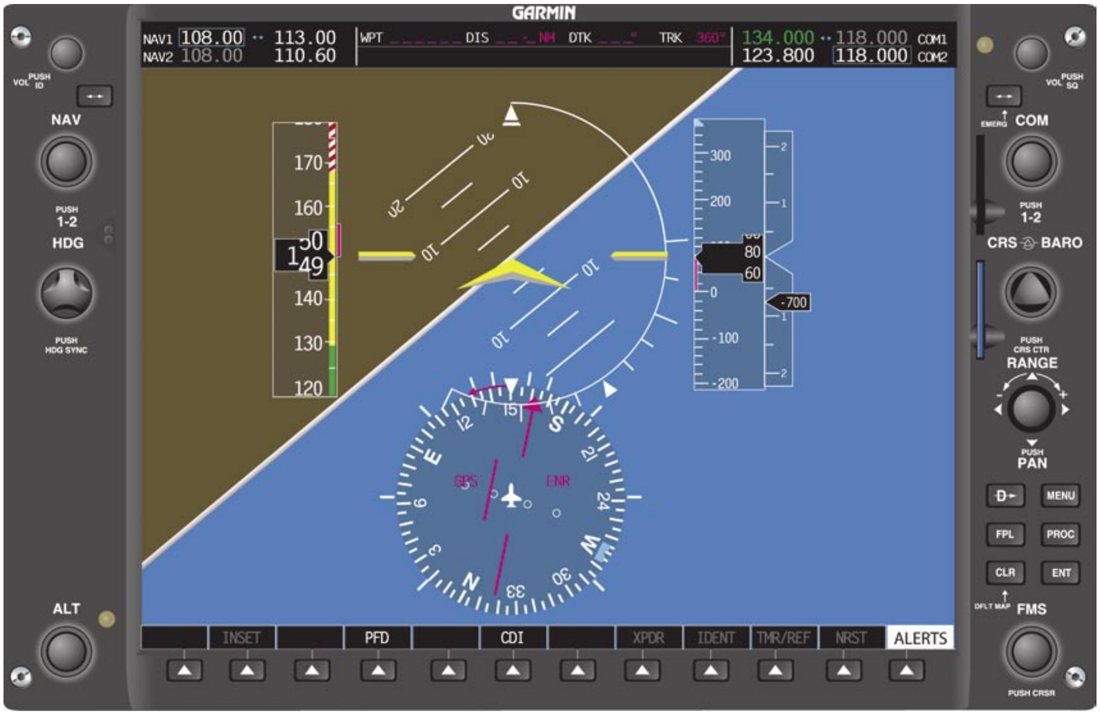
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT _ _ _ _ _ _ DIS _ _ . _ NM DTK _ _ _° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP FMS
PUSH CRSR
GPS
ENR
-700
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 5-71. *Aircraft Rolled Past 60°.*

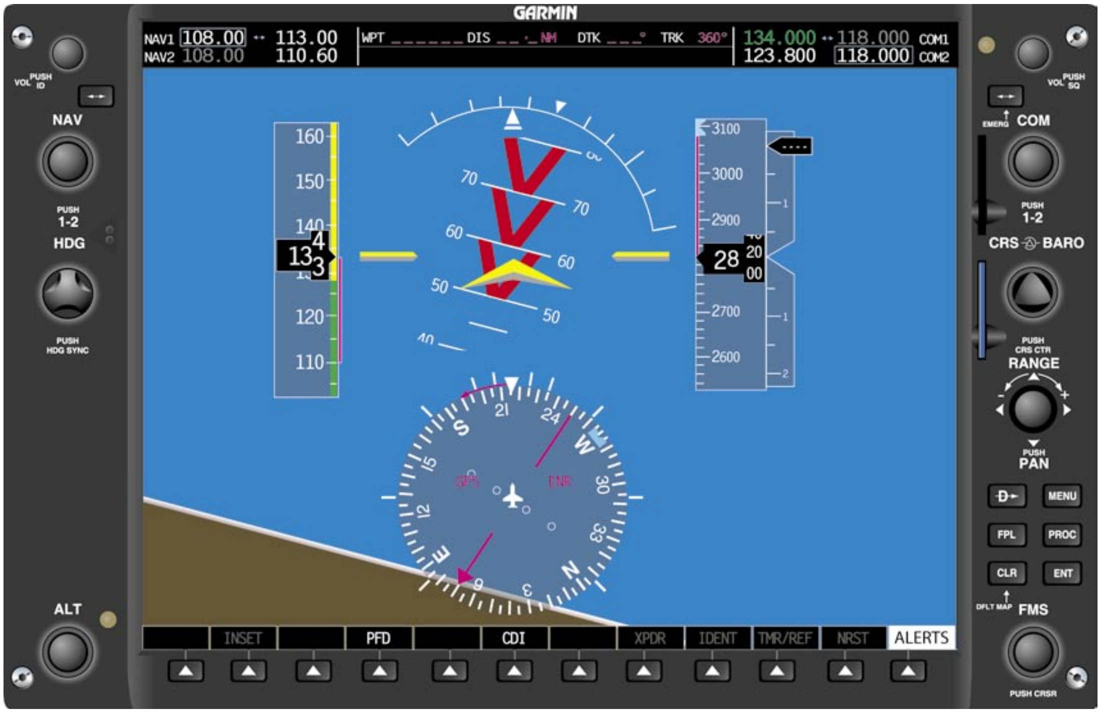
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT _ _ _ _ _ _ DIS _ _ . _ NM DTK _ _ _° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP FMS
PUSH CRSR
GPS
ENR
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 5-72. *Unusual Attitude Protection With Valid AHRS.*

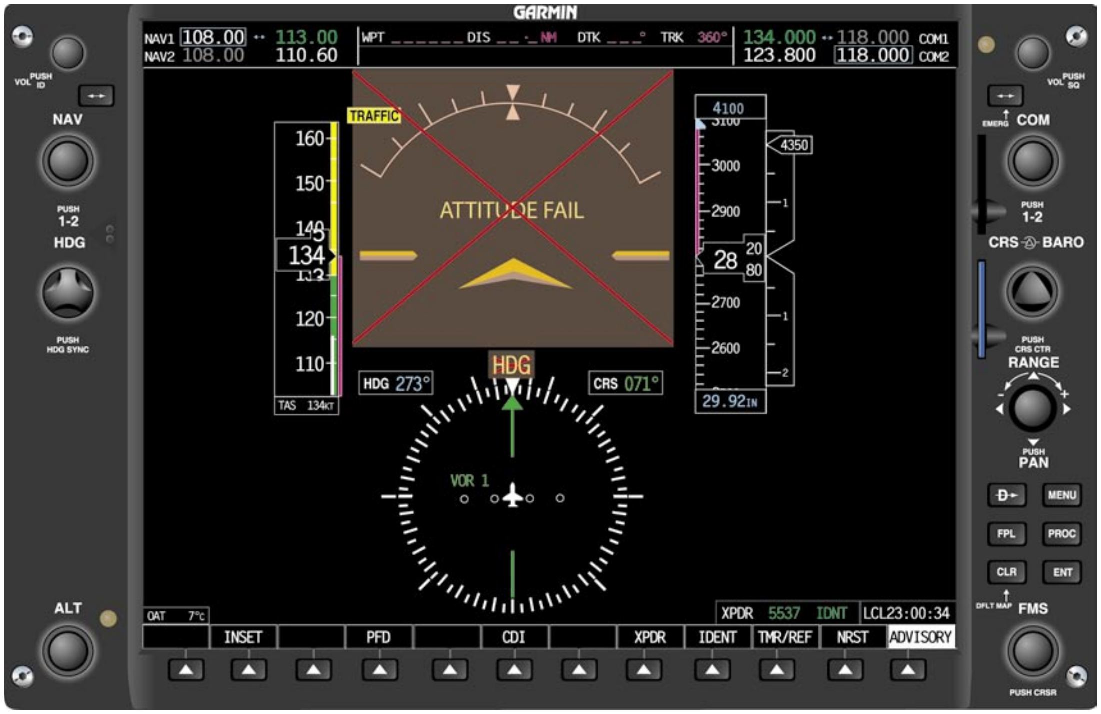
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT ______ DIS __._ NM DTK ___° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
TRAFFIC
ATTITUDE FAIL
160
150
140
134
120
110
TAS 134KT
4100
4350
3000
2900
28
2700
2600
29.92IN
HDG 273°
HDG
CRS 071°
VOR 1
OAT 7°C
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ADVISORY
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP FMS
PUSH CRSR

Figure 5-73. *AHRS Unit Failed.*

Figure 5-74 depicts an AHRS and ADC failure. In this failure scenario, there are no indications of the aircraft's attitude. The manufacturer recommends turning on the autopilot which is simply a wing leveler.

With a failure of the primary instrumentation on the PFD, the only references available are the stand-by instruments. The standby instrumentation consists of an analog ASI, attitude indicator, altimeter, and magnetic compass. There is no standby turn coordinator installed.

In extreme nose-high or nose-low pitch attitudes, as well as high bank angles, the analog attitude indicator has the potential to tumble, rendering it unusable.

Autopilot Usage

The autopilot is equipped with inputs from a turn coordinator installed behind the MFD screen. This turn coordinator is installed solely for the use of the autopilot to facilitate the roll mode. Roll mode, which is simply a wing leveler. This protection will always be available, barring a failure of the turn coordinator (to aid the pilot if the aircraft attains an unusual attitude).

NOTE: The pilot is not able to gain access to the turn coordinator. This instrument is installed behind the MFD panel. *[Figure 5-75]*

Most EFD equipped aircraft are coming from the factory with autopilots installed. However, the purchaser of the aircraft can specify if an autopilot is to be installed. Extreme caution should be utilized when flying an EFD equipped aircraft without an autopilot in IMC with an AHRS and ADC failure.

The autopilot should be utilized to reduce workload, which affords the pilot more time to monitor the flight. Utilization of the autopilot also decreases the chances of entry into an unusual attitude.

Flying an EFD equipped aircraft without the use of an autopilot has been shown to increase workload and decrease situational awareness for pilots first learning to flying the new system.

Common Errors Leading to Unusual Attitudes

The following errors have the potential to disrupt a pilot's situational awareness and lead to unusual attitudes.

1. Improper trimming techniques. A failure to keep the aircraft trimmed for level flight at all times can turn a momentary distraction into an emergency situation if the pilot stops cross-checking.
2. Poor crew resource management (CRM) skills. Failure to perform all single-pilot resource management duties efficiently. A major cause of CRM related accidents comes from the failure of the pilot to maintain an organized flight deck. Items that are being utilized for the flight portion should be neatly arranged for easy access. A disorganized flight deck

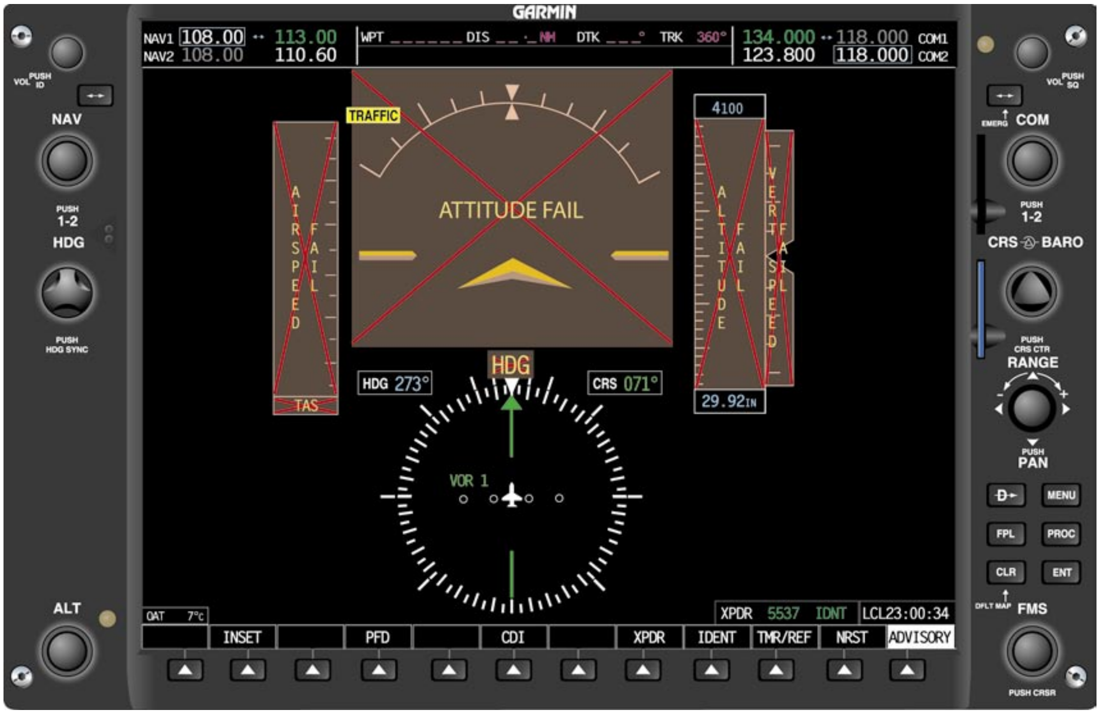
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT ______ DIS __._NM DTK ___° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
TRAFFIC
ATTITUDE FAIL
AIRSPEED FAIL
TAS
ALTITUDE FAIL
VERT SPEED FAIL
4100
29.92IN
HDG 273°
HDG
CRS 071°
VOR 1
OAT 7°C
XPDR 5537 IDNT LCL23:00:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ADVISORY
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP FMS
PUSH CRSR

Figure 5-74. *AHRS ADC Failure.*

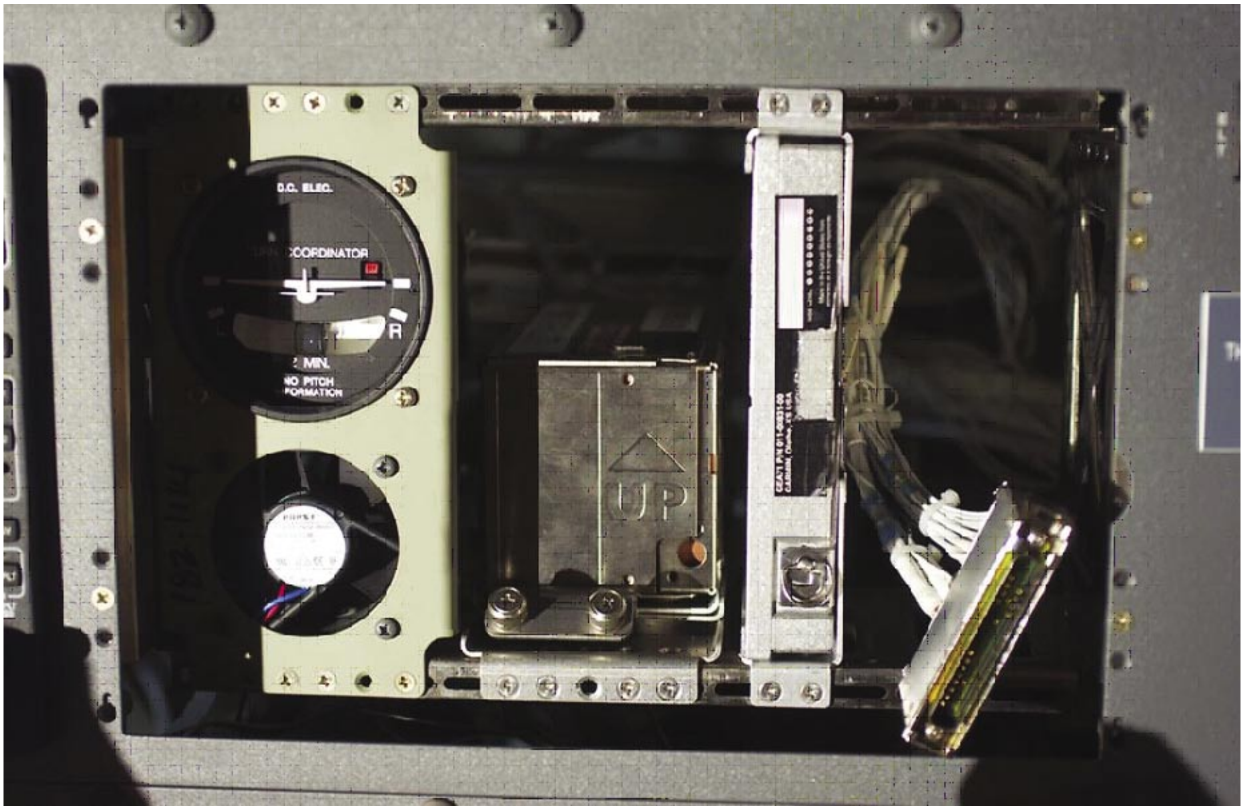
D.C. ELEC.
COORDINATOR
R
2 MIN.
NO PITCH
FORMATION
UP

Figure 5-75. *This autopilot requires roll information from a turn coordinator.*

can lead to a distraction that causes the pilot to cease cross-checking the instruments long enough to enter an unusual attitude.

3. Fixation is displayed when a pilot focuses far too much attention on one instrument because he or she perceives something is wrong or a deviation is occurring. It is important for the instrument pilot to remember that a cross-check of several instruments for corroboration is more valuable than checking a single instrument.
4. Attempting to recover by sensory sensations other than sight. Recovery by instinct almost always leads to erroneous corrections due to the illusions that are prevalent during instrument flight.
5. Failure to practice basic attitude instrument flying. When a pilot does not fly instrument approach procedures or even basic attitude instrument flying maneuvers for long periods of time, skill levels diminish. Pilots should avoid flying in IMC if they are not proficient. They should seek a qualified instructor to receive additional instruction prior to entry into IMC.

Instrument Takeoff

The reason for learning to fly by reference to instruments alone is to expand a pilot's abilities to operate an aircraft in visibility less than VFR. Another valuable maneuver to learn is the instrument takeoff. This maneuver requires the pilot to maneuver the aircraft during the takeoff roll by reference to flight instruments alone with no outside visual reference. With practice, this maneuver becomes as routine as a standard rate turn.

The reason behind practicing instrument takeoffs is to reduce the disorientation that can occur during the transitional phase of quickly moving the eyes from the outside references inside to the flight instruments.

One EFD system currently offers what is trademarked as synthetic vision. Synthetic vision is a three-dimensional computer-generated representation of the terrain that lies ahead of the aircraft. The display shows runways as well as a depiction of the terrain features based on a GPS terrain database. Similar to a video game, the display generates a runway the pilot can maneuver down in order to maintain directional control. As long as the pilot tracks down the computer-generated runway, the aircraft will remain aligned with the actual runway.

Not all EFD systems have such an advanced visioning system. With all other systems, the pilot needs to revert to the standard procedures for instrument takeoffs. Each aircraft may require a modification to the maneuver; therefore, always obtain training on any new equipment to be used.

In order to accomplish an instrument takeoff, the aircraft needs to be maneuvered on the centerline of the runway facing the direction of departure with the nose or tail wheel straight. Assistance from the instructor may be necessary if the pilot has been taxiing while wearing a view limiting device. Lock the tail wheel, if so equipped, and hold the brakes firmly to prevent the aircraft from creeping. Cross-check the heading indicator on the PFD with the magnetic compass and adjust for any deviations noted on the compass card. Set the heading to the nearest 5° mark closest to the runway heading. This allows the pilot to quickly detect any deviations from the desired heading and allows prompt corrective actions during the takeoff roll. Using the omnibearing select (OBS) mode on the GPS, rotate the OBS selector until the needle points to the runway heading. This adds additional situational awareness during the takeoff roll. Smoothly apply power to generate sufficient rudder authority for directional control. Release the brakes and continue to advance the power to the takeoff setting.

As soon as the brakes are released, any deviation in heading needs to be corrected immediately. Avoid using brakes to control direction as this increases the takeoff roll, as well as provides the potential of overcontrolling the aircraft.

Continuously cross-check the ASI and the heading indicator as the aircraft accelerates. As the aircraft approaches 15-25 knots below the rotation speed, smoothly apply aft elevator pressure to increase the pitch attitude to the desired takeoff attitude (approximately 7° for most small airplanes). With the pitch attitude held constant, continue to cross-check the flight instruments and allow the aircraft to fly off of the runway. Do not pull the aircraft off of the runway. Pulling the aircraft off of the runway imposes left turning tendencies due to P-Factor, which will yaw the aircraft to the left and destabilize the takeoff.

Maintain the desired pitch and bank attitudes by referencing the attitude indicator and cross-check the VSI tape for an indication of a positive rate of climb. Take note of the magenta 6-second altimeter trend indicator. The trend should show positive. Barring turbulence, all trend indications should be stabilized. The airspeed trend indicator should not be visible at this point if the airspeed is being held constant. An activation of the airspeed trend indicator shows that the pitch attitude is not being held at the desired value and, therefore, the airspeed is changing. The desired performance is to be climbing at a constant airspeed and vertical speed rate. Use the ASI as the primary instrument for the pitch indication.

Once the aircraft has reached a safe altitude (approximately 100 feet for insufficient runway available for landing should an engine failure occur) retract the landing gear and flaps while referencing the ASI and attitude indicator to maintain the desired pitch. As the configuration is changed, an increase in aft control pressure is needed in order to maintain the desired pitch attitude. Smoothly increase the aft control pressure to compensate for the change in configuration. Anticipate the changes and increase the rate of cross-check. The airspeed tape and altitude tape increases while the VSI tape is held constant. Allow the aircraft to accelerate to the desired climb speed. Once the desired climb speed is reached, reduce the power to the climb power setting as printed in the POH/AFM. Trim the aircraft to eliminate any control pressures.

Common Errors in Instrument Takeoffs

Common errors associated with the instrument takeoff include, but are not limited to, the following:

1. Failure to perform an adequate flight deck check before the takeoff. Pilots have attempted instrument takeoff with inoperative airspeed indicators (pitot tube obstructed), controls locked, and numerous other oversights due to haste or carelessness. It is imperative to cross-check the ASI as soon as possible. No airspeed will be indicated until 20 knots of true airspeed is generated in some systems.
2. Improper alignment on the runway. This may result from improper brake applications, allowing the airplane to creep after alignment, or from alignment with the nosewheel or tailwheel cocked. In any case, the result is a built-in directional control problem as the takeoff starts.
3. Improper application of power. Abrupt applications of power complicate directional control. Power should be applied in a smooth and continuous manner to arrive at the takeoff power setting within approximately 3 seconds.
4. Improper use of brakes. Incorrect seat or rudder pedal adjustment, with feet in an uncomfortable position, frequently causes inadvertent application of brakes and excessive heading changes.
5. Overcontrolling rudder pedals. This fault may be caused by late recognition of heading changes, tension on the controls, misinterpretation of the heading indicator (and correcting in the wrong direction), failure to appreciate changing effectiveness of rudder control as the aircraft accelerates, and other factors. If heading changes are observed and corrected instantly with small movement of the rudder pedals, swerving tendencies can be reduced.
6. Failure to maintain attitude after becoming airborne. If the pilot reacts to seat-of-the-pants sensations when the airplane lifts off, pitch control is guesswork. The pilot may either allow excessive pitch or apply excessive forward-elevator pressure, depending on the reaction to trim changes.
7. Inadequate cross-check. Fixations are likely during the trim changes, attitude changes, gear and flap retractions, and power changes. Once an instrument or a control input is applied, continue the cross-check and note the effect control during the next cross-check sequence.
8. Inadequate interpretation of instruments. Failure to understand instrument indications immediately indicates that further study of the maneuver is necessary.

Basic Instrument Flight Patterns

After attaining a reasonable degree of proficiency in basic maneuvers, apply these skills to the various combinations of individual maneuvers. The practice flight patterns, beginning on page 5-30, are directly applicable to operational instrument flying.

STBY PWR
20
10
10
20
20
10
10
20

Chapter 6

Helicopter Attitude Instrument Flying

Introduction

Attitude instrument flying in helicopters is essentially visual flying with the flight instruments substituted for the various reference points on the helicopter and the natural horizon. Control changes, required to produce a given attitude by reference to instruments, are identical to those used in helicopter visual flight rules (VFR) flight, and pilot thought processes are the same. Basic instrument training is intended to be a building block toward attaining an instrument rating.

Flight Instruments

When flying a helicopter with reference to the flight instruments, proper instrument interpretation is the basis for aircraft control. Skill, in part, depends on understanding how a particular instrument or system functions, including its indications and limitations (see Chapter 3, Flight Instruments). With this knowledge, a pilot can quickly interpret an instrument indication and translate that information into a control response.

Instrument Flight

To achieve smooth, positive control of the helicopter during instrument flight, three fundamental skills must be developed. They are instrument cross-check, instrument interpretation, and aircraft control.

Instrument Cross-Check

Cross-checking, sometimes referred to as scanning, is the continuous and logical observation of instruments for attitude and performance information. In attitude instrument flying, an attitude is maintained by reference to the instruments, which produces the desired result in performance. Due to human error, instrument error, and helicopter performance differences in various atmospheric and loading conditions, it is difficult to establish an attitude and have performance remain constant for a long period of time. These variables make it necessary to constantly check the instruments and make appropriate changes in the helicopter's attitude. The actual technique may vary depending on what instruments are installed and where they are installed, as well as pilot experience and proficiency level. This discussion concentrates on the six basic flight instruments. *[Figure 6-1]*

At first, there may be a tendency to cross-check rapidly, looking directly at the instruments without knowing exactly what information is needed. However, with familiarity and practice, the instrument cross-check reveals definite trends during specific flight conditions. These trends help a pilot control the helicopter as it makes a transition from one flight condition to another.

When full concentration is applied to a single instrument, a problem called fixation is encountered. This results from a natural human inclination to observe a specific instrument carefully and accurately, often to the exclusion of other instruments. Fixation on a single instrument usually results in poor control. For example, while performing a turn, there is a tendency to watch only the turn-and-slip indicator instead of including other instruments in the cross-check. This fixation on the turn-and-slip indicator often leads to a loss of altitude through poor pitch-and-bank control. Look at each

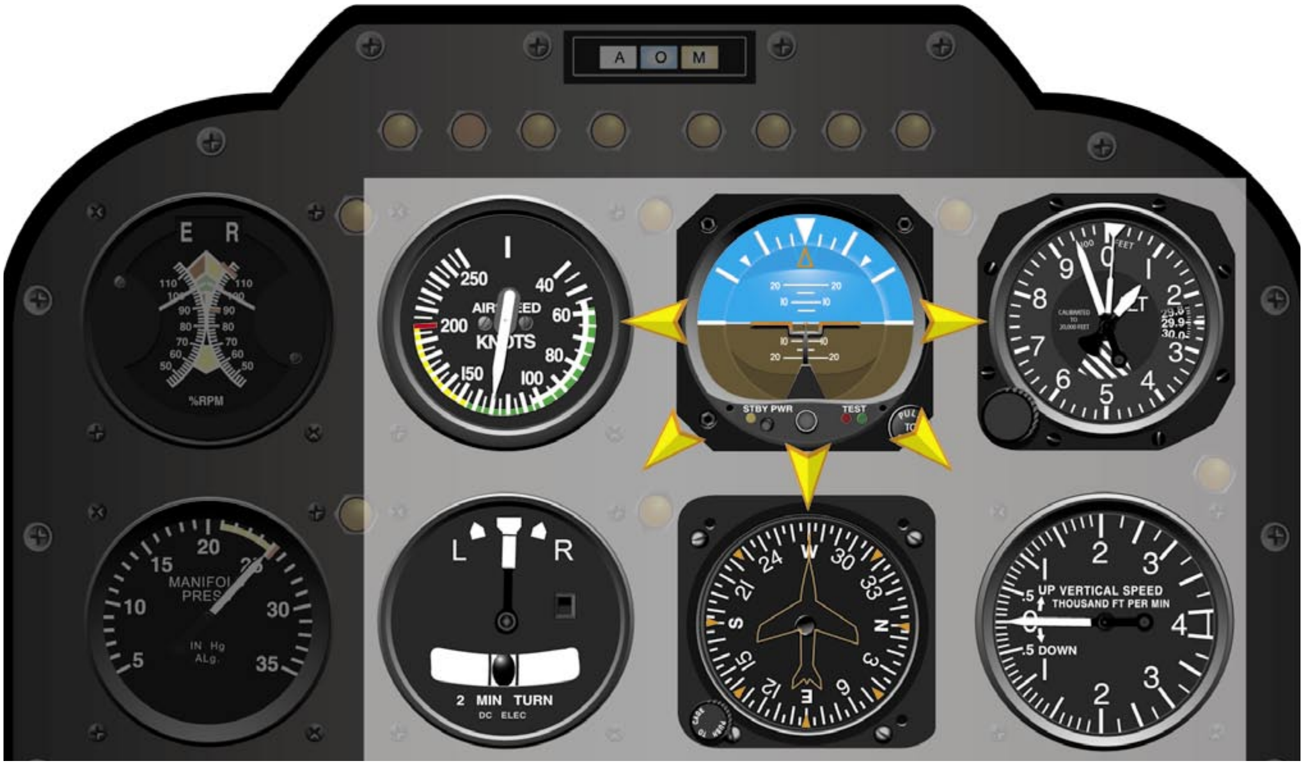

Figure 6-1. *In most situations, the cross-check pattern includes the attitude indicator between the cross-check of each of the other instruments. A typical cross-check might progress as follows: attitude indicator, altimeter, attitude indicator, vertical speed indicator, attitude indicator, heading indicator, attitude indicator, and so on.*

instrument only long enough to understand the information it presents, and then proceed to the next one. Similarly, too much emphasis can be placed on a single instrument, instead of relying on a combination of instruments necessary for helicopter performance information. This differs from fixation in that other instruments are included in a cross-check, but too much attention is placed on one particular instrument.

During performance of a maneuver, there is sometimes a failure to anticipate significant instrument indications following attitude changes. For example, during level off from a climb or descent, a pilot may concentrate on pitch control, while forgetting about heading or roll information. This error, called omission, results in erratic control of heading and bank.

In spite of these common errors, most pilots can adapt well to flight by instrument reference after instruction and practice. Many find that they can control the helicopter more easily and precisely by instruments.

Instrument Interpretation

The flight instruments together give a picture of what is happening. No one instrument is more important than the next; however, during certain maneuvers or conditions, those instruments that provide the most pertinent and useful information are termed primary instruments. Those which back up and supplement the primary instruments are termed supporting instruments. For example, since the attitude indicator is the only instrument that provides instant and direct aircraft attitude information, it should be considered primary during any change in pitch or bank attitude. After the new attitude is established, other instruments become primary, and the attitude indicator usually becomes the supporting instrument.

Aircraft Control

Controlling a helicopter is the result of accurately interpreting the flight instruments and translating these readings into correct control responses. Aircraft control involves adjustment to pitch, bank, power, and trim in order to achieve a desired flight path.

Pitch attitude control is controlling the movement of the helicopter about its lateral axis. After interpreting the helicopter's pitch attitude by reference to the pitch instruments (attitude indicator, altimeter, airspeed indicator, and vertical speed indicator (VSI)), cyclic control adjustments are made to affect the desired pitch attitude. In this chapter, the pitch attitudes depicted are approximate and vary with different helicopters.

Bank attitude control is controlling the angle made by the lateral tilt of the rotor and the natural horizon, or the movement of the helicopter about its longitudinal axis. After interpreting the helicopter's bank instruments (attitude indicator, heading indicator, and turn indicator), cyclic control adjustments are made to attain the desired bank attitude.

Power control is the application of collective pitch with corresponding throttle control, where applicable. In straight-and-level flight, changes of collective pitch are made to correct for altitude deviation if the error is more than 100 feet, or the airspeed is off by more than 10 knots. If the error is less than that amount, a pilot should use a slight cyclic climb or descent.

In order to fly a helicopter by reference to the instruments, it is important to know the approximate power settings required for a particular helicopter in various load configurations and flight conditions.

Trim, in helicopters, refers to the use of the cyclic centering button, if the helicopter is so equipped, to relieve all possible cyclic pressures. Trim also refers to the use of pedal adjustment to center the ball of the turn indicator. Pedal trim is required during all power changes.

The proper adjustment of collective pitch and cyclic friction helps a pilot relax during instrument flight. Friction should be adjusted to minimize overcontrolling and to prevent creeping, but not applied to such a degree that control movement is limited. In addition, many helicopters equipped for instrument flight contain stability augmentation systems or an autopilot to help relieve pilot workload.

Straight-and-Level Flight

Straight-and-level unaccelerated flight consists of maintaining the desired altitude, heading, airspeed, and pedal trim.

Pitch Control

The pitch attitude of a helicopter is the angular relation of its longitudinal axis to the natural horizon. If available, the attitude indicator is used to establish the desired pitch attitude. In level flight, pitch attitude varies with airspeed and center of gravity (CG). At a constant altitude and a stabilized airspeed, the pitch attitude is approximately level. *[Figure 6-2]*

Attitude Indicator

The attitude indicator gives a direct indication of the pitch attitude of the helicopter. In visual flight, attain the desired pitch attitude by using the cyclic to raise and lower the nose

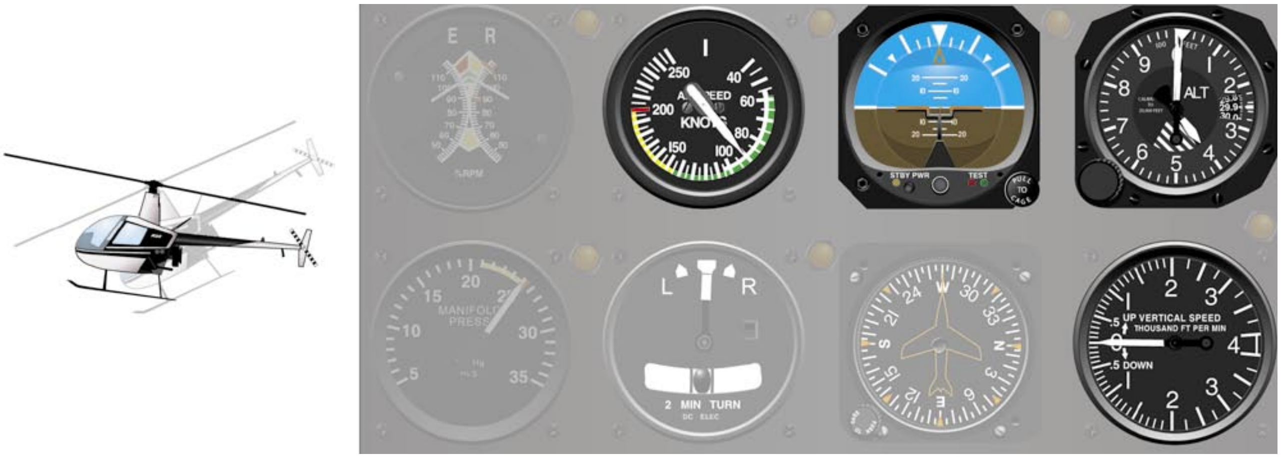

Figure 6-2. *The flight instruments for pitch control are the airspeed indicator, attitude indicator, altimeter, and vertical speed indicator.*

of the helicopter in relation to the natural horizon. During instrument flight, follow exactly the same procedure in raising or lowering the miniature aircraft in relation to the horizon bar.

There is some delay between control application and resultant instrument change. This is the normal control lag in the helicopter and should not be confused with instrument lag. The attitude indicator may show small misrepresentations of pitch attitude during maneuvers involving acceleration, deceleration, or turns. This precession error can be detected quickly by cross-checking the other pitch instruments.

If the miniature aircraft is properly adjusted on the ground, it may not require readjustment in flight. If the miniature aircraft is not on the horizon bar after level off at normal cruising airspeed, adjust it as necessary while maintaining level flight with the other pitch instruments. Once the miniature aircraft has been adjusted in level flight at normal cruising airspeed, leave it unchanged so it gives an accurate picture of pitch attitude at all times.

When making initial pitch attitude corrections to maintain altitude, the changes of attitude should be small and smoothly applied. The initial movement of the horizon bar should not exceed one bar width high or low. *[Figure 6-3]* If a further adjustment is required, an additional correction of one-half bar normally corrects any deviation from the desired altitude. This one-and-one-half bar correction is normally the maximum pitch attitude correction from level flight attitude.

After making the correction, cross-check the other pitch instruments to determine whether the pitch attitude change is sufficient. If additional correction is needed to return to altitude, or if the airspeed varies more than 10 knots from that desired, adjust the power.

Altimeter

The altimeter gives an indirect indication of the pitch attitude of the helicopter in straight-and-level flight. Since the altitude should remain constant in level flight, deviation from the desired altitude indicates a need for a change in pitch attitude and power as necessary. When losing altitude, raise the pitch attitude and adjust power as necessary. When gaining altitude, lower the pitch attitude and adjust power as necessary. Indications for power changes are explained in the next paragraph.

The rate at which the altimeter moves helps to determine pitch attitude. A very slow movement of the altimeter indicates a small deviation from the desired pitch attitude, while a

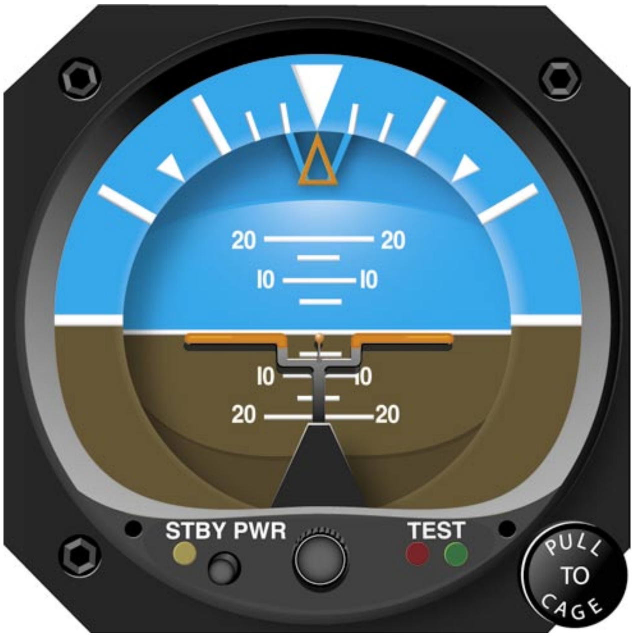

Figure 6-3. *The initial pitch correction at normal cruise is one bar width or less.*

fast movement of the altimeter indicates a large deviation from the desired pitch attitude. Make any corrective action promptly, with small control changes. Also, remember that movement of the altimeter should always be corrected by two distinct changes. The first is a change of attitude to stop the altimeter movement; the second is a change of attitude to return smoothly to the desired altitude. If altitude and airspeed are more than 100 feet and 10 knots low, respectively, apply power in addition to an increase of pitch attitude. If the altitude and airspeed are high by more than 100 feet and 10 knots, reduce power and lower the pitch attitude.

There is a small lag in the movement of the altimeter; however, for all practical purposes, consider that the altimeter gives an immediate indication of a change, or a need for change in pitch attitude. Since the altimeter provides the most pertinent information regarding pitch in level flight, it is considered primary for pitch.

Vertical Speed Indicator (VSI)

The VSI gives an indirect indication of the pitch attitude of the helicopter and should be used in conjunction with the other pitch instruments to attain a high degree of accuracy and precision. The instrument indicates zero when in level flight. Any movement of the needle from the zero position shows a need for an immediate change in pitch attitude to return it to zero. Always use the VSI in conjunction with the altimeter in level flight. If a movement of the VSI is detected, immediately use the proper corrective measures to return it to zero. If the correction is made promptly, there is usually little or no change in altitude. If the needle of the VSI does not indicate zero, the altimeter indicates a gain or loss of altitude.

The initial movement of the vertical speed needle is instantaneous and indicates the trend of the vertical movement of the helicopter. A period of time is necessary for the VSI to reach its maximum point of deflection after a correction has been made. This time element is commonly referred to as instrument lag. The lag is directly proportional to the speed and magnitude of the pitch change. When employing smooth control techniques and small adjustments in pitch attitude are made, lag is minimized, and the VSI is easy to interpret.

Overcontrolling can be minimized by first neutralizing the controls and allowing the pitch attitude to stabilize, then readjusting the pitch attitude by noting the indications of the other pitch instruments.

Occasionally, the VSI may be slightly out of calibration. This could result in the instrument indicating a slight climb or descent even when the helicopter is in level flight. If the instrument cannot be calibrated properly, this error must be taken into consideration when using the VSI for pitch control. For example, if a descent of 100 feet per minute (fpm) is the vertical speed indication when the helicopter is in level flight, use that indication as level flight. Any deviation from that reading would indicate a change in attitude.

Airspeed Indicator

The airspeed indicator gives an indirect indication of helicopter pitch attitude. With a given power setting and pitch attitude, the airspeed remains constant. If the airspeed increases, the nose is too low and should be raised. If the airspeed decreases, the nose is too high and should be lowered. A rapid change in airspeed indicates a large change in pitch attitude, and a slow change in airspeed indicates a small change in pitch attitude. There is very little lag in the indications of the airspeed indicator. If, while making attitude changes, there is some lag between control application and change of airspeed, it is most likely due to cyclic control lag. Generally, a departure from the desired airspeed, due to an inadvertent pitch attitude change, also results in a change in altitude. For example, an increase in airspeed due to a low pitch attitude results in a decrease in altitude. A correction in the pitch attitude regains both airspeed and altitude.

Bank Control

The bank attitude of a helicopter is the angular relation of its lateral axis to the natural horizon. To maintain a straight course in visual flight, keep the lateral axis of the helicopter level with the natural horizon. Assuming the helicopter is in coordinated flight, any deviation from a laterally level attitude produces a turn. *[Figure 6-4]*

Attitude Indicator

The attitude indicator gives a direct indication of the bank attitude of the helicopter. For instrument flight, the miniature aircraft and the horizon bar of the attitude indicator are substituted for the actual helicopter and the natural horizon. Any change in bank attitude of the helicopter is indicated instantly by the miniature aircraft. For proper interpretation of this instrument, imagine being in the miniature aircraft. If the helicopter is properly trimmed and the rotor tilts, a turn begins. The turn can be stopped by leveling the miniature aircraft with the horizon bar. The ball in the turn-and-slip indicator should always be kept centered through proper pedal trim.

The angle of bank is indicated by the pointer on the banking scale at the top of the instrument. *[Figure 6-5]* Small bank angles, which may not be seen by observing the miniature aircraft, can easily be determined by referring to the banking scale pointer.

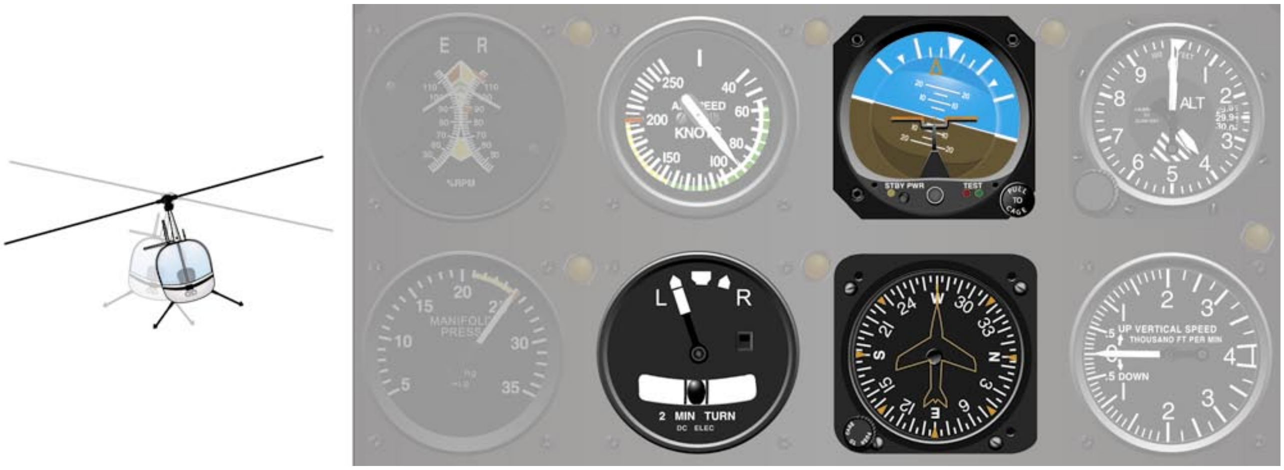

Figure 6-4. *The flight instruments used for bank control are the attitude, heading, and turn indicators.*

Pitch-and-bank attitudes can be determined simultaneously on the attitude indicator. Even though the miniature aircraft is not level with the horizon bar, pitch attitude can be established by observing the relative position of the miniature aircraft and the horizon bar.

The attitude indicator may show small misrepresentations of bank attitude during maneuvers that involve turns. This precession error can be detected immediately by closely cross-checking the other bank instruments during these maneuvers. Precession is normally noticed when rolling out of a turn. If, upon completion of a turn, the miniature aircraft is level and the helicopter is still turning, make a small change of bank attitude to center the turn needle and stop the movement of the heading indicator.

Heading Indicator

In coordinated flight, the heading indicator gives an indirect indication of a helicopter's bank attitude. When a helicopter is banked, it turns. When the lateral axis of a helicopter is level, it flies straight. Therefore, in coordinated flight when the heading indicator shows a constant heading, the helicopter is level laterally. A deviation from the desired heading indicates a bank in the direction the helicopter is turning. A small angle of bank is indicated by a slow change of heading; a large angle of bank is indicated by a rapid change of heading. If a turn is noticed, apply opposite cyclic until the heading indicator

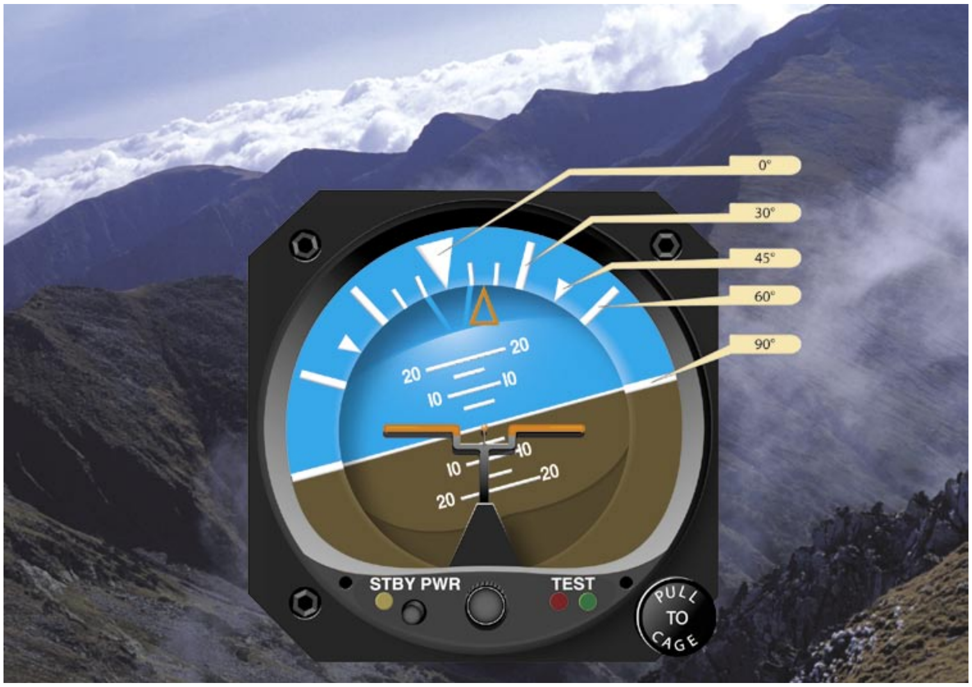

Figure 6-5. *The banking scale at the top of the attitude indicator indicates varying degrees of bank. In this example, the helicopter is banked approximately 15° to the right.*

indicates the desired heading, simultaneously ensuring the ball is centered. When making the correction to the desired heading, do not use a bank angle greater than that required to achieve a standard rate turn. In addition, if the number of degrees of change is small, limit the bank angle to the number of degrees to be turned. Bank angles greater than these require more skill and precision in attaining the desired results. During straight-and-level flight, the heading indicator is the primary reference for bank control.

Turn Indicator

During coordinated flight, the needle of the turn-and-slip indicator gives an indirect indication of the bank attitude of the helicopter. When the needle is displaced from the vertical position, the helicopter is turning in the direction of the displacement. Thus, if the needle is displaced to the left, the helicopter is turning left. Bringing the needle back to the vertical position with the cyclic produces straight flight. A close observation of the needle is necessary to accurately interpret small deviations from the desired position.

Cross-check the ball of the turn-and-slip indicator to determine if the helicopter is in coordinated flight. *[Figure 6-6]* If the rotor is laterally level and pedal pressure properly compensates for torque, the ball remains in the center. To center the ball, level the helicopter laterally by reference to the other bank instruments, then center the ball with pedal trim. Torque correction pressures vary as power changes are made. Always check the ball after such changes.

Common Errors During Straight-and-Level Flight

1. Failure to maintain altitude
2. Failure to maintain heading
3. Overcontrolling pitch and bank during corrections
4. Failure to maintain proper pedal trim
5. Failure to cross-check all available instruments

Power Control During Straight-and-Level Flight

Establishing specific power settings is accomplished through collective pitch adjustments and throttle control, where necessary. For reciprocating-powered helicopters, power indication is observed on the manifold pressure gauge. For turbine-powered helicopters, power is observed on the torque gauge. (Although most Instrument Flight Rules (IFR)-certified helicopters are turbine powered, depictions within this chapter use a reciprocating-powered helicopter as this is where training is most likely conducted.)

At any given airspeed, a specific power setting determines whether the helicopter is in level flight, in a climb, or in a descent. For example, cruising airspeed maintained with cruising power results in level flight. If a pilot increases the power setting and holds the airspeed constant, the helicopter climbs. Conversely, if the pilot decreases power and holds the airspeed constant, the helicopter descends.

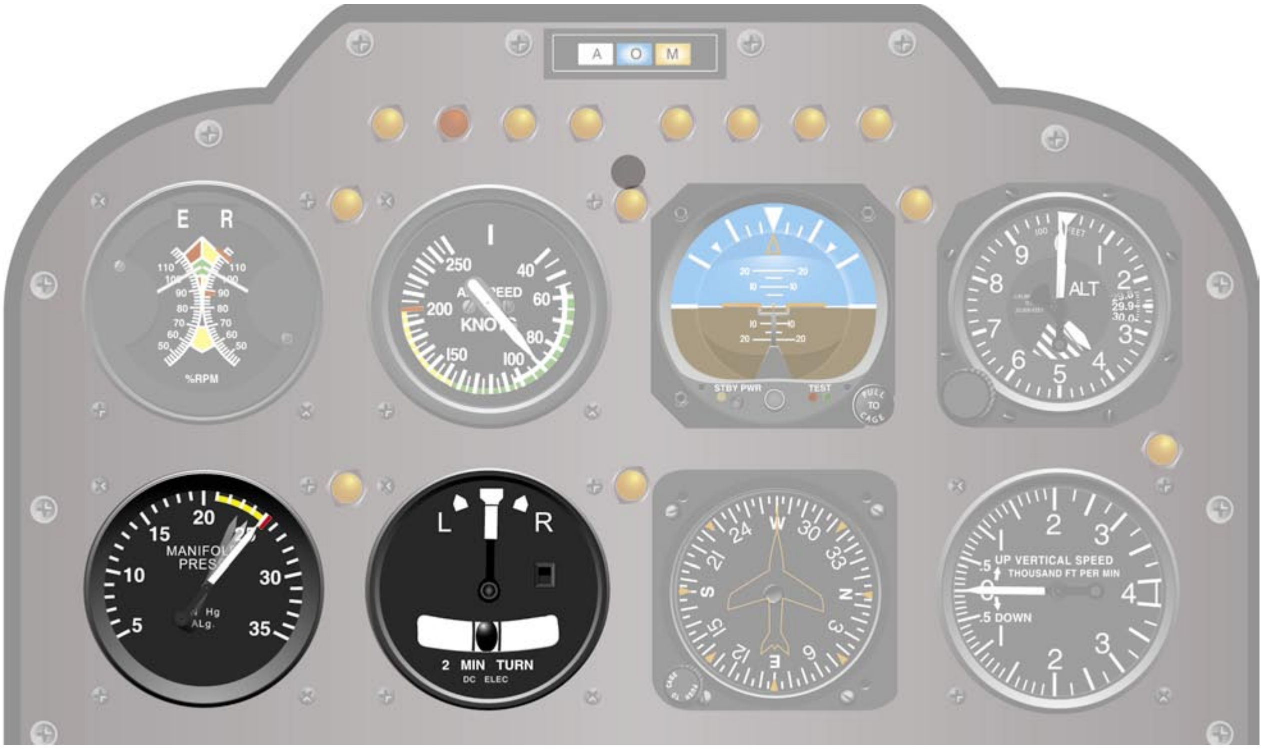

Figure 6-6. *Coordinated flight is indicated by centering of the ball.*

If the altitude is held constant, power determines the airspeed. For example, at a constant altitude, cruising power results in cruising airspeed. Any deviation from the cruising power setting results in a change of airspeed. When power is added to increase airspeed, the nose of the helicopter pitches up and yaws to the right in a helicopter with a counterclockwise main rotor blade rotation. *[Figure 6-7]* When power is reduced to decrease airspeed, the nose pitches down and yaws to the left. *[Figure 6-8]* The yawing effect is most pronounced in single-rotor helicopters, and is absent in helicopters with counter-rotating rotors. To counteract the yawing tendency of the helicopter, apply pedal trim during power changes.

To maintain a constant altitude and airspeed in level flight, coordinate pitch attitude and power control. The relationship between altitude and airspeed determines the need for a change in power and/or pitch attitude. If the altitude is constant and the airspeed is high or low, change the power to obtain the desired airspeed. During the change in power, make an accurate interpretation of the altimeter, then counteract any deviation from the desired altitude by an appropriate change of pitch attitude. If the altitude is low and the airspeed is high, or vice versa, a change in pitch attitude alone may return the helicopter to the proper altitude and airspeed. If both airspeed and altitude are low, or if both are high, changes in both power and pitch attitude are necessary.

To make power control easy when changing airspeed, it is necessary to know the approximate power settings for the various airspeeds at which the helicopter is flown. When the airspeed is to be changed by any appreciable amount, adjust the power so that it is over or under that setting necessary to maintain the new airspeed. As the power approaches the desired setting, include the manifold pressure in the cross-check to determine when the proper adjustment has been accomplished. As the airspeed is changing, adjust the pitch attitude to maintain a constant altitude. A constant heading should be maintained throughout the change. As the desired airspeed is approached, adjust power to the new cruising power setting and further adjust pitch attitude to maintain altitude. The instrument indications for straight-and-level flight at normal cruise and during the transition from normal cruise to slow cruise are illustrated in *Figures 6-9* and *6-10*. After the airspeed stabilizes at slow cruise, the attitude indicator shows an approximate level pitch attitude.

The altimeter is the primary pitch instrument during level flight, whether flying at a constant airspeed or during a change in airspeed. Altitude should not change during airspeed transitions, and the heading indicator remains the primary bank instrument. Whenever the airspeed is changed by an appreciable amount, the manifold pressure gauge is momentarily the primary instrument for power control. When the airspeed approaches the desired reading, the airspeed indicator again becomes the primary instrument for power control.

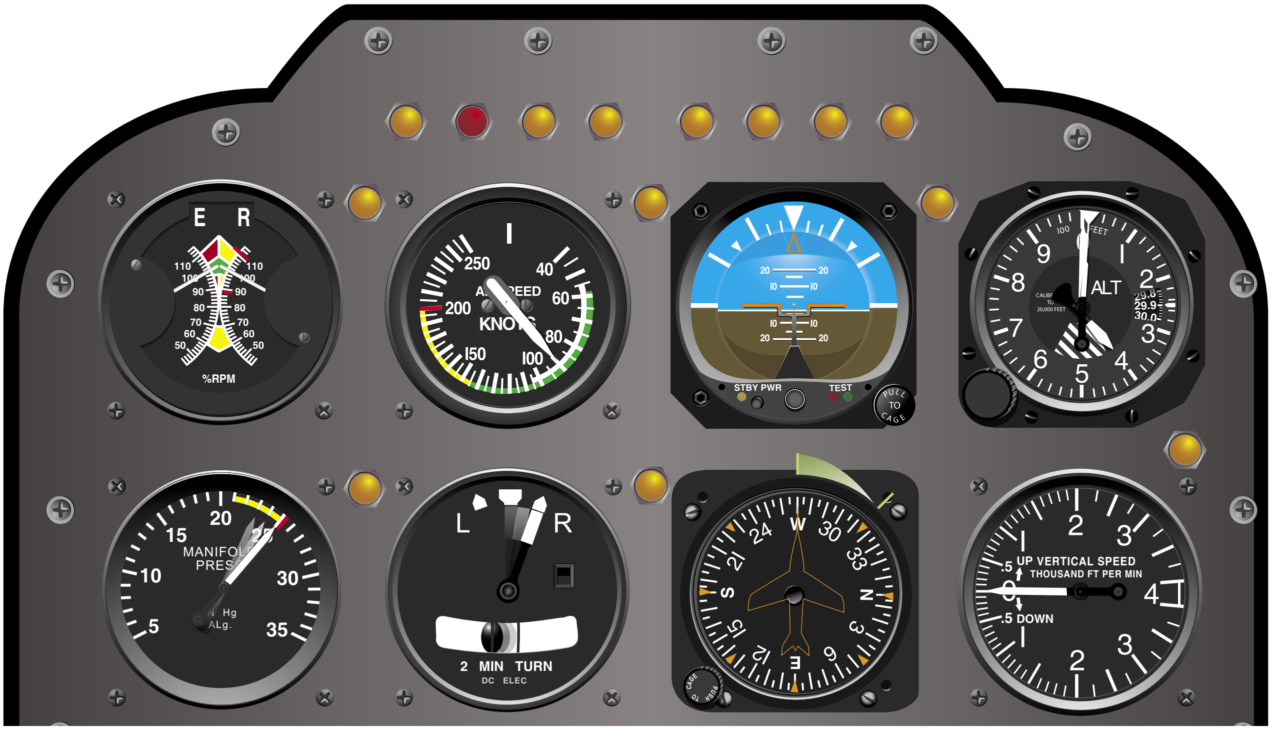

Figure 6-7. *Flight instrument indications in straight-and-level flight with power increasing.*

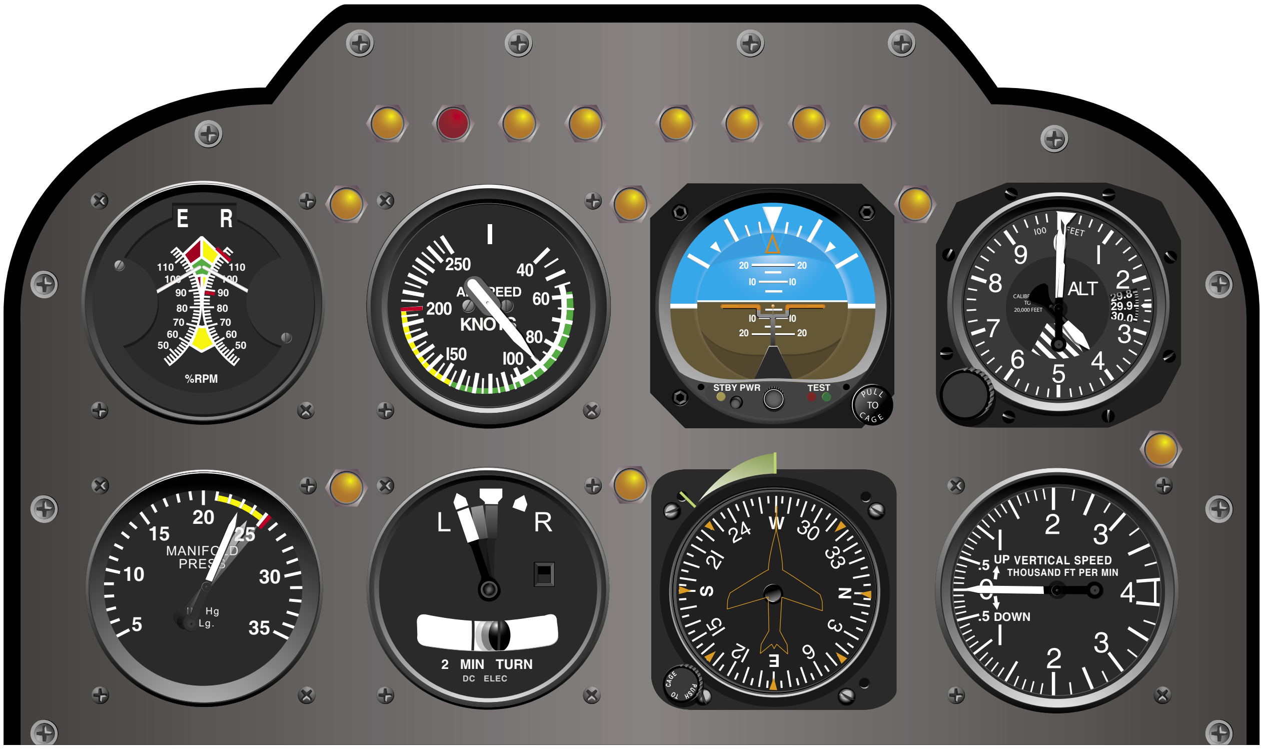
E R
%RPM
AIRSPEED
KNOTS
STBY PWR
TEST
PULL TO CAGE
ALT
MANIFOLD PRES
In. Hg
L R
2 MIN TURN
DC ELEC
UP VERTICAL SPEED
THOUSAND FT PER MIN
DOWN

Figure 6-8. *Flight instrument indications in straight-and-level flight with power decreasing.*

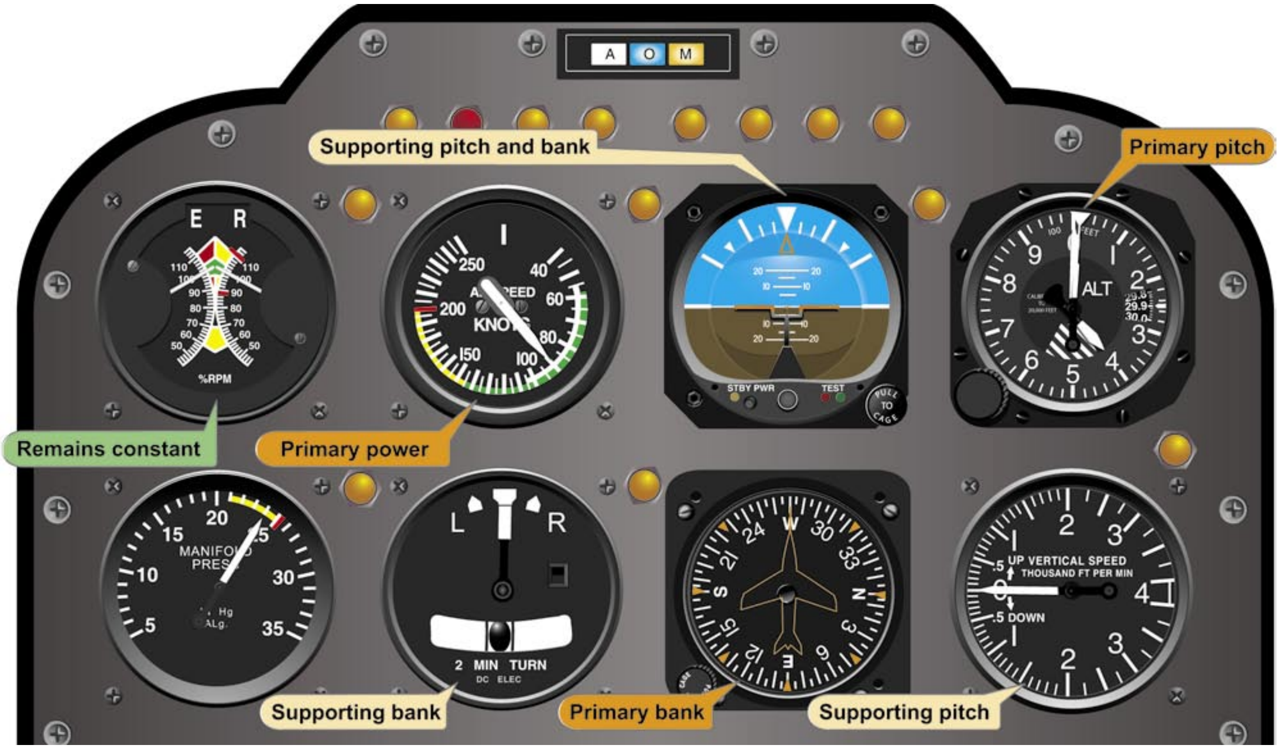
A O M
E R
%RPM
AIRSPEED
KNOTS
STBY PWR
TEST
PULL TO CAGE
ALT
MANIFOLD PRES
In. Hg
L R
2 MIN TURN
DC ELEC
UP VERTICAL SPEED
THOUSAND FT PER MIN
DOWN

Supporting pitch and bank

Primary pitch

Remains constant

Primary power

Supporting bank

Primary bank

Supporting pitch

Figure 6-9. *Flight instrument indications in straight-and-level flight at normal cruise speed.*

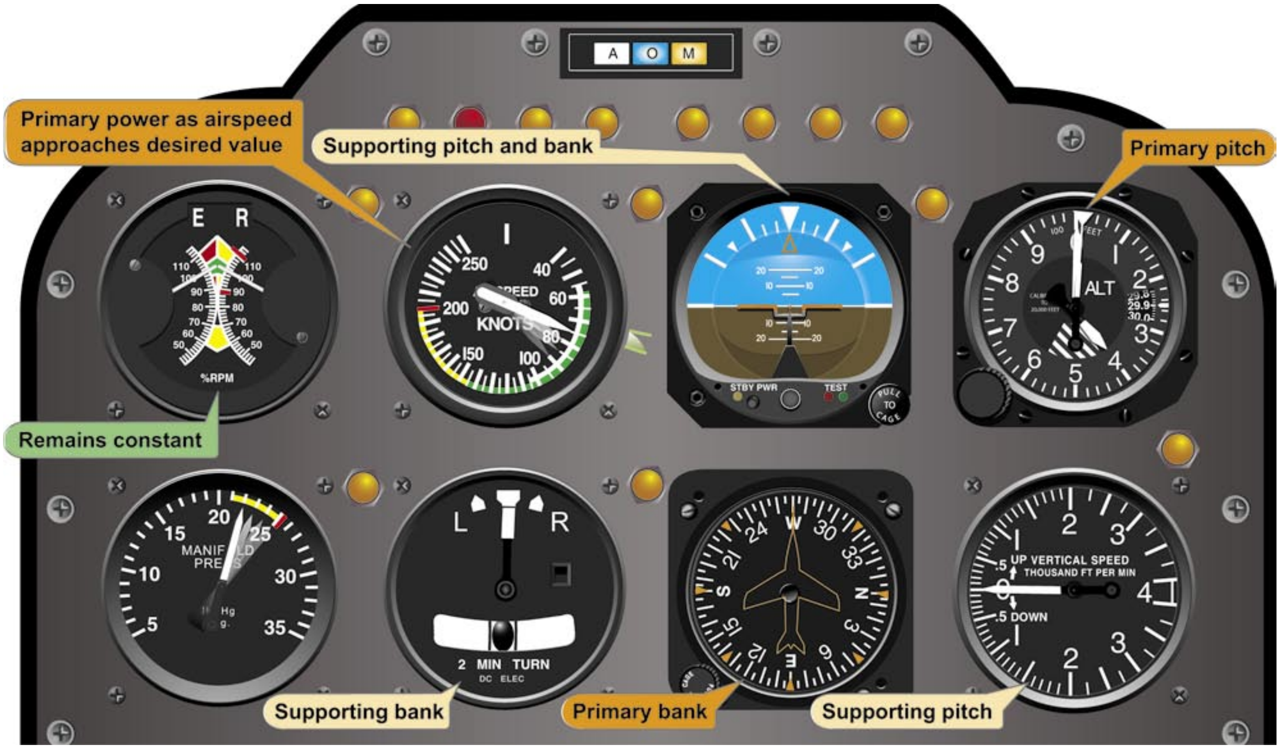
Primary power as airspeed approaches desired value
Supporting pitch and bank
Primary pitch
Remains constant
Supporting bank
Primary bank
Supporting pitch

Figure 6-10. *Flight instrument indications in straight-and-level flight with airspeed decreasing.*

To produce straight-and-level flight, the cross-check of the pitch-and-bank instruments should be combined with the power control instruments. With a constant power setting, a normal cross-check should be satisfactory. When changing power, the speed of the cross-check must be increased to cover the pitch-and-bank instruments adequately. This is necessary to counteract any deviations immediately.

Common Errors During Airspeed Changes

1. Improper use of power
2. Overcontrolling pitch attitude
3. Failure to maintain heading
4. Failure to maintain altitude
5. Improper pedal trim

Straight Climbs (Constant Airspeed and Constant Rate)

For any power setting and load condition, there is only one airspeed that gives the most efficient rate of climb. To determine this, consult the climb data for the type of helicopter being flown. The technique varies according to the airspeed on entry and whether a constant airspeed or constant rate climb is made.

Entry

To enter a constant airspeed climb from cruise airspeed when the climb speed is lower than cruise speed, simultaneously increase power to the climb power setting and adjust pitch attitude to the approximate climb attitude. The increase in power causes the helicopter to start climbing and only very slight back cyclic pressure is needed to complete the change from level to climb attitude. The attitude indicator should be used to accomplish the pitch change. If the transition from level flight to a climb is smooth, the VSI shows an immediate upward trend and then stops at a rate appropriate to the stabilized airspeed and attitude. Primary and supporting instruments for climb entry are illustrated in *Figure 6-11*.

When the helicopter stabilizes at a constant airspeed and attitude, the airspeed indicator becomes primary for pitch. The manifold pressure continues to be primary for power and should be monitored closely to determine if the proper climb power setting is being maintained. Primary and supporting instruments for a stabilized constant airspeed climb are shown in *Figure 6-12*.

The technique and procedures for entering a constant rate climb are very similar to those previously described for a constant airspeed climb. For training purposes, a constant

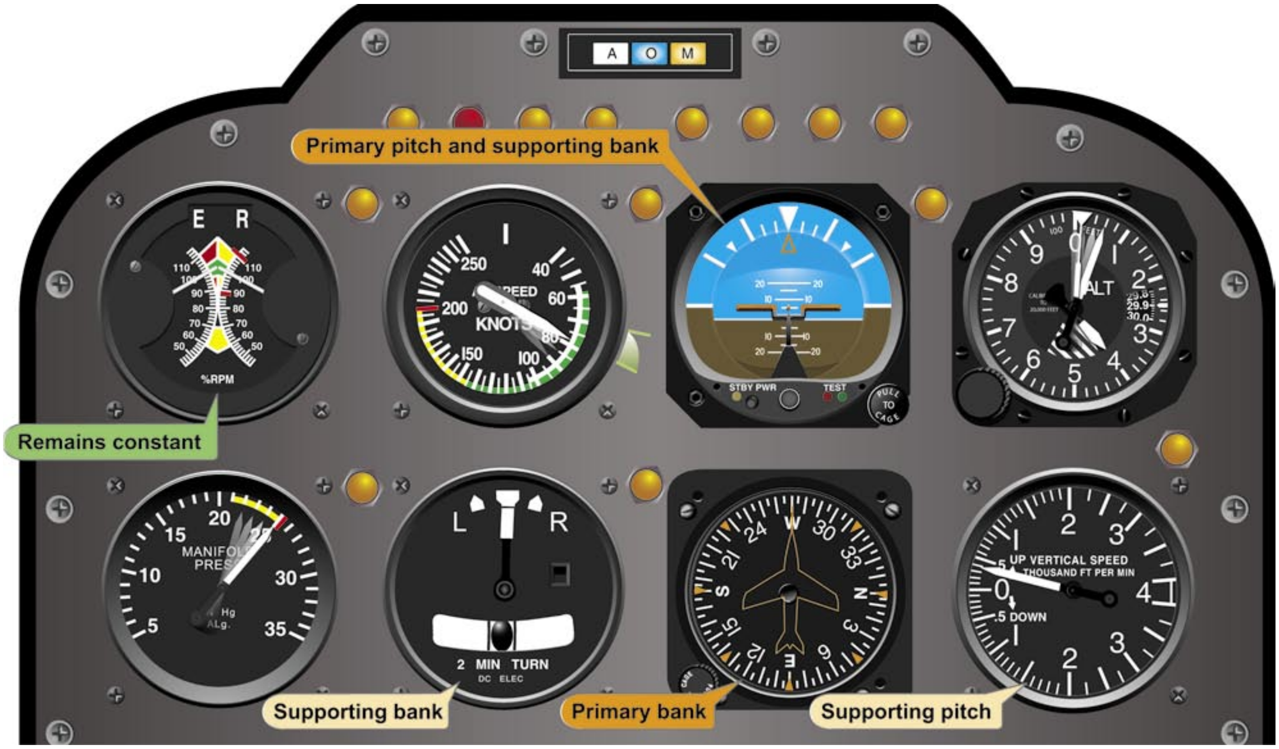
Primary pitch and supporting bank
Remains constant
Supporting bank
Primary bank
Supporting pitch

Figure 6-11. *Flight instrument indications during climb entry for a constant-airspeed climb.*

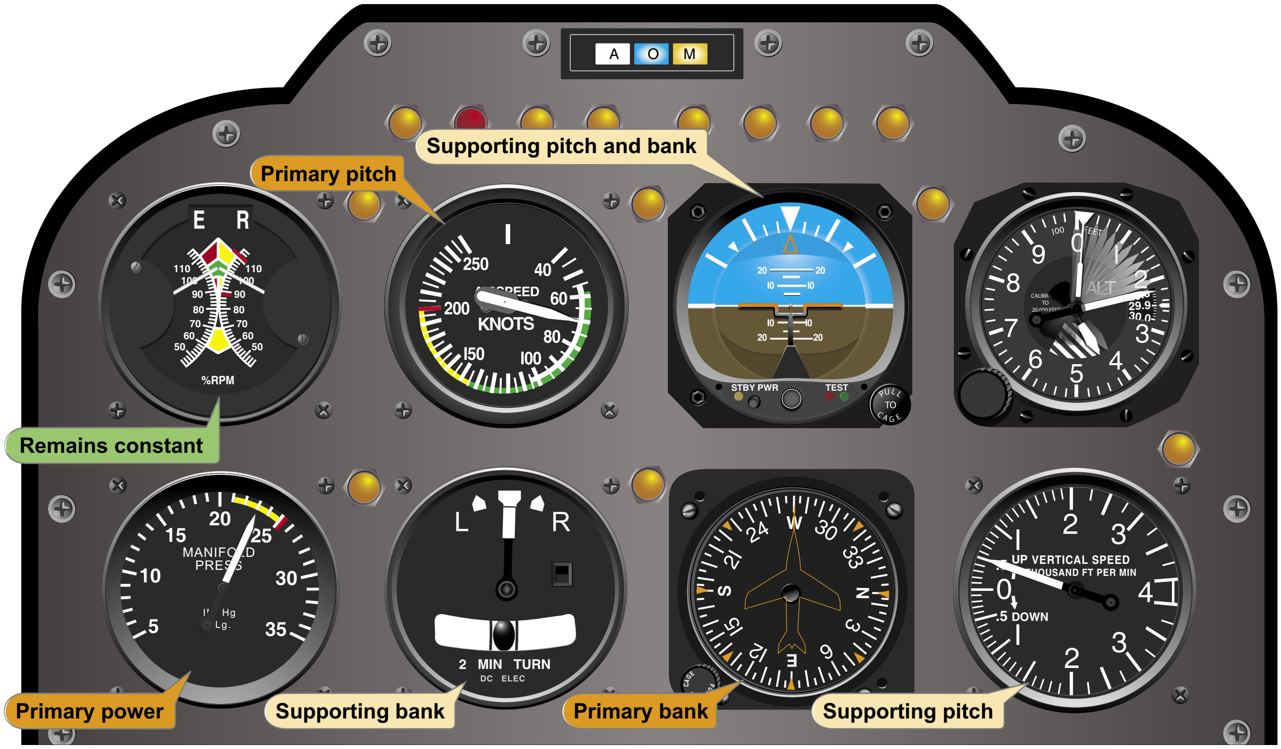
Supporting pitch and bank
Primary pitch
Remains constant
Primary power
Supporting bank
Primary bank
Supporting pitch

Figure 6-12. *Flight instrument indications in a stabilized constant-airspeed climb.*

rate climb is entered from climb airspeed. Use the rate appropriate for the particular helicopter being flown. Normally, in helicopters with low climb rates, 500 fpm is appropriate. In helicopters capable of high climb rates, use a rate of 1,000 fpm.

To enter a constant rate climb, increase power to the approximate setting for the desired rate. As power is applied, the airspeed indicator is primary for pitch until the vertical speed approaches the desired rate. At this time, the VSI becomes primary for pitch. Change pitch attitude by reference to the attitude indicator to maintain the desired vertical speed. When the VSI becomes primary for pitch, the airspeed indicator becomes primary for power. Primary and supporting instruments for a stabilized constant rate climb are illustrated in *Figure 6-13*. Adjust power to maintain desired airspeed. Pitch attitude and power corrections should be closely coordinated. To illustrate this, if the vertical speed is correct but the airspeed is low, add power. As power is increased, it may be necessary to lower the pitch attitude slightly to avoid increasing the vertical rate. Adjust the pitch attitude smoothly to avoid overcontrolling. Small power corrections are usually sufficient to bring the airspeed back to the desired indication.

Level Off

The level off from a constant airspeed climb must be started before reaching the desired altitude. Although the amount of lead varies with the type of helicopter being flown and pilot technique, the most important factor is vertical speed. As a rule of thumb, use 10 percent of the vertical velocity as the lead point. For example, if the rate of climb is 500 fpm, initiate the level off approximately 50 feet before the desired altitude. When the proper lead altitude is reached, the altimeter becomes primary for pitch. Adjust the pitch attitude to the level flight attitude for that airspeed. Cross-check the altimeter and VSI to determine when level flight has been attained at the desired altitude. If cruise airspeed is higher than climb airspeed, leave the power at the climb power setting until the airspeed approaches cruise airspeed, and then reduce it to the cruise power setting. The level off from a constant rate climb is accomplished in the same manner as the level off from a constant airspeed climb.

Straight Descents (Constant Airspeed and Constant Rate)

A descent may be performed at any normal airspeed the helicopter can attain, but the airspeed must be determined prior to entry. The technique is determined by the type of descent, a constant airspeed or a constant rate.

Entry

If airspeed is higher than descending airspeed, and a constant airspeed descent is desired, reduce power to a descent power setting and maintain a constant altitude using cyclic pitch control. This slows the helicopter. As the helicopter

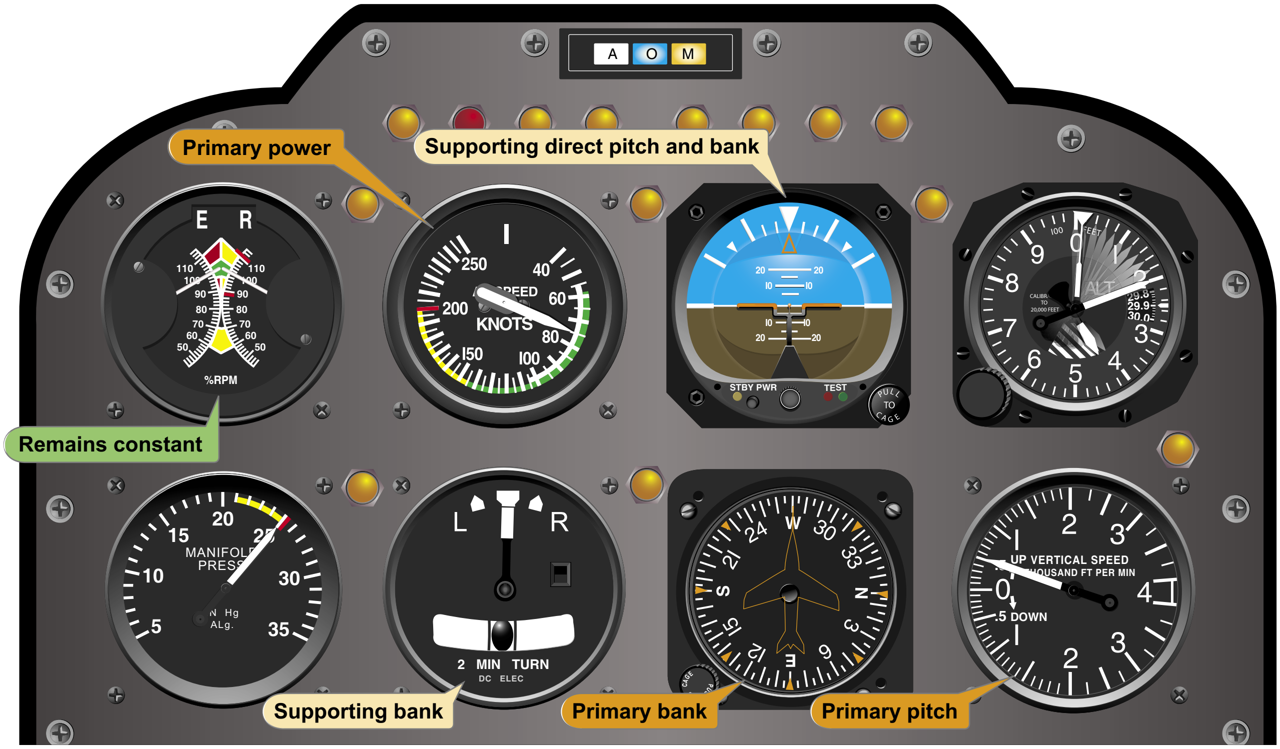

Figure 6-13. *Flight Instrument Indications in a Stabilized Constant-Rate Climb.*

approaches the descending airspeed, the airspeed indicator becomes primary for pitch and the manifold pressure is primary for power. Holding the airspeed constant causes the helicopter to descend. For a constant rate descent, reduce the power to the approximate setting for the desired rate. If the descent is started at the descending airspeed, the airspeed indicator is primary for pitch until the VSI approaches the desired rate. At this time, the VSI becomes primary for pitch, and the airspeed indicator becomes primary for power. Coordinate power and pitch attitude control as previously described on page 6-10 for constant rate climbs.

Level Off

The level off from a constant airspeed descent may be made at descending airspeed or at cruise airspeed, if this is higher than descending airspeed. As in a climb level off, the amount of lead depends on the rate of descent and control technique. For a level off at descending airspeed, the lead should be approximately 10 percent of the vertical speed. At the lead altitude, simultaneously increase power to the setting necessary to maintain descending airspeed in level flight. At this point, the altimeter becomes primary for pitch, and the airspeed indicator becomes primary for power.

To level off at an airspeed higher than descending airspeed, increase the power approximately 100 to 150 feet prior to reaching the desired altitude. The power setting should be that which is necessary to maintain the desired airspeed in level flight. Hold the vertical speed constant until approximately 50 feet above the desired altitude. At this point, the altimeter becomes primary for pitch and the airspeed indicator becomes primary for power. The level off from a constant rate descent should be accomplished in the same manner as the level off from a constant airspeed descent.

Common Errors During Straight Climbs and Descents

1. Failure to maintain heading
2. Improper use of power
3. Poor control of pitch attitude
4. Failure to maintain proper pedal trim
5. Failure to level off on desired altitude

Turns

Turns made by reference to the flight instruments should be made at a precise rate. Turns described in this chapter are those not exceeding a standard rate of 3° per second as indicated on the turn-and-slip indicator. True airspeed determines the angle of bank necessary to maintain a standard rate turn. A rule of thumb to determine the approximate angle of bank required for a standard rate turn is to use 15 percent of the airspeed. A simple way to determine this amount is to divide the airspeed by 10 and add one-half the result. For example, at 60 knots approximately 9° of bank is required (60 ÷ 10 = 6, 6 + 3 = 9); at 80 knots approximately 12° of bank is needed for a standard rate turn.

To enter a turn, apply lateral cyclic in the direction of the desired turn. The entry should be accomplished smoothly, using the attitude indicator to establish the approximate bank angle. When the turn indicator indicates a standard rate turn, it becomes primary for bank. The attitude indicator now becomes a supporting instrument. During level turns, the altimeter is primary for pitch, and the airspeed indicator is primary for power. Primary and supporting instruments for a stabilized standard rate turn are illustrated in *Figure 6-14*. If an increase in power is required to maintain airspeed, slight forward cyclic pressure may be required since the helicopter tends to pitch up as collective pitch is increased. Apply pedal trim, as required, to keep the ball centered.

To recover to straight-and-level flight, apply cyclic in the direction opposite the turn. The rate of roll-out should be the same as the rate used when rolling into the turn. As the turn recovery is initiated, the attitude indicator becomes primary for bank. When the helicopter is approximately level, the heading indicator becomes primary for bank as in straight-and-level flight. Cross-check the airspeed indicator and ball closely to maintain the desired airspeed and pedal trim.

Turn to a Predetermined Heading

A helicopter turns as long as its lateral axis is tilted; therefore, the recovery must start before the desired heading is reached. The amount of lead varies with the rate of turn and piloting technique.

As a guide, when making a 3° per second rate of turn, use a lead of one-half the bank angle. For example, if using a 12° bank angle, use half of that, or 6°, as the lead point prior to the desired heading. Use this lead until the exact amount required by a particular technique can be determined. The bank angle should never exceed the number of degrees to be turned. As in any standard rate turn, the rate of recovery should be the same as the rate of entry. During turns to predetermined headings, cross-check the primary and supporting pitch, bank, and power instruments closely.

Timed Turns

A timed turn is a turn in which the clock and turn-and-slip indicator are used to change heading a definite number of degrees in a given time. For example, using a standard rate turn, a helicopter turns 45° in 15 seconds. Using a half-

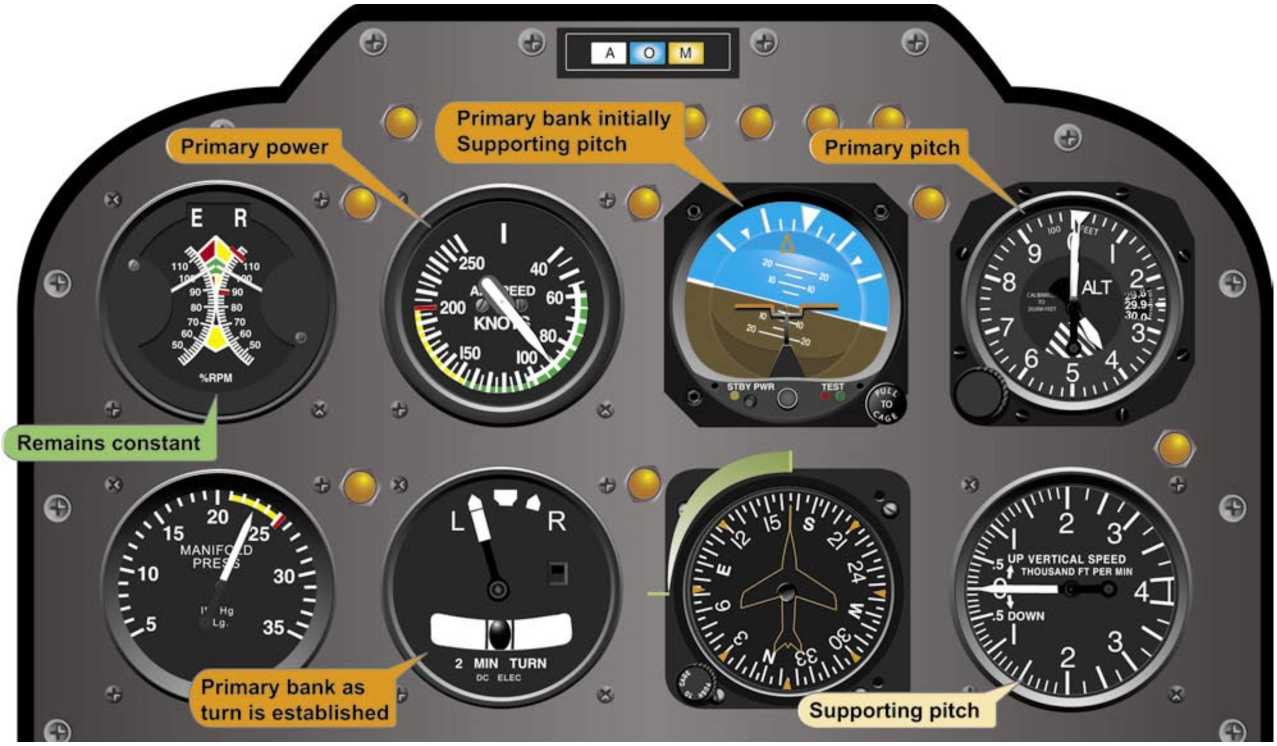

Figure 6-14. *Flight Instrument Indications in a Standard-Rate Turn to the Left.*

standard rate turn, the helicopter turns 45° in 30 seconds. Timed turns can be used if the heading indicator becomes inoperative.

Prior to performing timed turns, the turn coordinator should be calibrated to determine the accuracy of its indications. To do this, establish a standard rate turn by referring to the turn-and-slip indicator. Then, as the sweep second hand of the clock passes a cardinal point (12, 3, 6, or 9), check the heading on the heading indicator. While holding the indicated rate of turn constant, note the heading changes at 10-second intervals. If the helicopter turns more or less than 30° in that interval, a smaller or larger deflection of the needle is necessary to produce a standard rate turn. After the turn-and-slip indicator has been calibrated during turns in each direction, note the corrected deflections, if any, and apply them during all timed turns.

Use the same cross-check and control technique in making timed turns that is used to make turns to a predetermined heading, but substitute the clock for the heading indicator. The needle of the turn-and-slip indicator is primary for bank control, the altimeter is primary for pitch control, and the airspeed indicator is primary for power control. Begin the roll-in when the clock's second hand passes a cardinal point; hold the turn at the calibrated standard rate indication, or half-standard rate for small changes in heading; then begin the roll-out when the computed number of seconds has elapsed. If the roll-in and roll-out rates are the same, the time taken during entry and recovery need not be considered in the time computation.

If practicing timed turns with a full instrument panel, check the heading indicator for the accuracy of the turns. If executing turns without the heading indicator, use the magnetic compass at the completion of the turn to check turn accuracy, taking compass deviation errors into consideration.

Change of Airspeed in Turns

Changing airspeed in turns is an effective maneuver for increasing proficiency in all three basic instrument skills. Since the maneuver involves simultaneous changes in all components of control, proper execution requires a rapid cross-check and interpretation, as well as smooth control. Proficiency in the maneuver also contributes to confidence in the instruments during attitude and power changes involved in more complex maneuvers.

Pitch and power control techniques are the same as those used during airspeed changes in straight-and-level flight. As discussed previously, the angle of bank necessary for a given rate of turn is proportional to the true airspeed. Since the turns are executed at standard rate, the angle of bank must be varied in direct proportion to the airspeed change in order to maintain a constant rate of turn. During a reduction of airspeed, decrease the angle of bank and increase the pitch attitude to maintain altitude and a standard rate turn.

Altimeter and turn indicator readings should remain constant throughout the turn. The altimeter is primary for pitch control, and the turn needle is primary for bank control. Manifold pressure is primary for power control while the airspeed is changing. As the airspeed approaches the new indication, the airspeed indicator becomes primary for power control.

Two methods of changing airspeed in turns may be used. In the first method, airspeed is changed after the turn is established. In the second method, the airspeed change is initiated simultaneously with the turn entry. The first method is easier, but regardless of the method used, the rate of cross-check must be increased as power is reduced. As the helicopter decelerates, check the altimeter and VSI for needed pitch changes, and the bank instruments for needed bank changes. If the needle of the turn-and-slip indicator shows a deviation from the desired deflection, change the bank. Adjust pitch attitude to maintain altitude. When the airspeed approaches that desired, the airspeed indicator becomes primary for power control. Adjust the power to maintain the desired airspeed. Use pedal trim to ensure the maneuver is coordinated.

Until control technique is very smooth, frequently cross-check the attitude indicator to keep from overcontrolling and to provide approximate bank angles appropriate for the changing airspeeds.

Compass Turns

The use of gyroscopic heading indicators makes heading control very easy. However, if the heading indicator fails or the helicopter is not equipped with one, use the magnetic compass for heading reference. When making compass-only turns, a pilot needs to adjust for the lead or lag created by acceleration and deceleration errors so that the helicopter rolls out on the desired heading. When turning to a heading of north, the lead for the roll-out must include the number of degrees of latitude plus the lead normally used in recovery from turns. During a turn to a south heading, maintain the turn until the compass passes south the number of degrees of latitude, minus the normal roll-out lead. For example, when turning from an easterly direction to north, where the latitude is 30°, start the roll-out when the compass reads 37° (30° plus one-half the 15° angle of bank, or whatever amount is appropriate for the rate of roll-out). When turning from an easterly direction to south, start the roll-out when the magnetic compass reads 203° (180° plus 30° minus one-half the angle of bank). When making similar turns from a westerly direction, the appropriate points at which to begin the roll-out would be 323° for a turn to north, and 157° for a turn to south.

30° Bank Turn

A turn using 30° of bank is seldom necessary or advisable in instrument meteorological conditions (IMC), and is considered an unusual attitude in a helicopter. However, it is an excellent maneuver to practice to increase the ability to react quickly and smoothly to rapid changes of attitude. Even though the entry and recovery techniques are the same as for any other turn, it is more difficult to control pitch because of the decrease in vertical lift as the bank increases. Also, because of the decrease in vertical lift, there is a tendency to lose altitude and/or airspeed. Therefore, to maintain a constant altitude and airspeed, additional power is required. Do not initiate a correction, however, until the instruments indicate the need for one. During the maneuver, note the need for a correction on the altimeter and VSI, check the attitude indicator, and then make the necessary adjustments. After making a change, check the altimeter and VSI again to determine whether or not the correction was adequate.

Climbing and Descending Turns

For climbing and descending turns, the techniques described previously for straight climbs, descents, and standard rate turns are combined. For practice, simultaneously turn and start the climb or descent. The primary and supporting instruments for a stabilized constant airspeed left climbing turn are illustrated in *Figure 6-15*. The level off from a climbing or descending turn is the same as the level off from a straight climb or descent. To return to straight-and-level flight, stop the turn and then level off, or level off and then stop the turn, or simultaneously level off and stop the turn. During climbing and descending turns, keep the ball of the turn indicator centered with pedal trim.

Common Errors During Turns

1. Failure to maintain desired turn rate
2. Failure to maintain altitude in level turns
3. Failure to maintain desired airspeed
4. Variation in the rate of entry and recovery
5. Failure to use proper lead in turns to a heading
6. Failure to properly compute time during timed turns
7. Failure to use proper leads and lags during the compass turns
8. Improper use of power
9. Failure to use proper pedal trim

Unusual Attitudes

Any maneuver not required for normal helicopter instrument flight is an unusual attitude and may be caused by any one or combination of factors such as turbulence, disorientation, instrument failure, confusion, preoccupation with flight deck duties, carelessness in cross-checking, errors in instrument interpretation, or lack of proficiency in aircraft control. Due to the instability characteristics of the helicopter, unusual attitudes can be extremely critical. As soon as an unusual attitude is detected, make a recovery to straight-and-level flight as soon as possible with a minimum loss of altitude.

To recover from an unusual attitude, a pilot should correct bank-and-pitch attitude and adjust power as necessary. All components are changed almost simultaneously, with little lead of one over the other. A pilot must be able to perform this task with and without the attitude indicator. If the helicopter is in a climbing or descending turn, adjust bank, pitch, and power. The bank attitude should be corrected by referring to the turn-and-slip indicator and attitude indicator. Pitch attitude should be corrected by reference to the altimeter, airspeed indicator, VSI, and attitude indicator. Adjust power by referring to the airspeed indicator and manifold pressure.

Since the displacement of the controls used in recovery from unusual attitudes may be greater than those used for normal flight, make careful adjustments as straight-and-level flight is approached. Cross-check the other instruments closely to avoid overcontrolling.

Common Errors During Unusual Attitude Recoveries

1. Failure to make proper pitch correction
2. Failure to make proper bank correction
3. Failure to make proper power correction
4. Overcontrolling pitch and/or bank attitude
5. Overcontrolling power
6. Excessive loss of altitude

Emergencies

Emergencies during instrument flight are handled similarly to those occurring during VFR flight. A thorough knowledge of the helicopter and its systems, as well as good aeronautical knowledge and judgment, is the best preparation for emergency situations. Safe operations begin with preflight planning and a thorough preflight inspection. Plan a route of flight to include adequate landing sites in the event of an emergency landing. Make sure all resources, such as maps, publications, flashlights, and fire extinguishers are readily available for use in an emergency.

During any emergency, first fly the aircraft. This means ensure the helicopter is under control, and determine emergency

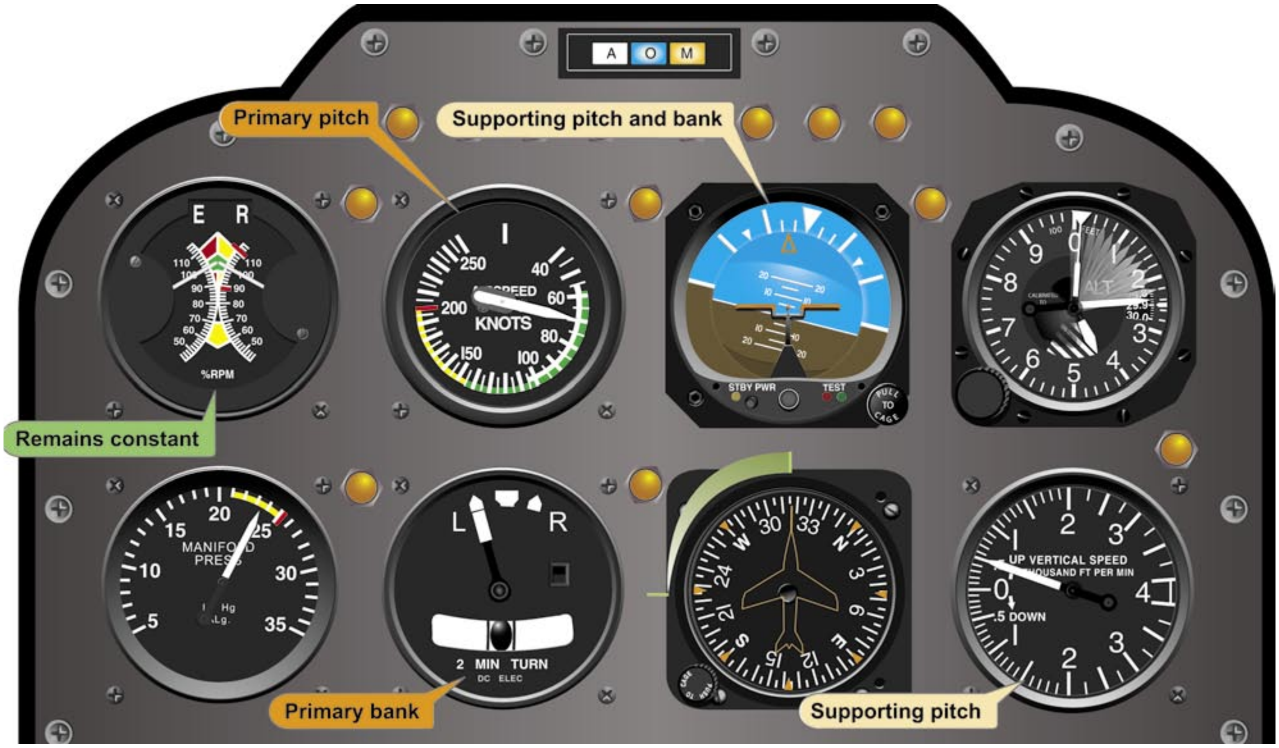

Figure 6-15. *Flight Instrument Indications for a Stabilized Left Climbing Turn at a Constant Airspeed.*

landing sites. Then perform the emergency checklist memory items, followed by items written in the rotorcraft flight manual (RFM). When all these items are under control, notify air traffic control (ATC). Declare any emergency on the last assigned ATC frequency. If one was not issued, transmit on the emergency frequency 121.5. Set the transponder to the emergency squawk code 7700. This code triggers an alarm or special indicator in radar facilities.

When experiencing most in-flight emergencies, such as low fuel or complete electrical failure, land as soon as possible. In the event of an electrical fire, turn off all nonessential equipment and land immediately. Some essential electrical instruments, such as the attitude indicator, may be required for a safe landing. A navigation radio failure may not require an immediate landing if the flight can continue safely. In this case, land as soon as practical. ATC may be able to provide vectors to a safe landing area. For specific details on what to do during an emergency, refer to the RFM for the helicopter.

Autorotations

Both straight-ahead and turning autorotations should be practiced by reference to instruments. This training ensures prompt corrective action to maintain positive aircraft control in the event of an engine failure.

To enter autorotation, reduce collective pitch smoothly to maintain a safe rotor RPM and apply pedal trim to keep the ball of the turn-and-slip indicator centered. The pitch attitude of the helicopter should be approximately level as shown by the attitude indicator. The airspeed indicator is the primary pitch instrument and should be adjusted to the recommended autorotation speed. The heading indicator is primary for bank in a straight-ahead autorotation. In a turning autorotation, a standard rate turn should be maintained by reference to the needle of the turn-and-slip indicator.

Common Errors During Autorotations

1. Uncoordinated entry due to improper pedal trim
2. Poor airspeed control due to improper pitch attitude
3. Poor heading control in straight-ahead autorotations
4. Failure to maintain proper rotor RPM
5. Failure to maintain a standard rate turn during turning autorotations

Servo Failure

Most helicopters certified for single-pilot IFR flight are required to have autopilots, which greatly reduces pilot workload. If an autopilot servo fails, however, resume manual control of the helicopter. The amount of workload increase depends on which servo fails. If a cyclic servo fails, a pilot may want to land immediately because the workload increases tremendously. If an antitorque or collective servo fails, continuing to the next suitable landing site might be possible.

Instrument Takeoff

The procedures and techniques described here should be modified as necessary to conform to those set forth in the operating instructions for the particular helicopter being flown. During training, instrument takeoffs should not be attempted except when receiving instruction from an appropriately certificated, proficient flight instructor pilot.

Adjust the miniature aircraft in the attitude indicator, as appropriate, for the aircraft being flown. After the helicopter is aligned with the runway or takeoff pad, to prevent forward movement of a helicopter equipped with a wheel-type landing gear, set the parking brakes or apply the toe brakes. If the parking brake is used, it must be unlocked after the takeoff has been completed. Apply sufficient friction to the collective pitch control to minimize overcontrolling and to prevent creeping. Excessive friction should be avoided since it limits collective pitch movement.

After checking all instruments for proper indications, start the takeoff by applying collective pitch and a predetermined power setting. Add power smoothly and steadily to gain airspeed and altitude simultaneously and to prevent settling to the ground. As power is applied and the helicopter becomes airborne, use the antitorque pedals initially to maintain the desired heading. At the same time, apply forward cyclic to begin accelerating to climbing airspeed. During the initial acceleration, the pitch attitude of the helicopter, as read on the attitude indicator, should be one- to two-bar widths low. The primary and supporting instruments after becoming airborne are illustrated in *Figure 6-16*. As the airspeed increases to the appropriate climb airspeed, adjust pitch gradually to climb attitude. As climb airspeed is reached, reduce power to the climb power setting and transition to a fully coordinated straight climb.

During the initial climb out, minor heading corrections should be made with pedals only until sufficient airspeed is attained to transition to fully coordinated flight. Throughout the instrument takeoff, instrument cross-check and interpretations must be rapid and accurate, and aircraft control positive and smooth.

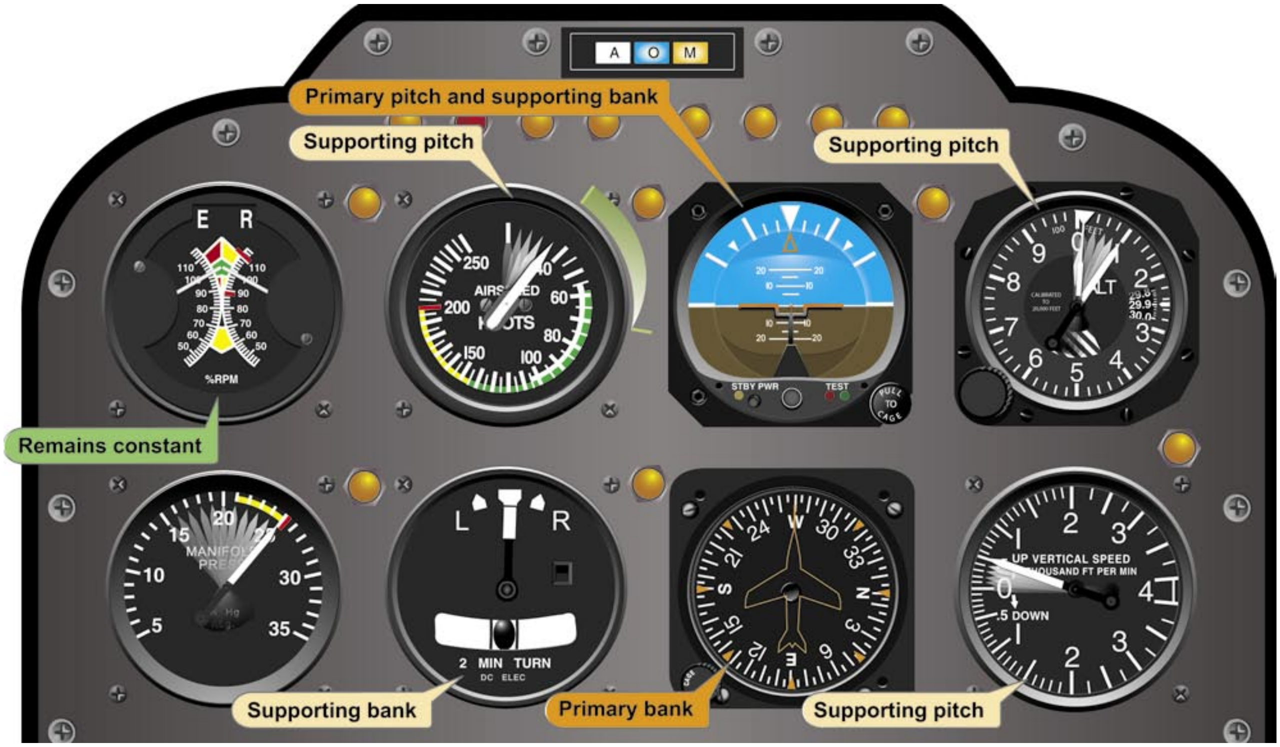
A O M
Primary pitch and supporting bank
Supporting pitch
Supporting pitch
Remains constant
Supporting bank
Primary bank
Supporting pitch

Figure 6-16. *Flight Instrument Indications During an Instrument Takeoff.*

Common Errors During Instrument Takeoffs

1. Failure to maintain heading
2. Overcontrolling pedals
3. Failure to use required power
4. Failure to adjust pitch attitude as climbing airspeed is reached

Changing Technology

Advances in technology have brought about changes in the instrumentation found in all types of aircraft, including helicopters. Electronic displays commonly referred to as "glass cockpits" are becoming more common. Primary flight displays (PFDs) and multi-function displays (MFDs) are changing not only what information is available to a pilot but also how that information is displayed.

Illustrations of technological advancements in instrumentation are described as follows. In *Figure 6-17,* a typical PFD depicts an aircraft flying straight-and-level at 3,000 feet and 100 knots. *Figure 6-18* illustrates a nose-low pitch attitude in a right turn. MFDs can be configured to provide navigation information such as the moving map in *Figure 6-19* or information pertaining to aircraft systems as in *Figure 6-20*.

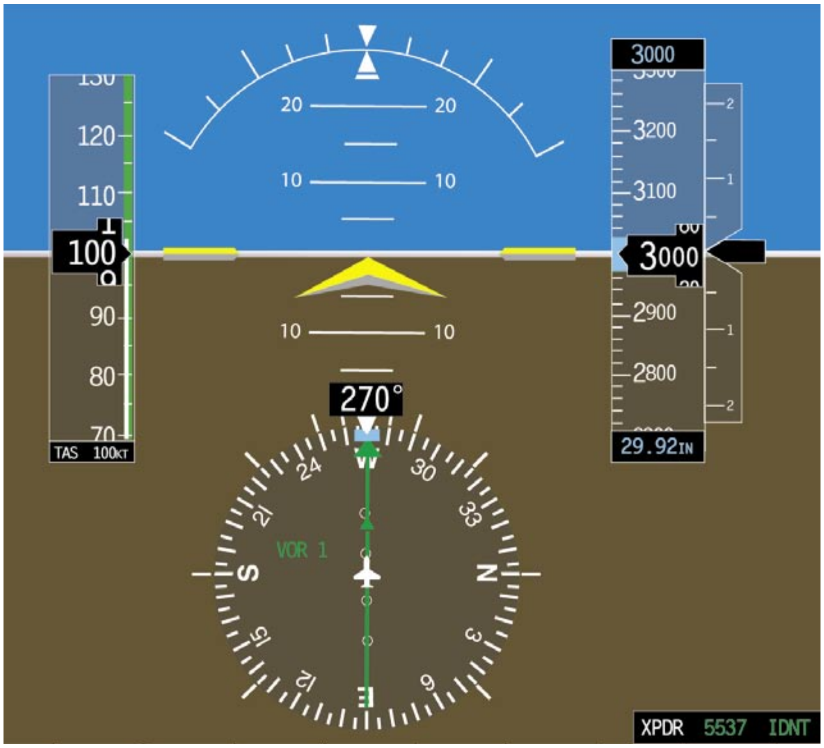
3000
130
120
110
100
90
80
70
TAS 100KT
20
10
3200
3100
3000
2900
2800
29.92IN
270°
VOR 1
XPDR 5537 IDNT

Figure 6-17. *PFD Indications During Straight-and-Level Flight.*

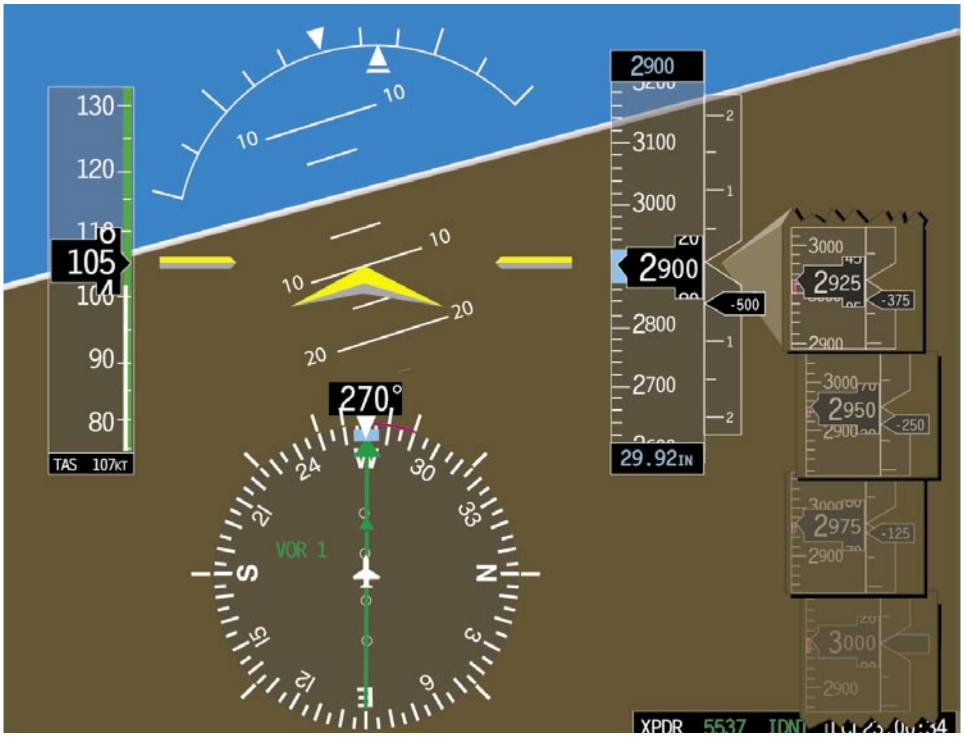
2900
130
120
105
90
80
TAS 107KT
10
20
3100
3000
2900
2800
2700
-500
29.92IN
270°
VOR 1
2925
-375
2950
-250
2975
-125
3000
XPDR 5537

Figure 6-18. *PFD Indications During a Nose-Low Pitch Attitude in a Right Turn.*

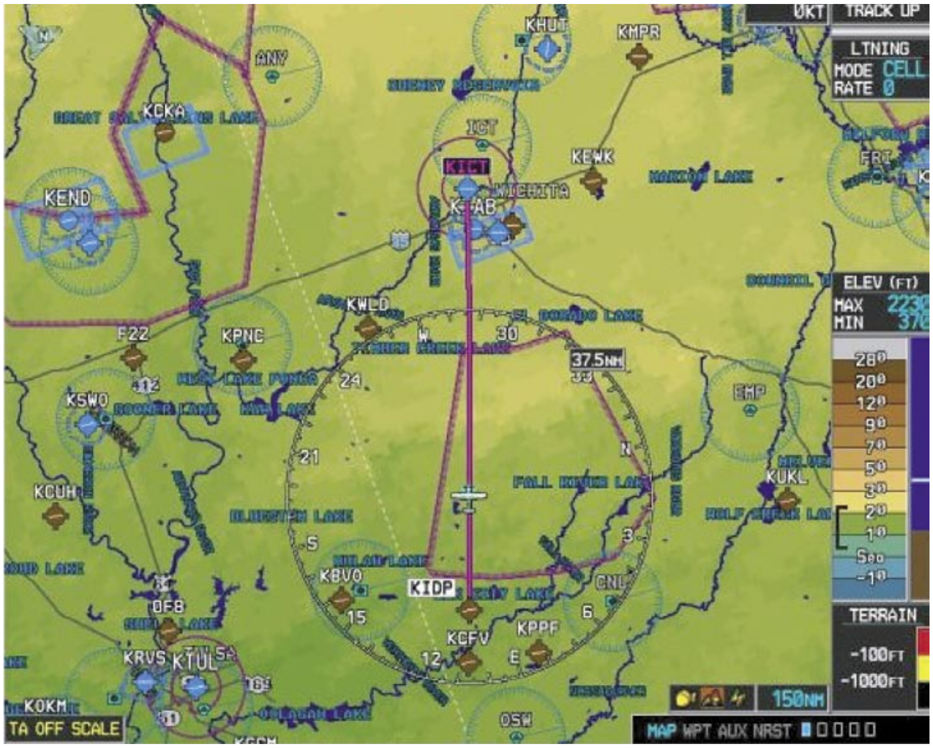
0KT
TRACK UP
LTNING
MODE CELL
RATE 0
KHUT
KMPR
ANY
KCKA
ICT
KICT
WICHITA
KEWK
KEND
KWLD
F22
KPNC
EL DORADO LAKE
37.5NM
KSWO
EMP
FALL RIVER LAKE
KUKL
KCUH
KBVO
KIDP
CNU
KCFV
KPPF
OF8
KRVS
KTUL
KOKM
OSW
TA OFF SCALE
ELEV (FT)
MAX 2230
MIN 370
2800
2000
1200
900
700
500
300
200
100
SEA
-100
TERRAIN
-100FT
-1000FT
150NM
MAP WPT AUX NRST

Figure 6-19. *MFD Display of a Moving Map.*

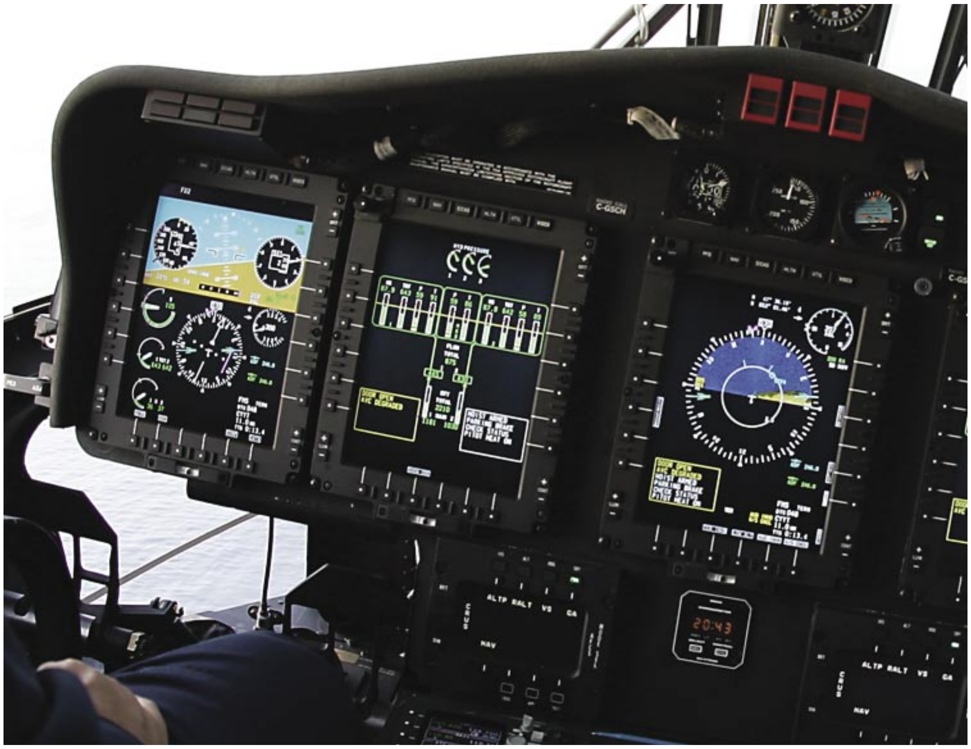
C-GSCN
ALTP RALT VS GA
NAV
20:43

Figure 6-20. *MFD Display of Aircraft Systems.*

Chapter 7

Navigation Systems

Introduction

This chapter provides the basic radio principles applicable to navigation equipment, as well as an operational knowledge of how to use these systems in instrument flight. This information provides the framework for all instrument procedures, including standard instrument departure procedures (SIDS), departure procedures (DPs), holding patterns, and approaches, because each of these maneuvers consists mainly of accurate attitude instrument flying and accurate tracking using navigation systems.

Basic Radio Principles

A radio wave is an electromagnetic (EM) wave with frequency characteristics that make it useful. The wave will travel long distances through space (in or out of the atmosphere) without losing too much strength. An antenna is used to convert electric current into a radio wave so it can travel through space to the receiving antenna, which converts it back into an electric current for use by a receiver.

How Radio Waves Propagate

All matter has a varying degree of conductivity or resistance to radio waves. The Earth itself acts as the greatest resistor to radio waves. Radiated energy that travels near the ground induces a voltage in the ground that subtracts energy from the wave, decreasing the strength of the wave as the distance from the antenna becomes greater. Trees, buildings, and mineral deposits affect the strength to varying degrees. Radiated energy in the upper atmosphere is likewise affected as the energy of radiation is absorbed by molecules of air, water, and dust. The characteristics of radio wave propagation vary according to the signal frequency and the design, use, and limitations of the equipment.

Ground Wave

A ground wave travels across the surface of the Earth. You can best imagine a ground wave's path as being in a tunnel or alley bounded by the surface of the Earth and by the ionosphere, which keeps the ground wave from going out into space. Generally, the lower the frequency, the farther the signal will travel.

Ground waves are usable for navigation purposes because they travel reliably and predictably along the same route day after day, and are not influenced by too many outside factors. The ground wave frequency range is generally from the lowest frequencies in the radio range (perhaps as low as 100 Hz) up to approximately 1,000 kHz (1 MHz). Although there is a ground wave component to frequencies above this, up to 30 MHz, the ground wave at these higher frequencies loses strength over very short distances.

Sky Wave

The sky wave, at frequencies of 1 to 30 MHz, is good for long distances because these frequencies are refracted or "bent" by the ionosphere, causing the signal to be sent back to Earth from high in the sky and received great distances away. *[Figure 7-1]* Used by high frequency (HF) radios in aircraft, messages can be sent across oceans using only 50 to 100 watts of power. Frequencies that produce a sky wave are not used for navigation because the pathway of the signal from transmitter to receiver is highly variable. The wave is "bounced" off of the ionosphere, which is always changing due to the varying amount of the sun's radiation reaching it (night/day and seasonal variations, sunspot activity, etc.). The sky wave is, therefore, unreliable for navigation purposes.

For aeronautical communication purposes, the sky wave (HF) is about 80 to 90 percent reliable. HF is being gradually replaced by more reliable satellite communication.

Space Wave

When able to pass through the ionosphere, radio waves of 15 MHz and above (all the way up to many GHz), are considered space waves. Most navigation systems operate with signals propagating as space waves. Frequencies above 100 MHz have nearly no ground or sky wave components. They are space waves, but (except for global positioning system (GPS)) the navigation signal is used before it reaches the ionosphere so the effect of the ionosphere, which can cause some propagation errors, is minimal. GPS errors caused by passage through the ionosphere are significant and are corrected for by the GPS receiver system.

Space waves have another characteristic of concern to users. Space waves reflect off hard objects and may be blocked if the object is between the transmitter and the receiver. Site and terrain error, as well as propeller/rotor modulation error in very high omnidirectional range (VOR) systems is caused by this bounce. Instrument landing system (ILS) course distortion is also the result of this phenomenon, which led to the need for establishment of ILS critical areas.

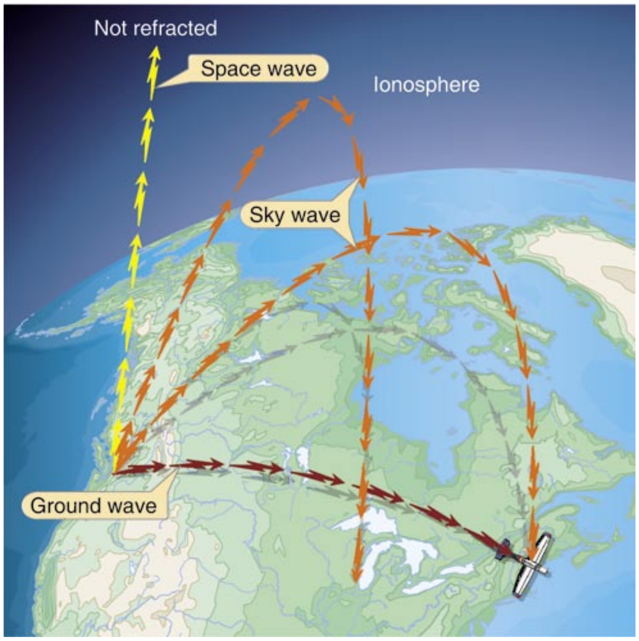

Figure 7-1. *Ground, Space, and Sky Wave Propogation.*

Generally, space waves are "line of sight" receivable, but those of lower frequencies will "bend" somewhat over the horizon. The VOR signal at 108 to 118 MHz is a lower frequency than distance measuring equipment (DME) at 962 to 1213 MHz. Therefore, when an aircraft is flown "over the horizon" from a VOR/DME station, the DME will normally be the first to stop functioning.

Disturbances to Radio Wave Reception

Static distorts the radio wave and interferes with normal reception of communications and navigation signals. Low-frequency airborne equipment such as automatic direction finder (ADF) and LORAN are particularly subject to static disturbance. Using very high frequency (VHF) and ultra-high frequency (UHF) frequencies avoids many of the discharge noise effects. Static noise heard on navigation or communication radio frequencies may be a warning of interference with navigation instrument displays. Some of the problems caused by precipitation static (P-static) are:

- Complete loss of VHF communications.
- Erroneous magnetic compass readings.
- Aircraft flying with one wing low while using the autopilot.
- High-pitched squeal on audio.
- Motorboat sound on audio.
- Loss of all avionics.
- Inoperative very-low frequency (VLF) navigation system.
- Erratic instrument readouts.
- Weak transmissions and poor radio reception.
- St. Elmo's Fire.

Traditional Navigation Systems

Nondirectional Radio Beacon (NDB)

The nondirectional radio beacon (NDB) is a ground-based radio transmitter that transmits radio energy in all directions. The ADF, when used with an NDB, determines the bearing from the aircraft to the transmitting station. The indicator may be mounted in a separate instrument in the aircraft panel. *[Figure 7-2]* The ADF needle points to the NDB ground station to determine the relative bearing (RB) to the transmitting station. It is the number of degrees measured clockwise between the aircraft's heading and the direction from which the bearing is taken. The aircraft's magnetic heading (MH) is the direction the aircraft is pointed with respect to magnetic north. The magnetic bearing (MB) is the direction to or from a radio transmitting station measured relative to magnetic north.

NDB Components

The ground equipment, the NDB, transmits in the frequency range of 190 to 535 kHz. Most ADFs will also tune the AM broadcast band frequencies above the NDB band (550 to 1650 kHz). However, these frequencies are not approved for navigation because stations do not continuously identify themselves, and they are much more susceptible to sky wave propagation especially from dusk to dawn. NDB stations are capable of voice transmission and are often used for transmitting the automated weather observing system (AWOS). The aircraft must be in operational range of the NDB. Coverage depends on the strength of the transmitting station. Before relying on ADF indications, identify the station by listening to the Morse code identifier. NDB stations are usually two letters or an alpha-numeric combination.

ADF Components

The airborne equipment includes two antennas, a receiver, and the indicator instrument. The "sense" antenna (non-directional) receives signals with nearly equal efficiency from all directions. The "loop" antenna receives signals better from two directions (bidirectional). When the loop and sense antenna inputs are processed together in the ADF radio, the result is the ability to receive a radio signal well in all directions but one, thus resolving all directional ambiguity. The indicator instrument can be one of four kinds: fixed-card ADF, rotatable compass-card ADF, or radio magnetic

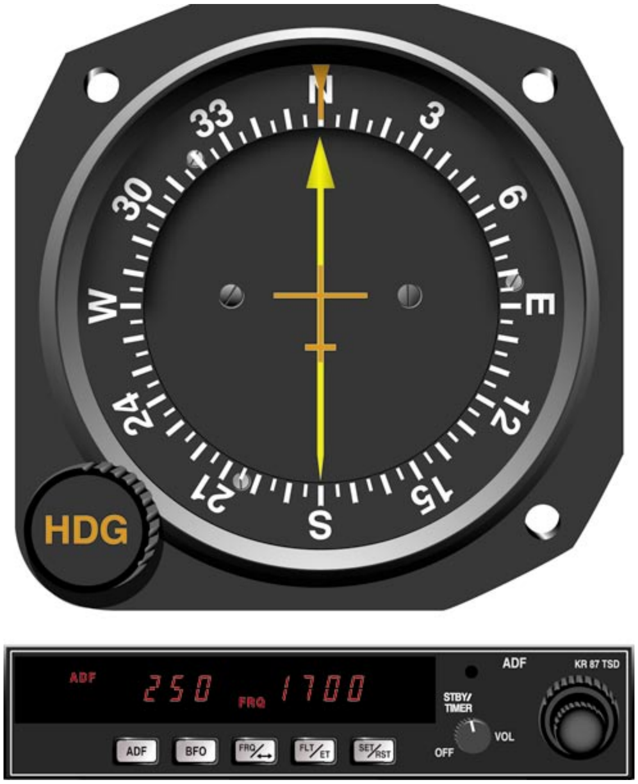

Figure 7-2. *ADF Indicator Instrument and Receiver.*

indicator (RMI) with either one needle or dual needle. Fixed-card ADF (also known as the relative bearing indicator (RBI)) always indicates zero at the top of the instrument, with the needle indicating the RB to the station. *Figure 7-3* indicates an RB of 135°; if the MH is 045°, the MB to the station is 180°. (MH + RB = MB to the station.)

The movable-card ADF allows the pilot to rotate the aircraft's present heading to the top of the instrument so that the head of the needle indicates MB to the station and the tail indicates MB from the station. *Figure 7-4* indicates a heading of 045°, MB to the station of 180°, and MB from the station of 360°.

The RMI differs from the movable-card ADF in that it automatically rotates the azimuth card (remotely controlled by a gyrocompass) to represent aircraft heading. The RMI has two needles, which can be used to indicate navigation information from either the ADF or the VOR receiver. When a needle is being driven by the ADF, the head of the needle indicates the MB TO the station tuned on the ADF receiver. The tail of the needle is the bearing FROM the station. When a needle of the RMI is driven by a VOR receiver, the needle indicates where the aircraft is radially with respect to the VOR station. The needle points to the bearing TO the station, as read on the azimuth card. The tail of the needle points to the radial of the VOR the aircraft is currently on or crossing. *Figure 7-5* indicates a heading of 005°, the MB to the station is 005°, and the MB from the station is 185°.

Function of ADF

The ADF can be used to plot your position, track inbound and outbound, and intercept a bearing. These procedures are used to execute holding patterns and nonprecision instrument approaches.

Orientation

The ADF needle points TO the station, regardless of aircraft heading or position. The RB indicated is thus the angular relationship between the aircraft heading and the station, measured clockwise from the nose of the aircraft. Think of the nose/tail and left/right needle indications, visualizing the ADF dial in terms of the longitudinal axis of the aircraft. When the needle points to 0°, the nose of the aircraft points directly to the station; with the pointer on 210°, the station is 30° to the left of the tail; with the pointer on 090°, the station is off the right wingtip. The RB alone does not indicate aircraft position. The RB must be related to aircraft heading in order to determine direction to or from the station.

Station Passage

When you are near the station, slight deviations from the desired track result in large deflections of the needle. Therefore, it is important to establish the correct drift correction angle as soon as possible. Make small heading corrections (not over 5°) as soon as the needle shows a deviation from course, until it begins to rotate steadily toward a wingtip position or shows erratic left/right oscillations. You

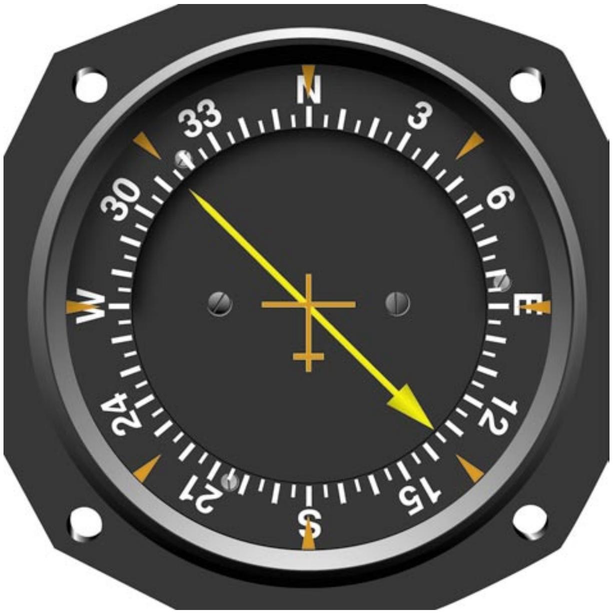

Figure 7-3. *Relative bearing (RB) on a fixed-card indicator. Note that the card always indicates 360°, or north. In this case, the relative bearing to the station is 135° to the right. If the aircraft were on a magnetic heading of 360°, then the magnetic bearing (MB) would also be 135°.*

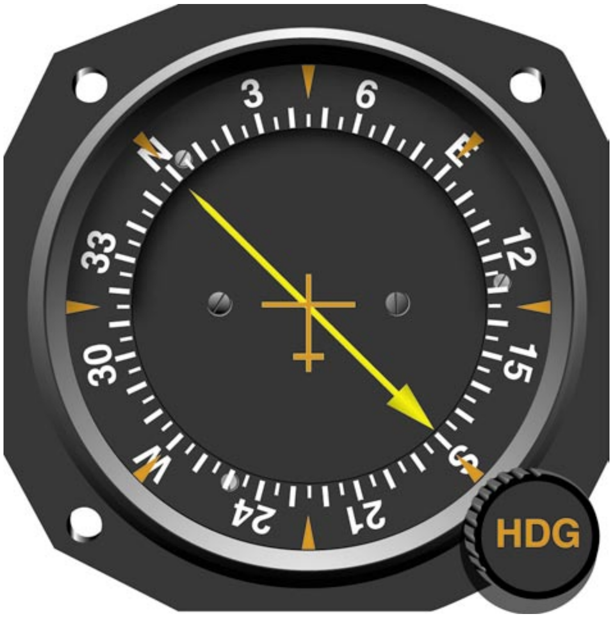

Figure 7-4. *Relative bearing (RB) on a movable-card indicator. By placing the aircraft's magnetic heading (MH) of 045° under the top index, the relative bearing (RB) of 135° to the right will also be the magnetic bearing (no wind conditions) which will take you to the transmitting station.*

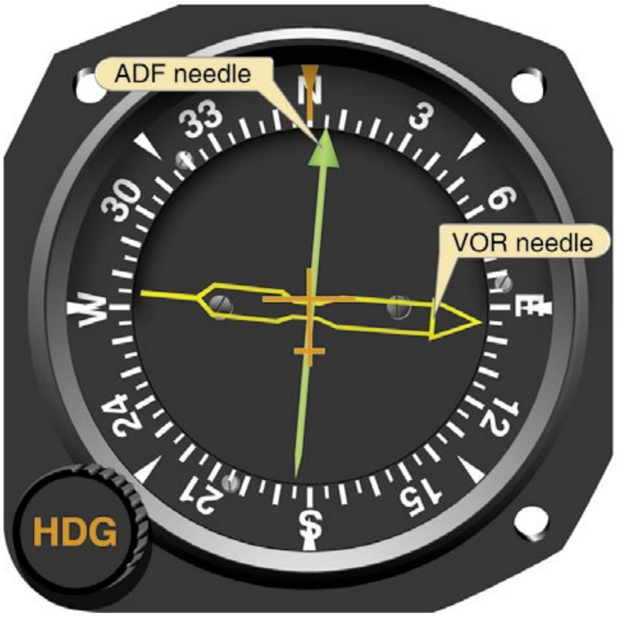

Figure 7-5. *Radio magnetic indicator (RMI). Because the aircraft's magnetic heading is automatically changed, the relative bearing (RB), in this case 095°, will indicate the magnetic bearing (095°) to the station (no wind conditions) and the magnetic heading that will take you there.*

are abeam a station when the needle points 90° off your track. Hold your last corrected heading constant and time station passage when the needle shows either wingtip position or settles at or near the 180° position. The time interval from the first indications of station proximity to positive station passage varies with altitude—a few seconds at low levels to 3 minutes at high altitude.

Homing

The ADF may be used to "home" in on a station. Homing is flying the aircraft on any heading required to keep the needle pointing directly to the 0° RB position. To home in on a station, tune the station, identify the Morse code signal, and then turn the aircraft to bring the ADF azimuth needle to the 0° RB position. Turns should be made using the heading indicator. When the turn is complete, check the ADF needle and make small corrections as necessary.

Figure 7-6 illustrates homing starting from an initial MH of 050° and an RB of 310°, indicating a 50° left turn is needed to produce an RB of zero. Turn left, rolling out at 50° minus 50° equals 360°. Small heading corrections are then made to zero the ADF needle.

If there is no wind, the aircraft will home to the station on a direct track over the ground. With a crosswind, the aircraft will follow a circuitous path to the station on the downwind side of the direct track to the station.

Tracking

Tracking uses a heading that will maintain the desired track to or from the station regardless of crosswind conditions. Interpretation of the heading indicator and needle is done to maintain a constant MB to or from the station.

To track inbound, turn to the heading that will produce a zero RB. Maintain this heading until off-course drift is indicated by displacement of the needle, which will occur if there is a crosswind (needle moving left = wind from the left; needle moving right = wind from the right). A rapid rate of bearing change with a constant heading indicates either a strong crosswind or close proximity to the station or both. When there is a definite (2° to 5°) change in needle reading, turn in the direction of needle deflection to intercept the initial MB. The angle of interception must be greater than the number of degrees of drift, otherwise the aircraft will slowly drift due to the wind pushing the aircraft. If repeated often enough, the track to the station will appear circular and the distance greatly increased as compared to a straight track. The intercept angle depends on the rate of drift, the aircraft speed, and station proximity. Initially, it is standard to double the RB when turning toward your course.

For example, if your heading equals your course and the needle points 10° left, turn 20° left, twice the initial RB. *[Figure 7-7]* This will be your intercept angle to capture the RB. Hold this heading until the needle is deflected 20° in the opposite direction. That is, the deflection of the needle equals the interception angle (in this case 20°). The track has been intercepted, and the aircraft will remain on track as long as the RB remains the same number of degrees as the wind correction angle (WCA), the angle between the desired track and the heading of the aircraft necessary to keep the aircraft tracking over the desired track. Lead the interception to avoid overshooting the track. Turn 10° toward the inbound course. You are now inbound with a 10° left correction angle.

NOTE: In *Figure 7-7*, for the aircraft closest to the station, the WCA is 10° left and the RB is 10° right. If those values do not change, the aircraft will track directly to the station. If you observe off-course deflection in the original direction, turn again to the original interception heading. When the desired course has been re-intercepted, turn 5° toward the inbound course, proceeding inbound with a 15° drift correction. If the initial 10° drift correction is excessive, as shown by needle deflection away from the wind, turn to parallel the desired course and let the wind drift you back on course. When the needle is again zeroed, turn into the wind with a reduced drift correction angle.

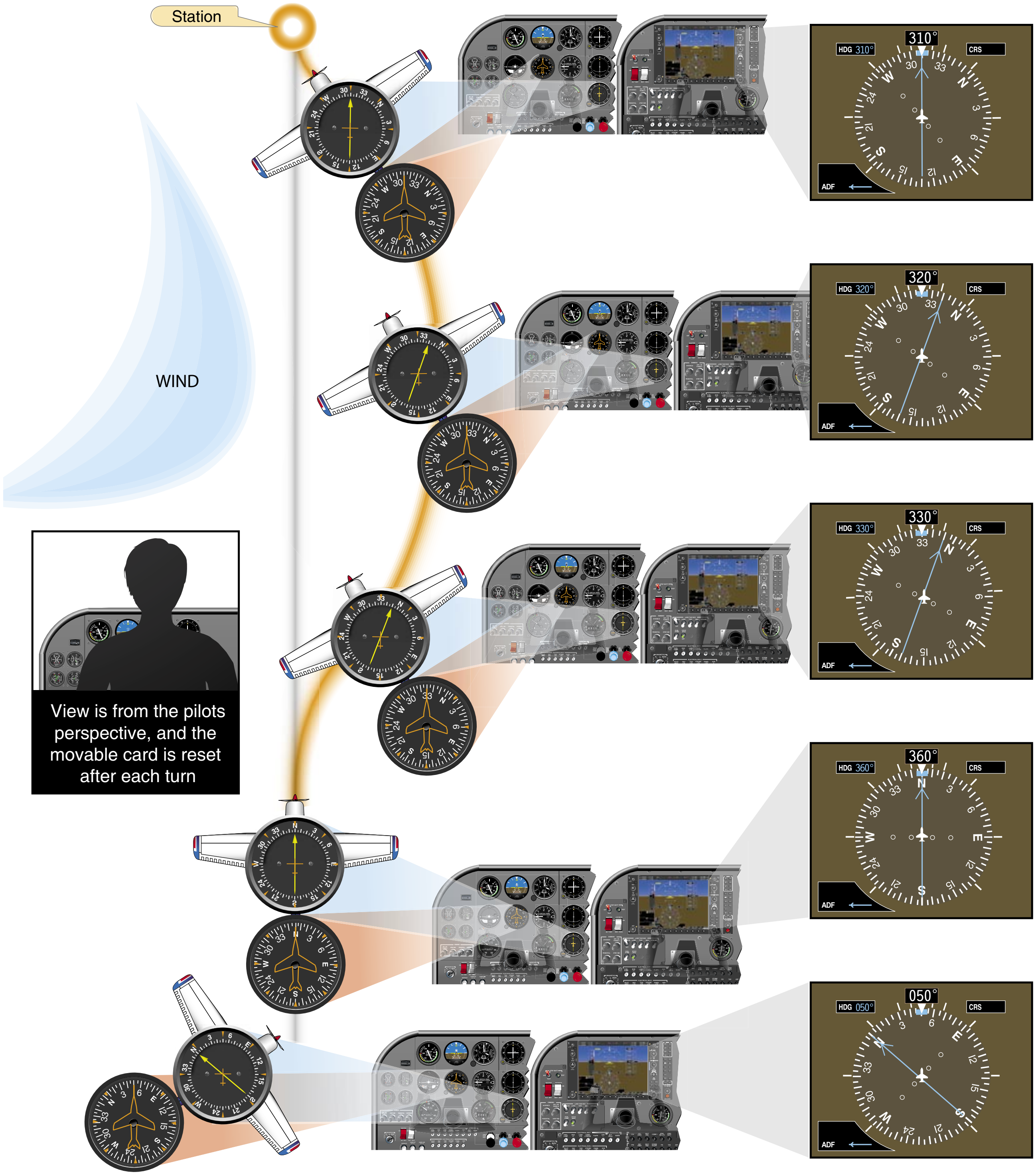
Station
WIND
310°
HDG 310°
CRS
ADF
320°
HDG 320°
CRS
ADF
330°
HDG 330°
CRS
ADF
360°
HDG 360°
CRS
ADF
050°
HDG 050°
CRS
ADF

View is from the pilots perspective, and the movable card is reset after each turn

Figure 7-6. *ADF Homing With a Crosswind.*

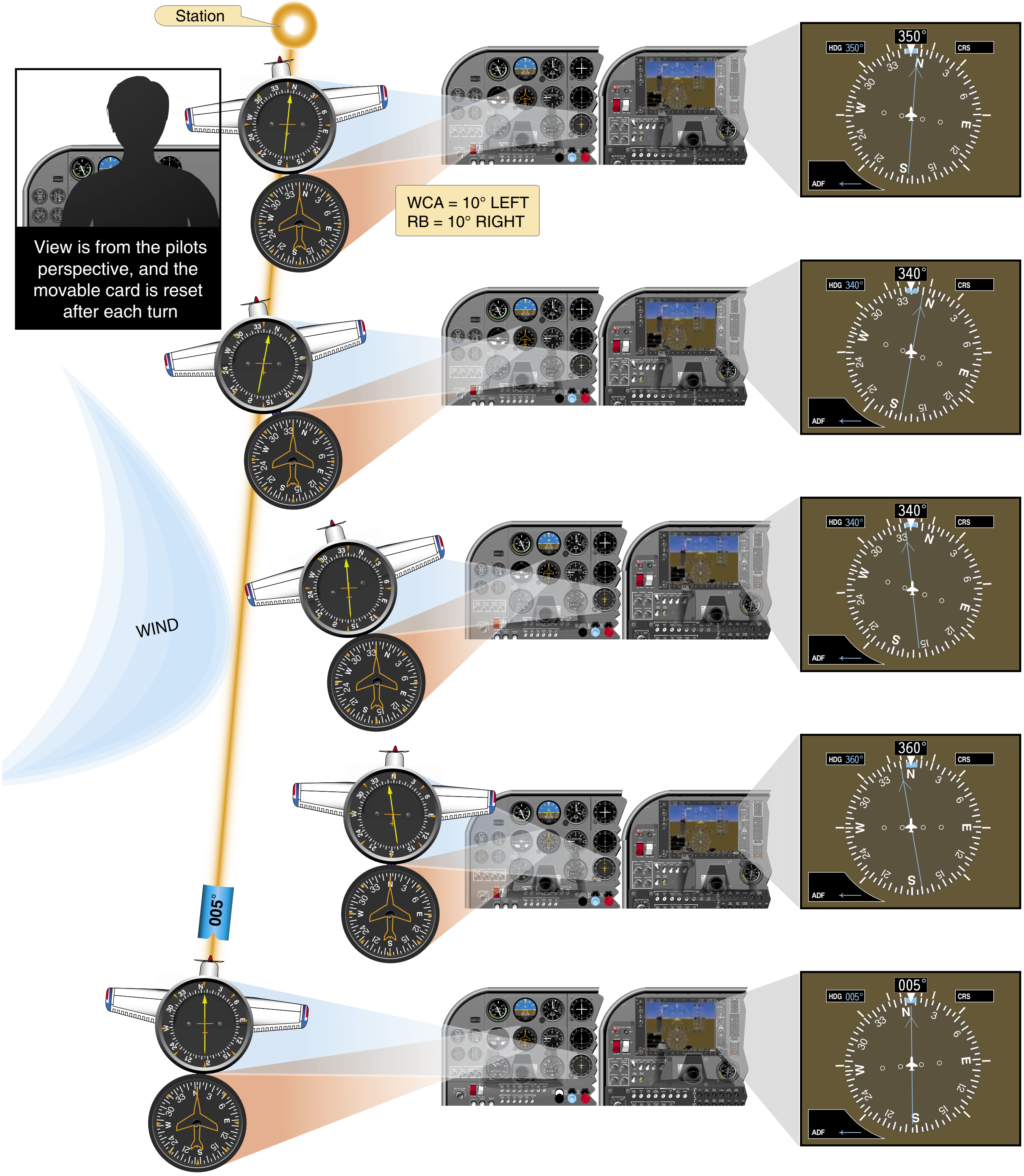
Station
View is from the pilots perspective, and the movable card is reset after each turn
WCA = 10° LEFT
RB = 10° RIGHT
WIND
005°
350°
HDG 350°
CRS
ADF
340°
HDG 340°
CRS
ADF
340°
HDG 340°
CRS
ADF
360°
HDG 360°
CRS
ADF
005°
HDG 005°
CRS
ADF

Figure 7-7. *ADF Tracking Inbound.*

To track outbound, the same principles apply: needle moving left = wind from the left, needle moving right = wind from the right. Wind correction is made toward the needle deflection. The only exception is while the turn to establish the WCA is being made, the direction of the azimuth needle deflections is reversed. When tracking inbound, needle deflection decreases while turning to establish the WCA, and needle deflection increases when tracking outbound. Note the example of course interception and outbound tracking in *Figure 7-8.*

Intercepting Bearings

ADF orientation and tracking procedures may be applied to intercept a specified inbound or outbound MB. To intercept an inbound bearing of 355°, the following steps may be used. *[Figure 7-9]*

1. Determine your position in relation to the station by paralleling the desired inbound bearing. In this case, turn to a heading of 355°. Note that the station is to the right front of the aircraft.
2. Determine the number of degrees of needle deflection from the nose of the aircraft. In this case, the needle's RB from the aircraft's nose is 40° to the right. A rule of thumb for interception is to double this RB amount as an interception angle (80°).
3. Turn the aircraft toward the desired MB the number of degrees determined for the interception angle which as indicated (in two above) is twice the initial RB (40°), or in this case 80°. Therefore, the right turn will be 80° from the initial MB of 355°, or a turn to 075° magnetic (355° + 80° + 075°).
4. Maintain this interception heading of 075° until the needle is deflected the same number of degrees "left" from the zero position as the angle of interception 080°, (minus any lead appropriate for the rate at which the bearing is changing).
5. Turn left 80° and the RB (in a no wind condition and with proper compensation for the rate of the ADF needle movement) should be 0°, or directly off the nose. Additionally, the MB should be 355° indicating proper interception of the desired course.

NOTE: The rate of an ADF needle movement or any bearing pointer for that matter will be faster as aircraft position becomes closer to the station or waypoint (WP).

Interception of an outbound MB can be accomplished by the same procedures as for the inbound intercept, except that it is necessary to substitute the 180° position for the zero position on the needle.

Operational Errors of ADF

Some of the common pilot-induced errors associated with ADF navigation are listed below to help you avoid making the same mistakes. The errors are:

1. Improper tuning and station identification. Many pilots have made the mistake of homing or tracking to the wrong station.
2. Positively identifying any malfunctions of the RMI slaving system or ignoring the warning flag.
3. Dependence on homing rather than proper tracking. This commonly results from sole reliance on the ADF indications, rather than correlating them with heading indications.
4. Poor orientation, due to failure to follow proper steps in orientation and tracking.
5. Careless interception angles, very likely to happen if you rush the initial orientation procedure.
6. Overshooting and undershooting predetermined MBs, often due to forgetting the course interception angles used.
7. Failure to maintain selected headings. Any heading change is accompanied by an ADF needle change. The instruments must be read in combination before any interpretation is made.
8. Failure to understand the limitations of the ADF and the factors that affect its use.
9. Overcontrolling track corrections close to the station (chasing the ADF needle), due to failure to understand or recognize station approach.
10. Failure to keep the heading indicator set so it agrees with the magnetic compass.

Very High Frequency Omnidirectional Range (VOR)

VOR is the primary navigational aid (NAVAID) used by civil aviation in the National Airspace System (NAS). The VOR ground station is oriented to magnetic north and transmits azimuth information to the aircraft, providing 360 courses TO or FROM the VOR station. When DME is installed with the VOR, it is referred to as a VOR/DME and provides both azimuth and distance information. When military tactical air navigation (TACAN) equipment is installed with the VOR, it is known as a VORTAC and provides both azimuth and distance information.

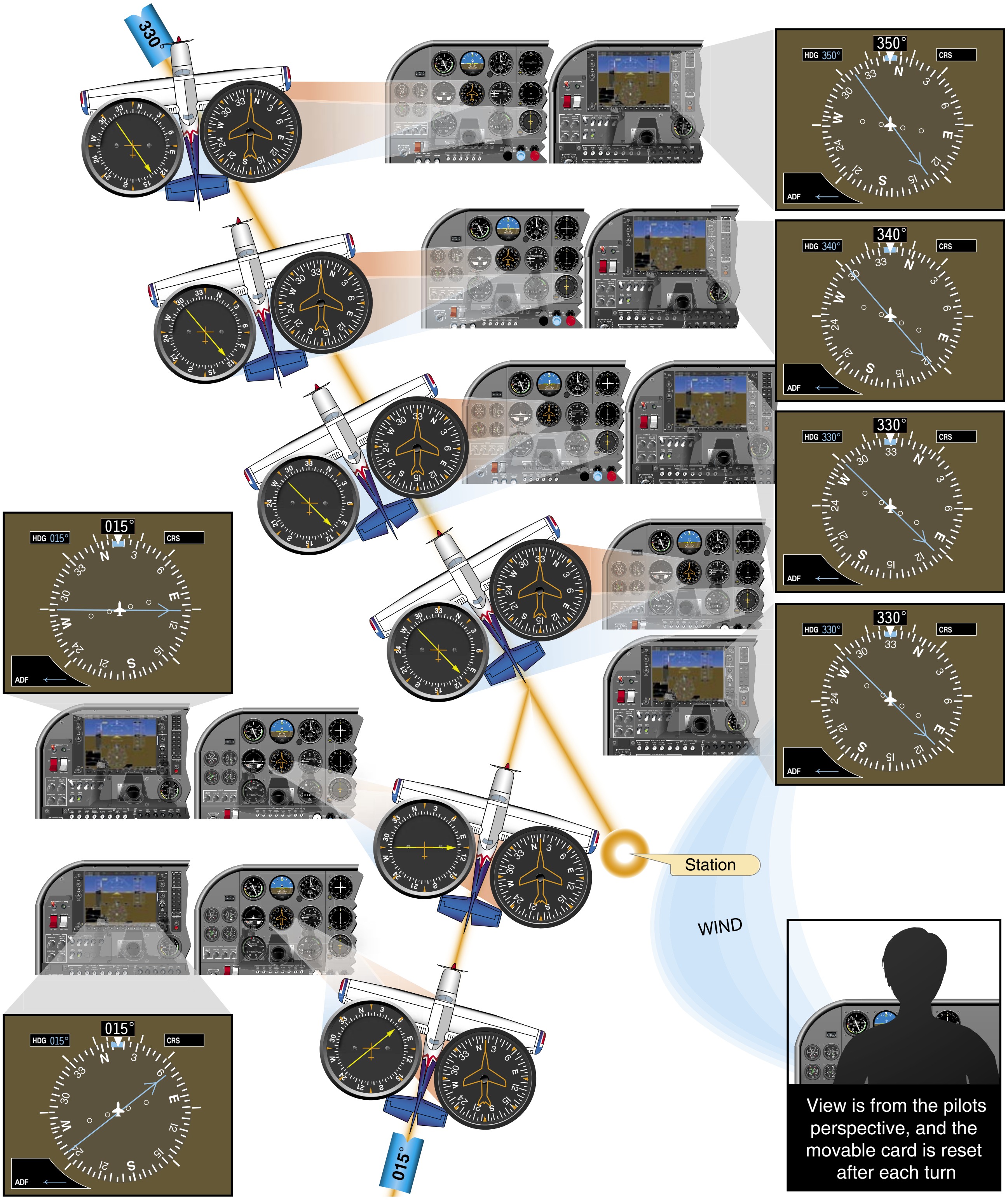
330°
HDG 350°
350°
CRS
ADF
HDG 340°
340°
CRS
ADF
HDG 330°
330°
CRS
ADF
HDG 330°
330°
CRS
ADF
HDG 015°
015°
CRS
ADF
Station
WIND
View is from the pilots perspective, and the movable card is reset after each turn
HDG 015°
015°
CRS
ADF
015°

Figure 7-8. *ADF Interception and Tracking Outbound.*

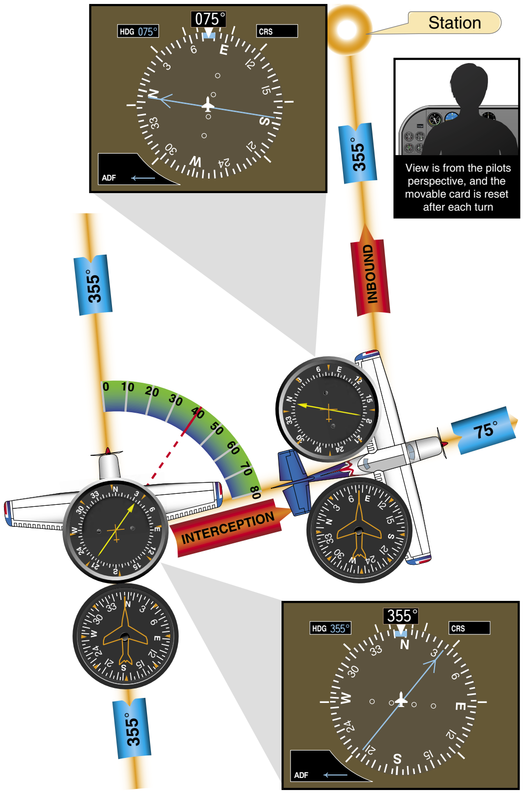

Figure 7-9. *Interception of Bearing.*

The courses oriented FROM the station are called radials. The VOR information received by an aircraft is not influenced by aircraft attitude or heading. *[Figure 7-10]* Radials can be envisioned to be like the spokes of a wheel on which the aircraft is on one specific radial at any time. For example, aircraft A (heading 180°) is inbound on the 360° radial; after crossing the station, the aircraft is outbound on the 180° radial at A1. Aircraft B is shown crossing the 225° radial. Similarly, at any point around the station, an aircraft can be located somewhere on a specific VOR radial. Additionally, a VOR needle on an RMI will always point to the course that will take you to the VOR station where conversely the ADF needle points to the station as a RB from the aircraft. In the example above, the ADF needle at position A would be pointed straight ahead, at A1 to the aircraft's 180° position (tail) and at B, to the aircraft's right.

The VOR receiver measures and presents information to indicate bearing TO or FROM the station. In addition to the navigation signals transmitted by the VOR, a Morse code signal is transmitted concurrently to identify the facility, as well as voice transmissions for communication and relay of weather and other information.

VORs are classified according to their operational uses. The standard VOR facility has a power output of approximately 200 watts, with a maximum usable range depending upon the aircraft altitude, class of facility, location of the facility, terrain conditions within the usable area of the facility, and other factors. Above and beyond certain altitude and distance limits, signal interference from other VOR facilities and a weak signal make it unreliable. Coverage is typically at least 40 miles at normal minimum instrument flight rules (IFR) altitudes. VORs with accuracy problems in parts of their service volume are listed in Notices to Airmen (NOTAMs) and in the Airport/Facility Directory (A/FD) under the name of the NAVAID.

VOR Components

The ground equipment consists of a VOR ground station, which is a small, low building topped with a flat white disc, upon which are located the VOR antennas and a fiberglass cone-shaped tower. *[Figure 7-11]* The station includes an automatic monitoring system. The monitor automatically turns off defective equipment and turns on the standby transmitter. Generally, the accuracy of the signal from the ground station is within 1°.

VOR facilities are aurally identified by Morse code, or voice, or both. The VOR can be used for ground-to-air communication without interference with the navigation signal. VOR facilities operate within the 108.0 to 117.95 MHz frequency band and assignment between 108.0 and 112.0

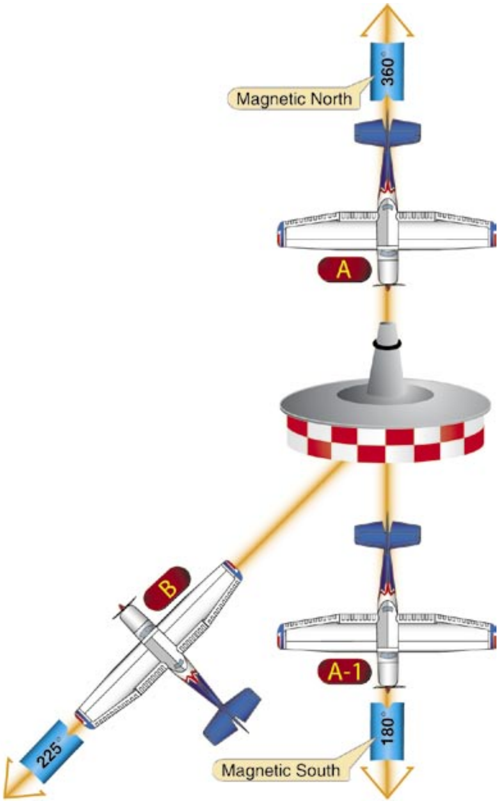

Figure 7-10. *VOR Radials.*

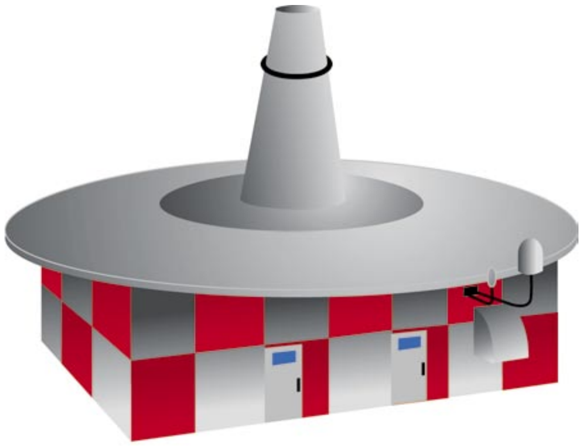

Figure 7-11. *VOR Transmitter (Ground Station).*

MHz is in even-tenth increments to preclude any conflict with ILS localizer frequency assignment, which uses the odd tenths in this range.

The airborne equipment includes an antenna, a receiver, and the indicator instrument. The receiver has a frequency knob to select any of the frequencies between 108.0 to 117.95 MHz. The On/Off/volume control turns on the navigation receiver and controls the audio volume. The volume has no effect on the operation of the receiver. You should listen to the station identifier before relying on the instrument for navigation.

VOR indicator instruments have at least the essential components shown in the instrument illustrated in *Figure 7-12.*

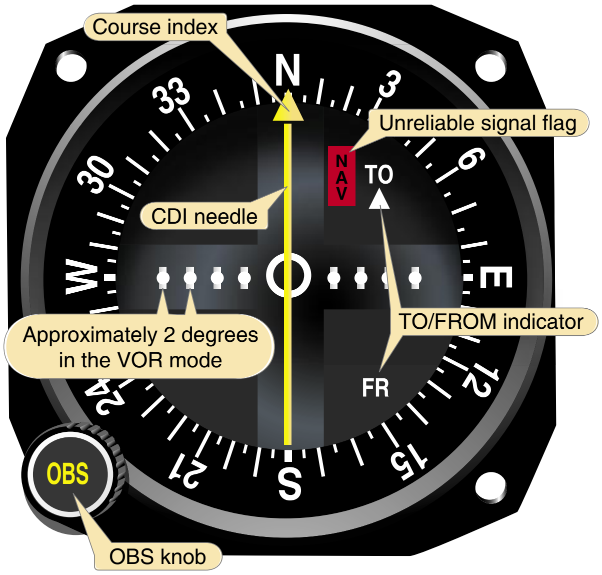
Course index
Unreliable signal flag
N
33
3
30
6
NAV
TO
CDI needle
W
E
Approximately 2 degrees in the VOR mode
TO/FROM indicator
FR
24
12
21
S
15
OBS
OBS knob

Figure 7-12. *The VOR Indicator Instrument.*

Omnibearing Selector (OBS)

The desired course is selected by turning the OBS knob until the course is aligned with the course index mark or displayed in the course window.

Course Deviation Indicator (CDI)

The deviation indicator is composed of an instrument face and a needle hinged to move laterally across the instrument face. The needle centers when the aircraft is on the selected radial or its reciprocal. Full needle deflection from the center position to either side of the dial indicates the aircraft is 12° or more off course, assuming normal needle sensitivity. The outer edge of the center circle is 2° off course; with each dot representing an additional 2°.

TO/FROM Indicator

The TO/FROM indicator shows whether the selected course if intercepted and flown will take the aircraft TO or FROM the station. It does not indicate whether the aircraft is heading to or from the station.

Flags or Other Signal Strength Indicators

The device that indicates a usable or an unreliable signal may be an "OFF" flag. It retracts from view when signal strength is sufficient for reliable instrument indications. Alternately, insufficient signal strength may be indicated by a blank or OFF in the TO/FROM window.

The indicator instrument may also be a horizontal situation indicator (HSI) which combines the heading indicator and CDI. *[Figure 7-13]* The combination of navigation information from VOR/Localizer (LOC) or from LORAN or GPS, with aircraft heading information provides a visual picture of the aircraft's location and direction. This decreases pilot workload especially with tasks such as course intercepts, flying a back-course approach, or holding pattern entry. (See

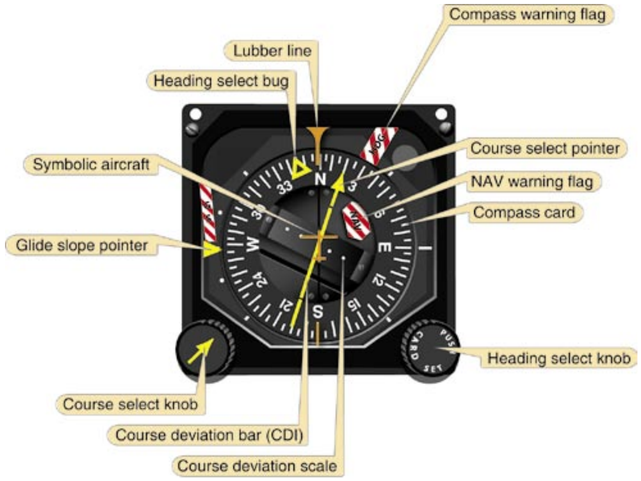
Compass warning flag
Lubber line
Heading select bug
Symbolic aircraft
Course select pointer
NAV warning flag
Compass card
Glide slope pointer
N
33
3
W
E
24
12
S
15
PUSH
CARD SET
Heading select knob
Course select knob
Course deviation bar (CDI)
Course deviation scale

Figure 7-13. *A Typical Horizontal Situation Indicator (HSI).*

Chapter 3, Flight Instruments, for operational characteristics.) *[Figure 7-14]*

Function of VOR

Orientation

The VOR does not account for the aircraft heading. It only relays the aircraft direction from the station and will have the same indications regardless of which way the nose is pointing. Tune the VOR receiver to the appropriate frequency of the selected VOR ground station, turn up the audio volume, and identify the station's signal audibly. Then, rotate the OBS to center the CDI needle and read the course under or over the index.

In *Figure 7-12*, 360° TO is the course indicated, while in *Figure 7-15*, 180° TO is the course. The latter indicates that the aircraft (which may be heading in any direction) is, at this moment, located at any point on the 360° radial (line from the station) except directly over the station or very close to it, as in *Figure 7-15*. The CDI will deviate from side to side as the aircraft passes over or nearly over the station because of the volume of space above the station where the zone of confusion exists. This zone of confusion is caused by lack of adequate signal directly above the station due to the radiation pattern of the station's antenna, and because the resultant of the opposing reference and variable signals is small and constantly changing.

The CDI in *Figure 7-15* indicates 180°, meaning that the aircraft is on the 180° or the 360° radial of the station. The TO/FROM indicator resolves the ambiguity. If the TO indicator is showing, then it is 180° TO the station. The FROM indication indicates the radial of the station the aircraft is presently on. Movement of the CDI from center, if it occurs at a relatively constant rate, indicates the aircraft is moving or drifting off the 180°/360° line. If the movement is rapid or fluctuating, this is an indication of impending station passage (the aircraft is near the station). To determine the aircraft's position relative to the station, rotate the OBS until FROM appears in the window, and then center the CDI needle. The index indicates the VOR radial where the aircraft is located. The inbound (to the station) course is the reciprocal of the radial.

If the VOR is set to the reciprocal of the intended course, the CDI will reflect reverse sensing. To correct for needle deflection, turn away from the needle. To avoid this reverse sensing situation, set the VOR to agree with the intended course.

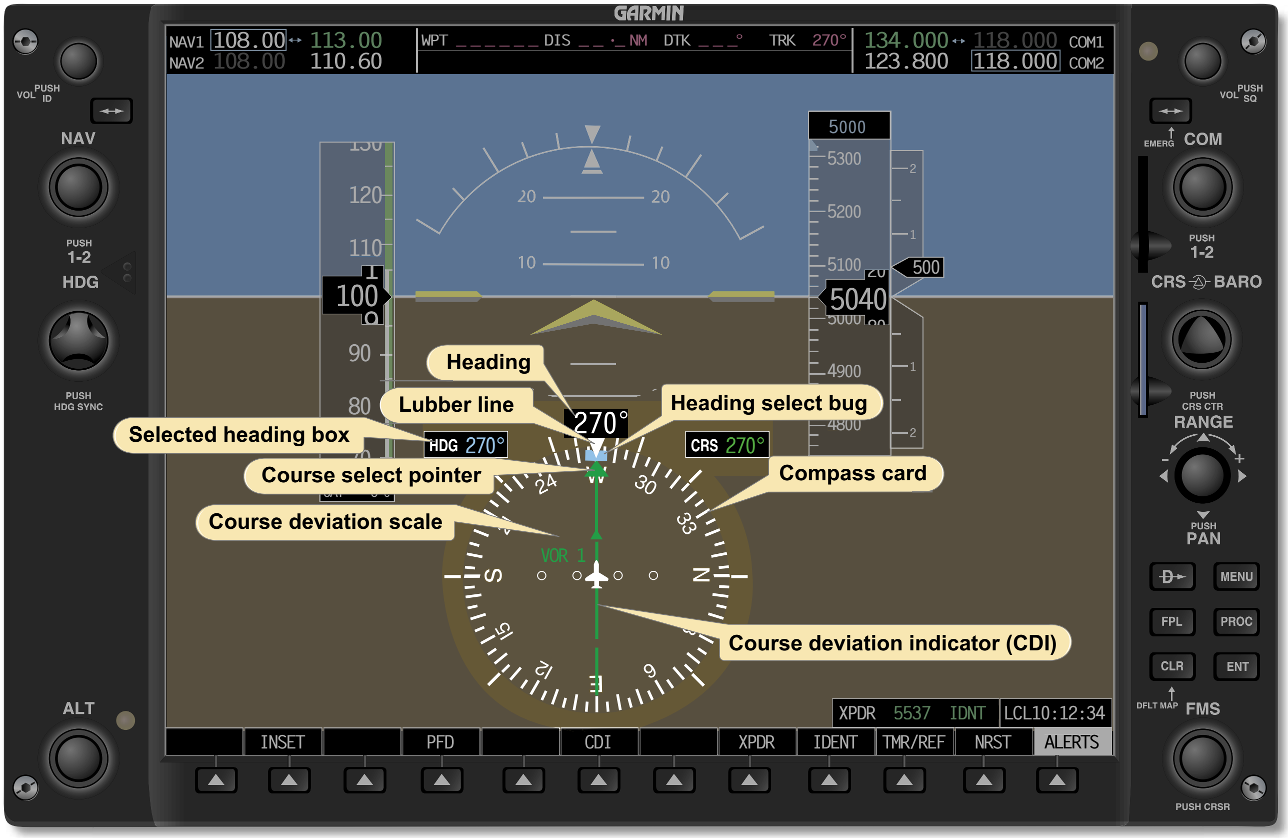
GARMIN
NAV1 108.00 113.00
NAV2 108.00 110.60
WPT DIS NM DTK TRK 270°
134.000 118.000 COM1
123.800 118.000 COM2
Heading
Lubber line
Heading select bug
Selected heading box
HDG 270°
270°
CRS 270°
Course select pointer
Compass card
Course deviation scale
VOR 1
Course deviation indicator (CDI)
XPDR 5537 IDNT LCL10:12:34
INSET
PFD
CDI
XPDR
IDENT
TMR/REF
NRST
ALERTS
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS

Figure 7-14. *An HSI display as seen on the pilot's primary flight display (PFD) on an electronic flight instrument. Note that only attributes related to the HSI are labeled.*

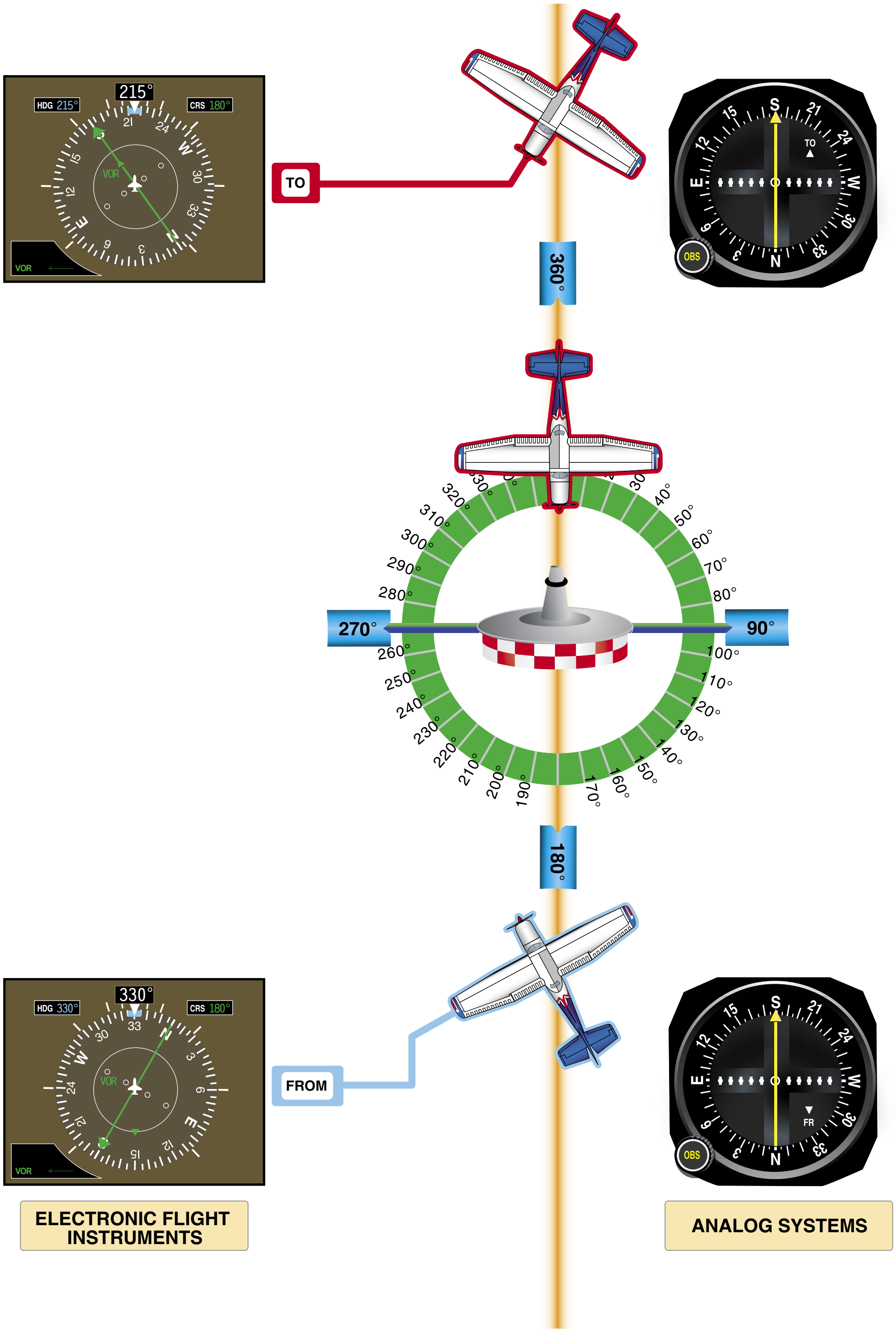
HDG 215°
215°
CRS 180°
VOR
TO
360°
TO
OBS
270°
90°
180°
FROM
HDG 330°
330°
CRS 180°
VOR
FR
OBS
ELECTRONIC FLIGHT INSTRUMENTS
ANALOG SYSTEMS

Figure 7-15. *CDI Interpretation. The CDI as typically found on analog systems (right) and as found on electronic flight instruments (left).*

A single NAVAID will allow a pilot to determine the aircraft's position relative to a radial. Indications from a second NAVAID are needed in order to narrow the aircraft's position down to an exact location on this radial.

Tracking TO and FROM the Station

To track to the station, rotate the OBS until TO appears, then center the CDI. Fly the course indicated by the index. If the CDI moves off center to the left, follow the needle by correcting course to the left, beginning with a 20° correction.

When flying the course indicated on the index, a left deflection of the needle indicates a crosswind component from the left. If the amount of correction brings the needle back to center, decrease the left course correction by half. If the CDI moves left or right now, it should do so much more slowly, and smaller heading corrections can be made for the next iteration.

Keeping the CDI centered will take the aircraft to the station. To track to the station, the OBS value at the index is not changed. To home to the station, the CDI needle is periodically centered, and the new course under the index is used for the aircraft heading. Homing will follow a circuitous route to the station, just as with ADF homing.

To track FROM the station on a VOR radial, you should first orient the aircraft's location with respect to the station and the desired outbound track by centering the CDI needle with a FROM indication. The track is intercepted by either flying over the station or establishing an intercept heading. The magnetic course of the desired radial is entered under the index using the OBS and the intercept heading held until the CDI centers. Then the procedure for tracking to the station is used to fly outbound on the specified radial.

Course Interception

If the desired course is not the one being flown, first orient the aircraft's position with respect to the VOR station and the course to be flown, and then establish an intercept heading. The following steps may be used to intercept a predetermined course, either inbound or outbound. Steps 1–3 may be omitted when turning directly to intercept the course without initially turning to parallel the desired course.

1. Determine the difference between the radial to be intercepted and the radial on which the aircraft is located (205° – 160° = 045°).
2. Double the difference to determine the interception angle, which will not be less than 20° nor greater than 90° (45° x 2 = 090°). 205° + 090° = 295° for the intercept)
3. Rotate the OBS to the desired radial or inbound course.
4. Turn to the interception heading.
5. Hold this heading constant until the CDI center, which indicates the aircraft is on course. (With practice in judging the varying rates of closure with the course centerline, pilots learn to lead the turn to prevent overshooting the course.)
6. Turn to the MH corresponding to the selected course, and follow tracking procedures inbound or outbound.

Course interception is illustrated in *Figure 7-16*.

VOR Operational Errors

Typical pilot-induced errors include:

1. Careless tuning and identification of station.
2. Failure to check receiver for accuracy/sensitivity.
3. Turning in the wrong direction during an orientation. This error is common until visualizing position rather than heading.
4. Failure to check the ambiguity (TO/FROM) indicator, particularly during course reversals, resulting in reverse sensing and corrections in the wrong direction.
5. Failure to parallel the desired radial on a track interception problem. Without this step, orientation to the desired radial can be confusing. Since pilots think in terms of left and right of course, aligning the aircraft position to the radial/course is essential.
6. Overshooting and undershooting radials on interception problems.
7. Overcontrolling corrections during tracking, especially close to the station.
8. Misinterpretation of station passage. On VOR receivers not equipped with an ON/OFF flag, a voice transmission on the combined communication and navigation radio (NAV/COM) in use for VOR may cause the same TO/FROM fluctuations on the ambiguity meter as shown during station passage. Read the whole receiver—TO/FROM, CDI, and OBS—before you make a decision. Do not utilize a VOR reading observed while transmitting.
9. Chasing the CDI, resulting in homing instead of tracking. Careless heading control and failure to bracket wind corrections make this error common.

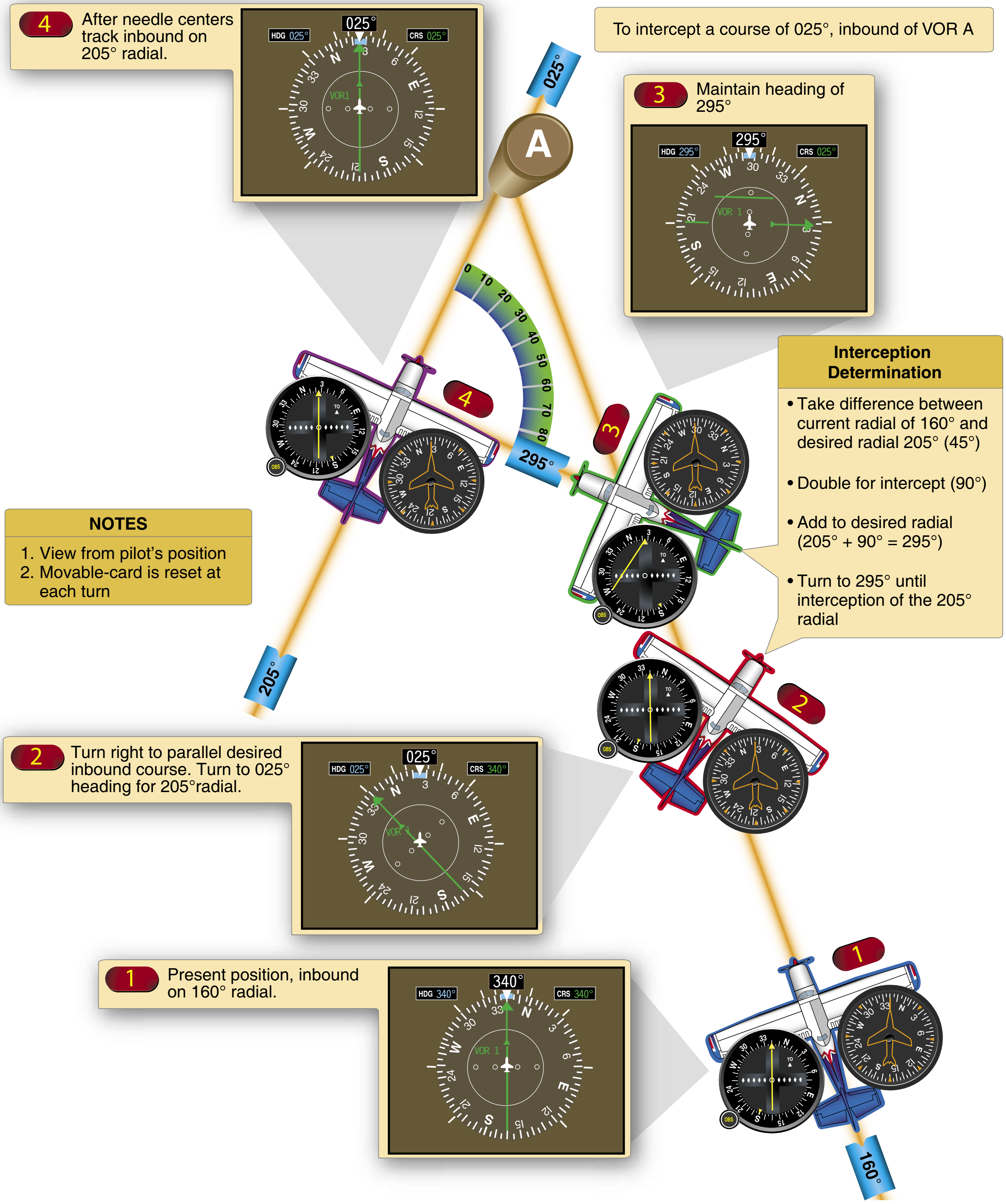
025°
A
205°
295°
160°
0
10
20
30
40
50
60
70
80
4
3
2
1
HDG 025°
025°
CRS 025°
VOR1
HDG 295°
295°
CRS 025°
VOR 1
HDG 025°
025°
CRS 340°
VOR 1
HDG 340°
340°
CRS 340°
VOR 1
OBS
TO

To intercept a course of 025°, inbound of VOR A

4 After needle centers track inbound on 205° radial.

3 Maintain heading of 295°

Interception Determination

- Take difference between current radial of 160° and desired radial 205° (45°)
- Double for intercept (90°)
- Add to desired radial (205° + 90° = 295°)
- Turn to 295° until interception of the 205° radial

NOTES

1. View from pilot's position
2. Movable-card is reset at each turn

2 Turn right to parallel desired inbound course. Turn to 025° heading for 205°radial.

1 Present position, inbound on 160° radial.

Figure 7-16. *Course Interception (VOR).*

VOR Accuracy

The effectiveness of the VOR depends upon proper use and adjustment of both ground and airborne equipment.

The accuracy of course alignment of the VOR is generally plus or minus 1°. On some VORs, minor course roughness may be observed, evidenced by course needle or brief flag alarm. At a few stations, usually in mountainous terrain, the pilot may occasionally observe a brief course needle oscillation, similar to the indication of "approaching station." Pilots flying over unfamiliar routes are cautioned to be on the alert for these vagaries, and in particular, to use the TO/FROM indicator to determine positive station passage.

Certain propeller revolutions per minute (RPM) settings or helicopter rotor speeds can cause the VOR CDI to fluctuate as much as plus or minus 6°. Slight changes to the RPM setting will normally smooth out this roughness. Pilots are urged to check for this modulation phenomenon prior to reporting a VOR station or aircraft equipment for unsatisfactory operation.

VOR Receiver Accuracy Check

VOR system course sensitivity may be checked by noting the number of degrees of change as the OBS is rotated to move the CDI from center to the last dot on either side. The course selected should not exceed 10° or 12° either side. In addition, Title 14 of the Code of Federal Regulations (14 CFR) part 91 provides for certain VOR equipment accuracy checks, and an appropriate endorsement, within 30 days prior to flight under IFR. To comply with this requirement and to ensure satisfactory operation of the airborne system, use the following means for checking VOR receiver accuracy:

1. VOR test facility (VOT) or a radiated test signal from an appropriately rated radio repair station.
2. Certified checkpoints on the airport surface.
3. Certified airborne checkpoints.

VOR Test Facility (VOT)

The Federal Aviation Administration (FAA) VOT transmits a test signal which provides users a convenient means to determine the operational status and accuracy of a VOR receiver while on the ground where a VOT is located. Locations of VOTs are published in the A/FD. Two means of identification are used. One is a series of dots and the other is a continuous tone. Information concerning an individual test signal can be obtained from the local flight service station (FSS.) The airborne use of VOT is permitted; however, its use is strictly limited to those areas/altitudes specifically authorized in the A/FD or appropriate supplement.

To use the VOT service, tune in the VOT frequency 108.0 MHz on the VOR receiver. With the CDI centered, the OBS should read 0° with the TO/FROM indication showing FROM or the OBS should read 180° with the TO/FROM indication showing TO. Should the VOR receiver operate an RMI, it would indicate 180° on any OBS setting.

A radiated VOT from an appropriately rated radio repair station serves the same purpose as an FAA VOT signal, and the check is made in much the same manner as a VOT with some differences.

The frequency normally approved by the Federal Communications Commission (FCC) is 108.0 MHz; however, repair stations are not permitted to radiate the VOR test signal continuously. The owner or operator of the aircraft must make arrangements with the repair station to have the test signal transmitted. A representative of the repair station must make an entry into the aircraft logbook or other permanent record certifying to the radial accuracy and the date of transmission.

Certified Checkpoints

Airborne and ground checkpoints consist of certified radials that should be received at specific points on the airport surface or over specific landmarks while airborne in the immediate vicinity of the airport. Locations of these checkpoints are published in the A/FD.

Should an error in excess of ±4° be indicated through use of a ground check, or ±6° using the airborne check, IFR flight shall not be attempted without first correcting the source of the error. No correction other than the correction card figures supplied by the manufacturer should be applied in making these VOR receiver checks.

If a dual system VOR (units independent of each other except for the antenna) is installed in the aircraft, one system may be checked against the other. Turn both systems to the same VOR ground facility and note the indicated bearing to that station. The maximum permissible variation between the two indicated bearings is 4°.

Distance Measuring Equipment (DME)

When used in conjunction with the VOR system, DME makes it possible for pilots to determine an accurate geographic position of the aircraft, including the bearing and distance TO or FROM the station. The aircraft DME transmits interrogating radio frequency (RF) pulses, which are received by the DME antenna at the ground facility. The signal triggers ground receiver equipment to respond to the interrogating aircraft. The airborne DME equipment measures the elapsed time between the interrogation signal sent by the aircraft and reception of the reply pulses from the ground station. This time measurement is converted into distance in nautical miles (NM) from the station.

Some DME receivers provide a groundspeed in knots by monitoring the rate of change of the aircraft's position relative to the ground station. Groundspeed values are accurate only when tracking directly to or from the station.

DME Components

VOR/DME, VORTAC, ILS/DME, and LOC/DME navigation facilities established by the FAA provide course and distance information from collocated components under a frequency pairing plan. DME operates on frequencies in the UHF spectrum between 962 MHz and 1213 MHz. Aircraft receiving equipment which provides for automatic DME selection assures reception of azimuth and distance information from a common source when designated VOR/DME, VORTAC, ILS/DME, and LOC/DME are selected. Some aircraft have separate VOR and DME receivers, each of which must be tuned to the appropriate navigation facility. The airborne equipment includes an antenna and a receiver.

The pilot-controllable features of the DME receiver include:

Channel (Frequency) Selector
Many DMEs are channeled by an associated VHF radio, or there may be a selector switch so a pilot can select which VHF radio is channeling the DME. For a DME with its own frequency selector, use the frequency of the associated VOR/DME or VORTAC station.

On/Off/Volume Switch
The DME identifier will be heard as a Morse code identifier with a tone somewhat higher than that of the associated VOR or LOC. It will be heard once for every three or four times the VOR or LOC identifier is heard. If only one identifier is heard about every 30 seconds, the DME is functional, but the associated VOR or LOC is not.

Mode Switch
The mode switch selects between distance (DIST) or distance in NMs, groundspeed, and time to station. There may also be one or more HOLD functions which permit the DME to stay channeled to the station that was selected before the switch was placed in the hold position. This is useful when you make an ILS approach at a facility that has no collocated DME, but there is a VOR/DME nearby.

Altitude
Some DMEs correct for slant-range error.

Function of DME

A DME is used for determining the distance from a ground DME transmitter. Compared to other VHF/UHF NAVAIDs, a DME is very accurate. The distance information can be used to determine the aircraft position or flying a track that is a constant distance from the station. This is referred to as a DME arc.

DME Arc

There are many instrument approach procedures (IAPs) that incorporate DME arcs. The procedures and techniques given here for intercepting and maintaining such arcs are applicable to any facility that provides DME information. Such a facility may or may not be collocated with the facility that provides final approach guidance.

As an example of flying a DME arc, refer to *Figure 7-17* and follow these steps:

1. Track inbound on the OKT 325° radial, frequently checking the DME mileage readout.
2. A 0.5 NM lead is satisfactory for groundspeeds of 150 knots or less; start the turn to the arc at 10.5 miles. At higher groundspeeds, use a proportionately greater lead.

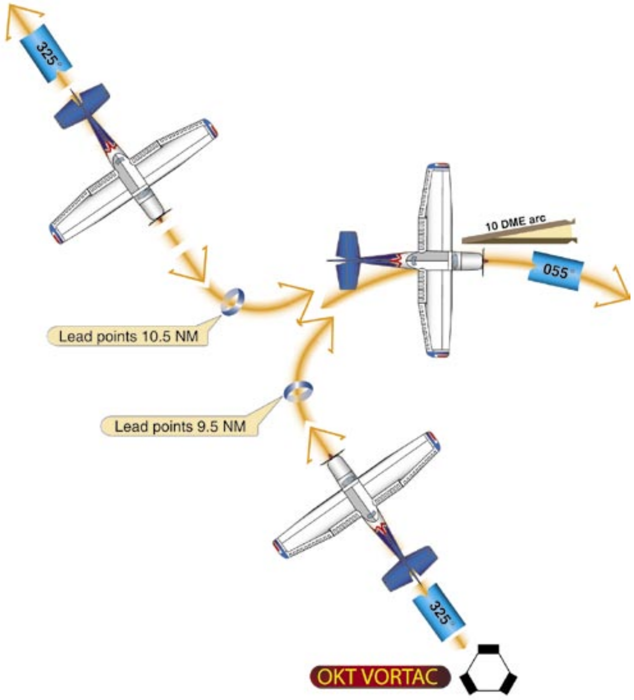

Figure 7-17. *DME Arc Interception.*

3. Continue the turn for approximately 90°. The roll-out heading will be 055° in a no wind condition.
4. During the last part of the intercepting turn, monitor the DME closely. If the arc is being overshot (more than 1.0 NM), continue through the originally planned roll-out heading. If the arc is being undershot, roll-out of the turn early.

The procedure for intercepting the 10 DME when outbound is basically the same, the lead point being 10 NM minus 0.5 NM, or 9.5 NM.

When flying a DME arc with wind, it is important to keep a continuous mental picture of the aircraft's position relative to the facility. Since the wind-drift correction angle is constantly changing throughout the arc, wind orientation is important. In some cases, wind can be used in returning to the desired track. High airspeeds require more pilot attention because of the higher rate of deviation and correction.

Maintaining the arc is simplified by keeping slightly inside the curve; thus, the arc is turning toward the aircraft and interception may be accomplished by holding a straight course. When outside the curve, the arc is "turning away" and a greater correction is required.

To fly the arc using the VOR CDI, center the CDI needle upon completion of the 90° turn to intercept the arc. The aircraft's heading will be found very near the left or right side (270° or 90° reference points) of the instrument. The readings at that side location on the instrument will give primary heading information while on the arc. Adjust the aircraft heading to compensate for wind and to correct for

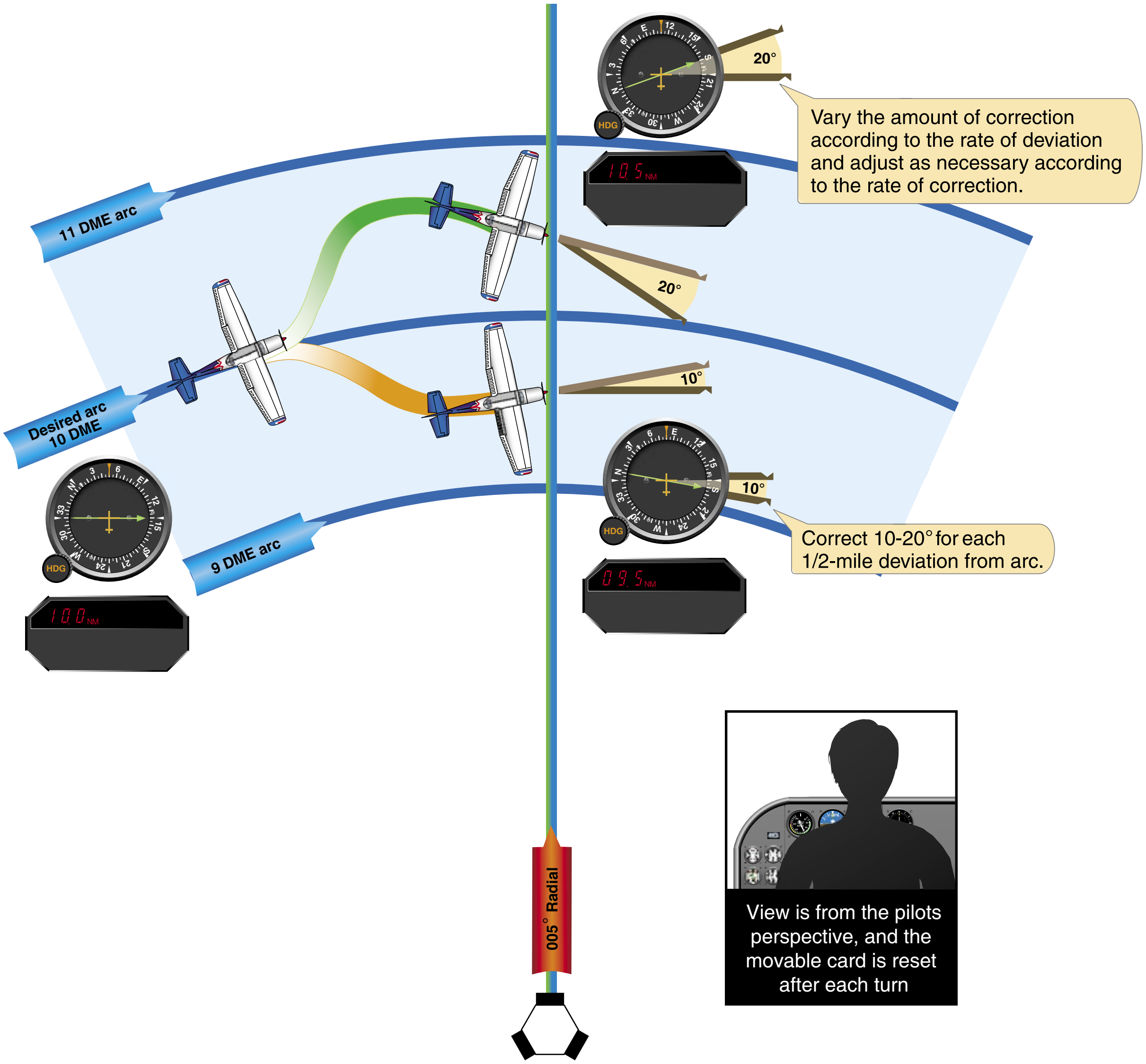

Vary the amount of correction according to the rate of deviation and adjust as necessary according to the rate of correction.

Correct 10-20° for each 1/2-mile deviation from arc.

View is from the pilots perspective, and the movable card is reset after each turn

Figure 7-18. *Using DME and RMI To Maintain an Arc.*

distance to maintain the correct arc distance. Recenter the CDI and note the new primary heading indicated whenever the CDI gets 2°–4° from center.

With an RMI, in a no wind condition, pilots should theoretically be able to fly an exact circle around the facility by maintaining an RB of 90° or 270°. In actual practice, a series of short legs are flown. To maintain the arc in *Figure 7-18*, proceed as follows:

1. With the RMI bearing pointer on the wingtip reference (90° or 270° position) and the aircraft at the desired DME range, maintain a constant heading and allow the bearing pointer to move 5°–10° behind the wingtip. This will cause the range to increase slightly.
2. Turn toward the facility to place the bearing pointer 5–10° ahead of the wingtip reference, and then maintain heading until the bearing pointer is again behind the wingtip. Continue this procedure to maintain the approximate arc.
3. If a crosswind causes the aircraft to drift away from the facility, turn the aircraft until the bearing pointer is ahead of the wingtip reference. If a crosswind causes the aircraft to drift toward the facility, turn until the bearing is behind the wingtip.
4. As a guide in making range corrections, change the RB 10°–20° for each half-mile deviation from the desired arc. For example, in no-wind conditions, if the aircraft is 1/2 to 1 mile outside the arc and the bearing pointer is on the wingtip reference, turn the aircraft 20° toward the facility to return to the arc.

Without an RMI, orientation is more difficult since there is no direct azimuth reference. However, the procedure can be flown using the OBS and CDI for azimuth information and the DME for arc distance.

Intercepting Lead Radials

A lead radial is the radial at which the turn from the arc to the inbound course is started. When intercepting a radial from a DME arc, the lead will vary with arc radius and ground speed. For the average general aviation aircraft, flying arcs such as those depicted on most approach charts at speeds of 150 knots or less, the lead will be under 5°. There is no difference between intercepting a radial from an arc and intercepting it from a straight course.

With an RMI, the rate of bearing movement should be monitored closely while flying the arc. Set the course of the radial to be intercepted as soon as possible and determine the approximate lead. Upon reaching this point, start the intercepting turn. Without an RMI, the technique for radial interception is the same except for azimuth information, which is available only from the OBS and CDI.

The technique for intercepting a localizer from a DME arc is similar to intercepting a radial. At the depicted lead radial (LR 223 or LR 212 in *Figures 7-19, 7-20,* and *7-21*), a pilot having a single VOR/LOC receiver should set it to the localizer frequency. If the pilot has dual VOR/LOC receivers, one unit may be used to provide azimuth information and the other set to the localizer frequency. Since these lead radials provide 7° of lead, a half-standard rate turn should be used until the LOC needle starts to move toward center.

DME Errors

A DME/DME fix (a location based on two DME lines of position from two DME stations) provides a more accurate aircraft location than using a VOR and a DME fix.

DME signals are line-of-sight; the mileage readout is the straight line distance from the aircraft to the DME ground facility and is commonly referred to as slant range distance. Slant range refers to the distance from the aircraft's antenna to the ground station (A line at an angle to the ground transmitter. GPS systems provide distance as the horizontal measurement from the WP to the aircraft. Therefore, at 3,000 feet and 0.5 miles the DME (slant range) would read 0.6 NM while the GPS distance would show the actual horizontal distance of .5 DME. This error is smallest at low altitudes and/or at long ranges. It is greatest when the aircraft is closer to the facility, at which time the DME receiver will display altitude (in NM) above the facility. Slant range error is negligible if the aircraft is one mile or more from the ground facility for each 1,000 feet of altitude above the elevation of the facility.

Area Navigation (RNAV)

Area navigation (RNAV) equipment includes VOR/DME, LORAN, GPS, and inertial navigation systems (INS). RNAV equipment is capable of computing the aircraft position, actual track, groundspeed, and then presenting meaningful information to the pilot. This information may be in the form of distance, cross-track error, and time estimates relative to the selected track or WP. In addition, the RNAV equipment installations must be approved for use under IFR. The Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM) should always be consulted to determine what equipment is installed, the operations that are approved, and the details of equipment use. Some aircraft may have equipment that allows input from more than one RNAV source, thereby providing a very accurate and reliable navigation source.

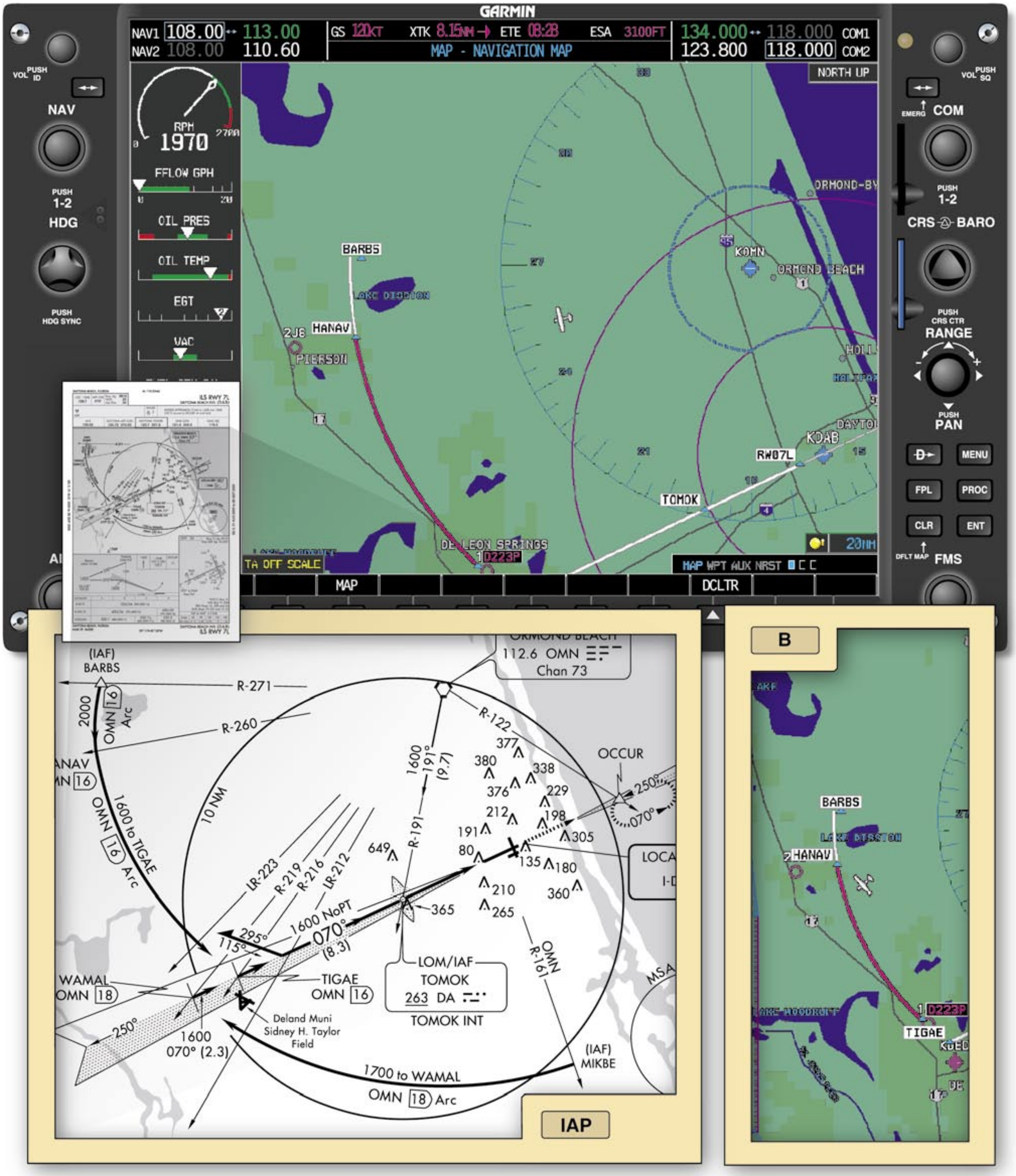
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
GS 120KT XTK 8.15NM ETE 08:28 ESA 3100FT
MAP - NAVIGATION MAP
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
NORTH UP
RPM 1970
FFLOW GPH
OIL PRES
OIL TEMP
EGT
VAC
BARBS
HANAV
2J6
PIERSON
LAKE DISSTON
KOMN
ORMOND BEACH
ORMOND-BY
KDAB
DAYTO
RW07L
TOMOK
DE LEON SPRINGS
D223P
TA OFF SCALE
20NM
MAP WPT AUX NRST
MAP
DCLTR
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
COM
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
ORMOND BEACH
112.6 OMN
Chan 73
(IAF)
BARBS
R-271
R-260
2000
OMN 16 Arc
HANAV
OMN 16
OMN 16 1600 to TIGAE Arc
10 NM
LR-223
R-219
R-216
LR-212
1600 NoPT
070°
(8.3)
295°
115°
1600
191°
(9.7)
R-191
R-122
OCCUR
250°
070°
LOCA
I-D
365
LOM/IAF
TOMOK
263 DA
TOMOK INT
TIGAE
OMN 16
WAMAL
OMN 18
250°
1600
070° (2.3)
Deland Muni
Sidney H. Taylor Field
1700 to WAMAL
OMN 18 Arc
(IAF)
MIKBE
OMN
R-161
MSA
IAP
B
TIGAE
KDED

Figure 7-19. *An aircraft is displayed heading southwest to intercept the localizer approach, using the 16 NM DME Arc off of ORM.*

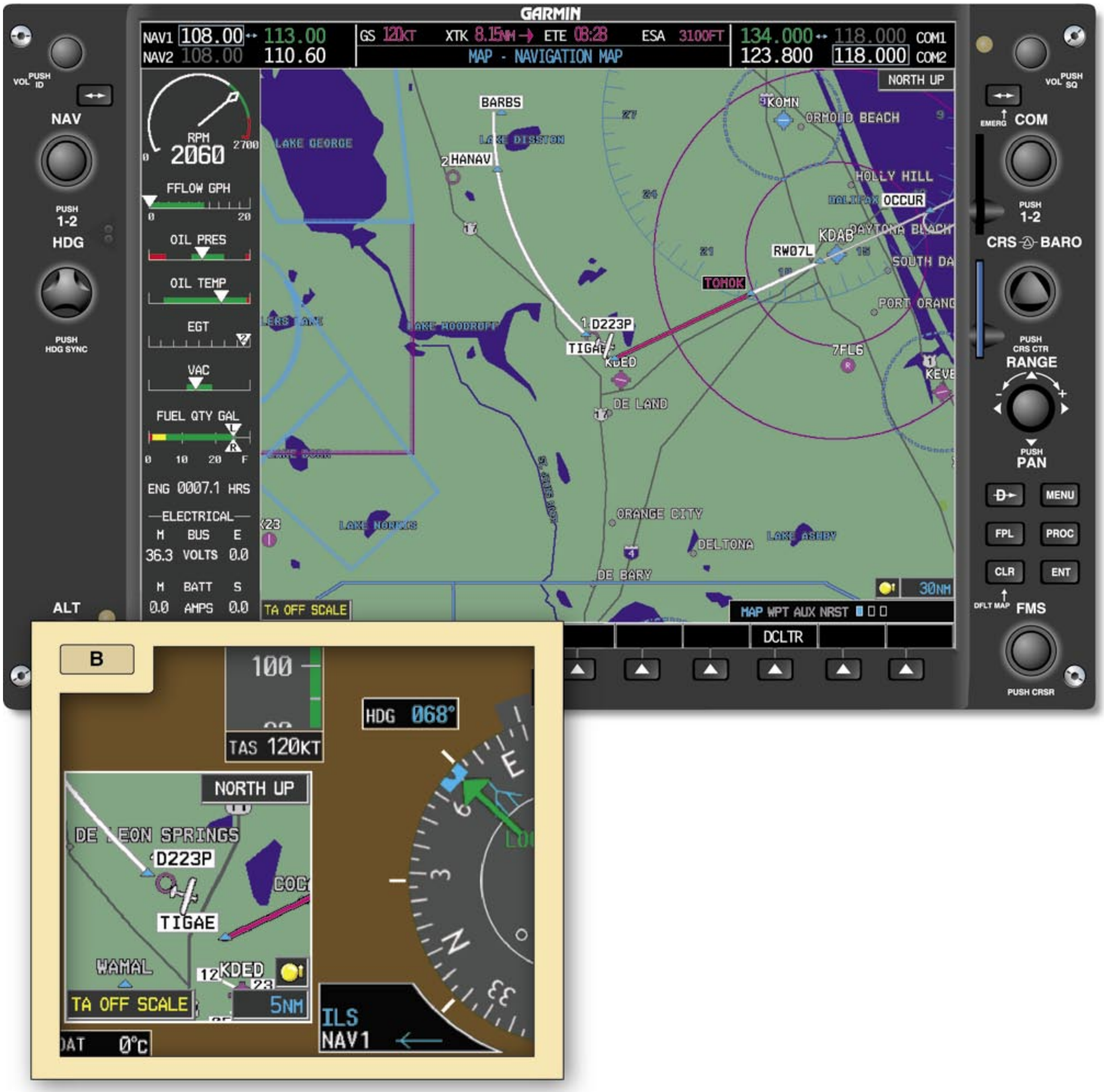
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
GS 120KT
XTK 8.15NM
ETE 08:28
ESA 3100FT
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
MAP - NAVIGATION MAP
NORTH UP
RPM 2060
FFLOW GPH
OIL PRES
OIL TEMP
EGT
VAC
FUEL QTY GAL
ENG 0007.1 HRS
ELECTRICAL
M BUS E
36.3 VOLTS 0.0
M BATT S
0.0 AMPS 0.0
TA OFF SCALE
BARBS
HANAV
LAKE GEORGE
KOMN
ORMOND BEACH
HOLLY HILL
OCCUR
KDAB
DAYTONA BEACH
RW07L
TOMOK
SOUTH DA
PORT ORANG
D223P
TIGAE
KDED
DE LAND
7FL6
ORANGE CITY
DELTONA
DE BARY
30NM
MAP WPT AUX NRST
DCLTR
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS
B
TAS 120KT
HDG 068°
NORTH UP
DE LEON SPRINGS
D223P
TIGAE
WAMAL
KDED
5NM
TA OFF SCALE
OAT 0°C
ILS
NAV1

Figure 7-20. *The same aircraft illustrated in Figure 7-19 shown on the ORM radial near TIGAE intersection turning inbound for the localizer.*

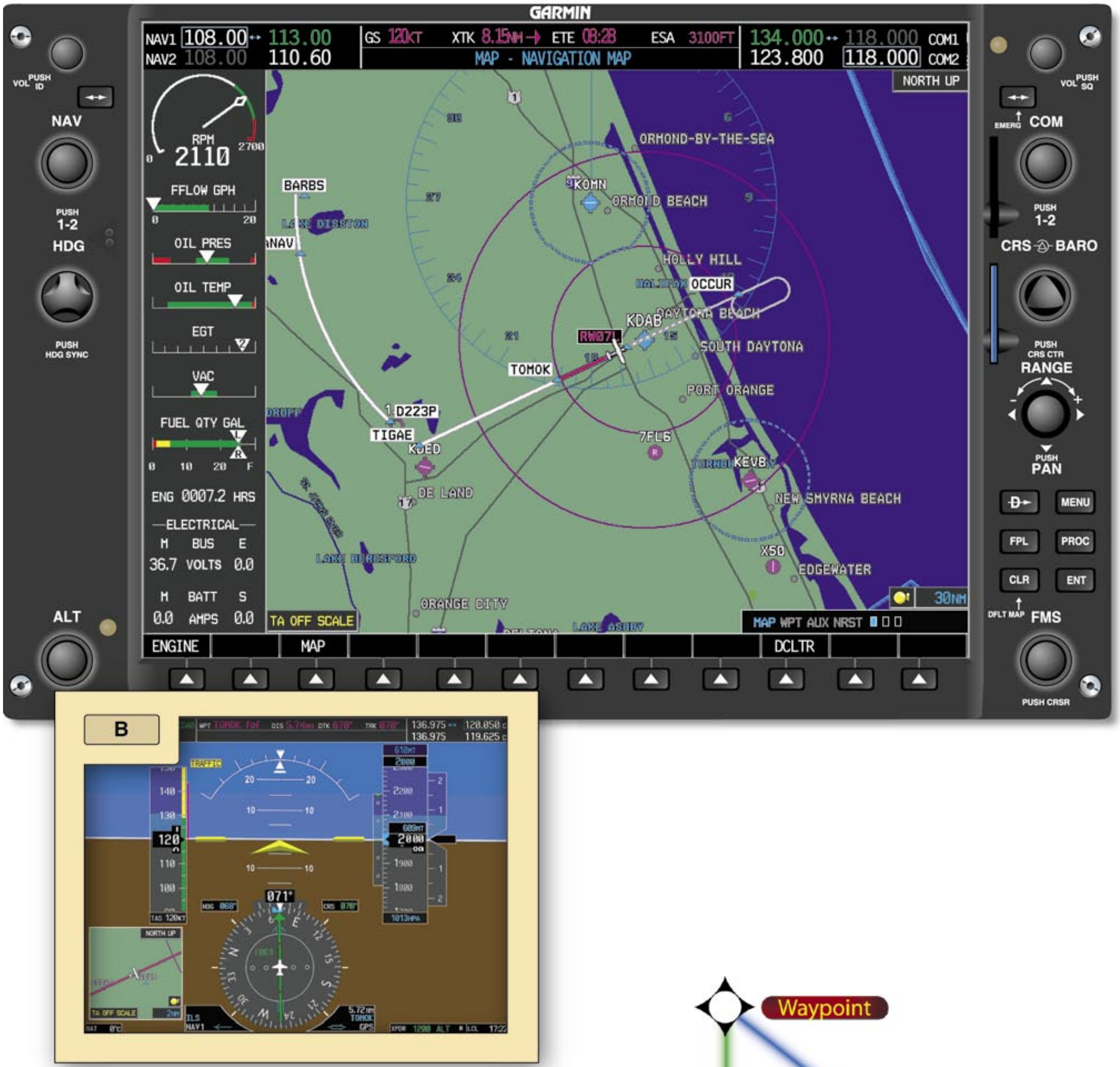
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
GS 120KT
XTK 8.15NM
ETE 08:28
ESA 3100FT
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
MAP - NAVIGATION MAP
NORTH UP
RPM 2110
FFLOW GPH
OIL PRES
OIL TEMP
EGT
VAC
FUEL QTY GAL
ENG 0007.2 HRS
ELECTRICAL
M BUS E
36.7 VOLTS 0.0
M BATT S
0.0 AMPS 0.0
ORMOND-BY-THE-SEA
KOMN
ORMOND BEACH
BARBS
LAKE DISSTON
HOLLY HILL
OCCUR
KDAB
DAYTONA BEACH
RW07L
SOUTH DAYTONA
TOMOK
PORT ORANGE
D223P
TIGAE
KDED
DE LAND
7FL6
KEVB
NEW SMYRNA BEACH
X50
EDGEWATER
LAKE BERESFORD
ORANGE CITY
TA OFF SCALE
30NM
MAP WPT AUX NRST
ENGINE
MAP
DCLTR
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
COM
EMERG
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR
VOL PUSH ID
VOL PUSH SQ
B
TOMOK
DIS 5.74NM
DTK 070°
TRK 070°
136.975
120.050
119.625
TRAFFIC
120
2000
071°
HDG 068°
CRS 070°
TAS 120KT
NORTH UP
TA OFF SCALE
ILS
NAV1
5.72NM
TOMOK
GPS
XPDR 1200 ALT
LCL 17:22
1013HPA
OAT 0°C

Figure 7-21. *Aircraft is illustrated inbound on the localizer course.*

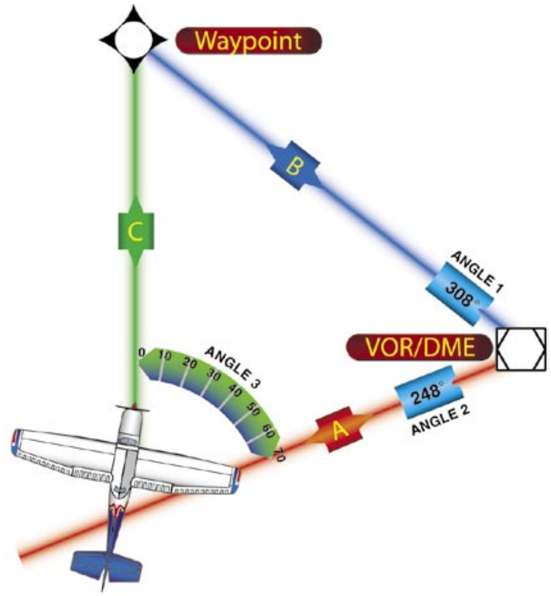
Waypoint
B
C
ANGLE 1
308°
VOR/DME
ANGLE 3
0 10 20 30 40 50 60 70
A
248°
ANGLE 2

Figure 7-22. *RNAV Computation.*

VOR/DME RNAV

VOR RNAV is based on information generated by the present VORTAC or VOR/DME system to create a WP using an airborne computer. As shown in *Figure 7-22*, the value of side A is the measured DME distance to the VOR/DME. Side B, the distance from the VOR/DME to the WP, and angle 1 (VOR radial or the bearing from the VORTAC to the WP) are values set in the flight deck control. The bearing from the VOR/DME to the aircraft, angle 2, is measured by the VOR receiver. The airborne computer continuously compares angles 1 and 2 and determines angle 3 and side C, which is the distance in NMs and magnetic course from the aircraft to the WP. This is presented as guidance information on the flight deck display.

VOR/DME RNAV Components

Although RNAV flight deck instrument displays vary among manufacturers, most are connected to the aircraft CDI with a switch or knob to select VOR or RNAV guidance. There is usually a light or indicator to inform the pilot whether VOR or RNAV is selected. *[Figure 7-23]* The display includes the WP, frequency, mode in use, WP radial and distance, DME distance, groundspeed, and time to station.

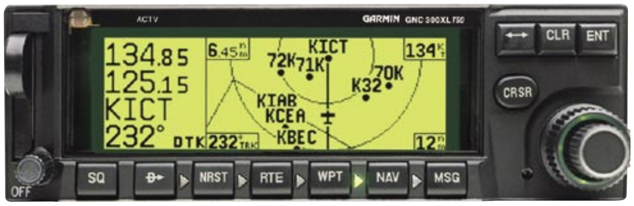

Figure 7-23. *Onboard RNAV receivers have changed significantly. Originally, RNAV receivers typically computed combined data from VOR, VORTAC, and/or DME. That is generally not the case now. Today, GPS such as the GNC 300 and the Bendix King KLS 88 LORAN receivers compute waypoints based upon embedded databases and aircraft positional information.*

Most VOR/DME RNAV systems have the following airborne controls:

1. Off/On/Volume control to select the frequency of the VOR/DME station to be used.
2. MODE select switch used to select VOR/DME mode, with:
   a. Angular course width deviation (standard VOR operation); or
   b. Linear cross-track deviation as standard (±5 NM full scale CDI).
3. RNAV mode, with direct to WP with linear cross-track deviation of ±5 NM.
4. RNAV/APPR (approach mode) with linear deviation of ±1.25 NM as full scale CDI deflection.
5. WP select control. Some units allow the storage of more than one WP; this control allows selection of any WP in storage.
6. Data input controls. These controls allow user input of WP number or ident, VOR or LOC frequency, WP radial and distance.

While DME groundspeed readout is accurate only when tracking directly to or from the station in VOR/DME mode, in RNAV mode the DME groundspeed readout is accurate on any track.

Function of VOR/DME RNAV

The advantages of the VOR/DME RNAV system stem from the ability of the airborne computer to locate a WP wherever it is convenient, as long as the aircraft is within reception range of both nearby VOR and DME facilities. A series of these WPs make up an RNAV route. In addition to the published routes, a random RNAV route may be flown under IFR if it is approved by air traffic control (ATC). RNAV DPs and standard terminal arrival routes (STARs) are contained in the DP and STAR booklets.

VOR/DME RNAV approach procedure charts are also available. Note in the VOR/DME RNAV chart excerpt shown in *Figure 7-24* that the WP identification boxes contain the following information: WP name, coordinates, frequency, identifier, radial distance (facility to WP), and reference facility elevation. The initial approach fix (IAF), final approach fix (FAF), and missed approach point (MAP) are labeled.

To fly a route or to execute an approach under IFR, the RNAV equipment installed in the aircraft must be approved for the appropriate IFR operations.

In vertical navigation (VNAV) mode, vertical guidance is provided, as well as horizontal guidance in some installations. A WP is selected at a point where the descent begins, and another WP is selected where the descent ends. The RNAV equipment computes the rate of descent relative to the groundspeed; on some installations, it displays vertical guidance information on the GS indicator. When using this type of equipment during an instrument approach, the pilot must keep in mind that the vertical guidance information provided is not part of the nonprecision approach. Published nonprecision approach altitudes must be observed and complied with, unless otherwise directed by ATC.

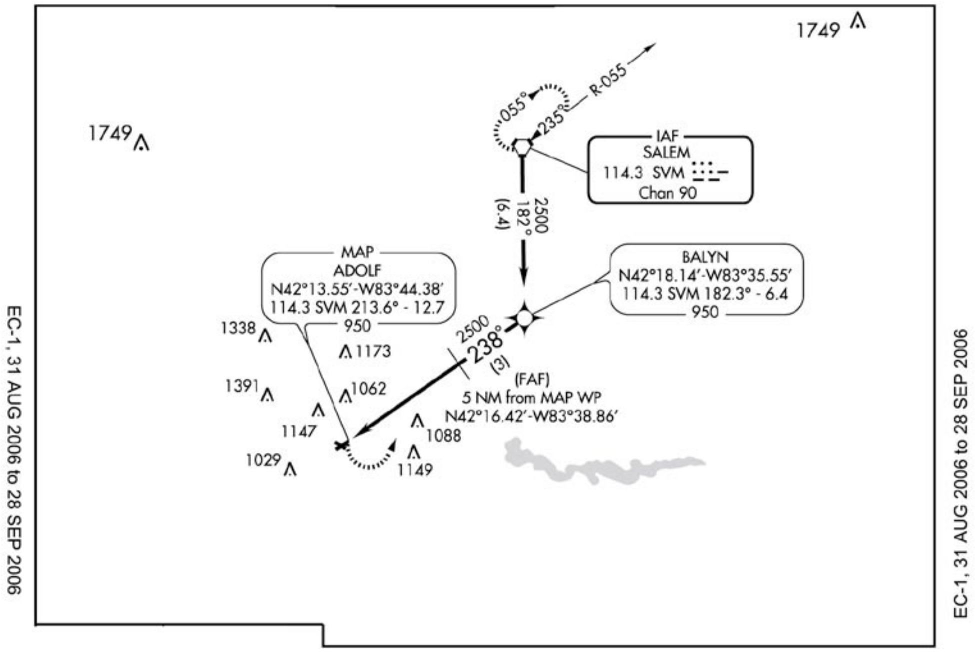
1749
1749
R-055
055°
235°
IAF
SALEM
114.3 SVM
Chan 90
2500
182°
(6.4)
MAP
ADOLF
N42°13.55′-W83°44.38′
114.3 SVM 213.6° - 12.7
950
BALYN
N42°18.14′-W83°35.55′
114.3 SVM 182.3° - 6.4
950
1338
1173
1391
1062
1147
1029
1149
1088
2500
238°
(3)
(FAF)
5 NM from MAP WP
N42°16.42′-W83°38.86′
EC-1, 31 AUG 2006 to 28 SEP 2006
EC-1, 31 AUG 2006 to 28 SEP 2006

Figure 7-24. *VOR/DME RNAV Rwy 25 Approach (Excerpt).*

To fly to a WP using RNAV, observe the following procedure *[Figure 7-25]*:

1. Select the VOR/DME frequency.
2. Select the RNAV mode.
3. Select the radial of the VOR that passes through the WP (225°).
4. Select the distance from the DME to the WP (12 NM).
5. Check and confirm all inputs, and center the CDI needle with the TO indicator showing.
6. Maneuver the aircraft to fly the indicated heading plus or minus wind correction to keep the CDI needle centered.
7. The CDI needle will indicate distance off course of 1 NM per dot; the DME readout will indicate distance in NM from the WP; the groundspeed will read closing speed (knots) to the WP; and the time to station (TTS) will read time to the WP.

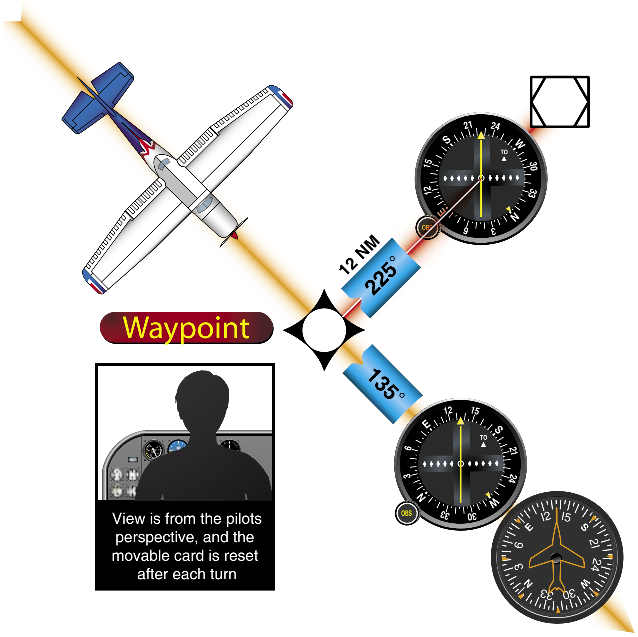
12 NM
225°
Waypoint
135°

View is from the pilots perspective, and the movable card is reset after each turn

Figure 7-25. *Aircraft/DME/Waypoint Relationship.*

VOR/DME RNAV Errors

The limitation of this system is the reception volume. Published approaches have been tested to ensure this is not a problem. Descents/approaches to airports distant from the VOR/DME facility may not be possible because, during the approach, the aircraft may descend below the reception altitude of the facility at that distance.

Long Range Navigation (LORAN)

LORAN uses a network of land-based transmitters to provide an accurate long-range navigation system. The FAA and the United States Coast Guard (USCG) arranged the stations into chains. The signal from station is a carefully structured sequence of brief RF pulses centered at 100 kHz. At that frequency, signals travel considerable distances as ground waves, from which accurate navigation information is available. The airborne receiver monitors all of the stations within the selected chain, then measures the arrival time difference (TD) between the signals. All of the points having the same TD from a station pair create a line of position

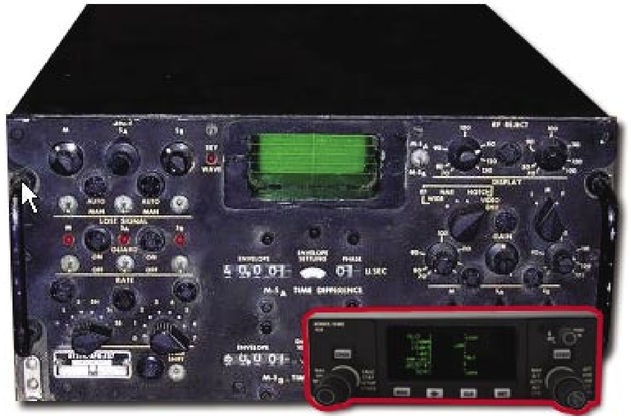

Figure 7-26. *A control panel from a military aircraft after LORAN was first put into use. The receiver is remotely mounted and weighs over 25 pounds. Its size is about six times that of the LORAN fully integrated receiver.*

(LOP). The aircraft position is determined at the intersection of two or more LOPs. Then the computer converts the known location to latitude and longitude coordinates. *[Figure 7-26]*

While continually computing latitude/longitude fixes, the computer is able to determine and display:

1. Track over the ground since last computation;
2. Groundspeed by dividing distance covered since last computation by the time since last computation (and averaging several of these);
3. Distance to destination;
4. Destination time of arrival; and
5. Cross-track error.

The Aeronautical Information Manual (AIM) provides a detailed explanation of how LORAN works. LORAN is a very accurate navigation system if adequate signals are received. There are two types of accuracy that must be addressed in any discussion of LORAN accuracy.

Repeatable accuracy is the accuracy measured when a user notes the LORAN position, moves away from that location, then uses the LORAN to return to that initial LORAN position. Distance from that initial position is the error. Propagation and terrain errors will be essentially the same as when the first position was taken, so those errors are factored out by using the initial position. Typical repeatable accuracy for LORAN can be as good as 0.01 NM, or 60 feet, if the second position is determined during the day and within a short period of time (a few days).

Absolute accuracy refers to the ability to determine present position in space independently, and is most often used by pilots. When the LORAN receiver is turned on and position is determined, absolute accuracy applies. Typical LORAN absolute accuracy will vary from about 0.1 NM to as much as 2.5 NM depending on distance from the station, geometry of the TD LOP crossing angles, terrain and environmental conditions, signal-to-noise ratio (signal strength), and some design choices made by the receiver manufacturer.

Although LORAN use diminished with the introduction of Global Navigation Satellite Systems such as the United States' GPS, its use has since increased. Three items aided in this resurgence:

- In 1996, a commission called the Gore Commission evaluated GPS' long-term use as a sole navigation aid. Although GPS was hailed originally as the eventual sole NAVAID, which would replace the need for most currently existing NAVAIDs by the year 2020, the Commission questioned single-link failure potential and its effect on the NAS. For this reason, the forecasted decommissioning of the VOR has been amended and their expectant lifecycle extended into the future. Additionally, the use of LORAN continues to be evaluated for facilitating carrying GPS corrective timing signals.
- The GPS is controlled by the DOD presenting certain unforecasted uncertainties for commercial use on an uninterrupted basis.

As a result of these and other key factors, it was determined that LORAN would remain. In recognition of GPS vulnerabilities as a GNSS, there are plans to maintain other systems that could provide en route and terminal accuracy such as LORAN. Therefore as LORAN is further modernized it's a possibility that it may be used to augment GPS and provide backup to GPS during unlikely but potential outages. Or if combined with GPS and other systems such as newer miniaturized low-cost inertial navigation systems (INS), superior accuracy and seamless backup will always be available.

LORAN Components

The LORAN receiver incorporates a radio receiver, signal processor, navigation computer, control/display, and antenna. When turned on, the receivers go through an initialization or warm-up period, then inform the user they are ready to be programmed. LORAN receivers vary widely in their appearance, method of user programming, and navigation information display. Therefore, it is necessary to become familiar with the unit, including programming and output interpretation. The LORAN operating manual should be in the aircraft at all times and available to the pilot. IFR-approved LORAN units require that the manual be aboard and that the pilot be familiar with the unit's functions, before flight.

Function of LORAN

After initialization, select for the present location WP (the airport), and select GO TO in order to determine if the LORAN is functioning properly. Proper operation is indicated by a low distance reading (0 to 0.5 NM). The simplest mode of navigation is referred to as GO TO: you select a WP from one of the databases and choose the GO TO mode. Before use in flight, verify that the latitude and longitude of the chosen WP is correct by reference to another approved information source. An updatable LORAN database that supports the appropriate operations (e.g., en route, terminal, and instrument approaches) is required when operating under IFR.

In addition to displaying bearing, distance, time to the WP, and track and speed over the ground, the LORAN receiver may have other features such as flight planning (WP sequential storage), emergency location of several nearest airports, vertical navigation capabilities, and more.

LORAN Errors

System Errors

LORAN is subject to interference from many external sources, which can cause distortion of or interference with LORAN signals. LORAN receiver manufacturers install "notch filters" to reduce or eliminate interference. Proximity to 60 Hz alternating current power lines, static discharge, P-static, electrical noise from generators, alternators, strobes, and other onboard electronics may decrease the signal-to-noise ratio to the point where the LORAN receiver's performance is degraded.

Proper installation of the antenna, good electrical bonding, and an effective static discharge system are the minimum requirements for LORAN receiver operation. Most receivers have internal tests that verify the timing alignment of the receiver clock with the LORAN pulse, and measure and display signal-to-noise ratio. A signal will be activated to alert the pilot if any of the parameters for reliable navigation are exceeded on LORAN sets certified for IFR operations.

LORAN is most accurate when the signal travels over sea water during the day and least accurate when the signal comes over land and large bodies of fresh water or ice at night; furthermore, the accuracy degrades as distance from the station increases. However, LORAN accuracy is generally better than VOR accuracy.

Operational Errors

Some of the typical pilot-induced errors of LORAN operation are:

1. Use of a nonapproved LORAN receiver for IFR operations. The pilot should check the aircraft's POH/AFM LORAN supplement to be certain the unit's functions are well understood (this supplement must be present in the aircraft for approved IFR operations). There should be a copy of FAA Form 337, Major Repair and Alteration, present in the aircraft's records, showing approval of use of this model LORAN for IFR operations in this aircraft.
2. Failure to double-check the latitude/longitude values for a WP to be used. Whether the WP was accessed from the airport, NDB, VOR, or intersection database, the values of latitude and longitude should still be checked against the values in the A/FD or other approved source. If the WP data is entered in the user database, its accuracy must be checked before use.
3. Attempting to use LORAN information with degraded signals.

Advanced Technologies

Global Navigation Satellite System (GNSS)

The Global Navigation Satellite System (GNSS) is a constellation of satellites providing a high-frequency signal which contains time and distance that is picked up by a receiver thereby. *[Figure 7-27]* The receiver which picks up multiple signals from different satellites is able to triangulate its position from these satellites.

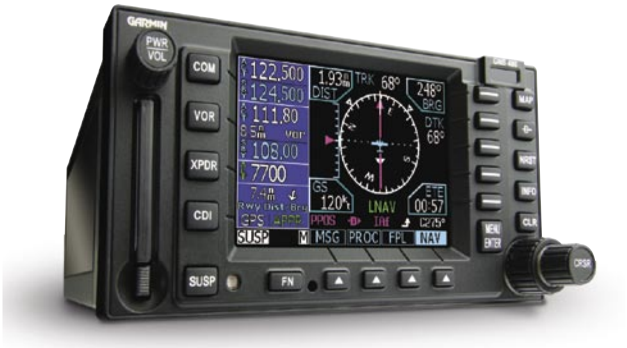

Figure 7-27. *A typical example (GNS 480) of a stand-alone GPS receiver and display.*

Three GNSSs exist today: the GPS, a United States system; the Russian GNSS (GLONASS); and Galileo, a European system.

1. GLONASS is a network of 24 satellites, which can be picked up by any GLONASS receiver, allowing the user to pinpoint their position.
2. Galileo planned to be a network of 30 satellites that continuously transmit high-frequency radio signals containing time and distance data that can be picked up by a Galileo receiver with operational expectancy by 2013.
3. The GPS came on line in 1992 with 24 satellites, and today utilizes 30 satellites.

Global Positioning System (GPS)

The GPS is a satellite-based radio navigation system, which broadcasts a signal that is used by receivers to determine precise position anywhere in the world. The receiver tracks multiple satellites and determines a measurement that is then used to determine the user location. *[Figure 7-28]*

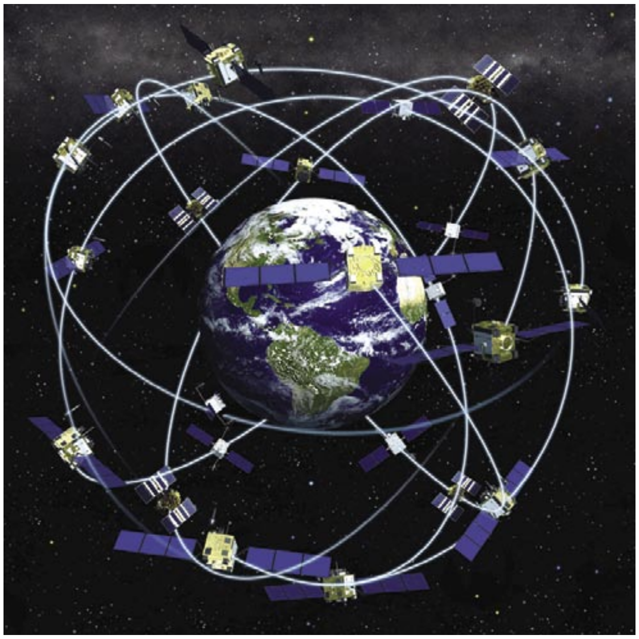

Figure 7-28. *Typical GPS Satellite Array.*

The Department of Defense (DOD) developed and deployed GPS as a space-based positioning, velocity, and time system. The DOD is responsible for operation of the GPS satellite constellation, and constantly monitors the satellites to ensure proper operation. The GPS system permits Earth-centered coordinates to be determined and provides aircraft position referenced to the DOD World Geodetic System of 1984 (WGS-84). Satellite navigation systems are unaffected by weather and provide global navigation coverage that fully meets the civil requirements for use as the primary means of navigation in oceanic airspace and certain remote areas. Properly certified GPS equipment may be used as a supplemental means of IFR navigation for domestic en route, terminal operations, and certain IAPs. Navigational values, such as distance and bearing to a WP and groundspeed, are computed from the aircraft's current position (latitude and longitude) and the location of the next WP. Course guidance is provided as a linear deviation from the desired track of a Great Circle route between defined WPs.

GPS may not be approved for IFR use in other countries. Prior to its use, pilots should ensure that GPS is authorized by the appropriate countries.

GPS Components

GPS consists of three distinct functional elements: space, control, and user.

The space element consists of over 30 Navstar satellites. This group of satellites is called a constellation. The space element consists of 24 Navigation System using Timing and Ranging (NAVSTAR) satellites in 6 orbital planes. The satellites in each plane are spaced 60° apart for complete coverage and are located (nominally) at about 11,000 miles above the Earth. The planes are arranged so that there are always five satellites in view at any time on the Earth. Presently, there are at least 31 Block II/IIA/IIR and IIR-M satellites in orbit with the additional satellites representing replacement satellites (upgraded systems) and spares. Recently, the Air Force received funding for procurement of 31 Block IIF satellites. The GPS constellation broadcasts a pseudo-random code timing signal and data message that the aircraft equipment processes to obtain satellite position and status data. By knowing the precise location of each satellite and precisely matching timing with the atomic clocks on the satellites, the aircraft receiver/processor can accurately measure the time each signal takes to arrive at the receiver and, therefore, determine aircraft position.

The control element consists of a network of ground-based GPS monitoring and control stations that ensure the accuracy of satellite positions and their clocks. In its present form, it has five monitoring stations, three ground antennas, and a master control station.

The user element consists of antennas and receiver/processors on board the aircraft that provide positioning, velocity, and precise timing to the user. GPS equipment used while operating under IFR must meet the standards set forth in Technical Standard Order (TSO) C-129 (or equivalent); meet the airworthiness installation requirements; be "approved" for that type of IFR operation; and be operated in accordance with the applicable POH/AFM or flight manual supplement.

An updatable GPS database that supports the appropriate operations (e.g., en route, terminal, and instrument approaches) is required when operating under IFR. The aircraft GPS navigation database contains WPs from the geographic areas where GPS navigation has been approved for IFR operations. The pilot selects the desired WPs from the database and may add user-defined WPs for the flight.

Equipment approved in accordance with TSO C-115a, visual flight rules (VFR), and hand-held GPS systems do not meet the requirements of TSO C-129 and are not authorized for IFR navigation, instrument approaches, or as a principal instrument flight reference. During IFR operations, these units (TSO C-115a) may be considered only an aid to situational awareness.

Prior to GPS/WAAS IFR operation, the pilot must review appropriate NOTAMs and aeronautical information. This information is available on request from an Automated Flight Service Station. The FAA will provide NOTAMs to advise pilots of the status of the WAAS and level of service available.

Function of GPS

GPS operation is based on the concept of ranging and triangulation from a group of satellites in space which act as precise reference points. The receiver uses data from a minimum of four satellites above the mask angle (the lowest angle above the horizon at which it can use a satellite).

The aircraft GPS receiver measures distance from a satellite using the travel time of a radio signal. Each satellite transmits a specific code, called a course/acquisition (CA) code, which contains information about satellite position, the GPS system time, and the health and accuracy of the transmitted data. Knowing the speed at which the signal traveled (approximately 186,000 miles per second) and the exact broadcast time, the distance traveled by the signal can be computed from the arrival time. The distance derived from this method of computing distance is called a pseudo-range because it is not a direct measurement of distance, but a measurement based on time. In addition to knowing the distance to a satellite, a receiver needs to know the satellite's exact position in space, its ephemeris. Each satellite transmits information about its exact orbital location. The GPS receiver uses this information to establish the precise position of the satellite.

Using the calculated pseudo-range and position information supplied by the satellite, the GPS receiver/processor mathematically determines its position by triangulation from several satellites. The GPS receiver needs at least four satellites to yield a three-dimensional position (latitude, longitude, and altitude) and time solution. The GPS receiver computes navigational values (distance and bearing to a WP, groundspeed, etc.) by using the aircraft's known latitude/longitude and referencing these to a database built into the receiver.

The GPS receiver verifies the integrity (usability) of the signals received from the GPS constellation through receiver autonomous integrity monitoring (RAIM) to determine if a satellite is providing corrupted information. RAIM needs a minimum of five satellites in view, or four satellites and a barometric altimeter baro-aiding to detect an integrity anomaly. For receivers capable of doing so, RAIM needs six satellites in view (or five satellites with baro-aiding) to isolate a corrupt satellite signal and remove it from the navigation solution.

Generally, there are two types of RAIM messages. One type indicates that there are not enough satellites available to provide RAIM and another type indicates that the RAIM has detected a potential error that exceeds the limit for the current phase of flight. Without RAIM capability, the pilot has no assurance of the accuracy of the GPS position.

Aircraft using GPS navigation equipment under IFR for domestic en route, terminal operations, and certain IAPs, must be equipped with an approved and operational alternate means of navigation appropriate to the flight. The avionics necessary to receive all of the ground-based facilities appropriate for the route to the destination airport and any required alternate airport must be installed and operational. Ground-based facilities necessary for these routes must also be operational. Active monitoring of alternative navigation equipment is not required if the GPS receiver uses RAIM for integrity monitoring. Active monitoring of an alternate means of navigation is required when the RAIM capability of the GPS equipment is lost. In situations where the loss of RAIM capability is predicted to occur, the flight must rely on other approved equipment, delay departure, or cancel the flight.

GPS Substitution

IFR En Route and Terminal Operations

GPS systems, certified for IFR en route and terminal operations, may be used as a substitute for ADF and DME receivers when conducting the following operations within the United States NAS.

1. Determining the aircraft position over a DME fix. This includes en route operations at and above 24,000 feet mean sea level (MSL) (FL 240) when using GPS for navigation.
2. Flying a DME arc.
3. Navigating TO/FROM an NDB/compass locator.

4. Determining the aircraft position over an NDB/compass locator.
5. Determining the aircraft position over a fix defined by an NDB/compass locator bearing crossing a VOR/LOC course.
6. Holding over an NDB/compass locator.

GPS Substitution for ADF or DME

Using GPS as a substitute for ADF or DME is subject to the following restrictions:

1. This equipment must be installed in accordance with appropriate airworthiness installation requirements and operated within the provisions of the applicable POH/AFM, or supplement.
2. The required integrity for these operations must be provided by at least en route RAIM, or equivalent.
3. WPs, fixes, intersections, and facility locations to be used for these operations must be retrieved from the GPS airborne database. The database must be current. If the required positions cannot be retrieved from the airborne database, the substitution of GPS for ADF and/or DME is not authorized
4. Procedures must be established for use when RAIM outages are predicted or occur. This may require the flight to rely on other approved equipment or require the aircraft to be equipped with operational NDB and/or DME receivers. Otherwise, the flight must be rerouted, delayed, canceled, or conducted under VFR.
5. The CDI must be set to terminal sensitivity (1 NM) when tracking GPS course guidance in the terminal area.
6. A non-GPS approach procedure must exist at the alternate airport when one is required. If the non-GPS approaches on which the pilot must rely require DME or ADF, the aircraft must be equipped with DME or ADF avionics as appropriate.
7. Charted requirements for ADF and/or DME can be met using the GPS system, except for use as the principal instrument approach navigation source.

NOTE: The following provides guidance, which is not specific to any particular aircraft GPS system. For specific system guidance, refer to the POH/AFM, or supplement, or contact the system manufacturer.

To Determine Aircraft Position Over a DME Fix:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. If the fix is identified by a five-letter name which is contained in the GPS airborne database, select either the named fix as the active GPS WP or the facility establishing the DME fix as the active GPS WP. When using a facility as the active WP, the only acceptable facility is the DME facility which is charted as the one used to establish the DME fix. If this facility is not in the airborne database, it is not authorized for use.
3. If the fix is identified by a five-letter name which is not contained in the GPS airborne database, or if the fix is not named, select the facility establishing the DME fix or another named DME fix as the active GPS WP.
4. When selecting the named fix as the active GPS WP, a pilot is over the fix when the GPS system indicates the active WP.
5. If selecting the DME providing facility as the active GPS WP, a pilot is over the fix when the GPS distance from the active WP equals the charted DME value, and the aircraft is established on the appropriate bearing or course.

To Fly a DME Arc:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. Select from the airborne database the facility providing the DME arc as the active GPS WP. The only acceptable facility is the DME facility on which the arc is based. If this facility is not in your airborne database, you are not authorized to perform this operation.
3. Maintain position on the arc by reference to the GPS distance instead of a DME readout.

To Navigate TO or FROM an NDB/Compass Locator:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. Select the NDB/compass locator facility from the airborne database as the active WP. If the chart depicts the compass locator collocated with a fix of the same name, use of that fix as the active WP in place of the compass locator facility is authorized.
3. Select and navigate on the appropriate course to or from the active WP.

To Determine Aircraft Position Over an NDB/ Compass Locator:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. Select the NDB/compass locator facility from the airborne database. When using an NDB/compass locator, the facility must be charted and be in the airborne database. If the facility is not in the airborne database, pilots are not authorized to use a facility WP for this operation.
3. A pilot is over the NDB/compass locator when the GPS system indicates arrival at the active WP.

To Determine Aircraft Position Over a Fix Made up of an NDB/Compass Locator Bearing Crossing a VOR/LOC Course:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. A fix made up by a crossing NDB/compass locator bearing is identified by a five-letter fix name. Pilots may select either the named fix or the NDB/compass locator facility providing the crossing bearing to establish the fix as the active GPS WP. When using an NDB/compass locator, that facility must be charted and be in the airborne database. If the facility is not in the airborne database, pilots are not authorized to use a facility WP for this operation.
3. When selecting the named fix as the active GPS WP, pilot is over the fix when the GPS system indicates the pilot is at the WP.
4. When selecting the NDB/compass locator facility as the active GPS WP, pilots are over the fix when the GPS bearing to the active WP is the same as the charted NDB/compass locator bearing for the fix flying the prescribed track from the non-GPS navigation source.

To Hold Over an NDB/Compass Locator:

1. Verify aircraft GPS system integrity monitoring is functioning properly and indicates satisfactory integrity.
2. Select the NDB/compass locator facility from the airborne database as the active WP. When using a facility as the active WP, the only acceptable facility is the NDB/compass locator facility which is charted. If this facility is not in the airborne database, its use is not authorized.
3. Select nonsequencing (e.g., "HOLD" or "OBS") mode and the appropriate course in accordance with the POH/AFM, or supplement.
4. Hold using the GPS system in accordance with the POH/AFM, or supplement.

IFR Flight Using GPS

Preflight preparations should ensure that the GPS is properly installed and certified with a current database for the type of operation. The GPS operation must be conducted in accordance with the FAA-approved POH/AFM or flight manual supplement. Flightcrew members must be thoroughly familiar with the particular GPS equipment installed in the aircraft, the receiver operation manual, and the POH/AFM or flight manual supplement. Unlike ILS and VOR, the basic operation, receiver presentation to the pilot and some capabilities of the equipment can vary greatly. Due to these differences, operation of different brands, or even models of the same brand of GPS receiver under IFR should not be attempted without thorough study of the operation of that particular receiver and installation. Using the equipment in flight under VFR conditions prior to attempting IFR operation will allow further familiarization.

Required preflight preparations should include checking NOTAMs relating to the IFR flight when using GPS as a supplemental method of navigation. GPS satellite outages are issued as GPS NOTAMs both domestically and internationally. Pilots may obtain GPS RAIM availability information for an airport by specifically requesting GPS aeronautical information from an automated flight service station (AFSS) during preflight briefings. GPS RAIM aeronautical information can be obtained for a 3-hour period: the estimated time of arrival (ETA), and 1 hour before to 1 hour after the ETA hour, or a 24-hour time frame for a specific airport. FAA briefers will provide RAIM information for a period of 1 hour before to 1 hour after the ETA, unless a specific timeframe is requested by the pilot. If flying a published GPS departure, the pilot should also request a RAIM prediction for the departure airport. Some GPS receivers have the capability to predict RAIM availability. The pilot should also ensure that the required underlying ground-based navigation facilities and related aircraft equipment appropriate to the route of flight, terminal operations, instrument approaches for the destination, and alternate airports/heliports will be operational for the ETA. If the required ground-based facilities and equipment will not be available, the flight should be rerouted, rescheduled, canceled, or conducted under VFR.

Except for programming and retrieving information from the GPS receiver, planning the flight is accomplished in a similar manner to conventional NAVAIDs. Departure WP, DP, route, STAR, desired approach, IAF, and destination airport are entered into the GPS receiver according to the manufacturer's instructions. During preflight, additional information may be entered for functions such as ETA, fuel planning, winds aloft, etc.

When the GPS receiver is turned on, it begins an internal process of test and initialization. When the receiver is initialized, the user develops the route by selecting a WP or series of WPs, verifies the data, and selects the active flight plan. This procedure varies widely among receivers made by different manufacturers. GPS is a complex system, offering little standardization between receiver models. It is the pilot's responsibility to be familiar with the operation of the equipment in the aircraft.

The GPS receiver provides navigational values such as track, bearing, groundspeed, and distance. These are computed from the aircraft's present latitude and longitude to the location of the next WP. Course guidance is provided between WPs. The pilot has the advantage of knowing the aircraft's actual track over the ground. As long as track and bearing to the WP are matched up (by selecting the correct aircraft heading), the aircraft is going directly to the WP.

GPS Instrument Approaches

There is a mixture of GPS overlay approaches (approaches with "or GPS" in the title) and GPS stand-alone approaches in the United States.

NOTE: GPS instrument approach operations outside the United States must be authorized by the appropriate country authority.

While conducting these IAPs, ground-based NAVAIDs are not required to be operational and associated aircraft avionics need not be installed, operational, turned on, or monitored; however, monitoring backup navigation systems is always recommended when available.

Pilots should have a basic understanding of GPS approach procedures and practice GPS IAPs under visual meteorological conditions (VMC) until thoroughly proficient with all aspects of their equipment (receiver and installation) prior to attempting flight in instrument meteorological conditions (IMC). *[Figure 7-29]*

All IAPs must be retrievable from the current GPS database supplied by the manufacturer or other FAA-approved source. Flying point to point on the approach does not assure compliance with the published approach procedure. The proper RAIM sensitivity will not be available and the CDI sensitivity will not automatically change to 0.3 NM. Manually setting CDI sensitivity does not automatically change the RAIM sensitivity on some receivers. Some existing nonprecision approach procedures cannot be coded for use with GPS and will not be available as overlays.

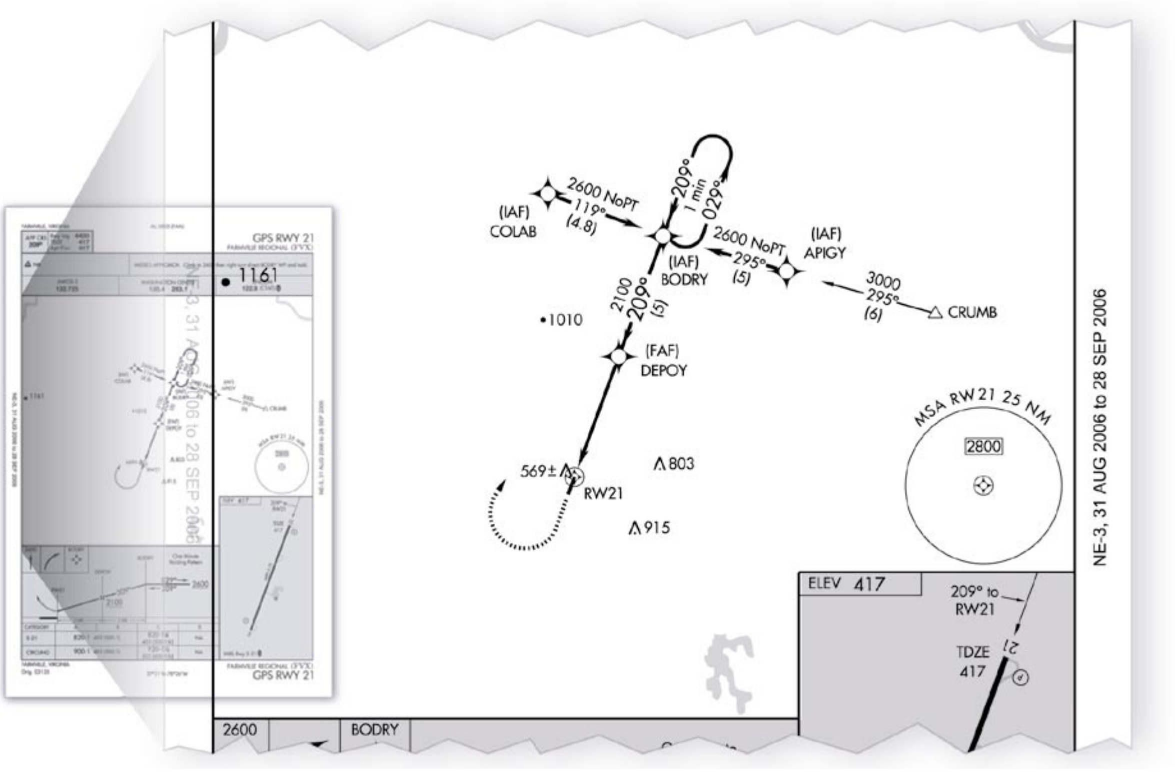

Figure 7-29. *A GPS Stand-Alone Approach.*

GPS approaches are requested and approved by ATC using the GPS title, such as "GPS RWY 24" or "RNAV RWY 35." Using the manufacturer's recommended procedures, the desired approach and the appropriate IAF are selected from the GPS receiver database. Pilots should fly the full approach from an initial approach waypoint (IAWP) or feeder fix unless specifically cleared otherwise. Randomly joining an approach at an intermediate fix does not ensure terrain clearance.

When an approach has been loaded in the flight plan, GPS receivers will give an "arm" annunciation 30 NM straight line distance from the airport/heliport reference point. The approach mode should be "armed" when within 30 NM distance so the receiver will change from en route CDI (±5 NM) and RAIM (±2 NM) sensitivity to ±1 NM terminal sensitivity. Where the IAWP is within 30 NM, a CDI sensitivity change will occur once the approach mode is armed and the aircraft is within 30 NM. Where the IAWP is beyond the 30 NM point, CDI sensitivity will not change until the aircraft is within 30 NM even if the approach is armed earlier. Feeder route obstacle clearance is predicated on the receiver CDI and RAIM being in terminal CDI sensitivity within 30 NM of the airport/heliport reference point; therefore, the receiver should always be armed no later than the 30 NM annunciation.

Pilots should pay particular attention to the exact operation of their GPS receivers for performing holding patterns and in the case of overlay approaches, operations such as procedure turns. These procedures may require manual intervention by the pilot to stop the sequencing of WPs by the receiver and to resume automatic GPS navigation sequencing once the maneuver is complete. The same WP may appear in the route of flight more than once and consecutively (e.g., IAWP, final approach waypoint (FAWP), missed approach waypoint (MAWP) on a procedure turn). Care must be exercised to ensure the receiver is sequenced to the appropriate WP for the segment of the procedure being flown, especially if one or more fly-over WPs are skipped (e.g., FAWP rather than IAWP if the procedure turn is not flown). The pilot may need to sequence past one or more fly-overs of the same WP in order to start GPS automatic sequencing at the proper place in the sequence of WPs.

When receiving vectors to final, most receiver operating manuals suggest placing the receiver in the nonsequencing mode on the FAWP and manually setting the course. This provides an extended final approach course in cases where the aircraft is vectored onto the final approach course outside of any existing segment which is aligned with the runway. Assigned altitudes must be maintained until established on a published segment of the approach. Required altitudes at WPs outside the FAWP or step-down fixes must be considered. Calculating the distance to the FAWP may be required in order to descend at the proper location.

When within 2 NM of the FAWP with the approach mode armed, the approach mode will switch to active, which results in RAIM and CDI sensitivity changing to the approach mode. Beginning 2 NM prior to the FAWP, the full scale CDI sensitivity will change smoothly from ±1 NM to ±0.3 NM at the FAWP. As sensitivity changes from ±1 NM to ±0.3 NM approaching the FAWP, and the CDI not centered, the corresponding increase in CDI displacement may give the impression the aircraft is moving further away from the intended course even though it is on an acceptable intercept heading. If digital track displacement information (cross-track error) is available in the approach mode, it may help the pilot remain position oriented in this situation. Being established on the final approach course prior to the beginning of the sensitivity change at 2 NM will help prevent problems in interpreting the CDI display during ramp-down. Requesting or accepting vectors, which will cause the aircraft to intercept the final approach course within 2 NM of the FAWP, is not recommended.

Incorrect inputs into the GPS receiver are especially critical during approaches. In some cases, an incorrect entry can cause the receiver to leave the approach mode. Overriding an automatically selected sensitivity during an approach will cancel the approach mode annunciation. If the approach mode is not armed by 2 NM prior to the FAWP, the approach mode will not become active at 2 NM prior to the FAWP and the equipment will flag. In these conditions, the RAIM and CDI sensitivity will not ramp down, and the pilot should not descend to minimum descent altitude (MDA), but fly to the MAWP and execute a missed approach. The approach active annunciator and/or the receiver should be checked to ensure the approach mode is active prior to the FAWP.

A GPS missed approach requires pilot action to sequence the receiver past the MAWP to the missed approach portion of the procedure. The pilot must be thoroughly familiar with the activation procedure for the particular GPS receiver installed in the aircraft and must initiate appropriate action after the MAWP. Activating the missed approach prior to the MAWP will cause CDI sensitivity to change immediately to terminal (±1 NM) sensitivity, and the receiver will continue to navigate to the MAWP. The receiver will not sequence past the MAWP. Turns should not begin prior to the MAWP. If the missed approach is not activated, the GPS receiver will display an extension of the inbound final approach course and the along track distance (ATD) will increase from the MAWP until it is manually sequenced after crossing the MAWP.

Missed approach routings in which the first track is via a course rather than direct to the next WP require additional action by the pilot to set the course. Being familiar with all of the required inputs is especially critical during this phase of flight.

Departures and Instrument Departure Procedures (DPs)

The GPS receiver must be set to terminal (±1 NM) CDI sensitivity and the navigation routes contained in the database in order to fly published IFR charted departures and DPs. Terminal RAIM should be provided automatically by the receiver. (Terminal RAIM for departure may not be available unless the WPs are part of the active flight plan rather than proceeding direct to the first destination.) Certain segments of a DP may require some manual intervention by the pilot, especially when radar vectored to a course or required to intercept a specific course to a WP. The database may not contain all of the transitions or departures from all runways and some GPS receivers do not contain DPs in the database. It is necessary that helicopter procedures be flown at 70 knots or less since helicopter departure procedures and missed approaches use a 20:1 obstacle clearance surface (OCS), which is double the fixed-wing OCS. Turning areas are based on this speed also. Missed approach routings in which the first track is via a course rather than direct to the next WP require additional action by the pilot to set the course. Being familiar with all of the required inputs is especially critical during this phase of flight.

GPS Errors

Normally, with 30 satellites in operation, the GPS constellation is expected to be available continuously worldwide. Whenever there are fewer than 24 operational satellites, GPS navigational capability may not be available at certain geographic locations. Loss of signals may also occur in valleys surrounded by high terrain, and any time the aircraft's GPS antenna is "shadowed" by the aircraft's structure (e.g., when the aircraft is banked).

Certain receivers, transceivers, mobile radios, and portable receivers can cause signal interference. Some VHF transmissions may cause "harmonic interference." Pilots can isolate the interference by relocating nearby portable receivers, changing frequencies, or turning off suspected causes of the interference while monitoring the receiver's signal quality data page.

GPS position data can be affected by equipment characteristics and various geometric factors, which typically cause errors of less than 100 feet. Satellite atomic clock inaccuracies, receiver/processors, signals reflected from hard objects (multi-path), ionospheric and tropospheric delays, and satellite data transmission errors may cause small position errors or momentary loss of the GPS signal.

System Status

The status of GPS satellites is broadcast as part of the data message transmitted by the GPS satellites. GPS status information is also available by means of the United States Coast Guard navigation information service: (703) 313-5907, or on the internet at http://www.navcen.uscg.gov/. Additionally, satellite status is available through the Notice to Airmen (NOTAM) system.

The GPS receiver verifies the integrity (usability) of the signals received from the GPS constellation through receiver autonomous integrity monitoring (RAIM) to determine if a satellite is providing corrupted information. At least one satellite, in addition to those required for navigation, must be in view for the receiver to perform the RAIM function; thus, RAIM needs a minimum of five satellites in view, or four satellites and a barometric altimeter (baro-aiding) to detect an integrity anomaly. For receivers capable of doing so, RAIM needs six satellites in view (or five satellites with baro-aiding) to isolate the corrupt satellite signal and remove it from the navigation solution.

RAIM messages vary somewhat between receivers; however, there are two most commonly used types. One type indicates that there are not enough satellites available to provide RAIM integrity monitoring and another type indicates that the RAIM integrity monitor has detected a potential error that exceeds the limit for the current phase of flight. Without RAIM capability, the pilot has no assurance of the accuracy of the GPS position.

Selective Availability. Selective Availability (SA) is a method by which the accuracy of GPS is intentionally degraded. This feature is designed to deny hostile use of precise GPS positioning data. SA was discontinued on May 1, 2000, but many GPS receivers are designed to assume that SA is still active. New receivers may take advantage of the discontinuance of SA based on the performance values in ICAO Annex 10, and do not need to be designed to operate outside of that performance.

GPS Familiarization

Pilots should practice GPS approaches under visual meteorological conditions (VMC) until thoroughly proficient with all aspects of their equipment (receiver and installation) prior to attempting flight by IFR in instrument meteorological conditions (IMC). Some of the tasks which the pilot should practice are:

1. Utilizing the receiver autonomous integrity monitoring (RAIM) prediction function;
2. Inserting a DP into the flight plan, including setting terminal CDI sensitivity, if required, and the conditions under which terminal RAIM is available for departure (some receivers are not DP or STAR capable);
3. Programming the destination airport;
4. Programming and flying the overlay approaches (especially procedure turns and arcs);
5. Changing to another approach after selecting an approach;
6. Programming and flying "direct" missed approaches;
7. Programming and flying "routed" missed approaches;
8. Entering, flying, and exiting holding patterns, particularly on overlay approaches with a second WP in the holding pattern;
9. Programming and flying a "route" from a holding pattern;
10. Programming and flying an approach with radar vectors to the intermediate segment;
11. Indication of the actions required for RAIM failure both before and after the FAWP; and
12. Programming a radial and distance from a VOR (often used in departure instructions).

Differential Global Positioning Systems (DGPS)

Differential global positioning systems (DGPS) are designed to improve the accuracy of global navigation satellite systems (GNSS) by measuring changes in variables to provide satellite positioning corrections.

Because multiple receivers receiving the same set of satellites produce similar errors, a reference receiver placed at a known location can compute its theoretical position accurately and can compare that value to the measurements provided by the navigation satellite signals. The difference in measurement between the two signals is an error that can be corrected by providing a reference signal correction.

As a result of this differential input accuracy of the satellite system can be increased to meters. The Wide Area Augmentation System (WAAS) and Local Area Augmentation System (LAAS) are examples of differential global positioning systems.

Wide Area Augmentation System (WAAS)

The WAAS is designed to improve the accuracy, integrity, and availability of GPS signals. WAAS allows GPS to be used, as the aviation navigation system, from takeoff through Category I precision approaches. The International Civil Aviation Organization (ICAO) has defined Standards for satellite-based augmentation systems (SBAS), and Japan and Europe are building similar systems that are planned to be interoperable with WAAS: EGNOS, the European Geostationary Navigation Overlay System, and MSAS, the Japanese Multifunctional Transport Satellite (MTSAT) Satellite-based Augmentation System. The result will be a worldwide seamless navigation capability similar to GPS but with greater accuracy, availability, and integrity.

Unlike traditional ground-based navigation aids, WAAS will cover a more extensive service area in which surveyed wide-area ground reference stations are linked to the WAAS network. Signals from the GPS satellites are monitored by these stations to determine satellite clock and ephemeris corrections. Each station in the network relays the data to a wide-area master station where the correction information is computed. A correction message is prepared and uplinked to a geostationary satellite (GEO) via a ground uplink and then broadcast on the same frequency as GPS to WAAS receivers within the broadcast coverage area. *[Figure 7-30]*

In addition to providing the correction signal, WAAS provides an additional measurement to the aircraft receiver, improving the availability of GPS by providing, in effect, an additional GPS satellite in view. The integrity of GPS is improved through real-time monitoring, and the accuracy is improved by providing differential corrections to reduce errors. *[Figure 7-31]* As a result, performance improvement is sufficient to enable approach procedures with GPS/WAAS glide paths. At this time the FAA has completed installation of 25 wide area ground reference systems, two master stations, and four ground uplink stations.

General Requirements

WAAS avionics must be certified in accordance with TSO-C145A, Airborne Navigation Sensors Using the GPS Augmented by the WAAS; or TSO-146A for stand-alone systems. GPS/WAAS operation must be conducted in accordance with the FAA-approved aircraft flight manual (AFM) and flight manual supplements. Flight manual supplements must state the level of approach procedure that the receiver supports.

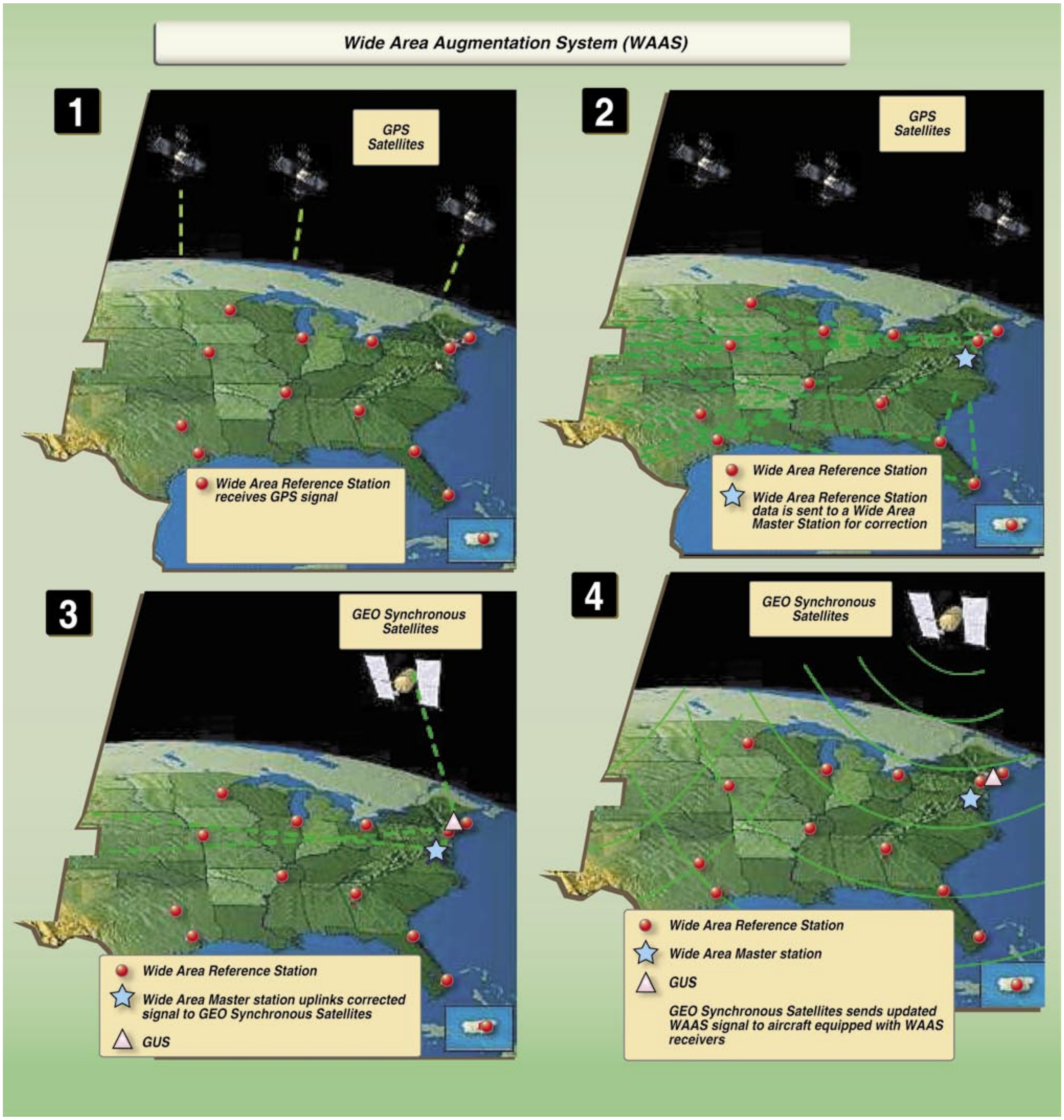
Wide Area Augmentation System (WAAS)
1
GPS Satellites
Wide Area Reference Station receives GPS signal
2
GPS Satellites
Wide Area Reference Station
Wide Area Reference Station data is sent to a Wide Area Master Station for correction
3
GEO Synchronous Satellites
Wide Area Reference Station
Wide Area Master station uplinks corrected signal to GEO Synchronous Satellites
GUS
4
GEO Synchronous Satellites
Wide Area Reference Station
Wide Area Master station
GUS
GEO Synchronous Satellites sends updated WAAS signal to aircraft equipped with WAAS receivers

Figure 7-30. *WAAS Satellite Representation.*

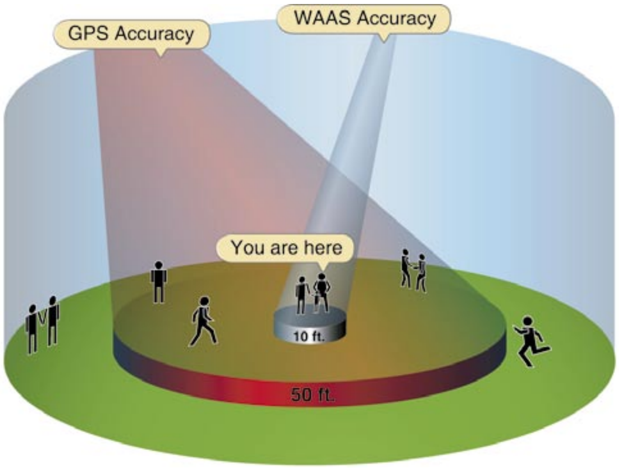
GPS Accuracy
WAAS Accuracy
You are here
10 ft.
50 ft.

Figure 7-31. *WAAS Satellite Representation.*

Instrument Approach Capabilities

WAAS receivers support all basic GPS approach functions and will provide additional capabilities with the key benefit to generate an electronic glide path, independent of ground equipment or barometric aiding. This eliminates several problems such as cold temperature effects, incorrect altimeter setting or lack of a local altimeter source, and allows approach procedures to be built without the cost of installing ground stations at each airport. A new class of approach procedures which provide vertical guidance requirements for precision approaches has been developed to support satellite navigation use for aviation applications. These new procedures called Approach with Vertical Guidance (APV) include approaches such as the LNAV/VNAV procedures presently being flown with barometric vertical navigation.

Local Area Augmentation System (LAAS)

LAAS is a ground-based augmentation system which uses a GPS reference facility located on or in the vicinity of the airport being serviced. This facility has a reference receiver that measures GPS satellite pseudo-range and timing and retransmits the signal. Aircraft landing at LAAS-equipped airports are able to conduct approaches to Category I level and above for properly equipped aircraft. *[Figures 7-32* and *7-33]*

Inertial Navigation System (INS)

Inertial Navigation System (INS) is a system that navigates precisely without any input from outside of the aircraft. It is fully self-contained. The INS is initialized by the pilot, who enters into the system the exact location of the aircraft on the ground before the flight. The INS is also programmed with WPs along the desired route of flight.

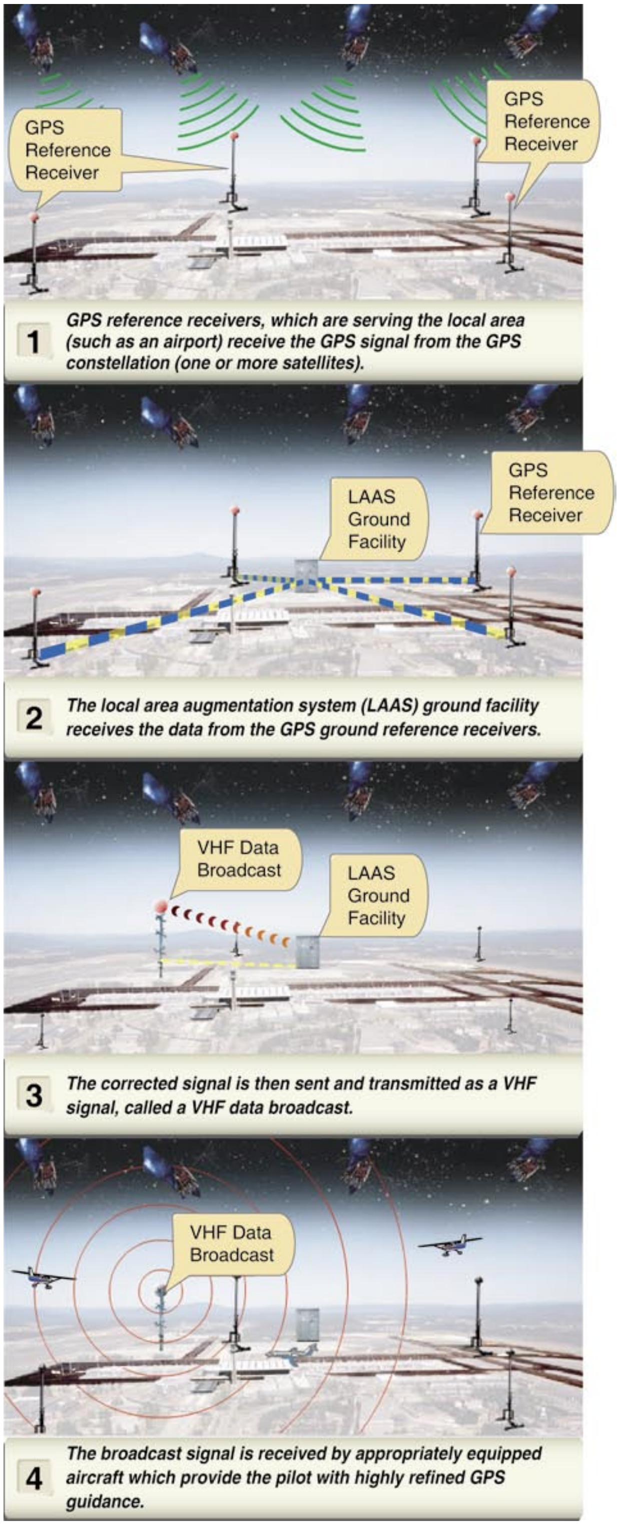
GPS Reference Receiver
GPS Reference Receiver
1
GPS reference receivers, which are serving the local area (such as an airport) receive the GPS signal from the GPS constellation (one or more satellites).
LAAS Ground Facility
GPS Reference Receiver
2
The local area augmentation system (LAAS) ground facility receives the data from the GPS ground reference receivers.
VHF Data Broadcast
LAAS Ground Facility
3
The corrected signal is then sent and transmitted as a VHF signal, called a VHF data broadcast.
VHF Data Broadcast
4
The broadcast signal is received by appropriately equipped aircraft which provide the pilot with highly refined GPS guidance.

Figure 7-32. *LAAS Representation.*

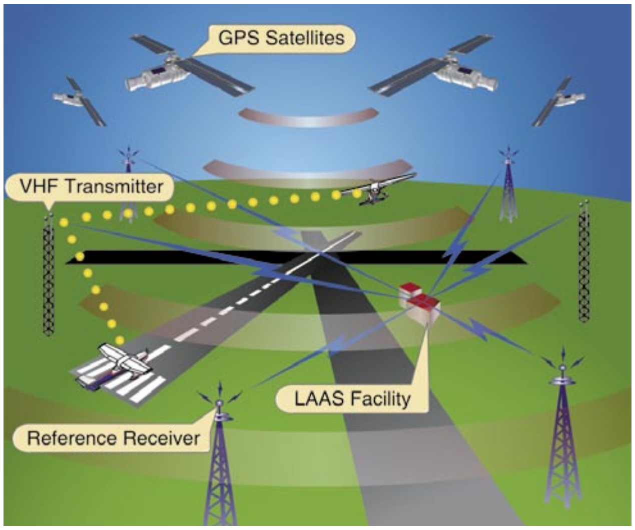

Figure 7-33. *LAAS Representation.*

INS Components

INS is considered a stand-alone navigation system, especially when more than one independent unit is onboard. The airborne equipment consists of an accelerometer to measure acceleration—which, when integrated with time, gives velocity—and gyros to measure direction.

Later versions of the INS, called inertial reference systems (IRS) utilize laser gyros and more powerful computers; therefore, the accelerometer mountings no longer need to be kept level and aligned with true north. The computer system can handle the added workload of dealing with the computations necessary to correct for gravitational and directional errors. Consequently, these newer systems are sometimes called strap down systems, as the accelerometers and gyros are strapped down to the airframe, rather than being mounted on a structure that stays fixed with respect to the horizon and true north.

INS Errors

The principal error associated with INS is degradation of position with time. INS computes position by starting with accurate position input which is changed continuously as accelerometers and gyros provide speed and direction inputs. Both accelerometers and gyros are subject to very small errors; as time passes, those errors probably accumulate.

While the best INS/IRS display errors of 0.1 to 0.4 NM after flights across the North Atlantic of 4 to 6 hours, smaller and less expensive systems are being built that show errors of 1 to 2 NM per hour. This accuracy is more than sufficient for a navigation system that can be combined with and updated by GPS. The synergy of a navigation system consisting of an INS/IRS unit in combination with a GPS resolves the errors and weaknesses of both systems. GPS is accurate all the time it is working but may be subject to short and periodic outages. INS is made more accurate because it is continually updated and continues to function with good accuracy if the GPS has moments of lost signal.

Instrument Approach Systems

Most navigation systems approved for en route and terminal operations under IFR, such as VOR, NDB, and GPS, may also be approved to conduct IAPs. The most common systems in use in the United States are the ILS, simplified directional facility (SDF), localizer directional aid (LDA), and microwave landing system (MLS). These systems operate independently of other navigation systems. There are new systems being developed, such as WAAS and LAAS. Other systems have been developed for special use.

Instrument Landing Systems (ILS)

The ILS system provides both course and altitude guidance to a specific runway. The ILS system is used to execute a precision instrument approach procedure or precision approach. *[Figure 7-34]* The system consists of the following components:

1. A localizer providing horizontal (left/right) guidance along the extended centerline of the runway.
2. A glide slope (GS) providing vertical (up/down) guidance toward the runway touchdown point, usually at a 3° slope.
3. Marker beacons providing range information along the approach path.
4. Approach lights assisting in the transition from instrument to visual flight.

The following supplementary elements, though not specific components of the system, may be incorporated to increase safety and utility:

1. Compass locators providing transition from en route NAVAIDs to the ILS system and assisting in holding procedures, tracking the localizer course, identifying the marker beacon sites, and providing a FAF for ADF approaches.
2. DME collocated with the GS transmitter providing positive distance-to-touchdown information or DME associated with another nearby facility (VOR or stand-alone), if specified in the approach procedure.

ILS approaches are categorized into three different types of approaches based on the equipment at the airport and the experience level of the pilot. Category I approaches provide for approach height above touchdown of not less than 200 feet. Category II approaches provide for approach to a height above

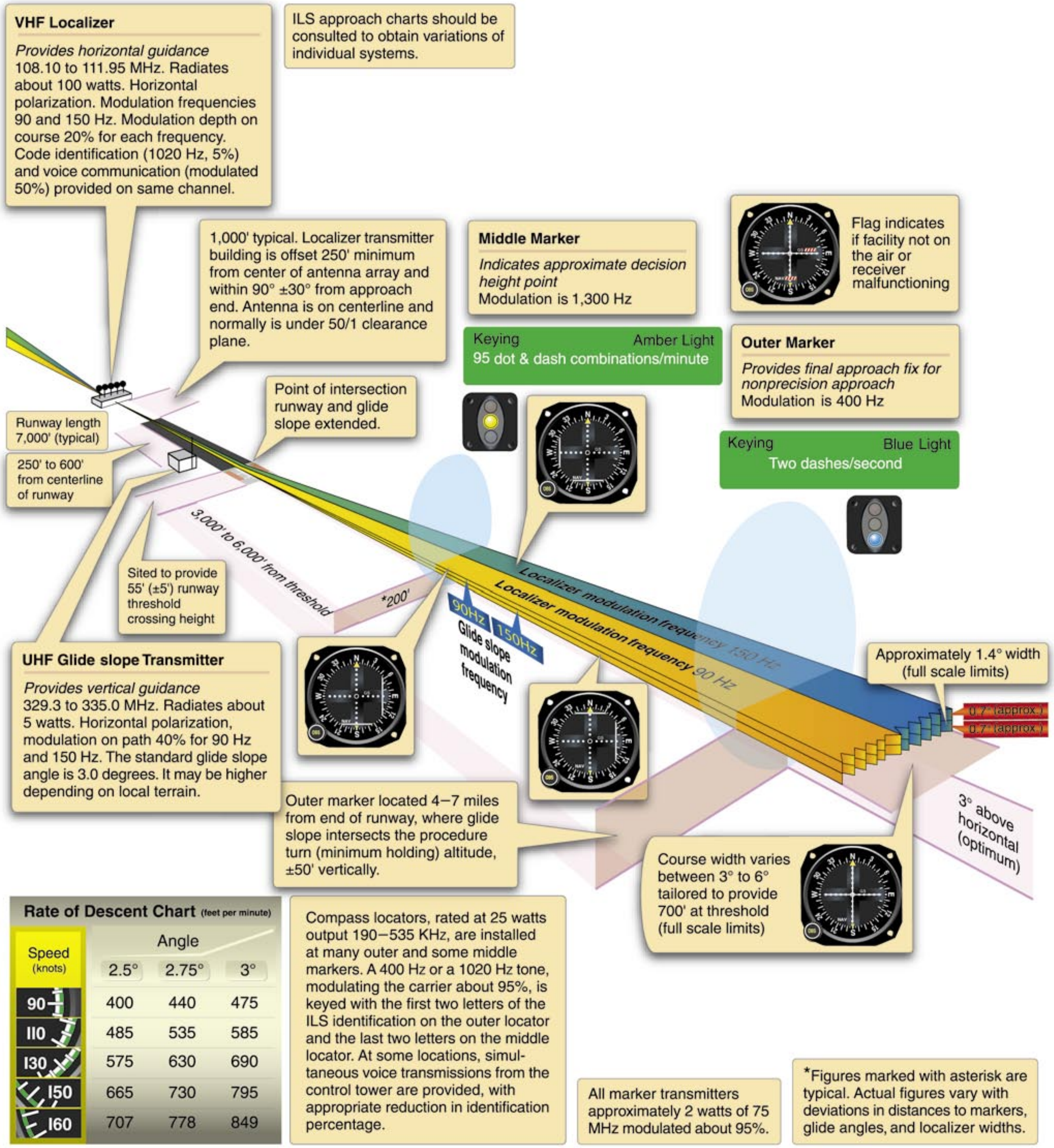
3,000' to 6,000' from threshold
*200'
90Hz
150Hz
Glide slope modulation frequency
Localizer modulation frequency 150 Hz
Localizer modulation frequency 90 Hz
0.7° (approx.)
0.7° (approx.)
3° above horizontal (optimum)

VHF Localizer

Provides horizontal guidance
108.10 to 111.95 MHz. Radiates about 100 watts. Horizontal polarization. Modulation frequencies 90 and 150 Hz. Modulation depth on course 20% for each frequency. Code identification (1020 Hz, 5%) and voice communication (modulated 50%) provided on same channel.

ILS approach charts should be consulted to obtain variations of individual systems.

1,000' typical. Localizer transmitter building is offset 250' minimum from center of antenna array and within 90° ±30° from approach end. Antenna is on centerline and normally is under 50/1 clearance plane.

Middle Marker

Indicates approximate decision height point
Modulation is 1,300 Hz

Keying Amber Light
95 dot & dash combinations/minute

Flag indicates if facility not on the air or receiver malfunctioning

Outer Marker

Provides final approach fix for nonprecision approach
Modulation is 400 Hz

Keying Blue Light
Two dashes/second

Runway length 7,000' (typical)

250' to 600' from centerline of runway

Point of intersection runway and glide slope extended.

Sited to provide 55' (±5') runway threshold crossing height

UHF Glide slope Transmitter

Provides vertical guidance
329.3 to 335.0 MHz. Radiates about 5 watts. Horizontal polarization, modulation on path 40% for 90 Hz and 150 Hz. The standard glide slope angle is 3.0 degrees. It may be higher depending on local terrain.

Approximately 1.4° width (full scale limits)

Outer marker located 4–7 miles from end of runway, where glide slope intersects the procedure turn (minimum holding) altitude, ±50' vertically.

Course width varies between 3° to 6° tailored to provide 700' at threshold (full scale limits)

Rate of Descent Chart (feet per minute)

Speed (knots)	Angle 2.5°	2.75°	3°
90	400	440	475
110	485	535	585
130	575	630	690
150	665	730	795
160	707	778	849

Compass locators, rated at 25 watts output 190–535 KHz, are installed at many outer and some middle markers. A 400 Hz or a 1020 Hz tone, modulating the carrier about 95%, is keyed with the first two letters of the ILS identification on the outer locator and the last two letters on the middle locator. At some locations, simultaneous voice transmissions from the control tower are provided, with appropriate reduction in identification percentage.

All marker transmitters approximately 2 watts of 75 MHz modulated about 95%.

*Figures marked with asterisk are typical. Actual figures vary with deviations in distances to markers, glide angles, and localizer widths.

Figure 7-34. *Instrument Landing Systems.*

touchdown of not less than 100 feet. Category III approaches provide lower minimums for approaches without a decision height minimum. While pilots need only be instrument rated and the aircraft be equipped with the appropriate airborne equipment to execute Category I approaches, Category II and III approaches require special certification for the pilots, ground equipment, and airborne equipment.

ILS Components

Ground Components

The ILS uses a number of different ground facilities. These facilities may be used as a part of the ILS system, as well as part of another approach. For example, the compass locator may be used with NDB approaches.

Localizer

The localizer (LOC) ground antenna array is located on the extended centerline of the instrument runway of an airport, located at the departure end of the runway to prevent it from being a collision hazard. This unit radiates a field pattern, which develops a course down the centerline of the runway toward the middle markers (MMs) and outer markers (OMs), and a similar course along the runway centerline in the opposite direction. These are called the front and back courses, respectively. The localizer provides course guidance, transmitted at 108.1 to 111.95 MHz (odd tenths only), throughout the descent path to the runway threshold from a distance of 18 NM from the antenna to an altitude of 4,500 feet above the elevation of the antenna site. *[Figure 7-35]*

The localizer course width is defined as the angular displacement at any point along the course between a full "fly-left" (CDI needle fully deflected to the left) and a full "fly-right" indication (CDI needle fully deflected to the right). Each localizer facility is audibly identified by a three-letter designator, transmitted at frequent regular intervals. The ILS identification is preceded by the letter "I" (two dots). For example, the ILS localizer at Springfield, Missouri transmits the identifier ISGF. The localizer includes a voice feature on its frequency for use by the associated ATC facility in issuing approach and landing instructions.

The localizer course is very narrow, normally 5°. This results in high needle sensitivity. With this course width, a full-scale deflection shows when the aircraft is 2.5° to either side of the centerline. This sensitivity permits accurate orientation to the landing runway. With no more than one-quarter scale deflection maintained, the aircraft will be aligned with the runway.

Glide Slope (GS)

GS describes the systems that generate, receive, and indicate the ground facility radiation pattern. The glide path is the straight, sloped line the aircraft should fly in its descent from where the GS intersects the altitude used for approaching the FAF, to the runway touchdown zone. The GS equipment is housed in a building approximately 750 to 1,250 feet down the runway from the approach end of the runway, and between 400 and 600 feet to one side of the centerline.

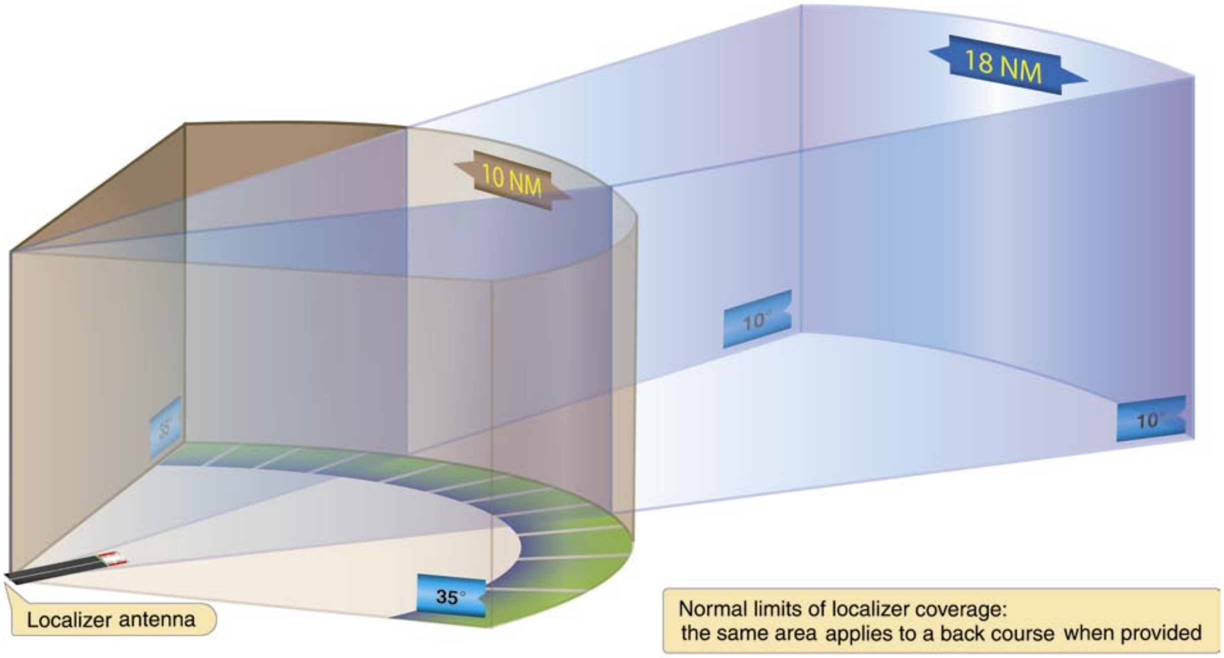
18 NM
10 NM
10°
10°
35°
35°
Localizer antenna
Normal limits of localizer coverage:
the same area applies to a back course when provided

Figure 7-35. *Localizer Coverage Limits.*

The course projected by the GS equipment is essentially the same as would be generated by a localizer operating on its side. The GS projection angle is normally adjusted to 2.5° to 3.5° above horizontal, so it intersects the MM at about 200 feet and the OM at about 1,400 feet above the runway elevation. At locations where standard minimum obstruction clearance cannot be obtained with the normal maximum GS angle, the GS equipment is displaced farther from the approach end of the runway if the length of the runway permits; or, the GS angle may be increased up to 4°.

Unlike the localizer, the GS transmitter radiates signals only in the direction of the final approach on the front course. The system provides no vertical guidance for approaches on the back course. The glide path is normally 1.4° thick. At 10 NM from the point of touchdown, this represents a vertical distance of approximately 1,500 feet, narrowing to a few feet at touchdown.

Marker Beacons

Two VHF marker beacons, outer and middle, are normally used in the ILS system. *[Figure 7-36]* A third beacon, the inner, is used where Category II operations are certified. A marker beacon may also be installed to indicate the FAF on the ILS back course.

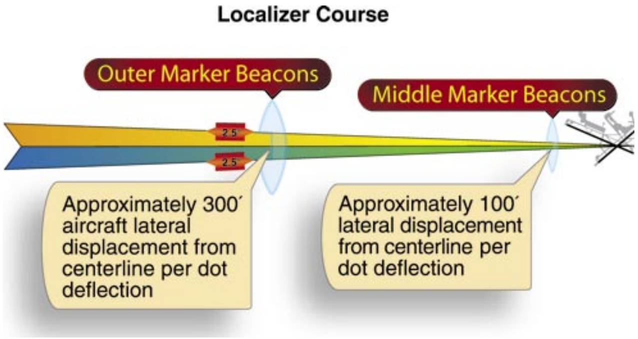

Approximately 300′ aircraft lateral displacement from centerline per dot deflection

Approximately 100′ lateral displacement from centerline per dot deflection

Figure 7-36. *Localizer receiver indications and aircraft displacement.*

The OM is located on the localizer front course 4–7 miles from the airport to indicate a position at which an aircraft, at the appropriate altitude on the localizer course, will intercept the glide path. The MM is located approximately 3,500 feet from the landing threshold on the centerline of the localizer front course at a position where the GS centerline is about 200 feet above the touchdown zone elevation. The inner marker (IM), where installed, is located on the front course between the MM and the landing threshold. It indicates the point at which an aircraft is at the decision height on the glide path during a Category II ILS approach. The back-course marker, where installed, indicates the back-course FAF.

Compass Locator

Compass locators are low-powered NDBs and are received and indicated by the ADF receiver. When used in conjunction with an ILS front course, the compass locator facilities are collocated with the outer and/or MM facilities. The coding identification of the outer locator consists of the first two letters of the three-letter identifier of the associated LOC. For example, the outer locator at Dallas/Love Field (DAL) is identified as “DA.” The middle locator at DAL is identified by the last two letters “AL.”

Approach Lighting Systems (ALS)

Normal approach and letdown on the ILS is divided into two distinct stages: the instrument approach stage using only radio guidance, and the visual stage, when visual contact with the ground runway environment is necessary for accuracy and safety. The most critical period of an instrument approach, particularly during low ceiling/visibility conditions, is the point at which the pilot must decide whether to land or execute a missed approach. As the runway threshold is approached, the visual glide path will separate into individual lights. At this point, the approach should be continued by reference to the runway touchdown zone markers. The ALS provides lights that will penetrate the atmosphere far enough from touchdown to give directional, distance, and glide path information for safe visual transition.

Visual identification of the ALS by the pilot must be instantaneous, so it is important to know the type of ALS before the approach is started. Check the instrument approach chart and the A/FD for the particular type of lighting facilities at the destination airport before any instrument flight. With reduced visibility, rapid orientation to a strange runway can be difficult, especially during a circling approach to an airport with minimum lighting facilities, or to a large terminal airport located in the midst of distracting city and ground facility lights. Some of the most common ALS systems are shown in *Figure 7-37*.

A high-intensity flasher system, often referred to as “the rabbit,” is installed at many large airports. The flashers consist of a series of brilliant blue-white bursts of light flashing in sequence along the approach lights, giving the effect of a ball of light traveling towards the runway. Typically, “the rabbit” makes two trips toward the runway per second.

Runway end identifier lights (REIL) are installed for rapid and positive identification of the approach end of an instrument runway. The system consists of a pair of synchronized flashing lights placed laterally on each side of the runway threshold facing the approach area.

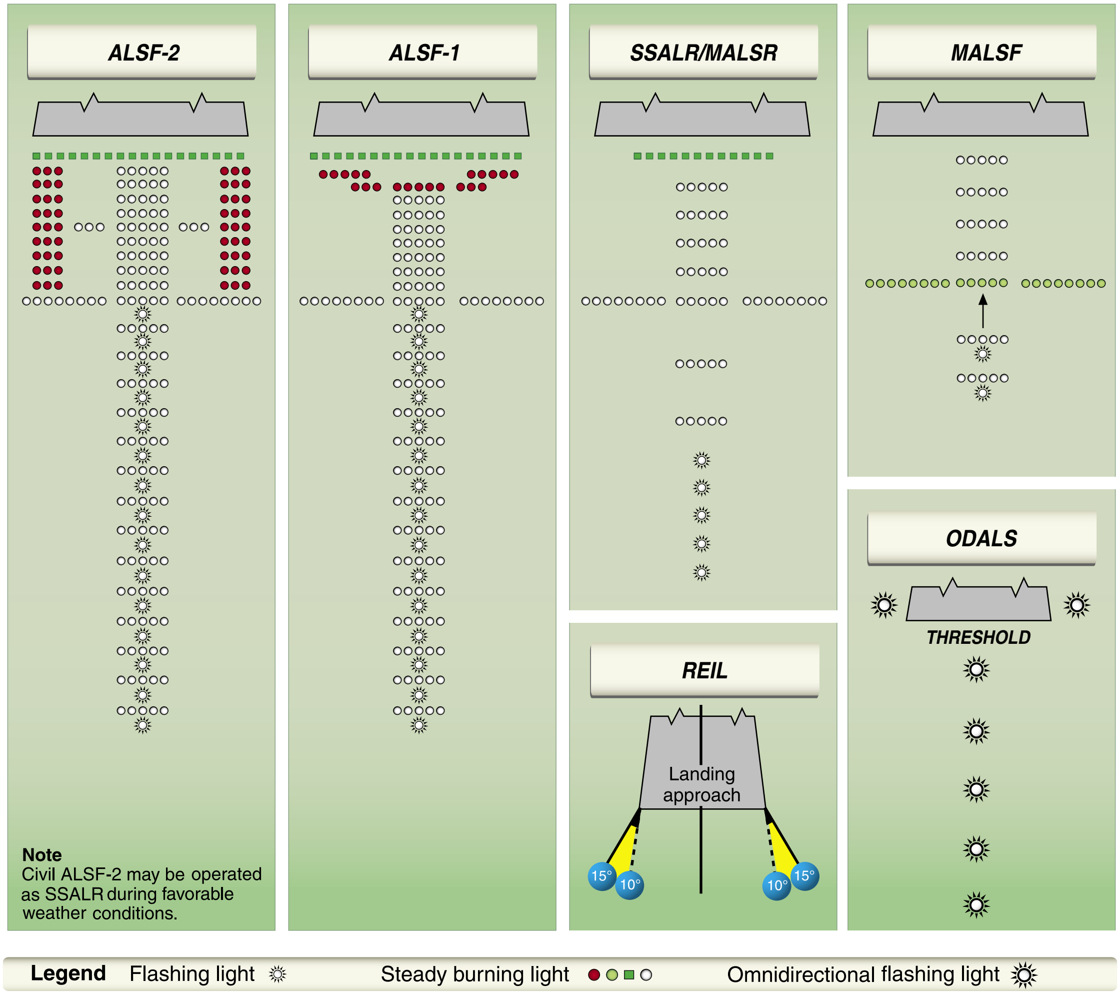

Note
Civil ALSF-2 may be operated as SSALR during favorable weather conditions.

ALSF—Approach light system with sequenced flashing lights
SSALR—Simplified short approach light system with runway alignment indicator lights
MALSR—Medium intensity approach light system with runway alignment indicator lights
REIL—Runway end identification lights
MALSF—Medium intensity approach light system with sequenced flashing lights (and runway alignment)
ODALS—Omnidirectional approach light system

Figure 7-37. *Precision and Nonprecision ALS Configuration.*

The visual approach slope indicator (VASI) gives visual descent guidance information during the approach to a runway. The standard VASI consists of light bars that project a visual glide path, which provides safe obstruction clearance within the approach zone. The normal GS angle is 3°; however, the angle may be as high as 4.5° for proper obstacle clearance. On runways served by ILS, the VASI angle normally coincides with the electronic GS angle. Visual left/right course guidance is obtained by alignment with the runway lights. The standard VASI installation consists of either 2-, 3-, 4-, 6-, 12-, or 16-light units arranged in downwind and upwind light bars. Some airports serving long-bodied aircraft have three-bar VASIs which provide two visual glidepaths to the same runway. The first glide path encountered is the same as provided by the standard VASI. The second glide path is about 25 percent higher than the first and is designed for the use of pilots of long-bodied aircraft.

The basic principle of VASI is that of color differentiation between red and white. Each light projects a beam having a white segment in the upper part and a red segment in the lower part of the beam. From a position above the glide path the pilot sees both bars as white. Lowering the aircraft with respect to the glide path, the color of the upwind bars changes from white to pink to red. When on the proper glide path, the landing aircraft will overshoot the downwind bars and undershoot the upwind bars. Thus the downwind (closer) bars are seen as white and the upwind bars as red. From a position below the glide path, both light bars are seen as red. Moving up to the glide path, the color of the downwind

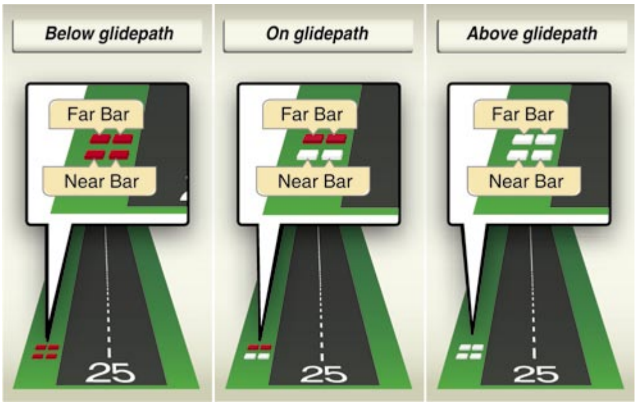

Figure 7-38. *Standard two-bar VASI.*

bars changes from red to pink to white. When below the glide path, as indicated by a distinct all-red signal, a safe obstruction clearance might not exist. A standard two-bar VASI is illustrated in *Figure 7-38*.

ILS Airborne Components

Airborne equipment for the ILS system includes receivers for the localizer, GS, marker beacons, ADF, DME, and the respective indicator instruments.

The typical VOR receiver is also a localizer receiver with common tuning and indicating equipment. Some receivers have separate function selector switches, but most switch between VOR and LOC automatically by sensing if odd tenths between 108 and 111.95 MHz have been selected. Otherwise, tuning of VOR and localizer frequencies is accomplished with the same knobs and switches, and the CDI indicates "on course" as it does on a VOR radial.

Though some GS receivers are tuned separately, in a typical installation the GS is tuned automatically to the proper frequency when the localizer is tuned. Each of the 40 localizer channels in the 108.10 to 111.95 MHz band is paired with a corresponding GS frequency.

When the localizer indicator also includes a GS needle, the instrument is often called a cross-pointer indicator. The crossed horizontal (GS) and vertical (localizer) needles are free to move through standard five-dot deflections to indicate position on the localizer course and glide path.

When the aircraft is on the glide path, the needle is horizontal, overlying the reference dots. Since the glide path is much narrower than the localizer course (approximately 1.4° from full up to full down deflection), the needle is very sensitive to displacement of the aircraft from on-path alignment. With the proper rate of descent established upon GS interception, very small corrections keep the aircraft aligned.

The localizer and GS warning flags disappear from view on the indicator when sufficient voltage is received to actuate the needles. The flags show when an unstable signal or receiver malfunction occurs.

The OM is identified by a low-pitched tone, continuous dashes at the rate of two per second, and a purple/blue marker beacon light. The MM is identified by an intermediate tone, alternate dots and dashes at the rate of 95 dot/dash combinations per minute, and an amber marker beacon light. The IM, where installed, is identified by a high-pitched tone, continuous dots at the rate of six per second, and a white marker beacon light. The back-course marker (BCM), where installed, is identified by a high-pitched tone with two dots at a rate of 72 to 75 two-dot combinations per minute, and a white marker beacon light. Marker beacon receiver sensitivity is selectable as high or low on many units. The low-sensitivity position gives the sharpest indication of position and should be used during an approach. The high-sensitivity position provides an earlier warning that the aircraft is approaching the marker beacon site.

ILS Function

The localizer needle indicates, by deflection, whether the aircraft is right or left of the localizer centerline, regardless of the position or heading of the aircraft. Rotating the OBS has no effect on the operation of the localizer needle, although it is useful to rotate the OBS to put the LOC inbound course under the course index. When inbound on the front course, or outbound on the back course, the course indication remains directional. (See *Figure 7-39,* aircraft C, D, and E.)

Unless the aircraft has reverse sensing capability and it is in use, when flying inbound on the back course or outbound on the front course, heading corrections to on-course are made opposite the needle deflection. This is commonly described as "flying away from the needle." (See *Figure 7-39*, aircraft A and B.) Back course signals should not be used for an approach unless a back course approach procedure is published for that particular runway and the approach is authorized by ATC.

Once you have reached the localizer centerline, maintain the inbound heading until the CDI moves off center. Drift corrections should be small and reduced proportionately as the course narrows. By the time you reach the OM, your drift correction should be established accurately enough on a well-executed approach to permit completion of the approach, with heading corrections no greater then 2°.

The heaviest demand on pilot technique occurs during descent from the OM to the MM, when you maintain the localizer course, adjust pitch attitude to maintain the

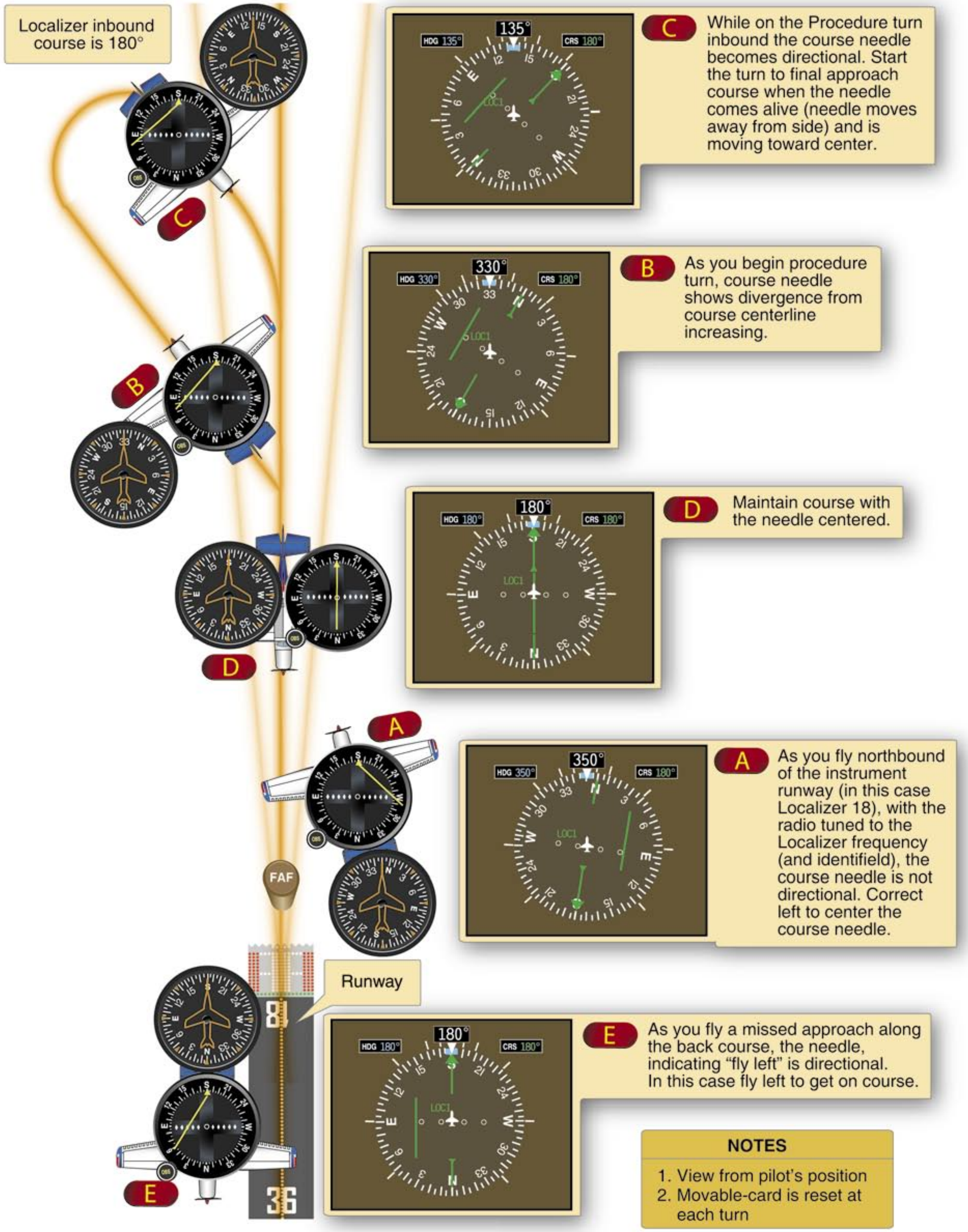

Localizer inbound course is 180°

C While on the Procedure turn inbound the course needle becomes directional. Start the turn to final approach course when the needle comes alive (needle moves away from side) and is moving toward center.

B As you begin procedure turn, course needle shows divergence from course centerline increasing.

D Maintain course with the needle centered.

A As you fly northbound of the instrument runway (in this case Localizer 18), with the radio tuned to the Localizer frequency (and identifield), the course needle is not directional. Correct left to center the course needle.

E As you fly a missed approach along the back course, the needle, indicating "fly left" is directional. In this case fly left to get on course.

NOTES

1. View from pilot's position
2. Movable-card is reset at each turn

HDG 135°
135°
CRS 180°
LOC1
HDG 330°
330°
CRS 180°
HDG 180°
180°
CRS 180°
HDG 350°
350°
CRS 180°
FAF
Runway
18
36
C
B
D
A
E
OBS

Figure 7-39. *Localizer Course Indications. To follow indications displayed in the aircraft, start from A and proceed through E.*

proper rate of descent, and adjust power to maintain proper airspeed. Simultaneously, the altimeter must be checked and preparation made for visual transition to land or for a missed approach. You can appreciate the need for accurate instrument interpretation and aircraft control within the ILS as a whole, when you notice the relationship between CDI and glide path needle indications, and aircraft displacement from the localizer and glide path centerlines.

Deflection of the GS needle indicates the position of the aircraft with respect to the glide path. When the aircraft is above the glide path, the needle is deflected downward. When the aircraft is below the glide path, the needle is deflected upward. *[Figure 7-40]*

ILS Errors

The ILS and its components are subject to certain errors, which are listed below. Localizer and GS signals are subject to the same type of bounce from hard objects as space waves.

1. Reflection. Surface vehicles and even other aircraft flying below 5,000 feet above ground level (AGL) may disturb the signal for aircraft on the approach.
2. False courses. In addition to the desired course, GS facilities inherently produce additional courses at higher vertical angles. The angle of the lowest of these false courses will occur at approximately 9°–12°. An aircraft flying the LOC/GS course at a constant altitude would observe gyrations of both the GS needle and GS warning flag as the aircraft passed through the various false courses. Getting established on one of these false courses will result in either confusion (reversed GS needle indications) or in the need for a very high descent rate. However, if the approach is conducted at the altitudes specified on the appropriate approach chart, these false courses will not be encountered.

Marker Beacons

The very low power and directional antenna of the marker beacon transmitter ensures that the signal will not be received any distance from the transmitter site. Problems with signal reception are usually caused by the airborne receiver not being turned on, or by incorrect receiver sensitivity.

Some marker beacon receivers, to decrease weight and cost, are designed without their own power supply. These units utilize a power source from another radio in the avionics stack, often the ADF. In some aircraft, this requires the ADF to be turned on in order for the marker beacon receiver to function, yet no warning placard is required. Another

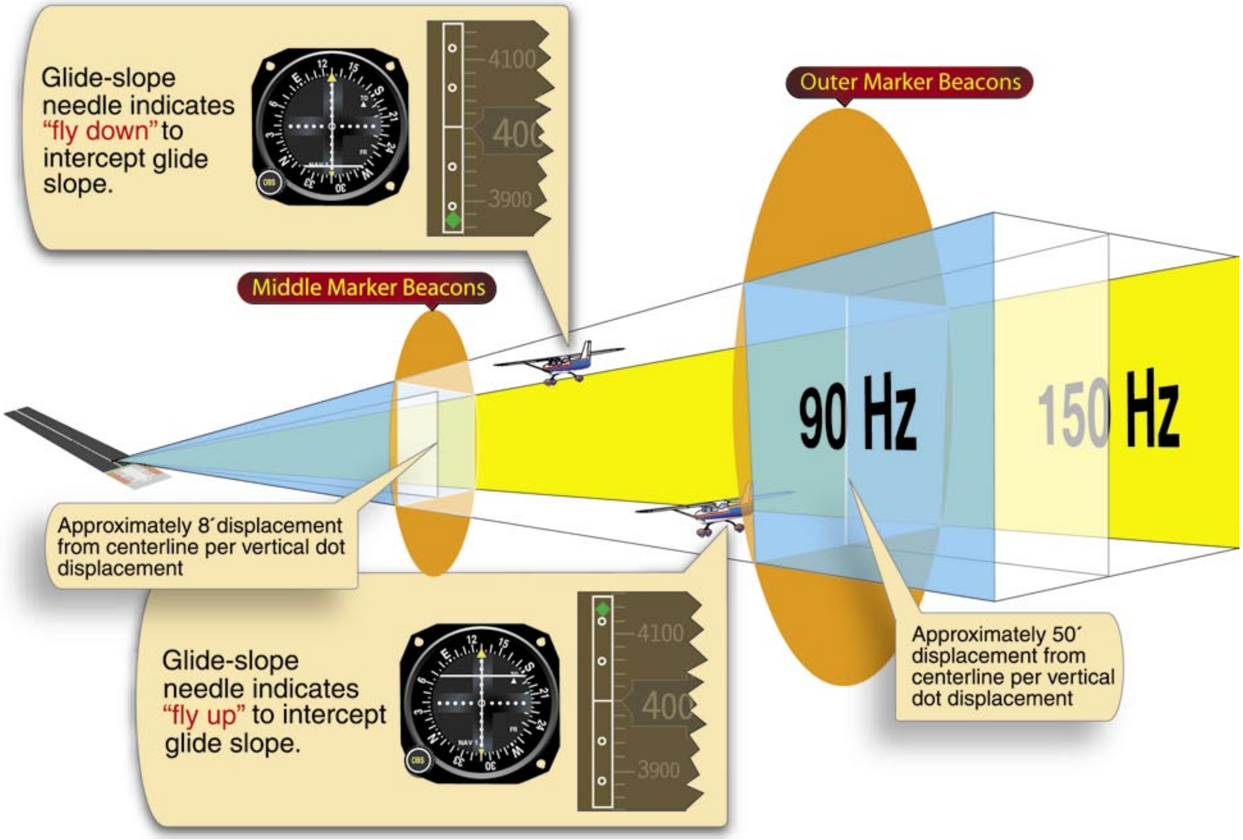
Glide-slope needle indicates "fly down" to intercept glide slope.
4100
400
3900
Middle Marker Beacons
Outer Marker Beacons
90 Hz
150 Hz
Approximately 8´ displacement from centerline per vertical dot displacement
Approximately 50´ displacement from centerline per vertical dot displacement
Glide-slope needle indicates "fly up" to intercept glide slope.
4100
400
3900

Figure 7-40. *Illustrates a GS receiver indication and aircraft displacement. An analog system is on the left and the same indication on the Garmin PFD on the right.*

source of trouble may be the "High/Low/Off" three-position switch, which both activates the receiver and selects receiver sensitivity. Usually, the "test" feature only tests to see if the light bulbs in the marker beacon lights are working. Therefore, in some installations, there is no functional way for the pilot to ascertain the marker beacon receiver is actually on except to fly over a marker beacon transmitter, and see if a signal is received and indicated (e.g., audibly, and visually via marker beacon lights).

Operational Errors

1. Failure to understand the fundamentals of ILS ground equipment, particularly the differences in course dimensions. Since the VOR receiver is used on the localizer course, the assumption is sometimes made that interception and tracking techniques are identical when tracking localizer courses and VOR radials. Remember that the CDI sensing is sharper and faster on the localizer course.
2. Disorientation during transition to the ILS due to poor planning and reliance on one receiver instead of on all available airborne equipment. Use all the assistance available; a single receiver may fail.
3. Disorientation on the localizer course, due to the first error noted above.
4. Incorrect localizer interception angles. A large interception angle usually results in overshooting, and possible disorientation. When intercepting, if possible, turn to the localizer course heading immediately upon the first indication of needle movement. An ADF receiver is an excellent aid to orient you during an ILS approach if there is a locator or NDB on the inbound course.
5. Chasing the CDI and glide path needles, especially when you have not sufficiently studied the approach before the flight.

Simplified Directional Facility (SDF)

The SDF provides a final approach course similar to the ILS localizer. The SDF course may or may not be aligned with the runway and the course may be wider than a standard ILS localizer, resulting in less precision. Usable off-course indications are limited to 35° either side of the course centerline. Instrument indications in the area between 35° and 90° from the course centerline are not controlled and should be disregarded.

The SDF must provide signals sufficient to allow satisfactory operation of a typical aircraft installation within a sector which extends from the center of the SDF antenna system to distances of 18 NM covering a sector 10° either side of centerline up to an angle 7° above the horizontal. The angle of convergence of the final approach course and the extended runway centerline must not exceed 30°. Pilots should note this angle since the approach course originates at the antenna site, and an approach continued beyond the runway threshold would lead the aircraft to the SDF offset position rather than along the runway centerline.

The course width of the SDF signal emitted from the transmitter is fixed at either 6° or 12°, as necessary, to provide maximum flyability and optimum approach course quality. A three-letter identifier is transmitted in code on the SDF frequency; there is no letter "I" (two dots) transmitted before the station identifier, as there is with the LOC. For example, the identifier for Lebanon, Missouri, SDF is LBO.

Localizer Type Directional Aid (LDA)

The LDA is of comparable utility and accuracy to a localizer but is not part of a complete ILS. The LDA course width is between 3° and 6° and thus provides a more precise approach course than an SDF installation. Some LDAs are equipped with a GS. The LDA course is not aligned with the runway, but straight-in minimums may be published where the angle between the runway centerline and the LDA course does not exceed 30°. If this angle exceeds 30°, only circling minimums are published. The identifier is three letters preceded by "I" transmitted in code on the LDA frequency. For example, the identifier for Van Nuys, California, LDA is I-BUR.

Microwave Landing System (MLS)

The MLS provides precision navigation guidance for exact alignment and descent of aircraft on approach to a runway. It provides azimuth, elevation, and distance. Both lateral and vertical guidance may be displayed on conventional course deviation indicators or incorporated into multipurpose flight deck displays. Range information can be displayed by conventional DME indicators and also incorporated into multipurpose displays. *[Figure 7-41]*

The system may be divided into five functions, which are approach azimuth, back azimuth, approach elevation, range; and data communications. The standard configuration of MLS ground equipment includes an azimuth station to perform functions as indicated above. In addition to providing azimuth navigation guidance, the station transmits basic data, which consists of information associated directly with the operation of the landing system, as well as advisory data on the performance of the ground equipment.

Approach Azimuth Guidance

The azimuth station transmits MLS angle and data on one of 200 channels within the frequency range of 5031 to 5091

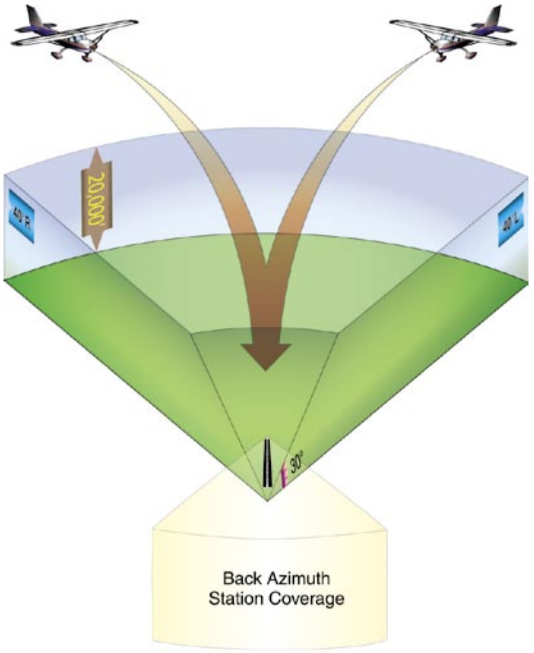

Figure 7-41. *MLS Coverage Volumes, 3-D Representation.*

MHz. The equipment is normally located about 1,000 feet beyond the stop end of the runway, but there is considerable flexibility in selecting sites. For example, for heliport operations the azimuth transmitter can be collocated with the elevation transmitter. The azimuth coverage extends laterally at least 40° on either side of the runway centerline in a standard configuration, in elevation up to an angle of 15° and to at least 20,000 feet, and in range to at least 20 NM.

MLS requires separate airborne equipment to receive and process the signals from what is normally installed in general aviation aircraft today. It has data communications capability, and can provide audible information about the condition of the transmitting system and other pertinent data such as weather, runway status, etc. The MLS transmits an audible identifier consisting of four letters beginning with the letter M, in Morse code at a rate of at least six per minute. The MLS system monitors itself and transmits ground-to-air data messages about the system's operational condition. During periods of routine or emergency maintenance, the coded identification is missing from the transmissions. At this time there are only a few systems installed.

Required Navigation Performance

RNP is a navigation system that provides a specified level of accuracy defined by a lateral area of confined airspace in which an RNP-certified aircraft operates. The continuing growth of aviation places increasing demands on airspace capacity and emphasizes the need for the best use of the available airspace. These factors, along with the accuracy of modern aviation navigation systems and the requirement for increased operational efficiency in terms of direct routings and track-keeping accuracy, have resulted in the concept of required navigation performance—a statement of the navigation performance accuracy necessary for operation within a defined airspace. RNP can include both performance and functional requirements, and is indicated by the RNP type. These standards are intended for designers, manufacturers, and installers of avionics equipment, as well as service providers and users of these systems for global operations. The minimum aviation system performance specification (MASPS) provides guidance for the development of airspace and operational procedures needed to obtain the benefits of improved navigation capability. *[Figure 7-42]*

The RNP type defines the total system error (TSE) that is allowed in lateral and longitudinal dimensions within a particular airspace. The TSE, which takes account of navigation system errors (NSE), computation errors, display errors and flight technical errors (FTE), must not exceed the specified RNP value for 95 percent of the flight time on any part of any single flight. RNP combines the accuracy standards laid out in the ICAO Manual (Doc 9613) with specific accuracy requirements, as well as functional and performance standards, for the RNAV system to realize a system that can meet future air traffic management requirements. The functional criteria for RNP address the need for the flight paths of participating aircraft to be both predictable and repeatable to the declared levels of accuracy. More information on RNP is contained in subsequent chapters.

The term RNP is also applied as a descriptor for airspace, routes, and procedures (including departures, arrivals, and IAPs). The descriptor can apply to a unique approach procedure or to a large region of airspace. RNP applies to navigation performance within a designated airspace, and includes the capability of both the available infrastructure (navigation aids) and the aircraft.

RNP type is used to specify navigation requirements for the airspace. The following are ICAO RNP Types: RNP-1.0, RNP-4.0, RNP-5.0, and RNP-10.0. The required performance is obtained through a combination of aircraft capability and the level of service provided by the corresponding navigation infrastructure. From a broad perspective:

Aircraft Capability + Level of Service = Access

In this context, aircraft capability refers to the airworthiness certification and operational approval elements (including avionics, maintenance, database, human factors, pilot procedures, training, and other issues). The level of service element refers to the NAS infrastructure, including published routes, signal-in-space performance and availability, and air

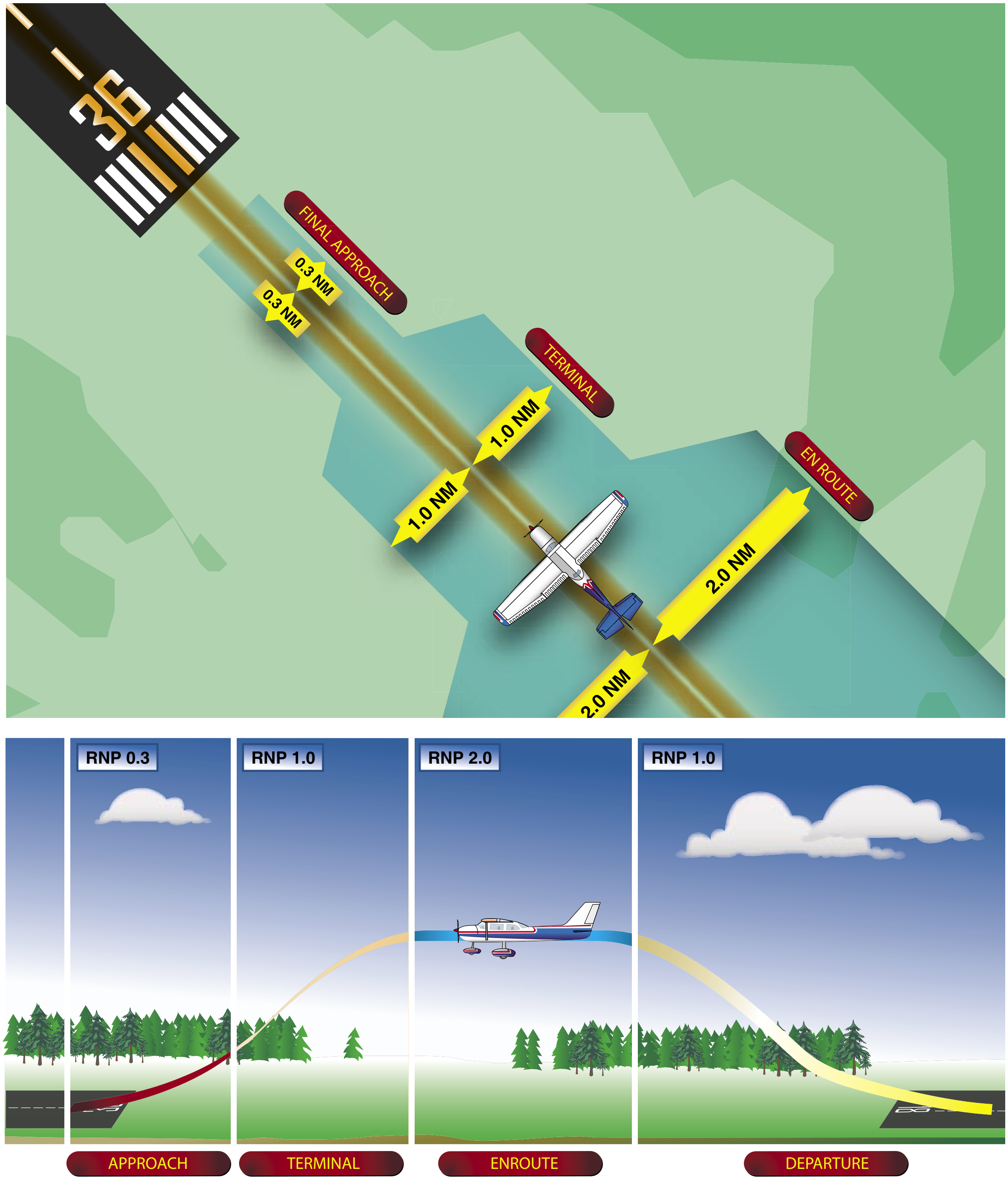
36
FINAL APPROACH
0.3 NM
0.3 NM
TERMINAL
1.0 NM
1.0 NM
EN ROUTE
2.0 NM
2.0 NM
RNP 0.3
RNP 1.0
RNP 2.0
RNP 1.0
APPROACH
TERMINAL
ENROUTE
DEPARTURE

Figure 7-42. *Required Navigation Performance.*

traffic management. When considered collectively, these elements result in providing access. Access provides the desired benefit (airspace, procedures, routes of flight, etc.). RNP levels are actual distances from the centerline of the flight path, which must be maintained for aircraft and obstacle separation. Although additional FAA-recognized RNP levels may be used for specific operations, the United States currently supports three standard RNP levels:

- RNP 0.3 – Approach
- RNP 1.0 – Departure, Terminal
- RNP 2.0 – En route

RNP 0.3 represents a distance of 0.3 NM either side of a specified flight path centerline. The specific performance that is required on the final approach segment of an instrument approach is an example of this RNP level. At the present time, a 0.3 RNP level is the lowest level used in normal RNAV operations. Specific airlines, using special procedures, are approved to use RNP levels lower than RNP 0.3, but those levels are used only in accordance with their approved operations specifications (OpsSpecs). For aircraft equipment to qualify for a specific RNP type, it must maintain navigational accuracy at least 95 percent of the total flight time.

Flight Management Systems (FMS)

A flight management system (FMS) is not a navigation system in itself. Rather, it is a system that automates the tasks of managing the onboard navigation systems. FMS may perform other onboard management tasks, but this discussion is limited to its navigation function.

FMS is an interface between flight crews and flight-deck systems. FMS can be thought of as a computer with a large database of airport and NAVAID locations and associated data, aircraft performance data, airways, intersections, DPs, and STARs. FMS also has the ability to accept and store numerous user-defined WPs, flight routes consisting of departures, WPs, arrivals, approaches, alternates, etc. FMS can quickly define a desired route from the aircraft's current position to any point in the world, perform flight plan computations, and display the total picture of the flight route to the crew.

FMS also has the capability of controlling (selecting) VOR, DME, and LOC NAVAIDs, and then receiving navigational data from them. INS, LORAN, and GPS navigational data may also be accepted by the FMS computer. The FMS may act as the input/output device for the onboard navigation systems, so that it becomes the "go-between" for the crew and the navigation systems.

Function of FMS

At startup, the crew programs the aircraft location, departure runway, DP (if applicable), WPs defining the route, approach procedure, approach to be used, and routing to alternate. This may be entered manually, be in the form of a stored flight plan, or be a flight plan developed in another computer and transferred by disk or electronically to the FMS computer. The crew enters this basic information in the control/display unit (CDU). *[Figure 7-43]*

Once airborne, the FMS computer channels the appropriate NAVAIDs and takes radial/distance information, or channels two NAVAIDs, taking the more accurate distance information. FMS then indicates position, track, desired heading, groundspeed and position relative to desired track. Position information from the FMS updates the INS. In more sophisticated aircraft, the FMS provides inputs to the HSI, RMI, glass flight deck navigation displays, head-up display (HUD), autopilot, and autothrottle systems.

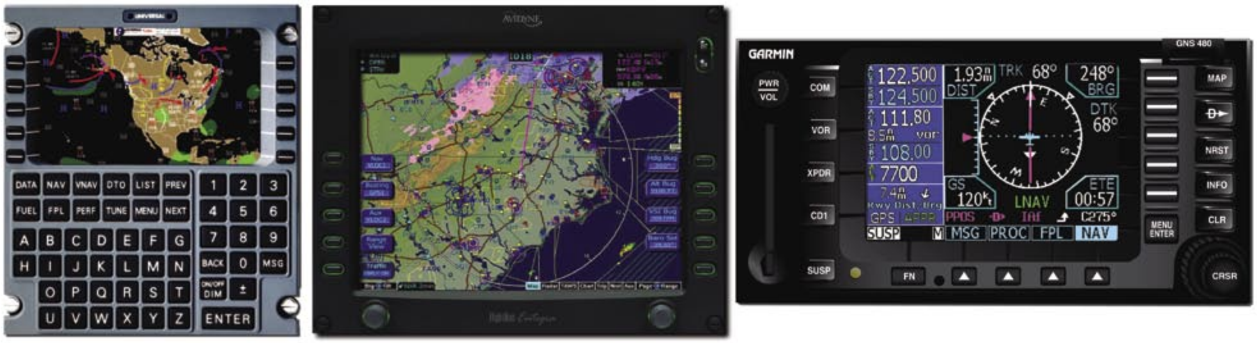

Figure 7-43. *Typical Display and Control Unit(s) in General Aviation. The Universal UNS-1 (left) controls and integrates all other systems. The Avidyne (center) and Garmin systems (right) illustrate and are typical of completely integrated systems. Although the Universal CDU is not typically found on smaller general aviation aircraft, the difference in capabilities of the CDUs and stand-alone sytems is diminishing each year.*

Head-Up Display (HUD)

The HUD is a display system that provides a projection of navigation and air data (airspeed in relation to approach reference speed, altitude, left/right and up/down GS) on a transparent screen between the pilot and the windshield. Other information may be displayed, including a runway target in relation to the nose of the aircraft. This allows the pilot to see the information necessary to make the approach while also being able to see out the windshield, which diminishes the need to shift between looking at the panel to looking outside. Virtually any information desired can be displayed on the HUD if it is available in the aircraft's flight computer, and if the display is user definable. *[Figure 7-44]*

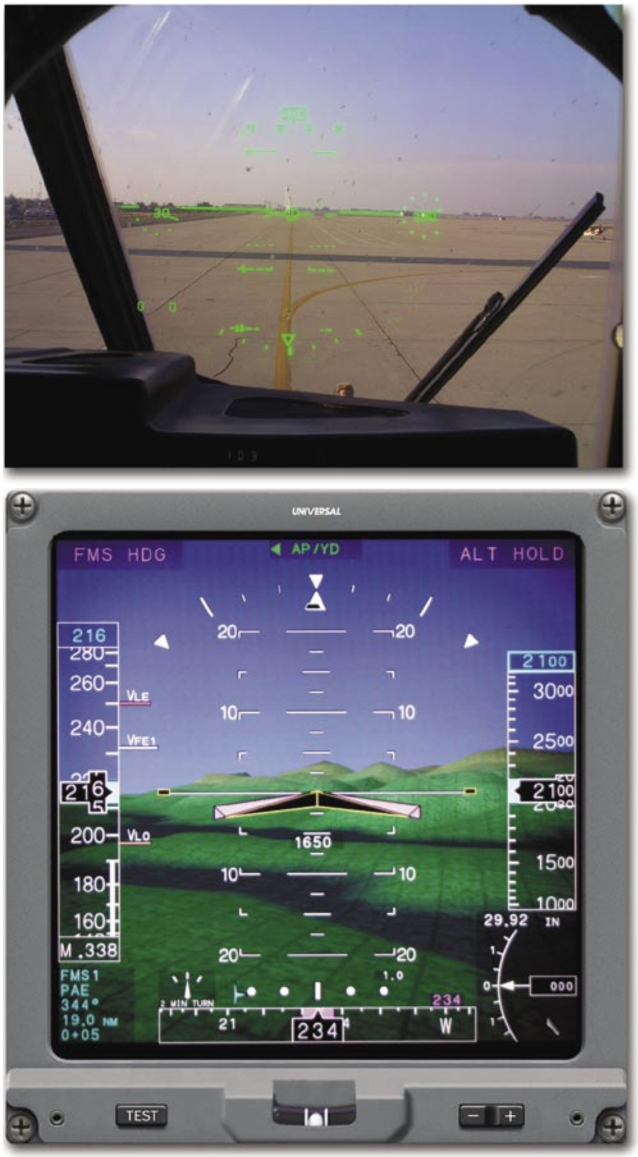

Figure 7-44. *Example of a Head-Up Display (top) and a Head-Down Display (bottom). The head-up display presents information in front of the pilot along his/her normal field of view while a head-down display may present information beyond the normal head-up field of view.*

Radar Navigation (Ground Based)

Radar works by transmitting a pulse of RF energy in a specific direction. The return of the echo or bounce of that pulse from a target is precisely timed. From this, the distance traveled by the pulse and its echo is determined and displayed on a radar screen in such a manner that the distance and bearing to this target can be instantly determined. The radar transmitter must be capable of delivering extremely high power levels toward the airspace under surveillance, and the associated radar receiver must be able to detect extremely small signal levels of the returning echoes.

The radar display system provides the controller with a map-like presentation upon which appear all the radar echoes of aircraft within detection range of the radar facility. By means of electronically generated range marks and azimuth-indicating devices, the controller can locate each radar target with respect to the radar facility, or can locate one radar target with respect to another.

Another device, a video-mapping unit, generates an actual airway or airport map and presents it on the radar display equipment. Using the video-mapping feature, the air traffic controller not only can view the aircraft targets, but can see these targets in relation to runways, navigation aids, and hazardous ground obstructions in the area. Therefore, radar becomes a NAVAID, as well as the most significant means of traffic separation.

In a display presenting perhaps a dozen or more targets, a primary surveillance radar system cannot identify one specific radar target, and it may have difficulty "seeing" a small target at considerable distance—especially if there is a rain shower or thunderstorm between the radar site and the aircraft. This problem is solved with the Air Traffic Control Radar Beacon System (ATCRBS), sometimes called secondary surveillance radar (SSR), which utilizes a transponder in the aircraft. The ground equipment is an interrogating unit, in which the beacon antenna is mounted so it rotates with the surveillance antenna. The interrogating unit transmits a coded pulse sequence that actuates the aircraft transponder. The transponder answers the coded sequence by transmitting a preselected coded sequence back to the ground equipment, providing a strong return signal and positive aircraft identification, as well as other special data such as aircraft altitude.

Functions of Radar Navigation

The radar systems used by ATC are air route surveillance radar (ARSR), airport surveillance radar (ASR), and precision approach radar (PAR) and airport surface detection equipment

(ASDE). Surveillance radars scan through 360° of azimuth and present target information on a radar display located in a tower or center. This information is used independently or in conjunction with other navigational aids in the control of air traffic.

ARSR is a long-range radar system designed primarily to cover large areas and provide a display of aircraft while en route between terminal areas. The ARSR enables air route traffic control center (ARTCC) controllers to provide radar service when the aircraft are within the ARSR coverage. In some instances, ARSR may enable ARTCC to provide terminal radar services similar to but usually more limited than those provided by a radar approach control.

ASR is designed to provide relatively short-range coverage in the general vicinity of an airport and to serve as an expeditious means of handling terminal area traffic through observation of precise aircraft locations on a radarscope. Nonprecision instrument approaches are available at airports that have an approved surveillance radar approach procedure. ASR provides radar vectors to the final approach course and then azimuth information to the pilot during the approach. In addition to range (distance) from the runway, the pilot is advised of MDA, when to begin descent, and when the aircraft is at the MDA. If requested, recommended altitudes will be furnished each mile while on final.

PAR is designed to be used as a landing aid displaying range, azimuth, and elevation information rather than as an aid for sequencing and spacing aircraft. PAR equipment may be used as a primary landing aid, or it may be used to monitor other types of approaches. Two antennas are used in the PAR array, one scanning a vertical plane, and the other scanning horizontally. Since the range is limited to 10 miles, azimuth to 20°, and elevation to 7°, only the final approach area is covered. The controller's scope is divided into two parts. The upper half presents altitude and distance information, and the lower half presents azimuth and distance.

PAR is a system in which a controller provides highly accurate navigational guidance in azimuth and elevation to a pilot. Pilots are given headings to fly to direct them to and keep their aircraft aligned with the extended centerline of the landing runway. They are told to anticipate glide path interception approximately 10–30 seconds before it occurs and when to start descent. The published decision height (DH) is given only if the pilot requests it. If the aircraft is observed to deviate above or below the glide path, the pilot is given the relative amount of deviation by use of terms "slightly" or "well" and is expected to adjust the aircraft's rate of descent/ascent to return to the glide path. Trend information is also issued with respect to the elevation of the aircraft and may be modified by the terms "rapidly" and "slowly" (e.g., "well above glide path, coming down rapidly"). Range from touchdown is given at least once each mile. If an aircraft is observed by the controller to proceed outside of specified safety zone limits in azimuth and/or elevation and continue to operate outside these prescribed limits, the pilot will be directed to execute a missed approach or to fly a specified course unless the pilot has the runway environment (runway, approach lights, etc.) in sight. Navigational guidance in azimuth and elevation is provided to the pilot until the aircraft reaches the published decision altitude (DA)/DH. Advisory course and glide path information is furnished by the controller until the aircraft passes over the landing threshold, at which point the pilot is advised of any deviation from the runway centerline. Radar service is automatically terminated upon completion of the approach.

Airport Surface Detection Equipment

Radar equipment is specifically designed to detect all principal features on the surface of an airport, including aircraft and vehicular traffic, and to present the entire image on a radar indicator console in the control tower. It is used to augment visual observation by tower personnel of aircraft and/or vehicular movements on runways and taxiways.

Radar Limitations

1. It is very important for the aviation community to recognize the fact that there are limitations to radar service and that ATC controllers may not always be able to issue traffic advisories concerning aircraft which are not under ATC control and cannot be seen on radar.
2. The characteristics of radio waves are such that they normally travel in a continuous straight line unless they are "bent" by abnormal atmospheric phenomena such as temperature inversions; reflected or attenuated by dense objects such as heavy clouds, precipitation, ground obstacles, mountains, etc.; or screened by high terrain features.
3. Primary radar energy that strikes dense objects will be reflected and displayed on the operator's scope, thereby blocking out aircraft at the same range and greatly weakening or completely eliminating the display of targets at a greater range.
4. Relatively low altitude aircraft will not be seen if they are screened by mountains or are below the radar beam due to curvature of the Earth.
5. The amount of reflective surface of an aircraft will determine the size of the radar return. Therefore, a small light airplane or a sleek jet fighter will be more difficult to see on primary radar than a large commercial jet or military bomber.

6. All ARTCC radar in the conterminous United States and many ASR have the capability to interrogate Mode C and display altitude information to the controller from appropriately equipped aircraft. However, a number of ASR do not have Mode C display capability; therefore, altitude information must be obtained from the pilot.

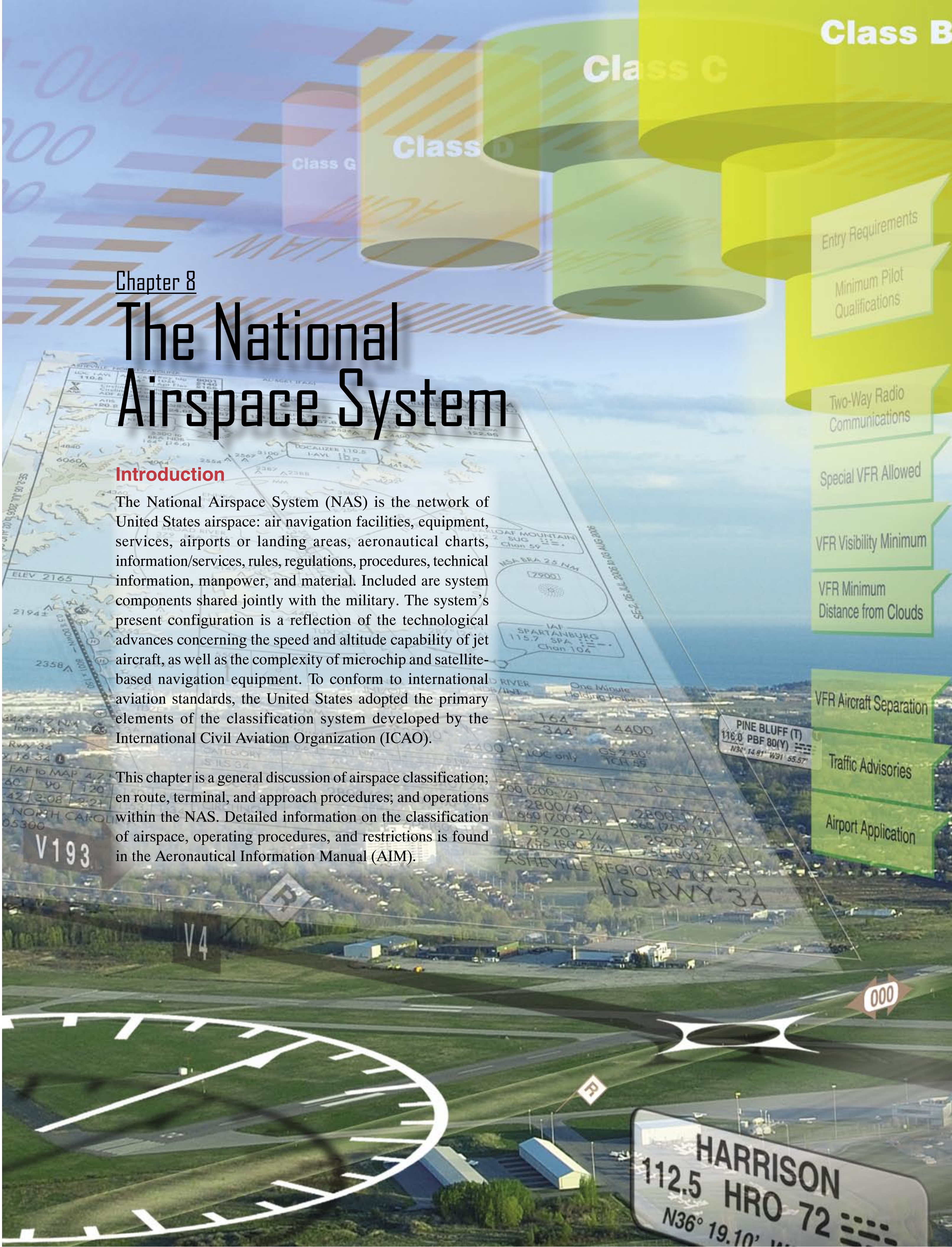
Class B
Class C
Class D
Class G
Entry Requirements
Minimum Pilot Qualifications
Two-Way Radio Communications
Special VFR Allowed
VFR Visibility Minimum
VFR Minimum Distance from Clouds
VFR Aircraft Separation
Traffic Advisories
Airport Application
PINE BLUFF (T)
116.0 PBF 80(Y)
V193
V4
ASHEVILLE REGIONAL (AVL)
ILS RWY 34
000
HARRISON
112.5 HRO 72
N36° 19.10'

Chapter 8

The National Airspace System

Introduction

The National Airspace System (NAS) is the network of United States airspace: air navigation facilities, equipment, services, airports or landing areas, aeronautical charts, information/services, rules, regulations, procedures, technical information, manpower, and material. Included are system components shared jointly with the military. The system's present configuration is a reflection of the technological advances concerning the speed and altitude capability of jet aircraft, as well as the complexity of microchip and satellite-based navigation equipment. To conform to international aviation standards, the United States adopted the primary elements of the classification system developed by the International Civil Aviation Organization (ICAO).

This chapter is a general discussion of airspace classification; en route, terminal, and approach procedures; and operations within the NAS. Detailed information on the classification of airspace, operating procedures, and restrictions is found in the Aeronautical Information Manual (AIM).

Airspace Classification

Airspace in the United States *[Figure 8-1]* is designated as follows:

1. Class A. Generally, airspace from 18,000 feet mean sea level (MSL) up to and including flight level (FL) 600, including the airspace overlying the waters within 12 nautical miles (NM) of the coast of the 48 contiguous states and Alaska. Unless otherwise authorized, all pilots must operate their aircraft under instrument flight rules (IFR).
2. Class B. Generally, airspace from the surface to 10,000 feet MSL surrounding the nation's busiest airports in terms of airport operations or passenger enplanements. The configuration of each Class B airspace area is individually tailored, consists of a surface area and two or more layers (some Class B airspace areas resemble upside-down wedding cakes), and is designed to contain all published instrument procedures once an aircraft enters the airspace. An air traffic control (ATC) clearance is required for all aircraft to operate in the area, and all aircraft that are so cleared receive separation services within the airspace.
3. Class C. Generally, airspace from the surface to 4,000 feet above the airport elevation (charted in MSL) surrounding those airports that have an operational control tower, are serviced by a radar approach control, and have a certain number of IFR operations or passenger enplanements. Although the configuration of each Class C area is individually tailored, the airspace usually consists of a surface area with a 5 NM radius, an outer circle with a 10 NM radius that extends from 1,200 feet to 4,000 feet above the airport elevation and an outer area. Each aircraft must establish two-way radio communications with the ATC facility providing air traffic services prior to entering the airspace and thereafter maintain those communications while within the airspace.
4. Class D. Generally, that airspace from the surface to 2,500 feet above the airport elevation (charted in MSL) surrounding those airports that have an operational control tower. The configuration of each Class D airspace area is individually tailored and when instrument procedures are published, the airspace will normally be designed to contain the procedures. Arrival extensions for instrument approach procedures (IAPs) may be Class D or Class E airspace. Unless otherwise authorized, each aircraft must establish two-way radio communications with the ATC facility providing air traffic services prior to entering the airspace and thereafter maintain those communications while in the airspace.
5. Class E. Generally, if the airspace is not Class A, B, C, or D, and is controlled airspace, then it is Class E airspace. Class E airspace extends upward from either the surface or a designated altitude to the overlying or adjacent controlled airspace. When designated as a surface area, the airspace will be configured to contain all instrument procedures. Also in this class are federal airways, airspace beginning at either 700 or 1,200 feet above ground level (AGL) used to transition to and from the terminal or en route environment, and en route domestic and offshore airspace areas designated below 18,000 feet MSL. Unless designated at a lower altitude, Class E airspace begins at 14,500 MSL over the United States, including that airspace overlying the waters within 12 NM of the coast of the 48 contiguous states and Alaska, up to but not including 18,000 feet MSL, and the airspace above FL 600.
6. Class G. Airspace not designated as Class A, B, C, D, or E. Class G airspace is essentially uncontrolled by ATC except when associated with a temporary control tower.

Special Use Airspace

Special use airspace is the designation for airspace in which certain activities must be confined, or where limitations may be imposed on aircraft operations that are not part of those activities. Certain special use airspace areas can create limitations on the mixed use of airspace. The special use airspace depicted on instrument charts includes the area name or number, effective altitude, time and weather conditions of operation, the controlling agency, and the chart panel location. On National Aeronautical Charting Group (NACG) en route charts, this information is available on one of the end panels.

Prohibited areas contain airspace of defined dimensions within which the flight of aircraft is prohibited. Such areas are established for security or other reasons associated with the national welfare. These areas are published in the Federal Register and are depicted on aeronautical charts. The area is charted as a "P" followed by a number (e.g., "P-123").

Restricted areas are areas where operations are hazardous to nonparticipating aircraft and contain airspace within which the flight of aircraft, while not wholly prohibited, is subject to restrictions. Activities within these areas must be confined because of their nature, or limitations may be imposed upon aircraft operations that are not a part of those activities, or both. Restricted areas denote the existence of unusual, often invisible, hazards to aircraft (e.g., artillery firing, aerial

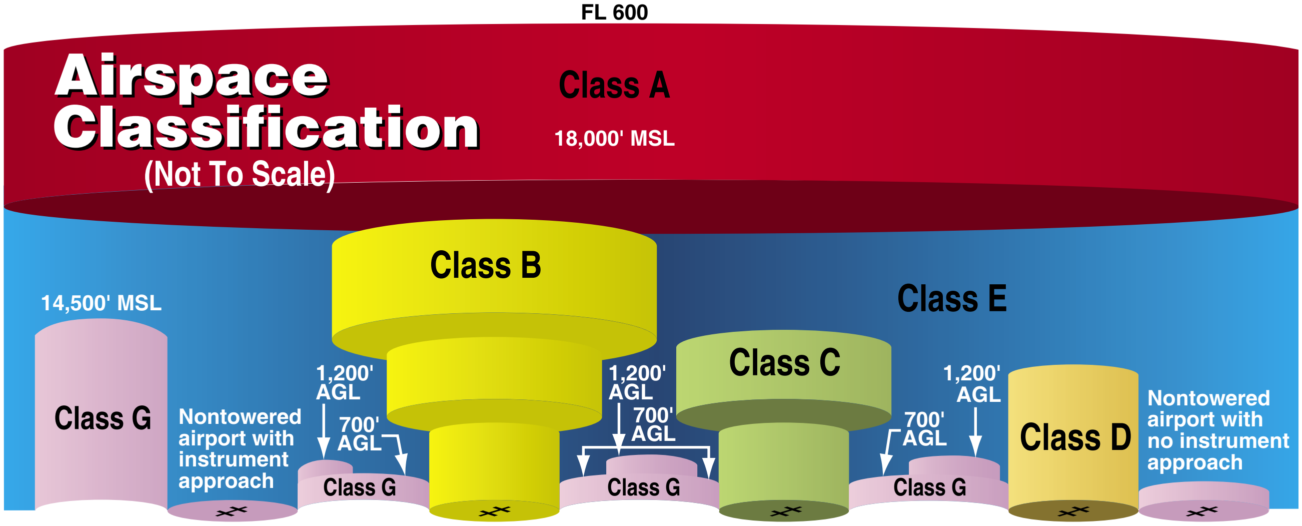
FL 600
Airspace Classification
(Not To Scale)
Class A
18,000' MSL
14,500' MSL
Class B
Class E
Class C
Class G
Nontowered airport with instrument approach
1,200' AGL
700' AGL
Class G
1,200' AGL
700' AGL
Class G
700' AGL
1,200' AGL
Class G
Class D
Nontowered airport with no instrument approach

	Class A	Class B	Class C	Class D	Class E	Class G
Entry Requirements	ATC clearance	ATC clearance	Prior two-way communications	Prior two-way communications	Prior two-way communications*	Prior two-way communications*
Minimum Pilot Qualifications	Instrument Rating	Private or Student certification—local restrictions apply.	Student certificate	Student certificate	Student certificate	Student certificate
Two-Way Radio Communications	Yes	Yes	Yes	Yes	Yes, under IFR flight plan*	Yes*
Special VFR Allowed	No	Yes	Yes	Yes	Yes	N/A
VFR Visibility Minimum	N/A	3 statute miles	3 statute miles	3 statute miles	3 statute miles**	1 statute mile†
VFR Minimum Distance from Clouds	N/A	Clear of clouds	500' below, 1,000' above, 2,000' horizontal	500' below, 1,000' above, 2,000' horizontal	500' below,** 1,000' above, 2,000' horizontal	Clear of clouds†
VFR Aircraft Separation	N/A	All	IFR aircraft	Runway operations	None	None
Traffic Advisories	Yes	Yes	Yes	Workload permitting	Workload permitting	Workload permitting
Airport Application	N/A	• Radar • Instrument approaches • Weather • Control tower • High density	• Radar • Instrument approaches • Weather • Control tower	• Instrument approaches • Weather • Control tower	• Instrument approaches • Weather	• Control tower

*Exception: temporary tower or control tower present
**True only below 10,000 feet
† True only during day at or below 1,200 feet AGL (see 14 CFR part 91)

AGL—above ground level
FL—flight level
MSL—mean sea level

Figure 8-1. *Airspace Classifications.*

gunnery, or guided missiles). IFR flights may be authorized to transit the airspace and are routed accordingly. Penetration of restricted areas without authorization from the using or controlling agency may be extremely hazardous to the aircraft and its occupants. ATC facilities apply the following procedures when aircraft are operating on an IFR clearance (including those cleared by ATC to maintain visual flight rules (VFR)-On-Top) via a route that lies within joint-use restricted airspace:

1. If the restricted area is not active and has been released to the Federal Aviation Administration (FAA), the ATC facility will allow the aircraft to operate in the restricted airspace without issuing specific clearance for it to do so.
2. If the restricted area is active and has not been released to the FAA, the ATC facility will issue a clearance which will ensure the aircraft avoids the restricted airspace.

Restricted areas are charted with an “R” followed by a number (e.g., “R-5701”) and are depicted on the en route chart appropriate for use at the altitude or FL being flown.

Warning areas are similar in nature to restricted areas; however, the United States government does not have sole jurisdiction over the airspace. A warning area is airspace of defined dimensions, extending from 12 NM outward from the coast of the United States, containing activity that may be hazardous to nonparticipating aircraft. The purpose of such areas is to warn nonparticipating pilots of the potential danger. A warning area may be located over domestic or international waters or both. The airspace is designated with a “W” followed by a number (e.g., “W-123”).

Military operations areas (MOAs) consist of airspace with defined vertical and lateral limits established for the purpose of separating certain military training activities from IFR traffic. Whenever an MOA is being used, nonparticipating IFR traffic may be cleared through an MOA if IFR separation can be provided by ATC. Otherwise, ATC will reroute or restrict nonparticipating IFR traffic. MOAs are depicted on sectional, VFR terminal area, and en route low altitude charts and are not numbered (e.g., “Boardman MOA”).

Alert areas are depicted on aeronautical charts with an “A” followed by a number (e.g., “A-123”) to inform nonparticipating pilots of areas that may contain a high volume of pilot training or an unusual type of aerial activity. Pilots should exercise caution in alert areas. All activity within an alert area shall be conducted in accordance with regulations, without waiver, and pilots of participating aircraft, as well as pilots transiting the area, shall be equally responsible for collision avoidance.

Military Training Routes (MTRs) are routes used by military aircraft to maintain proficiency in tactical flying. These routes are usually established below 10,000 feet MSL for operations at speeds in excess of 250 knots. Some route segments may be defined at higher altitudes for purposes of route continuity. Routes are identified as IFR (IR), and VFR (VR), followed by a number. MTRs with no segment above 1,500 feet AGL are identified by four number characters (e.g., IR1206, VR1207, etc.). MTRs that include one or more segments above 1,500 feet AGL are identified by three number characters (e.g., IR206, VR207). IFR Low Altitude En Route Charts depict all IR routes and all VR routes that accommodate operations above 1,500 feet AGL. IR routes are conducted in accordance with IFR regardless of weather conditions.

Temporary flight restrictions (TFRs) are put into effect when traffic in the airspace would endanger or hamper air or ground activities in the designated area. For example, a forest fire, chemical accident, flood, or disaster-relief effort could warrant a TFR, which would be issued as a Notice to Airmen (NOTAM).

National Security Areas (NSAs) consist of airspace with defined vertical and lateral dimensions established at locations where there is a requirement for increased security and safety of ground facilities. Flight in NSAs may be temporarily prohibited by regulation under the provisions of Title 14 of the Code of Federal Regulations (14 CFR) part 99 and prohibitions will be disseminated via NOTAM.

Federal Airways

The primary means for routing aircraft operating under IFR is the federal airways system.

Each federal airway is based on a centerline that extends from one NAVAID/waypoint/fix/intersection to another NAVAID/waypoint/fix/intersection specified for that airway. A federal airway includes the airspace within parallel boundary lines four NM to each side of the centerline. As in all instrument flight, courses are magnetic, and distances are in NM. The airspace of a federal airway has a floor of 1,200 feet AGL, unless otherwise specified. A federal airway does not include the airspace of a prohibited area.

Victor airways include the airspace extending from 1,200 feet AGL up to, but not including 18,000 feet MSL. The airways are designated on Sectional and IFR low altitude en route charts with the letter “V” followed by a number (e.g., “V23”). Typically, Victor airways are given odd numbers when oriented north/south and even numbers when oriented east/west. If more than one airway coincides on a route segment, the numbers are listed serially (e.g., “V287-495-500”). *[Figure 8-2]*

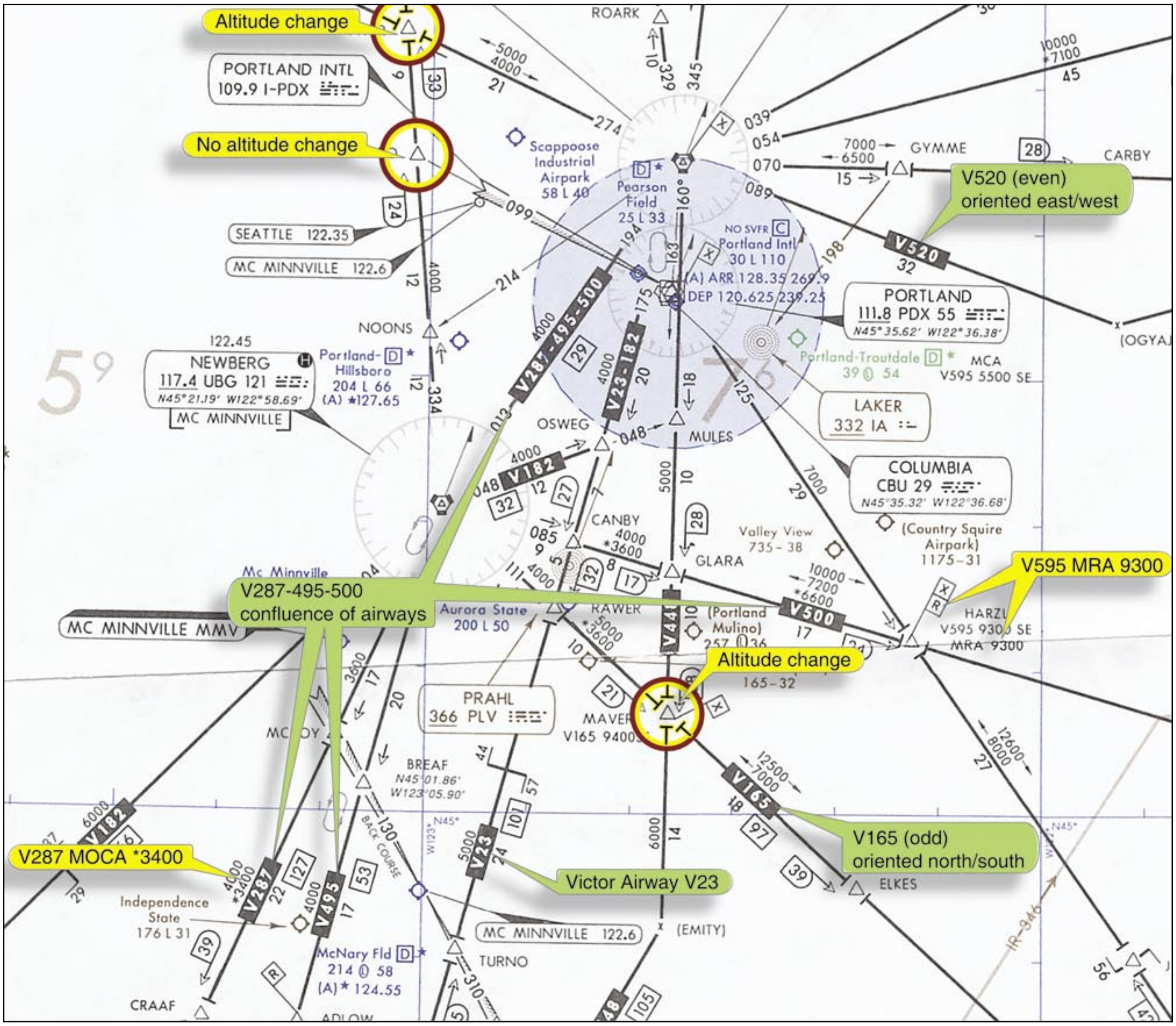
Altitude change
PORTLAND INTL
109.9 I-PDX
No altitude change
SEATTLE 122.35
MC MINNVILLE 122.6
NOONS
122.45
NEWBERG
117.4 UBG 121
N45°21.19' W122°58.69'
MC MINNVILLE
Portland-Hillsboro
204 L 66
(A) ★127.65
ROARK
Scappoose Industrial Airpark
58 L 40
Pearson Field
25 L 33
NO SVFR
Portland Intl
30 L 110
(A) ARR 128.35 269.9
DEP 120.625 239.25
GYMME
CARBY
V520 (even)
oriented east/west
V520
PORTLAND
111.8 PDX 55
N45°35.62' W122°36.38'
(OGYA)
Portland-Troutdale
39 54
MCA
V595 5500 SE
LAKER
332 IA
COLUMBIA
CBU 29
N45°35.32' W122°36.68'
(Country Squire Airpark)
1175-31
Valley View
735-38
OSWEG
MULES
V287-495-500
V23-182
V182
CANBY
GLARA
V595 MRA 9300
HARZL
V595 9300 SE
MRA 9300
V500
(Portland Mulino)
Altitude change
V287-495-500
confluence of airways
MC MINNVILLE MMV
Mc Minnville
Aurora State
200 L 50
RAWER
V441
PRAHL
366 PLV
MAVER
V165 9400
BREAF
N45°01.86'
W123°05.90'
V165
V165 (odd)
oriented north/south
ELKES
V287 MOCA *3400
V182
Independence State
176 L 31
V287
V495
V23
Victor Airway V23
MC MINNVILLE 122.6
(EMITY)
McNary Fld
214 58
(A)★ 124.55
TURNO
CRAAF
ADLOW
IR-346

Figure 8-2. *Victor Airways and Charted IFR Altitudes.*

Jet routes exist only in Class A airspace, from 18,000 feet MSL to FL 450, and are depicted on high-altitude en route charts. The letter "J" precedes a number to label the airway (e.g., J12).

RNAV routes have been established in both the low-altitude and the high-altitude structures in recent years and are depicted on the en route low and high chart series. High altitude RNAV routes are identified with a "Q" prefix (except the Q-routes in the Gulf of Mexico) and low altitude RNAV routes are identified with a "T" prefix. RNAV routes and data are depicted in aeronautical blue.

In addition to the published routes, a random RNAV route may be flown under IFR if it is approved by ATC. Random RNAV routes are direct routes, based on area navigation capability, between waypoints defined in terms of latitude/longitude coordinates, degree-distance fixes, or offsets from established routes/airways at a specified distance and direction.

Radar monitoring by ATC is required on all random RNAV routes. These routes can only be approved in a radar environment. Factors that will be considered by ATC in approving random RNAV routes include the capability to provide radar monitoring, and compatibility with traffic volume and flow. ATC will radar monitor each flight; however, navigation on the random RNAV route is the responsibility of the pilot.

Other Routing

Preferred IFR routes have been established between major terminals to guide pilots in planning their routes of flight, minimizing route changes and aiding in the orderly management of air traffic on federal airways. Low and high altitude preferred routes are listed in the Airport/Facility Directory (A/FD). To use a preferred route, reference the departure and arrival airports; if a routing exists for your flight, then airway instructions will be listed.

Tower En Route Control (TEC) is an ATC program that uses overlapping approach control radar services to provide IFR clearances. By using TEC, a pilot is routed by airport control towers. Some advantages include abbreviated filing procedures and reduced traffic separation requirements. TEC is dependent upon the ATC's workload, and the procedure varies among locales.

The latest version of Advisory Circular (AC) 90-91, North American Route Program (NRP), provides guidance to users of the NAS for participation in the NRP. All flights operating at or above FL 290 within the conterminous United States and Canada are eligible to participate in the NRP, the primary purpose of which is to allow operators to plan minimum time/cost routes that may be off the prescribed route structure. NRP aircraft are not subject to route-limiting restrictions (e.g., published preferred IFR routes) beyond a 200 NM radius of their point of departure or destination.

IFR En Route Charts

The objective of IFR en route flight is to navigate within the lateral limits of a designated airway at an altitude consistent with the ATC clearance. Your ability to fly instruments safely and competently in the system is greatly enhanced by understanding the vast array of data available to the pilot on instrument charts. The NACG maintains and produces the charts for the United States government.

En route high-altitude charts provide aeronautical information for en route instrument navigation (IFR) at or above 18,000 feet MSL. Information includes the portrayal of Jet and RNAV routes, identification and frequencies of radio aids, selected airports, distances, time zones, special use airspace, and related information. Established Jet routes from 18,000 feet MSL to FL 450 use NAVAIDs not more than 260 NM apart. The charts are revised every 56 days.

To effectively depart from one airport and navigate en route under instrument conditions a pilot needs the appropriate IFR en route low-altitude chart(s). The IFR low altitude en route chart is the instrument equivalent of the Sectional chart. When folded, the cover of the NACG en route chart displays an index map of the United States showing the coverage areas. Cities near congested airspace are shown in black type and their associated area chart is listed in the box in the lower left-hand corner of the map coverage box. Also noted is an explanation of the off-route obstruction clearance altitude (OROCA). The effective date of the chart is printed on the other side of the folded chart. Information concerning MTRs is also included on the chart cover. The en route charts are revised every 56 days.

When the NACG en route chart is unfolded, the legend is displayed and provides information concerning airports, NAVAIDs, communications, air traffic services, and airspace.

Airport Information

Airport information is provided in the legend, and the symbols used for the airport name, elevation, and runway length are similar to the sectional chart presentation. Associated city names are shown for public airports only. FAA identifiers are shown for all airports. ICAO identifiers are also shown for airports outside of the contiguous United States. Instrument approaches can be found at airports with blue or green symbols, while the brown airport symbol denotes airports that do not have instrument approaches. Stars are used to indicate the part-time nature of tower operations, ATIS frequencies, part-time or on request lighting facilities, and part-time airspace classifications. A box after an airport name with a "C" or "D" inside indicates Class C and D airspace, respectively, per *Figure 8-3*. D

Charted IFR Altitudes

The minimum en route altitude (MEA) ensures a navigation signal strong enough for adequate reception by the aircraft navigation (NAV) receiver and obstacle clearance along the airway. Communication is not necessarily guaranteed with MEA compliance. The obstacle clearance, within the limits of the airway, is typically 1,000 feet in non-mountainous areas and 2,000 feet in designated mountainous areas. MEAs can be authorized with breaks in the signal coverage; if this is the case, the NACG en route chart notes "MEA GAP" parallel to the affected airway. MEAs are usually bidirectional; however, they can be single-directional. Arrows are used to indicate the direction to which the MEA applies.

The minimum obstruction clearance altitude (MOCA), as the name suggests, provides the same obstruction clearance as an MEA; however, the NAV signal reception is ensured only within 22 NM of the closest NAVAID defining the route. The MOCA is listed below the MEA and indicated on NACG charts by a leading asterisk (e.g., "*3400"—see *Figure 8-2*, V287 at bottom left).

The minimum reception altitude (MRA) identifies the lowest altitude at which an intersection can be determined from an off-course NAVAID. If the reception is line-of-sight based, signal coverage will only extend to the MRA or above. However, if the aircraft is equipped with distance measuring equipment (DME) and the chart indicates the intersection can

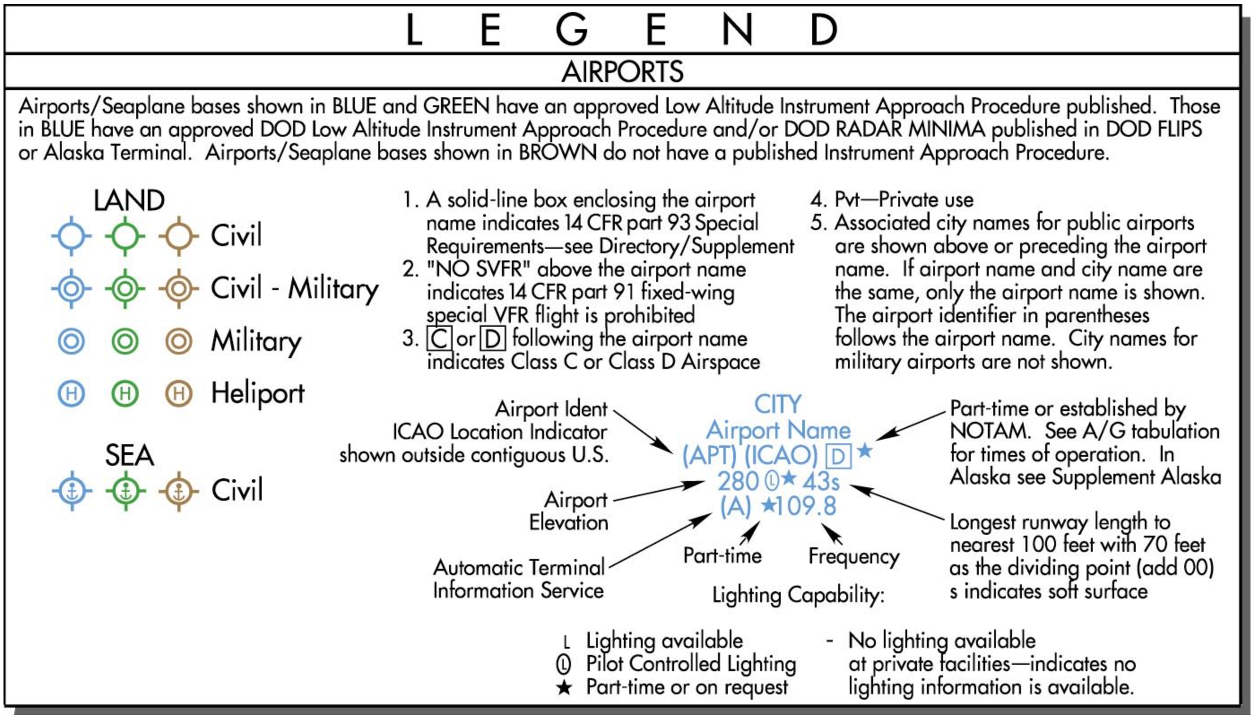

LEGEND

AIRPORTS

Airports/Seaplane bases shown in BLUE and GREEN have an approved Low Altitude Instrument Approach Procedure published. Those in BLUE have an approved DOD Low Altitude Instrument Approach Procedure and/or DOD RADAR MINIMA published in DOD FLIPS or Alaska Terminal. Airports/Seaplane bases shown in BROWN do not have a published Instrument Approach Procedure.

1. A solid-line box enclosing the airport name indicates 14 CFR part 93 Special Requirements—see Directory/Supplement
2. "NO SVFR" above the airport name indicates 14 CFR part 91 fixed-wing special VFR flight is prohibited
3. C or D following the airport name indicates Class C or Class D Airspace
4. Pvt—Private use
5. Associated city names for public airports are shown above or preceding the airport name. If airport name and city name are the same, only the airport name is shown. The airport identifier in parentheses follows the airport name. City names for military airports are not shown.

LAND
Civil
Civil - Military
Military
Heliport
SEA
Civil
CITY
Airport Name
(APT) (ICAO) D ★
280 Ⓛ★ 43s
(A) ★109.8
Airport Ident
ICAO Location Indicator shown outside contiguous U.S.
Airport Elevation
Automatic Terminal Information Service
Part-time
Frequency
Lighting Capability:
Part-time or established by NOTAM. See A/G tabulation for times of operation. In Alaska see Supplement Alaska
Longest runway length to nearest 100 feet with 70 feet as the dividing point (add 00) s indicates soft surface
L Lighting available
Ⓛ Pilot Controlled Lighting
★ Part-time or on request
- No lighting available at private facilities—indicates no lighting information is available.

Figure 8-3. *En Route Airport Legend.*

be identified with such equipment, the pilot could define the fix without attaining the MRA. On NACG charts, the MRA is indicated by the symbol and the altitude preceded by "MRA" (e.g., "MRA 9300"). *[Figure 8-2]*

The minimum crossing altitude (MCA) will be charted when a higher MEA route segment is approached. The MCA is usually indicated when a pilot is approaching steeply rising terrain, and obstacle clearance and/or signal reception is compromised. In this case, the pilot is required to initiate a climb so the MCA is reached by the time the intersection is crossed. On NACG charts, the MCA is indicated by the symbol , and the Victor airway number, altitude, and the direction to which it applies (e.g. "V24 8000 SE").

The maximum authorized altitude (MAA) is the highest altitude at which the airway can be flown with assurance of receiving adequate navigation signals. Chart depictions appear as "MAA-15000."

When an MEA, MOCA, and/or MAA change on a segment other than at a NAVAID, a sideways "T" ⊣ is depicted on the chart. If there is an airway break without the symbol, one can assume the altitudes have not changed (see the upper left area of *Figure 8-2*). When a change of MEA to a higher MEA is required, the climb may commence at the break, ensuring obstacle clearance. *[Figure 8-4]*

Navigation Features

Types of NAVAIDs

Very high frequency omnidirectional ranges (VORs) are the principal NAVAIDs that support the Victor and Jet airways. Many other navigation tools are also available to the pilot. For example, nondirectional beacons (NDBs) can broadcast signals accurate enough to provide stand-alone approaches, and DME allows the pilot to pinpoint a reporting point on the airway. Though primarily navigation tools, these NAVAIDs can also transmit voice broadcasts.

Tactical air navigation (TACAN) channels are represented as the two- or three-digit numbers following the three-letter identifier in the NAVAID boxes. The NACG terminal procedures provide a frequency-pairing table for the TACAN-only sites. On NACG charts, very-high frequencies and ultra-high frequencies (VHF/UHF) NAVAIDs (e.g., VORs) are depicted in black, while low frequencies and medium frequencies (LF/MF) are depicted as brown. *[Figure 8-5]*

Identifying Intersections

Intersections along the airway route are established by a variety of NAVAIDs. An open triangle △ indicates the location of an ATC reporting point at an intersection. If the triangle is solid, ▲ a report is compulsory. *[Figure 8-4]*

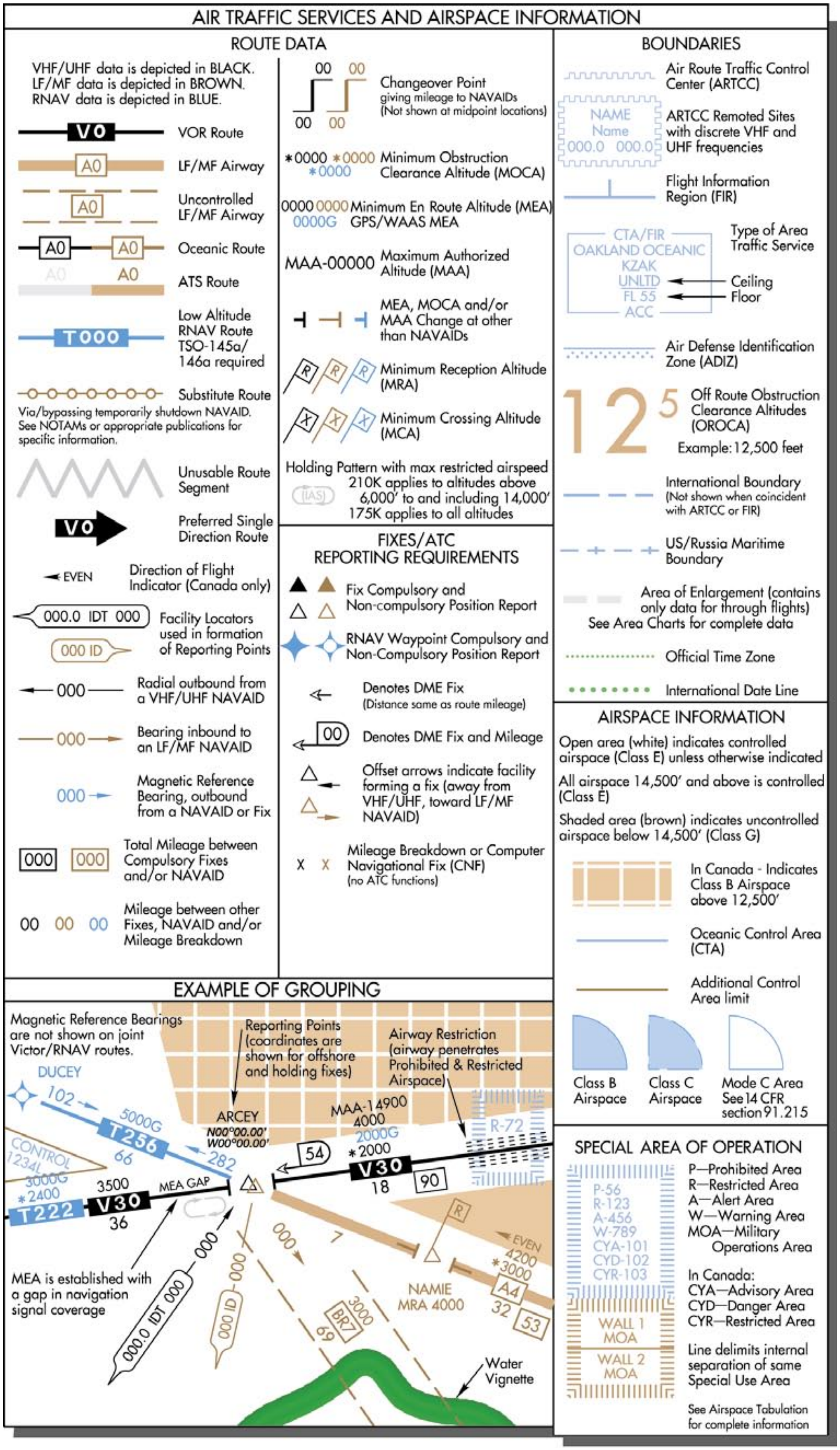
AIR TRAFFIC SERVICES AND AIRSPACE INFORMATION
ROUTE DATA
VHF/UHF data is depicted in BLACK.
LF/MF data is depicted in BROWN.
RNAV data is depicted in BLUE.
V0
VOR Route
A0
LF/MF Airway
A0
Uncontrolled LF/MF Airway
A0
Oceanic Route
A0
ATS Route
T000
Low Altitude RNAV Route TSO-145a/146a required
Substitute Route
Via/bypassing temporarily shutdown NAVAID. See NOTAMs or appropriate publications for specific information.
Unusable Route Segment
V0
Preferred Single Direction Route
EVEN
Direction of Flight Indicator (Canada only)
000.0 IDT 000
000 ID
Facility Locators used in formation of Reporting Points
000
Radial outbound from a VHF/UHF NAVAID
000
Bearing inbound to an LF/MF NAVAID
000
Magnetic Reference Bearing, outbound from a NAVAID or Fix
000 000
Total Mileage between Compulsory Fixes and/or NAVAID
00 00 00
Mileage between other Fixes, NAVAID and/or Mileage Breakdown
00 00
00 00
Changeover Point giving mileage to NAVAIDs (Not shown at midpoint locations)
*0000 *0000
*0000
Minimum Obstruction Clearance Altitude (MOCA)
0000 0000
0000G
Minimum En Route Altitude (MEA)
GPS/WAAS MEA
MAA-00000
Maximum Authorized Altitude (MAA)
MEA, MOCA and/or MAA Change at other than NAVAIDs
R
Minimum Reception Altitude (MRA)
X
Minimum Crossing Altitude (MCA)
Holding Pattern with max restricted airspeed
IAS
210K applies to altitudes above 6,000′ to and including 14,000′
175K applies to all altitudes
FIXES/ATC REPORTING REQUIREMENTS
Fix Compulsory and Non-compulsory Position Report
RNAV Waypoint Compulsory and Non-Compulsory Position Report
Denotes DME Fix (Distance same as route mileage)
00
Denotes DME Fix and Mileage
Offset arrows indicate facility forming a fix (away from VHF/UHF, toward LF/MF NAVAID)
X X
Mileage Breakdown or Computer Navigational Fix (CNF) (no ATC functions)
BOUNDARIES
Air Route Traffic Control Center (ARTCC)
NAME
Name
000.0 000.0
ARTCC Remoted Sites with discrete VHF and UHF frequencies
Flight Information Region (FIR)
CTA/FIR
OAKLAND OCEANIC
KZAK
UNLTD
FL 55
ACC
Type of Area Traffic Service
Ceiling
Floor
Air Defense Identification Zone (ADIZ)
12 5
Off Route Obstruction Clearance Altitudes (OROCA)
Example: 12,500 feet
International Boundary (Not shown when coincident with ARTCC or FIR)
US/Russia Maritime Boundary
Area of Enlargement (contains only data for through flights)
See Area Charts for complete data
Official Time Zone
International Date Line
AIRSPACE INFORMATION
Open area (white) indicates controlled airspace (Class E) unless otherwise indicated
All airspace 14,500′ and above is controlled (Class E)
Shaded area (brown) indicates uncontrolled airspace below 14,500′ (Class G)
In Canada - Indicates Class B Airspace above 12,500′
Oceanic Control Area (CTA)
Additional Control Area limit
Class B Airspace
Class C Airspace
Mode C Area See 14 CFR section 91.215
SPECIAL AREA OF OPERATION
P-56
R-123
A-456
W-789
CYA-101
CYD-102
CYR-103
P—Prohibited Area
R—Restricted Area
A—Alert Area
W—Warning Area
MOA—Military Operations Area
In Canada:
CYA—Advisory Area
CYD—Danger Area
CYR—Restricted Area
WALL 1 MOA
WALL 2 MOA
Line delimits internal separation of same Special Use Area
See Airspace Tabulation for complete information
EXAMPLE OF GROUPING
Magnetic Reference Bearings are not shown on joint Victor/RNAV routes.
DUCEY
102
5000G
T256
66
282
CONTROL 1234L
3000G
*2400
T222
3500
V30
36
MEA GAP
Reporting Points (coordinates are shown for offshore and holding fixes)
ARCEY
N00°00.00′
W00°00.00′
54
Airway Restriction (airway penetrates Prohibited & Restricted Airspace)
MAA-14900
4000
2000G
*2000
V30
18
90
R-72
R
EVEN
4200
*3000
A4
32
53
7
NAMIE
MRA 4000
000
000
000 ID
000.0 IDT 000
3000
BR7
69
MEA is established with a gap in navigation signal coverage
Water Vignette

Figure 8-4. *Legend From En Route Low Attitude Chart, Air Traffic Services and Airspace Information Section.*

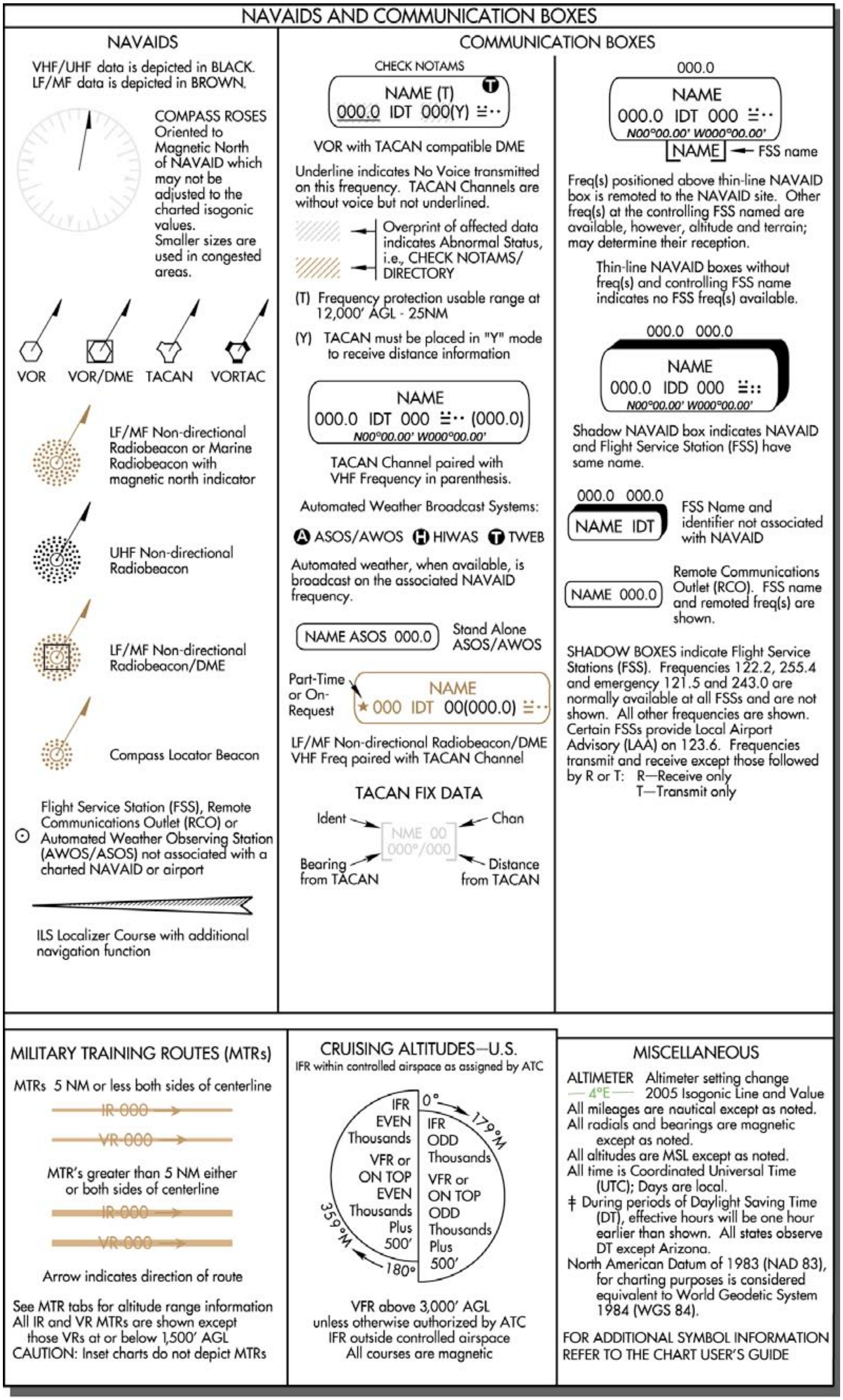

NAVAIDS AND COMMUNICATION BOXES

NAVAIDS

VHF/UHF data is depicted in BLACK.
LF/MF data is depicted in BROWN.

COMPASS ROSES Oriented to Magnetic North of NAVAID which may not be adjusted to the charted isogonic values. Smaller sizes are used in congested areas.

VOR
VOR/DME
TACAN
VORTAC

LF/MF Non-directional Radiobeacon or Marine Radiobeacon with magnetic north indicator

UHF Non-directional Radiobeacon

LF/MF Non-directional Radiobeacon/DME

Compass Locator Beacon

Flight Service Station (FSS), Remote Communications Outlet (RCO) or Automated Weather Observing Station (AWOS/ASOS) not associated with a charted NAVAID or airport

ILS Localizer Course with additional navigation function

COMMUNICATION BOXES

CHECK NOTAMS
NAME (T)
000.0 IDT 000(Y)

VOR with TACAN compatible DME

Underline indicates No Voice transmitted on this frequency. TACAN Channels are without voice but not underlined.

Overprint of affected data indicates Abnormal Status, i.e., CHECK NOTAMS/ DIRECTORY

(T) Frequency protection usable range at 12,000′ AGL - 25NM

(Y) TACAN must be placed in "Y" mode to receive distance information

NAME
000.0 IDT 000 (000.0)
N00°00.00′ W000°00.00′

TACAN Channel paired with VHF Frequency in parenthesis.

Automated Weather Broadcast Systems:

Ⓐ ASOS/AWOS Ⓗ HIWAS Ⓣ TWEB

Automated weather, when available, is broadcast on the associated NAVAID frequency.

NAME ASOS 000.0

Stand Alone ASOS/AWOS

Part-Time or On-Request
NAME
★000 IDT 00(000.0)

LF/MF Non-directional Radiobeacon/DME VHF Freq paired with TACAN Channel

TACAN FIX DATA

Ident
Chan
NME 00
000°/000
Bearing from TACAN
Distance from TACAN

000.0
NAME
000.0 IDT 000
N00°00.00′ W000°00.00′
NAME
FSS name

Freq(s) positioned above thin-line NAVAID box is remoted to the NAVAID site. Other freq(s) at the controlling FSS named are available, however, altitude and terrain; may determine their reception.

Thin-line NAVAID boxes without freq(s) and controlling FSS name indicates no FSS freq(s) available.

000.0 000.0
NAME
000.0 IDD 000
N00°00.00′ W000°00.00′

Shadow NAVAID box indicates NAVAID and Flight Service Station (FSS) have same name.

000.0 000.0
NAME IDT

FSS Name and identifier not associated with NAVAID

Remote Communications Outlet (RCO). FSS name and remoted freq(s) are shown.

SHADOW BOXES indicate Flight Service Stations (FSS). Frequencies 122.2, 255.4 and emergency 121.5 and 243.0 are normally available at all FSSs and are not shown. All other frequencies are shown. Certain FSSs provide Local Airport Advisory (LAA) on 123.6. Frequencies transmit and receive except those followed by R or T: R—Receive only
T—Transmit only

MILITARY TRAINING ROUTES (MTRs)

MTRs 5 NM or less both sides of centerline

IR-000
VR-000

MTR's greater than 5 NM either or both sides of centerline

IR-000
VR-000

Arrow indicates direction of route

See MTR tabs for altitude range information
All IR and VR MTRs are shown except those VRs at or below 1,500′ AGL
CAUTION: Inset charts do not depict MTRs

CRUISING ALTITUDES—U.S.

IFR within controlled airspace as assigned by ATC

0°
179°M
180°
359°M
IFR EVEN Thousands
VFR or ON TOP EVEN Thousands Plus 500′
IFR ODD Thousands
VFR or ON TOP ODD Thousands Plus 500′

VFR above 3,000′ AGL unless otherwise authorized by ATC
IFR outside controlled airspace
All courses are magnetic

MISCELLANEOUS

ALTIMETER Altimeter setting change
— 4°E — 2005 Isogonic Line and Value
All mileages are nautical except as noted.
All radials and bearings are magnetic except as noted.
All altitudes are MSL except as noted.
All time is Coordinated Universal Time (UTC); Days are local.
‡ During periods of Daylight Saving Time (DT), effective hours will be one hour earlier than shown. All states observe DT except Arizona.
North American Datum of 1983 (NAD 83), for charting purposes is considered equivalent to World Geodetic System 1984 (WGS 84).

FOR ADDITIONAL SYMBOL INFORMATION REFER TO THE CHART USER'S GUIDE

Figure 8-5. *Legend From En Route Low Attitude Chart.*

NDBs, localizers, and off-route VORs are used to establish intersections. NDBs are sometimes collocated with intersections, in which case passage of the NDB would mark the intersection. A bearing to an off-route NDB also can provide intersection identification. A localizer course used to identify an intersection is depicted by a feathered arrowhead symbol on the en route chart. 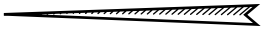 If feathered markings appear on the left-hand side of the arrowhead, 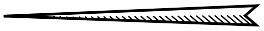 a back course (BC) signal is transmitted. On NACG en route charts, the localizer symbol is only depicted to identify an intersection.

Off-route VORs remain the most common means of identifying intersections when traveling on an airway. Arrows depicted next to the intersection indicate the NAVAID to be used for identification. Another means of identifying an intersection is with the use of DME. A hollow arrowhead indicates DME is authorized for intersection identification. If the DME mileage at the intersection is a cumulative distance of route segments, the mileage is totaled and indicated by a D-shaped symbol with a mileage number inside. 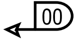

[Figure 8-4] Approved IFR GPS units can also be used to report intersections.

Other Route Information

DME and GPS provide valuable route information concerning such factors as mileage, position, and groundspeed. Even without this equipment, information is provided on the charts for making the necessary calculations using time and distance. The en route chart depicts point-to-point distances on the airway system. Distances from VOR to VOR are charted with a number inside of a box. 000 To differentiate distances when two airways coincide, the word "TO" with the three-letter VOR identifier appear to the left of the distance boxes. TO PDX 97

VOR changeover points (COPs) are depicted on the charts by this symbol: 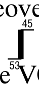

The numbers indicate the distance at which to change the VOR frequency. The frequency change might be required due to signal reception or conflicting frequencies. If a COP does not appear on an airway, the frequency should be changed midway between the facilities. A COP at an intersection may indicate a course change.

Occasionally an "x" will appear at a separated segment of an airway that is not an intersection. The "x" is a mileage breakdown or computer navigation fix and may indicate a course change.

Today's computerized system of ATC has greatly reduced the need for holding en route. However, published holding patterns are still found on charts at junctures where ATC has deemed it necessary to enable traffic flow. When a holding pattern is charted, the controller may provide the holding direction and the statement "as published." *[Figure 8-4]*

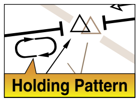

Boundaries separating the jurisdiction of Air Route Traffic Control Centers (ARTCC) are depicted on charts with blue serrations. The name of the controlling facility is printed on the corresponding side of the division line. ARTCC remote sites are depicted as blue serrated boxes and contain the center name, sector name, and the sector frequency. *[Figure 8-4]*

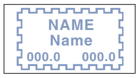

Weather Information and Communication Features

En route NAVAIDs also provide weather information and serve communication functions. When a NAVAID is shown as a shadowed box, an automated flight service station (AFSS) of the same name is directly associated with the facility. If an AFSS is located without an associated NAVAID, the shadowed box is smaller and contains only the name and identifier. The AFSS frequencies are provided above the box. (Frequencies 122.2 and 255.4, and emergency frequencies 121.5 and 243.0 are not listed.)

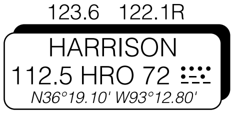

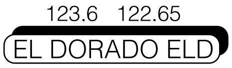

A Remote Communications Outlet (RCO) associated with a NAVAID is designated by a thin-lined box with the controlling AFSS frequency above the box, and the name under the box. Without an associated facility, the thin-lined RCO box contains the AFSS name and remote frequency.

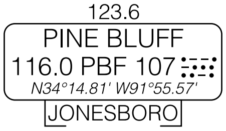

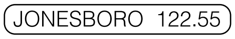

Automated Surface Observing Station (ASOS), Automated Weather Observing Station (AWOS), Hazardous Inflight Weather Advisory Service (HIWAS) and Transcribed Weather Broadcast (TWEB) are continuously transmitted over selected NAVAIDs and depicted in the NAVAID box. ASOS/AWOS are depicted by a white "A", HIWAS by a "H" and TWEB broadcasts by a "T" in a solid black circle in the upper right or left corner.

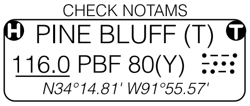

New Technologies

Technological advances have made multifunction displays and moving maps more common in newer aircraft. Even older aircraft are being retrofitted to include "glass" in the flight deck. *[Figure 8-6]* Moving maps improve pilot situational awareness by providing a picture of aircraft location in

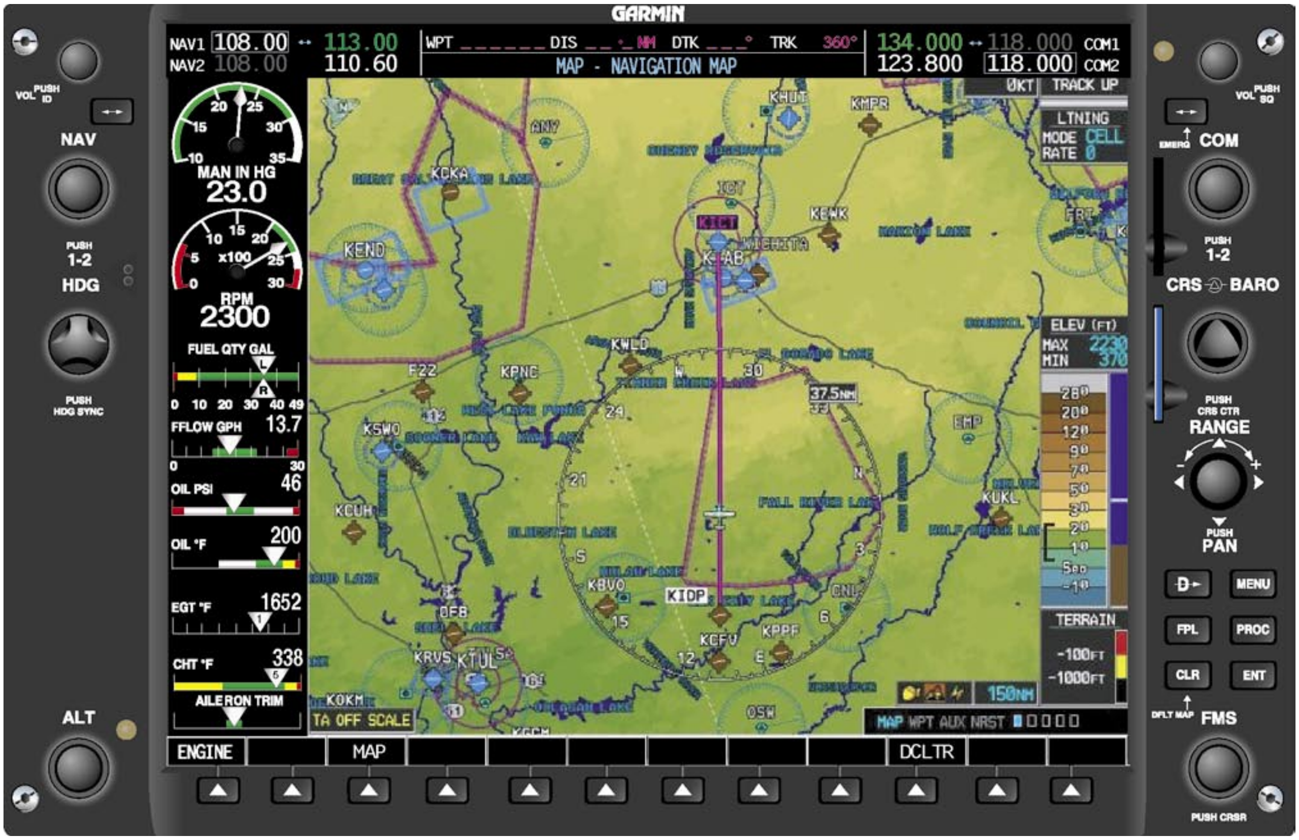
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT ______ DIS __._ NM DTK ___° TRK 360°
MAP - NAVIGATION MAP
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
0KT TRACK UP
LTNING MODE CELL RATE 0
ELEV (FT) MAX 2230 MIN 370
TERRAIN -100FT -1000FT
MAN IN HG 23.0
RPM 2300
FUEL QTY GAL
FFLOW GPH 13.7
OIL PSI 46
OIL °F 200
EGT °F 1652
CHT °F 338
AILERON TRIM
TA OFF SCALE
150NM
MAP WPT AUX NRST
ENGINE
MAP
DCLTR
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
VOL PUSH SQ
EMERG
COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR

Figure 8-6. *Moving Map Display.*

relation to NAVAIDS, waypoints, airspace, terrain, and hazardous weather. GPS systems can be certified for terminal area and en route use as well as approach guidance.

Additional breakthroughs in display technology are the new electronic chart systems or electronic flight bags that facilitate the use of electronic documents in the general aviation flight deck. *[Figure 8-7]* An electronic chart or flight bag is a self-powered electronic library that stores and displays en route charts and other essential documents on a screen. These electronic devices can store the digitized United States terminal procedures, en route charts, the complete airport facility directory, in addition to Title 14 of the Code of Federal Regulations (14 CFR) and the AIM. Full touch-screen based computers allow pilots to view airport approach and area charts electronically while flying. It replaces paper charts as well as other paper materials including minimum equipment lists (MELs), standard operating procedures (SOPs), standard instrument departures (SIDs), standard terminal arrival routes (STARs), checklists, and flight deck manuals. As with paper flight publications, the electronic database needs to be current to provide accurate information regarding NAVAIDS, waypoints, and terminal procedures. Databases are updated every 28 days and are available from various commercial vendors. Pilots should be familiar with equipment operation, capabilities, and limitations prior to use.

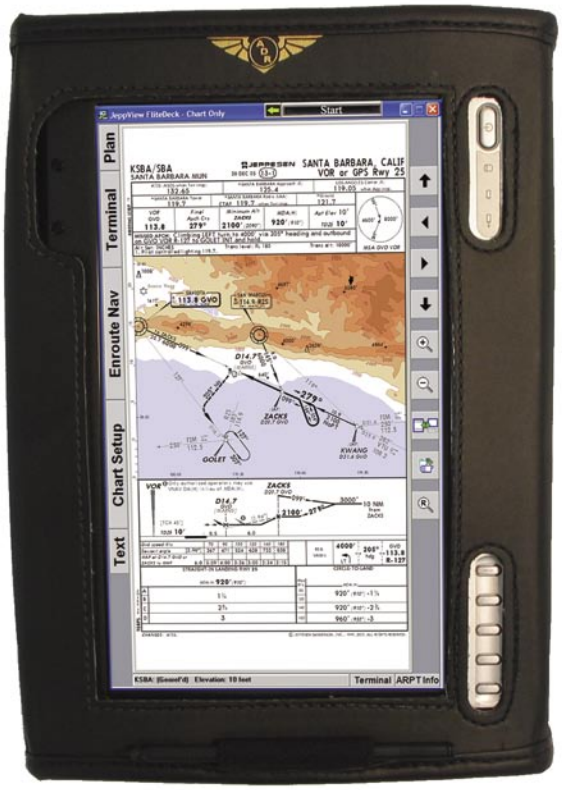
JeppView FliteDeck - Chart Only
Start
Plan
Terminal
Enroute Nav
Chart Setup
Text
KSBA/SBA
SANTA BARBARA, CALIF
VOR or GPS Rwy 25
Terminal ARPT Info

Figure 8-7. *Example of an Electronic Flight Bag.*

Terminal Procedures Publications

While the en route charts provide the information necessary to safely transit broad regions of airspace, the United States Terminal Procedures Publication (TPP) enables pilots to guide their aircraft in the airport area. Whether departing or arriving, these procedures exist to make the controllers' and pilots' jobs safer and more efficient. Available in booklets by region (published by NACG), the TPP includes approach procedures, STARs, Departure Procedures (DPs), and airport diagrams.

Departure Procedures (DPs)

There are two types of DPs, Obstacle Departure Procedures (ODP) and SIDs. *[Figure 8-8]* Both types of DPs provide obstacle clearance protection to aircraft in instrument meteorological conditions (IMC), while reducing communications and departure delays. DPs are published in text and/or charted graphic form. Regardless of the format, all DPs provide a way to depart the airport and transition to the en route structure safely. When possible, pilots are strongly encouraged to file and fly a DP at night, during marginal visual meteorological conditions (VMC) and IMC.

All DPs provide obstacle clearance provided the aircraft crosses the end of the runway at least 35 feet AGL; climbs to 400 feet above airport elevation before turning; and climbs at least 200 feet per nautical mile (FPNM), unless a higher climb gradient is specified to the assigned altitude. ATC may vector an aircraft off a previously assigned DP; however, the 200 FPNM or the FPNM specified in the DP is required.

Textual ODPs are listed by city and airport in the IFR Take-Off Minimums and DPs Section of the TPP. SIDs are depicted in the TPP following the approach procedures for the airport.

Standard Terminal Arrival Routes (STARs)

STARs depict prescribed routes to transition the instrument pilot from the en route structure to a fix in the terminal area from which an instrument approach can be conducted. If a pilot does not have the appropriate STAR, write "No STAR" in the flight plan. However, if the controller is busy, the pilot might be cleared along the same route and, if necessary, the controller will have the pilot copy the entire text of the procedure.

STARs are listed alphabetically at the beginning of the NACG booklet. *Figure 8-9* shows an example of a STAR, and the legend for STARs and DPs printed in NACG booklets.

Instrument Approach Procedure (IAP) Charts

The IAP chart provides the method to descend and land safely in low visibility conditions. The FAA establishes an IAP after thorough analyses of obstructions, terrain features, and navigational facilities. Maneuvers, including altitude changes, course corrections, and other limitations, are prescribed in the IAP. The approach charts reflect the criteria associated with the United States Standard for Terminal Instrument Approach Procedures (TERPs), which prescribes standardized methods for use in designing instrument flight procedures.

In addition to the NACG, other governmental and corporate entities produce approach procedures. The United States military IAPs are established and published by the Department of Defense and are available to the public upon request. Special IAPs are approved by the FAA for individual operators and are not available to the general public. Foreign country standard IAPs are established and published according to the individual country's publication procedures. The information presented in the following sections will highlight features of the United States Terminal Procedures Publications.

The instrument approach chart is divided into six main sections, which include the margin identification, pilot briefing (and notes), plan view, profile view, landing minimums, and airport diagram. *[Figure 8-10]* An examination of each section follows.

Margin Identification

The margin identification, at the top and bottom of the chart, depicts the airport location and procedure identification. The civil approach plates are organized by city, then airport name and state. For example, Orlando Executive in Orlando, Florida is alphabetically listed under "O" for Orlando. Military approaches are organized by airport name first.

The chart's amendment status appears below the city and state in the bottom margin. The amendment number is followed by the five-digit julian-date of the last chart change."05300" is read, "the 300th day of 2005". At the center of the top margin is the FAA chart reference number and the approving authority. At the bottom center, the airport's latitude and longitude coordinates are provided.

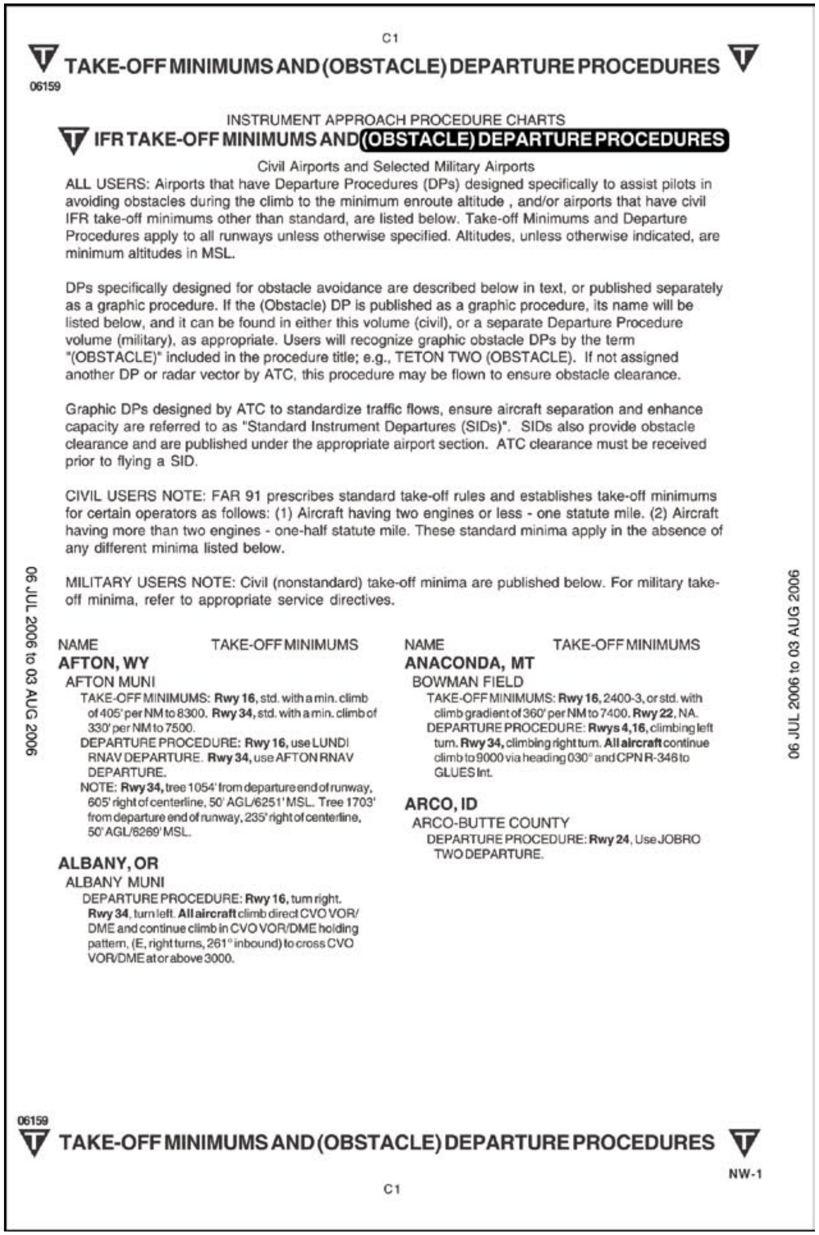

C1

TAKE-OFF MINIMUMS AND (OBSTACLE) DEPARTURE PROCEDURES

06159

INSTRUMENT APPROACH PROCEDURE CHARTS

IFR TAKE-OFF MINIMUMS AND (OBSTACLE) DEPARTURE PROCEDURES

Civil Airports and Selected Military Airports

ALL USERS: Airports that have Departure Procedures (DPs) designed specifically to assist pilots in avoiding obstacles during the climb to the minimum enroute altitude , and/or airports that have civil IFR take-off minimums other than standard, are listed below. Take-off Minimums and Departure Procedures apply to all runways unless otherwise specified. Altitudes, unless otherwise indicated, are minimum altitudes in MSL.

DPs specifically designed for obstacle avoidance are described below in text, or published separately as a graphic procedure. If the (Obstacle) DP is published as a graphic procedure, its name will be listed below, and it can be found in either this volume (civil), or a separate Departure Procedure volume (military), as appropriate. Users will recognize graphic obstacle DPs by the term "(OBSTACLE)" included in the procedure title; e.g., TETON TWO (OBSTACLE). If not assigned another DP or radar vector by ATC, this procedure may be flown to ensure obstacle clearance.

Graphic DPs designed by ATC to standardize traffic flows, ensure aircraft separation and enhance capacity are referred to as "Standard Instrument Departures (SIDs)". SIDs also provide obstacle clearance and are published under the appropriate airport section. ATC clearance must be received prior to flying a SID.

CIVIL USERS NOTE: FAR 91 prescribes standard take-off rules and establishes take-off minimums for certain operators as follows: (1) Aircraft having two engines or less - one statute mile. (2) Aircraft having more than two engines - one-half statute mile. These standard minima apply in the absence of any different minima listed below.

MILITARY USERS NOTE: Civil (nonstandard) take-off minima are published below. For military take-off minima, refer to appropriate service directives.

NAME TAKE-OFF MINIMUMS

AFTON, WY

AFTON MUNI

TAKE-OFF MINIMUMS: **Rwy 16**, std. with a min. climb of 405' per NM to 8300. **Rwy 34**, std. with a min. climb of 330' per NM to 7500.

DEPARTURE PROCEDURE: **Rwy 16**, use LUNDI RNAV DEPARTURE. **Rwy 34**, use AFTON RNAV DEPARTURE.

NOTE: **Rwy 34**, tree 1054' from departure end of runway, 605' right of centerline, 50' AGL/6251' MSL. Tree 1703' from departure end of runway, 235' right of centerline, 50' AGL/6269' MSL.

ALBANY, OR

ALBANY MUNI

DEPARTURE PROCEDURE: **Rwy 16**, turn right. **Rwy 34**, turn left. **All aircraft** climb direct CVO VOR/DME and continue climb in CVO VOR/DME holding pattern, (E, right turns, 261° inbound) to cross CVO VOR/DME at or above 3000.

NAME TAKE-OFF MINIMUMS

ANACONDA, MT

BOWMAN FIELD

TAKE-OFF MINIMUMS: **Rwy 16**, 2400-3, or std. with climb gradient of 360' per NM to 7400. **Rwy 22**, NA.

DEPARTURE PROCEDURE: **Rwys 4,16**, climbing left turn. **Rwy 34**, climbing right turn. **All aircraft** continue climb to 9000 via heading 030° and CPN R-346 to GLUES Int.

ARCO, ID

ARCO-BUTTE COUNTY

DEPARTURE PROCEDURE: **Rwy 24**, Use JOBRO TWO DEPARTURE.

06 JUL 2006 to 03 AUG 2006

06 JUL 2006 to 03 AUG 2006

06159

TAKE-OFF MINIMUMS AND (OBSTACLE) DEPARTURE PROCEDURES

NW-1

C1

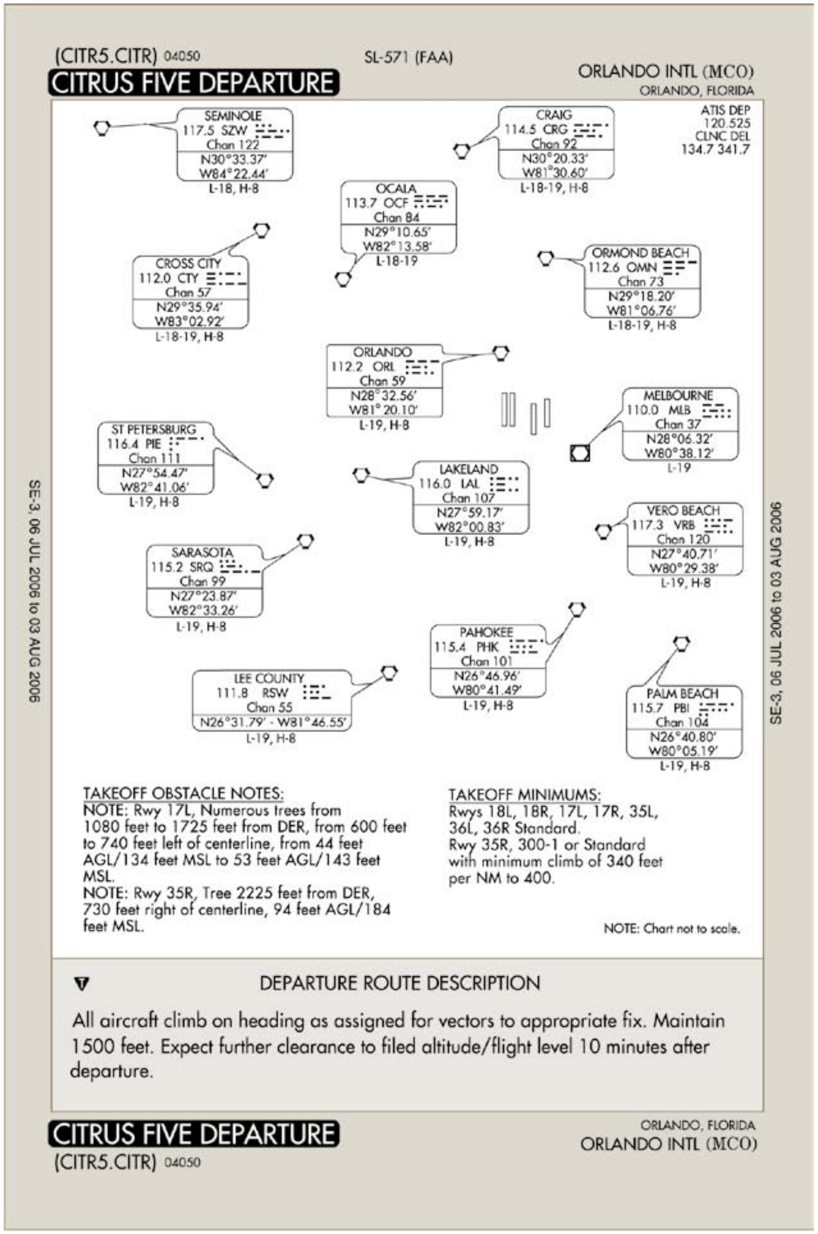

(CITR5.CITR) 04050

CITRUS FIVE DEPARTURE

SL-571 (FAA)

ORLANDO INTL (MCO)
ORLANDO, FLORIDA

ATIS DEP
120.525
CLNC DEL
134.7 341.7
SEMINOLE
117.5 SZW
Chan 122
N30°33.37'
W84°22.44'
L-18, H-8
OCALA
113.7 OCF
Chan 84
N29°10.65'
W82°13.58'
L-18-19
CRAIG
114.5 CRG
Chan 92
N30°20.33'
W81°30.60'
L-18-19, H-8
CROSS CITY
112.0 CTY
Chan 57
N29°35.94'
W83°02.92'
L-18-19, H-8
ORMOND BEACH
112.6 OMN
Chan 73
N29°18.20'
W81°06.76'
L-18-19, H-8
ORLANDO
112.2 ORL
Chan 59
N28°32.56'
W81°20.10'
L-19, H-8
MELBOURNE
110.0 MLB
Chan 37
N28°06.32'
W80°38.12'
L-19
ST PETERSBURG
116.4 PIE
Chan 111
N27°54.47'
W82°41.06'
L-19, H-8
LAKELAND
116.0 LAL
Chan 107
N27°59.17'
W82°00.83'
L-19, H-8
VERO BEACH
117.3 VRB
Chan 120
N27°40.71'
W80°29.38'
L-19, H-8
SARASOTA
115.2 SRQ
Chan 99
N27°23.87'
W82°33.26'
L-19, H-8
PAHOKEE
115.4 PHK
Chan 101
N26°46.96'
W80°41.49'
L-19, H-8
LEE COUNTY
111.8 RSW
Chan 55
N26°31.79' - W81°46.55'
L-19, H-8
PALM BEACH
115.7 PBI
Chan 104
N26°40.80'
W80°05.19'
L-19, H-8

TAKEOFF OBSTACLE NOTES:

NOTE: Rwy 17L, Numerous trees from 1080 feet to 1725 feet from DER, from 600 feet to 740 feet left of centerline, from 44 feet AGL/134 feet MSL to 53 feet AGL/143 feet MSL.

NOTE: Rwy 35R, Tree 2225 feet from DER, 730 feet right of centerline, 94 feet AGL/184 feet MSL.

TAKEOFF MINIMUMS:

Rwys 18L, 18R, 17L, 17R, 35L, 36L, 36R Standard.

Rwy 35R, 300-1 or Standard with minimum climb of 340 feet per NM to 400.

NOTE: Chart not to scale.

DEPARTURE ROUTE DESCRIPTION

All aircraft climb on heading as assigned for vectors to appropriate fix. Maintain 1500 feet. Expect further clearance to filed altitude/flight level 10 minutes after departure.

CITRUS FIVE DEPARTURE

(CITR5.CITR) 04050

ORLANDO, FLORIDA
ORLANDO INTL (MCO)

SE-3, 06 JUL 2006 to 03 AUG 2006

SE-3, 06 JUL 2006 to 03 AUG 2006

Figure 8-8. *Obstacle Departure Procedures (ODP) and Standard Instrument Departures (SID).*

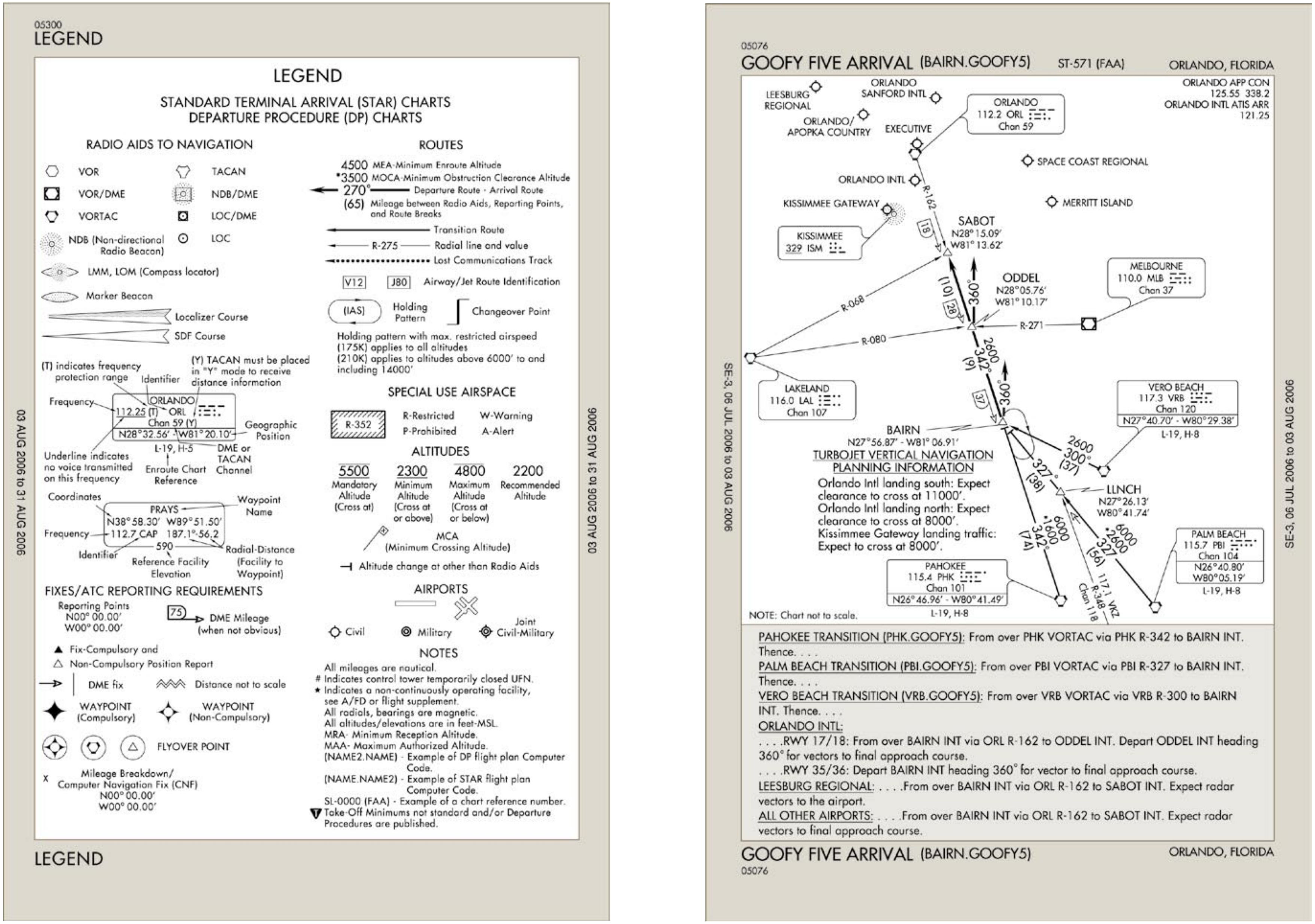
05300
LEGEND
LEGEND
STANDARD TERMINAL ARRIVAL (STAR) CHARTS
DEPARTURE PROCEDURE (DP) CHARTS
RADIO AIDS TO NAVIGATION
VOR
TACAN
VOR/DME
NDB/DME
VORTAC
LOC/DME
NDB (Non-directional Radio Beacon)
LOC
LMM, LOM (Compass locator)
Marker Beacon
Localizer Course
SDF Course
(T) indicates frequency protection range
Identifier
(Y) TACAN must be placed in "Y" mode to receive distance information
Frequency
112.25 (T) ORLANDO ORL Chan 59 (Y)
N28°32.56' W81°20.10'
Geographic Position
L-19, H-5
Underline indicates no voice transmitted on this frequency
Enroute Chart Reference
DME or TACAN Channel
Coordinates
PRAYS
Waypoint Name
N38°58.30' W89°51.50'
Frequency
112.7 CAP 187.1°-56.2
590
Identifier
Reference Facility Elevation
Radial-Distance (Facility to Waypoint)
FIXES/ATC REPORTING REQUIREMENTS
Reporting Points N00°00.00' W00°00.00'
75
DME Mileage (when not obvious)
Fix-Compulsory and
Non-Compulsory Position Report
DME fix
Distance not to scale
WAYPOINT (Compulsory)
WAYPOINT (Non-Compulsory)
FLYOVER POINT
Mileage Breakdown/ Computer Navigation Fix (CNF) N00°00.00' W00°00.00'
ROUTES
4500 MEA-Minimum Enroute Altitude
*3500 MOCA-Minimum Obstruction Clearance Altitude
270° Departure Route - Arrival Route
(65) Mileage between Radio Aids, Reporting Points, and Route Breaks
Transition Route
R-275 Radial line and value
Lost Communications Track
V12
J80
Airway/Jet Route Identification
(IAS)
Holding Pattern
Changeover Point
Holding pattern with max. restricted airspeed
(175K) applies to all altitudes
(210K) applies to altitudes above 6000' to and including 14000'
SPECIAL USE AIRSPACE
R-352
R-Restricted
P-Prohibited
W-Warning
A-Alert
ALTITUDES
5500 Mandatory Altitude (Cross at)
2300 Minimum Altitude (Cross at or above)
4800 Maximum Altitude (Cross at or below)
2200 Recommended Altitude
MCA (Minimum Crossing Altitude)
Altitude change at other than Radio Aids
AIRPORTS
Civil
Military
Joint Civil-Military
NOTES
All mileages are nautical.
# Indicates control tower temporarily closed UFN.
★ Indicates a non-continuously operating facility, see A/FD or flight supplement.
All radials, bearings are magnetic.
All altitudes/elevations are in feet-MSL.
MRA- Minimum Reception Altitude.
MAA- Maximum Authorized Altitude.
(NAME2.NAME) - Example of DP flight plan Computer Code.
(NAME.NAME2) - Example of STAR flight plan Computer Code.
SL-0000 (FAA) - Example of a chart reference number.
Take-Off Minimums not standard and/or Departure Procedures are published.
03 AUG 2006 to 31 AUG 2006
03 AUG 2006 to 31 AUG 2006
LEGEND
05076
GOOFY FIVE ARRIVAL (BAIRN.GOOFY5)
ST-571 (FAA)
ORLANDO, FLORIDA
ORLANDO APP CON 125.55 338.2
ORLANDO INTL ATIS ARR 121.25
LEESBURG REGIONAL
ORLANDO SANFORD INTL
ORLANDO/ APOPKA COUNTRY
EXECUTIVE
ORLANDO 112.2 ORL Chan 59
SPACE COAST REGIONAL
ORLANDO INTL
MERRITT ISLAND
KISSIMMEE GATEWAY
KISSIMMEE 329 ISM
R-162
18
SABOT N28°15.09' W81°13.62'
ODDEL N28°05.76' W81°10.17'
MELBOURNE 110.0 MLB Chan 37
R-068
(10)
28
360°
R-271
R-080
2600
342°
(9)
LAKELAND 116.0 LAL Chan 107
37
360°
VERO BEACH 117.3 VRB Chan 120
N27°40.70' - W80°29.38'
L-19, H-8
BAIRN N27°56.87' - W81°06.91'
TURBOJET VERTICAL NAVIGATION PLANNING INFORMATION
Orlando Intl landing south: Expect clearance to cross at 11000'.
Orlando Intl landing north: Expect clearance to cross at 8000'.
Kissimmee Gateway landing traffic: Expect to cross at 8000'.
2600
300°
(37)
327°
(38)
LLNCH N27°26.13' W80°41.74'
6000
*1600
342°
(74)
6000
*2600
327°
(56)
117.1 VKZ R-348 Chan 118
PAHOKEE 115.4 PHK Chan 101
N26°46.96' - W80°41.49'
L-19, H-8
PALM BEACH 115.7 PBI Chan 104
N26°40.80' W80°05.19'
L-19, H-8
NOTE: Chart not to scale.
SE-3, 06 JUL 2006 to 03 AUG 2006
SE-3, 06 JUL 2006 to 03 AUG 2006
GOOFY FIVE ARRIVAL (BAIRN.GOOFY5)
ORLANDO, FLORIDA
05076

PAHOKEE TRANSITION (PHK.GOOFY5): From over PHK VORTAC via PHK R-342 to BAIRN INT. Thence. . . .

PALM BEACH TRANSITION (PBI.GOOFY5): From over PBI VORTAC via PBI R-327 to BAIRN INT. Thence. . . .

VERO BEACH TRANSITION (VRB.GOOFY5): From over VRB VORTAC via VRB R-300 to BAIRN INT. Thence. . . .

ORLANDO INTL:

. . . .RWY 17/18: From over BAIRN INT via ORL R-162 to ODDEL INT. Depart ODDEL INT heading 360° for vectors to final approach course.

. . . .RWY 35/36: Depart BAIRN INT heading 360° for vector to final approach course.

LEESBURG REGIONAL: . . . .From over BAIRN INT via ORL R-162 to SABOT INT. Expect radar vectors to the airport.

ALL OTHER AIRPORTS: . . . .From over BAIRN INT via ORL R-162 to SABOT INT. Expect radar vectors to final approach course.

Figure 8-9. *DP Chart Legend and STAR.*

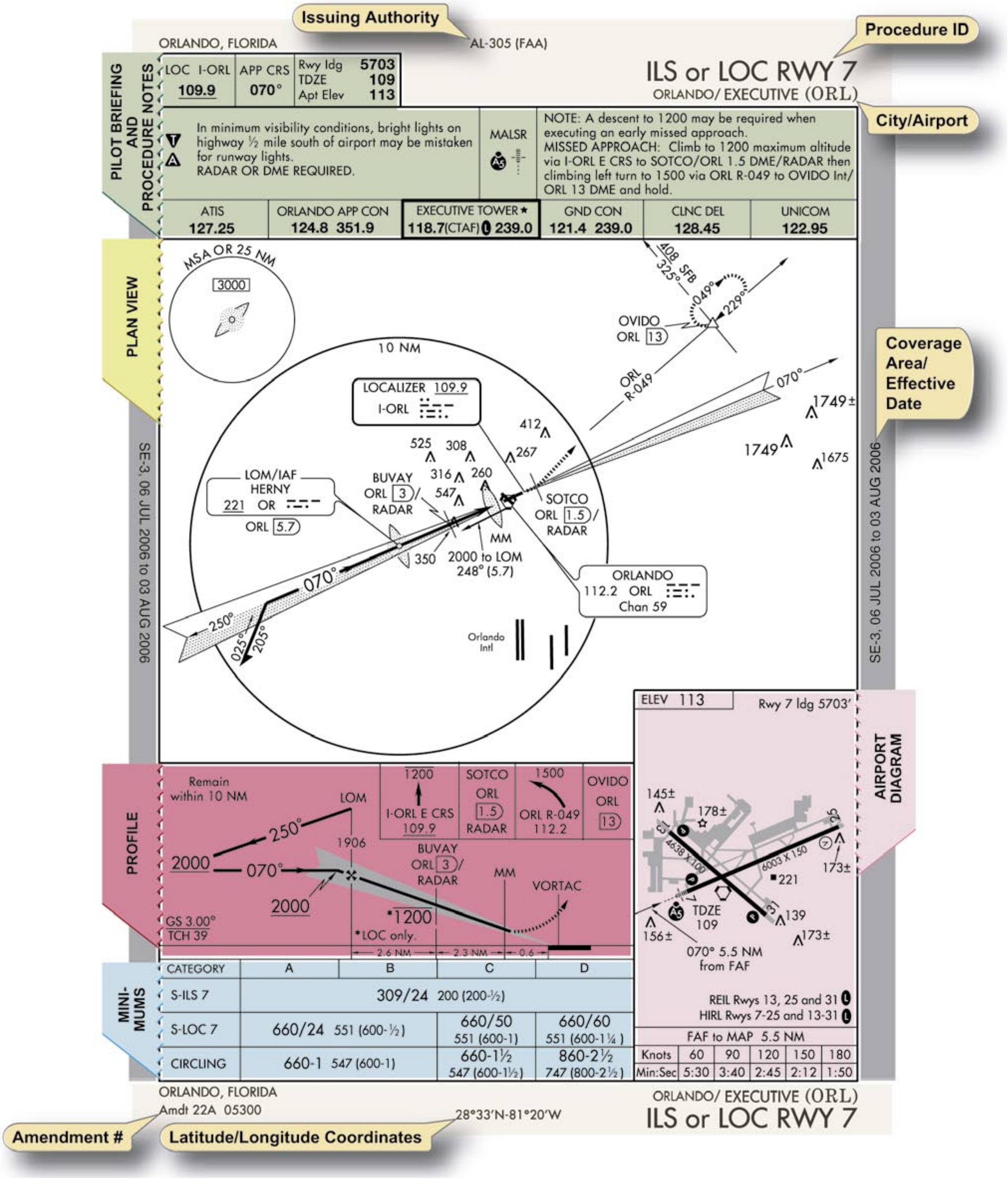
Issuing Authority
Procedure ID
City/Airport
Coverage Area/ Effective Date
Amendment #
Latitude/Longitude Coordinates
PILOT BRIEFING AND PROCEDURE NOTES
PLAN VIEW
PROFILE
MINI-MUMS
AIRPORT DIAGRAM
ORLANDO, FLORIDA
AL-305 (FAA)
ILS or LOC RWY 7
ORLANDO/ EXECUTIVE (ORL)
LOC I-ORL 109.9
APP CRS 070°
Rwy ldg 5703
TDZE 109
Apt Elev 113
MALSR
SE-3, 06 JUL 2006 to 03 AUG 2006
MSA OR 25 NM
3000
10 NM
LOCALIZER 109.9
I-ORL
LOM/IAF HERNY 221 OR
ORL 5.7
BUYAV ORL 3 / RADAR
SOTCO ORL 1.5 / RADAR
ORLANDO 112.2 ORL Chan 59
OVIDO ORL 13
ORL R-049
408 SFB 325°
049°
229°
070°
250°
025°
205°
2000 to LOM 248° (5.7)
MM
Orlando Intl
525
308
316
260
547
412
267
350
1749±
1749
1675
Remain within 10 NM
LOM
1906
2000
250°
070°
2000
GS 3.00°
TCH 39
*1200
*LOC only.
BUYAV ORL 3 / RADAR
MM
VORTAC
2.6 NM
2.3 NM
0.6
1200 I-ORL E CRS 109.9
SOTCO ORL 1.5 RADAR
1500 ORL R-049 112.2
OVIDO ORL 13
ELEV 113
Rwy 7 ldg 5703'
145±
178±
4638 X 100
6003 X 150
221
173±
TDZE 109
139
156±
173±
7
25
13
31
070° 5.5 NM from FAF
ORLANDO, FLORIDA
Amdt 22A 05300
28°33'N-81°20'W
ORLANDO/ EXECUTIVE (ORL)
ILS or LOC RWY 7

In minimum visibility conditions, bright lights on highway ½ mile south of airport may be mistaken for runway lights.
RADAR OR DME REQUIRED.

NOTE: A descent to 1200 may be required when executing an early missed approach.
MISSED APPROACH: Climb to 1200 maximum altitude via I-ORL E CRS to SOTCO/ORL 1.5 DME/RADAR then climbing left turn to 1500 via ORL R-049 to OVIDO Int/ ORL 13 DME and hold.

ATIS	ORLANDO APP CON	EXECUTIVE TOWER★	GND CON	CLNC DEL	UNICOM
127.25	124.8 351.9	118.7(CTAF) Ⓛ 239.0	121.4 239.0	128.45	122.95

CATEGORY	A	B	C	D
S-ILS 7	309/24 200 (200-½)			
S-LOC 7	660/24 551 (600-½)		660/50 551 (600-1)	660/60 551 (600-1¼)
CIRCLING	660-1 547 (600-1)		660-1½ 547 (600-1½)	860-2½ 747 (800-2½)

REIL Rwys 13, 25 and 31 Ⓛ
HIRL Rwys 7-25 and 13-31 Ⓛ

FAF to MAP 5.5 NM

Knots	60	90	120	150	180
Min:Sec	5:30	3:40	2:45	2:12	1:50

Figure 8-10. *Instrument Approach Chart.*

The procedure chart title (top and bottom margin area of *Figure 8-10*) is derived from the type of navigational facility providing final approach course guidance. A runway number is listed when the approach course is aligned within 30º of the runway centerline. This type of approach allows a straight-in landing under the right conditions. The type of approach followed by a letter identifies approaches that do not have straight-in landing minimums. Examples include procedure titles at the same airport, which have only circling minimums. The first approach of this type created at the airport will be labeled with the letter A, and the lettering will continue in alphabetical order (e.g., "VOR-A or "LDA-B"). The letter designation signifies the expectation is for the procedure to culminate in a circling approach to land. As a general rule, circling-only approaches are designed for one of the two following reasons:

- The final approach course alignment with the runway centerline exceeds 30º.
- The descent gradient is greater than 400 feet per NM from the FAF to the threshold crossing height (TCH). When this maximum gradient is exceeded, the circling-only approach procedure may be designed to meet the gradient criteria limits.

Further information on this topic can be found in the Instrument Procedures Handbook, Chapter 5, under Approach Naming Conventions.

To distinguish between the left, right, and center runways, an "L," "R," or "C" follows the runway number (e.g., "ILS RWY 16R"). In some cases, an airport might have more than one circling approach, shown as VOR-A, VOR/DME-B, etc.

More than one navigational system separated by a slash indicates more than one type of equipment is required to execute the final approach (e.g., VOR/DME RWY 31). More than one navigational system separated by "or" indicates either type of equipment may be used to execute the final approach (e.g., VOR or GPS RWY 15). Multiple approaches of the same type, to the same runway and using the same guidance, have an additional letter from the end of the alphabet, number, or term in the title (e.g., ILS Z RWY 28, SILVER ILS RWY 28, or ILS 2 RWY 28). VOR/DME RNAV approaches are identified as VOR/DME RNAV RWY (runway number). Helicopters have special IAPs, designated with COPTER in the procedure identification (e.g., COPTER LOC/DME 25L). Other types of navigation systems may be required to execute other portions of the approach prior to intercepting the final approach segment or during the missed approach.

The Pilot Briefing

The pilot briefing is located at the top of the chart and provides the pilot with information required to complete the published approach procedure. Included in the pilot briefing are the NAVAID providing approach guidance, its frequency, the final approach course, and runway information. A notes section contains additional procedural information. For example, a procedural note might indicate restrictions for circling maneuvers. Some other notes might concern a local altimeter setting and the resulting change in the minimums. The use of RADAR may also be noted in this section. Additional notes may be found in the plan view.

When a triangle containing a "T" (T) appears in the notes section, it signifies the airport has nonstandard IFR takeoff minimums. Pilots should refer to the DPs section of the TPP to determine takeoff minimums.

When a triangle containing an "A" (A) appears in the notes section, it signifies the airport has nonstandard IFR alternate minimums. Civil pilots should refer to the Alternate Minimums Section of the TPP to determine alternate minimums. Military pilots should refer to appropriate regulations.

When a triangle containing an "A" NA (ANA) appears in the notes area, it signifies that Alternate Minimums are Not Authorized due to unmonitored facility or the absence of weather reporting service.

Communication frequencies are listed in the order in which they would be used during the approach. Frequencies for weather and related facilities are included, where applicable, such as automatic terminal information service (ATIS), automated surface observing system (ASOS), automated weather observation system (AWOS), and AFSSs.

The Plan View

The plan view provides a graphical overhead view of the procedure, and depicts the routes that guide the pilot from the en route segments to the initial approach fix (IAF). *[Figure 8-10]* During the initial approach, the aircraft has departed the en route phase of flight and is maneuvering to enter an intermediate or final segment of the instrument approach. An initial approach can be made along prescribed routes within the terminal area, which may be along an arc, radial, course, heading, radar vector, or a combination thereof. Procedure turns and high altitude teardrop penetrations are initial approach segments. Features of the plan view, including the procedure turn, obstacle elevation, minimum safe altitude (MSA), and procedure track, are depicted in *Figure 8-11*. Terrain will be depicted in the plan view portion of all IAPs if the terrain within the plan view exceeds 4,000 feet above the airport elevation, or if within a 6 nautical mile radius of the airport reference point the terrain rises at least 2,000 feet above the airport elevation.

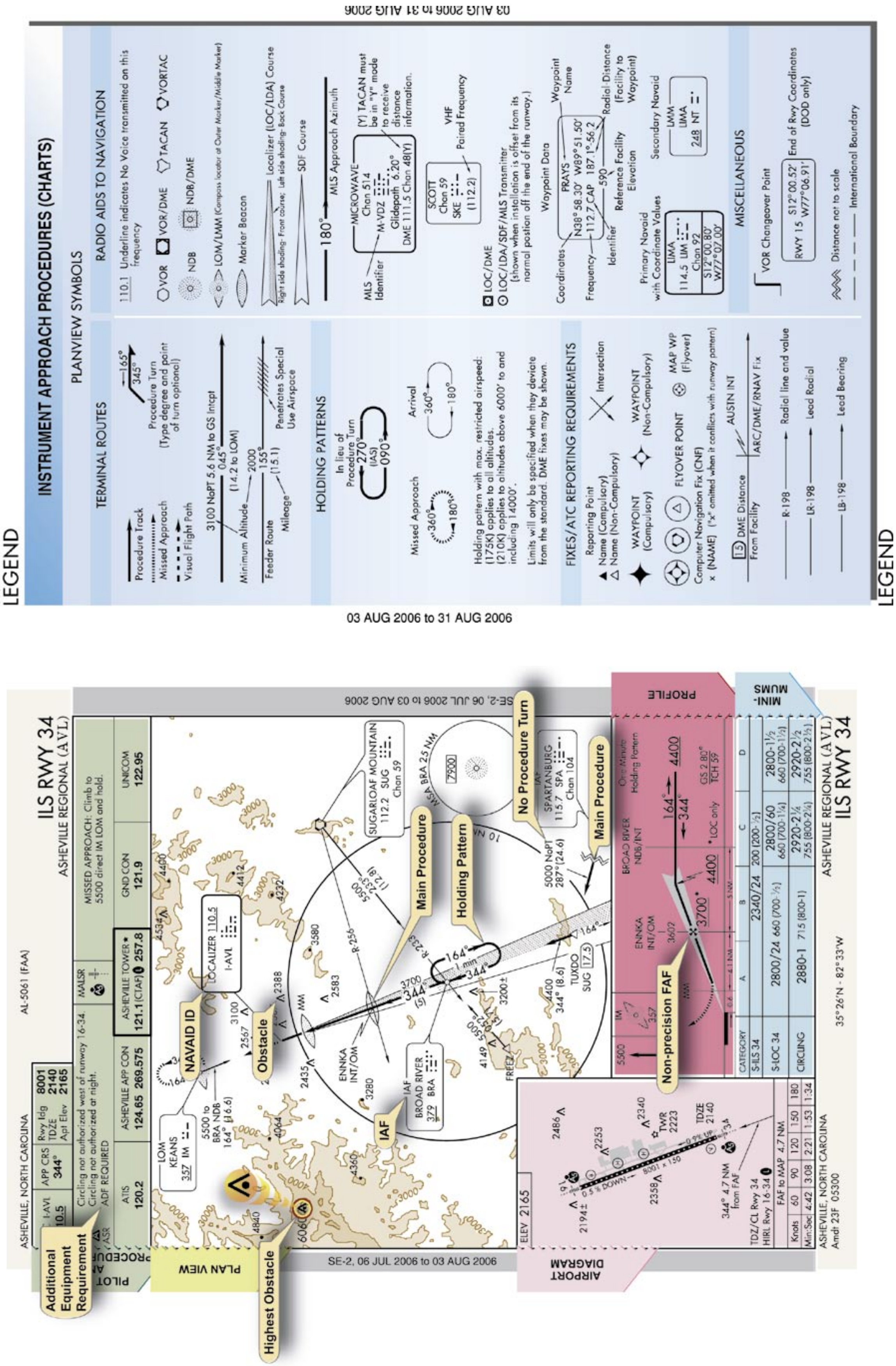
ASHEVILLE, NORTH CAROLINA
AL-5061 (FAA)
ILS RWY 34
ASHEVILLE REGIONAL (AVL)
LOC I-AVL 110.5
APP CRS 344°
Rwy ldg 8001
TDZE 2140
Apt Elev 2165
MALSR
Circling not authorized west of runway 16-34. Circling not authorized at night. ADF REQUIRED
MISSED APPROACH: Climb to 5500 direct IM LOM and hold.
ATIS 120.2
ASHEVILLE APP CON 124.65 269.575
ASHEVILLE TOWER 121.1 (CTAF) 257.8
GND CON 121.9
UNICOM 122.95
Additional Equipment Requirement
Highest Obstacle
NAVAID ID
Obstacle
IAF
Main Procedure
Holding Pattern
No Procedure Turn
Main Procedure
Non-precision FAF
PILOT BRIEFING
PLAN VIEW
AIRPORT DIAGRAM
PROFILE
MINIMUMS
SE-2, 06 JUL 2006 to 03 AUG 2006
ASHEVILLE, NORTH CAROLINA Amdt 23F 05300
35°26'N - 82°33'W
ASHEVILLE REGIONAL (AVL)
ILS RWY 34
06047
LEGEND
INSTRUMENT APPROACH PROCEDURES (CHARTS)
PLANVIEW SYMBOLS
TERMINAL ROUTES
RADIO AIDS TO NAVIGATION
HOLDING PATTERNS
FIXES/ATC REPORTING REQUIREMENTS
MISCELLANEOUS
03 AUG 2006 to 31 AUG 2006

Figure 8-11. *IAP Plan View and Symbol Legends.*

Some NACG charts contain a reference or distance circle with a specified radius (10 NM is most common). Normally, approach features within the plan view are shown to scale; however, only the data within the reference circle is always drawn to scale.

Concentric dashed circles, or concentric rings around the distance circle, are used when the information necessary to the procedure will not fit to scale within the limits of the plan view area. They serve as a means to systematically arrange this information in its relative position outside and beyond the reference circle. These concentric rings are labeled en route facilities and feeder facilities.

The primary airport depicted in the plan view is drawn with enough detail to show the runway orientation and final approach course alignment. Airports other than the primary approach airport are not normally depicted in the NACG plan view.

Known spot elevations are indicated on the plan view with a dot in MSL altitude. The largest dot and number combination indicates the highest elevation. An inverted "V" with a dot in the center depicts an obstacle. ⋀ The highest obstacle is indicated with a bolder, larger version of the same symbol. *[Figure 8-11]*

The MSA circle appears in the plan view, except in approaches for which the Terminal Arrival Area (TAA) format is used or appropriate NAVAIDs (e.g., VOR or NDB) are unavailable. The MSA is provided for emergency purposes only and guarantees 1,000 feet obstruction clearance in the sector indicated with reference to the bearings in the circle. For conventional navigation systems, the MSA is normally based on the primary omnidirectional facility (NAVAID) on which the IAP is predicated. The MSA depiction on the approach chart contains the facility identifier of the NAVAID used to determine the MSA altitudes. For RNAV approaches, the MSA is based on the runway waypoint for straight-in approaches, or the airport waypoint for circling approaches. For GPS approaches, the MSA center header will be the missed approach waypoint. The MSL altitudes appear in boxes within the circle, which is typically a 25 NM radius unless otherwise indicated. The MSA circle header refers to the letter identifier of the NAVAID or waypoint that describes the center of the circle.

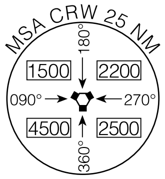

NAVAIDs necessary for the completion of the instrument procedure include the facility name, letter identifier, and Morse code sequence. They may also furnish the frequency, Morse code, and channel. A heavy-lined NAVAID box depicts the primary NAVAID used for the approach. An "I" in front of the NAVAID identifier (in *Figure 8-11*, "I-AVL") listed in the NAVAID box indicates a localizer. The requirement for an ADF, DME, or RADAR in the approach is noted in the plan view.

Intersections, fixes, radials, and course lines describe route and approach sequencing information. The main procedure or final approach course is a thick, solid line. ▬▬▬ A DME arc, which is part of the main procedure course, is also represented as a thick, solid line. ▬▬▬ A feeder route is depicted with a medium line ——— and provides heading, altitude, and distance information. (All three components must be designated on the chart to provide a navigable course.) Radials, such as lead radials, are shown by thin lines. ——— The missed approach track is drawn using a thin, hash marked line with a directional arrow. ||||||||||||||||||||||||||► A visual flight path segment appears as a thick dashed line with a directional arrow. - - - - - - - -➔ Visual Flight Path IAFs are charted IAF when associated with a NAVAID or when freestanding.

The missed approach holding pattern track is represented with a thin-dashed line. When collocated, the missed approach holding pattern and procedure turn holding pattern are indicated as a solid, black line. Arrival holding patterns are depicted as thin, solid lines.

Terminal Arrival Area (TAA)

The design objective of the TAA procedure is to provide a transition method for arriving aircraft with GPS/RNAV equipment. TAAs will also eliminate or reduce the need for feeder routes, departure extensions, and procedure turns or course reversal. The TAA is controlled airspace established in conjunction with the standard or modified RNAV approach configurations.

The standard TAA has three areas: straight-in, left base, and right base. The arc boundaries of the three areas of the TAA are published portions of the approach and allow aircraft to transition from the en route structure direct to the nearest IAF. When crossing the boundary of each of these areas or when released by ATC within the area, the pilot is expected to proceed direct to the appropriate waypoint IAF for the approach area being flown. A pilot has the option in all areas of proceeding directly to the holding pattern.

The TAA has a "T" structure that normally provides a NoPT for aircraft using the approach. *[Figure 8-12]* The TAA provides the pilot and air traffic controller with an efficient method for routing traffic from the en route to the terminal structure. The basic "T" contained in the TAA normally aligns the procedure on runway centerline, with the missed

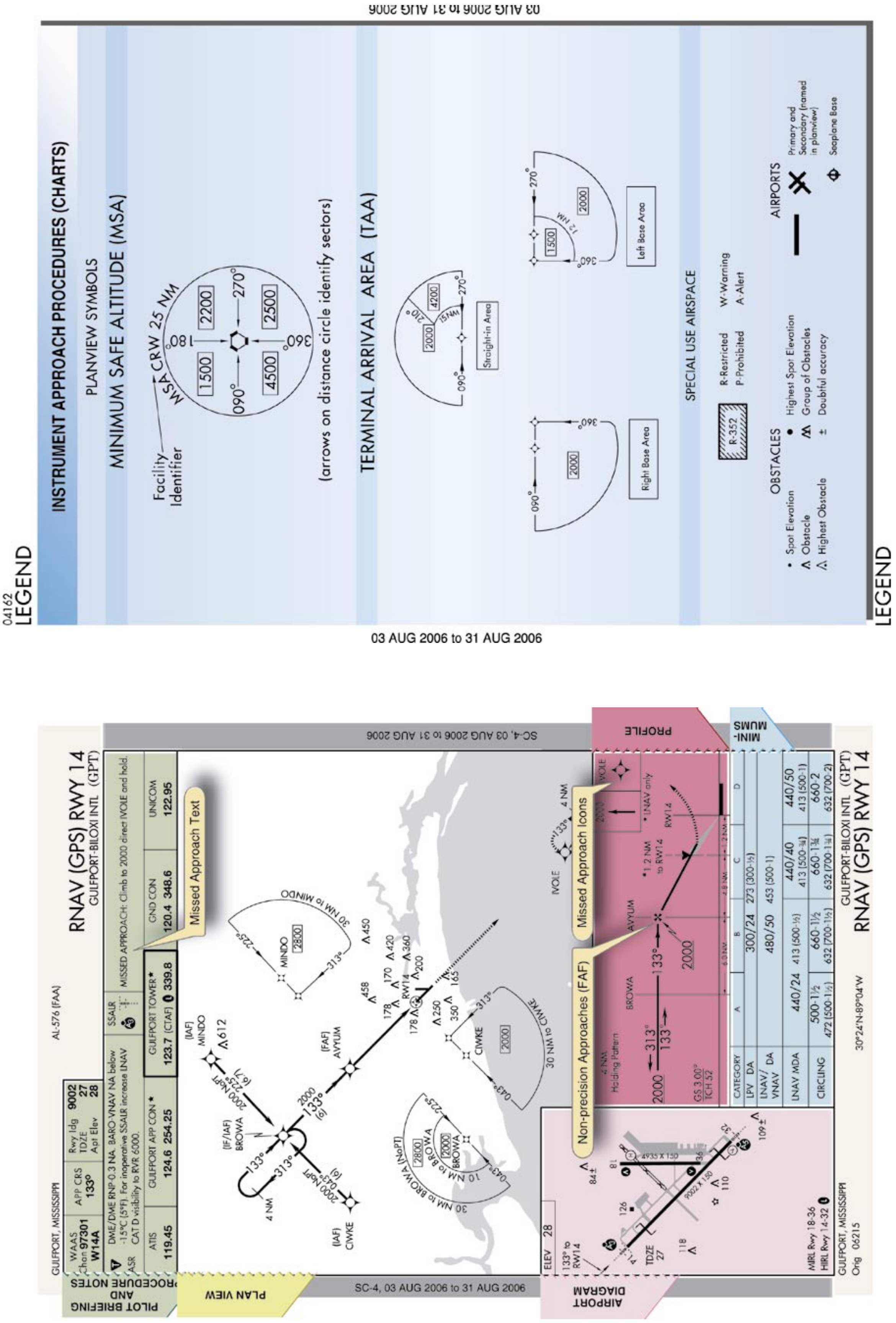
PILOT BRIEFING
PROCEDURE NOTES AND
PLAN VIEW
AIRPORT DIAGRAM
PROFILE
MINI-MUMS
Missed Approach Text
Missed Approach Icons
Non-precision Approaches (FAF)
RNAV (GPS) RWY 14
GULFPORT-BILOXI INTL (GPT)
GULFPORT, MISSISSIPPI
AL-576 (FAA)
SC-4, 03 AUG 2006 to 31 AUG 2006
WAAS Chan 97301 W14A
APP CRS 133°
Rwy ldg 9002
TDZE 27
Apt Elev 28
MISSED APPROACH: Climb to 2000 direct IVOLE and hold.
ATIS 119.45
GULFPORT APP CON 124.6 254.25
GULFPORT TOWER 123.7 (CTAF) 339.8
GND CON 120.4 348.6
UNICOM 122.95
(IAF) CIWKE
(IF/IAF) BROWA
(IAF) MINDO
(FAF) AYYUM
IVOLE
30 NM to MINDO
30 NM to CIWKE
30 NM to BROWA (NoPT)
10 NM to BROWA
CATEGORY A B C D
LPV DA 300/24 273 (300-½)
LNAV/VNAV DA 480/50 453 (500-1)
LNAV MDA 440/24 413 (500-½) 440/40 413 (500-¾) 440/50 413 (500-1)
CIRCLING 500-1½ 472 (500-1½) 660-1½ 632 (700-1½) 660-1¾ 632 (700-1¾) 660-2 632 (700-2)
30°24'N-89°04'W
Orig 06215
04162
LEGEND
INSTRUMENT APPROACH PROCEDURES (CHARTS)
PLANVIEW SYMBOLS
MINIMUM SAFE ALTITUDE (MSA)
Facility Identifier
MSA CRW 25 NM
(arrows on distance circle identify sectors)
TERMINAL ARRIVAL AREA (TAA)
Right Base Area
Straight-in Area
Left Base Area
SPECIAL USE AIRSPACE
R-352
R-Restricted
P-Prohibited
W-Warning
A-Alert
OBSTACLES
• Spot Elevation
Λ Obstacle
Highest Obstacle
Highest Spot Elevation
Group of Obstacles
± Doubtful accuracy
AIRPORTS
Primary and Secondary (named in planview)
Seaplane Base
03 AUG 2006 to 31 AUG 2006

Figure 8-12. *Basic "T" Design of Terminal Arrival Area (TAA) and Legend.*

approach point (MAP) located at the threshold, the FAF 5 NM from the threshold, and the intermediate fix (IF) 5 NM from the FAF.

In order to accommodate descent from a high en route altitude to the initial segment altitude, a hold in lieu of a procedure turn provides the aircraft with an extended distance for the necessary descent gradient. The holding pattern constructed for this purpose is always established on the center IAF waypoint. Other modifications may be required for parallel runways, or special operational requirements. When published, the RNAV chart will depict the TAA through the use of icons representing each TAA associated with the RNAV procedure. These icons are depicted in the plan view of the approach, generally arranged on the chart in accordance with their position relative to the aircraft's arrival from the en route structure.

Course Reversal Elements in Plan View and Profile View

Course reversals included in an IAP are depicted in one of three different ways: a 45°/180° procedure turn, a holding pattern in lieu of procedure turn, or a teardrop procedure. The maneuvers are required when it is necessary to reverse direction to establish the aircraft inbound on an intermediate or final approach course. Components of the required procedure are depicted in the plan view and the profile view. The maneuver must be completed within the distance and at the minimum altitude specified in the profile view. Pilots should coordinate with the appropriate ATC facility relating to course reversal during the IAP.

Procedure Turns

A procedure turn barbed arrow 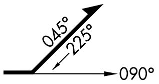 indicates the direction or side of the outbound course on which the procedure turn is made. *[Figure 8-13]* Headings are provided for course reversal using the 45° procedure turn. However, the point at which the turn may be commenced, and the type and rate of turn is left to the discretion of the pilot. Some of the options are the 45° procedure turn, the racetrack pattern, the teardrop procedure turn, or the 80°/260° course reversal. The absence of the procedure turn barbed arrow in the plan view indicates that a procedure turn is not authorized for that procedure. A maximum procedure turn speed of not greater than 200 knots indicated airspeed (KIAS) should be observed when turning outbound over the IAF and throughout the procedure turn maneuver to ensure staying within the obstruction clearance area. The normal procedure turn distance is 10 NM. This may be reduced to a minimum of 5 NM where only Category A or helicopter aircraft are operated, or increased to as much as 15 NM to accommodate high performance aircraft. Descent below the procedure turn altitude begins after the aircraft is established on the inbound course.

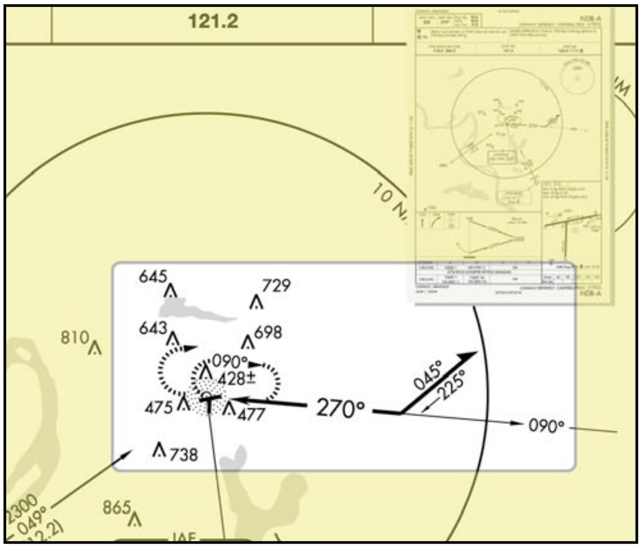

Figure 8-13. *45° Procedure Turn.*

The procedure turn is *not* required when the symbol "NoPT" appears, when radar vectoring to the final approach is provided, when conducting a timed approach, or when the procedure turn is not authorized. Pilots should contact the appropriate ATC facility when in doubt if a procedure turn is required.

Holding in Lieu of Procedure Turn

A holding pattern in lieu of a procedure turn may be specified for course reversal in some procedures. *[Figure 8-14]* In such cases, the holding pattern is established over an intermediate fix or a final approach fix (FAF). The holding pattern distance

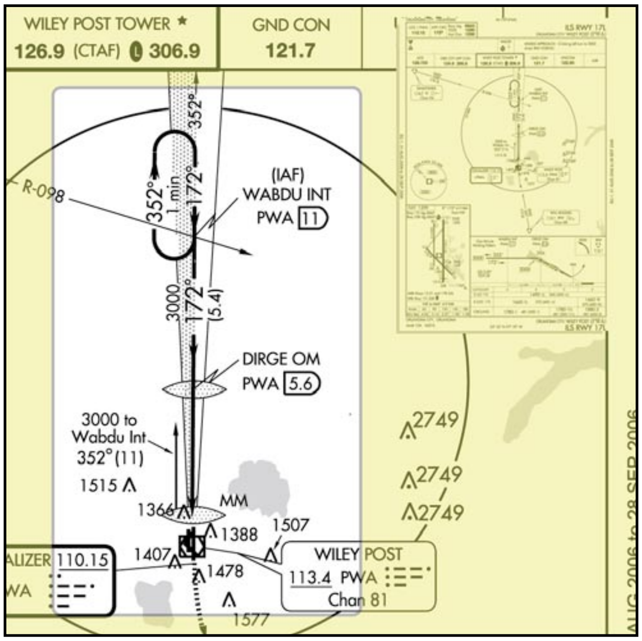

Figure 8-14. *Holding in Lieu of Procedure Turn.*

or time specified in the profile view must be observed. Maximum holding airspeed limitations as set forth for all holding patterns apply. The holding pattern maneuver is completed when the aircraft is established on the inbound course after executing the appropriate entry. If cleared for the approach prior to returning to the holding fix and the aircraft is at the prescribed altitude, additional circuits of the holding pattern are neither necessary nor expected by ATC. If pilots elect to make additional circuits to lose excessive altitude or to become better established on course, it is their responsibility to advise ATC upon receipt of their approach clearance. When holding in lieu of a procedure turn, the holding pattern must be followed, except when RADAR VECTORING to the final approach course is provided or when NoPT is shown on the approach course.

Teardrop Procedure

When a teardrop procedure turn is depicted and a course reversal is required, unless otherwise authorized by ATC, this type of procedure must be executed. *[Figure 8-15]* The teardrop procedure consists of departure from an IAF on the published outbound course followed by a turn toward and intercepting the inbound course at or prior to the intermediate fix or point. Its purpose is to permit an aircraft to reverse direction and lose considerable altitude within reasonably limited airspace. Where no fix is available to mark the beginning of the intermediate segment, it shall be assumed to commence at a point 10 NM prior to the FAF. When the facility is located on the airport, an aircraft is considered to be on final approach upon completion of the penetration turn. However, the final approach segment begins on the final approach course 10 NM from the facility.

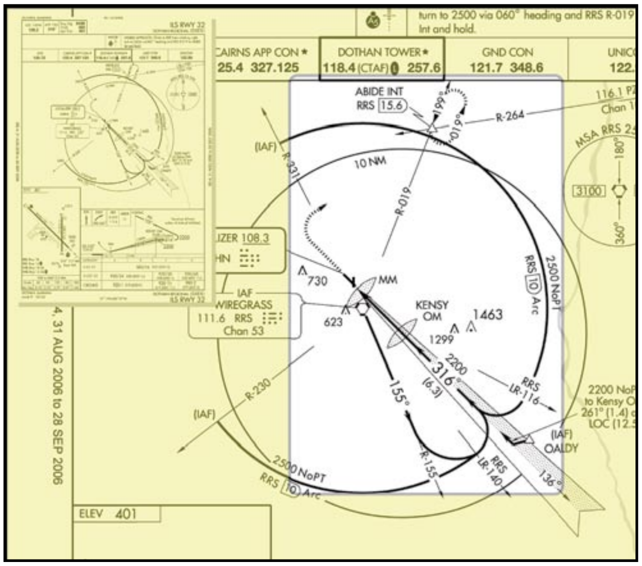

Figure 8-15. *Teardrop Procedure.*

The Profile View

The profile view is a depiction of the procedure from the side and illustrates the vertical approach path altitudes, headings, distances, and fixes. *[Figures 8-10, 8-11,* and *8-12]* The view includes the minimum altitude and the maximum distance for the procedure turn, altitudes over prescribed fixes, distances between fixes, and the missed approach procedure. The profile view aids in the pilot's interpretation of the IAP. The profile view is not drawn to scale. *[Figures 8-10, 8-11, 8-12,* and *8-16]*

The precision approach glide slope (GS) intercept altitude is a minimum altitude for GS interception after completion of the procedure turn, illustrated by an altitude number and "zigzag" line. It applies to precision approaches, and except where otherwise prescribed, also applies as a minimum altitude for crossing the FAF when the GS is inoperative or not used. Precision approach profiles also depict the GS angle of descent, threshold-crossing height (TCH), and GS altitude at the outer marker (OM).

For nonprecision approaches, a final descent is initiated and the final segment begins at either the FAF or the final approach point (FAP). The FAF is identified by use of the Maltese cross symbol in the profile view. ✖ *[Figure 8-11]* When no FAF is depicted, the final approach point is the point at which the aircraft is established inbound on the final approach course. *[Figure 8-16]*

Stepdown fixes in nonprecision procedures are provided between the FAF and the airport for authorizing a lower minimum descent altitude (MDA) after passing an obstruction. Stepdown fixes can be identified by NAVAID, NAVAID fix, waypoint or radar, and are depicted by a hash marked line. ¦ Normally, there is only one stepdown fix between the FAF and the MAP, but there can be several. If the stepdown fix cannot be identified for any reason, the minimum altitude at the stepdown fix becomes the MDA for the approach. However, circling minimums apply if they are higher than the stepdown fix minimum altitude, and a circling approach is required.

The visual descent point (VDP) is a defined point on the final approach course of a nonprecision straight-in approach procedure. A normal descent from the MDA to the runway touchdown point may be commenced, provided visual reference is established. The VDP is identified on the profile view of the approach chart by the symbol "V." *[Figure 8-12]*

The MAP varies depending upon the approach flown. For the ILS, the MAP is at the decision altitude/decision height (DA/DH). For nonprecision procedures, the pilot determines

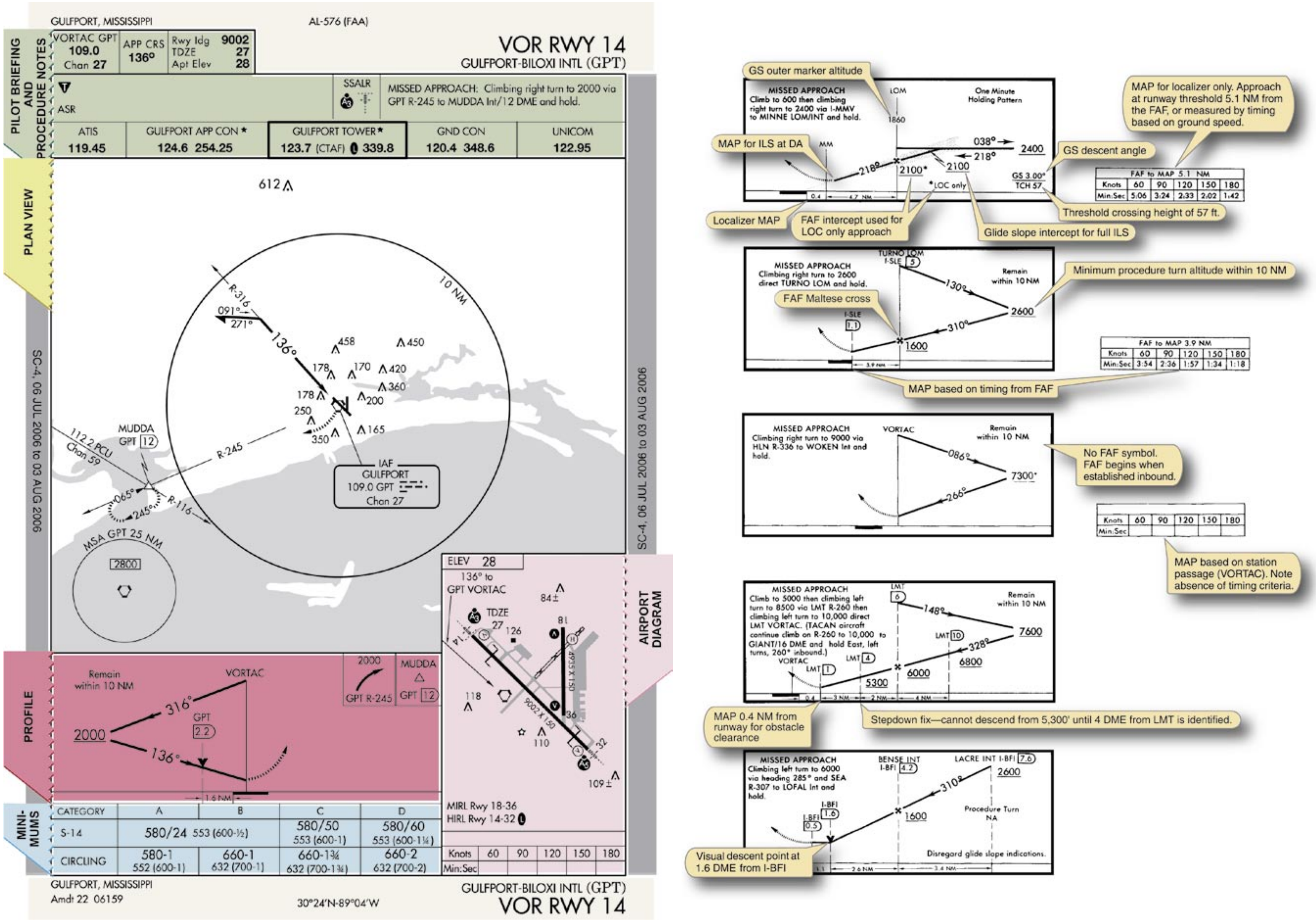
GULFPORT, MISSISSIPPI
AL-576 (FAA)
VOR RWY 14
GULFPORT-BILOXI INTL (GPT)
PILOT BRIEFING AND PROCEDURE NOTES
VORTAC GPT 109.0 Chan 27
APP CRS 136°
Rwy ldg 9002
TDZE 27
Apt Elev 28
ASR
SSALR
MISSED APPROACH: Climbing right turn to 2000 via GPT R-245 to MUDDA Int/12 DME and hold.
ATIS 119.45
GULFPORT APP CON ★ 124.6 254.25
GULFPORT TOWER ★ 123.7 (CTAF) 339.8
GND CON 120.4 348.6
UNICOM 122.95
PLAN VIEW
SC-4, 06 JUL 2006 to 03 AUG 2006
10 NM
R-316
091°
271°
136°
R-245
R-116
MUDDA GPT 12
112.2 PCU Chan 59
065°
245°
MSA GPT 25 NM
2800
IAF GULFPORT 109.0 GPT Chan 27
612
458
450
420
360
200
170
178
250
350
165
PROFILE
Remain within 10 NM
2000
316°
136°
GPT 2.2
VORTAC
1.6 NM
MUDDA GPT 12
GPT R-245
AIRPORT DIAGRAM
ELEV 28
136° to GPT VORTAC
TDZE 27
126
84±
18
4935 X 150
9002 X 150
36
32
14
118
110
109±
MIRL Rwy 18-36
HIRL Rwy 14-32
MINI-MUMS
CATEGORY
A
B
C
D
S-14
580/24 553 (600-½)
580/50 553 (600-1)
580/60 553 (600-1¼)
CIRCLING
580-1 552 (600-1)
660-1 632 (700-1)
660-1¾ 632 (700-1¾)
660-2 632 (700-2)
Knots 60 90 120 150 180
Min:Sec
GULFPORT, MISSISSIPPI
Amdt 22 06159
30°24'N-89°04'W
GULFPORT-BILOXI INTL (GPT)
VOR RWY 14
GS outer marker altitude
MISSED APPROACH Climb to 600 then climbing right turn to 2400 via I-MMV to MINNE LOM/INT and hold.
LOM
1860
One Minute Holding Pattern
MAP for ILS at DA
MM
218°
2100*
038°
218°
2100
2400
*LOC only
GS 3.00°
TCH 57
0.4
4.7 NM
GS descent angle
MAP for localizer only. Approach at runway threshold 5.1 NM from the FAF, or measured by timing based on ground speed.
FAF to MAP 5.1 NM
Knots 60 90 120 150 180
Min:Sec 5:06 3:24 2:33 2:02 1:42
Threshold crossing height of 57 ft.
Localizer MAP
FAF intercept used for LOC only approach
Glide slope intercept for full ILS
MISSED APPROACH Climbing right turn to 2600 direct TURNO LOM and hold.
TURNO LOM I-SLE 5
130°
Remain within 10 NM
2600
310°
1600
I-SLE 1.1
3.9 NM
FAF Maltese cross
Minimum procedure turn altitude within 10 NM
FAF to MAP 3.9 NM
Knots 60 90 120 150 180
Min:Sec 3:54 2:36 1:57 1:34 1:18
MAP based on timing from FAF
MISSED APPROACH Climbing right turn to 9000 via HLN R-336 to WOKEN Int and hold.
VORTAC
Remain within 10 NM
086°
7300'
266°
No FAF symbol. FAF begins when established inbound.
Knots 60 90 120 150 180
Min:Sec
MAP based on station passage (VORTAC). Note absence of timing criteria.
MISSED APPROACH Climb to 5000 then climbing left turn to 8500 via LMT R-260 then climbing left turn to 10,000 direct LMT VORTAC. (TACAN aircraft continue climb on R-260 to 10,000 to GIANT/16 DME and hold East, left turns, 260° inbound.)
LMT 6
148°
Remain within 10 NM
7600
LMT 10
328°
6800
LMT 4
6000
VORTAC
LMT 1
5300
0.4
3 NM
2 NM
4 NM
MAP 0.4 NM from runway for obstacle clearance
Stepdown fix—cannot descend from 5,300' until 4 DME from LMT is identified.
MISSED APPROACH Climbing left turn to 6000 via heading 285° and SEA R-307 to LOFAL Int and hold.
BENSE INT I-BFI 4.2
LACRE INT I-BFI 7.6
2600
310°
1600
I-BFI 1.6
I-BFI 0.5
Procedure Turn NA
Disregard glide slope indications.
1.1
2.6 NM
3.4 NM
Visual descent point at 1.6 DME from I-BFI

Figure 8-16. *More IAP Profile View Features.*

the MAP by timing from FAF when the approach aid is away from the airport, by a fix or NAVAID when the navigation facility is located on the field, or by waypoints as defined by GPS or VOR/DME RNAV. The pilot may execute the MAP early, but pilots should, unless otherwise cleared by ATC, fly the IAP as specified on the approach plate to the MAP at or above the MDA or DA/DH before executing a turning maneuver.

A complete description of the missed approach procedure appears in the pilot briefing section. *[Figure 8-16]* Icons indicating what is to be accomplished at the MAP are located in the profile view. When initiating a missed approach, the pilot will be directed to climb straight ahead (e.g., "Climb to 2,000") or commence a turning climb to a specified altitude (e.g., "Climbing right turn to 2,000"). In some cases, the procedure will direct the pilot to climb straight ahead to an initial altitude, then turn or enter a climbing turn to the holding altitude (e.g., "Climb to 900, then climbing right turn to 2,500 direct ABC VOR and hold").

When the missed approach procedure specifies holding at a facility or fix, the pilot proceeds according to the missed approach track and pattern depicted on the plan view. An alternate missed approach procedure may also be issued by ATC. The textual description will also specify the NAVAID(s) or radials that identify the holding fix.

The profile view also depicts minimum, maximum, recommended, and mandatory block altitudes used in approaches. The minimum altitude is depicted with the altitude underscored. 2500 On final approach, aircraft are required to maintain an altitude at or above the depicted altitude until reaching the subsequent fix. The maximum altitude will be depicted with the altitude overscored, 4300 and aircraft must remain at or below the depicted altitude. Mandatory altitudes will be depicted with the altitude both underscored and overscored, 5500 and altitude is to be maintained at the depicted value. Recommended altitudes are advisory altitudes and are neither over- nor underscored. When an over- or underscore spans two numbers, a mandatory block altitude is indicated, and aircraft are required to maintain altitude within the range of the two numbers. *[Figures 8-11* and *8-12]*

The Vertical Descent Angle (VDA) found on nonprecision approach charts provides the pilot with information required to establish a stabilized approach descent from the FAF or stepdown fix to the threshold crossing height (TCH). *[Figure 8-17]* Pilots can use the published angle and estimated or actual ground speed to find a target rate of descent using the rate of descent table in the back of the TPP.

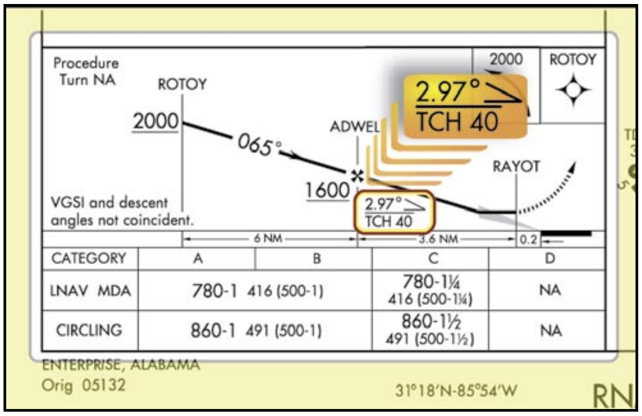

Figure 8-17. *Vertical Descent Angle (VDA).*

Landing Minimums

The minimums section sets forth the lowest altitude and visibility requirements for the approach, whether precision or nonprecision, straight-in or circling, or radar vectored. When a fix is incorporated in a nonprecision final segment, two sets of minimums may be published, depending upon how the fix can be identified. Two sets of minimums may also be published when a second altimeter source is used in the procedure. The minimums ensure that final approach obstacle clearance is provided from the start of the final segment to the runway or MAP, whichever occurs last. The same minimums apply to both day and night operations unless different minimums are specified in the Notes section of the pilot briefing. Published circling minimums provide obstacle clearance when pilots remain within the appropriate area of protection. *[Figure 8-18]*

Minimums are specified for various aircraft approach categories based upon a value 1.3 times the stalling speed of the aircraft in the landing configuration at maximum certified gross landing weight. If it is necessary to maneuver at speeds in excess of the upper limit of a speed range for a category, the minimums for the next higher category should be used. For example, an aircraft that falls into category A, but is circling to land at a speed in excess of 91 knots, should use approach category B minimums when circling to land. *[Figure 8-19]*

The minimums for straight-in and circling appear directly under each aircraft category. *[Figure 8-19]* When there is no solid division line between minimums for each category on the rows for straight-in or circling, the minimums apply to the two or more categories.

The terms used to describe the minimum approach altitudes differ between precision and nonprecision approaches.

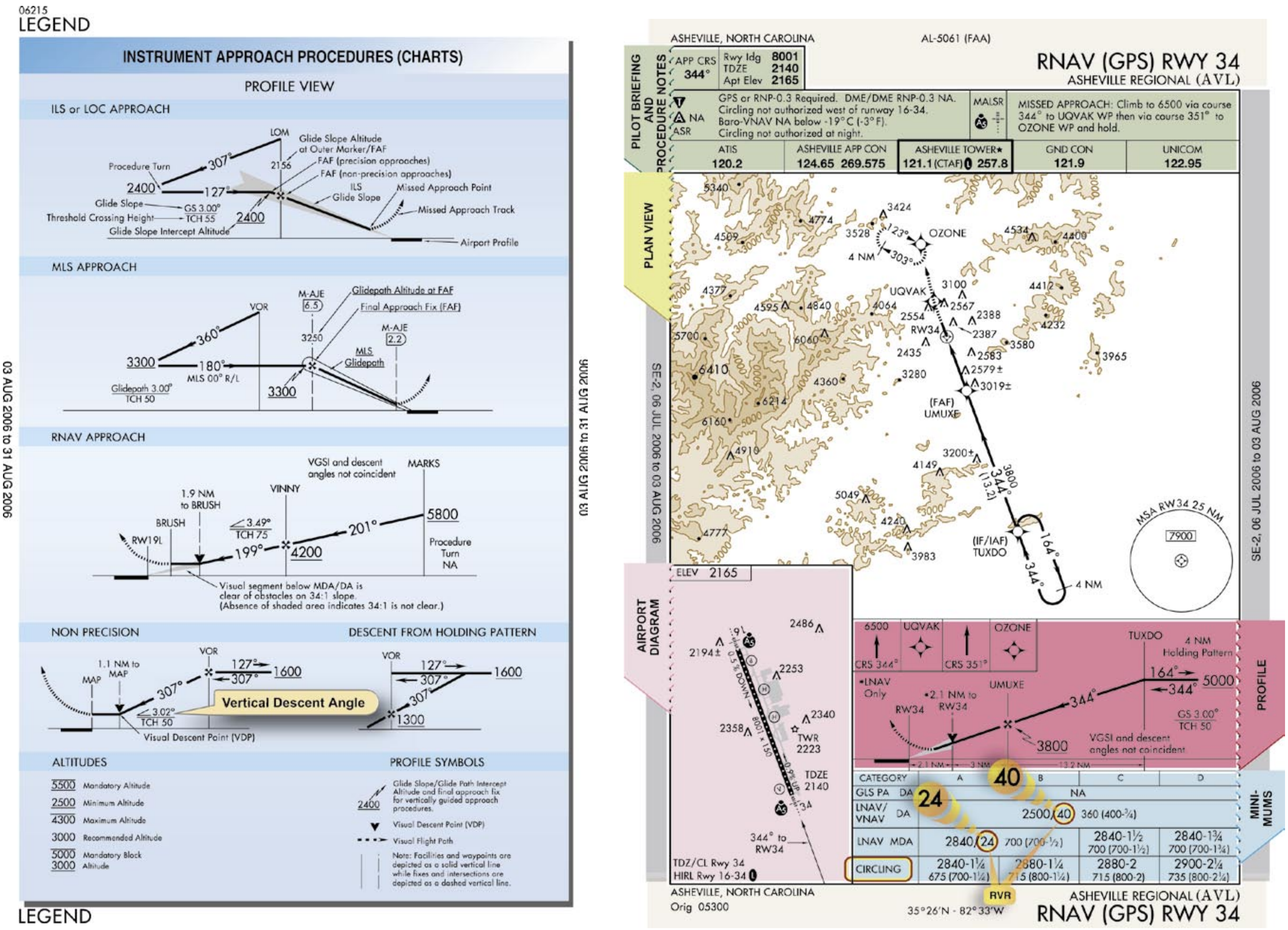
06215
LEGEND
INSTRUMENT APPROACH PROCEDURES (CHARTS)
PROFILE VIEW
ILS or LOC APPROACH
Procedure Turn
LOM
Glide Slope Altitude at Outer Marker/FAF
FAF (precision approaches)
FAF (non-precision approaches)
ILS Glide Slope
Missed Approach Point
Missed Approach Track
Airport Profile
Glide Slope
Threshold Crossing Height
Glide Slope Intercept Altitude
GS 3.00°
TCH 55
MLS APPROACH
VOR
M-AJE
Glidepath Altitude at FAF
Final Approach Fix (FAF)
MLS Glidepath
MLS 00° R/L
Glidepath 3.00°
TCH 50
RNAV APPROACH
VGSI and descent angles not coincident
MARKS
VINNY
1.9 NM to BRUSH
BRUSH
RW19L
3.49°
TCH 75
Procedure Turn NA
Visual segment below MDA/DA is clear of obstacles on 34:1 slope.
(Absence of shaded area indicates 34:1 is not clear.)
NON PRECISION
DESCENT FROM HOLDING PATTERN
1.1 NM to MAP
MAP
3.02°
TCH 50
Visual Descent Point (VDP)
Vertical Descent Angle
ALTITUDES
5500 Mandatory Altitude
2500 Minimum Altitude
4300 Maximum Altitude
3000 Recommended Altitude
5000 3000 Mandatory Block Altitude
PROFILE SYMBOLS
Glide Slope/Glide Path Intercept Altitude and final approach fix for vertically guided approach procedures.
Visual Descent Point (VDP)
Visual Flight Path
Note: Facilities and waypoints are depicted as a solid vertical line while fixes and intersections are depicted as a dashed vertical line.
LEGEND
03 AUG 2006 to 31 AUG 2006
ASHEVILLE, NORTH CAROLINA
AL-5061 (FAA)
RNAV (GPS) RWY 34
ASHEVILLE REGIONAL (AVL)
PILOT BRIEFING AND PROCEDURE NOTES
APP CRS 344°
Rwy ldg 8001
TDZE 2140
Apt Elev 2165
GPS or RNP-0.3 Required. DME/DME RNP-0.3 NA.
Circling not authorized west of runway 16-34.
Baro-VNAV NA below -19°C (-3°F).
Circling not authorized at night.
MALSR
MISSED APPROACH: Climb to 6500 via course 344° to UQVAK WP then via course 351° to OZONE WP and hold.
ATIS 120.2
ASHEVILLE APP CON 124.65 269.575
ASHEVILLE TOWER★ 121.1 (CTAF) 257.8
GND CON 121.9
UNICOM 122.95
PLAN VIEW
OZONE
UQVAK
RW34
(FAF) UMUXE
(IF/IAF) TUXDO
MSA RW34 25 NM
7900
SE-2, 06 JUL 2006 to 03 AUG 2006
AIRPORT DIAGRAM
ELEV 2165
0.5% DOWN
0.9% UP
8001 x 150
TWR 2223
TDZE 2140
344° to RW34
TDZ/CL Rwy 34
HIRL Rwy 16-34
PROFILE
4 NM Holding Pattern
GS 3.00°
TCH 50
VGSI and descent angles not coincident
•LNAV Only
•2.1 NM to RW34
MINI-MUMS
CATEGORY A B C D
GLS PA DA NA
LNAV/VNAV DA 2500/40 360 (400-¾)
LNAV MDA 2840/24 700 (700-½) 2840-1½ 700 (700-1½) 2840-1¾ 700 (700-1¾)
CIRCLING 2840-1¼ 675 (700-1¼) 2880-1¼ 715 (800-1¼) 2880-2 715 (800-2) 2900-2¼ 735 (800-2¼)
24
40
RVR
ASHEVILLE, NORTH CAROLINA
Orig 05300
35°26'N - 82°33'W
ASHEVILLE REGIONAL (AVL)
RNAV (GPS) RWY 34

Figure 8-18. *IAP Profile Legend.*

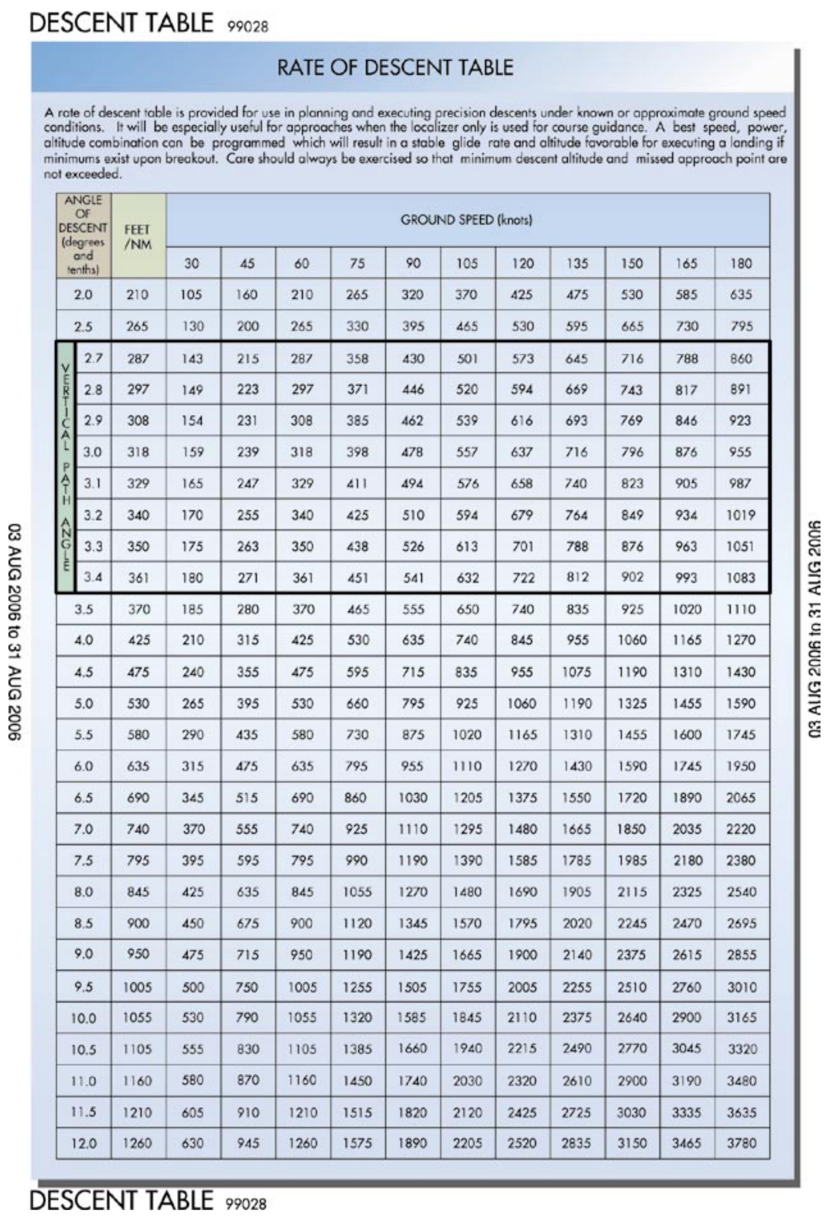

DESCENT TABLE 99028

03 AUG 2006 to 31 AUG 2006

RATE OF DESCENT TABLE

A rate of descent table is provided for use in planning and executing precision descents under known or approximate ground speed conditions. It will be especially useful for approaches when the localizer only is used for course guidance. A best speed, power, altitude combination can be programmed which will result in a stable glide rate and altitude favorable for executing a landing if minimums exist upon breakout. Care should always be exercised so that minimum descent altitude and missed approach point are not exceeded.

ANGLE OF DESCENT (degrees and tenths)	FEET /NM	GROUND SPEED (knots)										
		30	45	60	75	90	105	120	135	150	165	180
2.0	210	105	160	210	265	320	370	425	475	530	585	635
2.5	265	130	200	265	330	395	465	530	595	665	730	795
VERTICAL PATH ANGLE 2.7	287	143	215	287	358	430	501	573	645	716	788	860
2.8	297	149	223	297	371	446	520	594	669	743	817	891
2.9	308	154	231	308	385	462	539	616	693	769	846	923
3.0	318	159	239	318	398	478	557	637	716	796	876	955
3.1	329	165	247	329	411	494	576	658	740	823	905	987
3.2	340	170	255	340	425	510	594	679	764	849	934	1019
3.3	350	175	263	350	438	526	613	701	788	876	963	1051
3.4	361	180	271	361	451	541	632	722	812	902	993	1083
3.5	370	185	280	370	465	555	650	740	835	925	1020	1110
4.0	425	210	315	425	530	635	740	845	955	1060	1165	1270
4.5	475	240	355	475	595	715	835	955	1075	1190	1310	1430
5.0	530	265	395	530	660	795	925	1060	1190	1325	1455	1590
5.5	580	290	435	580	730	875	1020	1165	1310	1455	1600	1745
6.0	635	315	475	635	795	955	1110	1270	1430	1590	1745	1950
6.5	690	345	515	690	860	1030	1205	1375	1550	1720	1890	2065
7.0	740	370	555	740	925	1110	1295	1480	1665	1850	2035	2220
7.5	795	395	595	795	990	1190	1390	1585	1785	1985	2180	2380
8.0	845	425	635	845	1055	1270	1480	1690	1905	2115	2325	2540
8.5	900	450	675	900	1120	1345	1570	1795	2020	2245	2470	2695
9.0	950	475	715	950	1190	1425	1665	1900	2140	2375	2615	2855
9.5	1005	500	750	1005	1255	1505	1755	2005	2255	2510	2760	3010
10.0	1055	530	790	1055	1320	1585	1845	2110	2375	2640	2900	3165
10.5	1105	555	830	1105	1385	1660	1940	2215	2490	2770	3045	3320
11.0	1160	580	870	1160	1450	1740	2030	2320	2610	2900	3190	3480
11.5	1210	605	910	1210	1515	1820	2120	2425	2725	3030	3335	3635
12.0	1260	630	945	1260	1575	1890	2205	2520	2835	3150	3465	3780

DESCENT TABLE 99028

03 AUG 2006 to 31 AUG 2006

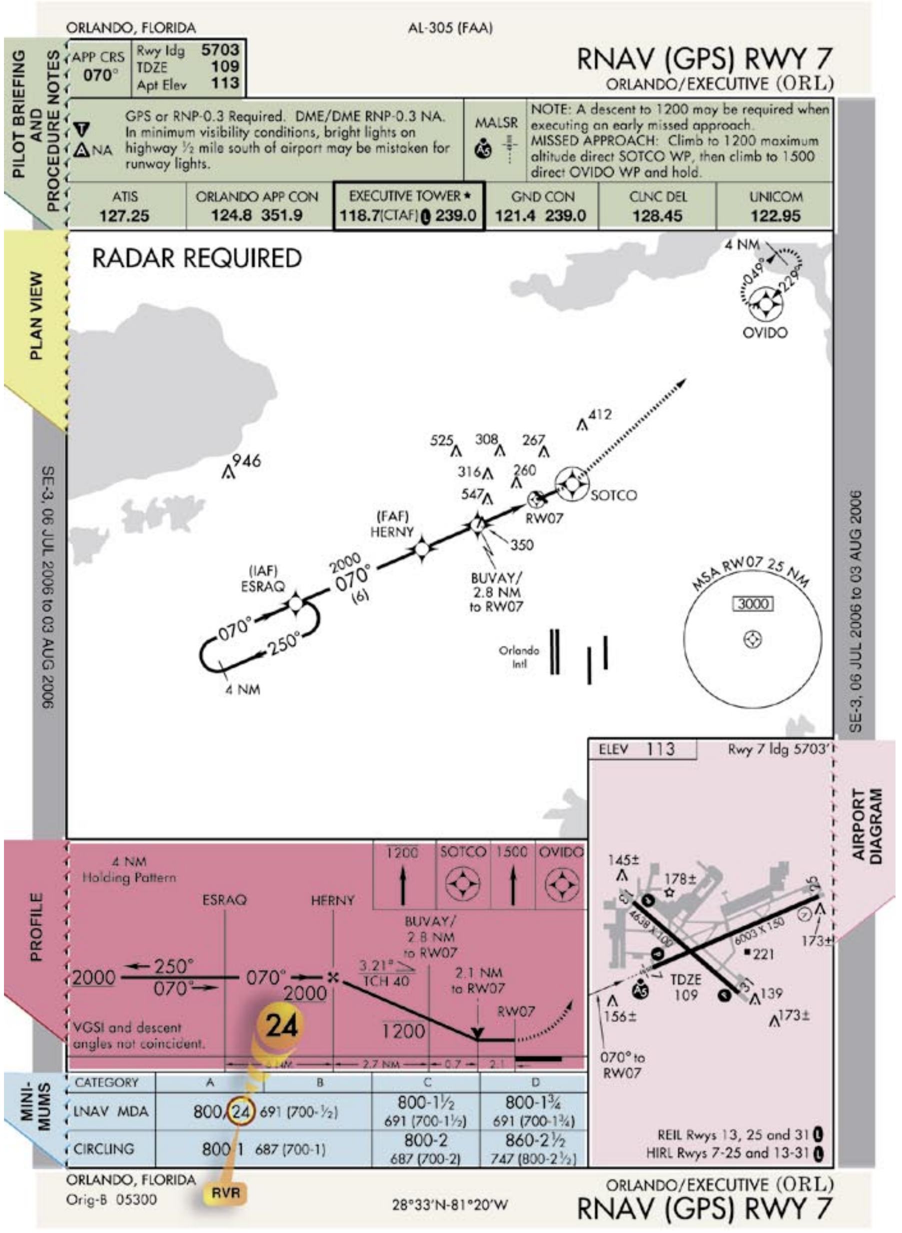

ORLANDO, FLORIDA

AL-305 (FAA)

RNAV (GPS) RWY 7
ORLANDO/EXECUTIVE (ORL)

PILOT BRIEFING AND PROCEDURE NOTES

APP CRS 070°

Rwy ldg 5703
TDZE 109
Apt Elev 113

GPS or RNP-0.3 Required. DME/DME RNP-0.3 NA. In minimum visibility conditions, bright lights on highway ½ mile south of airport may be mistaken for runway lights.

MALSR

NOTE: A descent to 1200 may be required when executing an early missed approach.
MISSED APPROACH: Climb to 1200 maximum altitude direct SOTCO WP, then climb to 1500 direct OVIDO WP and hold.

ATIS	ORLANDO APP CON	EXECUTIVE TOWER ★	GND CON	CLNC DEL	UNICOM
127.25	124.8 351.9	118.7(CTAF) Ⓛ 239.0	121.4 239.0	128.45	122.95

PLAN VIEW

RADAR REQUIRED

PROFILE

4 NM Holding Pattern

ESRAQ

HERNY

1200 SOTCO 1500 OVIDO

BUVAY/ 2.8 NM to RW07

2000 ←250° 070°→

070°→ 2000

3.21° TCH 40

2.1 NM to RW07

1200

RW07

VGSI and descent angles not coincident.

2.7 NM 0.7 2.1

24

MINIMUMS

CATEGORY	A	B	C	D
LNAV MDA	800/24 691 (700-½)		800-1½ 691 (700-1½)	800-1¾ 691 (700-1¾)
CIRCLING	800-1 687 (700-1)		800-2 687 (700-2)	860-2½ 747 (800-2½)

RVR

REIL Rwys 13, 25 and 31 Ⓛ
HIRL Rwys 7-25 and 13-31 Ⓛ

ORLANDO, FLORIDA
Orig-B 05300

28°33'N-81°20'W

ORLANDO/EXECUTIVE (ORL)
RNAV (GPS) RWY 7

Figure 8-19. *Descent Rate Table.*

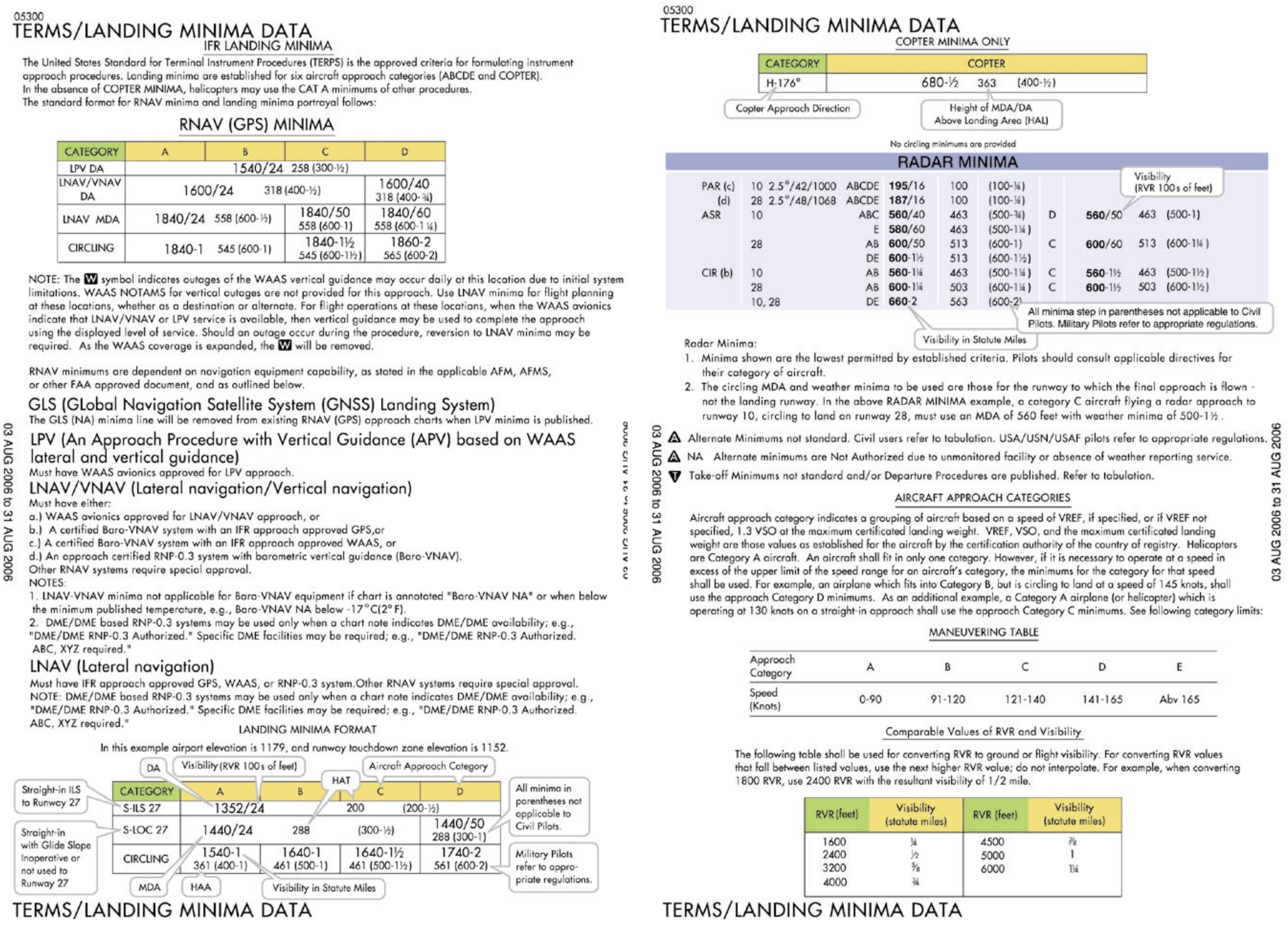

05300

TERMS/LANDING MINIMA DATA

IFR LANDING MINIMA

The United States Standard for Terminal Instrument Procedures (TERPS) is the approved criteria for formulating instrument approach procedures. Landing minima are established for six aircraft approach categories (ABCDE and COPTER).
In the absence of COPTER MINIMA, helicopters may use the CAT A minimums of other procedures.
The standard format for RNAV minima and landing minima portrayal follows:

RNAV (GPS) MINIMA

CATEGORY	A	B	C	D
LPV DA	1540/24 258 (300-½)			
LNAV/VNAV DA	1600/24 318 (400-½)			1600/40 318 (400-¾)
LNAV MDA	1840/24 558 (600-½)		1840/50 558 (600-1)	1840/60 558 (600-1¼)
CIRCLING	1840-1 545 (600-1)		1840-1½ 545 (600-1½)	1860-2 565 (600-2)

NOTE: The W symbol indicates outages of the WAAS vertical guidance may occur daily at this location due to initial system limitations. WAAS NOTAMS for vertical outages are not provided for this approach. Use LNAV minima for flight planning at these locations, whether as a destination or alternate. For flight operations at these locations, when the WAAS avionics indicate that LNAV/VNAV or LPV service is available, then vertical guidance may be used to complete the approach using the displayed level of service. Should an outage occur during the procedure, reversion to LNAV minima may be required. As the WAAS coverage is expanded, the W will be removed.

RNAV minimums are dependent on navigation equipment capability, as stated in the applicable AFM, AFMS, or other FAA approved document, and as outlined below.

GLS (GLobal Navigation Satellite System (GNSS) Landing System)

The GLS (NA) minima line will be removed from existing RNAV (GPS) approach charts when LPV minima is published.

LPV (An Approach Procedure with Vertical Guidance (APV) based on WAAS lateral and vertical guidance)

Must have WAAS avionics approved for LPV approach.

LNAV/VNAV (Lateral navigation/Vertical navigation)

Must have either:
a.) WAAS avionics approved for LNAV/VNAV approach, or
b.) A certified Baro-VNAV system with an IFR approach approved GPS,or
c.) A certified Baro-VNAV system with an IFR approach approved WAAS, or
d.) An approach certified RNP-0.3 system with barometric vertical guidance (Baro-VNAV).
Other RNAV systems require special approval.
NOTES:
1. LNAV-VNAV minima not applicable for Baro-VNAV equipment if chart is annotated "Baro-VNAV NA" or when below the minimum published temperature, e.g., Baro-VNAV NA below -17°C(2°F).
2. DME/DME based RNP-0.3 systems may be used only when a chart note indicates DME/DME availability; e.g., "DME/DME RNP-0.3 Authorized." Specific DME facilities may be required; e.g., "DME/DME RNP-0.3 Authorized. ABC, XYZ required."

LNAV (Lateral navigation)

Must have IFR approach approved GPS, WAAS, or RNP-0.3 system.Other RNAV systems require special approval.
NOTE: DME/DME based RNP-0.3 systems may be used only when a chart note indicates DME/DME availability; e.g., "DME/DME RNP-0.3 Authorized." Specific DME facilities may be required; e.g., "DME/DME RNP-0.3 Authorized. ABC, XYZ required."

LANDING MINIMA FORMAT

In this example airport elevation is 1179, and runway touchdown zone elevation is 1152.

CATEGORY	A	B	C	D
S-ILS 27	1352/24		200 (200-½)	
S-LOC 27	1440/24	288	(300-½)	1440/50 288 (300-1)
CIRCLING	1540-1 361 (400-1)	1640-1 461 (500-1)	1640-1½ 461 (500-1½)	1740-2 561 (600-2)

DA
Visibility (RVR 100s of feet)
Aircraft Approach Category
HAT
Straight-in ILS to Runway 27
Straight-in with Glide Slope Inoperative or not used to Runway 27
All minima in parentheses not applicable to Civil Pilots.
Military Pilots refer to appropriate regulations.
MDA
HAA
Visibility in Statute Miles

TERMS/LANDING MINIMA DATA

03 AUG 2006 to 31 AUG 2006

05300

TERMS/LANDING MINIMA DATA

COPTER MINIMA ONLY

CATEGORY	COPTER
H-176°	680-½ 363 (400-½)

Copter Approach Direction
Height of MDA/DA Above Landing Area (HAL)

No circling minimums are provided

RADAR MINIMA

PAR (c)	10 2.5°/42/1000	ABCDE	**195**/16	100	(100-¼)				
(d)	28 2.5°/48/1068	ABCDE	**187**/16	100	(100-¼)				
ASR	10	ABC	**560**/40	463	(500-¾)	D	**560**/50	463	(500-1)
		E	**580**/60	463	(500-1¼)				
	28	AB	**600**/50	513	(600-1)	C	**600**/60	513	(600-1¼)
		DE	**600**-1½	513	(600-1½)				
CIR (b)	10	AB	**560**-1¼	463	(500-1¼)	C	**560**-1½	463	(500-1½)
	28	AB	**600**-1¼	503	(600-1¼)	C	**600**-1½	503	(600-1½)
	10, 28	DE	**660**-2	563	(600-2)				

Visibility (RVR 100s of feet)
All minima step in parentheses not applicable to Civil Pilots. Military Pilots refer to appropriate regulations.
Visibility in Statute Miles

Radar Minima:
1. Minima shown are the lowest permitted by established criteria. Pilots should consult applicable directives for their category of aircraft.
2. The circling MDA and weather minima to be used are those for the runway to which the final approach is flown - not the landing runway. In the above RADAR MINIMA example, a category C aircraft flying a radar approach to runway 10, circling to land on runway 28, must use an MDA of 560 feet with weather minima of 500-1½.

A Alternate Minimums not standard. Civil users refer to tabulation. USA/USN/USAF pilots refer to appropriate regulations.

A NA Alternate minimums are Not Authorized due to unmonitored facility or absence of weather reporting service.

T Take-off Minimums not standard and/or Departure Procedures are published. Refer to tabulation.

AIRCRAFT APPROACH CATEGORIES

Aircraft approach category indicates a grouping of aircraft based on a speed of VREF, if specified, or if VREF not specified, 1.3 VSO at the maximum certificated landing weight. VREF, VSO, and the maximum certificated landing weight are those values as established for the aircraft by the certification authority of the country of registry. Helicopters are Category A aircraft. An aircraft shall fit in only one category. However, if it is necessary to operate at a speed in excess of the upper limit of the speed range for an aircraft's category, the minimums for the category for that speed shall be used. For example, an airplane which fits into Category B, but is circling to land at a speed of 145 knots, shall use the approach Category D minimums. As an additional example, a Category A airplane (or helicopter) which is operating at 130 knots on a straight-in approach shall use the approach Category C minimums. See following category limits:

MANEUVERING TABLE

Approach Category	A	B	C	D	E
Speed (Knots)	0-90	91-120	121-140	141-165	Abv 165

Comparable Values of RVR and Visibility

The following table shall be used for converting RVR to ground or flight visibility. For converting RVR values that fall between listed values, use the next higher RVR value; do not interpolate. For example, when converting 1800 RVR, use 2400 RVR with the resultant visibility of 1/2 mile.

RVR (feet)	Visibility (statute miles)	RVR (feet)	Visibility (statute miles)
1600	¼	4500	⅞
2400	½	5000	1
3200	⅝	6000	1¼
4000	¾		

TERMS/LANDING MINIMA DATA

03 AUG 2006 to 31 AUG 2006

03 AUG 2006 to 31 AUG 2006

Figure 8-20. *Terms/Landing Minima Data.*

Precision approaches use decision height (DH), which is referenced to the height above threshold elevation (HAT). Nonprecision approaches use MDA, referenced to "feet MSL." The MDA is also referenced to HAT for straight-in approaches, or height above airport (HAA) for circling approaches. On NACG charts, the figures listed parenthetically are for military operations and are not used in civil aviation.

Visibility figures are provided in statute miles or runway visual range (RVR), which is reported in hundreds of feet. RVR is measured by a transmissometer, which represents the horizontal distance measured at points along the runway. It is based on the sighting of either high intensity runway lights or on the visual contrast of other targets, whichever yields the greater visual range. RVR is horizontal visual range, not slant visual range, and is used in lieu of prevailing visibility in determining minimums for a particular runway. It is illustrated in hundreds of feet if less than a mile (i.e., "24" is an RVR of 2,400 feet). *[Figures 8-19* and *8-20]*

Visibility figures are depicted after the DA/DH or MDA in the minimums section. If visibility in statute miles is indicated, an altitude number, hyphen, and a whole or fractional number appear; for example, 530-1, which indicates "530 feet MSL' and 1 statute mile visibility. This is the descent minimum for the approach. The RVR value is separated from the minimum altitude with a slash, such as "1065/24," which indicates 1,065 feet MSL and an RVR of 2,400 feet. If RVR is prescribed for the procedure, but not available, a conversion table is used to provide the equivalent visibility in this case, of 1/2 statute mile visibility. *[Figure 8-20]* The conversion table is also available in the TPP.

When an alternate airport is required, standard IFR alternate minimums apply. For aircraft other than helicopters, precision approach procedures require a 600-feet ceiling and two statute miles visibility; nonprecision approaches require an 800-feet ceiling and two statute miles visibility. Helicopter alternate minimums are a ceiling that is 200 feet above the minimum for the approach to be flown and visibility of at least one statute mile, but not less than the minimum visibility for the approach to be flown. When a black triangle with a white "A" appears in the notes section of the pilot briefing, it indicates non-standard IFR alternate minimums exist for the airport. If an "NA" appears after the "A," ▲NA then alternate minimums are not authorized. This information is found in the beginning of the TPP.

In addition to the COPTER approaches, instrument-equipped helicopters may fly standard approach procedures. The required visibility minimum may be reduced to one-half the published visibility minimum for category A aircraft, but in no case may it be reduced to less than 1/4 mile or 1,200 feet RVR.

Two terms are specific to helicopters. Height above landing (HAL) means height above a designated helicopter landing area used for helicopter IAPs. "Point in space approach" refers to a helicopter IAP to a MAP more than 2,600 feet from an associated helicopter landing area.

Airport Sketch /Airport Diagram

The airport sketch, located on the bottom right side of the chart, includes many helpful features. IAPs for some of the larger airports devote an entire page to an airport diagram. Airport sketch information concerning runway orientation, lighting, final approach bearings, airport beacon, and obstacles all serve to guide the pilot in the final phases of flight. See *Figure 8-21* for a legend of airport diagram/airport sketch features (see also *Figure 8-10* for an example of an airport diagram).

The airport elevation is indicated in a separate box at the top left of the airport sketch. The touchdown zone elevation (TDZE), which is the highest elevation within the first 3,000 feet of the runway, is designated at the approach end of the procedure's runway.

Beneath the airport sketch is a time and speed table when applicable. The table provides the distance and the amount of time required to transit the distance from the FAF to the MAP for selected groundspeeds.

The approach lighting systems and the visual approach lights are depicted on the airport sketch. White on black symbols 🅥 are used for identifying pilot-controlled lighting (PCL). Runway lighting aids are also noted (e.g., REIL, HIRL), as is the runway centerline lighting (RCL). *[Figure 8-22]*

The airport diagram shows the paved runway configuration in solid black, while the taxiways and aprons are shaded gray. Other runway environment features are shown, such as the runway identification, dimensions, magnetic heading, displaced threshold, arresting gear, usable length, and slope.

Inoperative Components

Certain procedures can be flown with inoperative components. According to the Inoperative Components Table, for example, an ILS approach with a malfunctioning Medium Intensity Approach Lighting System with Runway Alignment Indicator Lights (MALSR = MALS with RAIL) can be flown if the minimum visibility is increased by 1/4 mile. *[Figure 8-23]* A note in this section might read, "Inoperative Table does not apply to ALS or HIRL Runway 13L."

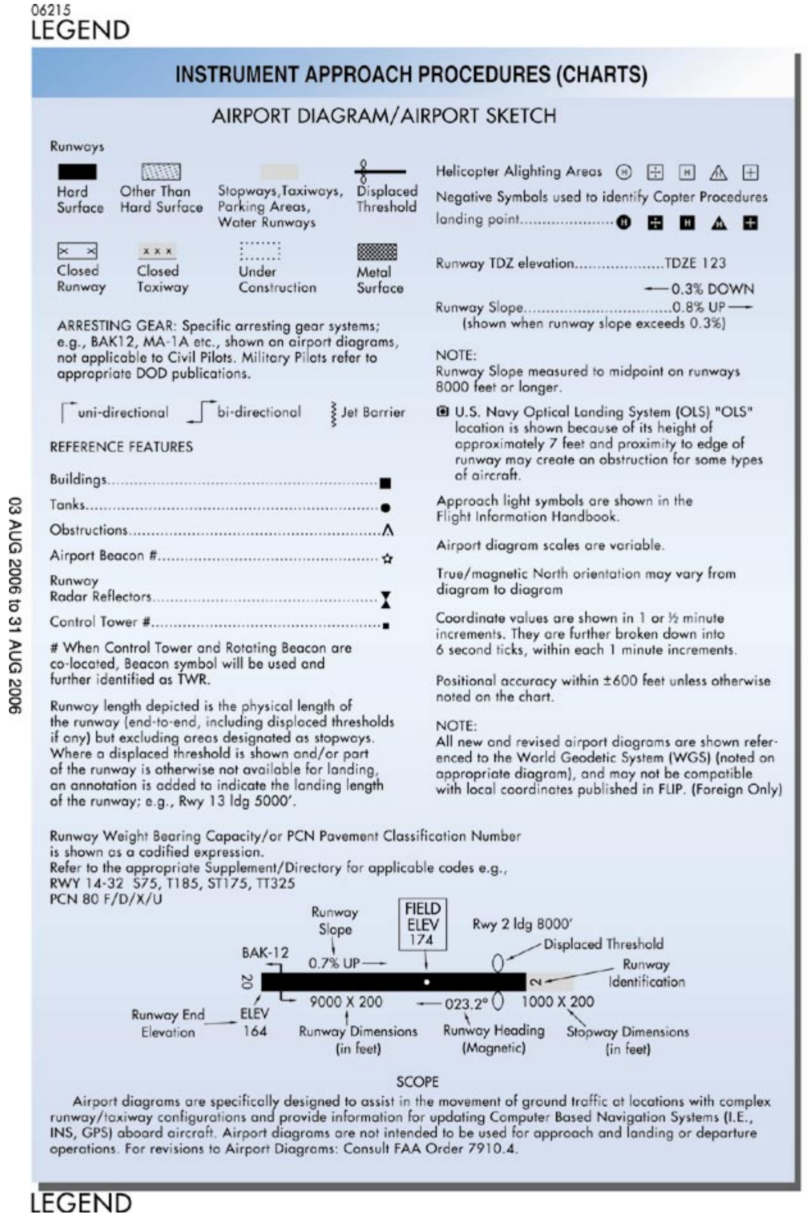

06215
LEGEND

03 AUG 2006 to 31 AUG 2006

INSTRUMENT APPROACH PROCEDURES (CHARTS)

AIRPORT DIAGRAM/AIRPORT SKETCH

Runways

■ Hard Surface	▨ Other Than Hard Surface	▭ Stopways, Taxiways, Parking Areas, Water Runways	Displaced Threshold
☒ Closed Runway	x x x Closed Taxiway	⬚ Under Construction	▩ Metal Surface

ARRESTING GEAR: Specific arresting gear systems; e.g., BAK12, MA-1A etc., shown on airport diagrams, not applicable to Civil Pilots. Military Pilots refer to appropriate DOD publications.

uni-directional bi-directional Jet Barrier

REFERENCE FEATURES

Buildings...■
Tanks...●
Obstructions...Λ
Airport Beacon #..☆
Runway Radar Reflectors.....................................⧗
Control Tower #..▪

When Control Tower and Rotating Beacon are co-located, Beacon symbol will be used and further identified as TWR.

Runway length depicted is the physical length of the runway (end-to-end, including displaced thresholds if any) but excluding areas designated as stopways. Where a displaced threshold is shown and/or part of the runway is otherwise not available for landing, an annotation is added to indicate the landing length of the runway; e.g., Rwy 13 ldg 5000′.

Helicopter Alighting Areas Ⓗ ⊞ ▣ ⚠ ⊞

Negative Symbols used to identify Copter Procedures landing point.......................● ⊞ ■ ▲ ⊞

Runway TDZ elevation....................TDZE 123

Runway Slope..................................0.8% UP ⟶ ⟵ 0.3% DOWN
(shown when runway slope exceeds 0.3%)

NOTE:
Runway Slope measured to midpoint on runways 8000 feet or longer.

U.S. Navy Optical Landing System (OLS) "OLS" location is shown because of its height of approximately 7 feet and proximity to edge of runway may create an obstruction for some types of aircraft.

Approach light symbols are shown in the Flight Information Handbook.

Airport diagram scales are variable.

True/magnetic North orientation may vary from diagram to diagram

Coordinate values are shown in 1 or ½ minute increments. They are further broken down into 6 second ticks, within each 1 minute increments.

Positional accuracy within ±600 feet unless otherwise noted on the chart.

NOTE:
All new and revised airport diagrams are shown referenced to the World Geodetic System (WGS) (noted on appropriate diagram), and may not be compatible with local coordinates published in FLIP. (Foreign Only)

Runway Weight Bearing Capacity/or PCN Pavement Classification Number is shown as a codified expression.
Refer to the appropriate Supplement/Directory for applicable codes e.g.,
RWY 14-32 S75, T185, ST175, TT325
PCN 80 F/D/X/U

Runway Slope
FIELD ELEV 174
Rwy 2 ldg 8000′
BAK-12
0.7% UP
Displaced Threshold
20
2
Runway Identification
Runway End Elevation
ELEV 164
9000 X 200
023.2°
1000 X 200
Runway Dimensions (in feet)
Runway Heading (Magnetic)
Stopway Dimensions (in feet)

SCOPE

Airport diagrams are specifically designed to assist in the movement of ground traffic at locations with complex runway/taxiway configurations and provide information for updating Computer Based Navigation Systems (I.E., INS, GPS) aboard aircraft. Airport diagrams are not intended to be used for approach and landing or departure operations. For revisions to Airport Diagrams: Consult FAA Order 7910.4.

03 AUG 2006 to 31 AUG 2006

LEGEND

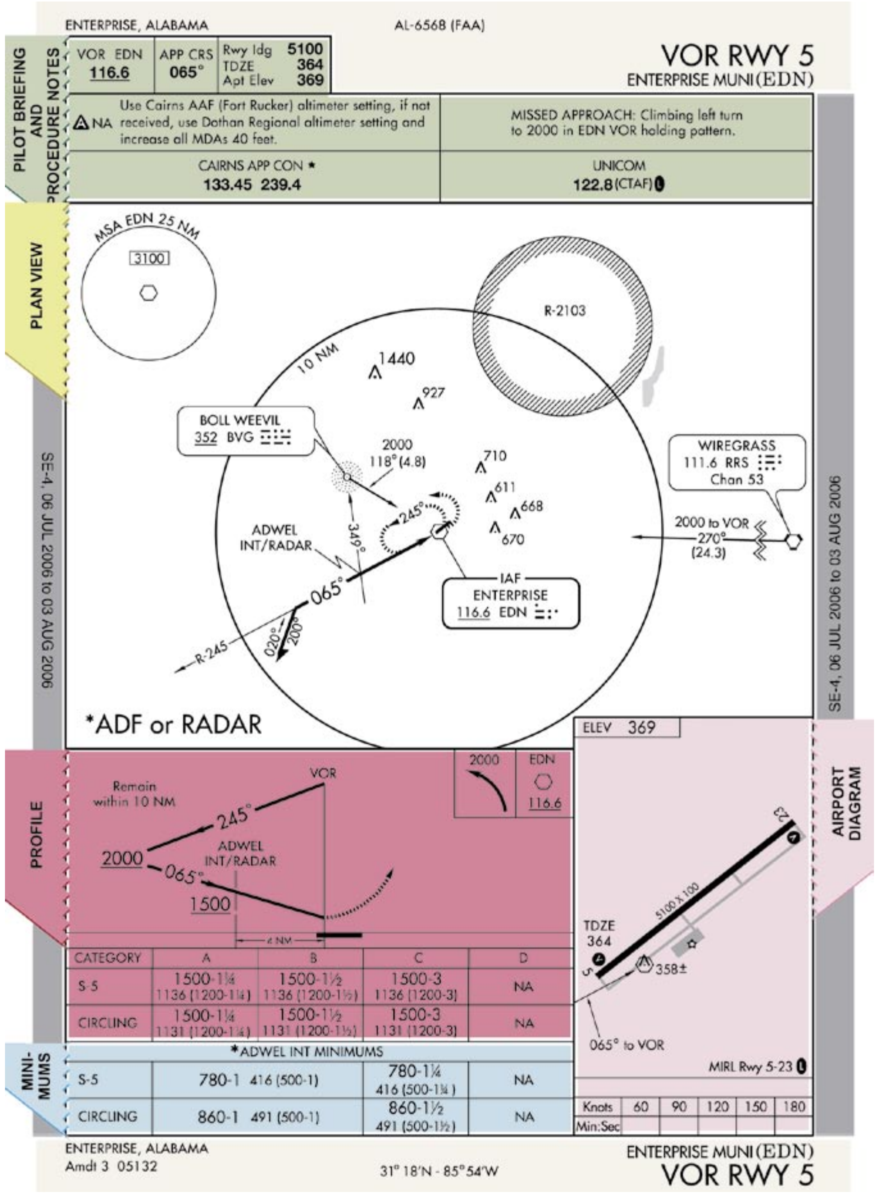

ENTERPRISE, ALABAMA AL-656B (FAA)

VOR RWY 5
ENTERPRISE MUNI (EDN)

PILOT BRIEFING AND PROCEDURE NOTES

VOR EDN	APP CRS	Rwy ldg	5100
116.6	065°	TDZE	364
		Apt Elev	369

⚠NA Use Cairns AAF (Fort Rucker) altimeter setting, if not received, use Dothan Regional altimeter setting and increase all MDAs 40 feet.

MISSED APPROACH: Climbing left turn to 2000 in EDN VOR holding pattern.

CAIRNS APP CON ★
133.45 239.4

UNICOM
122.8 (CTAF) L

PLAN VIEW

SE-4, 06 JUL 2006 to 03 AUG 2006

MSA EDN 25 NM
3100
R-2103
10 NM
1440
927
BOLL WEEVIL
352 BVG
2000
118° (4.8)
710
611
668
670
WIREGRASS
111.6 RRS
Chan 53
2000 to VOR
270°
(24.3)
ADWEL
INT/RADAR
349°
245°
065°
IAF
ENTERPRISE
116.6 EDN
R-245
020°
200°

*ADF or RADAR

PROFILE

Remain within 10 NM
VOR
2000
245°
ADWEL
INT/RADAR
065°
1500
4 NM
2000
EDN
116.6

MINIMUMS

CATEGORY	A	B	C	D
S-5	1500-1¼ 1136 (1200-1¼)	1500-1½ 1136 (1200-1½)	1500-3 1136 (1200-3)	NA
CIRCLING	1500-1¼ 1131 (1200-1¼)	1500-1½ 1131 (1200-1½)	1500-3 1131 (1200-3)	NA

*ADWEL INT MINIMUMS

	A	B	C	D
S-5	780-1 416 (500-1)		780-1¼ 416 (500-1¼)	NA
CIRCLING	860-1 491 (500-1)		860-1½ 491 (500-1½)	NA

AIRPORT DIAGRAM

ELEV 369
23
5100 X 100
TDZE 364
5
358±
065° to VOR
MIRL Rwy 5-23 L

Knots	60	90	120	150	180
Min:Sec					

SE-4, 06 JUL 2006 to 03 AUG 2006

ENTERPRISE, ALABAMA
Amdt 3 05132

31° 18′N - 85° 54′W

ENTERPRISE MUNI (EDN)
VOR RWY 5

Figure 8-21. *Airport Legend and Diagram.*

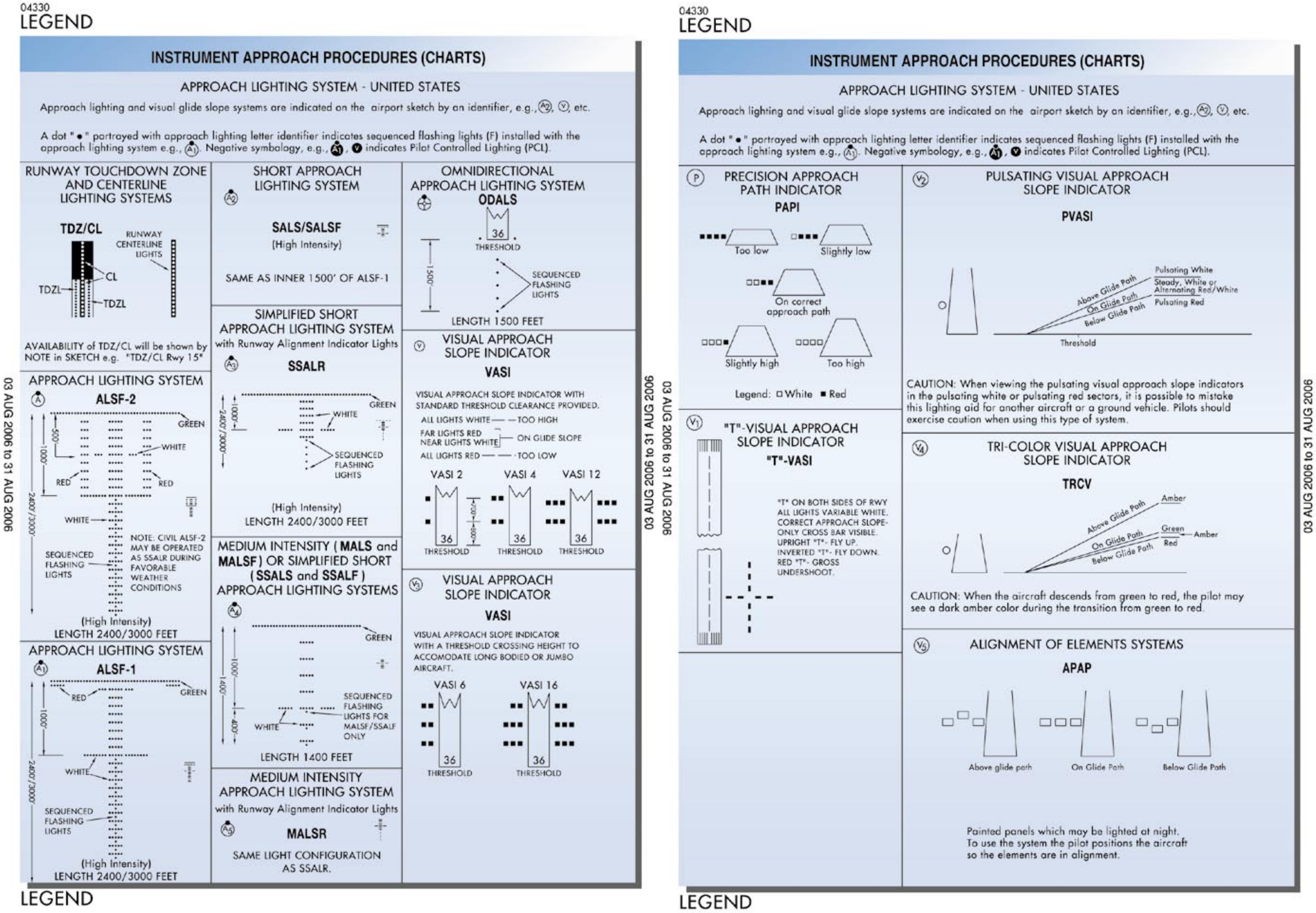

04330
LEGEND

INSTRUMENT APPROACH PROCEDURES (CHARTS)

APPROACH LIGHTING SYSTEM - UNITED STATES

Approach lighting and visual glide slope systems are indicated on the airport sketch by an identifier, e.g., Ⓐ2, Ⓥ, etc.

A dot "•" portrayed with approach lighting letter identifier indicates sequenced flashing lights (F) installed with the approach lighting system e.g., Ⓐ1. Negative symbology, e.g., Ⓐ1, Ⓥ indicates Pilot Controlled Lighting (PCL).

RUNWAY TOUCHDOWN ZONE AND CENTERLINE LIGHTING SYSTEMS

TDZ/CL

RUNWAY CENTERLINE LIGHTS
CL
TDZL
TDZL

AVAILABILITY of TDZ/CL will be shown by NOTE in SKETCH e.g. "TDZ/CL Rwy 15"

APPROACH LIGHTING SYSTEM

Ⓐ ALSF-2

500'
1000'
2400'/3000'
GREEN
WHITE
RED
RED
WHITE
SEQUENCED FLASHING LIGHTS

NOTE: CIVIL ALSF-2 MAY BE OPERATED AS SSALR DURING FAVORABLE WEATHER CONDITIONS

(High Intensity)
LENGTH 2400/3000 FEET

APPROACH LIGHTING SYSTEM

Ⓐ1 ALSF-1

1000'
2400'/3000'
RED
GREEN
WHITE
SEQUENCED FLASHING LIGHTS

(High Intensity)
LENGTH 2400/3000 FEET

SHORT APPROACH LIGHTING SYSTEM

Ⓐ2 SALS/SALSF
(High Intensity)

SAME AS INNER 1500' OF ALSF-1

SIMPLIFIED SHORT APPROACH LIGHTING SYSTEM
with Runway Alignment Indicator Lights

Ⓐ3 SSALR

1000'
2400'/3000'
GREEN
WHITE
SEQUENCED FLASHING LIGHTS

(High Intensity)
LENGTH 2400/3000 FEET

MEDIUM INTENSITY (**MALS** and **MALSF**) OR SIMPLIFIED SHORT (**SSALS** and **SSALF**) APPROACH LIGHTING SYSTEMS

Ⓐ4

1000'
1400'
400'
GREEN
WHITE
SEQUENCED FLASHING LIGHTS FOR MALSF/SSALF ONLY

LENGTH 1400 FEET

MEDIUM INTENSITY APPROACH LIGHTING SYSTEM
with Runway Alignment Indicator Lights

Ⓐ5 MALSR

SAME LIGHT CONFIGURATION AS SSALR.

OMNIDIRECTIONAL APPROACH LIGHTING SYSTEM
ODALS

36
THRESHOLD
1500'
SEQUENCED FLASHING LIGHTS

LENGTH 1500 FEET

Ⓥ VISUAL APPROACH SLOPE INDICATOR
VASI

VISUAL APPROACH SLOPE INDICATOR WITH STANDARD THRESHOLD CLEARANCE PROVIDED.

ALL LIGHTS WHITE — TOO HIGH
FAR LIGHTS RED
NEAR LIGHTS WHITE — ON GLIDE SLOPE
ALL LIGHTS RED — TOO LOW

VASI 2
VASI 4
VASI 12
700'
800'
36
THRESHOLD
36
THRESHOLD
36
THRESHOLD

Ⓥ3 VISUAL APPROACH SLOPE INDICATOR
VASI

VISUAL APPROACH SLOPE INDICATOR WITH A THRESHOLD CROSSING HEIGHT TO ACCOMODATE LONG BODIED OR JUMBO AIRCRAFT.

VASI 6
VASI 16
36
THRESHOLD
36
THRESHOLD

LEGEND

03 AUG 2006 to 31 AUG 2006

04330
LEGEND

INSTRUMENT APPROACH PROCEDURES (CHARTS)

APPROACH LIGHTING SYSTEM - UNITED STATES

Approach lighting and visual glide slope systems are indicated on the airport sketch by an identifier, e.g., Ⓐ2, Ⓥ, etc.

A dot "•" portrayed with approach lighting letter identifier indicates sequenced flashing lights (F) installed with the approach lighting system e.g., Ⓐ1. Negative symbology, e.g., Ⓐ1, Ⓥ indicates Pilot Controlled Lighting (PCL).

Ⓟ PRECISION APPROACH PATH INDICATOR
PAPI

■■■■ Too low
□■■■ Slightly low
□□■■ On correct approach path
□□□■ Slightly high
□□□□ Too high

Legend: □ White ■ Red

Ⓥ1 "T"-VISUAL APPROACH SLOPE INDICATOR
"T"-VASI

"T" ON BOTH SIDES OF RWY
ALL LIGHTS VARIABLE WHITE.
CORRECT APPROACH SLOPE-ONLY CROSS BAR VISIBLE.
UPRIGHT "T"- FLY UP.
INVERTED "T"- FLY DOWN.
RED "T"- GROSS UNDERSHOOT.

Ⓥ2 PULSATING VISUAL APPROACH SLOPE INDICATOR
PVASI

Above Glide Path
On Glide Path
Below Glide Path
Pulsating White
Steady, White or Alternating Red/White
Pulsating Red
Threshold

CAUTION: When viewing the pulsating visual approach slope indicators in the pulsating white or pulsating red sectors, it is possible to mistake this lighting aid for another aircraft or a ground vehicle. Pilots should exercise caution when using this type of system.

Ⓥ4 TRI-COLOR VISUAL APPROACH SLOPE INDICATOR
TRCV

Above Glide Path
On Glide Path
Below Glide Path
Amber
Green
Amber
Red

CAUTION: When the aircraft descends from green to red, the pilot may see a dark amber color during the transition from green to red.

Ⓥ5 ALIGNMENT OF ELEMENTS SYSTEMS
APAP

Above glide path
On Glide Path
Below Glide Path

Painted panels which may be lighted at night.
To use the system the pilot positions the aircraft so the elements are in alignment.

LEGEND

03 AUG 2006 to 31 AUG 2006

Figure 8-22. *Approach Lighting Legend.*

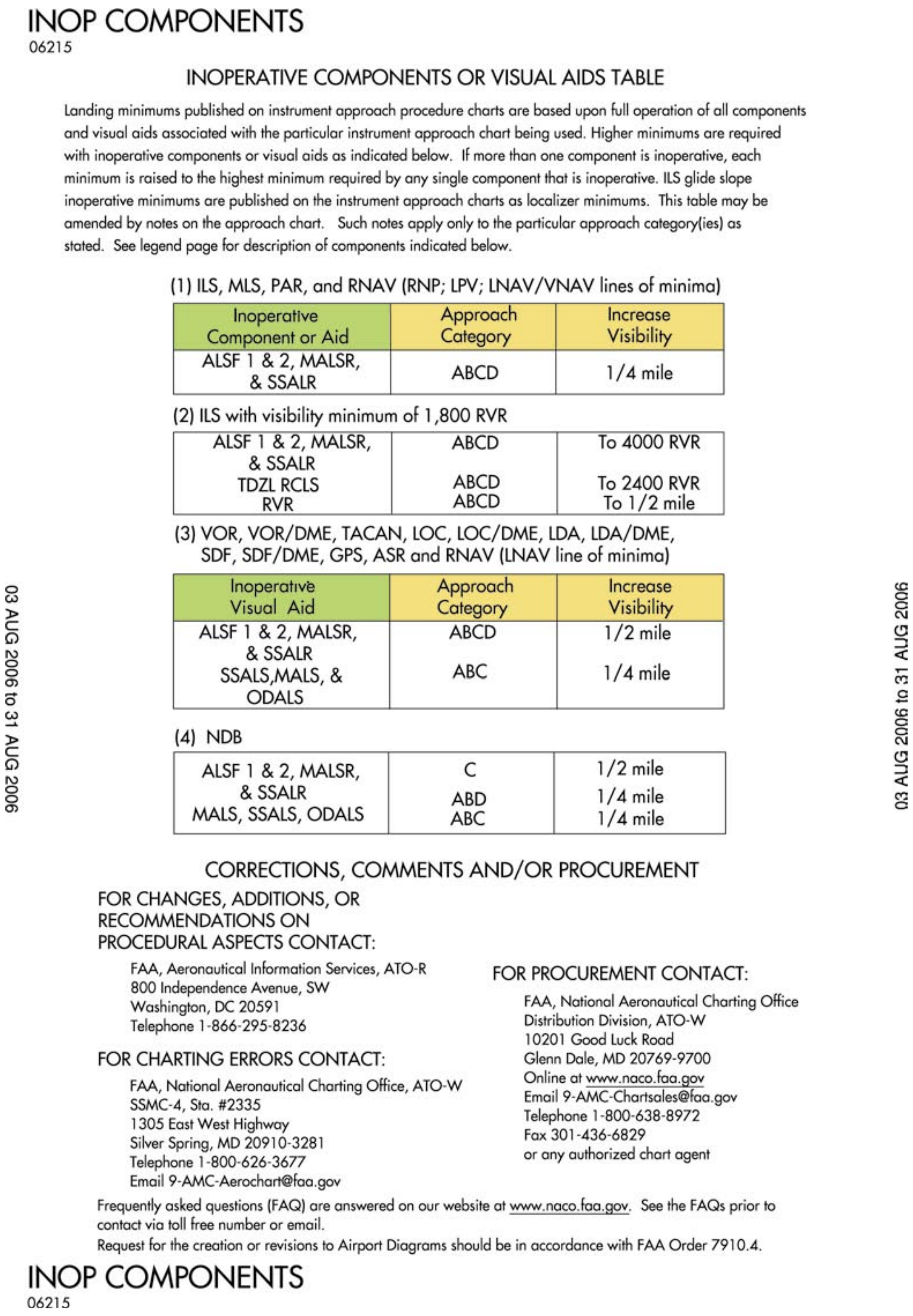

INOP COMPONENTS

06215

INOPERATIVE COMPONENTS OR VISUAL AIDS TABLE

Landing minimums published on instrument approach procedure charts are based upon full operation of all components and visual aids associated with the particular instrument approach chart being used. Higher minimums are required with inoperative components or visual aids as indicated below. If more than one component is inoperative, each minimum is raised to the highest minimum required by any single component that is inoperative. ILS glide slope inoperative minimums are published on the instrument approach charts as localizer minimums. This table may be amended by notes on the approach chart. Such notes apply only to the particular approach category(ies) as stated. See legend page for description of components indicated below.

(1) ILS, MLS, PAR, and RNAV (RNP; LPV; LNAV/VNAV lines of minima)

Inoperative Component or Aid	Approach Category	Increase Visibility
ALSF 1 & 2, MALSR, & SSALR	ABCD	1/4 mile

(2) ILS with visibility minimum of 1,800 RVR

ALSF 1 & 2, MALSR, & SSALR	ABCD	To 4000 RVR
TDZL RCLS	ABCD	To 2400 RVR
RVR	ABCD	To 1/2 mile

(3) VOR, VOR/DME, TACAN, LOC, LOC/DME, LDA, LDA/DME, SDF, SDF/DME, GPS, ASR and RNAV (LNAV line of minima)

Inoperative Visual Aid	Approach Category	Increase Visibility
ALSF 1 & 2, MALSR, & SSALR	ABCD	1/2 mile
SSALS,MALS, & ODALS	ABC	1/4 mile

(4) NDB

ALSF 1 & 2, MALSR, & SSALR	C	1/2 mile
	ABD	1/4 mile
MALS, SSALS, ODALS	ABC	1/4 mile

03 AUG 2006 to 31 AUG 2006

03 AUG 2006 to 31 AUG 2006

CORRECTIONS, COMMENTS AND/OR PROCUREMENT

FOR CHANGES, ADDITIONS, OR RECOMMENDATIONS ON PROCEDURAL ASPECTS CONTACT:

FAA, Aeronautical Information Services, ATO-R
800 Independence Avenue, SW
Washington, DC 20591
Telephone 1-866-295-8236

FOR CHARTING ERRORS CONTACT:

FAA, National Aeronautical Charting Office, ATO-W
SSMC-4, Sta. #2335
1305 East West Highway
Silver Spring, MD 20910-3281
Telephone 1-800-626-3677
Email 9-AMC-Aerochart@faa.gov

FOR PROCUREMENT CONTACT:

FAA, National Aeronautical Charting Office
Distribution Division, ATO-W
10201 Good Luck Road
Glenn Dale, MD 20769-9700
Online at www.naco.faa.gov
Email 9-AMC-Chartsales@faa.gov
Telephone 1-800-638-8972
Fax 301-436-6829
or any authorized chart agent

Frequently asked questions (FAQ) are answered on our website at www.naco.faa.gov. See the FAQs prior to contact via toll free number or email.

Request for the creation or revisions to Airport Diagrams should be in accordance with FAA Order 7910.4.

INOP COMPONENTS

06215

Figure 8-23. *IAP Inoperative Components Table.*

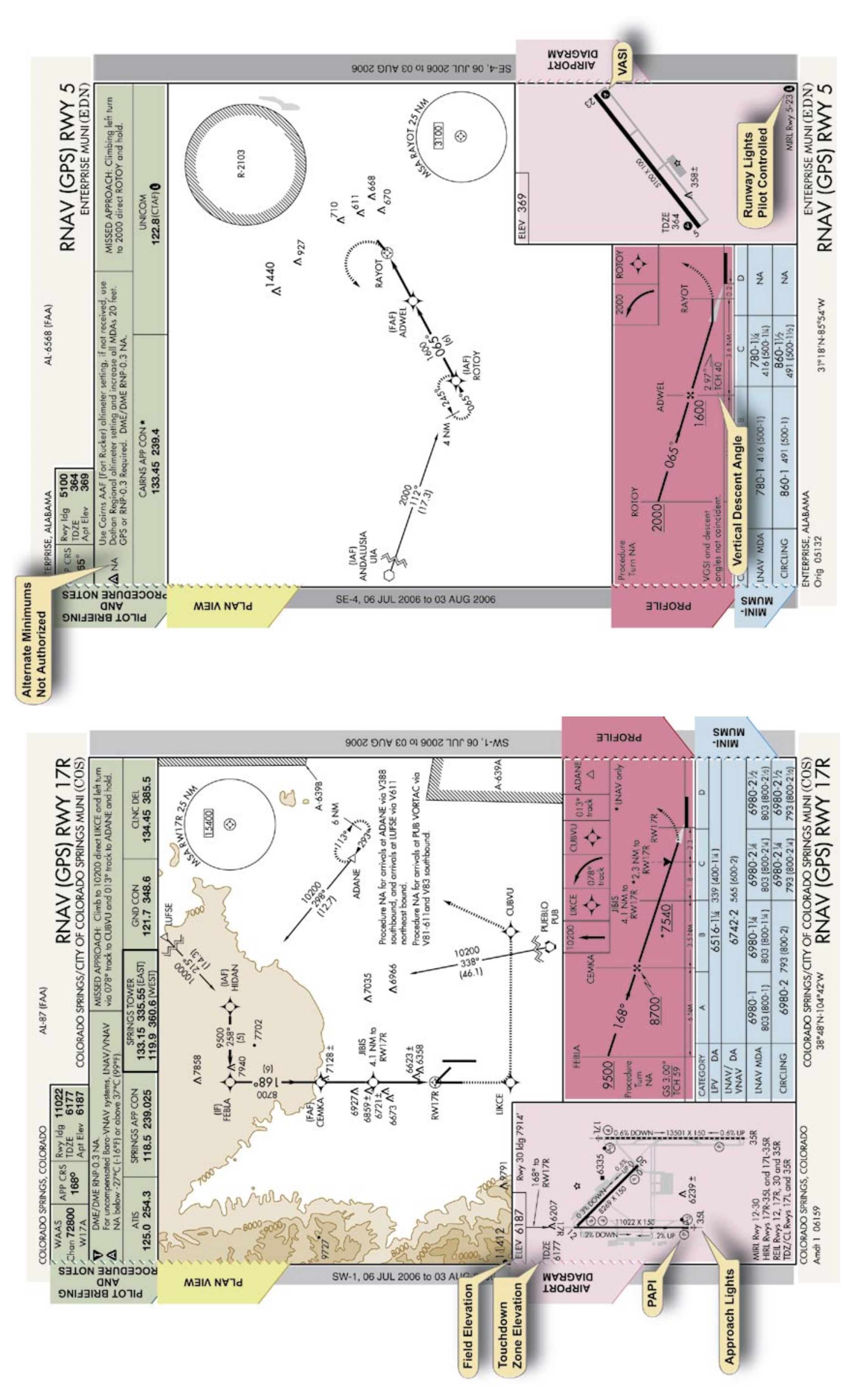
RNAV (GPS) RWY 17R
COLORADO SPRINGS/CITY OF COLORADO SPRINGS MUNI (COS)
COLORADO SPRINGS, COLORADO
AL-87 (FAA)
SW-1, 06 JUL 2006 to 03 AUG 2006
PILOT BRIEFING AND PROCEDURE NOTES
PLAN VIEW
AIRPORT DIAGRAM
PROFILE
MINI-MUMS
ELEV 6187
TDZE 6177
Field Elevation
Touchdown Zone Elevation
PAPI
Approach Lights
RNAV (GPS) RWY 5
ENTERPRISE MUNI (EDN)
ENTERPRISE, ALABAMA
AL-6568 (FAA)
SE-4, 06 JUL 2006 to 03 AUG 2006
ELEV 369
TDZE 364
Alternate Minimums Not Authorized
Vertical Descent Angle
Runway Lights Pilot Controlled
VASI

Figure 8-24. *RNAV Instrument Approach Charts.*

RNAV Instrument Approach Charts

To avoid unnecessary duplication and proliferation of approach charts, approach minimums for unaugmented GPS, Wide Area Augmentation System (WAAS), Local Area Augmentation System (LAAS), will be published on the same approach chart as lateral navigation/vertical navigation (LNAV/VNAV). Other types of equipment may be authorized to conduct the approach based on the minima notes in the front of the TPP approach chart books. Approach charts titled "RNAV RWY XX" may be used by aircraft with navigation systems that meet the required navigational performance (RNP) values for each segment of the approach. *[Figure 8-24]*

The chart may contain as many as four lines of approach minimums: global landing system (GLS), WAAS and LAAS, LNAV/VNAV, LNAV, and circling. LNAV/VNAV is an instrument approach with lateral and vertical guidance with integrity limits similar to barometric vertical navigation (BARO VNAV).

RNAV procedures that incorporate a final approach stepdown fix may be published without vertical navigation on a separate chart also titled RNAV. During a transition period when GPS procedures are undergoing revision to a new title, both RNAV and GPS approach charts and formats will be published. ATC clearance for the RNAV procedure will authorize a properly certificated pilot to utilize any landing minimums for which the aircraft is certified.

Chart terminology will change slightly to support the new procedure types:

1. DA replaces the term DH. DA conforms to the international convention where altitudes relate to MSL and heights relate to AGL. DA will eventually be published for other types of IAPs with vertical guidance, as well. DA indicates to the pilot that the published descent profile is flown to the DA (MSL), where a missed approach will be initiated if visual references for landing are not established. Obstacle clearance is provided to allow a momentary descent below DA while transitioning from the final approach to the missed approach. The aircraft is expected to follow the missed approach instructions while continuing along the published final approach course to at least the published runway threshold waypoint or MAP (if not at the threshold) before executing any turns.
2. MDA will continue to be used only for the LNAV and circling procedures.
3. Threshold crossing height (TCH) has been traditionally used in precision approaches as the height of the GS above threshold. With publication of LNAV/VNAV minimums and RNAV descent angles, including graphically depicted descent profiles, TCH also applies to the height of the "descent angle," or glide path, at the threshold. Unless otherwise required for larger type aircraft, which may be using the IAP, the typical TCH will be 30 to 50 feet.

The minima format changes slightly:

1. Each line of minima on the RNAV IAP will be titled to reflect the RNAV system applicable (e.g., GLS, LNAV/VNAV, and LNAV). Circling minima will also be provided.
2. The minima title box will also indicate the nature of the minimum altitude for the IAP. For example: DA will be published next to the minima line title for minimums supporting vertical guidance, and MDA will be published where the minima line supports only lateral guidance. During an approach where an MDA is used, descent below MDA is not authorized.
3. Where two or more systems share the same minima, each line of minima will be displayed separately.

For more information concerning government charts, the NACG can be contacted by telephone, or via their internet address at:

National Aeronautical Charting Group
Telephone 800-626-3677
http://naco.faa.gov/

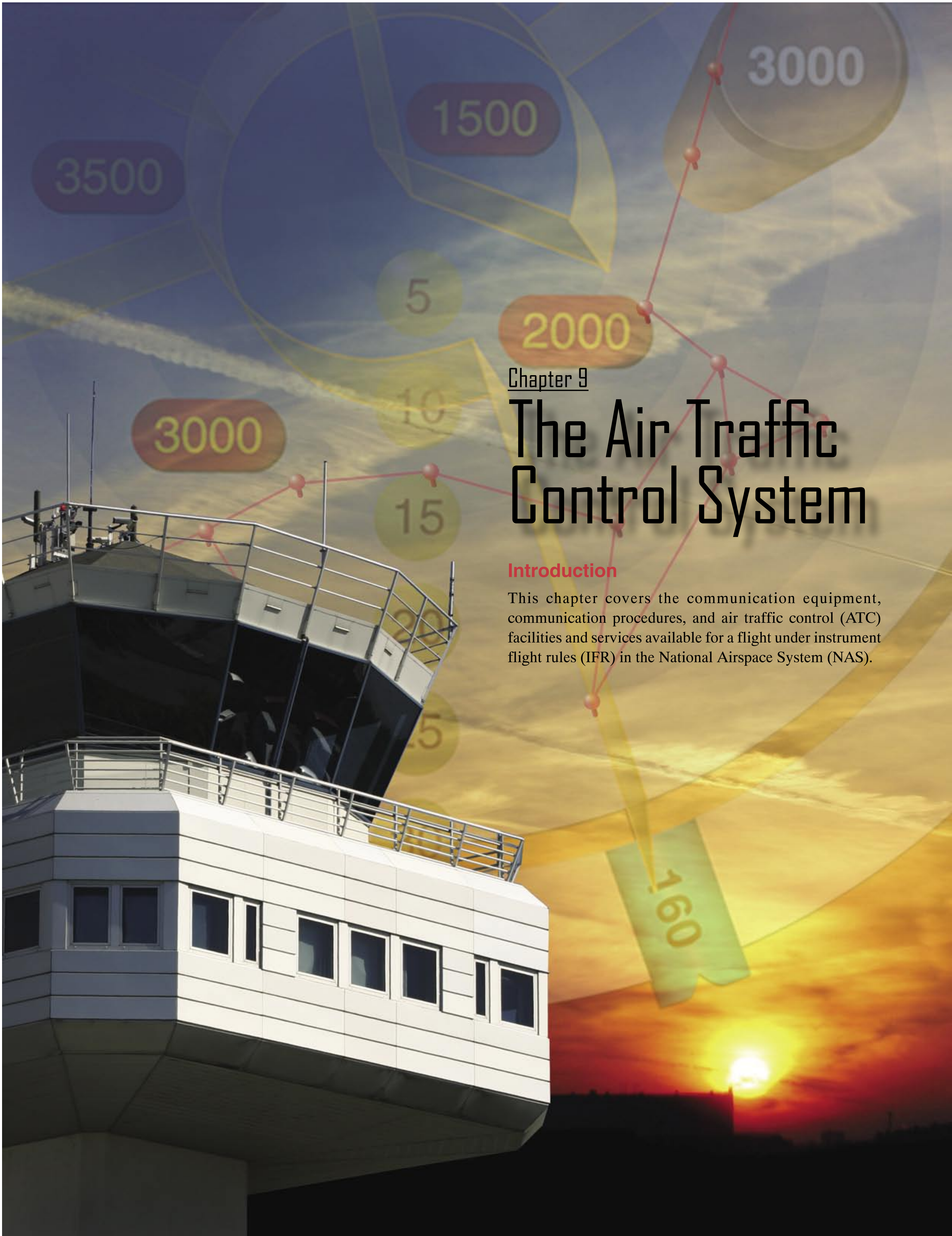
3000
1500
3500
5
2000
10
3000
15
20
25
160

Chapter 9

The Air Traffic Control System

Introduction

This chapter covers the communication equipment, communication procedures, and air traffic control (ATC) facilities and services available for a flight under instrument flight rules (IFR) in the National Airspace System (NAS).

Communication Equipment

Navigation/Communication (NAV/COM) Equipment

Civilian pilots communicate with ATC on frequencies in the very high frequency (VHF) range between 118.000 and 136.975 MHz. To derive full benefit from the ATC system, radios capable of 25 kHz spacing are required (e.g., 134.500, 134.575, 134.600). If ATC assigns a frequency that cannot be selected, ask for an alternative frequency.

Figure 9-1 illustrates a typical radio panel installation, consisting of a communications transceiver on the left and a navigational receiver on the right. Many radios allow the pilot to have one or more frequencies stored in memory and one frequency active for transmitting and receiving (called simplex operation). It is possible to communicate with some automated flight service stations (AFSS) by transmitting on 122.1 MHz (selected on the communication radio) and receiving on a VHF omnidirectional range (VOR) frequency (selected on the navigation radio). This is called duplex operation.

An audio panel allows a pilot to adjust the volume of the selected receiver(s) and to select the desired transmitter. *[Figure 9-2]* The audio panel has two positions for receiver selection, cabin speaker, and headphone (some units might have a center "off" position). Use of a hand-held microphone and the cabin speaker introduces the distraction of reaching for and hanging up the microphone. A headset with a boom microphone is recommended for clear communications. The microphone should be positioned close to the lips to reduce

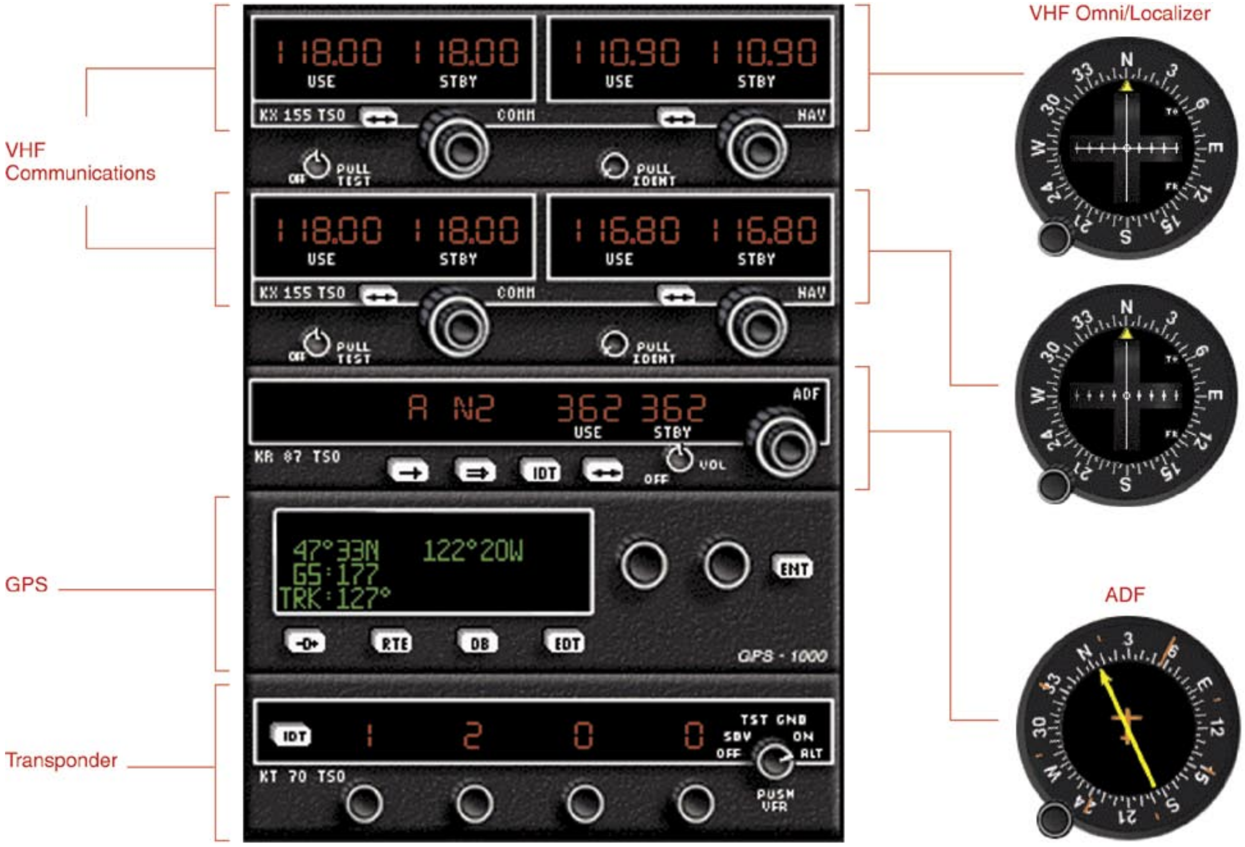
VHF Omni/Localizer
VHF Communications
GPS
ADF
Transponder
118.00 USE 118.00 STBY
KX 155 TSO
COMM
NAV
110.90 USE 110.90 STBY
PULL TEST
PULL IDENT
116.80 USE 116.80 STBY
ANT 362 USE 362 STBY
KR 87 TSO
IDT
VOL
47°33N 122°20W
GS:177
TRK:127°
RTE
DB
EDT
ENT
GPS-1000
KT 70 TSO
1 2 0 0
TST GND SBY ON OFF ALT
PUSH VFR

Figure 9-1. *Typical NAV/COM Installation.*

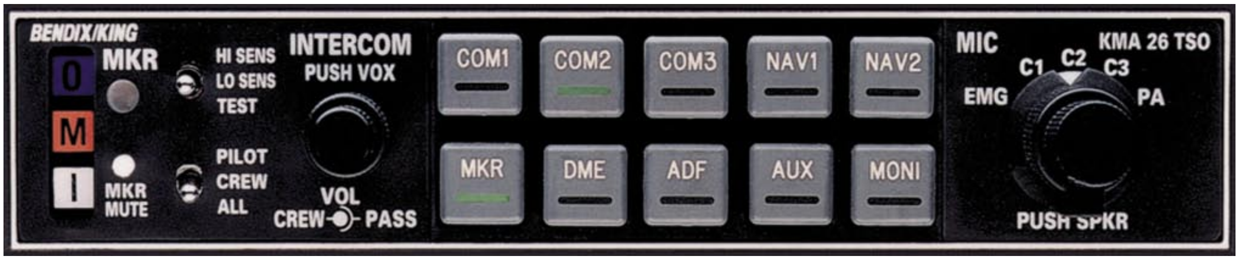
BENDIX/KING
MKR
O
M
I
HI SENS
LO SENS
TEST
INTERCOM
PUSH VOX
PILOT
CREW
ALL
MKR MUTE
VOL
CREW PASS
COM1
COM2
COM3
NAV1
NAV2
MKR
DME
ADF
AUX
MONI
MIC
KMA 26 TSO
C1
C2
C3
EMG
PA
PUSH SPKR

Figure 9-2. *Audio Panel.*

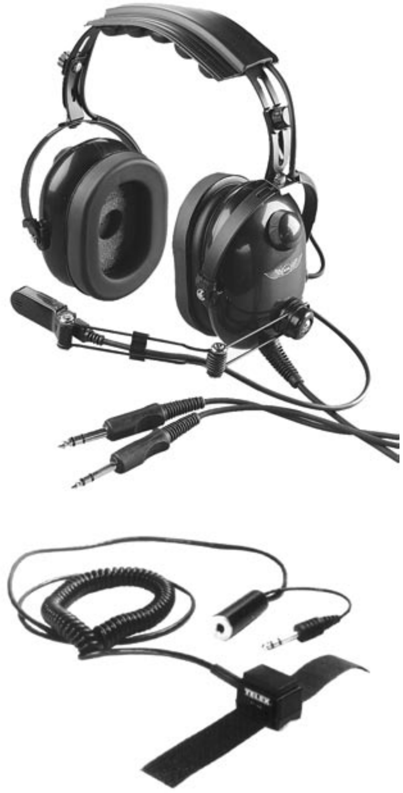

Figure 9-3. *Boom Microphone, Headset, and Push-To-Talk Switch.*

the possibility of ambient flight deck noise interfering with transmissions to the controller. Headphones deliver the received signal directly to the ears; therefore, ambient noise does not interfere with the pilot's ability to understand the transmission. *[Figure 9-3]*

Switching the transmitter selector between COM1 and COM2 changes both transmitter and receiver frequencies. It is necessary only when a pilot wants to monitor one frequency while transmitting on another. One example is listening to automatic terminal information service (ATIS) on one receiver while communicating with ATC on the other. Monitoring a navigation receiver to check for proper identification is another reason to use the switch panel.

Most audio switch panels also include a marker beacon receiver. All marker beacons transmit on 75 MHz, so there is no frequency selector.

Figure 9-4 illustrates an increasingly popular form of NAV/COM radio; it contains a global positioning system (GPS) receiver and a communications transceiver. Using its navigational capability, this unit can determine when a flight crosses an airspace boundary or fix and can automatically select the appropriate communications frequency for that location in the communications radio.

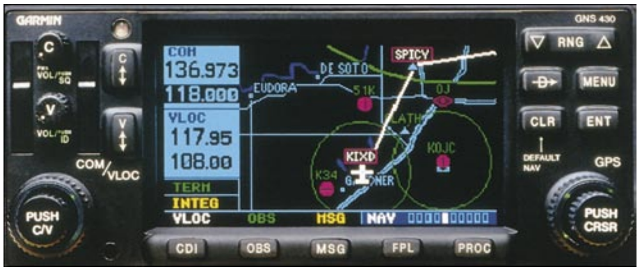

Figure 9-4. *Combination GPS-Com Unit.*

Radar and Transponders

ATC radars have a limited ability to display primary returns, which is energy reflected from an aircraft's metallic structure. Their ability to display secondary returns (transponder replies to ground interrogation signals) makes possible the many advantages of automation.

A transponder is a radar beacon transmitter/receiver installed in the instrument panel. ATC beacon transmitters send out interrogation signals continuously as the radar antenna rotates. When an interrogation is received by a transponder, a coded reply is sent to the ground station where it is displayed on the controller's scope. A reply light on the transponder panel flickers every time it receives and replies to a radar interrogation. Transponder codes are assigned by ATC.

When a controller asks a pilot to "ident" and the ident button is pushed, the return on the controller's scope is intensified for precise identification of a flight. When requested, briefly push the ident button to activate this feature. It is good practice for pilots to verbally confirm that they have changed codes or pushed the ident button.

Mode C (Altitude Reporting)

Primary radar returns indicate only range and bearing from the radar antenna to the target; secondary radar returns can display altitude, Mode C, on the control scope if the aircraft is equipped with an encoding altimeter or blind encoder. In either case, when the transponder's function switch is in the ALT position the aircraft's pressure altitude is sent to the controller. Adjusting the altimeter's Kollsman window has no effect on the altitude read by the controller.

Transponders, when installed, must be ON at all times when operating in controlled airspace; altitude reporting is required by regulation in Class B and Class C airspace and inside a 30-mile circle surrounding the primary airport in Class B airspace. Altitude reporting should also be ON at all times.

Communication Procedures

Clarity in communication is essential for a safe instrument flight. This requires pilots and controllers to use terms that are understood by both—the Pilot/Controller Glossary in the Aeronautical Information Manual (AIM) is the best source of terms and definitions. The AIM is revised twice a year and new definitions are added, so the glossary should be reviewed frequently. Because clearances and instructions are comprised largely of letters and numbers, a phonetic pronunciation guide has been developed for both. *[Figure 9-5]*

ATCs must follow the guidance of the Air Traffic Control Manual when communicating with pilots. The manual presents the controller with different situations and prescribes precise terminology that must be used. This is advantageous for pilots because once they have recognized a pattern or format they can expect future controller transmissions to follow that format. Controllers are faced with a wide variety of communication styles based on pilot experience, proficiency, and professionalism.

Pilots should study the examples in the AIM, listen to other pilots communicate, and apply the lessons learned to their own communications with ATC. Pilots should ask for clarification of a clearance or instruction. If necessary, use plain English to ensure understanding, and expect the controller to reply in the same way. A safe instrument flight is the result of cooperation between controller and pilot.

Communication Facilities

The controller's primary responsibility is separation of aircraft operating under IFR. This is accomplished with ATC facilities which include the AFSS, airport traffic control tower (ATCT), terminal radar approach control (TRACON), and air route traffic control center (ARTCC).

Automated Flight Service Stations (AFSS)

A pilot's first contact with ATC is usually through AFSS, either by radio or telephone. AFSSs provide pilot briefings, receive and process flight plans, relay ATC clearances, originate Notices to Airmen (NOTAMs), and broadcast aviation weather. Some facilities provide En Route Flight Advisory Service (EFAS), take weather observations, and advise United States Customs and Immigration of international flights.

Telephone contact with Flight Service can be obtained by dialing 1-800-WX-BRIEF. This number can be used anywhere in the United States and connects to the nearest AFSS based on the area code from which the call originates. There are a variety of methods of making radio contact: direct transmission, remote communication outlets (RCOs), ground communication outlets (GCOs), and by using duplex transmissions through navigational aids (NAVAIDs). The best source of information on frequency usage is the Airport/Facility Directory (A/FD) and the legend panel on sectional charts also contains contact information.

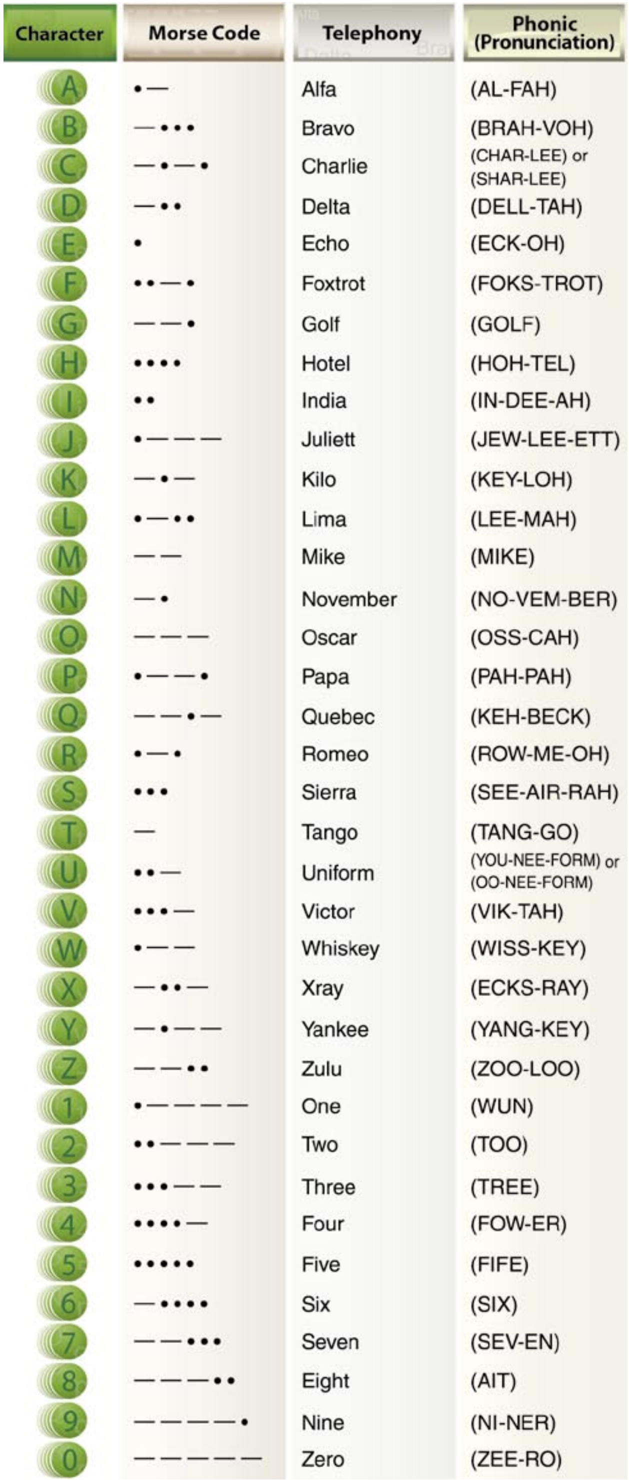

Character	Morse Code	Telephony	Phonic (Pronunciation)
A	•—	Alfa	(AL-FAH)
B	—•••	Bravo	(BRAH-VOH)
C	—•—•	Charlie	(CHAR-LEE) or (SHAR-LEE)
D	—••	Delta	(DELL-TAH)
E	•	Echo	(ECK-OH)
F	••—•	Foxtrot	(FOKS-TROT)
G	——•	Golf	(GOLF)
H	••••	Hotel	(HOH-TEL)
I	••	India	(IN-DEE-AH)
J	•———	Juliett	(JEW-LEE-ETT)
K	—•—	Kilo	(KEY-LOH)
L	•—••	Lima	(LEE-MAH)
M	——	Mike	(MIKE)
N	—•	November	(NO-VEM-BER)
O	———	Oscar	(OSS-CAH)
P	•——•	Papa	(PAH-PAH)
Q	——•—	Quebec	(KEH-BECK)
R	•—•	Romeo	(ROW-ME-OH)
S	•••	Sierra	(SEE-AIR-RAH)
T	—	Tango	(TANG-GO)
U	••—	Uniform	(YOU-NEE-FORM) or (OO-NEE-FORM)
V	•••—	Victor	(VIK-TAH)
W	•——	Whiskey	(WISS-KEY)
X	—••—	Xray	(ECKS-RAY)
Y	—•——	Yankee	(YANG-KEY)
Z	——••	Zulu	(ZOO-LOO)
1	•————	One	(WUN)
2	••———	Two	(TOO)
3	•••——	Three	(TREE)
4	••••—	Four	(FOW-ER)
5	•••••	Five	(FIFE)
6	—••••	Six	(SIX)
7	——•••	Seven	(SEV-EN)
8	———••	Eight	(AIT)
9	————•	Nine	(NI-NER)
0	—————	Zero	(ZEE-RO)

Figure 9-5. *Phonetic Pronunciation Guide.*

The briefer sends a flight plan to the host computer at the ARTCC (Center). After processing the flight plan, the computer will send flight strips to the tower, to the radar facility that will handle the departure route, and to the Center controller whose sector the flight first enters. *Figure 9-6* shows a typical strip. These strips are delivered approximately 30 minutes prior to the proposed departure time. Strips are delivered to en route facilities 30 minutes before the flight is expected to enter their airspace. If a flight plan is not opened, it will "time out" 2 hours after the proposed departure time.

When departing an airport in Class G airspace, a pilot receives an IFR clearance from the AFSS by radio or telephone. It contains either a clearance void time, in which case an aircraft must be airborne prior to that time, or a release time. Pilots should not take-off prior to the release time. Pilots can help the controller by stating how soon they expect to be airborne. If the void time is, for example, 10 minutes past the hour and an aircraft is airborne at exactly 10 minutes past the hour, the clearance is void—a pilot must take off prior to the void time. A specific void time may be requested when filing a flight plan.

ATC Towers

Several controllers in the tower cab are involved in handling an instrument flight. Where there is a dedicated clearance delivery position, that frequency is found in the A/FD and on the instrument approach chart for the departure airport. Where there is no clearance delivery position, the ground controller performs this function. At the busiest airports, pre-taxi clearance is required; the frequency for pre-taxi clearance can be found in the A/FD. Taxi clearance should be requested not more than 10 minutes before proposed taxi time.

It is recommended that pilots read their IFR clearance back to the clearance delivery controller. Instrument clearances can be overwhelming when attempting to copy them verbatim, but they follow a format that allows a pilot to be prepared when responding "Ready to copy." The format is: clearance limit (usually the destination airport); route, including any departure procedure; initial altitude; frequency (for departure control); and transponder code. With the exception of the transponder code, a pilot knows most of these items before engine start. One technique for clearance copying is writing C-R-A-F-T.

Assume an IFR flight plan has been filed from Seattle, Washington to Sacramento, California via V-23 at 7,000 feet. Traffic is taking off to the north from Seattle-Tacoma (Sea-Tac) airport and, by monitoring the clearance delivery frequency, a pilot can determine the departure procedure being assigned to southbound flights. The clearance limit is the destination airport, so write "SAC" after the letter C. Write "SEATTLE TWO – V23" after R for Route, because departure control issued this departure to other flights. Write "7" after the A, the departure control frequency printed on the approach charts for Sea-Tac after F, and leave the space after the letter T blank—the transponder code is generated by computer and can seldom be determined in advance. Then, call clearance delivery and report "Ready to copy."

As the controller reads the clearance, check it against what is already written down; if there is a change, draw a line through that item and write in the changed item. Chances are the changes are minimal, and most of the clearance is copied before keying the microphone. Still, it is worthwhile to develop clearance shorthand to decrease the verbiage that must be copied (see Appendix 1).

Pilots are required to have either the text of a departure procedure (DP) or a graphic representation (if one is available), and should review it before accepting a clearance. This is another reason to find out ahead of time which DP is in use. If the DP includes an altitude or a departure control frequency, those items are not included in the clearance.

The last clearance received supersedes all previous clearances. For example, if the DP says "Climb and maintain 2,000 feet, expect higher in 6 miles," but upon contacting the departure controller a new clearance is received: "Climb and maintain 8,000 feet," the 2,000 feet restriction has been canceled. This rule applies in both terminal and Center airspace.

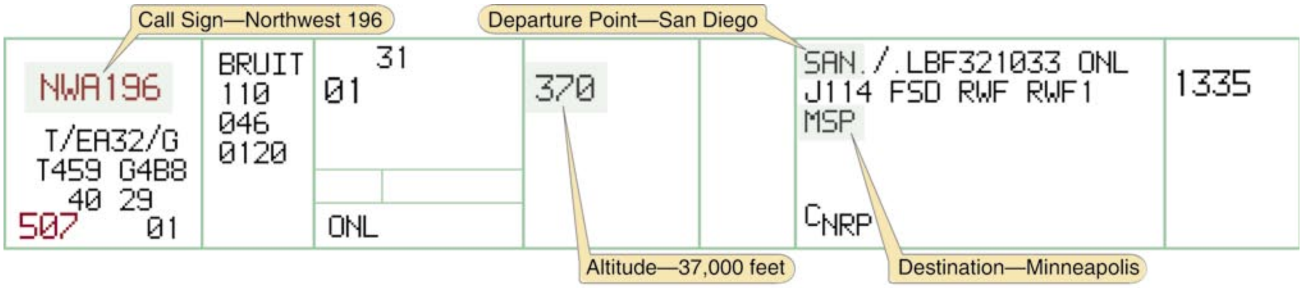

Figure 9-6. *Flight Strip.*

When reporting ready to copy an IFR clearance before the strip has been received from the Center computer, pilots are advised "clearance on request." The controller initiates contact when it has been received. This time can be used for taxi and pre-takeoff checks.

The local controller is responsible for operations in the Class D airspace and on the active runways. At some towers, designated as IFR towers, the local controller has vectoring authority. At visual flight rules (VFR) towers, the local controller accepts inbound IFR flights from the terminal radar facility and cannot provide vectors. The local controller also coordinates flights in the local area with radar controllers. Although Class D airspace normally extends 2,500 feet above field elevation, towers frequently release the top 500 feet to the radar controllers to facilitate overflights. Accordingly, when a flight is vectored over an airport at an altitude that appears to enter the tower controller's airspace, there is no need to contact the tower controller—all coordination is handled by ATC.

The departure radar controller may be in the same building as the control tower, but it is more likely that the departure radar position is remotely located. The tower controller will not issue a takeoff clearance until the departure controller issues a release.

Terminal Radar Approach Control (TRACON)

TRACONs are considered terminal facilities because they provide the link between the departure airport and the en route structure of the NAS. Terminal airspace normally extends 30 nautical miles (NM) from the facility, with a vertical extent of 10,000 feet; however, dimensions vary widely. Class B and Class C airspace dimensions are provided on aeronautical charts. At terminal radar facilities the airspace is divided into sectors, each with one or more controllers, and each sector is assigned a discrete radio frequency. All terminal facilities are approach controls and should be addressed as "Approach" except when directed to do otherwise (e.g., "Contact departure on 120.4").

Terminal radar antennas are located on or adjacent to the airport. *Figure 9-7* shows a typical configuration. Terminal controllers can assign altitudes lower than published procedural altitudes called minimum vectoring altitudes (MVAs). These altitudes are not published or accessible to pilots, but are displayed at the controller's position, as shown in *Figure 9-8*. However, when pilots are assigned an altitude that seems to be too low, they should query the controller before descending.

When a pilot accepts a clearance and reports ready for takeoff, a controller in the tower contacts the TRACON for a release. An aircraft is not cleared for takeoff until the departure controller can fit the flight into the departure flow. A pilot may have to hold for release. When takeoff clearance is received, the departure controller is aware of the flight and is waiting for a call. All of the information the controller needs is on the departure strip or the computer screen there is no need to repeat any portion of the clearance to that controller. Simply establish contact with the facility when instructed to do so by the tower controller. The terminal facility computer picks up the transponder and initiates tracking as soon as it detects the

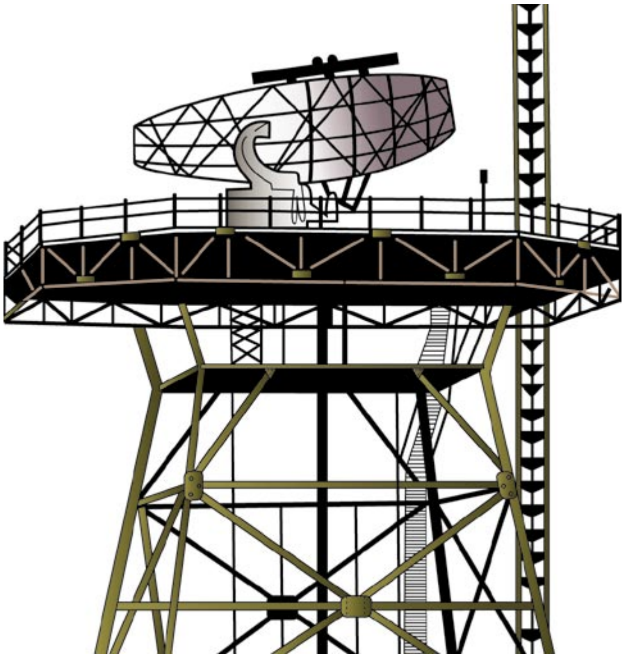

Figure 9-7. *Combined Radar and Beacon Antenna.*

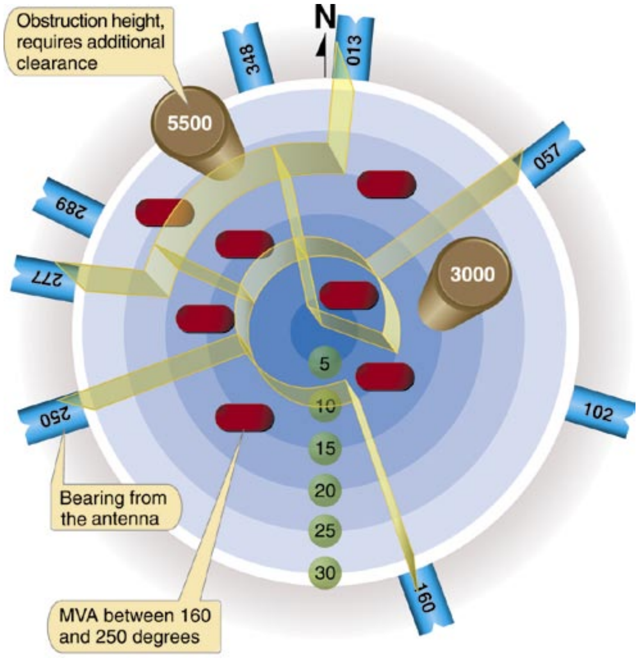
Obstruction height, requires additional clearance
N
348
013
5500
057
289
277
3000
5
10
15
20
25
30
250
102
160
Bearing from the antenna
MVA between 160 and 250 degrees

Figure 9-8. *Minimum Vectoring Altitude Chart.*

assigned code. For this reason, the transponder should remain on standby until takeoff clearance has been received.

The aircraft appears on the controller's radar display as a target with an associated data block that moves as the aircraft moves through the airspace. The data block includes aircraft identification, aircraft type, altitude, and airspeed.

A TRACON controller uses Airport Surveillance Radar (ASR) to detect primary targets and Automated Radar Terminal Systems (ARTS) to receive transponder signals; the two are combined on the controller's scope. *[Figure 9-9]*

At facilities with ASR-3 equipment, radar returns from precipitation are not displayed as varying levels of intensity, and controllers must rely on pilot reports and experience to provide weather avoidance information. With ASR-9 equipment, the controller can select up to six levels of intensity. Light precipitation does not require avoidance tactics but precipitation levels of moderate, heavy or extreme should cause pilots to plan accordingly. Along with precipitation the pilot must additionally consider the temperature, which if between -20° and +5° C will cause icing even during light precipitation. The returns from higher levels of intensity may obscure aircraft data blocks, and controllers may select the higher levels only on pilot request. When uncertainty exists about the weather ahead, ask the controller if the facility can display intensity levels—pilots of small aircraft should avoid intensity levels 3 or higher.

Tower En Route Control (TEC)

At many locations, instrument flights can be conducted entirely in terminal airspace. These TEC routes are generally for aircraft operating below 10,000 feet, and they can be found in the A/FD. Pilots desiring to use TEC should include that designation in the remarks section of the flight plan.

Pilots are not limited to the major airports at the city pairs listed in the A/FD. For example, a tower en route flight from an airport in New York (NYC) airspace could terminate at any airport within approximately 30 miles of Bradley International (BDL) airspace, such as Hartford (HFD). *[Figure 9-10]*

A valuable service provided by the automated radar equipment at terminal radar facilities is the Minimum Safe Altitude Warnings (MSAW). This equipment predicts an aircraft's position in 2 minutes based on present path of flight—the controller issues a safety alert if the projected path encounters terrain or an obstruction. An unusually rapid descent rate on a nonprecision approach can trigger such an alert.

Air Route Traffic Control Center (ARTCC)

ARTCC facilities are responsible for maintaining separation between IFR flights in the en route structure. Center radars (Air Route Surveillance Radar (ARSR)) acquire and track transponder returns using the same basic technology as terminal radars. *[Figure 9-11]*

Earlier Center radars display weather as an area of slashes (light precipitation) and Hs (moderate rainfall), as illustrated in *Figure 9-12*. Because the controller cannot detect higher levels of precipitation, pilots should be wary of areas showing moderate rainfall. Newer radar displays show weather as three levels of blue. Controllers can select the level of weather to be displayed. Weather displays of higher levels of intensity can make it difficult for controllers to see aircraft data blocks, so pilots should not expect ATC to keep weather displayed continuously.

Center airspace is divided into sectors in the same manner as terminal airspace; additionally, most Center airspace is divided by altitudes into high and low sectors. Each sector has a dedicated team of controllers and a selection of radio frequencies, because each Center has a network of remote transmitter/receiver sites. All Center frequencies can be found in the back of the A/FD in the format shown in *Figure 9-13;* they are also found on en route charts.

Each ARTCC's area of responsibility covers several states; when flying from the vicinity of one remote communication site toward another, expect to hear the same controller on different frequencies.

Center Approach/Departure Control

The majority of airports with instrument approaches do not lie within terminal radar airspace, and when operating to or from these airports pilots communicate directly with the Center controller. Departing from a tower-controlled airport, the tower controller provides instructions for contacting the appropriate Center controller. When departing an airport without an operating control tower, the clearance includes instructions such as "Upon entering controlled airspace, contact Houston Center on 126.5." Pilots are responsible for terrain clearance until reaching the controller's MVA. Simply hearing "Radar contact" does not relieve a pilot of this responsibility.

If obstacles in the departure path require a steeper-than-standard climb gradient (200 FPNM), then the controller advises the pilot. However, it is the pilot's responsibility to check the departure airport listing in the A/FD to determine if there are trees or wires in the departure path. When in doubt, ask the controller for the required climb gradient.

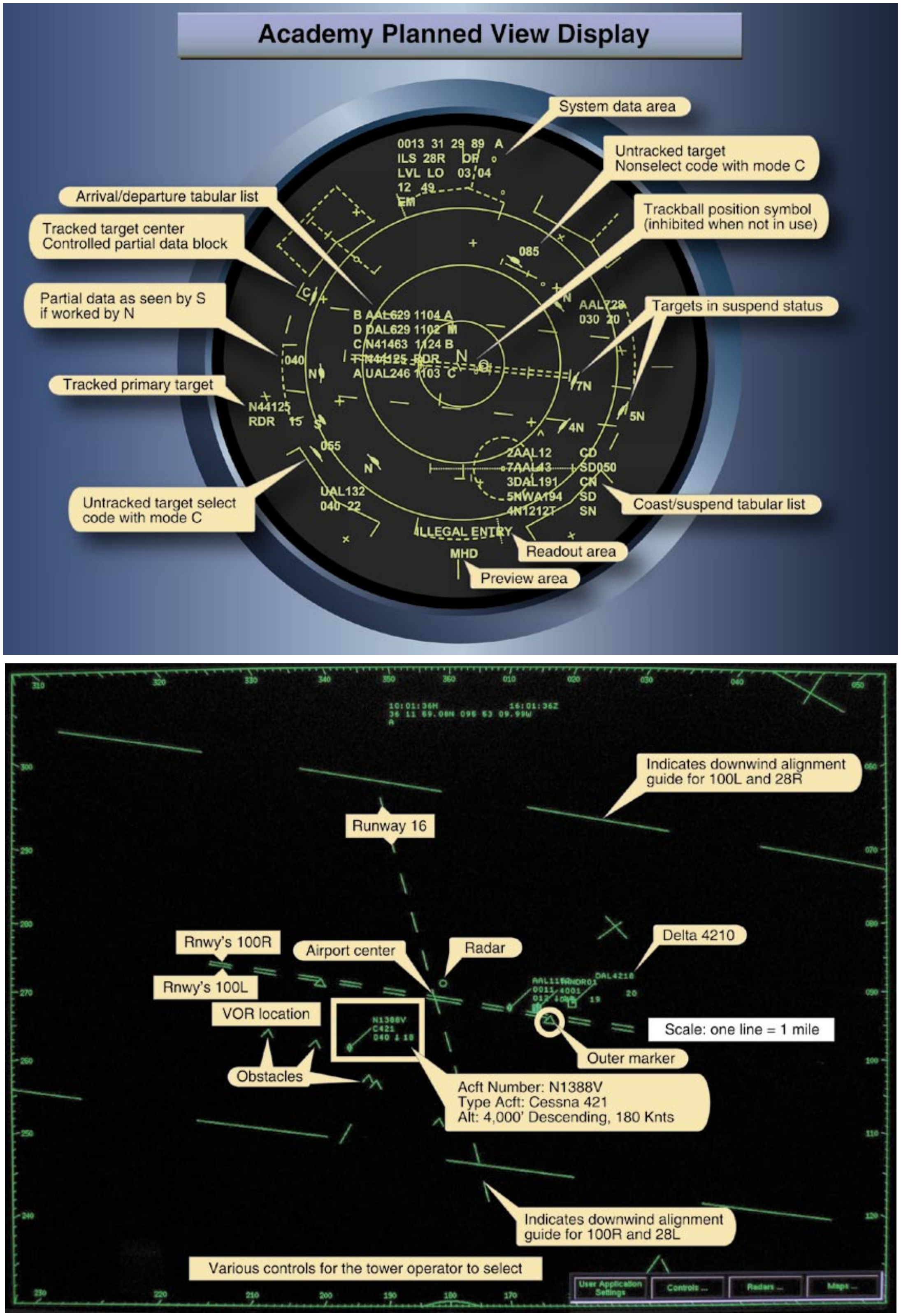
Academy Planned View Display
System data area
Untracked target
Nonselect code with mode C
Arrival/departure tabular list
Trackball position symbol
(inhibited when not in use)
Tracked target center
Controlled partial data block
Partial data as seen by S
if worked by N
Targets in suspend status
Tracked primary target
Untracked target select
code with mode C
Coast/suspend tabular list
Readout area
Preview area
Indicates downwind alignment
guide for 100L and 28R
Runway 16
Rnwy's 100R
Airport center
Radar
Delta 4210
Rnwy's 100L
VOR location
Scale: one line = 1 mile
Outer marker
Obstacles
Acft Number: N1388V
Type Acft: Cessna 421
Alt: 4,000' Descending, 180 Knts
Indicates downwind alignment
guide for 100R and 28L
Various controls for the tower operator to select

Figure 9-9. *The top image is a display as seen by controllers in an Air Traffic Facility. The one illustrated is an ARTS III (Automated Radar Terminal System). The display shown provides an explanation of the symbols in the graphic. The lower figure is an example of the Digital Bright Radar Indicator Tower Equipment (DBRITE) screen as seen by tower personnel. It provides tower controllers with a visual display of the airport surveillance radar, beacon signals, and data received from ARTS III. The display shown provides an explanation of the symbols in the graphic.*

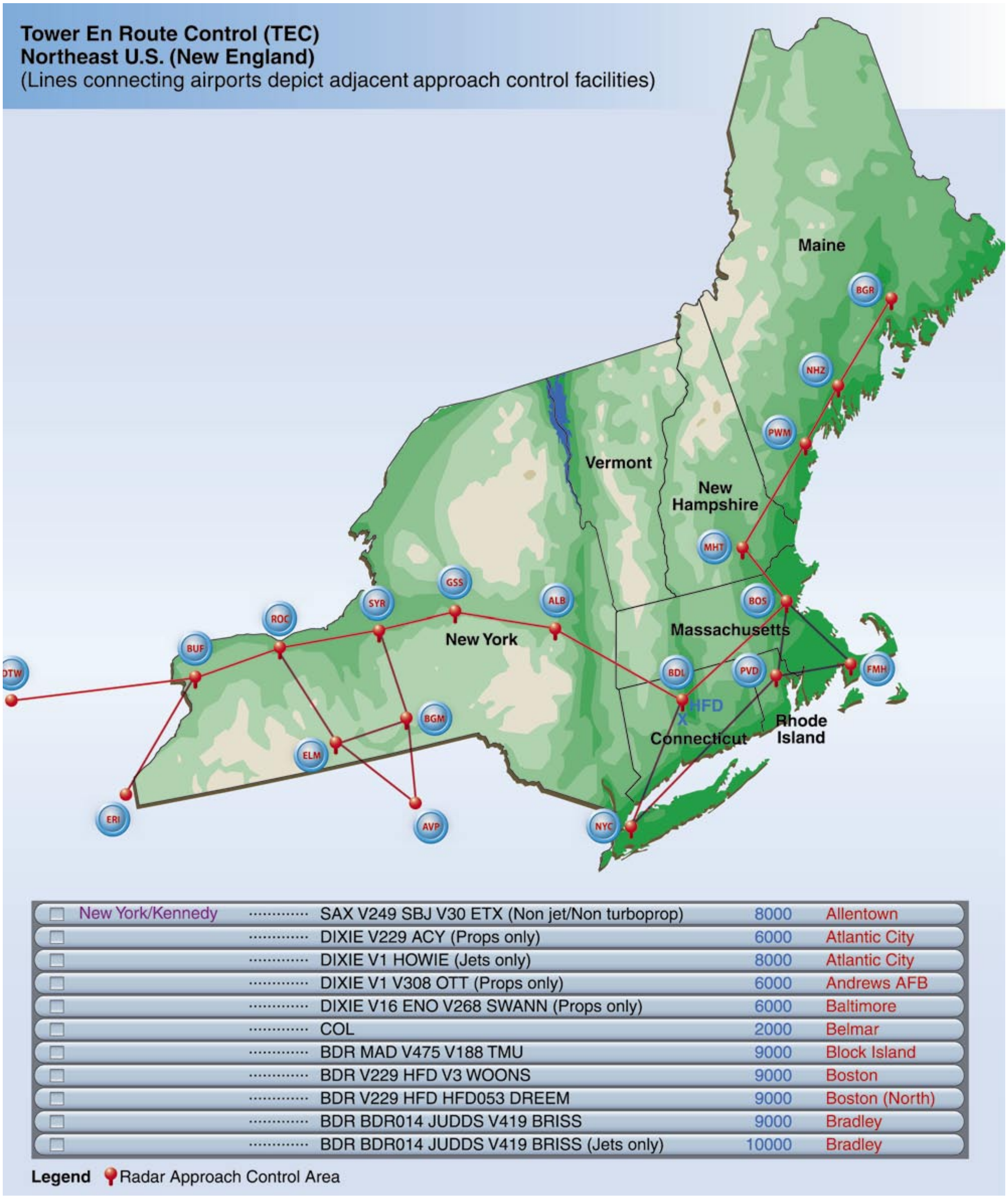
Maine
BGR
NHZ
PWM
Vermont
New Hampshire
MHT
GSS
SYR
ROC
BUF
DTW
ALB
BOS
New York
Massachusetts
BDL
PVD
FMH
HFD
X
BGM
Connecticut
Rhode Island
ELM
ERI
AVP
NYC

Tower En Route Control (TEC)
Northeast U.S. (New England)
(Lines connecting airports depict adjacent approach control facilities)

☐	New York/Kennedy	....	SAX V249 SBJ V30 ETX (Non jet/Non turboprop)	8000	Allentown
☐		....	DIXIE V229 ACY (Props only)	6000	Atlantic City
☐		....	DIXIE V1 HOWIE (Jets only)	8000	Atlantic City
☐		....	DIXIE V1 V308 OTT (Props only)	6000	Andrews AFB
☐		....	DIXIE V16 ENO V268 SWANN (Props only)	6000	Baltimore
☐		....	COL	2000	Belmar
☐		....	BDR MAD V475 V188 TMU	9000	Block Island
☐		....	BDR V229 HFD V3 WOONS	9000	Boston
☐		....	BDR V229 HFD HFD053 DREEM	9000	Boston (North)
☐		....	BDR BDR014 JUDDS V419 BRISS	9000	Bradley
☐		....	BDR BDR014 JUDDS V419 BRISS (Jets only)	10000	Bradley

Legend 📍Radar Approach Control Area

Figure 9-10. *A Portion of the New York Area Tower En Route List. (From the A/FD)*

Figure 9-11. *Center Radar Displays.*

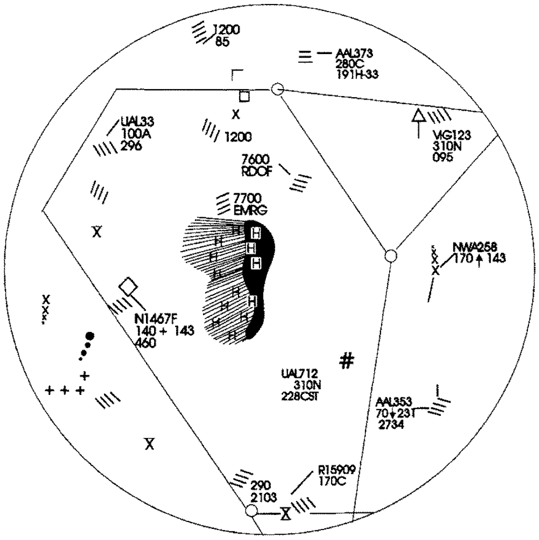

Figure 9-12. *A Center Controller's Scope.*

A common clearance in these situations is "When able, proceed direct to the Astoria VOR…" The words "when able" mean to proceed to the waypoint, intersection, or NAVAID when the pilot is able to navigate directly to that point using onboard available systems providing proper guidance, usable signal, etc. If provided such guidance while flying VFR, the pilot remains responsible for terrain and obstacle clearance. Using the standard climb gradient, an aircraft is 2 miles from the departure end of the runway before it is safe to turn (400 feet above ground level (AGL)). When a Center controller issues a heading, a direct route, or says "direct when able," the controller becomes responsible for terrain and obstruction clearance.

Another common Center clearance is "Leaving (altitude) fly (heading) or proceed direct when able." This keeps the terrain/obstruction clearance responsibility in the flight deck until above the minimum IFR altitude. A controller cannot issue an IFR clearance until an aircraft is above the minimum IFR altitude unless it is able to climb in VFR conditions.

On a Center controller's scope, 1 NM is about 1/28 of an inch. When a Center controller is providing Approach/Departure control services at an airport many miles from the radar antenna, estimating headings and distances is very difficult. Controllers providing vectors to final must set the range on their scopes to not more than 125 NM to provide the greatest possible accuracy for intercept headings. Accordingly, at locations more distant from a Center radar antenna, pilots should expect a minimum of vectoring.

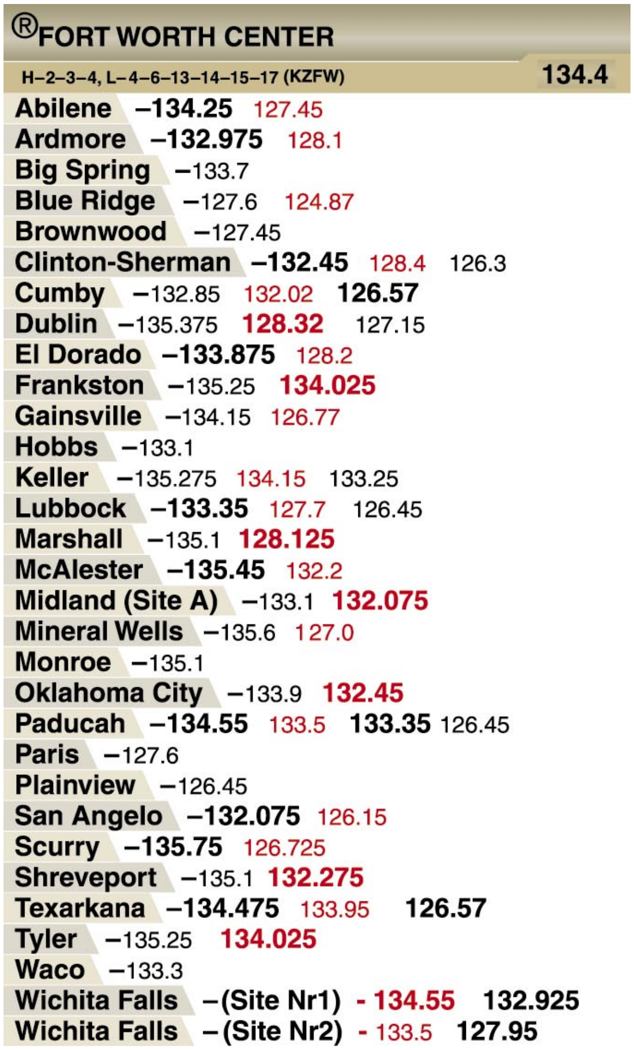

®FORT WORTH CENTER

H–2–3–4, L–4–6–13–14–15–17 (KZFW) 134.4

Abilene	–134.25	127.45		
Ardmore	–132.975	128.1		
Big Spring	–133.7			
Blue Ridge	–127.6	124.87		
Brownwood	–127.45			
Clinton-Sherman	–132.45	128.4	126.3	
Cumby	–132.85	132.02	126.57	
Dublin	–135.375	128.32	127.15	
El Dorado	–133.875	128.2		
Frankston	–135.25	134.025		
Gainsville	–134.15	126.77		
Hobbs	–133.1			
Keller	–135.275	134.15	133.25	
Lubbock	–133.35	127.7	126.45	
Marshall	–135.1	128.125		
McAlester	–135.45	132.2		
Midland (Site A)	–133.1	132.075		
Mineral Wells	–135.6	127.0		
Monroe	–135.1			
Oklahoma City	–133.9	132.45		
Paducah	–134.55	133.5	133.35	126.45
Paris	–127.6			
Plainview	–126.45			
San Angelo	–132.075	126.15		
Scurry	–135.75	126.725		
Shreveport	–135.1	132.275		
Texarkana	–134.475	133.95	126.57	
Tyler	–135.25	134.025		
Waco	–133.3			
Wichita Falls	–(Site Nr1)	- 134.55	132.925	
Wichita Falls	–(Site Nr2)	- 133.5	127.95	

Figure 9-13. *Center Symbology.*

ATC Inflight Weather Avoidance Assistance

ATC Radar Weather Displays

ATC radar systems are able to display areas of precipitation by sending out a beam of radio energy that is reflected back to the radar antenna when it strikes an object or moisture which may be in the form of rain drops, hail, or snow. The larger the object, or the denser its reflective surface, the stronger the return will be. Radar weather processors indicate the intensity of reflective returns in terms of decibels with respect to the radar reflectively factor (dBZ).

ATC systems cannot detect the presence or absence of clouds. ATC radar systems can often determine the intensity of a precipitation area, but the specific character of that area (snow, rain, hail, VIRGA, etc.) cannot be determined. For this reason, ATC refers to all weather areas displayed on ATC radar scopes as "precipitation."

All ATC facilities using radar weather processors with the ability to determine precipitation intensity describes the intensity to pilots as:

1. "LIGHT"	(< 30 dBZ)
2. "MODERATE"	(30 to 40 dBZ)
3. "HEAVY"	(>40 to 50 dBZ)
4. "EXTREME"	(>50 dBZ)

ARTCC controllers do not use the term "LIGHT" because their systems do not display "LIGHT" precipitation intensities. ATC facilities that, due to equipment limitations, cannot display the intensity levels of precipitation, will describe the location of the precipitation area by geographic position, or position relative to the aircraft. Since the intensity level is not available, the controller states, "INTENSITY UNKNOWN."

ARTCC facilities normally use a Weather and Radar Processor (WARP) to display a mosaic of data obtained from multiple NEXRAD sites. The WARP processor is only used in ARTCC facilities.

There is a time delay between actual conditions and those displayed to the controller. For example, the precipitation data on the ARTCC controller's display could be up to 6 minutes old. When the WARP is not available, a secondary system, the narrowband ARSR is utilized. The ARSR system can display two distinct levels of precipitation intensity that is described to pilots as "MODERATE" (30 to 40 dBZ) and "HEAVY to EXTREME" (>40 dBZ).

ATC radar systems cannot detect turbulence. Generally, turbulence can be expected to occur as the rate of rainfall or intensity of precipitation increases. Turbulence associated with greater rates of rainfall/precipitation is normally more severe than any associated with lesser rates of rainfall/precipitation. Turbulence should be expected to occur near convective activity, even in clear air. Thunderstorms are a form of convective activity that implies severe or greater turbulence. Operation within 20 miles of thunderstorms should be approached with great caution, as the severity of turbulence can be markedly greater than the precipitation intensity might indicate.

Weather Avoidance Assistance

ATC's first duty priority is to separate aircraft and issue safety alerts. ATC provides additional services to the extent possible, contingent upon higher priority duties and other factors including limitations of radar, volume of traffic, frequency congestion, and workload. Subject to the above factors/limitations, controllers issue pertinent information on weather or chaff areas; and if requested, assist pilots, to the extent possible, in avoiding areas of precipitation. Pilots should respond to a weather advisory by acknowledging the advisory and, if desired, requesting an alternate course of action, such as:

1. Request to deviate off course by stating the direction and number of degrees or miles needed to deviate from the original course;
2. Request a change of altitude; or
3. Request routing assistance to avoid the affected area. Because ATC radar systems cannot detect the presence or absence of clouds and turbulence, such assistance conveys no guarantee that the pilot will not encounter hazards associated with convective activity. Pilots wishing to circumnavigate precipitation areas by a specific distance should make their desires clearly known to ATC at the time of the request for services. Pilots must advise ATC when they can resume normal navigation.

IFR pilots shall not deviate from their assigned course or altitude without an ATC clearance. Plan ahead for possible course deviations because hazardous convective conditions can develop quite rapidly. This is important to consider because the precipitation data displayed on ARTCC radar scopes can be up to 6 minutes old and thunderstorms can develop at rates exceeding 6,000 feet per minute (fpm). When encountering weather conditions that threaten the safety of the aircraft, the pilot may exercise emergency authority as stated in 14 CFR part 91, section 91.3 should an immediate deviation from the assigned clearance be necessary and time does not permit approval by ATC.

Generally, when weather disrupts the flow of air traffic, greater workload demands are placed on the controller. Requests for deviations from course and other services should be made as far in advance as possible to better assure the controller's ability to approve these requests promptly. When requesting approval to detour around weather activity, include the following information to facilitate the request:

1. The proposed point where detour commences;
2. The proposed route and extent of detour (direction and distance);
3. The point where original route will be resumed;
4. Flight conditions (IMC or VMC);
5. Whether the aircraft is equipped with functioning airborne radar; and
6. Any further deviation that may become necessary.

To a large degree, the assistance that might be rendered by ATC depends upon the weather information available to controllers. Due to the extremely transitory nature of hazardous weather, the controller's displayed precipitation information may be of limited value.

Obtaining IFR clearance or approval to circumnavigate hazardous weather can often be accommodated more readily in the en route areas away from terminals because there is usually less congestion and, therefore, greater freedom of action. In terminal areas, the problem is more acute because of traffic density, ATC coordination requirements, complex departure and arrival routes, and adjacent airports. As a consequence, controllers are less likely to be able to accommodate all requests for weather detours in a terminal area. Nevertheless, pilots should not hesitate to advise controllers of any observed hazardous weather and should specifically advise controllers if they desire circumnavigation of observed weather.

Pilot reports (PIREPs) of flight conditions help define the nature and extent of weather conditions in a particular area. These reports are disseminated by radio and electronic means to other pilots. Provide PIREP information to ATC regarding pertinent flight conditions, such as:

1. Turbulence;
2. Visibility;
3. Cloud tops and bases; and
4. The presence of hazards such as ice, hail, and lightning.

Approach Control Facility

An approach control facility is a terminal ATC facility that provides approach control service in the terminal area. Services are provided for arriving and departing VFR and IFR aircraft and, on occasion, en route aircraft. In addition, for airports with parallel runways with ILS or LDA approaches, the approach control facility provides monitoring of the approaches.

Approach Control Advances

Precision Runway Monitor (PRM)

Over the past few years, a new technology has been installed at airports that permits a decreased separation distance between parallel runways. The system is called a Precision Runway Monitor (PRM) and is comprised of high-update radar, high-resolution ATC displays, and PRM-certified controllers. *[Figure 9-14]*

Figure 9-14. *High Resolution ATC Displays Used in PRM.*

Precision Runway Monitor (PRM) Radar

The PRM uses a Monopulse Secondary Surveillance Radar (MSSR) that employs electronically scanned antennas. Because the PRM has no scan rate restrictions, it is capable of providing a faster update rate (up to 0.5 second) over conventional systems, thereby providing better target presentation in terms of accuracy, resolution, and track prediction. The system is designed to search, track, process, and display SSR-equipped aircraft within airspace of over 30 miles in range and over 15,000 feet in elevation. Visual and audible alerts are generated to warn controllers to take corrective actions.

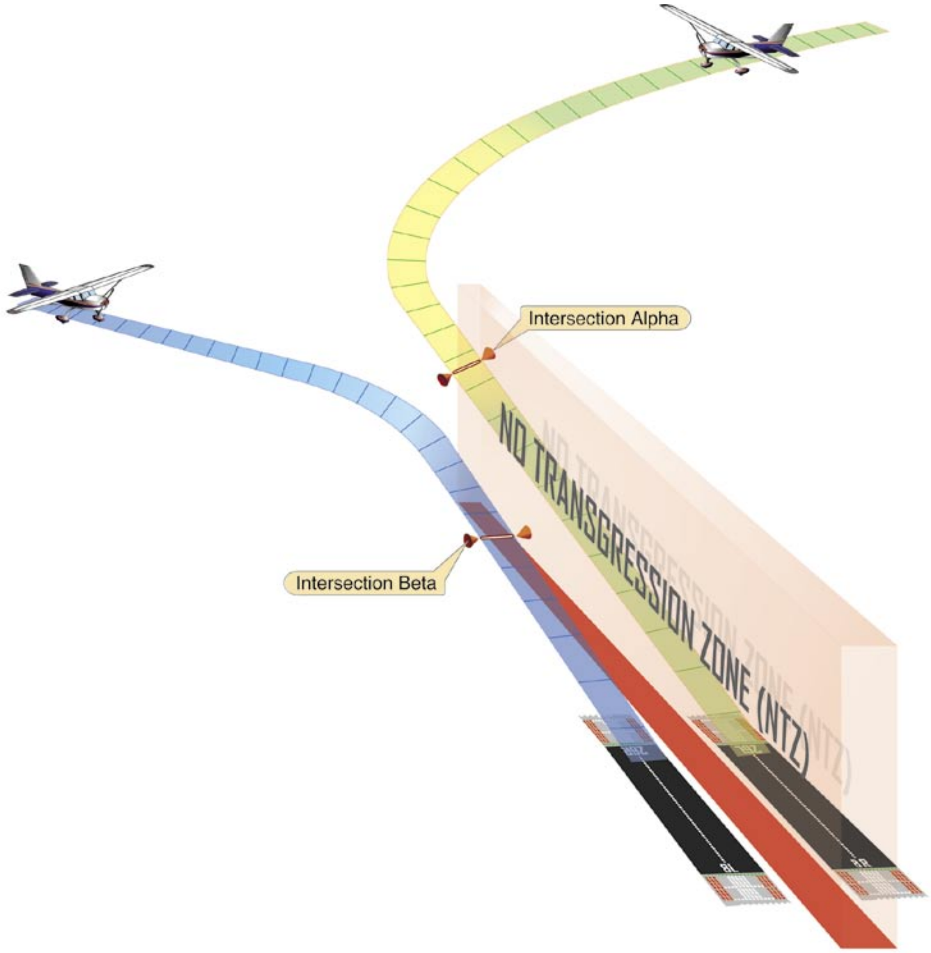

Figure 9-15. *Aircraft Management Using PRM. (Note the no transgression zone (NTZ) and how the aircraft are separated.)*

PRM Benefits

Typically, PRM is used with dual approaches with centerlines separated less than 4,300 feet but not less than 3,000 feet (under most conditions). *[Figure 9-15]* Separating the two final approach courses is a No Transgression Zone (NTZ) with surveillance of that zone provided by two controllers, one for each active approach. The system tracking software provides PRM monitor controllers with aircraft identification, position, speed, projected position, as well as visual and aural alerts.

Control Sequence

The IFR system is flexible and accommodating if pilots do their homework, have as many frequencies as possible written down before they are needed, and have an alternate in mind if the flight cannot be completed as planned. Pilots should familiarize themselves with all the facilities and services available along the planned route of flight. *[Figure 9-16]* Always know where the nearest VFR conditions can be found, and be prepared to head in that direction if the situation deteriorates.

A typical IFR flight, with departure and arrival at airports with control towers, would use the ATC facilities and services in the following sequence:

1. AFSS: Obtain a weather briefing for a departure, destination and alternate airports, and en route conditions, and then file a flight plan by calling 1-800-WX-BRIEF.
2. ATIS: Preflight complete, listen for present conditions and the approach in use.
3. Clearance Delivery: Prior to taxiing, obtain a departure clearance.
4. Ground Control: Noting that the flight is IFR, receive taxi instructions.
5. Tower: Pre-takeoff checks complete, receive clearance to takeoff.
6. Departure Control: Once the transponder "tags up" with the ARTS, the tower controller instructs the pilot to contact Departure to establish radar contact.

7. ARTCC: After departing the departure controller's airspace, aircraft is handed off to Center, who coordinates the flight while en route. Pilots may be in contact with multiple ARTCC facilities; they coordinate the hand-offs.
8. EFAS/HIWAS: Coordinate with ATC before leaving their frequency to obtain inflight weather information.
9. ATIS: Coordinate with ATC before leaving their frequency to obtain ATIS information.
10. Approach Control: Center hands off to approach control where pilots receive additional information and clearances.
11. Tower: Once cleared for the approach, pilots are instructed to contact tower control; the flight plan is canceled by the tower controller upon landing.

A typical IFR flight, with departure and arrival at airports without operating control towers, would use the ATC facilities and services in the following sequence:

1. AFSS: Obtain a weather briefing for departure, destination, and alternate airports, and en route conditions, and then file a flight plan by calling 1-800-WX-BRIEF. Provide the latitude/longitude description for small airports to ensure that Center is able to locate departure and arrival locations.
2. AFSS or UNICOM: ATC clearances can be filed and received on the UNICOM frequency if the licensee has made arrangements with the controlling ARTCC; otherwise, file with AFSS via telephone. Be sure all preflight preparations are complete before filing. The clearance includes a clearance void time. Pilots must be airborne prior to the void time.
3. ARTCC: After takeoff, establish contact with Center. During the flight, pilots may be in contact with multiple ARTCC facilities; ATC coordinates the hand-offs.
4. EFAS/HIWAS: Coordinate with ATC before leaving their frequency to obtain in-flight weather information.
5. Approach Control: Center hands off to approach control where pilots receive additional information and clearances. If a landing under visual meteorological conditions (VMC) is possible, pilots may cancel their IFR clearance before landing.

Letters of Agreement (LOA)

The ATC system is indeed a system, and very little happens by chance. As a flight progresses, controllers in adjoining sectors or adjoining Centers coordinate its handling by telephone or by computer. Where there is a boundary between the airspace controlled by different facilities, the location and altitude for hand-off is determined by Letters of Agreement (LOA) negotiated between the two facility managers. This information is not available to pilots in any Federal Aviation Administration (FAA) publication. For this reason, it is good practice to note on the en route chart the points at which hand-offs occur. Each time a flight is handed-off to a different facility, the controller knows the altitude and location—this was part of the hand-off procedure.

Communications Facility	Description	Frequency
Airport Advisory Area "[AFSS name] RADIO"	AFSS personnel provide traffic advisories to pilots operating within 10 miles of the airport.	123.6 MHz.
UNICOM "[airport name] UNICOM"	Airport advisories from an airport without an operating control tower or AFSS.	Listed in A/FD under the city name; also on sectional charts in airport data block.
Air Route Traffic Control Center (ARTCC) "CENTER"	En route radar facilities that maintain separation between IFR flights, and between IFR flights and known VFR flights. Centers provide VFR traffic advisories on a workload permitting basis.	Listed in A/FD and on instrument enroute charts.
Approach/Departure Control "[airport name] APPROACH" (unless otherwise advised)	Positions at a terminal radar facility responsible for handling of IFR flights to and from the primary airport (where Class B airspace exists).	Listed in A/FD; also on sectional charts in the communications panel and on terminal area charts.
Automatic Terminal Information Service (ATIS)	Continuous broadcast of audio tape prepared by ATC controller containing wind direction and velocity, temperature, altimeter setting, runway and approach in use, and other information of interest to pilots.	Listed in A/FD under the city name; also on sectional charts in airport data block and in the communications panel, and on terminal area charts.
Clearance Delivery "[airport name] CLEARANCE"	Control tower position responsible for transmitting departure clearances to IFR flights.	Listed on instrument approach procedure charts.
Common Traffic Advisory Frequency (CTAF)	CTAF provides a single frequency for pilots in the area to use for contacting the facility and/or broadcasting their position and intentions to other pilots.	Listed in A/FD; also on sectional charts in the airport data block (followed by a white C on a blue or magenta background). At airports with no tower, CTAF is 122.9, the "MULTICOM" frequency.
Automated Flight Service Station (AFSS) "[facility name] RADIO"	Provides information and services to pilots, using remote communications outlets (RCOs) and ground communications outlets (GCOs).	Listed in A/FD and sectional charts, both under city name and in a separate listing of AFSS frequencies. On sectional charts, listed above the VOR boxes, or in separate boxes when remote.
Ground Control "[airport name] GROUND"	At tower-controlled airports, a position in the tower responsible for controlling aircraft taxiing to and from the runways.	Listed in A/FD under city name.
Hazardous Inflight Weather Advisory Service (HIWAS)	Continuous broadcast of forecast hazardous weather conditions on selected NAVAIDs. No communication capability.	Black circle with white "H" in VOR frequency box; notation in A/FD airport listing under "Radio Aids to Navigation."
MULTICOM "[airport name] TRAFFIC"	Intended for use by pilots at airports with no radio facilities. Pilots should use self-announce procedures given in the AIM.	122.9 MHz. A/FD shows 122.9 as CTAF; also on sectional charts 122.9 is followed by a white C on a dark background, indicating CTAF.
Tower "[airport name] TOWER"	"Local" controller responsible for operations on the runways and in Class B, C, or D airspace surrounding the airport.	Listed in A/FD under city name; also on sectional and terminal control area charts in the airport data block and communications panel.
En Route Flight Advisory Service (EFAS) "FLIGHT WATCH"	For inflight weather information.	122.0 MHz (0600-2200 local time)

Figure 9-16. *ATC Facilities, Services, and Radio Call Signs.*

(IAF)
COLIG
3000
181° (2.7)
BIDPE
(IAF)
ALEPE
3000
271°
(5)
204
(IAF)
CAPUN
3000
091°
(5)
3000
033°
(5)
2300
181°
(5)
VULCAN
VUZ
(FAF)
BASKN
995
965
822
961
915±
866
RW18
835
951
1427
1403
1398
1137
1049
1219
1073
1287
1105
1185
FLIGHT PLAN
N1230A
C182/G
140
DTW
1300
DCT LFD DCT CMI
CMI-WILLARD AIRPORT SAVOY, IL
01 42
WARREN SMITH, 123 MAIN STREET DETROIT, MI 48123 217-555-1212 DTW
2
04 20
ALN
BLUE/WHITE
CLOSE VFR FLIGHT PLAN WITH
FSS ON ARRIVAL
FAA Form 7233-1 (8-82)
18

Chapter 10

IFR Flight

Introduction

This chapter is a discussion of conducting a flight under instrument flight rules (IFR). It also explains the sources for flight planning, the conditions associated with instrument flight, and the procedures used for each phase of IFR flight: departure, en route, and approach. The chapter concludes with an example of an IFR flight which applies many of the procedures discussed in the chapter.

Sources of Flight Planning Information

The following resources are available for a pilot planning a flight conducted under instrument flight rules (IFR).

National Aeronautical Charting Group (NACG) publications:

- IFR en route charts
- area charts
- United States (U.S.) Terminal Procedures Publications (TPP)

The Federal Aviation Administration (FAA) publications:

- AIM
- Airport/Facility Directory (A/FD)
- Notices to Airmen Publication (NTAP) for flight planning in the National Airspace System (NAS)

Pilots should also consult the Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM) for flight planning information pertinent to the aircraft to be flown.

A review of the contents of all the listed publications will help determine which material should be referenced for each flight. As a pilot becomes more familiar with these publications, the flight planning process becomes quicker and easier.

Aeronautical Information Manual (AIM)

The AIM provides the aviation community with basic flight information and air traffic control (ATC) procedures used in the United States NAS. An international version called the Aeronautical Information Publication contains parallel information, as well as specific information on the international airports used by the international community.

Airport/Facility Directory (A/FD)

The A/FD contains information on airports, communications, and navigation aids pertinent to IFR flight. It also includes very-high frequency omnidirectional range (VOR) receiver checkpoints, automated flight service station (AFSS), weather service telephone numbers, and air route traffic control center (ARTCC) frequencies. Various special notices essential to flight are also included, such as land-and-hold-short operations (LAHSO) data, the civil use of military fields, continuous power facilities, and special flight procedures.

In the major terminal and en route environments, preferred routes have been established to guide pilots in planning their routes of flight, to minimize route changes, and to aid in the orderly management of air traffic using the federal airways. The A/FD lists both high and low altitude preferred routes.

Notices to Airmen Publication (NTAP)

The NTAP is a publication containing current Notices to Airmen (NOTAMs) which are essential to the safety of flight, as well as supplemental data affecting the other operational publications listed. It also includes current Flight Data Center (FDC) NOTAMs, which are regulatory in nature, issued to establish restrictions to flight or to amend charts or published instrument approach procedures (IAPs).

POH/AFM

The POH/AFM contain operating limitations, performance, normal and emergency procedures, and a variety of other operational information for the respective aircraft. Aircraft manufacturers have done considerable testing to gather and substantiate the information in the aircraft manual. Pilots should refer to it for information relevant to a proposed flight.

IFR Flight Plan

As specified in Title 14 of the Code of Federal Regulations (14 CFR) part 91, no person may operate an aircraft in controlled airspace under IFR unless that person has filed an IFR flight plan. Flight plans may be submitted to the nearest AFSS or air traffic control tower (ATCT) either in person, by telephone (1-800-WX-BRIEF), by computer (using the direct user access terminal system (DUATS)), or by radio if no other means are available. Pilots should file IFR flight plans at least 30 minutes prior to estimated time of departure to preclude possible delay in receiving a departure clearance from ATC. The AIM provides guidance for completing and filing FAA Form 7233-1, Flight Plan. These forms are available at flight service stations (FSSs), and are generally found in flight planning rooms at airport terminal buildings. *[Figure 10-1]*

Filing in Flight

IFR flight plans may be filed from the air under various conditions, including:

1. A flight outside controlled airspace before proceeding into IFR conditions in controlled airspace.
2. A VFR flight expecting IFR weather conditions en route in controlled airspace.

In either of these situations, the flight plan may be filed with the nearest AFSS or directly with the ARTCC. A pilot who files with the AFSS submits the information normally entered during preflight filing, except for "point of departure," together with present position and altitude. AFSS then relays this information to the ARTCC. The ARTCC will then clear the pilot from present position or from a specified navigation fix.

A pilot who files directly with the ARTCC reports present position and altitude, and submits only the flight plan information normally relayed from the AFSS to the ARTCC. Be aware that traffic saturation frequently prevents ARTCC personnel from accepting flight plans by radio. In such cases, a pilot is advised to contact the nearest AFSS to file the flight plan.

Cancelling IFR Flight Plans

An IFR flight plan may be cancelled any time a pilot is operating in VFR conditions outside Class A airspace by stating "cancel my IFR flight plan" to the controller or air-to-ground station. After cancelling an IFR flight plan, the pilot should change to the appropriate air-to-ground frequency, transponder code as directed, and VFR altitude/flight level.

ATC separation and information services (including radar services, where applicable) are discontinued when an IFR flight plan is cancelled. If VFR radar advisory service is desired, a pilot must specifically request it. Be aware that other procedures may apply when cancelling an IFR flight plan within areas such as Class C or Class B airspace.

When operating on an IFR flight plan to an airport with an operating control tower, a flight plan is cancelled automatically upon landing. If operating on an IFR flight plan to an airport without an operating control tower, the pilot is responsible for cancelling the flight plan. This can be done by telephone after landing if there is no operating FSS or other means of direct communications with ATC. When there is no FSS or air-to-ground communications are not possible below a certain altitude, a pilot may cancel an IFR flight plan while still airborne and able to communicate with ATC by radio. If using this procedure, be certain the remainder of the flight can be conducted under VFR. It is essential that IFR flight plans be cancelled expeditiously. This allows other IFR traffic to utilize the airspace.

Clearances

An ATC clearance allows an aircraft to proceed under specified traffic conditions within controlled airspace for the purpose of providing separation between known aircraft.

Examples

A flight filed for a short distance at a relatively low altitude in an area of low traffic density might receive a clearance as follows:

> "Cessna 1230 Alpha, cleared to Doeville airport direct, cruise 5,000."

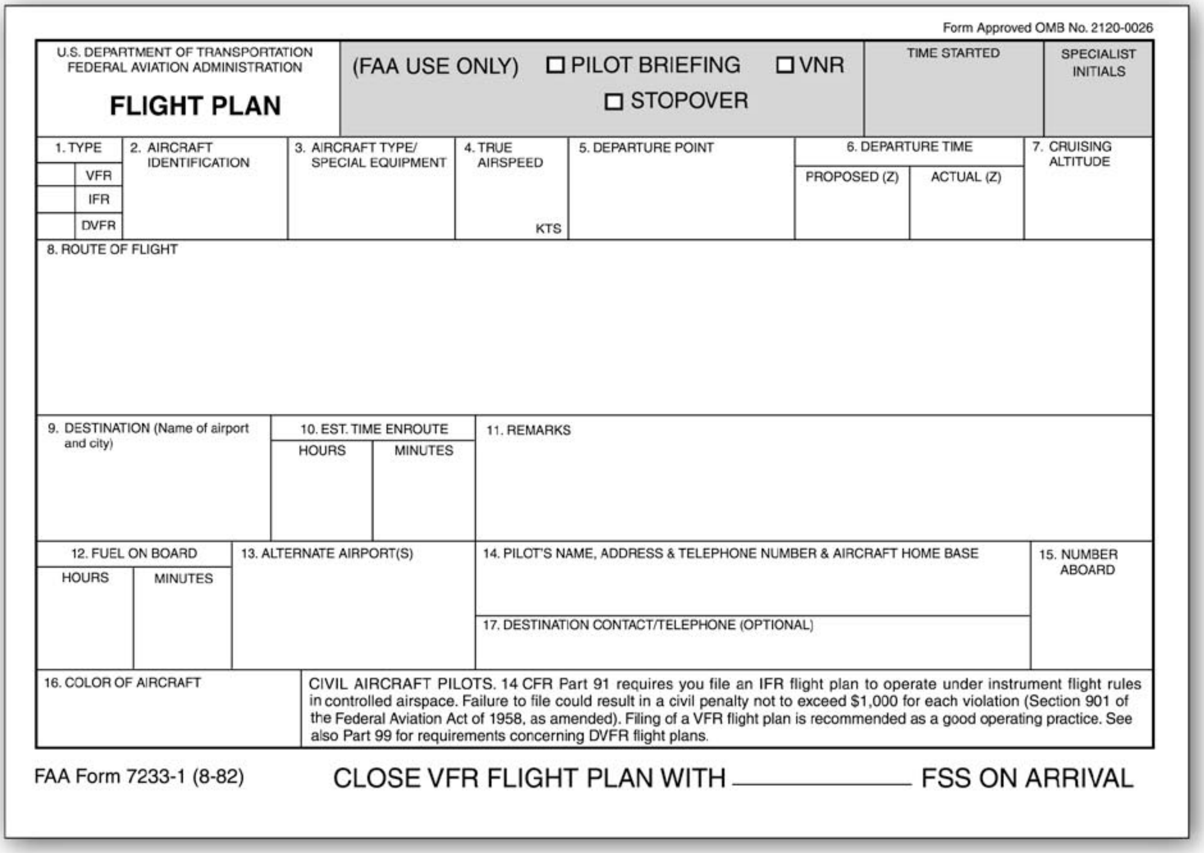

Form Approved OMB No. 2120-0026

U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION
FLIGHT PLAN

(FAA USE ONLY) ☐ PILOT BRIEFING ☐ VNR ☐ STOPOVER

TIME STARTED

SPECIALIST INITIALS

1. TYPE
☐ VFR
☐ IFR
☐ DVFR

2. AIRCRAFT IDENTIFICATION

3. AIRCRAFT TYPE/ SPECIAL EQUIPMENT

4. TRUE AIRSPEED KTS

5. DEPARTURE POINT

6. DEPARTURE TIME PROPOSED (Z) ACTUAL (Z)

7. CRUISING ALTITUDE

8. ROUTE OF FLIGHT

9. DESTINATION (Name of airport and city)

10. EST. TIME ENROUTE HOURS MINUTES

11. REMARKS

12. FUEL ON BOARD HOURS MINUTES

13. ALTERNATE AIRPORT(S)

14. PILOT'S NAME, ADDRESS & TELEPHONE NUMBER & AIRCRAFT HOME BASE

15. NUMBER ABOARD

17. DESTINATION CONTACT/TELEPHONE (OPTIONAL)

16. COLOR OF AIRCRAFT

CIVIL AIRCRAFT PILOTS. 14 CFR Part 91 requires you file an IFR flight plan to operate under instrument flight rules in controlled airspace. Failure to file could result in a civil penalty not to exceed $1,000 for each violation (Section 901 of the Federal Aviation Act of 1958, as amended). Filing of a VFR flight plan is recommended as a good operating practice. See also Part 99 for requirements concerning DVFR flight plans.

FAA Form 7233-1 (8-82)

CLOSE VFR FLIGHT PLAN WITH ________ FSS ON ARRIVAL

Figure 10-1. *Flight Plan Form.*

The term "cruise" in this clearance means a pilot is authorized to fly at any altitude from the minimum IFR altitude up to and including 5,000 feet, and may level off at any altitude within this block of airspace. A climb or descent within the block may be made at the pilot's discretion. However, once a pilot reports leaving an altitude within the block, the pilot may not return to that altitude without further ATC clearance.

When ATC issues a cruise clearance in conjunction with an unpublished route, an appropriate crossing altitude will be specified to ensure terrain clearance until the aircraft reaches a fix, point, or route where the altitude information is available. The crossing altitude ensures IFR obstruction clearance to the point at which the aircraft enters a segment of a published route or IAP.

Once a flight plan is filed, ATC will issue the clearance with appropriate instructions, such as the following:

> "Cessna 1230 Alpha is cleared to Skyline airport via the Crossville 055 radial, Victor 18, maintain 5,000. Clearance void if not off by 1330."

Or a more complex clearance, such as:

> "Cessna 1230 Alpha is cleared to Wichita Mid-continent airport via Victor 77, left turn after takeoff, proceed direct to the Oklahoma City VORTAC. Hold west on the Oklahoma City 277 radial, climb to 5,000 in holding pattern before proceeding on course. Maintain 5,000 to CASHION intersection. Climb to and maintain 7,000. Departure control frequency will be 121.05, Squawk 0412."

Clearance delivery may issue the following "abbreviated clearance" which includes a departure procedure (DP):

> "Cessna 1230 Alpha, cleared to La Guardia as filed, RINGOES 8 departure Phillipsburg transition, maintain 8,000. Departure control frequency will be 120.4, Squawk 0700."

This clearance may be readily copied in shorthand as follows:

> "CAF RNGO8 PSB M80 DPC 120.4 SQ 0700."

The information contained in this DP clearance is abbreviated using clearance shorthand (see appendix 1). The pilot should know the locations of the specified navigation facilities, together with the route and point-to-point time, before accepting the clearance.

The DP enables a pilot to study and understand the details of a departure before filing an IFR flight plan. It provides the information necessary to set up communication and navigation equipment and be ready for departure before requesting an IFR clearance.

Once the clearance is accepted, a pilot is required to comply with ATC instructions. A clearance different from that issued may be requested if the pilot considers another course of action more practicable or if aircraft equipment limitations or other considerations make acceptance of the clearance inadvisable.

A pilot should also request clarification or amendment, as appropriate, any time a clearance is not fully understood or considered unacceptable for safety of flight. The pilot is responsible for requesting an amended clearance if ATC issues a clearance that would cause a pilot to deviate from a rule or regulation or would place the aircraft in jeopardy.

Clearance Separations

ATC will provide the pilot on an IFR clearance with separation from other IFR traffic. This separation is provided:

1. Vertically—by assignment of different altitudes.
2. Longitudinally—by controlling time separation between aircraft on the same course.
3. Laterally—by assignment of different flight paths.
4. By radar—including all of the above.

ATC does not provide separation for an aircraft operating:

1. Outside controlled airspace.
2. On an IFR clearance:
    a) With "VFR-On-Top" authorized instead of a specific assigned altitude.
    b) Specifying climb or descent in "VFR conditions."
    c) At any time in VFR conditions, since uncontrolled VFR flights may be operating in the same airspace.

In addition to heading and altitude assignments, ATC will occasionally issue speed adjustments to maintain the required separations. For example:

> "Cessna 30 Alpha, slow to 100 knots."

A pilot who receives speed adjustments is expected to maintain that speed plus or minus 10 knots. If for any reason the pilot is not able to accept a speed restriction, the pilot should advise ATC.

At times, ATC may also employ visual separation techniques to keep aircraft safely separated. A pilot who obtains visual contact with another aircraft may be asked to maintain visual separation or to follow the aircraft. For example:

> "Cessna 30 Alpha, maintain visual separation with that traffic, climb and maintain 7,000."

The pilot's acceptance of instructions to maintain visual separation or to follow another aircraft is an acknowledgment that the aircraft will be maneuvered as necessary, to maintain safe separation. It is also an acknowledgment that the pilot accepts the responsibility for wake turbulence avoidance.

In the absence of radar contact, ATC will rely on position reports to assist in maintaining proper separation. Using the data transmitted by the pilot, the controller follows the progress of each flight. ATC must correlate the pilots' reports to provide separation; therefore, the accuracy of each pilot's report can affect the progress and safety of every other aircraft operating in the area on an IFR flight plan.

Departure Procedures (DPs)

Instrument departure procedures are preplanned instrument flight rule (IFR) procedures, which provide obstruction clearance from the terminal area to the appropriate en route structure and provide the pilot with a way to depart the airport and transition to the en route structure safely. Pilots operating under 14 CFR part 91 are strongly encouraged to file and fly a DP when one is available. *[Figure 10-2]*

There are two types of DPs, Obstacle Departure Procedures (ODP), printed either textually or graphically, and Standard Instrument Departures (SID), always printed graphically. All DPs, either textual or graphic, may be designed using either conventional or RNAV criteria. RNAV procedures will have RNAV printed in the title, e.g., SHEAD TWO DEPARTURE (RNAV).

Obstacle Departure Procedures (ODP)

ODPs provide obstruction clearance via the least onerous route from the terminal area to the appropriate en route structure. ODPs are recommended for obstruction clearance and may be flown without ATC clearance unless an alternate departure procedure (SID or radar vector) has been specifically assigned by ATC. Graphic ODPs will have (OBSTACLE) printed in the procedure title, e.g., GEYSR THREE DEPARTURE (OBSTACLE), CROWN ONE DEPARTURE (RNAV)(OBSTACLE).

Standard Instrument Departures

Standard Instrument Departures (SID) are air traffic control (ATC) procedures printed for pilot/controller use in graphic form to provide obstruction clearance and a transition from the terminal area to the appropriate en route structure. SIDs are primarily designed for system enhancement and to reduce pilot/controller workload. ATC clearance must be received prior to flying a SID.

ODPs are found in section C of each booklet published regionally by the NACG, TPP, along with "IFR Take-off Minimums" while SIDs are collocated with the approach procedures for the applicable airport. Additional information on the development of DPs can be found in paragraph 5-2-7 of the AIM. However, the following points are important to remember.

1. The pilot of IFR aircraft operating from locations where DP procedures are effective may expect an ATC clearance containing a DP. The use of a DP requires pilot possession of at least the textual description of the approved DP.
2. If a pilot does not possess a preprinted DP or for any other reason does not wish to use a DP, he or she is expected to advise ATC. Notification may be accomplished by filing "NO DP" in the remarks section of the filed flight plan, or by advising ATC.
3. If a DP is accepted in a clearance, a pilot must comply with it.

Radar Controlled Departures

On IFR departures from airports in congested areas, a pilot will normally receive navigational guidance from departure control by radar vector. When a departure is to be vectored immediately following takeoff, the pilot will be advised before takeoff of the initial heading to be flown. This information is vital in the event of a loss of two-way radio communications during departure.

The radar departure is normally simple. Following takeoff, contact departure control on the assigned frequency when advised to do so by the control tower. At this time departure control verifies radar contact, and gives headings, altitude, and climb instructions to move an aircraft quickly and safely out of the terminal area. A pilot is expected to fly the assigned headings and altitudes until informed by the controller of the aircraft's position with respect to the route given in the clearance, whom to contact next, and to "resume own navigation."

Departure control will provide vectors to either a navigation facility, or an en route position appropriate to the departure clearance, or transfer to another controller with further radar surveillance capabilities. *[Figure 10-2]*

A radar controlled departure does not relieve the pilot of responsibilities as pilot-in-command. Be prepared before takeoff to conduct navigation according to the ATC clearance, with navigation receivers checked and properly tuned. While under radar control, monitor instruments to ensure continuous orientation to the route specified in the clearance, and record the time over designated checkpoints.

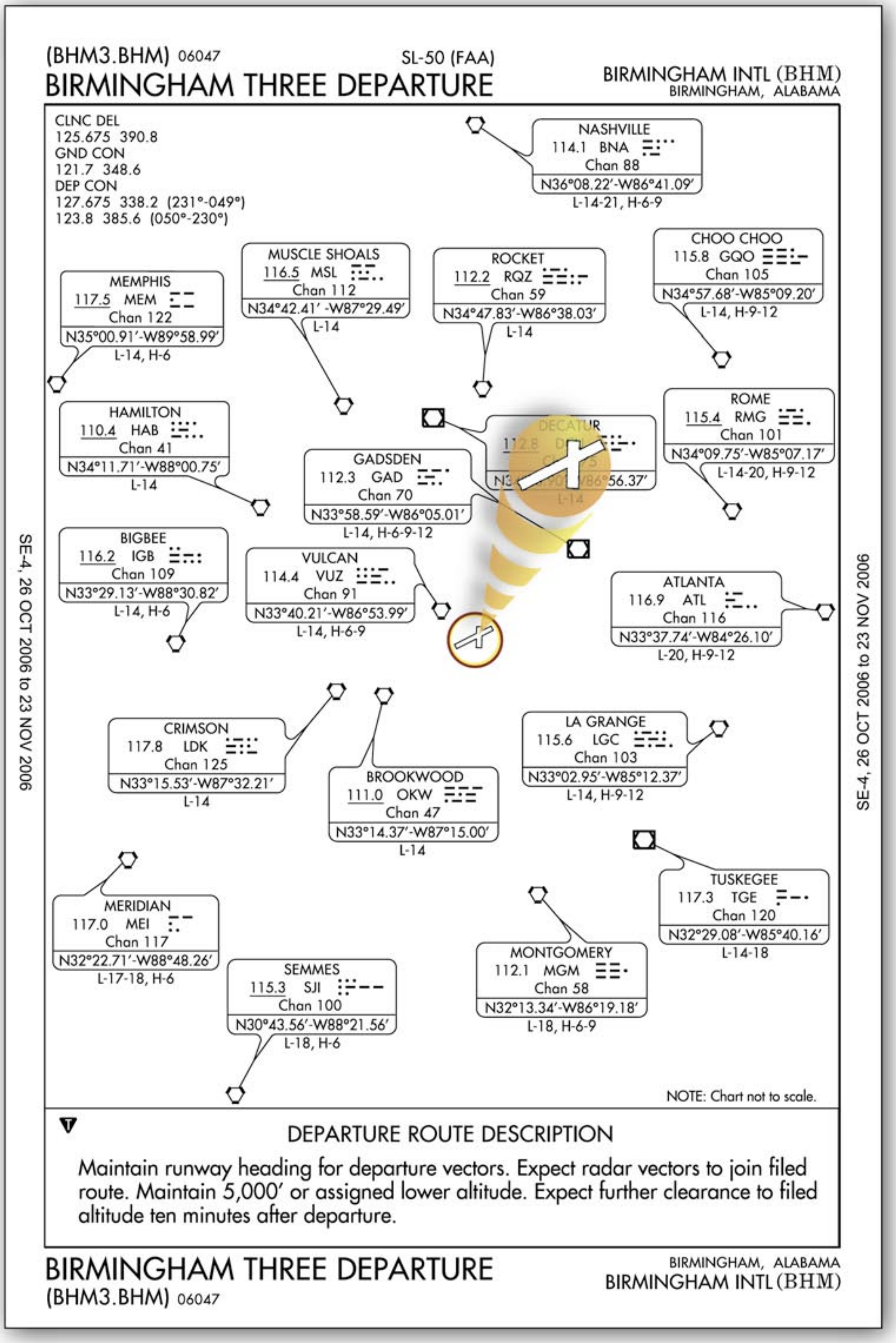

(BHM3.BHM) 06047 SL-50 (FAA)

BIRMINGHAM THREE DEPARTURE

BIRMINGHAM INTL (BHM)
BIRMINGHAM, ALABAMA

CLNC DEL
125.675 390.8
GND CON
121.7 348.6
DEP CON
127.675 338.2 (231°-049°)
123.8 385.6 (050°-230°)

NASHVILLE
114.1 BNA
Chan 88
N36°08.22′-W86°41.09′
L-14-21, H-6-9
MEMPHIS
117.5 MEM
Chan 122
N35°00.91′-W89°58.99′
L-14, H-6
MUSCLE SHOALS
116.5 MSL
Chan 112
N34°42.41′ -W87°29.49′
L-14
ROCKET
112.2 RQZ
Chan 59
N34°47.83′-W86°38.03′
L-14
CHOO CHOO
115.8 GQO
Chan 105
N34°57.68′-W85°09.20′
L-14, H-9-12
HAMILTON
110.4 HAB
Chan 41
N34°11.71′-W88°00.75′
L-14
DECATUR
112.8
W86°56.37′
L-14
ROME
115.4 RMG
Chan 101
N34°09.75′-W85°07.17′
L-14-20, H-9-12
GADSDEN
112.3 GAD
Chan 70
N33°58.59′-W86°05.01′
L-14, H-6-9-12
BIGBEE
116.2 IGB
Chan 109
N33°29.13′-W88°30.82′
L-14, H-6
VULCAN
114.4 VUZ
Chan 91
N33°40.21′-W86°53.99′
L-14, H-6-9
ATLANTA
116.9 ATL
Chan 116
N33°37.74′-W84°26.10′
L-20, H-9-12
CRIMSON
117.8 LDK
Chan 125
N33°15.53′-W87°32.21′
L-14
BROOKWOOD
111.0 OKW
Chan 47
N33°14.37′-W87°15.00′
L-14
LA GRANGE
115.6 LGC
Chan 103
N33°02.95′-W85°12.37′
L-14, H-9-12
TUSKEGEE
117.3 TGE
Chan 120
N32°29.08′-W85°40.16′
L-14-18
MERIDIAN
117.0 MEI
Chan 117
N32°22.71′-W88°48.26′
L-17-18, H-6
SEMMES
115.3 SJI
Chan 100
N30°43.56′-W88°21.56′
L-18, H-6
MONTGOMERY
112.1 MGM
Chan 58
N32°13.34′-W86°19.18′
L-18, H-6-9
SE-4, 26 OCT 2006 to 23 NOV 2006
SE-4, 26 OCT 2006 to 23 NOV 2006

NOTE: Chart not to scale.

DEPARTURE ROUTE DESCRIPTION

Maintain runway heading for departure vectors. Expect radar vectors to join filed route. Maintain 5,000′ or assigned lower altitude. Expect further clearance to filed altitude ten minutes after departure.

BIRMINGHAM THREE DEPARTURE
(BHM3.BHM) 06047

BIRMINGHAM, ALABAMA
BIRMINGHAM INTL (BHM)

Figure 10-2. *Departure Procedure (DP)*.

Departures From Airports Without an Operating Control Tower

When departing from airports that have neither an operating tower nor an FSS, a pilot should telephone the flight plan to the nearest ATC facility at least 30 minutes before the estimated departure time. If weather conditions permit, depart VFR and request IFR clearance as soon as radio contact is established with ATC.

If weather conditions make it undesirable to fly VFR, telephone clearance request. In this case, the controller would probably issue a short-range clearance pending establishment of radio contact, and might restrict the departure time to a certain period. For example:

> "Clearance void if not off by 0900."

This would authorize departure within the allotted period and permit a pilot to proceed in accordance with the clearance. In the absence of any specific departure instructions, a pilot would be expected to proceed on course via the most direct route.

En Route Procedures

Procedures en route will vary according to the proposed route, the traffic environment, and the ATC facilities controlling the flight. Some IFR flights are under radar surveillance and controlled from departure to arrival, and others rely entirely on pilot navigation.

Where ATC has no jurisdiction, it does not issue an IFR clearance. It has no control over the flight, nor does the pilot have any assurance of separation from other traffic.

ATC Reports

All pilots are required to report unforecast weather conditions or other information related to safety of flight to ATC. The pilot-in-command of each aircraft operated in controlled airspace under IFR shall report as soon as practical to ATC any malfunctions of navigational, approach, or communication equipment occurring in flight:

1. Loss of VOR, tactical air navigation (TACAN) or automatic direction finder (ADF) receiver capability.
2. Complete or partial loss of instrument landing system (ILS) receiver capability.
3. Impairment of air-to-ground communications capability.

The pilot-in-command shall include within the report (1) Aircraft identification, (2) Equipment affected, (3) Degree to which the pilot to operate under IFR within the ATC system is impaired, and (4) Nature and extent of assistance desired from ATC.

Position Reports

Position reports are required over each compulsory reporting point (shown on the chart as a solid triangle) along the route being flown regardless of altitude, including those with a VFR-on-top clearance. Along direct routes, reports are required of all IFR flights over each point used to define the route of flight. Reports at reporting points (shown as an open triangle) are made only when requested by ATC. A pilot should discontinue position reporting over designated reporting points when informed by ATC that the aircraft is in "RADAR CONTACT." Position reporting should be resumed when ATC advises "RADAR CONTACT LOST" or "RADAR SERVICE TERMINATED."

Position reports should include the following items:

1. Identification
2. Position
3. Time
4. Altitude or flight level (include actual altitude or flight level when operating on a clearance specifying VFR-on-top)
5. Type of flight plan (not required in IFR position reports made directly to ARTCCs or approach control)
6. ETA and name of next reporting point
7. The name only of the next succeeding reporting point along the route of flight
8. Pertinent remarks

En route position reports are submitted normally to the ARTCC controllers via direct controller-to-pilot communications channels, using the appropriate ARTCC frequencies listed on the en route chart.

Whenever an initial contact with a controller is to be followed by a position report, the name of the reporting point should be included in the call-up. This alerts the controller that such information is forthcoming. For example:

> "Atlanta Center, Cessna 1230 Alpha at JAILS intersection."
>
> "Cessna 1230 Alpha Atlanta Center."
>
> "Atlanta Center, Cessna 1230 Alpha at JAILS intersection, 5,000, estimating Monroeville at 1730."

Additional Reports

In addition to required position reports, the following reports should be made to ATC without a specific request.

1. At all times:
    a) When vacating any previously assigned altitude or flight level for a newly assigned altitude or flight level
    b) When an altitude change will be made if operating on a clearance specifying VFR-on-top
    c) When unable to climb/descend at a rate of at least 500 feet per minute (fpm)
    d) When an approach has been missed (Request clearance for specific action (to alternative airport, another approach, etc.))
    e) Change in average true airspeed (at cruising altitude) when it varies by 5 percent or ten knots (whichever is greater) from that filed in the flight plan
    f) The time and altitude upon reaching a holding fix or point to which cleared
    g) When leaving any assigned holding fix or point

       NOTE - The reports in (f) and (g) may be omitted by pilots of aircraft involved in instrument training at military terminal area facilities when radar service is being provided.
    h) Any loss in controlled airspace of VOR, TACAN, ADF, low frequency navigation receiver capability, GPS anomalies while using installed IFR-certified GPS/GNSS receivers, complete or partial loss of ILS receiver capability, or impairment of air/ground communications capability. Reports should include aircraft identification, equipment affected, degree to which the capability to operate under IFR in the ATC system is impaired, and the nature and extent of assistance desired from ATC.
    i) Any information relating to the safety of flight.
2. When not in radar contact:
    a) When leaving the final approach fix inbound on final approach (nonprecision approach), or when leaving the outer marker or fix used in lieu of the outer marker inbound on final approach (precision approach).
    b) A corrected estimate at any time it becomes apparent that an estimate as previously submitted is in error in excess of 3 minutes.

Any pilot who encounters weather conditions that have not been forecast, or hazardous conditions which have been forecast, is expected to forward a report of such weather to ATC.

Planning the Descent and Approach

ATC arrival procedures and flight deck workload are affected by weather conditions, traffic density, aircraft equipment, and radar availability.

When landing at an airport with approach control services and where two or more IAPs are published, information on the type of approach to expect will be provided in advance of arrival or vectors will be provided to a visual approach. This information will be broadcast either on automated terminal information service (ATIS) or by a controller. It will not be furnished when the visibility is 3 miles or more and the ceiling is at or above the highest initial approach altitude established for any low altitude IAP for the airport.

The purpose of this information is to help the pilot plan arrival actions; however, it is not an ATC clearance or commitment and is subject to change. Fluctuating weather, shifting winds, blocked runway, etc., are conditions that may result in changes to the approach information previously received. It is important for a pilot to advise ATC immediately if he or she is unable to execute the approach or prefers another type of approach.

If the destination is an airport without an operating control tower, and has automated weather data with broadcast capability, the pilot should monitor the automated surface observing system/automated weather observing system (ASOS/AWOS) frequency to ascertain the current weather for the airport. ATC should be advised that weather information has been received and what the pilot's intentions are.

When the approach to be executed has been determined, the pilot should plan for and request a descent to the appropriate altitude prior to the initial approach fix (IAF) or transition route depicted on the IAP. When flying the transition route, a pilot should maintain the last assigned altitude until ATC gives the instructions "cleared for the approach." Lower altitudes can be requested to bring the transition route altitude closer to the required altitude at the initial approach fix. When ATC uses the phrase "at pilot's discretion" in the altitude information of a clearance, the pilot has the option to start a descent at any rate, and may level off temporarily at any intermediate altitude. However, once an altitude has been vacated, return to that altitude is not authorized without a clearance. When ATC has not used the term "at pilot's discretion" nor imposed any descent restrictions, initiate descent promptly upon acknowledgment of the clearance.

Descend at an optimum rate (consistent with the operating characteristics of the aircraft) to 1,000 feet above the assigned altitude. Then attempt to descend at a rate of between 500 and

1,500 fpm until the assigned altitude is reached. If at anytime a pilot is unable to maintain a descent rate of at least 500 fpm, advise ATC. Also advise ATC if it is necessary to level off at an intermediate altitude during descent. An exception to this is when leveling off at 10,000 feet mean sea level (MSL) on descent, or 2,500 feet above airport elevation (prior to entering a Class B, Class C, or Class D surface area) when required for speed reduction.

Standard Terminal Arrival Routes (STARs)

Standard terminal arrival routes (as described in Chapter 8) have been established to simplify clearance delivery procedures for arriving aircraft at certain areas having high density traffic. A STAR serves a purpose parallel to that of a DP for departing traffic. *[Figure 10-3]* The following points regarding STARs are important to remember:

1. All STARs are contained in the TPP, along with the IAP charts for the destination airport. The AIM also describes STAR procedures.
2. If the destination is a location for which STARs have been published, a pilot may be issued a clearance containing a STAR whenever ATC deems it appropriate. To accept the clearance, a pilot must possess at least the approved textual description.
3. It is the pilot's responsibility to either accept or refuse an issued STAR. If a STAR will not or cannot be used, advise ATC by placing "NO STAR" in the remarks section of the filed flight plan or by advising ATC.
4. If a STAR is accepted in a clearance, compliance is mandatory.

Substitutes for Inoperative or Unusable Components

The basic ground components of an ILS are the localizer, glide slope, outer marker, middle marker, and inner marker (when installed). A compass locator or precision radar may be substituted for the outer or middle marker. Distance measuring equipment (DME), VOR, or nondirectional beacon (NDB) fixes authorized in the standard IAP or surveillance radar may be substituted for the outer marker.

Additionally, IFR-certified global positioning system (GPS) equipment, operated in accordance with Advisory Circular (AC) 90-94, Guidelines for Using Global Positioning System Equipment for IFR En Route and Terminal Operations and for Nonprecision Instrument Approaches in the United States National Airspace System, may be substituted for ADF and DME equipment, except when flying NDB IAP. Specifically, GPS can be substituted for ADF and DME equipment when:

1. Flying a DME arc;
2. Navigating TO/FROM an NDB;
3. Determining the aircraft position over an NDB;
4. Determining the aircraft position over a fix made up of a crossing NDB bearing;
5. Holding over an NDB;
6. Determining aircraft position over a DME fix.

Holding Procedures

Depending upon traffic and weather conditions, holding may be required. Holding is a predetermined maneuver which keeps aircraft within a specified airspace while awaiting further clearance from ATC. A standard holding pattern uses right turns, and a nonstandard holding pattern uses left turns. The ATC clearance will always specify left turns when a nonstandard pattern is to be flown.

Standard Holding Pattern (No Wind)

In a standard holding pattern with no winds, *[Figure 10-4]* the aircraft follows the specified course inbound to the holding fix, turns 180° to the right, flies a parallel straight course outbound for 1 minute, turns 180° to the right, and flies the inbound course to the fix.

Standard Holding Pattern (With Wind)

A standard symmetrical holding pattern cannot be flown when winds exist. In those situations, the pilot is expected to:

1. Compensate for the effect of a known wind except when turning.
2. Adjust outbound timing to achieve a 1-minute (1-1/2 minutes above 14,000 feet) inbound leg.

Figure 10-5 illustrates the holding track followed with a left crosswind. The effect of wind is counteracted by applying drift corrections to the inbound and outbound legs and by applying time allowances to the outbound leg.

Holding Instructions

If an aircraft arrives at a clearance limit before receiving clearance beyond the fix, ATC expects the pilot to maintain the last assigned altitude and begin holding in accordance with the charted holding pattern. If no holding pattern is charted and holding instructions have not been issued, enter a standard holding pattern on the course on which the aircraft approached the fix and request further clearance as soon as possible. Normally, when no delay is anticipated, ATC will issue holding instructions at least 5 minutes before the

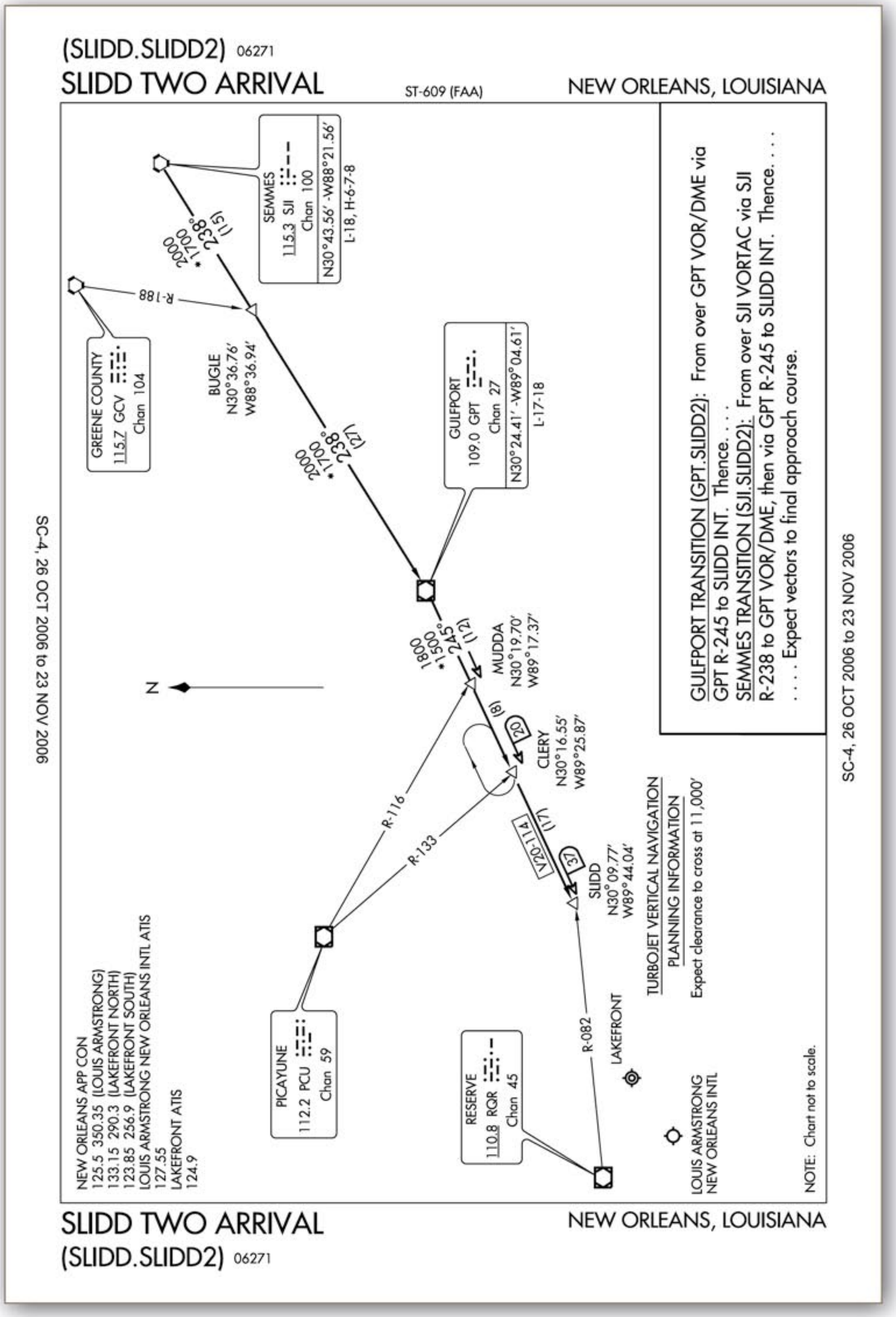
(SLIDD.SLIDD2) 06271
SLIDD TWO ARRIVAL
ST-609 (FAA)
NEW ORLEANS, LOUISIANA
SC-4, 26 OCT 2006 to 23 NOV 2006
GULFPORT TRANSITION (GPT.SLIDD2): From over GPT VOR/DME via GPT R-245 to SLIDD INT. Thence. . . .
SEMMES TRANSITION (SJI.SLIDD2): From over SJI VORTAC via SJI R-238 to GPT VOR/DME, then via GPT R-245 to SLIDD INT. Thence. . . .
. . . . Expect vectors to final approach course.
SEMMES
115.3 SJI
Chan 100
N30°43.56' -W88°21.56'
L-18, H-6-7-8
2000
*1700
238°
(15)
R-188
GREENE COUNTY
115.7 GCV
Chan 104
BUGLE
N30°36.76'
W88°36.94'
2000
*1700
238°
(27)
GULFPORT
109.0 GPT
Chan 27
N30°24.41' -W89°04.61'
L-17-18
N
1800
*1500
245°
(12)
MUDDA
N30°19.70'
W89°17.37'
(8)
20
CLERY
N30°16.55'
W89°25.87'
R-116
R-133
V20-114
(17)
37
SLIDD
N30°09.77'
W89°44.04'
TURBOJET VERTICAL NAVIGATION
PLANNING INFORMATION
Expect clearance to cross at 11,000'
R-082
LAKEFRONT
PICAYUNE
112.2 PCU
Chan 59
RESERVE
110.8 RQR
Chan 45
LOUIS ARMSTRONG
NEW ORLEANS INTL
NEW ORLEANS APP CON
125.5 350.35 (LOUIS ARMSTRONG)
133.15 290.3 (LAKEFRONT NORTH)
123.85 256.9 (LAKEFRONT SOUTH)
LOUIS ARMSTRONG NEW ORLEANS INTL ATIS
127.55
LAKEFRONT ATIS
124.9
NOTE: Chart not to scale.
SLIDD TWO ARRIVAL
(SLIDD.SLIDD2) 06271
NEW ORLEANS, LOUISIANA

Figure 10-3. *Standard Terminal Arrival Route (STAR).*

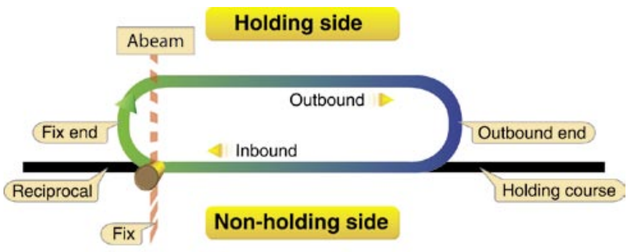

Standard pattern: Right turns (illustrated)
Non-standard pattern: Left turns

Figure 10-4. *Standard Holding Pattern—No Wind.*

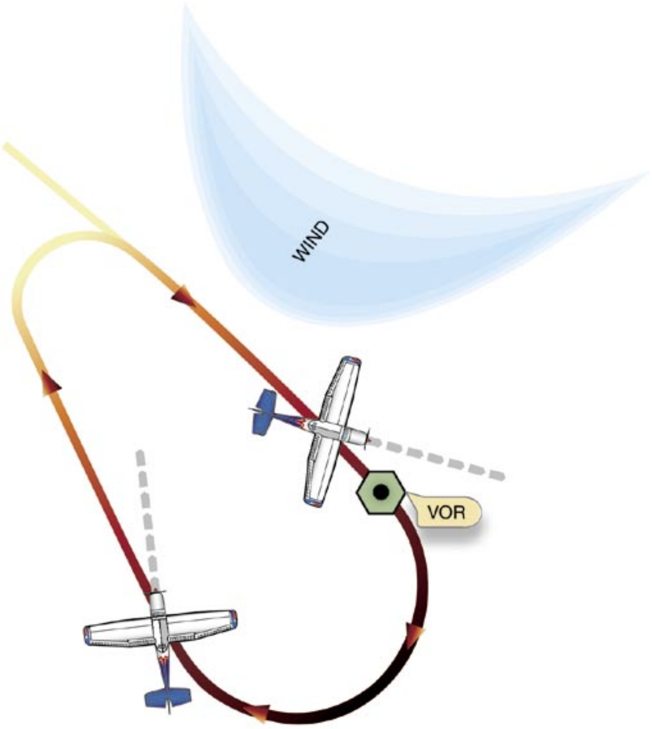

Figure 10-5. *Drift Correction in Holding Pattern.*

estimated arrival at the fix. Where a holding pattern is not charted, the ATC clearance will specify the following:

1. Direction of holding from the fix in terms of the eight cardinal compass points (N, NE, E, SE, etc.)
2. Holding fix (the fix may be omitted if included at the beginning of the transmission as the clearance limit)
3. Radial, course, bearing, airway, or route on which the aircraft is to hold.
4. Leg length in miles if DME or area navigation (RNAV) is to be used (leg length will be specified in minutes on pilot request or if the controller considers it necessary).
5. Direction of turn, if left turns are to be made, because the pilot requests or the controller considers it necessary.
6. Time to expect-further-clearance (EFC) and any pertinent additional delay information.

ATC instructions will also be issued whenever:

1. It is determined that a delay will exceed 1 hour.
2. A revised EFC is necessary.
3. In a terminal area having a number of navigation aids and approach procedures, a clearance limit may not indicate clearly which approach procedures will be used. On initial contact, or as soon as possible thereafter, approach control will advise the pilot of the type of approach to expect.
4. Ceiling and/or visibility is reported as being at or below the highest "circling minimums" established for the airport concerned. ATC will transmit a report of current weather conditions and subsequent changes, as necessary.
5. An aircraft is holding while awaiting approach clearance, and the pilot advises ATC that reported weather conditions are below minimums applicable to the operation. In this event, ATC will issue suitable instructions to aircraft desiring either to continue holding while awaiting weather improvement or proceed to another airport.

Standard Entry Procedures

The entry procedures given in the AIM evolved from extensive experimentation under a wide range of operational conditions. The standardized procedures should be followed to ensure that an aircraft remains within the boundaries of the prescribed holding airspace.

When a speed reduction is required, start the reduction when 3 minutes or less from the holding fix. Cross the holding fix initially at or below the maximum holding airspeed (MHA). The purpose of the speed reduction is to prevent overshooting the holding airspace limits, especially at locations where adjacent holding patterns are close together.

All aircraft may hold at the following altitudes and maximum holding airspeeds:

Altitude Mean Sea Level (MSL)	*Airspeed (KIAS)*
Up to 6,000 feet	200
6,001 – 14,000 feet	230
14,001 feet and above	265

The following are exceptions to the maximum holding airspeeds:

1. Holding patterns from 6,001 to 14,000 feet may be restricted to a maximum airspeed of 210 knots indicated airspeed (KIAS). This nonstandard pattern is depicted by an icon.
2. Holding patterns may be restricted to a maximum airspeed of 175 KIAS. This nonstandard pattern is depicted by an icon. Holding patterns restricted to 175 KIAS are generally found on IAPs applicable to category A and B aircraft only.
3. Holding patterns at Air Force airfields only—310 KIAS maximum, unless otherwise depicted.
4. Holding patterns at Navy airfields only—230 KIAS maximum, unless otherwise depicted.
5. The pilot of an aircraft unable to comply with maximum airspeed restrictions should notify ATC.

While other entry procedures may enable the aircraft to enter the holding pattern and remain within protected airspace, the parallel, teardrop, and direct entries are the procedures for entry and holding recommended by the FAA. Additionally, paragraph 5-3-7 in the AIM should be reviewed. *[Figure 10-6]*

1. Parallel Procedure. When approaching the holding fix from anywhere in sector (a), fly to the fix. Afterwards, turn to a heading to parallel the holding course outbound. Fly outbound for 1 minute, turn in the direction of the holding pattern through more than 180°, and return to the holding fix or intercept the holding course inbound.
2. Teardrop Procedure. When approaching the holding fix from anywhere in sector (b), the teardrop entry procedure would be to fly to the fix, turn outbound to a heading for a 30° teardrop entry within the pattern (on the holding side) for a period of 1 minute, then turn in the direction of the holding pattern to intercept the inbound holding course.
3. Direct Entry Procedure. When approaching the holding fix from anywhere in sector (c), the direct entry procedure would be to fly directly to the fix and turn to follow the holding pattern.

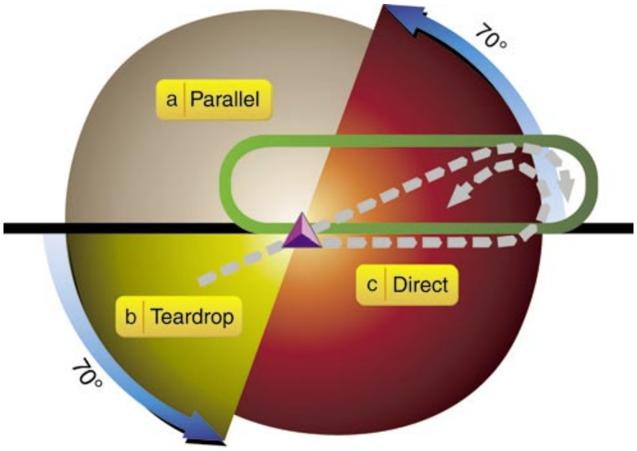

Figure 10-6. *Holding Pattern Entry Procedures.*

A pilot should make all turns during entry and while holding at:

1. 3° per second, or
2. 30° bank angle, or
3. A bank angle provided by a flight director system.

Time Factors

The holding pattern entry time reported to ATC is the initial time of arrival over the fix. Upon entering a holding pattern, the initial outbound leg is flown for 1 minute at or below 14,000 feet MSL, and for 1-1/2 minutes above 14,000 feet MSL. Timing for subsequent outbound legs should be adjusted as necessary to achieve proper inbound leg time. The pilot should begin outbound timing over or abeam the fix, whichever occurs later. If the abeam position cannot be determined, start timing when the turn to outbound is completed. *[Figure 10-7]*

Time leaving the holding fix must be known to ATC before succeeding aircraft can be cleared to the vacated airspace. Leave the holding fix:

1. When ATC issues either further clearance en route or approach clearance;
2. As prescribed in 14 CFR part 91 (for IFR operations; two-way radio communications failure, and responsibility and authority of the pilot-in-command); or
3. After the IFR flight plan has been cancelled, if the aircraft is holding in VFR conditions.

DME Holding

The same entry and holding procedures apply to DME holding, but distances (nautical miles) are used instead of time values. The length of the outbound leg will be specified by the controller, and the end of this leg is determined by the DME readout.

Approaches

Compliance With Published Standard Instrument Approach Procedures

Compliance with the approach procedures shown on the approach charts provides necessary navigation guidance information for alignment with the final approach courses,

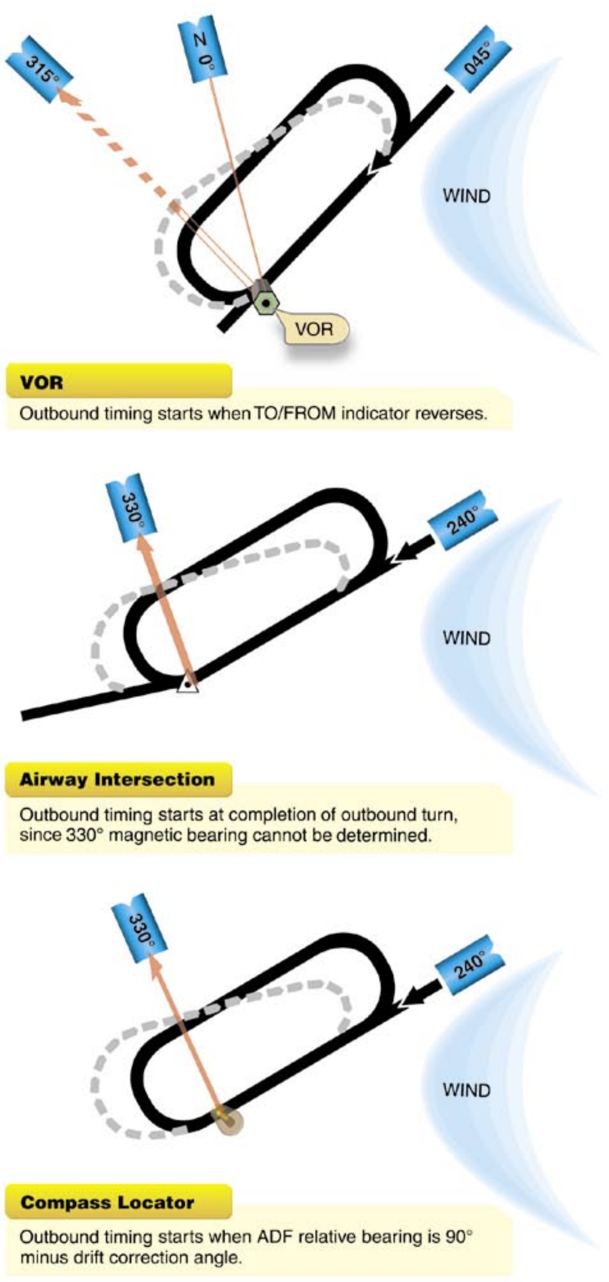
315°
N 0°
045°
WIND
VOR
VOR
Outbound timing starts when TO/FROM indicator reverses.
330°
240°
WIND
Airway Intersection
Outbound timing starts at completion of outbound turn, since 330° magnetic bearing cannot be determined.
330°
240°
WIND
Compass Locator
Outbound timing starts when ADF relative bearing is 90° minus drift correction angle.

Figure 10-7. *Holding—Outbound Timing.*

as well as obstruction clearance. Under certain conditions, a course reversal maneuver or procedure turn may be necessary. However, this procedure is not authorized when:

1. The symbol "NoPT" appears on the approach course on the plan view of the approach chart.
2. Radar vectoring is provided to the final approach course.
3. A holding pattern is published in lieu of a procedure turn.
4. Executing a timed approach from a holding fix.
5. Otherwise directed by ATC.

Instrument Approaches to Civil Airports

Unless otherwise authorized, when an instrument letdown to an airport is necessary, the pilot should use a standard IAP prescribed for that airport. IAPs are depicted on IAP charts and are found in the TPP.

ATC approach procedures depend upon the facilities available at the terminal area, the type of instrument approach executed, and the existing weather conditions. The ATC facilities, navigation aids (NAVAIDs), and associated frequencies appropriate to each standard instrument approach are given on the approach chart. Individual charts are published for standard approach procedures associated with the following types of facilities:

1. Nondirectional beacon (NDB)
2. Very-high frequency omnirange (VOR)
3. Very-high frequency omnirange with distance measuring equipment (VORTAC or VOR/DME)
4. Localizer (LOC)
5. Instrument landing system (ILS)
6. Localizer-type directional aid (LDA)
7. Simplified directional facility (SDF)
8. Area navigation (RNAV)
9. Global positioning system (GPS)

An IAP can be flown in one of two ways: as a full approach or with the assistance of radar vectors. When the IAP is flown as a full approach, pilots conduct their own navigation using the routes and altitudes depicted on the instrument approach chart. A full approach allows the pilot to transition from the en route phase, to the instrument approach, and then to a landing with minimal assistance from ATC. This type of procedure may be requested by the pilot but is most often used in areas without radar coverage. A full approach also provides the pilot with a means of completing an instrument approach in the event of a communications failure.

When an approach is flown with the assistance of radar vectors, ATC provides guidance in the form of headings and altitudes which position the aircraft to intercept the final approach. From this point, the pilot resumes navigation, intercepts the final approach course, and completes the approach using the IAP chart. This is often a more expedient method of flying the approach, as opposed to the full approach, and allows ATC to sequence arriving traffic. A pilot operating in radar contact can generally expect the assistance of radar vectors to the final approach course.

Approach to Airport Without an Operating Control Tower

Figure 10-8 shows an approach procedure at an airport without an operating control tower. When approaching such a facility, the pilot should monitor the AWOS/ASOS if available for the latest weather conditions. When direct communication between the pilot and controller is no longer required, the ARTCC or approach controller will issue a clearance for an instrument approach and advise "change to advisory frequency approved." When the aircraft arrives on a "cruise" clearance, ATC will not issue further clearance for approach and landing.

If an approach clearance is required, ATC will authorize the pilot to execute his or her choice of standard instrument approach (if more than one is published for the airport) with the phrase "Cleared for the approach" and the communications frequency change required, if any. From this point on, there will be no contact with ATC. The pilot is responsible for closing the IFR flight plan before landing, if in VFR conditions, or by telephone after landing.

Unless otherwise authorized by ATC, a pilot is expected to execute the complete IAP shown on the chart.

Approach to Airport With an Operating Tower, With No Approach Control

When an aircraft approaches an airport with an operating control tower, but no approach control, ATC will issue a clearance to an approach/outer fix with the appropriate information and instructions as follows:

1. Name of the fix
2. Altitude to be maintained
3. Holding information and expected approach clearance time, if appropriate
4. Instructions regarding further communications, including:
    a) facility to be contacted
    b) time and place of contact
    c) frequency/ies to be used

If ATIS is available, a pilot should monitor that frequency for information such as ceiling, visibility, wind direction and velocity, altimeter setting, instrument approach, and runways in use prior to initial radio contact with the tower. If ATIS is not available, ATC will provide weather information from the nearest reporting station.

Approach to an Airport With an Operating Tower, With an Approach Control

Where radar is approved for approach control service, it is used to provide vectors in conjunction with published IAPs. Radar vectors can provide course guidance and expedite traffic to the final approach course of any established IAP. *Figure 10-9* shows an IAP chart with maximum ATC facilities available.

Approach control facilities that provide this radar service operate in the following manner:

1. Arriving aircraft are either cleared to an outer fix most appropriate to the route being flown with vertical separation and, if required, given holding information; or,
2. When radar hand-offs are effected between ARTCC and approach control, or between two approach control facilities, aircraft are cleared to the airport, or to a fix so located that the hand-off will be completed prior to the time the aircraft reaches the fix.
    a) When the radar hand-offs are utilized, successive arriving flights may be handed off to approach control with radar separation in lieu of vertical separation.
    b) After hand-off to approach control, an aircraft is vectored to the appropriate final approach course.
3. Radar vectors and altitude/flight levels are issued as required for spacing and separating aircraft; do not deviate from the headings issued by approach control.
4. Aircraft are normally informed when it becomes necessary to be vectored across the final approach course for spacing or other reasons. If approach course crossing is imminent and the pilot has not been informed that the aircraft will be vectored across the final approach course, the pilot should query the controller. The pilot is not expected to turn inbound on the final approach course unless an approach clearance has been issued. This clearance is normally issued with the final vector for interception of the final approach course, and the vector enables the pilot to establish the aircraft on the final approach course prior to reaching the final approach fix.
5. Once the aircraft is established inbound on the final approach course, radar separation is maintained with other aircraft, and the pilot is expected to complete the approach using the NAVAID designated in the clearance (ILS, VOR, NDB, GPS, etc.) as the primary means of navigation.

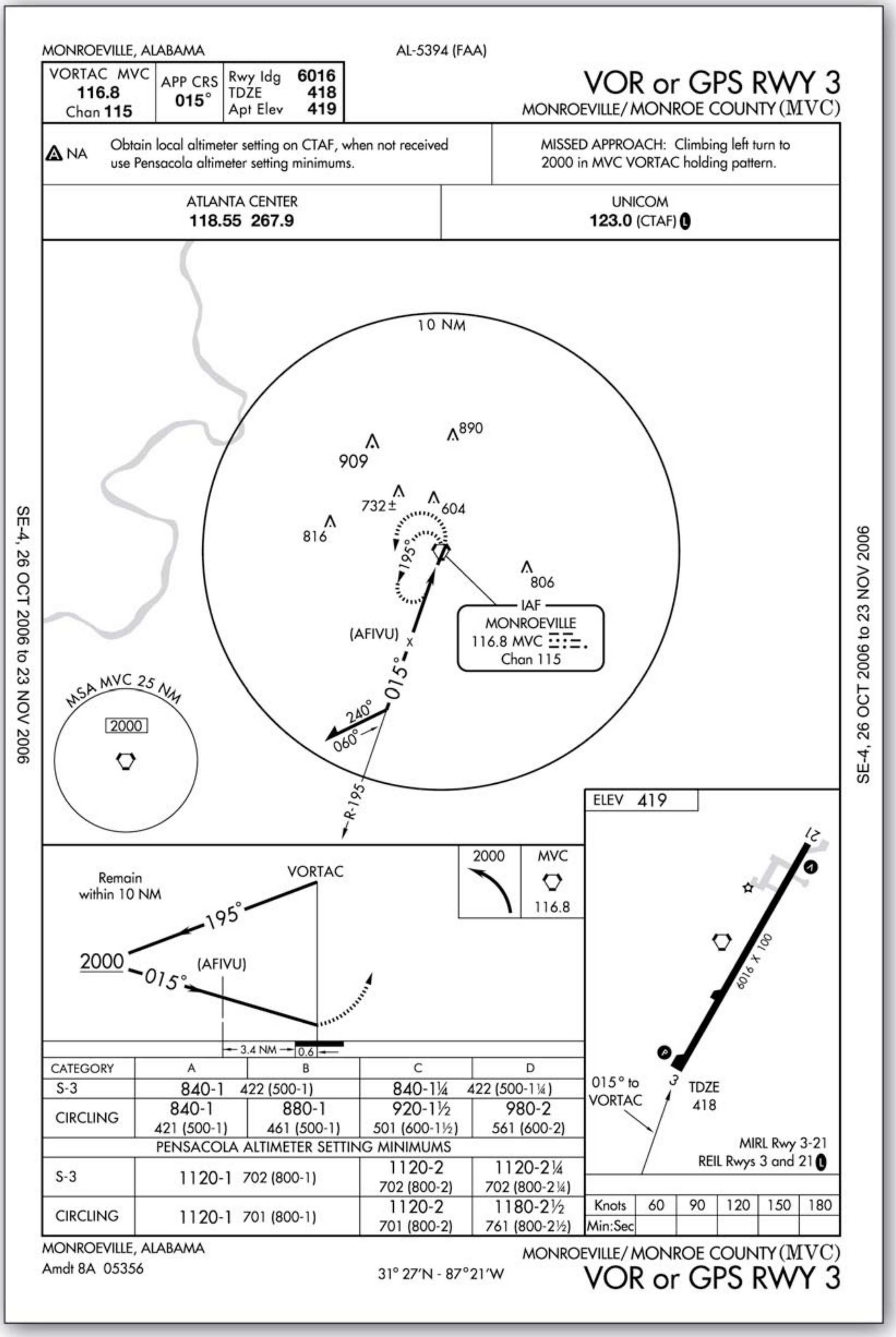
MONROEVILLE, ALABAMA
AL-5394 (FAA)
VORTAC MVC
116.8
Chan 115
APP CRS
015°
Rwy ldg 6016
TDZE 418
Apt Elev 419
VOR or GPS RWY 3
MONROEVILLE/ MONROE COUNTY (MVC)
A NA
Obtain local altimeter setting on CTAF, when not received use Pensacola altimeter setting minimums.
MISSED APPROACH: Climbing left turn to 2000 in MVC VORTAC holding pattern.
ATLANTA CENTER
118.55 267.9
UNICOM
123.0 (CTAF) L
10 NM
890
909
732±
604
816
195°
806
IAF
MONROEVILLE
116.8 MVC
Chan 115
(AFIVU)
015°
240°
060°
R-195
MSA MVC 25 NM
2000
SE-4, 26 OCT 2006 to 23 NOV 2006
SE-4, 26 OCT 2006 to 23 NOV 2006
ELEV 419
21
A
6016 X 100
P
3
TDZE
418
015° to VORTAC
MIRL Rwy 3-21
REIL Rwys 3 and 21 L
Remain within 10 NM
VORTAC
195°
2000
015°
(AFIVU)
3.4 NM
0.6
2000
MVC
116.8

CATEGORY	A	B	C	D
S-3	840-1 422 (500-1)		840-1¼ 422 (500-1¼)	
CIRCLING	840-1 421 (500-1)	880-1 461 (500-1)	920-1½ 501 (600-1½)	980-2 561 (600-2)
PENSACOLA ALTIMETER SETTING MINIMUMS				
S-3	1120-1 702 (800-1)		1120-2 702 (800-2)	1120-2¼ 702 (800-2¼)
CIRCLING	1120-1 701 (800-1)		1120-2 701 (800-2)	1180-2½ 761 (800-2½)

Knots	60	90	120	150	180
Min:Sec					

MONROEVILLE, ALABAMA
Amdt 8A 05356
31° 27′N - 87°21′W
MONROEVILLE/ MONROE COUNTY (MVC)
VOR or GPS RWY 3

Figure 10-8. *Monroeville, AL (MVC) VOR or GPS Rwy 3 Approach: An Approach Procedure at an Airport Without an Operating Control Tower.*

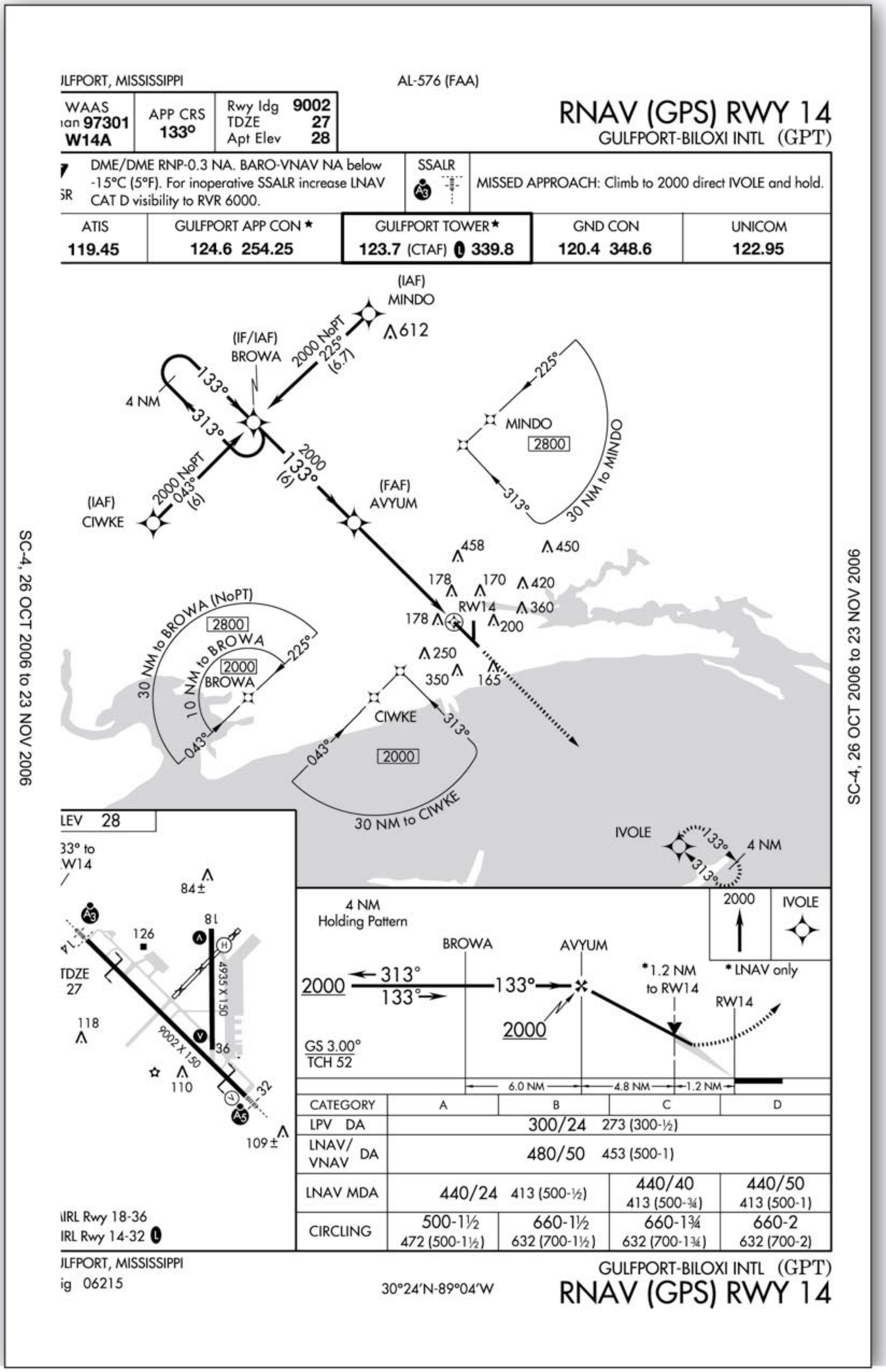

JLFPORT, MISSISSIPPI AL-576 (FAA)

WAAS ıan 97301 W14A	APP CRS 133°	Rwy ldg 9002 TDZE 27 Apt Elev 28	RNAV (GPS) RWY 14 GULFPORT-BILOXI INTL (GPT)

DME/DME RNP-0.3 NA. BARO-VNAV NA below -15°C (5°F). For inoperative SSALR increase LNAV CAT D visibility to RVR 6000.

SSALR

MISSED APPROACH: Climb to 2000 direct IVOLE and hold.

ATIS	GULFPORT APP CON ★	GULFPORT TOWER ★	GND CON	UNICOM
119.45	124.6 254.25	123.7 (CTAF) Ⓛ 339.8	120.4 348.6	122.95

CATEGORY	A	B	C	D
LPV DA	300/24 273 (300-½)			
LNAV/ VNAV DA	480/50 453 (500-1)			
LNAV MDA	440/24 413 (500-½)		440/40 413 (500-¾)	440/50 413 (500-1)
CIRCLING	500-1½ 472 (500-1½)	660-1½ 632 (700-1½)	660-1¾ 632 (700-1¾)	660-2 632 (700-2)

IRL Rwy 18-36
IRL Rwy 14-32 Ⓛ

JLFPORT, MISSISSIPPI GULFPORT-BILOXI INTL (GPT)
ig 06215 30°24′N-89°04′W RNAV (GPS) RWY 14

SC-4, 26 OCT 2006 to 23 NOV 2006
(IAF) MINDO
612
(IF/IAF) BROWA
2000 NoPT 225° (6.7)
4 NM
133°
313°
2000 NoPT 043° (6)
(IAF) CIWKE
2000 133° (6)
(FAF) AVYUM
MINDO
2800
225°
313°
30 NM to MINDO
458
450
178
170
420
RW14
360
200
250
350
165
30 NM to BROWA (NoPT)
2800
10 NM to BROWA
2000
BROWA
225°
043°
CIWKE
313°
2000
30 NM to CIWKE
IVOLE
133°
313°
4 NM
ELEV 28
133° to RW14
84±
TDZE 27
126
18
4935 X 150
118
9002 X 150
36
110
32
109±
4 NM Holding Pattern
BROWA
AVYUM
2000
313°
133°
*1.2 NM to RW14
RW14
GS 3.00°
TCH 52
6.0 NM
4.8 NM
1.2 NM
2000
IVOLE
* LNAV only

Figure 10-9. *Gulfport, MS (GPT) ILS or LOC Rwy 14 Approach: An Instrument Procedure Chart With Maximum ATC Facilities Available.*

6. After passing the final approach fix inbound, the pilot is expected to proceed direct to the airport and complete the approach, or to execute the published missed approach procedure.
7. Radar service is automatically terminated when the landing is completed or when the pilot is instructed to change to advisory frequency at uncontrolled airports, whichever occurs first.

Radar Approaches

With a radar approach, the pilot receives course and altitude guidance from a controller who monitors the progress of the flight with radar. This is an option should the pilot experience an emergency or distress situation.

The only airborne radio equipment required for radar approaches is a functioning radio transmitter and receiver. The radar controller vectors the aircraft to align it with the runway centerline. The controller continues the vectors to keep the aircraft on course until the pilot can complete the approach and landing by visual reference to the surface. There are two types of radar approaches: Precision (PAR) and Surveillance (ASR).

A radar approach may be given to any aircraft upon request and may be offered to pilots of aircraft in distress or to expedite traffic; however, an ASR might not be approved unless there is an ATC operational requirement, or in an unusual or emergency situation. Acceptance of a PAR or ASR by a pilot does not waive the prescribed weather minimums for the airport or for the particular aircraft operator concerned. The decision to make a radar approach when the reported weather is below the established minimums rests with the pilot.

PAR and ASR minimums are published on separate pages in the FAA Terminal Procedures Publication (TPP). *Figure 10-10*.

Precision Approach (PAR) is one in which a controller provides highly accurate navigational guidance in azimuth and elevation to a pilot.

The controller gives the pilot headings to fly that direct the aircraft to, and keep the aircraft aligned with, the extended centerline of the landing runway. The pilot is told to anticipate glide path interception approximately 10 to 30 seconds before it occurs and when to start descent. The published decision height (DH) will be given only if the pilot requests it. If the aircraft is observed to deviate above or below the glide path, the pilot is given the relative amount of deviation by use of terms "slightly" or "well" and is expected to adjust the aircraft's rate of descent/ascent to return to the glide path. Trend information is also issued with respect to the elevation of the aircraft and may be modified by the terms "rapidly" and "slowly"; e.g., "well above glide path, coming down rapidly."

Range from touchdown is given at least once each mile. If an aircraft is observed by the controller to proceed outside of specified safety zone limits in azimuth and/or elevation and continue to operate outside these prescribed limits, the pilot will be directed to execute a missed approach or to fly a

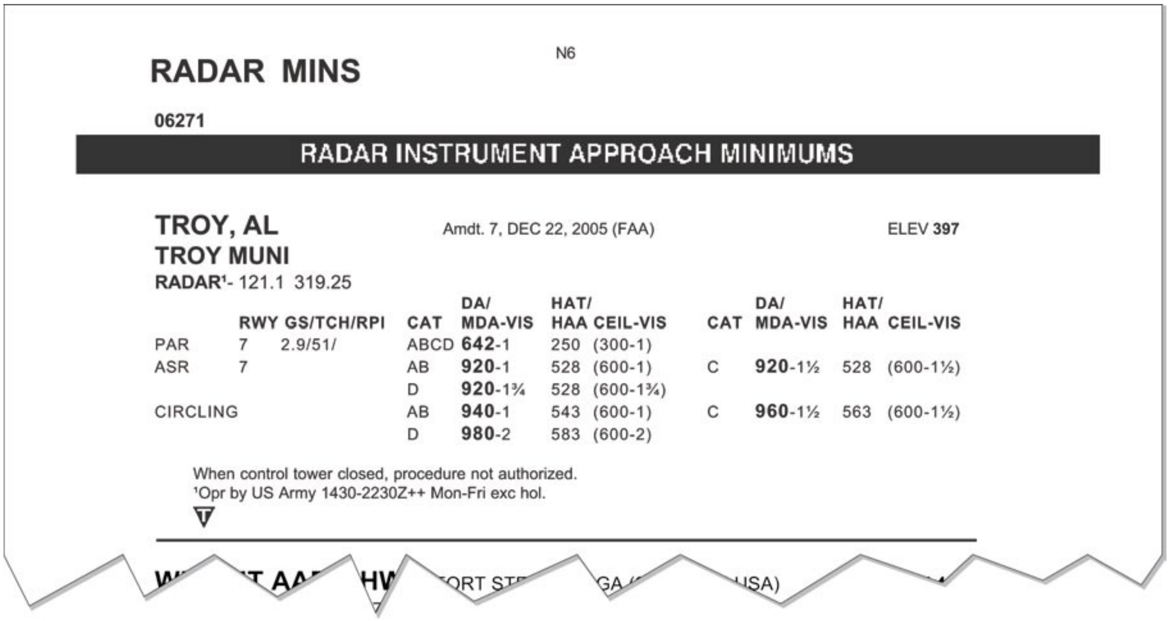

RADAR MINS

N6

06271

RADAR INSTRUMENT APPROACH MINIMUMS

TROY, AL Amdt. 7, DEC 22, 2005 (FAA) ELEV **397**

TROY MUNI

RADAR[1]- 121.1 319.25

	RWY	GS/TCH/RPI	CAT	DA/ MDA-VIS	HAT/ HAA	CEIL-VIS	CAT	DA/ MDA-VIS	HAT/ HAA	CEIL-VIS
PAR	7	2.9/51/	ABCD	**642**-1	250	(300-1)				
ASR	7		AB	**920**-1	528	(600-1)	C	**920**-1½	528	(600-1½)
			D	**920**-1¾	528	(600-1¾)				
CIRCLING			AB	**940**-1	543	(600-1)	C	**960**-1½	563	(600-1½)
			D	**980**-2	583	(600-2)				

When control tower closed, procedure not authorized.

[1]Opr by US Army 1430-2230Z++ Mon-Fri exc hol.

Figure 10-10. *Radar Instrument Approach Minimums for Troy, AL.*

specified course unless the pilot has the runway environment (runway, approach lights, etc.) in sight. Navigational guidance in azimuth and elevation is provided to the pilot until the aircraft reaches the published DH. Advisory course and glide path information is furnished by the controller until the aircraft passes over the landing threshold. At this point the pilot is advised of any deviation from the runway centerline. Radar service is automatically terminated upon completion of the approach.

Surveillance Approach (ASR) is one in which a controller provides navigational guidance in azimuth only.

The controller furnishes the pilot with headings to fly to align the aircraft with the extended centerline of the landing runway. Since the radar information used for a surveillance approach is considerably less precise than that used for a precision approach, the accuracy of the approach will not be as great and higher minimums will apply. Guidance in elevation is not possible but the pilot will be advised when to commence descent to the Minimum Descent Altitude (MDA) or, if appropriate, to an intermediate step-down fix Minimum Crossing Altitude and subsequently to the prescribed MDA. In addition, the pilot will be advised of the location of the Missed Approach Point (MAP) prescribed for the procedure and the aircraft's position each mile on final from the runway, airport or heliport or MAP, as appropriate.

If requested by the pilot, recommended altitudes will be issued at each mile, based on the descent gradient established for the procedure, down to the last mile that is at or above the MDA. Normally, navigational guidance will be provided until the aircraft reaches the MAP.

Radar service is automatically terminated at the completion of a radar approach.

No-Gyro Approach is available to a pilot under radar control who experiences circumstances wherein the directional gyro or other stabilized compass is inoperative or inaccurate. When this occurs, the pilot should so advise ATC and request a no-gyro vector or approach. The pilot of an aircraft not equipped with a directional gyro or other stabilized compass who desires radar handling may also request a no-gyro vector or approach. The pilot should make all turns at standard rate and should execute the turn immediately upon receipt of instructions. For example, "TURN RIGHT," "STOP TURN." When a surveillance or precision approach is made, the pilot will be advised after the aircraft has been turned onto final approach to make turns at half standard rate.

Radar Monitoring of Instrument Approaches

PAR facilities operated by the FAA and the military services at some joint-use (civil and military) and military installations monitor aircraft on instrument approaches and issue radar advisories to the pilot when weather is below VFR minimums (1,000 and 3), at night, or when requested by a pilot. This service is provided only when the PAR Final Approach Course coincides with the final approach of the navigational aid and only during the operational hours of the PAR. The radar advisories serve only as a secondary aid since the pilot has selected the navigational aid as the primary aid for the approach.

Prior to starting final approach, the pilot will be advised of the frequency on which the advisories will be transmitted. If, for any reason, radar advisories cannot be furnished, the pilot will be so advised.

Advisory information, derived from radar observations, includes information on:

1. Passing the final approach fix inbound (nonprecision approach) or passing the outer marker or fix used in lieu of the outer marker inbound (precision approach).
2. Trend advisories with respect to elevation and/or azimuth radar position and movement will be provided.
3. If, after repeated advisories, the aircraft proceeds outside the PAR safety limit or if a radical deviation is observed, the pilot will be advised to execute a missed approach unless the prescribed visual reference with the surface is established.

Radar service is automatically terminated upon completion of the approach. *[Figure 10-11]*

Timed Approaches From a Holding Fix

Timed approaches from a holding fix are conducted when many aircraft are waiting for an approach clearance. Although the controller will not specifically state "timed approaches are in progress," the assigning of a time to depart the FAF inbound (nonprecision approach), or the outer marker or fix used in lieu of the outer marker inbound (precision approach), indicates that timed approach procedures are being utilized.

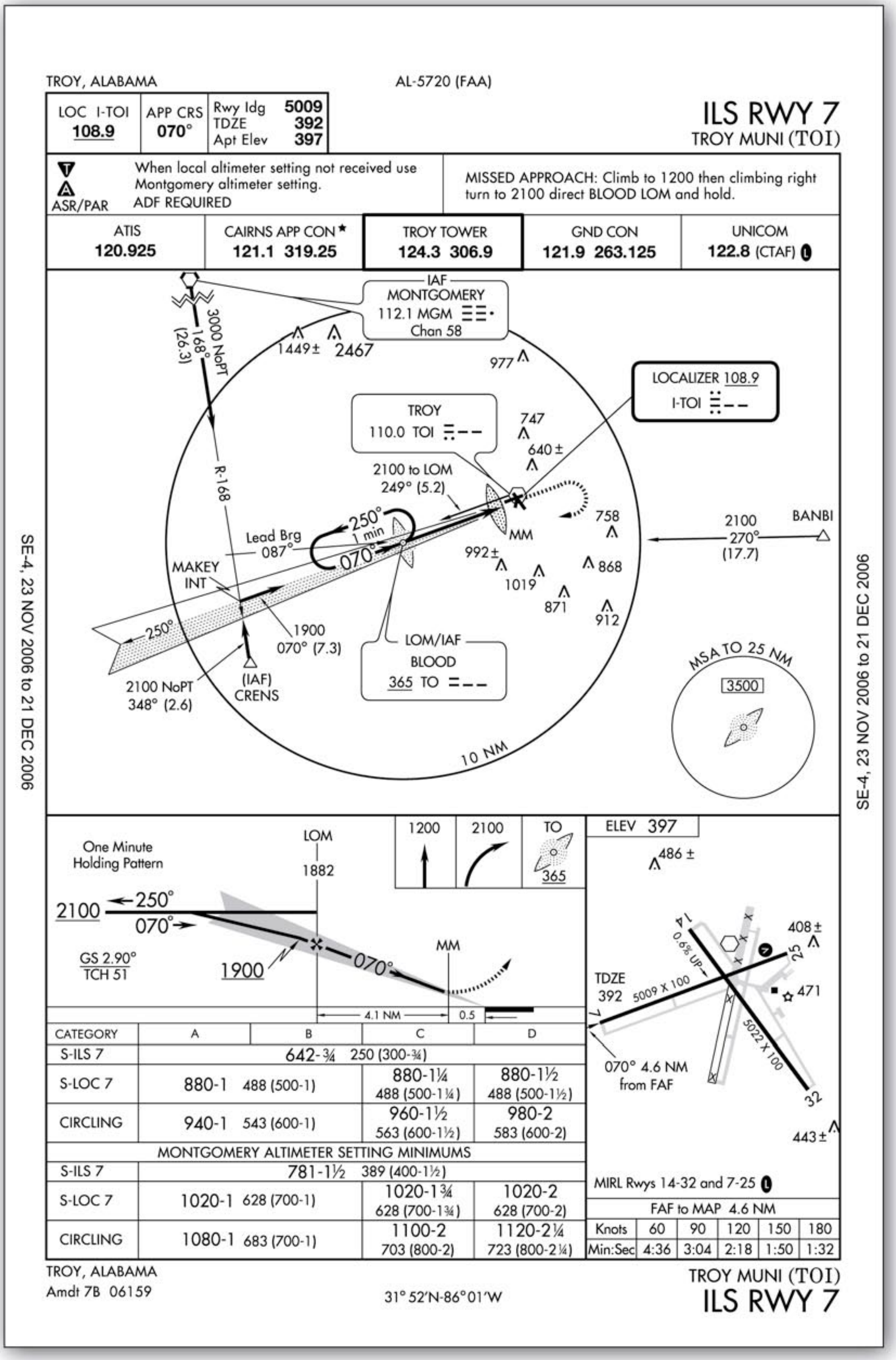

TROY, ALABAMA AL-5720 (FAA)

LOC I-TOI	APP CRS	Rwy ldg	5009
108.9	070°	TDZE	392
		Apt Elev	397

ILS RWY 7
TROY MUNI (TOI)

When local altimeter setting not received use Montgomery altimeter setting.
ASR/PAR ADF REQUIRED

MISSED APPROACH: Climb to 1200 then climbing right turn to 2100 direct BLOOD LOM and hold.

ATIS	CAIRNS APP CON ★	TROY TOWER	GND CON	UNICOM
120.925	121.1 319.25	124.3 306.9	121.9 263.125	122.8 (CTAF)

IAF
MONTGOMERY
112.1 MGM
Chan 58
3000 NoPT
168°
(26.3)
1449±
2467
977
LOCALIZER 108.9
I-TOI
TROY
110.0 TOI
747
640±
2100 to LOM
249° (5.2)
R-168
250°
1 min
Lead Brg
087°
070°
MM
758
2100
270°
(17.7)
BANBI
992±
868
MAKEY
INT
1019
871
912
250°
1900
070° (7.3)
LOM/IAF
BLOOD
365 TO
MSA TO 25 NM
3500
2100 NoPT
348° (2.6)
(IAF)
CRENS
10 NM
One Minute Holding Pattern
LOM
1882
1200
2100
TO
365
2100
250°
070°
GS 2.90°
TCH 51
1900
070°
MM
4.1 NM
0.5
ELEV 397
486±
14
0.6% UP
408±
25
TDZE
392
5009 X 100
471
7
5022 X 100
070° 4.6 NM from FAF
32
443±
MIRL Rwys 14-32 and 7-25
SE-4, 23 NOV 2006 to 21 DEC 2006
SE-4, 23 NOV 2006 to 21 DEC 2006

CATEGORY	A	B	C	D
S-ILS 7	642-¾ 250 (300-¾)			
S-LOC 7	880-1 488 (500-1)		880-1¼ 488 (500-1¼)	880-1½ 488 (500-1½)
CIRCLING	940-1 543 (600-1)		960-1½ 563 (600-1½)	980-2 583 (600-2)
MONTGOMERY ALTIMETER SETTING MINIMUMS				
S-ILS 7	781-1½ 389 (400-1½)			
S-LOC 7	1020-1 628 (700-1)		1020-1¾ 628 (700-1¾)	1020-2 628 (700-2)
CIRCLING	1080-1 683 (700-1)		1100-2 703 (800-2)	1120-2¼ 723 (800-2¼)

FAF to MAP 4.6 NM

Knots	60	90	120	150	180
Min:Sec	4:36	3:04	2:18	1:50	1:32

TROY, ALABAMA
Amdt 7B 06159

31°52'N-86°01'W

TROY MUNI (TOI)
ILS RWY 7

Figure 10-11. *ILS RWY 7 Troy, AL.*

In lieu of holding, the controller may use radar vectors to the final approach course to establish a distance between aircraft that will ensure the appropriate time sequence between the FAF and outer marker, or fix used in lieu of the outer marker and the airport. Each pilot in the approach sequence will be given advance notice of the time they should leave the holding point on approach to the airport. When a time to leave the holding point is received, the pilot should adjust the flight path in order to leave the fix as closely as possible to the designated time.

Timed approaches may be conducted when the following conditions are met:

1. A control tower is in operation at the airport where the approaches are conducted.
2. Direct communications are maintained between the pilot and the Center or approach controller until the pilot is instructed to contact the tower.
3. If more than one missed approach procedure is available, none require a course reversal.
4. If only one missed approach procedure is available, the following conditions are met:
   a) Course reversal is not required; and
   b) Reported ceiling and visibility are equal to or greater than the highest prescribed circling minimums for the IAP.
5. When cleared for the approach, pilots should not execute a procedure turn.

Approaches to Parallel Runways

Procedures permit ILS instrument approach operations to dual or triple parallel runway configurations. A parallel approach is an ATC procedure that permits parallel ILS approach to airports with parallel runways separated by at least 2,500 feet between centerlines. Wherever parallel approaches are in progress, pilots are informed that approaches to both runways are in use.

Simultaneous approaches are permitted to runways:

1. With centerlines separated by 4,300 to 9,000 feet;
2. That are equipped with final monitor controllers;
3. That require radar monitoring to ensure separation between aircraft on the adjacent parallel approach course.

The approach procedure chart will include the note "simultaneous approaches authorized RWYS 14L and 14R," identifying the appropriate runways. When advised that simultaneous parallel approaches are in progress, pilots must advise approach control immediately of malfunctioning or inoperative components.

Parallel approach operations demand heightened pilot situational awareness. The close proximity of adjacent aircraft conducting simultaneous parallel approaches mandates strict compliance with all ATC clearances and approach procedures. Pilots should pay particular attention to the following approach chart information: name and number of the approach, localizer frequency, inbound course, glide slope intercept altitude, DA/DH, missed approach instructions, special notes/procedures, and the assigned runway location and proximity to adjacent runways. Pilots also need to exercise strict radio discipline, which includes continuous monitoring of communications and the avoidance of lengthy, unnecessary radio transmissions.

Side-Step Maneuver

ATC may authorize a side-step maneuver to either one of two parallel runways that are separated by 1,200 feet or less, followed by a straight-in landing on the adjacent runway. Aircraft executing a side-step maneuver will be cleared for a specified nonprecision approach and landing on the adjacent parallel runway. For example, "Cleared ILS runway 7 left approach, side-step to runway 7 right." The pilot is expected to commence the side-step maneuver as soon as possible after the runway or runway environment is in sight. Landing minimums to the adjacent runway will be based on nonprecision criteria and therefore higher than the precision minimums to the primary runway, but will normally be lower than the published circling minimums.

Circling Approaches

Landing minimums listed on the approach chart under "CIRCLING" apply when it is necessary to circle the airport, maneuver for landing, or when no straight-in minimums are specified on the approach chart. *[Figure 10-11]*

The circling minimums published on the instrument approach chart provide a minimum of 300 feet of obstacle clearance in the circling area. *[Figure 10-12]* During a circling approach, the pilot should maintain visual contact with the runway of intended landing and fly no lower than the circling minimums until positioned to make a final descent for a landing. It is important to remember that circling minimums are only minimums. If the ceiling allows it, fly at an altitude that more nearly approximates VFR traffic pattern altitude. This will make any maneuvering safer and bring the view of the landing runway into a more normal perspective.

Figure 10-13 shows patterns that can be used for circling approaches. Pattern "A" can be flown when the final approach

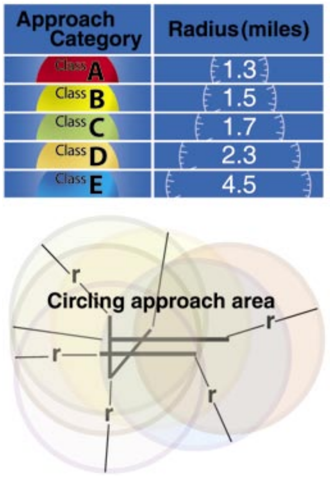

Approach Category	Radius (miles)
Class A	1.3
Class B	1.5
Class C	1.7
Class D	2.3
Class E	4.5

Defining size of areas, radii (r) vary with the approach category

Figure 10-12. *Circling Approach Area Radii.*

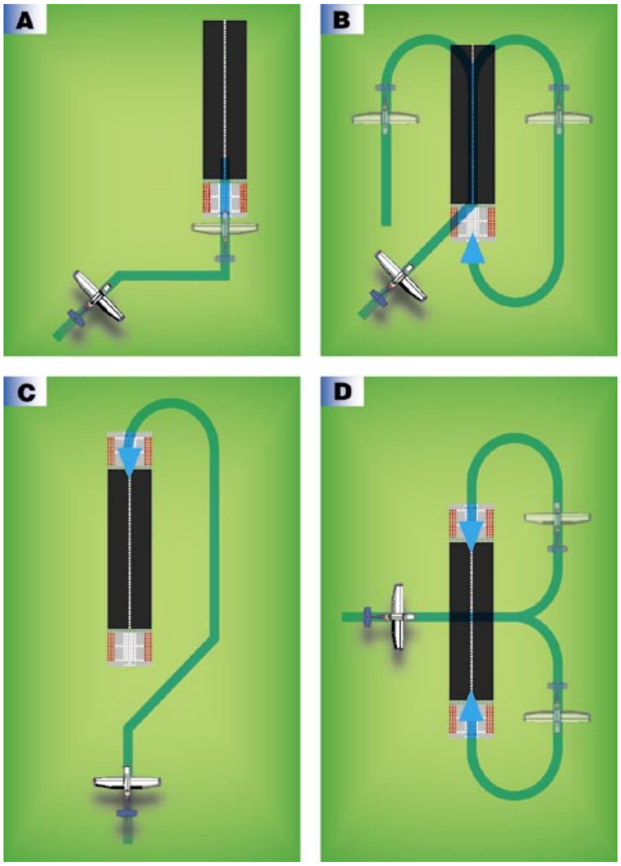

Figure 10-13. *Circling Approaches.*

course intersects the runway centerline at less than a 90° angle, and the runway is in sight early enough to establish a base leg. If the runway becomes visible too late to fly pattern "A," circle as shown in "B." Fly pattern "C" if it is desirable to land opposite the direction of the final approach, and the runway is sighted in time for a turn to downwind leg. If the runway is sighted too late for a turn to downwind, fly pattern "D." Regardless of the pattern flown, the pilot must maneuver the aircraft to remain within the designated circling area. Refer to section A ("Terms and Landing Minima Data") in the front of each TPP for a description of circling approach categories.

The criteria for determining the pattern to be flown are based on personal flying capabilities and knowledge of the performance characteristics of the aircraft. In each instance, the pilot must consider all factors: airport design, ceiling and visibility, wind direction and velocity, final approach course alignment, distance from the final approach fix to the runway, and ATC instructions.

IAP Minimums

Pilots may not operate an aircraft at any airport below the authorized MDA or continue an approach below the authorized DA/DH unless:

1. The aircraft is continuously in a position from which a descent to a landing on the intended runway can be made at a normal descent rate using normal maneuvers;
2. The flight visibility is not less than that prescribed for the approach procedure being used; and
3. At least one of the following visual references for the intended runway is visible and identifiable to the pilot:
    a) Approach light system
    b) Threshold
    c) Threshold markings
    d) Threshold lights
    e) Runway end identifier lights (REIL)
    f) Visual approach slope indicator (VASI)
    g) Touchdown zone or touchdown zone markings
    h) Touchdown zone lights
    i) Runway or runway markings
    j) Runway lights

Missed Approaches

A missed approach procedure is formulated for each published instrument approach and allows the pilot to return to the airway structure while remaining clear of obstacles. The procedure is shown on the approach chart in text and graphic form. Since the execution of a missed approach occurs when the flight deck workload is at a maximum, the procedure should be studied and mastered before beginning the approach.

When a missed approach procedure is initiated, a climb pitch attitude should be established while setting climb power. Configure the aircraft for climb, turn to the appropriate heading, advise ATC that a missed approach is being executed, and request further clearances.

If the missed approach is initiated prior to reaching the missed approach point (MAP), unless otherwise cleared by ATC, continue to fly the IAP as specified on the approach chart. Fly to the MAP at or above the MDA or DA/DH before beginning a turn.

If visual reference is lost while circling-to-land from an instrument approach, execute the appropriate missed approach procedure. Make the initial climbing turn toward the landing runway and then maneuver to intercept and fly the missed approach course.

Pilots should immediately execute the missed approach procedure:

1. Whenever the requirements for operating below DA/DH or MDA are not met when the aircraft is below MDA, or upon arrival at the MAP and at any time after that until touchdown;
2. Whenever an identifiable part of the airport is not visible to the pilot during a circling maneuver at or above MDA; or
3. When so directed by ATC.

Landing

According to 14 CFR part 91, no pilot may land when the flight visibility is less than the visibility prescribed in the standard IAP being used. ATC will provide the pilot with the current visibility reports appropriate to the runway in use. This may be in the form of prevailing visibility, runway visual value (RVV), or runway visual range (RVR). However, only the pilot can determine if the flight visibility meets the landing requirements indicated on the approach chart. If the flight visibility meets the minimum prescribed for the approach, then the approach may be continued to a landing. If the flight visibility is less than that prescribed for the approach, then the pilot must execute a missed approach, regardless of the reported visibility.

The landing minimums published on IAP charts are based on full operation of all components and visual aids associated with the instrument approach chart being used. Higher minimums are required with inoperative components or visual aids. For example, if the ALSF-1 approach lighting system were inoperative, the visibility minimums for an ILS would need to be increased by one-quarter mile. If more than one component is inoperative, each minimum is raised to the highest minimum required by any single component that is inoperative. ILS glide slope inoperative minimums are published on instrument approach charts as localizer minimums. Consult the "Inoperative Components or Visual Aids Table" (printed on the inside front cover of each TPP), for a complete description of the effect of inoperative components on approach minimums.

Instrument Weather Flying

Flying Experience

The more experience a pilot has in VFR and IFR flight, the more proficient a pilot becomes. VFR experience can be gained by flying in terminal areas with high traffic activity. This type of flying forces the pilot to polish the skill of dividing his or her attention between aircraft control, navigation, communications, and other flight deck duties. IFR experience can be gained through night flying which also promotes both instrument proficiency and confidence. The progression from flying at night under clear, moonlit conditions to flying at night without moonlight, natural horizon, or familiar landmarks teaches a pilot to trust the aircraft instruments with minimal dependence upon what can be seen outside the aircraft. It is a pilot's decision to proceed with an IFR flight or to wait for more acceptable weather conditions.

Recency of Experience

Currency as an instrument pilot is an equally important consideration. No person may act as pilot in command of an aircraft under IFR or in weather conditions less than VFR minimums unless he or she has met the requirements of part 91. Remember, these are minimum requirements.

Airborne Equipment and Ground Facilities

Regulations specify minimum equipment for filing an IFR flight plan. It is the pilot's responsibility to determine the adequacy of the aircraft and navigation/communication (NAV/COM) equipment for the proposed IFR flight. Performance limitations, accessories, and general condition of the equipment are directly related to the weather, route, altitude, and ground facilities pertinent to the flight, as well as to the flight deck workload.

Weather Conditions

In addition to the weather conditions that might affect a VFR flight, an IFR pilot must consider the effects of other weather phenomena (e.g., thunderstorms, turbulence, icing, and visibility).

Turbulence

Inflight turbulence can range from occasional light bumps to extreme airspeed and altitude variations that make aircraft control difficult. To reduce the risk factors associated with turbulence, pilots must learn methods of avoidance, as well as piloting techniques for dealing with an inadvertent encounter.

Turbulence avoidance begins with a thorough preflight weather briefing. Many reports and forecasts are available to assist the pilot in determining areas of potential turbulence. These include the Severe Weather Warning (WW), SIGMET (WS), Convective SIGMET (WST), AIRMET (WA), Severe Weather Outlook (AC), Center Weather Advisory (CWA), Area Forecast (FA), and Pilot Reports (UA or PIREPs). Since thunderstorms are always indicative of turbulence, areas of known and forecast thunderstorm activity will always be of interest to the pilot. In addition, clear air turbulence (CAT) associated with jet streams, strong winds over rough terrain, and fast moving cold fronts are good indicators of turbulence.

Pilots should be alert while in flight for the signposts of turbulence. For example, clouds with vertical development such as cumulus, towering cumulus, and cumulonimbus are indicators of atmospheric instability and possible turbulence. Standing lenticular clouds lack vertical development but indicate strong mountain wave turbulence. While en route, pilots can monitor hazardous inflight weather advisory service (HIWAS) broadcast for updated weather advisories, or contact the nearest AFSS or En Route Flight Advisory Service (EFAS) for the latest turbulence-related PIREPs.

To avoid turbulence associated with strong thunderstorms, circumnavigate cells by at least 20 miles. Turbulence may also be present in the clear air above a thunderstorm. To avoid this, fly at least 1,000 feet above the top for every 10 knots of wind at that level, or fly around the storm. Finally, do not underestimate the turbulence beneath a thunderstorm. Never attempt to fly under a thunderstorm. The possible results of turbulence and wind shear under the storm could be disastrous.

When moderate to severe turbulence is encountered, aircraft control is difficult, and a great deal of concentration is required to maintain an instrument scan. *[Figure 10-14]* Pilots should immediately reduce power and slow the aircraft to the recommended turbulence penetration speed as described in the POH/AFM. To minimize the load factor imposed on the aircraft, the wings should be kept level and the aircraft's pitch attitude should be held constant. The aircraft is allowed to fluctuate up and down, because maneuvering to maintain a constant altitude only increases the stress on the aircraft. If necessary, the pilot should advise ATC of the fluctuations and request a block altitude clearance. In addition, the power should remain constant at a setting that will maintain the recommended turbulence penetration airspeed.

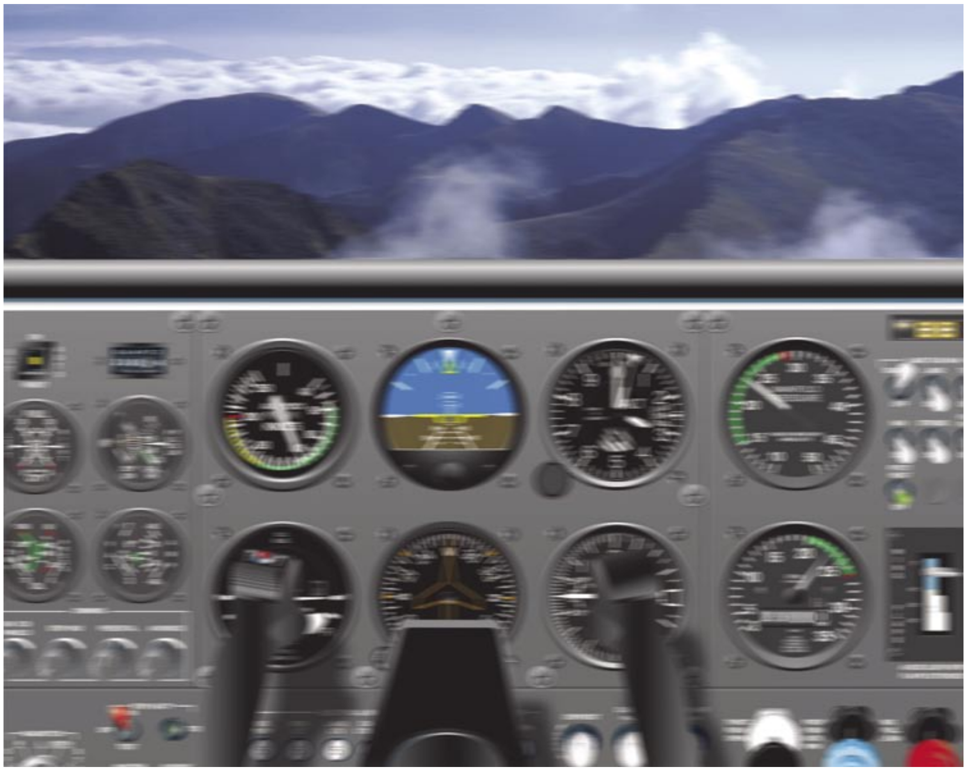

Figure 10-14. *Maintaining an instrument scan in severe turbulence can be difficult.*

The best source of information on the location and intensity of turbulence are PIREPs. Therefore, pilots are encouraged to familiarize themselves with the turbulence reporting criteria found in the AIM, which also describes the procedure for volunteering PIREPs relating to turbulence.

Structural Icing

The very nature of flight in Instrument Meteorological Conditions means operating in visible moisture such as clouds. At the right temperatures, this moisture can freeze on the aircraft, causing increased weight, degraded performance, and unpredictable aerodynamic characteristics. Understanding, avoidance, and early recognition followed by prompt action are the keys to avoiding this potentially hazardous situation.

Structural icing refers to the accumulation of ice on the exterior of the aircraft and is broken down into three classifications: rime ice, clear ice, and mixed ice. For ice to form, there must be moisture present in the air, and the air must be cooled to a temperature of 0° C (32° F) or less. Aerodynamic cooling can lower the surface temperature of an airfoil and cause ice to form on the airframe even though the ambient temperature is slightly above freezing.

Rime ice forms if the droplets are small and freeze immediately when contacting the aircraft surface. This type of ice usually forms on areas such as the leading edges of wings or struts. It has a somewhat rough-looking appearance and a milky-white color.

Clear ice is usually formed from larger water droplets or freezing rain that can spread over a surface. This is the most dangerous type of ice since it is clear, hard to see, and can change the shape of the airfoil.

Mixed ice is a mixture of clear ice and rime ice. It has the bad characteristics of both types and can form rapidly. Ice particles become embedded in clear ice, building a very rough accumulation. The table in *Figure 10-15* lists the temperatures at which the various types of ice will form.

Outside Air Temperature Range	Icing Type
0 °C to -10 °C	Clear
-10 °C to -15 °C	Mixed clear and rime
-15 °C to -20°C	Rime

Figure 10-15. *Temperature Ranges for Ice Formation.*

Structural icing is a condition that can only get worse. Therefore, during an inadvertent icing encounter, it is important the pilot act to prevent additional ice accumulation. Regardless of the level of anti-ice or deice protection offered by the aircraft, the first course of action should be to leave the area of visible moisture. This might mean descending to an altitude below the cloud bases, climbing to an altitude that is above the cloud tops, or turning to a different course. If this is not possible, then the pilot must move to an altitude where the temperature is above freezing. Pilots should report icing conditions to ATC and request new routing or altitude if icing will be a hazard. Refer to the AIM for information on reporting icing intensities.

Fog

Instrument pilots must learn to anticipate conditions leading to the formation of fog and take appropriate action early in the progress of the flight. Before a flight, close examination of current and forecast weather should alert the pilot to the possibility of fog formation. When fog is a consideration, pilots should plan adequate fuel reserves and alternate landing sites. En route, the pilot must stay alert for fog formation through weather updates from EFAS, ATIS, and ASOS/AWOS sites.

Two conditions will lead to the formation of fog. Either the air is cooled to saturation, or sufficient moisture is added to the air until saturation occurs. In either case, fog can form when the temperature/dewpoint spread is 5° or less. Pilots planning to arrive at their destination near dusk with decreasing temperatures should be particularly concerned about the possibility of fog formation.

Volcanic Ash

Volcanic eruptions create volcanic ash clouds containing an abrasive dust that poses a serious safety threat to flight operations. Adding to the danger is the fact that these ash clouds are not easily discernible from ordinary clouds when encountered at some distance from the volcanic eruption.
When an aircraft enters a volcanic ash cloud, dust particles and smoke may become evident in the cabin, often along with the odor of an electrical fire. Inside the volcanic ash cloud, the aircraft may also experience lightning and St. Elmo's fire on the windscreen. The abrasive nature of the volcanic ash can pit the windscreens, thus reducing or eliminating forward visibility. The pitot-static system may become clogged, causing instrument failure. Severe engine damage is probable in both piston and jet-powered aircraft.

Every effort must be made to avoid volcanic ash. Since volcanic ash clouds are carried by the wind, pilots should plan their flights to remain upwind of the ash-producing volcano. Visual detection and airborne radar are not considered a reliable means of avoiding volcanic ash clouds. Pilots witnessing volcanic eruptions or encountering volcanic ash should immediately pass this information along in the form of a pilot report. The National Weather Service monitors volcanic eruptions and estimates ash trajectories. This information is passed along to pilots in the form of SIGMETs.

As for many other hazards to flight, the best source of volcanic information comes from PIREPs. Pilots who witness a volcanic eruption or encounter volcanic ash in flight should immediately inform the nearest agency. Volcanic Ash Forecast Transport and Dispersion (VAFTAD) charts are also available; these depict volcanic ash cloud locations in the atmosphere following an eruption, and also forecast dispersion of the ash concentrations over 6- and 12-hour time intervals. See AC 00-45, Aviation Weather Services.

Thunderstorms

A thunderstorm packs just about every weather hazard known to aviation into one vicious bundle. Turbulence, hail, rain, snow, lightning, sustained updrafts and downdrafts, and icing conditions are all present in thunderstorms. Do not take off in the face of an approaching thunderstorm or fly an aircraft that is not equipped with thunderstorm detection in clouds or at night in areas of suspected thunderstorm activity. *[Figure 10-16]*

There is no useful correlation between the external visual appearance of thunderstorms and the severity or amount of turbulence or hail within them. All thunderstorms should be considered hazardous, and thunderstorms with tops above 35,000 feet should be considered extremely hazardous.

Figure 10-16. *A thunderstorm packs just about every weather hazard known to aviation into one vicious bundle.*

Weather radar, airborne or ground based, will normally reflect the areas of moderate to heavy precipitation (radar does not detect turbulence). The frequency and severity of turbulence generally increases with the radar reflectivity closely associated with the areas of highest liquid water content of the storm. A flight path through an area of strong or very strong radar echoes separated by 20 to 30 miles or less may not be considered free of severe turbulence.

The probability of lightning strikes occurring to aircraft is greatest when operating at altitudes where temperatures are between -5 ° C and +5 ° C. In addition, an aircraft flying in the clear air near a thunderstorm is also susceptible to lightning strikes. Thunderstorm avoidance is always the best policy.

Wind Shear

Wind shear can be defined as a change in wind speed and/or wind direction in a short distance. It can exist in a horizontal or vertical direction and occasionally in both. Wind shear can occur at all levels of the atmosphere but is of greatest concern during takeoffs and landings. It is typically associated with thunderstorms and low-level temperature inversions; however, the jet stream and weather fronts are also sources of wind shear.

As *Figure 10-17* illustrates, while an aircraft is on an instrument approach, a shear from a tailwind to a headwind causes the airspeed to increase and the nose to pitch up with a corresponding balloon above the glide path. A shear from a headwind to a tailwind has the opposite effect, and the aircraft will sink below the glide path.

A headwind shear followed by a tailwind/downdraft shear is particularly dangerous because the pilot has reduced power and lowered the nose in response to the headwind shear. This leaves the aircraft in a nose-low, power-low configuration when the tailwind shear occurs, which makes recovery more difficult, particularly near the ground. This type of wind shear scenario is likely while making an approach in the face of an oncoming thunderstorm. Pilots should be alert for indications of wind shear early in the approach phase and be ready to initiate a missed approach at the first indication. It may be impossible to recover from a wind shear encounter at low altitude.

To inform pilots of hazardous wind shear activity, some airports have installed a Low-Level Wind Shear Alert System (LLWAS) consisting of a centerfield wind indicator and several surrounding boundary-wind indicators. With

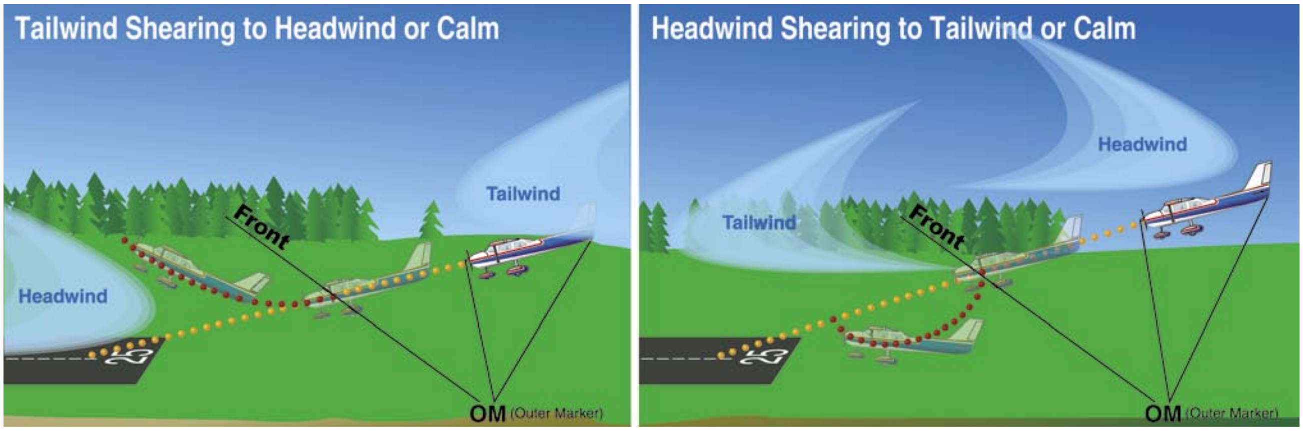

Figure 10-17. *Glide slope Deviations Due to Wind Shear Encounter.*

this system, controllers are alerted of wind discrepancies (an indicator of wind shear possibility) and provide this information to pilots. A typical wind shear alert issued to a pilot would be:

> "Runway 27 arrival, wind shear alert, 20 knot loss 3 mile final, threshold wind 200 at 15"

In plain language, the controller is advising aircraft arriving on runway 27 that at about 3 miles out they can expect a wind shear condition that will decrease their airspeed by 20 knots and possibly encounter turbulence. Additionally, the airport surface winds for landing runway 27 are reported as 200° at 15 knots.

Pilots encountering wind shear are encouraged to pass along pilot reports. Refer to AIM for additional information on wind shear PIREPs.

VFR-On-Top

Pilots on IFR flight plans operating in VFR weather conditions may request VFR-on-top in lieu of an assigned altitude. This permits them to select an altitude or flight level of their choice (subject to any ATC restrictions).

Pilots desiring to climb through a cloud, haze, smoke, or other meteorological formation and then either cancel their IFR flight plan or operate VFR-on-top may request a climb to VFR-on-top. The ATC authorization will contain a top report (or a statement that no top report is available) and a request to report upon reaching VFR-on-top. Additionally, the ATC authorization may contain a clearance limit, routing, and an alternative clearance if VFR-on-top is not reached by a specified altitude.

A pilot on an IFR flight plan, operating in VFR conditions, may request to climb/descend in VFR conditions. When operating in VFR conditions with an ATC authorization to "maintain VFR-on-top/maintain VFR conditions," pilots on IFR flight plans must:

1. Fly at the appropriate VFR altitude as prescribed in 14 CFR part 91.
2. Comply with the VFR visibility and distance-from-cloud criteria in 14 CFR part 91.
3. Comply with instrument flight rules applicable to this flight (minimum IFR altitudes, position reporting, radio communications, course to be flown, adherence to ATC clearance, etc.).

Pilots operating on a VFR-on-top clearance should advise ATC before any altitude change to ensure the exchange of accurate traffic information.

ATC authorization to "maintain VFR-on-top" is not intended to restrict pilots to operating only above an obscuring meteorological formation (layer). Rather, it permits operation above, below, between layers, or in areas where there is no meteorological obstruction. It is imperative pilots understand, however, that clearance to operate "VFR-on-top/VFR conditions" does not imply cancellation of the IFR flight plan.

Pilots operating VFR-on-top/VFR conditions may receive traffic information from ATC on other pertinent IFR or VFR aircraft. However, when operating in VFR weather conditions, it is the pilot's responsibility to be vigilant to see and avoid other aircraft.

This clearance must be requested by the pilot on an IFR flight plan. VFR-on-top is not permitted in certain areas, such as Class A airspace. Consequently, IFR flights operating VFR-on-top must avoid such airspace.

VFR Over-The-Top

VFR over-the-top must not be confused with VFR-on-top. VFR-on-top is an IFR clearance that allows the pilot to fly VFR altitudes. VFR over-the-top is strictly a VFR operation in which the pilot maintains VFR cloud clearance requirements while operating on top of an undercast layer. This situation might occur when the departure airport and the destination airport are reporting clear conditions, but a low overcast layer is present in between. The pilot could conduct a VFR departure, fly over the top of the undercast in VFR conditions, then complete a VFR descent and landing at the destination. VFR cloud clearance requirements would be maintained at all times, and an IFR clearance would not be required for any part of the flight.

Conducting an IFR Flight

To illustrate some of the concepts introduced in this chapter, follow along on a typical IFR flight from the Birmingham International Airport (BHM), Birmingham, Alabama to Gulfport-Biloxi International Airport (GPT), Gulfport, Mississippi. *[Figure 10-18]* For this trip, a Cessna 182 with a call sign of N1230A will be flown. The aircraft is equipped with dual navigation and communication radios, a transponder, and a GPS system approved for IFR en route, terminal, and approach operations.

Preflight

The success of the flight depends largely upon the thoroughness of the preflight planning. The evening before the flight, pay close attention to the weather forecast and begin planning the flight.

The Weather Channel indicates a large, low-pressure system has settled in over the Midwest, pulling moisture up from the Gulf of Mexico and causing low ceilings and visibility with little chance for improvement over the next couple of days. To begin planning, gather all the necessary charts and materials, and verify everything is current. This includes en route charts, approach charts, DPs, STAR charts, the GPS database, as well as an A/FD, some navigation logs, and the aircraft's POH/AFM. The charts cover both the departure and arrival airports and any contingency airports that will be needed if the flight cannot be completed as planned. This is also a good time for the pilot to consider recent flight experience, pilot proficiency, fitness, and personal weather minimums to fly this particular flight.

Check the A/FD to become familiar with the departure and arrival airport, and check for any preferred routing between BHM and GPT. Next, review the approach charts and any DP or STAR that pertains to the flight. Finally, review the en route charts for potential routing, paying close attention to the minimum en route and obstacle clearance altitudes.

After this review, select the best option. For this flight, the Birmingham Three Departure *[Figure 10-2]* to Brookwood VORTAC, V 209 to Kewanee VORTAC, direct to Gulfport using GPS would be a logical route. An altitude of 4,000 feet meets all the regulatory requirements and falls well within the performance capabilities of the aircraft.

Next, call 1-800-WX-BRIEF to obtain an outlook-type weather briefing for the proposed flight. This provides forecast conditions for departure and arrival airports, as well as the en route portion of the flight including forecast winds aloft. This also is a good opportunity to check the available NOTAMs.

The weather briefer confirms the predictions of the weather channel giving forecast conditions that are at or near minimum landing minimums at both BHM and GPT for the proposed departure time. The briefer provides NOTAM information for GPT indicating that the localizer to runway 32 is scheduled to be out of service and that runway 18/36 is closed until further notice. Also check for temporary flight restrictions (TFRs) along the proposed route.

After receiving a weather briefing, continue flight planning and begin to transfer some preliminary information onto the navigation log, listing each fix along the route and the distances, frequencies, and altitudes. Consolidating this information onto an organized navigation log will keep the workload to a minimum during the flight.

Next, obtain a standard weather briefing online for the proposed route. A check of current conditions indicates low IFR conditions at both the departure airport and the destination, with visibility of one-quarter mile:

SURFACE WEATHER OBSERVATIONS
METAR KBHM 111155Z VRB04KT ¼ SM FG –RA VV004 06/05 A2994 RMK A02 SLP140

METAR KGPT 111156Z 24003KT ¼ SM FG OVC001 08/07 A2962 RMK A02 SLP033

The small temperature/dewpoint spread is causing the low visibility and ceilings. Conditions should improve later in the day as temperatures increase. A check of the terminal forecast confirms this theory:

TERMINAL FORECASTS
TAF KBHM 111156Z 111212 VRB04KT ¼ SM FG VV004 TEMPO1316 ¾ SM OVC004

FM1600 VRB05KT 2SM BR OVC007 TEMPO 1720 3SM DZ BKN009

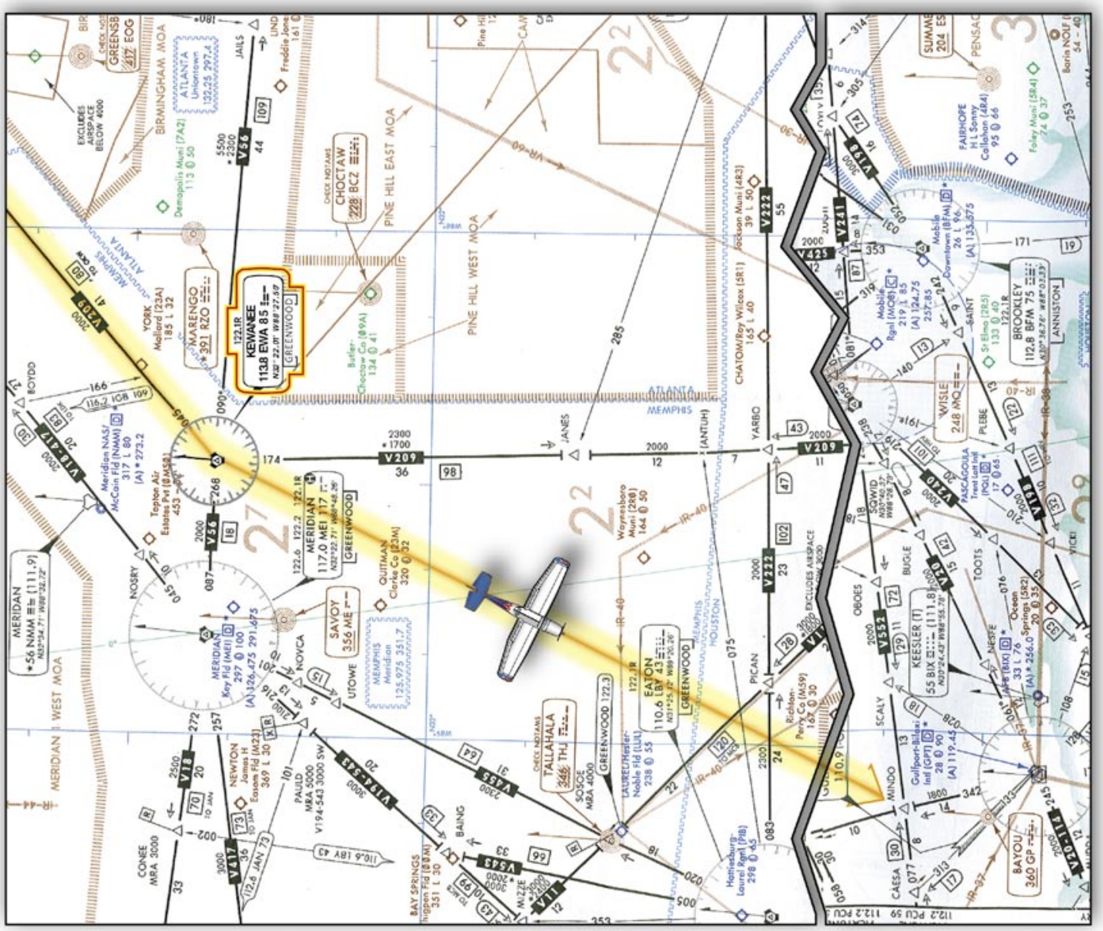
KEWANEE
113.8 EWA 85
MERIDIAN
117.0 MEI 117
EATON
110.6 LBY 43
KEESLER (T)
55 BX
BROOKLEY
112.8 BFM 75
V209
V222
PINE HILL EAST MOA
PINE HILL WEST MOA
MERIDIAN 1 WEST MOA
BIRMINGHAM MOA

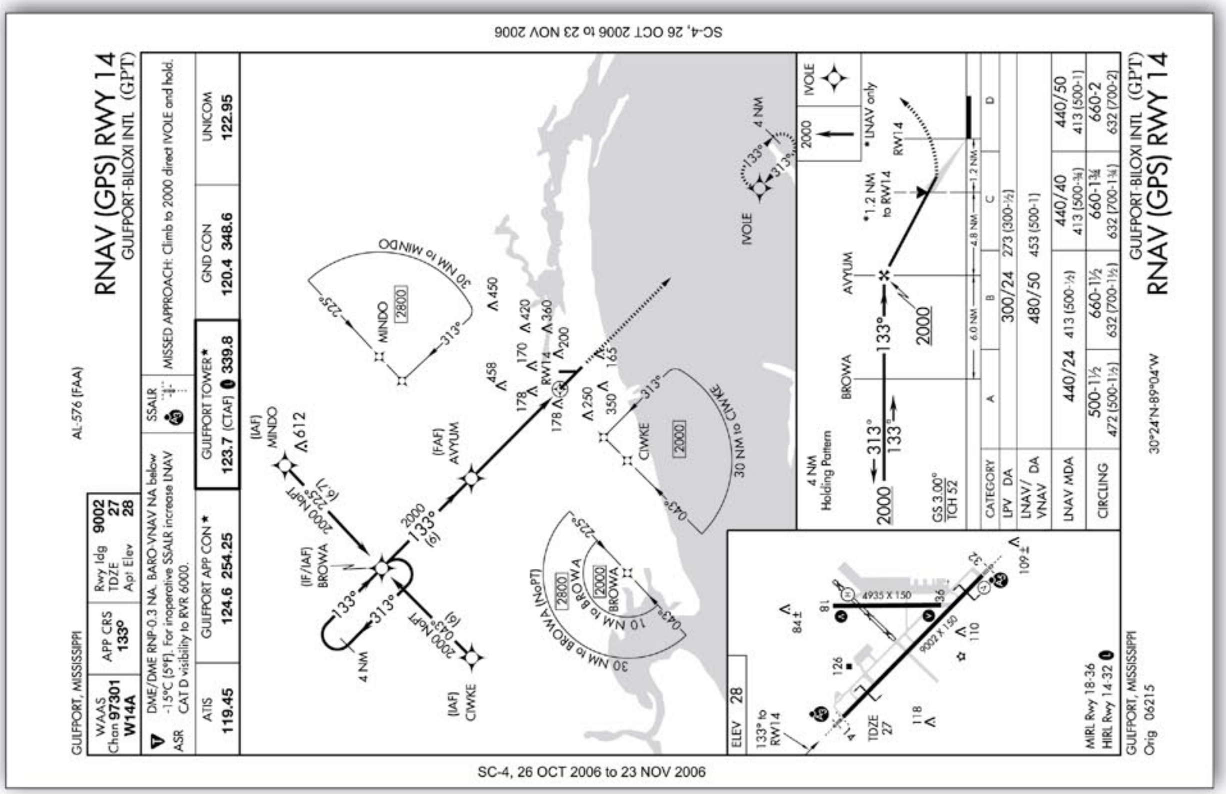
GULFPORT, MISSISSIPPI
AL-576 (FAA)
RNAV (GPS) RWY 14
GULFPORT-BILOXI INTL (GPT)
WAAS Chan 97301 W14A
APP CRS 133°
Rwy ldg 9002
TDZE 27
Apt Elev 28
MISSED APPROACH: Climb to 2000 direct IVOLE and hold.
ATIS 119.45
GULFPORT APP CON 124.6 254.25
GULFPORT TOWER 123.7 (CTAF) 339.8
GND CON 120.4 348.6
UNICOM 122.95
(IAF) CIWKE
(IF/IAF) BROWA
(IAF) MINDO
(FAF) AVYUM
IVOLE
RW14
ELEV 28
CATEGORY A B C D
LPV DA 300/24 273 (300-½)
LNAV/VNAV DA 480/50 453 (500-1)
LNAV MDA 440/24 413 (500-½) 440/40 413 (500-¾) 440/50 413 (500-1)
CIRCLING 500-1½ 472 (500-1½) 660-1½ 632 (700-1½) 660-1¾ 632 (700-1¾) 660-2 632 (700-2)
30°24'N-89°04'W
SC-4, 26 OCT 2006 to 23 NOV 2006

Figure 10-18. *Route Planning.*

FM2000 22008KT 3SM –RA OVC015 TEMP 2205 3SM –RA OVC025 FM0500 23013KT P6SM OVC025

FM0800 23013KT P6SM BKN030 PROB40 1012 2SM BR OVC030

TAF KGPT 111153Z 111212 24004KT ¼ SM FG OVC001 BECMG 1317 3SM BR 0VC004

FM1700 24010KT 4SM –RA OVC006 FM0400 24010 5SM SCT080 TEMPO 0612 P6SM SKC

In addition to the terminal forecast, the area forecast also indicates gradual improvement along the route. Since the terminal forecast only provides information for a 5-mile radius around a terminal area, checking the area forecast provides a better understanding of the overall weather picture along the route, as well as potential hazards:

SYNOPSIS AND VFR CLOUDS/WEATHER FORECASTS SYNOPSIS… AREA OF LOW PRESSURE CNTD OV AL RMNG GENLY STNRY BRNGNG MSTR AND WD SPRD IFR TO E TN. ALF…LOW PRES TROF ACRS CNTR PTN OF THE DFW FA WILL GDLY MOV EWD DURG PD.

NRN LA, AR, NRN MS
SWLY WND THRUT THE PD. 16Z CIG OVC006. SCT –SHRA. OTLK… IFR SRN ½ … CIG SCT – BKN015 TOPS TO FL250 SWLY WND THRUT THE PD. 17Z AGL BKN040. OTLK…MVFR CIG VIS.

LA MS CSTL WTRS
CIG OVC001 – OVC006. TOPS TO FL240. VIS ¼ - ¾ SM FG. SWLY WND. 16Z CIG OVC010 VIS 2 SM BR. OCNL VIS 3-5SM –RN BR OVC009. OTLK…MVFR CIG VIS.

FL
CIG BKN020 TOPS TO FL180. VIS 1–3 SM BR. SWLY WND. 18Z BRK030. OTLK…MVFR CIG.

At this time, there are no SIGMETs or PIREPs reported. However, there are several AIRMETs, one for IFR conditions, one for turbulence that covers the entire route, and another for icing conditions which covers an area just north of the route:

WAUS44 KKCI 111150

DFWS WA 0111150

AIRMET SIERRA FOR IFR VALID UNTIL 111800

AIRMET IFR...OK TX LA AR MS AL FL

TS IMPLY SEV OR GTR TURB SEV ICE LLWS AND IFR CONDS.

NON MSL HGHTS DENOTED BY AGL OR CIG.

A recheck of NOTAMs for Gulfport confirms that the localizer to runway 32 is out of service until further notice and runway 18/36 is closed. If runway 6 is planned for the departure, confirm that the climb restriction for the departure can be met.

GPT 12/006 GPT LOC OS UFN

GPT 12/008 GPT MIRL RWY 18/36 OS UFN

Since the weather is substantially better to the east, Pensacola Regional Airport is a good alternate with current conditions and a forecast of marginal VFR.

METAR KPNS 111150Z 21010Z 3SM BKN014 OVC025 09/03 A2973

TAF KPNS 111152Z 111212 22010KT 3 SM BR OVC020 BECMG 1317 4 SM BR OVC025

FM1700 23010KT 4SM –RA OVC030

FM 0400 25014KT 5SM OVC050 TEMPO1612 P6SM OVC080

If weather minimums are below a pilot's personal minimums, a delay in departure to wait for improved conditions is a good decision. This time can be used to complete the navigation log which is the next step in planning an IFR flight. *[Figure 10-19]*

Use the POH/AFM to compute a true airspeed, cruise power setting, and fuel burn based on the forecast temperatures aloft and cruising pressure altitude. Also, compute weight-and-balance information and determine takeoff and landing distances. There will be a crosswind if weather conditions require a straight-in landing on runway 14 at GPT. Therefore, compute the landing distance assuming a 10-knot crosswind and determine if the runway length is adequate to allow landing. Determine the estimated flight time and fuel burn using the winds aloft forecast and considering Pensacola Regional Airport as an alternate airport. With full tanks, the flight can be made nonstop with adequate fuel for flight to the destination, alternate, and the reserve requirement.

Next, check the surface analysis chart which shows where the pressure systems will be found. The weather depiction chart shows areas of IFR conditions and can be used to find areas

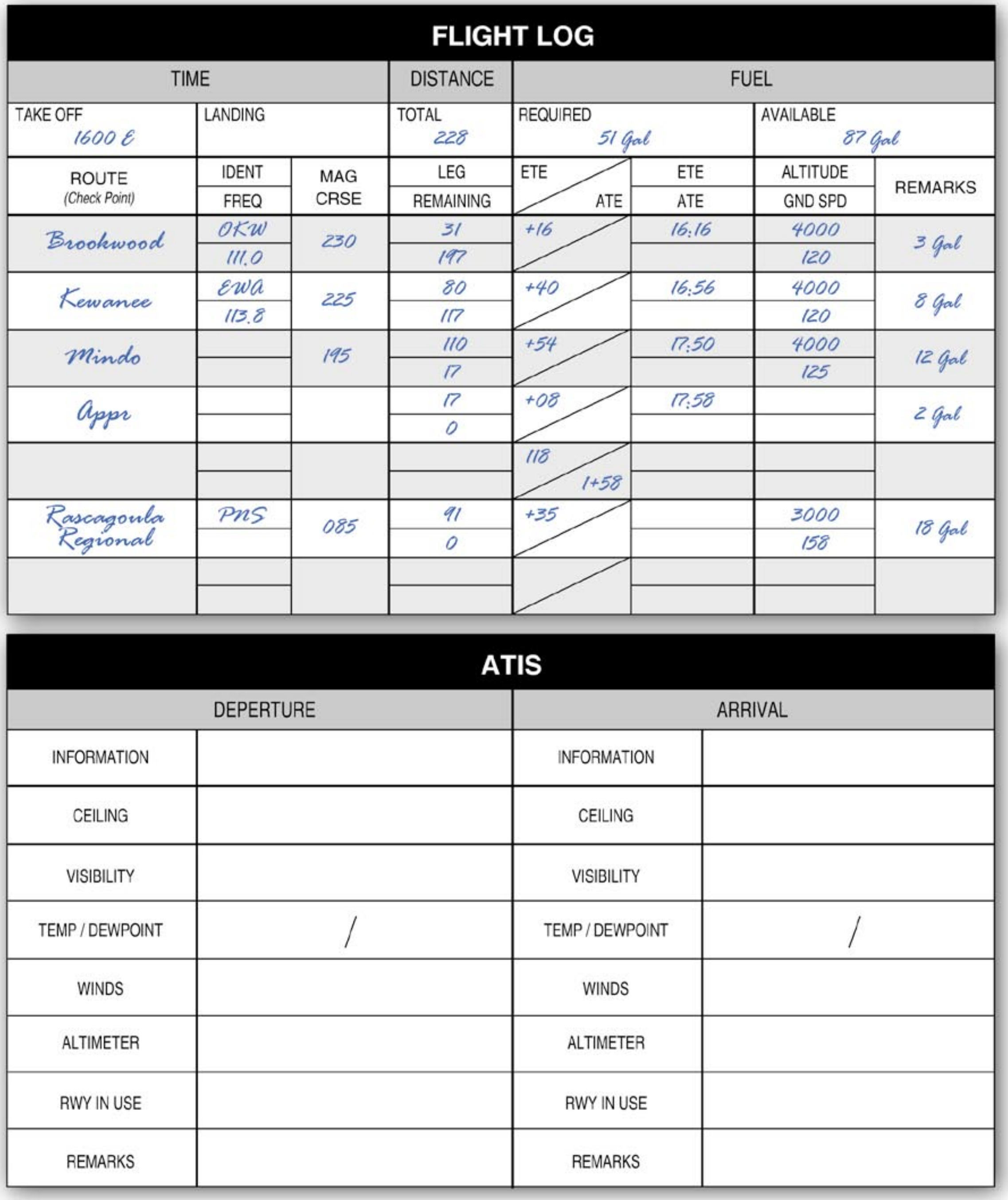

FLIGHT LOG

TIME		DISTANCE	FUEL	
TAKE OFF 1600 E	LANDING	TOTAL 228	REQUIRED 51 Gal	AVAILABLE 87 Gal

ROUTE (Check Point)	IDENT FREQ	MAG CRSE	LEG REMAINING	ETE ATE	ETE ATE	ALTITUDE GND SPD	REMARKS
Brookwood	OKW 111.0	230	31 197	+16	16:16	4000 120	3 Gal
Kewanee	EWA 113.8	225	80 117	+40	16:56	4000 120	8 Gal
Mindo		195	110 17	+54	17:50	4000 125	12 Gal
Appr			17 0	+08	17:58		2 Gal
				118 1+58			
Rascagoula Regional	PNS	085	91 0	+35		3000 158	18 Gal

ATIS

DEPERTURE		ARRIVAL	
INFORMATION		INFORMATION	
CEILING		CEILING	
VISIBILITY		VISIBILITY	
TEMP / DEWPOINT	/	TEMP / DEWPOINT	/
WINDS		WINDS	
ALTIMETER		ALTIMETER	
RWY IN USE		RWY IN USE	
REMARKS		REMARKS	

Figure 10-19. *Navigation Log.*

of improving conditions. These charts provide information a pilot will need should a diversion to VFR conditions be required. For this flight, the radar depicts precipitation along the route, and the latest satellite photo confirms what the weather depiction chart showed.

When the navigation log is finished, complete the flight plan in preparation for filing with flight service. *[Figure 10-20]*

Call an AFSS for an updated weather briefing, Birmingham INTL airport is now reporting 700 overcast with 3 miles visibility, and Gulfport-Biloxi is now 400 overcast with 2 miles visibility. The alternate, Pensacola Regional Airport, continues to report adequate weather conditions with 2,000 overcast and 3 miles visibility in light rain.

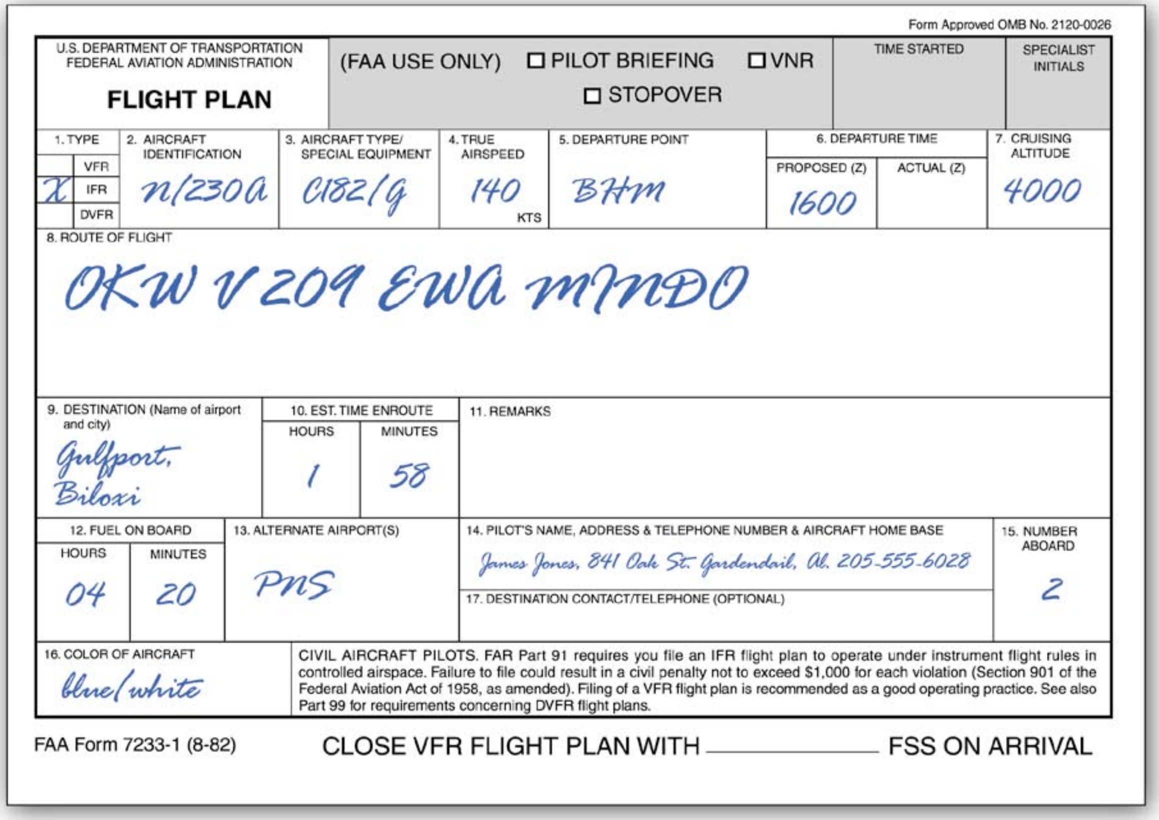

Form Approved OMB No. 2120-0026

U.S. DEPARTMENT OF TRANSPORTATION
FEDERAL AVIATION ADMINISTRATION

FLIGHT PLAN

(FAA USE ONLY) ☐ PILOT BRIEFING ☐ VNR ☐ STOPOVER

TIME STARTED

SPECIALIST INITIALS

1. TYPE: VFR / X IFR / DVFR
2. AIRCRAFT IDENTIFICATION: n1230a
3. AIRCRAFT TYPE/ SPECIAL EQUIPMENT: C182/G
4. TRUE AIRSPEED: 140 KTS
5. DEPARTURE POINT: BHM
6. DEPARTURE TIME: PROPOSED (Z) 1600 ACTUAL (Z)
7. CRUISING ALTITUDE: 4000
8. ROUTE OF FLIGHT: OKW V 209 EWA MINDO
9. DESTINATION (Name of airport and city): Gulfport, Biloxi
10. EST. TIME ENROUTE: HOURS 1 MINUTES 58
11. REMARKS
12. FUEL ON BOARD: HOURS 04 MINUTES 20
13. ALTERNATE AIRPORT(S): PNS
14. PILOT'S NAME, ADDRESS & TELEPHONE NUMBER & AIRCRAFT HOME BASE: James Jones, 841 Oak St. Gardendail, Al. 205-555-6028
15. NUMBER ABOARD: 2
16. COLOR OF AIRCRAFT: blue/white
17. DESTINATION CONTACT/TELEPHONE (OPTIONAL)

CIVIL AIRCRAFT PILOTS. FAR Part 91 requires you file an IFR flight plan to operate under instrument flight rules in controlled airspace. Failure to file could result in a civil penalty not to exceed $1,000 for each violation (Section 901 of the Federal Aviation Act of 1958, as amended). Filing of a VFR flight plan is recommended as a good operating practice. See also Part 99 for requirements concerning DVFR flight plans.

FAA Form 7233-1 (8-82) CLOSE VFR FLIGHT PLAN WITH ______ FSS ON ARRIVAL

Figure 10-20. *Flight Plan Form.*

Several pilot reports have been submitted for light icing conditions; however, all the reports are north of the route of flight and correspond to the AIRMET that was issued earlier. No pilot reports have included cloud tops, but the area forecast predicted cloud tops to flight level 240. Since the weather conditions appear to be improving, a flight plan can be filed using the completed form.

Analyze the latest weather minimums to determine if they exceed personal minimums. With the absence of icing reported along the route and steadily rising temperatures, structural icing should not be a problem. Make a note to do an operational check of the pitot heat during preflight and to take evasive action immediately should even light icing conditions be encountered in flight. This may require returning to BHM or landing at an intermediate spot before reaching GPT. The go/no-go decision will be constantly reevaluated during the flight.

Once at the airport, conduct a thorough preflight inspection. A quick check of the logbooks indicates all airworthiness requirements have been met to conduct this IFR flight including an altimeter, static, and transponder test within the preceding 24 calendar months. In addition, a log on the clipboard indicates the VOR system has been checked within the preceding 30 days. Turn on the master switch and pitot heat, and quickly check the heating element before it becomes too hot. Then, complete the rest of the walk-around procedure. Since this will be a flight in actual IFR conditions, place special emphasis on IFR equipment during the walk-around, including the alternator belt and antennas. After completing the preflight, organize charts, pencils, paper, and navigation log in the flight deck for quick, easy access. This is also the time to enter the planned flight into the GPS.

Departure

After starting the engine, tune in ATIS and copy the information to the navigation log. The conditions remain the same as the updated weather briefing with the ceiling at 700 overcast, and visibility at 3 miles. Call clearance delivery to receive a clearance:

> "Clearance Delivery, Cessna 1230A IFR to Gulfport Biloxi with information Kilo, ready to copy."

> "Cessna 1230A is cleared to Gulfport-Biloxi via the Birmingham Three Departure, Brookwood, Victor 209 Kewanee then direct Mindo, Gulfport. Climb and maintain 4,000. Squawk 0321."

Read back the clearance and review the DP. Although a departure frequency was not given in the clearance, note that on the DP, the departure control frequency is listed as 123.8 for the southern sector. Since a departure from runway 24 is anticipated, note the instruction to climb to 2,100 prior to turning. After tuning in the appropriate frequencies and setting up navigation equipment for the departure routing, contact ground control (noting that this is IFR) and receive the following clearance:

> "Cessna 1230A taxi to runway 24 via taxiway Mike."

Read back the clearance and aircraft call sign. After a review of the taxi instructions on the airport diagram, begin to taxi and check the flight instruments for proper indications.

Hold short of runway 24 and complete the before takeoff checklist and engine run-up. Advise the tower when ready for takeoff. The tower gives the following clearance:

> "Cessna 30A cleared for takeoff runway 24. Caution wake turbulence from 737 departing to the northwest."

Taxi into position. Note the time off on the navigation log, verify that the heading indicator and magnetic compass are in agreement, the transponder is in the ALT position, all the necessary lights, equipment, and pitot heat are on. Start the takeoff roll. To avoid the 737's wake turbulence, make note of its lift off point and take off prior to that point.

En Route

After departure, climb straight ahead to 2,100 feet as directed by the Birmingham Three Departure. While continuing a climb to the assigned altitude of 4,000 feet, the following instructions are received from the tower:

> "Cessna 30A contact Departure."

Acknowledge the clearance and contact departure on the frequency designated by the DP. State the present altitude so the departure controller can check the encoded altitude against indicated altitude:

> "Birmingham Departure Cessna 1230A climbing through 2,700 heading 240."

Departure replies:

> "Cessna 30A proceed direct to Brookwood and resume own navigation. Contact Atlanta Center on 134.05."

Acknowledge the clearance, contact Atlanta Center and proceed direct to Brookwood VORTAC, using the IFR-approved GPS equipment. En route to Kewanee, VORTAC Atlanta Center issues the following instructions:

> "Cessna 1230A contact Memphis Center on 125.975."

Acknowledge the instructions and contact Memphis Center with aircraft ID and present altitude. Memphis Center acknowledges contact:

> "Cessna 1230A, Meridian altimeter is 29.87. Traffic at your 2 o'clock and 6 miles is a King Air at 5,000 climbing to 12,000."

Even when on an IFR flight plan, pilots are still responsible for seeing and avoiding other aircraft. Acknowledge the call from Memphis Center and inform them of negative contact with traffic due to IMC.

> "Roger, altimeter setting 29.87. Cessna 1230A is in IMC negative contact with traffic."

Continue the flight, and at each fix note the arrival time on the navigation log to monitor progress.

To get an update of the weather at the destination and issue a pilot report, contact the FSS servicing the area. To find the nearest AFSS, locate a nearby VOR and check above the VOR information box for a frequency. In this case, the nearest VOR is Kewanee VORTAC which lists a receive-only frequency of 122.1 for Greenwood FSS. Request a frequency change from Memphis and then attempt to contact Greenwood on 122.1 while listening over the Kewanee VORTAC frequency of 113.8:

> "Greenwood Radio Cessna 1230A receiving on frequency 113.8, over."

> "Cessna 30A, this is Greenwood, go ahead."

> "Greenwood Radio, Cessna 30A is currently 30 miles south of the Kewanee VORTAC at 4,000 feet en route to Gulfport. Requesting an update of en route conditions and current weather at GPT, as well as PNS."

> "Cessna 30A, Greenwood Radio, current weather at Gulfport is 400 overcast with 3 miles visibility in light rain. The winds are from 140 at 7 and the altimeter is 29.86. Weather across your route is generally IFR in light rain with ceilings ranging from 300 to 1,000 overcast with visibilities between 1 and 3 miles. Pensacola weather is much better with ceilings now at 2,500 and visibility 6 miles. Checking current NOTAMs at GPT shows the localizer out of service and runway 18/36 closed."

> "Roger, Cessna 30A copies the weather. I have a PIREP when you are ready to copy."

> "Cessna 30A go ahead with your PIREP."

> "Cessna 30A is a Cessna 182 located on the Kewanee 195° radial at 30 miles level at 4,000 feet. I am currently in IMC conditions with a smooth ride. Outside air temperature is plus 1° Celsius. Negative icing."

> "Cessna 30A thank you for the PIREP."

With the weather check and PIREP complete, return to Memphis Center:

> "Memphis Center, Cessna 1230A is back on your frequency."

> "Cessna 1230A, Memphis Center, roger, contact Houston Center now on frequency 126.8."

> "Roger, contact Houston Center frequency 126.8, Cessna 1230A."

> "Houston Center, Cessna 1230A level at 4,000 feet."

> "Cessna 30A, Houston Center area altimeter 29.88."

Arrival

40 miles north of Gulfport, tune in ATIS on number two communication radio. The report reveals there has been no change in the weather and ATIS is advertising ILS runway 14 as the active approach.

Houston Center completes a hand off to Gulfport approach control with instructions to contact approach:

> "Gulfport Approach, Cessna 1230A level 4,000 feet with information TANGO. Request GPS Runway 14 approach."

> "Cessna 30A, Gulfport Approach, descend and maintain 3,000 feet."

> "Descend to 3,000, Cessna 30A."

Begin a descent to 3,000 and configure your navigation radios for the approach. The GPS will automatically change from the en route mode to the terminal mode. This change will affect the sensitivity of the CDI. Tune in the VORTAC frequency of 109.0 on the number one navigation radio, and set in the final approach course of 133° on the OBS. This setup will help with situational awareness should the GPS lose signal.

> "Cessna 30A your position is 7 miles from MINDO, maintain 3,000 feet until MINDO, cleared for the GPS runway 14 approach."

Read back the clearance and concentrate on flying the aircraft. At MINDO descend to 2,000 as depicted on the approach chart. At BROWA turn to the final approach course of 133°. Just outside the Final Approach Way Point (FAWP) AVYUM, the GPS will change to the approach mode and the CDI will become even more sensitive. Gulfport approach control issues instructions to contact Gulfport tower:

> "Cessna 30A contact Tower on 123.7."

> "123.7, Cessna 30A."

> "Tower, Cessna 1230A outside AVYUM on the GPS runway 14."

> "Cessna 30A Gulfport Tower, the ceiling is now 600 overcast and the visibility is 4 miles."

> "Cleared to land runway 14, Cessna 30A."

Continue the approach, complete the appropriate checklists, cross AVYUM, and begin the final descent. At 700 feet MSL visual contact with the airport is possible. Slow the aircraft and configure it to allow a normal descent to landing. As touch down is completed, Gulfport Tower gives further instructions:

> "Cessna 30A turn left at taxiway Bravo and contact ground on 120.4."

> "Roger, Cessna 30A."

Taxi clear of the runway and complete the appropriate checklists. The Tower will automatically cancel the IFR flight plan.

ATTITUDE FAIL
AIRSPEED FAIL
TRAFFIC
HDG
NORTH UP
INSET
PFD
OBS
CDI
DME
XPDR
IDENT
COLORADO
NEW MEXICO
ALBUQUERQUE
ENGINE

Chapter 11

Emergency Operations

Introduction

Changing weather conditions, air traffic control (ATC), the aircraft, and the pilot are all variables that make instrument flying an unpredictable and challenging operation. The safety of the flight depends upon the pilot's ability to manage these variables while maintaining positive aircraft control and adequate situational awareness. This chapter discusses the recognition and suggested remedies for such abnormal and emergency events related to unforecasted, adverse weather; aircraft system malfunctions; communication/navigation system malfunctions; and loss of situational awareness.

Unforecast Adverse Weather

Inadvertent Thunderstorm Encounter

A pilot should avoid flying through a thunderstorm of any intensity. However, certain conditions may be present that could lead to an inadvertent thunderstorm encounter. For example, flying in areas where thunderstorms are embedded in large cloud masses may make thunderstorm avoidance difficult, even when the aircraft is equipped with thunderstorm detection equipment. Therefore, pilots must be prepared to deal with an inadvertent thunderstorm penetration. At the very least, a thunderstorm encounter subjects the aircraft to turbulence that could be severe. The pilot and passengers should tighten seat belts and shoulder harnesses and secure any loose items in the cabin.

As with any emergency, the first order of business during an inadvertent thunderstorm encounter must be to fly the aircraft. The pilot workload is heavy; therefore, increased concentration is necessary to maintain an instrument scan. If a pilot inadvertently enters a thunderstorm, it is better to maintain a course straight through the thunderstorm rather than turning around. A straight course minimizes the amount of time in the thunderstorm and turning maneuvers only increase structural stress on the aircraft.

Reduce power to a setting that maintains a speed at the recommended turbulence penetration speed as described in the Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM), and try to minimize additional power adjustments. Concentrate on maintaining a level attitude while allowing airspeed and altitude to fluctuate. Similarly, if using the autopilot, disengage the altitude hold and speed hold modes, as they only increase the aircraft's maneuvering—thereby increasing structural stress.

During a thunderstorm encounter, the potential for icing also exists. As soon as possible, turn on anti-icing/deicing equipment and carburetor heat, if equipped. Icing can be rapid at any altitude and may lead to power failure and/or loss of airspeed indication.

Lightning is also present in a thunderstorm and can temporarily blind a pilot. To reduce this risk, turn up flight deck lights to the highest intensity, concentrate on the flight instruments, and resist the urge to look outside.

Inadvertent Icing Encounter

Because icing is unpredictable in nature, pilots may find themselves in icing conditions even though they have done everything practicable to avoid it. In order to stay alert to this possibility while operating in visible moisture, pilots should monitor the outside air temperature (OAT).

The effects of ice on aircraft are cumulative—thrust is reduced, drag increases, lift lessens, and weight increases. The results are an increase in stall speed and a deterioration of aircraft performance. In extreme cases, two to three inches of ice can form on the leading edge of the airfoil in less than 5 minutes. It takes only 1/2 inch of ice to reduce the lifting power of some aircraft by 50 percent and increases the frictional drag by an equal percentage.

A pilot can expect icing when flying in visible precipitation, such as rain or cloud droplets, and the temperature is between +02 and -10° Celsius. When icing is detected, a pilot should do one of two things, particularly if the aircraft is not equipped with deicing equipment: leave the area of precipitation or go to an altitude where the temperature is above freezing. This "warmer" altitude may not always be a lower altitude. Proper preflight action includes obtaining information on the freezing level and the above-freezing levels in precipitation areas.

If neither option is available, consider an immediate landing at the nearest suitable airport. Even if the aircraft is equipped with anti-icing/deicing equipment, it is not designed to allow aircraft to operate indefinitely in icing conditions. Anti-icing/deicing equipment gives a pilot more time to get out of the icing conditions. Report icing to ATC and request new routing or altitude. Be sure to report the type of aircraft, and use the following terms when reporting icing to ATC:

1. Trace. Ice becomes perceptible. Rate of accumulation is slightly greater than sublimation. Deicing/anti-icing equipment is not utilized unless encountered for an extended period of time (over 1 hour).
2. Light. The rate of accumulation may create a problem if flight is prolonged in this environment (over 1 hour). Occasional use of deicing/anti-icing equipment removes/prevents accumulation. It does not present a problem if deicing/anti-icing equipment is used.
3. Moderate. The rate of accumulation is such that even short encounters become potentially hazardous and use of deicing/anti-icing equipment or flight diversion is necessary.
4. Severe. The rate of accumulation is such that deicing/anti-icing equipment fails to reduce or control the hazard. Immediate flight diversion is necessary.

Early ice detection is critical and is particularly difficult during night flight. Use a flashlight to check for ice accumulation on the wings. At the first indication of ice accumulation, take action to get out of the icing conditions. Refer to the POH/AFM for the proper use of anti-icing/deicing equipment.

Figure 11-1. *St. Elmo's Fire is harmless but may affect both communication and navigation radios, especially the lower frequencies such as those used on the ADF.*

Precipitation Static

Precipitation static, often referred to as P-static, occurs when accumulated static electricity is discharged from the extremities of the aircraft. This discharge has the potential to create problems for the instrument pilot. These problems range from the serious, such as erroneous magnetic compass readings and the complete loss of very high frequency (VHF) communications to the annoyance of high-pitched audio squealing and St. Elmo's fire. *[Figure 11-1]*

Precipitation static is caused when an aircraft encounters airborne particles during flight (e.g., rain or snow), and develops a negative charge. It can also result from atmospheric electric fields in thunderstorm clouds. When a significant negative voltage level is reached, the aircraft discharges it, which can create electrical disturbances. This electrical discharge builds with time as the aircraft flies in precipitation. It is usually encountered in rain, but snow can cause the same effect. As the static buildup increases, the effectiveness of both communication and navigation systems decreases to the point of potential unusability.

To reduce the problems associated with P-static, the pilot should ensure the aircraft's static wicks are properly maintained and accounted for. Broken or missing static wicks should be replaced before an instrument flight. *[Figure 11-2]*

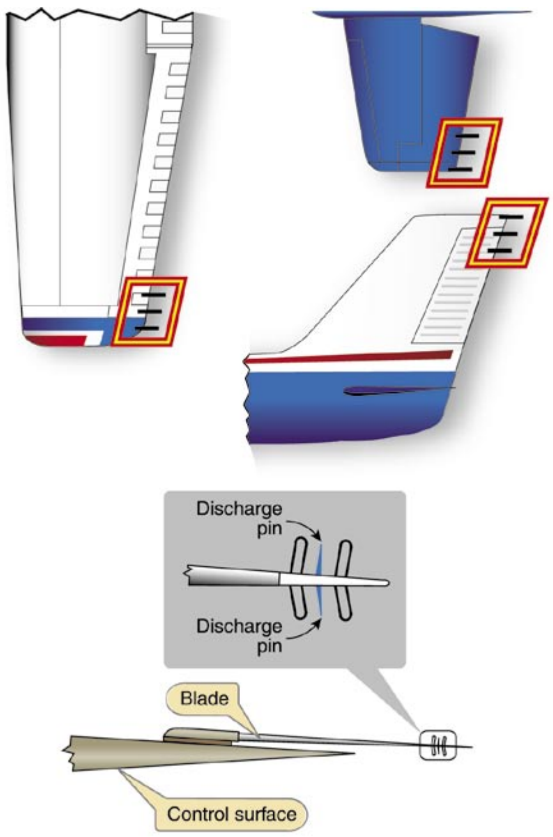

Figure 11-2. *One example of a static wick installed on aircraft control surface to bleed off static charges built up during flight. This will prevent static buildup and St. Elmo's fire by allowing the static electricity to dissipate harmlessly.*

Aircraft System Malfunctions

Preventing aircraft system malfunctions that might lead to an inflight emergency begins with a thorough preflight inspection. In addition to those items normally checked prior to a visual flight rules (VFR) flight, pilots intending to fly under instrument flight rules (IFR) should pay particular attention to the alternator belt, antennas, static wicks, anti-icing/deicing equipment, pitot tube, and static ports.

During taxi, verify the operation and accuracy of all flight instruments. In addition, during the run-up, verify that the operation of the pneumatic system(s) is within acceptable parameters. It is critical that all systems are determined to be operational before departing into IFR conditions.

Electronic Flight Display Malfunction

When a pilot becomes familiar and comfortable with the new electronic displays, he or she also tends to become more reliant on the system. The system then becomes a primary source of navigation and data acquisition instead of the supplementary source of data as initially intended.

Complete reliance on the moving map for navigation becomes a problem during a failure of one, more, or all of the flight display screens. Under these conditions, the systems revert to a composite mode (called reversionary), which eliminates the moving map display and combines the PFD with the engine indicating system. *[Figure 11-3]* If a pilot has relied on the display for navigation information and situational awareness, he or she lacks any concept of critical data such as the aircraft's position, the nearest airport, or proximity to other aircraft.

The electronic flight display is a supplementary source of navigation data and does not replace en route charts. To maintain situational awareness, a pilot must follow the flight on the en route chart while monitoring the PFD. It is important for the pilot to know the location of the closest airport as well as surrounding traffic relative to the location of his or her aircraft. This information becomes critical should the electronic flight display fail.

For the pilot who utilizes the electronic database as a substitute for the Airport Facilities Directory, screen failure or loss of electrical power can mean the pilot is no longer able to access airport information. Once the pilot loses the ability to call up airport information, aeronautical decision-making is compromised.

Alternator/Generator Failure

Depending upon the aircraft being flown, an alternator failure is indicated in different ways. Some aircraft use an ammeter

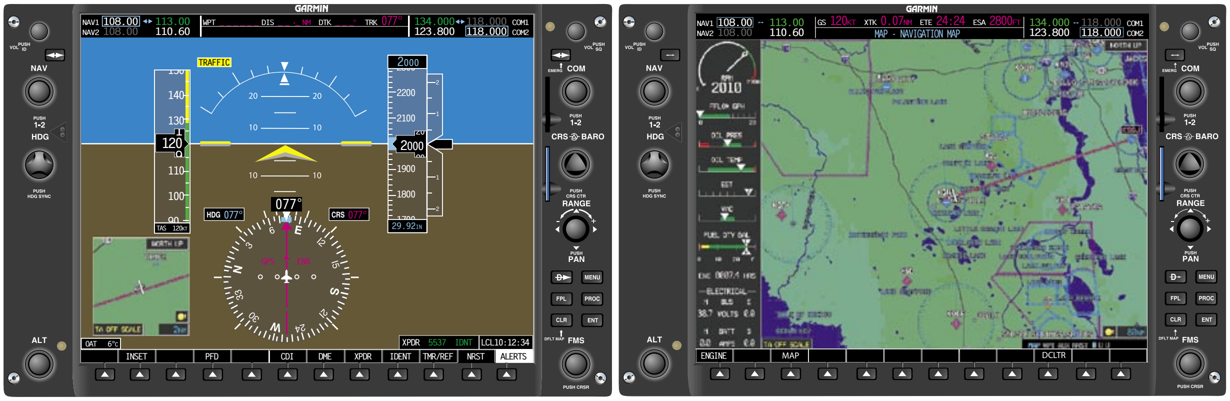

Normal Mode

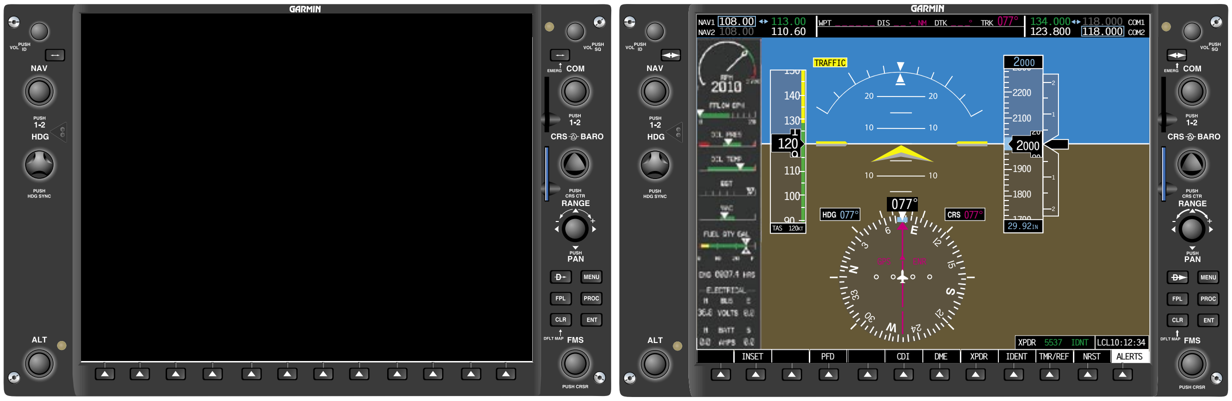

Display Failure / Reversionary Mode

Figure 11-3. *G1000 PFD display in normal mode and in the reversionary mode activated upon system failure.*

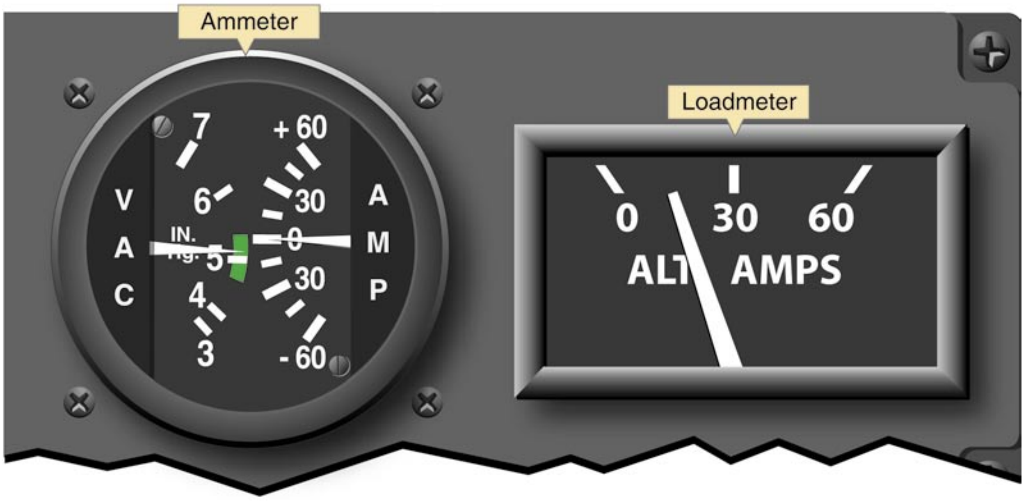

Figure 11-4. *Ammeter (left) and Loadmeter (right).*

that indicates the state of charge or discharge of the battery. *[Figure 11-4]* A positive indication on the ammeter indicates a charge condition; a negative indication reveals a discharge condition. Other aircraft use a load meter to indicate the load being carried by the alternator. *[Figure 11-4]*

Sometimes an indicator light is also installed in the aircraft to alert the pilot to an alternator failure. On some aircraft such as the Cessna 172, the light is located on the lower left side making it difficult to see its illumination if charts are open Ensure that these safety indicators are visible during flight.

When a loss of the electrical charging system is experienced, the pilot has approximately 40 minutes of battery life remaining before the system fails entirely. The time mentioned is an approximation and should not be relied upon as specific to all aircraft. In addition, the battery charge that exists in a battery may not be full, altering the time available before electrical exhaustion occurs. At no time should a pilot consider continuing a flight once the electrical charging system has failed. Land at the nearest suitable airport.

Techniques for Electrical Usage

Master Battery Switch

One technique for conserving the main battery charge is to fly the aircraft to the airport of intended landing while operating with minimal power. If a two-position battery master/alternator rocker switch *[Figure 11-5]* is installed, it can be utilized to isolate the main battery from the electrical system and conserve power.

Operating on the Main Battery

While en route to the airport of intended landing, reduce the electrical load as much as practical. Turn off all unnecessary electrical items such as duplicate radios, non-essential lighting, etc. If unable to turn off radios, lights, etc. manually, consider pulling circuit breakers to isolate those pieces of equipment from the electrical system. Maximum time of useful voltage may be between 30 and 40 minutes and is influenced by many factors, which degrade the useful time.

Loss of Alternator/Generator for Electronic Flight Instrumentation

With the increase in electrical components being installed in modern technically advanced aircraft, the power supply and the charging system need increased attention and understanding. Traditional round dial aircraft do not rely as heavily on electrical power for the primary six-pack instrumentation. Modern electronic flight displays utilize the electrical system to power the AHRS, ADC, engine indicating system (EIS), etc. A loss of an alternator or generator was considered an abnormality in traditionally equipped aircraft;

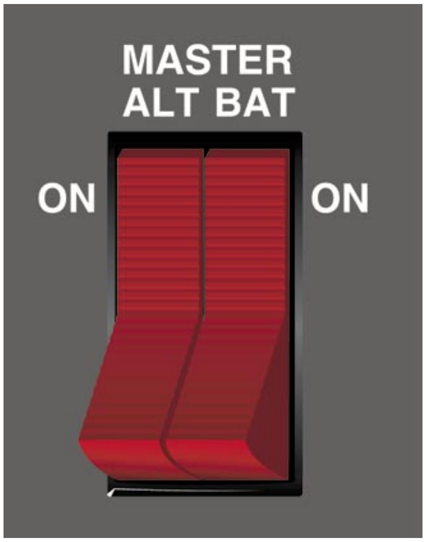

Figure 11-5. *Double Rocker Switch Seen on Many Aircraft.*

however, a failure of this magnitude is considered an emergency in technically advanced aircraft.

Due to the increased demand for electrical power, it is necessary for manufacturers to install a standby battery in conjunction with the primary battery. The standby battery is held in reserve and kept charged in case of a failure of the charging system and a subsequent exhaustion of the main battery. The standby battery is brought online when the main battery voltage is depleted to a specific value, approximately 19 volts. Generally, the standby battery switch must be in the ARM position for this to occur but pilots should refer to the aircraft flight manual for specifics on an aircraft's electrical system. The standby battery powers the essential bus and allows the primary flight display (PFD) to be utilized.

The essential bus usually powers the following components:

1. AHRS (Attitude and Heading Reference System)
2. ADC (Air Data Computer)
3. PFD (Primary Flight Display)
4. Navigation Radio #1
5. Communication Radio #1
6. Standby Indicator Light

Techniques for Electrical Usage

Standby Battery

One technique for conserving the main battery charge is to fly the aircraft to the airport of intended landing while using the standby battery. A two-position battery master/alternator rocker switch is installed on most aircraft with electronic flight displays, which can be utilized to isolate the main battery from the electrical system. By switching the MASTER side off, the battery is taken offline and the standby battery comes online to power the essential bus. However, the standby battery switch must be in the ARM position for this to occur. *[Figure 11-6]* Utilization of the standby battery first reserves the main battery for use when approaching to land. With this technique, electrical power may be available for the use of flaps, gear, lights, etc. Do not rely on any power to be available after the standby battery has exhausted itself. Once the charging system has failed, flight with a powered electrical system is not guaranteed.

Operating on the Main Battery

While en route to the airport of intended landing, reduce the electrical load as much as practical. Turn off all unnecessary electrical items such as duplicate radios, non-essential lighting, etc. If unable to turn off radios, lights, etc., manually, consider pulling circuit breakers to isolate those pieces of equipment from the electrical system. Keep in mind that once the standby battery has exhausted its charge, the flight deck may become very dark depending on what time of day the failure occurs. The priority during this emergency situation is landing the aircraft as soon as possible without jeopardizing safety.

A standby attitude indicator, altimeter, airspeed indicator (ASI) and magnetic compass are installed in each aircraft for use when the PFD instrumentation is unavailable. *[Figure 11-7]* These would be the only instruments left available to the pilot. Navigation would be limited to pilotage and dead reckoning unless a hand-held transceiver with a GPS/navigation function is onboard.

Once an alternator failure has been detected, the pilot must reduce the electrical load on the battery and land as soon as practical. Depending upon the electrical load and condition of the battery, there may be sufficient power available for 45 minutes of flight—or for only a matter of minutes. Pilots should also know which systems on the aircraft are electric and those that continue to operate without electrical power. Pilots can attempt to troubleshoot alternator failure by following the established alternator failure procedure published in the POH/AFM. If the alternator cannot be reset, advise ATC of the situation and inform them of the impending electrical failure.

Analog Instrument Failure

A warning indicator or an inconsistency between indications on the attitude indicator and the supporting performance

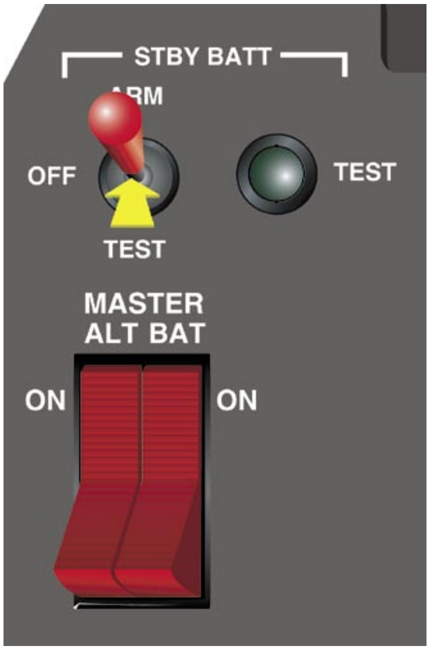

Figure 11-6. *Note the double rocker switch and the standby battery switch in this aircraft. The standby battery must be armed to work correctly; arming should be done prior to departure.*

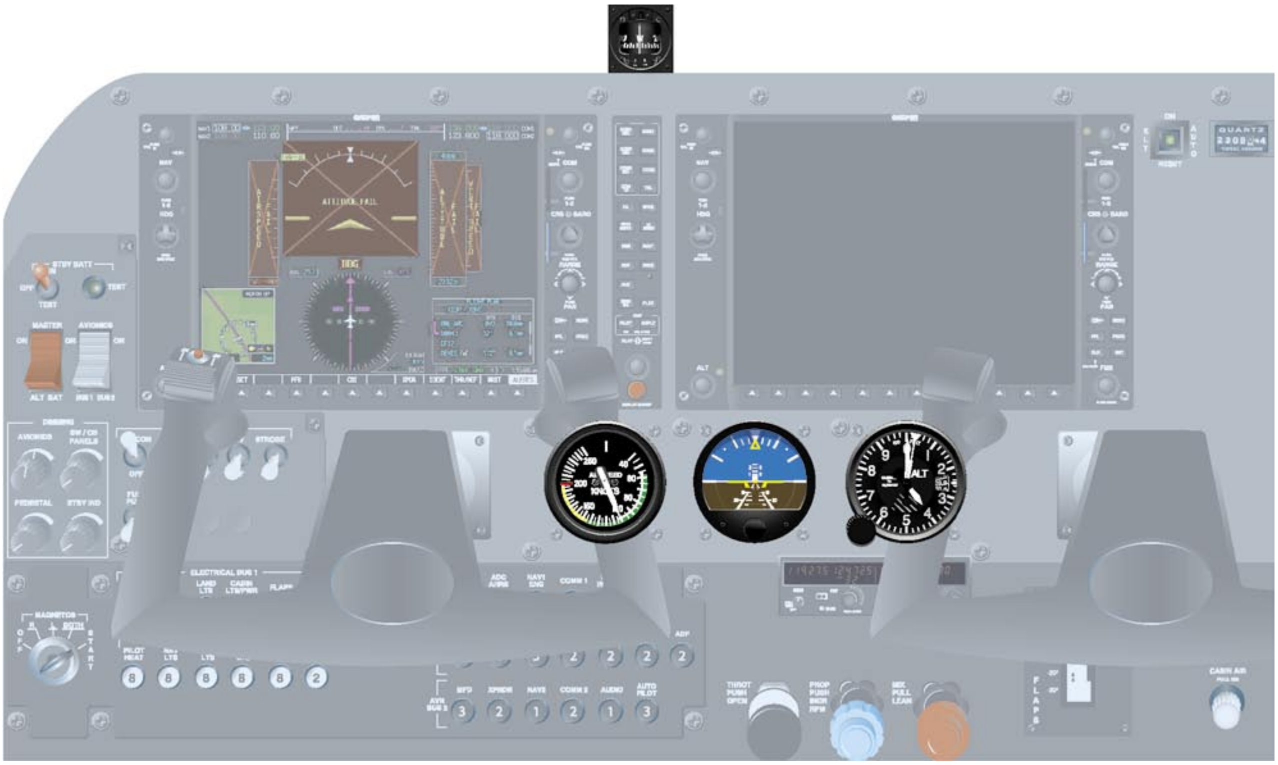

Figure 11-7. *Emergency Instrumentation Available to the Pilot on Electronic Flight Instrumented Aircraft.*

instruments usually identifies system or instrument failure. Aircraft control must be maintained while identifying the failed component(s). Expedite the cross-check and include all flight instruments. The problem may be individual instrument failure or a system failure affecting multiple instruments.

One method of identification involves an immediate comparison of the attitude indicator with the rate-of-turn indicator and vertical speed indicator (VSI). Along with providing pitch-and-bank information, this technique compares the static system with the suction or pressure system and the electrical system. Identify the failed component(s) and use the remaining functional instruments to maintain aircraft control.

Attempt to restore the inoperative component(s) by checking the appropriate power source, changing to a backup or alternate system, and resetting the instrument if possible. Covering the failed instrument(s) may enhance a pilot's ability to maintain aircraft control and navigate the aircraft. Usually, the next step is to advise ATC of the problem and, if necessary, declare an emergency before the situation deteriorates beyond the pilot's ability to recover.

Pneumatic System Failure

One possible cause of instrument failure is a loss of the suction or pressure source. This pressure or suction is supplied by a vacuum pump mechanically driven off the engine. Occasionally these pumps fail, leaving the pilot with inoperative attitude and heading indicators.

Figure 11-8 illustrates inoperative vacuum driven attitude and heading indicators which can fail progressively. As the gyroscopes slow down they may wander, which, if connected to the autopilot and/or flight director, can cause incorrect movement or erroneous indications. In *Figure 11-8,* the aircraft is actually level and at 2,000 feet MSL. It is not in a turn to the left which the pilot may misinterpret if he or she fails to see the off or failed flags. If that occurs, the pilot may transform a normally benign situation into a hazardous situation. Again, good decision-making by the pilot only occurs after a careful analysis of systems.

Many small aircraft are not equipped with a warning system for vacuum failure; therefore, the pilot should monitor the system's vacuum/pressure gauge. This can be a hazardous situation with the potential to lead the unsuspecting pilot into a dangerous unusual attitude which would require a partial panel recovery. It is important that pilots practice instrument flight without reference to the attitude and heading indicators in preparation for such a failure.

Pitot/Static System Failure

A pitot or static system failure can also cause erratic and unreliable instrument indications. When a static system

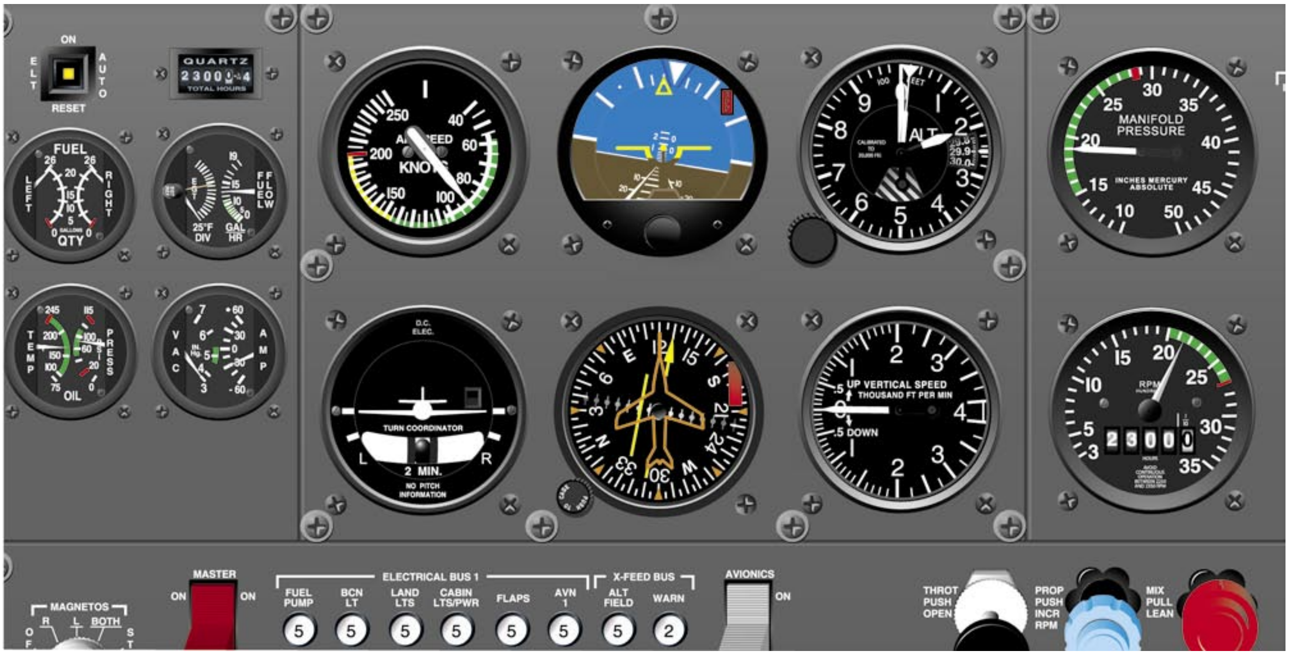
ON
ELT
AUTO
RESET
QUARTZ
TOTAL HOURS
FUEL
LEFT
RIGHT
QTY
TEMP
OIL
PRESS
VAC
AMP
AIRSPEED
KNOTS
ALT
MANIFOLD PRESSURE
INCHES MERCURY ABSOLUTE
D.C. ELEC.
TURN COORDINATOR
L
R
2 MIN.
NO PITCH INFORMATION
UP VERTICAL SPEED THOUSAND FT PER MIN
DOWN
RPM
MASTER
ON
ON
MAGNETOS
R
L
BOTH
ELECTRICAL BUS 1
FUEL PUMP
BCN LT
LAND LTS
CABIN LTS/PWR
FLAPS
AVN 1
X-FEED BUS
ALT FIELD
WARN
AVIONICS
ON
THROT PUSH OPEN
PROP PUSH INCR RPM
MIX PULL LEAN

Figure 11-8. *Vacuum Failure.*

problem occurs, it affects the ASI, altimeter, and the VSI. In most aircraft, provisions have been made for the pilot to select an alternate static source. Check the POH/AFM for the location and operation of the alternate static source. In the absence of an alternate static source, in an unpressurized aircraft, the pilot could break the glass on the VSI. The VSI is not required for instrument flight, and breaking the glass provides the altimeter and the ASI a source of static pressure. This procedure could cause additional instrument errors.

Communication/Navigation System Malfunction

Avionics equipment has become very reliable, and the likelihood of a complete communications failure is remote. However, each IFR flight should be planned and executed in anticipation of a two-way radio failure. At any given point during a flight, the pilot must know exactly what route to fly, what altitude to fly, and when to continue beyond a clearance limit. Title 14 of the Code of Federal Regulations (14 CFR) part 91 describes the procedures to be followed in case of a two-way radio communications failure. If operating in VFR conditions at the time of the failure, the pilot should continue the flight under VFR and land as soon as practicable. If the failure occurs in IFR conditions, or if VFR conditions cannot be maintained, the pilot must continue the flight:

1. Along the route assigned in the last ATC clearance received;
2. If being radar vectored, by the direct route from the point of radio failure to the fix, route, or airway specified in the vector clearance;
3. In the absence of an assigned route, by the route that ATC has advised may be expected in a further clearance; or
4. In the absence of an assigned route or a route that ATC has advised may be expected in a further clearance, by the route filed in the flight plan.

The pilot should maintain the highest of the following altitudes or flight levels for the route segment being flown:

1. The altitude or flight level assigned in the last ATC clearance received;
2. The minimum altitude (converted, if appropriate, to minimum flight level as prescribed in part 91 for IFR operations); or
3. The altitude or flight level ATC has advised may be expected in a further clearance.

In addition to route and altitude, the pilot must also plan the progress of the flight to leave the clearance limit.

1. When the clearance limit is a fix from which an approach begins, commence descent or descent and approach as close as possible to the expect-further-clearance time if one has been received. If an expect-further-clearance time has not been received, commence descent or descent and approach as close as possible to the estimated time of arrival as calculated from the filed or amended (with ATC) estimated time en route.

2. If the clearance limit is not a fix from which an approach begins, leave the clearance limit at the expect-further-clearance time if one has been received. If no expect-further-clearance time has been received, leave the clearance limit upon arrival over it, and proceed to a fix from which an approach begins and commence descent or descent and approach as close as possible to the estimated time of arrival as calculated from the filed or amended (with ATC) estimated time en route. *[Figure 11-8]*

While following these procedures, set the transponder to code 7600 and use all means possible to reestablish two-way radio communication with ATC. This includes monitoring navigational aids (NAVAIDs), attempting radio contact with other aircraft, and attempting contact with a nearby automated flight service station (AFSS).

GPS Nearest Airport Function

Procedures for accessing the nearest airport information vary by the type of display installed in an aircraft. Pilots can obtain information relative to the nearest airport by using the PFD, MFD, or the nearest function on the GPS receiver. The following examples are based on a popular system. Pilots should become familiar with the operational characteristics of the equipment to be used.

Nearest Airports Using the PFD

With the advancements in electronic databases, diverting to alternate airports has become easier. Simply by pressing a soft key on the PFD, pilots can access information for up to 25 of the nearest airports that meet the criteria set in the systems configuration page. *[Figure 11-9]* Pilots are able to specify what airports are acceptable for their aircraft requirements based on landing surface and length of runway.

When the text box opens, the flashing cursor is located over the nearest airport that meets the criteria set in the auxiliary setup page as shown in *Figure 11-10*. Scrolling through the 25 airports is accomplished by turning the outer FMS knob, which is located on the lower right corner of the display screen. Turning the FMS knob clockwise moves the blinking cursor to the next closest airport. By continuing to turn the knob, the pilot is able to scroll through all 25 nearest airports. Each airport box contains the information illustrated in *Figure 11-11,* which the pilot can utilize to determine which airport best suits their individual needs.

Additional Information for a Specific Airport

In addition to the information that is presented on the first screen, the pilot can view additional information as shown in *Figure 11-12* by highlighting the airport identifier and then pressing the enter key.

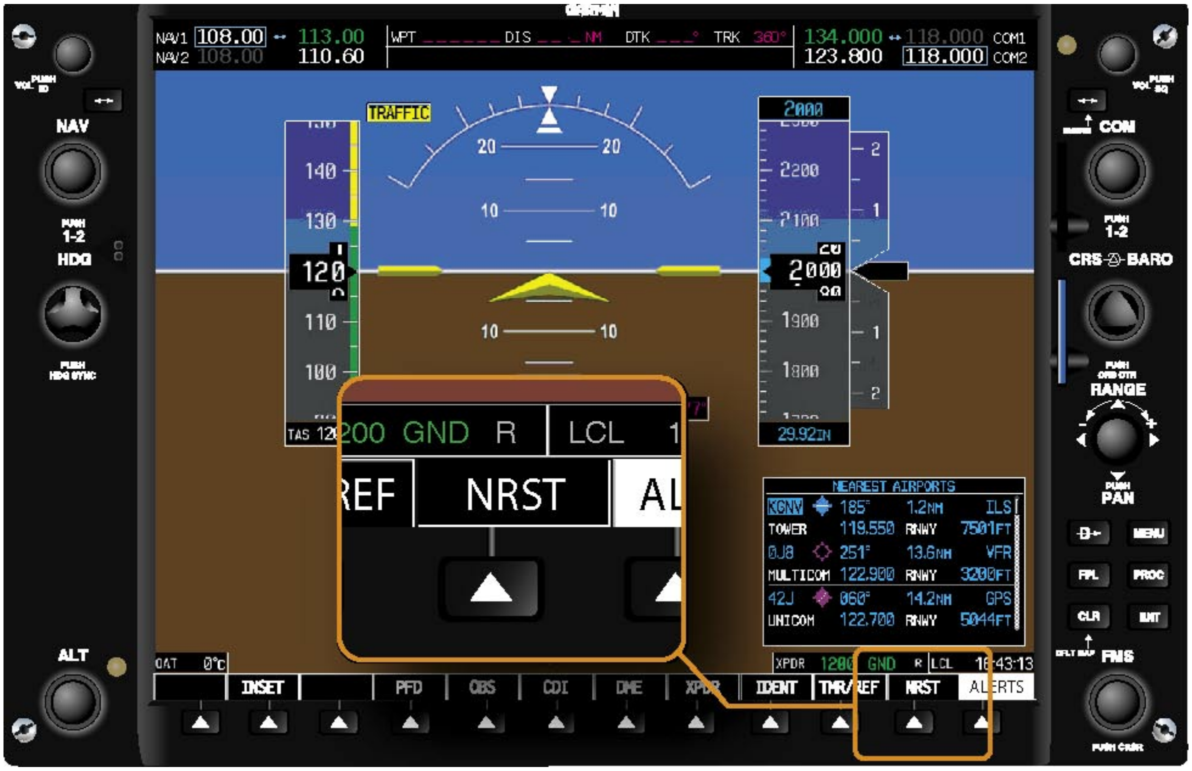
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK ° TRK 360°
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
FMS
MENU
FPL
PROC
CLR
ENT
TRAFFIC
2000
2200
2100
1900
1800
29.92IN
140
130
120
110
100
TAS
200 GND R LCL
REF NRST AL
NEAREST AIRPORTS
KGNV 185° 1.2NM ILS
TOWER 119.550 RNWY 7501FT
0J8 251° 13.6NM VFR
MULTICOM 122.900 RNWY 3200FT
42J 060° 14.2NM GPS
UNICOM 122.700 RNWY 5044FT
OAT 0°C
XPDR 1200 GND R LCL 16:43:13
INSET
PFD
OBS
CDI
DME
XPDR
IDENT
TMR/REF
NRST
ALERTS

Figure 11-9. *The default soft key menu that is displayed on the PFD contains a "NRST" (Nearest Airport) soft key. Pressing this soft key opens a text box which displays the nearest 25 airports.*

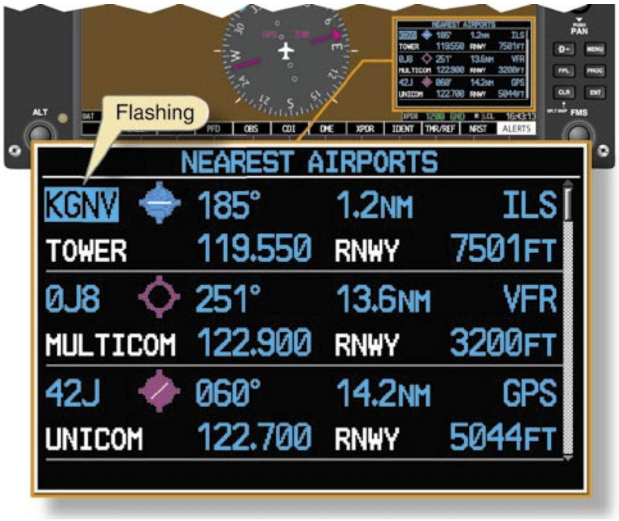

NEAREST AIRPORTS
KGNV 185° 1.2NM ILS
TOWER 119.550 RNWY 7501FT
0J8 251° 13.6NM VFR
MULTICOM 122.900 RNWY 3200FT
42J 060° 14.2NM GPS
UNICOM 122.700 RNWY 5044FT

Figure 11-10. *An enlargement of the box shown in the lower right of Figure 11-9. Note that KGNV would be flashing.*

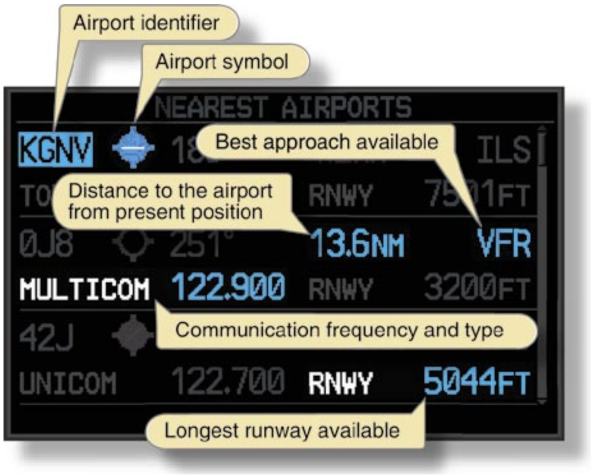
Airport identifier
Airport symbol
NEAREST AIRPORTS
KGNV
Best approach available
ILS
Distance to the airport from present position
RNWY
0J8
251°
13.6NM
VFR
MULTICOM
122.900
RNWY
3200FT
42J
Communication frequency and type
UNICOM
122.700
RNWY
5044FT
Longest runway available

Figure 11-11. *Information shown on the nearest airport page.*

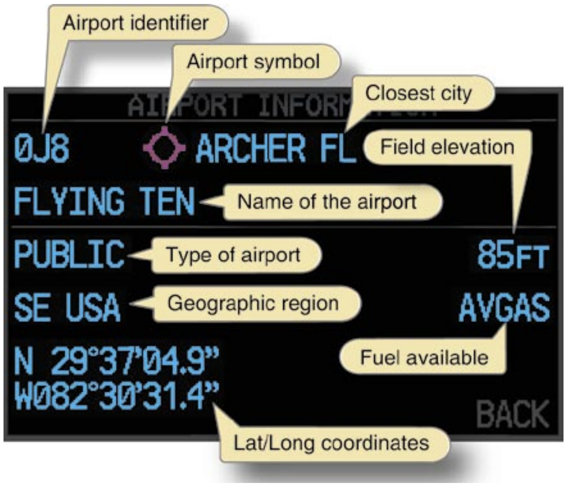
Airport identifier
Airport symbol
Closest city
0J8
ARCHER FL
Field elevation
FLYING TEN
Name of the airport
PUBLIC
Type of airport
85FT
SE USA
Geographic region
AVGAS
N 29°37'04.9"
W082°30'31.4"
Fuel available
BACK
Lat/Long coordinates

Figure 11-12. *Information shown on the additional information page that will aid the pilot in making a more informed decision about which airport to choose when diverting.*

From this menu or the previous default nearest airport screen, the pilot is able to activate the Direct-To function, which provides a direct GPS course to the airport. In addition, the pilot can auto-tune communication frequencies by highlighting the appropriate frequency and then pressing the enter key. The frequency is placed in the stand-by box of either COM1 or COM2, whichever frequency has the cyan box around it.

Nearest Airports Using the MFD

A second way to determine the nearest airport is by referencing the NRST Page Group located on the MFD. This method provides additional information to the pilot; however, it may require additional steps to view. *[Figure 11-13]*

Navigating the MFD Page Groups

Most display systems are designed for ease of navigation through the different screens on the MFD. Notice the various page groups in the lower right-hand corner of the MFD screen. Navigation through these four page groups is accomplished by turning the outer FMS knob clockwise. *[Figure 11-14]*

Within each page group are specific pages that provide additional information pertaining to that specific group. Once the desired page group and page is selected, the MFD remains in that configuration until the page is changed or the CLR button is depressed for more than 2 seconds. Holding the CLR button returns the display to the default moving map page.

Nearest Airport Page Group

The nearest airport page contains specific areas of interest for the airport selected. *[Figure 11-15]* The pilot is furnished information regarding runways, frequencies, and types of approaches available.

Nearest Airports Page Soft Keys

Figure 11-16 illustrates four specific soft keys that allow the pilot to access independent windows of the airport page. Selection of each of these windows can also be accomplished by utilizing the MENU hard key.

The soft keys and functions are as follows: Scroll through each section with the cursor, then press enter to accept the selection.

1. APT. Allows the user access to scroll through the 25 nearest airports. The white arrow indicates which airport is selected. The INFORMATION window is slaved to the white arrow. The INFORMATION window decodes the airport identifier. Scroll through the 25 airports by turning the outer FMS knob.

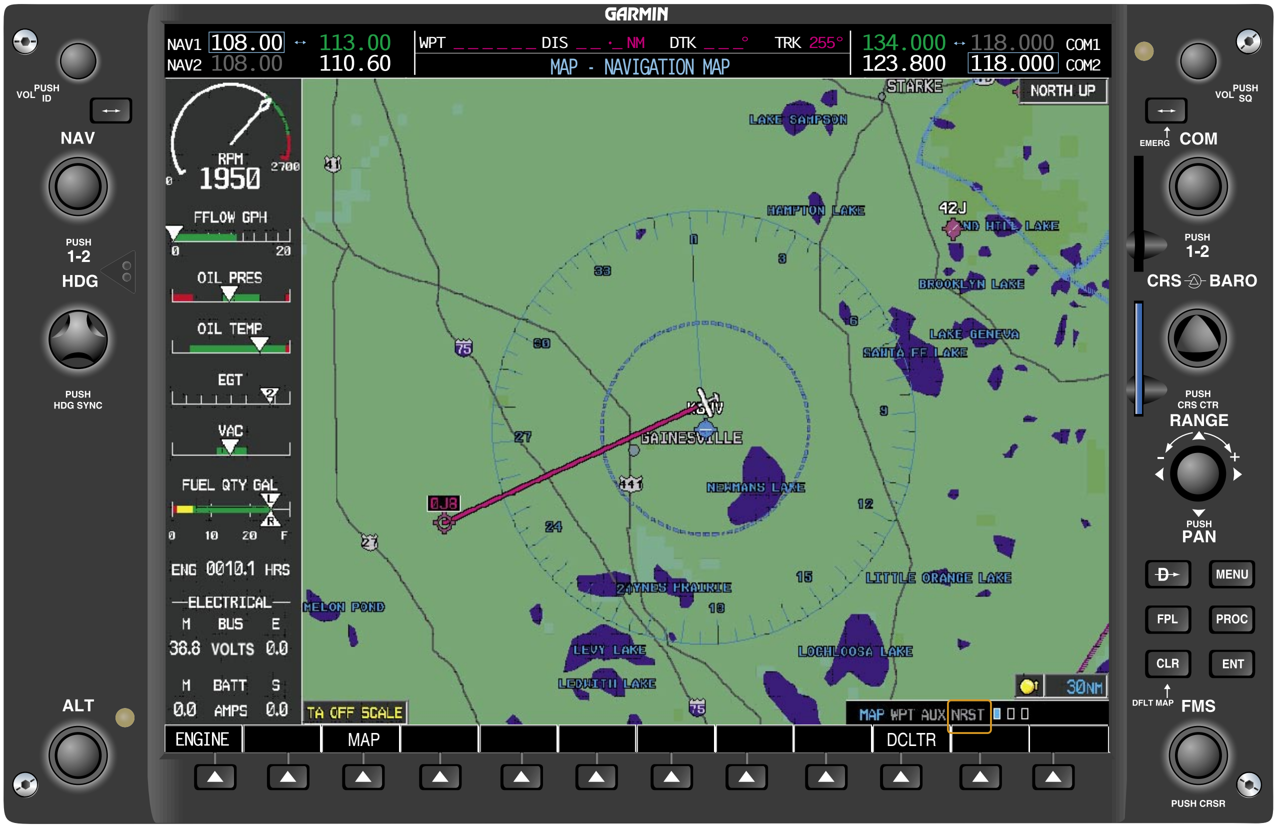

Figure 11-13. *The MFD is another means of viewing the nearest airports.*

2. RNWY. Moves the cursor into the Runways section and allows the user to scroll through the available runways at a specific airport that is selected in conjunction with the APT soft key. A green arrow indicates additional runways to view.
3. FREQ. Moves the cursor into the Frequencies section and allows the pilot to highlight and auto-tune the frequency into the selected standby box.
4. APR. Moves the cursor into the Approach section and allows the pilot to review approaches and load them into the flight plan. When the APR soft key is selected, an additional soft key appears. The LD APR (Load Approach) soft key must be pressed once the desired instrument approach procedure has been highlighted. Once the soft key is pressed, the screen changes to the PROC Page Group. From this page the pilot is able to choose the desired approach, the transition, and choose the option to activate the approach or just load it into the flight plan.

Situational Awareness

Situational awareness (SA) is not simply a mental picture of aircraft location; rather, it is an overall assessment of each element of the environment and how it affects a flight. On one end of the SA spectrum is a pilot who is knowledgeable of every aspect of the flight; consequently, this pilot's decision-making is proactive. With good SA, this pilot is able to make decisions well ahead of time and evaluate several different options. On the other end of the SA spectrum is a pilot who is missing important pieces of the puzzle: "I knew exactly where I was when I ran out of fuel." Consequently, this pilot's decision-making is reactive. With poor SA, a pilot

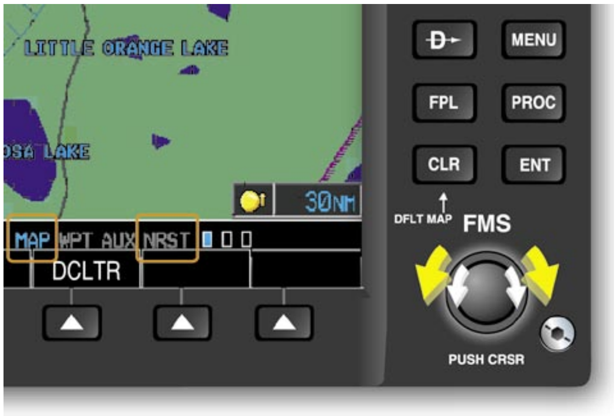

Figure 11-14. *Page Groups. As the FMS outer knob is rotated, the current page group is indicated by highlighting the specific group indicator. Notice that the MAP page group is highlighted.*

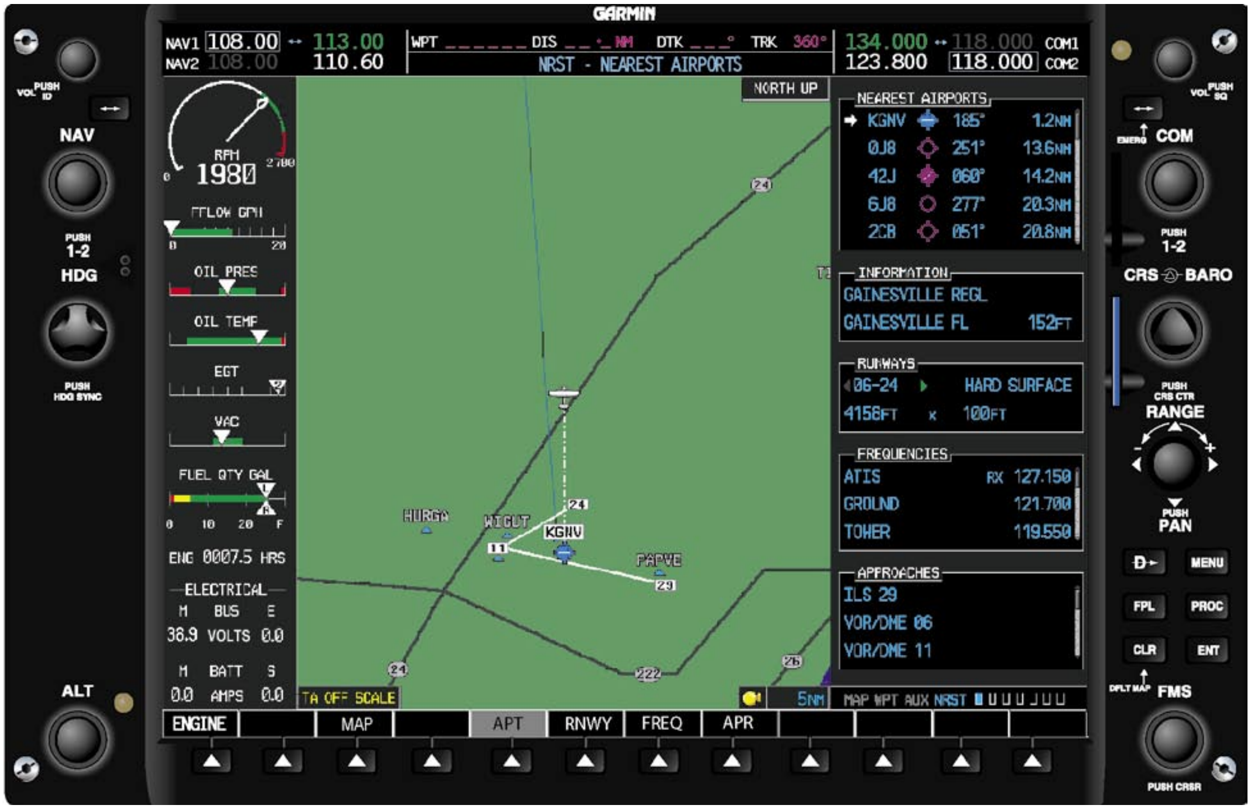
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT DIS NM DTK TRK 360°
NRST - NEAREST AIRPORTS
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
NORTH UP
RPM 1980
FFLOW GPH
OIL PRES
OIL TEMP
EGT
VAC
FUEL QTY GAL
ENG 0007.5 HRS
ELECTRICAL
M BUS E
36.9 VOLTS 0.0
M BATT S
0.0 AMPS 0.0
NEAREST AIRPORTS
KGNV 185° 1.2NM
0J8 251° 13.6NM
42J 060° 14.2NM
6J8 277° 20.3NM
2CB 051° 20.8NM
INFORMATION
GAINESVILLE REGL
GAINESVILLE FL 152FT
RUNWAYS
06-24 HARD SURFACE
4158FT x 100FT
FREQUENCIES
ATIS RX 127.150
GROUND 121.700
TOWER 119.550
APPROACHES
ILS 29
VOR/DME 06
VOR/DME 11
HURGA
WIGLT
KGNV
PAPVE
TA OFF SCALE
5NM
MAP WPT AUX NRST
ENGINE MAP APT RNWY FREQ APR
NAV
HDG
ALT
COM
CRS BARO
RANGE
PAN
MENU
FPL
PROC
CLR
ENT
FMS

Figure 11-15. *The page group of nearest airports has been selected.*

lacks a vision of future events and is forced to make decisions quickly, often with limited options.

During a typical IFR flight, a pilot operates at varying levels of SA. For example, a pilot may be cruising to his or her destination with a high level of SA when ATC issues an unexpected standard terminal arrival route (STAR). Since the pilot was not expecting the STAR and is not familiar with it, SA is lowered. However, after becoming familiar with the STAR and resuming normal navigation, the pilot returns to a higher level of SA.

Factors that reduce SA include: distractions, unusual or unexpected events, complacency, high workload, unfamiliar situations, and inoperative equipment. In some situations, a loss of SA may be beyond a pilot's control. For example, a pneumatic system failure and associated loss of the attitude and heading indicators could cause a pilot to find his or her aircraft in an unusual attitude. In this situation, established procedures must be used to regain SA.

Pilots should be alert to a loss of SA anytime they are in a reactive mindset. To regain SA, reassess the situation and seek additional information from other sources, such as the navigation instruments or ATC.

Traffic Avoidance

Electronic flight displays have the capability of displaying transponder-equipped aircraft on the MFD as well as the inset map on the PFD. However, due to the limitations of the systems, not all traffic is displayed. Some TIS units display only eight intruding targets within the service volume. The normal service volume has altitude limitations of 3,500 feet below the aircraft to 3,500 feet above the aircraft. The lateral limitation is 7 NM. *[Figure 11-17]* Pilots unfamiliar with the limitations of the system may rely on the aural warnings to alert them to approaching traffic.

In addition to an outside visual scan of traffic, a pilot should incorporate any Traffic Information electronically displayed such as TIS. This innovation in traffic alerting reinforces and adds synergy to the ability to see and avoid. However, it is an aid and not a replacement for the responsibilities of the pilot. Systems such as TIS provide a visual representation of nearby traffic and displays a symbol on the moving map display with relative information about altitude, vertical trends, and direction of flight. *[Figure 11-18]*

It is important to remember that most systems display only a specific maximum number of targets allowed. Therefore, it

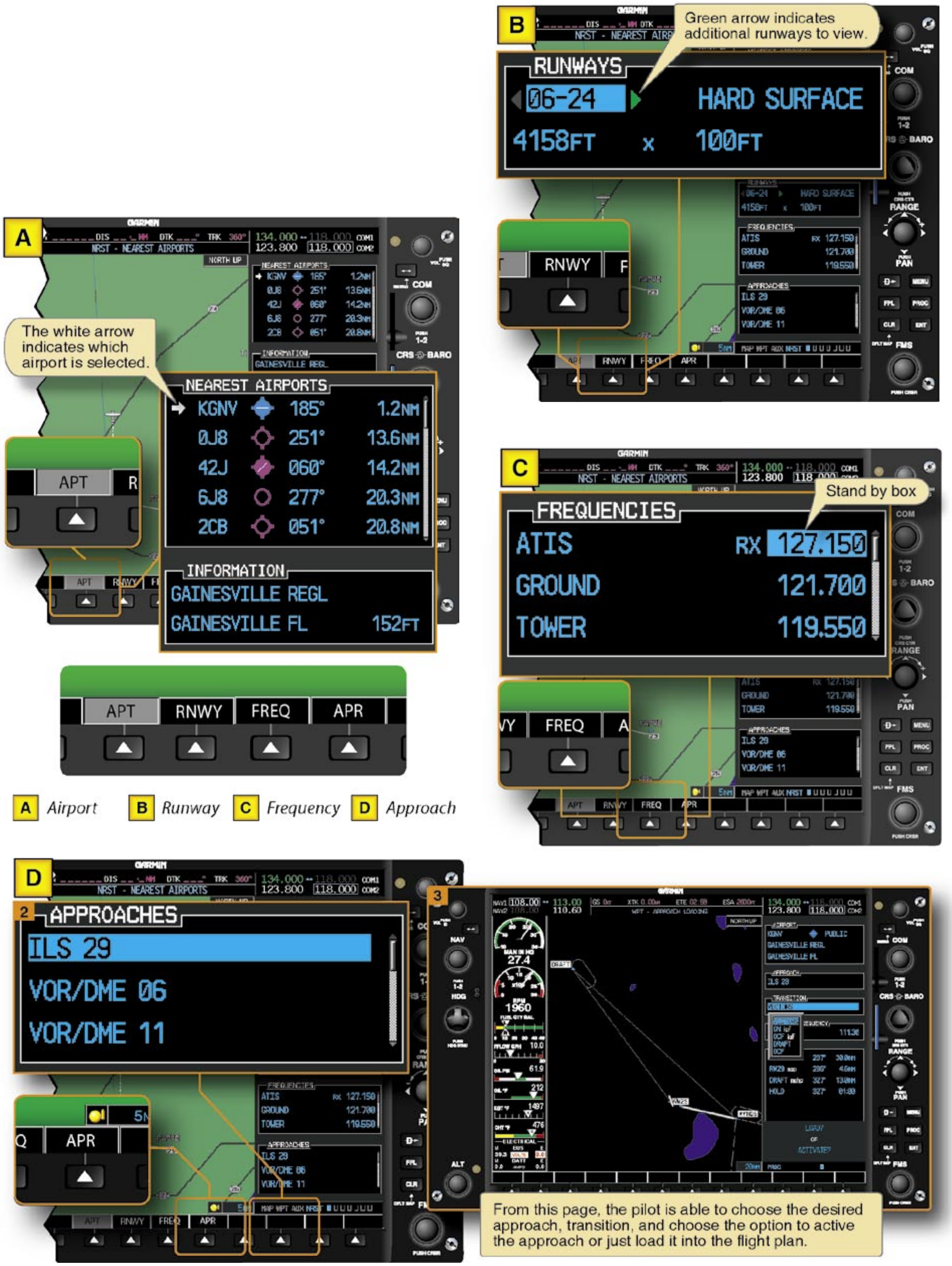

A
B
C
D
2
3
A Airport
B Runway
C Frequency
D Approach
APT
RNWY
FREQ
APR
RUNWAYS
06-24
HARD SURFACE
4158FT x 100FT
NEAREST AIRPORTS
INFORMATION
GAINESVILLE REGL
GAINESVILLE FL 152FT
FREQUENCIES
ATIS RX 127.150
GROUND 121.700
TOWER 119.550
APPROACHES
ILS 29
VOR/DME 06
VOR/DME 11

NEAREST AIRPORTS		
KGNV	185°	1.2NM
0J8	251°	13.6NM
42J	060°	14.2NM
6J8	277°	20.3NM
2CB	051°	20.8NM

Green arrow indicates additional runways to view.

The white arrow indicates which airport is selected.

Stand by box

From this page, the pilot is able to choose the desired approach, transition, and choose the option to active the approach or just load it into the flight plan.

Figure 11-16. *The four soft keys at the bottom of the MFD are airport (A), runway (B), frequency (C), and approach (D).*

does not mean that the targets displayed are the only aircraft in the vicinity. The system displays only the closest aircraft. In addition, the system does not display aircraft that are not equipped with transponders. The display may not show any aircraft; however, a Piper Cub with no transponder could be flying in the area. TIS coverage can be sporadic and is not available in some areas of the United States. Traffic advisory software is to be utilized only for increased situational awareness and not the sole means of traffic avoidance. There is no substitute for a good visual scan of the surrounding sky.

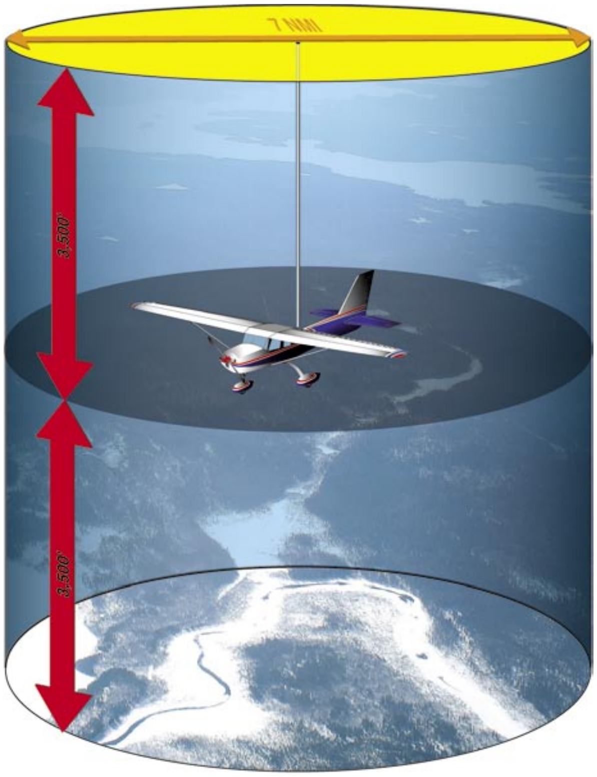
7 NM
3,500'
3,500'

Figure 11-17. *The Area Surrounding the Aircraft for Coverage Using TIS.*

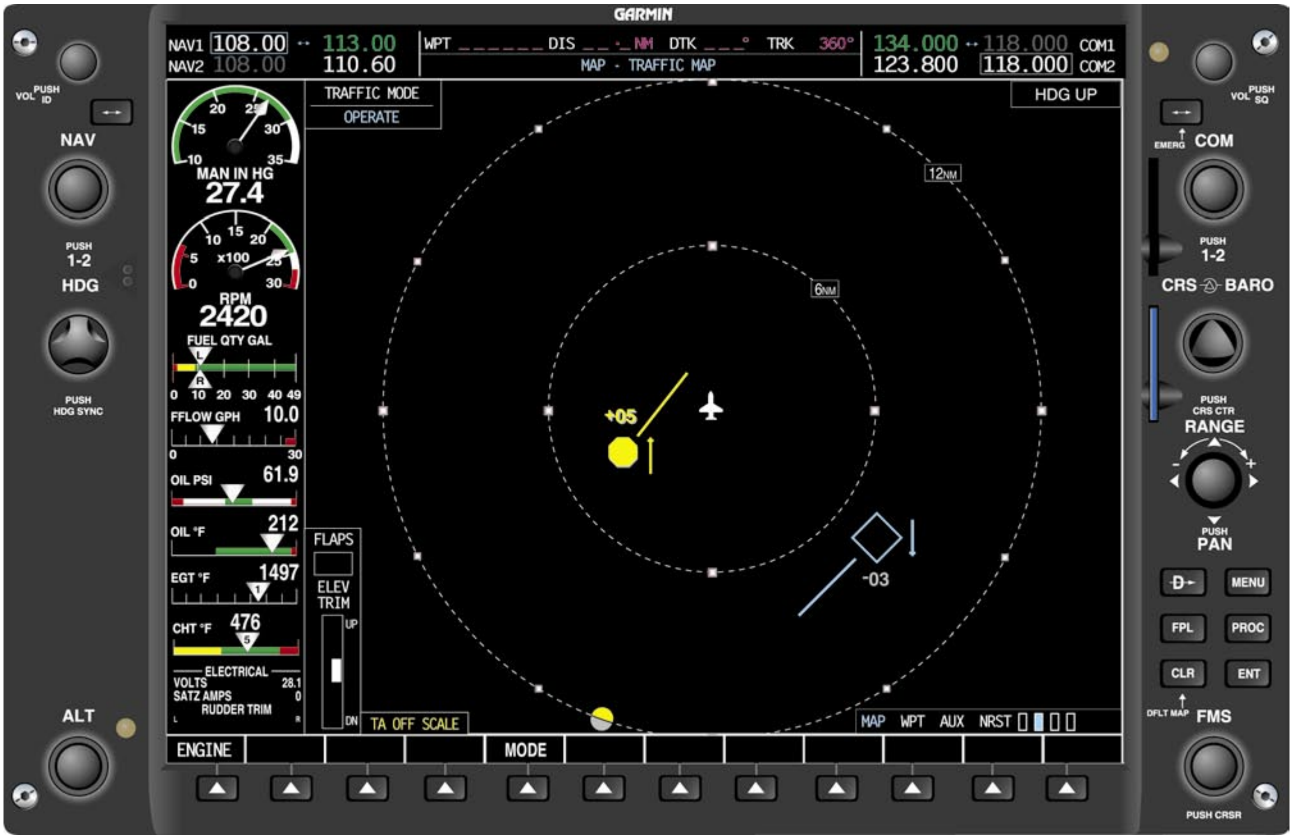
GARMIN
NAV1 108.00 ↔ 113.00
NAV2 108.00 110.60
WPT ______ DIS __._NM DTK ___° TRK 360°
MAP - TRAFFIC MAP
134.000 ↔ 118.000 COM1
123.800 118.000 COM2
TRAFFIC MODE
OPERATE
HDG UP
12NM
6NM
+05
-03
MAN IN HG
27.4
RPM
2420
FUEL QTY GAL
FFLOW GPH 10.0
OIL PSI 61.9
OIL °F 212
EGT °F 1497
CHT °F 476
ELECTRICAL
VOLTS 28.1
SATZ AMPS 0
RUDDER TRIM
FLAPS
ELEV TRIM
UP
DN
TA OFF SCALE
MAP WPT AUX NRST
ENGINE
MODE
VOL PUSH ID
NAV
PUSH 1-2
HDG
PUSH HDG SYNC
ALT
VOL PUSH SQ
EMERG COM
PUSH 1-2
CRS BARO
PUSH CRS CTR
RANGE
PUSH PAN
MENU
FPL
PROC
CLR
ENT
DFLT MAP
FMS
PUSH CRSR

Figure 11-18. *A Typical Display on Aircraft MFD When Using TIS.*

Appendix A

Clearance Shorthand

The following shorthand system is recommended by the Federal Aviation Administration (FAA). Applicants for the instrument rating may use any shorthand system, in any language, which ensures accurate compliance with air traffic control (ATC) instructions. No shorthand system is required by regulation and no knowledge of shorthand is required for the FAA Knowledge Test; however, because of the vital need for reliable communication between the pilot and controller, clearance information should be unmistakably clear.

The following symbols and contractions represent words and phrases frequently used in clearances. Most are used regularly by ATC personnel. By practicing this shorthand, omitting the parenthetical words, you will be able to copy long clearances as fast as they are read.

Example: CAF $\xrightarrow{\text{M}}$ RH RV V18 ↑40 SQ 0700 DPC 120.4
Cleared as filed, maintain runway heading for radar vector to Victor 18, climb to 4,000, squawk 0700, departure control frequency is 120.4.

Words and Phrases	**Shorthand**
Above	ABV
Above (Altitude, Hundreds of Feet)	$\underline{70}$
Adjust speed to 250 knots	250 K
Advise	ADZ
After (Passing)	<
Airway (Designation)	V26
Airport	A
Alternate Instructions	( )
Altitude 6,000–17,000	60-170
And	&
Approach	AP
Approach Control	APC
Area Navigation	RNAV
Arriving	↓
At	@
At or Above	$\underline{\uparrow}$
At or Below	$\overline{\downarrow}$
(ATC) Advises	CA
(ATC) Clears or Cleared	C
(ATC) Requests	CR
Back Course	BC
Bearing	BR
Before (Reaching, Passing)	>
Below	BLO
Below (Altitude, Hundreds of Feet)	$\overline{70}$
Center	CTR
Clearance Void if Not Off By (Time)	v<
Cleared as Filed	CAF
Cleared to Airport	A
Cleared to Climb/Descend at Pilot's Discretion	PD
Cleared to Cross	X
Cleared to Depart From the Fix	D
Cleared to the Fix	F
Cleared to Hold and Instructions Issued	H
Cleared to Land	L
Cleared to the Outer Marker	O
Climb to (Altitude, Hundreds of Feet)	↑70
Contact Approach	CT
Contact (Denver) Approach Control	(den
Contact (Denver) Center	(DEN
Course	CRS
Cross	X
Cruise	→
Delay Indefinite	DLI
Depart (Direction, if Specified)	T→( )
Departure Control	DPC
Descend To (Altitude, Hundreds of Feet)	↓70
Direct	DR
Direction (Bound)	
Eastbound	EB
Westbound	WB
Northbound	NB
Southbound	SB
Inbound	IB
Outbound	OB
DME Fix (Mile)	$\boxed{21}$
Each	EA
Enter Control Area	△
Estimated Time of Arrival	ETA
Expect	EX
Expect-Further-Clearance	EFC

Fan Marker	FM
Final	F
Final Approach	FA
Flight Level	FL
Flight Planned Route	FPR
For Further Clearance	FFC
For Further Headings	FFH
From	FM
Ground	GND
GPS Approach	GPS
Heading	HDG
Hold (Direction)	H-W
Holding Pattern	⬭
ILS Approach	ILS
Increase Speed 30 Knots	+30 K
Initial Approach	I
Instrument Departure Procedure	DP
Intersection	XN
Join or Intercept Airway/Jet Route/Track or Course	$\overline{>}$
Left Turn After Takeoff	↰
Locator Outer Marker	LOM
Magnetic	M
Maintain	~~M~~>
Maintain VFR Conditions On Top	$\underline{\text{VFR}}$
Middle Compass Locator	ML
Middle Marker	MM
Missed Approach	MA
Nondirectional Beacon Approach	NDB
Out of (Leave) Control Area	△↗
Outer Marker	OM
Over (Station)	$\underline{\text{OKC}}$
On Course	OC
Precision Approach Radar	PAR
Procedure Turn	PT
Radar Vector	RV
Radial (080° Radial)	080R
Reduce Speed 20 Knots	-20 K
Remain This Frequency	RTF
Remain Well to Left Side	LS
Remain Well to Right Side	RS
Report Crossing	RX
Report Departing	RD
Report Leaving	RL
Report on Course	R-CRS
Report Over	RO
Report Passing	RP
Report Reaching	RR
Report Starting Procedure Turn	RSPT
Reverse Course	RC
Right Turn After Takeoff	↱
Runway Heading	RH
Runway (Number)	RY18
Squawk	SQ
Standby	STBY
Straight-in Approach	SI
Surveillance Radar Approach	ASR
Takeoff (Direction)	T→N
Tower	Z
Turn Left	TL
Turn Right	TR
Until	/
Until Advised (By)	UA
Until Further Advised	UFA
VFR Conditions On Top	OTP
Via	VIA
Victor (Airway Number)	V14
Visual Approach	VA
VOR	⊙
VOR Approach	VR
VORTAC	Ⓣ
While in Control Area	△

Appendix B

Instrument Training Lesson Guide

Introduction

Flight instructors may use this guide in the development of lesson plans. The lessons are arranged in a logical learning sequence and use the building-block technique. Each lesson includes ground training appropriate to the flight portion of the lesson. It is vitally important that the flight instructor brief the student on the objective of the lesson and how it will be accomplished. Debriefing the student's performance is also necessary to motivate further progress. To ensure steady progress, student pilots should master the objective of each lesson before advancing to the next lesson. Lessons should be arranged to take advantage of each student's knowledge and skills.

Flight instructors must monitor progress closely during training to guide student pilots in how to properly divide their attention. The importance of this division of attention or "cross-check" cannot be overemphasized. Cross-check and proper instrument interpretation are essential components of "attitude instrument flying" that enables student pilots to accurately visualize the aircraft's attitude at all times.

When possible, each lesson should incorporate radio communications, basic navigation, and emergency procedures so the student pilot is exposed to the entire IFR experience with each flight. Cross-reference the Instrument Training Lesson Guide with this handbook and the Instrument Practical Test Standards for a comprehensive instrument rating training program.

Lesson 1—Ground and flight evaluation of student's knowledge and performance

Aircraft systems
Aircraft performance
Preflight planning
Use of checklists
Basic flight maneuvers
Radio communications procedures
Navigation systems

Lesson 2—Preflight preparation and flight by reference to instruments

Ground Training

Instrument system preflight procedures
Attitude instrument flying
Fundamental instrument skills
Instrument cross-check techniques

Flight Training

Aircraft and instrument preflight inspection
Use of checklists
Fundamental instrument skills
Basic flight maneuvers
Instrument approach (demonstrated)
Postflight procedures

Lesson 3—Flight instruments and human factors

Ground Training

Human factors
Flight instruments and systems
Aircraft systems
Navigation instruments and systems

Flight Training

Aircraft and instrument preflight inspection
Radio communications
Checklist procedures
Attitude instrument flying
Fundamental instrument skills
Basic flight maneuvers
Spatial disorientation demonstration
Navigation systems
Postflight procedures

Lesson 4—Attitude instrument flying

Ground Training

Human factors
Flight instruments and systems

Aircraft systems
Navigation instruments and systems
Attitude instrument flying
Fundamental instrument skills
Basic flight maneuvers

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Attitude instrument flying
Fundamental instrument skills
Basic flight maneuvers
Spatial disorientation
Navigation
Postflight procedures

Lesson 5—Aerodynamic factors and basic flight maneuvers

Ground Training

Basic aerodynamic factors
Basic instrument flight patterns
Emergency procedures

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Basic instrument flight patterns
Emergency procedures
Navigation
Postflight procedures

Lesson 6—Partial panel operations

Ground Training

ATC system
Flight instruments
Partial panel operations

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Basic instrument flight patterns
Emergency procedures
Partial panel practice
Navigation
Postflight procedures

Lesson 7—Recovery from unusual attitudes

Ground Training

Attitude instrument flying
ATC system
NAS overview

Flight Training

Preflight
Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Instrument takeoff
Navigation
Partial panel practice
Recovery from unusual attitudes
Postflight procedures

Lesson 8—Navigation systems

Ground Training

ATC clearances
Departure procedures
IFR en route charts

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Intercepting and tracking
Holding
Postflight procedures

Lesson 9—Review and practice

Ground Training

Aerodynamic factors
Flight instruments and systems
Attitude instrument flying
Navigation systems
NAS
ATC
Emergency procedures

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Review and practice as determined by the flight instructor

Instrument takeoff
Radio communications
Navigation systems
Emergency procedures
Postflight procedures

Lessons 10 through 19—Orientation, intercepting, tracking, and holding using each navigation system installed in the aircraft

Ground Training

Preflight planning
Navigation systems
NAS
ATC
Emergencies

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Departure procedures
En route navigation
Terminal operations
Partial panel operation
Instrument approach
Missed approach
Approach to a landing
Postflight procedures

Lessons 20 and 21—Cross-country flights

Ground Training

Preflight planning
Aircraft performance
Navigation systems
NAS
ATC
Emergencies

Flight Training

Emergency procedures
Partial panel operation
Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Departure procedures
En route navigation
Terminal operations
Instrument approach
Missed approach
Approach to a landing
Postflight procedures

Lessons 22 and 23—Review and practice

Ground Training

Human factors
Aerodynamic factors
Flight instruments and systems
Attitude instrument flying
Basic flight maneuvers
Navigation systems
NAS
ATC
Emergency operations

Flight Training

Aircraft and instrument preflight inspection
Checklist procedures
Radio communications
Review and practice as determined by the flight instructor
Instrument takeoff
Partial panel operations
Unusual attitude recoveries
Radio communications
Navigation systems
Emergency procedures
Postflight procedures

Lessons 24 and subsequent—Practical test preparation

Ground Training

Title 14 of the Code of Federal Regulations (14 CFR) parts 61, 71, 91, 95, and 97
Instrument Flying Handbook
Practical test standards
Administrative requirements
Equipment requirements
Applicant's requirements

Flight Training

Review and practice until the student can consistently perform all required tasks in accordance with the appropriate practical test standards.

NOTE: It is the recommending instructor's responsibility to ensure that the applicant meets 14 CFR part 61 requirements and is prepared for the practical test, including: training, knowledge, experience, and the appropriate instructor endorsements.

Glossary

Absolute accuracy. The ability to determine present position in space independently, and is most often used by pilots.

Absolute altitude. The actual distance between an aircraft and the terrain over which it is flying.

Absolute pressure. Pressure measured from the reference of zero pressure, or a vacuum.

A.C. Alternating current.

Acceleration error. A magnetic compass error apparent when the aircraft accelerates while flying on an easterly or westerly heading, causing the compass card to rotate toward North.

Accelerometer. A part of an inertial navigation system (INS) that accurately measures the force of acceleration in one direction.

ADF. See automatic direction finder.

ADI. See attitude director indicator.

ADM. See aeronautical decision-making.

ADS–B. See automatic dependent surveillance–broadcast.

Adverse yaw. A flight condition at the beginning of a turn in which the nose of the aircraft starts to move in the direction opposite the direction the turn is being made, caused by the induced drag produced by the downward-deflected aileron holding back the wing as it begins to rise.

Aeronautical decision-making (ADM). A systematic approach to the mental process used by pilots to consistently determine the best course of action in response to a given set of circumstances.

A/FD. See Airport/Facility Directory.

Agonic line. An irregular imaginary line across the surface of the Earth along which the magnetic and geographic poles are in alignment, and along which there is no magnetic variation.

Aircraft approach category. A performance grouping of aircraft based on a speed of 1.3 times the stall speed in the landing configuration at maximum gross landing weight.

Air data computer (ADC). An aircraft computer that receives and processes pitot pressure, static pressure, and temperature to calculate very precise altitude, indicated airspeed, true airspeed, and air temperature.

AIRMET. Inflight weather advisory issued as an amendment to the area forecast, concerning weather phenomena of operational interest to all aircraft and that is potentially hazardous to aircraft with limited capability due to lack of equipment, instrumentation, or pilot qualifications.

Airport diagram. The section of an instrument approach procedure chart that shows a detailed diagram of the airport. This diagram includes surface features and airport configuration information.

Airport/Facility Directory (A/FD). An FAA publication containing information on all airports, communications, and NAVAIDs.

Airport surface detection equipment (ASDE). Radar equipment specifically designed to detect all principal features and traffic on the surface of an airport, presenting the entire image on the control tower console; used to augment visual observation by tower personnel of aircraft and/or vehicular movements on runways and taxiways.

Airport surveillance radar (ASR). Approach control radar used to detect and display an aircraft's position in the terminal area.

Airport surveillance radar approach. An instrument approach in which ATC issues instructions for pilot compliance based on aircraft position in relation to the final approach course and the distance from the end of the runway as displayed on the controller's radar scope.

Air route surveillance radar (ARSR). Air route traffic control center (ARTCC) radar used primarily to detect and display an aircraft's position while en route between terminal areas.

Air route traffic control center (ARTCC). Provides ATC service to aircraft operating on IFR flight plans within controlled airspace and principally during the en route phase of flight.

Airspeed indicator. A differential pressure gauge that measures the dynamic pressure of the air through which the aircraft is flying. Displays the craft's airspeed, typically in knots, to the pilot.

Air traffic control radar beacon system (ATCRBS). Sometimes called secondary surveillance radar (SSR), which utilizes a transponder in the aircraft. The ground equipment is an interrogating unit, in which the beacon antenna is mounted so it rotates with the surveillance antenna. The interrogating unit transmits a coded pulse sequence that actuates the aircraft transponder. The transponder answers the coded sequence by transmitting a preselected coded sequence back to the ground equipment, providing a strong return signal and positive aircraft identification, as well as other special data.

Airway. An airway is based on a centerline that extends from one navigation aid or intersection to another navigation aid (or through several navigation aids or intersections); used to establish a known route for en route procedures between terminal areas.

Alert area. An area in which there is a high volume of pilot training or an unusual type of aeronautical activity.

Almanac data. Information the global positioning system (GPS) receiver can obtain from one satellite which describes the approximate orbital positioning of all satellites in the constellation. This information is necessary for the GPS receiver to know what satellites to look for in the sky at a given time.

ALS. See approach lighting system.

Alternate airport. An airport designated in an IFR flight plan, providing a suitable destination if a landing at the intended airport becomes inadvisable.

Alternate static source valve. A valve in the instrument static air system that supplies reference air pressure to the altimeter, airspeed indicator, and vertical speed indicator if the normal static pickup should become clogged or iced over.

Altimeter setting. Station pressure (the barometric pressure at the location the reading is taken) which has been corrected for the height of the station above sea level.

AME. See aviation medical examiner.

Amendment status. The circulation date and revision number of an instrument approach procedure, printed above the procedure identification.

Ammeter. An instrument installed in series with an electrical load used to measure the amount of current flowing through the load.

Aneroid. The sensitive component in an altimeter or barometer that measures the absolute pressure of the air. It is a sealed, flat capsule made of thin disks of corrugated metal soldered together and evacuated by pumping all of the air out of it.

Aneroid barometer. An instrument that measures the absolute pressure of the atmosphere by balancing the weight of the air above it against the spring action of the aneroid.

Angle of attack. The acute angle formed between the chord line of an airfoil and the direction of the air striking the airfoil.

Anti-ice. Preventing the accumulation of ice on an aircraft structure via a system designed for that purpose.

Approach lighting system (ALS). Provides lights that will penetrate the atmosphere far enough from touchdown to give directional, distance, and glide path information for safe transition from instrument to visual flight.

Area chart. Part of the low-altitude en route chart series, this chart furnishes terminal data at a larger scale for congested areas.

Area navigation (RNAV). Allows a pilot to fly a selected course to a predetermined point without the need to overfly ground-based navigation facilities, by using waypoints.

ARSR. See air route surveillance radar.

ARTCC. See air route traffic control center.

ASDE. See airport surface detection equipment.

ASOS. See automated surface observing station.

ASR. See airport surveillance radar.

ATC. Air Traffic Control.

ATCRBS. See air traffic control radar beacon system.

ATIS. See automatic terminal information service.

Atmospheric propagation delay. A bending of the electromagnetic (EM) wave from the satellite that creates an error in the GPS system.

Attitude and heading reference systems (AHRS). System composed of three-axis sensors that provide heading, attitude, and yaw information for aircraft. AHRS are designed to replace traditional mechanical gyroscopic flight instruments and provide superior reliability and accuracy.

Attitude director indicator (ADI). An aircraft attitude indicator that incorporates flight command bars to provide pitch and roll commands.

Attitude indicator. The foundation for all instrument flight, this instrument reflects the airplane's attitude in relation to the horizon.

Attitude instrument flying. Controlling the aircraft by reference to the instruments rather than by outside visual cues.

Autokinesis. Nighttime visual illusion that a stationary light is moving, which becomes apparent after several seconds of staring at the light.

Automated Weather Observing System (AWOS). Automated weather reporting system consisting of various sensors, a processor, a computer-generated voice subsystem, and a transmitter to broadcast weather data.

Automated Surface Observing Station (ASOS). Weather reporting system which provides surface observations every minute via digitized voice broadcasts and printed reports.

Automatic dependent surveillance–broadcast (ADS-B). A device used in aircraft that repeatedly broadcasts a message that includes position (such as latitude, longitude, and altitude), velocity, and possibly other information.

Automatic direction finder (ADF). Electronic navigation equipment that operates in the low- and medium-frequency bands. Used in conjunction with the ground-based nondirectional beacon (NDB), the instrument displays the number of degrees clockwise from the nose of the aircraft to the station being received.

Automatic terminal information service (ATIS). The continuous broadcast of recorded non-control information in selected terminal areas. Its purpose is to improve controller effectiveness and relieve frequency congestion by automating repetitive transmission of essential but routine information.

Aviation medical examiner (AME). A physician with training in aviation medicine designated by the Civil Aerospace Medical Institute (CAMI).

AWOS. See automated weather observing system.

Azimuth card. A card that may be set, gyroscopically controlled, or driven by a remote compass.

Back course (BC). The reciprocal of the localizer course for an ILS. When flying a back-course approach, an aircraft approaches the instrument runway from the end at which the localizer antennas are installed.

Baro-aiding. A method of augmenting the GPS integrity solution by using a non-satellite input source. To ensure that baro-aiding is available, the current altimeter setting must be entered as described in the operating manual.

Barometric scale. A scale on the dial of an altimeter to which the pilot sets the barometric pressure level from which the altitude shown by the pointers is measured.

BC. See back course.

Block altitude. A block of altitudes assigned by ATC to allow altitude deviations; for example, "Maintain block altitude 9 to 11 thousand."

Cage. The black markings on the ball instrument indicating its neutral position.

Calibrated. The instrument indication compared with a standard value to determine the accuracy of the instrument.

Calibrated orifice. A hole of specific diameter used to delay the pressure change in the case of a vertical speed indicator.

Calibrated airspeed. The speed at which the aircraft is moving through the air, found by correcting IAS for instrument and position errors.

CAS. Calibrated airspeed.

CDI. Course deviation indicator.

Changeover point (COP). A point along the route or airway segment between two adjacent navigation facilities or waypoints where changeover in navigation guidance should occur.

Circling approach. A maneuver initiated by the pilot to align the aircraft with a runway for landing when a straight-in landing from an instrument approach is not possible or is not desirable.

Class A airspace. Airspace from 18,000 feet MSL up to and including FL 600, including the airspace overlying the waters within 12 NM of the coast of the 48 contiguous states and Alaska; and designated international airspace beyond 12 NM of the coast of the 48 contiguous states and Alaska within areas of domestic radio navigational signal or ATC radar coverage, and within which domestic procedures are applied.

Class B airspace. Airspace from the surface to 10,000 feet MSL surrounding the nation's busiest airports in terms of IFR operations or passenger numbers. The configuration of each Class B airspace is individually tailored and consists of a surface area and two or more layers, and is designed to contain all published instrument procedures once an aircraft enters the airspace. For all aircraft, an ATC clearance is required to operate in the area, and aircraft so cleared receive separation services within the airspace.

Class C airspace. Airspace from the surface to 4,000 feet above the airport elevation (charted in MSL) surrounding those airports having an operational control tower, serviced by radar approach control, and having a certain number of IFR operations or passenger numbers. Although the configuration of each Class C airspace area is individually tailored, the airspace usually consists of a 5 NM radius core surface area that extends from the surface up to 4,000 feet above the airport elevation, and a 10 NM radius shelf area that extends from 1,200 feet to 4,000 feet above the airport elevation.

Class D airspace. Airspace from the surface to 2,500 feet above the airport elevation (charted in MSL) surrounding those airports that have an operational control tower. The configuration of each Class D airspace area is individually tailored, and when instrument procedures are published, the airspace is normally designed to contain the procedures.

Class E airspace. Airspace that is not Class A, Class B, Class C, or Class D, and is controlled airspace.

Class G airspace. Airspace that is uncontrolled, except when associated with a temporary control tower, and has not been designated as Class A, Class B, Class C, Class D, or Class E airspace.

Clean configuration. A configuration in which all flight control surfaces have been placed to create minimum drag. In most aircraft this means flaps and gear retracted.

Clearance. ATC permission for an aircraft to proceed under specified traffic conditions within controlled airspace, for the purpose of providing separation between known aircraft.

Clearance delivery. Control tower position responsible for transmitting departure clearances to IFR flights.

Clearance limit. The fix, point, or location to which an aircraft is cleared when issued an air traffic clearance.

Clearance on request. An IFR clearance not yet received after filing a flight plan.

Clearance void time. Used by ATC, the time at which the departure clearance is automatically canceled if takeoff has not been made. The pilot must obtain a new clearance or cancel the IFR flight plan if not off by the specified time.

Clear ice. Glossy, clear, or translucent ice formed by the relatively slow freezing of large, supercooled water droplets.

Compass course. A true course corrected for variation and deviation errors.

Compass locator. A low-power, low- or medium-frequency (L/MF) radio beacon installed at the site of the outer or middle marker of an ILS.

Compass rose. A small circle graduated in 360° increments, printed on navigational charts to show the amount of compass variation at different locations, or on instruments to indicate direction.

Computer navigation fix. A point used to define a navigation track for an airborne computer system such as GPS or FMS.

Concentric rings. Dashed-line circles depicted in the plan view of IAP charts, outside of the reference circle, that show en route and feeder facilities.

Cone of confusion. A cone-shaped volume of airspace directly above a VOR station where no signal is received, causing the CDI to fluctuate.

Control and performance. A method of attitude instrument flying in which one instrument is used for making attitude changes, and the other instruments are used to monitor the progress of the change.

Control display unit. A display interfaced with the master computer, providing the pilot with a single control point for all navigations systems, thereby reducing the number of required flight deck panels.

Controlled airspace. An airspace of defined dimensions within which ATC service is provided to IFR and VFR flights in accordance with the airspace classification. It includes Class A, Class B, Class C, Class D, and Class E airspace.

Control pressures. The amount of physical exertion on the control column necessary to achieve the desired attitude.

Convective weather. Unstable, rising air found in cumiliform clouds.

Convective SIGMET. Weather advisory concerning convective weather significant to the safety of all aircraft, including thunderstorms, hail, and tornadoes.

Coordinated flight. Flight with a minimum disturbance of the forces maintaining equilibrium, established via effective control use.

COP. See changeover point.

Coriolis illusion. The illusion of rotation or movement in an entirely different axis, caused by an abrupt head movement, while in a prolonged constant rate turn that has ceased stimulating the brain's motion sensing system.

Crew resource management (CRM). The effective use of all available resources—human, hardware, and information.

Critical areas. Areas where disturbances to the ILS localizer and glide slope courses may occur when surface vehicles or aircraft operate near the localizer or glide slope antennas.

CRM. See crew resource management.

Cross-check. The first fundamental skill of instrument flight, also known as "scan," the continuous and logical observation of instruments for attitude and performance information.

Cruise clearance. An ATC clearance issued to allow a pilot to conduct flight at any altitude from the minimum IFR altitude up to and including the altitude specified in the clearance. Also authorizes a pilot to proceed to and make an approach at the destination airport.

Current induction. An electrical current being induced into, or generated in, any conductor that is crossed by lines of flux from any magnet.

DA. See decision altitude.

D.C. Direct current.

Dark adaptation. Physical and chemical adjustments of the eye that make vision possible in relative darkness.

Deceleration error. A magnetic compass error that occurs when the aircraft decelerates while flying on an easterly or westerly heading, causing the compass card to rotate toward South.

Decision altitude (DA). A specified altitude in the precision approach, charted in feet MSL, at which a missed approach must be initiated if the required visual reference to continue the approach has not been established.

Decision height (DH). A specified altitude in the precision approach, charted in height above threshold elevation, at which a decision must be made either to continue the approach or to execute a missed approach.

Deice. The act of removing ice accumulation from an aircraft structure.

Density altitude. Pressure altitude corrected for nonstandard temperature. Density altitude is used in computing the performance of an aircraft and its engines.

Departure procedure (DP). Preplanned IFR ATC departure, published for pilot use, in textual and graphic format.

Deviation. A magnetic compass error caused by local magnetic fields within the aircraft. Deviation error is different on each heading.

DGPS. Differential global positioning system.

DH. See decision height.

Differential Global Positioning System (DGPS). A system that improves the accuracy of Global Navigation Satellite Systems (GNSS) by measuring changes in variables to provide satellite positioning corrections.

Direct indication. The true and instantaneous reflection of aircraft pitch-and-bank attitude by the miniature aircraft, relative to the horizon bar of the attitude indicator.

Direct User Access Terminal System (DUATS). A system that provides current FAA weather and flight plan filing services to certified civil pilots, via personal computer, modem, or telephone access to the system. Pilots can request specific types of weather briefings and other pertinent data for planned flights.

Distance circle. See reference circle.

Distance measuring equipment (DME). A pulse-type electronic navigation system that shows the pilot, by an instrument-panel indication, the number of nautical miles between the aircraft and a ground station or waypoint.

DME. See distance measuring equipment.

DME arc. A flight track that is a constant distance from the station or waypoint.

DOD. Department of Defense.

Doghouse. A turn-and-slip indicator dial mark in the shape of a doghouse.

Domestic Reduced Vertical Separation Minimum (DRVSM). Additional flight levels between FL 290 and FL 410 to provide operational, traffic, and airspace efficiency.

Double gimbal. A type of mount used for the gyro in an attitude instrument. The axes of the two gimbals are at right angles to the spin axis of the gyro, allowing free motion in two planes around the gyro.

DP. See departure procedure.

Drag. The net aerodynamic force parallel to the relative wind, usually the sum of two components: induced drag and parasite drag.

Drag curve. The curve created when plotting induced drag and parasite drag.

DUATS. See direct user access terminal system.

Duplex. Transmitting on one frequency and receiving on a separate frequency.

Eddy currents. Current induced in a metal cup or disc when it is crossed by lines of flux from a moving magnet.

EFAS. See En Route Flight Advisory Service.

EFC. See expect-further-clearance.

Electronic flight display (EFD). For the purpose of standardization, any flight instrument display that uses LCD or other image-producing system (Cathode Ray Tube [CRT], etc.)

Elevator illusion. The sensation of being in a climb or descent, caused by the kind of abrupt vertical accelerations that result from up- or downdrafts.

Emergency. A distress or urgent condition.

Emphasis error. The result of giving too much attention to a particular instrument during the cross-check, instead of relying on a combination of instruments necessary for attitude and performance information.

EM wave. Electromagnetic wave.

Encoding altimeter. A special type of pressure altimeter used to send a signal to the air traffic controller on the ground, showing the pressure altitude the aircraft is flying.

En route facilities ring. Depicted in the plan view of IAP charts, a circle which designates NAVAIDs, fixes, and intersections that are part of the en route low altitude airway structure.

En Route Flight Advisory Service (EFAS). An en route weather-only AFSS service.

En route high-altitude charts. Aeronautical charts for en route instrument navigation at or above 18,000 feet MSL.

En route low-altitude charts. Aeronautical charts for en route IFR navigation below 18,000 feet MSL.

Equivalent airspeed. Airspeed equivalent to CAS in standard atmosphere at sea level. As the airspeed and pressure altitude increase, the CAS becomes higher than it should be, and a correction for compression must be subtracted from the CAS.

Expect-further-clearance (EFC). The time a pilot can expect to receive clearance beyond a clearance limit.

FAA. Federal Aviation Administration.

FAF. See final approach fix.

False horizon. Inaccurate visual information for aligning the aircraft, caused by various natural and geometric formations that disorient the pilot from the actual horizon.

Federal airways. Class E airspace areas that extend upward from 1,200 feet to, but not including, 18,000 feet MSL, unless otherwise specified.

Feeder facilities. Used by ATC to direct aircraft to intervening fixes between the en route structure and the initial approach fix.

Final approach. Part of an instrument approach procedure in which alignment and descent for landing are accomplished.

Final approach fix (FAF). The fix from which the IFR final approach to an airport is executed, and which identifies the beginning of the final approach segment. An FAF is designated on government charts by a Maltese cross symbol for nonprecision approaches, and a lightning bolt symbol for precision approaches.

Fixating. Staring at a single instrument, thereby interrupting the cross-check process.

FL. See flight level.

Flight configurations. Adjusting the aircraft control surfaces (including flaps and landing gear) in a manner that will achieve a specified attitude.

Flight director indicator (FDI). One of the major components of a flight director system, it provides steering commands that the pilot (or the autopilot, if coupled) follows.

Flight level (FL). A measure of altitude (in hundreds of feet) used by aircraft flying above 18,000 feet with the altimeter set at 29.92" Hg.

Flight management system (FMS). Provides pilot and crew with highly accurate and automatic long-range navigation capability, blending available inputs from long- and short-range sensors.

Flight path. The line, course, or track along which an aircraft is flying or is intended to be flown.

Flight patterns. Basic maneuvers, flown by reference to the instruments rather than outside visual cues, for the purpose of practicing basic attitude flying. The patterns simulate maneuvers encountered on instrument flights such as holding patterns, procedure turns, and approaches.

Flight strips. Paper strips containing instrument flight information, used by ATC when processing flight plans.

FMS. See flight management system.

Form drag. The drag created because of the shape of a component or the aircraft.

Fundamental skills. Pilot skills of instrument cross-check, instrument interpretation, and aircraft control.

Glide slope (GS). Part of the ILS that projects a radio beam upward at an angle of approximately 3° from the approach end of an instrument runway. The glide slope provides vertical guidance to aircraft on the final approach course for the aircraft to follow when making an ILS approach along the localizer path.

Glide slope intercept altitude. The minimum altitude of an intermediate approach segment prescribed for a precision approach that ensures obstacle clearance.

Global landing system (GLS). An instrument approach with lateral and vertical guidance with integrity limits (similar to barometric vertical navigation (BRO VNAV).

Global navigation satellite systems (GNSS). Satellite navigation systems that provide autonomous geo-spatial positioning with global coverage. It allows small electronic receivers to determine their location (longitude, latitude, and altitude) to within a few meters using time signals transmitted along a line of sight by radio from satellites.

GNSS. See global navigation satellite systems.

Global positioning system (GPS). Navigation system that uses satellite rather than ground-based transmitters for location information.

Goniometer. As used in radio frequency (RF) antenna systems, a direction-sensing device consisting of two fixed loops of wire oriented 90° from each other, which separately sense received signal strength and send those signals to two rotors (also oriented 90°) in the sealed direction-indicating instrument. The rotors are attached to the direction-indicating needle of the instrument and rotated by a small motor until minimum magnetic field is sensed near the rotors.

GPS. See global positioning system.

GPS Approach Overlay Program. An authorization for pilots to use GPS avionics under IFR for flying designated existing nonprecision instrument approach procedures, with the exception of LOC, LDA, and SDF procedures.

Graveyard spiral. The illusion of the cessation of a turn while still in a prolonged, coordinated, constant rate turn, which can lead a disoriented pilot to a loss of control of the aircraft.

Great circle route. The shortest distance across the surface of a sphere (the Earth) between two points on the surface.

Ground proximity warning system (GPWS). A system designed to determine an aircraft's clearance above the Earth and provides limited predictability about aircraft position relative to rising terrain.

Groundspeed. Speed over the ground, either closing speed to the station or waypoint, or speed over the ground in whatever direction the aircraft is going at the moment, depending upon the navigation system used.

GS. See glide slope.

GWPS. See ground proximity warning system.

HAA. See height above airport.

HAL. See height above landing.

HAT. See height above touchdown elevation.

Hazardous attitudes. Five aeronautical decision-making attitudes that may contribute to poor pilot judgment: antiauthority, impulsivity, invulnerability, machismo, and resignation.

Hazardous Inflight Weather Advisory Service (HIWAS). Service providing recorded weather forecasts broadcast to airborne pilots over selected VORs.

Head-up display (HUD). A special type of flight viewing screen that allows the pilot to watch the flight instruments and other data while looking through the windshield of the aircraft for other traffic, the approach lights, or the runway.

Height above airport (HAA). The height of the MDA above the published airport elevation.

Height above landing (HAL). The height above a designated helicopter landing area used for helicopter instrument approach procedures.

Height above touchdown elevation (HAT). The DA/DH or MDA above the highest runway elevation in the touchdown zone (first 3,000 feet of the runway).

HF. High frequency.

Hg. Abbreviation for mercury, from the Latin hydrargyrum.

HIWAS. See Hazardous Inflight Weather Advisory Service.

Holding. A predetermined maneuver that keeps aircraft within a specified airspace while awaiting further clearance from ATC.

Holding pattern. A racetrack pattern, involving two turns and two legs, used to keep an aircraft within a prescribed airspace with respect to a geographic fix. A standard pattern uses right turns; nonstandard patterns use left turns.

Homing. Flying the aircraft on any heading required to keep the needle pointing to the 0° relative bearing position.

Horizontal situation indicator (HSI). A flight navigation instrument that combines the heading indicator with a CDI, in order to provide the pilot with better situational awareness of location with respect to the courseline.

HSI. See horizontal situation indicator.

HUD. See head-up display.

Human factors. A multidisciplinary field encompassing the behavioral and social sciences, engineering, and physiology, to consider the variables that influence individual and crew performance for the purpose of optimizing human performance and reducing errors.

Hypoxia. A state of oxygen deficiency in the body sufficient to impair functions of the brain and other organs.

IAF. See initial approach fix.

IAP. See instrument approach procedures.

IAS. See indicated airspeed.

ICAO. See International Civil Aviation Organization.

Ident. Air Traffic Control request for a pilot to push the button on the transponder to identify return on the controller's scope.

IFR. See instrument flight rules.

ILS. See instrument landing system.

ILS categories. Categories of instrument approach procedures allowed at airports equipped with the following types of instrument landing systems:

> **ILS Category I:** Provides for approach to a height above touchdown of not less than 200 feet, and with runway visual range of not less than 1,800 feet.
>
> **ILS Category II:** Provides for approach to a height above touchdown of not less than 100 feet and with runway visual range of not less than 1,200 feet.
>
> **ILS Category IIIA:** Provides for approach without a decision height minimum and with runway visual range of not less than 700 feet.
>
> **ILS Category IIIB:** Provides for approach without a decision height minimum and with runway visual range of not less than 150 feet.
>
> **ILS Category IIIC:** Provides for approach without a decision height minimum and without runway visual range minimum.

IMC. See instrument meteorological conditions.

Indicated airspeed (IAS). Shown on the dial of the instrument airspeed indicator on an aircraft. Directly related to calibrated airspeed (CAS), IAS includes instrument errors and position error.

Indirect indication. A reflection of aircraft pitch-and-bank attitude by the instruments other than the attitude indicator.

Induced drag. Drag caused by the same factors that produce lift; its amount varies inversely with airspeed. As airspeed decreases, the angle of attack must increase, in turn increasing induced drag.

Induction icing. A type of ice in the induction system that reduces the amount of air available for combustion. The most commonly found induction icing is carburetor icing.

Inertial navigation system (INS). A computer-based navigation system that tracks the movement of an aircraft via signals produced by onboard accelerometers. The initial location of the aircraft is entered into the computer, and all subsequent movement of the aircraft is sensed and used to keep the position updated. An INS does not require any inputs from outside signals.

Initial approach fix (IAF). The fix depicted on IAP charts where the instrument approach procedure (IAP) begins unless otherwise authorized by ATC.

Inoperative components. Higher minimums are prescribed when the specified visual aids are not functioning; this information is listed in the Inoperative Components Table found in the United States Terminal Procedures Publications.

INS. See inertial navigation system.

Instantaneous vertical speed indicator (IVSI). Assists in interpretation by instantaneously indicating the rate of climb or descent at a given moment with little or no lag as displayed in a vertical speed indicator (VSI).

Instrument approach procedures (IAP). A series of predetermined maneuvers for the orderly transfer of an aircraft under IFR from the beginning of the initial approach to a landing or to a point from which a landing may be made visually.

Instrument flight rules (IFR). Rules and regulations established by the Federal Aviation Administration to govern flight under conditions in which flight by outside visual reference is not safe. IFR flight depends upon flying by reference to instruments in the flight deck, and navigation is accomplished by reference to electronic signals.

Instrument landing system (ILS). An electronic system that provides both horizontal and vertical guidance to a specific runway, used to execute a precision instrument approach procedure.

Instrument meteorological conditions (IMC). Meteorological conditions expressed in terms of visibility, distance from clouds, and ceiling less than the minimums specified for visual meteorological conditions, requiring operations to be conducted under IFR.

Instrument takeoff. Using the instruments rather than outside visual cues to maintain runway heading and execute a safe takeoff.

Interference drag. Drag generated by the collision of airstreams creating eddy currents, turbulence, or restrictions to smooth flow.

International Civil Aviation Organization (ICAO). The United Nations agency for developing the principles and techniques of international air navigation, and fostering planning and development of international civil air transport.

International standard atmosphere (IAS). A model of standard variation of pressure and temperature.

Inversion illusion. The feeling that the aircraft is tumbling backwards, caused by an abrupt change from climb to straight-and-level flight while in situations lacking visual reference.

Inverter. A solid-state electronic device that converts D.C. into A.C. current of the proper voltage and frequency to operate A.C. gyro instruments.

Isogonic lines. Lines drawn across aeronautical charts to connect points having the same magnetic variation.

IVSI. See instantaneous vertical speed indicator.

Jet route. A route designated to serve flight operations from 18,000 feet MSL up to and including FL 450.

Jet stream. A high-velocity narrow stream of winds, usually found near the upper limit of the troposphere, which flows generally from west to east.

KIAS. Knots indicated airspeed.

Kollsman window. A barometric scale window of a sensitive altimeter used to adjust the altitude for the altimeter setting.

LAAS. See local area augmentation system.

Lag. The delay that occurs before an instrument needle attains a stable indication.

Land as soon as possible. ATC instruction to pilot. Land without delay at the nearest suitable area, such as an open field, at which a safe approach and landing is assured.

Land as soon as practical. ATC instruction to pilot. The landing site and duration of flight are at the discretion of the pilot. Extended flight beyond the nearest approved landing area is not recommended.

Land immediately. ATC instruction to pilot. The urgency of the landing is paramount. The primary consideration is to ensure the survival of the occupants. Landing in trees, water, or other unsafe areas should be considered only as a last resort.

LDA. See localizer-type directional aid.

Lead radial. The radial at which the turn from the DME arc to the inbound course is started.

Leans, the. A physical sensation caused by an abrupt correction of a banked attitude entered too slowly to stimulate the motion sensing system in the inner ear. The abrupt correction can create the illusion of banking in the opposite direction.

Lift. A component of the total aerodynamic force on an airfoil and acts perpendicular to the relative wind.

Lines of flux. Invisible lines of magnetic force passing between the poles of a magnet.

L/MF. See low or medium frequency.

LMM. See locator middle marker.

Load factor. The ratio of a specified load to the total weight of the aircraft. The specified load is expressed in terms of any of the following: aerodynamic forces, inertial forces, or ground or water reactions.

Loadmeter. A type of ammeter installed between the generator output and the main bus in an aircraft electrical system.

LOC. See localizer.

Local area augmentation system (LAAS). A differential global positioning system (DGPS) that improves the accuracy of the system by determining position error from the GPS satellites, then transmitting the error, or corrective factors, to the airborne GPS receiver.

Localizer (LOC). The portion of an ILS that gives left/right guidance information down the centerline of the instrument runway for final approach.

Localizer-type directional aid (LDA). A NAVAID used for nonprecision instrument approaches with utility and accuracy comparable to a localizer but which is not a part of a complete ILS and is not aligned with the runway. Some LDAs are equipped with a glide slope.

Locator middle marker (LMM). Nondirectional radio beacon (NDB) compass locator, collocated with a middle marker (MM).

Locator outer marker (LOM). NDB compass locator, collocated with an outer marker (OM).

LOM. See locator outer marker.

Long range navigation (LORAN). An electronic navigational system by which hyperbolic lines of position are determined by measuring the difference in the time of reception of synchronized pulse signals from two fixed transmitters. LORAN A operates in the 1750 to 1950 kHz frequency band. LORAN C and D operate in the 100 to 110 kHz frequency band.

LORAN. See long range navigation.

Low or medium frequency. A frequency range between 190–535 kHz with the medium frequency above 300 kHz. Generally associated with nondirectional beacons transmitting a continuous carrier with either a 400 or 1,020 Hz modulation.

Lubber line. The reference line used in a magnetic compass or heading indicator.

MAA. See maximum authorized altitude.

Mach number. The ratio of the true airspeed of the aircraft to the speed of sound in the same atmospheric conditions, named in honor of Ernst Mach, late 19th century physicist.

Mach meter. The instrument that displays the ratio of the speed of sound to the true airspeed an aircraft is flying.

Magnetic bearing (MB). The direction to or from a radio transmitting station measured relative to magnetic north.

Magnetic heading (MH). The direction an aircraft is pointed with respect to magnetic north.

Mandatory altitude. An altitude depicted on an instrument approach chart with the altitude value both underscored and overscored. Aircraft are required to maintain altitude at the depicted value.

Mandatory block altitude. An altitude depicted on an instrument approach chart with two underscored and overscored altitude values between which aircraft are required to maintain altitude.

MAP. See missed approach point.

Margin identification. The top and bottom areas on an instrument approach chart that depict information about the procedure, including airport location and procedure identification.

Marker beacon. A low-powered transmitter that directs its signal upward in a small, fan-shaped pattern. Used along the flight path when approaching an airport for landing, marker beacons indicate both aurally and visually when the aircraft is directly over the facility.

Maximum altitude. An altitude depicted on an instrument approach chart with overscored altitude value at which or below aircraft are required to maintain altitude.

Maximum authorized altitude (MAA). A published altitude representing the maximum usable altitude or flight level for an airspace structure or route segment.

MB. See magnetic bearing.

MCA. See minimum crossing altitude.

MDA. See minimum descent altitude.

MEA. See minimum en route altitude.

Mean sea level. The average height of the surface of the sea at a particular location for all stages of the tide over a 19-year period.

MFD. See multi-function display.

MH. See magnetic heading.

MHz. Megahertz.

Microwave landing system (MLS). A precision instrument approach system operating in the microwave spectrum which normally consists of an azimuth station, elevation station, and precision distance measuring equipment.

Mileage breakdown. A fix indicating a course change that appears on the chart as an "x" at a break between two segments of a federal airway.

Military operations area (MOA). Airspace established for the purpose of separating certain military training activities from IFR traffic.

Military training route (MTR). Airspace of defined vertical and lateral dimensions established for the conduct of military training at airspeeds in excess of 250 knots indicated airspeed (KIAS).

Minimum altitude. An altitude depicted on an instrument approach chart with the altitude value underscored. Aircraft are required to maintain altitude at or above the depicted value.

Minimum crossing altitude (MCA). The lowest allowed altitude at certain fixes an aircraft must cross when proceeding in the direction of a higher minimum en route altitude (MEA).

Minimum descent altitude (MDA). The lowest altitude (in feet MSL) to which descent is authorized on final approach, or during circle-to-land maneuvering in execution of a nonprecision approach.

Minimum en route altitude (MEA). The lowest published altitude between radio fixes that ensures acceptable navigational signal coverage and meets obstacle clearance requirements between those fixes.

Minimum obstruction clearance altitude (MOCA). The lowest published altitude in effect between radio fixes on VOR airways, off-airway routes, or route segments, which meets obstacle clearance requirements for the entire route segment and which ensures acceptable navigational signal coverage only within 25 statute (22 nautical) miles of a VOR.

Minimum reception altitude (MRA). The lowest altitude at which an airway intersection can be determined.

Minimum safe altitude (MSA). The minimum altitude depicted on approach charts which provides at least 1,000 feet of obstacle clearance for emergency use within a specified distance from the listed navigation facility.

Minimum vectoring altitude (MVA). An IFR altitude lower than the minimum en route altitude (MEA) that provides terrain and obstacle clearance.

Minimums section. The area on an IAP chart that displays the lowest altitude and visibility requirements for the approach.

Missed approach. A maneuver conducted by a pilot when an instrument approach cannot be completed to a landing.

Missed approach point (MAP). A point prescribed in each instrument approach at which a missed approach procedure shall be executed if the required visual reference has not been established.

Mixed ice. A mixture of clear ice and rime ice.

MLS. See microwave landing system.

MM. Middle marker.

MOA. See military operations area.

MOCA. See minimum obstruction clearance altitude.

Mode C. Altitude reporting transponder mode.

MRA. See minimum reception altitude.

MSA. See minimum safe altitude.

MSL. See mean sea level.

MTR. See military training route.

Multi-function display (MFD). Small screen (CRT or LCD) in an aircraft that can be used to display information to the pilot in numerous configurable ways. Often an MFD will be used in concert with a Primary Flight Display.

MVA. See minimum vectoring altitude.

NACG. See National Aeronautical Charting Group.

NAS. See National Airspace System.

National Airspace System (NAS). The common network of United States airspace—air navigation facilities, equipment and services, airports or landing areas; aeronautical charts, information and services; rules, regulations and procedures, technical information; and manpower and material.

National Aeronautical Charting Group (NACG). A Federal agency operating under the FAA, responsible for publishing charts such as the terminal procedures and en route charts.

National Route Program (NRP). A set of rules and procedures designed to increase the flexibility of user flight planning within published guidelines.

National Security Area (NSA). Areas consisting of airspace of defined vertical and lateral dimensions established at locations where there is a requirement for increased security and safety of ground facilities. Pilots are requested to voluntarily avoid flying through the depicted NSA. When it is necessary to provide a greater level of security and safety, flight in NSAs may be temporarily prohibited. Regulatory prohibitions are disseminated via NOTAMs.

National Transportation Safety Board (NTSB). A United States Government independent organization responsible for investigations of accidents involving aviation, highways, waterways, pipelines, and railroads in the United States. NTSB is charged by congress to investigate every civil aviation accident in the United States.

NAVAID. Naviagtional aid.

NAV/COM. Navigation and communication radio.

NDB. See nondirectional radio beacon.

NM. Nautical mile.

NOAA. National Oceanic and Atmospheric Administration.

No-gyro approach. A radar approach that may be used in case of a malfunctioning gyro-compass or directional gyro. Instead of providing the pilot with headings to be flown, the controller observes the radar track and issues control instructions "turn right/left" or "stop turn," as appropriate.

Nondirectional radio beacon (NDB). A ground-based radio transmitter that transmits radio energy in all directions.

Nonprecision approach. A standard instrument approach procedure in which only horizontal guidance is provided.

No procedure turn (NoPT). Term used with the appropriate course and altitude to denote that the procedure turn is not required.

NoPT. See no procedure turn.

Notice to Airmen (NOTAM). A notice filed with an aviation authority to alert aircraft pilots of any hazards en route or at a specific location. The authority in turn provides means of disseminating relevant NOTAMs to pilots.

NRP. See National Route Program.

NSA. See National Security Area.

NTSB. See National Transportation Safety Board.

NWS. National Weather Service.

Obstacle departure procedures (ODP). Obstacle clearance protection provided to aircraft in instrument meteorological conditions (IMC).

ODP. See obstacle departure procedures.

OM. Outer marker.

Omission error. The failure to anticipate significant instrument indications following attitude changes; for example, concentrating on pitch control while forgetting about heading or roll information, resulting in erratic control of heading and bank.

Optical illusion. A misleading visual image. For the purpose of this handbook, the term refers to the brain's misinterpretation of features on the ground associated with landing, which causes a pilot to misread the spatial relationships between the aircraft and the runway.

Orientation. Awareness of the position of the aircraft and of oneself in relation to a specific reference point.

Otolith organ. An inner ear organ that detects linear acceleration and gravity orientation.

Outer marker. A marker beacon at or near the glide slope intercept altitude of an ILS approach. It is normally located four to seven miles from the runway threshold on the extended centerline of the runway.

Overcontrolling. Using more movement in the control column than is necessary to achieve the desired pitch-and bank condition.

Overpower. To use more power than required for the purpose of achieving a faster rate of airspeed change.

P-static. See precipitation static.

PAPI. See precision approach path indicator.

PAR. See precision approach radar.

Parasite drag. Drag caused by the friction of air moving over the aircraft structure; its amount varies directly with the airspeed.

PFD. See primary flight display.

PIC. See pilot-in-command.

Pilot-in-command (PIC). The pilot responsible for the operation and safety of an aircraft.

Pilot report (PIREP). Report of meteorological phenomena encountered by aircraft.

Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM). FAA-approved documents published by the airframe manufacturer that list the operating conditions for a particular model of aircraft.

PIREP. See pilot report.

Pitot pressure. Ram air pressure used to measure airspeed.

Pitot-static head. A combination pickup used to sample pitot pressure and static air pressure.

Plan view. The overhead view of an approach procedure on an instrument approach chart. The plan view depicts the routes that guide the pilot from the en route segments to the IAF.

POH/AFM. See Pilot's Operating Handbook/Airplane Flight Manual.

Point-in-space approach. A type of helicopter instrument approach procedure to a missed approach point more than 2,600 feet from an associated helicopter landing area.

Position error. Error in the indication of the altimeter, ASI, and VSI caused by the air at the static system entrance not being absolutely still.

Position report. A report over a known location as transmitted by an aircraft to ATC.

Precession. The characteristic of a gyroscope that causes an applied force to be felt, not at the point of application, but 90° from that point in the direction of rotation.

Precipitation static (P-static). A form of radio interference caused by rain, snow, or dust particles hitting the antenna and inducing a small radio-frequency voltage into it.

Precision approach. A standard instrument approach procedure in which both vertical and horizontal guidance is provided.

Precision approach path indicator (PAPI). A system of lights similar to the VASI, but consisting of one row of lights in two- or four-light systems. A pilot on the correct glide slope will see two white lights and two red lights. See VASI.

Precision approach radar (PAR). A type of radar used at an airport to guide an aircraft through the final stages of landing, providing horizontal and vertical guidance. The radar operator directs the pilot to change heading or adjust the descent rate to keep the aircraft on a path that allows it to touch down at the correct spot on the runway.

Precision runway monitor (PRM). System allows simultaneous, independent Instrument Flight Rules (IFR) approaches at airports with closely spaced parallel runways.

Preferred IFR routes. Routes established in the major terminal and en route environments to increase system efficiency and capacity. IFR clearances are issued based on these routes, listed in the A/FD except when severe weather avoidance procedures or other factors dictate otherwise.

Pressure altitude. Altitude above the standard 29.92" Hg plane.

Prevailing visibility. The greatest horizontal visibility equaled or exceeded throughout at least half the horizon circle (which is not necessarily continuous).

Primary and supporting. A method of attitude instrument flying using the instrument that provides the most direct indication of attitude and performance.

Primary flight display (PFD). A display that provides increased situational awareness to the pilot by replacing the traditional six instruments used for instrument flight with an easy-to-scan display that provides the horizon, airspeed, altitude, vertical speed, trend, trim, rate of turn among other key relevant indications.

PRM. See precision runway monitor.

Procedure turn. A maneuver prescribed when it is necessary to reverse direction to establish an aircraft on the intermediate approach segment or final approach course.

Profile view. Side view of an IAP chart illustrating the vertical approach path altitudes, headings, distances, and fixes.

Prohibited area. Designated airspace within which flight of aircraft is prohibited.

Propeller/rotor modulation error. Certain propeller RPM settings or helicopter rotor speeds can cause the VOR course deviation indicator (CDI) to fluctuate as much as ±6°. Slight changes to the RPM setting will normally smooth out this roughness.

Rabbit, the. High-intensity flasher system installed at many large airports. The flashers consist of a series of brilliant blue-white bursts of light flashing in sequence along the approach lights, giving the effect of a ball of light traveling towards the runway.

Radar. Radio Detection And Ranging.

Radar approach. The controller provides vectors while monitoring the progress of the flight with radar, guiding the pilot through the descent to the airport/heliport or to a specific runway.

Radials. The courses oriented from a station.

Radio or radar altimeter. An electronic altimeter that determines the height of an aircraft above the terrain by measuring the time needed for a pulse of radio-frequency energy to travel from the aircraft to the ground and return.

Radio frequency (RF). A term that refers to alternating current (AC) having characteristics such that, if the current is input to antenna, an electromagnetic (EM) field is generated suitable for wireless broadcasting and/or communications.

Radio magnetic indicator (RMI). An electronic navigation instrument that combines a magnetic compass card with two bearing pointers (typically). Generally, one pointer is for the ADF while the other is for an RNAV or VOR navigation system. The pointers are typically different colors and of different widths for ease of identification. Sometimes a function switch is provided to allow the #2 pointer to be slaved to either a VOR or RNAV system. The card of the RMI acts as a gyro-stabilized magnetic compass (usually corrected for north via a flux valve) and shows the magnetic heading the aircraft is flying.

Radio wave. An electromagnetic wave (EM wave) with frequency characteristics useful for radio transmission.

RAIM. See receiver autonomous integrity monitoring.

Random RNAV routes. Direct routes, based on area navigation capability, between waypoints defined in terms of latitude/longitude coordinates, degree-distance fixes, or offsets from established routes/airways at a specified distance and direction.

Ranging signals. Transmitted from the GPS satellite, these allow the aircraft's receiver to determine range (distance) from each satellite.

RB. See relative bearing.
RBI. See relative bearing indicator.

RCO. See remote communications outlet.

Receiver autonomous integrity monitoring (RAIM). A system used to verify the usability of the received GPS signals and warns the pilot of any malfunction in the navigation system. This system is required for IFR-certified GPS units.

Recommended altitude. An altitude depicted on an instrument approach chart with the altitude value neither underscored nor overscored. The depicted value is an advisory value.

Receiver-transmitter (RT). A system that permits selection of a unique channel or frequency whereupon a signal (typically communication) can be transmitted and received.

Reduced vertical separation minimum (RVSM). Reduces the vertical separation between flight level (FL) 290–410 from 2,000 feet to 1,000 feet and makes six additional FLs available for operation. Also see DRVSM.

Reference circle (also, distance circle). The circle depicted in the plan view of an IAP chart that typically has a 10 NM radius, within which chart the elements are drawn to scale.

Regions of command. The "regions of normal and reversed command" refers to the relationship between speed and the power required to maintain or change that speed in flight.

REIL. See runway end identifier lights.

Relative bearing (RB). The angular difference between the aircraft heading and the direction to the station, measured clockwise from the nose of the aircraft.

Relative bearing indicator (RBI). Also known as the fixed-card ADF, zero is always indicated at the top of the instrument and the needle indicates the relative bearing to the station.

Relative wind. Direction of the airflow produced by an object moving through the air. The relative wind for an airplane in flight flows in a direction parallel with and opposite to the direction of flight; therefore, the actual flight path of the airplane determines the direction of the relative wind.

Remote communications outlet (RCO). An unmanned communications facility that is remotely controlled by air traffic personnel.

Required navigation performance (RNP). A specified level of accuracy defined by a lateral area of confined airspace in which an RNP-certified aircraft operates.

Restricted area. Airspace designated under 14 CFR part 73 within which the flight of aircraft, while not wholly prohibited, is subject to restriction.

Reverse sensing. The VOR needle appearing to indicate the reverse of normal operation.

RF. Radio frequency.

Rhodopsin. The photosensitive pigments that initiate the visual response in the rods of the eye.

Rigidity. The characteristic of a gyroscope that prevents its axis of rotation tilting as the Earth rotates.

Rime ice. Rough, milky, opaque ice formed by the instantaneous freezing of small supercooled water droplets.

Risk. The future impact of a hazard that is not eliminated or controlled.

RMI. See radio magnetic indicator.

RNAV. See area navigation.

RNP. See required navigation performance.

Runway end identifier lights (REIL). A pair of synchronized flashing lights, located laterally on each side of the runway threshold, providing rapid and positive identification of the approach end of a runway.

Runway visibility value (RVV). The visibility determined for a particular runway by a transmissometer.

Runway visual range (RVR). The instrumentally derived horizontal distance a pilot should be able to see down the runway from the approach end, based on either the sighting of high-intensity runway lights, or the visual contrast of other objects.

RVR. See runway visual range.

RVV. See runway visibility value.

SA. See selective availability.

St. Elmo's Fire. A corona discharge which lights up the aircraft surface areas where maximum static discharge occurs.

Satellite ephemeris data. Data broadcast by the GPS satellite containing very accurate orbital data for that satellite, atmospheric propagation data, and satellite clock error data.

Scan. The first fundamental skill of instrument flight, also known as "cross-check;" the continuous and logical observation of instruments for attitude and performance information.

SDF. See simplified directional facility.

Selective availability (SA). A satellite technology permitting the Department of Defense (DOD) to create, in the interest of national security, a significant clock and ephemeris error in the satellites, resulting in a navigation error.

Semicircular canal. An inner ear organ that detects angular acceleration of the body.

Sensitive altimeter. A form of multipointer pneumatic altimeter with an adjustable barometric scale that allows the reference pressure to be set to any desired level.

SIDS. See standard instrument departure procedures.

SIGMET. The acronym for Significant Meteorological information. A weather advisory issued concerning weather significant to the safety of all aircraft.

Signal-to-noise ratio. An indication of signal strength received compared to background noise, which is a measure of how adequate the received signal is.

Simplex. Transmission and reception on the same frequency.

Simplified directional facility (SDF). A NAVAID used for nonprecision instrument approaches. The final approach course is similar to that of an ILS localizer; however, the SDF course may be offset from the runway, generally not more than 3°, and the course may be wider than the localizer, resulting in a lower degree of accuracy.

Single-pilot resource management (SRM). The ability for crew or pilot to manage all resources effectively to ensure the outcome of the flight is successful.

Situational awareness. Pilot knowledge of where the aircraft is in regard to location, air traffic control, weather, regulations, aircraft status, and other factors that may affect flight.

Skidding turn. An uncoordinated turn in which the rate of turn is too great for the angle of bank, pulling the aircraft to the outside of the turn.

Skin friction drag. Drag generated between air molecules and the solid surface of the aircraft.

Slant range. The horizontal distance from the aircraft antenna to the ground station, due to line-of-sight transmission of the DME signal.

Slaved compass. A system whereby the heading gyro is "slaved to," or continuously corrected to bring its direction readings into agreement with a remotely located magnetic direction sensing device (usually this is a flux valve or flux gate compass).

Slipping turn. An uncoordinated turn in which the aircraft is banked too much for the rate of turn, so the horizontal lift component is greater than the centrifugal force, pulling the aircraft toward the inside of the turn.

Small airplane. An airplane of 12,500 pounds or less maximum certificated takeoff weight.

Somatogravic illusion. The misperception of being in a nose-up or nose-down attitude, caused by a rapid acceleration or deceleration while in flight situations that lack visual reference.

Spatial disorientation. The state of confusion due to misleading information being sent to the brain from various sensory organs, resulting in a lack of awareness of the aircraft position in relation to a specific reference point.

Special use airspace. Airspace in which flight activities are subject to restrictions that can create limitations on the mixed use of airspace. Consists of prohibited, restricted, warning, military operations, and alert areas.

SRM. See single-pilot resource management.

SSR. See secondary surveillance radar.

SSV. See standard service volume.

Standard holding pattern. A holding pattern in which all turns are made to the right.

Standard instrument departure procedures (SIDS). Published procedures to expedite clearance delivery and to facilitate transition between takeoff and en route operations.

Standard rate turn. A turn in which an aircraft changes its direction at a rate of 3° per second. The turn indicators are typically 2 minute or 4 minute instruments. In a 2 minute instrument, if the needle is one needle width either side of the center alignment mark, the turn is 3° per second and the turn takes 2 minutes to execute a 360° turn. In a 4 minute instrument, the same turn takes two widths deflection of the needle to achieve 3° per second. The 4 minute turn instrument is usually found on high performance aircraft.

Standard service volume (SSV). Defines the limits of the volume of airspace which the VOR serves.

Standard terminal arrival route (STAR). A preplanned IFR ATC arrival procedure published for pilot use in graphic and/or textual form.

STAR. See standard terminal arrival route.

Static longitudinal stability. The aerodynamic pitching moments required to return the aircraft to the equilibrium angle of attack.

Static pressure. Pressure of air that is still, or not moving, measured perpendicular to the surface of the aircraft.

Steep turns. In instrument flight, any turn greater than standard rate; in visual flight, anything greater than a 45° bank.

Stepdown fix. The point after which additional descent is permitted within a segment of an IAP.

Strapdown system. An INS in which the accelerometers and gyros are permanently "strapped down" or aligned with the three axes of the aircraft.

Stress. The body's response to demands placed upon it.

Structural icing. The accumulation of ice on the exterior of the aircraft.

Suction relief valve. A relief valve in an instrument vacuum system required to maintain the correct low pressure inside the instrument case for the proper operation of the gyros.

Synchro. A device used to transmit indications of angular movement or position from one location to another.

Synthetic vision. A realistic display depiction of the aircraft in relation to terrain and flight path.

TAA. See terminal arrival area.

TACAN. See tactical air navigation.

Tactical air navigation (TACAN). An electronic navigation system used by military aircraft, providing both distance and direction information.

TAWS. See terrain awareness and warning system.

TCAS. See traffic alert collision avoidance system.

TCH. See threshold crossing height.

TDZE. See touchdown zone elevation.

TEC. See Tower En Route Control.

Technique. The manner in which procedures are executed.

Temporary flight restriction (TFR). Restriction to flight imposed in order to:

1. Protect persons and property in the air or on the surface from an existing or imminent flight associated hazard;
2. Provide a safe environment for the operation of disaster relief aircraft;
3. Prevent an unsafe congestion of sightseeing aircraft above an incident;
4. Protect the President, Vice President, or other public figures; and,
5. Provide a safe environment for space agency operations.

Pilots are expected to check appropriate NOTAMs during flight planning when conducting flight in an area where a temporary flight restriction is in effect.

Tension. Maintaining an excessively strong grip on the control column, usually resulting in an overcontrolled situation.

Terminal Instrument Approach Procedure (TERP). Prescribes standardized methods for use in designing instrument flight procedures.

Terminal arrival area (TAA). A procedure to provide a new transition method for arriving aircraft equipped with FMS and/or GPS navigational equipment. The TAA contains a "T" structure that normally provides a NoPT for aircraft using the approach.

TERP. See terminal instrument approach procedure.

Terrain Awareness and Warning System (TAWS). A timed-based system that provides information concerning potential hazards with fixed objects by using GPS positioning and a database of terrain and obstructions to provide true predictability of the upcoming terrain and obstacles.

TFR. See temporary flight restriction.

Threshold crossing height (TCH). The theoretical height above the runway threshold at which the aircraft's glide slope antenna would be if the aircraft maintains the trajectory established by the mean ILS glide slope or MLS glide path.

Thrust (aerodynamic force). The forward aerodynamic force produced by a propeller, fan, or turbojet engine as it forces a mass of air to the rear, behind the aircraft.

Time and speed table. A table depicted on an instrument approach procedure chart that identifies the distance from the FAF to the MAP, and provides the time required to transit that distance based on various groundspeeds.

Timed turn. A turn in which the clock and the turn coordinator are used to change heading a definite number of degrees in a given time.

TIS. See traffic information service.

Title 14 of the Code of Federal Regulations (14 CFR). The federal aviation regulations governing the operation of aircraft, airways, and airmen.

Touchdown zone elevation (TDZE). The highest elevation in the first 3,000 feet of the landing surface, TDZE is indicated on the instrument approach procedure chart when straight-in landing minimums are authorized.

Tower En Route Control (TEC). The control of IFR en route traffic within delegated airspace between two or more adjacent approach control facilities, designed to expedite traffic and reduce control and pilot communication requirements.

TPP. See United States Terminal Procedures Publication.

Tracking. Flying a heading that will maintain the desired track to or from the station regardless of crosswind conditions.

Traffic Alert Collision Avoidance System (TCAS). An airborne system developed by the FAA that operates independently from the ground-based Air Traffic Control system. Designed to increase flight deck awareness of proximate aircraft and to serve as a "last line of defense" for the prevention of mid-air collisions.

Traffic information service (TIS). A ground-based service providing information to the flight deck via data link using the S-mode transponder and altitude encoder to improve the safety and efficiency of "see and avoid" flight through an automatic display that informs the pilot of nearby traffic.

Transcribed Weather Broadcast (TWEB). Meteorological and aeronautical data recorded on tapes and broadcast over selected NAVAIDs. Generally, the broadcast contains route-oriented data with specially prepared NWS forecasts, inflight advisories, and winds aloft. It also includes selected current information such as weather reports (METAR/SPECI), NOTAMs, and special notices.

Transponder. The airborne portion of the ATC radar beacon system.

Transponder code. One of 4,096 four-digit discrete codes ATC assigns to distinguish between aircraft.

Trend. Immediate indication of the direction of aircraft movement, as shown on instruments.

Trim. Adjusting the aerodynamic forces on the control surfaces so that the aircraft maintains the set attitude without any control input.

TWEB. See Transcribed Weather Broadcast.

True airspeed. Actual airspeed, determined by applying a correction for pressure altitude and temperature to the CAS.

UHF. See ultra-high frequency.

Ultra-high frequency (UHF). The range of electromagnetic frequencies between 962 MHz and 1213 MHz.

Uncaging. Unlocking the gimbals of a gyroscopic instrument, making it susceptible to damage by abrupt flight maneuvers or rough handling.

Underpower. Using less power than required for the purpose of achieving a faster rate of airspeed change.

United States Terminal Procedures Publication (TPP). Booklets published in regional format by the NACO that include DPs, STARs, IAPs, and other information pertinent to IFR flight.

Unusual attitude. An unintentional, unanticipated, or extreme aircraft attitude.

User-defined waypoints. Waypoint location and other data which may be input by the user, this is the only GPS database information that may be altered (edited) by the user.

Variation. Compass error caused by the difference in the physical locations of the magnetic north pole and the geographic north pole.

VASI. See visual approach slope indicator.

VDP. See visual descent point.

Vectoring. Navigational guidance by assigning headings.

Venturi tube. A specially shaped tube attached to the outside of an aircraft to produce suction to allow proper operation of gyro instruments.

Vertical speed indicator (VSI). A rate-of-pressure change instrument that gives an indication of any deviation from a constant pressure level.

Very-high frequency (VHF). A band of radio frequencies falling between 30 and 300 MHz.

Very-high frequency omnidirectional range (VOR). Electronic navigation equipment in which the flight deck instrument identifies the radial or line from the VOR station, measured in degrees clockwise from magnetic north, along which the aircraft is located.

Vestibule. The central cavity of the bony labyrinth of the ear, or the parts of the membranous labyrinth that it contains.

VFR. See visual flight rules.

VFR-on-top. ATC authorization for an IFR aircraft to operate in VFR conditions at any appropriate VFR altitude.

VFR over-the-top. A VFR operation in which an aircraft operates in VFR conditions on top of an undercast.

Victor airways. Airways based on a centerline that extends from one VOR or VORTAC navigation aid or intersection, to another navigation aid (or through several navigation aids or intersections); used to establish a known route for en route procedures between terminal areas.

Visual approach slope indicator (VASI). A visual aid of lights arranged to provide descent guidance information during the approach to the runway. A pilot on the correct glide slope will see red lights over white lights.

Visual descent point (VDP). A defined point on the final approach course of a nonprecision straight-in approach procedure from which normal descent from the MDA to the runway touchdown point may be commenced, provided the runway environment is clearly visible to the pilot.

Visual flight rules (VFR). Flight rules adopted by the FAA governing aircraft flight using visual references. VFR operations specify the amount of ceiling and the visibility the pilot must have in order to operate according to these rules. When the weather conditions are such that the pilot can not operate according to VFR, he or she must use instrument flight rules (IFR).

Visual meteorological conditions (VMC). Meteorological conditions expressed in terms of visibility, distance from cloud, and ceiling meeting or exceeding the minimums specified for VFR.

VMC. See visual meteorological conditions.

VOR. See very-high frequency omnidirectional range.

VORTAC. A facility consisting of two components, VOR and TACAN, which provides three individual services: VOR azimuth, TACAN azimuth, and TACAN distance (DME) at one site.

VOR test facility (VOT). A ground facility which emits a test signal to check VOR receiver accuracy. Some VOTs are available to the user while airborne, while others are limited to ground use only.

VOT. See VOR test facility.

VSI. See vertical speed indicator.

WAAS. See wide area augmentation system.

Warning area. An area containing hazards to any aircraft not participating in the activities being conducted in the area. Warning areas may contain intensive military training, gunnery exercises, or special weapons testing.

Waypoint. A designated geographical location used for route definition or progress-reporting purposes and is defined in terms of latitude/longitude coordinates.

WCA. See wind correction angle.

Weather and radar processor (WARP). A device that provides real-time, accurate, predictive and strategic weather information presented in an integrated manner in the National Airspace System (NAS).

Weight. The force exerted by an aircraft from the pull of gravity.

Wide area augmentation system (WAAS). A differential global positioning system (DGPS) that improves the accuracy of the system by determining position error from the GPS satellites, then transmitting the error, or corrective factors, to the airborne GPS receiver.

Wind correction angle (WCA). The angle between the desired track and the heading of the aircraft necessary to keep the aircraft tracking over the desired track.

Work. A measurement of force used to produce movement.

Zone of confusion. Volume of space above the station where a lack of adequate navigation signal directly above the VOR station causes the needle to deviate.

Index

A

above ground level 3-31, 7-44, 8-2, 9-10
absolute accuracy 7-25
acceleration in cruise flight 2-10
acute fatigue 1-11, 1-12, 1-13
additional reports 10-7
adjust 4-20
advanced technologies 7-26
advanced technology systems 3-28
adverse yaw.. 2-11, 2-12
aeronautical decision-making (ADM)1-1, 1-12, 1-15, 1-17
aeronautical information manual (AIM)............. 9-4, 10-2
agonic line 3-12
air data computer (ADC) 3-22
air route surveillance radar (ARSR) 7-49, 9-7
air route traffic control center (ARTC) 7-50, 9-4, 9-7, 10-2
air traffic control (ATC)........................... 1-15, 9-1, 11-1
  inflight weather avoidance assistance 9-11
  radar weather displays 9-11
air traffic control radar beacon system (ATCRBS)......7-49
air traffic control towers..9-5
aircraft approach categories 8-23
aircraft control...6-3
aircraft system malfunctions 11-3
airplane trim 4-8
airport diagram 8-27
airport information 8-6
airport sketch...8-27
airport surface detection equipment (ASDE)..... 7-49, 7-50
airport surveillance radar (ASR)......................... 7-49, 9-7
Airport/Facility Directory (A/FD)........1-9, 7-10, 8-6, 10-2
airspace classification 8-1
  class A through G 8-2
airspeed color codes 3-10
airspeed
  indicated 3-9
  indicator..4-6, 5-5, 5-37, 6-5
  calibrated 3-9
  equivalent 3-9
  true...3-9
airspeed changes
  common errors..6-10
airspeed indicators 4-26, 5-29, 5-61
  maximum allowable airspeed...................................3-10
alcohol..1-12
alternate airport 8-27
alternator/generator failure..11-5
altimeter 5-36, 6-4
amendment status...8-12
analog pictorial displays 3-22
anti-ice...2-12
approach lighting systems (ALS 7-40
approach to stall 5-26
altimeter
  errors..3-4
  cold weather 3-5
  enhancements (encoding)..3-7
analog instrument failure 11-6
angle of attack 2-2, 2-6
approach azimuth guidance..7-45
approach control advances 9-12
approach control facility 9-12
approach to airport
  without an operating control tower 10-14
  with control tower, no approach control 10-14
  with control tower and approach control................10-14
approaches...10-12
  missed 10-21
  parallel runways 10-20
  radar...10-17
  timed, from a holding fix.......................................10-18
area navigation (RNAV) 7-19
arrival 10-33
atmosphere 2-4
  layers of the atmosphere..2-5
attitude and heading reference system (AHRS)...........3-22
attitude director indicator (ADI) 3-23

attitude
  indicator....... 4-4, 4-5, 4-7, 5-2, 5-6, 5-34, 5-37, 6-3, 6-5
  control.........4-3
  instrument flying 4-1, 4-21, 6-1
autokinesis.........1-7
automated flight service stations (AFSS) 9-4
automated radar terminal systems (ARTS) 9-7
automated surface observing station (ASOS).........8-10
automated terminal information service (ATIS).........10-8
automated weather observing station (AWOS).........8-10
automatic dependent surveillance-broadcast (ADS-B)3-28
automatic direction finder (ADF) 3-16, 7-3, 10-7
  function of 7-4
  operational errors.........7-8
automatic terminal information service (ATIS). 1-15, 8-16
automatic weather observing system (AWOS).........7-3
autopilot systems.........3-24
autorotations.........6-17
    common errors.........6-17
azimuth card.........7-4

B

back courses (BC).........7-39
bank control 4-4, 4-7, 4-20, 5-6, 5-37, 6-5
baro-aiding.........7-28
barometric vertical navigation (BARO VNAV) 8-32
basic aerodynamics (review of) 2-2
  relative wind.........2-2
  angle of attack 2-2
basic instrument flight patterns......... 5-30, 5-61
basic radio principles 7-2
blockage considerations.........3-2
  indications of pitot tube blockage 3-3
  indications from static port blockage 3-3
  effects of flight conditions.........3-3

C

calibrated......... 5-2, 6-14
calibrated orifice 3-8
center approach/departure control.........9-7
certified checkpoints 7-16
changeover points (COPs) 8-10
changing technology 6-18
charted IFR altitudes.........8-6
chronic fatigue 1-13
circling approaches 10-20
circling approach pattern.........5-32
class D airspace......... 8-2, 9-6
clean configuration.........5-11
clear ice 2-13, 10-24
clearances.........10-3
  separations.........10-4
clearance delivery 10-4
clearance on request.........9-6
clearance void time 9-5
climbing
  while accelerating.........1-8
  while turning 1-8
climbs......... 2-10, 5-14
common errors
  fixation 4-27
  omission 4-28
  emphasis.........4-28
communication equipment.........9-2
communication facilities.........9-4
communication procedures.........9-4
communication/navigation system malfunction 11-8
compass
  course 3-13
  locator......... 7-28, 7-40
  turns......... 5-21, 5-53, 6-15
compass rose 3-12, 5-25
computer navigation fix 8-10
concentric rings.........8-18
conducting an IFR flight.........10-27
constant airspeed climb
  from cruise airspeed 5-46
  from established airspeed.........5-47
constant rate climbs.........5-47
control
  characteristics.........2-7
  and performance.........4-2
  instruments 4-2, 4-18
  sequence 9-13
control display unit (CDU) 3-26
control pressures 5-3
coordinated......... 5-6, 6-5
coordination of rudder and aileron controls.........2-11
coriolis illusion.........1-6
course interception.........7-14
course reversal elements
  plan view 8-20
  profile view 8-20
crew resource management (CRM) 1-14
critical areas 7-2
cross-check......... 4-3, 4-20
  common errors.........4-11
cruise clearance.........10-4
current induction 3-15

D

dark adaptation 1-3
DECIDE model 1-17
decision height (DH) 3-31, 7-50, 8-21
deice 2-13
density altitude 2-5
departure 10-31
departure procedures (DPs) 7-1, 7-33, 8-12, 10-5
  instrument 7-33
departures
  airports without an operating control tower 10-7
  radar controlled 10-5
descents 5-16, 5-49
deviation 3-12
differential global positioning systems (DGPS) 7-34
direct indication 5-2, 5-34, 6-3
directional 7-42
distance circle 8-18
distance measuring equipment (DME) 7-16, 8-7
  arc 7-17
  components 7-17
  errors 7-19
  function of 7-17
diving
  or rolling beyond the vertical plane 1-8
  while turning 1-8
DOD 3-22
doghouse 3-21
domestic reduced vertical
    separation minimum (DRVSM) 3-7
double gimbal 3-18
drag 2-3
drag curves 2-6
dry air vacuum pump 3-17
duplex 9-2
dynamic pressure type instruments 3-8

E

ears 1-4
  otolith organs 1-4, 1-5
  semicircular canals 1-4
eddy currents 2-3, 3-14
electrical systems 3-18
electronic flight display (EFD) 4-1
electronic flight instrument systems 3-27
elevator illusion 1-6
emergencies 6-16
en route 10-7, 10-32
en route flight advisory service (EFAS) 9-4
en route high-altitude charts 8-6
en route procedures 10-7
encoding altimeter 3-7
entry 5-14, 5-46, 5-49, 6-10
equipment 1-14
eyes 1-2

F

false horizon 1-7
fatigue 1-12
featureless terrain illusion 1-9
federal airways 8-4
feeder facilities 8-18
feet per minute (fpm) 4-6
filing in flight 10-2
final approach fix (FAF) 7-23
final approach waypoint (FAWP) 7-32
flight director indicator (FDI) 3-23
flight instruments 2-16, 3-1, 6-2
flight levels (FL) 3-7
flight management systems (FMS) 3-25
  function of 7-48
flight patterns 5-30
flight planning information, sources 10-2
flight strips 9-5
flight support systems 3-22
flight path 2-2, 2-6
four step process used to change attitude 4-20
flux gate compass 3-14
flying experience 10-22
fog 1-9, 10-24
form drag 2-4
four forces 2-2
fundamental skills
  of attitude instrument flying 4-24
  instrument cross-check 4-10
  instrument flight 6-2

G

glide slope 7-39
glide slope intercept altitude 10-20
global landing system (GLS) 8-32
global navigation satellite system (GNSS) 7-26
global positioning system (GPS) 3-27, 7-27
  components 7-27
  errors 7-33
  familiarization 7-34
  function of 7-28
  instrument approaches 7-31
  nearest airport function 11-9
  substitution 7-28
graveyard spiral 1-6
ground lighting illusions 1-9

ground proximity warning system (GPWS) 3-34
ground speed 7-19
ground wave...7-2
gyroscopic systems, power sources 3-16
  pneumatic systems..3-16
  vacuum pump systems 3-17
gyroscopic instruments
  attitude indicator...3-18

H

hazard identification...1-13
hazardous attitudes...1-18
  and antidotes...1-18
Hazardous Inflight Weather Advisory Service
(HIWAS) 8-10
haze 1-9
head up display (HUD) 3-34, 7-49
heading... 5-13, 5-44
heading indicators 3-19, 4-7, 5-7, 5-38, 6-6
height above airport (HAA) 8-27
height above landing (HAL) 8-27
height above threshold elevation (HAT).....................8-27
helicopter trim 4-10
holding 10-9
  DME 10-13
  instructions 10-10
  patterns 7-1
  procedures 10-10
homing 7-5
horizontal situation indicator (HSI) 3-22, 5-38
human
  factors 1-1
  resources...1-14

I

IAP minimums..10-21
ICAO cold temperature error table 3-6
ICAO Standard Atmosphere 2-5
icing...2-12
  types of 2-13
identifying intersections...8-7
IFR en route and terminal operations..........................7-28
IFR en route charts..8-6
IFR flight plan...10-2
  canceling...10-3
IFR Flight using GPS..7-30
illusions leading to spatial disorientation.......................1-5
IMSAFE Checklist...1-13
indirect indication 5-3, 5-6, 6-4
induced drag..2-3
induction icing 2-13
inertia navigation systems (INS).................................7-36
  components...7-37
  errors...7-37
initial approach fix (IAF)......................... 7-23, 8-16, 8-18
inoperative components 8-27
instantaneous vertical speed indicator (IVSI)........ 3-8, 4-6
instrument
  approach capabilities 7-36
  approach systems..7-37
  cross-check 4-10, 4-24, 6-2
instrument approach procedures (IAPs)...... 7-17, 8-2, 8-12
instrument approach procedures, compliance with....10-12
instrument approaches
  to civil airports 10-13
  radar monitoring of...10-18
instrument flight..6-2
instrument flight rules (IFR) 3-1, 10-1
instrument interpretation 6-3
instrument landing systems (ILS) 7-37
  components...7-39
  errors...7-44
  function...7-42
instrument takeoffs.................................. 5-29, 5-60, 6-17
  common errors.................................... 5-29, 5-61, 6-18
instrument weather flying 10-22
integrated flight control system 3-24
intercepting lead radials 7-19
interference drag 2-3
international civil aviation organization (ICAO)... 2-5, 8-1
international standard atmosphere (ISA) 2-5
inversion illusion...1-6
inverted-V cross-check...4-11
inverter 3-18
isogonic lines 3-12

J

jet routes..8-5

K

Kollsman window 3-4, 9-3

L

lag.. 5-5, 6-5
land as soon as possible 6-17
land as soon as practical..6-17
land immediately...6-17
landing..10-22
landing minimums 8-23
large airplanes 2-10
Law of Inertia..2-9

Law of Momentum 2-4
Law of Reaction 2-4
layers of the atmosphere 2-5
lead radial 7-19
leans, the 1-5
learning methods
  control and performance.............................. 4-2, 4-17
  primary and supporting 4-2
letters of agreement (LOA) 9-14
leveling off.....................................5-16, 5-17, 5-48, 5-50
lift 2-2, 2-6
lines of magnetic flux..3-11
load factor 2-11
local area augmentation system (LAAS) 7-36, 8-32
localizer (LOC) 7-39
localizer type directional aid (LDA) 7-45
long range navigation (LORAN) 7-3, 7-24
  components...7-25
  errors...7-26
  function of 7-26
loss of alternator/generator for electronic flight instrumentation 11-5
lubber line 3-11

M

mach number...3-10
machmeters 3-10
magnetic compass, basic aviation 3-11
  induced errors...3-12
magnetic bearing (MB) 7-3
magnetic heading (MH) 7-3
magnetism 3-10
margin identification..8-12
marker beacons 7-37, 7-40, 7-44
maximum authorized altitude (MAA) 8-7
mean sea level (MSL) 10-9
medical factors 1-12
  acute fatigue...1-12
  alcohol 1-12
  chronic fatigue...1-13
  fatigue...1-12
microwave landing system (MLS)..................................7-45
middle markers (MMs) 7-39
mileage breakdown 8-10
military operations areas (MOAs)...................................8-4
military training routes (MTRs)......................................8-4
minimum crossing altitude (MCA)..................................8-7
minimum descent altitude (MDA) 7-32, 8-21
minimum en route altitude (MEA) 8-6
minimum obstruction clearance altitude (MOCA) 8-6
minimum reception altitude (MRA) 8-6
minimum safe altitude (MSA) 8-16, 8-18
minimum vectoring altitudes (MVAs) 9-6
minimums section 8-23
missed approach point (MAP) 7-23
missed approach procedure..8-23
missed approach waypoint (MAHWP)...........................7-32
mixed ice 2-14
Mode C..3-7
  altitude reporting 9-3
models for practicing ADM
  perceive, process, perform 1-17
  DECIDE model, the 1-17
monopulse secondary surveillance radar (MSSR).......9-12
multi-function display (MFD)................ 3-27, 3-28, 11-12
  navigating page groups..11-10
  nearest airports, using..11-10

N

National Aeronautical Charting Group (NACG) 8-2
National Airspace System (NAS) 8-1, 9-1
National Security Areas (NSA).......................................8-4
National Transportation Safety Board (NTSB 2-16
nautical miles (NM) 7-17
navigation/communication (NAV/COM) equipment.......... .... 9-2, 9-3
navigation features 8-7
navigation instruments 4-2, 4-19
nearest airport page group..11-10
nearest airports page soft keys 11-10
nerves 1-5
new technologies...8-10
Newton's First Law of Motion Law of Inertia 2-4
Newton's Second Law of Motion Law of Momentum..2-4
Newton's Third Law of Motion Law of Reaction 2-4
no-gyro approach 10-18
nondirectional beacon (NDB) 7-3, 8-7
nonprecision approach 9-7
nonstandard pressure on an altimeter............................3-6
normal command 2-7
North American Route Program (NRP)..........................8-6
nose high attitudes...5-27
nose low attitudes...5-28
notices to airmen (NOTAM) 7-10, 8-4

O

obstical clearance surface 7-32
obstacle departure procedures (ODP) 8-12
off-route obstruction clearance altitude (OROCA 8-6
operating on the main battery 11-5
operational errors 7-45
optical illusions 1-9

  featureless terrain illusion 1-9
  fog...1-9
  ground lighting illusions...1-9
  haze..1-9
  how to prevent landing errors due to optical illusions 1-9
  runway width illusion...1-9
  runway and terrain slopes illusion.............................1-9
  water refraction 1-9
orientation 1-2
oscillation error 3-14
otolith organs 1-4, 1-5
outer markers (OMs)..7-39
outside air temperature (OAT) 11-2
overcontrolling.. 4-7, 5-4, 6-3
overpower 5-9

P

parasite drag...2-3
partial panel flight 5-36
performance instruments.............................. 4-2, 4-19
physiological and psychological factors 1-11
pilot briefing...8-12
Pilot's Operating Handbook/Airplane Flight Manual (POH/AFM)..3-3
pilot/static instruments 3-3
pilot/static systems 3-2
  failure 11-7
pitch control 4-4, 4-20-21, 5-2, 5-34, 6-3
pitch/power relationship..2-6
pitot pressure 3-2
pitot-static head 3-2
plan view 8-16
  course reversal elements..8-20
planning the descent and approach 10-8
preflight...10-2
profile view 8-21
pneumatic systems 3-16
  failure 11-7
POH/AFM..10-2
position
  error 3-3
  reports..10-7
postural considerations...1-7
power...2-10, 5-13, 5-25, 5-45
  control.................................. 4-4, 4-8, 4-21, 5-8, 5-39
  settings 5-9, 5-39
precession...3-16
  error 5-7, 6-6
precipitation static (P-static) 11-3
precision approach 7-37
precision approach path indicator (PAPI) 1-10
precision approach radar (PAR) 7-49, 10-17
precision runway monitor (PRM) 9-12
  RADAR 9-12
  benefits 9-12
preferred IFR routes 8-5
pressure
  altitude 2-5, 9-3
  density 2-5
  indicating systems 3-18
preventing landing errors due to optical illusions..........1-9
primary
  bank 4-23
  pitch 4-22
  power 4-23
  yaw 4-23
primary and supporting method 4-4, 4-21
primary flight display (PFD).......................................3-27
  additional information for specific airport 11-11
  nearest airports, using...11-10
procedure turn 7-32, 8-20
  holding in lieu of 8-20
  standard 45° 5-30
  80/260...5-31
profile view 8-21
propeller icing 2-16
propeller/rotor modulation error 7-2

R

racetrack pattern...5-30
radar 9-3
  limitations...7-50
  transponders 9-3
radar controlled departures 10-5
radar navigation (ground based) 7-49
  functions of...7-49
radials...7-10
radio altimeter 3-30
radio frequency (RF 7-16
radio magnetic indicator (RMI) 3-15, 7-4
radio wave 7-2
radius of turn 2-11
rate of turns 2-10
receiver autonomous integrity monitoring (RAIM).....7-28
receiver-transmitter (RT)..3-31
rectangular cross-check..4-11
reduced vertical separation minimum (RVSM) 3-7
reference circle..8-18

regions of command.......2-7
  normal command.......2-7
  reversed command.......2-8
relative bearing (RB).......7-3
relative wind....... 2-2, 2-4
remote communications outlet (RCO).......8-10
remote indicating compass.......3-15
repeatable accuracy.......7-25
required navigation performance.......7-46
required navigation instrument system inspection.......3-34
reversal of motion.......1-8
RNAV instrument approach charts.......8-32
reversed command.......2-8
reverse sensing.......7-12
rhodopsin.......1-2
rigidity.......3-16
rime ice.......2-14
risk.......1-13
risk analysis.......1-13
RNAV (See area navigation) runway width
illusion....... 1-9, 1-10
  and terrain slopes illusion....... 1-9, 1-10
runway end identifier lights (REIL).......7-40
runway visual range (RVR)....... 8-27, 10-22
runway visual value (RVV).......10-22

S

safety systems.......3-30
scanning techniques.......4-24
selected radial cross-check....... 4-11, 4-24
selective availability (SA).......7-33
semicircular canals.......1-4
sensitive altimeter.......3-3
  principle of operation.......3-3
sensory systems for orientation.......1-2
servo failure.......6-17
side-step maneuver.......10-20
simplex.......9-2
simplified directional facility (SDF).......7-45
single-pilot resource management (SRM).......1-14
situational awareness....... 1-14, 11-11
skin friction drag.......2-3
sky wave.......7-2
slip/skid indicator.......5-39
slow-speed flight.......2-8
small airplanes.......2-9
somatogravic illusion.......1-6
space wave.......7-2
spatial disorientation.......1-2
  coping with spatial disorientation.......1-8
  demonstration of spatial disorientation.......1-7
special use airspace.......8-2
speed stability.......2-7
St. Elmo's Fire....... 7-3, 11-3
stall warning systems.......2-16
standard entry procedures.......10-11
standard holding pattern
  no wind.......10-9
  with wind.......10-9
standard instrument departure procedures (SID).......10-5
standard rate of turn....... 2-11, 5-19, 5-51
  establishing.......5-51
  common errors.......5-51
standard terminal arrival routes (STAR)....... 8-12, 10-9
standby battery.......11-6
static longitudinal stability.......2-8
static pressure.......3-2
steep turns....... 5-22, 5-53
stepdown fixes.......8-21
straight-and-level flight.......4-22, 5-2, 5-34, 6-3
  airspeed changes....... 5-11, 5-40
  common errors.......6-7
  power control during.......6-7
straight climbs and descents....... 5-14, 5-46
  common errors.......5-17, 5-50, 6-13
stress.......1-11
structural icing....... 2-13, 10-24
suction relief valve.......3-17
synchro.......3-15
synthetic vision.......3-27
system status.......7-33
systems preflight procedures
  before engine start.......3-36
  after engine start.......3-37
  taxiing and takeoff.......3-37
  engine shut down.......3-37

T

tactical air navigation (TACAN).......7-8, 8-7, 10-7
tailplane stall symptoms.......2-16
task management.......1-15
teardrop
  patterns.......5-31
  procedure.......8-21
techniques.......5-1
  for electrical usage....... 11-5, 11-6
  master battery switch.......11-5
  operating on the main battery....... 11-5, 11-6
temporary flight restrictions (TFRs).......8-4
tension....... 4-10, 4-13
terminal arrival area (TAA).......8-18
terminal instrument approach procedures (TERPs).......8-12

Terminal Procedures Publications (TPP) 8-12
terminal radar approach control (TRACON) 9-6
terrain alerting systems 3-34
terrain awareness and warning systems (TAWS).........3-34
threshold crossing height (TCH).................................8-32
thrust 2-2, 2-3, 2-6, 2-10
thunderstorm encounter, inadvertent 11-2
thunderstorms... 9-11, 10-25
tilting to right or left..1-8
time factors...10-12
time and speed table..8-27
timed turns 5-21, 5-53, 6-13
Title 14 of the Code of Federal Regulations
(14 CFR) 1-12, 3-2, 7-16, 8-4, 8-11, 10-2, 11-8
touch down zone elevation (TDZE)............................8-27
Tower En Route Control (TEC) 8-6, 9-7
tracking 7-5
    to and from the station..7-14
Traditional navigation systems......................................7-3
traffic
    advisory systems 3-31
    alert systems 3-31
    alert and collision avoidance system (TCAS) 3-31
    avoidance...11-14
    avoidance systems..3-31
    information system (TIS) 3-31
transcribed weather broadcast (TWEB)......................8-10
transponder..9-3
    codes 9-3
trend indicators..4-26
trim................ 2-8, 4-8, 4-10, 4-20, 5-12, 5-13, 5-26, 5-45
    control.. 4-8, 5-43
turbulence..10-23
turn indicator 3-20, 6-7
turn rate indicator 5-38
turn-and-slip indicator... 3-20, 5-8
turns..2-10, 5-19, 5-51, 6-13
    change of airspeed 5-24, 6-14
    climbing and descending............................... 5-24, 6-15
    common errors.. 5-25, 6-15
    compass 5-21, 5-53, 6-15
    coordinator 3-21, 4-8, 5-7
    to predetermined headings 5-20, 5-52, 6-13
    radius of 2-11
    rate of..2-10
    standard rate 2-11, 5-19, 5-51
    steep..5-22
    timed...5-21, 5-53, 6-13
turn-and-slip indicator.................................3-20, 4-8, 5-8
types of icing..2-13
types of NAVAIDS...8-7

U

ultra high frequency (UHF) 7-3
uncaging..5-29
underpower 5-39
unforecast adverse weather 11-2
unusual attitude 5-26, 6-16
    common errors.................................5-28, 5-58, 6-16
    recognizing 5-27
    recovery from 5-26, 5-55

V

vacuum pump systems 3-17
variation 3-12
vectoring 9-6
venturi tubes...3-16
vertical card magnetic compass 3-14
vertical speed indicator (VSI) 3-8, 4-5, 5-4, 6-5
very high frequency (VHF)..9-2
very high frequency omni-directional
range (VOR)..7-8
    accuracy 7-16
    function of 7-12
    operational errors...7-14
    receiver accuracy check 7-16
vestibular.. 1-2, 1-4, 1-5
vestibular illusions 1-5
VFR Over-The-Top..10-27
VFR-On-Top 10-26
Victor airways 8-4
visual approach slope indicator (VASI) 7-41
visual descent point (VDP) 8-21
visual flight rules (VFR) 2-1, 3-1, 4-16, 6-1
visual illusions 1-7
visual meteorological conditions
(VMC)..1-3, 7-31, 7-34, 9-14
volcanic ash..10-24
VOR/DME RNAV 7-23
    components..7-23
    errors..7-24
    function of 7-23
VOR test facility (VOT) 7-16
VMC (See visual meteorological conditions)

W

water refraction 1-9
waypoint..7-8
weather and radar processor (WARP)..........................9-11
weather avoidance assistance.......................................9-11
weather conditions 10-22
weather information and communication features.......8-10

weight.. 2-2, 2-3
wet type vacuum pump 3-17
wide area augmentation system (WAAS) 7-34
windshields 2-16, 2-17
wing, the...2-2
wind correction angle (WCA).......................................7-5
wind shear 10-25
work 2-10

Z

zone of confusion..7-12

Introduction to Laboratory Techniques

Pavia
Kriz
Lampman
Engel

Harcourt *Custom Publishers*
Harcourt College Publishers
Fort Worth Philadelphia San Diego New York Orlando
Austin San Antonio Toronto Montreal Sydney Tokyo

Copyright © 1998 by Harcourt Brace & Company

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Requests for permission to make copies of any part of the work should be mailed to: Permissions Department, Harcourt Brace & Company, 6277 Sea Harbor Drive, Orlando, Florida 32887-6777.

Publisher: John Vondeling
Developmental Editor: Sandra Kiselica
Cover Designer: Kathleen Flanagan
Cover Photo Credit: Reginald Wickham Photography
Production Services: York Production Services
Product Manager: Angus McDonald

Printed in the United States of America
Introduction to Organic Laboratory Techniques: Small Scale Approach
ISBN: 0-03-024519-2
Library of Congress Catalog Card Number: 94-066238
90123 039 10987654

"WARNINGS ABOUT SAFETY PRECAUTIONS

Some of the experiments contained in this Laboratory Manual involve a degree of risk on the part of the instructor and student. Although performing the experiments is generally safe for the college laboratory, unanticipated and potentially dangerous reactions are possible for a number of reasons, such as improper measurement or handling of chemicals, improper use of laboratory equipment, failure to follow laboratory safety procedures, and other causes. Neither the Publisher nor the Authors can accept any responsibility for personal injury or property damage resulting from the use of this publication."

Contents

	Laboratory Glassware	31
Technique 1	Measurement of Volume and Weight	588
Technique 2	Heating and Cooling Methods	600
Technique 3	Reaction Methods	612
Technique 4	Filtration	634
Technique 5	Crystallization: Purification of Solids	648
Technique 6	Physical Constants: Melting Points, Boiling Points, Density	665
Technique 7	Extractions, Separations, and Drying Agents	685
Technique 8	Simple Distillation	706
Technique 9	Vacuum Distillation, Manometers	717
Technique 10	Fractional Distillation, Azeotropes	732
Technique 11	Steam Distillation	753
Technique 12	Column Chromatography	792
Technique 13	High-Performance Liquid Chromatography (HPLC)	787
Technique 14	Thin-Layer Chromatography	792
Technique 15	Gas Chromatography	808
Technique 16	Sublimation	824
Technique 17	Polarimetry	829
Technique 18	Refractometry	836

Technique 19 Preparation of Samples for Spectroscopy 842
Technique 20 Guide to the Chemical Literature 861
Appendix 1 Tables of Unknown and Derivatives 878
Appendix 2 Procedures for Preparing Derivatives 888
Appendix 3 Infrared Spectroscopy 892
Appendix 4 Nuclear Magnetic Resonance Spectroscopy 911
Appendix 5 Carbon-13 Nuclear Magnetic Resonance Spectroscopy 931
Appendix 6 Index of Spectra 940
Index 943

LABORATORY GLASSWARE

Since your glassware is expensive and you are responsible for it, you will want to give it proper care and respect. If you read this section carefully and follow the procedures presented here, you may be able to avoid some unnecessary expense. You may also save time, since cleaning problems and replacing broken glassware are time-consuming.

For those of you who are unfamiliar with the equipment found in an organic laboratory or who are uncertain about how such equipment should be treated, this section will provide some useful information. Topics such as cleaning glassware, caring for glassware when using corrosive or caustic reagents, and assembling components from your organic laboratory kit are included. At the end of this chapter are illustrations and names of most of the equipment you are likely to find in your drawer or locker.

CLEANING GLASSWARE

Glassware can be cleaned easily if you clean it immediately. It is good practice to
do your "dish washing" right away. With time, the organic tarry materials left in a con-
tainer begin to attack the surface of the glass. The longer you wait to clean glassware, the
more extensively this interaction will have progressed. Cleaning is then more difficult, be-
cause water will no longer wet the surface of the glass as effectively. If you are not able
to wash your glassware immediately after use, you should soak the dirty pieces in soapy
water. A half-gallon plastic container provides a convenient vessel in which to soak and
wash your glassware. The use of a plastic container also helps to prevent the loss of small
pieces of equipment.

Various soaps and detergents are available for washing glassware. They should be
tried first when washing dirty glassware. Organic solvents can also be used, since the
residue remaining in dirty glassware is likely to be soluble in some organic solvent. After
the solvent has been used, the flask probably will have to be washed with soap and wa-
ter to remove the residual solvent. When you use solvents in cleaning glassware, use cau-
tion since the solvents are hazardous (see the chapter entitled "Laboratory Safety"). You
should try to use fairly small amounts of a solvent for cleaning purposes. Usually less
than 5 mL will be sufficient. Acetone is commonly used, but it is expensive. Your **wash**
acetone can be used effectively several times before it is "spent." Once your acetone is
spent, dispose of it as directed by your instructor. If acetone does not work, other organic
solvents such as methylene chloride or toluene can be used in the same way as acetone.

Caution: Acetone is very flammable. Do not use it around flames.

For troublesome stains and residues that insist on adhering to the glass in spite of your best efforts, a mixture of sulfuric acid and nitric acid can be used. Cautiously add about 20 drops of concentrated sulfuric acid and five drops of concentrated nitric acid to the flask.

Caution: You must wear safety glasses when you are using this cleaning solution. Do not allow the solution to come into contact with your skin or your clothing. It will cause severe burns and create holes in your clothing. It is also possible that the acids will react with the residue in the container.

Swirl the acid mixture in the container for a few minutes. If necessary, place the
glassware in a warm water bath and heat cautiously to accelerate the cleaning process.
Continue heating until any sign of a reaction ceases. When the cleaning procedure is com-
pleted, decant the mixture into an appropriate waste container.

Caution: Do not pour the acid solution into a waste container that is intended for organic wastes.

Rinse the piece of glassware thoroughly with water and then wash with soap and water. For most common organic chemistry applications, any stains that survive this treatment are not likely to cause difficulty in subsequent laboratory procedures.

If the glassware is contaminated with stopcock grease, rinse the glassware with a small amount (1–2 mL) of methylene chloride. Discard the rinse solution into an appropriate waste container. Once the grease is removed, wash the glassware with soap or detergent and water.

DRYING GLASSWARE

The easiest way to dry glassware is to allow it to stand overnight. Vials, flasks, and beakers should be stored upside down on a piece of paper towel to permit the water to drain from them. Drying ovens can be used to dry glassware if they are available, and if they are not being used for other purposes. Rapid drying can be achieved by rinsing the glassware with acetone and air-drying it or placing it in an oven. First, thoroughly drain the glassware of water. Then rinse it with one or two *small* portions (1–2 mL) of acetone. Do not use any more acetone than is suggested here. Return the used acetone to a waste acetone container for recycling. After you rinse the glassware with acetone, dry it by placing it in a drying oven for a few minutes or allow it to air-dry at room temperature. The acetone can also be removed by aspirator suction. In some laboratories, it may be possible to dry the glassware by blowing a *gentle* stream of dry air into the container. (Your laboratory instructor will indicate if you should do this.) Before drying the glassware with air, you should make sure that the air line is not filled with oil. Otherwise, the oil will be blown into the container, and you will have to clean it again. It is not necessary to blast the acetone out of the glassware with a wide-open stream of air; a gentle stream of air is just as effective and will not startle other people in the room.

You should not dry your glassware with a paper towel unless the towel is lint-free. Most paper will leave lint on the glass that can interfere with subsequent procedures performed in the equipment. Sometimes it is not necessary to dry a piece of equipment thoroughly. For example, if you are going to place water or an aqueous solution in a container, it does not need to be completely dry.

GROUND-GLASS JOINTS

It is likely that the glassware in your organic kit has **standard-taper ground-glass joints.** Joints may be either inner (male) or outer (female). Each joint is ground to a precise size, which is designated by the symbol ₮ followed by two numbers. A common joint size in most organic glassware kits is $\$$ 19/22. The first number indicates the diameter (in millimeters) of the joint at its widest point, and the second number refers to its length. One advantage of standard-taper joints is that the pieces fit together snugly and form a good seal. In addition, standard-taper joints allow all glassware components with the same joint size to be connected, thus permitting the assembly of a wide variety of apparatus. One disadvantage of glassware with ground-glass joints, however, is that it is very expensive.

It is important to make sure no solid or liquid is on the joint surfaces. Such material will lessen the efficiency of the seal, and the joints may leak. Also, if the apparatus is to be heated, material caught between the joint surfaces will increase the tendency for the joints to stick. If the joint surfaces are coated with liquid or adhering solid, you should wipe them with a cloth or lint-free paper towel before assembling.

Another disadvantage of glassware with ground-glass joints is that the joints have some tendency to stick together. To prevent this sticking, you can use a lubricant. Typical lubricants include a variety of hydrocarbon-based stopcock greases, such as Lubriseal and Apiezon. Silicone greases are also used. To apply the grease, coat the inner (male) joint with a *very thin* film of lubricant. You should be careful not to apply too much grease, since excess grease may contaminate the materials contained within the apparatus. A properly lubricated joint appears clear, with no striations. Contamination due to excess grease is a particular problem with silicone grease, which is much harder to remove than hydrocarbon-based greases. Because of this possible contamination, many chemists prefer to use grease only when necessary. The joints usually seal well without grease. In some situations, grease is necessary, including distillations under reduced pressure—when joints must be airtight—and in reactions involving strong caustic solutions (see below, "Etching Glassware"). Whenever there is doubt, it never hurts to use grease. Before lubricating the joints, always make sure that they are free of any adhering liquid or solid.

SEPARATING GROUND-GLASS JOINTS

The most important thing you can do to prevent ground-glass joints from becoming "frozen" or stuck together is to disassemble the glassware as soon as possible after a procedure is completed. Even when this precaution is followed, ground-glass joints may become stuck tightly together. The same is true of glass stoppers in bottles. Since smaller items of glassware may be small and very fragile, it is relatively easy to break a piece of glassware when trying to pull two pieces apart. If the pieces do not separate easily, you must be careful when you try to pull them apart. The best way is to hold the two pieces, with both hands touching, as close as possible to the joint. With a firm grasp, try to loosen the joint with a slight twisting motion (do not twist very hard). If this does not work, try to pull your hands apart without pushing sideways on the glassware.

If it is not possible to pull the pieces apart, the following methods may help. A frozen joint can sometimes be loosened if you tap it *gently* with the wooden handle of a spatula. Then, try to pull it apart as already described. If this procedure fails, you may try heating the joint in hot water or a steam bath. If this heating fails, the instructor may be able to advise you. As a last resort, you may try heating the joint in a flame. You should not try this unless the apparatus is hopelessly stuck, because heating by flame often causes the joint to expand rapidly and crack or break. If you use a flame, make sure the joint is clean and dry. Heat the outer part of the joint slowly, in the yellow portion of a low flame, until it expands and breaks away from the inner section. Heat the joint very slowly and carefully, or it may break.

ETCHING GLASSWARE

Glassware that has been used for reactions involving strong bases such as sodium hydroxide or sodium alkoxides must be cleaned thoroughly *immediately* after use. If these caustic materials are allowed to remain in contact with the glass, they will etch the glass permanently. The etching makes later cleaning more difficult, since dirt particles may become trapped within the microscopic surface irregularities of the etched glass. Furthermore, the glass is weakened, so the lifetime of the glassware is decreased. If caustic materials are allowed to come into contact with ground-glass joints without being removed promptly, the joints will become fused or "frozen." It is extremely difficult to separate fused joints without breaking them.

ASSEMBLING THE APPARATUS

Care must be taken when assembling the glass components into the desired apparatus. You should always remember that Newtonian physics applies to chemical apparatus, and unsecured pieces of glassware are certain to respond to gravity. You should always clamp the glassware securely to a ring stand. Throughout this textbook, the illustrations of the various glassware arrangements include the clamps that attach the apparatus to a ring stand. You should assemble your apparatus using the clamps as shown in the illustrations.

ATTACHING RUBBER TUBING TO EQUIPMENT

When you attach rubber tubing to the glass apparatus or when you insert glass tubing into rubber stoppers, you should lubricate the rubber tubing or the rubber stopper with either water or glycerin beforehand. Without such lubrication, it can be difficult to attach rubber tubing to the sidearms of items of glassware such as condensers and filter flasks. Furthermore, glass tubing may break when it is inserted into rubber stoppers. Water is a good lubricant for most purposes. Do not use water as a lubricant when it might contaminate the reaction. Glycerin is a better lubricant than water and should be used when there is considerable friction between the glass and rubber. If glycerin is the lubricant, be careful not to use too much.

DESCRIPTION OF EQUIPMENT

The figures that follow on pages 37–39 include glassware and equipment that are commonly used in the organic laboratory. Your glassware and equipment may vary slightly from the pieces shown.

The components of the organic kit recommended for use in this textbook are shown in the figure on page 37. Notice that most of the joints in these pieces of glassware are $ 19/22. The organic kits used in your laboratory may have different joint sizes, but these kits will work as well with the experiments in this book as the glassware recommended in the figure.

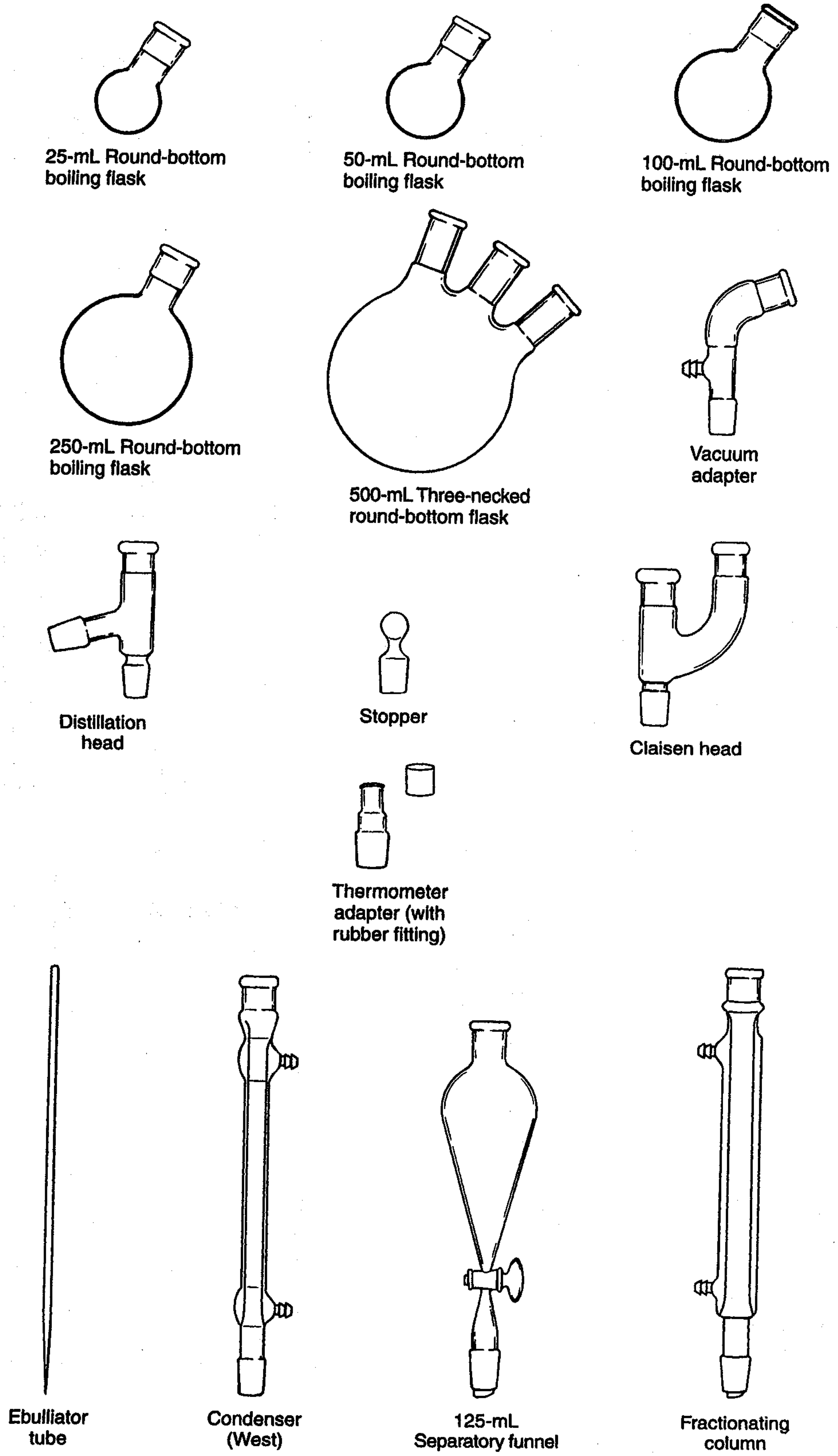
25-mL Round-bottom boiling flask
50-mL Round-bottom boiling flask
100-mL Round-bottom boiling flask
250-mL Round-bottom boiling flask
500-mL Three-necked round-bottom flask
Vacuum adapter
Distillation head
Stopper
Claisen head
Thermometer adapter (with rubber fitting)
Ebulliator tube
Condenser (West)
125-mL Separatory funnel
Fractionating column

Components of the organic laboratory kit.

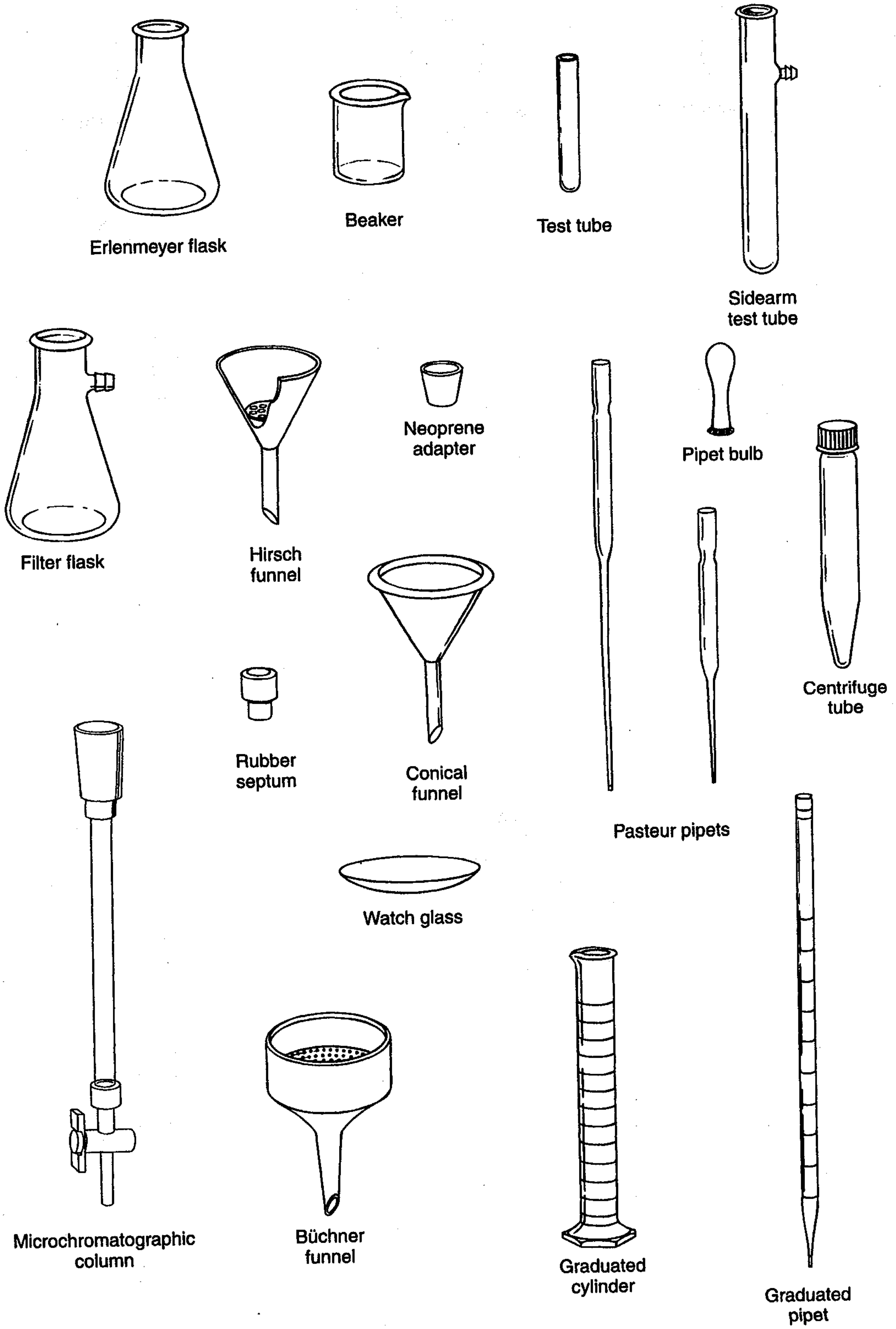
Erlenmeyer flask
Beaker
Test tube
Sidearm test tube
Filter flask
Hirsch funnel
Neoprene adapter
Pipet bulb
Centrifuge tube
Rubber septum
Conical funnel
Pasteur pipets
Watch glass
Microchromatographic column
Büchner funnel
Graduated cylinder
Graduated pipet

Equipment commonly used in the organic laboratory.

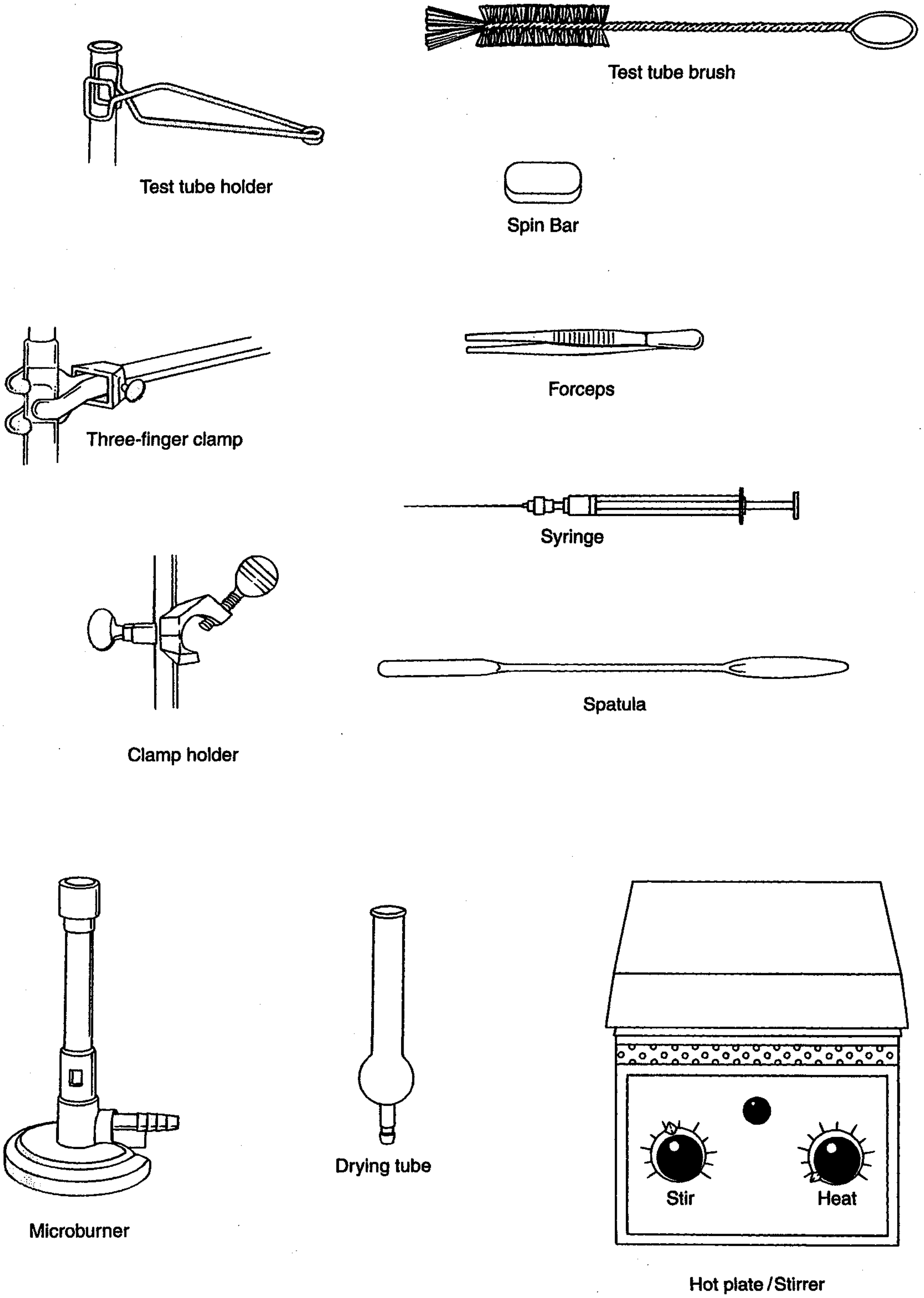

Test tube holder

Test tube brush

Spin Bar

Three-finger clamp

Forceps

Syringe

Clamp holder

Spatula

Microburner

Drying tube

Hot plate / Stirrer

Equipment commonly used in the organic laboratory.

PART TWO

The Techniques

TECHNIQUE 1

Measurement of Volume and Weight

Performing successful organic chemistry experiments requires the ability to measure solids and liquids accurately. This ability involves both selecting the proper measuring device and using this device correctly.

Liquids to be used for an experiment will usually be found in small containers in a hood. For most of the experiments in this textbook, a graduated cylinder, a dispensing pump, or a graduated pipet will be used for measuring the volume of a liquid. You will usually transfer the required volume of liquid to a round-bottom flask or an Erlenmeyer flask.

In cases where the liquid is a limiting reagent, you should pre-weigh **(tare)** the container before dispensing the liquid into the container. When the container is reweighed, you obtain the actual weight for the volume of liquid you have dispensed. The laboratory procedure usually specifies when you should weigh the liquid. When transferring the liquid to a round-bottom flask, place the flask in a beaker and tare both the flask and the beaker. The beaker keeps the round-bottom flask in an upright position and prevents spills from occurring. It is also advisable to tare a glass stopper, along with the beaker and flask, so that you can stopper the flask immediately after dispensing the liquid and prevent vapors from escaping into the air.

In cases where the liquid is not the limiting reagent, you may calculate the weight of the liquid from the volume you have delivered and the density of the liquid. Usually, densities are provided in the experimental procedures. You may calculate the weight from the following relationship:

$$\text{Weight (g)} = \text{Density (g/mL)} \times \text{Volume (mL)}$$

When using a graduated cylinder to measure small volumes (< 3 mL) of a limiting reagent, it is important to transfer the liquid **quantitatively** from the cylinder to the reaction vessel. This means that all of the liquid should be transferred. When the reagent is poured from the graduated cylinder, a small amount of liquid will remain in the cylinder. This remaining liquid can be transferred by pouring a small amount of the solvent being used in the reaction into the cylinder and then pouring this solution into the reaction vessel. If no solvent is used in the reaction, most of the remaining liquid in the cylinder can be transferred by using a Pasteur pipet.

Using a small amount of solvent to transfer a liquid quantitatively can also be applied in other situations. For example, if your product is dissolved in a solvent and the procedure instructs you to transfer the reaction mixture from a round-bottom flask to a separatory funnel, after pouring most of the liquid into the funnel, a small amount of solvent could be used to transfer the rest of the product quantitatively.

Solids to be used for an experiment will usually be found near the balance. When an accurate measurement is required, solids must be weighed on a balance that reads at least to the nearest decigram (0.01 g). To weigh a solid, place your round-bottom flask or

Erlenmeyer flask in a small beaker and take these with you to the balance. Place a piece of paper that has been folded once on the balance pan. The folded paper will enable you to pour the solid into the flask without spilling. Use your spatula to aid the transfer of the solid to the paper. Never weigh directly into the flask and never pour, dump, or shake a material from a bottle. While still at the balance, carefully transfer the solid from the paper to your flask. The flask should be placed in a beaker while transferring the solid. The beaker acts as a trap for any material that fails to make it into the container. It also supports the round-bottom flask so that it does not fall over. It is not necessary to obtain the exact amount specified in the experimental procedure, and trying to be exact requires too much time at the balance. For example, if you obtained 1.52 g of a solid, rather than the 1.50 g specified in a procedure, the actual amount weighed should be recorded in your notebook. Use the amount you weighed to calculate the theoretical yield, if this solid is the limiting agent.

Careless dispensing of liquids and solids is a hazard in any laboratory. When reagents are spilled, you may be subjected to an unnecessary health or fire hazard. In addition, you may waste expensive chemicals, destroy balance pans and clothing, and damage the environment. Always clean up any spills immediately.

1.1 GRADUATED CYLINDERS

Graduated cylinders are most often used to measure liquids for the experiments in this book (see Fig. 1.1). The most common sizes are 10 mL, 25 mL, 50 mL, and 100 mL, but it is possible that not all of these will be available in your laboratory. Volumes from

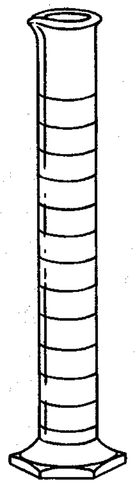

Figure 1.1 Graduated cylinder.

about 2 to 100 mL can be measured with reasonably good accuracy provided that the correct cylinder is used. You should use the *smallest* cylinder available that can hold all of the liquid that is being measured. For example, if a procedure calls for 4.5 mL of a reagent, use a 10 mL graduated cylinder. Using a large cylinder in this case will result in a less accurate measurement. Furthermore, using any cylinder to measure less than 10% of the total capacity of that cylinder will likely result in an inaccurate measurement. Always remember that whenever using a graduated cylinder to measure the volume of a limiting reagent, you must weigh the liquid to determine the amount used accurately. You should use a graduated pipet, a dispensing pump, or an automatic pipet for accurate transfer of liquids with a volume of less than 2 mL.

If the storage container is reasonably small (< 1.0 L) and has a narrow neck, you may pour most of the liquid into the graduated cylinder and use a Pasteur pipet to adjust to the final line. If the storage container is large (> 1.0 L) or has a wide mouth, two strategies are possible. First, you may use a pipet to transfer the liquid to the graduated cylinder. Alternatively, you may pour some of the liquid into a beaker first and then pour this liquid into a graduated cylinder. Use a Pasteur pipet to adjust to the final line. Remember that you should not take more than you need. Excess material should never be returned to the storage bottle. Unless you can convince someone else to take it, it must be poured into the appropriate waste container. You should be frugal in your estimation of amounts needed.

1.2 DISPENSING PUMPS

Dispensing pumps are simple to operate, chemically inert, and quite accurate. Since the plunger assembly is made of Teflon, the dispensing pump may be used with most corrosive liquids and organic solvents. Dispensing pumps come in a variety of sizes, ranging from 1 mL to 300 mL. When used correctly, dispensing pumps can be used to deliver accurate volumes ranging from 0.1 mL to the maximum capacity of the pump. The pump is attached to a bottle containing the liquid being dispensed. The liquid is drawn up from this reservoir into the pump assembly through a piece of inert plastic tubing.

Dispensing pumps are somewhat difficult to adjust to the proper volume. Normally, the instructor or assistant will carefully adjust the unit to deliver the proper amount of liquid. As shown in Figure 1.2, the plunger is pulled up as far as it will travel to draw in the liquid from the glass reservoir. To expel the liquid from the spout into a container, you slowly guide the plunger down. With low-viscosity liquids, the weight of the plunger will expel the liquid. With more viscous liquids, however, you may need to push the plunger gently to deliver the liquid into a container. Remove the last drop of liquid on the end of the spout by touching the tip on the interior wall of the container. When the liquid being transferred is a limiting reagent or when you need to know the weight precisely, you should weigh the liquid to determine the amount accurately.

As you pull up the plunger, look to see if the liquid is being drawn up into the pump unit. Some volatile liquids may not be drawn up in the expected manner, and you will observe an air bubble. Air bubbles are commonly observed when the pump has not been

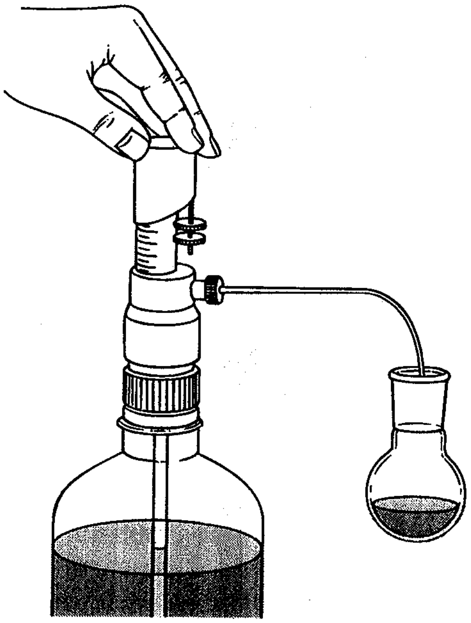

Figure 1.2 Use of a dispensing pump.

used for a while. The air bubble can be removed from the pump by dispensing, and discarding, several volumes of liquid to "reprime" the dispensing pump. Also check to see if the spout is filled completely with liquid. An accurate volume will not be dispensed unless the spout is filled with liquid before you lift up the plunger.

1.3 GRADUATED PIPETS

A widely used measuring device is the graduated serological pipet. These *glass* pipets are available commercially in a number of sizes. "Disposable" graduated pipets may be used many times and discarded only when the graduations become too faint to be seen. A good assortment of these pipets consists of the following:

1.00-mL pipets calibrated in 0.01-mL divisions (1 in 1/100 mL)

2.00-mL pipets calibrated in 0.01-mL divisions (2 in 1/100 mL)

5.0-mL pipets calibrated in 0.1-mL divisions (5 in 1/10 mL)

Never draw liquids into the pipets using mouth suction. A pipet pump or a pipet bulb, not a rubber dropper bulb, must be used to fill pipets. Two types of pipet pumps and

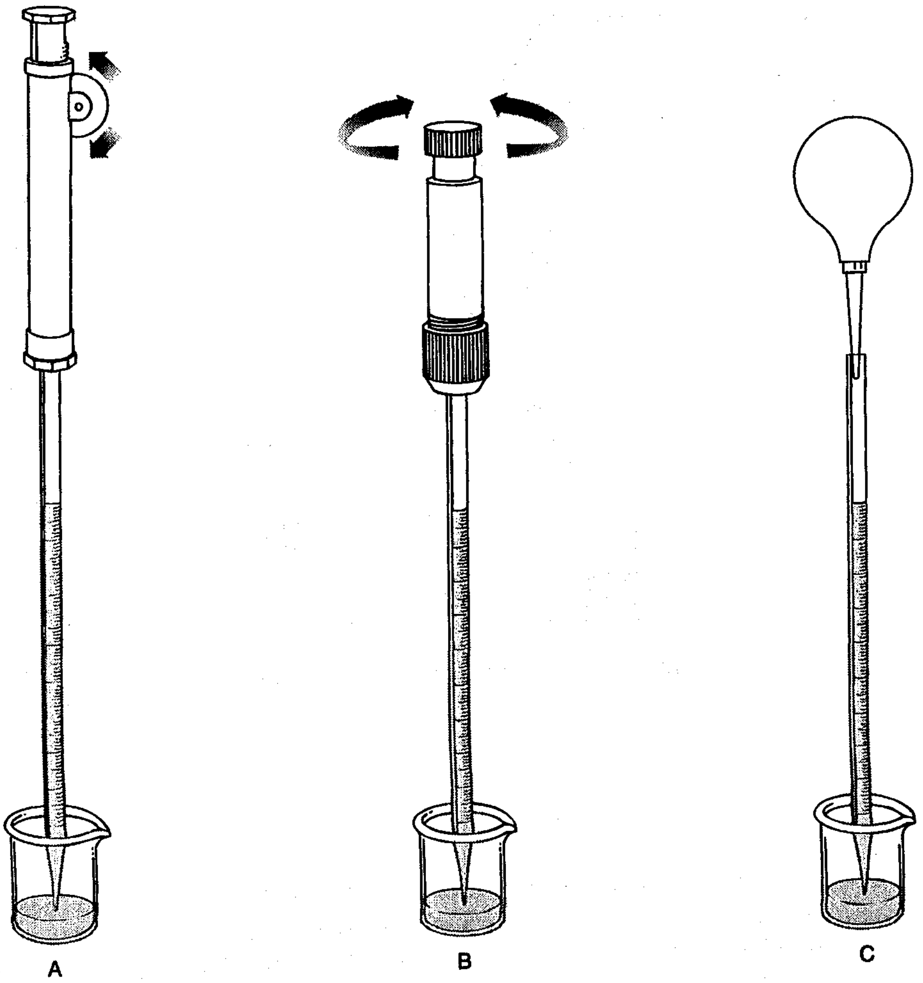

Figure 1.3 Pipet pumps and pipet bulb.

a pipet bulb are shown in Figure 1.3. A pipet fits snugly into the pipet pump, and the pump can be controlled to deliver precise volumes of liquids. Control of the pipet pump is accomplished by rotating a knob on the pump. Suction created when the knob is turned draws the liquid into the pipet. Liquid is expelled from the pipet by turning the knob in the opposite direction. The pump works satisfactorily with organic, as well as aqueous, solutions.

The style of pipet pump shown in Figure 1.3A is available in four sizes. The top of the pipet must be inserted securely into the pump and held there with one hand to obtain an adequate seal. The other hand is used to load and release the liquid. The pipet pump shown in Figure 1.3B may also be used with graduated pipets. With this style of pipet, the top of the pipet is held securely by a rubber O-ring, and it is easily handled with one hand. You should be certain that the pipet is held securely by the O-ring before using it. Disposable pipets may not fit tightly in the O-ring because they often have smaller diameters than nondisposable pipets.

An alternative, and less expensive, approach is to use a rubber pipet bulb, shown in Figure 1.3C. Use of the pipet bulb is made more convenient by inserting a plastic automatic pipet tip into a rubber pipet bulb.[1] The tapered end of the pipet tip fits snugly into the end of a pipet. Drawing the liquid into the pipet is made easy, and it is also convenient to remove the pipet bulb and place a finger over the pipet opening to control the flow of liquid.

The calibrations printed on graduated pipets are reasonably accurate, but you should practice using the pipets in order to achieve this accuracy. When accurate quantities of liquids are required, the best technique is to weigh the reagent that has been delivered from the pipet.

The following description, along with Figure 1.4, illustrates how to use a graduated pipet. Insert the end of the pipet firmly into the pipet pump. Rotate the knob of the pipet pump in the correct direction (counterclockwise or up) to fill the pipet. Fill the pipet to a point just above the uppermost mark and then reverse the direction of rotation of the knob to allow the liquid to drain from the pipet until the meniscus is adjusted to the 0.00-mL mark. Move the pipet to the receiving vessel. Rotate the knob of the pipet pump (clockwise or down) to force the liquid from the pipet. Allow the liquid to drain from the pipet until the meniscus arrives at the mark corresponding to the volume that you wish to dispense. Be sure to touch the tip of the pipet to the inside of the container before withdrawing the pipet. Remove the pipet and drain the remaining liquid into a waste receiver. Avoid transferring the entire contents of the pipet when measuring volumes with a pipet. Remember that in order to achieve the greatest possible accuracy with this method, you should deliver volumes as a *difference* between two marked calibrations.

Pipets may be obtained in a number of styles, but only three types will be described here (Figure 1.5). One type of graduated pipet is calibrated "to deliver" (TD) its total capacity when the last drop is blown out. This style of pipet, shown in Figure 1.5A, is probably the most common type of graduated pipet in use in the laboratory; it is designated by two rings at the top. Of course, one does not need to transfer the entire volume to a container. In order to deliver a more accurate volume, you should transfer an amount less than the total capacity of the pipet using the graduations on the pipet as a guide.

Another type of graduated pipet is shown in Figure 1.5B. This pipet is calibrated to deliver its total capacity when the meniscus is located on the last graduation mark near the bottom of the pipet. For example, the pipet shown in the Figure 1.5B delivers 10.0 mL of liquid when it has been drained to the point where the meniscus is located on the 10.0-mL mark. With this type of pipet, you must not drain the entire pipet or blow it out. In contrast, notice that the pipet discussed in Figure 1.5A has its last graduation at 0.90 mL. The last 0.10-mL volume is blown out to give the 1.00-mL volume.

A nongraduated volumetric pipet is shown in Figure 1.5C. It is easily identified by the large bulb in the center of the pipet. This pipet is calibrated so that it will retain its last drop after the tip is touched on the side of the container. It must not be blown out.

---

[1] This technique was described in Deckey, G. "A Versatile and Inexpensive Pipet Bulb." *Journal of Chemical Education, 57* (July 1980): 526.

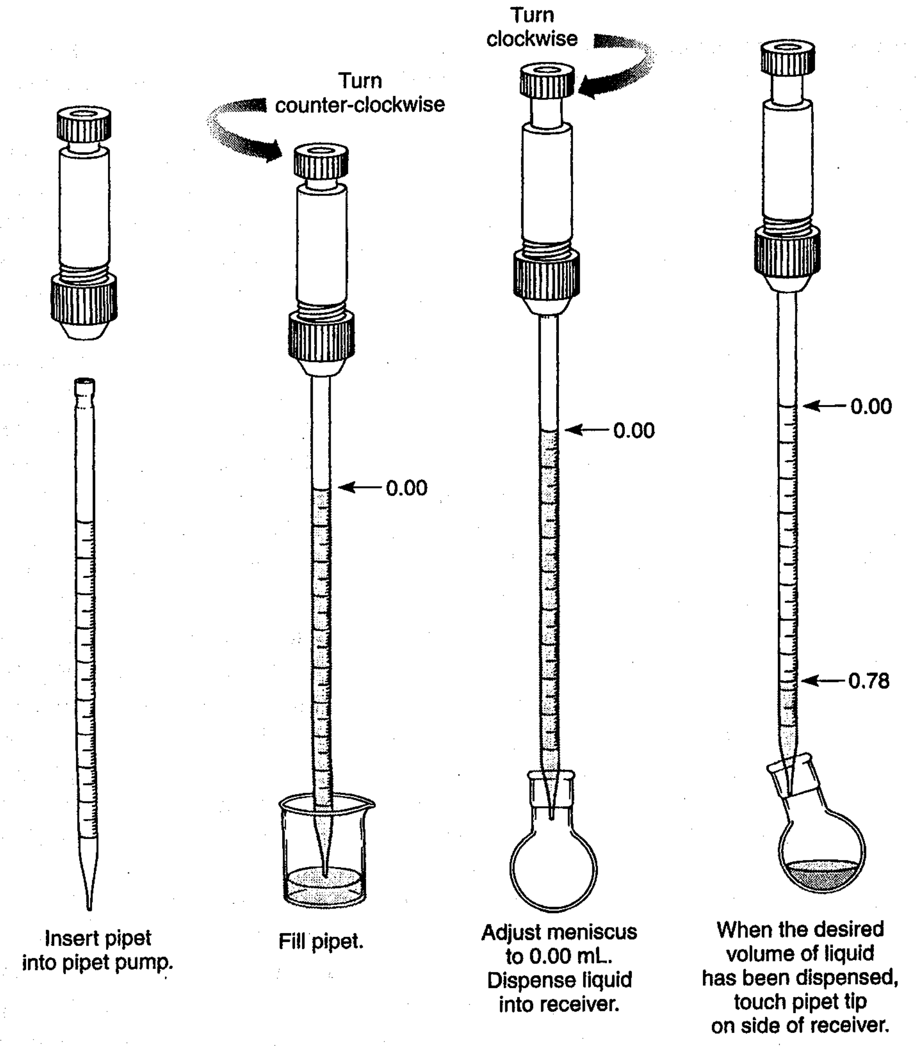
Turn
counter-clockwise
Turn
clockwise
0.00
0.00
0.00
0.78
Insert pipet
into pipet pump.
Fill pipet.
Adjust meniscus
to 0.00 mL.
Dispense liquid
into receiver.
When the desired
volume of liquid
has been dispensed,
touch pipet tip
on side of receiver.

Figure 1.4 Use of a graduated pipet. (The figure shows, as an illustration, the technique required to deliver a volume of 0.78 mL from a 1.00-mL pipet.)

These pipets often have a single colored band at the top that identifies it as a "touch-off" pipet. The color of the band is keyed to its total volume. This type of pipet is commonly used in analytical chemistry.

1.4 PASTEUR PIPETS

The Pasteur pipet is shown in Figure 1.6A with a 2-mL rubber bulb attached. There are two sizes of Pasteur pipets: a short one ($5\frac{3}{4}$ inch), which is shown in the figure, and a long one (9 inch). It is important that the pipet bulb fit securely. You should not use a medicine dropper bulb because of its small capacity. A Pasteur pipet is an indispensable

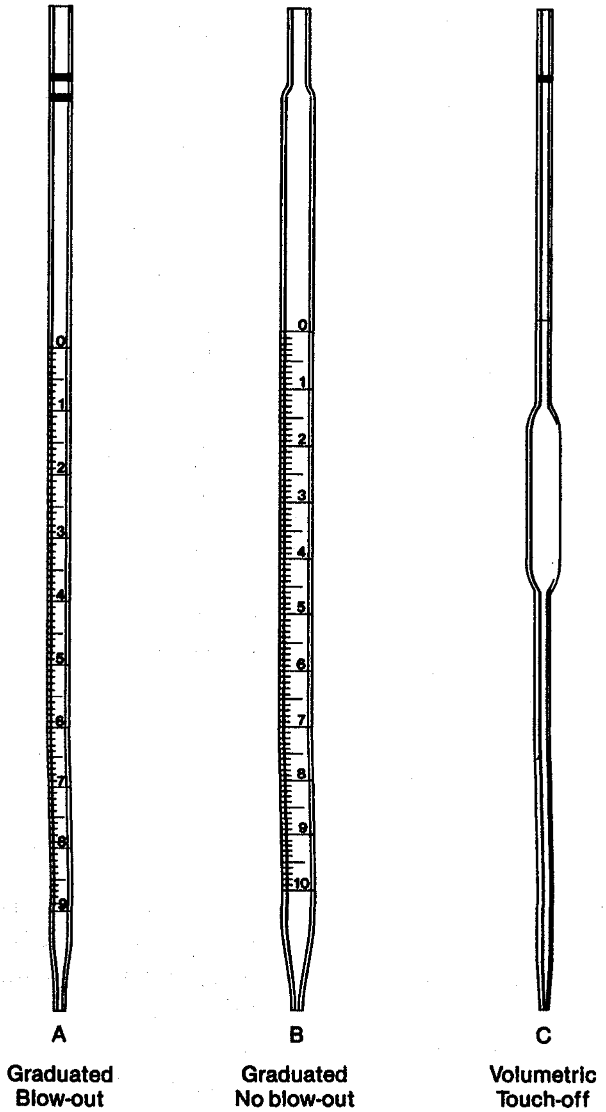

Figure 1.5 Pipets.

piece of equipment for the routine transfer of liquids. It is also used for separations (Technique 7). Pasteur pipets may be packed with cotton for use in gravity filtration (Technique 4) or packed with an adsorbent for small-scale column chromatography (Technique 12). Although they are considered disposable, you should be able to clean them for reuse as long as the tip remains unchipped.

A Pasteur pipet may be supplied by your instructor for dropwise addition of a particular reagent to a reaction mixture. For example, concentrated sulfuric acid is often dispensed in this way. When sulfuric acid is transferred, you should take care to avoid getting the acid into the rubber or latex dropper bulb.

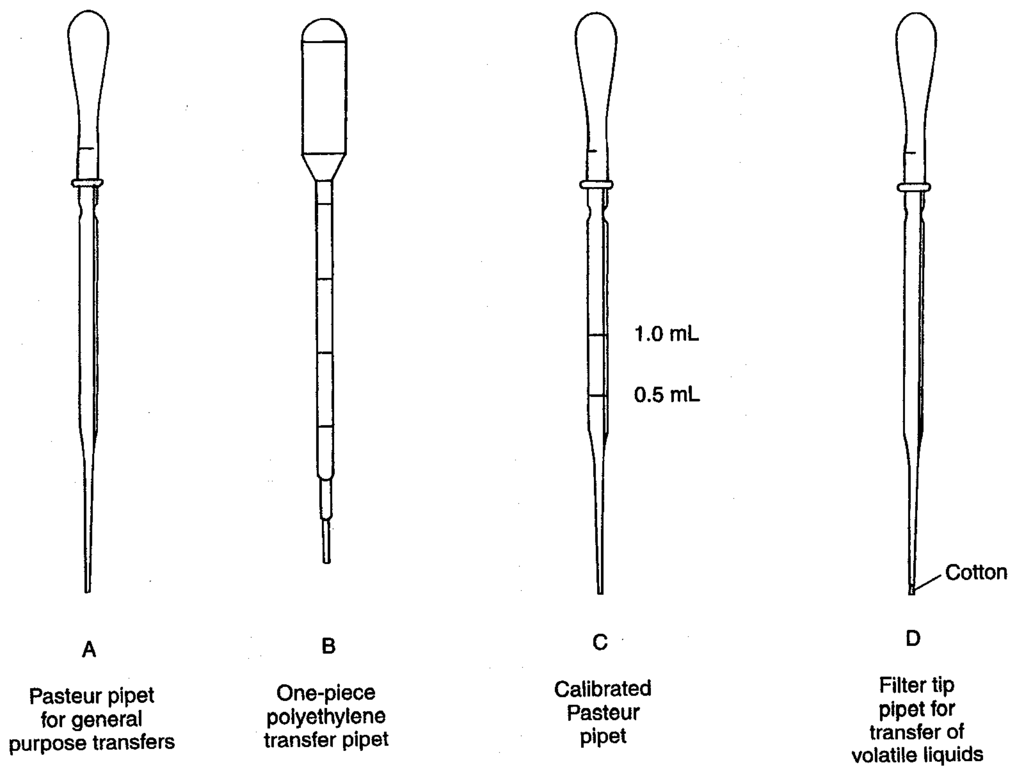
1.0 mL
0.5 mL
Cotton
A
Pasteur pipet for general purpose transfers
B
One-piece polyethylene transfer pipet
C
Calibrated Pasteur pipet
D
Filter tip pipet for transfer of volatile liquids

Figure 1.6 Pasteur and transfer pipets.

One may avoid the rubber dropper bulb entirely by using one-piece transfer pipets made entirely of polyethylene (Figure 1.6B). These plastic pipets are available in 1- or 2-mL sizes. They come from the manufacturers with approximate calibration marks stamped on them. These pipets can be used with all aqueous solutions and most organic liquids. They cannot be used with a few organic solvents.

Pasteur pipets may be calibrated for use in operations where the volume does not need to be known precisely. Examples include measurement of solvents needed for extraction and for washing a solid obtained following crystallization. A calibrated Pasteur pipet is shown in Figure 1.6C. It is suggested that you calibrate several $5\frac{3}{4}$-inch pipets using the following procedure. On a balance, weigh 0.5 grams (0.5 mL) of water into a small test tube. Select a short Pasteur pipet and attach a rubber bulb. Squeeze the rubber bulb before inserting the tip of the pipet into the water. Try to control how much you depress the bulb so that, when the pipet is placed into the water and the bulb is completely released, only the desired amount of liquid is drawn into the pipet. When the water has been drawn up, place a mark with an indelible marking pen at the position of the meniscus. A more durable mark can be made by scoring the pipet with a file. Repeat this procedure with 1.0 gram of water, and make a 1-mL mark on the same pipet.

Your instructor may provide you with a calibrated Pasteur pipet and bulb for transferring liquids where an accurate volume is not required. The pipet may be used to transfer a volume of 1.5 mL or less. You may find that the instructor has taped a test tube to the side of the storage bottle. The pipet is stored in the test tube with that particular reagent.

> **Note:** You should not assume that a certain number of drops equals a 1-mL volume. The common rule that 20 drops equal 1 mL, often used for a buret, does not hold true for a Pasteur pipet!

A Pasteur pipet may be packed with cotton to create a filter-tip pipet as shown in Figure 1.6D. This pipet is prepared by the instructions given in Technique 4, Section 4.6, page 645. Pipets of this type are very useful in transferring volatile solvents during extractions and in filtering small amounts of solid impurities from solutions. A filter-tip pipet is very useful for removing small particles from a solution of a sample prepared for NMR analysis.

1.5 SYRINGES

Syringes may be used to add a pure liquid or a solution to a reaction mixture. They are especially useful when anhydrous conditions must be maintained. The needle is inserted through a septum, and the liquid is added to the reaction mixture. Caution should be used with disposable syringes as they often use solvent-soluble rubber gaskets on the plungers. A syringe should be cleaned carefully after each use by drawing acetone or another volatile solvent into it and expelling the solvent with the plunger. Repeat this procedure several times to clean the syringe thoroughly. Draw air through the barrel with an aspirator to dry the syringe.

Syringes are usually supplied with volume graduations inscribed on the barrel. Large-volume syringes are not accurate enough to be used for measuring liquids in small scale experiments. A small microliter syringe, such as that used in gas chromatography, delivers a very precise volume.

1.6 AUTOMATIC PIPETS

Automatic pipets are commonly used in microscale organic laboratories and in biochemistry laboratories. One type of adjustable automatic pipet is shown in Figure 1.7. The automatic pipet is very accurate with aqueous solutions, but it is not as accurate with organic liquids. They are available in different sizes and can deliver accurate volumes ranging from 0.10 mL to 1.0 mL. These pipets are very expensive and must be shared by the entire laboratory. If they are used in your laboratory, your instructor will give directions for the correct use of the automatic pipet.

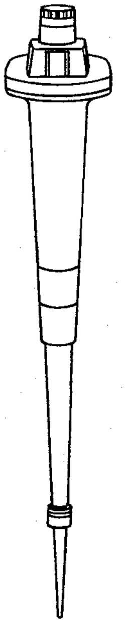

Figure 1.7 The adjustable automatic pipet.

1.7 MEASURING VOLUMES WITH BEAKERS AND ERLENMEYER FLASKS

Beakers and Erlenmeyer flasks both have graduations inscribed on them. Beakers and flasks can be used to give only a *crude* approximation of the volume. They are much less precise than graduated cylinders for measuring volume. You should use a graduated cylinder, dispensing pump, or graduated pipet for accurate measurement of liquids.

1.8 BALANCES

Solids and some liquids will need to be weighed on a balance that reads to at least the nearest decigram (0.01 g). A top-loading balance (see Fig. 1.8) works well if the balance pan is covered with a plastic draft shield. The shield has a flap that opens to allow access to the balance pan. An analytical balance (see Fig. 1.9) may also be used. This type of balance will weigh to the nearest tenth of a milligram (0.0001 g) when provided with a glass draft shield.

Modern electronic balances have a tare device that automatically subtracts the weight of a container or a piece of paper from the combined weight to give the weight of the sample. With solids, it is easy to place a piece of paper on the balance pan, press the tare device so that the paper appears to have zero weight, and then add your solid until the balance gives the weight you desire. You can then transfer the weighed solid to a container. You should al-

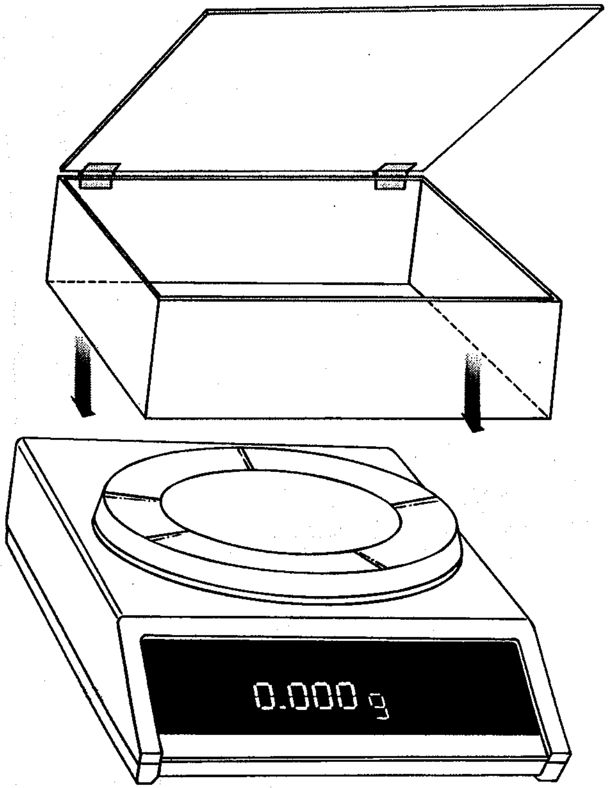

Figure 1.8 A top-loading balance with plastic draft shield.

ways use a spatula to transfer a solid and never pour material from a bottle. In addition, solids must be weighed on paper and not directly on the balance pan. Remember to clean any spills.

With liquids, you should weigh the flask to determine the tare weight, transfer the liquid with a graduated cylinder, dispensing pump, or graduated pipet into the flask, and then reweigh it. With liquids, it is usually necessary to weigh only the limiting reagent. The other liquids may be transferred using a graduated cylinder, dispensing pump, or graduated pipet. Their weights can be calculated by knowing the volumes and densities of the liquids.

PROBLEMS

1. What measuring device would you use to measure the volume under each of the conditions described below? In some cases, there may be more than one answer to the question.
  (a) 25 mL of a solvent needed for a crystallization
  (b) 2.4 mL of a liquid needed for a reaction
  (c) 1 mL of a solvent needed for an extraction

2. Assume that the liquid used in Question 1b is a limiting reagent for a reaction. What should you do after measuring the volume?

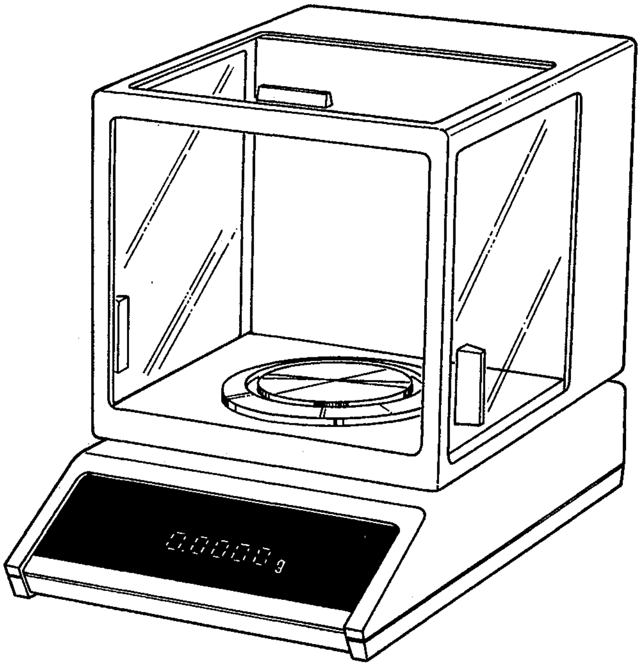

Figure 1.9 An analytical balance with glass draft shield.

3. Calculate the weight of a 2.5-mL sample of each of the following liquids:
   - **(a)** Diethyl ether (ether)
   - **(b)** Methylene chloride (dichloromethane)
   - **(c)** Acetone
4. A laboratory procedure calls for 1.46 g of acetic anhydride. Calculate the volume of this reagent needed in the reaction.

TECHNIQUE 2

Heating and Cooling Methods

Most organic reaction mixtures need to be heated in order to complete the reaction. In general chemistry, you used a Bunsen burner for heating because nonflammable aqueous solutions were used. In an organic laboratory, however, the student must heat nonaqueous solutions that may contain *highly flammable* solvents. You *should not heat organic mixtures with a Bunsen burner* unless you are directed by your laboratory instructor. Open flames present a potential fire hazard. Whenever possible you should use one of the alternative heating methods, as described in the following sections.

2.1 HEATING MANTLES

A useful source of heat for most experiments in this textbook is the heating mantle, illustrated in Figure 2.1. The heating mantle shown here consists of a ceramic heating shell with electric heating coils embedded within the shell. The temperature of a heating mantle is regulated with the heat controller. Although it is difficult to monitor the actual temperature of the heating mantle, the controller is calibrated so that it is fairly easy to duplicate approximate heating levels after one has gained some experience with this apparatus. Reactions or distillations requiring relatively high temperatures can be easily performed with a heating mantle. For temperatures in the range of 50–80°C, you should use a water bath (Section 2.3) or a steam bath (Section 2.8).

In the center of the heating mantle shown in Figure 2.1 is a well which can accomodate round-bottom flasks of several different sizes. Some heating mantles, however, are designed to fit only specific sizes of round-bottom flasks. Some heating mantles are also made to be used with a magnetic stirrer so that the reaction mixture can be heated and stirred at the same time. Figure 2.2 shows a reaction mixture being heated with a heating mantle.

Heating mantles are very easy to use and safe to operate. The metal housing is grounded to prevent electrical shock if liquid is spilled into the well; however, flammable liquids may ignite if spilled into the well of a hot heating mantle.

Caution: You should be very careful to avoid spilling liquids into the well of the heating mantle. The surface of the ceramic shell may be very hot and could cause the liquid to ignite.

Raising and lowering the apparatus is a much more rapid method of changing the temperature within the flask than changing the temperature with the controller. For this

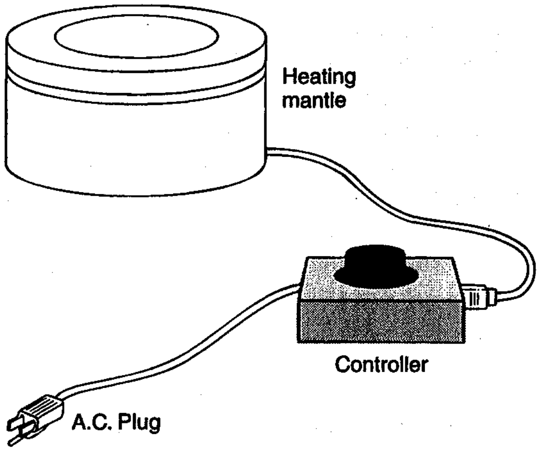

Figure 2.1 Heating mantle.

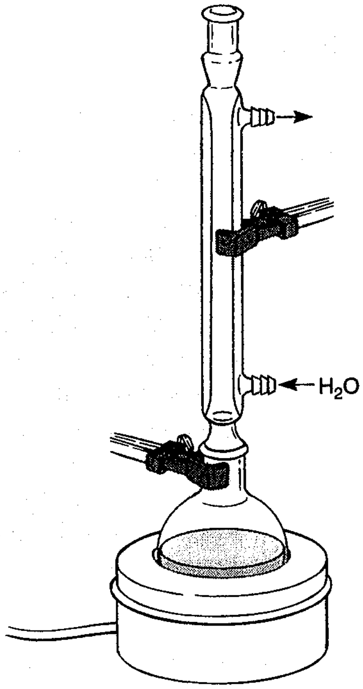

Figure 2.2 Heating with a heating mantle.

reason, the entire apparatus should be clamped above the heating mantle so that it can be raised quickly if overheating occurs. Some laboratories may provide a lab jack or blocks of wood which can be placed under the heating mantle. In this case, the heating mantle itself is lowered and the apparatus remains clamped in the same position.

There are two situations in which it is relatively easy to overheat the reaction mixture. The first situation occurs when a larger heating mantle is used to heat a relatively small flask. You should be very careful when doing this. Many laboratories provide heating mantles of different sizes to prevent this situation from happening. The second situation occurs when one is first bringing the reaction mixture to a boil. In order to bring the mixture to a boil as rapidly as possible, the heat controller is often turned up higher than it will need to be set in order to keep the mixture boiling. When the mixture begins boiling very rapidly, turn the controller to a lower setting and raise the apparatus until the mixture boils less rapidly. As the temperature of the heating mantle cools down, lower the apparatus until the flask is resting on the bottom of the well.

2.2 HOTPLATES

Hotplates are a very convenient source of heat; however, it is difficult to monitor the actual temperature, and changes in temperature occur somewhat slowly. Care must be

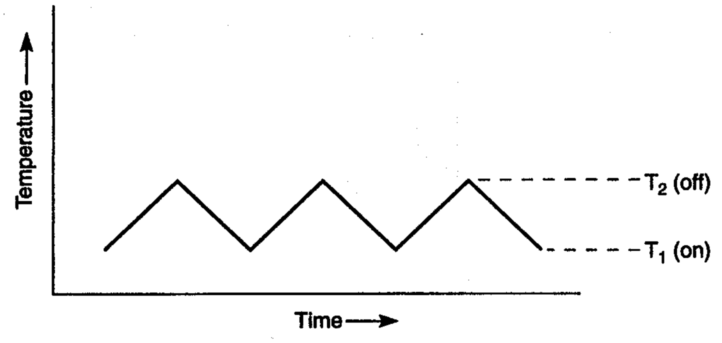

Figure 2.3 Temperature response for a hotplate with a thermostat.

taken with flammable solvents to ensure against fires caused by "flashing," when solvent vapors come into contact with the hotplate surface. One should never evaporate large quantities of a solvent by this method; the fire hazard is too large.

Some hotplates *heat constantly* at a given setting. They have no thermostat and you will have to control the temperature manually, either by removing the container being heated, or by adjusting the temperature up or down until a balance point is found. Some hotplates have a thermostat to control the temperature. A good thermostat will maintain a very even temperature. With many hotplates, however, the temperature may vary greatly (>10°–20°C), depending upon whether the heater is in its "on" cycle or its "off" cycle. These hotplates will have a cycling (or oscillating) temperature, as shown in Figure 2.3. They too will have to be adjusted continually to maintain even heat.

Some hotplates also have built-in magnetic stirring motors that enable the reaction mixture to be stirred and heated at the same time. Their use is described in Section 3.5, p. 621.

2.3 WATER BATH WITH HOTPLATE/STIRRER

A hot water bath is a very effective heat source when a temperature below 80°C is required. A beaker (250 mL or 400 mL) is partially filled with water and heated on a hot plate. A thermometer is clamped into position in the water bath. You may need to cover the water bath with aluminum foil to prevent evaporation, especially at higher temperatures. The water bath is illustrated in Figure 2.4. A mixture can be stirred with a magnetic stir bar (Technique 3, Section 3.5, p. 621). A hot water bath has some advantage over a heating mantle in that the temperature in the bath is uniform. In addition, it is sometimes easier to establish a lower temperature with a water bath than with other heating devices. Finally, the temperature of the reaction mixture will be closer to the temperature of the water, which allows for more precise control of the reaction conditions.

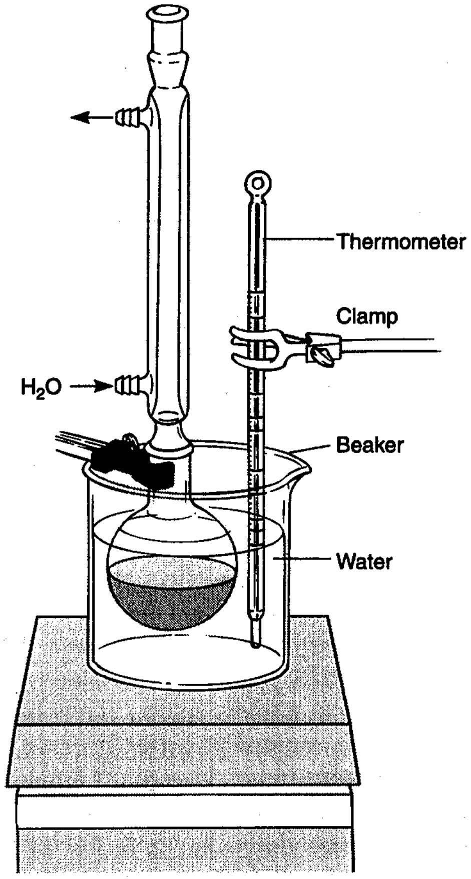

Figure 2.4 Water bath with hotplate/stirrer.

2.4 OIL BATH WITH HOTPLATE/STIRRER

In some laboratories oil baths may be available. An oil bath can be used when one is carrying out a distillation or is heating a reaction mixture that needs a temperature above 100°C. An oil bath can be heated most conveniently with a hotplate, and a *heavy-wall* beaker[1] provides a suitable container for the oil. A thermometer is clamped into position in the oil bath. In some laboratories the oil may be heated electrically by an immersion coil. Because oil baths have a high heat capacity and heat slowly, it is advisable to heat the oil bath partially before the actual time at which it is to be used.

An oil bath with ordinary mineral oil cannot be used above 200 to 220°C. Above this temperature the oil bath may "flash," or suddenly burst into flame. A hot oil fire is not extinguished easily. If the oil starts smoking, it may be near its flash temperature; discontinue heating. Old oil, which is dark, is more likely to flash than new oil is. Also, hot oil causes bad burns. Water should be kept away from a hot oil bath, since water in the oil will cause it to splatter. Never use an oil bath when it is obvious that there is water in the oil. If there should be water present, replace the oil before using the heating bath. An oil bath has only a finite lifetime. New oil is clear and colorless but, after extended use, becomes dark brown and gummy from oxidation.

---

[1] It is very dangerous to use a thin-wall beaker for an oil bath. Breakage due to heating can occur, spilling hot oil everywhere!

Besides ordinary mineral oil, a variety of other types of oils can be used in an oil bath. Silicone oil does not begin to decompose at as low a temperature as does mineral oil. When silicone oil is heated high enough to decompose, however, its vapors are far more hazardous than mineral oil vapors. The polyethylene glycols may be used in oil baths. They are water-soluble, which makes cleaning up after using an oil bath much easier than with mineral oil. One may select any one of a variety of polymer sizes of polyethylene glycol, depending on the temperature range required. The polymers of large molecular weight are often solid at room temperature. Wax may also be used for higher temperatures, but this material also becomes solid at room temperature. Some workers prefer to use a material that solidifies when not in use since it minimizes both storage and spillage problems. Vegetable shortening is occasionally used in heating baths.

2.5 ALUMINUM BLOCK WITH HOTPLATE/STIRRER

Although aluminum blocks are most commonly used in microscale organic laboratories, they can also be used with the smaller round-bottom flasks used in standard-scale experiments.[2] The aluminum block shown in Figure 2.5A can be used to hold 25-, 50-, or 100-mL round-bottom flasks, as well as a thermometer. Heating will occur more rapidly if the flask fits all the way into the hole; however, heating is also effective if the flask only partially fits into the hole. The aluminum block with smaller holes, as shown in Figure 2.5B, is designed for microscale glassware. It will hold a conical vial, a Craig tube or small test tubes, and a thermometer.

There are several advantages to heating with an aluminum block. The metal heats very quickly, high temperatures can be obtained, and you can cool the aluminum rapidly by removing it with crucible tongs and immersing it in cold water. Aluminum blocks are also inexpensive or can be fabricated readily in a machine shop.

Figure 2.6 shows a reaction mixture being heated with an aluminum block on a hotplate/stirrer unit. The thermometer in the figure is used to determine the temperature of the aluminum block. *Do not use a mercury thermometer:* use a thermometer containing a liquid other than mercury or use a metal dial thermometer that can be inserted into a smaller diameter hole drilled into the side of the block.[3] Make sure that the thermometer fits loosely in the hole, or it may break. Secure the thermometer with a clamp.

---

[2] The use of solid aluminum heating devices was developed by Siegfried Lodwig at Centralia College, Centralia, WA: Lodwig, S. N., *Journal of Chemical Education, 66* (1989): 77.

[3] Garner, C. M. "A Mercury-Free Alternative for Temperature Measurement in Aluminum Blocks." *Journal of Chemical Education, 68* (1991): A244.

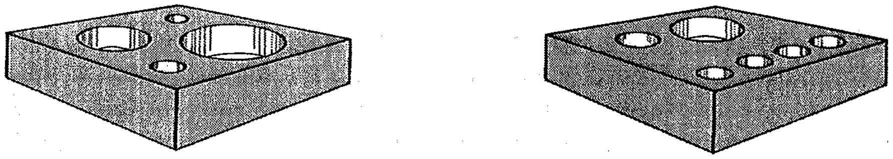

A. Large holes for 25-, 50- or 100-mL round-bottom flasks.

B. Small holes for Craig tube and 3-mL and 5-mL conical vials, and small test tubes.

Figure 2.5 Aluminum heating blocks.

As already mentioned, aluminum blocks are often used in the microscale organic laboratory. The use of an aluminum block to heat a microscale reflux apparatus is shown in Figure 2.7. The reaction vessel in the figure is a conical vial, which is used in many microscale experiments. Also shown in Figure 2.7 is a split aluminum collar that may be used when very high temperatures are required. The collar is split to facilitate easy placement around a 5-mL conical vial. The collar helps to distribute heat further up the wall of the vial.

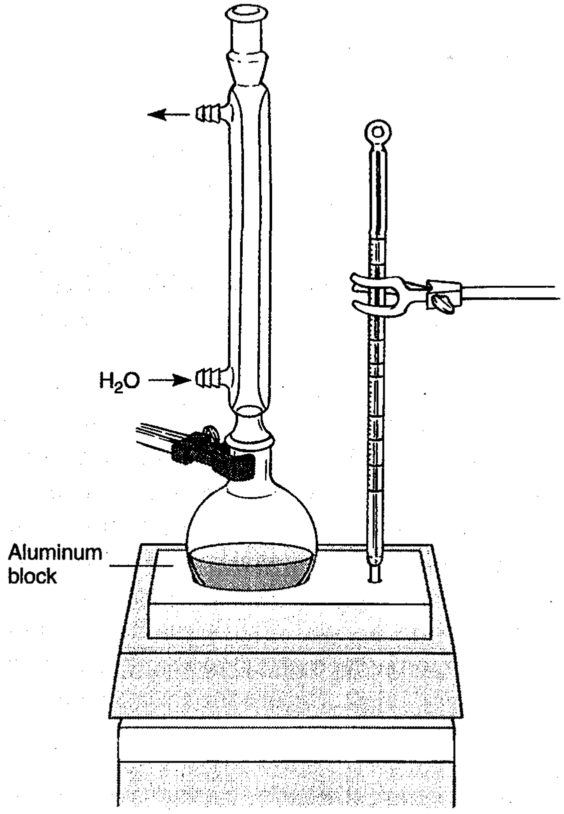

Figure 2.6 Heating with an aluminum block.

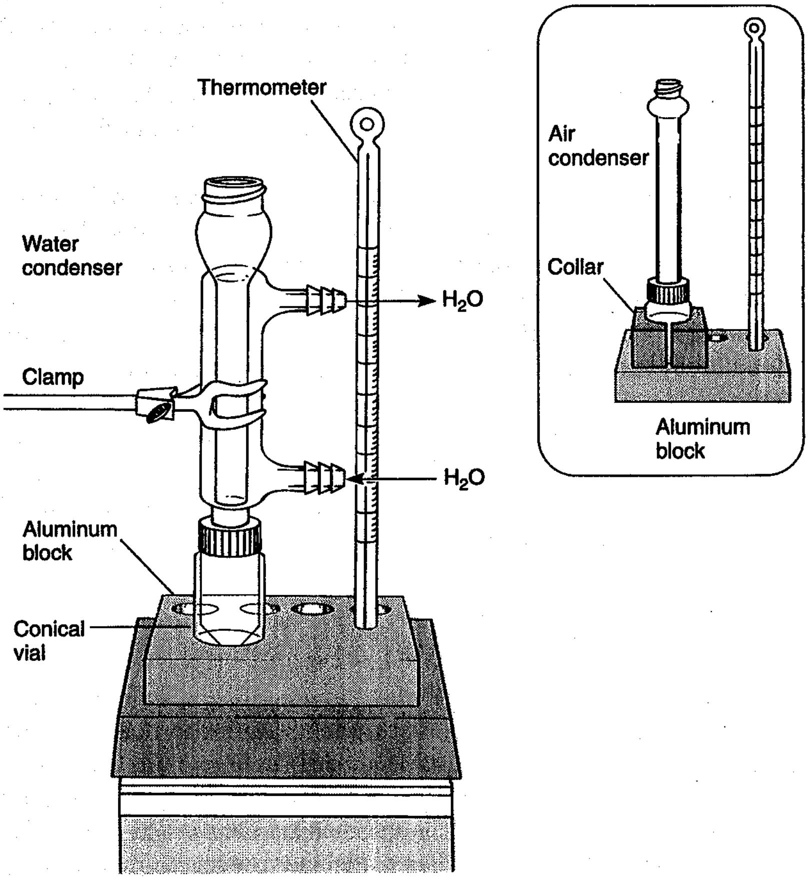

Figure 2.7 Heating with an aluminum block (microscale).

You should first calibrate the aluminum block so that you have an approximate idea where to set the control on the hotplate to achieve a desired temperature. Place the aluminum block on the hotplate and insert a thermometer into the small hole in the block. Select five equally spaced temperature settings, including the lowest and highest settings, on the heating control of the hotplate. Set the dial to the first of these settings and monitor the temperature recorded on the thermometer. When the thermometer reading arrives at a constant value,[4] record this final temperature, along with the dial setting. Repeat this procedure with the remaining four settings. Using these data, prepare a calibration curve for future reference.

It is a good idea to use the same hotplate each time, as it is very likely that two hotplates of the same type may give different temperatures with identical settings. Record the identification number printed on the unit that you are using in your notebook to ensure that you always use the same hotplate.

---

[4] See, however, Technique 2, Section 2.2, p. 602.

For many experiments, you can determine what the approximate setting on the hotplate should be from the boiling point of the liquid being heated. Because the temperature inside the flask is lower than the aluminum block temperature, you should add at least 20°C to the boiling point of the liquid and set the aluminum block at this higher temperature. In fact, you may need to raise the temperature even higher than this value in order to bring the liquid to a boil.

Many organic mixtures need to be stirred as well as heated to achieve satisfactory results. To stir a mixture, place a magnetic stir bar (Technique 3, Fig. 3.10A, p. 621) in a round-bottom flask containing the reaction mixture as shown in Figure 2.8A. If the mixture is to be heated as well as stirred, attach a water condenser as shown in Figure 2.6. With the combination stirrer/hotplate unit, it is possible to stir and heat a mixture simultaneously. Many reactions in this textbook are stirred continuously during the course of the reaction. With conical vials, a magnetic spin vane must be used to stir mixtures (Fig. 3.10B, p. 621). This is shown in Figure 2.8B. More uniform stirring will be obtained if the flask or vial is placed in the aluminum block so that it is centered on the hotplate. Mixing may also be achieved by boiling the mixture. A boiling stone (Section 3.6, p. 621) should be added when a mixture is boiled without magnetic stirring.

2.6 SAND BATH WITH HOTPLATE/STIRRER

The sand bath is used in some microscale laboratories to heat organic mixtures. It can also be used as a heat source in some standard-scale experiments. Sand provides a clean way of distributing heat to a reaction mixture. To prepare a sand bath, place about a 1-cm depth of sand in a crystallizing dish and then set the dish on a hotplate/stirrer unit. The apparatus is shown in Figure 2.9. Clamp the thermometer into position in the sand bath. You should calibrate the sand bath in a manner similar to that used with the aluminum block (see previous section). Because sand heats more slowly than an aluminum block, you will need to begin heating the sand bath well before using it.

Do not heat the sand bath much above 200°C or you may break the dish. If you need to heat at very high temperatures, you should use a heating mantle or an aluminum block rather than a sand bath. With sand baths, it may be necessary to cover the dish with alu-

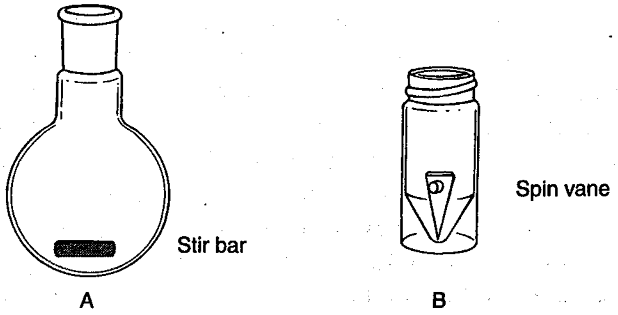

Figure 2.8 Methods of stirring in a round-bottom flask or conical vial.

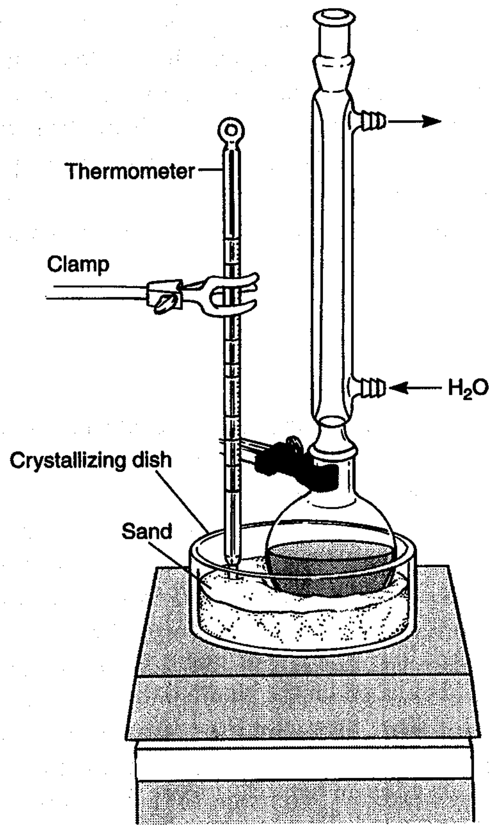

Figure 2.9 Heating with a sand bath.

minum foil to achieve a temperature near 200°C. Because of the relatively poor heat conductivity of sand, a temperature gradient is established within the sand bath. It is warmer near the bottom of the sand bath and cooler near the top for a given setting on the hotplate. To make use of this gradient, you may find it convenient to bury the flask or vial in the sand to heat a mixture more rapidly. Once the mixture is boiling, you can then slow the rate of heating by raising the flask or vial. These adjustments may be made easily and do not require a change in the setting on the hotplate.

2.7 FLAMES

The simplest technique for heating mixtures is to use a Bunsen burner. Because of the high danger of fires, however, the use of the Bunsen burner should be strictly limited to those cases for which the danger of fire is low or for which no reasonable alternative source of heat is available. A flame should generally be used only to heat aqueous solutions or solutions with very high boiling points. You should always check with your instructor about using a burner. If you use a burner at your bench, great care should be taken to ensure that others in the vicinity are not using flammable solvents.

In heating a flask with a Bunsen burner, you will find that using a wire gauze can produce more even heating over a broader area. The wire gauze, when placed under the object being heated, spreads the flame to keep the flask from being heated in one small area only.

Bunsen burners may be used to prepare capillary micropipets for thin-layer chromatography or to prepare other pieces of glassware requiring an open flame. For these purposes, burners should be used in designated areas in the laboratory and not at your laboratory bench.

2.8 STEAM BATHS

The steam cone or steam bath is a good source of heat when temperatures around 100°C are needed. Steam baths are used to heat reaction mixtures and solvents needed for crystallization. A steam cone and a portable steam bath are shown in Figure 2.10. These methods of heating have the disadvantage that water vapor may be introduced, through condensation of steam, into the mixture being heated. A slow flow of steam may minimize this difficulty.

Because water condenses in the steam line when it is not in use, it is necessary to purge the line of water before the steam will begin to flow. This purging should be accomplished before the flask is placed on the steam bath. The steam flow should be started with a high rate to purge the line; then, the flow should be reduced to the desired rate. When using a portable steam bath, be certain that condensate (water) is drained into a sink. Once the steam bath or cone is heated, a slow steam flow will maintain the temperature of the mixture being heated. There is no advantage to having a Vesuvius on your desk! An excessive steam flow may cause problems with condensation in the flask. This condensation problem can often be avoided by selecting the correct place at which to locate the flask on top of the steam bath.

The top of the steam bath consists of several flat concentric rings. The amount of heat delivered to the flask being heated can be controlled by selecting the correct sizes of

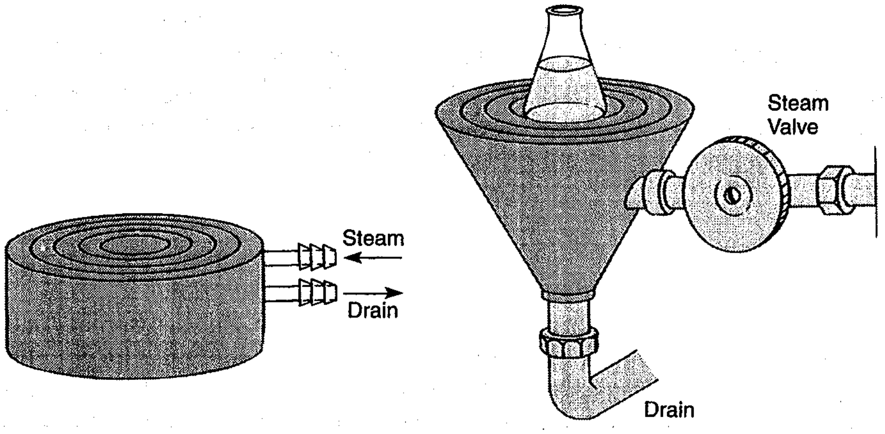

Figure 2.10 Steam bath and steam cone.

these rings. Heating is most efficient when the largest opening that will still support the flask is used. Heating large flasks on a steam bath while using the smallest opening leads to slow heating and wastes laboratory time.

2.9 COLD BATHS

At times, you may need to cool an Erlenmeyer flask or round-bottom flask below room temperature. A cold bath is used for this purpose. The most common cold bath is an **ice bath,** which is a highly convenient source of 0°C temperatures. An ice bath requires water along with ice to work well. If an ice bath is made up of only ice, it is not a very efficient cooler since the large pieces of ice do not make good contact with the flask. Enough water should be present with ice so that the flask is surrounded by water but not so much that the temperature is no longer maintained at 0°C. In addition, if too much water is present, the buoyancy of a flask resting in the ice bath may cause it to tip over. There should be enough ice in the bath to allow the flask to rest firmly.

For temperatures somewhat below 0°C, you may add some solid sodium chloride to the ice-water bath. The ionic salt lowers the freezing point of the ice, so that temperatures in the range of 0 to −10°C can be reached. The lowest temperatures are reached with ice-water mixtures that contain relatively little water.

A temperature of −78.5°C can be obtained with solid carbon dioxide or dry ice. However, large chunks of dry ice do not provide uniform contact with a flask being cooled. A liquid such as isopropyl alcohol is mixed with small pieces of dry ice to provide an efficient cooling mixture. Acetone and ethanol can be used in place of isopropyl alcohol. Be careful when handling dry ice because it can inflict severe frostbite. Extremely low temperatures can be obtained with liquid nitrogen (−195.8°C).

PROBLEMS

1. What would be the preferred heating device(s) in each of the following situations?
   (a) Reflux a solvent with a 56°C boiling point
   (b) Reflux a solvent with a 110°C boiling point
   (c) Distillation of a substance that boils at 220°C
2. Obtain the boiling points for the following compounds by using a handbook (Technique 20, Section 20.1, p. 861). In each case, suggest a heating device(s) that should be used for refluxing the substance.
   (a) Butyl benzoate
   (b) 1-Pentanol
   (c) 1-Chloropropane
3. What type of bath would you use to get a temperature of −10°C?
4. Obtain the melting point and boiling point for benzene and ammonia from a handbook (Technique 20, Section 20.1 p. 861) and answer the following questions.
   (a) A reaction was conducted in benzene as the solvent. Because the reaction was very exothermic, the mixture was cooled in a salt-ice bath. This was a bad choice. Why?
   (b) What bath should be used for a reaction that is conducted in liquid ammonia as the solvent?

TECHNIQUE 3

Reaction Methods

The successful completion of an organic reaction requires the chemist to be familiar with a variety of laboratory methods. These methods include operating safely, choosing and handling solvents correctly, heating reaction mixtures, adding liquid reagents, maintaining anhydrous conditions in the reaction, and collecting gaseous products. Several techniques that are used in bringing a reaction to a successful conclusion are treated in this chapter.

3.1 SOLVENTS

Organic solvents must be handled safely. Always remember that organic solvents are all at least mildly toxic and that many are flammable. You should become thoroughly familiar with the introductory chapter, "Laboratory Safety."

Read "Laboratory Safety," pages 5–22.

TABLE 3.1. Common Organic Solvents

Solvent	bp (°C)	Solvent	bp (°C)
Hydrocarbons		Ethers	
Pentane	36	**Ether** (Diethyl)	35
Hexane	69	**Dioxane***	101
Benzene*	80	**1,2-Dimethoxyethane**	83
Toluene	111	Others	
Hydrocarbon Mixtures		Acetic acid	118
Petroleum ether	30–60	Acetic anhydride	140
Ligroin	60–90	**Pyridine**	115
Chlorocarbons		**Acetone**	56
Methylene chloride	40	**Ethyl acetate**	77
Chloroform*	61	Dimethylformamide	153
Carbon tetrachloride*	77	Dimethylsulfoxide	189
Alcohols			
Methanol	65		
Ethanol	78		
Isopropyl alcohol	82		

NOTE: Boldface type indicates flammability.

*Suspect carcinogen (see p. 21).

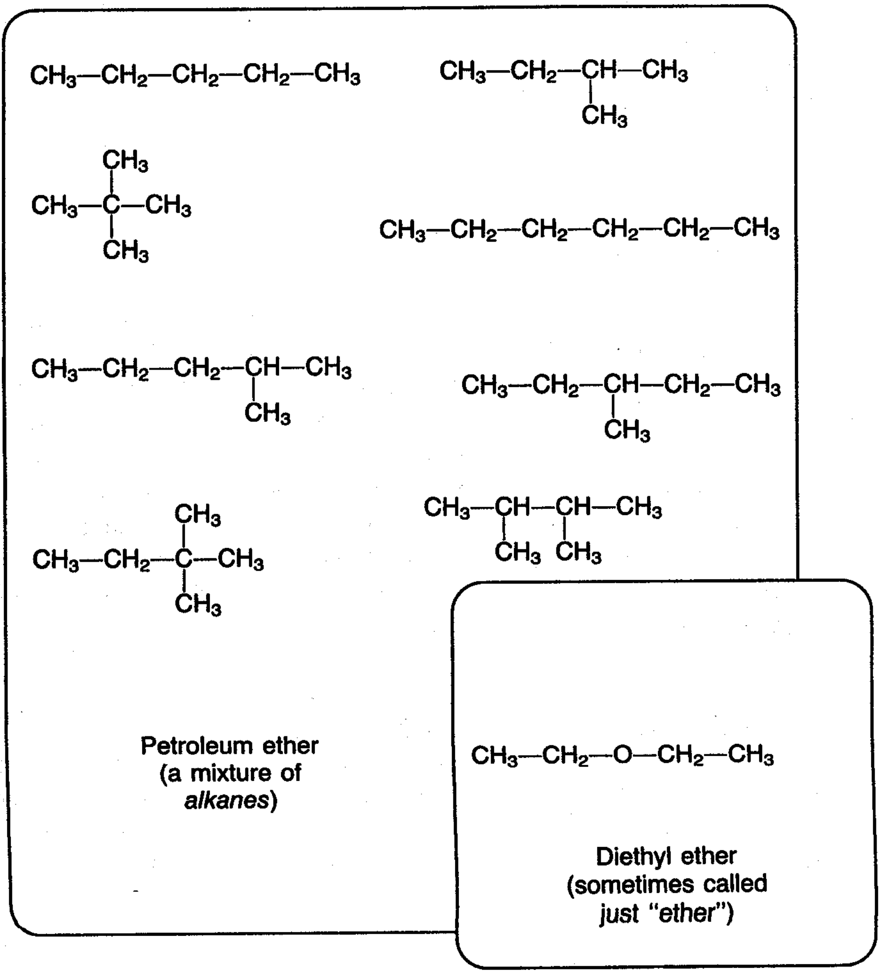

Figure 3.1 A comparison between "ether" (diethyl ether) and "petroleum ether."

The most common organic solvents are listed in Table 3.1 along with their boiling points. Solvents marked in boldface type will burn. Ether, pentane, and hexane are especially dangerous; if they are combined with the correct amount of air they will explode.

The terms **petroleum ether** and **ligroin** are often confusing. Petroleum ether is a mixture of hydrocarbons with isomers of formulas C_5H_{12} and C_6H_{14} predominating. Petroleum ether is not an ether at all because there are no oxygen-bearing compounds in the mixture. In organic chemistry, an ether is usually a compound containing an oxygen atom to which two alkyl groups are attached. Figure 3.1 shows some of the hydrocarbons that appear commonly in petroleum ether. Use special care when instructions call for either **ether** or **petroleum ether;** the two must not become accidentally confused. Confusion is particularly easy when one is selecting a container of solvent from the supply shelf.

Ligroin, or high-boiling petroleum ether, is like petroleum ether in composition except that compared with petroleum ether, ligroin generally includes higher-boiling alkane isomers. Depending on the supplier, ligroin may have different boiling ranges. While some brands of ligroin have boiling points ranging from about 60°C to about 90°C, other brands have boiling points ranging from about 60°C to about 75°C. The boiling point ranges of petroleum ether and ligroin are often included on the labels of the containers.

3.2 GROUND-GLASS JOINTS

The glassware in your organic kit has **standard-taper ground-glass joints.** For example, the Claisen head in Figure 3.2 consists of an inner (male) ground-glass joint at the bottom and two outer (female) joints at the top. Each end is ground to a precise size which is designated by the symbol Ŧ followed by two numbers. A common joint size in standard-scale glassware is Ŧ 19/22. The first number indicates the diameter (in millimeters) of the joint at its widest point, and the second number refers to its length (see Fig. 3.2). One advantage of standard-taper joints is that the pieces fit together snugly and form a good seal. In addition, standard-taper joints allow all glassware components with the same joint size to be connected, thus permitting the assembly of a wide variety of apparatus. One disadvantage of glassware with ground-glass joints, however, is that it is very expensive.

It is important to make sure no solid is on the joint surfaces. Such material will lessen the efficiency of the seal, and the joints may leak. Also, if the apparatus is to be heated, material caught between the joint surfaces will increase the tendency for the joints to stick. If the joint surfaces are coated with liquid or adhering solid, you should wipe them with a cloth or lint-free paper towel before assembling. It is sometimes advisable to lubricate joints lightly with stopcock grease before assembling them (see p. 34).

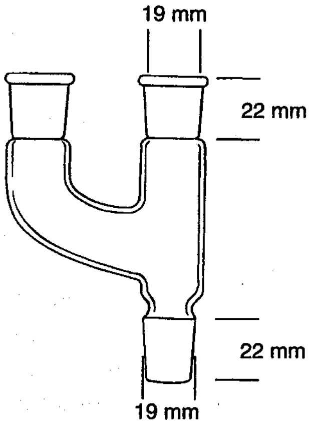

Figure 3.2 Illustration of Ŧ 19/22 inner and outer joints showing dimensions.

The most important thing you can do to prevent ground-glass joints from becoming "frozen" or stuck together is to disassemble the glassware as soon as possible after a procedure is completed. Even when this precaution is followed, ground-glass joints may become stuck tightly together. The same is true of glass stoppers in bottles or flasks. If the pieces do not separate easily, you must be careful when you try to pull them apart. The best way is to hold the two pieces, with both hands touching, as close as possible to the joint. With a firm grasp, try to loosen the joint with a slight twisting motion (do not twist very hard). If this does not work, try to pull your hands apart without pushing sideways on the glassware. If it is not possible to pull the pieces apart, see page 34 for other options.

3.3 ASSEMBLING THE APPARATUS

Care must be taken when assembling the glass components into the desired apparatus. You should always remember that Newtonian physics applies to chemical apparatus, and unsecured pieces of glassware are certain to respond to gravity.

Assembling an apparatus in the correct manner requires that the individual pieces of glassware are connected to each other securely and the entire apparatus is held in the correct position. This can be accomplished by using **adjustable metal clamps** or a combination of adjustable metal clamps and **plastic joint clips.**

Two types of adjustable metal clamps are shown in Figure 3.3. Although these two types of clamps can usually be interchanged, the extension clamp is more commonly used to hold round-bottom flasks in place and the three-finger clamp is frequently used to clamp condensers. Both types of clamps must be attached to a ring stand using a clamp holder, shown in Figure 3.3C.

Securing Standard-Scale Apparatus Assemblies

It is possible to assemble an apparatus using only adjustable metal clamps. An apparatus used to perform a distillation is shown in Figure 3.4. It is held together securely

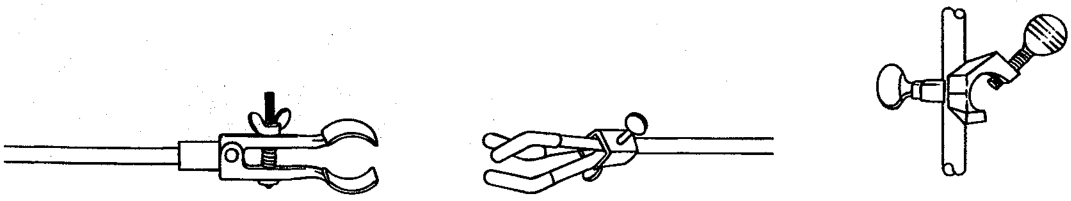

A. Extension clamp B. Three finger clamp C. Clamp holder

Figure 3.3 Adjustable metal clamps.

with three metal clamps. Because of the size of the apparatus and its geometry, the various clamps would likely be attached to three different ring stands. This apparatus would be somewhat difficult to assemble, because one must ensure that the individual pieces stay together while securing and adjusting the clamps required to hold the entire apparatus in place. In addition, one must be very careful not to bump any part of the apparatus or the ring stands after the apparatus is assembled.

A more convenient alternative is to use a combination of metal clamps and plastic joint clips. A plastic joint clip is shown in Figure 3.5A. These clips are very easy to use (they just "clip" on), will withstand temperatures up to 140°C, and are quite durable. They hold together two pieces of glassware that are connected by ground glass joints, as shown in Figure 3.5B. These clips come in different sizes to fit ground glass joints of different sizes and they are color-coded for each size.

When used in combination with metal clamps, the plastic joint clips make it much easier to assemble most apparatus in a secure manner. There is less chance of dropping the glassware while assembling the apparatus, and once the apparatus is set up it is more secure. Figure 3.6 shows the same distillation apparatus held in place with both adjustable metal clamps and plastic joint clips.

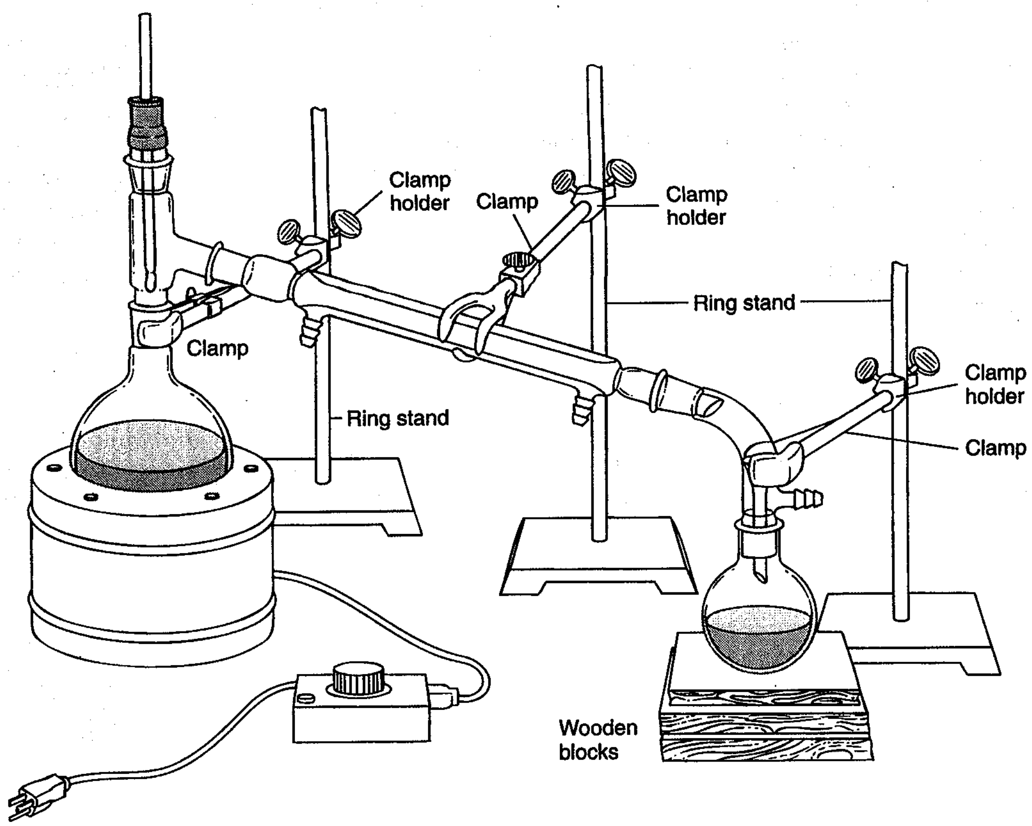

Figure 3.4 Distillation apparatus secured with metal clamps.

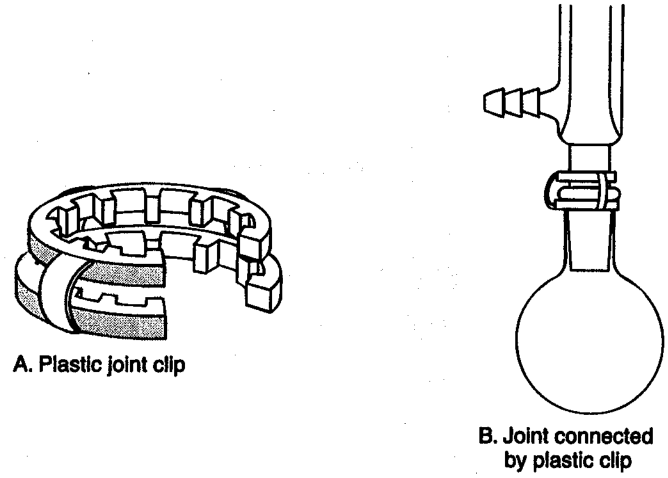

A. Plastic joint clip

B. Joint connected by plastic clip

Figure 3.5 Plastic joint clip.

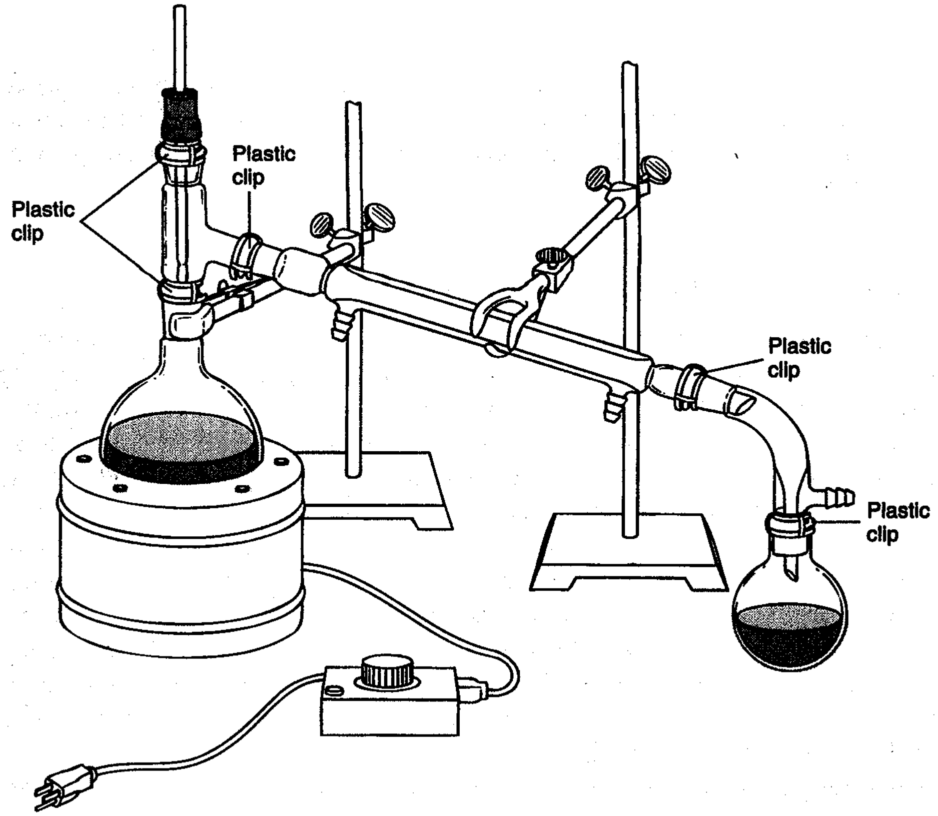
Plastic clip
Plastic clip
Plastic clip
Plastic clip

Figure 3.6 Distillation apparatus secured with metal clamps and plastic joint clips.

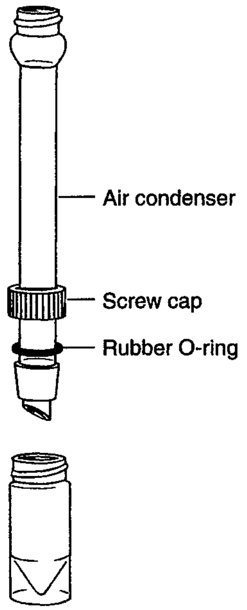

Figure 3.7 A microscale standard-taper joint assembly.

To assemble this apparatus, one would first connect all of the individual pieces together using the plastic clips. The entire apparatus is then connected to the ring stands using the adjustable metal clamps. Note that only two ring stands are required and the wooden blocks are not needed.

Securing Microscale Apparatus Assemblies

The glassware in most microscale kits is made with standard-taper ground joints. The most common joint size is ₸ 14/10. Some microscale glassware with ground-glass joints also has threads cast into the outside surface of the outer joints (see top of air condenser in Figure 3.7). The threaded joint allows the use of a plastic screw cap with a hole in the top to fasten two pieces of glassware together securely. The plastic cap is slipped over the inner joint of the upper piece of glassware, followed by a rubber O-ring (see Figure 3.7). The O-ring should be pushed down so that it fits snugly on top of the ground-glass joint. The inner ground-glass joint is then fitted into the outer joint of the bottom piece of glassware. The screw cap is tightened, without excessive force, to attach the entire apparatus firmly together. The O-ring provides an additional seal that makes this joint air-tight. With this connecting system, it is unnecessary to use any type of grease to seal the joint. The O-ring *must be used* to obtain a good seal and to lessen the chances of breaking the glassware when you tighten the plastic cap.

Microscale glassware connected together in this fashion can be assembled very easily. The entire apparatus is held together securely, and usually only one metal clamp is required to hold the apparatus onto a ring stand.

3.4 HEATING UNDER REFLUX

Often we wish to heat a mixture for a long time and to leave it untended. A **reflux apparatus** (see Fig. 3.8) allows such heating. It also keeps the solvent from being lost by evaporation. A condenser is attached to the boiling flask.

Condenser. The **water-jacketed condenser** shown in the figure consists of two concentric tubes with the outer cooling tube sealed onto the inner tube. The vapors rise within the inner tube and water circulates through the outer tube. The circulating water removes heat from the vapors and condenses them. The figure also shows a typical microscale apparatus for heating small quantities of material under reflux (Fig. 3.8B).

When using a water-jacketed condenser, the direction of the water flow should be such that the condenser will fill with cooling water. The water should enter the bottom of the condenser and leave from the top. The water should flow fast enough to withstand any changes in pressure in the water lines, but it should not flow any faster than absolutely necessary. An excessive flow rate greatly increases the chance of a flood, and high water pressure may force the hose from the condenser. Cooling water should be flowing before heating is begun! If the water is to remain flowing overnight, it is advisable to fasten the

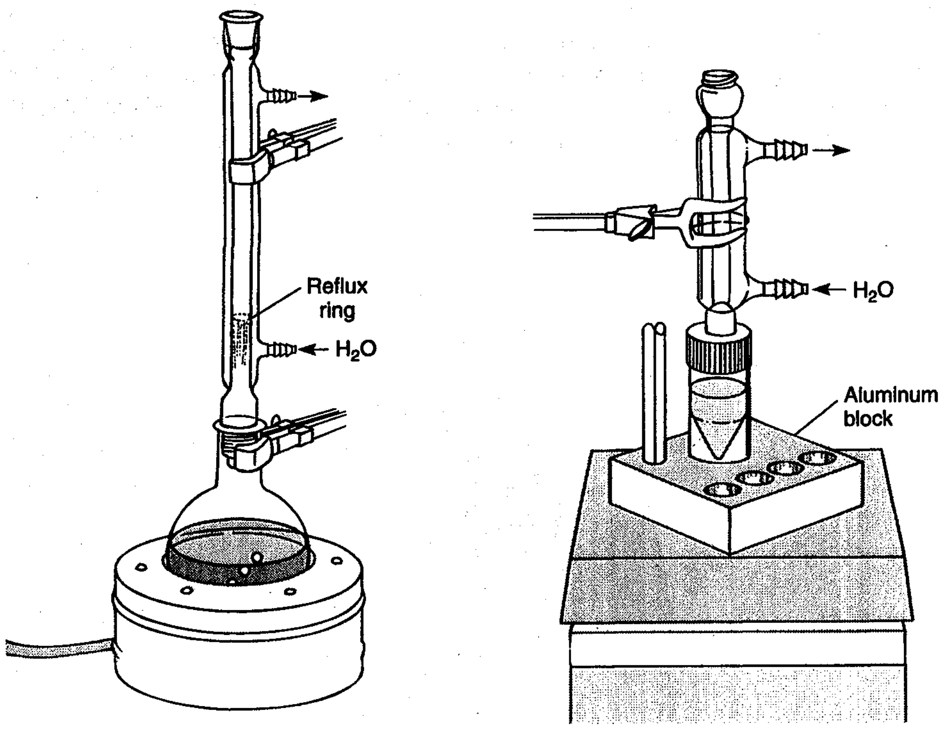

A. Reflux apparatus for standard-scale reactions, using a heating mantle and water-jacketed condenser.

B. Reflux apparatus for microscale reactions, using a hotplate, aluminum block, and water-jacketed condenser.

Figure 3.8 Heating under reflux.

rubber tubing securely with wire to the condenser. If a flame is used as a source of heat (see Cautions, p. 609), it is wise to use a wire gauze beneath the flask to provide an even distribution of heat from the flame. In most cases, a heating mantle, water bath, oil bath, aluminum block, sand bath, or steam bath is preferred over a flame.

Stirring. When heating a solution, always use a magnetic stirrer or a boiling stone (see Sections 3.5 and 3.6) to keep the solution from "bumping" (see next section).

Rate of Heating. If the heating rate has been correctly adjusted, the liquid being heated under reflux will travel only part way up the condenser tube before condensing. Below the condensation point, solvent will be seen running back into the flask; above it, the interior of the condenser will appear dry. The boundary between the two zones will be clearly demarcated, and a **reflux ring** or a ring of liquid will appear there. The reflux ring can be seen in Figure 3.8A. In heating under reflux, the rate of heating should be adjusted so that the reflux ring is no higher than a third to a half the distance to the top of the condenser. With microscale experiments, the quantities of vapor rising in the condenser frequently are so small that a clear reflux ring cannot be seen. In those cases, the heating rate must be adjusted so that the liquid boils smoothly but not so rapidly that solvent can escape the condenser. With such small volumes, the loss of even a small amount of solvent can affect the reaction. With standard-scale reactions, the reflux ring is much easier to see, and one can adjust the heating rate more easily.

Tended Reflux. It is possible to heat small amounts of a solvent under reflux in an Erlenmeyer flask. By heating gently, the evaporated solvent will condense in the relatively cold neck of the flask and return to the solution. This technique (see Fig. 3.9) requires constant attention. The flask must be swirled frequently and removed from the heating source for a short period if the boiling becomes too vigorous. When heating is in progress, the reflux ring should not be allowed to rise into the neck of the flask.

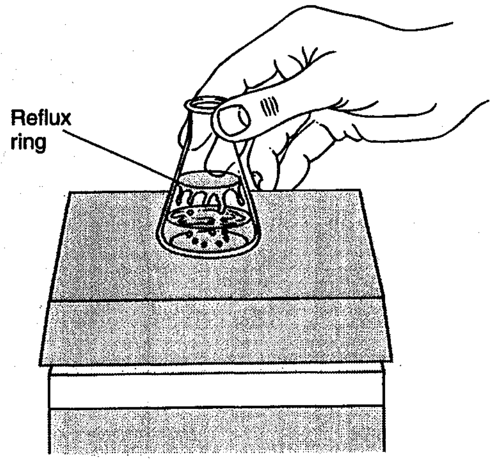

Figure 3.9 Tended reflux of small quantities on a hotplate.

3.5 STIRRING METHODS

When a solution is heated, there is a danger that it may become superheated. When this happens, very large bubbles sometimes erupt violently from the solution; this is called **bumping.** Bumping must be avoided because it brings with it the risk that material may be lost from the apparatus, that a fire might start, or that the apparatus may break.

Magnetic stirrers are used to prevent bumping because they produce turbulence in the solution. The turbulence breaks up the large bubbles that form in boiling solutions. An additional purpose for using a magnetic stirrer is to stir the reaction to ensure that all the reagents are thoroughly mixed. A magnetic stirring system consists of a magnet that is rotated by an electric motor. The rate at which this magnet rotates can be adjusted by a potentiometric control. A small magnet, which is coated with a nonreactive material such as Teflon or glass, is placed in the flask. The magnet within the flask rotates in response to the rotating magnetic field caused by the motor-driven magnet. The result is that the inner magnet stirs the solution as it rotates. A very common type of magnetic stirrer includes the stirring system within a hotplate. This type of hotplate-stirrer permits one to heat the reaction and stir it simultaneously.

Magnetic stirring bars are available in several sizes and shapes. For standard-scale apparatus, magnetic stirring bars of various sizes and shapes are available. For microscale apparatus, a **magnetic spin vane** is often used. It is designed to contain a tiny bar magnet and to have a shape that conforms to the conical bottom of a reaction vial. A small Teflon-coated magnetic stirring bar works well with very small round-bottom boiling flasks. Small stirring bars of this type (often sold as "disposable" stirring bars) can be obtained very cheaply. A variety of magnetic stirring bars is illustrated in Figure 3.10.

There is also a variety of simple techniques that may be used to stir a liquid mixture in a centrifuge tube or conical vial. A thorough mixing of the components of a liquid can be achieved by repeatedly drawing the liquid into a Pasteur pipet and then ejecting the liquid back into the container by pressing sharply on the dropper bulb. Liquids can also be stirred effectively by placing the flattened end of a spatula into the container and twirling it rapidly.

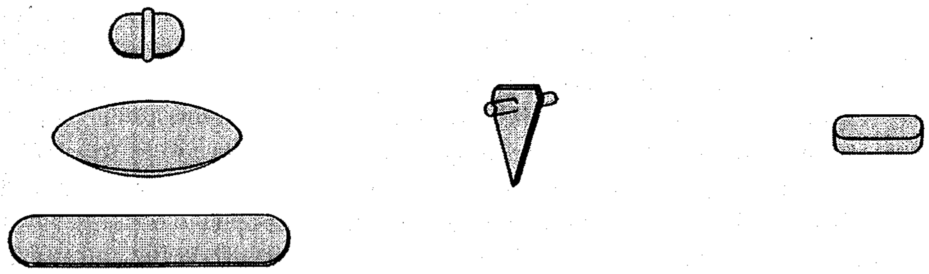

A. Standard-sized magnetic stirring bars.

B. Microscale magnetic spin vane.

C. Small magnetic stirring bar ("disposable" type).

Figure 3.10 Magnetic stirring bars.

3.6 BOILING STONES

A **boiling stone,** also known as a **boiling chip** or **Boileezer,** is a small lump of porous material that produces a steady stream of fine air bubbles when it is heated in a solvent. This stream of bubbles and the turbulence that accompanies it breaks up the large bubbles of gases in the liquid. In this way, it reduces the tendency of the liquid to become superheated, and it promotes the smooth boiling of the liquid. The boiling stone decreases the chances for bumping.

Two common types of boiling stones are carborundum and marble chip. Carborundum boiling stones are more inert and the pieces are usually quite small, suitable for most small-scale applications. If available, carborundum boiling stones are preferred for most purposes. Marble chips may dissolve in strong acid solutions and the pieces are larger. The advantage of marble chips is that they are cheaper.

Because boiling stones act to promote the smooth boiling of liquids, you should always make certain that a boiling stone has been placed in a liquid *before* heating is begun. If you wait until the liquid is hot, it may have become superheated. Adding a boiling stone to a superheated liquid will cause all the liquid to try to boil at once. The liquid, as a result, would erupt entirely out of the flask or froth violently.

As soon as boiling ceases in a liquid containing a boiling stone, the liquid is drawn into the pores of the boiling stone. When this happens, the boiling stone no longer can produce a fine stream of bubbles; it is spent. You may have to add a new boiling stone if you have allowed boiling to stop for a long period.

Wooden applicator sticks are used in some applications. They function in the same manner as boiling stones. Occasionally, glass beads are used. Their presence also causes sufficient turbulence in the liquid to prevent bumping.

3.7 ADDITION OF LIQUID REAGENTS

Liquid reagents and solutions are added to a reaction by several means, some of which are shown in Figure 3.11. The most common type of assembly for standard-scale experiments is shown in Figure 3.11A. In this apparatus, a separatory funnel is attached to the sidearm of a Claisen head adapter. The separatory funnel must be equipped with a standard-taper, ground-glass joint to be used in this manner. The liquid is stored in the separatory funnel (which is called an **addition funnel** in this application) and is added to the reaction. The rate of addition is controlled by adjusting the stopcock. When it is being used as an addition funnel, the upper opening must be kept open to the atmosphere. If the upper hole is stoppered, a vacuum will develop in the funnel and will prevent the liquid from passing into the reaction vessel. Because the funnel is open to the atmosphere, there is a danger that atmospheric moisture can contaminate the liquid reagent as it is being added. To prevent this outcome, a drying tube (see Section 3.8) may be attached to the upper opening of the addition funnel. The drying tube allows the funnel to maintain atmospheric pressure without allowing the passage of water vapor into the reaction. For

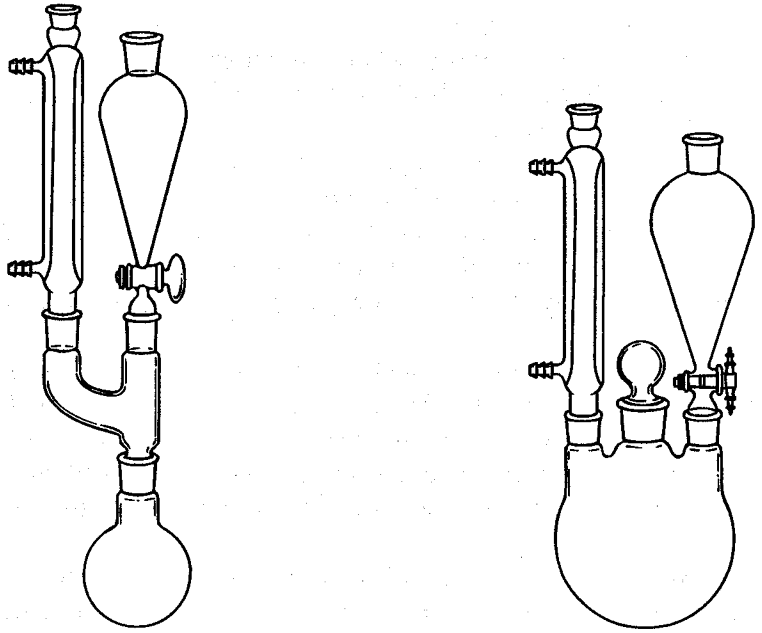

A. Standard-scale equipment, using a separatory funnel as an addition funnel.

B. Standard-scale, for larger amounts.

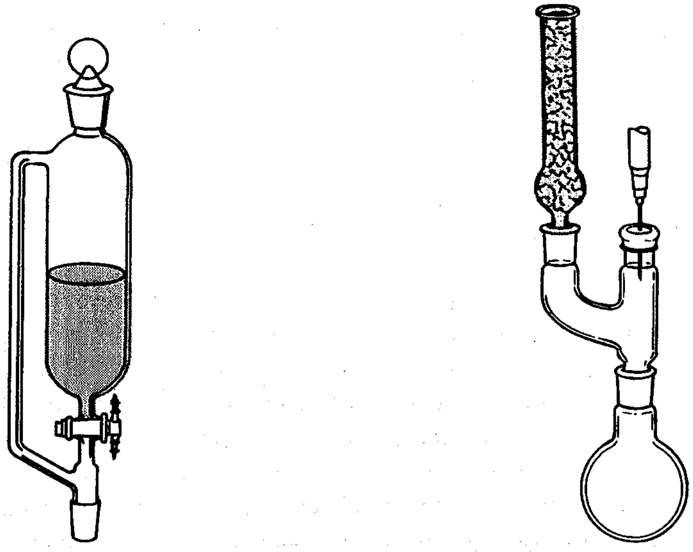

C. A pressure equalizing addition funnel.

D. Addition with a hypodermic syringe inserted through a rubber septum.

Figure 3.11 Methods for adding liquid reagents to a reaction.

reactions that are particularly sensitive to moisture, it is also advisable to attach a second drying tube to the top of the condenser.

Another standard-scale assembly, suitable for larger amounts of material, is shown in Figure 3.11B. Drying tubes may also be used with this apparatus to prevent contamination from atmospheric moisture.

Figure 3.11C shows an alternative type of addition funnel that is useful for reactions that must be maintained under an atmosphere of inert gas. This is the **pressure-equalizing addition funnel.** With this glassware, the upper opening is stoppered. The sidearm allows the pressure above the liquid in the funnel to be in equilibrium with the pressure in the rest of the apparatus, and it allows the inert gas to flow over the top of the liquid as it is being added.

With either type of standard-scale addition funnel, you can control the rate of addition of the liquid by carefully adjusting the stopcock. Even after careful adjustment, changes in pressure can occur, causing the flow rate to change. In some cases, the stopcock can become clogged. It is important, therefore, to monitor the addition rate carefully and to refine the adjustment of the stopcock as needed to maintain the desired rate of addition.

A fourth method, shown in Figure 3.11D, is suitable for use in microscale and some standard-scale experiments where the reaction should be kept isolated from the atmosphere. In this approach, the liquid is kept in a hypodermic syringe. The syringe needle is inserted through a rubber septum, and the liquid is added dropwise from the syringe. The septum seals the apparatus from the atmosphere, which makes this technique useful for reactions that are conducted under an atmosphere of inert gas or where anhydrous conditions must be maintained. The drying tube is used to protect the reaction mixture from atmospheric moisture.

3.8 DRYING TUBES

With certain reactions, atmospheric moisture must be prevented from entering the reaction vessel. A **drying tube** can be used to maintain anhydrous conditions within the apparatus. Two types of drying tubes are shown in Figure 3.12. The typical drying tube is prepared by placing a small, loose plug of glass wool or cotton into the constriction at the end of the tube nearest the ground-glass joint or hose connection. The plug is tamped gently with a glass rod or piece of wire to place it in the correct position. A drying agent, typically calcium sulfate ("Drierite") or calcium chloride (see Technique 7, Section 7.8, p. 696), is poured on top of the plug to the approximate depth shown in Figure 3.12. Another loose plug of glass wool or cotton is placed on top of the drying agent to prevent the solid material from falling out of the drying tube. The drying tube is then attached to the flask or condenser.

Air that enters the apparatus must pass through the drying tube. The drying agent absorbs any moisture from air passing through it, so that air entering the reaction vessel has had the water vapor removed from it.

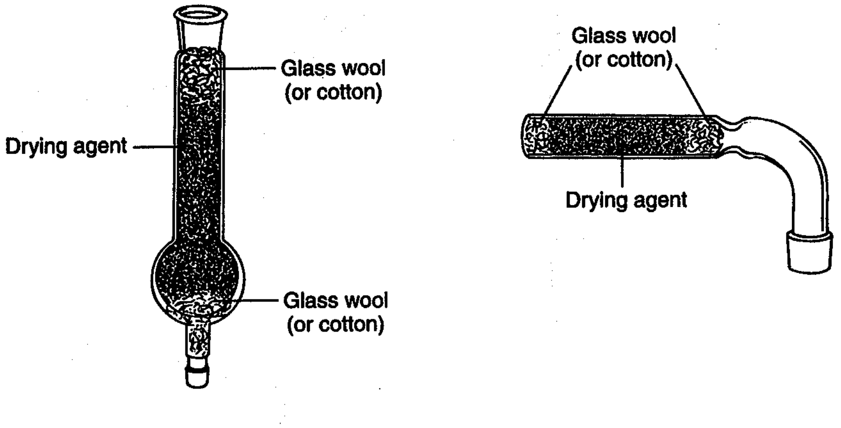

A. Standard-sized drying tube.

B. Microscale drying tube.

Figure 3.12 Drying tubes.

3.9 CAPTURING NOXIOUS GASES

Many organic reactions involve the production of a noxious gaseous product. The gas may be corrosive, such as hydrogen chloride, hydrogen bromide, or sulfur dioxide; or it may be toxic, such as carbon monoxide. The safest way to avoid exposure to these gases is to conduct the reaction in a ventilated hood where the gases can be safely drawn away by the ventilation system.

In many instances, however, it is quite safe and efficient to conduct the experiment on the laboratory bench, away from the hood. This is particularly true when the gases are soluble in water. Some techniques for capturing noxious gases are presented in this section.

A. EXTERNAL GAS TRAPS

One approach to capturing gases is to prepare a trap that is separate from the reaction apparatus. The gases are carried from the reaction to the trap by means of tubing. There are several variations on this type of trap. With standard-scale reactions, a trap using an inverted funnel placed in a beaker of water is used. A piece of glass tubing, inserted through a thermometer adapter attached to the reaction apparatus, is connected to flexible tubing. The tubing is attached to a conical funnel. The funnel is clamped in place inverted over a beaker of water. The funnel is clamped so that its lip *almost touches* the water surface, but is not placed below the surface of the water. With this arrangement, water cannot be sucked back into the reaction if the pressure in the reaction vessel changes suddenly. This type of trap can also be used in microscale applications. An example of the inverted funnel type of gas trap is shown in Figure 3.13.

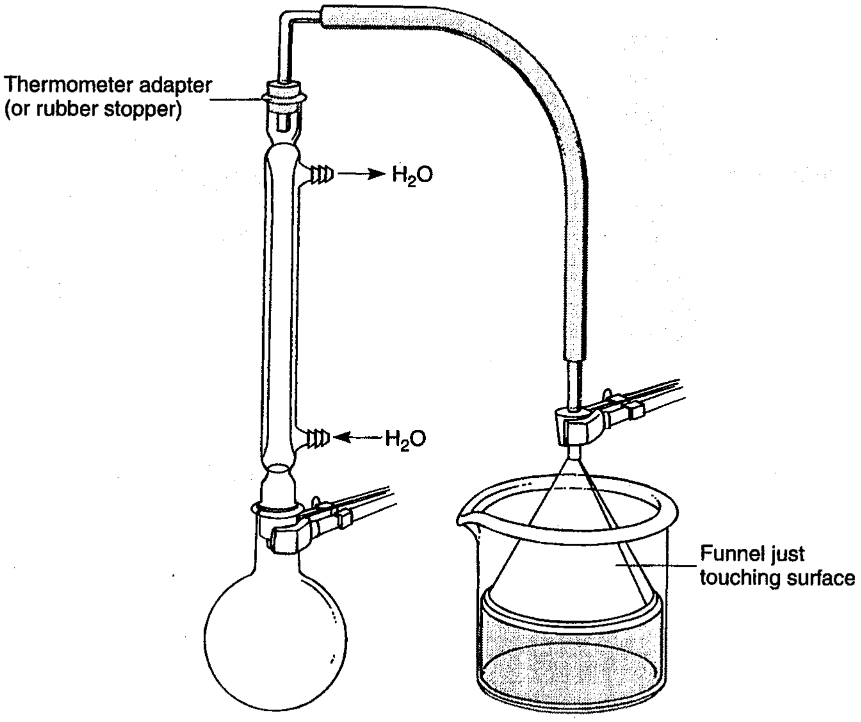

Figure 3.13 Inverted-funnel gas trap.

One method that works well for standard-scale and microscale experiments is to place a thermometer adapter into the opening in the reaction apparatus. A Pasteur pipet is inserted upside-down through the adapter, and a piece of flexible tubing is fitted over the narrow tip. It might be helpful to break the Pasteur pipet before using it for this purpose, so that only the narrow tip and a short section of the barrel is used. The other end of the flexible tubing is placed through a large plug of moistened glass wool in a test tube. The water in the glass wool absorbs the water-soluble gases. This method is shown in Figure 3.14.

B. DRYING-TUBE METHOD

Some standard-scale and most microscale experiments have the advantage that the amounts of gases produced are very small. Hence, it is easy to trap them and prevent them from escaping into the laboratory room. You can take advantage of the water solubility of corrosive gases such as hydrogen chloride, hydrogen bromide, and sulfur dioxide. A simple technique is to attach the drying tube (see Fig. 3.12) to the top of the reaction flask or condenser. The drying tube is filled with moistened glass wool. The moisture in the glass wool absorbs the gas, preventing its escape. To prepare this type of gas trap, fill the drying tube with glass wool and then add water dropwise to the glass wool until it has

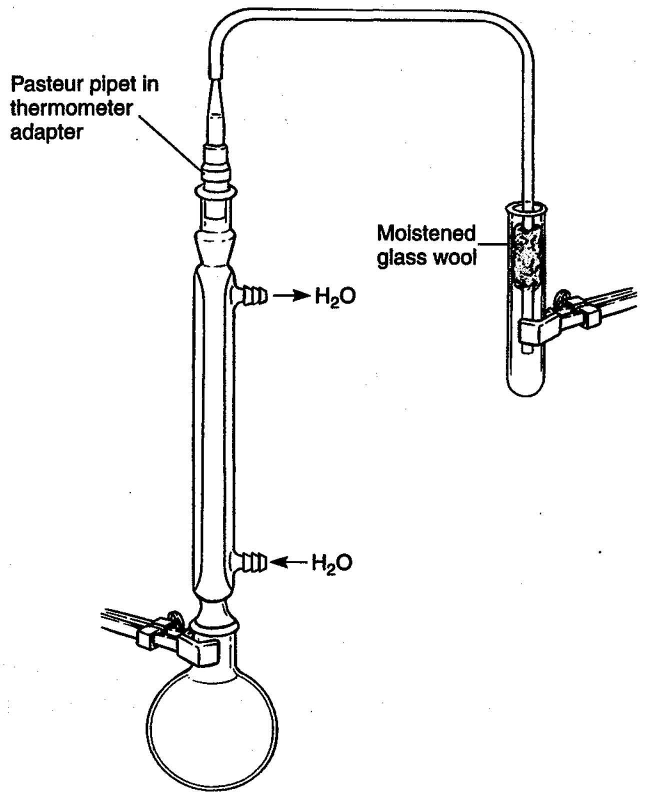

Figure 3.14 External gas trap.

been moistened to the desired degree. Moistened cotton can also be used, although cotton will absorb so much water that it is easy to plug the drying tube.

When using glass wool in a drying tube, moisture from the glass wool must not be allowed to drain from the drying tube into the reaction. It is best to use a drying tube that has a constriction between the part where the glass wool is placed and the neck, where the joint is attached (see Fig. 3.12B). The constriction acts as a partial barrier preventing the water from leaking into the neck of the drying tube. Make certain not to make the glass wool too moist. When it is necessary to use the drying tube shown in Figure 3.12A as a gas trap and it is essential that water not be allowed to enter the reaction flask, the modification shown in Figure 3.15 should be used. The rubber tubing between the thermometer adapter and the drying tube should be heavy enough to prevent crimping.

C. REMOVAL OF NOXIOUS GASES USING AN ASPIRATOR

An aspirator can be used to remove noxious gases from the reaction. The simplest approach is to clamp a disposable Pasteur pipet so that its tip is placed well into the condenser atop the reaction flask. An inverted funnel clamped over the apparatus can also be used. The

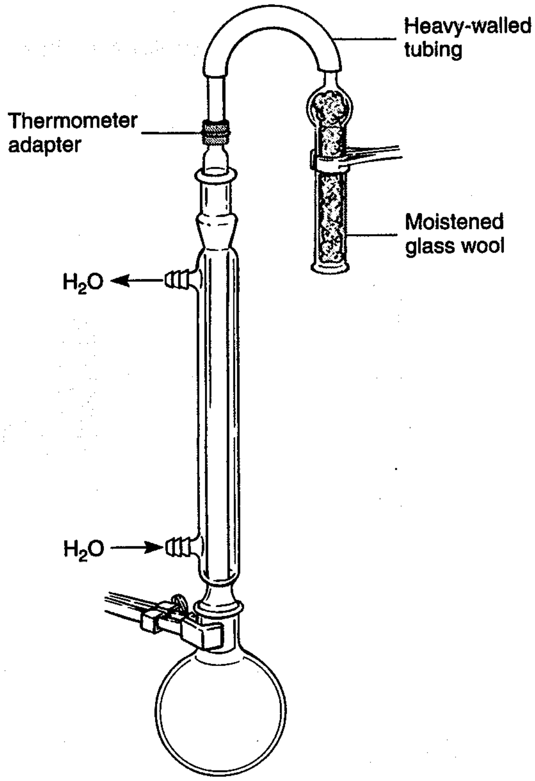
Heavy-walled tubing
Thermometer adapter
Moistened glass wool
H_2O
H_2O

Figure 3.15 Drying tube used to capture evolved gases.

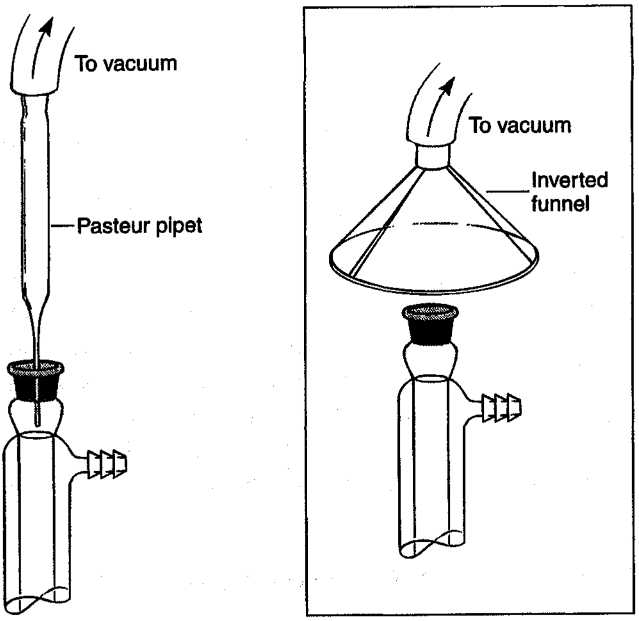
To vacuum
Pasteur pipet
To vacuum
Inverted funnel

Figure 3.16 Removal of noxious gases under vacuum. (The inset shows an alternative assembly, using an inverted funnel in place of the Pasteur pipet.)

pipet or funnel is attached to an aspirator with flexible tubing. A trap should be placed between the pipet or funnel and the aspirator. As gases are liberated from the reaction, they rise into the condenser. The vacuum draws the gases away from the apparatus. Both types of systems are shown in Figure 3.16. In the special case where the noxious gases are soluble in water, connecting a water aspirator to the pipet or funnel removes the gases from the reaction and traps them in the flowing water without the need for a separate gas trap.

3.10 COLLECTING GASEOUS PRODUCTS

In Section 3.9, means of removing unwanted gaseous products from the reaction system were examined. Some experiments produce gaseous products that you must collect and analyze. Methods to collect gaseous products are all based on the same principle. The gas is carried through tubing from the reaction to the opening of a flask or a test tube, which has been filled with water and is inverted in a container of water. The gas is allowed to bubble into the inverted collection tube (or flask). As the collection tube fills with gas, the water is displaced into the water container. If the collection tube is graduated, as in a graduated cylinder or a centrifuge tube, you can monitor the quantity of gas produced in the reaction.

If the inverted gas collection tube is constructed from a piece of glass tubing, a rubber septum can be used to close the upper end of the container. This type of collection tube is shown in Figure 3.17. A sample of the gas can be removed using a gas-tight syringe equipped with a needle. The gas that is removed can be analyzed by gas chromatography (see Technique 14).

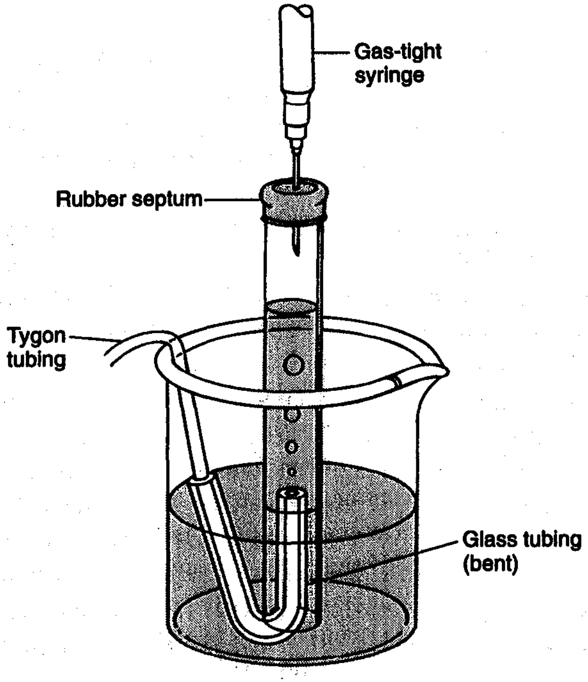

Figure 3.17 Gas collection tube, with rubber septum.

In Figure 3.17, a piece of glass tubing is attached to the free end of the flexible hose. This piece of glass tubing sometimes makes it easier to fix the open end in the proper position in the opening of the collection tube or flask. The other end of the flexible tubing is attached to a piece of glass tubing or a Pasteur pipet that has been inserted into a Thermometer adapter.

3.11 EVAPORATION OF SOLVENTS

In many experiments, it is necessary to remove excess solvent from a solution. An obvious approach is to allow the container to stand unstoppered in the hood for several hours until the solvent has evaporated. This method is generally not practical, however, and a quicker, more efficient means of evaporating solvents must be used.

Caution: You must always evaporate solvents in the hood.

This can be done by evaporating the solvent from an open Erlenmeyer flask (Fig. 3.18A and B). Such an evaporation must be conducted in a hood, since many solvent vapors are toxic or flammable. A boiling stone must be used. A gentle stream of air directed toward the surface of the liquid will remove vapors that are in equilibrium with the solution and accelerate the evaporation. A Pasteur pipet connected by a short piece of rubber tubing to the compressed air line will act as a convenient air nozzle (Fig. 3.18A). A tube or an inverted funnel connected to an aspirator may also be used (Fig. 3.18B). In this case, vapors are removed by suction. It is better to use an Erlenmeyer flask than a beaker for this procedure, since deposits of solid will usually build up on the sides of the beaker where the solvent evaporates. The refluxing action in an Erlenmeyer flask does not allow this buildup. If a hotplate is used as the heat source, care must be taken with flammable solvents to ensure against fires caused by "flashing," when solvent vapors come into contact with the hotplate surface.

It is also possible to remove low-boiling solvents under reduced pressure (Figure 3.18C). In this method, the solution is placed in a filter flask, along with a wooden applicator stick. The flask is stoppered, and the sidearm is connected to an aspirator (by a trap), as described in Technique 4, Section 4.3, p. 640. Under reduced pressure, the solvent begins to boil. The wooden stick serves the same function as a boiling stone. By this method, solvent can be evaporated from a solution without using much heat. This technique is often used when heating the solution might decompose thermally sensitive substances. The method has the disadvantage that, when low-boiling solvents are used, solvent evaporation cools the flask below the freezing point of water. When this happens, a layer of frost forms on the outside of the flask. Since frost is insulating, it must be removed to keep evaporation proceeding at a reasonable rate. Frost is best removed by one of two methods: either the flask is placed in a bath of warm water (with constant swirling) or it is heated on the steam bath (again with swirling). Either method promotes efficient heat transfer.

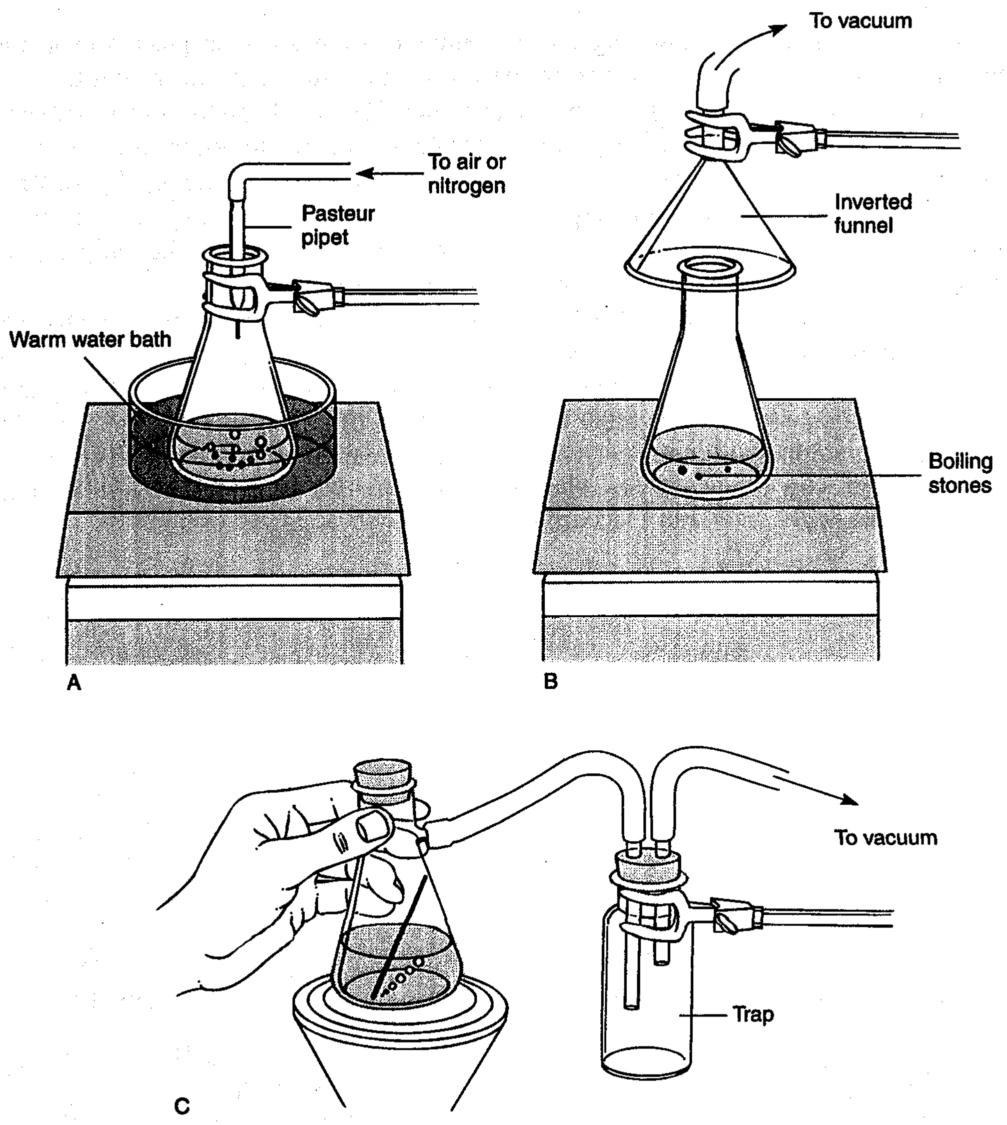
To air or nitrogen
Pasteur pipet
Warm water bath
A
To vacuum
Inverted funnel
Boiling stones
B
To vacuum
Trap
C

Figure 3.18 Evaporation of solvents (heat source can be varied among those shown.)

Large amounts of a solvent should be removed by distillation (see Technique 8). **One should never evaporate ether solutions to dryness,** except on a steam bath or by the reduced-pressure method. The tendency of ether to form explosive peroxides is a serious potential hazard. If peroxides should be present, the large and rapid temperature increase in the flask once the ether evaporates could bring about the detonation of any residual peroxides. The temperature of a steam bath is not high enough to cause such a detonation.

Small-Scale Methods

A simple means of evaporating a small amount of solvent is to place a centrifuge tube in a warm water bath. The heat from the water bath will warm the solvent to a temperature at which it can evaporate within a short time. The heat from the water can be adjusted to provide the best rate of evaporation, but the liquid should not be allowed to boil vigorously. The evaporation rate can be increased by allowing a stream of dry air or nitrogen to be directed into the centrifuge tube (Fig. 3.19A). The moving gas stream will sweep the vapors from the tube and accelerate the evaporation. As an alternative, a vacuum can be applied above the tube to draw away solvent vapors.

A convenient water bath suitable for microscale methods can be constructed by placing the aluminum collars, which are generally used with aluminum heating blocks, into a 150-mL beaker (Fig. 3.19B). In some cases, it may be necessary to round off the sharp edges of the collars with a file in order to allow them to fit properly into the beaker. Held by the aluminum collars, the conical vial will stand securely in the beaker. This assembly can be filled with water and placed on a hotplate for use in the evaporation of small amounts of solvent.

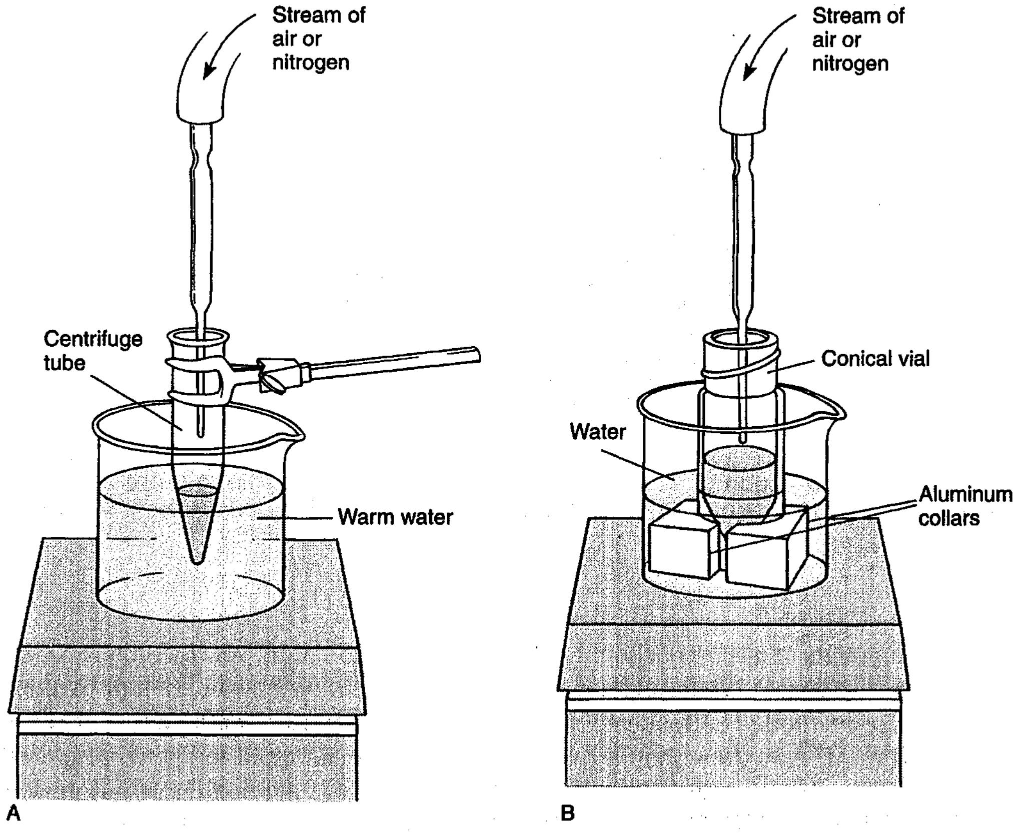

Figure 3.19 Evaporation of solvents (small-scale methods).

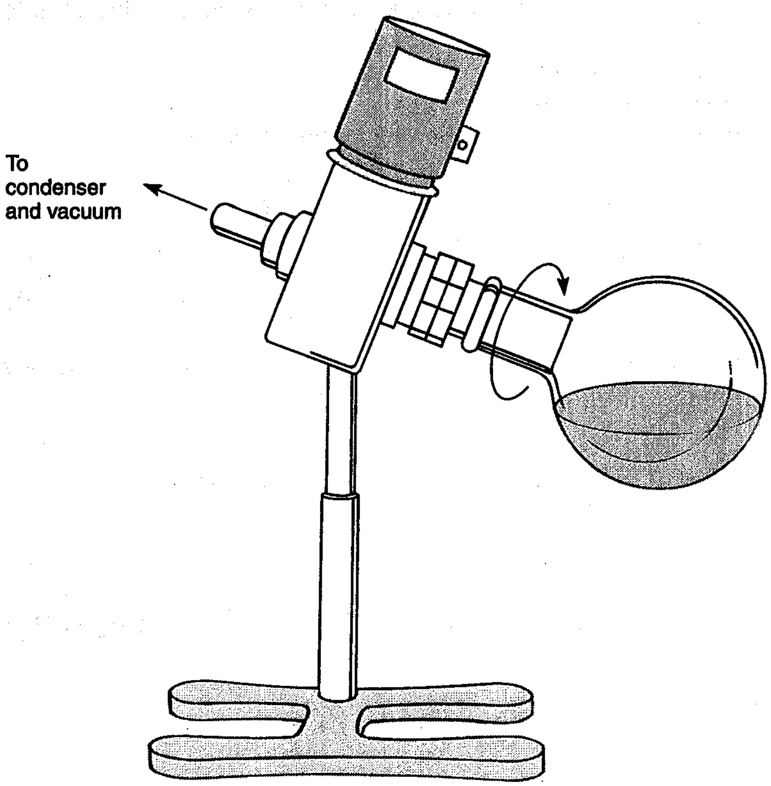

Figure 3.20 Rotary evaporator.

3.12 ROTARY EVAPORATOR

In the research laboratory, solvents are evaporated under reduced pressure using a **rotary evaporator.** This is a motor-driven device, which is designed for rapid evaporation of solvents, with heating, while minimizing the possibility of bumping. A vacuum is applied to the flask, and the motor spins the flask. The rotation of the flask spreads a thin film of the liquid over the surface of the glass. This accelerates evaporation. The rotation also agitates the solution sufficiently to reduce the problem of bumping. A water bath can be placed under the flask to warm the solution and increase the vapor pressure of the solvent. One can select the speed at which the flask is rotated and the temperature of the water bath to attain the desired evaporation rate. A rotary evaporator is shown in Figure 3.20.

PROBLEMS

1. What is the difference between
   (a) ether and petroleum ether?
   (b) ether and diethyl ether?
   (c) ligroin and petroleum ether?

2. What is the best type of stirring device to use for stirring a reaction that takes place in
   (a) a conical vial?
   (b) a 10-mL round-bottom flask?
   (c) a 250-mL round-bottom flask?
3. Should you use a drying tube for the following reaction? Explain.

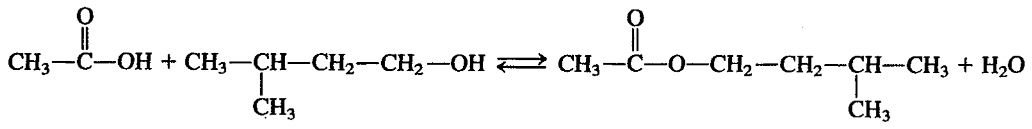

$$CH_3-\overset{O}{\overset{\|}{C}}-OH + CH_3-\underset{CH_3}{\underset{|}{CH}}-CH_2-CH_2-OH \rightleftarrows CH_3-\overset{O}{\overset{\|}{C}}-O-CH_2-CH_2-\underset{CH_3}{\underset{|}{CH}}-CH_3 + H_2O$$

4. For which of the following reactions should you use a trap to collect noxious gases?

(a) 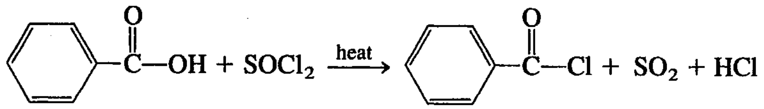

(b) 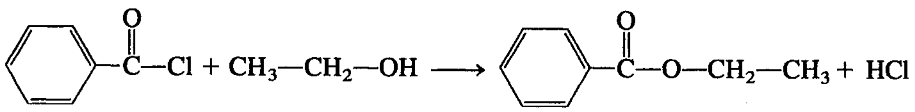

(c) $C_{12}H_{22}O_{11} + H_2O \longrightarrow 4\ CH_3-CH_2-OH + 4\ CO_2$
(Sucrose)

(d) 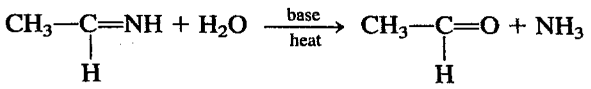

$$CH_3-\underset{H}{\underset{|}{C}}{=}NH + H_2O \xrightarrow[\text{heat}]{\text{base}} CH_3-\underset{H}{\underset{|}{C}}{=}O + NH_3$$

TECHNIQUE 4

Filtration

Filtration is a technique used for two main purposes. The first is to remove solid impurities from a liquid. The second is to collect a desired solid from the solution from which it was precipitated or crystallized. Several different kinds of filtration are commonly used: two general methods include gravity filtration and vacuum (or suction) filtration. The various filtration techniques and their applications are summarized in Table 4.1. These techniques are discussed in more detail in the following sections.

4.1 GRAVITY FILTRATION

The most familiar filtration technique is probably filtration of a solution through a paper filter held in a funnel, allowing gravity to draw the liquid through the paper. This technique is useful only when the volume of mixture to be filtered is greater than 10 mL. For

TABLE 4.1. Filtration Methods

Method	Application	Section
GRAVITY FILTRATION		
Filter cones	The volume of liquid to be filtered is about 10 mL or greater, and the solid collected in the filter is saved.	4.1A
Fluted filters	The volume of liquid to be filtered is greater than about 10 mL, and solid impurities are removed from a solution; often used in crystallization procedures.	4.1B
Filtering pipets	Used with volumes less than about 10 mL to remove solid impurities from a liquid.	4.1C
VACUUM FILTRATION		
Büchner funnels	Primarily used to collect a desired solid from a liquid when the volume is greater than about 10 mL; used frequently to collect the crystals obtained from crystallization.	4.3
Hirsch funnels	Used in the same way as Büchner funnels, except the volume of liquid is usually smaller (1–10 mL)	4.3
FILTERING MEDIA	Used to remove finely divided impurities.	4.4
FILTER TIP PIPETS	May be used to remove a small amount of solid impurities from a small volume (1–2 mL) of liquid; also useful for pipetting volatile liquids, especially in extraction procedures.	4.6
CRAIG TUBES	Used to collect a small amount of crystals resulting from crystallizations in which the volume of the solution is less than 2 mL.	4.7
CENTRIFUGATION	Although not strictly a filtration technique, centrifugation may be used to remove suspended impurities from a liquid (1–25 mL).	4.8

some small-scale procedures a more suitable technique, which also makes use of gravity, is to use a Pasteur (or disposable) pipet with a cotton or glass wool plug (called a filtering pipet).

A. FILTER CONES

This filtration technique is most useful when the solid material being filtered from a mixture is to be collected and used later. The filter cone, because of its smooth sides, can easily be scraped free of collected solids. Because of the many folds, fluted filter paper, described in the next section, cannot be scraped easily. The filter cone is likely to be used only when a relatively large volume (greater than 10 mL) is being filtered and when a Büchner funnel (Section 4.3) is not appropriate.

The filter cone is prepared as indicated in Figure 4.1. It is then placed into a funnel of an appropriate size. With filtrations using a simple filter cone, solvent may form seals between the filter and the funnel and between the funnel and the lip of the receiving flask. When a seal forms, the filtration stops because the displaced air has no possibility of escaping. To avoid the solvent seal, you can insert a small piece of paper, a paper clip, or some other bent wire between the funnel and the lip of the flask to let the displaced air escape. As an alter-

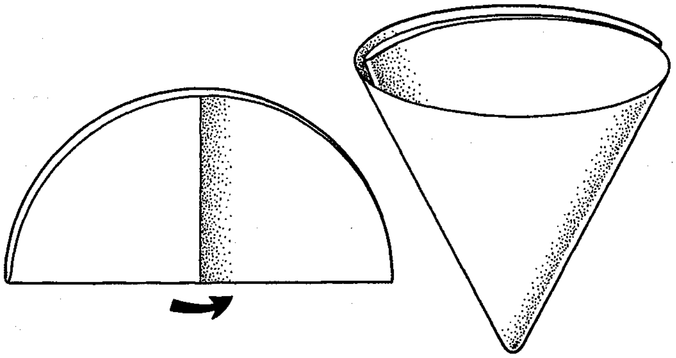

Figure 4.1 Folding a filter cone.

native, you can support the funnel by a clamp fixed *above* the flask, rather than by placing it on the neck of the flask. A gravity filtration using a filter cone is shown in Figure 4.2.

B. FLUTED FILTERS

This filtration method is also most useful when filtering a relatively large amount of liquid. Because a fluted filter is used when the desired material is expected to remain in solution, this filter is used to remove undesired solid materials, such as dirt particles, decolorizing charcoal, and undissolved impure crystals. A fluted filter is often used to filter a hot solution saturated with a solute during a crystallization procedure.

The technique for folding a fluted filter paper is shown in Figure 4.3. An advantage of a fluted filter is that it increases the speed of filtration, which occurs for two reasons.

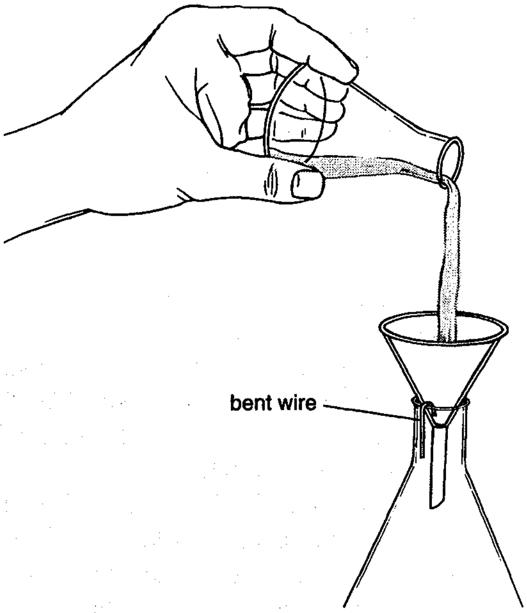

Figure 4.2 Gravity filtration with a filter cone.

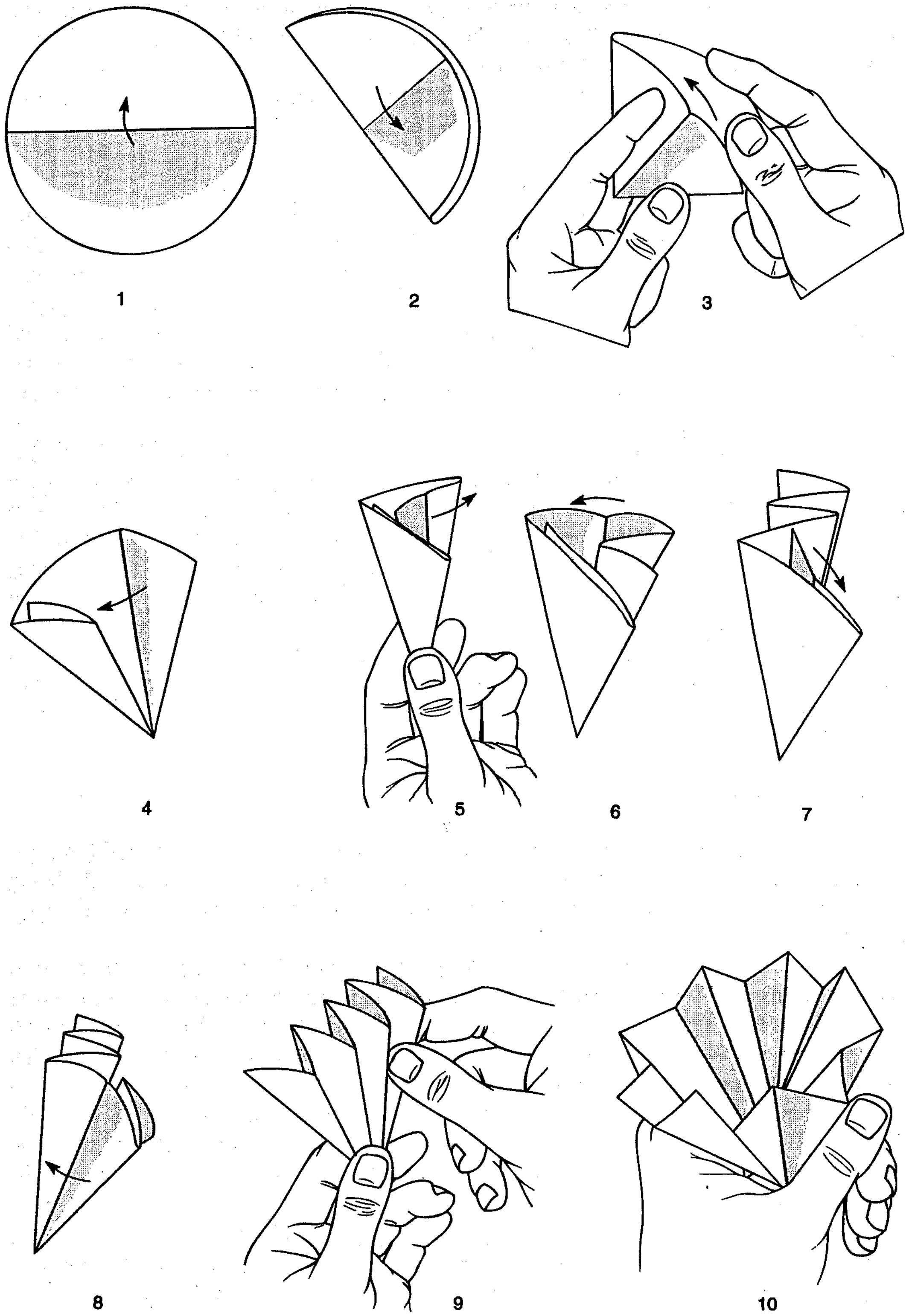
1
2
3
4
5
6
7
8
9
10

Figure 4.3 Folding a fluted filter paper, or origami at work in the organic lab.

First, it increases the surface area of the filter paper through which the solvent seeps; second, it allows air to enter the flask along its sides to permit rapid pressure equalization. If pressure builds up in the flask from hot vapors, filtering slows down. This problem is especially pronounced with filter cones. The fluted filter tends to reduce this problem considerably, but it may be a good idea to clamp the funnel above the receiving flask or to use a piece of paper, paper clip, or wire between the funnel and the lip of the flask as an added precaution against solvent seals.

Filtration with a fluted filter is relatively easy to perform when the mixture is at room temperature. However, when it is necessary to filter a hot solution saturated with a dissolved solute, a number of steps must be taken to ensure that the filter does not become clogged by solid material accumulated in the stem of the funnel or in the filter paper. When the hot saturated solution comes in contact with a relatively cold funnel (or a cold flask, for that matter), the solution is cooled and may become supersaturated. If crystallization then occurs in the filter, either the crystals will fail to pass through the filter paper or they will clog the stem of the funnel.

To keep the filter from clogging, use one of the following four methods. The first is to use a short-stemmed or a stemless funnel. With these funnels, it is less likely that the stem of the funnel will become clogged by solid material. The second method is to keep the liquid to be filtered at or near its boiling point at all times. The third way is to preheat the funnel by pouring hot solvent through it before the actual filtration. This keeps the cold glass from causing instantaneous crystallization. And fourth, it is helpful to keep the **filtrate** (filtered solution) in the receiver hot enough to continue boiling *slightly* (by setting it on a hotplate, for example). The refluxing solvent heats the receiving flask and the funnel stem and washes them clean of solids. This boiling of the filtrate also keeps the liquid in the funnel warm.

C. FILTERING PIPETS

A filtering pipet is a microscale technique most often used to remove solid impurities from a liquid with a volume less than 10 mL. It is important that the mixture being filtered be at or near room temperature because it is difficult to prevent premature crystallization in a hot solution saturated with a solute.

To prepare this filtration device, a small piece of cotton is inserted into the top of a Pasteur (disposable) pipet and pushed down to the beginning of the lower constriction in the pipet, as shown in Figure 4.4. It is important that enough cotton is used to collect all the solid being filtered; however, the amount of cotton used should not be so large that the flow rate through the pipet is significantly restricted. For the same reason, the cotton should not be packed too tightly. The cotton plug can be pushed down with a long thin object such as a glass stirring rod or a wooden applicator stick. It is advisable to wash the cotton plug by passing about 1 mL of solvent (usually the same solvent that is to be filtered) through the filter.

In some cases, such as when filtering a strongly acidic mixture or when performing a very rapid filtration to remove dirt or impurities of large particle size from a solution, it may be better to use glass wool in place of the cotton. The disadvantage in using glass

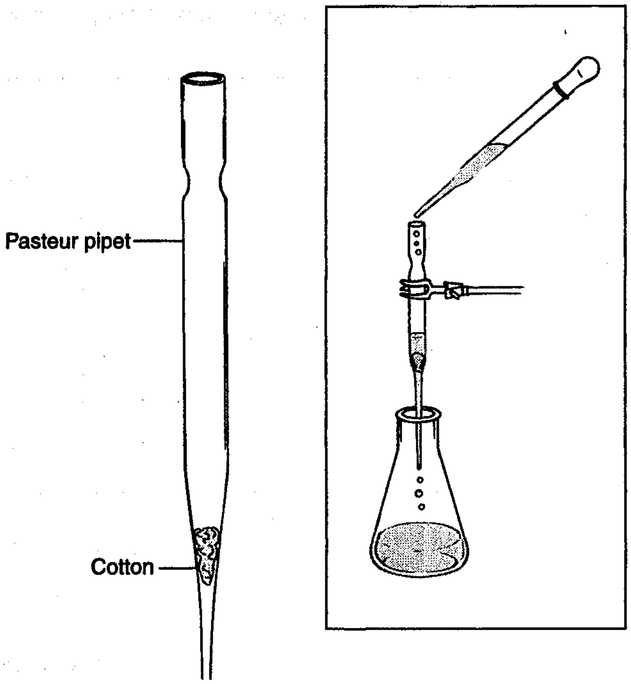

Figure 4.4 Filtering pipet.

wool is that the fibers do not pack together as tightly, and small particles will pass through the filter more easily.

To conduct a filtration (with either a cotton or glass wool plug), the filtering pipet is clamped so that the filtrate will drain into an appropriate container. The mixture to be filtered is usually transferred to the filtering pipet with another Pasteur pipet. If a small volume of liquid is being filtered (less than 1 or 2 mL), it is advisable to rinse the filter and plug with a small amount of solvent after the last of the filtrate has passed through the filter. The rinse solvent is then combined with the original filtrate. If desired, the rate of filtration can be increased by gently applying pressure to the top of the pipet using a pipet bulb.

Depending on the amount of solid being filtered and the size of the particles (small particles are more difficult to remove by filtration), it may be necessary to put the filtrate through a second filtering pipet. This should be done with a new filtering pipet rather than the one already used.

4.2 FILTER PAPER

Many kinds and grades of filter paper are available. The paper must be correct for a given application. In choosing filter paper, you should be aware of its various properties. **Porosity** is a measure of the size of the particles that can pass through the paper.

TABLE 4.2. Some Common Qualitative Filter Paper Types and Approximate Relative Speeds and Retentivities

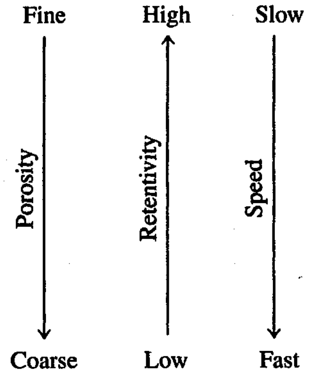

Speed	Type (by number)		
	E&D	*S&S*	*Whatman*
Very slow	610	576	5
Slow	613	602	3
Medium	615	597	2
Fast	617	595	1
Very fast	—	604	4

Highly porous paper does not remove small particles from solution; paper with low porosity removes very small particles. **Retentivity** is a property that is the opposite of porosity. Paper with low retentivity does not remove small particles from the filtrate. The **speed** of filter paper is a measure of the time it takes a liquid to drain through the filter. Fast paper allows the liquid to drain quickly; with slow paper, it takes much longer to complete the filtration. Since all these properties are related, fast filter paper usually has a low retentivity and high porosity, and slow filter paper usually has high retentivity and low porosity.

Table 4.2 compares some commonly available qualitative filter paper types and ranks them according to porosity, retentivity, and speed. Eaton-Dikeman (E&D), Schleicher and Schuell (S&S), and Whatman are the most common brands of filter paper. The numbers in the table refer to the grades of paper used by each company.

4.3 VACUUM FILTRATION

Vacuum, or suction, filtration is more rapid than gravity filtration and is most often used to collect solid products resulting from precipitation or crystallization. This technique is used primarily when the volume of liquid being filtered is more than 1–2 mL. With smaller volumes, use of the Craig tube (Section 4.7) is the preferred technique. In a vacuum filtration, a receiver flask with a sidearm, a **filter flask,** is used. For small-scale laboratory work, the most useful sizes of filter flasks range from 50 mL to 500 mL, depending on the volume of liquid being filtered. The sidearm is connected by *heavy-walled* rubber tubing (see Technique 9, Fig. 9.2, p. 720) to a source of vacuum. Thin-walled tubing will collapse under vacuum, due to atmospheric pressure on its outside walls, and will seal the vacuum source from the flask. Because this apparatus is unstable and can tip over easily, it should be clamped, as shown in Figure 4.5.

Caution: It is essential that the filter flask be clamped.

Two types of funnels are useful for vacuum filtration, the Büchner funnel and the Hirsch funnel. The **Büchner funnel** is used for filtering larger amounts of solid from solution. Büchner funnels are usually made from polypropylene or porcelain. A Büchner funnel (see Fig. 4.5) is sealed to the filter flask by a rubber stopper or a filter (Neoprene) adapter. The flat bottom of the Büchner funnel is covered with an unfolded piece of circular filter paper. To prevent the escape of solid materials from the funnel, you must be certain that the filter paper fits the funnel exactly. It must cover all the holes in the bottom of the funnel but not extend up the sides. Before beginning the filtration, it is advisable to moisten the paper with a small amount of solvent. The moistened filter paper adheres more strongly to the bottom of the funnel and prevents unfiltered mixture from passing around the edges of the filter paper.

The **Hirsch funnel,** which is shown in Figure 4.5B and C, operates on the same principle as the Büchner funnel, but it is usually smaller and its sides are sloped rather

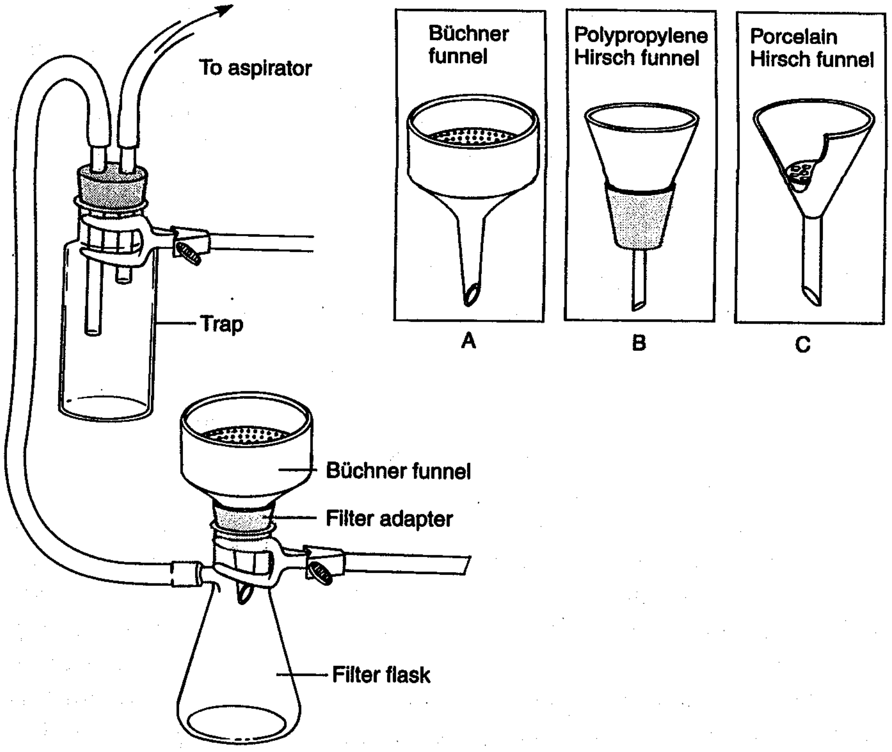

Figure 4.5 Vacuum filtration.

than vertical. The polypropylene Hirsch funnel (see Fig. 4.5B) is sealed to a 50-mL filter flask by a small section of Gooch tubing. This Hirsch funnel has a built-in adapter that forms a tight seal with some 25-mL filter flasks without the Gooch tubing. A polyethylene fritted disk fits into the bottom of the funnel. To prevent the holes in this disk from becoming clogged with solid material, the funnel should always be used with a circular filter paper that has the same diameter (1.27-cm) as the polyethylene disk. With a polypropylene Hirsch funnel, it is also important to moisten the paper with a small amount of solvent before beginning the filtration.

The porcelain Hirsch funnel is sealed to the filter flask with a rubber stopper or a Neoprene adapter. In this Hirsch funnel, the filter paper must also cover all the holes in the bottom but must not extend up the sides.

Because the filter flask is attached to a source of vacuum, a solution poured into a Büchner funnel or Hirsch funnel is literally "sucked" rapidly through the filter paper. For this reason, a vacuum filtration is generally not used to separate fine particles such as decolorizing charcoal, since the small particles would likely be pulled through the filter paper. However, this problem can be alleviated when desired by the use of specially prepared filter beds (see Section 4.4).

4.4 FILTERING MEDIA

It is occasionally necessary to use specially prepared filter beds to separate fine particles when using vacuum filtration. Often, very fine particles either pass right through a paper filter or they clog it so completely that the filtering stops. This is avoided by using a substance called Filter Aid, or Celite. This material is also called **diatomaceous earth** because of its source. It is a finely divided inert material derived from the microscopic shells of dead diatoms (a type of phytoplankton that grows in the sea).

Caution: LUNG IRRITANT
When using Filter Aid, take care not to breathe the dust.

Filter Aid will not clog the fiber pores of filter paper. It is **slurried,** mixed with a solvent to form a rather thin paste, and filtered through a Hirsch or Büchner funnel (with filter paper in place) until a layer of diatoms about 2–3 mm thick is formed on top of the filter paper. The solvent in which the diatoms were slurried is poured from the filter flask, and, if necessary, the filter flask is cleaned before the actual filtration is begun. Finely divided particles can now be suction-filtered through this layer and will be caught in the Filter Aid. This technique is used for removing impurities, not for collecting a product. The filtrate (filtered solution) is the desired material in this procedure. If the material caught in the filter were the desired material, you would have to try to separate the product from all those diatoms! Filtration with Filter Aid is not appropriate when the desired substance is likely to precipitate or crystallize from solution.

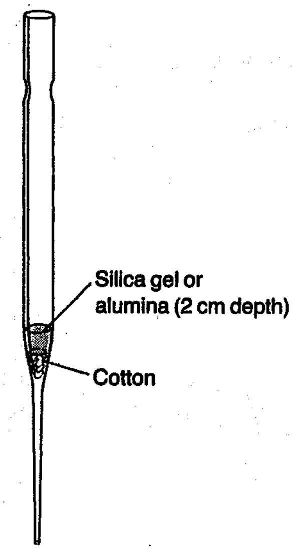

Figure 4.6 Pasteur pipet with filtering media.

In microscale work, it may sometimes be more convenient to use a column prepared with a Pasteur pipet to separate fine particles from a solution. The Pasteur pipet is packed with alumina or silica gel, as shown in Figure 4.6.

4.5 THE ASPIRATOR

The most common source of vacuum (approximately 10–20 mmHg) in the laboratory is the water aspirator, or "water pump," illustrated in Figure 4.7. This device passes water rapidly past a small hole to which a sidearm is attached. The water pulls air in

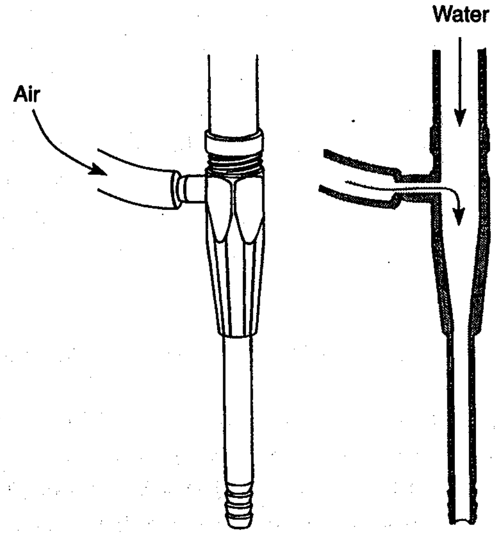

Figure 4.7 Aspirator.

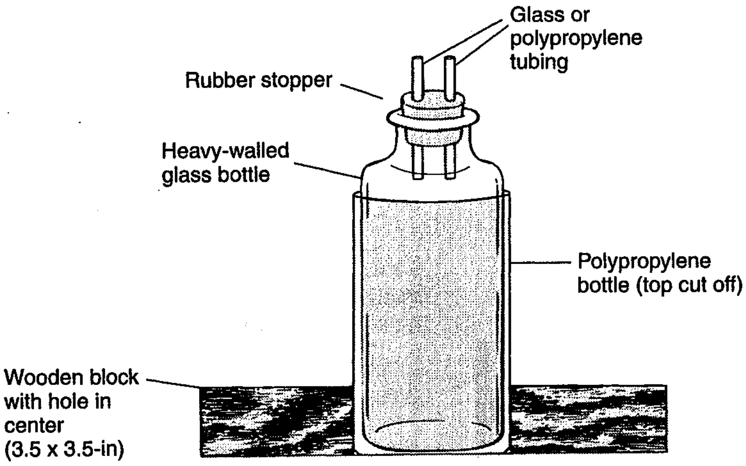
Glass or polypropylene tubing
Rubber stopper
Heavy-walled glass bottle
Polypropylene bottle (top cut off)
Wooden block with hole in center (3.5 x 3.5-in)

Figure 4.8 Simple aspirator trap and holder.

through the sidearm. This phenomenon, called the Bernoulli effect, causes a reduced pressure along the side of the rapidly moving water stream and creates a partial vacuum in the sidearm.

The aspirator works most effectively when the water is turned on to the fullest extent.

A water aspirator can never lower the pressure beyond the vapor pressure of the water used to create the vacuum. Hence, there is a lower limit to the pressure (on cold days) of 9–10 mmHg. A water aspirator does not provide as high a vacuum in the summer as in the winter, due to this water-temperature effect.

A trap must be used with an aspirator. One type of trap is illustrated in Figure 4.5. Another method for securing this type of trap is shown in Figure 4.8. This simple holder can be constructed from readily available material and can be placed anywhere on the laboratory bench. Although not often needed, a trap can prevent water from contaminating your experiment. If the water pressure in the laboratory drops suddenly, the pressure in the filter flask may suddenly become lower than the pressure in the water aspirator. This would cause water to be drawn from the aspirator stream into the filter flask and contaminate the filtrate or even the material in the filter. The trap stops this reverse flow. A similar flow will occur if the water flow at the aspirator is stopped before the tubing connected to the aspirator sidearm is disconnected.

Always disconnect the tubing before stopping the aspirator.

If a "back-up" begins, disconnect the tubing as rapidly as possible before the trap fills with water. Some chemists like to fit a stopcock into the stopper on top of the trap. A three-hole stopper is required for this purpose. With a stopcock in the trap, the system can be vented before the aspirator is shut off. Then, water cannot back up into the trap.

Aspirators do not work well if too many people use the water line at the same time because the water pressure is lowered. Also, the sinks at the ends of the lab benches or the lines that carry away the water flow may have a limited capacity for draining the resultant water flow from too many aspirators. Care must be taken to avoid floods.

4.6 FILTER-TIP PIPET

The filter-tip pipet, illustrated in Figure 4.9, has two common uses. The first is to remove a small amount of solid, such as dirt or filter paper fibers, from a small volume of liquid (1–2 mL). A filter-tip pipet is very useful for removing small particles from solutions of samples prepared for NMR analysis. It can also be helpful when using a Pasteur pipet to transfer a highly volatile liquid, especially during an extraction procedure (see Technique 7, Section 7.6, p. 693).

Preparing a filter-tip pipet is similar to preparing a filtering pipet, except that a much smaller amount of cotton is used. A very tiny piece of cotton is loosely shaped into a ball and placed into the large end of a Pasteur pipet. Using a wire with a diameter slightly smaller than the inside diameter of the narrow end of the pipet, the ball of cotton is pushed to the bottom of the pipet. If it becomes difficult to push the cotton, you have probably started with too much cotton; if the cotton slides through the narrow end with little resistance, you probably have not used enough.

To use a filter-tip pipet as a filter, the mixture is drawn up into the Pasteur pipet using a pipet bulb and then expelled. With this procedure, a small amount of solid will be captured by the cotton. However, very fine particles, such as activated charcoal, cannot

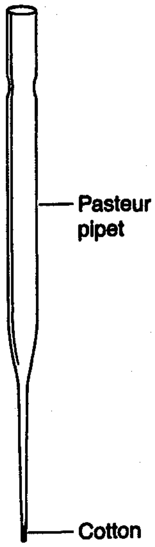

Figure 4.9 Filter-tip pipet.

be removed efficiently with a filter-tip pipet, and this technique is not effective in removing more than a trace amount of solid from a liquid.

Transferring many organic liquids with a Pasteur pipet can be a somewhat difficult procedure for two reasons. First, the liquid may not adhere well to the glass. Second, as you handle the Pasteur pipet, the temperature of the liquid in the pipet increases slightly, and the increased vapor pressure may tend to "squirt" the liquid out the end of the pipet. This problem can be particularly troublesome when separating two liquids during an extraction procedure. The purpose of the cotton plug in this situation is to slow the rate of flow through the end of the pipet so that you can control the movement of liquid in the Pasteur pipet more easily.

4.7 CRAIG TUBES

The **Craig tube,** illustrated in Figure 4.10, is used primarily to separate crystals from a solution after a microscale crystallization procedure has been performed (Technique 5, Section 5.4, p. 658). Although it may not be a filtration procedure in the traditional sense, the outcome is similar. The outer part of the Craig tube is similar to a test tube, except that the diameter of the tube becomes wider part of the way up the tube, and the glass is ground at this point so that the inside surface is rough. The inner part (plug) of the Craig tube may be made of Teflon or glass. If this part is glass, the end of the plug is also ground. With either a glass or a Teflon inner plug, there is only a partial seal where the plug and the outer tube come together. Liquid may pass through, but solid will not. This is where the solution is separated from the crystals.

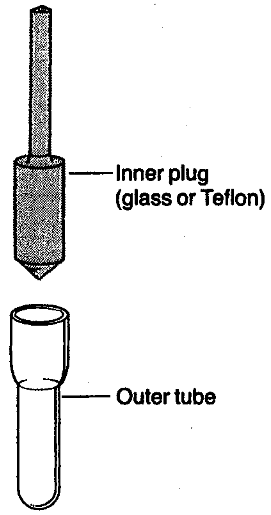

Figure 4.10 Craig tube (2 mL).

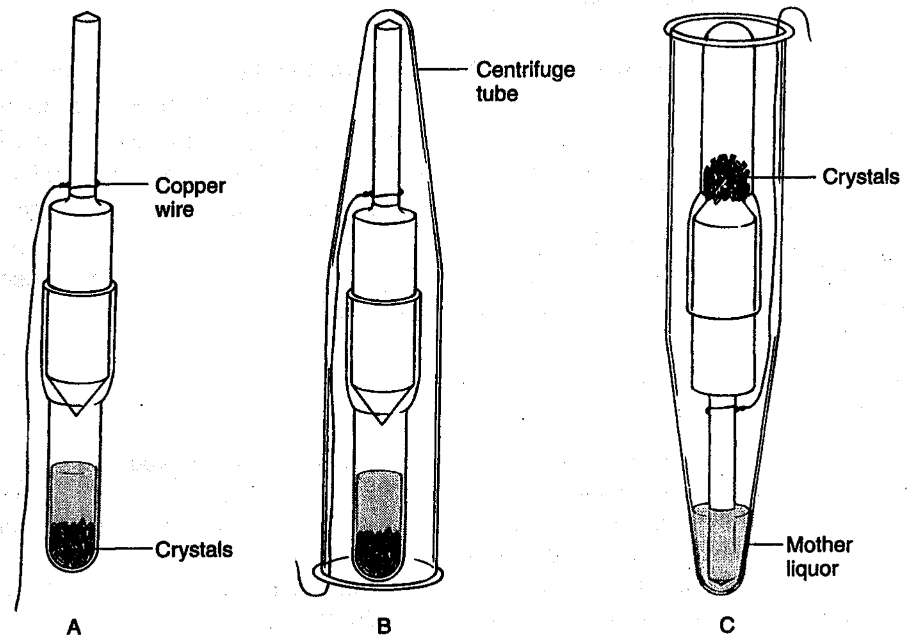

Figure 4.11 Separation with a Craig tube.

After crystallization has been completed in the outer Craig tube, replace the inner plug (if necessary) and connect a thin copper wire or strong thread to the narrow part of the inner plug, as indicated in Figure 4.11A. While holding the Craig tube in an upright position, a plastic centrifuge tube is placed over the Craig tube so that the bottom of the centrifuge tube rests on top of the inner plug, as shown in Figure 4.11B. The copper wire should extend just below the lip of the centrifuge tube and is now bent upward around the lip of the centrifuge tube. This apparatus is then turned over so that the centrifuge tube is in an upright position. The Craig tube is spun in a centrifuge (be sure it is balanced by placing another tube filled with water on the opposite side of the centrifuge) for several minutes until the **mother liquor** (solution from which the crystals grew) goes to the bottom of the centrifuge tube and the crystals collect on the end of the inner plug (see Fig. 4.11C). Depending on the consistency of the crystals and the speed of the centrifuge, the crystals may spin down to the inner plug, or (if you are unlucky) they may remain at the other end of the Craig tube. If the latter situation occurs, it may be helpful to centrifuge the Craig tube longer, or, if this problem is anticipated, to stir the crystal and solution mixture with a spatula or stirring rod before centrifugation.

Using the copper wire, the Craig tube is then pulled out of the centrifuge tube. If the crystals collected on the end of the inner plug, it is now a simple procedure to remove the plug and scrape the crystals with a spatula onto a watch glass, a clay plate, or a piece of smooth paper. Otherwise, it will be necessary to scrape the crystals from the inside surface of the outer part of the Craig tube.

4.8 CENTRIFUGATION

Sometimes centrifugation is more effective in removing solid impurities than are conventional filtration techniques. Centrifugation is particularly effective in removing suspended particles which are so small that the particles would pass through most filtering devices. Another situation in which centrifugation may be useful is when the mixture must be kept hot to prevent premature crystallization while the solid impurities are removed.

Centrifugation is performed by placing the mixture in one or two centrifuge tubes (be sure to balance the centrifuge) and centrifuging for several minutes. The supernatant liquid is then decanted (poured off) or removed with a Pasteur pipet.

PROBLEMS

1. In each of the following situations, what type of filtration device would you use?
  - **(a)** Remove powdered decolorizing charcoal from 20 mL of solution.
  - **(b)** Collect crystals obtained from crystallizing a substance from about 1 mL of solution.
  - **(c)** Remove a very small amount of dirt from 1 mL of liquid.
  - **(d)** Isolate 2.0 g of crystals from about 50 mL of solution after performing a crystallization.
  - **(e)** Remove dissolved colored impurities from about 3 mL of solution.
  - **(f)** Remove solid impurities from 5 mL of liquid at room temperature.

TECHNIQUE 5

Crystallization: Purification of Solids

Organic compounds that are solid at room temperature are usually purified by crystallization. The general technique involves dissolving the material to be crystallized in a *hot* solvent (or solvent mixture) and cooling the solution slowly. The dissolved material has a decreased solubility at lower temperatures and will separate from the solution as it is cooled. This phenomenon is called either **crystallization** if the crystal growth is relatively slow and selective or **precipitation** if the process is rapid and nonselective. Crystallization is an equilibrium process and produces very pure material. A small seed crystal is formed initially, and it then grows layer by layer in a reversible manner. In a sense, the crystal "selects" the correct molecules from the solution. In precipitation, the crystal lattice is formed so rapidly that impurities are trapped within the lattice. Therefore, any attempt at purification with too rapid a process should be avoided.

The method of crystallization described in detail in this chapter is called **standard-scale crystallization.** This technique, which is carried out with an Erlenmeyer flask to dissolve the material and a Büchner funnel to filter the crystals, is normally used when the weight of solid to be crystallized is more than 0.1 g. Another method, which is performed with a Craig tube, is used with smaller amounts of solid. Referred to as **microscale crystallization,** this technique is discussed briefly in Section 5.4.

5.1 SOLUBILITY

The first problem in performing a crystallization is selecting a solvent in which the material to be crystallized shows the desired solubility behavior. In the ideal case, the material should be sparingly soluble at room temperature and yet quite soluble at the boiling point of the solvent selected. The solubility curve should be steep, as can be seen in line A of Figure 5.1. A curve with a low slope (line B, Fig. 5.1) would not cause significant crystallization when the temperature of the solution was lowered. A solvent in which the material is very soluble at all temperatures (line C, Fig. 5.1) also would not be a suitable crystallization solvent. The basic problem in performing a crystallization is to select a solvent (or mixed solvent) that provides a steep solubility-vs.-temperature curve for the material to be crystallized. A solvent that allows the behavior shown in line A is an ideal crystallization solvent.

The solubility of organic compounds is a function of the polarities of both the solvent and the **solute** (dissolved material). A general rule is "like dissolves like." If the solute is very polar, a very polar solvent is needed to dissolve it; if it is nonpolar, a nonpolar solvent is needed. Usually, compounds having functional groups that can form hydrogen bonds (for example, —OH, —NH—, —COOH, —CONH—) will be more soluble in hydroxylic solvents such as water or methanol than in hydrocarbon solvents such as toluene or hexane. However, if the functional group is not a major part of the molecules, this solubility behavior may be reversed. For instance, dodecyl alcohol $CH_3(CH_2)_{10}CH_2OH$ is almost insoluble in water; its 12-carbon chain causes it to behave more like a hydrocarbon than an alcohol. The list in Table 5.1 gives an approximate order for decreasing polarity of organic functional groups.

The stability of the crystal lattice also affects solubility. Other things being equal, the higher the melting point (the more stable the crystal), the less soluble the compound. For instance, *p*-nitrobenzoic acid (mp 242°C) is, by a factor of 10, less soluble in a fixed amount of ethanol than the *ortho* (mp 147°C) and *meta* (mp 141°C) isomers.

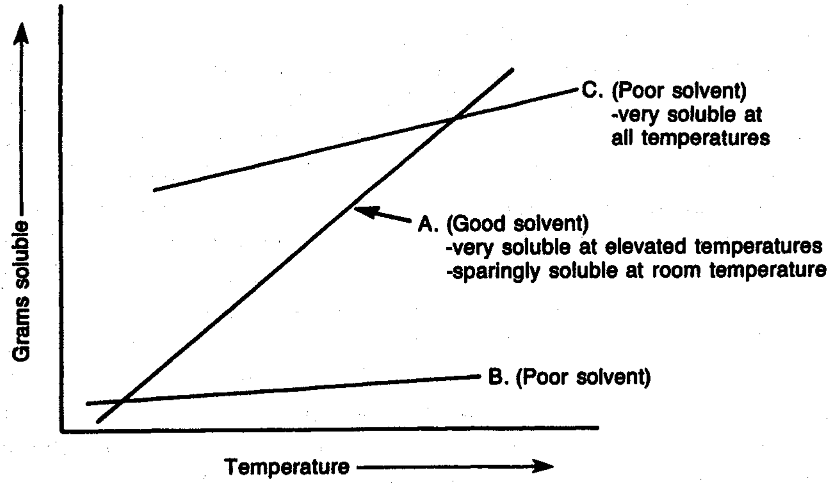

Figure 5.1 Graph of solubility vs. temperature.

TABLE 5.1. Solvents, in Decreasing Order of Polarity

Decreasing Polarity (Approximate) ↓	H_2O	Water
	RCOOH	Organic acids (acetic acid)
	$RCONH_2$	Amides (*N,N*-dimethylformamide)
	ROH	Alcohols (methanol, ethanol)
	RNH_2	Amines (triethylamine, pyridine)
	RCOR	Aldehydes, ketones (acetone)
	RCOOR	Esters (ethyl acetate)
	RX	Halides ($CH_2Cl_2 > CHCl_3 > CCl_4$)
	ROR	Ethers (diethyl ether)
	ArH	Aromatics (benzene, toluene)
	RH	Alkanes (hexane, petroleum ether)

5.2 THEORY OF CRYSTALLIZATION

A successful crystallization depends on a large difference between the solubility of a material in a hot solvent and its solubility in the same solvent when it is cold. When the impurities in a substance are equally soluble in both the hot and the cold solvent, an effective purification is not easily achieved through crystallization. A material can be purified by crystallization when both the desired substance and the impurity have similar solubilities, but only when the impurity represents a small fraction of the total solid. The desired substance will crystallize on cooling, but the impurities will not.

For example, consider a case in which the solubilities of substance A and its impurity B are both 1 g/100 mL of solvent at 20°C and 10 g/100 mL of solvent at 100°C. In an impure sample of A, the composition is given to be 9 g of A and 2 g of B for this particular example. In the calculations for this example, it is assumed that the solubilities of both A and B are unaffected by the presence of the other substance. One hundred mL of solvent is used in each crystallization to make the calculations easier to understand. Normally, the minimum amount of solvent required to dissolve the solid would be used.

At 20°C, this total amount of material will not dissolve in 100 mL of solvent. However, if the solvent is heated to 100°C, all 11 g dissolve. The solvent has the capacity to dissolve 10 g of A *and* 10 g of B at this temperature. If the solution is cooled to 20°C, only 1 g of each solute can remain dissolved, so 8 g of A and 1 g of B crystallize, leaving 2 g of material in the solution. This crystallization is shown in Figure 5.2. The solution that remains after a crystallization is called the **mother liquor.** If the process is now repeated by treating the crystals with 100 mL of fresh solvent, 7 g of A will crystallize again, leaving 1 g of A and 1 g of B in the mother liquor. As a result of these operations, 7 g of pure A are obtained, but with the loss of 4 g of material (2 g of A plus 2 g of B). Again, this second crystallization step is illustrated in Figure 5.2. The final result illustrates an important aspect of crystallization—it is wasteful. Nothing can be done to prevent this waste; some A must be lost along with the impurity B for the method to be successful. Of course, if the impurity B were *more* soluble than A in the solvent, the

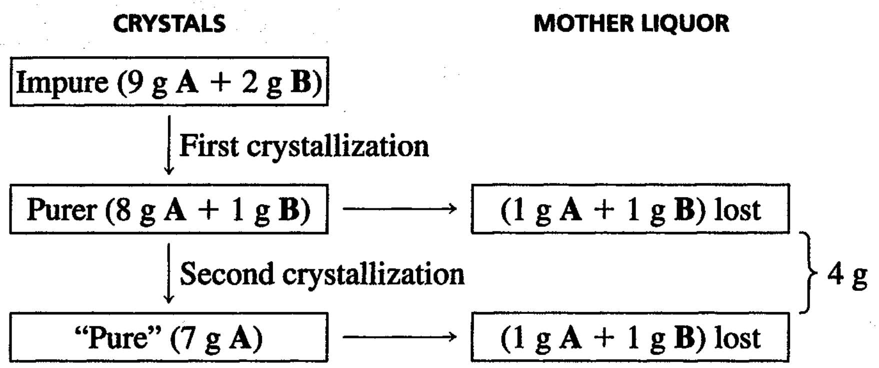

Figure 5.2 Purification of a mixture by crystallization.

losses would be reduced. Losses could also be reduced if the impurity were present in *much smaller* amounts than the desired material.

Note that in the preceding case, the method operated successfully because A was present in substantially larger quantity than its impurity B. If there had been a 50/50 mixture of A and B initially, no separation would have been achieved. In general, a crystallization is successful only if there is a *small* amount of impurity. As the amount of impurity increases, the loss of material must also increase. Two substances with nearly equal solubility behavior, present in equal amounts, cannot be separated. If the solubility behavior of two components present in equal amounts is different, however, a separation or purification is frequently possible.

In the preceding example, two crystallization procedures were performed. Normally this is not necessary; however, when it is, the second crystallization is more appropriately called **recrystallization.** As illustrated in this example, a second crystallization results in purer crystals, but the yield is lower.

In some experiments in this book, you will be instructed to cool the crystallizing mixture in an ice-water bath before collecting the crystals by filtration. Cooling the mixture increases the yield by decreasing the solubility of the substance; however, even at this reduced temperature, some of the product will be soluble in the solvent. It is not possible to recover all your product in a crystallization procedure.

5.3 STANDARD-SCALE CRYSTALLIZATION—BÜCHNER FUNNEL

The crystallization technique described in this section is used when the weight of solid to be crystallized is more than 0.1 g. The four main steps in a standard-scale crystallization are

1. Dissolving the solid
2. Removing insoluble impurities (when necessary)
3. Crystallization
4. Isolation of crystals

These steps are illustrated in Figure 5.3. It should be pointed out that a microscale crystallization with a Craig tube involves the same four steps, although the apparatus and procedures are somewhat different (see Section 5.4).

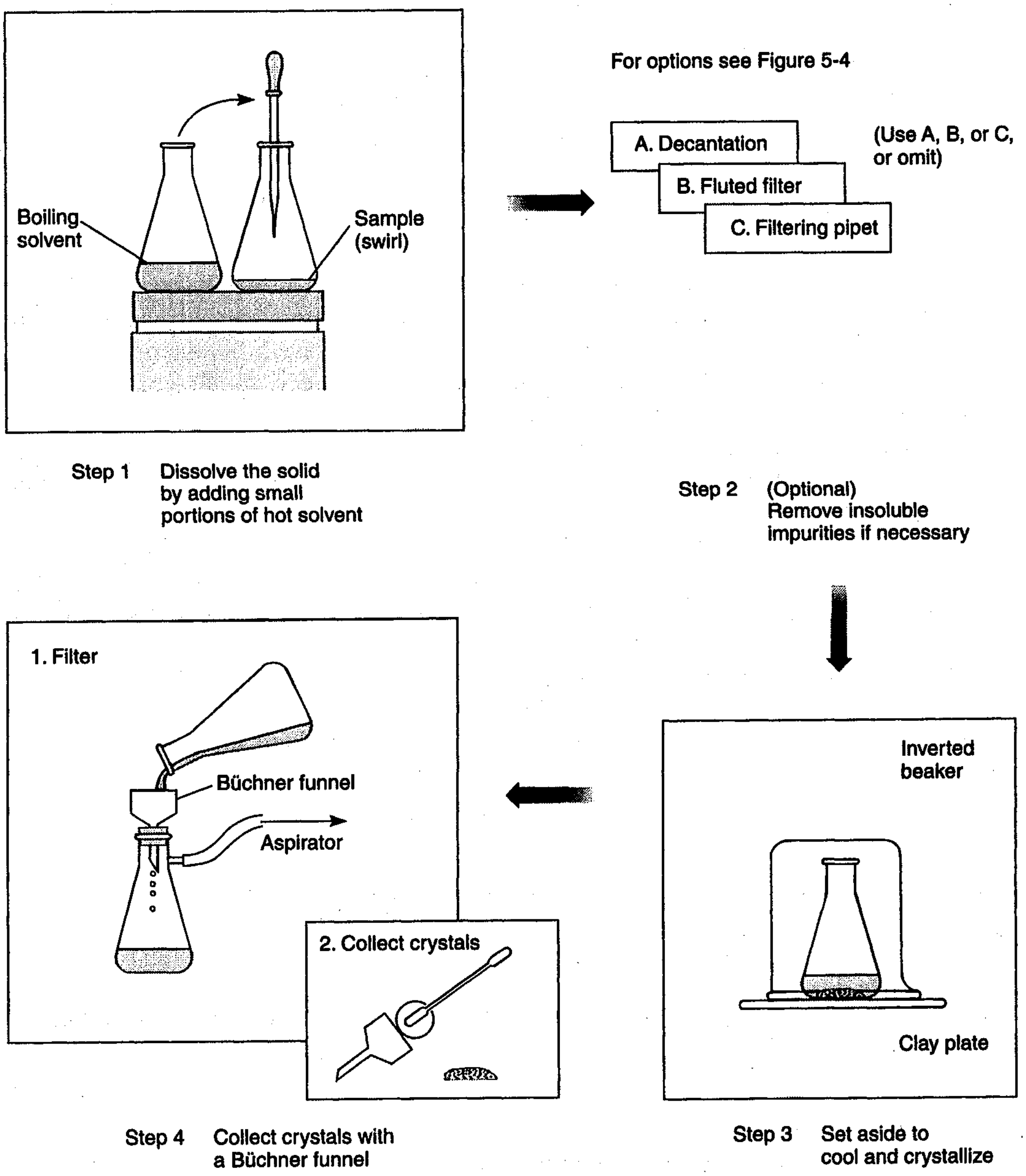
Boiling solvent
Sample (swirl)
For options see Figure 5-4
A. Decantation
B. Fluted filter
C. Filtering pipet
(Use A, B, or C, or omit)
Step 1 Dissolve the solid by adding small portions of hot solvent
Step 2 (Optional) Remove insoluble impurities if necessary
Inverted beaker
Clay plate
Step 3 Set aside to cool and crystallize
1. Filter
Büchner funnel
Aspirator
2. Collect crystals
Step 4 Collect crystals with a Büchner funnel

Figure 5.3 Steps in a standard-scale cystallization (no decolorization).

A. DISSOLVING THE SOLID

To minimize losses of material to the mother liquor, it is desirable to *saturate* the boiling solvent with solute. This solution, when cooled, will return the maximum possible amount of solute as crystals. To achieve this high return, the solvent is brought to its boiling point, and the solute is dissolved in the *minimum amount (!) of boiling solvent.* For this procedure, it is advisable to maintain a container of boiling solvent (on a hotplate). From this container, a small portion (about 1–2 mL) of the solvent is added to the Erlenmeyer flask containing the solid to be crystallized, and this mixture is heated while swirling occasionally until it resumes boiling. Do not heat the flask containing the solid until after you have added the first portion of solvent. If the solid does not dissolve in the first portion of boiling solvent, then another small portion of boiling solvent is added to the flask. The mixture is heated again until it resumes boiling. If the solid dissolves, no more solvent is added. But if the solid has not dissolved, another portion of boiling solvent is added, as before, and the process is repeated until the solid dissolves. It is important to stress that the portions of solvent added each time are small, so that only the *minimum* amount of solvent necessary for dissolving the solid is added. It is also important to emphasize that the procedure requires the addition of solvent to solid. You must never add portions of solid to a fixed quantity of boiling solvent. By this latter method, it may be impossible to tell when saturation has been achieved.

In many of the experiments in this textbook, a specified amount of solvent for a given weight of solid will be recommended. In these cases, you should use the amount specified rather than the minimum amount of solvent necessary for dissolving the solid. The amount of solvent recommended has been selected to provide the optimum conditions for good crystal formation.

Occasionally, you can encounter an impure solid that contains small particles of insoluble impurities, pieces of dust, or paper fibers that will not dissolve in the hot crystallizing solvent. A common error is to add too much of the hot solvent in an attempt to dissolve these small particles, without realizing that they are not soluble. In such cases, you must be careful not to add too much solvent.

It is sometimes necessary to decolorize the solution by adding activated charcoal or by passing the solution through a column containing alumina or silica gel (see Section 5.6, Parts A and B, and Technique 12, Section 12.14, p. 782). (Note: Often, there may be a small amount of colored material that will remain in solution during the crystallization step. When you believe that this may be the case, omit the decolorizing step.)

B. REMOVING INSOLUBLE IMPURITIES

It is necessary to use one of the following three methods only if insoluble material remains in the hot solution or if decolorizing charcoal has been used.

Caution: Indiscriminate use of the procedure can lead to needless loss of your product.

Decantation is the easiest method of removing solid impurities and should be considered first. If filtration is required, a filtering pipet is used when the volume of liquid to be filtered is less than 10 mL (see Technique 4, Section 4.1, Part C, p. 638), and you should use gravity filtration through a fluted filter when the volume is 10 mL or greater (see Technique 4, Section 4.1, Part B, p. 636). These three methods are illustrated in Figure 5.4, and each one is discussed below.

Decantation. If the solid particles are relatively large in size or they easily settle to the bottom of the flask, it may be possible to separate the hot solution from the impurities by carefully pouring off the liquid, leaving the solid behind. This is accomplished most easily by holding a glass stirring rod along the top of the flask and tilting the flask so that the liquid pours out along one end of the glass rod into another container. A technique similar in principle to decantation, which may be easier to perform with smaller amounts of liquid, is to use a **preheated Pasteur pipet** to remove the hot solution. With this method, it may be helpful to place the tip of the pipet against the bottom of the flask when removing the last portion of solution. The small space between the tip of the pipet and the inside surface of the flask prevents solid material from being drawn into the pipet. An easy way to preheat the pipet is to draw up a small portion of hot *solvent* (not the *solution* being transferred) into the pipet and expel the liquid. Repeat this process several times.

Fluted Filter. This method is the most effective way to remove solid impurities when the volume of liquid is greater than 10 mL or when decolorizing charcoal has been used (see Technique 4, Section 4.1, Part B, p. 636). You should first add a small amount of extra solvent to the hot mixture. This action helps to prevent crystal formation in the filter paper or the stem of the funnel during the filtration. The funnel is then fitted with a fluted filter and installed at the top of the Erlenmeyer flask to be used for the actual filtration. It is advisable to place a small piece of wire between the funnel and the mouth of the flask to relieve any increase in pressure caused by hot filtrate.

The Erlenmeyer flask containing the funnel and fluted paper is placed on top of a hotplate (low setting). The liquid to be filtered is brought to its boiling point and poured through the filter in portions. (If the volume of the mixture is less than 10 mL, it may be more convenient to transfer the mixture to the filter with a preheated Pasteur pipet.) It is necessary to keep the solutions in both flasks at their boiling temperatures to prevent premature crystallization. The refluxing action of the filtrate keeps the funnel warm and reduces the chance that the filter will clog with crystals that may have formed during the filtration. With low-boiling solvents, be aware that some solvent may be lost through evaporation. Consequently, extra solvent must be added to make up for this loss. If crystals begin to form in the filter during filtration, a minimum amount of boiling solvent is added to redissolve the crystals and to allow the solution to pass through the funnel. If the volume of liquid being filtered is less than 10 mL, a small amount of hot solvent should be used to rinse the filter after all the filtrate has been collected. The rinse solvent is then combined with the original filtrate.

After the filtration, it may be necessary to remove extra solvent by evaporation until the solution is once again saturated at the boiling point of the solvent (see Technique 3, Section 3.11, p. 630).

Filtering Pipet. If the volume of solution after dissolving the solid in hot solvent is less than 10 mL, gravity filtration with a filtering pipet may be used to remove solid

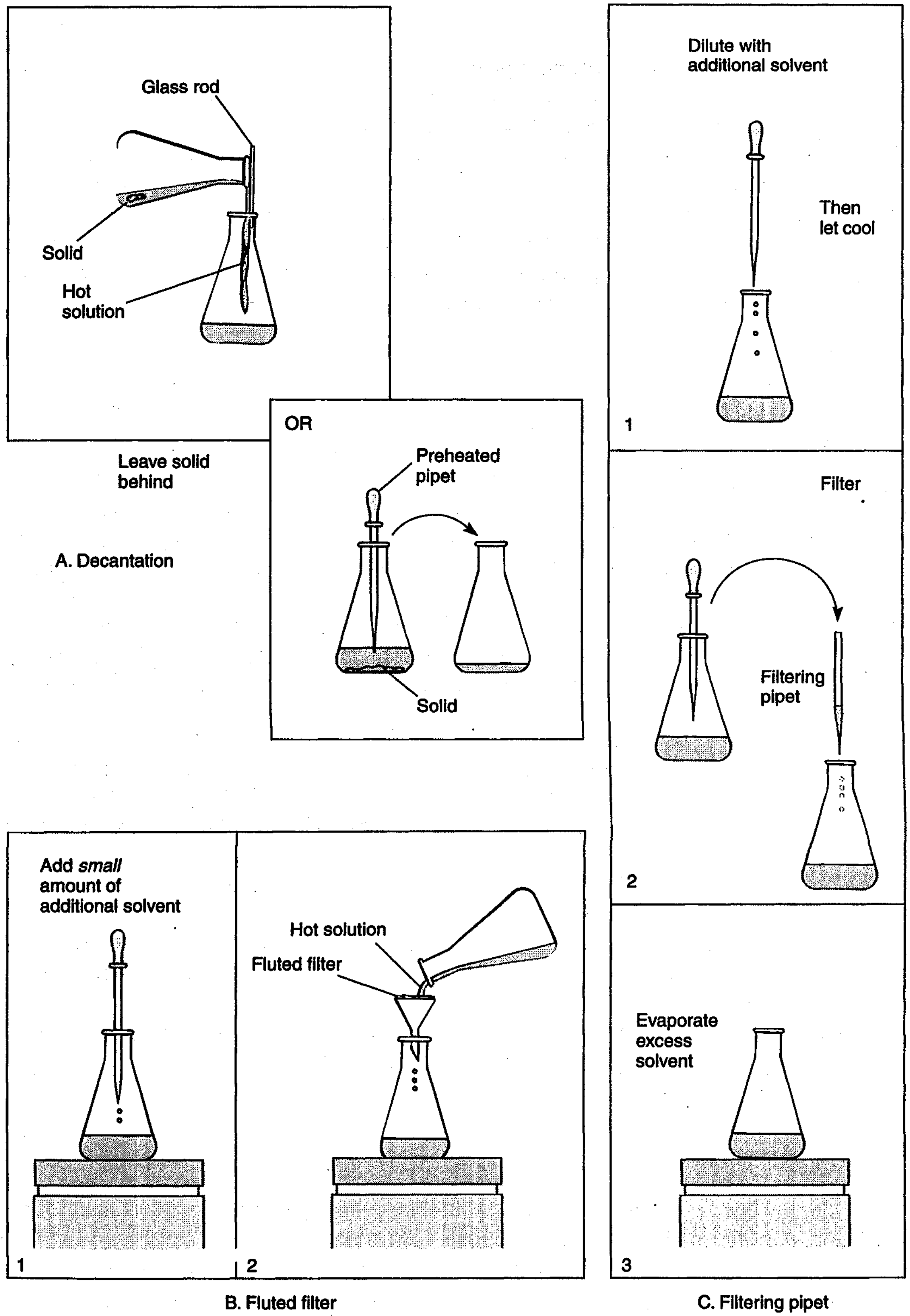
Glass rod
Solid
Hot solution
Leave solid behind
OR
Preheated pipet
Solid
A. Decantation
Dilute with additional solvent
Then let cool
1
Filter
Filtering pipet
2
Add *small* amount of additional solvent
1
Hot solution
Fluted filter
2
Evaporate excess solvent
3
B. Fluted filter
C. Filtering pipet

Figure 5.4 Methods for removing insoluble impurities in a semi-microscale crystallization.

impurities. However, using a filtering pipet to filter a hot solution saturated with solute can be difficult without premature crystallization. The best way to prevent this from occurring is to add enough solvent to dissolve the desired product at room temperature (be sure not to add too much solvent) and perform the filtration at room temperature, as described in Technique 4, Section 4.1, Part C, p. 638. After filtration, the excess solvent is evaporated by boiling until the solution is saturated at the boiling point of the mixture (see Technique 3, Section 3.11, p. 630). If powdered decolorizing charcoal was used, it will probably be necessary to perform two filtrations with a filtering pipet to remove all of the charcoal, or else a fluted filter can be used.

C. CRYSTALLIZATION

An Erlenmeyer flask, not a beaker, should be used for crystallization. The large open top of a beaker makes it an excellent dust catcher. The narrow opening of the Erlenmeyer flask reduces contamination by dust and allows the flask to be stoppered if it is to be set aside for a long period. Mixtures set aside for long periods must be stoppered after cooling to room temperature to prevent evaporation of solvent. If all the solvent evaporates, no purification is achieved, and the crystals originally formed become coated with the dried contents of the mother liquor. Even if the time required for crystallization to occur is relatively short, it is advisable to cover the top of the Erlenmeyer flask with a small watch glass or inverted beaker to prevent evaporation of solvent while the solution is cooling to room temperature.

The chances of obtaining pure crystals are improved if the solution cools to room temperature slowly. When the volume of solution is 10 mL or less, the solution is likely to cool more rapidly than is desired. This can be prevented by placing the flask on a surface that is a poor heat conductor and covering the flask with a beaker to provide a layer of insulating air. Appropriate surfaces include a clay plate or several pieces of filter paper on top of the laboratory bench. It may also be helpful to use a clay plate that has been warmed slightly on a hotplate or in an oven.

After crystallization has occurred, it is sometimes desirable to cool the flask in an ice-water bath. Because the solute is less soluble at lower temperatures, this will increase the yield of crystals.

If a cooled solution does not crystallize, it will be necessary to induce crystallization. Several techniques are described in Section 5.7, Part A.

D. ISOLATION OF CRYSTALS

After the flask has been cooled, the crystals are collected by vacuum filtration through a Büchner (or Hirsch) funnel (see Technique 4, Section 4.3, p. 640, and Fig. 4.5). The crystals should be washed with a small amount of *cold* solvent to remove any mother liquor adhering to their surface. Hot or warm solvent will dissolve some of the crystals. The crystals should then be left for a short time (usually 5–10 minutes) in the funnel, where air, as it passes, will dry them free of most of the solvent. It is often wise to cover the Büchner funnel with an oversize filter paper or towel during this air-drying. This precaution prevents accumulation of dust in the crystals. When the crystals are nearly dry, they should be gently scraped off (so paper fibers are not removed with the crystals) the filter paper onto a watch glass or clay plate for further drying (see Section 5.8).

The four steps in a standard scale crystallization are summarized in Figure 5.5

A. Dissolving the Solid

1. Find a solvent with a steep solubility-vs-temperature characteristic. (Done by trial and error using small amounts of material or by consulting a handbook.)
2. Heat the desired solvent to its boiling point.
3. Dissolve the solid in a **minimum** of boiling solvent in a flask.
4. If necessary, add decolorizing charcoal or decolorize the solution on a silica gel or alumina column.

B. Removing Insoluble Impurities

1. Decant or remove the solution with a Pasteur pipet, or
2. Filter the hot solution through a fluted filter, a filtering pipet, or a filter tip pipet to remove insoluble impurities or charcoal.

NOTE: If no decolorizing charcoal has been added or if there are no undissolved particles, Part B should be omitted.

C. Crystallizing

1. Allow the solution to cool.
2. If crystals appear, cool the mixture in an ice-water bath (if desired) and go to Part D. If crystals do not appear, go to the next step.
3. Inducing crystallization.
   - (a) Scratch the flask with a glass rod.
   - (b) Seed the solution with original solid, if available.
   - (c) Cool the solution in an ice-water bath.
   - (d) Evaporate excess solvent and allow the solution to cool again.

D. Collecting and Drying

1. Collect crystals by vacuum filtration using a Büchner funnel.
2. Rinse crystals with a small portion of **cold** solvent.
3. Continue suction until crystals are nearly dry.
4. Drying.
   - (a) Air-dry the crystals, or
   - (b) Place the crystals in a drying oven, or
   - (c) Dry the crystals *in vacuo.*

Figure 5.5 Steps in a crystallization.

5.4 MICROSCALE CRYSTALLIZATION—CRAIG TUBE

In most microscale experiments, the amount of solid to be crystallized is small enough (generally less than 0.1 g) that a **Craig tube** (see Technique 4, Fig. 4.10, p. 646) is the preferred method for crystallization. The main advantage of the Craig tube is that it minimizes the number of transfers of solid material, thus resulting in a greater yield of crystals. Also, the separation of the crystals from the mother liquor with the Craig tube is very efficient, and little time is required for drying the crystals. The steps involved are, in principle, the same as those performed when a crystallization is accomplished with an Erlenmeyer flask and a Büchner funnel.

The solid is transferred to the Craig tube, and small portions of hot solvent are added to the tube while the mixture is stirred with a spatula and heated. If there are any insoluble impurities present, they can be removed with a filter-tip pipet. The inner plug is then inserted into the Craig tube and the hot solution is cooled slowly to room temperature. When the crystals have formed, the Craig tube is placed into a centrifuge tube and the crystals are separated from the mother liquor by centrifugation (see Technique 4, Section 4.7, p. 646). The crystals are then scraped off the end of the inner plug or from inside the Craig tube onto a watch glass or piece of paper. Minimal drying will be necessary (see Section 5.8).

5.5 SELECTING A SOLVENT

A solvent that dissolves little of the material to be crystallized when it is cold but a great deal of the material when it is hot is a good solvent for the crystallization. With compounds that are well known, such as the compounds that are either isolated or prepared in this textbook, the correct crystallization solvent is already known through the experiments of earlier researchers. In such cases, the chemical literature can be consulted to determine which solvent should be used. Sources such as handbooks or tables frequently provide this information. Quite often, the correct crystallization solvents are indicated in the experimental procedures in this textbook.

When the appropriate solvent is not known, select a solvent for crystallization by experimenting with various solvents and a very small amount of the material to be crystallized. Experiments are conducted on a small test-tube scale before the entire quantity of material is committed to a particular solvent. Such trial-and-error methods are common when one is trying to purify a solid material that has not been previously studied.

When choosing a crystallization solvent, take care not to pick one whose boiling point is higher than the melting point of the substance to be crystallized. If the boiling point of the solvent is high, the solid may melt in the solvent rather than dissolve. In such a case, the solid may **oil out.** Oiling out occurs when the solid substance melts and forms a liquid that is not soluble in the solvent. On cooling, the liquid refuses to crystallize; rather, it becomes a supercooled liquid, or oil. Oils may solidify if the temperature is lowered, but often they will not crystallize. A solidified oil becomes an amorphous solid or a hardened mass—a condition that does not result in the purification of the substance. It can be very difficult to deal with oils when trying to obtain a pure substance. You must try to redissolve them and hope that they will precipitate as crystals with slow, careful cooling. During the cooling period, it may be helpful to scratch the glass container where the oil is present with a glass stirring rod that has not been fire-polished. Seeding the oil as it cools with a small sample of the original solid is another technique sometimes helpful in working with difficult oils. Other methods of inducing crystallization are discussed in Section 5.7.

One additional criterion for selecting the correct crystallization solvent is the **volatility** of that solvent. Volatile solvents are those that have low boiling points or evaporate easily. A solvent with a low boiling point may be removed from the crystals through evaporation without much difficulty. It will be difficult to remove a solvent with a high boiling point from the crystals without heating them under vacuum.

Table 5.2 lists common crystallization solvents. The solvents used most commonly are listed first in the table.

5.6 DECOLORIZATION

Small amounts of highly colored impurities may make the original crystallization solution appear colored; this color can often be removed by **decolorization,** either by using activated charcoal (often called Norit) or by passing the solution through a column packed with alumina or silica gel. A decolorizing step should be performed only if the color is due to impurities, not to the color of the desired product, and if the color is significant. Small amounts of colored impurities will remain in solution during crystalliza-

TABLE 5.2 Common Solvents for Crystallization

	Boils (°C)	Freezes (°C)	Soluble in H_2O	Flammability
Water	100	0	+	–
Methanol	65	*	+	+
95% Ethanol	78	*	+	+
Ligroin	60–90	*	–	+
Toluene	111	*	–	+
Chloroform†	61	*	–	–
Acetic acid	118	17	+	+
Dioxane†	101	11	+	+
Acetone	56	*	+	+
Diethyl ether	35	*	Slightly	++
Petroleum ether	30–60	*	–	++
Methylene chloride	41	*	–	–
Carbon tetrachloride†	77	*	–	–

*Lower than 0°C.

†Suspected carcinogen.

tion, making the decolorizing step unnecessary. The use of activated charcoal is described separately for standard scale and microscale crystallizations, and the column technique, which can be used with both crystallization techniques, is also described.

PART A. STANDARD-SCALE—POWDERED CHARCOAL

As soon as the solute is dissolved in the minimum amount of boiling solvent, the solution is allowed to cool slightly and a small amount of Norit (powdered charcoal) is added to the mixture. The Norit adsorbs the impurities. When performing a crystallization in which the filtration is performed with a fluted filter, powdered Norit should be added because it has a larger surface area and can remove impurities more effectively. A reasonable amount of Norit would be what could be held on the end of a small spatula, or about 0.1–0.2 g. If too much Norit is used, it will adsorb product as well as impurities. A small amount of Norit should be used, and its use should be repeated if necessary. (It is difficult to determine if the initial amount added is sufficient until after the solution is filtered, because the suspended particles of black charcoal will obscure the color of the liquid.) Caution should be exercised so that the solution does not froth or erupt when the finely divided charcoal is added. The mixture is boiled for several minutes and then filtered by gravity, using a fluted filter (see Section 5.3 and Technique 4, Section 4.1, Part B, p. 636), and the crystallization is carried forward as described in Section 5.3.

The Norit preferentially absorbs the colored impurities and removes them from the solution. The technique seems to be most effective with hydroxylic solvents. In using Norit, be careful not to breathe the dust. Normally, small quantities are used so that little risk of lung irritation exists.

B. DECOLORIZATION ON A COLUMN

Another method for decolorizing a solution is to pass the solution through a column containing alumina or silica gel. The adsorbent removes the colored impurities while allowing the desired material to pass through (see Technique 4, Fig. 4.6, p. 643, and Technique 12, Section 12.14, p. 782). If this technique is used, it will be necessary to dilute the solution with additional solvent to prevent crystallization from occurring during the process. The excess solvent must be evaporated after the solution is passed through the column (Technique 3, Section 3.11, p. 630), and the crystallization procedure is continued as described in Section 5.3.

C. MICROSCALE—PELLETIZED NORIT

If the crystallization is being performed in a Craig tube, it is advisable to use pelletized Norit. Although this is not as effective in removing impurities as powdered Norit, it is easier to carry out the subsequent filtration, and the amount of pelletized Norit required is more easily determined because you can see the solution as it is being decolorized. Again, the Norit is added to the hot solution (the solution should not be boiling) after the solid has dissolved. This should be performed in a test tube rather than in a Craig tube. About 0.02 g is added, and the mixture is boiled for a minute or so to see if more Norit is required. More Norit is added, if necessary, and the liquid is boiled again. It is important not to add too much pelletized Norit because the Norit will also adsorb some of the desired material, and it is possible that not all the color can be removed no matter how much is added. The decolorized solution is then removed with a preheated filter-tip pipet (see Section 5.4 and Technique 4, Section 4.6, p. 645) to filter the mixture and transferred to a Craig tube for crystallization as described in Section 5.4.

5.7 INDUCING CRYSTALLIZATION

If a cooled solution does not crystallize, several techniques may be used to induce crystallization. Although identical in principle, the actual procedures vary slightly when performing standard scale and microscale crystallizations.

A. STANDARD-SCALE

In the first technique, you should try scratching the inside surface of the flask vigorously with a glass rod that *has not been* fire-polished. The motion of the rod should be vertical (in and out of the solution) and should be vigorous enough to produce an audible scratching. Such scratching often induces crystallization, although the effect is not well understood. The high-frequency vibrations may have something to do with initiating crystallization; or perhaps—a more likely possibility—small amounts of solution dry by evaporation on the side of the flask, and the dried solute is pushed into the solution. These small amounts of material provide "seed crystals," or nuclei, on which crystallization may begin.

A second technique that can be used to induce crystallization is to cool the solution in an ice bath. This method decreases the solubility of the solute.

A third technique is useful when small amounts of the original material to be crystallized are saved. The saved material can be used to "seed" the cooled solution. A small crystal dropped into the cooled flask often will start the crystallization—this is called **seeding.**

If all these measures fail to induce crystallization, it is likely that too much solvent was added. The excess solvent must then be evaporated (Technique 3, Section 3.11, p. 630) and the solution allowed to cool.

B. MICROSCALE

The strategy is basically the same as described for standard scale crystallizations. Scratching vigorously with a glass rod *should be avoided,* however, because the Craig tube is fragile and expensive. Scratching *gently* is allowed.

Another measure is to dip a spatula or glass stirring rod into the solution and allow the solvent to evaporate so that a small amount of solid will form on the surface of the spatula or glass rod. When placed back into the solution, the solid will seed the solution. A small amount of the original material, if some was saved, may also be used to seed the solution.

A third technique is to cool the Craig tube in an ice-water bath. This method may also be combined with either of the previous suggestions.

If none of these measures is successful, it is possible that too much solvent is present, and it may be necessary to evaporate some of the solvent (Technique 3, Section 3.11, p. 630) and allow the solution to cool again.

5.8 DRYING CRYSTALS

The most common method of drying crystals involves placing them on a watch glass, a clay plate, or a piece of paper and allowing them to dry in air. While the advantage of this method is that heat is not required, thus reducing the danger of decomposition or melting, exposure to atmospheric moisture may cause the hydration of strongly **hygroscopic** materials. A hygroscopic substance is one that absorbs moisture from the air. When crystals are allowed to air-dry, they should be covered with a watch glass or inverted beaker to protect them from air-borne dust particles.

Another method of drying crystals is to place the crystals on a watch glass, a clay plate, or a piece of absorbent paper in an oven. Although this method is simple, some possible difficulties deserve mention. Crystals that sublime readily should not be dried in an oven because they might vaporize and disappear. Care should be taken that the temperature of the oven does not exceed the melting point of the crystals. Remember that the melting point of crystals is lowered by the presence of solvent; allow for this melting-point depression when selecting a suitable oven temperature. Some materials decompose on exposure to heat, and they should not be dried in an oven. Finally, when many different samples are being dried in the same oven, crystals might be lost due to confusion or reaction with another person's sample. It is important to label the crystals when they are placed in the oven.

A third method, which requires neither heat nor exposure to atmospheric moisture, is drying *in vacuo.* Two procedures are illustrated in Figure 5.6.

Procedure A. In this method a desiccator is used. The sample is placed under vacuum in the presence of a drying agent. Two potential problems must be noted. The first deals with samples that sublime readily. Under vacuum, the likelihood of sublimation is increased. The second problem deals with the vacuum desiccator itself. Because the surface area of glass that is under vacuum is large, there is some danger that the desiccator could implode. A vacuum desiccator should never be used unless it has been placed within a protective metal container (cage). If a cage is not available, the desiccator can be wrapped with electrical or duct tape. If you use an aspirator as a source of vacuum, you should use a water trap (see Technique 4, Fig. 4.5, p. 641).

Procedure B. This method can be accomplished with a round-bottom flask and a thermometer adapter equipped with a short piece of glass tubing, as illustrated in Figure 5.6B. The glass tubing is connected by vacuum tubing to either an aspirator or a vacuum pump. A convenient alternative using a sidearm test tube is also shown in Figure 5.6B. With either apparatus, install a water trap when an aspirator is used.

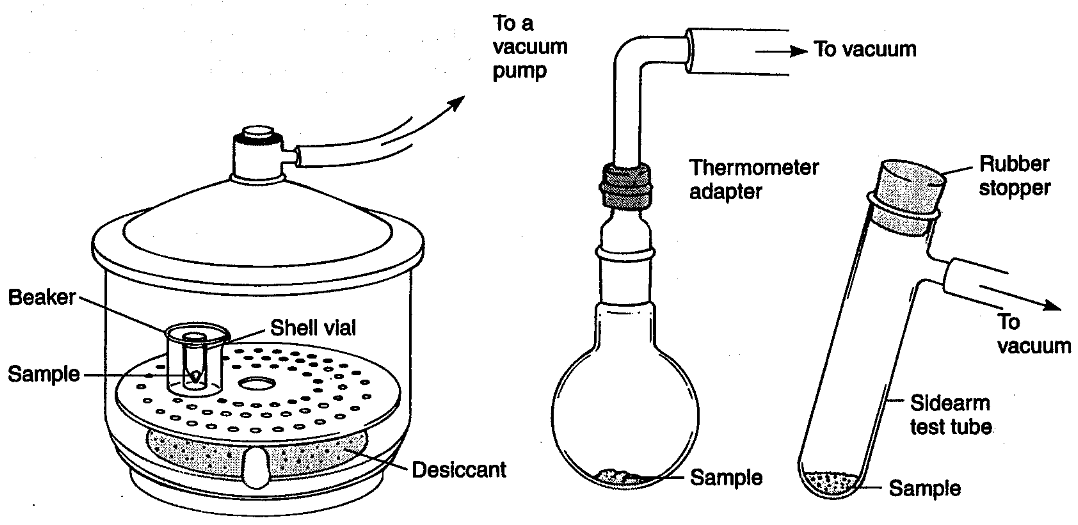
To a
vacuum
pump
Beaker
Shell vial
Sample
Desiccant
To vacuum
Thermometer
adapter
Sample
Rubber
stopper
To
vacuum
Sidearm
test tube
Sample

A. Desiccator.

B. Round-bottom flask or sidearm test tube.

Figure 5.6 Methods for drying crystals in a vacuum.

5.9 MIXED SOLVENTS

Often, the desired solubility characteristics for a particular compound are not found in a single solvent. In these cases, a mixed solvent may be used. You simply select a first solvent in which the solute is soluble and a second solvent, miscible with the first, in which the solute is relatively insoluble. The compound is dissolved in a minimum amount of the boiling solvent in which it is soluble. Following this, the second hot solvent is added to the boiling mixture, dropwise, until the mixture barely becomes cloudy. The cloudiness indicates precipitation. At this point, more of the first solvent should be added. Just enough is added to clear the cloudy mixture. At that point, the solution is saturated, and as it cools, crystals should separate. Common solvent mixtures are listed in Table 5.3.

TABLE 5.3. Common Solvent Pairs for Crystallization

Methanol–Water	Ether–Acetone
Ethanol–Water	Ether–Petroleum ether
Acetic acid–Water	Toluene–Ligroin
Acetone–Water	Methylene chloride–Methanol
Ether–Methanol	Dioxane*–Water

*Suspected carcinogen.

It is important not to add an excess of the second solvent or to cool the solution too rapidly. Either of these actions may cause the solute to oil out, or separate as a viscous liquid. If this happens, reheat the solution and add more of the first solvent.

PROBLEMS

1. Listed below are solubility-vs.-temperature data for an organic substance A dissolved in water.

Temperature (°C)	**Solubility of A in 100 mL of Water**
0	1.5 g
20	3.0 g
40	6.5 g
60	11.0 g
80	17.0 g

(a) Graph the solubility of A vs. temperature. Use the data given in the table. Connect the data points with a smooth curve.

(b) Suppose 0.1 g of A and 1.0 mL of water were mixed and heated to 80°C. Would all the substance A dissolve?

(c) The solution prepared in (b) is cooled. At what temperature will crystals of A appear?

(d) Suppose the cooling described in (c) were continued to 0°C. How many grams of A would come out of solution? Explain how you obtained your answer.

2. What would be likely to happen if a hot saturated solution were filtered by vacuum filtration using a Büchner funnel? (*Hint:* The mixture will cool as it comes in contact with the Büchner funnel.)

3. A compound you have prepared is reported in the literature to have a pale yellow color. When dissolving the substance in hot solvent to purify it by crystallization, the resulting solution is yellow. Should you use decolorizing charcoal before allowing the hot solution to cool? Explain your answer.

4. While performing a crystallization, you obtain a light tan solution after dissolving your crude product in hot solvent. A decolorizing step is determined to be unnecessary, and there are no solid impurities present. Should you perform a filtration to remove impurities before allowing the solution to cool? Why or why not?

5. (a) Draw a graph of a cooling curve (temperature vs. time) for a solution of a solid substance that shows no supercooling effects. Assume that the solvent does not freeze.

(b) Repeat the instructions in (a) for a solution of a solid substance that shows some supercooling behavior but eventually yields crystals if the solution is cooled sufficiently.

6. A solid substance A is soluble in water to the extent of 1 g/100 mL of water at 25°C and 10 g/100 mL of water at 100°C. You have a sample that contains 10 g of A and an impurity B.

(a) Assuming that 0.2 g of the impurity B is present along with 10 g of A, describe how you could purify A if B is completely insoluble in water.

(b) Assuming that 0.2 g of the impurity B is present along with 10 g of A, describe how you could purify A if B had the same solubility behavior as A. Would one crystallization produce absolutely pure A?

(c) Assume that 3 g of the impurity B is present along with 10 g of A. Describe how you would purify A if B had the same solubility behavior as A. Each time, use the correct amount of water to just dissolve the solid. Would one crystallization produce absolutely pure A? How many crystallizations would be needed to produce pure A? How much A would have been recovered when the crystallizations had been completed?

TECHNIQUE 6

Physical Constants: Melting Points, Boiling Points, Density

6.1 PHYSICAL PROPERTIES

The physical properties of a compound are those properties that are intrinsic to a given compound when it is pure. Often, a compound may be identified simply by determining a number of its physical properties. The most commonly recognized physical properties of a compound include its color, melting point, boiling point, density, refractive index, molecular weight, and optical rotation. Modern chemists would include the various types of spectra (infrared, nuclear magnetic resonance, mass, and ultraviolet-visible) among the physical properties of a compound. A compound's spectra do not vary from one pure sample to another. In this chapter, we look at methods of determining the melting point, boiling point, and density of compounds. Refractive index, optical rotation, and spectra are considered separately in their own technique chapters.

Many reference books list the physical properties of substances. Useful works for finding lists of values for the nonspectroscopic physical properties include:

The Merck Index
The CRC Handbook of Chemistry and Physics
The Dictionary of Organic Compounds
Lange's Handbook of Chemistry
CRC Handbook of Tables for Organic Compound Identification

Complete citations for these references may be found in Technique 20 (Guide to the Chemical Literature, p. 861). Although the *CRC Handbook* has very good tables, it adheres strictly to IUPAC nomenclature. For this reason, it may be easier to use one of the other references, particularly *The Merck Index,* in your first attempt to locate information. *The Dictionary of Organic Compounds* is a multivolume work. A trip to the reference shelves of your library is required for you to use it, but it is a very complete source book.

PART A. MELTING POINTS

6.2 THE MELTING POINT

The melting point of a compound is used by the organic chemist not only to identify it, but also to establish its purity. A small amount of material is heated *slowly* in a special apparatus equipped with a thermometer or thermocouple, a heating bath or heating coil, and a magnifying eyepiece for observing the sample. Two temperatures are noted. The first is the point at which the first drop of liquid forms among the crystals; the second is the point at which the whole mass of crystals turns to a *clear* liquid. The melting point is recorded by giving this range of melting. You might say, for example, that the melting point of a substance is 51–54°C. That is, the substance melted over a 3° range.

The melting point indicates purity in two ways. First, the purer the material, the higher its melting point. Second, the purer the material, the narrower the melting point range. Adding successive amounts of an impurity to a pure substance generally causes its melting point to decrease in proportion to the amount of impurity. Looking at it another way, adding impurities lowers the freezing point. The freezing point, a colligative property, is simply the melting point (solid ⟶ liquid) approached from the opposite direction (liquid ⟶ solid).

Figure 6.1 is a graph of the usual melting-point behavior of mixtures of two substances, A and B. The two extremes of the melting range (the low and high temperatures) are shown for various mixtures of the two. The upper curves indicate the temperature at which all the sample has melted. The lower curves indicate the temperature at which melting is observed to begin. With pure compounds, melting is sharp and without any range. This is shown at the left- and right-hand edges of the graph. If you begin with pure A, the melting point decreases as impurity B is added. At some point, a minimum temperature, or **eutectic,** is reached, and the melting point begins to increase to that of substance B. The vertical distance between the lower and upper curves represents the melting range. Notice that for mixtures that contain relatively small amounts of impurity (<15%) and

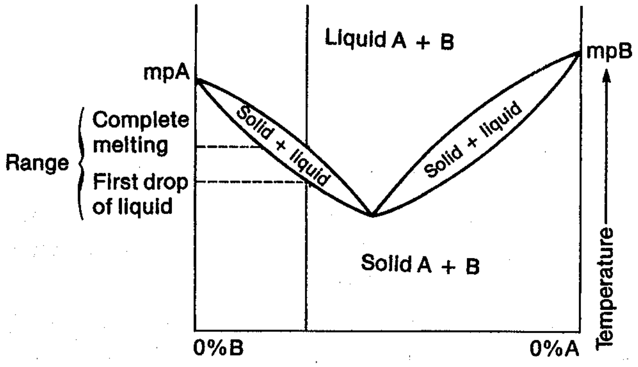

Figure 6.1 Melting point–composition curve.

are not close to the eutectic, the melting range increases as the sample becomes less pure. The range indicated by the lines in Figure 6.1 represents the typical behavior.

We can generalize the behavior shown in Figure 6.1. Pure substances melt with a narrow range of melting. With impure substances, the melting range becomes wider, and the entire melting range is lowered. Be careful to note, however, that at the minimum point of the melting point–composition curves, the mixture often forms a eutectic, which also melts sharply. Not all binary mixtures form eutectics, and some caution must be exercised in assuming that every binary mixture follows the previously described behavior. Some mixtures may form more than one eutectic, others might not form even one. In spite of these variations, both the melting point and its range are useful indications of purity, and they are easily determined by simple experimental methods.

6.3 MELTING POINT THEORY

Figure 6.2 is a phase diagram describing the usual behavior of a two-component mixture (A + B) on melting. The behavior on melting depends on the relative amounts of A and B in the mixture. If A is a pure substance (no B), then A melts sharply at its melting point t_A. This is represented by point A on the left side of the diagram. When B is a pure substance, it melts at t_B; its melting point is represented by point B on the right side of the diagram. At either point A or point B, the pure solid passes cleanly, with a narrow range, from solid to liquid.

In mixtures of A and B, the behavior is different. Using Figure 6.2, consider a mixture of 80% A and 20% B on a mole-per-mole basis (that is, mole percentage). The melting point of this mixture is given by t_M at point M on the diagram. That is, adding B to A has lowered the melting point of A from t_A to t_M. It has also expanded the melting range. The temperature t_M corresponds to the **upper limit** of the melting range.

Lowering the melting point of A by adding impurity B comes about in the following way. Substance A has the lower melting point in the phase diagram shown, and if

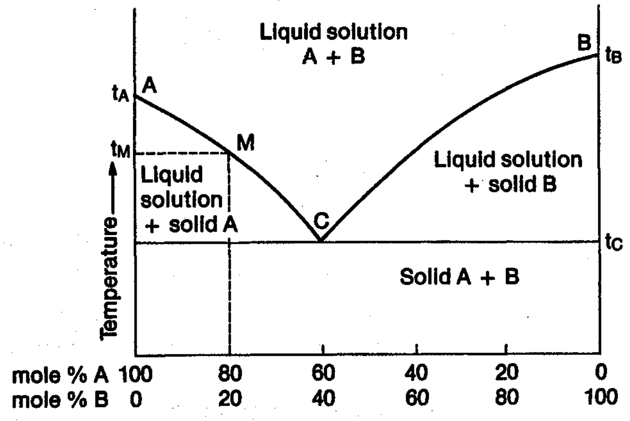

Figure 6.2 Phase diagram for melting in a two-component system.

heated, it begins to melt first. As A begins to melt, solid B begins to dissolve in the liquid A that is formed. When solid B dissolves in liquid A, the melting point is depressed. To understand this, consider the melting point from the opposite direction. When a liquid at a high temperature cools, it reaches a point at which it solidifies, or "freezes." The temperature at which a liquid freezes is identical to its melting point. Recall that the freezing point of a liquid can be lowered by adding an impurity. Because the freezing point and the melting point are identical, lowering the freezing point corresponds to lowering the melting point. Therefore, as more impurity is added to a solid, its melting point becomes lower. There is, however, a limit to how far the melting point can be depressed. You cannot dissolve an infinite amount of the impurity substance in the liquid. At some point, the liquid will become saturated with the impurity substance. The solubility of B in A has an upper limit. In Figure 6.2, the solubility limit of B in liquid A is reached at point C, the **eutectic point.** The melting point of the mixture cannot be lowered below t_C, the melting temperature of the eutectic.

Now consider what happens when the melting point of a mixture of 80% A and 20% B is approached. As the temperature is increased, A begins to "melt." This is not really a visible phenomenon in the beginning stages; it happens before liquid is visible. It is a softening of the compound to a point at which it can begin to mix with the impurity. As A begins to soften, it dissolves B. As it dissolves B, the melting point is lowered. The lowering continues until all B is dissolved or until the eutectic composition (saturation) is reached. When the maximum possible amount of B has been dissolved, actual melting begins, and one can observe the first appearance of liquid. The initial temperature of melting will be below t_A. The amount below t_A at which melting begins is determined by the amount of B dissolved in A, but will never be below t_C. Once all B has been dissolved, the melting point of the mixture begins to rise as more A begins to melt. As more A melts, the semisolid solution is diluted by more A, and its melting point rises. While all this is happening, you can observe *both* solid and liquid in the melting-point capillary. Once all A has begun to melt, the composition of the mixture M becomes uniform and will reach 80% A and 20% B. At this point, the mixture finally melts sharply, giving a clear solution. The maximum melting point will be t_M, because t_A is depressed by the impurity B that is present. The lower end of the melting range will always be t_C; however, melting will not always be observed at this temperature. An observable melting at t_C comes about only when a large amount of B is present. Otherwise, the amount of liquid formed at t_C will be too small to observe. Therefore, the melting behavior that is actually observed will have a smaller range, as shown in Figure 6.1.

6.4 MIXED MELTING POINTS

The melting point can be used as supporting evidence in identifying a compound in two different ways. Not only may the melting points of the two individual compounds be compared, but a special procedure called a **mixed melting point** may also be performed. The mixed melting point requires that an authentic sample of the same compound be available from another source. In this procedure, the two compounds (authentic and suspected) are finely pulverized and mixed together in equal quantities. Then the melting point of the mixture is determined. If there is a melting-point depression, or if the range of melting is expanded by a large amount, compared to the individual substances, you may conclude that one compound has acted as an impurity toward the other and that they are not the same compound. If there is no lowering of the melting point for the mixture (the melting point is identical with those of pure A and pure B), then A and B are almost certainly the same compound.

6.5 PACKING THE MELTING POINT TUBE

Melting points are usually determined by heating the sample in a piece of thin-walled capillary tubing (1 mm × 100 mm) that has been sealed at one end. To pack the tube, press the open end gently into a *pulverized* sample of the crystalline material. Crystals will stick in the open end of the tube. The amount of solid pressed into the tube should correspond to a column no more than 1–2 mm high. To transfer the crystals to the closed end of the tube, drop the capillary tube, closed end first, down a $\frac{2}{3}$-m length of glass tubing, which is held upright on the desk top. When the capillary tube hits the desk top, the crystals will pack down into the bottom of the tube. This procedure is repeated if necessary. Tapping the capillary on the desk top with fingers is not recommended, because it is easy to drive the small tubing into a finger if the tubing should break.

Some commercial melting-point instruments have a built-in vibrating device that is designed to pack capillary tubes. With these instruments, the sample is pressed into the open end of the capillary tube, and the tube is placed in the vibrator slot. The action of the vibrator will transfer the sample to the bottom of the tube and pack it tightly.

6.6 DETERMINING THE MELTING POINT—THE THIELE TUBE

There are two principal types of melting-point apparatus available: the Thiele tube and commercially available, electrically heated instruments. The Thiele tube, shown in Figure 6.3, is the simpler device. It is a glass tube designed to contain a heating oil (mineral oil or silicone oil) and a thermometer to which a capillary tube containing the sample is attached. The shape of the Thiele tube allows convection currents to form in the oil when it is heated. These currents maintain a uniform temperature distribution throughout the oil in the tube. The sidearm of the tube is designed to generate these convection currents and thus transfer the heat from the flame evenly and rapidly throughout the oil. The sample, which is in a capillary tube attached to the thermometer, is held by a rubber band or a thin slice of rubber tubing. It is important that this rubber band be above the level of the oil (allowing for expansion of the oil on heating), so that the oil does not soften the rubber and allow the capillary tubing to fall into the oil. If a cork or a rubber stopper is used to hold the thermometer, a triangular wedge should be sliced in it to allow pressure equalization.

The Thiele tube is usually heated by a microburner. During the heating, the rate of temperature increase should be regulated. Hold the burner by its cool base, and, using a low flame, move the burner slowly back and forth along the bottom of the arm of the

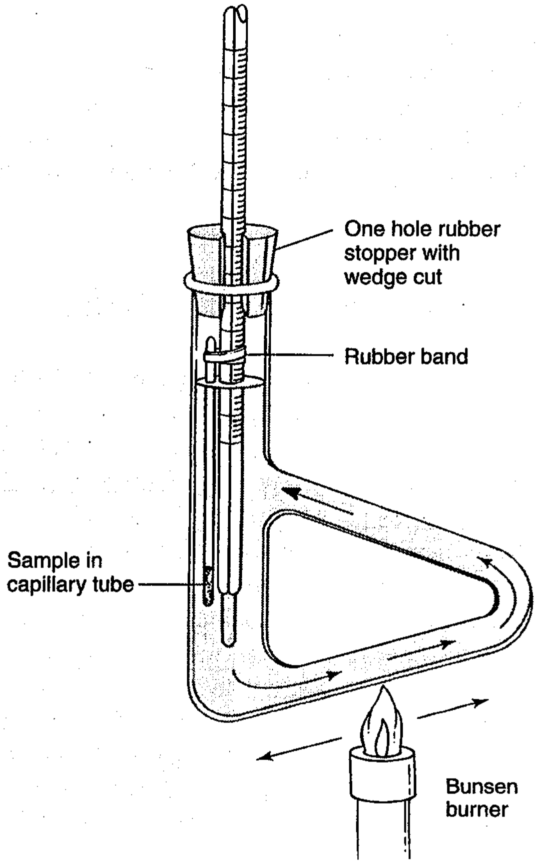

Figure 6.3 Thiele tube.

Thiele tube. If the heating is too fast, remove the burner for a few seconds, and then resume heating. The rate of heating should be *slow* near the melting point (about 1°C per min.) to ensure that the temperature increase is not faster than the rate at which heat can be transferred to the sample being observed. At the melting point, it is necessary that the mercury in the thermometer and the sample in the capillary tube be at temperature equilibrium.

6.7 DETERMINING THE MELTING POINT—ELECTRICAL INSTRUMENTS

Three types of electrically heated melting point instruments are illustrated in Figure 6.4. In each case, the melting-point tube is filled as described in Section 6.5 and placed in a holder located just behind the magnifying eyepiece. The apparatus is operated by moving the switch to the ON position, adjusting the potentiometric control dial for the

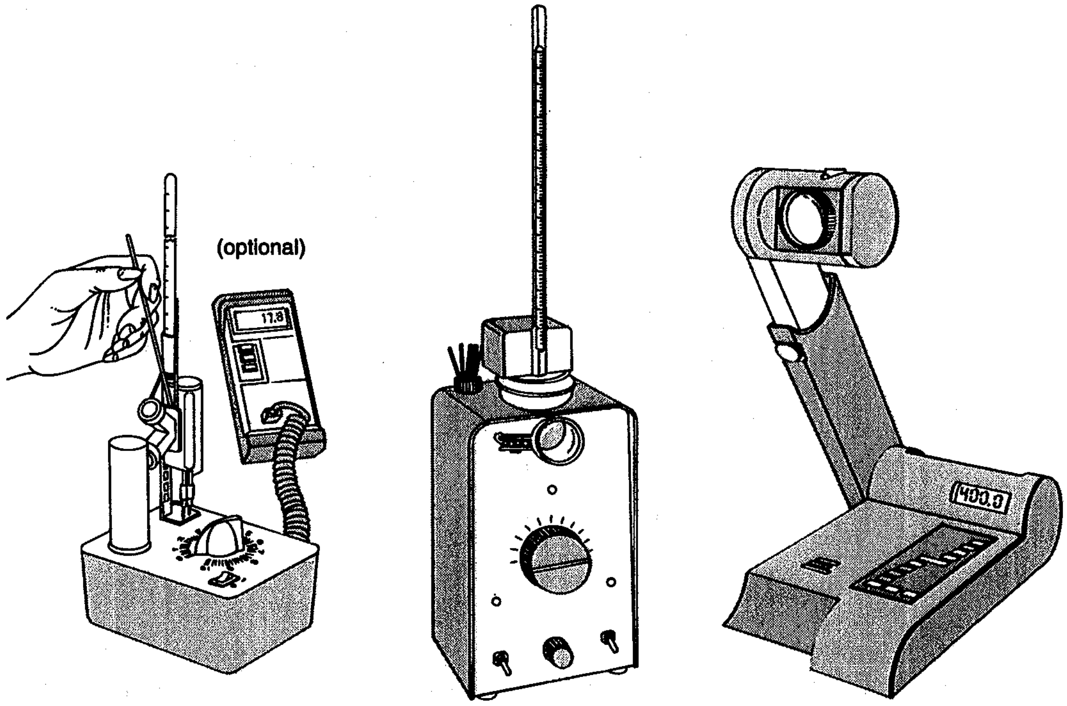

Figure 6.4 Melting-point apparatus.

desired rate of heating, and observing the sample through the magnifying eyepiece. The temperature is read from a thermometer, or in the most modern instruments, from a digital display attached to a thermocouple. Your instructor will demonstrate and explain the type used in your laboratory.

Most electrically heated instruments do not heat or increase the temperature of the sample linearly. Although the rate of increase may be linear in the early stages of heating, it usually decreases and leads to a constant temperature at some upper limit. The upper-limit temperature is determined by the setting of the heating control. Thus, a family of heating curves is usually obtained for various control settings, as shown in Figure 6.5. The four hypothetical curves shown (1–4) might correspond to different control settings. For a compound melting at temperature t_1, the setting corresponding to curve 3 would be ideal. In the beginning of the curve, the temperature is increasing too rapidly to allow determination of an accurate melting point, but after the change in slope, the temperature increase will have slowed to a more usable rate.

If the melting point of the sample is unknown, you can often save time by preparing two samples for melting-point determination. With one sample, you can rapidly determine a crude melting-point value. Then, repeat the experiment more carefully using the second sample. For the second determination, you already have an approximate idea of what the melting-point temperature should be and a proper rate of heating can be chosen.

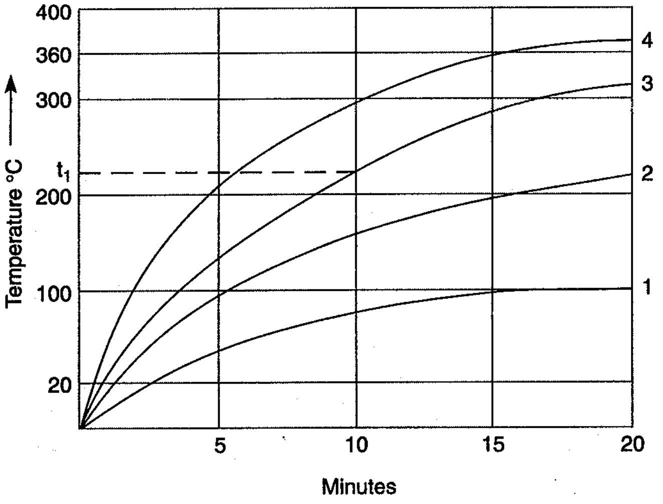

Figure 6.5 Heating-rate curves.

When measuring temperatures above 150°C, thermometer errors can become significant. For an accurate melting point with a high melting solid, you may wish to apply a **stem correction** to the thermometer as described in Section 6.13. An even better solution is to calibrate the thermometer as described in Section 6.12.

6.8 DECOMPOSITION, DISCOLORATION, SOFTENING, SHRINKAGE, AND SUBLIMATION

Many solid substances undergo some degree of unusual behavior before melting. At times it may be difficult to distinguish these types of behavior from actual melting. You should learn, through experience, how to recognize melting and how to distinguish it from decomposition, discoloration, and particularly softening and shrinkage.

Some compounds decompose on melting. This decomposition is usually evidenced by discoloration of the sample. Frequently, this decomposition point is a reliable physical property to be used in lieu of an actual melting point. Such decomposition points are indicated in tables of melting points by placing the symbol *d* immediately after the listed temperature. An example of a decomposition point is thiamine hydrochloride, whose melting point would be listed as 248°d, indicating that this substance melts with decomposition at 248°C. When decomposition is a result of reaction with the oxygen in air, it may be avoided by determining the melting point in a sealed, evacuated melting-point tube.

Figure 6.6 shows two simple methods of evacuating a packed tube. Method A uses an ordinary melting-point tube, while Method B constructs the melting-point tube from a disposable Pasteur pipet. Before using Method B, be sure to determine that the tip of the pipet will fit into the sample holder in your melting-point instrument.

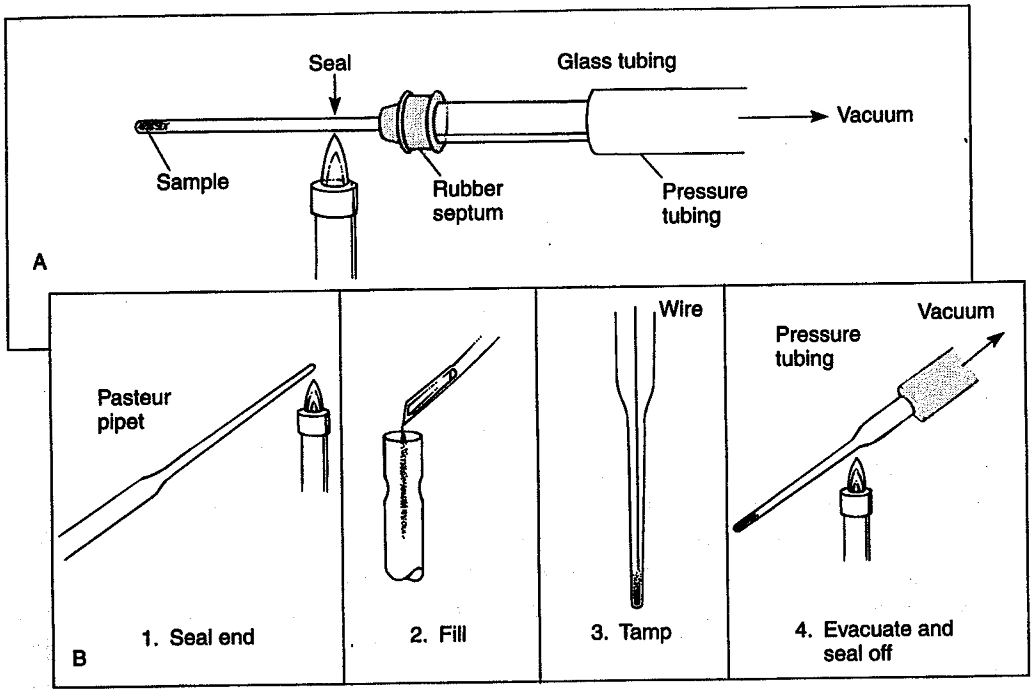

Figure 6.6 Evacuation and sealing of a melting-point capillary.

Method A. In Method A, a hole is punched through a rubber septum using a large pin or a small nail, and the capillary tube is inserted from the inside, sealed end first. The septum is placed over a piece of glass tubing connected to a vacuum line. After evacuating the tube, the upper end of the tube may be sealed by heating and pulling it closed.

Method B. In Method B, the thin section of a 9-inch Pasteur pipet is used to construct the melting-point tube. Carefully seal the tip of the pipet using a flame. Be sure to hold the tip *upward* as you seal it. This will prevent water vapor from condensing inside the pipet. When the sealed pipet has cooled, the sample may be added through the open end using a microspatula. A small wire may be used to compress the sample into the closed tip. (If your melting-point apparatus has a vibrator, it may be used in place of the wire to simplify the packing.) When the sample is in place, the pipet is connected to the vacuum line with tubing and evacuated. The evacuated sample tube is sealed by heating it with a flame and pulling it closed.

Some substances begin to decompose *below* their melting points. Thermally unstable substances may undergo elimination reactions or anhydride formation reactions during heating. The decomposition products formed represent impurities in the original sample, so the melting point of the substance may be lowered due to their presence.

It is normal for many compounds to soften or shrink immediately before melting. Such behavior represents not decomposition but a change in the crystal structure or a mixing with impurities. Some substances "sweat," or release solvent of crystallization, before melting. These changes do not indicate the beginning of melting. Actual melting begins

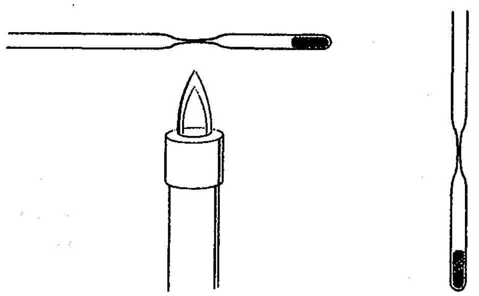

Figure 6.7 Sealing a tube for a substance that sublimes.

when the first drop of liquid becomes visible, and the melting range continues until the temperature at which all the solid has been converted to the liquid state. With experience, you soon learn to distinguish between softening, or "sweating," and actual melting. If you wish, the temperature of the onset of softening or sweating may be reported as a part of your melting point rage: 211°C (softens), 223–225°C (melts).

Some solid substances have such a high vapor pressure that they sublime at or below their melting points. In many handbooks, the sublimation temperature is listed along with the melting point. The symbols *sub, subl,* and sometimes *s* are used to designate a substance that sublimes. In such cases, the melting-point determination must be performed in a sealed capillary tube to avoid loss of the sample. The simplest way to accomplish sealing a packed tube is to heat the open end of the tube in a flame and pull it closed with a tweezers or forceps. A better way, although more difficult to master, is to heat the center of the tube in a small flame, rotating it about its axis, keeping the tube straight, until the center collapses. If this is not done quickly, the sample may melt or sublime while you are working. With the smaller chamber, the sample will not be able to migrate to the cool top of the tube that may be above the viewing area. Figure 6.7 illustrates the method.

PART B. BOILING POINTS

6.9 THE BOILING POINT

As a liquid is heated, the vapor pressure of the liquid increases to the point where it just equals the applied pressure (usually atmospheric pressure). At this point, the liquid is observed to boil. The normal boiling point is measured at 760 mmHg (760 torr) or 1 atm. At a lower applied pressure, the vapor pressure needed for boiling is also lowered, and the liquid boils at a lower temperature. The relation between applied pressure and temperature of boiling for a liquid is determined by its vapor pressure–temperature behavior. Figure 6.8 is an idealization of the typical vapor pressure–temperature behavior of a liquid.

Because the boiling point is sensitive to pressure, it is important to record the barometric pressure when determining a boiling point if the determination is being conducted at an elevation significantly above or below sea level. Normal atmospheric variations may

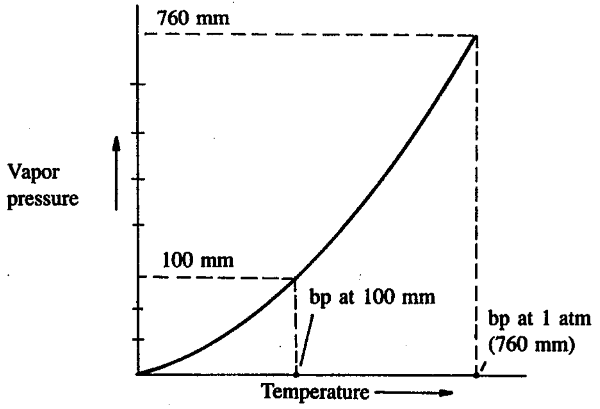

Figure 6.8 The vapor pressure–temperature curve for a typical liquid.

affect the boiling point, but they are usually of minor importance. However, if a boiling point is being monitored during the course of a vacuum distillation (Technique 9) that is being performed with an aspirator or a vacuum pump, the variation from the atmospheric value will be especially marked. In these cases, it is quite important to know the pressure as accurately as possible.

As a rule of thumb, the boiling point of many liquids drops about 0.5°C for a 10-mm decrease in pressure when in the vicinity of 760 mmHg. At lower pressures, a 10°C drop in boiling point is observed for each halving of the pressure. For example, if the observed boiling point of a liquid is 150°C at 10-mm pressure, then the boiling point would be about 140°C at 5 mmHg.

A more accurate estimate of the change in boiling point with a change of pressure can be made by using a **nomograph.** In Figure 6.9, a nomograph is given and a method is described for using it to obtain boiling points at various pressures when the boiling point is known at some other pressure.

6.10 DETERMINING BOILING POINTS—STANDARD-SCALE METHODS

Experimental methods of determining the boiling point of a liquid are easily available. When you have large quantities of material, you can simply record the boiling point (or boiling range) as indicated on a thermometer as the substance distills during a simple distillation (see Technique 8, Section 8.3, page 710). Alternatively, you may find it convenient to use a direct method, shown in Figure 6.10. With this method, the bulb of the thermometer can be immersed in vapor from the boiling liquid for a period long enough to allow it to equilibrate and give a good temperature reading.

The liquid can be brought to its boiling point very quickly when a test tube is used. Select a test tube that is long and narrow. You should avoid a large space between the sides of the test tube and the thermometer. Place the bulb of the thermometer as close as possible to the boiling liquid without actually touching it. You should use a small, inert

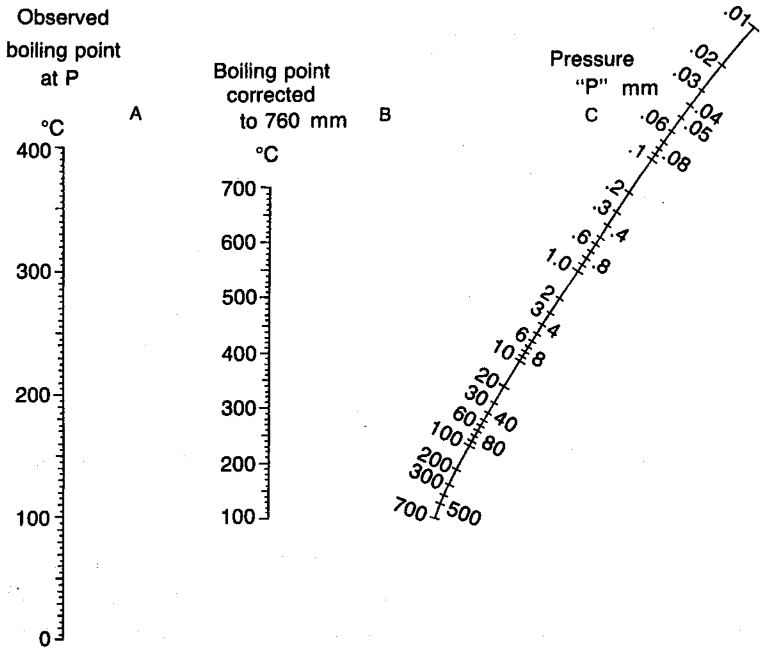

Figure 6.9 Pressure–temperature alignment nomograph. How to use the nomograph: Assume a reported boiling point of 100°C at 1 mm. To determine the boiling point at 18 mm, connect 100°C (column A) to 1 mm (column C) with a transparent plastic rule and observe where this line intersects column B (about 280°C). This value would correspond to the normal boiling point. Next, connect 280°C (column B) with 18 mm (column C) and observe where this intersects column A (151°C). The approximate boiling point will be 151°C at 18 mm. Reprinted by courtesy of MC/B Manufacturing Chemists, Inc.

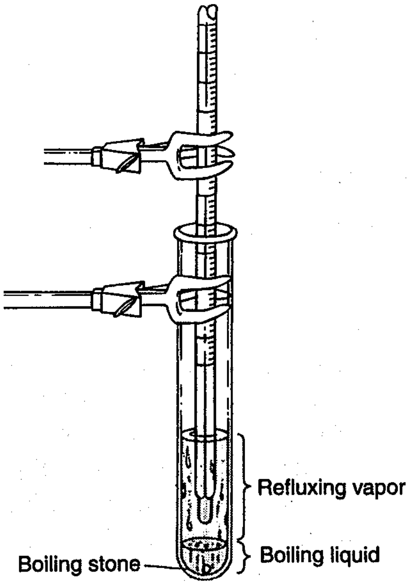

Figure 6.10 Standard-scale method of determining the boiling point.

carborundum (black) boiling stone. Do not use a marble or calcium carbonate (white) boiling chip. If safe operation permits, the best heating source is a small microburner. The flame will heat the liquid to boiling very quickly. The liquid must boil vigorously, such that you see a reflux ring and drops of liquid condensing on the sides of the test tube. You must, however, watch carefully so that you do not heat so much as to cause the liquid to boil out of the test tube, contact the flame, and cause a fire. The temperature reading on the thermometer must remain constant at its highest observed value. If the temperature continues to rise, the liquid has not yet reached its true boiling point.

6.11 DETERMINING BOILING POINTS—MICROSCALE METHODS

When you have smaller amounts of material, you can carry out a microscale or semi-microscale determination of the boiling point by using the apparatus shown in Figure 6.11.

Semi-Microscale Method. To carry out the semi-microscale determination, a piece of 5-mm glass tubing sealed at one end is attached to a thermometer with a rubber band

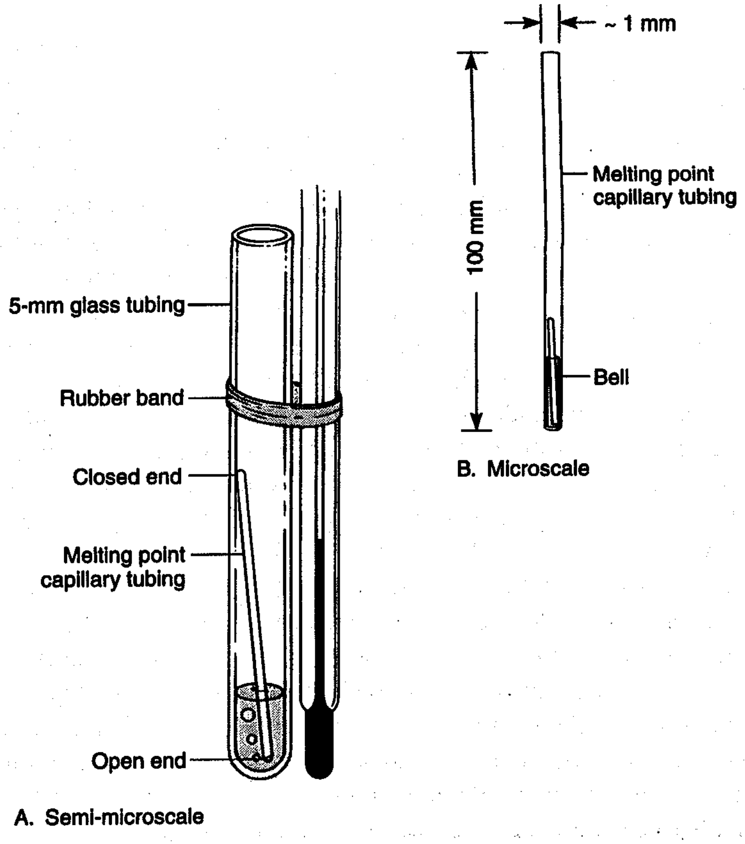

A. Semi-microscale

B. Microscale

Figure 6.11 Boiling-point determinations.

or a thin slice of rubber tubing. The liquid whose boiling point is being determined is introduced with a Pasteur pipet into this piece of tubing and a short piece of melting-point capillary (sealed at one end) is dropped in with the open end down. The whole unit is then placed in a Thiele tube. The rubber band should be placed above the level of the oil in the Thiele tube; otherwise, the band may soften in the hot oil. When positioning the band, keep in mind that the oil will expand when heated. Next, the Thiele tube is heated in the same fashion as described in Section 6.6 for determining a melting point. Heating is continued until a rapid and continuous stream of bubbles emerges from the inverted capillary. At this point, you should stop heating. Soon, the stream of bubbles slows down and stops. When the bubbles stop, the liquid enters the capillary tube. The moment at which the liquid enters the capillary tube corresponds to the boiling point of the liquid, and the temperature is recorded.

Microscale Method. In microscale experiments, there is often too little product available to use the semi-microscale method described above. However, the method can be scaled down in the following manner. The liquid is placed in a 1-mm melting-point capillary tube to a depth of about 4–6 mm. Use a syringe or a Pasteur pipet that has had its tip drawn thinner to transfer the liquid into the capillary tube. It may be necessary to use a centrifuge to transfer the liquid to the bottom of the tube. Next, prepare an appropriately sized inverted capillary, or **bell.**

The easiest way to prepare a bell is to use a commercial micropipet, such as a 10-μL Drummond "microcap." These are available in vials of 50 or 100 microcaps and are very inexpensive. To prepare the bell, cut the microcap in half with a file or scorer and then seal one end by inserting it a small distance into a flame, turning it on its axis until the opening closes.

If microcaps are not available, a piece of 1-mm open-end capillary tubing (same size as a melting-point capillary) can be rotated along its axis in a flame while being held horizontally. Use your index finger and thumbs to rotate the tube; do not change the distance between your two hands while rotating. When the tubing is soft, remove it from the flame and pull it to a thinner diameter. When pulling, keep the tube straight by *moving both your hands and your elbows outward* by about 4 inches. Hold the pulled tube in place a few moments until it cools. Using the edge of a file or your fingernail, break out the thin center section. Seal one end of the thin section in the flame; then, break it to a length that is about one and one-half times the height of your sample liquid (6–9 mm). Be sure the break is done squarely. Invert the bell (open end down) and place it in the capillary tube containing the sample liquid. Push the bell to the bottom with a fine copper wire if it adheres to the side of the capillary tube. A centrifuge may be used if you prefer. Figure 6.12 shows the construction method for the bell and the final assembly.

Place the microscale assembly in a standard melting-point apparatus (or a Thiele tube if an electrical apparatus is not available) to determine the boiling point. Heating is continued until a rapid and continuous stream of bubbles emerges from the inverted capillary. At this point, stop heating. Soon, the stream of bubbles slows down and stops. When the bubbles stop, the liquid enters the capillary tube. The moment at which the liquid enters the capillary tube corresponds to the boiling point of the liquid, and the temperature is recorded.

Explanation of the Method. During the initial heating, the air trapped in the inverted bell expands and leaves the tube, giving rise to a stream of bubbles. When the liquid begins boiling, most of the air has been expelled; the bubbles of gas are due to the boiling action of the liquid. Once the heating is stopped, most of the vapor pressure left in the bell comes from the vapor of the heated liquid that seals its open end. There is always vapor in equilibrium with a heated liquid. If the temperature of the liquid is above its boiling point, the pressure of the trapped vapor will either exceed or equal the atmospheric pressure. As the liquid cools, its vapor pressure decreases. When the vapor pressure drops just below atmospheric pressure (just below the boiling point), the liquid is forced into the capillary tube.

Difficulties. Three problems are common to this method. The first arises when the liquid is heated so strongly that it evaporates or boils away. The second arises when the liquid is not heated above its boiling point before heating is discontinued. If the heating is stopped at any point below the actual boiling point of the sample, the liquid enters the bell *immediately,* giving an apparent boiling point that is too low. Be sure that a continuous stream of bubbles, too fast for individual bubbles to be distinguished, is observed before lowering the temperature. Also, be sure that the bubbling action decreases slowly before the liquid enters the bell. If your melting-point apparatus has fine enough control and fast response, you can actually begin heating again and force the liquid out of the bell be-

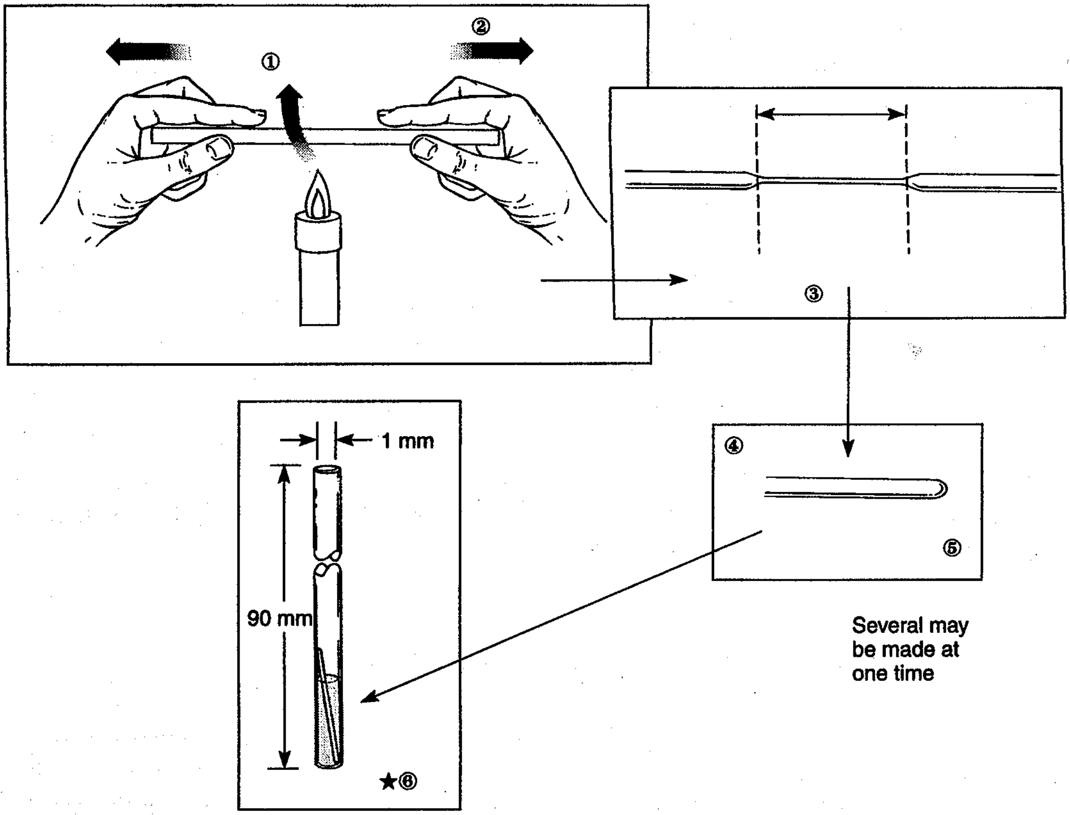

1. Rotate in flame until soft;
2. Remove from flame and pull;
3. Break pulled section out;
4. Seal one end;
5. Break to length;
6. Place bell in tube.

Figure 6.12 Construction of microcapillary bell for microscale boiling-point determination.

fore it becomes completely filled with the liquid. This allows a second determination to be performed on the same sample. The third problem is that the bell may be so light that the bubbling action of the liquid causes the bell to move up the capillary tube. This problem can sometimes be solved by using a longer (heavier) bell or by sealing the bell so that a larger section of solid glass is formed at the sealed end of the bell.

When measuring temperatures above 150°C, thermometer errors can become significant. For an accurate boiling point with a high-boiling liquid, you may wish to apply a **stem correction** to the thermometer as described in Section 6.13, or to calibrate the thermometer as described in Section 6.12.

PART C. THERMOMETER CALIBRATION AND CORRECTION

6.12 THERMOMETER CALIBRATION

When a melting-point or boiling-point determination has been completed, you expect to obtain a result that exactly duplicates the result recorded in a handbook or in the original literature. It is not unusual, however, to find a discrepancy of several degrees from the literature value. Such a discrepancy does not necessarily indicate that the experiment was incorrectly performed or that the material is impure; rather, it may indicate that the thermometer used for the determination was slightly in error. Most thermometers do not measure the temperature with perfect accuracy.

To determine accurate values, you must calibrate the thermometer that is used. This calibration is done by determining the melting points of a variety of standard substances with the thermometer. A plot is drawn of the observed temperature vs. the published value of each standard substance. A smooth line is drawn through the points to complete the chart. A correction chart prepared in this way is shown in Figure 6.13. This chart is used to correct any melting point determined with that particular thermometer. Each thermometer requires its own calibration curve. A list of suitable standard substances for calibrating thermometers is provided in Table 6.1. The standard substances, of course, must be pure in order for the corrections to be valid.

6.13 THERMOMETER STEM CORRECTIONS

Three types of thermometers are available: bulb immersion, stem immersion (partial immersion), and total immersion. **Bulb immersion** thermometers are calibrated by the manufacturer to give correct temperature readings when only the bulb (not the rest of the thermometer) is placed in the medium to be measured. **Stem immersion** thermometers are calibrated to give correct temperature readings when they are immersed to a specified depth in the medium to be measured. Stem immersion thermometers are easily rec-

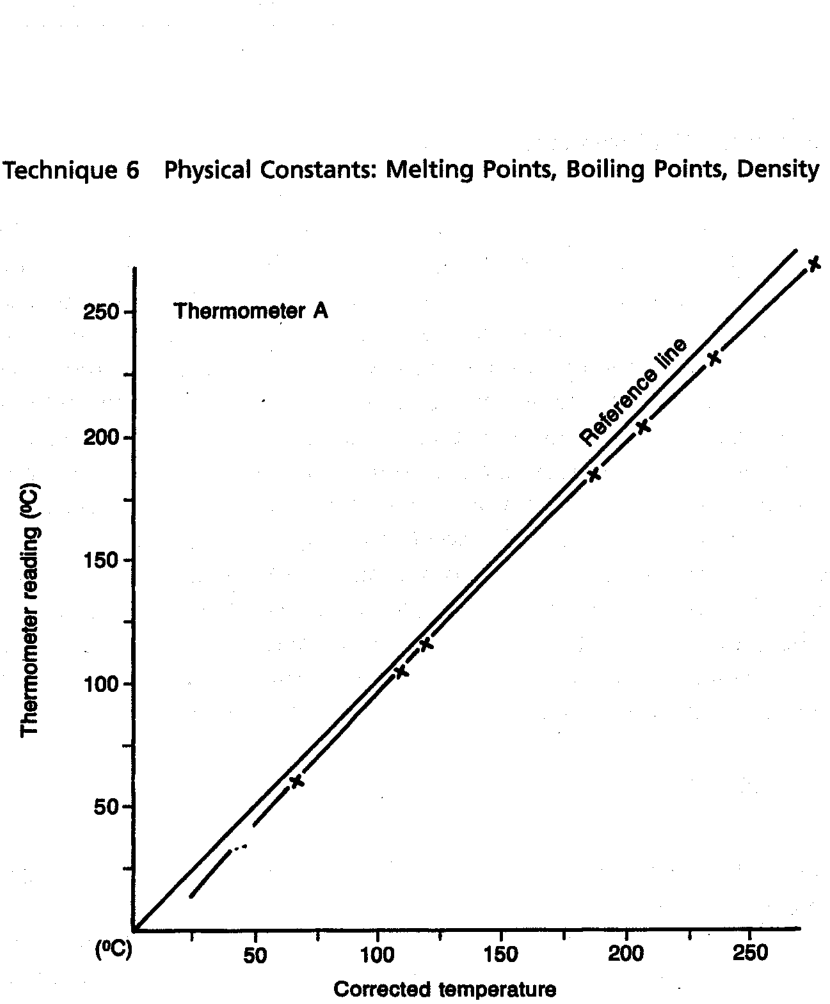

Figure 6.13 Thermometer calibration curve.

ognized because the manufacturer always scores a mark, or immersion ring, completely around the stem at the specified depth of immersion. The immersion ring is normally found below any of the temperature calibrations. **Total immersion** thermometers are calibrated when the entire thermometer is immersed in the medium to be measured. The three types of thermometer are often marked on the back (opposite side from the calibrations) by the words *bulb, immersion,* or *total,* but this may vary from one manufacturer to another. Because total immersion thermometers are less expensive, they are the type you are most likely to find in the laboratory.

TABLE 6.1 Melting-Point Standards

Compound	**Melting Point (°C)**
Ice (solid–liquid water)	0
Acetanilide	115
Benzamide	128
Urea	132
Succinic acid	189
3,5-Dinitrobenzoic acid	205

Manufacturers design total immersion thermometers to read correctly only when they are immersed totally in the medium to be measured. The entire mercury thread must be covered. Because this situation is rare, a **stem correction** should be added to the observed temperature. This correction, which is positive, can be fairly large when high temperatures are being measured. Keep in mind, however, that if your thermometer has been calibrated for its desired use (such as described in Section 6.12 for a melting-point apparatus), a stem correction should not be necessary for any temperature within the calibration limits. You are most likely to want a stem correction when you are performing a distillation. If you determine a melting point or boiling point using an uncalibrated, total-immersion thermometer, you will also want to use a stem correction.

When you wish to make a stem correction for a total immersion thermometer, the formula given below may be used. It is based on the fact that the portion of the mercury thread in the stem is cooler than the portion immersed in the vapor or the heated area around the thermometer. The mercury will not have expanded in the cool stem to the same extent as in the warmed section of the thermometer. The equation used is

$$(0.000154)(T - t_1)(T - t_2) = \text{correction to be added to } T \text{ observed.}$$

1. The factor 0.00154 is a constant, the coefficient of expansion for the mercury in the thermometer.
2. The term $T - t_1$ corresponds to the length of the mercury thread not immersed in the heated area. It is convenient to use the temperature scale on the thermometer itself for this measurement rather than an actual length unit. T is the observed temperature, and t_1 is the *approximate* place where the heated part of the stem ends and the cooler part begins.
3. The term $T - t_2$ corresponds to the difference between the temperature of the mercury in the vapor T and the temperature of the mercury in the air outside the heated area (room temperature). The term T is the observed temperature, and t_2 is measured by hanging another thermometer so that the bulb is close to the stem of the main thermometer.

Figure 6.14 shows how to apply this method for a distillation. By the formula given above, it can be shown that high temperatures are more likely to require a stem correction and that low temperatures need not be corrected. The calculations given below illustrate this point.

Example 1		**Example 2**	
$T = 200°\text{C}$		$T = 100°\text{C}$	
$t_1 = 0°\text{C}$		$t_1 = 0°\text{C}$	
$t_2 = 35°\text{C}$		$t_2 = 35°\text{C}$	
$(0.000154)(200)(165) = 5.1°$	stem correction	$(0.000154)(100)(65) = 1.0°\text{C}$	stem correction
$200°\text{C} + 5°\text{C} = 205°\text{C}$	corrected temp	$100°\text{C} + 1°\text{C} = 101°\text{C}$	corrected temp

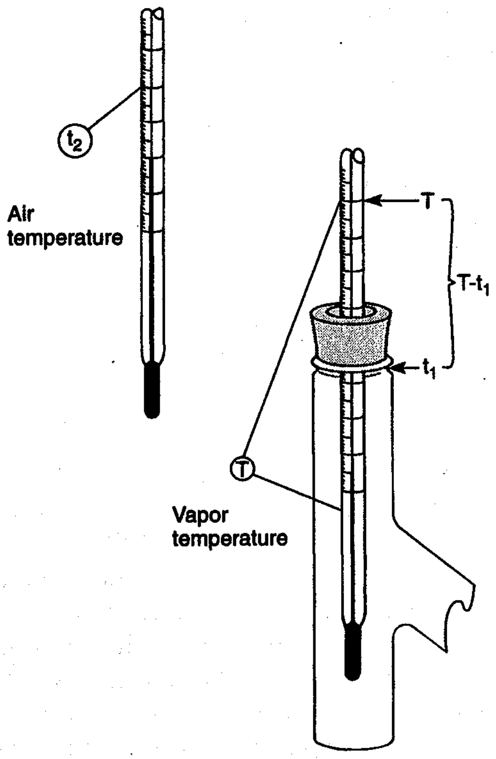

Figure 6.14 Measurement of a thermometer stem correction during distillation.

PART D. DENSITY

6.14 DENSITY

Density is defined as mass per unit volume and is generally expressed in units of grams per milliliter (g/mL) for a liquid and grams per cubic centimeter (g/cm^3) for a solid.

$$\text{Density} = \frac{\text{Mass}}{\text{Volume}} \quad \text{or} \quad D = \frac{M}{V}$$

In organic chemistry, density is most commonly used in converting the weight of liquid to a corresponding volume, or vice versa. It is often easier to measure a volume of a liquid than to weigh it. As a physical property, density is also useful for identifying liquids in much the same way that boiling points are used.

Although precise methods that allow the measurement of the densities of liquids at the microscale level have been developed, they are often difficult to perform. An approx-

TABLE 6.2 Densities Determined by the Automatic Pipet Method (g/mL)

Substance	bp	lit	100 μL
Water	100	1.000	1.01
Hexane	69	0.660	0.66
Acetone	56	0.788	0.77
Dichloromethane	40	1.330	1.27
Diethyl ether	35	0.713	0.67

imate method for measuring densities can be found in using a 100-μL (0.100-mL) automatic pipet (Technique 1, Section 1.6, p. 597). Clean, dry, and pre-weigh one or more small vials (including their caps and liners) and record their weights. You should handle these vials with a tissue in order to avoid getting your fingerprints on them. Adjust the automatic pipet to deliver 100 μL and fit it with a clean, new tip. Use the pipet to deliver 100 μL of the unknown liquid to each of your tared vials. Cap them so that the liquid does not evaporate. Reweigh the vials and use the weight of the 100 μL of liquid delivered to calculate a density for each case. It is recommended that from three to five determinations be performed, that the calculations be performed to three significant figures, and that all the calculations be averaged to obtain the final result. This determination of the density will be accurate to within two significant figures. Table 6.2 compares some literature values with those that could be obtained by this method.

PROBLEMS

1. Two substances, A and B, have the same melting point. How can you determine if they are the same without using any form of spectroscopy? Explain in detail.

2. Using Figure 6.5, determine which heating curve would be most appropriate for a substance with a melting point of about 150°C.

3. What steps can you take to determine the melting point of a substance that sublimes before it melts?

4. Using the pressure–temperature alignment chart in Figure 6.9, answer the following questions.

(a) What is the normal boiling point (at 760 mmHg) for a compound that boils at 150°C at 10 mmHg pressure?

(b) At what temperature would the compound in (a) boil if the pressure were 40 mmHg?

(c) A compound was distilled at atmospheric pressure and had a boiling point of 285°C. What would be the approximate boiling range for this compound at 15 mmHg?

5. Calculate the corrected boiling point for nitrobenzene by using the method given in Section 6.13. The boiling point was determined using an apparatus similar to that shown in Figure 6.10. The observed boiling point was 205°C. The reflux ring in the test tube just reached up to the 0°C mark on the thermometer. A second thermometer suspended alongside the test tube, at a slightly higher level than the one inside, gave a reading of 35°C.

6. Suppose that you had calibrated the thermometer in your melting-point apparatus against a series of melting-point standards. After reading the temperature and converting it using the calibration chart, should you also apply a stem correction? Explain.

7. The density of a liquid was determined by the automatic pipet method. A 100-μL automatic pipet was used. The liquid had a mass of 0.082 g. What was the density in grams per milliliter of the liquid?

8. A compound melting at 134°C was suspected to be either aspirin (mp 135°C) or urea (mp 133°C). Explain how you could determine whether one of these two suspected compounds was identical to the unknown compound without using any form of spectroscopy.

9. An unknown compound gave a melting point of 230°C. When the molten liquid solidified, the melting point was redetermined and found to be 131°C. Give a possible explanation for this discrepancy.

10. During the micro boiling-point determination of an unknown liquid, heating was discontinued at 154°C and the liquid immediately began to enter the inverted bell. Heating was begun again at once, and the liquid was forced out of the bell. Heating was again discontinued at 165°C, at which time a very rapid stream of bubbles emerged from the bell. On cooling, the rate of bubbling gradually diminished until the liquid reached a temperature of 161°C, and entered and filled the bell. Explain this sequence of events. What was the boiling point of the liquid?

TECHNIQUE 7

Extractions, Separations, and Drying Agents

7.1 EXTRACTION

Transferring a solute from one solvent into another is called **extraction,** or more precisely, liquid–liquid extraction. The solute is extracted from one solvent into the other because the solute is more soluble in the second solvent than in the first. The two solvents must not be **miscible** (mix freely), and they must form two separate **phases** or layers, in order for this procedure to work. Extraction is used in many ways in organic chemistry. Many **natural products** (organic chemicals that exist in nature) are present in animal and plant tissues having high water content. Extracting these tissues with a water-immiscible solvent is useful for isolating natural products. Often, diethyl ether (commonly referred to as "ether") is used for this purpose. Sometimes, alternative water-immiscible solvents such as hexane, petroleum ether, ligroin, and methylene chloride are used. For instance, caffeine, a natural product, can be extracted from an aqueous tea solution by shaking it successively with several portions of methylene chloride.

A generalized extraction process, using a specialized piece of glassware called a separatory funnel, is illustrated in Figure 7.1. The first solvent contains a mixture of black and white molecules (Fig. 7.1A). A second solvent that is not miscible with the first is added. After the separatory funnel is capped and shaken, the layers separate. In this ex-

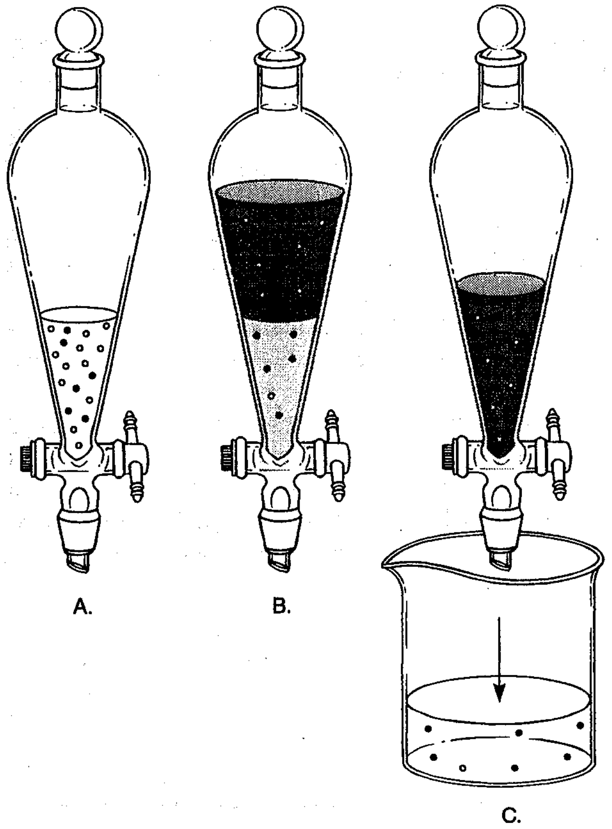

A. Solvent 1 contains a mixture of molecules (black and white).

B. After shaking with solvent 2 (shaded), most of the white molecules have been extracted into the new solvent. The white molecules are more soluble in the second solvent, while the black molecules are more soluble in the original solvent.

C. With removal of the lower phase, the black and white molecules have been partially separated.

Figure 7.1 The extraction process.

ample the second solvent (shaded) is less dense than the first, so it becomes the top layer (Fig. 7.1B). Because of differences in physical properties, the white molecules are more soluble in the second solvent, while the black molecules are more soluble in the first solvent. Most of the white molecules are in the upper layer, but there are some black molecules there, too. Likewise, most of the black molecules are in the lower layer. However, there are still a few white molecules in this lower phase. The lower phase may be separated from the upper phase by opening the stopcock at the bottom of the separatory funnel and allowing the lower layer to drain into a beaker (Fig. 7.1C). In this example, notice that it was not possible to effect a complete separation of the two types of molecules with a single extraction. This is a common occurrence in organic chemistry.

Many substances are soluble in both water and organic solvents. Water can be used to extract or "wash" water-soluble impurities from an organic reaction mixture. To carry out a "washing" operation, you add water and an immiscible organic solvent to the reaction mixture contained in a separatory funnel. After stoppering the funnel and shaking it, you allow the organic layer and the aqueous (water) layer to separate from each other. A water wash removes highly polar and water-soluble materials, such as sulfuric acid, hydrochloric acid, or sodium hydroxide from the organic layer. The washing operation helps to purify the desired organic compound present in the original reaction mixture.

7.2 DISTRIBUTION COEFFICIENT

When a solution (solute A in solvent 1) is shaken with a second solvent (solvent 2) with which it is not miscible, the solute distributes itself between the two liquid phases. When the two phases have separated again into two distinct solvent layers, an equilibrium will have been achieved such that the ratio of the concentrations of the solute in each layer defines a constant. The constant, called the **distribution coefficient** (or partition coefficient) K, is defined by

$$K = \frac{C_2}{C_1}$$

where C_1 and C_2 are the concentrations at equilibrium, in grams per liter or milligrams per milliliter of solute A in solvent 1 and in solvent 2, respectively. This relationship is a ratio of two concentrations and is independent of the actual amounts of the two solvents mixed. The distribution coefficient has a constant value for each solute considered and depends on the nature of the solvents used in each case.

Not all the solute will be transferred to solvent 2 in a single extraction unless K is very large. Usually, it takes several extractions to remove all the solute from solvent 1. In extracting a solute from a solution, it is always better to use several small portions of the second solvent than to make a single extraction with a large portion. Suppose, as an illustration, a particular extraction proceeds with a distribution coefficient of 10. The system consists of 5.0 g of organic compound dissolved in 100 mL of water (solvent 1). In this illustration, the effectiveness of three 50-mL extractions with ether (solvent 2) is compared with one 150-mL extraction with ether. In the first 50-mL extraction, the amount extracted into the ether layer is given by the following calculation. The amount of compound remaining in the aqueous phase is given by x.

$$K = 10 = \frac{C_2}{C_1} = \frac{\left(\dfrac{5.0 - x}{50}\ \dfrac{\text{g}}{\text{mL ether}}\right)}{\left(\dfrac{x}{100}\ \dfrac{\text{g}}{\text{mL H}_2\text{O}}\right)}\ ;\ 10 = \frac{(5.0 - x)(100)}{50x}$$

$$\begin{aligned} 500x &= 500 - 100x \\ 600x &= 500 \\ x &= 0.83 \text{ g remaining in the aqueous phase} \\ 5.0 - x &= 4.17 \text{ g in the ether layer} \end{aligned}$$

As a check on the calculation, it is possible to substitute the value 0.83 g for x in the original equation and demonstrate that the concentration in the ether layer divided by the concentration in the water layer equals the distribution coefficient.

$$\frac{\left(\dfrac{5.0 - x}{50}\ \dfrac{\text{g}}{\text{mL ether}}\right)}{\left(\dfrac{x}{100}\ \dfrac{\text{g}}{\text{mL H}_2\text{O}}\right)} = \frac{\dfrac{4.17}{50}}{\dfrac{0.83}{100}} = \frac{0.083 \text{ g/mL}}{0.0083 \text{ g/mL}} = 10 = K$$

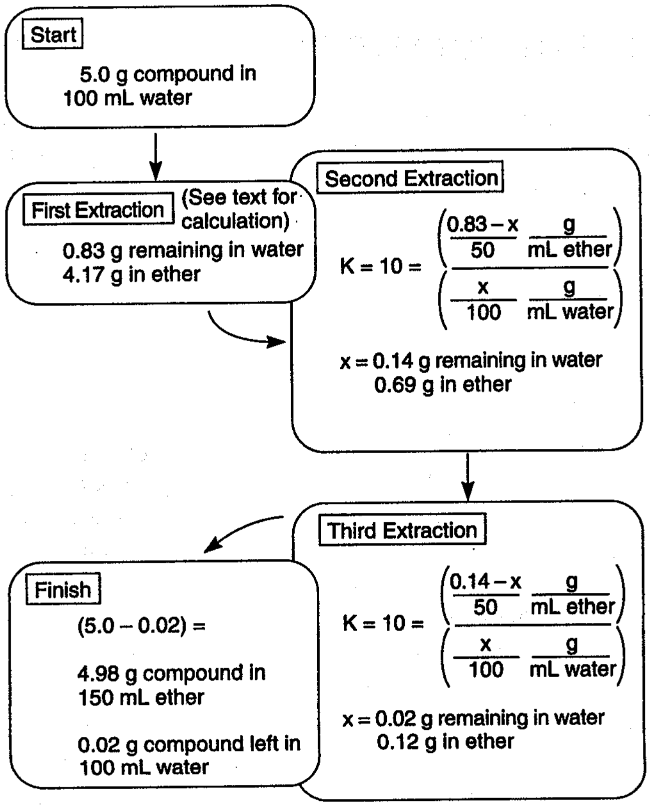
Start
5.0 g compound in
100 mL water
First Extraction
(See text for calculation)
0.83 g remaining in water
4.17 g in ether
Second Extraction
K = 10 = (0.83 − x / 50 g / mL ether) / (x / 100 g / mL water)
x = 0.14 g remaining in water
0.69 g in ether
Third Extraction
K = 10 = (0.14 − x / 50 g / mL ether) / (x / 100 g / mL water)
x = 0.02 g remaining in water
0.12 g in ether
Finish
(5.0 − 0.02) =
4.98 g compound in
150 mL ether
0.02 g compound left in
100 mL water

Figure 7.2 The result of extraction of 5.0 g of compound in 100 mL of water by three successive 50-mL portions of ether. Compare this result with that of Figure 7.3.

The second extraction with another 50-mL portion of fresh ether is performed on the aqueous phase which now contains 0.83 g of the solute. The amount of solute extracted is given by the calculation shown in Figure 7.2. Also shown in the figure is a calculation for a third extraction with another 50-mL portion of ether. This third extraction will transfer 0.12 g of solute into the ether layer, leaving 0.02 g of solute remaining in the water layer. A total of 4.98 g of solute will be extracted into the combined ether layers, and 0.02 g will remain in the aqueous phase.

Figure 7.3 shows the result of a *single* extraction with 150 mL of ether. As shown there, 4.69 g of solute was extracted into the ether layer, leaving 0.31 g of compound in the aqueous phase. One can see that three successive 50-mL ether extractions (Fig. 7.2) succeeded in removing 0.29 g more solute from the aqueous phase than using one 150-mL portion of either (Fig. 7.3). This differential represents 5.8% of the total material.

7.3 CHOOSING AN EXTRACTION METHOD AND A SOLVENT

Three types of apparatus are used for extractions: conical vials, centrifuge tubes, and separatory funnels. These are shown in Figure 7.4. Conical vials may be used with volumes

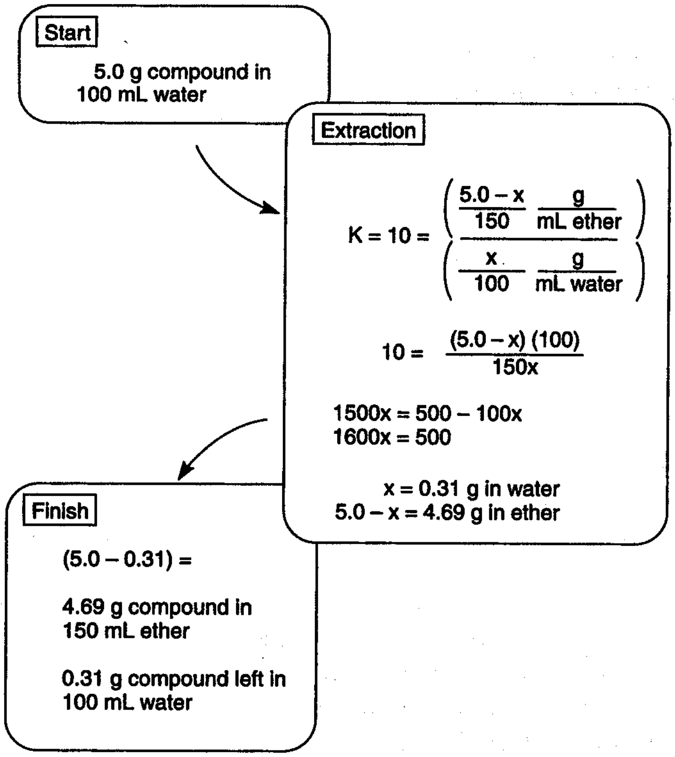
Start
5.0 g compound in 100 mL water
Extraction
$K = 10 = \dfrac{\left(\dfrac{5.0 - x}{150}\ \dfrac{g}{\text{mL ether}}\right)}{\left(\dfrac{x}{100}\ \dfrac{g}{\text{mL water}}\right)}$
$10 = \dfrac{(5.0 - x)(100)}{150x}$
$1500x = 500 - 100x$
$1600x = 500$
$x = 0.31$ g in water
$5.0 - x = 4.69$ g in ether
Finish
$(5.0 - 0.31) =$
4.69 g compound in 150 mL ether
0.31 g compound left in 100 mL water

Figure 7.3 The result of extraction of 5.0 g of compound in 100 mL of water with one 150-mL portion of ether. Compare this result with that of Figure 7.2

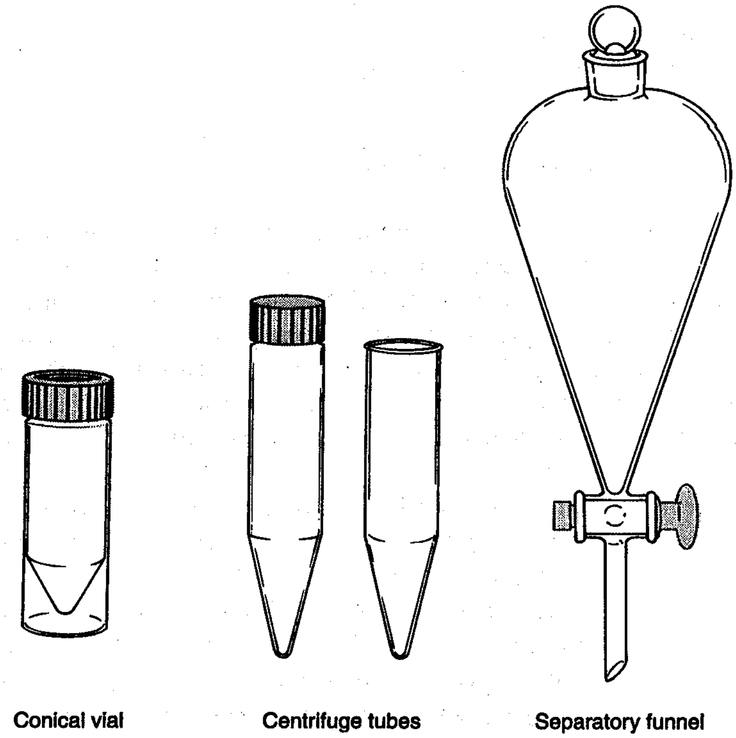
Conical vial
Centrifuge tubes
Separatory funnel

Figure 7.4 Apparatus used for extraction.

TABLE 7.1 Densities of Common Extraction Solvents

Solvent	Density (g/mL)
Ligroin	0.67–0.69
Diethyl ether	0.71
Toluene	0.87
Water	1.00
Methylene chloride	1.330

of less than 4 mL; volumes of up to 10 mL may be handled in centrifuge tubes. A centrifuge tube equipped with a screwcap is particularly useful for extractions. The separatory funnel is used in large-scale reactions. Each type of equipment is discussed in a separate section.

Most extractions consist of an aqueous phase and an organic phase. In order to extract a substance from an aqueous phase, an organic solvent that is not miscible with water must be used. Table 7.1 lists a number of the common organic solvents that are not miscible with water and are used for extractions.

Those solvents that have a density less than that of water (1.00 g/mL) will separate as the top layer when shaken with water. Those solvents that have a greater density than water will separate into the lower layer. For instance, diethyl ether ($d = 0.71$ g/mL) when shaken with water will form the upper layer, whereas methylene chloride ($d = 1.33$ g/mL) will form the lower layer. When performing an extraction, slightly different methods are used when you wish to separate the lower layer (whether or not it is the aqueous layer or the organic layer) than when you wish to separate the upper layer.

7.4 THE SEPARATORY FUNNEL

A separatory funnel is illustrated in Figure 7.5. It is the piece of equipment used for carrying out extractions with medium to large quantities of material, and it will be used frequently in the procedures in this text. To fill the separatory funnel, support it in an iron ring attached to a ring stand. Since it is easy to break a separatory funnel by "clanking" it against the metal ring, pieces of rubber tubing are often attached to the ring to cushion the funnel as shown in Figure 7.5. These are short pieces of tubing, cut to a length of about 3 cm and slit open along their length. When slipped over the inside of the ring, they cushion the funnel in its resting place.

When beginning an extraction, the first step is to close the stopcock. (Don't forget!) Using a powder funnel (wide bore) placed in the top of the separatory funnel, fill it with both the solution to be extracted and the extraction solvent. Swirl the funnel gently by holding it by its upper neck, and then stopper it. Pick up the separatory funnel with two hands and hold it as shown in Figure 7.6. Hold then stopper in place firmly because the two immiscible liquids will build pressure when they mix, and this pressure may force the stopper out of the separatory funnel. To release this pressure, vent the funnel by hold-

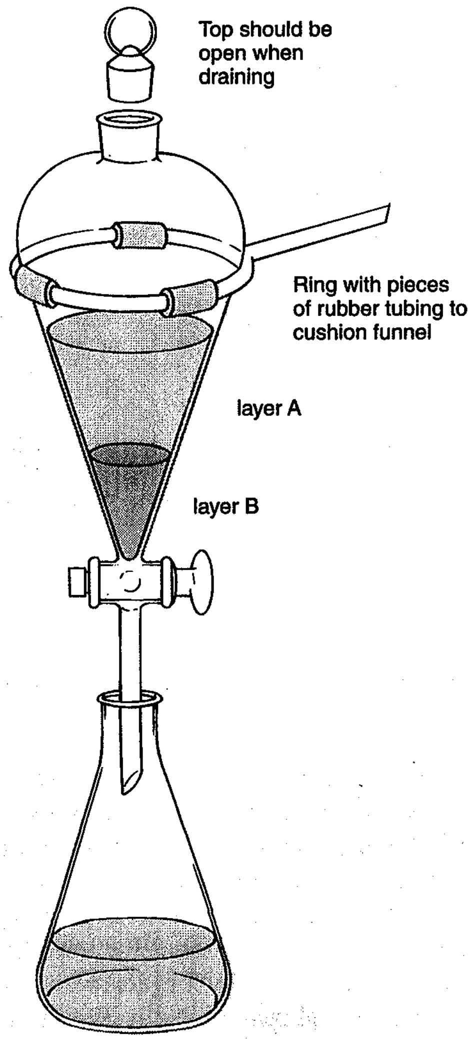
Top should be open when draining
Ring with pieces of rubber tubing to cushion funnel
layer A
layer B

Figure 7.5 The separatory funnel.

ing it upside down (hold the stopper securely) and slowly open the stopcock. Usually the rush of vapors out of the opening can be heard. Continue shaking and venting until the "whoosh" is no longer audible. Now continue shaking the mixture gently for about one minute. This can be done by inverting the funnel in a rocking motion repeatedly or, if the formation of an emulsion is not a problem (see Section 7.9, p. 699), by shaking the funnel more vigorously for less time.

> There is an art to shaking and venting a separatory funnel correctly, and it usually seems awkward to the beginner. The technique is best learned by observing a person, such as your instructor, who is thoroughly familiar with the separatory funnel's use.

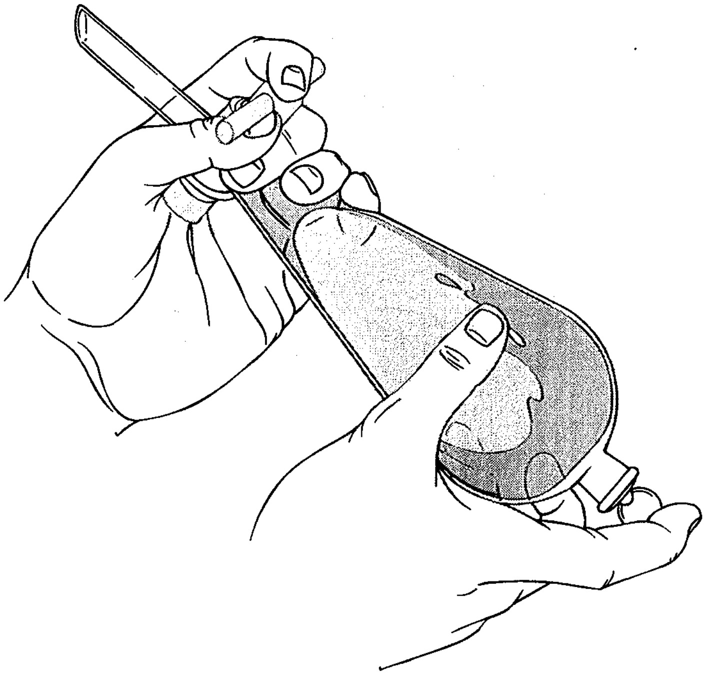

Figure 7.6 Correct way of shaking and venting the separatory funnel.

When you have finished mixing the liquids, place the separatory funnel in the iron ring and remove the top stopper immediately. The two immiscible solvents separate into two layers after a short time, and they can be separated from one another by draining most of the lower layer through the stopcock.[1] Allow a few minutes to pass so that any of the lower phase adhering to the inner glass surfaces of the separatory funnel can drain down. Open the stopcock again and allow the remainder of the lower layer to drain until the interface between the upper and lower phases just begins to enter the bore of the stopcock. At this moment, close the stopcock and remove the remaining upper layer by pouring it from the top opening of the separatory funnel.

To minimize contamination of the two layers, the lower layer should always be drained from the bottom of the separatory funnel and the upper layer poured out from the top of the funnel.

When methylene chloride is used as the extracting solvent with an aqueous phase, it will settle to the bottom and be removed through the stopcock. The aqueous layer remains in the funnel. A second extraction of the remaining aqueous layer with fresh methylene chloride may be needed.

With a diethyl ether (ether) extraction of an aqueous phase, the organic layer will form on top. Remove the lower aqueous layer through the stopcock and pour the upper ether layer from the top of the separatory funnel. Pour the aqueous phase back into the separatory funnel and extract it a second time with fresh ether. The combined organic phases must be dried using a suitable drying agent (Section 7.8) before the solvent is removed.

---

[1] A common error is to try to drain the separatory funnel without removing the top stopper. Under this circumstance, the funnel will not drain since a partial vacuum is created in the space above the liquid.

7.5 HOW DO YOU DETERMINE WHICH ONE IS THE ORGANIC LAYER?

A common problem that you might encounter during an extraction is trying to determine which of the two layers is the organic layer and which is the aqueous (water) layer. The most common situation is when the aqueous layer is on the bottom in the presence of an upper organic layer consisting of ether, ligroin, petroleum ether, or hexane (see densities in Table 7.1). However, the aqueous layer will be on the top when you use methylene chloride as a solvent (again, see Table 7.1). Although the textbook frequently identifies the expected relative positions of the organic and aqueous layers, sometimes their actual positions are reversed. Surprises usually occur in situations where the aqueous layer contains a high concentration of sulfuric acid or a dissolved ionic compound, such as sodium chloride. Dissolved substances greatly increase the density of the aqueous layer, which may lead to the aqueous layer being found on the bottom even when coexisting with a relatively dense organic layer such as methylene chloride.

Always keep both layers until you have actually isolated the desired compound or until you are certain where your desired substance is located.

To determine if a particular layer is an aqueous one, add a few drops of water to the layer. Observe closely as you add the water to see where it goes. If the layer is water, then the drops of added water will dissolve in the aqueous layer and increase its volume. If the added water forms droplets or a new layer, however, you can assume that the suspected aqueous layer is actually organic. You can use a similar procedure to identify a suspected organic layer. This time, try adding more of the solvent, such as methylene chloride. The organic layer should increase in size, without separation of a new layer, if the tested layer is actually organic.

7.6 THE CONICAL VIAL

The 5-mL conical vial is the most useful piece of equipment for carrying out extractions with small amounts of material. Before you begin you should check the capped conical vial for leaks. Leaks are especially serious with a conical vial since you are dealing with a small volume of liquid from the very beginning. To check for leaks, place some water in the conical vial, place the Teflon liner in the cap, and screw the cap securely onto the conical vial. Shake the vial vigorously and check for leaks. Conical vials used for extractions must not be chipped on the top edge of the vial. The cap and liner seal against this top edge and, if there are chips, the seal may not be adequate. If there is a leak, try tightening the cap or replacing the liner with a new one. The liners are a sandwich with two sides, a hard Teflon surface (thin) on one side and a softer rubber surface (thick) on the other. Sometimes the rubber surface will provide a better seal to the top edge than the Teflon surface. If all else fails, replace the glass vial.

As an alternative to shaking, the contents of a conical vial can be mixed by spinning your microspatula in the vial for several minutes (twirl it between your thumb and your index finger). Another mixing technique involves repeatedly drawing the mixture into a disposable Pasteur pipet and squirting it rapidly back into the vial for a period of several minutes.

The conical vial, with a capacity of 5 mL of liquid, is commonly used in microscale procedures. In this section, we consider explicitly the method whereby the lower layer is removed by a disposable Pasteur pipet. A concrete example would be the extraction of the desired product from an aqueous layer using methylene chloride (*d* = 1.33 g/mL) as the extraction solvent.

Always remember to place a conical vial in a small beaker when it is not being held. Although the vials have a flat bottom, they can tip easily and roll off the bench.

Figure 7.7 shows the steps in using a conical vial for a separation. Figure 7.7A shows the solution to be extracted. The extraction solvent is added, and the vial is capped and

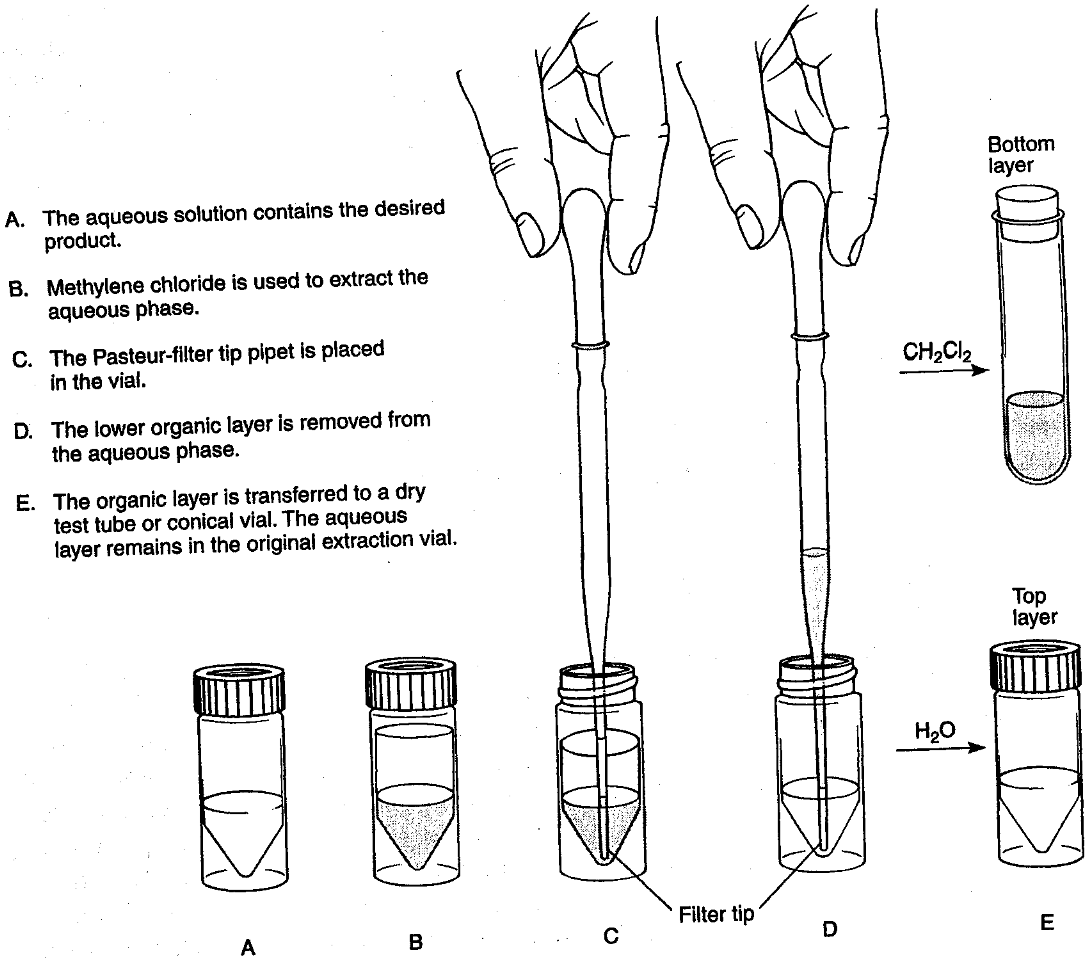
A. The aqueous solution contains the desired product.
B. Methylene chloride is used to extract the aqueous phase.
C. The Pasteur-filter tip pipet is placed in the vial.
D. The lower organic layer is removed from the aqueous phase.
E. The organic layer is transferred to a dry test tube or conical vial. The aqueous layer remains in the original extraction vial.

Figure 7.7 Extraction of an aqueous solution using a solvent more dense than water: Methylene chloride.

shaken (with occasional venting to relieve pressure). Figure 7.7B shows the mixture after the immiscible layers have reseparated. The lower layer is removed using a disposable Pasteur pipet (Fig. 7.7C) and transferred to a *dry* storage container such as a test tube. This operation requires a little practice. You must learn to depress the rubber bulb just the right amount to remove only the lower layer, and to leave the upper layer behind. To avoid spillage, it is best to have the storage test tube or another vial held in the same hand as the source vial as shown in Figure 7.8.

Many workers prefer to use a **filter tip pipet** when separating the layers. A filter tip pipet is a Pasteur pipet that has been modified to have a loose plug of cotton in the end. The filter tip pipet both fills and empties more slowly than an unmodified Pasteur pipet. This allows for better control when trying to separate the two layers in the conical vial. A *very small amount* of cotton or glass wool is used to make the filter plug, and it is pushed in place with a short length of copper wire. Care must be taken to avoid compressing the cotton or glass wool so tightly that liquid will not pass through the tip of the pipet.

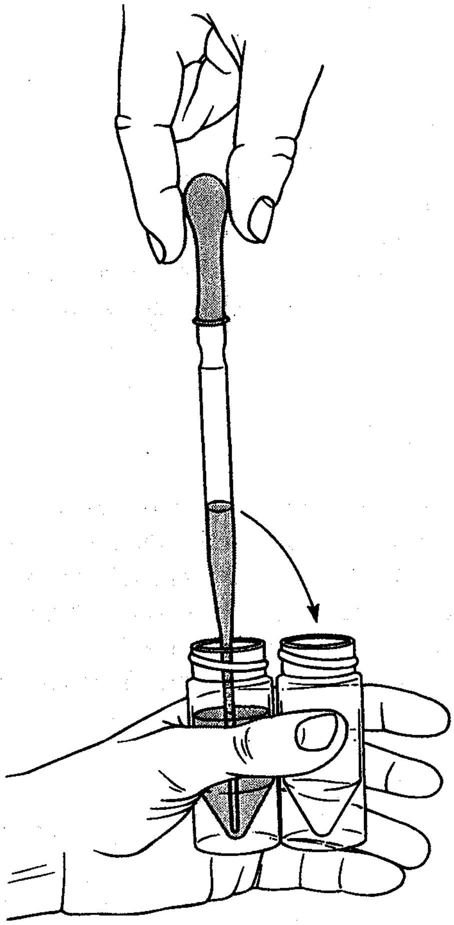

Figure 7.8 Method of holding vials while transferring liquids.

In Figure 7.7, methylene chloride, a solvent more dense than water, was illustrated. If you are extracting with a solvent less dense than water (for example, diethyl ether), you will want to save the upper layer. In that case, the extraction sequence shown in Figure 7.7 will have the positions of the water and the solvent reversed at stage E of the separation. The water layer will be in the test tube, and the ether in the conical vial.

7.7 THE SCREW-CAP CENTRIFUGE TUBE

If you require an extraction that uses a larger volume than a conical vial can accommodate (5 mL), a centrifuge tube can often be used. A commonly available size of centrifuge tube has a volume of about 15 mL, and is supplied with a screw cap. In performing an extraction with a screwcapped centrifuge tube, the same procedures outlined for the conical vial (Section 7.6) are used. As is the case for a conical vial, the tapered bottom of the centrifuge tube makes it easy to withdraw the lower layer with a Pasteur pipet.

A centrifuge tube has a large advantage over other methods of extraction. If an emulsion (Section 7.9) forms, you can use a centrifuge to aid in the separation of the layers.

You should check the capped centrifuge tube for leaks by filling it with water and shaking it vigorously. If it leaks, try replacing the cap with a different one. A **vortex mixer,** if available, provides an alternative to shaking the tube. In fact, a vortex mixer works well with a variety of containers, including small flasks, test tubes, conical vials, and centrifuge tubes. You start the mixing action on a vortex mixer by holding the test tube or other container on one of the neoprene pads. The unit mixes the sample by high-frequency vibration.

7.8 DRYING AGENTS

After an organic solvent has been shaken with an aqueous solution, it will be "wet"; that is, it will have dissolved some water even though its miscibility with water is not great. The amount of water dissolved varies from solvent to solvent: diethyl ether represents a solvent in which a fairly large amount of water dissolves (about 1.5% by weight). To remove water from a wet ether solution obtained from an extraction, a **drying agent** is used. A drying agent is an insoluble, *anhydrous* inorganic salt that acquires waters of hydration when exposed to moist air or a wet solution:

$$\underset{\substack{\textbf{Anhydrous}\\ \textbf{drying agent}}}{\overset{\textbf{Insoluble}}{Na_2SO_4(s)}} + \text{WET SOLUTION } (nH_2O) \longrightarrow \underset{\substack{\textbf{Hydrated}\\ \textbf{drying agent}}}{\overset{\textbf{Insoluble}}{Na_2SO_4 \cdot nH_2O\ (s)}} + \text{DRY SOLUTION}$$

The insoluble drying agent is placed directly into the solution, where it acquires water molecules and becomes hydrated. If enough drying agent is used, all of the water can be removed from a wet solution, making it "dry," or free of water. The following anhydrous salts are commonly used: sodium sulfate, magnesium sulfate, calcium chloride, calcium sulfate (Drierite), and potassium carbonate.

These drying agents differ in their properties and applications. Two major properties are defined for a drying agent: capacity and completeness. **Capacity** refers to the amount of water a drying agent absorbs per unit weight (usually per gram). **Completeness** refers to a compound's effectiveness in removing all of the water from a solution by the time equilibrium has been reached with the solution. Sodium and magnesium sulfates both absorb a large amount of water (capacity), but magnesium sulfate usually leaves less water in the solution (completeness). You can equalize this difference by drying with several smaller portions of sodium sulfate, where you might use one large portion of magnesium sulfate. A third sulfate compound, calcium sulfate, dries a solution completely, but has a low capacity (you would have to use a lot more of it than of magnesium sulfate, but a single drying might suffice).

Many magnesium compounds are good drying agents, but magnesium ion is a strong Lewis acid, and it can sometimes cause rearrangements of sensitive compounds such as epoxides. Calcium chloride is a good drying agent but cannot be used with many compounds containing oxygen or nitrogen because it forms complexes. Calcium chloride absorbs methanol and ethanol in addition to water, so it is also useful for removing these solvents when they are present as impurities or they make a solution "wet." Potassium carbonate is a base and is used for drying solutions of basic substances. Potassium hydroxide is often used to dry solutions of aromatic amines.

Anhydrous sodium sulfate is the most widely used drying agent. The granular variety is recommended, because it is easier to remove the dried solution from it than from the powdered variety. Sodium sulfate is mild and effective. It will remove water from most common solvents, with the possible exception of diethyl ether, in which case a prior drying with saturated salt solution may be advised (see p. 699). Sodium sulfate must be used at room temperature to be effective; it cannot be used with boiling solutions. Table 7.2 compares the various common drying agents.

Large-Scale Reactions. To dry a large amount of solution, you should add enough granular anhydrous sodium sulfate to give a 1–3 mm layer on the bottom of the flask, depending on the volume of the solution. Dry the solution for at least 15 minutes, occasionally swirling the flask. The mixture is dry if it appears clear and shows the common signs of a dry solution given in Table 7.3.

If the solution remains cloudy after treatment with the first batch of drying agent, add more drying agent and repeat the drying procedure. If the drying agent clumps badly, with no drying agent that will flow freely when the flask is swirled, you should transfer the solution to a clean, dry flask and add a fresh portion of drying agent. If droplets of water are present, you should also transfer the solution to a new container and add a fresh quantity of drying agent. You should not add more drying agent if a puddle (water layer) forms. Instead, separate the layers, using a separatory funnel if necessary, and then add fresh drying agent to the organic layer, placed in a new container. If the first batch of drying agent does not completely dry the solution, it is not at all unusual to have to transfer (decant) the solution to a clean, dry flask and add a fresh portion of drying agent.

TABLE 7.2 Common Drying Agents

	Acidity	Hydrated	Capacity*	Completeness†	Rate‡	Use
Magnesium sulfate	Neutral	$MgSO_4 \cdot 7H_2O$	High	Medium	Rapid	General
Sodium sulfate	Neutral	$Na_2SO_4 \cdot 7H_2O$ $Na_2SO_4 \cdot 10H_2O$	High	Low	Medium	General
Calcium chloride	Neutral	$CaCl_2 \cdot 2H_2O$ $CaCl_2 \cdot 6H_2O$	Low	High	Rapid	Hydrocarbons Halides
Calcium sulfate (Drierite)	Neutral	$CaSO_4 \cdot \frac{1}{2}H_2O$ $CaSO_4 \cdot 2H_2O$	Low	High	Rapid	General
Potassium carbonate	Basic	$K_2CO_3 \cdot 1\frac{1}{2}H_2O$ $K_2CO_3 \cdot 2H_2O$	Medium	Medium	Medium	Amines, esters, bases, ketones
Potassium hydroxide	Basic	—	—	—	Rapid	Amines only
Molecular sieves (3 or 4 Å)	Neutral	—	High	Extremely high	—	General

*Amount of water removed per given weight of drying agent.

†Refers to amount of H_2O still in solution at equilibrium with drying agent.

‡Refers to rate of action (drying).

When the solution is dry, the drying agent should be removed by using decantation (pouring carefully to leave the drying agent behind). With granular sodium sulfate, decantation is quite easy to perform because of the size of the drying agent particles. If a powdered drying agent, such as magnesium sulfate, is used, it may be necessary to use a gravity filtration (Technique 4, Section 4.1B, p. 636) to remove the drying agent. The solvent is removed by distillation (Technique 8, Section 8.3, p. 710) or evaporation (Technique 3, Section 3.11, p. 630)

Microscale Reactions. Before attempting to dry an organic layer, check closely to see that there are no visible signs of water. If you see droplets of water in the organic layer or water droplets clinging to the sides of the conical vial or test tube, transfer the organic layer

TABLE 7.3 Common Signs That Indicate a Solution Is Dry

1. There are no visible water droplets on the side of flask or suspended in solution.
2. There is not a separate layer of liquid or a "puddle."
3. The solution is clear, not cloudy. Cloudiness indicates water is present.
4. The drying agent (or a portion of it) flows freely on the bottom of the container when stirred or swirled and does not "clump" together as a solid mass.

with a *dry* Pasteur pipet to a *dry* container before adding any drying agent. Now add one spatulaful of granular anhydrous sodium sulfate (or other drying agent) from the V-grooved end of a microspatula into a solution contained in a conical vial or test tube. If all the drying agent "clumps," add another spatulaful of sodium sulfate. Dry the solution for at least 15 minutes. Stir the mixture occasionally with a spatula during that period. The mixture is dry if there are no visible signs of water and the drying agent flows freely in the container when stirred with a microspatula. The solution should not be cloudy. Add more drying agent if necessary. You should not add more drying agent if a "puddle" (water layer) forms or if drops of water are visible. Instead, you should transfer the organic layer to a dry container before adding fresh drying agent. When dry, use a *dry* Pasteur pipet or a *dry* filter-tip pipet (Technique 4, Section 4.6, p. 645) to remove the solution from the drying agent and transfer the solution to a *dry* conical vial. Rinse the drying agent with a small amount of fresh solvent and transfer this solvent to the vial containing the solution. Remove the solvent by evaporation using heat and a stream of air or nitrogen (Technique 3, Section 3.11, p. 630).

An alternative method of drying an organic phase is to pass it through a filtering pipet (Technique 4, Section 4.1C, p. 638) that has been packed with a small amount (ca. 2 cm) of drying agent. Again, the solvent is removed by evaporation.

Saturated Salt Solution. At room temperature, diethyl ether (ether) dissolves 1.5% by weight of water, and water dissolves 7.5% of ether. Ether, however, dissolves a much smaller amount of water from a saturated aqueous sodium chloride solution. Hence, the bulk of water in ether, or ether in water, can be removed by shaking it with a saturated aqueous sodium chloride solution. A solution of high ionic strength is usually not compatible with an organic solvent and forces separation of it from the aqueous layer. The ether phase (organic layer) will be on top, and the saturated sodium chloride solution will be on the bottom (d = 1.2 g/mL). After removing the organic phase from the aqueous sodium chloride, dry the organic layer completely with sodium sulfate or with one of the other drying agents listed in Table 7.2.

7.9 EMULSIONS

An **emulsion** is a colloidal suspension of one liquid in another. Minute droplets of an organic solvent are often held in suspension in an aqueous solution when the two are mixed or shaken vigorously; these droplets form an emulsion. This is especially true if any gummy or viscous material was present in the solution. Emulsions are often encountered in performing extractions. Emulsions may require a long time to separate into two layers and are a nuisance to the organic chemist.

Fortunately, several techniques may be used to break a difficult emulsion once it has formed.

1. Often, an emulsion will break up if it is allowed to stand for some time. Patience is important here. Gently stirring with a stirring rod or spatula may also be useful.
2. If one of the solvents is water, adding a saturated aqueous sodium chloride solution will help destroy the emulsion. This makes the aqueous and organic layers less compatible, thereby forcing separation.

3. With microscale and many small-scale experiments, the mixture may be transferred to a centrifuge tube. The emulsion will often break during centrifugation. Remember to place another tube filled with water on the opposite of the centrifuge to balance it.
4. Adding a very small amount of a water-soluble detergent may also help. This method has been used in the past for combating oil spills. The detergent helps to solubilize the tightly bound oil droplets.
5. Gravity filtration (see Technique 4, Section 4.1, p. 634) may help to destroy an emulsion by removing a gummy polymeric substance. With large-scale reactions, you might try filtering the mixture through a fluted filter (Technique 4, Section 4.1B, p. 636) or a piece of cotton. With microscale reactions, a filtering pipet may work (Technique 4, Section 4.1C, p. 638). In many cases, once the gum is removed, the emulsion breaks up rapidly.
6. If you are using a separatory funnel, you might try to use a gentle swirling action in the funnel to help break an emulsion. Gently stirring with a stirring rod may also be useful.

When you know through prior experience that a mixture may form a difficult emulsion, you should avoid shaking the mixture vigorously. When using conical vials for extractions, it may be better to use a magnetic spin vane for mixing and not shake the mixture at all. When using separatory funnels, extractions should be performed with gentle swirling instead of shaking or with several gentle inversions of the separatory funnel. Do not shake the separatory funnel vigorously in these cases. It is important that you use a longer extraction period if the more gentle techniques described in this paragraph are being employed. Otherwise, you will not transfer all the material from the first phase to the second one.

7.10 PURIFICATION AND SEPARATION METHODS

In nearly all the synthetic experiments undertaken in this textbook, a series of operations involving extractions is used after the actual reaction has been concluded. These extractions form an important part of the purification. Using them, the desired product is separated from unreacted starting materials or from undesired side products in the reaction mixture. These extractions may be grouped into three categories, depending on the nature of the impurities they are designed to remove.

The first category involves extracting or "washing" an organic mixture with water. Water washes are designed to remove highly polar materials, such as inorganic salts, strong acids or bases, and low-molecular-weight, polar substances including alcohols, carboxylic acids, and amines. Many organic compounds containing fewer than five carbons are water-soluble. Water extractions are also used immediately following extractions of a mixture with either acid or base to ensure that all traces of acid or base have been removed.

The second category concerns extraction of an organic mixture with a dilute acid, usually 5 to 10% hydrochloric acid. Acid extractions are intended to remove basic impurities, especially such basic impurities as organic amines. The bases are converted to their corresponding cationic salts by the acid used in the extraction. If an amine is one of the reactants, or if pyridine or another amine is a solvent, such an extraction might be used to remove any excess amine present at the end of a reaction.

$$RNH_2 + HCl \longrightarrow \underset{\text{(Water-soluble salt)}}{RNH_3^+Cl^-}$$

Cationic salts are usually soluble in the aqueous solution, and they are thus extracted from the organic material. A water extraction may be used immediately following the acid extraction to ensure that all traces of the acid have been removed from the organic material.

The third category is extraction of an organic mixture with a dilute base, usually 5% sodium bicarbonate, although extractions with dilute sodium hydroxide can also be used. Such basic extractions are intended to convert acidic impurities, such as organic acids, to their corresponding anionic salts. For example, in the preparation of an ester, a sodium bicarbonate extraction might be used to remove any excess carboxylic acid that is present.

$$\underset{(pKa \sim 5)}{RCOOH} + NaHCO_3 \longrightarrow \underset{\text{(Water-soluble salt)}}{RCOO^-Na^+} + H_2O + CO_2$$

Anionic salts, being highly polar, are soluble in the aqueous phase. As a result, these acidic impurities are extracted from the organic material into the basic solution. A water extraction may be used after the basic extraction to ensure that all the base has been removed from the organic material.

Occasionally, phenols may be present in a reaction mixture as impurities, and removing them by extraction may be desired. Because phenols, although they are acidic, are about 10^5 times less acidic than carboxylic acids, basic extractions may be used to separate phenols from carboxylic acids by a careful selection of the base. If sodium bicarbonate is used as a base, carboxylic acids are extracted into the aqueous base, but phenols are not. Phenols are not sufficiently acidic to be deprotonated by the weak base, bicarbonate. Extraction with sodium hydroxide, on the other hand, extracts both carboxylic acids and phenols into the aqueous basic solution, because hydroxide ion is a sufficiently strong base to deprotonate phenols.

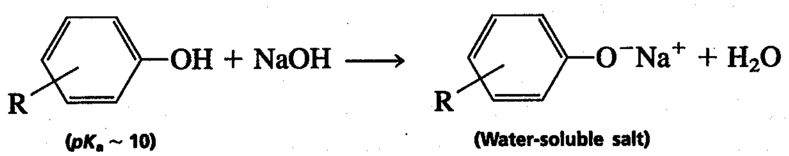

It would be a useful exercise for you to examine the instructions for some of the preparative experiments given in the textbook. While you are examining these procedures, you should try to identify which impurities are being removed at each extraction step. Mixtures of acidic, basic, and neutral compounds are easily separated by extraction techniques. One such example is shown in Figure 7.9.

Materials that have been extracted can be regenerated by neutralizing the extraction reagent. If an acidic material has been extracted with aqueous base, the material can be

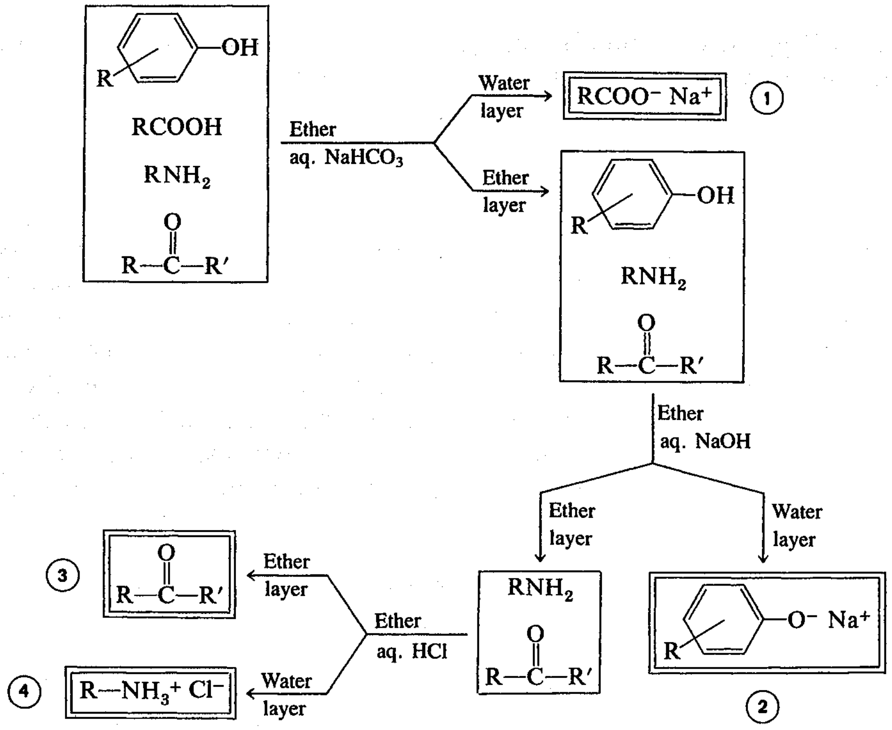

Figure 7.9 Separating a four-component mixture by extraction.

regenerated by acidifying the extract until the solution becomes acidic to blue litmus. The material will separate from the acidified solution. A basic material can be recovered from the acidic extract by adding base to the extract. These substances can then be removed from the neutralized aqueous solutions by extraction with an organic solvent such as ether. After the ether phase is dried with a drying agent, evaporation of the ether yields the isolated compounds.

7.11 CONTINUOUS SOLID–LIQUID EXTRACTION

The technique of liquid–liquid extraction was described in Sections 7.1–7.7. In this section, solid–liquid extraction is described. Solid–liquid extraction is often used to extract a solid natural product from a natural source, such as a plant. A solvent is chosen that selectively dissolves the desired compound but that leaves behind the undesired insoluble solid. A continuous solid–liquid extraction apparatus, called a Soxhlet extractor, is commonly used in a research laboratory (Fig. 7.10).

As shown in Figure 7.10, the solid to be extracted is placed in a thimble made from filter paper, and the thimble is inserted into the central chamber. A low boiling solvent, such as diethyl ether, is placed in the round-bottom distilling flask and is heated to reflux. The vapor rises through the left sidearm into the condenser where it liquefies. The con-

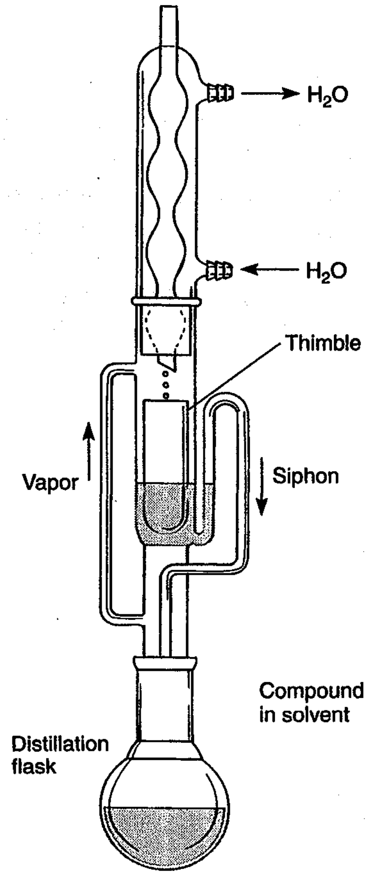

Figure 7.10 *Continuous solid–liquid extraction using a Soxhlet extractor.*

densate (liquid) drips into the thimble containing the solid. The hot solvent begins to fill the thimble and extracts the desired compound from the solid. Once the thimble is filled with solvent, the sidearm on the right acts as a siphon, and the solvent, which now contains the dissolved compound, drains back into the distillation flask. The vaporization, condensation, extraction, siphoning process is repeated hundreds of times, and the desired product is concentrated in the distillation flask. The product is concentrated in the flask, because it has a boiling point higher than that of the solvent or because it is a solid.

7.12 CONTINUOUS LIQUID–LIQUID EXTRACTION

When a product is very soluble in water, it is often difficult to extract using the techniques described in Sections 7.4–7.7, because of an unfavorable distribution coefficient. In this case, you need to extract the aqueous solution numerous times with fresh batches of an immiscible organic solvent to remove the desired product from water. A less labor-intensive technique involves the use of a continuous liquid–liquid extraction apparatus. One type of extractor, used with solvents that are less dense than water, is shown in Figure 7.11. Diethyl ether is usually the solvent of choice.

The aqueous phase is placed in the extractor, which is then filled with diethyl ether up to the sidearm. The round-bottom distillation flask is partially filled with ether. The ether is heated to reflux in the round-bottom flask, and the vapor is liquefied in the water-cooled condenser. The ether drips into the central tube, passes through the porous sintered glass tip, and flows through the aqueous layer. The solvent extracts the desired compound from the aqueous phase, and the ether is recycled back into the round-bottom flask. The product is concentrated in the flask. The extraction is rather inefficient and must be placed in operation for at least 24 hours to remove the compound from the aqueous phase.

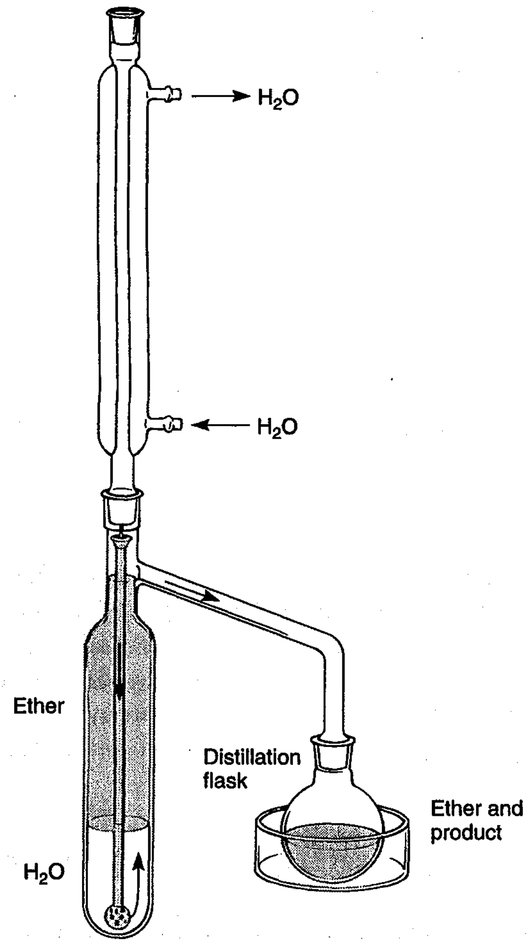

Figure 7.11 Continuous liquid–liquid extraction using a solvent less dense than water.

PROBLEMS

1. Suppose solute A has a distribution coefficient of 1.0 between water and diethyl ether. Demonstrate that if 100 mL of a solution of 5.0 g of A in water were extracted with two 25-mL portions of ether, a smaller amount of A would remain in the water than if the solution were extracted with one 50-mL portion of ether.

2. Write an equation to show how you could recover the parent compounds from their respective salts (1, 2, and 4) shown in Figure 7.9.

3. Aqueous hydrochloric acid was used *after* the sodium bicarbonate and sodium hydroxide extractions in the separation scheme shown in Figure 7.9. Is it possible to use this reagent earlier in the separation scheme so as to achieve the same overall result? If so, explain where you would perform this extraction.

4. Using aqueous hydrochloric acid, sodium bicarbonate, or sodium hydroxide solutions, devise a separation scheme using the style shown in Figure 7.9 to separate the following two-component mixtures. All the substances are soluble in ether. Also indicate how you would recover each of the compounds from their respective salts.

(a) Give two different methods for separating this mixture.

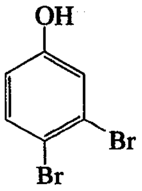

$(CH_3CH_2CH_2CH_2)_3N$

(b) Give two different methods for separating this mixture.

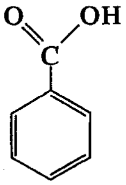

$CH_3CH_2CH_2CH_2CH_2CH_2OH$

(c) Give one method for separating this mixture.

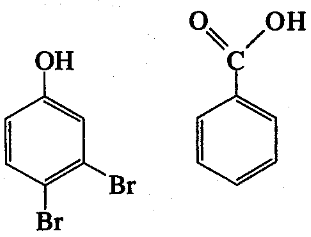

5. Solvents other than those in Table 7.1 may be used for extractions. Determine the relative positions of the organic layer and the aqueous layer in a conical vial or separatory funnel after shaking each of the following solvents with an aqueous phase. Find the densities for each of these solvents in a handbook (Technique 20. Section 20.1, p. 861).

(a) 1,1,1-Trichloroethane
(b) Hexane

6. A student prepares ethyl benzoate by the reaction of benzoic acid with ethanol using a sulfuric acid catalyst. The following compounds are found in the crude reaction mixture: ethyl benzoate (major component), benzoic acid, ethanol, and sulfuric acid. Using a handbook, obtain the solubility properties in water for each of these compounds (Technique 20, Section 20.1, p. 861). Indicate how you would remove benzoic acid, ethanol, and sulfuric acid from ethyl benzoate. At some point in the purification, you should also use an aqueous sodium bicarbonate solution.

7. Calculate the weight of water that could be removed from a wet organic phase using 50.0 mg of magnesium sulfate. Assume that it gives the hydrate listed in Table 7.2.

8. Explain the following laboratory instructions in a procedure:
(a) "Wash the organic layer with 5.0 mL of 5% aqueous sodium bicarbonate."
(b) "Extract the aqueous layer three times with 2-mL portions of methylene chloride."

9. Just prior to drying an organic layer with a drying agent, you notice water droplets in the organic layer. What should you do next?

10. What should you do if there is some question about which layer is the organic one during an extraction procedure?

11. Saturated aqueous sodium chloride (d = 1.2 g/mL) is added to the following mixtures in order to dry the organic layer. Which layer is likely to be on the bottom in each case?

(a) Sodium chloride layer or a layer containing a high-density organic compound dissolved in methylene chloride (d = 1.4 g/mL)

(b) Sodium chloride layer or a layer containing a low-density organic compound dissolved in methylene chloride (d = 1.1 g/mL)

TECHNIQUE 8

Simple Distillation

Distillation is the process of vaporizing a liquid, condensing the vapor, and collecting the condensate in another container. This technique is very useful for separating a liquid mixture when the components have different boiling points, or when one of the components will not distill. It is one of the principal methods of purifying a liquid. Four basic distillation methods are available to the chemist: simple distillation, vacuum distillation (distillation at reduced pressure), fractional distillation, and steam distillation. This technique chapter will discuss simple distillation. Vacuum distillation will be discussed in Technique 9. Fractional distillation will be discussed in Technique 10, and steam distillation will be discussed in Technique 11.

8.1 THE EVOLUTION OF DISTILLATION EQUIPMENT

There are probably more types and styles of distillation apparatus than exist for any other technique in chemistry. Over the centuries, chemists have devised just about every conceivable design. The earliest known types of distillation apparatus were the **alembic** and the **retort** (Fig. 8.1). They were used by alchemists in the Middle Ages and the

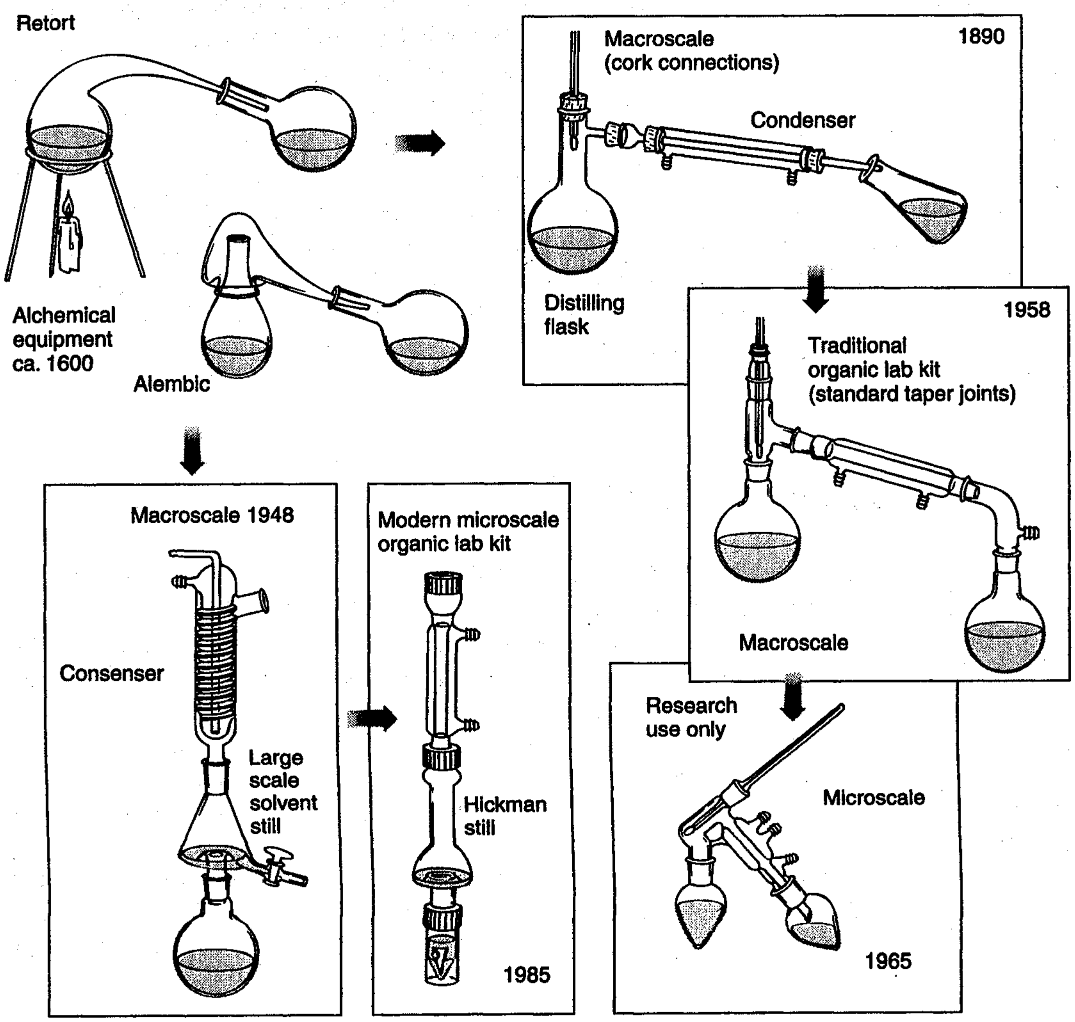
Retort
Alchemical equipment ca. 1600
Alembic
Macroscale (cork connections)
1890
Condenser
Distilling flask
1958
Traditional organic lab kit (standard taper joints)
Macroscale
Macroscale 1948
Consenser
Large scale solvent still
Modern microscale organic lab kit
Hickman still
1985
Research use only
Microscale
1965

Figure 8.1 Some stages in the evolution of distillation equipment from alchemical equipment (dates represent approximate time of use).

Renaissance, and probably even earlier by Arabic chemists. Most other distillation equipment has evolved as variations on these designs.

Figure 8.1 shows several stages in the evolution of distillation equipment as it relates to the organic laboratory. It is not intended to be a complete history; rather, it is representative. Up until recent years, equipment based on the retort design was common in the laboratory. Although the retort itself was still in use early in this century, it had evolved by that time into the distillation flask and water-cooled condenser combination. This early equipment was connected with drilled corks. By 1958, most introductory laboratories were beginning to use "organic lab kits" that included glassware connected by standard-taper glass joints. The original lab kits contained large ₸ 24/40 joints. Within a short time, they became smaller with ₸ 19/22 and even ₸ 14/20 joints. These later kits are still being used today in many "macroscale" laboratory courses such as yours.

In the 1960s, researchers developed even smaller versions of these kits for working at the "microscale" level (in Fig. 8.1, see the box labeled 1965, research use only), but this glassware is generally too expensive to use in an introductory laboratory. However, in the mid 1980s, several groups developed a different style of microscale distillation equipment based on the alembic design (see the box labeled 1985, Hickman still). This new microscale equipment has Ŧ 14/10 standard-taper joints, threaded outer joints with screwcap connectors, and an internal O-ring for a compression seal. Microscale equipment similar to this is now used in many introductory courses. The advantages of this glassware are that there is less material used (lower cost), lower personal exposure to chemicals, and less waste generated. Because both types of equipment are in use today, after we describe macroscale equipment, we will also show the equivalent microscale distillation apparatus.

8.2 DISTILLATION THEORY

In the traditional distillation of a pure substance, vapor rises from the distillation flask and comes into contact with a thermometer that records its temperature. The vapor then passes through a condenser, which reliquefies the vapor and passes it into the receiving flask. The temperature observed during the distillation of a **pure substance** remains constant throughout the distillation so long as both vapor *and* liquid are present in the system (see Fig. 8.2A). When a **liquid mixture** is distilled, often the temperature does not remain constant but increases throughout the distillation. The reason for this is that the composition of the vapor that is distilling varies continuously during the distillation (see Fig. 8.2B).

For a liquid mixture, the composition of the vapor in equilibrium with the heated solution is different from the composition of the solution itself. This is shown in Figure 8.3, which is a phase diagram of the typical vapor–liquid relation for a two-component system (A + B).

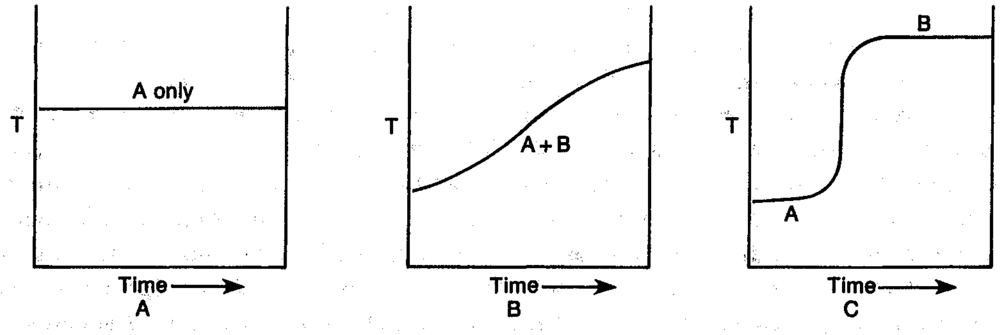

Figure 8.2 Three types of temperature behavior during a simple distillation. (A) A single pure component. (B) Two components of similar boiling points. (C) Two components with widely differing boiling points. Good separations are achieved in A and C.

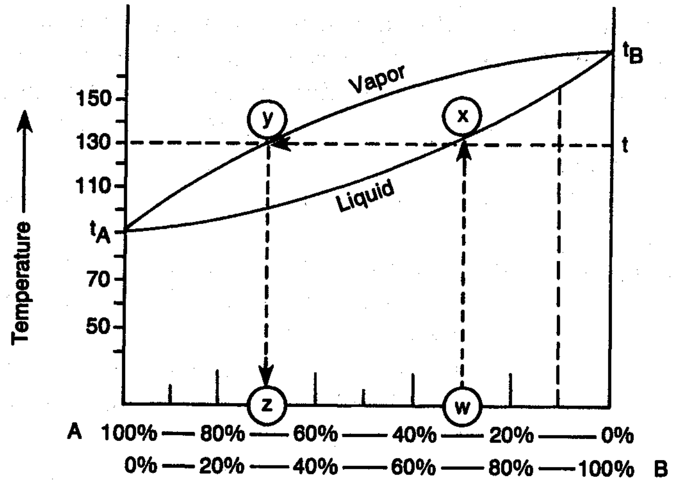

Figure 8.3 Phase diagram for a typical liquid mixture of two components.

On this diagram, horizontal lines represent constant temperatures. The upper curve represents vapor composition, and the lower curve represents liquid composition. For any horizontal line (constant temperature), like that shown at *t*, the intersections of the line with the curves give the compositions of the liquid and the vapor that are in equilibrium with each other at that temperature. In the diagram, at temperature *t*, the intersection of the curve at *x* indicates that liquid of composition *w* will be in equilibrium with vapor of composition *z*, which corresponds to the intersection at *y*. Composition is given as a mole percentage of A and B in the mixture. Pure A, which boils at temperature t_A, is represented at the left. Pure B, which boils at temperature t_B, is represented on the right. For either pure A or pure B, the vapor and liquid curves meet at the boiling point. Thus, either pure A or pure B will distill at a constant temperature (t_A or t_B). Both the vapor and the liquid must have the same composition in either of these cases. This is not the case for mixtures of A and B.

A mixture of A and B of composition *w* will have the following behavior when heated. The temperature of the liquid mixture will increase until the boiling point of the mixture is reached. This corresponds to following line *wx* from *w* to *x*, the boiling point of the mixture *t*. At temperature *t* the liquid begins to vaporize, which corresponds to line *xy*. The vapor has the composition corresponding to *z*. In other words, the first vapor obtained in distilling a mixture of A and B does not consist of pure A. It is richer in A than the original mixture but still contains a significant amount of the higher boiling component B, *even from the very beginning of the distillation.* The result is that it is never possible to separate a mixture completely by a simple distillation. However, in two cases it is possible to get an acceptable separation into relatively pure components. In the first case, if the boiling points of A and B differ by a large amount (>100°), and if the distillation is carried out carefully, it will be possible to get a fair separation of A and B. In the second case, if A contains a fairly small amount of B (<10%), a reasonable separation of A from B can be achieved. When the boiling-point differences are not large, and when highly pure components are desired, it is necessary to perform a **fractional distil-**

lation. Fractional distillation is described in Technique 10, where the behavior during a simple distillation is also considered in detail. Note only that as vapor distills from the mixture of composition *w* (Fig. 8.3), it is richer in A than is the solution. Thus, the composition of the material left behind in the distillation becomes richer in B (moves to the right from *w* toward pure B in the graph). A mixture of 90% B (dotted line on the right side in Fig. 8.3) has a higher boiling point than at *w*. Hence, the temperature of the liquid in the distillation flask will increase during the distillation, and the composition of the distillate will change (as is shown in Fig. 8.2B).

When two components that have a large boiling-point difference are distilled, the temperature remains constant while the first component distills. If the temperature remains constant, a relatively pure substance is being distilled. After the first substance distills, the temperature of the vapors rises, and the second component distills, again at a constant temperature. This is shown in Figure 8.2C. A typical application of this type of distillation might be an instance of a reaction mixture containing the desired component A (bp 140°C) contaminated with a small amount of undesired component B (bp 250°C) and mixed with a solvent such as diethyl ether (bp 36°C). The ether is removed easily at low temperature. Pure A is removed at a higher temperature and collected in a separate receiver. Component B can then be distilled, but it usually is left as a residue and not distilled. This separation is not difficult and represents a case where simple distillation might be used to advantage.

8.3 SIMPLE DISTILLATION—STANDARD APPARATUS

For a simple distillation, the apparatus shown in Figure 8.4 is used. Six pieces of specialized glassware are used:

1. Distilling flask
2. Distillation head
3. Thermometer adapter
4. Water condenser
5. Vacuum takeoff adapter
6. Receiving flask

The apparatus is usually heated electrically, using a heating mantle. The distilling flask, condenser, and vacuum adapter should be clamped. Two different methods of clamping this apparatus were shown in Technique 3 (Fig. 3.4, page 616 and Fig. 3.6, page 617). The receiving flask should be supported by removable wooden blocks or a wire gauze on an iron ring attached to a ring stand. The various components are each discussed below, along with some other important points.

Distilling Flask. The distilling flask should be a round-bottom flask. This type of flask is designed to withstand the required input of heat, and to accommodate the boiling action. It gives a maximized heating surface. The size of the distilling flask should be chosen so that it is never filled more than two-thirds full. When the flask is filled beyond this point, the neck constricts and "chokes" the boiling action, resulting in bumping. The sur-

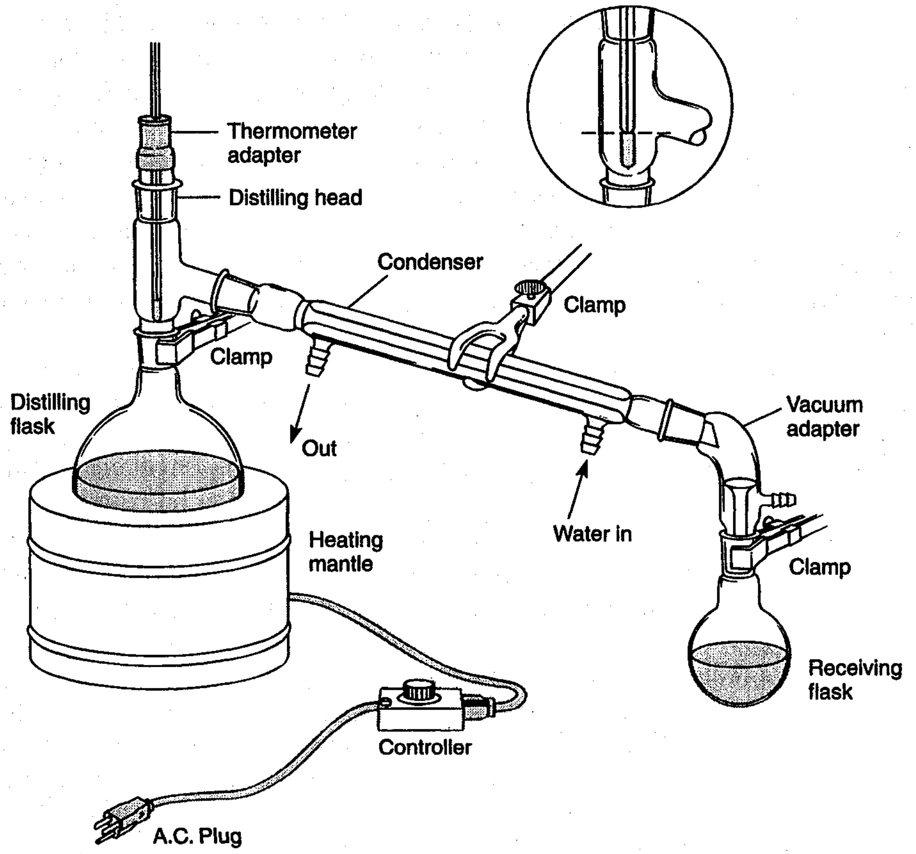
Thermometer adapter
Distilling head
Condenser
Clamp
Clamp
Distilling flask
Out
Vacuum adapter
Heating mantle
Water in
Clamp
Receiving flask
Controller
A.C. Plug

Figure 8.4 Distillation with the standard macroscale lab kit.

face area of the boiling liquid should be kept as large as possible. However, too large a distilling flask should also be avoided. With too large a flask, the **holdup** is excessive; the holdup is the amount of material that cannot distill since some vapor must fill the empty flask. When you cool the apparatus at the end, this material drops back into the distilling flask.

Boiling Stones. A boiling stone (Technique 3, Section 3.6, page 622) should be used during distillation to prevent bumping. As an alternative, the liquid being distilled may be rapidly stirred using a magnetic stirrer and stir bar (Technique 3, Section 3.5, page 621). If you forget a boiling stone, cool the mixture before adding it. If you add a boiling stone to a hot superheated liquid, it may "erupt" into vigorous boiling, breaking your apparatus and spilling hot solvent everywhere.

Grease. In most cases, it is unnecessary to grease standard-taper joints for a simple distillation. The grease makes cleanup more difficult, and it may contaminate your product.

Distillation Head. The distillation head directs the distilling vapors into the condenser and allows the connection of a thermometer via the thermometer adapter. The thermometer should be positioned in the distillation head so that it is placed directly in the stream of vapor that is distilling. This can be accomplished if the entire bulb of the thermometer is positioned *below* the side arm of the distilling head (see the circular inset in Fig. 8.4). The entire bulb must be immersed in the vapor to achieve an accurate temperature reading. When distilling, you should be able to see a reflux ring (Technique 3, Section 3.4, page 619) positioned well above both the thermometer bulb and the bottom of the side arm.

Thermometer Adapter. The thermometer adapter connects to the top of the distillation head (see Fig. 8.4). There are two parts to the thermometer adapter: a glass joint with an open rolled edge on the top, and a rubber adapter that fits over the rolled edge and holds the thermometer. The thermometer fits in a hole in the top of the rubber adapter and can be adjusted upward and downward by sliding it in the hole. Adjust the bulb to a point below the side arm.

Water Condenser. The joint between the distillation head and the water condenser is the joint most prone to leak in this entire apparatus. Since the distilling liquid is both hot and vaporized when it reaches this joint, it will leak out of any small opening between the two joint surfaces. The odd angle of the joint, neither vertical or horizontal, also makes a good connection more difficult. Be sure this joint is well sealed. If possible, use one of the plastic joint clips described in Technique 3, Figure 3.5, page 617. Otherwise, adjust your clamps to be sure that the joint surfaces are pressed together, and not pulled apart.

The condenser will remain full of cooling water only if the water flows *upward,* not downward. The water input hose should be connected to the lower opening in the jacket, and the exit hose should be attached to the upper opening. Place the other end of the exit hose in a sink. A moderate water flow will perform a good deal of cooling. A high rate of water flow may cause the tubing to pop off the joints and cause a flood. If you hold the exit hose horizontally and point the end into a sink, the flow rate is correct if the water stream continues horizontally for about two inches before bending downward.

If a distillation apparatus is to be left untended for a period of time, it is a good idea to wrap copper wire around the ends of the tubing and twist it tight. This will help to prevent the hoses from popping off of the connectors if there is an unexpected water pressure change.

Vacuum Adapter. In a simple distillation the vacuum adapter is not connected to a vacuum, but is left open. It is merely an opening to the outside air so that pressure does not build up in the distillation system. If you plug this opening, you will have a **closed system** (no outlet). It is always dangerous to heat a closed system. Enough pressure can build up in the closed system so that it can explode. The vacuum adapter, in this case, merely directs the distillate into the receiving, or collection, flask.

If the substance you are distilling is water sensitive, you can attach a calcium chloride drying tube to the vacuum connection to protect the freshly distilled liquid from atmospheric water vapor. Air that enters the apparatus will have to pass through the calcium chloride and be dried. Depending on the severity of the problem, drying agents other than calcium chloride may also be used.

The vacuum adapter has a disturbing tendency to obey the laws of Newtonian physics and fall off the slanted condenser onto the desk and break. If they are available, it is a good idea to use plastic joint clips on both ends of this piece. The top clip will secure the vacuum adapter to the condenser, and the bottom clip will secure the receiving flask, preventing it from falling.

Rate of Heating. The rate of heating for the distillation can be adjusted to the proper rate of **takeoff,** the rate at which distillate leaves the condenser, by watching drops of liquid emerge from the bottom of the vacuum adapter. A rate of from one to three drops per second is considered a proper rate of takeoff for most applications. At a greater rate, equilibrium is not established within the distillation apparatus, and the separation may be poor. A slower rate of takeoff is also unsatisfactory since the temperature recorded on the thermometer is not maintained by a constant vapor stream, thus leading to an inaccurately low observation of the boiling point.

Receiving Flask. The receiving flask collects the distilled liquid, and is usually a round-bottom flask. If the liquid you are distilling is extremely volatile, and there is danger of losing some of it to evaporation, it is sometimes advisable to cool the receiving flask in an ice-water bath.

Fractions. The material being distilled is called the **distillate.** Frequently a distillate is collected in contiguous portions, called **fractions.** This is accomplished by replacing the collection flask with clean ones at regular intervals. If a small amount of liquid is collected at the beginning of a distillation and not saved or used further, it is called a **forerun.** Subsequent fractions will have higher boiling ranges, and each fraction should be labeled with its correct boiling range when the fraction is taken. For a simple distillation of a pure material, most of the material will be collected in a single large **midrun** fraction, with only a small forerun. In some small-scale distillations, the volume of the forerun will be so small that you will not be able to collect it separately from the midrun fraction. The material left behind is called the **residue.** It is usually advised that you discontinue a distillation before the distilling flask becomes empty. Typically, the residue becomes increasingly dark in color during distillation, and it frequently contains thermal decomposition products. In addition, a dry residue may explode on overheating, or the flask may melt or crack when it becomes dry. Don't distill until the distilling flask is completely dry!

8.4 MICROSCALE AND SEMI-MICROSCALE EQUIPMENT

When you wish to distill quantities that are smaller than 4–5 mL, different equipment is required. What you use depends on how small a quantity you wish to distill.

A. SEMI-MICROSCALE

One possibility is to use equipment identical in style to that used with conventional macroscale procedures, but to "downsize" it using ₸ 14/10 joints. The major manufacturers do make distillation heads and vacuum takeoff adapters with ₸ 14/10 joints. This equipment will allow you to handle quantities of from 5 to 15 mL. An example of such a "semi-microscale" apparatus is given in Figure 8.5. Although the manufacturers make ₸ 14/10 condensers, the condenser has been left out in this example. This can be done if the material to be distilled is not extremely volatile or is high boiling. It is also possible to omit the condenser if you not have a large amount of material and can cool the receiving flask in an ice-water bath as shown in the figure.

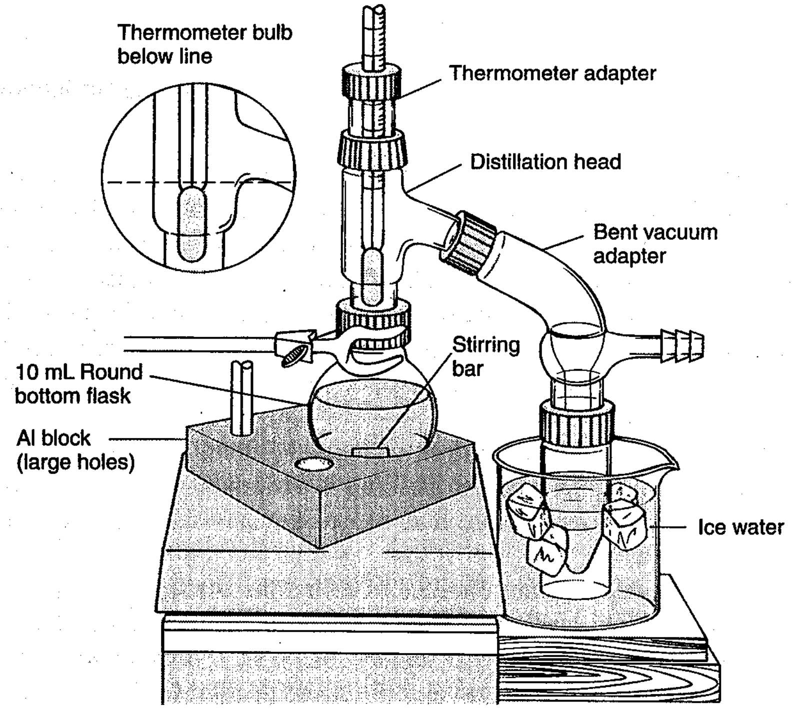
Thermometer bulb below line
Thermometer adapter
Distillation head
Bent vacuum adapter
Stirring bar
10 mL Round bottom flask
Al block (large holes)
Ice water

Figure 8.5 Semi-microscale distillation.

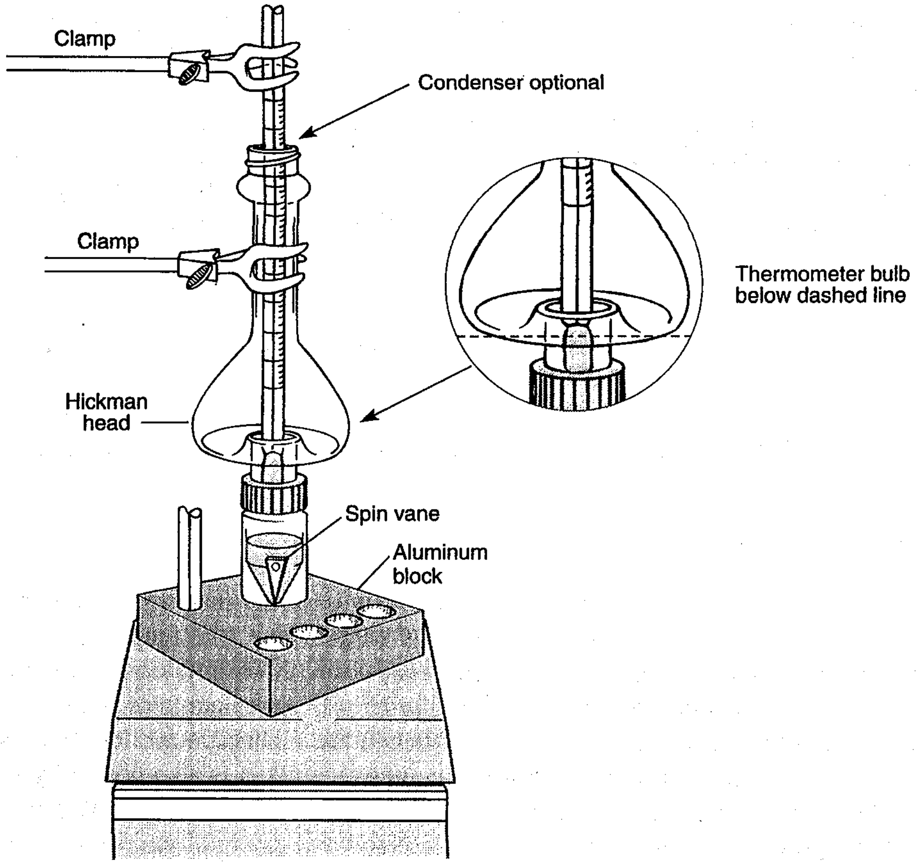
Clamp
Condenser optional
Clamp
Thermometer bulb below dashed line
Hickman head
Spin vane
Aluminum block

Figure 8.6 Basic microscale distillation.

B. MICROSCALE—STUDENT EQUIPMENT

Figure 8.6 shows the typical distillation setup for those students who are taking a microscale laboratory course. Instead of a distillation head, condenser, and vacuum takeoff, this equipment uses a single piece of glassware called a **Hickman head.** The Hickman head provides a "short path" for the distilled liquid to travel before it is collected. The liquid is boiled, moves upward through the central stem of the Hickman head, condenses on the walls of the "chimney," and then runs down the sides into the circular well surrounding the stem. With very volatile liquids, a condenser can be placed on top of the Hickman head to improve its efficiency. The apparatus shown uses a 5-mL conical vial as the distilling flask, meaning that this apparatus can distill from 1 to 3 mL of liquid. Unfortunately, the well in most Hickman heads holds only about 0.5–1.0 mL. That means the well must be emptied several times. This is done using a disposable Pasteur pipet as shown in Figure 8.7. The figure shows two different styles of Hickman head. The one with the side port makes removal of the distillate easier.

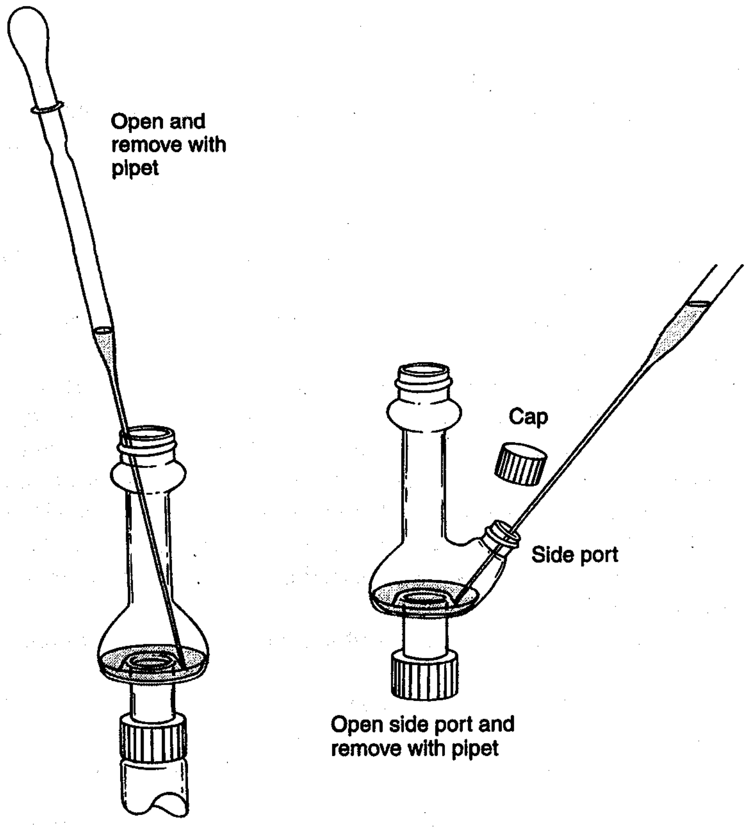

Figure 8.7 Two styles of Hickman head.

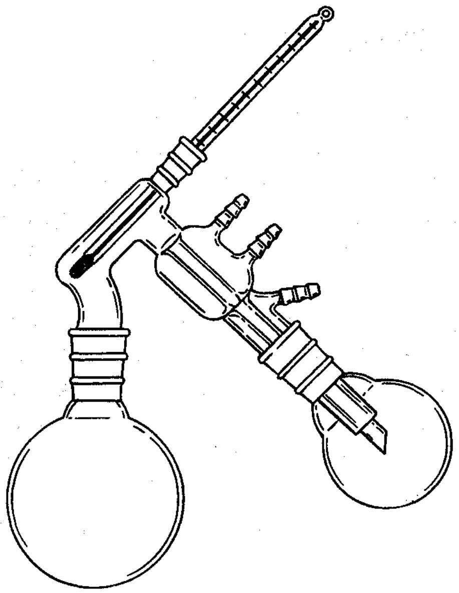

Figure 8.8 A research-style short-path distillation apparatus.

C. MICROSCALE—RESEARCH EQUIPMENT

Figure 8.8 shows a very-well designed research-style, short-path distillation head. Note how the equipment has been "unitized," eliminating several joints, and decreasing the holdup.

PROBLEMS

1. Using Figure 8.3, answer the following questions.

(a) What is the molar composition of the vapor in equilibrium with a boiling liquid that has a composition of 60% A and 40% B?

(b) A sample of vapor has the composition 50% A and 50% B. What is the composition of the boiling liquid that produced this vapor?

2. Use an apparatus similar to that shown in Figure 8.4, and assume that the round-bottom flask holds 100 mL and the distilling head has an internal volume of 12 mL in the vertical section. At the end of a distillation, vapor would fill this volume, but it could not be forced through the system. No liquid would remain in the distillation flask. Assuming this holdup volume of 112 mL, use the ideal gas law and assume a boiling point of 100°C (760 mm Hg) to calculate the number of milliliters of liquid ($d = 0.9$ g/mL, $MW = 200$) that would recondense into the distillation flask upon cooling.

3. Explain the significance of a horizontal line connecting a point on the lower curve with a point on the upper curve (such as line xy) in Figure 8.3.

4. Using Figure 8.3, determine the boiling point of a liquid having a molar composition of 50% A and 50% B.

5. Where should the thermometer bulb be located in

(a) a microscale distillation apparatus using a Hickman head?

(b) a macroscale distillation apparatus using a distilling head, condenser, and vacuum takeoff adapter?

6. Under what conditions can a good separation be achieved with a simple distillation?

TECHNIQUE 9

Vacuum Distillation, Manometers

Vacuum distillation (distillation at reduced pressure) is used for compounds that have high boiling points (above 200°C). Such compounds often undergo thermal decomposition at the temperatures required for their distillation at atmospheric pressure. The boiling point of a compound is lowered substantially by reducing the applied pressure. Vacuum distillation is also used for compounds that, when heated, might react with the oxygen present in air. It is also used when it is more convenient to distill at a lower temperature because of experimental limitations. For instance, a heating device may have difficulty heating to a temperature in excess of 250°C.

The effect of pressure on the boiling point is discussed more thoroughly in Technique 6 (Section 6.9, p. 674). A nomograph is given (Fig. 6.9, p. 676) that allows you to estimate the boiling point of a liquid at a pressure different from the one at which it is reported. For example, a liquid reported to boil at 200°C at 760 mmHg would be expected to boil at 90°C at 20 mmHg. This is a significant decrease in temperature, and it would be advantageous to use a vacuum distillation if any problems were to be expected. Counterbalancing this advantage, however, is the fact that separations of liquids of different boiling points may not be as effective with a vacuum distillation as with a simple distillation.

9.1 STANDARD SCALE METHODS

When working with glassware that is to be evacuated, you should wear safety glasses at all times. *There is always danger of an implosion.*

Caution: Safety glasses must be worn at all times during vacuum distillation.

It is a good idea to work in a hood when performing a vacuum distillation. If the experiment will involve high temperatures (>220°C) for distillation, or an extremely low pressure (<0.1 mmHg), for your own safety you should definitely work in a hood, behind a shield.

A basic apparatus similar to the one shown in Figure 9.1 may be used for vacuum distillations. The major differences to be found when comparing this assembly to one for simple distillation (Fig. 8.4, p. 711) are that a Claisen head has been inserted between the distillation flask and the distilling head, and that the opening to the atmosphere has been replaced by a connection **A** to a vacuum source. In addition, an air inlet tube **B** has been added to the top of the Claisen head. When connecting to a vacuum source, an aspirator (Technique 4, Section 4.5, p. 643), a mechanical vacuum pump (Section 9.6, p. 728), or a "house" vacuum system (one piped directly to the laboratory bench) may be used. The

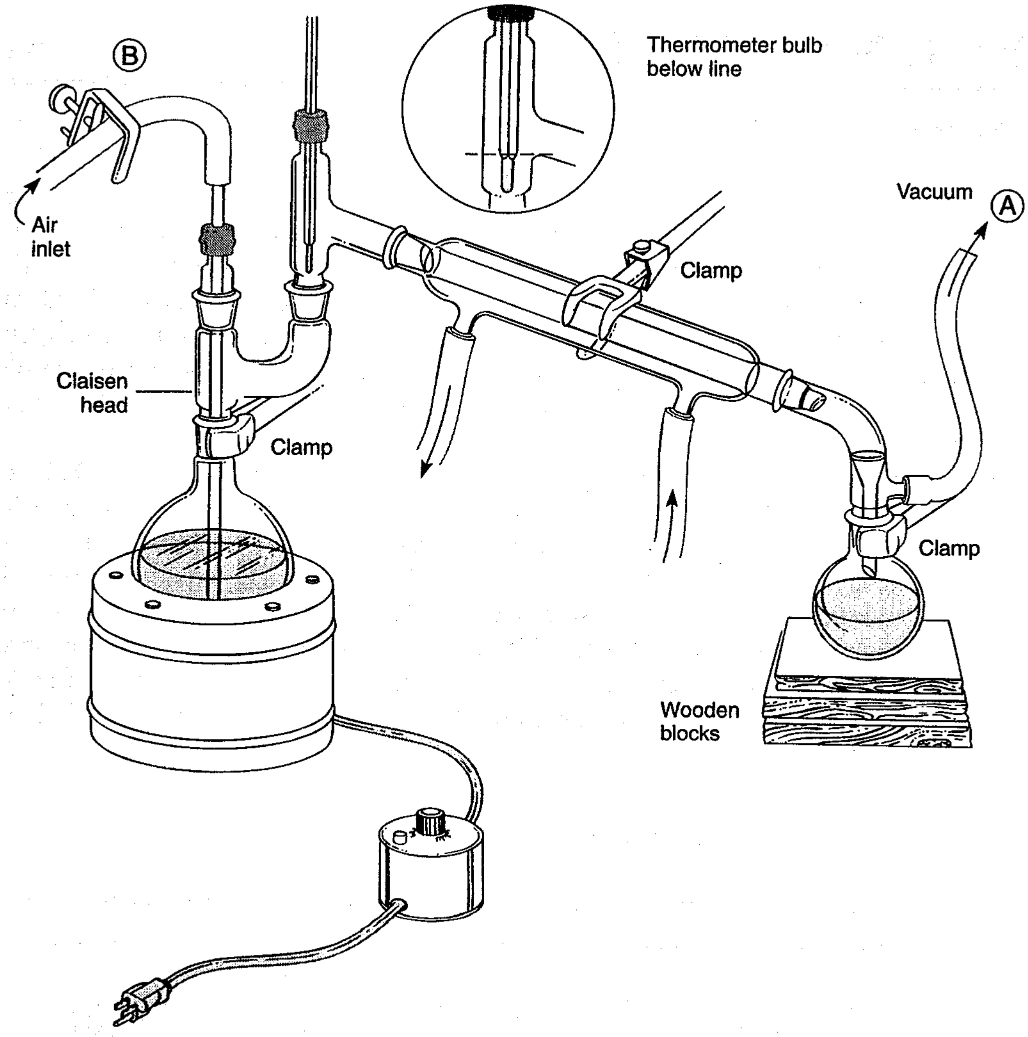
Ⓑ
Air inlet
Claisen head
Clamp
Thermometer bulb below line
Clamp
Vacuum Ⓐ
Clamp
Wooden blocks

Figure 9.1 Macroscale vacuum distillation using the standard organic laboratory kit.

aspirator is probably the simplest of these sources and the vacuum source most likely to be available. However, if pressures below 10–20 mmHg are required, a mechanical vacuum pump must be used.

Assembling the Apparatus

When assembling an apparatus for vacuum distillation, it is important to check the following points.

Glassware. Before assembly, check all glassware to be sure there are no cracks and that there are no chips in the standard-taper joints. Cracked glassware may break when evacuated. Joints that have chips may not be air-tight and they will leak.

Greasing Joints. With standard scale equipment it is necessary to grease all standard-taper joints lightly. Take care not to use too much grease. Grease can become a very serious contaminant if it oozes out the bottom of the joints into your system. Apply a small amount of grease (thin film) completely around the top of the *inner* joint; then mate the joints and press or turn them slightly to spread the grease evenly. If you have used the correct amount of grease, it will not ooze out the bottom of the joint; rather the entire joint will appear clear and without striations or uncovered areas.

Claisen Head. The Claisen head is placed between the distilling flask and the distilling head to help prevent material from "bumping over" into the condenser.

Ebulliator Tube. The air inlet tube on top of the Claisen head is called an ebulliator (*ebb-u-lay-tor*) tube. Using the screw clamp **B** on the attached heavy-wall tubing (see pressure tubing below), the ebulliator is adjusted to admit a slow continuous stream of air bubbles into the distillation flask while you are distilling. Since boiling stones will not work in a vacuum, these bubbles keep the solution stirred and help to prevent bumping. The ebulliator tube is drawn to a point at its lower end. The end of the tube should be adjusted so that it is just above the bottom of the distilling flask.

Most standard ground-glass kits contain an ebulliator tube. If one is not available, an ebulliator can be prepared easily by heating a section of glass tubing and drawing it out about 3 cm. The glass is then scored in the middle of this drawn-out section and broken, making two tubes at once. In Figure 9.1, the ebulliator is inserted into a thermometer adapter. If you do not have a second thermometer adapter, a one-hole rubber stopper may be used, placing the stopper directly into the joint on top of the Claisen head.

Wooden Applicator Sticks. An alternative to an ebulliator tube that is sometimes used is the wooden pine splint or wooden applicator stick. Air is trapped in the pores of the wood. Under vacuum, the stick will emit a slow stream of bubbles to stir the solution. The disadvantage is that, each time you open the system, you must use a new stick.

Thermometer Placement. Be sure that the thermometer is positioned so that the entire mercury bulb is below the sidearm in the distilling head (see the circular insert in Fig. 9.1). If it is placed higher, it may not be surrounded by a constant stream of vapor from the material being distilled. If the thermometer is not exposed to a continuous stream of vapor, it may not reach temperature equilibrium. As a result, the temperature reading would be incorrect (low).

Joint Clips. If plastic joint clips are available (Technique 3, Figure 3.5, p. 617), they should be used to secure the greased joints, particularly those on either side of the condenser, and the one at the bottom of the vacuum adapter where the receiving flask is attached.

Pressure Tubing. The connection to the vacuum source **A** is made using pressure tubing. Pressure tubing (also called vacuum tubing), unlike the more common thin-walled tubing used to carry water or gas, has heavy walls and will not collapse inward when it is evacuated. A comparison of the two types of tubing is shown in Figure 9.2.

Make doubly sure that any connections to pressure tubing are tight. If a tight connection cannot be made, you may have the wrong size of tubing (either the rubber tubing or the glass tubing to which it is attached). Keep the lengths of pressure tubing relatively short. The pressure tubing should be relatively new and without cracks. If the tubing shows cracks when you stretch it or bend it, it may be old and leak air into the system. Replace any tubing that shows its age.

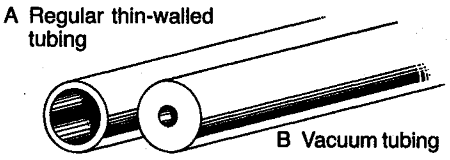

Figure 9.2 Comparison of tubing.

Rubber Stoppers. Always use soft rubber stoppers in a vacuum apparatus; corks will not give an air-tight seal. Rubber stoppers harden with age and use. If a rubber stopper is not soft (will not squeeze), discard it. Glass tubing should fit securely into any rubber stoppers. If you can move the tubing up and down with only gentle force, it is too loose and you should obtain a larger size.

Receiving Flask. When more than one fraction is expected from a vacuum distillation, it is considered good practice to have several pre-weighed receiver flasks, including the original, available before the distillation begins. Such preparation permits the rapid changing of receiving flasks during the distillation. The pre-weighing allows easy calculation of the weight of distillate in each fraction without the need to transfer the distillate to yet another flask.

To change receiving flasks, heating must be stopped, and the system vented at both ends, before replacing a flask. Complete directions for this procedure are given in the next section.

Vacuum Traps. When performing a vacuum distillation, it is customary to place a "trap" in the line that connects to the vacuum source. Two common trap arrangements are shown in Figures 9.3 and 9.4. This type of trap is essential if an aspirator or a house vacuum is used as the source of vacuum. A mechanical vacuum pump requires a different type of trap (see Fig. 9.8, p. 728). Variations in pressure are to be expected when using an aspiratory or a house vacuum. With an aspirator, if the pressure drops low enough, the vacuum in the system will draw water from the aspirator into the connecting line. The trap allows you to see this happening and take corrective action (i.e., prevent water from entering the distillation apparatus). The correct action for anything but a small amount of water is to "vent" the system. This can be accomplished by opening the screw clamp **C** at the top of the trap to let air into the system. This is also the way air is admitted into the system at the end of the distillation.

Caution: Note also that it is always necessary to vent the system *before* the aspirator is stopped. If you fail to vent the system, water may be drawn into it, contaminating your product. Be sure, however, that you vent *both ends* of the system. After venting the vacuum trap, you should immediately open the screw clamp on top of the ebulliator tube.

The trap, which contains a large volume, also acts as a buffer to pressure changes, evening out small variations in the line. In the house vacuum system, it prevents oil and water (often present in house lines) from entering your system.

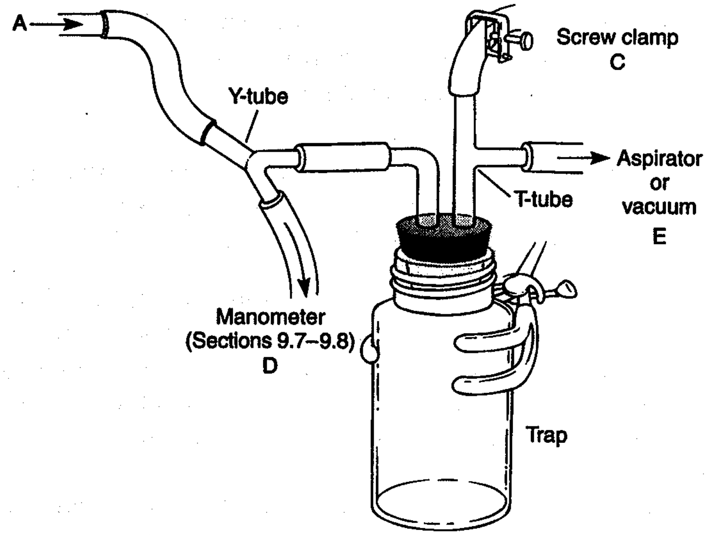
A
Y-tube
Screw clamp
C
T-tube
Aspirator
or
vacuum
E
Manometer
(Sections 9.7–9.8)
D
Trap

Figure 9.3 Vacuum trap using a gas bottle. The assembly connects to Figure 9.1 by joining the tubing at the point marked A. (The Y-tube connection to a manometer is optional.)

Manometer Connection. A manometer allows measurement of pressure. A Y-tube (or T-tube) connection **D** is shown in the line from the apparatus to the trap. This branching connection is optional, but is required if you wish to monitor the actual pressure of your system when using a manometer. The operation of manometers is discussed in

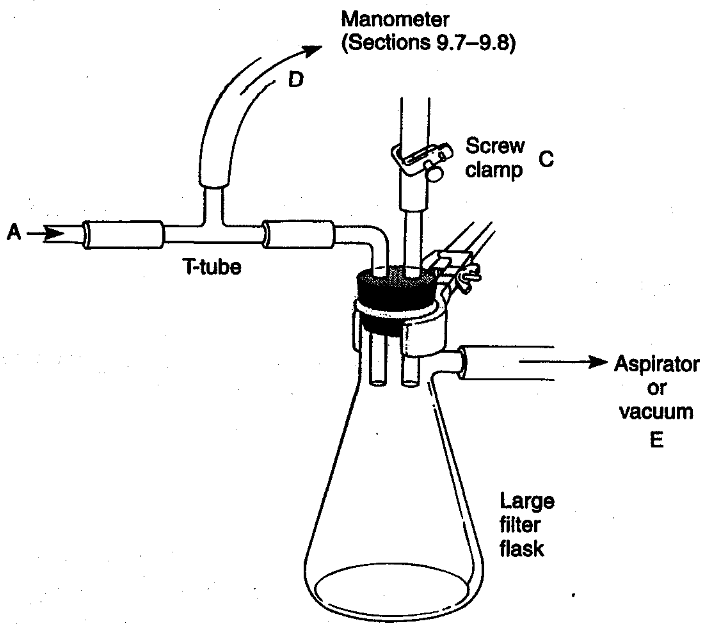
Manometer
(Sections 9.7–9.8)
D
Screw clamp
C
A
T-tube
Aspirator
or
vacuum
E
Large
filter
flask

Figure 9.4 Vacuum trap using a heavy-wall filter flask. The assembly connects to Figure 9.1 by joining the tubing at the point A. (The T-tube connection to the manometer is optional.)

Sections 9.7 and 9.8. A suitable manometer should be included in the system at least part of the time during the distillation to measure the pressure at which the distillation is being conducted. A boiling point is of little value if the pressure is not known! After use, the manometer can be removed if a screw clamp is used to close the connection.

Caution: The manometer must be vented very slowly to prevent a rush of mercury from breaking out the end of the tubing.

A manometer is also very useful in troubleshooting your system. It can be attached to the aspirator or house vacuum to determine the working pressure. In this way a defective aspirator (not uncommon) can be spotted and replaced. When you connect your apparatus, you can adjust all the joints and connections to obtain the best working pressure *before* you begin to distill. Generally a working pressure of from 25–50 mmHg is adequate for the procedures in this text.

Aspirators. In many labs the most convenient source of vacuum for a reduced-pressure distillation is the aspirator. The aspirator, or other vacuum source, is attached to the trap. The aspirator can theoretically pull a vacuum equal to the vapor pressure of the water flowing through it. The vapor pressure of flowing water depends on its temperature (24 mm Hg at 25°C; 18 mm Hg at 20°C; 9 mm Hg at 10°C). However, in the typical laboratory, the pressures attained are higher than expected due to reduced water pressure when many students are using their aspirators simultaneously. Good laboratory practice requires that only a few students on a given bench use the aspirator at the same time. It may be necessary to establish a schedule for aspirator use, or at least to have some students wait until others are finished.

House Vacuum. As stated for aspirators, depending on the capacity of the system, it may not be possible for everyone to use the vacuum system at once. Students may have to take turns or work in rotation. A typical house vacuum system will have a base pressure of about 35–100 mmHg when it is not overloaded.

9.2 VACUUM DISTILLATION: STEPWISE DIRECTIONS

The procedures in applying vacuum distillation are described in this section.

Caution: Safety glasses must be worn at all times during vacuum distillation.

Evacuating the Apparatus

1. Assemble the apparatus shown in Figure 9.1 as discussed in Section 9.1, and attach a trap (either Fig. 9.3 or 9.4). The connection is made at the points labeled **A**. Next, attach the trap to either an aspirator or a house vacuum system at point **E**. Do not close any clamps at this time.

2. Weigh each empty receiving flask to be used in collecting the various fractions during the distillation.
3. Concentrate the material to be distilled in an Erlenmeyer flask or beaker by removing all volatile solvents, such as ether, using a steam bath or a water bath in the hood. Use boiling stones and a stream of air to help the solvent removal.
4. Remove the distilling flask from the vacuum distillation apparatus, remove the grease by wiping with a towel, and transfer the concentrate to the flask, using a funnel. Complete the transfer by rinsing with a *small* amount of solvent. Again, concentrate the material until no additional volatile solvent can be removed (boiling will cease). The flask should be no more than half-full after concentration. Regrease the joint and reattach the flask to the distilling apparatus. Make sure all joints are tight.
5. On the trap assembly (Figure 9.3 or 9.4) open the clamp at **C** and attach a manometer at point **D**.
6. Turn on the aspirator (or house vacuum) (Fig. 9.3) to the maximum extent.
7. Tighten the screw clamp at **B** (Fig. 9.1) until the tubing is nearly closed.
8. Going back to the trap (Fig. 9.3), slowly tighten the screw clamp at point **C**. Watch the bubbling action of the ebulliator tube to see that it is not too vigorous or too slow. Any volatile solvents you could not remove during concentration will be removed now. Once the loss of volatiles slows down, close screw clamp **C** to the fullest extent.
9. Adjust the ebulliator tube at **B** until a fine steady stream of bubbles is formed.
10. Wait a few minutes and then record the pressure obtained.
11. If the pressure is not satisfactory, check all connections to see that they are tight. Gently twist any hoses to snug them down. Press down on any rubber stoppers. Check the fit of all glass tubing. Press any joints together until they appear evenly greased and well joined. If you crimp the rubber tubing between the apparatus and the trap with your hand and the pressure decreases, you will know that there is a leak in the glassware assembly. If there is no change, the problem may be with the aspirator or the trap. Readjust the ebulliator screw clamp at **B** if necessary.

Do not proceed until you have a good vacuum. Ask your instructor for help if necessary.

12. Once your vacuum has been established, record the pressure, and the manometer may be removed for use by another student if necessary. Place a screw clamp ahead of the manometer at **D** and tighten it. With careful venting, the manometer may now be removed.

Beginning Distillation

13. Raise the heat source into position with wooden blocks, or other means, and begin to heat.
14. Increase the temperature. Eventually a reflux ring will contact the thermometer bulb, and distillation will begin.

15. Record the temperature range and the pressure range (if the manometer is still connected) during the distillation. The distillate should be collected at a rate of about 1 drop per second.
16. If the reflux ring is in the Claisen head but will not rise into the distilling head, it may be necessary to insulate these pieces by wrapping them with glass wool and aluminum foil (shiny side in). The insulation should aid the distillate to pass into the condenser.
17. The boiling point should be relatively constant so long as the pressure is constant. A rapid increase in pressure may be due to increased use of the aspirators in the lab (or additional connections to the house vacuum). It could also be due to decomposition of the material being distilled. Decomposition will produce a dense white fog in the distilling flask. If this happens, reduce the temperature of the heat source, or remove it, and *stand back* until the system cools. When the fog subsides you can investigate the cause.

Changing Receiving Flasks

18. To change receiving flasks during distillation when a new component begins to distill (higher boiling point at the same pressure), carefully open the clamp on top of the trap assembly at **C** and immediately lower the heat source.

Watch the ebulliator for excessive backup! It may also be necessary to open the clamp at B.

19. Remove the wooden blocks, or other support under the receiving flask, release the clamp, and replace the flask with a clean, pre-weighed receiver. Use a small amount of grease, if necessary, to reestablish a good seal.
20. Reclose the clamp at **C** and allow several minutes for the system to reestablish the reduced pressure. If you opened the ebulliator screw clamp at **B** you will have to close and readjust it. Bubbling will not recommence until any liquid is drawn back out of the ebulliator. This liquid may have been forced into the ebulliator when the vacuum was interrupted.
21. Raise the heating source back into position under the distilling flask and continue with the distillation.
22. When the temperature falls at the thermometer, this usually indicates that distillation is complete. If a significant amount of liquid remains, however, the bubbling may have stopped, the pressure may have risen, the heating source may not be hot enough, or perhaps insulation of the distillation head is required. Adjust accordingly.

Shutdown

23. At the end of the distillation, remove the heat source, and slowly open the screw clamps at **C** and **B**. When the system is vented you may shut off the aspirator or house vacuum and disconnect the tubing.

24. Remove the receiving flask and clean all glassware as soon as possible after disassembly (let it cool a bit) to keep the ground-glass joints from sticking.

If you used grease, thoroughly clean all grease off the joints, or it will contaminate your samples in other procedures.

9.3 ROTARY FRACTION COLLECTORS

With the types of apparatus we have discussed previously, the vacuum must be stopped to remove fractions when a new substance (fraction) begins to distill. Quite a few steps are required to perform this change, and it is quite inconvenient when there are several fractions to be collected. Two pieces of semi-microscale apparatus that are designed to alleviate the difficulty of collecting fractions while working under vacuum are shown in Figure 9.5. The collector, which is shown to the right, is sometimes called a "cow" because of its appearance. With these rotary fraction collecting devices, all you need to do is rotate the device to collect fractions.

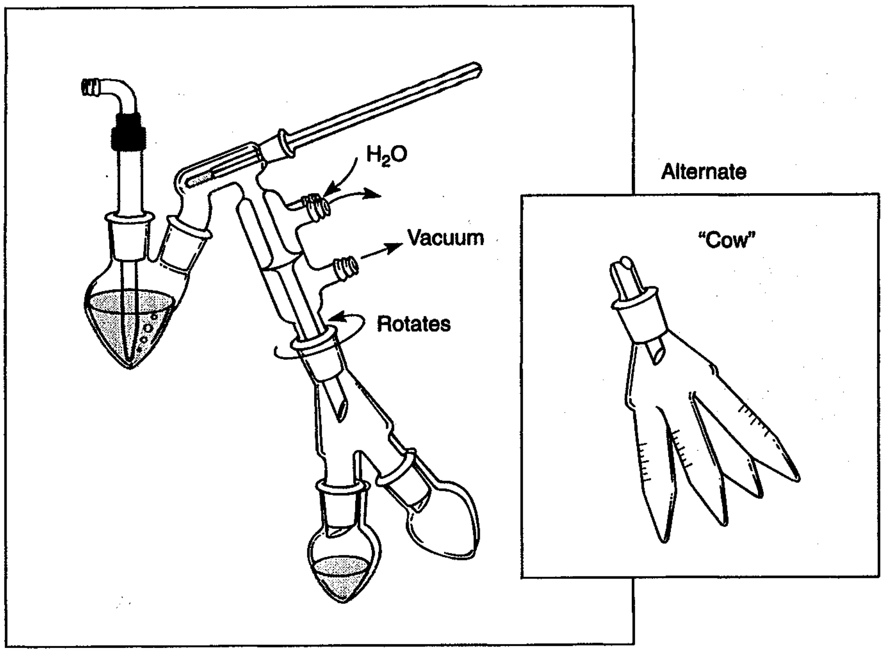

Figure 9.5 Rotary fraction collector.

9.4 MICROSCALE METHODS—STUDENT APPARATUS

Figure 9.6 shows the type of vacuum distillation equipment that would be used by a student enrolled in a microscale laboratory program. This apparatus, which uses a 5-mL conical vial as a distilling flask, can distill from 1 to 3 mL of liquid. The Hickman head replaces the Claisen head, distilling head, condenser, and receiving flask with a single piece of glassware.

9.5 BULB-TO-BULB DISTILLATION

The ultimate in microscale methods is to use a bulb-to-bulb distillation apparatus. This apparatus is shown in Figure 9.7. The sample to be distilled is placed in the

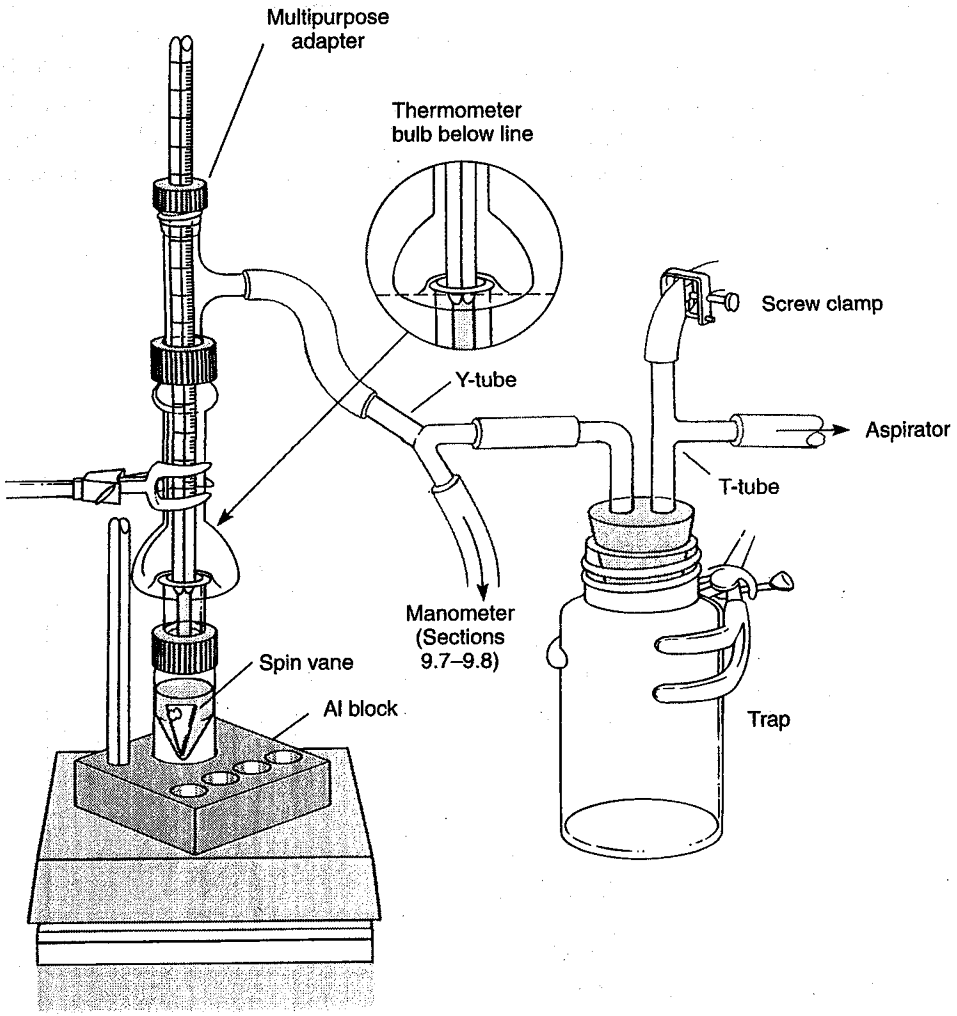
Multipurpose adapter
Thermometer bulb below line
Screw clamp
Y-tube
Aspirator
T-tube
Manometer (Sections 9.7–9.8)
Spin vane
Al block
Trap

Figure 9.6 Reduced-pressure microscale distillation.

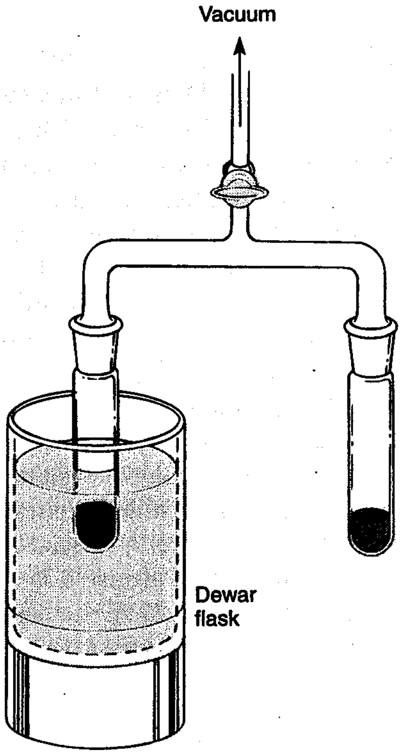

Figure 9.7 Bulb-to-bulb distillation.

glass container attached to one of the arms of the apparatus. The sample is frozen solid, usually by using liquid nitrogen, but dry ice in 2-propanol or an ice–salt-water mixture may also be used. The coolant container shown in the figure is a **Dewar flask.** The Dewar flask has a double wall with the space between the walls evacuated and sealed. A vacuum is a very good thermal insulator, and there is little heat loss from the cooling solution.

After freezing the sample, the entire apparatus is evacuated by opening the stopcock. When the evacuation is complete, the stopcock is closed, and the Dewar flask is removed. The sample is allowed to thaw and then it is frozen again. This freeze–thaw–freeze cycle removes any air or gases that were trapped in the frozen sample. Next, the stopcock is opened to evacuate the system again. When the second evacuation is complete, the stopcock is closed, and the Dewar flask is moved to the other arm to cool the empty container. As the sample warms, it will vaporize, travel to the other side, and be frozen or liquefied by the cooling solution. This transfer of the liquid from one arm to the other may take quite a while, but *no heating is required.*

The bulb-to-bulb distillation is most effective when liquid nitrogen is used as coolant and when the vacuum system can achieve a pressure of 10^{-3} mmHg or lower. This requires a vacuum pump; an aspirator cannot be used.

9.6 THE MECHANICAL VACUUM PUMP

The aspirator is not capable of yielding pressures below about 5 mmHg. This is the vapor pressure of water at 0°C, and water freezes at this temperature. A more realistic value of pressure for an aspirator is about 20 mmHg. When pressures below 20 mmHg are required, a vacuum pump will have to be employed. Figure 9.8 illustrates a mechanical vacuum pump and its associated glassware. The vacuum pump operates on a principle similar to that of the aspirator, but the vacuum pump uses a high-boiling oil, rather than water, to remove air from the attached system. The oil used in a vacuum pump, a silicone oil or a high molecular weight hydrocarbon-based oil, has a very low vapor pressure, and very low system pressures can be achieved. A good vacuum pump, with new oil, can achieve pressures of 10^{-3} or 10^{-4} mmHg. Instead of discarding the oil as it is used, it is recycled continuously through the system.

A cooled trap is required when using a vacuum pump. This trap protects the oil in the pump from any vapors that may be present in the system. If vapors from organic solvents, or from the organic compounds being distilled, dissolve in the oil, the oil's vapor pressure will increase, rendering it less effective. A special type of vacuum trap is illustrated in Figure 9.8. It is designed to fit into an insulated Dewar flask so that the coolant will last for a long period. At a minimum, this flask should be filled with ice water, but a dry ice–acetone mixture or liquid nitrogen is required to achieve lower temperatures and

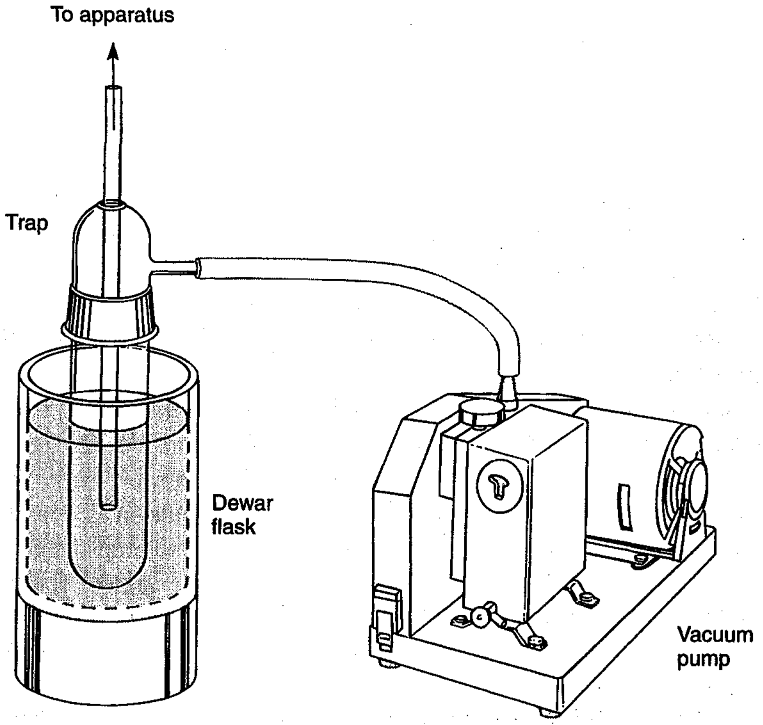

Figure 9.8 A vacuum pump and its trap.

better protect the oil. Often two traps are used; the first trap contains ice water and the second trap dry ice–acetone or liquid nitrogen. The first trap liquefies low-boiling vapors that might freeze or solidify in the second trap and block it.

9.7 THE CLOSED-END MANOMETER

The principal device used to measure pressures in a vacuum distillation is the **closed-end manometer.** Two basic types are shown in Figures 9.9 and 9.10. The manometer shown in Figure 9.9 is widely used because it is relatively easy to construct. It consists of a U-tube that is closed at one end and mounted on a wooden support. You can construct the manometer from 9-mm glass capillary tubing and fill it, as shown in Figure 9.11.

Caution: Mercury is a very toxic metal with cumulative effects. Because mercury has a high vapor pressure, it must not be spilled in the laboratory. You must not touch it with your skin. Seek immediate help from an instructor in case of a spill or if you break a manometer. Spills must be cleaned immediately.

A small filling device is connected to the U-tube with pressure tubing. The U-tube is evacuated with a good vacuum pump; then the mercury is introduced by tilting the mer-

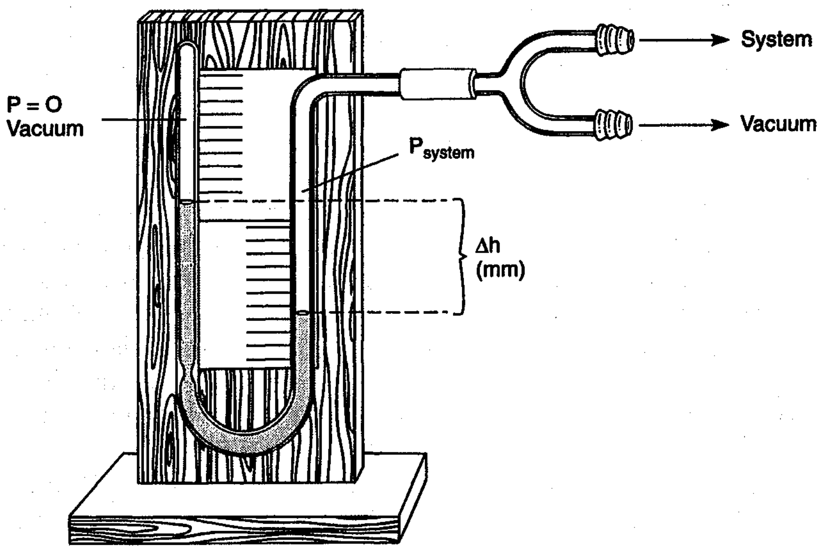

Figure 9.9 A simple U-tube manometer.

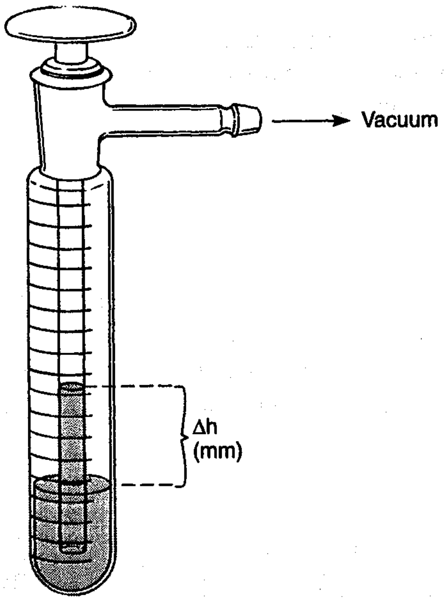

Figure 9.10 Commercial "stick" manometer.

cury reservoir. The entire filling operation should be conducted in a shallow pan in order to contain any spills that might occur. Enough mercury should be added to form a column about 20 cm in total length. When the vacuum is interrupted by admitting air, the mercury is forced by atmospheric pressure to the end of the evacuated tube. The manometer is then ready for use. The constriction shown in Figure 9.11 helps to protect the manometer against breakage when the pressure is released. Be sure that the column of mercury is long enough to pass through this constriction.

When an aspirator or any other vacuum source is used, a manometer can be connected into the system. As the pressure is lowered, the mercury rises in the right tube and drops in the left tube until Δh corresponds to the approximate pressure of the system (see Fig. 9.9).

$$\Delta h = (P_{\text{system}} - P_{\text{reference arm}}) = (P_{\text{system}} - 10^{-3}\ \text{mmHg}) \approx P_{\text{system}}$$

A short piece of metric ruler or a piece of graph paper ruled in millimeter squares is mounted on the support board to allow Δh to be read. No addition or subtraction is necessary, because the reference pressure (created by the initial evacuation when filling) is approximately zero (10^{-3} mmHg) when referred to readings in the 10–50 mmHg range. To determine the pressure, count the number of millimeter squares beginning at the top of the mercury column on the left and continuing downward to the top of the mercury column on the right. This is the height difference Δh, and it gives the pressure in the system directly.

A commercial counterpart to the U-tube manometer is shown in Figure 9.10. With this manometer, the pressure is given by the difference in the mercury levels in the inner and outer tubes.

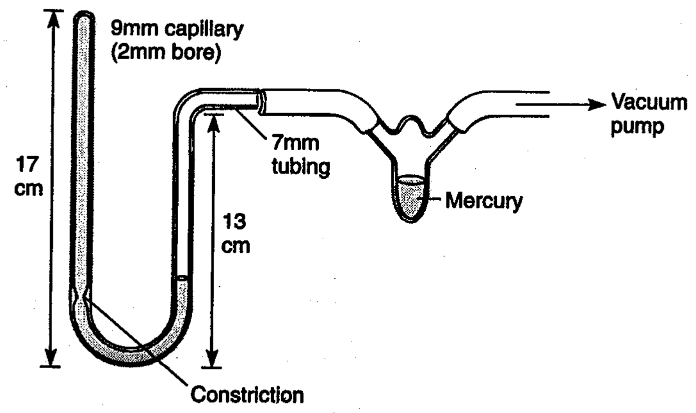

Figure 9.11 Filling a U-tube manometer.

The manometers described here have a range of about 1–150 mmHg in pressure. They are convenient to use when an aspirator is the source of vacuum. For high-vacuum systems (pressures below 1 mmHg), a more elaborate manometer or an electronic measuring device must be used. These devices will not be discussed here.

9.8 CONNECTING AND USING A MANOMETER

The most common use of a closed-end manometer is to monitor pressure during a reduced-pressure distillation. The manometer is placed in a vacuum distillation system, as shown in Figure 9.12. Generally, an aspirator is the source of vacuum. Both the manometer and the distillation apparatus should be protected by a trap from possible backups in the water line. Alternatives to the trap arrangements shown in Figure 9.12 appear in Figures 9.3 and 9.4. Notice in each case that the trap has a device (screw clamp or stopcock) for opening the system to the atmosphere. This is especially important in using a manometer, because you should always make pressure changes slowly. If this is not done, there is a danger of spraying mercury throughout the system, breaking the manometer, or spurting mercury into the room. In the closed-end manometer, if the system is opened suddenly, the mercury rushes to the closed end of the U-tube. The mercury rushes with such speed and force that the end will be broken out of the manometer. Air should be admitted *slowly* by opening the valve cautiously. In a similar fashion, the valve should be closed slowly when the vacuum is being started, or mercury may be forcefully drawn into the system through the open end of the manometer.

If the pressure in a reduced-pressure distillation is lower than desired, it is possible to adjust it by means of a **bleed valve.** The stopcock can serve this function in Figure 9.12 if it is opened only a small amount. In those systems with a screw clamp on the trap (Figs. 9.3 and 9.4), remove the screw clamp from the trap valve and attach the base of a Tirrill-style Bunsen burner. The needle valve in the base of the burner can be used to adjust precisely the amount of air that is admitted (bled) to the system and, hence, control the pressure.

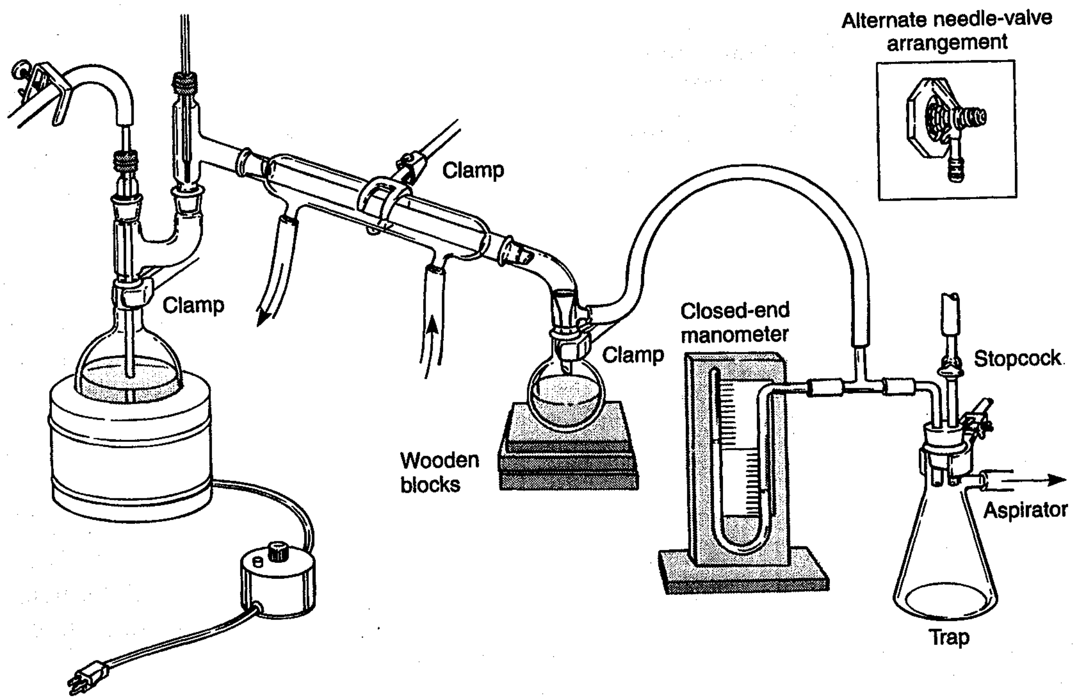

Figure 9.12 Connecting a manometer to the system. To construct a "bleed," the needle valve may replace the stopcock.

PROBLEMS

1. Give some reasons that would lead you to purify a liquid by using vacuum distillation rather than by using simple distillation.
2. When using an aspirator as a source of vacuum in a vacuum distillation, do you turn off the aspirator before venting the system? Explain.
3. A compound was distilled at atmospheric pressure and had a boiling range of 310–325°C. What would be the approximate boiling range of this liquid if it was distilled under vacuum at 20 mmHg?
4. Boiling stones generally do not work when performing a vacuum distillation. What substitutes may be used?
5. What is the purpose of the trap that is used during a vacuum distillation performed with an aspirator?

TECHNIQUE 10

Fractional Distillation, Azeotropes

Simple distillation, described in Technique 8, works well for most routine separation and purification procedures for organic compounds. When boiling-point differences of components to be separated are not large, however, **fractional distillation** must be used to achieve a good separation.

Part A. Fractional Distillation

10.1 DIFFERENCES BETWEEN SIMPLE AND FRACTIONAL DISTILLATION

When an ideal solution of two liquids, such as benzene (bp 80°C) and toluene (bp 110°C), is distilled by simple distillation, the first vapor produced will be enriched in the lower-boiling component (benzene). However, when that initial vapor is condensed and analyzed, the distillate will not be pure benzene. The boiling point difference of benzene and toluene (30°C) is too small to achieve a complete separation by simple distillation. Following the principles outlined in Technique 8, Section 8.2 (pp. 708–710), and using the vapor-liquid composition curve given in Figure 10.1, you can see what would happen if you started with a equimolar mixture of benzene and toluene.

Following the dashed lines shows that an equimolar mixture (50 mole percent benzene) would begin to boil at about 91°C and, far from being 100% benzene, the distillate would contain about 74 mole percent benzene and 26 mole percent toluene. As the distillation continued, the composition of the undistilled liquid would move in the direction of A′ (there would be increased toluene, due to removal of more benzene than toluene), and the corresponding vapor would contain a progressively smaller amount of benzene. In effect, the temperature of the distillation would continue to increase throughout the distillation (as in Figure 8.2B, p. 708), and it would be impossible to obtain any fraction that consisted of pure benzene.

Suppose, however, that we are able to collect a small quantity of the first distillate that was 74 mole percent benzene and redistill it. Using Figure 10.1, we can see that this liquid would begin to boil at about 84°C and would give an initial distillate containing 90

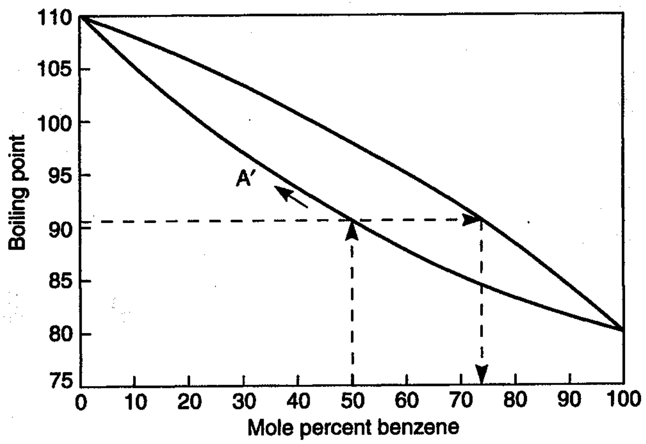

Figure 10.1 The vapor-liquid composition curve for mixtures of benzene and toluene.

mole percent of benzene. If we were experimentally able to continue taking small fractions at the beginning of each distillation, and redistill them, we would eventually reach a liquid with a composition of nearly 100 mole per cent benzene. However, since we took only a small amount of material at the beginning of each distillation, we would have lost most of the material we started with. To recapture a reasonable amount of benzene, we would have to process each of the fractions left behind in the same way as our early fractions. As each of them was partially distilled, the material advanced would become progressively richer in benzene, while that left behind would become progressively richer in toluene. It would require thousands (maybe millions) of such microdistillations to separate benzene from toluene.

Obviously, the procedure just described would be very tedious; fortunately, it need not be performed in usual laboratory practice. **Fractional distillation** accomplishes the same result. You simply have to use a column inserted between the distillation flask and the distilling head, as shown in Figure 10.2. This **fractionating column** is filled, or **packed,** with a suitable material such as a stainless steel sponge. This packing allows a mixture of benzene and toluene to be subjected continuously to many vaporization–condensation cycles as the material moves up the column. With each cycle within the column, the composition of the vapor is progressively enriched in the lower-boiling component (benzene).

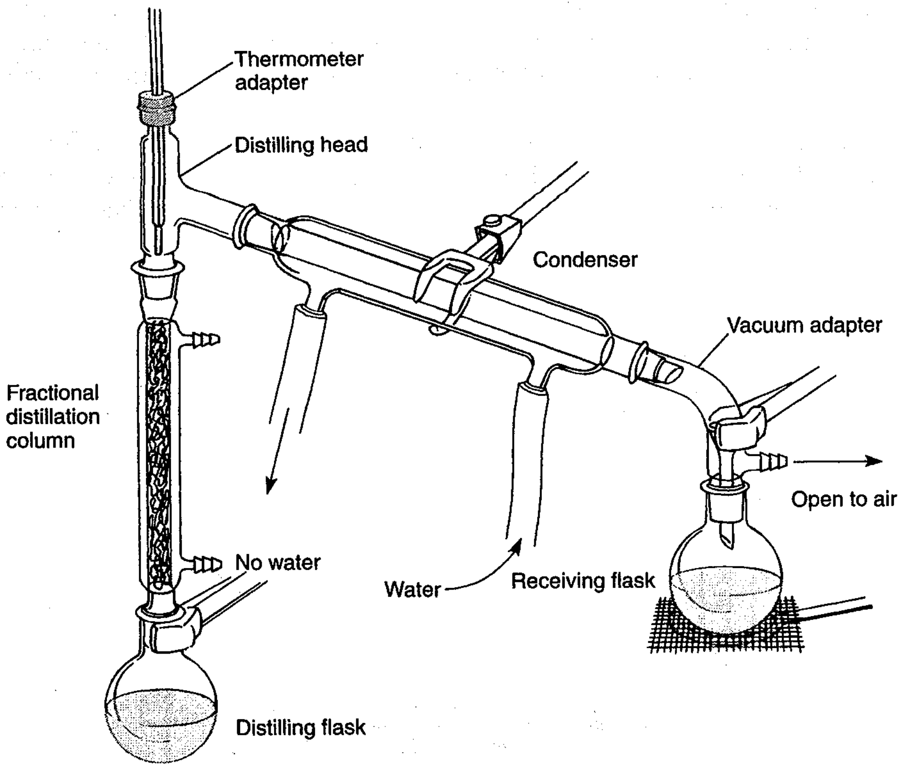

Figure 10.2 Fractional distillation apparatus.

Nearly pure benzene (bp 80°C) finally emerges from the top of the column, condenses, and passes into the receiving head or flask. This process continues until all the benzene is removed. The distillation must be carried out slowly to ensure that numerous vaporization–condensation cycles occur. When nearly all the benzene has been removed, the temperature begins to rise and a small amount of a second fraction, which contains some benzene and toluene, may be collected. When the temperature reaches 110°C, the boiling point of pure toluene, the vapor is condensed and collected as the third fraction. A plot of boiling-point versus volume of condensate (distillate) would resemble Figure 8.2C (p. 708). This separation would be much better than that achieved by simple distillation (Figure 8.2B (p. 708).

10.2 VAPOR–LIQUID COMPOSITION DIAGRAMS

A vapor–liquid composition phase diagram like the one in Figure 10.3 can be used to explain the operation of a fractionating column with an **ideal solution** of two liquids, A and B. An ideal solution is one in which the two liquids are chemically similar, miscible (mutually soluble) in all proportions, and do not interact. Ideal solutions obey **Raoult's Law.** Raoult's Law is explained in detail in Section 10.3.

The phase diagram relates the compositions of the boiling liquid (lower curve) and its vapor (upper curve) as a function of temperature. Any horizontal line drawn across the diagram (a constant-temperature line) intersects the diagram in two places. These intersections relate the vapor composition to the composition of the boiling liquid that pro-

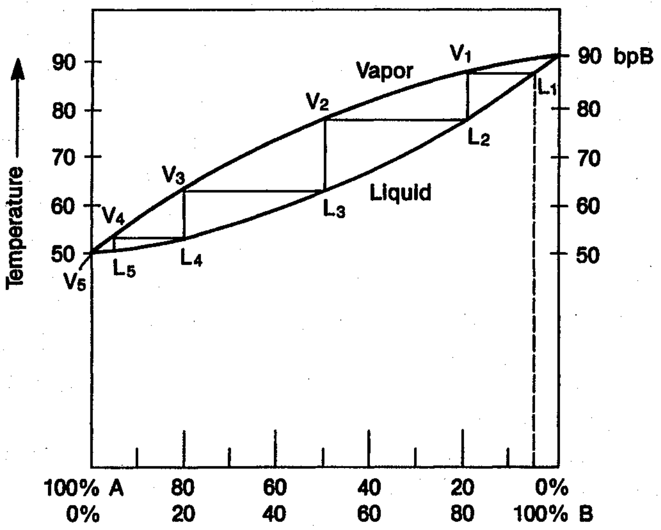

Figure 10.3 Phase diagram for a fractional distillation of an ideal two-component system.

duces that vapor. By convention, composition is expressed either in **mole fraction** or in **mole percentage.** The mole fraction is defined as follows:

$$\text{Mole fraction A} = N_A = \frac{\text{Moles A}}{\text{Moles A} + \text{Moles B}}$$

$$\text{Mole fraction B} = N_B = \frac{\text{Moles B}}{\text{Moles A} + \text{Moles B}}$$

$$N_A + N_B = 1$$

$$\text{Mole percentage A} = N_A \times 100$$

$$\text{Mole percentage B} = N_B \times 100$$

The horizontal and vertical lines shown in Figure 10.3 represents the processes that occur during a fractional distillation. Each of the **horizontal lines** (L_1V_1, L_2V_2, etc.) represents the **vaporization** step of a given vaporization–condensation cycle and represents the composition of the vapor in equilibrium with liquid at a given temperature. For example, at 63°C a liquid with a composition of 50% A (L_3 on the diagram) would yield vapor of composition 80% A (V_3 on diagram) at equilibrium. The vapor is richer in the lower-boiling component A than the original liquid was.

Each of the **vertical lines** (V_1L_2, V_2L_3, etc.) represents the **condensation** step of a given vaporization–condensation cycle. The composition does not change as the temperature drops on condensation. The vapor at V_3, for example, condenses to give a liquid (L_4 on the diagram) of composition 80% A with a drop in temperature from 63 to 53°C.

In the example shown in Figure 10.3, pure A boils at 50°C and pure B boils at 90°C. These two boiling points are represented at the left- and right-hand edges of the diagram, respectively. Now consider a solution that contains only 5% of A but 95% of B. (Remember that these are *mole* percentages.) This solution is heated (following the dashed line) until it is observed to boil at L_1 (87°C). The resulting vapor has composition V_1 (20% A, 80% B). The vapor is richer in A than the original liquid, but it is by no means pure A. In a simple distillation apparatus, this vapor would be condensed and passed into the receiver in a very impure state. However, with a fractionating column in place, the vapor is condensed in the **column** to give liquid L_2 (20% A, 80% B). Liquid L_2 is immediately revaporized (bp 78°C) to give a vapor of composition V_2 (50% A, 50% B), which is condensed to give liquid L_3. Liquid L_3 is revaporized (bp 63°C) to give vapor of composition V_3 (80% A, 20% B), which is condensed to give liquid L_4. Liquid L_4 is revaporized (bp 53°C) to give vapor of composition V_4 (95% A, 5% B). This process continues to V_5, which condenses to give nearly pure liquid A. The fractionating process follows the stepped lines in the figure downward and to the left.

As this process continues, all of liquid A is removed from the distillation flask or vial, leaving nearly pure B behind. If the temperature is raised, liquid B may be distilled as a nearly pure fraction. Fractional distillation will have achieved a separation of A and B, a separation that would have been nearly impossible with simple distillation. Notice that the boiling point of the liquid becomes lower each time it vaporizes. Because the temperature at the bottom of a column is normally higher than the temperature at the top, successive vaporizations occur higher and higher in the column as the composition of the distillate approaches that of pure A. This process is illustrated in Figure 10.4, where the composition of the liquids, their boiling points, and the composition of the vapors present are shown alongside the fractionating column.

10.3 RAOULT'S LAW

Two liquids (A and B) that are miscible and that do not interact form an **ideal solution** and follow Raoult's Law. The law states that the partial vapor pressure of component A in the solution (P_A) equals the vapor pressure of pure A (P_A^0) times its mole frac-

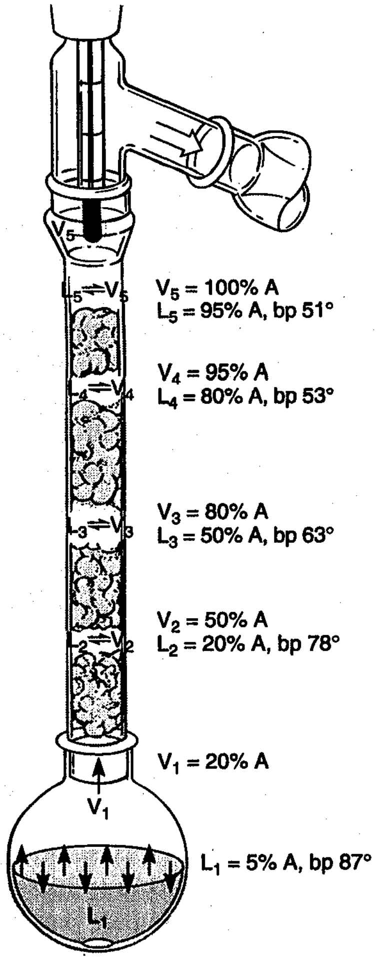

V_5
$L_5 \rightleftharpoons V_5$
V_5 = 100% A
L_5 = 95% A, bp 51°
$L_4 \rightleftharpoons V_4$
V_4 = 95% A
L_4 = 80% A, bp 53°
$L_3 \rightleftharpoons V_3$
V_3 = 80% A
L_3 = 50% A, bp 63°
$L_2 \rightleftharpoons V_2$
V_2 = 50% A
L_2 = 20% A, bp 78°
V_1 = 20% A
V_1
L_1 = 5% A, bp 87°
L_1

Figure 10.4 Vaporization-condensation in a fractionation column.

tion (N_A) (Eq.1). A similar expression can be written for component B (Eq. 2). The mole fractions N_A and N_B were defined in Section 10.2.

$$\text{Partial vapor pressure of A in solution} = P_A = (P_A^0)(N_A) \qquad (1)$$

$$\text{Partial vapor pressure of B in solution} = P_B = (P_B^0)(N_B) \qquad (2)$$

P_A^0 is the vapor pressure of pure A, independent of B. P_B^0 is the vapor pressure of B, independent of A. In a mixture of A and B, the partial vapor pressures are added to give the total vapor pressure above the solution (Eq. 3). When the total pressure (sum of the partial pressures) equals the applied pressure, the solution boils.

$$P_{total} = P_A + P_B = P_A^0 N_A + P_B^0 N_B \qquad (3)$$

The composition of A and B in the vapor produced is given by Equations 4 and 5.

$$N_A \text{ (vapor)} = \frac{P_A}{P_{total}} \qquad (4)$$

$$N_B \text{ (vapor)} = \frac{P_B}{P_{total}} \qquad (5)$$

Several exercises involving applications of Raoult's Law are illustrated in Figure 10.5. Note, particularly in the result from Equation 4, that the vapor is richer ($N_A = 0.67$) in the lower-boiling (higher vapor pressure) component A than it was before vaporization ($N_A = 0.50$). This proves mathematically what was described in Section 10.2.

The consequences of Raoult's Law for distillations are shown schematically in Figure 10.6. In Part A the boiling points are identical (vapor pressures the same), and no separation is attained regardless of how the distillation is conducted. In Part B a fractional distillation is required, while in Part C a simple distillation provides an adequate separation.

When a solid B (rather than another liquid) is dissolved in a liquid A, the boiling point is increased. In this extreme case, the vapor pressure of B is negligible, and the vapor will be pure A no matter how much solid B is added. Consider a solution of salt in water.

$$P_{total} = P^0_{water} N_{water} + P^0_{salt} N_{salt}$$

$$P^0_{salt} = 0$$

$$P_{total} = P^0_{water} N_{water}$$

A solution whose mole fraction of water is 0.7 will not boil at 100°C, because $P_{total} = (760)(0.7) = 532$ mmHg and is less than atmospheric pressure. If the solution is heated to 110°C, it will boil because $P_{total} = (1085)(0.7) = 760$ mmHg. Although the solution must be heated at 110°C to boil it, the vapor is pure water and has a boiling-point temperature of 100°C. (The vapor pressure of water at 110°C can be looked up in a handbook; it is 1085 mmHg.)

Consider a solution at 100°C where $N_A = 0.5$ and $N_B = 0.5$.

1. What is the partial vapor pressure of A in the solution if the vapor pressure of pure A at 100°C is 1020 mmHg?

   Answer: $P_A = P^{\circ}_A N_A = (1020)(0.5) = 510$ mmHg

2. What is the partial vapor pressure of B in the solution if the vapor pressure of pure B at 100°C is 500 mmHg?

   Answer: $P_B = P^{\circ}_B N_B = (500)(0.5) = 250$ mmHg

3. Would the solution boil at 100°C if the applied pressure were 760 mmHg?

   Answer: Yes. $P_{total} = P_A + P_B = (510 + 250) = 760$ mmHg

4. What is the composition of the vapor at the boiling point?

   Answer: The boiling point is 100°C.

$$N_A\text{ (vapor)} = \frac{P_A}{P_{total}} = 510/760 = 0.67$$

$$N_B\text{ (vapor)} = \frac{P_B}{P_{total}} = 250/760 = 0.33$$

Figure 10.5 Sample calculations with Raoult's Law.

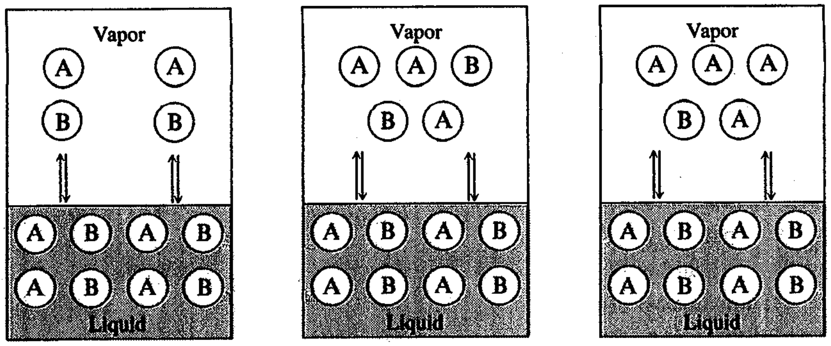

$N_A = N_B$
$P_A{}^{\circ} = P_B{}^{\circ}$
$P_A{}^{\circ}N_A = P_B{}^{\circ}N_B$
Equal amounts of A and B in vapor—no separation
A

$N_A = N_B$
$P_A{}^{\circ} > P_B{}^{\circ}$ ($bp_A < bp_B$)
$P_A{}^{\circ}N_A > P_B{}^{\circ}N_B$
More A in vapor than B—some separation
B

$N_A = N_B$
$P_A{}^{\circ} >>> P_B{}^{\circ}$
$P_A{}^{\circ}N_A >>> P_B{}^{\circ}N_B$
Much more A in vapor than B—good separation
C

Figure 10.6 Consequences of Raoult's Law. (A) Boiling points (vapor pressures) are identical—no separation. (B) Boiling points somewhat less for A than for B—requires fractional distillation. (C) Boiling points much less for A than for B—simple distillation will suffice.

10.4 COLUMN EFFICIENCY

A common measure of the efficiency of a column is given by its number of **theoretical plates.** The number of theoretical plates in a column is related to the number of vaporization–condensation cycles that occur as a liquid mixture travels through it. Using the example mixture in Figure 10.3, if the first distillate (condensed vapor) had the composition at L_2 when starting with liquid of composition L_1, the column would be said to have *one theoretical plate.* This would correspond to a simple distillation, or one vaporization–condensation cycle. A column would have two theoretical plates if the first distillate had the composition at L_3. The two-theoretical-plate column essentially carries out "two simple distillations." According to Figure 10.3, *five theoretical plates* would be required to separate the mixture that started with composition L_1. Notice that this corresponds to the number of "steps" that need to be drawn in the figure to arrive at a composition of 100% A.

Most columns do not allow distillation in discrete steps, as indicated in Figure 10.3. Instead, the process is *continuous,* allowing the vapors to be continuously in contact with liquid of changing composition as they pass through the column. Any material can be used to pack the column as long as it can be wetted by the liquid and as long as it does not pack so tightly that vapor cannot pass.

The approximate relationship between the number of theoretical plates needed to separate an ideal two-component mixture and the difference in boiling points is given in Table 10.1. Notice that more theoretical plates are required as the boiling-point differences between the components decrease. For instance, a mixture of A (bp 130°C) and B (bp 166°C) with a boiling-point difference of 36°C would be expected to require a column with a minimum of five theoretical plates.

TABLE 10.1 Theoretical Plates Required to Separate Mixtures, Based on Boiling-Point Differences of Components

Boiling-Point Difference	Number of Theoretical Plates
108	1
72	2
54	3
43	4
36	5
20	10
10	20
7	30
4	50
2	100

10.5 TYPES OF FRACTIONATING COLUMNS AND PACKINGS

Several types of fractionating columns are shown in Figure 10.7. The Vigreux column, shown in Part A, has indentations that incline downward at angles of 45° and are in pairs on opposites sides of the column. The projections into the column provide increased possibilities for condensation and for the vapor to equilibrate with the liquid. Vigreux columns are popular in cases where only a small number of theoretical plates are required. They are not very efficient (a 20-cm column might have only 2.5 theoretical plates), but they allow for rapid distillation and have a small **holdup** (the amount of liquid retained by the column). A column packed with a stainless steel sponge is a more effective fractionating column than a Vigreux column, but not by a large margin. Glass beads, or glass helices, can also be used as a packing material, and they have a slightly greater efficiency yet. The air condenser or the water condenser can be used as an improvised column if an actual fractionating column is unavailable. If a condenser is packed with glass beads, glass helices, or sections of glass tubing, the packing must be held in place by inserting a small plug of stainless steel sponge into the bottom of the condenser.

The most effective type of column is the **spinning-band column.** In the most elegant form of this device, a tightly fitting, twisted platinum screen or a Teflon rod with helical threads is rotated rapidly inside the bore of the column (Fig. 10.8). A spinning-band column that is available for microscale work is shown in Figure 10.9. This spinning-band column has a band about 2–3 cm in length and provides 4–5 theoretical

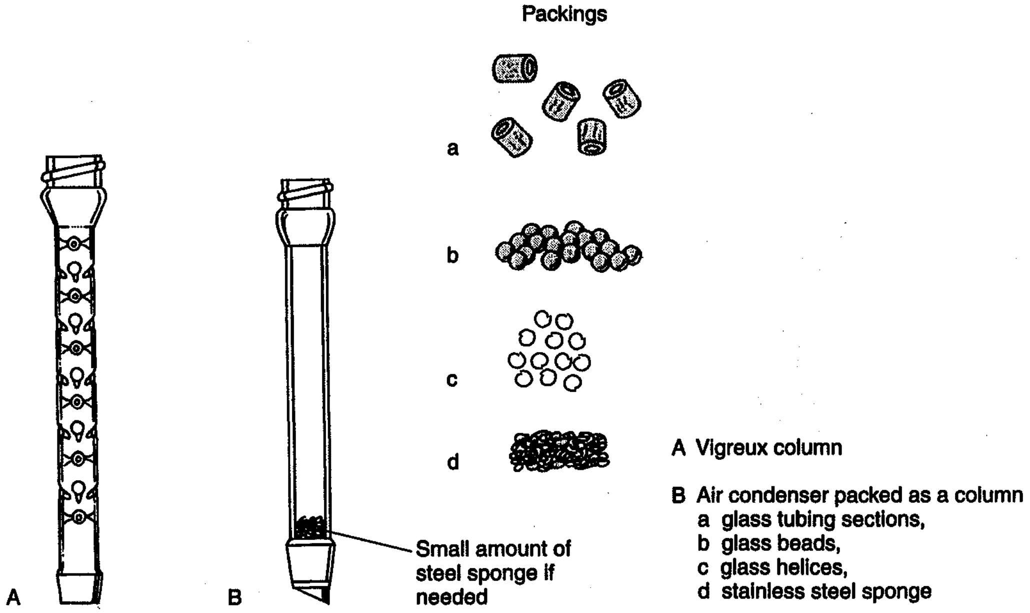

Figure 10.7 Columns for fractional distillation.

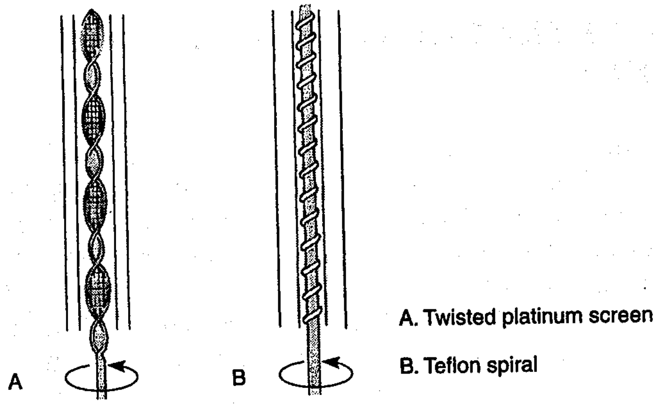
A
B
A. Twisted platinum screen
B. Teflon spiral

Figure 10.8 Bands for spinning-band columns.

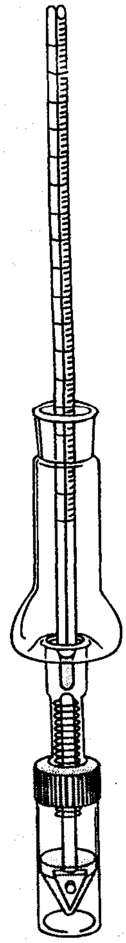

Figure 10.9 A commercially available microscale spinning-band column.

plates. It can separate 1–2 mL of a mixture with a 30°C boiling-point difference. Larger research models of this spinning-band column can provide as many as 20 or 30 theoretical plates and can separate mixtures with a boiling-point difference of as little as 5–10°C.

Manufacturers of fractionating columns often offer them in a variety of lengths. Because the efficiency of a column is a function of its length, longer columns have more theoretical plates than shorter ones. It is common to express efficiency of a column in a unit called **HETP,** the **H**eight of a column that is **E**quivalent to one **T**heoretical **P**late. HETP is usually expressed in units of cm/plate. When the height of the column (in centimeters) is divided by this value, the total number of theoretical plates is specified.

10.6 FRACTIONAL DISTILLATION: METHODS AND PRACTICE

In performing a fractional distillation, the column should be clamped in a vertical position. The distillation should be conducted as slowly as possible, but the rate of distillation should be steady enough to produce a constant temperature reading at the thermometer.

Many fractionating columns must be insulated so that temperature equilibrium is maintained at all times. Additional insulation will not be required for columns that have an evacuated outer jacket, but those that do not can benefit from being wrapped in insulation.

Glass wool and aluminum foil (shiny side in) are often used for insulation. You can wrap the column with glass wool and then use a wrapping of the aluminum foil to keep it in place. An especially effective method is to make an insulation blanket by placing a layer of glass wool or cotton between two rectangles of aluminum foil, placed shiny side in. The sandwich is bound together with duct tape. This blanket, which is reusable, can be wrapped around the column and held in place with twist ties or tape.

The **reflux ratio** is defined as the ratio of the number of drops of distillate that return to the distillation flask compared to the number of drops of distillate collected. In an efficient column, the reflux ratio should equal or exceed the number of theoretical plates. A high reflux ratio ensures that the column will achieve temperature equilibrium and achieve its maximum efficiency. This ratio is not easy to determine; in fact, it is impossible to determine when using a Hickman head, and it should not concern a beginning student. In some cases, the **throughput,** or **rate of takeoff,** of a column may be specified. This is expressed as the number of milliliters of distillate that can be collected per unit of time, usually as mL/min.

Standard Scale Apparatus. Figure 10.2 illustrates a fractional distillation assembly that can be used for larger-scale distillations. It has a glass-jacketed column that is packed with a stainless steel sponge. This apparatus would be common in situations where quantities of liquid in excess of 10 mL were to be distilled.

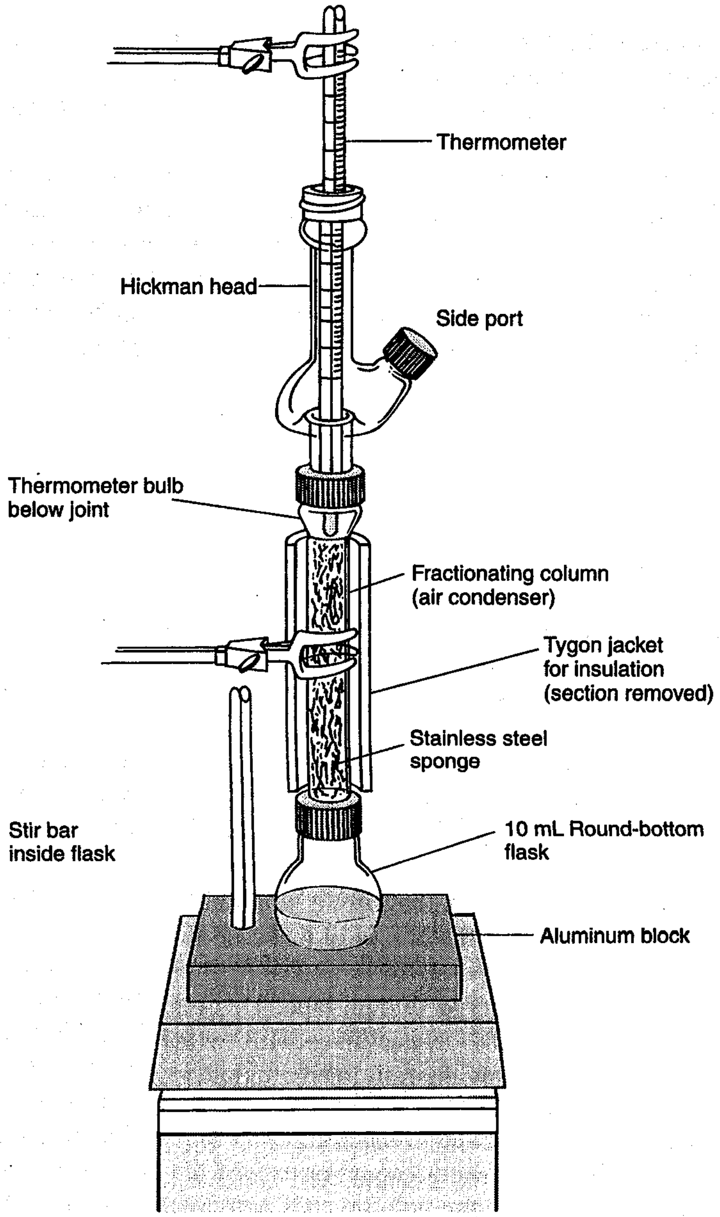
Thermometer
Hickman head
Side port
Thermometer bulb below joint
Fractionating column (air condenser)
Tygon jacket for insulation (section removed)
Stainless steel sponge
Stir bar inside flask
10 mL Round-bottom flask
Aluminum block

Figure 10.10 Microscale apparatus for fractional distillation.

Microscale Apparatus. The apparatus shown in Figure 10.10 is the one you are most likely to use in the microscale laboratory. If your laboratory is one of the better equipped ones, you may have access to spinning-band columns like those shown in Figure 10.9.

Part B. Azeotropes

10.7 NONIDEAL SOLUTIONS: AZEOTROPES

Some mixtures of liquids, because of attractions or repulsions between the molecules, do not behave ideally; they do not follow Raoult's Law. There are two types of vapor–liquid composition diagrams that result from this nonideal behavior: **minimum-boiling-point** and **maximum-boiling-point** diagrams. The minimum or maximum points in these diagrams correspond to a constant-boiling mixture called an **azeotrope.** An azeotrope is a mixture with a fixed composition that cannot be altered by either simple or fractional distillation. An azeotrope behaves as if it were a pure compound, and it distills from the beginning to the end of its distillation at a constant temperature, giving a distillate of constant (azeotropic) composition. The vapor in equilibrium with an azeotropic liquid has the same composition as the azeotrope. Because of this, an azeotrope is represented as a *point* on a vapor–liquid composition diagram.

A. MINIMUM-BOILING-POINT DIAGRAMS

A minimum-boiling-point azeotrope results from a slight incompatibility (repulsion) between the liquids being mixed. This incompatibility leads to a higher-than-expected combined vapor pressure from the solution. This higher combined vapor pressure brings about a lower boiling point for the mixture than is observed for the pure components. The most common two-component mixture that gives a minimum-boiling-point azeotrope is the ethanol–water system shown in Figure 10.11. The azeotrope at V_3 has a composition

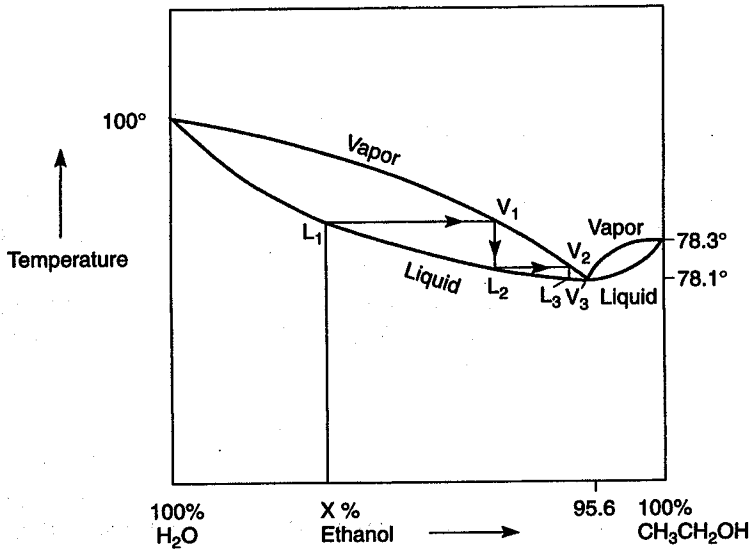

Figure 10.11 Ethanol–water minimum-boiling-point phase diagram.

of 96% ethanol–4% water and a boiling point of 78.1°C. This boiling point is not much lower than that of pure ethanol (78.3°C), but it means that it is impossible to obtain pure ethanol from the distillation of any ethanol–water mixture that contains more than 4% water. Even with the best fractionating column, you cannot obtain 100% ethanol. The remaining 4% of water can be removed by adding benzene and removing a different azeotrope, the ternary benzene–water–ethanol azeotrope (bp 65°C). Once the water is removed, the excess benzene is removed as an ethanol–benzene azeotrope (bp 68°C). The resulting material is free of water and is called "absolute" ethanol.

The fractional distillation of an ethanol–water mixture of composition X can be described as follows. The mixture is heated (follow line XL_1) until it is observed to boil at L_1. The resulting vapor at V_1 will be richer in the lower-boiling component, ethanol, than the original mixture.[1] The condensate at L_2 is vaporized to give V_2. The process continues, following the lines to the right, until the azeotrope is obtained at V_3. The liquid that distills is not pure ethanol, but it has the azeotropic composition of 96% ethanol and 4% water, and it distills at 78.1°C. The azeotrope, which is richer in ethanol than the original mixture, continues to distill. As it distills, the percentage of water left behind in the distillation flask continues to increase. When all the ethanol has been distilled (as the azeotrope), pure water remains behind in the distillation flask, and it distills at 100°C.

If the azeotrope obtained by the preceding procedure is redistilled, it distills from the beginning to the end of the distillation at a constant temperature of 78.1°C as if it were a pure substance. There is no change in the composition of the vapor during the distillation.

Some common minimum-boiling azeotropes are given in Table 10.2. Numerous other azeotropes are formed in two- and three-component systems; such azeotropes are common. Water forms azeotropes with many substances; therefore, water must be carefully removed with **drying agents** whenever possible before compounds are distilled. Extensive azeotropic data are available in references such as the *Handbook of Chemistry and Physics.*[2]

B. MAXIMUM-BOILING-POINT DIAGRAMS

A maximum-boiling point azeotrope results from a slight attraction between the component molecules. This attraction leads to lower combined vapor pressure than expected in the solution. The lower combined vapor pressures cause a higher boiling point than what would be characteristic for the components. A two-component maximum-boiling-point azeotrope is illustrated in Figure 10.12. Because the azeotrope has a higher boil-

---

[1] Keep in mind that this distillate is not pure ethanol but is an ethanol–water mixture.

[2] More examples of azeotropes, with their compositions and boiling points, can be found in the *CRC Handbook of Chemistry and Physics;* also in L.H. Horsley, ed., *Advances in Chemistry Series,* no. 116. Azeotropic Data, III (Washington: American Chemical Society, 1973).

TABLE 10.2 Common Minimum-Boiling Azeotropes

Azeotrope	Composition (Weight percentage)	Boiling Point (°C)
Ethanol–water	95.6% C_2H_5OH, 4.4% H_2O	78.17
Benzene–water	91.1% C_6H_6, 8.9% H_2O	69.4
Benzene–water–ethanol	74.1% C_6H_6, 7.4% H_2O, 18.5% C_2H_5OH	64.9
Methanol–carbon tetrachloride	20.6% CH_3OH, 79.4% CCl_4	55.7
Ethanol–benzene	32.4% C_2H_5OH, 67.6% C_6H_6	67.8
Methanol–toluene	72.4% CH_3OH, 27.6% $C_6H_5CH_3$	63.7
Methanol–benzene	39.5% CH_3OH, 60.5% C_6H_6	58.3
Cyclohexane–ethanol	69.5% C_6H_{12}, 30.5% C_2H_5OH	64.9
2-Propanol–water	87.8% $(CH_3)_2CHOH$, 12.2% H_2O	80.4
Butyl acetate–water	72.9% $CH_3COOC_4H_9$, 27.1% H_2O	90.7
Phenol–water	9.2% C_6H_5OH, 90.8% H_2O	99.5

ing point than any of the components, it will be concentrated in the distillation flask as the distillate (pure B) is removed. The distillation of a solution of composition X would follow to the right along the lines in Figure 10.12. Once the composition of the material remaining in the flask has reached that of the azeotrope, the temperature will rise, and the azeotrope will begin to distill. The azeotrope will continue to distill until all the material in the distillation flask has been exhausted.

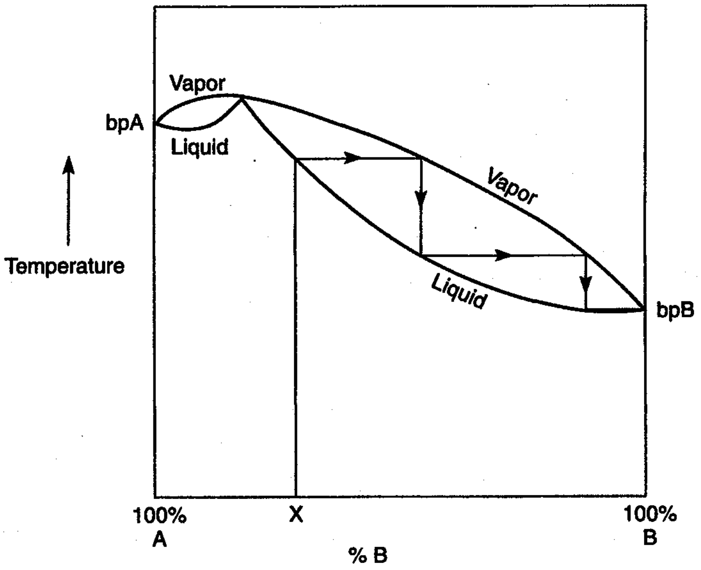

Figure 10.12 A maximum-boiling-point phase diagram.

TABLE 10.3 Maximum-Boiling Azeotropes

Azeotrope	Composition (Weight percentage)	Boiling Point (°C)
Acetone–chloroform	20.0% CH_3COCH_3, 80.0% $CHCl_3$	64.7
Chloroform–methyl ethyl ketone	17.0% $CHCl_3$, 83.0% $CH_3COCH_2CH_3$	79.9
Hydrochloric acid	20.2% HCl, 79.8% H_2O	108.6
Acetic acid–dioxane	77.0% CH_3COCH, 23.0% $C_4H_8O_2$	119.5
Benzaldehyde–phenol	49.0% C_6H_5CHO, 51.0% C_6H_5OH	185.6

Some maximum-boiling-point azeotropes are listed in Table 10.3. They are not nearly as common as minimum-boiling-point azeotropes.[3]

C. GENERALIZATIONS

There are some generalizations that can be made about azeotropic behavior. They are presented here without explanation, but you should be able to verify them by thinking through each case using the phase diagrams given. (Note that pure A is always to the left of the azeotrope in these diagrams, while pure B is to the right of the azeotrope.)

Minimum-Boiling-Point Azeotropes

Initial Composition	Experimental Result
to left of azeotrope	azeotrope distills first, pure A second
azeotrope	unseparable
to right of azeotrope	azeotrope distills first, pure B second

Maximum-Boiling Point Azeotropes

Initial Composition	Experimental Result
to left of azeotrope	pure A distills first, azeotrope second
azeotrope	unseparable
to right of azeotrope	pure B distills first, azeotrope second

---

[3] See Footnote 2.

10.8 AZEOTROPIC DISTILLATION: APPLICATIONS

There are numerous examples of chemical reactions in which the amount of product is low because of an unfavorable equilibrium. An example is the direct acid-catalyzed esterification of a carboxylic acid with an alcohol:

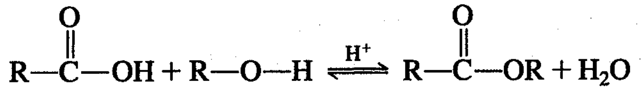

$$R{-}\overset{\overset{\displaystyle O}{\|}}{C}{-}OH + R{-}O{-}H \overset{H^+}{\rightleftharpoons} R{-}\overset{\overset{\displaystyle O}{\|}}{C}{-}OR + H_2O$$

Because the equilibrium does not favor formation of the ester, it must be shifted to the right, in favor of the product, by using an excess of one of the starting materials. In most cases, the alcohol is the least expensive reagent and is the material used in excess. Isopentyl acetate (Experiment 8) and methyl salicylate (Experiment 10) are examples of esters prepared by using one of the starting materials in excess.

Another way of shifting the equilibrium to the right is to remove one of the products from the reaction mixture as it is formed. In the above example, water can be removed as it is formed by **azeotropic distillation.** A common large-scale method is to use the Dean–Stark water separator shown in Figure 10.13A. In this technique, an in-

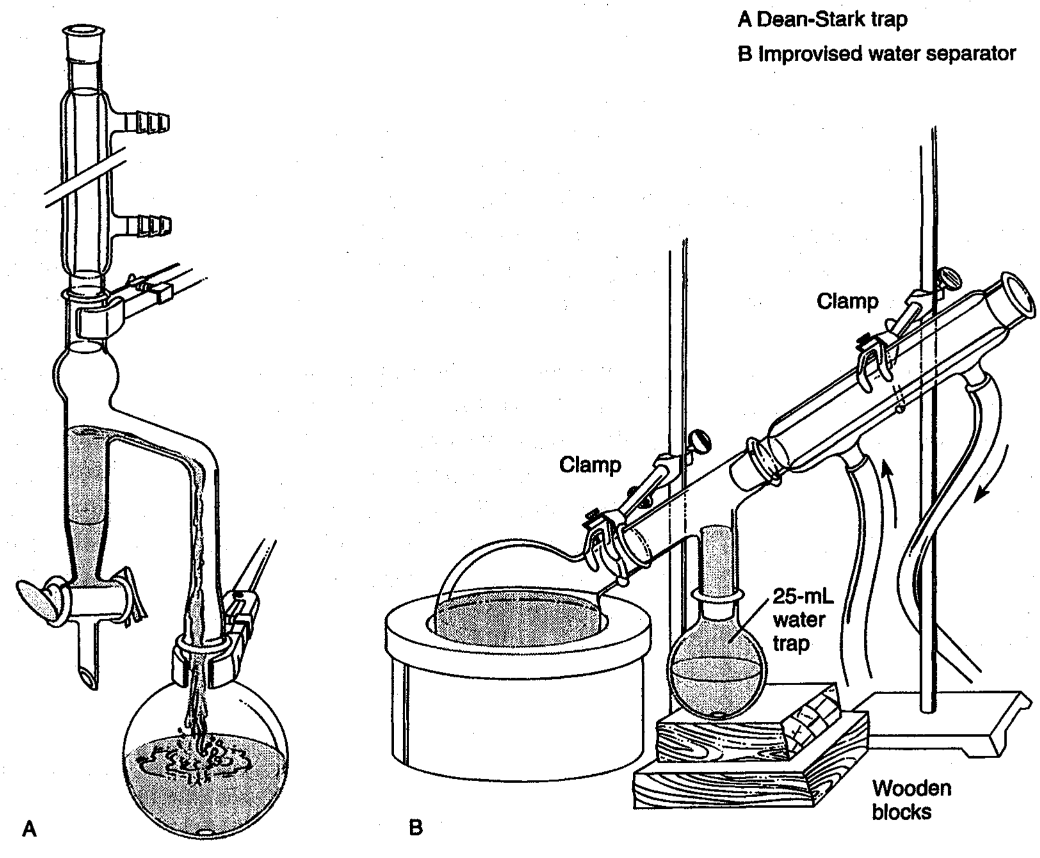

Figure 10.13 Large-scale water separators.

ert solvent, commonly benzene or toluene, is added to the reaction mixture contained in the round-bottom flask. The side-arm of the water separator is also filled with this solvent. If benzene is used, as the mixture is heated under reflux, the benzene–water azeotrope (bp 69.4°C, Table 10.2) distills out of the flask.[4] When the vapor condenses, it enters the side-arm directly below the condenser, and water separates from the benzene–water condensate; benzene and water mix as vapors, but they are not miscible as cooled liquids. Once the water (lower phase) separates from the benzene (upper phase), liquid benzene overflows from the side-arm back into the flask. The cycle is repeated continuously until no more water forms in the side-arm. You may calculate the weight of water that should theoretically be produced and compare this value with the amount of water collected in the side-arm. Because the density of water is 1.0, the volume of water collected can be compared directly with the calculated amount, assuming 100% yield.

An improvised water separator, constructed from the components found in the traditional organic kit, is shown in Figure 10.13B. Although this requires the condenser to be placed in a nonvertical position, it works quite well.

At the microscale level, water separation can be achieved using a standard distillation assembly with a water condenser and a Hickman head (Fig. 10.14). The side-ported variation of the Hickman head is the most convenient one to use for this purpose, but it is not essential. In this variation, you simply remove all the distillate (both solvent and water) several times during the course of the reaction. Use a Pasteur pipet to remove the distillate, as shown in Technique 8 (Fig. 8.7, p. 715). Because both the solvent and water are removed in this procedure, it may be desirable to add more solvent from time to time, adding it through the condenser with a Pasteur pipet.

The most important consideration in using azeotropic distillation to prepare an ester (described on p. 749) is that the azeotrope containing water must have a **lower boiling point** than the alcohol used. With ethanol, the benzene–water azeotrope boils at a much lower temperature (69.4°C) than ethanol (78.3°C), and the technique previously described works well. With higher-boiling-point alcohols, azeotropic distillation works well because of the large boiling-point difference between the azeotrope and the alcohol.

With methanol (bp 65°C), however, the boiling point of the benzene–water azeotrope is actually *higher* by about 5°C, and methanol distills first. Thus, in esterifications involving methanol, a totally different approach must be taken. For example, you can mix the carboxylic acid, methanol, the acid catalyst, and *1,2-dichloroethane* in a conventional reflux apparatus (Technique 3, Fig. 3.8, p. 619) without a water separator. During the reaction, water separates from the 1,2-dichloroethane because it is not miscible; however, the remainder of the components are soluble, so the reaction can continue. The equilibrium is shifted to the right by the “removal” of water from the reaction mixture.

---

[4] Actually, with ethanol, a lower-boiling-point, three-component azeotrope distills at 64.9°C (see Table 10.2). It consists of benzene–water–ethanol. Because some ethanol is lost in the azeotropic distillation, a large excess of ethanol is used in esterification reactions. The excess also helps to shift the equilibrium to the right.

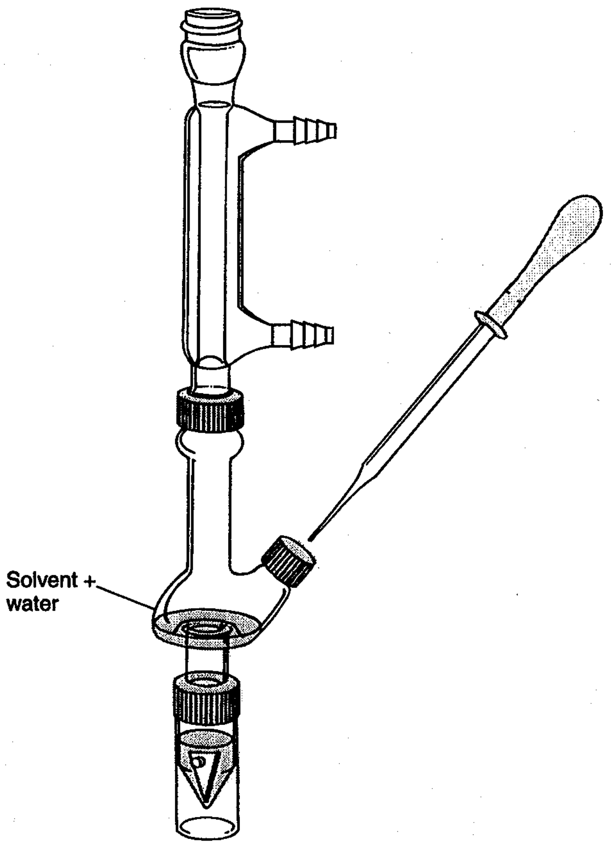

Figure 10.14 Microscale water separator (both layers are removed).

Azeotropic distillation is also used in other types of reactions, such as ketal or acetal formation, and in enamine formation.

Acetal Formation

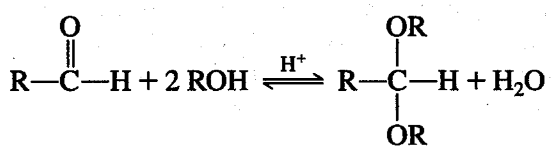

$$R\text{—}\overset{\displaystyle O}{\overset{\|}{C}}\text{—}H + 2\,ROH \underset{}{\overset{H^+}{\rightleftharpoons}} R\text{—}\underset{\displaystyle OR}{\underset{|}{\overset{\displaystyle OR}{\overset{|}{C}}}}\text{—}H + H_2O$$

Enamine Formation

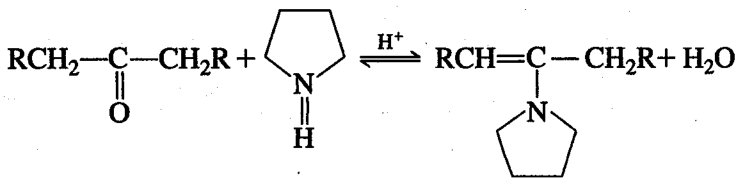

PROBLEMS

1. In the accompanying chart are approximate vapor pressures for benzene and toluene at various temperatures.

Temp (°C)	mmHg	Temp (°C)	mmHg
Benzene 30	120	Toluene 30	37
40	180	40	60
50	270	50	95
60	390	60	140
70	550	70	200
80	760	80	290
90	1010	90	405
100	1340	100	560
		110	760

(a) What is the mole fraction of each component if 3.9 g of benzene C_6H_6 is dissolved in 4.6 g of toluene C_7H_8?

(b) Assuming that this mixture is ideal, that is, it follows Raoult's Law, what is the partial vapor pressure of benzene in this mixture at 50°C?

(c) Estimate to the nearest degree the temperature at which the vapor pressure of the solution equals 1 atm (bp of the solution).

(d) Calculate the composition of the vapor (*mole fraction of each component*) that is in equilibrium in the solution at the boiling point of this solution.

(e) Calculate the composition in weight percentage of the vapor that is in equilibrium with the solution.

2. Estimate how many theoretical plates are needed to separate a mixture that has a mole fraction of B equal to 0.70 (70% B) in Figure 10.3.

3. Two moles of sucrose are dissolved in 8 moles of water. Assume that the solution follows Raoult's Law and that the vapor pressure of sucrose is negligible. The boiling point of water is 100°C. The distillation is carried out at 1 atm (760 mmHg).

(a) Calculate the vapor pressure of the solution when the temperature reaches 100°C.

(b) What temperature would be observed during the entire distillation?

(c) What would be the composition of the distillate?

(d) If a thermometer were immersed below the surface of the liquid of the boiling flask, what temperature would be observed?

4. Explain why the boiling point of a two-component mixture rises slowly throughout a simple distillation when the boiling-point differences are not large.

5. Given the boiling points of several known mixtures of A and B (mole fractions are known) and the vapor pressures of A and B in the pure state (P_A^0 and P_B^0) at these same temperatures, how would you construct a boiling-point-composition phase diagram for A and B? Give a stepwise explanation.

6. Describe the behavior on distillation of a 98% ethanol solution through an efficient column. Refer to Figure 10.11.

7. Construct an approximate boiling-point-composition diagram for a benzene-methanol system. The mixture shows azeotropic behavior (see Table 10.2). Include on the graph the boiling points of pure benzene and pure methanol and the boiling point of the azeotrope. Describe the behavior for a mixture that is initially rich in benzene (90%) and then for a mixture that is initially rich in methanol (90%).

8. Construct an approximate boiling-point-composition diagram for an acetone–chloroform system, which forms a maximum boiling azeotrope (Table 10.3). Describe the behavior on distillation of a mixture that is initially rich in acetone (90%), then describe the behavior of a mixture that is initially rich in chloroform (90%).

9. Two components have boiling points of 130 and 150°C. Estimate the number of theoretical plates needed to separate these substances in a fractional distillation.

10. A spinning-band column has an HETP of 0.25 in./plate. If the column has 12 theoretical plates, how long is it?

TECHNIQUE 11

Steam Distillation

The simple, vacuum, and fractional distillations described in Techniques 8, 9, and 10 are applicable to completely soluble (miscible) mixtures only. When liquids are *not* mutually soluble (immiscible), they can also be distilled, but with a somewhat different result. A mixture of immiscible liquids will boil at a lower temperature than the boiling points of any of the separate components as pure compounds. When steam is used to provide one of the immiscible phases, the process is called **steam distillation.** The advantage of this technique is that the desired material distills at a temperature below 100°C. Thus, if unstable or very high-boiling substances are to be removed from a mixture, decomposition is avoided. Because all gases mix, the two substances can mix in the vapor and codistill. Once the distillate is cooled, the desired component, which is not miscible, separates from the water. Steam distillation is used widely in isolating liquids from natural sources. It is also used in removing a reaction product from a tarry reaction mixture.

11.1 DIFFERENCES BETWEEN DISTILLATION OF MISCIBLE AND IMMISCIBLE MIXTURES

$$\text{MISCIBLE LIQUIDS} \quad P_{\text{total}} = P_{\text{A}}^{0}N_{\text{A}} + P_{\text{B}}^{0}N_{\text{B}} \qquad (1)$$

Two liquids A and B that are mutually soluble (miscible), and that do not interact, form an ideal solution and follow Raoult's Law, as shown in Equation 1. Note that the vapor pressures of pure liquids P^0_A and P^0_B are not added directly to give the total pressure P_{total} but are reduced by the respective mole fractions N_A and N_B. The total pressure above a miscible or homogeneous solution will depend on P^0_A and P^0_B and also N_A and N_B. Thus, the composition of the vapor will also depend on *both* the vapor pressures and the mole fractions of each component.

$$\text{IMMISCIBLE LIQUIDS} \quad P_{total} = P^0_A + P^0_B \tag{2}$$

In contrast, when two mutually insoluble (immiscible) liquids are "mixed" to give a heterogeneous mixture, each exerts its own vapor pressure, independently of the other, as shown in Equation 2. The mole fraction term does not appear in this equation, because the compounds are not miscible. You simply add the vapor pressures of the pure liquids P^0_A and P^0_B at a given temperature to obtain the total pressure above the mixture. When the total pressure equals 760 mmHg, the mixture boils. The composition of the vapor from an immiscible mixture, in contrast to the miscible mixture, is determined only by the vapor pressures of the two substances codistilling. Equation 3 defines the composition of the vapor from an immiscible mixture. Calculations involving this equation are given in Section 11.2.

$$\frac{\text{Moles A}}{\text{Moles B}} = \frac{P^0_A}{P^0_B} \tag{3}$$

A mixture of two immiscible liquids boils at a lower temperature than the boiling points of either component. The explanation for this behavior is like that given for minimum-boiling-point azeotropes (Technique 10, Section 10.7). Immiscible liquids behave as they do because an extreme incompatibility between the two liquids leads to higher combined vapor pressures than Raoult's Law would predict. The higher combined vapor pressures cause a lower boiling point for the mixture than for either single component. Thus, you may think of steam distillation as a special type of azeotropic distillation in which the substance is completely insoluble in water.

The differences in behavior of miscible and immiscible liquids, where it is assumed that P^0_A equals P^0_B, are shown in Figure 11.1. Note that with miscible liquids, the composition of the vapor depends on the relative amounts of A and B present (Fig. 11.1A). Thus, the composition of the vapor must change during a distillation. In contrast, the composition of the vapor with immiscible liquids is independent of the amounts of A and B present (Fig. 11.1B). Hence, the vapor composition must remain *constant* during the distillation of such liquids, as predicted by Equation 3. Immiscible liquids act as if they were being distilled simultaneously from separate compartments, as shown in Figure 11.1B, even though in practice they are "mixed" during a steam distillation. Because all gases mix, they do give rise to a homogeneous vapor and codistill.

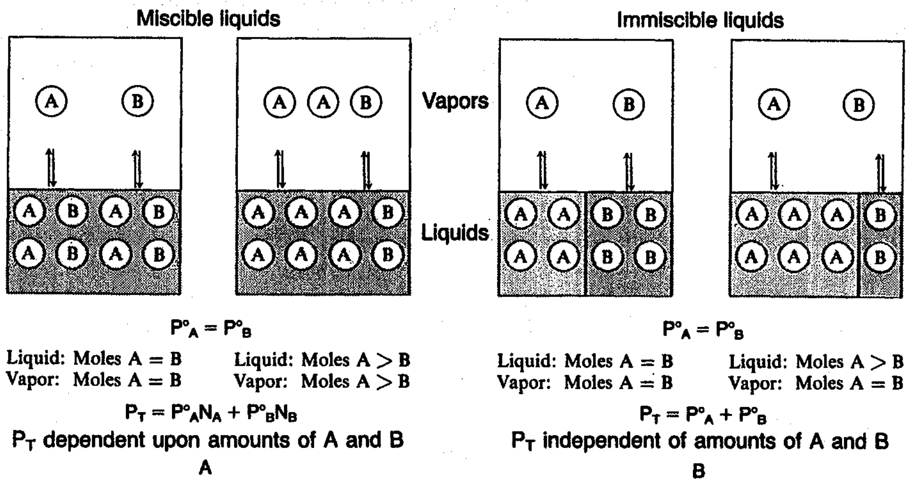
Miscible liquids
Immiscible liquids
Vapors
Liquids
A
B
P°A = P°B
Liquid: Moles A = B
Vapor: Moles A = B
Liquid: Moles A > B
Vapor: Moles A > B
PT = P°ANA + P°BNB
PT dependent upon amounts of A and B
A
P°A = P°B
Liquid: Moles A = B
Vapor: Moles A = B
Liquid: Moles A > B
Vapor: Moles A = B
PT = P°A + P°B
PT independent of amounts of A and B
B

Figure 11.1 Total pressure behavior for miscible and immiscible liquids. (A) Ideal miscible liquids follow Raoult's Law: P_T depends on the mole fractions and vapor pressures of A and B. (B) Immiscible liquids do not follow Raoult's Law: P_T depends only on the vapor pressures of A and B.

11.2 IMMISCIBLE MIXTURES: CALCULATIONS

The composition of the distillate is constant during a steam distillation, as is the boiling point of the mixture. The boiling points of steam-distilled mixtures will always be below the boiling point of water (bp 100°C) as well as the boiling point of any of the other substances distilled. Some representative boiling points and compositions of steam distillates are given in Table 11.1. Note that the higher the boiling point of a pure substance, the more closely the temperature of the steam distillate approaches, but does not exceed, 100°C. This is a reasonably low temperature, and it avoids the decomposition that might result at high temperatures with a simple distillation.

TABLE 11.1 Boiling Points and Compositions of Steam Distillates

Mixture	Boiling Point of Pure Substance (°C)	Boiling Point of Mixture (°C)	Composition (% Water)
Benzene–water	80.1	69.4	8.9%
Toluene–water	110.6	85.0	20.2%
Hexane–water	69.0	61.6	5.6%
Heptane–water	98.4	79.2	12.9%
Octane–water	125.7	89.6	25.5%
Nonane–water	150.8	95.0	39.8%
1-Octanol–water	195.0	99.4	90.0%

For immiscible liquids, the molar proportions of two components in a distillate equal the ratio of their vapor pressures in the boiling mixture, as given in Equation 3. When Equation 3 is rewritten for an immiscible mixture involving water, Equation 4 results. Equation 4 can be modified by substituting the relation moles = (weight/molecular weight) to give Equation 5.

$$\frac{\text{Moles substance}}{\text{Moles water}} = \frac{P^0_{\text{substance}}}{P^0_{\text{water}}} \tag{4}$$

$$\frac{\text{Wt substance}}{\text{Wt water}} = \frac{(P^0_{\text{substance}})(\text{Molecular weight}_{\text{substance}})}{(P^0_{\text{water}})(\text{Molecular weight}_{\text{water}})} \tag{5}$$

A sample calculation using this equation is given in Figure 11.2. Notice that the result of this calculation is very close to the experimental value given in Table 11.1.

11.3 STEAM DISTILLATION: METHODS

Two methods for steam distillation are in general use in the laboratory: the **direct method** and the **live steam method.** In the first method, steam is generated *in situ* (in place) by heating a distillation flask containing the compound and water. In the second method, steam is generated outside and is passed into the distillation flask using an inlet tube.

Problem How many grams of water must be distilled to steam distill 1.55 g of 1-octanol from an aqueous solution? What will be the composition (wt %) of the distillate? The mixture distills at 99.4°C.

Answer The vapor pressure of water at 99.4°C must be obtained from the CRC Handbook (= 744 mmHg).

(a) Obtain the partial pressure of 1-octanol.

$$P^\circ_{\text{1-octanol}} = P_{\text{total}} - P^\circ_{\text{water}}$$
$$P^\circ_{\text{1-octanol}} = (760 - 744) = 16 \text{ mmHg}$$

(b) Obtain the composition of the distillate.

$$\frac{\text{wt 1-octanol}}{\text{wt water}} = \frac{(16)(130)}{(744)(18)} = 0.155 \text{ g/g-water}$$

(c) Clearly, 10 g of water must be distilled.

$$(0.155 \text{ g/g-water})(10 \text{ g-water}) = 1.55 \text{ g 1-octanol}$$

(d) Calculate the **weight** percentages.

$$\text{1-octanol} = 1.55 \text{ g}/(10 \text{ g} + 1.55 \text{ g}) = 13.4\%$$
$$\text{water} = 10 \text{ g}/(10 \text{ g} + 1.55 \text{ g}) = 86.6\%$$

Figure 11.2 Sample calculations for a steam distillation.

A. DIRECT METHOD

Standard Scale. A large-scale direct method steam distillation is illustrated in Figure 11.3. Although a heating mantle may be used, it is probably best to use a flame with this method, because a large volume of water must be heated rapidly. A boiling stone must be used to prevent bumping. The separatory funnel allows more water to be added during the course of the distillation.

Distillate is collected as long as it is either cloudy or milky white in appearance. Cloudiness indicates that an immiscible liquid is separating. When the distillate runs clear in the distillation, it is usually a sign that only water is distilling. However, there are some steam distillations where the distillate is never cloudy, even though material has codistilled. You must observe carefully, and be sure to collect enough distillate that all of the organic material codistills.

Microscale. The direct method of steam distillation is the only one suitable for microscale reactions. Steam is produced in the conical vial or distillation flask (*in situ*) by heating water to its boiling point in the presence of the compound to be distilled. This method works well for small amounts of materials. A microscale steam distillation apparatus is shown in Figure 11.4. Water and the compound to be distilled are placed in the flask and heated. A stirring bar or a boiling stone should be used to prevent bumping. The vapors of the water and the desired compound codistill when they are heated. They are

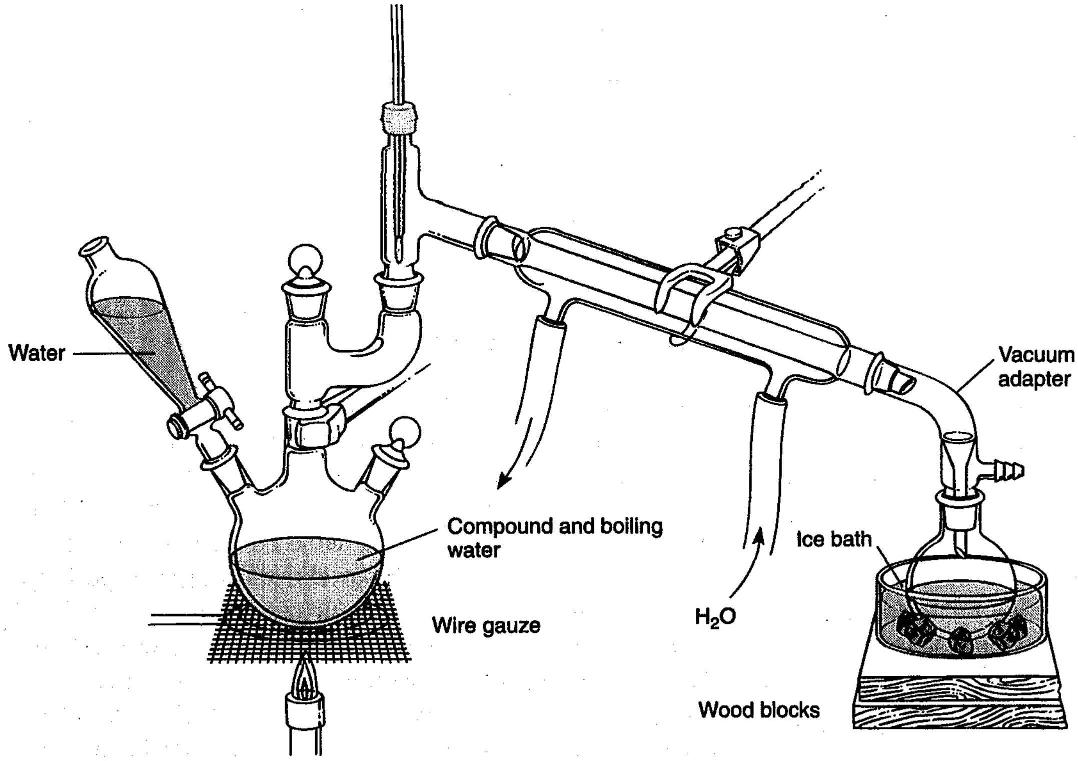

Figure 11.3 Macroscale direct steam distillation.

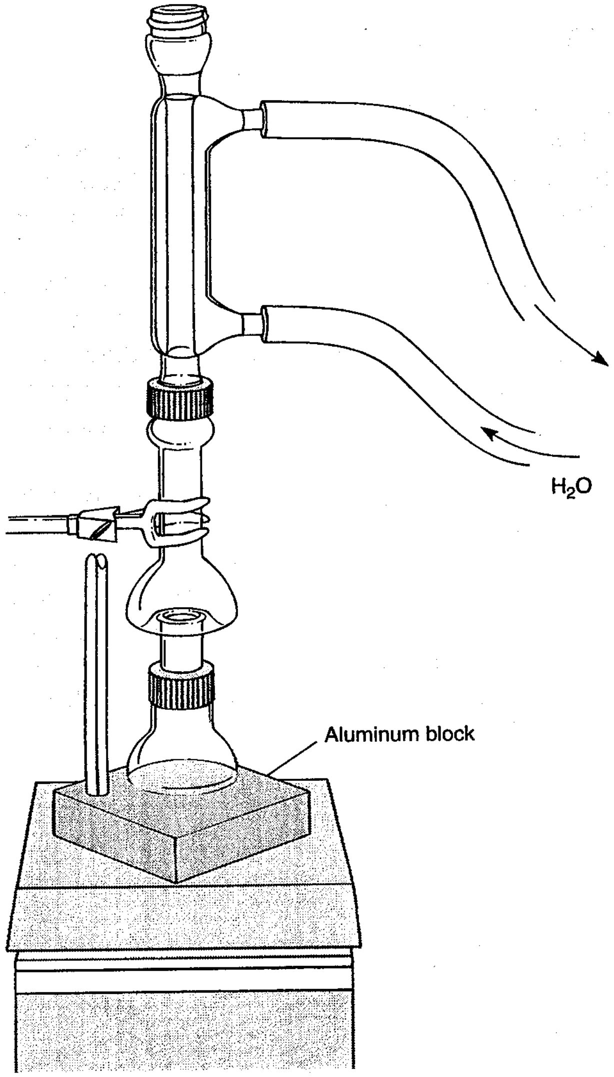

Figure 11.4 Microscale steam distillation.

condensed and collect in the Hickman head. When the Hickman head fills, the distillate is removed with a Pasteur pipet and placed in another vial for storage. For the typical microscale experiment, it will be necessary to fill the well and remove the distillate three or four times. All of these distillate fractions are placed in the same storage container. The efficiency in collecting the distillate can sometimes be improved if the inside walls of the Hickman head are rinsed several times into the well. A Pasteur pipet is used to perform the rinsing. Distillate is withdrawn from the well, and then it is used to wash the walls of the Hickman head all the way around the head. After the walls have been washed and when the well is full, the distillate can be withdrawn and transferred to the storage container. It may be necessary to add more water during the course of the distillation. More water is added (remove the condenser if used) through the center of the Hickman head by using a Pasteur pipet.

Semi-Microscale. The apparatus shown in Figure 8.5, page 714, may also be used to perform a steam distillation at the microscale level or slightly above. This apparatus avoids the need to empty the collected distillate during the course of the distillation as is required when a Hickman head is used.

B. LIVE STEAM METHOD

Standard Scale. A large-scale steam distillation using the live steam method is shown in Figure 11.5. If steam lines are available in the laboratory, they may be attached directly to the steam trap (purge them first to drain water). If steam lines are not available, an external steam generator (see inset) must be prepared. The external generator usually will require a flame to produce steam at a rate fast enough for the distillation. When the distillation is first

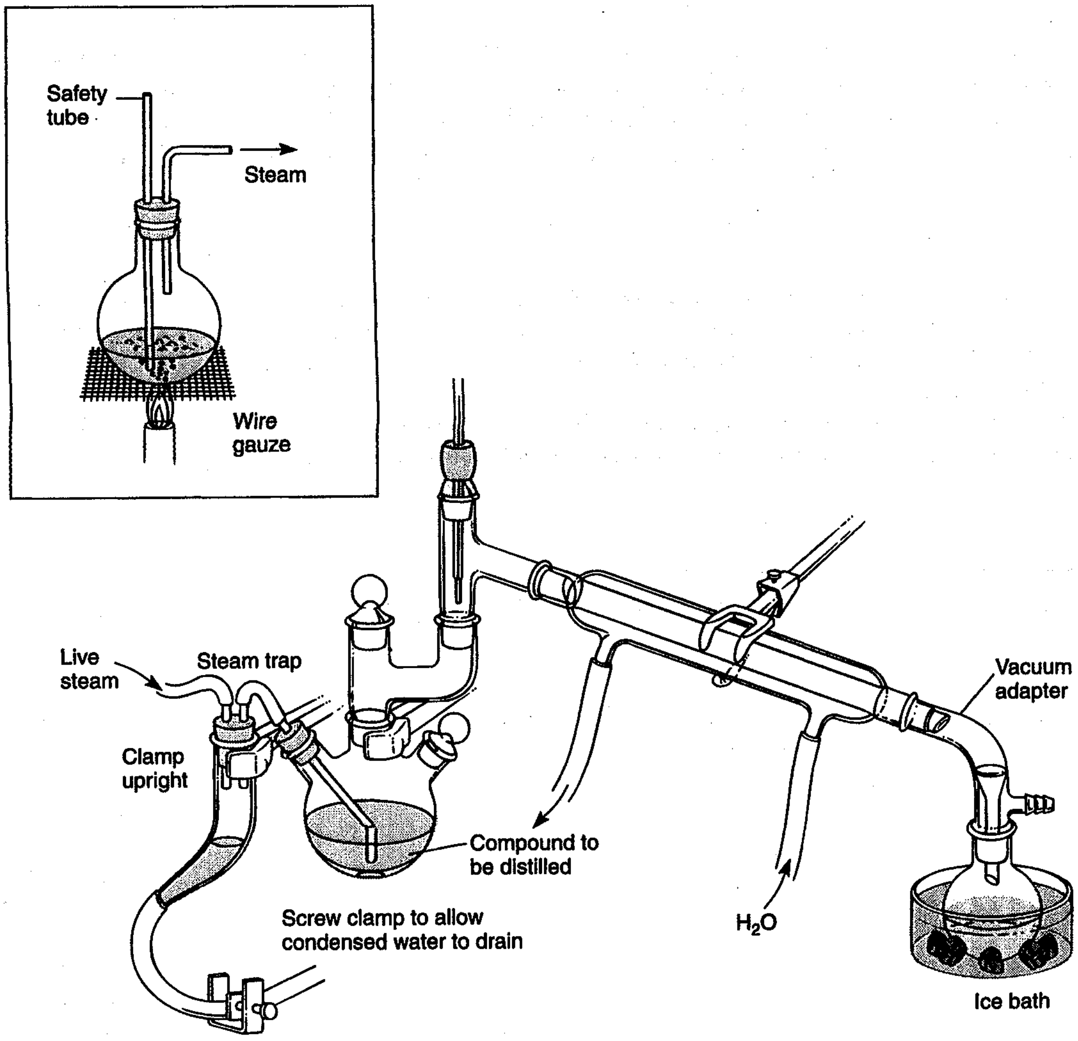
Safety tube
Steam
Wire gauze
Live steam
Steam trap
Clamp upright
Compound to be distilled
Screw clamp to allow condensed water to drain
H_2O
Vacuum adapter
Ice bath

Figure 11.5 Macroscale steam distillation using live steam.

started, the clamp at the bottom of the steam trap is left open. The steam lines will have a large quantity of condensed water in them until they are well heated. When the lines become hot and condensation of steam ceases, the clamp may be closed. Occasionally, the clamp will have to be reopened to remove condensate. In this method, the steam agitates the mixture as it enters the bottom of the flask, and a stirrer or boiling stone is not required.

Caution. Hot steam can produce very severe burns.

Sometimes it is helpful to heat the three-necked distilling flask with a heating mantle (or flame) to prevent excessive condensation at that point. Steam must be admitted at a fast enough rate for you to see the distillate condensing as a milky white fluid in the condenser. The vapors that codistill will separate on cooling to give this cloudiness. When the condensate becomes clear, the distillation is near the end. The flow of water through the condenser should be faster than in other types of distillation to help cool the vapors. Make sure the vacuum adapter remains cool to the touch. An ice bath may be used to cool the receiving flask if desired. When the distillation is to be stopped, the screw clamp on the steam trap should be opened, and the steam inlet tube must be removed from the three-necked flask. If this is not done, liquid will back up into the tube and steam trap.

PROBLEMS

1. Calculate the weight of benzene codistilled with each gram of water and the percentage composition of the vapor produced during a steam distillation. The boiling point of the mixture is 69.4°C. The vapor pressure of water at 69.4°C is 227.7 mmHg. Compare the result with the data in Table 11.1.

2. Calculate the approximate boiling point of a mixture of bromobenzene and water at atmospheric pressure. A table of vapor pressures of water and bromobenzene at various temperatures is given.

	Vapor Pressures (mmHg)	
Temperature (°C)	**Water**	**Bromobenzene**
93	588	110
94	611	114
95	634	118
96	657	122
97	682	127
98	707	131
99	733	136

3. Calculate the weight of nitrobenzene that codistills (bp 99°C of mixture) with each gram of water during a steam distillation. You may need the data given in Problem 2.
4. A mixture of *p*-nitrophenol and *o*-nitrophenol can be separated by steam distillation. The *o*-nitrophenol is steam volatile and the *para* isomer is not volatile. Explain. Base your answer on the ability of the isomers to form hydrogen bonds internally.

TECHNIQUE 12

Column Chromatography

The most modern and sophisticated methods of separating mixtures that the organic chemist has available all involve **chromatography.** Chromatography is defined as the separation of a mixture of two or more different compounds or ions by distribution between two phases, one of which is stationary and the other moving. Various types of chromatography are possible, depending on the nature of the two phases involved: **solid–liquid** (column, thin-layer, and paper), **liquid–liquid,** (high-performance liquid), and **gas–liquid** (vapor–phase) chromatographic methods are common.

All chromatography works on much the same principle as solvent extraction (Technique 7). Basically, the methods depend on the differential solubilities or adsorptivities of the substances to be separated relative to the two phases between which they are to be partitioned. In this chapter, column chromatography, a solid–liquid method, is considered. High-performance liquid chromatography is discussed in Technique 13. Thin-layer chromatography is examined in Technique 14; gas chromatography, a gas–liquid method, is discussed in Technique 15.

12.1 ADSORBENTS

Column chromatography is a technique based on both adsorptivity and solubility. It is a solid–liquid phase-partitioning technique. The solid may be almost any material that does not dissolve in the associated liquid phase; the solids used most commonly are silica gel $SiO_2 \cdot xH_2O$, also called silicic acid, and alumina $Al_2O_3 \cdot xH_2O$. These compounds are used in their powdered or finely ground forms (usually 200- to 400-mesh).

Most alumina used for chromatography is prepared from the impure ore bauxite $Al_2O_3 \cdot xH_2O + Fe_2O_3$. The bauxite is dissolved in hot sodium hydroxide and filtered to remove the insoluble iron oxides; the alumina in the ore forms the soluble amphoteric hydroxide $Al(OH)_4^-$. The hydroxide is precipitated by CO_2, which reduces the pH, as $Al(OH)_3$. When heated, the $Al(OH)_3$ loses water to form pure alumina Al_2O_3.

$$\text{Bauxite (crude)} \xrightarrow{\text{hot NaOH}} Al(OH)_4^-(aq) + Fe_2O_3 \text{ (insoluble)}$$

$$Al(OH)_4^-(aq) + CO_2 \longrightarrow Al(OH)_3 + HCO_3^-$$

$$2\ Al(OH)_3 \xrightarrow{\text{heat}} Al_2O_3(s) + 3\ H_2O$$

Alumnia prepared in this way is called **basic alumina** because it still contains some hydroxides. Basic alumina cannot be used for chromatography of compounds that are base-sensitive. Therefore, it is washed with acid to neutralize the base, giving **acid-washed alumina.** This material is unsatisfactory unless it has been washed with enough water to remove *all* the acid; on being so washed, it becomes the best chromatographic material, called **neutral alumina.** If a compound is acid-sensitive, either basic or neutral alumina must be used. You should be careful to ascertain what type of alumnia is being used for chromatography. Silica gel is not available in any form other than that suitable for chromatography.

12.2 INTERACTIONS

If powdered or finely ground alumina (or silica gel) is added to a solution containing an organic compound, some of the organic compound will *adsorb* onto or adhere to the fine particles of alumina. Many kinds of intermolecular forces cause organic molecules to bind to alumnia. These forces vary in strength according to their type. Nonpolar compounds bind to the alumina using only van der Waals forces. These are weak forces, and nonpolar molecules do not bind strongly unless they have extremely high molecular weights. The most important interactions are those typical of polar organic compounds. Either these forces are of the dipole–dipole type or they involve some direct interaction (coordination, hydrogen-bonding, or salt formation). These types of interactions are illustrated in Figure 12.1, which for convenience shows only a portion of the alumina structure. Similar interactions occur with silica gel. The strengths of such interactions vary in the approximate order:

Salt formation > Coordination > Hydrogen-bonding > Dipole–dipole > van der Waals

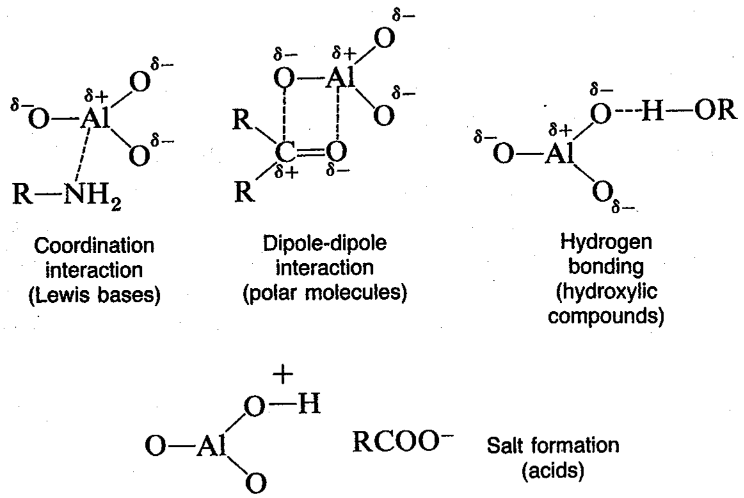

Figure 12.1 Possible interactions of organic compounds with alumina.

Strength of interaction varies among compounds. For instance, a strongly basic amine would bind more strongly than a weakly basic one (by coordination). In fact, strong bases and strong acids often interact so strongly that they *dissolve* alumina to some extent. You can use the following rule of thumb:

> **The more polar the functional group, the stronger the bond to alumina (or silica gel).**

A similar rule holds for solubility. Polar solvents dissolve polar compounds more effectively than nonpolar solvents: nonpolar compounds are dissolved best by nonpolar solvents. Thus, the extent to which any given solvent can wash an adsorbed compound from alumina depends almost directly on the relative polarity of the solvent. For example, although a ketone adsorbed on alumina might not be removed by hexane, it might be removed completely by chloroform. For any adsorbed material, a kind of **distribution** equilibrium can be envisioned between the adsorbent material and the solvent. This is illustrated in Figure 12.2.

The distribution equilibrium is *dynamic,* with molecules constantly *adsorbing* from the solution and *desorbing* into it. The average number of molecules remaining adsorbed on the solid particles at equilibrium depends both on the particular molecule (RX) involved and the dissolving power of the solvent with which the adsorbent must compete.

12.3 PRINCIPLE OF COLUMN CHROMATOGRAPHIC SEPARATION

The dynamic equilibrium mentioned previously, and the variations in the extent to which different compounds adsorb on alumina or silica gel, underlie a versatile and ingenious method for *separating* mixtures of organic compounds. In this method, the mixture

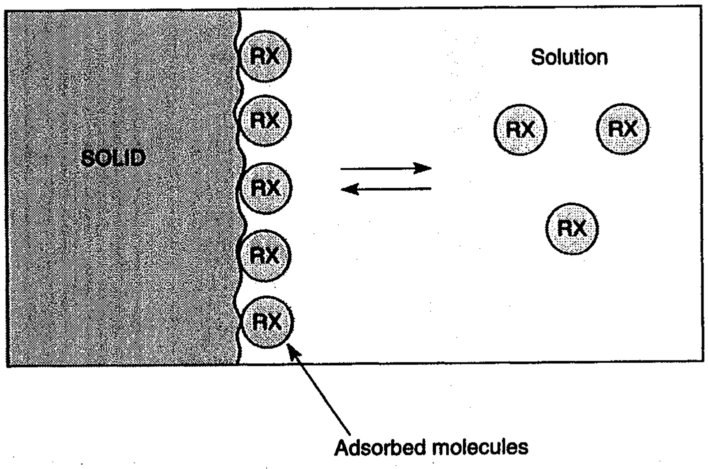

Figure 12.2 Dynamic adsorption equilibrium.

of compounds to be separated is introduced onto the top of a cylindrical glass column (Figure 12.3) **packed** or filled with fine alumina particles (stationary solid phase). The adsorbent is continuously washed by a flow of solvent (moving phase) passing through the column.

Initially, the components of the mixture adsorb onto the alumina particles at the top of the column. The continuous flow of solvent through the column **elutes,** or washes, the solutes off the alumina and sweeps them down the column. The solutes (or materials to be separated) are called **eluates** or **elutants;** and the solvents are called **eluents.** As the solutes pass down the column to fresh alumina, new equilibria are established between the adsorbent, the solutes, and the solvent. The constant equilibration means that different compounds will move down the column at differing rates depending on their relative affinity for the adsorbent on one hand, and for the solvent on the other. Because the number of alumina particles is large, because they are closely packed, and because fresh solvent is being added continuously, the number of equilibrations between adsorbent and solvent that the solutes experience is enormous.

As the components of the mixture are separated, they begin to form moving bands (or zones), each band containing a single component. If the column is long enough and the other parameters (column diameter, adsorbent, solvent, and flow rate) are correctly chosen, the bands separate from one another, leaving gaps or bands of pure solvent in between. As each band (solvent and solute) passes out the bottom of the column, it can be collected before the next band arrives. If the parameters mentioned are poorly chosen, the various bands either overlap or coincide, in which case either a poor separation or no separation at all is the result. A successful chromatographic separation is illustrated in Figure 12.4.

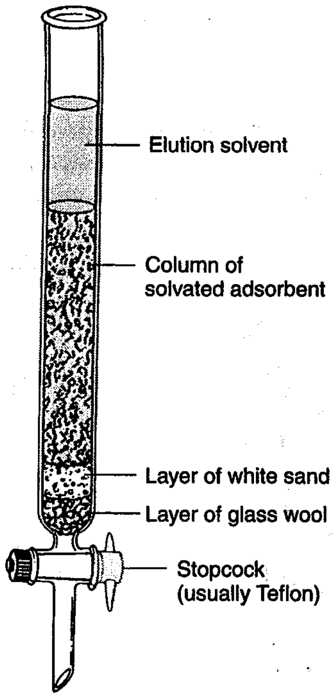

Figure 12.3 Chromatographic column.

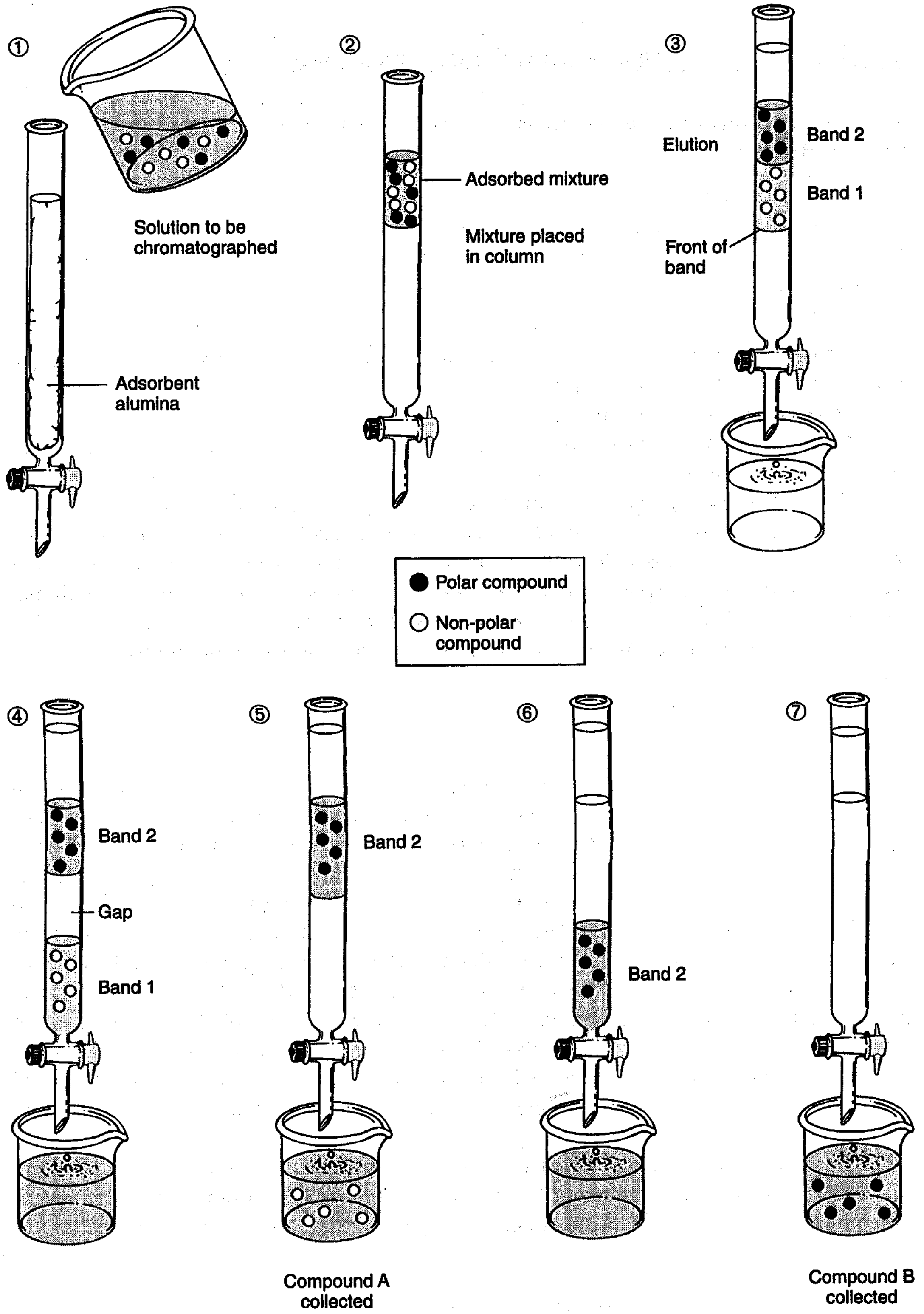
①
Solution to be chromatographed
Adsorbent alumina
②
Adsorbed mixture
Mixture placed in column
③
Elution
Band 2
Band 1
Front of band
● Polar compound
○ Non-polar compound
④
Band 2
Gap
Band 1
⑤
Band 2
⑥
Band 2
⑦
Compound A collected
Compound B collected

Figure 12.4 Sequence of steps in a chromatographic separation.

12.4 PARAMETERS AFFECTING SEPARATION

The versatility of column chromatography results from the many factors that can be adjusted. These include:

1. Adsorbent chosen
2. Polarity of the column or solvents chosen
3. Size of the column (both length and diameter) relative to the amount of material to be chromatographed
4. Rate of elution (or flow)

By carefully choosing the conditions, almost any mixture can be separated. This technique has even been used to separate optical isomers. An optically active solid-phase adsorbent was used to separate the enantiomers.

Two fundamental choices for anyone attempting a chromatographic separation are the kind of adsorbent and the solvent system. In general, nonpolar compounds pass through the column faster than polar compounds, because they have a smaller affinity for the adsorbent. If the adsorbent chosen binds all the solute molecules (both polar and nonpolar) strongly, they will not move down the column. On the other hand, if too polar a solvent is chosen, all the solutes (polar and nonpolar) may simply be washed through the column, with no separation taking place. The adsorbent and the solvent should be chosen so that neither is favored excessively in the equilibrium competition for solute molecules.[1]

A. ADSORBENTS

In Table 12.1, various kinds of adsorbents (solid phases) used in column chromatography are listed. The choice of adsorbent often depends on the types of compounds to be separated. Cellulose, starch, and sugars are used for polyfunctional plant and animal materials (natural products) that are very sensitive to acid–base interactions. Magnesium silicate is often used for separating acetylated sugars, steroids, and essential oils. Silica gel and Florisil are relatively mild toward most compounds and are widely used for a variety of functional groups—hydrocarbons, alcohols, ketones, esters, acids, azo compounds, amines. Alumina is the most widely used adsorbent and is obtained in the three forms mentioned in Section 12.1: acidic, basic, and neutral. The pH of acidic or acid-washed alumina is approximately 4. This adsorbent is particularly useful for separating acidic materials such as carboxylic acids and amino acids. Basic alumina has a pH of 10 and is useful in separating amines. Neutral alumina can be used to separate a variety of nonacidic and nonbasic materials.

The approximate strength of the various adsorbents listed in Table 12.1 is also given. The order is only approximate and therefore it may vary. For instance, the strength, or separating abilities, of alumina and silica gel largely depend on the amount of water pre-

---

[1] Often the chemist uses thin-layer chromatography (TLC), which is described in Technique 14, to arrive at the best choices of solvents and adsorbents for the best separation. The TLC experimentation can be performed quickly and with extremely small amounts (microgram quantities) of the mixture to be separated. This saves significant time and materials.

TABLE 12.1 Solid Adsorbents for Column Chromatography

Paper	**Increasing strength of binding interactions toward polar compounds** ↓
Cellulose	
Starch	
Sugars	
Magnesium silicate	
Calcium sulfate	
Silicic acid	
Silica gel	
Florisil	
Magnesium oxide	
Aluminum oxide (Alumina)*	
Activated charcoal (Norit)	

*Basic, acid-washed, and neutral.

sent. Water binds very tightly to either adsorbent, taking up sites on the particles that could otherwise be used for equilibration with solute molecules. If one adds water to the adsorbent, it is said to have been **deactivated.** Anhydrous alumina or silica gel are said to be highly **activated.** High activity is usually avoided with these adsorbents. Use of the highly active forms of either alumina or silica gel, or of the acidic or basic forms of alumina, can often lead to molecular rearrangement or decomposition in certain types of solute compounds.

The chemist can select the degree of activity that is appropriate to carry out a particular separation. To accomplish this, highly activated alumina is mixed thoroughly with a precisely measured quantity of water. The water partially hydrates the alumina and thus reduces its activity. By carefully determining the amount of water required, the chemist can have available an entire spectrum of possible activities.

B. SOLVENTS

In Table 12.2, some common chromatographic solvents are listed along with their relative ability to dissolve polar compounds. Sometimes, a single solvent can be found that will separate all the components of a mixture. Sometimes, a mixture of solvents can be found that will achieve separation. More often, you must start elution with a nonpolar solvent to remove relatively nonpolar compounds from the column and then gradually increase the solvent polarity to force compounds of greater polarity to come down the column, or elute. The approximate order in which various classes of compounds elute by this procedure is given in Table 12.3. In general, nonpolar compounds travel through the column faster (elute first), and polar compounds travel more slowly (elute last). However, molecular weight is also a factor in determining the order of elution. A nonpolar compound of high molecular weight travels more slowly than a nonpolar compound of low molecular weight, and it may even be passed by some polar compounds.

TABLE 12.2 Solvents (Eluents) for Chromatography

Petroleum ether	
Cyclohexane	
Carbon tetrachloride*	
Toluene	
Chloroform*	
Methylene chloride	**Increasing polarity and “solvent power” toward polar functional groups**
Diethyl ether	
Ethyl acetate	
Acetone	
Pyridine	
Ethanol	
Methanol	
Water	
Acetic acid	

*Suspected carcinogens.

When the polarity of the solvent has to be changed during a chromatographic separation, some precautions must be taken. Rapid changes from one solvent to another are to be avoided (especially when silica gel or alumina is involved). Usually, small percentages of a new solvent are mixed slowly into the one in use until the percentage reaches the desired level. If this is not done, the column packing often “cracks” as a result of the heat liberated when alumina or silica gel is mixed with a solvent. The solvent solvates the adsorbent, and the formation of a weak bond generates heat.

$$\text{Solvent} + \text{Alumina} \longrightarrow (\text{Alumina} \cdot \text{Solvent}) + \text{Heat}$$

TABLE 12.3 Elution Sequence for Compounds

Hydrocarbons	**Fastest (will elute with nonpolar solvent)**
Olefins	
Ethers	
Halocarbons	
Aromatics	
Ketones	**Order of elution**
Aldehydes	
Esters	
Alcohols	
Amines	
Acids, strong bases	**Slowest (need a polar solvent)**

Often, enough heat is generated locally to evaporate the solvent. The formation of vapor creates bubbles, which forces a separation of the column packing; this is called **cracking.** A cracked column does not produce a good separation, because it has discontinuities in the packing. The way in which a column is packed or filled is also very important in preventing cracking.

That the solvent itself has a tendency to adsorb on the alumina is an important factor in how compounds move down the column. The solvent can displace the adsorbed compound if it is more polar than the compound and, hence, can move it down the column. Thus, a more polar solvent not only dissolves more compound but also is effective in removing the compound from the alumina, because it displaces the compound from its site of adsorption.

Certain solvents should be avoided with alumina or silica gel, especially with the acidic, basic, and highly active forms. For instance, with any of these adsorbents, acetone dimerizes *via* an aldol condensation to give diacetone alcohol. Mixtures of esters *transesterify* (exchange their alcoholic portions) when ethyl acetate or an alcohol is the eluent. Finally, the most active solvents (pyridine, methanol, water, and acetic acid) dissolve and elute some of the adsorbent itself. Generally, try to avoid going to solvents more polar than diethyl ether or methylene chloride in the eluent series (Table 12.2).

C. COLUMN SIZE AND ADSORBENT QUANTITY

The column size and the amount of adsorbent must also be selected correctly to separate a given amount of sample well. As a rule of thumb, the amount of adsorbent should be 25 to 30 times, by weight, the amount of material to be separated by chromatography. Furthermore, the column should have a height-to-diameter ratio of about 8:1. Some typical relations of this sort are given in Table 12.4.

Note, as a caution, that the difficulty of the separation is also a factor in determining the size and length of the column to be used and in the amount of adsorbent needed. Compounds that do not separate easily may require larger columns and more adsorbent than specified in Table 12.4. For easily separated compounds, a smaller column and less adsorbent may suffice.

TABLE 12.4 Size of Column and Amount of Adsorbent for Typical Sample Sizes

Amount of Sample (g)	Amount of Adsorbent (g)	Column Diameter (mm)	Column Height (mm)
0.01	0.3	3.5	30
0.10	3.0	7.5	60
1.00	30.0	16.0	130
10.00	300.0	35.0	280

D. FLOW RATE

The rate at which solvent flows through the column is also significant in the effectiveness of a separation. In general, the time the mixture to be separated remains on the column is directly proportional to the extent of equilibration between stationary and moving phases. Thus, similar compounds eventually separate if they remain on the column long enough. The time a material remains on the column depends on the flow rate of the solvent. If the flow is too slow, however, the dissolved substances in the mixture may diffuse faster than the rate at which they move down the column. Then the bands grow wider and more diffuse, and the separation becomes poor.

12.5 PACKING THE COLUMN: TYPICAL PROBLEMS

The most critical operation in column chromatography is packing (filling) the column with adsorbent. The **column packing** must be evenly packed and free of irregularities, air bubbles, and gaps. As a compound travels down the column, it moves in an advancing zone, or **band.** It is important that the leading edge, or **front,** of this band be horizontal, or perpendicular to the long axis of the column. If two bands are close together and do not have horizontal band fronts, it is impossible to collect one band while completely excluding the other. The leading edge of the second band begins to elute before the first band has finished eluting. This condition can be seen in Figure 12.5. There

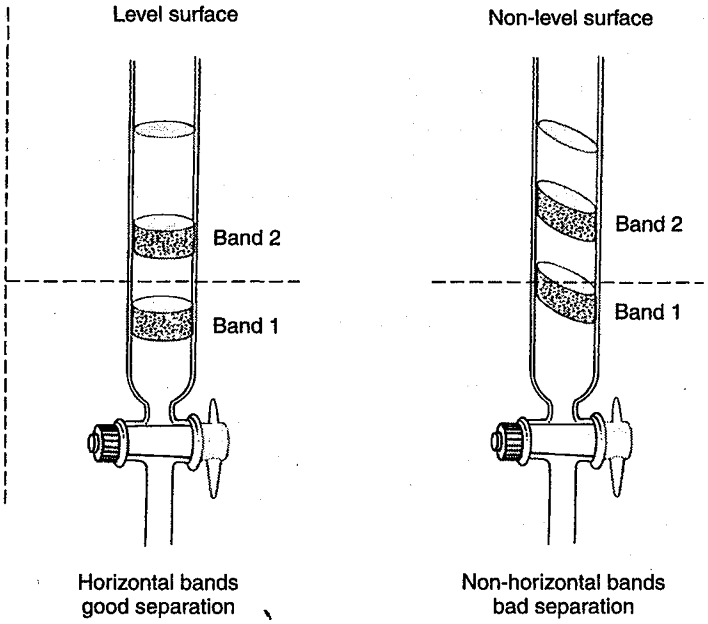

Figure 12.5 Comparison of horizontal and nonhorizontal band fronts.

are two main reasons for this problem. First, if the top surface edge of the adsorbent packing is not level, nonhorizontal bands result. Second, bands may be nonhorizontal if the column is not held in an exactly vertical position in both planes (front-to-back and side-to-side). When you are preparing a column, you must watch both these factors carefully.

Another phenomenon, called **streaming** or **channeling,** occurs when part of the band front advances ahead of the major part of the band. Channeling occurs if there are any cracks or irregularities in the adsorbent surface or any irregularities caused by air bubbles in the packing. A part of the advancing front moves ahead of the rest of the band by flowing through the channel. Two examples of channeling are shown in Figure 12.6.

12.6 PACKING THE COLUMN: STANDARD-SCALE AND SEMI-MICROSCALE METHODS

The following methods are used to avoid problems resulting from uneven packing and column irregularities. These procedures should be followed carefully in preparing a chromatography column. Failure to pay close attention to the preparation of the column may well affect the quality of the separation.

Preparation of a column involves two distinct stages. In the first stage, a support base on which the packing will rest is prepared. This must be done so that the packing, a finely divided material, does not wash out of the bottom of the column. In the second stage, the column of adsorbent is deposited on top of the supporting base.

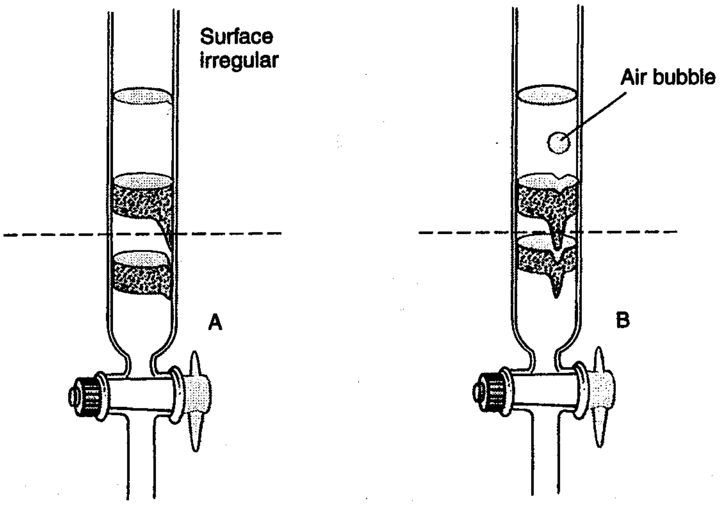

Figure 12.6 Channeling complications.

A. PREPARING THE SUPPORT BASE

Standard-Scale Columns. For large-scale applications, clamp a chromatography column upright (vertically). The column (Fig. 12.3) is a piece of cylindrical glass tubing with a stopcock attached at one end. The stopcock usually has a Teflon plug, because stopcock grease (used on glass plugs) dissolves in many of the organic solvents used as eluents. Stopcock grease in the eluent will contaminate the eluates.

Instead of a stopcock, attach a piece of flexible tubing to the bottom of the column, with a screw clamp used to stop or regulate the flow (Fig. 12.7). When a screw clamp is used, care must be taken that the tubing is not dissolved by the solvents that will pass through the column during the experiment. Rubber, for instance, dissolves in chloroform, benzene, methylene chloride, toluene, or tetrahydrofuran (THF). Tygon tubing dissolves (actually, the plasticizer is removed) in many solvents, including benzene, methylene chloride, chloroform, ether, ethyl acetate, toluene, and THF. Polyethylene tubing is the best choice for use at the end of a column, because it is inert with most solvents.

Next, the column is partially filled with a quantity of solvent, usually a nonpolar solvent like hexane, and a support for the finely divided adsorbent is prepared in the following way. A loose plug of glass wool is tamped down into the bottom of the column with a long glass rod until all entrapped air is forced out as bubbles. Take care not to plug the column totally by tamping the glass wool too hard. A small layer of clean white sand is formed on top of the glass wool by pouring sand into the column. The column is tapped to level the surface of the sand. Any sand adhering to the side of the column is washed down with a small quantity of solvent. The sand forms a base that supports the column

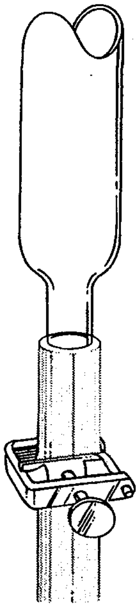

Figure 12.7 Tubing with screw clamp to regulate solvent flow on a chromatography column.

of adsorbent and prevents it from washing through the stopcock. The column is packed in one of two ways: by the slurry method or by the dry pack method.

Semi-Microscale Columns. An alternative apparatus for small-scale column chromatography is a commercial column, such as the one shown in Figure 12.8. This type of column is made of glass and has a solvent-resistant plastic stopcock at the bottom.[1] The stopcock assembly contains a filter disc to support the adsorbent column. An optional upper fitting, also made of solvent-resistant plastic, serves as a solvent reservoir. The column shown in Figure 12.8 is equipped with the solvent reservoir. This type of column is available in a variety of lengths, ranging from 100 to 300 mm. Because the column has a built-in filter disc, it is not necessary to prepare a support base before the adsorbent is added.

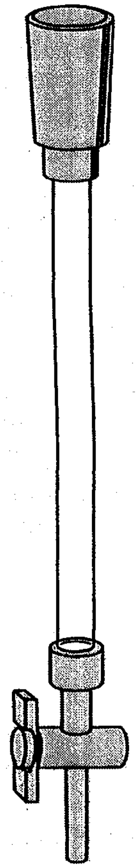

Figure 12.8 Commercial semi-microscale chromatography column. (The column is shown equipped with an optional solvent reservoir.)

---

[1] Note to the instructor: With certain organic solvents, we have found that the "solvent-resistant plastic" stopcock may tend to dissolve! We recommend that instructors test their equipment with the solvents that they intend to use before the beginning of the laboratory class.

B. DEPOSITING THE ADSORBENT

Slurry Method

In the slurry method, the adsorbent is packed into the column as a mixture of a solvent and an undissolved solid. The slurry is prepared in a separate container by adding the solid adsorbent, a little at a time, to a quantity of the solvent. This order of addition (adsorbent added to solvent) should be followed strictly, because the adsorbent solvates and liberates heat. If the solvent is added to the adsorbent, it may boil away almost as fast as it is added due to heat evolved. This will be especially true if ether or another low-boiling solvent is used. When this happens, the final mixture will be uneven and lumpy. Enough adsorbent is added to the solvent, and mixed by swirling the container, to form a thick, but flowing, slurry. The container should be swirled until the mixture is homogenous and relatively free of entrapped air bubbles.

For standard-sized or semi-microscale columns, the procedure is as follows. When the slurry has been prepared, the column is filled about half-full with solvent, and the stopcock is opened to allow solvent to drain slowly into a large beaker. The slurry is mixed by swirling and is then poured in portions into the top of the draining column (a wide-necked funnel may be useful here). Be sure to swirl the slurry thoroughly before each addition to the column. The column is tapped constantly and *gently* on the side, during the pouring operation, with the fingers or with a pencil fitted with a rubber stopper. A short piece of large-diameter pressure tubing may also be used for tapping. The tapping promotes even settling and mixing and gives an evenly packed column free of air bubbles. Tapping is continued until all the material has settled, showing a well-defined level at the top of the column. Solvent from the collecting beaker may be re-added to the slurry if it becomes too thick to be poured into the column at one time. In fact, the collected solvent should be cycled through the column several times to ensure that settling is complete and that the column is firmly packed. The downward flow of solvent tends to compact the adsorbent. You should take care never to let the column run dry during packing.

Dry Pack Method 1

In the first of the dry pack methods introduced here, the column is filled with solvent and allowed to drain *slowly.* The dry adsorbent is added, a little at a time, while the column is tapped gently with a pencil, finger, or glass rod.

Standard-Scale Columns. A plug of cotton is placed at the base of the column, and an even layer of sand is formed on top (see p. 764). The column is filled about half-full with solvent, and the solid adsorbent is added carefully from a beaker, while the solvent is allowed to flow slowly from the column. As the solid is added, the column is tapped as described for the slurry method in order to ensure that the column is packed evenly. When the column has the desired length, no more adsorbent is added. This method also produces an evenly packed column. Solvent should be cycled through this column (for standard scale applications) several times before each use. The same portion of solvent that has drained from the column during the packing is used to cycle through the column.

Semi-Microscale Columns. The procedure to fill a commercial semi-microscale column is essentially the same as that used to fill a Pasteur pipet (Section 12.7). The commercial column has the advantage that it is much easier to control the flow of solvent from the column during the filling process, because the stopcock can be adjusted appropriately. It is not necessary to use a cotton plug or to deposit a layer of sand before adding the adsorbent. The presence of the fritted disc at the base of the column acts to prevent adsorbent from escaping from the column.

Dry Pack Method 2

Standard-Scale Columns. Standard-scale columns can also be packed by a dry pack method that is similar to the microscale methods described in Section 12.7. The disadvantages described for the microscale method also apply to the macroscale method. This method is not recommended for use with silica gel or alumina, because the combination leads to uneven packing, air bubbles, and cracking, especially if a solvent that has a highly exothermic heat of solvation is used.

Semi-Microscale Columns. The Dry Pack Method 2 is similar to that described for Pasteur pipets (Section 12.7), except that the plug of cotton is not required. The flow rate of solvent through the column can be controlled using the stopcock, which is part of the column assembly (see Fig. 12.8).

12.7 PACKING THE COLUMN: MICROSCALE METHODS

As with standard-scale or semi-microscale columns, the procedures described in this section should be followed carefully in preparing a microscale chromatography column. Failure to pay close attention to the details of these procedures may adversely affect the quality of the separation.

Again, preparation of a column involves two distinct stages: preparation of the support base and filling the column with adsorbent.

A. PREPARING THE SUPPORT BASE

For microscale applications, select a Pasteur pipet ($5\frac{3}{4}$-inch) and clamp it upright (vertically). In order to reduce the amount of solvent needed to fill the column, break off most of the tip of the pipet. Place a small ball of cotton in the pipet and tamp it into position using a glass rod or a piece of wire. Take care not to plug the column totally by tamping the cotton too hard. The correct position of the cotton is shown in Figure 12.9. A microscale chromatography column is packed by one of the dry pack methods described in Part B of this section.

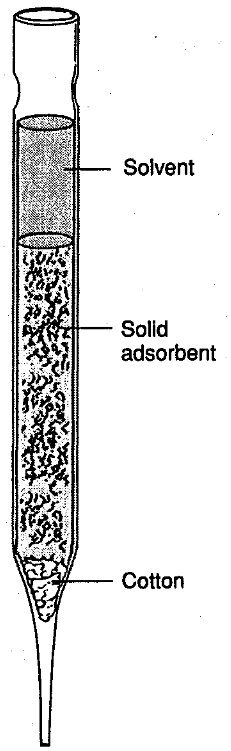

Figure 12.9 Microscale chromatography column.

B. DEPOSITING THE ADSORBENT

The slurry method is not recommended as a microscale method for use with Pasteur pipets. On a very small scale, it is too difficult to pack the column with the slurry without losing the solvent before the packing has been completed. Microscale columns should be packed by the one of the dry pack methods, as described in the following discussion.

Dry Pack Method 1

To fill a microscale column, fill the Pasteur pipet (with the cotton plug, prepared as described in Section A) about half-full with solvent. Using a microspatula, add the solid adsorbent slowly to the solvent in the column. As you add the solid, tap the column *gently* with a pencil, a finger, or a glass rod. The tapping promotes even settling and mixing and gives an evenly packed column free of air bubbles. As the adsorbent is added, solvent flows out of the Pasteur pipet. Because the adsorbent must not be allowed to dry during the packing process, you must use a means of controlling the solvent flow. If a piece of small-diameter plastic tubing is available, it can be fitted over the narrow tip of the Pasteur pipet. The flow rate can then be controlled using a screw clamp. A simple approach to controlling the flow rate is to use a finger over the top of the Pasteur pipet, much as you control the flow of liquid in a volumetric pipet. Continue adding the adsorbent slowly, with constant tapping, until the level of the adsorbent has reached the desired level. As you pack the column, be careful not to let the column run dry. The final column should appear as shown in Figure 12.9.

Dry Pack Method 2

An alternative dry pack method for microscale columns is to fill the Pasteur pipet with *dry* adsorbent, without any solvent. A plug of cotton is positioned in the bottom of the Pasteur pipet. The desired amount of adsorbent is added slowly, and the pipet tapped constantly, until the level of adsorbent has reached the desired height. Figure 12.9 can be used as a guide to judge the correct height of the column of adsorbent. When the column is packed, added solvent is allowed to percolate through the adsorbent until the entire column is moistened. The solvent is not added until just before the column is to be used.

This method is not recommended for use with silica gel nor for experiments where a very careful separation is required.

This method is useful when the adsorbent is alumina, but it does not produce satisfactory results with silica gel. Even with alumina, poor separations can arise due to uneven packing, air bubbles, and cracking, especially if a solvent that has a highly exothermic heat of solvation is used.

12.8 APPLYING THE SAMPLE TO THE COLUMN

The solvent (or solvent mixture) used to pack the column is normally the least polar elution solvent that can be used during chromatography. The compounds to be chromatographed are not highly soluble in the solvent. If they were, they would probably have a greater affinity for the solvent than for the adsorbent and would pass right through the column without equilibrating with the stationary phase.

The first elution solvent, however, is generally not a good solvent to use in preparing the sample to be placed on the column. Because the compounds are not highly soluble in nonpolar solvents, it takes a large amount of the initial solvent to dissolve the compounds, and it is difficult to get the mixture to form a narrow band on top of the column. A narrow band is ideal for an optimum separation of components. For the best separation, therefore, the compound is applied to the top of the column undiluted if it is a liquid, or in a *very small* amount of highly polar solvent if it is a solid. Water must not be used to dissolve the initial sample being chromatographed, because it reacts with the column packing.

In adding the sample to the column, use the following procedure. Lower the solvent level to the top of the adsorbent column by draining the solvent from the column. Add the sample (either a pure liquid or a solution) to form a small layer on top of the adsorbent. A Pasteur pipet is convenient for adding the sample to the column. Take care not to disturb the surface of the adsorbent. This is best accomplished by touching the pipet to the inside of the glass column and slowly draining it so as to allow the sample to spread into a thin film, which slowly descends to cover the entire adsorbent surface. Drain the pipet close to the surface of the adsorbent. When all the sample has been added, drain this small layer of liquid into the column until the top surface of the column *just begins* to dry. Then add a small layer of the chromatographic solvent carefully with a Pasteur pipet, again being careful not to disturb the surface. Drain this small layer of solvent into the column until the top surface of the column just dries. Add another small layer of fresh solvent, if necessary, and repeat the process until it is clear that the sample is strongly adsorbed on the top of the column. If the sample is colored and the fresh layer of solvent acquires some of this color, the sample has not been properly adsorbed. Once the sample has been properly applied, you can protect the level surface of the adsorbent by carefully filling the top of the column with solvent and sprinkling clean, white sand into the column so as to form a small protective layer on top of the adsorbent. For microscale applications, this layer of sand is not required.

Separations are often better if the sample is allowed to stand a short time on the column before elution. This allows a true equilibrium to be established. In columns that stand for too long, however, the adsorbent often compacts or even swells, and the flow can become annoyingly slow. Diffusion of the sample to widen the bands also becomes a problem if a column is allowed to stand over an extended period. For microscale chromatography, using Pasteur pipets, there is no stopcock, and it is not possible to stop the flow. In this case, it is not considered necessary to allow the column to stand.

12.9 ELUTION TECHNIQUES

Solvents for analytical and preparative chromatography should be pure reagents. Commercial-grade solvents often contain small amounts of residue, which remains when the solvent is evaporated. For normal work, and for relatively easy separations that take only small amounts of solvent, the residue usually presents few problems. For large-scale work, commercial-grade solvents may have to be redistilled before use. This is especially true for hydrocarbon solvents, which tend to have more residue than other solvent types. Most of the experiments in this laboratory manual have been designed to avoid this particular problem.

One usually begins elution of the products with a nonpolar solvent, like hexane or petroleum ether. The polarity of the elution solvent can be increased gradually by adding successively greater percentages of either ether or toluene (for instance, 1, 2, 5, 10, 15, 25, 50, 100%) or some other solvent of greater solvent power (polarity) than hexane. The transition from one solvent to another should not be too rapid in most solvent changes. If the two solvents to be changed differ greatly in their heats of solvation in binding to the adsorbent, enough heat can be generated to crack the column. Ether is especially troublesome in this respect, as it has both a low boiling point and a relatively high heat of solvation. Most organic compounds can be separated on silica gel or alumina using hexane–ether or hexane–toluene combinations for elution, and following these by pure methylene chloride. Solvents of greater polarity are usually avoided for the various reasons mentioned previously. In microscale work, the usual procedure is to use only one solvent for the chromatography.

The flow of solvent through the column should not be too rapid, or the solutes will not have time to equilibrate with the adsorbent as they pass down the column. If the rate of flow is too low, or stopped for a period, diffusion can become a problem—the solute band will diffuse, or spread out, in all directions. In either of these cases, separation will be poor. As a general rule (and only an approximate one), most standard scale columns are run with flow rates ranging from 5 to 50 drops of effluent per minute; a steady flow of solvent is usually avoided. Microscale columns made from Pasteur pipets do not have a means of controlling the solvent flow rate, but commercial semimicroscale columns are equipped with stopcocks. The solvent flow rate in this type of column can be adjusted in a manner similar to that used with larger columns. To avoid diffusion of the bands, do not stop the column and do not set it aside overnight.

12.10 RESERVOIRS

When large quantities of solvent are used in a chromatographic separation, it is often convenient to use a solvent reservoir to forestall having to add small portions of fresh solvent continually. The simplest type of reservoir, a feature of many columns, is created by fusing the top of the column to a round-bottom flask (Figure 12.10A). If the column has a standard-taper joint at its top, a reservoir can be created by joining a standard-taper separatory funnel to the column (Figure 12.10B). In this arrangement, the stopcock is left open, and no stopper is placed in the top of the separatory funnel. A third common arrangement is shown in Figure 12.10C. A separatory funnel is filled with solvent; its stopper is wetted with solvent and put *firmly* in place. The funnel is inserted into the empty filling space at the top of the chromatographic column, and the stopcock is opened. Solvent flows out of the funnel, filling the space at the top of the column until the solvent level is well above the outlet of the separatory funnel. As solvent drains from the column, this arrangement automatically refills the space at the top of the column by allowing air to enter through the stem of the separatory funnel. Some semi-microscale columns, such as that shown in Figure 12.8, are equipped with a solvent reservoir that fits onto the top of the column. It functions just like the reservoirs described in this section.

For a microscale chromatography, the portion of the Pasteur pipet above the adsorbent is used as a reservoir of solvent. Fresh solvent, as needed, is added by means of another Pasteur pipet. When it is necessary to change solvent, the new solvent is also added in this manner. In some cases, the chromatography may proceed too slowly; the rate of solvent flow can be accelerated by attaching a rubber dropper bulb to the top of the Pasteur pipet column and squeezing *gently*. The additional air pressure forces the solvent through the column more rapidly. If this technique is used, however, care must be taken to remove the rubber bulb from the column before releasing it. Otherwise, air may be drawn up through the bottom of the column, destroying the column packing.

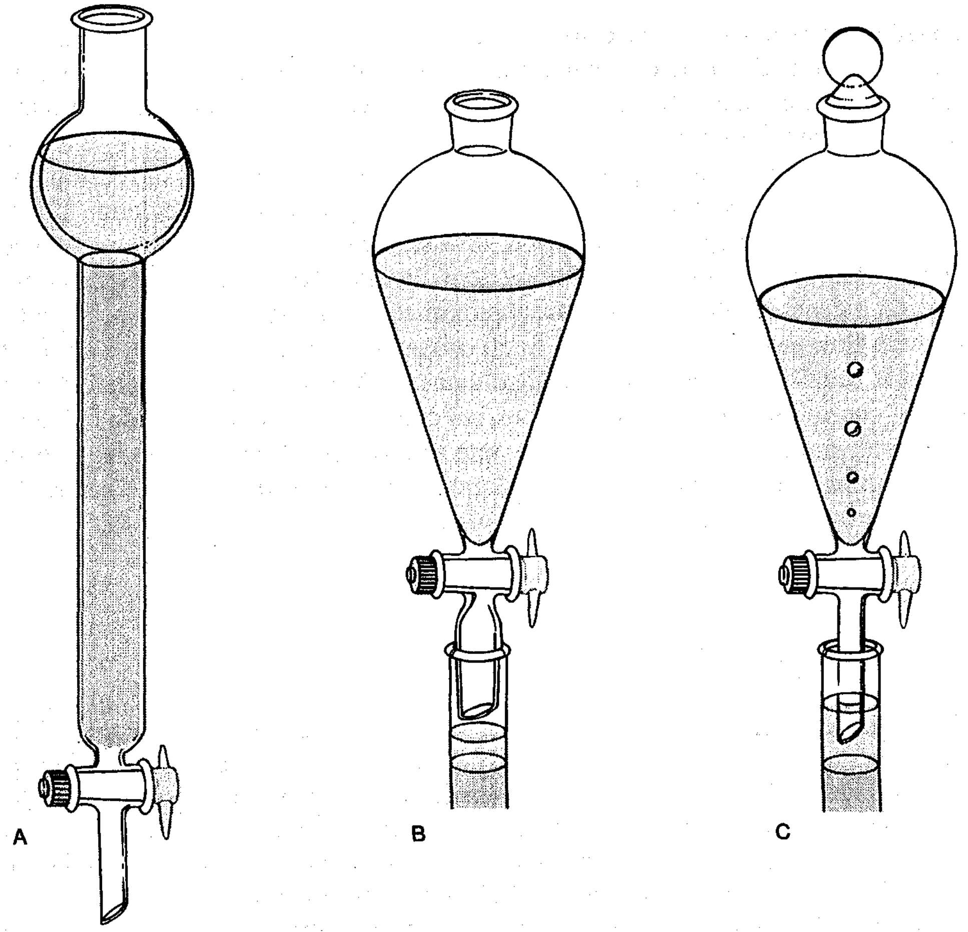

Figure 12.10 Various types of solvent-reservoir arrangements for chromatographic columns.

12.11 MONITORING THE COLUMN

It is a happy instance when the compounds to be separated are colored. The separation can then be followed visually and the various bands collected separately as they elute from the column. For the majority of organic compounds, however, this lucky circumstance does not exist, and other methods must be used to determine the positions of the bands. The most common method of following a separation of colorless compounds is to collect **fractions** of constant volume in pre-weighed flasks, to evaporate the solvent from each fraction, and to reweigh the flask plus any residue. A plot of fraction number *versus* the weight of the residues after evaporation of solvent gives a plot like that in Figure 12.11. Clearly, fractions 2 through 7 (Peak 1) may be combined as a single compound, and so can fractions 8 through 11 (Peak 2) and 12 through 15 (Peak 3). The size of the fractions collected (1, 10, 100, or 500 mL) depends on the size of the column and the ease of separation.

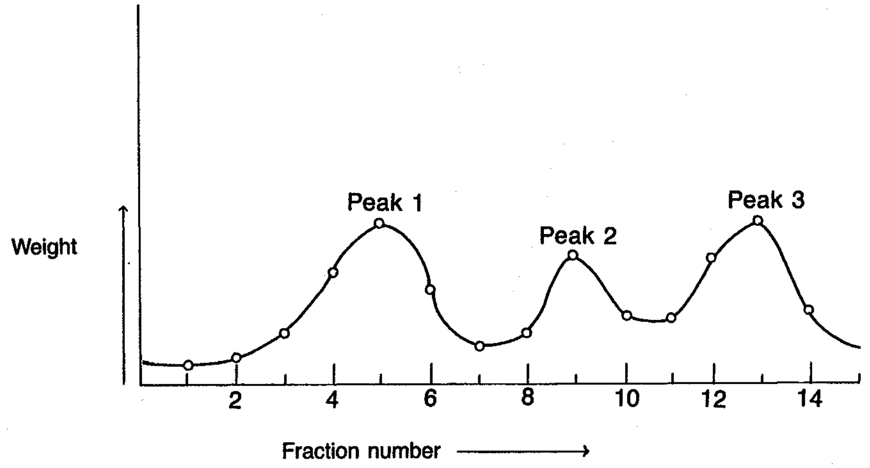

Figure 12.11 Typical elution graph.

Another common method of monitoring the column is to mix an inorganic phosphor into the adsorbent used to pack the column. When the column is illuminated with an ultraviolet light, the adsorbent treated in this way fluoresces. However, many solutes have the ability to **quench** the fluorescence of the indicator phosphor. In areas in which solutes are present, the adsorbent does not fluoresce, and a dark band is visible. In this type of column, the separation can also be followed visually.

Thin-layer chromatography is often used to monitor a column. This method is described in Technique 14 (Section 14.10, p. 804). Several sophisticated instrumental and spectroscopic methods, which we shall not detail, can also monitor a chromatographic separation.

12.12 TAILING

Often, when a single solvent is used for elution, an elution curve (weight versus fraction) like that shown as a solid line in Figure 12.12 is observed. An ideal elution curve is shown by dashed lines. In the nonideal curve, the compound is said to be **tailing.** Tailing can interfere with the beginning of a curve or a peak of a second component and lead to a poor separation. One way to avoid this is to increase the polarity of the solvent constantly while eluting. In this way, at the tail of the peak, where the solvent polarity is increasing, the compound will move slightly faster than at the front and allow the tail to squeeze forward, forming a more nearly ideal band.

12.13 RECOVERING THE SEPARATED COMPOUNDS

In recovering each of the separated compounds of a chromatographic separation when they are solids, the various correct fractions are combined and evaporated. If the combined fractions contain sufficient material, they may be purified by recrystallization. If the

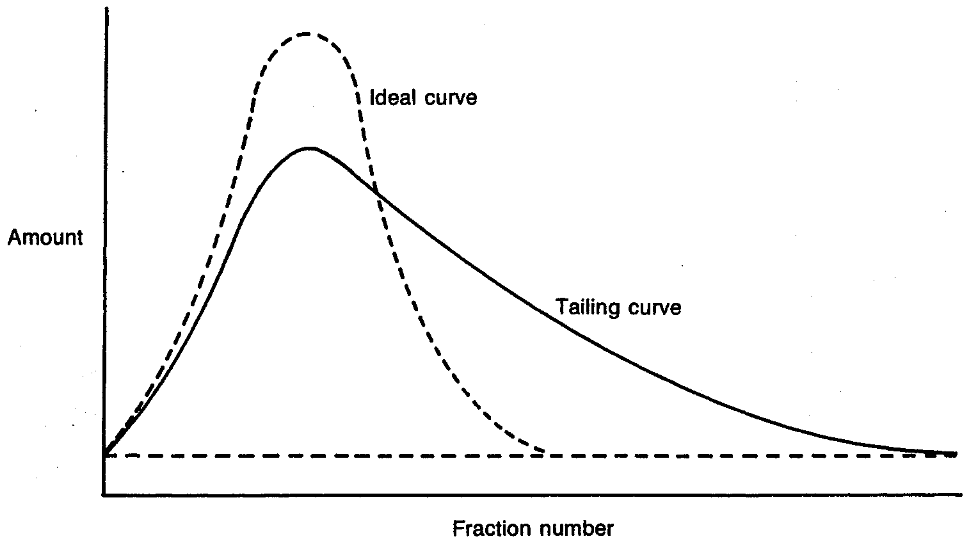

Figure 12.12 Elution curves: One ideal and one that "tails."

compounds are liquids, the correct fractions are combined, and the solvent is evaporated. If sufficient material has been collected, liquid samples can be purified by distillation. The combination of chromatography–crystallization or chromatography–distillation usually yields very pure compounds. For microscale applications, the amount of sample collected is too small to allow a purification by crystallization or distillation. The samples that are obtained after the solvent has been evaporated are considered to be sufficiently pure, and no additional purification is attempted.

12.14 DECOLORIZATION BY COLUMN CHROMATOGRAPHY

A common outcome of organic reactions is the formation of a product that is contaminated by highly colored impurities. Very often these impurities are highly polar, and they have a high molecular weight, as well as being colored. The purification of the desired product requires that these impurities be removed. Section 5.6 of Technique 5 (pp. 659–661) details methods of decolorizing an organic product. In most cases, these methods involve the use of a form of activated charcoal, or Norit.

An alternative, which is applied conveniently in microscale experiments, is to remove the colored impurity by column chromatography. Because of the polarity of the impurities, the colored components are strongly adsorbed on the stationary phase of the column, and the less polar desired product passes through the column and is collected.

Microscale decolorization of a solution on a chromatography column requires that a column be prepared in a Pasteur pipet, using either alumina or silica gel as the adsorbent (Section 12.6). The sample to be decolorized is diluted to the point where crystallization within the column will not take place, and it is then passed through the column in the usual manner. The desired compound is collected as it exits the column, and the excess solvent is removed by evaporation (Technique 3, Section 3.11, p. 630).

12.15 GEL CHROMATOGRAPHY

The stationary phase in gel chromatography consists of a cross-linked polymeric material. Molecules are separated according to their *size* by their ability to penetrate a sieve-like structure. Molecules permeate the porous stationary phase as they move down the column. Small molecules penetrate the porous structure more easily than large ones. Thus, the large molecules move through the column faster than the smaller ones and elute first. The separation of molecules by gel chromatography is depicted in Figure 12.13. With adsorption chromatography using materials such as alumina or silica, the order is usually the reverse. Small molecules (of low molecular weight) pass through the column *faster* than large molecules (of high molecular weight) because large molecules are more strongly attracted to the polar stationary phase.

Equivalent terms used by chemists for the gel-chromatography technique are **gel filtration** (biochemistry term), **gel-permeation chromatography** (polymer chemistry term),

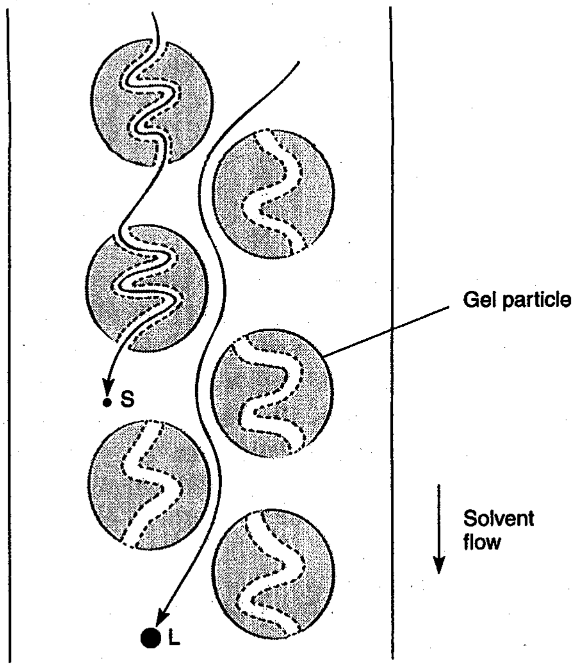

Figure 12.13 Gel chromatography: Comparison of the paths of large (L) and small (S) molecules through the column during the same interval of time.

and **molecular sieve chromatography. Size-exclusion chromatography** is a general term for the technique, and it is perhaps the most descriptive term for what occurs on a molecular level.

Sephadex is one of the most popular materials for gel chromatography. It is widely used by biochemists for separating proteins, nucleic acids, enzymes, and carbohydrates. Most often, water or aqueous solutions of buffers are used as the moving phase. Chemically, Sephadex is a polymeric carbohydrate that has been cross-linked. The degree of cross-linking determines the size of the "holes" in the polymer matrix. In addition, the hydroxyl groups on the polymer can adsorb water, which causes the material to swell. As it expands, "holes" are created in the matrix. Several different gels are available from manufacturers, each with its own set of characteristics. For example, a typical Sephadex gel, such as G-75, can separate molecules in the molecular-weight (MW) range 3000 to 70,000. Assume for the moment that one has a four-component mixture containing compounds with molecular weights of 10,000, 20,000, 50,000, and 100,000. The 100,000-MW compound would pass through the column first, because it cannot penetrate the polymer matrix. The 50,000-, 20,000-, and 10,000-MW compounds penetrate the matrix to varying degrees and would be separated. The molecules would elute in the order given (decreasing order of molecular weights). The gel separates on the basis of molecular size and configuration, rather than molecular weight.

Sephadex LH-20 has been developed for nonaqueous solvents. Some of the hydroxyl groups have been alkylated, and thus the material can swell under both aqueous and nonaqueous conditions (it now has "organic" character). This material can be used with several organic solvents, such as alcohol, acetone, methylene chloride, and aromatic hydrocarbons.

Another type of gel is based on a polyacrylamide structure (Bio-Gel P and Poly-Sep AA). A portion of a polyacrylamide chain is shown below.

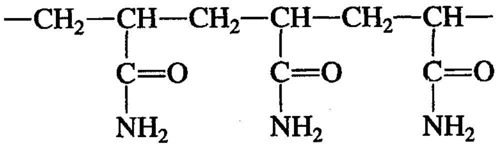

$$-CH_2-\underset{\underset{NH_2}{|}}{\underset{C=O}{\underset{|}{CH}}}-CH_2-\underset{\underset{NH_2}{|}}{\underset{C=O}{\underset{|}{CH}}}-CH_2-\underset{\underset{NH_2}{|}}{\underset{C=O}{\underset{|}{CH}}}-$$

Gels of this type can also be used in water and some polar organic solvents. They tend to be more stable than Sephadex, especially under acidic conditions. Polyacrylamides can be used for many biochemical applications involving macromolecules. For separating synthetic polymers, cross-linked polystyrene beads (copolymer of styrene and divinyl benzene) find common application. Again, the beads are swollen before use. Common organic solvents can be used to elute the polymers. As with other gels, the higher-molecular-weight compounds elute before the lower-molecular-weight compounds.

12.16 FLASH CHROMATOGRAPHY

One of the drawbacks to column chromatography is that, for large-scale preparative separations, the time required to complete a separation may be very long. Furthermore, the resolution that is possible for a particular experiment tends to deteriorate as the time for the experiment grows longer. This latter effect arises because the bands of compounds that move very slowly through a column tend to "tail."

A technique that can be useful in overcoming these problems has been developed. This technique, called **flash chromatography,** is actually a very simple modification of an ordinary column chromatography. In flash chromatography, the adsorbent is packed into a relatively short glass column, and air pressure is used to force the solvent through the adsorbent.

The apparatus used for flash chromatography is shown in Figure 12.14. The glass column is fitted with a Teflon stopcock at the bottom to control the flow rate of solvent. A plug of glass wool is placed in the bottom of the column to act as a support for the adsorbent. A layer of sand may also be added on top of the glass wool. The column is filled with adsorbent using the dry pack method. When the column has been filled, a fitting is attached to the top of the column, and the entire apparatus is connected to a source of high-pressure air or nitrogen. The fitting is designed so that the pressure applied to the top of the column can be adjusted precisely. The source of the high-pressure air is often a specially adapted air pump.

A typical column would use silica gel adsorbent (particle size = 40 to 63 μm) packed to a height of 5 inches in a glass column of 20-mm diameter. The pressure applied to the column would be adjusted to achieve a solvent flow rate such that the solvent level in the column would decrease by about 2 in/min. This system would be appropriate to separate the components of a 250-mg sample.

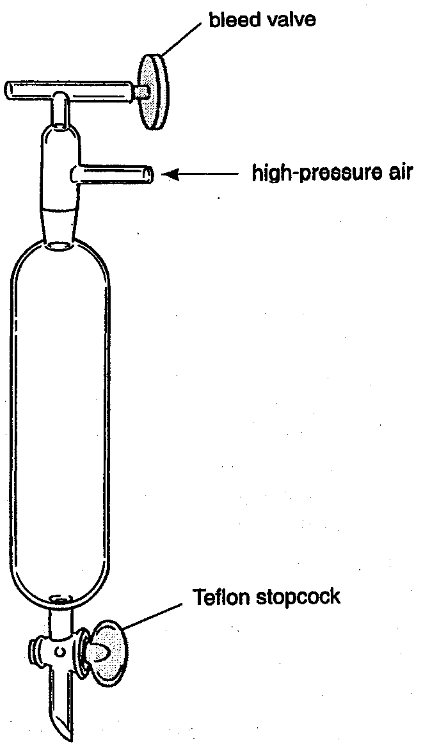

Figure 12.14 Apparatus for flash chromatography.

The high-pressure air forces the solvent through the column of adsorbent at a rate that is much greater than what would be achieved if the solvent flowed through the column under the force of gravity. Because the solvent is caused to flow faster, the time required for substances to pass through the column is reduced. By itself, simply applying air pressure to the column might reduce the clarity of the separation, because the components of the mixture would not have time to establish themselves into distinctly separate bands. However, in flash chromatography, you can use a much finer adsorbent than would be used in ordinary chromatography. With a much smaller particle size for the adsorbent, the surface area is increased, and the resolution possible thereby improves.

A simple variation on this idea does not use air pressure. Instead, the lower end of the column is inserted into a stopper, which is fitted into the top of a suction flask. Vacuum is applied to the system, and the vacuum acts to draw the solvent through the adsorbent column. The overall effect of this variation is similar to that obtained when air pressure is applied to the top of the column.

REFERENCES

Deyl, Z., Macek, K., and Janák, J. *Liquid Column Chromatography.* Amsterdam: Elsevier, 1975.

Heftmann, E. *Chromatography.* 3rd ed. New York: Van Nostrand Reinhold, 1975.

Jacobson, B.M. "An Inexpensive Way to Do Flash Chromatography." *Journal of Chemical Education, 65* (May 1988): 459.

Still, W. C., Kahn, M., and Mitra, A. "Rapid Chromatographic Technique for Preparative Separations with Moderate Resolution." *Journal of Organic Chemistry, 43* (1978): 2923.

PROBLEMS

1. A sample was placed on a chromatography column. Methylene chloride was used as the eluting solvent. All of the components eluted off the column, but no separation was observed. What must have been happening during this experiment? How would you change the experiment in order to overcome this problem?

2. You are about to purify an impure sample of naphthalene by column chromatography. What solvent should you use to elute the sample?

3. Consider a sample that is a mixture composed of biphenyl, benzoic acid, and benzyl alcohol. Predict the order of elution of the components in this mixture. Assume that the chromatography uses a silica column, and the solvent system is based on cyclohexane, with an increasing proportion of methylene chloride being added as a function of time.

4. An orange compound was added to the top of a chromatography column. Solvent was added immediately, with the result that the entire volume of solvent in the solvent reservoir turned orange. No separation could be obtained from the chromatography experiment. What went wrong?

5. A yellow compound, dissolved in methylene chloride, is added to a chromatography column. The elution is begun using petroleum ether as the solvent. After 6 L of solvent had passed through the column, the yellow band still had not traveled down the column appreciably. What should be done to make this experiment work better?

6. You have 0.50 g of a mixture that you wish to purify by column chromatography. How much adsorbent should you use to pack the column? Estimate the appropriate column diameter and height.

7. In a particular sample, you wish to collect the component with the *highest* molecular weight as the *first* fraction. What chromatographic technique should you use?

8. A colored band shows an excessive amount of tailing as it passes through the column. What can you do to rectify this problem?

9. How would you monitor the progress of a column chromatography when the sample is colorless? Describe at least two methods.

TECHNIQUE 13

High-Performance Liquid Chromatography (HPLC)

The separation that can be achieved is greater if the column packing used in column chromatography is made more dense by using an adsorbent that has a small particle size. The solute molecules encounter a much larger surface area on which they can be adsorbed as they pass through the column packing. At the same time, the solvent spaces between the particles are reduced in size. As a result of this tight packing, equilibrium between the liquid and solid phases can be established very rapidly with a fairly short column, and the degree of separation is markedly improved. The disadvantage of making the column packing more dense is that the solvent flow rate becomes very slow or even stops. Gravity is not strong enough to pull the solvent through a tightly packed column.

A modern technique can be applied to obtain much better separations with tightly packed columns. A pump forces the solvent through the column packing. As a result, solvent flow rate is increased and the advantage of better separation is retained. This technique, called **high-performance liquid chromatography (HPLC),** is becoming widely applied to problems where separations by ordinary column chromatography are unsatisfactory. Because the pump often provides pressures in excess of 1000 pounds per square inch (psi), this method is also known as **high-pressure liquid chromatography.** High pressures are not required, however, and satisfactory separations can be achieved with pressures as low as 100 psi.

The basic design of an HPLC instrument is shown in Figure 13.1. The instrument contains the following essential components:

1. Solvent reservoir
2. Solvent filter and degasser
3. Pump
4. Pressure gauge
5. Sample injection system
6. Column
7. Detector
8. Amplifier and electronic controls
9. Chart recorder

There may be other variations on this simple design. Some instruments have heated ovens in order to maintain the column at a specified temperature, fraction collectors, and microprocessor-controlled data-handling systems. Additional filters for the solvent and

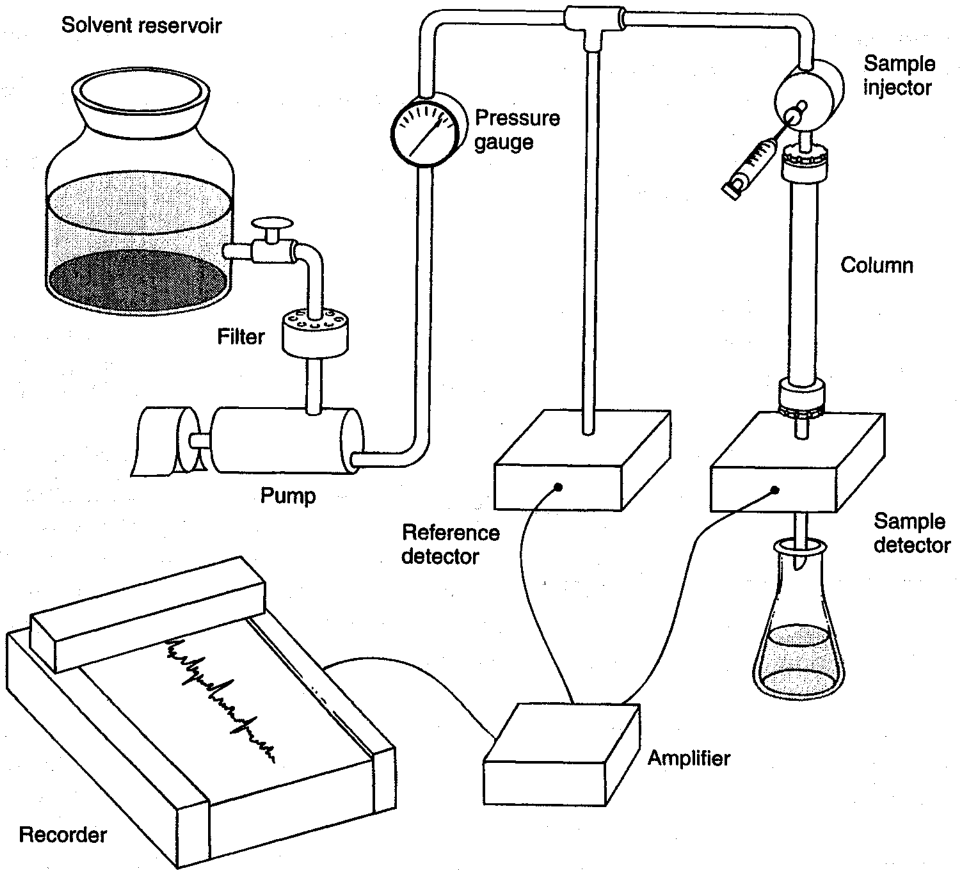
Solvent reservoir
Pressure gauge
Sample injector
Column
Filter
Pump
Reference detector
Sample detector
Amplifier
Recorder

Figure 13.1 *Schematic diagram of a high-performance liquid chromatography (HPLC).*

sample may also be included. You may find it interesting to compare this schematic diagram with that shown in Technique 15 (Fig. 15.2, p. 809) for a gas chromatography instrument. Many of the essential components are common to both types of instruments.

13.1 ADSORBENTS AND COLUMNS

The most important factor to consider when choosing a set of experimental conditions is the nature of the material packed into the column. You must also consider the size of the column that will be selected. The chromatography column is generally packed with silica or alumina adsorbents. Unlike column chromatography, however, the adsorbents used for HPLC have a much smaller particle size. Typically, particle size ranges from 5 to 20 μm in diameter for HPLC; it is on the order of 100 μm for column chromatography.

The adsorbent is packed into a column that can withstand the elevated pressures typical of this type of experiment. Generally, the column is constructed of stainless steel, although some columns that are constructed of a rigid polymeric material ("PEEK"—Poly Ether Ether Ketone) are available commercially. A strong column is required to withstand the high pressures that may be used. The columns are fitted with stainless steel connectors, which ensure a pressure-tight fit between the column and the tubing that connects the column to the other components of the instrument.

Columns that fulfill a large number of specialized purposes are available. In this chapter, we consider only the four most important types of columns:

1. Normal-phase chromatography
2. Reversed-phase chromatography
3. Ion-exchange chromatography
4. Size exclusion chromatography

In most types of chromatography, the adsorbent is more polar than the mobile phase. For example, the solid packing material, which may be either silica or alumina, has a stronger affinity for polar molecules than does the solvent. As a result, the molecules in the sample adhere strongly to the solid phase, and their progress down the column is much slower than the rate at which solvent moves through the column. The time required for a substance to move through the column can be altered by changing the polarity of the solvent. In general, as the solvent becomes more polar, substances move faster through the column. This type of behavior is known as **normal-phase chromatography.** In HPLC, you inject a sample onto a normal-phase column and elute it by varying the polarity of the solvent, much as you do with ordinary column chromatography. Disadvantages of normal-phase chromatography are that retention times tend to be long, and bands have a tendency to "tail."

These disadvantages can be ameliorated by selecting a column in which the solid support is *less polar* than the moving solvent phase. In this type of chromatography, known as **reversed-phase chromatography,** the silica column packing is treated with alkylating agents. As a result, nonpolar alkyl groups are bonded to the silica surface, making the adsorbent nonpolar. The alkylating agents that are used most commonly can attach methyl ($—CH_3$), octyl ($—C_8H_{17}$), or octadecyl ($—C_{18}H_{37}$) groups to the silica surface. The latter variation, where an 18-carbon chain is attached to the silica, is the most popular. This type of column is known as a **C-18 column.** The bonded alkyl groups have an effect similar to that which would be produced by an extremely thin organic solvent layer coating the surface of the silica particles. The interactions that take place between the substances dissolved in the solvent and the stationary phase thus become more like those observed in a liquid–liquid extraction. The solute particles distribute themselves between the two "solvents"—that is, between the moving solvent and the organic coating on the silica. The longer the alkyl groups that are bonded to the silica, the more effective the alkyl groups are as they interact with solute molecules.

Reversed-phase chromatography is widely used because the rate at which solute molecules exchange between moving phase and stationary phase is very rapid, which means that substances pass through the column relatively quickly. Furthermore, problems arising from the "tailing" of peaks are reduced. A disadvantage of this type of column, however, is that the chemically bonded solid phases tend to decompose. The organic groups are slowly hydrolyzed from the surface of the silica, which leaves a normal silica surface exposed. Thus, the chromatographic process that takes place on the column slowly shifts from a reversed-phase to a normal-phase separation mechanism.

Another type of solid support that is sometimes used in reversed-phase chromatography is organic polymer beads. These beads present a surface to the moving phase that is largely organic in nature.

For solutions of ions, a column that is packed with an ion-exchange resin is used. This type of chromatography is known as **ion-exchange chromatography.** The ion-exchange resin that is chosen can be either an anion-exchange resin or a cation-exchange resin, depending upon the nature of the sample being examined.

A fourth type of column is known as a **size exclusion column** or a **gel filtration column.** The interaction that takes place on this type of column is similar to that described in Technique 12, Section 12.15, page 783.

13.2 COLUMN DIMENSIONS

The dimensions of the column that you use depend upon the application. For analytical applications, a typical column is constructed of tubing that has an inside diameter of between 4 and 5 mm, although analytical columns with inside diameters of 1 or 2 mm are also available. A typical analytical column has a length of about 7.5 to 30 cm. This type of column is suitable for the separation of a 0.1- to 5-mg sample. With columns of smaller diameter, it is possible to perform an analysis with samples smaller than 1 *micro*gram.

High-performance liquid chromatography is an excellent analytical technique, but the separated compounds may also be isolated. The technique can be used for preparative experiments. Just as in column chromatography, the fractions can be collected into individual receiving containers as they pass through the column. The solvents can be evaporated from these fractions, allowing you to isolate separated components of the original mixture. Samples that range in size from 5 to 100 mg can be separated on a **semiprep column.** The dimensions of a semiprep column are typically 8 mm inside diameter and 10 cm in length. A semiprep column is a practical choice when you wish to use the same column for both analytical and preparative separations. A semiprep column is small enough to provide reasonable sensitivity in analyses, but it is also capable of handling moderate-size samples when you need to isolate the components of a mixture. Even larger samples can be separated using a **preparative column.** This type of column is useful when you wish to collect the components of a mixture and then use the pure samples for additional study (e.g., for a subsequent chemical reaction or for spectroscopic analysis). A preparative column may be as large as 20 mm in inside diameter and 30 cm in length. A preparative column can handle samples as large as 1 g per injection.

13.3 SOLVENTS

The choice of solvent used for an HPLC run depends on the type of chromatographic separation selected. For a normal-phase separation, the solvent is selected based on its polarity. The criteria described in Technique 12, Section 12.4B, page 767, are used. A solvent of very low polarity might be pentane, petroleum ether, hexane, or carbon tetrachloride; a solvent of very high polarity might be water, acetic acid, methanol, or 1-propanol.

For a reversed-phase experiment, a less polar solvent causes solutes to migrate *faster.* For example, for a mixed methanol–water solvent, as the percentage of methanol in the solvent increases (solvent becomes less polar), the time required to elute the components of a mixture from a column decreases. The behavior of solvents as eluents in a reversed-phase chromatography would be the reverse of the order shown in Table 12.2 on page 768.

If a single solvent (or solvent mixture) is used for the entire separation, the chromatogram is said to be **isochratic.** Special electronic devices are available with HPLC instruments that allow you to program changes in the solvent composition from the beginning to the end of the chromatography. These are called **gradient elution systems.** With gradient elution, the time required for a separation may be shortened considerably.

The need for pure solvents is especially acute with HPLC. The narrow bore of the column and the very small particle size of the column packing require that solvents be particularly pure and free of insoluble residue. In most cases, the solvents must be filtered through ultrafine filters and **degassed** (have dissolved gases removed) before they can be used.

The solvent gradient is chosen so that the eluting power of the solvent increases over the duration of the experiment. The result is that components of the mixture that tend to move very slowly through the column are caused to move faster as the eluting power of the solvent gradually increases. The instrument can be programmed to change the composition of the solvent following a linear gradient or a nonlinear gradient, depending upon the specific requirements of the separation.

13.4 DETECTORS

A flow-through **detector** must be provided to determine when a substance has passed through the column. In most applications, the detector detects either the change in index of refraction of the liquid as its composition changes or the presence of solute by its absorption of ultraviolet or visible light. The signal generated by the detector is amplified and treated electronically in a manner similar to that found in gas chromatography (Technique 15, Section 15.6, p. 814).

A detector that responds to changes in the index of refraction of the solution may be considered the most universal of the HPLC detectors. The refractive index of the liquid passing through the detector changes slightly, but significantly, as the liquid changes from pure solvent to a liquid where the solvent contains some type of organic solute. This change in refractive index can be detected and compared to the refractive index of pure solvent. The difference in index values is then recorded as a peak on a chart. A disadvantage of this type of detector is that it must respond to very small changes in refractive index. As a result, the detector tends to be unstable and difficult to balance.

When the components of the mixture have some type of absorption in the ultraviolet or visible regions of the spectrum, a detector that is adjusted to detect absorption at a particular wavelength of light can be used. This type of detector is much more stable, and the readings tend to be more reliable. Unfortunately, many organic compounds do not absorb ultraviolet light, and this type of detector cannot be used.

REFERENCES

Karger, Barry L. "HPLC: Early and Recent Perspectives." *Journal of Chemical Education, 74* (January 1997): 45.

Meyer, Veronika R. *Practical High-performance Liquid Chromatography.* Chichester: John Wiley & Sons, Ltd., 1988.

Rubinson, K. A. "Liquid Chromatography." Chapter 14 in *Chemical Analysis.* Boston: Little, Brown and Co., 1987.

Snyder, Lloyd R. "Modern Practice of Liquid Chromatography." *Journal of Chemical Education, 74* (January 1997): 37.

Van Arman, Scott A., and Thomsen, Marcus W. "HPLC for Undergraduate Introductory Laboratories." *Journal of Chemical Education, 74* (January 1997): 49.

PROBLEMS

1. For a mixture of biphenyl, benzoic acid, and benzyl alcohol, predict the order of elution and describe any differences that you would expect for a normal-phase HPLC experiment (in hexane solvent) compared with a reversed-phase experiment (in tetrahydrofuran–water solvent).

2. How would the *gradient elution program* differ between normal-phase and reversed-phase chromatography?

TECHNIQUE 14

Thin-Layer Chromatography

Thin-layer chromatography (TLC) is a very important technique for the rapid separation and qualitative analysis of small amounts of material. It is ideally suited for the analysis of mixtures and reaction products in small-scale and microscale experiments. The technique is closely related to column chromatography. In fact, TLC can be considered simply column chromatography *in reverse,* with the solvent ascending the adsorbent, rather than descending. Because of this close relationship to column chromatography, and because the principles governing the two techniques are similar, Technique 12, on column chromatography, should be read before reading this one.

14.1 PRINCIPLES OF THIN-LAYER CHROMATOGRAPHY

Like column chromatography, TLC is a solid–liquid partitioning technique. However, the moving liquid phase is not allowed to percolate down the adsorbent; it is caused to *ascend* a thin layer of adsorbent coated onto a backing support. The most typical backing is a glass plate, but other materials are also used. A thin layer of the adsorbent is spread onto the plate and allowed to dry. A coated and dried plate of glass is called a **thin-layer plate** or a **thin-layer slide.** (The reference to *slide* comes about because microscope slides are often used to prepare small thin-layer plates.) When a thin-layer plate is placed upright in a vessel that contains a shallow layer of solvent, the solvent ascends the layer of adsorbent on the plate by capillary action.

In TLC, the sample is applied to the plate before the solvent is allowed to ascend the adsorbent layer. The sample is usually applied as a small spot near the base of the plate; this technique is often referred to as **spotting.** The plate is spotted by repeated applications of a sample solution from a small capillary pipet. When the filled pipet touches the plate, capillary action delivers its contents to the plate, and a small spot is formed.

As the solvent ascends the plate, the sample is partitioned between the moving liquid phase and the stationary solid phase. During this process, you are **developing,** or **running,** the thin-layer plate. In development, the various components in the applied mixture are separated. The separation is based on the many equilibrations the solutes experience between the moving and the stationary phases. (The nature of these equilibrations was thoroughly discussed in Technique 12, Sections 12.2 and 12.3, pp. 762–765.) As in column chromatography, the least polar substances advance faster than the most polar substances. A separation results from the differences in the rates at which the individual components of the mixture advance upward on the plate. When many substances are present in a mixture, each has its own characteristic solubility and adsorptivity properties, depending on the functional groups in its structure. In general, the stationary phase is strongly polar and strongly binds polar substances. The moving liquid phase is usually less polar than the adsorbent and most easily dissolves substances that are less polar or even nonpolar. Thus, substances that are the most polar travel slowly upward, or not at all, and nonpolar substances travel more rapidly if the solvent is sufficiently nonpolar.

When the thin-layer plate has been developed, it is removed from the developing tank and allowed to dry until it is free of solvent. If the mixture that was originally spotted on the plate was separated, there will be a vertical series of spots on the plate. Each spot corresponds to a separate component or compound from the original mixture. If the components of the mixture are colored substances, the various spots will be clearly visible after development. More often, however, the "spots" will not be visible because they correspond to colorless substances. If spots are not apparent, they can be made visible only if a **visualization method** is used. Often, spots can be seen when the thin-layer plate is held under ultraviolet light; the ultraviolet lamp is a common visualization method. Also common is the use of iodine vapor. The plates are placed in a chamber containing iodine crystals and left to stand for a short time. The iodine reacts with the various compounds adsorbed on the plate to give colored complexes that are clearly visible. Because iodine often changes the compounds by reaction, the components of the mixture cannot be recovered from the plate when the iodine method is used. (Other methods of visualization are discussed in Section 14.7.)

14.2 COMMERCIALLY PREPARED TLC PLATES

The most convenient type of TLC plate is prepared commercially and sold in a ready-to-use form. Many manufacturers supply glass plates precoated with a durable layer of silica gel or alumina. More conveniently, plates are also available that have either a flexible plastic backing or an aluminum backing. The most common types of commercial TLC plates are composed of plastic sheets that are coated with silica gel and polyacrylic acid, which serves as a binder. A fluorescent indicator may be mixed with the silica gel. The indicator renders the spots due to the presence of compounds in the sample visible under ultraviolet light (see Section 14.7). Although these plates are relatively expensive when compared with plates prepared in the laboratory, they are far more convenient to use, and they provide more consistent results. The plates are manufactured quite uniformly. Because the plastic backing is flexible, they have the additional advantage that the coating does not flake off the plates easily. The plastic sheets can also be cut with a pair of scissors to whatever size may be required.

14.3 PREPARATION OF THIN-LAYER SLIDES AND PLATES

The two adsorbent materials used most often for TLC are alumina G (aluminum oxide) and silica gel G (silicic acid). The G designation stands for gypsum (calcium sulfate). Calcined gypsum $CaSO_4 \cdot \frac{1}{2}H_2O$ is better known as plaster of Paris. When exposed to water or moisture, gypsum sets in a rigid mass $CaSO_4 \cdot 2H_2O$, which binds the adsorbent together and to the glass plates used as a backing support. In the adsorbents used for TLC, about 10%–13% by weight of gypsum is added as a binder. The adsorbent materials are otherwise like those used in column chromatography; the adsorbents used in column chromatography have a larger particle size, however. The material for thin-layer work is a fine powder. The small particle size, along with the added gypsum, makes it impossible to use silica gel G or alumina G for column work. In a column, these adsorbents generally set so rigidly that solvent virtually stops flowing through the column.

A. MICROSCOPE SLIDE TLC PLATES

For qualitative work such as identifying the number of components in a mixture or trying to establish that two compounds are identical, small TLC plates made from microscope slides are especially convenient. Coated microscope slides are easily made by dipping the slides into a container holding a slurry of the adsorbent material. Although numerous solvents can be used to prepare a slurry, methylene chloride is probably the best choice. It has the two advantages of low boiling point (40°C) and inability to cause the adsorbent to set or form lumps. The low boiling point means that it is not necessary to dry the coated slides in an oven. Its inability to cause the gypsum binder to set means that slurries made with it are stable for several days. The layer of adsorbent formed is fragile, however, and must be treated carefully. For this reason, some persons prefer to add a small amount of methanol to the methylene chloride to enable the gypsum to set more firmly.

The methanol solvates the calcium sulfate much as water does. More durable plates can be made by dipping plates into a slurry prepared from water. These plates must be oven-dried before use. Also, a slurry prepared from water must be used soon after its preparation. If it is not, it begins to set and form lumps. Thus, an aqueous slurry must be prepared immediately before use; it cannot be used after it has stood for any length of time. For microscope slides, a slurry of silica gel G in methylene chloride is not only convenient but also adequate for most purposes.

Preparing the Slurry. The slurry is most conveniently prepared in a 4-oz wide-mouthed screwcap jar. About 3 mL of methylene chloride is required for each gram of silica gel G. For a smooth slurry without lumps, the silica gel should be added to the solvent while the mixture is being either stirred or swirled. Adding solvent to the adsorbent usually causes lumps to form in the mixture. When the addition is complete, the cap should be placed on the jar tightly and the jar shaken vigorously to ensure thorough mixing (use protective gloves). The slurry may be stored in the tightly capped jar until it is to be used. More methylene chloride may have to be added to replace evaporation losses.

Caution: Avoid breathing silica dust or methylene chloride, prepare and use the slurry in a hood, and avoid getting methylene chloride or the slurry mixture on your skin. Use protective gloves.

Preparing the Slides. If new microscope slides are available, you can use them without any special treatment. However, it is more economical to reuse or recycle microscope slides. Wash the slides with soap and water, rinse with water, and then rinse with 50% aqueous methanol. Allow the plates to dry thoroughly on paper towels. They should be handled by the edges, because fingerprints on the plate surface will make it difficult for the adsorbent to bind to the glass.

Coating the Slides. The slides are coated with adsorbent by dipping them into the container of slurry. You can coat two slides simultaneously by sandwiching them together before dipping them in the slurry.

Perform the coating operation under a hood. Use gloves.

Shake the slurry vigorously just before dipping the slides. Because the slurry settles on standing, it should be mixed in this way before each set of slides is dipped. The depth of the slurry in the jar should be about 3 inches, and the plates should be dipped into the slurry until only about 0.25 inches at the top remains uncoated. The dipping operation should be performed smoothly. The plates may be held at the top (see Fig. 14.1), where they will not be coated. They are dipped into the slurry and withdrawn with a slow and steady motion. The dipping operation takes about 2 seconds. Some practice may be required to get the correct timing. After dipping, replace the cap on the jar, and hold the plates for a minute until most of the solvent has evaporated. Separate the plates and place them on paper towels to complete the drying.

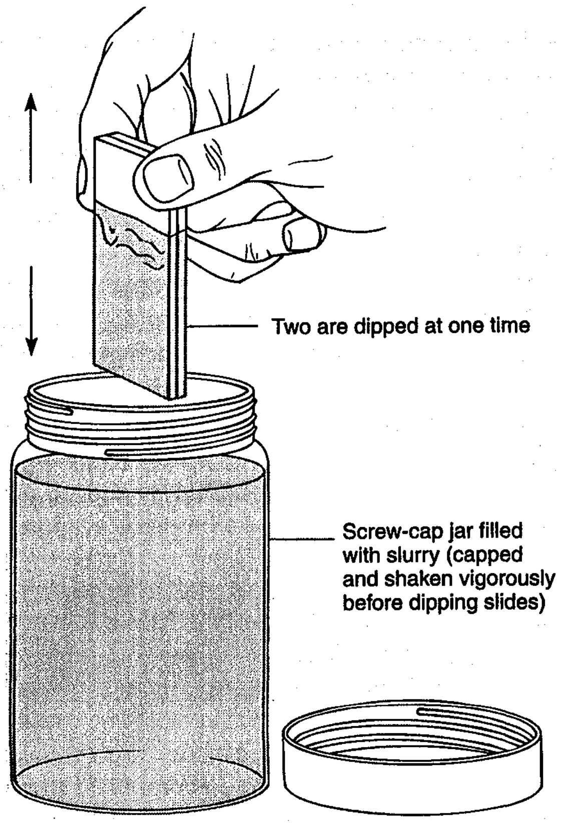
Two are dipped at one time
Screw-cap jar filled with slurry (capped and shaken vigorously before dipping slides)

Figure 14.1 Dipping slides to coat them (gloves recommended).

The plates should have an even coating; there should be no streaks and no thin spots where glass shows through the adsorbent. The plates should not have a thick and lumpy coating.

Two conditions cause thin and streaked plates. First, the slurry may not have been mixed thoroughly before the dipping operation; the adsorbent might then have settled to the bottom of the jar, and the thin slurry at the top would not have coated the slides properly. Second, the slurry simply may not have been thick enough; more silica gel G must then be added to the slurry until the consistency is correct. If the slurry is too thick, the coating on the plates will be thick, uneven, and lumpy. To correct this, dilute the slurry with enough solvent to achieve the proper consistency.

Plates with an unsatisfactory coating may be wiped clean with a paper towel and redipped. Take care to handle the plates only from the top or by the sides, to avoid getting fingerprints on the glass surface.

B. LARGER THIN-LAYER PLATES

For separations involving large amounts of material, or for difficult separations, it may be necessary to use larger thin-layer plates. Plates with dimensions up to 20–25 cm^2

are common. With larger plates, it is desirable to have a more durable coating, and a water slurry of the adsorbent should be used to prepare them. If silica gel is used, the slurry should be prepared in the ratio about 1 g silica gel G to each 2 mL of water. The glass plate used for the thin-layer plate should be washed, dried, and placed on a sheet of newspaper. Place two strips of masking tape along two edges of the plate. Use more than one layer of masking tape if a thicker coating is desired on the plate. A slurry is prepared, shaken well, and poured along one of the untaped edges of the plate.

Observe the precautions stated on page 795.

A heavy piece of glass rod, long enough to span the taped edges, is used to level and spread the slurry over the plate. While the rod is resting on the tape, it is pushed along the plate from the end at which the slurry was poured toward the opposite end of the plate. This is illustrated in Figure 14.2. After the slurry is spread, the masking tape strips are removed, and the plates are dried in a 110°C oven for about 1 hour. Plates of 10–25 cm^2 are easily prepared by this method. Larger plates present more difficulties. Many laboratories have a commercially manufactured spreading machine that makes the entire operation simpler.

14.4 SAMPLE APPLICATION: SPOTTING THE PLATES

Preparing a Micropipet

To apply the sample that is to be separated to the thin-layer plate, one uses a micropipet. A micropipet is easily made from a short length of thin-walled capillary tubing like that used for melting-point determinations. The capillary tubing is heated at its midpoint with a microburner and rotated until it is soft. When the tubing is soft, the heated

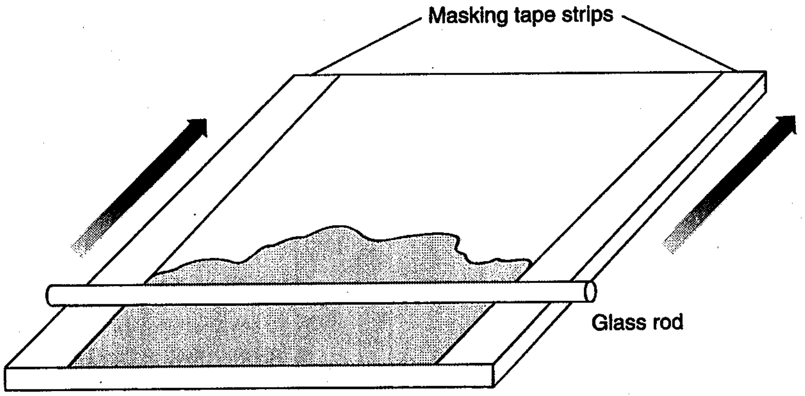

Figure 14.2 Preparing a large plate.

portion of the tubing is drawn out until a constricted portion of tubing 4–5 cm long is formed. After cooling, the constricted portion of tubing is scored at its center with a file or scorer and broken. The two halves yield two capillary micropipets. Figure 14.3 shows how to make such pipets.

Spotting the Plate

To apply a sample to the plate, begin by placing about 1 mg of a solid test substance, or one drop of a liquid test substance, in a small container like a watch glass or a test tube. Dissolve the sample in a few drops of a volatile solvent. Acetone or methylene chloride is usually a suitable solvent. If a solution is to be tested, it can often be used directly. The small capillary pipet, prepared as described, is filled by dipping the pulled end into the solution to be examined. Capillary action fills the pipet. Empty the pipet by touching it *lightly* to the thin-layer plate at a point about 1 cm from the bottom (Fig. 14.4). The spot must be high enough that it does not dissolve in the developing solvent. It is important to touch the plate very lightly and not to gouge a hole in the adsorbent. When the pipet touches the plate, the solution is transferred to the plate as a small spot. The pipet should be touched to the plate *very briefly* and then removed. If the pipet is held to the plate, its entire contents will be delivered to the plate. Only a small amount of material is needed. It is often helpful to blow gently on the plate as the sample is applied. This helps to keep the spot small by evaporating the solvent before it can spread out on the plate. The smaller the spot formed, the better the separation obtainable. If needed, additional material can be applied to the plate by repeating the spotting procedure. You should repeat the procedure with several small amounts, rather than apply one large amount. The solvent should be allowed to evaporate between applications. If the spot is not small (about 2 mm in diameter), a new plate should be prepared. The capillary pipet may be used several times if it is rinsed between uses. It

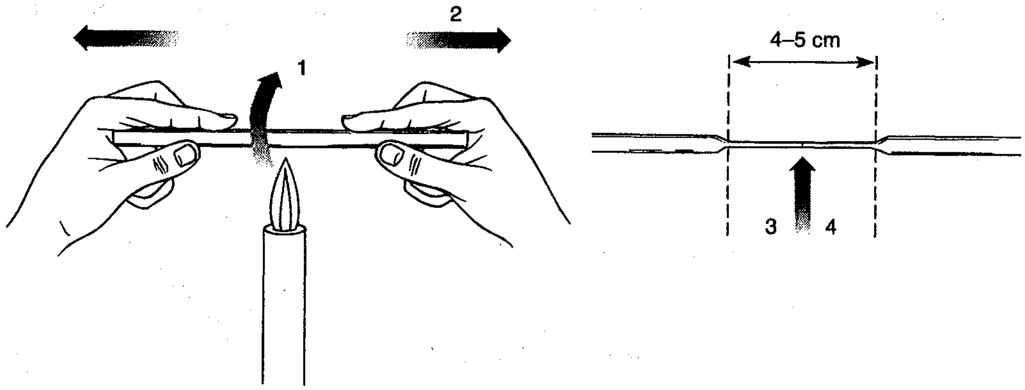

① Rotate in flame until soft.
② Remove from flame and pull.
③ Score lightly in center of pulled section.
④ Break in half to give two pipets.

Figure 14.3 Construction of two capillary micropipets.

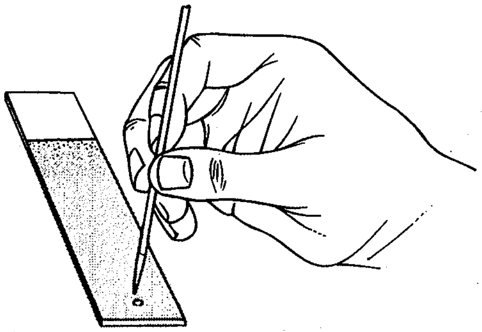

Figure 14.4 Spotting the plate with a drawn capillary pipet.

is repeatedly dipped into a small portion of solvent to rinse it and touched to a paper towel to empty it.

As many as three different spots may be applied to a microscope-slide TLC plate. Each spot should be about 1 cm from the bottom of the plate, and all spots should be evenly spaced, with one spot in the center of the plate. Due to diffusion, spots often increase in diameter as the plate is developed. To keep spots containing different materials from merging, and to avoid confusing the samples, do not place more than three spots on a single plate. Larger plates can accommodate many more samples.

14.5 DEVELOPING (RUNNING) TLC PLATES

Preparing a Development Chamber

A convenient development chamber for microscope-slide TLC plates can be made from a 4-oz wide-mouthed jar. An alternative development chamber can be constructed from a beaker, using aluminum foil to cover the opening. The inside of the jar or beaker should be lined with a piece of filter paper, cut so that it does not quite extend around the inside of the jar. A small vertical opening (2–3 cm) should be left in the filter paper for observing the development. Before development, the filter paper inside the jar or beaker should be thoroughly moistened with the development solvent. The solvent-saturated liner helps to keep the chamber saturated with solvent vapors, thereby speeding the development. Once the liner is saturated, the level of solvent in the bottom of the development chamber is adjusted to a depth of about 5 mm, and the chamber is capped (or covered with aluminum foil) and set aside until it is to be used. A correctly prepared development chamber (with slide in place) is shown in Figure 14.5.

Developing the TLC Plate

Once the spot has been applied to the thin-layer plate and the solvent has been selected (see Sections 14.4–14.6), the plate is placed in the chamber for development. The

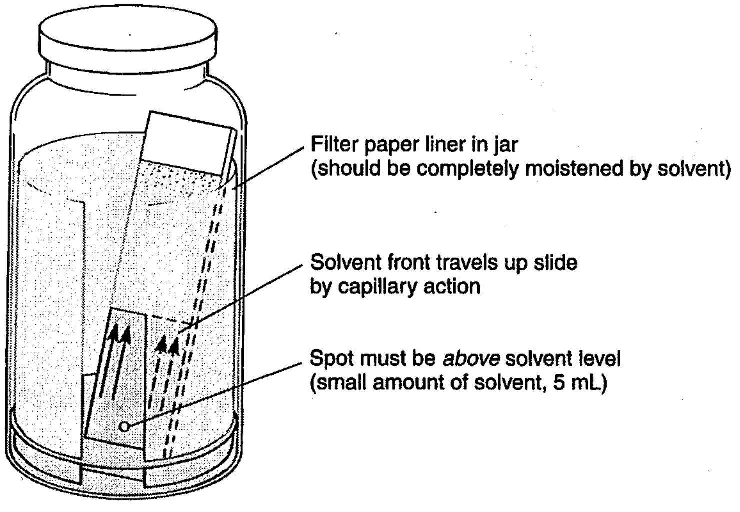
Filter paper liner in jar
(should be completely moistened by solvent)
Solvent front travels up slide
by capillary action
Spot must be *above* solvent level
(small amount of solvent, 5 mL)

Figure 14.5 Development chamber with thin-layer plate undergoing development.

plate must be placed in the chamber carefully so that none of the coated portion touches the filter paper liner. In addition, the solvent level in the bottom of the chamber must not be above the spot that was applied to the plate, or the spotted material will dissolve in the pool of solvent instead of undergoing chromatography. Once the plate has been placed correctly, replace the cap on the developing chamber and wait for the solvent to advance up the plate by capillary action. This generally occurs rapidly, and you should watch carefully. As the solvent rises, the plate becomes visibly moist. When the solvent has advanced to within 5 mm of the end of the coated surface, the plate should be removed, and the position of the solvent front should be marked *immediately* by scoring the plate along the solvent line with a *pencil.* The solvent front must not be allowed to travel beyond the end of the coated surface. The plate should be removed before this happens. The solvent will not actually advance beyond the end of the plate, but spots allowed to stand on a completely moistened plate on which the solvent is not in motion expand by diffusion. Once the plate has dried, any visible spots should be outlined on the plate with a pencil. If no spots are apparent, a visualization method (Section 14.7) may be needed.

14.6 CHOOSING A SOLVENT FOR DEVELOPMENT

The development solvent used depends on the materials to be separated. You may have to try several solvents before a satisfactory separation is achieved. Because microscope slides can be prepared and developed rapidly, an empirical choice is usually not hard to make. A solvent that causes all the spotted material to move with the solvent front is too polar. One that does not cause any of the material in the spot to move is not polar enough. As a guide to the relative polarity of solvents, consult Table 12.2 in Technique 12 (p. 768).

Methylene chloride and toluene are solvents of intermediate polarity and good choices for a wide variety of functional groups to be separated. For hydrocarbon materials, good first choices are hexane, petroleum ether (ligroin), or toluene. Hexane or petroleum ether with varying proportions of toluene or ether gives solvent mixtures of moderate polarity that are useful for many common functional groups. Polar materials may require ethyl acetate, acetone, or methanol.

A rapid way to determine a good solvent is to apply several sample spots to a single plate. The spots should be placed a minimum of 1 cm apart. A capillary pipet is filled with a solvent and gently touched to one of the spots. The solvent expands outward in a circle. The solvent front should be marked with a pencil. A different solvent is applied to each spot. As the solvents expand outward, the spots expand as concentric rings. From the appearance of the rings, you can judge approximately the suitability of the solvent. Several types of behavior experienced with this method of testing are shown in Figure 14.6.

14.7 VISUALIZATION METHODS

It is fortunate when the compounds separated by TLC are colored because the separation can be followed visually. More often than not, however, the compounds are colorless. In that case, the separated materials must be made visible by some reagent or some method that makes the separated compounds visible. Reagents that give rise to colored spots are called **visualization reagents.** Methods of viewing that make the spots apparent are **visualization methods.**

The visualization reagent used most often is iodine. Iodine reacts with many organic materials to form complexes that are either brown or yellow. In this visualization method, the developed and dried TLC plate is placed in a 4-oz wide-mouthed screwcap jar along with a few crystals of iodine. The jar is capped and gently warmed on a steam bath or a hot plate at low heat. The jar fills with iodine vapors and the spots begin to appear. When the spots are sufficiently intense, the plate is removed from the jar, and the spots are outlined with a pencil. The spots are not permanent. Their appearance results from the formation of complexes the iodine makes with the organic substances. As the iodine sublimes off the plate, the spots fade. Hence, they should be marked immediately. Nearly all

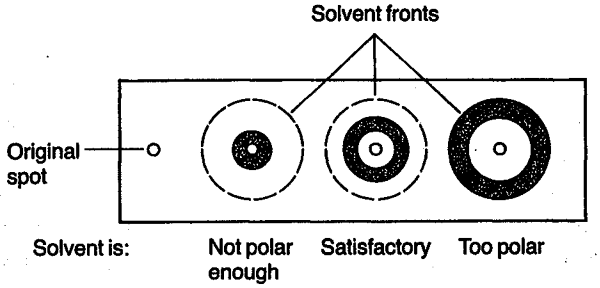

Figure 14.6 Concentric ring method of testing solvents.

compounds except saturated hydrocarbons and alkyl halides form complexes with iodine. The intensities of the spots do not accurately indicate the amount of material present, except in the crudest way.

The second most common method of visualization is by an ultraviolet (UV) lamp. Under UV light, compounds often look like bright spots on the plate. This often suggests the structure of the compound because certain types of compounds shine very brightly under UV light because they fluoresce.

Another method that provides good results involves adding a fluorescent indicator to the adsorbent used to coat the plates. A mixture of zinc and cadmium sulfides is often used. When treated in this way and held under UV light, the entire plate fluoresces. However, dark spots appear on the plate where the separated compounds are seen to quench this fluorescence.

In addition to the preceding methods, several chemical methods are available that either destroy or permanently alter the separated compounds through reaction. Many of these methods are specific for particular functional groups.

Alkyl halides can be visualized if a dilute solution of silver nitrate is sprayed on the plates. Silver halides are formed. These halides decompose if exposed to light, giving rise to dark spots (free silver) on the TLC plate.

Most organic functional groups can be made visible if they are charred with sulfuric acid. Concentrated sulfuric acid is sprayed on the plate, which is then heated in an oven at 110°C to complete the charring. Permanent spots are thus created.

Colored compounds can be prepared from colorless compounds by making derivatives before spotting them on the plate. An example of this is the preparation of 2,4-dinitrophenylhydrazones from aldehydes, and ketones to produce yellow and orange compounds. You may also spray the 2,4-dinitrophenylhydrazine reagent on the plate after the ketones or aldehydes have separated. Red and yellow spots form where the compounds are located. Other examples of this method are using ferric chloride for visualizing phenols and using bromocresol green for detecting carboxylic acids. Chromium trioxide, potassium dichromate, and potassium permanganate can be used for visualizing compounds that are easily oxidized. *p*-Dimethylaminobenzaldehyde easily detects amines. Ninhydrin reacts with amino acids to make them visible. Numerous other methods and reagents available from various supply outlets are specific for certain types of functional groups. These visualize only the class of compounds of interest.

14.8 PREPARATIVE PLATES

If you use large plates (Section 14.3B), materials can be separated and the separated components can be recovered individually from the plates. Plates used in this way are called **preparative plates.** For preparative plates, a thick layer of adsorbent is generally used. Instead of being applied as a spot or a series of spots, the mixture to be separated is applied as a line of material about 1 cm from the bottom of the plate. As the plate is developed, the separated materials form bands. After development, you can observe the separated bands, usually by UV light, and outline the zones in pencil. If the method of visualization is destructive, most of the plate is covered with paper to protect it, and the reagent is applied only at the extreme edge of the plate.

Once the zones have been identified, the adsorbent in those bands is scraped from the plate and extracted with solvent to remove the adsorbed material. Filtration removes the adsorbent, and evaporation of the solvent gives the recovered component from the mixture.

14.9 THE R_f VALUE

Thin layer chromatography conditions include:

1. Solvent system
2. Adsorbent
3. Thickness of the adsorbent layer
4. Relative amount of material spotted

Under an established set of such conditions, a given compound always travels a fixed distance relative to the distance the solvent front travels. This ratio of the distance the compound travels to the distance the solvent travels is called the ***R_f* value.** The symbol R_f stands for "retardation factor," or "ratio-to-front," and it is expressed as a decimal fraction:

$$R_f = \frac{\textit{Distance traveled by substance}}{\textit{Distance traveled by solvent front}}$$

When the conditions of measurement are completely specified, the R_f value is constant for any given compound, and it corresponds to a physical property of that compound.

The R_f value can be used to identify an unknown compound, but like any other identification based on a single piece of data, the R_f value is best confirmed with some additional data. Many compounds can have the same R_f value, just as many compounds have the same melting point.

It is not always possible, in measuring an R_f value, to duplicate exactly the conditions of measurement another researcher has used. Therefore, R_f values tend to be of more use to a single researcher in one laboratory than they are to researchers in different laboratories. The only exception to this is when two researchers use TLC plates from the same source, as in commercial plates, or know the *exact* details of how the plates were prepared. Nevertheless, the R_f value can be a useful guide. If exact values cannot be relied on, the relative values can provide another researcher with useful information about what to expect. Anyone using published R_f values will find it a good idea to check them by comparing them with standard substances whose identity and R_f values are known.

To calculate the R_f value for a given compound, measure the distance that the compound has traveled from the point at which it was originally spotted. For spots that are not too large, measure to the center of the migrated spot. For large spots, the measurement should be repeated on a new plate, using less material. For spots that show tailing, the measurement is made to the "center of gravity" of the spot. This first distance measurement is then divided by the distance the solvent front has traveled from the same original spot. A sample calculation of the R_f values of two compounds is illustrated in Figure 14.7.

14.10 THIN-LAYER CHROMATOGRAPHY APPLIED IN ORGANIC CHEMISTRY

Thin-layer chromatography has several important uses in organic chemistry. It can be used in the following applications:

1. To establish that two compounds are identical
2. To determine the number of components in a mixture
3. To determine the appropriate solvent for a column chromatographic separation
4. To monitor a column chromatographic separation
5. To check the effectiveness of a separation achieved on a column, by crystallization or by extraction
6. To monitor the progress of a reaction

In all these applications, TLC has the advantage that only small amounts of material are necessary. Material is not wasted. With many of the visualization methods, less than a tenth of a microgram (10^{-7} g) of material can be detected. On the other hand, samples as large as a milligram may be used. With preparative plates that are large (about 9 in. on a side) and have a relatively thick coating of adsorbent (>500 μm), it is often possible to separate from 0.2 to 0.5 g of material at one time. The main disadvantage of TLC is that volatile materials cannot be used, because they evaporate from the plates.

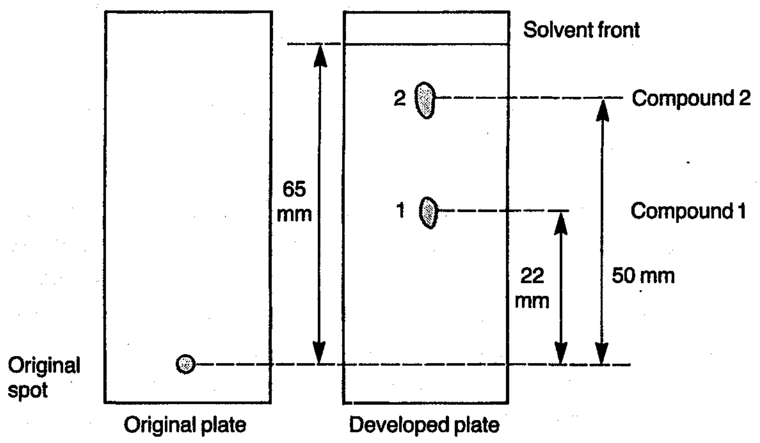

$R_f\text{(compound 1)} = \frac{22}{65} = 0.34 \qquad R_f\text{(compound 2)} = \frac{50}{65} = 0.77$

Figure 14.7 Sample calculation of R_f values.

Thin-layer chromatography can establish that two compounds suspected to be identical are in fact identical. Simply spot both compounds side by side on a single plate and develop the plate. If both compounds travel the same distance on the plate (have the same R_f), they are probably identical. If the spot positions are not the same, the compounds are definitely not identical. It is important to spot both compounds *on the same plate.* This is especially important with hand-dipped microscope slides. Because they vary widely from plate to plate, no two plates have exactly the same thickness of adsorbent. If you use commercial plates, this precaution is not necessary, although it is nevertheless a good idea.

Thin-layer chromatography can establish whether a sample is a single substance or a mixture. A single substance gives a single spot no matter what solvent is used to develop the plate. On the other hand, the number of components in a mixture can be established by trying various solvents on a mixture. A word of caution should be given. It may be difficult, in dealing with compounds of very similar properties, isomers for example, to find a solvent that will separate the mixture. Inability to achieve a separation is not absolute proof that a sample is a single pure substance. Many compounds can be separated only by *multiple developments* of the TLC slide with a fairly nonpolar solvent. In this method, you remove the plate after the first development and allow it to dry. After being dried, it is placed in the chamber again and developed once more. This effectively doubles the length of the slide. At times, several developments may be necessary.

When a mixture is to be separated, you can use TLC to choose the best solvent to separate it if column chromatography is contemplated. You can try various solvents on a plate coated with the same adsorbent as will be used in the column. The solvent that resolves the components best will probably work well on the column. These small-scale experiments are quick, use very little material, and save time that would be wasted by attempting to separate the entire mixture on the column. Similarly, TLC plates can **monitor** a column. A hypothetical situation is shown in Figure 14.8. A solvent was found that

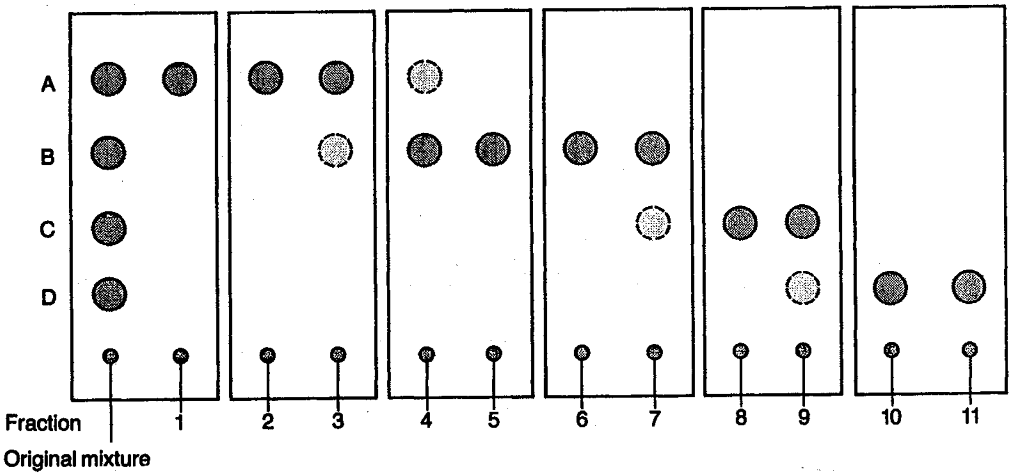

Figure 14.8 Monitoring a column.

would separate the mixture into four components (A–D). A column was run using this solvent, and 11 fractions of 15 mL each were collected. Thin-layer analysis of the various fractions showed that Fractions 1–3 contained Component A; Fractions 4–7, Component B; Fractions 8–9, Component C; and Fractions 10–11, Component D. A small amount of cross-contamination was observed in Fractions 3, 4, 7, and 9.

In another TLC example, a researcher found a product from a reaction to be a mixture. It gave two spots, A and B, on a TLC slide. After the product was crystallized, the crystals were found by TLC to be pure A, whereas the mother liquor was found to have a mixture of A and B. The crystallization was judged to have purified A satisfactorily.

Finally, it is often possible to monitor the progress of a reaction by TLC. At various points during a reaction, samples of the reaction mixture are taken and subjected to TLC analysis. An example is given in Figure 14.9. In this case, the desired reaction was the conversion of A to B. At the beginning of the reaction (0 hr), a TLC slide was prepared that was spotted with pure A, pure B, and the reaction mixture. Similar slides were prepared at 0.5, 1, 2, and 3 hours after the start of the reaction. The slides showed that the reaction was complete in 2 hours. When the reaction was run longer than 2 hours, a new compound, side-product C, began to appear. Thus, the optimum reaction time was judged to be 2 hours.

14.11 PAPER CHROMATOGRAPHY

Paper chromatography is often considered to be related to thin-layer chromatography. The experimental techniques are somewhat like those of TLC, but the principles are more closely related to those of extraction. Paper chromatography is actually a liquid–

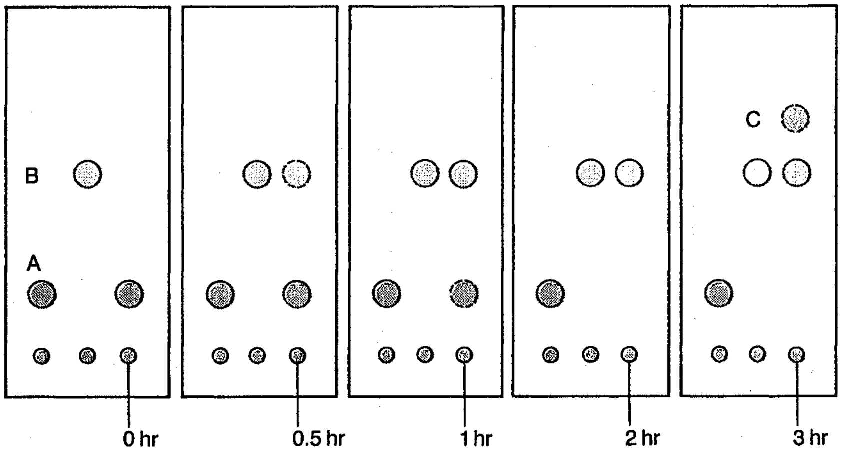

Figure 14.9 Monitoring a reaction.

liquid partitioning technique, rather than a solid–liquid technique. For paper chromatography, a spot is placed near the bottom of a piece of high-grade filter paper (Whatman No. 1 is often used). Then the paper is placed in a developing chamber. The development solvent ascends the paper by capillary action and moves the components of the spotted mixture upward at differing rates. Although paper consists mainly of pure cellulose, the cellulose itself does not function as the stationary phase. Rather, the cellulose absorbs water from the atmosphere, especially from an atmosphere saturated with water vapor. Cellulose can absorb up to about 22% of water. It is this water adsorbed on the cellulose that functions as the stationary phase. To ensure that the cellulose is kept saturated with water, many development solvents used in paper chromatography contain water as a component. As the solvent ascends the paper, the compounds are partitioned between the stationary water phase and the moving solvent. Because the water phase is stationary, the components in a mixture that are most highly water-soluble, or those that have the greatest hydrogen-bonding capacity, are the ones that are held back and move most slowly. Paper chromatography applies mostly to highly polar compounds or to those that are polyfunctional. The most common use of paper chromatography is for sugars, amino acids, and natural pigments. Because filter paper is manufactured consistently, R_f values can often be relied on in paper chromatographic work. However, R_f values are customarily measured from the leading edge (top) of the spot—not from its center, as is customary in TLC.

PROBLEMS

1. A student spots an unknown sample on a TLC plate and develops it in dichloromethane solvent. Only one spot, for which the R_f value is 0.95, is observed. Does this indicate that the unknown material is a pure compound? What can be done to verify the purity of the sample?

2. You and another student were each given an unknown compound. Both samples contained colorless material. You each prepared your own hand-dipped TLC plates and developed the plate using the same solvent. Each of you obtained a single spot of $R_f = 0.75$. Were the samples that you and the other student were assigned necessarily the same substance? How could you prove unambiguously that they were identical, using TLC?

3. Consider a sample that is a mixture composed of biphenyl, benzoic acid, and benzyl alcohol. The sample is spotted on a TLC plate and developed in a dichloromethane–cyclohexane solvent mixture. Predict the *relative* R_f values for the three components in the sample. *Hint:* See Table 12.3.

4. Calculate the R_f value of a spot that travels 5.7 cm, with a solvent front that travels 13 cm.

5. A student spots an unknown sample on a TLC plate and develops it in pentane solvent. Only one spot, for which the R_f value is 0.05, is observed. Is the unknown material a pure compound? What can be done to verify the purity of the sample?

6. A *colorless* unknown substance is spotted on a TLC plate and developed in the correct solvent. The spots do not appear when visualization with a UV lamp or iodine vapors is attempted. What could you do in order to visualize the spots if the compound is

- **(a)** an alkyl halide
- **(b)** a ketone
- **(c)** an amino acid
- **(d)** a sugar

TECHNIQUE 15

Gas Chromatography

Gas chromatography is one of the most useful instrumental tools for separating and analyzing organic compounds that can be vaporized without decomposition. Common uses include testing the purity of a substance and separating the components of a mixture. The relative amounts of the components in a mixture may also be determined. In some cases, gas chromatography can be used to identify a compound. In microscale work, it can also be used as a preparative method to isolate pure compounds from a small amount of a mixture.

Gas chromatography resembles column chromatography in principle, but it differs in three respects. First, the partitioning processes for the compounds to be separated are carried out between a **moving gas phase** and a **stationary liquid phase.** (Recall that in column chromatography the moving phase is a liquid and the stationary phase is a solid adsorbent.) A second difference is that the temperature of the gas system can be controlled, because the column is contained in an insulated oven. And third, the concentration of any given compound in the gas phase is a function of its vapor pressure only. Because gas chromatography separates the components of a mixture primarily on the basis of their vapor pressures (or boiling points), this technique is also similar in principle to fractional distillation. In microscale work, it is sometimes used to separate and isolate compounds from a mixture; fractional distillation would normally be used with larger amounts of material.

Gas chromatography (GC) is also known as vapor-phase chromatography (VPC) and as gas–liquid partition chromatography (GLPC). All three names, as well as their indicated abbreviations, are often found in the literature of organic chemistry. In reference to the technique, the last term, GLPC, is the most strictly correct and is preferred by most authors, but GC is used most often.

15.1 THE GAS CHROMATOGRAPH

The apparatus used to carry out a gas–liquid chromatographic separation is generally called a **gas chromatograph.** A typical student-model gas chromatograph, the GOW-MAC model 69–350, is illustrated in Figure 15.1. A schematic block diagram of a basic gas chromatograph is shown in Figure 15.2. The basic elements of the apparatus are apparent. In short, the sample is injected into the chromatograph, and it is immediately vaporized in a heated injection chamber and introduced into a moving stream of gas, called the **carrier gas.** The vaporized sample is then swept into a column filled with particles coated with a liquid adsorbent. The column is contained in a temperature-controlled oven. As the sample passes through the column, it is subjected to many gas–liquid partitioning processes, and the components are separated. As each component leaves the column, its presence is detected by an electrical detector that generates a signal that is recorded on a strip chart recorder.

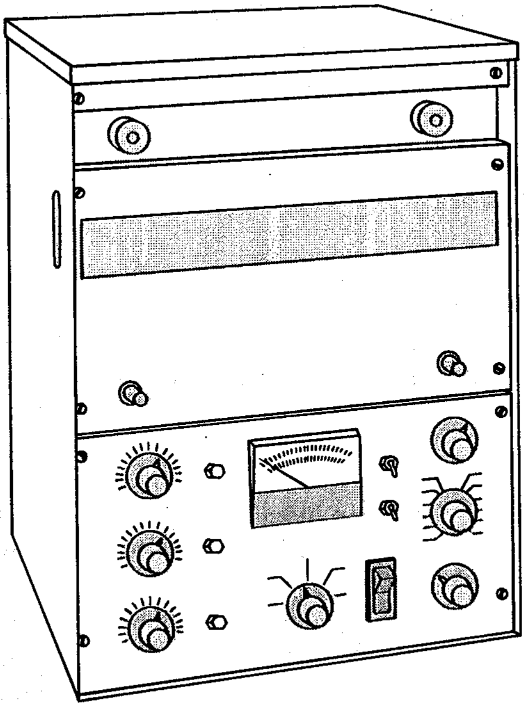

Figure 15.1 Gas chromatograph.

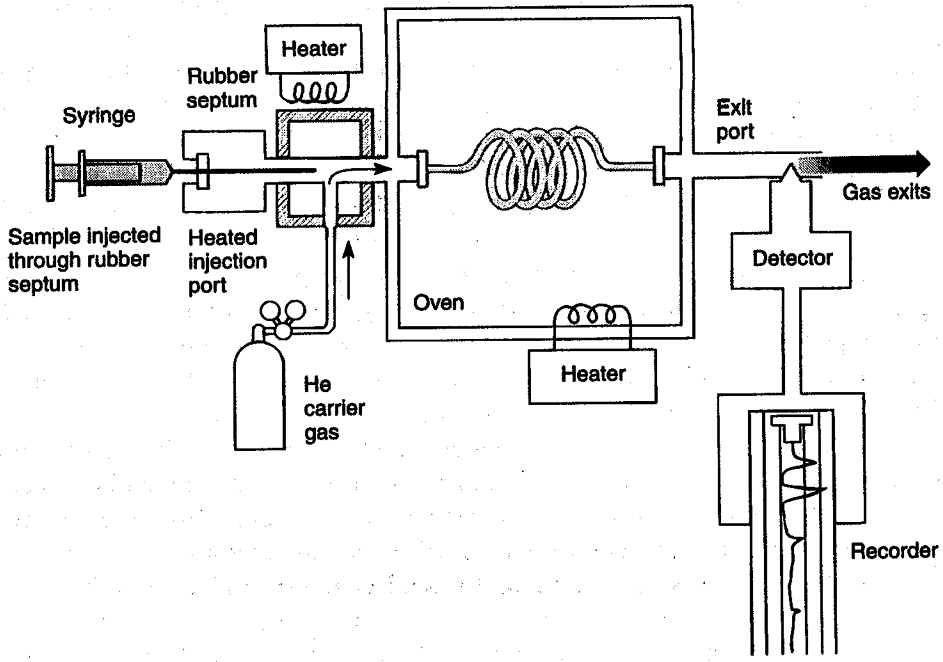
Heater
Rubber septum
Syringe
Exit port
Gas exits
Sample injected through rubber septum
Heated injection port
Detector
Oven
Heater
He carrier gas
Recorder

Figure 15.2 Schematic diagram of gas chromatograph.

Many modern instruments are also equipped with a microprocessor, which can be programmed to change parameters, such as the temperature of the oven, while a mixture is being separated on a column. With this capability, it is possible to optimize the separation of components and to complete a run in a relatively short time.

15.2 THE COLUMN

The heart of the gas chromatograph is the packed column. This column is usually made of copper or stainless steel tubing, but sometimes glass is used. The most common diameters of tubing are $\frac{1}{8}$ in. (3 mm) and $\frac{1}{4}$ in. (6 mm). To construct a column, cut a piece of tubing to the desired length and attach the proper fittings on each of the two ends to connect it to the apparatus. The most common length is 4–12 ft, but some columns may be up to 50 ft in length.

The tubing (column) is then packed with the **stationary phase.** The material chosen for the stationary phase is usually a liquid, a wax, or a low-melting solid. This material should be relatively nonvolatile, that is, it should have a low vapor pressure and a high boiling point. Liquids commonly used are high-boiling hydrocarbons, silicone oils, waxes, and polymeric esters, ethers, and amides. Some typical substances are listed in Table 15.1.

The liquid phase is usually coated onto a **support material.** A common support material is crushed firebrick. Many methods exist for coating the high-boiling liquid phase onto the support particles. The easiest is to dissolve the liquid (or low-melting wax or solid) in a volatile solvent like methylene chloride (bp 40°C). The firebrick (or other support) is added to this solution, which is then slowly evaporated (rotary evaporator) so as to leave each particle of support material evenly coated. Other support materials are listed in Table 15.2.

In the final step, the liquid-phase-coated support material is packed into the tubing as evenly as possible. The tubing is bent or coiled so that it fits into the oven of the gas chromatograph with its two ends connected to the gas entrance and exit ports.

Selection of a liquid phase usually revolves about two factors. First, most of them have an upper temperature limit above which they cannot be used. Above the specified limit of temperature, the liquid phase itself will begin to "bleed" off the column. Second, the materials to be separated must be considered, For polar samples, it is usually best to use a polar liquid phase; for nonpolar samples, a nonpolar liquid phase is indicated. The liquid phase performs best when the substances to be separated *dissolve* in it.

Most researchers today buy packed columns from commercial sources, rather than pack their own. A wide variety of types and lengths are available.

Alternatives to packed columns are Golay or glass capillary columns of diameters 0.1–0.2 mm. With these columns, no solid support is required, and the liquid is coated directly on the inner walls of the tubing. Liquid phases commonly used in glass capillary columns are similar in composition to those used in packed columns. They include DB-1 (similar to SE-30), DB-17 (similar to DC-710), and DB-WAX (similar to Carbowax 20M).

TABLE 15.1 Typical Liquid Phases

		Type	Composition	Maximum Temperature (°C)	Typical Use
Increasing polarity ↓	Apiezons (L, M, N, etc.)	Hydrocarbon greases (varying MW)	Hydrocarbon mixtures	250–300	Hydrocarbons
	SE–30	Methyl silicone rubber	Like silicone oil, but cross-linked	350	General applications
	DC–200	Silicone oil (R = CH_3)	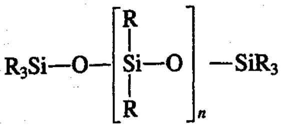 $R_3Si{-}O{-}\left[\overset{R}{\underset{R}{Si}}{-}O\right]_n{-}SiR_3$	225	Aldehydes, ketones, halocarbons
	DC–710	Silicone oil (R = CH_3) (R′ = C_6H_5)	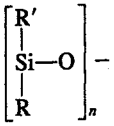 $\left[\overset{R'}{\underset{R}{Si}}{-}O\right]_n{-}$	300	General applications
	Carbowaxes (400–20M)	Polyethylene glycols (varying chain lengths)	Polyether $HO{-}(CH_2CH_2{-}O)_n{-}CH_2CH_2OH$	Up to 250	Alcohols, ethers, halocarbons
	DEGS	Diethylene glycol succinate	Polyester 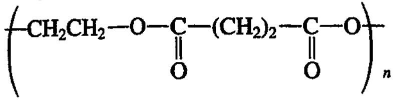 $\left(CH_2CH_2{-}O{-}\underset{\overset{\|}{O}}{C}{-}(CH_2)_2{-}\underset{\overset{\|}{O}}{C}{-}O\right)_n$	200	General applications

TABLE 15.2 Typical Solid Supports

Crushed firebrick
Nylon beads
Glass beads
Silica
Alumina
Charcoal
Molecular sieves

Chromosorb T
  (Teflon beads)
Chromosorb P
  (Pink diatomaceous earth,
  highly absorptive, pH 6–7)
Chromosorb W
  (White diatomaceous earth,
  medium absorptivity, pH 8–10)
Chromosorb G
  (like the above,
  low absorptivity, pH 8.5)

The length of a capillary column is usually very long, typically 50–100 ft. Because of the length and small diameter, there is increased interaction between the sample and the stationary phase. Gas chromatographs equipped with these small-diameter columns are able to separate components more effectively than instruments using packed columns.

15.3 PRINCIPLES OF SEPARATION

After a column is selected, packed, and installed, the **carrier gas** (usually helium, argon, or nitrogen) is allowed to flow through the column supporting the liquid phase. The mixture of compounds to be separated is introduced into the carrier gas stream, where its components are equilibrated (or partitioned) between the moving gas phase and the stationary liquid phase (Figure 15.3). The latter is held stationary because it is adsorbed onto the surfaces of the support material.

The sample is introduced into the gas chromatograph by a microliter syringe. It is injected as a liquid or as a solution through a rubber septum into a heated chamber, called the **injection port,** where it is vaporized and mixed with the carrier gas. As this mixture reaches the column, which is heated in a controlled oven, it begins to equilibrate between the liquid and gas phases. The length of time required for a sample to move through the column is a function of how much time it spends in the vapor phase and how much time it spends in the liquid phase. The more time it spends in the vapor phase, the faster it gets to the end of the column. In most separations, the components of a sample have similar solubilities in the liquid phase. Therefore, the time the different compounds spend in the vapor phase is primarily a function of their vapor pressure, and the more volatile component arrives at the end of the column first, as illustrated in Figure 15.3. By selecting the correct temperature of the oven and the correct liquid phase, the compounds in the injected mixture travel through the column at different rates and are separated.

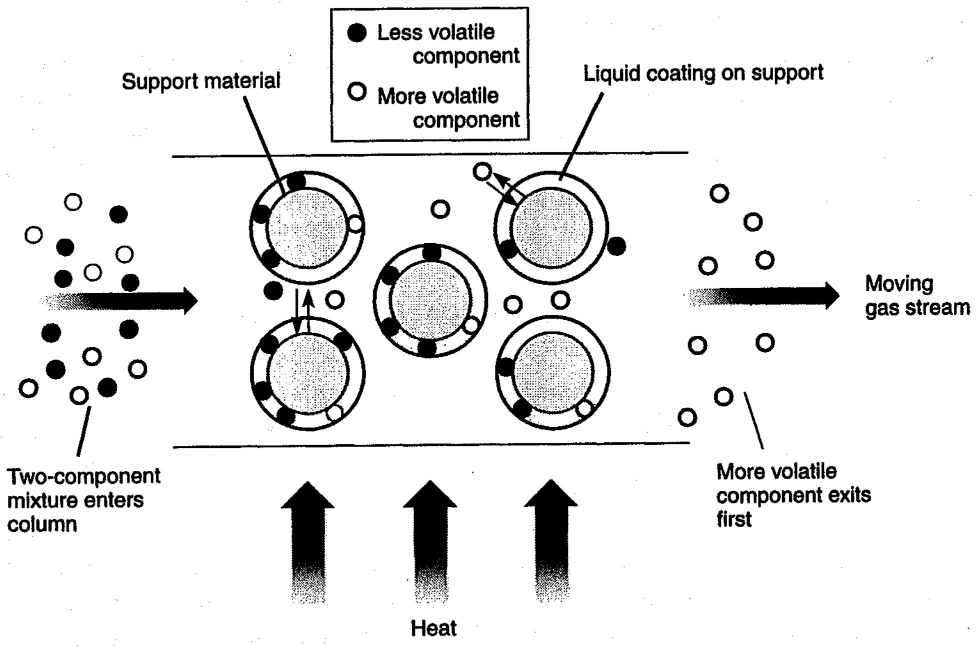
● Less volatile component
○ More volatile component
Support material
Liquid coating on support
Moving gas stream
Two-component mixture enters column
More volatile component exits first
Heat

Figure 15.3 The separation process.

15.4 FACTORS AFFECTING SEPARATION

Several factors determine the rate at which a given compound travels through a gas chromatograph. First of all, compounds of low boiling point will generally travel through the gas chromatograph faster than compounds of higher boiling point. This is because the column is heated, and low-boiling compounds always have higher vapor pressures than compounds of higher boiling point. In general, therefore, for compounds with the same functional group, the higher the molecular weight the longer the retention time. For most molecules, the boiling point increases as the molecular weight increases. If the column is heated to a temperature that is too high, however, the entire mixture to be separated is flushed through the column at the same rate as the carrier gas, and no equilibration takes place with the liquid phase. On the other hand, at too low a temperature, the mixture dissolves in the liquid phase and never revaporizes. Thus, it is retained on the column.

Second, the rate of flow of the carrier gas is important. The carrier gas must not move so rapidly that molecules of the sample in the vapor phase cannot equilibrate with those dissolved in the liquid phase. This may result in poor separation between components in the injected mixture. If the rate of flow is too slow, however, the bands broaden significantly, leading to poor resolution (see Section 15.8).

The third factor is the choice of liquid phase used in the column. The molecular weights, functional groups, and polarities of the component molecules in the mixture to be separated must be considered when a liquid phase is being chosen. One generally uses a different type of material for hydrocarbons, for instance, than for esters. The materials to be separated should *dissolve* in the liquid. The useful temperature limit of the liquid phase selected must also be considered.

Fourth, the length of the column is important. Compounds that resemble one another closely, in general, require longer columns than dissimilar compounds. Many kinds of isomeric mixtures fit into the "difficult" category. The components of isomeric mixtures are so much alike that they travel through the column at very similar rates. You need a longer column, therefore, to take advantage of any differences that may exist.

15.5 ADVANTAGES OF GAS CHROMATOGRAPHY

All factors that have been mentioned must be adjusted by the chemist for any mixture to be separated. Considerable preliminary investigation is often required before a mixture can be separated successfully into its components by gas chromatography. Nevertheless, the advantages of the technique are many.

First, many mixtures can be separated by this technique when no other method is adequate. Second, as little as 1–10 μL ($1\ \mu L = 10^{-6}$ L) of a mixture can be separated by this technique. This advantage is particularly important when working at the microscale level. Third, when gas chromatography is coupled with an electronic recording device (see following discussion), the amount of each component present in the separated mixture can be estimated quantitatively.

The range of compounds that can be separated by gas chromatography extends from gases, such as oxygen (bp −183°C) and nitrogen (bp −196°C), to organic compounds with boiling points over 400°C. The only requirement for the compounds to be separated is that they have an appreciable vapor pressure at a temperature at which they can be separated and that they be thermally stable at this temperature.

15.6 MONITORING THE COLUMN (THE DETECTOR)

To follow the separation of the mixture injected into the gas chromatograph, it is necessary to use an electrical device called a **detector.** Two types of detectors in common use are the **thermal conductivity detector (TCD)** and the **flame ionization detector (FID).**

The thermal conductivity detector is simply a hot wire placed in the gas stream at the column exit. The wire is heated by constant electrical voltage. When a steady stream of carrier gas passes over this wire, the rate at which it loses heat and its electrical conductance have constant values. When the composition of the vapor stream changes, the rate of heat flow from the wire, and hence its resistance, changes. Helium, which has a higher thermal conductivity than most organic substances, is a common carrier gas. Thus, when a substance elutes in the vapor stream, the thermal conductivity of the moving gases will be lower than with helium alone. The wire then heats up, and its resistance decreases.

A typical TCD operates by difference. Two detectors are used: one exposed to the actual effluent gas and the other exposed to a reference flow of carrier gas only. To achieve this situation, a portion of the carrier gas stream is diverted *before* it enters the injection port. The diverted gas is routed through a reference column into which no sample has been admitted. The detectors mounted in the sample and reference columns are arranged so as to form the arms of a Wheatstone bridge circuit, as shown in Figure 15.4. As long as the carrier gas alone flows over both detectors, the circuit is in balance. However, when a sample elutes from the sample column, the bridge circuit becomes unbalanced, creating an electrical signal. This signal can be amplified and used to activate a strip chart recorder. The recorder is an instrument that plots, by means of a moving pen, the unbalanced bridge current versus time on a continuously moving roll of chart paper. This record of detector response (current) versus time is called a **chromatogram.** A typical gas chromatogram is illustrated in Figure 15.5. Deflections of the pen are called **peaks.**

When a sample is injected, some air (CO_2, H_2O, N_2, and O_2) is introduced along with the sample. The air travels through the column almost as rapidly as the carrier gas; as it passes the detector, it causes a small pen response, thereby giving a peak, called the **air peak.** At later times (t_1, t_2, t_3), the components also give rise to peaks on the chromatogram as they pass out of the column and past the detector.

In a flame ionization detector, the effluent from the column is directed into a flame produced by the combustion of hydrogen, as illustrated in Figure 15.6. As organic compounds burn in the fame, ion fragments are produced which collect on the ring above the flame. The resulting electrical signal is amplified and sent to a recorder in a similar manner to a TCD, except that a FID does not produce an air peak. The main advantage of the FID is that it is more sensitive and can be used to analyze smaller quantities of sample. Also, because a FID does not respond to water, a gas chromatograph with this detector can be used to analyze aqueous solutions. Two disadvantages are that it is more difficult

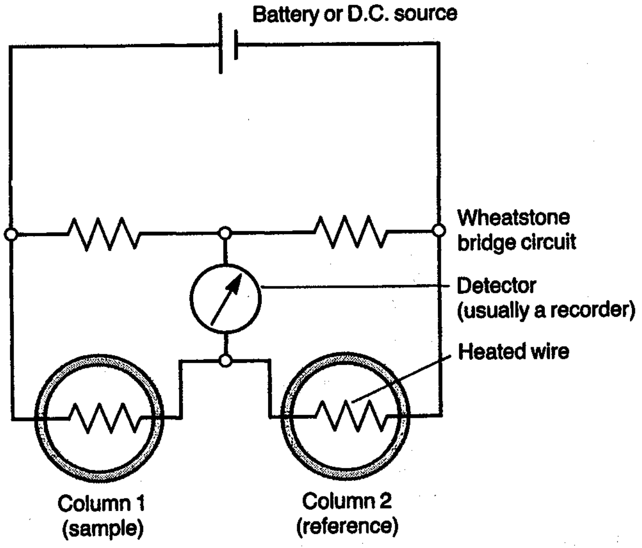

Figure 15.4 Typical thermal conductivity detector.

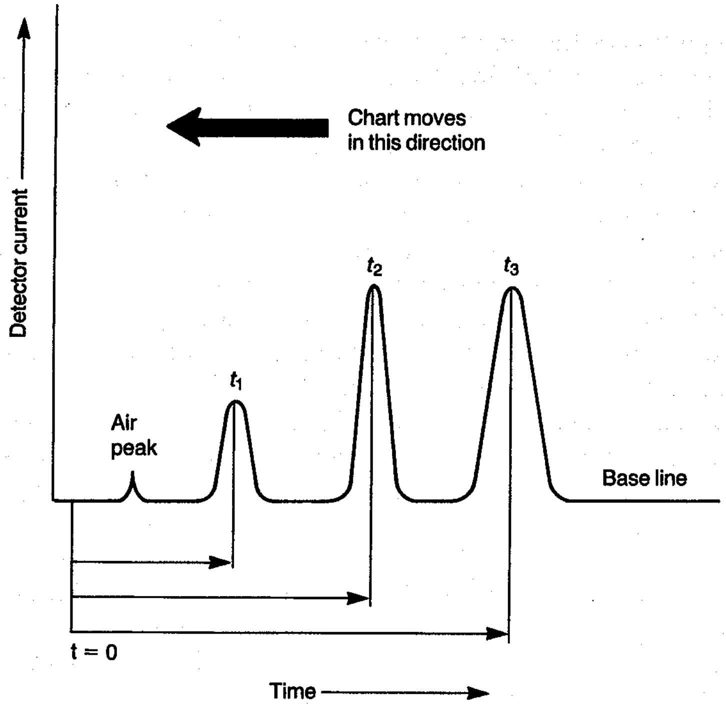

Figure 15.5 Typical chromatograph.

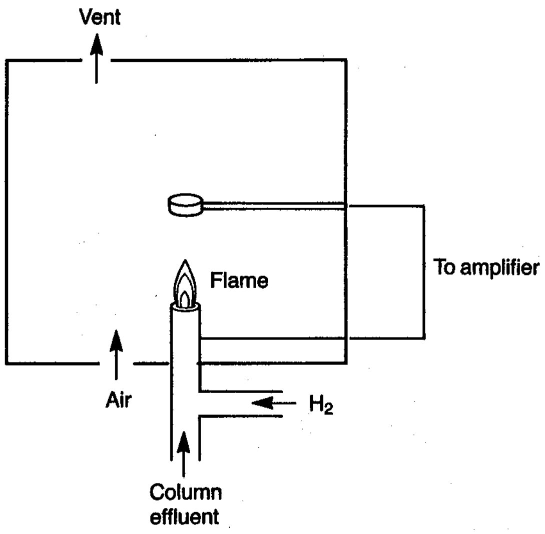
Vent
Flame
To amplifier
Air
H_2
Column
effluent

Figure 15.6 Flame ionization detector.

to operate and the detection process destroys the sample. Therefore, an FID gas chromatograph cannot be used to do preparative work, which is often desired in the microscale laboratory.

15.7 RETENTION TIME

The period following injection that is required for a compound to pass through the column is called the **retention time** of that compound. For a given set of constant conditions (flow rate of carrier gas, column temperature, column length, liquid phase, injection port temperature, carrier), the retention time of any compound is always constant (much like the R_f value in thin-layer chromatography, as described in Technique 14, Section 14.9, p. 803). The retention time is measured from the time of injection to the time of maximum pen deflection (detector current) for the component being observed. This value, when obtained under controlled conditions, can identify a compound by a direct comparison of it with values for known compounds determined under the same conditions. For easier measurement of retention times, most strip chart recorders are adjusted to move the paper at a rate that corresponds to time divisions calibrated on the chart paper. The retention times (t_1, t_2, t_3) are indicated in Figure 15.5 for the three peaks illustrated.

15.8 POOR RESOLUTION AND TAILING

The peaks in Figure 15.5 are well **resolved.** That is, the peaks are separated from one another, and between each pair of adjacent peaks the tracing returns to the **baseline.** In Figure 15.7, the peaks overlap and the resolution is not good. Poor resolution is often caused by using too much sample, too high a column temperature, too short a column, a liquid phase that does not discriminate well between the two components, a column with too large a diameter, or, in short, almost any wrongly adjusted parameter. When peaks are poorly resolved, it is more difficult to determine the relative amount of each component. Methods for determining the relative percentages of each component are given in Section 15.11.

Another desirable feature illustrated by the chromatogram in Figure 15.5 is that each peak is symmetrical. A common example of an unsymmetrical peak is one in which **tailing** has occurred, as shown in Figure 15.8. Tailing usually results from injecting too much sample into the gas chromatograph. Another cause of tailing occurs with polar compounds,

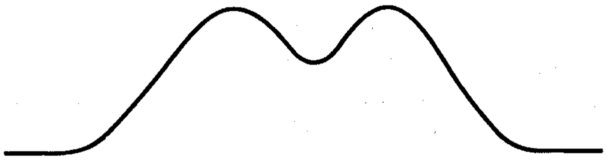

Figure 15.7 Poor resolution or peaks overlap.

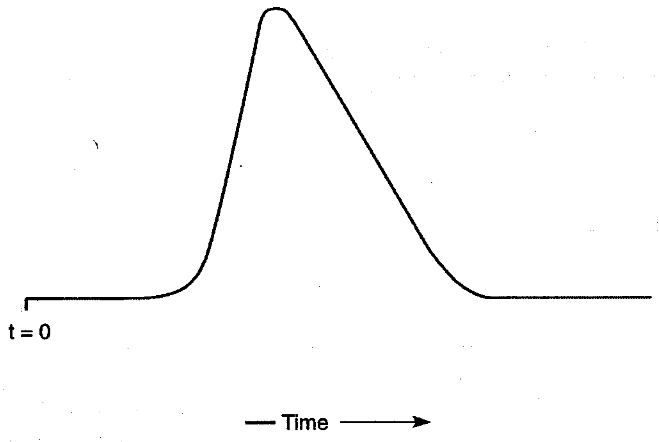

Figure 15.8 Tailing.

such as alcohols and aldehydes. These compounds may be temporarily adsorbed on column walls or areas of the support material that are not adequately coated by the liquid phase. Therefore, they do not leave in a band, and tailing results.

15.9 QUALITATIVE ANALYSIS

A disadvantage of the gas chromatograph is that it gives no information whatever about the identities of the substances it has separated. The little information it does provide is given by the retention time. It is hard to reproduce this quantity from day to day, however, and exact duplications of separations performed last month may be difficult to make this month. It is usually necessary to **calibrate** the column each time it is used. That is, you must run pure samples of all known and suspected components of a mixture individually, just before chromatographing the mixture, to obtain the retention time of each known compound. As an alternative, each suspected component can be added, one by one, to the unknown mixture while the operator looks to see which peak has its intensity increased relative to the unmodified mixture. Another solution is to *collect* the components individually as they emerge from the gas chromatograph. Each component can then be identified by other means, such as by infrared or nuclear magnetic resonance spectroscopy or by mass spectrometry.

15.10 COLLECTING THE SAMPLE

For gas chromatographs with a thermal conductivity detector, it is possible to collect samples that have passed through the column. One uses a form of cooled trap connected to the outlet of the column as a collecting device. The simplest form of trap is just a U-shaped tube dipped into a bath of coolant (ice water, liquid nitrogen, or dry ice–acetone), as shown in Figure 15.9.

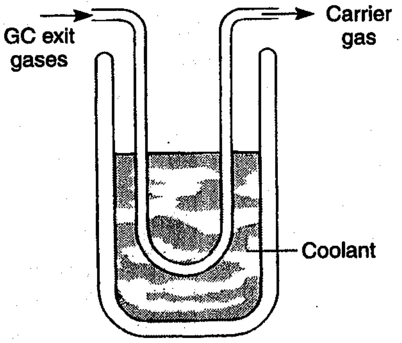

Figure 15.9 Collection trap.

For instance, if the coolant is liquid nitrogen (bp −196°C) and the carrier gas is helium (bp −269°C), compounds boiling above the temperature of liquid nitrogen generally are condensed or trapped at the bottom of the U-shaped tube, while the carrier gas continues to flow. However, if the compound leaving is frozen in the form of a fine mist, it may be carried through the trap even though it is frozen. When this happens, special trapping arrangements are required. To collect each component of the mixture, one must change the trap each time a new peak begins to appear on the chromatogram. Since even a gas chromatograph with large-diameter columns can rarely handle more than about 0.5 mL of a mixture at a time, a chemist or student having 50 mL of a mixture would have to make 50 to 100 separate injections, changing the trap for each component as it comes off the column. Fortunately, there are accessories that can be attached to a gas chromatograph to do this task automatically.

A microscale method uses a gas collection tube (see Fig. 15.10), which is included in most microscale glassware kits. A collection tube is joined to the exit port of the col-

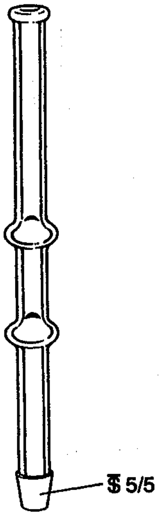

Figure 15.10 Gas chromatograph collection tube.

umn by inserting the ₸ 5/5 inner joint into a metal adapter, which is connected to the exit port. When a sample is eluted from the column in the vapor state, it is cooled by the connecting adapter and the gas collection tube and condenses in the collection tube. The gas collection tube is removed from the adapter when the recorder indicates that the desired sample has completely passed through the column. After the first sample has been collected, the process can be repeated with another gas collection tube.

To isolate the liquid, the tapered joint of the collection tube is inserted into a 0.1-mL conical vial, which has a ₸ 5/5 outer joint. The assembly is placed into a test tube, as illustrated in Figure 15.11. During centrifugation, the sample is forced into the bottom of the conical vial. After disassembling the apparatus, the liquid can be removed from the vial with a syringe for a boiling-point determination or analysis by infrared spectroscopy. If a determination of the sample weight is desired, the empty conical vial and cap should be tared and reweighed after the liquid has been collected. It is advisable to dry the gas collection tube and the conical vial in an oven before use in order to prevent contamination by water or other solvents used in cleaning this glassware.

15.11 QUANTITATIVE ANALYSIS

The area under a gas chromatograph peak is proportional to the amount (moles) of compound eluted. Hence, the molar percentage composition of a mixture can be approximated by comparing relative peak areas. This method of analysis assumes that the de-

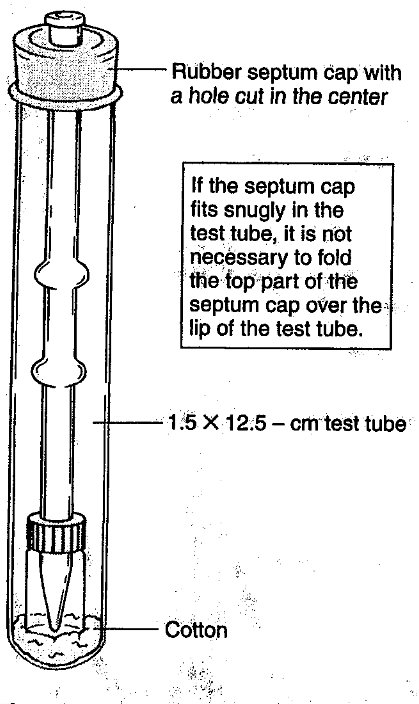
Rubber septum cap with a hole cut in the center
If the septum cap fits snugly in the test tube, it is not necessary to fold the top part of the septum cap over the lip of the test tube.
1.5 X 12.5 – cm test tube
Cotton

Figure 15.11 Gas chromatograph collection tube and 0.1-mL conical vial.

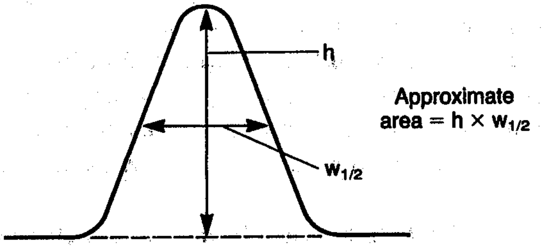

Figure 15.12 Triangulation of a peak.

tector is equally sensitive to all compounds eluted and that it gives a linear response with respect to amount. Nevertheless, it gives reasonably accurate results.

The simplest method of measuring the area of a peak is by geometrical approximation, or triangulation. In this method, you multiply the height *h* of the peak above the baseline of the chromatogram by the width of the peak at half of its height $w_{1/2}$. This is illustrated in Figure 15.12. The baseline is approximated by drawing a line between the two side-arms of the peak. This method works well only if the peak is symmetrical. If the peak has tailed or is unsymmetrical, it is best to cut out the peaks with scissors and weigh the pieces of paper on an **analytical balance.** Because the weight per area of a piece of good chart paper is reasonably constant from place to place, the ratio of the areas is the same as the ratio of the weights. To obtain a percentage composition for the mixture, first add all the peak areas (weights). Then, to calculate the percentage of any component in the mixture, divide its individual area by the total area and multiply the result by 100. A sample calculation is illustrated in Figure 15.13. If peaks overlap (see Fig. 15.7), either

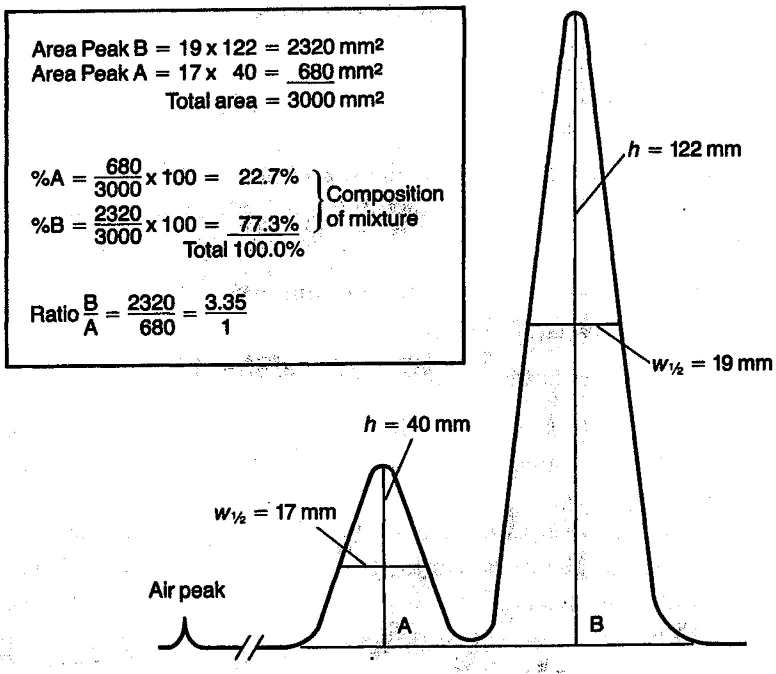

Area Peak B = 19 x 122 = 2320 mm²
Area Peak A = 17 x 40 = $\underline{680}$ mm²
Total area = 3000 mm²

$$\left.\begin{array}{l}\%A = \frac{680}{3000} \times 100 = 22.7\% \\ \%B = \frac{2320}{3000} \times 100 = \underline{77.3\%}\end{array}\right\} \text{Composition of mixture}$$

Total 100.0%

$$\text{Ratio}\ \frac{B}{A} = \frac{2320}{680} = \frac{3.35}{1}$$

Figure 15.13 Sample percentage composition calculation.

[illegible] 27 29

$W_{1/2} = 9$
60 = h

the gas chromatographic conditions must be readjusted to achieve better resolution of the peaks or the peak shape must be estimated.

There are various instrumental means, which are built into recorders, of detecting the amounts of each sample automatically. One method uses a separate pen that produces a trace that integrates the area under each peak. Another method employs an electronic device that automatically prints out the area under each peak and the percentage composition of the sample.

For the experiments in this textbook, we have assumed that the detector is equally sensitive to all compounds eluted. Compounds with different functional groups or with widely varying molecular weights, however, produce different responses with both TCD and FID gas chromatographs. With a TCD, the responses are different because not all compounds have the same thermal conductivity. Different compounds analyzed with a FID gas chromatograph also give different responses because the detector response varies with the type of ions produced. For both types of detectors, it is possible to calculate a **response factor** for each compound in a mixture. Response factors are usually determined by making up an equimolar mixture of two compounds, one of which is considered to be the reference. The mixture is separated on a gas chromatograph, and the relative percentages are calculated using one of the methods described previously. From these percentages you can determine a response factor for the compound being compared to the reference. If you do this for all the components in a mixture, you can then use these correction factors to make more accurate calculations of the relative percentages for the compounds in the mixture.

Consider the following example, which illustrates how response factors are determined. In this example, an equimolar mixture of benzene, hexane, and ethyl acetate is prepared and analyzed using a flame-ionization gas chromatograph. The peak areas obtained are:

Benzene	966463
Hexane	831158
Ethyl Acetate	1449695

In most cases, benzene is taken as the standard, and its response factor is defined to be equal to 1.00. Calculation of the response factors for the other components of the test mixture proceeds as follows:

Benzene	966463/966463 = 1.00 (by definition)
Hexane	831158/966463 = 0.86
Ethyl Acetate	1449695/966463 = 1.50

Notice that the response factors calculated in this example are *molar* response factors. It is necessary to correct these values by the relative molecular weights of each substance to obtain *weight* response factors.

When you use a flame-ionization gas chromatograph for quantitative analysis, it is first necessary to determine the response factors for each component of the mixture being analyzed. For a quantitative analysis, it is likely that you will have to convert molar response factors into *weight* response factors. Next, the chromatography experiment using the unknown samples is performed. The observed peak areas for each component are corrected using the response factors in order to arrive at the correct weight percentage of each component in the sample.

PROBLEMS

1. **(a)** A sample consisting of 1-bromopropane and 1-chloropropane is injected into a gas chromatograph equipped with a nonpolar column. Which compound has the shorter retention time? Explain your answer.

(b) If the same sample were run several days later with the conditions as nearly the same as possible, would you expect the retention times to be identical to those obtained the first time? Explain.

2. Using triangulation, calculate the percentage of each component in a mixture composed of two substances, A and B. The chromatogram is shown in Figure 15.14.

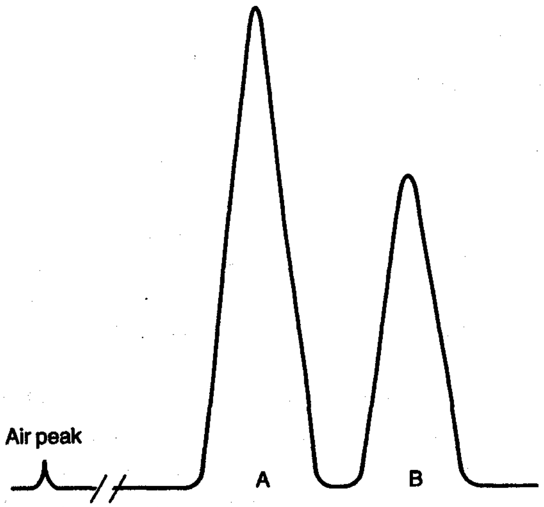

Figure 15.14 Chromatogram for Problem 2.

3. Make a photocopy of the chromatogram in Figure 15.14. Cut out the peaks and weigh them on an analytical balance. Use the weights to calculate the percentage of each component in the mixture. Compare your answer to what you calculated in Problem 2.

4. What would happen to the retention time of a compound if the following changes were made?

(a) Decrease the flow rate of the carrier gas

(b) Increase the temperature of the column

(c) Increase the length of the column

TECHNIQUE 16

Sublimation

In Technique 6, the influence of temperature on the change in vapor pressure of a liquid was considered (see Fig. 6.8, p. 675). It was shown that the vapor pressure of a liquid increases with temperature. Because the boiling point of a liquid occurs when its vapor pressure is equal to the applied pressure (normally atmospheric pressure), the vapor pressure of a liquid equals 760 mmHg at its boiling point. The vapor pressure of a solid also varies with temperature. Because of this behavior, some solids can pass directly into the vapor phase without going through a liquid phase. This process is called **sublimation.** Because the vapor can be resolidified, the overall vaporization–solidification cycle can be used as a purification method. The purification can be successful only if the impurities have significantly lower vapor pressures than the material being sublimed.

16.1 VAPOR PRESSURE BEHAVIOR OF SOLIDS AND LIQUIDS

In Figure 16.1, vapor pressure curves for solid and liquid phases for two different substances are shown. Along lines *AB* and *DF,* the sublimation curves, the solid and vapor are at equilibrium. To the left of these lines, the solid phase exists, and to the right of these lines, the vapor phase is present. Along lines *BC* and *FG,* the liquid and vapor are at equilibrium. To the left of these lines, the liquid phase exists, and to the right, the vapor is present. The two substances vary greatly in their physical properties, as shown in Figure 16.1.

In the first case (Fig. 16.1A), the substance shows normal change-of-state behavior on being heated, going from solid to liquid to gas. The dashed line, which represents an atmospheric pressure of 760 mmHg, is located *above* the melting point *B* in Figure 16.1A. Thus, the applied pressure (760 mmHg) is *greater* than the vapor pressure of the solid–liquid phase at the melting point. Starting at *A,* as the temperature of the solid is raised, the vapor pressure increases along *AB* until the solid is observed to melt at *B.* At *B* the vapor pressures of *both* the solid and liquid are identical. As the temperature continues to rise, the vapor pressure will increase along *BC* until the liquid is observed to boil at *C.* The description given is for the "normal" behavior expected for a solid substance. All three states (solid, liquid, and gas) are observed sequentially during the change in temperature.

In the second case (Fig. 16.1B), the substance develops enough vapor pressure to vaporize completely at a temperature below its melting point. The substance shows a solid-to-gas transition only. The dashed line is now located *below* the melting point *F* of this substance. Thus, the applied pressure (760 mmHg) is *less* than the vapor pressure of the solid–liquid phase at the melting point. Starting at *D,* the vapor pressure of the solid rises as the temperature increases along line *DF.* However, the vapor pressure of the solid reaches atmospheric pressure (point *E*) *before* the melting point at *F* is attained. Therefore, sublimation occurs at *E.* No melting behavior will be observed at atmospheric pressure for

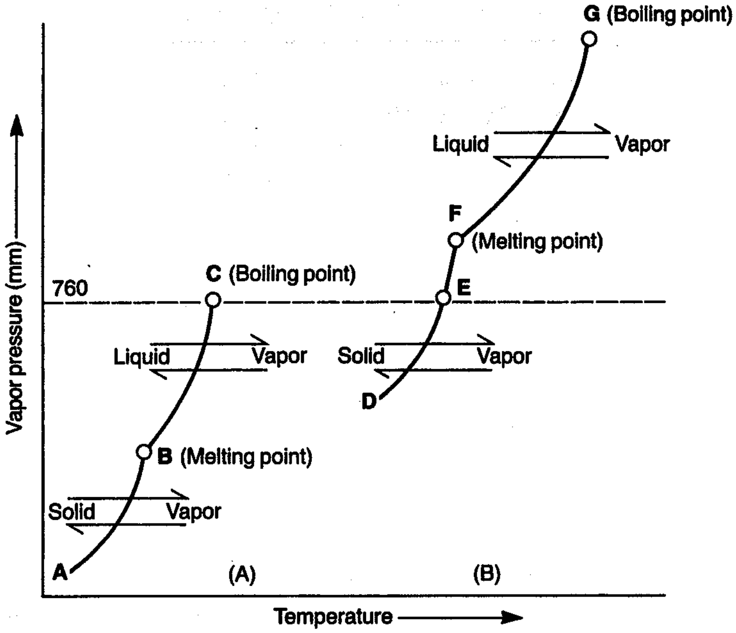

Figure 16.1 Vapor pressure curves for solids and liquids. (A) Substance shows normal solid to liquid to gas transitions at 760-mmHg pressure. (B) Substance shows a solid to gas transition at 760-mmHg pressure.

this substance. For a melting point to be reached and the behavior along line *FG* to be observed, an applied pressure greater than the vapor pressure of the substance at point *F* would be required. This could be achieved by using a sealed pressure apparatus.

The sublimation behavior just described is relatively rare for substances at atmospheric pressure. Several compounds exhibiting this behavior—carbon dioxide, perfluorocyclohexane, and hexachloroethane—are listed in Table 16.1. Notice that these compounds have vapor pressures *above* 760 mmHg at their melting points. In other words, their vapor pressures reach 760 mmHg below their melting points and they sublime rather than melt. Anyone trying to determine the melting point of hexachloroethane at atmospheric pressure will see vapor pouring from the end of the melting point tube! With a sealed capillary tube, the melting point of 186°C is observed.

16.2 SUBLIMATION BEHAVIOR OF SOLIDS

Sublimation is usually a property of relatively nonpolar substances that also have highly symmetrical structures. Symmetrical compounds have relatively high melting points and high vapor pressures. The ease with which a substance can escape from the solid state is determined by the strength of intermolecular forces. Symmetrical molecular structures have a relatively uniform distribution of electron density and a small dipole moment. A smaller dipole moment means a higher vapor pressure because of lower electrostatic attractive forces in the crystal.

TABLE 16.1 Vapor Pressures of Solids at Their Melting Points

Compound	Vapor Pressure of Solid at MP (mmHg)	Melting Point (°C)
Carbon dioxide	3876 (5.1 atm)	−57
Perfluorocyclohexane	950	59
Hexachloroethane	780	186
Camphor	370	179
Iodine	90	114
Naphthalene	7	80
Benzoic acid	6	122
p-Nitrobenzaldehyde	0.009	106

Solids sublime if their vapor pressures are greater than atmospheric pressure at their melting points. Some compounds with the vapor pressures at their melting points are listed in Table 16.1. The first three entries in the table were discussed in Section 16.1. At atmospheric pressure they would sublime rather than melt, as shown in Figure 16.1B.

The next four entries in Table 16.1 (camphor, iodine, naphthalene, and benzoic acid) exhibit typical change-of-state behavior (solid, liquid, and gas) at atmospheric pressure, as shown in Figure 16.1A. These compounds sublime readily under reduced pressure, however. Vacuum sublimation is discussed in Section 16.3.

Compared with many other organic compounds, camphor, iodine, and naphthalene have relatively high vapor pressures at relatively low temperatures. For example, they have a vapor pressure of 1 mmHg at 42, 39, and 53°C, respectively. Although this vapor pressure does not seem very large, it is high enough to lead, after a time, to **evaporation** of the solid from an open container. Mothballs (naphthalene and 1,4-dichlorobenzene) show this behavior. When iodine stands in a closed container over a period of time, you can observe movement of crystals from one part of the container to another.

Although chemists often refer to any solid–vapor transition as sublimation, the process described for camphor, iodine, and naphthalene is really an **evaporation** of a solid. Strictly speaking, a sublimation point is like a melting point or a boiling point. It is defined as the point at which the vapor pressure of the solid *equals* the applied pressure. Many liquids readily evaporate at temperatures far below their boiling points. It is, however, much less common for solids to evaporate. Solids that readily sublime (evaporate) must be stored in sealed containers. When the melting point of such a solid is being determined, some of the solid may sublime and collect toward the open end of the melting-point tube while the rest of the sample melts. To solve the sublimation problem, one seals the capillary tube or rapidly determines the melting point. It is possible to use the sublimation behavior to purify a substance. For example, at atmospheric pressure, camphor can be readily sublimed, just below its melting point at 175°C. At 175°C the vapor pressure of camphor is 320 mmHg. The vapor solidifies on a cool surface.

16.3 VACUUM SUBLIMATION

Many organic compounds sublime readily under reduced pressure. When the vapor pressure of the solid equals the applied pressure, sublimation occurs, and the behavior is identical to that shown in Figure 16.1B. The solid phase passes directly into the vapor phase. From the data given in Table 16.1, you should expect camphor, naphthalene, and benzoic acid to sublime at or below the respective applied pressures of 370, 7, and 6 mmHg. In principle, you can sublime *p*-nitrobenzaldehyde (last entry in the table), but it would not be practical because of the low applied pressure required.

16.4 SUBLIMATION METHODS

Sublimation can be used to purify solids. The solid is warmed until its vapor pressure becomes high enough for it to vaporize and condense as a solid on a cooled surface placed closely above. Three types of apparatus are illustrated in Figure 16.2. Since all of the parts fit securely, they are all capable of holding a vacuum. Chemists usually perform vacuum sublimations because most solids undergo the solid to gas transition only at low pressures. Reduction of pressure also helps to prevent thermal decomposition of substances that would require high temperatures to sublime at ordinary pressures. One end of a piece of rubber pressure tubing is attached to the apparatus and the other end is attached to an aspirator or to the house vacuum system or to a vacuum pump.

A sublimation is probably best carried out using one of the pieces of microscale equipment shown in Figures 16.2A and B. It is recommended that the laboratory instructor make available either one type or the other to be used on a communal basis. Each of the apparatus shown employs a central tube (closed on one end) filled with ice-cold water that serves as a condensing surface. The tube is filled with ice chips and a minimum of water. If the cooling water becomes warm before the sublimation is completed, a Pasteur pipet can be used to remove the warm water. The tube is then refilled with more ice-cold water. Warm water is undesirable because the vapor will not condense efficiently to form a solid as readily on a warm surface as it would on a cold surface. A poor recovery of solid results.

The apparatus shown in Figure 16.2C can be constructed from a side-arm test tube, a neoprene adapter, and a piece of glass tubing sealed at one end. Alternatively, a 15 × 125-mm test tube may be used instead of the piece of glass tubing. The test tube is inserted into a #2 neoprene adapter using a little water as a lubricant. All pieces must fit securely in order to be able to obtain a good vacuum and to avoid water being drawn into the side-arm test tube around the rubber adapter. In order to achieve an adequate seal, the side-arm test tube may need to be flared somewhat. A complete description of the assembly of the apparatus, similar to that shown in Figure 16.2C, is given in Experiment 5A, p. 73. Detailed procedures for using this apparatus are also given in that experiment.

A flame is the preferred heating device because the sublimation will occur more quickly than with other heating devices. The sublimation will be finished before the ice water warms significantly. The burner can be held by its cool base (not the hot barrel!) and moved up and down the sides of the outer tube to "chase" any solid that has formed

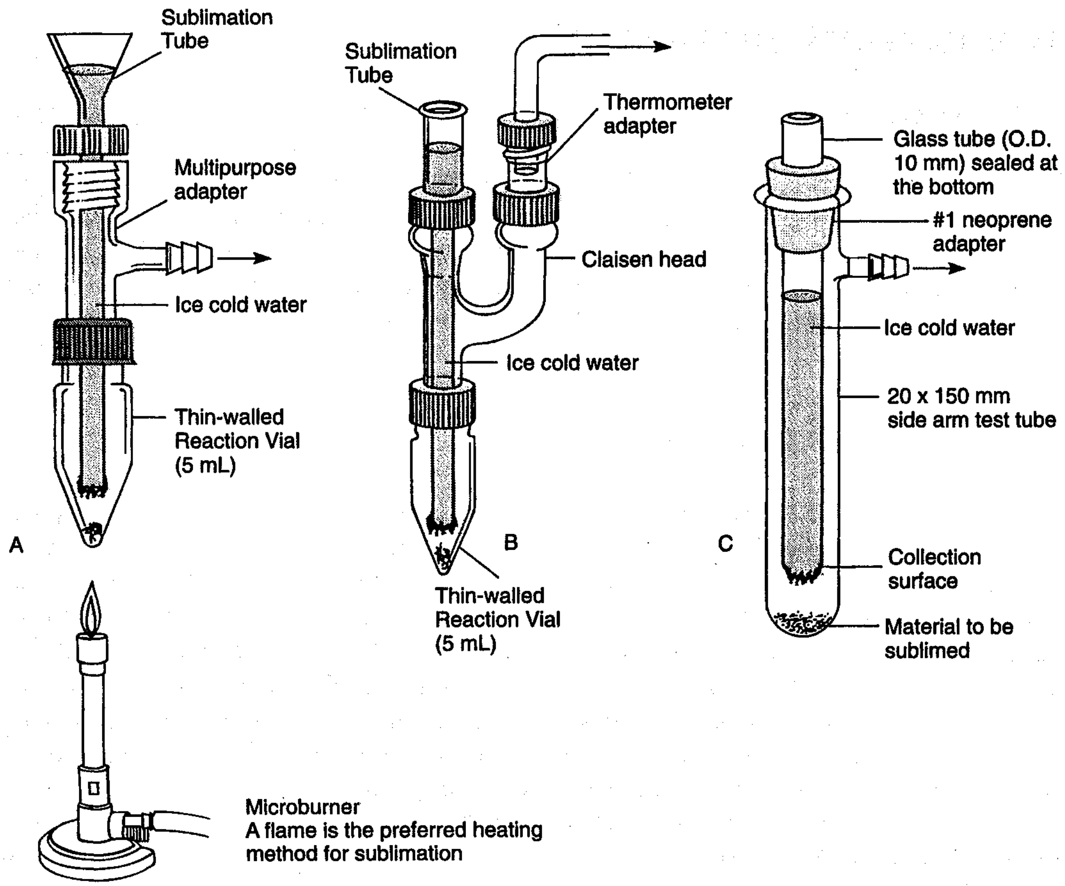
Sublimation Tube
Multipurpose adapter
Ice cold water
Thin-walled Reaction Vial (5 mL)
A
Sublimation Tube
Thermometer adapter
Claisen head
Ice cold water
B
Thin-walled Reaction Vial (5 mL)
C
Glass tube (O.D. 10 mm) sealed at the bottom
#1 neoprene adapter
Ice cold water
20 x 150 mm side arm test tube
Collection surface
Material to be sublimed
Microburner
A flame is the preferred heating method for sublimation

Figure 16.2 Sublimation apparatus.

on the sides toward the cold tube in the center. When using the apparatus shown in Figures 16.2A and B with a flame, you will need to use a thin-wall vial. Thicker glass can shatter when heated with a flame.

Remember that while performing a sublimation, it is important to keep the temperature below the melting point of the solid. After sublimation, the material that has collected on the cooled surface is recovered by removing the central tube (cold-finger) from the apparatus. Take care in removing this tube to avoid dislodging the crystals that have collected. The deposit of crystals is scraped from the inner tube with a spatula. If reduced pressure has been used, the pressure must be released carefully to keep a blast of air from dislodging the crystals.

16.5 ADVANTAGES OF SUBLIMATION

One advantage of sublimation is that no solvent is used and therefore none needs to be removed later. Sublimation also removes occluded material, like molecules of solvent, from the sublimed substance. For instance, caffeine (sublimes at 178°C, melts at 236°C) absorbs water gradually from the atmosphere to form a hydrate. During sublimation, this water is lost, and anhydrous caffeine is obtained. If too much solvent is present in a sample to be sublimed, however, instead of becoming lost, it condenses on the cooled surface and thus interferes with the sublimation.

Sublimation is a faster method of purification than crystallization but not as selective. Similar vapor pressures are often a factor in dealing with solids that sublime; consequently, little separation can be achieved. For this reason, solids are far more often purified by crystallization. Sublimation is most effective in removing a volatile substance from a nonvolatile compound, particularly a salt or other inorganic material. Sublimation is also effective in removing highly volatile bicyclic or other symmetrical molecules from less volatile reaction products. Examples of volatile bicyclic compounds are borneol, isoborneol, and camphor.

PROBLEMS

1. Why is solid carbon dioxide called dry ice? How does it differ from solid water in behavior?
2. Under what conditions can you have *liquid* carbon dioxide?
3. A solid substance has a vapor pressure of 800 mmHg at its melting point (80°C). Describe how the solid behaves as the temperature is raised from room temperature to 80°C, while the atmospheric pressure is held constant at 760 mmHg.
4. A solid substance has a vapor pressure of 100 mmHg at the melting point (100°C). Assuming an atmospheric pressure of 760 mmHg, describe the behavior of this solid as the temperature is raised from room temperature to its melting point.
5. A substance has a vapor pressure of 50 mmHg at the melting point (100°C). Describe how you would experimentally sublime this substance.

TECHNIQUE 17

Polarimetry

17.1 NATURE OF POLARIZED LIGHT

Light has a dual nature because it shows properties of both waves and particles. The wave nature of light can be demonstrated by two experiments: polarization and interference. Of the two, polarization is the more interesting to organic chemists, because they can take advantage of polarization experiments to learn something about the structure of an unknown molecule.

Ordinary white light consists of wave motion in which the waves have a variety of wavelengths and vibrate in all possible planes perpendicular to the direction of propagation. Light can be made to be **monochromatic** (of one wavelength or color) by using fil-

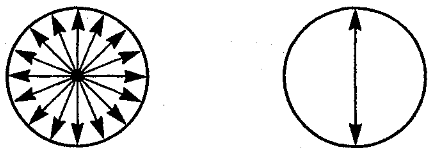

Figure 17.1 Ordinary versus plane-polarized light.

ters or special light sources. Frequently, a sodium lamp (sodium D line = 5893 Å) is used. Although the light from this lamp consists of waves of only one wavelength, the individual light waves still vibrate in all possible planes perpendicular to the beam. If we imagine that the beam of light is aimed directly at the viewer, ordinary light can be represented by showing the edges of the planes oriented randomly around the path of the beam, as in the left part of Figure 17.1.

A Nicol prism, which consists of a specially prepared crystal of Iceland spar (or calcite), has the property of serving as a screen that can restrict the passage of light waves. Waves that are vibrating in one plane are transmitted; those in all other planes are rejected (either refracted in another direction or absorbed). The light that passes through the prism is called **plane-polarized light,** and it consists of waves that vibrate in only one plane. A beam of plane-polarized light aimed directly at the viewer can be represented by showing the edges of the plane oriented in one particular direction, as in the right portion of Figure 17.1.

Iceland spar has the property of **double refraction,** that is, it can split, or doubly refract, an entering beam of ordinary light into two separate emerging beams of light. Each of the two emerging beams (labeled A and B in Fig. 17.2) has only a single plane of vibration, and the plane of vibration in Beam A is perpendicular to the plane of Beam B. In other words, the crystal has separated the incident beam of ordinary light into two beams of plane-polarized light, with the plane of polarization of Beam A perpendicular to the plane of Beam B.

To generate a single beam of plane-polarized light, one can take advantage of the double-refracting property of Iceland spar. A Nicol prism, invented by the Scottish physicist William Nicol, consists of two crystals of Iceland spar cut to specified angles and ce-

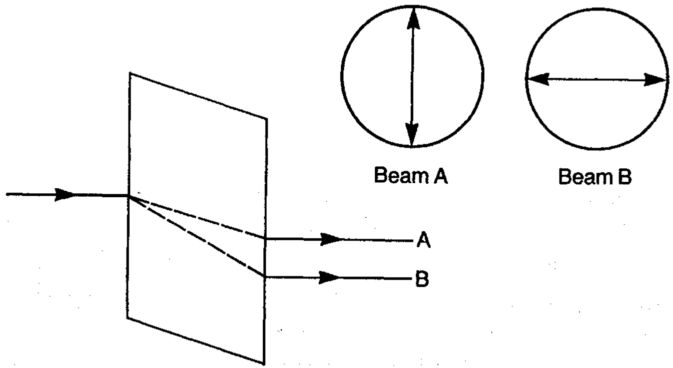

Figure 17.2 Double refraction.

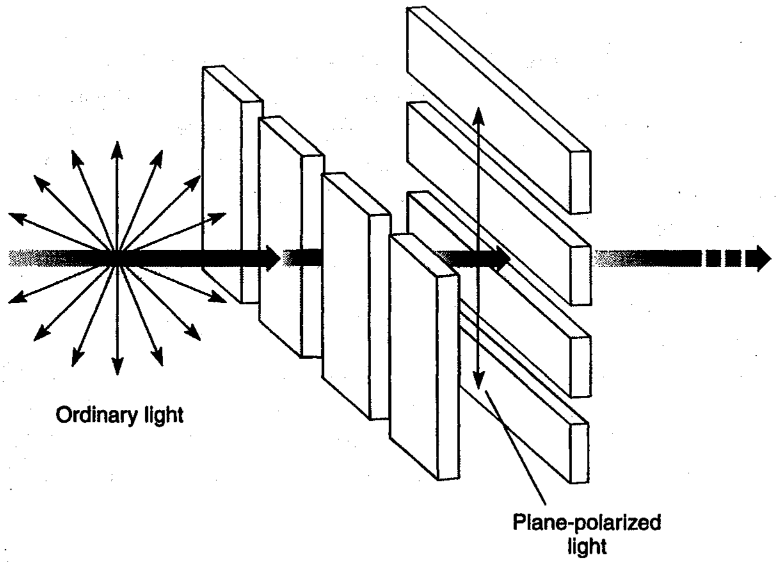

Figure 17.3 The picket-fence analogy.

mented by Canada balsam. This prism transmits one of the two beams of plane-polarized light while reflecting the other at a sharp angle so that it does not interfere with the transmitted beam. Plane-polarized light can also be generated by a Polaroid filter, a device invented by E. H. Land, an American physicist. Polaroid filters consist of certain types of crystals, embedded in transparent plastic and capable of producing plane-polarized light.

After passing through a first Nicol prism, plane-polarized light can pass through a second Nicol prism, but only if the second prism has its axis oriented so that it is *parallel* to the incident light's plane of polarization. Plane-polarized light is *absorbed* by a Nicol prism that is oriented so that its axis is *perpendicular* to the incident light's plane of polarization. These situations can be illustrated by the picket-fence analogy, as shown in Figure 17.3. Plane-polarized light can pass through a fence whose slats are oriented in the proper direction but is blocked out by a fence whose slats are oriented perpendicularly.

An **optically active substance** is one that interacts with polarized light to rotate the plane of polarization through some angle α. Figure 17.4 illustrates this phenomenon.

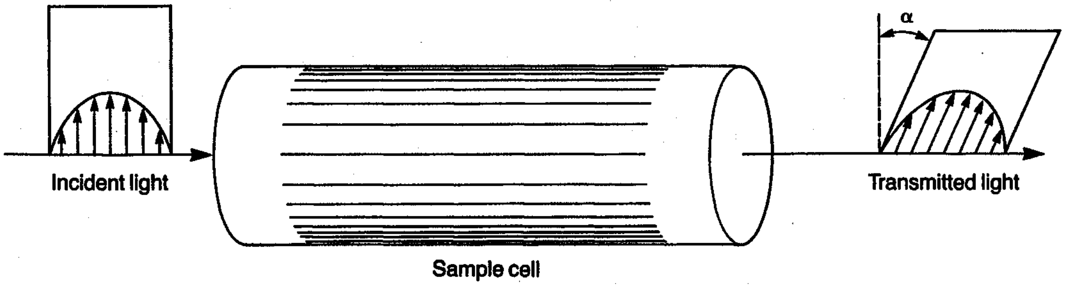

Figure 17.4 Optical activity.

17.2 THE POLARIMETER

An instrument called a **polarimeter** is used to measure the extent to which a substance interacts with polarized light. A schematic diagram of a polarimeter is shown in Figure 17.5. The light from the source lamp is polarized by being passed through a fixed Nicol prism, called a polarizer. This light passes through the sample, with which it may or may not interact to have its plane of polarization rotated in one direction or the other. A second, rotatable Nicol prism, called the analyzer, is adjusted to allow the maximum amount of light to pass through. The number of degrees and the direction of rotation required for this adjustment are measured to give the **observed rotation** α.

So that data determined by several persons under different conditions can be compared, a standardized means of presenting optical rotation data is necessary. The most common way of presenting such data is by recording the **specific rotation** $[\alpha]_\lambda^t$, which has been corrected for differences in concentration, cell path length, temperature, solvent, and wavelength of the light source. The equation defining the specific rotation of a compound in solution is

$$[\alpha]_\lambda^t = \frac{\alpha}{cl}$$

where α = observed rotation in degrees, c = concentration in grams per milliliter of solution, l = length of sample tube in decimeters, λ = wavelength of light (usually indicated as "D," for the sodium D line), and t = temperature in degrees Celsius. For pure liquids, the density d of the liquid in grams per milliliter replaces c in the preceding formula. You may occasionally want to compare compounds of different molecular weights, so a **molecular rotation,** based on moles instead of grams, is more convenient than a specific rotation. The molecular rotation M_λ^t is derived from the specific rotation $[\alpha]_\lambda^t$ by

$$M_\lambda^t = \frac{[\alpha_\lambda^t] \times \text{Molecular weight}}{100}$$

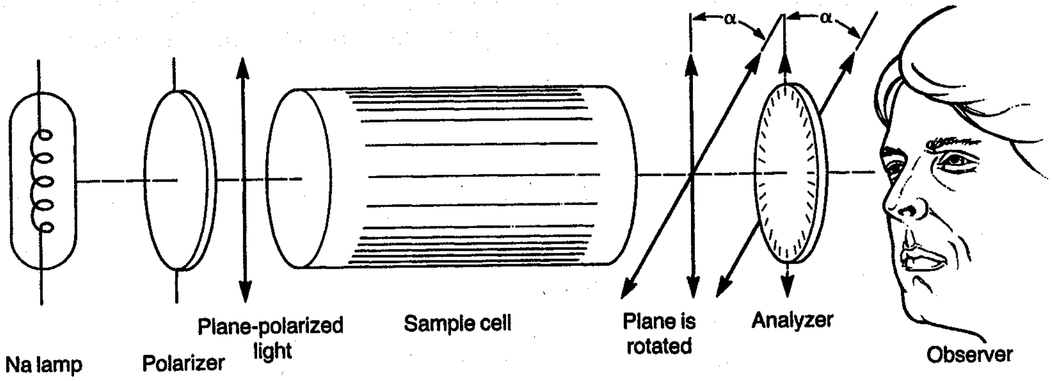

Figure 17.5 Schematic diagram of a polarimeter.

Usually, measurements are made at 25°C with the sodium D line as a light source; consequently, specific rotations are reported as $[\alpha]_D^{25}$.

Polarimeters which are now available incorporate electronics to determine the angle of rotation of chiral molecules. These instruments are essentially automatic. The only real difference between an automatic polarimeter and a manual one is that a light detector replaces the eye. No visual observation of any kind is made with an automatic instrument. A microprocessor adjusts the analyzer until the light reaching the detector is at a minimum. The angle of rotation is displayed digitally in an LCD window, including the sign of rotation. The simplest instrument is equipped with a sodium lamp that gives rotations based on the D line of sodium (589 nm). More expensive instruments use a tungsten lamp and filters so that wavelengths can be varied over a range of values. Using the latter instrument, a chemist can observe rotations at different wavelengths.

17.3 THE SAMPLE CELLS

It is important for the solution whose optical rotation is to be determined to contain no suspended particles of dust or dirt that might disperse the incident polarized light. Therefore, you must clean the sample cell carefully and make certain that there are no air bubbles trapped in the path of the light. The sample cells contain an enlarged ring near one end, in which the air bubbles may be trapped. Sample cells are available in various lengths, with 0.5 and 1.0 dm being the most common.

The sample is generally prepared by dissolving 0.1–0.5 g of the substance to be studied in 25 mL of solvent, usually water, ethanol, or methylene chloride (chloroform was used in the past). If the specific rotation of the substance is very high or very low, you may need to make the concentration of the solution respectively lower or higher, but usually this is determined after first trying a concentration range such as that suggested previously. The sample cell shown in Figure 17.6 is filled with solution. It is then tilted upward and tapped until the air bubbles move into the enlarged ring. It is important not to get fingerprints on the glass endplate in reassembling the cell.

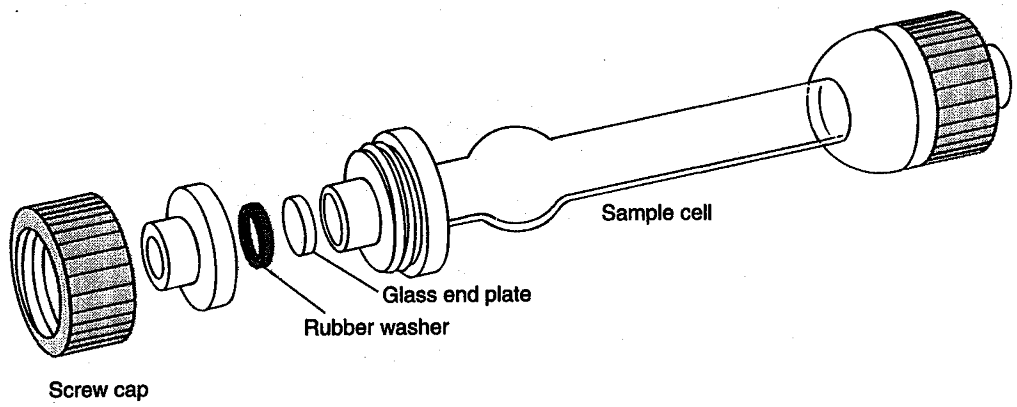

Figure 17.6 Polarimeter cell assembly.

17.4 OPERATION OF THE POLARIMETER

The procedures given here for preparing the cells and for operating the instrument are appropriate for the Zeiss polarimeter with the circular scale; other models of polarimeter are operated similarly. It is necessary before beginning the experiments to turn the power switch to the ON position and wait 5–10 minutes until the sodium lamp is properly warmed.

The instrument should be checked initially by making a zero reading with a sample cell filled only with solvent. If the zero reading does not correspond with the zero-degree calibration mark, then the difference in readings must be used to correct all subsequent readings. The reading is determined by laying the sample tube in the cradle, enlarged end up (making sure that there are no air bubbles in the light path), closing the cover, and turning the knob until the proper angle of the analyzer is reached. Most instruments, including the Zeiss polarimeter, are of the double-field type, in which the eye sees a split field whose sections must be matched in light intensity. The value of the angle through which the plane of polarized light has been rotated (if any) is read directly from the scale that can be seen through the eyepiece directly below the split-field image. Figure 17.7 shows how this split field might appear.

The cell containing the solution of the sample is then placed in the polarimeter, and the observed angle of rotation is measured in the same way. Be sure to record not only the numerical value of the angle of rotation in degrees but also the direction of rotation. Rotations clockwise are due to **dextrorotatory** substances and are indicated by the sign "+." Rotations counterclockwise are due to **levorotatory** substances and are indicated by the sign "−." It is best, in making a determination, to take several readings, including readings for which the actual value was approached from both sides. In other words, where the actual reading might be +75°, first approach this reading upward from a reading near zero; *on the next measurement approach this reading downward from an angle greater* than +75°. Duplicating readings and approaching the observed value from both sides reduce errors. The readings are then averaged to get the observed rotation α. This rotation is then corrected by the appropriate factors, according to the formulas in Section 17.2, to provide the specific rotation. The specific rotation is always reported as a function of temperature, indicating the wavelength by "D" if a sodium lamp is used and reporting the concentration and solvent used. For example $[\alpha]_D^{20} = +43.8°$ ($c = 7.5$ g/100 mL in absolute ethanol).

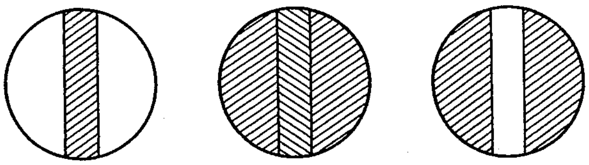

Adjustment incorrect

Adjustment correct

Adjustment incorrect

Figure 17.7 Split-field image in the polarimeter.

17.5 OPTICAL PURITY

When you prepare a sample of an enantiomer by a resolution method, the sample is not always 100% of a single enantiomer. It frequently is contaminated by residual amounts of the opposite stereoisomer. If you know the amount of each enantiomer in a mixture, you can calculate the **optical purity.** Some chemists prefer to use the term **enantiomeric excess (ee)** rather than optical purity. The two terms can be used interchangeably. The percentage enantiomeric excess or optical purity is calculated as follows:

$$\%\ \text{Optical purity} = \frac{\text{moles one enantiomer} - \text{moles of other enantiomer}}{\text{total moles of both enantiomers}} \times 100$$

$$\%\ \text{Optical purity} = \%\ \text{Enantiomeric excess (ee)}$$

Often, it is difficult to apply the equation shown above because you do not know the exact amount of each enantiomer present in a mixture. It is far easier to calculate the optical purity (enantiomeric excess) by using the observed specific rotation of the mixture and dividing it by the specific rotation of the pure enantiomer. Values for the pure enantiomers can sometimes be found in literature sources.

$$\%\ \text{Optical purity} = \%\ \text{Enantiomeric excess} = \frac{\text{Observed specific rotation}}{\text{Specific rotation of pure enantiomer}} \times 100$$

This latter equation only holds true for mixtures of two chiral molecules that are mirror images of each other (enantiomers). If some other chiral substance is present in the mixture as an impurity, then the actual optical purity will deviate from the value calculated.

In a racemic (±) mixture, there is no excess enantiomer and the optical purity (enantiomeric excess) is zero; in a completely resolved material, the optical purity (enantiomeric excess) is 100%. A compound that is $x\%$ optically pure contains $x\%$ of one enantiomer and $(100 - x)\%$ of a racemic mixture.

Once the optical purity (enantiomeric excess) is known, the relative percentages of the each of the enantiomers can be calculated easily. If the predominant form in the impure, optically active mixture is assumed to be the (+) enantiomer, the percentage of the (+) enantiomer is

$$\left[x + \left(\frac{100 - x}{2}\right)\right]\%$$

and the percentage of the (−) enantiomer is $[(100 - x)/2]\%$. The relative percentages of (+) and (−) forms in a partially resolved mixture of enantiomers can be calculated as shown next. Consider a partially resolved mixture of camphor enantiomers. The specific rotation for pure (+)-camphor is +43.8° in absolute ethanol, but the mixture shows a specific rotation of +26.3°.

$$\text{Optical purity} = \frac{+26.3^\circ}{+43.8^\circ} \times 100 = 60\% \text{ optically pure}$$

$$\%\ (+)\text{ enantiomer} = 60 + \left(\frac{100 - 60}{2}\right) = 80\%$$

$$\%\ (-)\text{ enantiomer} = \left(\frac{100 - 60}{2}\right) = 20\%$$

Notice that the difference between these two calculated values equals the optical purity or enantiomeric excess (ee).

PROBLEMS

1. Calculate the specific rotation of a substance that is dissolved in a solvent (0.4 g/mL) and that has an observed rotation of −10° as determined with a 0.5-dm cell.
2. Calculate the observed rotation for a solution of a substance (2.0 g/mL) that is 80% optically pure. A 2-dm cell is used. The specific rotation for the optically pure substance is +20°.
3. What is the optical purity of a partially racemized product if the calculated specific rotation is −8° and the pure enantiomer has a specific rotation of −10°? Calculate the percentage of each of the enantiomers in the partially racemized product.

TECHNIQUE 18

Refractometry

The **refractive index** is a useful physical property of liquids. Often, a liquid can be identified from a measurement of its refractive index. The refractive index can also provide a measure of the purity of the sample being examined. This is accomplished by comparing the experimentally measured refractive index with the value reported in the literature for an ultrapure sample of the compound. The closer the measured sample's value to the literature value, the purer the sample.

18.1 THE REFRACTIVE INDEX

The refractive index has as its basis the fact that light travels at a different velocity in condensed phases (liquids, solids) than in air. The refractive index n is defined as the ratio of the velocity of light in air to the velocity of light in the medium being measured:

$$n = \frac{V_{\text{air}}}{V_{\text{liquid}}} = \frac{\sin \theta}{\sin \phi}$$

It is not difficult to measure the ratio of the velocities experimentally. It corresponds to ($\sin \theta / \sin \phi$), where θ is the angle of incidence for a beam of light striking the surface of the medium and ϕ is the angle of refraction of the beam of light *within* the medium. This is illustrated in Figure 18.1.

The refractive index for a given medium depends on two variable factors. First, it is *temperature*-dependent. The density of the medium changes with temperature; hence, the speed of light in the medium also changes. Second, the refractive index is *wavelength*-dependent. Beams of light with different wavelengths are refracted to different extents in the same medium and give different refractive indices for that medium. It is usual to report refractive indices measured at 20°C, with a sodium discharge lamp as the source of illumination. The sodium lamp gives off yellow light of 589-nm wavelength, the so-called sodium D line. Under these conditions, the refractive index is reported in the following form

$$n_D^{20} = 1.4892$$

The superscript indicates the temperature, and the subscript indicates that the sodium D line was used for the measurement. If another wavelength is used for the determination, the D is replaced by the appropriate value, usually in nanometers (1 nm = 10^{-9} m).

Notice that the hypothetical value reported above has four decimal places. It is easy to determine the refractive index to within several parts in 10,000. Therefore, n_D is a very

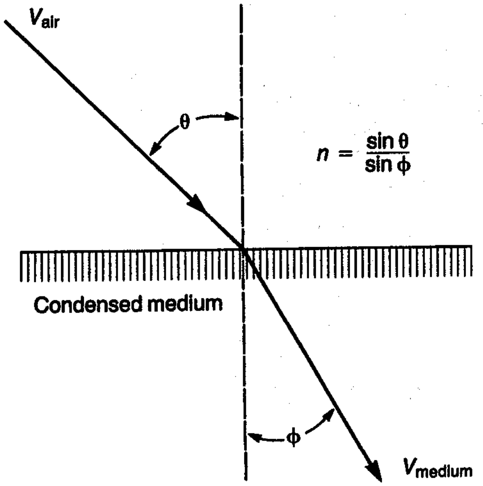

Figure 18.1 The refractive index.

accurate physical constant for a given substance and can be used for identification. However, it is sensitive to even small amounts of impurity in the substance measured. Unless the substance is purified *extensively*, you will not usually be able to reproduce the last two decimal places given in a handbook or other literature source. Typical organic liquids have refractive index values between 1.3400 and 1.5600.

18.2 THE ABBÉ REFRACTOMETER

The instrument used to measure the refractive index is called a **refractometer.** Although many styles of refractometer are available, by far the most common instrument is the Abbé refractometer. This style of refractometer has the following advantages:

1. White light may be used for illumination; the instrument is compensated, however, so that the index of refraction obtained is actually that for the sodium D line.
2. The prisms can be temperature-controlled.
3. Only a small sample is required (a few drops of liquid using the standard method, or about 5 μL using a modified technique).

A common type of Abbé refractometer is shown in Figure 18.2.

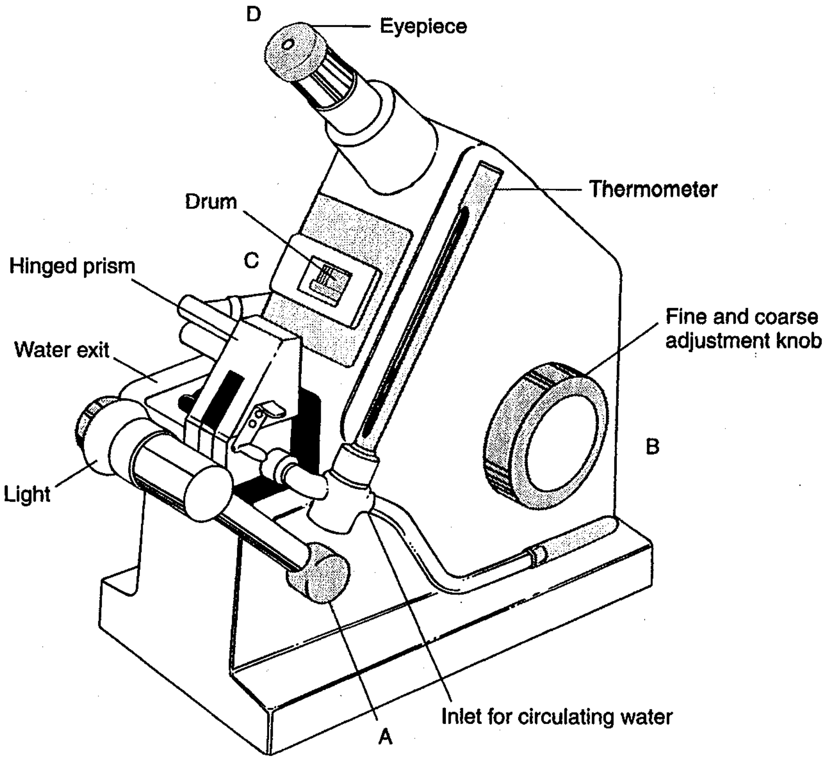

Figure 18.2 Abbé refractometer (Bausch and Lomb Abbé 3L).

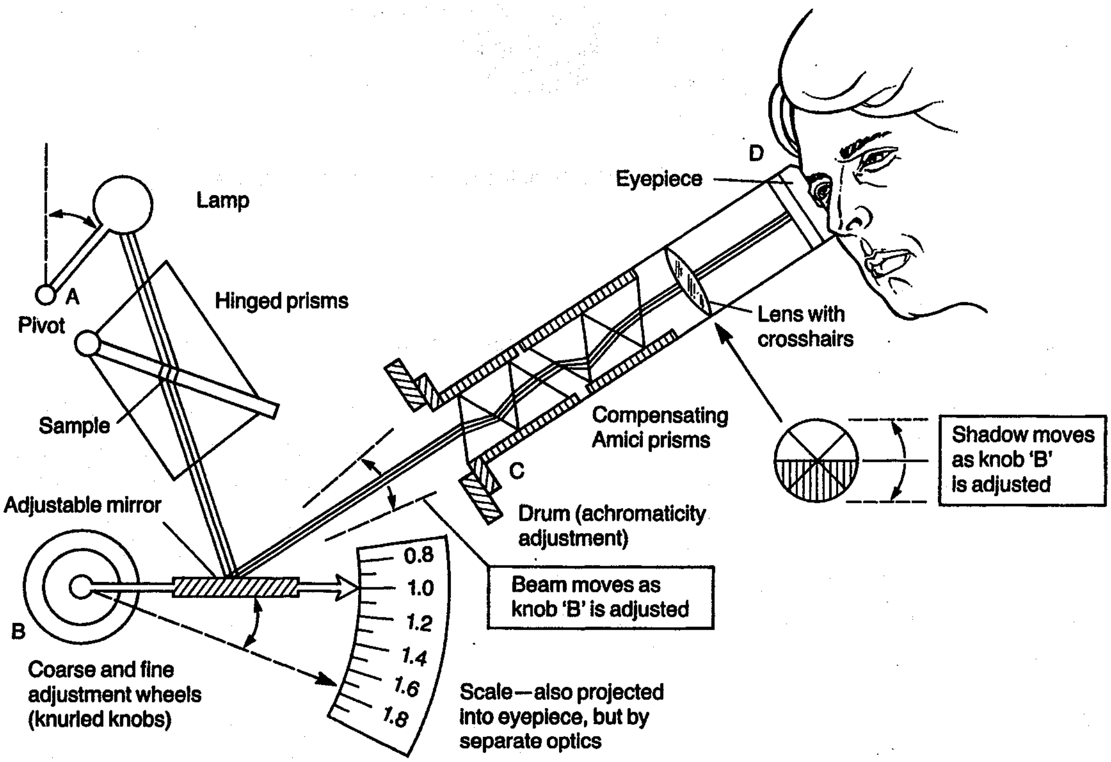
Lamp
A
Pivot
Hinged prisms
Sample
Adjustable mirror
B
Coarse and fine
adjustment wheels
(knurled knobs)
0.8
1.0
1.2
1.4
1.6
1.8
Scale—also projected
into eyepiece, but by
separate optics
Beam moves as
knob 'B' is adjusted
Drum (achromaticity
adjustment)
C
Compensating
Amici prisms
Eyepiece
D
Lens with
crosshairs
Shadow moves
as knob 'B'
is adjusted

Figure 18.3 Simplified diagram of a refractometer.

The optical arrangement of the refractometer is very complex; a simplified diagram of the internal workings is given in Figure 18.3. The letters *A, B, C,* and *D* label corresponding parts in both Figures 18.2 and 18.3. A complete description of refractometer optics is too difficult to attempt here, but Figure 18.3 gives a simplified diagram of the essential operating principles.

Using the standard method, the sample to be measured is introduced between the two prisms. If it is a free-flowing liquid, it may be introduced into a channel along the side of the prisms, injected from a Pasteur pipet. If it is a viscous sample, the prisms must be opened (they are hinged) by lifting the upper one; a few drops of liquid are applied to the lower prism with a Pasteur pipet or a wooden applicator. If a Pasteur pipet is used, take care not to touch the prisms, because they become scratched easily. When the prisms are closed, the liquid should spread evenly to make a thin film. With highly volatile samples, the remaining operations must be performed rapidly. Even when the prisms are closed, evaporation of volatile liquids can readily occur.

Next, you turn on the light and look into the eyepiece *D*. The hinged lamp is adjusted to give the maximum illumination to the visible field in the eyepiece. The light rotates at pivot *A*.

Rotate the coarse and fine adjustment knobs at *B* until the dividing line between the light and dark halves of the visual field coincide with the center of the cross hairs

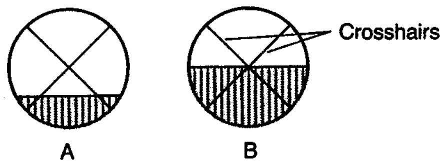

Figure 18.4 (A) Refractometer incorrectly adjusted. (B) Correct adjustment.

(Fig. 18.4). If the cross hairs are not in sharp focus, adjust the eyepiece to focus them. If the horizontal line dividing the light and dark areas appears as a colored band, as in Figure 18.5, the refractometer shows **chromatic aberration** (color dispersion). This can be adjusted with the knob labeled *C* drum. This knurled knob rotates a series of prisms, called Amici prisms, that color-compensate the refractometer and cancel out dispersion. Adjust the knob to give a sharp, uncolored division between the light and dark segments. When you have adjusted everything correctly (as in Fig. 18.4B), read the refractive index. In the instrument described here, press a small button on the left side of the housing to make the scale visible in the eyepiece. In other refractometers, the scale is visible at all times, frequently through a separate eyepiece.

Occasionally, the refractometer will be so far out of adjustment that it may be difficult to measure the refractive index of an unknown. When this happens, it is wise to place a pure sample of known refractive index in the instrument, set the scale to the correct value of refractive index, and adjust the controls for the sharpest line possible. Once this is done, it is easier to measure an unknown sample. It is especially helpful to perform this procedure prior to measuring the refractive index of a highly volatile sample.

There are many styles of refractometer, but most have adjustments similar to those described here.

In the procedure described above, several drops of liquid are required to obtain the refractive index. In some experiments, you may not have enough sample to use this standard method. It is possible to modify the procedure so that a reasonably accurate refractive index can be obtained on about 5 μL of liquid. Instead of placing the sample directly onto the prism, the sample is applied to a small piece of lens paper. The lens paper can be

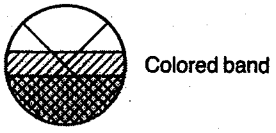

Figure 18.5 Refractometer showing chromatic aberration (color dispersion). The dispersion is incorrectly adjusted.

conveniently cut with a hand-held paper punch,[1] and the paper disc (0.6 cm diameter) is placed in the center of the bottom prism of the refractometer. To avoid scratching the prism, a forceps or tweezers with plastic tips should be used to handle the disc. About 5 μL of liquid is carefully placed on the lens paper using a microliter syringe. After closing the prisms, the refractometer is adjusted as described above and the refractive index is read. With this method, the horizontal line dividing the light and dark areas may not be as sharp as it is in the absence of the lens paper. It may also be impossible to eliminate color dispersion completely. Nonetheless, the refractive index values determined by this method are usually within 10 parts in 10,000 of the values determined by the standard procedure.

18.3 CLEANING THE REFRACTOMETER

In using the refractometer, you should always remember that if the prisms are scratched, the instrument will be ruined.

Do not touch the prisms with any hard object.

This admonition includes Pasteur pipets and glass rods.

When measurements are completed, the prisms should be cleaned with ethanol or petroleum ether. *Soft* tissues are moistened with the solvent, and the prisms are wiped *gently*. When the solvent has evaporated from the prism surfaces, the prisms should be locked together. The refractometer should be left with the prisms closed to avoid collection of dust in the space between them. The instrument should also be turned off when it is no longer in use.

18.4 TEMPERATURE CORRECTIONS

Most refractometers are designed so that circulating water at a constant temperature can maintain the prisms at 20°C. If this temperature-control system is not used, or if the water is not at 20°C, a temperature correction must be made. Although the magnitude of the temperature correction may vary from one class of compound to another, a value of 0.00045 per degree Celsius is a useful approximation for most substances. The index of refraction of a substance *decreases* with *increasing* temperature. Therefore, add the correction to the observed n_D value for temperatures higher than 20°C and subtract it for temperatures lower than 20°C. For example, the reported n_D value for nitrobenzene is 1.5529. One would observe a value at 25°C of 1.5506. The temperature correction would be made as follows:

$$n_D^{20} = 1.5506 + 5(0.00045) = 1.5529$$

---

[1] In order to cut the lens paper more easily, place several sheets between two pieces of heavier paper, such as that used for file folders.

PROBLEMS

1. A solution consisting of isobutyl bromide and isobutyl chloride is found to have a refractive index of 1.3931 at 20°C. The refractive indices at 20°C of isobutyl bromide and isobutyl chloride are 1.4368 and 1.3785, respectively. Determine the molar composition (in percent) of the mixture by assuming a linear relation between the refractive index and the molar composition of the mixture.

2. The refractive index of a compound at 16°C is found to be 1.3982. Correct this refractive index to 20°C.

TECHNIQUE 19

Preparation of Samples for Spectroscopy

Modern organic chemistry requires sophisticated scientific instruments. Most important among these instruments are the two spectroscopic instruments: the infrared (IR) and nuclear magnetic resonance (NMR) spectrometers. These instruments are indispensable to the modern organic chemist in proving the structures of unknown substances, in verifying that reaction products are indeed the predicted ones, and in characterizing organic compounds. The theory underlying these instruments can be found in most standard lecture textbooks in organic chemistry. Additional information, including correlation charts, to help in interpreting spectra are found in this textbook in Appendix 3 (Infrared Spectroscopy), Appendix 4 (Nuclear Magnetic Resonance Spectroscopy) and Appendix 5 (Carbon-13 Nuclear Magnetic Resonance Spectroscopy). This technique chapter concentrates on the preparation of samples for these spectroscopic methods. Part A covers techniques used in infrared spectroscopy, and Part B describes sample preparation for nuclear magnetic resonance spectroscopy.

Part A. Infrared Spectroscopy

19.1 INTRODUCTION

To determine the infrared spectrum of a compound, one must place it in a sample holder or cell. In infrared spectroscopy this immediately poses a problem. Glass, quartz, and plastics absorb strongly throughout the infrared region of the spectrum (any compound with covalent bonds usually absorbs) and cannot be used to construct sample cells. Ionic substances must be used in cell construction. Metal halides (sodium chloride, potassium bromide, silver chloride) are commonly used for this purpose.

Sodium Chloride Cells. Single crystals of sodium chloride are cut and polished to give plates that are transparent throughout the infrared region. These plates are then used to fabricate cells that can be used to hold *liquid* samples. Because sodium chloride is water-soluble, samples must be *dry* before a spectrum can be obtained. In general, sodium chloride plates are preferred for most applications involving liquid samples.

Silver Chloride Cells. Cells may be constructed of silver chloride. These plates may be used for *liquid* samples that contain small amounts of water, because silver chloride is water-insoluble. However, because water absorbs in the infrared region, as much water as possible should be removed, even when using silver chloride. Silver chloride plates must be stored in the dark, and they cannot be used with compounds that have an amino functional group. Amines react with silver chloride.

Solid Samples. A *solid* sample is usually held in place by making a potassium bromide pellet that contains a small amount of dispersed compound. A solid sample may also be suspended in mineral oil, which absorbs only in specific regions of the infrared spectrum. Another method is to dissolve the solid compound in an appropriate solvent and place the solution between two sodium chloride or silver chloride plates.

19.2 LIQUID SAMPLES—NaCl PLATES

The simplest method of preparing the sample, if it is a liquid, is to place a thin layer of the liquid between two sodium chloride plates that have been ground flat and polished. This is the method of choice when you need to determine the infrared spectrum of a pure liquid. A spectrum determined by this method is referred to as a **neat** spectrum. No solvent is used. The polished plates are expensive because they are cut from a large, single crystal of sodium chloride. Salt plates break easily, and they are water-soluble.

Preparing the Sample. Obtain two sodium chloride plates and a holder from the desiccator where they are stored. Moisture from fingers will mar and occlude the polished surfaces. Samples that contain water will destroy the plates.

The plates should be touched only on their edges. Be certain to use a sample that is dry or free from water.

Add a drop of the liquid to the surface of one plate, then place the second plate on top. The pressure of this second plate causes the liquid to spread out and form a thin, capillary film between the two plates. As shown in Figure 19.1, set the plates between the bolts in a holder and place the metal ring carefully on the salt plates. Use the hex nuts to hold the salt plates in place.

Do not overtighten the nuts or the salt plates will cleave or split.

Tighten the nuts firmly, but do not use any force at all to turn them. Spin them with the fingers until they stop; then turn them just another fraction of a full turn, and they will be tight enough. If the nuts have been tightened carefully, you should observe a transparent film of sample (a uniform wetting of the surface). If a thin film has not been obtained, either loosen one or more of the hex nuts and adjust them so that a uniform film is obtained or add more sample.

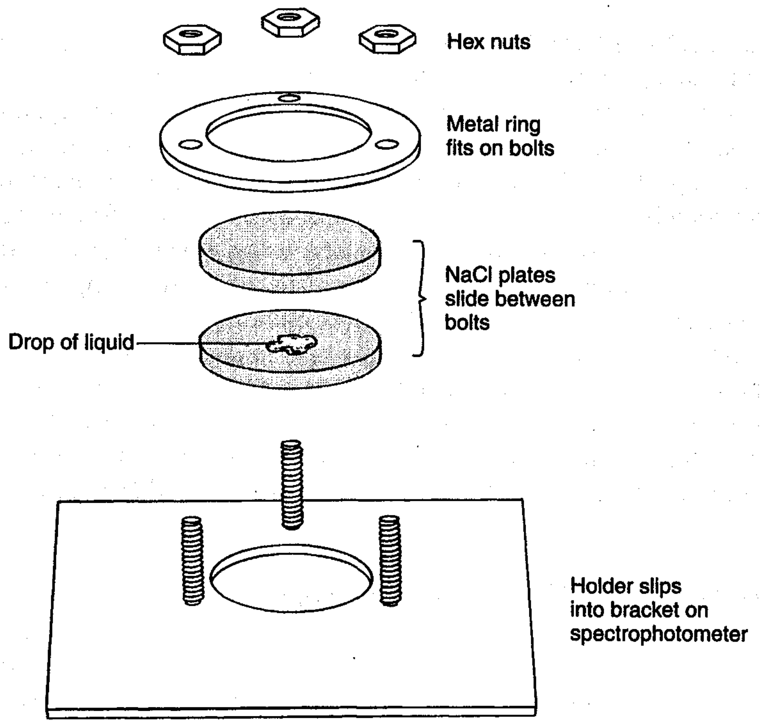
Hex nuts
Metal ring fits on bolts
NaCl plates slide between bolts
Drop of liquid
Holder slips into bracket on spectrophotometer

Figure 19.1 Salt plates and holder.

The thickness of the film obtained between the two plates is a function of two factors: (1) the amount of liquid placed on the first plate (one drop, two drops, etc.) and (2) the pressure used to hold the plates together. If more than one or two drops of liquid have been used, it will probably be too much, and the resulting spectrum will show strong absorptions that are off the scale of the chart paper. Only enough liquid to wet both surfaces is needed.

If the sample has a very low viscosity, you may find that the capillary film is too thin to produce a good spectrum. Another problem you may find is that the liquid is so volatile that the sample evaporates before the spectrum can be determined. In these cases, you may need to use the silver chloride plates discussed in Section 19.3, or a solution cell described in Section 19.5. Often, you can obtain a reasonable spectrum by assembling the cell quickly and running the spectrum before the sample runs out of the salt plates or evaporates.

Determining the Infrared Spectrum. Slide the holder into the slot in the sample beam of the spectrophotometer. Determine the spectrum according to the instructions provided by your instructor. In some cases, your instructor may ask you to calibrate your spectrum. If this is the case, refer to Section 19.8.

Cleaning and Storing the Salt Plates. Once the spectrum has been determined, demount the holder and rinse the salt plates with methylene chloride (or *dry* acetone). (Keep the plates away from water!) Use a soft tissue, moistened with the solvent, to wipe the plates. If some of your compound remains on the plates, you may observe a shiny surface. Continue to clean the plates with solvent until no more compound remains on the surfaces of the plates.

Caution: Avoid direct contact with methylene chloride. Return the salt plates and holder to the desiccator for storage. Use gloves.

19.3 LIQUID SAMPLES—AgCl PLATES

The mini-cell[1] shown in Figure 19.2 may also be used with liquids. The cell assembly consists of a two-piece threaded body, an O-ring, and two silver chloride plates. The plates are flat on one side, and there is a circular depression (0.025 mm or 0.10 mm deep) on the other side of the plate. The advantages of using silver chloride plates are that they may be used with wet samples or solutions. A disadvantage is that silver chloride darkens when exposed to light for extended periods. They also scratch more easily than salt plates and react with amines.

Preparing the Sample. Silver chloride plates should be handled in the same way as salt plates. Unfortunately, they are smaller and thinner (about like a contact lens) than salt plates, and care must be taken to not lose them! Remove them from the light-tight container with care. It is difficult to tell which side of the plate has the slight circular depression. Your instructor may have etched a letter on each plate to indicate which side is the flat one. If you are going to determine the infrared spectrum of a pure liquid (neat

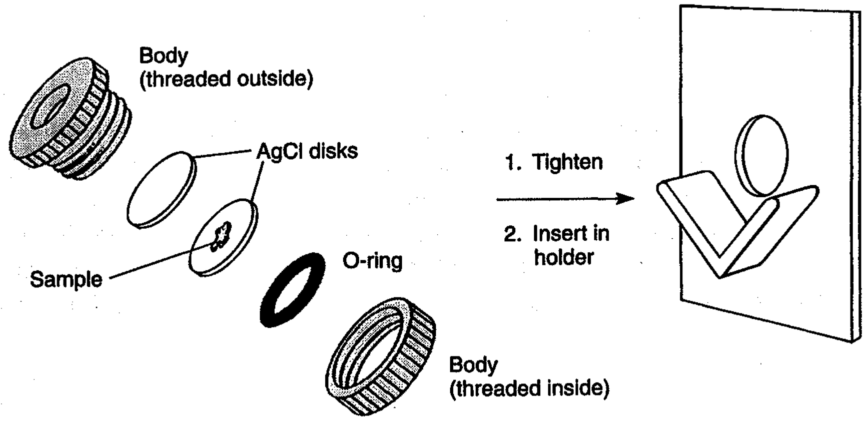

Figure 19.2 AgCl mini-liquid cell and V-mount holder.

---

[1] The Wilks Mini-Cell liquid sample holder is available from the Foxboro Company, 151 Woodward Avenue, South Norwalk, CT 06856. We recommend the AgCl cell windows with a 0.10-mm depression, rather than the 0.025-mm depression.

spectrum), you should select the flat side of each silver chloride plate. Insert the O-ring into the cell body as shown in Figure 19.2, place the plate into the cell body with the flat surface up, and add one drop or less of liquid to the plate.

Do not use amines with AgCl plates.

Place the second plate on top of the first with the flat side down. The orientation of the silver chloride plates is shown in Figure 19.3A. This arrangement is used to obtain a capillary film of your sample. Screw the top of the mini-cell into the body of the cell so that the silver chloride plates are held firmly together. A tight seal forms because AgCl deforms under pressure.

Other combinations may be used with these plates. For example, you may vary the sample path length by using the orientations shown in Figures 19.3B and 19.3C. If you add your sample to the 0.10-mm depression of one plate and cover it with the flat side of the other one, you obtain a path length of 0.10 mm (Fig. 19.3B). This arrangement is useful for analyzing volatile or low-viscosity liquids. Placement of the two plates with their depressions toward each other gives a path length of 0.20 mm (Fig. 19.3C). This orientation may be used for a solution of a solid (or liquid) in carbon tetrachloride (Section 19.5B).

Determining the Spectrum. Slide the V-mount holder shown in Figure 19.2 into the slot on the infrared spectrophotometer. Set the cell assembly in the V-mount holder and determine the infrared spectrum of the liquid.

Cleaning and Storing the AgCl Plates. Once the spectrum has been determined, the cell assembly holder should be demounted and the AgCl plates rinsed with methylene chloride or acetone. Do not use tissue to wipe the plates as they scratch easily. AgCl plates are light-sensitive. Store the plates in a light-tight container.

19.4 SOLID SAMPLES—KBr PELLETS

The easiest method of preparing a solid sample is to make a potassium bromide (KBr) pellet. When KBr is placed under pressure, it melts, flows, and seals the sample into a solid solution, or matrix. Because potassium bromide does not absorb in the infrared spectrum, a spectrum can be obtained on a sample without interference.

Preparing the Sample. Remove the agate mortar and pestle from the desiccator for use in preparing the sample. (Take care of them, they are expensive.) Grind 1 mg (0.001 g) of the solid sample for 1 minute in the agate mortar. At this point, the particle

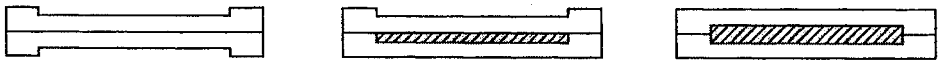

A. Capillary film B. 0.10–mm path length C. 0.20–mm path length

Figure 19.3 Path length variations for AgCl plates.

size will become so small that the surface of the solid appears shiny. Add 80 mg (0.080 g) of *powdered* potassium bromide and grind the mixture for about 30 seconds with the pestle. Scrape the mixture into the middle with a spatula and grind the mixture again for about 15 seconds. This grinding operation helps to mix the sample thoroughly with the KBr. You should work as rapidly as possible, because KBr absorbs water. The sample and KBr must be finely ground or the mixture will scatter the infrared radiation excessively. Using your spatula, heap the mixture in the center of the mortar. Return the bottle of potassium bromide to the desiccator where it is stored when it is not in use.

The sample and potassium bromide should be weighed on an analytical balance the first few times that a pellet is prepared. After some experience, you can estimate these quantities quite accurately by eye.

Making a Pellet Using a KBr Hand Press. Two methods are commonly used to prepare KBr pellets. The first method uses the hand press apparatus shown in Figure 19.4.[2]

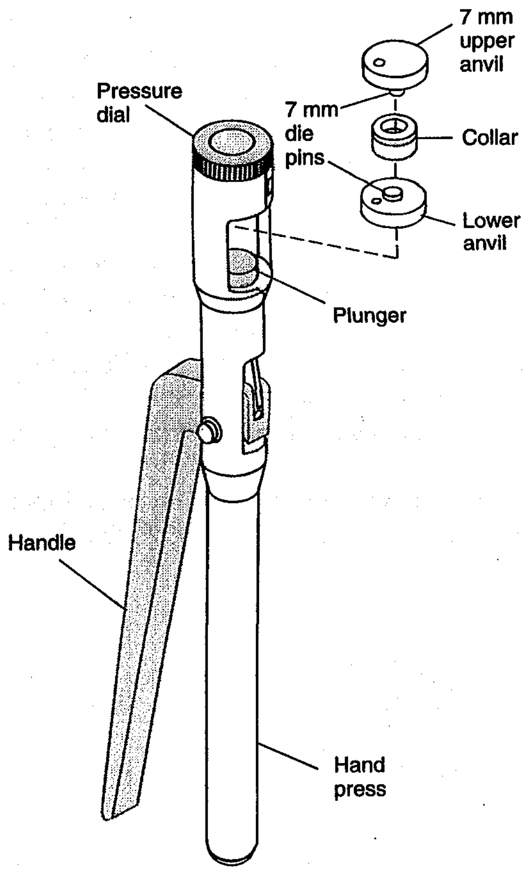

Figure 19.4 Making a KBr pellet with a hand press.

---

[2] KBr Quick Press unit is available from Wilmad Glass Company, Inc., Route 40 and Oak Road, Buena, NJ 08310.

Remove the die set from the storage container. Take extreme care to avoid scratching the polished surfaces of the die set. Place the anvil with the shorter die pin (lower anvil in Figure 19.4) on a bench. Slip the collar over the pin. Remove about one-fourth of your KBr mixture with a spatula and transfer it into the collar. The powder may not cover the head of the pin completely but do not be concerned about this. Place the anvil with the longer die pin into the collar so that the die pin comes into contact with the sample. Never press the die set unless it contains a sample.

Lift the die set carefully by holding onto the lower anvil so that the collar stays in place. If you are careless with this operation, the collar may move enough to allow the powder to escape. Open the handle of the hand press slightly, tilt the press back a bit, and insert the die set into the press. Make sure that the die set is seated against the side wall of the chamber. Close the handle. It is imperative that the die set be seated against the side wall of the chamber so that the die is centered in the chamber. Pressing the die in an off-centered position can bend the anvil pins.

With the handle in the closed position, rotate the pressure dial so that the upper ram of the hand press just touches the upper anvil of the die assembly. Tilt the unit back so that the die set does not fall out of the press. Open the handle and rotate the pressure dial clockwise about one-half turn. Slowly compress the KBr mixture by closing the handle. The pressure should be no greater than that exerted by a very firm handshake. Do not apply excessive pressure or the dies may be damaged. If in doubt, rotate the pressure dial counterclockwise to lower the pressure. If the handle closes too easily, open the handle, rotate the pressure dial clockwise, and compress the sample again. Compress the sample for about 60 seconds.

After this time, tilt the unit back so that the die set does not fall out of the hand press. Open the handle and carefully remove the die set from the unit. Turn the pressure dial counterclockwise about one full turn. Pull the die set apart and inspect the KBr pellet. Ideally, the pellet should appear clear like a piece of glass, but usually it will be translucent or somewhat opaque. There may be some cracks or holes in the pellet. The pellet will produce a good spectrum, even with imperfections, as long as light can travel through the pellet.

Making a Pellet with a KBr Minipress. The second method of preparing a pellet uses the minipress apparatus shown in Figure 19.5. Obtain a ground KBr mixture as

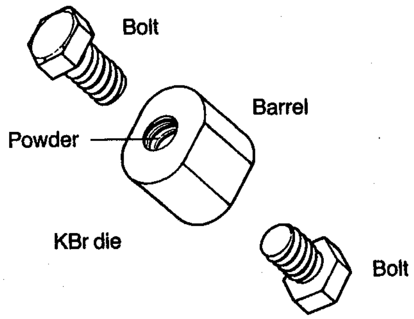

Figure 19.5 Making a KBr pellet with a minipress.

described above in "Preparing the Sample," and transfer a portion of the finely ground powder (usually not more than half) into a die that compresses it into a translucent pellet. As shown in Figure 19.5, the die consists of two stainless steel bolts and a threaded barrel. The bolts have their ends ground flat. To use this die, screw one of the bolts into the barrel, but not all the way; leave one or two turns. Carefully add the powder with a spatula into the open end of the partly assembled die and tap it lightly on the bench top to give an even layer on the face of the bolt. While keeping the barrel upright, carefully screw the second bolt into the barrel until it is finger-tight. Insert the head of the bottom bolt into the hexagonal hole in a plate bolted to the bench top. This plate keeps the head of one bolt from turning. The top bolt is tightened with a torque wrench to compress the KBr mixture. Continue to turn the torque wrench until you hear a loud click (the ratchet mechanism makes softer clicks) or until you reach the appropriate torque value (20 ft-lb). If you tighten the bolt beyond this point, you may twist the head off one of the bolts. Leave the die under pressure for about 60 seconds; then reverse the ratchet on the torque wrench or pull the torque wrench in the opposite direction to open the assembly. When the two bolts are loose, hold the barrel horizontally and carefully remove the two bolts. You should observe a clear or translucent KBr pellet in the center of the barrel. Even if the pellet is not totally transparent, you should be able to obtain a satisfactory spectrum as long as light passes through the pellet.

Determining the Infrared Spectrum. To obtain the spectrum, slide the holder appropriate for the type of die that you are using into the slot on the infrared spectrophotometer. Set the die containing the pellet in the holder so that the sample is centered in the optical path. Obtain the infrared spectrum. If you are using a double-beam instrument, you may be able to compensate (at least partially) for a marginal pellet by placing a wire screen or attenuator in the reference beam, thereby balancing the lowered transmittance of the pellet.

Problems with an Unsatisfactory Pellet. If the pellet is unsatisfactory (too cloudy to pass light), one of several things may have been wrong:

1. The KBr mixture may not have been ground finely enough, and the particle size may be too big. The large particle size creates too much light scattering.
2. The sample may not be dry.
3. Too much sample may have been used for the amount of KBr taken.
4. The pellet may be too thick; that is, too much of the powdered mixture was put into the die.
5. The KBr may have been "wet" or have acquired moisture from the air while the mixture was being ground in the mortar.
6. The sample may have a low melting point. Low-melting solids not only are difficult to dry but also melt under pressure. You may need to dissolve the compound in a solvent and run the spectrum in solution (Section 19.5).

Cleaning and Storing the Equipment. After you have determined the spectrum, punch the pellet out of the die with a wooden applicator stick (a spatula should not be used as it may scratch the dies). Remember that the polished faces of the die set must not be scratched or they become useless. Pull a piece of Kimwipe through the die unit to remove all the sample. Also wipe any surfaces with a Kimwipe. *Do not wash the dies with* *water.* Check with your instructor to see if there are additional instructions for cleaning the die set. Return the dies to the storage container. Wash the mortar and pestle with water and acetone, dry them carefully with paper towels, and return them to the desiccator. Return the KBr powder to its desiccator.

19.5 SOLID SAMPLES—SOLUTION SPECTRA

Method A—Solution Between Salt (NaCl) Plates

For substances that are soluble in carbon tetrachloride, a quick and easy method for determining the spectra of solids is available. Dissolve as much solid as possible in 0.1 mL of carbon tetrachloride. Place one or two drops of the solution between sodium chloride plates in precisely the same manner as that used for pure liquids (Section 19.2). The spectrum is determined as described for pure liquids using salt plates (Section 19.2). You should work as quickly as possible. If there is a delay, the solvent will evaporate from between the plates before the spectrum is recorded. Because the spectrum contains the absorptions of the solute superimposed on the absorptions of carbon tetrachloride, it is important to remember that any absorption that appears near 800 cm^{-1} may be due to the stretching of the C—Cl bond of the solvent. Information contained to the right of about 900 cm^{-1} is not usable in this method. There are no other interfering bands for this solvent (see Fig. 19.6), and any other absorptions can be attributed to your sample. Chloroform solutions should not be studied by this method, because the solvent has too many interfering absorptions (see Fig. 19.7).

Caution: Carbon tetrachloride is a hazardous solvent. Work under the hood! Use gloves.

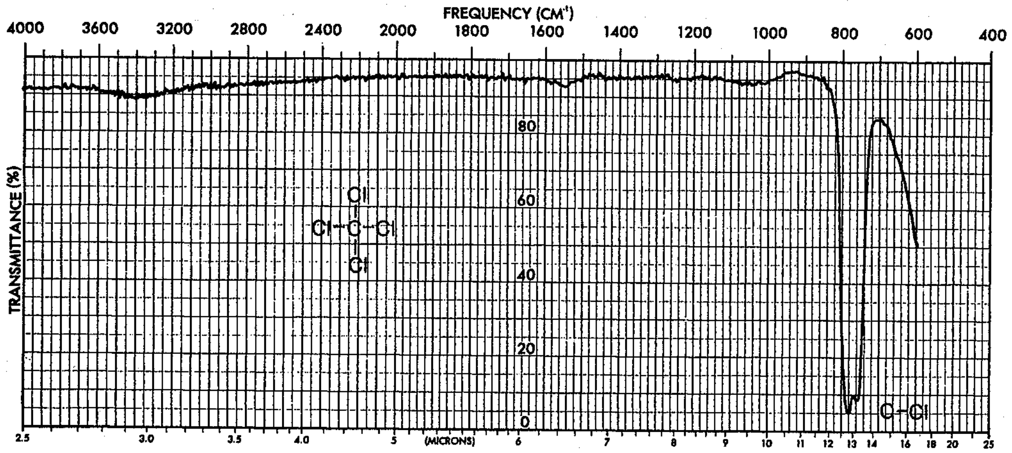

Figure 19.6 Infrared spectrum of carbon tetrachloride.

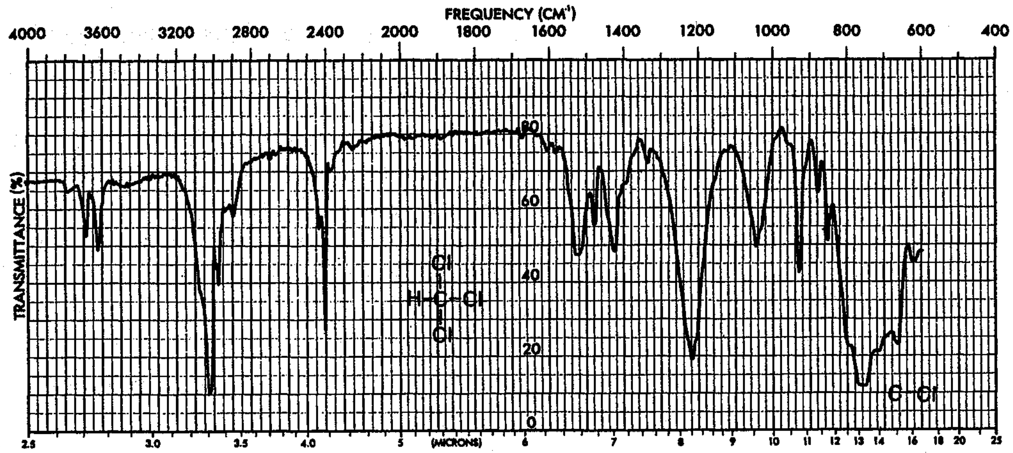

Figure 19.7 Infrared spectrum of chloroform.

Carbon tetrachloride, besides being toxic, is suspected of being a carcinogen. In spite of the health problems associated with its use, there is no suitable alternative solvent for infrared spectroscopy. Other solvents have too many interfering infrared absorption bands. Handle carbon tetrachloride very carefully to minimize the adverse health effects. The spectroscopic-grade carbon tetrachloride should be stored in a glass stoppered bottle in a hood. A Pasteur pipet should be attached to the bottle, possibly by storing it in a test tube taped to the side of the bottle. All sample preparation should be conducted in a hood. Rubber or plastic gloves should be worn. The cells should also be cleaned in the hood. All carbon tetrachloride used in preparing samples should be disposed of in an appropriately marked waste container.

Method B—AgCl Mini-Cell

The AgCl mini-cell described in Section 19.3 may be used to determine the infrared spectrum of a solid dissolved in carbon tetrachloride. Prepare a 5–10% solution (5–10 mg in 0.1 mL) in carbon tetrachloride. If it is not possible to prepare a solution of this concentration because of low solubility, dissolve as much solid as possible in the solvent. Following the instructions given in Section 19.3, position the AgCl plates as shown in Figure 19.3C to obtain the maximum possible path length of 0.20 mm. When the cell is tightened firmly, it will not leak.

As indicated in Method A, the spectrum will contain the absorptions of the dissolved solid superimposed on the absorptions of carbon tetrachloride. A strong absorption appears near 800 cm^{-1} for C—Cl stretch in the solvent. No useful information may be obtained for the sample to the right of about 900 cm^{-1}, but other bands that appear in the spectrum will belong to your sample. Read the safety material provided in Method A. Carbon tetrachloride is toxic, and it should be used under a hood.

Care should be taken in cleaning the AgCl plates. Because AgCl plates scratch easily, they should not be wiped with tissue. Rinse them with methylene chloride and keep them in a dark place. Amines will destroy the plates.

Method C—Solution Cells (NaCl)

The spectra of solids may also be determined in a type of permanent sample cell called a **solution cell.** (The infrared spectra of liquids may also be determined in this cell.) The solution cell, shown in Figure 19.8, is made from two salt plates, mounted with a Teflon spacer between them to control the thickness of the sample. The top sodium chloride plate has two holes drilled in it so that the sample can be introduced into the cavity between the two plates. These holes are extended through the face plate by two tubular extensions designed to hold Teflon plugs, which seal the internal chamber and prevent evaporation. The tubular extensions are tapered so that a syringe body (Luer lock without a needle) will fit snugly into them from the outside. The cells are thus filled from a syringe; usually, they are held upright and filled from the bottom entrance port.

These cells are very expensive, and you should try either Method A or B before using solution cells. If you do need them, obtain your instructor's permission and receive instruction before using the cells. The cells are purchased in matched pairs, with identical path lengths. Dissolve a solid in a suitable solvent, usually carbon tetrachloride, and add the solution to one of the cells **(sample cell)** as described in the previous paragraph. The pure solvent, identical to that used to dissolve the solid, is placed in the other cell **(reference cell).** The spectrum of the solvent is subtracted from the spectrum of the so-

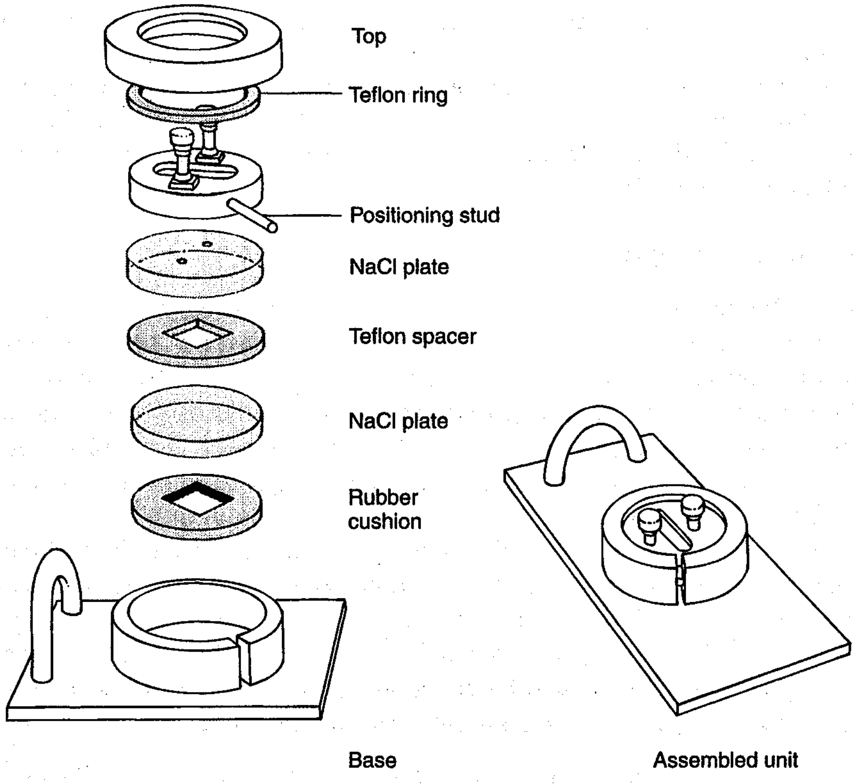

Figure 19.8 Solution cell.

lution (not always completely), and a spectrum of the solute is thus provided. For the solvent compensation to be as exact as possible and to avoid contamination of the reference cell, it is essential that one cell be used as a reference and that the other cell be used as a sample cell without ever being interchanged. After the spectrum is determined, it is important to clean the cells by flushing them with clean solvent. They should be dried by passing dry air through the cell.

Solvents most often used in determining infrared spectra are carbon tetrachloride, chloroform, and carbon disulfide. The spectra of these substances are shown in Figures 19.6, 19.7, and 19.9. A 5–10% solution of solid in one of these solvents usually gives a good spectrum. Carbon tetrachloride and chloroform are suspected carcinogens. However, because there are no suitable alternative solvents, these compounds must be used in infrared spectroscopy. The procedure outlined on page 851 for carbon tetrachloride should be followed. This procedure serves equally well for chloroform.

Before you use the solution cells, you must obtain the instructor's permission and instruction on how to fill and clean the cells.

19.6 SOLID SAMPLES—NUJOL MULLS

If an adequate KBr pellet cannot be obtained, or if the solid is insoluble in a suitable solvent, the spectrum of a solid may be determined as a Nujol mull. In this method, finely grind about 5 mg of the solid sample in an agate mortar with a pestle. Then add one or two drops of Nujol mineral oil (white) and grind the mixture to a very fine dispersion. The solid is not dissolved in the Nujol; it is actually a suspension. This mull is then placed between two salt plates using a rubber policeman. Mount the salt plates in the holder in the same way as for liquid samples (Section 19.2).

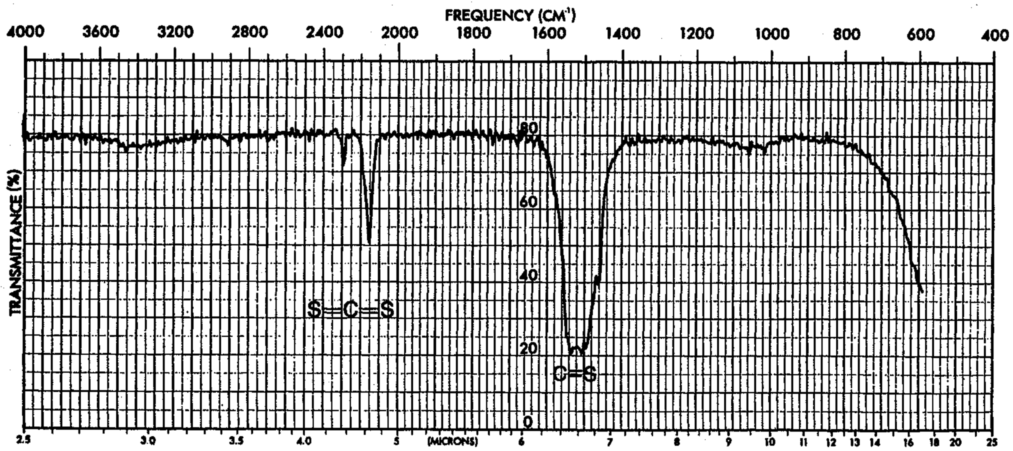

Figure 19.9 Infrared spectrum of carbon disulfide.

Nujol is a mixture of high-molecular-weight hydrocarbons. Hence, it has absorptions in the C—H stretch and CH_2 and CH_3 bending regions of the spectrum (Fig. 19.10). Clearly, if Nujol is used, no information can be obtained in these portions of the spectrum. In interpreting the spectrum, you must ignore these Nujol peaks. It is important to label the spectrum immediately after it was determined, noting that it was determined as a Nujol mull. Otherwise, you might forget that the C—H peaks belong to Nujol and not to the dispersed solid.

19.7 RECORDING THE SPECTRUM

The instructor will describe how to operate the infrared spectrophotometer, because the controls vary considerably, depending on the manufacturer, model of the instrument, and type. For example, some instruments involve pushing only a few buttons, while others use a more complicated computer interface system.

In all cases, it is important that the sample, the solvent, the type of cell or method used, and any other pertinent information be written on the spectrum immediately after the determination. This information may be important, and it is easily forgotten if not recorded. You may also need to calibrate the instrument (Section 19.8).

19.8 CALIBRATION

For some instruments, the frequency scale of the spectrum must be calibrated so that you know the position of each absorption peak precisely. You can recalibrate by recording a very small portion of the spectrum of polystyrene over the spectrum of your sample. The complete spectrum of polystyrene is shown in Figure 19.11. The most important

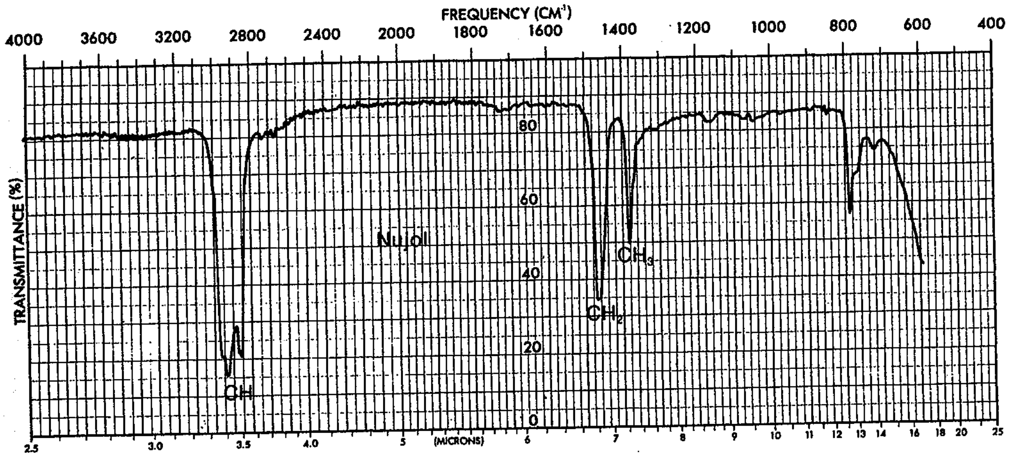

Figure 19.10 Infrared spectrum of Nujol (mineral oil).

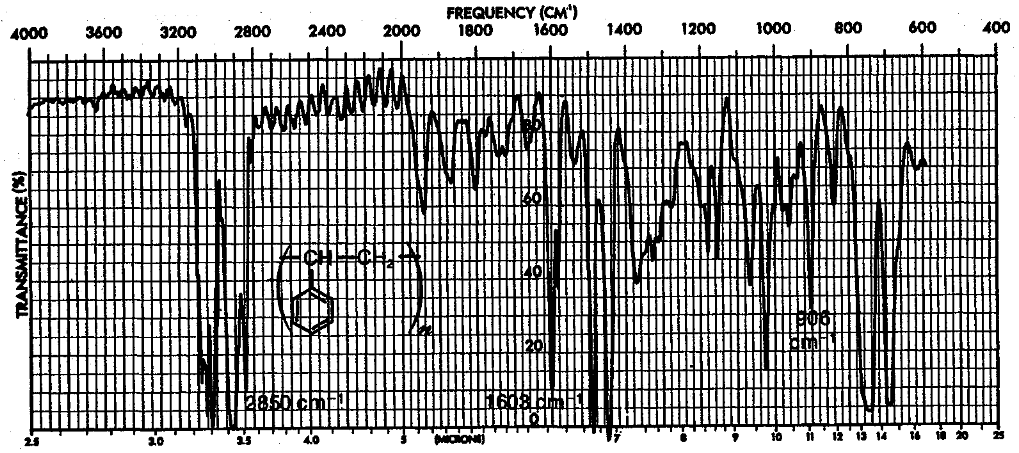

Figure 19.11 Infrared spectrum of polystyrene (thin film).

of these peaks is at 1603 cm^{-1}; other useful peaks are at 2850 and 906 cm^{-1}. After you record the spectrum of your sample, substitute a thin film of polystyrene for the sample cell and record the **tips** (not the entire spectrum) of the most important peaks over the sample spectrum.

It is always a good idea to calibrate a spectrum when the instrument uses chart paper with a preprinted scale. It is difficult to align the paper properly so that the scale matches the absorption lines precisely. You often need to know the precise values for certain functional groups (for example, the carbonyl group). Calibration is essential in these cases.

With computer-interfaced instruments, the instrument does not need to be calibrated. With this type of instrument, the spectrum and scale are printed on blank paper at the same time. The instrument has an internal calibration that ensures that the positions of the absorptions are known precisely and that they are placed at the proper positions on the scale. With this type of instrument, it is often possible to print a list of the locations of the major peaks as well as to obtain the complete spectrum of your compound.

Part B. Nuclear Magnetic Resonance (NMR)

19.9 PREPARING A SAMPLE FOR PROTON NMR

The NMR sample tubes used in most instruments are approximately 0.5 cm × 18 cm in overall dimension and are fabricated of uniformly thin glass tubing. These tubes are very fragile and expensive, so care must be taken to avoid breaking the tubes.

To prepare the solution you must first choose the appropriate solvent. The solvent should not have NMR absorption peaks of its own, that is, no protons. Carbon tetrachloride CCl_4 fits this requirement and can be used in some instruments. However, because

Fourier transform (FT) NMR spectrometers require deuterium to stabilize (lock) the field (Section 19.10), organic chemists usually use deuterated chloroform $CDCl_3$ as a solvent. This solvent dissolves most organic compounds and is relatively inexpensive. You can use this solvent with any NMR instrument. You should not use normal chloroform $CHCl_3$, because the solvent contains a proton. Deuterium ^2H does not absorb in the proton region and is thus "invisible," or not seen, in the proton NMR spectrum. Use deuterated chloroform to dissolve your sample unless you are instructed by your instructor to use another solvent, such as carbon tetrachloride CCl_4.

Routine Sample Preparation Using Deuterated Chloroform

1. Most organic liquids and low-melting solids will dissolve in deuterated chloroform. However, you should first determine if your sample will dissolve in ordinary $CHCl_3$ before using the deuterated solvent. If your sample does not dissolve in chloroform, consult your instructor about a possible alternate solvent, or consult the section entitled "Nonroutine Sample Preparation" later in this section.

Caution: Chloroform, deuterated chloroform, and carbon tetrachloride are all toxic solvents. In addition, they may be carcinogenic substances (see p. 21). Use protective gloves.

2. If you are using an FT-NMR spectrometer, add 30 mg (0.030 g) of your liquid or solid sample to a tared conical vial or test tube. Use a Pasteur pipet to transfer a liquid or a spatula to transfer a solid. Non-FT instruments usually require a more concentrated solution in order to obtain an adequate spectrum. Typically, a 10–30% sample concentration (weight/weight) is used.
3. With the help of your instructor, transfer about 0.5 mL of the deuterated chloroform with a *clean and dry Pasteur pipet* to your sample. Swirl the test tube or conical vial to help dissolve the sample. At this point, the sample should have completely dissolved. Add a little more solvent, if necessary, to dissolve the sample fully.
4. Transfer the solution to the NMR tube using a clean and dry Pasteur pipet. Be careful when transferring the solution into the NMR tube so that you avoid breaking the edge of the fragile NMR tube. It is best to hold the NMR tube and the container with the solution in the same hand when making the transfer.
5. Once the solution has been transferred to the NMR tube, use a clean pipet to add enough deuterated chloroform to bring the total solution height[1] to about 35 mm total (Fig. 19.12). In some cases, you will need to add a small amount of tetramethylsilane (TMS) as a reference substance (Section 19.11). Check with your instructor to see if you need to add TMS to your sample. Deuterated chloroform has a small amount of $CHCl_3$ impurity, which gives rise to a low-intensity peak in the NMR spectrum at 7.27 ppm. This impurity may also help you to "reference" your spectrum.

---

[1] Your instructor may recommend a different height than the one specified here (35 mm).

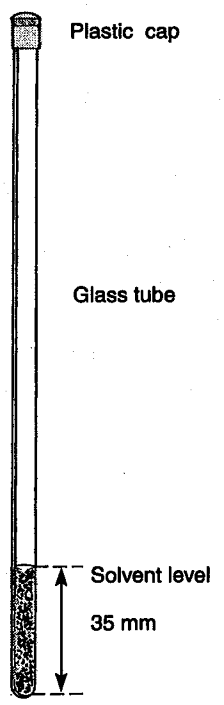

Figure 19.12 NMR sample tube.

6. Cap the NMR tube. Do this firmly, but not too tightly. If you jam the cap on, you may have trouble removing it later without breaking the end off of the very thin glass tube. Make sure that the cap is on straight. Invert the NMR tube several times to mix the contents.
7. You are now ready to record the NMR spectrum of your sample. Insert the NMR tube into its holder and adjust its depth by using the gauge provided to you. See Section 19.12.

Cleaning the NMR Tube

1. Carefully uncap the tube so that you do not break it. Turn the tube upside-down and hold it vertically over a beaker. Shake the tube up and down gently so that the contents of the tube empties into the beaker.
2. Partially refill the NMR tube with acetone using a Pasteur pipet. Carefully replace the cap and invert the tube several times to rinse it.
3. Remove the cap and drain the tube as before. Place the open tube upside-down in a beaker with a Kimwipe or paper towel placed in the bottom of the beaker. Leave the tube standing in this position for at least one laboratory period so that the acetone completely evaporates. Alternatively, you may place the beaker and NMR tube in an oven for at least 2 hours. If you need to use the NMR tube before the acetone has fully evaporated, attach a piece of pressure tubing to the tube, and pull a vacuum with an aspirator. After several

minutes, the acetone should have fully evaporated. Because acetone contains protons, you must not use the NMR tube until the acetone has evaporated completely.

4. Once the acetone is evaporated, place the clean tube and its cap (do not cap the tube) in its storage container and place it in your desk. The storage container will prevent the tube from being crushed.

Nonroutine Sample Preparation. With highly polar substances you may find that your sample will not dissolve in deuterated chloroform. If this is the case, you may be able to dissolve the sample in deuterium oxide, D_2O. Spectra determined in D_2O often show a small peak at about 5 ppm because of OH impurity. If the sample compound has acidic hydrogens, they may *exchange* with D_2O, leading to the appearance of an OH peak in the spectrum and the *loss* of the original absorption from the acidic proton, owing to the exchanged hydrogen. In many cases, this will also alter the splitting patterns of a compound.

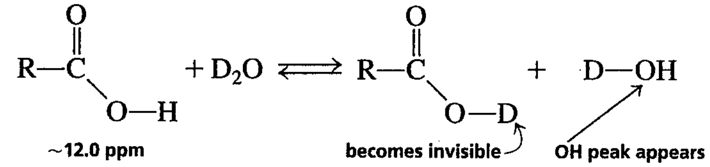

$$R{-}C(=O){-}O{-}H + D_2O \rightleftharpoons R{-}C(=O){-}O{-}D + D{-}OH$$

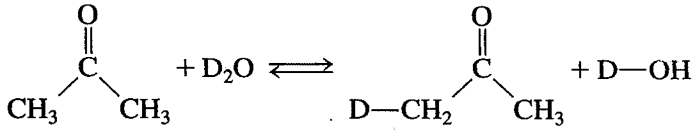

$$CH_3C(=O)CH_3 + D_2O \rightleftharpoons D{-}CH_2C(=O)CH_3 + D{-}OH$$

$$CH_3CH_2OH + D_2O \rightleftharpoons CH_3CH_2OD + D{-}OH$$

Most solid carboxylic acids do not dissolve in CCl_4, $CDCl_3$, or even D_2O. In such cases, a small piece of sodium metal is added to about 1 mL of D_2O. The acid is then dissolved in this solution. The resulting basic solution enhances the solubility of the carboxylic acid. In such a case, the hydroxyl proton of the carboxylic acid cannot be observed in the NMR spectrum, because it exchanges with the solvent. A large DOH peak is observed, however, due to the exchange and the H_2O impurity in the D_2O solvent.

When the above solvents fail, other special solvents can be used. Acetone, acetonitrile, dimethylsulfoxide, pyridine, benzene, and dimethylformamide can be used if you are not interested in the region or regions of the NMR spectrum in which they give rise to absorption. The deuterated (but expensive) analogues of these compounds are also used in special instances (for example, acetone-d_6, dimethylsulfoxide-d_6, dimethylformamide-d_7, and benzene-d_6).[2] If the sample is not sensitive to acid, trifluoroacetic acid (which has no protons with $\delta < 12$) can be used. You must be aware that these solvents often lead to different chemical shift values from those determined in CCl_4 or $CDCl_3$. Variations of as much as 0.5–1.0 ppm have been observed. In fact, it is sometimes possible, by switching to pyridine, benzene, acetone, or dimethylsulfoxide as solvents, to separate peaks that overlap when CCl_4 or $CDCl_3$ solutions are used.

---

[2] Unisol, a commercial mixture of dimethylsulfoxide-d_6 and $CDCl_3$, dissolves most carboxylic acids.

Health Hazards Associated with NMR Solvents. Carbon tetrachloride, chloroform (and chloroform-d), and benzene (and benzene-d_6) are hazardous solvents. Besides being highly toxic, they are also suspected carcinogens. In spite of these health problems, these solvents are commonly used in NMR spectroscopy, because there are no suitable alternatives. These solvents are used because they contain no protons and because they are excellent solvents for most organic compounds. Therefore, you must learn to handle these solvents with great care to minimize the hazard. These solvents should be stored either under a hood or in septum-capped bottles. If the bottles have screwcaps, a pipet should be attached to each bottle. A recommended way of attaching the pipet is to store it in a test tube taped to the side of the bottle. Septum-capped bottles can be used only by withdrawing the solvent with a hypodermic syringe that has been designated solely for this use. All samples should be prepared under a hood and solutions should be disposed of in an appropriately designated waste container that is stored under the hood. Wear rubber or plastic gloves when preparing or discarding samples.

19.10 PREPARING A SAMPLE FOR CARBON-13 NMR

Section 19.9 describes the technique for preparing samples for proton NMR. Much of what is described there also applies to carbon NMR. There are some differences, however, in determining a carbon spectrum. Fourier transform instruments require a deuterium signal to stabilize (lock) the field. Therefore, the solvents must contain deuterium. Deuterated chloroform $CDCl_3$ is used most commonly for this purpose because of its relatively low cost. Other deuterated solvents may also be used.

Because of the low natural abundance of carbon-13 in a sample, you often need to acquire multiple scans over a long period (Appendix 5, Section CMR.1, p. 931). You can save considerable time by using a relatively concentrated sample.

Modern FT-NMR spectrometers allow chemists to obtain both the proton and carbon NMR spectra of the same sample in the same NMR tube. After changing several parameters in the program operating the spectrometer, you can obtain both spectra without removing the sample from the probe. The only real difference is that a proton spectrum may be obtained after a few scans, whereas the carbon spectrum may require several thousand scans to obtain a suitable spectrum.

19.11 REFERENCE SUBSTANCES

Proton NMR. To provide the internal reference standard, tetramethylsilane (TMS) must be added to the sample solution. This substance has the formula $(CH_3)_4Si$. By universal convention, the chemical shifts of the protons in this substance are defined as 0.00 ppm (0.00 δ). The spectrum should be shifted so that the TMS signal appears at this position on precalibrated paper.

The concentration of TMS in the sample should range from 1 to 3%. Some people prefer to add one to two drops of TMS to the sample just before determining the spectrum. Because TMS has 12 equivalent protons, not much of it needs to be added. A Pasteur pipet or a syringe may be used for the addition. It is far easier to have available in the laboratory a prepared solvent that already contains TMS. Deuterated chloroform and carbon tetrachloride often have TMS added to them. Because TMS is highly volatile (bp 26.5°C), such solutions should be stored, tightly stoppered, in a refrigerator. Tetramethylsilane itself is best stored in a refrigerator as well.

Tetramethylsilane does not dissolve in D_2O. For spectra determined in D_2O, a different internal standard, sodium 2,2-dimethyl-2-silapentane-5-sulfonate, must be used. This standard is water-soluble and gives a resonance peak at 0.00 ppm (0.00 δ).

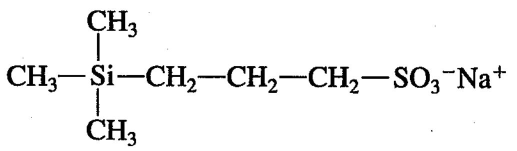

$$CH_3-\overset{\displaystyle CH_3}{\underset{\displaystyle CH_3}{\overset{|}{\underset{|}{Si}}}}-CH_2-CH_2-CH_2-SO_3^-Na^+$$

Sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS)

Carbon NMR. TMS may be added as an internal reference standard where the chemical shift of the methyl carbon is defined as 0.00 ppm. Alternatively, you may use the center peak of the $CDCl_3$ pattern, which is found at 77.0 ppm. This pattern can be observed as a small "triplet" near 77.0 ppm in a number of the spectra given in Appendix 5. (For example, see Fig. CMR.3 on p. 933).

In most instances, the instructor or some qualified laboratory assistant will actually record your NMR spectrum. If you are permitted to operate the NMR spectrometer, the instructor will provide instructions. Because the controls of NMR spectrometers vary, *depending on the make or model of the instrument, we shall not try to describe these* controls.

Do not operate the NMR spectrometer unless you have been properly instructed.

PROBLEMS

1. Comment on the suitability of running the infrared spectrum under each of the following conditions. If there is a problem with the conditions given, provide a suitable alternative method.

(a) A neat spectrum of a liquid with a boiling point of 150°C is determined using salt plates.
(b) A neat spectrum of a liquid with a boiling point of 35°C is determined using salt plates.
(c) A KBr pellet is prepared with a compound that melts at 200°C.
(d) A KBr pellet is prepared with a compound that melts at 30°C.
(e) A solid aliphatic hydrocarbon compound is determined as a Nujol mull.
(f) Silver chloride plates are used to determine the spectrum of aniline.
(g) Sodium chloride plates are selected to run the spectrum of a compound that contains some water.

2. Describe the method that you should employ to determine the proton NMR spectrum of a carboxylic acid, which is insoluble in *all* the common organic solvents that your instructor is likely to make available.
3. In order to save money, a student uses chloroform instead of deuterated chloroform to run a carbon-13 NMR spectrum. Is this a good idea?
4. Look up the solubilities for the following compounds and decide whether you would select deuterated chloroform or deuterated water to dissolve the substances for NMR spectroscopy.
   (a) Glycerol (1,2,3-propanetriol)
   (b) 1,4-Diethoxybenzene
   (c) Propyl pentanoate (propyl ester of pentanoic acid)
5. What would happen if you ran a proton NMR spectrum without any TMS in the sample?

TECHNIQUE 20

Guide to the Chemical Literature

Often, it may be necessary to go beyond the information contained in the typical organic chemistry textbook and to use reference material in the library. At first glance, using library materials may seem formidable because of the numerous sources the library contains. If, however, you adopt a systematic approach, the task can prove rather useful. This description of various popular sources and an outline of logical steps to follow in the typical literature search should be helpful.

20.1 LOCATING PHYSICAL CONSTANTS: HANDBOOKS

To find information on routine physical constants, such as melting points, boiling points, indices of refraction, and densities, you should first consider a handbook. Examples of suitable handbooks are

S. Budavari, ed. *The Merck Index.* 12th ed. Whitehouse Station, NJ: Merck & Co., 1996.
J. A. Dean, ed. *Lange's Handbook of Chemistry.* 14th ed. New York: McGraw-Hill, 1992.
D. R. Lide, ed. *CRC Handbook of Chemistry and Physics.* 76th ed. Boca Raton, FL: CRC Press, 1993.

The *Handbook of Chemistry and Physics* is the handbook consulted most often. For organic chemistry, however, *The Merck Index* is probably better suited. *The Merck Index* also contains literature references on the isolation, structure determination, and synthesis of a substance, along with its molecular formula, elemental analysis, and certain properties of medicinal interest (e.g., toxicity and medicinal and veterinary uses).

A more complete handbook is

J. Buckingham, ed., *Dictionary of Organic Compounds.* New York: Chapman & Hall/Methuen, 1982–1992.

This is a revised version of an earlier four-volume handbook edited by I. M. Heilbron and H. M. Bunbury. In its present form, it consists of seven volumes with 10 supplements.

20.2 GENERAL SYNTHETIC METHODS

Many standard introductory textbooks in organic chemistry provide tables that summarize most of the common reactions, including side reactions, for a given class of compounds. These books also describe alternative methods of preparing compounds.

W. H. Brown. *Organic Chemistry,* 2nd ed. Philadelphia: Saunders College Publishing, 1997.
F. A. Carey. *Organic Chemistry.* 3rd ed. New York: McGraw-Hill, 1996.
S. Ege. *Organic Chemistry.* 3rd ed. Lexington, MA: D. C. Heath, 1994.
R. J. Fessenden and J. S. Fessenden. *Organic Chemistry.* 5th ed. Pacific Grove, CA: Brooks/Cole, 1994.
M. A. Fox and J. K. Whitesell. *Organic Chemistry,* 2nd ed. Boston: Jones and Bartlett, 1997.
G. M. Loudon. *Organic Chemistry.* 3rd ed. Menlo Park, CA: Benjamin/Cummings, 1995.
J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992. Appendix B.
J. McMurry. *Organic Chemistry.* 4th ed. Pacific Grove, CA: Brooks/Cole, 1996.
R. T. Morrison and R. N. Boyd. *Organic Chemistry.* 6th ed. Englewood Cliffs, NJ: Prentice Hall, 1992.
S. H. Pine. *Organic Chemistry.* 5th ed. New York: McGraw-Hill, 1987.
T. W. G. Solomons. *Organic Chemistry.* 6th ed. New York: John Wiley, 1996.
A. Streitwieser, C. H. Heathcock, and E. M. Kosower. *Introduction to Organic Chemistry.* 4th ed. New York: Macmillan, 1992.
K. P. C. Vollhardt and N. E. Schore. *Organic Chemistry.* 2nd ed. New York: W. H. Freeman, 1994.
L. G. Wade, Jr. *Organic Chemistry.* 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1995.

20.3 SEARCHING THE CHEMICAL LITERATURE

If the information you are seeking is not available in any of the handbooks mentioned in Section 20.1 or if you are searching for more detailed information than they can provide, then a proper literature search is in order. Although an examination of standard textbooks can provide some help, you often must use all the resources of the library, including journals, reference collections, and abstracts. The following sections of this chapter outline how the various types of sources should be used and what sort of information can be obtained from them.

The methods for searching the literature discussed in this chapter use mainly printed materials. Modern search methods also make use of computerized databases and are discussed in Section 20.11. These are vast collections of data and bibliographic materials that can be scanned very rapidly from remote computer terminals. Although computerized searching is widely available, its use may not be readily accessible to undergraduate students. The following references provide excellent introductions to the literature of organic chemistry:

C. Carr. "Teaching and Using Chemical Information." *Journal of Chemical Education, 70* (September 1993): 719.

J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992. Appendix A.

A. N. Somerville. "Information Sources for Organic Chemistry, 1: Searching by Name Reaction and Reaction Type." *Journal of Chemical Education, 68* (July 1991): 553.

A. N. Somerville. "Information Sources for Organic Chemistry, 2: Searching by Functional Group." *Journal of Chemical Education, 68* (October 1991): 842.

A. N. Somerville. "Information Sources for Organic Chemistry, 3: Searching by Reagent." *Journal of Chemical Education, 69* (May 1992): 379.

G. Wiggins. *Chemical Information Sources.* New York: McGraw Hill, 1990. Integrates printed materials and computer sources of information.

20.4 COLLECTIONS OF SPECTRA

Collections of infrared, nuclear magnetic resonance, and mass spectra can be found in the following catalogs of spectra:

A. Cornu and R. Massot. *Compilation of Mass Spectral Data.* 2nd ed. London: Heyden and Sons, Ltd., 1975.

High-Resolution NMR Spectra Catalog. Palo Alto, CA: Varian Associates. Volume 1, 1962; Volume 2, 1963.

L. F. Johnson and W. C. Jankowski. *Carbon-13 NMR Spectra.* New York: John Wiley, 1972.

C. J. Pouchert. *Aldrich Library of Infrared Spectra.* 3rd ed. Milwaukee: Aldrich Chemical Co., 1981.

C. J. Pouchert. *Aldrich Library of FT-IR Spectra.* Milwaukee: Aldrich Chemical Co., 1985.

C. J. Pouchert. *Aldrich Library of NMR Spectra,* 2nd ed. Milwaukee: Aldrich Chemical Co., 1983.

C. J. Pouchert and J. Behnke. *Aldrich Library of* ^{13}C *and* ^{1}H *FT NMR Spectra.* Milwaukee: Aldrich Chemical Co., 1993.

Sadtler Standard Spectra. Philadelphia: Sadtler Research Laboratories. Continuing collection.

E. Stenhagen, S. Abrahamsson, and F. W. McLafferty. *Registry of Mass Spectral Data.* New York: John Wiley-Interscience, 1974. Four-volume set.

The American Petroleum Institute has also published collections of infrared, nuclear magnetic resonance, and mass spectra.

20.5 ADVANCED TEXTBOOKS

Much information about synthetic methods, reaction mechanisms, and reactions of organic compounds is available in any of the many current advanced textbooks in organic chemistry. Examples of such books are

F. A. Carey and R. J. Sundberg. *Advanced Organic Chemistry. Part A. Structure and Mechanisms; Part B. Reactions and Synthesis.* 3rd ed. New York: Plenum Press, 1990.

W. Carruthers. *Some Modern Methods of Organic Synthesis.* 3rd ed. Cambridge, U.K.: Cambridge University Press, 1986.
E. J. Corey and Xue-Min Cheng. *The Logic of Chemical Synthesis.* New York: John Wiley, 1989.
L. F. Fieser and M. Fieser. *Advanced Organic Chemistry.* New York: Reinhold, 1961.
I. L. Finar. *Organic Chemistry.* 6th ed. London: Longman Group, Ltd., 1986.
H. O. House. *Modern Synthetic Reactions.* 2nd ed. Menlo Park, CA: W. H. Benjamin, 1972.
J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992.
C. R. Noller. *Chemistry of Organic Compounds.* 3rd ed. Philadelphia: W. B. Saunders, 1965.
M. B. Smith. *Organic Synthesis.* New York: McGraw-Hill, 1994.
J. C. Stowell. *Intermediate Organic Chemistry.* 2nd ed. New York: John Wiley, 1993.
S. Warren. *Organic Synthesis: The Disconnection Approach.* New York: John Wiley, 1982.

These books often contain references to original papers in the literature for students wanting to follow the subject further. Consequently, you obtain not only a review of the subject from such a textbook but also a key reference that is helpful toward a more extensive literature search. The textbook by March is particularly useful for this purpose.

20.6 SPECIFIC SYNTHETIC METHODS

Anyone interested in locating information about a particular method of synthesizing a compound should first consult one of the many general textbooks on the subject. Useful ones are

N. Anand, J. S. Bindra, and S. Ranganathan. *Art in Organic Synthesis.* 2nd ed. New York: John Wiley, 1988.
D. Barton and W. D. Ollis, eds. *Comprehensive Organic Chemistry.* Oxford: Pergamon Press, 1979. Six-volume set.
C. A. Buehler and D. E. Pearson. *Survey of Organic Syntheses.* New York: John Wiley-Interscience, 1970 and 1977. Two-volume set.
F. A. Carey and R. J. Sundberg. *Advanced Organic Chemistry. Part B. Reactions and Synthesis.* 3rd ed. New York: Plenum Press, 1990.
Compendium of Organic Synthetic Methods. New York: John Wiley-Interscience, 1971–1995. This is a continuing series, now in eight volumes.
L. F. Fieser and M. Fieser. *Reagents for Organic Synthesis.* New York: John Wiley-Interscience, 1967–1994. This is a continuing series, now in 17 volumes.
T. W. Greene and P. G. M. Wuts. *Protective Groups in Organic Synthesis.* 2nd ed. New York: John Wiley, 1991.
H. O. House. *Modern Synthetic Reactions.* 2nd ed. Menlo Park, CA: W. H. Benjamin, 1972.
R. C. Larock. *Comprehensive Organic Transformations.* New York: VCH Press, 1989.
J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992.
B. P. Mundy and M. G. Ellerd. *Name Reactions and Reagents in Organic Synthesis.* New York: John Wiley, 1988.
S. Patai, ed. *The Chemistry of the Functional Groups.* London: Interscience, 1964–present. This series consists of many volumes, each one specializing in a particular functional group.

A. I. Vogel. Revised by members of the School of Chemistry, Thames Polytechnic. *Vogel's Textbook of Practical Organic Chemistry, Including Qualitative Organic Analysis.* 5th ed. London: Longman Group, Ltd., 1989.
R. B. Wagner and H. D. Zook. *Synthetic Organic Chemistry.* New York: John Wiley, 1956.

More specific information, including actual reaction conditions, exists in collections specializing in organic synthetic methods. The most important of these is

Organic Syntheses. New York: John Wiley, 1921–present. Published annually.
Organic Syntheses, Collective Volumes. New York: John Wiley, 1941–1993.
  Vol. 1, 1941, Annual Volumes 1–9
  Vol. 2, 1943, Annual Volumes 10–19
  Vol. 3, 1955, Annual Volumes 20–29
  Vol. 4, 1963, Annual Volumes 30–39
  Vol. 5, 1973, Annual Volumes 40–49
  Vol. 6, 1988, Annual Volumes 50–59
  Vol. 7, 1990, Annual Volumes 60–64
  Vol. 8, 1993, Annual Volumes 65–69

It is much more convenient to use the collective volumes where the earlier annual volumes of *Organic Syntheses* are combined in groups of nine or ten in the first six collective volumes (Vol. 1–6), and then in groups of five for the next two volumes (Vol. 7 and 8). Useful indices are included at the end of each of the collective volumes that classify methods according to the type of reaction, type of compound prepared, formula of compound prepared, preparation or purification of solvents and reagents, and use of various types of specialized apparatus.

The main advantage of using one of the *Organic Syntheses* procedures is that they have been tested to make sure that they work as written. Often, an organic chemist will adapt one of these tested procedures to the preparation of another compound. One of the features of the advanced organic textbook by March is that it includes references to specific preparative methods contained in *Organic Syntheses.*

More advanced material on organic chemical reactions and synthetic methods may be found in any one of a number of annual publications that review the original literature and summarize it. Examples include

Advances in Organic Chemistry: Methods and Results. New York: John Wiley, 1960–present.
Annual Reports of the Chemical Society, Section B. London: Chemical Society, 1905–present. Specifically, the section on *Synthetic Methods.*
Annual Reports in Organic Synthesis. Orlando, FL: Academic Press, 1985–1995.
Progress in Organic Chemistry. New York: John Wiley, 1952–1973.
Organic Reactions. New York: John Wiley, 1942–present.

Each of these publications contains a great many citations to the appropriate articles in the original literature.

20.7 ADVANCED LABORATORY TECHNIQUES

The student who is interested in reading about more advanced techniques than those described in this textbook, or in more complete descriptions of techniques, should consult one of the advanced textbooks specializing in organic laboratory techniques. Besides focusing on apparatus construction and the performance of complex reactions, these books also provide advice on purifying reagents and solvents. Useful sources of information on organic laboratory techniques include

R. B. Bates and J. P. Schaefer. *Research Techniques in Organic Chemistry.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

A. J. Krubsack. *Experimental Organic Chemistry.* Boston: Allyn and Bacon, 1973.

J. Leonard, B. Lygo, and G. Procter. *Advanced Practical Organic Chemistry,* 2nd ed. London: Chapman and Hall, 1995.

R. S. Monson, *Advanced Organic Synthesis: Methods and Techniques.* New York: Academic Press, 1971.

Techniques of Chemistry. New York: John Wiley, 1970–present. Currently 21 volumes. The successor to *Technique of Organic Chemistry,* this series covers experimental methods of chemistry, such as purification of solvents, spectral methods, and kinetic methods.

A. Weissberger, et al., eds. *Technique of Organic Chemistry.* 3rd ed. New York: John Wiley-Interscience, 1959–1969. This work is in 14 volumes.

K. B. Wiberg. *Laboratory Technique in Organic Chemistry.* New York: McGraw-Hill, 1960.

J. W. Zubrick. *The Organic Chem Lab Survival Manual: A Student's Guide to Techniques.* 4th ed. New York: John Wiley, 1997.

Numerous works and some general textbooks specialize in particular techniques. The above list is only representative of the most common books in this category. The following books deal specifically with micro and semi-microscale techniques.

N. D. Cheronis. "Micro and Semimicro Methods." In: A Weissberger, ed. *Technique of Organic Chemistry,* Volume 6. New York: John Wiley-Interscience, 1954.

N. D. Cheronis and T. S. Ma. *Organic Functional Group Analysis by Micro and Semimicro Methods.* New York: John Wiley-Interscience, 1964.

T. S. Ma and V. Horak. *Microscale Manipulations in Chemistry.* New York: John Wiley-Interscience, 1976.

20.8 REACTION MECHANISMS

As with the case of locating information on synthetic methods, you can obtain a great deal of information about reaction mechanisms by consulting one of the common textbooks on physical organic chemistry. The textbooks listed here provide a general description of mechanisms, but they do not contain specific literature citations. Very general textbooks include

A. Miller. *Writing Reaction Mechanisms in Organic Chemistry.* San Diego: Academic Press, 1992.

P. Sykes. *A Guidebook to Mechanism in Organic Chemistry.* 6th ed. London: Longman Group, Ltd., 1986.

More advanced textbooks include

F. A. Carey and R. J. Sundberg. *Advanced Organic Chemistry. Part A. Structure and Mechanisms.* 3rd ed. New York: Plenum Press, 1990.
L. P. Hammett. *Physical Organic Chemistry: Reaction Rates, Equilibria, and Mechanisms.* 2nd ed. New York: McGraw-Hill, 1970.
J. Hine. *Physical Organic Chemistry.* 2nd ed. New York: McGraw-Hill, 1962.
C. K. Ingold. *Structure and Mechanism in Organic Chemistry.* 2nd ed. Ithaca, NY: Cornell University Press, 1969.
N. S. Isaacs. *Physical Organic Chemistry.* 2nd ed. New York: John Wiley, 1995.
R. A. Y. Jones. *Physical and Mechanistic Organic Chemistry.* 2nd ed. Cambridge, U.K.: Cambridge University Press, 1984.
T. H. Lowry and K. S. Richardson. *Mechanism and Theory in Organic Chemistry.* 3rd ed. New York: Harper & Row, 1987.
J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992.
J. W. Moore and R. G. Pearson. *Kinetics and Mechanism.* 3rd ed. New York: John Wiley, 1981.

These books include extensive bibliographies that permit the reader to delve more deeply into the subject.

Most libraries also subscribe to annual series of publications that specialize in articles dealing with reaction mechanisms. Among these are

Advances in Physical Organic Chemistry. London: Academic Press, 1963–present.
Annual Reports of the Chemical Society. Section B. London: Chemical Society, 1905–present. Specifically, the section on *Reaction Mechanisms.*
Organic Reaction Mechanisms. Chichester, U.K.: John Wiley, 1965–present.
Progress in Physical Organic Chemistry. New York: Interscience, 1963–present.

These publications provide the reader with citations from the original literature that can be very useful in an extensive literature search.

20.9 ORGANIC QUALITATIVE ANALYSIS

Experiment 57 contains a procedure for identifying organic compounds through a series of chemical tests and reactions. Occasionally, you might require a more complete description of analytical methods or a more complete set of tables of derivatives. Textbooks specializing in organic qualitative analysis should fill this need. Examples of sources for such information include

N. D. Cheronis and J. B. Entriken. *Identification of Organic Compounds: A Student's Text Using Semimicro Techniques.* New York: Interscience, 1963.
D. J. Pasto and C. R. Johnson. *Laboratory Text for Organic Chemistry: A Source Book of Chemical and Physical Techniques.* Englewood Cliffs, NJ: Prentice-Hall, 1979.
Z. Rappoport, ed. *Handbook of Tables for Organic Compound Identification.* 3rd ed. Cleveland: Chemical Rubber Co., 1967.

R. L. Shriner, R. C. Fuson, D. Y. Curtin, and T. C. Morrill. *The Systematic Identification of Organic Compounds: A Laboratory Manual.* 6th ed. New York: John Wiley, 1980.

A. I. Vogel, *Elementary Practical Organic Chemistry. Part 2. Qualitative Organic Analysis.* 2nd ed. New York: John Wiley, 1966.

A. I. Vogel. Revised by members of the School of Chemistry, Thames Polytechnic. *Vogel's Textbook of Practical Organic Chemistry, Including Qualitative Organic Analysis.* 5th ed. London: Longman Group, Ltd., 1989.

20.10 BEILSTEIN AND CHEMICAL ABSTRACTS

One of the most useful sources of information about the physical properties, synthesis, and reactions of organic compounds is *Beilsteins Handbuch der Organischen Chemie.* This is a monumental work, initially edited by Friedrich Konrad Beilstein, and updated through several revisions by the Beilstein Institute in Frankfurt am Main, Germany. The original edition (the *Hauptwerk,* abbreviated H) was published in 1918 and covers completely the literature to 1909. Five supplementary series (*Ergänzungswerken*) have been published since that time. The first supplement (*Erstes Ergänzungswerk,* abbreviated E I) covers the literature from 1910–1919; the second supplement (*Zweites Ergänzungswerk,* E II) covers 1920–1929; the third supplement (*Drittes Ergänzungswerk,* E III) covers 1930–1949; the fourth supplement (*Viertes Ergänzungswerk,* E IV) covers 1950–1959; and the fifth supplement (in English) covers 1960–1979. Volumes 17–27 of supplementary series III and IV, covering heterocyclic compounds, are combined in a joint issue, E III/IV. Supplementary series III, IV, and V are not complete, so the coverage of *Handbuch der Organischen Chemie* can be considered complete to 1929, with partial coverage to 1979.

Beilsteins Handbuch der Organischen Chemie, usually referred to simply as *Beilstein,* also contains two types of cumulative indices. The first of these is a name index (*Sachregister*) and the second is a formula index (*Formelregister*). These indices are particularly useful for a person wishing to locate a compound in *Beilstein.*

The principal difficulty in using *Beilstein* is that it is written in German through the fourth supplement. The fifth supplement is in English. Although some reading knowledge of German is useful, you can obtain information from the work by learning a few key phrases. For example, *Bildung* is "formation" or "structure." *Darst* or *Darstellung* is "preparation," K_P or *Siedepunkt* is "boiling point," and *F* or *Schmelzpunkt* is "melting point." Furthermore, the names of some compounds in German are not cognates of the English names. Some examples are *Apfelsäure* for "malic acid" (*säure* means "acid"), *Harnstoff* for "urea," *Jod* for "iodine," and *Zimtsäure* for "cinnamic acid." If you have access to a German-English dictionary for chemists, many of these difficulties can be overcome. The best such dictionary is

A. M. Patterson. *German–English Dictionary for Chemists.* 3rd ed. New York: John Wiley, 1959.

Beilstein is organized according to a very sophisticated and complicated system. To locate a compound in *Beilstein,* you can learn all the intricacies of this system. However, most students do not wish to become experts on *Beilstein* to this extent. A simpler, though slightly less reliable, method is to look for the compound in the formula index that accompanies the second supplement. Under the molecular formula, you will find the names of compounds that have that formula. After that name will be a series of numbers that indicate the pages and volume in which that compound is listed. Suppose, as an example, that you are searching for information on *p*-nitroaniline. This compound has the molecular formula $C_6H_6N_2O_2$. Searching for this formula in the formula index to the second supplement, you find

4-Nitro-anilin **12** 711, **I** 349, **II** 383

This information tells us that *p*-nitroaniline is listed in the main edition (*Hauptwerk*) in volume 12, page 711. Locating this particular volume, which is devoted to isocyclic monoamines, we turn to page 711 and find the beginning of the section on *p*-nitroaniline. At the left side of the top of this page, we find "Syst. No. 1671." This is the system number given to compounds in this part of Volume 12. The system number is useful, as it can help you find entries for this compound in subsequent supplements. The organization of *Beilstein* is such that all entries on *p*-nitroaniline in each of the supplements will be found in Volume 12. The entry in the formula index also indicates that material on this compound may be found in the first supplement on page 349 and in the second supplement on page 383. On page 349 of volume 12 of the first supplement, there is a heading, "**XII, 710–712,**" and on the left is "Syst. No. 1671." Material on *p*-nitroaniline is found in each supplement on a page that is headed with the volume and page of the *Hauptwerk* in which the same compound is found. On page 383 of volume 12 of the second supplement, the heading in the center of the top of the page is "**H12, 710–712.**" On the left you find "Syst. No. 1671." Again, because *p*-nitroaniline appeared in volume 12, page 711, of the main edition, you can locate it by searching through volume 12 of any supplement until you find a page with the heading corresponding to volume 12, page 711. Because the third and fourth supplements are not complete, there is no comprehensive formula index for these supplements. However, you can still find material on *p*-nitroaniline by using the system number and the volume and page in the main work. In the third supplement, because the amount of information available has grown so much since the early days of Beilstein's work, volume 12 has now expanded so that it is found in several bound parts. However, you select the part that includes system number 1671. In this part of volume 12, you look through the pages until you find a page headed "Syst. No. 1671/H 711." The information on *p*-nitroaniline is found on this page (p. 1580). If volume 12 of the fourth supplement were available, you would go on in the same way to locate more recent data on *p*-nitroaniline. This example is meant to illustrate how you can locate information on particular compounds without having to learn the Beilstein system of classification. You might do well to test your ability at finding compounds in *Beilstein* as we have described here.

Guidebooks to using *Beilstein*, which include a description of the Beilstein system, are recommended for anyone who wants to work extensively with *Beilstein*. Among such sources are

E. H. Huntress. *A Brief Introduction to the Use of Beilsteins Handbuch der Organischen Chemie.* 2nd ed. New York: John Wiley, 1938.

How to Use Beilstein. Beilstein Institute, Frankfurt am Main. Berlin: Springer-Verlag.

O. Weissbach. *The Beilstein Guide: A Manual for the Use of Beilsteins Handbuch der Organischen Chemie.* New York: Springer-Verlag, 1976.

Beilstein reference numbers are listed in such handbooks as *CRC Handbook of Chemistry and Physics* and *Lange's Handbook of Chemistry.* Additionally, Beilstein numbers are included in the *Aldrich Catalog Handbook of Fine Chemicals,* issued by the Aldrich Chemical Company. If the compound you are seeking is listed in one of these handbooks, you will find that using *Beilstein* is simplified.

Another very useful publication for finding references for research on a particular topic is *Chemical Abstracts,* published by the Chemical Abstracts Service of the American Chemical Society. *Chemical Abstracts* contains abstracts of articles appearing in more than 10,000 journals from virtually every country conducting scientific research. These abstracts list the authors, the journal in which the article appeared, the title of the paper, and a short summary of the contents of the article. Abstracts of articles that appeared originally in a foreign language are provided in English, with a notation indicating the original language.

To use *Chemical Abstracts,* you must know how to use the various indices that accompany it. At the end of each volume there appears a set of indices, including a formula index, a general subject index, a chemical substances index, an author index, and a patent index. The listings in each index refer the reader to the appropriate abstract according to the number assigned to it. There are also collective indices that combine all the indexed material appearing in a 5-year period (10-year period before 1956). In the collective indices, the listings include the volume number as well as the abstract number.

For material after 1929, *Chemical Abstracts* provides the most complete coverage of the literature. For material before 1929, use *Beilstein* before consulting *Chemical Abstracts. Chemical Abstracts* has the advantage that it is written entirely in English. Nevertheless, most students perform a literature search to find a relatively simple compound. Finding the desired entry for a simple compound is much easier in *Beilstein* than in *Chemical Abstracts.* For simple compounds, the indices in *Chemical Abstracts* are likely to contain very many entries. To locate the desired information, you must comb through this multitude of listings—potentially a very time-consuming task.

The opening pages of each index in *Chemical Abstracts* contain a brief set of instructions on using that index. If you want a more complete guide to *Chemical Abstracts,* consult a textbook designed to familiarize you with these abstracts and indices. Two such books are

CAS Printed Access Tools: A Workbook. Washington, D.C.: Chemical Abstracts Service, American Chemical Society, 1977.

How to Search Printed CA. Washington, D.C.: Chemical Abstracts Service, American Chemical Society, 1989.

Chemical Abstracts Service maintains a computerized database that permits users to search through *Chemical Abstracts* very rapidly and thoroughly. This service, which is called *CA Online,* is described in Section 20.11. *Beilstein* is also available for online searching by computer.

20.11 COMPUTER ONLINE SEARCHING

You can search a number of chemistry databases online by using a computer and modem. Many academic and industrial libraries can access these databases through their computers. One organization that maintains a large number of databases is the Scientific and Technical Information Network (STN International). The fee charged to the library for this service depends on the total time used in making the search, the type of information being asked for, the time of day when the search is being conducted, and the type of database being searched.

The Chemical Abstracts Service database (*CA Online*) is one of many databases available on STN. It is particularly useful to chemists. Unfortunately, this database extends back only to about 1967, although some earlier references are available. Searches earlier than 1967 must be made with printed abstracts (Section 20.10). An online search is much faster than searching in the printed abstracts. In addition, you can tailor the search in a number of ways by using keywords and the Chemical Abstracts Substance Registry Number (CAS number) as part of the search routine. For the more common organic compounds, you can easily obtain CAS numbers from the catalogues of most of the companies that supply chemicals. Another advantage of performing an online search is that the *Chemical Abstracts* files are updated much more quickly than the printed versions of abstracts. This means that your search is more likely to reveal the most current information available.

Other useful databases available from STN include *Beilstein* and *CASREACTS*. As described in Section 20.10, *Beilstein* is very useful to organic chemists. Currently, there are over 3.5 million compounds listed in the database. You can use the CAS Registry Numbers to help in a search that has the potential of going back to 1830. *CASREACTS* is a chemical reactions database derived from over 100 journals covered by *Chemical Abstracts,* starting in 1985. With this database, you can specify a starting material and a product using the CAS Registry Numbers. Further information on *CA Online, Beilstein, CASREACTS,* and other databases can be obtained from the following references:

J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure.* 4th ed. New York: John Wiley, 1992. Appendix A contains a summary.

A. N. Somerville. "Information Sources for Organic Chemistry, 2: Searching by Functional Group." *Journal of Chemical Education, 68* (October 1991): 842.

A. N. Somerville. "Subject Searching of Chemical Abstracts Online." *Journal of Chemical Education, 70* (March 1993): 200.

G. Wiggins. *Chemical Information Sources.* New York: McGraw Hill, 1990. Integrates printed materials and computer sources of information.

20.12 SCIENTIFIC JOURNALS

Ultimately, someone wanting information about a particular area of research will be required to read articles from the scientific journals. These journals are of two basic types: review journals and primary scientific journals. Journals that specialize in review articles summarize all the work that bears on the particular topic. These articles may focus on the contributions of one particular researcher but often consider the contributions of many researchers to the subject. These articles also contain extensive bibliographies, which refer you to the original research articles. Among the important journals devoted, at least partly, to review articles are

Accounts of Chemical Research
Angewandte Chemie (International Edition, in English)
Chemical Reviews
Chemical Society Reviews (formerly known as *Quarterly Reviews*)
Nature
Science

The details of the research of interest appear in the primary scientific journals. Although there are thousands of journals published in the world, a few important journals specializing in articles dealing with organic chemistry might be mentioned here. These are

Canadian Journal of Chemistry
Chemische Berichte
Journal of the American Chemical Society
Journal of the Chemical Society, Chemical Communications
Journal of the Chemical Society, Perkin Transactions (Parts I and II)
Journal of Organic Chemistry
Journal of Organometallic Chemistry
Liebigs Annalen der Chemie
Organometallics
Synthesis
Tetrahedron
Tetrahedron Letters

20.13 TOPICS OF CURRENT INTEREST

The following journals and magazines are good sources for topics of educational and current interest. They specialize in news articles and focus on current events in chemistry or in science in general. Articles in these journals (magazines) can be useful in keeping you abreast of developments in science that are not part of your normal specialized scientific reading.

American Scientist
Chemical and Engineering News
Chemistry in Britain

Chemistry and Industry
Chemtech
Discover
Journal of Chemical Education
Nature
New Scientist
Omni
Science
Science Digest
Scientific American
SciQuest (formerly *Chemistry*)

20.14 HOW TO CONDUCT A LITERATURE SEARCH

The easiest method to follow in searching the literature is to begin with secondary sources and then go to the primary sources. In other words, you would try to locate material in a textbook, *Beilstein,* or *Chemical Abstracts.* From the results of that search, you would then consult one of the primary scientific journals.

A literature search that ultimately requires you to read one or more papers in the scientific journals is best conducted if you can identify a particular paper central to the study. Often, you can obtain this reference from a textbook or a review article on the subject. If this is not available, a search through *Beilstein* is required. A search through one of the handbooks that provides *Beilstein* reference numbers (see Section 20.10) may be helpful. Searching through *Chemical Abstracts* would be considered the next logical step. From these sources, you should be able to identify citations from the original literature on the subject.

Additional citations may be found in the references cited in the journal article. In this way, the background leading to the research can be examined. It is also possible to conduct a search forward in time from the date of the journal article through the *Science Citation Index.* This publication provides the service of listing articles and the papers in which these articles were cited. Although the *Science Citation Index* consists of several types of indices, the *Citation Index* is most useful for the purposes described here. A person who knows of a particular key reference on a subject can examine the *Science Citation Index* to obtain a list of papers that have used that seminal reference in support of the work described. The *Citation Index* lists papers by their senior author, journal, volume, page, and date, followed by citations of papers that have referred to that article, author, journal, volume, page, and date of each. The *Citation Index* is published in annual volumes, with quarterly supplements issued during the current year. Each volume contains a complete list of the citations of the key articles made during that year. A disadvantage is that *Science Citation Index* has been available only since 1961. An additional disadvantage is that you may miss journal articles on the subject of interest if they failed to cite that particular key reference in their bibliographies—a reasonably likely possibility.

You can, of course, conduct a literature search by a "brute force" method, by beginning the search with *Beilstein* or even with the indices in *Chemical Abstracts.* However, the task can be made much easier by performing a computer search (Section 20.11) or by starting with a book or an article of general and broad coverage, which can provide a few citations for starting points in the search.

The following guides to using the chemical literature are provided for the reader who is interested in going further into this subject.

R. T. Bottle and J. F. B. Rowland, eds. *Information Sources in Chemistry,* 4th ed. New York: Bowker-Saur, 1992.
M. G. Mellon. *Chemical Publications.* 5th ed. New York: McGraw-Hill, 1982.
R. E. Maizell. *How to Find Chemical Information: A Guide for Practicing Chemists, Teachers, and Students,* 2nd ed. New York: John Wiley-Interscience, 1987.
G. Wiggins. *Chemical Information Sources.* New York: McGraw Hill, 1990. Integrates printed materials and computer sources of information.

PROBLEMS

1. Using the *Merck Index* discussed in Section 20.1, find and draw structures for the following compounds:
  (a) atropine
  (b) quinine
  (c) saccharin
  (d) benzo[a]pyrene (benzpyrene)
  (e) itaconic acid
  (f) adrenosterone
  (g) chrysanthemic acid (chrysanthemumic acid)
  (h) cholesterol
  (i) vitamin C (ascorbic acid)

2. Find the melting points for the following compounds in the *Handbook of Chemistry and Physics* or *Lange's Handbook of Chemistry* (Section 20.1):
  (a) biphenyl
  (b) 4-bromobenzoic acid
  (c) 3-nitrophenol

3. Find the boiling point for each compound in the references listed in Problem 2:
  (a) octanoic acid at reduced pressure
  (b) 4-chloroacetophenone at atmosphere and reduced pressure
  (c) 2-methyl-2-heptanol

4. Find the index of refraction n_D and density for the liquids listed in Problem 3.

5. Using the references given in Problem 2, give the specific rotations $[\alpha]_D$ for the enantiomers of camphor.

6. Read the section on carbon tetrachloride in the *Merck Index* and list some of the health hazards for this compound.

7. Find the following compounds in the formula index for the *Second Supplement of Beilstein* (Section (20.10). (i) List the page numbers from the Main Work and the Supplements (First and Second). (ii) Using these page numbers, look up the System Number (Syst. No.) and the Main Work number (Hauptwerk number, H) for each compound in the Main Work, and the First and

Second supplements. In some cases, a compound may not be found in all three places. (iii) Now use the System Number and Main Work number to find each of these compounds in the Third and Fourth Supplements. List the page numbers where these compounds are found.

- **(a)** 2,5-hexanedione (acetonylacetone)
- **(b)** 3-nitroacetophenone
- **(c)** 4-*tert*-butylcyclohexanone
- **(d)** 4-phenylbutanoic acid (4-phenylbutyric acid, γ-phenylbuttersäure)

8. Using the *Science Citation Index* (Section 20.14), list five research papers by complete title and journal citation for each of the following chemists who have been awarded the Nobel Prize. Use the *Five-Year Cumulation Source Index* for the years 1980–1984 as your source.

- **(a)** H. C. Brown
- **(b)** R. B. Woodward
- **(c)** D. J. Cram
- **(d)** G. Olah

9. The reference book by J. March is listed in Section 20.2. Using Appendix B in this book, give two methods for preparing the following functional groups. You will need to provide equations.

- **(a)** carboxylic acids
- **(b)** aldehydes
- **(c)** esters (Carboxylic esters)

10. *Organic Synthesis* is described in Section 20.6. There are currently eight collective volumes in the series, each with its own index. Find the compounds listed below and provide the equations for preparing each compound.

- **(a)** 2-methylcyclopentane-1,3-dione
- **(b)** *cis*-Δ^4-tetrahydrophthalic anhydride (listed as tetrahydrophthalic anhydride)

11. Provide four ways that may be used to oxidize an alcohol to an aldehyde. Give complete literature references for each method as well as equations. Use the *Compendium of Organic Synthetic Methods* or *Survey of Organic Syntheses* by Buehler and Pearson (Section 20.6).

Appendices

APPENDIX 1

Tables of Unknowns and Derivatives

More extensive tables of unknowns may be found in Z. Rappoport, ed. *Handbook of Tables for Organic Compound Identification,* 3rd ed. Cleveland: Chemical Rubber Co., 1967

ALDEHYDES

Compound	BP	MP	Semi-carbazone*	2,4-Dinitro-phenyl-hydrazone*
Ethanal (acetaldehyde)	21	—	162	168
Propanal (propionaldehyde)	48	—	89	148
Propenal (acrolein)	52	—	171	165
2-Methylpropanal (isobutyraldehyde)	64	—	125	187
Butanal (butyraldehyde)	75	—	95	123
3-Methylbutanal (isovaleraldehyde)	92	—	107	123
Pentanal (valeraldehyde)	102	—	—	106
2-Butenal (crotonaldehyde)	104	—	199	190
2-Ethylbutanal (diethylacetaldehyde)	117	—	99	95
Hexanal (caproaldehyde)	130	—	106	104
Heptanal (heptaldehyde)	153	—	109	108
2-Furaldehyde (furfural)	162	—	202	212
2-Ethylhexanal	163	—	254	114
Octanal (caprylaldehyde)	171	—	101	106
Benzaldehyde	179	—	222	237
Phenylethanal (phenylacetaldehyde)	195	33	153	121
2-Hydroxybenzaldehyde (salicylaldehyde)	197	—	231	248
4-Methylbenzaldehyde (*p*-tolualdehyde)	204	—	234	234
3,7-Dimethyl-6-octenal (citronellal)	207	—	82	77
2-Chlorobenzaldehyde	213	11	229	213
4-Methoxybenzaldehyde (*p*-anisaldehyde)	248	2.5	210	253
trans-Cinnamaldehyde	250 d.	—	215	255
3,4-Methylenedioxybenzaldehyde (piperonal)	263	37	230	266 d.
2-Methoxybenzaldehyde (*o*-anisaldehyde)	245	38	215 d.	254
4-Chlorobenzaldehyde	214	48	230	254
3-Nitrobenzaldehyde	—	58	246	293
4-Dimethylaminobenzaldehyde	—	74	222	325
Vanillin	285 d.	82	230	271

ALDEHYDES *(Cont.)*

Compound	BP	MP	Semi-carbazone*	2,4-Dinitro-phenyl-hydrazone*
4-Nitrobenzaldehyde	—	106	221	320 d.
4-Hydroxybenzaldehyde	—	116	224	280 d.
(±)-Glyceraldehyde	—	142	160 d.	167

NOTE: "*d*" indicates "*decomposition.*"
*See Appendix 2, "Procedures for Preparing Derivatives."

KETONES

Compound	BP	MP	Semi-carbazone*	2,4-Dinitro-phenyl hydrazone*
2-Propanone (acetone)	56	—	187	126
2-Butanone (methyl ethyl ketone)	80	—	146	117
3-Methyl-2-butanone (isopropyl methyl ketone)	94	—	112	120
2-Pentanone (methyl propyl ketone)	101	—	112	143
3-Pentanone (diethyl ketone)	102	—	138	156
Pinacolone	106	—	157	125
4-Methyl-2-pentanone (isobutyl methyl ketone)	117	—	132	95
2,4-Dimethyl-3-pentanone (diisopropyl ketone)	124	—	160	95
2-Hexanone (methyl butyl ketone)	128	—	125	106
4-Methyl-3-penten-2-one (mesityl oxide)	130	—	164	205
Cyclopentanone	131	—	210	146
2,3-Pentanedione	134	—	122 (mono) 209 (di)	209
2,4-Pentanedione (acetylacetone)	139	—	—	122 (mono) 209 (di)
4-Heptanone (dipropyl ketone)	144	—	132	75
2-Heptanone (methyl amyl ketone)	151	—	123	89
Cyclohexanone	156	—	166	162
2,6-Dimethyl-4-heptanone (diisobutyl ketone)	168	—	122	92
2-Octanone	173	—	122	58
Cycloheptanone	181	—	163	148
2,5-Hexanedione (acetonylacetone)	191	−9	185 (mono) 224 (di)	257 (di)
Acetophenone (methyl phenyl ketone)	202	20	198	238

KETONES *(Cont.)*

Compound	BP	MP	Semi-carbazone*	2,4-Dinitro-phenyl hydrazone*
Phenyl-2-propanone (phenylacetone)	216	27	198	156
Propiophenone (ethyl phenyl ketone)	218	21	182	191
4-Methylacetophenone	226	—	205	258
2-Undecanone	231	12	122	63
4-Chloroacetophenone	232	12	204	236
4-Phenyl-2-butanone (benzylacetone)	235	—	142	127
4-Chloropropiophenone	—	36	176	223
4-Phenyl-3-buten-2-one	—	37	187	227
4-Methoxyacetophenone	258	38	198	228
Benzophenone	305	48	167	238
4-Bromoacetophenone	225	51	208	230
2-Acetonaphthone	—	54	235	262
Desoxybenzoin	320	60	148	204
3-Nitroacetophenone	202	80	257	228
9-Fluorenone	345	83	234	283
Benzoin	344	136	206	245
4-Hydroxypropiophenone	—	148	—	229
(±)-Camphor	205	179	237	177

*See Appendix 2, "Procedures for Preparing Derivatives."

CARBOXYLIC ACIDS

Compound	BP	MP	*p*-Toluidide*	Anilide*	Amide*
Formic acid	101	8	53	47	43
Acetic acid	118	17	148	114	82
Propenoic acid (acrylic acid)	139	13	141	104	85
Propanoic acid (propionic acid)	141	—	124	103	81
2-Methylpropanoic acid (isobutyric acid)	154	—	104	105	128
Butanoic acid (butyric acid)	162	—	72	95	115
2-Methylpropenoic acid (methacrylic acid)	163	16	—	87	102
Trimethylacetic acid (pivalic acid)	164	35	—	127	178
Pyruvic acid	165 d.	14	109	104	124
3-Methylbutanoic acid (isovaleric acid)	176	—	109	109	135

CARBOXYLIC ACIDS *(Cont.)*

Compound	BP	MP	*p*-Toluidide*	Anilide*	Amide*
Pentanoic acid (valeric acid)	186	—	70	63	106
2-Methylpentanoic acid	186	—	80	95	79
2-Chloropropanoic acid	186	—	124	92	80
Dichloroacetic acid	194	6	153	118	98
Hexanoic acid (caproic acid)	205	—	75	95	101
2-Bromopropanoic acid	205	24	125	99	123
Octanoic acid (caprylic acid)	237	16	70	57	107
Nonanoic acid	254	12	84	57	99
Decanoic acid (capric acid)	268	32	78	70	108
4-Oxopentanoic acid (levulinic acid)	246	33	108	102	108 d.
Dodecanoic acid (lauric acid)	299	43	87	78	100
3-Phenylpropanoic acid (hydrocinnamic acid)	279	48	135	98	105
Bromoacetic acid	208	50	—	131	91
Tetradecanoic acid (myristic acid)	—	54	93	84	103
Trichloroacetic acid	198	57	113	97	141
Hexadecanoic acid (palmitic acid)	—	62	98	90	106
Chloroacetic acid	189	63	162	137	121
Octadecanoic acid (stearic acid)	—	69	102	95	109
trans-2-Butenoic acid (crotonic acid)	—	72	132	118	158
Phenylacetic acid	—	77	136	118	156
2-Methoxybenzoic acid (*o*-anisic acid)	200	101	—	131	129
2-Methylbenzoic acid (*o*-toluic acid)	—	104	144	125	142
Nonanedioic acid (azelaic acid)	—	106	201 (di)	107 (mono) 186 (di)	93 (mono) 175 (di)
3-Methylbenzoic acid (*m*-toluic acid)	263 s.	110	118	126	94
(±)-Phenylhydroxyacetic acid (mandelic acid)	—	118	172	151	133
Benzoic acid	249	122	158	163	130
2-Benzoylbenzoic acid	—	127	—	195	165
Maleic acid	—	130	142 (di)	198 (mono) 187 (di)	172 (mono) 260 (di)
Decanedioic acid (sebacic acid)	—	133	201 (di)	122 (mono) 200 (di)	170 (mono) 210 (di)
Cinnamic acid	300	133	168	153	147

NOTE: "*s*" indicates "*sublimination*"; "*d*" indicates "*decomposition*."

*See Appendix 2, "Procedures for Preparing Derivatives."

CARBOXYLIC ACIDS (*Cont.*)

Compound	BP	MP	*p*-Toluidide*	Anilide*	Amide*
2-Chlorobenzoic acid	—	140	131	118	139
3-Nitrobenzoic acid	—	140	162	155	143
2-Aminobenzoic acid (anthranilic acid)	—	146	151	131	109
Diphenylacetic acid	—	148	172	180	167
2-Bromobenzoic acid	—	150	—	141	155
Benzilic acid	—	150	190	175	154
Hexanedioic acid (adipic acid)	—	152	239	151 (mono) 241 (di)	125 (mono) 220 (di)
Citric acid	—	153	189 (tri)	199 (tri)	210 (tri)
4-Chlorophenoxyacetic acid	—	158	—	125	133
2-Hydroxybenzoic acid (salicylic acid)	—	158	156	136	142
5-Bromo-2-hydroxybenzoic acid (5-bromosalicylic acid)	—	165	—	222	232
Methylenesuccinic acid (itaconic acid)	—	166 d.	—	152 (mono)	191 (di)
(+)-Tartaric acid	—	169	—	180 (mono) 264 (di)	171 (mono) 196 (di)
4-Chloro-3-nitrobenzoic acid	—	180	—	131	156
4-Methylbenzoic acid (*p*-toluic acid)	—	180	160	145	160
4-Methoxybenzoic acid (*p*-anisic acid)	280	184	186	169	167
Butanedioic acid (succinic acid)	235 d.	188	180 (mono) 255 (di)	143 (mono) 230 (di)	157 (mono) 260 (di)
3-Hydroxybenzoic acid	—	201	163	157	170
3,5-Dinitrobenzoic acid	—	202	—	234	183
Phthalic acid	—	210 d.	150 (mono) 201 (di)	169 (mono) 253 (di)	144 (mono) 220 (di)
4-Hydroxybenzoic acid	—	214	204	197	162
Pyridine-3-carboxylic acid (nicotinic acid)	—	236	150	132	128
4-Nitrobenzoic acid	—	240	204	211	201
4-Chlorobenzoic acid	—	242	—	194	179
Fumaric acid	—	300	—	233 (mono) 314 (di)	270 (mono) 266 (di)

NOTE: "*d*" indicates "*decomposition.*"

*See Appendix 2, "Procedures for Preparing Derivatives."

PHENOLS†

Compound	BP	MP	α-Naphthylurethane*	Bromo Derivative* *Mono*	*Di*	*Tri*	*Tetra*
2-Chlorophenol	176	7	120	48	76	—	—
3-Methylphenol (*m*-cresol)	203	12	128	—	—	84	—
2-Methylphenol (*o*-cresol)	191	32	142	—	56	—	—
2-Methoxyphenol (guaiacol)	204	32	118	—	—	116	—
4-Methylphenol (*p*-cresol)	202	34	146	—	49	—	198
Phenol	181	42	133	—	—	95	—
4-Chlorophenol	217	43	166	33	90	—	—
2,4-Dichlorophenol	210	45	—	68	—	—	—
4-Ethylphenol	219	45	128	—	—	—	—
2-Nitrophenol	216	45	113	—	117	—	—
2-Isopropyl-5-methylphenol (thymol)	234	51	160	55	—	—	—
3,4-Dimethylphenol	225	64	141	—	—	171	—
4-Bromophenol	238	64	169	—	—	95	—
3,5-Dimethylphenol	220	68	109	—	—	166	—
2,5-Dimethylphenol	212	75	173	—	—	178	—
1-Naphthol (α-naphthol)	278	96	152	—	105	—	—
2-Hydroxyphenol (catechol)	245	104	175	—	—	—	192
3-Hydroxyphenol (resorcinol)	281	109	275	—	—	112	—
4-Nitrophenol	—	112	150	—	142	—	—
2-Naphthol (β-naphthol)	286	121	157	84	—	—	—
1,2,3-Trihydroxybenzene (pyrogallol)	309	133	—	—	158	—	—
4-Phenylphenol	305	164	—	—	—	—	—

*See Appendix 2, "Procedures for Preparing Derivatives."

†Also check:
Salicylic acid (2-hydroxybenzoic acid)
Esters of salicylic acid (salicylates)
Salicylaldehyde (2-hydroxybenzaldehyde)
4-Hydroxybenzaldehyde
4-Hydroxypropiophenone
3-Hydroxybenzoic acid
4-Hydroxybenzoic acid
4-Hydroxybenzophenone

PRIMARY AMINES†

Compound	BP	MP	Benzamide*	Picrate*	Acetamide*
t-Butylamine	46	—	134	198	101
Propylamine	48	—	84	135	—

PRIMARY AMINES† *(Cont.)*

Compound	BP	MP	Benzamide*	Picrate*	Acetamide*
Allylamine	56	—	—	140	—
sec-Butylamine	63	—	76	139	—
Isobutylamine	69	—	57	150	—
Butylamine	78	—	42	151	—
Cyclohexylamine	135	—	149	—	104
Furfurylamine	145	—	—	150	—
Benzylamine	184	—	105	194	60
Aniline	184	—	163	198	114
2-Methylaniline (*o*-toluidine)	200	—	144	213	110
3-Methylaniline (*m*-toluidine)	203	—	125	200	65
2-Chloroaniline	208	—	99	134	87
2,6-Dimethylaniline	216	11	168	180	177
2-Methoxyaniline (*o*-anisidine)	225	6	60	200	85
3-Chloroaniline	230	—	120	177	74
2-Ethoxyaniline (*o*-phenetidine)	231	—	104	—	79
4-Chloro-2-methylaniline	241	29	142	—	140
4-Ethoxyaniline (*p*-phenetidine)	250	2	173	69	137
4-Methylaniline (*p*-toluidine)	200	43	158	182	147
2-Ethylaniline	210	47	147	194	111
2,5-Dichloroaniline	251	50	120	86	132
4-Methoxyaniline (*p*-anisidine)	—	58	154	170	130
4-Bromoaniline	245	64	204	180	168
2,4,5-Trimethylaniline	—	64	167	—	162
4-Chloroaniline	—	70	192	178	179
2-Nitroaniline	—	72	110	73	92
Ethyl *p*-aminobenzoate	—	89	148	—	110
o-Phenylenediamine	258	102	301 (di)	208	185 (di)
2-Methyl-5-nitroaniline	—	106	186	—	151
2-Chloro-4-nitroaniline	—	108	161	—	139
3-Nitroaniline	—	114	157	143	155
4-Chloro-2-nitroaniline	—	118	—	—	104
2,4,6-Tribromoaniline	300	120	200	—	232 (mono) 127 (di)
2-Methyl-4-nitroaniline	—	130	—	—	202
2-Methoxy-4-nitroaniline	—	138	149	—	153
p-Phenylenediamine	267	140	128 (mono) 300 (di)	—	162 (mono) 304 (di)
4-Nitroaniline	—	148	199	100	215
4-Aminoacetanilide	—	162	—	—	304
2,4-Dinitroaniline	—	180	202	—	120

*See Appendix 2, "Procedures for Preparing Derivatives."

†Also check: 4-Aminobenzoic acid and its esters.

SECONDARY AMINES

Compound	BP	MP	Benzamide*	Picrate*	Acetamide*
Diethylamine	56	—	42	155	—
Diisopropylamine	84	—	—	140	—
Pyrrolidine	88	—	Oil	112	—
Piperidine	106	—	48	152	—
Dipropylamine	110	—	—	75	—
Morpholine	129	—	75	146	—
Diisobutylamine	139	—	—	121	86
N-Methylcyclohexylamine	148	—	85	170	—
Dibutylamine	159	—	—	59	—
Benzylmethylamine	184	—	—	117	—
N-Methylaniline	196	—	63	145	102
N-Ethylaniline	205	—	60	132	54
N-Ethyl-*m*-toluidine	221	—	72	—	—
Dicyclohexylamine	256	—	153	173	103
N-Benzylaniline	298	37	107	48	58
Indole	254	52	68	—	157
Diphenylamine	302	52	180	182	101
N-Phenyl-1-naphthylamine	335	62	152	—	115

*See Appendix 2, "Procedures for Preparing Derivatives."

TERTIARY AMINES†

Compound	BP	MP	Picrate*	Methiodide*
Triethylamine	89	—	173	280
Pyridine	115	—	167	117
2-Methylpyridine (α-picoline)	129	—	169	230
3-Methylpyridine (β-picoline)	144	—	150	92
Tripropylamine	157	—	116	207
N,N-Dimethylbenzylamine	183	—	93	179
N,N-Dimethylaniline	193	—	163	228 d.
Tributylamine	216	—	105	186
N,N-Diethylaniline	217	—	142	102
Quinoline	237	—	203	133

NOTE: "*d*" indicates "*decomposition.*"

*See Appendix 2, "Procedures for Preparing Derivatives."

†Also check: Nicotinic acid and its esters.

ALCOHOLS

Compound	BP	MP	3,5-Dinitrobenzoate*	Phenylurethane*
Methanol	65	—	108	47
Ethanol	78	—	93	52
2-Propanol (isopropyl alcohol)	82	—	123	88
2-Methyl-2-propanol (*t*-butyl alcohol)	83	26	142	136
2-Propen-1-ol (allyl alcohol)	97	—	49	70
1-Propanol	97	—	74	57
2-Butanol (*sec*-butyl alcohol)	99	—	76	65
2-Methyl-2-butanol (*t*-pentyl alcohol)	102	−8.5	116	42
2-Methyl-3-butyn-2-ol	104	—	112	—
2-Methyl-1-propanol (isobutyl alcohol)	108	—	87	86
2-Propyn-1-ol (propargyl alcohol)	114	—	—	—
3-Pentanol	115	—	101	48
1-Butanol	118	—	64	61
2-Pentanol	119	—	62	—
3-Methyl-3-pentanol	123	—	96	43
2-Methoxyethanol	124	—	—	(113)†
2-Chloroethanol	129	—	95	51
3-Methyl-1-butanol (isoamyl alcohol)	130	—	70	31
4-Methyl-2-pentanol	132	—	65	143
1-Pentanol	138	—	46	46
Cyclopentanol	140	—	115	132
2-Ethyl-1-butanol	146	—	51	—
2,2,2-Trichloroethanol	151	—	142	87
1-Hexanol	157	—	58	42
Cyclohexanol	160	—	113	82
(2-Furyl)-methanol (furfuryl alcohol)	170	—	80	45
1-Heptanol	176	—	47	60
2-Octanol	179	—	32	114
1-Octanol	195	—	61	74
3,7-Dimethyl-1,6-octadien-3-ol (linalool)	196	—	—	66
Benzyl alcohol	204	—	113	77
1-Phenylethanol	204	20	92	95
2-Phenylethanol	219	—	108	78
1-Decanol	231	7	57	59
3-Phenylpropanol	236	—	45	92
1-Dodecanol (lauryl alcohol)	—	24	60	74
3-Phenyl-2-propen-1-ol (cinnamyl alcohol)	250	34	121	90
1-Tetradecanol (myristyl alcohol)	—	39	67	74
(−)-Menthol	212	41	158	111
1-Hexadecanol (cetyl alcohol)	—	49	66	73

ALCOHOLS *(Cont.)*

Compound	BP	MP	3,5-Dinitrobenzoate*	Phenylurethane*
1-Octadecanol (stearyl alcohol)	—	59	77	79
Diphenylmethanol (benzhydrol)	288	68	141	139
Benzoin	—	133	—	165
Cholesterol	—	147	—	168
(+)-Borneol	—	208	154	138

*See Appendix 2, "Procedures for Preparing Derivatives."

†α-Naphthylurethane.

ESTERS

Compound	BP	MP
Methyl formate	34	—
Ethyl formate	54	—
Vinyl acetate	72	—
Ethyl acetate	77	—
Methyl propanoate (methyl propionate)	80	—
Methyl acrylate	80	—
2-Propyl acetate (isopropyl acetate)	85	—
Ethyl chloroformate	93	—
Methyl 2-methylpropanoate (methyl isobutyrate)	93	—
2-Propenyl acetate (isopropenyl acetate)	94	—
2-(2-Methylpropyl acetate (*t*-butyl acetate)	98	—
Ethyl acrylate	99	—
Ethyl propanoate (ethyl propionate)	99	—
Methyl methacrylate	100	—
Methyl trimethylacetate (methyl pivalate)	101	—
Propyl acetate	102	—
Methyl butanoate (methyl butyrate)	102	—
2-Butyl acetate (*sec*-butyl acetate)	111	—
Methyl 3-methylbutanoate (methyl isovalerate)	117	—
Ethyl butanoate (ethyl butyrate)	120	—
Butyl acetate	127	—
Methyl pentanoate (methyl valerate)	128	—
Methyl chloroacetate	130	—
Ethyl lactate	154	—
Ethyl hexanoate (ethyl caproate)	168	—
Methyl acetoacetate	170	—
Dimethyl malonate	180	—
Ethyl acetoacetate	181	—
Diethyl oxalate	185	—
Methyl benzoate	199	—
Ethyl octanoate (ethyl caprylate)	207	—
Ethyl cyanoacetate	210	—
Ethyl benzoate	212	—
Diethyl succinate	217	—
Methyl phenylacetate	218	—
Diethyl fumarate	219	—
Methyl salicylate	222	—
Diethyl maleate	225	—
Ethyl phenylacetate	229	—
Ethyl salicylate	234	—
Dimethyl suberate	268	—
Ethyl cinnamate	271	—
Diethyl phthalate	298	—
Dibutyl phthalate	340	—
Methyl cinnamate	—	36
Phenyl salicylate	—	42
Methyl *p*-chlorobenzoate	—	44
Ethyl *p*-nitrobenzoate	—	56
Phenylbenzoate	314	69

ESTERS *(Cont.)*

Compound	BP	MP	Compound	BP	MP
Ethyl 3-methylbutanoate (ethyl isovalerate)	132	—	Methyl *m*-nitrobenzoate	—	78
Pentyl acetate (*n*-amyl acetate)	142	—	Methyl *p*-bromobenzoate	—	81
3-Methylbutyl acetate (isoamyl acetate)	142	—	Ethyl *p*-aminobenzoate	—	90
Ethyl chloroacetate	143	—	Methyl *p*-nitrobenzoate	—	94

APPENDIX 2

Procedures for Preparing Derivatives

Caution: Some of the chemicals used in preparing derivatives are suspected carcinogens. The list of suspected carcinogens on page 21 should be consulted before beginning any of these procedures. Care should be exercised in handling these substances.

ALDEHYDES AND KETONES

Semicarbazones. Place 0.5 mL of a 2*M* stock solution of semicarbazide hydrochloride (or 0.5 mL of a solution prepared by dissolving 1.11 g of semicarbazide hydrochloride [MW = 111.5] in 5 mL of water) in a small test tube. Add an estimated 1 millimole (mmol) of the unknown compound to the test tube. If the unknown does not dissolve in the solution, or if the solution becomes cloudy, add enough methanol (maximum of 10 mL) to dissolve the solid and produce a clear solution. If a solid or cloudiness remains after adding 10 mL of methanol, do not add any more methanol and continue this procedure with the solid present. Using a Pasteur pipet, add 10 drops of pyridine and heat the mixture in a hot-water bath (about 60°C) for about 10–15 minutes. By that time, the product should have begun to crystallize. Collect the product by vacuum filtration. The product can be recrystallized from ethanol if necessary.

Semicarbazones (Alternative Method). Dissolve 0.25 g of semicarbazide hydrochloride and 0.38 g of sodium acetate in 1.3 mL of water. Then dissolve 0.25 g of the unknown in 2.5 mL of ethanol. Mix the two solutions together in a 25-mL Erlenmeyer flask and heat the mixture to boiling for about 5 minutes. After heating, place the reaction flask in a beaker of ice and scratch the sides of the flask with a glass rod to induce crystallization of the derivative. Collect the derivative by vacuum filtration and recrystallize it from ethanol.

2,4-Dinitrophenylhydrazones. Place 10 mL of a solution of 2,4-dinitrophenylhydrazine (prepared as described for the classification test in Experiment 57D) in a test tube and add an estimated 1 mmol of the unknown compound. If the unknown is a solid, it should be dissolved in the minimum amount of 95% ethanol or 1,2-dimethoxyethane before it is added. If crystallization is not immediate, gently warm the solution for a minute in a hot-water bath (90°C) and then set it aside to crystallize. Collect the product by vacuum filtration.

CARBOXYLIC ACIDS

Working in a hood, place 0.5 g of the acid and 2 mL of thionyl chloride into a small round-bottom flask. Add a magnetic stir bar and attach a condenser and a drying tube packed with calcium chloride. While stirring, heat the reaction mixture almost to boiling with a heating mantle with a magnetic stirrer or a hotplate/stirrer. *If the mixture does not turn color,* heat the solution for 30 minutes. Allow the mixture to cool to room temperature. *If the mixture begins to turn color,* remove the heat source and stir the mixture for 45 minutes at room temperature. Use the mixture for one of the following three procedures.

Amides. Working in a hood, pour the reaction mixture into a beaker containing 10 mL of ice cold concentrated ammonium hydroxide and stir it vigorously. When the reaction is complete, collect the product by vacuum filtration and recrystallize it from water or from water–ethanol, using the mixed-solvents method (Technique 5, Section 5.9).

Anilides. Dissolve 1 g of aniline in 25 mL of methylene chloride in a 125-mL Erlenmeyer flask. Using a Pasteur pipet, carefully add the reaction mixture to this solution. Warm the mixture for an additional 5 minutes on a hotplate, unless a significant color change occurs. *If a color change occurs,* discontinue heating, add a magnetic stirring bar, and stir the mixture for 20 minutes at room temperature. Then transfer the methylene chloride solution to a small separatory funnel and wash it sequentially with 5 mL of water, 5 mL of 5% hydrochloric acid, 5 mL of 5% sodium hydroxide, and a second 5-mL portion of water (the methylene chloride solution should be the bottom layer). Dry the methylene chloride layer over a small amount of anhydrous sodium sulfate. Decant the methylene chloride layer away from the drying agent into a small flask and evaporate the methylene chloride on a warm hotplate in the hood. Use a stream of air or nitrogen to speed up the evaporation. Recrystallize the product from water or from ethanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

***p*-Toluidides.** Use the same procedure as that described for the anilide but substitute *p*-toluidine for aniline.

PHENOLS

α-Naphthylurethanes. Follow the procedure given later for preparing phenylurethanes from alcohols but substitute α-naphthylisocyanate for phenylisocyanate.

Bromo Derivatives. First, if a stock brominating solution is not available, prepare one by dissolving 0.75 g of potassium bromide in 5 mL of water and adding 0.5 g of bromine. Dissolve 0.1 g of the phenol in 1 mL of methanol or 1,2-dimethoxyethane; then add 1 mL of water. Add 1 mL of the brominating mixture to the phenol solution and swirl the mixture vigorously. Then, continue adding the brominating solution dropwise while swirling, until the color of the bromine reagent persists. Finally, add 3–5 mL of water and shake the mixture vigorously. Collect the precipitated product by vacuum filtration and wash it well with water. Recrystallize the derivative from methanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

AMINES

Acetamides. Place an estimated 1 mmol of the amine and 0.5 mL of acetic anhydride in a small Erlenmeyer flask. Heat the mixture for about 5 minutes; then add 5 mL of water and stir the solution vigorously to precipitate the product and hydrolyze the excess acetic anhydride. If the product does not crystallize, it may be necessary to scratch the walls of the flask with a glass rod. Collect the crystals by vacuum filtration and wash them with several portions of cold 5% hydrochloric acid. Recrystallize the derivative from methanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

Aromatic amines, or those amines that are not very basic, may require pyridine (2 mL) as a solvent and a catalyst for the reaction. If pyridine is used, a longer period of heating is required (up to 1 hour), and the reaction should be carried out in an apparatus equipped with a reflux condenser. After reflux, the reaction mixture must be extracted with 5–10 mL of 5% sulfuric acid to remove the pyridine.

Benzamides. Using a centrifuge tube suspend an estimated 1 mmol of the amine in 1 mL of 10% sodium hydroxide solution and add 0.5 g of benzoyl chloride. Cap the tube and shake the mixture vigorously for about 10 minutes. After shaking, add enough dilute hydrochloric acid to bring the pH of the solution to pH 7 or 8. Collect the precipitate by vacuum filtration, wash it thoroughly with cold water, and recrystallize it from ethanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

Benzamides (Alternative Method). Dissolve 0.5 g of the amine in a solution of 2.5 mL of pyridine and 5 mL of toluene. Add 0.5 mL of benzoyl chloride to the solution and heat the mixture under reflux about 30 minutes. Pour the cooled reaction mixture into 50 mL of water and stir the mixture vigorously to hydrolyze the excess benzoyl chloride. Separate the toluene layer and wash it, first with 3 mL of water and then with 3 mL of 5% sodium carbonate. Dry the toluene over anhydrous sodium sulfate, decant the toluene into a small flask, and remove the toluene by evaporation on a hotplate in the hood. Use a stream of air or nitrogen to speed up the evaporation. Recrystallize the benzamide from ethanol or ethanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

Picrates. Dissolve 0.2 g of the unknown in about 5 mL of ethanol and add 5 mL of a saturated solution of picric acid in ethanol. Heat the solution to boiling and then allow it to cool slowly. Collect the product by vacuum filtration and rinse it with a small amount of cold ethanol.

Methiodides. Mix equal-volume quantities of the amine and methyl iodide in a test tube (about 0.25 mL is sufficient) and allow the mixture to stand for several minutes. Then heat the mixture gently under reflux for about 5 minutes. The methiodide should crystallize on cooling. If it does not, you can induce crystallization by scratching the walls of the test tube with a glass rod. Collect the product by vacuum filtration and recrystallize it from ethanol or ethyl acetate.

ALCOHOLS

3,5-Dinitrobenzoates. *Liquid Alcohols.* Dissolve 0.25 g of 3,5-dinitrobenzoyl chloride[1] in 0.25 mL of the alcohol and heat the mixture for about 5 minutes. Allow the mixture to cool and add 1.5 mL of a 5% sodium carbonate solution and 1 mL of water. Stir the mixture vigorously and crush any solid that forms. Collect the product by vacuum filtration and wash it with cold water. Recrystallize the derivative from ethanol–water, using the mixed-solvent method (Technique 5, Section 5.9).

Solid Alcohols. Dissolve 0.25 g of the alcohol in 1.5 mL of dry pyridine and add 0.25 g of 3,5-dinitrobenzoyl chloride.[2] Heat the mixture under reflux for 15 minutes. Pour the cooled reaction mixture into a cold mixture of 2.5 mL of 5% sodium carbonate and 2.5 mL of water. Keep the solution cooled in an ice bath until the product crystallizes and stir it vigorously during the entire period. Collect the product by vacuum filtration, wash it with cold water, and recrystallize it from ethanol–water, using the mixed-solvents method (Technique 5, Section 5.9).

Phenylurethanes. Place 0.25 g of the *anhydrous* alcohol in a dry test tube and add 0.25 mL of phenylisocyanate (α-naphthylisocyanate for a phenol). If the compound is a phenol, add one drop of pyridine to catalyze the reaction. If the reaction is not spontaneous, heat the mixture in a hot-water bath (90°C) for 5–10 minutes. Cool the test tube in a beaker of ice and scratch the tube with a glass rod to induce crystallization. Decant the liquid from the solid product or, if necessary, collect the product by vacuum filtration. Dissolve the product in 2.5–3 mL of hot ligroin or hexane and filter the mixture by gravity (preheat funnel) to remove any unwanted and insoluble diphenylurea present. Cool the filtrate to induce crystallization of the urethane. Collect the product by vacuum filtration.

ESTERS

We recommend that esters be characterized by spectroscopic methods whenever possible. A derivative of the alcohol part of an ester can be prepared with the procedure described below. For other derivatives, consult a comprehensive textbook. Several are listed in Experiment 57 (p. 485).

---

[1] This is an acid chloride and undergoes hydrolysis readily. The purity of this reagent should be checked before its use by determining its melting point. When the carboxylic acid is present, the melting point will be high.

[2] See Footnote 1.

3,5-Dinitrobenzoates. Place 2.0 mL of the ester and 1.5 g of 3,5-dinitrobenzoic acid in a small round-bottom flask. Add 4 drops of concentrated sulfuric acid and a boiling stone to the flask and attach a condenser. If the boiling point of the ester is below 150°C, heat under reflux while stirring for 30–45 minutes. If the boiling point of the ester is above 150°C, heat the mixture at about 150°C for 30–45 minutes. Cool the mixture and transfer it to a small separatory funnel. Add 20 mL of ether. Extract the ether layer two times with 10 mL of 5% aqueous sodium carbonate (save the ether layer). Wash the organic layer with 10 mL of water and dry the ether solution over magnesium sulfate. Evaporate the ether in a hot-water bath in the hood. Use a stream of air or nitrogen to speed the evaporation. Dissolve the residue, usually an oil, in 4 mL of boiling ethanol and add water dropwise until the mixture becomes cloudy. Cool the solution to induce crystallization of the derivative.

APPENDIX 3

Infrared Spectroscopy

Almost any compound having covalent bonds, whether organic or inorganic, will be found to absorb frequencies of electromagnetic radiation in the infrared region of the spectrum. The infrared region of the electromagnetic spectrum lies at wavelengths longer than those associated with visible light, which includes wavelengths from approximately 400 to 800 nm (1 nm = 10^{-9} m), but at wavelengths shorter than those associated with radio waves, which have wavelengths longer than 1 cm. For chemical purposes, we are interested in the *vibrational* portion of the infrared region. This portion is defined as that which includes radiations with wavelengths (λ) between 2.5 and 15 μm (1 μm = 10^{-6} m). The relation of the infrared region to others included in the electromagnetic spectrum is illustrated in Figure IR.1.

As with other types of energy absorption, molecules are excited to a higher energy state when they absorb infrared radiation. The absorption of infrared radiation is, like other

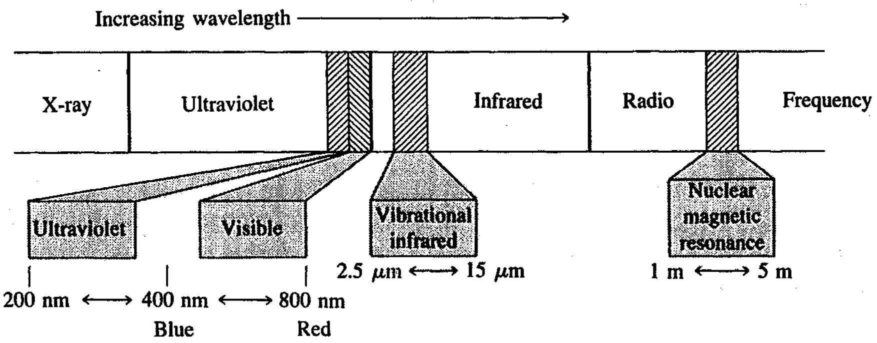

Figure IR.1 Portion of electromagnetic spectrum, showing relation of vibrational infrared to other types of radiation.

absorption processes, a quantized process. Only selected frequencies (energies) of infrared radiation are absorbed by a molecule. The absorption of infrared radiation corresponds to energy changes on the order of 8–40 kJ/mole (2–10 kcal/mole). Radiation in this energy range corresponds to the range encompassing the stretching and bending vibrational frequencies of the bonds in most covalent molecules. In the absorption process, those frequencies of infrared radiation that match the natural vibrational frequencies of the molecule in question are absorbed, and the energy absorbed increases the *amplitude* of the vibrational motions of the bonds in the molecule.

Most chemists refer to the radiation in the vibrational infrared region of the electromagnetic spectrum by units called **wavenumbers** (*ν*). Wavenumbers are expressed in reciprocal centimeters (cm^{-1}) and are easily computed by taking the reciprocal of the wavelength (λ) expressed in centimeters. This unit has the advantage, for those performing calculations, of being directly proportional to energy. Thus, the vibrational infrared region of the spectrum extends from about 4000 to 650 cm^{-1} (or wavenumbers).

Wavelengths (μm) and wavenumbers (cm^{-1}) can be interconverted by the following relationships:

$$\text{cm}^{-1} = \frac{1}{(\mu\text{m})} \times 10{,}000$$

$$\mu\text{m} = \frac{1}{(\text{cm}^{-1})} \times 10{,}000$$

IR.1 USES OF THE INFRARED SPECTRUM

Because every type of bond has a different natural frequency of vibration, and because the same type of bond in two different compounds is in a slightly different environment, no two molecules of different structure have exactly the same infrared absorption pattern, or **infrared spectrum.** Although some of the frequencies absorbed in the two cases might be the same, in no case of two different molecules will their infrared spectra (the patterns of absorption) be identical. Thus, the infrared spectrum can be used for molecules much as a fingerprint can be used for people. By comparing the infrared spectra of two substances thought to be identical, you can establish whether or not they are in fact identical. If the infrared spectra of two substances coincide peak for peak (absorption for absorption), in most cases, the substances are identical.

A second and more important use of the infrared spectrum is that it gives structural information about a molecule. The absorptions of each type of bond (N—H, C—H, O—H, C—X, C═O, C—O, C—C, C═C, C≡C, C≡N, and so on) are regularly found only in certain small portions of the vibrational infrared region. A small range of absorption can be defined for each type of bond. Outside this range, absorptions will normally be due to some other type of bond. Thus, for instance, any absorption in the range 3000 ± 150 cm^{-1} will almost always be due to the presence of a CH bond in the molecule; an absorption in the range 1700 ± 100 cm^{-1} will normally be due to the presence of a C═O bond (carbonyl group) in the molecule. The same type of range applies to each type of bond. The

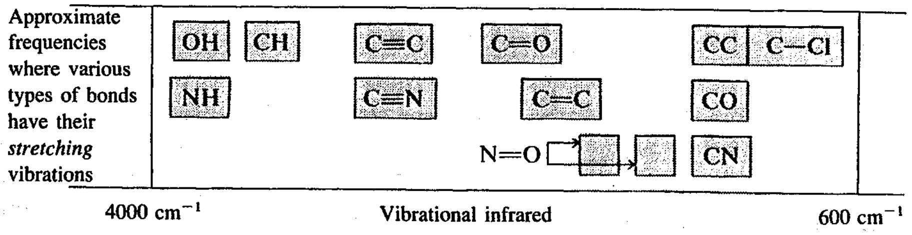
Approximate frequencies where various types of bonds have their *stretching* vibrations
OH
CH
C≡C
C=O
CC
C—Cl
NH
C≡N
C=C
CO
N=O
CN
4000 cm⁻¹
Vibrational infrared
600 cm⁻¹

Figure IR.2 Approximate regions in which various common types of bonds absorb. (Bending and twisting and other types of bond vibration have been omitted for clarity).

way these are spread out over the vibrational infrared is illustrated schematically in Figure IR.2. It is a good idea to remember this general scheme for future convenience.

IR.2 MODES OF VIBRATION

The simplest types, or **modes,** of vibrational motion in a molecule that are **infrared-active,** that is, give rise to absorptions, are the stretching and bending modes.

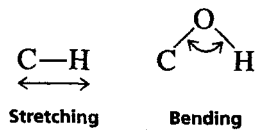

Other, more complex types of stretching and bending are also active, however. To introduce several words of terminology, the normal modes of vibration for a methylene group are shown on page 895.

In any group of three or more atoms—at least two of which are identical—there are *two* modes of stretching or bending: the symmetric mode and the asymmetric mode. Examples of such groups are $—CH_3$, $—CH_2—$, $—NO_2$, $—NH_2$, and anhydrides $(CO)_2O$. For the anhydride, owing to asymmetric and symmetric modes of stretch, this functional group gives *two* absorptions in the C=O region. A similar phenomenon is seen for amino groups, where primary amines usually have *two* absorptions in the NH stretch region, whereas secondary amines R_2NH have only one absorption peak. Amides show similar bands. There are two strong N=O stretch peaks for a nitro group, which are caused by asymmetric and symmetric stretching modes.

IR.3 WHAT TO LOOK FOR IN EXAMINING INFRARED SPECTRA

The instrument that determines the absorption spectrum for a compound is called an **infrared spectrophotometer.** The spectrophotometer determines the relative strengths and positions of all the absorptions in the infrared region and plots this information on a piece of calibrated chart paper. This plot of absorption intensity versus wavenumber or wavelength is referred to as the **infrared spectrum** of the compound. A typical infrared spectrum, that of methyl isopropyl ketone, is shown in Figure IR.3.

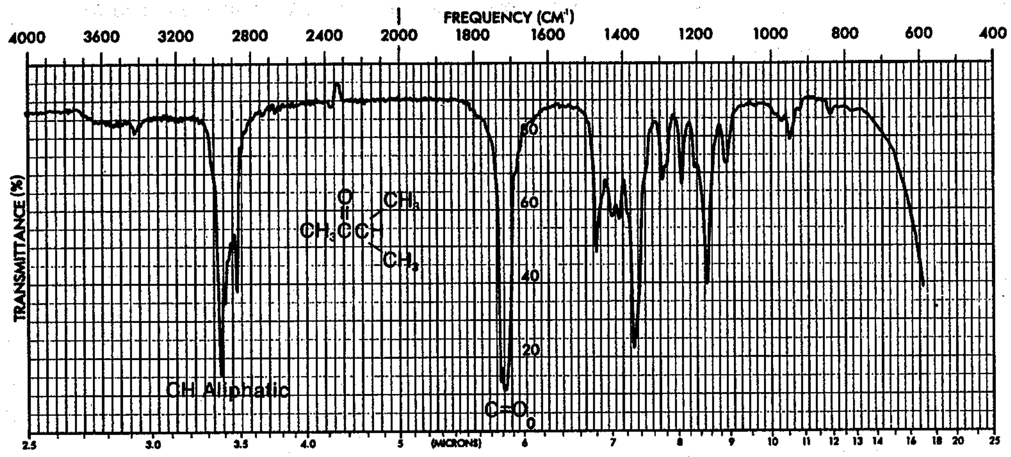

Figure IR.3 Infrared spectrum of methyl isopropyl ketone (neat liquid, salt plates.)

The strong absorption in the middle of the spectrum corresponds to C═O, the carbonyl group. Note that the C═O peak is quite intense. In addition to the characteristic position of absorption, the **shape** and **intensity** of this peak are also unique to the C═O bond. This is true for almost every type of absorption peak; both shape and intensity characteristics can be described, and these characteristics often make it possible to distinguish the peak in a confusing situation. For instance, to some extent both C═O and C═C bonds absorb in the same region of the infrared spectrum:

$$\begin{array}{ll} C{=}O & 1850\text{–}1630\ \text{cm}^{-1} \\ C{=}C & 1680\text{–}1620\ \text{cm}^{-1} \end{array}$$

However, the C═O bond is a strong absorber, whereas the C═C bond generally absorbs only weakly. Hence, a trained observer would not normally interpret a strong peak at 1670

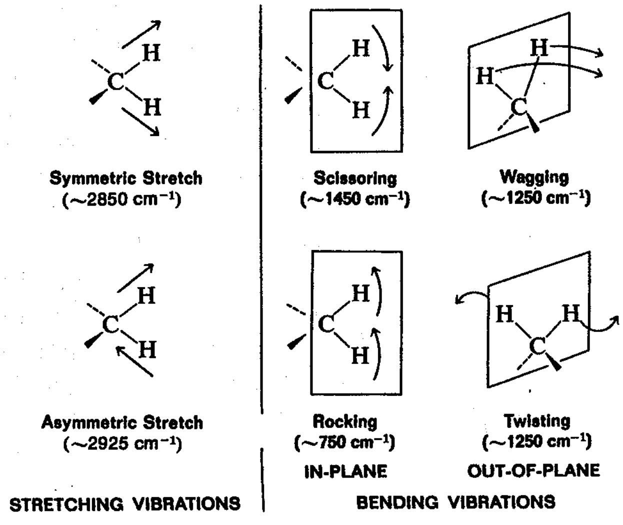
H
C
H
Symmetric Stretch
(~2850 cm⁻¹)
Scissoring
(~1450 cm⁻¹)
Wagging
(~1250 cm⁻¹)
Asymmetric Stretch
(~2925 cm⁻¹)
Rocking
(~750 cm⁻¹)
Twisting
(~1250 cm⁻¹)
IN-PLANE
OUT-OF-PLANE
STRETCHING VIBRATIONS
BENDING VIBRATIONS

cm^{-1} to be a carbon–carbon double bond nor a weak absorption at this frequency to be due to a carbonyl group.

The shape of a peak often gives a clue to its identity as well. Thus, although the NH and OH regions of the infrared overlap,

OH 3650–3200 cm^{-1}
NH 3500–3300 cm^{-1}

NH usually gives a **sharp** absorption peak (absorbs a very narrow range of frequencies), and OH, when it is in the NH region, usually gives a **broad** absorption peak. Primary amines give *two* absorptions in this region, whereas alcohols give only one.

Therefore, while you are studying the sample spectra in the pages that follow, you should also notice shapes and intensities. They are as important as the frequency at which an absorption occurs, and you must train your eye to recognize these features. Often, in the literature of organic chemistry, you will find absorptions referred to as strong (s), medium (m), weak (w), broad, or sharp. The author is trying to convey some idea of what the peak looks like without actually drawing the spectrum. Although the intensity of an absorption often provides useful information about the identity of a peak, you should be aware that the relative intensities of all the peaks in the spectrum are dependent on the amount of sample that is used and the sensitivity setting of the instrument. Therefore, the *actual* intensity of a particular peak may vary from spectrum to spectrum, and you must pay attention to *relative* intensities.

IR.4 CORRELATION CHARTS AND TABLES

To extract structural information from infrared spectra, you must know the frequencies or wavelengths at which various functional groups absorb. Infrared **correlation tables** present as much information as is known about where the various functional groups absorb. The books listed at the end of this appendix present extensive lists of correlation tables (see page 911). Sometimes the absorption information is given in a chart, called a **correlation chart.** A simplified correlation table is given in Table IR.1.

Although you may think assimilating the mass of data in Table IR.1 will be difficult, it is not if you make a modest start and then gradually increase your familiarity with the data. An ability to interpret the fine details of an infrared spectrum will follow. This is most easily accomplished by first establishing the broad visual patterns of Figure IR.2 firmly in mind. Then, as a second step, a "typical absorption value" can be memorized for each of the functional groups in this pattern. This value will be a single number that can be used as a pivot value for the memory. For instance, start with a simple aliphatic ketone as a model for all typical carbonyl compounds. The typical aliphatic ketone has a carbonyl absorption of 1715 ± 10 cm^{-1}. Without worrying about the variation, memorize 1715 cm^{-1} as the base value for carbonyl absorption. Then, learn the extent of the carbonyl range and the visual pattern of how the different kinds of carbonyl groups are arranged throughout this region. See, for instance, Figure IR.15, which gives typical val-

TABLE IR.1 A Simplified Correlation Table

	Type of Vibration	Frequency (cm^{-1})	Intensity
C—H	Alkanes (stretch)	3000–2850	s
	—CH_3 (bend)	1450 and 1375	m
	—CH_2— (bend)	1465	m
	Alkenes (stretch)	3100–3000	m
	(bend)	1700–1000	s
	Aromatics (stretch)	3150–3050	s
	(out-of-plane bend)	1000–700	s
	Alkyne (stretch)	ca. 3300	s
	Aldehyde	2900–2800	w
		2800–2700	w
C—C	Alkane Not interpretatively useful		
C=C	Alkene	1680–1600	m–w NO
	Aromatic	1600–1400	m–w yess
C≡C	Alkyne	2250–2100	m–w NO
C=O	Aldehyde	1740–1720	s yes
	Ketone (acyclic)	1725–1705	s
	Carboxylic acid	1725–1700	s
	Ester	1750–1730	s
	Amide	1700–1640	s
	Anhydride	ca. 1810	s
		ca. 1760	s
C—O	Alcohols, ethers, esters, carboxylic acids	1300–1000	s 2 wek 1s
O—H	Alcohol, phenols		
	Free	3650–3600	m
	H-Bonded	3400–3200	m
	Carboxylic acids	3300–2500	m
N—H	Primary and secondary amines	ca. 3500	m
C≡N	Nitriles	2260–2240	m
N=O	Nitro (R—NO_2)	1600–1500	s
		1400–1300	s
C—X	Fluoride	1400–1000	s
	Chloride	800–600	s
	Bromide, iodide	< 600	s

NOTE: s, strong; m, medium; w, weak.

TABLE IR.2 Base Values for Absorptions of Bonds

OH	3600 cm^{-1}	C≡C	2150 cm^{-1}
NH	3500 cm^{-1}	C=O	1715 cm^{-1}
CH	3000 cm^{-1}	C=C	1650 cm^{-1}
C≡N	2250 cm^{-1}	C—O	1100 cm^{-1}

ues for carbonyl compounds. Also learn how factors like ring size (when the functional group is contained in a ring) and conjugation affect the base values (that is, in which direction the values are shifted). Learn the trends—always remembering the base value (1715 cm^{-1}). It might prove useful as a beginning to memorize the base values in Table IR.2 for this approach. Notice that there are only eight values.

IR.5 ANALYZING A SPECTRUM (OR WHAT YOU CAN TELL AT A GLANCE)

In trying to analyze the spectrum of an unknown, you should concentrate first on trying to establish the presence (or absence) of a few major functional groups. The most conspicuous peaks are C=O, C—H, N—H, C—O, C=C, C≡C, C≡N, and NO_2. If they are present, they give immediate structural information. Do not try to analyze in detail the CH absorptions near 3000 cm^{-1}; almost all compounds have these absorptions. Do not worry about subtleties of the exact type of environment in which the functional group is found. A checklist of the important gross features follows:

1. Is a carbonyl group present?
The C=O group gives rise to a strong absorption in the region 1820–1600 cm^{-1}. The peak is often the strongest in the spectrum and of medium width. You can't miss it.

2. If C=O is present, check the following types. (If it is absent, go to 3.)

ACIDS	Is OH also present? **Broad** absorption near 3300–2500 cm^{-1} (usually overlaps C—H).
AMIDES	Is NH also present? Medium absorption near 3500 cm^{-1}, sometimes a double peak, equivalent halves.
ESTERS	Is C—O also present? Medium intensity absorptions near 1300–1000 cm^{-1}.
ANHYDRIDES	Have *two* C=O absorptions near 1810 and 1760 cm^{-1}.

ALDEHYDES	Is aldehyde CH present? Two weak absorptions near 2850 and 2750 cm^{-1} on the right side of CH absorptions.
KETONES	The above five choices have been eliminated.
3. If $C{=}O$ is absent	
ALCOHOLS or PHENOLS	Check for OH. **Broad** absorption near 3600–3300 cm^{-1}. Confirm this by finding $C{-}O$ near 1300–1000 cm^{-1}.
AMINES	Check for NH. Medium absorption(s) near 3500 cm^{-1}.
ETHERS	Check for $C{-}O$ (and absence of OH) near 1300–1000 cm^{-1}.
4. Double Bonds or Aromatic Rings or Both	$C{=}C$ is a **weak** absorption near 1650 cm^{-1}. Medium to strong absorptions in the region 1650–1450 cm^{-1} often imply an aromatic ring. Confirm the above by consulting the CH region. Aromatic and vinyl CH occur to the left of 3000 cm^{-1} (aliphatic CH occurs to the right of this value).
5. Triple Bonds	$C{\equiv}N$ is a medium, sharp absorption near 2250 cm^{-1}. $C{\equiv}C$ is a weak but sharp absorption near 2150 cm^{-1}. Check also for acetylenic CH near 3300 cm^{-1}.
6. Nitro Groups	*Two* strong absorptions 1600–1500 cm^{-1} and 1390–1300 cm^{-1}.
7. Hydrocarbons	None of the above is found. Main absorptions are in CH region near 3000 cm^{-1}. Very simple spectrum, only other absorptions near 1450 cm^{-1} and 1375 cm^{-1}.

The beginning student should resist the idea of trying to assign or interpret *every* peak in the spectrum. You simply will not be able to do this. Concentrate first on learning the principal peaks and recognizing their presence or absence. This is best done by carefully studying the illustrative spectra in the section that follows.

Note: In describing the shifts of absorption peaks or their relative positions, we have used the phrases "to the left" and "to the right." This was done to simplify descriptions of peak positions. The meaning is clear, because all spectra are conventionally presented left to right from 4000 to 600 cm^{-1}.

IR.6 SURVEY OF THE IMPORTANT FUNCTIONAL GROUPS

Alkanes

Spectrum is usually simple with few peaks.

C—H Stretch occurs around 3000 cm^{-1}.

(a) In alkanes (except strained ring compounds), absorption always occurs to the right of 3000 cm^{-1}.

(b) If a compound has vinylic, aromatic, acetylenic, or cyclopropyl hydrogens, the CH absorption is to the left of 3000 cm^{-1}.

CH_2 Methylene groups have a characteristic absorption at approximately 1450 cm^{-1}.

CH_3 Methyl groups have a characteristic absorption at approximately 1375 cm^{-1}.

C—C Stretch—not interpretatively useful—has many peaks.

The spectrum of decane is shown in Figure IR.4.

Alkenes

=C—H Stretch occurs to the left of 3000 cm^{-1}.

=C—H Out-of-plane ("oop") bending at 1000–650 cm^{-1}.

The C—H out-of-plane absorptions often allow you to determine the type of substitution pattern on the double bond, according to the number of absorptions and their positions. The correlation chart in Figure IR.5 shows the positions of these bands.

C═C Stretch 1675–1600 cm^{-1}, often weak.

Conjugation moves C═C stretch to the right.

Symmetrically substituted bonds, as in 2,3-dimethyl-2-butene, do not absorb in the infrared region (no dipole change). Highly substituted double bonds are often vanishingly weak in absorption.

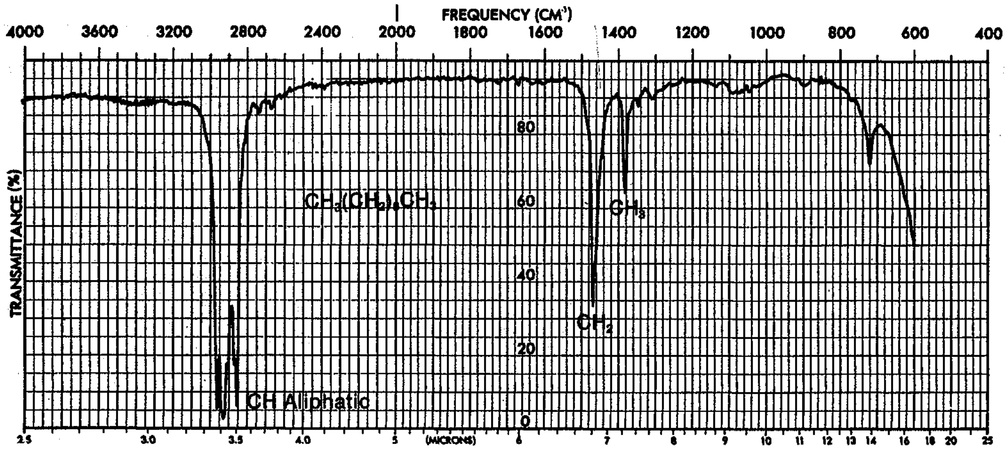

Figure IR.4 Infrared spectrum of decane (neat liquid, salt plates).

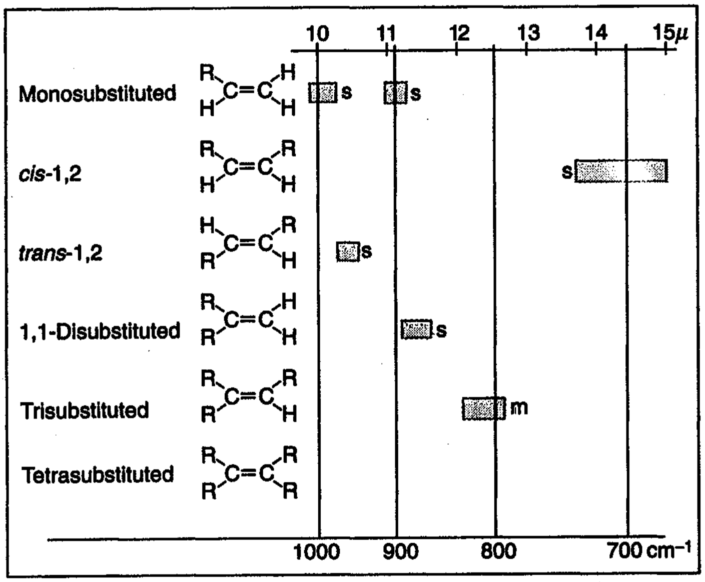

Figure IR.5 The C—H out-of-plane bending vibrations for substituted alkenes.

The spectrum of styrene is shown in Figure IR.6. The spectrum of 4-methylcyclohexene in shown in Experiment 27.

Aromatic Rings

=C—H Stretch is always to the left of 3000 cm^{-1}.

=C—H Out-of-plane (oop) bending at 900 to 690 cm^{-1}.

The CH out-of-plane absorptions often allow you to determine the type of ring substitution by their numbers, intensities, and positions. The correlation chart in Figure IR.7A indicates the positions of these bands.

The patterns are generally reliable—most particularly reliable for rings with alkyl substituents, least for polar substituents.

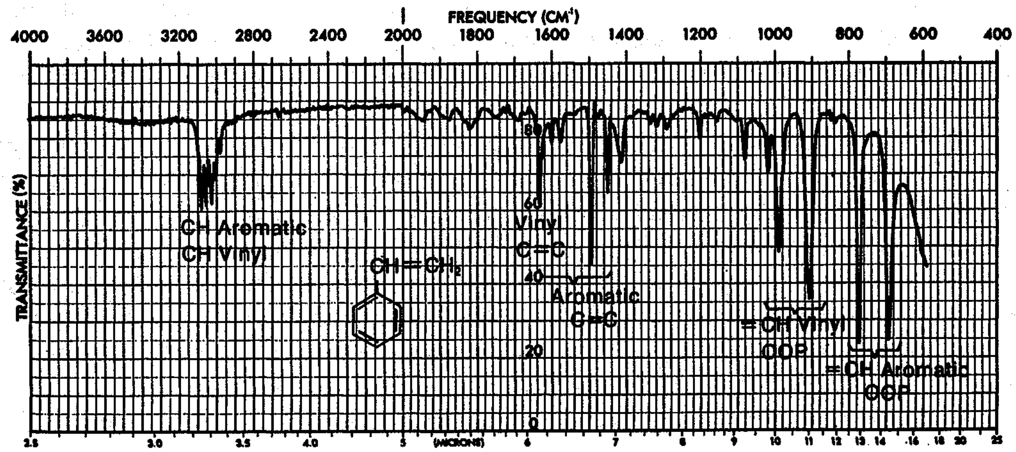

Figure IR.6 Infrared spectrum of styrene (neat liquid, salt plates).

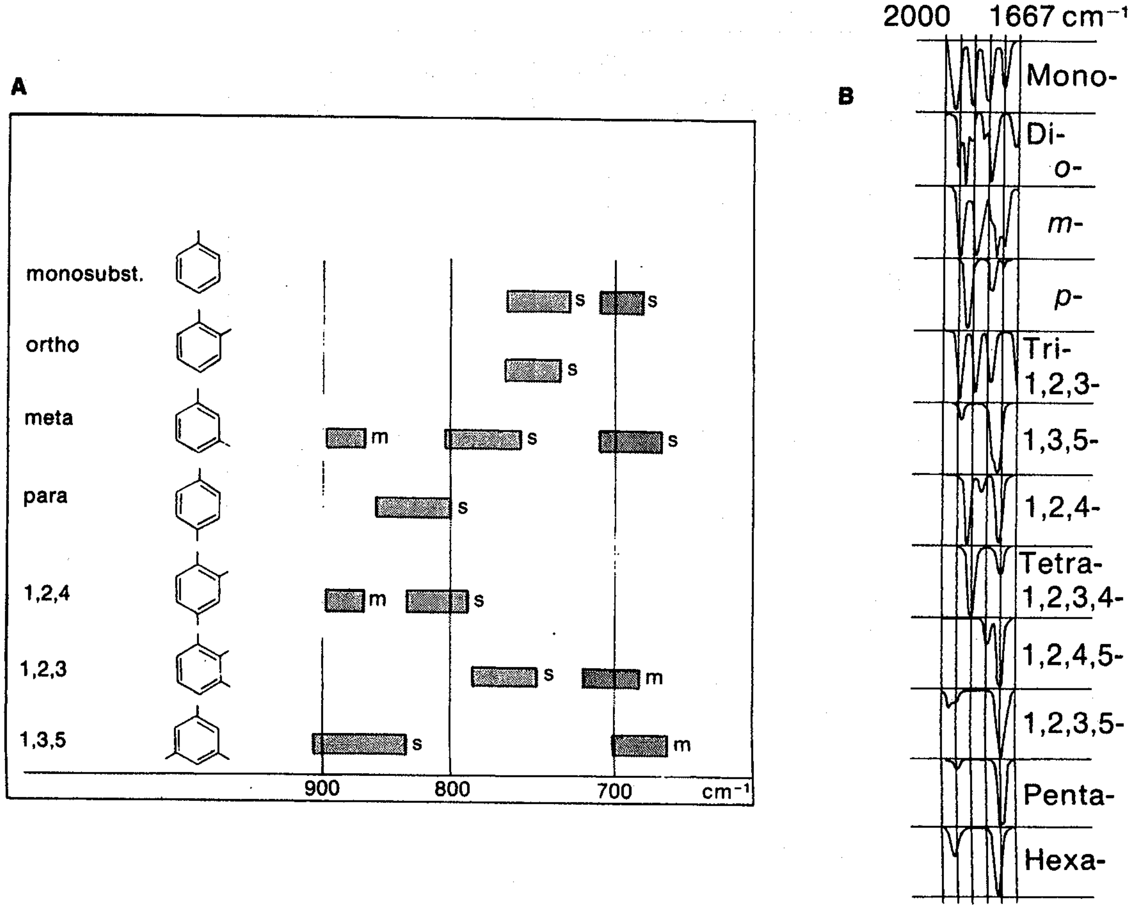

Figure IR.7 The C—H out-of-plane bending vibrations for substituted benzenoid compounds. (B) The 2000–1667 cm^{-1} region for substituted benzenoid compounds. (From John R. Dyer, *Applications of Absorption Spectroscopy of Organic Compounds.* Englewood Cliffs, NJ: Prentice-Hall, 1965.)

Ring Absorptions (C═C). There are often four sharp absorptions that occur in pairs at 1600 and 1450 cm^{-1} and are characteristic of an aromatic ring. See, for example, the spectra of anisole (Fig. IR.11), benzonitrile (Fig. IR.14), and methyl benzoate (Fig. IR.18).

There are many weak combination and overtone absorptions that appear between 2000 and 1667 cm^{-1}. The relative shapes and numbers of these peaks can be used to tell whether an aromatic ring is monosubstituted or di-, tri-, tetra-, penta-, or hexasubstituted. Positional isomers can also be distinguished. Because the absorptions are weak, these bands are best observed by using neat liquids or concentrated solutions. If the compound has a high-frequency carbonyl group, this absorption overlaps the weak overtone bands, so that no useful information can be obtained from analyzing this region. The various patterns that are obtained in this region are shown in Figure IR.7B.

The spectra of styrene and *o*-dichlorobenzene are shown in Figures IR.6 and IR.8.

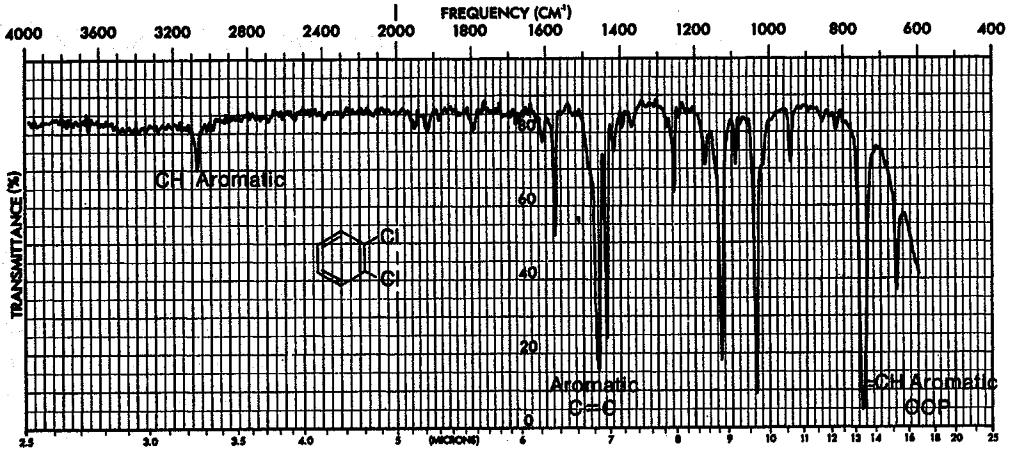

Figure IR.8 Infrared spectrum of *o*-dichlorobenzene (neat liquid, salt plates).

Alkynes

$\equiv$C—H Stretch is usually near 3300 cm^{-1}.

C$\equiv$C Stretch is near 2150 cm^{-1}.
Conjugation moves C$\equiv$C stretch to the right.
Disubstituted or symmetrically substituted triple bonds give either no absorption or weak absorption.

The spectrum of propargyl alcohol is shown in Figure IR.9.

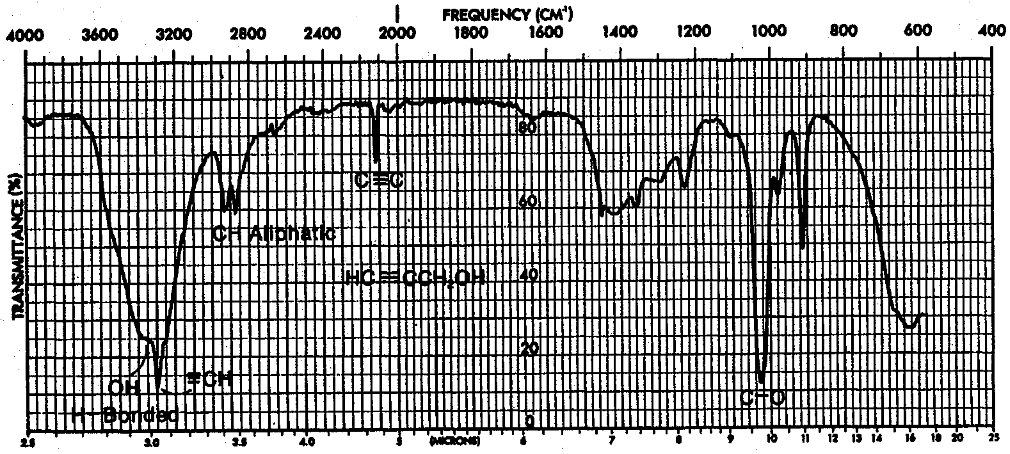

Figure IR.9 Infrared spectrum of propargyl alcohol (neat liquid, salt plates).

Alcohols and Phenols

O—H Stretch is a sharp peak at 3650–3600 cm^{-1} if no hydrogen bonding takes place. (This is usually observed only in dilute solutions.)

If there is hydrogen bonding (usually in neat or concentrated solutions), the absorption is **broad** and occurs more to the right at 3500–3200 cm^{-1}, sometimes overlapping C—H stretch absorptions.

C—O Stretch is usually in the range 1300–1000 cm^{-1}.

Phenols are like alcohols. The 2-naphthol shown in Figure IR.10 has some molecules hydrogen-bonded and some free. The spectrum of 4-methylcyclohexanol is given in Experiment 27. This alcohol, which was determined neat, would also have had a free OH spike to the left of its hydrogen-bonded band if it had been determined in dilute solution. The solution spectra of borneol and isoborneol are shown in Experiment 20.

Ethers

C—O The most prominent band is due to C—O stretch at 1300–1000 cm^{-1}. Absence of C=O and O—H bands is required to be sure C—O stretch is not due to an alcohol or ester. Phenyl and vinyl ethers are found in the left portion of the range, aliphatic ethers in the right. (Conjugation with the oxygen moves the absorption to the left.)

The spectrum of anisole is shown in Figure IR.11.

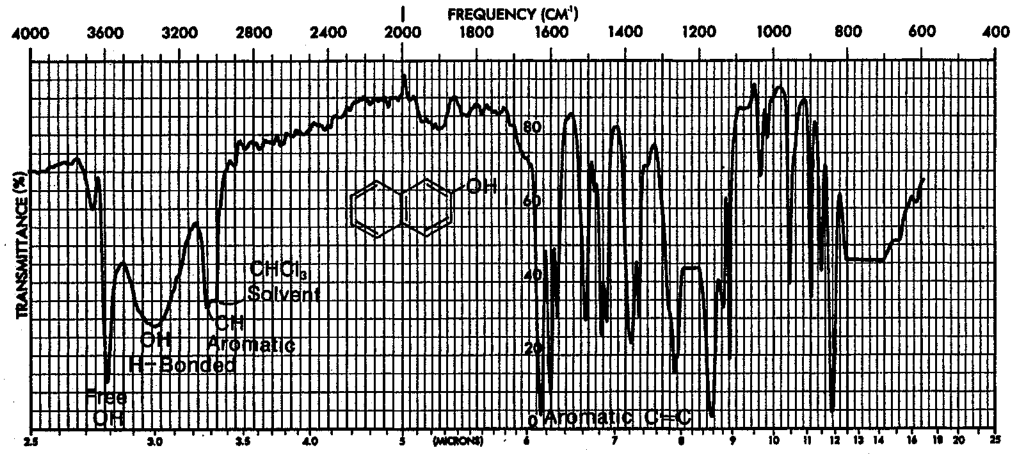

Figure IR.10 Infrared spectrum of 2-naphthol, showing both free and hydrogen-bonded OH ($CHCl_3$ solution).

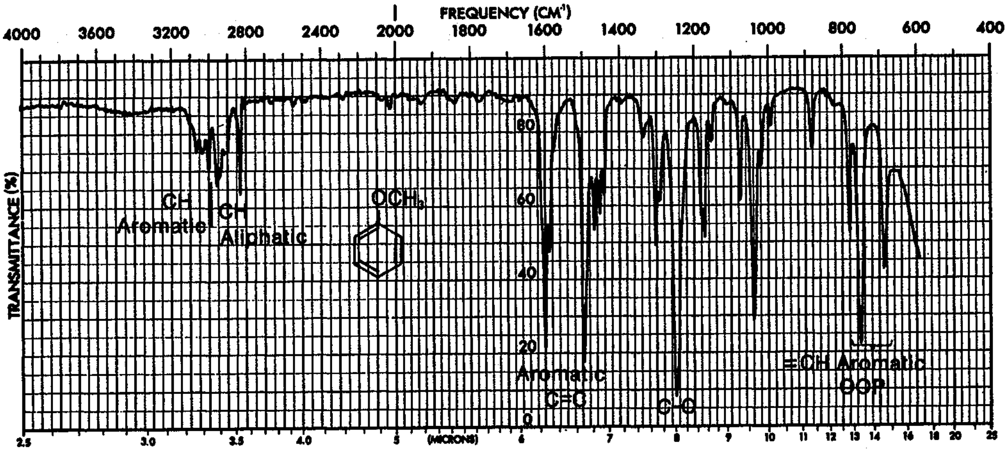

Figure IR.11 Infrared spectrum of anisole (neat liquid, salt plates).

Amines

N—H	Stretch occurs in the range of 3500–3300 cm^{-1}. Primary amines have *two* bands typically 30 cm^{-1} apart. Secondary amines have one band, often vanishingly weak. Tertiary amines have no NH stretch.
C—N	Stretch is weak and occurs in the range of 1350–1000 cm^{-1}.
N—H	Scissoring mode occurs in the range of 1640–1560 cm^{-1} (broad). An out-of-plane bending absorption can sometimes be observed at about 800 cm^{-1}.

The spectrum of *n*-butylamine is shown in Figure IR.12.

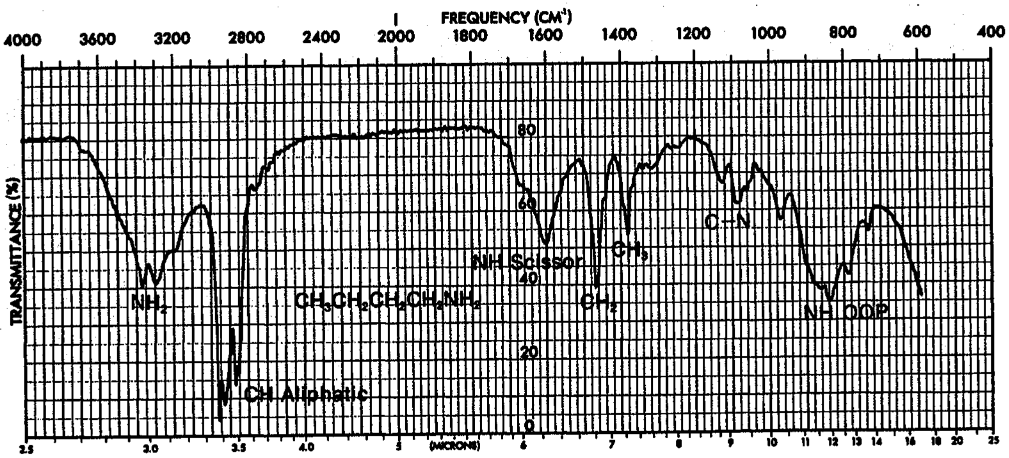

Figure IR.12 Infrared spectrum of *n*-butylamine (neat liquid, salt plates).

Nitro Compounds

N=O Stretch is usually two strong bonds at 1600–1500 cm^{-1} and 1390–1300 cm^{-1}.

The spectrum of nitrobenzene is shown in Figure IR.13.

Nitriles

C≡N Stretch is a sharp absorption near 2250 cm^{-1}.
Conjugation with double bonds or aromatic rings moves the absorption to the right.

The spectrum of benzonitrile is shown in Figure IR.14.

Carbonyl Compounds

The carbonyl group is one of the most strongly absorbing groups in the infrared region of the spectrum. This is mainly due to its large dipole moment. It absorbs in a variety of compounds (aldehydes, ketones, acids, esters, amides, anhydrides, and so on) in the range of 1850–1650 cm^{-1}. In Figure IR.15 the normal values for the various types of carbonyl groups are compared. In the sections that follow, each type is examined separately.

Aldehydes

C=O Stretch at approximately 1725 cm^{-1} is normal.
Aldehydes *seldom* absorb to the left of this value.
Conjugation moves the absorption to the right.

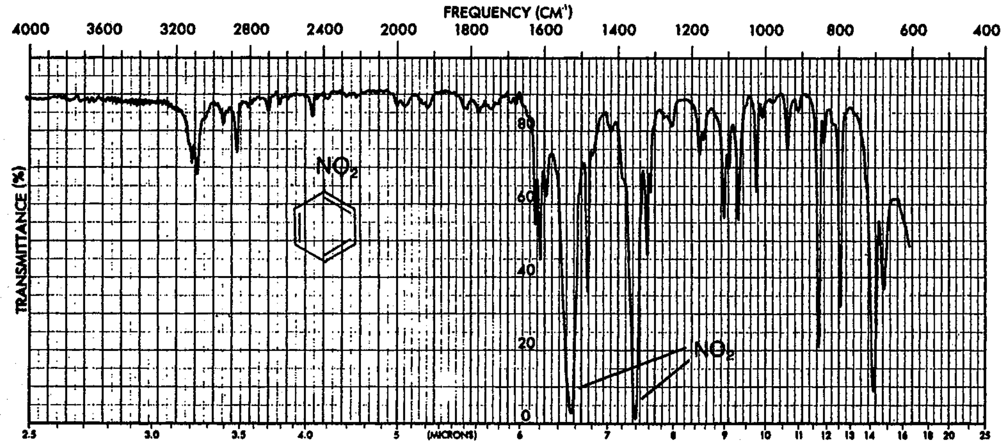

Figure IR.13 Infrared spectrum of nitrobenzene, neat.

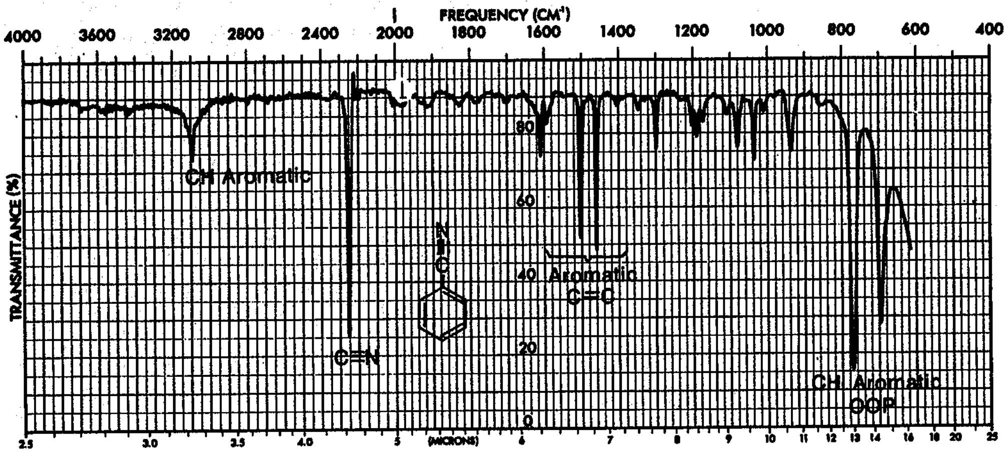

Figure IR.14 Infrared spectrum of benzonitrile (neat liquid, salt plates).

C—H Stretch, aldehyde hydrogen (—CHO), consists of *weak* bands at about 2750 and 2850 cm^{-1}. Note that the CH stretch in alkyl chains does not usually extend this far to the right.

The spectrum of nonanal is shown in Figure IR.16. In addition, the spectrum of benzaldehyde is shown in Experiment 40.

Ketones

C=O Stretch at approximately 1715 cm^{-1} is normal.
Conjugation moves the absorption to the right.
Ring strain moves the absorption to the left in cyclic ketones.

The spectra of methyl isopropyl ketone and mesityl oxide are shown in Figure IR.3 and IR.17. The spectrum of camphor is shown in Experiment 20.

1810	1760	1735	1725	1715	1710	1690	cm^{-1}
Anhydride (Band 1)	Anhydride (Band 2)	Esters	Aldehydes	Ketones	Carboxylic acids	Amides	

Figure IR.15 Normal values (±10 cm^{-1}) for various types of carbonyl groups.

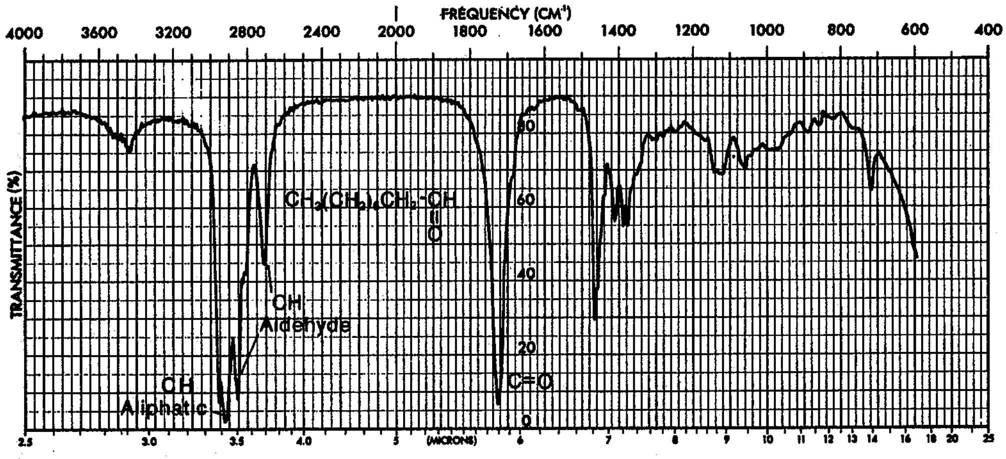

Figure IR.16 Infrared spectrum of nonanal (neat liquid, salt plates).

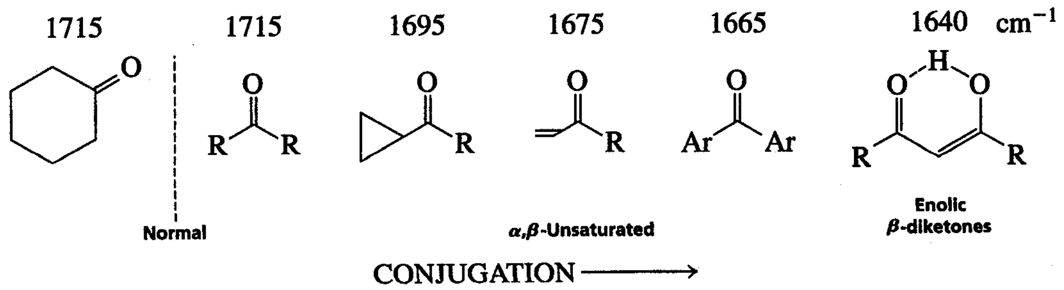
1715
1715
1695
1675
1665
1640 cm⁻¹
O
R
Ar
H
Normal
α,β-Unsaturated
Enolic
β-diketones
CONJUGATION ⟶

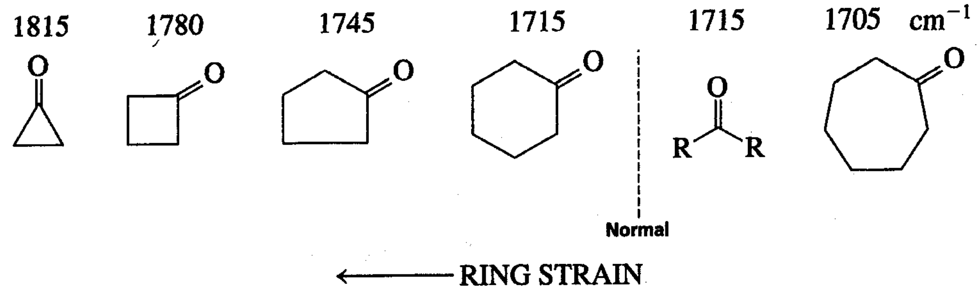
1815
1780
1745
1715
1715
1705 cm⁻¹
O
R
Normal
⟵ RING STRAIN

Acids

O—H Stretch, usually **very broad** (strongly hydrogen-bonded) at 3300–2500 cm^{-1}, often interferes with C—H absorptions.

C=O Stretch, broad, 1730–1700 cm^{-1}.
Conjugation moves the absorption to the right.

C—O Stretch, in range of 1320–1210 cm^{-1}, strong.

The spectrum of benzoic acid is shown in Experiment 30B.

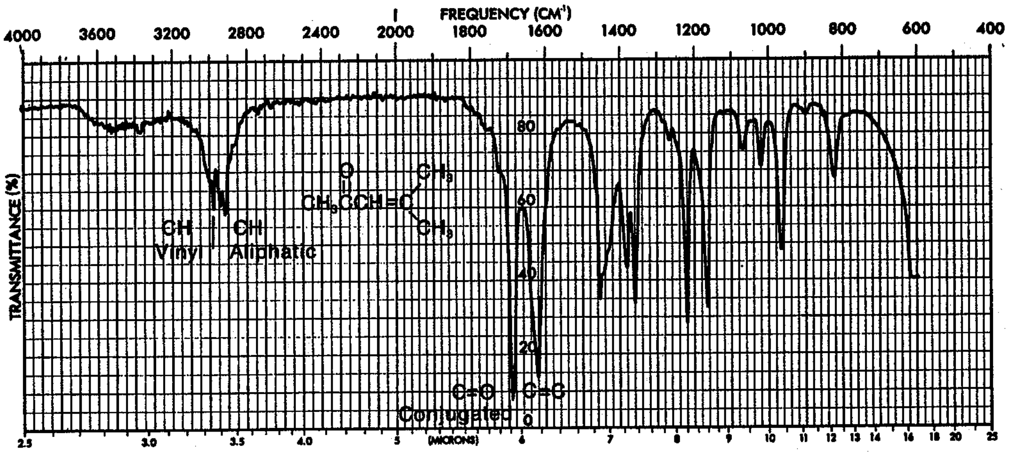

Figure IR.17 Infrared spectrum of mesityl oxide (neat liquid, salt plates).

Esters ($R{-}\overset{O}{\overset{\|}{C}}{-}OR'$)

C=O Stretch occurs at about 1735 cm^{-1} in normal esters.

(a) Conjugation in the R part moves the absorption to the right.

(b) Conjugation with the O in the R′ part moves the absorption to the left.

(c) Ring strain (lactones) moves the absorption to the left.

C—O Stretch, two bands or more, one stronger than the others, is in the range of 1300–1000 cm^{-1}.

The spectrum of methyl benzoate is shown in Figure IR.18. The spectra of isopentyl acetate and methyl salicylate are shown in Experiments 8 and 10.

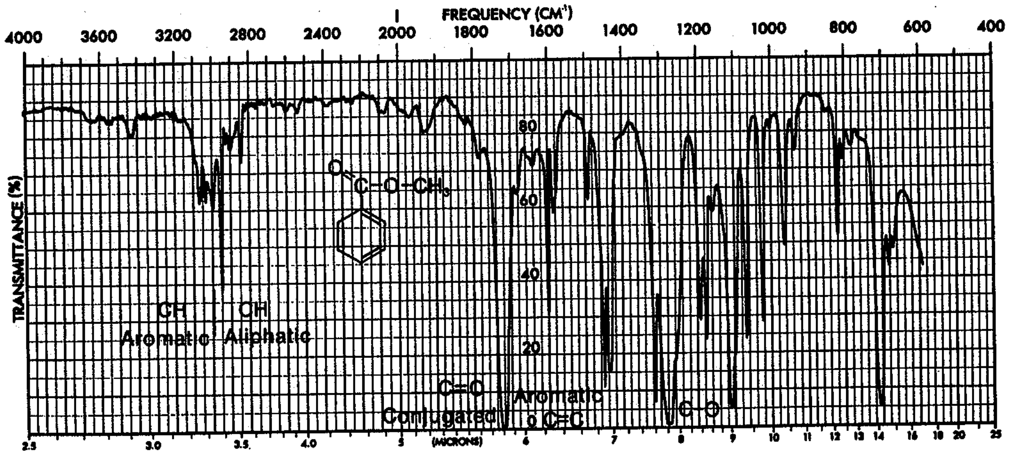

Figure IR.18 Infrared spectrum of methyl benzoate (neat liquid, salt plates).

Amides

C=O	Stretch is at approximately 1670–1640 cm^{-1}. Conjugation and ring size (lactams) have the usual effects.
N—H	Stretch (if monosubstituted or unsubstituted) 3500–3100 cm^{-1}. Unsubstituted amides have two bands (—NH_2) in this region.
N—H	Bending around 1640–1550 cm^{-1}.

The spectrum of benzamide is shown in Figure IR.19.

Anhydrides

C=O	Stretch always has *two* bands: 1830–1800 cm^{-1} and 1775–1740 cm^{-1}. Unsaturation moves the absorptions to the right. Ring strain (cyclic anhydrides) moves the absorptions to the left.
C—O	Stretch is at 1300–900 cm^{-1}. The spectrum of *cis*-norbornane-5,6-*endo*-dicarboxylic anhydride is shown in Experiment 51.

Halides

It is often difficult to determine either the presence or the absence of a halide in a compound by infrared spectroscopy. The absorption bands cannot be relied on, especially if the spectrum is being determined with the compound dissolved in CCl_4 or $CHCl_3$ solution.

C—F	Stretch, 1350–960 cm^{-1}.
C—Cl	Stretch, 850–500 cm^{-1}.
C—Br	Stretch, to the right of 667 cm^{-1}.
C—I	Stretch, to the right of 667 cm^{-1}.

The spectra of carbon tetrachloride and chloroform are shown in Technique 19, Figures 19.6 and 19.7, respectively (pp. 850–851).

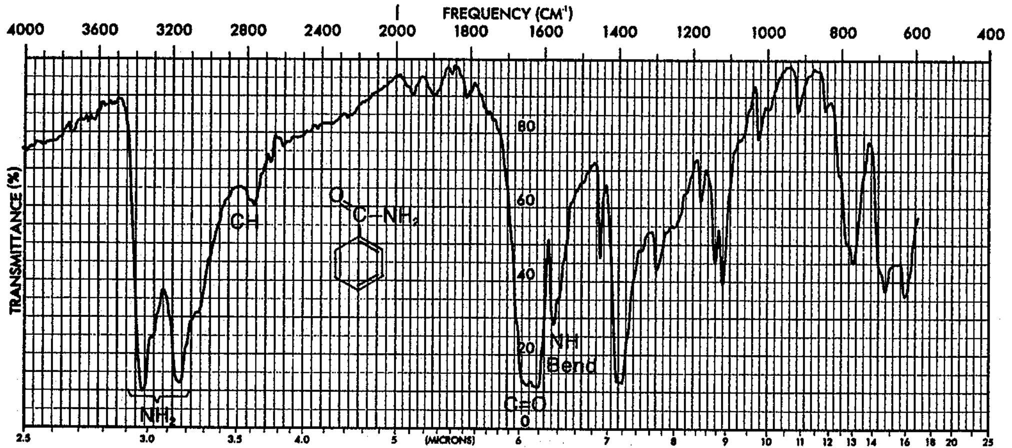

Figure IR.19 Infrared spectrum of benzamide (solid phase, KBr).

REFERENCES

Bellamy, L. J. *The Infra-red Spectra of Complex Molecules.* 3rd ed. New York: Methuen, 1975.

Colthup, N. B., Daly, L. H., and Wiberly, S. E. *Introduction to Infrared and Raman Spectroscopy.* 3rd ed. San Diego, CA: Academic Press, 1990.

Dyer, J. R. *Applications of Absorption Spectroscopy of Organic Compounds.* Englewood Cliffs, NJ: Prentice-Hall, 1965.

Lin-Vien, D., Colthup, N. B., Fateley, W. G., and Grasselli, J. G. *Infrared and Raman Characteristic Frequencies of Organic Molecules.* San Diego, CA: Academic Press, 1991.

Nakanishi, K., and Soloman, P. H. *Infrared Absorption Spectroscopy.* 2nd ed. San Francisco: Holden-Day, 1977.

Pavia, D. L., Lampman, G. M., and Kriz, G. S. *Introduction to Spectroscopy: A Guide for Students of Organic Chemistry.* 2nd ed. Philadelphia: W. B. Saunders, 1996.

Silverstein, R. M., Bassler, G. C., and Morrill, T. C. *Spectrometric Identification of Organic Compounds.* 5th ed. New York: Wiley, 1991.

APPENDIX 4

Nuclear Magnetic Resonance Spectroscopy

NMR.1 THE RESONANCE PHENOMENON

Nuclear magnetic resonance (NMR) spectroscopy is an instrumental technique that allows the number, type, and relative positions of certain atoms in a molecule to be determined. This type of spectroscopy applies only to those atoms that have nuclear magnetic moments because of their nuclear spin properties. Although many atoms meet this requirement, hydrogen atoms ($^{1}_{1}H$) and carbon atoms ($^{13}_{6}C$) are of the greatest interest to the organic chemist. Atoms of the ordinary isotopes of carbon ($^{12}_{6}C$) and oxygen ($^{16}_{8}O$) do not have nuclear magnetic moments, and ordinary nitrogen atoms ($^{14}_{7}N$), although they do have magnetic moments, generally fail to show typical NMR behavior for other reasons. The same is true of the halogen atoms, except for fluorine ($^{19}_{9}F$), which does show active NMR behavior. The hydrogen nucleus is discussed in this appendix and carbon-13 NMR is described in Appendix 5.

Nuclei of NMR-active atoms placed in a magnetic field can be thought of as tiny bar magnets. In hydrogen, which has two allowed nuclear spin states ($+\frac{1}{2}$ and $-\frac{1}{2}$), either the nuclear magnets of individual atoms can be aligned with the magnetic field (spin $+\frac{1}{2}$), or they can be opposed to it (spin $-\frac{1}{2}$). A slight majority of the nuclei are aligned with the field, as this spin orientation constitutes a slightly lower-energy spin state. If radiofrequency waves of the appropriate energy are supplied, nuclei aligned with the field can absorb this radiation and reverse their direction of spin or become reoriented so that the nuclear magnet opposes the applied magnetic field (Fig. NMR.1).

The frequency of radiation required to induce spin conversion is a direct function of the strength of the applied magnetic field. When a spinning hydrogen nucleus is placed

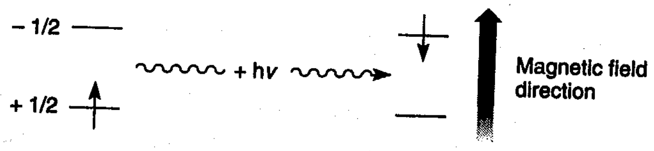

Figure NMR.1 The NMR absorption process.

in a magnetic field, the nucleus begins to precess with angular frequency ω, much like a child's toy top. This processional motion is depicted in Figure NMR.2. The angular frequency of nuclear precession ω increases as the strength of the applied magnetic field is increased. The radiation that must be supplied to induce spin conversion in a hydrogen nucleus of spin $+\frac{1}{2}$ must have a frequency that just matches the angular precessional frequency ω. This is called the resonance condition, and spin conversion is said to be a resonance process.

For the average proton (hydrogen atom), if a magnetic field of approximately 14,000 gauss is applied, radiofrequency radiation of 60 MHz (60,000,000 cycles per second) is required to induce a spin transition. Fortunately, the magnetic-field strength required to induce the various protons in a molecule to absorb 60-MHz radiation varies from proton to proton within the molecule and is a sensitive function of the immediate *electronic* environment of each proton. The typical proton nuclear magnetic resonance spectrometer supplies a basic radiofrequency radiation of 60 MHz to the sample being measured and *increases* the strength of the applied magnetic field over a range of several parts per million from the basic field strength. As the field increases, various protons come into resonance (absorb 60-MHz energy), and a resonance signal is generated for each proton. An NMR spectrum is a plot of the strength of the magnetic field versus the intensity of the absorptions. A typical NMR spectrum is shown in Figure NMR.3.

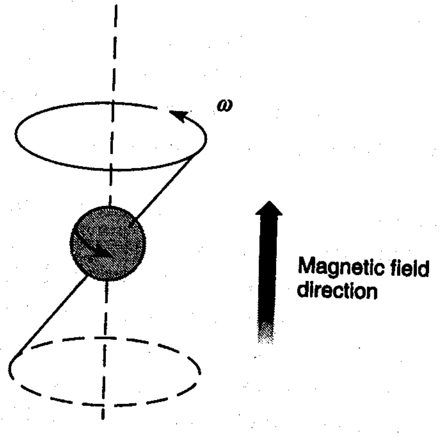

Figure NMR.2 Precessional motion of a spinning nucleus in an applied magnetic field.

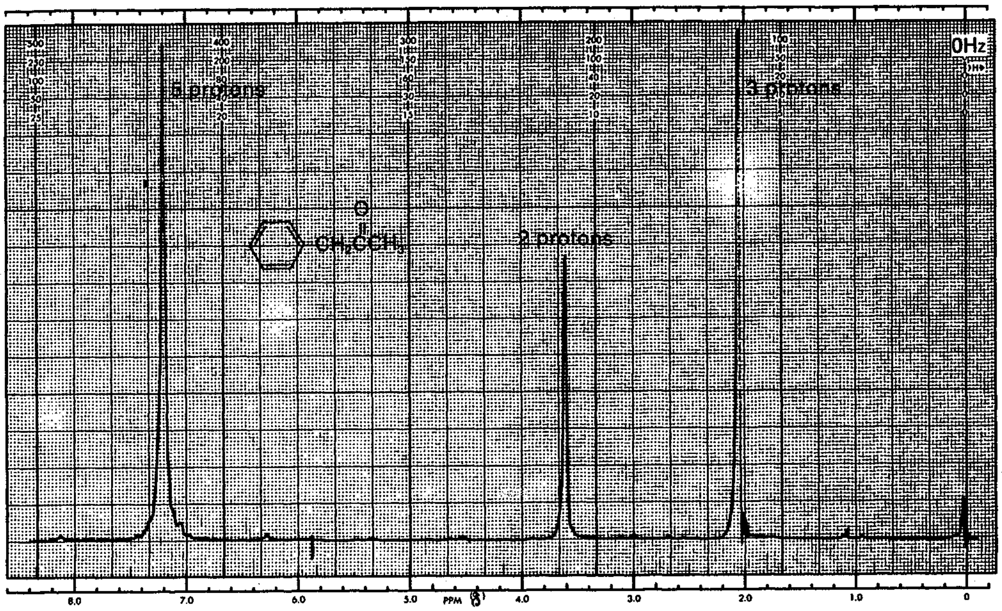

Figure NMR.3 NMR spectrum of phenylacetone (the absorption peak at the far right is caused by the added reference substance, TMS).

NMR.2 THE CHEMICAL SHIFT

The differences in the applied field strengths at which the various protons in a molecule absorb 60-MHz radiation are extremely small. The different absorption positions amount to a difference of only a few parts per million (ppm) in the magnetic field strength. Because it is experimentally difficult to measure the precise field strength at which each proton absorbs to less than one part in a million, a technique has been developed whereby the *difference* between two absorption positions is measured directly. To achieve this measurement, a standard reference substance is used, and the positions of the absorptions of all other protons are measured relative to the values for the reference substance. The reference substance that has been universally accepted is tetramethylsilane $(CH_3)_4Si$, which is also called TMS. The proton resonances in this molecule appear at a higher field strength than the proton resonances in most all other molecules do, and all the protons of TMS have resonance at the same field strength.

To give the position of absorption of a proton a quantitative measurement, a parameter called the **chemical shift** (δ) has been defined. One δ unit corresponds to a 1-ppm change in the magnetic field strength. To determine the chemical shift value for the various protons in a molecule, the operator determines an NMR spectrum of the molecule with a small quantity of **TMS** added directly to the sample. That is, both spectra are determined *simultaneously*. The TMS absorption is adjusted to correspond to the $\delta = 0$ ppm position on the recording chart, which is calibrated in δ units, and the δ values of the absorption peaks for all other protons can be read directly from the chart.

Because the NMR spectrometer increases the magnetic field as the pen moves from left to right on the chart, the TMS absorption appears at the extreme right edge of the spectrum ($\delta = 0$ ppm) or at the *upfield* end of the spectrum. The chart is calibrated in δ units (or ppm), and most other protons absorb at a lower field strength (or *downfield*) from TMS.

Because the frequency at which a proton precesses, and hence the frequency at which it absorbs radiation, is directly proportional to the strength of the applied magnetic field, a second method of measuring an NMR spectrum is possible. You could hold the magnetic field strength constant and vary the frequency of the radiofrequency radiation supplied. Thus, a given proton could be induced to absorb *either* by increasing the field strength, as described earlier, or alternatively, by decreasing the frequency of the radiofrequency oscillator. A 1-ppm decrease in the frequency of the oscillator would have the same effect as a 1-ppm increase in the magnetic field strength. For reasons of instrumental design, it is simpler to vary the strength of the magnetic field than to vary the frequency of the oscillator. Most instruments operate on the former principle. Nevertheless, the recording chart is calibrated not only in δ units but in Hertz (Hz) as well (1 ppm = 60 Hz when the frequency is 60 MHz), and the chemical shift is customarily defined and computed using Hertz rather than gauss:

$$\delta = \text{Chemical shift} = \frac{\text{Observed shift from TMS (in Hz)}}{\text{60 MHz}} = \frac{\text{Hz}}{\text{MHz}} = \text{ppm}$$

Although the equation defines the chemical shift for a spectrometer operating at 14,100 gauss and 60 MHz, the chemical shift value that is calculated is *independent* of the field strength. For instance, at 23,500 gauss the oscillator frequency would have to be 100 MHz. Although the observed shifts from TMS (in Hertz) would be larger at this field strength, the divisor of the equation would be 100 MHz, instead of 60 MHz, and δ would turn out to be identical under either set of conditions.

NMR.3 CHEMICAL EQUIVALENCE—INTEGRALS

All the protons in a molecule that are in chemically identical environments often exhibit the same chemical shift. Thus, all the protons in tetramethylsilane (TMS) or all the protons in benzene, cyclopentane, or acetone have their own respective resonance values all at the same δ value. Each compound gives rise to a single absorption peak in its NMR spectrum. The protons are said to be **chemically equivalent.** On the other hand, molecules that have sets of protons that are chemically distinct from one another may give rise to an absorption peak from each set.

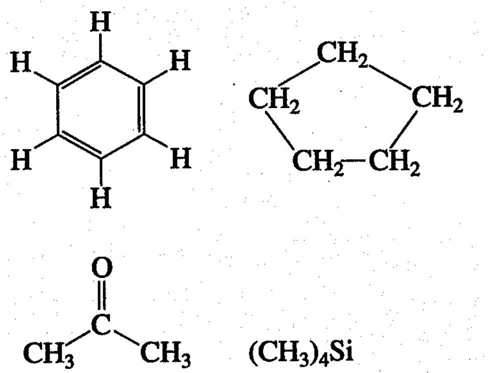

Molecules giving rise to one NMR absorption peak—all protons chemically equivalent

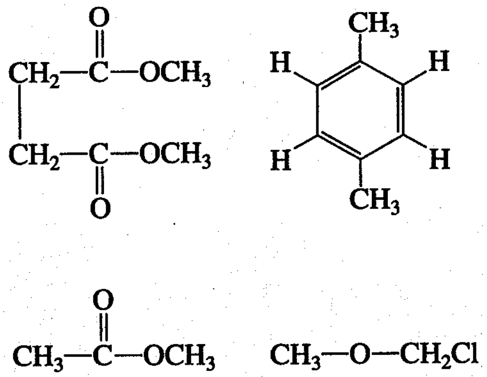

Molecules giving rise to two NMR absorption peaks—two different sets of chemically equivalent protons

The NMR spectrum given in Figure NMR.3 is that of phenylacetone, a compound having *three* chemically distinct types of protons:

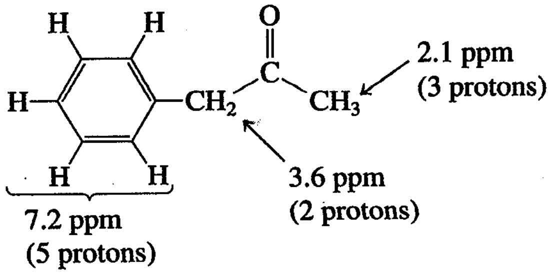

You can immediately see that the NMR spectrum furnishes valuable information on this basis alone. In fact, the NMR spectrum can not only distinguish how many types of protons a molecule has but also can reveal *how many* of each type are contained within the molecule.

In the NMR spectrum, the area under each peak is proportional to the number of hydrogens generating that peak. Hence, in the case of phenylacetone, the area ratio of the three peaks is 5:2:3, the same as the ratio of the numbers of each type of hydrogen. The NMR spectrometer can electronically "integrate" the area under each peak. It does this by tracing over each peak a vertically rising line, which rises in height by an amount proportional to the area under the peak. Shown in Figure NMR.4 is an NMR spectrum of benzyl acetate, with each of the peaks integrated in this way.

It is important to note that the height of the integral line does not give the absolute number of hydrogens; it gives the *relative* numbers of each type of hydrogen. For a given integral to be of any use, there must be a second integral to which it is referred. The benzyl acetate case gives a good example of this. The first integral rises for 55.5 divisions on the chart paper, the second for 22.0 divisions, and the third for 32.5 divisions. These num-

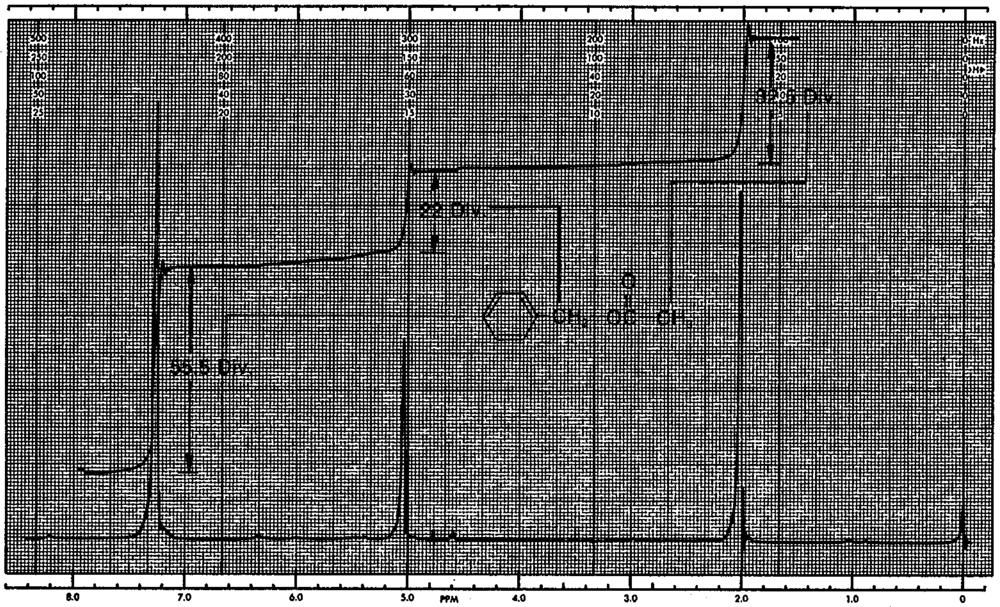

Figure NMR.4 Determination of the integral ratios for benzyl acetate.

bers are relative and give the *ratios* of the various types of protons. You can find these ratios by dividing each of the larger numbers by the smallest number:

$$\frac{55.5 \text{ div}}{22.0 \text{ div}} = 2.52 \qquad \frac{22.0 \text{ div}}{22.0 \text{ div}} = 1.00 \qquad \frac{32.5 \text{ div}}{22.0 \text{ div}} = 1.48$$

Thus, the number ratio of the protons of each type is 2.52:1.00:1.48. If you assume that the peak at 5.1 ppm is really caused by two hydrogens, and if you assume that the integrals are slightly in error (this can be as much as 10%), then you can arrive at the true ratios by multiplying each figure by two and rounding off; we then get 5:2:3. Clearly the peak at 7.3 ppm, which integrates for 5, arises from the resonance of the aromatic ring protons, and the peak at 2.0 ppm, which integrates for 3, is caused by the methyl protons. The two-proton resonance at 5.1 ppm arises from the benzyl protons. Notice then that the integrals give the simplest ratios, but not necessarily the true ratios, of the number of protons of each type.

NMR.4 CHEMICAL ENVIRONMENT AND CHEMICAL SHIFT

If the resonance frequencies of all protons in a molecule were the same, NMR would be of little use to the organic chemist. However, not only do different types of protons have different chemical shifts, they also have a value of chemical shift that characterizes the type of proton they represent. Every type of proton has only a limited range of δ values over which it gives resonance. Hence, the numerical value of the chemical shift for a proton indicates the *type of proton* originating the signal, just as the infrared frequency suggests the type of bond or functional group. Notice, for instance, that the aromatic protons of both phenylacetone (Fig. NMR.3) and benzyl acetate (Fig. NMR.4) have resonance near 7.3 ppm and that both methyl groups attached directly to a carbonyl group have a resonance of approximately 2.1 ppm. Aromatic protons characteristically have resonance near 7–8 ppm, and acetyl groups (the methyl protons) have their resonance near 2 ppm. These values of chemical shift are diagnostic. Notice also how the resonance of the benzyl (—CH_2—) protons comes at a higher value of chemical shift (5.1 ppm) in benzyl acetate than in phenylacetone (3.6 ppm). Being attached to the electronegative element, oxygen, these protons are more deshielded (see Section NMR.5) than the protons in phenylacetone. A trained chemist would have readily recognized the probable presence of the oxygen by the chemical shift shown by these protons.

It is important to learn the ranges of chemical shifts over which the most common types of protons have resonance. Figure NMR.5 is a correlation chart that contains the most essential and frequently encountered types of protons. For the beginner, it is often difficult to memorize a large body of numbers relating to chemical shifts and proton types. You need actually do this only crudely. It is more important to "get a feel" for the regions and the types of protons than to know a string of factual numbers.

The values of chemical shift given in Figure NMR.5 can be easily understood in terms of two factors: local diamagnetic shielding and anisotropy. These two factors are discussed in Sections NMR.5 and NMR.6.

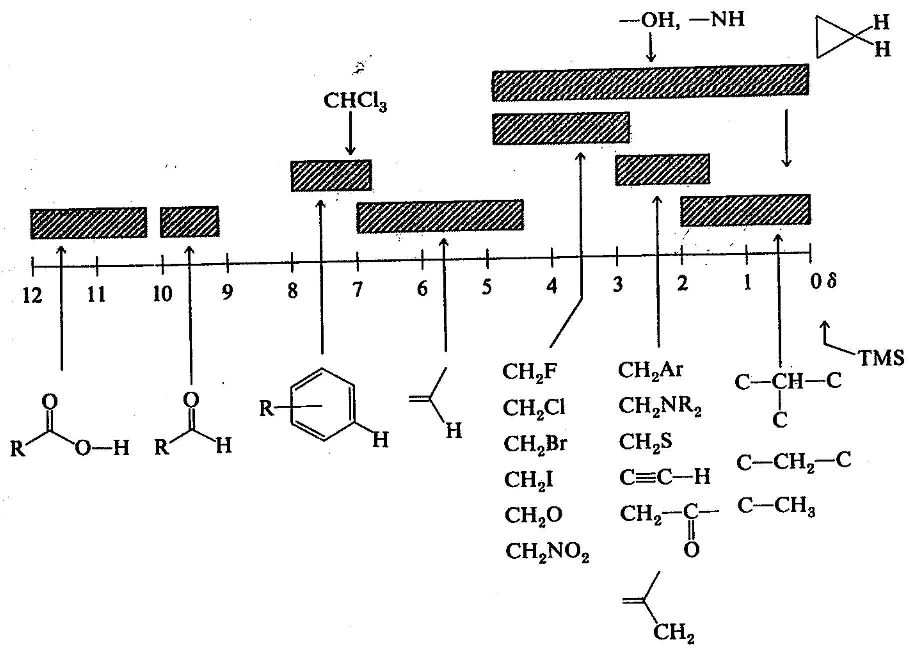

Figure NMR.5 Simplified correlation chart for proton chemical shift values.

NMR.5 LOCAL DIAMAGNETIC SHIELDING

The trend of chemical shifts that is easiest to explain is that involving electronegative elements substituted on the same carbon to which the protons of interest are attached. The chemical shift simply increases as the electronegativity of the attached element increases. This is illustrated in Table NMR.1 for several compounds of the type CH_3X.

Multiple substituents have a stronger effect than a single substituent. The influence of the substituent, an electronegative element having little effect on protons that are more than three carbons away, drops off rapidly with distance. These effects are illustrated in Table NMR.2.

Electronegative substituents attached to a carbon atom, because of their electron-withdrawing effects, reduce the valence electron density around the protons attached to that carbon. These electrons *shield* the proton from the applied magnetic field. This effect, called **local diamagnetic shielding,** occurs because the applied magnetic field induces the valence electrons to circulate and thus to generate an induced magnetic field, which *opposes* the applied field. This is illustrated in Figure NMR.6. Electronegative substituents on carbon reduce the local diamagnetic shielding in the vicinity of the attached protons because they reduce the electron density around those protons. Substituents that produce this effect are said to *deshield* the proton. The greater the electronegativity of the substituent, the more the deshielding of the protons and, hence, the greater the chemical shift of those protons.

NMR.6 ANISOTROPY

Figure NMR.5 clearly shows that several types of protons have chemical shifts not easily explained by simple consideration of the electronegativity of the attached groups. Consider, for instance, the protons of benzene or other aromatic systems. Aryl protons generally have a chemical shift that is as large as that for the proton of chloroform. Alkenes, alkynes, and aldehydes also have protons whose resonance values are not in line with the expected magnitude of any electron-withdrawing effects. In each of these cases, the effect is due to the presence of an unsaturated system (π electrons) in the vicinity of the proton in question. In benzene, for example, when the π electrons in the aromatic ring system are placed in a magnetic field, they are induced to circulate around the ring. This circulation is called a **ring current.** Moving electrons (the ring current) generate a mag-

TABLE NMR.1 Dependence of Chemical Shift of CH_3X on the Element X

Compound CH_3X	CH_3F	CH_3OH	CH_3Cl	CH_3Br	CH_3I	CH_4	$(CH_3)_4Si$
Element X	F	O	Cl	Br	I	H	Si
Electronegativity of X	4.0	3.5	3.1	2.8	2.5	2.1	1.8
Chemical shift (ppm)	4.26	3.40	3.05	2.68	2.16	0.23	0

TABLE NMR.2 Substitution Effects

	$CHCl_3$	CH_2Cl_2	CH_3Cl	$—C\underline{H}_2Br$	$—C\underline{H}_2—CH_2Br$	$—C\underline{H}_2—CH_2CH_2Br$
δ (ppm)	7.27	5.30	3.05	3.30	1.69	1.25

netic field much like that generated in a loop of wire through which a current is induced to flow. The magnetic field covers a spatial volume large enough to influence the shielding of the benzene hydrogens. This is illustrated in Figure NMR.7. The benzene hydrogens are deshielded by the **diamagnetic anisotropy** of the ring. An applied magnetic field is non-uniform (anisotropic) in the vicinity of a benzene molecule because of the labile electrons in the ring that interact with the applied field. Thus, a proton attached to a benzene ring is influenced by *three* magnetic fields: the strong magnetic field applied by the electromagnets of the NMR spectrometer and two weaker fields, one due to the usual shielding by the valence electrons around the proton and the other due to the anisotropy generated by the ring system electrons. It is this anisotropic effect that gives the benzene protons a greater chemical shift than is expected. These protons just happen to lie in a **deshielding** region of this anisotropic field. If a proton were placed in the center of the ring rather than on its periphery, the proton would be shielded, because the field lines would have the opposite direction.

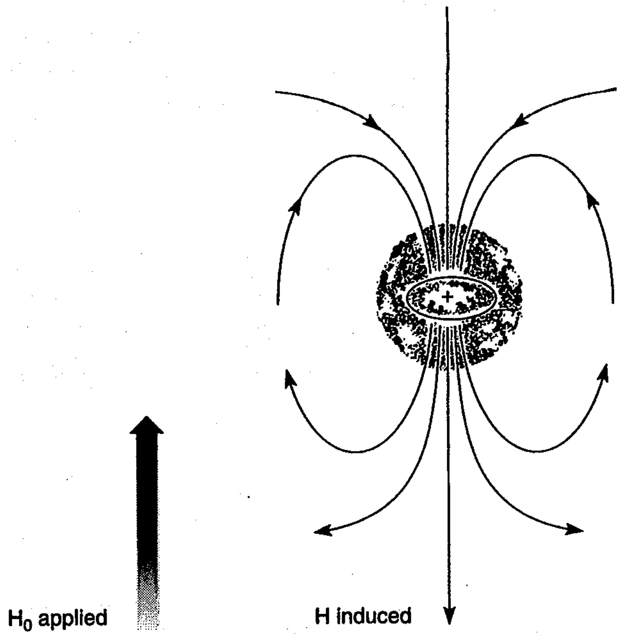

Figure NMR.6 Local diamagnetic shielding of a proton due to its valence electrons.

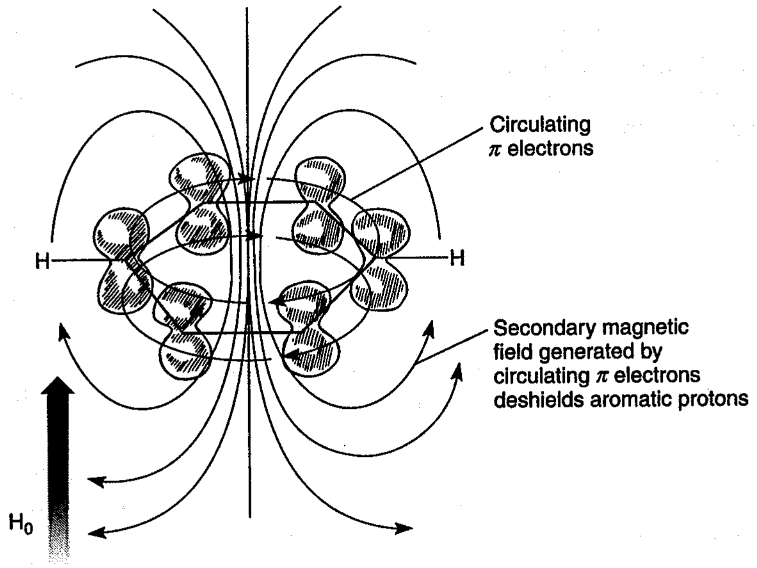
Circulating
π electrons
H
H
Secondary magnetic field generated by circulating π electrons deshields aromatic protons
H_0

Figure NMR.7 Diamagnetic anisotropy in benzene.

All groups in a molecule that have π electrons generate secondary anisotropic fields. In acetylene, the magnetic field generated by induced circulation of π electrons has a geometry such that the acetylene hydrogens are **shielded.** Hence, acetylenic hydrogens come at a higher field than expected. The shielding and deshielding regions due to the various π electron functional groups have characteristic shapes and directions; they are illustrated in Figure NMR.8. Protons falling within the cones are shielded, and those falling outside the conical areas are deshielded. Because the magnitude of the anisotropic field diminishes with distance, beyond a certain distance anisotropy has essentially no effect.

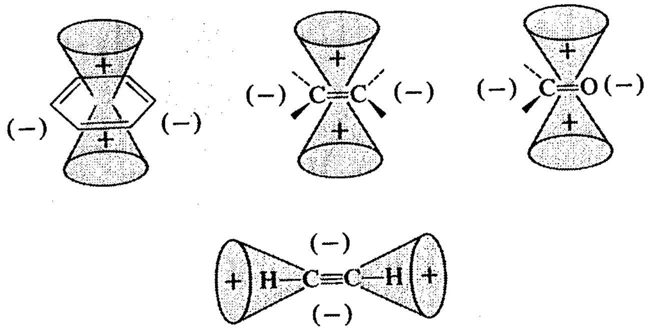

Figure NMR.8 Anisotropy caused by the presence of π electrons in some common multiple-bond systems.

NMR.7 SPIN–SPIN SPLITTING (*N* + 1 RULE)

We have already considered how the chemical shift and the integral (peak area) can give information about the number and type of hydrogens contained in a molecule. A third type of information available from the NMR spectrum is derived from spin–spin splitting. Even in simple molecules, each type of proton rarely gives a single resonance peak. For instance, in 1,1,2-trichloroethane there are two chemically distinct types of hydrogen:

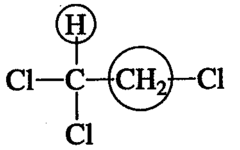

$$\mathrm{Cl{-}\overset{\textcircled{H}}{\underset{Cl}{C}}{-}\textcircled{CH_2}{-}Cl}$$

From information given this far, you would predict *two* resonance peaks in the NMR spectrum of 1,1,2-trichloroethane with an area ratio (integral ratio) of 2:1. In fact, the NMR spectrum of this compound has *five* peaks. A group of three peaks (called a **triplet**) exists at 5.77 ppm and a group of two peaks (called a **doublet**) is found at 3.95 ppm. The spectrum is shown in Figure NMR.9. The methine (CH) resonance (5.77 ppm) is split into a triplet, and the methylene resonance (3.95 ppm) is split into a doublet. The area under the three triplet peaks is *one*, relative to an area of *two* under the two doublet peaks.

This phenomenon is called **spin–spin splitting.** Empirically, spin–spin splitting can be explained by the "*n* + 1 rule." Each type of proton "senses" the number of equivalent protons (*n*) on the carbon atom or atoms next to the one to which it is bonded, and its resonance peak is split into *n* + 1 components.

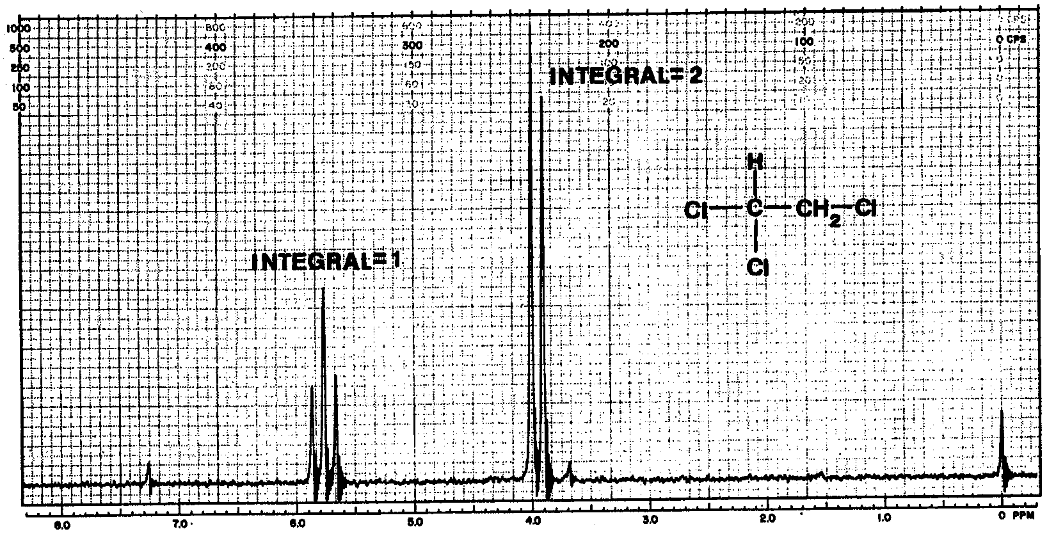

Figure NMR.9 NMR spectrum of 1,1,2-trochloroethane (courtesy of Varian Associates).

Let's examine the case at hand, 1,1,2-trichloroethane, using the $n + 1$ rule. First, the lone methine hydrogen is situated next to a carbon bearing two methylene protons. According to the rule, it has two equivalent neighbors ($n = 2$) and is split into $n + 1 = 3$ peaks (a triplet). The methylene protons are situated next to a carbon bearing only one methine hydrogen. According to the rule, they have one neighbor ($n = 1$) and are split into $n + 1 = 2$ peaks (a doublet).

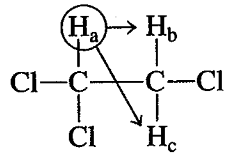

Two neighbors give a triplet ($n + 1 = 3$) (area = 1)

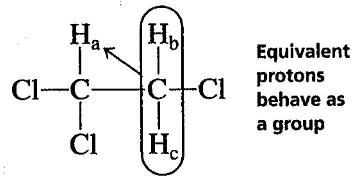

One neighbor gives a doublet ($n + 1 = 2$) (area = 2)

The spectrum of 1,1,2-trichloroethane can be explained easily by the interaction, or coupling, of the spins of protons on adjacent carbon atoms. The position of absorption of proton H_a is affected by the spins of protons H_b and H_c attached to the neighboring (adjacent) carbon atom. If the spins of these protons are aligned with the applied magnetic field, the small magnetic field generated by their nuclear spin properties will augment the strength of the field experienced by the first-mentioned proton H_a. The proton H_a will thus be *deshielded.* If the spins of H_b and H_c are opposed to the applied field, they will decrease the field experienced by proton H_a. It will then be *shielded.* In each of these situations, the absorption position of H_a will be altered. Among the many molecules in the solution, you will find all the various possible spin combinations for H_b and H_c; hence, the NMR spectrum of the molecular solution will give *three* absorption peaks (a triplet) for H_a because H_b and H_c have three different possible spin combinations (Fig. NMR.10). By a similar analysis, it can be seen that protons H_b and H_c should appear as a doublet.

Some common splitting patterns that can be predicted by the $n + 1$ rule and that are frequently observed in a number of molecules are shown in Figure NMR.11. Notice particularly the last entry, where *both* methyl groups (six protons in all) function as a unit and split the methine proton into a septet ($6 + 1 = 7$).

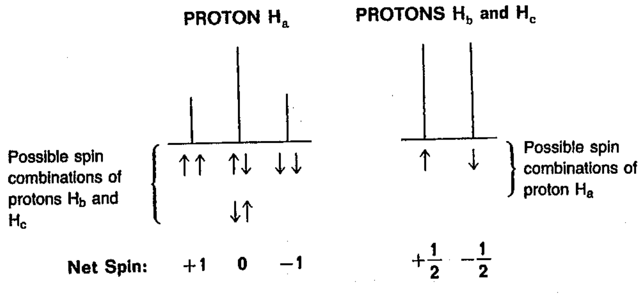
PROTON H_a
PROTONS H_b and H_c
Possible spin combinations of protons H_b and H_c
↑↑ ↑↓ ↓↓
↓↑
↑ ↓
Possible spin combinations of proton H_a
Net Spin: +1 0 −1
$+\frac{1}{2}$ $-\frac{1}{2}$

Figure NMR.10 Analysis of the spin–spin splitting pattern of 1,1,2-trichloroethane.

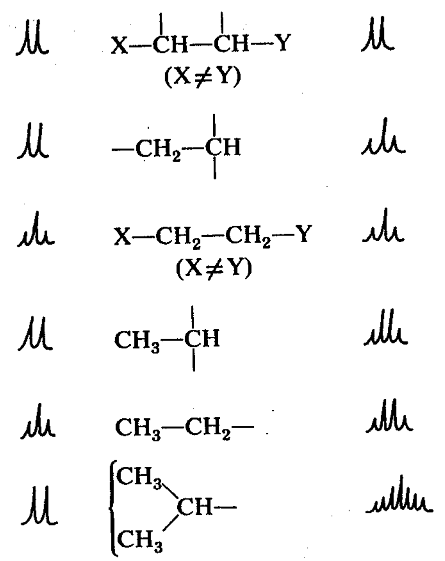

Figure NMR.11 Some common splitting patterns.

NMR.8 THE COUPLING CONSTANT

The quantitative amount of spin–spin interaction between two protons can be defined by the **coupling constant.** The spacing between the component peaks in a simple multiplet is called the coupling constant *J*. This distance is measured on the same scale as the chemical shift and is expressed in Hertz (Hz).

For the interaction of most aliphatic protons in acyclic systems, the magnitudes of the coupling constants are always near 7.5 Hz. See, for instance, the NMR spectrum of 1,1,2-trichloroethane in Figure NMR.9, where the coupling constant is approximately 6 Hz. Different magnitudes of *J* are found for different types of protons. For instance, the *cis* and *trans* protons substituted on a double bond commonly have values where $J_{trans} \cong$ 17 Hz and $J_{cis} \cong$ 10 Hz are typical coupling constants. In ordinary compounds, coupling constants may range anywhere from 0 to 18 Hz. The magnitude of *J* often provides structural clues. You can usually distinguish, for example, between a *cis* olefin and a *trans* olefin on the basis of the observed coupling constants for the vinyl protons. The approximate values of some representative coupling constants are given in Table NMR.3.

NMR.9 MAGNETIC EQUIVALENCE

In the example of spin–spin splitting in 1,1,2-trichloroethane, notice that the two protons H_b and H_c, which are attached to the same carbon atom, do not split one another.

TABLE NMR.3 Representative Coupling Constants and Approximate Values (Hz)

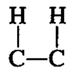	6–8	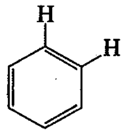	*ortho* 6–10	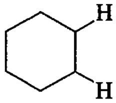	a,a 8–14 a,e 0–7 e,e 0–5
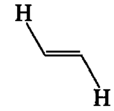	11–18				
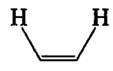	6–15	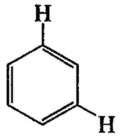	*meta* 1–4	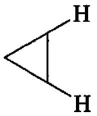	*cis* 6–12 *trans* 4–8
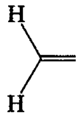	0–5	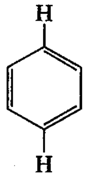	*para* 0–2	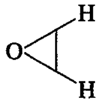	*cis* 2–5 *trans* 1–3
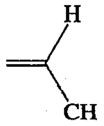	4–10	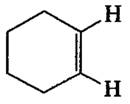	8–11	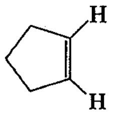	5–7
H—C=C—CH	0–3				

They behave as an integral group. Actually the two protons H_b and H_c *are* coupled to one another; however, for reasons we cannot explain fully here, protons that are attached to the same carbon and both of which have the *same chemical shift* do not show spin–spin splitting. Another way of stating this is that protons coupled to the same extent to *all* other protons in a molecule do not show spin–spin splitting. Protons that have the same chemical shift and are coupled equivalently to all other protons are *magnetically equivalent* and do not show spin–spin splitting. Thus, in 1,1,2-trichloroethane, protons H_b and H_c have the same value of δ and are coupled by the same value of J to proton H_a. They are magnetically equivalent.

It is important to differentiate magnetic equivalence and chemical equivalence. Note the following two compounds:

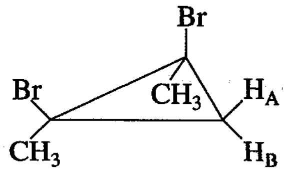

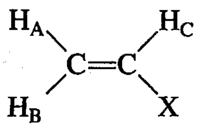

In the cyclopropane compound, the two geminal hydrogens are chemically equivalent; however, they are not magnetically equivalent. Proton H_A is on the same side of the ring as the two halogens. Proton H_B is on the same side of the ring as the two methyl groups. Protons H_A and H_B will have different chemical shifts, will couple to one another, and will show spin-spin splitting. Two doublets will be seen for H_A and H_B. For cyclopropane rings, $J_{geminal}$ is usually around 5 Hz.

Another situation in which protons are chemically equivalent but not magnetically equivalent exists in the vinyl compound. In this example, protons A and B are chemically equivalent but not magnetically equivalent. H_A and H_B have different chemical shifts. In addition, a second distinction can be made between H_A and H_B in this type of compound. Each has a different coupling constant with H_C. The constant J_{AC} is a *cis* coupling constant, and J_{BC} is a *trans* coupling constant. Whenever two protons have different coupling constants relative to a third proton, they are not magnetically equivalent. In the vinyl compound, H_A and H_B do not act as a group to split proton H_C. Each proton acts independently. Thus, H_B splits H_C with coupling constant J_{BC} into a doublet, and then H_A splits each of the components of the doublet into doublets with coupling constant J_{AC}. In such a case, the NMR spectrum must be analyzed graphically, splitting by splitting. An NMR spectrum of a vinyl compound is shown in Figure NMR.12. The graphical analysis of the vinyl portion of the NMR spectrum is in Figure NMR.13.

NMR.10 AROMATIC COMPOUNDS

The NMR spectra of protons on aromatic rings are often too complex to explain by simple theories. However, some simple generalizations can be made that are useful in an-

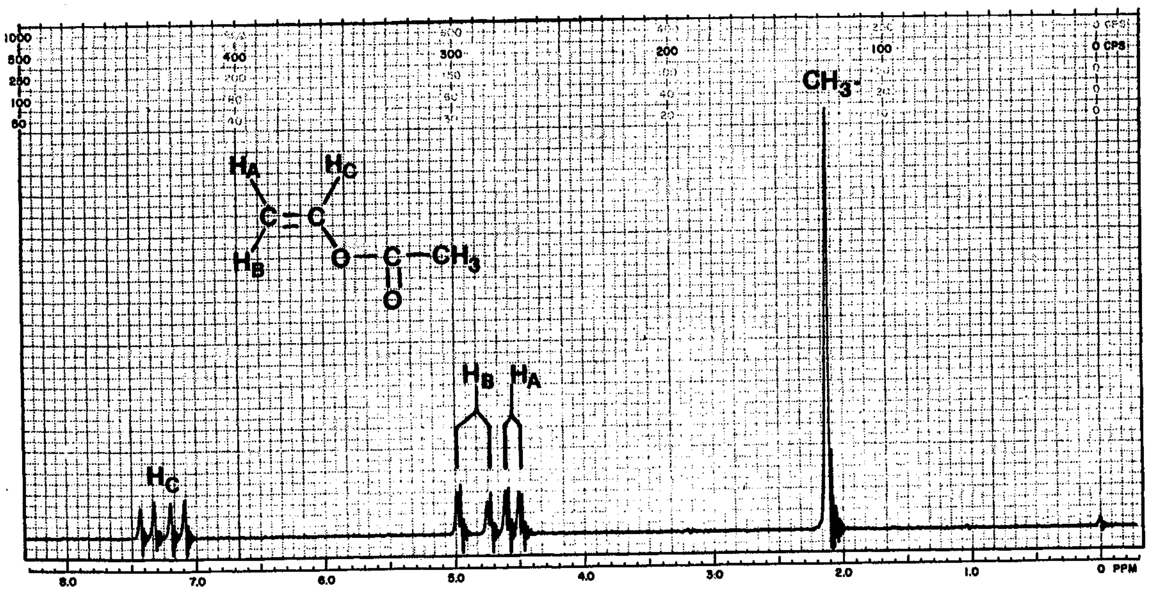

Figure NMR.12 NMR spectrum of vinyl acetate (courtesy of Varian Associates).

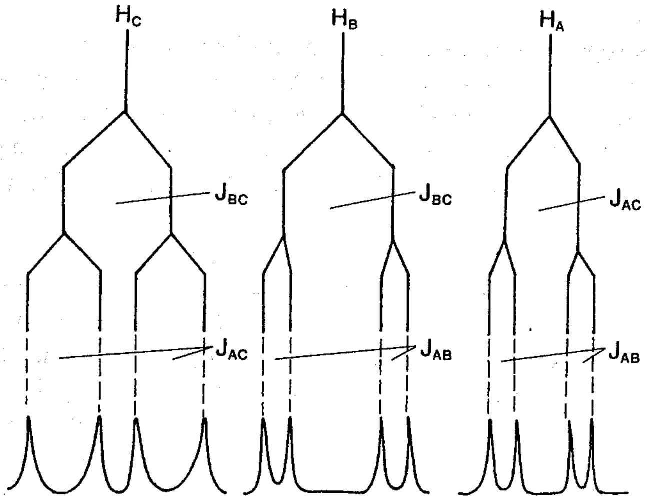
H_C
H_B
H_A
J_{BC}
J_{BC}
J_{AC}
J_{AC}
J_{AB}
J_{AB}

Figure NMR.13 Analysis of the splittings in vinyl acetate.

alyzing the aromatic region of the NMR spectrum. First of all, most aromatic protons have resonance near 7.0 ppm. In monosubstituted rings in which the ring substituent is an alkyl group, all the ring protons often have chemical shifts that are very nearly identical, and the five ring protons may appear as if they gave rise to an overly broad singlet (Fig. NMR.14A). If an electronegative group is attached to the ring, all the ring protons are

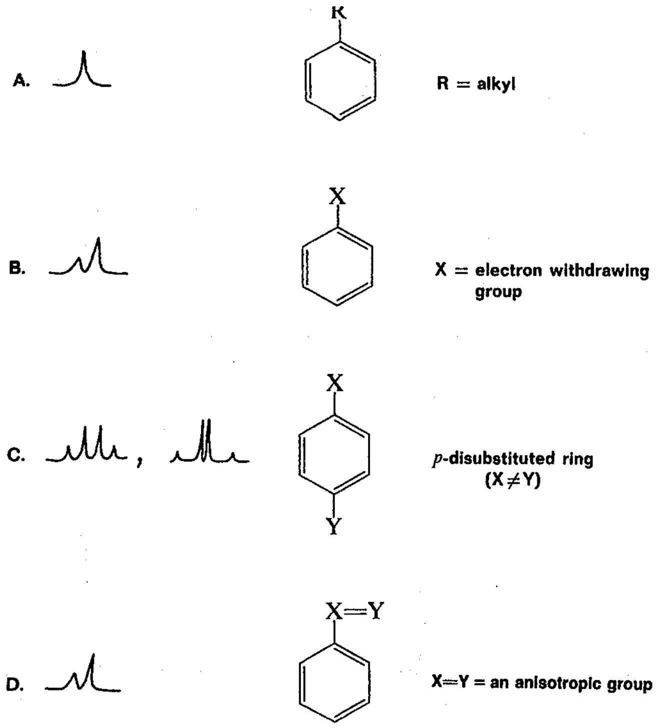
A.
K
R = alkyl
B.
X
X = electron withdrawing group
C.
X
Y
p-disubstituted ring (X≠Y)
D.
X=Y
X=Y = an anisotropic group

Figure NMR.14 Some common aromatic patterns.

shifted downfield from where they would appear in benzene. However, often the *ortho* protons are shifted more than the others, as they are more affected by the group. This often gives rise to an absorption pattern like that in Figure NMR.14B. In a *para*-disubstituted ring with two substituents X and Y that are identical, all the protons in the ring are chemically and magnetically equivalent, and a singlet is observed. If X is different from Y in electronegativity, however, a pattern like that shown in the left side of Figure NMR.14C is often observed, clearly identifying a *p*-disubstituted ring. If X and Y are more nearly similar, a pattern more like the one on the right is observed. In monosubstituted rings that have a carbonyl group or a double bond attached directly to the ring, a pattern like that in Figure NMR.14D is not uncommon. In this case, the *ortho* protons of the ring are influenced by the anisotropy of the π systems that make up the CO and CC double bonds and are deshielded by them. In other types of substitution, such as *ortho* or *meta*, or polysubstituted ring systems, the patterns may be much more complicated and require an advanced analysis.

NMR.11 PROTONS ATTACHED TO ATOMS OTHER THAN CARBON

Protons attached to atoms other than carbon often have a widely variable range of absorptions. Several of these groups are tabulated in Table NMR.4. In addition, under the usual conditions of determining an NMR spectrum, protons on heteroelements normally do not couple with protons on adjacent carbon atoms to give spin–spin splitting. This is primarily because such protons often exchange very rapidly with those of the solvent medium. The absorption position is variable because these groups also undergo varying degrees of hydrogen bonding in solutions of different concentrations. The amount of hydrogen bonding that occurs with a proton radically affects the valence electron density around that proton and produces correspondingly large changes in the chemical shift. The absorption peaks for protons that have hydrogen bonding or are undergoing exchange are frequently broad relative to other singlets and can often be recognized on that basis. For a different reason, called quadrupole broadening, protons attached to nitrogen atoms often show an extremely broad resonance peak, often almost indistinguishable from the baseline.

TABLE NMR.4 Typical Ranges for Groups with Variable Chemical Shift

Acids	RCOOH	10.5–12.0 ppm
Phenols	ArOH	4.0–7.0
Alcohols	ROH	0.5–5.0
Amines	RNH_2	0.5–5.0
Amides	$RCONH_2$	5.0–8.0
Enols	CH═CH—OH	≥15

NMR.12 SPECTRA AT HIGHER FIELD STRENGTH

Occasionally, the 60-MHz spectrum of an organic compound, or a portion of it, is almost undecipherable because the chemical shifts of several groups of protons are all very similar. In these cases, all the proton resonances occur in the same area of the spectrum, and peaks often overlap so extensively that individual peaks and splittings cannot be extracted. One of the ways in which such a situation can be simplified is by using a spectrometer that operates at a higher frequency. Although both 60- and 90-MHz instruments are quite common, it is not unusual to find instruments with operating frequencies of 100, 220, 300 MHz, or even higher.

Although NMR coupling constants are not dependent on the frequency or the field strength of operation of the NMR spectrometer, chemical shifts in Hertz are dependent on these parameters. This circumstance can often be used to simplify an otherwise undecipherable spectrum. Suppose, for instance, that a compound contained three multiplets: a quartet and two triplets derived from groups of protons with very similar chemical shifts. At 60 MHz these peaks might overlap, as illustrated in Figure NMR.15, and simply give an unresolved envelope of absorption.

Figure NMR.15 also shows the spectrum of the same compound at two higher field strengths (frequencies). In redetermining the spectrum at higher field strengths, the coupling constants do not change, but the chemical shifts in Hertz (not ppm) of the proton groups (H_A, H_B, H_C) responsible for the multiplets do increase. It should be noted that at 220 MHz the individual multiplets are cleanly separated and resolved.

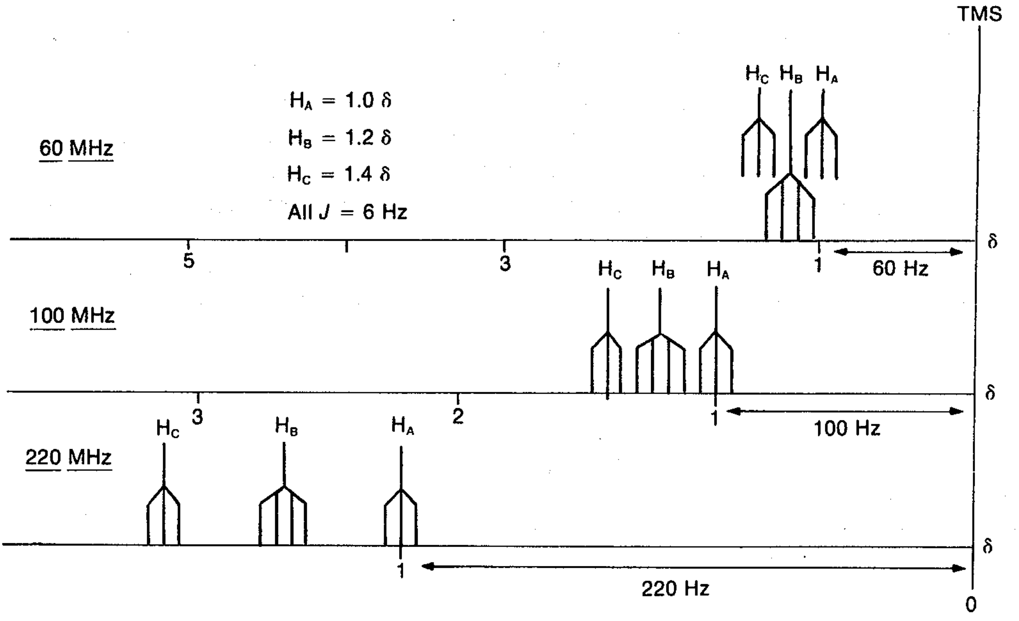

Figure NMR.15 A comparison of the spectrum of a compound with overlapping multiplets at 60 MHz with spectra of the same compound also determined at 100 MHz and 220 MHz. The drawing is to scale.

NMR.13 CHEMICAL SHIFT REAGENTS

Researchers have known for some time that interactions between molecules and solvents, such as those due to hydrogen bonding, can cause large changes in the resonance positions of certain types of protons (e.g., hydroxyl and amino). They have also known that the resonance positions of some groups of protons can be greatly affected by changing from the usual NMR solvents such as CCl_4 and $CDCl_3$ to solvents like benzene, which impose local anisotropic effects on surrounding molecules. In many cases, it was possible to resolve partially overlapping multiplets by such a solvent change. However, the use of chemical shift reagents for this purpose dates from about 1969. Most of these chemical shift reagents are organic complexes of paramagnetic rare earth metals from the lanthanide series of elements. When these metal complexes are added to the compound whose spectrum is being determined, profound shifts in the resonance positions of the various groups of protons are observed. The direction of the shift (upfield or downfield) depends primarily on which metal is being used. Complexes of europium, erbium, thulium, and ytterbium shift resonances to lower field; complexes of cerium, praseodymium, neodymium, samarium, terbium, and holmium generally shift resonances to higher field. The advantage of using such reagents is that shifts similar to those observed at higher field can be induced without the purchase of an expensive higher field instrument.

Of the lanthanides, europium is probably the most commonly used metal. Two of its widely used complexes are *tris*-(dipivalomethanato)europium and *tris*-(6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octanedionato)europium. These are frequently abbreviated $Eu(dpm)_3$ and $Eu(fod)_3$, respectively.

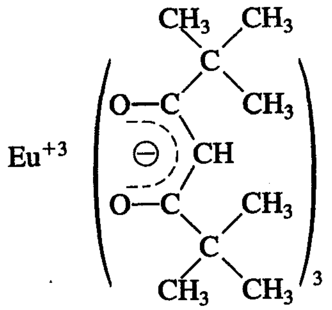
Eu(dpm)$_3$ or Eu(thd)$_3$

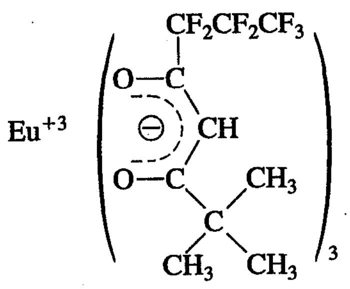
Eu(fod)$_3$

These lanthanide complexes produce spectral simplifications in the NMR spectrum of any compound that has a relatively basic pair of electrons (unshared pair) that can coordinate with Eu^{3+}. Typically, aldehydes, ketones, alcohols, thiols, ethers, and amines will all interact:

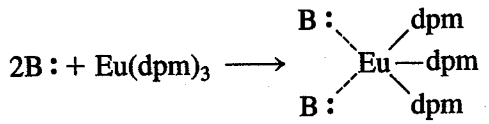

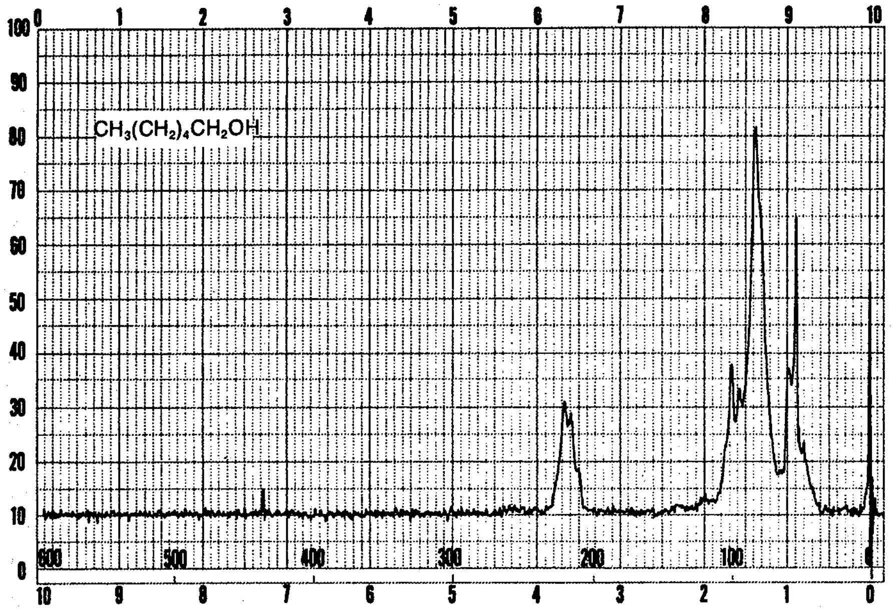

Figure NMR.16 The normal 60-MHz spectrum of hexanol (courtesy of Aldrich Chemical Co.).

The amount of shift that a given group of protons will experience depends (1) on the distance separating the metal (Eu^{3+}) and that group of protons, and (2) on the concentration of the shift reagent in the solution. Because of the latter dependence, it is necessary when reporting a lanthanide-shifted spectrum to report the number of mole equivalents of shift reagent used or its molar concentration.

The distance factor is illustrated in the spectra of hexanol, which are given in Figures NMR.16 and NMR.17. In the absence of shift reagent, the normal spectrum is obtained

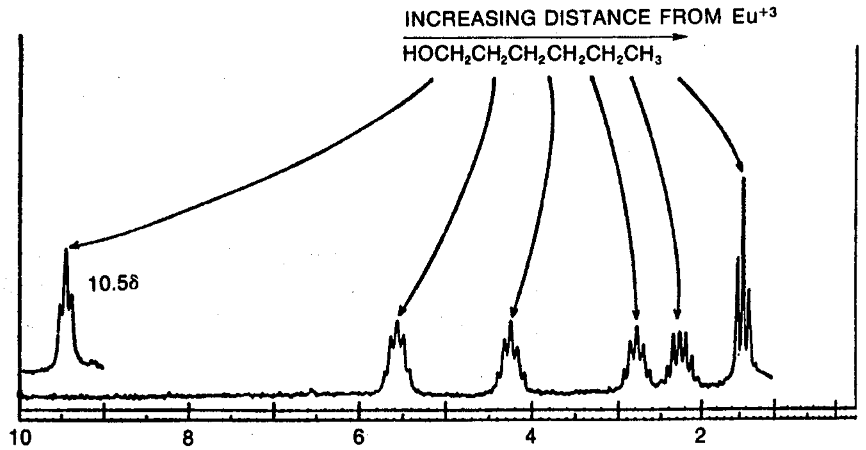

Figure NMR.17 The 100-MHz spectrum of hexanol with 0.29 mole equivalents of $Eu(dpm)_3$. (Reprinted with permission from J.K.M. Sanders and D.H. Williams, *Chem. Commun.*, 422[1970].)

(Figure NMR.16). Only the triplet of the terminal methyl group and the triplet of the methylene group next to the hydroxyl are resolved in the spectrum. The other protons (aside from OH) are found together in a broad unresolved group. With shift reagent added (Figure NMR.17), each of the methylene groups is clearly separated and resolved into the proper multiplet structure. The spectrum is first order and simplified; all the splittings are explained by the $n + 1$ rule.

One final consequence of using a shift reagent should be noted. Notice in Figure NMR.17 that the multiplets are not as nicely resolved into sharp peaks as you might expect. This is due to the fact that shift reagents cause a small amount of peak broadening. At high shift reagent concentrations this problem becomes serious, but at most useful concentrations the amount of broadening experienced is tolerable.

REFERENCES

Friebolin, H. *Basic One- and Two-Dimensional NMR Spectroscopy.* New York: VCH Publishers, 1991.

Jackman, L. M., and Sternhell, S. *Applications of Nuclear Magnetic Resonance in Organic Chemistry.* 2nd ed. New York: Pergamon Press, 1969.

Paudler, W. W. *Nuclear Magnetic Resonance.* New York: John Wiley, 1987.

Pavia, D. L., Lampman, G. M., and Kriz, G. S. *Introduction to Spectroscopy.* 2nd ed. Philadelphia: W. B. Saunders, 1996.

Sanders, J. K. M., and Hunter, B. K. *Modern NMR Spectroscopy.* 2nd ed. Oxford: Oxford University Press, 1993.

Silverstein, R. M., Bassler, G. C., and Morrill, T. C. *Spectrometric Identification of Organic Compounds.* 5th ed. New York: John Wiley 1991.

APPENDIX 5

Carbon-13 Nuclear Magnetic Resonance Spectroscopy

CMR.1 CARBON-13 NUCLEAR MAGNETIC RESONANCE

Carbon-12, the most abundant isotope of carbon, does not possess spin ($I = 0$); it has both an even atomic number and an even atomic weight. The second principal isotope of carbon, ^{13}C, however, does have the nuclear spin property ($I = \frac{1}{2}$). ^{13}C atom resonances are not easy to observe, due to a combination of two factors. First, the natural abundance of ^{13}C is low; only 1.08% of all carbon atoms are ^{13}C. Second, the magnetic moment μ of ^{13}C is low. For these two reasons, the resonances of ^{13}C are about *6000 times weaker* than those of hydrogen. With special Fourier transform instrumental techniques, which are not discussed here, it is possible to observe ^{13}C nuclear magnetic resonance (carbon-13) spectra on samples that contain only the natural abundance of ^{13}C.

The most useful parameter derived from carbon-13 spectra is the chemical shift. Integrals are unreliable and are not necessarily related to the relative numbers of ^{13}C atoms present in the sample. Hydrogens that are attached to ^{13}C atoms cause spin–spin splitting, but spin–spin interaction between adjacent carbon atoms is rare. With its low natural abundance (0.0108), the probability of finding two ^{13}C atoms adjacent to one another is extremely low.

CMR.2 COMPLETELY COUPLED ^{13}C SPECTRA

Figure CMR.1 shows the carbon-13 spectrum of ethyl phenylacetate. Consider first the upper trace shown in the figure. Chemical shifts, just as in proton NMR, are reported by the number of ppm (δ units) that the peak is shifted downfield from TMS. Keep in mind, however, that it is a ^{13}C atom of the methyl group of TMS that is being observed, not the 12 methyl hydrogens. Notice the extent of the scale. Although the chemical shifts of protons encompass a range of only about 20 ppm, ^{13}C chemical shifts cover an extremely wide range of up to 200 ppm! Under these circumstances, even adjacent —CH_2— carbons in a long hydrocarbon chain generally have their own distinct resonance peaks, and these peaks are clearly resolved. It is unusual to find any two carbon atoms in a molecule having resonance at the same chemical shift unless these two carbon atoms are equivalent by symmetry.

Returning to the upper spectrum in Figure CMR.1, you can see that the first quartet downfield from TMS (14.2 ppm) corresponds to the carbon of the methyl group. It is

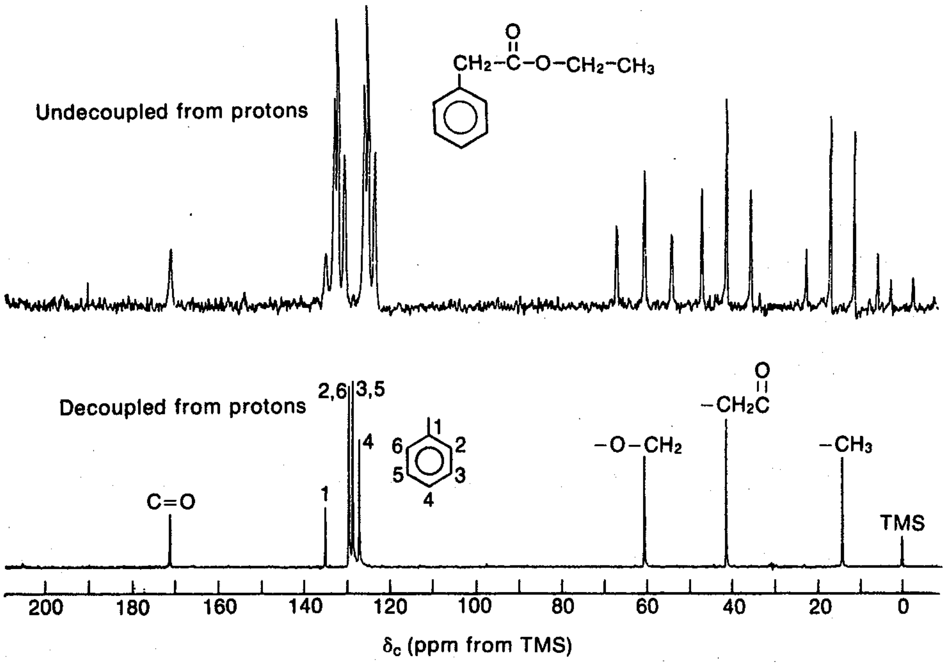

Figure CMR.1 Carbon-13 spectra of ethyl phenylacetate. (From Rodig, O.R., Bell, C.E., Jr., and Clark, A.K., *Organic Chemistry Laboratory; Standard and Microscale Experiments,* 1990 by Saunders College Publishing, Philadelphia, PA. Reprinted by permission of the publisher.)

split into a quartet (J = 127 Hz) by the three attached hydrogen atoms. In addition, although it cannot be seen on the scale of this spectrum, each of the quartet lines is split into a *closely spaced* triplet (J = ca 1 Hz). This additional fine splitting is caused by the two protons on the adjacent —CH_2— group. These are geminal couplings (H—C—^{13}C) of a type that commonly occurs in carbon-13 spectra, with coupling constants that are generally small (J = 0–2 Hz). The quartet is caused by **direct coupling** (^{13}C—H). Direct coupling constants are larger, usually about 100–200 Hz and are more obvious on the scale in which the spectrum is presented.

There are two —CH_2— groups in ethyl phenylacetate. The one corresponding to the ethyl —CH_2— group is found further downfield (60.6 ppm), as this carbon is deshielded by the attached oxygen. It is a triplet because of the two attached hydrogens. Again, although it is not seen in this unexpanded spectrum, each of the triplet peaks is finely split into a quartet by the three hydrogens on the adjacent methyl group. The benzyl —CH_2— carbon is the intermediate triplet (41.4 ppm). Furthest downfield is the carbonyl group carbon (171.1 ppm). On the scale of presentation, it is a singlet (no directly attached hydrogens), but because of the adjacent benzyl —CH_2— group, it is actually split finely into a triplet. The aromatic ring carbons also appear in the spectrum, and they have resonances over the range from 127 ppm to 136 ppm.

CMR.3 BROAD-BAND DECOUPLED ^{13}C SPECTRA

Although the splittings in a simple molecule such as ethyl phenylacetate yield interesting structural information, namely the number of hydrogens attached to each carbon (as well as those adjacent if the spectrum is expanded), for large molecules the carbon-13 spectrum becomes very complex due to these splittings, and the splitting patterns often overlap. It is customary, therefore, to decouple *all* the protons in the molecule by irradiating them simultaneously with a broad spectrum of frequencies in the proper range. This type of spectrum is said to be **completely decoupled.** The completely decoupled spectrum is much simpler and, for larger molecules, much easier to interpret. The decoupled spectrum of ethyl phenylacetate is presented in the lower trace of Figure CMR.1.

In the completely decoupled carbon-13 spectrum, each peak represents a different carbon atom. If two carbons are represented by a single peak, they must be equivalent by symmetry. Thus, the carbons at positions 2 and 6 of the aromatic ring of ethyl phenylacetate give a single peak, and the carbons at positions 3 and 5 also give a single peak in the lower spectrum of Figure CMR.1.

CMR.4 CHEMICAL SHIFTS

Just as is the case for proton spectra, the chemical shift of each carbon indicates both its type and its structural environment. In fact, a correlation chart can be presented for ^{13}C chemical shift ranges, similar to the correlation chart for proton resonances shown in Figure NMR.5. Figure CMR.2 gives typical chemical shift ranges for the types of carbon resonances.

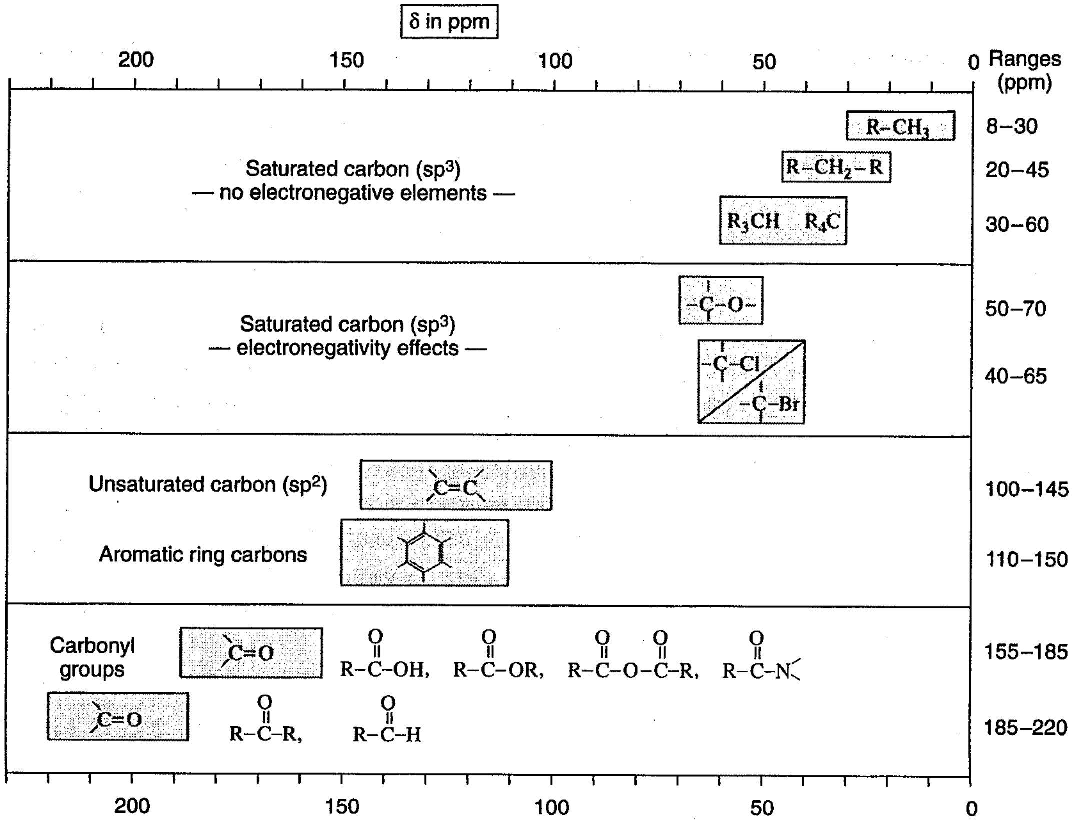

Figure CMR.2 A correlation chart for ^{13}C chemical shifts (ppm from TMS).

Electronegativity, hybridization, and anisotropy effects all influence ^{13}C chemical shifts, just as they do for protons, but in a more complex fashion. These factors are not discussed in any detail here, but note that the —CH_2— group carbon attached to oxygen in ethyl phenylacetate has a larger chemical shift than the —CH_2— carbon of the benzyl group. Note also that the carbonyl carbon appears relatively far downfield, probably due to an anisotropy effect.

CMR.5 SOME SAMPLE SPECTRA

The following spectra illustrate some of the effects that can be observed in carbon-13 spectra. The spectrum of 2,2-dimethylbutane is presented in Figure CMR.3. Notice that, although this compound has six total carbon atoms, 2,2-dimethylbutane shows only four peaks in the carbon-13 spectrum along with the solvent peaks ($CDCl_3$) and TMS. The carbon-13 atoms that are equivalent appear at the same chemical shift value. Thus, a single methyl carbon **a** appears at highest field (8.8 ppm), while the three equivalent methyl carbons **b** appear at 28.9 ppm. The quaternary carbon **c** gives rise to the small peak of 30.4 ppm, whereas the methylene carbon **d** appears at 36.5 ppm. The relative sizes of

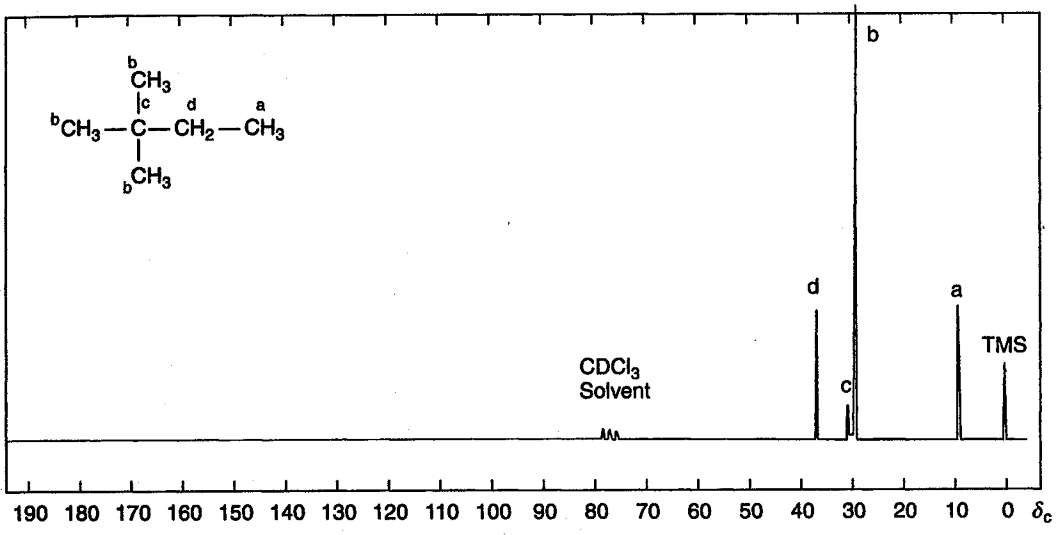

Figure CMR.3 Carbon-13 spectrum of 2,2-dimethylbutane. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

peaks gives some idea of the numbers of each type of carbon atom present in the molecule. For example, in Figure CMR.3 notice that the peak at 28.9 ppm (**b**) is much larger than the other peaks. A characteristic of proton-decoupled ^{13}C NMR spectra is that carbon atoms that do not have hydrogens attached to them generally appear as weak peaks. Thus, the quaternary carbon at 30.4 ppm (**c**) is very weak (see Section CMR.6).

The presence of an electronegative element should deshield a carbon atom closest to it, as is illustrated in the cases of bromocyclohexane (Fig. CMR.4) and cyclohexanol

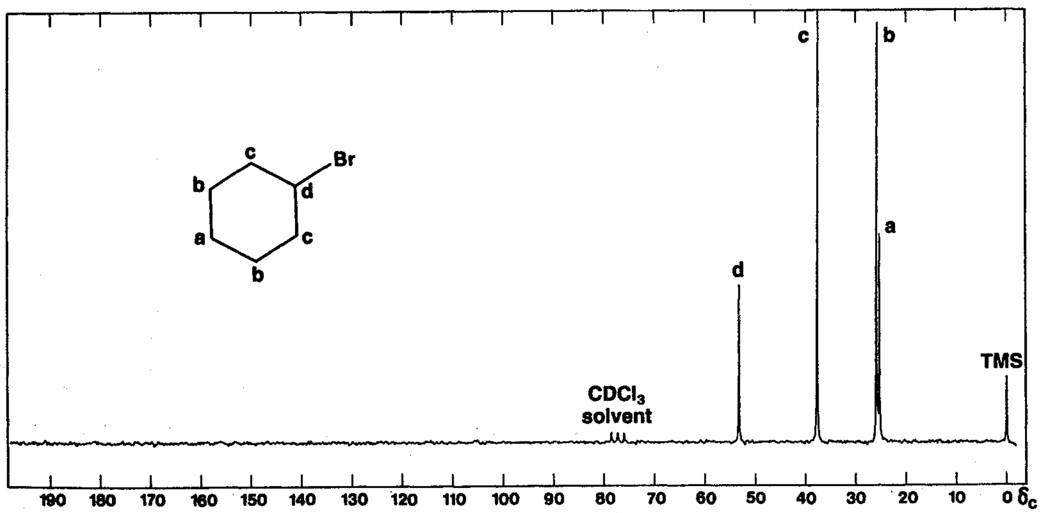

Figure CMR.4 Carbon-13 spectrum of bromocyclohexane. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

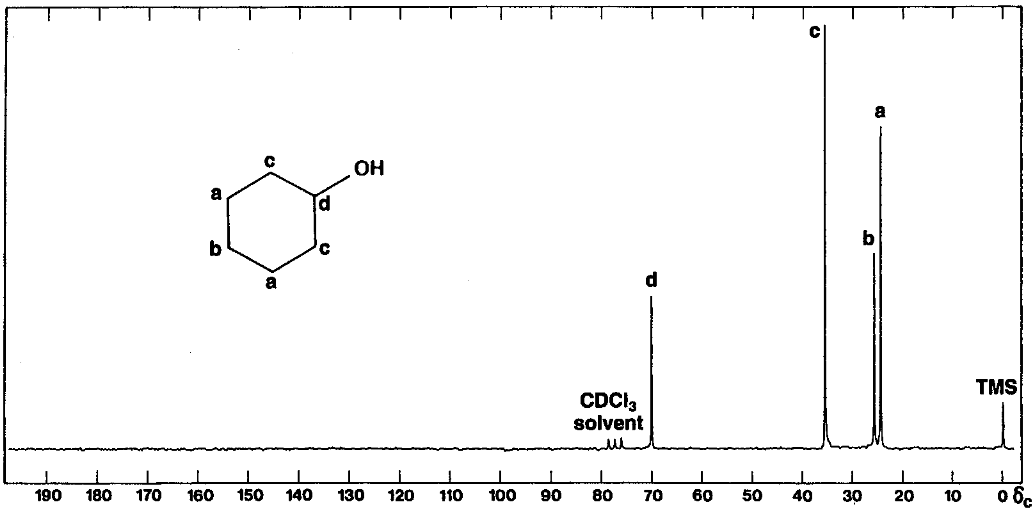

Figure CMR.5 Carbon-13 spectrum of cyclohexanol. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

(Fig. CMR.5). The carbon bearing the bromine in bromocyclohexane appears at 53.0 ppm; the carbon bearing the hydroxyl group of cyclohexanol appears at 70.0 ppm. In each of these cases, note that as the ring carbons are located farther away from the electronegative element, their resonances appear at higher field. A carbon attached to a double bond appears deshielded, due to diamagnetic anisotropy. This effect can be seen in the spectrum of cyclohexene (Fig. CMR.6). The carbon atoms of the double bond appear at 127.2

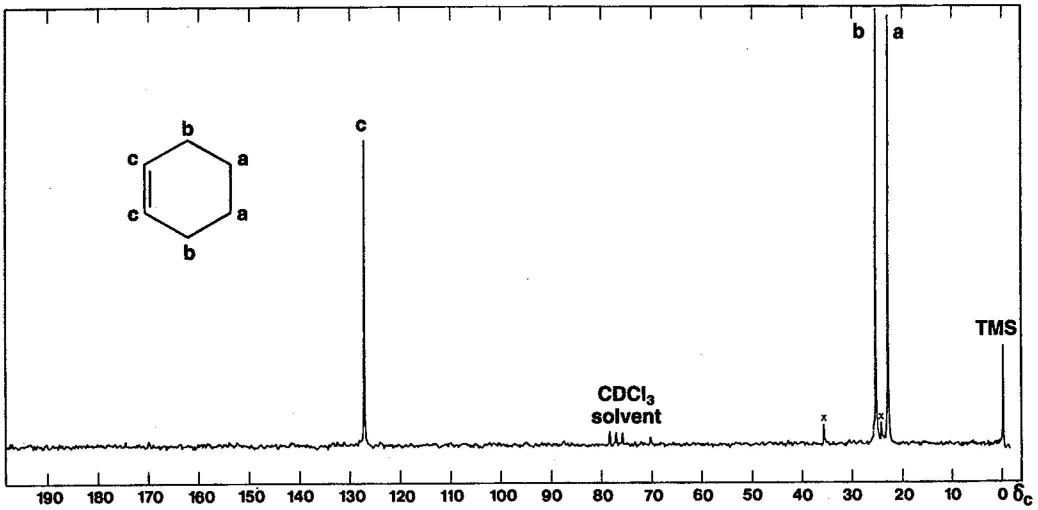

Figure CMR.6 Carbon-13 spectrum of cyclohexene. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

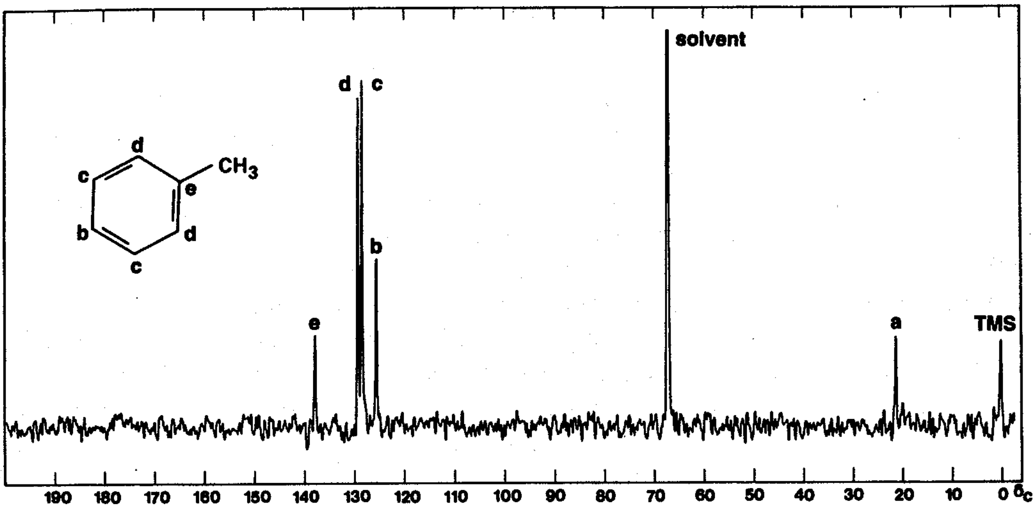

Figure CMR.7 Carbon-13 spectrum of toluene. (The peak at 67.4 is due to the solvent, dioxane.) (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

ppm. Again, it can be seen that as carbon atoms are located farther from the double bond, their resonances appear at higher field. The effect of diamagnetic anisotropy can be seen in the spectrum of toluene (Fig. CMR.7), where the carbon atoms of the aromatic ring appear at low field (125.5–137.7 ppm). Finally, the strong deshielding experienced by the carbon atom of a carbonyl group can be seen in the carbon-13 spectrum of cyclohexanone (Fig. CMR.8). The carbon atom appears at a chemical shift of 211.3 ppm.

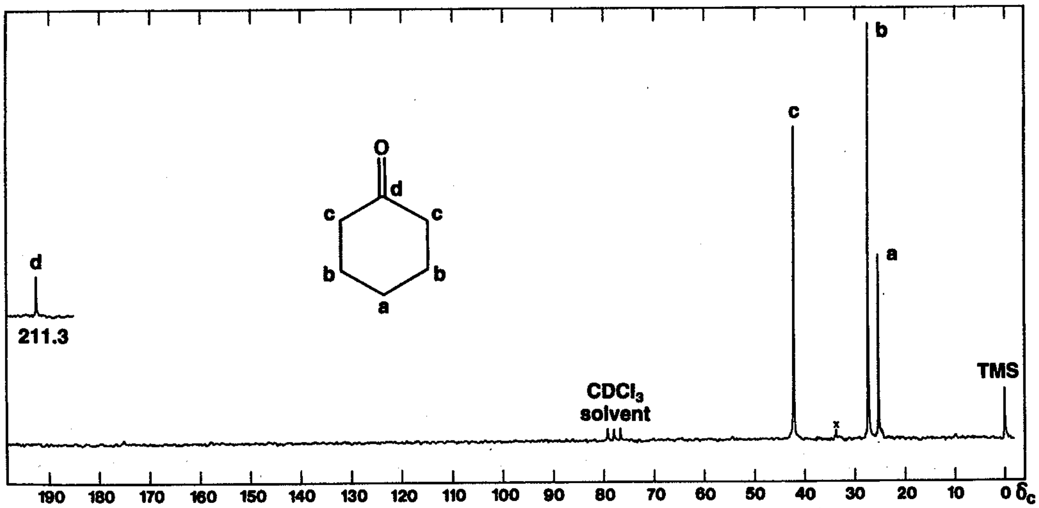

Figure CMR.8 Carbon-13 spectrum of cyclohexanone. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

CMR.6 NUCLEAR OVERHAUSER EFFECT

As mentioned previously, integrals (areas under peaks) are not as reliable for carbon spectra as they are for hydrogen spectra. This is due in part to the **nuclear Overhauser effect.** This effect operates when two dissimilar adjacent atoms (in this case carbon and hydrogen) both exhibit spins and are NMR active. The atoms can influence the NMR absorption intensities of each other. The effect can be either positive or negative, but in the case of carbon-13 interacting with hydrogen, the effect is positive. As a result, carbon-13 NMR absorptions vary in intensity with respect to the number of hydrogen atoms that are directly attached to the carbon atom being observed. In general, the more hydrogens that are attached to a given carbon, the stronger its NMR absorption. Other factors also influence the absorption intensities (they are related to molecular relaxation phenomena), so the number of attached hydrogens can only be taken as a single factor influencing absorption intensity; this is often a very helpful factor in deciding which carbon to assign to a given absorption. In Figure CMR.1 note the low intensity of the carbonyl carbon (172 ppm), and in Figure CMR.7 note the low intensity of the ring carbon to which the methyl group is attached (138 ppm). The carbonyl peak in cyclohexanone (Fig. CMR.8) is also weak. None of these carbons has attached hydrogens.

CMR.7 AN EXAMPLE OF SYMMETRY

As one example of the utility of carbon-13 experiments, consider the cases of the isomers 1,2- and 1,3-dichlorobenzene. Although these isomers could be difficult to dis-

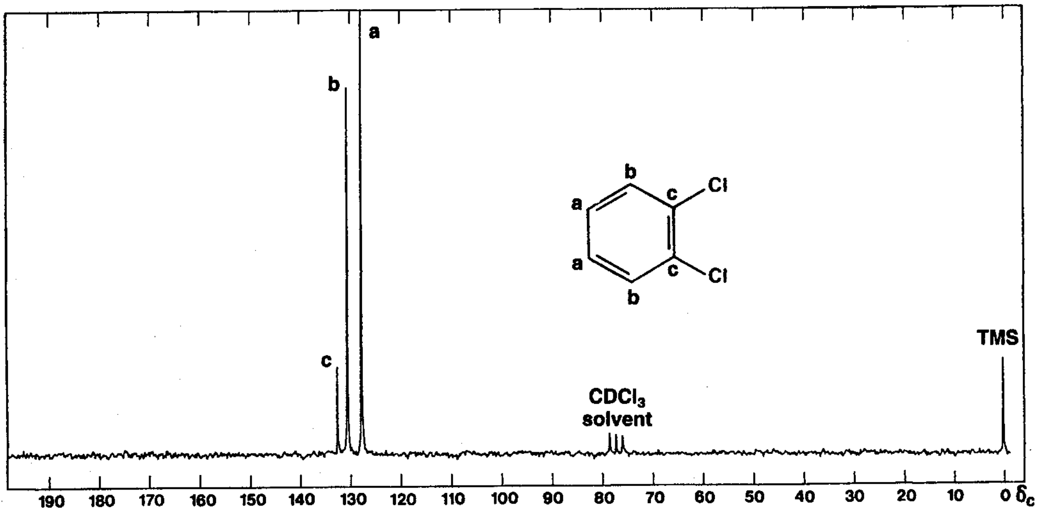

Figure CMR.9 Carbon-13 spectrum of 1,2-dichlorobenzene. (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

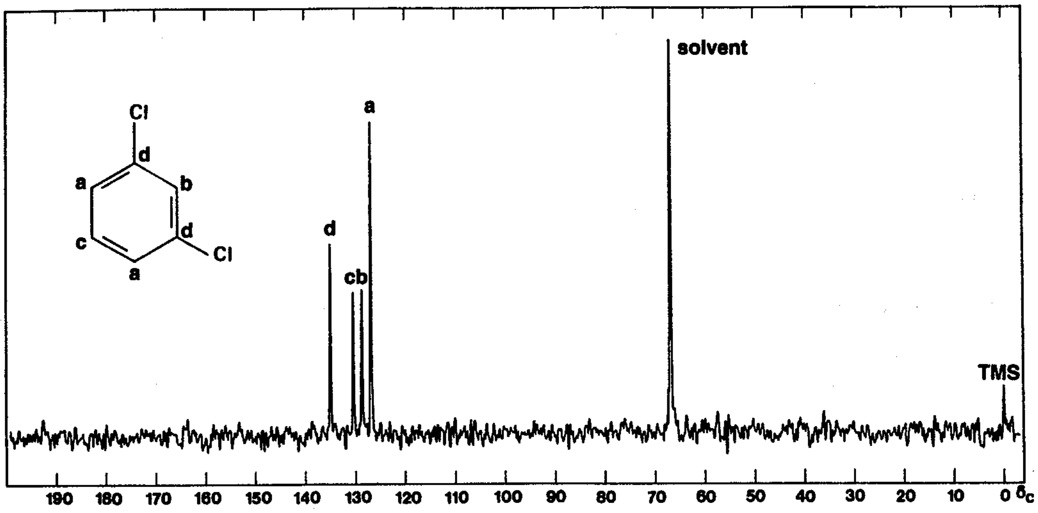

Figure CMR.10 Carbon-13 spectrum of 1,3-dichlorobenzene. (The peak at 67.4 ppm is due to the solvent, dioxane.) (From Johnson, L.F., and Jankowski, W.C., *Carbon-13 NMR Spectra: A Collection of Assigned, Coded, and Indexed Spectra,* 1972 by John Wiley and Sons, New York. Reprinted by permission of the publisher.)

tinguish from one another on the basis of their boiling points or their infrared spectra, each can be identified clearly by their carbon-13 spectra. 1,2-Dichlorobenzene has a plane of symmetry that gives it only three different types of carbon atoms. 1,3-Dichlorobenzene has a plane of symmetry that gives it four different types of carbon atoms. The proton-decoupled carbon-13 spectra of these two compounds are shown in Figures CMR.9 and CMR.10, respectively. It is easy to see the differences in the carbon-13 spectra of these two isomers.

REFERENCES

Breitmaier, E., and Voelter, W. *Carbon-13 NMR Spectroscopy.* 3rd ed. New York: VCH Publishers, 1987.

Friebolin, H. *Basic One- and Two-Dimensional NMR Spectroscopy.* New York: VCH Publishers, 1991.

Levy, G. C., Lichter, R. L., and Nelson, G. L. *Carbon-13 Nuclear Magnetic Resonance for Organic Chemists.* 2nd ed. New York: John Wiley, 1980.

Paudler, W. W. *Nuclear Magnetic Resonance.* New York: John Wiley, 1987.

Pavia, D. L., Lampman, G. M., and Kriz, G. S. *Introduction to Spectroscopy.* 2nd ed. Philadelphia: W. B. Saunders, 1996.

Sanders, J. K. M., and Hunter, B. K. *Modern NMR Spectroscopy.* 2nd ed. Oxford: Oxford University Press, 1993.

Silverstein, R. M., Bassler, G. C., and Morrill, T. C. *Spectrometric Identification of Organic Compounds.* 5th ed. New York: John Wiley, 1991.

APPENDIX 6

Index of Spectra

Infrared Spectra

2-Acetylcyclohexanone 407
p-Aminobenzoic acid 387
Anisole 905
Benzaldehyde 368
Benzamide 910
Benzil 371
Benzilic acid 373
Benzocaine 394
Benzoic acid 287
Benzoin 368
Benzonitrile 907
Borneol 183
n-Butyl bromide 209
n-Butylamine 905
Camphor 183
Carbon disulfide 853
Carbon tetrachloride 850
Carvone 196
Chloroform 851
Decane 900
1,2-Dichlorobenzene 903
7,7-Dichloronorcarane 259
N,N-Diethyl-*m*-toluamide 150
6-Ethoxycarbonyl-3,5-diphenyl-2-cyclohexenone 417
Isoborneol 183
Isopentyl acetate 99
Limonene 196
Mesityl oxide 909
Methyl benzoate 909
Methyl isopropyl ketone 895

Methyl *m*-nitrobenzoate 300
Methyl salicylate 106
4-Methylcyclohexanol 252
4-Methylcyclohexene 251
Mineral oil 854
2-Naphthol 904
Nitrobenzene 906
Nonanal 908
cis-Norbornene-5,6-*endo*-dicarboxylic anhydride 435
Nujol 854
Paraffin oil 854
t-Pentyl chloride 210
Polystyrene 855
Propargyl alcohol 903
Styrene 901
Sulfanilamide 383
Tetraphenylcyclopentadienone 397
1,2,3,4-Tetraphenylnaphthalene 441
Triphenylmethanol 285

^{1}H NMR Spectra

2-Acetylcyclohexanone 407
Benzocaine 394
Benzyl acetate 916
Borneol 185
Camphor 185
Carvone 197
7,7-Dichloronorcarane 259
Ethyl 3-hydroxybutanoate 575
1-Hexanol 930
Isoborneol 185
Limonene 197
Methyl salicylate 107
1-Methylcyclohexanol 275
trans-2-Methylcyclohexanol 274

Phenylacetone 913
1,2,3,4-Tetraphenylnaphthalene 441
1,1,2-Trichloroethane 921
Vinyl acetate 925

^{13}C NMR Spectra

Borneol 186
Bromocyclohexane 935
Camphor 186
Carvone 199
Cyclohexanol 936
Cyclohexanone 937
Cyclohexene 936
1,2-Dichlorobenzene 938
1,3-Dichlorobenzene 939
7,7-Dichloronorcarane 260
2,2-Dimethylbutane 935
Ethyl phenylacetate 932
Isoborneol 187
Toluene 937

Ultraviolet-Visible Spectra

Benzophenone 447
Naphthalene 447

Mixtures

Borneol and isoborneol 187
t-Butyl chloride and *t*-butyl bromide 220
1-Chlorobutane and 1-bromobutane 220

INDEX

A

Acetamides, 524
  preparation, 890
p-Acetamidobenzenesulfonyl chloride, 380
Acetaminophen, 53, 60, 79
  isolation, 42
  preparation, 60
  structure, 42
Acetanilide, 53
  chlorosulfonation, 378
  preparation, 57
Acetate esters,
  preparation, 100
Acetic acid
  hazards, 13
Acetone
  hazards, 13
  wash, 32
Acetophenone
  aldol condensation, 409
Acetyl chloride test, 523, 525
Acetylation
  p-aminophenol, 60
  aniline, 57
  aromatic substrates, 306
  C-4 and C-5 alcohols, 100
  cyclohexanone, 405
  isopentyl alcohol, 96
  salicylic acid, 49
  p-Toluidine, 383
2-Acetylcyclohexanone
  calculation of enol content, 406
  infrared spectrum, 407
  NMR spectrum, 407
  preparation, 398
(S)-(+)-O-Acetylmandelic Acid,
  use as a chiral resolving agent, 568
Acetylsalicylic acid
  preparation, 49
  See also Aspirin
Acid hydrazides, 532
Acids
  removal by extraction, 700
  See also Carboxylic Acids
Activated adsorbent, 766
Activated charcoal, 660
Actual yield, 27
Acylglycerol, 111
Adapter
  Neoprene, 38
  thermometer, 37, 712
  vacuum, 37, 712
Addition funnel, 622
Addition methods
  liquid reagents, 622
  Addition reactions
  benzyne to tetraphenylcyclopentadienone, 436
  bromine to 4-methylcyclohexene, 251
  bromine to unknowns, 504
  diborane to an alkene, 261, 263, 266
  dichlorocarbene to cyclohexene, 253
  mercuric acetate to an alkene, 261, 265, 270
  water to an alkene, 261
Adsorbent interactions, 762
Air drying, 662
Albumins, 553
Alcohol
  detection, 230
  derivatives, 528, 886, 891
  esterification, 100
  identification tests, 524
  infrared spectroscopy, 528, 904
  NMR spectroscopy, 528, 927
  oxidation, 234
  spectroscopy, 528
  tables of unknowns, 886
Aldehydes
  derivatives, 514, 878, 888
  identification tests, 508
  infrared spectroscopy, 513, 906
  NMR spectroscopy, 514, 921
  spectroscopy, 513
  tables of unknowns, 878
Aldol condensation, 395, 409, 413
  acetophenone, 409
  benzaldehyde, 409
  crossed, 409
Aldrich Library of Infrared Spectra, 863
Aldrich Library of NMR Spectra, 863
Aldrin, 428
Alkaloid, 64
Alkanes
  infrared spectroscopy, 900
Alkenes
  C-H out-of-plane bending, 901
  identification tests, 503
  infrared spectroscopy, 506, 900
  NMR spectroscopy, 506, 920, 923
Alkyl chlorides
  preparation, 221, 225
  rate of hydrolysis, 221
Alkyl halides
  identification tests, 496
  reactivities, 200
  synthesis, 204
  See also Halides
Alkynes
  identification tests, 503
  infrared spectroscopy, 506, 903
  NMR spectroscopy, 506, 880
Alumina, 761
Alumina G, 794
Aluminum collars, 606
Aluminum heating block, 605
Amide derivatives, 516
  preparation, 889
  Amide formation, 383
  p-aminophenol, 60
  3-nitrophthalic acid, 455
  m-toluic acid, 145
Amide hydrolysis, 383
Amides
  infrared spectroscopy, 910
  preparation, 57

Amines
  derivatives, 524, 884, 890
  identification tests, 520
  infrared spectroscopy, 523, 905
  NMR spectroscopy, 523, 886
  spectroscopy, 523
  tables of unknowns, 883
p-Aminobenzoic acid
  action, 377
  esterification, 392
  infrared spectrum, 387
  preparation, 383
p-Aminophenol, 60
  acetylation, 60
5-Aminophthalhydrazide
  preparation, 455
Analgesic drugs
  isolation, 42
Analgesics, 46, 53, 79
  composition, 56
  essay, 53
  thin-layer chromatography, 79
  Anesthetics
  essay, 388
  preparation, 392
  structures, 391
Anhydrides
  infrared spectroscopy, 910
Anilide derivatives, 517
  preparation, 889
Aniline
  acetylation, 57
Anisaldehyde
  aldol condensation, 409
Anisole
  infrared spectrum, 905
Anisotropy in NMR spectroscopy, 918
Anthranilic acid
  benzyne formation, 436
Anti-inflammatory, 46
Antipyretic, 46, 53
APC tablets, 54, 66
Apparatus
  assembly, 35
  identification key, 39
  securing with joint clips, 617
Aromatic compounds
  C-H out-of-plane bending, 902
  Friedel-Crafts acylation, 306
  identification tests, 507
Aromatic compounds (*continued*)
  infrared spectroscopy, 507, 901
  NMR spectroscopy, 508, 879, 884
Aromatic halides
  carbonation, 288
Aromatic rings
  spectroscopy, 507
Aromatic substitution
  acetanilide, 301
  acylation, 306
  aniline, 301
  anisole, 301
  chlorosulfonation, 378
  diazo coupling, 317, 318
  methyl benzoate, 297
  nitration, 297
  relative reactivities, 301
Aromaticity
  detection, 507
Aspartame, 535
  structure, 546
Aspirator, 643
Aspirator trap, 641, 644, 721, 732
Aspirin, 46, 49, 56, 79
  buffered, 48
  combination tablets, 48
  essay, 46
  isolation, 42
  preparation, 49
  structure, 42, 47
  tablets, 48
Automatic pipet, 597
Azeotropes, 745
  composition, 747, 748
  definition, 745
  maximum boiling point, 748
  minimum boiling point, 745
Azeotropic distillation
  applications, 749
  microscale apparatus, 751

B

Baeyer test, 505
Balance
  analytical, 600
  draft shield, 599
  top-loading, 598
Banana oil
  preparation, 96
Barfoed's test, 543
Bases
  removal by extraction, 700
Basic hydrolysis of esters test, 530
Beilstein, 868
Beilstein test, 497
Benedict's test, 542, 556, 559
Benzalacetophenone
  preparation, 409
Benzaldehyde
  aldol condensation, 409
  benzoin condensation, 365
  infrared spectrum, 368
Benzamide
  infrared spectrum, 910
Benzamides, 524
  preparation, 890
Benzene
  hazards, 19
Benzil
  aldol condensation, 395
  infrared spectrum, 371
  preparation, 369
  rearrangement, 371
Benzilic acid
  infrared spectrum, 373
  preparation, 371
Benzilic acid rearrangement, 371
Benzocaine
  infrared spectrum, 394
  NMR spectrum, 394
  preparation, 392
Benzoic acid
  infrared spectrum, 287
  preparation, 276
  structure, 546
Benzoin
  coenzyme synthesis, 365
  infrared spectrum, 368
  oxidation, 369
  preparation, 365
Benzoin condensation, 365
Benzonitrile
  infrared spectrum, 907
Benzophenone
  photoreduction, 442
  ultraviolet spectrum, 447
β-Benzopinacolone
  preparation, 450
Benzpinacol
  preparation, 442
  rearrangement, 450
Benzyl acetate
  NMR spectrum, 875

N-Benzylamides, 532
Benzyltriphenylphosphonium chloride
  preparation, 460
Benzyne
  preparation, 436
Bial test, 540
Bio-Gel P, 784
Biodegradable, 129
Bioluminescence, 452
Biphenyl, 279
Boileezer, 622
Boiling chip, 622
Boiling points, 674
  construction of microcapillary bell, 678
  determination, 675
  effect of pressure on, 674
  microscale determination, 677
  pressure correction, 675
  pressure-temperature nomograph, 676
  standard scale determination, 675
Boiling stones, 622
Borneol
  carbon-13 NMR spectrum, 186
  infrared spectrum, 183
  NMR spectrum, 185
  oxidation, 176
Borneol and isoborneol
  NMR spectrum of mixture, 187
Bottles
  labels, 13
Breathalyzer, 230, 233
Bromination
  acetanilide, 301
  aniline, 301
  anisole, 301
  phenols, 518, 520
  unsaturated compounds, 251, 504
Bromine in carbon tetrachloride test, 251, 504
Bromine in water test, 518
Bromo derivatives of phenols, 520
  preparation, 890
Bromobenzene
  Grignard reaction, 276
Bromocyclohexane
  carbon-13 NMR spectrum, 895
Bubbler trap, 124
Büchner funnel, 38, 641
Buffered aspirin, 48
Bumping, 622
Burns, 10
Butane
  conformations, 424
1-Butanol
  nucleophilic substitution, 204, 211
2-Butanol
  nucleophilic substitution, 211
2-Butene
  cis and trans isomers, 426
tert-Butyl chloride and *tert*-butyl bromide
  NMR spectrum of mixture, 220
n-Butyl alcohol
  nucleophilic substitution, 204, 211
sec-Butyl alcohol
  nucleophilic substitution, 211
tert-Butyl alcohol,
  nucleophilic substitution, 211
n-Butyl bromide
  infrared spectrum, 209
  preparation, 204
n-Butylamine
  infrared spectrum, 905

C

C-4 and C-5 alcohols
  esterification, 100
Caffeine, 54
  essay, 63
  isolation from a tea bag, 74
  isolation from Mountain Dew syrup, 76
  isolation from tea, 68
  isolation from tea leaves, 71
  structure, 546
Camphor
  carbon-13 NMR spectrum, 186
  infrared spectrum, 183
  NMR spectrum, 185
Carbene addition
  cyclohexene, 253
Carbocations
  rearrangement, 291
Carbohydrates
  identification, 536
Carbon disulfide
  infrared spectrum, 853
Carbon tetrachloride
  hazards, 19
  infrared spectrum, 850
Carbon-13 NMR spectroscopy, 306, 308
  chemical shifts, 933
  correlation table, 894
  coupled spectra, 892
  coupling constants, 932
  decoupled spectra, 932
  effect of symmetry, 898
  index of spectra, 939
  nuclear Overhauser effect, 938
  sample preparation, 859
Carbonation
  aromatic halides, 288
Carbonyl compounds
  infrared spectroscopy, 513, 516, 531, 906
Carboxylic acids
  derivatives, 516, 880, 889
  identification tests, 514
  infrared spectroscopy, 516, 908
  NMR spectroscopy, 516, 886
  preparation, 288
  spectroscopy, 516
  tables of unknowns, 880
  titration, 516
Carcinogen
  definition, 7
  list, 21
Carcinogenic substances, 21
β-Carotene
  chromatography, 352
  isolation, 352
  structure, 350
Carotenoids, 350
Carrier gas, 808, 812
Carvone
  carbon-13 NMR spectrum, 199
  gas chromatography, 192
  infrared spectrum, 196
  isolation, 192
  NMR spectrum, 197
  structure, 192

CAS Number, 11
Casein, 551
  micelle, 552
Celite, 642
Centrifugation, 648
  balancing, 647
  centrifuge tube, 648
  Craig tube, 647
Centrifuge tube, 38
  use in extraction, 696
Chalcones, 409
Channeling, 771
Charcoal
  activated, 660
Chemical Abstracts, 868
Chemical equivalence, 914
Chemical literature, 861
  advanced textbooks, 863
  Beilstein, 868
  CA Online, 871
  Chemical Abstracts, 868
  collections of spectra, 863
  computer online searching, 871
  handbooks, 861
  qualitative organic analysis, 867
  reaction mechanisms, 866
  Science Citation Index, 873
  scientific journals, 871
  search methods, 862, 873
  synthetic methods, 862, 864
Chemical shift, 911
Chemical shift reagents, 888
Chemiluminescence
  definition, 452
  experiment (luminol), 455
  tests with luminol, 459
Chemotherapy, 375
Chiral reduction, 570
Chiral resolving agent, 568
Chiral shift reagent, 574
Chlordane, 428
1-Chlorobutane and 1-bromobutane
  NMR spectrum of mixture, 220
Chloroform
  hazards, 19
  infrared spectrum, 851
Chloroform-d, 856
Chlorophyll
  chromatography, 352
Chlorophyll (*continued*)
  isolation, 352
  structure, 352
Chlorosulfonation
  acetanilide, 378
Chlorosulfonic acid, 135
Cholesterol, 114
Chromatic aberration, 840
Chromatography
  gas, 808
  gel, 783
  high performance liquid chromatography, 787
  HPLC, 787
  HPLC apparatus, 788
  liquid, 787
  paper, 806
  reversed phase, 789
  thin-layer, 792
  See also Column, Gas, Paper, HPLC & Thin Layer, chromatographies
Chromatography Column
  microscale, 38
Chromic acid
  oxidation of borneol, 173
Chromic acid test, 510, 527
Cinnamaldehyde
  Wittig reaction, 460
Claisen head, 37, 719
Clamp holder, 39
Clamps, 39
  adjustable, 615
  extension, 615
  three-finger, 615
Classification tests
  See specific functional group
Cleaning solution
  preparation, 32
  safety precautions, 32
CMR spectroscopy, 891, 931
  See also Carbon-13 NMR spectroscopy
Cocaine, 388
Coenzyme, 365
Coffee
  caffeine content, 65
  decaffeination, 66
  discovery, 63
Cold bath, 611
  dry ice, 611
  ice-salt, 611
  liquid nitrogen, 611
Cold-finger condenser, 828
Column
  chromatographic, 764
  chromatographic, microscale, 776
  fractionating, 37, 741
  spinning-band, 741
  Vigreux, 741
Column chromatography, 761
  adsorbent quantity, 769
  adsorbents, 761, 766, 767
  adsorption equilibrium, 763
  apparatus, 764
  column size, 769
  depositing the adsorbent, 774, 776, 777
  dye mixtures, 342
  elution sequence for compounds, 768
  elution techniques, 778
  flash chromatography, 784
  flow rate, 770
  interactions, 762
  monitoring, 780, 805
  packing the column, 771
  preparing the support base, 772, 775
  principles, 763
  reservoirs, 779
  sample application, 777
  semi-microscale column, 773
  semi-microscale methods, 771
  sequence of steps, 765
  solvents, 767, 768
  spinach, 352
  standard scale methods, 771
  typical problems, 770
  use in decolorization, 782
Combination pain relievers, 48
Computational chemistry, 418
Computer interfaced methods, 227, 244
  oxidation of alcohols, 244
  use in kinetics, 227, 244
Condensation reactions,
  aldol, 395
  benzoin, 365
  enamine, 398
  luminol preparation, 455
  Wittig reaction, 460
Condenser, 37
  cold-finger, 828
  water, 37, 619, 712

Conical vials
  use in extraction, 693
Conjugate addition, 413
Cooling methods, 600
Correlation chart
  NMR spectroscopy, 917
Correlation table
  carbon-13 NMR spectroscopy, 934
  infrared spectroscopy, 897
Coupling constant, 923
Cracking, 152
Craig tube, 646, 658
  centrifugation, 647
Crown ethers, 256
Crystallization, 648
  common solvents, 659
  Craig tube, 658
  Hirsch funnel, 656
  inducing crystal formation, 661
  microscale, 658
  mixed solvents, 663
  removing insoluble impurities, 653
  scratching to induce crystallization, 661
  seeding, 661
  solvent pairs, 663
  solvent selection, 649, 658
  standard scale methods, 651
  stepwise methods, 657
  theory, 650
Crystals
  drying methods, 662
Cyano groups
  detection, 500
  spectroscopy, 500
Cycloaddition
  See Diels-Alder reaction
Cyclohexane
  chair and boat conformations, 425
  substituted rings, 426
Cyclohexanol
  carbon-13 NMR spectrum, 935
Cyclohexanone
  acetylation, 398
  carbon-13 NMR spectrum, 937
  enamine reaction, 398
Cyclohexene
  carbene addition, 253
  carbon-13 NMR spectrum, 936
Cyclopentadiene
  preparation, 432

D

DDT, 428
Dean-Stark water separator, 749
Decaffeinated coffee, 66
Decaffeination, 66
Decane
  infrared spectrum, 900
Decolorization, 659
  by column chromatography, 782
  using a column, 660
Decomposition point, 672
Dehydration
  4-methylcyclohexanol, 248
Density
  determination, 683
Derivative formation
  for unknowns, 888
  osazones from sugars, 543
Derivatives
  methods of preparation, 888
  tables of, 878
  See also specific functional groups
Desiccator, 663
Detection of alcohol
  infrared spectroscopic method, 233
Detergents
  essay, 127
  preparation, 135
  problems, 132
  test, 138
Deuteriochloroform, 856
Deuterium oxide, 858
Deuterochloroform
  hazards, 19
Dextrorotatory, 834
Diatomaceous earth, 642
Diazonium coupling, 317, 318
Diazotization, 317
  nitrous acid test for amines, 521
  sulfanilic acid, 318
Dibenzyl ketone
  aldol condensation, 395
Diborane, 263, 266
1,2-Dichlorobenzene
  carbon-13 NMR spectrum, 898
  infrared spectrum, 903
1,3-Dichlorobenzene
  carbon-13 NMR spectrum, 938
Dichlorocarbene, 253
Dichloromethane
  See Methylene Chloride
7,7-Dichloronorcarane
  carbon-13 NMR spectrum, 260
  infrared spectrum, 259
  NMR spectrum, 259
  preparation, 253
Dictionary of Organic Compounds, 861
Dicyclopentadiene
  cracking, 432
  structure, 433
Dieldrin, 428
Diels-Alder reaction, 432, 436
  benzyne, 436
  cyclopentadiene, 432
  essay, 427
  maleic anhydride, 432
  tetraphenylcyclopentadienone, 436
Diet soft drinks
  HPLC analysis, 546
Diethyl ether, 613
  hazards, 19
N,N-Diethyl-m-toluamide,
  action, 143
  infrared spectrum, 150
  preparation, 145
1,2-Dimethoxyethane
  hazards, 19
2,2-Dimethylbutane
  carbon-13 NMR spectrum, 935
3,5-Dinitrobenzoates, 520, 528, 532
  preparation, 891
2,4-Dinitrophenylhydrazine test, 509
2,4-Dinitrophenylhydrazones, 514
  preparation, 889
Dioxane
  hazards, 19

1,4-Diphenyl-1,3-butadiene
  preparation, 460
Disparlure, 141
Dispensing pumps, 590, 591
Disposable pipets
  See Pasteur pipets
Distillation
  bulb-to-bulb, 726
  Hickman head, 715
  phase diagrams, 709, 735, 745, 747
  reduced pressure, 717
  removing fractions, 713
  simple, 706
  steam, 753
  theory, 708, 735
  vacuum, 717
  See also Fractional, Simple, Steam, and Vacuum Distillations
Distillation Head, 37, 711
  Claisen, 37
  Hickman, 715
Distribution coefficient, 687
Diuretic, 64
DNA
  structure, 67
Double bonds
  detection, 503
  spectroscopy, 506
Downfield, 913
Drugs
  analysis, 79
  identification, 77, 79
  thin-layer chromatography, 79
Dry ice, 611
Dry solutions
  common signs, 698
Drying agents, 696
  table, 698
Drying tubes, 39, 624
DSS, 860
Dye mixtures
  chromatography, 342
Dyes
  azo, 316, 317, 318, 339
  essay, 313
  FD&C, 338, 340
  food, 337
  methods of dyeing, 330
  natural, 313
  synthetic, 313, 316, 323

Dyes (*continued*)
  triphenylmethane, 316
  types, 325

E

E1/E2 Reactions
  See Elimination Reactions
Ebulliator, 719
Ebulliator tube, 37
Electrophilic aromatic substitution
  See Aromatic substitution
Elemental analysis, 496
Elimination reactions
  4-methylcyclohexanol (E1), 248
Eluates, 764
Eluents, 764
Elutants, 764
Emulsions, 699
Enamine
  preparation, 403
  cyclohexanone, 398
  theory, 399
Enantiomeric excess, 835
Energy transfer, 447
Enol content determination, 406
Essential oils, 165
  allspice oil, 169
  caraway oil, 169, 192
  cinnamon oil, 169
  clove oil, 169
  cumin oil, 169
  fennel oil, 169
  isolation, 169
  spearmint oil, 192
Esterification
  p-aminobenzoic acid, 392
  C-4 and C-5 alcohols, 100
  isopentyl alcohol, 96
  salicylic acid, 49, 103
  vanillin, 579
Esters
  derivatives, 532
  essay, 93
  hydrolysis, 108, 530
  identification tests, 529
  infrared spectroscopy, 531, 909
  NMR spectroscopy, 532
  preparation, 116
  spectroscopy, 531
  tables of unknowns, 887

Ethanol, 155
  95 percent, 87
  denatured, 87
  essay, 85
  hazards, 19
  preparation, 88
Ether
  See Diethyl Ether
Ethers
  infrared spectroscopy, 904
6-Ethoxycarbonyl-3,5-diphenyl-2-cyclohexenone
  infrared spectrum, 417
  preparation, 413
Ethyl 3-hydroxybutanoate
  NMR spectrum, 575
(S)-(+)-Ethyl 3-hydroxybutanoate
  preparation, 570
  optical purity determination, 574
Ethyl acetoacetate
  chiral reduction, 570
  condensation reaction, 413
Ethyl *p*-aminobenzoate
  See Benzocaine
Ethyl phenylacetate
  carbon-13 NMR spectrum, 931
Ethylene glycol dimethyl ether, 19
Eutectic, 666
Evaporation to dryness, 630
Extraction, 685
  apparatus, 689
  choosing a method, 688
  determining organic layer, 693
  liquid-liquid, 703
  solid-liquid, 702
  solvent selection, 688
  theory, 687
  use in purification, 700
  use of centrifuge tube, 696
  use of conical vial, 693
  use of separatory funnel, 101, 690
Eye safety, 5

F

Fabric types, 324
Fabrics
  dyeing, 323, 330
Fat replacers, 114

Fats
  fatty acid composition, 113
  synthetic, 114
  unsaturated, 112
Fats and oils
  analysis, 116
  composition, 113
  essay, 110
  saponification, 133
Fatty acids, 111
  isomerization, 114
  molecular modeling, 581
  saturated, 581
  unsaturated, 581
  trans-unsaturated, 114, 581
trans-Fatty acids, 114, 581
Fermentation
  apparatus, 573
  essay, 85
  ethyl acetoacetate, 570
  sucrose, 85, 88
Ferric chloride test, 50, 51, 513, 518
Ferric hydroxamate test, 530
Ferrous hydroxide test, 499
Fiber types, 324
Fibers
  dyeing, 330
Filter Aid, 642
Filter flask, 38, 640
Filter paper
  cones, 635
  fluted, 636
  porosity, 639
  retentivity, 640
  speed, 639
  types, 640
Filter tip pipet, 596, 645
Filtering pipet, 591, 638, 654
Filtration
  Craig tube, 646
  filtering pipet, 638
  fluted filter, 654
  gravity, 634
  in crystallization, 653
  methods, 634
  suction, 640
  vacuum, 640
Fireflies
  essay, 452
Fires, 6
First Aid, 10
  acid burns, 10
First Aid (*continued*)
  base burns, 10
  burns, 10
  cuts, 10
Flames, 8
  heating with, 609
Flash chromatography, 784
Flask
  three-necked, 37
Flavors
  artificial, 93, 95
  composition, 93
  essay, 93
  synthetic, 93, 95
Fluorescence, 444
Fluted filter paper, 636, 654
Folic acid, 377
Food colors
  approved list, 340
  chromatography, 342
  essay, 337
Force field, 418
Forerun, 713
Fraction collectors, 725
Fractional distillation, 732
  columns, 741
  methods, 743
  microscale apparatus, 744
  spinning-band column, 741
  standard scale apparatus, 734
  See also Distillation
Fractionating columns, 37, 741, 744
  types of packing, 741
Fractions, 713, 780
Fragrances
  composition, 93
  essay, 93
Friedel-Crafts reaction, 306
Frozen joints, 34
Fructose, 89
Funnels
  Büchner, 38
  methods of preheating, 654

G

Gallic acid, 69
Gas chromatograph, 809
Gas chromatography, 808
  alkyl halides, 211
  apparatus, 809
  calculation of peak areas, 820
Gas chromatography (*continued*)
  carrier gas, 808, 812
  carvone, 192
  collection trap, 819
  columns, 810
  detector, 814
  fats and oils, 116
  flame ionization detector, 814
  gasoline, 159
  liquid phase, 810, 811
  1-methyl- and 2-methylcyclohexanols, 273
  monitoring, 814
  peak areas, 820
  principles, 812
  qualitative use, 818
  quantitative analysis, 820
  resolution, 817
  response factors, 822
  retention time, 817
  sample collection, 818
  sample collection tube, 819
  stationary phase, 808, 810
  support material, 810, 812
  thermal conductivity detector, 814
Gas chromatography collection tube, 819
Gas trap, 625
Gases,
  collection, 629
  traps for noxious gases, 625
Gasoline
  composition, 151
  gas chromatography, 159
  oxygenated, 155
  sample gas chromatograms, 163
Gel chromatography, 783
Glassware, 31
  assembly, 33
  care, 32
  cleaning, 32
  drying, 33
  etching, 35
  frozen joints, 34
  identification key, 37, 38, 39
  illustrations, 37, 38, 39
  lubrication, 34
Gloves
  protective, 7
Glucose, 87, 89
Glyceride, 111

- Gossyplure, 141
- Gradient elution, 791
- Graduated cylinders, 589
- Gravity filtration, 634
- Grignard reaction, 276
  - apparatus, 281
  - experiments, 276
  - starting, 281
- Grignard reactions, 288
- Ground state, 444
- Ground-glass joints, 33, 614
- Gyplure, 140

H

- Half-life, 224, 239
- Halides
  - detection, 496
  - elemental analysis, 502
  - infrared spectroscopy, 910
  - *See also* Alkyl Halides
- Handbook of Chemistry and Physics, 861
- Hazards
  - solvents, 6
- Heating mantles, 601
- Heating methods, 600
  - aluminum block, 605
  - Bunsen burner, 609
  - evaporation to dryness, 630
  - flames, 609
  - heating mantles, 601
  - hot plates, 605
  - oil baths, 604
  - reflux, 619
  - sand bath, 608
  - solvents, 612
  - steam bath, 610
  - water baths, 603
- Heating under reflux, 602
- Heme
  - structure, 55
- Hemoglobin, 55
- HETP, 743
- Hexane
  - hazards, 20
- 1-Hexanol
  - NMR spectrum, 929
  - NMR spectrum with shift reagent, 930
- Hickman head, 715
  - removing fractions, 715
- High performance liquid chromatography, 787
  - adsorbents, 788
  - apparatus, 788
  - columns, 788
  - detectors, 791
  - diet soft drinks, 546
  - gradient elution, 791
  - ion-exchange chromatography, 789
  - isochratic, 791
  - normal-phase chromatography, 789
  - reversed-phase chromatography, 789
  - size exclusion chromatography, 789
  - solvents, 790
- High pressure liquid chromatography, 787
- Hinsberg test, 523
- Hirsch funnel, 38, 641, 656
- Holdup, 711, 741
- Hormones
  - insect, 431
- Hot plate/stirrer, 39
- Hot plates, 602, 603, 604, 608
- HPLC, 787
  - *See also* High Performance Liquid Chromatography
- HPLC apparatus, 788
- Hydration
  - alkenes, 261
  - anti-Markovnikov, 261
  - Markovnikov, 261
- Hydrazides, 532
- Hydroboration-oxidation
  - alkenes, 261
- Hydrogenation
  - apparatus, 124
  - methyl oleate, 121
  - oils, 112
- Hydrolysis
  - amides, 378
  - ester unknowns, 530
  - methyl salicylate, 108
- Hydrolysis of alkyl chlorides
  - kinetics by computer interface method, 227
  - kinetics by titration, 225
- Hydroxamic acid test, 530
- Hygroscopic, 662

I

- Ibuprofen, 55, 79
  - isolation, 42
  - structure, 42
- Ice bath, 611
- Ice-salt bath, 611
- Ideal solution, 735, 737
- Identification of unknowns, 485
- Ignition test, 507
- Index of refraction, 836
- Index of spectra, 940
- Infrared spectroscopy, 892
  - alcohols, 528, 904
  - aldehydes, 513, 906
  - alkanes, 900
  - alkene C-H out-of-plane bending, 901
  - alkenes, 506, 900
  - alkynes, 506, 903
  - amides, 910
  - amines, 523, 905
  - anhydrides, 910
  - aromatic C-H out-of-plane bending, 902
  - aromatic compounds, 507, 901
  - base values, 898
  - carbonyl compounds, 513, 516, 531, 906
  - characterization of polymers, 483
  - correlation table, 897
  - effects of conjugation, 907, 908, 909
  - effects of ring size, 908
  - esters, 531, 909
  - ethers, 904
  - halides, 910
  - index of spectra, 900
  - interpretation, 894–899
  - KBr hand press, 847
  - KBr minipress, 848
  - KBr pellets, 846
  - ketones, 513, 907
  - liquid samples, 843
  - modes of vibration, 894
  - neat liquids, 843
  - nitriles, 500, 906
  - nitro compounds, 500, 906
  - Nujol mulls, 853
  - phenols, 519, 904
  - salt plates, 843
  - sample preparation, 843, 845, 846, 847, 848, 853

Infrared spectroscopy (*continued*)
  silver chloride cells, 845
  solution cell, 852
  solution spectra, 850
  solution spectra, silver chloride mini-cell, 851
  solvent spectra, 850-854
  spectrum calibration, 854
  survey of functional groups, 900–910
  theory, 892
  uses, 893
Injection port, 812
Insect attractants, 139
Insect repellents, 139, 143
  preparation, 145
Insecticides
  essay, 427
Internal conversion, 447
Intersystem crossing, 444
Iodoform test, 512, 528
"Iso-octane," 153
Isoamyl acetate
  See Isopentyl acetate
Isoborneol
  carbon-13 NMR spectrum, 187
  infrared spectrum, 183
  NMR spectrum, 185
  preparation, 176
Isochratic, 791
Isolation experiment, 25
Isolation experiment
  β-carotene from spinach, 352
  caffeine from tea, 68
  carvone from caraway and spearmint oils, 192
  casein from milk, 557
  chlorophyll from spinach, 352
  essential oils from spices, 169
  lactose from milk, 558
Isopentyl acetate
  infrared spectrum, 99
  preparation, 96
  esterification, 96
Isoprene rule, 167

J
Joint clips, 616
Journals, 871
Juvenile hormone, 431

K
KBr pellet
  hand press, 847
  minipress, 848
KBr pellet preparation, 846
Keto-enol tautomerism, 398
Ketones
  derivatives, 514, 879, 888
  identification tests, 508
  infrared spectroscopy, 513, 907
  NMR spectroscopy, 514, 920
  spectroscopy, 513
  tables of unknowns, 879
Ketoprofen, 56, 79
Kinetics
  alcohol oxidation, 234
  alkyl chloride hydrolysis, 221
  determination by computer-interfaced method, 227
  determination by titration, 225
Knocking, 154

L
Labels
  commercial bottles, 13
  sample, 31
Laboratory notebook, 23
Laboratory records, 25
Lactose
  isolation from milk, 555
  mutarotation, 560
  structure, 553, 555
Lange's Handbook of Chemistry, 861
Lanthanide shift reagent, 929
  structure, 577
LD50, 12
Lethal Dose, 12
Levorotatory, 834
Ligroin, 613
  hazards, 20
Limiting reagent, 27
Limonene
  infrared spectrum, 196
  NMR spectrum, 197
Liquid chromatography, 787
Liquids
  addition of reagents, 622
  boiling point determination, 674
  density determination, 683
  measurement, 588
Literature of chemistry, 861
Lucas test, 526
Luciferin, 452
Luminol
  chemiluminescence tests, 459
  preparation, 455

M
Macroscale
  definition, 4
Magnetic equivalence, 924
Magnetic spin bar, 39
Magnetic spin vane, 608, 621
Magnetic stirrers, 608, 621
Malathion, 430
Maltose, 86
Manometer, 729
  bleed valve, 731
  closed-end, 729
  connection, 731
  construction, 729
  filling, 731
  trap, 732
Markovnikov's rule, 262
Material Safety Data Sheets, 11
Measurement
  liquids, 588
Melting point, 666
  behavior, 666
  capillary, 669
  composition diagram, 666
  corrections, 680
  decomposition, 672
  depression, 666
  determination, 669, 670
  discoloration, 672
  electrical apparatus, 670
  mixed, 668
  packing tubes, 669
  phase diagram, 667
  range, 666
  sealed tubes, 673
  shrinkage of crystals, 672
  softening, 672
  standards, 681
  sublimation, 672
  theory, 667
Melting point tube, 669
  sealing methods, 673
Merck Index, 861
Mesityl oxide
  infrared spectrum, 909

- Methanol
  - hazards, 20
  - MSDS sheet, 14
- Methemoglobinemia, 54
- Methiodides, 524
  - preparation, 891
- *trans*-2-Methylcyclohexanol
  - NMR spectrum, 274
- 2-Methyl-2-propanol
  - nucleophilic substitution, 211
- Methyl benzoate
  - infrared spectrum, 909
  - nitration, 297
- Methyl isopropyl ketone
  - infrared spectrum, 895
- Methyl m-nitrobenzoate
  - infrared spectrum, 300
  - preparation, 297
- Methyl oleate
  - hydrogenation, 121
- Methyl orange
  - preparation, 318
- Methyl salicylate, 108
  - hydrolysis, 108
  - infrared spectrum, 106
  - NMR spectrum, 107
  - preparation, 103
- Methyl stearate
  - preparation, 121
- Methyl tert-butyl ether, 155
- 3-Methyl-1-butanol
  - *esterification*, 96
- 1-Methylcyclohexanol
  - NMR spectrum, 275
  - preparation, 270
- 2-Methylcyclohexanol
  - preparation, 266
- 4-Methylcyclohexanol
  - dehydration, 248
  - infrared spectrum, 252
- 1-Methylcyclohexene
  - hydration, 261
  - hydroboration-oxidation, 261
  - oxymercuration, 261
- 4-Methylcyclohexene
  - bromine addition, 251
  - infrared spectrum, 251
  - preparation, 248
- Methylene chloride
  - hazards, 20
- Micelle, 130
  - casein, 551
  - soap, 130
- Michael addition, 413
- Micro boiling point determination, 677
  - bell preparation, 678
- Microchromatographic Column, 38
- Micropipet, 797
- Microscale
  - definition, 5
- Milk
  - essay, 549
  - isolation of casein, 555
  - isolation of lactose, 555
- Mineral oil
  - infrared spectrum, 854
- Minimization, 418, 420
- Mixed solvents, 663
- Mixtures
  - separation, 700
- Mole fraction, 736
- Mole percentage, 736
- Molecular mechanics, 418
  - essay, 417
  - limitations, 421
- Molecular modeling
  - essay, 417
  - experiment, 423
  - fatty acids, 581
- Molecular rotation, 832
- Molecular sieve chromatography, 784
- *Molisch test*, 539
- Monoglyme, 19
- Mordants, 314
- Mother liquor, 650
- MSDS, 11
  - sample pages, 14
- Mucic acid test, 545, 559
- Mutarotation, 556, 560

N

- n + 1 rule, 920
- *N,N*-Dimethylaniline
  - diazo coupling, 318
- Naphthalene
  - ultraviolet spectrum, 447
- 2-Naphthol
  - infrared spectrum, 904
- α-Naphthylurethanes, 520
  - preparation, 889
- Naproxen, 56, 79
- Natural dyes, 315
- Neat liquids, 843
- Neoprene adapter, 38
- NFPA Rating, 11
- Nitration
  - methyl benzoate, 297
- Nitriles
  - infrared spectroscopy, 500, 906
- Nitro compounds
  - infrared spectroscopy, 500, 906
- Nitro groups
  - detection, 499
  - spectroscopy, 500
- 3-Nitrobenzaldehyde
  - aldol condensation, 409
- Nitrobenzene
  - infrared spectrum, 906
- Nitrogen
  - elemental analysis, 502
  - liquid, 611
- 5-Nitrophthalhydrazide
  - preparation, 455
  - reduction, 455
- 3-Nitrophthalic acid
  - amide formation, 455
- Nitrous acid test, 521
- NMR spectroscopy, 911
  - alcohols, 528, 927
  - aldehydes, 514, 880
  - alkenes, 506, 883, 921
  - alkynes, 506, 920
  - amines, 523, 886
  - anisotropy, 879
  - aromatic compounds, 508, 918, 925
  - aromatic rings, 923
  - carbon-13, 306, 308, 891
  - carboxylic acids, 516, 927
  - chemical equivalence, 874
  - chemical shift, 913
  - chemical shift reagents, 928
  - common splitting patterns, 923
  - correlation chart, 918
  - coupling constant, 921
  - esters, 532
  - high-field NMR, 927
  - index of spectra, 940
  - integrals, 914
  - ketones, 514, 880
  - magnetic equivalence, 924
  - n + 1 rule, 880
  - optical purity determination, 574

- NMR spectroscopy (*continued*)
  - phenols, 519, 886
  - protons on heteroatoms, 886
  - quantitative use, 219
  - reference substances, 859
  - ring current, 918
  - sample preparation, 855
  - sample tube, 855, 857
  - shielding, 916
  - shift reagents, 888
  - solvents, 855-859
  - spin-spin splitting, 920
  - structure proof, 579
  - theory, 871
  - use of shift reagents, 574
- Nonanal
  - infrared spectrum, 908
- Nonideal solutions, 745
- *cis-Norbornene-5,6-endo-dicarboxylic anhydride* infrared spectrum, 435
  - preparation, 432
- Norit, 660
  - pelletized, 660
- Notebook, 23
  - advance preparation, 24
  - calculations, 27
  - format, 24
  - laboratory records, 25
  - sample pages, 29, 30
- Nuclear magnetic resonance
  - *See* NMR Spectroscopy
- Nuclear Overhauser effect, 938
- Nucleophilic substitution
  - competing nucleophiles, 211
  - kinetic study, 221
  - *n*-butyl alcohol (SN2), 204
  - *tert*-pentyl chloride (SN1), 204
  - preparation of alkyl halides, 204, 225
  - reactivities of alkyl halides, 200
  - tests for reactivities, 200
- Nujol
  - infrared spectrum, 854
  - mull, 853
- NutraSweet, 535
- Nylon
  - preparation, 477

O

- Observed rotation, 832
- Octadecanoic acid
  - conformations, 585
- (E)-Octadecenoic acid
  - conformations, 583
- (Z)-Octadecenoic acid
  - conformations, 583
- Octane ratings, 156
- Odor
  - stereochemical theory, 189
- Oil baths, 604
- Oiling out, 658
- Oils
  - fatty acid composition, 113
  - vegetable, 112
- Olean, 114
- Oleic acid
  - conformations, 583
- *trans*-Oleic acid
  - conformations, 583
- Olestra, 114
- Optical activity, 831, 832
- Optical purity, 835
  - NMR determination, 574
- Organic Reactions, 865
- Organic Syntheses, 865
- Origami, 637
- Osazone formation, 543
- Oven drying, 662
- Oxidation
  - alcohols, 234, 527
  - aldehydes, 510
  - alkenes, 505
  - benzoin, 369
  - borneol, 176
  - by chromic acid, 176
  - by nitric acid, 369
  - by potassium permanganate, 505
  - by sodium dichromate, 234, 510, 527
  - permanganate, 383
- Oxidation of alcohols
  - kinetics by computer-interfaced methods, 244
  - kinetics by ultraviolet-visible spectroscopy, 242
  - kinetics by visible spectroscopy, 241
- Oximes, 514
- Oxymercuration,
  - alkenes, 261

P

- PABA
  - *See p*-Aminobenzoic Acid
- Palladium on charcoal, 121
- Paper chromatography, 806
  - food colors, 342
  - powdered drink mixes, 342
- Paraffin oil
  - infrared spectrum, 854
- Parathion, 430
- Partial vapor pressure, 738
- Partition coefficient, 687
- Pasteur pipet, 38, 596
  - calibrated, 596
- PEL, 12
- Penicillin, 376
- Pentane
  - hazards, 20
- *tert*-Pentyl alcohol
  - nucleophilic substitution, 204
- *tert*-Pentyl chloride
  - infrared spectrum, 210
  - preparation, 204
- Percentage yield, 27
- Petroleum
  - essay, 151
- Petroleum ether, 613
  - hazards, 20
- Phase-transfer catalysis, 253
- Phenacetin, 53, 54
  - structure, 63
- Phenols
  - derivatives, 519, 883, 889
  - identification tests, 517
  - infrared spectroscopy, 519, 904
  - NMR spectroscopy, 519, 927
  - spectroscopy, 519
  - tables of unknowns, 883
- Phenylacetone
  - NMR spectrum, 873
- α-Phenylethylamine
  - optical purity determination, 568
  - resolution of enantiomers, 563
- Phenylmagnesium bromide, 276, 280
  - preparation, 276
- Phenylpropanoids
  - essay, 165
- Phenylurethanes, 529
  - preparation, 891
- Pheromones, 95, 139
  - alarm, 142
  - essay, 139
  - primer, 139, 141

- Pheromones (*continued*)
  - recognition, 143
  - recruiting, 143
  - releaser, 139
  - trail, 143
- Phosphorescence, 445
- Photochemistry, 442, 452, 455
- Photoreduction
  - benzophenone, 442
- Picrates, 524
  - preparation, 890
- Pinacol
  - preparation, 291
- Pinacol rearrangement, 291, 450
- Pinacolone
  - preparation, 291
- Piperonaldehyde
  - aldol condensation, 409
- Pipet
  - automatic, 597
  - filter tip, 645
  - filtering, 591, 654
  - graduated, 38, 591
  - Pasteur, 38, 596
- Pipet pump, 592
- Plastic joint clips, 616
- Plasticizer, 469
- Plastics
  - essay, 467
  - recycling codes, 474
- Polarimeter, 832
  - operation, 834
- Polarimeter cell, 833
- Polarimetry, 829
  - carvone, 196
  - (S)-(+)-ethyl 3-hydroxybutanoate, 570
  - mutarotation of lactose, 560
  - resolution of α-phenylethylamine, 563
- Polarized light, 829
- Pollution
  - petroleum, 157
- Poly-Sep AA, 784
- Polyamide
  - preparation, 477
- Polyester
  - preparation, 478
- Polymerization, 152
- Polymers
  - addition, 467, 469, 471
  - condensation, 467, 468, 470, 473
- Polymers (*continued*)
  - cross-linked, 468
  - disposability problems, 474
  - essay, 467
  - infrared spectroscopy, 483
  - preparation, 477
  - recycling codes, 474
  - types, 467
- Polystyrene
  - infrared spectrum, 855
  - preparation, 481
- Polyunsaturated fats, 112
- Potassium permanganate test, 251, 505
- Pregnancy
  - precautions, 7
- Preparative experiment, 24
- Product development control, 180
- Product purification,
  - by extraction, 700
- Propargyl alcohol
  - infrared spectrum, 903
- Prostaglandins, 47
- Protective groups, 379, 383
- Purification of solids, 648
- Pyrethrins, 431
- Pyridine
  - hazards, 20

Q

- Qualitative organic analysis, 485, 867
- Queen substance, 141
- Quenching, 447

R

- Racemic mixture, 835
- Radiationless transition, 445
- Raoult's Law, 735
  - immiscible liquids, 754
  - miscible liquids, 737, 753
- Rate constant
  - determination, 221, 234
- Reaction methods, 612
- Reagents
  - addition of liquids, 622
- Reagents for Organic Synthesis, 864
- Rearrangement
  - benzilic acid, 371
  - benzpinacol, 450
  - carbocation, 291, 450
  - pinacol, 291, 450
- Reducing sugar tests, 541
- Reducing sugars, 541
- Reduction
  - by ferrous hydroxide, 499
  - by hydrogen, 121
  - by sodium borohydride, 176, 270
  - by sodium dithionite, 455
  - by yeast, 570
  - camphor, 176
  - chiral, 570
  - ethyl acetoacetate, 570
  - methyl oleate, 121
  - nitro group, 455, 499
  - photoreduction, 442
- Reflux, 619
- Reflux apparatus, 619
- Reflux ring, 620
- Reforming, 154
- Refractive index, 836
- Refractive index
  - temperature corrections, 841
- Refractometer
  - apparatus, 838, 839
  - cleaning, 841
  - operation, 838
- Refractometry, 836
  - microscale method, 840
- Regioselective, 266
- Regiospecific, 266
- Resolution of enantiomers, 563
- Retention time, 817
- Retinal, 347
- Reversed phase chromatography, 789
- Rf value, 803, 807
- Rhodopsin, 347
- Right-to-Know Laws, 10
- Ring current in NMR spectroscopy, 918
- Robinson annelation, 401
- Rotary evaporator, 633
- Round-bottom flask, 37
- Rubber septum, 38

S

- Saccharine, 534
  - structure, 546
- Safety, 5
  - eye, 5
  - fires, 6
  - first aid, 10
  - flames, 8

Safety (*continued*)
  general, 5
  organic solvents, 6
  solvent disposal, 7
  solvents, 6
Safety glasses, 5
Salicylamide, 54, 79
Salicylic acid, 49, 103, 108
  acetylation, 49
  esterification, 103
  preparation, 108
Salt plates, 843
Sample vials, 31
Samples,
  containers, 31
  labeling, 31
Sand bath, 608
Saponification, 127, 133
Science Citation Index, 873
Scratching
  to induce crystallization, 661
*See*d crystals, 661
*See*ding
  to induce crystallization, 661
Seliwanoff test, 540
Semicarbazones, 514
  preparation, 888
Separation
  by column chromatography, 765
Separation methods, 700
Separation scheme, 25, 702
Separatory funnel, 37, 690, 691
  method of holding, 692
  use in extraction, 101
Sephadex, 784
Septum
  rubber, 38
Sex attractants, 140
Shielding, 877
Shift reagents, 574, 929
Side products, 23
Side reactions, 23
Side-arm test tube, 38
Silica gel, 761
Silica gel G, 794
Silver nitrate test, 200, 498, 515
Simple distillation, 706
  apparatus, 710
  large-scale apparatus, 711
  methods, 710, 715
  microscale apparatus, 714, 715
Simple distillation (*continued*)
  removing fractions, 713
  semi-microscale apparatus, 714
  See also Distillation
Singlet state, 444
Size-exclusion chromatography, 784
Slurry, 642, 774
Small scale
  definition, 4
SN1/SN2 reactions
  See Nucleophilic Substitution
Soap
  cleaning action, 130
  essay, 127
  preparation, 133
  tests, 138
Soap making, 127
Sodium 2,2-dimethyl-2-silapentane-5-sulfonate, 860
Sodium borohydride, 176, 263, 265
Sodium cyclamate, 535
Sodium D line, 832
Sodium fusion tests, 500
Sodium iodide test, 200, 499
Sodium lauryl sulfate, 131
  preparation, 135
Solids
  measurement, 588
  melting point determination, 666
  purification by sublimation, 824
Solubility rules, 649
Solubility tests, 491
Solution cell
  infrared spectroscopy, 852
  NMR spectroscopy, 855
Solvent disposal, 7
Solvent evaporation,
  methods, 630
  reduced pressure, 631
  rotary evaporator, 633
Solvents
  boiling points, 612
  densities, 690
  for crystallization, 659
  hazards, 6
  heating methods, 600
  mixed, 663
  relative polarities, 650
  safety, 6
Sorbitol, 534
Soxhlet extractor, 703
Spatula, 39
Spearmint oil, 192
Specific rotation, 832
Spectra catalogues, 863
Spectroscopy
  catalogues of spectra, 863
  index of spectra, 900
  sample preparation, 842, 846, 850, 855
  See also Infrared, NMR, or Carbon-13 NMR Spectroscopies
Spices, 169
Spin bar, 39
Spin vane, 621
Spin-spin splitting, 921
Spinning-band column, 741
Spotting TLC plates, 798
Standard scale
  definition, 4
Standard taper
  definition, 33, 614
Standard taper joints
  microscale, 618
Starch, 86
  test for, 544
Starch-iodine test, 52, 544
Steam baths, 610
Steam cones, 610
Steam distillation, 171, 753
  essential oils, 169, 171
  methods, 756, 757, 759
  microscale apparatus, 758
  spices, 169
  standard scale apparatus, 757, 759
  theory, 753
  trap, 759
  See also Distillation
Stearic acid
  conformations, 585
Stem corrections, 680
Steric approach control, 180
Stirring bars, 608, 621
Stopper
  ground glass, 37
Strain energy, 418
Streaming, 771
Styrene
  infrared spectrum, 901

- Sublimation, 824
  - advantages, 828
  - apparatus, 828
  - caffeine, 73
  - methods, 827
  - on melting, 672
  - theory, 824
  - vacuum, 827
- Sucrose, 89, 533
  - fermentation, 88
  - hydrolysis, 545
- Suction filtration, 640
  - *See also* Vacuum filtration
- Sugars
  - identification, 536
  - reducing, 541
- Sulfa drugs
  - essay, 375
  - preparation, 378
  - tests on bacteria (*See* Instructor's Manual)
  - types, 376
- Sulfanilamide
  - action, 377
  - infrared spectrum, 383
  - preparation, 378
- Sulfanilic acid
  - diazotization, 318
- Sulfonation
  - lauryl alcohol, 131, 135
- Sulfur
  - elemental analysis, 502
- Sweeteners
  - essay, 533
- Synthetic dyes, 315
- Syringe, 39, 597

T

- Tables of derivatives, 878
- Tables of unknowns, 878
- Tailing, 781, 803
- Tannins, 68, 69
- Tare weight, 588
- Tea
  - caffeine content, 65
  - discovery, 64
  - extraction, 68
- Technique of Organic Chemistry, 866
- Terpenes
  - essay, 165
- Test tube
  - side arm, 38
- Tetracycline, 376
- Tetraethyllead, 154
- Tetrahydrofuran,
  - hazards, 20
- Tetramethylsilane, 859, 913
- Tetraphenylcyclopentadienone
  - Diels-Alder reaction, 436
  - infrared spectrum, 397
  - preparation, 395
- 1, 2, 3, 4-Tetraphenylnaphthalene
  - infrared spectrum, 441
  - NMR spectrum, 441
  - preparation, 436
- Theobromine, 64
- Theophylline, 64
- Theoretical plates, 740
- Theoretical yield, 27
- Thermometer,
  - bulb immersion, 680
  - dial, 605
  - stem corrections, 680
  - stem immersion, 680
  - total immersion, 681
- Thermometer adapter, 37, 712
- Thermometer calibration, 680, 681
- Thermoplastic, 469
- Thermoset, 467
- Thiamine
  - catalytic action, 360, 365
  - essay, 360
  - mechanism of action, 360
- Thiele tube, 669
- Thin-layer chromatography, 792
  - analgesics, 79
  - chemical applications, 804
  - coating slides, 795
  - commercially prepared plates, 794
  - development chambers, 799
  - food dyes, 342
  - formation of camphor, 173
  - large plates, 796, 802
  - micropipet preparation, 797
  - monitoring a reaction, 173
  - oxidation of borneol, 173
  - preparative, 802
  - principles, 792
  - Rf values, 803
  - sample application, 798
  - slide preparation, 794
  - slurry preparation, 795
- Thin-layer chromatography, (*continued*)
  - solvent choice, 800
  - spinach, 352
  - spotting plates, 798
  - visualization methods, 801
- Three-necked flask, 37
- TLV, 12
- TMS, 859, 872
- Tollens test, 511
- Toluene
  - carbon-13 NMR spectrum, 936
  - hazards, 20
- *m*-Toluic acid
  - amide formation, 145
- *p*-Toluidide derivatives, 517
  - preparation, 889
- *p*-Toluidine
  - acetylation, 383
- Transesterification, 116
- Trap
  - acidic gases, 148, 208, 215, 310, 625
  - aspirator, 641
  - manometer, 732
  - steam distillation, 759
  - vacuum, 721, 732
  - vacuum distillation, 726, 732
  - vacuum pump, 728
- Triangulation of gas chromatography peaks, 820, 821
- 1,1,2-Trichloroethane
  - NMR spectrum, 881
- Triglyceride, 111
- Triphenylmethanol
  - infrared spectrum, 285
  - preparation, 276
- Triphenylphosphine
  - Wittig reaction, 460
- Triple bonds
  - detection, 503
  - spectroscopy, 506
- Triplet state, 444
- Tubing
  - thin-walled, 720
  - vacuum, 720

U

- Ultraviolet-visible spectrophotometry
  - use in kinetics, 242

Ultraviolet-visible spectroscopy
  index of spectra, 940
Unknown compounds
  identification, 485
  tables of, 878
  See also specific functional groups
α,β-Unsaturated ketones
  addition reactions, 413
Unsaturation tests, 126, 251, 503
Upfield, 914

V

Vacuum Adapter, 37, 712
Vacuum distillation, 717
  acetophenones, 311
  bulb-to-bulb, 726
  fraction collectors, 725
  methods, 717
  methyl salicylate, 103
  microscale apparatus, 726
  standard scale apparatus, 718
  stepwise directions, 722
  See also Distillation
Vacuum filtration, 640
  apparatus, 641
Vacuum pump, 728
Vacuum pump trap, 728
Vacuum tubing, 720
Vanillin
  esterification, 579
Vapor phase chromatography
  See Gas Chromatography
Varian NMR Spectra Catalog, 863
Vigreux column, 741
Vinyl acetate
  NMR spectrum, 925
Visible spectrophotometer, 241
Visible spectroscopy
  use in kinetics, 241, 244
Vision
  chemistry of, 347
Visualization reagents for TLC, 801
Vitamin, 365
Vitamin A, 349

W

Wash acetone, 32
Waste Disposal, 7
  acids and bases, 8
  broken glass, 8
  halogenated solvents, 8
  heavy metals, 8
Waste Disposal (*continued*)
  non-halogenated solvents, 8
  solids, 8
Water condenser, 712
Water pump, 643
Water separator
  azeotropic, 749
Wavenumbers, 893
Weighing
  liquids, 588
  solids, 588
Weight percent recovery, 28
Wintergreen
  See Methyl Salicylate
Wittig reaction, 460

X

Xanthines, 64

Y

Yeast
  use as a reducing agent, 570
  use in fermentation, 88
Yield, 27
  actual, 27
  calculation, 27
  theoretical, 27
Ylide, 361, 461

BUILDING A DIY VIDEO TRANSMITTER

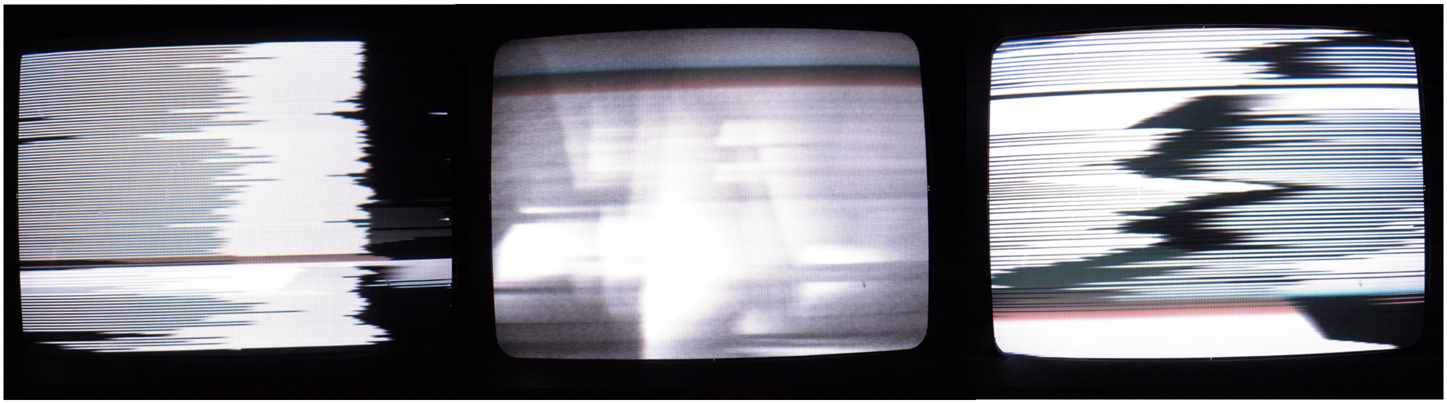

WHAT YOU NEED

- Blank PCB Copper Plate (jameco.com part #169279)
- Capacitors: 2x 0.1uf
  1x 100u
  2x 6-70pf Variable Caps (jameco.com part #32855)
- Resistors: 27kOhm
  10kOh
- Transistor-MPSA18 (jameco.com part #210681)
- 1k potentiometer
- Small amount (~4”) of magnet wire 24 AWG or close (jameco.com part #2098419)
- Sand paper
- Stranded hookup wire
- Hacksaw or dremel tool
-Headphones with 1/8” plug (from iPod or MP3 player)
- Hot glue or super glue
- Soldering Iron
- CRT “tube” TV with an antenna (the transmitter won’t transmit to newer flat screen TV’s)

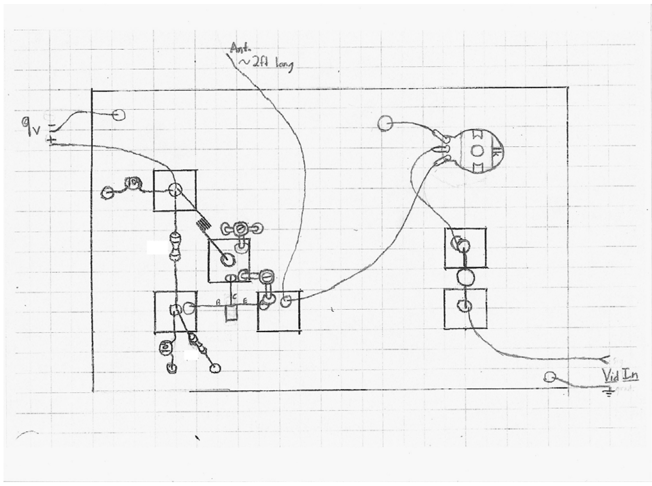

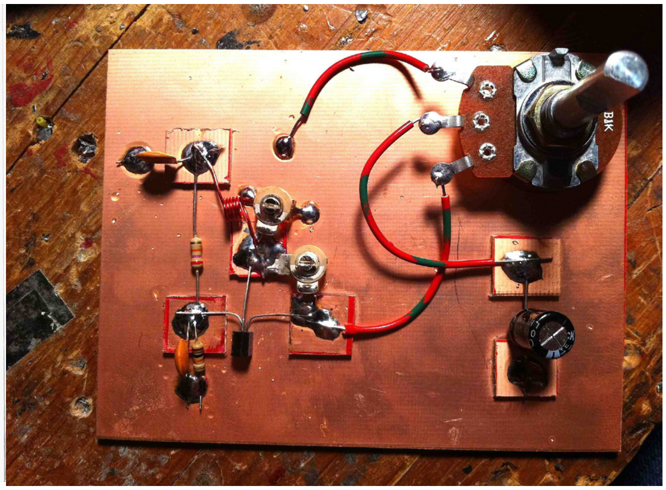

INTRODUCTION

This micro video transmitter was designed by the DIY radio and micro TV artist Tetsuo Kogawa. His website (http://anarchy.translocal.jp/) is full of great info on DIY radio, and all credit for this circuit goes to him.The transmitter can be used to transmit a relatively clear video signal or abstract visualizations, depending on the desired output. It will usually function at a distance of about 20 feet. If you have any previous experience with electronics or circuit building, you will notice that this method of building is different. The circuit is constructed on top of a copper “ground plate” which is common in radio applications. This tutorial uses step-by-step sketches and photos created based on Tetsuo Kogawa’s design.

BUILDING THE TRANSMITTER

1. PREPARING THE COPPER PLATE

- Take the bare copper plate and cut off 1inch of material using a dremel tool or hacksaw.
- Use this material to cut out 6 square ½ in. x ½ in. pieces.
- Arrange the 6 pieces on the larger copper plate as shown in the image below…

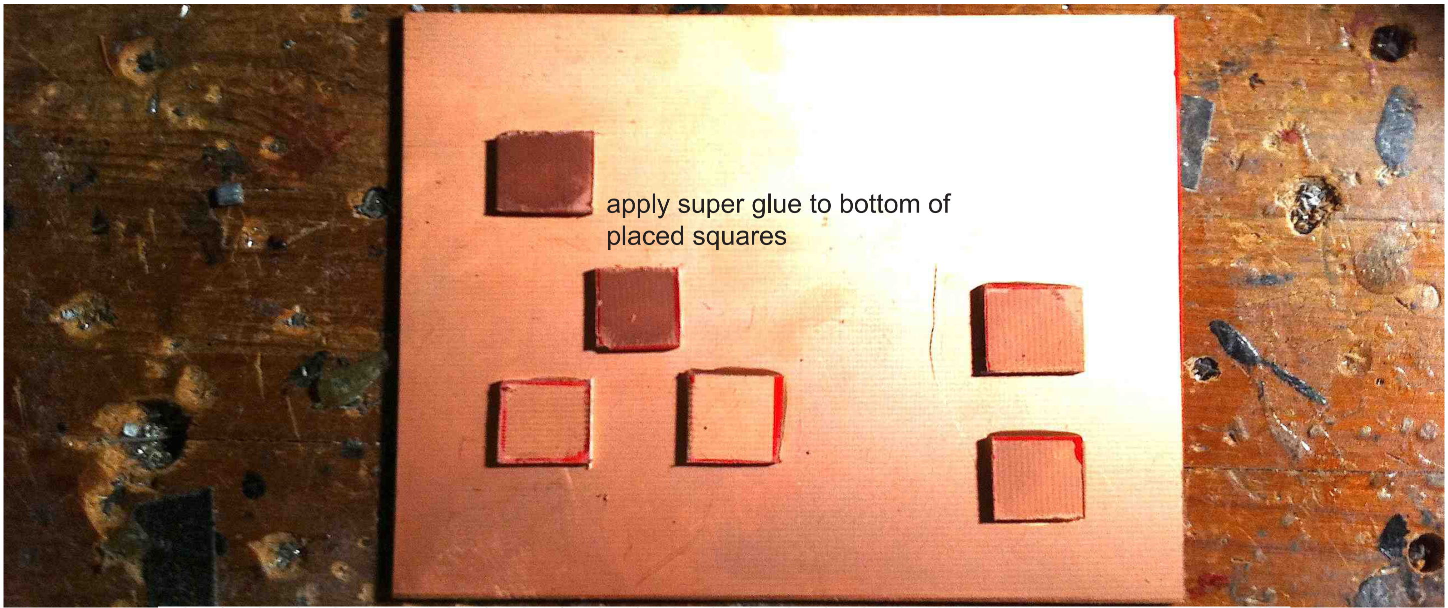

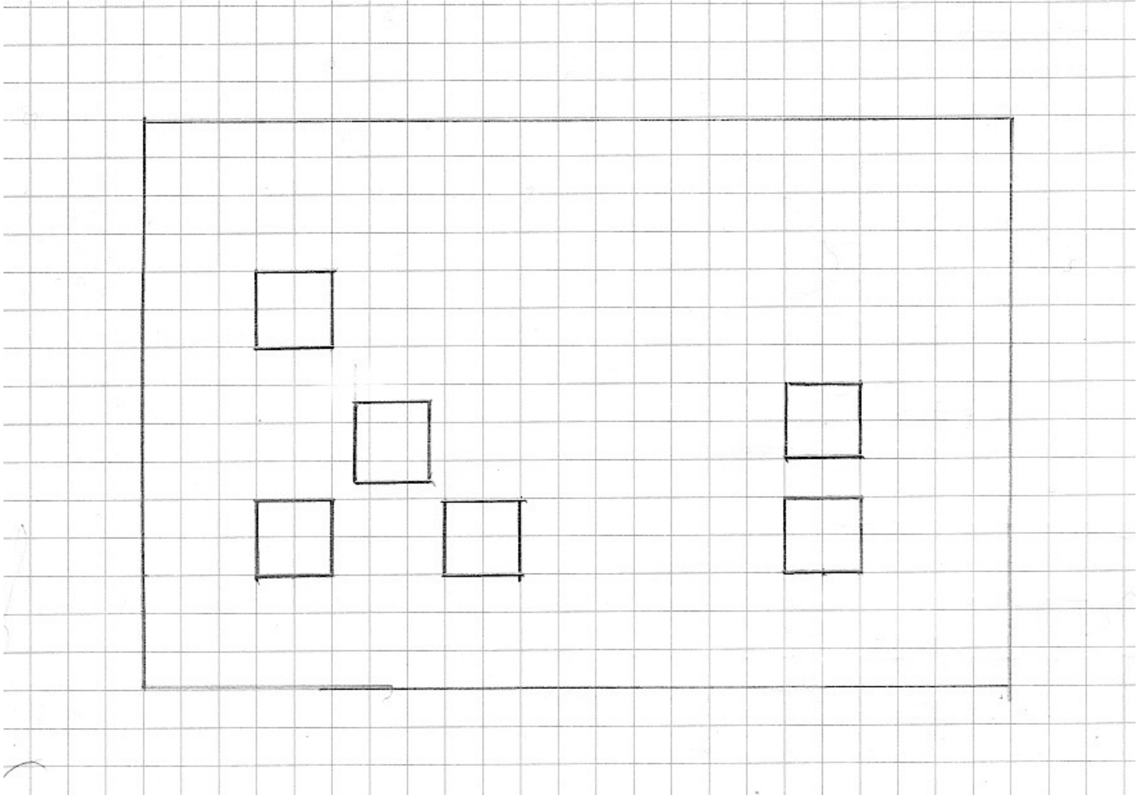

****NOTE**** Be sure to arrange them so that all the component leads will be able to reach the squares. Look ahead to see what components will go where so you know exactly how to space them.

- Once you have them arranged, put a dab of hot or super glue on the bottom of each square and stick them in place on the copper plate.

2. ADDING THE 100uF CAPACITOR (the cylinder with “100uF” written on the side)

- Melt a blob of solder onto the two squares furthest to the right. To do this, apply your iron to each square to transfer heat for an extended period of time until the solder flows onto the surface.

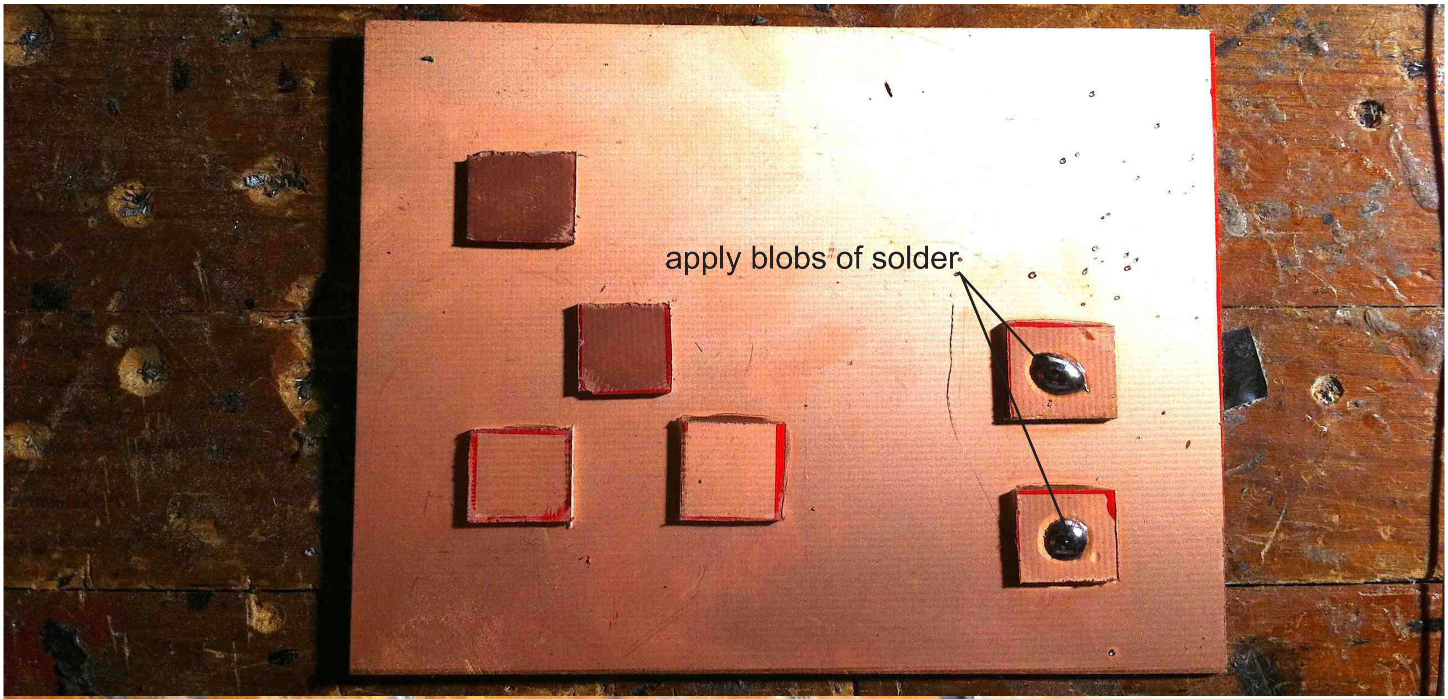

apply blobs of solder

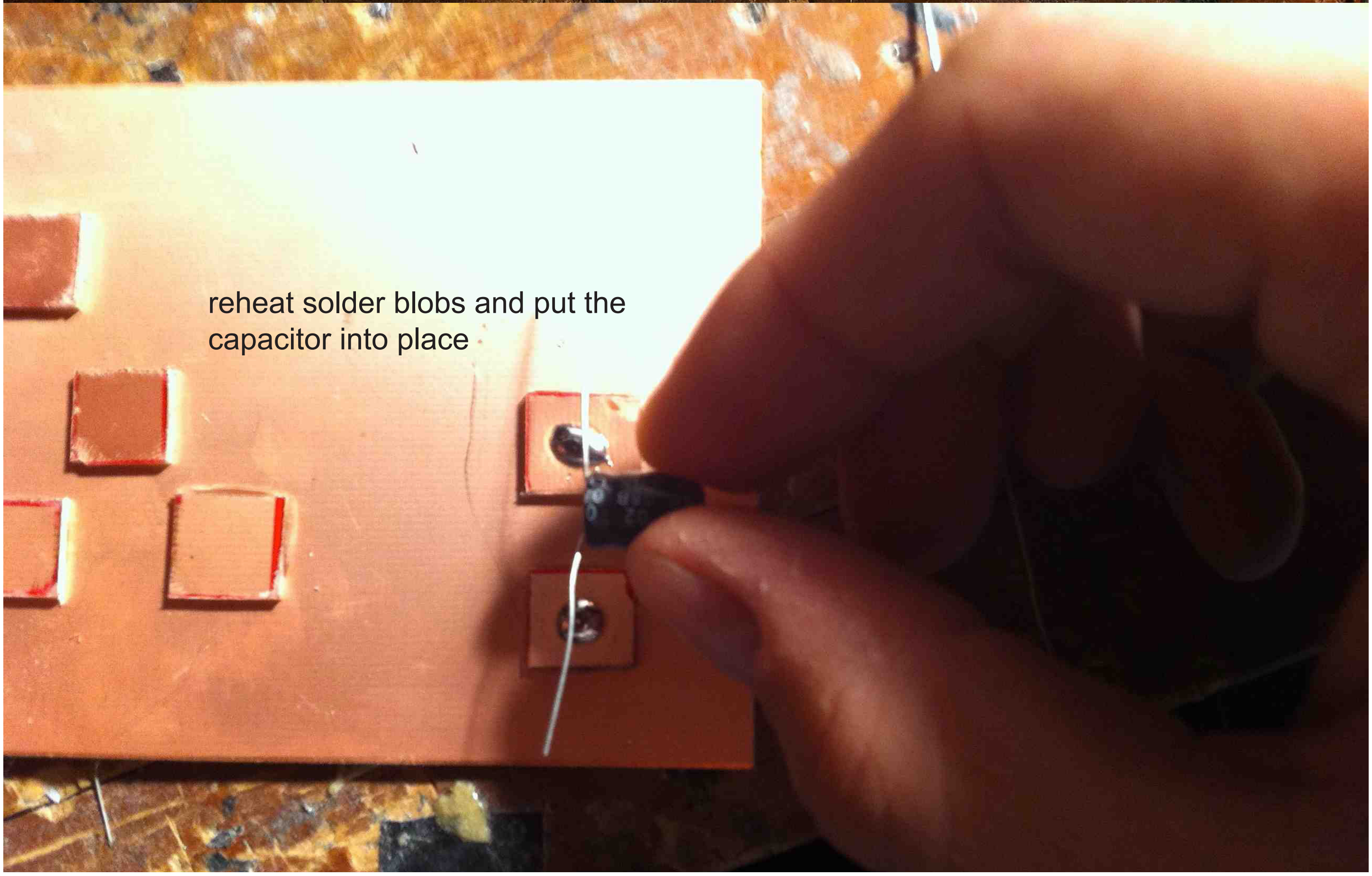

reheat solder blobs and put the capacitor into place

- Bend the legs out of the 100uf capacitor and straddle them between the two squares making sure that the leads are only touching the two small squares (not touching the ground plate).
- Cut the excess off of the leads.
- One at a time, re-heat the solder blobs and place the capacitor leads into them, allowing them to cool down until they stay in place.

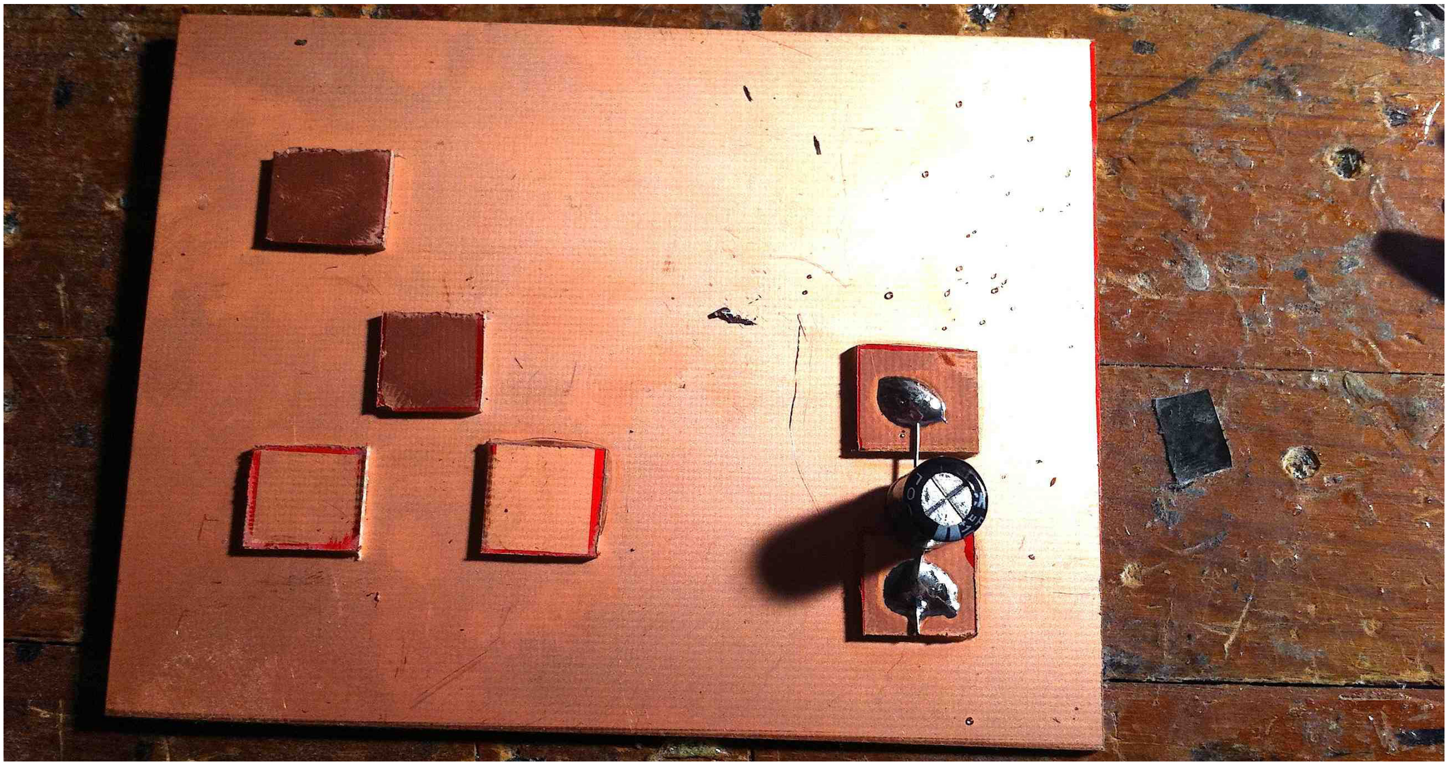

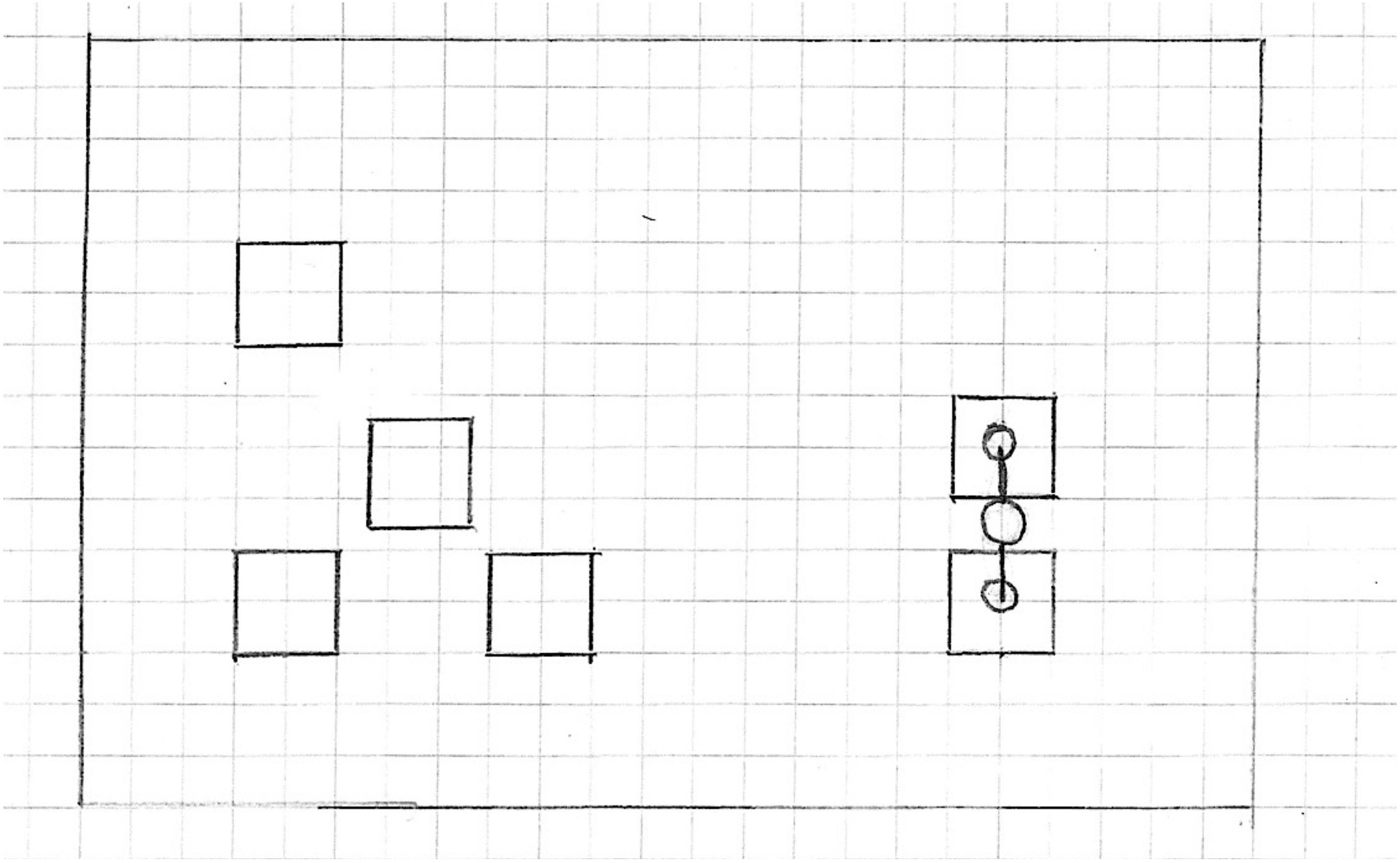

*****NOTE***** - This capacitor is polarized meaning that it has to be facing in the appropriate direction. Make sure you solder the NEGATIVE lead to the LOWER of the two squares. The negative side has the shorter lead and a vertical stripe along the side of the capacitor.

3. ADDING THE .1uF CAPACITORS (the two discs with “104” written on them)

- These capacitors will be positioned between the upper and lower left most squares and the ground plate. Melt a blob of solder on these two squares and two blobs on the ground plate beside each. (Reference the pictures for exact positioning).
- Re-heat the blobs and position the capacitors accordingly as in step 2.

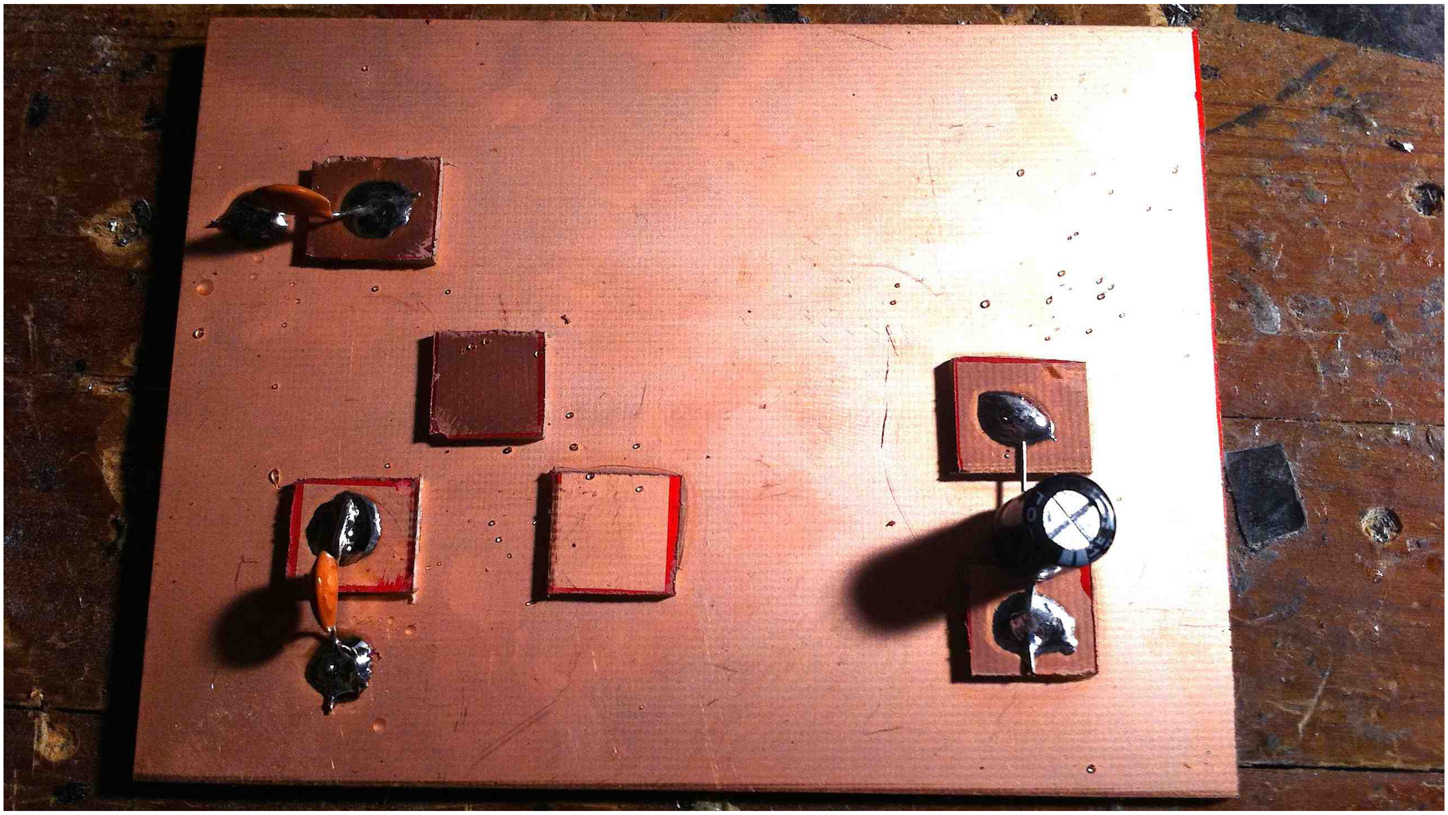

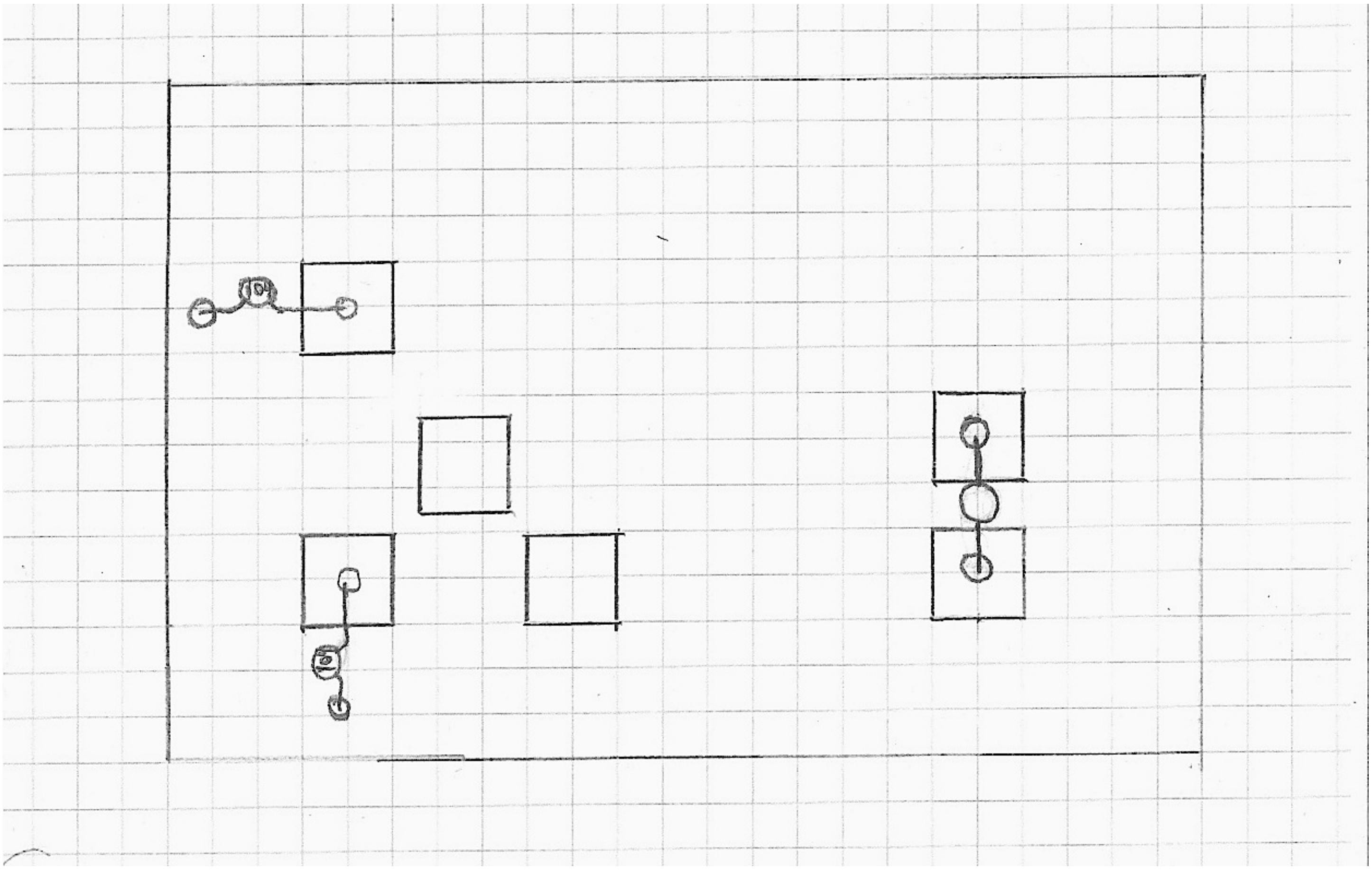

4. ADDING THE RESISTORS (10k and 27k)

- Solder the 10k resistor (brown, black and orange stripes) between the bottom left square and ground.
- Solder the 27k resistor (red, violate and orange stripes) between the two left squares.

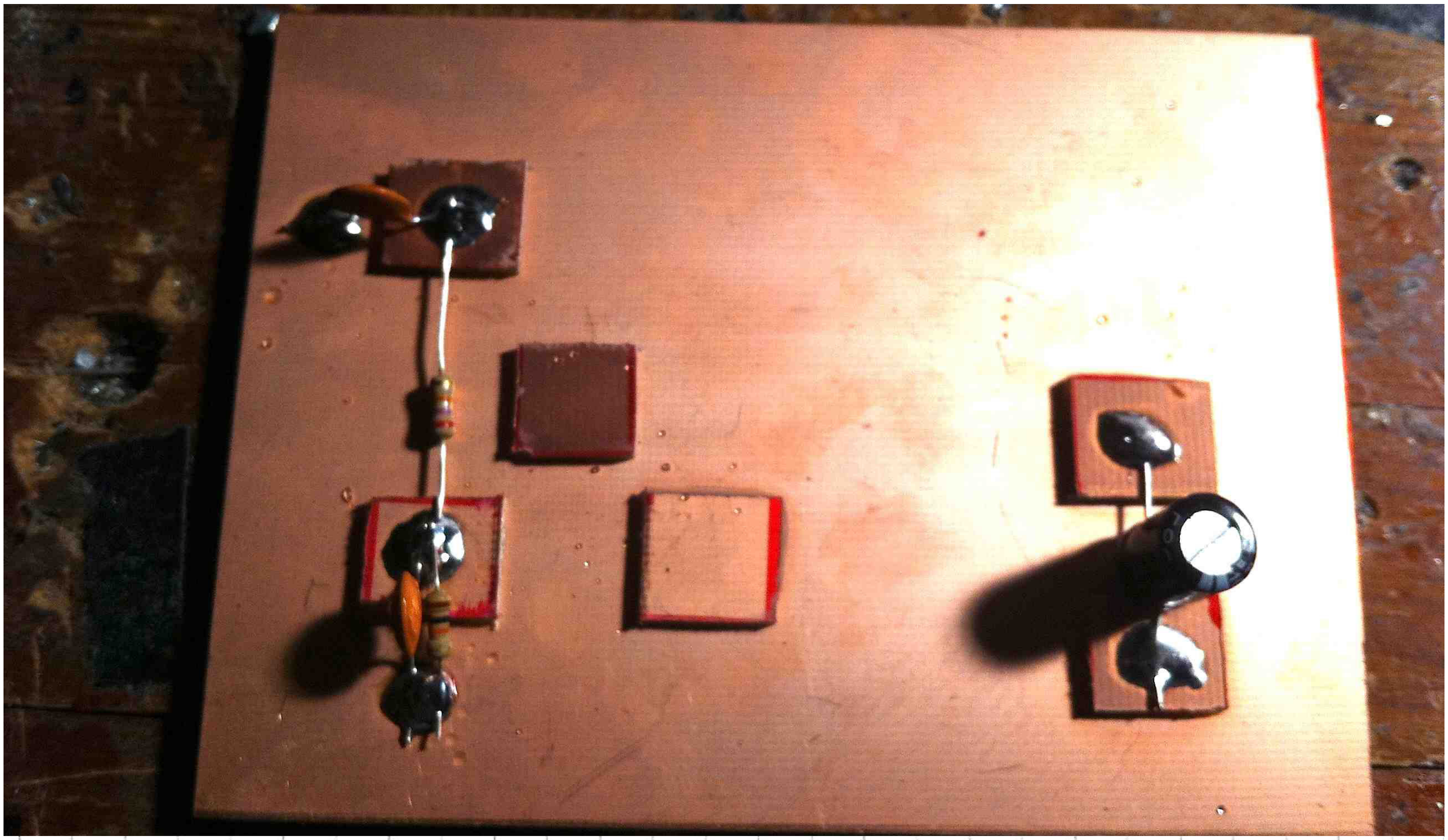

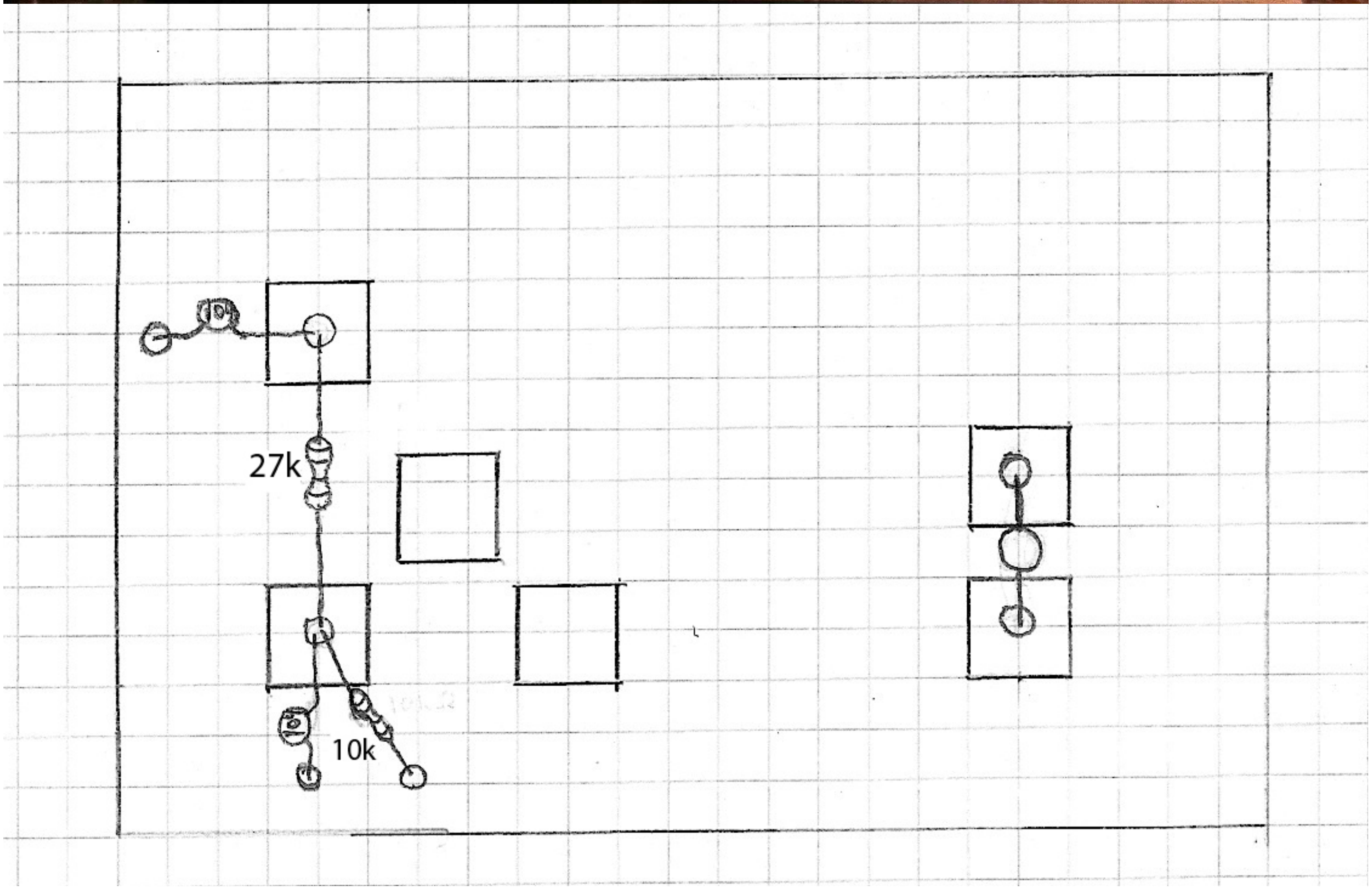

5. ADDING THE COIL

- Cut about 4” of the magnet wire (24 AWG or close)
- Use the plug from your headphones to wrap the wire around. Make 5 turns.

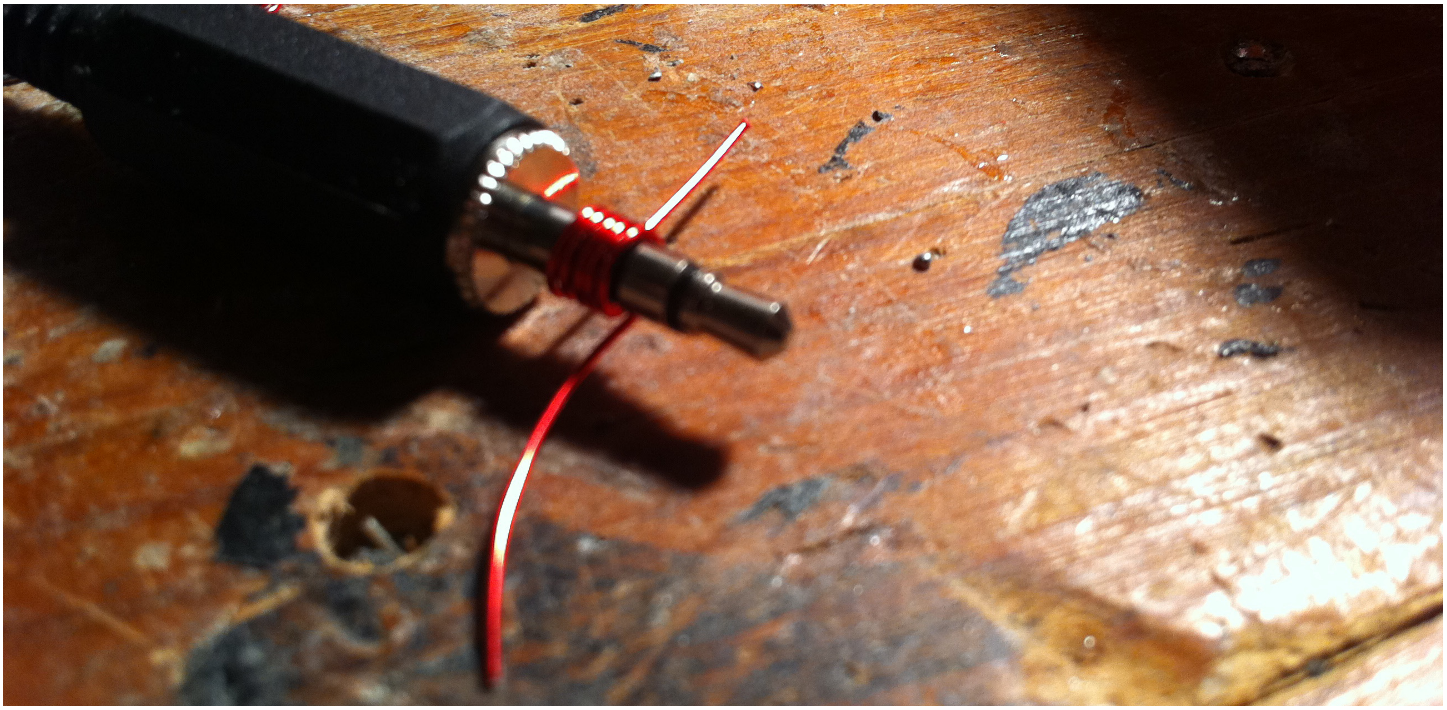

- With the wire still wrapped around the headphone plug, take a small piece of sand paper and rub off the colored insulation of each end until the bare copper is exposed (If the copper is not exposed, you won’t be able to make a connection).

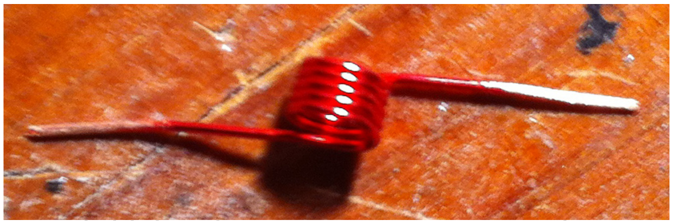

- Solder the coil into place bridging the two squares as seen below

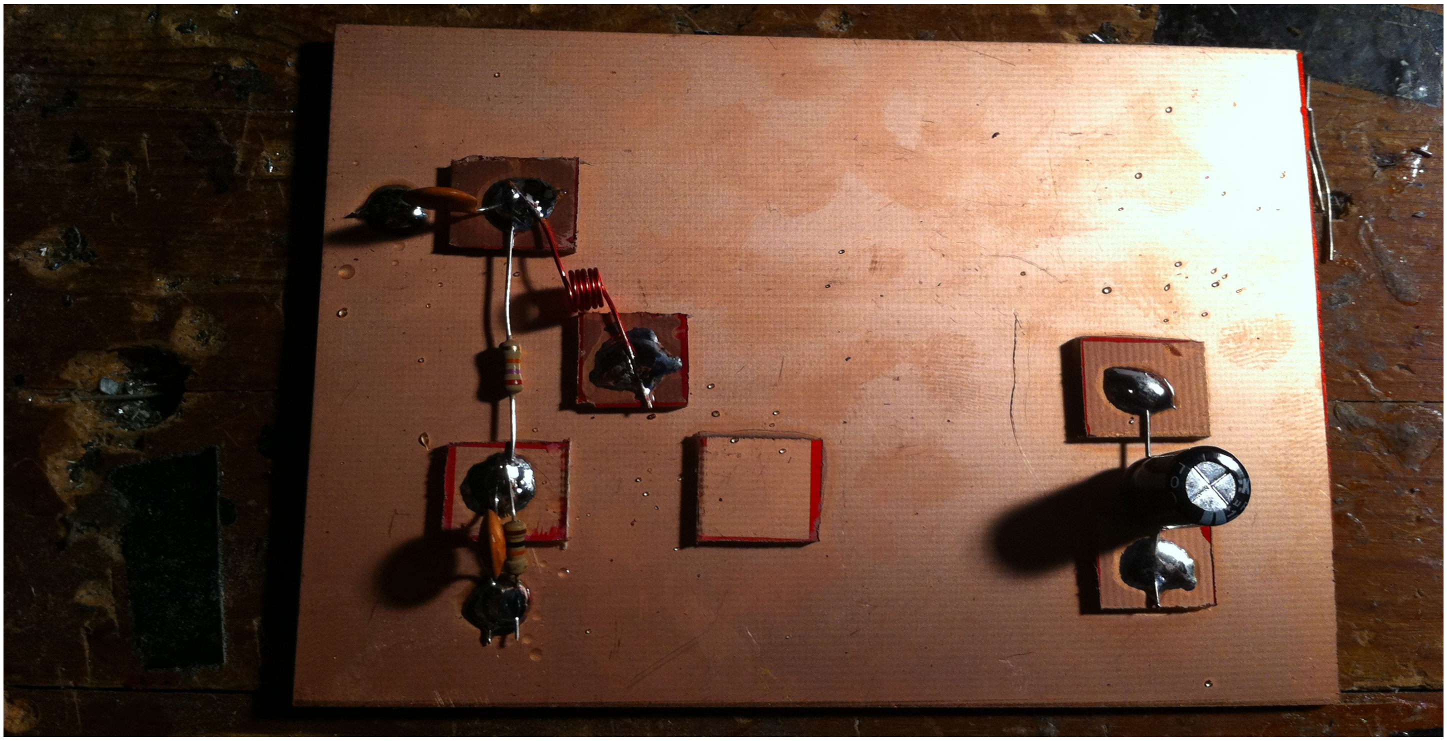

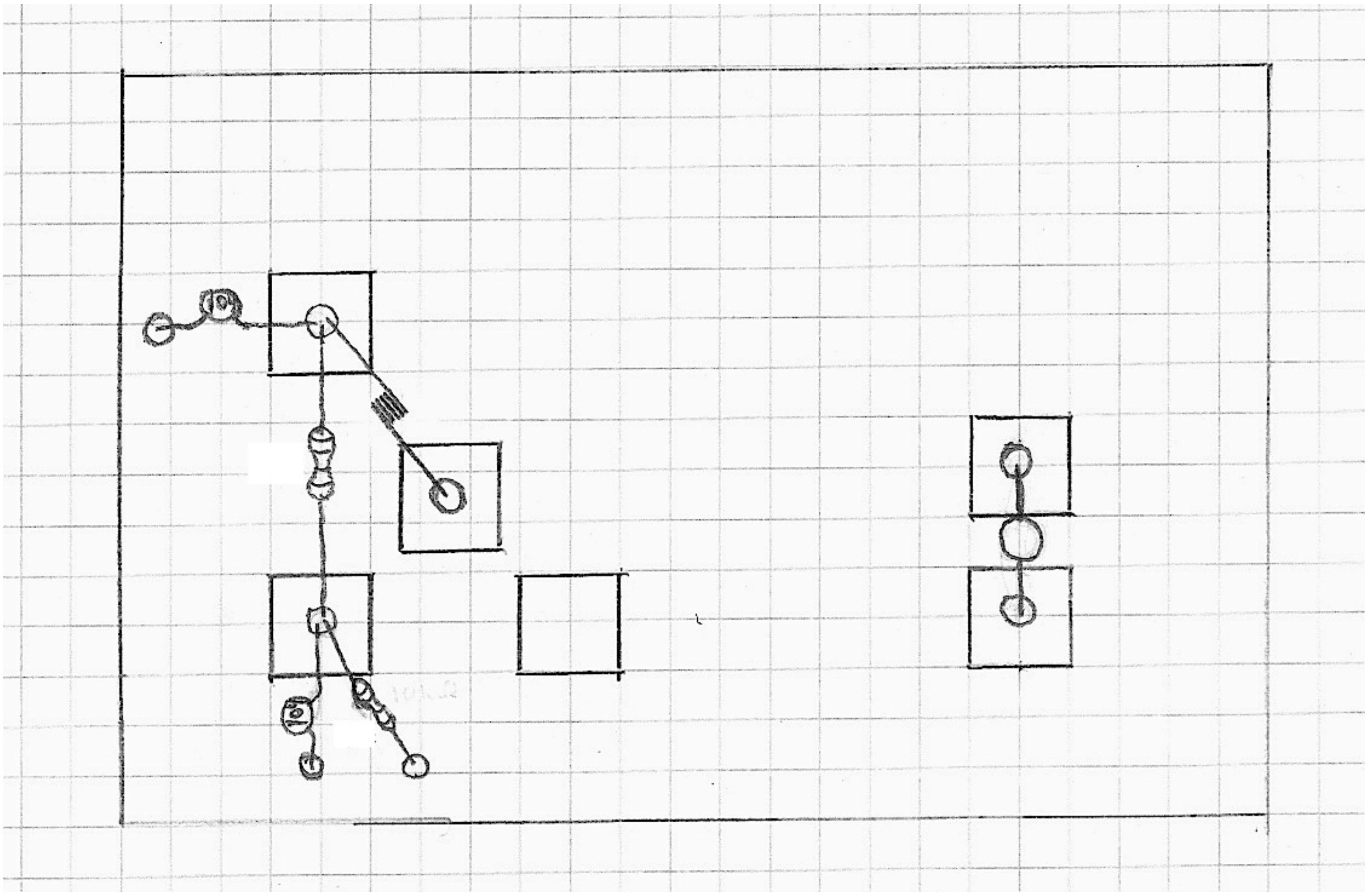

6. ADDING THE VARIABLE CAPACITORS

NOTE Once the component leads start to pile up on each other on each square, it helps to add additional solder.

- The first variable capacitor will be soldered between the top square of the three that make up the bottom triangle and the ground plate.

- Solder the “nose” of the capacitor (the lead that is different from the other two) to the square, and the other two side leads to the ground plate.

NOTE You will have to bend the leads in a certain way in order to get everything positioned properly. Just play around with it till you get it.

- The second variable capacitor will bridge between top and right squares of the triangle. Only two leads are needed here so use your clippers to remove one of the side leads, leaving the nose and only one side. (See the pic below)

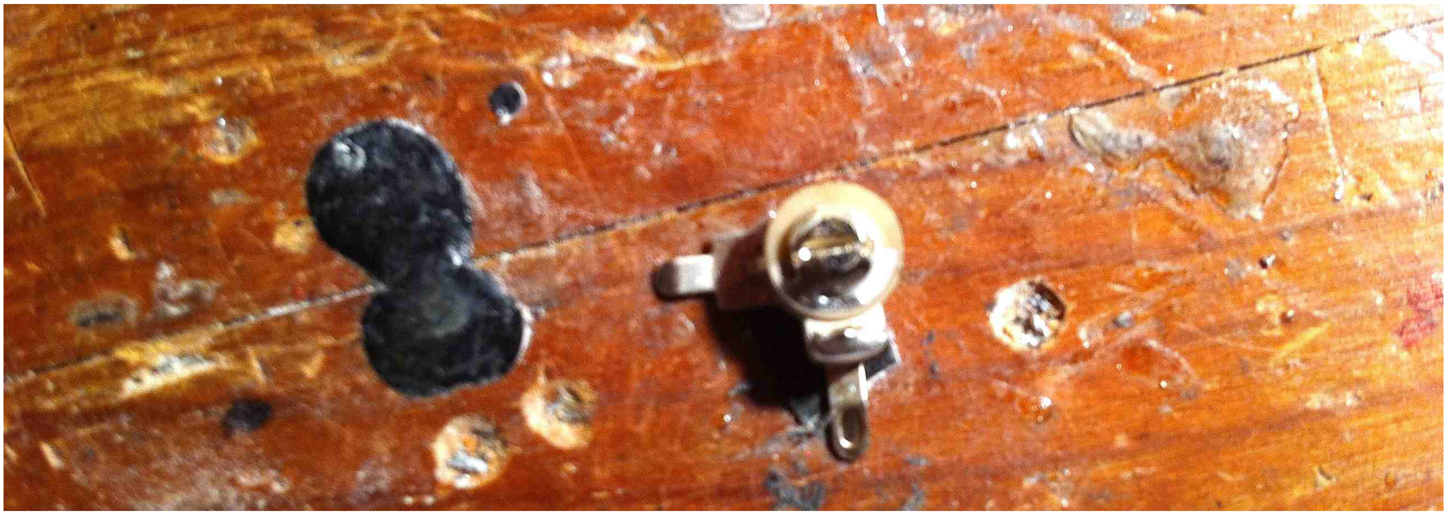

The variable capacitors soldered into place…

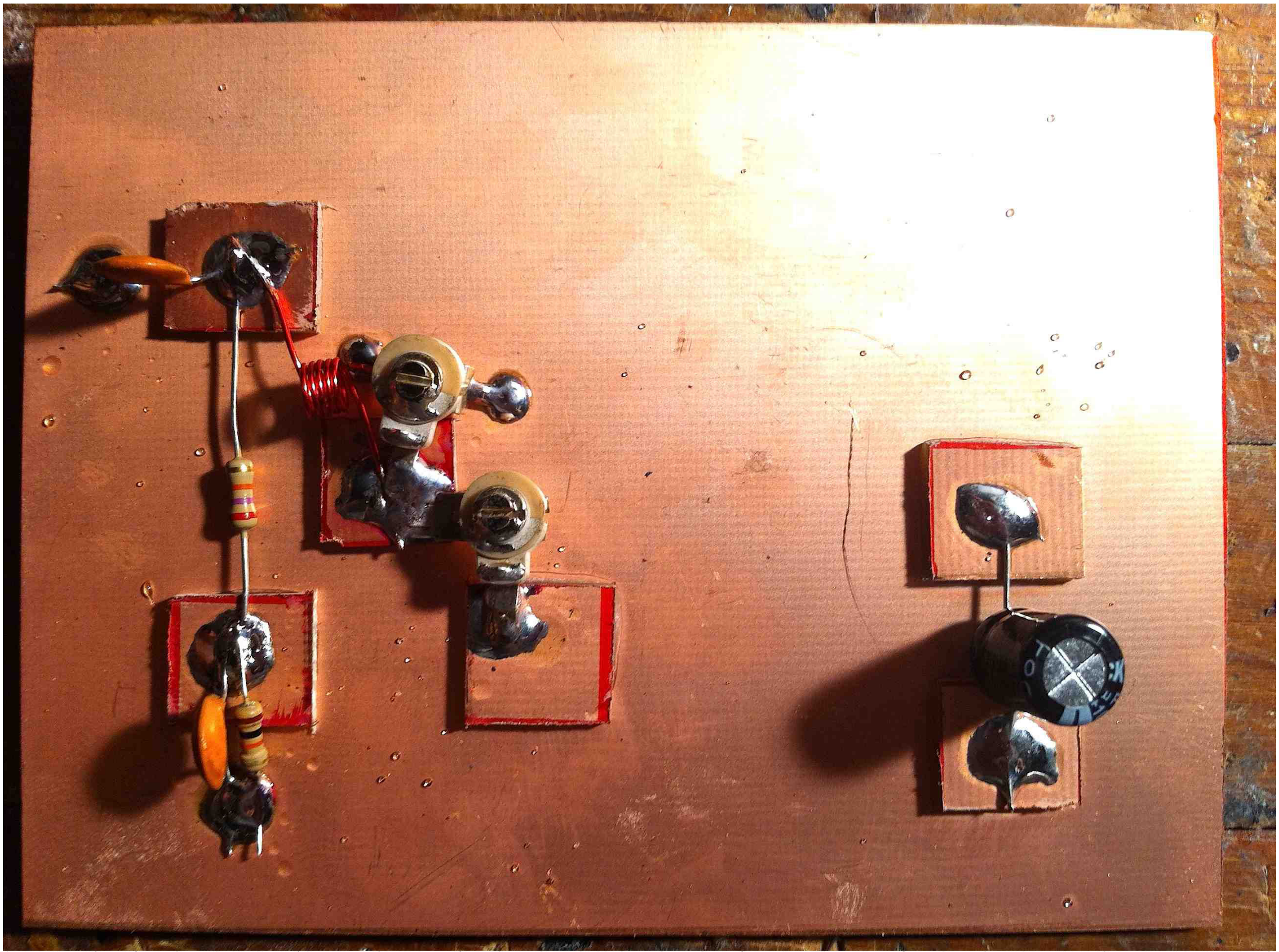

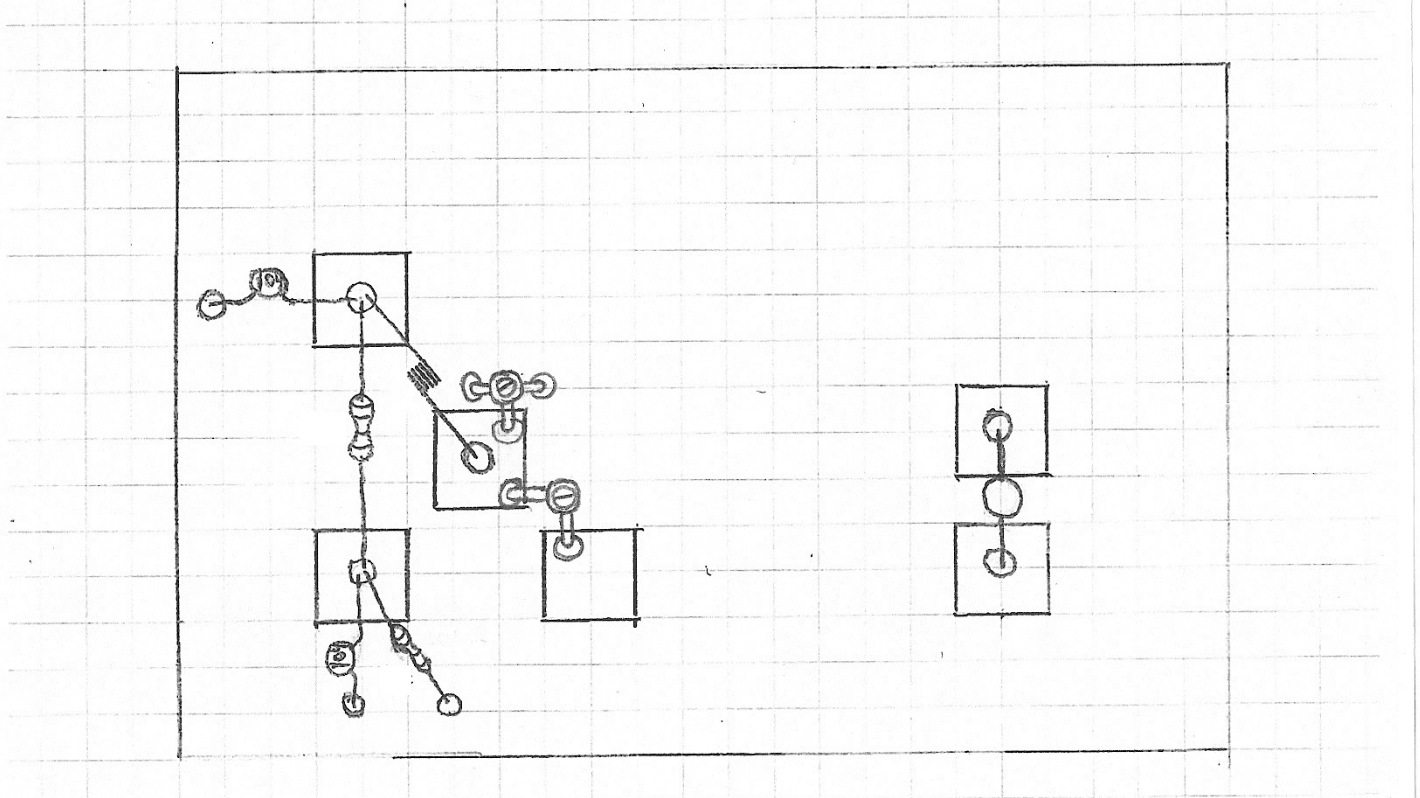

7. ADDING THE TRANSISTOR (MPSA18)

NOTE This transistor is sensitive to heat and can be destroyed by the heat from your iron. Be sure to follow the necessary precautions below. But don't worry if you do kill one, they are only ~$.10 a piece.

- The three leads of the transistor all do specific things and must go to the appropriate places. Here is a pic of the orientation of this transistor…

C=Collector B=Base E=Emitter

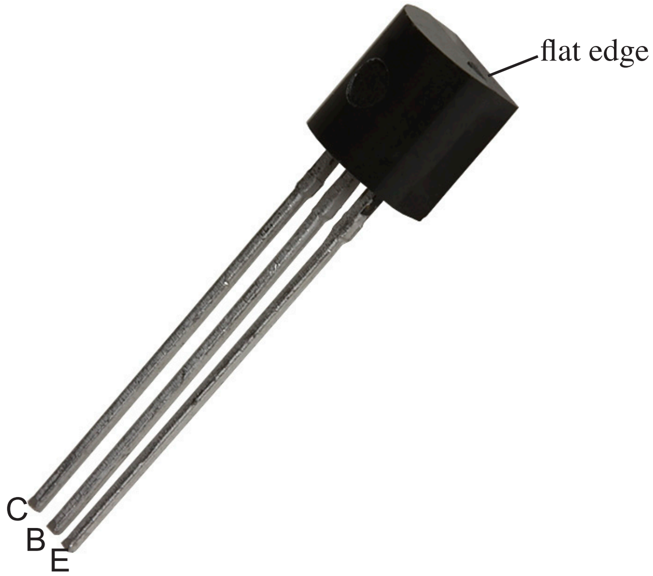
flat edge
C
B
E

- Each lead from the transistor will be soldered to the three different squares that make up the triangle, so you will have to bend the leads accordingly.
- The B lead (base) goes to the left square of the triangle, the C lead (collector) goes to the top square, and the E lead (emitter) goes to the right square, as in the image below.

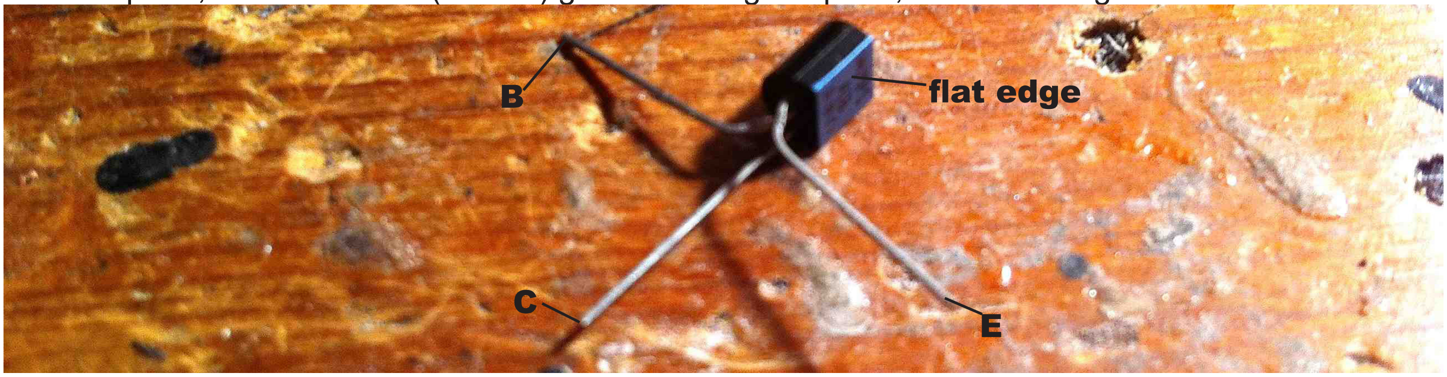
B
flat edge
C
E

- Before you solder the transistor in place, you need to have a "heat sink" attached to protect it. This will draw some of the heat from the transistor to something else. To do this, simply take an alligator clip and attach it to the lead you are soldering…

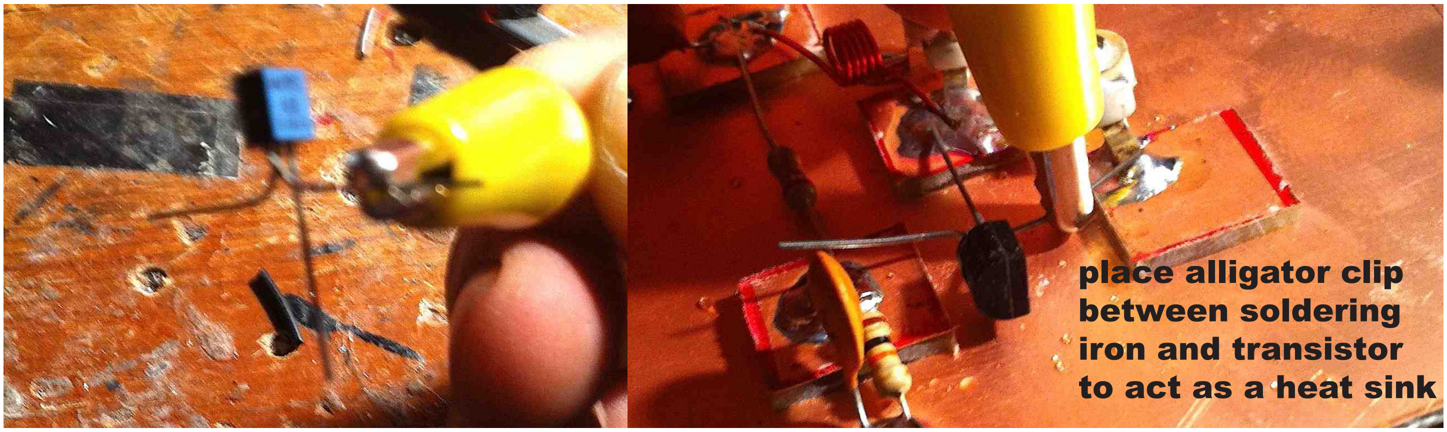
place alligator clip between soldering iron and transistor to act as a heat sink

- Solder each lead in place attaching the clip lead to each one before doing so.

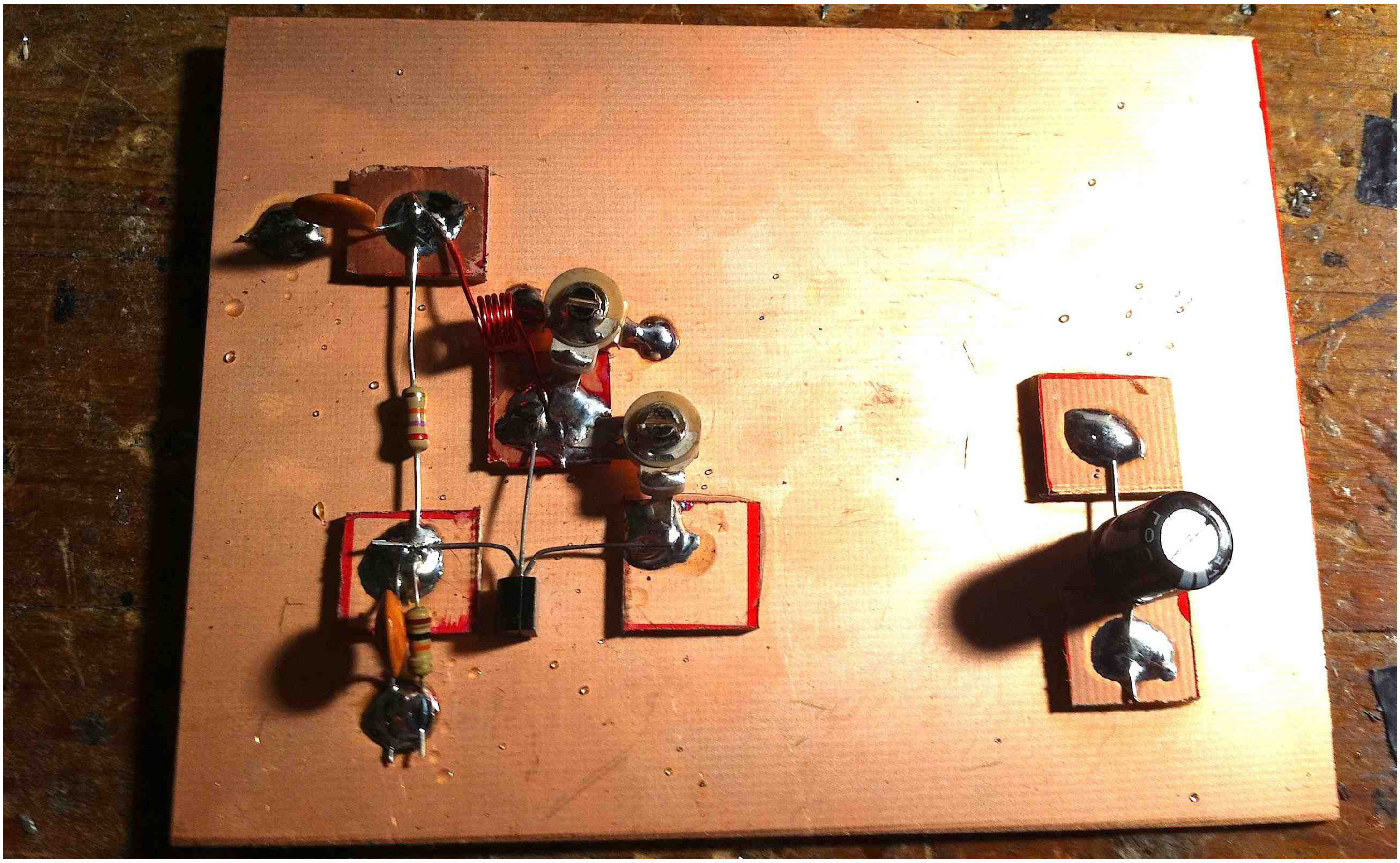

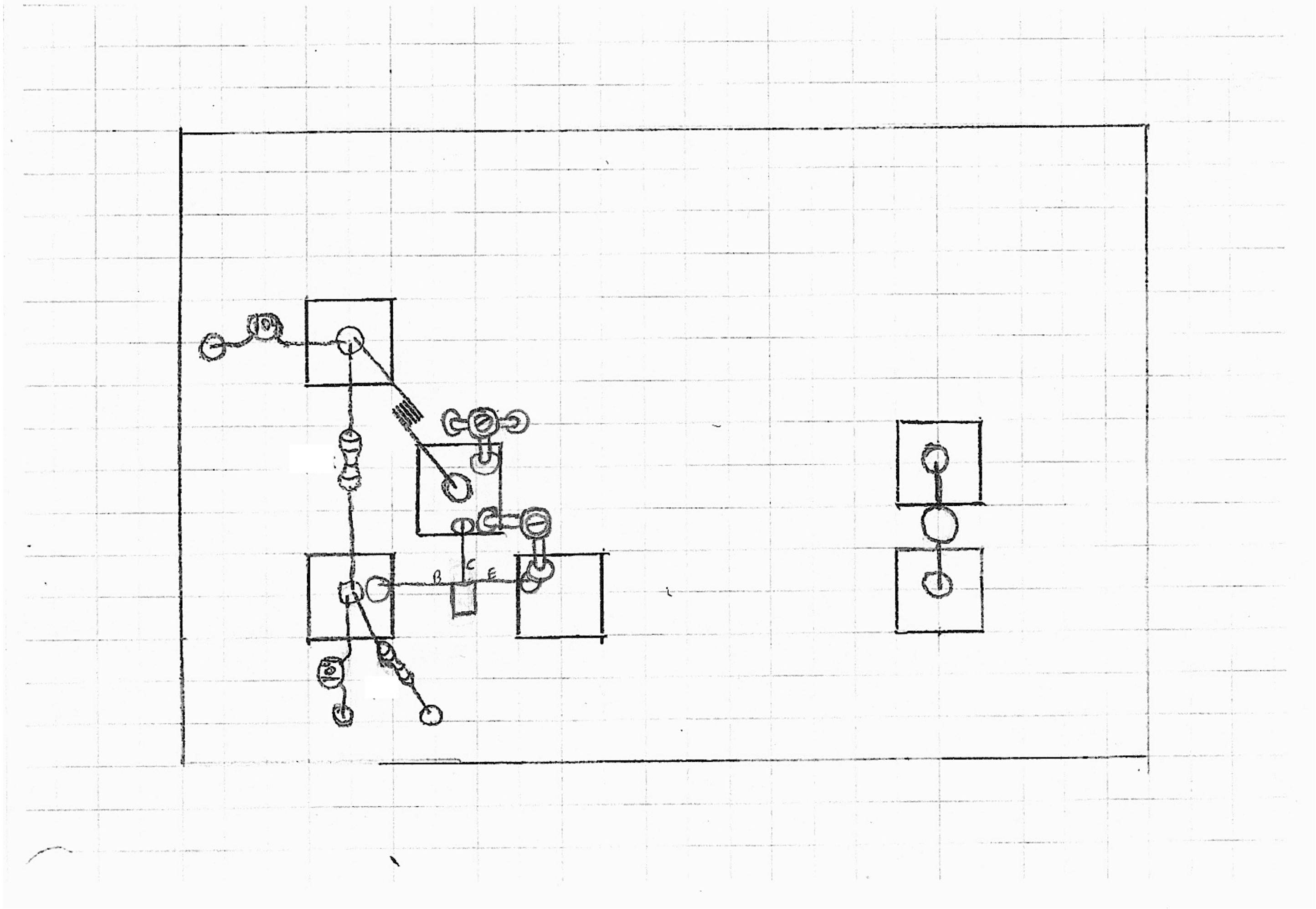

8. ADDING THE 1k POTENTIOMETER

- Put a bit of super glue or hot glue on the bottom of the potentiometer and stick it to the upper right corner with the leads facing towards the left of the plate.
- Take a small amount of stranded hookup wire and strip both ends. Solder one side to the upper most lead and the other side to the ground plate.
- Repeat the process by connecting the middle lead of the potentiometer to the upper most of the two right squares. Then connect the lower lead of the potentiometer to the right most square of the triangle.

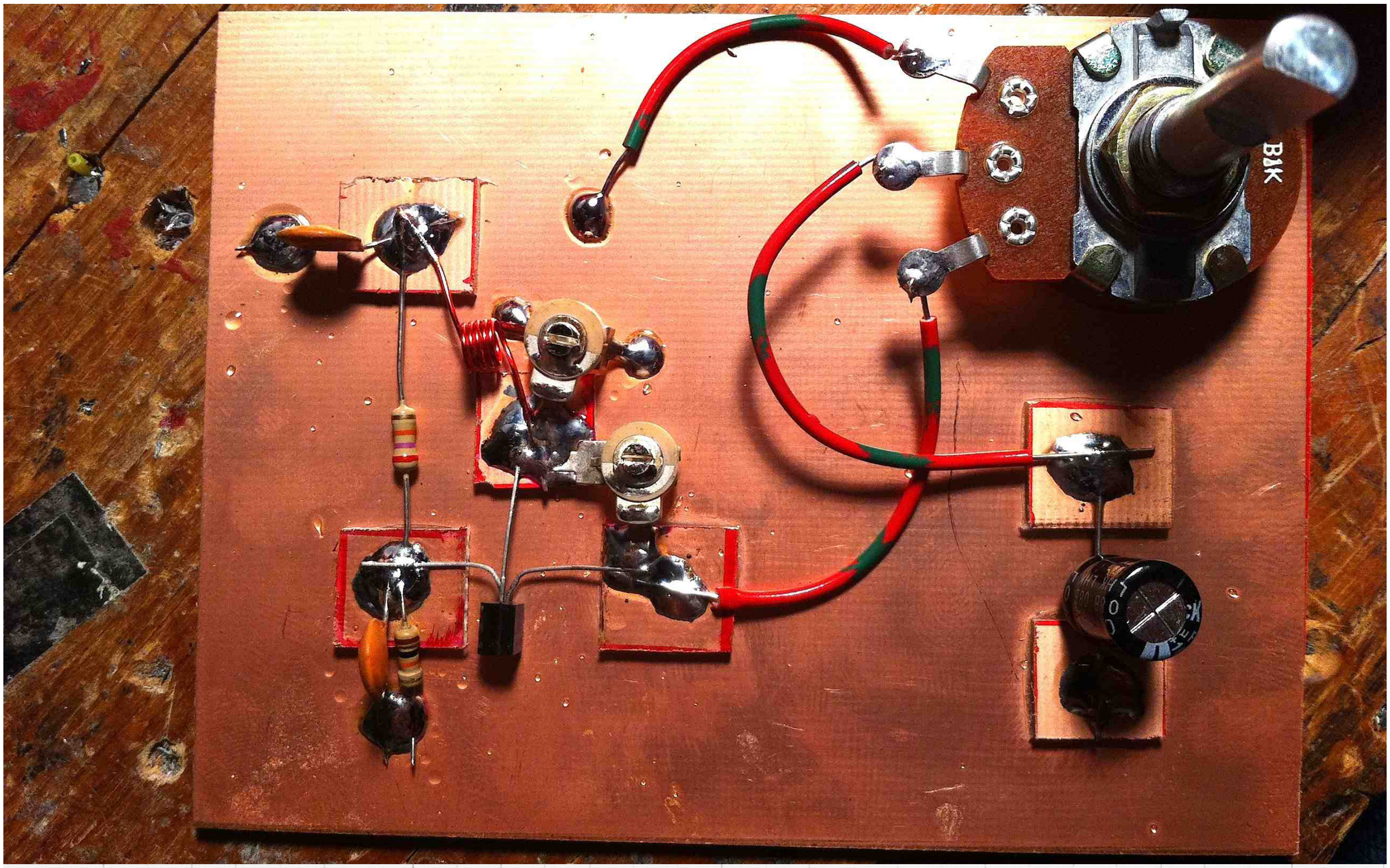

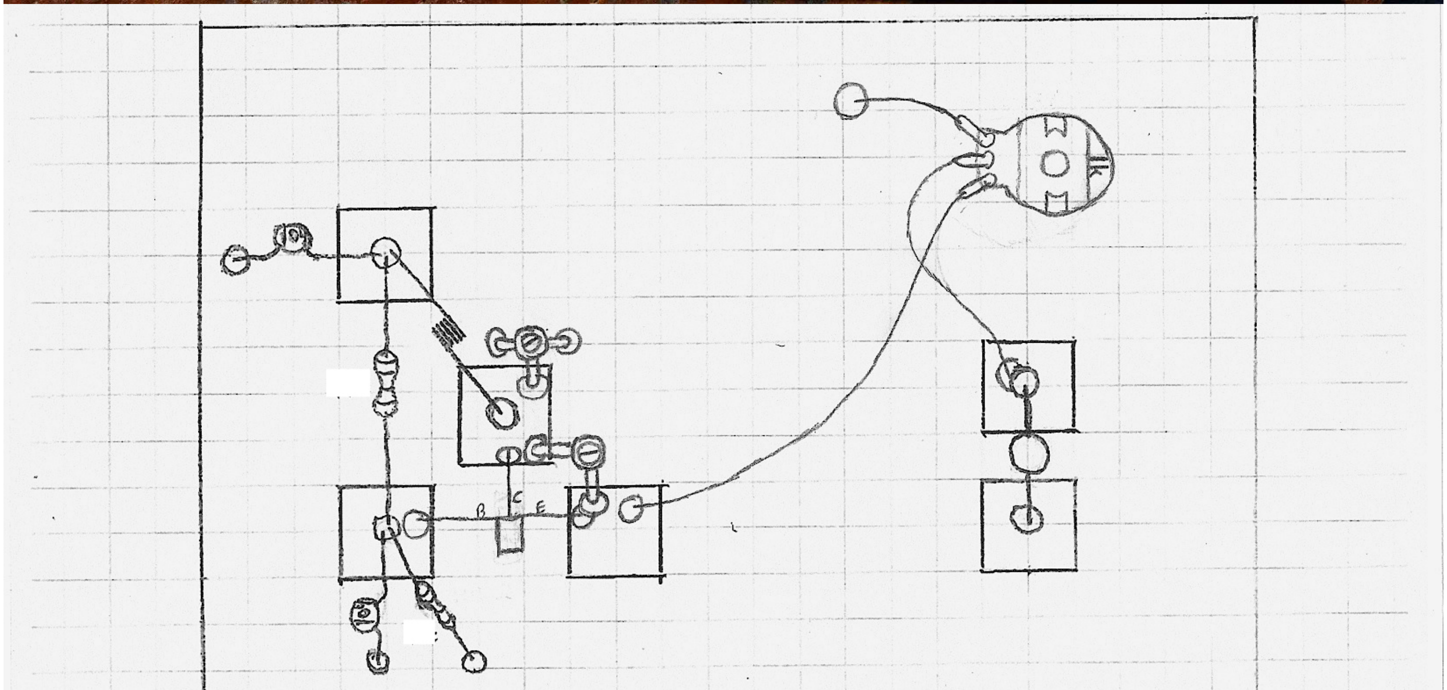

9. ADDING THE INPUT WIRES

- Cut two small lengths (~4-5”) of your stranded wire and strip both sides
- Solder one of the wires to the ground plate (this will be the ground input from your video source)
- Solder the other wire to the bottom square of the two right squares. (This will be the signal input from your video source)

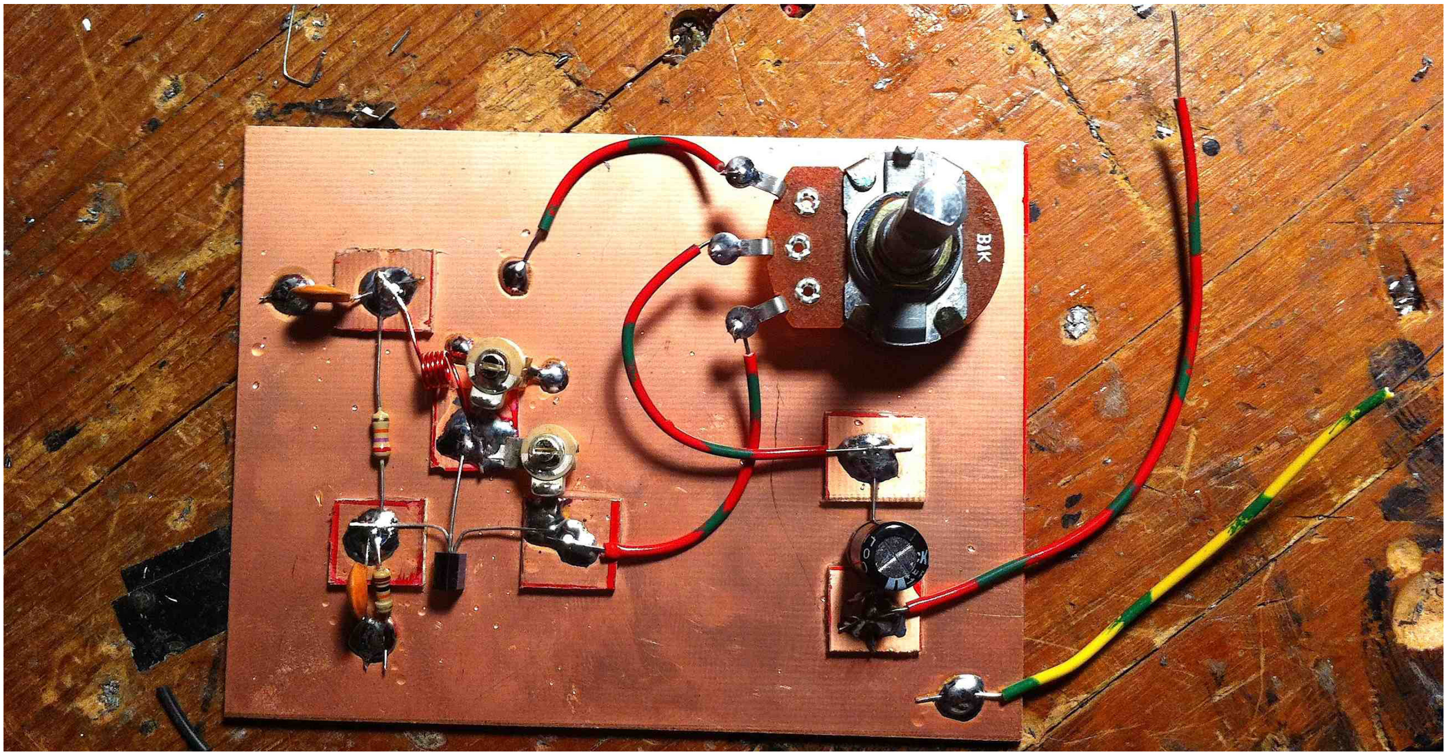

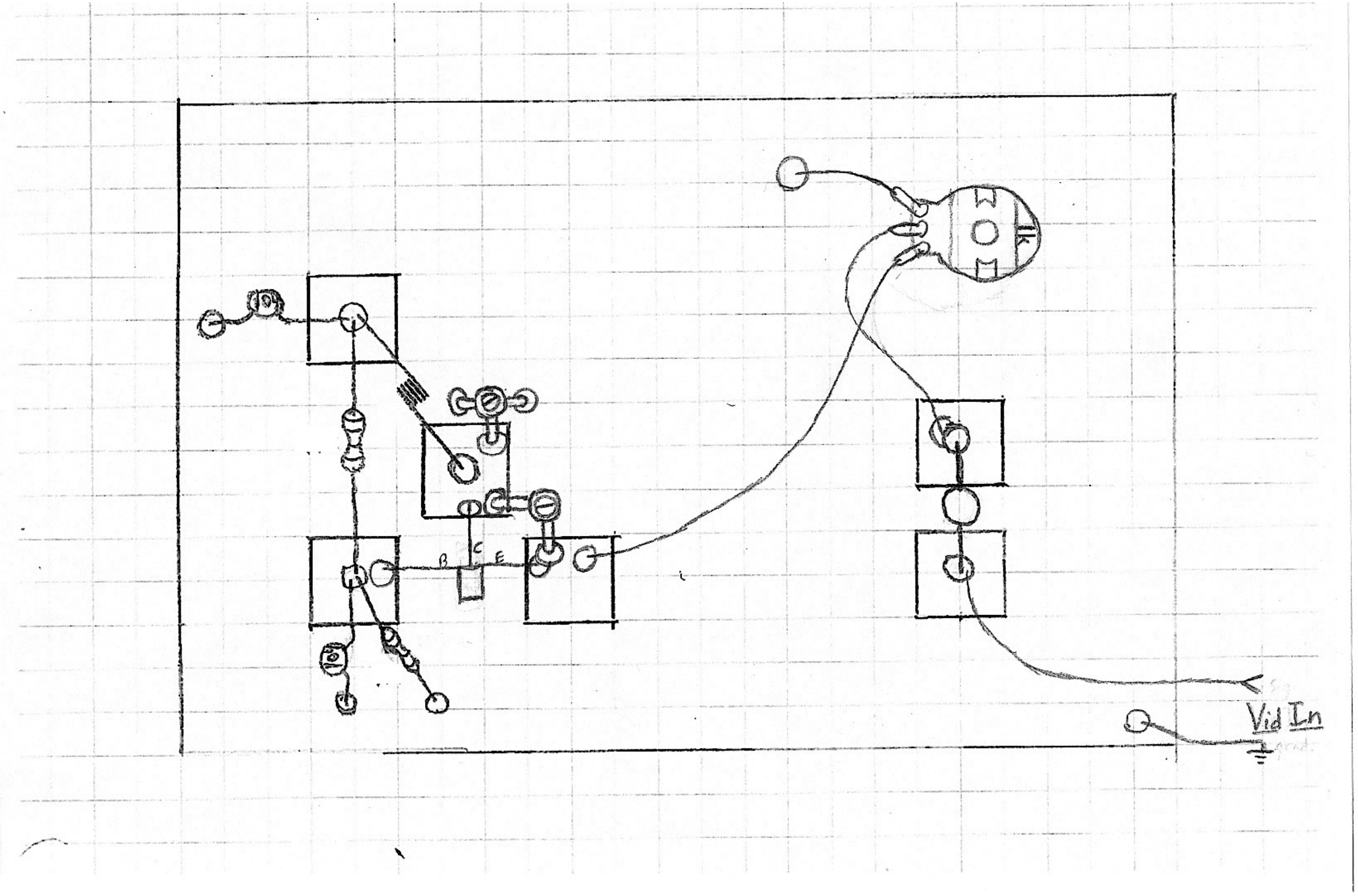

10. ADDING THE BATTERY INPUT WIRES

- Just like the last step, cut two shot pieces of wire and strip both ends.
- Solder one of the wires to the ground plate (this will be the negative input from your 9volt battery).
- Solder the other wire to the upper left square (this will be the positive input from your 9volt battery).

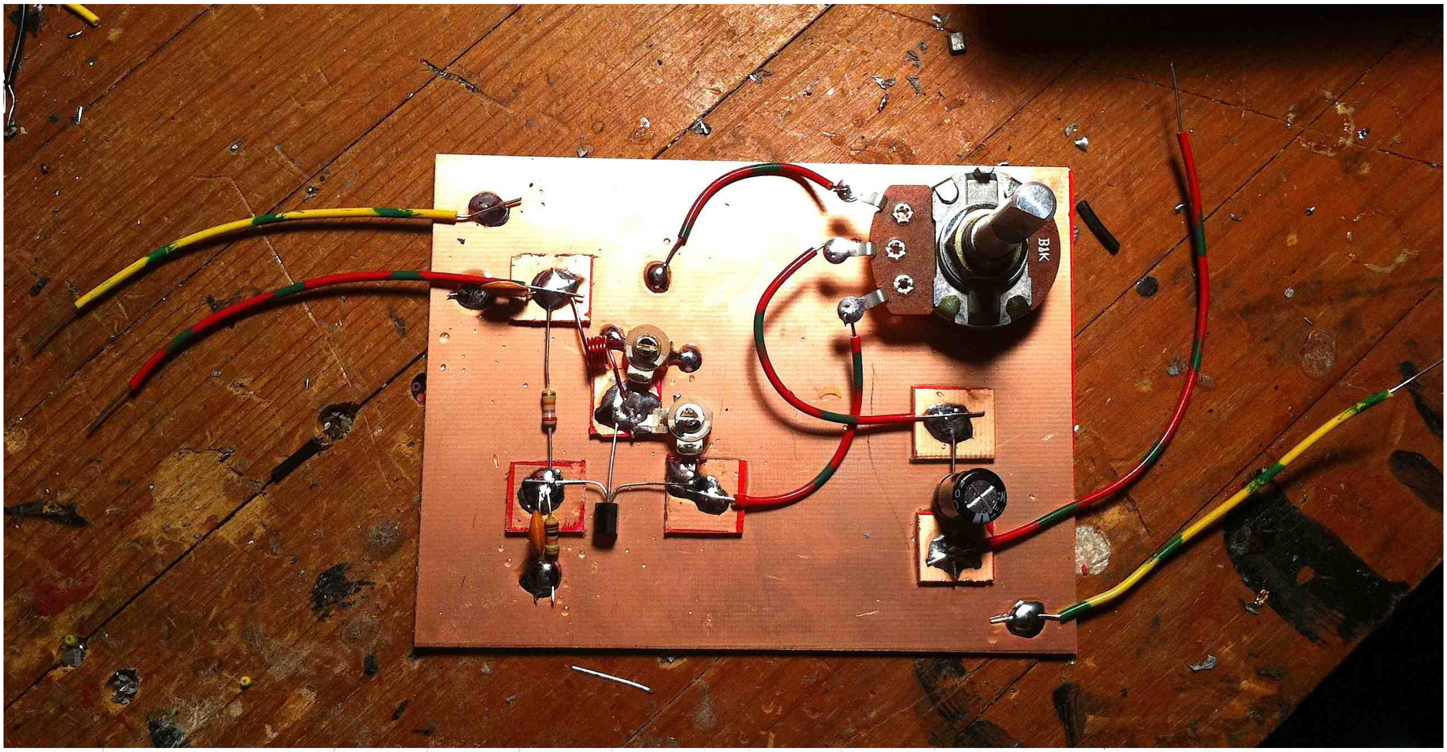

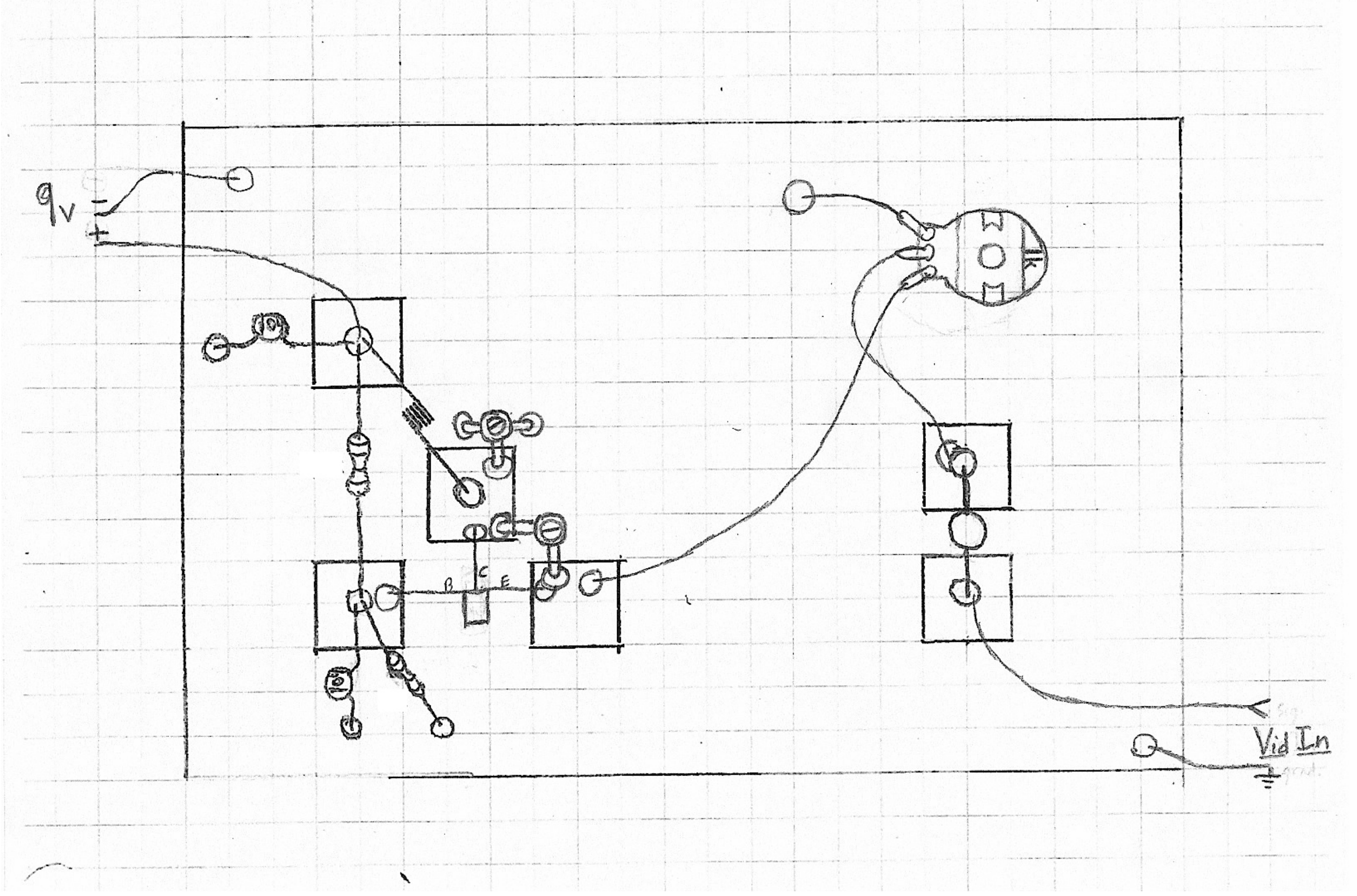

11. ADDING THE ANTENNA

- Cut about 2 feet of wire and strip one end.
- Solder the wire to the right most square of the triangle.

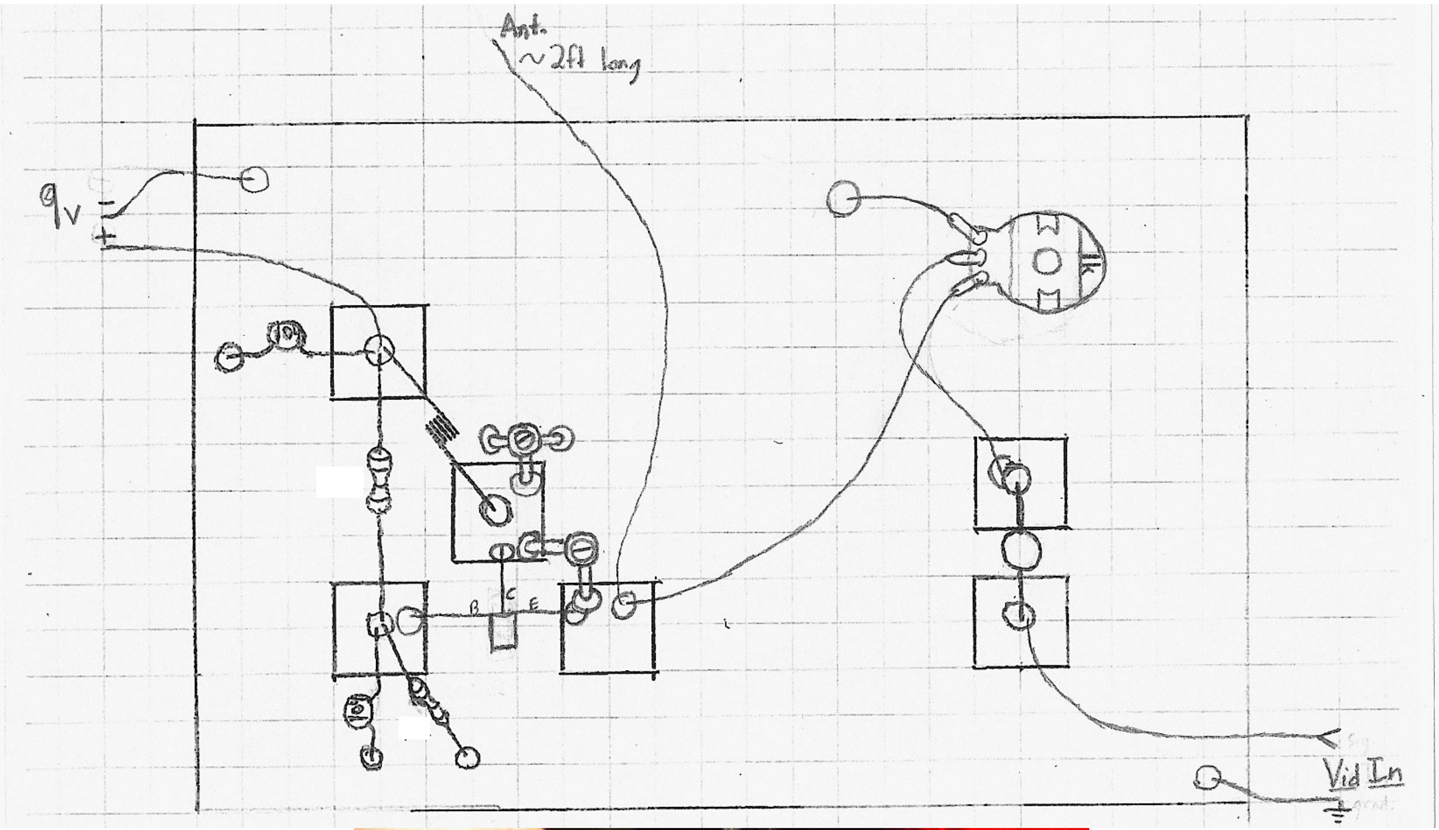
Ant.
~2ft long
9v
-
+
B
C
E
sig.
Vid In
gnd.

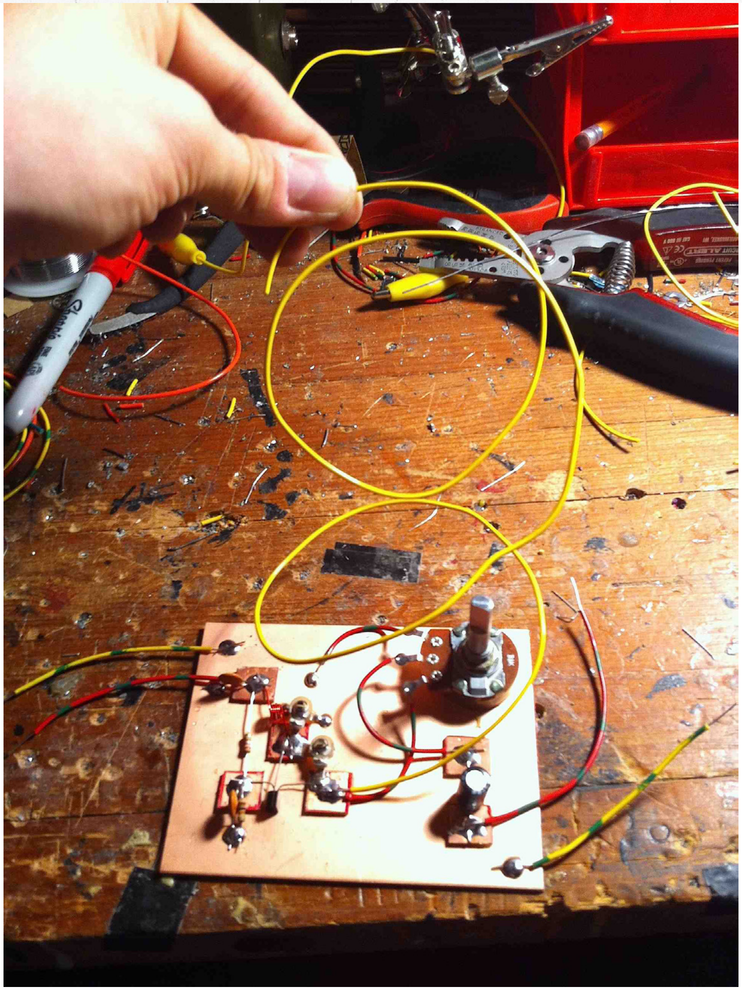

OPERATING

- Find a video source (DVD player, VHS player, iPod etc.) that has a composite video output.
- Use an alligator clip lead to connect the ground from your video source to the ground input of the transmitter. Using another alligator clip lead, attach the signal output from your source to the signal input of the transmitter.
- Use two more alligator clips to connect the positive and negative of a 9v battery to the positive and negative battery inputs of the transmitter.
- Turn on a CRT "tube style" TV and tune it to channel 2 or 3.
- Turn your potentiometer to somewhere around 50%
- Using a small screwdriver, tune the upper most variable cap slowly until you see the TV make a flicker or change.
- Tune the second variable capacitor until you start to see some strong signal coming through
- Tune the potentiometer and the two variable capacitors until you get the imagery you wan

Send any comments, questions or suggestions to yaktronix.online@gmail.com

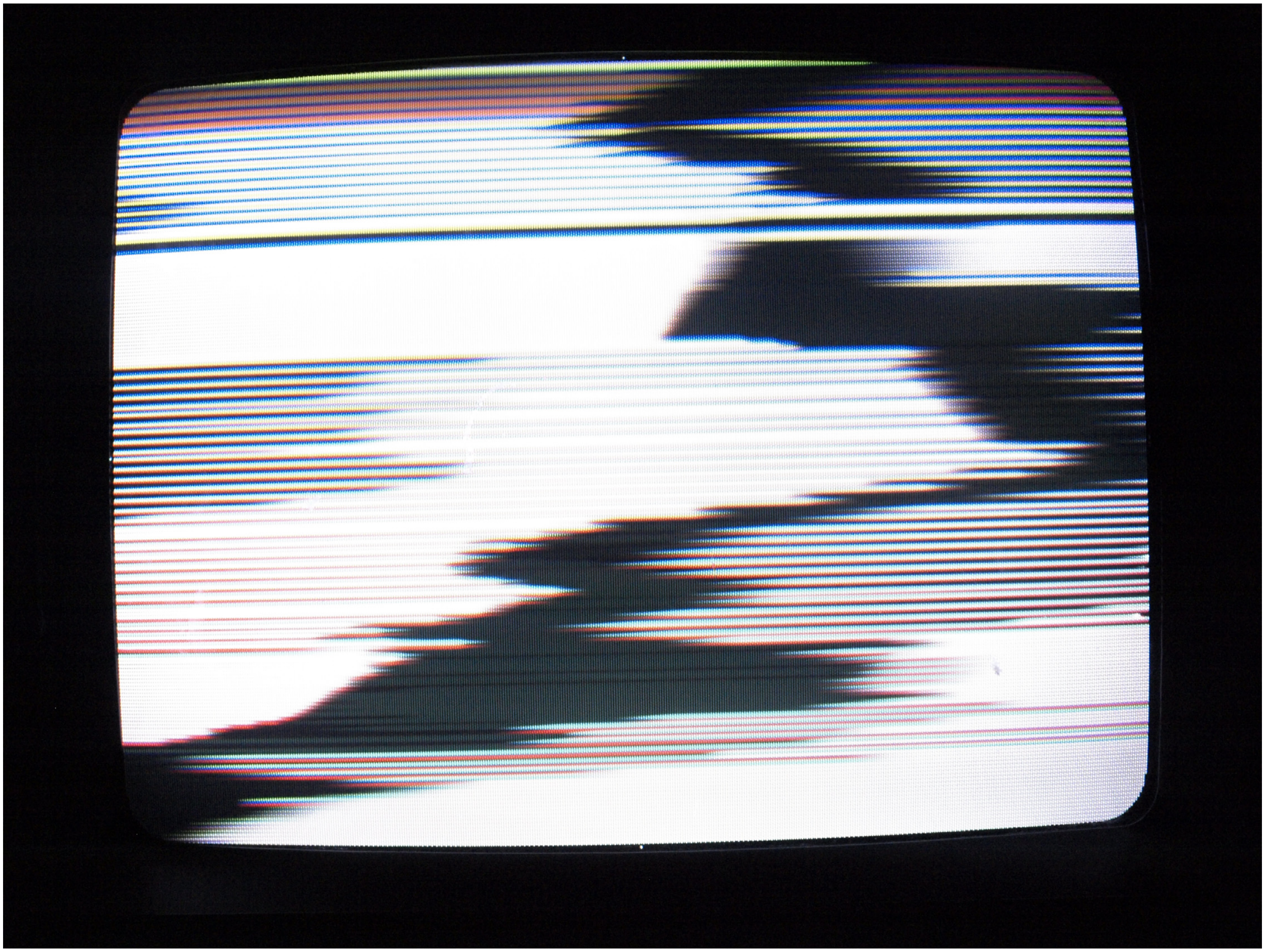

Fighting the pollution by public radio broadcast antennas

Un texte en français :	La pollution radio par les émetteurs FM
Een tekst in het Nederlands:	De radiovervuiling door FM-stralen
A petition:	www.petition-electrosmog.be

The electromagnetic pollution has now reached such a level that not only hypersensitive persons are being harmed. The debate most often focuses on cell phones, their relay antennas, Wi-Fi routers and similar technologies. It is often pointed out that domestic appliances like power saving lamps and cheap electronic devices, are harmful too. Recently, I came to the conclusion that I too had problems with electromagnetic radiations. Yet after a few months investigating, I was surprised to find out that far out the main cause of my problems was nothing I had heard about. A 100 MHz FM public radio broadcast antenna is located atop a local hospital and burns the whole city with an astounding force of 2 V/m (pike value).

I found other persons that are harmed. For example people who develop severe headaches or tiredness after a few hours working in a given place. They have no problems in other places, that are less exposed to the FM radio waves. It seems to me that people who live in strongly exposed habitations often develop psychiatric problems. My guess is that probably several thousand persons in the city are severely affected. Maybe tens of thousands are affected but without getting obviously sick.

Most puzzling is the fact that this hammering by radio waves is completely useless. It wouldn't change anything to the quality of the reception by the listeners in the city, if the force of the radio waves was divided by 100 or preferably 1,000. The purpose of those antennas is to emit far away in the countryside. That's why they are located atop hills or tall buildings and they have a huge power. But, due to their rudimentary construction, quite much of that power is wasted and hits the nearby city.

It is legal to hit the ground with such force. Whether I measure in rooms or in the open, it is at worst a little below the norms. Hence the problem is with the norms. They keep being too high, despite the efforts of many scholars and health organizations. The situation is catastrophic. If the people did understand the harm that is done to their children, revolts would break open.

Below is a picture of a cheap calculator that I transformed. It contains no more battery nor solar power cell. Instead it draws its energy from a 1 meter long antenna. Anyplace in the city center, in direct view of the emitter, the power of the radio waves is strong enough to feed the device. I made this because nobody was frightened by the figures I was reporting, except some friends that are into electronics. Everybody knows that a car can kill even at 30 km/h speed but almost nobody has any knowledge about radio waves figures. When I showed the calculator, everybody understood and some whitened. (The calculator also works in some classrooms where I tried it out. The emitter was visible through the windows... Should I warn the parents of the students that failed their exams?)

Powered by 000webhost

This didn't help anyway, so I invented another device: the "snake". The picture below shows six of them, an early version.

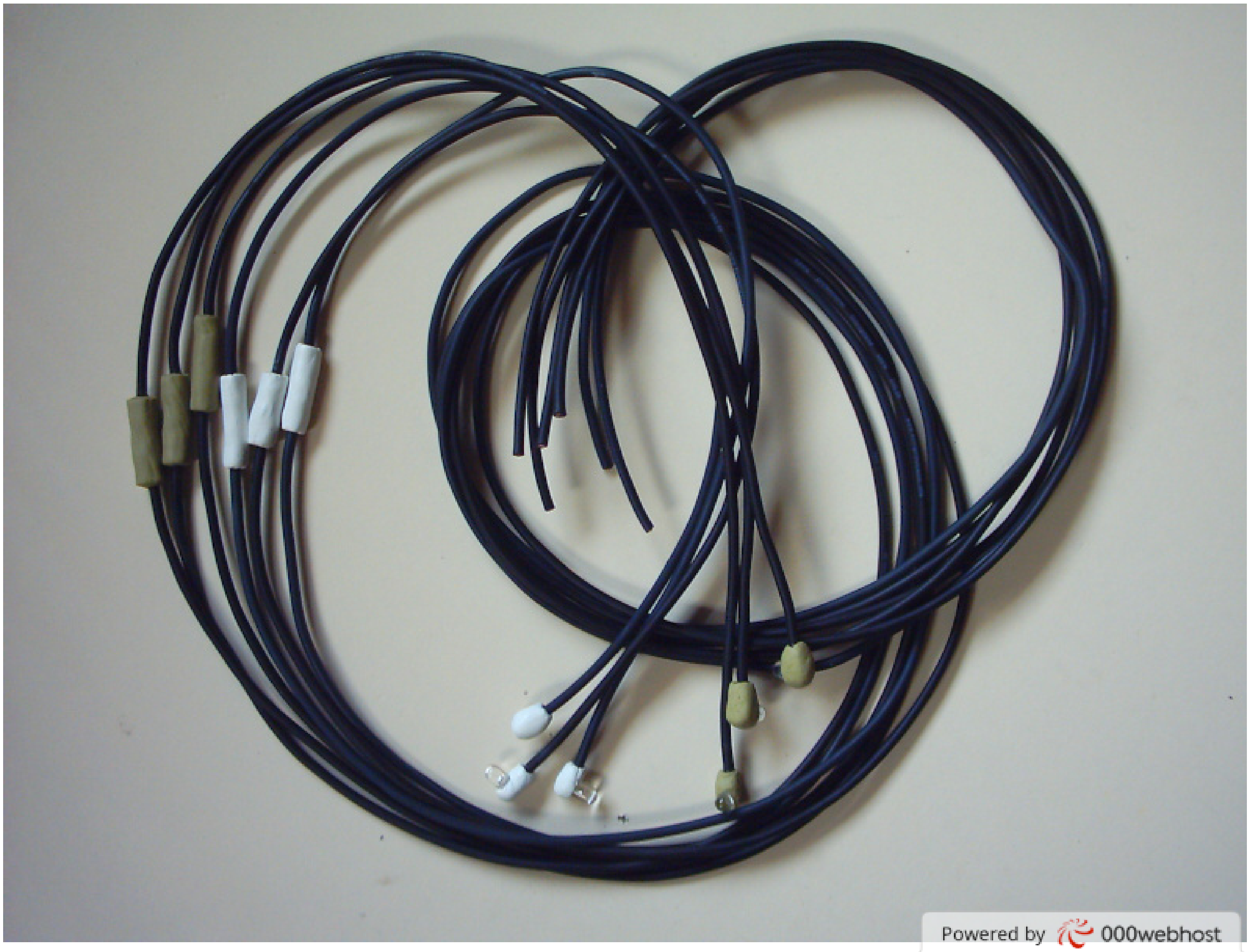

Powered by 000webhost

You just hang the snake in the flux of radio waves and the LED lamp at one end will bright up, clearly visible in daytime and enough to light and read a text at night.

The snake is less impressive than the calculator but it is much cheaper and easier to make and it can be sent in an envelope. I made tens of them and sent them to politicians, newspapers, universities... I gave some to local people, together with a user guide on a sheet of paper. Just like the calculator, the snake contains no battery. The lamp lights up solely on the power of the radio waves. The picture below shows a later version, that can be rolled up in an envelope that fits the conditions to be sent with only one stamp.

Powered by 000webhost

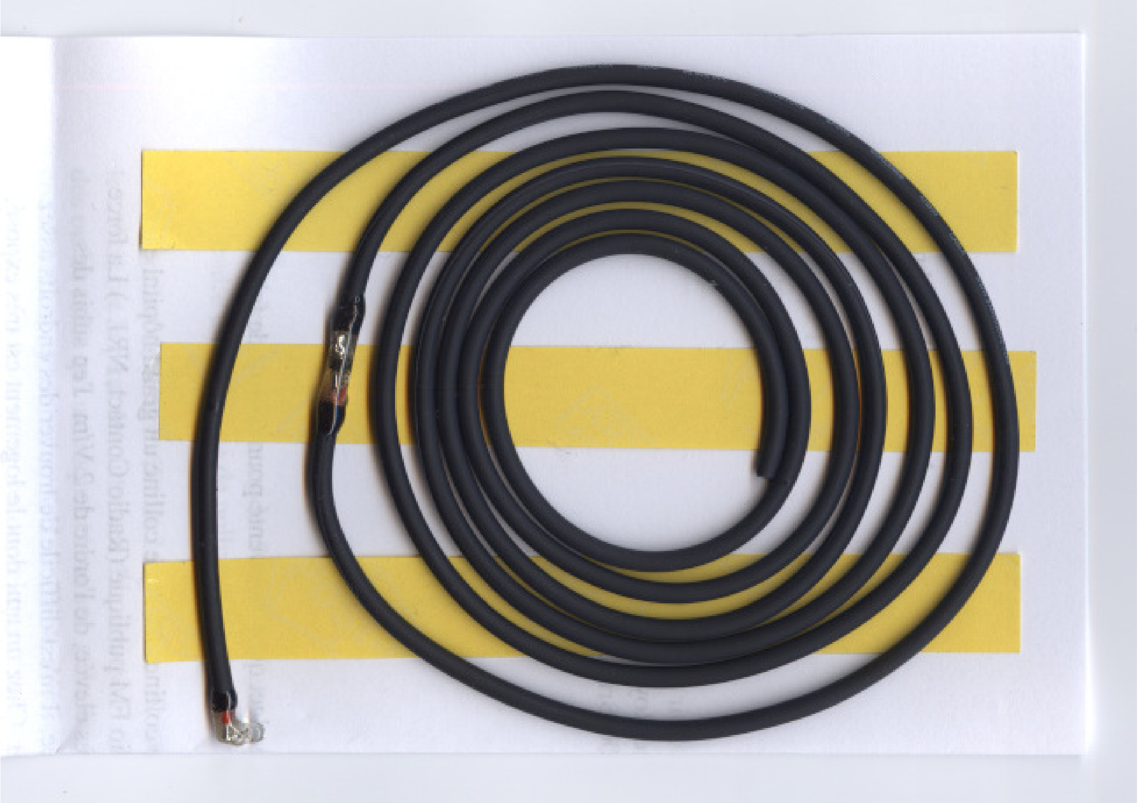

However mad it is to send such power on a city, nothing has been done yet against this. It's like an 19th century chemical plant sending its pollutants all around. Some complain but others argue that closing the factory would mean people laid off. Some then reply that being employed to feed children that die from ugly sicknesses may be nonsense... And so on, except for the fact that the chemical factory indeed had a purpose while the strength of the radio waves we're talking about, has no serious purpose at all.

Something common to all heavy sources of electromagnetic pollution is that it wouldn't cost much to make them harmless. For example, high tension power lines can be made harmless by feeding them DC current instead of AC current. This requires some more heavy electronics at the input and the output but the price of it is a detail in the whole. I studied the schematics of power saving lamps and simply noticed that they were conceived by people who do not understand the propagation of radio waves. Adding the necessary components to make the lamps radio-silent would almost not increase their prices. Enforcing proper norms will not change our way of life, it will just stop to favor the sociopathics amongst the industrials and lobbyists.

So, what can I do to help people understand the situation? They need to appropriate the scientific knowledge involved. Most assume that the necessary studies will be made by labs and then the governments will adapt the laws accordingly... That would be OK if it wasn't for the trillions of money involved. A minister in Belgium tried to prevent sugar dispensers in schools, she immediately got a made-up scandal on her shoulders. All she ultimately managed to get was to have some health advice notices glued on the sacred dispensers...

Flows of observations are now available, showing the effects of radio waves that match the norms: brains of rats destroyed, people unable to work, eggs that don't hatch, quarreling birds, increased risk of cancer, cows aging abnormally... I myself had seeds germinate and saw them die quicker when closer to a power saving lamp. Their strategy to cope with this is simple: "OK, there does exist scientific evidence of some danger... but there also does exist evidence that there is no danger at all! So let's balance the whole..." The norms are lowered but kept above what cell phone operators need for the current systems. The fact in itself of making a balance between scientific evidence is nonsense. Either a study proves a given danger and it can be verified, or bust. But anyway, what are those studies that tend to prove that there would be no danger? Lots of them are simply payed for by the cell phone industry... Yet some are perfectly serious, for example those concluding that radiations within the norms will not heat the brain or the body abnormally. This is perfectly true. There is no problem of scientific fairness involved with these studies. But, is a bullet harmless because it only very slightly heats your bones? A more sophisticated approach is to pick out studies that have borderline conclusions: "there may be a problem but it is not clear..." The study will be assumed to conclude that there is no problem at all, while it just was an inconclusive experiment like many are in science.

Powered by 000webhost

(When I'm exposed, I tend to compulsively perform simple tasks. What if a study examines if being exposed hampers the ability to perform that kind of tasks?)

Then, there is the problem that I'm going to make people sick. Indeed, electromagnetic hypersensitivity has a psychological component. Once people understand that they are victim to the radio waves and what symptoms the radio waves cause, they can overreact and become very sick when they start feeling the symptoms or just when they know there is a strong radiation force around. The radio waves induce a physiological stress, that hampers the proper functioning of the brain. When you add a psychological stress, due to the awareness of the problem, the summation of the two multiplies... Therapies for hypersensitive persons imply to dislearn them to overreact. Some friends of mine just refuse to start talking about electromagnetic sensitivity because they don't want to embark in the sickness. They do so with everything, meanwhile they eat very good food and have lots of pleasure in life. It works... with people with a strong health and a tad of selfishness.

When you are in a cold wind and you feel a pain in the neck, you know it is due to the wind and you take cover. Almost nobody can do this with the symptoms of an exposition to strong radio waves. You feel tired, some brain fog, like if your head was in a clamp, a headache, or simply you make stupid errors... You will blame yourself or the location, possibly the noisy people around... while in fact you are mainly undergoing electromagnetic sensitivity. The home of a friend is strongly exposed. After about twenty minutes I can nor more find my words and speak correctly, unless I wrap my head in aluminum foil. There is a fair chance that a while ago, I would have blamed my friend for being annoying (he is) to the point of making me speechless. There also is a fair chance that my friend is annoying because he hangs in that electromagnetic pollution everyday. Since he lives there, he became an alcoholic, lost his job, manages to keep no girlfriend... (Everything I mention here can have other causes. Quite often, such causes just add up. For example, the person blaming the noisy people around, would maybe have felt nothing if the noisy people weren't there. Like you don't feel your bad knee when you carry no bag.)

The impact of the electromagnetic pollution is severe. It is a far too heavy, and pointless, contribution to our burdens and losses. But how do you establish, scientifically, be it as an amateur, that there indeed is a problem? My initial approach can be an illustration of this. As my home is uninhabitable, I spend most of the day working in local libraries. I noticed that I could work quite well at some places and not at all at other places. Then I saw an article taped on a board, about the effects of cell phone and Wi-Fi radiations. Okay... I made a spreadsheet, with a row for each location and a note from 0 to 10 according to my ability to work there. Although I knew it was a weak approach, I then filled in each row the strength of the closest Wi-Fi router, as listed by my laptop. There was no correlation... but this proved not at all that the Wi-Fi was innocent. Weeks later I made a nonsensical but working little assembly that made the variability of the radio waves hearable. This time I got a correlation. Mild, but obvious. The ability to work is horizontally while the intensity of the radio signals is vertically:

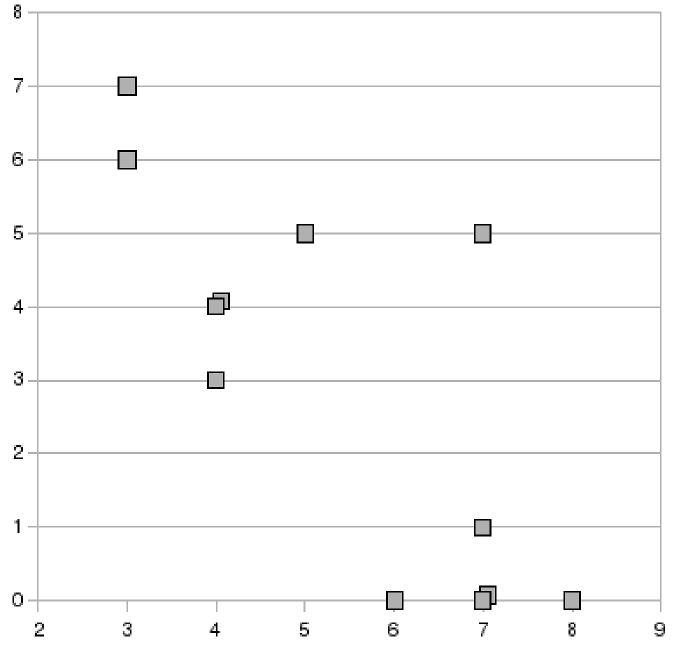

It quickly became obvious that I was measuring no strong signal in the GHz bands. I had no doubt that there were cell phone and Wi-Fi emissions coming from all directions and echoing all around but their strength was obviously

Powered by 000webhost

weak. The problem was with lower frequencies. I accused the fluorescent lamps for a while, till I finally understood that there was only one strong source, hitting me everywhere; those 100 MHz public radio broadcasts. The picture below shows the final correlation. Only one dot disagrees with the theory:

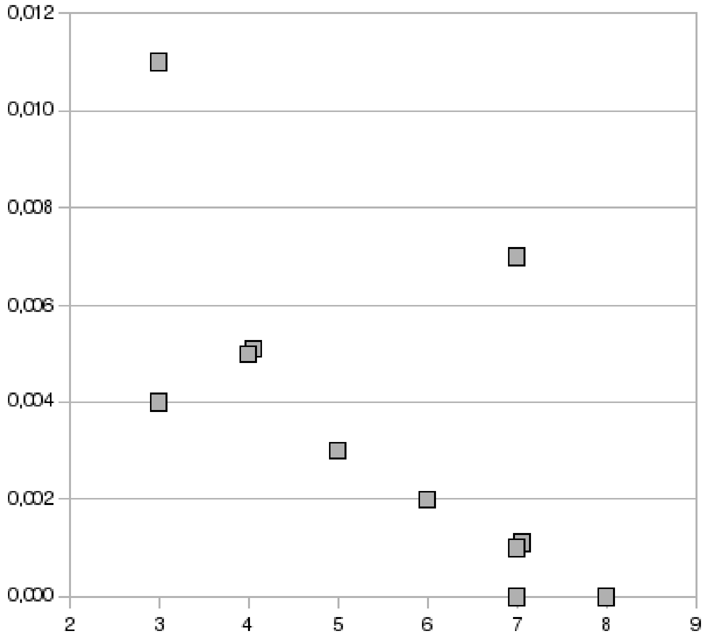

I was heavily disturbed by the radio waves, simply unable to work, long before I started puzzling on the subject. This is a strong proof that the matter is not just psychological. Reciprocally, once I understood the psychological component, I tried to use it positively, to counterbalance the effects of the electrosmog. I did get some results but I was never able to work correctly in the most exposed places. It is by no means "just psychological". It is useless "to be willing to make a small effort". I did efforts that frightened other people. One thing that does clearly help is eating a lot of sugar (Organic jam on rice bread...) (by the way, is that why sugar dispensers are placed everywhere kids are supposed to think and learn, to counterbalance the effects of the nearby cell phone relay antennas?) Alcohol does also help and I'm afraid this is why people seem to become alcoholics in exposed habitations.

Also, it does sometimes happen that I feel like undergoing strong radio waves but when I pop out my measuring device, I measure nothing serious. I have a natural tendency to be knocked-out... The radio waves are far out my heaviest problem but they are not always responsible.

I started studying other sources of radio pollution. I had an obvious benefit of removing every power saving lamp in my home. I put ferrite cores on most power chords. On one power chord that I sometimes use to feed my laptop inside my shielded tent, I had to put ten ferrites till I felt no more direct effect. Something very interesting is that till then I had no problem when using my cell phone. You guessed correctly: I started having problems, to the point that I never more hold it against my ear. I always use an earplug linked to the cell phone by a wire with two ferrite cores. I suppose that since I started avoiding the exposure to strong radio waves (which significantly improved my life and ability to work), my brains are no more constantly knocked out. So they get knocked out if I hold my cell phone against my head. And I feel it. This is one more synergy between the public broadcasts emitters and the cell phone industry: one hides the other away. 'Want to prove that inhabitants close to a relay antenna are no worse than those further away? Do the experiment in a city that is strongly exposed to public broadcasts...

So: buy a measuring tool or have a friend assemble one for you. And start comparing your problems and those of the people around you, with the level of electromagnetic pollution. If you can, make large-scale experiments with tens of persons, under advice of somebody with knowledge in statistics (to build an expressive statistic yet completely false, is *easy*). We need many such experiments, till it becomes unbearable for local interlocutors to negate the problem and not to dim the emissions. Step by step, we need to conquer our world back. Some enterprises and some countries have already set their norms at responsible levels.

One very important parameter is the rate at which the problems build up and then disappear, when exposed. Some people claim that they instantly feel when their cell phone is going to ring. I never could verify this. A friend claims that his cat reacts when he is going to get a text message on his cell phone. I tried several times to see the cat change

Powered by 000webhost

behavior while making my cell phone emit close to him but I got nothing. So I suppose that the cat reacts to the tone of the cell phone and my friend brain-confuses the timing. (I don't pretend nobody can feel a text message arriving, I'm just saying that I never could verify.) What I'm sure about is that my problems need minutes to build up, most often some ten or twenty minutes, the problem possibly becoming really painful after an hour. Once the problem built up, it needs much more time to diminish, maybe four times more. I often had to sleep a whole night on it. One day I "burned" my brains by holding a very polluting power saving lamp about 3 decimeters away from my head. I only noticed the problem after half an hour. But then it was horrible, like if a flame had been sweeping inside my head. Really just like the ache of a burn wound. I had a strong nausea and headache, that calmed down only hours later. I identified a few persons whose problems are obviously due to the electromagnetic pollution, in all cases the same applies: it needs time to build up. This means you cannot simply walk in and out of an exposed zone and claim you feel no difference. Also, you cannot get into a safe zone and claim you feel no better. If you were already poisoned, you will stay that way a long while. The only sure approach is to stay really long in a safe zone, preferably spend the night there, then go working in a polluted area. Try to be aware of what happens to you. The next day go working in a safe zone and try to see the difference. (This is a goldmine to build inconclusive studies.)

Everybody is sensitive to electromagnetic waves. The question is about the level of the sensibility. That's why people who get seriously sick with the current levels of exposition are called "hypersensitive". They are more sensitive than the average, sometimes with very painful and debilitating consequences.

There is a whole ladder of symptoms. This is roughly the ladder for me, exposed to the 100 MHz FM broadcasts:

0.01	V/m	No problem
0.1	V/m	No more able to concentrate and work
1	V/m	head in a clamp, maybe difficulties to talk
10	V/m	Tried it once for 2 hours, had to stay the whole next day in bed with an awful headache.

The serious problems begin for me at 0.1 V/m. Yet I know some people that wouldn't be called hypersensitive and that get problems similar to mine at about 0.3 V/m. A friend spent with me the two hours under the 10 V/m radiation and claimed he endured no problem... but he made surprising errors on the way back. It is very difficult for me, in my city, to find locations where the force is around 0.01 V/m. Values between 0.1 and 0.3 are common. It is easy to find building with rooms reaching 1 V/m. The force is quite predictable, according to the location and the orientation of the building, how much it emerges out of other buildings... So, because many people are being affected and because the radiation is reasonably easy to measure or to estimate, making statistics linking common health problems with the level of exposition is quite doable.

Hypersensitive persons would be 1% of the population. The need to protect this minority is enough to stop those powerful broadcasts from reaching the ground. I don't understand why I have to add that everybody is being harmed, be it to a lesser degree. I met official people whose job is to help hypersensitive persons. They've seen their pain and anguish by themselves... but they feel clueless as to how to stop the emissions. It's a very strange situation, when even the people close to the steer and convinced of a problem, can't do anything to stop it.

It has been claimed that the hypersensitivity of some persons is caused by an accumulation of toxic metals or organic pollutants. In such cases, antioxidants can be of immediate help and detoxifications would greatly help. (Note that common medical analysis like looking for mercury and the like in samples of blood, urine or hair, will reveal nothing even under severe intoxication, because blood and hair renew constantly. Only a sudden intoxication can be revealed by such analysis. The toxics slowly accumulated in standing body cells like the brain... Best method seem to be to take a well-tuned quantity of chelating molecules and then perform an analysis of the blood or urine.)

Cell phone and Wi-Fi radiations are different, for several reasons: they don't spread the same way in the body, they are pulsated and their frequency make they will not target the same macromolecules in the body. It would seem that you need 3 times more force of FM broadcast waves than cell phone or Wi-Fi waves, to have a same global impact...

I was frightened when I understood that under those 2 V/m force of radiation, an electric current in the order of 1 mA oscillates through the body. Such a current at a high frequency of 100 MHz will interact with the body quite differently than a DC or low frequency AC current. You cannot make direct comparisons... Anyway, 100 mA of DC current can be enough to kill a person by electrocution. 1 mA of low frequency current is enough to feel the current as a tickle. What are the consequences of such a current on the chemical reactions in the liver, on the production of

Powered by 000webhost

hormones by glands everywhere in the body? What if the person being exposed is a pregnant woman, with the baby just at the worst place in the flux of current? The length of an "antenna" matters much to allow it to build up current. Teenagers have the "optimal" length for 100 MHz radio waves... One thing I'm sure about is that if the government decided that from now on the school kids will be latched electrodes on the body and a current of 1 mA send through several hours a day, whatever the frequency, the parents will not accept. But this *is* happening, through radio waves, and I could feel the neuroleptic-like effect of it. (By the way, this implies that those experiments that show brains of rats destroyed by cell phones waves within the norms, apply only to a lesser degree to human brains, because of the difference in size. Rats are close to the optimal length to be harmed by cell phone waves... The optimal wavelength to dammage teenagers is that of those 100 MHz public broadcasts...)

The radiations harm me more inside buildings than outside. I may get more sick under 0.2 V/m in a room than under 2 V/m directly exposed to an antenna. I don't know exactly why. Here are four possibilities:

- The roofs atop buildings act like prisms and send the waves towards the inhabitants beneath. Radio waves coming from above would be more harmful than when coming from aside.
- The radio waves resonate and echo inside the building and build up in metallic structures. This creates halos of proximity that will inject more power inside nearby living bodies. For example, while simply wrapping aluminum foil around my head, it is obvious that some designs of shielding will make me even more sick than using no aluminum at all. The same way, when some parts of the shielding of my tent wear off and stop to be conducting, I can get serious problems till I replace or complement the faulty area.
- Buildings contain their own sources of electromagnetic pollution. They add their effects to those of the pollution from the outside...
- Close to metallic structures, the electromagnetic field can become rotating. This because waves with equivalent strength come from different directions with different phase shifts.

Again, the four possibilities mentioned above are potentially very suitable to build experiments that demonstrate there is no problem. Perform the experiment in an anechoic place with the waves coming from aside and no other radiation than the one from the experiment... Just the opposite from real world situations but you can claim to be serious at avoiding parasitic variables...

If you want to find the FM emitters somewhere, fmscan.org is the reference I used till now.

This is the schematic of the "snake":

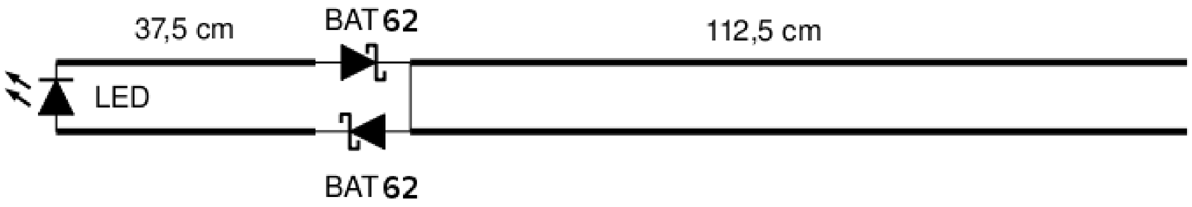

BAT62 detection diodes are no more produced. **BAT15** diodes work fine but they wear out; after a few months they stop functioning. **SMS7630** diodes are great but very little and mechanically fragile. **MMSD701T1G** diodes are sturdy and powerful; an excellent choice for a beginner. Such SMD diodes do also work for cell phone frequencies, which allows to test out a snake with a calling cell phone pushed against it. But any detection diodes that can manage 100 MHz will do.

The LED I'm currently using is the **L-7113SEC-H** . It lights up with a low tension and a very low current (the bluer a LED, the more tension it needs). Its color is red yet close to orange hence it is easily seen by the human eye (the eye is most sensitive to green, yellow and orange). The beam is quite narrow so when the LED is directed towards somebody's eyes it will appear quite bright.

For the lengths of 37.5 and 112.5 centimeters, any electric wire with two copper conductors will do. Audio signal wire is a practical solution. Use the shielding as one of the two conductors. The lengths of the two segments must not be precise. What matters is that the total length of the snake be 1.5 meters. Do not hesitate to try out if a little

Powered by 000webhost

longer or shorter snake gives better results.

A schematic of my current probes, that I connect to a standard multimeter, measuring Volts DC. The measure displayed by the multimeter must be multiplied by 10. When using a 200.0 mV scale, just read while forgetting the dot:

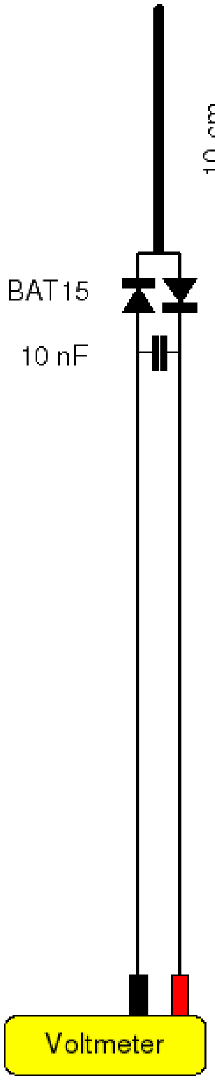

This is a picture of a device that tries to pack the whole in a neat gadget. The electronic circuit is a €7 digital Volts display with a scale of 200.0 mV. I soldered away the jumper that makes the dot be displayed. The antenna is only 7 cm long, so an RMS value is displayed (for roughly steady radio waves). The whole cost about €12. Hold the antenna perpendicular to the direction the waves come from and try to hold the device away from conducting objects like your body or metallic objects. You cannot closely trust such simple device yet it is very handy and adequate to compare the radiation level in different places:

Powered by 000webhost

Powered by 000webhost

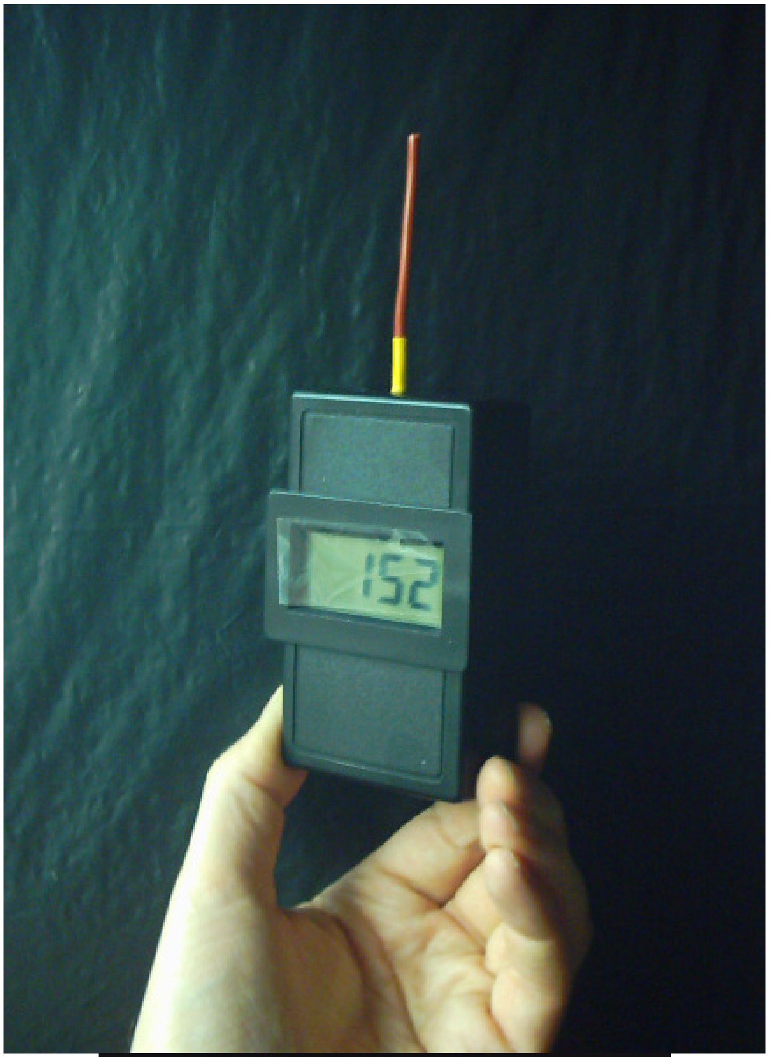

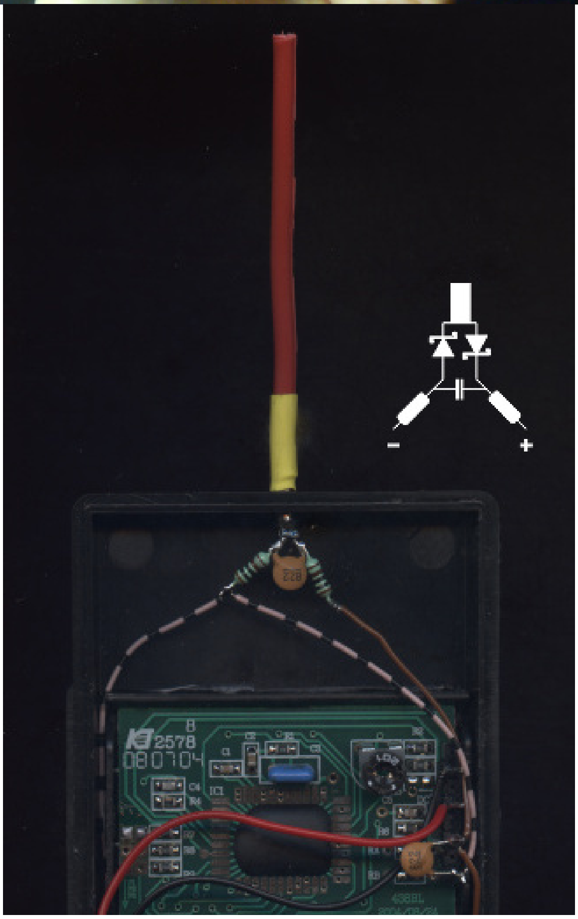

Powered by 000webhost

Those probes and device, quite easily allow to know the direction some 100 MHz radiation come from. Hold the thing at some decimeters from your body, roughly at the height of the belly. Your body and the device form a directional antenna... When the device is towards the emitter, it will display up to twice the value, while when it is hidden from the emitter by your body, the measure will be very low.

If you measure something strong, yet by turning in different directions the device always shows approximately the same value, then you are probably undergoing low-frequency radio waves like short wave broadcasts.

If the emitter is close, like a power saving lamp or a wrongly build power supply, converge towards it simply using the fact that the closer you get, the stronger the measure.

Such measuring tools are handy to find out power chords and appliances that pollute, yet then the measure displayed doesn't mean much. Once you are really close to an emitter, the measuring device and the emitter interact and the power flowing through the measuring device can be tremendous even though the device is only weakly polluting. You are in the "halo of proximity" of the device... Best example is most quality laptops, which emit almost no radio waves. I never could notice a problem while using my laptop when it is plugged to nothing... Yet if I hold a measuring tool against the screen I get frightening data... A few decimeters further, where my head can be, the measuring tool tells there is nothing... So, the high figures close to the laptop indeed tell that strong currents flow through it, but competent people made things such that those currents harm nobody...

If you want a global measure, away from the effects of your body, hold the device above your head. Or place it on some plastic or cardboard box and walk some distance away.

The windows are a common entry point of radio waves and electric wires can be awfull ducts, yet in some cases I noticed that metallic structures unrelated with electricity, like the copper tubes of water heating radiators, were the main ducts of radio pollution towards a room.

Below is a device I assembled to focus on 1.8 GHz cell phone radio waves. Close to the nearby cell phone relay antenna, in direct view of it, I could not measure much more than 0.1 V/m (pike value). While this is still too much, it is by no means comparable to the frightening 2 V/m intensity of the FM broadcasts. I also went measuring in places where people complain about problems similar to mine. They thought the reason was the nearby cell phone relay antennas. But I could measure no significant intensity of cell phone radio waves at all, while the force of the FM broadcasts was above 0.2 V/m...

Powered by 000webhost

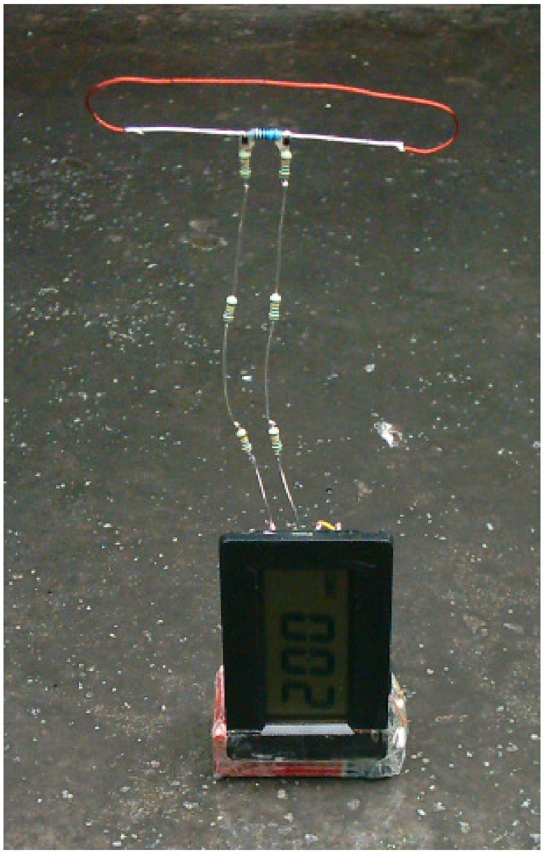

After about a year experimenting, the most simple protection I could come up with against 100 MHz broadcasts is a headband of aluminum, much the shape and size of a sweatband. I hide it inside a bonnet. I have no idea as to how and why it helps, aside from the general fact that any conducting structure will change something to the propagation of high frequency currents. Don't make it too high, a 3 to 4 centimeters height is OK. To try this out, you can just fold in a circle a 1 meter long wrap of aluminum and compress it to a flattened ring that fits your head. If you intent to re-use it, then you must strengthen it with lengths of tape before (inside) and after (outside) the ring is constituted. Make sure the tape does not hamper a perfect electric contact all around the ring. This protection is not perfect but it really helps.

When the field strength is above 0.1 V/m, I need a complementary vertical band in order to be able to work.

This also is a cheap way to make statistics. Compare days with the headband to days without the headband. In order to be scientific, you must ask somebody else to put the headband inside your bonnet every morning, at random. Every morning, the other person writes down if there was a conducting headband inside the bonnet or not. Every evening, you write a comment down to how your day was (headache? brain fog? errors? doomed day?) Then after a while you compare the two. At worst, if you have to do it alone, buy two identical bonnets, push a plastic headband inside one and an aluminum headband inside the other, and spend each day without knowing which one you picked. At the end of the day, write your appreciation down before checking and writing down which bonnet you picked.

Powered by 000webhost

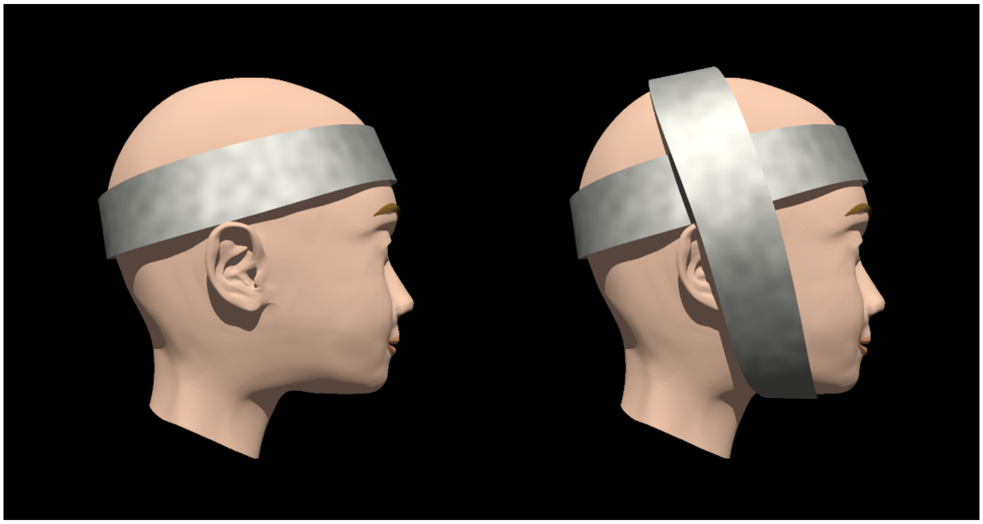

A friend who is a physician points out that in many administrations, be they public or private, the employees are being persecuted. (That often happens when the work done by the administration is mostly useless and only serves to justify the social position and the salaries of a few people.) A statistic about the health impact of radio waves in such administrations could be of few significance, because the people have all kinds of health problems due to the harassment. The addition to this by the radio waves may be comparatively negligible. (Yet another way to build a statistic that shows few or no impact of the radio waves: perform it in a place where people bow under humilations...)

I was quite frightened to discover that the symptoms caused by radio waves, as I see them around me, are very close to those of "radiation poisoning". This is part of nuclear medicine. When a victim is exposed to a short and strong burst of radioactivity or any other ionizing radiation, its body will quickly react like if it had been poisoned. A "low" poisoning by radiations yields headaches, tiredness, difficulties to think... while stronger poisoning can lead to the person die in a few days like if it was heavily burn. The exposure to ionizing radiations is measured in "Gray". 1 Gy leads to a low poisoning. So I computed out how much energy a person gets from the local FM emitter, like the person I know that get sick at their desk. The result is 0.01 Gy for 8 hours of exposition, which is quite close.

Radio waves can induce all the typical problems caused by that ionisinig radiations but the energy levels involved can be quite different. For example, ionizing radiations cause burns without increasing the temperature of the body, while radio waves can only burn if they are powerful enough to heat the body. Radio waves can induce cancer but again the energy required is ways higher, like spending ten years using a cell phone against the head, several hours a day.

Body cells can be attacked by all kinds of means. Ionizing radiations lead to direct damage of the cell membrane. Radio waves disturb the molecular cell gates operate. The cell will always react the same way, as for poisoning: it will close the cell gates, to try to shun itself from the outside world and from the harm. This has many consequences. The cell is less able to fulfill its purpose inside the body. It uses resources to try to protect its internal parts. It will accumulate waste. That's why people get tired, less efficient or even sick. When the aggression made to the cell is deemed too strong, the cell will auto-destruct. That's the way radiations and chemicals can be used against cancer, making the cancer cell suicide.

The cells can be helped to better recover from the state of shock. Also they can be made to go less likely into the state of shock. Medication and health advice can be helpful for this.

A link has been proposed between autism and radio pollution. The growth of the human brain is very complex, with cells traveling from one side of the brain towards the other to find their place. This involves a tremendous amount of communication and coordination between the cells. What if those cells are too often put into a state of shock?

These are two scientific publications about the link between autism and radio waves :

http://www.ncbi.nlm.nih.gov/pubmed/16530334

http://www.ncbi.nlm.nih.gov/pubmed/14962625

Powered by 000webhost

Publications about how radio waves disturb cells :

http://www.ncbi.nlm.nih.gov/pubmed/10860806

http://www.ncbi.nlm.nih.gov/pubmed/12379225

Links

http://www.teslabel.be
http://www.clag.be
http://www.001.be.cx
http://www.criirem.org
http://www.next-up.org
http://www.robindestoits.org
http://www.beperkdestraling.org
http://www.stopumts.nl

A petition: www.petition-electrosmog.be

Eric Brasseur - October 3 2009 till October 16 2015

Powered by 000webhost

▬▬▬ ▬▬●● ●●▬▬▬ ▬●▬● ●▬▬● ●●▬

SWR Bridge 1 to 1000 Mhz

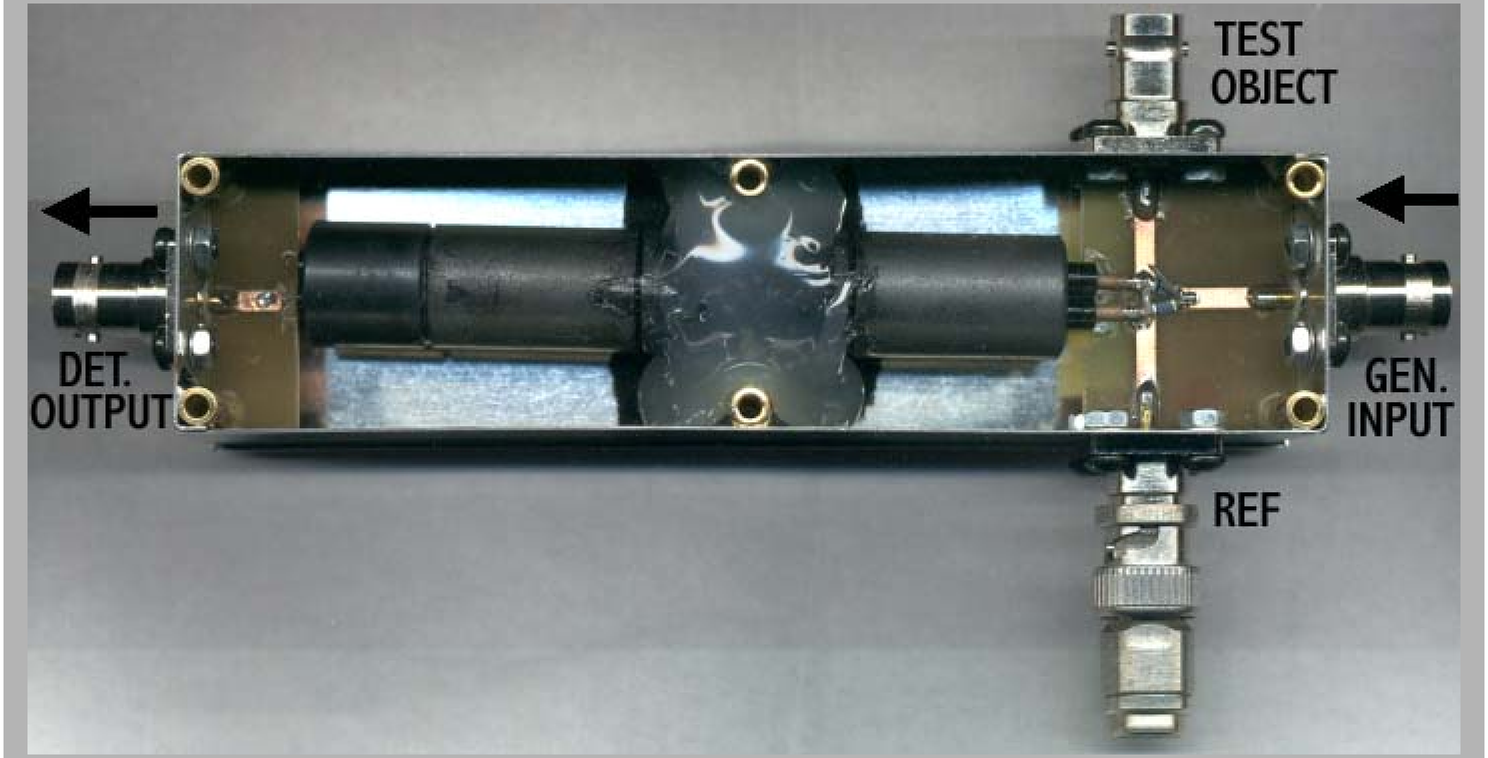
TEST
OBJECT
DET.
OUTPUT
GEN.
INPUT
REF

This SWR bridge is not my own invention, OZ3SW and OZ9MO has improved the design after DJ7VY construction in ukw-berichte 1/83

This is a construction set made for the members of Frederikssunds EDR Club, all I did was a bit of gluing and soldering

Thanks for all the hardware help, parts help, calibrating and measuring ! you know who I mean !!

It is possible to measure return loss up to 1000 Mhz, the return loss range is 0-50 to 0-40dB depending on frequency.

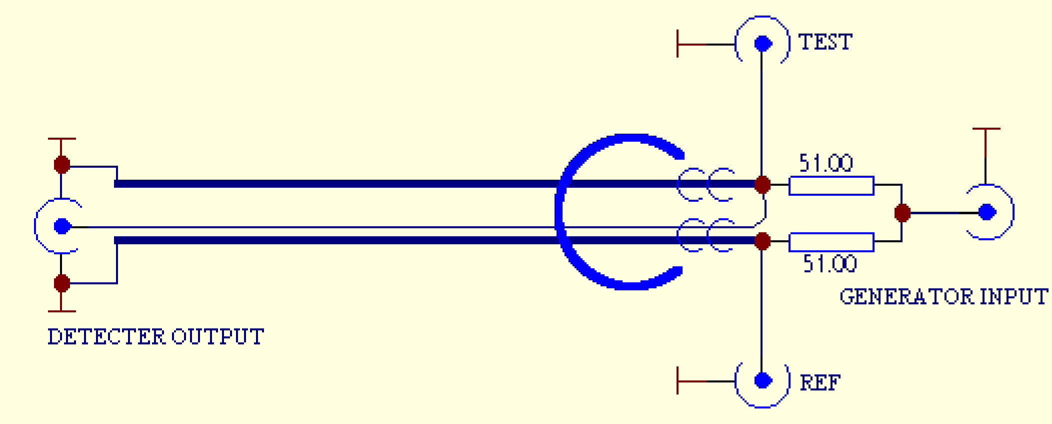
TEST
51.00
51.00
GENERATOR INPUT
DETECTER OUTPUT
REF

The schematic

The small toroids inside the 4 big ones, goes all the way inside,

They are glued with 10 sec. fast glue called cyano.

BACK

MICROWAVE MIND CONTROL

by Tim Rifat

UK INTELLIGENCE FORCES AND MICROWAVE MIND CONTROL
MIND CONTROL

As a scientist researching into microwave weapons used on the general public, evidence that the GM900 microwave network as used by Vodaphone and British Telecom, is a major health hazard, has come to my attention. Defence Intelligence Agency (DIA) documents entitled: Biological Effects Of Electromagnetic Radiation (Radiowaves and Microwaves) Eurasian Communist Countries, show that microwave frequencies similar to those of the cellular phones can cause health problems in the following areas:

* Blood.
* Cardiovascular System.
* Cells.
* Central Nervous System.
* Digestive System.
* Glands.
* Metabolism.
* Reproduction.
* Visual System.
* Internal Sound Perception.

Equipment to test the frequency and intensity of microwave phones shows that they produce signals similar to microwave ovens. Two Vodaphone cellular phones were producing 100mW/cm2 and 50 mW/cm2 respectively. The danger level for microwave ovens is 5mW/cm2 and the Russians regularly used 10 mW/cm2 for weapon research. These Vodaphones will therefore cause significant health problems if used. A variety of cell phones can be tested to see how dangerous they are.

The GM1800 system used by Orange uses higher frequency microwaves. Australian scientists found this caused cancer in mice exposed to radiation of this type. The documents in my possession show that microwaves pass deeply into the body and can have a negative effect on the brain and body.

The DIA research dates from 1976 and shows the dangers were known about over twenty years ago. The Soviets used the frequencies and intensities used by mobile phones, 1800 MHz and 900 MHz, as weapons. It is a rule of the intelligence community that you hide things in plain view, getting the public to accept microwave mind control weapons which effect their behaviour under the guise of mobile phones was a stroke of genius. Getting the public to pay for these microwave mind control devices, so their brains and behaviour can be damaged, to make them more docile and easy to control, was pure diabolical genius.

UK INTELLIGENCE FORCES AND MICROWAVE MIND CONTROL

Microwave weapons that turn people into stressed, confused, submissive zombies are being used in Britain‘s inner cities. Developed by the communists, microwave weapons similar to microwave ovens have since the 1980s, been targeted on inner city council estates. These weapons transmit extremely low frequency (ELF) signals which mimic natural brain waves; at the flick of a switch, all the people around these microwave transmitters are turned into submissive zombies who cannot think clearly, become depressed, apathetic and want to lounge around all day doing nothing: the inner city malaise found on Britain‘s streets. With the advent of new ELF detectors designed by the author‘s research teams, the mass mind control of the metropolitan UK population can be proved. ELF signals are officially only found at nuclear submarine communication stations, so the researchers were shocked to find microwave and UHF, mind control signals in city centres, the author‘s flat, being sent through mobile phones, the BT telephones and from the numerous transmitters that dot the country.

The massive increase in mobile phones has enabled the UK security forces to use this network of transmitters to beam mind control signals into the brains of anyone living near these transmitters. Microwave phones use pulse modulated microwaves of the correct intensity to pass through the skull into the brain and control behaviour. Microwave transmitters are therefore the perfect medium for the transmission of ELF signals to mind control the UK population.

Researchers are of the opinion that the inner city riots of the early eighties forced the Thatcher regime to deploy ELF mind control devices developed in the 1970s to turn the working class housing estates into total policing zones. In these zones, ELF transmitters turned the inhabitants into docile zombies. So successful was this technology that it was expanded to cover all major towns. Mass mind control of the UK public to make them submissive and obey authority was expanded, hand-in-hand with the mobile phone network and military and police microwave transmitters. Now the entire London conurbation is covered by UHF and microwave carrying mind numbing ELF.

Research into the use of microwave weapons and their use for mind control began in 1950s at the Tavistock Institute, one of Britain‛s leading psychiatric research establishments. The UK institute was researching into ways of mind controlling the British population without them knowing. The monkey submission response, whereby the dominant monkey caused submissive behaviour in the underlings, was the brain state of most interest to the British scientists. Having found this specific brain rhythm for docile submissive, zombie-like behaviour, it was then recorded and used as the template for the ELF signal beamed on UK microwave transmitters. Britain was the first discoverer of microwave technology, used for radar, in the 1940s and therefore had a commanding lead over everyone else in this field.

The 1970s brought a darker side to the story, with the news that the Russians were microwaving the US embassy in Moscow. One third of the staff eventually died of cancer from this microwave irradiation. Australian scientists proved in 1997, that tiny amounts of microwave produced by the very safest mobile phones, cause cancer in mice, when exposed to this radiation. Greenham Common was the first time that the Ministry of Defence was involved in the offensive microwaving of women protesters, the so called Greenham Common Women. This event showed that the UK security forces were willing to use lethal doses of microwave radiation on large groups of the general public. Equipment showed that the women peace protesters were subject to high levels of microwave radiation, some of them have subsequently died from cancer. Northern Ireland would have been a perfect proving ground for mass mind control technology. Research at the Tavistock institute enabled the Thatcher government to put into large scale use, the frequency which causes submission in humans. Radiating this ELF frequency on microwave transmitters, cellphones, BT phones and by use of other transmitters, the inner city population can be behaviourally controlled. My research group have ELF detectors, which can prove all manner of UK transmitters are being used to mind control the population. Since ELF is not found naturally, and is only used for nuclear submarine communication, its presence proves mind control weapons are being used on the general public. The author also has access to scientific documents showing the devastating effects of ELF broadcast by microwave, UHF and VHF on humans.

It is alleged that Marconi put this microwave technology into full scale production, and around thirty scientists and military personnel on the project who began to ask questions committed `suicide‛ under mysterious circumstances.

My research has found that microwave weapons are targeted on middle class troublemakers and researchers who cause problems for the establishment. Russian and American research has found that pulse modulated microwaves (as used for mobile phones) can, when modulated with ELF which mimics specific brain patterns change the behaviour of the victim at the flick of a switch. It has been found that UK security police, such as MI5 use the 450 MHz frequency used for this research (legally allowed to be used by the police) for behavioural control. A vast catalogue of mind control frequencies in the MHz range, FM radio, TV and mobile phone frequencies, have been measured, which are used in the UK for mind control and killing or disabling victims: 147, 153, 197, 199, 447, 453, 456, 466, 853, 883, 884, 887... Symptoms can be depression, befuddled thinking, loss of memory, stress, not being able to cope, manic behaviour, schizophrenia, nervous breakdowns, physical collapse, brain and nervous system damage, heart attacks, cancer...

An example was Brighton police van, parked outside the Brighton Pavilion which was used to beam UHF and microwaves at vagrants to clear them from the area. All Brighton‛s city centres are bombarded with microwaves when vagrants congregate to make them so ill they have to flee. The Evening Argus newspaper asked the police if the white van which was parked in the Brighton Pavilion grounds for months, was their van, they replied it was one of their ‚surveillance‛ vehicles. On Wednesday, 4th of February the author was passing Brighton Police station with a frequency counter which detects microwave weapons, and I detected a 452 MHz microwave beam being directed at two young offenders sitting on a wall opposite the station. I confronted the police with the reading, Robert Galloway, spokesperson for the police stated, „they were not interested in microwaves even if they were a danger to health.‟ The author has uncovered widespread use of microwave weapons in Brighton by the police and MI5 as a test programme to clear vagrants from the city centre.

Contacting researchers in this field of enquiry, they all complain of microwave-like symptoms, headache, nausea, giddiness, eye damage, ear problems... Readings I have taken show that the 750-1000 MHz range is used by the intelligence services for inducing nervous and physical collapse. Microwave ovens give off 1000 MHz.

On the Channel 4 programme, For the Love of…New World Order, shown on Monday night at 12pm, on April 6th, the author revealed some of these facts on television. The presenter, and researchers of this programme complained of severe headaches in the days preceding filming symptomatic of microwave attack. As the foremost non-military expert microwave weapons, remote viewing (psychic spying) and Psi-warfare, the author became aware of the microwave problem when he was targeted to stop his forthcoming book on remote viewing being released. This book on psychic spying, published by Century Books, is the first book by a scientist on the subject of Psi-warfare and the techniques the superpowers developed.

Subsequent to the filming of the Channel 4 programme, he was arrested by Brighton Police on February I5th outside their building and his frequency counter was confiscated to stop him recording the use of microwave weapons around the police station and in public places around the town. Even though no law was broken. It appears Brighton Police are worried that their use of microwave weapons on Brighton's vagrants is coming to light. The UK government has also bought up all the readily available frequency counters, these being the Watson FC-128 frequency counters, and the Optotronic devices, and the manufacturers now market a ‚modified' model, which most probably has a digital filter so that the microwave weapon frequencies given above, cannot be shown. The author's frequency counter has still not been returned. With such high-level manipulation of events it seems that microwave mind control is part of a secret policy which is being practised on the general public in this county. As part of a covert government policy of harassing and disabling critics, this means of attack is not covered by the law and highly deniable, the perfect intelligence device for civilian control. The author's flat is targeted with a 900 MHz beam which seems to be designed to cause massive neurological damage and produce tumours.

The BT phone network has the potential to be used for docilisation of the UK public. A 30-40 MHz signal is carried by the phone. When the earpiece is placed against the head, bone conduction carries the ELF component into the brain of the phone user. A ELF signal which can effect behaviour or health is therefore passed into all BT phone users. It appears from experts I have consulted on the subject that the digital phones have been designed to carry this mind control carrier frequency. Vodaphone have supplied a mobile phone to a friend of mine that pumps out an 847 MHz signal at over 100 mW/cm2, this is equivalent to 4x100 watt lightbulbs being turned on in the body of the user. This means that mobile phones can potentially be used to kill or so disable a person they no longer become a problem to the establishment. Another Vodaphone tested pumped out 50 mW/cm2. UK safe levels for microwave ovens are 5 mW/cm2 - which are not held near the head. There are reports by military researchers that 1 mW/cm2 will cause swelling of nerve cells exposed to microwave.

It is obvious that a total policing of the population by means of UHF, VHF, and microwave EM radiation, amongst others, is being used to keep the UK population in a docile and submissive mental state. Mental confusion can also be added to the signal for working class inner city areas to keep them docile and confused. A land of stultified zombies, who are mind controlled to make them docile and unable to think clearly. Subversives are habitually targeted with microwaves to drive them mad or make them fatally ill

On July 4th 1976 seven giant transmitters in the Ukraine, powered by the Chernobyl nuclear facility, pumped a 100 megawatt radio frequency at the West, which contained a 10 Hz ELF mind control frequency. According to a US scientist, Dr. Andrija Puharich, MD, the soviet pulses covered the human brain frequencies. With a Dr. Bob Beck Andrija proved that the Soviet transmissions were a weapon. He found that a 6.66 Hz frequency would cause depression and an 11 Hz signal would cause manic and riotous behaviour. transmissions could indeed entrain the human brain, and thereby induce behavioural modification populations can be mind controlled en masse by ELF transmissions. More importantly he found that an ELF signal could cause cancer at the flick of a switch. It did this by modifying the function of the RNA transferases so that amino acid sequences are scrambled and produce unnatural proteins.

MIND CONTROL

Torture is alive and well in the UK. MI5, the UK's secret police, regularly use Non-Lethal Weapons on any dissidents. Since MI5 have a well documented history of hating the Labour Party, and were instrumental in bringing down the Labour Prime Minister Harold Wilson, they naturally took to spying on, and discrediting, any group or person who did not follow their rabid right wing dogma. When Margaret Thatcher became Prime Minister of a right wing Conservative Government in 1979, she soon saw the power in recruiting MI5 to be her own secret police. During the Miner's strike in the early eighties, Arthur Scargill and his National Union of Miners nearly brought the Conservative Government down, with their year long strike. Margaret Thatcher used MI5 to spy on the Miners and to discredit Arthur Scargill. Leading police officers argued against the use of MI5 on the civilian population, they were ignored or replaced - if they proved too vociferous.

Margaret Thatcher was the first Prime Minister for many a year to sit in on the Joint Intelligence Committee�much meetings on a regular basis. Thatcher, as a trained scientist, would have been well aware of the usefulness of microwave weapons. They are impossible to detect unless you have a detector, dissidents have no idea these weapons exist, and best of all, they are totally deniable. It may be that MI5 kept this research secret from Margaret Thatcher, but the resources the UK has put into these fearsome, so-called non-lethal weapons, is extensive. UK intelligence runs a fleet of microwave weapon carrying vans, as well as portable microwave weapons that can be deployed near the dissident�much home. The vast expense can be disguised as communications equipment, for as we all know, microwave telephone communication is all the rage in our modern world. My research shows that the microwave telephone network also has the potential to be used as a major mind control weapon system to control the behaviour of the microwave phone users.

Developed under the Conservative government, which was in power from 1979-97. The UK secret, or ‚black‛ government, has at its disposal a fearsome array of mind control weapons. These abhorrent devices are euphemistically called ‚non-lethal weapons‛ by the UK military. In fact they kill you slowly by causing: nerve damage, cancers, mental collapse leading to suicides, or tissue failure, such as heart attacks due to the cooking effect of microwaves.

Precisely modulated microwave radiation is used to influence brain function. Human behaviour and reactions can be entirely controlled by using pulse modulated microwave EM radiation. Pulse modulated microwaves are useful as the carrier for the mind control signals as they are able to pass through the skull, which is rather resistant to low level EM. The massive number of microwave antennae that dot the country, some of which are used for the microwave phone network, all use pulse modulated microwaves, which makes their use for a strategic mind control device against the civilian population in times of trouble, or rioting, crucial to modifying the behaviour of the general population. In modern democracies it is no longer viable to shoot rioters, or torture dissidents by normal means, as the bad publicity is self defeating. Thus, microwave weapons have been developed by the UK‛s military intelligence as they leave no marks, or gaping wounds. This pulse modulated microwave carrier beam can then be used to carry signals. These signals are extremely low frequency recordings of brain electrical potentials, which have been recorded by neuro-medical researchers such as Dr Ross Adey.

Dr Ross Adey‛s research at the Brain Research Institute of the University of California, was funded by the CIA. In their Pandora project a catalogue of different brain signals for specific actions, emotions and pathological states of mind were recorded. It was found that when microwaves were used to fire these signals at victims‛ brains, they experienced the moods, behaviour, and the pathological states, carried by the signals. This meant that by mimicking natural brain frequencies, the human brain could be controlled remotely by use of extremely low frequency broadcast carried by pulse modulated microwave beams (ELF pulse modulated microwave remote mind control technology).

It is now possible to broadcast mind control commands directly into the brain by use of microwave beams. All that is needed is a catalogue of every specific brain frequency for each: mood, action and thought. These catalogues of excitation potentials are available from Russian neuro-medical research institutes, so any one with enough cash can have the same technology at their disposal, as UK military mind control groups (the psychological-enforcement arm of MI5). The Aum sect bought microwave weaponry via their 50,000 Russian converts.

Particular excitation potential, is then broadcast by pulse modulated microwave transmitter. This pulse modulated microwave beam has the ELF excitation potential frequency imprinted upon it. It was found that each behavioural set in humans had a distinctive frequency. There was one for: anger, suicide, hysteria, trauma, serial killing, paranoia, lust...etc. Intelligence operatives in the UK regularly park microwave transmitters outside targets‛ houses and beam specific mood inducing excitation potentials at the victims. To aid them, they have sophisticated millimetre wave scanners to look through the victims‛ walls, so they can see the targets‛ in their homes. Pulse modulated microwaves are regularly directed at the victims‛ brains, while other people in their homes are oblivious of what is going on. A leading conspiracy researcher, who looks into GCHQ at Cheltenham, found one of these vans, with two spheres on its roof parked in his road. When he took the number plate and through contacts checked who owned it, he found it was an MOD van. Hopefully he has found the microwave weapon system before it has done him too much harm.

These microwave weapons were developed allegedly at Marconi. When firing microwave beams through walls at one specific target, every material in the way of the microwave beam attenuates or modifies the intensity and frequency of the beam. Since precise frequencies and intensities are needed for mind control, very sophisticated microwave arrays and computer programmes had to be developed so that the microwave beam could be changed in response to the materials which lay between the target and the weapon, as the victim moved around the house. To do this, the reflectivity and refractivity of the materials between victim and weapon had to be analysed in real-time and fed to a computer which could change the microwave array in concert with the changing environment between victim and weapon, as the target moved around his/her home. Secondly, there had to be an automatic interrupter if another person walked in front of the beam. The victim needed to be driven mad or disabled, without anyone else being aware that he or she was being targeted. The technology for this was very complex but eventually it was perfected. Twenty five or more scientists and military personnel, associated with the Marconi project, then died in mysterious circumstances. Intelligence personnel regularly kill people to keep them quiet. Maybe they killed the entire research team to keep such a diabolical weapon secret. For if it were made public, the scandal would bring down the government. Whatever the real story, by the mid to late eighties, all the problems had been ironed out and these new smart microwave weapons were deployed on the UK s streets. Northern Ireland would have proved the perfect place to test them out. Pulse modulated microwave weapons had now come of age.

By this method, any mood or behavioural set, can be conditioned into the target‵s brain. Intelligence agents keep a log of the victim‵s behaviour to see if more intense ‚treatment‵ is needed and as a guide for future mind control projects. It is alleged that by this method, UFO and conspiracy researchers, are routinely driven to commit suicide. having the excitation potential for suicide beamed into your brain day and night by microwave mind control weapons, soon resets the brain into a cycle of depression that spirals out of control ending in suicide. Many intelligence and technical officers in the UK, who have spilled the beans, or could be a potential leak, are driven to commit suicide by the special mind control teams run by MI5. If you look at the long list of UK military and intelligence experts who inexplicably commit suicide, one can see an underlying mind control logic that drives them to kill themselves.

Behavioural reinforcement is used in a synergistic way with the mind control. It was found that the effect of the microwave beams could be greatly enhanced by external reinforcement. Intelligence community personnel destroy the target‵s property, ruin their financial affairs spread vicious rumour, make sexual peccadilloes public knowledge - such as wearing women‵s underwear, while checking oneself - with an orange in one‵s mouth. External reinforcement of pulse modulated microwave mind control technology was found to be very effective. Intelligence chiefs are now in seventh heaven; if some one becomes a problem they get the ‚suicide mind control team‵ parked outside their house. Within weeks, the victim kills himself. This is very pleasing to the intelligence mandarins, as suicides are easy to explain away - even if the victim was a highly placed politician or military man.

If the military intelligence agency does not wish you to commit suicide, they can drive you mad. This is done by beaming the excitation potential of a particular pathological mental state at your brain while you are at home. To aid in this, the intelligence operatives can place sounds and speech in the target victim‵s brain. This inter cerebral hearing is used to drive the victim mad, as no one else can hear the voices transmitted into the brain of the target. Transmission of auditory data directly into the targets‵ brains using microwave carrier beams is now common practise. Instead of using excitation potentials, one uses a transducer to modify the spoken word into ELF audiograms, that are then superimposed on the pulse modulated microwave beam.

Discrediting well known people who are causing problems for the shadowy elite, by driving them mad, seems to be standard operating procedure for the intelligence community. Victims are subject to pulse modulated microwaves which carry different types of madness and behavioural aberrations, encoded as ELF excitation potentials. This makes the troublesome high profile person, display manic or insane behaviour that discredits them. Examples of this technique are allegedly: David Icke, Fergie, Princess Diana... Outside environmental reinforcement, by use of media agents in league with MI5, makes it assured that the high profile person‵s mind controlled madness, will be put in the worst possible light to discredit them. In this way, high profile subversives who cannot be wrongly imprisoned (unlike the Carl Bridgewater prisoners, Birmingham Six, Guilford Four...) in the normal way by Britain‵s police, or killed by assassination squads, such as the Pegasus group, are made to look ridiculous. Public humiliation is the finest weapon the authorities have, to make harmless, a potential well known figure, who is causing trouble.

If the VIP needs to be made temporarily ill, microwave beams containing the signal the brain gives off during a vicious bout of flu can be fired at the victim. This causes the target to display all the symptoms of flu, even though they have not caught the virus. Major Ed Dames, the remote viewing specialist, who has close links with the US secret military, alluded to this device on a US TV programme.

The intelligence agents can also use low level microwaves to cause mental and physical confusion that leads to illness. Beaming microwaves at victims makes them fatigued, damages their immune system, causes neurological damage that effects their thinking, and ability to carry out tasks, induces premature ageing, cancer and cataracts. Sussex Police regularly use this low level microwave to clear drunks from city centres. While researching this matter, I was subject to high levels of microwave >10mW/cm2 whenever I approached their headquarters, whilst MI5 irradiated my home continually with microwaves. It seems that I touched a raw nerve while researching mind control weapon usage in the UK.

Organisations that irritate the authorities have their building turned into a ‚hot spot' by microwave transmitters, so the staff all suffer sick building syndrome caused by microwave damage. Or the staff of the target organisation have their behaviour changed to cause discord. UFO Reality, a leading UFO and conspiracy magazine have complained of mysterious health problems, while another New Age group complained of eye damage, nausea and headaches brought on whenever they had a meeting, which may have been caused by microwaves.

Heating the victim to death, by microwave cooking is caused by increasing the field intensity of the radiation, to cause local hot spots in the victims' eyes and gall bladder, which have poor circulation, so cannot carry away the heat.

Irradiating the optic nerve of the victim with the same signal that is sent to the brain by this nerve, causes the nerve tissue to overload. In this way, subversives can be blinded by the intelligence community without them knowing what has occurred. An American researcher complained of this problem to me and warned me that this weapon system was being used on my person by MI5.

Neurological research has found that the brain has specific frequencies for each voluntary movement called preparatory sets. When you pick up an object, there is a specific preparatory set for this action. By firing at your chest a microwave beam containing the ELF signals given off by the heart, this organ can be put into a chaotic state, the so called heart attack. In this way, high profile leaders of political parties, who are prone to heart attacks, can be killed off -before they cause any trouble. Neil Kinnock's Labour government was allegedly cheated out of an election victory by postal vote rigging in twenty key marginal seats. When a new even more electable Labour leader was found, it is rumoured that John Smith, the then Labour leader, was prompted to have a fatal heart attack, while walking in the country with his family, by means of a concealed microwave device which operated on the Vagus nerve to bring about a massive heart attack. Since MI5 have a long history of naked hatred toward the Labour Party, there may be some truth in the above, though no hard evidence has yet been found.

Paralysis can be induced in the target by use of this method of broadcasting preparatory sets encoded on microwave beams. A pulse modulated microwave beam, carrying an ELF signal, which is identical to the one in the motor neurone centre of the brain, is used to jam the victim's motor co-ordination. This is analogous to radar jamming, using a more powerful signal at the same frequency to swamp out the enemy's radar. Motor neurone preparatory set potentials are jammed by a bigger signal carried by a microwave carrier beam, that literally overloads the brain, so it cannot control the body. Pulse modulated microwave weapons which broadcast the ELF preparatory sets of the motor cortex at the victim, will paralyse the victim without killing them. Breathing and heartbeat are involuntary actions controlled by another set of frequencies in another part of the brain.

This technique can be used to abduct people for secret government mind control experiments, under the guise of alien abduction. A microwave beam of this nature will paralyse the victim, so they can be bundled into a black helicopter and airlifted away for experimentation.

Once the procedure is complete, hypnosis can be used to plant false memories of alien abduction. In this way, real alien abductions can be used by the authorities to enable them to obtain a limitless supply of guinea pigs for their mind control experiments.

Real memories of government involvement are erased electronically. This technique clears all short-term memories from the victim's consciousness by broadcasting microwave beams at the target which carry the signals used for memory retention. When you remember something, it is first stored in your short-term memory. After approximately twelve hours, this short-term memory is converted in the brain to long-term memory, after which you remember this information for the rest of your life. If this conversion from short-term memory to long-term memory does not occur, the data is lost. We all see details around us, but try to recall the decor in a restaurant you ate at some weeks ago and you will see how tenuous memory is. Microwave radiation of a specific frequency can interfere with the transfer of memories from short to long-term memory. Microwave radiation of a specific frequency can interfere with the synapses of the brain. By interfering with the connections between brain cells, memory of people can be disrupted. In this way, Seal special forces assassins can be brain wiped after a mission, so they have no idea of the target they killed. Using hypnosis, false memories can then be planted in the brain, so the gap left by the real memory is papered over.

The latest advance in electronic mind control were discussed in my previous articles in Nexus, but for those who missed the ESP of Espionage, this equipment uses special types of microwave beams called MASERs. These are the laser equivalent of microwave beams. These MASER beams have been used to develop something called synthetic telepathy. This is the ability to read peoples' minds from a distance. Electronic scanning of victims' brains by monitoring the electromagnetic (EM) emissions from peoples' brains and using amongst other things, the brain waves (as measured on an EEG), to read the victim's subvocalised thoughts.

Head of US Special Forces Major-General Schaknow, talked about synthetic telepathy during a lecture in July 1992 at Fort Bragg North Carolina. The US military is hard at work perfecting synthetic telepathy. In synthetic telepathy, the weak electromagnetic signals in the brain associated with subvocalised thought, are connected to a computer by use of electrodes, or in more advanced mechanisms by MASER beams. Sophisticated computer systems have learnt to read the subvocalised thoughts in the brain, by associating a specific brain excitation potential, with a particular word. In this case, only one specific language can be decoded, as each word in a language has a specific set of frequencies that must be discovered. Once the donkey work of finding the specific frequencies for all the words in a language has been programmed into a super computer, which can carry out massive parallel processing, fuzzy logic software is used to match this with real world excitation potential associated with subvocalised thought obtained from thousands of abductees, who are used to calibrate the synthetic telepathy devices. GCHQ Cheltenham, the intelligence gathering arm of MI5, possess the advanced computer systems needed for synthetic telepathy.

Synthetic telepathy detects the l5Hz, 5 milliwatt auditory cortex brain emissions, that are linked with the excitation potentials in the brain associated with subvocalised thought. New technology, involving low frequency microwaves and RF, has enabled devices to be built which can scan through walls and look inside peoples' bodies like X-rays. This enables security personnel to see a target in his own home and to track him throughout the house. Further to this, being able to see inside the victim's head, would allow computer controlled targeting of specific brain centres in the victim's brain, even when he was walking around the house.

A scan of the specific brain emissions given off when the victim subvocalises using an array of pulsed frequency MASERs fired at the specific brain centres of the subversive, while he resides in his own home, enables the victim to be scanned. By firing an array of ELF pulse modulated MASERs, which scan up and down the window of frequency emissions given off by subvocalised thought, interference effects can be measured in the MASER beam. The victim's ELF brain emissions will interact constructively or destructively with the pulsed frequency MASER carrying ELF in the ELF window associated with subvocalised thoughts. If we fire an array of pulsed MASERs, which are out of phase with each other, extraneous noise can be filtered out in the digital domain. Since the converging ELF modulated MASERs are being effected by the low level emissions in the victim's brain, the shifts in the ELF pulsed signal going into the subversive's brain can be detected. A simplistic version of this would be the LASER beam shone at the window of the person that is being bugged. The vibrations in the window cause modulations in the LASER that can be converted into electrical signals and hence into sound. In this way the subvocalised thoughts in the victim's brain can be read. Having already built up a library of excitation potential signatures for differing words and groupings of words, a sophisticated computer can begin to decode the emission signatures into word streams. In this way the subvocalised thoughts of the victim can be stored in the memory of a supercomputer and analysed to give a read out of what the target is thinking..

Using ELF audiograms carried by a single pulse-modulated Maser, subvocalised thoughts can be placed in the victim's brain. This enables UK synthetic telepathy operators the ability to enter into conversations with the subversive to drive him mad or to bring up key words which will get the victim thinking about the information they wish to find. Visual cortex excitation potentials can also be broadcast into the victim's brain so that illusory images can be projected into their brain to drive them mad, or to programme them to commit suicide.

My research has led me to uncover a truly nation-wide mind control weapon system. Each UK police station is equipped with a vast array of microwave antennae. The Sussex Police headquarters has a two hundred foot antenna. The building is surrounded by a fence and is off limits to the public. This microwave complex sits in the middle of a council housing estate, which means the people in this area are being bathed in low level microwaves. New research has shown that low level microwaves give mice cancer, since the emitters tested by the scientists, were the new ultra safe mobile phones which give of much smaller intensities than the police antennae, the Police microwave system is a serious health hazard. It seems that low level microwave radiation excites the hydrogen bonds in the cell and can interfere with meiosis. This causes cell division to go wrong, which leads to cancerous cells and hence tumours.

Secondly, the police have been granted the exclusive use of the 450 MHz microwave frequency range. This is exactly the frequency used by Dr Ross Adey, the CIA mind control expert, in his experiments on behavioural modification. It seems the police have the exclusive use of this mind control frequency and a vast array of antennae to broadcast this frequency all over the country. Very useful for mass mind control in times of emergency. Adey found that by using 0.75mW/cm2 intensity of pulse modulated microwave at a frequency of 450 MHz, it was discovered that an ELF modulation could be used to control all aspects of human behaviour. The Sussex police headquarters is connected to CCTV, closed circuit television cameras throughout the town. Some of these cameras have microwave telemetry devices that could easily be used to broadcast this frequency. The large antennae that bracket the town could also be used.

The most insidious aspect of this, is that the entire mobile phone network could easily be used to control the behaviour of the phone users. By use of my microwave detector, I have found that mobile phones of the newer type, give off a pulse modulated microwave signal of around 0.75mW/cm2 at the earpiece. This may be coincidence, but it is exactly the intensity required form behavioural control as found by Dr Ross Adey, the pioneer of microwave mind control. So in theory, an ELF signal could be added to the microwave network to feed a precise behavioural pattern into every mobile phone user in the UK. If their were widespread riots, the ability to broadcast behavioural stimuli to mollify all the mobile phone users in the country would prove useful. Since mobile phone users are generally middle class, it means authority has a useful method of controlling the behaviour of the key voters. Microwave carrier beams are perfect for transmitting the excitation potential of docility to the phone user to keep them servile in times of trouble. When no ELF signal is broadcast the phone acts in a completely different manner on behaviour in humans. In this case the microwave phones causes the neurones to release calcium ions which makes the user tired irritable and when stresses likely to emotional outbursts such as road rage.

In Brighton, the local MI5 headquarters has a large microwave array on its roof. Secret bunkers in the area also have large microwave arrays above them. It is child‘s play to transmit an ELF modulated signal to be broadcast by the entire mobile phone network - if need be. By this means the entire mobile phone users can be behaviourally modified. At the cost of developing cancer from low level microwave exposure from the phones they constantly use, stressing the neural network by constant calcium ion efflux and interference with bioelectric fields.

It seems strange that a few milliCuries of ionising radiation will get the National Radiological Protection Board excited, whilst high intensity cancer inducing non ionising microwave radiation, goes unchecked. So high are levels near transmitters, that litter the countryside, that light bulbs will explode near them. If the intelligence community is using microwaves on a large scale as mass mind control weapons, then the NRPB‘s silence is easily explainable. To add to this, the numerous microwave detectors that were cheaply available to check leakage from microwave ovens are no longer made by any company in the country. It seems the UK authorities do not want the population to be able to detect when they are being microwave.

In conclusion, it can be seen that the UK intelligence and police, have a dizzy array of high-tech mind control devices. They regularly target their own populations and thousands of people are made ill by microwave weapons. With the advent of synthetic telepathy, ‚CCTV of the mind‘ becomes a reality.

With the election of a New Labour government unaware of microwave weapons and untainted, as they had no part in their development and deployment, it is certain that Tony Blair‘s government will outlaw these abhorrent microwave weapons - as they have done land mines. The only problem will be convincing the Labour government that MI5‘s microwave communication equipment is in fact microwave mind control weaponry. With the total secrecy the MI5 organisation operates under, one can be sure they will be mendacious as to how their £200 million pound budget is used to hone their leading edge microwave mind control torture weapons that are used on the public. Even more disturbing is the use of EMP and microwave weapons by the police on anyone researching into this area.

REFERENCES ELECTRONIC MIND CONTROL

[1] Non-Lethality: John B Alexander, the Pentagon‘s Penguin, by Armen Victorian, Lobster June 1993. This is an excellent article on the US black government‘s ‚hard-man‘ of the mind control world -recommended for those who really would like to know how these people think and work.
UFO Reality No 1-3, Jon King, Mind Control.
Mind Control and UFO‘s, by Alex Constantine, Feb 1996.

[2] UFO Reality No 4.

[3] Enigma Vol 2, It‘s All In The Mind, T Rifat.

[4] Ross W Adey: Neurophysiologic Effects of Radiofrequency and Microwave Radiation, Bulletin of the New York Academy of Medicine, Vo1.55 No 1 1, Dec 1979; The Influences of Impressed Electrical Fields at EEG Frequencies on Brain and Behaviour, in Behaviour and Brain Electrical Activity, Burch, N and Altshuler, H.I., eds, Plenum press, 1975; Effects of Modulated Very High Frequency Fields on Specific Brain Rhythms in Cats, Brain Research, Vol 58, 1973; Spectral Analysis of Low Frequency Components in the Electrical Activity of the Hippocampus During Learning, Electroencephalography and Clinical Neurophysiology, Vol 23, 1967.

Dr Ross Adey, formally of the Brain Research centre at the University of Southern California, Los Angeles, now at Loma Linda University Medical School, Loma Linda, California. He worked on the CIA Pandora Project and was concerned with inducing specific behaviour modification by electromagnetic means as well as induction of calcium efflux events to interfere with brain function - the so called ‚confusion weaponry‘.
C.S. Blackman is a researcher who worked in the US on theoretical research that underlies the electronic RMCT devices now deployed world-wide.
Electronics World and Wireless World: The healing face of electromagnetic fields; 1993.

[5]. Martin Cannon, The Controllers.

[6]. Secret and Suppressed, Jim Keith.

[7]. Psychic Dictatorship in the USA, Alex Constantine.

[8]. Encounters, Vol 7, Psi Spies, Marie-Louise Small. Cosmic Top Secret: America‘s Secret UFO Program,1992 William Hamilton III.

[9] Encounters, Vol 10, Room with a Remote View, by Richard Forsyth. This is an article about the author. http://www.fastnet.co.uk/pms ; Paranormal Management Systems‘ web site. For an information pack on RV/RI, send £3 plus a SAE to PMS, PO Box 2749, Brighton, BN2 2DR, UK. Correspondence and normal courses are available, phone 01273 690424 to book for courses. We also offer consultancy on mind control problems.

Nexus Vol 3 No 6, Remote Viewing: The ESP of Espionage, by T Rifat. Nexus Vol 4 No 1 The Esp of Espionage, Part 1, by T Rifat.

Nexus Vol 4 No 2, The Esp of Espionage, by T Rifat. These three articles cover the theory of RV, RI and ESP in very great detail. recommended for the serious RV/RI student. For more information contact the author.

Alien Encounters No 1 1 Alien and Government Mind control, T Rifat. Sightings Nol2, Losing Your Mind! Mind control Technology, by T Rifat. UFO Reality No 8, Mind Control, Big Brother Is All In The Mind, T Rifat. Enigma 6, Mind Control, by T Rifat.

Fortean Times 95, Police State of Mind, by David Guyatt. US Synthetic Telepathy info supplied by David Guyatt.

Guyatt, David. „Some Aspects of Antipersonnel Electromagnetic Weapons (paper prepared for ICRC Symposium for the medical profession.

ELECTRONIC SURVEYING INSTRUMENTS

A Review of Principles, Problems and Procedures

J. M. RÜEGER

MONOGRAPH 18

SCHOOL OF
SURVEYING AND SPATIAL
INFORMATION SYSTEMS

THE UNIVERSITY OF NEW SOUTH WALES UNSW SYDNEY NSW 2052 AUSTRALIA

J. M. Rüeger

Electronic Surveying Instruments

A Review of Principles, Problems and Procedures

First Edition

Reprint 2010

With 81 Figures and 6 Tables

Monograph 18
School of Surveying and Spatial Information Systems
The University of New South Wales
UNSW SYDNEY NSW 2052
Australia

Published by
School of Surveying and Spatial Information Systems
The University of New South Wales
UNSW SYDNEY NSW 2052
Australia

Received October 2003

© Copyright 2003, 2010

No part may be reproduced without written permission

First published: November 2003
Reprinted with corrections: October 2010

National Library of Australia
Card Number and ISBN
0-7334-2083-4

Preface

The field equipment used in daily surveying practice has changed considerably over the last 15 to 20 years of the 20th century. The traditional optical-mechanical instruments gave way to the newer electro-optical-mechanical devices. The new technology was not always understood nor were the field procedures always amended to suit it. Sometimes, field procedures were changed incorrectly because of a lack of understanding of the new instrumentation. It is hoped that this text will improve the understanding and use of the electronic equipment that is used today.

This text evolved from more advanced undergraduate and postgraduate courses taught by the author at the University of New South Wales (UNSW, Sydney, Australia) as well as from continuing professional development courses on electronic theodolites and digital (bar code) levels. It covers theoretical and practical aspects of modern field surveying instruments, their operating principles and their differences from traditional (non-electronic) field surveying equipment. Not discussed are methods involving the Global Positioning System (GPS), photogrammetry and remote sensing and the details of electronic data recording and coding. The reader is referred to other textbooks on these topics. The principles of electronic distance measurement (EDM) are not covered either as a modern and detailed textbook is available (Rüeger 1990 1996). Some specialised instruments using EDM, however, are discussed, as is reflectorless distance measurement.

This text was written for readers that had some previous exposure to field surveying. In particular, it is assumed that the reader is familiar with the basic principles and operation of optical theodolites and spirit levels. If not, the reader should consult a standard textbook on surveying and read the chapters on theodolites, levels and on the field operations with such before reading this text. This book suits professionals, who went through tertiary courses before the widespread introduction of electronic surveying instruments, current students who wish to broaden their understanding of field instruments, and readers of standard textbooks on 'surveying', who are not satisfied with the, usually very limited, information on instruments and procedures provided by these texts.

After an introduction, Chapter 2 introduces a number of hand-held instruments, such as hand-held distance meters, compasses, clinometers (and combinations thereof) as well as different types of electronic barometers. The latter double as altimeters and can be used for heighting.

Chapter 3 discusses digital (bar code) levelling instruments and rotating laser levels. Particular emphasis is placed on the various errors that may occur in levelling with digital levels and laser levels. This chapter also contains useful advice for all professionals involved in precise levelling with digital levels.

Chapter 4 introduces the different types of electronic theodolites and tacheometers, including gyro, laser, tracking, scanning and reflectorless instruments. Electronic circle reading systems and electronic level sensors are discussed in some detail, as are their inherent errors. Some notes on the practical use of these instruments follow.

Chapter 5 presents a number of dedicated instruments or techniques that were developped for specific tasks. Reflectorless EDM is discussed in some detail, highlighting some of the problems that can occur and giving some advice on its use. Laser scanners are mentioned as is the often ignored technique of motorised levelling. The chapter concludes with sections on airborne laser profilers, laser trackers and a laser measurement system. The latter two are employed in metrology.

Exercises are given at the end of each chapter. These exercises are designed to further the understanding of the principles and concepts discussed in the text. The solutions are not provided; they can be worked out from the information given in the text.

Sydney, Spring 2003 J. M. Rüeger

Reprint 2010

The author is greatly indebted to the readers that have communicated errors in the first editions, in particular A. Young. All known errors have been corrected in this reprint of the first edition.

Sydney, Spring 2010 J. M. Rüeger

Contents

1. Introduction 1

2. Hand-Held Equipment 5
2.1 Electronic Pocket Tapes 5
2.1.1 Ultrasound Sensor 6
2.1.2 Electro-Optical Sensor 7
2.2 Hand-Held Angle-Angle-Distance Devices 8
2.2.1 Fluxgate Compass 10
2.2.2 Clinometers 12
2.2.3 Use of Hand-Held Angle-Angle-Range Devices 13
2.3 Electronic Barometers (and Altimeters) 15
2.3.1 Working Principles 15
2.3.1.1 Vibration Pressure Sensor 15
2.3.1.2 Capacitive Pressure Sensor 17
2.3.1.3 Piezo-Resistive Pressure Sensor 18
2.3.2 Performance of Digital Barometers 19
2.3.3 Heighting with Digital Barometers/Altimeters 20
2.4 Exercises 22

3. Levelling Equipment 27
3.1 Digital (Bar Code) Levelling Instruments 27
3.1.1 Introduction 27
3.1.2 How do Digital Levels Differ from Conventional Automatic Levels? 29

3.1.2.1 DC Power 29
3.1.2.2 Staffs 30
3.1.2.3 Electronic Staff Reading and Distance Measurement 30
3.1.2.4 Illumination of Staff 30
3.1.2.5 Compensator 30
3.1.2.6 Optical or Electronic Horizontal Circle 30
3.1.2.7 Electronic Data Recording 31
3.1.2.8 Computer Interface 31
3.1.2.9 Real-Time Correction of Instrument Errors 31
3.1.2.10 On-Board Computing 31
3.1.3 Principle of the Wild NA2000 32
3.1.3.1 Basic Design and Working Principle 32
3.1.3.2 Measuring Sequence 34
3.1.3.3 Signal Processing 35
3.1.3.4 Accuracy of Bar Code Levels 37
3.1.4 Bar Code Staffs 39
3.1.4.1 Leica Normal Bar Code Staffs 39
3.1.4.2 Leica Invar Bar Code Staffs 40
3.1.4.3 Errors of Normal Bar Code Staffs 40
3.1.4.4 Errors of Invar Bar Code Staffs 40
3.1.5 Errors in Precise Levelling with Bar Code (Digital) Levels 40
3.1.5.1 Errors in Bar Code (Digital) Levels 41
3.1.5.2 Errors of Invar Bar Code Staffs 46
3.1.5.3 External Influences on Line of Sight 47
3.1.5.4 Errors of the Observer 48
3.2 Laser Levels 49
3.2.1 Construction of a Rotating Laser Level 50
3.2.2 Design of a Two-Axis Compensator of a Rotating Laser Level 53
3.2.3 Major Errors of Rotating Laser Levels 53
3.2.3.1 Compensation Error 53
3.2.3.2 Adjustment Conditions of Rotating Laser Levels 54
3.2.3.3 Misalignment of Rotating Mirror 55
3.2.3.4 Secondary Height Error 56
3.2.4 Calibration of Rotating Laser Levels 56
3.2.5 Computation of Heights, Earth Curvature and Refraction 57
3.3 Exercises 59

4. Electronic Theodolites and Tacheometers 65
4.1 Types of Electronic Theodolites and Tacheometers 65

4.1.1	Electronic Theodolites	65
4.1.2	Electronic Tacheometers	66
4.1.3	Gyro-Theodolites	67
4.1.4	Motorised Electronic Tacheometers	68
4.1.4.1	One-Person-Survey-Systems	70
4.1.4.2	Automatic Monitoring Systems	71
4.1.5	Motorised Electronic Video Theodolites	73
4.1.6	Laser Theodolites	75
4.1.7	Tracking Electronic Tacheometers	76
4.1.8	Reflectorless Electronic Tacheometers	76
4.1.9	Scanning Electronic Tacheometers	77
4.2	Features of Electronic Theodolites and Tacheometers	78
4.2.1	Circle Types	78
4.2.1.1	Incremental Circles	79
4.2.1.2	Quasi-Absolute Circles	80
4.2.1.3	Absolute Circles	80
4.2.2	Diametrical Circle Reading	81
4.2.3	Electronic Level Sensors versus Vertical Circle Compensators	82
4.2.3.1	Definition of an Electronic Level Sensor	82
4.2.3.2	Definition of a Vertical Circle Compensator	82
4.2.4	Data Recording	83
4.2.5	On-Board Computing	84
4.2.6	Power Supply	84
4.2.7	Communication with Personal Computers	85
4.2.8	Automatic Focussing, Searching and Pointing	86
4.2.8.1	Automatic Focussing	86
4.2.8.2	Automatic Searching	86
4.2.8.3	Automatic Pointing	87
4.2.9	Computer Assisted Instrument Calibration	89
4.2.10	Modern Vertical Axis Systems	90
4.3	Circle Reading System of the Electronic Tacheometer Nikon DTM-1	91
4.3.1	Coarse Measurement	92
4.3.2	Fine Measurement	93
4.3.3	Sensing of the Direction of Rotation	95
4.4	Electro-Optical Level Sensors	95
4.4.1	Nikon DTM-1 Single-Axis Level Sensor	97
4.4.2	Leica TPS1100 Series Dual-Axis Level Sensor	97
4.4.3	Kern E2 Dual-Axis Level Sensor	98
4.4.4	Principle of the Determination of the Theodolite's Dislevelment	100

4.5 Properties of Electronic Level Sensors 102
4.5.1 Compensation Error 103
4.5.2 Scale Factor Error 103
4.5.3 Setting Error 104
4.5.3.1 Random Component 104
4.5.3.2 Systematic Component 105
4.5.4 Vertical Circle Index Error and Correction 105
4.6 Circle Graduation Errors 106
4.6.1 Periodic Errors 107
4.6.2 Interpolation Errors 107
4.7 Some Notes on the Use of Electronic Theodolites and Tacheometers 109
4.7.1 Levelling of Electronic Theodolites and Tacheometers 110
4.7.2 Selection of Measuring Unit 110
4.7.3 Use of Umbrella 111
4.7.3.1 Electronic Theodolites 111
4.7.3.2 Electronic Tacheometers 112
4.7.4 Automatic Power-Off 112
4.7.5 Two-Way Radios and Cellular (Mobile) Phones 112
4.7.6 Real-Time Corrections to Measured Data 112
4.7.7 Direction Measurements 113
4.7.7.1 Single-Face versus Dual-Face Measurements 114
4.7.7.2 Sequence of Direction Measurements 115
4.7.7.3 Elimination of Periodic Errors 115
4.7.7.4 Circle Setting 115
4.7.8 Zenith Angle Measurements 116
4.7.9 Simultaneous Direction, Zenith Angle and Distance Measurements 116
4.7.10 Plumbing of Structures 117
4.7.10.1 Plumbing Shafts and Wells 117
4.7.10.2 Plumbing a Building 118
4.8 Exercises 118

5. Specialised Equipment 121
5.1 Reflectorless Electronic Distance Measurement 121
5.1.1 Basic Properties 122
5.1.2 Beam Size and Shape 123
5.1.3 Multiple Target Problem 125
5.1.3.1 Pulse Distance Meters 125
5.1.3.2 Phase Measuring Distance Meters 127

5.1.4 Influence of the Material and Colour of Target 128
5.1.5 Target Orientation 128
5.1.6 Measurement to Corners 129
5.1.7 Some Notes on the Use of Reflectorless EDM 130
5.2 Laser Scanners 132
5.3 Motorised Levelling 134
5.4 Laser Profiling and Laser Altimeters 137
5.4.1 Laser Profiling on Land 137
5.4.2 Depth Sounders 139
5.5 Single Beam Laser Trackers for Metrology 140
5.6 Reflectorless Single Station Laser Measurement System 143
5.7 Exercises 144

References **147**

Subject Index **153**

1. Introduction

Surveying technology did change considerably during the last 30 years due to developments in computers, electro-optics and satellite technology, to name just the major influences. It has been stated (Rüeger 1995b) that surveying practice does not make full use of the new (and more efficient and cheaper) opportunities. For reasons of efficiency and economics, electronic data recording and electronic post-processing will be mandatory in any large scale survey such as setting-out, detail and contour surveys. In consequence, a number of field techniques and surveying instruments have become obsolete over the last decades. Rüeger (1995b) discussed these aspects in more detail.

The changes in the surveying instrument market indicate the new technology bought by geomatic engineers and surveyors. For simplicity, the periodical instrument surveys published by the American journal Point of Beginning (POB) are used to show the trends in the purchasing of instruments. Although these surveys reflect the US market, the detected changes would be similar to those in South-East Asia (apart of vernier theodolites, naturally).

In 1982, the journal POB reported 46 optical theodolites and 3 electronic theodolites on the market. Ten years later, POB listed 18 optical theodolites and 16 electronic theodolites. It is unlikely, that optical theodolites will be manufactured very much longer, considering that an increasing number of second hand optical theodolites becomes available.

The journal POB reported 24 electronic distance meters in 1981 and 25 in 1993. Although the number of models did not change in the 12 years, one would expect a significant drop in the numbers sold. POB listed 11 fully electronic tacheometers ('total stations') in 1983 whilst listing 68 different models in 1994. It is obvious that the electronic tacheometers slowly replace stand-alone theodolites and stand-alone distance meters.

As far as tilting and automatic levelling instruments are concerned, 27 were reported by POB in 1984 and 37 in 1994. Six digital (bar code) levels were reported for the first time in 1994. Most noticeable are the 51 different models of rotating laser levels listed in 1994. Although some rotating laser levels did exist in 1984 (but were not included in the POB survey in that year), the number of models has certainly grown significantly over the last ten years. New is that all traditional surveying instrument manufacturers now cater for this market.

The above market surveys demonstrate that the electronic tacheometer has become a universal surveying instrument, replacing theodolites, distance meters and levels. The dumpy levels on building sites are clearly being replaced by the rotating laser levels. Where levelling instruments are still required for special applications, digital (bar code) levels are being used. The market for GPS receivers will grow further once millimetre accuracy can be achieved at the same price, with the same ease and with the same speed as with electronic tacheometers. GPS receivers, however, will never work in buildings, underground and under trees, to give a few examples.

Another trend in instrumentation, which could be documented with the aid of surveys on the purchase of data recorders, is electronic data recording. This goes hand-in-hand with the other trend of using personal computers for the processing of all field data and the production of plans, maps and data bases. In the opposite sense, set-out data are pre-computed in the office and transferred in electronic form to the field instrument for the setting-out. Electronic data capture is essential if data are to be transferred in electronic form from the field to the office or if they are to be processed in the field.

The trend towards one-person-survey-systems should also be noted. For example, the rotating laser levels on building sites are very powerful one-person levelling systems. Robotic electronic tacheometers can find and track (moving) reflectors and can be operated remotely by the surveyor at the reflector. They provide efficient one-person-systems for detail and contour surveys as well as setting-out work. They are also ideally suited for repeated surveys of the same network as, for example, in around-the-clock deformation measurements of unstable terrain and structures.

Electronic distance measurement (EDM) to passive targets (now referred to as 'reflectorless' EDM) has been possible for quite some time. Stand-alone distance meters, able to measure to non-cooperative targets, became available in 1986. Zeiss released the first 'reflectorless' electronic tacheometer ('total station') in 1993. Broad acceptance by the surveying profession followed once Leica, Topcon and Trimble released their reflectorless instruments in 1998, followed by Nikon in 1999. The range of reflectorless EDM depends heavily on the type of surface that is measured to and its orientation to the measuring beam. At the time of writing, 50 - 100 m seem to be typical maximum distances to 'natural' targets.

Laser scanners are another type of instrument that measure distances to passive targets. In this case, the laser beam is deviated vertically and horizontally to scan the surrounding space in a grid pattern. The step interval can be selected. Some instruments can scan around the full circle (360°), others only in a limited sector. The typical results of a laser scan are the x, y, z coordinates of a large number of points 'seen' by the scanner. The precision of these coordinates depends on the precision of the angular measurements (directions, zenith angles) and the precision of the distance measurements. The range and accuracy specifications vary greatly between the different products, from ±0.5 mm at 5 m to ±50 mm at 1000 m at the time of writing.

Geomatic engineers and surveyors must consider the above trends and check if their current operations make the best use of current technology. If not, they may provide slower, more expensive or less sophisticated services to their customers. Also, geomatic engineers and surveyors must ensure that their personnel is competent in the use of the new technology. Although some basic principles of traditional instruments also apply to modern instruments, some new aspects arise. For example, do you know how to calibrate a rotating laser level or the electronic level sensor of an electronic tacheometer? Do you know why it is better to have diametrical circle reading on the vertical circle (rather than on the horizontal circle) of electronic tacheometers or on what kind of survey jobs the dual-axis level sensors (of electronic tacheometers) are most useful? Do you know the properties of the compensators used in automatic levels, digital levels and rotating laser levels? The answer to these and other questions can be found in this text.

2. Hand-Held Equipment

Hand-held instruments are used in many ways in geomatic engineering and surveying. For example, they may provide the traditional angle-angle-distance data with lesser accuracy but faster than electronic tacheometers. Such applications include reconnaissance surveys and surveys in support of land information systems or geographic information systems. On other occasions, a geomatic engineer or surveyor needs to measure short distances fast, such as in the case of building surveys. Finally, the ambient temperature, pressure and humidity is required in the connection with all electronic distance measurements. Electronic instruments are available for all these and other applications. Some types of instruments are briefly reviewed below.

2.1 Electronic 'Pocket Tapes'

The 'electronic pocket tapes' may replace the traditional steel pocket tapes and 30 m steel tapes. They do require a surface for the (passive) reflection of the transmitted beam. Because they measure the average distance to the 'foot print' at the target surface, they work best, if the entire area 'illuminated' by the beam (the 'foot print') is flat and at a right angle to the beam. Measuring to concave or convex corners of walls will not give the true distance. Also, any other objects in the beam between the instrument and the reflecting surface can cause a false reading (too short).

Two types are discussed further, one based on an ultrasound sensor, the other one based on the well-known principle of electronic distance measurement.

2.1.1 Ultrasound Sensor

The *Accutape-S ultrasonic distance estimator* (see Figure 2.1, on left) is one example of an ultrasound sensor. It can measure distances from 0.6 m to 13 m with a resolution of 1 cm or 1 inch. The precision is quoted as ±1% + 1 digit which is equivalent to ±(10 mm + 1000 ppm). It operates between 4°C and 38°C and weighs 57 g. The instruments uses an electro-acoustic transducer (made by Polaroid) to transmit and receive sound waves. The 'flight time' for a pulse to travel to the reflecting surface and back is measured. Knowledge of the speed of sound c permits to compute the distance d as follows:

$$d = \frac{(c \times \text{flight time})}{2} \tag{2.1}$$

Schwarz (1985) did derive the following equation for the speed of sound c (in m/s):

$$c = 20.0533 \sqrt{\frac{T}{1 - 0.3780\,(e/p)}} \tag{2.1a}$$

where T is the atmospheric temperature (in kelvin; = 273.15 + t), t is the atmospheric temperature in degrees Celsius (°C), e is the atmospheric partial water vapour pressure (in hPa) and p the atmospheric pressure (in hPa). The speed of sound is 325.30 m/s at –10°C and 354.86 m/s at +40°C, for example.

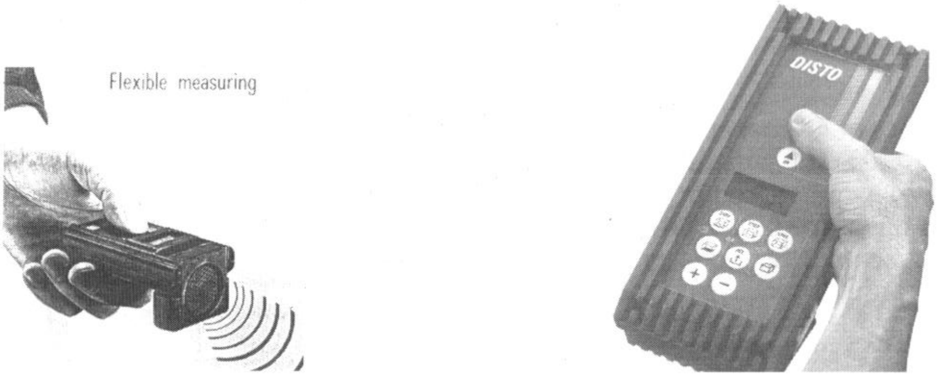

Figure 2.1: *Accutape-S* ultrasonic distance estimator (on left) and hand-held (electro-optical) laser meter *DISTO* by Leica (on right).

As the speed of sound depends greatly on the density of the medium (air), the *Accutape-S* is temperature compensated. Some systematic errors occur, however, with changes in pressure (altitude) and humidity. Sonic signals are not reflected by absorbing materials such as curtains, upholstery and carpets. Rough surfaces return an average distance. Because of the internal measurement of temperature (for the computation of the most appropriate speed of sound), the device must be given enough time to settle at the ambient temperature and should not be exposed to the sun or other heat sources.

2.1.2 Electro-Optical Sensor

As an example of an electro-optical sensor, the Leica hand-held laser meter 'DISTO' is discussed (see Figure 2.1, on right). The first DISTO model was released in 1994. The instrument is 235 mm long. That is much longer than the ultrasound device discussed above. It uses an visible laser diode distance meter to measure the distance to surfaces. This distance meter uses the same phase-measurement principle as most other electronic distance meters. (The unit length of the original DISTO is 3 m.) The range is quoted as 0.2 m to 30 m, without reflector. The resolution is 1 mm. Test measurements have established that all systematic errors (additive constant, periodic errors, scale errors) of the particular DISTO tested were well below the 2 mm level (Witte & Yang 1994). The fifth DISTO series was released in 2002. For the DISTO *lite*[5] and the *classic*[5], the operating range is listed as 0.2-200 m and the accuracy as ±3 mm at 2σ (±1.5 mm at 1σ). The display resolution is 1 mm below 100 m and 10 mm above 100 m. (Tests have shown that these two models are programmed not to measure distances longer than 200 m. Special target plates are required to achieve the maximum distance.)

The DISTO does not have an input facility for the atmospheric parameters. Like with all other electro-optical distance meters, the scale of the measurements does change at a rate of about 1 mm/100 m per 10°C change in atmospheric temperature and per 33 hPa change in atmospheric pressure. Whilst this might be ignored for distances of 0.2 m to 30 m, corrections might be appropriate for distances of 100 m. At temperatures higher than the reference ('standard') temperature of the instrument, distances are measured too short; a positive correction has to be applied. At pressures lower than the reference ('standard') pressure of the instrument, distances are also measured too short; a positive correction is required.

According to the manufacturer, the Leica DISTO measures the correct distance at 23°C, 953 hPa and 60% relative humidity. When ignoring the influence of humidity, the following correction equation can be derived from the standard conditions and the carrier wavelength of 635 nm given by the manufacturer:

$$d_{corr} = d_{meas} + \{(259.7 - \frac{80.91\ p}{T}) (10^{-6})\ d_{meas}\} \qquad (2.2)$$

where T is the atmospheric temperature (in kelvin; = 273.15 + t), t is the atmospheric temperature in degrees Celsius (°C), and p the atmospheric pressure (in hPa).

The visible laser diode projects a red dot onto the reflecting surface, so that the user knows exactly to what point the distance is being measured to. The range can be extended to 100 m (or 200 m for newer models) by using a (brown) target plate (made of reflective sheeting). As the laser dot cannot be seen by eye at longer distances, Leica sells a telescope attachment, which clips onto the DISTO and is held in place by magnets. Also, freehand operation may not provide a stable pointing at longer distances. DISTOs can be screwed onto camera tripods if better beam stability is required.

Since 1995, the instrument is available in some specialised versions, such as the Data DISTO RS232 for operation from computers for control measurements and as Data DISTO GSI for the attachment to the Leica electronic theodolites. In 1996, DISOFTmini followed. This data acquisition system links a DISTO to a penpad computer for the computer-assisted survey of buildings. A data transfer to CAD systems is possible.

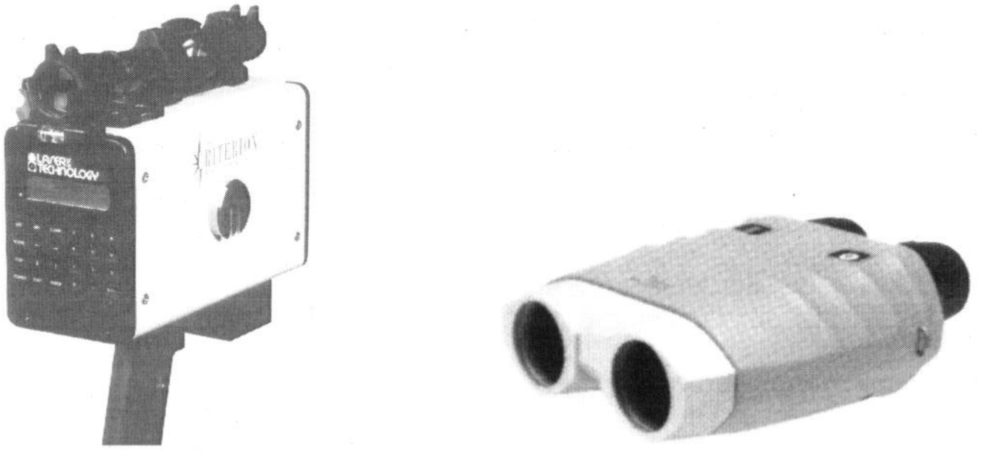

Figure 2.2: Criterion hand-held Survey Laser by Laser Technology (on left) and Leica Locator (on right)

2.2 Hand-Held Angle-Angle-Distance Devices

A number of manufacturers provide binocular-type systems, that can measure the distance to a point as well as the magnetic bearing and the inclination of the corresponding line of sight. Such systems provide data similar to those of the electronic tacheometers, namely a slope distance, a horizontal angle (referenced against magnetic north in this case) and a vertical angle. In consequence, the horizontal distance and the height difference to a point can be calculated on-board. Together with the magnetic bearing, a three-dimensional positioning is possible. An RS-232 interface for the external storage and processing of the data is usually available.

Instrument	Criterion 300	LEM 300-W Laser Meter	Leica Locator	Laser Scout 1000-C
Distance				
Accuracy	±100 mm	±100 mm	±1 m	±200 mm
Minimum Distance	6 m		25 m	2 m
Range to Poles	275 m	<100 m		1200 m
Range to Buildings	455 m	300 m	1500 m	1500 m
Range to Triple Prism	12200 m	3000 m		3500 m
Measuring Time	0.3 s	???	0.3 s	0.5 s
Azimuth				
Accuracy	±0.3°	???	0.6°	0.2°
Measuring Time	1.0 s	???	0.2 s	???
Stabilisation	±15°(floating)	???	3-D sensor	±20°(gimbals)
Inclination				
Range	1x±60°	1x±65 grads	2x±45°	1x±60°
Accuracy	±0.2°	±0.1 grad	0.2°	0.1°
Measuring Time	0.75 s	???	???	???
RS-232 Port	yes	yes	yes	yes
Manufacturer	Laser Technology 7399 S Tucson Way Englewood Co 80112 USA	Jenoptik GmbH D-07739 Jena Germany	Leica Geosystems CH-9435 Heerbrugg Switzerland	Dr. Johannes Riegl Wiener Strasse 2 A-3580 Horn Austria

Table 2.1: Technical data of a selection of laser range finders

These hand-held instruments have many uses for mapping, profiling, estimating, sampling, volume, inventory, topographic, relocation and GIS/LIS measurements. They are often used to connect additional features to points which were coordinated by GPS techniques. Because the instruments measure bearings directly, no other points need to be visible, at least in principle. Specialised instruments have been designed for the timber industry. Having measured the distance to a tree, the instruments can easily determine the height (measure elevation angle to the top and the bottom) and the diameter (measure azimuth to left and right edge), after some trigonometric computations. Some application aspects are discussed by Carr (1992). Other versions of the same type of instruments permit the measurement of the velocity of objects, such as cars. Jenoptik, for example, supplies an instrument (Video-LAVEG Laser Velocity Guard) that measures speeds of 0 to 320 km/h (to 1 km/h) at distances up to 350 m and more. All particulars are overlaid on a video image of the scene for documentation purposes.

The key data of a selection of typical instruments is given in Table 2.1. Two specific instruments are depicted in Figure 2.2. All instruments use pulsed diode lasers for the distance measurements, usually operating in the near infrared (wavelengths of about 900 nm). The distance is derived from the flight time of the pulses to and from the reflecting surface (see Eq. (2.1)). The quoted distance range is for daylight measurements. At night, like with all other electro-optical distance meters, the range can be extended up to the double of the daytime range. Further information on pulse distance meters may be found in Rüeger (1990 1996). It should be noted that freehand operation of these devices may not be able to sufficiently stabilyse the beam on a distant target. In these cases, some support may be required to keep the beam on the target. The components used for the azimuth and the tilt measurements are discussed below in more detail.

2.2.1 Fluxgate Compass

All hand-held angle-angle-distance devices described above use so-called flux-gate compasses for the measurement of the magnetic bearing. This technology is not new; it has been used for many years in aircraft navigation. However, the flux-gate compasses have only recently become sufficiently small for inclusion in surveying instrumentation.

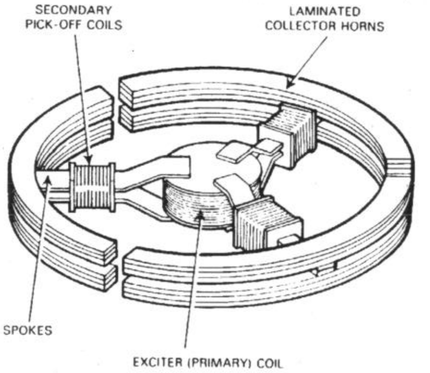

Figure 2.3: General construction of a flux detector (after Pallett 1972)

The flux-gate compass is a magnetometer which has been designed to determine the direction of the horizontal component of the vector of the earth's magnetic field. The design shown in Figure 2.3 has been the primary sensor design for more than 30 years. The flux sensor is usually gimballed so that it remains horizontal even if the host (or host vehicle) is tilted forwards and/or sideways. Naturally, the sensor has no degree of freedom about its vertical axis. To damp the vibrations induced by the host (or host vehicle), the sensor module is usually surrounded by a fluid ('floated'). The accuracy is typically 0.1°.

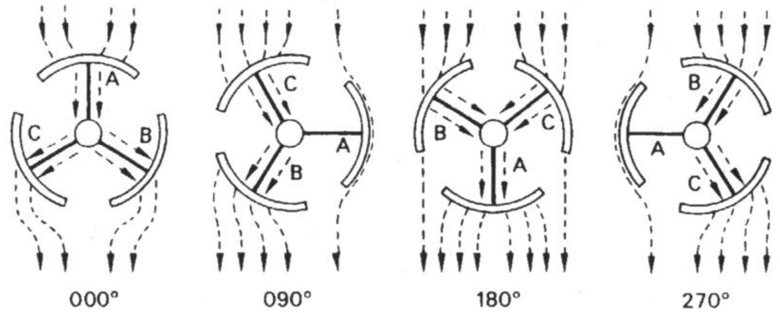

Figure 2.4: Path of the Earth's magnetic field through a flux detector and the signals induced in the pick-off coils on the three spokes (after Pallett 1972)

The primary coil generates an magnetic field (e.g. at 400 Hz) in the three spokes. The magnetic flux is carried by the upper and the lower part of the spokes. As the secondary (pick-off) coils include the forward and return paths of the induced magnetic flux, the total flux measured by the pick-off (at 800 Hz) will be zero in the absence of the Earth's magnetic field. As soon as the Earth magnetic field adds to the induced magnetic flux, the total flux through the secondary coil becomes non-zero and a voltage is generated in the pick-off coil due to electro-magnetic induction. The voltages from the three pick-offs are then combined to give the angle of the Earth's magnetic field with respected to the coordinate system defined by the three spokes. Figure 2.4 shows how the orientation of the flux gate sensor in the Earth's magnetic field changes the flux in the three spokes. A pick-off generates a maximum signal, when its measuring axis (spoke) is parallel to the Earth's field, and a minimum signal, when it is at right angle to it. In aircrafts, the flux detector elements were connected to the heading indicator in the cockpit. For hand-held instruments, the pick-off voltages have to be digitised and processed further by the on-board computer.

The traditional flux gate compasses are relatively large and heavy and must be stabilised in a horizontal plane. The gimbals or pivots used for the latter purpose restrict the allowable along- and across-tilts of the device. Rather than stabilise the traditional two-dimensional sensor in a horizontal position, it is possible to use a triad of three orthogonal single-axis flux gate sensors. Such a triad can measure the three components of the Earth's magnetic field in the case coordinate system. To derive the bearing of the sensor's case in a horizontal plane, the along- and across-sensor tilts must be measured. (The traditional flux-gate compass does not require this additional information as such. But in combination with the distance meter, at least the tilt in the forward direction must be measured anyway.)

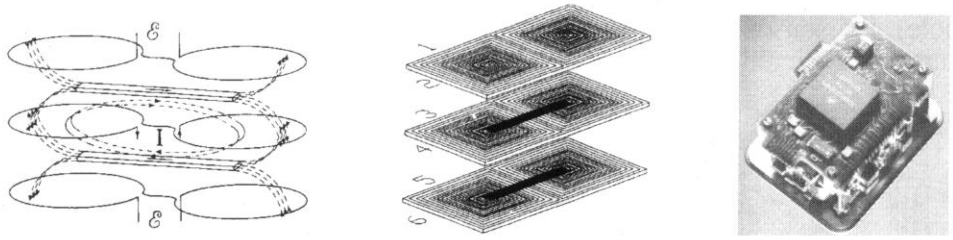

Figure 2.5: Principle (on left) and possible design (in middle) of a planar parallel-gated fluxgate sensor (after Vincueria et al. 1994). On the right, the three-dimensional digital magnetic compass (DMC) designed by Leica for their 'Vector' and 'Geovid' hand-held devices (after Leica 1995).

Figure 2.5 shows the basic principle a possible design on the left (Vincueria et al 1994). The (planar) induction coil is sandwiched between two (planar) exciting coils. Two cores made from amorphous low magnetostrictive ribbons are placed between the layers with the coils. In the absence of an external magnetic field, the induction coil would register a total flux of zero, which means that no induction would occur. The diagram in the middle of Figure 2.5 shows a practical design with four exciting coils (Numbers 2 to 5) and two secondary (pick-off) coils (Numbers 1 and 6) (Vincueria et al 1994). Leica does manufacture a three-axis flux-gate sensor (Digital Magnetic Compass, DMC) triad for their binocular based measuring systems. The device is specified to operate at along- and across-tilts of ±35°. Two integrated inclination sensors determine the two components of the dislevelment of the sensor and correct the sensor reading in real-time.

2.2.2 Clinometers

Most manufacturers of hand-held angle-angle-distance devices are not specific about the tilt sensors used. All instruments require a clinometer reading in direction to the target for the reduction of the measured slope distances to horizontal. The instruments which use a triad of single-axis flux sensors require an additional tilt sensor for the measurement of the cross tilt for the correction of the azimuth readout.

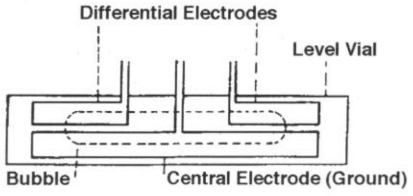

Figure 2.6: Plan view of an electrolytic level sensor.

Only Riegl quotes a 'high-quality electrolytic level sensor' with a measuring range of ±60° and a resolution of 0.1°. Figure 2.6 shows the basic design of such a sensor. The electrolytic level sensors resemble the plate levels in theodolites. The glass vials are filled with an electrolyte, which is conductive. If the air bubble is centred, both differential electrodes will be covered equally by the liquid. This means that the current in both leads will be the same. If the bubble moves to one side, one differential electrode will carry a larger current than the other. The difference between the two currents is a measure of the tilt. As an alternative, the readout electronic can be based on the differential change of the capacitance.

All tiltmeters and clinometers have a number of errors requiring calibration or adjustment. The most important corrections are the additive constant (for the zero error) and the scale correction (for the vertical-angle-dependent (linear) error). The former can be easily tested by the user through the measurement of the *same* slope, forward and backward. The latter requires a comparison against true slopes and is less easily established. Errors in the additive constant of the clinometer can be eliminated in the field by measuring the *same physical slopes* (or height differences) in both directions and by taking the mean.

2.2.3 Use of Hand-Held Angle-Angle-Range Devices

The hand-held angle-angle-distance devices are promoted for the measurement to non-cooperative targets such as poles, trees, bushes, houses etc. It is clear that the measurement accuracy of the devices cannot be realised if the dimensions of the object are larger than the measurement accuracy. All distances are measured to the closest surface (on the object), and not to the centre of the object. Also, one has to make sure that the foot print of the laser beam falls fully onto the object. If a number of objects are 'hit' by the beam, the instrument may display an average distance.

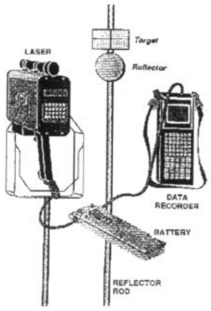

Figure 2.7: Equipment used by the Forest Service of the US Department of Agriculture for the reconnaissance surveys of single lane, low-use roads. (after Moll 1993)

The fluxgate instruments can be used to build up survey networks as it was done in the past with magnetic compasses. The compass traverses are well known. The definition of the forward, backward and instrument stations becomes very important, if measurements from different stations are to be linked. Figure 2.7 shows some ancillary equipment trialed by the Forest Service of the US Department of Agriculture for the survey of forest roads. The use of (supporting) poles at the instrument and the target stations, and the use of (plastic) reflectors, ensures that the marked points can be re-occupied with an accuracy that is better than that of the final product. The data are recorded for later processing on a personal computer. The reflector rod carries a target plate at the same offset as the telescope above the distance meter. Equal instrument heights and reflector heights ensure that the forward and backward elevation differences are compatible. Having an 'active' reflector and a well defined instrument station (on rod) and target (on rod) delivers the best accuracy in all three components.

All types of flux-gate compasses make use of the Earth's magnetic field to obtain orientations (magnetic bearings). In consequence, they are subject to the same effects as the older magnetic compasses. If the measured orientations are to be referenced against true north (geographic north), then the value for the *magnetic declination* (the angle between the true north (TN) and the magnetic north MN) must be considered. The declination value must be current for the time and location of the measurement site. Most instruments have a facility to enter this value. The problem is that the Earth's magnetic field is not constant. It is well known that the magnetic declination is subject to a secular (long-term) variation, a diurnal variation (periodic error over one day) of about 8 minutes of arc, an annual variation (periodic error of less than one minute of arc) and to irregular variations due to magnetic storms. The latter can reach one degree or more. In 1970, the magnetic declination was 11.50 degrees in Newcastle (NSW) and the secular variation +1.7 minutes of arc per year (Finlayson 1973). In Carnarvon (WA) and again for 1970, the values were –1.15 degrees and –1.1 minutes per year, respectively.

If the measured orientations are to be referenced against the grid north (GN) of a map projection, then the *grid convergence* (the angle between the true north (TN) and the grid north, GN) has to be considered in addition. In practice, it might be easier to calibrate an angle-angle-distance device on a few lines of known azimuth or bearing, depending on requirements.

Less predictable are local disturbances of the magnetic field. These 'local attractions' fall into two categories, namely those due to the station environment and those introduced by the observer. The magnetic field at a particular set-up is affected by metallic objects in close proximity (such as passing or stationary cars, transmission line towers, steel street light poles, steel structures, reinforced concrete structures, to give just a few examples), mineral deposits and direct currents. The errors caused by these objects are not predictable, but are constant at a particular point. Measuring bearings forward and backwards on lines allows to detect station specific errors, unless both end points are affected. The interference caused by the observer, such as from the metal frame of glasses, the watch, the buckle of a belt, the keys and/or the pocket knife will be constant for a particular observer (if the readings are always taken the same way). The measuring devices can compensate for these errors, either as part of the declination or as a separate entry.

The problem of local disturbances should not be underestimated. In urban areas, for example, it would be extremely difficult to obtain correct magnetic bearings. In these cases, where local disturbances are unavoidable, the measured bearings can be processed as measured directions, as all bearings observed from one station would be wrong by the same amount.

2.3 Electronic Barometers (and Altimeters)

Electronic (or 'digital') barometers have two main uses in geomatic engineering. Firstly, they can be used to measure the atmospheric pressure required for the reduction of electronic distance measurements. In this case, a read-out in hectopascal (hPa, previously known as millibar, mb) or in another pressure unit is required. Secondly, barometers double as altimeters, in which case a read-out in metres (or in another length unit) is required.

2.3.1 Working Principles

All electronic barometers are built around a pressure sensor with its associated analogue and digital circuits. The devices feature on-board computers for the conversion between different units, the corrections for the calibration values and the computation of the elevations from the measured absolute pressure. The most common sensor principles are now outlined.

2.3.1.1 Vibration Pressure Sensor

Vibration pressure sensors make use of the change in the resonance frequency of, typically, a quartz beam or a metallic cylinder under varying load. They are the most accurate of the three sensor types discussed here. Figure 2.8 shows one implementation of each type. In the vibrating quartz sensor, a quartz beam is placed in an oscillator circuit and is made to oscillate at its fundamental resonance frequency by piezoelectric means. This quartz beam is compressed by a lever arm which moves with the pressure in the steel bellows. The changing force on the vibrating beam causes a change in the resonance frequency. The frequency of the oscillator circuit follows this varying resonance frequency. In consequence, the changes in the oscillator frequency become a measure for the changes in the atmospheric pressure in the bellows. The frequency can be measured as in frequency counters.

Vibrating beam pressure sensors show a good repeatability and very small hysteresis effects. (Hysteresis is present if the pressure reading at a particular pressure is different depending on the pressure being reached by a pressure reduction or by a pressure increase. In the presence of hysteresis, the pressure reading is, slightly, dependent on the preceding pressure trend.) As with frequency counters, the resolution can be increased by increasing the length of the counting interval. Because quartz crystals change their resonance frequency also with temperature, the resonance frequency of a second (equivalent) quartz is monitored for temperature compensation purposes. The long-term stability is quite good. The vibrating quartz beam pressure sensors can be classified as precision pressure sensors (Sudau 1994).

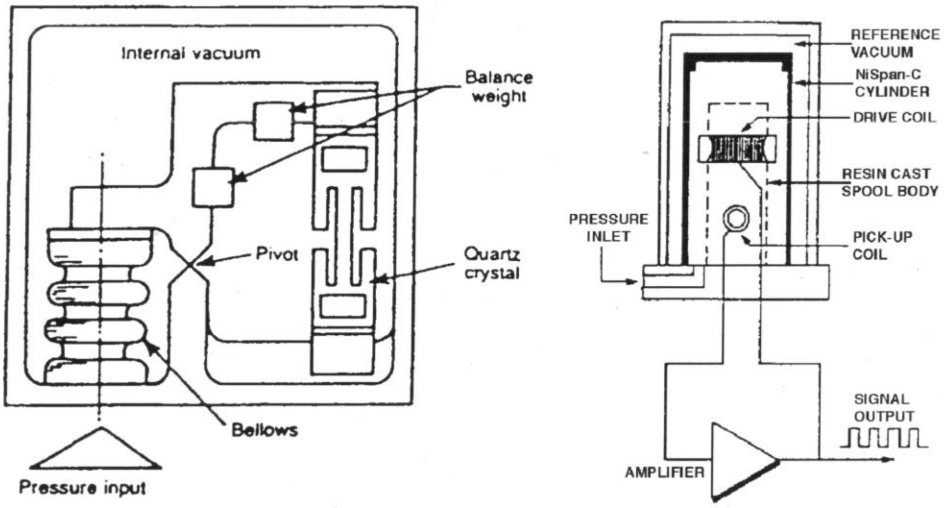

Figure 2.8: Two vibration pressure sensors. An oscillating quartz sensor (made by Paroscientific) is shown on the left and a oscillating cylinder sensor (made by Solartron) on the right. (after Sudau 1994)

The vibrating cylinder pressure sensor has a NiFe cylinder as a sensor element. A vacuum is present between the outer shell and the metallic cylinder. The drive coil causes the cylinder to vibrate at its resonance frequency. The pick-off coil is used to monitor the vibrations of the cylinder wall and to synchronise the frequency of the drive coil (in a feed-back loop). The inside of the cylinder is exposed to the pressure to be measured. As with the quartz sensor, the resonance frequency is measured and converted to the corresponding pressure by the on-board computer. The vibrating cylinder devices feature a very good repeatability, a small hysteresis and small long-term drifts. They are basically insensitive to accelerations but are susceptible to humidity (Sudau 1994). As the humidity enters into the cylinder, it is in contact with the inside of the cylinder and the drive and pick-off coils and, thus, may alter the readings.

2.3.1.2 Capacitive Pressure Sensor

The capacitive sensors are based on the principle of a capacitor, where a change in the width of the gap between the two electrodes causes a change in capacitance. Capacitive sensors are, typically, less accurate than the vibration sensors. In the case of a pressure sensor, one electrode is fixed whereas the other one is exposed to the pressure to be measured. Figures 2.9 and 2.10 show the arrangement used in the AIR-HB-1A hand-held digital barometer/altimeter. The device uses a dual-diaphragm, aneroid pressure sensor. It is highly sensitive to pressure changes and has a low susceptibility to shock and vibration. The dependence on temperature is small, almost linear and, thus, easily calibrated.

Figure 2.9 shows the two NiSpan-C diaphragms on opposite sides of a ceramic disk. The inside of the sensor is in vacuum. The two capacitors are electrically in parallel. Mechanically, they are differential. This back-to-back arrangement provides large capacitance changes in function of pressure and low susceptibility to vibrations and shocks. The sensor is placed in an oscillator circuit, the frequency of which varies with the applied pressure. For error compensation purposes, a high and a low reference capacitor and a temperature sensitive capacitor are also switched sequentially into the oscillator circuit. The frequency measured with each of the four elements is then used by the microprocessor to derive the pressure reading. The transfer function, calibrated during manufacturing, provides the temperature compensated pressure and the elevation reading (Whitaker & Call 1987; Pike & Bargen 1976; Pike et al. 1983; Whitaker & Call 1986; Pascoe & Rüeger 1989).

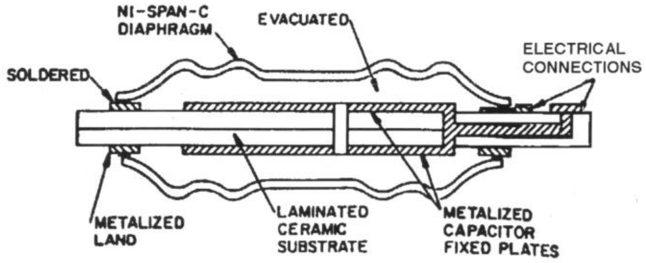

Figure 2.9: Capacitive pressure transducer of the AIR-HB-1A hand-held digital barometer/altimeter (after Pike et al. 1983)

According to Sudau (1994), capacitive pressure sensors feature a good resolution, a good repeatability, a good long-term stability and a good dynamic behaviour. The non-linearity between pressure change and capacitance change can be compensated by the microprocessor. The susceptibility to temperature can be calibrated and corrected for by the on-board microprocessor. See Pascoe & Rüeger (1989) for some test data.

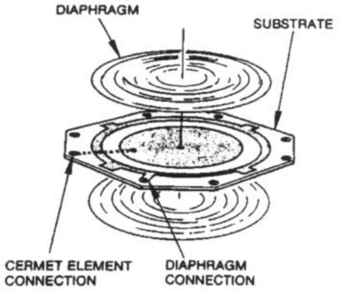
DIAPHRAGM
SUBSTRATE
CERMET ELEMENT CONNECTION
DIAPHRAGM CONNECTION

Figure 2.10: Dual-diaphragm, capacitance, aneroid pressure sensor of the AIR-HB-1A hand-held digital barometer/altimeter (after Whitaker & Call, 1987)

2.3.1.3 Piezo-Resistive Pressure Sensor

The piezoresistive sensors are the least accurate of the three types of sensors discussed here. They consist of a silicon membrane on a silicon substrate. A vacuum exists between the membrane and substrate. On top of the silicon membrane, four resistance sensors (strain gauges), in cardinal directions, are implanted. The strain gauges are connected in a Wheatstone bridge circuit. A change in the pressure causes a strain in the silicon membrane. This changes the resistance of the four gauges through changes in the specific resistance of the semiconductor material and yields a bridge voltage which is proportional to the applied pressure.

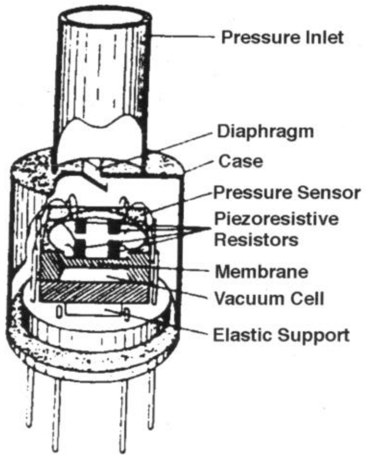
Pressure Inlet
Diaphragm
Case
Pressure Sensor
Piezoresistive Resistors
Membrane
Vacuum Cell
Elastic Support

Figure 2.11: Construction of a piezoresistive pressure sensor (after Sudau 1994)

The silicon chips can be as small as 4 x 4 mm, with a membrane thickness of 25 micrometres and a membrane diameter of 3 mm. As shown in Figure 2.11, the measuring chip is not directly exposed to the pressure to be measured. The external pressure is applied to an outer membrane ('diaphragm' in Fig. 2.11), which pressurises the oil inside the device. The oil applies the pressure to the silicon chip. The simple construction of piezoresistive sensor, together with their long service life and their miniaturisation, have lead to a multitude of applications in a number of fields. The temperature sensitivity of the resistance measurement is largely eliminated by the bridge circuit with the four strain gauges. The temperature dependence of the pressure reading must be compensated by other means.

2.3.2 Performance of Digital Barometers

Sudau (1994) summarised the performance of the three types of sensors discussed earlier. Sudau's table is reproduced in Table 2.2 in modified form. Rather than give the performance data in percent of the full scale (measuring range) of the devices, they have been converted to hPa by assuming a range from 600 to 1100 hPa (full scale = 500 hPa) for the devices. The performance as an altimeter can be estimated from the vertical lapse rate of the atmospheric pressure of about 1 hPa per 10 m.

The vibration sensors are intrinsically digital devices (frequency output) whereas the other two types are analogue devices, which require analogue to digital conversion. Most devices designed for outdoor use and having a display for the pressure (and elevation) readings seem to be of the piezoresistive and capacitive kind. Other devices provide the data only through a serial (RS232) computer interface and, thus, must be connected to a personal computer for the readout of the pressure or altitude.

Specification	Type of Sensor		
	piezoresistive	capacitive	vibration
Resolution (hPa)	∼0.1	<0.03	>0.000025
Linearity (hPa)	<2.5	<0.25	0.03
Repeatability (hPa)	0.15–2.0	<0.10	<0.03
Hysteresis (hPa)	0.15–0.5	<0.10	<0.03
Overall Accuracy (hPa)	∼2.5	0.10	0.05
Long-term Stability (hPa/year)	2.5-8.5	0.25-0.50	<0.05

Table 2.2: Performance data of three types of pressure sensors (after Sudau 1994)

The specifications of some self-contained hand-held barometer/altimeter devices are given in Table 2.3. The data have been taken from the respective instrument brochures. Prices start from about AUS $350 and are, typically, lower than the prices of a mechanical aneroid barometer/altimeter of equal precision.

Manufacturer	Eschenbach Optic Germany	Atmospheric Instrument Research (AIR) USA	Thommen Switzerland	Väisälä Finland
Model	6941 Alpin EL	HB-1A	Altitronic Professional	PA 11
Range (m)	-500 to 10000	-700 to 4200	-700 to 10000	
(hPa)		600 to 1100	250 to 1100	800 to 1050
Resolution (m)	0.5	0.1	1	
(hPa)	0.1	0.1	0.1	0.1
Tolerance	±8 m	±0.5 hPa	±1.0 hPa ±10m (>500m) ±2m (<500m)	±0.3 hPa
Temperature Range (°C)	-20 to +50	5 to 40 (Option: -25 to +50)	-30 to +60	+5 to +55
Temperature output	no	no	yes (Pt-1000) ±0.3°C	no
Sensor	?	1x capacitive	1x piezoresistive	3x capacitive
Weight (g)	?	280	170	1700
Size (mm)	?	30x91x145	131x68x25	72x144x250
Supply Voltage (V)	?	9	9	4.8

Table 2.3: Specification of some commercial hand-held barometers and altimeters

2.3.3 Heighting with Digital Barometers/Altimeters

Based on the height formula for a standard atmosphere, the microprocessors of digital barometers can easily compute and display elevations. Kahmen & Faig (1988, Eq. 11.7.2) give the following formula for the height H (in metres) in a standard atmosphere:

$$H\,(m) = \frac{T_o}{a}\left(1 - (p/p_o)^{1/n}\right) \qquad (2.3)$$

where
$n = 5.256$
p_o = mean sea level pressure = 1013.25 hPa
T_o = mean air temperature at sea level = 288 K (~15°C)
a = vertical lapse rate of temperature = 0.0065°C/m

Substitution in Eq. (2.3) of $T_o = 288.15$ K and other numerical values as given above leads to the following equation used by Atmospheric Instrument Research (AIR) for their HB-1A barometer/altimeter:

$$H\,(m) = 44330.77\,\left(1.0 - \left(\frac{p}{p_o}\right)^{0.19026}\right) \qquad (2.4)$$

where p (in hPa) is the pressure measured at the elevation H. The above equation is based on many assumptions, namely that the temperature at the mean height (H/2) is $\{T_o - a\,(H/2)\}$, the humidity is zero, the mean gravity is equal to g_o and that the sea level pressure is indeed p_o. These conditions are usually not fulfilled. In consequence, correction formulæ have been used in the past to compensate for the deviation of the real atmosphere from the standard one in the computation of barometric height differences. Details on these formulae may be found in sections on barometric heighting in surveying textbooks. (See Kahmen & Faig, 1988, Sect. 11, for example.)

Practical tests have shown that a simple interpolation (proportional to the height difference) of the barometric heights between the points of given elevation can provide quite accurate elevations when using electronic barometers. The interpolation formula is given by Kahmen & Faig (1988, Eq. 11.7.7) as follows;

$$H_i - H_A = (H_E - H_A)\,\frac{h_i - h_A}{h_E - h_A} = \Delta H_{A\text{->}E}\,\frac{\Sigma\Delta h_{A\text{->}i}}{\Sigma\Delta h_{A\text{->}E}} \qquad (2.5)$$

where $\Sigma\Delta h_{A\text{->}i}$ is the measured elevation difference from the starting Point A to a Point i, $\Sigma h_{A\text{->}E}$ the measured total elevation difference between the fixed Points A and E and ΔH the given elevation difference between the fixed Points A and E.

Lechner el al. (1987) reported on a test of motorised levelling with three different electronic barometers/altimeters (two types of piezoresistive devices) in two networks. (The speed of travel was about 30 km/h between set-ups.) The first network featured height differences of 20 to 80 m over 5 to 12 km, the second one a height difference of 250 m over 6 km. The root mean square error of the interpolated barometric heights were, on average, ±0.4 m and ±1.0 in the first and second network, respectively. The worst errors in the interpolated heights did not exceed 1.1 m and 2.8 m in the first and second network. Sudau (1994) quoted results of the interpolation method in barometric heighting with the more precise vibrating quartz beam instruments. A height precision of ±0.06 m was achieved between bench marks 450 m apart and over a height difference of 43 m. It follows that barometric levelling (motorised or otherwise) with digital barometers is a powerful and simple method, particularly if points of given heights are available at the edge of the area to be surveyed.

The interpolation method is based on the assumption that the atmospheric pressure does not change with time. Unfortunately, this is usually not the case. Even in stable weather conditions, the pressure falls by about 3 hPa between 9 h and 15 h and rises a similar amount between 15 h and 9 h on the next day. A smaller rise and fall occurs during the night and early in the morning (Bureau of Meteorology 1990). Considering a pressure change of 1.145 hPa per 10 m height difference (at 500 m above sea level), this diurnal pressure change translates into a 'change' of 26 m in the altimeter reading. (At 2500 m above sea level, the lapse rate of pressure is 0.9428 hPa per 10 m.)

Larger pressure changes occur during weather changes. When a high is moving away or a depression is approaching, the pressure can easily drop by more than 7 hPa over 24 h (Bureau of Meteorology 1990). This is equivalent to a change in the altimeter reading of 74 m (at 500 m above sea level). The interpolation method with a single altimeter is therefore only appropriate during stable weather conditions and over short periods of time. Best results are achieved on uniform slopes and when starting and finishing on points with given heights at the highest and lowest points of the height 'traverse'. Naturally, changes of pressure due to weather can be monitored by a second, stationary barometer if the barometric levelling extends over a longer period of time.

If given elevations are not available at both ends of a barometric heighting run, then the normal formula of barometric heighting and the normal observation procedures should be employed. The latter includes the use of a stationary barometer to monitor the changes, due to weather, of the atmospheric pressure with time. Refer to textbooks (e.g. Kahmen & Faig 1988) for full details.

2.4 Exercises

2.1 The measurements taken with the Accutape-S ultrasonic distance 'Estimator' are corrected internally for temperature. It is assumed that these devices are calibrated for 20°C, sea level pressure (1013.25 hPa) and zero humidity.

(a) Investigate the error in a 30 m distance measured at +35°C and 100% relative humidity (56 hPa partial water vapour pressure).

(b) What is the error in a distance of 100 m measured at 20°C at an altitude of 3000 m?

Hint: Consider Eqs. (2.1), (2.2) and (2.4).

2.2 Discuss the relative merits of ultrasonic, electro-optical and steel 'pocket tapes' for the three-dimensional (length, width, height) measurement of the internal size of rooms and buildings.

2.3 The Leica 'DISTO' *classic*[5] measures distances correctly at 23°C and 953 hPa. The instrument is accurate to ±3 mm at these conditions.

(a) What error would you expect in a 100 m distance measured at +30°C and 890 hPa?

(b) Would you correct the instrument reading in any way? Why?

2.4 The manufacturer of the Criterion 300 hand-held 'Survey Laser' (angle-angle-range device) specifies the distance accuracy as ±100 mm, the bearing accuracy as ±0.3° and the vertical angle (inclination) accuracy as ±0.2°.

(a) At what range does the accuracy of the computed height difference reach ±100 mm? (Assume a horizontal line of sight.)

(b) At what range does the accuracy of both components (Easting and Northing) of the horizontal coordinates (computed from bearing and horizontal distance) of a target point attain ±100 mm? (Assume a point due East or due North from the instrument station.)

2.5 You have been asked to map (in the Map Grid of Australia 1994, MGA94) a stand of old trees on a development site. Four survey marks with given MGA94 eastings and northings are available on the site. You are using a hand-held angle-angle-range device with a flux gate compass for the purpose. How do you get the MGA94 bearings (required for the calculation of the coordinates) from the measured magnetic bearings?

2.6 You have been asked to determine the height and the diameters of the stem and spread of some trees with a hand-held angle-angle-range device.

(a) What do you measure, exactly, to each tree?

(b) How do you calculate the required tree dimensions? Derive the necessary formulae!

(c) What precision of the three elements would you expect when using a Leica Geovid at 50 m from the tree?

2.7 Standing on a survey mark, *you* measure the height difference to the top of a power line pole with a hand-held binocular based angle-angle-range device similar to the one shown in Figure 2.2 (on right). The height of the survey mark is 238.44 m above sea level (Australian Height Datum 1971, AHD71). The height difference between the device and the top of the pole is displayed as –45.6 m. What is the AHD71 height of the top of the pole?

2.8 You have mounted a reflector on a ranging rod at a height above ground which corresponds to your eyelevel. You stand on a Mark A. Your assistant holds the ranging rod on Mark B. You read a vertical angle of +10.8° with a binocular based angle-angle-range device. You then move to Mark B and your assistant to Mark A. From B, you measure the vertical angle from B to A as –12.2°C.

(a) What do you conclude?

(b) What would you do next?

2.9 List and discuss those variations of the magnetic field of the Earth which affect the magnetic bearings measured with flux-gate compasses.

2.10 The vibrating cylinder pressure sensors are said to be susceptible to humidity, whereas the vibrating quartz beam sensors (of equal precision) are not. Based on the Figures 2.8, explain the reason for this difference.

2.11 Compare the practical results of the interpolation method in barometric levelling with digital barometers/altimeters reported by Lechner (1987) and Suday (1994) with the performance parameters of the respective devices. What performance parameter corresponds best with the test results?

2.12 The motorised barometric levelling with one digital altimeter was used to determine the elevations of a number of points along a road. Given the 'true' heights of the Point 40 and the Point 51, compute the elevations of the Points 41 to 50.

Point	Given (m)	Measured (m)	Adjusted (m)	Point	Given (m)	Measured (m)	Adjusted (m)
40	256.85	251.7	256.85	46		456.7	______
41		292.6	______	47		484.3	______
42		327.3	______	48		513.9	______
43		351.8	______	49		545.2	______
44		390.1	______	50		561.5	______
45		421.4	______	51	586.28	573.3	586.28

2.13 Barometric levelling was executed with a digital barometer along a forest road. Compute the elevations of the Stations 61 to 65 based on the given elevations of Stations 60 and 66.

Point	Given Elevation (AHD) (m)	Measured Pressure (hPa)	_______ ___	Adjusted Elevation (AHD) (m)
60	1946.75	799.2	______	1946.75
61		789.7	______	_______
62		785.2	______	_______
63		780.8	______	_______
64		774.9	______	_______
65		770.1	______	_______
66	2339.42	760.5	______	2339.42

3. Levelling Equipment

3.1 Digital Levelling Instruments

3.1.1 Introduction

Digital levelling instruments ('bar code levels') were introduced in 1990 by the surveying equipment manufacturer Wild (now Leica) of Switzerland. Two other manufacturers (Topcon and Zeiss) released digital levels in 1994. Sokkia followed in 1998. Each of the manufacturers uses a proprietary design of the bar code graduations on the respective staff, so that staffs cannot be interchanged between instruments of different manufacturers. Digital levels simplify the traditional levelling by not having to read the levelling staff through the telescope and permit electronic data recording. In consequence and because larger sighting distances can be used in non-precise levelling applications, digital levelling instruments have made levelling faster. However, levelling with digital levels is still labour intensive as somebody has to operate the instrument and somebody has to hold the staff.

Considering the wide-spread use of one-person rotating laser levels on building and construction sites (discussed in the next section) and the replacement of line levelling by EDM-Height Traversing (Rüeger 1995, Rüeger & Brunner 1981) with electronic tacheometers in all but high precision 'levelling' applications, the traditional use of levelling instruments is likely to decline. The largest use of digital levels may well be in precise levelling applications in geodesy and engineering (e.g. monitoring of structures and machines) as the other possible applications are likely to be taken over by more efficient techniques.

In 2003, one manufacturer (Topcon Japan) offers three models of digital levels, two manufacturers (Leica (Wild) Switzerland, Trimble (Zeiss)

USA/Germany) offer two instruments each, and one manufacturer (Sokkia Japan) offers one. The manufacturers that sell at least two models, provide one or more for normal levelling (with fibre glass, aluminium or wooden staffs) and one for precise levelling (with invar staffs).

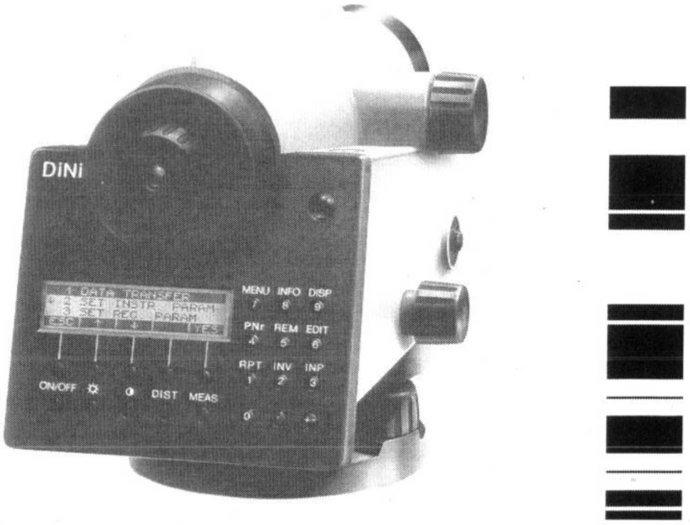
DiNi
DATA TRANSFER
SET INSTR PARAM
SET REC PARAM
ESC
YES
MENU INFO DISP
PNr REM EDIT
RPT INV INP
ON/OFF
DIST MEAS

Figure 3.1: Zeiss (Trimble) DiNi10 digital levelling instrument (on left) and Zeiss bar code (on right)

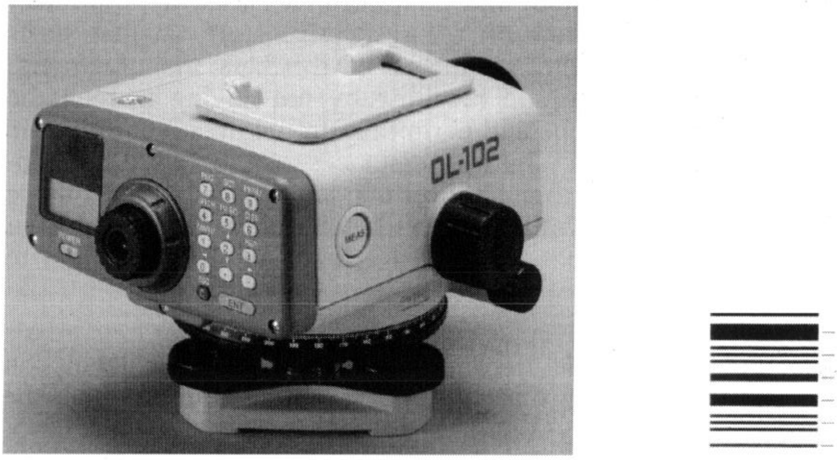

Figure 3.2: Topcon DL-102 digital levelling instrument (on left) and Topcon bar code (on right)

The precision listed by manufacturers for 1 km double run levelling (after the German DIN Standard 18723) with the less accurate instruments (and the corresponding (normal) bar code staffs) is as follows:

Leica DNA10 $s_{1\text{km double run}} = \pm 1.5$ mm
Sokkia SDL30 $s_{1\text{km double run}} = \pm 1.0$ mm

Topcon DL-102 $s_{1km\ double\ run}$ = ±1.0 mm
Topcon DL-103 $s_{1km\ double\ run}$ = ±1.8 mm
Wild NA2002 $s_{1km\ double\ run}$ = ±1.5 mm (now Leica)
Zeiss DiNi20/21 $s_{1km\ double\ run}$ = ±1.3 mm (now Trimble)

The precision listed by manufacturers for 1 km double run levelling (after the German DIN Standard 18723) with the more accurate instruments (and the corresponding invar bar code staffs) is as follows:
Leica DNA03 $s_{1km\ double\ run}$ = ±0.3 mm
Topcon DL-101/C $s_{1km\ double\ run}$ = ±0.4 mm
Wild NA3003 $s_{1km\ double\ run}$ = ±0.4 mm (now Leica)
Zeiss DiNi10/11 $s_{1km\ double\ run}$ = ±0.3 mm (now Trimble)

3.1.2 How do Digital Levels Differ from Conventional Automatic Levels?

3.1.2.1 DC Power

When using invar staffs, having no power means that no readings can be taken. When using non-invar staffs (like sectional fibre glass staffs), the digital levels can still be used as a normal level as the non-invar bar code staffs have a normal graduation on the back.
Having no power also means that no data recording is possible. It is evident that enough charged batteries must be taken into the field if no loss of time is to take place because of a lack of power.

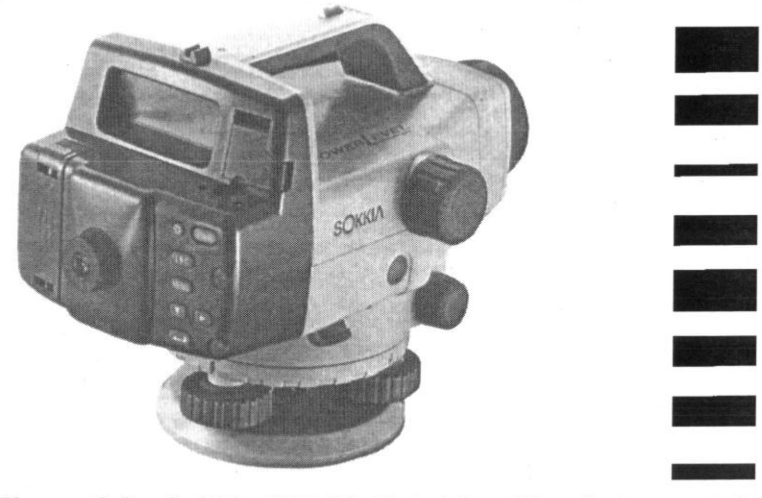

Figure 3.3: Sokkia SDL30 digital levelling instrument (on left) and Sokkia bar code (on right)

3.1.2.2 Staffs

Digital levels require special bar code staffs, which are make specific. The staff and the digital level must be of the same make. Normal bar code staffs are made from fibreglass, aluminium or wood and have normal graduations on the back face. They are usually sectional or folding staffs. Invar bar code staffs do not have a normal staff face on the back. The invar ribbon is typically mounted inside an aluminium frame which has a steel base plate at the bottom.

3.1.2.3 Electronic Staff Reading and Distance Measurement

Digital levels not only measure the staff reading but also the distance between the instrument and the face of the staff. The latter is measured with an accuracy of a few centimetres. Both values are displayed on the read-out screen.

The electronic reading is non-personal and faster and allows to change the units of measurement between metres and feet very easily (No need to change the staff as in the past!) and the resolution of the readings can be selected.

3.1.2.4 Illumination of the Staff

The illumination of the staff is more important than in the past, as the electronic measurement requires a good contrast. An artificial (warm) light source is required for indoor or poor light conditions. (The early Wild (Leica) and the Sokkia instruments use only the infrared image of the staff for the processing.) It has been found that problems occur when observing against a bright background, against the sun and in the case of shadow patterns on the staff.

3.1.2.5 Compensator

In 1995, all digital levels are equipped with compensators similar to those in the classic automatic levels where the entire optical path of the telescope is deviated. Electronic level sensors (similar to those in electronic theodolites) could be used, but this would no longer permit the use of the instrument as a normal automatic level.

3.1.2.6 Optical or Electronic Horizontal Circle

Most bar code levels have an internal or external circle for rough angular measurements. This circle is typically read by eye as in traditional automatic levels.

In 1995, Zeiss did release a digital level with an electronic horizontal circle reading ('*Total Level Station*' DiNi10T) where the directions are measured to 6" and the distances to 1 cm/10 m. In this case, the instrument can be used for detail and contour surveys like electronic tacheometers, as long as the ground is reasonably level. It is to be noted that correct pointing to the centre of the staff is then essential as is the set-up of the staff exactly in the centre of the staffs base plate. Strictly speaking, the distance is measured to the front of the staff whereas the (fibreglass) staff is usually set-up in the centre of its cross-section.

3.1.2.7 Electronic Data Recording

This is the most important advantage of digital levels. Electronic booking speeds up the measuring process and avoids booking errors. Also, no additional person is required as 'booker'. The electronic data can be recorded on the following media, depending on instrument:

Removable memory cards: These cards may be make specific or may be PC cards, formerly known as PCMCIA-cards (**P**ersonal **C**omputer **M**emory **C**ard **I**nternational **A**ssociation). The latter may be read by inserting them into a PCMCIA port of a computer. The former need an additional memory card reader for the up or down loading of the data.

Internal (RAM) Memory: with RS-232 port for the up and down loading of data.

Hard Disk of External Notebook Computer: linked to the instrument via a RS-232 port.

3.1.2.8 Computer Interface

All digital levels feature an two-way RS-232 port for the up and down loading of data and/or the remote operation of the digital level from a computer.

3.1.2.9 Real-Time Correction of Instrument Errors

Having the measurements in electronic form and a computer on board makes it possible to correct instrumental errors by (real-time) computation rather than by physical adjustment. This requires that the instrumental errors be determined by the manufacturer and/or the user, stored in a non-volatile memory on board the instrument and removed from the raw observations by computation, prior to displaying and recording the measurements.

The manufacturers provide on-board program routines which guide the user through the determination and storage of user accessible instrument corrections, such as the (vertical) collimation error. Other on-line error corrections are only accessible to the manufacturer. This may include corrections for temperature dependent effects, which are corrected on the basis of an internally measured temperature.

3.1.2.10 On-Board Computing

The powerful computers on-board digital levels can be programmed to assist with some calculations which were previously done by hand. Manufacturers may provide on-board programs in support of:

Line Levelling: For example, carrying forward of elevations from a starting point, summation of backsight and foresight distances, height transfer to intermediate sights.

Two-Peg Test: Calibration, storage and correction for the vertical collimation error.

Users should note, that **it is impossible to adjust or calibrate the collimation error of digital and other automatic levels without error**. The two-peg test gives the collimation error at a particular instant and temperature only. The collimation error changes all the time. Experiments have shown changes of 8" between two set-ups and changes of 1" per °C! Digital levels were found to be more sensitive to temperature changes than optical automatic levels.

3.1.3 Principle of the Wild (Leica) NA200x/300x Instruments

As the WILD NA2000 was the first digital levelling instrument (released in 1990), it will be used to introduce the principle of operation of such instruments. Other makes of digital levels do employ slightly different procedures for the decoding of the bar code staffs. Most operational aspects are the same, however. The principle of the Topcon digital levels is explained in Rüeger (2000) and that of the Zeiss (Trimble) digital levels in Feist et al. (1995), for example.

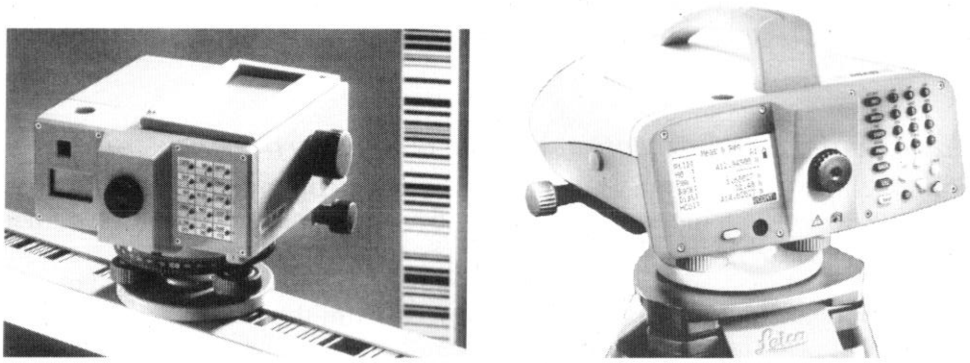

Figure 3.4: Wild NA2002 digital level (on left), Leica DNA03 precision digital level (on right), Leica fibre glass staff (in middle) and Leica invar staff (under the instrument on the left)

3.1.3.1 Basic Design and Working Principle

The digital level Wild NA2000 by Leica uses the same optical and mechanical components as a traditional automatic level. (For example, the compensator is the same as in the Wild NA2 and NA24 automatic levels.) In consequence it can also be used like a conventional automatic level. Figure 3.5 shows the main components of this digital level. The image of the staff's bar code is deviated onto the detector array by the beam splitter. The beam splitter doubles as an optical filter which deviates the infrared part of the spectrum onto the detector. The maximum sensitivity of the Silicon detector is in the near infrared. The visible part of the spectrum proceeds to the reticule of the level.

The linear detector array is 6.5 mm long and features 256 pixels. The spacing and the size of the pixels is 25 µm. The viewing angle of the telescope is about 2 degrees. At the shortest measuring distance of 1.8 m, a 70 mm long staff section is projected onto the array. At the maximum measuring distance of 100 m, the detector array sees a staff section of 3.5 m length. (This arrangement is quite different from that in the Zeiss DiNi10/20 levels, where only 15 cm above and 15 cm below the line of sight are used for the measurement.) It should be noted that the *white code elements* are detected and used in the processing. They produce a light pattern on the photodiode array; the black lines become the gaps between recorded signals.

The focussing and pointing to the staff is manual. The position of the focussing lens is a measure of the distance to the staff and is recorded by a linear encoder. The focussing lens travels about 14 mm within the focussing range from 1.8 m to 100 m. The position of the compensator is checked by a monitoring system. This compensator monitor generates an error message whenever the compensator is outside its specified operating range. This check ensures that the instrument is levelled with the spot bubble before measurements are taken. The linear photodiode (CCD) array converts the received bar code image into an analogue video signal.

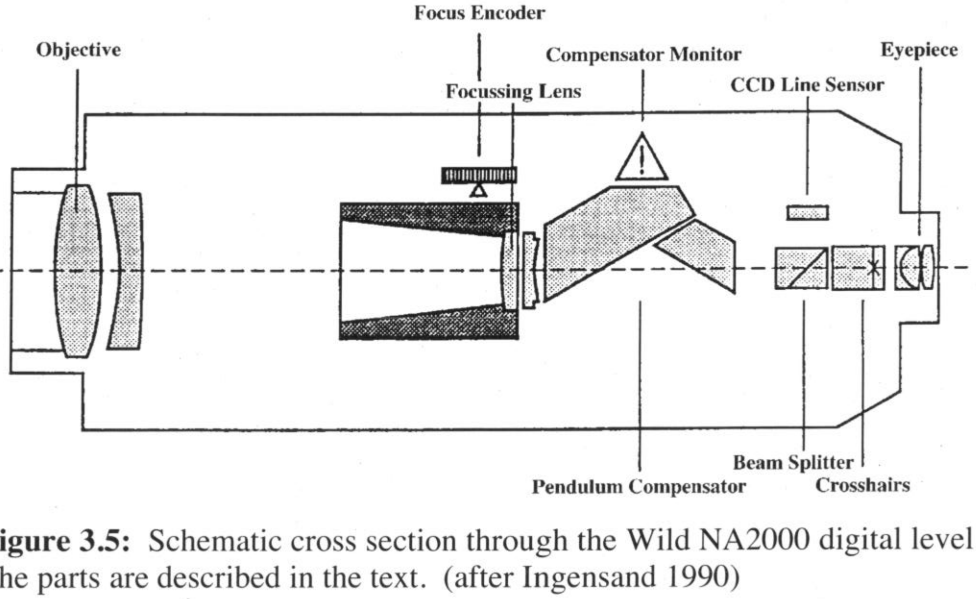

Figure 3.5: Schematic cross section through the Wild NA2000 digital level. The parts are described in the text. (after Ingensand 1990)

A block diagram of the electronics of the instrument is shown in Figure 3.6. A micro-processor performs the real-time image processing required. For the computation of the correlations the processor is supported by a gate array. The read-out electronic amplifies and digitises the video signal. It uses an 8-bit analogue-to-digital converter which provides 256 grey values for each of the 256 pixels.

3.1.3.2 Measuring Sequence

After pressing the measuring button, the position of the focussing lens is read and the proper functioning of the compensator is checked. The integration time for the image recording is selected according to the measured signal intensity. The image of the bar code on the photo diode array is then measured and stored. This takes 0.004 to 1.0 seconds.

The *coarse measurement* follows. Starting with the recorded position of the focussing lens (see d_f in Fi. 3.8), an approximate value for the staff reading and the image scale (distance) is determined. This is basically done by trial-and-error, namely by comparing the recorded image pattern with the simulated image pattern, which is calculated by the computer for a large number of different distances and staff readings. To assess the goodness of fit of the real and the simulated patterns for the different distance and staff readings, the correlation coefficient is computed. This (1-bit) correlation process takes 0.3 to 1.0 seconds.

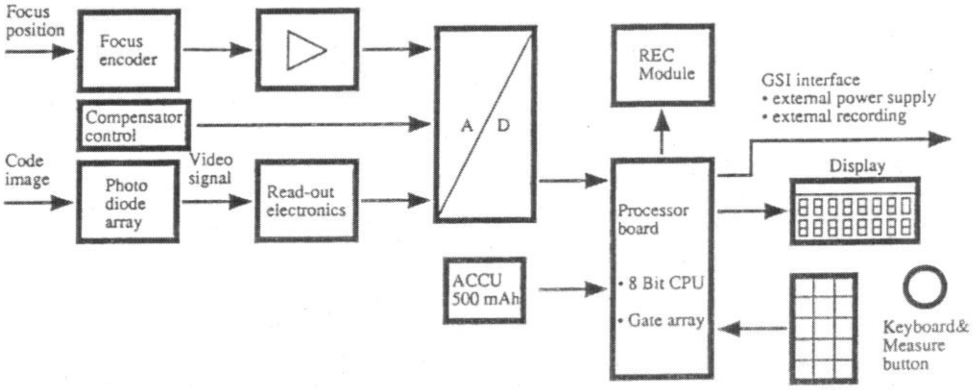
Focus position
Focus encoder
Compensator control
Code image
Photo diode array
Video signal
Read-out electronics
A / D
ACCU 500 mAh
REC Module
Processor board
• 8 Bit CPU
• Gate array
GSI interface
• external power supply
• external recording
Display
Keyboard& Measure button

Figure 3.6: Block diagram of the electronic section of the Wild NA2000 digital level (after Ingensand 1990)

During the final *fine measurement*, the staff reading and the image scale (distance) are determined with full 8-bit correlation between measured and simulated readings of the 256 pixels. The final staff reading and distance are obtained through the application of stored calibration values. The time required for the fine optimisation depends on distance and the quality of the optical signal. The time varies from 0.5 to 1.0 seconds. Depending on the selected measuring mode, the measurements are then processed further and/or displayed and/or stored.

3.1.3.3 *Signal Processing*

The digital level compares the measured grey values of the 256 pixels of the photo detector with the simulated values computed from the known code pattern. Two parameters must be optimised through the correlation process, namely the *staff reading* and the *scale of the staff image*. Changes in the height difference h lead to a *shift of the bar code* image on the photo detector array. Changes in the distance d between the instrument and the staff change the *scale of the bar code* on the photo detector array.

To optimise the two unknown parameters, a two-dimensional correlation function has to be used. In the Wild NA2000, the following correlation function ρ_{PQ} (d,h) is computed between the measured signal Q(y) and the reference signal P(d,y+h) for all 256 pixel values of the array (Ingensand 1990):

$$\rho_{PQ}(d,h) = \frac{1}{N} \sum_{i=0}^{N} Q_i(y) \times P_i(d,y+h) \qquad (3.1)$$

Figure 3.7 shows a typical result of the correlation result over the measuring range of d = 1.8 – 100 m and h = 0.0 – 4.05 m. (The position of the code relative to the detector is denoted by y.) The distance (d_o in Fig. 3.7) and staff reading (h_o in Fig. 3.7) corresponding to the peak of the correlation function are the solutions for the unknown parameters d and h.

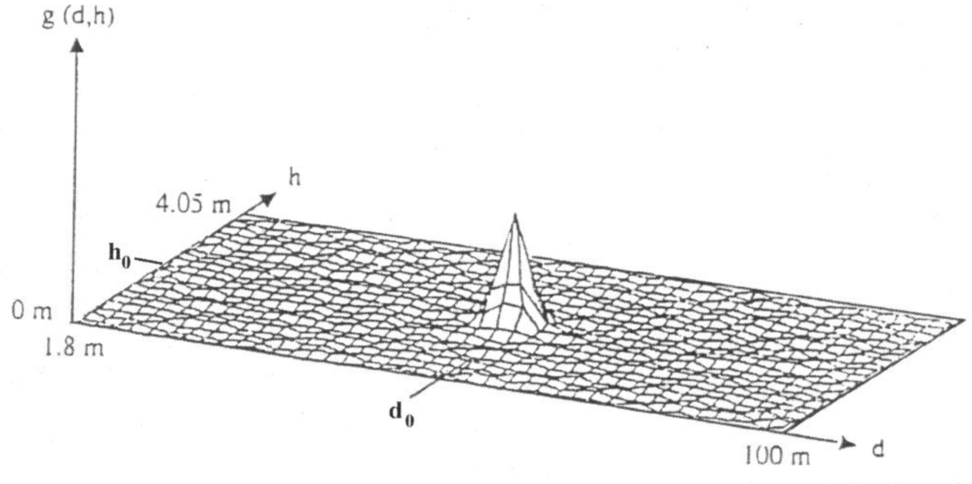

Figure 3.7: Plot of a typical result of the image correlation (ρ) in function of the distance (d) and the staff reading (h) (after Ingensand 1990)

In principle, the entire measuring range of d from 1.8 m to 100 m and of h from 0.0 m to 4.05 m must be checked to find the maximum of the correlation function. To scan the entire measuring range, the correlation coefficient would have to be computed about 50 000 times. To reduce the amount of computing (and measuring time), the search for the maximum correlation is divided in two stages.

The *coarse search* covers a band of about 80% of the total search area. The search band covers the total range of staff readings and is centred about the distance value d_f derived from the linear encoder on the focussing lens. To further reduce the amount of computing, the coarse search is carried out with a 1-bit signal of each pixel only. The original 8-bit signal of each pixel is reduced to 1-bit on the basis of a computed threshold level. In consequence, the multiplications in the above formula can be replaced by an equivalent exclusive-not-or-logic.

The correlation coefficient is computed for each intersection of the distance-height grid within the search band. A peak appears where the reference signal coincides with the measurement signal. This correlation peak is significantly above the noise level.

The *fine search* is used to find the correlation peak with the highest resolution. Within the square fine search area the correlation coefficient is computed with the full 8-bit signal values of each pixel. As the signal levels in the measurement and reference signal are different, the correlation function is normalised as follows (Ingensand 1990):

$$\rho_{PQ}\,(d,h) = \frac{\frac{1}{N}\sum\limits_{i=0}^{N-1} Q_i \times P_i - \bar{Q} \times \bar{P}}{\sqrt{\frac{1}{N}\sum\limits_{i=0}^{N-1} Q_i^2 - \bar{Q}}\;\sqrt{\frac{1}{N}\sum\limits_{i=0}^{N-1} P_i^2 - \bar{P}}} \qquad (3.2)$$

The normalisation through Eq. (3.2) has the added advantage that the correlation coefficient ρ will be in the usual range, that is between –1.0 and +1.0, and that the quality of the measurements can be assessed.

A few other aspects must be considered during the image processing:

(1) The inhomogeneous light intensity of the bar code image must be analysed and taken into account.

(2) Missing code elements have to be detected and omitted from the subsequent correlation computations. (Partial coverage of the staff can lead to missing elements.) Tests have shown that, for sighting distances longer than 5 m, up to 30% of the staff image can be missing without loss of accuracy and reliability. It does not matter where on the image the obstruction occurs.

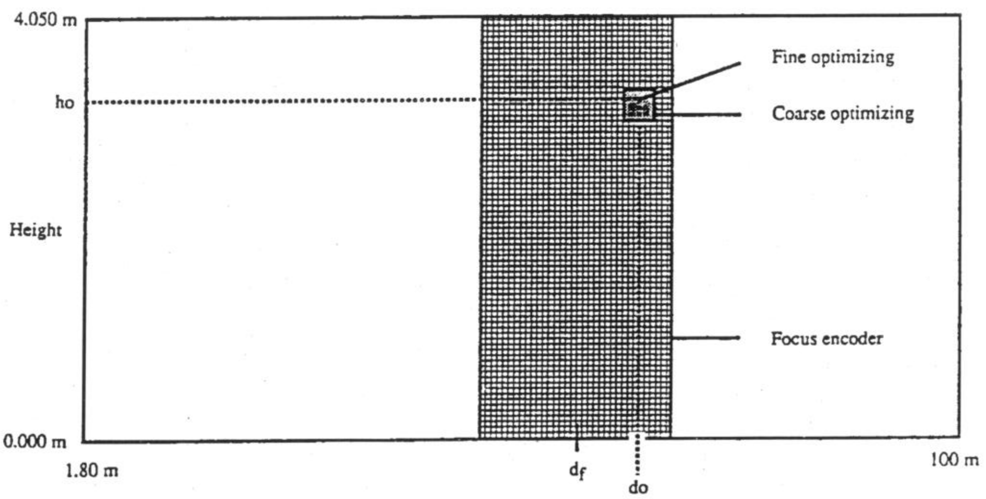

Figure 3.8: Definition of the coarse and fine search areas for the correlation maximum about the coarse distance d_f measured by the focus lens encoder. The final estimates of d and h are denoted by h_o and d_o. (after Ingensand 1990)

3.1.3.4 Accuracy of Bar Code Levels

The accuracy of the measurements depends on the following *internal* aspects of the measuring system (instrument and staff):

- accuracy of the code on the staff
- quantisation/sampling errors/noise
- quality of the optics
- quality and stability of the linear CCD array

The following *external* aspects also affect the accuracy of the system:

- Accuracy of the *Focussing*: Factory tests with the Wild NA2000 have shown that the accuracy of the staff reading is largely independent from the accuracy of the manual focussing. Proper focussing, however, leads to faster measurements. Since other makes of digital levels are more susceptible to poor focussing, it is advisable to properly focus at all times. One of the Topcon digital levels can be bought with an autofocus system.

- The *pointing* of the vertical hair to the centre of the staff is not critical for the reading of the staff. (Pointing becomes important, however, as soon as directions are also measured.) Over 100 m, the photo diode array views only 14.0 mm of the total width of 50 mm of the bar code pattern. Over 2 m, the photodiode array sees only 0.3 mm of the total bar code width of 50 mm. The correct orientation of the staff is also not critical. Even at an erroneous staff orientation of 45°, the apparent width of the graduation still exceed the width of the pixels.

- *Atmospheric Turbulence*: Heat shimmer reduces the contrast of the image of the bar code pattern and distorts the code pattern. Because of the averaging over a large number of graduations, the influence of turbulence is less than in conventional instruments.

- *Tripod Vibrations:* Image vibrations can be caused by the vibrations of the compensator (e.g. due to a set-up near a busy road). The effect of vibrations on measurements depends on the amplitude and the frequency of the vibrations and the speed of the readout of the photodiode array. As the whole image oscillates as a block, averaging over a large number of graduations does not help. (However, increasing the number of measurements does!)
  Like other automatic levels, there are occasions when digital levels cannot be used on vibrating platforms or ground. This case arises when the frequency of the vibrations is close to the resonance frequency of the level's pendulum compensator.

- *Staff Graduation Errors:* The errors of the individual bars are averaged because of the large number of graduations processed. However, the scale error of the staff is not eliminated!

- *Illumination* of the staff is critical. When levelling in the open air, the 'colour' of the light reflected of the staff cannot be influenced. When using artificial lighting, the spectrum of the light source should be similar to that of the sun.
  The intensity variation between full sun exposure, overcast conditions and dusk and dawn must be compensated by a variable integration time of the sensor array. The integration time is varied between 4 ms in full sun and 1 second in poor light. Inhomogeneous illumination of the staff (caused by partial shadowing) is also considered in the processing.

- *Partial Covering* of the bar code is a particular challenge to the processing software: There is no prior knowledge of where, how and how much of the code is blocked out. The software has to recognise missing patterns and ignores those sections in the correlation process. For reasons of reliability, a block out percentage of 30% has been selected as a maximum (above 5 m sighting distance). At least 30 elements (70 mm) of the bar code must be projected onto the sensor array to permit a unique height determination. At sighting distances below 5 m, no block out is acceptable.

- The *repeatability of the staff reading* is distance dependent. A number of tests have been carried out to establish the precision of staff readings. The table below gives the peak-to-peak variation of 20 observations with a Wild NA2000 (Müller 1990). The table shows that sighting distances should not exceed 50-60 m in sunny conditions.

Sighting Distance	10 m	30 m	60 m	80m
peak-to-peak variation				
at +18°C, overcast	0.0 mm	0.2 mm	0.4 mm	0.8 mm
at +27°C, heat shimmer	0.2 mm	0.5 mm	2.8 mm	not possible

Schauerte (1992) established the precision (in millimetres) of the staff reading with a Wild NA2000 digital level as ±(0.021 + 0.012 D[m]), where D is the sighting distance (in metres, not exceeding 35 m).

- The *repeatability of the distance measurement* is also distance dependent.

For the Wild NA2000 the repeatability of the distance measurements to the staff was once reported as <±3 mm for distances below 20 m, <±10 mm for sighting distances less than 40 m and ±10-30 mm for distances between 50 m and 100 m. These values give an indication of the trend with distance.

3.1.4 Bar Code Staffs

Originally, only glass fibre staffs (Wild GKNLE4) were available for the Wild NA2000. These staffs come in three sections of 1.35 m length each. The three plug-in sections give a total length of 4.05 m. Later, Leica added the invar staffs Wild GPCL3 and GPCL2 as options for the more accurate versions of the Wild NA2000 (e.g. Wild NA3000) and aluminium and wooden staffs for normal levelling. (Zeiss offers folding bar code staffs for their digital levels.)

3.1.4.1 Leica Normal Bar Code Staffs

The basic bar code element is 2.025 mm wide. The 4 m fibre glass staff features 2000 elements over 4050 mm. Leica uses a pseudo-stochastic code for their staffs. (Topcon and Zeiss use different coding principles.) On the reverse side, standard metric (10 mm) or imperial (0.1 feet) faces are available. The Wild instruments and staffs permit unique measurements for sighting distances between 1.8 m and 100 m. The manufacturer claims a coefficient of expansion of less than 10 ppm/°C for the fibre glass staffs. Based on practical measurements, the manufacturer reported coefficients of expansions of <8 ppm/°C and <0.2 ppm / (percent relative humidity).

The temperature coefficient of the fibre glass staff is worse than that of the traditional wooden staffs but better than that of aluminium staffs. On the other hand, the humidity coefficient of (poorly painted) wooden staffs can be much worse. Leica also sells 4 m wooden and aluminium staffs. The manufacturer quotes a coefficient of expansion of 10-20 ppm/°C for the former (depending on humidity) and 24 ppm/°C for the latter.

3.1.4.2 Leica Invar Bar Code Staffs

The bar code invar staffs are constructed in the same way as all modern invar staffs, namely with an invar ribbon recessed in an aluminium housing and a hardened and rust-free base plate. The usual length of the invar staffs supplied by Leica is 3.0 m or 2.0 m. (For industrial surveys, thin staffs of 60 cm, 92 cm and 182 cm length are also supplied.) The coefficient of expansion of these staffs is less than 1 ppm/°C. The errors of the individual bar code graduations (the basic interval is again 2.025 mm) are less than 10 μm because the bar code is manufactured with the aid of laser interferometry.

3.1.4.3 Errors of Normal Bar Code Staffs

The errors of the bar code staffs are no different from those of the traditional levelling staffs. These errors are listed for convenience. See surveying textbooks for a detailed discussion.

- Zero error
- Random errors of individual graduations
- Systematic graduation errors, e.g. scale errors in function of humidity, temperature and time
- Junction errors (of sectional, telescopic and folding staffs)
  Practical investigations have found errors of 0.1 mm to 1.0 mm across junctions of the Wild fibre glass sectional staff (!!!)
- Maladjustment of spot bubble
- Other errors such as bending, kinks

3.1.4.4 Errors of Invar Bar Code Staffs

The errors of invar bar code staffs are similar to those of classical invar staffs, which are also produced by modern manufacturing techniques (laser engraving under control of a laser interferometer). The support structure of modern invar staffs is made from aluminium and, thus, not sensitive to humidity changes.

It is expected that the coefficient of expansion is below the 1 ppm/°C level and the random graduation errors at the 10 μm level. Maurer & Schnädelbach (1995) quote an average coefficient of expansion of 0.75 ppm/°C for invar staffs manufactured after 1992. Although the coefficient of expansion is small, it varies from one staff to the next and needs to be calibrated for high precision levelling over large height differences.

3.1.5 Errors in Levelling with Digital Levels

The errors of digital levels are similar to those of automatic levels. Only a listing of most errors is given below. The readers are referred to textbooks for a complete discussion. The following books are suggested for further reading on this topic: Cooper, M. A. R. 1982, Modern Theodolites and

Levels, Granada, London, ISBN 0-246-11502-5, Sections 9.6 and 10, pp. 216-234 and Deumlich, F. 1982. Surveying Instruments, Walter de Gruyter, Berlin, ISBN 3-11-007765-5, pp. 179-211. The listing of the errors in levelling is based on the tables published by Schauerte (1993).

For general purpose levelling most errors of automatic and digital levelling instruments can be ignored. The collimation error must be considered, however, if the sighting distances in the foresight and backsight are not strictly the same. The scale errors of staffs and the refraction errors must be considered if large height differences are levelled. When using automatic and digital levels for precise levelling, all error sources must be considered. In consequence, the operator should be familiar with the properties of such levelling instruments.

3.1.5.1 Errors of Bar Code (Digital) Levels

- *Electronic Collimation Error*

Remedy: Observe strictly from the middle between the two adjacent staff set-ups (or determine the collimation error before and after the measurements and post-correct all measurements using the recorded sighting distances.) Note that this error is temperature and time dependent!

There is only one sure way of eliminating the effect of the collimation error on the height difference measured from one set-up: **set up in the middle**! This allows to use the same focus in BS and FS and avoids mechanical shocks between BS and FS.

Furthermore, the measuring procedures must be selected so that temperature changes at one set-up are minimised. If the **instrument is kept in the shade** at all times, there should not be any temperature induced changes of the collimation error between BS and FS.

- *Maladjustment of the Spot Bubble*

Remedy: At the beginning of a levelling run, **adjust the circular bubble** if it moves relative to its case when the instrument is turned around its vertical axis.

- *Compensation and Height Error* (leads to the 'obliquity of horizon')

Remedy: Observe in such a way that the error generates a systematic effect and eliminate it through an appropriate observation techniques such as the 'red trousers' method

BFFB (odd numbered set-ups)
FBBF (even numbered set-ups)

or the modified classical 'Wild-Zeiss' method

BFFB (at all set-ups)

(underlined = *level the spot bubble whilst the instrument is pointing to this staff*, B = *backsight*, F = *foresight*)

- *Measurements near the Top and Bottom of Staffs*

Digital levels are programmed to be able to measure near the bottom and top end of staffs since ordinary levels can measure to any graduation on a levelling staff. Users have found that digital levels can measure to a staff, even if the line of sight is slightly above the top end (or slightly below the bottom end) of a staff. Since the digital levels can use only half the measuring sector when measuring to the ends of staffs, the precision of staff readings near the ends of staffs may be reduced and systematic errors may occur.

Figure 3.9a shows some experimental results with a Leica NA3003 near the bottom end of a Leica invar staff. At 10 m sighting distance, the maximum error is about 0.4 mm and at 30 m, the error may be as large as 1.2 mm (even though the instrument displays staff readings to 0.01 mm). Figure 3.9b shows the field of view of the level should extend to a maximum of 20% (of the field of view diameter) above the highest useable staff graduation.

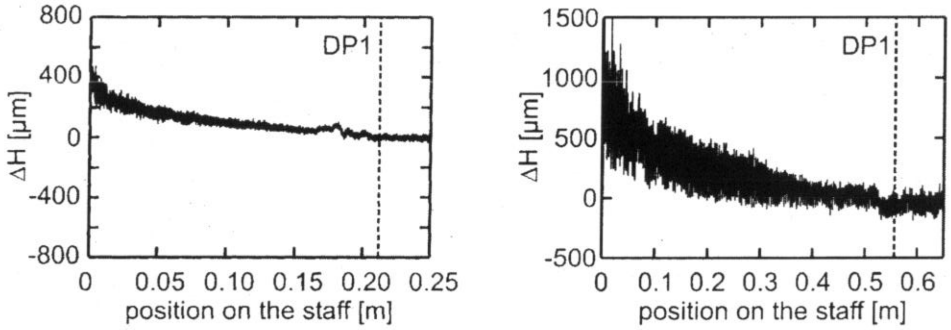

Figure 3.9a: Errors when measuring to the bottom end of a Leica invar staff with a digital level Leica NA3003 at a sighting distance of 10 m (left) and 30 m (right). For measurements to the right of the DP1 line, the full measuring sector falls on the staff. (after Woschitz 2003)

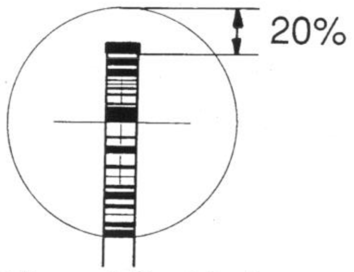

Figure 3.9b: Maximum reading on staffs as suggested by Leica for precise levelling with the NA2000/3000 series of digital levels (Leica 1996)

Figure 3.9c shows some experimental results with a Topcon DL-101C near the top end of a Topcon invar staff. At 10 m sighting distance, the maximum error is about 0.8 mm and at 30 m, the error may be as large as 0.7 mm. Note that, at 10 m, the staff readings are too small near the top end of the staff and, at 30 m, too large. Woschitz (2003) recorded a gross error of more than 2 metres during a similar test at the bottom end of the Topcon staff. Figure 3.9c (after Woschitz 2003) indicates what range of staff readings should be avoided with Topcon digital levels near the top of 3 m Topcon digital levelling staffs.

Similar tests by Woschitz (2003) with a Zeiss(Trimble) DiNi11 digital level showed a maximum error of about 0.2 mm at 10 m and no errors at 30 m. Ingensand (2001) reported additional errors (to 0.2 mm) of Zeiss (Trimble) DiNi levels near the top end of staffs caused by artificial lighting in tunnels.

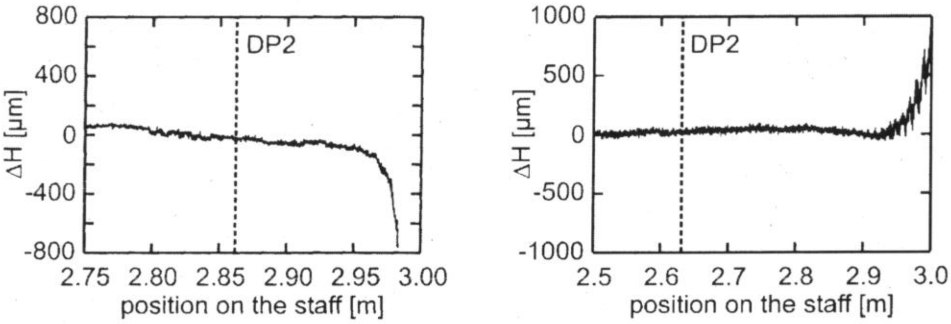

Figure 3.9c: Errors when measuring to the bottom end of a Topcon invar staff with a digital level Topcon DL-101C at a sighting distance of 10 m (left) and 30 m (right). For measurements to the left of the DP2 line, the full measuring sector falls on the staff. (after Woschitz 2003)

- *Systematic Staff Reading Errors*

Soon after the release of the first digital levels, periodic staff reading errors were noted at sighting distances of 15 m and 7.5 m. The effect was thoroughly investigated later by Woschitz (2003). Figure 3.9d shows how the errors of the staff reading change with height in line with the projected pixel size on the staff. The largest errors occur when the fundamental staff interval coincides with the (projected) pixel interval. The largest peak-to-peak variation of 0.81 mm occurs at 14.98 m. Similar effects are found at sighting distances of around 30 m and around 7.5 m. Equivalent tests in the 'extended system accuracy mode' showed reduced peak-to-peak variations of 0.30 mm (at 15 m). At other sighting distances, the peak-to-peak variation is much less, namely 0.1 to 0.2 mm in 'standard system accuracy mode'. Clearly, these sighting distances should be avoided when using the Wild/Leica NA200x/NA300x series of instruments for precise levelling.

Woschitz (2003) also investigated the systematic errors in Topcon digital levels. Some experimental results are shown in Fig. 3.9e. The error patterns at 5 m (and below) and 11 m (and above) are quite different. It is known that the Topcon digital levels change the way they derive the staff reading at about 9 m (Rüeger 2000).

Figure 3.9e seems to indicate that the short range measuring mode is free of systematic errors. At 7.0 m, the Topcon level seems to switch occasionally to the long range measuring mode. At 9.0 m, the level seems to switch continuously between short and long range mode. Woschitz determined a mean staff reading offset between the two modes of 0.19 mm. At sighting distances of 11 m and more, the Topcon levels show a periodic error pattern with a period of 300 mm and a peak-to-peak variation of 0.18 mm.

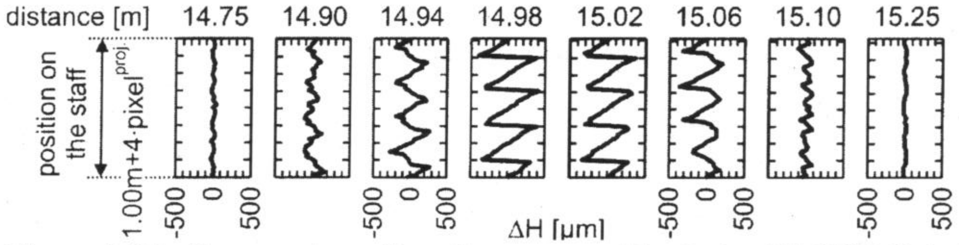

Figure 3.9d: Systematic staff reading errors with a Leica NA3003 digital level in 'standard system accuracy mode', at sighting distances near 15 m. (after Woschitz 2003)

The amplitude of the sine pattern decreases with distance and cannot be distinguished any more at about 30 m. In practice, sighting distances between, say, 7 m and 11 m should be avoided in precise levelling since one measurement may differ from another by 0.36 mm because of the switching between the two processing modes and the 'cyclic error' of the long range mode.

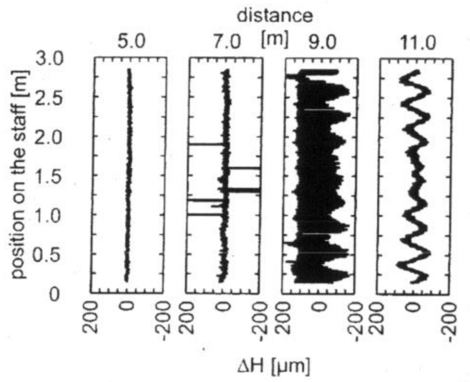

Figure 3.9e: Systematic staff reading errors with a Topcon DL-101C digital level at sighting distances from 5 m to 11 m. (after Woschitz 2003)

A Zeiss/Trimble DiNi11 was also investigated by Woschitz (2003). Periodic errors similar to those in Fig. 3.9d were found, but with much smaller amplitudes. Periodic patterns were found at 5 m, 9, 11 m, 19 m and 30 m. The maximum peak-to-peak variation of 0.1 mm was recorded at a sighting distance of 19 m. These errors are sufficiently small to be ignored in levelling practice.

- *Influences on the Setting Behaviour of the Compensator* (wind, traffic)

Remedy: Increase the number of staff readings: observe at least two backsight readings (BS) and two foresight readings (FS) so that the possible influences become evident in the two height difference (BS-FS) values. In precise levelling, one would measure, for example, BB FFFF BB or even B FF BB FF B.

- *Influence of Magnetic Fields* caused by direct or alternating currents

The magnetic field of the Earth has to be considered at all times. Additional fields are present near electrical installations (e.g. transformer stations). The effect of the Earth's magnetic field on compensators of levelling instruments was first reported in 1981. Laboratory tests at a field strength of 23000 nT (nanoTesla) established the following systematic errors of particular instruments for N-S levelling runs due to the Earth's magnetic field: Zeiss Ni2 - 0.2 mm/km; Wild NAK2 –> 0.5 mm/km; Zeiss Ni002–> 1.5 mm/km; Zeiss Ni1 –> 2.2 mm/km. Since 1982, all manufacturers have reduced the susceptibility of their instruments to magnetic fields by better shielding and/or the use of non-magnetic materials for the compensators. Post-1982 automatic levels (such as all digital levels) are likely to be much less affected by the Earth's magnetic field. Recent tests with a Wild NA3000 found a residual magnetic effect of 'only' 0.15 mm/km on N-S lines.

Remedy: Test in a laboratory if magnetic fields have an effect on the collimation error.

- *Temperature Effects on the Compensator and/or the Photo Detector/CCD Array* (e.g. due to lack of temperature acclimatisation or sun exposure during transport from set-up to set-up). Temperature may change the electronic collimation error and/or generate a vertical shift of the line of sight.

Remedy:

–> Shade the instrument at each set-up (and also during transport) with an umbrella.

–> Increase the number of measurements to the staffs so that jumps in the compensator's rest position can be detected in multiple BS-FS values.

–> Measure strictly from the middle between the forward and backward staff positions and use time-symmetric measurements for the BS and the FS, such as the 'red trousers' method or the modified 'Wild-Zeiss' method (see above).

- *Sinking of the Tripod*

Remedy: Employ a measurement sequence which features time-symmetric BS and FS measurements and do not set up instruments on warm bitumen surfaces.

3.1.5.2 Errors of Invar Bar Code Staffs

- *Graduation Errors in Invar Bar Code*

Remedy: As modern invar staffs are manufactured under laser interferometer control, only small random errors are to be expected. Because of the averaging over a large number of individual code elements, this small effect is reduced even further. In consequence, line-by-line calibration is not required.

- *Difference of the Zero Errors of the two Staffs* (in precise levelling)

Remedy: Use an even number of instrument set-ups between the bench marks (BMs) and always use the same staff for the connection to BMs. Do not change the staffs for the backwards run.

- *Change of the Length of the Bar Code due to Temperature*

(leads to variable height errors which depend on the staff reading)
Remedy:
– Calibrate the composite coefficient of expansion of the bar code scale (as the sum of the expansions of the invar ribbon and the aluminium case and changes in spring tension) with the staff in a vertical position. Alternatively, the data of a factory calibration may be used or an average value of +0.75ppm/°C (after Maurer & Schnädelbach 1995).
– Measure the temperature of the invar ribbon (e.g. at three heights above ground) of each staff and at each staff set-up and correct (post-mission) all staff readings accordingly using the appropriate coefficient of expansion.

- *Sinking of the Staff during the Observations from one Instrument Set-Up*

Remedy: Ensure a stable marking of the change points. Employ an observation schedule with time–symmetric BS-FS measurements.

- *Sinking of the Staff during the Move of the Instrument to the next Set-Up*

Remedy: Ensure a stable marking of the change points. This effect becomes evident in a comparison of the total height differences obtained in the forward and the backward runs. It cannot be eliminated by any other mean than a double run (–>,<–) levelling.

- *Skewness and Non-Flatness of the Foot Plate of the Staff ('Shoe')*

Remedy: Test the flatness of the foot plate of the staff. If large variations are detected, get foot plate changed. Minimise this effect (and the possible eccentricity between the optimal set-up point and the middle of the foot plate) by setting up of the staff at the point where the centre line of the invar ribbon intersects the foot plate. The manufacturer may provide a ring to achieve a correct set-up. (Not possible for BMs in walls.)

- *Maladjusted Spot Bubble of the Staff*

Remedy: Test the state of the adjustment prior to the measurements and adjust the spot bubble of the staff, if required. With invar levelling staffs, insure that the staff is set up so that the axis of the invar ribbon is in the centre of the benchmark or change plate. Refer also to Fig. 3.17 in the exercise section at the end of this chapter.

- *Movement of the Staff* (due to Wind or Traffic)

Remedy: It is essential to support of staff in two perpendicular directions. Buy optional struts or use two ranging rods (at right angle) for the purpose.

3.1.5.3 External Influences on Line of Sight

- *Grazing Rays (–> Levelling Refraction)*

Digital levels are likely to be more susceptible to refraction than other levels because they use graduations below the line of sight for the measurements. Apart of this, the digital levels suffer from the same refraction effects as all other levels. These effects may be summarised as follows (Angus-Leppan 1984):

– Levelling on a Plain: The refraction effect can amount to 1 mm over 50 m but is cancelled if the sighting distances to the foresight and the backsight staffs are strictly the same.

– Levelling on E–W Slopes: When going uphill, the foresight is more affected by refraction than the backsight. The effect is worst on gentle slopes where the slope still permits to maintain the tolerated maximum sighting distance. Errors in of 20 to 30 mm per 100 m height difference may occur. The signs of the errors in the rise and the descend over a hill are likely to be opposite so that the error may cancel in the height difference between points on both sides of the hill.

– Levelling on N-S Slopes: The refraction effects in ascend and descend no longer cancel because the sun is predominantly north or south, depending on the hemisphere. The accumulated errors vary depending on slope, season, climate and latitude; they can reach 50 mm to 1000 mm per 1000 km N-S levelling.

Recommendation: The horizontal line of sight should be more than 1 m above ground with the Wild NA2000/3000, as is uses ±0.18 m per 10 m above and below the horizontal line of sight. Reduce the sighting distance, if this condition is not fulfilled. This procedure ensures that the refraction effect with the Wild NA2000/NA3000 is not larger than with conventional levels. The refraction errors listed above may be reduced, but not cancelled, by this procedure.

- *Terrain below the Line of Sight*

Remedy: Rough surfaces (like grass) are preferred. Avoid measuring across bitumen and concrete surfaces.

- *Obstruction of Bar Code Section in Use*

Remedy: Ensure free observation sector. With the Wild NA2000/NA3000 no obstructions are tolerated at sighting distances of less than 5 m.

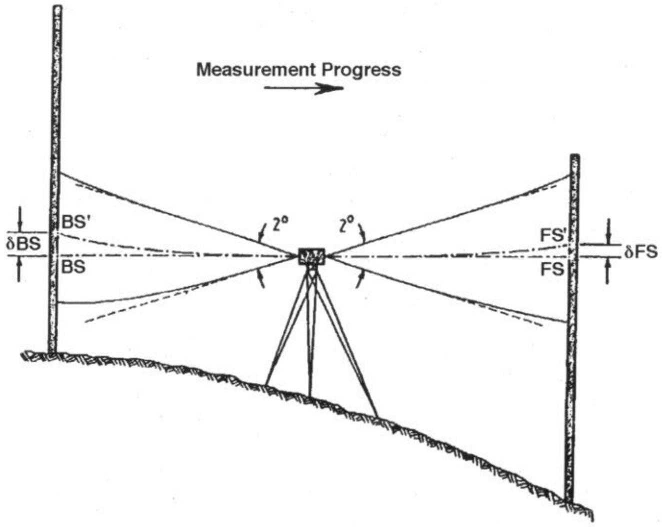

Figure 3.9f: Refraction effect on levelling on a slope. The errors caused by refraction are denoted by dBS and dFS. BS and FS are the true backsight and true foresight readings, respectively. BS' and FS' are the erroneous actual readings. The 2° angles indicate the viewing angle of the Wild NA2000/NA3000 digital levels. (after Schauerte 1993)

- *Sun Exposure* (–> heat shimmer)

Remedy: Select the optimal time of the day, the best season and the most appropriate weather conditions for any precise levelling.

- *Insufficient Illumination of the Staff*

Remedy: In poor light outdoors or under artificial lighting indoors (in tunnels, halls, buildings) the staffs must be illuminated with a lamp with 'daylight' spectrum.

3.1.5.4 Errors of the Observer

- *Inaccurate Focussing*

Remedy: The tolerance range for the focussing increases with distance. At sighting distances of less than 5 m, focussing errors can lead to 0.1 mm errors in height in Wild (Leica) NA200x/NA300x instruments. Woschitz (2003) noted that poor focussing with Zeiss (Trimble) levels increases the periodic error in these instruments. For example, a peak-to-peak variation of 0.21 mm is mentioned for a pointing of 0.5 m in front or behind the staff over 10.2 m.

- *Recording of Erroneous Observations* or omission of the recording of additional information

Remedy: Book such occurrences in handwritten notes or through the storage of additional information in the measurement file.

3.2 Laser Levels

Laser levels in general and rotating laser levels in particular have become an important surveying tool in building, construction and agriculture. In fact, rotating laser levels have almost replaced the classical levels on building and construction sites. This is reflected by the fact that most manufacturers of surveying equipment now produce rotating and other laser levels in great numbers. Although they are more expensive than the normal levelling instruments, they are much more efficient as they provide effectively a one-person levelling system. The rotating laser level consist of a tripod mounted instrument which, after levelling, rotates (about a vertical axis) a visible or infrared beam in a horizontal plane. It typically rotates at a rate of a few hundred revolutions per minute (rpm). Some models have adjustable rotation rates (e.g. between 0 and 500 rpm). In the case of visible (red) beams, the red dot can be seen on any surfaces 'hit' by the beam. For laser levels, that transmit an infrared beam, a special detector has to be used to detect the beam. A similar (or sometimes the same) detector can be used to increase the range and the precision of the visible beam instruments.

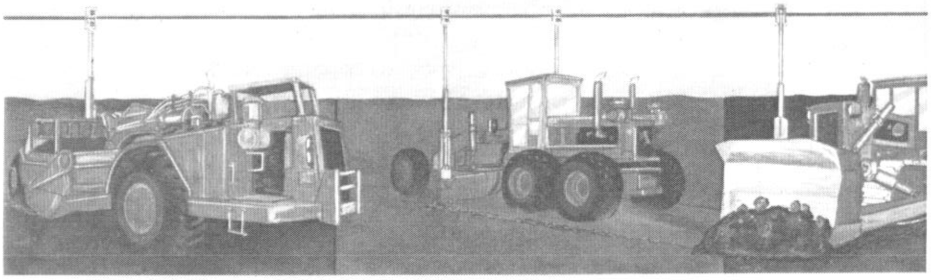

Figure 3.10: Use of laser levels in construction and agriculture (after Laser Alignment Inc 1986)

This detector is usually moved up and down a wall or a levelling staff until the detector is centred on the horizontal laser plane. Most detectors provide up-down guidance in visual and acoustic form to assist the centring and to provide better precision. For example, the detector may generate a fast beeping noise, if it has to be moved down to the laser plane, a slow beeping noise, if it has to be moved up, and a continuous tone if it is centred on the horizontal plane. Some detectors may provide different sensitivity settings. A high sensitivity provides higher accuracy under good measuring conditions (e.g. short range, indoor). For their LPD2 detector, Leica quotes a setting precision of ±0.8 mm for the high sensitivity setting and ±2.4 mm for the low sensitivity setting (long range, outdoors). The overall accuracy of a well adjusted rotating laser level is quoted by Leica as ±2 mm at 50 m and ±4 mm at 100 m. Some building site rotating laser levels can be mounted horizontally to provide a vertical reference plane.

Outside building sites, laser levels can be used to guide or control precision grading machines in construction work (excavators, graders, bulldozers, scrapers) with horizontal, single or dual grade surfaces. These laser levels typically operate over longer ranges than those on building sites and can be set to mark non-horizontal planes. The detectors are mounted on the cutting tools and the results displayed in the drivers cabin. Another use of these devices is for land forming in agriculture, again with horizontal, single or dual grade surfaces.

The range of laser levels depends on the output power of the lasers or laser diodes used, on the atmospheric conditions and on the sensitivity of the laser detector used; it can be as much as 450 m. All systems have a much larger range at night. At longer distances, the effects of earth curvature and, in particular, refraction must be considered. This aspect will be discussed later.

Laser levels come in many forms and are not restricted to rotating laser levels. Linear laser levels are used for pipe laying, for example. They fit into pipes and project a laser beam at a selectable gradient. Pipe laying lasers provide not only the levelling but also the alignment of the pipes.

The simplest form of laser levels are hand-held and project a laser beam parallel to the principle axis of an attached tubular level. If the bubble of the tubular level is centred, the beam is horizontal. Mounting the pocket laser level on a (horizontal) swivel base permits to project a horizontal plane. Attachment of a right angle beam splitter even allows the setting out of right angles. The figure below shows one of these simple devices.

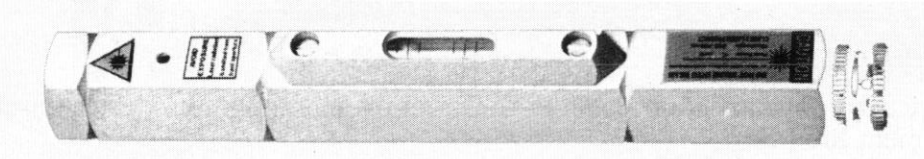

Figure 3.11: Pocket Laser Level with a visible laser diode emitting at 635 nm, 150 mm long, daytime range 30 m (after Leica Instruments Australia)

3.2.1 Construction of a Rotating Laser Level

Figure 3.12 shows the cross-section through a typical rotating laser level (Telamat by the German company Theis). The infrared beam is produced by a GaAlAs diode (nominal wavelength 780 nm, output power 0.3 mW) at the bottom of the housing. This diode is in the focal point of the two-lens objective directly above it. The figure shows also one of the two adjustment screws for the diode. (If the instrument is out of adjustment, these adjustment screws can be used to make the laser beam horizontal in two perpendicular planes.) The laser beam travels straight up, through the lens system and to the pentagonal prism. This prism deviates the beam by 90° and rotates at a few revolutions per second. A belt connects the rotating part to the drive motor. The horizontal beam leaves through the glass plates on all sides. These glass plates protect the rotating part form the elements.

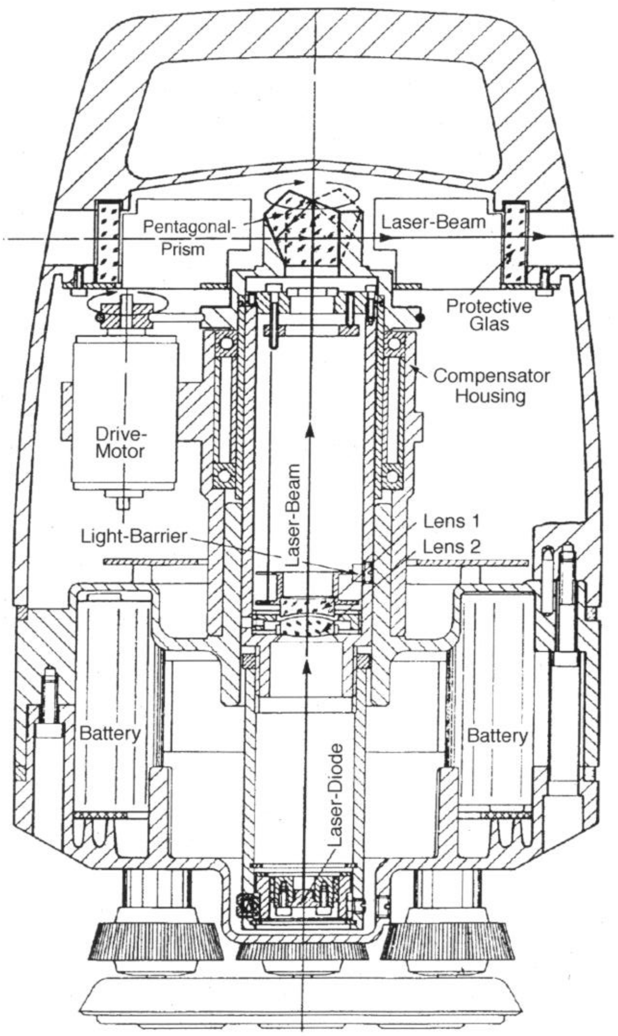
Pentagonal-Prism
Laser-Beam
Protective Glas
Compensator Housing
Drive-Motor
Laser-Beam
Lens 1
Lens 2
Light-Barrier
Battery
Battery
Laser-Diode

Figure 3.12: Cross-Section through a rotating laser level Theis Telamat (after Wüller 1988)

Figure 3.12 also shows that the upper (concave) lens is suspended on wires, which are attached below the rotating prism. The hanging lens acts like a two-dimensional pendulum and is used as a so-called compensator. Like in the traditional automatic levelling instruments and the newer bar code levels, the compensator compensates the laser beam for any small non-verticality of the case of the instrument, which is typically levelled by a spot bubble (circular level) only. Light barriers check if the pendulum is freely hanging. Most instruments will not operate until the instrument is levelled accurately with the aid of the spot bubble.

The instrument shown has rechargeable batteries on-board, which can be charged with the appropriate charger and cables. Most manufacturers also provide cable connections and, sometimes, DC/DC converters for running the laser levels of 12 V car batteries. The range of the level is quoted as at least 120 m.

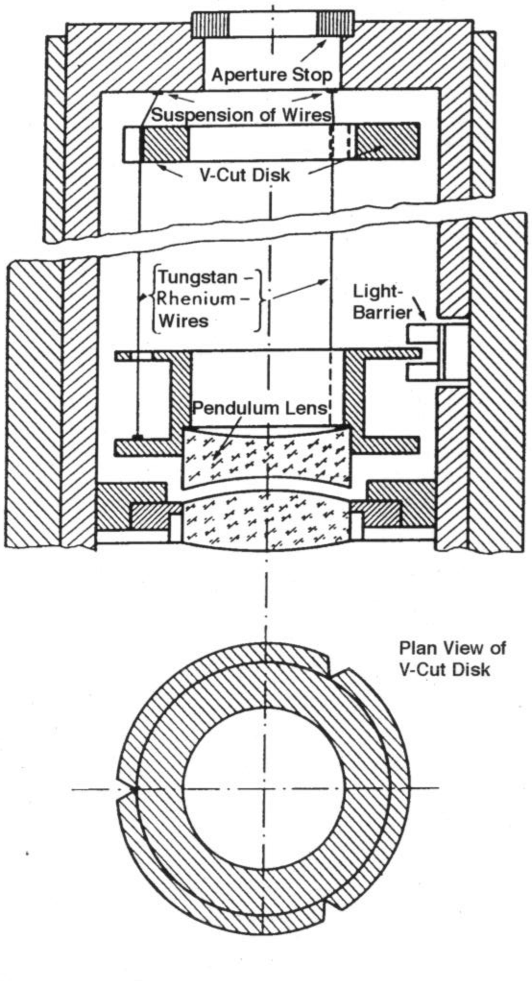
Aperture Stop
Suspension of Wires
V-Cut Disk
Tungsten – Rhenium – Wires
Light-Barrier
Pendulum Lens
Plan View of V-Cut Disk

Figure 3.13: Cross-Section through the compensator of the Theis laser level Telamat (after Wüller 1988)

3.2.2 Design of a Two-Axis Compensator of a Rotating Laser Level

A more detailed diagram of the compensator of the Theis laser level 'Telamat' is given in Fig. 3.13. Figure 3.13 shows that three wires support the pendulum lens. The correct spacing and orientation of the wires is provided by the disk with the three V-cuts. The pendulum features magnetic damping and has a working range of 0.5 degrees of arc.

Figure 3.14 explains the working principle of the two-axis compensator. The beam axis of the laser diode is not coincident with the plumbline through the diode because of the small residual dislevelment α of the instrument after levelling with the spot bubble. In consequence, the laser beam travels at an angle of α to the plumbline towards the suspension point P of the concave pendulum lens. (In a perfectly levelled instrument, the point P would be strictly above the diode.) The laser beam is then refracted in the pendulum lens. Because of the type of suspension used, the plane of the pendulum lens is at right angle to the line between the diode and the suspension point P. The centre of the lens is directly below the suspension point P. Knowing this, the focal point of the lens can be plotted at the focal distance f from the centre of the lens. Because the incident beam from the diode is parallel to the optical axis of the lens (see 'f' in figure), the laser beam exits in a direction defined by the focal point and the point of intersection between the plane of the lens and the incident laser beam. It follows from the figure that the laser beam is 'bent' by an angle β (rather than the dislevelment angle α) at the lens.

3.2.3 Major Errors of Rotating Laser Levels

3.2.3.1 Compensation Error

The compensator of the laser level corrects most of the dislevelment. Some small residual error $(\alpha - \beta)$ remains. This *compensation error* $(\alpha - \beta)$ causes the laser beam rotate in an oblique plane at an angle of $(\alpha - \beta)$ against the horizon. This is sometimes referred to as *obliquity of horizon*. The relationship between the α and β follows from Figure 3.14 as (Wüller 1988):

$$x = PL \sin\alpha = f \tan\beta \qquad (3.3)$$

$$\beta = \arctan \{(PL/f) \sin\alpha \qquad (3.4)$$

$$\alpha - \beta = \alpha - \arctan \{(PL/f) \sin\alpha\} \qquad (3.5)$$

Equation 3.5 shows that β becomes zero for zero dislevelment α. Considering the close agreement between tan and sin values for small angles, the angle β would always be close to α and the compensation error very small. Further, the designer can select the fraction (PL/f) in such a way that the compensation error is reduced further. However, as Fig. 3.14 and Eq. (3.5) show, the distance PL has to be close to the focal distance f. In the case of the WILD LNA2, the obliquity of horizon is adjusted if ($\alpha - \beta$) exceeds 1.25 mm over 30 m (8.6 seconds of arc).

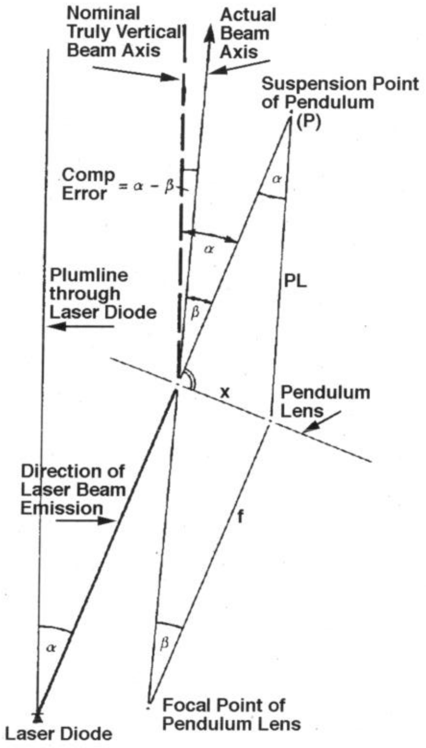
Nominal
Truly Vertical
Beam Axis
Actual
Beam
Axis
Suspension Point
of Pendulum
(P)
Comp
Error
$= \alpha - \beta$
α
α
Plumline
through
Laser Diode
β
PL
x
Pendulum
Lens
Direction of
Laser Beam
Emission
f
α
β
Laser Diode
Focal Point of
Pendulum Lens

Figure 3.14: Two-axis compensator principle in Theis rotating laser level Telamat (after Wüller 1988)

3.2.3.2 Adjustment Conditions of Rotating Laser Level

Before the adjustment conditions of a rotation laser level can be introduced, three axes need to be defined. The *centre of the beam* is the centre of gravity of the beam intensity in the plane perpendicular to ray. The laser *beam axis* is the line between the centre of the beam at the target and the centre of the beam in the (rotating) pentagonal prism (or mirror). (This laser beam axis corresponds to the 'line of sight' of ordinary levelling instruments.) The *rotation axis* is the axis of rotation of the pentagonal prism or mirror assembly and corresponds to the 'vertical axis' of theodolites.

An rotating laser level is in perfect adjustment when:

- the axis of the pentagonal prism coincides with the rotation axis,
- the radiation axis of the (infrared or visible) laser diode coincides with the rotation axis,

- the optical axis of the biconvex lens coincides with the rotation axis.

The first two adjustment conditions are illustrated below in Figs. 3.15 and 3.16. The non-coincidence of the two respective axes is described by an linear eccentricity error and an angular misalignment. It is evident from Fig. 3.15 that, both, the eccentricity and the angular misalignment of prism and laser beam axes have no effect on the definition of a horizontal plane.

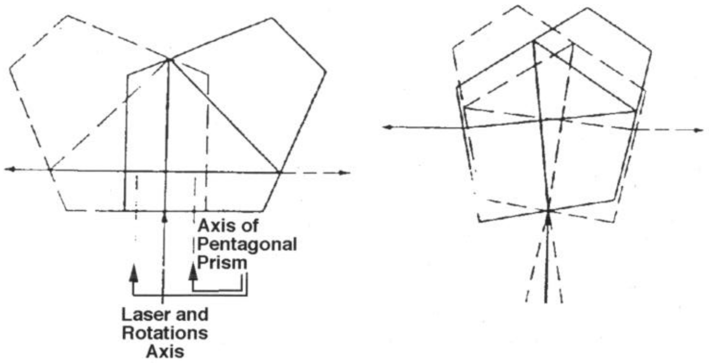

Figure 3.15: Effects of the non-coincidence of the axis of the pentagonal prism and the rotation axis. Left: Effect of the eccentricity of the axis of the pentagonal prism relative to the rotation and the laser axes; Right: Effect of the angular misalignment between the axis of the pentagonal prism and the laser and rotation axes. (after Noack 1989)

The non-coincidence of the laser beam and the rotation axes in Fig. 3.16 are more significant. The eccentricity e of the laser beam relative to the rotation axis of the prism causes the horizontal plane projected by the laser beam to change its height by 2e over a 360° rotation of the prism. In the Theis Telamat, this *height error* is minimised by an aperture stop below the rotating prism. The angular misalignment ε of the laser beam relative to the rotation axis of the prism causes a tilt of the projected laser plane of the same angle ε and a height shift of the beam of 2k over a 360° rotation of the prism. The beam shift of 2k can be computed from the dimensions of the pentagonal prism and the incident angle ε. Considering the principle of the compensator shown in Fig. 3.14, is evident that ε corresponds to the angle β and is very close to the angle α (the dislevelment of the laser level). After accurate levelling of the laser level with a spot bubble, ε should not exceed a few minutes of arc and would cause a minimal height error k.

3.2.3.3 Misalignment of a Rotating Mirror

If the rotating laser level is equipped with a rotating mirror (at 45° to the rotating axis) rather than a rotating pentagonal prism, then any error in the 45° angle will generate an error which is analogous to a collimation error in a ordinary level. The laser beam will travel on a conical rather than plane surface. The Spectra Physics EL-1 Electronic Level System is such an instrument.

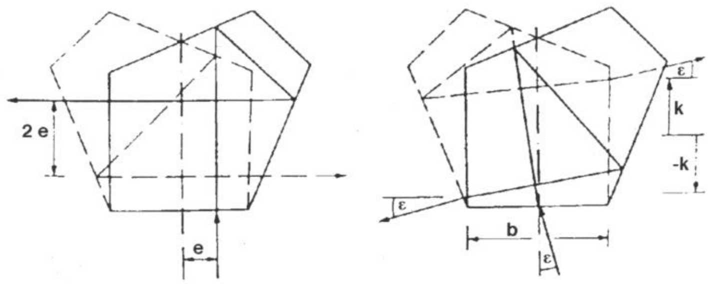

Figure 3.16: Effect of the non-coincidence of the laser beam and the rotation axis. Left: Effect of eccentricity e of laser beam relative to the rotation axis. The resulting height error amounts to ±e. Right: Effect of the angular misalignment ε of the laser beam relative to the rotation axis. The magnitude of the resulting height error is ±k. (after Noack 1989)

3.2.3.4 Secondary Height Error

Because the intensity distribution across the originally vertical laser beam is not uniform, different parts of the beam will be used by the laser detectors depending on azimuth to derive the up-down information. For the Theis Telamat the peak-to-peak variation due to this error is stated as less than 0.5 mm.

3.2.4 Calibration of Rotating Laser Levels

As the main part of rotating laser levels cannot be rotated about a vertical axis and as two components (x-y) of the compensator must be checked, the tests known from the traditional levels (e.g. Two-Peg-Test) cannot be used. The procedure outlined here only gives the principle for the testing and adjustment of the horizontal laser plane. The manufacturers of particular instruments will prescribe the exact procedures to be used.

- Set up the tripod 30 m from a house wall and make the tripod's top surface truly horizontal (in two vertical planes) by using a carpenters level. To check the latter, all readings with the carpenters level should be taken in 'two faces' (and the mean readings used). Extend and retract the adjustable tripod legs to get the tripod top truly horizontal.

- Define two vertical planes through the laser level which are at right angles to each other. Use features of the laser level to distinguish between the x and the y planes.

- Point the positive x-axis to the wall, level the instrument and switch it on. Move the detector up and down until it is centred on the beam. Mark this position, say with X1. Rotate the laser level by 180° (till the negative x axis points to the wall). Do not touch the footscrews in this process. Mark the position of the beam on the wall with X2. For the laser level Wild LNA2, the manufacturer suggests that the x-component of the instrument be adjusted if the distance between X1 and X2 exceeds 2.5 mm (at 30 m). To do so, mark the half-way point between X1 and X2, centre the detector on it and adjust the x-component of the compensator as specified by the manufacturer until the beam is centred on the detector.

- Turn the laser level by 90° and level it by centring the spot bubble. Proceed with the adjustment of the y-component of the compensator as discussed above for the x-component.

3.2.5 Computation of Heights, Earth Curvature and Refraction

It is known from the theory of trigonometric heighting (see Rüeger 1990, pp. 108 to 114, for example) that the (ground) elevation H_T of a target station can be obtained from the (ground) elevation H_{LL} of the laser level station for a line of sight which is strictly horizontal at the laser level station as follows:

$$H_T = H_{LL} + h_{LL} - h_T + \frac{1}{2R} d^2 - \frac{k}{2R} d^2 \tag{3.6}$$

It is important to realise that the above formula gives 'heights above sea level' or 'levelled heights'. This means that water between points of the same elevation would not flow. As in traditional levelling, the laser level is usually not set up over a mark of known height. Instead, the elevation of the laser beam plane (traditionally referred to as 'collimation height') is determined backward, by setting up the laser detector on a mark of known height. In the absence of earth curvature and refraction effects, the collimation height is obtained as

$$\text{Collimation Height} = \text{Elevation of reference mark} + \text{height of detector over reference mark} \tag{3.7}$$

Therefore, the formula for the (ground) elevation H_T at the target point can be simplified to

$$H_T = \text{Collimation Height} - h_T + \frac{1}{2R} d^2 - \frac{k}{2R} d^2 \tag{3.8}$$

where h_T is the height of the detector above the mark (ground) at the target station, R is the earth's radius (e.g. 6370100 m for New South Wales) and k is the coefficient of refraction. The third term on the right hand side is the correction for *earth curvature.* This correction is positive because the spherical earth curves away from the horizontal laser beam. The laser plane is a parallel to a tangential plane to the spherical earth at the laser level station. The forth term on the right hand side corrects for the effects of *vertical refraction.* On clear sunny days close to the ground, the coefficient of refraction k is strongly negative (down to k = –3). This means that the laser plane is no longer a plane but a paraboloid pointing upwards. The offset from the horizontal plane increases with the square of the distance from the laser level. On clear nights, the coefficient of refraction is strongly positive (up to k = +3) , which means that the laser plane is bent downwards, towards the ground. The table below gives an indication of the magnitude of the effects of earth curvature and refraction.

Distance	50 m	100 m	200 m	300 m	400 m	500 m
	mm	mm	mm	mm	mm	mm
Earth Curv Corr	+0.2	+0.8	+3.1	+7.1	+12.6	+19.6
Refr Corr k=–3	+0.6	+2.4	+9.4	+21.2	+37.7	+58.9
Refr Corr k=–2	+0.4	+1.6	+6.3	+14.1	+25.1	+39.2
Refr Corr k=–1	+0.2	+0.8	+3.1	+7.1	+12.6	+19.6
Refr Corr k=0	0.0	0.0	0.0	0.0	0.0	0.0
Refr Corr k=+1	–0.2	–0.8	–3.1	–7.1	–12.6	–19.6
Refr Corr k=+2	–0.4	–1.6	–6.3	–14.1	–25.1	–39.2
Refr Corr k=+3	–0.6	–2.4	–9.4	–21.2	–37.7	–58.9

Table 3.1: Sample corrections of the elevation H_T of the target point for earth curvature and refraction in function of the distance from the rotating laser level. The coefficient of refraction is denoted by k.

Equation (3.8) can be written in a form where the corrections for refraction and earth curvature are applied directly to the height h_T of the laser detector at the target station

$$H_T = \text{Collimation Height} - \{h_T - \frac{1}{2R} d^2 + \frac{k}{2R} d^2\} \qquad (3.9)$$

This form shows that earth curvature causes the h_T readings always to be too large. Vertical refraction causes the h_T readings to be too small when the coefficient of refraction is positive (typically at night) and too large if the coefficient of refraction k is negative (typically during daytime).

On building sites, where the operating radius of laser levels if often below 50 m, the effects of refraction and earth curvature can be ignored. At a radius of 100 m, the corrections are still below the typical system accuracy of laser levels (±4 - 6 mm). At 400 m radius, the corrections do reach and exceed the laser level system accuracy (about ±16 to 24 mm).

For practical reasons, the above corrections are usually ignored even for the large radii used in land forming operations, for example. The effect of omitting these corrections are best explained with the aid of an example. Let us assume that a grader is used to prepared a horizontal surface of 800 by 800 m. In this case, the cutting blade of the grader would be set so that $h_T = h_{LL}$, assuming that the laser level is set up in the middle of the surface. In the absence of refraction, the graded surface would be strictly parallel to the laser plane. This means, in terms of sea level elevations, that the ground at 400 m from the level would be 12.6 mm 'higher' than the ground at the laser level. Water would therefore flow towards the laser level.

If we assume that this earth work was carried out during the night with the laser level set up very close to the ground (k = +3.0), then the ground at 400 m from the laser level could be 25 mm 'lower' than the ground near the laser level. If one were to check the night-time grading later on a clear sunny day, with the laser level again close to the ground (k = -3.0), then one might note that the ground at 400 m is 75.4 mm (2x37.7 mm) 'lower' than expected, that is 75.4 mm lower than the ground would have been if graded on a sunny day. (If the grading were carried out on a sunny day with h_T set equal to h_{LL}, then the ground at 400 m from the laser level would be 50 mm too high.)

The fact that the laser beam may wander 100 mm up and down between day and night over 500 m due to refraction must be considered if some work is carried out during the day and some during the night. *Refraction effects can be reduced by setting the rotating laser level as high as possible above ground and by measuring during windy, overcast days or around sunrise and sunset, when the coefficient of refraction is close to zero.*

3.3 Exercises

3.1 Name and discuss three key differences between digital (bar code) and conventional automatic levelling instruments.

3.2 Like conventional automatic levels, digital levels use compensators to make the line of sight horizontal even though the vertical axis might be out of plumb by a small amount.
Name three errors caused by the compensator and how they may be 'eliminated' in precise levelling.

3.3 The figure below highlights the effect of the contact point between invar staff and bench mark not being exactly in the axis of the invar ribbon. The actual staff reading L' does not correspond to the vertical separation L between the line of sight and the top of the bench mark.

(a) Derive an equation for the staff reading error δh as a function of L, e and the staff tilt α noting that $L' = L + \delta h$.

(b) What is the error δh if $L = 2.600$ m, $\alpha = 10'$ and $e = 32$ mm?

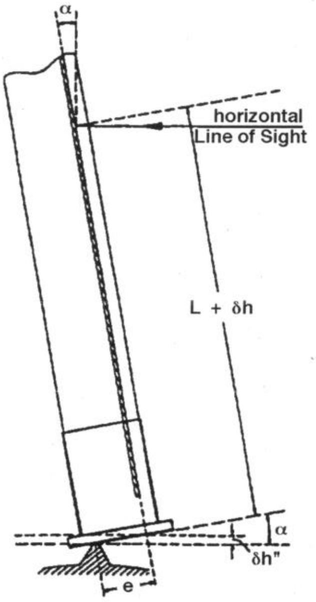

Figure 3.17: Error δh in a reading ($L' = L + \delta h$) on an invar staff due to the tilt of the staff (α) and the eccentricity (e) between the contact point (on the bench mark) and the axis of the invar ribbon. (after Schauerte 1995)

3.4 You have been asked to determine the heights of three bench marks (BMs), namely BM A at the foot of a hill on the eastern side, BM B on top of the hill and BM C at the foot of the hill on the western side. On a clear summer day, you then level from A to B, from B to C and from C back to A around the base of the hill. The misclosure in the level loop is consistent with the accuracy of the equipment and the techniques used. However, you are concerned about possible refraction effect.

(a) Explain why refraction can affect the height of BM B even though the loop misclosure was found to be effectively zero.

(b) By how much could the height of BM B be in error if BM B is 500 m higher than the BMs A and C?

3.5 You instruct a survey party to level a bench mark at the bottom of an open cut mine relative to a bench mark (of given height) on the rim of the mine. The field party uses a digital level with an aluminium bar code staff on a day with an ambient temperature of +40°C. The height difference of about 500 m is measured forward and backwards. The height of the BM at the bottom of the pit is obtained as H = 23.148 m.

(a) Six months later, the level run is repeated with the same equipment, but at an ambient temperature of +10°C. Assuming that both bench marks were not subjected to ground movements or settlements, what elevation would you expect to find for the bench mark at the bottom of the pit?

(b) Assuming that the staffs are at nominal length at +20°C, what would the most likely height of the bench mark at the bottom of the pit be?

(c) The mine management wants you to determine all BM heights to ±10 mm or better. What type of staff will you use in future:

(i) on the (natural) plain around the mine pit?

(ii) to level down into the pit?

Explain your selection.

3.6 In the context of the levelling of the Australian continent in the 1960s, the line from Brisbane to Cairns (about 2000 km) was levelled. Large differences were found when connecting to tide gauges along the coast. Figure 3.18 shows the mean sea level heights as determined from a free continental adjustment (solid line) and from unadjusted orthometrically corrected observed heights (non-adjusted).

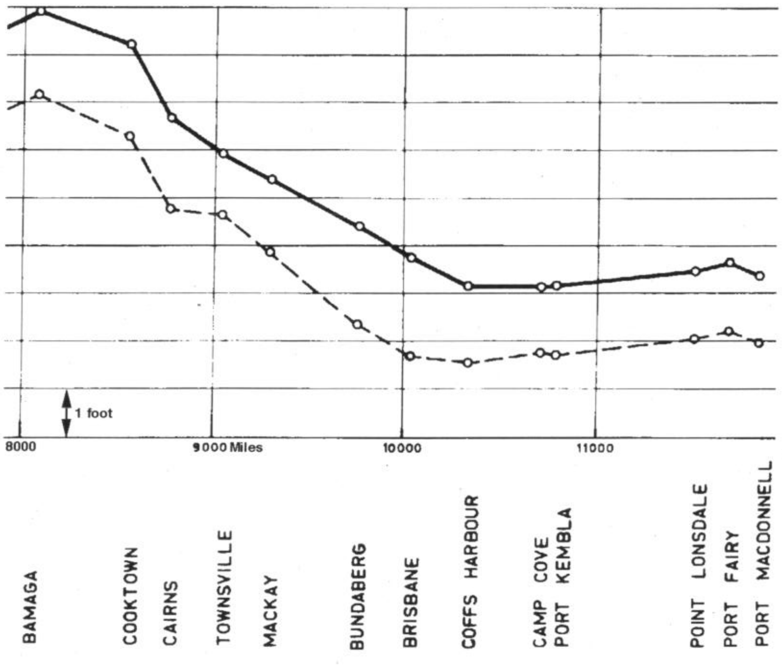

Figure 3.18: Mean sea level heights as determined from a free continental adjustment (solid line) and from unadjusted orthometrically corrected observed heights (non-adjusted) (after Roelse et al. 1975, Annex C)

Explain the likely magnitude of refraction and magnetic field errors, assuming that first generation automatic levels (e.g. Zeiss Ni2) and long sighting distances (90 m) were used.

3.7 Any levelling of importance is done in double runs, that is forward and backward. This provides the only rigorous check in levelling where the over-determination (redundancy) of measurements is, by default, quite poor. Apart from improving the precision and quality assurance, double runs also eliminate some systematic levelling errors.
Explain why the 'sinking of the staff during the move of the instrument to the next set-up' is cancelled (on average) in the mean of the forward and backward height difference.
(Hint: Draw two consecutive instrument set-ups (and three consecutive staff set-ups) on sloping ground and assume some staff readings. Analyse the height differences in the forward and return direction, noting that height differences are computed as backsight BS minus foresight FS and that the staff common to both instrument set-ups sinks by a certain amount during the move on the instrument.)

3.8 Describe a few possible surveying tasks that could be carried out on a building site with a rotating laser level.

3.9 On a building site, you have been asked to check the level of the form work of a reinforced concrete plate. To do so, you set up the rotating laser level on the form work and you adjust the laser detector to the laser plane at Bench Mark A (Reduced Level = 67.045 m), getting a staff reading of 4.613 m. Similarly, at Bench Mark B (Reduced Level = 68.919 m) a staff reading of 2.738 m is obtained.
(a) What is the collimation height of your laser level?
(b) What staff reading do you expect to get for points on the formwork if the bottom edge of the concrete plate is at a reduced level of 70.100 m?

3.10 On the basis of Fig. 3.14, explain why the laser beam travels in an oblique plane in the presence of a compensation error.

3.11 Draw a cross-section through a rotating laser level in the middle of a field of 500 m radius. Using a horizontal scale of 1:10000 and a vertical scale of 1:2, plot:
(a) how the spherical earth deviates from a tangential plane through the laser level,
(b) the curve described by the laser beam in case of a coefficient of refraction of k = +3.0,
(c) the curve described by the laser beam in case of a coefficient of refraction of k = –3.0.

3.12 Figure 3.19 shows two orientations of a rotating laser level, before and after rotation by 180° about a plumbline on a strictly horizontal tripod top.

(a) in both diagrams, draw the ray path through the laser level from the erroneous positions of the laser diode (see empty diode symbols).

(b) Explain how the tilts of the laser beams exiting to the right from the rotating prism differ in the two diagrams.

(c) Discuss how this difference can be used to calibrate the rotating laser

(d) Elaborate on why it is essential that the tripod head is perfectly horizontal (= level) during such a calibration.

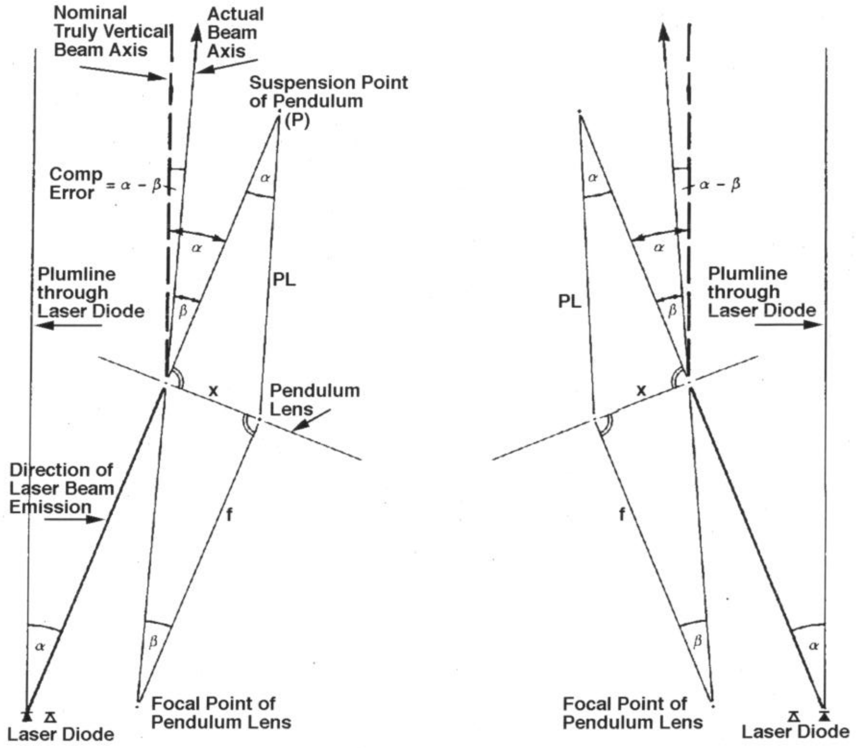
Nominal Truly Vertical Beam Axis
Actual Beam Axis
Suspension Point of Pendulum (P)
Comp Error = α − β
α
β
PL
Plumline through Laser Diode
x
Pendulum Lens
Direction of Laser Beam Emission
f
Laser Diode
Focal Point of Pendulum Lens
α − β
PL
Plumline through Laser Diode
x
f
Focal Point of Pendulum Lens
Laser Diode

Figure 3.19: Two orientations of a rotating laser level, before and after rotation by 180° about the plumbline, on a strictly horizontal tripod top

3.13 You instructed a crew with earth moving equipment to level a field of 400 m radius around a laser level set up. The work was then carried out at night.

Later you check the work by geodetic levelling (with 50 m sighting distances) from the laser set-up to the edge of the field 400 m away.

What height difference, obtained in your levelling run, between the centre and the edge of the (nominally horizontal) field would you tolerate?

Give reasons!

4. Electronic Theodolites and Tacheometers

4.1 Types of Electronic Theodolites and Tacheometers

In the discussion below, the term 'theodolite' is used in a very broad sense. The term is applied to all devices which measure horizontal directions and zenith angles by electronic means.

This chapter deals only with the angular measurement section of electronic theodolites and tacheometers. For a complete discussion of the electronic distance measurement (EDM) aspects of electronic tacheometers, the reader is referred to the EDM textbook by Rüeger (1990 1996). The special aspects of 'reflectorless' EDM are, however, discussed in Sections 4.1.8 and 5.1.

4.1.1 Electronic Theodolites

The electronic theodolites are replacing the optical theodolites which do require a visual reading of the glass circles and do not permit real-time electronic data recording. Electronic theodolites are easier to use as the circles are read electronically and the angles displayed on a read-out screen. All electronic and optical theodolites can measure horizontal directions and zenith angles as well as distances by stadia (fixed-hair vertical staff tacheometry with the aid of the stadia hairs on the reticle).

4.1.2 Electronic Tacheometers

The electronic tacheometers (sometimes referred to as 'total stations') are replacing the older optical tacheometers (optical theodolites specialised for optical distance measurement). The latter permitted optical distance measurement with an accuracy of between ±10 mm (horizontal staff tacheometry) and ±100 mm (vertical staff tacheometry), depending on the principle used. Horizontal staff tacheometry was never widely used in Australia. Electronic tacheometers are electronic theodolites with a built-in electronic distance meter. They measure horizontal directions, zenith angles and slope distances. Using the on-board computer, height differences and horizontal distances can be computed and displayed. The electronic tacheometers can measure distances more easily and more accurately than optical tacheometers. Also, much longer distances can be measured, typically 1 to 2 km to a single prism reflector. Because of the on-board distance meter, most electronic tacheometers do not have stadia hairs on the reticle.

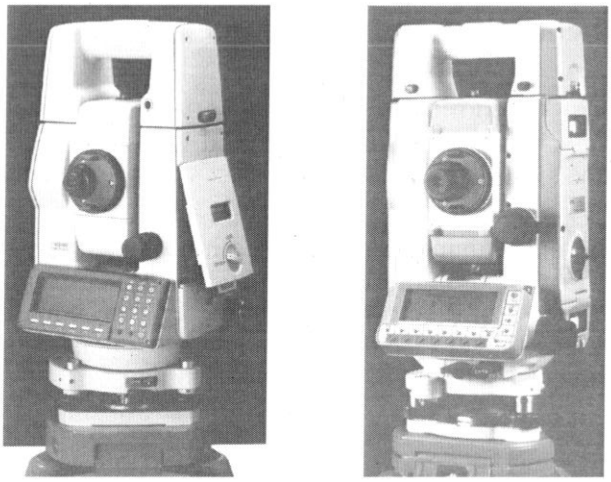

Figure 4.1: Two electronic tacheometers; Topcon GTS-701 on the left and Nikon DTM-720 on the right. Both instruments store data on PC-cards which slide into the side panels.

The first electronic tacheometers (or 'total stations') were the Zeiss Reg ELTA 14 (released in 1970) and the AGA Geodimeter 700 (released in 1971). These instruments were quite large and heavy. At that time, data recording was on paper tape (like in telex machines) with the Reg ELTA 14 and on (audio) cassette tapes with the Geodimeter 700. A second generation of instruments was released around 1978 by Hewlett-Packard, Wild and Zeiss, with the Zeiss Elta 2 being the first instrument with slide-in memory modules (RAM) for data recording. Today, all manufacturers of surveying instruments are producing electronic tacheometers and electronic theodolites. Most instruments sold are, however, electronic tacheometers because they allow the three-dimensional coordination of points from a single instrument set-up. (Naturally, all professionals will survey important points from two different stations, for quality assurance purposes and to improve the precision by taking the mean of the two solutions.)

Electronic tacheometers have become the workhorses of surveyors and geomatic engineers for detail and contour surveys, for control surveys and for setting-out work. EDM-Height Traversing with electronic tacheometers can replace the traditional line levelling with levelling instruments (see Rüeger 1995 and Rüeger & Brunner 1981). On slopes, EDM-Height Traversing is much less affected by refraction than line levelling. Some survey work presently carried out with electronic tacheometers will eventually be replaced by GPS survey techniques, at least where GPS signals can be received. This change is likely to happen as soon as GPS techniques can compete in terms of ease of operation, efficiency, quality assurance and cost.

4.1.3 Gyro-Theodolites

The integral gyrotheodolites and the gyroscope attachments are replacing the earlier (magnetic) compass theodolites and theodolite attachments. Compass theodolites were capable of measuring magnetic bearings to ±1 minute of arc at best. Tubular compass attachments are still available for some instruments to give magnetic bearings and, in fact, may be used for the provisional orientation of gyrotheodolites to true north. The magnetic declination must be considered when using a magnetic compass for this purpose.

After approximate orientation to geographic ('true') north, the north-seeking gyroscope determines the direction to geographic north. Once the horizontal direction to north is established, other points can be sighted to. The clockwise angle between the direction to north and the direction to another point is called azimuth. The azimuth is obtained simply by taking the difference of the directions to the point and to north.

Two types of gyrotheodolites are available. The most precise gyrotheodolites are integrated instruments, such as the Sokkia AGP1 Auto Gyro Station and the DMT Gyromat *2000* by the Deutsche Montan Technologie. Both gyrotheodolites are based on the Gyromat design and give true north to ±6 seconds of arc in about 6 minutes of time. These precision gyrotheodolites are semi-automatic and are used widely underground in tunnels or mines. In long tunnels, the measurement of the azimuths of well distributed lines in traverses greatly improves the lateral precision (across-track) of such underground traverses. For example, the Gyromat 2000 gyrotheodolites were successfully used in the 42 km long tunnel between England and France.

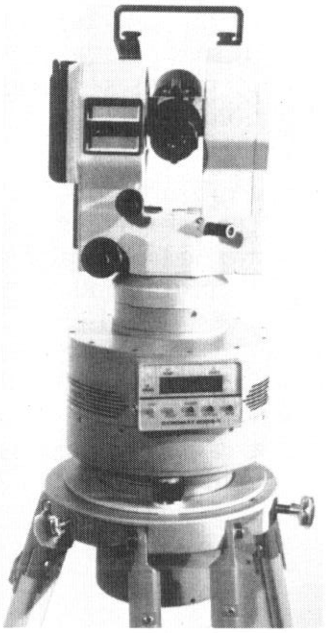

Figure 4.2: Gyromat 2000 with attached Zeiss electronic theodolite. Please note that this particular gyrotheodolite does not fit standard tribrachs or tripods.

Gyro attachments were available for a number of optical and, later, electronic theodolites. These were typically mounted on a bridge between the two standards of the host theodolites. The Wild GAK1 gyro attachment gives true north to ±20 seconds of arc in 20 minutes. A similar accuracy is quoted for the Sokkia GP1-2 Gyro Station (Electronic tacheometer SET 3 with GP1 Gyroscope attachment). The measurements with the gyro attachments are manual and require some skills of the observer.

4.1.4 Motorised Electronic Tacheometers

Motorised (servo-driven) electronic tacheometers and theodolites can be configured and used in many ways. Figure 4.3 gives an indication of the various implementations of servo-technology using the example of the Geodimeter Model 600. This servo-instrument offers different operating modes. 'Conventional Surveying' with a servo-instrument means that the use of the horizontal and vertical motion clamps is a thing of the past; the instrument is turned by the motors under control of the observer through the pointing knobs. Repeated measurements to targets can be programmed so that the instruments turns itself from one target to the next. Only the fine pointing is done by the observer. 'Surveying with autolock' with the Geodimeter 600 requires a tracker unit for the instrument and a special reflector, which assists the tracking. Once the instrument is pointed approximately to the reflector unit, pointing is automatic; no focussing is required and the reflector is followed automatically. Geodimeter claims that the automatic pointing is more accurate than the visual pointing by an observer.

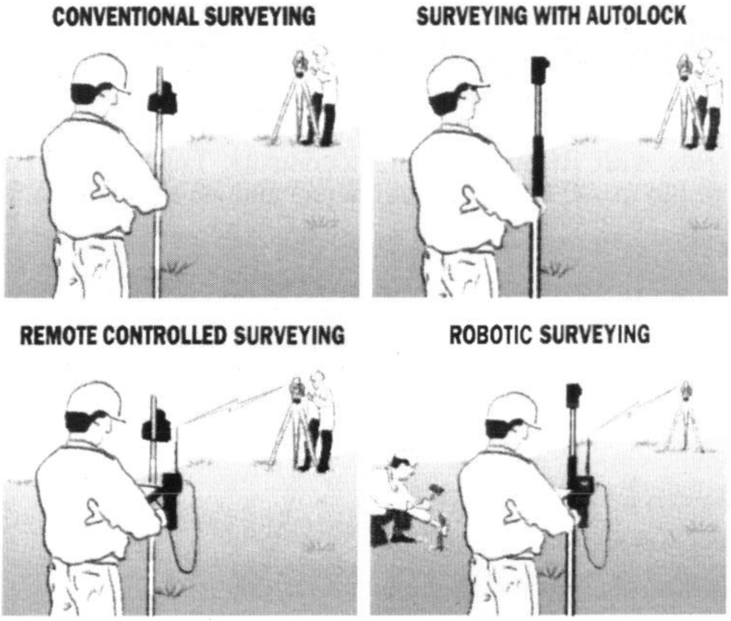

Figure 4.3: Different options of automation with a motorised electronic tacheometer (after Geotronics 1994)

'Remote controlled surveying' means that the instrument is controlled from the reflector and that point coding, recording and calculations are carried out at the reflector. All the instrument operator has to do is to point to the reflector. A radio-link connects the instrument and the reflector in this case. 'Robotic surveying' requires a tracking unit and a communication unit for the instrument and a special reflector, a radio unit and a keyboard for the reflector. The unmanned instrument will automatically locate and follow the reflector. Pointing is automatic. A search routing will pick-up the reflector again should it be lost during tracking. Surveys can be carried out at night, as the instrument operation is fully automatic. (Light is naturally required for the operator at the reflector.)

In 2003, motorised electronic tacheometers are supplied by Geotronics, Leica, Sokkia, Topcon and Trimble (now incorporating Geodimeter and Zeiss) for some or all of the modes of operation discussed above. Two important applications of motorised theodolites and tacheometers are discussed below.

4.1.4.1 One-Person-Survey-Systems

In this case, the operator carries the reflector around and inputs point information; the instrument maintains pointing to the moving prism and executes measurements when instructed through a radio or optical communication link. The one-person-survey-systems can be used for detail and contour surveys as well as for setting out work. The great advantage of these systems is that the expertise of the operator is used at the reflector, that is at the point surveyed or set out. The routine work at the instrument is taken up by the robotic instrument. In 2001, such systems are available from Leica (TPS1100 series with RCS1100 remote control), Topcon (AP-L1A and AP-L1) and Trimble (5600 Series (Geodimeter) and Zeiss Elta S10/S20 'Space').

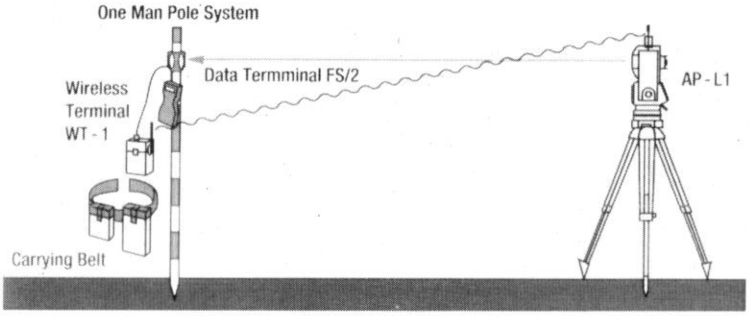

Figure 4.4: Implementation of the Topcon one-person-survey-system (after Topcon 1994)

The basic components of the Topcon one-person-survey-system are depicted in Figure 4.4. The figure shows the instrument, the 360° multiple prism assembly, the reflector pole, the radio communication units at both ends, the data terminal and the carrying belt for the battery packs at the reflector. The Topcon AP-L1 has a horizontal and vertical search facility, which can locate and point to a standard prism around 360°. The prism is found faster if additional information (approximate bearing) is communicated from the reflector operator via the radio link. The AP-L1 telescope axis is shared (coaxially) by the visual telescope, the EDM beam and the tracker. The fine pointing is carried out in steps of 2.5". Moving prisms can be tracked up to angular speeds of 10°/s (63.5 km/h at 100 m). The data terminal allows the remote operation of the instrument and stores the measured data, which are transmitted through the radio link. The usual field computations can also be carried out.

Investigations (Höglund 1994) have shown that one-person-survey systems allow a 30% increase in the number of points surveyed per hour and the survey of 200 points per person hour (against 77 points per person hour for a conventional electronic tacheometers with a two-person field party). When setting out with a two-person field party, the time spent per set-out point is half of that of conventional systems. The higher cost of a one-person-survey-system can be recovered in less than one year because of the saving in staff costs.

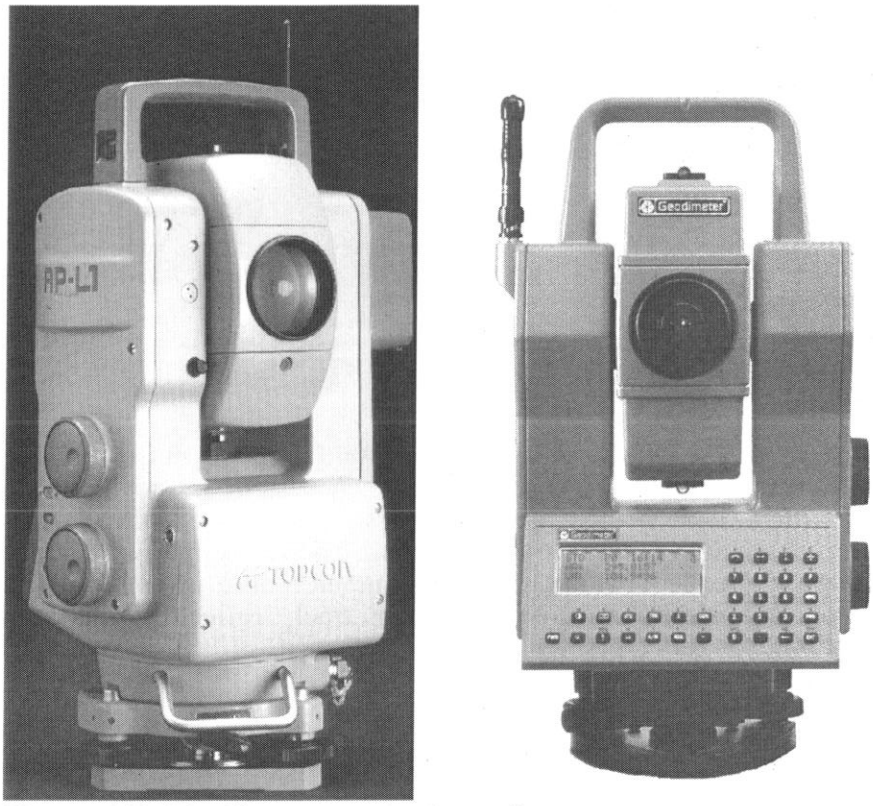

Figure 4.5: Two instrument stations of one-person-survey systems: Topcon AP-L1 on left and Geodimeter 600 on the right.

4.1.4.2 Automatic Monitoring Systems

Fully automatic systems (surveying robots) are being used for the periodic or semi-continuous monitoring of structures (buildings, bridges, walls, tunnels, railway lines), natural slopes, rock faces and mine pit walls. These systems are usually running under the control of an external computer. The latter is also used for data recording and data processing. Typically, the systems are trained by manually pointing to all targets in the desired sequence. This information is then later used to find the points again during the routine measurements. A fine pointing might be executed during the measurements in each epoch, either with the use of a separate tracker, of the EDM beam or with an on-board CCD-camera with associated image processing and pattern recognition software. A search might be launched if a point is not in the field of view after the approximate pointing.

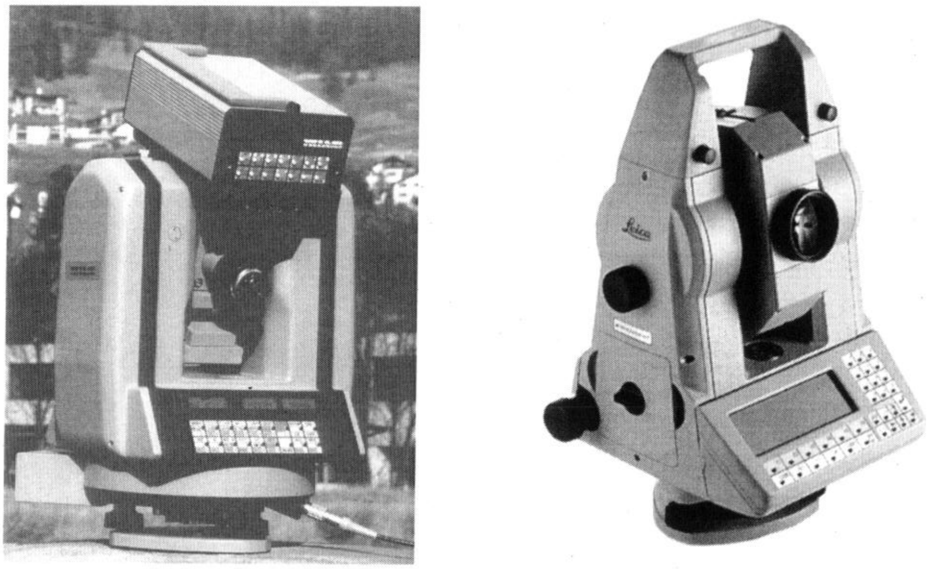

Figure 4.6: Wild TM3000D with mounted Wild DI3000 pulse distance meter on left and Leica TCA1800 on the right. A system as shown on the left was operating at UNSW using Leica's APS software. The TM3000 does not have any slow motion screws and clamps. The telescope and alidade movements are controlled by the external computer or a joy stick. The instrument on the right is motorised, can track reflectors and can automatically point to reflectors.

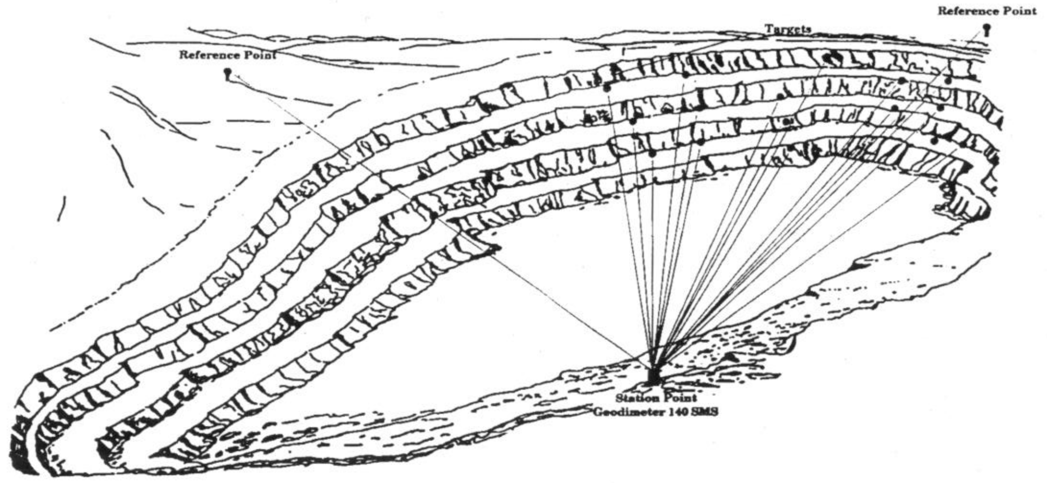

Figure 4.7: Principle of the measurement lay-out in slope monitoring (after Geotronics 1989)

With the Wild TM3000D, distances are measured to mm accuracy and the directions and zenith angles (using EDM pointing) to 2–5" accuracy. The inherent angular accuracy of the TM3000 can only be realised with the use of an on-board video camera. Leica offers a software suite (APS Win) for the different configurations of its **A**utomatic **P**olar measuring **S**ystem (APS). This software supports deformation measurements, machine guidance, surface scanning and general purpose monitoring and can be adapted to other repetitive measuring tasks. The accuracy of the point determination depends on the accuracy of the hardware as well as on the sophistication of the processing software, in particular in the area of refraction corrections and corrections for instrument errors.

All motorised electronic tacheometers that feature automatic pointing and, possibly, automatic target searching, can be used for automatic monitoring tasks, provided that the control software for the field operation is available. In 2003, instruments such as the Sokkia SET4110M, the Topcon GMT-100/100L, the Trimble 5600 (ex Geodimeter?) / 3600 (ex Zeiss?) series of instruments and the Leica TCRA1100 series are available and have been (or could be) programmed for monitoring tasks.

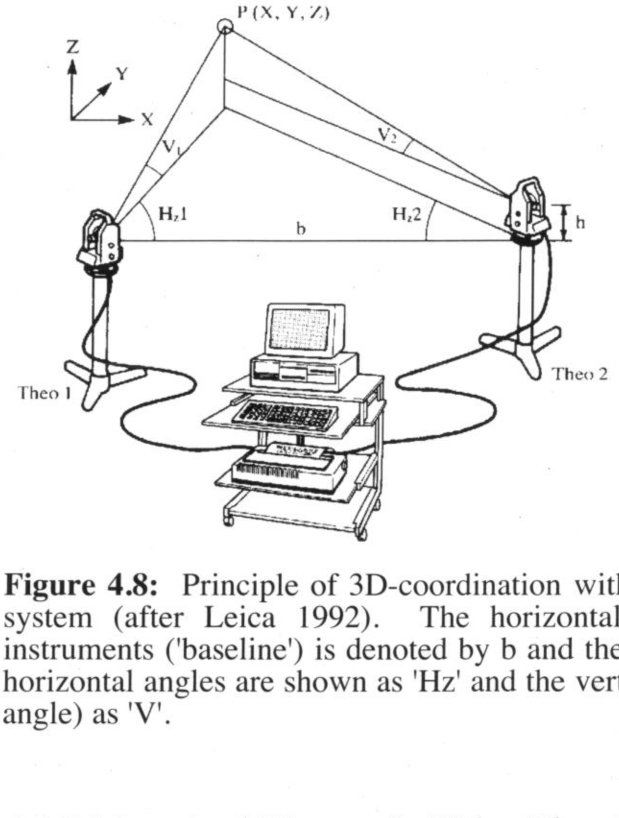

Figure 4.8: Principle of 3D-coordination with a two-theodolite measuring system (after Leica 1992). The horizontal distance between the two instruments ('baseline') is denoted by b and the height difference by h. The horizontal angles are shown as 'Hz' and the vertical angles (90° minus zenith angle) as 'V'.

4.1.5 Motorised Electronic Video Theodolites

In video theodolites, the objective projects the image of the target not only onto the reticle but also onto the image plane of an internal CCD camera. The two ray paths are split by a beam splitter in the telescope. For search purposes, an alternative wide angle image can be switched through to the CCD camera. The CCD array of the Wild TM3000V is 8.8 mm by 6.6 mm with 500 x 582 pixels. The pixel size is 17 μm by 11 μm. The video signal is taken from a special port on the video theodolite and input into a frame grabber on a personal computer for further digital image processing. Object recognition software has to be used to locate a target on the image. Once a target has been identified on the image, the coordinates of the centre of the target can be computed. Knowing (from calibration measurements) the angular offsets of the CCD array's pixels from the centre of the cross hairs, the angular offset of the target from the crosshairs can be found and added to the instrument's readings of the horizontal direction and the zenith angle to give the angular measurements to the target.

This process can be very accurate (e.g. ±1"), so that the full resolution of the instrument is effective. In the 1990s, motorised electronic video theodolites were released by Kern and Wild (now Leica in both cases), namely the Kern E2SE and the Wild TM3000V. With video theodolites, the size, shape and illumination of the targets is very important. Sometimes, reflective targets are used and illuminated with a light on-board of the video theodolite. The Wild TM3000V could be equipped with a light which projects an infrared beam along the telescope axis. On other occasions, a strong laser beam was projected by a laser theodolite on to the object to me measured. The laser dot could then be used as a target by other (video) theodolites. Focussing was not automatic. Once a target was pointed to manually (i.e. in a 'training' session), the focus position was recorded by the computer and then reset (by a servo motor) during subsequent production runs.

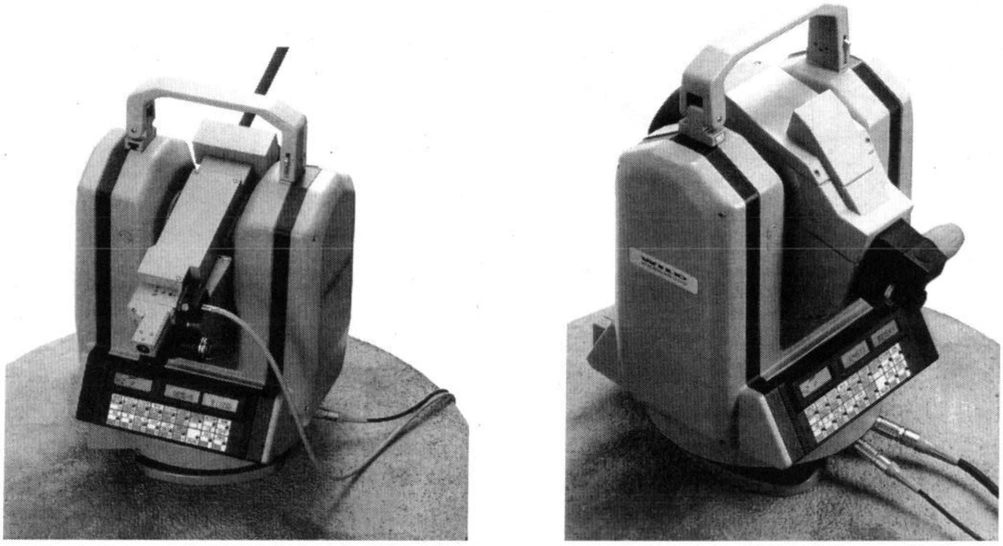

Figure 4.9: Motorised laser precision theodolite Wild TM3000L (on left) and Wild TM3000V motorised precision video theodolite (on right). A special version of the video theodolite (TM3000V/D) permits the attachment of a distance meter.

Figure 4.8 shows a typical application of motorised video theodolites. After initial training, the two (or preferable more) motorised electronic video-theodolites can take repeated measurements to the points of interest on an object. The three-dimensional coordinates are obtained by intersection. The length of the baseline is not directly measured. The positions of the two (or more) theodolites are derived from measurements to marks of given coordinates or from the measurements to the object and to a scale bar of known length. The precision of the resulting object coordinates depends on the distance from the instruments, the intersection angle, the quality of the targets and the stability of the instrument set-up. Angular measurements to one second of arc can give positions to 0.01 mm over 2 m, for example.

In 2003, the video theodolites are no longer available. Depending on application, the video theodolites are now replaced by motorised electronic tacheometers (reflective targets required), by scanning reflectorless motorised electronic tacheometers or by dedicated laser scanners. The laser scanners may be able to detect traditional (non-reflecting) survey targets on objects and structures if their reflectance is sufficiently different from the surface around them.

4.1.6 Laser Theodolites

Laser theodolites have been available before the time of electronic theodolites, either as integrated systems or as a laser eyepiece attachment to standard optical theodolites. The same laser attachments can naturally be fitted to electronic theodolites and tacheometers. Laser theodolites project a visible laser beam along its line of sight (He-Ne gas laser or visible laser diode). They can be used for machine guidance, alignment or as a targeting device for non-contact measurements of structures and objects. If a motorised laser theodolite is combined with a motorised video theodolite, the system can be programmed to scan the surface of an object or structure, with the video theodolite tracking the laser dot after an initial pointing to the laser dot.

Motorised laser precision theodolites may use an on-board semiconductor laser diode for industrial applications (close range) or a helium-neon gas laser for the monitoring or measurement of larger structures (for example in civil engineering). The latter is usually attached to the motorised instrument through a glass fibre connection. The instrument shown in Fig. 4.9 is linked to such an external HeNe laser. The laser theodolites may adjust the laser output power to mach the intensity of the laser dot to the brightness of the object surface. In order to generate a laser dot of uniform intensity, the laser beam is sometimes rotated inside specialised instruments. For example, the TM3000L rotates the laser beam at 30 rpm to give a circle of 0.68 mm diameter at 5 m.

At the turn of the century, lasers are being used less and less for machine guidance. In part, they are being replaced by motorised electronic tacheometers that can measure to reflectors mounted in critical positions on machines. This allows guidance along curves as well as straight lines. If more than one reflector is mounted on the machine, the motorised electronic tacheometer can provide information on the heading (azimuth), pitch (along-track tilt) and roll (across-track tilt) of the machine, besides horizontal and vertical position. In open air, lasers are also replaced by guidance systems based on the Global Positioning System (GPS).

4.1.7 Tracking Electronic Tacheometers

Before the appearance of the one-person-survey-systems, a number of manufacturers did develop specialised tracking electronic tacheometers ('Autotrackers'). These instruments can follow a moving reflector once an initial manual pointing has been carried out. They measure horizontal directions, zenith angles and slope distance continuously and can transmit these data to the reflector, if required. These instruments are typically used in near-shore hydrographic surveys. In this case, the instrument is set up on a shore station of known coordinates and orientated to give a bearing read-out. Bearing and (horizontal) distance are transmitted to the vessel with the reflector, where the vessel position is computed and merged with the data from the echo sounder or other hydrographic instruments.

Typical instruments were the Geodimeter 140T and the Krupp Atlas 'Polarfix'. A number of other manufacturers offered similar systems. The latter had a range of 10 km, an angular precision of ±(50 mm + 50 ppm) and a distance precision of ±(150 mm + 50 ppm). These systems were also used for the position fixing of other vessels, tracking of buoys, tracking of land vehicles, tracking of helicopters, position fixing of construction vehicles and the dynamic measurement of roads and other linear features of terrain. By the turn of the century, these tracking tasks were taken over by GPS systems or appropriately configured main-stream motorised electronic tacheometers.

4.1.8 Reflectorless Electronic Tacheometers

This section gives an introduction into the historical development of electronic distance measurement (EDM) to non-cooperative targets. The particular aspects of reflectorless EDM with current surveying instrumentation are discussed later in Section 5.1. 'Reflectorless' electronic distance measurement (EDM) to passive ('non-cooperative') targets has been used in surveying for quite some time. The first geodetic pulse distance meter became available in 1982: The Fennel 2000 was a stand-alone distance meter, able to measure to non-cooperative targets. Military rangefinders with 1 - 10 m resolution have been around even longer. Zeiss released the first 'reflectorless' electronic tacheometer ('total station') in 1993. Broad acceptance by the surveying profession followed once Leica, Topcon and Trimble released their reflectorless instruments in 1998, followed by Nikon in 1999. The range of reflectorless EDM depends heavily on the type of surface that is measured to and its orientation to the measuring beam. At the time of writing, 50 - 100 m seem to be typical maximum distances to 'natural' targets although Zeiss (now Trimble) quoted a reflectorless range of their RecElta RL electronic tacheometer of 400 m.

The 'reflectorless' instruments can also be used to measure, in 'reflectorless' mode, to the traditional EDM glass prism reflectors. This allows the measurement of distances of a few kilometres, a range that was once reserved for the earlier class of medium range distance meters. Please refer to Section 5.1 for further information on reflectorless EDM.

4.1.9 Scanning Electronic Tacheometers

It has been mentioned before that motorised 'reflectorless' electronic tacheometers can be programmed to measure distances at predetermined values of horizontal directions and zenith angles. This would permit to create a three-dimensional model *of the visible surfaces* surrounding the instrument. As electronic tacheometers are optimised for accuracy rather than speed, special devices have been developed for pure scanning applications. These devices are sometimes advertised as *Laser Radar*, *Ladar* and *Lidar* with reference to the well-known *radar* (RAdio Detection And Ranging). Some of these scanners resemble motorised electronic tacheometers, others resemble the laser scanners used for laser shows in the entertainment industry. In the latter case, the (pulse) distance meter is usually installed with the beam axis straight up. The beam is then reflected by a scanning mirror with two degrees of freedom, with one vertical and one horizontal axis. Angular encoders on the two axes measure the equivalent of horizontal directions and zenith angles.

Three-dimensional scanners can be used for architectural surveys, volumetric measurements, object recognition, process control and collision protection. An early three-dimensional laser scanner (3D-Ladar) by Ibeo had a distance precision of ±20 mm and an angular resolution of 2^{14} in a full circle (80 seconds of arc). The range was up to 500 m and the rate of distance measurements 4600 per second (4600 Hz; Ibeo 1991). Special devices have also been developed for the monitoring of the wear of the lining of converters and ladles in steel making. In this case, distances are measured to the red-hot (1400°C) refractory linings inside the converters and ladles. The system manufactured by VTT Electronics (Oulu, Finland) gives an accuracy of 2 mm over 20 m with a measuring rate of one point per second (Ailisto et al 1996).

By the turn of the century, laser scanners had developed as a special class of instruments for the reflectorless measurement of indoor and outdoor spaces, objects and structures. These instruments can scan a three-dimensional space at high speed and provide 3D coordinates of all objects and surfaces in view. Powerful postprocessing software is then able to convert the data into plans, 3D-representations and computer aided design (CAD) files. This class of specialised instruments is discussed in Section 5.2.

It is mentioned again that any robotic electronic tacheometer, that is motorised and permits reflectorless EDM, can be programmed to scan a three-dimensional space like a dedicated laser scanner. However, the main-stream robotic electronic tacheometers will take much longer for the measurements. Also, it is likely that the range of main-stream robotic electronic tacheometers is smaller and that no powerful postprocessing software is available (to convert the data files to CAD files, for example) for this application of electronic tacheometers.

4.2 Features of Electronic Theodolites and Tacheometers

As far as the telescope and the mechanical construction are concerned, the electronic theodolites differ very little from second order optical theodolites. In consequence, most instrumental errors are the same and most operational procedures remain unchanged. In some respects, however, the electronic theodolites and tacheometers differ significantly from the instruments with optical circles. Some important differences and similarities are now discussed to provide a better understanding of the newer technology and its use.

4.2.1 Circle Types

The horizontal and vertical circles are read electro-optically and the readings are displayed electronically. This means that power is absolutely necessary because the circles cannot be read manually. It is preferable that the essential measurements are displayed simultaneously, for example the horizontal direction, the zenith angle and the slope distance (H, V, SD). The circle graduations used in electronic theodolites are etched onto glass disks using photolithographic techniques. They have approximately the same diameter as in optical theodolites. The Topcon ET-1 features a circle diameter of 80 mm, for example. In other disciplines, electro-optical angle measuring devices are referred to as optical encoders. Three basic types of electronic circles ('angle encoders') can be distinguished.

4.2.1.1 Incremental Circles

Pure incremental circles feature only a uniform radial line pattern. An electro-optical incremental circle is shown in Figure 4.11 (on the right). The pattern is a series of equidistant light and dark fields. The individual graduations cannot be identified as they have no 'labels'. The changes in the horizontal direction and the zenith angle are obtained by counting the number of light and dark fields passing through an electro-optical read-out during the rotation of the theodolite and telescope, respectively, and by multiplying the number of cycles with the angular equivalent of one graduation. The instrument resets the count to zero after switch-on. So, in principle, the zero of the circle (circle orientation) is lost at turn-off or in the case of a loss of power. (Some modern instruments store the last horizontal circle reading into non-volatile RAM in the on-board computer and reset this value after switch-off. If the instrument is rotated in any way whilst turned off, the horizontal circle will have a wrong orientation after re-powering the instrument.) It should be noted that most modern instruments do not have pure incremental circles. The reading of the electro-optical incremental circles is discussed in more detail later, in another section.

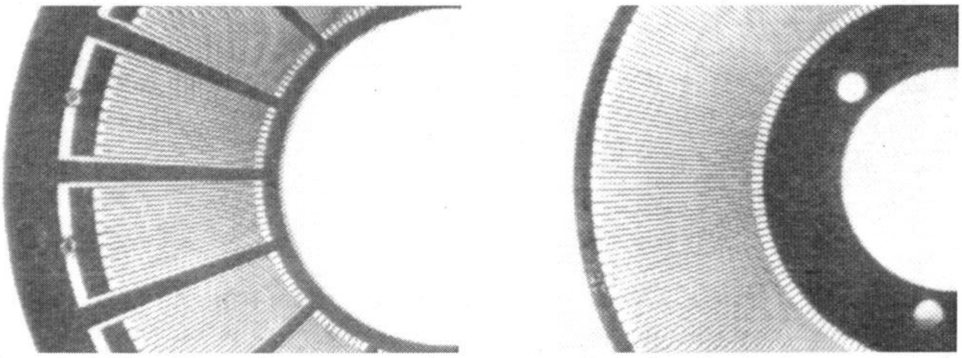

Figure 4.10: Part of a rotor and a stator of a circular Inductosyn (after Inductosyn Corp. and Walcher 1985)

Incremental circles do not have to be of the electro-optical type. Some Geodimeter (now Trimble) instruments are equipped with an inductive circle reading system that operates on the Inductosyn principle. Two parallel disks (rotor and stator) carry equivalent conducting square wave patterns. (Refer to Fig. 4.10.) The square wave patterns are applied to glass disks using the same principles as in the manufacture of printed circuits. One of the disks is fixed. The other one moves with the alidade (horizontal circle) or the telescope (vertical circle). When turning one disk against its opposite, the phase of the measuring signal changes. This phase change is averaged over the entire square wave pattern around the circle and is, thus, free of any circle eccentricity errors and circle graduation errors. Inductive circles reading systems can be as accurate as 2" and are less susceptible to contamination by dust than optical circles. On the other hand, they are more susceptible to electro-magnetic interference. The manufacturing tolerances are quite tight. Measuring the phase to one part in one thousand gives a resolution of 1.3 seconds of arc. Inductosyn systems are widely used for radar antennas, in engineering-works, machine-tools and coordinate measuring machines.

4.2.1.2 Quasi-Absolute Circles

In this context, quasi-absolute circles are defined as incremental circles with one additional reference mark. The reference mark (see Figure 4.11, at top, on the right) allows to start the cycle count at an absolute reference. This means that the circle readings before a power loss can be recovered after recommencement of the measurements (unless the instrument was instructed to set a specific circle reading in a particular direction before the power loss). To initialise the vertical circle, the telescope is typically tilted through the horizon (zenith angle of 90°); to initialise the horizontal circle, the user has to rotate the alidade through (a maximum of) 360°. It should be noted that all vertical circles must be at least quasi-absolute as the zenith angles could not be referenced to the zenith otherwise.

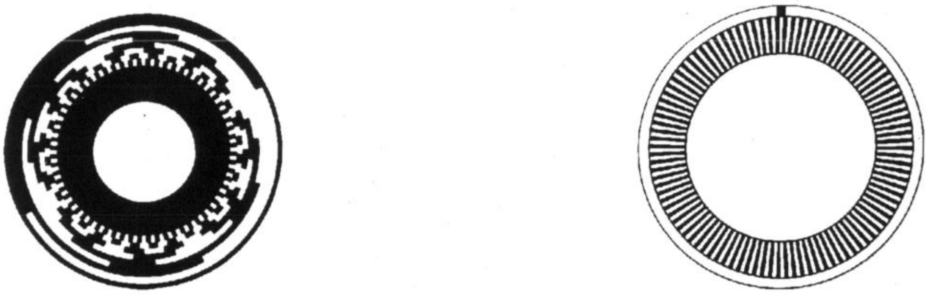

Figure 4.11: Basic layout of an absolute circle (on the left) and an incremental circle (on the right), after BEI (1987). The incremental circle also carries a reference mark (at the top). The innermost graduation of the absolute circle is equidistant, like the graduation of the incremental circle. Each graduation on the absolute circle can be identified by the unique binary code on the seven outer tracks.

4.2.1.3 Absolute Circles

On absolute circles, each circle graduation can be uniquely identified. Figure 4.11 (left) shows an absolute circle, with the incremental (equidistant) graduation on the innermost track. The outer tracks carry a binary pattern, which fully identifies each of the basic graduations. In Figure 4.11, seven code tracks are visible. This code has a resolution of 128 and, thus, fully identifies the 64 graduations on the innermost track. It is obvious that absolute circles with binary patterns must have a readout system for each track. This is expensive in manufacturing.

Not all electronic theodolites and tacheometers with absolute circles have a binary pattern for the identification of the incremental high resolution graduation. The Wild T2000/T3000 theodolites from Leica use a rotating quasi-absolute circle, which is read against the fixed part and the rotating part of the instrument. This gives the (absolute) angle between the fixed readout and the moving readout. Since the Wild TC1600 family of instruments, Leica uses a bar code pattern on the circle which uniquely identifies each position. (See Figure 4.12.) The processing (image matching) is similar to that of the Leica bar code levels.

Instruments with absolute circles do not 'lose' the circle readings in case of a power loss, be it due to a change of battery, the automatic power-off facility of modern instruments or an accidental switch off.

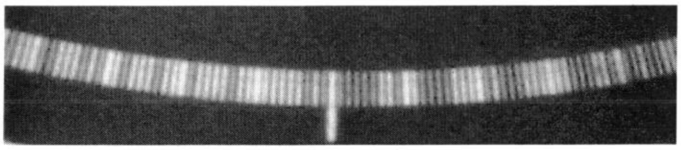

Figure 4.12: Part of the bar code circle of a Wild TC1600.

4.2.2 Diametrical Circle Reading

Circle eccentricities are a very important source of error in angular measurements. In the case of horizontal direction measurement, they can be eliminated by measuring in two faces. In zenith angle measurement, observation in FL and FR do not eliminate the circle eccentricity completely. A number of techniques can be used to eliminate the circle eccentricities in single face measurements with more accurate theodolites and tacheometers. (The coincidence reading micrometer was used in optical (one-second) theodolites to read the circle simultaneously at two points 180° apart and, thus, to eliminate the circle eccentricity by reading the mean). The elimination of circle eccentricities through appropriate instrument design is more important for zenith angle measurement, because the user cannot eliminate the effect by measurement in two faces.

For more accurate work (like the surveys previously done by optical 'one-second theodolites'), most surveyors and geomatic engineers use only instruments which eliminate the circle eccentricity one way or another. Several techniques are presently employed to 'eliminate' the effects of circle eccentricity on angular measurements:

- Some instruments average the circle reading all around the circle. This effectively eliminates the eccentricity error, as the sum of the error around the circle is zero. This averaging approach is used by Geodimeter with the Inductosyn circle reading and by Leica with the rotating circles in the T2000/T3000 instruments.

- Most accurate electronic theodolites and tacheometers employ the traditional diametrical circle reading, where the circles are read separately and simultaneously on opposite sides of the circle. The two readings are then merged electronically.
- Some other instruments read the circles only on one side, but use a factory calibration of the circle eccentricity to correct all measurements on-line before the readings are displayed and stored. Such instruments were the Wild TC1010/1610/1600 electronic tacheometers, for example. The accuracy of the on-line correction depends on the accuracy of the original calibration as well as the stability of the circle mounting over the life of the instrument.

4.2.3 Electronic Level Sensors versus Vertical Circle Compensators

4.2.3.1 Definition of an Electronic Level Sensor

A tilt sensor measures the dislevelment of the vertical axis in digital form and the associated computer corrects at least the zenith angles accordingly. Instruments with single-axis level sensors (parallel to the telescope) can only correct the zenith angles. Instruments with dual-axis level sensors are better for precise measurements or steep sights because they can correct the zenith angles and the horizontal directions for the two components of the measured dislevelment of the vertical axis. Most electronic theodolites and tacheometers use level sensors. Precise instruments feature dual-axis level sensors. With a dual-axis level sensor, the inclination of the vertical axis can, sometimes, be displayed and used for the levelling of the instrument.

4.2.3.2 Definition of a Vertical Circle Compensator

The optical reading of the vertical circle is deviated physically to compensate for the dislevelment (parallel to the plane of the vertical circle) of the vertical axis. This ensures that the zenith angle readings are truly referenced to the zenith. Some electronic theodolites and tacheometers may use vertical circle compensators. The Wild (Leica) TC1600 was such an instrument.

As defined above, the difference between a compensator and a level sensor is that the mechanism of a compensator optically deviates the vertical circle reading whereas a level sensor measures the dislevelment of the vertical axis and corrects the vertical circle reading in real-time by the on-board computer. Both approaches ensure that the correct zenith angle is obtained even when the vertical axis is not truly vertical. Before the introduction of vertical circle compensators, altitude bubbles were used to set the vertical circle reading index to horizontal. As this had to be done before every single zenith angle reading, it was less convenient, less accurate and quite time consuming.

Two types of compensators were used in optical instruments. One used a mechanical pendulum to define the vertical direction and the other the surface of a suitable liquid to define a horizontal plane. Vertical circle compensators are different from those in levels in the sense that they deviate only the ray path of the vertical circle readings and not the entire telescope image. The vertical circle compensators were (almost) always uni-directional. The Kern (now Leica) DKM-2AM did have a dual-axis compensator, but the trunnion axis component of the inclination of the vertical axis had to be read-out with an additional device (trunnion axis micrometer) and the corrections to the direction measurements applied by hand calculation.

As compensators are being phased out in electronic theodolites and tacheometers (and replaced by independent level sensors) they are not discussed any further in this context. Level sensors are discussed later in more detail. Instruments without level sensors (or compensators) are not up to the standard required by geomatic engineers and surveyors as they do not relate the zenith angles directly to the plumbline. In this case, the zenith angles refer to the mechanical vertical axis, the verticality of which depends entirely on the accuracy of the levelling of the instrument.

4.2.4 Data Recording

Electronic data recording saves a lot of time by replacing the manual booking (and the 'booker' assisting the observer). Data recording also avoids the well known gross errors experienced in manual booking, at least as far as the actual electronic measurements (horizontal direction, slope distance, zenith angle) are concerned. Additional input (e.g. height of instrument, height of reflector, point number, point description) through the keyboard is still subject to the well known problems.

Most electronic theodolites and tacheometers can electronically store data on one or more of the following media:

- on-board RAM
- (computer) industry standard memory (PCMCIA) cards (**P**ersonal **C**omputer **M**emory **C**ard **I**ndustry **A**ssociation), now known as PC-cards
- make specific memory cards and modules
- make specific external data recorders ('electronic field books')
- external notebook or laptop computers

The data stored in one of the four first devices need to be down-loaded to a computer for further processing. Additional cables and/or devices and software are required for this purpose. PC cards may be inserted into laptop and desktop computers, that feature a PC card port, for the transfer of data. Data can also be up-loaded, for example coordinates for setting-out work. It should be noted that some instruments, that accept PC cards, cannot record data on-line onto the cards. In these cases, the data are recorded onto RAM and have to be transferred to the PC card later, for example before moving the instrument or at the end of a day.

Unfortunately, all makes of instruments (or even different types of the same make) have different recording formats. There are also some differences between makes and types on how a full data set can be recorded (e.g. one or more key strokes). The sophistication of entering point numbers and other attributes of surveyed points varies greatly. Alpha-numeric entry is sometimes restricted or not possible.

4.2.5 On-Board Computing

Some on-board computing is available on all modern instruments. Sometimes, software modules are offered as optional items by manufacturers. A number of instruments support the personal computer industry's operating systems (initially DOS, now Windows) and programs written by the user. Some examples of what pre-stored programs can do are:

- working out the mean of FL and FR observations, and executing the error analysis in direction measurement with multiple arcs. To be useful, such programs must be able to cope with the standard procedure of direction measurement, namely the sequence FL_1–FL_2–FL_3– ... –FL_n–FR_n– ... –FR_3–FR_2–FR_1.
- computation of the co-ordinates and the elevation of the instrument station using resection (free stationing). The horizontal positioning should work with directions only as well as with directions and distances and should allow and use redundant data. The precision of the determined coordinates should be displayed. The determination of the elevation should be possible with zenith angles only or with zenith angles and slope distances and should allow and use redundant data. The precision of the determined elevation should be displayed.
- computation of bearings, for example for the purpose of orientating the instrument to 'north'.

The many on-board programs, which use the electronic distance measurement (EDM) part of the instruments, are not listed here. Some additional computations are usually available for the guided instrument calibration routines. More on these below.

4.2.6 Power Supply

All electronic theodolites and tacheometers need power to operate. Depending on the make and the type of instrument, direct current is supplied by one or more of the following sources:

- internal rechargeable (nickel-cadmium, NiCd, or nickel-metal-hydride, NiMH) batteries. These are make and type specific. The voltage varies between makes and, sometimes, different types of the same make. Common voltages are: 6 V, 7.8 V, 8.4 V, 12 V. Some surveying instrument manufacturers are now using widely available video camera ('camcorder') batteries as standard power sources for their instruments.
- external rechargeable NiCd batteries. These are also make and, sometimes, type specific.
- lead-acid 12 V car or motor bike batteries. These can be attached to electronic tacheometers and theodolites with the conversion cables supplied by the manufacturers. These make and, sometimes, type specific 12V DC battery cables may include a dc-dc converter to reduce the voltage to the level required by the instrument. As it is easy to connect the 12 V batteries the 'wrong-way-round', the 12 V battery cables should always include reverse polarity protection.

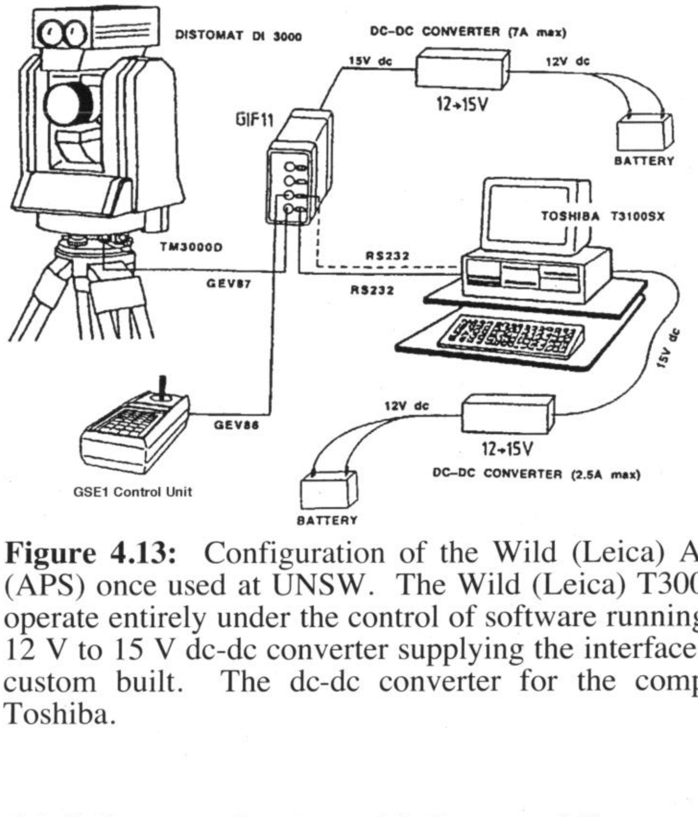
DISTOMAT DI 3000
DC-DC CONVERTER (7A max)
15V dc
12V dc
12→15V
GIF11
BATTERY
TOSHIBA T3100SX
TM3000D
RS232
RS232
GEV87
15V dc
12V dc
GEV86
12→15V
DC-DC CONVERTER (2.5A max)
GSE1 Control Unit
BATTERY

Figure 4.13: Configuration of the Wild (Leica) Automatic Polar System (APS) once used at UNSW. The Wild (Leica) T3000D/DI3000 system did operate entirely under the control of software running on the computer. The 12 V to 15 V dc-dc converter supplying the interface unit (Wild GIF11) was custom built. The dc-dc converter for the computer was supplied by Toshiba.

4.2.7 Communication with Personal Computers

A two-way communication between an instrument and a computer is often possible via a RS-232C interface. Some newer instruments support DOS commands, at least for the import and the export of files. If the instrument's command language is known, the instrument can be instructed to carry out functions which are usually triggered by buttons on the keyboard.

An example for the implementation of a communication link between an electronic tacheometer and a computer is shown in Figure 4.13. The figure shows the connections between a motorised theodolite (Wild TM3000D) with mounted pulse distance meter (Wild DI3000), a computer and a joystick unit (Wild GSE1). In this case, an interface unit (Wild GIF11) is placed between the instrument and the computer. The cable between the interface and the instrument carries not only the supply current but also the two-way communication. As the figure shows, the instrument can be run either from the computer or from the control unit (GSE1). The selection between the two modes is by either a switch on the interface (GIF11) or by software selection. (The latter requires the second RS232 link between the interface and the computer.) A 'joystick', or the arrow keys on the computer keyboard, are required for any manual pointing since the instrument does not feature slow motion screws. All data are recorded on the hard disk of the laptop computer.

4.2.8 Automatic Focussing, Searching and Pointing

4.2.8.1 Automatic Focussing

The first instruments with an automatic focus were the video theodolites. These instruments could 'remember' the position of the focussing lens only after training. Today, auto-focus as in still and video cameras is available in a few electronic theodolites and tacheometers, such as the Topçon 600AF Series and the Pentax R-100 and R-300 Series. The first Auto Focus Levels (Pentax AFL240/280/320) were released earlier in 1996. According to Pentax, the phase contrast method is used by their autofocus system.

4.2.8.2 Automatic Searching

Before an automatic pointing in the narrow sense (see the following section) can be executed by a robotic electronic tacheometer, the glass prism reflector has to be 'found' first and, typically, brought into the field of view of the telescope. Robotic electronic tacheometers, that can find a prism in a 360° sector around the instrument, are available from a number of manufacturers, such as Geodimeter (ATS, 600 Series, now Trimble), Leica (TPS1100 Series with Powersearch, Topcon (GPT8000A, GTS-800A, APL1A), Trimble (5600), Zeiss (Elta S10 'Space', now Trimble). Most instruments transmit a special beam for the fast acquisition of a (reflecting) target and rotate with a maximum speed of, typically, 45°/s. Typical acquisition times are less that 15 s over a 360° sector. Most systems allow to define smaller search sectors to speed up the search, where possible. Naturally, the search times can also be minimised by continuously following (tracking) of a (usually 360°) reflector.

Some systems require an 'active' reflector assembly, that emits a structured infrared light. Others do not absolutely need an 'active' reflector assembly, but achieve the target acquisition faster with one. The use of an active reflector assembly has the advantage that the robotic instrument does not accidentally lock onto the wrong reflector nor, for that matter, onto another 'foreign' reflecting surface in the landscape. The drawback is that batteries are required at the reflector end.

4.2.8.3 Automatic Pointing

Various forms of automatic pointing were implemented in earlier motorised electronic tacheometers, such as:

- pointing under computer control (either on-board or external) to previously established (and 'trained') points,
- recognition of the image of a target on an internal CCD camera,
- tracking (with a video theodolite) the laser dot projected by a laser theodolite onto a surface, and
- tracking the infrared beam of the EDM instrument which is returned by a reflector.

With time, manufacturers started to implement pointing systems that operate separately from the distance meter and tracking systems but use the same reflector to return the transmitted beam. Typically, these systems allow to determine the angular offset between the line of sight of a roughly pointed telescope and the reflecting target, and then use the servo motors to 'move' the cross hairs onto the (centre of the) reflector. The coarse pointing may be done manually (through the gun sight) or with an automatic search or tracking system.

Since only Leica has explained the principle of it's 'Automatic Target Recognition' (ATR) system in any detail (Haag et al. 1997), this system is discussed further. The Leica ATR uses a separate (coaxial) infrared beam and a built-in two-dimensional CCD sensor. Since this beam is of fixed focus, no focussing is required. On return from the prism, the ATR beam is split from the EDM beam and the visual image and projected onto the CCD sensor. To speed up the processing, the pixel values are summed along columns and rows. Histograms of the column-by-column values and the row-by-row values will show peaks at the centre coordinates of the returned ATR beam. Checks for multiple prisms in the ATR beam are made and the beam power and the exposure times of the CCD array are optimised for background brightness and distance. Once the coordinates of the return beam on the sensor are known, the offsets from the 'centre' of the CCD sensor can be computed. Using the servo motors, the telescope is then automatically pointed very close to the centre of the prism. This is necessary to allow electronic distance measurements and to permit a visual check by the observer.

Figure 4.14a shows the configuration during a typical measurement. The field of view of the telescope is indicated by the letter 'A', the EDM beam diameter by 'D' and the diameter of the initial ATR beam by 'C'. The ATR beam diameter is about one third of the visual field of view. If no return beam is received on the initial coarse pointing ('C' in Fig. 4.14a), the servo motors are used to drive the ATR beam around the initial pointing. This is indicated by the letter 'E' in Fig. 4.14a. The letter 'B' shows the total field of view covered by ATR, after initial pointing and fine search.

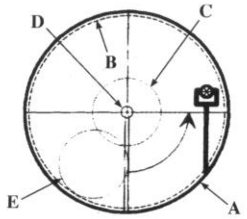

Figure 4.14a: Principle of Leica's Automatic Target Recognition (ATR) of a Leica TCA1800 for measurements to a reflector. The letters are explained in the text. (after Haag et al. 1997)

For a measurement to a reflector, the telescope is pointed with the gun sight. No focussing is required. The observer then starts the measuring routine. The instrument checks if the reflector is within the ATR field of view ('C' in Fig. 4.14a). If not, the ATR beam is driven in a spiral pattern around 'C', as indicated by 'D' and 'E' in Fig. 4.14a. After that, the automatic pointing occurs and the distance is measured. This is followed by the measurement of the direction and the zenith angle, which are then corrected for the offsets of the ATR return beam from the centre of the CCD array and 'frozen' in a buffer memory. Finally, the distance and the 'frozen' angles are recorded. The buffer memory is then cleared.

The accuracy of Leica's automatic pointing (ATR) varies between instruments. The measuring principle discussed above applies to the most accurate Leica instruments. For the Leica TCA2003 and TDA5005 precision instruments, the ATR pointing precision is quoted as 1 mm for distances less than 200 m and 2-3 mm for distances between 200 and 500 m. For the Leica TCA1800, the ATR pointing precision is given as 2 mm to 400 m. The Leica TPS700 and TPS1100 Series of instruments are likely to feature a less sophisticated ATR procedure, since the ATR precision is given as 3 mm below 300 m and 'equivalent to angular accuracy' at longer distances.

All Leica instruments with ATR provide an on-board routing to calibrate the CCD sensor's origin (or 'zero' in Fig. 4.14b) in relation to the centre of the normal crosshairs. This requires careful manual measurements in two faces to a reflector at a specified distance. Figure 4.14b shows the two calibration values that have to be determined. Once determined, the calibration values are applied to all subsequent ATR measurements. Haag et al. (1997) make a few suggestions for precision ATR measurements: Let the instrument acclimatise to ambient temperature, use only reflectors with clean glass surfaces, do a very careful ATR calibration when mixing ATR and manual pointing to targets in the same arc of directions. The latter is best done on site before critical direction measurements and, possibly, between arcs.

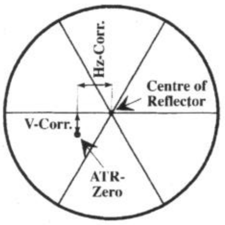

Figure 4.14b: Calibration of CCD array for the offset of its 'centre' or 'zero' from the centre of the (normal) crosshairs. (after Haag et al. 1997)

4.2.9 Computer Assisted Instrument Calibration

Most instruments feature built-in computerised measuring routines for the calibration of some instrumental errors. Such routines may determine and store the following instrument errors or constants (for later real-time correction): horizontal collimation error, index error of vertical circle, additive constant(s) of single-axis (or dual-axis) level sensors, trunnion axis tilt, calibration values of automatic pointing systems (e.g. V-Corr and Hz-Corr in Fig. 4.14b).

These computer assisted instrument calibrations allow some 'mechanical adjustments' to be replaced by 'electronic adjustments'. Rather than physically adjust components into their nominal positions, as in the past, the errors are now determined, then stored on-board (in non-volatile RAM) and corrected in real-time during the measurements.

Similar to the physical adjustments of the older optical instruments never being perfect, the corrections determined by the pre-programmed routines are only as good as the measurements taken. The stored corrections are thus not error free. Users are reminded that some instrument errors do change with time, temperature and shocks. In consequence, instrumental errors must still be removed by appropriate field procedures (e.g. measurement in two faces) if the specified accuracy of an instrument is to be realised, even if the instrument applies the stored corrections in real-time. Also, the users are well advised to check the stored corrections before starting single face measurements. An inexperienced observer may have stored calibration values that are totally wrong!

4.2.10 *Modern Vertical Axis Systems*

The construction of the vertical axes of theodolites has changed considerably over the last 50 years. The left side of Figure 4.14c shows the modern bearing used for the vertical axes of the Zeiss (Oberkochen) theodolites. The weight of the upper part is carried by the load bearing (No. 3 in Fig. 4.14c). The stability of the vertical axis is, however, provided by the second and cylindrical bearing (No. 1 in Fig. 4.14c), which extends over the full height of the bearing. This gives a better stability (less wobble) of the vertical axis than two single ball bearings (one at the top, and one at the bottom) could provide. Figure 4.14c also shows a second cylindrical bearing (No. 2 in the figure) for the rotation of the horizontal circle about the vertical axis.

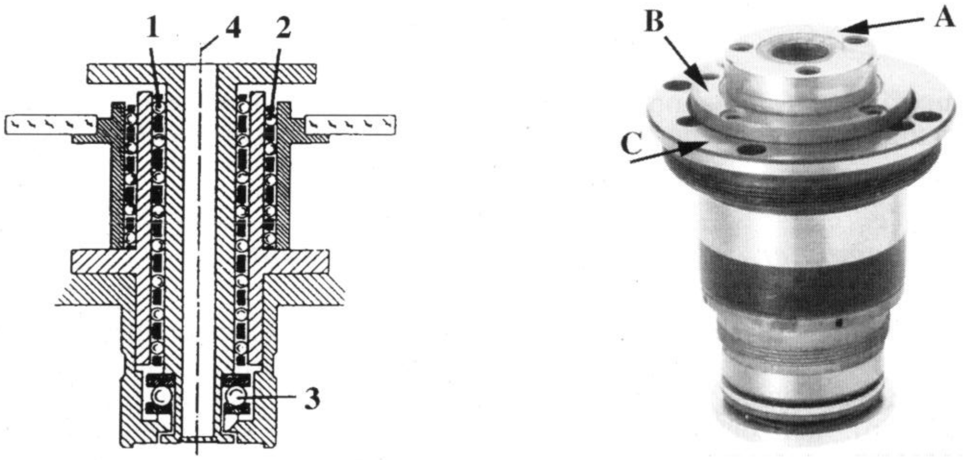

Figure 4.14c: Modern vertical axis systems: Zeiss (Oberkochen) on the left and Topcon ETL-1 on the right. The parts shown on the left are: 1 = ball cage for the cylindrical vertical axis bearing, 2 = ball cage of the bearing of the horizontal circle, 3 = ball bearing carrying the full load of the alidade, 4 = vertical axis. The parts shown on the right are: A = upper plate (on which the rotating upper structure of the theodolite is mounted), B = lower plate (with carries the horizontal circle), C = non-rotating part (attached to the base of the theodolite).

The vertical axis assembly of a Topcon ETL-1 electronic theodolite, shown on the right hand side of Fig. 4.14c, gives a visual impression of a modern design. This instrument was still designed as a 'double-centre', 'double-plate' or 'repetition' theodolite, with two clamps and two slow motion drives for the horizontal motion about the vertical axis. The three layers of holes for mounting screws are testimony to that. The top most level (see 'A' in Fig. 4.14c) carries the upper part of the theodolite, the middle one (see 'B' in Fig. 4.14c) the horizontal circle and the lowest ('C' in Fig. 4.14c) is attached to the (non-rotating) base of the instrument. The upper set of clamp and slow motion drive acts between 'A' and 'B', the lower between 'B' and 'C', allowing the circle to move with the upper part of the theodolite.

Most modern electronic theodolites do no longer have a second bearing, that allows the rotation of the horizontal circle against the base of the theodolite (as explained for the axis systems shown in Fig 4.14c). This means that the horizontal circle cannot be physically turned any more (other than by turning the instrument 120° inside its tribrach) and cannot be set to a given reading (other than by electronic means).

4.3 Circle Reading System of the Electronic Tacheometer Nikon DTM-1

The Nikon DTM-1 uses an electro-optical circle reading. The light and dark increments are produced by the equidistant bars and equidistant gaps. The horizontal circle of the DTM-1 is a fully incremental circle. The vertical circle, by default, is a quasi - absolute circle. The basic principle of the horizontal circle reading is shown in Fig. 4.15. The figure shows that diametrical circle readings are used to eliminate the residual circle eccentricities. The circle of the DTM-1 has a diameter of about 92 mm and carries 10000 graduations. The radius of the circle is 46 mm and the circumference 289.0 mm. Considering the number of graduations over the circumference, the width of the gaps and the bars can be calculated as 14 μm. The radius of an atom is 0.0001 μm and may be used for comparison.

Figure 4.16 outlines the components of the circle reading system of the Nikon DTM-1. Many other instruments operate on very similar principles. On the left of the figure, the two electro-optical systems that read the circle are shown. A light emitting diode (usually infrared) is mounted in the focal point of a lens. The circle is, therefore, illuminated by parallel rays that continue through the glass circle, through the index patterns (masks on glass) onto the photodiodes. According to the mask shown in Fig. 4.15, six separate photodiodes are required to convert the light received in the six windows into electrical currents.

Whenever the readout optics move against the fixed circle (as it happens with the horizontal circle) or whenever the circle rotates through the fixed readout optics (as in the case of the vertical circle), the intensity of the light received by the photodiodes A and B varies. When an opaque field of the mask fully covers a transparent field of the circle, no light is received (= minimum of the sine curve). When a transparent field of the mask fully covers a transparent field of the circle, the maximum of light is received (= maximum of the sine curve). Between the two states, the signal will change, approximately, as a sine curve.

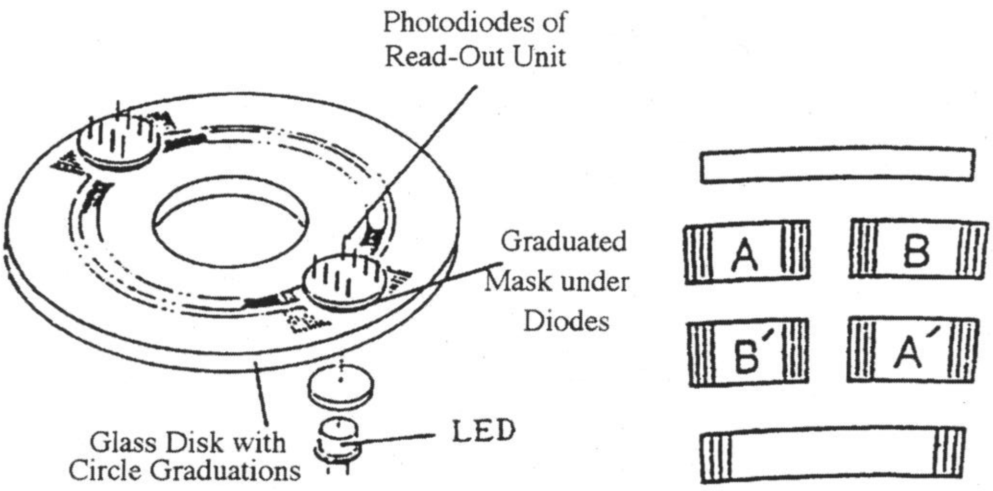

Figure 4.15: On left: View of the circle reading system of the Nikon DTM-1. A light emitting diode and a collimating lens are below the circle. The graduated masks and the photodiodes are above the circle. On right: graduations on the index masks. The top window is blank. This transparent part is used to monitor (and regulate) the light output of the LEDs. A and B refer to the SIN and COS masks, respectively. 'A' and B' duplicate the masks A and B, most likely to reduce systematic errors by a symmetrical lay-out. The bottom window is used on the (quasi-absolute) vertical circle for the detection of the 'zero'- (reference) mark.

The processing of the two signals is shown in Fig. 4.16, namely those of the A (sine) and the B (cosine) masks from both sides. The cosine signal has a 90° delay against the sine signal. The two signals from the other side of the circle feature phase shifts of 180° (-sin) and 270° (-cos). The output signals of the photodiodes are first amplified and then fed into a differential amplifier. The two differential amplifiers take the differences of the two input signals and amplify the result:

$$\sin\phi - \sin(\phi + 180°) = \sin\phi - (-\sin\phi) = 2\sin\phi \qquad (4.1)$$

$$\sin(\phi + 90°) - \sin(\phi + 270°) = \cos\phi - (-\cos\phi) = 2\cos\phi \qquad (4.2)$$

The above equations show that taking the difference of the two input signals already doubles the amplitude. The resulting sine and cosine signals are split for the execution of the coarse and the fine measurement.

4.3.1 Coarse Measurement

The sine and the cosine signals are converted to square waves for the coarse measurement. The coarse measurement is obtained by counting the number of periods (full cycles) which occur between the start and the end of the rotation to be measured. The basic resolution of the coarse measurement is the angular equivalent of the period of the sine and cosine signals, namely 400 grads/10000 = 0.04 grads = 0.04 gon. The coarse measurement of any instrument or telescope rotation is obtained simply by multiplication of the number of counts with 0.0400 grads. The question on how to find out the sense of the rotation (clockwise: positive coarse measurement; anticlockwise: negative coarse measurement) will be addressed later.

The resolution of the coarse measurement (about 130 seconds of arc) is not accurate enough. To get a better resolution, it is necessary to determine the fraction ϕ of a graduation interval between a circle graduation and a graduation on the mask. This process is called interpolation or fine measurement.

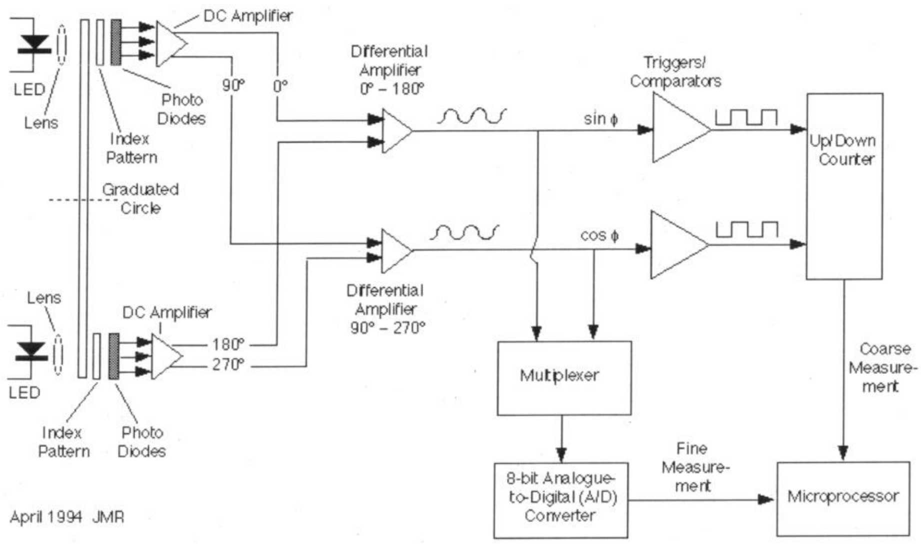
DC Amplifier
LED
Lens
Photo Diodes
Index Pattern
Graduated Circle
90°
0°
Differential Amplifier 0° – 180°
sin ϕ
Triggers/ Comparators
Up/Down Counter
Differential Amplifier 90° – 270°
cos ϕ
Lens
DC Amplifier
180°
270°
LED
Index Pattern
Photo Diodes
Multiplexer
Coarse Measure-ment
Fine Measure-ment
8-bit Analogue-to-Digital (A/D) Converter
Microprocessor
April 1994 JMR

Figure 4.16: Block diagram of the circle reading system of the Nikon DTM-1.

4.3.2 Fine Measurement

The sine and cosine signals are multiplexed for the fine measurement. This means that the sine and the cosine signals are processed one after the other in a time sharing mode. The two signals are available in analogue form as voltage levels, which change for any introduced rotation. For digital processing, the voltage has to be converted to a number. This is achieved with an analogue-to-digital converter. In the DTM-1, the maximum voltage is converted in 256 increments (2^8). The 8–bit analogue-to-digital (A/D) converter changes the amplitudes of the sine and cosine signals to numbers (e_0, e_{90}) between 0 and 256. Assuming that the maximum signal voltage is very close to the maximum input voltage of the converter, the least count of the fine measurement is 1.6^{cc} or 0.00016 grads or 0.00016 gon or 0.16 mgon or 0.52". At each circle position, the phase angle ϕ between the circle and mask graduations must be evaluated. The following relationship is used:

$$\tan \phi = \sin \phi / \cos \phi \qquad (4.3)$$

Using e_0 (the digitised amplitude of the sin ϕ signal) and e_{90} (the digitised amplitude of the cos ϕ signal) and solution for ϕ leads to

$$\phi = \arctan (e_0/e_{90}) \qquad (4.4)$$

The unit of ϕ depends on how the trigonometric function is computed. The circle interpolation ϕ must be converted to grads to be consistent with the previous derivations. Because of the incremental nature of the circles of the Nikon DTM-1, the display of the angular values has to be referenced to the horizontal circle position at switch-on and to the vertical circle reading at the time of the pick up of the reference mark (quasi-absolute circle).

The *displayed direction* value at an alidade position (x) may be derived as Direction (x) = 0.0000 grads (= direction at switch-on) + {number of pulses counted (positive or negative) x 0.04 grad} + ϕ(x) – ϕ(at switch on), where ϕ(x) and ϕ(at switch-on) are the fine measurements (interpolation values) at the arbitrary (x) and 'switch on' alidade positions, respectively. Before the horizontal direction is displayed, it is converted into the unit requested by the user (degrees-minutes-seconds, decimal degrees, mils).

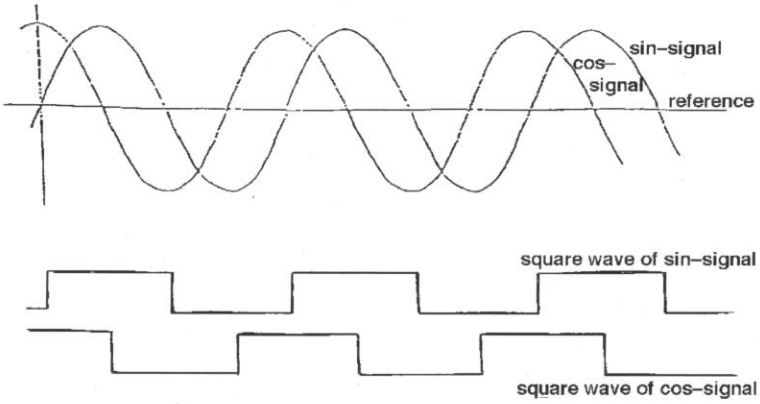

Figure 4.17: Direction Sensing (Up-Down) logic of the Nikon DTM-1. The horizontal axis represents the angle of rotation, positive to the right.

The zenith angle is initialised by tilting the telescope through the horizon. When doing so, the instrument picks up a reference mark on the quasi-absolute circle and 'zeros' the circle reading (setting $z = 100.0000$ grads). The *displayed zenith angle* value at a telescope position (y) can then be derived as $z(y) = z(ref) (= 100.0000 \text{ grads}) + \phi(y) - \phi(ref) + \{\text{number of pulses } (+ \text{ or } -) \text{ from reference mark} \times 0.04 \text{ grad}\}$ where $\phi(y)$ and $\phi(ref)$ are the fine measurements (interpolation values) at the arbitrary (y) and 'reference' telescope positions, respectively. Before the zenith angle is displayed, it is converted into the unit requested by the user (degrees-minutes-seconds, decimal degrees, mils).

4.3.3 Sensing of the Direction of Rotation

The only problem left is find out whether the coarse measurement pulses need to be counted positive or negative by the up/down counter. A logic is required for the detection of the clockwise (+) and the counter clockwise (–) rotation. The basic principle of the direction sensing is shown in Fig. 4.17. As discussed before, the sinusoidal signals are converted to square signals by the two triggers. The rotation sensing logic uses the square signals only. Based on the 'high' and 'low' states of the two waves, it determines if the sine-signal leads the cosine signal or vice versa. Two possible implementations are discussed below. In reality, the sense of the direction can be derived four times over the period of the square signal, namely for each rise and fall of the two signals. This provides redundant data.

In Fig. 4.17, moving to the right on the time axis corresponds to a clockwise rotation of the alidade. The logic makes use of the fact that a zero crossing (rise) of the sine signal is followed by a negative zero crossing (fall) of the cosine signal. In other words, if the sine is high and a fall of the cosine is recorded, we know that the alidade rotates clockwise. The counted pulses are thus positive and must be added by the counter.

On the other hand, a movement to the left in Fig. 4.17 refers to an anticlockwise rotation of the alidade. A positive zero crossing (rise) of the sine signal is followed by a positive zero crossing (rise) of the cosine signal. In other words, if the sine is high and a rise of the cosine signal occurs, the logic concludes that the alidade is turned anticlockwise and that the counted interval must be subtracted.

4.4 Electro-Optical Level Sensors

Level sensors are commonly used in modern theodolites and tacheometers. They provide the actual value of the dislevelment and simplify the circle reading optics. Typically, a level sensor uses a liquid surface to provide a horizontal plane and measures the tilt of the vertical axis opto-electronically. A controlling microprocessor is used to compute the tilt corrections automatically. When enabled, the displayed zenith angle readings (and, for dual-axis level sensors, the horizontal directions) are largely free of the effects of a small residual dislevelment of the vertical axis.

There are two types of level sensors, namely single-axis and dual-axis devices. To permit the real-time correction of the horizontal directions for a small dislevelment of the vertical axis, the instrument must be equipped with a dual-axis level sensor. Instruments with dual-axis level sensors are becoming more common because they are very useful whenever highly accurate directions are required to points with steep lines of sight. It is well known that, at zenith angles of 45°, a 10" levelling error (in trunnion axis direction) causes a 10" error in the horizontal direction. (For even steeper sights, the errors in the directions exceed those in the levelling.) This cannot be eliminated by measuring in two faces but can be removed by an instrument with a dual-axis level sensor and associated software.

Early instruments to use dual-axis level sensors were the Hewlett-Packard HP3820A (1978), Kern E2 (1983), Zeiss Elta 3 (1986), Geodimeter 420/440 (1986) and the Wild T2002/T3000. As dual-axis level sensors are useful whenever steep sights are involved in direction measurements (for example in engineering surveying), most modern electronic tacheometers and theodolites feature such devices.

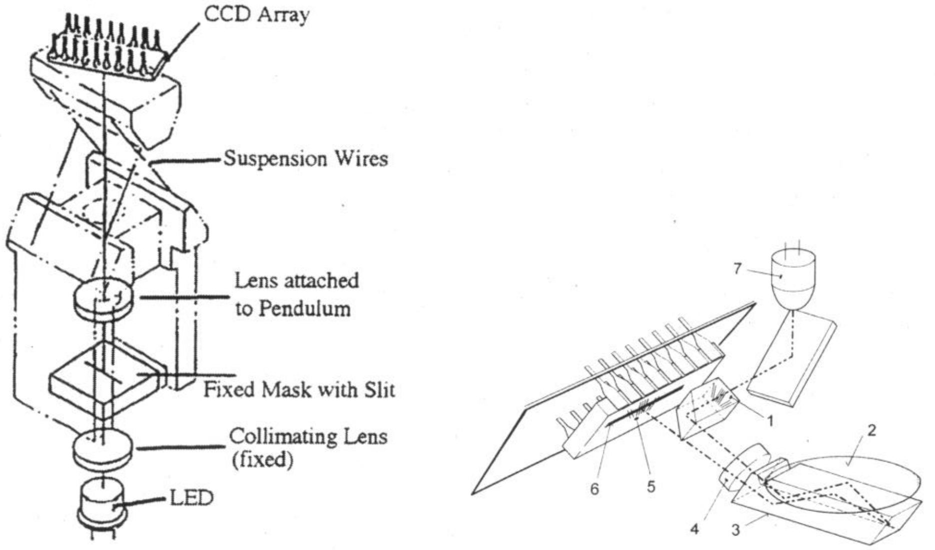

Figure 4.18: Single-axis pendulum level sensor of the Nikon DTM-1 electronic tacheometer on the left and dual-axis liquid level sensor of Leica TPS1100 series on the right. The numbered items are defined in the text. (Figure on the right after Zeiske 1999.)

4.4.1 Nikon DTM-1 Single-Axis Level Sensor

The Nikon DTM-1 electronic tacheometer uses a pendulum–type level sensor. Like most other pendulum sensors, it is a single-axis device. It uses a light emitting diode (LED) to illuminate a transparent slit on a fixed mask. The light bar is projected through a lens onto a CCD array. The lens is attached to a conventional mechanical pendulum system that hangs vertically under the effect of gravity. The position of the light bar on the CCD array is converted into a reading of the tilt of the vertical axis parallel to vertical circle. This value is then used to correct the zenith angle reading into the true zenith. (Refer to Eqs. (4.15) and (4.17).)

4.4.2 Leica TPS1100 Series Dual-Axis Level Sensor

The Leica TPS1100 series of instruments incorporates liquid level dual-axis level sensors. The layout of this level sensor is shown on the right of Fig. 4.18. This particular design was developed to allow to install the liquid horizon in the centre of the instrument, that is in the vertical axis. This has the advantage that the rotations of the upper part of the theodolite exert minimal centrifugal forces on the liquid. In consequence, this level sensor requires less time to settle after a rotation than a level sensor mounted (as usual) in the standards of a theodolite (see Fig. 4.19 for example of the latter design). Also, centrally mounted level sensors are not affected by temperature induced deformations in the standards (Zeiske 1999).

On the right of Fig. 4.18, the ray path through the Leica TPS1100 level sensor is depicted. The prism ('1' in Fig. 4.18) carries a line pattern that is illuminated by a light emitting diode (LED, see '7' in figure). The line pattern is projected through a lens ('4') and a deviation prism ('3') onto the liquid horizon, where it is reflected twice at the surface of the liquid. The image ('5') of the line pattern falls on the linear CCD-array ('6'). Because of the use of V-type lines at both ends of the line pattern (besides the parallel lines in the centre), it is possible to measure both components of the tilt of the vertical axis with one linear ('single-axis') CCD sensor. A longitudinal tilt of the vertical axis (i.e. parallel to the vertical plane defined by the rotating telescope) changes the distance between the V-type lines of the pattern. A traverse tilt (in the vertical plane parallel to the trunnion axis) moves the centre of gravity of the line pattern on the CCD sensor.

Leica quotes the 'setting accuracy' of the level sensors of the TPS1101 and TPS1102 instruments as 0.5". Since no distinction is being made between the two components of the dual-axis level sensor, one may assume that they are equally accurate. For comparison, the angular accuracy of the TPS1001 instruments is quoted as ±1.5" (after ISO 17123-3). It is appropriate that the accuracy specification of the level sensor is only a third of the angular accuracy rating.

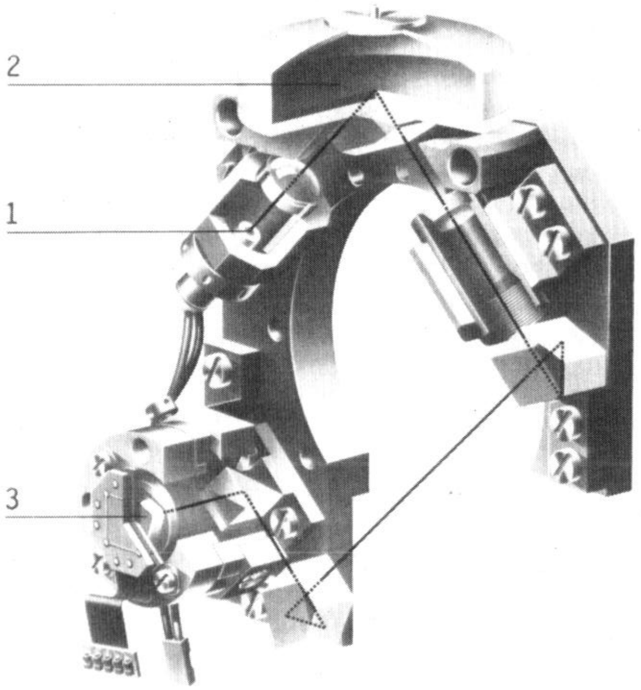

Figure 4.19: Cross section through the dual-axis level sensor of the Kern E2 electronic theodolite (after Lardelli 1988). A light emitting diode (1) emits a light beam that is focussed to infinity by a lens with a short focal length. The parallel rays are reflected downwards at the top surface of the liquid horizon (2). The reflected beam is then focussed onto a position sensitive quadrant photodiode (3) by a second lens with much longer focal length. To save space, the read-out ray path is bent three times.

4.4.3 Kern E2 Dual-Axis Level Sensor

The dual-axis level sensor of the electronic theodolite Kern E2 is similar to the proven liquid compensator of the Kern DKM-2A one-second theodolite. (Some Zeiss electronic theodolites use the same device in a slightly modified form.) Figure 4.19 shows a three-dimensional view of the device and explains the basic principle. Because the liquid horizon is intrinsically two-dimensional, the light beam is reflected in two dimensions, one component being parallel to the plane of the light path (parallel to the vertical circle) and the other perpendicular to it (parallel to the trunnion axis).

The quadrant photodiode measures the x-y coordinates of the light spot created by the reflected rays on the diode relative to the centre of the diode.

The two components of the diode readout are a measure of the two components of the vertical axis tilt, namely the tilt parallel to the vertical plane described by the telescope axis rotating about the trunnion axis and the tilt parallel to the vertical plane through the trunnion axis. The output currents (i_1, i_2, i_3, i_4) of the diode are used to derive the x and y coordinates of the light dot.

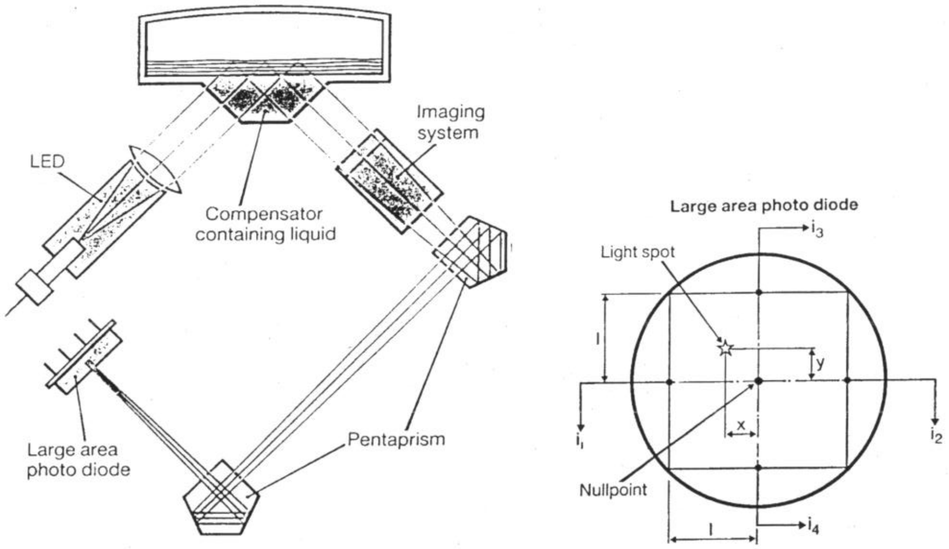

Figure 4.20: Side view (parallel to the plane of the vertical circle) of the Kern E2 dual-axis level sensor (on the left) and the principle of the two-component read-out on the position sensitive quadrant photodiode (on the right). The photodiode has five electrical connections, one in the centre and one each (i_1, i_2, i_3, i_4) for the four quadrants. (after George 1984)

In simple terms, the derivation can be given as follows:

$$i_1\, k\,(l - x) = i_2\, k\,(l + x) \qquad \text{and} \qquad i_3\, k\,(l - y) = i_4\, k\,(l + y) \qquad (4.5)$$

where l is one half of the side length of the photosensitive area of the diode and k a constant. Solving for x and y leads to

$$x = l\left(\frac{i_1 - i_2}{i_1 + i_2}\right) \qquad \text{and} \qquad y = l\left(\frac{i_3 - i_4}{i_3 + i_4}\right) \qquad (4.6)$$

The units of x and y are the same as that of l, namely a length unit. Angular units are obtained by dividing x and y by the focal length of the lens system between the liquid horizon and the photodiode. Any errors in the length of the focal distance used for this purpose will cause *scale errors* in the angular values and, eventually, in the displayed vertical axis tilts. In Fig. 4.20 it is assumed that the optical axis of the lens system intersects the centre of the photodiode. In practice, this is not exactly the case. Small corrections *(the additive constants of the level sensor)* need to be applied to the measured x-y values to reference them to the point where the optical axis intersects the photodiode.

Changing the inclination of the vertical axis (see Fig. 4.20) causes a movement of the light dot on the position sensitive photodiode. This information is used to compute the two-components of the vertical axis tilt. The next section discusses this principle in more detail for a simplified ray path.

4.4.4 Principle of the Determination of the Theodolite's Dislevelment

The unfolded ray path of a liquid level sensor is shown in Figure 4.21 for three different cases. The sensor design is a simplified version of the type shown in Section 4.4.3. In all three figures it is assumed that the optical axis of the lens on the photodiode side intersects the photodiode in the centre. Figure 4.21a depicts the ideal case where the plumbline is coincident with the vertical axis and the symmetry axis of the level sensor. In consequence, the light travels along the optical axes and is projected into the centre of the position sensitive photodiode. The angle between the symmetry axis and the optical axes of both lens systems is denoted by ϕ.

Figure 4.21b gives the same view as Fig. 4.21a, but for a dislevelled theodolite or tacheometer. The instrument is out of plumb by an angle α. The angles between the optical axes of the two lens systems and the symmetry axis of the level sensor (and the inclined vertical axis) are still the same (ϕ). Because of the tilt, the liquid horizon is no longer perpendicular to the vertical axis. This affects the reflection of the rays. The incident angle at the liquid horizon is $(\phi - \alpha)$ as is the emerging angle. The figure shows that the optical axis and the reflected rays are at an angle of 2α to each other. In consequence, the light dot is no longer in the centre of the photodiode but rather $2 \times \alpha \times f$ towards the vertical axis, f being the focal length of the photodiode optics.

The configuration, after a rotation of the alidade by 180° about the vertical axis, is shown in Fig. 4.21c. In the absence of a wobble error in the vertical axis bearing, the direction of the vertical axis will still be inclined at an angle α. The incident angle of the rays on the liquid horizon is now $(\phi + \alpha)$, because the rays are coming from the right. The emerging angle is also $(\phi + \alpha)$, whereas the angle between the optical axis on the photodiode side and the plumbline is now $(\phi - \alpha)$. The figure shows that the optical axis and the reflected rays are at an angle of 2α at the photodiode. The light dot is again not in the centre of the diode but rather at a distance of $2 \times \alpha \times f$ away from the vertical axis. If the coordinates of the light dot on the photodiode are measured in face left and in face right, then the change of the coordinates between the two faces will be equivalent to 4 times the dislevelment of the vertical axis. This rule applies to both components.

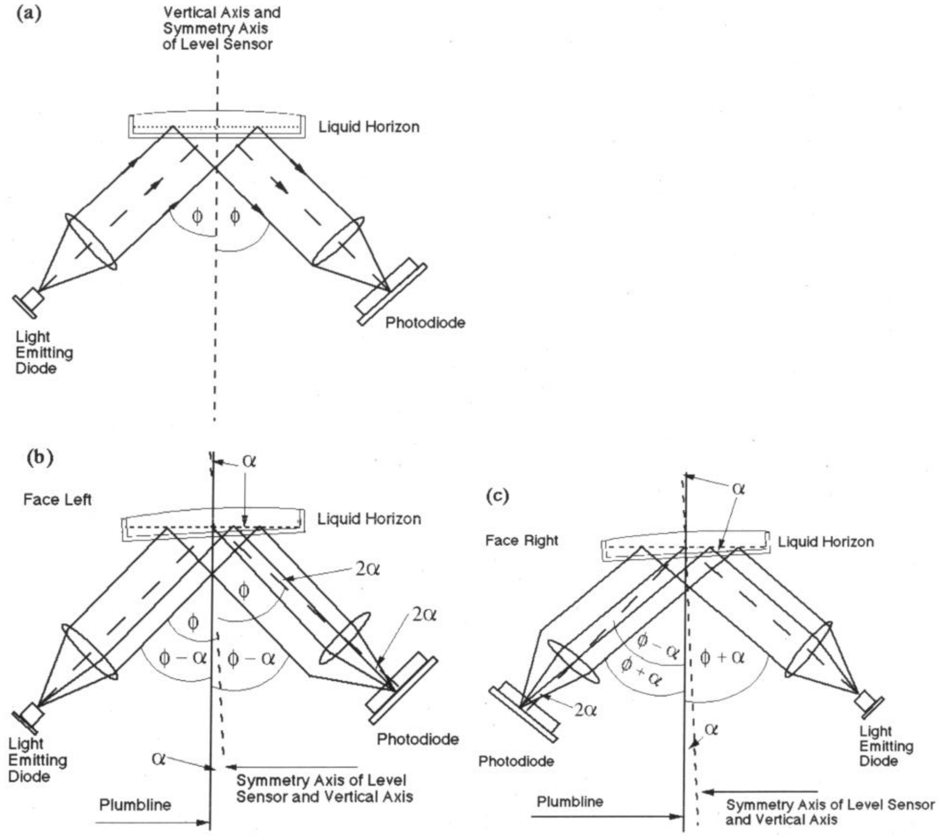
(a)
Vertical Axis and Symmetry Axis of Level Sensor
Liquid Horizon
ϕ
ϕ
Light Emitting Diode
Photodiode
(b)
Face Left
α
Liquid Horizon
2α
ϕ
ϕ
2α
$\phi - \alpha$
$\phi - \alpha$
Light Emitting Diode
α
Photodiode
Plumbline
Symmetry Axis of Level Sensor and Vertical Axis
(c)
Face Right
α
Liquid Horizon
$\phi - \alpha$
$\phi + \alpha$
$\phi + \alpha$
2α
α
Photodiode
Light Emitting Diode
Plumbline
Symmetry Axis of Level Sensor and Vertical Axis

Figure 4.21: Working principle of (one component of) a liquid level sensor

As only coordinates on the photodiodes are recorded, the angle α has to be expressed in terms of these measurements. Let S_{FL} be the coordinate reading on the photo diode in face left, S_{FR} the coordinate reading on the photo diode in face right and S_{REF} the coordinate reading for the optical axis. S_{REF} should be zero if the theodolite is truly vertical and if no index error (additive constant) of the level sensor is present. In face left, the following formula applies in the right angled triangle between the centre of the lens, the centre of the photodiode and the light dot:

$$2\alpha = \arctan\left(\frac{S_{FL} - S_{REF}}{f}\right) \tag{4.7}$$

$$\alpha = (S_{FL} - S_{REF}) \times \frac{206265''}{2\,f} \tag{4.8a}$$

$$\alpha = S_{FL} \times C - S_{REF} \times C \tag{4.8b}$$

where ($S_{FL} - S_{REF}$) is positive, S_{FL}, S_{REF} and the focal length f are taken in mm and where C is the scale factor which converts the photodiode output into the (single) vertical axis tilt α, in seconds of arc. In face right, the following relationship holds:

$$-\alpha = (S_{FR} - S_{REF}) \times \frac{206265''}{2\,f} \tag{4.9}$$

$$-\alpha = S_{FR} \times C - S_{REF} \times C \tag{4.10}$$

Taking the difference between Eqs. (4.8) and (4.10) gives the measured component of the dislevelment α of the vertical axis

$$2\alpha = S_{FL} \times C - S_{REF} \times C - S_{FR} \times C - S_{REF} \times C \tag{4.11}$$

$$\alpha = (S_{FL} \times C - S_{FR} \times C) / 2 \tag{4.12}$$

Adding Eqs. (4.8) and (4.10) gives the additive constant ($S_{REF} \times C$) of the measured component of the level sensor

$$0 = S_{FL} \times C - S_{REF} \times C + S_{FR} \times C - S_{REF} \times C \tag{4.13}$$

$$0 = S_{FL} \times C + S_{FR} \times C - 2S_{REF} \times C$$

$$S_{REF} \times C = (S_{FL} \times C + S_{FR} \times C) / 2 \tag{4.14}$$

It follows that it is very easy to determined the additive constants of level sensors: Read the level sensor in two faces and take the mean. Also, a theodolite or tacheometer is truly vertical, if the level sensor reads the same value ($S_{REF} \times C$) in all orientations of the alidade.

Please note that each coordinate is individually multiplied with C as it is the product of S_i and C that is displayed by instruments.

4.5 Properties of Electronic Level Sensors

The error patterns of mechano-optical vertical circle compensators and compensators in automatic and digital levels are similar. In theodolites, the errors are reduced, because the theodolites can be (and should be) levelled more accurately than levelling instruments. The error patterns of electronic level sensors and vertical circle compensators are also similar because the basic sensor elements are the same, namely mechanical pendulums and liquid horizons.

4.5.1 Compensation Error ε

For the traditional vertical circle compensators, the compensation error ε is defined as the residual inclination of the vertical circle reading system caused by under- or over-compensation of the compensator. It is a function of the dislevelment α of the vertical axis. Figure 4.22 demonstrates the principle of the compensation error of mechano-optical compensators and applies equally to FL and FR observations.

The compensation error is caused by the positioning tolerances (errors) in manufacturing, the inaccuracy of the optical-mechanical design and the effects of changing gravity. The compensation error may also depend on ambient temperature. The magnitude of the compensator error ε in classical instruments with compensators is usually less than 1%-5% of the dislevelment α.

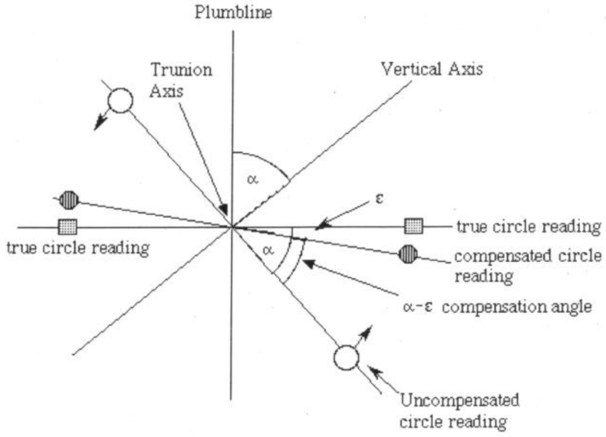
Plumbline
Trunion Axis
Vertical Axis
α
ε
true circle reading
true circle reading
α
compensated circle reading
α-ε compensation angle
Uncompensated circle reading

Figure 4.22: Principle of the compensation error ε in mechanical vertical circle compensators. The dislevelment of vertical axis is denoted by α.

4.5.2 Scale Factor Error ΔC

In electronic level sensors, the scale factor error causes errors that are equivalent to the compensation errors of the vertical circle compensators. As before, the errors in the measured inclination of the vertical axis, as well as in the corrections to zenith angles and directions computed from these tilts, are a function of the dislevelment α. The scale factor C of level sensors has been defined previously (Eqs. (4.8a) and (4.8b)). It is a function of the focal length f of the sensor's optics.

An error ΔC in C may be due to the manufacturing tolerances (positioning of lens, emitting diode and photo diode) or due to the conversion of the electronic read-out of the photodiode output current to length units. The magnitude of the scale factor error in electronic level sensors is typically less than 1%-2% (of α).

The scale factor errors are dangerous (as are the compensation errors), because the measured zenith angles are affected but not the vertical circle index error. This error cannot be detected in the usual way (by checking the sum of FL and FR zenith angle readings). *The scale factor error (as well as the compensation error)* ***cannot*** *be eliminated by measurement in two faces!*

4.5.3 Setting Error

4.5.3.1 Random Component

This reflects the setting accuracy of the level sensor after a disturbance (such as a rotation of the alidade or another instrument manipulation). This repeatability is typically better than 3" – 1". The setting accuracy specified by manufacturers for some instruments is given as a guide:

Sokkia SET 2C/3C	±1"	Leica TPS1100	±0.5"
Wild T2000	±0.3"	Nikon DTM-821	±1.0"
Sokkia SET 5C	±5"	Topcon GTS-603C	1.0"
Zeiss Elta 2/3	±0.6–1.0"	Zeiss Elta S10/S20	±0.3"

It is unfortunate that some manufacturers (e.g. Sokkia, Trimble) do not specify the accuracy of their level sensors. The presumption is that, in these cases, the specifications of the angular accuracy include the level sensor uncertainty.

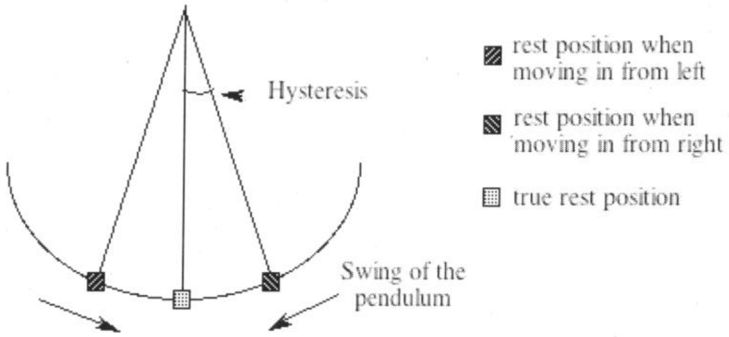
▨ rest position when moving in from left
▧ rest position when moving in from right
▦ true rest position
Hysteresis
Swing of the pendulum

Figure 4.23: Principle of mechanical hysteresis ('compensator drag')

The setting time of the level sensors depends on their design. The mechanical devices (pendulum) need about 1 second of time to reach an equilibrium position. The liquid level devices require longer times, typically

3 to 5 seconds. Even longer times are required in cold weather, because of the change in the viscosity of the liquid. It is likely that level sensors located in the vertical axis (see right of Fig. 4.18) need less time to settle. The observer should wait long enough before reading and/or recording measurements. This is essential for all zenith angle measurements. If dual-axis level sensors are used to correct horizontal direction measurements, then the same waiting period applies for horizontal angle measurements.

4.5.3.2 Systematic Component

Mechanical hysteresis exists if the final rest position of the pendulum depends on the direction from which the pendulum moves into its rest position. See Fig. 4.23. The error caused by the mechanical hysteresis is a function of the inclination of the vertical axis (α) and of ambient temperature. In the last version of the classic Wild T2 theodolite, a push button ensured that the compensator always 'moved in' from the same side. The hysteresis effect will then be similar for all readings and, thus, become part of the vertical circle index error. Some experiments did establish the magnitude of the hysteresis effect as less than 3" - 25" for a selection of optical third order theodolites. *Liquid level compensators and level sensors are not affect by hysteresis.*

4.5.4 Vertical Circle Index Error and Correction

As optical theodolites, electronic theodolites and tacheometers have an index error of the vertical circle reading system and require a vertical circle index correction (VCIC). This error has two components, which might be explained with reference to an instrument, which reads the zenith angle in the horizon through the trunnion axis. Firstly, the line between the 90° and the 270° mark on the circle might not be exactly parallel to the line of sight. Secondly, the reading mechanism might not be exactly in a horizontal plane trough the trunnion axis. The two components cannot be separated; the sum is determined and corrected for. The previously discussed additive constant of the electronic level sensor constitutes an additional (third) component of the total VCIC. If the level sensor correction is enabled and the VCIC is determined in the usual way, the resulting value will include the additive constant of the level sensor (component parallel to telescope axis). If the level sensor is 'switched off' during the determination of the VCIC, then the resulting VCIC value will not include the additive constant of the level sensor. See Eqs. (4.15) and (4.17) for proof.

The (total) vertical circle index error may be affected by ageing (long-term changes), shocks, lateral tilts (cross-tilts), sticking of a mechanical compensator level sensor and temperature. Only the last effect is discussed further.

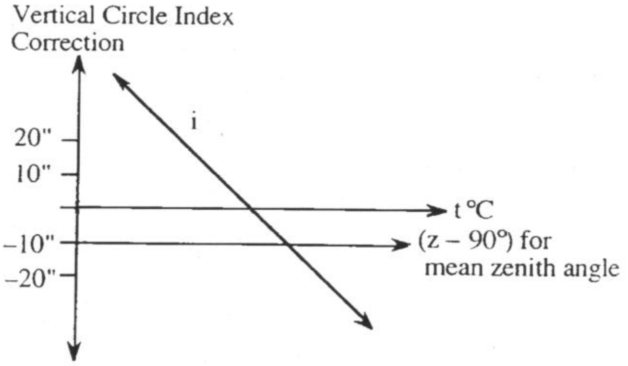

Figure 4.24: Temperature dependence of the index error and temperature independence of the mean (of two faces) zenith angle (z) for a particular theodolite

Experiments have shown that the vertical circle index correction (VCIC) and/or level sensor correction can be strongly dependent on temperature. Some examples of test results may illustrate the magnitude of the problem in electronic tacheometers and level sensors:

KERN E2	+0.30"/°C
WILD T2000	–0.88"/°C
KERN DKM 2A	+0.91"/°C

As the ambient temperature changes as a function of time, the temperature effects caused by the level sensors in the zenith angle measurements can be reduced by appropriate field procedures. The linear changes with time of the index or level sensor correction may be eliminated by measuring zenith angles in the following sequence: $z_{FL} - z_{FR} - z_{FR} - z_{FL}$.

4.6 Circle Graduation Errors

There are two types of circle graduation errors, namely random errors and systematic errors. The systematic errors are usually referred to as periodic errors. All systematic circle graduation errors are caused by imperfections in the devices that produced the master circles and those used for the transfer to the instrument's circles.

First and second order theodolites read circles at diametrically opposite points on the circles. In these cases, the actual *graduation error* is of little interest. What is of interest is the sum of the opposite errors, namely the *diameter error*. In some precision theodolites, the circles are read in four positions, 90° apart.

4.6.1 Periodic Errors

Periodic errors can be modelled with sine and cosine functions. Periods vary from 360° down to multiples of the basic circle graduation interval (from 180° downwards for instruments with diametrical circle reading). Most tests are usually restricted the fundamental wave period (360° or 180°) and some of its harmonics.

The periodic errors are traditionally determined by measuring the same angle between two (or more) collimators at different zeros of the (horizontal) circle. Other methods involve comparison against a laser interferometer or against a high precision theodolite. An example of the latter technique is shown in Fig. 4.25. Periodic circle graduation errors can be eliminated to a certain degree by appropriate observation procedures. These are discussed later.

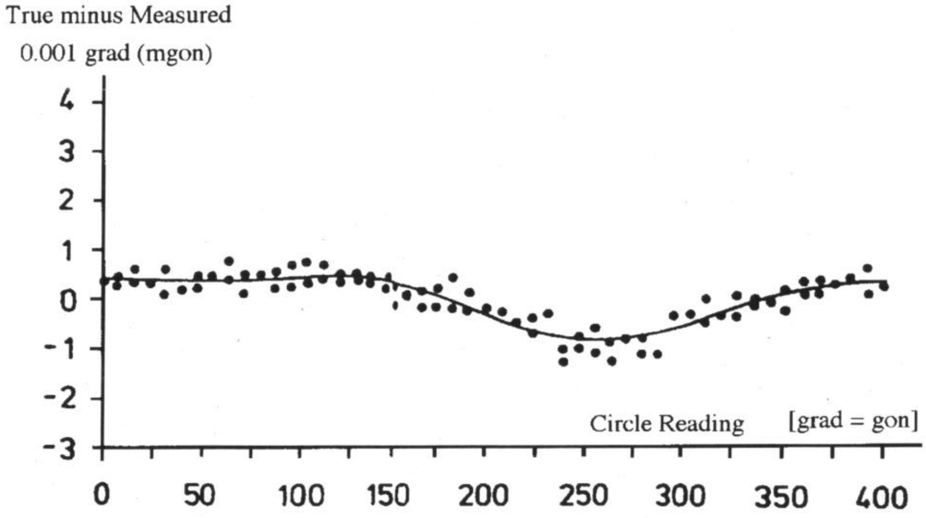

Figure 4.25: Periodic error of a specific electronic tacheometer Zeiss Elta 6. As the instrument does not feature diametrical circle reading, the fundamental period is 360° (or 400 grads; after Hennes & Witte 1991)

4.6.2 Interpolation Errors

Interpolation errors are the systematic errors of the fine measurement during the circle reading. As discussed previously, the circle interpolation is the measurement of the position of the reading index between two graduations of the circle. The interpolation errors are equivalent to the errors of the micrometer reading ('micrometer errors') in the classical instruments with visual circle reading.

The error patterns across the interpolation interval depend on the type of the interpolation system used. They may be due to errors in the index patterns on the mask, unequal illumination, a deviation of the actual signal from the (assumed) sinusoidal wave shape, and any errors in the spacing of the multiple reading indices (e.g. the spacing of the diodes that generate the 0°, 90°, 180° and 270° signals).

The interpolation errors can be determined in laboratories and modelled by sine and cosine curves (Fourier series). Figures 4.26a and 4.26b show the results of such tests and the fitted curves. If the calibration is done by the manufacturer, then there is a possibility of real-time correction (used in Zeiss ETh3, for example). However, if the calibration is done by the user, then post-correction is possible, in principle, but usually too cumbersome to be practical. Fortunately, the interpolation errors are usually very small. The table gives the results of some tests. For the horizontal direction measurement only, the interpolation errors could be reduced in the field, as the micrometer errors were with optical instruments. However, a physical change of the zero of the horizontal circle by a fraction of the interpolation interval is not practical and, in most cases, not possible.

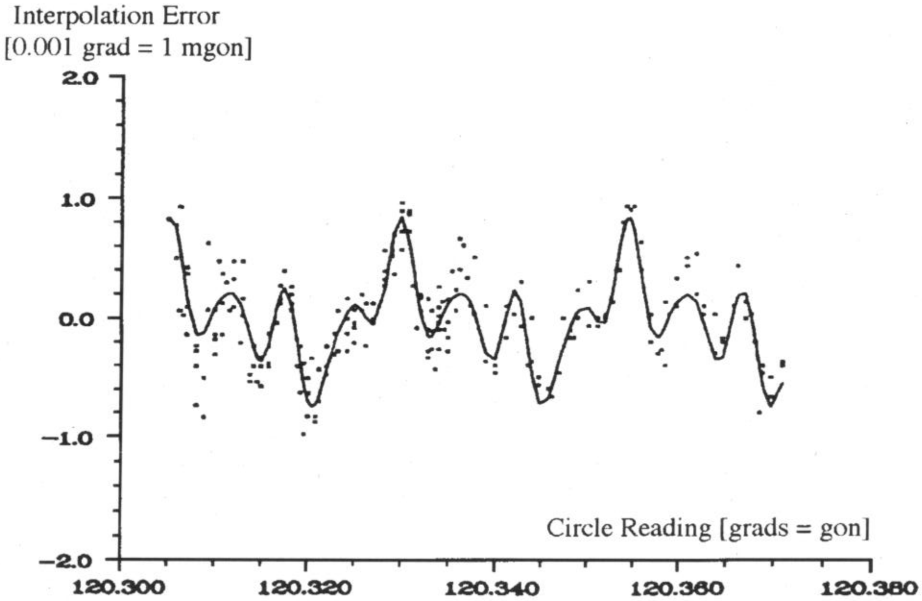

Figure 4.26a: Interpolation error of an electronic theodolite Zeiss ETh3. As the circle graduation interval is 82.3" (0.0254 grad), the interpolation error repeats a few times over the circle section shown. The peak-to-peak (p-p) variation is 5.2". (after Hennes and Witte 1991)

The interpolation error of another instrument (Nikon DTM-750) is shown in Fig. 4.26b. This electronic tacheometer features 16200 graduations over the full circle (360°) and, thus, has an interpolation interval of 80" exactly (in Fig. 4.26b shown as 24.69 mgon or 0.0247 grads). In this case, the peak-to-peak variation of the circle reading is 3.05" (or 0.94 mgon). The angular accuracy quoted by the manufacturer is ±2" (after the German industry standard DIN 18723). The curve shown in Fig. 4.26b (Schauerte et al. 1994) is the sum of a first order (period = 80") and a second order periodic error (period = 40") fitted by least squares. The corresponding amplitudes are 0.55" and 0.96", respectively.

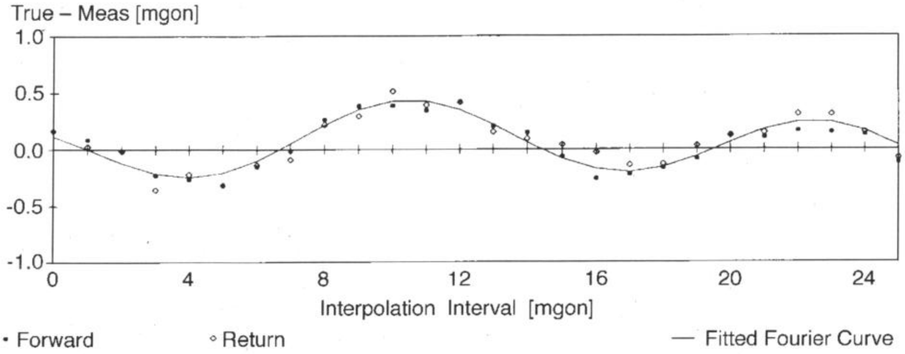

Figure 4.26b: Interpolation error of an electronic tacheometer Nikon DTM-750. Since the circle graduation interval is 80.0" (0.02469 grad, 24.69 mgon), the figure shows the interpolation error without repeats across one interpolation interval. The peak-to-peak (p-p) variation is 3.05". The dominant periodic error term has a period of 40.0" and an amplitude of 0.96". One second of arc is equivalent to 0.3 mgon or 0.0003 grads. (after Schauerte et al. 1994)

Instrument	Peak-Peak Variation	Interpolation Interval
Nikon DTM-750	3.0 "	80.0 "
Topcon GTS-6B	1.7 "	80.0 "
Topcon GTS-301	3.3 "	80.0 "
Zeiss ETh3	5.2 "	82.3 "

Table 4.1: Measured interpolation errors for a selection of individual instruments.

4.7 Some Notes on the Use of Electronic Theodolites and Tacheometers

It has been mentioned before that, in principle, the electronic theodolites and tacheometers should be used in the same way as the older optical instruments in general, and the one-second theodolites in particular. A few aspects have changed however and need to be considered. Only some aspects are discussed here. The aspects relating to electronic distance measurement are not covered. Readers may refer to the EDM textbook by Rüeger (1990, 1996), for example.

4.7.1 Levelling of Electronic Theodolites and Tacheometers

If an electronic tacheometer or theodolite is equipped with a dual-axis level sensor, and if the level sensor and the correction function for zenith angles and directions are enabled, then and only then is the levelling less critical than in the past. Because of the scale factor error of level sensors, it is suggested that the instruments be levelled to about ±20" to ±30" (one division on most plate levels) all the same.

Some instruments with dual-axis level sensors have displays like the one shown in Fig. 4.27. For instruments with identical displays on both sides, an operator can completely level an instrument without moving around the tripod. This saves time. The operator starts by levelling the fictitious bubble on the display in the first alidade orientation. After each turn on the footscrews, one should let the values settle for about three to five seconds (to a 'constant' value). This is equivalent to getting the values for the x and y tilts to zero. (Normally, it is not possible (nor necessary) to make the x and y tilts exactly zero as the pitch of the footscrews is not fine enough and the tripod often not stable enough.) The alidade is then rotated 180° about its vertical axis. If the fictitious bubble moves out of the centre position (and the x and/or y tilt values change), the levelling is changed half-way back (to the mean values/position of both 'faces'). The theodolite or tacheometer is truly vertical, if the level sensor reads the same values of x and y in all orientations of the rotating upper part ('alidade') of the theodolite.

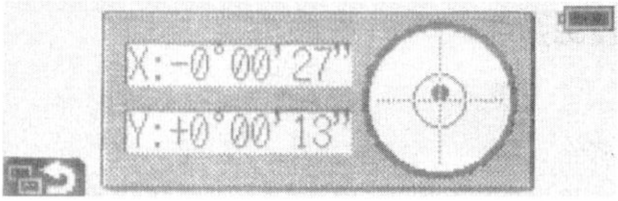

X:-0°00'27"
Y:+0°00'13"

Figure 4.27: Level sensor read-out and levelling display on an electronic tacheometer Topcon GTS-6

4.7.2 Selection of Measuring Units

Since the introduction of the electronic theodolites and tacheometers, the units of the angular (and distance) measurements can be changed easily by the observer. Typically, 400 grads (gon), 360° decimal, 360° sexagesimal

(° ' ") and 6400 mils are offered. Many users don't realise that the resolution can be quite different, depending on the selected unit. For example, the Wild TC1610 has the following resolutions: 1^{cc} (0.33"), 1", 0.0001° (0.36"), 0.001 mil (0.20"). Note that 1^{cc} = 0.1 mgon = 0.0001 gon = 0.0001 grad. Clearly, the best resolution is achieved in the military mil setting.

When recording data, users of electronic instruments should ascertain that the post-processing software can use the selected units. Unfortunately, this may greatly limit the choice available in the field. *When booking by hand* and doing the post-processing by hand, it is suggested to use a decimal unit, like decimal degrees (or grads(gon) for that matter). This simplifies the field reductions and avoids the cumbersome conversions from degrees, minutes and seconds to decimal degrees on calculators.

4.7.3 Use of Umbrella

There are many good reasons why an umbrella should be employed to provide shade for the observer, the instrument and the tripod. Some of these are given below.

4.7.3.1 Electronic Theodolites

- Australia has the highest incidence of skin cancer in the world. (One in 45 Australians is affected by melanoma!) Similarly, extensive exposure to sun can cause permanent eye damage. To provide shade for field parties is, thus, essential and an occupational health and safety (OHS) requirement.
- The operating temperatures of instruments are specified as ambient, shadow temperatures. In full sunlight, the inside of the instrument can be 20°C higher. Some instruments cut-out automatically when the upper limit of the specified operating temperature is reached. A shaded instrument may work whilst an unshaded may not!
- The pointing to targets, either manually or automatically, can be very difficult, if not impossible, when the sun is shining into the optics. This increases the time required for the pointing and may reduce the accuracy of the measurements.
- A change of the temperature of the instrument leads to temperature gradients through the mechanical and optical components. Temporary misalignments and changes of instrument errors occur. Typically, about 5 minutes per °C temperature change are required for an instrument to reach a dimensional equilibrium at a new temperature. This waiting time after temperature changes (e.g. after moving the instrument from an air-conditioned car into 40°C heat) should be observed before beginning high precision measurements. Proper acclimatisation of instruments is particularly important when using automatic pointing systems. In consequence, it is advisable not to leave instruments in a hot, stationary car;

rather, they should be unloaded and stored in a shadowed area in the open air (e.g. under an umbrella).

- The level sensors are usually mounted in one of the two standards and thus fully exposed to heating by the sun in the absence of an umbrella. Changing face alternates the level sensor between the sun and the shadow side, with all the negative effects on it temperature equilibrium.

4.7.3.2 Electronic Tacheometers

- Do not aim the telescope at the sun: You may burn out the transmitting and receiving diodes in doing so and, thus, permanently damage the instrument. Do not carry out sun azimuth observations with an electronic tacheometers unless a *sun filter* (approved by the manufacturer) is *always attached* to the objective. (Not on the eyepiece, as in the past!)

4.7.4 Automatic Power-Off

To extend the charge of batteries in the field, some instruments switch themselves off after being idle for a specified time. For *measurements of highest precision*, the instrument should **not** be turned off during measurements. This ensures that the temperature inside the instrument remains as constant as possible.

4.7.5 Two-Way Radios and Cellular (Mobile) Phones

It is advisable not to transmit from walkie-talkies and cellular phones during actual measurements. Since mobile phones respond to incoming calls without user input, they should be totally switched off when operating an electronic theodolite or tacheometer. Transmitting devices may interfere with measurements, leading to an erroneous result or to a total failure of the measurements. Instances of the latter were reported at UNSW with respect to distance measurements.

4.7.6 Real-Time Corrections to Measured Data

The vertical axis tilt(s) is(are) computed by electronic tacheometers and theodolites and used to correct the zenith angles (single-axis level sensors) or the zenith angles and the horizontal directions (dual-axis level sensors). These corrections are combined with other on-line corrections. For example, for face left:

$$z_{FL} = z'_{FL} + S^{TE}_{FL} \times C - S^{TE}_{REF} \times C + VCIC \qquad (4.15)$$

$$r_{FL} = r'_{FL} + \frac{c}{\sin(z_{FL})} + (S^{HA}_{FL} \times C - S^{HA}_{REF} \times C + TAI) \times \cot(z_{FL}) \tag{4.16}$$

where z is the corrected and z' the raw zenith angle measurement, r is the corrected and r' the raw horizontal direction measurement, VCIC is the vertical circle index error correction, c is the collimation error correction, TAI is the correction for the trunnion axis inclination and where the 'exponents' TE and HA denote the **te**lescope and the **h**orizontal **a**xis components of the level sensor measurements. In face right, the formulae are a function of the face right zenith angle:

$$z_{FR} = z'_{FR} + S^{TE}_{FR} \times C - S^{TE}_{REF} \times C + VCIC \tag{4.17}$$

$$r_{FR} = r'_{FR} + \frac{c}{\sin(z_{FR})} + (S^{HA}_{FR} \times C - S^{HA}_{REF} \times C + TAI) \times \cot(z_{FR}) \tag{4.18}$$

The Equations (4.15) and (4.17) highlight that the electronic theodolites and tacheometers have two 'index errors', one for the vertical circle reading mechanism (VCIC) and one for the level sensor (S_{REF}). Please note that any errors in S_{REF} will be cancelled if the mean of face left and face right is taken in zenith angle and horizontal direction measurements. Any errors in C will **not** be cancelled by this procedure.

The correct sign of the level sensor readings will depend on the construction of the level sensor and the position inside the instrument. When using a liquid level sensor for the measurement of the vertical axis tilt, it should be noted that it takes longer to settle down than a mechanical pendulum device, due to the viscosity of the liquid. It takes about 3-4 seconds to settle in normal conditions, longer in cold conditions. This waiting time must be observed (for zenith angles **and** direction measurements) if the above on-line corrections are enabled.

The Eqs. (4.16) and (4.18) are not defined when pointing to the zenith because of the division through zero. If Eqs. (4.15) and (4.17) are not applied, the displayed and recorded zenith angles will not refer to the plumbline but rather to the (inclined) vertical axis. The real-time corrections after Eqs. (4.15) and (4.17) may have to be disabled in cases, where the instrument is set-up on a moving or vibrating platform. Measurements are then taken in a coordinate system defined by the instrument's vertical axis and the zero direction of the horizontal circle.

4.7.7 Direction Measurements

The electronic theodolites and tacheometers are, as far as the mechanical design and the telescope are concerned, not very different from classical theodolites with visual reading of circles. In consequence, most of the existing rules of measurements with classical theodolites still apply. These accepted procedures are part of good professional practice and have been established over decades if not centuries.

The new technology used in electronic theodolites may require a revision of some of the accepted procedures, a drop of others and an introduction of new procedures. Only the observation rules that relate to the circle reading (resolution), the circle zero and the level sensors (replacing compensators and altitude bubbles) need to be reassessed. All other rules are still valid and should be adhered to.

4.7.7.1 Single-Face versus Dual-Face Measurements

The decision between single-face (FL only) and dual-face (FL & FR) measurements is the same as in the past. *Single-face observations* are acceptable whenever:

- the combined effects of all residual instrument errors are smaller than the required angular accuracy,
- the validity of the values of the corrections for instrumental errors stored on-board for real-time correction are confirmed in the field,
- the required angular accuracy is less than the resolution of the instrument and
- the gross errors can be detected by means other than a second measurement.

The fact that some instrument errors are corrected by computations in real-time, does not change much. In fact, the errors in the corrections for the physical errors can be quite large if the calibration measurements were not carried out properly. (In other words, the correction for an error might be a bigger problem than the error as such.) The stored corrections (like the mechanical adjustments in the past) may also change with temperature and time.

Single-face measurements are typically used in tacheometry, detail surveys and setting-out. When measuring in FL only, the speed of surveys increases considerably. However, the accuracy and reliability of single face measurements must be assured always. Quality assurance requires that all data are checked one way or another.

Dual-face measurements are executed whenever:

- the accuracy of the measurements is to approach the resolution of the instrument,
- as many as possible instrument errors are to be truly eliminated on the spot, and
- any gross errors are to be avoided or detected immediately by a second set of data.

Dual-face measurements are used for control surveys (traversing, trilateration) and for high precision surveys, usually in engineering applications. Dual-face measurements improve the reliability of the data considerably, as all points are pointed to at least twice. The errors of identification are reduced.

4.7.7.2 Sequence of Direction Measurements

If an arc of direction is to be measured to targets 1,2,3,4 .. ,t, then the accepted sequence of the direction observations is:

- {rotate alidade clockwise by 360°}
- in clockwise (cw) direction: FL(1)–>FL(2)–>FL(3)–>....–>FL(t)
- change face {and rotate alidade counterclockwise by 540°}
- in ccw direction (read from right): FR(1)<–FR(2)<–FR(3)<–....<–FR(t)

Actions listed in { } brackets are suggested for high precision measurements for the lubrication of the bearings. The sequence of the points is selected in such a way that the smallest sector is covered (FL(t) – FL(1) = minimum). This minimises the walking around the tripod or the pillar. The accepted sequence does also eliminate most of the effect of the tripod or pillar rotations (the so-called 'tripod twist'). The assumption for this is that the time spent on measuring to each point is about the same.

A *horizontal angle* is an arc of direction with two rays and is measured the same way. In traverses, the clockwise angle from the backsight to the foresight station is measured, irrespective of it being the smaller of the two possible angles or not.

4.7.7.3 Elimination of Periodic Errors

For high precision direction measurements, the periodic graduation errors and diameter errors should be eliminated by a physical change of the circle's orientation between multiple arcs. (A real *change of the zero* of the circle is required. Changing the circle reading through the keyboard *does not help*!) When measuring n arcs, the circle orientation should be changed physically by 180°/n for electronic theodolites with diametrical circle reading and 360°/n for electronic theodolites with a single index.

Unfortunately, most electronic theodolites and tacheometers do no longer feature a horizontal circle setting screw. The rotation (by 120°) of the whole instrument inside its tribrach between arcs is then the only feasible approach. This should be done for high precision direction measurements.

4.7.7.4 Circle Setting

With optical theodolites and manual booking, it was appropriate and beneficial to set the circle to a little bit more than zero to the left most target, in face left, in the first arc. For subsequent arcs, the 180°/n or 360°/n rule (see previous paragraph) was employed. If manual booking is used with electronic theodolites and tacheometers, this procedure is still appropriate, even if the (virtual) change of the circle reading is achieved through keyboard entry of the desired value.

When using *electronic instruments with electronic data recording*, however, setting of the circle to zero (or any other value) is a waste of time. It is suggested to start at whatever orientation the circle is after the set-up. When measuring multiple arcs of directions in precision work, the procedure outlined in the previous section on the *Elimination of Periodic Errors* should be used. If the circle cannot be rotated physically one way or another, multiple arcs might as well be measured with the same circle orientation and readings when using data recording. Entering a new starting value for each arc through the keyboard does not change any errors and is, thus, a waste of time.

4.7.8 Zenith Angle Measurements

Good levelling of the theodolite is required, despite (and because of) the fact that an electronic level sensor or a compensator is being used. For more precise work, all zenith angle readings should be done in a certain sequence, e.g. FL-FR-FR-FL, to offset time varying changes of the vertical circle index correction (VCIC) and the additive constant of the level sensor. In this routine, the corresponding FL and FR readings follow each other immediately. It can be assumed that any errors in the additive constant of the level sensor are the same and, thus, cancelled. Also, the vertical refraction is unlikely to change significantly in the few seconds that elapse between the FL and FR readings. The FL and FR readings are, thus, subject to the same refraction error.

Large magnetic fields can affect mechanical pendulum level sensors. A liquid level sensor is recommended for measurements in electro-magnetic fields. The liquid level sensors take longer to settle down due to the viscosity of the liquid. On average they take 3-4 seconds in normal conditions to settle, longer in cold conditions.

As the zero of the zenith angle is traditionally fixed, the only way to eliminate the periodic graduation or diameter errors is to determine them and to correct the measurements later in the office or on-line, through a linked personal computer. As the circle graduation errors of vertical circles are not easily determined, the calibration and the post-correction are usually not carried out. This is another reason why zenith angles, typically, are less precise than horizontal directions.

4.7.9 Simultaneous Direction, Zenith Angle and Distance Measurements

With electronic tacheometers, the cross hairs **must** be pointed to the centre of the reflector for a distance measurement. On longer distances, where the apex of the glass prism cannot be seen, the pointing occurs to a target plate symmetric to the reflector. It is convenient (also for recording purposes), to use the same pointing for the directions and the zenith angles. The errors caused in angular measurements by an erroneous pointing of the apex of the reflector are not discussed here. Please refer to Rüeger (1978, 1990, 1996), for example.

If the standard measuring routine for arcs of directions is followed (see above), the quality of the zenith angles will depend on the elapsed time (and any changes in the ambient temperature) between the corresponding FL and FR readings. The vertical refraction and the additive constant of the level sensor do change with time. This is because the ambient temperature and the vertical temperature gradient also change with time. Simultaneous direction, zenith angle and distance measurements might be possible for precise measurements, if the temperature and the refraction conditions remain the same. One such case arises in close range three-dimensional (3D) industrial measurements indoors. Otherwise, simultaneous direction, zenith angle and distance measurements are not recommended for precision work.

4.7.10 Plumbing of Structures

Theodolites can be used for the plumbing of structures in two major ways. The first technique establishes a point directly above the instrument, for example in a vertical shaft or well. The second is used to establish the verticality of the corners of a building.

4.7.10.1 Plumbing Shafts and Wells

Plumbing from the bottom to the top of a shaft or well is possible, if a diagonal ('broken') eyepiece is attached to the electronic theodolite or tacheometer. These attachments are offered as accessories for most instruments. To establish a point directly above the intersection of the vertical and trunnion axes of an instrument, level the instrument in the normal way, switch the level sensor and the tilt corrections on, turn the alidade to read 0° on the horizontal circle, tilt the telescope until a zenith angle of 0° 00' 00" is read. Mark the intersection of the line of sight on a suitable target at the top of the shaft or well. Then turn the alidade to read 90° on the horizontal circle, reset the telescope to a zenith angle of 0° 00' 00" and mark the intersection of the second line of sight on the target at the top. Turn the alidade to read 180° on the horizontal circle, reset the telescope to a zenith angle of 0° 00' 00" and mark the intersection of the third line of sight on the target at the top. Then turn the alidade to read 270° on the horizontal circle, reset the telescope to a zenith angle of 0° 00' 00" and mark the intersection of the fourth line of sight on the target at the top. These four measurements are taken in the same face. Because of instrumental errors, the line of sight travels on a cone when rotating the alidade about the vertical axis and 'draws' a circle on the target. The centre of the (imaginary) circle is the (error free) plumbing point. The accuracy of the plumbing depends on the accuracy of the level sensor, how accurately the plumbing points are marked and on how accurately the zenith angle can be set to zero (exactly). The latter also depends on the stability of the tripod.

4.7.10.2 Plumbing a Building

When checking the verticality of building corners, for example, instruments with dual-axis level sensors are preferred since they correct the direction measurements for (residual) tilts of the vertical axis. This is important for steep sights. (The diagonal eyepieces mentioned in the previous paragraph may also be useful.) To check the verticality of a corner in one direction, set up the instrument in line with one of the two facades and sufficiently offset from the corner to avoid zenith angles less than 45°. Level the instrument in the usual way, switch on the dual-axis level sensor and the tilt corrections, point (in FL) to the corner at the top, read the (corrected) direction, point (in FL) to the corner at the bottom, read the (corrected) direction, point (in FR) to the corner at the top, read the (corrected) direction, point (in FR) to the corner at the bottom and read the (corrected) direction. Calculate the mean direction to the top and the mean direction to the bottom. The building is vertical (at the corner and in the direction measured), if the two values are the same. The accuracy depends on the angular accuracy of the instrument, the pointing accuracy and the accuracy of the level sensor. The latter has to be multiplied with cot(z), since the tilts of the vertical axis, measured by the level sensor, are multiplied with cot(z) to derive the corrected directions. Refer to Eqs. (4.16) and (4.18) for details.

4.8 Exercises

4.1 List three important differences between optical and electronic tacheometers ('total stations') and briefly explain their significance.

4.2 Explain the major differences between two robotic surveying systems, namely a one-person-survey-system and an automatic monitoring system.

4.3 Describe two completely different types of incremental circles used in electronic tacheometers.

4.4 Explain how the field operations differ between instruments with electro-optical incremental and absolute horizontal circles.

4.5 Describe the key differences between vertical circle compensators and electronic level sensors in electronic tacheometers.

4.6 What are dual-axis level sensors good for? How valuable are these activities?

4.7 An electronic tacheometer, designed according to Fig. 4.15, features an incremental circle with 21600 graduations.

(a) What is the resolution of the coarse measurement on the circle?

(b) At what intervals would you expect the interpolation error to repeat itself?

(c) Assuming that the instrument employs a 6-bit analogue-to-digital converter, what would the resolution of a single circle reading be?

4.8 Based on Fig. 4.17, design two other logical tests for the determination of the sense of the rotation, one each for the clockwise and the counterclockwise movement.

4.9 Redraw the window (b) of Figure 4.20 for the case, where the vertical axis is tilted the other way (by the same amount).

4.10 Using Eqs. (4.12) and (4.14), investigate the effect of a 1% error in C on α and S_{REF}, respectively.

4.11 How significant is the systematic component of the setting error in liquid level sensors?

4.12 You have been measuring zenith angles with an electronic theodolite with a dual-axis liquid level sensor. The corrections for the dislevelment were 'switched-on'. At +20°C, you noted a total vertical circle index correction (VCIC) of –12". At +35°C, you found a VCIC of +2". What do you attribute this change to? Why?

4.13 Would you prefer to level a theodolite with the procedure outlined in Figure 4.27 or with the traditional method with the plate level? Give reasons!

4.14 (a) Considering Figs. 4.21(b) and 4.21(c), Eqs. (4.15) and (4.17) and assuming that $S_{REF}^{TE} = 0.000$, how do S_{FR}^{TE} and S_{FL}^{TE} relate to each other?

(b) Noting that the mean zenith angle is $\overline{z} = 0.5\,(z_{FL} + (360° - z_{FR}))$, how would an error of 2% in C affect $\overline{z}$ for $S_{FL}^{TE} \times C = 20''$ and for $S_{FL}^{TE} \times C = 100''$, respectively?

4.15 Considering Eq. (4.16) and Eq. (4.18), $\overline{r} = 0.5\,(r_{FL} + (r_{FR} \pm 180°))$ and $z_{FR} = 360° - z_{FL}$ and assuming $c = 0.0''$, $TAI = 0.0''$, $S_{REF}^{HA} \times C = 0.0''$,

(a) prove that the level sensor readings $S_{FL}^{HA} \times C$ and $S_{FR}^{HA} \times C$ are not cancelled in $\overline{r}$, and

(b) discuss the effect of $S_{FL}^{HA} \times C = +10''$ on $\overline{r}$ for zenith angles z_{FL} of 90°, 70°, 50°, 30° and 10°.

4.16 You have been asked to check the verticality of a hight rise building. For this purpose, you have set up an electronic theodolite with a dual-axis level sensor. You are pointing in face left (FL) to a point at ground level on the building as a reference and you set the direction reading to 0°00'00". (Here, it is assumed that the zenith angle to this point is 90° exactly.) You then check the verticality of points above the reference (in FL) by tilting the telescope to zenith angles of 70°, 50° and 30°.

(a) Assuming that your instrument employs the Eqs. (4.16) and (4.18) to correct directions in real-time and that $c = +15''$, $TAI = -8''$, $S_{REF}^{HA} = 0.0$ exactly, $S_{FL}^{HA} \times C = -15''$, what direction would you read after tilting the telescope (without any alidade rotation) from the reference mark to $z=70°$, $z=50°$ and $z=30°$?

(b) What would you do next, at each zenith angle setting ($z=70°$, $50°$, $30°$) before assessing the verticality?

(c) Give reasons why you would repeat the process in face right (FR) afterwards?

5. Specialised Equipment

5.1 Reflectorless Electronic Distance Measurement

The evolution of reflectorless EDM in surveying has been discussed in Section 4.1.8. In this section, the special aspects of measuring to non-cooperative targets with modern instruments are discussed.

Since reflectorless EDM relies on the spurious reflectance of features of the landscape, the range is much reduced in comparison to measurements to normal EDM glass prism reflectors. To increase the range of reflectorless EDM instruments, manufacturers have to increase the power output of the emitting diodes, use the pulse distance meter principle, where short and powerful pulses are separated by idle times (low duty cycle), increase the sensitivity of the receiving diodes, focus the EDM beam (noting that most electronic tacheometers feature fixed focus distance meters) or any combination thereof. Reflectorless instruments measure the distance to the area ('footprint') on the target that is illuminated by the distance meter's beam. If the distance meter uses a visible beam, then the footprint may be visible trough the telescope, at least on close range If the distance meter uses an infrared beam, then the observer has to assume that the optical line of sight is properly aligned with the EDM beam. Should the EDM beam be out of alignment, then the distance is not measured to the point indicated by the crosshairs on the object. Note that some instruments transmit a visible light beam for target identification, even though the distance measurement is carried out with an infrared beam. In such cases, the same uncertainties arise as between the optical line of sight and an infrared EDM beam.

5.1.1 Basic Properties

Figure 5.1 shows some of the basic properties of reflectorless EDM. At the top left of Fig. 5.1, the ideal case for reflectorless EDM is shown where the beam hits a flat surface that is at a right angle to the central ray (i.e. angle of incidence is zero) and where the footprint of the beam is well within the surface. This is clearly the preferred configuration for reflectorless measurements. The average distance across the footprint will be measured. At the top right of Fig. 5.1, the flat surface is not at a right angle to the beam any more. The angle of incidence on the surface is no longer zero and the footprint is now an ellipse. It is likely that less light will be returned to the distance meter, when compared to the first case, and that the measuring range is reduced. Again, the average distance across the footprint is measured. Unfortunately, the centre of gravity of the illuminated area ('footprint') does not coincide with the point where the central ray intersects the surface. This error was calculated by Köhler (1994) for a Zeiss Reg Elta RL as 3 mm at 60 m for an (extreme) angle of incidence of 81°. For angles of incidence closer to zero, this error is insignificant.

The case of a rough (structured) surface is illustrated at the bottom left of Fig. 5.1. Again the average distance across the illuminated area is measured. Note that tape measurements would be to the 'highest' point of the structured surface rather than the average and, thus, will differ from reflectorless measurements. Reflectorless EDM measures to a point inside the structured surface. Shiny surfaces (bottom right of Fig. 5.1) may act as mirrors and reflect the EDM beam. Typically, the reflectorless measurement will fail to such surfaces. Should the *deflected* EDM beam hit another surface of sufficient reflectivity, then an (erroneous) distance may be measured to the other target surface, through the shiny surface. In consequence, glass and shiny surfaces should be avoided.

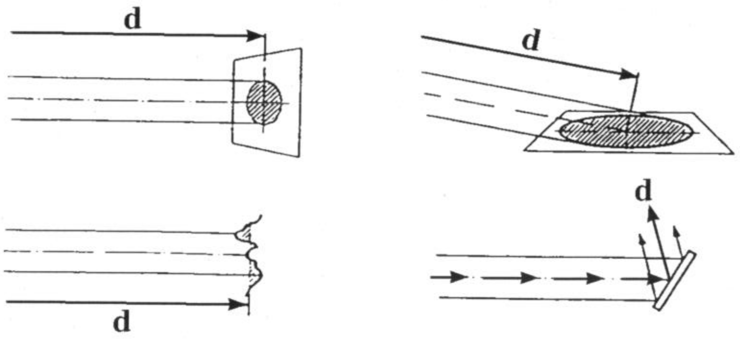

Figure 5.1: Basic properties of (reflectorless) distance measurements to natural targets (after Ibeo 1986)

Since a reflectorless EDM instrument illuminates a circular area (footprint) rather than a dimensionless point, all distance measurements to corners and rounded surfaces will be wrong to a certain extent. Distances to convex corners (on left in Fig. 5.2) will be measured too long and distances to concave corners (in the middle of Fig. 5.2) too short. The magnitude of the error depends on the angle of incidence, the diameter of the footprint and, thus, on the instrument used and the distance to the corner. The distance to rounded surfaces, such as cylinder and spheres (see right of Fig. 5.2), are also measured too long. Typically, the errors will be less than those to corners.

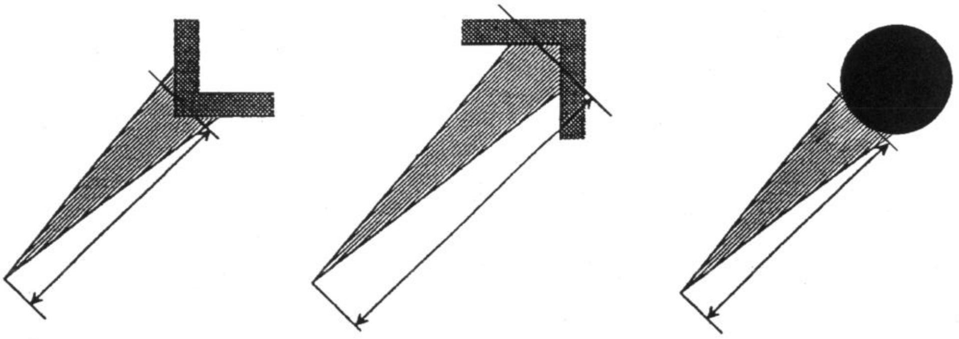

Figure 5.2: Measurement to corners and round surfaces (after Buchmann & Scherer 1994)

5.1.2 Beam Size and Shape

Because reflectorless EDM instruments do measure to the area that is illuminated by the instrument, it is important for the user to be fully aware of the dimensions and shape of the beam of the reflectorless distance meter when measuring to surfaces. Nikon, for their NPL-821 reflectorless pulse distance meter, write that the reflectorless EDM beam is circular and has the same diameter as the central circle present on the crosshairs of the instrument. The latter is 5.00 minutes of arc. The beam diameter, thus, increases linearly with distance. Nikon also states that focussing the telescope of the NPL-821 onto a target also focusses the reflectorless EDM beam. (As far as known, the other manufacturers use fixed focus EDM beams.) Leica quotes different beam cross-sections and dimensions for different distances for their instruments operating in reflectorless mode. Figure 5.3 gives the information for the reflectorless mode of the Leica 1100 Series of instruments. The reflectorless EDM beam is coaxial with the telescope when properly adjusted.

In normal EDM, the knowledge of the EDM beam dimensions is less critical, since the glass prism reflector has the same diameter as the objective lens of the instrument. As long as no strongly reflecting surfaces are located behind the prism, only that part of the EDM beam, that illuminates to prism, will be reflected back to the instrument and used for the distance measurement. In contrast to reflectorless EDM, it is quite clear in normal EDM to what point a measurement refers to, namely the intersection of the reflector's horizontal and vertical axes.

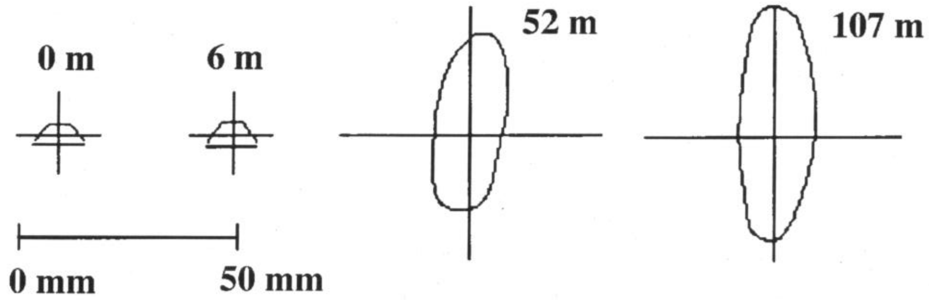

Figure 5.3: Beam shape and dimensions of Leica Series 1100 reflectorless electronic tacheometers (after Leica 2001)

The effective beam size and shape can be easily determined by the user. Since this information is essential for reflectorless measurements, tests should be carried out at a number of distances and vertically and horizontally. The principle is shown in Fig. 5.4.

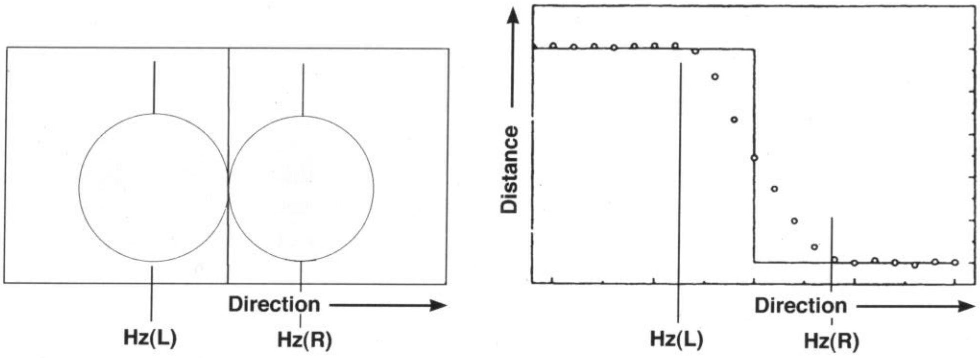

Figure 5.4: Scanning across an edge to determine the (horizontal) beam width of a reflectorless distance meter. Front view on the left, plan view on the right. On the right, the dots represent possible changes in distance measurements across the edge (after Meier-Hirmer 1996). The EDM footprints at the direction readings Hz(L) and Hz(R) are shown on the left.

To determine the horizontal width of the EDM beam, two blocks (for example, of wood) are arranged in such a way that their front faces are approximately perpendicular to the beam and separated by 0.1 m or so. On the right hand side of Fig. 5.4, the block on the right is closer than the block on the left. The view from the distance meter is shown on the left hand side of Fig. 5.4. The vertical edge that separates the left and the right block should be about vertical, as shown. To determine the horizontal beam width, one points the vertical crosshair to the edge and records the distance and the horizontal direction. (Since half the beam hits the back target and half the front, one expects to read the average distance to the two surfaces.) Then, the scan to the left starts by measuring distance and direction at appropriate angular intervals. Once the distance readings stabilise (at the distance reading to the back block on the left), the left hand scan is terminated. Point the vertical hair again to the edge, and record direction and distance. Then, start the scan to the right, using the same angular intervals as before. Once the distance readings stabilise (at the distance reading to the front block on the right), the right hand scan is terminated. As a check, point the vertical hair again to the edge, and record direction and distance. As Fig. 5.4 shows, the positions Hz(L) and Hz(R) are the closest points to the edge, where the beam is still fully on the front or back target. The beam width (in angular units) is obtained as Hz(R) - Hz(L).

The test should then be repeated in a vertical direction, across a horizontal edge. Afterwards, the beam width should also be determined at another distance. As a minimum, the beam width should be established at the shortest and longest distance experienced in daily practice. The beam width test discussed here allows to check if the EDM beam is properly aligned to the telescope's crosshairs. For a EDM beam, that is properly aligned in the horizontal direction, the direction readings to the edge (at the beginning, middle and end of the test) should be the mean of Hz(R) and Hz(L), to use the example of Fig. 5.4.

5.1.3 Multiple Target Problem

Since reflectorless instruments are usually pointed to non-cooperative targets in the landscape, it can happen that the EDM beam exceeds the diameter of the target. It is then possible that the part of the beam, that overlaps the intended target, is returned by a surface behind the target. A similar case arises if small reflecting surfaces (such as leaves), located between the instrument and the intended target, interfere with the beam. The effect of multiple targets on the distance displayed by the instrument is different for phase measuring reflectorless instruments (e.g. Leica instruments) and pulse reflectorless distance meters (e.g. Nikon and Topcon instruments, Zeiss Rec Elta RL).

5.1.3.1 Pulse Distance Meters

The effect of multiple targets for pulse distance meters varies with the distance between the two targets as well as the relative strength of the signals returned by the two targets. Figure 5.5 shows experimental results obtained by Meier-Hirmer (1995) with a Zeiss Reg Elta RL. The closer target was 22 m (distance s_0) from the instrument. The second one was set up behind the first one so that both targets were visible from the instrument. The more distant target was then shifted in steps (distance s_1) until it was 4.000 m behind the first target. Each time, the instrument was pointed to the edge between the two targets, focussing to the closer target in all cases. The back dots in Fig. 5.5 represent the measurements. In Fig. 5.5, the abscissa shows the distance ($s_1 - s_0$) between the two targets (in the direction of the EDM beam). The ordinate shows the difference between the EDM measurement (s) and the distance (s_0) to the first target. Since half of the footprint is on the forward target and half on the backward target, one expects to measure the average distance, that is $(s - s_0) = 0.5\ (s_1 - s_0)$. This line is shown in Fig. 5.5. The measurements follow the line up to a step interval of 0.3 m. After that, the distant target has lesser influence. As soon as the step interval exceeds 3.000 m, the distance meter readout approaches again the distance value (s_0) to the closer target.

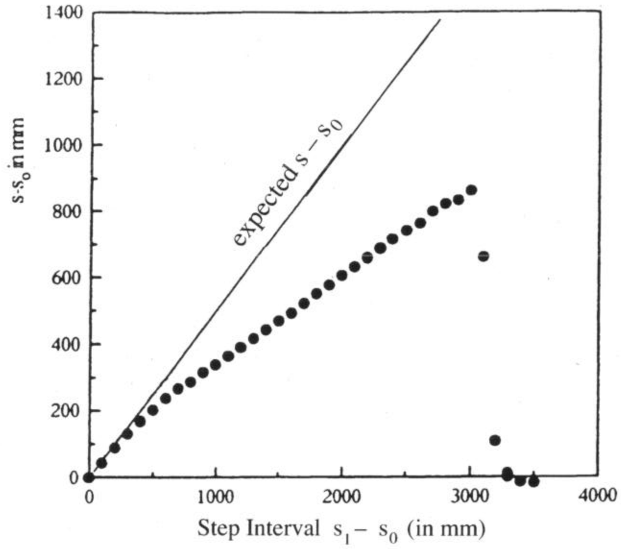

Figure 5.5: Measurements with a pulse distance meter (Zeiss Reg Elta RL) to two targets, the separation of which changes from 0 to 4 m. The closer target was steady at 22 m. See text for explanation. (adapted from Meier-Hirmer 1995)

The pattern shown in Fig. 5.5 may be caused by the pulse length of the Zeiss Reg Elta RL. Assuming a 50% return by each target, a type of average distance is measured if the double of the difference in the distances to the two targets is less than the distance equivalent of the length of the pulse transmitted by the instrument. This is because the leading edge of the return pulse from more distant target arrives at the instrument before the return pulse from the closer target has died down. Once this critical separation is exceeded, it is likely that only the distance to the closer target is measured. Experiments have established this critical separation at about 3 m for the Zeiss Rec Elta RL (Meier-Hirmer 1996 and Fig. 5.5) and at about 1.8 m for the Nikon NPL-821.

5.1.3.2 Phase Measuring Distance Meters

As with the pulse distance meters, the effect of multiple targets for reflectorless phase measuring distance meters varies with the distance between the two targets as well as the relative strength of the signals returned by the two targets. The effect was investigated by Scherer (1999) with a Leica TCRM electronic tacheometer. The experimental set up was similar to the one discussed in the previous section. The closer target was always at 4.3 m and the more distant target at distances from 4.3 to 7.3 m. At each target separation, Scherer (1999) carried out a lateral scan across the edge of the targets, like the test explained in Fig. 5.4. This means, when scanning to the right, the return from the distant target increased from 0% to 100%. Figure 5.6 depicts the results on two scans only. The left part of Fig, 5.6 shows the result of the scan with the second target at 6.6 m, the right part the scan with the second target at 5.2 m.

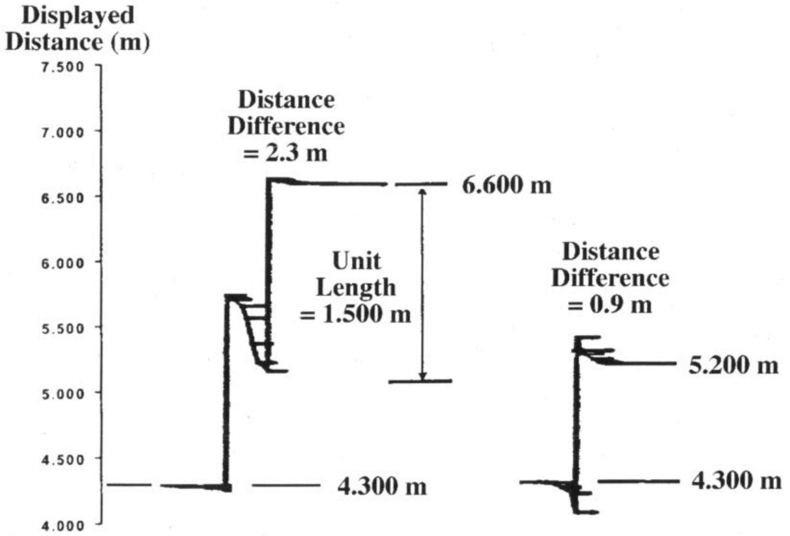

Figure 5.6: Scanning across two targets with a (phase measuring) Leica TCRM instrument. The first target is at 4.300 m. The second target is 2.300 m (on the left) and 0.900 m (on the right) behind the first. Scanning (from left to right) begins with the full beam on the closer target (on left) and continues then in equidistant lateral steps until the beam is fully on the back target. (from Scherer 1999)

It follows from Fig. 5.6 that the displayed distances are either close to the distance to the closer target or close to the distance to the more distant target, but never half way in between, if the separation does not exceed the unit length of the instrument (1.5 m in this case). The matter of concern is that the displayed distances may be shorter than the distance to the closer target, or longer than the distance to the distant target. Scherer (1999) reported magnitudes of such errors of up to 0.20 m. The behaviour of phase measuring distance meters is quite different from that of pulse distance meters. Clearly, it is best to avoid measuring to multiple targets!

5.1.4 Influence of the Material and Colour of the Target

The material, roughness, reflectivity and colour of a target affect the maximum distance and precision that can be achieved. Also, gross errors can occur in certain conditions.

The experiments by Kogoj (2001) established the following maximum distances with a Leica TCR307 to targets at 90° to the EDM beam (zero angle of incidence): sheet metal (60 m), render (85 m), wood (100 m) and styrofoam (110 m). When measuring to coloured paper with a Leica TCRM1100, which uses a visible (red) EDM beam, white, red and yellow gave the best return signal and blue and green the worst (Möser 2000).

Kogoj's (2001) experiments with a Leica TCR307 showed that the precision gets worse with distance. However, there was no noticeable difference in precision between the measurements to sheet metal, render, wood and styrofoam.

Möser (2000) investigated the variations in distance measured with a Leica TCR1100 to number of different target materials at the same physical distance. Errors of less than 0.5 mm were recorded for paper, (smooth) PVC, (smooth and rough) steel. Wood and wallpaper were within 1 mm. At 20 m, errors of 15 mm were recorded for acrylic glass, 7 mm for styrofoam and 4 mm for sandpaper. Clearly, measurements to (and through) glass should be avoided as should measurements to styrofoam and sandpaper.

5.1.5 Target Orientation

The target orientation (or, in other words, the angle of incidence) does have an effect on the maximum distances that can be achieved to a particular material. For sheet metal and render, Kogoj (2001) recorded a 5% and 6% reduction in the maximum distance, respectively, from zero to 45° angle of incidence with a Leica TCR307.

Another problem of target inclination is shown in Fig. 5.7 (Köhler 1994). The figure shows the distances measured in two faces with a pulse distance meter Zeiss Reg Elta RL to a target with different tilts about a horizontal axis. The angle of incidence is given on the abscissa (in grads). In the Zeiss Reg Elta RL, the EDM beam is transmitted coaxially. The receiver optics is, however, below the main telescope. Even though the misalignment is corrected, some adjustment errors remain. As Fig. 5.7 demonstrates, the mean of distance measurements in face left and face right is free of errors caused by the misalignment.

It should be noted that any reflectorless distance meter, with a misalignment between the line of sight and the EDM beam, will exhibit similar errors. Measuring in two faces to an target at an angle of incidence of say 45° can, therefore, be used to check for possible misalignments of the EDM beam. Rotations of the target about horizontal and vertical axes should be checked.

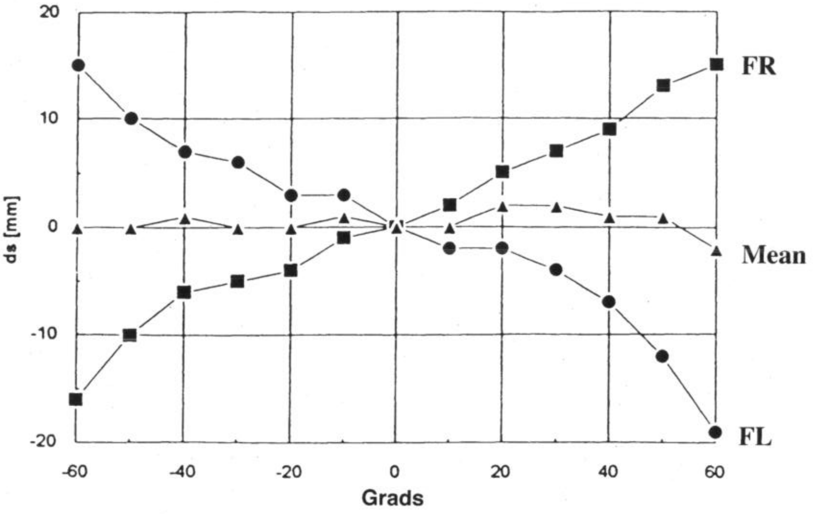

Figure 5.7: Errors in distance for reflectorless distance measurements with misaligned optics, measured for a Zeiss Reg Elta RL (after Köhler 1994)

5.1.6 Measurement to Corners

The fundamental problem of reflectorless EDM measurements to corners and similar shapes has already been outlined in Fig. 5.2. Since corners and other edges are one-dimensional features, the two-dimensional footprint of the EDM beam will (usually) not deliver the correct distance. Unfortunately, it is very often the edges and corners of structures and objects that are of interest to surveyors. Because of that, most manufacturers provide on-board software to remedy the problem. The solutions provided rely usually on the fact that corners are at the intersection of two planes or three planes. All one has to do is to measure in reflectorless mode to three (arbitrary and unmarked) points on one of the planes and then point to the corner. The on-board software will calculate the equation of the plane and, then, the slope distance (or coordinates) of the corner from the measured direction and zenith angle to the corner. No distance is measured to the actual corner.

The accuracy of the special measurements to corners depends on the flatness and size of the plane used for the corner definition and on where the three points are selected in relation to the corner. For example, the corner between two vertical walls of a room, halfway up the wall, is best determined by measuring to two points near the (two-dimensional) corner on the wall, one close to the floor and the other close to the ceiling, and a third point at the other end of the same wall. The accuracy of a corner determination can easily be checked by doing a second measurement using the second wall. Such a check is only useful if the corner point is (temporarily) marked.

On close range, the footprint of the reflectorless beam may be sufficiently small to allow direct measurements to corners. Small errors (see Fig. 5.2) will remain but may not be worse than the assumed flatness of walls or the verticality of wall corners.

5.1.7 Some Notes on the Use of Reflectorless EDM

(1) The target must be larger than the footprint of the reflectorless EDM beam. If the instrument used features a visible reflectorless EDM beam, note that the footprint used by the actual distance measurement may be (slightly) larger than the illuminated area seen by the eye.

(2) Avoid measuring to two targets at the same time. Make sure that there are no objects (e.g. leaves) in the beam between the instrument and the target. Make sure that there are no other reflecting surfaces behind the target.

(3) Since the knowledge of the actual beam width at different distances is essential, the user should establish the beam width using the test outlined in Section 5.1.2 and illustrated in Figs. 5.4.

(4) With instruments that do not employ a visible beam for the reflectorless measurements, it is essential that the user checks that the EDM beam is colinear with the line of sight of the instrument. Use the tests outlined in Sections 5.1.2 or 5.1.5 and illustrated in Figs. 5.4 or 5.7.

(5) Flat targets at 90° to the line of sight (at 0° angle of incidence) are preferred.

(6) Measure in two faces if there is any doubt on the accuracy of the alignment of the reflectorless EDM beam onto the telescope's line of sight.

(7) Do not measure to materials that give erroneous results, such as glass, acrylic glass, styrofoam, glazed ceramic tiles. Mirror like surfaces may deflect the beam to another surface (see Fig. 5.1).

(8) The range of reflectorless EDM is reduced when the incidence angle is large, the target is wet (see Item (7) above) and the target is in full sun light. Shading helps in the latter case.

(9) The range of all reflectorless EDM instruments can be increased considerably by measuring at night and by using cooperative rather than non-cooperative targets. Self-adhesive reflective sheeting targets are available from most manufacturers, in varying sizes (e.g. 20 by 20 mm, 40 by 40 mm, 60 by 60 mm).

(10) The additive constants of the normal distance meter and the reflectorless distance meter in the same instrument may not be the same. Calibrate the instrument separately for the two modes of operation, with the respective targets or reflectors.

(11) Non-cooperative ('natural') targets have a nil reflector constant. Most glass prism reflectors used in normal EDM have a substantial reflector constant. Never forget to change the reflector constant setting when switching between normal and reflectorless EDM. Frequent switching between reflectorless and normal EDM requires great discipline on the part of the operator.

(12) Reflectorless EDM to the normal glass prism reflectors is possible. The maximum measurable distance is typically double that of normal EDM. For example, for the Leica TPS1100 series, the reflectorless range is 5000 m (with slightly reduced precision) whereas the normal EDM range is 3000 m. For the Nikon NPL-350, the range increases from 2000 m to 5000 m.

Manufacturers may specify a minimum distance that must be observed when using glass prisms with reflectorless instruments (e.g. 10 m for Nikon NPL-350). Because of the high power of reflectorless instruments, glass prisms may reflect the EDM beam at points and surfaces other than the faces of the glass cube. Nikon suggests not to use prisms with scratches, a chipped apex or with thick edges for reflectorless EDM. Since the front surface of the prism can also reflect the beam, Nikon further suggests to mispoint the prism slightly for close range reflectorless measurements to glass prism reflectors. This is good advice for all reflectorless instruments.

(13) Reflectorless EDM through traffic (pedestrians or vehicles) can cause problems. If the cars have a similar reflectance than the intended target, and are close to the target, distances may be measured to the car and not the target, or an average distance may be displayed. It is best to measure distances across roads (or footpaths) multiple times, and to start the measurements when the line if sight is (temporarily) clear.

Nikon warns of possible errors with their NPL-350 reflectorless pulse distance meter, if the car is within three metres of the intended target. Considering Fig. 5.5, it can be inferred that an average distance will result within the three metres (length of transmitted pulse) and that the shorter distance (to the car) might be measured if the car is more than three metres in front of the target.

(14) Check the safety rating of the (lasing) diodes used in reflectorless instruments. Class I (Class 1) instruments are safe for any reasonable usage. Reflectorless instruments using visible (lasing) diodes are usually labelled as Class II (Class 2). Do not look into the transmitter (front lens) at close range.

5.2 Laser Scanners

The laser scanners are a further development of the scanning electronic tacheometers discussed in Section 4.1.9. They are optimised for speed of measurement and fast data acquisition and are supported by sophisticated software packages that allow a conversion of the data to formats readable by computer aided design packages. As discussed previously, the laser beam transmitted by the scanner is deviated vertically and horizontally to scan the surrounding space in a grid pattern. The step interval can be selected. Some instruments can scan around the full circle (360°), others only in a limited sector. The distances are measured reflectorless using the pulse method.

Company	Callidus	Cyra(Leica)	Riegl	Mensi
Country	Germany	USA	Austria	France
Model	Callidus 1.1	Cyrax 2500	LMS-Z210	GS 100
Wawelength	NIR	green (532 nm)	NIR (905 nm)	???
Safety Classification	Class 1	Class 2	Class 1	Class 2
Distance Range	0.15 - 150 m	1.5 - 100 m	2 - 350 m	2 - 100 m
Accuracy/Distance	±5 mm/32 m	±4 mm/50 m	±25 mm/200 m	±1mm/3-6 mm
Angle accuracy	±61"	±25"	±130"	6.6"
Hor Field of View	360°	40°	340°	360°
Vert Field of View	±90° & −60°	40°	80°	60°
Time for Scan	4-9 min	10 min	2 min	1000 points/s
Intensity Recording	no	yes	yes	yes
Camera	CCD	video	RGB channel	video
Can be centred	yes	no	yes	yes
Can be levelled	yes	no	yes	yes
Dimensions (mm)	460x300x300	400x330x430	435x210x210	320x420x280
Weight	13 kg	28.2 kg	13 kg	13 kg

Table 5.1: Major technical data of a selection of laser scanners

Table 5.1 gives some technical data of a small selection of laser scanning systems. Three systems are depicted in Fig. 5.8. The range and accuracy specifications of these and other systems vary greatly, from ±0.5 mm at 5 m to ±50 mm at 1000 m at the time of writing. Some systems employ infrared lasers, others visible ones. The latter have more stringent safety requirements. Of the systems listed in Table 5.1, the Mensi GS100 seems to be the most precise in its 'high precision' mode. Three of the four systems described in Table 5.1 (and all systems shown in Fig. 5.8) can be centred and levelled like conventional surveying equipment.

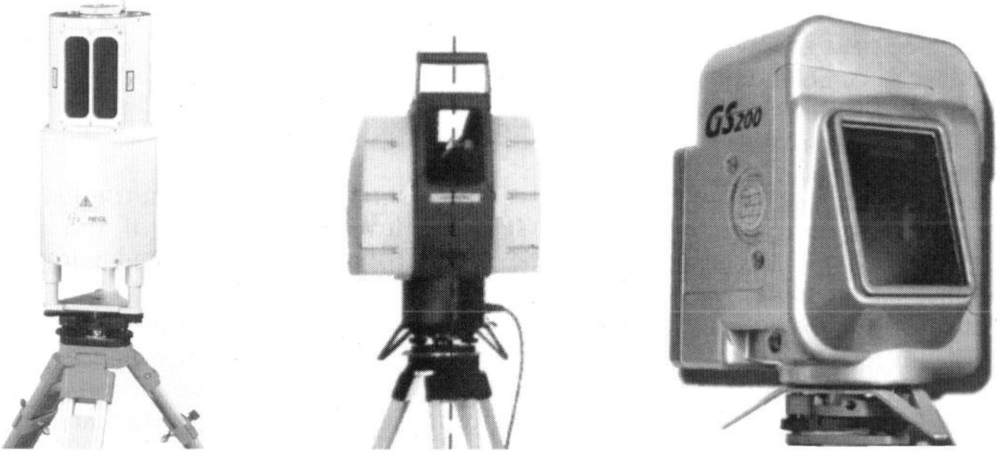

Figure 5.8: Riegl LMS-Z210 '3D-imaging-sensor' on the left, Leica HDS3000 'high definition surveying system' in the middle and Mensi GS200 'long range scanner' on the right.

The typical results of a laser scan are the x, y, z coordinates of the large number of points 'seen' by the scanner. The precision of these coordinates depends on the precision of the angular measurements (directions, zenith angles) and the precision of the distance measurements. Some laser scanners also record the intensity (signal strength) of the returned laser beam. This permits to generate something that resembles a photograph. A significant computing effort is required to convert the large number of measured coordinates ('point clouds') to useful plans and computed aided design (CAD) files. Some systems do record photographic images simultaneously. These can be used to render three-dimensional computer images, that may be derived, with realistic colours.

Since laser scanners mostly measure to non-cooperative targets, namely the natural and man-made features around the scanner, the problems associated with reflectorless EDM must be considered when interpreting laser scanner data. Refer to the discussions in Section 5.1 for details. In reflectorless EDM, an operator manually points to a passive target and, thus, can select suitable surfaces to a certain degree; in contrast, laser scanners measure in a grid pattern to whatever surface is found in the beam. Apart from being faster and less precise, laser scanners have typically a much narrower beam than the reflectorless EDM instruments. Some potential problems are reduced because of that, some others may be created.

Liechti & Harvey (2001) have investigated a few aspects of the Cyrax 2400 laser scanner, which features a beam diameter of 5-6 mm. They found, for example, that measurements to glass prism reflectors reveal significant errors in the distances measured, wire mesh fences in the scene (3 m in front of instrument) are being picked up (even though the wire is thinner than the laser beam) and water on surfaces can change the distance reading by up to 3 mm at 53 m.

5.3 Motorised Levelling

Motorised precision levelling is not a new technique. It has been used in the 1970s to remeasure the entire first order levelling network of the then German Democratic Republic. The method was greatly assisted with the development of the Zeiss Ni002 automatic level with its swivel eyepiece; with this instrument, the foresight and the backsight can be observed without any movement of the observer. A total of three vehicles is needed. The (invar) staffs are attached to cars and the instrument is mounted on a car. This method uses the same staffs as normal and normal sighting distances. In (East) Germany, the achieved one kilometre precision was ±0.3–0.4 mm Please note that this precision is much better than the precision specified for Class A (Order 1) vertical surveys in Australia.

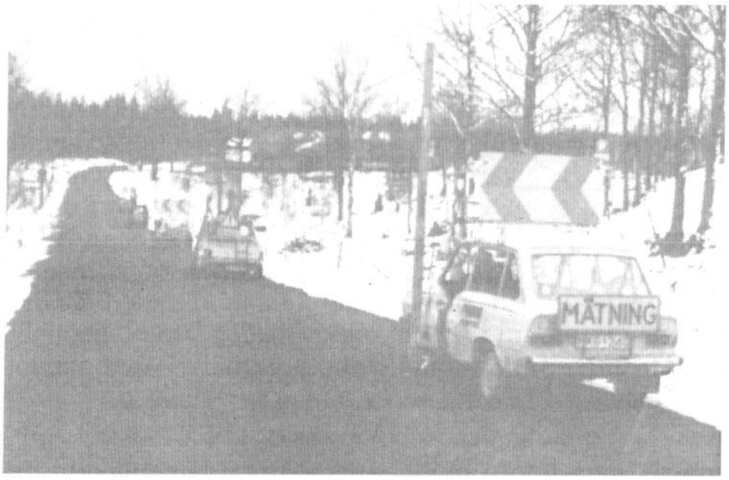

Figure 5.9: Motorised precise levelling in Sweden (after Becker 1979). In this case, two instrument vehicles and three staff vehicles can be seen.

The method was further developed by the National Survey of Sweden for the re-survey of 40 000 km of double run levelling lines, beginning in 1980.

The Swedish equipment is based on cars with automatic transmission, a one-ton truck for the instrument, an electronic calculator for the data recording, an instrument tripod with extra-long legs and changeable tripod shoes (like spikes and plates), a quick level tribrach, adjustable staff supports fitting a variety of vehicles as well as wind and rain protection. In Sweden, precise levelling is carried out between April and September and between 07.00 h and 17.00 h in all weather. Sighting distances do not exceed 50 m. The measurement tolerance (maximum permissible difference between the forward and the backward run height difference) is $2\ \text{mm}\ \sqrt{L}$, with L in km. It is only necessary to leave the vehicle when connecting to benchmarks. The Zeiss Ni002 was the preferred instrument for motorised levelling before the introduction of precise digital levels. The Swedish equipment is shown in Figs. 5.9 and 5.10 (Becker 1979).

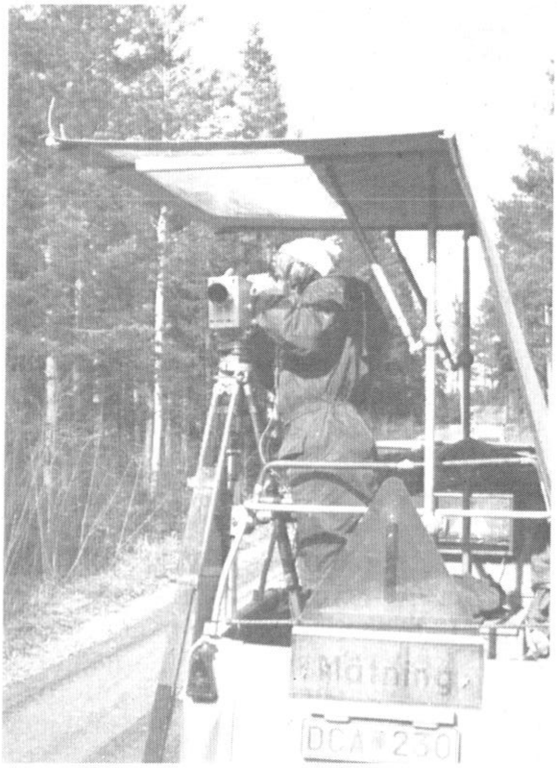

Figure 5.10: Instrument vehicle in motorised precise levelling in Sweden with Zeiss Ni002 (after Becker 1979)

The advantages of motorised levelling can be summarised as follows:

- better progress rates (quicker)
- less dangerous
- higher line of sight reduces refraction (avoid car exhausts)
- better stability of staff
- immediate on-board check of data

- less tedious than conventional levelling on foot

	Distance Levelled (km)	s_{1km} (mm √L)
East Germany (double run)	~5000	±0.38
Sweden (double run)	>2000	±0.32–0.55
Denmark (double run)	>46	±0.44
The Netherlands (double run)	>540	±0.58
West Germany (double run)	201	±0.24

Table 5.2: Precision of motorised double-run precise levelling (s_{1km}) over 1 km with classical automatic levels, usually of the Zeiss Ni002 type. The precision values refer to the mean of double-runs. (L in kilometres)

The Swedish National Survey did investigate the progress rates with different types of precise levelling. The following progress rates were found: on foot: 3.4 km / day, by bicycle: 5.5 km / day; fully motorised: 11.5 km / day. In Sweden, the cost of motorised levelling was about 50% of the cost of foot levelling, including the costs of setting the motorised system up. Initial tests in (West) Germany, with sighting distances of 25 m - 40 m and invar staffs attached to motor bikes, showed progress rates of 7.4 km / day when using the BFFB observation routine, and 9.4 km / day when using the BBFF routine. Table 5.2 gives a summary of the precision achieved with motorised levelling in various countries.

Since 1990, the digital levelling instruments of the types Wild NA2000 and NA3000 were also tested by the National Survey of Sweden for their suitability in motorised levelling. Because of the lack of a swivel eyepiece as a standard item, the broken eyepieces known from theodolites had to be used. The resulting image rotations were not found to be suitable for a production environment. Becker et al. (1994) also noted the problems caused by

- the fact that 70% of the measurement sector must be free (particularly when connecting to bench marks),
- measurements into a low sun (a longer sun shade did reduce but not eliminate the problem),
- shadow patterns on the staff from bare trees (measurements along some roads through forests were simply not possible),
- poor light (in spring and autumn, measurements had to stop half an hour earlier than with optical instruments due to poor lighting; time was also lost on heavily overcast days),
- rain, smoke and heat (measurements take longer to make),

- the increased workload of the observer (normally, in motorised levelling, the driver of the instrument vehicle carries out the bookkeeping).

A poor temperature stability of the line of sights of the digital levels NA2000 and NA3000 were reported by Becker et al. (1994). These authors also noted the longer set-up time required per station (in comparison to the Ni002) and the increase in the number of set-ups per kilometre (5% - 20%). They suggested a standard swivel eyepiece for the NA3000 and the removal of the lowest 30 cm of graduations on the bar code staff (to reduce the effects of differential refraction). Refer to Section 3.1.5.1 for more information on the errors of bar code levels.

5.4 Laser Profiling and Laser Altimeters

5.4.1 Laser Profiling on Land

Laser profilers for aircraft and helicopter operations have been available for a number of years. At the centre of a laser profiler is a high power pulse distance meter, usually based on a Nd:YAG laser operating at an infrared wavelength of 1064 nm. Such systems are available, for example, from:

- Azimuth, 13 Park Drive, Westford MA 01886, USA
- LH Systems / Leica Geosystems GIS & Mapping LLC, 2801 Burford Highway, Atlanta Georgia 30329, USA
- Measurements Technology and Systems, Avco Everett Research Laboratory Inc., Everett Massachusetts 02149, USA
- Optech Systems, 701 Petrolia Road, Downsville Ontario, Canada M3J 2N6

The early Avco Airborne Laser Mapping System operated at 5, 10 or 20 pulses per second. This translated to one measurement per 75 m on the ground at a ground speed of 275 km/h and at a pulse repetition rate of 10 Hz. The typical operating altitude was 350-600 m above ground. The vertical mapping accuracy was specified as ±1.4 m on land and ±0.15 m on water. The roughness of the ground coverage has, naturally, an effect on the accuracy. Most systems are, however, able to process multiple pulse returns, such as those for the shortest and the longest distance for each pulse. Otherwise, only an average distance in the footprint area is measured. The horizontal mapping accuracy was based on a microwave distance measuring system; the specified accuracy was ±3.2 m. The Avco system required 28 V DC supply and required 1.1 kW to run. The dimensions and weight were 1470 x 910 x 790 mm and 180 kg, respectively.

Besides the pulse distance meter, an aircraft positioning system and an attitude measurement system are required as well as powerful data recording facilities. Today, the former is typically a GPS receiver and provides the coordinates of the aircraft or helicopter. The attitude angles (roll-angle and pitch-angle) are required to transfer the coordinates of the GPS antenna to the laser system and to measure the deviation of the laser beam from the plumbline. The attitude information can be provided by an inertial navigation or reference system or, possibly, by an array of three to four GPS receivers. Rather than measure the attitude angles, the laser transmitter might be installed on a stabilised (gimballed) platform that is kept in a strictly levelled state irrespective of aircraft movements. This solution also requires inertial sensors for the platform stabilisation. Some laser profilers are able to scan the terrain (or water) sideways. In this case, a scanning mirror is required to deflect the laser beam off the vertical. The laser beam deflection introduced must be measured and recorded to be able to compute the ground position.

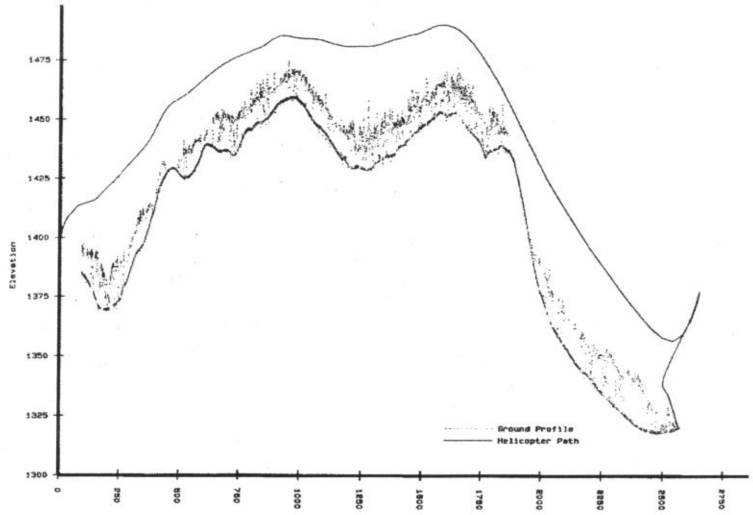

Figure 5.11: Longitudinal profile surveyed with a helicopter-borne laser profiler. The graduation intervals on the vertical and the horizontal axes are 25 m and 250 m, respectively. Flight time: 270 seconds. The continuous upper curve is the helicopter path and includes the take-off on the right hand side. The lower curve shows the ground level and the dots between ground and helicopter the upper foliage level. (from Nortech Surveys brochure)

The ALS50 Airborne Laser Scanner by LH Systems (Leica Geosystems) achieves rates of 52000 pulses per second, (lateral) scan rates of 70 Hz and a swath width of 75°. Flying altitudes above ground from 500 m to 4000 m are possible. The position of the helicopter or aircraft is determined by GPS and the roll, pitch and heading (azimuth) data by an inertial measuring system. The ALS50 can record up to three times per pulse; this permits to measure to tree tops and ground heights, for example. The system also records the signal intensity. This information can then be displayed in a photograph like form. The height accuracy is listed as 0.13 to 0.30 m and the horizontal accuracy as 0.11 to 0.46 m, depending on flying height and scan angle.

Figure 5.11 shows a terrain profile mapped with a helicopter based laser scanner operated by Nortech Surveys (of Calgary, Canada). The figure shows that the ground level below trees can be recovered in certain circumstances, as the system records the shortest and the longest distance measured from each pulse. The laser profiler used by Nortech Surveys has a high pulse repetition rate (4000 Hz). It is ideally suited for corridor surveys along proposed pipelines, transmission lines, roads or microwave links, either in a single profile or multiple profiles, 25 m apart. Areas are surveyed by surveying parallel profiles. Nortech Surveys quote progress rates of 160 km per day.

5.4.2 Depth Sounders

A number of agencies have developed laser profilers for hydrographic applications. These systems make use of the fact that the Nd:YAG laser's fundamental frequency of 1064 nm can be frequency doubled to the green wavelength of 532 nm. The infrared beam is reflected by the water. The green beam does penetrate the water (unless at a right angle to it) and is reflected at the sea floor. The difference of the two distances gives the depth of the water.

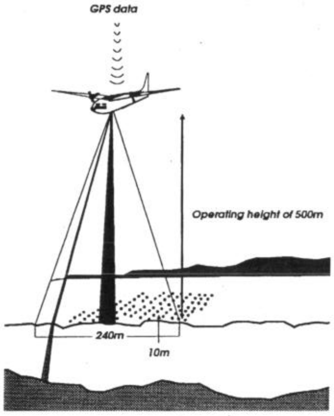

Figure 5.12: Principle of the Laser Airborne Depths Sounder (LADS) of the Royal Australian Navy.

A Scanning Hydrographic Operational Airborne Lidar Survey (SHOALS) system was developed by Optech in Canada for the US Army Corps of Engineers. It has a normal operating altitude of 200 m. The width of the swath of soundings is about 120 m, with 200 soundings per second. Maximum depths are 25 to 30 m in moderately clear waters. This can be doubled at night. The off-nadir scan angle is 15 to 25 degrees. The rotation rate can be varied between 0 and 5 Hz. The depth accuracy was estimated as ±0.3 m and the horizontal accuracy with 4 to 5 m.

A similar system was developed by BHP Engineering and Vision Systems for the Australian Navy. The Australian Laser Airborne Depth Sounder (LADS) is in operation with the Hydrographic service of the Royal Australian Navy. It is installed on board of a Fokker F27-200 Frienship aircraft and operates at altitudes of about 500 m. It can map the sea floor at a rate of 50 km^2 per hour. As the Fig. 5.12 shows, the swath is 240 m wide and individual soundings are spaced at 10 m intervals. Depths of 50 m can be measured. The aircraft flies at a speed of 270 km/h and can stay 7 hours in the air. Like the SHOALS system, it uses a Nd:YAG laser at two frequencies, with a pulse rate of 168 Hz, a pulse power of 1 MW and a pulse length of 5 ns. The return pulses are digitised at 2 ns intervals, which gives a resolution of 0.2 m in height. The positional accuracy is given as ±15 m; it includes the errors of the GPS positioning and the inaccuracy of the platform stabilisation. The laser platform is stabilised for roll, pitch and yaw (azimuth) with input from an inertial measurement system. Orientation data are also provided by a fluxgate compass. Data are stored on tape and post-processed on the ground, immediately after the return of the aircraft.

5.5 Single Beam Laser Trackers for Metrology

At one time, there were triple-beam and single-beam laser tracking interferometers on the market. Because of the more universal appeal and wider usage, only the latter type is discussed here. The single-beam laser tracking interferometers, of the types shown in Fig. 5.13, consist of a single laser interferometer, the beam of which is deflected in three-dimensions (about two axes, one vertical and the other horizontal) by a tilting and rotating mirror. This tracker has angular encoders fitted to both angular servo motors and, thus, delivers directions and zenith angle information very much like electronic theodolites. Together with the (slope) distances, three-dimensional coordinates can be computed for the target points. A feed-back loop ensures that the laser beam is kept on the (moving) reflector.

Figure 5.14 shows the measured quantities (on the right) and the electro-optical components of one device (on the left). The left part of Fig. 5.14 shows the interferometer and the servo loops for the servo mirror. The "Detector" near the "Interferometer's Beam Splitter" records the interference fringes between the transmitted and the returned laser beam. Whenever the reflector moves, constructive and destructive interference alternates. The number of bright fringes is counted by the "Position Electronics" and processed into the change of distance per time interval. A second beam splitter (between the "Interferometer's Beam Splitter" and the "Servo Mirror") deviates 15% of the returned laser light onto the "Position Sensor".

The latter is a position sensitive diode, which measures the position of the light spot. If the light moves out of the centre, the appropriate servo motors are switched on until the light spot is centred again. This correction mechanism occurs at a rate of 1 kHz in the case of the CMS3000 (Ruland 1993).

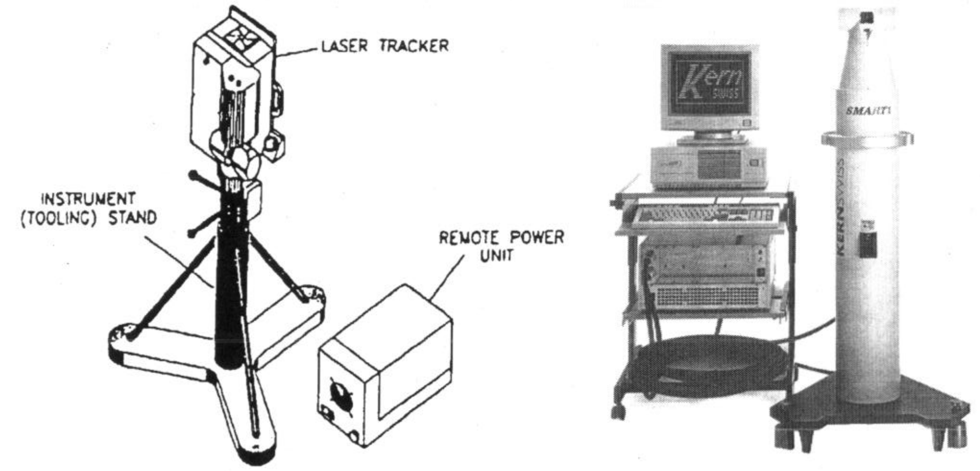

Figure 5.13: Chesapeake CMS3000 "Coordinate Measuring System" (on left, after Ruland 1993) and Leica "Smart 310" 3-D Laser Tracker (on right, after Kern/Leica 1990)

The interferometer of the laser tracker measures only changes in distance, not the absolute distance. To refer the distance to the tracker and its coordinate system, the reflector has to be placed on a reference mounting on the tracker or, alternatively, on two target points with a known distance in between (baseline). Another property of the laser interferometers (see Rüeger 1990 1996 for reference) is that the change of distance between two positions of the reflector can only be measured if the beam is constantly pointed to the reflector and if no beam interruptions occur. In the case of a beam interruption, the measurements have to start again from the reference point(s).

The Kern (now Leica) SMART 310 is one of the systems shown in Fig. 5.13. Its tracking range is ±240° horizontally and ±45° vertically. The working range is specified as 0.2 m to 25 m. The maximum tracking speed (at 2 m to 5 m from the tracker) is 2 m/s across the beam (horizontally and vertically) and 0.6 m/s along the beam. The resolution of the angular encoders is 0.2 mgon or 0.7". This corresponds to 0.03 mm per 10 m. The distance resolution is specified as ±0.16 μm. The coordinate accuracy in the object space depends on the tracking speed. It varies between ±20 ppm and ±50 ppm or ±20 μm/m to ±50 μm/m (three times standard deviation). The data are collected at a rate of 1000 points per second. The tracker weighs 30 kg and the control unit (under the personal computer in Figure 5.8) 15.5 kg.

The reflectors are mounted in precision fittings (usually a stainless steel sphere) for the attachment to reference and target points. This permits a centring accuracy that is compatible with the overall accuracy of the system. The reflectors are either of the corner cube type (like in EDM) or of the cat's-eye variety. The latter allows larger incidence angles.

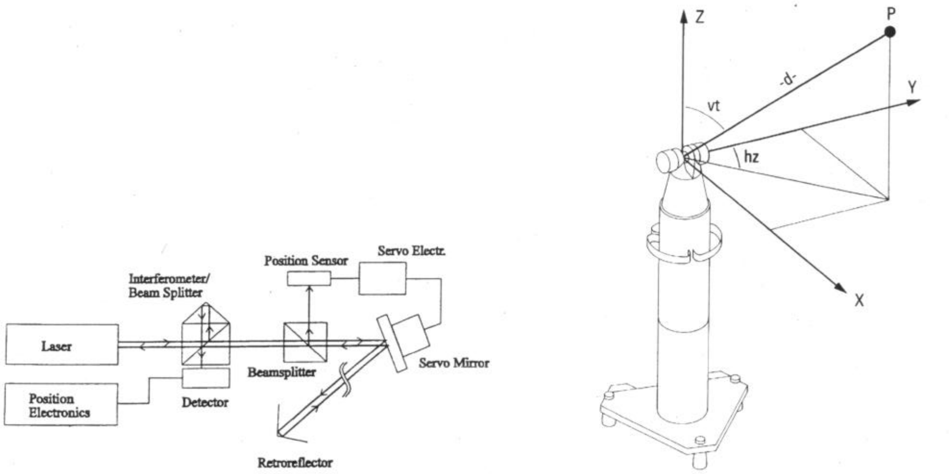

Figure 5.14: Block diagram of the servo-controlled tracking system of the Chesapeake CMS3000 Laser Tracker (after Ruland 1993) and the basic optical-mechanical lay-out of the Leica "Smart 310" 3-D Laser Tracker (after Leica 1993)

The performance of the Chesapeake CMS3000 laser tracker was evaluated comprehensively by Ruland (1993). Ruland recommends that all measurements are taken in "two faces" (for the same reasons as with theodolites) and that measurements are taken at different reflector orientations (about the reflector axis). The latter procedure will reduce the errors due to the eccentricities between the reflectors and their mounting spheres.

Since 1997, the two major suppliers of single beam laser trackers offer special models that feature an 'absolute distance meter' (ADM) besides the usual laser interferometer ('relative distance meter'). The absolute distance meters referred to here are nothing else than high precision electronic distance measurement (EDM) instruments. The Leica LTD-500 3D-Laser Tracking System (a further development of the SMART 310 shown in Fig. 5.14) uses a distance meter that operates at an infrared wavelength of 780 nm, has an accuracy of ±0.025 mm (1 σ), a resolution of 1 μm and a range of 2 - 35 m. The firm SMX supplies the Tracker 4500 with ADM. Its EDM section uses a HeNe gas laser, has a resolution of 0.5 μm, an accuracy of ±20 μm at close range and an accuracy of ±1.1 ppm at 'long' range (to 35 m).

The single-axis laser trackers are used for robot calibration, 'teach-in' of manufacturing robots, automatic tracking of robots, dynamic measurements of large objects, digitising of surfaces, inspection of antennas, quality control on master-gauges, jigs and tools in mechanical engineering, alignment and assembly of machines and for the calibration of coordinate measuring machines (CMM) and numerically controlled (NC) machine tools, to name just a few examples. The 3-D laser trackers compete with the multi-theodolite 3-D measurement systems. The trackers require that the measuring points are, temporarily, marked with a reflector. If the interferometer is being used, the reflector must be moved between all measurement points without beam interruptions. Clearly, they are most appropriate in applications where the item to be measured can move the reflector automatically, under computer control.

5.6 Reflectorless Single Station Laser Measurement Systems

The previously discussed single beam laser tracking interferometers and the single beam laser trackers with absolute distance measurement require that reflectors be used to mark the points on an object. The Coherent Laser Radar (CLR) MV100B by Metric Vision, and the Leica Laser Radar LR200 now allow to scan quickly and precisely large surfaces without reflectors. Since no reflectors need to be attached to the measured objects, sensitive and non-accessible surfaces can be measured. The Leica system measures three points per second in the most accurate ('metrology') mode. In the 'pseudo vision' mode, measuring rates of up to 1000 points per second are possible. Different models are available for maximum ranges of 8 m, 24 m and 60 m.

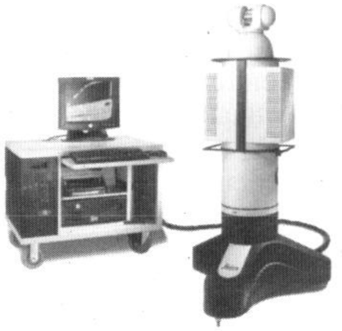

Figure 5.15: Leica Laser Radar System LR200 (after Leica 2002), similar to Metric Vision MV100B Coherent Laser Radar (CLR)

Leica specifies the coordinate accuracy (1 σ) for 'tooling ball targets' (highly polished perfect steel spheres), as ±0.05 mm for distances below 10 m and ±5 ppm for distances longer than 10 m. Like with all other reflectorless measuring systems (refer to Sections 4.1.8, 4.1.9, 5.1, 5.2), the quality of the data will depend on the type, reflectance, roughness and orientation of the surfaces. The usefulness of such a system depends on the sophistication of the software used to convert the measured data ('point clouds') into CAD plans and files and into coordinates of specially marked points (e.g. with tooling ball targets).

The distance meter of the Leica LR200 operates at a wavelength of 1.55 μm. A visible pointing laser is also available. According to Slotwinski et al. (1989) the distance meter uses a frequency modulated continuous wave (FMCW) laser, where the optical frequency of the laser diode is changed in a saw-tooth ('chirp') pattern by thermal tuning. The distance is derived from the difference in the frequencies of the transmitted and received IR wavelengths.

5.7 Exercises

5.1 Discuss the pros and cons of two methods of checking the alignment between the beam of a reflectorless EDM instrument and the line of sight of the telescope.

5.2 Discuss the different behaviours of reflectorless phase measuring and pulse distance meters, when measuring to two targets at the same time.

5.3 Explain how to determine the horizontal and vertical beams width of an infrared reflectorless distance meter.

5.4 Discuss the pros and cons of motorised precision levelling in comparison with precision levelling on foot.

5.5 Your firm is tendering for the coordination of the centre line of a two-lane road between two cities, 500 km apart. You are evaluating a laser profiling system.
(a) Discuss the pros and cons of this system for this particular application.
(b) What accuracy could be achieved?
(c) What surveying infrastructure would you have to put in place?
(d) Approximately, how long would the actual field survey take with the laser profiler?

5.6 (a) Determine how many soundings (at 10 m spacing) are made by the Australian Laser Airborne Depth Sounder in one second of time if flying at 75 m/s and at a height of 500 m above sea level.
(b) Hydrographic survey ships, in comparison, travel at a speed of, say, 10 m/s but take continuous soundings to the sea floor. How much longer does a survey by ship take, if it is assumed that the hydrographic survey vessel runs longitudinal profiles at 10 m spacing?
(c) If the same area were mapped by LADS and by ship (as discussed in

(b)), which survey would provide more information on the sea floor?

5.7 Discuss the pros and cons of a 3-D laser tracking interferometer, in comparison with a multi-theodolite 3-D measuring system, for
(a) the dimensional measurement of a steel structure of 1 x 1 x 1 m,
(b) the calibration of the movements of a manufacturing robot.

5.8 What are the similarities of (and differences between) laser scanners and reflectorless single station laser measurement systems?

References

Ailisto, H., Kostamovaara, J., Lammasniemi, J., Myllylä, R. 1996. Lidar Measures Up. Photonics Spectra, March 1996, 96-100.

Angus-Leppan, P. V. 1984. Refraction in Geodetic Levelling. In: Brunner, F. K. (ed.), Geodetic Refraction - Effects of Electromagnetic Wave Propagation Through the Atmosphere, Springer-Verlag, Berlin, 163-180.

Bannister, A., Raymond, S., Baker, R. 1992. Surveying. 6th edition. Longman Scientific & Technical, Harlow, England, ISBN 0-470-21845-2, 482 + XII pages

Becker, J.-M. 1979. Heutige Messverfahren beim Nivellement – Erfahrungen mit dem motorisierten Nivellement in Sweden (Current Measurement Procedures in Levelling – Experiences with the Motorised Levelling in Sweden). In: Vorträge zum Seminar Präzisionsnivellement, 7 Dec 1978, Aachen, Germany. Veröffentlichung des Geodätischen Instituts der Rheinisch-Westfälischen Technischen Hochschule Aachen, No. 26, 2-18

Becker, J.-M., Andersson, B., Ericksson, P.-O., Nordquist, A. A new Generation of Levelling Instruments: NA2000 and NA3000. In: Papers, Commission 5, 20th FIG Congress, Melbourne, Australia, 5-12 March 1994, TS509.1/1 - TS509.1/11

BEI. 1987. Optical Encoder Design Guide. Doc. No. 924-08027-002, BEI Motion Systems Company, Industrial Encoder Division, 7230 Hollister Avenue, Goleta California 93117-2819, USA, 15 pages.

Buchmann, P., Scherer, M. 1994. Building Surveys with a Computer Controlled 3D-Laser Scanner. Der Vermessungs-Ingenieur, 45(5): 210-215 (in German)

Bureau of Meteorology. 1990. The Aneroid Barometer ... and how to use it. Information leaflet, 2 pages.

Carr, B. 1992. Using Laser Technology for Forestry and Engineering Applications. The COMPILER, 10(4): 5-15.

Cooper, M. A. R. 1982. Modern Theodolites and Levels. 2nd edition. Granada, London, ISBN 0-246-11502-5, 258 + XI pages

Deumlich, F. 1982. Surveying Instruments. Walter de Gruyter, Berlin-New York, ISBN 3-11-007765-5, 316 pages

Feist, W., Gürtler, K., Marold, T., Rosenkranz, H. 1995. Die neuen Digitalnivelliere DiNi10 und DiNi20 (The new Digital Levels DiNi10 and DiNi20). Vermessungswesen - Raumordnung (VR), 57: 65-89 (in German)

Finlayson, D. M. 1973. Isomagnetic Maps of the Australian Region for Epoch 1970.0. Report 159, Bureau of Mineral Resources, Geology and Geophysics. Australian Government Publishing Service, Canberra.

George, J. 1984. Large Area Photodiode yields Pinpoint Accuracy. Eureka – Innovative Engineering Design, July/August 1984, reprint, 3 pages.

Geotronics AB. 1994. Geodimeter System 600. Publ. No. 571 710 351, 8 pages.

Geotronics AB. 1989. Geodimeter 140 SMS Surveying Robot – Slope Monitoring System Description. Publ. No. Geo - 0024, 17 pages

Haag, R., Bayer, G., Zimmermann, M., Scherrer, R. 1997. Surveying with the automatic Fine-Pointing of the Leica TCA1800. Vermessung-Photogrammetrie-Kulturtechnik (VPK), 95(7): 466-471 (in German)

Hennes, M., Witte, B. 1991. Genauigkeitsuntersuchungen an elektronischen Tachymetern und Theodoliten der Baureihe E von Zeiss (Investigation of the Accuracy of Electronic Tacheometers and Theodolites of the E-Series by Zeiss). Vermessungswesen - Raumordnung (VR), 53(1): 30-45

Höglund, R. 1994. Practical Use of an One Man Total Station. Proc., Commission 5, 20th FIG Congress, Melbourne, Australia, 5-12 March 1994, TS502.1/1 - TS502.1/9

Ibeo. 1991. Ladar: 2D, 2D Linear, 3D. IBEO Ingnieurbüro für Elektronik und Optik, Fahrenkrön 125, D-2000 Hamburg 71, Germany, 4 pages

Ibeo. 1986. Survey Instruments - Laser Technology - Sensors and Systems: Sensors for Industrial Applications. IBEO Ingnieurbüro für Elektronik und Optik, Fahrenkrön 125, D-2000 Hamburg 71, Germany, 28 pages

Ingensand, H. 2001. Systematic Influences on Practical Measurements with (Electronic) Tacheometers and Digital Levels. In: Heister, H., Staiger, R. (eds.) Quality Management in Geodetic Metrology, 54th DVW Seminar, Fulda, 19/20 November 2001, Schriftenreihe des DVW, Vol. 42/2001, Verlag Konrad Wittwer, Stuttgart, pp. 120-135

Ingensand, H. 1993. Wild NA2000 - NA3000: Technical Paper Digital Levels. Leica AG, CH-9435 Heerbrugg Switzerland, Publ. No. G1 220e-V.93, 14 pages.

Ingensand, H. 1990. Das Wild NA2000 - Das erste Digitale Nivellier der Welt. Allgemeine Vermessungs-Nachrichten (AVN), 97(6): 201-210 (in German)

Kahmen, H., Faig, W. 1988. Surveying. Walter de Gruyter, Berlin-New York, 578 +xvii pages, ISBN 3-11-008303-5

Kern/Leica. 1990. Smart 310 – Measuring Movements. Publication 206e – 4.90 FA/8TE, Kern & Co. Ltd. (now Leica AG, Photogrammetry and Metrology, CH-5035 Unterentfelden, Switzerland)

Kogoj, D. 2001. Performance of Electronic Distance Meters in Reflectorless Distance Measurement. Allgemeine Vermessungs-Nachrichten (AVN), 108(5): 186-190 (in German)

Köhler, M. 1994. Some Theoretical Foundations for Reflectorless Measurements with the Zeiss Tacheometer Rec Elta RL. Der Vermessungsingenieur, 45(5): 216-221 (in German)

Lardelli, A. 1988. ECDS2 – A Mobile 3D-Measuring System – The New Standard for Metrology and Quality Control. Publ. No. 261e - 11.88, Kern & Co Ltd (now: Leica AG, Photogrammetry and Metrology, CH-5035 Unterentfelden, Switzerland)

Laser Alignment Inc. 1986. Laser Beacon 5000 Series for Construction and Machine Control. Publ. No. LAI 2002 (1-22-86), Laser Alignment Inc., 2850 Thornhills S.E., Grand Rapids Michigan 49546, USA, 6 pages

Lechner, W., Sudau, A., Wanzke, H. 1987. Motorisiertes Nivellement mit Digitalbarometern (Motorised Levelling with Digital Barometers). Zeitschrift für Vermessungswesen (ZfV), 112(8): 385-396

Leica. 2002. Leica Laser Radar System LR2002 - Non-contact 3D laser metrology advances factory automation. Publ. No. 727 644en, 1.2002, 4 pages.

Leica. 2001. RL-EDM Measurements. TPS News, 11/2001.

Leica. 1996. Wild NA2002 - NA3003: Technical Report, Digital Levels. Publ. No. G1 - 220 - 0en - IV.96 - VD, 14 pages

Leica. 1995. Leica Vector. Publication No. A1-300-0en-I.95-RVA, 4 pages.

Leica. 1993. Smart 310 – 3D Laser Tracker. Publication U1 206 e – V.93, Leica AG, Photogrammetry and Metrology, CH-5035 Unterentfelden, Switzerland

Leica. 1992. Mobile 3-D Metrology for Industry – Measurably Better. Publ. U1 241e - XI.92, Leica AG, Photogrammetry and Metrology, CH-5035 Unterentfelden, Switzerland

Leica Instruments. Pocket Laser Level – Perfect Dimensional Control. Leica Instruments Pty Ltd, 45 Epping Road, North Ryde NSW 2113, Australia, 4 pages.

Liechti, D. D., Harvey, B. R. 2002. An Investigation into the effects of Reflecting Surface Material Properties on Terrestrial Laser Scanner Measurements. Geomatics Research Australia, No. 76, 1-22.

Maurer, W., Schnädelbach, K. 1995. Laserinterferometry - Ten Years Experience in Calibrating Invar Levelling Staffs. Proc., First Int. Symp. Appl. Laser Techniques in Geodesy and Mine Surveying, Ljubljana, September 1995, 9 pages

Meier-Hirmer, B. Meier-Hirmer, R. 1996. Investigations of the Zeiss Tacheometer Rec Elta RL for Measurements to Structured Targets. Allgemeine Vermessungs-Nachrichten (AVN), 103(6): 233-240 (In German)

Moll, J. 1993. Low-Volume Roads Survey Laser. Publ. No. 9377 1202-SDTDC, Forest Service, Technology and Development Program, US Department of Agriculture, Washington DC 20250, February 1993, 11 pages

Möser, M. 2000. Close Range Distance Measurements. Zeitschrift für Vermessungswesen (ZfV), 125(11): 377-381

Noack, G. 1989. Problems with the Levelling of Rotating Laser Levels for Area Levelling of High Accuracy. Vermessungstechnik (VT), 37(2): 59-61 (in German)

Pallett, E. H. J. 1972. Aircraft Instruments – Principles and Applications. Pitman Publishing, London, ISBN 0-273-31747-4.

Pascoe, R. W., Rüeger, J. M. 1989. Testing a Precise Hand-Held Digital Barometer. The Australian Surveyor, 34(5): 502-513

Pike, J. M., Bargen, D. W. 1976. The NCAR Digital Barometer. Bulletin of the American Meteorological Society, 57(9): 1106-1111

Pike, J. M., Brock, F. V., Semmer, S. R. 1983. Integrated Sensors for PAM II. Paper 17.5, 5th Symposium on Meteorological Observations and Instrumentation, Toronto, Canada, pp. 326-333

Rizos, C. 1997. Principles and Practice of GPS Surveying. Monograph No. 17, School of Surveying and Spatial Information Systems, University of New South Wales, Sydney, Australia, 565 pages.

Rüeger, J. M. 2000. The Topcon DL-101C Digital Level. The Australian Surveyor, 45(2): 62-70 (Trans Tasman Surveyor, 1(3): 62-70).

Rüeger, J. M. 1996. Electronic Distance Measurement – An Introduction. 4th ed., Springer-Verlag, Berlin-Heidelberg-New York, xvii + 266 pages, ISBN 3-540-61159-2

Rüeger, J. M. 1995. EDM-Height Traversing Refraction Correction and Experiences. The Australian Surveyor, 40(4): 48-56.

Rüeger, J. M. 1995b. Surveying Practice, Education, Textbooks and Instrument Design in the Light of Changing Instrumentation. The Australian Surveyor, 40(3): 197-204

Rüeger, J. M. 1990. Electronic Distance Measurement – An Introduction. 3rd ed., Springer-Verlag, Berlin-Heidelberg-New York, xvii + 266 pages, ISBN 3-540-51523-2

Rüeger, J. M. 1978. Misalignment of EDM Reflectors and its Effects. The Australian Surveyor, 29(1): 28-36

Rüeger, J. M., Brunner, F. K. 1981. Practical Results of EDM-Height Traversing. The Australian Surveyor, 30(6): 363-372.

Ruland, R. E. 1993. The Chesapeake Laser Tracker in Industrial Metrology. Proceedings of the Third International Workshop on Accelerator Alignment, European Laboratory for Particle Physics (CERN), Geneva, Switzerland, 28 September to 1 October 1993, I/101–I/118

Schauerte, W. 1993. Digitalnivelliere des Typs Leica NA 3000 und ihr Einsatz beim Präzisionsnivellement (Digital Levels of the Type Leica NA3000 and their Use for Precise Levelling). BDVI-Forum, 19(3): 125-138

Schauerte, W. 1995. Effects of Special Errors on NA-3000 Measurements. BDVI-Forum, 21(1): 17-28 (in German)

Schwarz, W. 1985. Zur Ermittlung der integralen Temperatur der Atmosphäre mit Ultraschall für Refraktionsbestimmungen im Nahbereich (On the Evaluation of the Integral Temperature of the Atmosphere with Ultrasound for the Determination of Refraction on Close Range). No. 38 of the Veröffentlichungen des Geodätischen Instituts der Rheinisch-Westfälischen Technischen Hochschule Aachen, Germany, 136 pages (ISSN 0515-0574)

Slotwinski, A. R., Goodwin, F. E., Simonson, D. L. 1989. Utilizing GaAlAs Laser Diodes as a Source for Frequency Modulated Continuous Wave (FMCW) Coherent Laser Radars. In: Figueroa, L. (ed), Proc., Laser Diode Technology and Applications, 18-20 January 1989, Los Angeles, California, SPIE, Vol. 1043: 245-251

Sudau, A. 1994. Untersuchungen zur Leistungsfähigkeit elektronischer Druckaufnehmer für präzise barometrische Höhenmessungen (Investigations into the Performance of Electronic Pressure Sensors for Precise Barometric Heighting). Series C (Dissertations), No. 434, German Geodetic Commission, C. H. Beck'sche Verlagsbuchhandlung, München, 142 pages

Topcon Corporation. 1994. Auto Tracking Total Station AP-L1. Publ. No. 94-08-36W TE.TS 16-1, 8 pages

Vincueria, I., Tudanca, M., Aroca, C., Lopez, E., Sanchez, M. C., Sanchez, P. 1994. Flux-Gate Sensor Based on Planar Technology. IEEE Transactions on Magnetics, 30(6): 5042-5045

Walcher, H. 1985. Winkel- und Wegmessung im Maschinenbau (Angle and Distance Measurement in Mechanical Engineering). 2nd ed. VDI-Verlag GmbH, Düsseldorf, 218 pages

Whitaker, S., Call, D. B. 1986. Intelligent Pressure Transducers. Paper, Symposium, Instrumentation Society of America, Houston, Texas, October 1986, 8 pages

Whitaker, S., Call, D.B. 1987. A New Hand-Held Barometer/Altimeter Offers Portable Accuracy. Paper 24.1, Extended Abstracts, Sixth Symposium on Meteorological Observations and Instrumentation, American Meteorological Society (AMS), New Orleans, 12-16 January 1987, pp. 4487-451

Witte, B., Yang, J. 1994. Das Hand-Lasermeter DISTO der Firma Leica (The Hand-Held Laser Meter DISTO by Leica). Vermessungswesen und Raumordnung (VR), 56(4+5): 284-289

Woschitz, H. 2003. System Calibration of Digital Levels: Calibration Facility, Procedures and Results. PhD Thesis, Dept of Engineering Geodesy and Measuring Systems, Technical University, Graz, Austria, 142 pages

Wüller, H. 1988. Entwicklung und Untersuchung eines Rotations-Nivellierinstrumentes und einer photoelektrischen Nivellierlatte zur Automatisierung des geometrischen Nivellements (Development and Evaluation of a Rotating Laser Level and of a Photo-Electric Levelling Staff for the Automation of Geodetic Levelling). No. 41 of the Veröffentlichungen des Geodätischen Instituts der Rheinisch-Westfälischen Technischen Hochschule Aachen, Germany, 138 pages (ISSN 0515-0574)

Zeiske, K. 1999. TPS1100 Professional Series – A new Electronic Tacheometer Generation from Leica. Vermessungswesen und Raumordnung (VR), 61(2): 82-90 (in German)

Subject Index

3D imaging sensor 133

angle-angle-distance devices 8
  use of 13
automatic focussing 86
automatic levels 29
automatic pointing 87
automatic monitoring system 71
automatic power off 112
automatic searching 86

bar code levelling instruments 27
bar code levels 27
  accuracy 37
  errors 41
bar code staffs 39
  errors 40, 46
barometer 15
  digital 15
  electronic 15
batteries 29, 112

calibration
  computer assisted 31, 89
capacitive pressure sensor 17
cellular phones 112
Chesapeake CMS3000 141, 142
circles 30
  absolute 80
  incremental 79
  quasi-absolute 80
circle graduation errors 106
      interpolation errors 107
      periodic errors 106
    circle reading 81, 91

Accutape-S 6
altimeter 15, 137
  coarse measurement 92
  diametrical 81
  direction of rotation 95
  fine measurement 93
  graduation errors 106
clinometer 12
communication with computer 31, 85
compensator 30, 53, 82
compensation error 53, 103
computer interface 31
computing
  on-board 31, 84
corrections
  of instrument errors 31, 112
Criterion 8

data recording 31, 83
depth sounders 139
diametrical circle reading 81
digital barometer 15
digital levelling instruments 27
digital levels 32
  errors 41
direction measurement 113
dislevelment 100

electronic barometer 15
  heighting with 20
  performance 19
  working principles 15
electronic data recording 30
electronic distance measurement (levelling)
  30

- electronic level sensor 82
  - properties 82, 102
- electronic pocket tape 5
  - electro-optical sensor 7
  - ultrasound sensor 6
- electronic staff reading (levelling) 30
- electronic theodolites 65
  - circles 79
  - features 78
  - levelling 110
  - use of 109, 111
- electronic tacheometers 65, 66
  - features 78
  - levelling 110
  - motorised 68
  - reflectorless 76, 121
  - scanning 77
  - tracking 76
  - use of 109, 112
- electro-optical level sensors 95
- errors in levelling with digital levels 40
- earth curvature 57
- exercises 22, 59, 118, 144

- fluxgate compass 10
- focussing
  - automatic 86

- graduation errors
  - circles 106
- Geodimeter 600 71
- Gyromat 2000 68
- gyro-theodolites 67

- hand-held equipment 5
- high definition surveying system 133

- Inductosyn 79
- instrument calibration 89
- instrument errors 31
- interface 85
- interpolation error 107
- invar bar code staffs 40

- Kern E2 98

- ladar 77, 132, 143
- laser altimeter 137
- laser levels 48
  - calibration 54, 56
  - computing heights, earth curvature and refraction 57
  - errors of 53, 54, 55, 56
- laser measurement system 143
- laser profiling 137
- laser scanners 132
- laser theodolites 75
- laser trackers 140
- Leica Automatic Polar System (APS) 85
- Leica invar bar code staff 40
- Leica normal bar code staff 39
- Leica DISTO 6
- Leica DNA03 32
- Leica HDS3000 133
- Leica Laser Radar System LR200 143
- Leica Locator 8
- Leica NA2000 32
- Leica NA3003 42, 44
- Leica Smart310 141, 142
- Leica TCA1800 72, 88
- Leica TCRM 127
- Leica TPS1100 96, 97
- level sensor
  - electro-optical 95, 97, 98
- levelling
  - electronic distance measurement 30
  - electronic theodolites 110
  - electronic tacheometers 110
  - errors 40
  - motorised 134
- levelling equipment 27
  - automatic levels 29
  - barcode levels 27
  - digital levels 27
- levelling staffs 30
- lidar 77, 132, 143
- line of sight 47

- measurements to corners 129
- measuring units 110
- Mensi GS100 133
- Mensi GS200 133
- MetricVision MV100B 143
- metrology 140, 143
- mobile phones 112
- motorised electronic tacheometer 68
- motorised electronic video theodolites 73
- motorised levelling 134

- Nikon DTM-1 91, 92, 93, 94, 96, 97
- Nikon DTM-720 66
- Nikon DTM-750 109

- on-board computing 84
- one-person-survey-system 70
- periodic errors 107
- personal computers
  - communication 31, 85

phase measuring distance meters 127
phones
  cellular 112
  mobile 112
piezo-restistive pressure sensor 18
plumbing
  buildings 118
  shafts and wells 117
pointing
  automatic 86
power off facility 112
power supply 29, 84
precise levelling
  errors 40
pressure sensors 15
  capacitive 17
  piezo-resistive 18
  vibration 15
profiling 137
pulse distance meters 125

radar 77, 132, 143
real-time corrections 31, 112
recording
  data 83
reflectorless distance measurement 76, 121, 122
  beam size and shape 123
  measurements to corners 129
  multiple targets 125
  target material and colour 128
  target orientation 128
  use of 130
reflectorless electronic tacheometers 76, 121, 130
refraction 57
Riegl LMS-Z21 133
rotating laser levels 49

scale factor error 103
searching,
  automatic 86
setting error 104
  random component 104
  systematic component 105
single beam laser trackers 140
Sokkia SDL30 29
staff 30
staff reading 30
  electronic 30

tacheometers 65
  electronic 65
  motorised 73
  scanning 77
  tracking 76
Theis Telamat 51, 52
theodolites 65
  digital 65
  electronic 65
  gyro 67
  laser 75
  motorised 68
  video 73
Topcon AP-L1 71
Topcon DL-101C 43, 44
Topcon DL-102 28
Topcon ETL-1 90
Topcon GTS-6 110
Topcon GTS-701 66
tracking electronic tacheometers 76
two-way radios 112

ultrasound 6
ultrasound distance meter 6
umbrellas 111
use of
  digital levels 40
  electronic theodolites and tacheometers 109
  reflectorless EDM 130

vertical axis systems 90
vertical circle compensator 82
vertical circle index error 105
vertical circle index correction 105
vibration pressure sensor 15
video theodolites 73

walkie-talkies 112
Wild NA2000 32, 33, 34, 35, 36, 42
Wild TC1600 81
Wild TM3000D 72
Wild TM3000L 74
Wild TM3000V 74

Zeiss DiNi 10 28
Zeiss Elta6 107
Zeiss ETh3 108
Zeiss Reg Elta RL 126, 129
Zeiss Ni002 135
zenith angle measurement 116

Publications from the

SCHOOL OF SURVEYING AND SPATIAL INFORMATION SYSTEMS
(formerly: SCHOOL OF GEOMATIC ENGINEERING)
THE UNIVERSITY OF NEW SOUTH WALES
ABN 57 195 873 179

To order, write or fax to:
Publications Officer, School of Surveying and Spatial Information Systems,
The University of New South Wales, UNSW SYDNEY NSW 2052, AUSTRALIA
Email: survsis@unsw.edu.au Fax: +61 - 2 - 9313 - 7493

NOTE ON MAIL ORDERS: ALL ORDERS MUST BE PREPAID.
NO REQUESTS FOR RETURNS ARE ACCEPTED.
CREDIT CARDS ARE ACCEPTED. CONDITIONS APPLY. SEE ORDER FORM FOR DETAILS.
WHEN PAYING BY *CHEQUE*, THE CHEQUES MUST BE PAYABLE TO 'SCHOOL OF SURVEYING & SIS, UNSW',
IN AUSTRALIAN DOLLARS AND DRAWN ON AN AUSTRALIAN BANK.

MONOGRAPHS

Australian prices include GST #. Overseas prices include delivery by UNSW's air-lifted mail service (~2-4 weeks to Europe and North America) #. Rates for air mail through Australia Post on application.

# For all (local and overseas) **orders, add $11.00 for processing and handling** (once per order).

(Prices effective Oct 2010)

		Price # Australia (incl. GST)	**Price #** Overseas
M1.	R. S. Mather, "The Theory and Geodetic Use of some Common Projections", (3rd edition), 125 pp, reprint 1998, ISBN 0-85839-007-8	$ 16.50	$ 15.00
M2.	R. S. Mather, "The Analysis of the Earth's Gravity Field", 172 pp, 1971 (no ISBN)	$ 8.80	$ 8.00
M8.	A. H. W. Kearsley, "Geodetic Surveying", 96 pp, revised reprint 1988, ISBN 0-85839-036-1	$ 13.20	$ 12.00
M11.	W. F. Caspary, "Concepts of Network and Deformation Analysis", 183 pp, 3rd corrected impression, 2000, ISBN 0-85839-044-2	$ 27.50	$ 25.00
M12.	F. K. Brunner, "Atmospheric Effects on Geodetic Space Measurements", 110 pp, 1988, ISBN 0-85839-048-5	$ 17.60	$ 16.00
M13.	B. R. Harvey, "Practical Least Squares and Statistics for Surveyors", (3rd revised and extended edition), 341 pp, 2006, Reprint 2009, ISBN 0-7334-2339-6	$ 40.00	$ 36.36
M14.	E. G. Masters and J. R. Pollard (Eds.), "Land Information Management", 269 pp, 1991, ISBN 0-85839-061-2	$ 22.00	$ 20.00
M15/1	E. G. Masters and J. R. Pollard (Eds.), "Land Information Management - Geographic Information Systems - Advance Remote Sensing Vol. 1", 295 pp, 1993, ISBN 0-85939-064-7	$ 33.00	$ 30.00
M15/2	E. G. Masters and J. R. Pollard (Eds.), "Land Information Management - Geographic Information Systems - Advance Remote Sensing Vol. 2", 376 pp, 1993, ISBN 0-85839-065-5	$ 33.00	$ 30.00
M16.	A. Stolz, "An Introduction to Geodesy", 2nd extended edition, 148 pp, 2001, Reprint 2009, ISBN 0-7334-1736-1	$ 24.20	$ 22.00
M18.	J. M. Rüeger, "Electronic Surveying Instruments – A Review of Principles, Problems and Procedures", 1st ed., 166 pp, 2003, Reprint 2010 ISBN 0-7334-2083-4	$ 25.30	$ 23.00

UNISURV REPORTS - S SERIES

(Prices effective Oct 2010)

Australian Prices *:	S8 - S20	$11.00 #	S29 onwards	$33.00 #
Overseas Prices **:	S8 - S20	$10.00 #	S29 onwards	$30.00 #

* Australian prices include GST. #

** Overseas prices include delivery by UNSW's air-lifted mail service (~2-4 weeks to Europe and North America). #
Rates for air mail through Australia Post on application.

# **For all** (local and overseas) **orders, add $11.00 for processing and handling.**

S17. C. Rizos, "The Role of the Gravity Field in Sea Surface Topography Studies", Unisurv S17, 299 pp, 1980, ISBN 0-85839-029-9

S19. R. Coleman, "A Geodetic Basis for Recovering Ocean Dynamic Information from Satellite Altimetry", Unisurv S19, 332 pp, 1981, ISBN 0-85839-029-9

S36. A. R. Marshall, "Network Design and Optimisation in Close Range Photogrammetry", Unisurv S36, 249 pp, 1989, ISBN 0-85839-054-X

S37. W. Jaroondhampinij, "A Model of Computerised Parcel-Based Land Information System for the Department of Lands, Thailand", Unisurv S37, 281 pp, 1989, ISBN 0-85839-055-8

S39. C. Bosloper, "Multipath and GPS Short Periodic Components of the Time Variation of the Differential Dispersive Delay", Unisurv S39, 214 pp, 1990, ISBN 0-85839-057-4

S40. J. M. Nolan, "Development of a Navigational System Utilising the Global Positioning System in a Real Time, Differential Mode", Unisurv S40, 163 pp, 1990, ISBN 0-85839-058-2

S41. R. T. Macleod, "The Resolution of Mean Sea Level Anomalies along the NSW Coastline Using the Global Positioning System", Unisurv S41, 278 pp, 1990, ISBN 0-85839-060-4

S42. D. A. Kinlyside, "Densification Surveys in New South Wales - Coping with Distortions", Unisurv S42, 209 pp, 1992, ISBN 0-85839-062-0

S43. A. H. W. Kearsley (ed.), Z. Ahmad, B. R. Harvey and A. Kasenda, "Contributions to Geoid Evaluations and GPS Heighting", Unisurv S43, 209 pp, 1993, ISBN 0-85839-063-9

S44. P. Tregoning, "GPS Measurements in the Australian and Indonesian Regions (1989-1993)", Unisurv S44, 134 + xiii pp, 1996, ISBN 0-85839-068-X

S45. W.-X. Fu, "A Study of GPS and Other Navigation Systems for High Precision Navigation and Attitude Determinations", Unisurv S45, 332 pp, 1996, ISBN 085839-069-8

S46. P. Morgan et al, "A Zero Order GPS Network for the Australia Region", Unisurv S46, 187 + xii pp, 1996, ISBN 0-85839-070-1

S47. Y. Huang, "A Digital Photogrammetry System for Industrial Monitoring", Unisurv S47, 145 + xiv pp, 1997, ISBN 0-85839-072-8

S48. K. Mobbs, "Tectonic Interpretation of the Papua New Guinea Region from Repeat Satellite Measurements", Unisurv S48, 256 + xv pp, 1997, ISBN 0-85839-073-6

S50. M. D. Subari, "Low-cost GPS Systems for Intermediate Surveying and Mapping Accuracy Applications", Unisurv S50, 179 + xiii pp, 1997, ISBN 0-85839-075-2

S51. L.-S. Lin, "Real-Time Estimation of Ionospheric Delay Using GPS Measurements", Unisurv S51, 199 + xix pp, 1997, ISBN 0-7334-1664-0

S53. D. B. Lemon, "The Nature and Management of Positional Relationships within a Local Government Geographic Information System", Unisurv S53, 273 + xvi pp, 1997, ISBN 0-7334-1678-0

S54. C. Ticehurst, "Development of Models for Monitoring the Urban Environment Using Radar Remote Sensing", Unisurv S54, 282 + xix pp, 1998, ISBN 0-7334-1679-9

S55. S. S. Boey, "A Model for Establishing the Legal Traceability of GPS Measurements for Cadastral Surveying in Australia", Unisurv S55, 186 + xi pp, 1999, ISBN 0-7334-0685-X

S56. P. Morgan and M. Pearse, "A First-Order Network for New Zealand", Unisurv S56, 134 + x pp, 1999, ISBN 0-7334-0685-8

S57. P. N. Tiangco, "A Multi-Parameter Radar Approach to Stand Structure and Forest Biomass Estimation", Unisurv S57, 319 + xxii pp, 2000, ISBN 0-7334-0786-2

S58. M. A. Syafi'i, "Object–Relational Database Management Systems (ORDBMS) for Managing Marine Spatial Data: ADCP Data Case Study", Unisurv S58, 123 + ix pp, 2000, ISBN 0-7334-1704-3

S59. X.-Q. Lu, "Strategies for Improving the Determination of Displacements of Sea Surface Temperature Patterns Using Consecutive AVHRR Thermal Images", Unisurv S59, 209 + xiii pp, 2000, ISBN 0-7334-1721-3

S60. G. Dickson, "GPS-Controlled Photography: The Design, Development and Evaluation of an Operational System Utilising Long-Range Kinematic GPS", Unisurv S60, 417 + x pp, 2000, ISBN 0-7334-1725-6

S61. J. Wang, "Modelling and Quality Control for Precise GPS and GLONASS Satellite Positioning", Unisurv S61, 171 + x pp, 2001, ISBN 0-7334-1766-3

S62. Y. Wang, "Knowledge-Based Road Extraction from Aerial Images", Unisurv S62, 178 + xi pp, 2001, ISBN 0-7334-1767-1

S63. L. Ge, "Development and Testing of Augmentations of Continuously-Operating GPS Networks to Improve their Spatial and Temporal Resolution", Unisurv S63, 230 + xvi pp, 2001, ISBN 0-7334-1841-4

S64. H.-Y. Chen, "A Study on Real-Time Medium-Range Carrier-Phase-Based GPS Multiple Reference Stations", Unisurv S64, 182 + xxiv pp, 2001, ISBN 0-7334-1842-2

S65. G. Y. K. Shea, "A Web-Based Approach to the Integration of Diverse Data Sources for GIS", Unisurv S65, 233 + xv pp, 2001, ISBN 0-7334-1845-7

S66. M. Mirbagheri, "Analysis of Interferometric SAR for Topographic Mapping", Unisurv S66, 135 + xvii pp, 2001, ISBN 0-7334-1856-2

S67. P. Wang, "Applying Two-Dimensional Kalman Filtering Techniques to Digital Elevation Models for Terrain Surface Modelling", Unisurv S67, 175 + xi pp, 2001, ISBN 0-7334-1857-0

S69. C. Satirapod, "Improving the GPS Data Processing Algorithm for Precise Static Relative Positioning", Unisurv S69, 131 + viii pp, 2002, ISBN 0-7334-1901-1

S70. R. Mason, "Developing Australian Spatial Data Policies – Existing Practices and Future Strategies", Unisurv S70, 258 + xv pp, 2002, ISBN 0-7334-1941-0

S71. C. Ogaja, "A Framework in Support of Structural Monitoring by Real Time Kinematic GPS and Multisensor Data", Unisurv S71, 191 + xiii pp, 2002, ISBN 0-7334-1958-5

S72. L. Dai, "Augmentation of GPS with GLONASS and Pseudolite Signals for Carrier Phase Based Kinematic Positioning", Unisurv S72, 188 + viii pp, 2002, ISBN 0-7334-1975-5

S73. C. Roberts, "A continuous Low-Cost GPS-Based Volcano Deformation Monitoring System in Indonesia", Unisurv S73, 271 + xvi pp, 2002, ISBN 0-7334-1976-3

S74. V. Janssen, "A Mixed-Mode GPS Network Processing Approach for Volcano Deformation Monitoring", Unisurv S74, 199 + viii pp, 2003, ISBN 0-7334-2059-1

S75. Y. H. Lu, "Automatic Building Extraction for 3D Terrain Reconstruction Using Image Interpretation Techniques", Unisurv S75, 175 + xii pp, 2004, ISBN 0-7334-2141-5

S76. Y. K. Lee, "Integration of GPS/Pseudolites/INS for High Precision Kinematic Positioning and Navigation", Unisurv S76, 200 + xiv pp, 2004, ISBN 0-7334-2149-0

Note: Since 1 January 2005, the School's PhD and MEng theses are nor longer published in this series. See www.ssis.unsw.edu.au for information on post 2004 theses.

School of Surveying and Spatial Information Systems
The University of New South Wales
ABN 57 195 873 179

ORDER FORM

To order, mail or fax this form to: Publications Officer, School of Surveying and Spatial Information Systems,
The University of New South Wales, UNSW SYDNEY NSW 2052, AUSTRALIA Fax: +61 - 2 - 9313 - 7493

Name and postal address of ordering person/organisation:

____

____

____

____

____

Email : ____________________ Fax : ____________________

Day time phone number : ______________ Date : __________

Note: All orders must be prepaid. No requests for returns are accepted.

I/We order herewith:

______________________________	$ ________
______________________________	$ ________
______________________________	$ ________
Postage and Handling (from 1 July 2004)	$ **11.00**
Total	$ ________

Payment Details:

☐ A cheque, drawn on an Australian Bank, in Australian Dollars and payable to 'School of Surveying & SIS, UNSW' is enclosed

☐ Please debit my credit card: ☐ Mastercard ☐ Visa

Name on Card: ____________________ Signature: ____________________

Card No: ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐ Expiry Date : ________

Poor Man's Electronics Web Page

Random Wire Antenna Project Links:

Back to main page
How to build it
How it works
Computer program
Homebrew arrestor

Random Wire Antenna

Homebrew lightning arrestor:

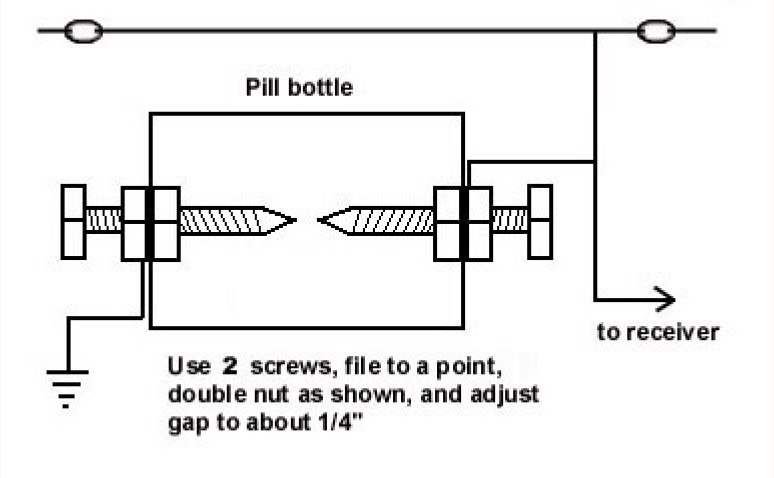
Pill bottle
to receiver
Use 2 screws, file to a point, double nut as shown, and adjust gap to about 1/4"

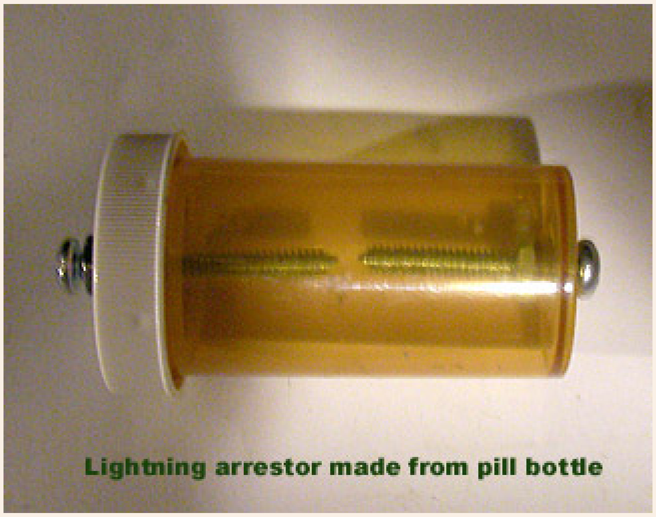

Lightning arrestor made from pill bottle

How to make it

Drill a hole in the bottom and in the lid of a 2-1/2" long by 1-1/4" dia. pill bottle. File the ends of two 3" #20 machine bolts into a sharp point. Then lock them in place on each end of the pill box with double nuts. Adjust the gap to about 1/4".

Strip about 3/4" of #14 copper stranded wire and loop it under the outside nut. Do this on both ends. The wires should connect to the antenna lead-in on one side and the ground lead wire on the other side.

Seal around the lid and connections with clear silicone caulking to make the unit water tight.

How it works

Electrical field intensity is a function of the amount of voltage divided by the surface area over which it acts. The diameter of the tip of the screws when filed down is quite small, so the surface area is quite small. Since field intensity is greater for a smaller area, it will take a relatively small voltage to raise the field intensity inside the pill bottle to the point where the air across the gap will ionize.

Once the air has ionized, it becomes a conductor; that is, a short circuit, and will bypass the static charges to ground.

Copyright © 2007 - Prof. W.S. Walker, West Virginia State University - All Rights Reserved.

Contact me

03W-8605-5.qxd 12/10/09 10:23 AM Page 1

Tektronix
Ch 1
Ch 2
Ch 3
Ch 4
Scale

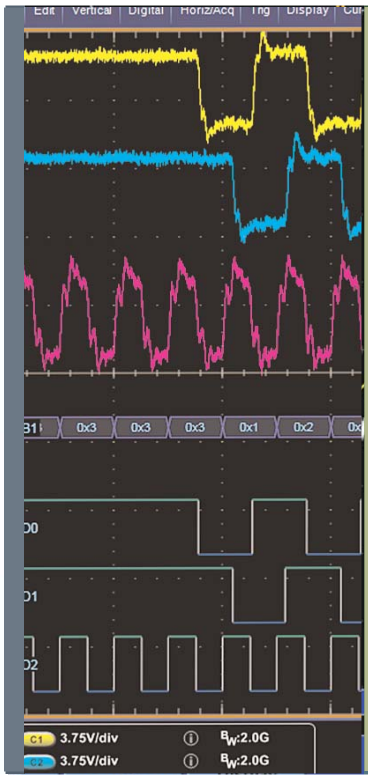
Edit
Vertical
Digital
Horiz/Acq
Trig
Display
0x3
0x3
0x3
0x1
0x2
D0
D1
D2
C1 3.75V/div
C2 3.75V/div
BW:2.0G
BW:2.0G

XYZs of Oscilloscopes

Primer

Tektronix®

Table of Contents

Introduction 4

Signal Integrity . . . . 5 - 6

- The Significance of Signal Integrity 5
- Why is Signal Integrity a Problem? 5
- Viewing the Analog Origins of Digital Signals 6

The Oscilloscope . . . . 7 - 12

- Understanding Waveforms & Waveform Measurements . .7
- Types of Waves 8
  - Sine Waves 9
  - Square and Rectangular Waves 9
  - Sawtooth and Triangle Waves 9
  - Step and Pulse Shapes 9
  - Periodic and Non-periodic Signals 9
  - Synchronous and Asynchronous Signals 9
  - Complex Waves 10
- Waveform Measurements 11
  - Frequency and Period 11
  - Voltage 11
  - Amplitude 11
  - Phase 11
  - Waveform Measurements with Digital Oscilloscopes 12

Types of Oscilloscopes .13 - 17

- Digital Storage Oscilloscopes 13
- Digital Phosphor Oscilloscopes 15
- Mixed Signal Oscilloscopes 16
- Digital Sampling Oscilloscopes 17

The Systems and Controls of an Oscilloscope .18 - 31

- Vertical System and Controls 19
  - Position and Volts per Division 19
  - Input Coupling 19
  - Bandwidth Limit 19
  - Bandwidth Enhancement 20
- Horizontal System and Controls 20
  - Acquisition Controls 20
  - Acquisition Modes 20
  - Types of Acquisition Modes 21
  - Starting and Stopping the Acquisition System 21
  - Sampling 22
  - Sampling Controls 22
  - Real-time Sampling Method 22
  - Equivalent-time Sampling Method 24
  - Position and Seconds per Division 26
  - Time Base Selections 26
  - Zoom/Pan 26
  - Search 26
  - XY Mode 26
  - Z Axis 26
  - XYZ Mode 26
- Trigger System and Controls 27
  - Trigger Position 29
  - Trigger Level and Slope 29
  - Trigger Sources 29
  - Trigger Modes 30
  - Trigger Coupling 30
  - Trigger Holdoff 30
- Display System and Controls 31
- Other Oscilloscope Controls 31
  - Math and Measurement Operations 31
  - Digital Timing and State Acquisitions 31

The Complete Measurement System . . . . 32 - 34
- Probes 32
- Passive Probes 33
- Active and Differential Probes 34
- Logicl Probes 34
- Specialty Probes 35
- Probe Accessories 35

Performance Terms and Considerations . . . . . 35 - 42
- Bandwidth 35
- Rise Time 36
- Sample Rate 37
- Waveform Capture Rate 38
- Record Length 38
- Triggering Capabilities 39
- Effective Bits 39
- Frequency Response 39
- Vertical Sensitivity 39
- Sweep Speed 39
- Gain Accuracy 39
- Horizontal Accuracy (Time Base) 39
- Vertical Resolution (Analog-to-Digital Converter) 39
- Timing Resolution (MSO) 40
- Connectivity 40
- Expandability 40
- Ease-of-use 42

Operating the Oscilloscope . . . . 43 - 45
- Proper Grounding 43
- Setting the Controls 43
- Calibrating the Instrument 44
- Connecting the Probes 44
- Compensating the Probes 44

Oscilloscope Measurement Techniques . . . . . . 46 - 48
- Voltage Measurements 46
- Time and Frequency Measurements 47
- Pulse Width and Rise Time Measurements 47
- Phase Shift Measurements 48
- Other Measurement Techniques 48

Written Exercises . . . . 49 - 54
- Part I
  - A. Vocabulary Exercises 49
  - B. Application Exercises 50
- Part II
  - A. Vocabulary Exercises 51
  - B. Application Exercises 52
- Answer Key 54

Glossary . . . . 55 - 58

Introduction

Nature moves in the form of a sine wave, be it an ocean wave, earthquake, sonic boom, explosion, sound through air, or the natural frequency of a body in motion. Energy, vibrating particles and other invisible forces pervade our physical universe. Even light – part particle, part wave – has a fundamental frequency, which can be observed as color.

Sensors can convert these forces into electrical signals that you can observe and study with an oscilloscope. Oscilloscopes enable scientists, engineers, technicians, educators and others to "see" events that change over time.

Oscilloscopes are indispensable tools for anyone designing, manufacturing or repairing electronic equipment. In today's fast-paced world, engineers need the best tools available to solve their measurement challenges quickly and accurately. As the eyes of the engineer, oscilloscopes are the key to meeting today's demanding measurement challenges.

The usefulness of an oscilloscope is not limited to the world of electronics. With the proper sensor, an oscilloscope can measure all kinds of phenomena. A sensor is a device that creates an electrical signal in response to physical stimuli, such as sound, mechanical stress, pressure, light, or heat. A microphone is a sensor that converts sound into an electrical signal. Figure 1 shows an example of scientific data that can be gathered by an oscilloscope.

Oscilloscopes are used by everyone from physicists to repair technicians. An automotive engineer uses an oscilloscope to correlate analog data from sensors with serial data from the engine control unit. A medical researcher uses an oscillscope to measure brain waves. The possibilities are endless.

The concepts presented in this primer will provide you with a good starting point in understanding oscilloscope basics and operation.

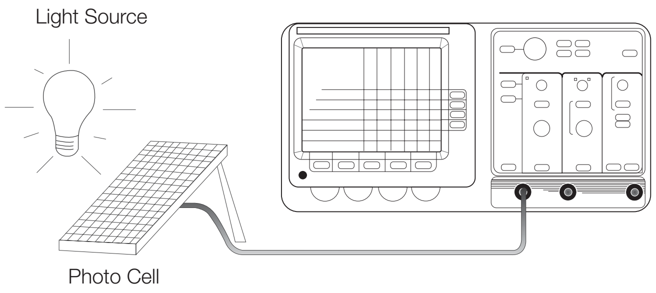

Figure 1. An example of scientific data gathered by an oscilloscope.

The glossary in the back of this primer will give you definitions of unfamiliar terms. The vocabulary and multiple-choice written exercises on oscilloscope theory and controls make this primer a useful classroom aid. No mathematical or electronics knowledge is necessary.

After reading this primer, you will be able to:

- Describe how oscilloscopes work
- Describe the differences between various oscilloscopes
- Describe electrical waveform types
- Understand basic oscilloscope controls
- Take simple measurements

The manual provided with your oscilloscope will give you more specific information about how to use the oscilloscope in your work. Some oscilloscope manufacturers also provide a multitude of application notes to help you optimize the oscilloscope for your application-specific measurements.

Should you need additional assistance, or have any comments or questions about the material in this primer, simply contact your Tektronix representative, or visit **www.tektronix.com**.

Signal Integrity

The Significance of Signal Integrity

The key to any good oscilloscope system is its ability to accurately reconstruct a waveform – referred to as signal integrity. An oscilloscope is analogous to a camera that captures signal images that we can then observe and interpret. Two key issues lie at the heart of signal integrity.

- When you take a picture, is it an accurate picture of what actually happened?
- Is the picture clear or fuzzy?
- How many of those accurate pictures can you take per second?

Taken together, the different systems and performance capabilities of an oscilloscope contribute to its ability to deliver the highest signal integrity possible. Probes also affect the signal integrity of a measurement system.

Signal integrity impacts many electronic design disciplines. But until a few years ago, it wasn't much of a problem for digital designers. They could rely on their logic designs to act like the Boolean circuits they were. Noisy, indeterminate signals were something that occurred in high-speed designs – something for RF designers to worry about. Digital systems switched slowly and signals stabilized predictably.

Processor clock rates have since multiplied by orders of magnitude. Computer applications such as 3D graphics, video and server I/O demand vast bandwidth. Much of today's telecommunications equipment is digitally based, and similarly requires massive bandwidth. So too does digital high-definition TV. The current crop of microprocessor devices handles data at rates up to 2, 3 and even 5 GS/s (gigasamples per second), while some DDR3 memory devices use clocks in excess of 2 GHz as well as data signals with 35 ps rise times.

Importantly, speed increases have trickled down to the common IC devices used in automobiles, consumer electronics, and machine controllers, to name just a few applications.

A processor running at a 20 MHz clock rate may well have signals with rise times similar to those of an 800 MHz processor. Designers have crossed a performance threshold that means, in effect, almost every design is a high-speed design.

Without some precautionary measures, high-speed problems can creep into otherwise conventional digital designs. If a circuit is experiencing intermittent failures, or if it encounters errors at voltage and temperature extremes, chances are there are some hidden signal integrity problems. These can affect time-to-market, product reliability, EMI compliance, and more. These high-speed problems can also impact the integrity of a serial data stream in a system, requiring some method of correlating specific patterns in the data with the observed characteristics of high-speed waveforms.

Why is Signal Integrity a Problem?

Let's look at some of the specific causes of signal degradation in today's digital designs. Why are these problems so much more prevalent today than in years past?

The answer is speed. In the "slow old days," maintaining acceptable digital signal integrity meant paying attention to details like clock distribution, signal path design, noise margins, loading effects, transmission line effects, bus termination, decoupling and power distribution. All of these rules still apply, but…

Bus cycle times are up to a thousand times faster than they were 20 years ago! Transactions that once took microseconds are now measured in nanoseconds. To achieve this improvement, edge speeds too have accelerated: they are up to 100 times faster than those of two decades ago.

This is all well and good; however, certain physical realities have kept circuit board technology from keeping up the pace. The propagation time of inter-chip buses has remained almost unchanged over the decades. Geometries have shrunk, certainly, but there is still a need to provide circuit board real estate for IC devices, connectors, passive components, and of course, the bus traces themselves. This real estate adds up to distance, and distance means time – the enemy of speed.

It's important to remember that the edge speed – rise time – of a digital signal can carry much higher frequency components than its repetition rate might imply. For this reason, some designers deliberately seek IC devices with relatively "slow" rise times.

The lumped circuit model has always been the basis of most calculations used to predict signal behavior in a circuit. But when edge speeds are more than four to six times faster than the signal path delay, the simple lumped model no longer applies.

Circuit board traces just six inches long become transmission lines when driven with signals exhibiting edge rates below four to six nanoseconds, irrespective of the cycle rate. In effect, new signal paths are created. These intangible connections aren't on the schematics, but nevertheless provide a means for signals to influence one another in unpredictable ways.

Sometimes even the errors introduced by the probe/instrument combination can provide a significant contribution to the signal being measured. However, by applying the "square root of the sum of the squares" formula to the measured value, it is possible to determine whether the device under test is approaching a rise/fall time failure. In addition, recent oscilloscope tools use special filtering techniques to de-embed the measurement system's effects on the signal, displaying edge times and other signal characteristics.

At the same time, the intended signal paths don't work the way they are supposed to. Ground planes and power planes, like the signal traces described above, become inductive and act like transmission lines; power supply decoupling is far less effective. EMI goes up as faster edge speeds produce shorter wavelengths relative to the bus length. Crosstalk increases.

In addition, fast edge speeds require generally higher currents to produce them. Higher currents tend to cause ground bounce, especially on wide buses in which many signals switch at once. Moreover, higher current increases the amount of radiated magnetic energy and with it, crosstalk.

Viewing the Analog Origins of Digital Signals

What do all these characteristics have in common? They are classic analog phenomena. To solve signal integrity problems, digital designers need to step into the analog domain. And to take that step, they need tools that can show them how digital and analog signals interact.

Digital errors often have their roots in analog signal integrity problems. To track down the cause of the digital fault, it's often necessary to turn to an oscilloscope, which can display waveform details, edges and noise; can detect and display transients; and can help you precisely measure timing relationships such as setup and hold times. Modern oscilloscopes can help to simplify the troubleshooting process by triggering on specific patterns in parallel or serial data streams and displaying the analog signal that corresponds in time with a specified event.

Understanding each of the systems within your oscilloscope and how to apply them will contribute to the effective application of the oscilloscope to tackle your specific measurement challenge.

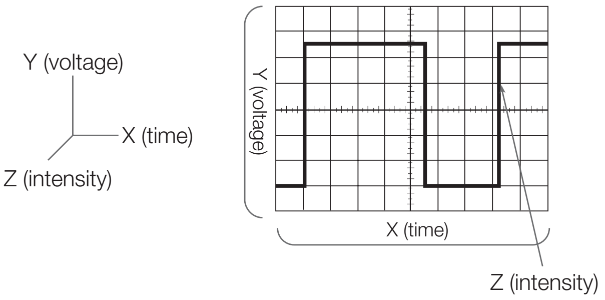
Y (voltage)
X (time)
Z (intensity)
Y (voltage)
X (time)
Z (intensity)

Figure 2. X, Y, and Z components of a displayed waveform.

The Oscilloscope

What is an oscilloscope and how does it work? This section answers these fundamental questions.

The oscilloscope is basically a graph-displaying device – it draws a graph of an electrical signal. In most applications, the graph shows how signals change over time: the vertical (Y) axis represents voltage and the horizontal (X) axis represents time. The intensity or brightness of the display is sometimes called the Z axis, as shown in Figure 2. In DPO oscilloscopes, the Z axis can be represented by color grading of the display, as seen in Figure 3.

This simple graph can tell you many things about a signal, such as:

- The time and voltage values of a signal
- The frequency of an oscillating signal
- The "moving parts" of a circuit represented by the signal
- The frequency with which a particular portion of the signal is occurring relative to other portions
- Whether or not a malfunctioning component is distorting the signal
- How much of a signal is direct current (DC) or alternating current (AC)
- How much of the signal is noise and whether the noise is changing with time

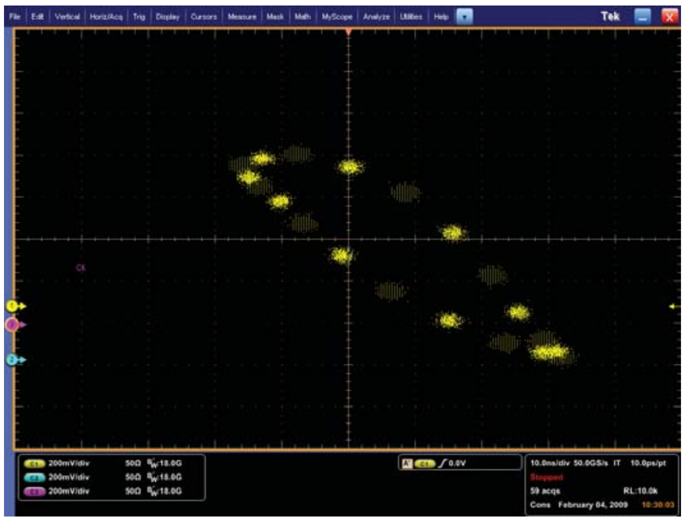
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 200mV/div 50Ω 18.0G
C2 200mV/div 50Ω 18.0G
C3 200mV/div 50Ω 18.0G
A C1 0.0V
10.0ns/div 50.0GS/s IT 10.0ps/pt
Stopped
59 acqs RL:10.0k
Cons February 04, 2009 10:30:03

Figure 3. Two offset clock patterns with Z axis intensity grading.

Understanding Waveforms and Waveform Measurements

The generic term for a pattern that repeats over time is a wave – sound waves, brain waves, ocean waves, and voltage waves are all repetitive patterns. An oscilloscope measures voltage waves. Remember as mentioned earlier, that physical phenomena such as vibrations or temperature or electrical phenomena such as current or power can be converted to a voltage by a sensor. One cycle of a wave is the portion of the wave that repeats. A waveform is a graphic representation of a wave. A voltage waveform shows time on the horizontal axis and voltage on the vertical axis.

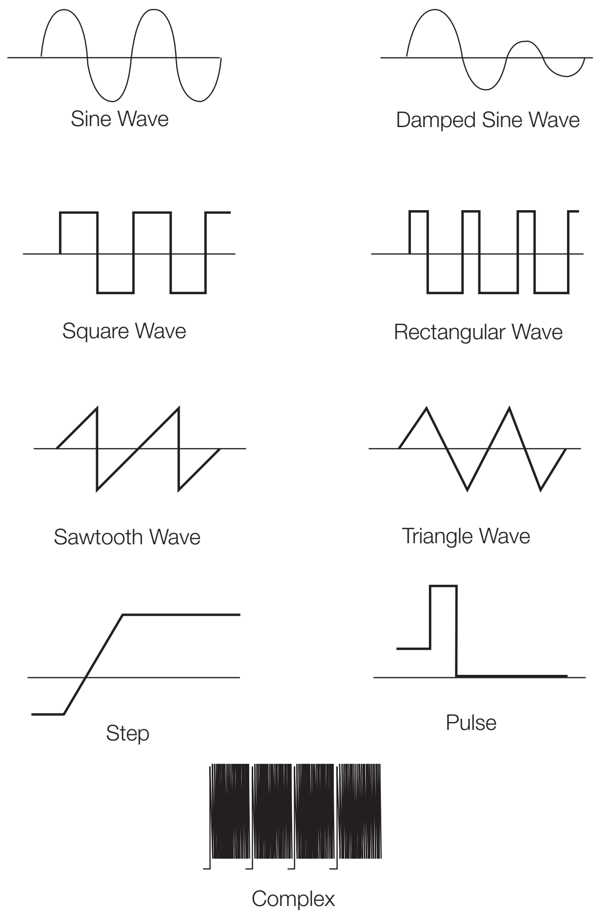
Sine Wave
Damped Sine Wave
Square Wave
Rectangular Wave
Sawtooth Wave
Triangle Wave
Step
Pulse
Complex

Figure 4. Common waveforms.

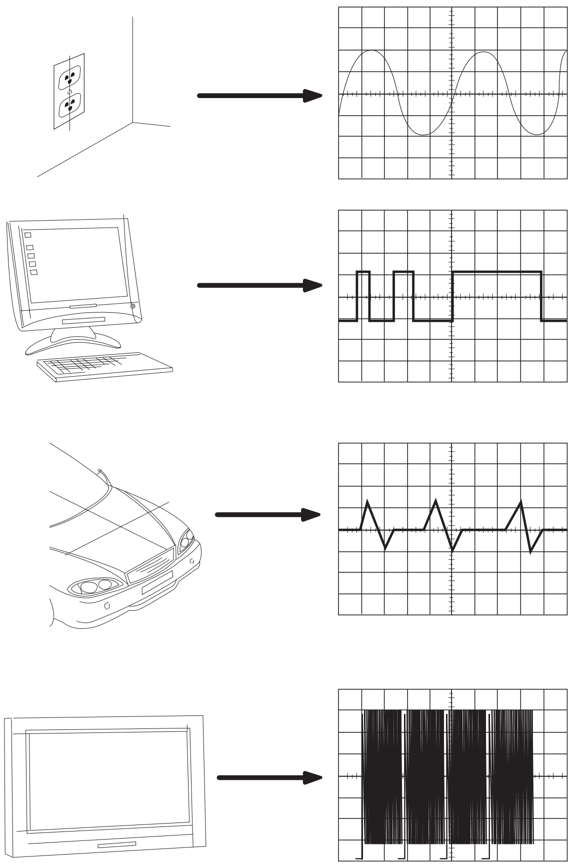

Figure 5. Sources of common waveforms.

Waveform shapes reveal a great deal about a signal. Any time you see a change in the height of the waveform, you know the voltage has changed. Any time there is a flat horizontal line, you know that there is no change for that length of time. Straight, diagonal lines mean a linear change – rise or fall of voltage at a steady rate. Sharp angles on a waveform indicate sudden change. Figure 4 shows common waveforms and Figure 5 displays sources of common waveforms.

Types of Waves

You can classify most waves into these types:

- Sine waves
- Square and rectangular waves
- Sawtooth and triangle waves
- Step and pulse shapes
- Periodic and non-periodic signals
- Synchronous and asynchronous signals
- Complex waves

Sine Waves

The sine wave is the fundamental wave shape for several reasons. It has harmonious mathematical properties – it is the same sine shape you may have studied in trigonometry class. The voltage in your wall outlet varies as a sine wave. Test signals produced by the oscillator circuit of a signal generator are often sine waves. Most AC power sources produce sine waves. (AC signifies alternating current, although the voltage alternates too. DC stands for direct current, which means a steady current and voltage, such as a battery produces.)

The damped sine wave is a special case you may see in a circuit that oscillates, but winds down over time.

Square and Rectangular Waves

The square wave is another common wave shape. Basically, a square wave is a voltage that turns on and off (or goes high and low) at regular intervals. It is a standard wave for testing amplifiers – good amplifiers increase the amplitude of a square wave with minimum distortion. Television, radio and computer circuitry often use square waves for timing signals.

The rectangular wave is like the square wave except that the high and low time intervals are not of equal length. It is particularly important when analyzing digital circuitry.

Sawtooth and Triangle Waves

Sawtooth and triangle waves result from circuits designed to control voltages linearly, such as the horizontal sweep of an analog oscilloscope or the raster scan of a television. The transitions between voltage levels of these waves change at a constant rate. These transitions are called ramps.

Step and Pulse Shapes

Signals such as steps and pulses that occur rarely, or non-periodically, are called single-shot or transient signals. A step indicates a sudden change in voltage, similar to the voltage change you would see if you turned on a power switch.

A pulse indicates sudden changes in voltage, similar to the voltage changes you would see if you turned a power switch on and then off again. A pulse might represent one bit of information traveling through a computer circuit or it might be a glitch, or defect, in a circuit. A collection of pulses traveling together creates a pulse train. Digital components in a computer communicate with each other using pulses. These pulses may be in the form of serial data stream or multiple signal lines may be used to represent a value in a parallel data bus. Pulses are also common in x-ray, radar, and communications equipment.

Periodic and Non-periodic Signals

Repetitive signals are referred to as periodic signals, while signals that constantly change are known as non-periodic signals. A still picture is analogous to a periodic signal, while a moving picture can be equated to a non-periodic signal.

Synchronous and Asynchronous Signals

When a timing relationship exists between two signals, those signals are referred to as synchronous. Clock, data and address signals inside a computer are an example of synchronous signals.

Asynchronous is a term used to describe those signals between which no timing relationship exists. Because no time correlation exists between the act of touching a key on a computer keyboard and the clock inside the computer, these are considered asynchronous.

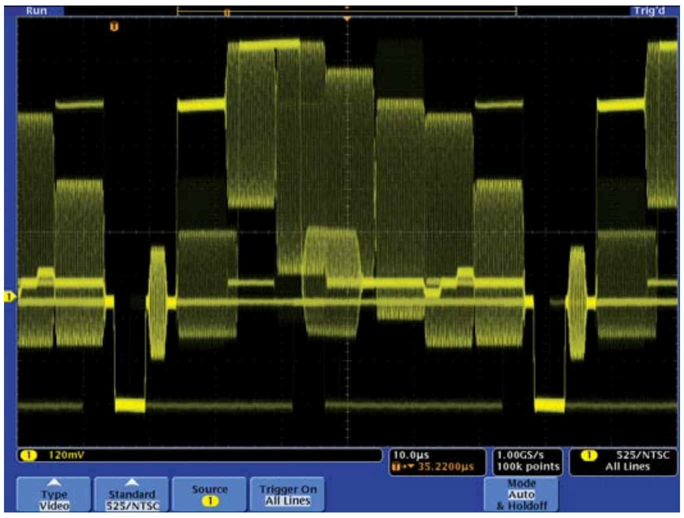

Figure 6. An NTSC composite video signal is an example of a complex wave.

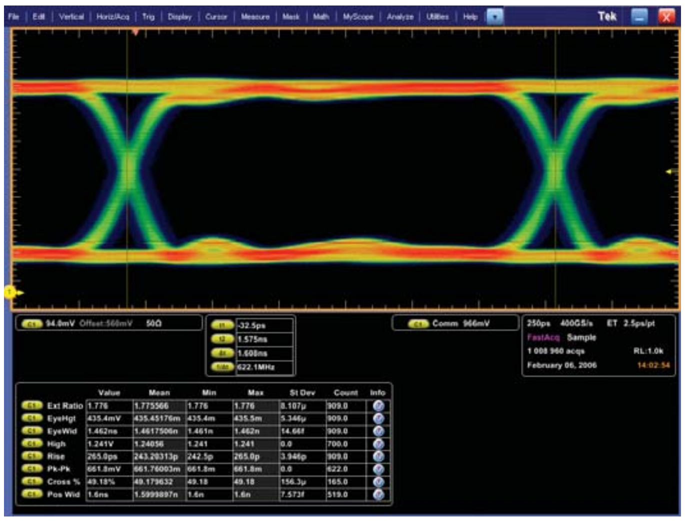

C1 94.0mV Offset:560mV 50Ω

-32.5ps
1.575ns
1.608ns
622.1MHz

C1 Comm 966mV

250ps 400GS/s ET 2.5ps/pt
FastAcq Sample
1 008 960 acqs RL:1.0k
February 06, 2006 14:02:54

	Value	Mean	Min	Max	St Dev	Count	Info
C1 Ext Ratio	1.776	1.775566	1.776	1.776	8.107µ	909.0	
C1 EyeHgt	435.4mV	435.45176m	435.4m	435.5m	5.346µ	909.0	
C1 EyeWid	1.462ns	1.4617506n	1.461n	1.462n	14.66f	909.0	
C1 High	1.241V	1.24056	1.241	1.241	0.0	700.0	
C1 Rise	265.0ps	243.20313p	242.5p	265.0p	3.946p	909.0	
C1 Pk-Pk	661.8mV	661.76003m	661.8m	661.8m	0.0	622.0	
C1 Cross %	49.18%	49.179632	49.18	49.18	156.3µ	165.0	
C1 Pos Wid	1.6ns	1.5999897n	1.6n	1.6n	7.573f	519.0	

Figure 7. 622 Mb/s serial data eye pattern.

Complex Waves

Some waveforms combine the characteristics of sines, squares, steps, and pulses to produce complex waveshapes. The signal information may be embedded in the form of amplitude, phase, and/or frequency variations. For example, although the signal in Figure 6 is an ordinary composite video signal, it is composed of many cycles of higher-frequency waveforms embedded in a lower-frequency envelope.

In this example, it is usually most important to understand the relative levels and timing relationships of the steps. To view this signal, you need an oscilloscope that captures the low-frequency envelope and blends in the higher-frequency waves in an intensity-graded fashion so that you can see their overall combination as an image that can be visually interpreted. Digital phosphor oscilloscopes are most suited to viewing complex waves, such as video signals, illustrated in Figure 6. Their displays provide the necessary frequency-of-occurrence information, or intensity grading, that is essential to understanding what the waveform is really doing.

Some oscilloscopes allow for displaying certain types of complex waveforms in special ways. For example, telecommunications data may be displayed as an eye pattern or a constellation diagram.

Telecommunications digital data signals can be displayed on an oscilloscope as a special type of waveform referred to as an eye pattern. The name comes from the similarity of the waveform to a series of eyes, as seen in Figure 7. Eye patterns are produced when digital data from a receiver is sampled and applied to the vertical input, while the data rate is used to trigger the horizontal sweep. The eye pattern displays one bit or unit interval of data with all possible edge transitions and states superimposed in one comprehensive view.

A constellation diagram is a representation of a signal modulated by a digital modulation scheme such as quadrature amplitude modulation or phase-shift keying.

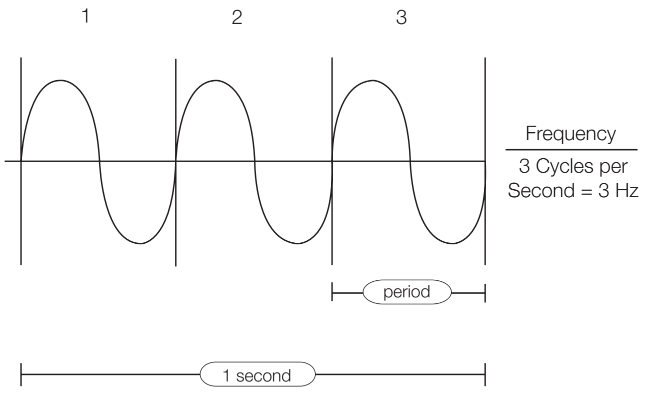

Figure 8. Frequency and period of a sine wave.

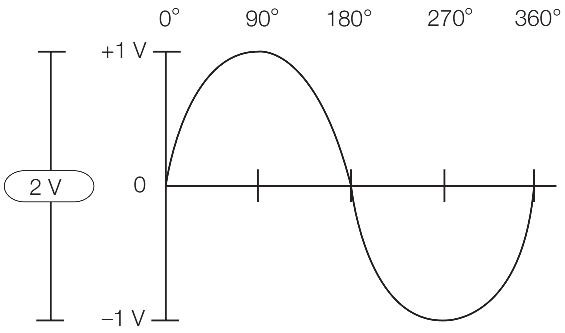

Figure 9. Amplitude and degrees of a sine wave.

Waveform Measurements

Many terms are used to describe the types of measurements that you make with your oscilloscope. This section describes some of the most common measurements and terms.

Frequency and Period

If a signal repeats, it has a frequency. The frequency is measured in Hertz (Hz) and equals the number of times the signal repeats itself in one second, referred to as cycles per second. A repetitive signal also has a period, which is the amount of time it takes the signal to complete one cycle. Period and frequency are reciprocals of each other, so that 1/period equals the frequency and 1/frequency equals the period. For example, the sine wave in Figure 8 has a frequency of 3 Hz and a period of 1/3 second.

Voltage

Voltage is the amount of electric potential, or signal strength, between two points in a circuit. Usually, one of these points is ground, or zero volts, but not always. You may want to measure the voltage from the maximum peak to the minimum peak of a waveform, referred to as the peak-to-peak voltage.

Amplitude

Amplitude refers to the amount of voltage between two points in a circuit. Amplitude commonly refers to the maximum voltage of a signal measured from ground, or zero volts. The waveform shown in Figure 9 has an amplitude of 1 V and a peak-to-peak voltage of 2 V.

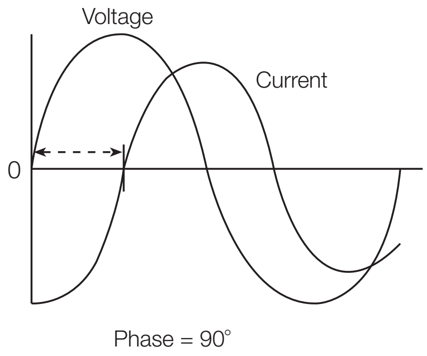

Figure 10. Phase shift.

Phase

Phase is best explained by looking at a sine wave. The voltage level of sine waves is based on circular motion. Given that a circle has 360°, one cycle of a sine wave has 360°, as shown in Figure 9. Using degrees, you can refer to the phase angle of a sine wave when you want to describe how much of the period has elapsed.

Phase shift describes the difference in timing between two otherwise similar signals. The waveform in Figure 10 labeled “current” is said to be 90° out of phase with the waveform labeled “voltage,” since the waves reach similar points in their cycles exactly 1/4 of a cycle apart (360°/4 = 90°). Phase shifts are common in electronics.

Waveform Measurements with Digital Oscilloscopes

Modern digital oscilloscopes have functions that make waveform measurements easier. They have front-panel buttons and/or screen-based menus from which you can select fully automated measurements. These include amplitude, period, rise/fall time, and many more. Many digital instruments also provide mean and RMS calculations, duty cycle, and other math operations. Automated measurements appear as on-screen alphanumeric readouts. Typically these readings are more accurate than is possible to obtain with direct graticule interpretation.

Examples of fully automated waveform measurements:

- Period
- Frequency
- Width +
- Width -
- Rise time
- Fall time
- Amplitude
- Extinction ratio
- Mean optical power
- Duty Cycle +
- Duty Cycle -
- Delay
- Phase
- Burst width
- Peak-to-peak
- Mean
- Cycle mean
- Cycle area
- High
- Low
- Minimum
- Maximum
- Overshoot +
- Overshoot -
- RMS
- Cycle RMS
- Jitter

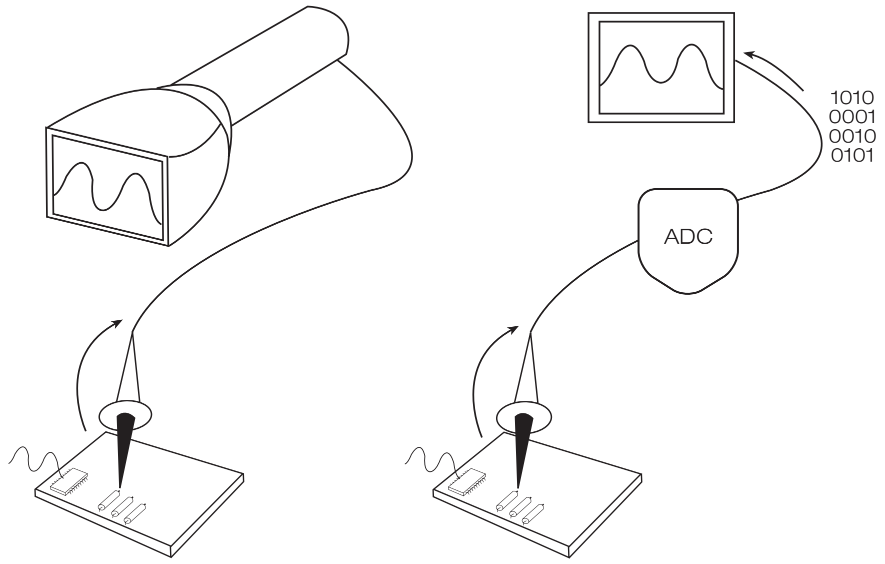

Analog Oscilloscopes
Trace Signals

Digital Oscilloscopes Samples
Signals and Construct Displays

Figure 11. Analog oscilloscopes trace signals, while digital oscilloscopes sample signals and construct displays.

The Types of Oscilloscopes

Electronic equipment can be classified into two categories: analog and digital. Analog equipment works with continuously variable voltages, while digital equipment works with discrete binary numbers that represent voltage samples. A conventional phonograph is an analog device, while a compact disc player is a digital device.

Oscilloscopes can be classified similarly – as analog and digital types. In contrast to an analog oscilloscope, a digital oscilloscope uses an analog-to-digital converter (ADC) to convert the measured voltage into digital information. It acquires the waveform as a series of samples, and stores these samples until it accumulates enough samples to describe a waveform. The digital oscilloscope then re-assembles the waveform for display on the screen, as seen in Figure 11.

Digital oscilloscopes can be classified into digital storage oscilloscopes (DSOs), digital phosphor oscilloscopes (DPOs), mixed signal oscilloscopes (MSOs), and digital sampling oscilloscopes.

The digital approach means that the oscilloscope can display any frequency within its range with stability, brightness, and clarity. For repetitive signals, the bandwidth of the digital oscilloscope is a function of the analog bandwidth of the front-end components of the oscilloscope, commonly referred to as the –3 dB point. For single-shot and transient events, such as pulses and steps, the bandwidth can be limited by the oscilloscope's sample rate. Please refer to the Sample Rate section under Performance Terms and Considerations for a more detailed discussion.

Digital Storage Oscilloscopes

A conventional digital oscilloscope is known as a digital storage oscilloscope (DSO). Its display typically relies on a raster-type screen rather than the luminous phosphor found in an older analog oscilloscope.

Digital storage oscilloscopes (DSOs) allow you to capture and view events that may happen only once – known as transients. Because the waveform information exists in digital form as a series of stored binary values, it can be analyzed, archived, printed, and otherwise processed, within the oscilloscope itself or by an external computer. The waveform need not be continuous; it can be displayed even when the signal disappears. Unlike analog oscilloscopes, digital storage oscilloscopes provide permanent signal storage and extensive waveform processing. However, DSOs typically have no real-time intensity grading; therefore, they cannot express varying levels of intensity in the live signal.

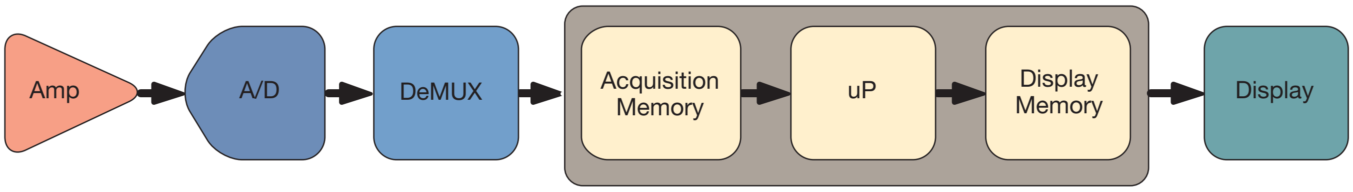

Figure 12. The serial-processing architecture of a digital storage oscilloscope (DSO).

Some of the subsystems that comprise DSOs are similar to those in analog oscilloscopes. However, DSOs contain additional data-processing subsystems that are used to collect and display data for the entire waveform. A DSO employs a serial-processing architecture to capture and display a signal on its screen, as shown in Figure 12. A description of this serial-processing architecture follows.

Serial-processing Architecture

Like an analog oscilloscope, a DSO's first (input) stage is a vertical amplifier. Vertical controls allow you to adjust the amplitude and position range at this stage. Next, the analog-to-digital converter (ADC) in the horizontal system samples the signal at discrete points in time and converts the signal's voltage at these points into digital values called sample points. This process is referred to as digitizing a signal.

The horizontal system's sample clock determines how often the ADC takes a sample. This rate is referred to as the sample rate and is expressed in samples per second (S/s).

The sample points from the ADC are stored in acquisition memory as waveform points. Several sample points may comprise one waveform point. Together, the waveform points comprise one waveform record. The number of waveform points used to create a waveform record is called the record length. The trigger system determines the start and stop points of the record.

The DSO's signal path includes a microprocessor through which the measured signal passes on its way to the display. This microprocessor processes the signal, coordinates display activities, manages the front panel controls, and more. The signal then passes through the display memory and is displayed on the oscilloscope screen.

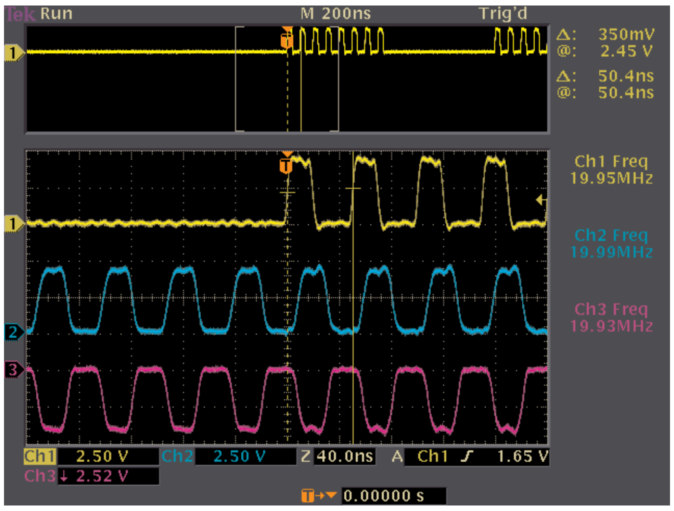

Figure 13. The digital storage oscilloscope delivers high-speed, single-shot acquisition across multiple channels, increasing the likelihood of capturing elusive glitches and transient events.

Depending on the capabilities of your oscilloscope, additional processing of the sample points may take place, which enhances the display. Pre-trigger may also be available, enabling you to see events before the trigger point. Most of today's digital oscilloscopes also provide a selection of automatic parametric measurements, simplifying the measurement process.

As shown in Figure 13, a DSO provides high performance in a single-shot, multi-channel instrument. DSOs are ideal for low-repetition-rate or single-shot, high-speed, multi-channel design applications. In the real world of digital design, an engineer usually examines four or more signals simultaneously, making the DSO a critical companion.

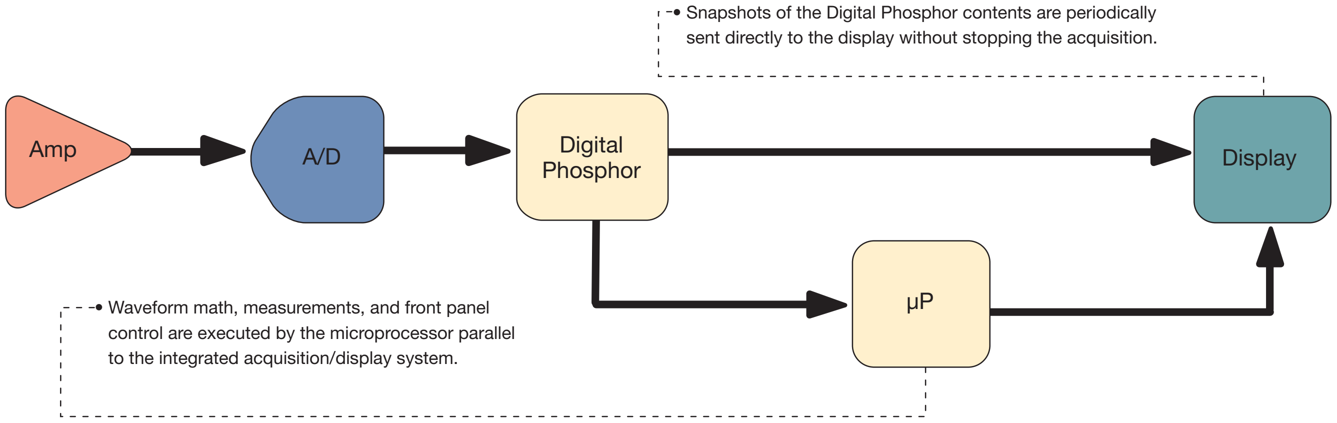
Snapshots of the Digital Phosphor contents are periodically sent directly to the display without stopping the acquisition.
Amp
A/D
Digital Phosphor
Display
μP
Waveform math, measurements, and front panel control are executed by the microprocessor parallel to the integrated acquisition/display system.

Figure 14. The parallel-processing architecture of a digital phosphor oscilloscope (DPO).

Digital Phosphor Oscilloscopes

The digital phosphor oscilloscope (DPO) offers a new approach to oscilloscope architecture. This architecture enables a DPO to deliver unique acquisition and display capabilities to accurately reconstruct a signal.

While a DSO uses a serial-processing architecture to capture, display and analyze signals, a DPO employs a parallel-processing architecture to perform these functions, as shown in Figure 14. The DPO architecture dedicates unique ASIC hardware to acquire waveform images, delivering high waveform capture rates that result in a higher level of signal visualization. This performance increases the probability of witnessing transient events that occur in digital systems, such as runt pulses, glitches and transition errors, and enables additional analysis capability. A description of this parallel-processing architecture follows.

Parallel-processing Architecture

A DPO's first (input) stage is similar to that of an analog oscilloscope – a vertical amplifier – and its second stage is similar to that of a DSO – an ADC. But, the DPO differs significantly from its predecessors following the analog-to-digital conversion.

For any oscilloscope – analog, DSO or DPO – there is always a holdoff time during which the instrument processes the most recently acquired data, resets the system, and waits for the next trigger event. During this time, the oscilloscope is blind to all signal activity. The probability of seeing an infrequent or low-repetition event decreases as the holdoff time increases.

It should be noted that it is impossible to determine the probability of capture by simply looking at the display update rate. If you rely solely on the update rate, it is easy to make the mistake of believing that the oscilloscope is capturing all pertinent information about the waveform when, in fact, it is not.

The digital storage oscilloscope processes captured waveforms serially. The speed of its microprocessor is a bottleneck in this process because it limits the waveform capture rate.

The DPO rasterizes the digitized waveform data into a digital phosphor database. Every 1/30th of a second – about as fast as the human eye can perceive it – a snapshot of the signal image that is stored in the database is pipelined directly to the display system. This direct rasterization of waveform data, and direct copy to display memory from the database, removes the data-processing bottleneck inherent in other architectures. The result is an enhanced "real-time" and lively display update. Signal details, intermittent events, and dynamic characteristics of the signal are captured in real-time. The DPO's microprocessor works in parallel with this integrated acquisition system for display management, measurement automation and instrument control, so that it does not affect the oscilloscope's acquisition speed.

A DPO faithfully emulates the best display attributes of an analog oscilloscope, displaying the signal in three dimensions: time, amplitude and the distribution of amplitude over time, all in real-time.

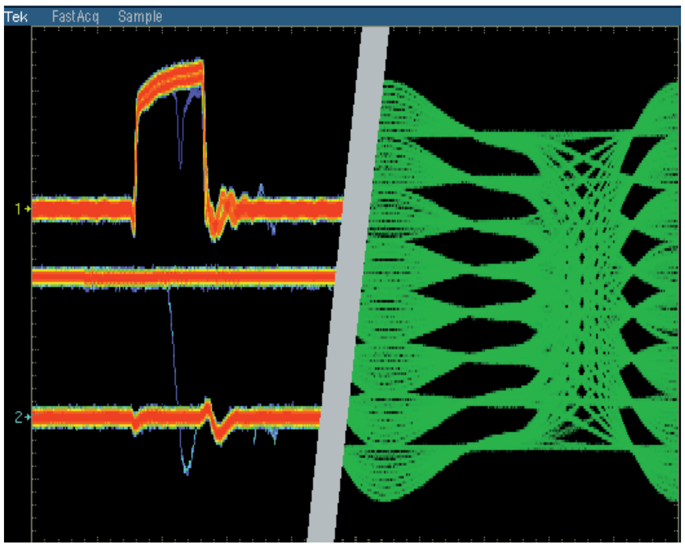

Figure 15. Some DPOs can acquire millions of waveform in just seconds, significantly increasing the probability of capturing intermittent and elusive events and revealing dynamic signal behavior.

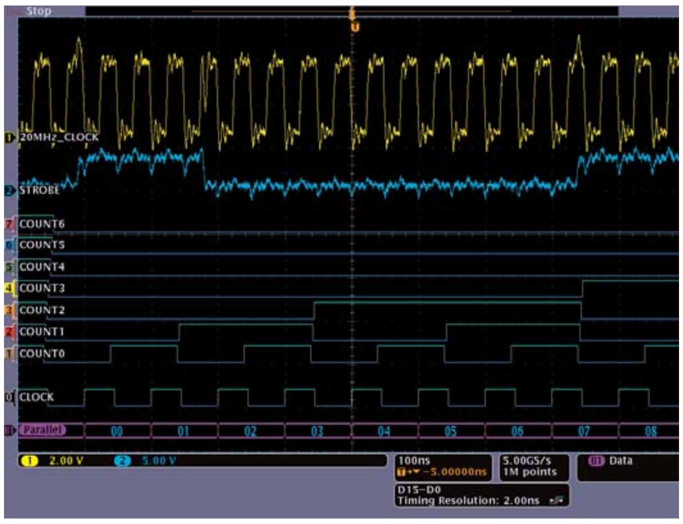

Figure 16. The MSO provides 16 integrated digital channels, enabling the ability to view and analyze time-correlated analog and digital signals.

Unlike an analog oscilloscope's reliance on chemical phosphor, a DPO uses a purely electronic digital phosphor that's actually a continuously updated database. This database has a separate "cell" of information for every single pixel in the oscilloscope's display. Each time a waveform is captured – in other words, every time the oscilloscope triggers – it is mapped into the digital phosphor database's cells. Each cell that represents a screen location and is touched by the waveform is reinforced with intensity information, while other cells are not. Thus, intensity information builds up in cells where the waveform passes most often.

When the digital phosphor database is fed to the oscilloscope's display, the display reveals intensified waveform areas, in proportion to the signal's frequency of occurrence at each point – much like the intensity grading characteristics of an analog oscilloscope. The DPO also allows the display of the varying frequency-of-occurrence information on the display as contrasting colors, unlike an analog oscilloscope. With a DPO, it is easy to see the difference between a waveform that occurs on almost every trigger and one that occurs, say, every 100th trigger.

Digital phosphor oscilloscopes (DPOs) break down the barrier between analog and digital oscilloscope technologies. They are equally suitable for viewing high and low frequencies, repetitive waveforms, transients, and signal variations in real-time. Only a DPO provides the Z (intensity) axis in real-time that is missing from conventional DSOs.

A DPO is ideal for those who need the best general-purpose design and troubleshooting tool for a wide range of applications, as seen in Figure 15. A DPO is exemplary for advanced analysis, communication mask testing, digital debug of intermittent signals, repetitive digital design and timing applications.

Mixed Signal Oscilloscopes

The mixed signal oscilloscope (MSO) combines the performance of a DPO with the basic functionality of a 16-channel logic analyzer, including parallel/serial bus protocol decoding and triggering. The MSO's digital channels view a digital signal as either a logic high or logic low, just like a digital circuit views the signal. This means as long as ringing, overshoot and ground bounce do not cause logic transitions, these analog characteristics are not of concern to the MSO. Just like a logic analyzer, a MSO uses a threshold voltage to determine if the signal is logic high or logic low.

The MSO is the tool of choice for quickly debugging digital circuits using its powerful digital triggering, high resolution acquisition capability, and analysis tools. The root cause of many digital problems is quicker to pinpoint by analyzing both the analog and digital representations of the signal, as shown in Figure 16, making a MSO ideal for verifying and debugging digital circuits.

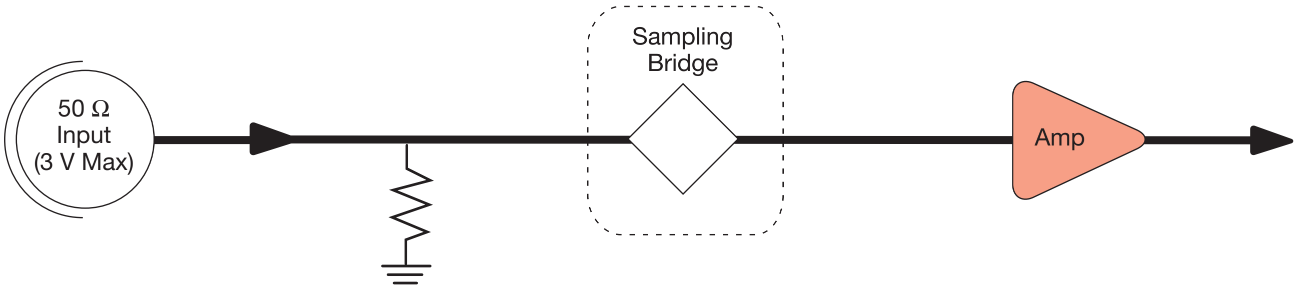

Figure 17. The architecture of a digital sampling oscilloscope.

Digital Sampling Oscilloscopes

In contrast to the digital storage and digital phosphor oscilloscope architectures, the architecture of the digital sampling oscilloscope reverses the position of the attenuator/amplifier and the sampling bridge, as shown in Figure 17. The input signal is sampled before any attenuation or amplification is performed. A low bandwidth amplifier can then be utilized after the sampling bridge because the signal has already been converted to a lower frequency by the sampling gate, resulting in a much higher bandwidth instrument.

The tradeoff for this high bandwidth, however, is that the sampling oscilloscope's dynamic range is limited. Since there is no attenuator/amplifier in front of the sampling gate, there is no facility to scale the input. The sampling bridge must be able to handle the full dynamic range of the input at all times. Therefore, the dynamic range of most sampling oscilloscopes is limited to about 1 V peak-to-peak. Digital storage and digital phosphor oscilloscopes, on the other hand, can handle 50 to 100 volts.

In addition, protection diodes cannot be placed in front of the sampling bridge as this would limit the bandwidth. This reduces the safe input voltage for a sampling oscilloscope to about 3 V, as compared to 500 V available on other oscilloscopes.

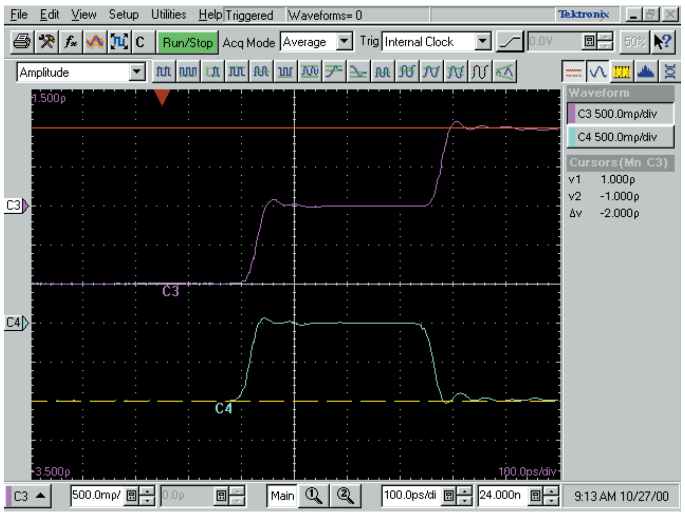

Acq Mode Average

Trig Internal Clock

C3 500.0mρ/div

C4 500.0mρ/div

Cursors(Mn C3)

v1 1.000ρ

v2 -1.000ρ

Δv -2.000ρ

C3 500.0mρ/

100.0ps/di

24.000n

9:13 AM 10/27/00

Figure 18. Time domain reflectometry (TDR) display from a digital sampling oscilloscope.

When measuring high-frequency signals, the DSO or DPO may not be able to collect enough samples in one sweep. A digital sampling oscilloscope is an ideal tool for accurately capturing signals whose frequency components are much higher than the oscilloscope's sample rate, as seen in Figure 18. This oscilloscope is capable of measuring signals of up to an order of magnitude faster than any other oscilloscope. It can achieve bandwidth and high-speed timing ten times higher than other oscilloscopes for repetitive signals. Sequential equivalent-time sampling oscilloscopes are available with bandwidths to 80 GHz.

The Systems and Controls of an Oscilloscope

This section briefly describes the basic systems and controls found on analog and digital oscilloscopes. Some controls differ between analog and digital oscilloscopes; your oscilloscope probably has additional controls not discussed here.

A basic oscilloscope consists of four different systems – the vertical system, horizontal system, trigger system, and display system. Understanding each of these systems will enable you to effectively apply the oscilloscope to tackle your specific measurement challenges. Recall that each system contributes to the oscilloscope's ability to accurately reconstruct a signal.

The front panel of an oscilloscope is divided into three main sections labeled vertical, horizontal, and trigger. Your oscilloscope may have other sections, depending on the model and type. See if you can locate these front-panel sections in Figure 19, and on your oscilloscope, as you read through this section.

When using an oscilloscope, you need to adjust three basic settings to accommodate an incoming signal:

- Vertical: The attenuation or amplification of the signal. Use the volts/div control to adjust the amplitude of the signal to the desired measurement range.
- Horizontal: The time base. Use the sec/div control to set the amount of time per division represented horizontally across the screen.
- Trigger: The triggering of the oscilloscope. Use the trigger level to stabilize a repeating signal, or to trigger on a single event.

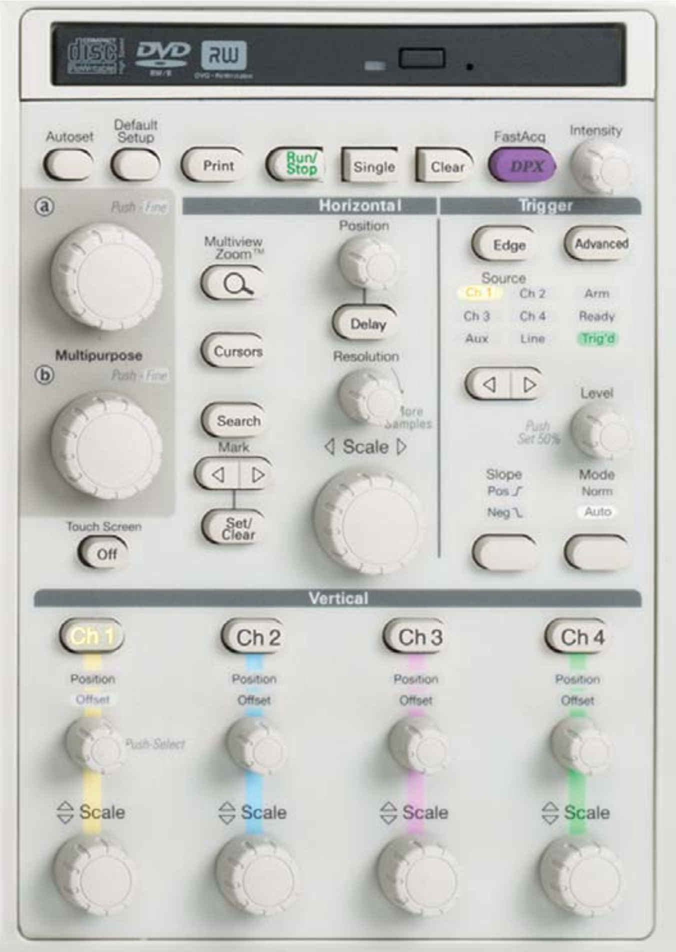
DVD RW
Autoset
Default Setup
Print
Run/Stop
Single
Clear
FastAcq
DPX
Intensity
Horizontal
Trigger
Multipurpose
Multiview Zoom
Position
Delay
Resolution
Cursors
Search
Mark
Set/Clear
Scale
Touch Screen
Off
Edge
Advanced
Source
Ch 1
Ch 2
Ch 3
Ch 4
Aux
Line
Arm
Ready
Trig'd
Level
Slope
Mode
Norm
Auto
Vertical
Ch 1
Ch 2
Ch 3
Ch 4
Position
Offset
Scale

Figure 19. Front-panel control section of an oscilloscope.

Common vertical controls include:

- Termination
  - 1M ohm
  - 50 ohm
- Coupling
  - DC
  - AC
  - GND
- Bandwidth
  - Limit
  - Enhancement
- Position
- Offset
- Invert – On/Off
- Scale
  - Fixed steps
  - Variable

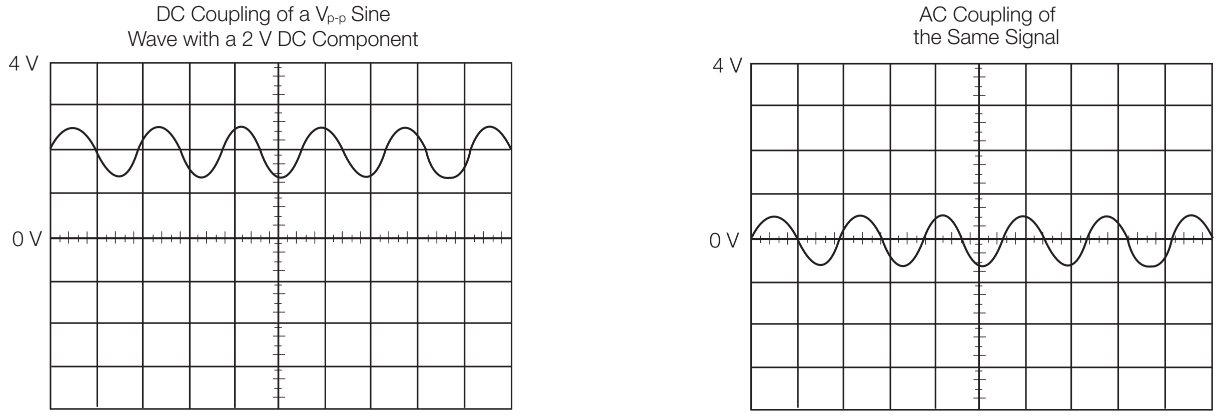

DC Coupling of a V_{p-p} Sine Wave with a 2 V DC Component

AC Coupling of the Same Signal

Figure 20. AC and DC input coupling.

Vertical System and Controls

Vertical controls can be used to position and scale the waveform vertically, set the input coupling, and adjust other signal conditioning.

Position and Volts per Division

The vertical position control allows you to move the waveform up and down exactly where you want it on the screen.

The volts-per-division setting (usually written as volts/div) is a scaling factor that varies the size of the waveform on the screen. If the volts/div setting is 5 volts, then each of the eight vertical divisions represents 5 volts and the entire screen can display 40 volts from bottom to top, assuming a graticule with eight major divisions. If the setting is 0.5 volts/div, the screen can display 4 volts from bottom to top, and so on. The maximum voltage you can display on the screen is the volts/div setting multiplied by the number of vertical divisions. Note that the probe you use, 1X or 10X, also influences the scale factor. You must divide the volts/div scale by the attenuation factor of the probe if the oscilloscope does not do it for you.

Often the volts/div scale has either a variable gain or a fine gain control for scaling a displayed signal to a certain number of divisions. Use this control to assist in taking rise time measurements.

Input Coupling

Coupling refers to the method used to connect an electrical signal from one circuit to another. In this case, the input coupling is the connection from your test circuit to the oscilloscope. The coupling can be set to DC, AC, or ground. DC coupling shows all of an input signal. AC coupling blocks the DC component of a signal so that you see the waveform centered around zero volts. Figure 20 illustrates this difference. The AC coupling setting is useful when the entire signal (alternating current + direct current) is too large for the volts/div setting.

The ground setting disconnects the input signal from the vertical system, which lets you see where zero volts is located on the screen. With grounded input coupling and auto trigger mode, you see a horizontal line on the screen that represents zero volts. Switching from DC to ground and back again is a handy way of measuring signal voltage levels with respect to ground.

Bandwidth Limit

Most oscilloscopes have a circuit that limits the bandwidth of the oscilloscope. By limiting the bandwidth, you reduce the noise that sometimes appears on the displayed waveform, resulting in a cleaner signal display. Note that, while eliminating noise, the bandwidth limit can also reduce or eliminate high-frequency signal content.

Bandwidth Enhancement

Some oscilloscopes may provide a DSP arbitrary equalization filter which can be used to improve the oscilloscope channel response. This filter extends the bandwidth, flattens the oscilloscope channel frequency response, improves phase linearity, and provides a better match between channels. It also decreases rise time and improves the time domain step response.

Horizontal System and Controls

An oscilloscope's horizontal system is most closely associated with its acquisition of an input signal – sample rate and record length are among the considerations here. Horizontal controls are used to position and scale the waveform horizontally.

Acquisition Controls

Digital oscilloscopes have settings that let you control how the acquisition system processes a signal. Look over the acquisition options on your digital oscilloscope while you read this description. Figure 21 shows you an example of an acquisition menu.

Acquisition Modes

Acquisition modes control how waveform points are produced from sample points. Sample points are the digital values derived directly from the analog-to-digital converter (ADC). The sample interval refers to the time between these sample points. Waveform points are the digital values that are stored in memory and displayed to construct the waveform. The time value difference between waveform points is referred to as the waveform interval.

Common horizontal controls include:

- Time Base
- XY
- Scale
- Trace Separation
- Record Length
- Resolution
- Sample Rate
- Trigger Position
- Zoom/Pan
- Search

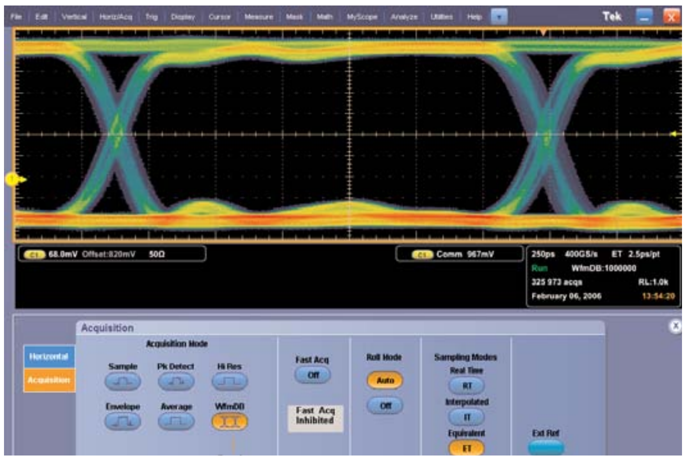
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursor
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 68.0mV Offset:820mV 50Ω
C1 Comm 967mV
250ps 400GS/s ET 2.5ps/pt
Run WfmDB:1000000
325 973 acqs RL:1.0k
February 06, 2006 13:54:20
Horizontal
Acquisition
Acquisition
Acquisition Mode
Sample
Pk Detect
Hi Res
Envelope
Average
WfmDB
Fast Acq
Off
Fast Acq Inhibited
Roll Mode
Auto
Off
Sampling Modes
Real Time
RT
Interpolated
IT
Equivalent
ET
Ext Ref

Figure 21. Example of an acquisition menu.

The sample interval and the waveform interval may, or may not, be the same. This fact leads to the existence of several different acquisition modes in which one waveform point is comprised of several sequentially acquired sample points.

Additionally, waveform points can be created from a composite of sample points taken from multiple acquisitions, which provides another set of acquisition modes. A description of the most commonly used acquisition modes follows.

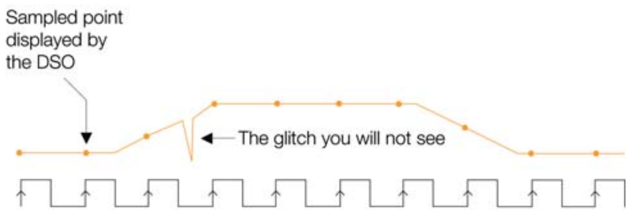
Sampled point displayed by the DSO
The glitch you will not see

Figure 22. Sample rate varies with time base settings - the slower the time based setting, the slower the sample rate. Some digital oscilloscopes provide peak detect mode to capture fast transients at slow sweep speeds.

Types of Acquisition Modes

- Sample Mode: This is the simplest acquisition mode. The oscilloscope creates a waveform point by saving one sample point during each waveform interval.
- Peak Detect Mode: The oscilloscope saves the minimum and maximum value sample points taken during two waveform intervals and uses these samples as the two corresponding waveform points. Digital oscilloscopes with peak detect mode run the ADC at a fast sample rate, even at very slow time base settings (slow time base settings translate into long waveform intervals) and are able to capture fast signal changes that would occur between the waveform points if in sample mode, as shown in Figure 22. Peak detect mode is particularly useful for seeing narrow pulses spaced far apart in time, as demonstrated in Figure 23.
- Hi-Res Mode: Like peak detect, hi res mode is a way of getting more information in cases when the ADC can sample faster than the time base setting requires. In this case, multiple samples taken within one waveform interval are averaged together to produce one waveform point. The result is a decrease in noise and an improvement in resolution for low-speed signals. The advantage of Hi Res Mode over Average is that Hi-Res Mode can be used even on a single shot event.
- Envelope Mode: Envelope mode is similar to peak detect mode. However, in envelope mode, the minimum and maximum waveform points from multiple acquisitions are combined to form a waveform that shows min/max accumulation over time. Peak detect mode is usually used to acquire the records that are combined to form the envelope waveform.

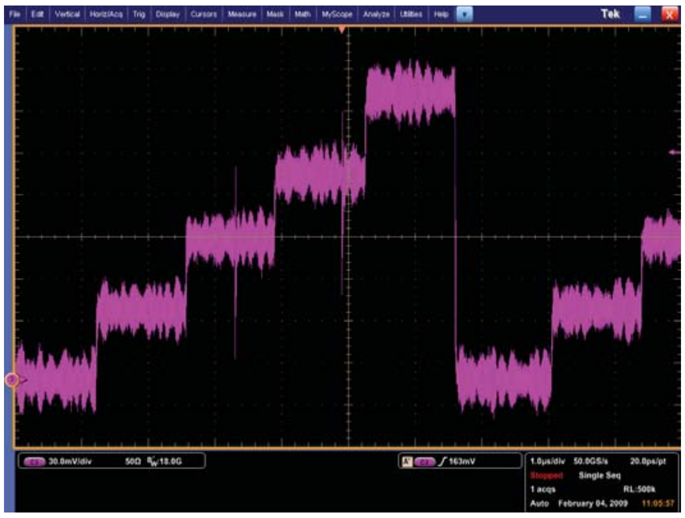

C3 30.0mV/div 50Ω B_W:18.0G

C3 ∫ 163mV

1.0µs/div 50.0GS/s 20.0ps/pt

Stopped Single Seq

1 acqs RL:500k

Auto February 04, 2009 11:05:57

Figure 23. Peak detect mode enables the oscilloscope to capture extremely brief transient anomalies.

- Average Mode: In average mode, the oscilloscope saves one sample point during each waveform interval as in sample mode. However, waveform points from consecutive acquisitions are then averaged together to produce the final displayed waveform. Average mode reduces noise without loss of bandwidth, but requires a repeating signal.
- Waveform Database Mode: In waveform database mode, the oscillocope accumulates a waveform database that provides a three-dimensional array of amplitude, time, and counts.

Starting and Stopping the Acquisition System

One of the greatest advantages of digital oscilloscopes is their ability to store waveforms for later viewing. To this end, there are usually one or more buttons on the front panel that allow you to start and stop the acquisition system so you can analyze waveforms at your leisure. Additionally, you may want the oscilloscope to automatically stop acquiring after one acquisition is complete or after one set of records has been turned into an envelope or average waveform. This feature is commonly called single sweep or single sequence and its controls are usually found either with the other acquisition controls or with the trigger controls.

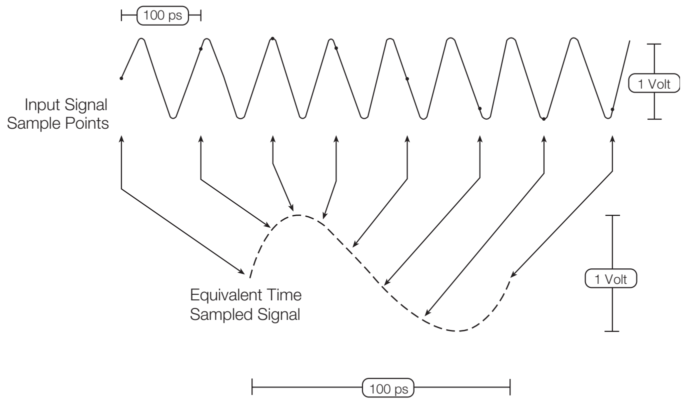

Figure 24. Basic sampling, showing sample points are connected by interpolation to produce a continuous waveform.

Sampling

Sampling is the process of converting a portion of an input signal into a number of discrete electrical values for the purpose of storage, processing and/or display. The magnitude of each sampled point is equal to the amplitude of the input signal at the instant in time in which the signal is sampled.

Sampling is like taking snapshots. Each snapshot corresponds to a specific point in time on the waveform. These snapshots can then be arranged in the appropriate order in time so as to reconstruct the input signal.

In a digital oscilloscope, an array of sampled points is reconstructed on a display with the measured amplitude on the vertical axis and time on the horizontal axis, as illustrated in Figure 24.

The input waveform in Figure 24 appears as a series of dots on the screen. If the dots are widely spaced and difficult to interpret as a waveform, the dots can be connected using a process called interpolation. Interpolation connects the dots with lines, or vectors. A number of interpolation methods are available that can be used to produce an accurate representation of a continuous input signal.

Sampling Controls

Some digital oscilloscopes provide you with a choice in sampling method – either real-time sampling or equivalent-time sampling. The acquisition controls available with these oscilloscopes will allow you to select a sample method to acquire signals. Note that this choice makes no difference for slow time base settings and only has an effect when the ADC cannot sample fast enough to fill the record with waveform points in one pass. Each sampling method has distinct advantages, depending on the kind of measurements being made.

Controls are typically available on modern oscilloscopes to give you the choice of three horizontal time base modes of operations. If you are simply doing signal exploration and want to interact with a lively signal, you will use the Automatic or interactive default mode that provides you with the liveliest display update rate. If you want a precise measurement and the highest real-time sample rate that will give you the most measurement accuracy, then the Constant Sample Rate mode is for you. It will maintain the highest sample rate and provide the best real-time resolution. The last mode is called the Manual mode because it ensures direct and independent control of the sample rate and record length.

Real-time Sampling Method

Real-time sampling is ideal for signals whose frequency range is less than half the oscilloscope's maximum sample rate. Here, the oscilloscope can acquire more than enough points in one "sweep" of the waveform to construct an accurate picture, as shown in Figure 25. Real-time sampling is the only way to capture fast, single-shot, transient signals with a digital oscilloscope.

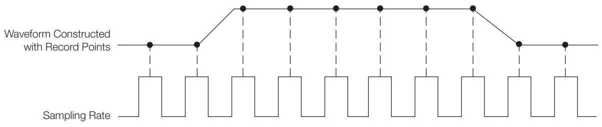

Figure 25. Real-time sampling method.

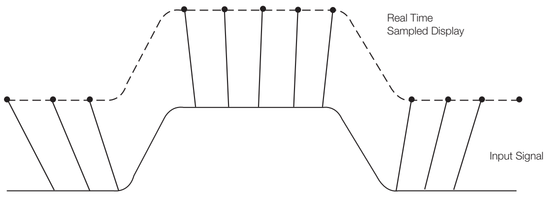

Figure 26. In order to capture this 10 ns pulse in real-time, the sample rate must be high enough to accurately define the edges.

Real-time sampling presents the greatest challenge for digital oscilloscopes because of the sample rate needed to accurately digitize high-frequency transient events, as shown in Figure 26. These events occur only once, and must be sampled in the same time frame that they occur.

If the sample rate isn't fast enough, high-frequency components can "fold down" into a lower frequency, causing aliasing in the display, as demonstrated in Figure 27. In addition, real-time sampling is further complicated by the high-speed memory required to store the waveform once it is digitized. Please refer to the Sample Rate and Record Length sections under Performance Terms and Considerations for additional detail regarding the sample rate and record length needed to accurately characterize high-frequency components.

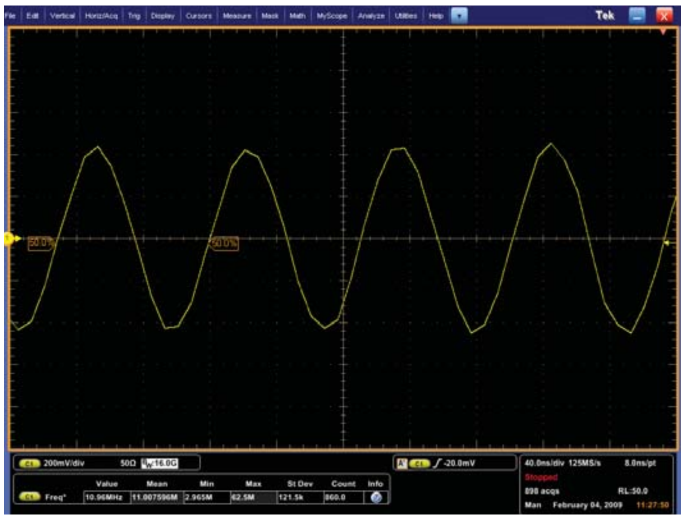

C1 200mV/div 50Ω BW:16.0G

A C1 ∫ -20.0mV

40.0ns/div 125MS/s 8.0ns/pt

Stopped

898 acqs RL:50.0

Man February 04, 2009 11:27:50

	Value	Mean	Min	Max	St Dev	Count	Info
C1 Freq*	10.96MHz	11.007596M	2.965M	62.5M	121.5k	860.0	

Figure 27. Undersampling of a 100 MHz sine wave introduces aliasing effects.

Real-time sampling with interpolation, dgital oscilloscopes take discrete samples of the signal that can be displayed. However, it can be difficult to visualize the signal represented as dots, especially because there can be only a few dots representing high-frequency portions of the signal. To aid in the visualization of signals, digital oscilloscopes typically have interpolation display modes.

In simple terms, interpolation "connects the dots" so that a signal that is sampled only a few times in each cycle can be accurately displayed. Using real-time sampling with interpolation, the oscilloscope collects a few sample points of the signal in a single pass in real-time mode and uses interpolation to fill in the gaps. Interpolation is a processing technique used to estimate what the waveform looks like based on a few points.

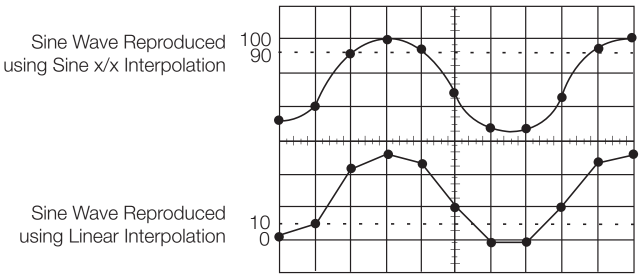

Figure 28. Linear and sin x/x interpolation.

Linear interpolation connects sample points with straight lines. This approach is limited to reconstructing straight-edged signals, as illustrated in Figure 28, which better lends itself to square waves.

The more versatile sin x/x interpolation connects sample points with curves, as shown in Figure 28. Sin x/x interpolation is a mathematical process in which points are calculated to fill in the time between the real samples. This form of interpolation lends itself to curved and irregular signal shapes, which are far more common in the real world than pure square waves and pulses. Consequently, sin x/x interpolation is the preferred method for applications where the sample rate is 3 to 5 times the system bandwidth.

Equivalent-time Sampling Method

When measuring high-frequency signals, the oscilloscope may not be able to collect enough samples in one sweep. Equivalent-time sampling can be used to accurately acquire signals whose frequency exceeds half the oscilloscope's sample rate, as illustrated in Figure 29. Equivalent-time digitizers (samplers) take advantage of the fact that most naturally occurring and man-made events are repetitive. Equivalent-time sampling constructs a picture of a repetitive signal by capturing a little bit of information from each repetition. The waveform slowly builds up like a string of lights, illuminating one-by-one. This allows the oscilloscope to accurately capture signals whose frequency components are much higher than the oscilloscope's sample rate.

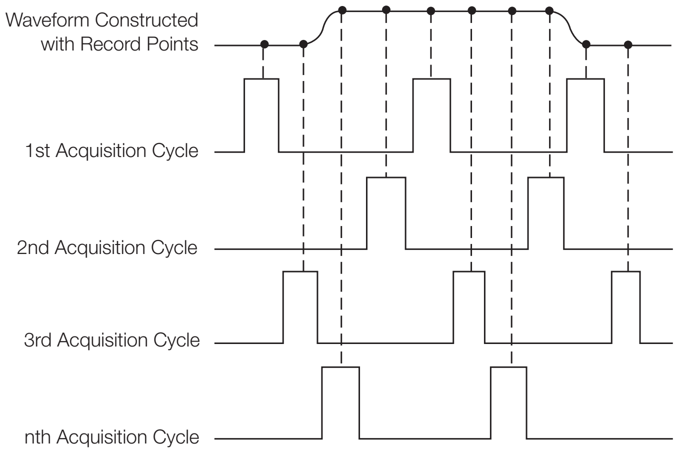

Figure 29. Some oscilloscopes use equivalent-time sampling to capture and display very fast, repetitive signals.

There are two types of equivalent-time sampling methods: random and sequential. Each has its advantages. Random equivalent-time sampling allows display of the input signal prior to the trigger point, without the use of a delay line. Sequential equivalent-time sampling provides much greater time resolution and accuracy. Both require that the input signal be repetitive.

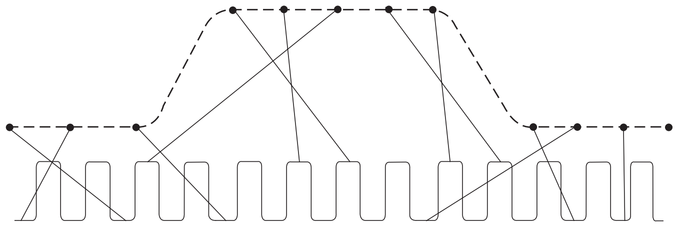

Figure 30. In random equivalent-time sampling, the sampling clock runs asynchronously with the input signal and the trigger.

Random Equivalent-time Sampling

Random equivalent-time digitizers (samplers) utilize an internal clock that runs asynchronously with respect to the input signal and the signal trigger, as illustrated in Figure 30. Samples are taken continuously, independent of the trigger position, and are displayed based on the time difference between the sample and the trigger. Although samples are taken sequentially in time, they are random with respect to the trigger – hence the name "random" equivalent-time sampling. Sample points appear randomly along the waveform when displayed on the oscilloscope screen.

The ability to acquire and display samples prior to the trigger point is the key advantage of this sampling technique, eliminating the need for external pretrigger signals or delay lines. Depending on the sample rate and the time window of the display, random sampling may also allow more than one sample to be acquired per triggered event. However, at faster sweep speeds, the acquisition window narrows until the digitizer cannot sample on every trigger. It is at these faster sweep speeds that very precise timing measurements are often made, and where the extraordinary time resolution of the sequential equivalent-time sampler is most beneficial. The bandwidth limit for random equivalent-time sampling is less than for sequential-time sampling.

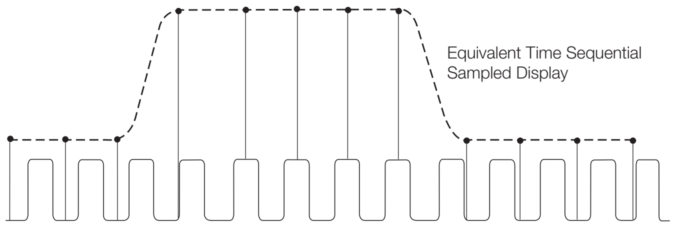

Figure 31. In sequential equivalent-time sampling, the single sample is taken for each recognized trigger after a time delay which is incremented after each cycle.

Sequential Equivalent-time Sampling

The sequential equivalent-time sampler acquires one sample per trigger, independent of the time/div setting, or sweep speed, as illustrated in Figure 31. When a trigger is detected, a sample is taken after a very short, but well-defined, delay. When the next trigger occurs, a small time increment – delta t – is added to this delay and the digitizer takes another sample. This process is repeated many times, with "delta t" added to each previous acquisition, until the time window is filled. Sample points appear from left to right in sequence along the waveform when displayed on the oscilloscope screen.

Technologically speaking, it is easier to generate a very short, very precise "delta t" than it is to accurately measure the vertical and horizontal positions of a sample relative to the trigger point, as required by random samplers. This precisely measured delay is what gives sequential samplers their unmatched time resolution. Since, with sequential sampling, the sample is taken after the trigger level is detected, the trigger point cannot be displayed without an analog delay line, which may, in turn, reduce the bandwidth of the instrument. If an external pretrigger can be supplied, bandwidth will not be affected.

Position and Seconds per Division

The horizontal position control moves the waveform left and right to exactly where you want it on the screen.

The seconds-per-division setting (usually written as sec/div) lets you select the rate at which the waveform is drawn across the screen (also known as the time base setting or sweep speed). This setting is a scale factor. If the setting is 1 ms, each horizontal division represents 1 ms and the total screen width represents 10 ms, or ten divisions. Changing the sec/div setting enables you to look at longer and shorter time intervals of the input signal.

As with the vertical volts/div scale, the horizontal sec/div scale may have variable timing, allowing you to set the horizontal time scale between the discrete settings.

Time Base Selections

Your oscilloscope has a time base, which is usually referred to as the main time base. Many oscilloscopes also have what is called a delayed time base – a time base with a sweep that can start (or be triggered to start) relative to a pre-determined time on the main time base sweep. Using a delayed time base sweep allows you to see events more clearly and to see events that are not visible solely with the main time base sweep.

The delayed time base requires the setting of a time delay and the possible use of delayed trigger modes and other settings not described in this primer. Refer to the manual supplied with your oscilloscope for information on how to use these features.

Zoom/Pan

Your oscilloscope may have special horizontal magnification settings that let you display a magnified section of the waveform on-screen. Some oscilloscopes add pan functions to the zoom capability. Knobs are used to adjust zoom factor or scale and the pan of the zoom box across the waveform.

Search

Some oscilloscopes offer search and mark capabilities, enabling you to quickly navigate through long acquisitions looking for user-defined events.

XY Mode

Most oscilloscopes have an XY mode that lets you display an input signal, rather than the time base, on the horizontal axis. This mode of operation opens up a whole new area of phase shift measurement techniques, explained in the Measurement Techniques section of this primer.

Z Axis

A digital phosphor oscilloscope (DPO) has a high display sample density and an innate ability to capture intensity information. With its intensity axis (Z axis), the DPO is able to provide a three-dimensional, real-time display similar to that of an analog oscilloscope. As you look at the waveform trace on a DPO, you can see brightened areas – the areas where a signal occurs most often. This display makes it easy to distinguish the basic signal shape from a transient that occurs only once in a while – the basic signal would appear much brighter. One application of the Z axis is to feed special timed signals into the separate Z input to create highlighted "marker" dots at known intervals in the waveform.

XYZ Mode with DPO and XYZ Record Display

Some DPOs can use the Z input to create an XY display with intensity grading. In this case, the DPO samples the instantaneous data value at the Z input and uses that value to qualify a specific part of the waveform. Once you have qualified samples, these samples can accumulate, resulting in an intensity-graded XYZ display. XYZ mode is especially useful for displaying the polar patterns commonly used in testing wireless communication devices – a constellation diagram, for example. Another method of displaying XYZ data is XYZ record display. In this mode the data from the acquisition memory is used rather than the DPO database.

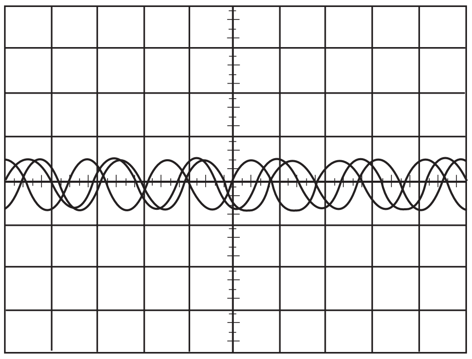

Figure 32. Untriggered display.

Trigger System and Controls

An oscilloscope's trigger function synchronizes the horizontal sweep at the correct point of the signal, essential for clear signal characterization. Trigger controls allow you to stabilize repetitive waveforms and capture single-shot waveforms.

The trigger makes repetitive waveforms appear static on the oscilloscope display by repeatedly displaying the same portion of the input signal. Imagine the jumble on the screen that would result if each sweep started at a different place on the signal, as illustrated in Figure 32.

Edge triggering, available in analog and digital oscilloscopes, is the basic and most common type. In addition to threshold triggering offered by both analog and digital oscilloscopes, many digital oscilloscopes offer numerous specialized trigger settings not offered by analog instruments. These triggers respond to specific conditions in the incoming signal, making it easy to detect, for example, a pulse that is narrower than it should be. Such a condition would be impossible to detect with a voltage threshold trigger alone.

Advanced trigger controls enable you to isolate specific events of interest to optimize the oscilloscope's sample rate and record length. Advanced triggering capabilities in some oscilloscopes give you highly selective control. You can trigger on pulses defined by amplitude (such as runt pulses), qualified by time (pulse width, glitch, slew rate, setup-and-hold, and time-out), and delineated by logic state or pattern (logic triggering).

Other advanced trigger functions include:

- **Pattern Lock Triggering:** Pattern lock triggering adds a new dimension to NRZ serial pattern triggering by enabling the oscilloscope to take synchronized acquisitions of a long serial test pattern with outstanding time base accuracy. Pattern lock triggering can be used to remove random jitter from long serial data patterns. Effects of specific bit transitions can be investigated, and averaging can be used with mask testing.
- **Serial Pattern Triggering:** Serial pattern triggering can be used to debug serial architectures. It provides a trigger on the serial pattern of an NRZ serial data stream with built-in clock recovery and correlates events across the physical and link layer. The instrument can recover the clock signal, identify transitions, and allow you to set the desired encoded words for the serial pattern trigger to capture.
- **A & B Triggering:** Some trigger systems offer multiple trigger types only on a single event (A event), with delayed trigger (B event) selection limited to edge type triggering and often do not provide a way to reset the trigger sequence if the B event doesn't occur. Modern oscilloscopes can provide the full suite of advanced trigger types on both A and B triggers, logic qualification to control when to look for these events, and reset triggering to begin the trigger sequence again after a specified time, state, or transition so that even events in the most complex signals can be captured.
- **Search & Mark Triggering:** Hardware triggers watch for one event type at a time, but Search can scan for multiple event types simultaneously. For example, scan for setup or hold time violations on multiple channels. Individual marks can be placed by Search indicating events that meet search criteria.
- **Trigger Correction:** Since the trigger and data acquisition systems share different paths there is some inherent time delay between the trigger position and the data acquired. This results in skew and trigger jitter. With a trigger correction system the instrument adjusts the trigger position and compensates for the difference of delay there is between the trigger path and the data acquisition path. This will eliminate virtually any trigger jitter at the trigger point. In this mode, the trigger point can be used as a measurement reference.

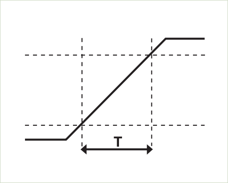

Slew Rate Triggering. High frequency signals with slew rates faster than expected or needed can radiate troublesome energy. Slew rate triggering surpasses conventional edge triggering by adding the element of time and allowing you to selectively trigger on fast or slow edges.

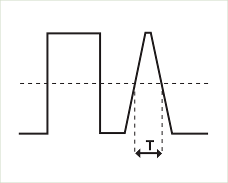

Glitch Triggering. Glitch triggering allows you to trigger on digital pulses when they are shorter or longer than a user-defined time limit. This trigger control enables you to examine the causes of even rare glitches and their effects on other signals

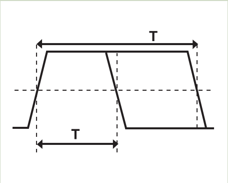

Pulse Width Triggering. Using pulse width triggering, you can monitor a signal indefinitely and trigger on the first occurrence of a pulse whose duration (pulse width) is outside the allowable limits.

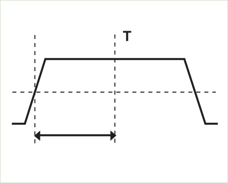

Time-out Triggering. Time-out triggering lets you trigger on an event without waiting for the trigger pulse to end, by triggering based on a specified time lapse.

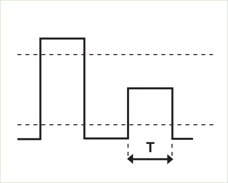

Runt Pulse Triggering. Runt triggering allows you to capture and examine pulses that cross one logic threshold, but not both.

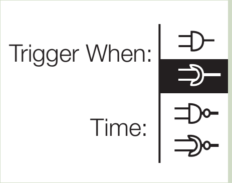

Logic Triggering. Logic triggering allows you to trigger on any logical combination of available input channels – especially useful in verifying the operation of digital logic.

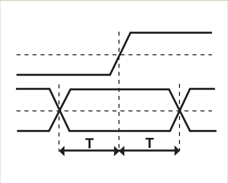

Setup-and-Hold Triggering. Only setup-and-hold triggering lets you deterministically trap a single violation of setup-and-hold time that would almost certainly be missed by using other trigger modes. This trigger mode makes it easy to capture specific signal quality and timing details when a synchronous data signal fails to meet setup-and-hold specifications.

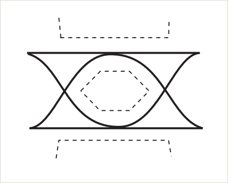

Communication Triggering. Optionally available on certain oscilloscope models, these trigger modes address the need to acquire a wide variety of Alternate-Mark Inversion (AMI), Code-Mark Inversion (CMI), and Non-Return to Zero (NRZ) communication signals.

Figure 33. Common trigger types.

- **Serial Triggering on Specific Standard Signals I²C, CAN, LIN, etc.)** - Some oscilloscopes provide the ability to trigger on specific signal types for standard serial data signals such as CAN, LIN, I²C, SPI, and others. The decode of these signal types is also available on many oscilloscopes today.
- **Parallel Bus Triggering** - Multiple parallel buses can be defined and displayed at one time to easily view decoded parallel bus data over time. By specifying which channels are the clock and data lines, you can create a parallel bus display on some oscilloscopes that automatically decodes bus content. Countless hours can be saved by using parallel bus triggers to simplify capture and analysis.

Optional trigger controls in some oscilloscopes are designed specifically to examine communications signals as well. Figure 33 highlights a few of these common trigger types in more detail. The intuitive user interface available in some oscilloscopes also allows rapid setup of trigger parameters with wide flexibility in the test setup to maximize your productivity.

Trigger Position

Horizontal trigger position control is only available on digital oscilloscopes. The trigger position control may be located in the horizontal control section of your oscilloscope. It actually represents the horizontal position of the trigger in the waveform record.

Varying the horizontal trigger position allows you to capture what a signal did before a trigger event, known as pre-trigger viewing. Thus, it determines the length of viewable signal both preceding and following a trigger point.

Digital oscilloscopes can provide pre-trigger viewing because they constantly process the input signal, whether or not a trigger has been received. A steady stream of data flows through the oscilloscope; the trigger merely tells the oscilloscope to save the present data in memory.
In contrast, analog oscilloscopes only display the signal – that is, write it on the CRT – after receiving the trigger. Thus, pre-trigger viewing is not available in analog oscilloscopes, with the exception of a small amount of pre-trigger provided by a delay line in the vertical system.

Pre-trigger viewing is a valuable troubleshooting aid. If a problem occurs intermittently, you can trigger on the problem, record the events that led up to it and, possibly, find the cause.

Trigger Level and Slope

The trigger level and slope controls provide the basic trigger point definition and determine how a waveform is displayed, as illustrated in Figure 34.

The trigger circuit acts as a comparator. You select the slope and voltage level on one input of the comparator. When the trigger signal on the other comparator input matches your settings, the oscilloscope generates a trigger.

The slope control determines whether the trigger point is on the rising or the falling edge of a signal. A rising edge is a positive slope and a falling edge is a negative slope. The level control determines where on the edge the trigger point occurs.

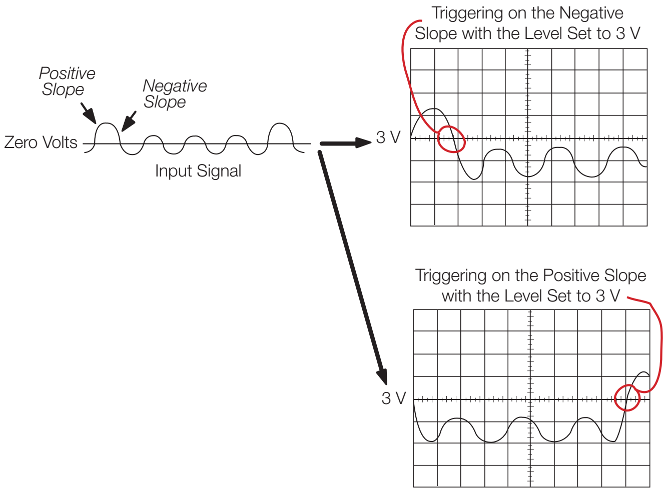
Triggering on the Negative Slope with the Level Set to 3 V
Positive Slope
Negative Slope
Zero Volts
Input Signal
3 V
Triggering on the Positive Slope with the Level Set to 3 V
3 V

Figure 34. Positive and negative slope triggering.

Trigger Sources

The oscilloscope does not necessarily need to trigger on the signal being displayed. Several sources can trigger the sweep:

- Any input channel
- An external source other than the signal applied to an input channel
- The power source signal
- A signal internally defined by the oscilloscope, from one or more input channels

Most of the time, you can leave the oscilloscope set to trigger on the channel displayed. Some oscilloscopes provide a trigger output that delivers the trigger signal to another instrument.

The oscilloscope can use an alternate trigger source, whether or not it is displayed, so you should be careful not to unwittingly trigger on channel 1 while displaying channel 2, for example.

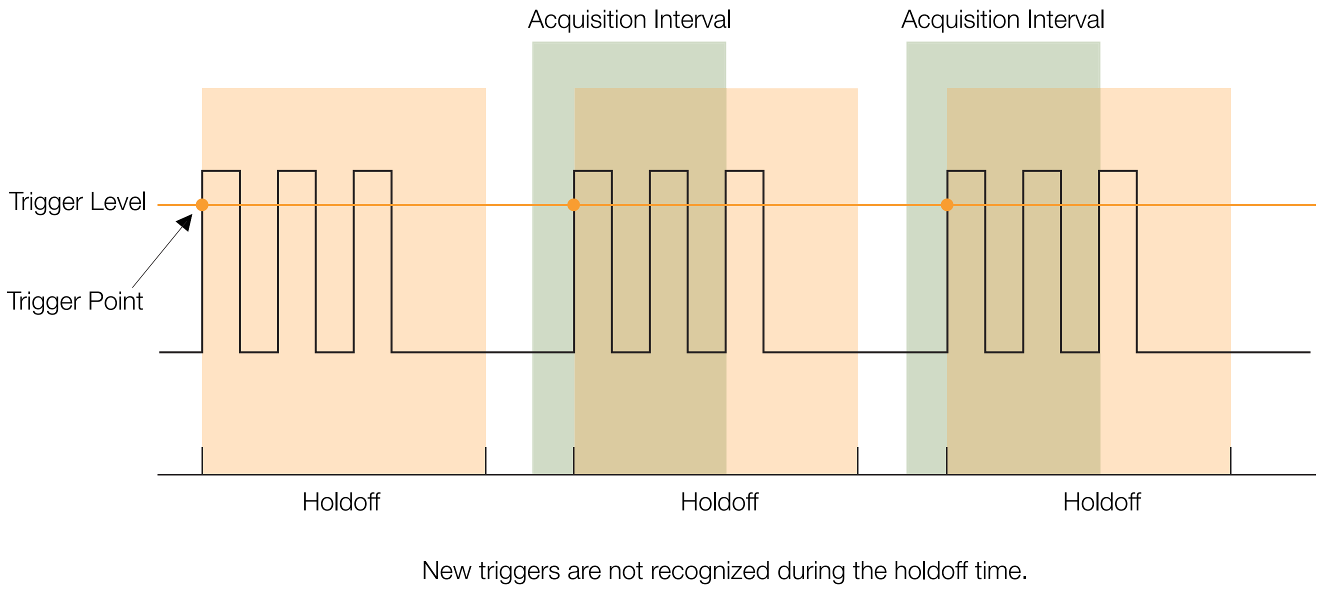

Figure 35. Trigger holdoff.

Trigger Modes

The trigger mode determines whether or not the oscilloscope draws a waveform based on a signal condition. Common trigger modes include normal and auto.

In normal mode the oscilloscope only sweeps if the input signal reaches the set trigger point; otherwise (on an analog oscilloscope) the screen is blank or (on a digital oscilloscope) frozen on the last acquired waveform. Normal mode can be disorienting since you may not see the signal at first if the level control is not adjusted correctly.

Auto mode causes the oscilloscope to sweep, even without a trigger. If no signal is present, a timer in the oscilloscope triggers the sweep. This ensures that the display will not disappear if the signal does not cause a trigger.

In practice, you will probably use both modes: normal mode because it lets you see just the signal of interest, even when triggers occur at a slow rate, and auto mode because it requires less adjustment.

Many oscilloscopes also include special modes for single sweeps, triggering on video signals, or automatically setting the trigger level.

Trigger Coupling

Just as you can select either AC or DC coupling for the vertical system, you can choose the kind of coupling for the trigger signal.

Besides AC and DC coupling, your oscilloscope may also have high frequency rejection, low frequency rejection, and noise rejection trigger coupling. These special settings are useful for eliminating noise from the trigger signal to prevent false triggering.

Trigger Holdoff

Sometimes getting an oscilloscope to trigger on the correct part of a signal requires great skill. Many oscilloscopes have special features to make this task easier.

Trigger holdoff is an adjustable period of time after a valid trigger during which the oscilloscope cannot trigger.
This feature is useful when you are triggering on complex waveform shapes, so that the oscilloscope only triggers on an eligible trigger point. Figure 35 shows how using trigger holdoff helps create a usable display.

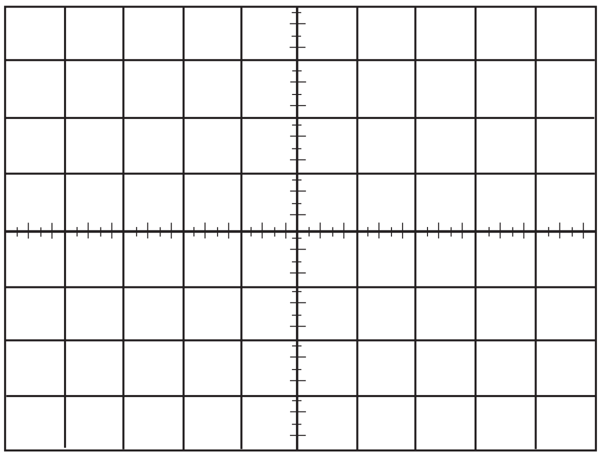

Figure 36. An oscilloscope graticule.

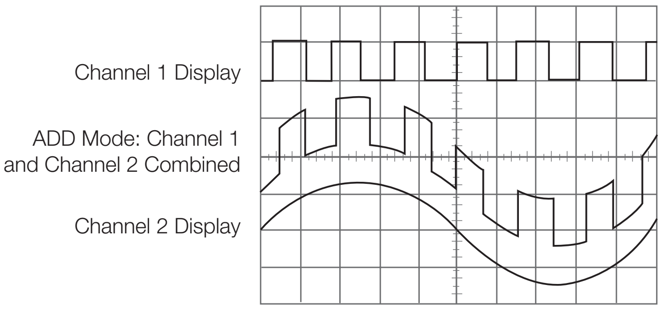

Figure 37. Adding channels.

Display System and Controls

An oscilloscope's front panel includes a display screen and the knobs, buttons, switches, and indicators used to control signal acquisition and display. As mentioned at the front of this section, front-panel controls are usually divided into vertical, horizontal and trigger sections. The front panel also includes input connectors.

Take a look at the oscilloscope display. Notice the grid markings on the screen – these markings create the graticule. Each vertical and horizontal line constitutes a major division. The graticule is usually laid out in an 8-by-10 or 10-by-10 division pattern. Labeling on the oscilloscope controls (such as volts/div and sec/div) always refers to major divisions. The tick marks on the center horizontal and vertical graticule lines, as shown in Figure 36, are called minor divisions. Many oscilloscopes display on the screen how many volts each vertical division represents and how many seconds each horizontal division represents.

Other Oscilloscope Controls

Math and Measurement Operations

Your oscilloscope may also have operations that allow you to add waveforms together, creating a new waveform display. Analog oscilloscopes combine the signals while digital oscilloscopes create new waveforms mathematically. Subtracting waveforms is another math operation. Subtraction with analog oscilloscopes is possible by using the channel invert function on one signal and then using the add operation. Digital oscilloscopes typically have a subtraction operation available. Figure 37 illustrates a third waveform created by combining two different signals.

Using the power of their internal processors, digital oscilloscopes offer many advanced math operations: multiplication, division, integration, Fast Fourier Transform, and more. This advanced signal processing capability can also perform functions such as the insertion of a filter block which can be used de-embed the characteristics of the fixture on the device under test or implement a filter block with desired frequency response such as a low pass filter. The processing block is flexible – not dedicated; it can perform as an arbitrary filter instead, for example for simulation of pre-emphasis/de-emphasis schemes.

Digital Timing and State Acquisitions

Digital channels provided by a mixed signal oscilloscope enable acquisition capabilities similar to those found on logic analyzers. There are two major digital acquisition techniques. The first technique is timing acquisition in which the MSO samples the digital signal at uniformly spaced times determined by the MSO's sample rate. At each sample point, the MSO stores the signal's logic state and creates a timing diagram of the signal. The second digital acquisition technique is state acquisition. State acquisition defines special times that the digital signal's logic state is valid and stable. This is common in synchronous and clocked digital circuits. A clock signal defines the time when the signal state is valid. For example, the input signal stable time is around the rising clock edge for a D-Flip-Flop with rising edge clocking. The output signal stable time is around the falling clock edge for a D-Flip-Flop with rising edge clocking. Since the clock period of a synchronous circuit may not be fixed, the time between state acquisitions may not be uniform as it is in a timing acquisition.

A mixed signal oscilloscope's digital channels acquire signals similar to how a logic analyzer acquires signals in timing acquisition mode. The MSO then decodes the timing acquisition into a clocked bus display, and event table which is similar to the logic analyzer's state acquisition display, providing you with important information during debug.

We have described the basic oscilloscope controls that a beginner needs to know about. Your oscilloscope may have other controls for various functions. Some of these may include:

- Automatic parametric measurements
- Measurement cursors
- Keypads for mathematical operations or data entry
- Printing capabilities
- Interfaces for connecting your oscilloscope to a computer or directly to the Internet

Look over the other options available to you and read your oscilloscope's manual to find out more about these other controls.

The Complete Measurement System

Probes

Even the most advanced instrument can only be as precise as the data that goes into it. A probe functions in conjunction with an oscilloscope as part of the measurement system. Precision measurements start at the probe tip. The right probes matched to the oscilloscope and the device-under-test (DUT) not only allow the signal to be brought to the oscilloscope cleanly, they also amplify and preserve the signal for the greatest signal integrity and measurement accuracy.

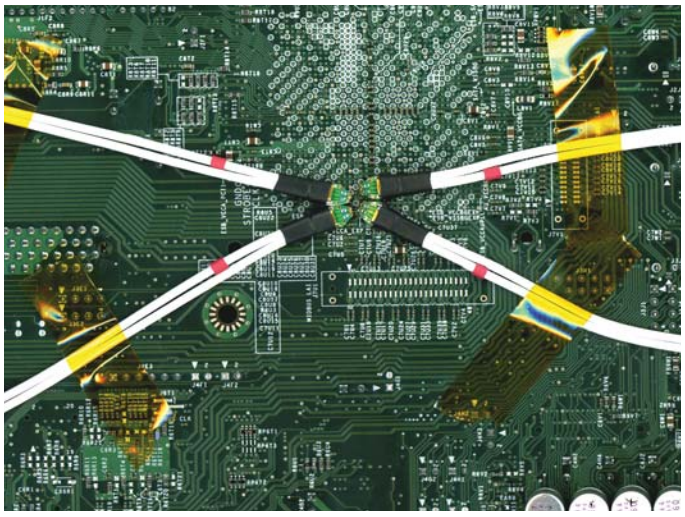

Figure 38. Dense devices and systems require small form factor probes.

To ensure accurate reconstruction of your signal, try to choose a probe that, when paired with your oscilloscope, exceeds the signal bandwidth by 5 times.

Probes actually become part of the circuit, introducing resistive, capacitive and inductive loading that inevitably alters the measurement. For the most accurate results, the goal is to select a probe with minimal loading. An ideal pairing of the probe with the oscilloscope will minimize this loading, and enable you to access all of the power, features and capabilities of your oscilloscope.

Another connectivity consideration is the probe's form factor. Small form factor probes provide easier access to today's densely packed circuitry, as shown in Figure 38.

A brief description of the types of probes follows. Please refer to the Tektronix ABCs of Probes primer for more information about this essential component of the overall measurement system.

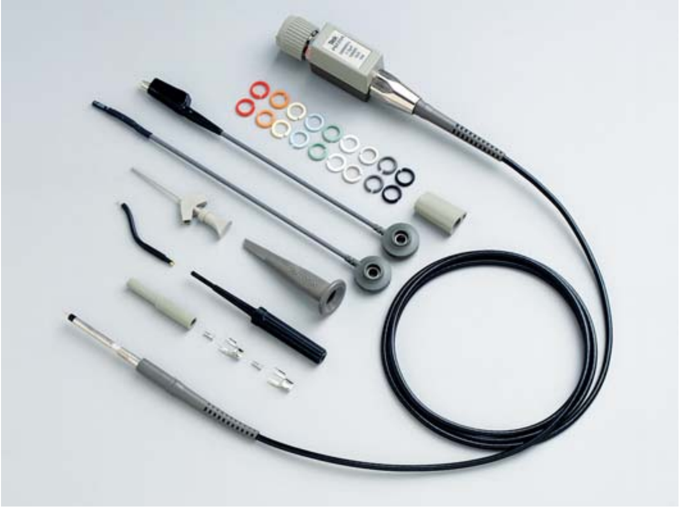

Figure 39. A typical passive probe with accessories.

Passive Probes

For measuring typical signal and voltage levels, passive probes provide ease-of-use and a wide range of measurement capabilities at an affordable price. The pairing of a passive voltage probe with a current probe will provide you with an ideal solution for measuring power.

Most passive probes have some attenuation factor, such as 10X, 100X, and so on. By convention, attenuation factors, such as for the 10X attenuator probe, have the X after the factor. In contrast, magnification factors like X10 have the X first.

The 10X (read as "ten times") attenuator probe reduces circuit loading in comparison to a 1X probe and is an excellent general-purpose passive probe. Circuit loading becomes more pronounced for higher frequency and/or higher impedance signal sources, so be sure to analyze these signal/probe loading interactions before selecting a probe. The 10X attenuator probe improves the accuracy of your measurements, but also reduces the signal's amplitude at the oscilloscope input by a factor of 10.

Because it attenuates the signal, the 10X attenuator probe makes it difficult to look at signals less than 10 millivolts peak-to-peak. The 1X probe is similar to the 10X attenuator probe but lacks the attenuation circuitry. Without this circuitry, more interference is introduced to the circuit being tested.

Use the 10X attenuator probe as your general-purpose probe, but keep the 1X probe accessible to measure slow-speed, low-amplitude signals. Some probes have a convenient feature for switching between 1X and 10X attenuation at the probe tip. If your probe has this feature, make sure you are using the correct setting before taking measurements.

Many oscilloscopes can automatically detect whether you are using a 1X or 10X probe and adjust their screen readouts accordingly. However with some oscilloscopes, you must set the type of probe you are using or read from the proper 1X or 10X marking on the volts/div control.

The 10X attenuator probe works by balancing the probe's electrical properties against the oscilloscope's electrical properties. Before using a 10X attenuator probe you need to adjust this balance for your particular oscilloscope. This adjustment is known as compensating the probe and is described in more detail in the Operating the Oscilloscope section of this primer.

Passive probes, such as the one shown in Figure 39, provide excellent general-purpose probing solutions. However, general-purpose passive probes cannot accurately measure signals with extremely fast rise times, and may excessively load sensitive circuits. The steady increase in signal clock rates and edge speeds demands higher speed probes with less loading effects. High-speed active and differential probes provide ideal solutions when measuring high-speed and/or differential signals.

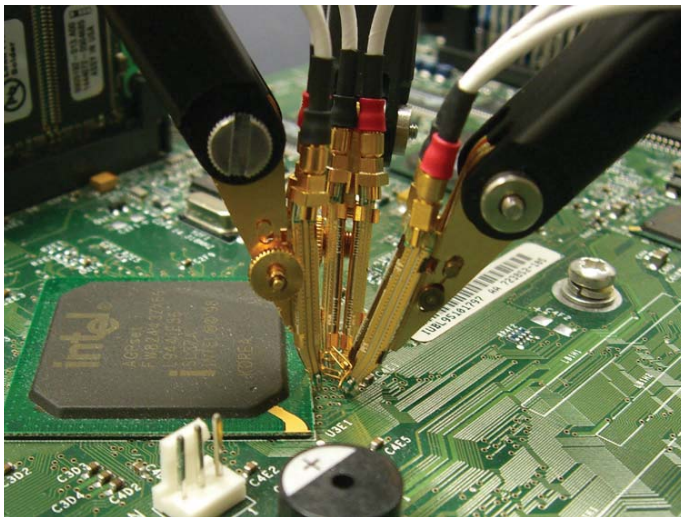

Figure 40. High-performance probes are critical when measuring the fast clocks and edges found in today's computer buses and data transmission lines.

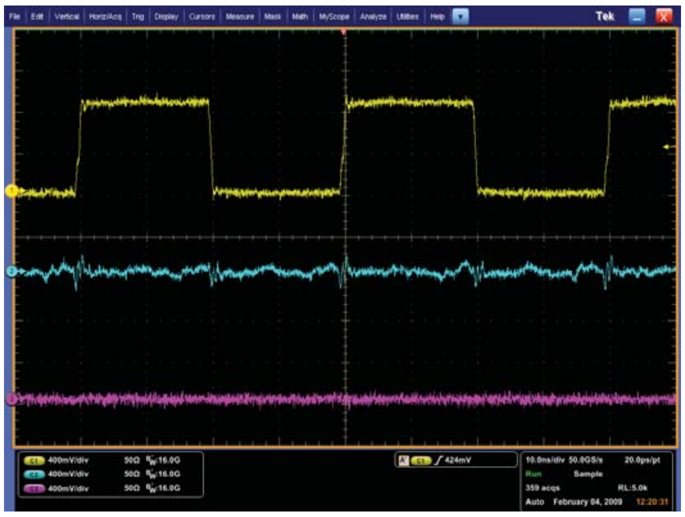

C1 400mV/div 50Ω B_W:16.0G
C2 400mV/div 50Ω B_W:16.0G
C3 400mV/div 50Ω B_W:16.0G

A' C1 ∫ 424mV

10.0ns/div 50.0GS/s 20.0ps/pt
Run Sample
359 acqs RL:5.0k
Auto February 04, 2009 12:20:31

Figure 41 Differential probes can separate common-mode noise from the signal content of interest in today's fast, low-voltage applications - especially important as digital signals continue to fall below typical noise thresholds found in integrated circuits.

Active and Differential Probes

Increasing signal speeds and lower-voltage logic families make accurate measurement results difficult to achieve. Signal fidelity and device loading are critical issues. A complete measurement solution at these high speeds includes high-speed, high-fidelity probing solutions to match the performance of the oscilloscope, as shown in Figure 40.

Active and differential probes use specially developed integrated circuits to preserve the signal during access and transmission to the oscilloscope, ensuring signal integrity. For measuring signals with fast rise times, a high-speed active or differential probe will provide more accurate results, as highlighted in Figure 41.

Newer probe types provide the advantage of being able to use one setup, and get three types of measurements without adjusting probe tip connections. These probes can make differential, single-ended and common mode measurements from the same probe setup.

Logic Probes

The logic probe shown in Figure 42 offers two eight-channel pods. Each channel ends with a probe tip featuring a recessed ground for simplified connection to the device-under-test. The coax on the first channel of each pod is colored blue making it easy to identify. The common ground uses an automotive style connector making it easy to create custom grounds for connecting to the device-under-test. When connecting to square pins, you can use an adapter that attaches to the probe head extending the probe ground flush with the probe tip so you can attach to a header. These probes offer outstanding electrical characteristics with minimal capacitive loading.

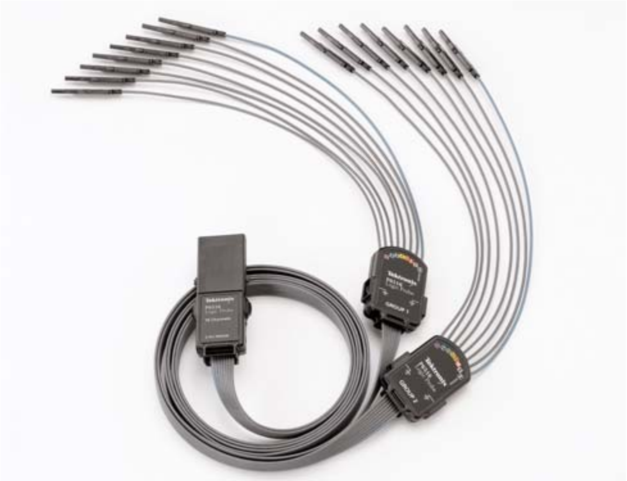

Figure 42. Logic probes for a mixed signal oscilloscope (MSO) simplify digital connectivity to your device.

Specialty Probes

In addition to the previously mentioned probe types, there are also a variety of other specialty probes and probing systems. These include current, high-voltage, and optical probes, just to name a few.

Probe Accessories

Many modern oscilloscopes provide special automated features built into the input and mating probe connectors. In the case of intelligent probe interfaces, the act of connecting the probe to the instrument notifies the oscilloscope about the probe's attenuation factor, which in turn scales the display so that the probe's attenuation is figured into the readout on the screen. Some probe interfaces also recognize the type of probe – that is, passive, active or current. The interface may act as a DC power source for probes. Active probes have their own amplifier and buffer circuitry that requires DC power.

Ground lead and probe tip accessories are also available to improve signal integrity when measuring high-speed signals. Ground lead adapters provide spacing flexibility between probe tip and ground lead connections to the DUT, while maintaining very short lead lengths from probe tip to DUT.

Please refer to the Tektronix ABCs of Probes primer for more information about probes and probe accessories.

Performance Terms and Considerations

As previously mentioned, an oscilloscope is analogous to a camera that captures signal images that we can observe and interpret. Shutter speed, lighting conditions, aperture and the ASA rating of the film all affect the camera's ability to capture an image clearly and accurately.

Like the basic systems of an oscilloscope, the performance considerations of an oscilloscope significantly affect its ability to achieve the required signal integrity.

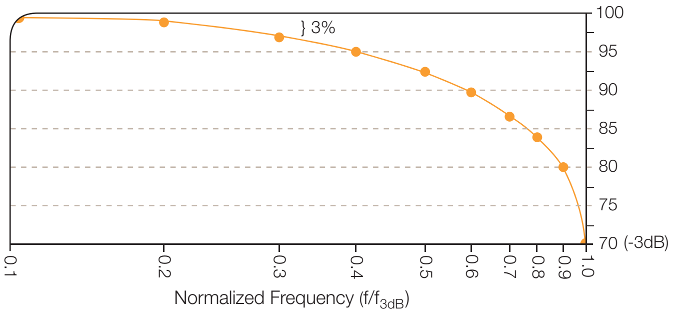

Figure 43. Oscilloscope bandwidth is the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the -3 dB point.

Learning a new skill often involves learning a new vocabulary. This idea holds true for learning how to use an oscilloscope. This section describes some useful measurement and oscilloscope performance terms. These terms are used to describe the criteria essential to choosing the right oscilloscope for your application. Understanding these terms will help you to evaluate and compare your oscilloscope with other models.

Bandwidth

Bandwidth determines an oscilloscope's fundamental ability to measure a signal. As signal frequency increases, the capability of the oscilloscope to accurately display the signal decreases. This specification indicates the frequency range that the oscilloscope can accurately measure.

Oscilloscope bandwidth is specified as the frequency at which a sinusoidal input signal is attenuated to 70.7% of the signal's true amplitude, known as the –3 dB point, a term based on a logarithmic scale, as seen in Figure 43.

Without adequate bandwidth, your oscilloscope will not be able to resolve high-frequency changes. Amplitude will be distorted. Edges will vanish. Details will be lost. Without adequate bandwidth, all the features, bells and whistles in your oscilloscope will mean nothing.

To determine the oscilloscope bandwidth needed to accurately characterize signal amplitude in your specific application, apply the "5 Times Rule."

$$\text{Oscilloscope Bandwidth} \geq \frac{\text{Highest Frequency}}{\text{Component of Signal}} \; x \; 5$$

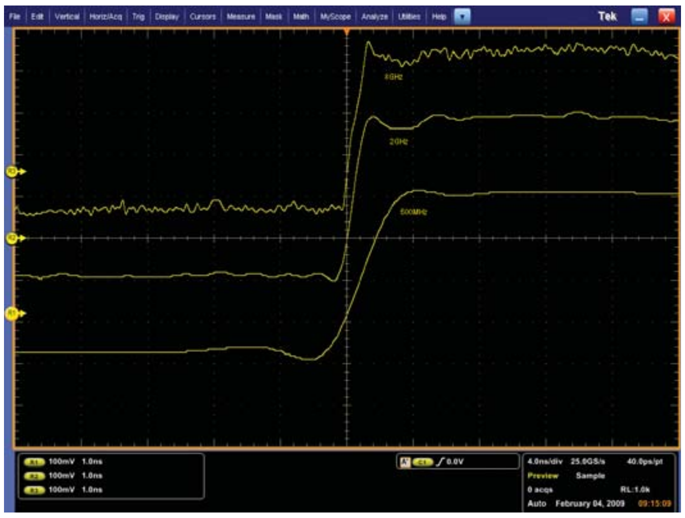
File Edit Vertical Horiz/Acq Trig Display Cursors Measure Mask Math MyScope Analyze Utilities Help
Tek
4GHz
2GHz
500MHz
R1 100mV 1.0ns
R2 100mV 1.0ns
R3 100mV 1.0ns
C1 0.0V
4.0ns/div 25.0GS/s 40.0ps/pt
Preview Sample
0 acqs RL:1.0k
Auto February 04, 2009 09:15:09

Figure 44. The higher the bandwidth, the more accurate the reproduction of your signal, as illustrated with a signal captured at 250 MHz, 1 GHz and 4 GHz bandwidth levels.

An oscilloscope selected using the 5 Times Rule will give you less than ±2% error in your measurements – typically sufficient for today's applications. However, as signal speeds increase, it may not be possible to achieve this rule of thumb. Always keep in mind that higher bandwidth will likely provide more accurate reproduction of your signal, as demonstrated in Figure 44.

Some oscilloscopes provide a method of enhancing the bandwidth through digital signal processing. A DSP arbitrary equalization filter can be used to improve the oscilloscope channel response. This filter extends the bandwidth, flattens the oscilloscope's channel frequency response, improves phase linearity, and provides a better match between channels. It also decreases rise time and improves the time-domain step response.

Rise Time

In the digital world, rise time measurements are critical. Rise time may be a more appropriate performance consideration when you expect to measure digital signals, such as pulses and steps. As shown in Figure 45, your oscilloscope must have sufficient rise time to accurately capture the details of rapid transitions.

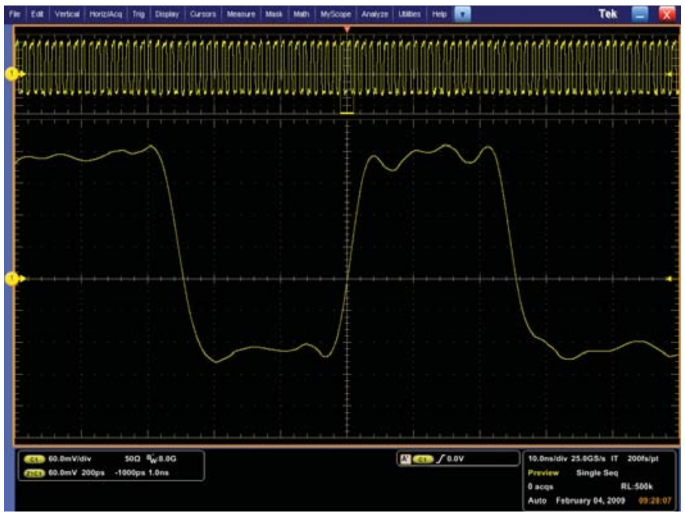
File Edit Vertical Horiz/Acq Trig Display Cursors Measure Mask Math MyScope Analyze Utilities Help
Tek
C1 60.0mV/div 50Ω BW:8.0G
Z1(C1) 60.0mV 200ps -1000ps 1.0ns
C1 0.0V
10.0ns/div 25.0GS/s IT 200fs/pt
Preview Single Seq
0 acqs RL:500k
Auto February 04, 2009 09:28:07

Figure 45. Rise time characterization of a high-speed digital signal.

Rise time describes the useful frequency range of an oscilloscope. To calculate the oscilloscope rise time required for your signal type, use the following equation:

$$\text{Oscilloscope Rise Time} \leq \text{Fastest Rise Time of Signal} \ x \ \frac{1}{5}$$

Note that this basis for oscilloscope rise time selection is similar to that for bandwidth. As in the case of bandwidth, achieving this rule of thumb may not always be possible given the extreme speeds of today's signals. Always remember that an oscilloscope with faster rise time will more accurately capture the critical details of fast transitions.

In some applications, you may know only the rise time of a signal. A constant allows you to relate the bandwidth and rise time of the oscilloscope, using the equation:

$$\text{Bandwidth} = \frac{k}{\text{Rise Time}}$$

where k is a value between 0.35 and 0.45, depending on the shape of the oscilloscope's frequency response curve and pulse rise time response. Oscilloscopes with a bandwidth of <1 GHz typically have a 0.35 value, while oscilloscopes with a bandwidth of > 1 GHz usually have a value between 0.40 and 0.45.

Logic Family	Typical Signal Rise Time	Calculated Signal Bandwidth
TTL	2 ns	175 MHz
CMOS	1.5 ns	230 MHz
GTL	1 ns	350 MHz
LVDS	400 ps	875 MHz
ECL	100 ps	3.5 GHz
GaAs	40 ps	8.75 GHz

Figure 46. Some logic families produce inherently faster rise times than others.

Some logic families produce inherently faster rise times than others, as illustrated in Figure 46.

Sample Rate

Sample rate – specified in samples per second (S/s) – refers to how frequently a digital oscilloscope takes a snapshot or sample of the signal, analogous to the frames on a movie camera. The faster an oscilloscope samples (i.e., the higher the sample rate), the greater the resolution and detail of the displayed waveform and the less likely that critical information or events will be lost, as shown in Figure 47. The minimum sample rate may also be important if you need to look at slowly changing signals over longer periods of time. Typically, the displayed sample rate changes with changes made to the horizontal scale control to maintain a constant number of waveform points in the displayed waveform record.

How do you calculate your sample rate requirements? The method differs based on the type of waveform you are measuring, and the method of signal reconstruction used by the oscilloscope.

In order to accurately reconstruct a signal and avoid aliasing, the Nyquist theorem states that the signal must be sampled at least twice as fast as its highest frequency component. This theorem, however, assumes an infinite record length and a continuous signal. Since no oscilloscope offers infinite record length and, by definition, glitches are not continuous, sampling at only twice the rate of highest frequency component is usually insufficient.

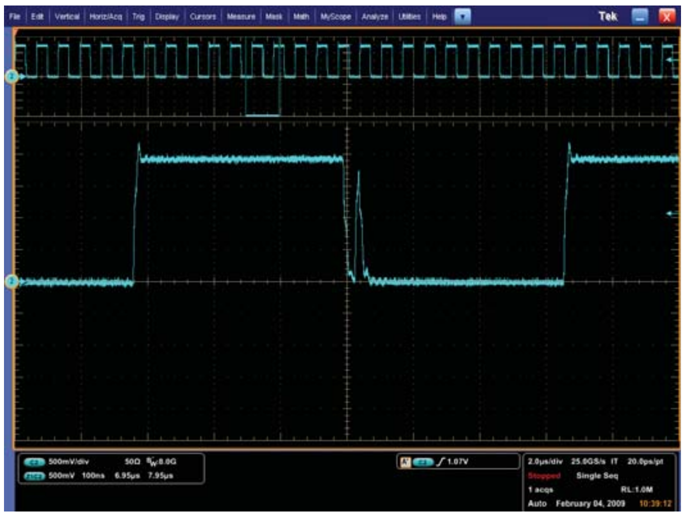

Figure 47. A higher sample rate provides greater signal resolution, ensuring that you'll see intermittent events.

In reality, accurate reconstruction of a signal depends on both the sample rate and the interpolation method used to fill in the spaces between the samples. Some oscilloscopes let you select either sin (x)/x interpolation for measuring sinusoidal signals, or linear interpolation for square waves, pulses and other signal types.

> For accurate reconstruction using sin(x)/x interpolation, your oscilloscope should have a sample rate at least 2.5 times the highest frequency component of your signal. Using linear interpolation, the sample rate should be at least 10 times the highest frequency signal component.

Some measurement systems with sample rates to 50 GS/s and bandwidths to 20 GHz have been optimized for capturing very fast, single-shot and transient events by oversampling up to 5 times the bandwidth.

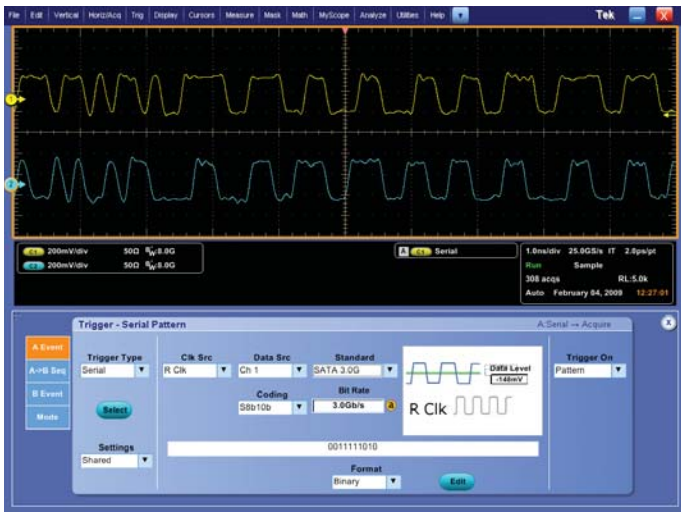
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
C1 200mV/div 50Ω BW:8.0G
C2 200mV/div 50Ω BW:8.0G
A C1 Serial
1.0ns/div 25.0GS/s IT 2.0ps/pt
Run Sample
308 acqs RL:5.0k
Auto February 04, 2009 12:27:01
A Event
A->B Seq
B Event
Mode
Select
Edit
R Clk

Trigger - Serial Pattern

A:Serial → Acquire

Trigger Type: Serial

Clk Src: R Clk

Data Src: Ch 1

Standard: SATA 3.0G

Coding: S8b10b

Bit Rate: 3.0Gb/s

Data Level: -148mV

Trigger On: Pattern

Settings: Shared

0011111010

Format: Binary

Figure 48. A DPO provides an ideal solution for non-repetitive, high-speed, multi-channel digital design applications.

Waveform Capture Rate

All oscilloscopes blink. That is, they open their eyes a given number of times per second to capture the signal, and close their eyes in between. This is the waveform capture rate, expressed as waveforms per second (wfms/s). While the sample rate indicates how frequently the oscilloscope samples the input signal within one waveform, or cycle, the waveform capture rate refers to how quickly an oscilloscope acquires waveforms.

Waveform capture rates vary greatly, depending on the type and performance level of the oscilloscope. Oscilloscopes with high waveform capture rates provide significantly more visual insight into signal behavior, and dramatically increase the probability that the oscilloscope will quickly capture transient anomalies such as jitter, runt pulses, glitches and transition errors.

Digital storage oscilloscopes (DSOs) employ a serial-processing architecture to capture from 10 to 5,000 wfms/s. Some DSOs provide a special mode that bursts multiple captures into long memory, temporarily delivering higher waveform capture rates followed by long processing dead times that reduce the probability of capturing rare, intermittent events.

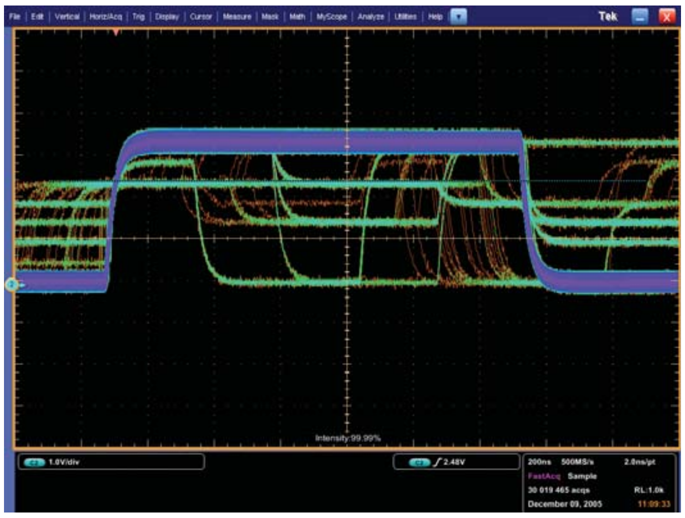
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursor
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
Intensity:99.99%
C1 1.0V/div
C1 2.48V
200ns 500MS/s 2.0ns/pt
FastAcq Sample
30 019 465 acqs RL:1.0k
December 09, 2005 11:09:33

Figure 49. A DPO enables a superior level of insight into signal behavior by delivering vastly greater waveform capture rates and three-dimensional display, making it the best general-purpose design and troubleshooting tool for a wide range of applications.

Most digital phosphor oscilloscopes (DPOs) employ a parallel-processing architecture to deliver vastly greater waveform capture rates. As seen in Figure 48, some DPOs can acquire millions of waveforms in just seconds, significantly increasing the probability of capturing intermittent and elusive events and allowing you to see the problems in your signal more quickly. Moreover, the DPO's ability to acquire and display three dimensions of signal behavior in real time – amplitude, time and distribution of amplitude over time – results in a superior level of insight into signal behavior, as shown in Figure 49.

Record Length

Record length, expressed as the number of points that comprise a complete waveform record, determines the amount of data that can be captured with each channel. Since an oscilloscope can store only a limited number of samples, the waveform duration (time) will be inversely proportional to the oscilloscope's sample rate.

$$\text{Time Interval} = \frac{\text{Record Length}}{\text{Sample Rate}}$$

Modern oscilloscopes allow you to select record length to optimize the level of detail needed for your application. If you are analyzing an extremely stable sinusoidal signal, you may need only a 500 point record length, but if you are isolating the causes of timing anomalies in a complex digital data stream, you may need a million points or more for a given record length, as demonstrated in Figure 50.

Triggering Capabilities

An oscilloscope's trigger function synchronizes the horizontal sweep at the correct point of the signal, essential for clear signal characterization. Trigger controls allow you to stabilize repetitive waveforms and capture single-shot waveforms.

Please refer to the Trigger section under Performance Terms and Considerations for more information regarding triggering capabilities.

Effective Bits

Effective bits represent a measure of a digital oscilloscope's ability to accurately reconstruct a sinewave signal's shape. This measurement compares the oscilloscope's actual error to that of a theoretical "ideal" digitizer. Because the actual errors include noise and distortion, the frequency and amplitude of the signal must be specified.

Frequency Response

Bandwidth alone is not enough to ensure that an oscilloscope can accurately capture a high frequency signal. The goal of oscilloscope design is a specific type of frequency response: Maximally Flat Envelope Delay (MFED). A frequency response of this type delivers excellent pulse fidelity with minimum overshoot and ringing. Since a digital oscilloscope is composed of real amplifiers, attenuators, ADCs, interconnects, and relays, MFED response is a goal that can only be approached. Pulse fidelity varies considerably with model and manufacturer.

Vertical Sensitivity

Vertical sensitivity indicates how much the vertical amplifier can amplify a weak signal – usually measured in millivolts (mV) per division. The smallest voltage detected by a general-purpose oscilloscope is typically about 1 mV per vertical screen division.

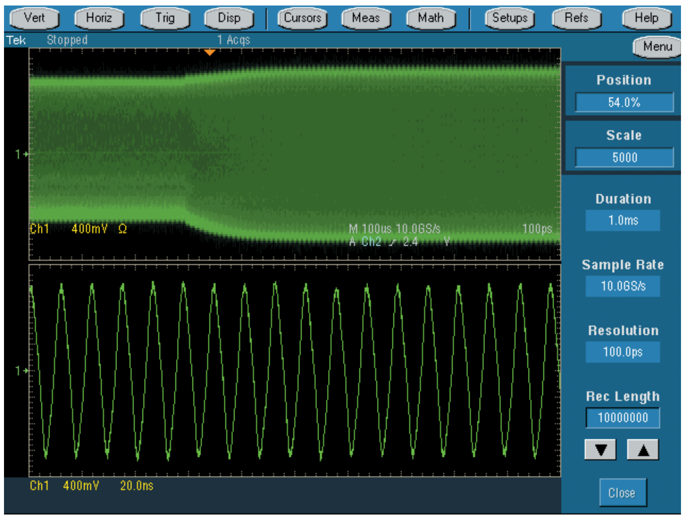

Position 54.0%
Scale 5000
Duration 1.0ms
Sample Rate 10.0GS/s
Resolution 100.0ps
Rec Length 10000000

Figure 50. Capturing the high frequency detail of this modulated 85 MHz carrier requires high resolution sampling (100 ps). Seeing the signal's complete modulation envelope requires a long time duration (1 ms). Using long record length (10 MB), the oscilloscope can display both.

Sweep Speed

Sweep speed indicates how fast the trace can sweep across the oscilloscope screen, enabling you to see fine details. The sweep speed of an oscilloscope is represented by time (seconds) per division.

Gain Accuracy

Gain accuracy indicates how accurately the vertical system attenuates or amplifies a signal, usually represented as a percentage error.

Horizontal Accuracy (Time Base)

Horizontal, or time base, accuracy indicates how accurately the horizontal system displays the timing of a signal, usually represented as a percentage error.

Vertical Resolution (Analog-to-Digital Converter)

Vertical resolution of the ADC, and therefore, the digital oscilloscope, indicates how precisely it can convert input voltages into digital values. Vertical resolution is measured in bits. Calculation techniques can improve the effective resolution, as exemplified with hi-res acquisition mode.

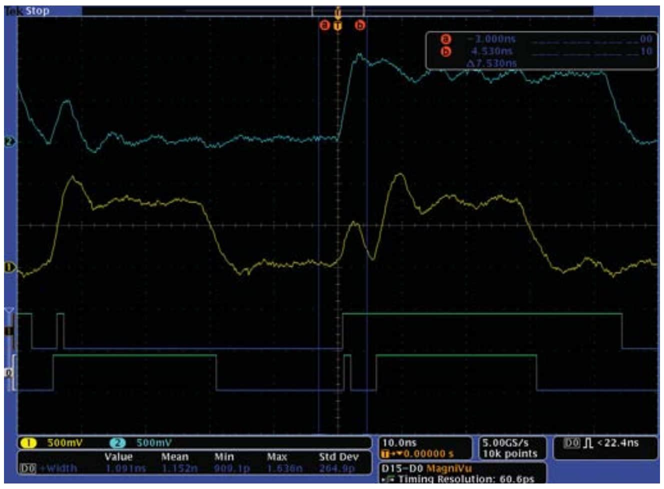
Tek Stop
a
T
b
a -3.000ns
b 4.530ns
Δ7.530ns
1 500mV
2 500mV
10.0ns
T▸▾0.00000 s
5.00GS/s
10k points
D0 ⎍ <22.4ns
D15–D0 MagniVu
Timing Resolution: 60.6ps

	Value	Mean	Min	Max	Std Dev
D0 +Width	1.091ns	1.152n	909.1p	1.636n	264.9p

Figure 51. The MSO provides 16 integrated digital channels, enabling the ability to view and analyze time-correlated analog and digital signals. A high speed timing acquisition provides more resolution to reveal narrow events such as glitches.

Timing Resolution (MSO)

An important MSO acquisition specification is the timing resolution used for capturing digital signals. Acquiring a signal with better timing resolution provides a more accurate timing measurement of when the signal changes. For example, a 500 MS/s acquisition rate has 2 ns timing resolution and the acquired signal edge uncertainty is 2 ns. A smaller timing resolution of 60.6 ps (16.5 GS/s) decreases the signal edge uncertainty to 60.6 ps and captures faster changing signals.

Some MSOs internally acquire digital signals with two types of acquisitions at the same time. The first acquisition is with standard timing resolution, and the second acquisition uses a high speed resolution. The standard resolution is used over a longer record length while the high speed timing acquisition offers more resolution around a narrow point of interest, as shown in Figure 51.

Connectivity

The need to analyze measurement results remains of utmost importance. The need to document and share information and measurement results easily and frequently has also grown in importance. The connectivity of an oscilloscope delivers advanced analysis capabilities and simplifies the documentation and sharing of results. As shown in Figure 52, standard interfaces (GPIB, RS-232, USB, Ethernet) and network communication modules enable some oscilloscopes to deliver a vast array of functionality and control.

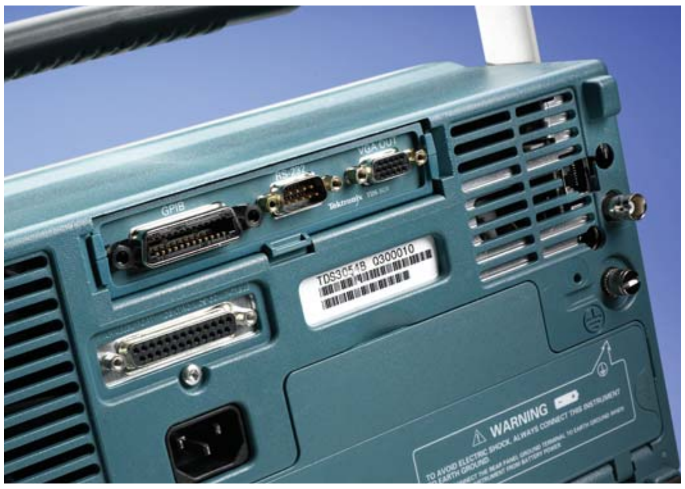
GPIB
RS-232
VGA OUT
Tektronix
TDS3054B 0300010
WARNING

Figure 52. Today's oscilloscopes provide a wide array of communications interfaces, such as a standard Centronics port and optional Ethernet/RS-232, GPIB/RS-232, and VGA/RS-232 modules. There is even a USB port (not shown) on the front panel.

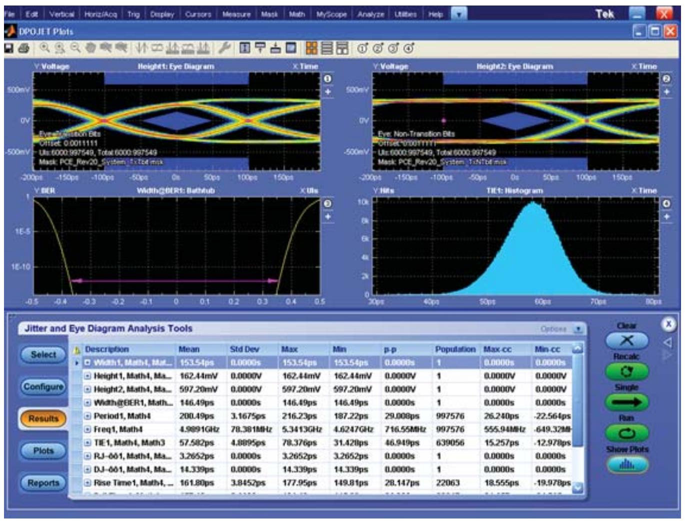
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
DPOJET Plots
Height1: Eye Diagram
Height2: Eye Diagram
Width@BER1: Bathtub
TIE1: Histogram
Y BER
X UIs
Y Hits
X Time
Select
Configure
Results
Plots
Reports
Options
Clear
Recalc
Single
Run
Show Plots

Jitter and Eye Diagram Analysis Tools

Description	Mean	Std Dev	Max	Min	p-p	Population	Max-cc	Min-cc
Width1, Math4, Mat...	153.54ps	0.0000s	153.54ps	153.54ps	0.0000s	1	0.0000s	0.0000s
Height1, Math4, Ma...	162.44mV	0.0000V	162.44mV	162.44mV	0.0000V	1	0.0000V	0.0000V
Height2, Math4, Ma...	597.20mV	0.0000V	597.20mV	597.20mV	0.0000V	1	0.0000V	0.0000V
Width@BER1, Math...	146.49ps	0.0000s	146.49ps	146.49ps	0.0000s	1	0.0000s	0.0000s
Period1, Math4	200.49ps	3.1675ps	216.23ps	187.22ps	29.008ps	997576	26.240ps	-22.564ps
Freq1, Math4	4.9891GHz	78.381MHz	5.3413GHz	4.6247GHz	716.55MHz	997576	555.94MHz	-649.32MH
TIE1, Math4, Math3	57.582ps	4.8895ps	78.376ps	31.428ps	46.949ps	639056	15.257ps	-12.978ps
RJ-δδ1, Math4, Ma...	3.2652ps	0.0000s	3.2652ps	3.2652ps	0.0000s	1	0.0000s	0.0000s
DJ-δδ1, Math4, Ma...	14.339ps	0.0000s	14.339ps	14.339ps	0.0000s	1	0.0000s	0.0000s
Rise Time1, Math4, ...	161.80ps	3.8452ps	177.95ps	149.81ps	28.147ps	22063	18.555ps	-19.978ps

Figure 53. Analysis software packages are specifically designed to meet jitter and eye measurement needs of today's high-speed digital designers.

Some advanced oscilloscopes also let you:

- Create, edit and share documents on the oscilloscope – all while working with the instrument in your particular environment
- Access network printing and file sharing resources
- Access the Windows® desktop
- Run third-party analysis and documentation software
- Link to networks
- Access the Internet
- Send and receive e-mail

Expandability

An oscilloscope should be able to accommodate your needs as they change. Some oscilloscopes allow you to:

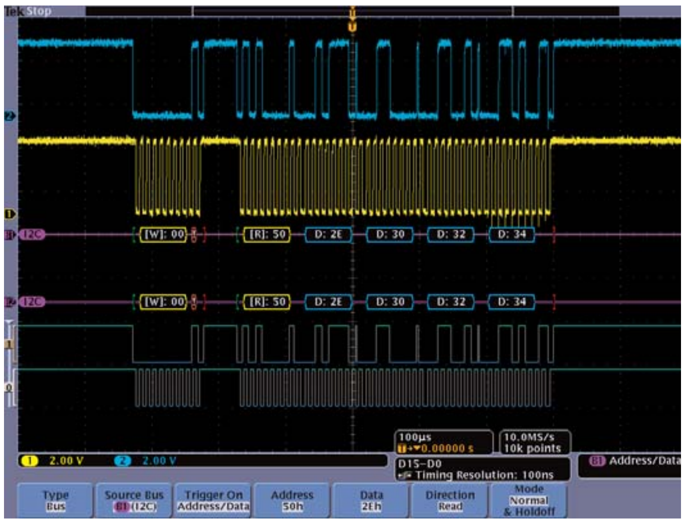
Stop
I2C
[W]: 00
[R]: 50
D: 2E
D: 30
D: 32
D: 34
100µs
10.0MS/s
10k points
2.00 V
2.00 V
D15–D0
Timing Resolution: 100ns
B1 Address/Data
Type
Bus
Source Bus
B1 (I2C)
Trigger On
Address/Data
Address
50h
Data
2Eh
Direction
Read
Mode
Normal
& Holdoff

Figure 54. Serial bus analysis is accelerated with automated trigger, decode, and search on serial packet context.

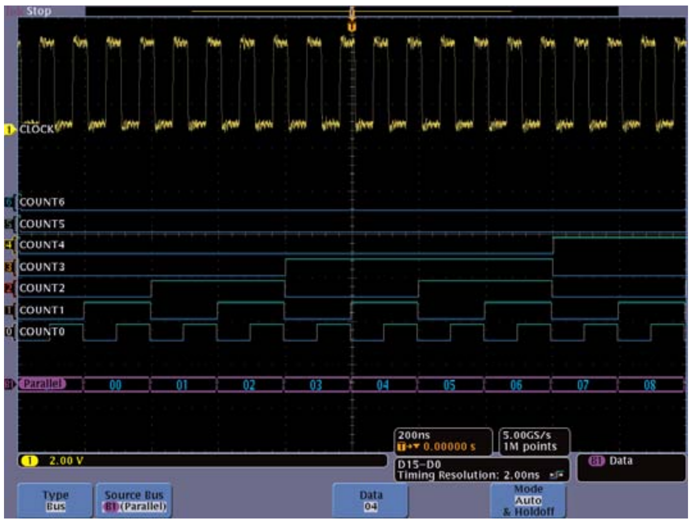
Stop
CLOCK
COUNT6
COUNT5
COUNT4
COUNT3
COUNT2
COUNT1
COUNT0
Parallel
00
01
02
03
04
05
06
07
08
200ns
5.00GS/s
1M points
2.00 V
D15–D0
Timing Resolution: 2.00ns
B1 Data
Type
Bus
Source Bus
B1 (Parallel)
Data
04
Mode
Auto
& Holdoff

Figure 55. Automatically trigger, decode, and search on clocked or unclocked parallel bus data.

- Add memory to channels to analyze longer record lengths
- Add application-specific measurement capabilities
- Complement the power of the oscilloscope with a full range of probes and modules
- Work with popular third-party analysis and productivity Windows-compatible software
- Add accessories, such as battery packs and rackmounts

Application modules and software may enable you to transform your oscilloscope into a highly specialized analysis tool capable of performing functions such as jitter and timing analysis, microprocessor memory system verification, communications standards testing, disk drive measurements, video measurements, power measurements and much more. Figures 53 - 58 highlight a few of these examples.

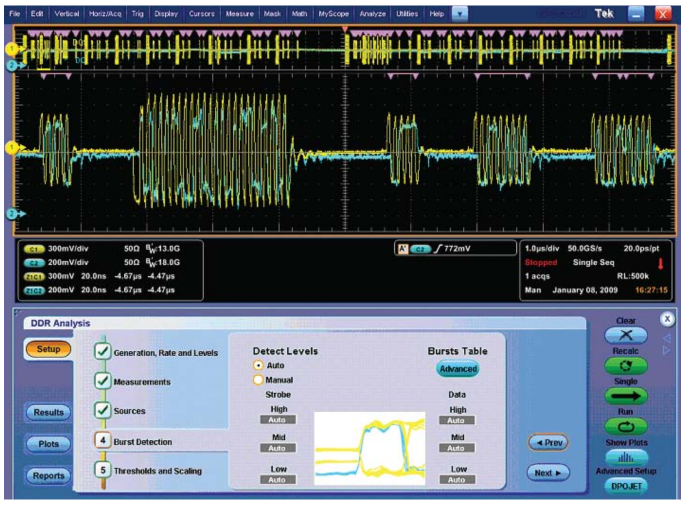
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
DDR Analysis
Setup
Results
Plots
Reports
Generation, Rate and Levels
Measurements
Sources
Burst Detection
Thresholds and Scaling
Detect Levels
Auto
Manual
Strobe
High
Mid
Low
Bursts Table
Advanced
Data
Clear
Recalc
Single
Run
Show Plots
Advanced Setup
DPOJET
Prev
Next

Figure 56. Advanced DDR analysis tools automate complex memory tasks like separating read/write bursts and performing JEDEC measurements.

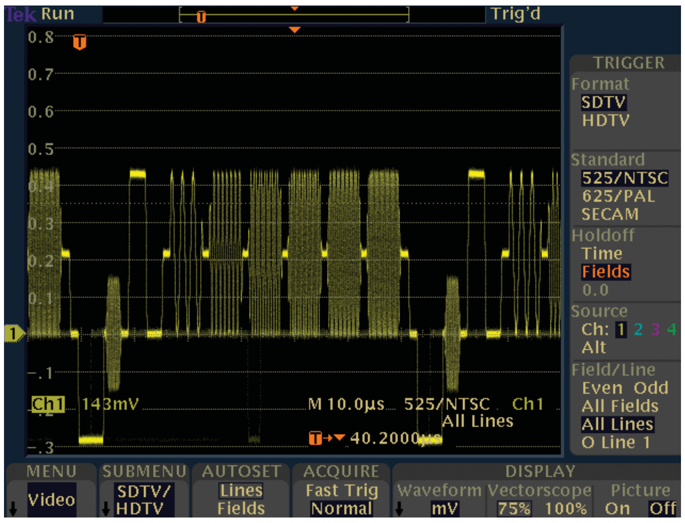
Run
Trig'd
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
−.1
−.3
Ch1 143mV
M 10.0µs
525/NTSC
Ch1
All Lines
40.2000µs
TRIGGER
Format
SDTV
HDTV
Standard
525/NTSC
625/PAL
SECAM
Holdoff
Time
Fields
0.0
Source
Ch: 1 2 3 4
Alt
Field/Line
Even Odd
All Fields
All Lines
0 Line 1
MENU
Video
SUBMENU
SDTV/
HDTV
AUTOSET
Lines
Fields
ACQUIRE
Fast Trig
Normal
DISPLAY
Waveform
mV
Vectorscope
75%
100%
Picture
On
Off

Figure 57. Video application modules make the oscilloscope a fast, tell-all tool for video troubleshooting.

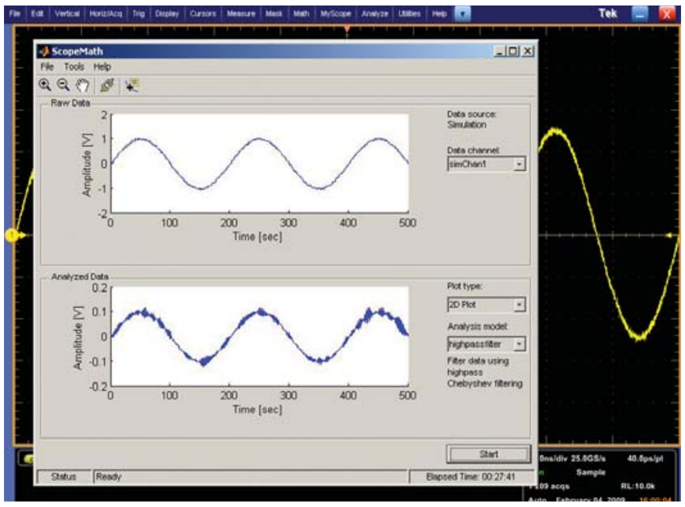
ScopeMath
File
Tools
Help
Raw Data
Amplitude [V]
Time [sec]
Data source:
Simulation
Data channel:
simChan1
Analyzed Data
Plot type:
2D Plot
Analysis model:
highpassfilter
Filter data using highpass Chebyshev filtering
Start
Status
Ready
Elapsed Time: 00:27:41

Figure 58. Advanced analysis and productivity software, such as MATLAB®, can be installed in Windows-based oscilloscopes to accomplish local signal analysis.

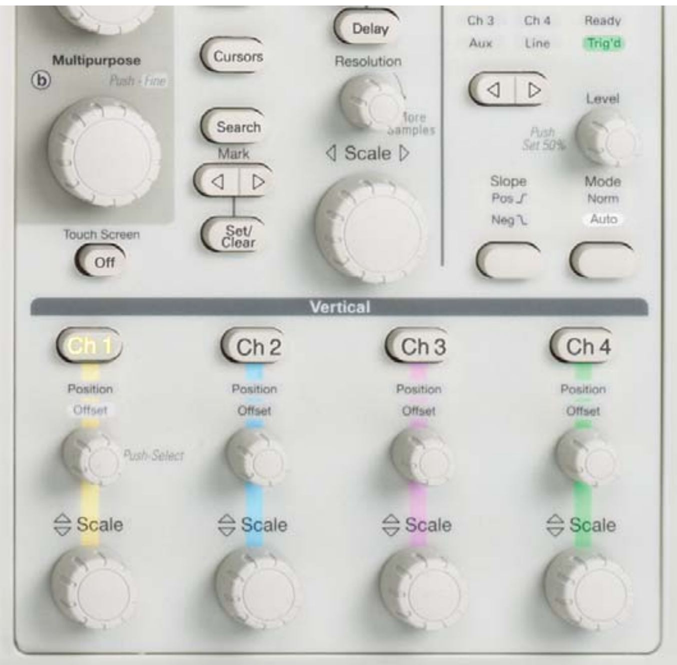
Multipurpose
Push - Fine
Touch Screen
Off
Cursors
Search
Mark
Set/Clear
Delay
Resolution
More Samples
Scale
Ch 3
Ch 4
Aux
Line
Ready
Trig'd
Level
Push Set 50%
Slope
Pos
Neg
Mode
Norm
Auto
Vertical
Ch 1
Ch 2
Ch 3
Ch 4
Position
Offset
Push-Select
Scale

Figure 59. Traditional, analog-style knobs control position, scale, intensity, etc. – precisely as you would expect.

Ease-of-Use

Oscilloscopes should be easy to learn and easy to use, helping you work at peak efficiency and productivity. Allowing you to focus on your design, rather than the measurement tools. Just as there is no one typical car driver, there is no one typical oscilloscope user. Regardless of whether you prefer a traditional instrument interface, or that of a Windows® interface, it is important to have flexibility in your oscilloscope's operation.

Many oscilloscopes offer a balance between performance and simplicity by providing the user with many ways to operate the instrument. A front-panel layout, in Figure 59, provides dedicated vertical, horizontal and trigger controls. An icon-rich graphical user interface, as shown in Figure 60, helps you understand and intuitively use advanced capabilities. Touch-sensitive displays solve issues with cluttered benches and carts, while providing access to clear, on-screen buttons, as seen in Figure 61. Online help provides a convenient, built-in reference manual. Intuitive controls allow even occasional oscilloscope users to feel as comfortable driving the oscilloscope as they do driving a car, while giving full-time users easy access to the oscilloscope's most advanced features. In addition, many oscilloscopes are portable, such as the one shown in Figure 62, making the oscilloscope efficient in many different operating environments – in the lab or in the field.

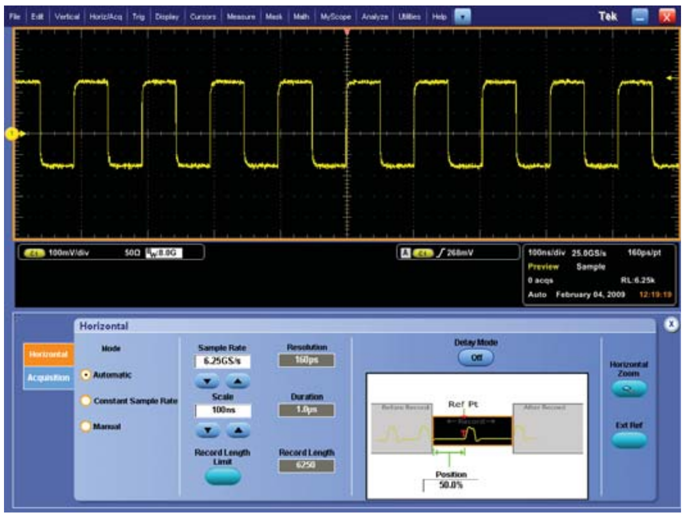
File
Edit
Vertical
Horiz/Acq
Trig
Display
Cursors
Measure
Mask
Math
MyScope
Analyze
Utilities
Help
Tek
100mV/div
100ns/div
Preview
Sample
Horizontal
Acquisition
Mode
Automatic
Constant Sample Rate
Manual
Sample Rate
6.25GS/s
Resolution
160ps
Scale
100ns
Duration
1.0µs
Record Length Limit
Record Length
6250
Delay Mode
Off
Ref Pt
Position
50.0%
Horizontal Zoom
Ext Ref

Figure 60. Use graphical control windows to access even the most sophisiticated functions with confidence and ease.

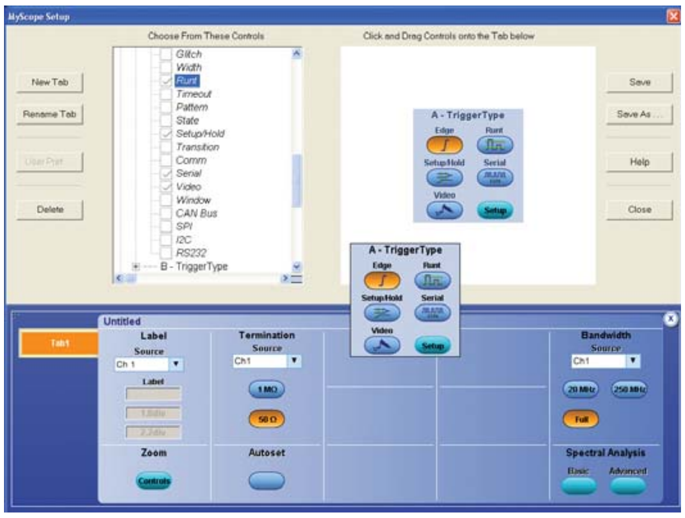
MyScope Setup
Choose From These Controls
Click and Drag Controls onto the Tab below
New Tab
Rename Tab
Delete
Save
Save As
Help
Close
Glitch
Width
Runt
Timeout
Pattern
State
Setup/Hold
Transition
Comm
Serial
Video
Window
CAN Bus
SPI
I2C
RS232
B - TriggerType
A - TriggerType
Edge
Runt
Setup/Hold
Serial
Video
Setup
Untitled
Tab1
Label
Source
Ch 1
Termination
1 MΩ
50 Ω
Zoom
Controls
Autoset
Bandwidth
20 MHz
250 MHz
Full
Spectral Analysis
Basic
Advanced

Figure 61. Touch-sensitive display naturally solves issues with cluttered benches and carts, while providing access to clear, on-screen buttons.

Figure 62. The portability of many oscilloscopes makes the instrument efficient in many operating environments.

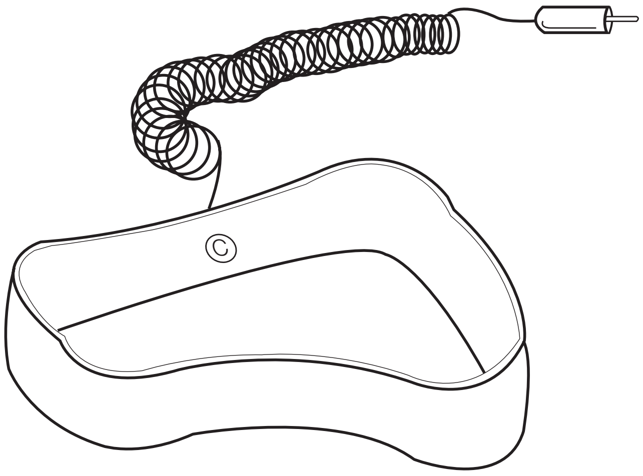

Figure 63. Typical wrist-type grounding strap.

Operating the Oscilloscope

This section briefly describes how to set up and start using an oscilloscope – specifically, how to properly ground the oscilloscope and yourself, set the oscilloscope controls, calibrate the oscilloscope, connect the probes, and compensate the probes.

Proper grounding is an important step when setting up to take measurements or work on a circuit. Proper grounding of the oscilloscope protects you from a hazardous shock and grounding yourself protects your circuits from damage.

Proper Grounding

To ground the oscilloscope means to connect it to an electrically neutral reference point, such as earth ground. Ground your oscilloscope by plugging its three-pronged power cord into an outlet grounded to earth ground.

Grounding the oscilloscope is necessary for safety. If a high voltage contacts the case of an ungrounded oscilloscope – any part of the case, including knobs that appear insulated – it can give you a shock. However, with a properly grounded oscilloscope, the current travels through the grounding path to earth ground rather than through you to earth ground.

Grounding is also necessary for taking accurate measurements with your oscilloscope. The oscilloscope needs to share the same ground as any circuits you are testing.

Some oscilloscopes do not require separate connection to earth ground. These oscilloscopes have insulated cases and controls, which keeps any possible shock hazard away from the user.

If you are working with integrated circuits (ICs), you also need to ground yourself. Integrated circuits have tiny conduction paths that can be damaged by static electricity that builds up on your body. You can ruin an expensive IC simply by walking across a carpet or taking off a sweater and then touching the leads of the IC. To solve this problem, wear a grounding strap, as shows in Figure 63. This strap safely sends static charges on your body to earth ground.

Setting the Controls

After plugging in the oscilloscope, take a look at the front panel. As described previously, the front panel is typically divided into three main sections labeled vertical, horizontal, and trigger. Your oscilloscope may have other sections, depending on the model and type.

Notice the input connectors on your oscilloscope – this is where you attach the probes. Most oscilloscopes have at least two input channels and each channel can display a waveform on the screen. Multiple channels are useful for comparing waveforms. As mentioned earlier, MSOs also have digital inputs as well.

Some oscilloscopes have AUTOSET and/or DEFAULT buttons that can set up the controls in one step to accommodate a signal. If your oscilloscope does not have this capability, it is helpful to set the controls to standard positions before taking measurements.

General instructions to manually set up the oscilloscope in standard positions are as follows:

- Set the oscilloscope to display channel 1
- Set the vertical volts/division scale and position controls to mid–range positions
- Turn off the variable volts/division
- Turn off all magnification settings
- Set the channel 1 input coupling to DC
- Set the trigger mode to auto
- Set the trigger source to channel 1
- Turn trigger holdoff to minimum or off
- Set the horizontal time/division and position controls to mid-range positions
- Adjust channel 1 volts/division such that the signal occupies as much of the 10 vertical divisions as possible without clipping or signal distortion

Calibrating the Instrument

In addition to proper oscilloscope setup, periodic instrument self-calibration is recommended for accurate measurements. Calibration is needed if the ambient temperature has changed more than 5° C (9° F) since the last self-calibration or once a week. In the oscilloscope menu this can sometimes be initiated as “Signal Path Compensation”. Refer to the manual that accompanied your oscilloscope for more detailed instructions.

Connecting the Probes

Now you are ready to connect a probe to your oscilloscope. A probe, if well-matched to the oscilloscope, will enable you to access all of the power and performance in the oscilloscope and will ensure the integrity of the signal you are measuring.

Measuring a signal requires two connections: the probe tip connection and the ground connection. Probes often come with a clip attachment for grounding the probe to the circuit under test. In practice, you attach the grounding clip to a known ground in the circuit, such as the metal chassis of a product you are repairing, and touch the probe tip to a test point in the circuit

Compensating the Probes

Passive attenuation voltage probes must be compensated to the oscilloscope. Before using a passive probe, you need to compensate it – to balance its electrical properties to a particular oscilloscope.

You should get into the habit of compensating the probe every time you set up your oscilloscope. A poorly adjusted probe can make your measurements less accurate. Figure 64 illustrates the effects on a 1 MHz test signal when using a probe that is not properly compensated.

Most oscilloscopes have a square wave reference signal available at a terminal on the front panel used to compensate the probe. General instructions to compensate the probe are as follows:

- Attach the probe to a vertical channel
- Connect the probe tip to the probe compensation, i.e. square wave reference signal
- Attach the ground clip of the probe to ground
- View the square wave reference signal
- Make the proper adjustments on the probe so that the corners of the square wave are square

Probe Undercompensated.

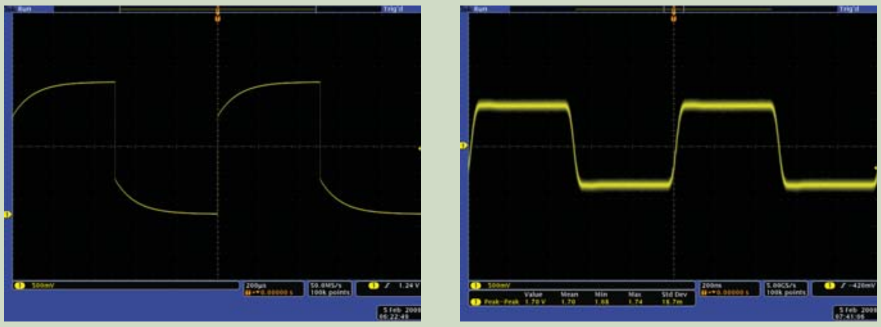

Probe Adjustment Signal

Note Reduced Amplitude

Probe Compensated Correctly.

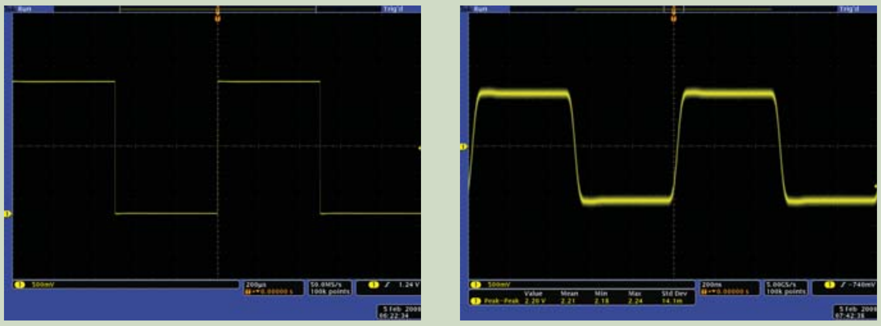

Probe Adjustment Signal

Note Proper Amplitude

Probe Overcompensated.

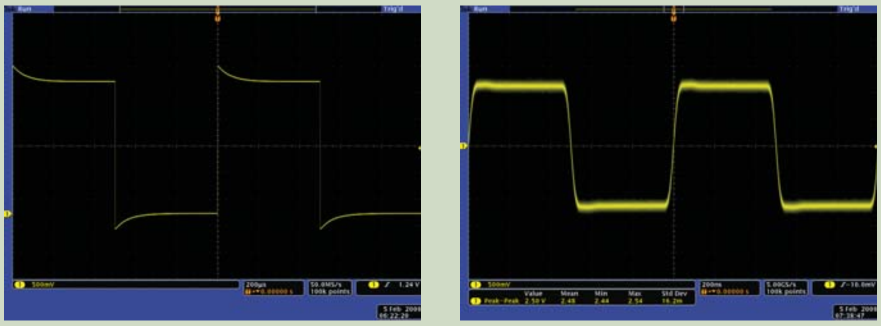

Probe Adjustment Signal

Note Increased Amplitude

Figure 64. The effects of improper probe compensation.

When you compensate the probe, always attach any accessory tips you will use and connect the probe to the vertical channel you plan to use. This will ensure that the oscilloscope has the same electrical properties as it does when you take measurements.

Oscilloscope Measurement Techniques

This section reviews basic measurement techniques. The two most basic measurements you can make are voltage and time measurements. Just about every other measurement is based on one of these two fundamental techniques.

This section discusses methods for taking measurements visually with the oscilloscope screen. This is a common technique with analog instruments, and also may be useful for "at-a-glance" interpretation of digital oscilloscope displays.

Note that most digital oscilloscopes include automated measurement tools that simplify and accelerate common analysis tasks, thus improving the reliability and confidence of your measurements. However, knowing how to make measurements manually as described here will help you understand and check the automatic measurements.

Voltage Measurements

Voltage is the amount of electric potential, expressed in volts, between two points in a circuit. Usually one of these points is ground (zero volts), but not always. Voltages can also be measured from peak-to-peak – from the maximum point of a signal to its minimum point. You must be careful to specify which voltage you mean.

The oscilloscope is primarily a voltage-measuring device. Once you have measured the voltage, other quantities are just a calculation away. For example, Ohm's law states that voltage between two points in a circuit equals the current times the resistance. From any two of these quantities you can calculate the third using the following formula:

$$\text{Voltage} = \text{Current} \; x \; \text{Resistance}$$

$$\text{Current} = \frac{\text{Voltage}}{\text{Resistance}}$$

$$\text{Resistance} = \frac{\text{Voltage}}{\text{Current}}$$

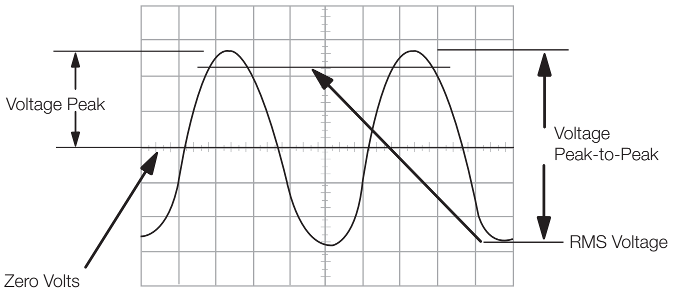

Figure 65. Voltage peak (V_p) and peak-to-peak voltage ($V_{p\text{-}p}$)

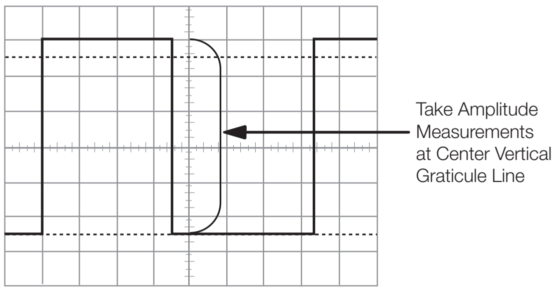
Take Amplitude Measurements at Center Vertical Graticule Line

Figure 66. Measure voltage on the center vertical graticule line.

Another handy formula is the power law, which states that the power of a DC signal equals the voltage times the current. Calculations are more complicated for AC signals, but the point here is that measuring the voltage is the first step toward calculating other quantities. Figure 65 shows the voltage of one peak (V_p) and the peak-to-peak voltage ($V_{p\text{-}p}$).

The most basic method of taking voltage measurements is to count the number of divisions a waveform spans on the oscilloscope's vertical scale. Adjusting the signal to cover most of the display vertically makes for the best voltage measurements, as shown in Figure 66. The more display area you use, the more accurately you can read the measurement.

Many oscilloscopes have cursors that let you make waveform measurements automatically, without having to count graticule marks. A cursor is simply a line that you can move across the display. Two horizontal cursor lines can be moved up and down to bracket a waveform's amplitude for voltage measurements, and two vertical lines move right and left for time measurements. A readout shows the voltage or time at their positions.

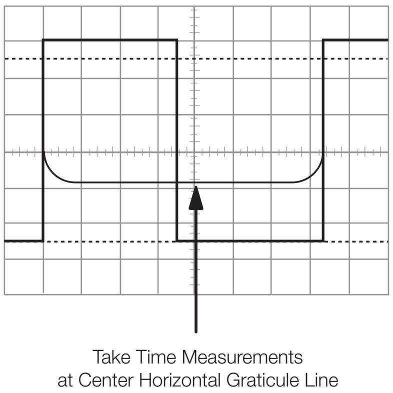

Figure 67. Measure time on the center horizontal graticule line.

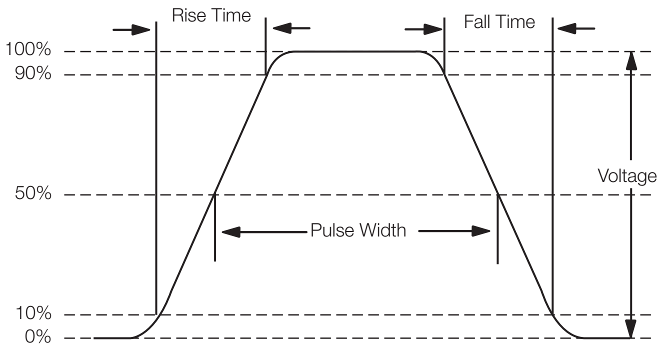

Figure 68. Rise time and pulse width measurement points.

Time and Frequency Measurements

You can make time measurements using the horizontal scale of the oscilloscope. Time measurements include measuring the period and pulse width of pulses. Frequency is the reciprocal of the period, so once you know the period, the frequency is one divided by the period. Like voltage measurements, time measurements are more accurate when you adjust the portion of the signal to be measured to cover a large area of the display, as illustrated in Figure 67.

Pulse Width and Rise Time Measurements

In many applications, the details of a pulse's shape are important. Pulses can become distorted and cause a digital circuit to malfunction, and the timing of pulses in a pulse train is often significant.

Standard pulse measurements are pulse rise time and pulse width. Rise time is the amount of time a pulse takes to go from a low to high voltage. By convention, the rise time is measured from 10% to 90% of the full voltage of the pulse. This eliminates any irregularities at the pulse's transition corners. Pulse width is the amount of time the pulse takes to go from low to high and back to low again. By convention, the pulse width is measured at 50% of full voltage. Figure 68 illustrates these measurement points.

Pulse measurements often require fine-tuning the triggering. To become an expert at capturing pulses, you should learn how to use trigger holdoff and how to set the digital oscilloscope to capture pretrigger data, as described in the Systems and Controls of an Oscilloscope section. Horizontal magnification is another useful feature for measuring pulses, since it allows you to see fine details of a fast pulse.

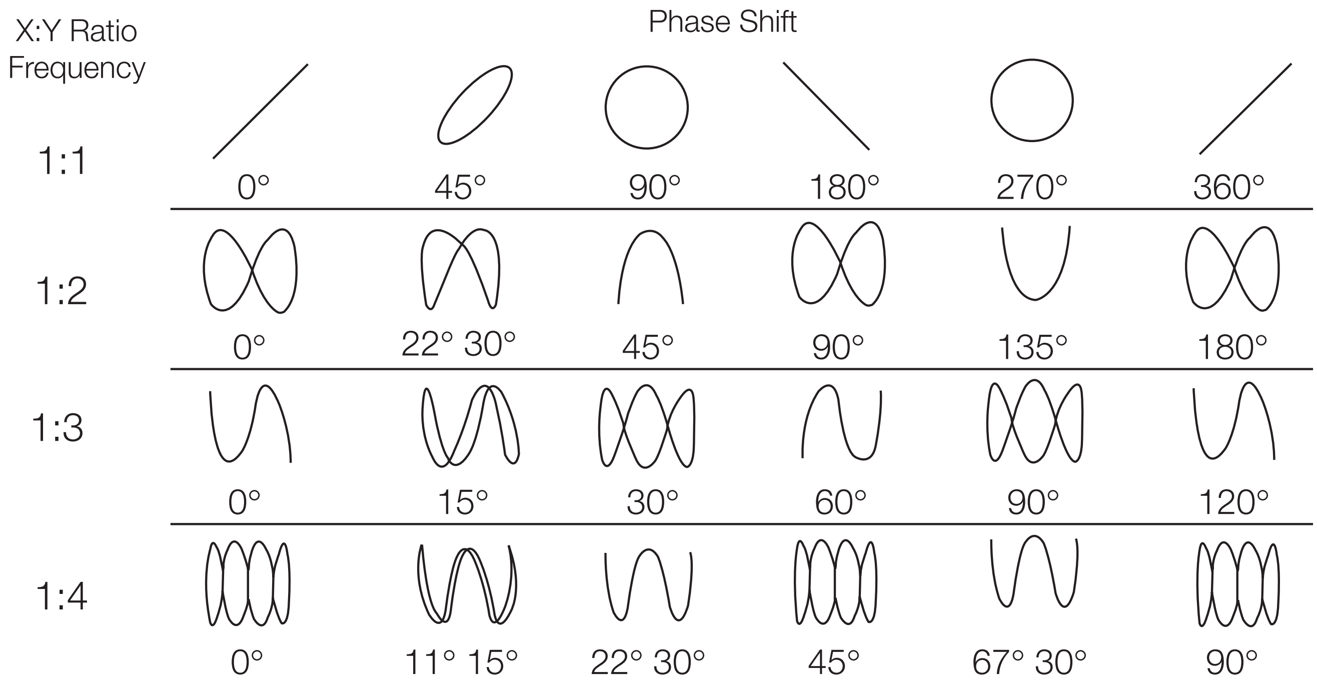

Figure 69. Lissajous patterns.

Phase Shift Measurements

One method for measuring phase shift – the difference in timing between two otherwise identical periodic signals – is to use XY mode. This measurement technique involves inputting one signal into the vertical system as usual and then another signal into the horizontal system – called an XY measurement because both the X and Y axis are tracing voltages. The waveform that results from this arrangement is called a Lissajous pattern (named for French physicist Jules Antoine Lissajous and pronounced LEE–sa–zhoo). From the shape of the Lissajous pattern, you can tell the phase difference between the two signals. You can also tell their frequency ratio. Figure 69 shows Lissajous patterns for various frequency ratios and phase shifts.

The XY measurement technique originated with analog oscilloscopes. DSOs may have difficulty creating real-time XY displays. Some DSOs create an XY image by accumulating triggered data points over time, then displaying two channels as an XY display. DPOs, on the other hand, are able to acquire and display a genuine XY mode image in real-time, using a continuous stream of digitized data. DPOs can also display an XYZ image with intensified areas. Unlike XY displays on DSOs and DPOs, these displays on analog oscilloscopes are typically limited to a few megahertz of bandwidth.

Other Measurement Techniques

This section has covered basic measurement techniques. Other measurement techniques involve setting up the oscilloscope to test electrical components on an assembly line, capturing elusive transient signals, and many others. The measurement techniques you will use will depend on your application, but you have learned enough to get started. Practice using your oscilloscope and read more about it. Soon its operation will be second nature to you.

Written Exercises

This section contains written exercises that cover information in this primer. These are divided into two parts, Part I and Part II, with vocabulary and application exercises for each.

Check how well you have absorbed the information in these sections by verifying your responses against the answer key at the end of this section on page 54.

Part I covers information presented in these sections:

- The Oscilloscope
- Performance Terms and Considerations

Part II covers information presented in sections:

- The Systems and Controls of an Oscilloscope
- Operating the Oscilloscope
- Measurement Techniques

Part I A: Vocabulary Exercise

Write the letter of the definitions in the right column next to the correct words in the left column.

	Term		**Definition**
1. ___	Acquisition	**A**	The unit of electric potential difference.
2. ___	Analog	**B**	A performance measurement indicating the precision of an ADC, measured in bits.
3. ___	Bandwidth	**C**	Term used when referring to degree points of a signal's period.
4. ___	Digital Phosphor	**D**	The number of times a signal repeats in one second.
5. ___	Frequency	**E**	The amount of time it takes a wave to complete one cycle.
6. ___	Glitch	**F**	A stored digital value that represents the voltage of a signal at a specific point in time on the display.
7. ___	Period	**G**	A common waveform shape that has a rising edge, a width, and a falling edge.
8. ___	Phase	**H**	A performance measurement indicating the rising edge speed of a pulse.
9. ___	Pulse	**I**	Oscilloscope circuitry that controls the timing of the sweep.
10. ___	Waveform Point	**J**	An intermittent spike in a circuit.
11. ___	Rise Time	**K**	A signal measured by an oscilloscope that only occurs once.
12. ___	Sample Point	**L**	The oscilloscope's process of collecting sample points from the ADC, processing them, and storing them in memory.
13. ___	Digital Storage	**M**	Something that operates with continuously changing values.
14. ___	Time Base	**N**	Digital oscilloscope that captures 3 dimensions of signal information in real-time.
15. ___	Transient	**O**	Digital oscilloscope with serial processing.
16. ___	ADC Resolution	**P**	A sine wave frequency range, defined by the – 3 dB point.
17. ___	Volt	**Q**	The raw data from an ADC used to calculate and display waveform points.

Part I B: Application Exercise

Circle the best answers for each statement. Some statements have more than one right answer.

1. With an oscilloscope you can:

a. Calculate the frequency of a signal.

b. Find malfunctioning electrical components.

c. Analyze signal details.

d. All the above.

2. The difference between analog and digitizing oscilloscopes is:

a. Analog oscilloscopes do not have on-screen menus.

b. Analog oscilloscopes apply a measurement voltage directly to the display system, while digital oscilloscopes first convert the voltage into digital values.

c. Analog oscilloscopes measure analogs, whereas digitizing oscilloscopes measure digits.

d. Analog oscilloscopes do not have an acquisition system.

3. An oscilloscope's vertical section does the following:

a. Acquires sample points with an ADC.

b. Starts a horizontal sweep.

c. Lets you adjust the brightness of the display.

d. Attenuates or amplifies the input signal.

4. The time base control of the oscilloscope does the following:

a. Adjusts the vertical scale.

b. Shows you the current time of day.

c. Sets the amount of time represented by the horizontal width of the screen.

d. Sends a clock pulse to the probe.

5. On an oscilloscope display:

a. Voltage is on the vertical axis and time is on the horizontal axis.

b. A straight diagonal trace means voltage is changing at a steady rate.

c. A flat horizontal trace means voltage is constant.

d. All the above.

6. All repeating waves have the following properties:

a. A frequency measured in Hertz.

b. A period measured in seconds.

c. A bandwidth measured in Hertz.

d. All the above.

7. If you probe inside a computer with an oscilloscope, you are likely to find the following types of signals:

a. Pulse trains.

b. Ramp waves.

c. Sine waves.

d. All the above.

8. When evaluating the performance of an analog oscilloscope, some things you might consider are:

a. The bandwidth.

b. The vertical sensitivity.

c. The ADC resolution.

d. The sweep speed.

9. The difference between digital storage oscilloscopes (DSO) and digital phosphor oscilloscopes (DPO) is:

a. The DSO has a higher bandwidth.

b. The DPO captures three dimensions of waveform information in real-time.

c. The DSO has a color display.

d. The DSO captures more signal details.

Part II A: Vocabulary Exercise

Write the letter of the definitions in the right column next to the correct words in the left column.

	Term		**Definition**
1. ___	Averaging Mode	**A**	The unintentional interaction of the probe and oscilloscope with the circuit being tested which distorts a signal.
2. ___	Circuit Loading	**B**	A conductor that connects electrical currents to the Earth.
3. ___	Compensation	**C**	A sampling mode in which the digital oscilloscope collects as many samples as it can as the signal occurs, then constructs a display, using interpolation if necessary.
4. ___	Coupling	**D**	A sampling mode in which the digital oscilloscope constructs a picture of a repetitive signal by capturing a little bit of information from each repetition.
5. ___	Earth Ground	**E**	A device that converts a specific physical quantity such as sound, pressure, strain, or light intensity into an electrical signal.
6. ___	Equivalent-Time	**F**	A test device for injecting a signal into a circuit input.
7. ___	Graticule	**G**	A processing technique used by digital oscilloscopes to eliminate noise in a displayed signal.
8. ___	Interpolation	**H**	The method of connecting two circuits together.
9. ___	Real Time	**I**	A “connect-the-dots” processing technique to estimate what a fast waveform looks like based on only a few sampled points.
10. ___	Signal Generator	**J**	The grid lines on a screen for measuring oscilloscope traces.
11. ___	Single Sweep	**K**	A trigger mode that triggers the sweep once, must be reset to accept another trigger event.
12. ___	Sensor	**L**	A probe adjustment for 10X attenuator probes that balances the electrical properties of the probe with the electrical properties of the oscilloscope.

Part II B: Application Exercise

Circle the best answers for each statement. Some statements have more than one right answer.

1. To operate an oscilloscope safely, you should:

a. Ground the oscilloscope with the proper three-pronged power cord.

b. Learn to recognize potentially dangerous electrical components.

c. Avoid touching exposed connections in a circuit being tested even if the power is off.

d. All the above.

2. Grounding an oscilloscope is necessary:

a. For safety reasons.

b. To provide a reference point for making measurements.

c. To align the trace with the screen's horizontal axis.

d. All the above.

3. Circuit loading is caused by:

a. An input signal having too large a voltage.

b. The probe and oscilloscope interacting with the circuit being tested.

c. A 10X attenuator probe being uncompensated.

d. Putting too much weight on a circuit.

4. Compensating a probe is necessary to:

a. Balance the electrical properties of the 10X attenuator probe with the oscilloscope.

b. Prevent damaging the circuit being tested.

c. Improve the accuracy of your measurements.

d. All the above.

5. The trace rotation control is useful for:

a. Scaling waveforms on the screen.

b. Detecting sine wave signals.

c. Aligning the waveform trace with the screen's horizontal axis on an analog oscilloscope.

d. Measuring pulse width.

6. The volts per division control is used to:

a. Scale a waveform vertically.

b. Position a waveform vertically.

c. Attenuate or amplify an input signal.

d. Set the numbers of volts each division represents.

7. Setting the vertical input coupling to ground does the following:

a. Disconnects the input signal from the oscilloscope.

b. Causes a horizontal line to appear with auto trigger.

c. Lets you see where zero volts is on the screen.

d. All the above.

8. The trigger is necessary to:

a. Stabilize repeating waveforms on the screen.

b. Capture single-shot waveforms.

c. Mark a particular point of an acquisition.

d. All the above.

9. The difference between auto and normal trigger mode is:

a. In normal mode the oscilloscope only sweeps once and then stops.

b. In normal mode the oscilloscope only sweeps if the input signal reaches the trigger point; otherwise the screen is blank.

c. Auto mode makes the oscilloscope sweep continuously even without being triggered.

d. All the above.

10. The acquisition mode that best reduces noise in a repeating signal is:

a. Sample mode.

b. Peak detect mode.

c. Envelope mode.

d. Averaging mode.

11. The two most basic measurements you can make with an oscilloscope are:

a. Time and frequency measurements.

b. Time and voltage measurements.

c. Voltage and pulse width measurements.

d. Pulse width and phase shift measurements.

12. If the volts/division is set at 0.5, the largest signal that can fit on the screen (assuming an 8 x 10 division screen) is:

a. 62.5 millivolts peak-to-peak.

b. 8 volts peak-to-peak.

c. 4 volts peak-to-peak.

d. 0.5 volts peak-to-peak.

13. If the seconds/division is set at 0.1 ms, the amount of time represented by the width of the screen is:

a. 0.1 ms.

b. 1 ms.

c. 1 second.

d. 0.1 kHz.

14. By convention, pulse width is measured:

a. At 10% of the pulse's peak-to-peak (pk-pk) voltage.

b. At 50% of the pulse's peak-to-peak (pk-pk) voltage.

c. At 90% of the pulse's peak-to-peak (pk-pk) voltage.

d. At 10% and 90% of the pulse's peak-to-peak (pk-pk) voltage.

15. You attach a probe to your test circuit but the screen is blank. You should:

a. Check that the screen intensity is turned up.

b. Check that the oscilloscope is set to display the channel that the probe is connected to.

c. Set the trigger mode to auto since norm mode blanks the screen.

d. Set the vertical input coupling to AC and set the volts/division to its largest value since a large DC signal may go off the top or bottom of the screen.

e. Check that the probe isn't shorted and make sure it is properly grounded.

f. Check that the oscilloscope is set to trigger on the input channel you are using.

g. All of the above.

Answer Key

This section provides the answers to all written exercises in the previous section.

Part IA: Vocabulary Exercise Answers

1. L	5. D	9. G	13. O
2. M	6. J	10. F	14. I
3. P	7. E	11. H	15. K
4. N	8. C	12. Q	16. B
			17. A

Part IB: Application Exercise Answers

1. D	3. D	5. D	7. A
2. B,D	4. C	6. A,B	8. A,B,D
			9. B

Part IIA: Vocabulary Exercise Answers

1. G	4. H	7. J	10. F
2. A	5. B	8. I	11. K
3. L	6. D	9. C	12. E

Part IIB: Application Exercise Answers

1. D	5. C	9. B,C	13. B
2. A,B	6. A,C,D	10. D	14. B
3. B	7. D	11. B	15. G
4. A,C	8. D	12. C	

Glossary

A

Acquisition Mode – Modes that control how waveform points are produced from sample points. Some types include sample, peak detect, hi res, envelope, average, and waveform data base.

Alternating Current (AC) – A signal in which the current and voltage vary in a repeating pattern over time. Also used to indicate signal coupling type.

Amplification – An increase in signal amplitude during its transmission from one point to another.

Amplitude – The magnitude of a quantity or strength of a signal. In electronics, amplitude usually refers to either voltage or power.

Analog-to-Digital Converter (ADC) – A digital electronic component that converts an electrical signal into discrete binary values.

Analog Oscilloscope – An instrument that creates a waveform display by applying the input signal (conditioned and amplified) to the vertical axis of an electron beam moving across a cathode-ray tube (CRT) screen horizontally from left to right. A chemical phosphor coated on the CRT creates a glowing trace wherever the beam hits.

Analog Signal – A signal with continuously variable voltages.

Attenuation – A decrease in signal amplitude during its transmission from one point to another.

Averaging – A processing technique used by digital oscilloscopes to reduce noise in a displayed signal.

B

Bandwidth – A frequency range, usually limited by –3 dB.

C

Circuit Loading – The unintentional interaction of the probe and oscilloscope with the circuit being tested, distorting the signal.

Compensation – A probe adjustment for passive attenuation probes that balances the capacitance of the probe with the capacitance of the oscilloscope.

Coupling – The method of connecting two circuits together. Circuits connected with a wire are directly coupled (DC); circuits connected through a capacitor or transformer are indirectly (AC) coupled.

Cursor – An on–screen marker that you can align with a waveform to make more accurate measurements.

D

Delayed Time Base – A time base with a sweep that can start (or be triggered to start) relative to a pre-determined time on the main time base sweep. Allows you to see events more clearly and to see events that are not visible solely with the main time base sweep.

Digital Signal – A signal whose voltage samples are represented by discrete binary numbers.

Digital Oscilloscope – A type of oscilloscope that uses an analog-to-digital converter (ADC) to convert the measured voltage into digital information. Types include: digital storage, digital phosphor, mixed signal, and digital sampling oscilloscopes.

Digital Phosphor Oscilloscope (DPO) – A type of digital oscilloscope that closely models the display characteristics of an analog oscilloscope while providing traditional digital oscilloscope benefits (waveform storage, automated measurements, etc.) The DPO uses a parallel-processing architecture to pass the signal to the raster-type display, which provides intensity-graded viewing of signal characteristics in real time. The DPO displays signals in three dimensions: amplitude, time and the distribution of amplitude over time.

Digital Sampling Oscilloscope – A type of digital oscilloscope that employs equivalent-time sampling method to capture and display samples of a signal, ideal for accurately capturing signals whose frequency components are much higher than the oscilloscope's sample rate.

Digital Signal Processing – The application of algorithms to improve the accuracy of measured signals.

Digital Storage Oscilloscope (DSO) – A digital oscilloscope that acquires signals via digital sampling (using an analog-to-digital converter). It uses a serial-processing architecture to control acquisition, user interface, and the raster display.

Digitize – The process by which an analog-to-digital converter (ADC) in the horizontal system samples a signal at discrete points in time and converts the signal's voltage at these points into digital values called sample points.

Direct Current (DC) – A signal with a constant voltage and/or current. Also used to indicate signal coupling type.

Division – Measurement markings on the oscilloscope graticule indicating major and minor marks.

E

Earth Ground – A conductor that will connect electrical currents to the Earth.

Effective Bits – A measure of a digital oscilloscope's ability to accurately reconstruct a sine wave signal's shape. This measurement compares the oscilloscope's actual error to that of a theoretical "ideal" digitizer.

Envelope – The outline of a signal's highest and lowest points acquired over many displayed waveform repetitions.

Equivalent-time Sampling – A sampling mode in which the oscilloscope constructs a picture of a repetitive signal by capturing a little bit of information from each repetition. Two types of equivalent-time sampling: random and sequential.

F

Focus – The analog oscilloscope control that adjusts the cathode-ray tube (CRT) electron beam to control the sharpness of the display.

Frequency – The number of times a signal repeats in one second, measured in Hertz (cycles per second). The frequency equals 1/period.

Frequency Response – Frequency response curves of an oscilloscope define the accuracy in amplitude representation of the input signal in function of the signals frequency. In order to obtain maximum signal fidelity, it is important that the oscilloscope has a flat (stable) frequency response across the entire specified oscilloscopes bandwidth.

G

Gain Accuracy – An indication of how accurately the vertical system attenuates or amplifies a signal, usually represented as a percentage error.

Gigahertz (GHz) – 1,000,000,000 Hertz; a unit of frequency.

Glitch – An intermittent, high-speed error in a circuit.

Graticule – The grid lines on a display for measuring oscilloscope traces.

Ground –

1. A conducting connection by which an electric circuit or equipment is connected to the earth to establish and maintain a reference voltage level.
2. The voltage reference point in a circuit.

H

Hertz (Hz) – One cycle per second; the unit of frequency.

Horizontal Accuracy (Time Base) – An indication of how accurately the horizontal system displays the timing of a signal, usually represented as a percentage error.

Horizontal Sweep – The action of the horizontal system that causes a waveform to be drawn.

I

Intensity Grading – Frequency-of-occurrence information that is essential to understanding what the waveform is really doing.

Interpolation – A "connect-the-dots" processing technique to estimate what a fast waveform looks like based on only a few sampled points. Two types: linear and sin x/x.

K

Kilohertz (kHz) – 1,000 Hertz; a unit of frequency.

L

Loading – The unintentional interaction of the probe and oscilloscope with the circuit being tested which distorts a signal.

Logic Analyzer – An instrument used to make the logic states of many digital signals visible over time. It analyzes the digital data and can represent the data as real-time software execution, data flow values, state sequences, etc.

M

Megahertz (MHz) – 1,000,000 Hertz; a unit of frequency.

Megasamples per second (MS/s) – A sample rate unit equal to one million samples per second.

Microsecond (µs) – A unit of time equivalent to 0.000001 seconds.

Millisecond (ms) – A unit of time equivalent to 0.001 seconds.

Mixed Signal Oscilloscope (MSO) – A type of digital oscilloscope that combines the basic functionality of a 16-channel logic analyzer with the trusted performance of a 4-channel digital phosphor oscilloscope.

N

Nanosecond (ns) – A unit of time equivalent to 0.000000001 seconds.

Noise – An unwanted voltage or current in an electrical circuit.

O

Oscilloscope – An instrument used to make voltage changes visible over time. The word oscilloscope comes from "oscillate," since oscilloscopes are often used to measure oscillating voltages.

P

Peak (V_p) – The maximum voltage level measured from a zero reference point.

Peak Detection – An acquisition mode available with digital oscilloscopes that enables you to observe signal details that may otherwise be missed, particularly useful for seeing narrow pulses spaced far apart in time.

Peak-to-peak (V_{p-p}) – The voltage measured from the maximum point of a signal to its minimum point.

Period – The amount of time it takes a wave to complete one cycle. The period equals 1/frequency.

Phase – The amount of time that passes from the beginning of a cycle to the beginning of the next cycle, measured in degrees.

Phase Shift – The difference in timing between two otherwise similar signals.

Pre-trigger Viewing – The ability of a digital oscilloscope to capture what a signal did before a trigger event. Determines the length of viewable signal both preceding and following a trigger point.

Probe – An oscilloscope input device, usually having a pointed metal tip for making electrical contact with a circuit element, a lead to connect to the circuit's ground reference, and a flexible cable for transmitting the signal and ground to the oscilloscope.

Pulse – A common waveform shape that has a fast rising edge, a width, and a fast falling edge.

Pulse Train – A collection of pulses traveling together.

Pulse Width – The amount of time the pulse takes to go from low to high and back to low again, conventionally measured at 50% of full voltage.

R

Ramps – Transitions between voltage levels of sine waves that change at a constant rate.

Raster – A type of display.

Real-time Sampling – A sampling mode in which the oscilloscope collects as many samples as possible from one triggered acquisition. Ideal for signals whose frequency range is less than half the oscilloscope's maximum sample rate.

Record Length – The number of waveform points used to create a record of a signal.

Rise Time – The time taken for the leading edge of a pulse to rise from its low to its high values, typically measured from 10% to 90%.

S

Sampling – The conversion of a portion of an input signal into a number of discrete electrical values for the purpose of storage, processing and/or display by an oscilloscope. Two types: real-time sampling and equivalent-time sampling.

Sample Point – The raw data from an ADC used to calculate waveform points.

Sample Rate – Refers to how frequently a digital oscilloscope takes a sample of the signal, specified in samples per second (S/s).

Sensor – A device that converts a specific physical quantity such as sound, pressure, strain, or light intensity into an electrical signal.

Signal Integrity – The accurate reconstruction of a signal, determined by the systems and performance considerations of an oscilloscope, in addition to the probe used to acquire the signal.

Signal Source – A test device used to inject a signal into a circuit input; the circuit's output is then read by an oscilloscope. Also known as a signal generator.

Sine Wave – A common curved wave shape that is mathematically defined.

Single Shot – A signal measured by an oscilloscope that only occurs once (also called a transient event).

Single Sweep – A trigger mode to display one triggered screen of a signal and then stop.

Slope – On a graph or an oscilloscope display, the ratio of a vertical distance to a horizontal distance. A positive slope increases from left to right, while a negative slope decreases from left to right.

Square Wave – A common wave shape consisting of repeating square pulses.

Sweep – One horizontal pass of an anlog oscilloscope's electron beam from left to right across the CRT screen.

Sweep Speed – Same as the time base.

T

Time Base – Oscilloscope circuitry that controls the timing of the sweep. The time base is set by the seconds/division control.

Trace – The visible shapes drawn on a CRT by the movement of the electron beam.

Transient – A signal measured by an oscilloscope that only occurs once (also called a single–shot event).

Trigger – The circuit that references a horizontal sweep on an oscilloscope.

Trigger Holdoff – A control that allows you to adjust the period of time after a valid trigger during which the oscilloscope cannot trigger.

Trigger Level – The voltage level that a trigger source signal must reach before the trigger circuit initiates a sweep.

Trigger Mode – A mode that determines whether or not the oscilloscope draws a waveform if it does not detect a trigger. Common trigger modes include normal and auto.

Trigger Slope – The slope that a trigger source signal must reach before the trigger circuit initiates a sweep.

V

Vertical Resolution (Analog-to-Digital Converter) – An indication of how precisely an analog-to-digital converter (ADC) in a digital oscilloscope can convert input voltages into digital values, measured in bits. Calculation techniques, such as hi res acquisition mode, can improve the effective resolution.

Vertical Sensitivity – An indication of how much the vertical amplifier can amplify a weak signal – usually measured in millivolts (mV) per division.

Volt – The unit of electric potential difference.

Voltage – The difference in electric potential, expressed in volts, between two points.

W

Wave – The generic term for a pattern that repeats over time. Common types include: sine, square, rectangular, sawtooth, triangle, step, pulse, periodic, non-periodic, synchronous, asynchronous.

Waveform – A graphic representation of a voltage varying over time.

Waveform Capture Rate – Refers to how quickly an oscilloscope acquires waveforms, expressed as waveforms per second (wfms/s).

Waveform Point – A digital value that represents the voltage of a signal at a specific point in time. Waveform points are calculated from sample points and stored in memory.

Writing Speed – The ability of an analog oscilloscope to provide a visible trace of the movement of a signal from one point to another. This ability is restrictive for low-repetition signals that have fast-moving details, such as digital logic signals.

XY Mode – A measuremenbt technique that involves inputting one signal into the vertical system, as usual, and one into the horizontal system to trace voltages on both the X and Y axis.

Z

Z Axis – The display attribute on an oscilloscope that shows brightness variations as the trace is formed.

Contact Tektronix:

ASEAN / Australasia	(65) 6356 3900
Austria	00800 2255 4835*
Balkans, Israel, South Africa and other ISE Countries	+41 52 675 3777
Belgium	00800 2255 4835*
Brazil	+55 (11) 3759 7600
Canada	1 (800) 833-9200
Central East Europe, Ukraine and the Baltics	+41 52 675 3777
Central Europe & Greece	+41 52 675 3777
Denmark	+45 80 88 1401
Finland	+41 52 675 3777
France	00800 2255 4835*
Germany	00800 2255 4835*
Hong Kong	400-820-5835
India	(91) 80-42922600
Italy	00800 2255 4835*
Japan	81 (3) 6714-3010
Luxembourg	+41 52 675 3777
Mexico, Central/South America & Caribbean	52 (55) 53 35 10 85
Middle East, Asia and North Africa	+41 52 675 3777
The Netherlands	00800 2255 4835*
Norway	800 16098
People's Republic of China	400-820-5835
Poland	+41 52 675 3777
Portugal	80 08 12370
Republic of Korea	82 (2) 6917-5000
Russia & CIS	+7 (495) 7484900
South Africa	+27 11 206 8360
Spain	00800 2255 4835*
Sweden	00800 2255 4835*
Switzerland	00800 2255 4835*
Taiwan	886 (2) 2722-9622
United Kingdom & Ireland	00800 2255 4835*
USA	1 (800) 833-9200

* European toll-free number. If not accessible, call: +41 52 675 3777

Contact List Updated 09 December 2009

For Further Information

Tektronix maintains a comprehensive, constantly expanding collection of application notes, technical briefs and other resources to help engineers working on the cutting edge of technology. Please visit **www.tektronix.com**

Copyright © 2009, Tektronix. All rights reserved. Tektronix products are covered by U.S. and foreign patents, issued and pending. Information in this publication supersedes that in all previously published material. Specification and price change privileges reserved. TEKTRONIX and TEK are registered trademarks of Tektronix, Inc. All other trade names referenced are the service marks, trademarks or registered trademarks of their respective companies.

12/09 JS 03W-8605-5
TEK1511

Tektronix®

Springer Handbook of Nanotechnology

Springer Handbook provides a concise compilation of approved key information on methods of research, general principles, and functional relationships in physics and engineering. The world's leading experts in the fields of physics and engineering will be assigned by one or several renowned editors to write the chapters comprising each volume. The content is selected by these experts from Springer sources (books, journals, online content) and other systematic and approved recent publications of physical and technical information.

The volumes will be designed to be useful as readable desk reference book to give a fast and comprehensive overview and easy retrieval of essential reliable key information, including tables, graphs, and bibliographies. References to extensive sources are provided.

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Springer Handbook of Nanotechnology

Bharat Bhushan (Ed.)

With 972 Figures and 71 Tables

Springer

Professor Bharat Bhushan
Nanotribology Laboratory
for Information Storage and MEMS/NEMS
The Ohio State University
206 W. 18th Avenue
Columbus, Ohio 43210-1107
USA

Library of Congress Cataloging-in-Publication Data
Springer handbook of nanotechnology / Bharat Bhushan (ed.)
p. cm.
Includes bibliographical references and index
ISBN 3-540-01218-4 (alk. paper)
1. Nanotechnology--Handbooks, manuals, etc. I. Bhushan, Bharat; 1949-

T174.7S67 2003
620′.5--dc22 2003064953

ISBN 3-540-01218-4
Spinger-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September, 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag is a part of Springer Science+Business Media

springeronline.com

©Springer-Verlag Berlin Heidelberg 2004
Printed in Germany

The use of designations, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Product liability: The publisher cannot guarantee the accuracy of any information about dosage and application contained in this book. In every individual case the user must check such information by consulting the relevant literature.

Production and typesetting: LE-TeX GbR, Leipzig
Handbook coordinator: Dr. W. Skolaut, Heidelberg
Typography, layout and illustrations: schreiberVIS, Seeheim
Cover design: eStudio Calamar Steinen, Barcelona
Cover production: *design&production* GmbH, Heidelberg
Printing and binding: Stürtz AG, Würzburg

Printed on acid-free paper
SPIN 10890790 62/3141/YL 5 4 3 2 1 0

Foreword by Neal Lane

Prof. Neal Lane

University Professor
Department of Physics and Astronomy and James A. Baker III Institute for Public Policy Rice University Houston, Texas USA

Served in the Clinton Administration as Assistant to the President for Science and Technology and Director of the White House Office of Science and Technology Policy (1998–2001) and, prior to that, as Director of the National Science Foundation (1993–1998). While at the White House, he was instrumental in creating NNI.

In a January 2000 speech at the California Institute of Technology, former President W. J. Clinton talked about the exciting promise of "nanotechnology" and the importance of expanding research in nanoscale science and engineering and in the physical sciences, more broadly. Later that month, he announced in his State of the Union Address an ambitious $ 497 million federal, multi-agency national nanotechnology initiative (NNI) in the fiscal year 2001 budget; and he made the NNI a top science and technology priority within a budget that emphasized increased investment in U.S. scientific research. With strong bipartisan support in Congress, most of this request was appropriated, and the NNI was born.

Nanotechnology is the ability to manipulate individual atoms and molecules to produce nanostructured materials and sub-micron objects that have applications in the real world. Nanotechnology involves the production and application of physical, chemical and biological systems at scales ranging from individual atoms or molecules to about 100 nanometers, as well as the integration of the resulting nanostructures into larger systems. Nanotechnology is likely to have a profound impact on our economy and society in the early 21st century, perhaps comparable to that of information technology or advances in cellular and molecular biology. Science and engineering research in nanotechnology promises breakthroughs in areas such as materials and manufacturing, electronics, medicine and healthcare, energy and the environment, biotechnology, information technology and national security. It is widely felt that nanotechnology will lead to the next industrial revolution.

Nanometer-scale features are built up from their elemental constituents. Micro- and nanosystems components are fabricated using batch-processing techniques that are compatible with integrated circuits and range in size from micro- to nanometers. Micro- and nanosystems include Micro/NanoElectroMechanical Systems (MEMS/NEMS), micromechatronics, optoelectronics, microfluidics and systems integration. These systems can sense, control, and activate on the micro/nanoscale and can function individually or in arrays to generate effects on the macroscale. Due to the enabling nature of these systems and the significant impact they can have on both the commercial and defense applications, industry as well as the federal government have taken special interest in seeing growth nurtured in this field. Micro- and nanosystems are the next logical step in the "silicon revolution".

The discovery of novel materials, processes, and phenomena at the nanoscale and the development of new experimental and theoretical techniques for research provide fresh opportunities for the development of innovative nanosystems and nanostructured materials. There is an increasing need for a multidisciplinary, systems-oriented approach to manufacturing micro/nanodevices which function reliably. This can only be achieved through the cross-fertilization of ideas from different disciplines and the systematic flow of information and people among research groups.

Nanotechnology is a broad, highly interdisciplinary, and still evolving field. Covering even the most important aspects of nanotechnology in a single book that reaches readers ranging from students to active researchers in academia and industry is an enormous challenge. To prepare such a wide-ranging book on nanotechnology, Professor Bhushan has harnessed his own knowledge and experience, gained in several industries and universities, and has assembled about 90 internationally recognized authors from three continents to write 38 chapters. The authors come from both academia and industry.

Professor Bharat Bhushan's comprehensive book is intended to serve both as a textbook for university courses as well as a reference for researchers. It is a timely addition to the literature on nanotechnology, which I anticipate will stimulate further interest in this important new field and serve as an invaluable resource to members of the international scientific and industrial community.

The Editor-in-Chief and his team are to be warmly congratulated for bringing together this exclusive, timely, and useful Nanotechnology Handbook.

Foreword by James R. Heath

Nanotechnology has become an increasingly popular buzzword over the past five years or so, a trend that has been fueled by a global set of publicly funded nanotechnology initiatives. Even as researchers have been struggling to demonstrate some of the most fundamental and simple aspects of this field, the term nanotechnology has entered into the public consciousness through articles in the popular press and popular fiction. As a consequence, the expectations of the public are high for nanotechnology, even while the actual public definition of nanotechnology remains a bit fuzzy.

Why shouldn't those expectations be high? The late 1990's witnessed a major information technology (IT) revolution and a minor biotechnology revolution. The IT revolution impacted virtually every aspect of life in the western world. I am sitting on an airplane at 30,000 feet at the moment, working on my laptop, as are about half of the other passengers on this plane. The plane itself is riddled with computational and communications equipment. As soon as we land, many of us will pull out cell phones, others will check email via wireless modem, some will do both. This picture would be the same if I was landing in Los Angeles, Beijing, or Capetown. I will probably never actually print this text, but will instead submit it electronically. All of this was unthinkable a dozen years ago. It is therefore no wonder that the public expects marvelous things to happen quickly. However, the science that laid the groundwork for the IT revolution dates back 60 years or more, with its origins in the fundamental solid state physics.

By contrast, the biotech revolution was relatively minor and, at least to date, not particularly effective. The major diseases that plagued mankind a quarter century ago are still here. In some third world countries, the average lifespan of individuals has actually decreased from where it was a full century ago. While the costs of electronics technologies have plummeted, health care costs have continued to rise. The biotech revolution may have a profound impact, but the task at hand is substantially more difficult to what was required for the IT revolution. In effect, the IT revolution was based on the advanced engineering of two-dimensional digital circuits constructed from relatively simple components – extended solids. The biotech revolution is really dependent upon the ability to reverse engineer three-dimensional analog systems constructed from quite complex components – proteins. Given that the basic science behind biotech is substantially younger than the science that has supported IT, it is perhaps not surprising that the biotech revolution has not really been a proper revolution yet, and it likely needs at least another decade or so to come to fruition.

Where does nanotechnology fit into this picture? In many ways, nanotechnology depends upon the ability to engineer two- and three-dimensional systems constructed from complex components such as macromolecules, biomolecules, nanostructured solids, etc. Furthermore, in terms of patents, publications, and other metrics that can be used to gauge the birth and evolution of a field, nanotech lags some 15–20 years behind biotech. Thus, now is the time that the fundamental science behind nanotechnology is being explored and developed. Nevertheless, progress with that science is moving forward at a dramatic pace. If the scientific community can keep up this pace and if the public sector will continue to support this science, then it is possible, and perhaps even likely, that in 20 years from now we may be speaking of the nanotech revolution.

The Nanotechnology Handbook is timely in assembling chapters in the broad field of nanotechnology with an emphasis on reliability. The handbook should be a valuable reference for experienced researchers as well as for a novice in the field.

Prof. James R. Heath

Department of Chemistry
Mail Code: 127-72
California Institute of Technology
Pasadena, CA 91125, USA

Worked in the group of Nobel Laureate Richard E. Smalley at Rice University (1984–88) and co-invented Fullerene molecules which led to a revolution in Chemistry including the realization of nanotubes. The work on Fullerene molecules was cited for the 1996 Nobel Prize in Chemistry. Later he joined the University of California at Los Angeles (1994–2002), and co-founded and served as a Scientific Director of The California Nanosystems Institute.

Preface

On December 29, 1959 at the California Institute of Technology, Nobel Laureate Richard P. Feynman gave a talk at the Annual meeting of the American Physical Society that has become one classic science lecture of the 20th century, titled "There's Plenty of Room at the Bottom." He presented a technological vision of extreme miniaturization in 1959, several years before the word "chip" became part of the lexicon. He talked about the problem of manipulating and controlling things on a small scale. Extrapolating from known physical laws, Feynman envisioned a technology using the ultimate toolbox of nature, building nanoobjects atom by atom or molecule by molecule. Since the 1980s, many inventions and discoveries in fabrication of nanoobjects have been a testament to his vision. In recognition of this reality, in a January 2000 speech at the same institute, former President W. J. Clinton talked about the exciting promise of "nanotechnology" and the importance of expanding research in nanoscale science and engineering. Later that month, he announced in his State of the Union Address an ambitious $ 497 million federal, multi-agency national nanotechnology initiative (NNI) in the fiscal year 2001 budget, and made the NNI a top science and technology priority. Nanotechnology literally means any technology done on a nanoscale that has applications in the real world. Nanotechnology encompasses production and application of physical, chemical and biological systems at size scales, ranging from individual atoms or molecules to submicron dimensions as well as the integration of the resulting nanostructures into larger systems. Nanofabrication methods include the manipulation or self-assembly of individual atoms, molecules, or molecular structures to produce nanostructured materials and sub-micron devices. Micro- and nanosystems components are fabricated using top-down lithographic and nonlithographic fabrication techniques. Nanotechnology will have a profound impact on our economy and society in the early 21st century, comparable to that of semiconductor technology, information technology, or advances in cellular and molecular biology. The research and development in nanotechnology will lead to potential breakthroughs in areas such as materials and manufacturing, nanoelectronics, medicine and healthcare, energy, biotechnology, information technology and national security. It is widely felt that nanotechnology will lead to the next industrial revolution.

Reliability is a critical technology for many micro- and nanosystems and nanostructured materials. No book exists on this emerging field. A broad based handbook is needed. The purpose of this handbook is to present an overview of nanomaterial synthesis, micro/nanofabrication, micro- and nanocomponents and systems, reliability issues (including nanotribology and nanomechanics) for nanotechnology, and industrial applications. The chapters have been written by internationally recognized experts in the field, from academia, national research labs and industry from all over the world.

The handbook integrates knowledge from the fabrication, mechanics, materials science and reliability points of view. This book is intended for three types of readers: graduate students of nanotechnology, researchers in academia and industry who are active or intend to become active in this field, and practicing engineers and scientists who have encountered a problem and hope to solve it as expeditiously as possible. The handbook should serve as an excellent text for one or two semester graduate courses in nanotechnology in mechanical engineering, materials science, applied physics, or applied chemistry.

We embarked on this project in February 2002, and we worked very hard to get all the chapters to the publisher in a record time of about 1 year. I wish to sincerely thank the authors for offering to write comprehensive chapters on a tight schedule. This is generally an added responsibility in the hectic work schedules of researchers today. I depended on a large number of reviewers who provided critical reviews. I would like to thank Dr. Phillip J. Bond, Chief of Staff and Under Secretary for Technology, US Department of Commerce, Washington, D.C. for suggestions for chapters as well as authors in the handbook. I would also like to thank my colleague, Dr. Huiwen Liu, whose efforts during the preparation of this handbook were very useful.

I hope that this handbook will stimulate further interest in this important new field, and the readers of this handbook will find it useful.

September 2003

Bharat Bhushan
Editor

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Editors Vita

Dr. Bharat Bhushan received an M.S. in mechanical engineering from the Massachusetts Institute of Technology in 1971, an M.S. in mechanics and a Ph.D. in mechanical engineering from the University of Colorado at Boulder in 1973 and 1976, respectively, an MBA from Rensselaer Polytechnic Institute at Troy, NY in 1980, Doctor Technicae from the University of Trondheim at Trondheim, Norway in 1990, a Doctor of Technical Sciences from the Warsaw University of Technology at Warsaw, Poland in 1996, and Doctor Honouris Causa from the Metal-Polymer Research Institute of National Academy of Sciences at Gomel, Belarus in 2000. He is a registered professional engineer (mechanical). He is presently an Ohio Eminent Scholar and The Howard D. Winbigler Professor in the Department of Mechanical Engineering, Graduate Research Faculty Advisor in the Department of Materials Science and Engineering, and the Director of the Nanotribology Laboratory for Information Storage & MEMS/NEMS (NLIM) at the Ohio State University, Columbus, Ohio. He is an internationally recognized expert of tribology on the macro- to nanoscales, and is one of the most prolific authors in the field. He is considered by some a pioneer of the tribology and mechanics of magnetic storage devices and a leading researcher in the fields of nanotribology and nanomechanics using scanning probe microscopy and applications to micro/nanotechnology. He has authored 5 technical books, 45 handbook chapters, more than 450 technical papers in referred journals, and more than 60 technical reports, edited more than 25 books, and holds 14 U.S. patents. He is founding editor-in-chief of World Scientific Advances in Information Storage Systems Series, CRC Press Mechanics and Materials Science Series, and Microsystem Technologies – Micro- & Nanosystems and Information Storage & Processing Systems (formerly called Journal of Information Storage and Processing Systems). He has given more than 250 invited presentations on five continents and more than 60 keynote/plenary addresses at major international conferences.

Dr. Bhushan is an accomplished organizer. He organized the first symposium on Tribology and Mechanics of Magnetic Storage Systems in 1984 and the first international symposium on Advances in Information Storage Systems in 1990, both of which are now held annually. He is the founder of an ASME Information Storage and Processing Systems Division founded in 1993 and served as the founding chair during 1993–1998. His biography has been listed in over two dozen Who's Who books including Who's Who in the World and has received more than a dozen awards for his contributions to science and technology from professional societies, industry, and U.S. government agencies. He is also the recipient of various international fellowships including the Alexander von Humboldt Research Prize for Senior Scientists, Max Planck Foundation Research Award for Outstanding Foreign Scientists, and the Fulbright Senior Scholar Award. He is a foreign member of the International Academy of Engineering (Russia), Belorussian Academy of Engineering and Technology and the Academy of Triboengineering of Ukraine, an honorary member of the Society of Tribologists of Belarus, a fellow of ASME, IEEE, and the New York Academy of Sciences, and a member of STLE, ASEE, Sigma Xi and Tau Beta Pi.

Dr. Bhushan has previously worked for the R & D Division of Mechanical Technology Inc., Latham, NY; the Technology Services Division of SKF Industries Inc., King of Prussia, PA; the General Products Division Laboratory of IBM Corporation, Tucson, AZ; and the Almaden Research Center of IBM Corporation, San Jose, CA.

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

List of Authors

Chong H. Ahn
University of Cincinnati
Department of Electrical and Computer Engineering and Computer Science
814 Rhodes Hall
Cincinnati, OH 45221-0030, USA
e-mail: *chong.ahn@uc.edu*

Boris Anczykowski
nanoAnalytics GmbH
Gievenbecker Weg 11
48149 Münster, Germany
e-mail: *anczykowski@nanoanalytics.com*

Massood Z. Atashbar
Western Michigan University
Department of Electrical and Computer Engineering
200 Union Street SE
Kalamazoo, MI 49008-5329, USA
e-mail: *massood.atashbar@wmich.edu*

Wolfgang Bacsa
Université Paul Sabatier
Laboratoire de Physique des Solides (LPST)
118 Route de Narbonne
31062 Toulouse Cedex 4, France
e-mail: *bacsa@lpst.ups-tlse.fr*

William Sims Bainbridge
National Science Foundation
Division of Information and Intelligent Systems
4201 Wilson Boulevard
Arlington, VA 22230, USA
e-mail: *wbainbri@nsf.gov*

Antonio Baldi
Institut de Microelectronica de Barcelona (IMB)
Centro National Microelectrónica (CNM-CSIC)
Campus Universitat Autonoma de Barcelona
08193 Barcelona, Spain
e-mail: *Antoni.baldi@cnm.es*

Philip D. Barnes
Ohio State University
Biomedical Engineering Center
1080 Carmack Road
Columbus, OH 43210, USA
e-mail: *d_skill@yahoo.com*

James D. Batteas
National Institute of Standards and Technology
Surface and Microanalysis Science Division
100 Bureau Drive Mailstop 8372
Gaithersburg, MD 20899-8372, USA
e-mail: *james.batteas@nist.gov*

Roland Bennewitz
McGill University
Physics Department
3600 rue University
Montreal, QC H3A 2T8, Canada
e-mail: *roland@physics.mcgill.ca*

Alan D. Berman
Monitor Venture Enterprises
241 S. Figueroa St. Suite 300
Los Angeles, CA 90012, USA
e-mail: *alan.berman.2001@anderson.ucla.edu*

Bharat Bhushan
The Ohio State University
Nanotribology Laboratory for Information Storage and MEMS/NEMS
206 W. 18th Avenue
Columbus, OH 43210-1107, USA
e-mail: *bhushan.2@osu.edu*

Gerd K. Binnig
IBM Zurich Research Laboratory
Micro-/Nanomechanics
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *gbi@zurich.ibm.com*

Marcie R. Black
Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: *marcie@alum.mit.edu*

Jean-Marc Broto
University Toulouse III
Laboratoire National
des Champs Magnétiques Pulsés (LNCMP)
143 Avenue de Rangueil
31432 Toulouse Cedex 4, France
e-mail: *broto@insa-tlse.fr*

Robert W. Carpick
University of Wisconsin-Madison
Department of Engineering Physics
1500 Engineering Drive
Madison, WI 53706-1687, USA
e-mail: *carpick@engr.wisc.edu*

Tsung-Lin Chen
National Chiao Tung University
Department of Mechanical Engineering
30050 Shin Chu, Taiwan
e-mail: *tsunglin@mail.nctu.edu.tw*

Yu-Ting Cheng
National Chiao Tung University
Department of Electronics Engineering
& Institute of Electronics
1001, Ta-Hsueh Road
300 HsinChu, Taiwan
e-mail: *ytcheng@faculty.nctu.edu.tw*

Giovanni Cherubini
IBM Zurich Research Laboratory
Storage Technologies
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *cbi@zurich.ibm.com*

Jin-Woo Choi
Louisiana State University
Department of Electrical
and Computer Engineering
102 South Campus Drive
Baton Rouge, LA 70803-5901, USA
e-mail: *choi@ece.lsu.edu*

Shawn J. Cunningham
WiSpry, Inc.
Colorado Springs Design Center
7150 Campus Drive, Suite 255
Colorado Springs, CO 80920, USA
e-mail: *shawn.cunningham@wispry.com*

Michel Despont
IBM Zurich Research Laboratory
Micro-/Nanomechanics
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *dpt@zurich.ibm.com*

Gene Dresselhaus
Massachusetts Institute of Technology
Francis Bitter Magnet Laboratory
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: *gene@mgm.mit.edu*

Mildred S. Dresselhaus
Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science and Department of Physics
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: *millie@mgm.mit.edu*

Martin L. Dunn
University of Colorado at Boulder
Department of Mechanical Engineering
Campus Box 427
Boulder, CO 80309, USA
e-mail: *martin.dunn@colorado.edu*

Urs T. Dürig
IBM Zurich Research Laboratory
Micro-/Nanomechanics
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *drg@zurich.ibm.com*

Evangelos Eleftheriou
IBM Zurich Research Laboratory
Storage Technologies
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *ele@zurich.ibm.com*

Mauro Ferrari
Ohio State University
Biomedical Engineering Center
1080 Carmack Road
Columbus, OH 43210-1002, USA
e-mail: *Ferrari.5@osu.edu*

Emmanuel Flahaut
Université Paul Sabatier
CIRIMAT (Centre Interuniversitaire de Recherche
et d'Ingénierie des Matériaux)
118 Route de Narbonne
31062 Toulouse Cedex 04, France
e-mail: *flahaut@chimie.ups-tlse.fr*

Lásló Forró
Swiss Federal Institute of Technology (EPFL)
Institute of Physics of Complex Matter
Ecublens
1015 Lausanne, Switzerland
e-mail: *laszlo.forro@epfl.ch*

Jane Frommer
IBM Almaden Research Center
Department of Science and Technology
650 Harry Road
San Jose, CA 95120, USA
e-mail: *frommer@Almaden.ibm.com*

Harald Fuchs
Universität Münster
Physikalisches Institut
Wilhelm-Klemm-Straße 10
48149 Münster, Germany
e-mail: *fuchsh@uni-muenster.de*

Franz J. Giessibl
Universität Augsburg
Lehrstuhl für Experimentalphysik VI
Universitätsstraße 1
86135 Augsburg, Germany
e-mail: *franz.giessibl@physik.uni-augsburg.de*

Enrico Gnecco
University of Basel
Department of Physics
Klingelbergstraße 82
4056 Basel, Switzerland
e-mail: *Enrico.Gnecco@unibas.ch*

Gérard Gremaud
Swiss Federal Institute of Technology (EPFL)
Institute of Physics of Complex Matter
Ecublens
1015 Lausanne, Switzerland
e-mail: *gremaud@epfl.ch*

Jason H. Hafner
Rice University
Department of Physics & Astronomy
PO BOX 1892
Houston, TX 77251-1892, USA
e-mail: *hafner@rice.edu*

Stefan Hengsberger
University of Applied Science of Fribourg
Bd de Pérolles
1705 Fribourg, Switzerland
e-mail: *stefan.hengsberger@eif.ch*

Peter Hinterdorfer
Johannes Kepler University of Linz
Institute for Biophysics
Altenbergerstraße 69
4040 Linz, Austria
e-mail: *peter.hinterdorfer@jku.at*

Roberto Horowitz
University of California at Berkeley
Department of Mechanical Engineering
5121 Etcheverry Hall
Berkeley, CA 94720-1742, USA
e-mail: *horowitz@me.berkeley.edu*

Hirotaka Hosoi
Japan Science and Technology Corporation
Innovation Plaza, Hokkaido
060-0819 Sapporo, Japan
e-mail: *hosoi@sapporo.jst-plaza.jp*

Jacob N. Israelachvili
University of California
Department of Chemical Engineering
and Materials Department
Santa Barbara, CA 93106, USA
e-mail: *Jacob@engineering.ucsb.edu*

Ghassan E. Jabbour
University of Arizona
Optical Sciences Center
1630 East University Boulevard
Tucson, AZ 85721, USA
e-mail: *gej@optics.arizona.edu*

Harold Kahn
Case Western Reserve University
Department of Materials Science and Engineering
10900 Euclid Avenue
Cleveland, OH 44106-7204, USA
e-mail: *kahn@cwru.edu*

András Kis
Swiss Federal Institute of Technology (EPFL)
Institute of Physics of Complex Matter
Ecublens
1015 Lausanne, Switzerland
e-mail: *andras@igahpse.epfl.ch*

Jané Kondev
Brandeis University
Physics Department
Waltham, MA 02454, USA
e-mail: *kondev@brandeis.edu*

Andrzej J. Kulik
Swiss Federal Institute of Technology (EPFL)
Institute of Physics of Complex Matter
1015 Lausanne, Switzerland
e-mail: *andrzej.kulik@epfl.ch*

Christophe Laurent
Université Paul Sabatier
CIRIMAT (Centre Interuniversitaire de Recherche
et d'Ingénierie des Matériaux)
118 Route de Narbonne
31062 Toulouse Cedex 04, France
e-mail: *laurent@chimie.ups-tlse.fr*

Stephen C. Lee
Ohio State University
Biomedical Engineering Center
1080 Carmack Road
Columbus, OH 43210-1002, USA
e-mail: *Lee@bme.ohio-state.edu*

Yunfeng Li
University of California at Berkeley
Department of Mechanical Engineering
5121 Etcheverry Hall
Berkeley, CA 94720-1740, USA
e-mail: *yunfeng@me.Berkeley.edu*

Liwei Lin
University of California at Berkeley
Mechanical Engineering Department
5126 Etcheverry
Berkeley, CA 94720-1740, USA
e-mail: *lwlin@me.berkeley.edu*

Yu-Ming Lin
Massachusetts Institute of Technology
Department of Electrical Engineering
and Computer Science
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: *yming@mgm.mit.edu*

Huiwen Liu
Ohio State University
Nanotribology Laboratory for Information Storage and MEMS/NEMS
3070 St John Ct 7
Columbus, OH 43210-1107, USA
e-mail: *liu.403@osu.edu*

Adrian B. Mann
Rutgers University
Department of Ceramics and Materials Engineering
607 Taylor Road
Piscataway, NJ 08854, USA
e-mail: *abmann@rci.rutgers.edu*

Othmar Marti
University of Ulm
Department of Experimental Physics
Albert-Einstein-Allee 11
89069 Ulm, Germany
e-mail: *Othmar.Marti@physik.uni-ulm.de*

Jack Martin
Analog Devices, Inc.
Micromachined Products Division
21 Osborn Street
Cambridge, MA 02139, USA
e-mail: *jack.martin@analog.com*

Brendan McCarthy
University of Arizona
Optical Sciences Center
1630 East University Boulevard
Tucson, AZ 85721, USA
e-mail: *bmccarthy@optics.arizona.edu*

Mehran Mehregany
Case Western Reserve University
Department of Electrical Engineering
and Computer Science
188 Bingham Building
Cleveland, OH 44106, USA
e-mail: *mxm31@cwru.edu*

Ernst Meyer
University of Basel
Institute of Physics
Klingelbergstraße 82
4056 Basel, Switzerland
e-mail: *Ernst.Meyer@unibas.ch*

Marc Monthioux
UPR A-8011 CNRS
Centre d'Elaboration des Matériaux
et d'Etudes Structurales (CEMES)
29 Rue Jeanne Marvig
31055 Toulouse Cedex 4, France
e-mail: *monthiou@cemes.fr*

Markus Morgenstern
University of Hamburg
Institute of Applied Physics
Jungiusstraße 11
20355 Hamburg, Germany
e-mail: *mmorgens@physnet.uni-hamburg.de*

Seizo Morita
Osaka University
Department of Electronic Engineering
Yamada-Oka 2-1
565-0871 Suita-Citiy, Osaka, Japan
e-mail: *smorita@ele.eng.osaka-u.ac.jp*

Koichi Mukasa
Hokkaido University
Nanoelectronics Laboratory
Nishi-8, Kita-13, Kita-ku
060-8628 Sapporo, Japan
e-mail: *mukasa@nano.eng.hokudai.ac.jp*

Martin H. Müser
University of Western Ontario
Department of Applied Mathematics
WSC 139, Faculty of Science
London, Ontario N6A 5B7, Canada
e-mail: *mmuser@uwo.ca*

Kenn Oldham
University of California at Berkeley
Department of Mechanical Engineering
5121 Etcheverry Hall
Berkeley, CA 94720-1740, USA
e-mail: *oldham@newton.berkeley.edu*

Hiroshi Onishi
Kanagawa Academy of Science and Technology
Surface Chemistry Laboratory
KSP East 404, 3-2-1 Sakado, Takatsu-ku,
Kawasaki-shi
213-0012 Kanagawa, Japan
e-mail: *oni@net.ksp.or.jp*

René M. Overney
University of Washington
Department of Chemical Engineering
Seattle, WA 98195-1750, USA
e-mail: *roverney@u.Washington.edu*

Alain Peigney
Université Paul Sabatier
CIRIMAT (Centre Inter-universitaire de Recherches
et d'Ingénierie des Matériaux) – UMR CNRS 5085
118 Route de Narbonne
31062 Toulouse Cedex 4, France
e-mail: *peigney@chimie.ups-tlse.fr*

Oliver Pfeiffer
University of Basel
Institute of Physics
Klingelbergstraß 82
4056 Basel, Switzerland
e-mail: *Oliver.Pfeiffer@stud.unibas.ch*

Rob Phillips
California Institute of Technology
Mechanical Engineering and Applied Physics
1200 California Boulevard
Pasadena, CA 91125, USA
e-mail: *phillips@aero.caltech.edu*

Haralampos Pozidis
IBM Zurich Research Laboratory
Storage Technologies
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *hap@zurich.ibm.com*

Prashant K. Purohit
California Institute of Technology
Mechanical Engineering
1200 California Boulevard
Pasadena, CA 91125, USA
e-mail: *prashant@caltech.edu*

Oded Rabin
Massachusetts Institute of Technology
Department of Chemistry
77 Massachusetts Avenue
Cambridge, MA 02139, USA
e-mail: *oded@mgm.mit.edu*

Françisco M. Raymo
University of Miami
Department of Chemistry
1301 Memorial Drive
Coral Gables, FL 33146-0431, USA
e-mail: *fraymo@miami.edu*

Manitra Razafinimanana
Université Paul Sabatier
Centre de Physique des Plasmas
et leurs Applications (CPPAT)
118 Route de Narbonne
31062 Toulouse Cedex, France
e-mail: *razafinimanana@cpat.ups-tlse.fr*

Mark O. Robbins
Johns Hopkins University
Department of Physics and Astronomy
3400 North Charles Street
Baltimore, MD 21218, USA
e-mail: *mr@jhu.edu*

John A. Rogers
University of Illinois
Department of Materials Science and Engineering
1304 W. Green Street
Urbana, IL 61801, USA
e-mail: *jrogers@uiuc.edu*

Mark Ruegsegger
Ohio State University
Biomedical Engineering Center
1080 Carmack Road
Columbus, OH 43210, USA
e-mail: *mark@bme.ohio-state.edu*

Marina Ruths
Åbo Akademi University
Department of Physical Chemistry
Porthansgatan 3–5
20500 Åbo, Finland
e-mail: *mruths@abo.fi*

Dror Sarid
University of Arizona
Optical Sciences Center
1630 East University Boulevard
Tucson, AZ 85721, USA
e-mail: *sarid@optics.arizona.edu*

Akira Sasahara
Kanagawa Academy of Science and Technology
Surface Chemistry Laboratory
KSP East 404, 3-2-1 Sakado, Takatsu-ku,
Kawasaki-shi
213-0012 Kanagawa, Japan
e-mail: *ryo@net.ksp.or.jp*

André Schirmeisen
University of Münster
Institute of Physics
Whilhem-Klemm-Straße 10
48149 Münster, Germany
e-mail: *schira@uni-muenster.de*

Alexander Schwarz
University of Hamburg
Institute of Applied Physics
Jungiusstraße 11
20355 Hamburg, Germany
e-mail: *aschwarz@physnet.uni-hamburg.de*

Udo D. Schwarz
Yale University
Department of Mechanical Engineering
15 Prospect Street
New Haven, CT 06510, USA
e-mail: *udo.schwarz@yale.edu*

Philippe Serp
Ecole Nationale Supèrieure d'Ingénieurs
en Arts Chimiques et Technologiques
Laboratoire de Catalyse, Chimie Fine et Polymères
118 Route de Narbonne
31077 Toulouse, France
e-mail: *Philippe.Serp@ensiacet.fr*

Bryan R. Smith
Ohio State University
Biomedical Engineering Center
1080 Carmack Road
Columbus, OH 43210, USA
e-mail: *bryan@bme.ohio-state.edu*

Anisoara Socoliuc
University of Basel
Institute of Physics
Klingelbergstraße 82
4056 Basel, Switzerland
e-mail: *A.Socoliuc@unibas.ch*

Yasuhiro Sugawara
Osaka University
Department of Applied Physics
Yamada-Oka 2-1
565-0871 Suita, Japan
e-mail: *sugawara@ap.eng.osaka-u.ac.jp*

George W. Tyndall
IBM Almaden Research Center
Science and Technology
6369 Didion Ct.
San Jose, CA 95123, USA
e-mail: *tyndallgw@netscape.net*

Peter Vettiger
IBM Zurich Research Laboratory
Manager Micro-/Nanomechanics
Säumerstraße 4
8803 Rüschlikon, Switzerland
e-mail: *pv@zurich.ibm.com*

Darrin J. Young
Case Western Reserve University
Electrical Engineering and Computer Science
10900 Euclid Avenue
Cleveland, OH 44106, USA
e-mail: *djy@po.cwru.edu*

Babak Ziaie
University of Minnesota
Department of Electrical
and Computer Engineering
200 Union Street SE
Minneapolis, MN 55455, USA
e-mail: *ziaie@ece.umn.edu*

Christian A. Zorman
Case Western Reserve University
Department of Electrical Engineering
and Computer Science
719 Glennan Building
Cleveland, OH 44106, USA
e-mail: *caz@po.cwru.edu*

Philippe K. Zysset
Technische Universität Wien
Institut für Leichtbau und Flugzeugbau (ILFB)
Gußhausstraße 27–29
1040 Wien, Austria
e-mail: *philippe.zysset@epfl.ch*

Contents

List of Tables XXIX
List of Abbreviations XXXIII

1 **Introduction to Nanotechnology** 1
- 1.1 Background and Definition of Nanotechnology 1
- 1.2 Why Nano? 2
- 1.3 Lessons from Nature 2
- 1.4 Applications in Different Fields 3
- 1.5 Reliability Issues of MEMS/NEMS 4
- 1.6 Organization of the Handbook 5
- References 5

Part A Nanostructures, Micro/Nanofabrication, and Micro/Nanodevices

2 **Nanomaterials Synthesis and Applications: Molecule-Based Devices** 9
- 2.1 Chemical Approaches to Nanostructured Materials 10
- 2.2 Molecular Switches and Logic Gates 14
- 2.3 Solid State Devices 22
- 2.4 Conclusions and Outlook 35
- References 36

3 **Introduction to Carbon Nanotubes** 39
- 3.1 Structure of Carbon Nanotubes 40
- 3.2 Synthesis of Carbon Nanotubes 45
- 3.3 Growth Mechanisms of Carbon Nanotubes 59
- 3.4 Properties of Carbon Nanotubes 63
- 3.5 Carbon Nanotube-Based Nano-Objects 68
- 3.6 Applications of Carbon Nanotubes 73
- References 86

4 **Nanowires** 99
- 4.1 Synthesis 100
- 4.2 Characterization and Physical Properties of Nanowires 110
- 4.3 Applications 131
- 4.4 Concluding Remarks 138
- References 138

5 **Introduction to Micro/Nanofabrication** 147
- 5.1 Basic Microfabrication Techniques 148
- 5.2 MEMS Fabrication Techniques 159

5.3 Nanofabrication Techniques 170
References........ 180

6 **Stamping Techniques for Micro and Nanofabrication: Methods and Applications**........ 185
6.1 High Resolution Stamps........ 186
6.2 Microcontact Printing 187
6.3 Nanotransfer Printing........ 190
6.4 Applications........ 193
6.5 Conclusions........ 200
References........ 200

7 **Materials Aspects of Micro- and Nanoelectromechanical Systems**.. 203
7.1 Silicon 203
7.2 Germanium-Based Materials 210
7.3 Metals 211
7.4 Harsh Environment Semiconductors 212
7.5 GaAs, InP, and Related III-V Materials 217
7.6 Ferroelectric Materials 218
7.7 Polymer Materials 219
7.8 Future Trends........ 220
References........ 221

8 **MEMS/NEMS Devices and Applications**........ 225
8.1 MEMS Devices and Applications........ 227
8.2 NEMS Devices and Applications 246
8.3 Current Challenges and Future Trends 249
References........ 250

9 **Microfluidics and Their Applications to Lab-on-a-Chip** 253
9.1 Materials for Microfluidic Devices and Micro/Nano Fabrication Techniques 254
9.2 Active Microfluidic Devices........ 257
9.3 Smart Passive Microfluidic Devices........ 262
9.4 Lab-on-a-Chip for Biochemical Analysis 270
References........ 276

10 **Therapeutic Nanodevices**........ 279
10.1 Definitions and Scope of Discussion 280
10.2 Synthetic Approaches: "top-down" versus "bottom-up" Approaches for Nanotherapeutic Device Components 285
10.3 Technological and Biological Opportunities........ 288
10.4 Applications for Nanotherapeutic Devices........ 307
10.5 Concluding Remarks: Barriers to Practice and Prospects 315
References........ 317

Part B Scanning Probe Microscopy

11 **Scanning Probe Microscopy – Principle of Operation, Instrumentation, and Probes** 325
11.1 Scanning Tunneling Microscope 327
11.2 Atomic Force Microscope 331
11.3 AFM Instrumentation and Analyses 347
References 364

12 **Probes in Scanning Microscopies** 371
12.1 Atomic Force Microscopy 372
12.2 Scanning Tunneling Microscopy 382
References 383

13 **Noncontact Atomic Force Microscopy and Its Related Topics** 385
13.1 Principles of Noncontact Atomic Force Microscope (NC-AFM) 386
13.2 Applications to Semiconductors 391
13.3 Applications to Insulators 397
13.4 Applications to Molecules 404
References 407

14 **Low Temperature Scanning Probe Microscopy** 413
14.1 Microscope Operation at Low Temperatures 414
14.2 Instrumentation 415
14.3 Scanning Tunneling Microscopy and Spectroscopy 419
14.4 Scanning Force Microscopy and Spectroscopy 433
References 442

15 **Dynamic Force Microscopy** 449
15.1 Motivation: Measurement of a Single Atomic Bond 450
15.2 Harmonic Oscillator: A Model System for Dynamic AFM 454
15.3 Dynamic AFM Operational Modes 455
15.4 Q-Control 464
15.5 Dissipation Processes Measured with Dynamic AFM 468
15.6 Conclusion 471
References 471

16 **Molecular Recognition Force Microscopy** 475
16.1 Ligand Tip Chemistry 476
16.2 Fixation of Receptors to Probe Surfaces 478
16.3 Single-Molecule Recognition Force Detection 479
16.4 Principles of Molecular Recognition Force Spectroscopy 482
16.5 Recognition Force Spectroscopy: From Isolated Molecules to Biological Membranes 484
16.6 Recognition Imaging 489
16.7 Concluding Remarks 491
References 492

Part C Nanotribology and Nanomechanics

17 **Micro/Nanotribology and Materials Characterization Studies Using Scanning Probe Microscopy** ... 497
17.1 Description of AFM/FFM and Various Measurement Techniques ... 499
17.2 Friction and Adhesion ... 507
17.3 Scratching, Wear, Local Deformation, and Fabrication/Machining ... 518
17.4 Indentation ... 526
17.5 Boundary Lubrication ... 530
17.6 Closure ... 538
References ... 539

18 **Surface Forces and Nanorheology of Molecularly Thin Films** ... 543
18.1 Introduction: Types of Surface Forces ... 544
18.2 Methods Used to Study Surface Forces ... 546
18.3 Normal Forces Between Dry (Unlubricated) Surfaces ... 550
18.4 Normal Forces Between Surfaces in Liquids ... 554
18.5 Adhesion and Capillary Forces ... 564
18.6 Introduction: Different Modes of Friction and the Limits of Continuum Models ... 569
18.7 Relationship Between Adhesion and Friction Between Dry (Unlubricated and Solid Boundary Lubricated) Surfaces ... 571
18.8 Liquid Lubricated Surfaces ... 580
18.9 Role of Molecular Shape and Surface Structure in Friction ... 591
References ... 594

19 **Scanning Probe Studies of Nanoscale Adhesion Between Solids in the Presence of Liquids and Monolayer Films** ... 605
19.1 The Importance of Adhesion at the Nanoscale ... 605
19.2 Techniques for Measuring Adhesion ... 606
19.3 Calibration of Forces, Displacements, and Tips ... 610
19.4 The Effect of Liquid Capillaries on Adhesion ... 612
19.5 Self-Assembled Monolayers ... 618
19.6 Concluding Remarks ... 624
References ... 624

20 **Friction and Wear on the Atomic Scale** ... 631
20.1 Friction Force Microscopy in Ultra-High Vacuum ... 632
20.2 The Tomlinson Model ... 636
20.3 Friction Experiments on Atomic Scale ... 638
20.4 Thermal Effects on Atomic Friction ... 642
20.5 Geometry Effects in Nanocontacts ... 646
20.6 Wear on the Atomic Scale ... 649
20.7 Molecular Dynamics Simulations of Atomic Friction and Wear ... 651
20.8 Energy Dissipation in Noncontact Atomic Force Microscopy ... 654
20.9 Conclusion ... 656
References ... 657

21 **Nanoscale Mechanical Properties – Measuring Techniques and Applications** 661
- 21.1 Local Mechanical Spectroscopy by Contact AFM 662
- 21.2 Static Methods – Mesoscopic Samples 667
- 21.3 Scanning Nanoindentation: An Application to Bone Tissue 674
- 21.4 Conclusions and Perspectives 682
- References 682

22 **Nanomechanical Properties of Solid Surfaces and Thin Films** 687
- 22.1 Instrumentation 688
- 22.2 Data Analysis 694
- 22.3 Modes of Deformation 702
- 22.4 Thin Films and Multilayers 707
- 22.5 Developing Areas 711
- References 712

23 **Atomistic Computer Simulations of Nanotribology** 717
- 23.1 Molecular Dynamics 718
- 23.2 Friction Mechanisms at the Atomic Scale 723
- 23.3 Stick–Slip Dynamics 732
- 23.4 Conclusions 734
- References 735

24 **Mechanics of Biological Nanotechnology** 739
- 24.1 Science at the Biology–Nanotechnology Interface 740
- 24.2 Scales at the Bio–Nano Interface 746
- 24.3 Modeling at the Nano–Bio Interface 752
- 24.4 Nature's Nanotechnology Revealed: Viruses as a Case Study 755
- 24.5 Concluding Remarks 760
- References 761

25 **Mechanical Properties of Nanostructures** 763
- 25.1 Experimental Techniques for Measurement of Mechanical Properties of Nanostructures 765
- 25.2 Experimental Results and Discussion 770
- 25.3 Finite Element Analysis of Nanostructures with Roughness and Scratches 778
- 25.4 Closure 785
- References 786

Part D Molecularly Thick Films for Lubrication

26 **Nanotribology of Ultrathin and Hard Amorphous Carbon Films** 791
- 26.1 Description of Commonly Used Deposition Techniques 795
- 26.2 Chemical Characterization and Effect of Deposition Conditions on Chemical Characteristics and Physical Properties 800

26.3 Micromechanical and Tribological Characterizations of Coatings Deposited by Various Techniques ... 805
References ... 827

27 **Self-Assembled Monolayers for Controlling Adhesion, Friction and Wear** ... 831
27.1 A Primer to Organic Chemistry ... 834
27.2 Self-Assembled Monolayers: Substrates, Head Groups, Spacer Chains, and End Groups ... 839
27.3 Tribological Properties of SAMs ... 841
27.4 Closure ... 856
References ... 857

28 **Nanoscale Boundary Lubrication Studies** ... 861
28.1 Lubricants Details ... 862
28.2 Nanodeformation, Molecular Conformation, and Lubricant Spreading ... 864
28.3 Boundary Lubrication Studies ... 866
28.4 Closure ... 880
References ... 881

29 **Kinetics and Energetics in Nanolubrication** ... 883
29.1 Background: From Bulk to Molecular Lubrication ... 885
29.2 Thermal Activation Model of Lubricated Friction ... 887
29.3 Functional Behavior of Lubricated Friction ... 888
29.4 Thermodynamical Models Based on Small and Nonconforming Contacts ... 890
29.5 Limitation of the Gaussian Statistics – The Fractal Space ... 891
29.6 Fractal Mobility in Reactive Lubrication ... 892
29.7 Metastable Lubricant Systems in Large Conforming Contacts ... 894
29.8 Conclusion ... 895
References ... 895

Part E Industrial Applications and Microdevice Reliability

30 **Nanotechnology for Data Storage Applications** ... 899
30.1 Current Status of Commercial Data Storage Devices ... 901
30.2 Opportunities Offered by Nanotechnology for Data Storage ... 907
30.3 Conclusion ... 918
References ... 919

31 **The “Millipede” – A Nanotechnology-Based AFM Data-Storage System** ... 921
31.1 The Millipede Concept ... 923
31.2 Thermomechanical AFM Data Storage ... 924
31.3 Array Design, Technology, and Fabrication ... 926

31.4 Array Characterization ... 927
31.5 *x*/*y*/*z* Medium Microscanner ... 929
31.6 First Write/Read Results with the 32×32 Array Chip ... 931
31.7 Polymer Medium ... 932
31.8 Read Channel Model ... 939
31.9 System Aspects ... 943
31.10 Conclusions ... 948
References ... 948

32 **Microactuators for Dual-Stage Servo Systems in Magnetic Disk Files** ... 951
32.1 Design of the Electrostatic Microactuator ... 952
32.2 Fabrication ... 962
32.3 Servo Control Design of MEMS Microactuator Dual-Stage Servo Systems ... 968
32.4 Conclusions and Outlook ... 978
References ... 979

33 **Micro/Nanotribology of MEMS/NEMS Materials and Devices** ... 983
33.1 Introduction to MEMS ... 985
33.2 Introduction to NEMS ... 988
33.3 Tribological Issues in MEMS/NEMS ... 989
33.4 Tribological Studies of Silicon and Related Materials ... 995
33.5 Lubrication Studies for MEMS/NEMS ... 1003
33.6 Component-Level Studies ... 1009
References ... 1017

34 **Mechanical Properties of Micromachined Structures** ... 1023
34.1 Measuring Mechanical Properties of Films on Substrates ... 1023
34.2 Micromachined Structures for Measuring Mechanical Properties ... 1024
34.3 Measurements of Mechanical Properties ... 1034
References ... 1037

35 **Thermo- and Electromechanics of Thin-Film Microstructures** ... 1039
35.1 Thermomechanics of Multilayer Thin-Film Microstructures ... 1041
35.2 Electromechanics of Thin-Film Microstructures ... 1061
35.3 Summary and Mention of Topics not Covered ... 1078
References ... 1078

36 **High Volume Manufacturing and Field Stability of MEMS Products** ... 1083
36.1 Manufacturing Strategy ... 1086
36.2 Robust Manufacturing ... 1087
36.3 Stable Field Performance ... 1102
References ... 1106

37 **MEMS Packaging and Thermal Issues in Reliability** 1111
37.1 MEMS Packaging 1111
37.2 Hermetic and Vacuum Packaging and Applications 1116
37.3 Thermal Issues and Packaging Reliability 1122
37.4 Future Trends and Summary 1128
References 1129

Part F Social and Ethical Implication

38 **Social and Ethical Implications of Nanotechnology** 1135
38.1 Applications and Societal Impacts 1136
38.2 Technological Convergence 1139
38.3 Major Socio-technical Trends 1141
38.4 Sources of Ethical Behavior 1143
38.5 Public Opinion 1145
38.6 A Research Agenda 1148
References 1149

Acknowledgements 1153
About the Authors 1155
Detailed Contents 1171
Subject Index 1189

List of Tables

Part A Nanostructures, Micro/Nanofabrication, and Micro/Nanodevices

3 Introduction to Carbon Nanotubes

Table 3.1 Different carbon morphologies 50
Table 3.2 Carbon nanofilament morphologies and some basic synthesis conditions 62
Table 3.3 Adsorption properties and sites of SWNTs and MWNTs 65
Table 3.4 Preparation and catalytic performances of some nanotube-supported catalysts................................ 77
Table 3.5 Applications for nanotube-based multifunctional materials ... 84

4 Nanowires

Table 4.1 Selected syntheses of nanowires by material 101

5 Introduction to Micro/Nanofabrication

Table 5.1 Typical dry etch chemistries.. 157
Table 5.2 Surface micromachined sacrificial/structural combinations 165

7 Materials Aspects of Micro- and Nanoelectromechanical Systems

Table 7.1 Selected materials for MEMS and NEMS................................. 204

9 Microfluidics and Their Applications to Lab-on-a-Chip

Table 9.1 Overview of the different polymer micro/nano fabrication techniques 257

10 Therapeutic Nanodevices

Table 10.1 Ideal characteristics of nanodevices...................................... 282
Table 10.2 Orthogonal conjugation chemistries 295

Part B Scanning Probe Microscopy

11 Scanning Probe Microscopy – Principle of Operation, Instrumentation, and Probes

Table 11.1 Comparison of Various Conventional Microscopes with SPMs ... 326
Table 11.2 Relevant properties of materials used for cantilevers............. 339
Table 11.3 Measured vertical spring constants and natural frequencies of triangular (V-shaped) cantilevers made of PECVD Si_3N_4 341
Table 11.4 Vertical and torsional spring constants of rectangular cantilevers made of Si and PECVD.................... 341
Table 11.5 Noise in Interferometers 353

13 **Noncontact Atomic Force Microscopy and Its Related Topics**
Table 13.1 Comparison between the adatom heights observed in an AFM image and the variety of properties for inequivalent adatoms 393

Part C Nanotribology and Nanomechanics

17 **Micro/Nanotribology and Materials Characterization Studies Using Scanning Probe Microscopy**
Table 17.1 Comparison of typical operating parameters in SFA, STM, and AFM/FFM used for micro/nanotribological studies 498
Table 17.2 Surface roughness and coefficients of friction of various samples in air 518

18 **Surface Forces and Nanorheology of Molecularly Thin Films**
Table 18.1 Types of surface forces in vacuum vs. in liquid (colloidal forces) 545
Table 18.2 Van der Waals interaction energy and force between macroscopic bodies of different geometries 551
Table 18.3 Electrical "double-layer" interaction energy and force between macroscopic bodies.. 556
Table 18.4 The three main tribological regimes characterizing the changing properties of liquids.. 570
Table 18.5 Effect of molecular shape and short-range forces on tribological properties.. 592

19 **Scanning Probe Studies of Nanoscale Adhesion Between Solids in the Presence of Liquids and Monolayer Films**
Table 19.1 Interactions evaluated by atomic force microscopy adhesion measurements... 620

21 **Nanoscale Mechanical Properties – Measuring Techniques and Applications**
Table 21.1 Summary of the mechanical properties of carbon nanotubes.. 672
Table 21.2 Absolute and relative changes of the indentation modulus with respect to the initial values under wet conditions 680

22 **Nanomechanical Properties of Solid Surfaces and Thin Films**
Table 22.1 Results for some experimental studies of multilayer hardness 710

24 **Mechanics of Biological Nanotechnology**
Table 24.1 Sizes of entities in comparison to the size of *E. coli*................ 748
Table 24.2 Time scales of various events in biological nanotechnology 750

25 **Mechanical Properties of Nanostructures**
Table 25.1 Hardness, elastic modulus, fracture toughness, and critical load results of the bulk single-crystal Si(100) and thin films of undoped polysilicon, SiO_2, SiC, Ni-P, and Au 770

Table 25.2 Summary of measured parameters from quasi-static bending tests 775
Table 25.3 Stresses and displacements for materials that are elastic, elastic-plastic, or elastic-perfectly plastic 784

Part D Molecularly Thick Films for Lubrication

26 Nanotribology of Ultrathin and Hard Amorphous Carbon Films

Table 26.1 Summary of the most commonly used deposition techniques . 796
Table 26.2 Experimental results from EELS and Raman spectroscopy 801
Table 26.3 Experimental results of FRS analysis 802
Table 26.4 Experimental results of physical properties 802
Table 26.5 Hardness, elastic modulus, fracture toughness, fatigue life, critical load during scratch, coefficient of friction of various DLC coatings 809

27 Self-Assembled Monolayers for Controlling Adhesion, Friction and Wear

Table 27.1 Relative electronegativity of selected elements 835
Table 27.2 Names and formulas of selected hydrocarbons 836
Table 27.3 Names and formulas of selected alcohols, ethers, phenols, and thiols 836
Table 27.4 Names and formulas of selected aldehydes and ketones 837
Table 27.5 Names and formulas of selected carboxylic acids and esters ... 838
Table 27.6 Names and formulas of selected organic nitrogen compounds 838
Table 27.7 Some examples of polar and nonpolar groups 839
Table 27.8 Organic groups listed in the increasing order of polarity 839
Table 27.9 Selected substrates and precursors used for formation of SAMs 840
Table 27.10 The roughness, thickness, tilt angles, and spacer chain lengths of SAMs 846
Table 27.11 Melting point of typical organic compounds similar to HDT and BPT SAMs 851
Table 27.12 Calculated and measured relative heights of HDT and MHA self-assembled monolayers 855
Table 27.13 Calculated and measured residual film thickness for SAMs under critical load 855
Table 27.14 Bond strengths of the chemical bonds in SAMs 856

28 Nanoscale Boundary Lubrication Studies

Table 28.1 Typical properties of Z-15 and Z-DOL 863

Part E Industrial Applications and Microdevice Reliability

30 Nanotechnology for Data Storage Applications

Table 30.1 A comparison of the parameters of HDD and ODD 903

Table 30.2 Perceived technical limits for hard disk drive technology........ 904
Table 30.3 A table of probe storage operating parameters 909
Table 30.4 Probe storage projected performance 909

31 The "Millipede" – A Nanotechnology-Based AFM Data-Storage System

Table 31.1 Areal density and storage capacity 947

32 Microactuators for Dual-Stage Servo Systems in Magnetic Disk Files

Table 32.1 Parameters of the electrostatic microactuator 960

33 Micro/Nanotribology of MEMS/NEMS Materials and Devices

Table 33.1 Dimensions and masses in perspective 985
Table 33.2 Selected bulk properties of SiC and Si(100)............................. 996
Table 33.3 Surface roughness and micro- and macroscale coefficients of friction of selected samples.. 997
Table 33.4 RMS, microfriction, microscratching/microwear, and nanoindentation hardness data for various virgin, coated, and treated silicon samples 999
Table 33.5 Summary of micro/nanotribological properties of the sample materials 1001
Table 33.6 Surface roughness parameters and microscale coefficient of friction 1009
Table 33.7 Coefficient of static friction measurements of MEMS devices and structures 1015

35 Thermo- and Electromechanics of Thin-Film Microstructures

Table 35.1 Stresses, curvature, and strains in a film/substrate system...... 1045
Table 35.2 Predicted and measured curvature per unit negative temperature change for gold/polysilicon microstructures as a function of the polysilicon thickness. 1053

37 MEMS Packaging and Thermal Issues in Reliability

Table 37.1 Summary of bonding mechanisms.. 1116
Table 37.2 The Maximum Likelihood Estimation for Mean Time To Failure (MTTF) 1127

List of Abbreviations

µCP	microcontact printing
µTAS	micro-total analysis systems
2-DEG	two-dimensional electron gas

A

A	adenine
ABS	air-bearing surface
ADEPTS	antibody directed enzyme-prodrug therapy
AFAM	atomic force acoustic microscopy
AFM	atomic force microscope/microscopy
AIDCN	2-amino-4,5-imidazoledicarbonitrile
AM	amplitude modulation
APCVD	atmospheric pressure chemical vapor deposition
ASA	anti-stiction agent
ATP	adenosine triphosphat

B

BE	boundary element
BioMEMS	biological or biomedical microelectromechanical systems
BP	bit pitch
BPI	bits per inch
bpsi	bits per square inch
BSA	bovine serum albumin

C

C	cytosine
CA	constant amplitude
CBA	cantilever beam array
CCVD	catalytic chemical vapor deposition
CDS	correlated double sampling
CDW	charge density wave
CE	capillary electrophoresis
CE	constant excitation mode
CFM	chemical force microscopy
CG	controlled geometry
CNT	carbon nanotube
COC	cyclic olefin copolymers
COF	chip-on-flex
COGs	cost of goods
CSM	continuous stiffness measurement
CTE	coefficient of thermal expansion
CTLs	cytoxic T-lymphocytes
CVD	chemical vapor deposition

D

DAS	dimer adatom stacking
DBR	distributed Bragg reflector
DCs	dendritic cells
DFB	distributed feedback
DFM	dynamic force microscopy
DFT	density functional theory
DLC	diamond-like carbon
DLP	digital light processing
DLVO	Derjaguin–Landau–Verwey–Overbeek
DMD	digital micromirror device
DMT	Derjaguin–Muller–Toporov
DOS	density of states
DPN	dip-pen nanolithography
DRIE	deep reactive ion etching
DSC	differential scanning calorimetry
DSP	digital signal processor
DT	diphteria toxin
DWNTs	double-wall nanotubes

E

EAM	embedded atom method
EBD	electron beam deposition
ECR-CVD	electron cyclotron resonance chemical vapor deposition
EDC	1-ethyl-3-(3-diamethylaminoprophyl) carbodiimide
EDP	ethylene diamine pyrocatechol
EDS	energy dispersive X-ray spectrometer
EELS	electron energy loss spectrometer/spectroscopy
EFC	electrostatic force constant
EFM	electric field gradient microscopy
EHD	electrohydrodynamic
EL	electro-luminiscence
EO	electro-osmosis
EOF	electro-osmotic flow
EPR	enhanced permeability and retention effect
ESD	electrostatic discharge

F

FAD	flavin adenine dinucleotide
FC	flip chip technique
FCA	filtered cathodic arc
FCP	force calibration plot
FD	finite difference
FE	finite element
FEM	finite element method/modeling

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

FET	field-effect transistor
FFM	friction force microscope/microscopy
FIB	focused ion beam
FID	free induction decay
FIM	field-ion microscope/microscopy
FKT	Frenkel-Kontorova-Tomlinson
FM	frequency modulation
FM-AFM	frequency modulation AFM
FMEA	failure mode effect analysis
FMM	force modulation mode
FM-SFM	frequency-modulation SFM
FS	force spectroscopy

G

G	guanine
GIO	grazing impact oscillator
GMR	giant magnetoresistance
GOD	glucose oxidase
Gox	flavoenzyme glucose oxidase

H

HARMEMS	high-aspect-ratio MEMS
HDD	hard disk drive
HF	hydrofluoric acid
HOP	highly oriented pyrolytic
HOPG	highly oriented pyrolytic graphite
HPMA	hydroxyl polymethacrylamide
HtBDC	hexa-tert-butyl-decacyclene
HTCS	high temperature superconductivity

I

IBD	ion beam deposition
IC	integrated circuit
ICAM-1	intercellular adhesion molecule-1
IFM	interfacial force microscope
ISE	indentation size effect
ITO	indium tin oxide

J

JKR	Johnson–Kendall–Roberts

K

KPFM	Kelvin probe force microscopy

L

LB	Langmuir–Blodgett
LCC	leadless chip carrier
LDOS	local density of states
LEDs	light-emitting diodes
LFA-1	leukocyte function-associated antigen-1
LFM	lateral force microscope
LFT	linear fractional transformation
LN	liquid nitrogen
LPCVD	low pressure chemical vapor deposition
LQG	linear quadratic Gaussian
LTR	loop-transfer recovery
LTSPM	low-temperature SPM
LTSTS	low-temperature scanning tunneling spectroscopy
LVDT	linear variable differential transformers

M

MAP	manifold absolute pressure
MD	molecular dynamics
MDS	molecular dynamics simulation
ME	metal evaporated
MEMS	microelectromechanical systems
MFM	magnetic field microscope/microscopy
MHA	16-mercaptohexadecanoic acid thiol
MHC	major histocompatibility complex
MHD	magnetohydrodynamic
MIM	metal/insulator/metal
MLE	maximum likelihood estimator
MOCVD	metalorganic CVD
MP	metal particle
MPTMS	mercaptopropyltrimethoxysilane
MRAM	magnetoresistive RAM
MRFM	magnetic resonance force microscopy
MRFM	molecular recognition force microscopy
MRI	magnetic resonance imaging
MTTF	mean time to failure
MWCNT	multiwall carbon nanotube

N

NA	nucleic acid
NBMN	3-nitrobenzal malonitrile
NC-AFM	noncontact atomic force microscopy
NCS	neocarzinostatin
NEMS	nanoelectromechanical systems
NMP	no moving part
NNI	National Nanotechnology Initiative
NSOM	near-field scanning optical microscope/microscopy
NTA	nitrilotriacetate
nTP	nanotransfer printing
NVRAM	nonvolatile random access memories

O

ODD	optical disk drives
OMVPE	organometallic vapor phase epitaxy
OT	optical tweezers
OTE	octadecyltrimethoxysilane
OTS	octadecyltrichlorosilane

OUM	Ovonyx unified memory

P

P/W	power to weight
PA	plasminogen activator
PAMAM	poly(amido) amine
PAPP	p-aminophenyl phosphate
PC	ploycarbonate
PC-RAM	phase change RAM
pDA	1,4-phenylenediamine
PDMS	polydimethylsiloxane
PDP	2-pyridyldithiopropionyl
PE	polyethylene
PECVD	plasma enhanced CVD
PEG	poly(ethylene glycol)
PES	photoemission spectroscopy
PES	position error signal
PET	poly(ethylene terephthalate)
PFDA	perfluorodecanoic acid
PFPE	perfluoropolyether
PL	photoluminescence
PMMA	poly(methylmethacrylate)
PS	polystyrene
PSG	phosphorus-doped glass
PSGL-1	P-selection glycoprotein ligand-1
PTFE	polytetrafluoroethylene
PZT	lead zirconate titanate

Q

QCM	quartz-crystal microbalance

R

RES	reticuloendothelial system
RF	radiofrequency
RH	relative humidity
RICM	reflection interface contrast microscopy
RIE	reactive ion etching
RLS	recursive least square algorithm
RPES	relative position error signal
RTP	rapid thermal processing

S

SACA	static advancing contact angle
SAED	selected area electron diffraction
SAM	self-assembling monolayer
SAM	scanning acoustic microscopy
SCM	scanning capitance microscopy
SCPM	scanning chemical potential
SEcM	scanning electrochemical microscopy
SEFM	scanning electrostatic force microscopy
SEM	scanning electron microscope/microscopy
SFA	surface forces apparatus
SFAM	scanning force acoustic microscopy
SFD	shear flow detachment
SFM	scanning force microscopy
SFS	scanning force spectroscopy
SICM	scanning ion conductance microscopy
SIMO	single-input–multi-output
SISO	single-input–single-output
SKPM	scanning Kelvin probe microscopy
SLAM	scanning local-acceleration microscopy
SMA	shape memory alloy
SMANCS	S-Methacryl-neocarzinostatin
SMM	scanning magnetic microscopy
SN	scanning nanoindentation
SNOM	scanning near-field optical microscopy
SPM	scanning probe microscopy
sPROMS	structurally programmable microfluidic system
SPS	spark plasma sintering
SRAM	static random access memory
ssDNA	single stranded DNA
SSNA	single-stranded nucleic acid molecule
SThM	scanning thermal microscopy
STM	scanning tunneling microscope/microscopy
SWCNT	single-wall carbon nanotubes
SWNT	single-wall nanotubes

T

T	thymine
TAAs	tumor associated antigens
TEM	transmission electron microscopy
TESP	tapping-mode etched silicon probe
TGA	thermo-gravimetric analysis
TIRM	total internal reflection microscopy
TMAH	tetramethyl ammonium hydroxide
TMAH	tetramethyl-aluminium hydroxide
TP	track pitch
TPI	tracks per inch
TRM/TMR	track mis-registration
T-SLAM	variable-temperature SLAM
TTF	tetrathiofulvane

U

UHV	ultrahigh vacuum

V

VCM	voice coil motor
VCO	voltage-controlled oscillator
VLS	vapor-liquid-solid
VLSI	very large-scale integration

X

XRD	X-ray diffraction

1. Introduction to Nanotechnology

> *A biological system can be exceedingly small. Many of the cells are very tiny, but they are very active; they manufacture various substances; they walk around; they wiggle; and they do all kinds of marvelous things – all on a very small scale. Also, they store information. Consider the possibility that we too can make a thing very small which does what we want – that we can manufacture an object that maneuvers at that level.*
>
> (From the talk "There's Plenty of Room at the Bottom", delivered by Richard P. Feynman at the annual meeting of the American Physical Society at the California Institute of Technology, Pasadena, CA, on December 29, 1959.)

by B. Bhushan

1.1 **Background and Definition of Nanotechnology** 1
1.2 **Why Nano?** 2
1.3 **Lessons from Nature** 2
1.4 **Applications in Different Fields** 3
1.5 **Reliability Issues of MEMS/NEMS** 4
1.6 **Organization of the Handbook**............... 5
References 5

1.1 Background and Definition of Nanotechnology

On Dec. 29, 1959, at the California Institute of Technology, Nobel Laureate Richard P. Feynman gave a talk at the annual meeting of the American Physical Society that has become one of the twentieth century's classic science lectures, titled "There's Plenty of Room at the Bottom" [1.1]. He presented a technological vision of extreme miniaturization several years before the word "chip" became part of the lexicon. He talked about the problem of manipulating and controlling things on a small scale. Extrapolating from known physical laws, Feynman envisioned a technology using the ultimate toolbox of nature, building nanoobjects atom by atom or molecule by molecule. Since the 1980s, many inventions and discoveries in the fabrication of nanoobjects have become a testament to his vision. In recognition of this reality, the National Science and Technology Council (NSTC) of the White House created the Interagency Working Group on Nanoscience, Engineering and Technology (IWGN) in 1998. In a January 2000 speech at the same institute, former President William J. Clinton talked about the exciting promise of nanotechnology and, more generally, the importance of expanding research in nanoscale science and technology. Later that month, he announced in his State of the Union Address an ambitious $ 497 million federal, multi-agency National Nanotechnology Initiative (NNI) in the fiscal year 2001 budget, and made it a top science and technology priority [1.2, 3]. The objective of this initiative was to form a broad-based coalition in which academe, the private sector, and local, state, and federal governments would work together to push the envelope of nanoscience and nanoengineering to reap nanotechnology's potential social and economic benefits.

Nanotechnology literally means any technology performed on a nanoscale that has applications in the real world. Nanotechnology encompasses the production and application of physical, chemical, and biological systems at scales ranging from individual atoms or molecules to submicron dimensions, as well as the integration of the resulting nanostructures into larger systems. Nanotechnology is likely to have a profound impact on our economy and society in the early twenty-first century, comparable to that of semiconductor technology, information technology, or cellular and molecular biology. Science and technology research in nanotechnology promises breakthroughs in such areas as materials and manufacturing, nanoelectronics, medicine and healthcare, energy, biotechnology, information technology, and national security. It is widely felt that nanotechnology will be the next industrial revolution.

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Nanometer-scale features are mainly built up from their elemental constituents. Chemical synthesis – the spontaneous self-assembly of molecular clusters (molecular self-assembly) from simple reagents in solution – or biological molecules (e.g., DNA) are used as building blocks for the production of three-dimensional nanostructures, including quantum dots (nanocrystals) of arbitrary diameter (about 10 to 10^5 atoms). A variety of vacuum deposition and nonequilibrium plasma chemistry techniques are used to produce layered nanocomposites and nanotubes. Atomically controlled structures are produced using molecular beam epitaxy and organo-metallic vapor phase epitaxy. Micro- and nanosystem components are fabricated using top-down lithographic and nonlithographic fabrication techniques and range in size from micro- to nanometers. Continued improvements in lithography for use in the production of nanocomponents have resulted in line widths as small as 10 nanometers in experimental prototypes. The nanotechnology field, in addition to the fabrication of nanosystems, provides the impetus to development of experimental and computational tools.

The micro- and nanosystems include micro/nanoelectromechanical systems (MEMS/NEMS) (e.g., sensors, actuators, and miniaturized systems comprising sensing, processing, and/or actuating functions), micromechatronics, optoelectronics, microfluidics, and systems integration. These systems can sense, control, and activate on the micro/nanoscale and function individually or in arrays to generate effects on the macroscale. The microsystems market in 2000 was about $ 15 billion, and, with a projected 10–20 % annual growth rate, it is expected to increase to more than $ 100 billion by the end of this decade. The nanosystems market in 2001 was about $ 100 million and the integrated nanosystems market is expected to be more than $ 25 billion by the end of this decade. Due to the enabling nature of these systems, and because of the significant impact they can have on the commercial and defense applications, venture capitalists, industry, as well as the federal government have taken a special interest in nurturing growth in this field. Micro- and nanosystems are likely to be the next logical step in the "silicon revolution."

1.2 Why Nano?

The discovery of novel materials, processes, and phenomena at the nanoscale, as well as the development of new experimental and theoretical techniques for research provide fresh opportunities for the development of innovative nanosystems and nanostructured materials. Nanosystems are expected to find various unique applications. Nanostructured materials can be made with unique nanostructures and properties. This field is expected to open new venues in science and technology.

1.3 Lessons from Nature

Nanotechnology is a new word, but it is not an entirely new field. Nature has many objects and processes that function on a micro- to nanoscale [1.2, 4]. The understanding of these functions can guide us in imitating and producing nanodevices and nanomaterials.

Billions of years ago, molecules began organizing themselves into the complex structures that could support life. Photosynthesis harnesses solar energy to support plant life. Molecular ensembles are present in plants, which include light harvesting molecules, such as chlorophyll, arranged within the cells on the nanometer to micrometer scales. These structures capture light energy, and convert it into the chemical energy that drives the biochemical machinery of plant cells. Live organs use chemical energy in the body. The flagella, a type of bacteria, rotates at over 10,000 RPM [1.5]. This is an example of a biological molecular machine. The flagella motor is driven by the proton flow caused by the electrochemical potential differences across the membrane. The diameter of the bearing is about 20–30 nm, with an estimated clearance of about 1 nm.

In the context of tribology, some biological systems have anti-adhesion surfaces. First, many plant leaves (such as lotus leaf) are covered by a hydrophobic cuticle, which is composed of a mixture of large hydrocarbon molecules that have a strong hydrophobia. Second, the surface is made of a unique roughness distribution [1.6, 7]. It has been reported that for some leaf surfaces, the roughness of the hydrophobic leaf surface decreases wetness, which is reflected in a greater contact angle of water droplets on such surfaces.

1.4 Applications in Different Fields

Science and technology continue to move forward in making the fabrication of micro/nanodevices and systems possible for a variety of industrial, consumer, and biomedical applications. A range of MEMS devices have been produced, some of which are commercially used [1.4, 8–12]. A variety of sensors are used in industrial, consumer, and biomedical applications. Various microstructures or microcomponents are used in micro-instruments and other industrial applications, such as micromirror arrays. Two of the largest "killer" industrial applications are accelerometers (about 85 million units in 2002) and digital micromirror devices (about $ 400 million in sales in 2001). Integrated capacitive-type, silicon accelerometers have been used in airbag deployment in automobiles since 1991 [1.13, 14]. Accelerometer technology was about a billion-dollar-a-year industry in 2001, dominated by Analog Devices followed by Motorola and Bosch. Commercial digital light processing (DLP) equipment using digital micromirror devices (DMD) were launched in 1996 by Texas Instruments for digital projection displays in portable and home theater projectors, as well as table-top and projection TVs [1.15, 16]. More than 1.5 million projectors were sold before 2002. Other major industrial applications include pressure sensors, inkjet printer heads, and optical switches. Silicon-based piezoresistive pressure sensors for manifold absolute pressure sensing for engines were launched in 1991 by NovaSensor, and their annual sales were about 25 million units in 2002. Annual sales of inkjet printer heads with microscale functional components were about 400 million units in 2002. Capacitive pressure sensors for tire pressure measurements were launched by Motorola. Other applications of MEMS devices include chemical sensors; gas sensors; infrared detectors and focal plane arrays for earth observations; space science and missile defense applications; pico-satellites for space applications; and many hydraulic, pneumatic, and other consumer products. MEMS devices are also being pursued in magnetic storage systems [1.17], where they are being developed for super-compact and ultrahigh recording-density magnetic disk drives. Several integrated head/suspension microdevices have been fabricated for contact recording applications [1.18, 19]. High-bandwidth, servo-controlled microactuators have been fabricated for ultrahigh track-density applications, which serve as the fine-position control element of a two-stage, coarse/fine servo system, coupled with a conventional actuator [1.20–23]. Millimeter-sized wobble motors and actuators for tip-based recording schemes have also been fabricated [1.24].

BIOMEMS are increasingly used in commercial and defense applications (e.g., [1.4,25–28]). Applications of BIOMEMS include biofluidic chips (otherwise known as microfluidic chips, bioflips, or simply biochips) for chemical and biochemical analyses (biosensors) in medical diagnostics (e.g., DNA, RNA, proteins, cells, blood pressure and assays, and toxin identification) and implantable pharmaceutical drug delivery. The biosensors, also referred to as lab-on-a-chip, integrate sample handling, separation, detection, and data analysis onto one platform. Biosensors are designed to either detect a single or class of (bio)chemicals or system-level analytical capabilities for a broad range of (bio)chemical species known as micro total analysis systems (μTAS). The chips rely on microfluidics and involve the manipulation of tiny amounts of fluids in microchannels using microvalves for various analyses. The test fluid is pumped into the chip generally using an external pump for analyses. Some chips have been designed with an integrated electrostatically actuated diaphragm-type micropump. Silicon-based, disposable blood-pressure sensor chips were introduced in the early 1990s by NovaSensor for blood pressure monitoring (about 20 million units in 2002). A variety of biosensors are manufactured by various companies, including ACLARA, Agilent Technologies, Calipertech, and I-STAT.

After the tragedy of Sept. 11, 2001, concern over biological and chemical warfare has led to the development of handheld units with bio- and chemical sensors for the detection of biological germs, chemical or nerve agents, mustard agents, and chemical precursors to protect subways, airports, the water supply, and the population [1.29].

Other BIOMEMS applications include minimal invasive surgery, such as endoscopic surgery, laser angioplasty, and microscopic surgery. Implantable artificial organs can also be produced.

Micro-instruments and micro-manipulators are used to move, position, probe, pattern, and characterize nanoscale objects and nanoscale features. Miniaturized analytical equipment includes gas chromatography and mass spectrometry. Other instruments include micro-STM, where STM stands for scanning tunneling microscope.

Examples of NEMS include nanocomponents, nanodevices, nanosystems, and nanomaterials, such as microcantilever with integrated sharp nanotips for STM and atomic force microscopy (AFM), AFM array (millipede) for data storage, AFM tips for nanolithography, dip-pen nanolithography for printing molecules, biological (DNA) motors, molecular gears, molecularly thick films (e.g., in giant magneto-resistive or GMR heads and magnetic media), nanoparticles, (e.g., nanomagnetic particles in magnetic media), nanowires, carbon nanotubes, quantum wires (QWRs), quantum boxes (QBs), and quantum transistors [1.30–34]. BIONEMS include nanobiosensors – a microarray of silicon nanowires, roughly a few nm in size, to selectively bind and detect even a single biological molecule, such as DNA or protein, by using nanoelectronics to detect the slight electrical charge caused by such binding, or a microarray of carbon nanotubes to electrically detect glucose, implantable drug-delivery devices – e.g., micro/nanoparticles with drug molecules encapsulated in functionized shells for a site-specific targeting application, and a silicon capsule with a nanoporous membrane filled with drugs for long term delivery, nanodevices for sequencing single molecules of DNA in the Human Genome Project, cellular growth using carbon nanotubes for spinal cord repair, nanotubes for nanostructured materials for various applications, such as spinal fusion devices, organ growth, and growth of artificial tissues using nanofibers.

Nanoelectronics can be used to build computer memory, using individual molecules or nanotubes to store bits of information, as well as molecular switches, molecular or nanotube transistors, nanotube flat-panel displays, nanotube integrated circuits, fast logic gates, switches, nanoscopic lasers, and nanotubes as electrodes in fuel cells.

1.5 Reliability Issues of MEMS/NEMS

There is an increasing need for a multidisciplinary, system-oriented approach to manufacturing micro/nanodevices that function reliably. This can only be achieved through the cross-fertilization of ideas from different disciplines and the systematic flow of information and people among research groups. Common potential failure mechanisms for MEMS/NEMS that need to be addressed in order to increase reliability are: adhesion, friction, wear, fracture, fatigue, and contamination. Surface micro/nanomachined structures often include smooth and chemically active surfaces. Due to the large surface area to volume ratio in MEMS/NEMS, they are particularly prone to stiction (high static friction) as part of normal operation. Fracture occurs when the load on a microdevice is greater than the strength of the material. Fracture is a serious reliability concern, particularly for the brittle materials used in the construction of these components, since it can immediately, or eventually, lead to catastrophic failures. Additionally, debris can be formed from the fracturing of microstructures, leading to other failure processes. For less brittle materials, repeated loading over a long period causes fatigue that would also lead to the breaking and fracturing of the device. In principle, this failure mode is relatively easy to observe and simple to predict. However, the materials properties of thin films are often not known, making fatigue predictions prone to error.

Many MEMS/NEMS devices operate near their thermal dissipation limit. They may encounter hot spots that can cause failures, particularly in weak structures such as diaphragms or cantilevers. Thermal stressing and relaxation caused by thermal variations can create material delamination and fatigue in cantilevers. In large temperature changes, as experienced in the space environment, bimetallic beams will also experience warping due to mismatched coefficients of thermal expansion. Packaging has been a big problem. The contamination, which probably happens in packaging and during storage, can also strongly influence the reliability of MEMS/NEMS. For example, a particulate dust landed on one of the electrodes of a comb drive can cause catastrophic failure. There are no MEMS/NEMS fabrication standards, which makes it difficult to transfer fabrication steps in MEMS/NEMS between foundaries.

Obviously, studies of determination and suppression of active failure mechanisms affecting this new and promising technology are critical to the high reliability of MEMS/NEMS and are determining factors in successful practical application.

Mechanical properties are known to exhibit a dependence on specimen size. Mechanical property evaluation of nanometer-scaled structures is carried out to help design reliable systems, since good mechanical properties are of critical importance in such applications. Some of the properties of interest are: Young's modulus of elasticity, hardness, bending strength, fracture toughness, and fatigue life. Finite element modeling is carried out to study the effects of surface roughness and scratches on stresses in nanostructures. When nanostructures are smaller than a fundamental physical length scale, conventional theory may no longer apply, and new phenomena may emerge. Molecular mechanics is used to simulate the behavior of a nano-object.

1.6 Organization of the Handbook

The handbook integrates knowledge from the fabrication, mechanics, materials science, and reliability points of view. Organization of the book is straightforward. The handbook is divided into six parts. This first part introduces the nanotechnology field, including an introduction to nanostructures, micro/nanofabrication and, micro/nanodevices. The second part introduces scanning probe microscopy. The third part provides an overview of nanotribology and nanomechanics, which will prepare the reader to understand the tribology and mechanics of industrial applications. The fourth part provides an overview of molecularly thick films for lubrication. The fifth part focuses on industrial applications and microdevice reliability. And the last part focuses on the social and ethical implications of nanotechnology.

References

1.1 R. P. Feynmann: There's plenty of room at the bottom, Eng. Sci. **23** (1960) 22–36, and www.zyvex.com/nanotech/feynman.html (1959)
1.2 I. Amato: Nanotechnology, www.ostp.gov/nstc/html/iwgn/iwgn.public.brochure/welcome.htm or www.nsf.gov/home/crssprgm/nano/nsfnnireports.htm (2000)
1.3 Anonymous: National nanotechnology initiative, www.ostp.gov/nstc/html/iwgn.fy01budsuppl/nni.pdf or www.nsf.gov/home/crssprgm/nano/nsfnnireports.htm (2000)
1.4 I. Fujimasa: *Micromachines: A New Era in Mechanical Engineering* (Oxford Univ. Press, Oxford 1996)
1.5 C. J. Jones, S. Aizawa: The bacterial flagellum and flagellar motor: Structure, assembly, and functions, Adv. Microb. Physiol. **32** (1991) 109–172
1.6 V. Bergeron, D. Quere: Water droplets make an impact, Phys. World **14** (May 2001) 27–31
1.7 M. Scherge, S. Gorb: *Biological Micro- and Nanotribology* (Springer, Berlin, Heidelberg 2000)
1.8 B. Bhushan: *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)
1.9 G. T. A. Kovacs: *Micromachined Transducers Sourcebook* (WCB McGraw-Hill, Boston 1998)
1.10 S. D. Senturia: *Microsystem Design* (Kluwer, Boston 2001)
1.11 T. R. Hsu: *MEMS and Microsystems* (McGraw-Hill, Boston 2002)
1.12 M. Madou: *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd edn. (CRC, Boca Raton 2002)
1.13 T. A. Core, W. K. Tsang, S. J. Sherman: Fabrication technology for an integrated surface-micromachined sensor, Solid State Technol. **36** (October 1993) 39–47
1.14 J. Bryzek, K. Peterson, W. McCulley: Micromachines on the march, IEEE Spectrum (May 1994) 20–31
1.15 L. J. Hornbeck, W. E. Nelson: Bistable deformable mirror device, OSA Technical Digest **8** (1988) 107–110
1.16 L. J. Hornbeck: A digital light processing(tm) update – Status and future applications (invited), Proc. Soc. Photo-Opt. Eng. **3634** (1999) 158–170
1.17 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)
1.18 H. Hamilton: Contact recording on perpendicular rigid media, J. Mag. Soc. Jpn. **15** (Suppl. S2) (1991) 481–483
1.19 T. Ohwe, Y. Mizoshita, S. Yonoeka: Development of integrated suspension system for a nanoslider with an MR head transducer, IEEE Trans. Magn. **29** (1993) 3924–3926
1.20 D. K. Miu, Y. C. Tai: Silicon micromachined scaled technology, IEEE Trans. Industr. Electron. **42** (1995) 234–239
1.21 L. S. Fan, H. H. Ottesen, T. C. Reiley, R. W. Wood: Magnetic recording head positioning at very high track densities using a microactuator-based, two-stage servo system, IEEE Trans. Ind. Electron. **42** (1995) 222–233
1.22 D. A. Horsley, M. B. Cohn, A. Singh, R. Horowitz, A. P. Pisano: Design and fabrication of an angular

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

microactuator for magnetic disk drives, J. Microelectromech. Syst. **7** (1998) 141–148

1.23 T. Hirano, L.S. Fan, D. Kercher, S. Pattanaik, T. S. Pan: HDD tracking microactuator and its integration issues, Proc. ASME Int. Mech. Eng. Congress, MEMS, New York 2000, ed. by A. P. Lee, J. Simon, F. K. Foster, R. S. Keynton (ASME, New York 2000) 449–452

1.24 L. S. Fan, S. Woodman: Batch fabrication of mechanical platforms for high-density data storage, 8th Int. Conf. Solid State Sensors and Actuators (Transducers '95)/Eurosensors IX, Stockholm (June, 1995) 434–437

1.25 P. Gravesen, J. Branebjerg, O. S. Jensen: Microfluidics – A review, J. Micromech. Microeng. **3** (1993) 168–182

1.26 C. Lai Poh San, E. P. H. Yap (Eds.): *Frontiers in Human Genetics* (World Scientific, Singapore 2001)

1.27 C. H. Mastrangelo, H. Becker (Eds.): *Microfluidics and BioMEMS*, Proc. SPIE **4560** (SPIE, Bellingham 2001)

1.28 H. Becker, L. E. Locascio: Polymer microfluidic devices, Talanta **56** (2002) 267–287

1.29 M. Scott: MEMS and MOEMS for national security applications, , Reliability, Testing, and Characterization of MEMS/MOEMS II, Proc. SPIE **4980** (SPIE, Bellingham 2003)

1.30 K. E. Drexler: *Nanosystems: Molecular Machinery, Manufacturing and Computation* (Wiley, New York 1992)

1.31 G. Timp (Ed.): *Nanotechnology* (Springer, Berlin, Heidelberg 1999)

1.32 E. A. Rietman: *Molecular Engineering of Nanosystems* (Springer, Berlin, Heidelberg 2001)

1.33 H. S. Nalwa (Ed.): *Nanostructured Materials and Nanotechnology* (Academic, San Diego 2002)

1.34 W. A. Goddard, D. W. Brenner, S. E. Lyshevski, G. J. Iafrate: *Handbook of Nanoscience, Engineering, and Technology* (CRC, Boca Raton 2003)

Part A

Nanostruc

Part A Nanostructures, Micro/Nanofabrication, and Micro/Nanodevices

2 Nanomaterials Synthesis and Applications: Molecule-Based Devices
Françisco M. Raymo, Coral Gables, USA

3 Introduction to Carbon Nanotubes
Marc Monthioux, Toulouse, France
Philippe Serp, Toulouse, France
Emmanuel Flahaut, Toulouse, France
Manitra Razafinimanana, Toulouse, France
Christophe Laurent, Toulouse, France
Alain Peigney, Toulouse, France
Wolfgang Bacsa, Toulouse, France
Jean-Marc Broto, Toulouse, France

4 Nanowires
Mildred S. Dresselhaus, Cambridge, USA
Yu-Ming Lin, Cambridge, USA
Oded Rabin, Cambridge, USA
Marcie R. Black, Cambridge, USA
Gene Dresselhaus, Cambridge, USA

5 Introduction to Micro/Nanofabrication
Babak Ziaie, Minneapolis, USA
Antonio Baldi, Barcelona, Spain
Massood Z. Atashbar, Kalamazoo, USA

6 Stamping Techniques for Micro and Nanofabrication: Methods and Applications
John A. Rogers, Urbana, USA

7 Materials Aspects of Micro- and Nanoelectromechanical Systems
Christian A. Zorman, Cleveland, USA
Mehran Mehregany, Cleveland, USA

8 MEMS/NEMS Devices and Applications
Darrin J. Young, Cleveland, USA
Christian A. Zorman, Cleveland, USA
Mehran Mehregany, Cleveland, USA

9 Microfluidics and Their Applications to Lab-on-a-Chip
Chong H. Ahn, Cincinnati, USA
Jin-Woo Choi, Baton Rouge, USA

10 Therapeutic Nanodevices
Stephen C. Lee, Columbus, USA
Mark Ruegsegger, Columbus, USA
Philip D. Barnes, Columbus, USA
Bryan R. Smith, Columbus, USA
Mauro Ferrari, Columbus, USA

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

2. Nanomaterials Synthesis and Applications: Molecule-Based Devices

The constituent components of conventional devices are carved out of larger materials relying on physical methods. This top-down approach to engineered building blocks becomes increasingly challenging as the dimensions of the target structures approach the nanoscale. Nature, on the other hand, relies on chemical strategies to assemble nanoscaled biomolecules. Small molecular building blocks are joined to produce nanostructures with defined geometries and specific functions. It is becoming apparent that nature's bottom-up approach to functional nanostructures can be mimicked to produce artificial molecules with nanoscaled dimensions and engineered properties. Indeed, examples of artificial nanohelices, nanotubes, and molecular motors are starting to be developed. Some of these fascinating chemical systems have intriguing electrochemical and photochemical properties that can be exploited to manipulate chemical, electrical, and optical signals at the molecular level. This tremendous opportunity has lead to the development of the molecular equivalent of conventional logic gates. Simple logic operations, for example, can be reproduced with collections of molecules operating in solution. Most of these chemical systems, however, rely on bulk addressing to execute combinational and sequential logic operations. It is essential to devise methods to reproduce these useful functions in solid-state configurations and, eventually, with single molecules. These challenging objectives are stimulating the design of clever devices that interface small assemblies of organic molecules with macroscaled and nanoscaled electrodes. These strategies have already produced rudimentary examples of diodes, switches, and transistors based on functional molecular components. The rapid and continuous progress of this exploratory research will, we hope, lead to an entire generation of molecule-based devices that might ultimately find useful applications in a variety of fields, ranging from biomedical research to information technology.

2.1 **Chemical Approaches to Nanostructured Materials** 10
  2.1.1 From Molecular Building Blocks to Nanostructures......................... 10
  2.1.2 Nanoscaled Biomolecules: Nucleic Acids and Proteins............ 10
  2.1.3 Chemical Synthesis of Artificial Nanostructures 12
  2.1.4 From Structural Control to Designed Properties and Functions............................. 12

2.2 **Molecular Switches and Logic Gates** 14
  2.2.1 From Macroscopic to Molecular Switches 15
  2.2.2 Digital Processing and Molecular Logic Gates 15
  2.2.3 Molecular AND, NOT, and OR Gates.. 16
  2.2.4 Combinational Logic at the Molecular Level 17
  2.2.5 Intermolecular Communication...... 18

2.3 **Solid State Devices**............................... 22
  2.3.1 From Functional Solutions to Electroactive and Photoactive Solids................. 22
  2.3.2 Langmuir–Blodgett Films 23
  2.3.3 Self-Assembled Monolayers........... 27
  2.3.4 Nanogaps and Nanowires............. 31

2.4 **Conclusions and Outlook** 35

References 36

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

2.1 Chemical Approaches to Nanostructured Materials

The fabrication of conventional devices relies on the assembly of macroscopic building blocks with specific configurations. The shapes of these components are carved out of larger materials by exploiting physical methods. This top-down approach to engineered building blocks is extremely powerful and can deliver effectively and reproducibly microscaled objects. This strategy becomes increasingly challenging, however, as the dimensions of the target structures approach the nanoscale. Indeed, the physical fabrication of nanosized features with subnanometer precision is a formidable technological challenge.

2.1.1 From Molecular Building Blocks to Nanostructures

Nature efficiently builds nanostructures by relying on chemical approaches. Tiny molecular building blocks are assembled with a remarkable degree of structural control in a variety of nanoscaled materials with defined shapes, properties, and functions. In contrast to the top-down physical methods, small components are connected to produce larger objects in these bottom-up chemical strategies. It is becoming apparent that the limitations of the top-down approach to artificial nanostructures can be overcome by mimicking nature's bottom-up processes. Indeed, we are starting to see emerge beautiful and ingenious examples of molecule-based strategies to fabricate chemically nanoscaled building blocks for functional materials and innovative devices.

2.1.2 Nanoscaled Biomolecules: Nucleic Acids and Proteins

Nanoscaled macromolecules play a fundamental role in biological processes [2.1]. Nucleic acids, for example, ensure the transmission and expression of genetic information. These particular biomolecules are linear polymers incorporating nucleotide repeating units (Fig. 2.1a). Each nucleotide has a phosphate bridge and a sugar residue. Chemical bonds between the phosphate of one nucleotide and the sugar of the next ensures the propagation of a polynucleotide strand from the $5'$ to the $3'$ end. Along the sequence of alternating sugar and phosphate fragments, an extended chain of robust covalent bonds involving carbon, oxygen, and phosphorous atoms forms the main backbone of the polymeric strand.

Every single nucleotide of a polynucleotide strand carries one of the four heterocyclic bases shown in Fig. 2.1b. For a strand incorporating 100 nucleotide repeating units, a total of 4^{100} unique polynucleotide sequences are possible. It follows that nature can fabricate a huge number of closely related nanostructures relying only on four building blocks. The heterocyclic bases appended to the main backbone of alternating phosphate and sugar units can sustain hydrogen bonding and $[\pi \cdots \pi]$ stacking interactions. Hydrogen bonds, formed between [N−H] donors and either N or O acceptors, encourage the pairing of adenine (A) with thymine (T) and of guanine (G) with cytosine (C). The stacking interactions involve attractive contacts between the extended π-surfaces of heterocyclic bases.

In the *B* conformation of deoxyribonucleic acid (DNA), the synergism of hydrogen bonds and $[\pi \cdots \pi]$ stacking glues pairs of complementary polynucleotide strands in fascinating double helical supermolecules (Fig. 2.1c) with precise structural control at the subnanometer level. The two polynucleotide strands wrap around a common axis to form a right-handed double helix with a diameter of ca. 2 nm. The hydrogen bonded and $[\pi \cdots \pi]$ stacked base pairs lie at the core of the helix with their π-planes perpendicular to the main axis of the helix. The alternating phosphate and sugar units define the outer surface of the double helix. In *B*-DNA, approximately ten base pairs define each helical turn corresponding to a rise per turn or helical pitch of ca. 3 nm. Considering that these molecules can incorporate up to approximately 10^{11} base pairs, extended end-to-end lengths spanning from only few nanometers to hundreds of meters are possible.

Nature's operating principles to fabricate nanostructures are not limited to nucleic acids. Proteins are also built joining simple molecular building blocks, the amino acids, by strong covalent bonds [2.1]. More precisely, nature relies on 20 amino acids differing in their side chains to assemble linear polymers, called polypeptides, incorporating an extended backbone of robust [C−N] and [C−C] bonds (Fig. 2.2a). For a single polymer strand of 100 repeating amino acid units, a total of 20^{100} unique combinations of polypeptide sequences are possible. Considering that proteins can incorporate more than one polypetide chain with over 4,000 amino acid residues each, it is obvious that nature can assemble an enormous number of different biomolecules relying on the same fabrication strategy and a relatively small pool of building blocks.

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

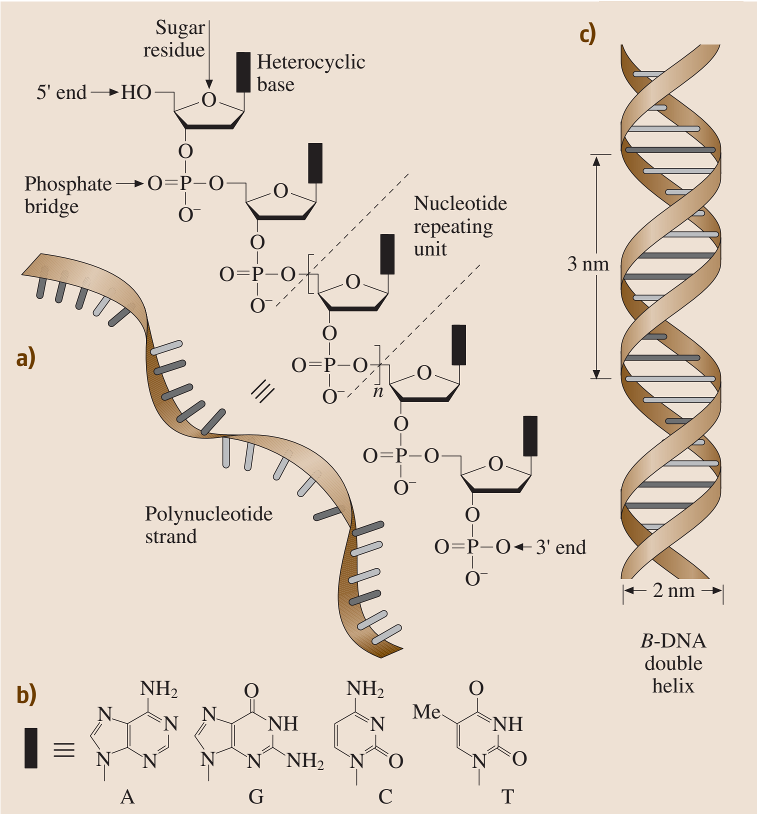
a)
Sugar residue
Heterocyclic base
5' end → HO
Phosphate → O=P–O bridge
O⁻
Nucleotide repeating unit
n
Polynucleotide strand
O=P–O ← 3' end
b)
NH₂
O
NH₂
O
Me
N
NH
NH₂
O
A
G
C
T
c)
3 nm
2 nm
B-DNA double helix

Fig. 2.1a–c A polynucleotide strand **(a)** incorporates alternating phosphate and sugar residues joined by covalent bonds. Each sugar carries one of four heterocyclic bases **(b)**. Noncovalent interactions between complementary bases in two independent polynucleotide strands encourage the formation of nanoscaled double helixes **(c)**

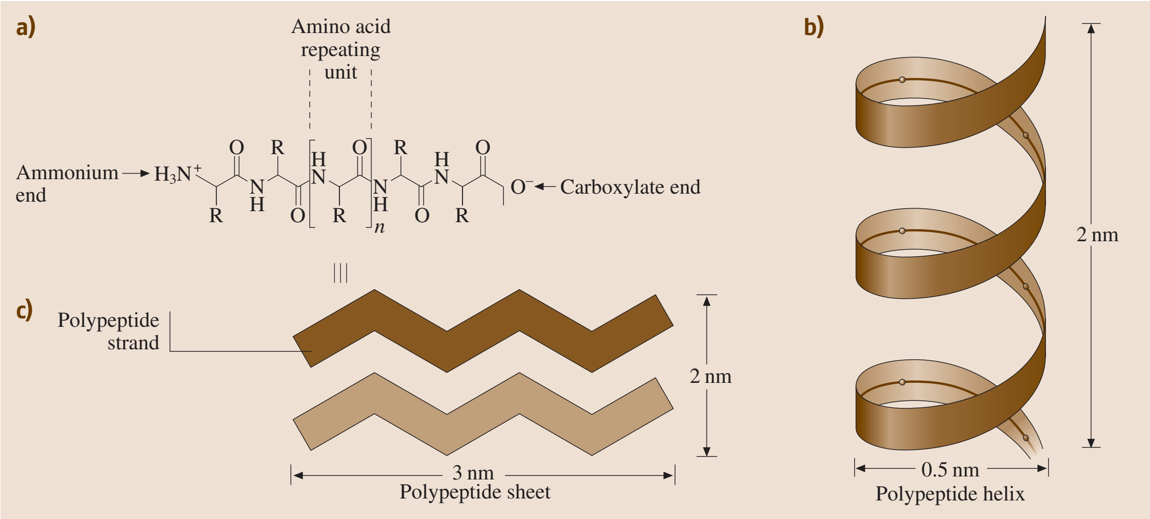
a)
Amino acid repeating unit
Ammonium → H₃N⁺ end
O
R
H
N
O
R
H
N
O
R
n
O⁻ ← Carboxylate end
c)
Polypeptide strand
2 nm
3 nm
Polypeptide sheet
b)
2 nm
0.5 nm
Polypeptide helix

Fig. 2.2a–c A polypeptide strand **(a)** incorporates amino acid residues differing in their side chains and joined by covalent bonds. Hydrogen bonding interactions curl a single polypeptide strand into a helical arrangement **(b)** or lock pairs of strands into nanoscaled sheets **(c)**

The covalent backbones of the polypeptide strands form the main skeleton of a protein molecule. In addition, myriad secondary interactions, involving noncovalent contacts between portions of the amino acid residues, control the arrangement of the individual polypeptide chains. Intrastrand hydrogen bonds curl single polypeptide chains around a longitudinal axis in a helical fashion to form tubular nanostructures ca. 0.5 nm wide and ca. 2 nm long (Fig. 2.2b). Similarly, interstrand hydrogen bonds can align from 2 up to 15 parallel or antiparallel polypeptide chains to form nanoscaled sheets with average dimensions of 2 × 3 nm (Fig. 2.2c). Multiple nanohelices and/or nanosheets combine into a unique three-dimensional arrangement dictating the overall shape and dimensions of a protein.

2.1.3 Chemical Synthesis of Artificial Nanostructures

Nature fabricates complex nanostructures relying on simple criteria and a relatively small pool of molecular building blocks. Robust chemical bonds join the basic components into covalent scaffolds. Noncovalent interactions determine the three-dimensional arrangement and overall shape of the resulting assemblies. The multitude of unique combinations possible for long sequences of chemically connected building blocks provides access to huge libraries of nanoscaled biomolecules.

Modern chemical synthesis has evolved considerably over the past few decades [2.2]. Experimental procedures to join molecular components with structural control at the picometer level are available. A multitude of synthetic schemes to encourage the formation of chemical bonds between selected atoms in reacting molecules have been developed. Furthermore, the tremendous progress of crystallographic and spectroscopic techniques has provided efficient and reliable tools to probe directly the structural features of artificial inorganic and organic compounds. It follows that designed molecules with engineered shapes and dimensions can be now prepared in a laboratory relying on the many tricks of chemical synthesis and the power of crystallographic and spectroscopic analyses.

The high degree of sophistication reached in this research area translates into the possibility of mimicking the strategies successfully employed by nature to fabricate chemically nanostructures [2.3]. Small molecular building blocks can be synthesized and joined covalently following routine laboratory procedures. It is even possible to design the stereoelectronic properties of the assembling components in order to shape the geometry of the final product with the assistance of noncovalent interactions. For example, five bipyridine building blocks (Fig. 2.3) can be connected in five synthetic steps to produce an oligobipyridine strand [2.4]. The five repeating units are bridged by [C−O] bonds and can chelate metal cations in the bay regions defined by their two nitrogen atoms. The spontaneous assembly of two organic strands in a double helical arrangement occurs in the presence of inorganic cations. In the resulting helicate, the two oligobipyridine strands wrap around an axis defined by five Cu(I) centers. Each inorganic cation coordinates two bipyridine units with a tetrahedral geometry imposing a diameter of ca. 0.6 nm on the nanoscaled helicate [2.5]. The overall length from one end of the helicate to the other is ca. 3 nm [2.6]. The analogy between this artificial double helix and the *B*-DNA double helix shown in Fig. 2.1c is obvious. In both instances, a supramolecular glue combines two independent molecular strands into nanostructures with defined shapes and dimensions.

The chemical synthesis of nanostructures can borrow nature's design criteria as well as its molecular building blocks. Amino acids, the basic components of proteins, can be assembled into artificial macrocycles. In the example of Fig. 2.4, eight amino acid residues are joined through the formation of [C−N] bonds in multiple synthetic steps [2.7]. The resulting covalent backbone defines a circular cavity with a diameter of ca. 0.8 nm [2.8]. In analogy to the polypeptide chains of proteins, the amino acid residues of this artificial oligopeptide can sustain hydrogen bonding interactions. It follows that multiple macrocycles can pile on top of each other to form tubular nanostructures. The walls of the resulting nanotubes are maintained in position by the cooperative action of at least eight primary hydrogen bonding contacts per macrocycle. These noncovalent interactions maintain the mean planes of independent macrocycles in an approximately parallel arrangement with a plane-to-plane separation of ca. 0.5 nm.

2.1.4 From Structural Control to Designed Properties and Functions

The examples in Figs. 2.3 and 2.4 demonstrate that modular building blocks can be assembled into target compounds with precise structural control at the picometer level through programmed sequences of synthetic steps. Indeed, modern chemical synthesis offers access to complex molecules with nanoscaled dimensions and, thus, provides cost-effective strategies for the production and characterization of billions of engineered nanostructures in parallel. Furthermore, the high degree of structural control is accompanied by the possibility of designing specific properties into the target nanostructures. Electroactive and photoactive components can be integrated chemically into functional molecular machines [2.9]. Extensive electrochemical investigations have demonstrated that inorganic and organic compounds can exchange electrons with macroscopic electrodes [2.10]. These studies have unraveled the processes responsible for the oxidation and reduction of numerous functional groups and indicated viable design criteria to adjust the ability of molecules to accept or donate electrons [2.11]. Similarly, detailed photochemical and photophysical investigations have elucidated the mechanisms responsible for the absorption and emission of photons at the molecular level [2.12]. The vast knowledge established on the interactions between light and molecules offers the opportunity to engineer chromophoric and fluorophoric functional groups with defined absorption and emission properties [2.11, 13].

The power of chemical synthesis to deliver functional molecules is, perhaps, better illustrated by the molecular motor shown in Fig. 2.5. The preparation of this [2]rotaxane requires 12 synthetic steps starting from known precursors [2.14]. This complex molecule incorporates a Ru(II)-trisbipyridine stopper bridged to a linear tetracationic fragment by a rigid triaryl spacer. The other end of the tetracationic portion is terminated by a bulky tetraarylmethane stopper. The bipyridinium unit of this dumbbell-shaped compound is encircled by a macrocyclic polyether. No covalent bonds join the macrocyclic and linear components.

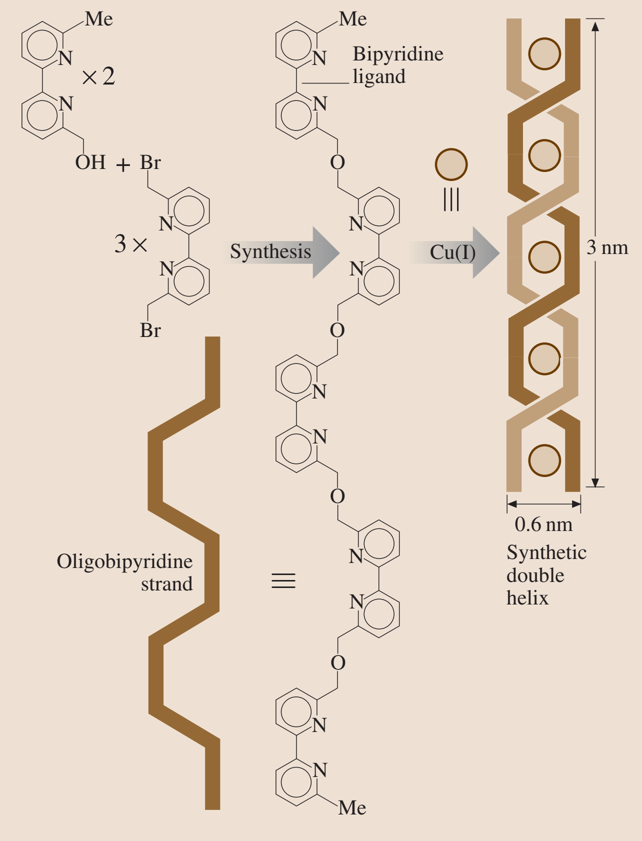

Fig. 2.3 An oligobipyridine strand can be synthesized joining five bipyridine subunits by covalent bonds. The tetrahedral coordination of pairs of bipyridine ligands by Cu(I) ions encourages the assembly two oligobipyridine strands into a double helical arrangement

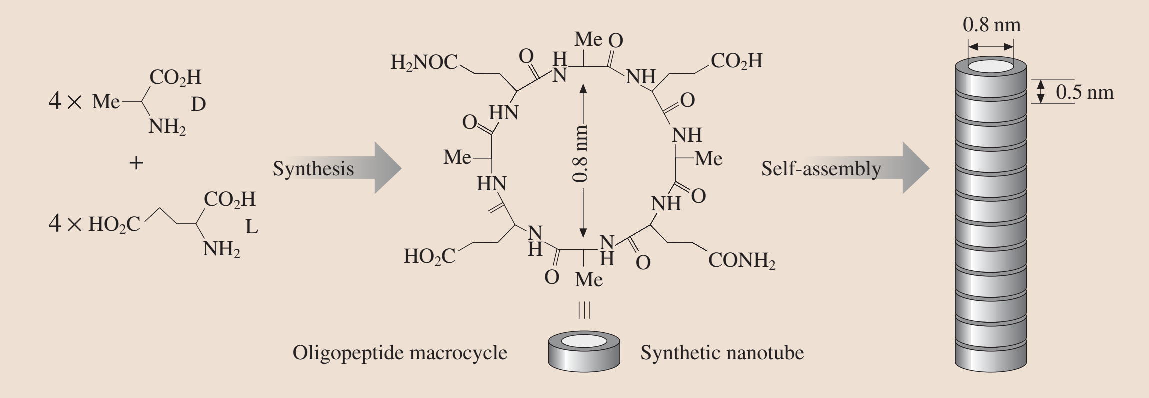

Fig. 2.4 Cyclic oligopeptides can be synthesized joining eight amino acid residues by covalent bonds. The resulting macrocycles self-assemble into nanoscaled tube-like arrays

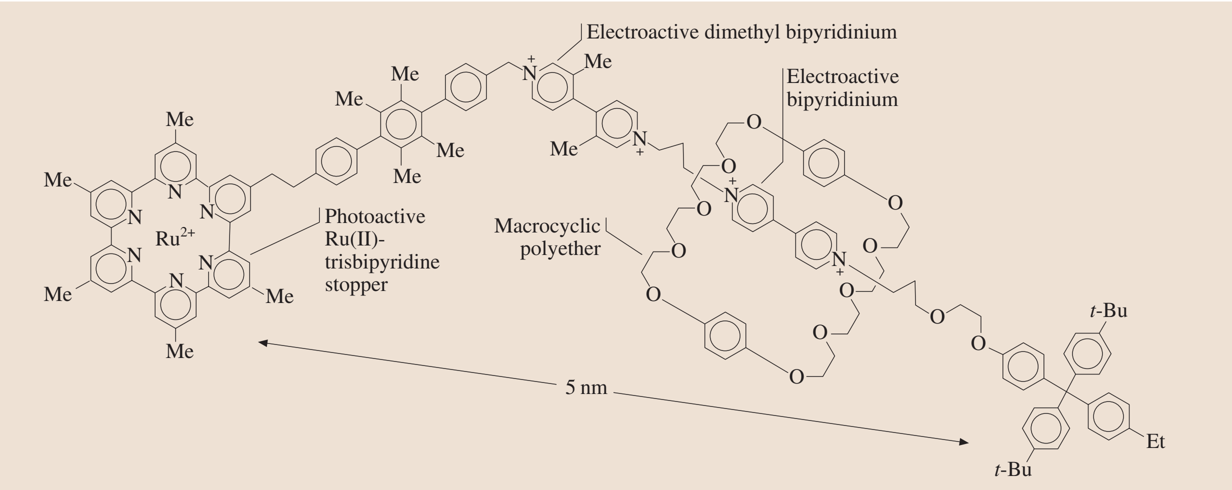
Electroactive dimethyl bipyridinium
Electroactive bipyridinium
Photoactive Ru(II)-trisbipyridine stopper
Macrocyclic polyether
Ru^{2+}
Me
N
O
5 nm
t-Bu
Et

Fig. 2.5 This nanoscaled [2]rotaxane incorporates a photoactive Ru(II)-trisbipyridine stopper and two electroactive bipyridinium units. Photoinduced electron transfer from the photoactive stopper to the encircled electroactive unit forces the macrocyclic polyether to shuttle to the adjacent bipyridinium dication

Rather, hydrogen bonding and [$\pi \cdots \pi$] stacking interactions maintain the macrocyclic polyether around the bipyridinium unit. In addition, mechanical constrains associated with the bulk of the two terminal stoppers prevent the macrocycle to slip off the thread. The approximate end-to-end distance for this [2]rotaxane is ca. 5 nm.

The bipyridinium and the 3,3′-dimethyl bipyridinium units within the dumbbell-shaped component undergo two consecutive and reversible monoelectronic reductions [2.14]. The two methyl substituents on the 3,3′-dimethyl bipyridinium dication make this electroactive unit more difficult to reduce. In acetonitrile, its redox potential is ca. 0.29 V more negative than that of the unsubstituted bipyridinium dication. Under irradiation at 436 nm in degassed acetonitrile, the excitation of the Ru(II)-trisbipyridine stopper is followed by electron transfer to the unsubstituted bipyridinium unit. In the presence of a sacrificial electron donor (triethanolamine) in solution, the photogenerated hole in the photoactive stopper is filled, and undesired back electron transfer is suppressed. The permanent and light-induced reduction of the dicationic bipyridinium unit to a radical cation depresses significantly the magnitude of the noncovalent interactions holding the macrocyclic polyether in position. As a result, the macrocycle shuttles from the reduced unit to the adjacent dicationic 3,3′-dimethyl bipyridinum. After the diffusion of molecular oxygen into the acetonitrile solution, oxidation occurs restoring the dicationic form of the bipyridinium unit and its ability to sustain strong noncovalent bonds. As a result, the macrocyclic polyether shuttles back to its original position. This amazing example of a molecular shuttle reveals that dynamic processes can be controlled reversibly at the molecular level relying on the clever integration of electroactive and photoactive fragments into functional and nanoscaled molecules.

2.2 Molecular Switches and Logic Gates

Everyday, we routinely perform dozens of switching operations. We turn on and off our personal computers, cellular phones, CD players, radios, or simple light bulbs at a click of a button. Every single time, our finger exerts a mechanical stimulation on a control device, namely a switch. The external stimulus changes the physical state of the switch closing or opening an electric circuit and enabling or preventing the passage of electrons.

Overall, the switch transduces a mechanical input into an electrical output.

2.2.1 From Macroscopic to Molecular Switches

The use of switching devices is certainly not limited to electric circuits. For example, a switch at the junction of a railroad can divert trains from one track to another. Similarly, a faucet in a lavatory pipe can block or release the flow of water. Of course, the nature of the control stimulations and the character of the final outcome vary significantly from case to case, but the operating principle behind each switching device is the same. In all cases, input stimulations reach the switch changing its physical state and producing a specific output.

The development of nanoscaled counterparts to conventional switches is expected to have fundamental scientific and technological implications. For instance, one can envisage practical applications for ultraminiaturized switches in areas ranging from biomedical research to information technology. The major challenge in the quest for nanoswitches, however, is the identification of reliable design criteria and operating principles for these innovative and fascinating devices. Chemical approaches to implement molecule-sized switches appear to be extremely promising. The intrinsically small dimensions of organic molecules coupled with the power of chemical synthesis are the main driving forces behind these exploratory investigations.

Certain organic molecules adjust their structural and electronic properties when stimulated with chemical, electrical, or optical inputs. Generally, the change is accompanied by an electrochemical or spectroscopic response. Overall, these nanostructures transduce input stimulations into detectable outputs and, appropriately, are called molecular switches [2.15, 16]. The chemical transformations associated with these switching processes are often reversible. The chemical system returns to the original state when the input signal is turned off. The interconverting states of a molecular switch can be isomers, an acid and its conjugated base, the oxidized and reduced forms of a redox active molecule, or even the complexed and uncomplexed forms of a receptor [2.9, 15, 16]. The output of a molecular switch can be a chemical, electrical, and/or optical signal that varies in intensity with the interconversion process. For example, changes in absorbance, fluorescence, pH, or redox potential can accompany the reversible transformation of a molecular switch.

2.2.2 Digital Processing and Molecular Logic Gates

In present computer networks, data are elaborated electronically by microprocessor systems [2.17] and are exchanged optically between remote locations [2.18]. Data processing and communication require the encoding of information in electrical and optical signals in the form of binary digits. Using arbitrary assumptions, logic thresholds can be established for each signal and, then, 0 and 1 digits can be encoded following simple conventions. Sequences of electronic devices manipulate the encoded bits executing logic functions as a result of basic switching operations.

The three basic AND, NOT, and OR operators combine binary inputs into binary outputs following precise logic protocols [2.17]. The NOT operator converts an input signal into an output signal. When the input is 0, the output is 1. When the input is 1, the output is 0. Because of the inverse relationship between the input and output values, the NOT gate is often called "inverter" [2.19]. The OR operator combines two input signals into a single output signal. When one or both inputs are 1, the output is 1. When both inputs are 0, the output is 0. The AND gate also combines two input signals into one output signal. In this instance, however, the output is 1 only when both inputs are 1. When at least one input is 0, the output is 0.

The output of one gate can be connected to one of the inputs of another operator. A NAND gate, for example, is assembled connecting the output of an AND operator to the input of a NOT gate. Now the two input signals are converted into the final output after two consecutive logic operations. In a similar fashion, a NOR gate can be assembled connecting the output of an OR operator to the input of a NOT gate. Once again, two consecutive logic operations determine the relation between two input signals and a single output. The NAND and NOR operations are termed universal functions because any conceivable logic operation can be implemented relying only on one of these two gates [2.17]. In fact, digital circuits are fabricated routinely interconnecting exclusively NAND or exclusively NOR operators [2.19].

The logic gates of conventional microprocessors are assembled interconnecting transistors, and their input and output signals are electrical [2.19]. But the concepts of binary logic can be extended to chemical, mechanical, optical, pneumatic, or any other type of signal. First it is necessary to design devices that can respond to these stimulations in the same way transistors respond to electrical signals. Molecular switches respond to a variety of input stimulations producing specific outputs and can, therefore, be exploited to implement logic functions [2.20, 21].

2.2.3 Molecular AND, NOT, and OR Gates

More than a decade ago, researchers proposed a potential strategy to execute logic operations at the molecular level [2.22]. Later, the analogy between molecular switches and logic gates was recognized in a seminal article [2.23], in which it was demonstrated that AND, NOT, and OR operations can be reproduced with fluorescent molecules. The pyrazole derivative **1** (Fig. 2.6) is a molecular NOT gate. It imposes an inverse relation between a chemical input (concentration of H^+) and an optical output (emission intensity). In a mixture of methanol and water, the fluorescence quantum yield of **1** is 0.13 in the presence of only 0.1 equivalents of H^+ [2.23]. The quantum yield drops to 0.003 when the equivalents of H^+ are 1,000. Photoinduced electron transfer from the central pyrazoline unit to the pendant benzoic acid quenches the fluorescence of the protonated form. Thus, a change in H^+ concentration (I) from a low to a high value switches the emission intensity (O) from a high to a low value. The inverse relationship between the chemical input I and the optical output O translates into the truth table of a NOT operation if a positive logic convention (low = 0, high = 1) is applied to both signals. The emission intensity is high (O = 1) when the concentration of H^+ is low (I = 0). The emission intensity is low (O = 0) when the concentration of H^+ is high (I = 1).

The anthracene derivative **2** (Fig. 2.6) is a molecular OR gate. It transduces two chemical inputs

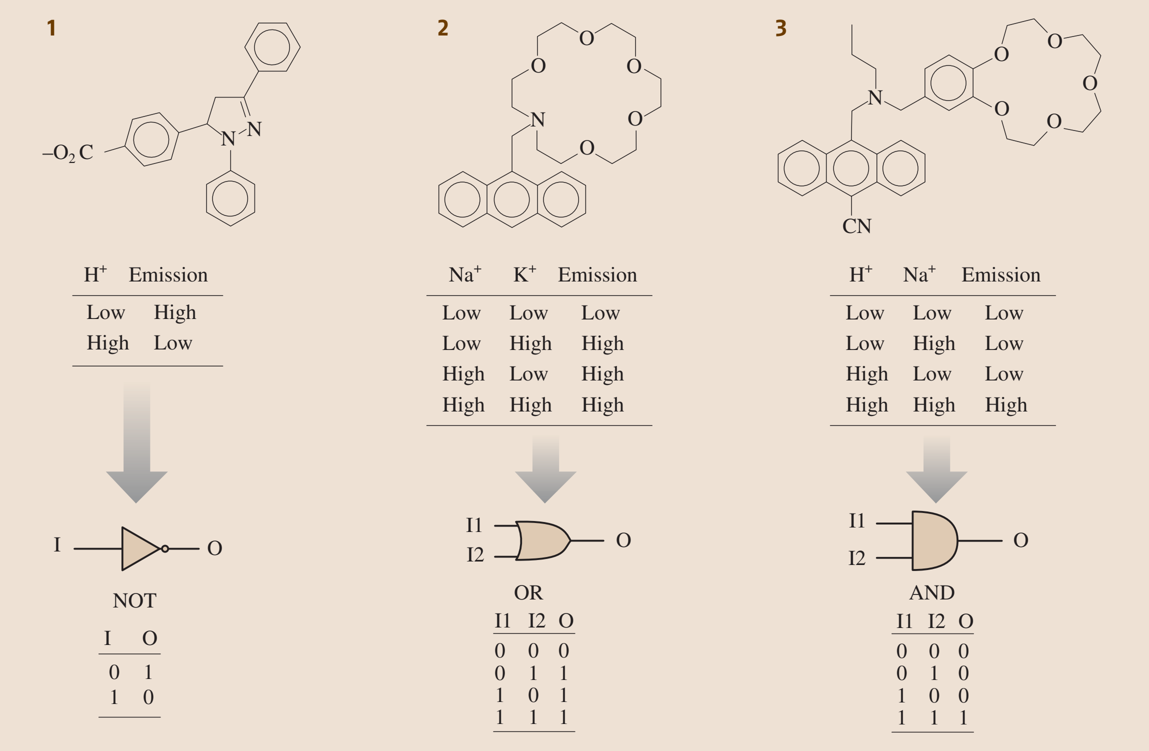

H^+	Emission
Low	High
High	Low

Na^+	K^+	Emission
Low	Low	Low
Low	High	High
High	Low	High
High	High	High

H^+	Na^+	Emission
Low	Low	Low
Low	High	Low
High	Low	Low
High	High	High

NOT

I	O
0	1
1	0

OR

I1	I2	O
0	0	0
0	1	1
1	0	1
1	1	1

AND

I1	I2	O
0	0	0
0	1	0
1	0	0
1	1	1

Fig. 2.6 The fluorescence intensity of the pyrazoline derivative **1** is high when the concentration of H^+ is low, and vice versa. The fluorescence intensity of the anthracene derivative **2** is high when the concentration of Na^+ and/or K^+ is high. The emission is low when both concentrations are low. The fluorescence intensity of the anthracene **3** is high only when the concentrations of H^+ and Na^+ are high. The emission is low in the other three cases. The signal transductions of the molecular switches **1**, **2**, and **3** translate into the truth tables of NOT, OR, and AND gates, respectively, if a positive logic convention is applied to all inputs and outputs (low = 0, high = 1)

(concentrations of Na^+ and K^+) into an optical output (emission intensity). In methanol, the fluorescence quantum yield is only 0.003 in the absence of metal cations [2.23]. Photoinduced electron transfer from the nitrogen atom of the azacrown fragment to the anthracene fluorophore quenches the emission. After the addition of 1,000 equivalents of either Na^+ or K^+, the quantum yield raises to 0.053 and 0.14, respectively. Similarly, the quantum yield is 0.14 when both metal cations are present in solution. The complexation of one of the two metal cations inside the azacrown receptor depresses the efficiency of the photoinduced electron transfer enhancing the fluorescence. Thus, changes in the concentrations of Na^+ (I1) and/or K^+ (I2) from low to high values switch the emission intensity (O) from a low to a high value. The relationship between the chemical inputs I1 and I2 and the optical output O translates into the truth table of an OR operation if a positive logic convention (low = 0, high = 1) is applied to all signals. The emission intensity is low (O = 0) only when the concentration of Na^+ and K^+ are low (I1 = 0, I2 = 0). The emission intensity is high (O = 1) for the other three input combinations.

The anthracene derivative **3** (Fig. 2.6) is a molecular AND gate. It transduces two chemical inputs (concentrations of H^+ and Na^+) into an optical output (emission intensity). In a mixture of methanol and *iso*-propanol, the fluorescence quantum yield is only 0.011 in the absence of H^+ or Na^+ [2.23]. Photoinduced electron transfer from either the tertiary amino group or the catechol fragment to the anthracene fluorophore quenches the emission. After the addition of either 100 equivalents of H^+ or 1,000 equivalents of Na^+, a modest change of the quantum yield to 0.020 and 0.011, respectively, is observed. Instead, the quantum yield increases to 0.068 when both species are present in solution. The protonation of the amino group and the insertion of the metal cation in the benzocrown ether receptor depress the efficiency of the photoinduced electron transfer processes enhancing the fluorescence. Thus, changes in the concentrations of H^+ (I1) and Na^+ (I2) from low to high values switch the emission intensity (O) from a low to a high value. The relationship between the chemical inputs I1 and I2 and the optical output O translates into the truth table of an AND operation if a positive logic convention (low = 0, high = 1) is applied to all signals. The emission intensity is high (O = 1) only when the concentration of H^+ and Na^+ are high (I1 = 1, I2 = 1). The emission intensity is low (O = 0) for the other three input combinations.

2.2.4 Combinational Logic at the Molecular Level

The fascinating molecular AND, NOT, and OR gates illustrated in Fig. 2.6 have stimulated the design of related chemical systems able to execute the three basic logic operations and simple combinations of them [2.20, 21]. Most of these molecular switches convert chemical inputs into optical outputs. But the implementation of logic operations at the molecular level is not limited to the use of chemical inputs. For example, electrical signals and reversible redox processes can be exploited to modulate the output of a molecular switch [2.24]. The supramolecular assembly **4** (Fig. 2.7) executes a XNOR function relying on these operating principles. The π-electron rich tetrathiafulvalene (TTF) guest threads the cavity of a π-electron deficient bipyridinium (BIPY) host. In acetonitrile, an absorption band associated with the charge-transfer interactions between the complementary π-surfaces is observed at 830 nm. Electrical stimulations alter the redox state of either the TTF or the BIPY units encouraging the separation of the two components of the complex and the disappearance of the charge-transfer band. Electrolysis at a potential of +0.5 V oxidizes the neutral TTF unit to a monocationic state. The now cationic guest is expelled from the cavity of the tetracationic host as a result of electrostatic repulsion. Consistently, the absorption band at 830 nm disappears. The charge-transfer band, however, is restored after the exhaustive back reduction of the TTF unit at a potential of 0 V. Similar changes in the absorption properties can be induced addressing the BIPY units. Electrolysis at −0.3 V reduces the dicationic BIPY units to their monocationic forms encouraging the separation of the two components of the complex and the disappearance of the absorption band. The original absorption spectrum is restored after the exhaustive back oxidation of the BIPY units at a potential of 0 V. Thus, this supramolecular system responds to electrical stimulations producing an optical output. One of the electrical inputs (I1) controls the redox state of the TTF unit switching between 0 and +0.5 V. The other (I2) determines the redox state of the bipyridinium units switching between −0.3 and 0 V. The optical output (O) is the absorbance of the charge-transfer band. A positive logic convention (low = 0, high = 1) can be applied to the input I1 and output O. A negative logic convention (low = 1, high = 0) can be applied to the input I2. The resulting truth table corresponds to that of a XNOR circuit (Fig. 2.7). The charge-transfer absorbance is high (O = 1) only when one voltage input is low and the other

4

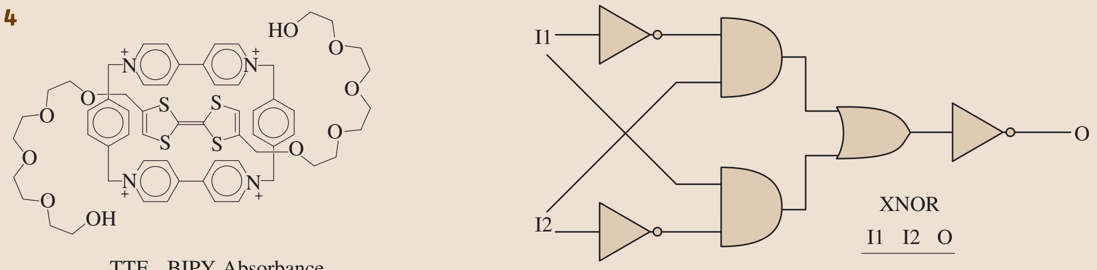

TTF	BIPY	Absorbance
Low	Low	High
Low	High	Low
High	Low	Low
High	High	High

XNOR

I1	I2	O
0	0	1
0	1	0
1	0	0
1	1	1

Fig. 2.7 The charge-transfer absorbance of the complex **4** is high when the voltage input addressing the tetrathiafulvalene (TTF) unit is low and that stimulating the bipyridinium (BIPY) units is high and vice versa. If a positive logic convention is applied to the TTF input and to the absorbance output (low = 0, high = 1) while a negative logic convention is applied to the BIPY input (low = 0, high = 1), the signal transduction of **4** translates into the truth table of a XNOR circuit

is high (I1 = 0, I2 = 0) or vice versa (I1 = 1, I2 = 1). It is important to note that the input string with both I1 and I2 equal to 1 implies that input potentials of +0.5 and −0.3 V are applied simultaneously to a solution containing the supramolecular assembly **4** and not to an individual complex. Of course, the concomitant oxidation of the TTF guest and reduction of the BIPY units in the very same complex would be unrealistic. In bulk solution, instead, some complexes are oxidized while others are reduced, leaving the average solution composition unaffected. Thus, the XNOR operation executed by this supramolecular system is a consequence of bulk properties and not a result of unimolecular signal transduction.

Optical inputs can be employed to operate the three-state molecular switch of Fig. 2.8 in acetonitrile solution [2.25]. This chemical system responds to three inputs producing two outputs. The three input stimulations are ultraviolet light (I1), visible light (I2), and the concentration of H^+ (I3). One of the two optical outputs is the absorbance at 401 nm (O1), which is high when the molecular switch is in the yellow-green state **6** and low in the other two cases. The other optical output is the absorbance at 563 nm (O2), which is high when the molecular switch is in the purple state **7** and low in the other two cases. The colorless spiropyran state **5** switches to the merocyanine form **7** upon irradiation with ultraviolet light. It switches to the protonated merocyanine from **6** when treated with H^+. The colored state **7** isomerizes back to **5** in the dark or upon irradiation with visible light. Alternatively, **7** switches to **6** when treated with H^+. The colored state **6** switches to **5**, when irradiated with visible light, and to **7**, after the removal of H^+. In summary, this three-state molecular switch responds to two optical inputs (I1 and I2) and one chemical input (I3) producing two optical outputs (O1 and O2). Binary digits can be encoded on each signal applying positive logic conventions (low = 0, high = 1). It follows that the three-state molecular switch converts input strings of three binary digits into output strings of two binary digits. The corresponding truth table (Fig. 2.8) reveals that the optical output O1 is high (O1 = 1) when only the input I3 is applied (I1 = 0, I2 = 0, I3 = 1), when only the input I2 is not applied (I1 = 1, I2 = 0, I3 = 0), or when all three inputs are applied (I1 = 1, I2 = 0, I3 = 0). The optical output O2 is high (O2 = 1) when only the input I1 is applied (I1 = 1, I2 = 0, I3 = 0) or when only the input I3 is not applied (I1 = 1, I2 = 0, I3 = 0). The combinational logic circuit (Fig. 2.8) equivalent to this truth table shows that all three inputs determine the output O1, while only I1 and I3 control the value of O2.

2.2.5 Intermolecular Communication

The combinational logic circuits in Figs. 2.7 and 2.8 are arrays of interconnected AND, NOT, and OR operators. The digital communication between these basic logic elements ensures the execution of a sequence of simple logic operations that results in the complex logic function processed by the entire circuit. It follows that

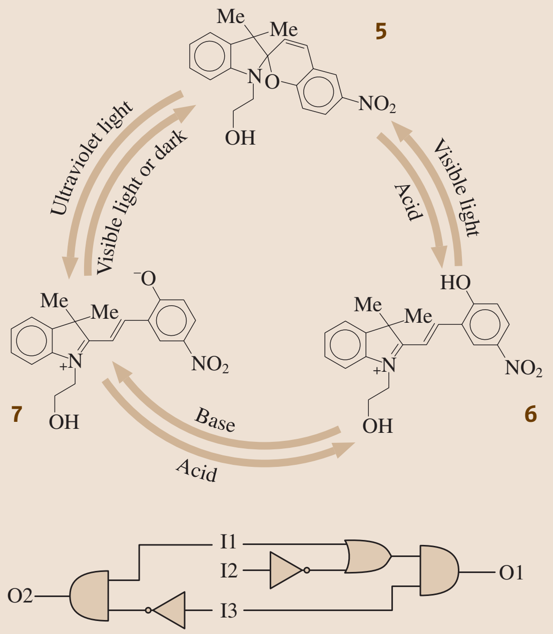

I1	I2	I3	O1	O2
0	0	0	0	0
0	0	1	1	0
0	1	0	0	0
1	0	0	0	1
0	1	1	0	0
1	0	1	1	0
1	1	0	0	1
1	1	0	1	0

Fig. 2.8 Ultraviolet light (I1), visible light (I2), and H^+ (I3) inputs induce the interconversion between the three states **5**, **6**, and **7**. The colorless state **5** does not absorb in the visible region. The yellow-green state **6** absorbs at 401 nm (O1). The purple state **7** absorbs at 563 nm (O2). The truth table illustrates the conversion of input strings of three binary digits (I1, I2, and I3) into output strings of two binary digits (O1 and O2) operated by this three-state molecular switch. A combinational logic circuit incorporating nine AND, NOT, and OR operators correspond to this particular truth table

the logic function of a given circuit can be adjusted altering the number and type of basic gates and their interconnection protocol [2.17]. This modular approach to combinational logic circuits is extremely powerful. Any logic function can be implemented connecting the appropriate combination of simple AND, NOT, and OR gates.

The strategies followed so far to implement complex logic functions with molecular switches are based on the careful design of the chemical system and on the judicious choice of the inputs and outputs [2.20, 21]. A specific sequence of AND, NOT, and OR operations is programmed in a single molecular switch. No digital communication between distinct gates is needed since they are built in the same molecular entity. Though extremely elegant, this strategy does not have the same versatility of a modular approach. A different molecule has to be designed, synthesized, and analyzed every single time a different logic function has to be realized. In addition, the degree of complexity that can be achieved with only one molecular switch is fairly limited. The connection of the input and output terminals of independent molecular AND, NOT, and OR operators, instead, would offer the possibility of assembling any combinational logic circuit from three basic building blocks.

In digital electronics, the communication between two logic gates can be realized connecting their terminals with a wire [2.19]. Methods to transmit binary data between distinct molecular switches are not so obvious and must be identified. Recently we developed two strategies to communicate signals between compatible molecular components. In one instance, a chemical signal is communicated between two distinct molecular switches [2.26]. They are the three-state switch illustrated in Fig. 2.8 and the two-state switch of Fig. 2.9. The merocyanine form **7** is a photogenerated base. Its p-nitrophenolate fragment, produced upon irradiation of the colorless state **5** with ultraviolet light, can abstract a proton from an acid present in the same solution. The resulting protonated form **6** is a photoacid. It releases a proton upon irradiation with visible light and can protonate a base co-dissolved in the same medium. The orange azopyridine **8** switches to the red-purple azopyridinium **9** upon protonation. This process is reversible, and the addition of a base restores the orange state **8**. It follows that photoinduced proton transfer can be exploited to communicate a chemical signal from **6** to **8** and from **9** to **7**. The two colored states **8** and **9** have different absorption properties in the visible region. In acetonitrile, the orange state **8** absorbs at 422 nm, and the red-purple state **9** absorbs at 556 nm. The changes in absorbance of these two bands can be exploited to monitor the photoinduced exchange of protons between the two communicating molecular switches.

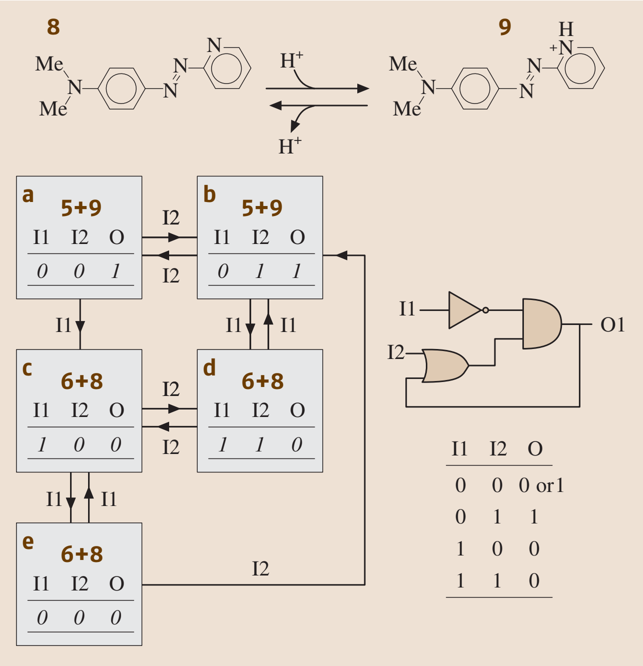

I1	I2	O
0	0	0 or 1
0	1	1
1	0	0
1	1	0

Fig. 2.9 The concentration of H^+ controls the reversible interconversion between the two states **8** and **9**. In response to ultraviolet (I1) and visible (I2) inputs, the three-state molecular switch in Fig. 2.7 modulates the ratio between these two forms and the absorbance (O) of **9** through photoinduced proton transfer. The truth table and sequential logic circuit illustrate the signal transduction behavior of the two communicating molecular switches. The interconversion between the five three-digit strings of input (I1 and I2) and output (O) data is achieved varying the input values in steps

The three-state molecular switch and the two-state molecular switch can be operated sequentially when dissolved in the same acetonitrile solution. In the presence of one equivalent of H^+, the two-state molecular switch is in state **9** and the absorbance at 556 nm is high ($O = 1$). Upon irradiation with ultraviolet light ($I1 = 0$), **5** switches to **7**. The photogenerated base deprotonates **9** producing **8** and **6**. As a result, the absorbance at 556 nm decreases ($O = 0$). Upon irradiation with visible light ($I2 = 1$), **6** switches to **5** releasing H^+. The result is the protonation of **8** to form **9** and restore the high absorbance at 556 nm ($O = 1$). In summary, the three-state molecular switch transduces two optical inputs ($I1 =$ ultraviolet light, $I2 =$ visible light) into a chemical signal (proton transfer) that is communicated to the two-state molecular switch and converted into a final optical output ($O =$ absorbance at 556 nm).

The logic behavior of the two communicating molecular switches is significantly different from those of the chemical systems illustrated in Figs. 2.6, 2.7, and 2.8 [2.26]. The truth table in Fig. 2.9 lists the four possible combinations of two-digit input strings and the corresponding one-digit output. The output digit O for the input strings 01, 10, and 11 can take only one value. In fact, the input string 01 is transduced into a 1, and the input strings 10 and 11 are converted into 0. Instead, the output digit O for the input string 00 can be either 0 or 1. The sequence of events leading to the input string 00 determines the value of the output. The boxes ***a–e*** in Fig. 2.9 illustrates this effect. They correspond to the five three-digit input/output strings. The transformation of one box into any of the other four is achieved in one or two steps by changing the values of I1 and/or I2. In two instances (***a*** and ***b***), the two-state molecular switch is in state **9**, and the output signal is high ($O = 1$). In the other three cases (***c***, ***d***, and ***e***), the two-state molecular switch is in state **8**, and the output signal is low ($O = 0$). The strings 000 (***e***) and 001 (***a***) correspond to the first entry of the truth table. They share the same input digits but differ in the output value. The string 000 (***e***) can be obtained only from the string 100 (***c***) varying the value of I1. Similarly, the string 001 (***a***) can be accessed only from the string 011 (***b***) varying the value of I2. In both transformations, the output digit remains unchanged. Thus, the value of O1 in the parent string is memorized and maintained in the daughter string when both inputs become 0. This memory effect is the fundamental operating principle of sequential logic circuits [2.17], which are used extensively to assemble the memory elements of modern microprocessors. The sequential logic circuit equivalent to the truth table of the two communicating molecular switches is also shown in Fig. 2.9. In this circuit, the input data I1 and I2 are combined through NOT, OR, and AND operators. The output of the AND gate O is also an input of the OR gate and controls, together with I1 and I2, the signal transduction behavior.

The other strategy for digital transmission between molecules is based on the communication of optical signals between the three-state molecular switch (Fig. 2.8) and fluorescent compounds [2.27,28]. In the optical network of Fig. 2.10, three optical signals travel from an excitation source to a detector after passing through two quartz cells. The first cell contains an equimolar acetonitrile solution of naphthalene, anthracene, and tetracene. The second cell contains an acetonitrile solu-

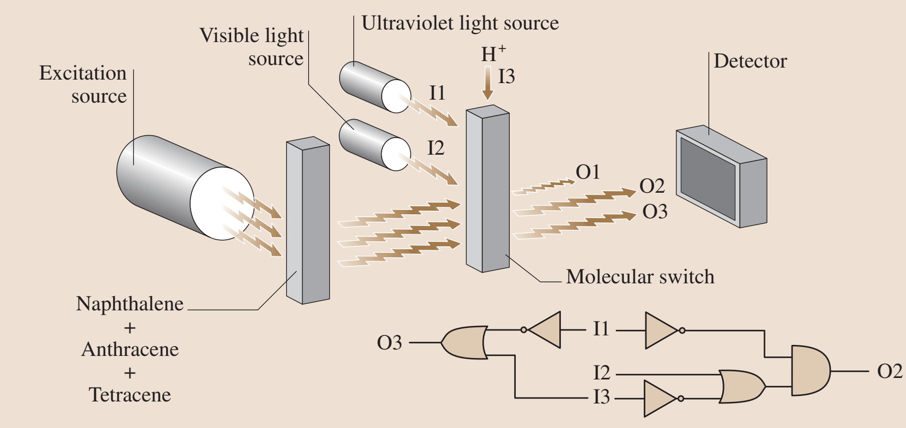

I1	I2	I3	O1	O2	O3
0	0	0	0	1	1
0	0	1	0	0	1
0	1	0	0	1	1
1	0	0	0	0	0
0	1	1	0	1	1
1	0	1	0	0	1
1	1	0	0	0	0
1	1	1	0	0	1

Fig. 2.10 The excitation source sends three monochromatic light beams (275, 357, and 441 nm) to a quartz cell containing an equimolar acetonitrile solution of naphthalene, anthracene and tetracene. The three fluorophores absorb the exciting beams and reemit at 305, 401, and 544 nm, respectively. The light emitted in the direction perpendicular to the exciting beams passes through another quartz cell containing an acetonitrile solution of the three-state molecular switch shown in Fig. 2.7. Ultraviolet (I1), visible (I2), and H^+ (I3) inputs control the interconversion between the three states of the molecular switch. They determine the intensity of the optical outputs reaching the detector and correspond to the naphthalene (O1), anthracene (O2), and tetracene (O3) emissions. The truth table and equivalent combinational logic illustrate the relation between the three inputs and the three outputs. The output O1 is always 0, and it is not influenced by the three inputs. Only two inputs determine the value of O3, while all of them control the output O2

tion of the three-state molecular switch. The excitation source sends three consecutive monochromatic light beams to the first cell stimulating the emission of the three fluorophores. The light emitted in the direction perpendicular to the exciting beam reaches the second cell. When the molecular switch is in state **5**, the naphthalene emission at 335 nm is absorbed and a low intensity output (O1) reaches the detector. Instead, the anthracene and tetracene emissions at 401 and 544 nm, respectively, pass unaffected and high intensity outputs (O2 and O3) reach the detector. When the molecular switch is in state **6**, the naphthalene and anthracene emissions are absorbed and only the tetracene emission reaches the detector ($O1 = 0$, $O2 = 0$, $O3 = 1$). When the molecular switch is state **7**, the emission of all three fluorophores is absorbed ($O1 = 0$, $O2 = 0$, $O3 = 0$). The interconversion of the molecular switch between the three states is induced addressing the second cell with ultraviolet (I1), visible (I2) and H^+ (I3) inputs. Thus, three independent optical outputs (O1, O2 and O3) can be modulated stimulating the molecular switch with two optical and one chemical input. The truth table in Fig. 2.10 illustrates the relation between the three inputs (I1, I2 and I3) and the three outputs (O1, O2 and O3), when positive logic conventions are applied to all signals. The equivalent logic circuit shows that all three inputs control the anthracene channel O2, but only I1 and I3 influence the tetracene channel O3. Instead, the intensity of the naphthalene channel O1 is always low, and it is not affected by the three inputs.

The operating principles of the optical network in Fig. 2.10 can be simplified to implement all-optical logic gates. The chemical input inducing the formation of the protonated form **6** of the molecular switch can be eliminated. The interconversion between the remaining two states **5** and **7** can be controlled relying exclusively on ultraviolet inputs. Indeed, ultraviolet irradiation induces the isomerization of the colorless form **5** to the colored species **7**, which reisomerizes to the original state in the dark. Thus, a single ultraviolet source is sufficient to control the switching from **5** to **7** and vice versa. On the basis of these considerations, all-optical NAND, NOR, and NOT gates can be implemented operating sequentially or in parallel from one to three independent switching elements [2.28]. For example, the all-optical network illustrated in Fig. 2.11 is a three-input NOR gate. A monochromatic optical signal travels from a visible source to a detector. Three switching elements are aligned along the path of the traveling light. They are quartz cells containing an acetonitrile solution of the mo-

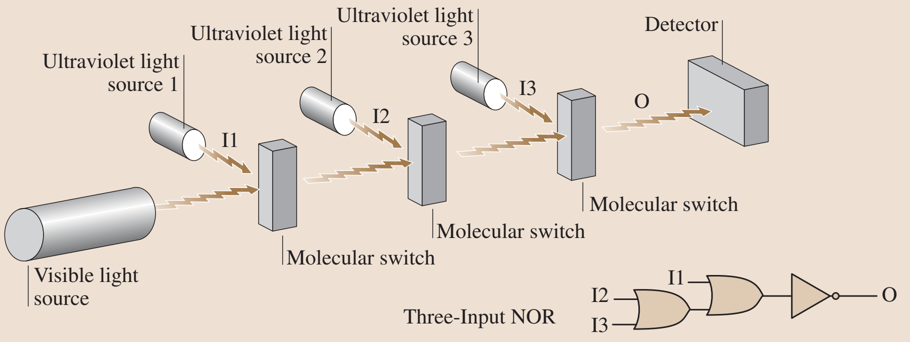

I1	I2	I3	O
0	0	0	1
0	0	1	0
0	1	0	0
1	0	0	0
0	1	1	0
1	0	1	0
1	1	0	0
1	1	1	0

Fig. 2.11 The visible source sends a monochromatic beam (563 nm) to the detector. The traveling light is forced to pass through three quartz cells containing the molecular switch illustrated in Fig. 2.7. The three switching elements are operated by independent ultraviolet inputs. When at least one of them is on, the associated molecular switch is in the purple form **7**, which can absorb and block the traveling light. The truth table and equivalent logic circuit illustrate the relation between the three inputs I1, I2, and I3 and the optical output O

lecular switch shown in Fig. 2.8. The interconversion of the colorless form **5** into the purple isomer **7** is induced stimulating the cell with an ultraviolet input. The reisomerization from **7** to **5** occurs spontaneously, as the ultraviolet sources is turned off. Using three distinct ultraviolet sources, the three switching elements can be controlled independently.

The colorless form **5** does not absorb in the visible region, while the purple isomer **7** has a strong absorption band at 563 nm. Thus, a 563 nm optical signal leaving the visible source can reach the detector unaffected only if all three switching elements are in the nonabsorbing state **5**. If one of the three ultraviolet inputs I1, I2, or I3 is turned on, the intensity of the optical output O drops to 3–4% of its original value. If two or three ultraviolet inputs are turned on simultaneously, the optical output drops to 0%. Indeed, the photogenerated state **7** absorbs and blocks the traveling light. Applying positive logic conventions to all signals, binary digits can be encoded in the three optical inputs and in the optical output. The resulting truth table is illustrated in Fig. 2.11. The output O is 1 only if all three inputs I1, I2, or I3 are 0. The output O is 0 if at least one of the three inputs I1, I2, or I3 is 1. This signal transduction corresponds to that executed by a three-input NOR gate, which is a combination of one NOT and two OR operators.

2.3 Solid State Devices

The fascinating chemical systems illustrated in Figs. 2.6–2.11 demonstrate that logic functions can be implemented relying on the interplay between designed molecules and chemical, electrical and/or optical signals [2.20, 21].

2.3.1 From Functional Solutions to Electroactive and Photoactive Solids

These molecular switches, however, are operated exclusively in solution and remain far from potential applications in information technology at this stage. The integration of liquid components and volatile organic solvents in practical digital devices is hard to envisage. Furthermore, the logic operations executed by these chemical systems rely on bulk addressing. Although the individual molecular components have nanoscaled dimensions, macroscopic collections of them are employed for digital processing. In some instances, the operating principles cannot even be scaled down to the unimolecular level. Often bulk properties are responsible for signal transduction. For example, a single fluorescent compound **2** cannot execute an OR operation. Its azacrown appendage can accommodate only one metal cation. As a result, an individual molecular switch can respond to only one of the two chemical inputs. It is a collection of numerous molecular switches dissolved in an organic solvent that responds to both inputs enabling an OR operation.

The development of miniaturized molecule-based devices requires the identification of methods to transfer the switching mechanisms developed in solution to the solid state [2.29]. Borrowing designs and fabrication strategies from conventional electronics, researchers are starting to explore the integration of molecular components into functional circuits and devices [2.30–33]. Generally, these strategies combine lithography and surface chemistry to assemble nanometer-thick organic films on the surfaces of microscaled or nanoscaled electrodes. Two main approaches for the deposition of organized molecular arrays on inorganic supports have emerged so far. In one instance, amphiphilic molecular building blocks are compressed into organized monolayers at air/water interfaces. The resulting films can be transferred on supporting solids employing the Langmuir–Blodgett technique [2.34]. Alternatively, certain molecules can be designed to adsorb spontaneously on the surfaces of compatible solids from liquid or vapor phases. The result is the self-assembly of organic layers on inorganic supports [2.35].

2.3.2 Langmuir–Blodgett Films

Films of amphiphilic molecules can be deposited on a variety of solid supports employing the Langmuir–Blodgett technique [2.34]. This method can be extended to electroactive compounds incorporating hydrophilic and hydrophobic groups. For example, the amphiphile **10** (Fig. 2.12) has a hydrophobic hexadecyl tail attached to a hydrophilic bipyridinium dication [2.36, 37]. This compound dissolves in mixtures of chloroform and methanol, but it is not soluble in moderately concentrated aqueous solutions of sodium perchlorate.

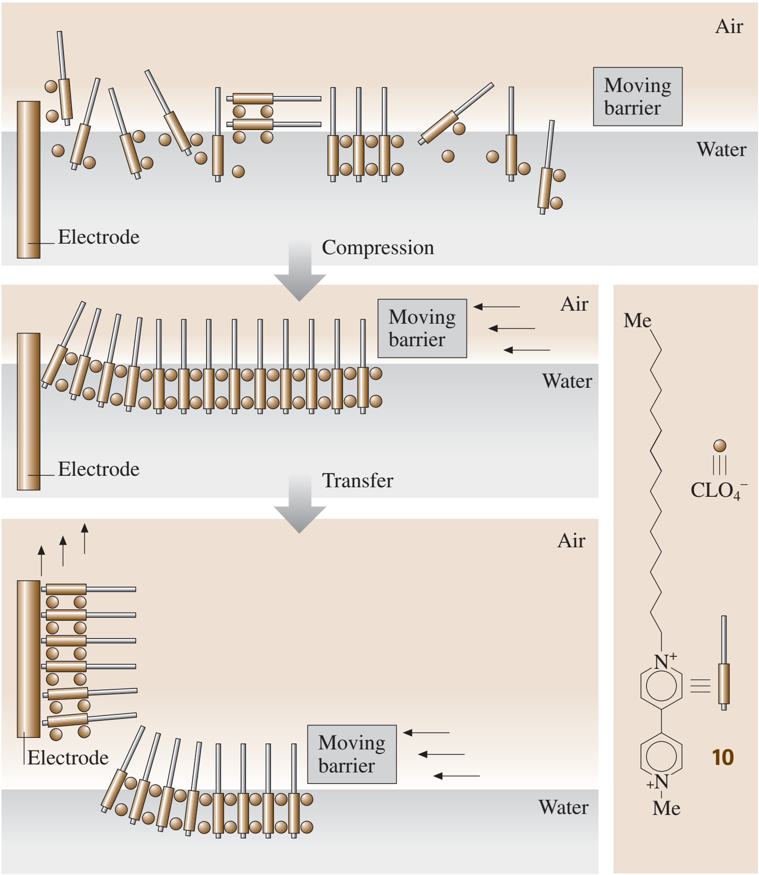

Fig. 2.12 The compression of the amphiphilic dication **10** with a moving barrier results in the formation of a packed monolayer at the air/water interface. The lifting of an electrode pre-immersed in the aqueous subphase encourages the transfer of part of the monolayer on the solid support

Thus the spreading of an organic solution of **10** on an aqueous sodium perchlorate subphase affords a collection of disorganized amphiphiles floating on the water surface (Fig. 2.12), after the organic solvent has evaporated. The molecular building blocks can be compressed into a monolayer with the aid of a moving barrier. The hydrophobic tails align away from the aqueous phase. The hydrophilic dicationic heads and the accompanying perchlorate counterions pack to form an organized monolayer at the air/water interface. The compression process can be monitored recording the surface pressure (π)-area per molecule (A) isotherm, which indicates a limiting molecular area of ca. 50 Å^2. This value is larger than the projected area of an oligomethylene chain. It correlates reasonably, however, with the overall area of a bipyridinium dication plus two perchlorate anions.

The monolayer prepared at the air/water interface (Fig. 2.12) can be transferred on the surface of a indium-tin oxide electrode pre-immersed in the aqueous phase. The slow lifting of the solid support drags the monolayer away from the aqueous subphase. The final result is the coating of the electrode with an organic film containing electroactive bipyridinium building blocks. The modified electrode can be integrated in a conventional electrochemical cell to probe the redox response of the electroactive layer. The resulting cyclic voltammograms reveal the characteristic waves for the first reduction process of the bipyridinium dications, confirming the successful transfer of the electroactive amphiphiles from the air/water interface to the electrode surface. The integration of the redox waves indicates a surface coverage of ca. 4×10^{10} mol cm^{-2}. This value corresponds to a molecular area of ca. 40 Å^2 and is in excellent agreement with the limiting molecular area of the π–A isotherm.

These seminal experiments demonstrate that electroactive amphiphiles can be organized into uniform monolayers at the air/water interface and then transferred efficiently on the surface of appropriate substrates to produce electrode/monolayer junctions. The resulting electroactive materials can become the functional components of molecule-based devices. For example, bipyridinium-based photodiodes can be fabricated following this approach [2.38, 39]. Their operating principles rely on photoinduced electron transfer from chromophoric units to bipyridinium acceptors. The electroactive and photoactive amphiphile **11** (Fig. 2.13) incorporates hydrophobic ferrocene and pyrene tails and a hydrophilic bipyridinium head. Chloroform solutions of **11** containing ten equivalents of arachidic acid can be spread on an aqueous calcium chloride subphase in a Langmuir trough. The amphiphiles can be compressed into a mixed monolayer, after the evaporation of the organic solvent. Pronounced steps in the corresponding π–A isotherm suggest that the bulky ferrocene and pyrene groups are squeezed away from the water surface. In the final arrangement, both photoactive groups align above the hydrophobic dication.

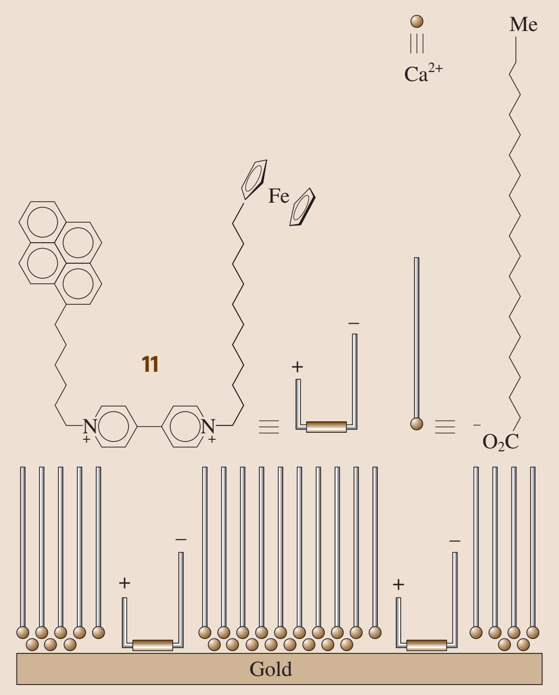

Fig. 2.13 Mixed monolayers of the amphiphile **11** and arachidic acid can be transferred from the air/water interface to the surface of an electrode to generate a molecule-based photodiode

A mixed monolayer of **11** and arachidic acid can be transferred from the air/water interface to the surface of a transparent gold electrode following the methodology illustrated for the system in Fig. 2.12. The coated electrode can be integrated in a conventional electrochemical cell. Upon irradiation at 330 nm under an inert atmosphere, an anodic photocurrent of ca. 2 nA develops at a potential of 0 V relative to a saturated calomel electrode. Indeed, the illumination of the electroactive monolayer induces the electron transfer from the pyrene appendage to the bipyridinium acceptor and then from the reduced acceptor to the electrode. A second intramolecular electron transfer from the ferrocene donor to the oxidized pyrene fills its photogenerated hole. Overall, a unidirectional flow of electrons across the monolayer/electrode junction is established under the influence of light.

The ability to transfer electroactive monolayers from air/water interfaces to electrode surfaces can be exploited to fabricate molecule-based electronic devices. In particular, arrays of interconnected electrode/monolayer/electrode tunneling junctions can be assembled combining the Langmuir–Blodgett technique with electron beam evaporation [2.33]. Fig. 2.14 illustrates a schematic representation of the resulting devices. Initially, parallel fingers are patterned on a silicon wafer with a silicon dioxide overlayer by electron beam evaporation. The bottom electrodes deposited on the support can be either aluminum wires covered by an aluminum oxide or n-doped silicon lines with silicon dioxide overlayers. Their widths are ca. 6 or 7 μm, respectively. The patterned silicon chip is immersed in the aqueous subphase of a Langmuir trough prior to monolayer formation. After the compression of electroactive amphiphiles at the air/water interface, the substrate is pulled out of the aqueous phase to encourage the transfer of the molecular layer on the parallel bottom electrodes as well as on the gaps between them. Then, a second set of electrodes orthogonal to the first is deposited through a mask by electron beam evaporation. They consist of a titanium underlayer plus an aluminum overlayer. Their thicknesses are ca. 0.05 and 1 μm, respectively, and their width is ca. 10 μm. In the final assembly, portions of the molecular layer become sandwiched between the bottom and top electrodes. The active areas of these electrode/monolayer/electrode junctions are ca. $60-70\,\mu\mathrm{m}^2$ and correspond to ca. 10^6 molecules.

The [2]rotaxane **12** (Fig. 2.14) incorporates a macrocyclic polyether threaded onto a bipyridinium-based backbone [2.40, 41]. The two bipyridinium dications are bridged by a m-phenylene spacer and terminated by tetraarylmethane appendages. These two bulky groups trap mechanically the macrocycle preventing its dis-

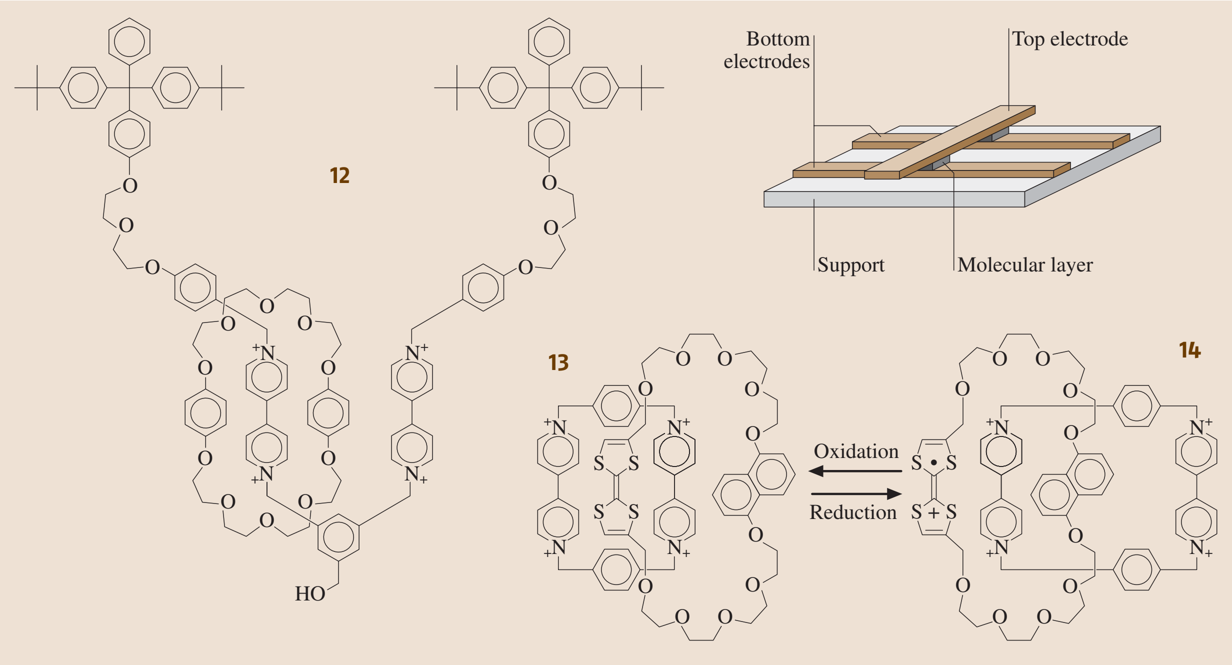

Fig. 2.14 The [2]rotaxane **12** and the [2]catenane **13** can be compressed into organized monolayers at air/water interfaces. The resulting monolayers can be transferred on the bottom electrodes of a patterned silicon support. After the deposition of a top electrode, electrode/monolayer/electrode junctions can be assembled. Note that only the portion of the monolayer sandwiched between the top and bottom electrodes is shown in the diagram. The oxidation of the tetrathiafulvalene unit of the [2]catenane **13** is followed by the circumrotation of the macrocyclic polyether to afford the [2]catenane **14**. The process is reversible, and the reduction of the cationic tetrathiafulvalene unit restores the original state

sociation from the tetracationic backbone. In addition, their hydrophobicity complements the hydrophilicity of the two bipyridinium dications imposing amphiphilic character on the overall molecular assembly. This compound does not dissolve in aqueous solutions and can be compressed into organized monolayers at air/water interfaces. The corresponding π–A isotherm reveals a limiting molecular area of ca. 130 Å^2. This large value is a consequence of the bulk associated with the hydrophobic tetraarylmethane tails and the macrocycle encircling the tetracationic backbone.

Monolayers of the [2]rotaxane **12** can be transferred from the air/water interface to the surfaces of the bottom aluminum/aluminum oxide electrodes of a patterned silicon chip with the hydrophobic tetraarylmethane groups pointing away from the supporting substrate. The subsequent assembly of a top titanium/aluminum electrode affords electrode/monolayer/electrode junctions. Their current/voltage signature can be recorded grounding the top electrode and scanning the potential of the bottom electrode. A pronounced increase in current is observed when the potential is lowered below -0.7 V. Under these conditions, the bipyridinium-centered LUMOs mediate the tunneling of electrons from the bottom to the top electrode leading to a current enhancement. A similar current profile is observed if the potential is returned to 0 and then back to -2 V. Instead, a modest increase in current in the opposite direction is observed when the potential is raised above $+0.7$ V. Presumably, this trend is a result of the participation of the phenoxy-centered HOMOs in the tunneling process. After a single positive voltage pulse, however, no current can be detected if the potential is returned to negative values. In summary, the positive potential scan suppresses irreversibly the conducting ability of the electrode/molecule/electrode junction. The behavior of this device correlates with the redox response of the [2]rotaxane **12** in solution. Cyclic voltammograms reveal reversible monoelectronic reductions of the bipyridinium dications. But they also show two irreversible oxidations associated, presumably, with the phenoxy rings of the macrocycle and tetraarylmethane groups. These observations suggest that a positive voltage pulse applied to the electrode/monolayer/electrode junction oxidizes irreversibly the sandwiched molecules suppressing their ability to mediate the transfer of electrons from the bottom to the top electrode under a negative bias.

The device incorporating the [2]rotaxane **13** can be exploited to implement simple logic operations [2.40]. The two bottom electrodes can be stimulated with voltage inputs (I1 and I2) while measuring a current output (O) at the common top electrode. When at least one of the two inputs is high (0 V), the output is low (< 0.7 nA). When both inputs are low (-2 V), the output is high (ca. 4 nA). If a negative logic convention is applied to the voltage inputs (low $= 1$, high $= 0$) and a positive logic convention is applied to the current output (low $= 0$, high $= 1$), the signal transduction behavior translates into the truth table of an AND gate. The output O is 1 only when both inputs are 1. Instead, an OR operation can be executed if the logarithm of the current is considered as the output. The logarithm of the current is -12 when both voltage inputs are 0 V. It raises to ca. -9 when one or both voltage inputs are lowered to -2 V. This signal transduction behavior translates into the truth table of an OR gate if a negative logic convention is applied to the voltage inputs (low $= 1$, high $= 0$) and a positive logic convention is applied to the current output (low $= 0$, high $= 1$). The output O is 1 when at least one of the two inputs is 1.

The [2]catenane **13** (Fig. 2.14) incorporates a macrocyclic polyether interlocked with a tetracationic cyclophane [2.42, 43]. Organic solutions of the hexafluorophosphate salt of this [2]catenane and six equivalents of the sodium salt of dimyristoylphosphatidic acid can be co-spread on the water surface of a Langmuir trough [2.44]. The sodium hexafluorophosphate formed dissolves in the supporting aqueous phase, while the hydrophilic bipyridinium cations and the amphiphilic anions remain at the interface. Upon compression, the anions align their hydrophobic tails away from the water surface forming a compact monolayer above the cationic bipyridinium derivatives. The corresponding π–A isotherm indicates limiting molecular areas of ca. 125 Å^2. This large value is a consequence of the bulk associated with the two interlocking macrocycles.

Monolayers of the [2]catenane **13** can be transferred from the air/water interface to the surfaces of the bottom n-doped silicon/silicon dioxide electrodes of a patterned silicon chip with the hydrophobic tails of the amphiphilic anions pointing away from the supporting substrate [2.45, 46]. The subsequent assembly of a top titanium/aluminum electrode affords electrode/monolayer/electrode arrays. Their junction resistance can be probed grounding the top electrode and maintaining the potential of the bottom electrode at $+0.1$ V. If a voltage pulse of $+2$ V is applied to the bottom electrode before the measurement, the junction resistance probed is ca. 0.7 GΩ. After a pulse of -2 V applied to the bottom electrode, the junction resistance probed at $+0.1$ V drops ca. 0.3 GΩ. Thus, alternating positive and negative voltage pulses can switch reversibly the junction resistance between high and low values. This intriguing behavior is a result of the redox and dynamic properties of the [2]catenane **13**.

Extensive spectroscopic and crystallographic studies [2.42, 43] demonstrated that the tetrathiafulvalene unit resides preferentially inside the cavity of the tetracationic cyclophane of the [2]catenane **13** (Fig. 2.14). Attractive $[\pi \cdots \pi]$ stacking interactions between the neutral tetrathiafulvalene and the bipyridinium dications are responsible for this co-conformation. Oxidation of the tetrathiafulvalene generates a cationic form that is expelled from the cavity of the tetracationic cyclophane. After the circumrotation of the macrocyclic polyether, the oxidized tetrathiafulvalene is exchanged with the neutral 1,5-dioxynaphthalene producing the [2]catenane **14** (Fig. 2.14). The reduction of the tetrathiafulvalene back to its neutral state is followed by the circumrotation of the macrocyclic polyether, which restores the original state **14**. The voltage pulses applied to the bottom electrode of the electrode/monolayer/junction oxidize and reduce the tetrathiafulvalene unit inducing the interconversion between the forms **13** and **14**. The difference in the stereoelectronic properties of these two states translates into distinct current/voltage signatures. Indeed, their ability to mediate the tunneling of electrons across the junction differs significantly. As a result, the junction resistance probed at a low voltage after an oxidizing pulse is significantly different from that determined under the same conditions after a reducing pulse.

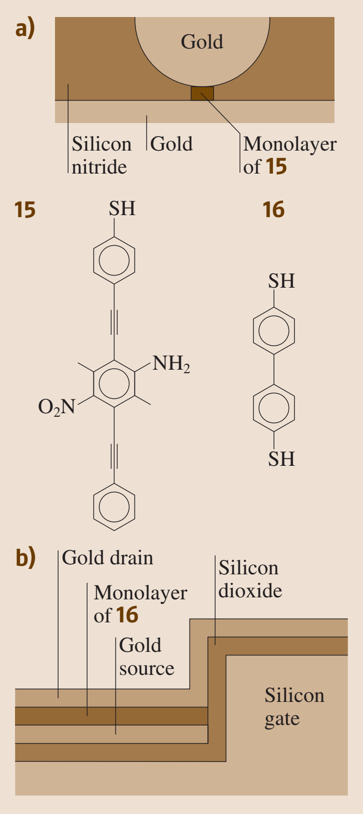

Fig. 2.15 **(a)** A monolayer of the thiol **15** is embedded between two gold electrodes maintained in position by a silicon nitride support. **(b)** A monolayer of the bisthiol **16** is located between two gold electrodes and stimulated by a silicon gate

2.3.3 Self-Assembled Monolayers

In the examples illustrated in Figs. 2.12–2.15, monolayers of amphiphilic and electroactive derivatives are assembled at air/water interfaces and then transferred on the surfaces of appropriate substrates. An alternative strategy to coat electrodes with molecular layers relies on the ability of certain compounds to adsorb spontaneously on solid supports from liquid or vapor phases [2.35]. In particular, the affinity of certain sulfurated functional groups for gold can be exploited to encourage the self-assembly of organic molecules on microscaled and nanoscaled electrodes.

The electrode/monolayer/electrode junction in Fig. 2.15a incorporates a molecular layer between two gold electrodes mounted on a silicon nitride support. This device can be fabricated combining chemical vapor deposition, lithography, anisotropic etching, and self-assembly [2.47]. Initially, a silicon wafer is coated with a 50 nm thick layer of silicon nitride by low pressure chemical vapor deposition. Then, a square of $400 \times 400\,\mu\text{m}$ is patterned on one side of the coated wafer by optical lithography and reactive ion etching. Anisotropic etching of the exposed silicon up to the other side of the wafer leaves a suspended silicon nitride membrane of $40 \times 40\,\mu\text{m}$. Electron beam lithography and reactive ion etching can be used to carve a bowl-shaped hole (diameter = 30–50 nm) in the membrane. Evaporation of gold on the membrane fills the pore producing a bowl-shaped electrode. Immersion of the substrate in a solution of the thiol **15** results in the self-assembly of a molecular layer on the narrow part of the bowl-shaped electrode. The subsequent evaporation of a gold film on the organic monolayer produces an electrode/monolayer/electrode junction (Fig. 2.15a) with a contact area of less than 2,000 nm^2 and approximately 1,000 molecules.

Under the influence of voltage pulses applied to one of the two gold electrodes in Fig. 2.15a, the conductivity of the sandwiched monolayer switches reversibly between low and high values [2.48]. In the initial state, the monolayer is in a low conducting mode. A current output of only 30 pA is detected, when a probing voltage of $+0.25$ V is applied to the bowl-shaped electrode. If the same electrode is stimulated with a short voltage pulse of $+5$ V, the monolayer switches to a high conducting mode. Now a current output of 150 pA is measured at the same probing voltage of $+0.25$ V. Repeated probing of the current output at various intervals of time indicates that the high conducting state is memorized by the molecule-based device, and it is retained for more than 15 min. The low conducting mode is restored after either a relatively long period of time or the stimulation of the bowl-shaped electrode with a reverse voltage pulse of -5 V. Thus the current output switches from a low to a high value, if a high voltage input is applied. It switches from a high to a low value, under the influence of a low voltage pulse. This behavior offers the opportunity to store and erase binary data in analogy to a conventional random access memory [2.17]. Binary digits can be encoded on the current output of the molecule-based device applying a positive logic convention (low $=0$, high $=1$). It follows that a binary 1 can be stored in the molecule-based device applying a high voltage input, and it can be erased applying a low voltage input [2.48].

The electrode/monolayer/electrode junctions in Fig. 2.14 and in Fig. 2.15a are two-terminal devices. A voltage input is applied to one terminal, and a current output is measured at the other. The implementation of molecule-based transistors, however, requires three terminals. A third electrode is needed to gate the current flowing between the source and drain terminals. These types of devices are much more versatile than their two-terminal counterparts for the implementation of logic functions. But the incorporation of molecular components in a three-terminal configuration is considerably more challenging. A remarkable example is the composite assembly in Fig. 2.15b [2.49]. This field-effect transistor integrates a monolayer of the bisthiol **16** and three electrodes. Conventional lithography and anisotropic etching are employed to define a vertical step in a doped silicon substrate. The resulting structure is covered with an insulating layer of silicon dioxide with a thickness of ca. 30 nm. Then gold is evaporated on the bottom part of the step to allow the self-assembly the bisthiol **16**. Under the influence of the gate voltage, the molecular components embedded in this device modulate the current flowing between the source and drain electrodes. The intensity of the drain current increases by five orders of magnitude if the gate voltage is lowered below -0.2 V. It is important to note, however, that only the molecules within 5 nm from the monolayer/insulator interface are responsible for conduction. Thus the current/voltage behavior of this three-terminal device is determined only by few thousand molecules.

Conventional transistors are the basic building blocks of digital circuits [2.19]. They are routinely interconnected with specific configurations to execute designed logic functions. The molecule-based transistor in Fig. 2.15b is no exception. It can be used to assemble real electronic circuits incorporating molecular components. The connection of the drain terminals of two molecule-based transistors affords an inverter [2.49]. In this basic circuit, an output voltage switches from a high (0 V) to a low value (-2 V) when an input voltage varies from a low (-2 V) to a high value (0 V) and vice versa. This signal transduction behavior translates into the truth table of a NOT gate, if a positive logic convention (low $=0$, high $=1$) is applied to the input and output voltages.

The ability of thiols to self-assemble on the surface of gold can be exploited to fabricate nanocomposite materials integrating organic and inorganic components. For example, the bisthiol **17** forms monolayers (Fig. 2.16a) on gold electrodes with surface coverages of ca. 4.1×10^{10} mol cm^2 [2.50, 51]. The formation of a thiolate–gold bond at one of the two thiol ends of **17** is responsible for adsorption. The remaining thiol group points away from the supporting surface and can be exploited for further functionalization. Gold nanoparticles adsorb on the molecular layer (Fig. 2.16b), once again, as a result of thiolate–gold bond formation. The immersion of the resulting material in a methanol solution of **17** encourages the adsorption of an additional organic layer (Fig. 2.16c) on the composite material. Following these procedures, up to ten alternating organic and inorganic layers can be deposited on the electrode surface. The resulting assembly can mediate the unidirectional electron transfer from the supporting electrode to redox active species in solution. For example, the cyclic voltammogram of the $[\mathrm{Ru(NH_3)_6}]^{3+/2+}$ couple recorded with a bare gold electrode reveals a reversible reduction process. In the presence of ten alternating molecular and nanoparticle layers on the electrode surface, the reduction potential shifts by ca. -0.2 V and the back oxidation wave disappears. The pronounced potential shift indicates that $[\mathrm{Ru(NH_3)_6}]^{3+}$ accepts electrons only after the surface-confined bipyridinium dications have been reduced. The lack of reversibility indicates that the back oxidation to the bipyridinium dications inhibits the transfer of electrons from the $[Ru(NH_3)_6]^{2+}$ to the electrode. Thus the electroactive multilayer allows the flow of electrons in one direction only in analogy to conventional diodes.

The current/voltage behavior of individual nanoparticles in Fig. 2.16b can be probed by scanning tunneling spectroscopy in an aqueous electrolyte under an inert atmosphere [2.52]. The platinum-iridium tip of a scanning tunneling microscope is positioned above one of the gold particles. The voltage of the gold substrate relative to the tip is maintained at -0.2 V while that relative to a reference electrode immersed in the same electrolyte is varied to control the redox state of the electroactive units. Indeed, the bipyridinium dications in the molecular layer can be reduced reversibly to a monocationic state. The resulting monocations can be reduced further and, once again, reversibly to a neutral form. Finally, the current flowing from the gold support to the tip of the scanning tunneling microscope is monitored as the tip–particle distance increases. From the distance dependence of the current, inverse length decays of ca. 16 and 7 nm^{-1} for the dicationic and monocationic states, respectively, of the molecular spacer can be determined. The dramatic decrease indicates that the reduction of the electroactive unit facilitates the tunneling of electrons through the gold/molecule/nanoparticle/tip junction. In summary, a change in the redox state of the bipyridinium components can be exploited to gate reversibly the current flowing through this nanoscaled device.

Similar nanostructured materials, combining molecular and nanoparticles layers, can be prepared on layers on indium-tin oxide electrodes following multistep procedures [2.53]. The hydroxylated surfaces of indium-tin oxide supports can be functionalized with 3-ammoniumpropylysilyl groups and then exposed to gold nanoparticles having a diameter of ca. 13 nm [2.54, 55]. Electrostatic interactions promote the adsorption of the nanoparticles on the organic layer (Fig. 2.17a). The treatment of the composite film with the bipyridinium cyclophane **18** produces an organic layer on the gold nanoparticles (Fig. 2.18b). Following this approach, alternating layers of inorganic nanoparticles and organic building blocks can be assembled on the indium-tin oxide support. Cyclic voltammograms of the resulting materials show the oxidation of the gold nanoparticles and the reduction of the bipyridinium units. The peak current for both processes increases with the number of alternating layers. Comparison of these values indicates that the ratio between the number of tetracationic cyclophanes and that of the nanoparticles is ca. 100 : 1.

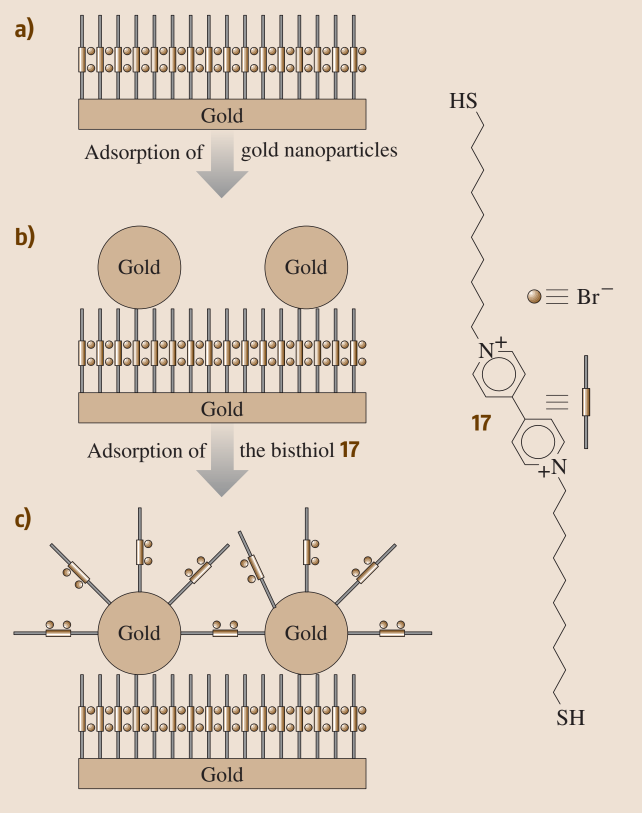

Fig. 2.16 **(a)** The bisthiol **18** self-assembles on gold electrodes as a result of thiolate–gold bond formation. **(b)** Gold nanoparticles adsorb spontaneously on the molecular layer. **(c)** Exposure of the composite assembly to a solution of **18** results in the formation of an additional molecular layer on the surface of the gold nanoparticles

The tetracationic cyclophane **18** binds dioxyarenes in solution [2.56, 57]. Attractive supramolecular forces between the electron deficient bipyridinium units and the electron rich guests are responsible for complexation. This recognition motif can be exploited to probe the ability of the composite films in Fig. 2.17b,c to sense electron rich analytes. In particular, hydroquinone is expected to enter the electron deficient cavities of the surface-confined cyclophanes. Cyclic voltammograms consistently reveal the redox waves associated with the reversible oxidation of hydroquinone even when very small amounts of the guest (ca. 1×10^5 M) are added to the electrolyte solution [2.54, 55]. No redox response can be detected with a bare indium–tin oxide electrode under otherwise identical conditions. The supramolecular association of the guest and the surface confined

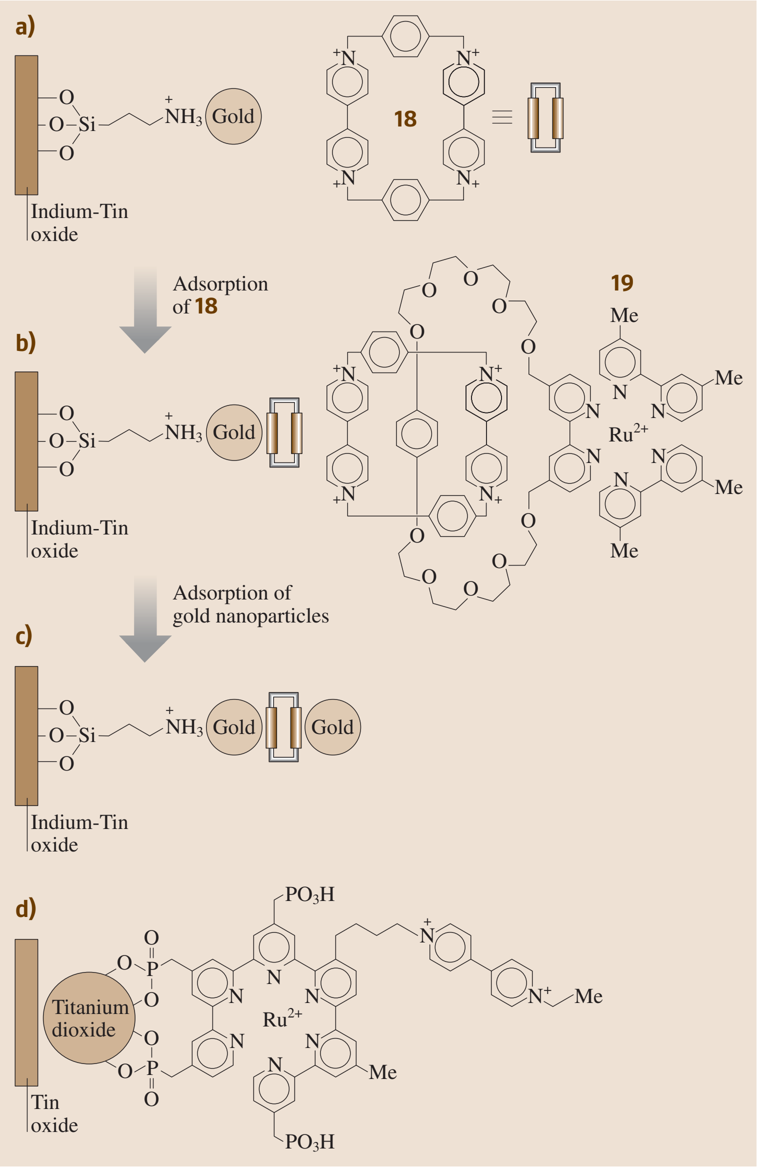
a)
O
O–Si
O
NH3
Gold
Indium-Tin oxide
18
Adsorption of 18
b)
Gold
Indium-Tin oxide
19
Me
Ru2+
Adsorption of gold nanoparticles
c)
Gold
Gold
Indium-Tin oxide
d)
PO3H
Titanium dioxide
Tin oxide
N+
Me

Fig. 2.17 **(a)** Gold nanoparticles assemble spontaneously on prefunctionalized indium-tin oxide electrodes. **(b)** Electrostatic interactions encourage the adsorption of the tetracationic cyclophane **18** on the surface-confined nanoparticles. **(c)** An additional layer of nanoparticles assembles on the cationic organic coating. Similar composite films can be prepared using the tetracationic [2]catenane **19** instead of the cyclophane **18**. **(d)** Phosphonate groups can be used to anchor molecular building blocks to titanium dioxide nanoparticles

cyclophanes increases the local concentration of hydroquinone at the electrode/solution interface enabling its electrochemical detection.

Following a related strategy, the [2]catenane **19** (Fig. 2.17) can be incorporated into similar composite arrays [2.58,59]. This interlocked molecule incorporates a Ru(II)/trisbipyridine sensitizer and two bipyridinium acceptors. Upon irradiation of the composite material at 440 nm, photoinduced electron transfer from the sensitizer to the appended acceptors occurs. The photogenerated hole in the sensitizer is filled after the transfer of an electron from a sacrificial electron donor present in the electrolyte solution. Under a positive voltage bias applied to the supporting electrode, an electron flow from the bipyridinium acceptors to the indium-tin oxide support is established. The resulting current switches between high and low values as the light source is turned on and off.

Another photoresponsive device, assembled combining inorganic nanoparticles with molecular building blocks, is illustrated in Fig. 2.17d. Phosphonate groups can be used to anchor a Ru(II)/trisbipyridine complex with an appended bipyridinium dication to titanium dioxide nanoparticles deposited on a doped tin oxide electrode [2.60, 61]. The resulting composite array can be integrated in a conventional electrochemical cell filled with an aqueous electrolyte containing triethanolamine. Under a bias voltage of -0.45 V and irradiation at 532 nm, 95% of the excited ruthenium centers transfer electrons to the titanium dioxide nanoparticles. The other 5% donate electrons to the bipyridinium dications. All the electrons transferred to the bipyridinium acceptors return to the ruthenium centers, while only 80% of those accepted by the nanoparticles return to the transition metal complexes. The remaining 15% reach the bipyridinium acceptors, while electron transfer from sacrificial triethanolamine donors fills the photogenerated holes left in the ruthenium sensitizers. The photoinduced reduction of the bipyridinium dication is accompanied by the appearance of the characteristic band of the radical cation in the absorption spectrum. This band persists for hours under open circuit conditions. But it fades in ca. 15 s under a voltage bias of $+1$ V, as the radical cation is oxidized back to the dicationic form. In summary, an optical stimulation accompanied by a negative voltage bias reduces the bipyridinium building block. The state of the photogenerated form can be read optically, recording the absorption spectrum in the visible region, and erased electrically, applying a positive voltage pulse.

2.3.4 Nanogaps and Nanowires

The operating principles of the electroactive and photoactive devices illustrated in Figs. 2.12–2.17 exploit the ability of small collections of molecular components to manipulate electrons and photons. Designed molecules are deposited on relatively large electrodes and can be addressed electrically and/or optically by controlling the voltage of the support and/or illuminating its surface. The transition from devices relying on collections of molecules to unimolecular devices requires the identification of practical methods to contact single molecules. This fascinating objective demands the rather challenging miniaturization of contacting electrodes to the nanoscale.

A promising approach to unimolecular devices relies on the fabrication of nanometer-sized gaps in metallic features followed by the insertion of individual molecules between the terminals of the gap. This strategy permits the assembly of nanoscaled three-terminal devices equivalent to conventional transistors [2.62–64]. A remarkable example is illustrated in Fig. 2.18a [2.62]. It incorporates a single molecule in the nanogap generated between two gold electrodes. Initially electron beam lithography is used to pattern a gold wire on a doped silicon wafer covered by an insulating silicon dioxide layer. Then the gold feature is broken by electromigration to generate the nanogap. The lateral size of the separated electrodes is ca. 100 nm and their thickness is ca. 15 nm. Scanning electron microcopy indicates that the facing surfaces of the separated electrodes are not uniform and that tiny gaps between their protrusions are formed. Current/voltage measurements suggest that the size of the smallest nanogap is ca. 1 nm. When the

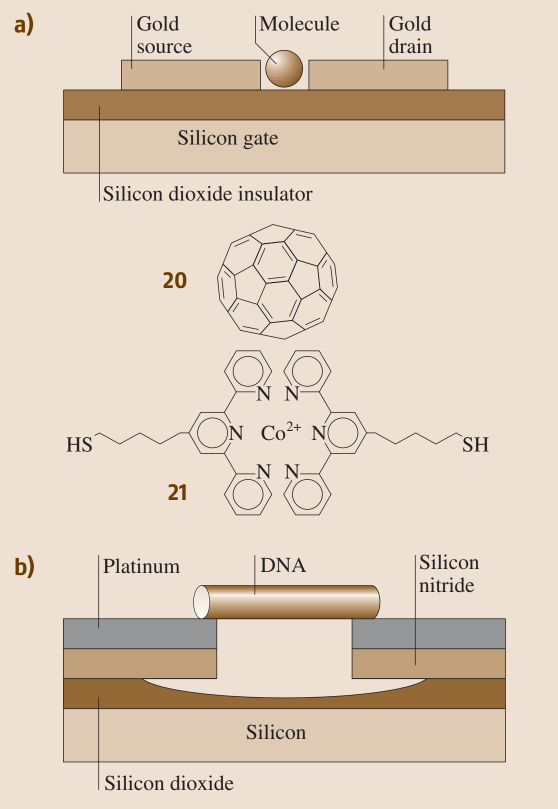

Fig. 2.18 **(a)** Nanoscaled transistors can be fabricated inserting a single molecule (**20** or **21**) between source and drain electrodes mounted on a silicon/silicon dioxide support. **(b)** A DNA nanowire can bridge nanoelectrodes suspended above a silicon dioxide support

breakage of the gold feature is preceded by the deposition of a dilute toluene solution of C_{60} (**20**), junctions with enhanced conduction are obtained. This particular molecule has a diameter of ca. 0.7 nm and can insert in the nanogap facilitating the flow of electrons across the junction.

The unique configuration of the molecule-based device in Fig. 2.18a can reproduce the functions of a conventional transistor [2.19] at the nanoscale. The two gold terminals of the junction are the drain and source of this nanotransistor, and the underlying silicon wafer is the gate. At a temperature of 1.5 K, the junction conductance is very small, when the gate bias is low, and increases in steps at higher voltages [2.62]. The conductance gap is a consequence of the finite energy required to oxidize/reduce the single C_{60} positioned in the junction. It is interesting that the zero-conductance window also changes with the gate voltage and can be opened and closed reversibly adjusting the gate bias.

A similar strategy can be employed to fabricate a nanoscaled transistor incorporating the Co(II) complex **21** shown in Fig. 2.18 [2.64]. In this instance, a silicon dioxide layer with a thickness of ca. 30 nm is grown thermally on a doped silicon substrate. Then a gold wire with a width of ca. 200 nm and a thicknesses of ca. 10–15 nm is patterned on the silicon dioxide overlayer by electron beam lithography. After extensive washing of the substrate with acetone and methylene chloride and cleaning with oxygen plasma, the gold wire is exposed to a solution of the bisthiol **21**. The formation of thiolate–gold bonds promotes the self-assembly of the molecular building block on the gold surface. At this point, electromigration-induced breakage produces a gap of 1–2 nm in the gold wire. The surface-confined bisthiol **21** is only 0.24 nm long and, therefore, it can insert in the nanogap producing an electrode/molecule/electrode junction.

The cobalt center in **21** can be oxidized/reduced reversibly between Co(II) and Co(III) [2.64]. When this electroactive molecule is inserted in a nanogap (Fig. 2.18a), its ability to accept and donate electrodes dictates the current/voltage profile of the resulting electrode/molecule/electrode junction. More precisely, no current flows across the junction below a certain voltage threshold. As the source voltage is raised above this particular value, the drain current increases in steps. The threshold associated with the source voltage varies in magnitude with the gate voltage. This intriguing behavior is a consequence of the finite energy necessary to oxidize/reduce the cobalt center and of a change in the relative stabilities of the oxidized and reduced forms Co(II) and Co(III) with the gate voltage. In summary, the conduction of the electrode/molecule/electrode junction can be tuned adjusting the voltage of the silicon support. The behavior of this molecule-based nanoelectronic device is equivalent to that of a conventional transistor [2.19]. In both instances, the gate voltage regulates the current flowing from the source to the drain.

The electromigration-induced breakage of preformed metallic features successfully produces nanogaps by moving apart two fragments of the same wire. Alternatively, nanogaps can be fabricated reducing the separation of the two terminals of much larger gaps. For example, gold electrodes separated by a distance of 20–80 nm can be patterned on a silicon/silicon dioxide substrate by electron beam lithography [2.65]. The relatively large gap between them can be reduced significantly by the electrochemical deposition of gold on the surfaces of both electrodes. The final result is the fabrication of two nanoelectrodes separated by ca. 1 nm and with a radius of curvature of 5–15 nm. The two terminals of this nanogap can be "contacted" by organic nanowires grown between them [2.66]. In particular, the electropolymerization of aniline produces polyaniline bridges between the gold nanoelectrodes. The conductance of the resulting junction can be probed immersing the overall assembly in an electrolyte solution. Employing a bipotentiostat, the bias voltage of the two terminals of the junction can be maintained at 20 mV, while their potentials are scanned relative to that of a silver/silver chloride reference electrode. Below ca. 0.15 V, the polymer wire is in an insulating state and the current flowing across the junction is less than 0.05 nA. At this voltage threshold, however, the current raises abruptly to ca. 30 nA. This value corresponds to a conductivity for the polymer nanojunction of 10–100 S cm^{-1}. When the potential is lowered again below the threshold, the current returns back to very low values. The abrupt decrease in current in the backward scan is observed at a potential that is slightly more negative than that causing the abrupt current increase in the forward scan. In summary, the conductance of this nanoscaled junction switches on and off as a potential input is switched above and below a voltage threshold.

It is interesting to note that the influence of organic bridges on the junction conductance can be exploited for chemical sensing. Nanogaps fabricated following a similar strategy but lacking the polyaniline bridge alter their conduction after exposure to dilute solutions of small organic molecules [2.67]. Indeed, the organic analytes dock into the nanogaps producing a marked decrease in the junction conductance. The magnitude of the conductance drop happens to be proportional to the analyte–nanoelectrode binding strength. Thus the presence of the analyte in solution can be detected probing the current/voltage characteristics of the nanogaps.

Nanogaps between electrodes patterned on silicon/silicon dioxide supports can be bridged also by DNA double strands [2.68, 69]. The device in Fig. 2.18b has a 10.4 nm long poly(G)–poly(C) DNA oligomer suspended between two nanoelectrodes. It can be fabricated patterning a 30 nm wide slit in a silicon nitride overlayer covering a silicon/silicon dioxide support by electron beam evaporation. Underetching the silicon dioxide layer leaves a silicon nitride finger, which can be sputtered with a platinum layer and chopped to leave a nanogap of 8 nm. At this point, a microdroplet of a dilute solution of DNA is deposited on the device and a bias of 5 V is applied between the two electrodes. Electrostatic forces encourage the deposition of a single DNA wire on top of the nanogap. As soon as the nanowire is in position, current starts to flow across the junction. The current/voltage signature of the electrode/DNA/electrode junction shows currents below 1 pA at low voltage biases. Under these conditions, the DNA nanowire is an insulator. Above a certain voltage threshold, however, the nanowire becomes conducting and currents up to 100 nA can flow across the junction through a single nanowire. Assuming that direct tunneling from electrode to electrode is extremely unlikely for a relatively large gap of 8 nm, the intriguing current/voltage behavior has to be a consequence of the participation of the molecular states in the electron transport process. Two possible mechanisms can be envisaged. Sequential hopping of the electrons between states localized in the DNA base pairs can allow the current flow above a certain voltage threshold. But this mechanism would presumably result in a Coulomb blockade voltage gap that is not observed experimentally. More likely, electronic states delocalized across the entire length of the DNA nanowire are producing a molecular conduction band. The off-set between the molecular conduction band and the Fermi levels of the electrodes is responsible for the insulating behavior at low biases. Above a certain voltage threshold, the molecular band and one of the Fermi levels align facilitating the passage of electrons across the junction.

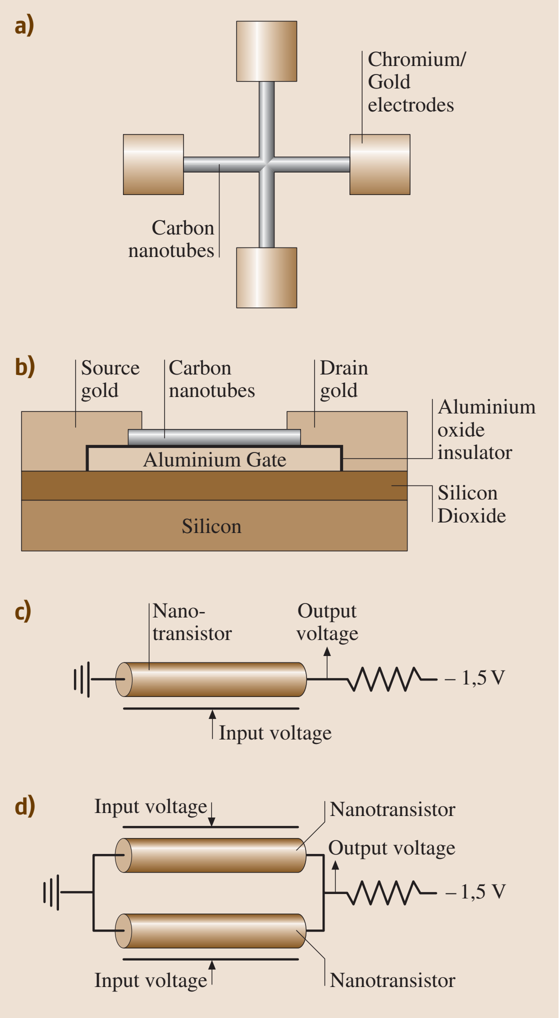

Fig. 2.19 **(a)** Nanoscaled junctions can be assembled on silicon/silicon dioxide supports crossing pairs of orthogonally arranged single-wall carbon nanotubes with chromium/gold electrical contacts at their ends. **(b)** Nanotransistors can be fabricated contacting the two ends of a single-wall carbon nanotube deposited on an aluminum/aluminum oxide gate with gold sources and drain. One or two nanotube transistors can be integrated into nanoscaled NOT **(c)** and NOR **(d)** logic gates

Carbon nanotubes are extremely versatile building blocks for the assembly of nanoscaled electronic devices. They can be used to bridge nanogaps [2.70–73] and assemble nanoscaled cross junctions [2.74–76]. In Fig. 2.19a, a single-wall carbon nanotube crosses over another one in an orthogonal arrangement [2.74]. Both nanotubes have electrical contacts at their ends.

The fabrication of this device involves three main steps. First, alignment marks for the electrodes are patterned on a silicon/silicon dioxide support by electron beam lithography. Then the substrate is exposed to a dichloromethane suspension of single-wall SWNT carbon nanotubes. After washing with isopropanol, crosses of carbon nanotubes in an appropriate alignment relative to the electrode marks are identified by tapping mode atomic force microscopy. Finally chromium/gold electrodes are fabricated on top of the nanotube ends, again, by electron beam lithography. The conductance of individual nanotubes can be probed by exploiting the two electric contacts at their ends. These two-terminal measurements reveal that certain nanotubes have metallic behavior, while others are semiconducting. It follows that three distinct types of cross junctions differing in the nature of their constituent nanotubes can be identified on the silicon/silicon dioxide support. Four terminal current/voltage measurements indicate that junctions formed by two metallic nanotubes have high conductance and ohmic behavior. Similarly, high junction conductance and ohmic behavior is observed when two semiconducting nanotubes cross. The current/voltage signature of junctions formed when a metallic nanotube crosses a semiconducting one are, instead, completely different. The metallic nanotube depletes the semiconducting one at the junction region producing a nanoscaled Schottky barrier with a pronounced rectifying behavior.

Similar fabrication strategies can be exploited to assemble nanoscaled counterparts of conventional transistors. The device in Fig. 2.19b is assembled patterning an aluminum finger on a silicon/silicon dioxide substrate by electron beam lithography [2.76]. After exposure to air, an insulating aluminum oxide layer forms on the aluminum finger. Then a dichloromethane suspension of single-wall carbon nanotubes is deposited on the resulting substrate. Atomic force microscopy can be used to select carbon nanotubes with a diameter of ca. 1 nm positioned on the aluminum finger. After registering their coordinates relative to alignment markers, gold contacts can be evaporated on their ends by electron beam lithography. The final assembly is a nanoscaled three-terminal device equivalent to a conventional field effect transistor [2.19]. The two gold contacts are the source and drain terminals, while the underlying aluminum finger reproduces the function of the gate. At a source to drain bias of ca. -1.3 V, the drain current jumps from ca. 0 to ca. 50 nA when the gate voltage is lowered from -1.0 to -1.3 V. Thus moderate changes in the gate voltage vary significantly the current flowing through the nanotube-based device in analogy to a conventional enhancement mode *p*-type field effect transistor [2.19].

The nanoscaled transistor in Fig. 2.18a has a microscaled silicon gate that extends under the entire chip [2.62, 64]. The configuration in Fig. 2.19b, instead, has nanoscaled aluminum gates for every single carbon nanotube transistor fabricated on the same support [2.76]. It follows that multiple nanoscaled transistors can be fabricated on the same chip and operated independently following this strategy. This unique feature offers the possibility of fabricating nanoscaled digital circuits by interconnecting the terminals of independent nanotube transistors. The examples in Fig. 2.19c,d illustrate the configurations of nanoscaled NOT and NOR gates implemented using one or two nanotube transistors. In Fig. 2.19c, an off-chip bias resistor is connected to the drain terminal of a single transistor while the source is grounded. A voltage input applied to the gate modulates the nanotube conductance altering the voltage output probed at the drain terminal. In particular, a voltage input of -1.5 V lowers the nanotube resistance (26 MΩ) below that of the bias resistor (100 MΩ). As a result, the voltage output drops to 0 V. When the voltage input is raised to 0 V, the nanotube resistance increases above that of the bias resistor and the voltage output becomes -1.5 V. Thus the output of this nanoelectronic device switches from a high (0 V) and to a low (-1.5 V) level as the input shifts from a low (-1.5 V) to a high (0 V) value. The inverse relation between input and output translates into a NOT operation if a negative logic convention (low = 1, high = 0) is applied to both signals.

In Fig. 2.15d, the source terminals of two independent nanotube transistors fabricated on the same chip are connected by a gold wire and grounded [2.76]. Similarly, the two drain terminals are connected by another gold wire and contacted to an off-chip bias resistors. The gate of each nanotube can be stimulated with a voltage input and the voltage output of the device can be probed at their interconnected drain terminals. When the resistance of at least one of the two nanotubes is below that of the resistor, the output is 0 V. When both nanotubes are in a nonconducting mode, the output voltage is -1.5 V. Thus if a low voltage input -1.5 V is applied to one or both transistors, the output is high (0 V). When both voltage inputs are high (0 V), the output is low (-1.5 V). If a negative logic convention (low = 1, high = 0) is applied to all signals, the signal transduction behavior translates in to a NOR operation.

2.4 Conclusions and Outlook

Nature builds nanostructured biomolecules relying on a highly modular approach [2.1]. Small building blocks are connected by robust chemical bonds to generate long strands of repeating units. The synergism of a multitude of attractive supramolecular forces determines the three-dimensional arrangement of the resulting polymeric chains and controls the association of independent strands into single and well-defined entities. Nucleic acids and proteins are two representative classes of biomolecules assembled with subnanometer precision through the subtle interplay of covalent and noncovalent bonds starting from a relatively small pool of nucleotide and amino acid building blocks.

The power of chemical synthesis [2.2] offers the opportunity of mimicking nature's modular approach to nanostructured materials. Following established experimental protocols, small molecular building blocks can be joined together relying on the controlled formation of covalent bonds between designed functional groups. Thus artificial molecules with nanoscaled dimensions can be assembled piece by piece with high structural control. Indeed, helical, tubular, interlocked, and highly branched nanostructures have been all prepared already exploiting this general strategy and the synergism of covalent and noncovalent bonds [2.3].

The chemical construction of nanoscaled molecules from modular building blocks also offers the opportunity for engineering specific properties in the resulting assemblies. In particular, electroactive and photoactive fragments can be integrated into single molecules. The ability of these functional subunits to accept/donate electrons and photons can be exploited to design nanoscaled electronic and photonic devices. Indeed, molecules that respond to electrical and optical stimulations producing detectable outputs have been designed already [2.16]. These chemical systems can be employed to control the interplay of input and output signals at the molecular level. Their conceptual analogy with the signal transduction operated by conventional logic gates in digital circuits is evident. In fact, electroactive and photoactive molecules able to reproduce AND, NOT, and OR operations as well as simple combinational of these basic logic functions are already a reality [2.20, 21].

Most of the molecular switches for digital processing developed so far rely on bulk addressing. In general, relatively large collections of functional molecules are addressed simultaneously in solution. The realization of molecule-based devices with reduced dimensions as well as practical limitations associated with liquid phases in potential applications are encouraging a transition from the solution to the solid state. The general strategy followed so far relies on the deposition of functional molecules on the surfaces of appropriate electrodes following either the Langmuir–Blodgett methodology [2.34] or self-assembly processes [2.35]. The combination of these techniques with the nanofabrication of insulating, metallic, and semiconducting features on appropriate supports has already allowed the realization of fascinating molecule-based devices [2.30–33, 53]. The resulting assemblies integrate inorganic and organic components and, in some instances, even biomolecules to execute specific functions. They can convert optical stimulations into electrical signals. They can execute irreversible and reversible switching operations. They can sense qualitatively and quantitatively specific analytes. They can reproduce the functions of conventional rectifiers and transistors. They can be integrated within functioning nanoelectronic devices capable of simple logic operations.

The remarkable examples of molecule-based materials and devices now available demonstrate the great potential and promise for this research area. At this stage, the only limit left to the design of functional molecules is the imagination of the synthetic chemist. All sort of molecular building blocks with tailored dimensions, shapes, and properties are more or less accessible with the assistance of modern chemical synthesis. Now, the major challenges are (1) to master the operating principles of the molecule-based devices that have been and continue to be assembled and (2) to expand and improve the fabrication strategies available to incorporate molecules into reliable device architectures. As we continue to gather further insights in these directions, design criteria for a wide diversity of molecule-based devices will emerge. It is not unrealistic to foresee the evolution of an entire generation of nanoscaled devices, based on engineered molecular components, that will find applications in a variety of fields ranging from biomedical research to information technology. Perhaps nature can once again illuminate our path, teaching us not only how to synthesize nanostructured molecules but also how to use them. After all, nature is replete with examples of extremely sophisticated molecule-based devices. From tiny bacteria to higher animals, we are all a collection of molecule-based devices.

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

References

2.1 D. Voet, J. G. Voet: *Biochemistry* (Wiley, New York 1995)

2.2 K. C. Nicolau, E. C. Sorensen: *Classics in Total Synthesis* (VCH, Weinheim 1996)

2.3 J.-M. Lehn: *Supramolecular Chemistry: Concepts and Perspectives* (VCH, Weinheim 1995)

2.4 M. M. Harding, U. Koert, J.-M. Lehn, A. Marquis-Rigault, C. Piguet, J. Siegel: Synthesis of unsubstituted and 4,4'-substituted oligobipyridines as ligand strands for helicate self-assembly, Helv. Chim. Acta **74** (1991) 594–610

2.5 J.-M. Lehn, A. Rigault, J. Siegel, B. Harrowfield, B. Chevrier, D. Moras: Spontaneous assembly of double-stranded helicates from oligobipyridine ligands and copper(I) cations: Structure of an inorganic double helix, Proc. Natl. Acad. Sci. USA **84** (1987) 2565–2569

2.6 J.-M. Lehn, A. Rigault: Helicates: Tetra- and pentanuclear double helix complexes of Cu(I) and poly(bipyridine) strands, Angew. Chem. Int. Ed. Engl. **27** (1988) 1095–1097

2.7 J. D. Hartgerink, J. R. Granja, R. A. Milligan, M. R. Ghadiri: Self-assembling peptide nanotubes, J. Am. Chem. Soc. **118** (1996) 43–50

2.8 M. R. Ghadiri, J. R. Granja, R. A. Milligan, D. E. McRee, N. Khazanovich: Self-assembling organic nanotubes based on a cyclic peptide architecture, Nature **366** (1993) 324–327

2.9 V. Balzani, A. Credi, F. M. Raymo, J. F. Stoddart: Artificial molecular machines, Angew. Chem. Int. Ed. **39** (2000) 3348–3391

2.10 A. J. Bard, L. R. Faulkner: *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York 2000)

2.11 V. Balzani (Ed.): *Electron Transfer in Chemistry* (Wiley-VCH, Weinheim 2001)

2.12 J. D. Coyle: *Principles and Applications of Photochemistry* (Wiley, New York 1988)

2.13 J. F. Stoddart (Ed.): Molecular machines, Acc. Chem. Res. **34** (2001) 409–522

2.14 P. R. Ashton, R. Ballardini, V. Balzani, A. Credi, K. R. Dress, E. Ishow, C. J. Kleverlaan, O. Kocian, J. A. Preece, N. Spencer, J. F. Stoddart, M. Venturi, S. Wenger: A photochemically driven molecular-level abacus, Chem. Eur. J. **6** (2000) 3558–3574

2.15 B. L Feringa (Ed.): Photochromism: memories and switches, Chem. Rev. **100** (2000) 1683–1890

2.16 B. L. Feringa (Ed.): *Molecular Switches* (Wiley-VCH, Weinheim 2001)

2.17 R. J. Mitchell: *Microprocessor Systems: An Introduction* (Macmillan, London 1995)

2.18 D. R. Smith: *Digital Transmission Systems* (Van Nostrand Reinhold, New York 1993)

2.19 S. Madhu: *Electronics: Circuits and Systems* (SAMS, Indianapolis 1985)

2.20 A. P. de Silva, N. D. McClenaghan, C. P. McCoy: Logic Gates. In: *Electron Transfer in Chemistry*, ed. by V. Balzani (Wiley-VCH, Weinheim 2001) pp. 156–185

2.21 F. M. Raymo: Digital processing and communication with molecular switches, Adv. Mater. **14** (2002) 401–414

2.22 A. Aviram: Molecules for memory, logic and amplification, J. Am. Chem. Soc. **110** (1988) 5687–5692

2.23 A. P. de Silva, H. Q. N. Gunaratne, C. P. McCoy: A molecular photoionic AND gate based on fluorescent signaling, Nature **364** (1993) 42–44

2.24 M. Asakawa, P. R. Ashton, V. Balzani, A. Credi, G. Mattersteig, O. A. Matthews, M. Montalti, N. Spencer, J. F. Stoddart, M. Venturi: Electrochemically induced molecular motions in pseudorotaxanes: A case of dual-mode (oxidative and reductive) dethreading, Chem. Eur. J. **3** (1997) 1992–1996

2.25 F. M. Raymo, S. Giordani: Signal processing at the molecular level, J. Am. Chem. Soc. **123** (2001) 4651–4652

2.26 F. M. Raymo, S. Giordani: Signal communication between molecular switches, Org. Lett. **3** (2001) 3475–3478

2.27 F. M. Raymo, S. Giordani: Digital communication through intermolecular fluorescence modulation, Org. Lett. **3** (2001) 1833–1836

2.28 F. M. Raymo, S. Giordani: All-optical processing with molecular switches, Proc. Natl. Acad. Sci. USA **99** (2002) 4941–4944

2.29 A. J. Bard: *Integrated Chemical Systems: A Chemical Approach to Nanotechnology* (Wiley, New York 1994)

2.30 R. M. Metzger: Electrical rectification by a molecule: The advent of unimolecular electronic devices, Acc. Chem. Res. **32** (1999) 950–957

2.31 C. Joachim, J. K. Gimzewski, A. Aviram: Electronics using hybrid-molecular and mono-molecular devices, Nature **408** (2000) 541–548

2.32 J. M. Tour: Molecular electronics. Synthesis and testing of components, Acc. Chem. Res. **33** (2000) 791–804

2.33 A. R. Pease, J. O. Jeppesen, J. F. Stoddart, Y. Luo, C. P. Collier, J. R. Heath: Switching devices based on interlocked molecules, Acc. Chem. Res. **34** (2001) 433–444

2.34 M. C. Petty: *Langmuir–Blodgett Films: An Introduction* (Cambridge Univ. Press, Cambridge 1996)

2.35 A. Ulman: *An Introduction to Ultrathin Organic Films* (Academic, Boston 1991)

2.36 C. Lee, A. J. Bard: Comparative electrochemical studies of *N*-Methyl-*N'*-hexadecyl viologen monomolecular films formed by irreversible adsorption and the Langmuir–Blodgett method, J. Electroanal. Chem. **239** (1988) 441–446

2.37 C. Lee, A. J. Bard: Cyclic voltammetry and Langmuir film isotherms of mixed monolayers of *N*-docosoyl-

N′-methyl viologen with arachidic acid, Chem. Phys. Lett. **170** (1990) 57–60

2.38 M. Fujihira, K. Nishiyama, H. Yamada: Photoelectrochemical responses of optically transparent electrodes modified with Langmuir–Blodgett films consisting of surfactant derivatives of electron donor, acceptor and sensitizer molecules, Thin Solid Films **132** (1985) 77–82

2.39 M. Fujihira: Photoelectric conversion with Langmuir–Blodgett films. In: *Nanostructures Based on Molecular Materials*, ed. by W. Göpel, C. Ziegler (VCH, Weinheim 1992) pp. 27–46

2.40 C. P. Collier, E. W. Wong, M. Belohradsky, F. M. Raymo, J. F. Stoddart, P. J. Kuekes, R. S. Williams, J. R. Heath: Electronically configurable molecular-based logic gates, Science **285** (1999) 391–394

2.41 E. W. Wong, C. P. Collier, M. Belohradsky, F. M. Raymo, J. F. Stoddart, J. R. Heath: Fabrication and transport properties of single-molecule-thick electrochemical junctions, J. Am. Chem. Soc. **122** (2000) 5831–5840

2.42 M. Asakawa, P. R. Ashton, V. Balzani, A. Credi, C. Hamers, G. Mattersteig, M. Montalti, A. N. Shipway, N. Spencer, J. F. Stoddart, M. S. Tolley, M. Venturi, A. J. P. White, D. J. Williams: A chemically and electrochemically switchable [2]catenane incorporating a tetrathiafulvalene unit, Angew. Chem. Int. Ed. **37** (1998) 333–337

2.43 V. Balzani, A. Credi, G. Mattersteig, O. A. Matthews, F. M. Raymo, J. F. Stoddart, M. Venturi, A. J. P. White, D. J. Williams: Switching of pseudorotaxanes and catenanes incorporating a tetrathiafulvalene unit by redox and chemical inputs, J. Org. Chem. **65** (2000) 1924–1936

2.44 M. Asakawa, M. Higuchi, G. Mattersteig, T. Nakamura, A. R. Pease, F. M. Raymo, T. Shimizu, J. F. Stoddart: Current/Voltage characteristics of monolayers of redox-switchable [2]catenanes on gold, Adv. Mater. **12** (2000) 1099–1102

2.45 C. P. Collier, G. Mattersteig, E. W. Wong, Y. Luo, K. Beverly, J. Sampaio, F. M. Raymo, J. F. Stoddart, J. R. Heath: A [2]catenane based solid-state electronically reconfigurable switch, Science **289** (2000) 1172–1175

2.46 C. P. Collier, J. O. Jeppesen, Y. Luo, J. Perkins, E. W. Wong, J. R. Heath, J. F. Stoddart: Molecular-based electronically switchable tunnel junction devices, J. Am. Chem. Soc. **123** (2001) 12632–12641

2.47 J. Chen, M. A. Reed, A. M. Rawlett, J. M. Tour: Large on-off ratios and negative differential resistance in a molecular electronic device, Science **286** (1999) 1550–1552

2.48 M. A. Reed, J. Chen, A. M. Rawlett, D. W. Price, J. M. Tour: Molecular random access memory cell, Appl. Phys. Lett. **78** (2001) 3735–3737

2.49 J. H. Schön, H. Meng, Z. Bao: Self-assembled monolayer organic field-effect transistors, Nature **413** (2001) 713–716

2.50 D. I. Gittins, D. Bethell, R. J. Nichols, D. J. Schiffrin: Redox-controlled multilayers of discrete gold particles: A novel electroactive nanomaterial, Adv. Mater. **9** (1999) 737–740

2.51 D. I. Gittins, D. Bethell, R. J. Nichols, D. J. Schiffrin: Diode-like electron transfer across nanostructured films containing a redox ligand, J. Mater. Chem. **10** (2000) 79–83

2.52 D. I. Gittins, D. Bethell, D. J. Schiffrin, R. J. Nichols: A nanometer-scale electronic switch consisting of a metal cluster and redox-addressable groups, Nature **408** (2000) 67–69

2.53 A. N. Shipway, M. Lahav, I. Willner: Nanostructured gold colloid electrodes, Adv. Mater. **12** (2000) 993–998

2.54 A. N. Shipway, M. Lahav, R. Blonder, I. Willner: Bis-bipyridinium cyclophane receptor–Au nanoparticle superstructure for electrochemical sensing applications, Chem. Mater. **11** (1999) 13–15

2.55 M. Lahav, A. N. Shipway, I. Willner, M. B. Nielsen, J. F. Stoddart: An enlarged bis-bipyridinum cyclophane–Au nanoparticle superstructure for selective electrochemical sensing applications, J. Electroanal. Chem. **482** (2000) 217–221

2.56 R. E. Gillard, F. M. Raymo, J. F. Stoddart: Controlling self-assembly, Chem. Eur. J. **3** (1997) 1933–1940

2.57 F. M. Raymo, J. F. Stoddart: From supramolecular complexes to interlocked molecular compounds, Chemtracts – Organic Chemistry **11** (1998) 491–511

2.58 M. Lahav, T. Gabriel, A. N. Shipway, I. Willner: Assembly of a Zn(II)-porphyrin-bipyridinium dyad and Au-nanoparticle superstructures on conductive surfaces, J. Am. Chem. Soc. **121** (1999) 258–259

2.59 M. Lahav, V. Heleg-Shabtai, J. Wasserman, E. Katz, I. Willner, H. Durr, Y. Hu, S. H. Bossmann: Photoelectrochemistry with integrated photosensitizer-electron acceptor Au-nanoparticle arrays, J. Am. Chem. Soc. **122** (2000) 11480–11487

2.60 G. Will, S. N. Rao, D. Fitzmaurice: Heterosupramolecular optical write-read-erase device, J. Mater. Chem. **9** (1999) 2297–2299

2.61 A. Merrins, C. Kleverlann, G. Will, S. N. Rao, F. Scandola, D. Fitzmaurice: Time-resolved optical spectroscopy of heterosupramolecular assemblies based on nanostructured TiO_2 films modified by chemisorption of covalently linked ruthenium and viologen complex components, J. Phys. Chem. B **105** (2001) 2998–3004

2.62 H. Park, J. Park, A. K. L. Lim, E. H. Anderson, A. P. Alivisatos, P. L. McEuen: Nanomechanical oscillations in a single C_{60} transistor, Nature **407** (2000) 57–60

2.63 W. Liang, M. P. Shores, M. Bockrath, J. R. Long, H. Park: Kondo resonance in a single-molecule transistor, Nature **417** (2002) 725–729

2.64 J. Park, A. N. Pasupathy, J. I. Goldsmith, C. Chang, Y. Yaish, J. R. Petta, M. Rinkoski, J. P. Sethna,

H. D. Abruna, P. L. McEuen, D. C. Ralph: Coulomb blockade and the Kondo effect in single-atom transistors, Nature **417** (2002) 722–725

2.65 C. Z. Li, H. X. He, N. J. Tao: Quantized tunneling current in the metallic nanogaps formed by electrodeposition and etching, Appl. Phys. Lett. **77** (2000) 3995–3997

2.66 H. He, J. Zhu, N. J. Tao, L. A. Nagahara, I. Amlani, R. Tsui: A conducting polymer nanojunction switch, J. Am. Chem. Soc. **123** (2001) 7730–7731

2.67 A. Bogozi, O. Lam, H. He, C. Li, N. J. Tao, L. A. Nagahara, I. Amlani, R. Tsui: Molecular adsorption onto metallic quantum wires, J. Am. Chem. Soc. **123** (2001) 4585–4590

2.68 A. Bezryadin, C. N. Lau, M. Tinkham: Quantum suppression of superconductivity in ultrathin nanowires, Nature **404** (2000) 971–974

2.69 D. Porath, A. Bezryadin, S. de Vries, C. Dekker: Direct measurement of electrical transport through DNA molecules, Nature **403** (2000) 635–638

2.70 S. J. Tans, M. H. Devoret, H. Dai, A. Thess, E. E. Smalley, L. J. Geerligs, C. Dekker: Individual single-wall carbon nanotubes as quantum wires, Nature **386** (1997) 474–477

2.71 A. F. Morpurgo, J. Kong, C. M. Marcus, H. Dai: Gate-controlled superconducting proximity effect in carbon nanotubes, Nature **286** (1999) 263–265

2.72 J. Nygård, D. H. Cobden, P. E. Lindelof: Kondo physics in carbon nanotubes, Nature **408** (2000) 342–346

2.73 W. Liang, M. Bockrath, D. Bozovic, J. H. Hafner, M. Tinkham, H. Park: Fabry–Perot interference in a nanotube electron waveguide, Nature **411** (2001) 665–669

2.74 M. S. Fuhrer, J. Nygård, L. Shih, M. Forero, Y.-G. Yoon, M. S. C. Mazzoni, H. J. Choi, J. Ihm, S. G. Louie, A. Zettl, P. L. McEuen: Crossed nanotube junctions, Science **288** (2000) 494–497

2.75 T. Rueckes, K. Kim, E. Joselevich, G. Y. Tseng, C.-L. Cheung, C. M. Lieber: Carbon nanotube-based nonvolatile random access memory for molecular computing, Science **289** (2000) 94–97

2.76 A. Bachtold, P. Hadley, T. Nakanishi, C. Dekker: Logic circuits with carbon nanotube transistors, Science **294** (2001) 1317–1320

3. Introduction to Carbon Nanotubes

Carbon nanotubes are among the amazing objects that science sometimes creates by accident, without meaning to, but that will likely revolutionize the technological landscape of the century ahead. Our society stands to be significantly influenced by carbon nanotubes, shaped by nanotube applications in every aspect, just as silicon-based technology still shapes society today. The world already dreams of space-elevators tethered by the strongest of cables, hydrogen-powered vehicles, artificial muscles, and so on – feasts that would be made possible by the emerging carbon nanotube science.

Of course, nothing is set in stone yet. We are still at the stage of possibilities and potential. The recent example of fullerenes – molecules closely related to nanotubes, whose importance was so anticipated that their discovery in 1985 brought a Nobel Prize to their finders in 1996 although few related applications have actually yet reached the market – teaches us to play the game of enthusiastic predictions with some caution. But in the case of carbon nanotubes, expectations are high. Taking again the example of electronics, the miniaturization of chips is about to reach its lowest limits. Are we going to accept that our video camera, computers, and cellular phones no longer decrease in size and increase in memory every six months? Surely not. Always going deeper, farther, smaller, higher is a characteristic unique to humankind, and which helps to explain its domination of the living world on earth. Carbon nanotubes can help us fulfill our expectation of constant technological progress as a source of better living.

In this chapter, after the structure, synthesis methods, growth mechanisms, and properties of carbon nanotubes will be described, an entire section will be devoted to nanotube-related nano-objects. Indeed, should pristine nanotubes reach any limitation in some area, their ready and close association to foreign atoms, molecules, and compounds offers the prospect of an even magnified set of properties. Finally, we will describe carbon nanotube applications supporting the idea that the future for the science and technology of carbon nanotubes looks very promising.

3.1 **Structure of Carbon Nanotubes** 40
  3.1.1 Single-Wall Nanotubes 40
  3.1.2 Multiwall Nanotubes 43

3.2 **Synthesis of Carbon Nanotubes** 45
  3.2.1 Solid Carbon Source-Based Production Techniques for Carbon Nanotubes 45
  3.2.2 Gaseous Carbon Source-Based Production Techniques for Carbon Nanotubes 52
  3.2.3 Miscellaneous Techniques 57
  3.2.4 Synthesis of Aligned Carbon Nanotubes 58

3.3 **Growth Mechanisms of Carbon Nanotubes** 59
  3.3.1 Catalyst-Free Growth 59
  3.3.2 Catalytically Activated Growth........ 60

3.4 **Properties of Carbon Nanotubes**............ 63
  3.4.1 Variability of Carbon Nanotube Properties 63
  3.4.2 General Properties 63
  3.4.3 SWNT Adsorption Properties 63
  3.4.4 Transport Properties 65
  3.4.5 Mechanical Properties.................. 67
  3.4.6 Reactivity 67

3.5 **Carbon Nanotube-Based Nano-Objects**... 68
  3.5.1 Hetero-Nanotubes 68
  3.5.2 Hybrid Carbon Nanotubes............. 68
  3.5.3 Functionalized Nanotubes............ 71

3.6 **Applications of Carbon Nanotubes**.......... 73
  3.6.1 Current Applications 73
  3.6.2 Expected Applications Related to Adsorption 76

References 86

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Carbon nanotubes have been synthesized for a long time as products from the action of a catalyst over the gaseous species originating from the thermal decomposition of hydrocarbons (see Sect. 3.2). Some of the first evidence that the nanofilaments thus produced were actually nanotubes – i. e., exhibiting an inner cavity – can be found in the transmission electron microscope micrographs published by *Hillert* et al. in 1958 [3.1]. This was of course related to and made possible by the progress in transmission electron microscopy. It is then likely that the carbon filaments prepared by *Hughes* and *Chambers* in 1889 [3.2], reported by *Maruyama* et al. [3.3] as probably the first patent ever deposited in the field, and whose preparation method was also based on the catalytically enhanced thermal cracking of hydrocarbons, were already carbon nanotube-related morphologies. The preparation of vapor-grown carbon fibers had actually been reported as early as more than one century ago [3.4, 5]. Since then, the interest in carbon nanofilaments/nanotubes has been recurrent, though within a scientific area almost limited to the carbon material scientist community. The reader is invited to consult the review published by *Baker* et al. [3.6] regarding the early works. The worldwide enthusiasm came unexpectedly in 1991, after the catalyst-free formation of nearly perfect concentric multiwall carbon nanotubes (c-MWNTs, see Sect. 3.1) was reported [3.7] as by-products of the formation of fullerenes by the electric-arc technique. But the real breakthrough occurred two years later, when attempts in situ to fill the nanotubes with various metals (see Sect. 3.5) led to the discovery – again unexpected – of single-wall carbon nanotubes (SWNTs) simultaneously by *Iijima* et al. [3.8] and *Bethune* et al. [3.9]. Single-wall carbon nanotubes were really new nano-objects many of whose properties and behaviors are quite specific (see Sect. 3.4). They are also beautiful objects for fundamental physics as well as unique molecules for experimental chemistry, still keeping some mystery since their formation mechanisms are a subject of controversy and are still debated (see Sect. 3.3). Potential applications seem countless, although few have reached a marketable status so far (see Sect. 3.6). Consequently, about five papers a day with carbon nanotubes as the main topic are currently published by research teams from around the world, an illustration of how extraordinarily active – and highly competitive – is this field of research. It is a quite unusual situation, similar to that of fullerenes, which, by the way, are again carbon nano-objects structurally closely related to nanotubes.

This is not, however, only about scientific exaltation. Economical aspects are leading the game to a greater and greater extent. According to experts, the world market is predicted to be more than 430 M $ in 2004 and estimated to grow to several b $ before 2009. That is serious business, and it will be closely related to how scientists and engineers will be able to deal with the many challenges found on the path from the beautiful, ideal molecule to the reliable – and it is hoped, cheap – manufactured product.

3.1 Structure of Carbon Nanotubes

It is simple to imagine a single-wall carbon nanotube (SWNT). Ideally, it is enough to consider a perfect graphene sheet (a graphene being the same polyaromatic mono-atomic layer made of an hexagonal display of sp^2 hybridized carbon atoms that genuine graphite is built up with), to roll it into a cylinder (Fig. 3.1) paying attention that the hexagonal rings put in contact join coherently, then to close the tips by two caps, each cap being a hemi-fullerene with the appropriate diameter (Fig. 3.2a–c).

3.1.1 Single-Wall Nanotubes

Geometrically, there is no restriction regarding the tube diameter. But calculations have shown that collapsing the single-wall tube into a flattened two-layer ribbon is energetically more favorable than maintaining the tubular morphology beyond a diameter value of ~ 2.5 nm [3.10]. On the other hand, it intuitively comes to mind that the shorter the radius of curvature, the higher the stress and the energetic cost, although SWNTs with diameters as low as 0.4 nm have been successfully synthesized [3.11]. A suitable energetic compromise is thus reached for ~ 1.4 nm, the most frequent diameter encountered regardless of the synthesis techniques (at least those based on solid carbon source) when conditions for high SWNT yields are used. There is no such restriction for nanotube length, which only depends on limitations brought by the preparation method and the specific conditions used for the synthesis (thermal gradients, res-

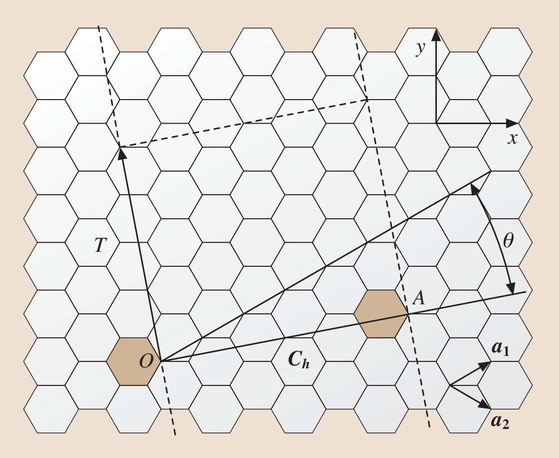

Fig. 3.1 Sketch of the way to make a single-wall carbon nanotube, starting from a graphene sheet (adapted from [3.12])

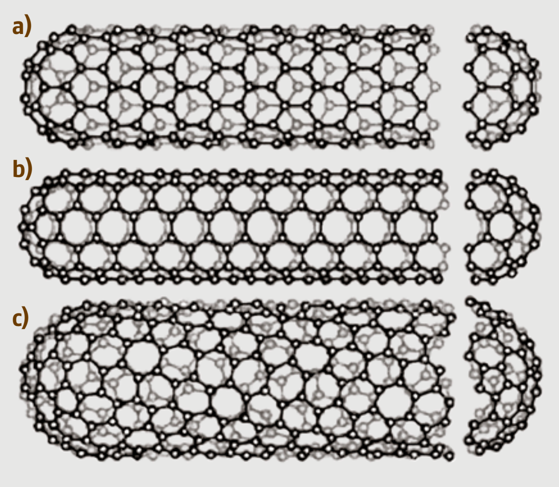

Fig. 3.2a–c Sketch of three different SWNT structures as examples for **(a)** a zig-zag-type nanotube, **(b)** an armchair-type nanotube, **(c)** a helical nanotube (adapted from [3.13])

idence time, etc.). Experimental data are consistent with these statements, since SWNTs wider than 2.5 nm are scarcely reported in literature, whatever the preparation method, while SWNT length can be in the micrometer or the millimeter range. These features make single-wall carbon nanotubes a unique example of single molecules with huge aspect ratios.

Two important consequences derive from the SWNT structure as described above:

1. All carbon atoms are involved in hexagonal aromatic rings only and are therefore in equivalent position, except at the nanotube tips where $6 \times 5 = 30$ atoms at each tip are involved in pentagonal rings (considering that adjacent pentagons are unlikely) – though not more, not less, as a consequence of the Euler's rule that also governs the fullerene structure. In case SWNTs are ideally perfect, their chemical reactivity will therefore be highly favored at the tube tips, at the very location of the pentagonal rings.
2. Though carbon atoms are involved in aromatic rings, the C=C bond angles are no longer planar as they should ideally be. This means that the hybridization of carbon atoms are no longer pure sp^2 but get some percentage of the sp^3 character, in a proportion that increases as the tube radius of curvature decreases. For example, the effect is the same as for the C_{60} fullerene molecules, whose radius of curvature is 0.35 nm, and the subsequent sp^3 character proportion about 30%. On the one hand, this is supposed to make the SWNT surface (though consisting of aromatic ring faces) a bit more reactive than regular, planar graphene, relatively speaking. On the other hand, this somehow induces a variable overlapping of the bands of density of states, thereby inducing a unique versatile electronic behavior (see Sect. 3.4).

As illustrated by Fig. 3.2, there are many ways to roll a graphene into a single-wall nanotube, some of the resulting nanotubes enabling symmetry mirrors both parallel and perpendicular to the nanotube axis (such as the SWNTs from Fig. 3.2a and 3.2b), some others not (such as the SWNT from Fig. 3.2c). By correspondence with the terms used for molecules, the latter are commonly called "chiral" nanotubes. "Helical" should be preferred, however, in order to respect the definition of chirality, which makes all chiral molecules unable to be superimposed on their own image in a mirror. The various ways to roll graphene into tubes are therefore mathematically defined by the vector of helicity $\boldsymbol{C}_\mathrm{h}$, and the angle of helicity θ, as follows (referring to Fig. 3.1):

$$\boldsymbol{OA} = \boldsymbol{C}_\mathrm{h} = n\boldsymbol{a}_1 + m\boldsymbol{a}_2$$

with

$$\boldsymbol{a}_1 = \frac{a\sqrt{3}}{2}\boldsymbol{x} + \frac{a}{2}\boldsymbol{y} \quad \text{and} \quad \boldsymbol{a}_2 = \frac{a\sqrt{3}}{2}\boldsymbol{x} - \frac{a}{2}\boldsymbol{y}$$
$$\text{where} \quad a = 2.46\,\text{Å}$$

and

$$\cos\theta = \frac{2n+m}{2\sqrt{n^2+m^2+nm}}$$

where n and m are the integers of the vector $\boldsymbol{OA}$ considering the unit vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$.

The vector of helicity $\boldsymbol{C}_\mathrm{h}(=\boldsymbol{OA})$ is perpendicular to the tube axis, while the angle of helicity θ is taken with respect to the so-called zig-zag axis, i. e., the vector of helicity that makes nanotubes of the "zig-zag" type (see below). The diameter D of the corresponding nanotube is related to $\boldsymbol{C}_\mathrm{h}$ by the relation:

$$D = \frac{|\boldsymbol{C}_\mathrm{h}|}{\pi} = \frac{a_\mathrm{CC}\sqrt{3(n^2+m^2+nm)}}{\pi}\,,$$

where

$$\underset{\text{(graphite)}}{1.41\,\text{Å}} \le a_{\mathrm{C=C}} \le \underset{(\mathrm{C}_{60})}{1.44\,\text{Å}}\,.$$

The C−C bond length is actually elongated by the curvature imposed, taking the average value for the C_{60} fullerene molecule as a reasonable upper limit, and the value for flat graphenes in genuine graphite as the lower limit (corresponding to an infinite radius of curvature). Since $\boldsymbol{C}_\mathrm{h}$, θ, and D are all expressed as a function of the integers n and m, they are sufficient to define any SWNT specifically, by noting them (n,m). Obtaining the values of n and m for a given SWNT is simple by counting the number of hexagons that separate the extremities of the $\boldsymbol{C}_\mathrm{h}$ vector following the unit vector $\boldsymbol{a}_1$ first, then $\boldsymbol{a}_2$ [3.12]. In the example of Fig. 3.1, the SWNT that will be obtained by rolling the graphene so that the two shaded aromatic cycles superimpose exactly is a (4,2) chiral nanotube. Similarly, SWNTs from Fig. 3.2a to 3.2c are (9,0), (5,5), and (10,5) nanotubes respectively, thereby providing examples of zig-zag–type SWNT (with an angle of helicity = 0°), armchair-type SWNT (with an angle of helicity of 30°) and a chiral SWNT, respectively. This also illustrates why the term "chiral" is sometimes inappropriate and should preferably be replaced by "helical". Armchair (n,n) nanotubes, though definitely achiral from a standpoint of symmetry, exhibit a "chiral angle" different from 0. "Zig-zag" and "armchair" qualifications for achiral nanotubes refer to the way carbon atoms are displayed at the edge of the nanotube cross section (Fig. 3.2a and 3.2b). Generally speaking, it is clear from Figs. 3.1 and 3.2a that having the vector of helicity perpendicular to any of the three overall C=C bond directions will provide zig-zag–type SWNTs, noted (n,0), while having the vector of helicity parallel to one of the three C=C bond directions will provide armchair-type SWNTs, noted (n,n). On the other hand, because of the sixfold symmetry of the graphene sheet, the angle of helicity θ for the chiral (n,m) nanotubes is such as $0 < \theta < 30°$. Figure 3.3 provides two examples of what chiral SWNTs look like, as seen by means of atomic force microscopy.

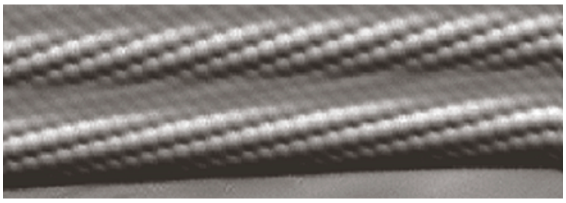

Fig. 3.3 Image of two neighboring chiral SWNTs within a SWNT bundle as seen by high resolution scanning tunneling microscopy (by courtesy of Prof. Yazdani, University of Illinois at Urbana, USA)

Planar graphenes in graphite have π electrons, which are satisfied by the stacking of graphenes that allows van der Waals forces to develop. Similar reasons make fullerenes gather and order into fullerite crystals and SWNTs into SWNT ropes (Fig. 3.4a). Spontaneously, SWNTs in ropes tend to arrange into an hexagonal array, which corresponds to the highest compactness achievable (Fig. 3.4b). This feature brings new periodicities

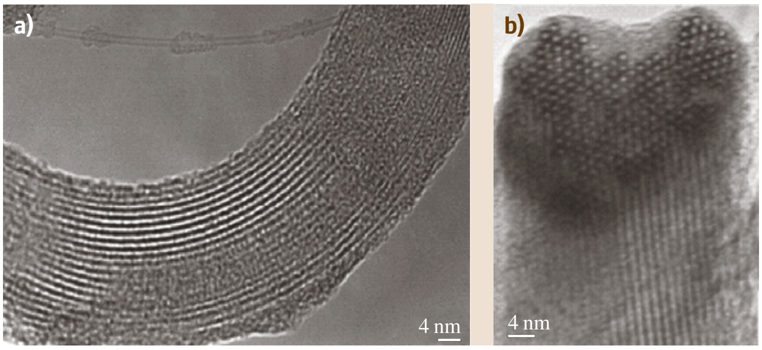

Fig. 3.4a,b High resolution transmission electron microscopy images of a SWNT rope **(a)** longitudinal view. At the top of the image an isolated single SWNT also appears. **(b)** cross section view (from [3.14])

with respect to graphite or turbostratic polyaromatic carbon crystals. Turbostatic structure corresponds to graphenes which are stacked with random rotations or translations instead of being piled up following sequential ABAB positions, as in graphite structure. This implies that no lattice atom plane exists anymore other than the graphene planes themselves (corresponding to the (001) atom plane family). These new periodicities make diffraction patterns specific and quite different from that of other sp^2-carbon-based crystals, although hk reflections, which account for the hexagonal symmetry of graphene plane, are still present. On the other hand, $00l$ reflections, which account for the stacking sequence of graphenes in regular, "multilayered" polyaromatic crystals that does not exist in SWNT ropes, are absent. Such a hexagonal packing of SWNTs within the ropes requires that SWNTs exhibit similar diameters, which is the common case for SWNTs prepared by the electric arc or the laser vaporization processes. SWNTs prepared these ways are actually about 1.35 nm wide (diameter value for a (10,10) tube, among others), for reasons still unclear but related to the growth mechanisms specific to the conditions provided by these techniques (see Sect. 3.3).

3.1.2 Multiwall Nanotubes

Building multiwall carbon nanotubes is a little bit more complex, since it has to deal with the various ways graphenes can be displayed and mutually arranged within a filament morphology. Considering the usual textural versatility of polyaromatic solids, a similar versatility can be expected. Likewise, diffraction patterns no longer differentiate from that of anisotropic polyaromatic solids. The easiest MWNT to imagine is the concentric type (c-MWNT), in which SWNTs with regularly increasing diameters are coaxially displayed according to a Russian-doll model into a multiwall nanotube (Fig. 3.5). Such nanotubes are generally formed either by the electric-arc technique (without need of any catalyst) or by catalyst-enhanced thermal cracking of gaseous hydrocarbons or CO disproportionation (see Sect. 3.2). The number of walls (or number of coaxial tubes) can be anything, starting from two, with no upper limit. The intertube distance is approximately that of the intergraphene distance in turbostratic, polyaromatic solids, i. e. 0.34 nm (as opposed to 0.335 nm in genuine graphite), since the increasing radius of curvature imposed on the concentric graphenes prevents the carbon atoms from being displayed as in graphite, i. e., each of the carbon atoms from a graphene facing alternatively either a ring center or a carbon atom from the neighboring graphene. But two cases allow such a nanotube to reach – totally or partially – the 3-D crystal periodicity of graphite. One is to consider a high number of concentric graphenes, i. e. concentric graphenes with long radius of curvature. In that case, the shift of the relative positions of carbon atoms from superimposed graphenes is so small with respect to that in graphite that some commensurability is possible. It may result in MWNTs with the association of both structures, i. e. turbostratic in the core and graphitic in the outer part [3.15]. The other case occurs for c-MWNTs exhibiting faceted morphologies, originating either from the synthesis process or more likely from

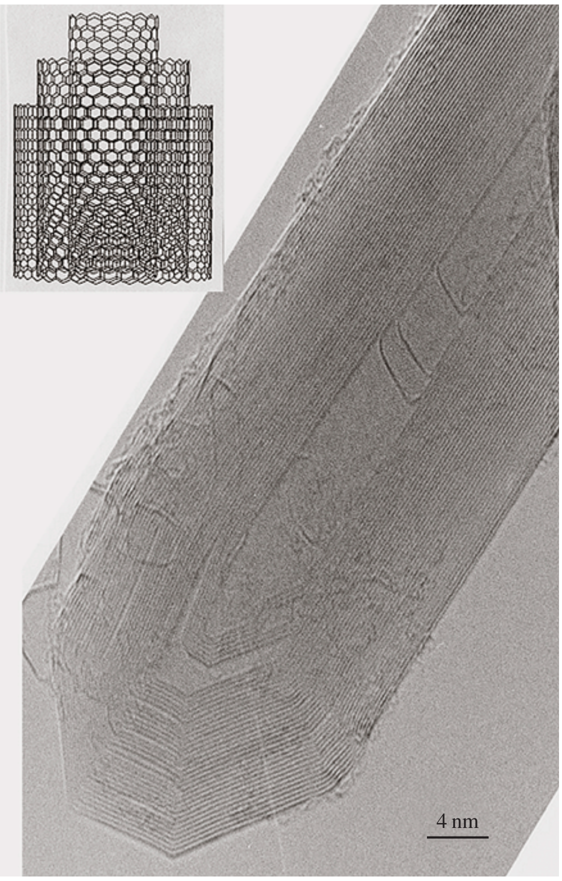

Fig. 3.5 High resolution transmission electron microscopy image (longitudinal view) of a concentric multiwall carbon nanotube (c-MWNT) prepared by electric arc. In *insert*, sketch of the Russian-doll-like display of graphenes

subsequent heat-treatments at high temperatures (e.g., 2,500 °C) in inert atmospheres. Obviously, facets allow graphenes to get back a flat arrangement of atoms – except at the junction between neighboring facets – in which the specific stacking sequence of graphite can develop.

Another frequent inner texture for multiwall carbon nanotubes is the so-called herringbone texture (h-MWNTs), in which graphenes make an angle with respect to the nanotube axis (Fig. 3.6). The angle value varies upon the processing conditions (e.g., the catalyst morphology or the atmosphere composition), from 0 (in which case the texture turns into that of a c-MWNT) to 90° (in that case, the filament is no longer a tube, see below), and the inner diameter varies so that the tubular feature can be lost [3.19], justifying that the latter can be called nanofibers rather than nanotubes. h-MWNTs are exclusively obtained by processes involving catalysts, generally catalyst-enhanced thermal cracking of hydrocarbons or CO disproportionation. One unresolved question is whether the herringbone texture, which actually describes the texture projection rather than the overall three-dimensional texture, originates from the scroll-like spiral arrangement of a single graphene ribbon or from the stacking of independent truncated-cone-like graphenes. Although the former is more likely for energetic reasons since providing a minimal and constant amount of transitorily unsatisfied bonds during the nanotube growth (similar to the well-known growth mechanism from a screw dislocation), the question is still debated.

Another common feature is the occurrence, at a variable frequency, of a limited amount of graphenes oriented perpendicular to the nanotube axis, thus forming the so-called "bamboo" texture. It cannot be a texture by its own but affects in a variable extent either c-MWNT (bc-MWNT) or h-MWNT (bh-MWNT) textures (Figs. 3.6 and 3.7). The question may be addressed whether such filaments, though hollowed, should still be called nanotubes, since the inner cavity is no longer opened all along the filament as it is for a genuine tube. They therefore are sometimes referred as "nanofibers" in literature.

Nanofilaments that definitely cannot be called nanotubes are those built from graphenes oriented perpendicular to the filament axis and stacked as piled-up plates. Although they actually correspond to h-MWNTs with the graphene/MWNT axis angle = 90°, the occurrence of inner cavity is no longer possible, and such filaments are therefore most often referred to as "platelet-nanofibers" in literature [3.19].

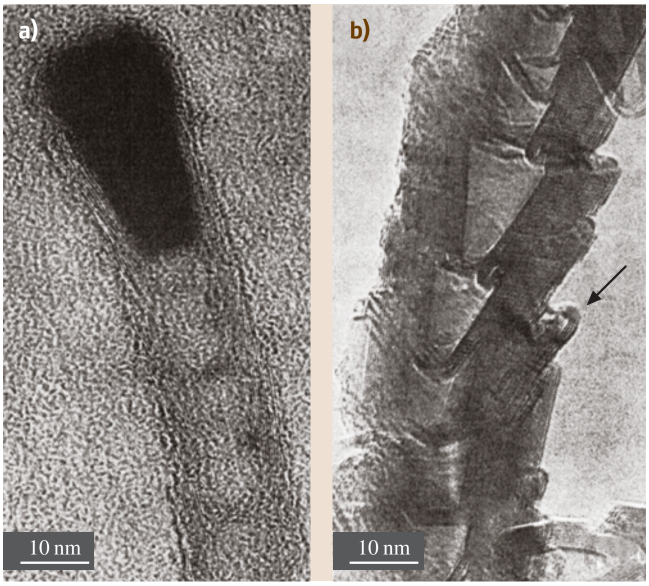

Fig. 3.6a,b One of the earliest high resolution transmission electron microscopy image of a herringbone (and bamboo) multi-wall nanotube (bh-MWNT, longitudinal view) prepared by CO disproportionation on Fe-Co catalyst. **(a)** as-grown. The nanotube surface is made of free graphene edges. **(b)** after 2,900 °C heat-treatment. Both the herringbone and the bamboo textures have become obvious. Graphene edges from the surface have buckled with neighbors (*arrow*), closing the access to the intergraphene spacing (adapted from [3.16])

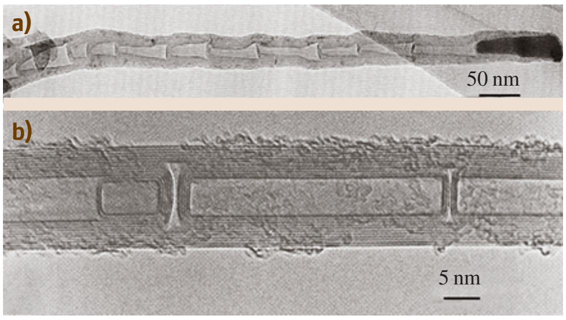

Fig. 3.7a,b Transmission electron microscopy images from bamboo-multi-wall nanotubes (longitudinal views). **(a)** low magnification of a bamboo-herringbone multi-wall nanotube (bh-MWNT) showing the nearly periodic feature of the texture, which is very frequent. (from [3.17]); **(b)** high resolution image of a bamboo-concentric multi-wall nanotube (bc-MWNT) (modified from [3.18])

As opposed to SWNTs, whose aspect ratio is so high that acceding the tube tips is almost impossible, aspect ratio for MWNTs (and carbon nanofibers) are generally lower and often allow one to image tube ends by transmission electron microscopy. Except for c-MWNTs from electric arc (see Fig. 3.5) that grow following a catalyst-free process, nanotube tips are frequently found associated with the catalyst crystal from which they have been formed.

MWNT properties (see Sect. 3.4) will obviously depend on the perfection and orientation of graphenes in the tube more than any other feature (e.g., the respective spiral angle of the constituting nanotubes for c-MWNTs has little importance). Graphene orientation is a matter of texture, as described above. Graphene perfection is a matter of nanotexture, which is commonly used to describe other polyaromatic carbon materials, and which is quantified by several parameters preferably obtained from high resolution transmission electron microscopy (Fig. 3.8). Both texture and nanotexture depend on the processing conditions. While the texture type is a permanent, intrinsic feature only able to go toward complete alteration upon severe degradation treatments (e.g., oxidation), however, nanotexture can be improved by subsequent thermal treatments at high temperatures (e.g., > 2,000 °C) or, reversibly, possibly degraded by chemical treatments (e.g., slight oxidation conditions).

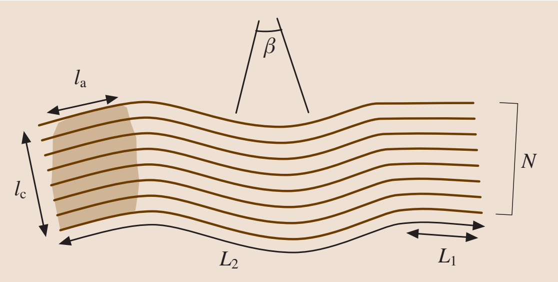

Fig. 3.8 Sketch explaining the various parameters obtained from high resolution (lattice fringe mode) transmission electron microscopy for quantifying nanotexture: L_1 is the average length of perfect (distortion-free) graphenes of coherent areas; N is the number of piled-up graphenes in the coherent (distortion-free) areas; L_2 is the average length of continuous though distorted graphenes within graphene stacks; β is the average distortion angle. L_1 and N are more or less related to L_a and L_c obtained from X-ray diffraction

3.2 Synthesis of Carbon Nanotubes

Producing carbon nanotubes so that the currently planned applications become marketable requires solving some problems that are more or less restrictive depending on the cases. One example is to specifically control the configuration (chirality), the purity, or the structural quality of SWNTs, with the production capacity adapted to the application. One condition would be to understand perfectly the mechanism of nanotube nucleation and growth, which remains the object of a lot of controversy in spite of an intense, worldwide experimental effort. This problem is partly explained by the lack of knowledge regarding several parameters controlling the synthesis conditions. For instance, neither the temperatures of nanotube condensation and formation nor the manner in which they are influenced by the synthesis parameters are known. Equally often unknown is the exact and accurate role of the catalysts in nanotube growth. Given the large number of experimental parameters and considering the large range of conditions that the several synthesis techniques correspond to, it is quite legitimate to think of more than one mechanism intervening in the nanotube formation.

3.2.1 Solid Carbon Source-Based Production Techniques for Carbon Nanotubes

Among the different SWNT production techniques, the three processes (laser ablation, solar energy, and electric arc) presented in this section have at least two common points: a high temperature (1,000 K < T < 6,000 K) medium and the fact that the carbon source originates from the erosion of solid graphite. In spite of these common points, both the morphologies of the carbon nanostructures and the SWNT yields can differ notably with respect to the experimental conditions.

Before being utilized for carbon nanotube synthesis, these techniques permitted the production of fullerenes. Laser vaporization of graphite was actually the very first method to demonstrate the existence of fullerenes, including the most popular one (because the most stable and therefore the most abundant), C_{60} [3.20]. On the other hand, the electric arc technique was (and still is) the first method to produce fullerenes in relatively large quantities [3.21]. As opposed to the fullerene formation, which requires the presence of carbon atoms in high temperature media and the absence of oxygen, the utilization of these techniques for the synthesis of nanotubes requires an additional condition, i. e. the presence of catalysts in one of the electrode or the target.

In case of these synthesis techniques requiring relatively high temperatures to sublime graphite, the different mechanisms such as the carbon molecule dissociation and the atom recombination processes take place at different time scales, from nanosecond to microsecond and even millisecond. The formation of nanotubes and other graphene-based products appears afterward with a relatively long delay.

The methods of laser ablation, solar energy, and electric arc are based on an essential mechanism, i. e. the energy transfer resulting from the interaction between either the target material and an external radiation (laser beam or radiation emanating from solar energy) or the electrode and the plasma (in case of electric arc). This interaction is at the origin of the target or anode erosion leading to the formation of a plasma, i. e. an electrically neutral ionized gas, composed of neutral atoms, charged particles (molecules and ionized species), and electrons. The ionization degree of this plasma, defined by the ratio $(n_e/n_e + n_o)$, where n_e and n_o are the density of electrons and that of neutral atoms respectively, highlights the importance of energy transfer between the plasma and the material. The characteristics of this plasma and notably the fields of temperature and of concentration of the various species present in the plasma thereby depend not only on the nature and composition of the target or that of the electrode but also on the energy transferred.

One of the advantages of these different synthesis techniques is the possibility of varying a large number of parameters, which allow the composition of the high temperature medium to be modified and, consequently, the most relevant parameters to be determined, in order to define the optimal conditions for the control of carbon nanotube formation. But a major drawback of these techniques – as for any other technique for the production of SWNTs – is that SWNTs never come pure, i. e. they are associated with other carbon phases and catalyst remnants. Although purification processes are proposed in literature and by some commercial companies to rid these undesirable phases, they are all based on oxidation (e.g. acidic) processes that are likely to affect deeply the SWNT structure [3.14]. Subsequent thermal treatments at ∼ 1,200 °C under inert atmosphere, however, succeed somewhat in recovering the former structure quality [3.24].

Laser Ablation

After the first laser was built in 1960, physicists immediately made use of it as a means of concentrating a large quantity of energy inside a very small volume within a relatively short time. The consequences of this energy input naturally depends on the characteristics of the device employed. During the interaction between the laser beam and the material, numerous phenomena superimpose and/or follow each other within the time range, each of these processes being sensitive to such different parameters as the laser beam characteristics, the incoming power density (also termed "fluence"), the nature of the target, and the environment surrounding it. For instance, the solid target can merely heat up, or melt, or vaporize depending on the power provided.

While this technique was used successfully to synthesize fullerene-related structures for the very first time [3.20], the synthesis of SWNTs by laser ablation came only ten years later [3.22].

Laser Ablation – Experimental Devices

Two types of laser devices are utilized nowadays for carbon nanotube production: lasers operating in pulsed mode on the one hand and lasers operating in continuous mode on the other hand, the latter generally providing a smaller influence.

An example of the layout indicating the principle of a laser ablation device is given in Fig. 3.9. A graphite pellet containing the catalyst is put in the middle of an inert gas-filled quartz tube placed in an oven maintained at a temperature of 1,200 °C [3.22, 23]. The energy of the laser beam focused on the pellet permits it to vaporize and sublime the graphite by uniformly bombarding its surface. The carbon species swept by a flow of neutral gas are thereafter deposited as soot in different regions: on the conical water-cooled copper collector, on the quartz tube walls, and on the backside of the pellet.

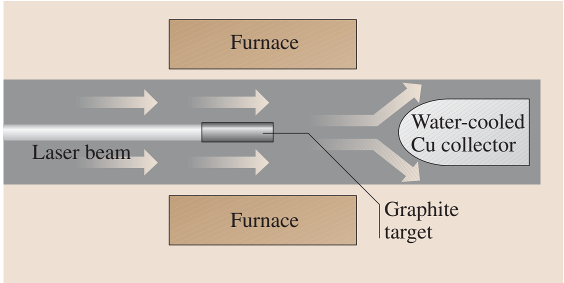

Fig. 3.9 Sketch of an early laser vaporization apparatus (adapted from [3.22, 23])

Various improvements have been made to this device in order to increase the production efficiency. For example, *Thess* et al. [3.25] employed a second pulsed laser that follows the initial impulsion but at a different frequency in order to ensure a more complete and efficient irradiation of the pellet. This second impulsion has the role of vaporizing the coarse aggregates issued from the first ablation and thereby making them participate in the active carbon feedstock involved in the nanotube growth. Other modifications were brought by *Rinzler* et al. [3.24], who inserted a second quartz tube of a smaller diameter coaxially disposed inside to the first one. This second tube has the role of reducing the vaporization zone and thereby permitting an increase in the quantity of sublimed carbon. They also arranged to place the graphite pellet on a revolving system so that the laser beam uniformly scans its whole surface.

Other groups have realized that, as far as the target contains both the catalyst and the graphite, the latter evaporates in priority and the pellet surface becomes more and more metal rich, resulting in a decrease of the efficiency in nanotube formation in the course of the process. To solve this problem, *Yudasaka* et al. [3.27] utilized two pellets facing each other, one entirely made from the graphite powder and the other from an alloy of transition metals (catalysts), and irradiated simultaneously.

A sketch of a synthesis reactor based on the vaporization of a target at a fixed temperature by a continuous CO_2 laser beam ($\lambda = 10.6\,\mu m$) is shown in Fig. 3.10 [3.26]. The power can be varied from 100 W to 1,600 W. The temperature of the target is measured with an optical pyrometer, and these measurements are used to regulate the laser power to maintain a constant vaporization temperature. The gas, heated by the contact with the target, acts as a local furnace and creates an extended hot zone, making an external furnace unnecessary. The gas is extracted through a silica pipe, and the solid products formed are carried away by the gas flow through the pipe and then collected on a filter. The synthesis yield is controlled by three parameters: the cooling rate of the medium where the active, secondary catalyst particles are formed, the residence time, and the temperature (in the 1,000–2,100 K range) at which SWNTs nucleate and grow [3.28].

But devices equipped with facilities to gather in situ data such as the target temperature are few and, generally speaking, among the numerous parameters of the laser ablation synthesis technique. The most studied are the nature of the target, the nature and concentration of the catalysts, the nature of the neutral gas flow, and the temperature of the outer oven (when any).

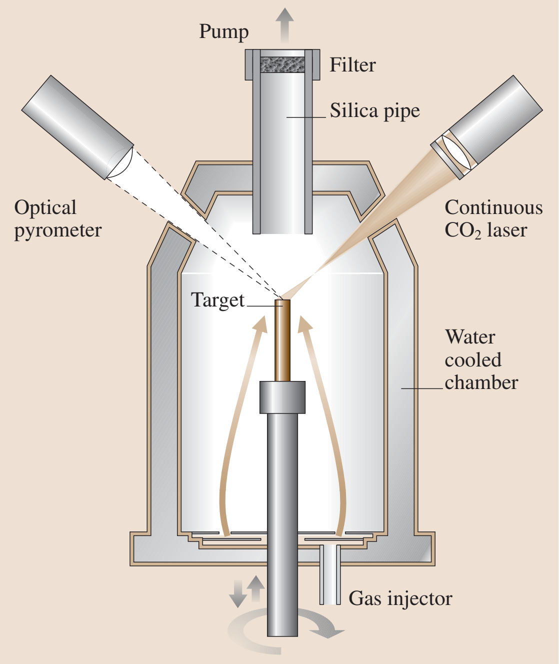

Fig. 3.10 Sketch of a synthesis reactor with a continuous CO_2 laser device (adapted from [3.26])

Laser Ablation – Results

In the absence of catalysts in the target, the soot collected mainly contains multiwall nanotubes (c-MWNTs). Their lengths can reach 300 nm. Their quantity and structure quality are dependent on the oven temperature. The best quality is obtained for an oven temperature set at 1,200 °C. At lower oven temperatures, the structure quality decreases, and the nanotubes start presenting many defects [3.23]. As soon as small quantities (few percents or less) of transition metals (Ni, Co) playing the role of catalysts are incorporated into the graphite pellet, products yielded undergo significant modifications, and SWNTs are formed instead of MWNTs. The yield of SWNTs strongly depends on the type of metal catalyst used and is seen to increase with furnace temperature, among other factors. The SWNTs have remarkably uniform diameter and they self-organize into rope-like crystallites 5–20 nm in diameter and tens to hundreds of micrometers in length (Fig. 3.11). The ends of all SWNTs appear to be perfectly closed with hemispherical end caps showing no evidence of any as-

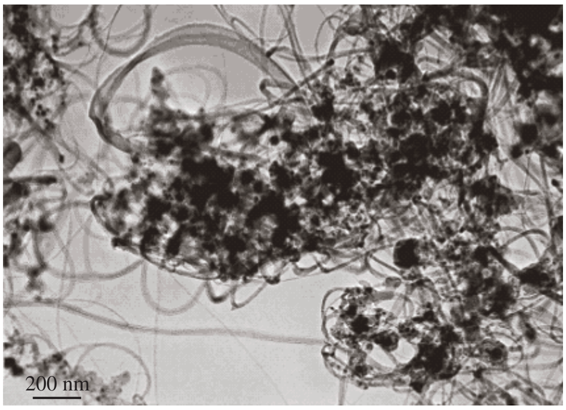

Fig. 3.11 Low magnification TEM images of a typical raw SWNT material obtained from the laser vaporization technique. Fibrous morphologies are SWNT bundles, and dark particles are catalyst remnants. Raw SWNT materials from electric arc exhibit a similar aspect (from [3.14])

sociated metal catalyst particle, although, as pointed out in Sect. 3.1, finding the two tips of a SWNT is rather challenging, considering the huge aspect ratio of the nanotube and their entanglement. Another feature of SWNTs produced with this technique is that they are supposedly "cleaner" than those produced employing other techniques, i. e., associated with a lower amount of an amorphous carbon phase, either coating the SWNTs or gathered into nano-particles. Such an advantage, however, stands only for synthesis conditions specifically set to ensure high quality SWNTs. It can no longer be true when conditions are such that high yields are preferred, and SWNTs from electric arc may appear cleaner than SWNTs from laser vaporization [3.14].

The laser vaporization technique is one of the three methods currently used to prepare SWNTs as commercial products. SWNTs prepared that way were first marketed by Carbon Nanotechnologies Inc. (Texas, USA), with prices as high as $ 1,000/g (raw materials) until December 2002. As a probable consequence of the impossibility to lower the amount of impurities in the raw materials, they have recently decided to focus on fabricating SWNTs by the HiPCo technique (see Sect. 3.2.2). Though laser-based method are generally considered not competitive in the long term for the low-cost production of SWNTs compared to CCVD-based methods (see Sect. 3.2.2), prediction of prices as low as $ 0.03 per gram of raw high concentration SWNT soot are expected in the near future (Acolt S.A., Switzerland).

Electric-Arc Method

Electric arcs between carbon electrodes have been studied as light sources and radiation standards for a very long time. Lately they have received renewed attention for their use in producing new fullerenes-related molecular carbon nanostructures such as genuine fullerenes or nanotubes. This technique was first brought to light by *Krätschmer* et al. [3.21] who utilized it to achieve the production of fullerenes in macroscopic quantities. In the course of investigating the other carbon nanostructures formed along with the fullerenes and more particularly the solid carbon deposit forming onto the cathode, Iijima [3.7] discovered the catalyst-free formation of perfect c-MWNT-type carbon nanotubes. Then, as mentioned in the introduction of this chapter, the catalyst-promoted formation of SWNTs was incidentally discovered after some amounts of transition metals were introduced into the anode in an attempt to fill the c-MWNTs with metals while they grow [3.8, 9]. Since then, a lot of work has been carried out by many groups using this technique to understand the mechanisms of nanotube growth as well as the role played by the catalysts (when any) for the synthesis of MWNTs and/or SWNTs [3.29–41].

Electric-Art Method – Experimental Devices

The principle of this technique is to vaporize carbon in the presence of catalysts (iron, nickel, cobalt, yttrium, boron, gadolinium, and so forth) under reduced atmosphere of inert gas (argon or helium). After the triggering of the arc between two electrodes, a plasma is formed consisting of the mixture of carbon vapor, the rare gas (helium or argon), and the vapors of catalysts. The vaporization is the consequence of the energy transfer from the arc to the anode made of graphite doped with catalysts. The anode erosion rate is more or less important depending on the power of the arc and also on the other experimental conditions. It is noteworthy that a high anode erosion does not necessarily lead to a high carbon nanotube production.

An example of a reactor layout is shown in Fig. 3.12. It consists of a cylinder of about 30 cm in diameter and about 1 m in height, equipped with diametrically opposed sapphire windows located so that they face the plasma zone in view of observing the arc. The reactor possesses two valves, one for carrying out the primary evacuation (0.1 Pa) of the chamber, the other permitting it to fill with a rare gas up to the desired working pressure.

Contrary to the solar energy technique, SWNTs are deposited (provided appropriate catalysts are used) in

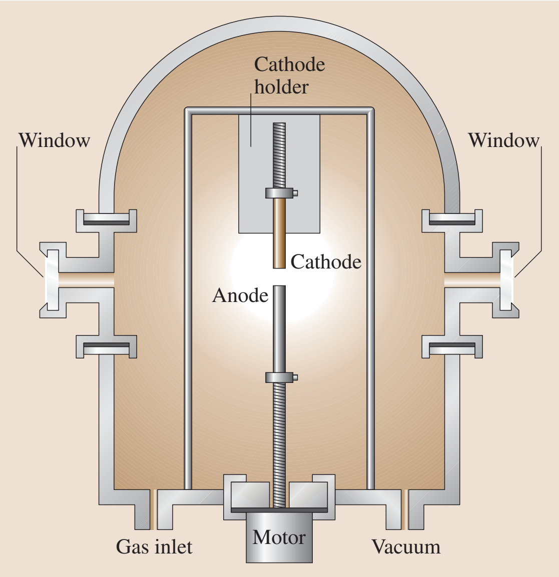

Fig. 3.12 Sketch of an electric arc reactor

different regions of the reactor: (1) the collaret, which forms around the cathode; (2) the web-like deposits found above the cathode; (3) the soot deposited all around the reactor walls and bottom. On the other hand, MWNTs are formed embedded in a hard deposit adherent to the cathode, whenever catalysts are used or not. The cathode deposits formed under the cathode. The formation of collaret and web is not systematic and depends on the experimental conditions as indicated in Table 3.1, as opposed to cathode deposit and soot that are systematically found.

Two graphite rods of few millimeters in diameter constitute the electrodes between which a potential difference is applied. The dimensions of these electrodes vary according to the authors. In certain cases, the cathode has a greater diameter than the anode in order to facilitate their alignment [3.32, 42]. Other authors utilize electrodes of the same diameter [3.41]. The whole device can be designed horizontally [3.33, 41] or vertically [3.34, 36–38]. The advantage of the latter is the symmetry brought by the verticality with respect to gravity, which facilitates computer modeling (regarding convection flows, for instance).

Two types of anodes can be utilized as soon as catalysts need to be introduced: (1) graphite anodes in which a coaxial hole is drilled several centimeters in length and in which catalyst and graphite powders are mixed; (2) graphite anodes in which the catalysts are homogeneously dispersed [3.43]. The former are by far the most popular, due to their ease of fabrication.

The optimization of the process regarding the nanotube yield and quality is attempted by studying the role of various parameters such as the type of doped anode (homogeneous or heterogeneous catalyst dispersion), the nature as well as the concentration of catalysts, the nature of the plasmagenic gas, the buffer gas pressure, the arc current intensity, and the distance between electrodes. Investigating the result of varying these parameters on the type and amount of carbon nanostructures formed is, of course, the preliminary work that has been done. Though electric arc reactors equipped with the related facilities are few (see Fig. 3.12), investigating the missing link, i. e. the effect of varying the parameters on the plasma characteristics (fields of concentration species and temperature), is likely to provide a more comprehensive understanding of the phenomena involved in nanotube formation. This has been recently developed by atomic and molecular optical emission spectroscopy [3.34, 36–39, 41].

Finally, mention has to be made of attempts to make the electric arc within liquid media such as liquid nitrogen [3.44] or water [3.45, 46], the goal being easier processing since they do not require pumping devices or a closed volume and therefore are likely to allow continuous synthesis. This adaptation has not, however, reached the state of mass production.

Electric-Arc Method – Results

In view of the numerous results obtained with this electric-arc technique, it appears clearly that both the nanotube morphology and the nanotube production efficiency strongly depend on the experimental conditions and, in particular, on the nature of the catalysts. It is worth noting that the products obtained do not consist solely in carbon nanotubes but also in nontubular forms of carbon such as nanoparticles, fullerene-like structures including C_{60}, poorly organized polyaromatic carbons, nearly amorphous nanofibers, multiwall shells, single-wall nanocapsules, and amorphous carbon as reported in Table 3.1 [3.35, 37, 38]. In addition, catalyst remnants are found all over the place, i. e. in the soot, collaret, web, and cathode deposit in various concentration. Generally, at pressure value of about 600 mbar of helium, for 80 A arc current and 1 mm electrode gap, the use of Ni/Y as coupled catalysts favors more particularly the synthesis of SWNTs [3.8, 33, 47]. For such conditions providing high SWNT yields, SWNT concentrations are higher in the collaret (in the range of 50–70 %), then in the web (∼ 50% or less), then in the soot. On the

Table 3.1 Different carbon morphologies obtained by changing the type of anode, the type of catalysts, and pressure value, in a series of arc-discharge experiments (electrode gap = 1 mm)

Catalyst (atom%) Arc conditions	**0.6Ni + 0.6Co (homogeneous anode) $P \sim 60$ kPa $I \sim 80$ A**	**0.6Ni + 0.6Co (homogeneous anode) $P \sim 40$ kPa $I \sim 80$ A**	**0.5Ni + 0.5Co $P \sim 60$ kPa $I \sim 80$ A**	**4.2Ni + 1Y $P \sim 60$ kPa $I \sim 80$ A**
Soot	• **MWNT** + **MWS** + **POPAC** or **Cn** ± catalysts $\phi \sim 3-35$ nm • NANF + catalysts • AC particles + catalysts • [*DWNT*], [*SWNT*], ropes or isolated, + POPAC	• **POPAC** and **AC** particles + **catalysts** $\phi \sim 2-20$ nm • NANF + catalysts $\phi \sim 5-20$ nm + MWS • [*SWNT*] $\phi \sim 1-1.4$ nm, distorted or damaged, isolated or ropes + Cn	• **AC** and POPAC particles + catalysts $\phi \sim 3-35$ nm • **NANF** + catalysts $\phi \sim 4-15$ nm • [*SWNT*] $\phi \sim 1.2$ nm, isolated or ropes	• **POPAC** and **AC** + particles + catalysts $\phi \leq 30$ nm • SWNT $\phi \sim 1.4$ nm, clean + Cn, short with tips, [*damaged*], isolated or ropes $\phi \leq 25$ nm • [*SWNC*] particles
Web	• [*MWNT*], **DWNT**, $\phi\, 2.7 - \mathbf{4} - 5.7$ nm SWNT $\phi\, 1.2-1.8$ nm, isolated or **ropes** $\phi < 15$ nm, + POPAC ± **Cn** • AC particles + catalysts $\phi \sim 3-40$ nm + MWS • [*NANF*]	None	None	• **SWNT**, $\phi \sim 1.4$ nm, isolated or **ropes** $\phi \leq 20$ nm, + **AC** • POPAC and AC particles + catalysts $\phi \sim 3 - \mathbf{10} - 40$ nm + MWS
Collaret	• **POPAC** and **SWNC** particles • Catalysts $\phi \sim 3-250$ nm, $< \mathbf{50}$ nm + MWS • SWNT $\phi\, 1-1.2$ nm, [*opened*], **distorted**, isolated or **ropes** $\phi < 15$ nm, + **Cn** • [*AC*] particles	• **AC** and **POPAC** particles + catalysts $\phi \sim 3-25$ nm • SWNT $\phi \sim 1-1.4$ nm clean + Cn, [*isolated*] or ropes $\phi < 25$ nm • Catalysts $\phi \sim 5-50$ nm + MWS, • [SWNC]	• **Catalysts** $\phi \sim 3-170$ nm + MWS • AC or POPAC particles + **catalysts** $\phi \sim 3-50$ nm • SWNT $\phi \sim 1.4$ nm clean + Cn isolated or ropes $\phi < 20$ nm	• **SWNT** $\phi \sim 1.4-2.5$ nm, clean + Cn, [*damaged*], isolated or **ropes** $\phi < 30$ nm • POPAC or AC particles + catalysts $\phi \sim 3-30$ nm • [*MWS*] + catalysts or catalyst-free.
Cathode deposit	• **POPAC** and SWNC particles • **Catalysts** $\phi \sim$ 5–300 nm MWS • MWNT $\phi < 50$ nm • [*SWNT*] $\phi \sim 1.6$ nm clean + Cn, isolated or ropes	• **POPAC** and **SWNC** particles + Cn • Catalysts $\phi \sim 20-100$ nm + MWS	• **MWS**, catalyst-free • **MWNT** $\phi < 35$ nm • POPAC and PSWNC particles • [*SWNT*], isolated or ropes • [*Catalysts*] $\phi \sim 3-30$ nm	• **SWNT** $\phi \sim 1.4-4.1$ nm, clean + **Cn**, short with tips, isolated or ropes $\phi \leq 20$ nm. • **POPAC** or **AC** particles + catalysts $\phi \sim 3-30$ nm • MWS + catalysts $\phi < 40$ nm or catalyst-free • [*MWNT*]

Abundant – Present – [*Rare*]

Glossary: AC: amorphous carbon; POPAC: poorly organised polyaromatic carbon; Cn: fullerene-like structure, including C_{60}; NANF: nearly amorphous nanofiber; MWS: multiwall shell; SWNT: single-wall nanotube; DWNT: double-wall nanotube, MWNT: multiwall nanotube; SWNC: single-wall nanocapsule.

other hand, c-MWNTs are found in the cathode deposit. SWNT lengths are micrometric and, typically, outer diameters range around 1.4 nm. Taking the latter conditions as a reference experiment (Table 3.1, column 4), Table 3.1 illustrates the consequence of changing the parameters. For instance (Table 3.1, column 3), using Ni/Co instead of Ni/Y as catalysts prevents the formation of SWNTs. But when the Ni/Co catalysts are homogeneously dispersed in the anode (Table 3.1, column 1), the formation of nanotubes is promoted again, but MWNTs with two or three walls prevail over SWNTs, among which DWNTs (double-wall nanotubes) make a majority. But it is enough to decrease the ambient pressure from 60 to 40 kPa (Table 3.1, column 2) for getting back to conditions where the nanotube formation is impeached.

Recent works have attempted to replace graphite powder (sp^2 hybridized carbon) by diamond powder (sp^3 hybridized carbon) to mix with the catalyst powder and fill hollowed-type graphite anodes. The result was an unexpected but quite significant increase (up to +230%) in the SWNT yield [3.39,40]. Such experiments reveal, as for the comparison between the results from using homogeneous instead of heterogeneous anodes that the physical phenomena (charge and heat transfers) that occurred in the anode during the arc are of utmost importance, which was neglected until now.

It is clear that while the use of rare earth element (such as Y) as a single catalyst does not provide the conditions to grow SWNTs, associating it with a transition metal (such as Ni/Y) seems to correspond to the most appropriate combinations, leading to the highest SWNT yields [3.42]. On the other hand, using a single rare earth element may lead to unexpected results, such as the anticipated closure of graphene edges from a c-MWNT wall with the neighboring graphene edges from the same wall side, leading to the preferred formation of telescope-like and open c-MWNTs able to contain nested Gd crystals [3.36, 38]. Such a need for bimetallic catalysts has just begun to be understood as a possible requirement to promote the transitory formation of nickel particles coated with yttrium carbide, whose lattice constants are somewhat commensurable with that of graphenes [3.48].

Figure 3.13 illustrates the kind of information provided by the analysis of the plasma by means of emission spectroscopy, i.e. radial temperature (Fig. 3.13a) or C_2 species concentration (Fig. 3.13b) profiles in the plasma. A common feature is that a huge vertical gradient (∼ 500 K/mm) rapidly establishes (at ∼ 0.5 mm from the center in the radial direction) from the bottom to the top of the plasma, as a probable consequence of convection phenomena (Fig. 3.13a). The zone of actual SWNT formation is beyond the limit of the volume analyzable in the radial direction, corresponding to colder areas. The C_2 concentration increases dramatically from the anode to the cathode and decreases dramatically in the radial direction (Fig. 3.13b). This demonstrates that C_2 moieties are secondary products resulting from the recombination of primary species formed from the anode. It also suggests that C_2 moieties may be the building blocks for MWNTs (formed at the cathode) but not for SWNTs [3.38, 40].

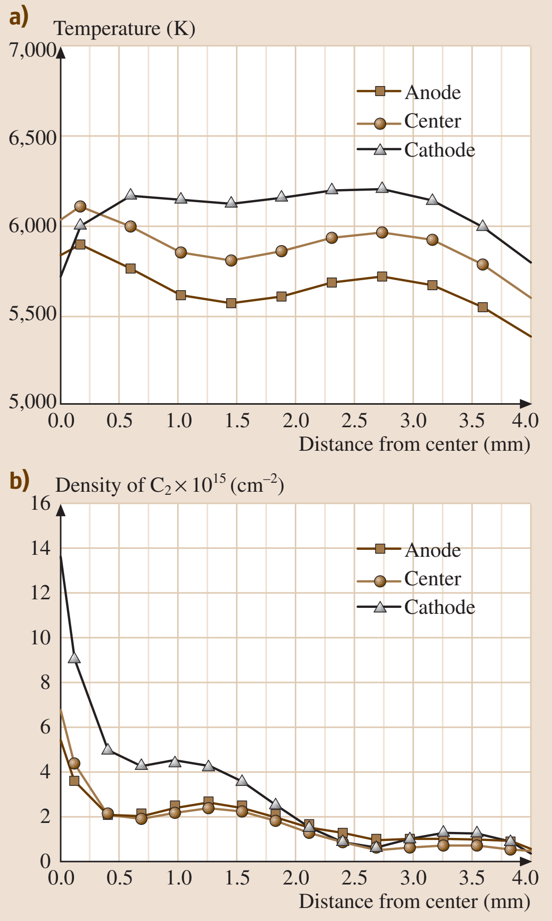

Fig. 3.13a,b Typical temperature **(a)** and C_2 concentration **(b)** profiles in the plasma at the anode surface (*square*), at the plasma center (*dot*), and at the cathode surface (*triangle*) for "standard" conditions (see text). Gradients are similar whatever the catalyst, although absolute value may vary

Although many aspects of it still need to be understood, the electric-arc method is one of the three methods currently used to produce SWNTs as commercial products. At the French company Nanoledge S.A. (Montpellier, France), for instance, current production (for 2003) reaches ~ 20 kg/year (raw SWNTs, i. e. not purified), with a marketed price of ~ 90 Euros/g, which is much cheaper than any other production method so far. A decrease to 2–5 Euros/g is expected for sometime before 2007, with a SWNT production capacity of 4–5 tons/year. Amazingly, raw SWNTs from electric arc proposed by Bucky USA (Texas, USA) are still proposed at a marketed price of $ 1,000/g.

Solar Furnace

Solar furnace devices were originally utilized by several groups to produce fullerenes [3.49–51]. *Heben* et al. [3.52] and *Laplaze* et al. [3.53] later modified their former devices in order to use them for carbon nanotube production. This modification consisted mainly in using more powerful ovens [3.54, 55].

Solar Furnace – Experimental Devices

The principle of this technique is again based on the sublimation in an inert gas of a mixture of graphite powder and catalysts formerly placed in a crucible. An example of such a device is shown in Fig. 3.14. The solar rays are collected by a plain mirror and reflected toward a parabolic mirror that focuses them directly onto a graphite pellet under controlled atmosphere. The high temperature of about 4,000 K permits both the carbon and the catalysts to vaporize. The vapors are then dragged by the neutral gas and condense onto the cold walls of the thermal screen. The reactor consists of a brass support cooled by water circulation, on which Pyrex® chambers with various shapes can be fixed (Fig. 3.14b). This support contains a watertight passage permitting the introduction of the neutral gas and a copper rod on which the target is mounted. The target is surrounded by a graphite tube that plays both the role of a thermal screen to reduce radiation losses (very important in the case of graphite) and the role of a duct to lead carbon vapors to a filter in order to avoid soot deposits on the Pyrex® chamber wall. A graphite crucible filled with powdered graphite (for fullerene synthesis) or a mixture of graphite and catalysts (for nanotube synthesis) is utilized in order to reduce the conduction losses.

These studies primarily investigated the target composition, the type and concentration of catalysts, the flow-rate, the composition and pressure of the plasmagenic gas inside the chamber, and the oven power. The objectives are similar to that of works carried out with the other solid-carbon-source-based processes. When possible, specific in situ diagnostics (pyrometry, optical emission spectroscopy, etc.) are also performed in order to understand better the role of the various parameters (temperature measurements at the crucible surface, along the graphite tube acting as thermal screen, C_2 radical concentration in the immediate vicinity of the crucible).

Solar Furnace – Results

Some results obtained by different groups concerning the influence of the catalysts can be summarized as follows. With Ni/Co and at low pressure, the sample collected contains mainly MWNTs with bamboo texture, carbon shells, and some bundles of SWNTs [3.54]. At higher pressures, only bundles of SWNTs are obtained with fewer carbon shells. With Ni/Y and at a high pressure, relatively long bundles of SWNTs are observed. With Co, SWNT bundles are obtained in the soot with SWNT diameters ranging from 1 to 2 nm. Laplaze et al. [3.54] observed very few nanotubes but a large quantity of carbon shells.

In order to proceed to a large-scale synthesis of single-wall carbon nanotubes, which is still a challenge for chemical engineers, *Flamant* et al. [3.56] recently demonstrated that solar energy–based synthesis is a versatile method to obtain SWNTs and can be scaled from 0.1–0.5 g/h to 10 g/h and then to 100 g/h productivity using existing solar furnaces. Experiments at medium scale (10-g/h, 50 kW solar power) have proven the feasibility of designing and building such a reactor and of the scaling-up method. Numerical simulation was meanwhile performed in order to improve the selectivity of the synthesis, in particular by controlling the carbon vapor cooling rate.

3.2.2 Gaseous Carbon Source-Based Production Techniques for Carbon Nanotubes

As mentioned in the introduction of this chapter, the catalysis-enhanced thermal cracking of gaseous carbon source (hydrocarbons, CO) – commonly referred to as catalytic chemical vapor deposition (CCVD) – has been known to produce carbon nanofilaments for a long time [3.4], so that reporting all the works published in the field since the beginning of the century is almost impossible. Until the 90s, however, carbon nanofilaments were mainly produced to act as a core substrate for the subsequent growth of larger (micrometric) carbon fibers – so-called vapor-grown carbon fibers – by means of thickening through mere catalyst-free CVD processes [3.57, 58]. We therefore are going to focus instead on the more recent attempts made to prepare genuine carbon nanotubes.

The synthesis of carbon nanotubes (either single- or multiwalled) by CCVD methods involves the catalytic decomposition of a carbon containing source on small metallic particles or clusters. This technique involves either an heterogeneous process if a solid substrate has a role or an homogeneous process if everything takes place in the gas phase. The metals generally used for these reactions are transition metals, such as Fe, Co, and Ni. It is a rather "low temperature" process compared to arc-discharge and laser-ablation methods, with the formation of carbon nanotubes typically occurring between 600 °C and 1,000 °C. Because of the low temperature, the selectivity of the CCVD method is generally better for the production of MWNTs with respect to graphitic particles and amorphous-like carbon, which remain an important part of the raw arc-discharge SWNT samples, for example. Both homogeneous and heterogeneous processes appear very sensitive to the nature and the structure of the catalyst used, as well as to the operating conditions. Carbon nanotubes prepared by CCVD methods are generally much longer (few tens to hundreds of micrometers) than those obtained by arc discharge (few micrometers). Depending on the experimental conditions, it is possible to grow dense arrays of nanotubes. It is a general statement that MWNTs from CCVD contain more structural defects (i. e., exhibit a lower nanotexture) than MWNTs from arc, due to the lower temperature reaction, which does not allow any structural rearrangements. These defects can be removed by subsequently applying heat treatments in vacuum or inert atmosphere to the products. Whether such a discrepancy is also true for SWNTs remains questionable. CCVD SWNTs are generally gathered into bundles that are generally of smaller diameter (few tens of nm) than their arc-discharge and laser-ablation counterparts (around 100 nm in diameter). CCVD provides reasonably good perspectives of large-scale and low-cost processes for the mass production of carbon nanotubes, a key point for their applications at the industrial scale.

A last word concerns the nomenclature. Because work in the field started more than one century ago, denomination of the carbon objects prepared by this method has changed with time with regard to the authors, research areas, and fashions. The same objects are found to be called vapor grown carbon fibers, nanofilaments, nanofibers, and from now on, nanotubes. Specifically regarding multilayered fibrous morphologies (since single-layered fibrous morphologies cannot be otherwise than SWNT anyway), the exact name should be vapor-grown carbon nanofilaments (VGCNF). Whether or not the filaments are then tubular is a matter of textural description, which should go with other textural features such as bamboo, herringbone, concentric, etc. (see Sect. 3.1.2). But the usage has decided otherwise. In the following, we will therefore use MWNTs for any hollowed nanofilament, whether they contain transversally oriented graphene walls or not. Any other nanofilament will be named "nanofiber."

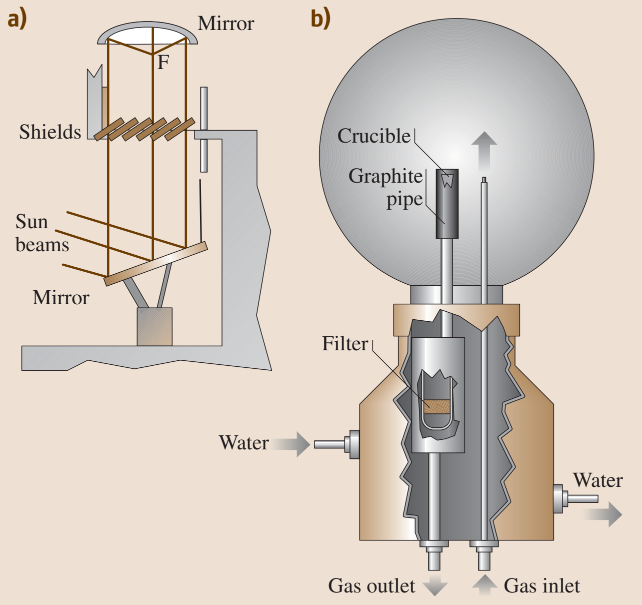

Fig. 3.14a,b Sketch of a solar energy reactor in use in Odeilho (France). **(a)** Gathering of sun rays, focused in F; **(b)** Example of Pyrex® chamber placed in **(a)** so that the graphite crucible is at the point F (adapted from [3.56])

Heterogeneous Processes

Heterogeneous CCVD processes are basically simple and consist, within a furnace heated to the desired temperature, in passing a gaseous flow containing a given proportion of a hydrocarbon (mainly CH_4, C_2H_2, C_2H_4, or C_6H_6, usually as a mixture with either H_2 or an inert gas such as Ar) over small transition metal particles (Fe, Co, Ni). The latter are previously deposited onto an inert substrate, for instance, by spraying a suspension of the metal particles on it, or any other method. The reaction is chemically defined as catalysis-enhanced thermal cracking

$$C_xH_y \rightarrow x\,C + y/2\,H_2 \,.$$

Catalysis-enhanced thermal cracking was used as early as the late 19th century (see beginning of chapter). Further extensive works before the 90s include those by Baker et al. [3.6, 59], or *Endo* et al. [3.60, 61]. Several review papers have been published since then, such as [3.62], in addition to many regular papers.

CO can be used instead of hydrocarbons, the reaction is then chemically defined as catalysis-enhanced disproportionation

$$2\,CO \rightarrow C + CO_2 \,.$$

Heterogeneous Processes – Experimental Devices

The ability of catalysis-enhanced CO disproportionation in making carbon nanofilaments was reported by *Davis* et al. [3.63] as early as 1953, probably for the first time. Extensive following works were performed by *Boehm* [3.64], *Audier* et al. [3.16, 65–67], and *Gadelle* et al. [3.68–71].

Although formation mechanisms for SWNTs and MWNTs can be quite different (see Sect. 3.3, or refer to a review article such as [3.72]), many of the parameters of the catalytic processes have a similar and important role on the type of nanotubes formed: the temperature, the duration of the treatment, the gas composition and flow rate, and of course the catalyst nature and size. At a given temperature, depending mainly on the nature of both the catalyst and the carbon-containing gas, the catalytic decomposition will take place at the surface of the metal particles, followed by a mass-transport of the freshly produced carbon either by surface or volume diffusion until the carbon concentration reaches the solubility limit, and the precipitation starts.

It is now agreed that the formation of CCVD carbon nanotubes occurs on metal particles of a very small size, typically in the nanometer-size range [3.72]. These catalytic metal particles are prepared mainly by reduction of transition metals compounds (salts, oxides) by H_2, prior to the nanotube formation step (where the carbon containing gas is required). It is possible, however, to produce these catalytic metal particles in situ in presence of the carbon source, allowing for a one-step process [3.73]. Because the control of the metal particle size is the key point (they have to be kept at a nanometric size), their coalescence is generally avoided by supporting them on an inert support such as an oxide (Al_2O_3, SiO_2, zeolites, $MgAl_2O_4$, MgO, etc.), or more rarely on graphite. A low concentration of the catalytic metal precursor is required to limit the coalescence of the metal particles that may happen during the reduction step.

There are two main ways for the preparation of the catalyst: (a) the impregnation of a substrate with a solution of a salt of the desired transition metal catalyst, and (b) the preparation of a solid solution of an oxide of the chosen catalytic metal in a chemically inert and thermally stable host oxide. The catalyst is then reduced to form the metal particles on which the catalytic decomposition of the carbon source will lead to carbon nanotube growth. In most cases, the nanotubes can then be separated from the catalyst (Fig. 3.15).

Heterogeneous Processes – Results with CCVD Involving Impregnated Catalysts

A lot of work was already done in this area even before the discovery of fullerenes and carbon nanotubes, but although the formation of tubular carbon structures by catalytic processes involving small metal particles was clearly identified, the authors did not focus on the preparation of SWNTs or MWNTs with respect to the other carbon species. Some examples will be given here to illustrate the most striking improvements obtained.

With the impregnation method, the process generally involves four different and successive steps: (1) the impregnation of the support by a solution of a salt (nitrate, chloride) of the chosen metal catalyst; (2) the drying and calcination of the supported catalyst to get the oxide of the catalytic metal; (3) the reduction in a H_2-containing atmosphere to make the catalytic metal particles and at last (4) the decomposition of a carbon-containing gas over the freshly prepared metal particles that will lead to the nanotube growth. For example, *Ivanov* et al. [3.74] have prepared nanotubes by the decomposition of C_2H_2 (pure or in mixture with H_2) on well-dispersed transition metal particles (Fe, Co, Ni, Cu) supported on graphite or SiO_2. Co-SiO_2 was found to be the best catalyst/support combination for the preparation of MWNTs, but most of the other combinations led to carbon filaments, sometimes covered with amorphous-like carbon. The same authors have developed a precipitation-ion-exchange method that provides a better dispersion of metals on silica compared to the classical impregnation technique. The same group has then proposed the use of a zeolite-supported Co catalyst [3.75, 76], resulting in very finely dispersed metal particles (from 1 to 50 nm in diameter). Only on this catalyst could they observe MWNTs with a diameter around 4 nm and only two or three walls. *Dai* et al. [3.77] have prepared SWNTs by CO disproportionation on nano-sized Mo particles. The diameters of the nanotubes obtained are closely related to that of the original particles and range from 1 to 5 nm. The nanotubes obtained by this method are free of amorphous carbon coating. It is also found that a synergetic effect occurs in the case of an alloy instead of the components alone, and one of the most striking examples is the addition of Mo to Fe [3.78] or Co [3.79].

Heterogeneous Processes – Results with CCVD Involving Solid-Solutions-Based Catalysts

A solid solution of two metal oxides is formed when some ions of one of the metals are found in substitution of the other metal ions. For example, Fe_2O_3 can be prepared in solid solution into Al_2O_3 to give a $Al_{2-2x}Fe_{2x}O_3$ solid solution. The use of a solid solution allows a perfect homogeneity of the dispersion of each oxide one in the other. These solid solutions can be prepared by different ways but the co-precipitation of mixed-oxalates and the combustion synthesis have been used mainly for the preparation of nanotubes. The synthesis of nanotubes by catalytic decomposition of CH_4 over $Al_{2-2x}Fe_{2x}O_3$ solid solutions was originated by *Peigney* et al. [3.73] and then studied extensively by the same group using different oxides such as spinel-based solid solutions ($Mg_{1-x}M_xAl_2O_4$ with M = Fe, Co, Ni, or a binary alloy [3.80, 81] or magnesia-based solid solutions [3.80, 82] ($Mg_{1-x}M_xO$, with M = Fe, Co or Ni)). Because of the very homogeneous dispersion of the catalytic oxide, it is possible to produce very small catalytic metal particles, at the high temperature required for the decomposition of CH_4 (which was chosen for its greater thermal stability compared to other hydrocarbons). The method proposed by these authors involves the heating of the solid solution from room temperature to a temperature between 850 °C and 1,050 °C in a mixture of H_2 and CH_4, typically containing 18 mol.% of CH_4. The nanotubes obtained depend clearly on the nature of both the transition metal (or alloy) used and the inert oxide (matrix), the latter because the Lewis acidity seems to play an important role [3.83]. For example, in the case of solid solutions containing around 10 wt% of Fe, the amount of carbon nanotubes obtained is decreasing in the following order depending of the matrixoxide: $MgO > Al_2O_3 > MgAl_2O_4$ [3.80]. In the case of MgO-based solid solutions the nanotubes can be very easily separated from the catalyst by dissolving it, in diluted HCl, for example [3.82]. The nanotubes obtained are typically gathered into small diameter bundles (less than 15 nm) with lengths up to 100 µm. The nanotubes are mainly SWNTs and DWNTs, with diameters ranging between 1 and 3 nm.

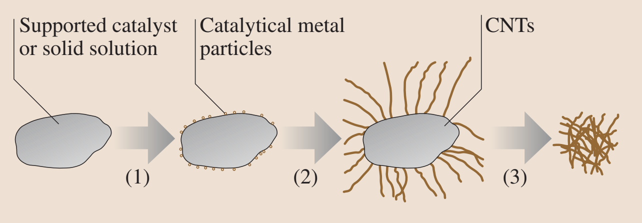
Supported catalyst or solid solution
Catalytical metal particles
CNTs
(1)
(2)
(3)

Fig. 3.15 Formation of nanotubes by the CCVD-based impregnation technique. (1) Formation of the catalytic metal particles by reduction of a precursor; (2) Catalytic decomposition of a carbon-containing gas, leading to the growth of carbon nanotubes (CNTs); (3) Removal of the catalyst to recover the CNTs (from [3.80])

Obtaining pure nanotubes by the CCVD method requires, as for all the other techniques, the removal of the catalyst. When a catalyst supported (impregnation) in a solid solution is used, the supporting – and catalytically inactive – oxide is the main impurity, both in weight and volume. When oxides such as Al_2O_3 or SiO_2 (or even combinations) are used, aggressive treatments involving hot caustic solutions (KOH, NaOH) for Al_2O_3 or the use of HF for SiO_2 are required. These treatments have no effect, however, on the other impurities such as other forms of carbon (amorphous-like carbon, graphitized carbon particles and shells, and so on). Oxidizing treatments (air oxidation, use of strong oxidants such as HNO_3, $KMnO_4$, H_2O_2) are thus required and allow for the removal of most of the unwanted forms of carbon but are resulting in a low yield of remaining carbon nanotubes, which are often quite damaged. *Flahaut* et al. [3.82] were the first to use a MgCoO solid solution to prepare SWNTs and DWNTs that could be easily separated without any damage by a fast and safe washing with an aqueous HCl solution.

In most cases, only very small quantities of catalyst (typically less than 500 mg) have been used, and most claims for "high yield" productions of nanotubes are based on laboratory experimental data, without taking into account all the technical problems related to the scaling up of a laboratory-scale CCVD reactor. At the present time, although the production of MWNTs is possible at an industrial scale, the production of SWNTs at an affordable cost is still a challenge.

Homogeneous Processes

The homogenous route, also called "floating catalyst method," differs from the other CCVD-based methods because it uses only gaseous species and does not require the presence of any solid phase in the reactor. The basic principle of this technique, similar to the other CCVD processes, is to decompose a carbon source (ethylene, xylene, benzene, carbon monoxide, etc.) on nanometric transition metallic (generally Fe, Co, or Ni) particles in order to obtain carbon nanotubes. The catalytic particles are formed directly in the reactor, however, and are not introduced before the reaction, as it happens in the supported CCVD, for instance.

Homogeneous Processes – Experimental Devices

The typical reactor used in this technique is a quartz tube placed in an oven to which the gaseous feedstock, containing the metal precursor, the carbon source, some hydrogen, and a vector gas (N_2, Ar, or He), is sent. The first zone of the reactor is kept at a lower temperature, and the second zone, where the formation of tubes occurs, is heated to 700–1,200 °C. The metal precursor is generally a metal-organic compound, such as a zero-valent carbonyl compound like $Fe(CO)_5$ [3.84], or a metallocene [3.85–87], for instance ferrocene, nickelocene or cobaltocene. It may be advantageous to make the reactor vertical, in order for the effect of gravity to be symmetrically dispatched on the gaseous volume inside the furnace and to help in maintaining for a while the solid products in fluidized bed.

Homogeneous Processes – Results

The metal-organic compound decomposes in the first zone of the reactor to generate the nanometric metallic particles that can catalyze the nanotubes formation. In the second part of the reactor, the carbon source is decomposed to atomic carbon which then is responsible for the formation of nanotubes.

This technique is quite flexible and both single-walled nanotubes [3.88] and multiwalled nanotubes [3.89] have been obtained depending on the carbon feedstock gas; it has also been exploited for some years in the production of vapor grown carbon nanofibers [3.90].

The main drawback of this type of process is again, as for heterogeneous processes, the difficulty to control the size of the metal nanoparticles, and thus the nanotube formation is often accompanied by the production of undesired carbon forms (amorphous carbon or polyaromatic carbon phases found as various phases or as coatings). In particular, encapsulated forms have been often found, as the result of the creation of metallic particles that are too big to be active for growing nanotubes (but are still effective for catalytically decomposing the carbon source) and be totally recovered by graphene layers.

The same kind of parameters as for heterogeneous processes have to be controlled in order to finely tune this process and obtain selectively the desired morphology and structure of the nanotubes formed, such as: the choice of the carbon source; the reaction temperature; the residence time; the composition of the incoming gaseous feedstock with a particular attention paid to the role played by hydrogen proportion, which can control the orientation of graphenes with respect to the nanotube axis thus switching from c-MWNT to h-MWNT [3.69]; and the ratio of the metallorganic precursor to the carbon source (for lower values, SWNTs are obtained [3.85]). As recently demonstrated, the overall process can be improved by adding other compounds such as ammonia or sulfur-containing species to the reactive gas phase. The former allows aligned nanotubes and mixed C-N nanotubes [3.91] to be obtained, the latter results in a significant increase in productivity [3.90].

It should be emphasized that only small productions have been achieved so far, and the scale up toward industrial exploitation seems quite difficult because of the large number of parameters that have to be considered. A critical one is to be able to increase the quantity of metallorganic compound that has to be sent in the reactor, as a requirement to increase the production, without obtaining too big particles. This problem has not yet been solved. An additional problem inherent in the process is the possibility of clogging the reactor due to the deposition of metallic nanoparticles on the reactor walls followed by carbon deposition.

A significant breakthrough concerning this technique could be the process developed at Rice University, the so-called HiPCo^{TM} process to produce SWNTs of very high purity [3.92]. This gas phase catalytic reaction uses carbon monoxide to produce, from $[Fe(CO)_5]$, a SWNT material claimed to be relatively free of by-products. The temperature and pressure conditions required are applicable to industrial plants. The company Carbon Nanotechnologies Inc. (Houston, TX, USA) actually sells raw SWNT materials prepared that way, at a marketed price of \$375/g, doubled if purified (2003 data). Other companies are more specialized in MWNTs such as Applied Sciences Inc. (Cedarville, Ohio, USA), which currently has a production facility of ~ 40 tons/year of ~ 100 nm large MWNTs (Pyrograf-III), or Hyperion Catalysis (Cambridge, MA, USA), which makes MWNT-based materials.

Templating

Another technique interesting to describe briefly, though definitely not suitable for mass production, is the templating technique. It is the second method only (the first being the electric-arc technique, when considering the formation of MWNTs on the cathode) able to synthesize carbon nanotubes without any catalyst. Any other work reporting the catalyst-free formation of nanotubes is likely to have been fooled by metallic impurities present in the reactor or by some other factors having brought a chemical gradient to the system. Another original aspect is that it allows aligned nanotubes to be obtained naturally, without the help of any subsequent alignment procedure. But recovering the nanotubes only requires the template to be removed (dissolved), which means the former alignment of the nanotubes is lost.

Templating – Experimental Devices

The principle of the technique is to deposit a solid carbon coating obtained from CVD method onto the walls of an appropriate porous substrate whose pores are displayed as parallel channels. The feedstock is again a hydrocarbon, as a common carbon source. The substrate can be alumina or zeolite for instance, however, which present natural channel pores, while the whole is heated to a temperature able to crack the hydrocarbon selected as carbon source (Fig. 3.16).

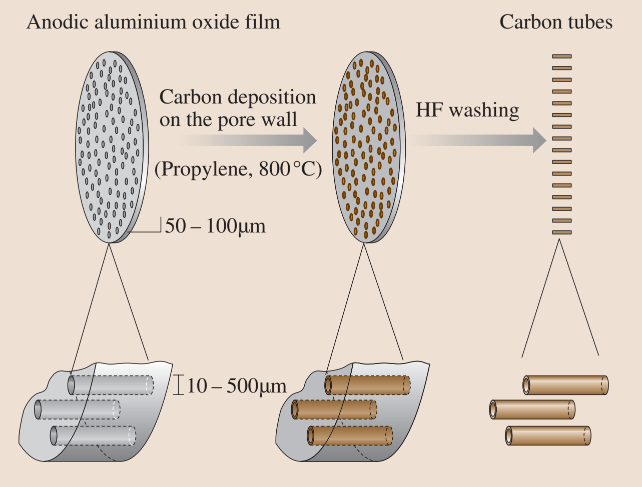
Anodic aluminium oxide film
Carbon tubes
Carbon deposition on the pore wall
(Propylene, 800 °C)
HF washing
50 – 100µm
10 – 500µm

Fig. 3.16 Principle of the templating technique for the catalyst-free formation of single-walled or concentric-type multi-walled carbon nanotubes (from [3.93])

Templating – Results

Provided the chemical vapor deposition mechanism (which is actually better described as a chemical vapor infiltration mechanism) is well controlled, the synthesis results in the coating of the channel pore walls by a variable number of graphenes. Both MWNTs (exclusively concentric type) or SWNTs can be obtained. The smallest SWNTs (diameters ∼ 0.4 nm) ever obtained mentioned in Sect. 3.1 have actually been synthesized by this technique [3.11]. Lengths are directly determined by channel lengths, i. e. by the thickness of the substrate plate. A main advantage is the purity of the tubes (no catalyst remnants, and little other carbon phases). On the other hand, the nanotube structure is not closed at both ends, which can be an advantage or a drawback, depending on the application. For instance, recovering the tubes requires the porous matrix to dissolve using one of the chemical treatment previously cited. The fact that tubes are open makes them even more sensitive to the acid attack.

3.2.3 Miscellaneous Techniques

In addition to the major techniques described in Sects. 3.2.1 and 3.2.2, many attempts can be found in literature to produce nanotubes by various ways, with a generally specific goal, such as looking for a low-cost or a catalyst-free production process. None has been sufficiently convincing so far to be presented as a serious alternative to the major processes described previously. Some examples are provided in the following.

Hsu et al. [3.94] have succeeded in preparing MWNTs (including coiled MWNTs, a peculiar morphology resembling a spring) by a catalyst-free (although Li was present) electrolytic method, by running a 3–5 A current between two graphite electrodes (a graphite crucible as the anode and a graphite rod as the cathode). The graphite crucible was filled with lithium chloride, while the whole was heated in air or argon at ∼ 600 °C. As with many other techniques, by-products such as encapsulated metal particles, carbon shells, amorphous carbon, and so on, are formed.

Cho et al. [3.95] have proposed a pure chemistry route, by the polyesterification of citric acid onto ethylen glycol at 50 °C, followed by a polymerization at 135 °C, then carbonized at 300 °C under argon, then oxidized at 400 °C in air. Despite the latter oxidation step, the solid product surprisingly contains short MWNTs, although obviously with a poor nanotexture. By-products such as carbon shells and amorphous carbon are also formed.

Li et al. [3.96] have also obtained short MWNTs by a catalyst-free (although Si is present) pyrolytic method which involves heating silicon carbonitride nanograins in a BN crucible to 1,200–1,900 °C in nitrogen within a graphite furnace. No details are given about the possible occurrence of by-products, but they are likely considering the complexity of the chemical system (Si-C-B-N) and the high temperatures involved.

Terranova et al. [3.98] have investigated the catalyzed reaction between a solid carbon source and atomic hydrogen. Graphite nanoparticles (∼ 20 nm) are sent with a stream of H_2 onto a Ta filament heated at 2,200 °C. The species produced, whatever they are, then hit a Si polished plate warmed to 900 °C and supporting transition metal particles. The whole chamber is under a dynamic vacuum of 40 torr. SWNTs are supposed to form according to the authors, although images are poorly convincing. One major drawback of the method, besides its complexity compared to others, is the difficulty of recovering the "nanotubes" from the Si substrates onto which they seem to be firmly bonded.

A last example is an attempt to prepare nanotubes by diffusion flame synthesis [3.99]. A regular gaseous hydrocarbon source (ethylene, . . .) added with ferrocene vapor is sent within a laminar diffusion flame made from air and CH_4 whose temperature range is 500–1,200 °C. SWNTs are actually formed, together with encapsulated metal particles, soot, and so on. In addition to a low yield, the SWNT structure is quite poor.

3.2.4 Synthesis of Aligned Carbon Nanotubes

Several applications (such as field-emission-based display, see Sect. 3.6) require that carbon nanotubes grow as highly aligned bunches, in highly ordered arrays, or located at specific positions. In that case, the purpose of the process is not mass production but controlled growth and purity, with subsequent control of nanotube morphology, texture, and structure. Generally speaking, the more promising methods for the synthesis of aligned nanotubes are based on CCVD processes, which involve molecular precursors as carbon source, and method of thermal cracking assisted by the catalytic activity of transition metal (Co, Ni, Fe) nanoparticles deposited onto solid supports. Few attempts have been made so far to obtain such materials with SWNTs. But the catalyst-free templating methods related to that described in Sect. 3.2.2 are not considered here, due to the lack of support after the template removal, which does not allow the former alignment to be maintained.

During the CCVD-growth, nanotubes can self-assemble into nanotube bunches aligned perpendicular to the substrate if the catalyst film on the substrate has a critical thickness [3.100], the driving forces for this alignment are the van der Waals interactions between the nanotubes, which allow them to grow perpendicularly to the substrates. If the catalyst nanoparticles are deposited onto a mesoporous substrate, the mesoscopic pores may also cause a certain effect on the alignment when the growth starts, thus controlling the growth direction of the nanotubes. Two kinds of substrates have been used so far in this purpose: mesoporous silica [3.101, 102] and anodic alumina [3.103].

Different methods have been reported in the literature for metal particles deposition onto substrates: (i) deposition of a thin film on alumina substrates from metallic-salt precursor impregnation followed by oxidation/reduction steps [3.104], (ii) embedding catalyst particles in mesoporous silica by sol-gel processes [3.101], (iii) thermal evaporation of Fe, Co, Ni, or Co-Ni metal alloys on SiO_2 or quartz substrates under high vacuum [3.105, 106], (iv) photolithographic patterning of metal-containing photoresist polymer with the aid of conventional black and white films as a mask [3.107] and, electrochemical deposition at the bottom of the pores in anodic aluminum oxide templates [3.103].

Depositing the catalyst nanoparticles onto a previously patterned substrate allows one to control the frequency of local occurrence and the display of the MWNT bunches formed. The as-produced materials mainly consist in arrayed, densely packed, freestanding, aligned MWNTs (Fig. 3.17), which is quite

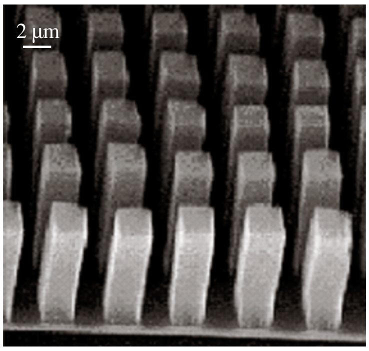

Fig. 3.17 Example of a free-standing MWNT array obtained from the pyrolysis of a gaseous carbon source over catalyst nanoparticles previously deposited onto a patterned substrate. Each square-base rod is a bunch of MWNTs aligned perpendicular to the surface (from [3.97])

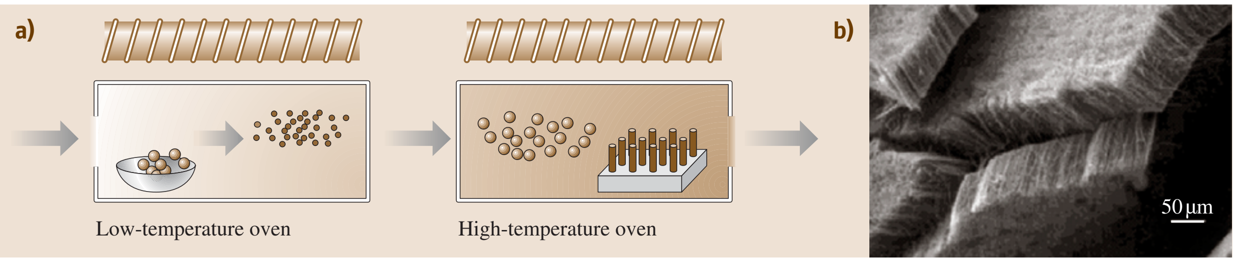

Fig. 3.18a–c Sketch of a double-furnace CCVD device for the organo-metallic/hydrocarbon co-pyrolysis process. In **(a)**: sublimation of the precursor; in **(b)**: decomposition of the precursor and MWNTs growth onto the substrate. **(c)** Example of the densely packed and aligned MWNT material obtained (from [3.108])

suitable for field emission-based applications for instance [3.97].

When a densely packed coating of vertically aligned MWNTs is desired (Fig. 3.18b), another route is the pyrolysis of hydrocarbons in the presence of organometallic precursor molecules like ferrocene or iron pentacarbonyl, operating in a dual furnace system (Fig. 3.18a). The organo-metallic precursor (e.g., ferrocene) is first sublimed at low temperature in the first furnace or is injected as a solution along with the hydrocarbon feedstock, then the whole components are pyrolyzed at higher temperature in the second furnace [3.88, 108–111]. The parameters that must be taken into account are the heating or feeding rate of ferrocene, the flow rate of the vector gas (Ar or N_2) and of the gaseous hydrocarbon, and the temperature of pyrolysis (650–1,050 °C). Generally speaking, the co-deposition process using [Fe(CO)$_5$] as catalyst source implies its thermal decomposition at elevated temperatures to produce atomic iron that deposits on the substrates in the hot zone of the reactor. As the nanotube growth occurs simultaneously with the introduction of [Fe(CO)$_5$], the temperatures chosen for the growth depend on the carbon feedstock utilized, e.g., they can vary from 750 °C for acetylene to 1,100 °C for methane. Mixtures of [FeCp$_2$] and xylene or [FeCp$_2$] and acetylene have also been used successfully for the production of freestanding MWNTs.

The nanotube yield and quality are directly linked with the amount and size of the catalyst particles, and since the planar substrates used do not exhibit high surface area, the dispersion of the metal can be a key step in the process. It has been observed that an etching pretreatment of the surface of the deposited catalyst thin film with NH_3 may be critical for an efficient growth of nanotubes by bringing the appropriate metal particle size distribution. It may also favor the alignment of MWNTs and prevent the formation of amorphous carbon from thermal cracking of acetylene [3.112].

3.3 Growth Mechanisms of Carbon Nanotubes

Growth mechanisms of carbon nanotubes are still debated. But researchers have been impressively imaginative and have made a number of hypotheses. One reason is that conditions allowing carbon nanofilaments to grow are very diverse, which means that related growing mechanisms are many. For given conditions, the truth is even probably a combination or a compromise between some of the proposals. Another reason is that phenomena are quite fast and difficult to observe in situ. It is generally agreed, however, that growth should occur so that the number of dangling bonds is limited, for energetic reasons.

3.3.1 Catalyst-Free Growth

As already mentioned, in addition to the templating technique which merely relates to chemical vapor infiltration mechanism for pyrolytic carbon, the growth of c-MWNT as a deposit onto the cathode in the electric-arc method is a rare example of catalyst-free carbon nanofilament growth. The driving forces are obviously related to the electric field, i. e. related to charge transfers from an electrode to the other via the particles contained in the plasma. How the MWNT nucleus forms is not clear, but once it has started, it may include the direct incorporation of C_2 species to the primary graphene structure as it was formerly proposed for fullerenes [3.113]. This is supported by recent C_2 radical concentration measurements that revealed an increasing concentration of C_2 from the anode being consumed to the growing cathode (Fig. 3.13), indicating, first, that C_2 are secondary species only, and second, that C_2 species could actually participate actively in the growth mechanism of c-MWNTs in the arc method.

3.3.2 Catalytically Activated Growth

Growth mechanisms involving catalysts are more difficult to ascertain, since they are more diverse. Although a more or less extensive contribution of a VLS (Vapor-Liquid-Solid [3.114]) mechanism is rather well admitted, it is quite difficult to find comprehensive and plausible explanations able to account for both the various conditions used and the various morphologies observed. What follows is an attempt to provide overall explanations of most of the phenomena, while remaining consistent with the experimental data. We did not consider hypotheses for which there are a lack of experimental evidence, such as the moving nano-catalyst mechanism, which proposed that dangling bonds from a growing SWNT may be temporarily stabilized by a nano-sized catalyst located at the SWNT tip [3.23], or the scooter mechanism, which proposed that C dangling bonds are temporarily stabilized by a single catalyst atom, moving all around the SWNT cross section edge, allowing subsequent C atom addition [3.115].

From the various results, it appears that the most important parameters are probably the thermodynamic conditions (only temperature will be considered here), the catalyst particle size, and the presence of a substrate. Temperature is critical and basically corresponds to the discrepancy between CCVD methods and solid-carbon-source based method.

Low Temperature Conditions

Low temperature conditions are typical from CCVD conditions, since nanotubes are frequently found to grow far below 1,000 °C. If conditions are such that the catalyst is a crystallized solid, the nanofilament is probably formed according to a mechanism close to a VLS mechanism, in which three steps are defined: (i) adsorption then decomposition of C-containing gaseous moieties at the catalyst surface; (ii) dissolution then diffusion of the C-species through the catalyst, thus forming a solid solution; (iii) back precipitation of solid carbon as nanotube walls. The texture is then determined by the orientation of the crystal faces relative to the filament axis (Fig. 3.19), as demonstrated beyond doubt by transmission electron microscopy images such as that in *Rodriguez* et al. [3.19]. This mechanism can therefore provide either c-MWNT, or h-MWNT, or platelet nanofiber.

If conditions are such that the catalyst is a liquid droplet, either because of higher temperature conditions or because of lower temperature melting catalyst, a mechanism similar to that above can still occur, really VLS (vapor = gaseous C species, liquid = molten catalyst, S = graphenes), but there are obviously no more crystal faces to orientate preferentially the rejected graphenes. Energy minimization requirements will therefore tend to make them concentric and parallel to the filament axis.

With large catalyst particles (or in the absence of any substrate), the mechanisms above will generally follow a "tip growth" scheme, i. e., the catalyst will move forward while the rejected carbon form the nanotube behind, whether there is a substrate or not. In that case,

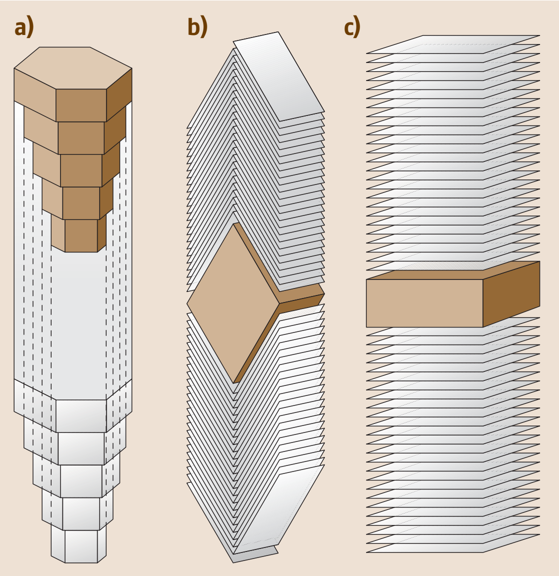

Fig. 3.19 Illustration of the relationships between the catalyst crystal outer morphology and the subsequent carbon nanofilament inner texture. Crystal are drawn from their images in the projected plane perpendicular to the electron beam in a transmission electron microscope, i. e. crystal morphology in the out-of-plane dimension is not ascertained (adapted from [3.19])

chances are high that one end is open. On the contrary, when catalyst particles deposited onto a substrate are small enough (nanoparticles) to be refrained to move by the interaction forces with the substrate, the growth mechanism will follow a "base growth" scheme, i. e. the carbon nanofilament grows away from the substrate leaving the catalyst nanoparticle attached to the substrate (Fig. 3.20).

The bamboo texture that affects both the herringbone and the concentric texture may reveal a specificity of the dissolution-rejection mechanism, i. e. a periodic, discontinuous dynamics of the phenomenon. Once the catalyst has reached the saturation threshold regarding its content in carbon, it expulses it quite suddenly. Then it becomes again able to incorporate a given amount of carbon without having any catalytic activity for a little while, then over-saturation is reached again, etc. Factors controlling such a behavior, however, are not determined.

It is then clear that, in any of the mechanisms above, 1 catalyst particle = 1 nanofilament. This explains why, although making SWNT by CCVD methods is possible, controlling the catalyst particle size is critical, since it thereby controls the subsequent nanofilament to be grown from it. Reaching a really narrow size distribution in CCVD is quite challenging, specifically as far as nanosizes are concerned when growing SWNTs is the goal. Only particles < 2 nm will be able to make it (Fig. 3.20), since larger SWNTs would energetically not be favored [3.10]. Another specificity of the CCVD method and from the related growth mechanisms is that the process can be effective all along the isothermal zone of the reactor furnace since continuously fed with a carbon-rich feedstock, which is generally in excess, with a constant composition at a given time of flight of the species. Coarsely, the longer the isothermal zone in gaseous carbon excess conditions, the longer the nanotubes. This is why the nanotube lengths can be much longer than that obtained by solid-carbon-source-based methods.

Table 3.2 provides guidelines for the relationships between the synthesis conditions and the type of nanotube formed.

High Temperature Conditions

High temperature conditions are typical from solid-carbon-source based methods such as electric arc, laser vaporization, or solar furnace (see Sect. 3.2). The huge temperatures involved (several thousands °C) atomize both the carbon source and the catalyst. Of course, catalyst-based SWNTs do not form in the area of highest temperatures (contrary to c-MWNT in the electric arc method), but the medium is made of atoms and radicals all mixed, some of which are likely to condense into the same droplet (since it will be liquid at the beginning). At some distance from the atomization zone, the medium is therefore made of carbon metal alloy droplets and of secondary carbon species that range from C_2 to higher order molecules such as corannulene which is made of a central pentagon surrounded by 5 hexagons. (The preferred formation of such a molecule can be explained by the former association of carbon atoms into a pentagon, because it is the fastest way to limit dangling bonds at low energetic cost, thereby providing a fixation site for other carbon atoms (or C_2) which also will tend to close into a ring, again to limit dangling bonds. Since adjacent pentagons are not energetically favoured, these cycles will be hexagons. Such a molecule is thought to be a probable precursor for fullerenes. Fullerenes are actually always produced, even in conditions which produce SWNTs.) The same event of saturation in C of the carbon-metal alloy droplet occurs as described in Sect. 3.3.2, resulting in the precipitation of excess C outside the particle upon the effect of the decreasing thermal gradient in the reactor, which decreases the solubility threshold of C in the metal [3.117]. Once the "inner" carbon atoms reach the catalyst particle surface, they meet the "outer" carbon species, including corannulene, that will con-

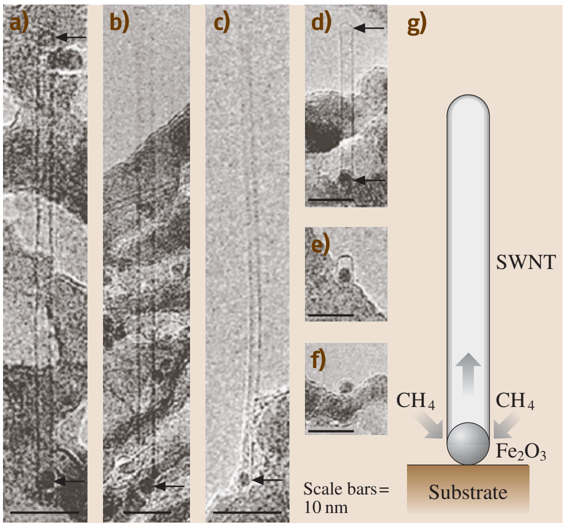

Fig. 3.20 High resolution transmission electron microscopy images of several SWNTs grown from iron-based nanoparticles by CCVD method, showing that particle sizes determine SWNT diameters in that case (adapted from [3.116])

Table 3.2 Guidelines indicating the relationships between possible carbon nanofilament morphologies and some basic synthesis conditions

		Increasing temperature . . . → . . . and physical state of catalyst			Substrate		Thermal gradient	
		Solid (crystallized)	**Liquid from melting**	**Liquid from clusters**	**Yes**	**No**	**Low**	**High**
Catalyst particle size	<∼ **3** nm	SWNT	SWNT	?	base-growth	tip-growth	long length	short length
	>∼ **3** nm	MWNT (c,h,b) platelet nanofiber	c-MWNT	SWNT	tip-growth			
Nanotube diameter		(heterogeneous related to catalyst particle size)		homogeneous (independent from particle size)				
Nanotube/particle		one nanotube/particle		several SWNTs/particle				

tribute to cap the merging nanotubes. Once formed and capped, nanotubes can grow both from the inner carbon atoms (Fig. 3.21a), according to the VLS mechanism proposed by *Saito* et al. [3.117], and from the outer carbon atoms, according the adatom mechanism proposed by *Bernholc* et al. [3.118]. In the latter, carbon atoms from the surrounding medium in the reactor are attracted then stabilized by the carbon/catalyst interface at the nanotube/catalyst surface contact, promoting their subsequent incorporation at the tube base. The growth mechanism therefore mainly follows the base growth scheme. But once the nanotubes are capped, C_2 species still remaining in the medium and which meet the growing nanotubes far from the nanotube/catalyst interface, may still incorporate the nanotubes from both the side wall or the tip, thereby giving rise to some proportion of Stone–Wales defects [3.40]. The occurrence of a nanometer-thick surface layer of yttrium carbide (onto the main Ni-containing catalyst core), whose some lattice distances are commensurable with that of the C−C distance in graphene, as recently revealed by *Gavillet* et al. [3.48], could possibly play a beneficial role in the stabilization effect at the nanotube/catalyst interface and explain why SWNT yield is enhanced by such bimetallic alloys (as opposed to single metal catalysts).

A major difference with the low temperature mechanisms described for CCVD methods is that many nanotubes are formed from a single, relatively large (∼ 10–50 nm) catalyst particle (Fig. 3.21b), whose size distribution is therefore not as critical as for the low

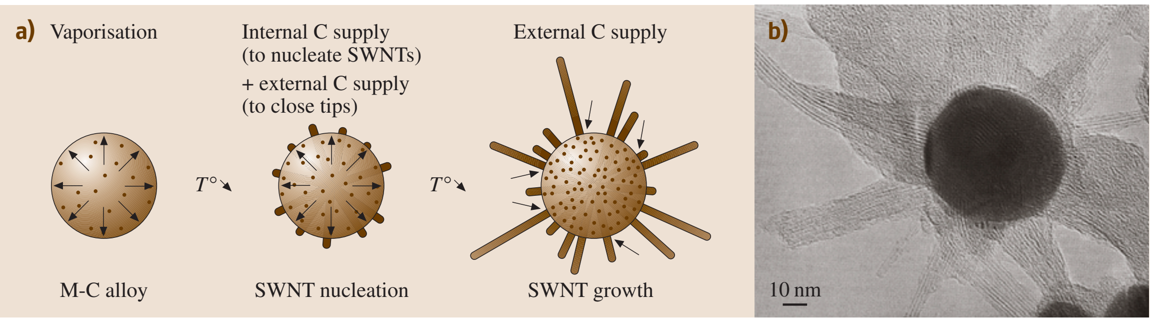

Fig. 3.21 **(a)** Mechanism proposed for SWNT growth (see text). **(b)** Transmission electron microscopy image of SWNT growing radial to a large Ni catalyst particle surface in the electric arc experiment. (modified from [3.17])

temperature mechanisms (particles that are too large, however, would induce polyaromatic shells instead of nanotubes). That is why SWNT diameters are much more homogeneous than for CCVD methods. Why the most frequent diameter is ∼ 1.4 nm is again a matter of energy balance. Single-wall nanotubes larger than ∼ 2.5 nm are not stable [3.10]. On the other hand, strain to the C−C angles increases as the radius of curvature decreases. Such diameter (1.4 nm) should therefore correspond to the best energetic compromise. Another difference with the low temperature mechanisms for CCVD is that temperature gradients are huge, and the gas phase composition surrounding the catalysts droplets is also subjected to rapid changes (as opposed to what could happen in a laminar flow of a gaseous feedstock whose carbon source is in excess). This explains why nanotubes from arc are generally shorter than nanotubes from CCVD, and why mass production by CCVD should be favored. In the latter, the catalytic role of the metallic particle can be recurrent as long as the conditions are maintained. In the former, surrounding conditions change continuously, and the window for efficient catalysis effect can be very narrow. Lowering temperature gradients in solid carbon source-based SWNT producing methods such as the electric-arc reactor should therefore be a way to increase the SWNT yield and length [3.119]. Amazingly, this is in opposition to fullerene production from arc.

3.4 Properties of Carbon Nanotubes

Carbon is a unique light atom that can form one-, two-, or threefold strong chemical bonds. The planar threefold configuration forms graphene planes that can, under certain growth conditions (see Sect. 3.3), adopt a tubular geometry. Properties of the so-called carbon nanotubes may change drastically depending on whether SWNTs or MWNTs are considered (see Sect. 3.1).

3.4.1 Variability of Carbon Nanotube Properties

Properties of MWNTs are generally not much different from that of regular polyaromatic solids (which may exhibit graphitic, or turbostratic, or intermediate crystallographic structure), and variations are then mainly driven by the textural type of the MWNTs considered (concentric, herringbone, bamboo) and the quality of the nanotexture (see Sect. 3.1), both of which control the extent of anisotropy. Actually, for polyaromatic solids, whose structural entities are built with stacked graphenes, bond strength is quite different depending on whether the in-plane direction is considered (which includes only very strong – covalent – and therefore very short – 0.142 nm bonds) or the direction perpendicular to it (which includes only very weak – van der Waals – and therefore very loose – ∼ 0.34 nm bonds). Such heterogeneity is not found in single (isolated) SWNTs. But the heterogeneity is back along with the related consequences when SWNTs associate into bundles. Therefore, properties – and applicability – for SWNTs may also change dramatically depending on whether single SWNT or SWNT ropes are involved.

In the following, we will emphasize SWNT properties, as far as their original structure often leads to original properties with respect to that of regular polyaromatic solids. But we will also sometimes cite properties of MWNTs for comparison.

3.4.2 General Properties

SWNT-type carbon nanotube diameters fall in the nanometer range and can be hundreds of micrometers long. SWNTs are narrower in diameter than the thinnest line able to be obtained by electron beam lithography. SWNTs are stable up to 750 °C in air (but start being damaged earlier through oxidation mechanisms, as demonstrated by their subsequent ability to be filled with molecules, see Sect. 3.5) and up to ∼ 1,500–1,800 °C in inert atmosphere beyond which they transform into regular, polyaromatic solids (i. e., phases built with stacked graphenes instead of single graphenes) [3.120]. They have half the mass density of aluminum. While the length of SWNTs can be macroscopic, the diameter has a molecular dimension. As a molecule, properties are closely influenced by the way atoms are displayed along the molecule direction. The physical and chemical behavior of SWNTs are therefore related to their unique structural features [3.121].

3.4.3 SWNT Adsorption Properties

An interesting feature of SWNTs is their very high surface area, the highest ever due to the fact that a single graphene sheet is probably the unique example of a material energetically stable in normal conditions while consisting of a single layer of atoms. Ideally, i.e. not considering SWNTs in bundles but isolated SWNTs, and provided the SWNTs have one end opened (by oxidation treatment for instance), the real surface area can be equal to that of a single, flat graphene, i.e. $\sim 2{,}700\,\mathrm{m^2/g}$ (accounting for both sides).

But practically, nanotubes – specifically SWNTs – are more often associated with many other nanotubes to form bundles, then fibers, films, papers, etc. than as a single entity. Such architectures develop various porosity ranges important for determining adsorption properties (the latter aspects are also developed in Sect. 3.6.2, focused on applications). It is thus most appropriate to discuss adsorption on SWNT-based materials in term of adsorption on the outer or inner surfaces of such bundles. Furthermore, theoretical calculations have predicted that the molecule adsorption on the surface or inside of the nanotube bundle is stronger than that on an individual tube. A similar situation exists for MWNTs where adsorption could occur either on or inside the tubes or between aggregated MWNTs. Additionally, it has been shown that the curvature of the graphene sheets constituting the nanotube walls can result in a lower heat for adsorption with respect to that of a planar graphene (see Sect. 3.1.1).

Accessible SWNT Surface Area

Various studies dealing with the adsorption of nitrogen on MWNTs [3.122, 123] and SWNTs [3.124] have highlighted the porous nature of these two materials. Pores in MWNTs can be divided mainly into inner hollow cavities of small diameter (narrowly distributed, mainly 3–10 nm) and aggregated pores (widely distributed, 20–40 nm), formed by interaction of isolated MWNTs. It is also worth noting that the ultra-strong nitrogen capillarity in the aggregated pores contributes to the major part of the total adsorption, showing that the aggregated pores are much more important than the inner cavities of the MWNTs for adsorption. Adsorption of N_2 has been studied on as-prepared and acid-treated SWNTs, and the results obtained point out the microporous nature of SWNT materials, as opposed to the mesoporous MWNT materials. Also, as opposed to isolated SWNTs (see above), surface areas well above $400\,\mathrm{m^2g^{-1}}$ have been measured for SWNT-bundle-containing materials, with internal surface areas of $300\,\mathrm{m^2g^{-1}}$ or higher.

The theoretical specific surface area of carbon nanotubes ranges over a broad scale, from 50 to $1{,}315\,\mathrm{m^2g^{-1}}$ depending on the number of walls, the diameter, or the number of nanotubes in a bundle of SWNTs [3.125]. Experimentally, the specific surface area of SWNTs is often larger than that of MWNTs. Typically, the total surface area of as-grown SWNTs ranges between 400 and $900\,\mathrm{m^2g^{-1}}$ (micropore volume, $0.15–0.3\,\mathrm{cm^3g^{-1}}$), whereas values ranging between 200 and $400\,\mathrm{m^2g^{-1}}$ for as-produced MWNTs are often reported. In the case of SWNTs, the diameter of the tubes and the number of tubes in the bundle will affect mainly the BET value. It is worth noting that opening/closing of the central canal noticeably contributes to the adsorption properties of nanotubes. In the case of MWNTs, chemical treatments such as KOH activation are efficient to develop a microporosity, and surface areas as high as $1{,}050\,\mathrm{m^2g^{-1}}$ have been reported [3.126]. Thus it appears that opening or cutting as well as chemical treatments (purification step for example) of carbon nanotubes can considerably affect their surface area and pore structure.

Adsorption Sites and Binding Energy of the Adsorbates

An important problem to solve when considering adsorption on nanotubes is the identification of the adsorption sites. Adsorption of gases in a SWNT bundle can occur inside the tubes (pore), in the interstitial triangular channels between the tubes, on the outer surface of the bundle, or in a groove formed at the contact between adjacent tubes on the outside of the bundle (Fig. 3.22). Starting from closed-end SWNTs, simple molecules can adsorb on the walls of the outer nanotubes of the bundle and preferably on the external grooves. For the more attractive sites, corresponding to the first adsorption stages, it seems that adsorption or condensation in the interstitial channel of SWNT bundles depends on the size of the molecule (or on the SWNT diameters) and on their interaction energies [3.127]. Then, adapted treatments to open the tubes will favour the gas adsorption (e.g., O_2, N_2 in the inner walls [3.128, 129]). For hydro-

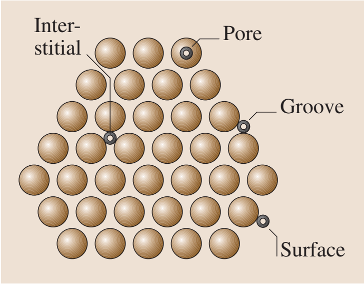

Fig. 3.22 Sketch of a SWNT bundle illustrating the four different adsorption sites

gen and other small molecules, computational methods have shown that, for open SWNTs, the pore, interstitial, and groove sites are energetically more favourable than the surface site, with respect to the increasing number of carbon nanotubes interacting with the adsorbed molecules [3.130].

For MWNTs, adsorption can occur in the aggregated pores, inside the tube or on the external walls. In the latter case, the presence of defects, as incomplete graphene layers, has to be taken into consideration. Although adsorption between the graphenes (intercalation) has been proposed in the case of hydrogen adsorption in h-MWNTs or platelet nanofibers [3.132], it is unlikely for many molecules due to steric effect and should not prevail for small molecules due to long diffusion path.

Few studies deal with adsorption sites in MWNTs, however it has been shown that butane adsorbs more in the case of MWNTs with smaller outside diameters, which is consistent with another statement that the strain brought to curved graphene surfaces affects sorption. Most of the butane adsorbs to the external surface of the MWNTs while only a small fraction of the gas condensed in the pores [3.133]. Comparative adsorption of krypton or of ethylene on MWNTs or on graphite has allowed scientists to determine the dependence of the adsorption and wetting properties on the specific morphology of the nanotubes. Higher condensation pressure and lower heat of adsorption were found on nanotubes with respect to graphite [3.134]. These differences result mainly from a decrease in the lateral interactions between the adsorbed molecules, related to the curvature of the graphene sheets.

A limited number of theoretical as well as experimental works exist on the values of the binding energies for gases on carbon nanotubes. If most of these studies report low binding energies on SWNTs, consistent with a physisorption, some experimental results, in particular in the case of hydrogen, are still controversial (see Sect. 3.6.2). In the case of platelet nanofibers, the initial dissociation of hydrogen on graphite edge sites, which constitute the majority of the nanofiber surface has been proposed [3.135]. For carbon nanotubes, a mechanism that involves H_2 dissociation on residual metal catalyst followed by H spillover and adsorption on the most reactive nanotube sites could be envisaged [3.136]. Doping nanotubes with alkali may enhance hydrogen adsorption, due to the charge transfer from the alkali metal to the nanotube, which polarizes the H_2 molecule and induces dipole interaction [3.137].

Generally speaking, the adsorbates can be either charge donors or acceptors to the nanotubes. The observed trends in the binding energies of gases with different van der Waals radii suggest that the groove sites of SWNTs are the preferred low coverage adsorption sites due to their higher binding energies. Finally, several studies have shown that, at low coverage, the binding energy of the adsorbate on SWNT is between 25% and 75% higher than the binding energy on a single graphene. This discrepancy can be attributed to an increase of the effective coordination in the binding sites, such as the groove sites, in SWNTs bundles [3.138, 139]. Representative results concerning the adsorption properties of SWNTs and MWNTs are summarized on Table 3.3.

Table 3.3 Adsorption properties and sites of SWNTs and MWNTs. Letters in column 5 refer to Fig. 3.22. Data of two last columns are from [3.131]

Type of nanotube	Porosity (cm^3g^{-1})	Surface area (m^2/g)	Binding energy of the adsorbate	Adsorption sites	Attractive potential per site (eV)	Surface area per site (m^2/g)
SWNT (bundle)	Microporous V_{micro}: 0.15–0.3	400–900	Low, mainly physisorption 25–75 % > graphite	Surface (A)	0.049	483
				Groove (B)	0.089	22
				Pore (C)	0.062	783
				Interstitial (D)	0.119	45
MWNT	Mesoporous	200–400	Physisorption	Surface Pore Aggregated pores	–	–

3.4.4 Transport Properties

The narrow diameter of SWNTs has a strong influence on its electronic excitations due to its small size compared to the characteristic length scale of low energy electronic excitations. Combined with the particular shape of the electronic band structure of graphene, carbon nanotubes are ideal quantum wires. The conduction and the valence band of the graphene touch on the 6 corners (K points) of the Brillouin zone [3.140]. By rolling the graphene to form a tube, new periodic boundary conditions for the electronic wave functions superimpose, which give rise to one-dimensional sub-bands: $C_n K = 2q$ where q is an integer and C_n the vector $na_1 + ma_2$ characteristic of the diameter and the helicity of the tube (see Sect. 3.1). If one of these sub-bands passes through one of the K points, the nanotube will be metallic, otherwise it will be semiconductor with a gap inversely proportional with the diameter. For a tube with a diameter of 3 nm, the gap falls in the range of the thermal energy at room temperature. As pointed out in Sect. 3.1 already, knowing (n, m) then allows one to predict whether or not the tube exhibits a metallic behavior. To summarize, we can say that the electronic structure of SWNTs depends both on the orientation of the honeycomb lattice with respect to the tube axis and the radius of curvature imposed to the bent graphene sheet. This explains why properties of MWNTs quickly become no different from regular, polyaromatic solids as the number of walls (graphenes) increases, since the radius of curvature increases meanwhile. In addition, the conduction occurs essentially through the external tube (for c-MWNTs), but the interactions with the internal coaxial tubes may make the electronic properties vary. For small diameter MWNTs exhibiting a low number of walls, typically two (DWNTs), the relative lattice orientation of the two superimposed graphenes remains determinant, and the resulting overall electronic behavior can either be metallic or semiconductor upon structural considerations [3.141]. Another consequence, valid for any nano-sized carbon nanotube whatever the number of walls, is that applying stresses to the tube is also likely to change its electronic behavior, as a consequence of the variation brought to the position of the sub-bands with respect to the K points of the Brillouin zone [3.142].

The details of electronic properties of SWNTs are not well understood yet, and numerous theoretical and experimental works have revealed fascinating effects [3.143]. An assembly of agglomerated tubes (ropes) increases the complexity. Moreover, investigating the electron properties of a single SWNT requires one to characterize and control the properties of the electrical contacts. Likewise, the interaction with the substrate influence the bulk transport properties a lot when SWNTs embedded in a medium (e.g., a composite) are considered. The conduction of charge carriers in metallic SWNTs is thought to be ballistic (i.e., independent from length) due to the expected small number of defects [3.144–146] with a predicted electric conductance twice the fundamental conductance unit $G_0 = 4e/h$ because of the existence of two propagating modes. Because of the reduced scattering, metallic SWNTs can transport huge current densities (max 10^9 A/cm^2) without being damaged, i.e. about three orders of magnitude higher than in Cu. On the other hand, in a one-dimensional system, electrons often localize due to structural defects or disorder. In case the contact resistance is too high, a tunnel barrier is formed, and the charge transport is dominated by tunneling effects. Several transport phenomena were reported in the literature of this regime: weak localization due to quantum interferences of the electronic wave functions that leads to an increase of the resistivity [3.147–149], Coulomb blockade at low temperature that characterizes by conductance oscillations as the gate voltage increases [3.148], superconductivity induced by superconductor contacts [3.150], or spin polarization effects induced by ferromagnetic contacts [3.151]. Furthermore, due to enhanced Coulomb interactions, one-dimensional metals are expected to show drastic changes in the density of states at the Fermi level and are described by a Luttinger liquid [3.148]. This behavior is expected to result in the variation of conductance vs. temperature that follows a power law, with a zero conductance at low temperature.

As mentioned previously, SWNTs are model systems to study one-dimensional charge transport phenomena. Due to the of mesoscopic dimension of the morphology, the one-dimensional structure of the nanotube has strong singularities in the electronic density of states (DOS) that fall in the energy range of visible light. As a consequence, visible light is strongly absorbed. It has been observed that flash illumination with a broadband light can lead to spontaneous burning of a macroscopic sample of agglomerated (i.e., ropes) carbon nanotubes in air and room temperature [3.152].

As a probable consequence of both the small number of defects (at least the kind of defects that oppose phonon transport) and the cylindrical topography, SWNTs exhibit a large phonon mean free path, which results in a high thermal conductivity. The thermal conductivity of SWNTs is comparable to that of a single, isolated graphene layer or high purity diamond [3.153] or possibly higher (~ 6,000 W/mK).

Finally, carbon nanotubes may also exhibit a positive or negative magneto-resistance depending on the current, the temperature, and the field, whether they are MWNTs [3.154] or SWNTs [3.149].

3.4.5 Mechanical Properties

While tubular nano-morphology is also observed for many two-dimensional solids, carbon nanotubes are unique through the particularly strong threefolded bonding (sp^2 hybridization of the atomic orbitals) of the curved graphene sheet, which is stronger than in diamond (sp^3 hybridization) as revealed by their difference in C−C bond length (0.142 vs. 0.154 nm for graphene and diamond respectively). This makes carbon nanotubes – SWNTs or c-MWNTs – particularly stable against deformations. The tensile strength of SWNTs can be 20 times that of steel [3.155] and has actually been measured equal to ~ 45 GPa [3.156]. Very high tensile strength values are also expected for ideal (defect-free) c-MWNTs, since combining perfect tubes concentrically is not supposed to be detrimental to the overall tube strength, provided the tube ends are well capped (otherwise, concentric tubes could glide relative to each other, inducing high strain). Tensile strength values as high as ~ 150 GPa have actually been measured for perfect MWNTs from electric arc [3.157], although the reason for such a high value compared to that measured for SWNTs is not clear. It probably reveals the difficulty in carrying out such measurements in a reliable manner. At least, flexural modulus, for perfect MWNTs should logically exhibit higher values than SWNTs [3.155], with a flexibility that decreases as the number of walls increases. On the other hand, measurements performed on defective MWNTs obtained from CCVD gave a range of 3–30 GPa [3.158]. Tensile modulus, also, reaches the highest values ever, 1 TPa for MWNTs [3.159], and possibly even higher for SWNTs, up to 1.3 TPa [3.160, 161]. Figure 3.23 illustrates how spectacularly defect-free carbon nanotubes could revolutionize the current panel of high performance fibrous materials.

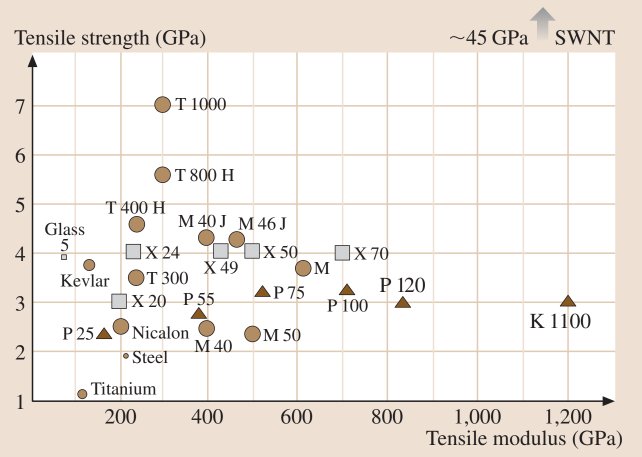

Fig. 3.23 Plot of the tensile strength versus tensile modulus for current fibrous materials, as compared to SWNTs. *Large circles* are PAN-based carbon fibers, which include the highest tensile strength fiber available on the market (T1000 from Torayca); *Triangles* are pitch-based carbon fibers, which include the highest tensile modulus fiber on the market (K1100 from Amoco)

3.4.6 Reactivity

The chemical reactivity of graphite, fullerenes, and carbon nanotubes exhibits some common features. Like any small object, carbon nanotubes have a large surface with which they can interact with their environment (see Sect. 3.4.1). It is worth noting, however, that the chemistry of nanotubes chemistry differs from that of regular polyaromatic carbon materials because of their unique shape, small diameter, and structural properties. Compared to graphite, perfect SWNTs have no chemically active dangling bonds (the reactivity of polyaromatic solids is known to occur mainly through graphene edges). Compared to fullerenes, the ratio of “weak” sites (C−C bonds involved in heterocycles) over strong sites (C−C bonds between regular hexagons) is only slightly different from 0 for ideally perfect tubes (as opposed to 1 in C_{60} fullerenes – C_{60} molecules have 12 pentagons (therefore accounting for $5 \times 12 = 30$ C−C bonds) and 20 hexagons, each of them having three C−C bonds not involved in an adjacent pentagon but shared with a neighboring hexagon (therefore accounting for $20 \times 3 \times 1/2 = 30$ C−C bonds involved in hexagons only). Although graphene faces are chemically relatively inert, the radius of curvature imposed on the graphene in nanotubes enforces the three formerly planar C−C bonds characteristic of the genuine sp^2 hybridization to accept distortions in bond angles that make them closer and closer to the situation of three of the four C−C bonds in diamond (characteristic of the genuine sp^3 hybridization) as the radius of curvature decreases. Even though it is not enough to bring a genuine chemical reactivity to the carbon atoms, a consequence is to consider that either nesting sites are created at the concave surface, or strong physisorption sites are created above each carbon atom of the convex surface, both with a bonding efficiency that increases as the nanotube diameter decreases.

As already pointed out in Sect. 3.1, the chemical reactivity of SWNTs (and c-MNWTs) is supposed to come mainly from the caps, since they contain six pentagons each, as opposed to the tube body, which supposedly contains hexagons only. Indeed, oxidizing treatments applied to carbon nanotubes (air oxidation, wet-chemistry oxidation) are known to open the nanotube tips rather selectively [3.162]. But the ability of SWNTs to be opened by oxidation methods then filled with foreign molecules such as fullerenes (see Sect. 3.5) reveals the occurrence of side defects [3.14], whose identification and occurrence were discussed then proposed to be an average of one Stone–Wales defect every 5 nm along the tube length, involving an amount of about 2% of the carbon atoms for the example of a regular (10,10) SWNT [3.163] – a Stone–Wales defect is made of four adjacent heterocycles, two pentagons and two heptagons, displayed by pairs in opposition. Such a defect actually allows localized double bonds to be formed between the carbon atoms involved in the defect (instead of the electrons participating to the delocalized electron cloud above the graphene as usual, enhancing the chemical reactivity, e.g. toward chlorocarbenes [3.163]). This means that the overall chemical reactivity of carbon nanotubesshould strongly depend on the way they are synthesized. For example, SWNTs prepared by the arc-discharge method are supposed to contain fewer structural defects compared to CCVD-synthesized SWNTs, which have a higher chemical reactivity. Of course, the reactivity of h-MWNT-type nanotubes is intrinsically higher, due to the occurrence of accessible graphene edges at the nanotube surface.

3.5 Carbon Nanotube-Based Nano-Objects

3.5.1 Hetero-Nanotubes

Hetero-nanotubes are defined here as carbon nanotubes whose carbon atoms, part or all of them, are substituted with another element while the overall honeycomb lattice-based graphene structure is maintained. Elements involved are, therefore, boron and/or nitrogen. Benefits can be new behaviors (e.g., BN nanotubes are electrical insulators), improved properties (e.g., regarding the resistance to oxidation), or better control of the properties. One current challenge is actually to control the processing so that the desired SWNT structure is selectively synthesized (metallic or semiconductor). In this regard, it was demonstrated that substituting some C atoms by N or B atoms lead to SWNTs whose electrical behavior is systematically metallic [3.164, 165].

Some examples of hetero-nanotubes, though mainly MWNTs, are found in literature, whose hetero-atom involved is generally nitrogen due to the ease of introducing some gaseous or solid nitrogen and/or boron containing species (e.g., N_2, NH_3, BN, HfB_2) in existing MWNT synthesis devices [3.164, 166] until complete substitution of carbon [3.167, 168]. An amazing result of such attempts to synthesize hetero-MWNTs was the subsequent formation of "multilayered" c-MWNTs, i. e. MWNTs whose constituting coaxial tubes were alternatively made of carbon graphenes and boron nitride graphenes [3.169]. On the other hand, examples of hetero-SWNTs are few. But the synthesis of B- or N-containing SWNTs has been recently reported [3.165, 170].

3.5.2 Hybrid Carbon Nanotubes

Hybrid carbon nanotubes are here defined as carbon nanotubes, SWNTs or MWNTs, whose inner cavity has been filled, partially or entirely, by foreign atoms, molecules, compounds, or crystals. Such hybrids are thereby noted as X@SWNT (or X@MWNT, if appropriate, where X is the atom, molecule, etc. involved) [3.171].

Motivation

Why filling carbon nanotubes? The very small inner cavity of nanotubes is an amazing tool to prepare and study the properties of confined nanostructures of any nature, such as salts, metals, oxides, gases, or even discrete molecules like C_{60}, for example. Because of the almost one-dimensional structure of carbon nanotubes (specifically considering SWNTs), different physical and/or chemical properties can be expected for the encapsulated foreign materials or possibly for the overall hybrid materials. Indeed, when the volume available inside a carbon nanotube is small enough, the foreign material can be made mainly of "surface atoms" of reduced coordinence. The overall former motivation was then to tentatively form metal nanowires likely to be interesting for electronic applications (quantum wires). Metals were the preferred fillers, and nanotubes were mostly considered merely as nano-molds, likely to be removed afterward. But it is likely that the removal of the SWNT "container" to liberate the confined one-dimensional structure may destroy or at least transform it because of the stabilizing effect brought by the interactions with the nanotube wall.

Filling nanotubes while they grow (in situ filling) was one of the pioneering methods. In most cases, however, the filling step is separated from the step of synthesis of the nanotubes. Three methods can then be distinguished: (a) wet chemistry procedures, and physical procedures, involving (b) capillarity filling by a molten material or (c) filling by a sublimated material.

Generally speaking, the estimation of the filling rate is problematic and most of the time is taken merely from the TEM observation, without any statistics on the number of tubes observed. Moreover, as far as SWNTs are concerned, the fact that the nanotubes are gathered into bundles makes difficult the observation of the exact number of filled tubes, as well as the estimation of the filled length for each tube.

Filling carbon nanotubes with materials that could not have been introduced directly is also possible. This can be accomplished by first filling the nanotubes with an appropriate precursor (i. e., able to sublime, or melt, or solubilize) that will later be transformed by chemical reaction or by a physical interaction such as electron beam irradiation, for example [3.172]. In the case of secondary chemical transformation, the reduction by H_2 is the most frequently used to obtain nanotubes filled with metals [3.173]. Sulfides can be obtained as well by using H_2S as reducing agent [3.173].

Because the inner diameter of SWNTs is generally smaller than that of MWNTs, it is more difficult to fill them, and the driving forces involved in the phenomena are not yet totally understood (see the review paper by [3.163]). Consequently, more than one hundred articles dealing with filled MWNTs were already published by the time this chapter was written. We, therefore, did not attempt to review all the works. We chose to cite the pioneering works, then to focus on the more recent works dealing with the more challenging topic of filling SWNTs.

In Situ Filling Method

Most of the first hybrid carbon nanotubes ever synthesized were directly obtained in the course of their processing. This concerns only MWNTs, prepared by the arc-discharge method in its early hours, for which the filling materials were easily introduced in the system by drilling a central hole in the anode and filling it with the heteroelements. First attempts were all reported the same year [3.174–178], with elements such as Pb, Bi, YC_2, TiC, etc. Later on, *Loiseau* et al. [3.179] showed that MWNTs could also be filled to several μm in length by elements such as Se, Sb, S, and Ge, but only as nanoparticles with elements such as Bi, B, Al and Te. Sulfur was suggested to play an important role during the in situ formation of filled MWNTs by arc-discharge [3.180]. This technique is no longer the preferred one because of the difficulty in controling the filling ratio and yield and in achieving mass production.

Wet Chemistry Filling Method

The wet chemistry method requires the opening of the nanotubes tips by chemical oxidation prior to the filling step. This is generally obtained by refluxing the nanotubes in diluted nitric acid [3.181–183], although other oxidizing liquid media may work as well, e.g. $[HCl + CrO_3]$ [3.184] or chlorocarbenes formed from the photolytic dissociation of $CHCl_3$ [3.163] as a rare example of a nonacidic liquid route for opening SWNTs. If a dissolved form (salt, oxide, etc.) of the desired metal is meanwhile introduced during the opening step, some of it will get inside the nanotubes. An annealing treatment (after washing and drying of the treated nanotubes) may then lead to the oxide or to the metal, depending on the annealing atmosphere [3.162]. Although the wet chemistry method looked promising because a wide variety of materials can be introduced within nanotubes this way, and because it operates at temperature not much different from room temperature, attention has to be paid to the oxidation method that has to be carried out as the first steps. The damages brought to the nanotubes by severe treatments (e.g., nitric acid) make them improper for use with SWNTs. Moreover, the filling yield is not very important, as a probable consequence of the presence of the solvent molecules that also enter the tube cavity, and the subsequent filled lengths rarely exceed 100 nm. Recent results have been obtained, however, using wet chemistry filling by *Mittal* et al. [3.184] who have filled SWNTs ropes by CrO_3 with an average yield of $\sim 20\%$.

Molten State Filling Method

The physical method involving a liquid (molten) phase is more restrictive, first, because some materials may start decomposing when they melt, and second because the melting point has to be compatible with the nanotubes, i. e. thermal treatment temperature should remain below the temperature of transformation or damaging of the nanotubes. Because the filling is occurring by capillarity, the surface tension threshold of the molten material was found to be 100–200 N/cm^2 [3.185]. But this threshold was proposed for MWNTs, whose inner diameters (5–10 nm) are generally larger than those of SWNTs (1–2 nm). In a typical filling experiment, the MWNTs are closely mixed with the desired amount of filler by gentle grinding, and the mixture is then vacuum-sealed in a silica ampoule. The ampoule is then slowly heated to a temperature above the melting point of the filler, then is slowly cooled. The use of this method does not require any opening of the nanotubes prior to the heat treatment. The mechanism of nanotube opening is not yet established clearly, but it is certainly related to the chemical aggressivity of the molten materials toward carbon and more precisely toward graphene defects in the tube structure (see Sect. 3.4.4).

Most of the works on SWNTs involving this method were performed by Sloan and coworkers at Oxford University [3.187], although other groups followed the same procedure [3.86, 181, 183]. Precursors to fill the nanotubes were mainly metals halides. Although little is known yet about the physical properties of halides crystallized within carbon nanotubes, the crystallization itself of molten salts within small diameter SWNTs has been studied in detail, revealing a strong interaction between the one-dimensional crystals and the surrounding graphene wall. For example, *Sloan* et al. [3.188] have described two-layer 4 : 4 coordinated KI crystals formed within SWNTs of around 1.4 nm diameter. These two-layer crystals are "all surface" and have no "internal" atoms. Significant lattice distortions occur compared to the bulk structure of KI, where the normal coordinence is 6 : 6 (meaning that each ion is surrounded by six identical closer neighbors). Indeed, the distance between two ions across the SWNT capillary is 1.4 times the same distance along the tube axis. This denotes an accommodation of the KI crystal to the confined space provided by the inner nanotube cavity in the constrained crystal direction (across the tube axis). This implies that the interactions between the ions and the surrounding carbon atoms are strong. The volume dimension available within the nanotubes thus somehow controls the crystal structure of inserted materials. For instance, the structure and orientation of encapsulated PbI_2 crystals inside their capillaries were found to arrange in different ways inside SWNTs and DWNTs, with respect to the diameter of the confining nanotubes [3.186]. In the case of SWNTs, most of the obtained encapsulated one-dimensional PbI_2 crystals exhibit a strong preferred orientation with their (110) planes aligning at an angle of ca. 60° to the SWNT axes as shown in Fig. 3.24a and 3.24b. Due to the extremely small diameter of the nanotube capillaries, individual crystallites are often only a few polyhedral layers thick, as outlined in Fig. 3.24d to 3.24h. As a result of lattice terminations enforced by capillary confinement, the edging polyhedra must be of reduced coordination, as indicated in Fig. 3.24g and 3.24h. In the case of PbI_2 formed inside DWNTs, similar crystal growth behavior was generally observed to occur in narrow nanotubes with diameters comparable to that of SWNTs. As the diameter of the encapsulating capillaries increases, however, different preferred orientations are frequently observed (Fig. 3.25). In this example, the PbI_2 crystal is oriented with the [121] direction parallel to the direction of the electron beam (Fig. 3.25a to 3.25d). If the PbI_2@DWNT hybrid is viewed "side-on" (as indicated by the arrow in Fig. 3.25e) polyhedral slabs are seen to arrange along the capillary, oriented at an angle of ca. 45° with respect to the tubule axis.

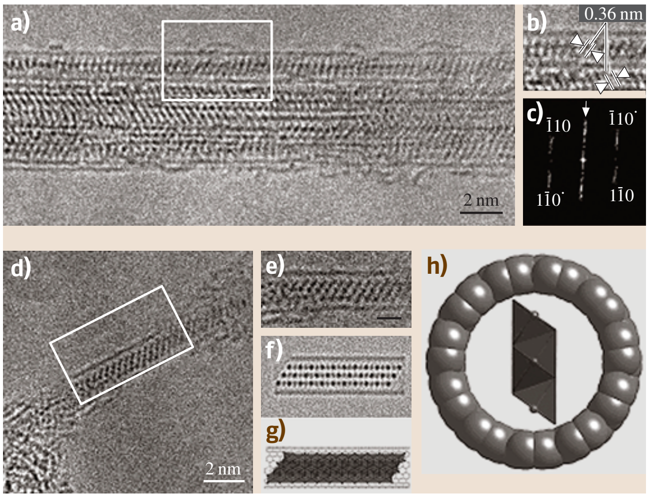

Fig. 3.24 HRTEM images and corresponding structural model for PbI_2 filled SWNTs (from [3.186])

Sublimation Filling Method

This method is even more restrictive than the previous one, since it is applicable only to a very limited number of compounds due to the requirement for the filling materials to sublimate within the temperature range of thermal stability of the nanotubes. Examples are therefore few. Actually, except for an attempt to fill SWNTs with $ZrCl_4$ [3.189], the only example published so far is the formation of C_{60}@SWNT (whose popular nickname is “peapods”), reported for the first time in 1998 [3.190], where regular ∼ 1.4 nm large SWNTs are filled with C_{60} fullerene molecule chains (Fig. 3.26a). Of course, the process requires a pre-opening of the SWNTs by some method, as discussed previously, typically either acid attack [3.191] or heat-treatment in air [3.192]. Then the opened SWNTs are put into a sealed glass tube together with fullerene powder and placed into a furnace heated above the sublimation temperature for fullerite ($>\sim 350\,°C$). Since there is no limitation related to Laplace’s law or the presence of solvent since only gaseous molecules are involved, filling efficiency may actually reach ∼ 100% [3.192].

Remarkable behaviors have been revealed since then, such as the ability of the C_{60} molecules to move freely within the SWNT cavity (Fig. 3.26b and 3.26c) upon random ionization effects from electron irradiation [3.193], or to coalesce into 0.7 nm large inner elongated capsules upon electron irradiation [3.194], or into a 0.7 nm large nanotube upon subsequent thermal treatment above 1,200 °C [3.193, 195]. Annealing of C_{60}@SWNT materials may thus appear as an efficient way to produce DWNTs with nearly constant inner (∼ 0.7 nm) and outer (∼ 1.4 nm) diameters. DWNTs produced this way are actually the smallest MWNTs synthesized so far.

Based on the successful synthesis of so-called endofullerenes previously achieved [3.13], further development of the process has recently led to the synthesis of more complex nanotube-based hybrid materials such as La_2@C_{80}@SWNTs [3.196], Gd@C_{82}@SWNTs [3.197], or $Er_xSc_{3-x}N$@C_{80}@SWNT [3.198] among other examples. These provide even more extended perspective regarding the potential applications of peapods, which are still being investigated.

Finally the last example to cite is the successful attempt to produce peapods by a related method, by using accelerated fullerene ions (instead of neutral gaseous molecules) to enforce the fullerenes to enter the SWNT structure [3.199].

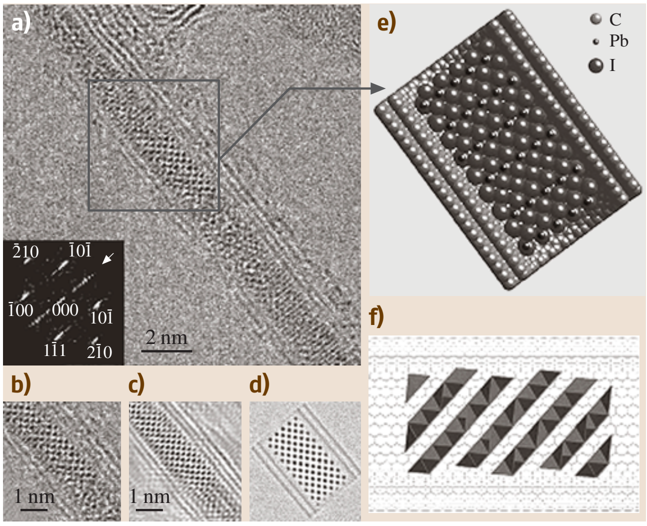

Fig. 3.25 HRTEM images (experimental and simulations), and corresponding structural model for a PbI_2 filled double-wall carbon nanotube (from [3.186])

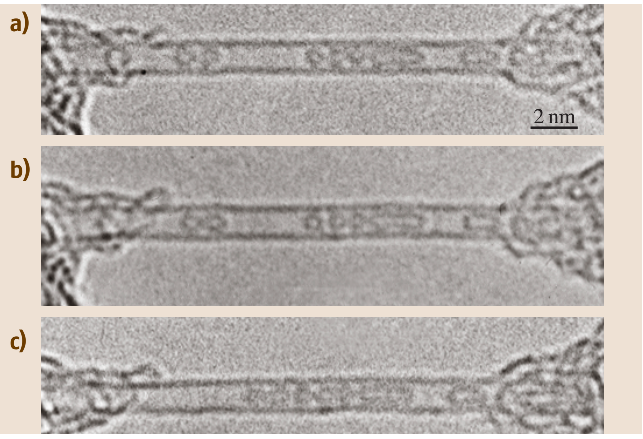

Fig. 3.26a–c HRTEM images of **(a)** example of five regular C_{60} molecules encapsulated together with two higher fullerenes (C_{120} and C_{180}) as distorted capsules (*on the right*) within a regular 1.4 nm diameter SWNT. **(a)**–**(c)** Example of the diffusion of the C_{60}-molecules along the SWNT cavity. The time between each image of the sequence is about 10 s. The fact that nothing occurs between **(a)** and **(b)** reveals the randomness of the ionisation events generated by the microscope electron beam and assumed to be responsible for the molecule displacement

3.5.3 Functionalized Nanotubes

Based on the reactivity of carbon nanotubes as discussed in Sect. 3.4.6, the functionalization reactions can be divided into two main groups. One is based on the chemical oxidation of the nanotubes (tips, structural defects) leading to carboxylic, carbonyl, and/or hydroxyl functions. These functions are then used for additional reactions, to attach oligomeric or polymeric functional entities. The second group is based on direct addition to the graphitic-like surface of the nanotubes (i. e. without any intermediate step). Examples of the latter reactions include oxidation or fluorination (an important first step for further functionalization with other organic groups). Properties and applications of functionalized nanotubes have been reviewed recently [3.200].

Oxidation of Carbon Nanotubes

Carbon nanotubes are often oxidized and therefore opened before chemical functionalization to increase their chemical reactivity (creation of dangling bonds). Chemical oxidation of nanotubes is mainly achieved using either wet chemistry or gaseous oxidants such as oxygen (typically air) or CO_2. Depending on the synthesis process, the oxidation resistance of the nanotubes can vary. When oxidation is carried out through gas phases, thermo-gravimetric analysis (TGA) is of great use to determine at which temperature the treatment should be applied. It is important to know that TGA accuracy increases as the heating rate diminishes, while the literature often provides TGA analyses obtained in nonoptimized conditions, leading to overestimated oxidation temperatures. The presence of catalyst remnants (metals, more rarely oxides), the type of nanotubes (SWNTs, c-MWNTs, h-MWNTs), and the oxidizing agent (air, O_2 is an inert gas, CO_2, etc.) as well as the flow rate used make difficult the comparison of published results. It is generally agreed, however, that amorphous carbon burns first, followed by SWNTs and then multiwall materials (shells, MWNTs), even if TGA is often unfortunately not able to clearly separate the different oxidation steps. The more common method used for nanotube oxidation is that of aqueous solutions of oxidizing reagents. The main one is nitric acid, either concentrated or diluted (around 3 moles per liter in most cases), but oxidants such as potassium dichromate ($K_2Cr_2O_7$), hydrogen peroxide (H_2O_2), or potassium permanganate ($KMnO_4$) are very often used as well. HCl, like HF, do not damage nanotubes because these acids are not oxidizing.

Functionalization of Oxidized Carbon Nanotubes

Carboxylic groups located at the nanotube tips can be coupled with different chemical groups. Oxidized nanotubes are usually reacted with thionyl chloride ($SOCl_2$) to generate the acyl chloride, even if a direct reaction is theoretically possible with alcohols or amines, for example. Reaction of SWNTs with octadecylamine (ODA) was reported by *Chen* et al. [3.201] after reaction of oxidized SWNTs with $SOCl_2$. The functionalized SWNTs have substantial solubility in chloroform ($CHCl_3$), dichloromethane (CH_2Cl_2), aromatic solvents, and carbon bisulfide (CS_2). Many other reactions between functionalized nanotubes (after reaction with $SOCl_2$) and amines have been reported in the literature and will not be reviewed here. Noncovalent reactions between the carboxylic groups of oxidized nanotubes and octadecylammonium ions are possible [3.202], providing solubility in tetrahydrofuran (THF) and CH_2Cl_2. Functionalization by glucosamine using similar procedures [3.203] allowed the water solubilization of SWNTs, which is of special interest when considering biological applications of functionalized nanotubes. Functionalization with lipophilic and hydrophilic dendra (with long alkyl chains and oligomeric poly(ethyleneglycol) groups) has been achieved by amination and esterification reactions [3.204], leading to solubility of the functionalized nanotubes in hexane, chloroform, and water. It is interesting to note that in the latter case, the functional groups could be removed just by modifying the pH of the solution (base- and acid-catalyzed hydrolysis reaction conditions, [3.205]). A last example is the possible interconnection of nanotubes via chemical functionalization. This has been recently achieved by *Chiu* et al. [3.206] using the already detailed acyl chloride way and a bifunctionalized amine to link the nanotubes by the formation of amide bonds.

Sidewall Functionalization of Carbon Nanotubes

The covalent functionalization of the nanotube walls is possible by means of fluorination reactions. It was first reported by *Mickelson* et al. [3.207], with F_2 gas (the nanotubes can then be defluorinated, if required, with anhydrous hydrazine). As recently reviewed by *Khabashesku* et al. [3.208], it is then possible to use these fluorinated nanotubes to carry out subsequent derivatization reactions. Thus, sidewall-alkylated nanotubes can be prepared by nucleophilic substitution (Grignard synthesis or reaction with alkyllithium precursors [3.209]). These alkyl sidewall groups can be removed by air oxidation. Electrochemical addition of aryl radicals (from the reduction of aryl diazonium salts) to nanotubes has also been reported by *Bahr* et al. [3.210]. Functionalization of the nanotube external wall by cycloaddition of nitrenes, addition of nuclephilic carbenes, or addition of radicals has been described by *Holzinger* et al. [3.211]. Electrophilic addition of dichlorocarbene to SWNTs occurs by reaction with the deactivated double bonds in the nanotube wall [3.212]. Silanization reactions are another way to functionalize nanotubes, although only tested with MWNTs. *Velasco-Santos* et al. [3.213] have reacted oxidized MWNTs with an organosilane ($RsiR'_3$, with R being an organo-functional group attached to silicon) and obtained organo-functional groups attached to the nanotubes by silanol groups.

Noncovalent sidewall functionalization of nanotubes is important because the covalent bonds are associated with changes from sp^2 to sp^3 in the hybridization of the carbon atoms involved, which corresponds to a loss of the graphite-like character. As a consequence, physical properties of functionalized nanotubes, specifically SWNTs, may be modified. One possibility for achieving noncovalent functionalization of nanotubes is to wrap the nanotubes in a polymer [3.214], which allows their solubilization (thus improving the processing possibilities) while preserving their physical properties.

Finally, it is worth keeping in mind that all these chemical reactions are not specific to nanotubes and may occur to most of the carbonaceous impurities present in the raw materials. This makes difficult the characterization of the functionalized samples and thus requires one to perform the experiments starting with very pure carbon nanotube samples, which is unfortunately not always the case for the results reported in the literature. On the other hand, purifying the nanotubes to start with may also bias the functionalization experiments since purification requires chemical treatments. But the demand for such products does exist already, and purified then fluorinated SWNTs can be found on the market at $ 900/g (Carbon Nanotechnologies Inc., 2003).

3.6 Applications of Carbon Nanotubes

Carbon nanotubes can be inert and can have a high aspect ratio, high tensile strength, low mass density, high heat conductivity, large surface area, and versatile electronic behavior including high electron conductivity. While these are the main characteristics of the properties for individual nanotubes, a large number of them can form secondary structures such as ropes, fibers, papers, thin films with aligned tubes, etc., with their own specific properties. The combination of these properties makes them ideal candidates for a large number of applications provided their cost is sufficiently low. The cost of carbon nanotubes depends strongly on both the quality and the production process. High quality single shell carbon nanotubes can cost 50–100 times more than gold. But synthesis of carbon nanotubes is constantly improving, and sale prizes fall by 50% per year. Application of carbon nanotubes is therefore a very fast moving area with new potential applications found every year, even several times per year. Making an exhaustive list is again a challenge we will not take. Below are listed applications considered "current" (Sect. 3.6.1) either because already marketed, or because up-scaling synthesis is being processed, or because prototypes are currently developed by profit-based companies. Other applications are said to be "expected" (Sect. 3.6.2).

3.6.1 Current Applications

High mechanical strength of carbon nanotubes makes them a near to ideal force sensor in scanning probe microscopy (SPM) with a higher durability and the ability to image surfaces with a high lateral resolution, the latter being a typical limitation of conventional force sensors (ceramic tips). The idea was first proposed and experimented by Dai et al. [3.77] using c-MWNTs. It was extended to SWNTs by *Hafner* et al. [3.215], since small diameters of SWNTs were supposed to bring higher resolution than MWNTs due to the extremely short radius of curvature of the tube end. But commercial nanotube-based tips (e.g., Piezomax, Middleton, WI, USA) use MWNTs for processing convenience. It is also likely that the flexural modulus of SWNTs is too low and therefore induces artifacts affecting the lateral resolution when scanning a rough surface. On the other hand, the flexural modulus of c-MWNTs is supposed to increase with the number of walls, but the radius of curvature of the tip meanwhile increases.

Near-Field Microscope Probes

SWNT or MWNT, such nanotube-based SPM tips also offer the perspective of being functionalized, in the prospect of making selective images based on chemical discrimination by "chemical force microscopy" (CFM). Chemical function imaging using functionalized nanotubes represents a huge step forward in CFM because the tip can be functionalized very accurately (ideally at the very nanotube tip only, where the reactivity is the highest), increasing the spatial resolution. The interaction between chemical species present at the end of the nanotube tip and a surface containing chemical functions can be recorded with great sensitivity, allowing the chemical mapping of molecules [3.216, 217].

Today the cost of nanotube-based SPM tips is in the range of $ 200/tip. Such a high cost is explained both

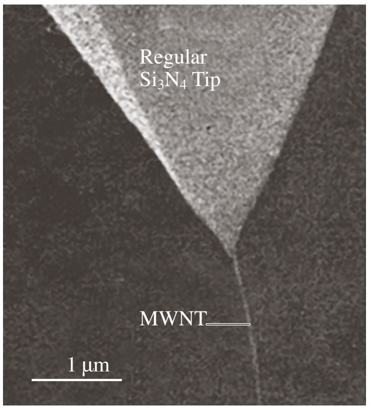

Fig. 3.27 Scanning electron microscopy image of a carbon nanotube (MWNT) mounted onto a regular ceramic tip as a probe for atomic force microscopy

by the processing difficulty, which ideally requires one to grow or mount a single MWNT in the appropriate direction at the tip of a regular SPM probe (Fig. 3.27), and by the need to individually control the tip quality. Such a market has been estimated at ∼ $ 20 M/year.

Field Emission-Based Devices

Based on a pioneering work by *de Heer* et al. [3.218], carbon nanotubes have been demonstrated to be efficient field emitters and are currently being incorporated in several applications, including flat panel display for television sets or computers (whose a first prototype was exhibited by Samsung in 1999) or any devices requiring an electron producing cathode, such as X-ray sources. The principle of a field-emission-based screen is demonstrated in Fig. 3.28a. Briefly, a potential difference is brought between emitting tips and an extraction grid so that electrons are taking out from the tips and sent onto an electron sensitive screen layer. Replacing the glass support and protection of the screen by some polymer-based material will even allow the develope of flexible screens. As opposed to regular (metallic) electron emitting tips, the structural perfection of carbon nanotubes allows higher electron emission stability, higher mechanical resistance, and longer life time. First of all, it allows energy savings since it needs lower (or no) heating temperature of the tips and requires much lower threshold voltage. As an illustration for the latter, producing a current density of 1 mA/cm^2 is possible for a threshold voltage of 3 V/µm with nanotubes, while it requires 20 V/µm for graphite powder and 100 V/µm for regular Mo or Si tips. The subsequent reductions in cost and energy consumption are estimated at 1/3 and 1/10 respectively. Generally speaking, the maximum current density obtainable ranges between 10^6 A/cm^2 and 10^8 A/cm^2 depending on the nanotubes involved (e.g., SWNT or MWNT, opened or capped, aligned or not) [3.219–221]. Although nanotube side-walls seem to emit as well as nanotube tips, many works have dealt (and are still dealing) with growing nanotubes perpendicular to the substrate surface as regular arrays (Fig. 3.28b). Besides, using SWNTs instead of MWNTs does not appear necessary for many of these applications when they are used as bunches. On the other hand, when considering single, isolated nanotubes, SWNTs are generally less preferable since they allow much lower electron doses than do MWNTs, although they are likely meanwhile to provide an even higher coherent source (an useful feature for devices such as electron microscopes or X-ray generators).

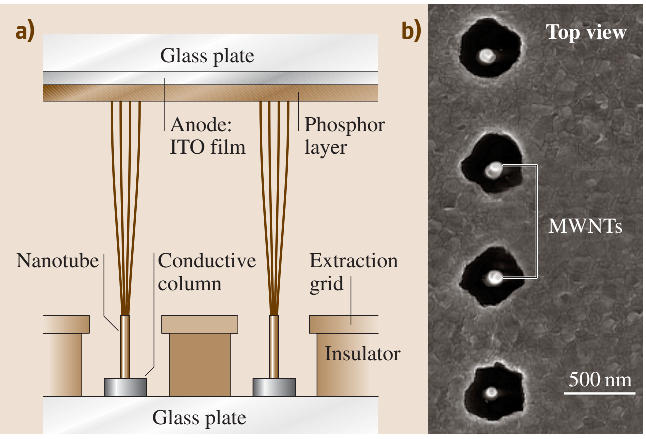

Fig. 3.28 **(a)** Principle of a field-emitter-based screen. **(b)** Scanning electron microscope image of a nanotube-based emitter system (*top view*). *Round dots* are MWNT tips seen through the holes corresponding to the extraction grid. By courtesy of *Legagneux* (Thales Research & Technology, Orsay, USA)

The market associated with this application is huge. With such major companies involved as Motorola, NEC, NKK, Samsung, Thales, Toshiba, etc., the first commercial flat TV sets and computers using nanotube-based screens are about to be seen in stores, and companies such as Oxford Instruments and Medirad are about to commercialize miniature X-ray generators for medical applications on the basis of nanotube-based cold cathodes developed by Applied Nanotech Inc.

Chemical Sensors

The electrical conductance of semiconductor SWNTs was recently demonstrated to be highly sensitive to the change in the chemical composition of the surrounding atmosphere at room temperature, due to the charges transfer between the nanotubes and the molecules from the gases adsorbed onto the SWNT surface. It has also been shown that there is a linear dependence between the concentration of the adsorbed gas and the difference in electrical properties, and that the adsorption is reversible. First tries involved NO_2 or NH_3 [3.222] and O_2 [3.223]. SWNT-based chemical NO_2 and NH_3 sensors are characterized by extremely short response time (Fig. 3.29), thus being different from conventionally used sensors [3.222, 224]. High sensitivity toward water or ammonia vapors has been measured on SWNT-SiO_2 composite [3.225]. This study indicated the presence of p-type SWNTs dispersed among the predominant metallic SWNTs, and that chemisorption of gases on the surface of the semiconductor SWNTs is responsible for the sensing action. The determination of CO_2 and O_2 concentrations on SWNT-SiO_2 composite has also been reported [3.226]. By doping nanotubes with palladium nanoparticles, *Kong* et al. [3.227] have also shown that the modified material could reveal the presence of hydrogen up to 400 ppm whereas the as-grown material was totally ineffective.

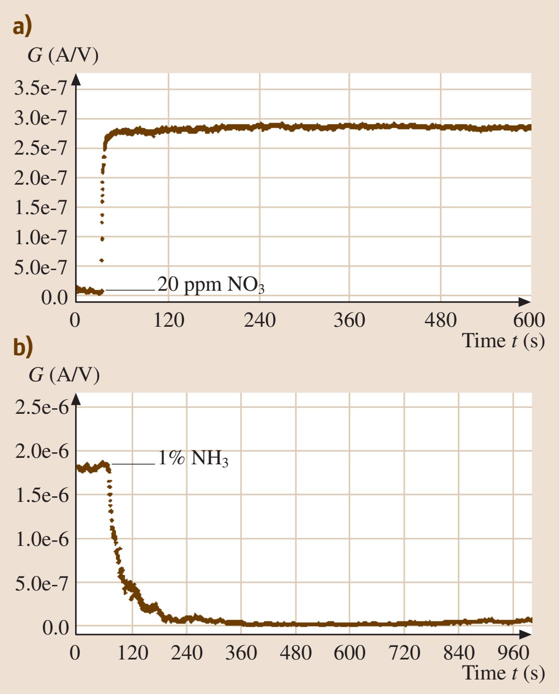

Fig. 3.29a,b Demonstration of the ability of SWNTs in detecting molecule traces in inert gases. **(a)** Increase in a single SWNT conductance when 20 ppm of NO_2 are added to an argon gas flow. **(b)** Same with 1% NH_3 added to the argon gas flow (from [3.222])

Generally speaking, the sensitivity of the new nanotube-based sensors is three orders of magnitude higher than that of standard solid state devices. In addition, the interest in using nanotubes as opposed to current sensors is the simplicity and the very small size of the system in which they can be placed, the fact that they can operate at room temperature, and their selectivity, which allows a limited number of sensor device architectures to be built for a variety of industrial purposes, while the current technology requires a large diversity of devices based on mixed metal oxide, opto-mechanics, catalytic beads, electrochemistry, etc. The market opportunity is expected to be $ 1.6 billion by 2006, including both the biological area and the chemical industrial area. Nanotube-based sensors are currently developed in both large and small companies, such as Nanomix (USA), for example.

Catalyst Support

Carbon materials are attractive supports in heterogeneous catalytic processes due to their ability to be tailored to specific needs: indeed, activated carbons are already currently employed as catalyst supports because of their high surface area, their stability at high temperatures (under not oxidizing atmosphere), and the possibility of controlling both their porous structure and the chemical nature of their surface [3.228, 229]. The attention has been brought to nano-sized fibrous morphologies of carbon that appeared during the last decade, showing novel interesting potentialities as supports [3.230]. Carbon nanofibers (also improperly called graphite nanofibers) and carbon nanotubes have been used successfully and shown to present, as catalyst-supporting materials, properties superior to that of such other regular catalyst-supports as activated carbon, soot, or graphite [3.231]. The use of graphite nanofibers as a direct catalyst for oxidative dehydrogenation has also been reported [3.232].

The morphology and size of carbon nanotubes, especially since they do present huge lengths vs. diameters (= aspect ratio), can play a significant role in catalytic applications due to their ability to disperse catalytically active metal particles. Their electronic properties are also of primary importance [3.233] since this conductive support may be exerting electronic perturbation as well as geometric constraint on the dispersed metal particles. A recent comparison between the interaction of transition metal atoms with carbon nanotube walls and that with graphite indicates major differences in bonding sites, magnetic moments, and charge-transfer direction [3.234]. Thus the possibility of strong metal-support interaction has to be taken into account. Their mechanical strength is also important and makes them resistant to attrition in view of recycling. Their external or internal surface presents strong hydrophobicity and strong adsorption ability toward organic molecules. Finally, for MWNT-based catalyst-supports, the relatively high surface area and the absence of microporosity (i.e., pores < 2 nm) associated with a high meso- and macropore volume (see Sect. 3.4.3) result in significant improvements of the catalytic activity, in the case of liquid-phase reactions when compared to catalysts supported on activated carbon. With nanotube supports, the mass transfer limitation of the reactants to the active sites is actually prevented due to the absence of microporosity, and the apparent contact time of the products with the catalyst is diminished, leading to more active and more selective catalytic effects.

The most widely used technique to prepare carbon nanotube-supported catalysts is the incipient wetness impregnation, in which the purified support is impregnated with a solution of the metal precursor, then dried, calcinated and/or reduced in order to obtain metal particles dispersed on the support. The chemical treatment and/or the modification of the carbon nanotube surface was found to be a useful tool to control its hydrophobic or hydrophilic character [3.236]. A strong metal/support interaction can thus be expected from the occurrence of functionalized groups created by the oxidation of the support surface, resulting in smaller particle sizes [3.237]. A more sophisticated technique to achieve the grafting of metal particles onto carbon nanotubes consists of functionalizing the outer surface of the tubes and then performing a chemical reaction with a metal complex, thus resulting in a good dispersion of the metallic particles (Fig. 3.30) [3.235].

Selected examples of some carbon nanotube-based catalysts together with the related preparation routes and catalytic activities are listed in Table 3.4.

The market is huge for this application since it most often concerns the "heavy" chemical industry. It implies and requires mass production of low-cost nanotubes, processed by methods other than solid-carbon-source-based methods (see Sect. 3.2.1). Such an application also requires some surface reactivity, making the h-MWNT type nanotubes with a poor nanotexture (see Sect. 3.1.2) the most suitable starting material for preparing such catalyst supports. Catalysis-enhanced thermal cracking of gaseous carbon precursors is therefore preferred, and pilot plants are already being built by major chemical industrial companies (e.g., Ato-Fina, France).

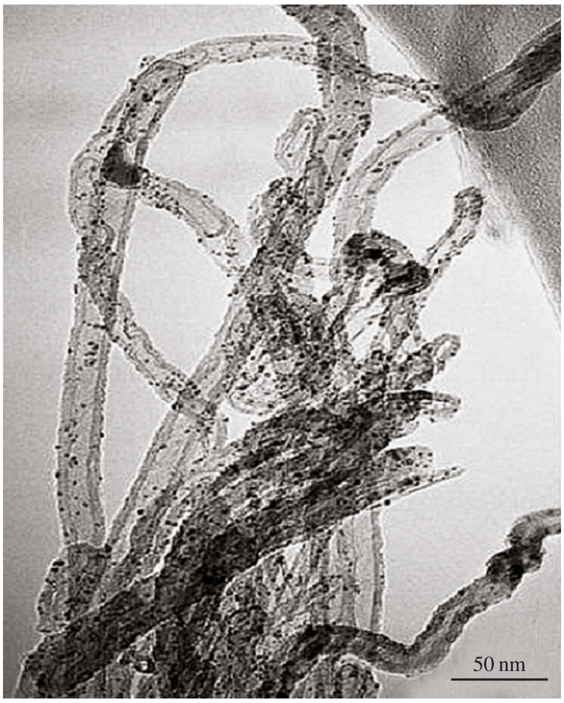

Fig. 3.30 Transmission electron microscopy image showing rhodium nanoparticles supported on MWNT surface (from [3.235])

3.6.2 Expected Applications Related to Adsorption

The adsorption and interaction of gases, liquids, or metal with carbon nanotubes has attracted much attention in the past several years. The applications resulting from the adsorption properties of carbon nanotubes can be arbitrarily divided into two groups. The first group is based on the consequences of molecule adsorption on nanotube electronic properties, whose main application is the chemical sensors (see Sect. 3.6.1). The second group includes gas storage, gas separation, or the use of carbon nanotubes as adsorbent and results from the morphological characteristics of carbon nanotubes (surface area, aspect ratio, and so forth). Among the three latter potential applications, the possibility of storing gases, and more particularly hydrogen, on carbon nanotubes has received most of the attention.

Table 3.4 Preparation and catalytic performances of some nanotube-supported catalysts

Catalyst	Preparation route	Catalytic reaction	Comments
Ru/MWNT+SWNT [3.230]	Liquid phase impregnation, no pretreatment of the tubes	Liquid phase cinnamaldehyde hydrogenation	Different kind of metal support interaction compared to activated carbon
Pt/MWNT electrodes [3.149]	Electrode-less plating with pre-functionalization of MWNT	Oxygen reduction for fuel cells applications	High electrocatalytic activity
Rh/MWNT [3.235]	Surface mediated organometallic synthesis, pre-functionalization of MWNT	Liquid phase hydroformylation and hydrogenation	Higher activity of Rh/MWNT compared to Rh/activated carbon
Ru-alkali/MWNT [3.238]	Liquid phase impregnation, no pretreatment of the tubes	Ammonia synthesis, gas phase reaction	Higher activity with MWNT than with graphite
Rh-phosphine/MWNT [3.239]	Liquid phase grafting from $[RhH(CO)(PPh_3)_3]$	Liquid phase hydroformylation	Highly active and regioselective catalyst

Gas Storage – Hydrogen

The development of lightweight and safe system for hydrogen storage is necessary for a wide use of highly efficient H_2-air fuel cells in transportation vehicles. The U.S. Department of Energy Hydrogen Plan has set a standard by providing a commercially significant benchmark for the amount of reversible hydrogen adsorption. This benchmark requires a system-weight efficiency (the ratio of H_2 weight to system weight) of 6.5 wt% hydrogen and a volumetric density of 63 kgH_2/m^3.

From now on, the failure to produce a practical storage system for hydrogen has prevented hydrogen from coming to the commercial forefront as a transportation fuel. The ideal hydrogen storage system needs to be light, compact, relatively inexpensive, safe, easy to use, and reusable without the need for regeneration. While research and development are continuing on such technologies as liquid hydrogen systems, compressed hydrogen systems, metal hydride systems, and superactivated carbon systems, all have serious disadvantages. For example, liquid hydrogen systems are very expensive, primarily because the hydrogen must be cooled to about −252 °C. For example, a liquid hydrogen system will cost about four times more than an equivalent amount of gasoline. Further, liquid hydrogen must be kept cooled to prevent it from boiling away, even when the vehicle is parked.

Compressed hydrogen is much cheaper than liquid hydrogen but also much bulkier.

Metal hydride systems store hydrogen as a solid in combination with other materials. For example, metal hydrides are produced by bathing a metal, such as palladium or magnesium, in hydrogen. The metal splits the dihydrogen gas molecules and binds the hydrogen atoms to the metal until released by heating. The disadvantages of a metal hydride system are its weight (typically about eight times more than an equivalent amount of liquid hydrogen or an equivalent amount of gasoline) and the need to warm it up to release the hydrogen.

Superactivated carbon is the basis of another system for storing hydrogen that initially showed commercial potential. Superactivated carbon is a material similar to the highly porous activated carbon used in water filters but can gently hold hydrogen molecules by physisorption at sub-zero temperatures. The colder the carbon, the less heat is needed to disturb the weak forces holding the carbon and hydrogen together. Again, a major disadvantage of such a system is that preventing the hydrogen from escaping requires it to constantly remain at very low temperatures, even when the vehicle is parked.

Consequently there still remains a great need for a material that can store hydrogen and is light, compact, relatively inexpensive, safe, easy to use, and reusable without regeneration. Recently some articles or patents concerning the very high, reversible adsorption of hydrogen in carbon nanotubes or platelet nanofibers have aroused tremendous interest in the research community, stimulating much experimental and theoretical works. Most of the works on hydrogen adsorption on carbon nanotubes have been recently reviewed [3.240–243].

A group from Northeastern University [3.132, 244] was the first to report the supposedly successful hydrogen storage in carbon layered nanostructures possessing some crystallinity. The authors claimed that, in platelet nanofibers (3–50 nm in width), hydrogen can be stored up to 75 wt%, meaning a C/H ratio of 1/9. Complete

Springer Handbook of Nanotechnology
B. Bhushan • © Springer 2004

hydrogen desorption occurs only at very high temperature. The same authors have recently pointed out the dramatic role of the treatment that has to be applied to the carbon nanofibers prior to H_2 adsorption [3.244]. This thermal treatment, though simple (1,000 °C in an inert gas), is a critical procedure in order to remove chemisorbed gases from the edge and step regions of the carbon structures. Failure to achieve this condition results in a dramatic decrease in the hydrogen adsorption performance of the materials. In this work, the authors have performed the experiments at room temperature, and 100 bar of H_2, hydrogen uptake between 20 and 40 wt% were recorded. The authors have also demonstrated by XRD measurements that structural perturbation of the material was produced following hydrogen treatment, and that it was manifested by an expansion of the graphitic interplane distance (from 0.34 before H_2 adsorption to 0.347 after H_2 uptake).

Hydrogen adsorption and desorption measurements on nanofibers similar to that used by *Rodriguez* et al. [3.19] were also performed by *Ahn* et al. [3.245]. The authors have measured a hydrogen uptake of 2.4 wt%. In fact, in their study, none of the graphite nanofiber material tested has indicated hydrogen storage abilities that exceed the values already reported for activated carbons.

Adsorption of H_2 on platelet nanofibers was computed by Monte Carlo simulations [3.246]. The graphene spacing was optimized to maximize the weight fraction of H_2 adsorbed. Given the results of their calculations, the authors concluded that no physically realistic graphite-hydrogen potential can account for the tremendous adsorption ability reported by the group from Penn State.

Rzepka et al. [3.247] have also used theoretical Monte Carlo calculation to calculate the amount of physically adsorbed hydrogen molecules in carbon slit pores of dimensions similar to that used by the Penn State group (i. e., 0.34 nm as the interplanar distance between graphenes in the nanofibers). They reached the conclusion that no hydrogen can be adsorbed at all. Even if the interplanar distance is assumed to expand during the adsorption process, the maximum calculated adsorption is 1 wt% for $d = 0.7$ nm. These authors concluded that the Penn State experimental results could not be attributed to some "abnormal capillary condensation effect" as initially claimed. Since then, nobody has been able to duplicate it experimentally.

Meanwhile, hydrogen storage experiments were also attempted on SWNTs. The first work was reported by *Dillon* et al. [3.248] who used temperature-programmed desorption measurements. A major drawback of their experimental procedure was that they used only 1 mg of *unpurified* soot containing 0.1% of SWNTs. They attributed the reversible hydrogen capacity of 0.01 wt% observed to the SWNTs only, therefore leading to a presumable capacity of 5–10 wt% for a pure nanotubes sample. TPD experiments shown H_2 desorption between −170 and +100 °C. Despite the weakness of the demonstration, the paper, which was published not much before that of Chambers et al. [3.132], has induced a worldwide excitement for the field.

Hydrogen adsorption measurements were also performed on SWNT samples of high purity [3.249]. At −193 °C, 160 bar of H_2 were admitted and the hydrogen adsorption was found to exceed 8 wt%. In this case, the authors proposed that hydrogen is first adsorbed on the outer surfaces of the crystalline material. But no data concerning H_2 desorption are given.

In recent reports, Dresselhaus and coworkers [3.240, 250] reported hydrogen storage in SWNTs (1.85 nm of average diameter) at room temperature. A hydrogen storage capacity of 4.2 wt% was achieved reproducibly under 100 bar for a SWNT-containing material of about 0.5 g that was previously soaked in hydrochloric acid and then heat-treated (500 °C) in vacuum. The purity of the sample was estimated from TGA and TEM observations to be about 50 to 60%. Moreover, 78% of the adsorbed hydrogen could be released under ambient pressure at room temperature, while the release of the residual stored hydrogen required some heating at 200 °C.

A valuable input to the topic was brought by *Hirscher* et al. [3.251], who demonstrated that several of the supposedly successful experiments regarding the storage of H_2 in SWNTs were actually misled by the hydrogen storage capacity of Ti nanoparticles originating from the sonoprobe frequently used at one step or another of the procedure, specifically when the SWNT material is previously purified.

Density functional theory (DFT) was used [3.252] to estimate H_2 adsorption in SWNTs. From their calculations, within the regime of operating conditions where adsorptive storage seems attractive, the storage properties of H_2 in a SWNT system appear to fall far short of the DOE target. In their model, rolling graphite into a cylindrical sheet does not significantly alter the nature of the carbon–H_2 interaction, in contradiction with the latest calculations, which indicate that the possibilities of physisorption increase as the radius of curvature of SWNTs decreases [3.253].

More recently, however, *Lee* et al. [3.254, 255] have also used DFT calculations to search for hydrogen adsorption sites and predict maximum storage capacity in SWNTs. They have found two chemisorption sites, at the exterior and the interior of the tube wall. Thus they predict a maximum hydrogen storage capacity that can exceed 14 wt% (160 kgH_2/m^3). The authors have also considered H_2 adsorption in multiwall carbon nanotubes and have predicted lower storage capacities for the latter.

Calculations are constrained, however, by the starting hypotheses. While considering the same (10,10) SWNT, although calculations based on DFT predict a 14.3% storage [3.254], calculations based on a geometrical model predict 3.3% [3.240], and calculations based on quantum mechanical molecular dynamics model predict 0.47% [3.256].

In conclusion, neither the experimental results, obviously biased by some problem in the procedures, nor theoretical results are yet able to demonstrate that an efficient storage of H_2 is achievable for carbon nanotubes, whatever the type. But the definitive failure statement cannot yet be claimed. Further efforts have to be made to enhance H_2 adsorption of these materials, in particular (i) by adjusting the surface properties that can be modified by chemical or mechanical treatments, and (ii) by adjusting the material structure (pore size and curvature). Whether the best carbon material will then be nanotube-based is another story.

Gas Storage – Gases Other than Hydrogen

Encouraged by the potential applications related to hydrogen adsorption, several research groups have tried to use carbon nanotubes as a stocking and transporting mean for other gases such as: oxygen, nitrogen, noble gases (argon and xenon), and hydrocarbons (methane, ethane, and ethylene). These studies have shown that carbon nanotubes could become the world's smallest gas cylinders combining low weight, easy transportability, and safe use with acceptable adsorbed quantities. Thanks to their nano-sizes, nanotubes might also be used in medicine where physically confining special gases (like ^{133}Xe for instance) prior to injection would be extremely useful.

Kusnetzova et al. [3.257] have conducted experiments with xenon and found that a very significant enhancement (280 times more, up to a molar ratio $N_{\text{Xe}}/N_{\text{C}} = 0.045$) can be achieved by opening the SWNT bundles by thermal activation at 800 °C. With this treatment the gas can be adsorbed inside the nanotubes and the rates of adsorption are also increased.

The possibility of storing argon in carbon nanotubes has been studied with encouraging results by *Gadd* et al. [3.258]: their experiments show that large amounts of argon can be trapped into catalytically grown MWNTs (20–150 nm) by hot isostatic pressing (HIP-ing) for 48 hours at 650 °C under an argon pressure of 1,700 bar. Energy dispersive X-ray spectroscopy was used to determine that the gas was located inside the tubes and not on their walls. Further studies determined the argon pressure inside the tubes at room temperature. The authors estimated this feature to be around 600 bars, indicating that the equilibrium pressure was attained in the tubes during the HIP-ing and that MWNTs would be a convenient material for storing that gas.

Gas Separation

As SWNTs or MWNTs have regular geometries that can, to some extent, be controllable, they could be used to develop precise separation tools. If the sorption mechanisms are known, it should be possible to control sorption of various gases through combinations of temperature, pressure, and nanotube morphology. Since the large-scale production of nanotubes is now constantly progressing and may result in low costs, accurate separation methods based on carbon nanotubes have started to be investigated.

A theoretical study has aimed to determine the effect of different factors such as the diameter of the tubes, the density and the type of the gas used on the flow of molecules inside the nanotubes. An atomistic simulation with methane, ethane, and ethylene [3.259] has shown that the molecule mobility decreases with the decrease of the diameter of the tube for each of the three gases. Ethane and ethylene have a smaller mobility due to the stronger interaction they seem to have with the nanotube walls. In another theoretical study on the possibility of hydrocarbon mixture separation on SWNT bundles, the authors conclude on the possibility of using carbon nanotubes to separate methane/n-butane and methane/isobutene mixtures [3.260], with an efficiency that increases as the average tube diameter decreases. An experimental work was also performed by the same group on the sorption of butane on MWNTs [3.133].

Simulation by grand canonical Monte Carlo for separation of hydrogen and carbon monoxide by adsorption on SWNT has also been reported [3.261]. In most of the situations studied, SWNTs are found to adsorb more CO than H_2, and an excellent separation effect could probably be obtained, again, by varying the SWNT average diameter.

Adsorbents

Carbon nanotubes were recently found to be able to adsorb some toxic gases such as dioxins [3.262], fluoride [3.263], lead [3.264], or alcohols [3.265] better than the materials used so far, such as activated carbon. These pioneering works opened a new field of applications as cleaning filters for many industrial processes having hazardous by-products. Adsorption of dioxins, which are very common and persistent carcinogenic by-products of many industrial processes, is a good example of the interest of nanotubes in this field. Ecological consciousness has imposed emission limits on dioxin-generating sources in many countries, but finding materials that can act as effective filters, even at extremely low concentrations, is a big issue. *Long* et al. [3.262] have found that nanotubes can attract and trap more dioxins than activated carbons or other polyaromatic materials that are currently used as filters. This improvement is probably due to the stronger interaction forces that exist between dioxin molecules and the nanotube curved surfaces compared to those for flat graphene sheets.

The adsorption capacity of Al_2O_3/MWNT for fluoride from water has been reported to be 13.5 times that of activated carbon and 4 times that of Al_2O_3 [3.263]. The same group has also reported an adsorption capacity of lead from water by MWNTs higher than that by activated carbon [3.264]. The possibility to using graphite nanofibers to purify water from alcohols has also been explored [3.265]. These experimental results suggest that carbon nanotubes may be promising adsorbents for the removal of polluting agents from water.

Bio-Sensors

Attaching molecules of biological interest to carbon nanotubes is an ideal way to realize nanometer-sized biosensors. Indeed, the electrical conductivity of such functionalized nanotubes would depend on modifications of the interaction of the probe with the studied media, because of chemical changes or as result of their interaction with target species. The science of attaching biomolecules to nanotubes is rather recent and was inspired by similar research in the fullerene area. Some results have already been patented, and what was looking like a dream a couple of years ago may become reality in the near future. The use of the internal cavity of nanotubes for drug delivery would be another amazing application, but little work has been carried out so far to investigate the harmfulness of nanotubes in the human body. MWNTs have been used by *Mattson* et al. [3.266] as a substrate for neuronal growth. They have compared the activity of untreated MWNTs with that of MWNTs coated with a bioactive molecule (4-hydroxynonenal) and observed that on these latter functionalized nanotubes, neurons elaborated multiple neurites. This is an important result illustrating the feasibility of using nanotubes as substrate for nerve cell growth.

Davis et al. [3.267] have immobilized different proteins (metallothionein, cytochrome c and c_3, β-lactamase I) in MWNTs and checked that these molecules were still catalytically active compared to the nonimmobilized ones. They have shown that confining a protein within a nanotube provides some protection toward the external environment. Protein immobilization via noncovalent sidewall functionalization was proposed by *Chen* et al. [3.268] by using a bifunctional molecule (1-pyrenebutanoic acid, succinimidyl ester). This molecule is maintained to the nanotube wall by the pyrenyl group, and amine groups or biological molecules can react with the ester function to form amid bonds. This method was also used to immobilize ferritin and streptavidin onto SWNTs. It has the main advantage of not modifying the SWNT wall and keeping unperturbed the sp^2 structure, maintaining the physical properties of the nanotubes. *Shim* et al. [3.269] have functionalized SWNTs with biotin and observed specific binding with streptavidin, illustrating biomolecular recognition possibilities. *Dwyer* et al. [3.270] have functionalized SWNTs by covalently coupling DNA strands to them using EDC(1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride) but did not test biomolecular recognition; other proteins such as bovine serum albumin (BSA) [3.271] have been attached to nanotubes using the same process (diimide-activated amidation with EDC) and most of the attached proteins remained bioactive. Instead of working with individual nanotubes (or more likely nanotube bundles in the case of SWNTs), *Nguyen* et al. [3.272] have functionalized nanotubes arrayed with a nucleic acid, still using EDC as the coupling agent, in order to facilitate the realization of biosensors based on protein-functionalized nanotubes. Recently *Azamian* et al. [3.273] have immobilized a series of biomolecules (cytochrome c, ferritin, and glucose oxydase) on SWNTs and observed that the use of EDC was not always necessary, indicating that the binding was predominantly noncovalent. In the case of glucose oxydase, they have tested the catalytic activity of functionalized nanotubes immobilized on a glassy carbon electrode and observed a tenfold greater catalytic response compared to that in the absence of modified SWNTs.

Functionalization of nanotubes with biomolecules is still in its infancy, and their use as biosensors may lead to practical applications earlier than expected. For example, functionalized nanotubes can be used as AFM tips (see Sect. 3.6.1) allowing to perform "chemical force microscopy" (CFM), allowing single-molecule measurements. Even in the case of nonfunctionalized CNT-based tips, important improvements have been obtained for the characterization of biomolecules (see the review by [3.215]).

Composites

Because of their exceptional morphological, electrical, thermal, and mechanical characteristics, carbon nanotubes are particularly promising materials as reinforcement in composite materials with metal, ceramic, or polymer matrix. Key basic issues include the good dispersion of the nanotubes, the control of the nanotube/matrix bonding, the densification of bulk composites and thin films, and the possibility of aligning the nanotubes. In addition, the nanotube type (SWNT, c-MWNT, h-MWNT, etc.) and origin (arc, laser, CCVD, etc.) is also an important variable since determining the structural perfection, surface reactivity, and aspect ratio of the reinforcement.

Considering the major breakthrough that carbon nanotubes are expected to make in the field, the following will give an overview of the current work on metal-, ceramic- and polymer-matrix composites reinforced with nanotubes. We will not consider nanotubes merely coated with another material will not be considered here. We discussed filled nanotubes in Sect. 3.5.2. Applications involving the incorporation of nanotubes to materials for purposes other than structural will be mentioned in Sect. 3.6.2.

Metal Matrix Composites

Nanotube-metal matrix composites are still rarely studied. The materials are generally prepared by standard powder metallurgy techniques, but the dispersion of the nanotubes is not optimal. Thermal stability and electrical and mechanical properties of the composites are investigated. The room temperature electrical resistivity of hot-pressed CCVD MWNTs-Al composites increases slightly by increasing the MWNT volume fraction [3.274]. The tensile strength and elongation of nonpurified arc-discharge MWNT-Al composites are only slightly affected by annealing at 873 K in contrast to that of pure Al [3.275]. The Young's modulus of nonpurified arc-discharge MWNTs-Ti composites is about 1.7 times that of pure Ti [3.276]. The formation of TiC, probably resulting from the reaction between amorphous carbon and the matrix, was observed, but the MWNTs themselves were not damaged. An increase of the Vickers hardness by a factor 5.5 over that of pure Ti was associated with the suppression of coarsening of the Ti grains, the TiC formation, and the addition of MWNTs with an extremely high Young's modulus. CCVD MWNTs-Cu composites [3.277] also show a higher hardness and a lower friction coefficient and wear loss. A deformation of 50–60 % of the composites was observed. A MWNTs-metallic glass Fe82P18 composite [3.278] with a good dispersion of the MWNTS was prepared by the rapid solidification technique.

Ceramic Matrix Composites

Carbon nanotube-containing ceramic-matrix composites are a bit more frequently studied, most efforts made to obtain tougher ceramics [3.279]. Composites can be processed following the regular processing route, in which the nanotubes are usually mechanically mixed with the matrix (or a matrix precursor) then densified using hot-pressing sintering of the slurry or powders. *Zhan* et al. [3.280] ball-milled SWNT bundles (from the HiPCo technique) and nanometric alumina powders, producing a fairly homogeneous dispersion, supposedly without damaging the SWNTs. On the other hand, interesting results were obtained using the spark plasma sintering (SPS) technique, which was able to prepare fully dense composites without damage to the SWNTs [3.281]. Other original composite processing include sol-gel route, by which bulk and thin film composites were prepared [3.282], and in situ SWNT growth in ceramic foams, using procedures closely related to that described in Sect. 3.2.2 [3.283].

Ma et al. [3.284] prepared MWNT-SiC composites by hot-pressing mixtures of CCVD MWNTs and nano-SiC powders. They claimed that the presence of the MWNTs provides an increase of about 10% of both the bending strength and fracture toughness, but the dispersion of the MWNTs looks poor. Several detailed studies [3.82, 285–288] have dealt with the preparation of nanotube-Fe-Al_2O_3, nanotube-Co-MgO, and nanotube-Fe/Co-$MgAl_2O_4$ composites by hot-pressing of the corresponding composite powders. The nanotubes were grown in situ in the starting powders by a CCVD method [3.73, 80, 289–293] and thus are very homogeneously dispersed between the metal-oxide grains, in a way that could be impossible to achieve by mechanical mixing. The nanotubes (mostly SWNTs and DWNTs) gather in long, branched bundles smaller than 50 nm in diameter, which appear to be very flexible. Depending on the matrix and hot-pressing temperature, a fraction of the nanotubes seems to be destroyed during hot-pressing in a primary vacuum. The increase of the quantity of nanotubes (ca. 2–12 wt%) leads to a refinement of the microstructure but also to a strong decrease in relative density (70–93 %). Probably because of this, the fracture strength and the toughness of the nanotube-containing composites are generally lower than those of the nanotube-free metal-oxide composites and only marginally higher than those of the corresponding ceramics. SEM observations revealed both the trapping of some of the nanotubes within the matrix grains or at grain boundaries and a relatively good wetting in the case of alumina. Most nanotubes are cut near the fracture surface after some pull-out and could contribute to a mechanical reinforcement.

Siegel et al. [3.294] claimed that the fracture toughness of 10 vol.% MWNTs-alumina hot-pressed composites is increased by 24% (to 4.2 MPa $m^{1/2}$) over that of pure alumina. But the density and grain size are not reported, so it is difficult to consider this result as an evidence for the beneficial role of the MWNTs themselves. A confirmation could be found in the work by *Sun* et al. [3.281] who developed a colloidal route to coat the nanotubes with alumina particles prior to another mixing step with alumina particles then densification by SPS. A very large gain in fracture toughness (calculated from Vickers indentation) is reported (from 3.7 MPa $m^{1/2}$ for pure Al_2O_3 to 4.9 MPa $m^{1/2}$ for an addition of 0.1 wt% of SWNT nanotubes, and to 9.7 MPa $m^{1/2}$ for an addition of 10 wt%). But it is possible that the beneficial input of the nanotubes is due to the previous coating of the nanotubes before sintering, resulting in an improved bonding with the matrix.

Bulk properties other than mechanical are also worth being investigated. Interestingly, the presence of well dispersed nanotubes confers an electrical conductivity to the otherwise insulating ceramic-matrix composites. The percolation threshold is very low – in the range 0.2–0.6 wt% of carbon – due to the very high aspect ratio of nanotubes. Varying the nanotube quantity in the starting powders allows the electrical conductivity to be controlled in the range 0.01–10 S cm^{-1}. Moreover, the extrusion at high temperatures in vacuum allows the nanotubes to be aligned in the matrix, thus inducing an anisotropy in the electrical conductivity [3.295]. One of the early works involved MWNTs in a superconducting $Bi_2Sr_2CaCu_2O_{8+\delta}$ matrix [3.296]. The application of SWNTs for thermal management in partially stabilized zirconia [3.297] and of MWNTs-lanthanum cobaltate composites for application as an oxygen electrode in zinc/air [3.298] were reported more recently. The friction coefficient was also shown to increase and the wear loss to decrease upon the increase in nanotube volume fraction in carbon/carbon composites [3.299].

The latter is one of the few examples in which carbon has been considered as a matrix. Another one is the work reported by *Andrews* et al. [3.300], who introduced 5 wt% of SWNTs into a pitch material before spinning the whole as a SWNT-reinforced pitch-fiber. Subsequent carbonization transformed it into a SWNT-reinforced carbon fiber. Due to the contribution of the SWNTs, which were aligned by the spinning stresses, the resulting gain in tensile strength, modulus, and electrical conductivity with respect to the nonreinforced carbon fiber was claimed to be 90, 150, and 340%, respectively.

Polymer Matrix Composites

Nanotube-polymer composites, first reported by *Ajayan* et al. [3.301] are now intensively studied, notably epoxy- and polymethylmethacrylate (PMMA)-matrix composites. Regarding the mechanical characteristics, the three key issues affecting the performance of a fiber-polymer composite are the strength and toughness of the fibrous reinforcement, its orientation, and a good interfacial bonding crucial for the load transfer to occur [3.302]. The ability of the polymer to form large-diameter helices around individual nanotubes favors the formation of a strong bond with the matrix [3.302]. Isolated SWNTs may be more desirable than MWNTs or bundles for dispersion in a matrix because of the weak frictional interactions between layers of MWNTs and between SWNTs in bundles [3.302]. The main mechanisms of load transfer are micromechanical interlocking, chemical bonding, and van der Waals bonding between the nanotubes and the matrix. A high interfacial shear stress between the fiber and the matrix will transfer the applied load to the fiber over a short distance [3.303]. SWNTs longer than 10–100 µm would be needed for significant load-bearing ability in the case of nonbonded SWNT-matrix interactions, whereas the critical length for SWNTs cross-linked to the matrix is only 1 µm [3.304]. Defects are likely to limit the working length of SWNTs [3.305], however.

The load transfer to MWNTs dispersed in an epoxy resin was much higher in compression than in tension [3.303]. It was proposed that all the walls of the MWNTs are stressed in compression, whereas only the outer walls are stressed in tension because all the inner tubes are sliding within the outer. Mechanical tests performed on 5 wt% SWNT-epoxy composites [3.306] showed that SWNTs bundles were pulled out of the matrix during the deformation of the material. The influence of the nanotube/matrix interfacial interaction was evidenced by *Gong* et al. [3.307]. It was also reported that coating regular carbon fiber with MWNTs prior to their dispersion into an epoxy matrix improves the interfacial load transfer, possibly by local stiffening of the matrix near interface [3.308].

As for ceramic matrix composites, the electrical characteristics of SWNT- and MWNT-epoxy composites are described with the percolation theory. Very low percolation threshold (below 1 wt%) is often reported [3.309–311]. Industrial epoxy loaded with 1 wt% unpurified CCVD-prepared SWNTs showed an increase in thermal conductivity of 70% and 125% at 40 K and at room temperature, respectively [3.309]. Also, the Vickers hardness rose with the SWNT loading up to a factor of 3.5 at 2 wt%. Thus both the thermal and mechanical properties of such composites are improved without the need to chemically functionalize the SWNTs.

Likewise, the conductivity of laser-prepared SWNT-PMMA composites increases with the load in SWNTs (1–8 wt% SWNTs) [3.302]. Thermogravimetric analysis shows that, compared to pure PMMA, the thermal degradation of PMMA films occurs at a slightly higher temperature when 26 wt% MWNTs are added [3.312]. Improving the wetting between the MWNTs and the PMMA by coating the MWNTs with poly(vinylidene fluoride) prior to melt-blending with PMMA resulted in an increased storage modulus [3.313]. The impact strength in aligned SWNT-PMMA composites was significantly increased with only 0.1 wt% of SWNTs, possibly because of a weak interfacial adhesion, and/or of the high flexibility of the SWNTs, and/or the pullout and sliding effects of individual SWNTs within bundles [3.314]. The transport properties of arc-discharge SWNTs-PMMA composite films (10 μm thick) were studied in great detail [3.315, 316]. The electrical conductivity increases by nine orders of magnitude from 0.1 to 8 wt% SWNTs. The room temperature conductivity is again well described with the standard percolation theory, confirming the good dispersion of the SWNTs in the matrix.

Polymer composites with other matrices include CCVD-prepared MWNT-polyvinyl alcohol [3.317], arc-prepared MWNT-polyhydroxyaminoether [3.318], arc-prepared MWNT-polyurethane acrylate [3.319, 320], SWNT-polyurethane acrylate [3.321], SWNT-polycarbonate [3.322], MWNT-polyaniline [3.323], MWNT-polystyrene [3.324], SWNT-polyethylene [3.325], CCVD-prepared MWNT-polyacrylonitrile. [3.326], MWNT-oxotitanium phthalocyanine [3.327], arc-prepared MWNT-poly(3-octylthiophene) [3.328], SWNT-poly(3-octylthiophene) [3.329], and CCVD MWNTs-poly(3-hexylthiophene) [3.330]. These works deal mainly with films 100–200 micrometer thick and aim to study the glass transition of the polymer, mechanical, and electrical characteristics as well as photoconductivity.

A great deal of work is also devoted to nano-tube-polymer composites for applications as materials for molecular optoelectronics, using primarily poly(m-phenylenevinylene-co-2,5-dioctoxy-p-phenylenevinylene) (PmPV) as the matrix. This conjugated polymer tends to coil, forming a helical structure. The electrical conductivity of the composite films (4–36 wt% MWNTs) is increased by eight orders of magnitude compared to that of PmPV [3.331]. Using the MWNT-PmPV composites as the electron-transport-layer in light-emitting diodes results in a significant increase in brightness [3.332]. The SWNTs act as a hole-trapping material blocking the holes in the composites, this being probably induced through long-range interactions within the matrix [3.333]. Similar investigations were carried out on arc-discharge SWNTs-polyethylene dioxythiophene (PEDOT) composite layers [3.334] and MWNTs-polyphenylenevinylene composites [3.335].

To conclude, two critical issues have to be considered for the application of nanotubes as reinforcement for advanced composites. One is to chose between SWNTs and MWNTs. The former seem more beneficial for the purpose of mechanical strengthening, provided they are isolated or arranged into cohesive yarns so that the load is able to be conveniently transferred from a SWNT to another. Unfortunately, despite recent advances [3.336–339], this is still a technical challenge. The other issue is to tailor the nanotube/matrix interface with respect to the matrix. This is a current topic for many laboratories involved in the field.

Multifunctional Materials

One of the major benefits expected from incorporating carbon nanotubes in other solid or liquid materials is bringing some electrical conductivity to them while not affecting the other properties or behaviors of these materials. As already mentioned in the previous section, the percolation threshold is reached at very low load with nanotubes. Tailoring the electrical conductivity of a bulk material is then achievable by adjusting the nanotube volume fraction in the formerly insulating material while not making this fraction too large anyway. As demonstrated by Maruyama [3.3], there are three regimes of interest regarding the electrical conductivity: (i) electrostatic discharge, e.g., to prevent fire or explosion hazard in combustible environments or perturbations in electronics, which requires electrical resistivity less than $10^{12}\ \Omega$ cm; (ii) electrostatic painting, which requires the material to be painted to be electrical conductive enough (electrical resistivity below $10^6\ \Omega$ cm) to prevent the charged paint droplets to be repelled; (iii) electromagnetic interference shielding, which is achieved for electrical resistivity lower than $10\ \Omega$ cm.

Materials are often asked to be multifunctional, e.g. having both high electrical conductivity and high toughness, or high thermal conductivity and high thermal stability, etc. The association of several materials, each of them bringing one of the desired features, therefore generally meets this need. Exceptional features and properties of carbon nanotubes make them likely to be a perfect multifunctional material in many cases. For instance, materials for satellites are often required to be electrical conductive, mechanically self-supporting, able to transport the excess heat away, sometimes be protected against electromagnetic interferences, etc., while exhibiting minimal weight and volume. All these properties should be possible with a single nanotube-containing composite material instead of complex multi-materials combining layers of polymers, aluminum, copper, etc. Table 3.5 provides an overview of various fields in which nanotube-based multifunctional materials should find an application.

Nano-Electronics

As reported in Sects. 3.1.1 and 3.4.4, SWNT nanotubes can be either metallic (with an electrical conductivity higher than that of copper), or semiconductor. This has inspired the design of several components for nanoelectronics. First, metallic SWNTs can be used as mere ballistic conductors. Moreover, as early as 1995,

Table 3.5 Applications for nanotube-based multifunctional materials (from [3.3]), by courtesy of B. Maruyama (WPAFB, Dayton, Ohio)

Fiber fraction	Applications system	Mechanical			Electrical			Thermal		Thermo-mechanical	
		Strength/ stiffness	Specific strength	through-thickness strength	Static dissipation	Surface Conduction[a]	EMI shielding	Service[b] temp.	conduction/ dissipation[c]	Dimensional Stability[d]	CTE reduction[e]
Low Volume fraction (fillers)											
Elastomers	Tires	×			×				×		
Thermo Plastics	Chip package				×				×		
	Electronics/ Housing	×					×	×	×		
Thermosets	Epoxy products	×	×	×		×				×	×
	Composites			×						×	
High Volume Fraction											
Structural composites	Space/aircraft components		×	×							
High conduction composites	Radiators	×							×	×	
	Heat exchangers	×						×	×		×
	EMI shield	×					×				

[a] For electrostatic painting, to mitigate lightning strikes on aircraft, etc.

[b] To increase service temperature rating of product

[c] To reduce operating temperatures of electronic packages

[d] Reduces warping

[e] Reduces microcracking damage in composites

realizing a rectifying diode by joining one metallic SWNT to one semiconductor SWNT (hetero-junction) was proposed by *Lambin* et al. [3.340], then later by *Chico* et al. [3.341] and *Yao* et al. [3.342]. Also, field-effect transistors (FET) can be built by the attachment of a semiconductor SWNT across two electrodes (source and drain) deposited on an insulating substrate that serves as a gate electrode [3.343, 344]. Recently the association of two such SWNT-based FETs has made a voltage inverter [3.345].

All the latter developments are fascinating and provide promising perspectives for nanotube-based electronics. But progress is obviously needed before making SWNT-based integrated circuits on a routine basis. A key point is the need to be able to prepare selectively either metallic or semiconductor nanotubes. Although a way has been proposed to destroy selectively metallic SWNTs from bundles of undifferentiated SWNTs [3.347], the method is not scalable and selective synthesis would be preferable. Also, defect-free nanotubes are required. Generally speaking, it relates to another major challenge, which is to be able to fabricate at industrial scale integrated circuits including nanometer-size components that only sophisticated imaging methods (AFM) are able to visualize.

Nano-Tools, Nano-Devices, Nano-Systems

Due to the ability of graphene to expand slightly when electrically charged, nanotubes have been found to act conveniently as actuators. *Kim* et al. [3.346] demonstrated it by designing "nano"-tweezers able to grab, manipulate, release nano-objects (the "nano"-bead having been handled for the demonstration was actually closer to micrometer than nanometer), and measure their electrical properties. This was made possible quite simply by depositing two non-interconnected gold coatings onto a pulled glass micropipette (Fig. 3.31), then attaching two MWNTs (or two SWNT-bundles) ∼ 20–50 nm in diameter to each of the gold electrodes. Applying a voltage (0–8.5 V) between the two electrodes then makes the tube tips to open and close reversibly in a controlled manner.

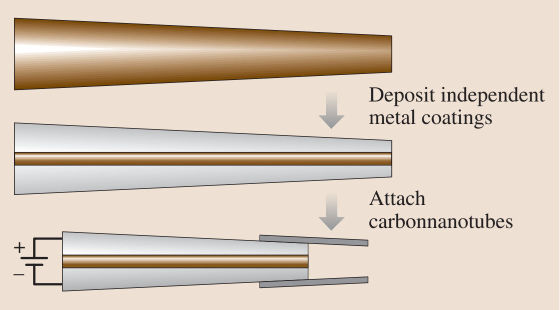

Fig. 3.31 Sketch explaining how the first nano-tweezers were designed. First is a glass micropipette (dark cone, *top*). Then two Au coatings (in grey, *middle*) are deposited so that they are not in contact. Then a voltage is applied to the electrodes (from [3.346])

A similar experiment, again rather simple, was proposed by *Baughman* et al. the same year (1999) [3.348], consisting in mounting two SWNT-based paper strips ("bucky-paper") on both sides of an insulating double-side tape. The two bucky-paper strips were previously loaded with Na^+ and Cl^-, respectively. When 1 V was applied between the two paper strips, both expand, but the strip loaded with Na^+ expands a bit more, forcing the whole system to bend. Though performed in a liquid environment, such a behavior has inspired the authors to predict a future for their system as "artificial muscles."

Another example of amazing nano-tools is the nano-thermometer proposed by *Gao* et al. [3.349]. A single MWNT was used, in that case, partially filled with liquid gallium. Upon the effect of temperature variations in the range 50–500 °C, the gallium goes up and down reversibly within the nanotube cavity at reproducible level with respect to the values of the temperature applied.

Of course, nano-tools such as nano-tweezers or nano-thermometers will hardly reach a commercial development so to justify industrial investments. But such experiments are more than amazing laboratory curiosities. They definitely demonstrate the ability of carbon nanotubes as parts for future nano-devices, including nanomechanics-based systems.

Supercapacitors

Supercapacitors include two electrodes immersed in an electrolyte (e.g., 6 M KOH), separated by an insulating ion-permeable membrane. Charging the capacitors is achieved by applying a potential between the two electrodes, making the cations and the anions moving toward the electrode oppositely charged. Suitable electrodes should exhibit a high electrical conductivity and a high surface area since the capacitance is proportional to it. Actually, the surface area should originate mainly from the appropriate combination of mesopores (to allow the electrolyte components to circulate well, which is related to the charging speed) and micropores (whose walls are the surface of attraction and fixation sites for the ions). Based on early works by

Niu et al. [3.350], such a combination was found to be reached with the specific architecture offered by packed and entangled h-MWNTs with a poor nanotexture (see Sect. 3.1.2). But, activation pre-treatments are necessary. For instance, a capacitor made from nanotubes with surface area $= 220\,\mathrm{m}^2/\mathrm{g}$ exhibited a capacitance of $20\,\mathrm{F/g}$, which increased up to $100\,\mathrm{F/g}$ after an activation treatment was applied to the nanotubes so that their surface area reaches $880\,\mathrm{m}^2/\mathrm{g}$ [3.126]. Alternatively, again due to their specific architecture induced by their huge aspect ratio, nanotubes can also be used as supports for conductive polymer coatings, e.g., polypyrrole or polyaniline [3.351], which otherwise would make a too dense phase (i. e. not allowing an easy circulation and penetration of ions). Supercapacitors built from such composites can overcome more than 2,000 charging cycles, with current density as high as $350\,\mathrm{mA/g}$ [3.352]. Capacitors including nanotubes already have capacitance as high as $180–200\,\mathrm{F/g}$, i. e. equivalent to that obtained with electrodes built from regular carbon materials, with the advantage of faster charging [3.126]. Current works will certainly lead to further optimization of both the nanotube material architecture and the nanotube-supported conductive polymers, giving reasonable perspectives for the commercial use of nanotubes as components for supercapacitors.

References

3.1 M. Hillert, N. Lange: The structure of graphite filaments, Zeitschr. Kristall. **111** (1958) 24–34

3.2 T. V. Hughes, C. R. Chambers: US Patent 405,480 (1889)

3.3 B. Maruyama, K. Alam: Carbon nanotubes and nanofibers in composite materials, SAMPE J. **38** (2002) 59–70

3.4 P. Schützenberger, L. Schützenberger: Sur quelques faits relatifs à l'histoire du carbone, C. R. Acad. Sci. Paris **111** (1890) 774–778

3.5 C. Pélabon, H. Pélabon: Sur une variété de carbone filamenteux, C. R. Acad. Sci. (Paris) **137** (1903) 706–708

3.6 R. T. K. Baker, P. S. Harris: The formation of filamentous carbon. In: *Chemistry and Physics of Carbon*, Vol. 14, ed. by P. L. Walker Jr., P. A. Thrower (Dekker, New York 1978) pp. 83–165

3.7 S. Iijima: Helical microtubules of graphite carbon, Nature **354** (1991) 56–58

3.8 S. Iijima, T. Ichihashi: Single-shell carbon nanotubes of 1-nm diameter, Nature **363** (1993) 603–605

3.9 D. S. Bethune, C. H. Kiang, M. S. de Vries, G. Gorman, R. Savoy, J. Vazquez, R. Bayers: Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls, Nature **363** (1993) 605–607

3.10 J. Tersoff, R. S. Ruoff: Structural properties of a carbon-nanotube crystal, Phys. Rev. Lett. **73** (1994) 676–679

3.11 N. Wang, Z. K. Tang, G. D. Li, J. S. Chen: Single-walled 4 Å carbon nanotube arrays, Nature **408** (2000) 50–51

3.12 N. Hamada, S. I. Sawada, A. Oshiyama: New one-dimensional conductors, graphitemicrotubules, Phys. Rev. Lett. **68** (1992) 1579–1581

3.13 M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund: *Science of Fullerenes and Carbon Nanotubes* (Academic, San Diego 1995)

3.14 M. Monthioux, B. W. Smith, B. Burteaux, A. Claye, J. Fisher, D. E. Luzzi: Sensitivity of single-wall nanotubes to chemical processing: An electron microscopy investigation, Carbon **39** (2001) 1261–1272

3.15 H. Allouche, M. Monthioux: Chemical vapor deposition of pyrolytic carbon onto carbon nanotubes II – Structure and texture, Carbon (2003) accepted

3.16 M. Audier, A. Oberlin, M. Oberlin, M. Coulon, L. Bonnetain: Morphology and crystalline order in catalytic carbons, Carbon **19** (1981) 217–224

3.17 Y. Saito: Nanoparticles and filled nanocapsules, Carbon **33** (1995) 979–988

3.18 P. J. F. Harris: *Carbon Nanotubes and Related Structures* (Cambridge Univ. Press, Cambridge 1999)

3.19 N. M. Rodriguez, A. Chambers, R. T. Baker: Catalytic engineering of carbon nanostructures, Langmuir **11** (1995) 3862–3866

3.20 H. W. Kroto, J. R. Heath, S.C. O'Brien, R. F. Curl, R. E. Smalley: C_{60} Buckminsterfullerene, Nature **318** (1985) 162–163

3.21 W. Krätschmer, L. D. Lamb, K. Fostiropoulos, D. R. Huffman: Solid C_{60}: A new form of carbon, Nature **347** (1990) 354–358

3.22 T. Guo, P. Nikolaev, A. G. Rinzler, D. Tomanek, D. T. Colbert, R. E. Smalley: Self-assembly of tubular fullerenes, J. Phys. Chem. **99** (1995) 10694–10697

3.23 T. Guo, P. Nikolaev, A. Thess, D. T. Colbert, R. E. Smalley: Catalytic growth of single-walled nanotubes by laser vaporisation, Chem. Phys. Lett. **243** (1995) 49–54

3.24 A. G. Rinzler, J. Liu, H. Dai, P. Nikolaev, C. B. Huffman, F. J. Rodriguez-Macias, P. J. Boul, A. H. Lu, D. Heymann, D. T. Colbert, R. S. Lee, J. E. Fischer, A. M. Rao, P. C. Eklund, R. E. Smalley: Large scale purification of single wall carbon nanotubes: Process, product and characterization, Appl. Phys. A **67** (1998) 29–37

3.25 A. Thess, R. Lee, P. Nikolaev, H. Dai, P. Petit, J. Robert, C. Xu, Y. H. Lee, S. G. Kim, D. T. Colbert, G. Scuseria, D. Tomanek, J. E. Fischer, R. E. Smalley: Crystalline ropes of metallic carbon nanotubes, Science **273** (1996) 487–493

3.26 L. M. de la Chapelle, J. Gavillet, J. L. Cochon, M. Ory, S. Lefrant, A. Loiseau, D. Pigache: A continuous wave CO_2 laser reactor for nanotube synthesis, Proc. Electronic Properties Novel Materials-XVI Int. Winterschool – AIP Conf. Proc., Melville 1999, ed. by H. Kuzmany, J. Fink, M. Mehring, S. Roth (Springer, Berlin, Heidelberg 1999) **486** 237– 240

3.27 M. Yudasaka, T. Komatsu, T. Ichihashi, S. Iijima: Single wall carbon nanotube formation by laser ablation using double targets of carbon and metal, Chem. Phys. Lett. **278** (1997) 102–106

3.28 M. Castignolles, A. Foutel-Richard, A. Mavel, J. L. Cochon, D. Pigache, A. Loiseau, P. Bernier: Combined experimental and numerical study of the parameters controlling the C-SWNT synthesis via laser vaporization, Proc. Electronic Properties of Novel Materials-XVI Int. Winterschool – AIP Conf. Proc., Melville 2002, ed. by H. Kuzmany, J. Fink, M. Mehring, S. Roth (Springer, Berlin, Heidelberg 2002) **633** 385–389

3.29 T. W. Ebbesen, P. M. Ajayan: Large-scale synthesis of carbon nanotubes, Nature **358** (1992) 220–221

3.30 D. Ugarte: Morphology and structure of graphitic soot particles generated in arc-discharge C_{60} production, Chem. Phys. Lett. **198** (1992) 596–602

3.31 T. W. Ebbesen: Carbon nanotubes, Ann. Rev. Mater. Sci. **24** (1994) 235–264

3.32 T. Beltz, J. Find, D. Herein, N. Pfänder, T. Rühle, H. Werner, M. Wohlers, R. Schlögl: On the production of different carbon forms by electric arc graphite evaporation, Ber. Bunsenges. Phys. Chem. **101** (1997) 712–725

3.33 C. Journet, W. K. Maser, P. Bernier, A. Loiseau, L. M. de la Chapelle, S. Lefrant, P. Deniard, R. Lee, J. E. Fischer: Large-scale production of single-walled carbon nanotubes by the electric-arc technique, Nature **388** (1997) 756–758

3.34 K. Saïdane, M. Razafinimanana, H. Lange, M. Baltas, A. Gleizes, J. J. Gonzalez: Influence of the carbon arc current intensity on fullerene synthesis, Proc. 24th Int. Conf. on Phenomena in Ionized Gases, Warsaw 1999, ed. by P. Pisarczyk, T. Pisarczyk, J. Wotowski, 203–204

3.35 H. Allouche, M. Monthioux, M. Pacheco, M. Razafinimanana, H. Lange, A. Huczko, T. P. Teulet, A. Gleizes, T. Sogabe: Physical characteristics of the graphite-electrode electric-arc as parameters for the formation of single-wall carbon nanotubes, Proc. Eurocarbon (Deutsche Keram. Ges., Berlin 2000) 1053–1054

3.36 M. Razafinimanana, M. Pacheco, M. Monthioux, H. Allouche, H. Lange, A. Huczko, A. Gleizes: Spectroscopic study of an electric arc with Gd and Fe doped anodes for the carbon nanotube formation, Proc. 25th Int. Conf. on Phenomena in Ionized Gases, Nagoya 2001, ed. by E. Goto (Nagoya Univ., Nagoya 2001) Extend. Abstr. **3** 297–298

3.37 M. Razafinimanana, M. Pacheco, M. Monthioux, H. Allouche, H. Lange, A. Huczko, P. Teulet, A. Gleizes, C. Goze, P. Bernier, T. Sogabe: Influence of doped graphite electrode in electric arc for the formation of single wall carbon nanotubes, Proc. 6th Eur. Conf. on Thermal Plasma Processes – Progress in Plasma Processing of Materials, New York 2000, ed. by P. Fauchais (Begell House, New York 2001) 649–654

3.38 M. Pacheco, H. Allouche, M. Monthioux, A. Razafinimanana, A. Gleizes: Correlation between the plasma characteristics and the morphology and structure of the carbon phases synthesised by electric arc discharge, Proc. 25th Biennial Conf. on Carbon, Lexington 2001 (Univ. of Kentucky, Lexington 2001) Extend. Abstr.(CD-Rom), Novel/14.1

3.39 M. Pacheco, M. Monthioux, M. Razafinimanana, L. Donadieu, H. Allouche, N. Caprais, A. Gleizes: New factors controlling the formation of single-wall carbon nanotubes by arc plasma, Proc. Carbon 2002 Int. Conf., Beijing 2002 (Tsinghua Univ., Beijing 2002) (CD-Rom/Oral/I014)

3.40 M. Monthioux, M. Pacheco, H. Allouche, M. Razafinimanana, N. Caprais, L. Donnadieu, A. Gleizes: New data about the formation of SWNTs by the electric arc method, Electronic Properties of Molecular Nanostructures, AIP Conf. Proc., Melville 2002, ed. by H. Kuzmany, J. Fink, M. Mehring, S. Roth (Springer, Berlin, Heidelberg 2002) 182–185

3.41 H. Lange, A. Huczko, M. Sioda, M. Pacheco, M. Razafinimanana, A. Gleizes: Influence of gadolinium on carbon arc plasma and formation of fullerenes and nanotubes, Plasma Chem. Plasma Process **22** (2002) 523–536

3.42 C. Journet: La production de nanotubes de carbone. Ph.D. Thesis (University of Montpellier II, Montpellier 1998)

3.43 T. Sogabe, T. Masuda, K. Kuroda, Y. Hirohaya, T. Hino, T. Ymashina: Preparation of B_4C-mixed graphite by pressureless sintering and its air oxidation behavior, Carbon **33** (1995) 1783–1788

3.44 M. Ishigami, J. Cumings, A. Zettl, S. Chen: A simple method for the continuous production of carbon nanotubes, Chem. Phys. Lett. **319** (2000) 457–459

3.45 Y. L. Hsin, K. C. Hwang, F. R. Chen, J. J. Kai: Production and in-situ metal filling of carbon nanotube in water, Adv. Mater. **13** (2001) 830–833

3.46 H. W. Zhu, X. S. Li, B. Jiang, C. L. Xu, C. L. Zhu, Y. F. Zhu, D. H. Wu, X. H. Chen: Formation of carbon nanotubes in water by the electric arc technique, Chem. Phys. Lett. **366** (2002) 664–669

3.47 W. K. Maser, P. Bernier, J. M. Lambert, O. Stephan, P. M. Ajayan, C. Colliex, V. Brotons, J. M. Planeix, B. Coq, P. Molinie, S. Lefrant: Elaboration and characterization of various carbon nanostructures, Synth. Met. **81** (1996) 243–250

3.48 J. Gavillet, A. Loiseau, J. Thibault, A. Maigné, O. Stéphan, P. Bernier: TEM study of the influence of the catalyst composition on the formation and growth of SWNT, Proc. Electronic Properties Novel Materials-XVI Int. Winterschool – AIP Conf. Proc., Melville 2002, ed. by H. Kuzmany, J. Fink, M. Mehring, S. Roth (Springer, Berlin, Heidelberg 2002) **633** 202–206

3.49 L. P. F. Chibante, A. Thess, J. M. Alford, M. D. Diener, R. E. Smalley: Solar generation of the fullerenes, J. Phys. Chem. **97** (1993) 8696–8700

3.50 C. L. Fields, J. R. Pitts, M. J. Hale, C. Bingham, A. Lewandowski, D. E. King: Formation of fullerenes in highly concentrated solar flux, J. Phys. Chem. **97** (1993) 8701–8702

3.51 P. Bernier, D. Laplaze, J. Auriol, L. Barbedette, G. Flamant, M. Lebrun, A. Brunelle, S. Della-Negra: Production of fullerenes from solar energy, Synth. Met. **70** (1995) 1455–1456

3.52 M. J. Heben, T. A. Bekkedhal, D. L. Schultz, K. M. Jones, A. C. Dillon, C. J. Curtis, C. Bingham, J. R. Pitts, A. Lewandowski, C. L. Fields: Production of single wall carbon nanotubes using concentrated sunlight, Proc. Symp. Recent Adv. Chem. Phys. Fullerenes Rel. Mater., Pennington 1996, ed. by K. M. Kadish, R. S. Ruoff (Electrochemical Society, Pennington 1996) INC **3** 803–811

3.53 D. Laplaze, P. Bernier, C. Journet, G. Vié, G. Flamant, E. Philippot, M. Lebrun: Evaporation of graphite using a solar furnace, Proc. 8th Int. Symp. Solar Concentrating Technol., Köln 1996, ed. by M. Becker, M. Balmer (C. F. Müller Verlag, Heidelberg 1997)

3.54 D. Laplaze, P. Bernier, W. K. Maser, G. Flamant, T. Guillard, A. Loiseau: Carbon nanotubes : The solar approach, Carbon **36** (1998) 685–688

3.55 T. Guillard, S. Cetout, L. Alvarez, J. L. Sauvajol, E. Anglaret, P. Bernier, G. Flamant, D. Laplaze: Production of carbon nanotubes by the solar route, Eur. Phys. J. **5** (1999) 251–256

3.56 G. Flamant, D. Luxembourg, C. Mas, D. Laplaze: Carbon nanotubes from solar energy; a chemical engineering approach of the scale up, Proc. Am. Inst. Chem. Eng. Ann. Meeting, Indianapolis 2002, in print

3.57 G. G. Tibbets, M. Endo, C. P. Beetz: Carbon fibers grown from the vapor phase: A novel material, SAMPE J. **22** (1989) 30

3.58 R. T. K. Baker: Catalytic growth of carbon filaments, Carbon **27** (1989) 315–323

3.59 R. T. K. Baker, P. S. Harris, R. B. Thomas, R. J. Waite: Formation of filamentous carbon from iron, cobalt, and chromium catalyzed decomposition of acetylene, J. Catal. **30** (1973) 86–95

3.60 T. Koyama, M. Endo, Y. Oyuma: Carbon fibers obtained by thermal decomposition of vaporized hydrocarbon, Jap. J. Appl. Phys. **11** (1972) 445–449

3.61 M. Endo, A. Oberlin, T. Koyama: High resolution electron microscopy of graphitizable carbon fiber prepared by benzene decomposition, Jap. J. Appl. Phys. **16** (1977) 1519–1523

3.62 N. M. Rodriguez: A review of catalytically grown carbon nanofibers, J. Mater. Res. **8** (1993) 3233–3250

3.63 W. R. Davis, R. J. Slawson, G. R. Rigby: An unusual form of carbon, Nature **171** (1953) 756

3.64 H. P. Boehm: Carbon from carbon monoxide disproportionation on nickel and iron catalysts; morphological studies and possible growth mechanisms, Carbon **11** (1973) 583–590

3.65 M. Audier, A. Oberlin, M. Coulon: Crystallographic orientations of catalytic particles in filamentous carbon; case of simple conical particles, J. Cryst. Growth **55** (1981) 546–549

3.66 M. Audier, M. Coulon: Kinetic and microscopic aspects of catalytic carbon growth, Carbon **23** (1985) 317–323

3.67 M. Audier, A. Oberlin, M. Coulon: Study of biconic microcrystals in the middle of carbon tubes obtained by catalytic disproportionation of CO, J. Cryst. Growth **57** (1981) 524–534

3.68 A. Thaib, G. A. Martin, P. Pinheiro, M. C. Schouler, P. Gadelle: Formation of carbon nanotubes from the carbon monoxide disproportionation reaction over CO/Al$_2$O′ and Co/SiO′ catalysts, Catal. Lett. **63** (1999) 135–141

3.69 P. Pinheiro, M. C. Schouler, P. Gadelle, M. Mermoux, E. Dooryhée: Effect of hydrogen on the orientation of carbon layers in deposits from the carbon monoxide disproportionation reaction over Co/Al$_2$O$_3$ catalysts, Carbon **38** (2000) 1469–1479

3.70 P. Pinheiro, P. Gadelle: Chemical state of a supported iron-cobalt catalyst during CO disproportionation. I. Thermodynamic study, J. Phys. Chem. Solids **62** (2001) 1015–1021

3.71 P. Pinheiro, P. Gadelle, C. Jeandey, J. L. Oddou: Chemical state of a supported iron-cobalt catalyst during CO disproportionation. II. Experimental study, J. Phys. Chem. Solids **62** (2001) 1023–1037

3.72 C. Laurent, E. Flahaut, A. Peigney, A. Rousset: Metal nanoparticles for the catalytic synthesis of carbon nanotubes, New J. Chem. **22** (1998) 1229–1237

3.73 A. Peigney, C. Laurent, F. Dobigeon, A. Rousset: Carbon nanotubes grown in situ by a novel catalytic method, J. Mater. Res. **12** (1997) 613–615

3.74 V. Ivanov, J. B. Nagy, P. Lambin, A. Lucas, X. B. Zhang, X. F. Zhang, D. Bernaerts, G. Van Tendeloo, S. Amelinckx, J. Van Landuyt: The study of nanotubules produced by catalytic method, Chem. Phys. Lett. **223** (1994) 329–335

3.75 V. Ivanov, A. Fonseca, J. B. Nagy, A. Lucas, P. Lambin, D. Bernaerts, X. B. Zhang: Catalytic production and purification of nanotubules having fullerene-scale diameters, Carbon **33** (1995) 1727–1738

3.76 K. Hernadi, A. Fonseca, J. B. Nagy, D. Bernaerts, A. Fudala, A. Lucas: Catalytic synthesis of carbon nanotubes using zeolite support, Zeolites **17** (1996) 416–423

3.77 H. Dai, A. G. Rinzler, P. Nikolaev, A. Thess, D. T. Colbert, R. E. Smalley: Single-wall nanotubes produced by metal-catalysed disproportionation of carbon monoxide, Chem. Phys. Lett. **260** (1996) 471–475
3.78 A. M. Cassel, J. A. Raymakers, J. Kong, H. Dai: Large scale CVD synthesis of single-walled carbon nanotubes, J. Phys. Chem. B **109** (1999) 6484–6492
3.79 B. Kitiyanan, W. E. Alvarez, J. H. Harwell, D. E. Resasco: Controlled production of single-wall carbon nanotubes by catalytic decomposition of CO on bimetallic Co-Mo catalysts, Chem. Phys. Lett. **317** (2000) 497–503
3.80 E. Flahaut: Synthèse par voir catalytique et caractérisation de composites nanotubes de carbone-metal-oxyde Poudres et matériaux denses. Ph.D. Thesis (Univers. Paul Sabatier, Toulouse 1999)
3.81 A. Govindaraj, E. Flahaut, C. Laurent, A. Peigney, A. Rousset, C. N. R. Rao: An investigation of carbon nanotubes obtained from the decomposition of methane over reduced $Mg_{1-x}M_xAl_2O_4$ spinel catalysts, J. Mater. Res. **14** (1999) 2567–2576
3.82 E. Flahaut, A. Peigney, C. Laurent, A. Rousset: Synthesis of single-walled carbon nanotube-Co-MgO composite powders and extraction of the nanotubes, J. Mater. Chem. **10** (2000) 249–252
3.83 J. Kong, A. M. Cassel, H. Dai: Chemical vapor deposition of methane for single-walled carbon nanotubes, Chem. Phys. Lett. **292** (1998) 567–574
3.84 R. Marangoni, P. Serp, R. Feurrer, Y. Kihn, P. Kalck, C. Vahlas: Carbon nanotubes produced by substrate free metalorganic chemical vapor deposition of iron catalyst and ethylene, Carbon **39** (2001) 443–449
3.85 R. Sen, A. Govindaraj, C. N. R. Rao: Carbon nanotubes by the metallocene route, Chem. Phys. Lett. **267** (1997) 276–280
3.86 Y. Y. Fan, H. M. Cheng, Y. L. Wei, G. Su, S. H. Shen: The influence of preparation parameters on the mass production of vapor grown carbon nanofibers, Carbon **38** (2000) 789–795
3.87 L. Ci, J. Wei, B. Wei, J. Liang, C. Xu, D. Wu: Carbon nanofibers and single-walled carbon nanotubes prepared by the floating catalyst method, Carbon **39** (2001) 329–335
3.88 O. A. Nerushev, M. Sveningsson, L. K. L. Falk, F. Rohmund: Carbon nanotube films obtained by thermal vapour deposition, J. Mater. Chem. **11** (2001) 1122–1132
3.89 F. Rohmund, L. K. L. Falk, F. E. B. Campbell: A simple method for the production of large arrays of aligned carbon nanotubes, Chem. Phys. Lett. **328** (2000) 369–373
3.90 G. G. Tibbets, C. A. Bernardo, D. W. Gorkiewicz, R. L. Alig: Role of sulfur in the production of carbon fibers in the vapor phase, Carbon **32** (1994) 569–576
3.91 W. Q. Han, P. Kholer-Riedlich, T. Seeger, F. Ernst, M. Ruhle, N. Grobert, W. K. Hsu, B. H. Chang, Y. Q. Zhu, H. W. Kroto, M. Terrones, H. Terrones: Aligned CN_x nanotubes by pyrolysis of ferrocene under NH_3 atmosphere, Appl. Phys. Lett. **77** (2000) 1807–1809
3.92 R. E. Smalley, J. H. Hafner, D. T. Colbert, K. Smith: Catalytic growth of single-wall carbon nanotubes from metal particles, US patent US19980601010903 (1998)
3.93 T. Kyotani, L. F. Tsai, A. Tomita: Preparation of ultrafine carbon tubes in nanochannels of an anodic aluminum oxide film, Chem. Mater. **8** (1996) 2109–2113
3.94 W. K. Hsu, J. P. Hare, M. Terrones, H. W. Kroto, D. R. M. Walton, P. J. F. Harris: Condensed-phase nanotubes, Nature **377** (1995) 687
3.95 W. S. Cho, E. Hamada, Y. Kondo, K. Takayanagi: Synthesis of carbon nanotubes from bulk polymer, Appl. Phys. Lett. **69** (1996) 278–279
3.96 Y. L. Li, Y. D. Yu, Y. Liang: A novel method for synthesis of carbon nanotubes: Low temperature solid pyrolysis, J. Mater. Res. **12** (1997) 1678–1680
3.97 S. Fan, M. Chapline, N. Franklin, T. Tombler, A. M. Cassel, H. Dai: Self-oriented regular arrays of carbon nanotubes and their field emission properties, Science **283** (1999) 512–514
3.98 M. L. Terranova, S. Piccirillo, V. Sessa, P. Sbornicchia, M. Rossi, S. Botti, D. Manno: Growth of single-walled carbon nanotubes by a novel technique using nanosized graphite as carbon source, Chem. Phys. Lett. **327** (2000) 284–290
3.99 R. L. Vander Wal, T. Ticich, V. E. Curtis: Diffusion flame synthesis of single-walled carbon nanotubes, Chem. Phys. Lett. **323** (2000) 217–223
3.100 Y. Y. Wei, G. Eres, V. I. Merkulov, D. H. Lowdens: Effect of film thickness on carbon nanotube growth by selective area chemical vapor deposition, Appl. Phys. Lett. **78** (2001) 1394–1396
3.101 W. Z. Li, S. S. Xie, L. X. Qian, B. H. Chang, B. S. Zou, W. Y. Zhou, R. A. Zha, G. Wang: Large scale synthesis of aligned carbon nanotubes, Science **274** (1996) 1701–1703
3.102 F. Zheng, L. Liang, Y. Gao, J. H. Sukamto, L. Aardahl: Carbon nanotubes synthesis using mesoporous silica templates, Nanolett. **2** (2002) 729–732
3.103 S. H. Jeong, O.-K. Lee, K. H. Lee, S. H. Oh, C. G. Park: Preparation of aligned carbon nanotubes with prescribed dimension: Template synthesis and sonication cutting approach, Chem. Mater. **14** (2002) 1859–1862
3.104 N. S. Kim, Y. T. Lee, J. Park, H. Ryu, H. J. Lee, S. Y. Choi, J. Choo: Dependence of vertically aligned growth of carbon nanotubes on catalyst, J. Phys. Chem. B **106** (2002) 9286–9290
3.105 C. J. Lee, D. W. Kim, T. J. Lee, Y. C. Choi, Y. S. Park, Y. H. Lee, W. B. Choi, N. S. Lee, G.-S. Park, J. M. Kim: Synthesis of aligned carbon nanotubes using thermal chemical vapor deposition, Chem. Phys. Lett. **312** (1999) 461–468
3.106 W. D. Zhang, Y. Wen, S. M. Liu, W. C. Tjiu, G. Q. Xu, L. M. Gan: Synthesis of vertically aligned carbon

nanotubes on metal deposited quartz plates, Carbon **40** (2002) 1981–1989
3.107 S. Huang, L. Dai, A. W. H. Mau: Controlled fabrication of large scale aligned carbon nanofiber/nanotube patterns by photolithography, Adv. Mater. **14** (2002) 1140–1143
3.108 R. Andrews, D. Jacques, A. M. Rao, F. Derbyshire, D. Qian, X. Fan, E. C. Dickey, J. Chen: Continous production of aligned carbon nanotubes: A step closer to commercial realization, Chem. Phys. Lett. **303** (1999) 467–474
3.109 C. N. R. Rao, R. Sen, B. C. Satishkumar, A. Govindaraj: Large aligned carbon nanotubes bundles from ferrocene pyrolysis, Chem. Commun. **15** (1998) 1525–1526
3.110 X. Zhang, A. Cao, B. Wei, Y. Li, J. Wei, C. Xu, D. Wu: Rapid growth of well-aligned carbon nanotube arrays, Chem. Phys. Lett. **362** (2002) 285–290
3.111 X. Zhang, A. Cao, Y. Li, C. Xu, J. Liang, D. Wu, B. Wei: Self-organized arrays of carbon nanotube ropes, Chem. Phys. Lett. **351** (2002) 183–188
3.112 K. S. Choi, Y. S. Cho, S. Y. Hong, J. B. Park, D. J. Kim: Effects of ammonia on the alignment of carbon nanotubes in metal-assisted chemical vapor deposition, J. Eur. Cer. Soc. **21** (2001) 2095–2098
3.113 M. Endo, H. W. Kroto: Formation of carbon nanofibers, J. Phys. Chem. **96** (1992) 6941–6944
3.114 R. S. Wagner: VLS mechanisms of crystal growth. In: *Whisker Technology*, ed. by P. Levit A. (Wiley, New York 1970) pp. 47–72
3.115 Y. H. Lee, S. G. Kim, D. Tomanek: Catalytic growth of single-wall carbon nanotubes: An ab initio study, Phys. Rev. Lett. **78** (1997) 2393–2396
3.116 H. Dai: Carbon Nanotubes: Synthesis, integration, and properties, Acc. Chem. Res. **35** (2002) 1035–1044
3.117 Y. Saito, M. Okuda, N. Fujimoto, T. Yoshikawa, M. Tomita, T. Hayashi: Single-wall carbon nanotubes growing radially from Ni fine particles formed by arc evaporation, Jpn. J. Appl. Phys. **33** (1994) L526–L529
3.118 J. Bernholc, C. Brabec, M. Buongiorno Nardelli, A. Malti, C. Roland, B. J. Yakobson: Theory of growth and mechanical properties of nanotubes, Appl. Phys. A **67** (1998) 39–46
3.119 M. Pacheco: Synthèse des nanotubes de carbone par arc electrique. Ph.D. Thesis (Université Toulouse III, Toulouse 2003)
3.120 K. Méténier, S. Bonnamy, F. Béguin, C. Journet, P. Bernier, L. M. de la Chapelle, O. Chauvet, S. Lefrant: Coalescence of single walled nanotubes and formation of multi-walled carbon nanotubes under high temperature treatments, Carbon **40** (2002) 1765–1773
3.121 P. G. Collins, P. Avouris: Nanotubes for electronics, Sci. Am. **283** (2000) 38–45
3.122 Q.-H. Yang, P. X. Hou, S. Bai, M. Z. Wang, H. M. Cheng: Adsorption and capillarity of nitrogen in aggregated multi-walled carbon nanotubes, Chem. Phys. Lett. **345** (2001) 18–24
3.123 S. Inoue, N. Ichikuni, T. Suzuki, T. Uematsu, K. Kaneko: Capillary condensation of N_2 on multiwall carbon nanotubes, J. Phys. Chem. **102** (1998) 4689–4692
3.124 M. Eswaramoorthy, R. Sen, C. N. R. Rao: A study of micropores in single-walled carbon nanotubes by the adsorption of gases and vapors, Chem. Phys. Lett. **304** (1999) 207–210
3.125 A. Peigney, Ch. Laurent, E. Flahaut, R. R. Bacsa, A. Rousset: Specific surface area of carbon nanotubes and bundles of carbon nanotubes, Carbon **39** (2001) 507–514
3.126 E. Frackowiak, S. Delpeux, K. Jurewicz, K. Szostak, D. Cazorla-Amoros, F. Béguin: Enhanced capacitance of carbon nanotubes through chemical activation, Chem. Phys. Lett. **336** (2002) 35–41
3.127 M. Muris, N. Dupont-Pavlosky, M. Bienfait, P. Zeppenfeld: Where are the molecules adsorbed on single-walled nanotubes?, Surf. Sci. **492** (2001) 67–74
3.128 A. Fujiwara, K. Ishii, H. Suematsu, H. Kataura, Y. Maniwa, S. Suzuki, Y. Achiba: Gas adsorption in the inside and outside of single-walled carbon nanotubes, Chem. Phys. Lett. **336** (2001) 205–211
3.129 C. M. Yang, H. Kanoh, K. Kaneko, M. Yudasaka, S. Iijima: Adsorption behaviors of HiPco single-walled carbon nanotubes aggregates for alcohol vapors, J. Phys. Chem. **106** (2002) 8994–8999
3.130 J. Zhao, A. Buldum, J. Han, J. P. Lu: Gas molecule adsorption in carbon nanotubes and nanotube bundles, Nanotechnology **13** (2002) 195–200
3.131 K. A. Williams, P. C. Eklund: Monte Carlo simulation of H_2 physisorption in finite diameter carbon nanotube ropes, Chem. Phys. Lett. **320** (2000) 352–358
3.132 A. Chambers, C. Park, R. T. K. Baker, N. Rodriguez: Hydrogen storage in graphite nanofibers, J. Phys. Chem. B. **102** (1998) 4253–4256
3.133 J. Hilding, E. A. Grulke, S. B. Sinnott, D. Qian, R. Andrews, M. Jagtoyen: Sorption of butane on carbon multiwall nanotubes at room temperature, Langmuir **17** (2001) 7540–7544
3.134 K. Masenelli-Varlot, E. McRae, N. Dupont-Pavlosky: Comparative adsorption of simple molecules on carbon nanotubes. Dependence of the adsorption properties on the nanotube morphology, Appl. Surf. Sci. **196** (2002) 209–215
3.135 D. J. Browning, M. L. Gerrard, J. B. Lakeman, I. M. Mellor, R. J. Mortimer, M. C. Turpin: Studies into the storage of hydrogen in carbon nanofibers: Proposal of a possible mechanism, Nanolett. **2** (2002) 201–205
3.136 F. H. Yang, R. T. Yang: Ab initio molecular orbital study of adsorption of atomic hydrogen on graphite: insight into hydrogen storage in carbon nanotubes, Carbon **40** (2002) 437–444
3.137 G. E. Froudakis: Why alkali-metal-doped carbon nanotubes possess high hydrogen uptake, Nanolett. **1** (2001) 531–533

3.138 H. Ulbricht, G. Moos, T. Hertel: Physisorption of molecular oxygen on single-wall carbon nanotube bundles and graphite, Phys. Rev. B **66** (2002) 075404-1–075404-7
3.139 H. Ulbricht, J. Kriebel, G. Moos, T. Hertel: Desorption kinetics and interaction of Xe with single-wall carbon nanotube bundles, Chem. Phys. Lett. **363** (2002) 252–260
3.140 R. Saito, G. Dresselhaus, M.S. Dresselhaus: *Physical Properties of Carbon Nanotubes* (Imperial College Press, London 1998)
3.141 A. Charlier, E. McRae, R. Heyd, M.F. Charlier, D. Moretti: Classification for double-walled carbon nanotubes, Carbon **37** (1999) 1779–1783
3.142 A. Charlier, E. McRae, R. Heyd, M.F. Charlier: Metal semi-conductor transitions under uniaxial stress for single- and double-walled carbon nanotubes, J. Phys. Chem. Solids **62** (2001) 439–444
3.143 H. Ajiki, T. Ando: Electronic states of carbon nanotubes, J. Phys. Soc. Jap. **62** (1993) 1255–1266
3.144 C.T. White, T.N. Todorov: Carbon nanotubes as long ballistic conductors, Nature **393** (1998) 240–242
3.145 S. Frank, P. Poncharal, Z.L. Wang, W.A. de Heer: Carbon nanotube quantum resistors, Science **280** (1998) 1744–1746
3.146 W. Liang, M. Bockrath, D. Bozovic, J.H. Hafner, M. Tinkham, H. Park: Fabry–Perot interference in a nanotube electron waveguide, Nature **411** (2001) 665–669
3.147 L. Langer, V. Bayot, E. Grivei, J.-P. Issi, J.-P. Heremans, C.H. Olk, L. Stockman, C. van Haesendonck, Y. Buynseraeder: Quantum transport in a multiwalled carbon nanotube, Phys. Rev. Lett. **76** (1996) 479–482
3.148 M. Bockrath, D.H. Cobden, J. Lu, A.G. Rinzler, R.E. Smalley, L. Balents, P.L. McEuen: Luttinger-liquid behaviour in carbon nanotubes, Nature **397** (1999) 598–601
3.149 K. Liu, S. Roth, G.S. Duesberg, G.T. Kim, D. Popa, K. Mukhopadhyay, R. Doome, J. B'Nagy: Antilocalization in multiwalled carbon nanotubes, Phys. Rev. B **61** (2000) 2375–2379
3.150 Y.A. Kasumov, R. Deblock, M. Kociak, B. Reulet, H. Bouchiat, I.I. Khodos, Y.B. Gorbatov, V.T. Volkov, C. Journet, M. Burghard: Supercurrents through single-walled carabon nanotubes, Science **284** (1999) 1508–1511
3.151 B.W. Alphenaar, K. Tsukagoshi, M. Wagner: Magnetoresistance of ferromagnetically contacted carbon nanotubes, Phys. Eng. **10** (2001) 499–504
3.152 P.M. Ajayan, M. Terrrones, A. de la Guardia, V. Hue, N. Grobert, B.Q. Wei, H. Lezec, G. Ramanath, T.W. Ebbesen: Nanotubes in a flash – Ignition and reconstruction, Science **296** (2002) 705
3.153 S. Berber, Y. Kwon, D. Tomanek: Unusually high thermal conductivity of carbon nanotubes, Phys. Rev. Lett. **84** (2000) 4613–4616
3.154 S.N. Song, X.K. Wang, R.P.H. Chang, J.B. Ketterson: Electronic properties of graphite nanotubules from galvanomagnetic effects, Phys. Rev. Lett. **72** (1994) 697–700
3.155 M.-F. Yu, O. Lourie, M.J. Dyer, K. Moloni, T.F. Kelley, R.S. Ruoff: Strength and breaking mechanism of multiwalled carbon nanotubes under tensile load, Science **287** (2000) 637–640
3.156 D.A. Walters, L.M. Ericson, M.J. Casavant, J. Liu, D.T. Colbert, K.A. Smith, R.E. Smalley: Elastic strain of freely suspended single-wall carbon nanotube ropes, Appl. Phys. Lett. **74** (1999) 3803–3805
3.157 B.G. Demczyk, Y.M. Wang, J. Cumingd, M. Hetamn, W. Han, A. Zettl, R.O. Ritchie: Direct mechanical measurement of the tensile strength and eleastic modulus of multiwalled carbon nanotubes, Mater. Sci. Eng. A **334** (2002) 173–178
3.158 R.P. Gao, Z.L. Wang, Z.G. Bai, W.A. De Heer, L.M. Dai, M. Gao: Nanomechanics of inidividual carbon nanotubes from pyrolytically grown arrays, Phys. Rev. Lett. **85** (2000) 622–625
3.159 M.M.J. Treacy, T.W. Ebbesen, J.M. Gibson: Exceptionally high Young's modulus observed for individual carbon nanotubes, Nature **381** (1996) 678–680
3.160 N. Yao, V. Lordie: Young's modulus of single-wall carbon nanotubes, J. Appl. Phys. **84** (1998) 1939–1943
3.161 O. Lourie, H.D. Wagner: Transmission electron microscopy observations of fracture of single-wall carbon nanotubes under axial tension, Appl. Phys. Lett. **73** (1998) 3527–3529
3.162 S.C. Tsang, Y.K. Chen, P.J.F. Harris, M.L.H. Green: A simple chemical method of opening and filling carbon nanotubes, Nature **372** (1994) 159–162
3.163 M. Montioux: Filling single-wall carbon nanotubes, Carbon **40** (2002) 1809–1823
3.164 W.K. Hsu, S.Y. Chu, E. Munoz-Picone, J.L. Boldu, Firth S., P. Franchi, B.P. Roberts, A. Shilder, H. Terrones, N. Grobert, Y.Q. Zhu, M. Terrones, M.E. McHenry, H.W. Kroto, D.R.M. Walton: Metallic behaviour of boron-containing carbon nanotubes, Chem. Phys. Lett. **323** (2000) 572–579
3.165 R. Czerw, M. Terrones, J.C. Charlier, X. Blasé, B. Foley, R. Kamalakaran, N. Grobert, H. Terrones, D. Tekleab, P.M. Ajayan, W. Blau, M. Rühle, D.L. Caroll: Identification of electron donor states, in N-doped carbon nanotubes, Nanolett. **1** (2001) 457–460
3.166 O. Stephan, P.M. Ajayan, C. Colliex, P. Redlich, J.M. Lambert, P. Bernier, P. Lefin: Doping graphitic and carbon nanotube structures with boron and nitrogen, Science **266** (1994) 1683–1685
3.167 A. Loiseau, F. Willaime, N. Demoncy, N. Schramchenko, G. Hug, C. Colliex, H. Pascard: Boron nitride nanotubes, Carbon **36** (1998) 743–752
3.168 C.C. Tang, L.M. de la Chapelle, P. Li, Y.M. Liu, H.Y. Dang, S.S. Fan: Catalytic growth of nanotube

and nanobamboo structures of boron nitride, Chem. Phys. Lett. **342** (2001) 492–496

3.169 K. Suenaga, C. Colliex, N. Demoncy, A. Loiseau, H. Pascard, F. Willaime: Synthesis of nanoparticles and nanotubes with well separated layers of boron-nitride and carbon, Science **278** (1997) 653–655

3.170 D. Golberg, Y. Bando, L. Bourgeois, K. Kurashima, T. Sato: Large-scale synthesis and HRTEM analysis of single-walled B-and N-doped carbon nanotube bundles, Carbon **38** (2000) 2017–2027

3.171 B. Burteaux, A. Claye, B.W. Smith, M. Monthioux, D.E. Luzzi, J.E. Fischer: Abundance of encapsulated C_{60} in single-wall carbon nanotubes, Chem. Phys. Lett. **310** (1999) 21–24

3.172 D. Ugarte, A. Châtelain, W.A. de Heer: Nanocapillarity and chemistry in carbon nanotubes, Science **274** (1996) 1897–1899

3.173 J. Cook, J. Sloan, M.L.H. Green: Opening and filling carbon nanotubes, Fuller. Sci. Technol. **5** (1997) 695–704

3.174 P.M. Ajayan, S. Iijima: Capillarity-induced filling of carbon nanotubes, Nature **361** (1993) 333–334

3.175 P.M. Ajayan, T.W. Ebbesen, T. Ichihashi, S. Iijima, K. Tanigaki, H. Hiura: Opening carbon nanotubes with oxygen and implications for filling, Nature **362** (1993) 522–525

3.176 S. Seraphin, D. Zhou, J. Jiao, J.C. Withers, R. Loufty: Yttrium carbide in nanotubes, Nature **362** (1993) 503

3.177 S. Seraphin, D. Zhou, J. Jiao, J.C. Withers, R. Loufty: Selective encapsulation of the carbides of yttrium and titanium into carbon nanoclusters, Appl. Phys. Lett. **63** (1993) 2073–2075

3.178 R.S. Ruoff, D.C. Lorents, B. Chan, R. Malhotra, S. Subramoney: Single-crystal metals encapsulated in carbon nanoparticles, Science **259** (1993) 346–348

3.179 A. Loiseau, H. Pascard: Synthesis of long carbon nanotubes filled with Se, S, Sb, and Ge by the arc method, Chem. Phys. Lett. **256** (1996) 246–252

3.180 N. Demoncy, O. Stephan, N. Brun, C. Colliex, A. Loiseau, H. Pascard: Filling carbon nanotubes with metals by the arc discharge method: The key role of sulfur, Eur. Phys. J. B **4** (1998) 147–157

3.181 C.H. Kiang, J.S. Choi, T.T. Tran, A.D. Bacher: Molecular nanowires of 1 nm diameter from capillary filling of single-walled carbon nanotubes, J. Phys. Chem. B **103** (1999) 7449–7551

3.182 Z.L. Zhang, B. Li, Z.J. Shi, Z.N. Gu, Z.Q. Xue, L.M. Peng: Filling of single-walled carbon nanotubes with silver, J. Mater. Res. **15** (2000) 2658–2661

3.183 A. Govindaraj, B.C. Satishkumar, M. Nath, C.N.R. Tao: Metal nanowires and intercalated metal layers in single-walled carbon nanotubes bundles, Chem. Mater. **12** (2000) 202–205

3.184 J. Mittal, M. Monthioux, H. Allouche: Synthesis of SWNT-based hybrid nanomaterials from photolysis-enhanced chemical processes, Chem. Phys. Lett. **339** (2001) 311–318

3.185 E. Dujardin, T.W. Ebbesen, H. Hiura, K. Tanigaki: Capillarity and wetting of carbon nanotubes, Science **265** (1994) 1850–1852

3.186 E. Flahaut, J. Sloan, K.S. Coleman, V.C. Williams, S. Friedrichs, N. Hanson, M.L.H. Green: 1D p-block halide crystals confined into single walled carbon nanotubes, Proc. Mater. Res. Soc. Symp. **633** (2001) A13.15.1–A13.15.6

3.187 J. Sloan, A.I. Kirkland, J.L. Hutchison, M.L.H. Green: Integral atomic layer architectures of 1D crystals inserted into single walled carbon nanotubes, Chem. Commun. (2002) 1319–1332

3.188 J. Sloan, M.C. Novotny, S.R. Bailey, G. Brown, C. Xu, V.C. Williams, S. Friedrichs, E. Flahaut, R.L. Callender, A.P.E. York, K.S. Coleman, M.L.H. Green, R.E. Dunin-Borkowski, J.L. Hutchison: Two layer 4 : 4 co-ordinated KI crystals grown within single walled carbon nanotubes, Chem. Phys. Lett. **329** (2000) 61–65

3.189 G. Brown, S.R. Bailey, J. Sloan, C. Xu, S. Friedriechs, E. Flahaut, K.S. Coleman, J.L. Hutchinson, R.E. Dunin-Borkowski, M.L.H. Green: Electron beam induced in situ clusterisation of 1D $ZrCl_4$ chains within single-walled carbon nanotubes, Chem. Commun. **9** (2001) 845–846

3.190 B.W. Smith, M. Monthioux, D.E. Luzzi: Encapsulated C_{60} in carbon nanotubes, Nature **396** (1998) 323–324

3.191 B.W. Smith, D.E. Luzzi: Formation mechanism of fullerene peapods and coaxial tubes: A path to large scale synthesis, Chem. Phys. Lett. **321** (2000) 169–174

3.192 K. Hirahara, K. Suenaga, S. Bandow, H. Kato, T. Okazaki, H. Shinohara, S. Iijima: One-dimensional metallo-fullerene crystal generated inside single-walled carbon nanotubes, Phys. Rev. Lett. **85** (2000) 5384–5387

3.193 B.W. Smith, M. Monthioux, D.E. Luzzi: Carbon nanotube encapsulated fullerenes: A unique class of hybrid material, Chem. Phys. Lett. **315** (1999) 31–36

3.194 D.E. Luzzi, B.W. Smith: Carbon cage structures in single wall carbon nanotubes: A new class of materials, Carbon **38** (2000) 1751–1756

3.195 S. Bandow, M. Takisawa, K. Hirahara, M. Yudasoka, S. Iijima: Raman scattering study of double-wall carbon nanotubes derived from the chains of fullerenes in single-wall carbon nanotubes, Chem. Phys. Lett. **337** (2001) 48–54

3.196 B.W. Smith, D.E. Luzzi, Y. Achiba: Tumbling atoms and evidence for charge transfer in $La_2@C_{80}@$SWNT, Chem. Phys. Lett. **331** (2000) 137–142

3.197 K. Suenaga, M. Tence, C. Mory, C. Colliex, H. Kato, T. Okazaki, H. Shinohara, K. Hirahara, S. Bandow, S. Iijima: Element-selective single atom imaging, Science **290** (2000) 2280–2282

3.198 D.E. Luzzi, B.W. Smith, R. Russo, B.C. Satishkumar, F. Stercel, N.R.C. Nemes: Encapsulation of metallofullerenes and metallocenes in carbon nanotubes, Proc. Electronic Properties of Novel Materials-XVI Int. Winterschool – AIP Conf. Proc., Melville 2001, ed. by

H. Kuzmany, J. Fink, M. Mehring, S. Roth (Springer, Berlin, Heidelberg 2001) **591** 622–626

3.199 G.H. Jeong, R. Hatakeyama, T. Hirata, K. Tohji, K. Motomiya, N. Sato, Y. Kawazoe: Structural deformation of single-walled carbon nanotubes and fullerene encapsulation due to magnetized plasma ion irradiation, Appl. Phys. Lett. **79** (2001) 4213–4215

3.200 Y.P. Sun, K. Fu, Y. Lin, W. Huang: Functionalized carbon nanotubes: Properties and applications, Acc. Chem. Res. **35** (2002) 1095–1104

3.201 J. Chen, M.A. Hamon, M. Hui, C. Yongsheng, A.M. Rao, P.C. Eklund, R.C. Haddon: Solution properties of single-walled carbon nanotubes, Science **282** (1998) 95–98

3.202 J. Chen, A.M. Rao, S. Lyuksyutov, M.E. Itkis, M.A. Hamon, H. Hu, R.W. Cohn, P.C. Eklund, D.T. Colbert, R.E. Smalley, R.C. Haddon: Dissolution of full-length single-walled carbon nanotubes, J. Phys. Chem. B **105** (2001) 2525–2528

3.203 F. Pompeo, D.E. Resasco: Water solubilization of single-walled carbon nanotubes by functionalization with glucosamine, Nanolett. **2** (2002) 369–373

3.204 Y.P. Sun, W. Huang, Y. Lin, K. Fu, A. Kitaygorodskiy, L.A. Riddle, Y.J. Yu, D.L. Carroll: Soluble dendron-functionalized carbon nanotubes: Preparation, characterization, and properties, Chem. Mater. **13** (2001) 2864–2869

3.205 K. Fu, W. Huang, Y. Lin, L.A. Riddle, D.L. Carroll, Y.P. Sun: Defunctionalization of functionalized carbon nanotubes, Nanolett. **1** (2001) 439–441

3.206 P.W. Chiu, G.S. Duesberg, U. Dettlaff-Weglikowska, S. Roth: Interconnection of carbon nanotubes by chemical functionalization, Appl. Phys. Lett. **80** (2002) 3811–3813

3.207 E.T. Mickelson, C.B. Huffman, A.G. Rinzler, R.E. Smalley, R.H. Hauge, J.L. Margrave: Fluorination of single-wall carbon nanotubes, Chem. Phys. Lett. **296** (1998) 188–194

3.208 V.N. Khabashesku, W.E. Billups, J.L. Margrave: Fluorination of single-wall carbon nanotubes and subsequent derivatization reactions, Acc. Chem. Res. **35** (2002) 1087–1095

3.209 P.J. Boul, J. Liu, E.T. Michelson, C.B. Huffman, L.M. Ericson, I.W. Chiang, K.A. Smith, D.T. Colbert, R.H. Hauge, J.L. Margrave, R.E. Smalley: Reversible side-wall functionalization of buckytubes, Chem. Phys. Lett. **310** (1999) 367–372

3.210 J.L. Bahr, J. Yang, D.V. Kosynkin, M.J. Bronikowski, R.E. Smalley, J.M. Tour: Functionalization of carbon nanotubes by electrochemical reduction of aryl diazonium salts: A bucky paper electrode, J. Am. Chem. Soc. **123** (2001) 6536–6542

3.211 M. Holzinger, O. Vostrowsky, A. Hirsch, F. Hennrich, M. Kappes, R. Weiss, F. Jellen: Sidewall functionalization of carbon nanotubes, Angew. Chem. Int. Ed. **40** (2001) 4002–4005

3.212 Y. Chen, R.C. Haddon, S. Fang, A.M. Rao, P.C. Eklund, W.H. Lee, E.C. Dickey, E.A. Grulke, J.C. Pendergrass, A. Chavan, B.E. Haley, R.E. Smalley: Chemical attachment of organic functional groups to single-walled carbon nanotube material, J. Mater. Res. **13** (1998) 2423–2431

3.213 C. Velasco-Santos, A.L. Martinez-Hernandez, M. Lozada-Cassou, A. Alvarez-Castillo, V.M. Castano: Chemical functionalization of carbon nanotubes through an organosilane, Nanotechnology **13** (2002) 495–498

3.214 A. Star, J.F. Stoddart, D. Steuerman, M. Diehl, A. Boukai, E.W. Wong, X. Yang, S.W. Chung, H. Choi, J.R. Heath: Preparation and properties of polymer-wrapped single-walled carbon nanotubes, Angew. Chem. Int. Ed. **41** (2002) 1721–1725

3.215 J.H. Hafner, C.L. Cheung, A.T. Wooley, C.M. Lieber: Structural and functional imaging with carbon nanotube AFM probes, Progr. Biophys. Molec. Biol. **77** (2001) 73–110

3.216 S.S. Wong, E. Joselevich, A.T. Woodley, C.L. Cheung, C.M. Lieber: Covalently functionalized nanotubes as nanometre-size probes in chemistry and biology, Nature **394** (1998) 52–55

3.217 C.L. Cheung, J.H. Hafner, C.M. Lieber: Carbon nanotube atomic force microscopy tips: Direct growth by chemical vapor deposition and application to high-resolution imaging, Proc. Nat. Acad. Sci. (2000) 973809–973813

3.218 W.A. de Heer, A. Châtelain, D. Ugarte: A carbon nanotube field-emission electron source, Science **270** (1995) 1179–1180

3.219 J.M. Bonard, J.P. Salvetat, T. Stockli, W.A. de Heer, L. Forro, A. Chatelain: Field emission from single-wall carbon nanotube films, Appl. Phys. Lett. **73** (1998) 918–920

3.220 W. Zhu, C. Bower, O. Zhou, G. Kochanski, S. Jin: Large curent density from carbon nanotube field emitters, Appl. Phys. Lett. **75** (1999) 873–875

3.221 Y. Saito, R. Mizushima, T. Tanaka, K. Tohji, K. Uchida, M. Yumura, S. Uemura: Synthesis, structure, and field emission of carbon nanotubes, Fuller. Sci. Technol. **7** (1999) 653–664

3.222 J. Kong, N.R. Franklin, C. Zhou, M.G. Chapline, S. Peng, K. Cho, H. Dai: Nanotube molecular wire as chemical sensors, Science **287** (2000) 622–625

3.223 P.G. Collins, K. Bradley, M. Ishigami, A. Zettl: Extreme oxygen sensitivity of electronic properties of carbon nanotubes, Science **287** (2000) 1801–1804

3.224 H. Chang, J.D. Lee, S.M. Lee, Y.H. Lee: Adsorption of NH_3 and NO_2 molecules on carbon nanotubes, Appl. Phys. Lett. **79** (2001) 3863–3865

3.225 O.K. Varghese, P.D. Kichambre, D. Gong, K.G. Ong, E.C. Dickey, C.A. Grimes: Gas sensing characteristics of multi-wall carbon nanotubes, Sens. Actuators B **81** (2001) 32–41

3.226 K.G. Ong, K. Zeng, C.A. Grimes: A wireless, passive carbon nanotube-based gas sensor, IEEE Sens. J. **2/2** (2002) 82–88

3.227 J. Kong, M. G. Chapline, H. Dai: Functionalized carbon nanotubes for molecular hydrogen sensors, Adv. Mater. **13** (2001) 1384–1386

3.228 F. Rodriguez-Reinoso: The role of carbon materials in heterogeneous catalysis, Carbon **36** (1998) 159–175

3.229 E. Auer, A. Freund, J. Pietsch, T. Tacke: Carbon as support for industrial precious metal catalysts, Appl. Catal. A **173** (1998) 259–271

3.230 J. M. Planeix, N. Coustel, B. Coq, B. Botrons, P. S. Kumbhar, R. Dutartre, P. Geneste, P. Bernier, P. M. Ajayan: Application of carbon nanotubes as supports in heterogeneous catalysis, J. Am. Chem. Soc. **116** (1994) 7935–7936

3.231 B. Coq, J. M. Planeix, V. Brotons: Fullerene-based materials as new support media in heterogeneous catalysis by metals, Appl. Catal. A **173** (1998) 175–183

3.232 G. Mestl, N. I. Maksimova, N. Keller, V. V. Roddatis, R. Schlögl: Carbon nanofilaments in heterogeneous catalysis: An industrial application for new carbon materials?, Angew. Chem. Int. Ed. Engl. **40** (2001) 2066–2068

3.233 J. E. Fischer, A. T. Johnson: Electronic properties of carbon nanotubes, Current Opinion Solid State Mater. Sci. **4** (1999) 28–33

3.234 M. Menon, A. N. Andriotis, G. E. Froudakis: Curvature dependence of the metal catalyst atom interaction with carbon nanotubes walls, Chem. Phys. Lett. **320** (2000) 425–434

3.235 R. Giordano, Ph. Serp, Ph. Kalck, Y. Kihn, J. Schreiber, C. Marhic, J.-L. Duvail: Preparation of Rhodium supported on carbon canotubes catalysts via surface mediated organometallic reaction, Eur. J. Inorg. Chem. (2003) 610–617

3.236 T. Kyotani, S. Nakazaki, W.-H. Xu, A. Tomita: Chemical modification of the inner walls of carbon nanotubes by HNO_3 oxidation, Carbon **39** (2001) 782–785

3.237 Z. J. Liu, Z. Y. Yuan, W. Zhou, L. M. Peng, Z. Xu: Co/carbon nanotubes monometallic system: The effects of oxidation by nitric acid, Phys. Chem. Chem. Phys. **3** (2001) 2518–2521

3.238 H.-B. Chen, J. D. Lin, Y. Cai, X. Y. Wang, J. Yi, J. Wang, G. Wei, Y. Z. Lin, D. W. Liao: Novel multi-walled nanotubes-supported and alkali-promoted Ru catalysts for ammonia synthesis under atmospheric pressure, Appl. Surf. Sci. **180** (2001) 328–335

3.239 Y. Zhang, H. B. Zhang, G. D. Lin, P. Chen, Y. Z. Yuan, K. R. Tsai: Preparation, characterization and catalytic hydroformylation properties of carbon nanotubes-supported Rh-phosphine catalyst, Appl. Catal. A **187** (1999) 213–224

3.240 M. S. Dresselhaus, K. A. Williams, P. C. Eklund: Hydrogen adsorption in carbon materials, Mater. Res. Soc. Bull. (November 1999) 45–50

3.241 H.-M. Cheng, Q.-H. Yang, C. Liu: Hydrogen storage in carbon nanotubes, Carbon **39** (2001) 1447–1454

3.242 G. G. Tibbetts, G. P. Meisner, C. H. Olk: Hydrogen storage capacity of carbon nanotubes, filaments, and vapor-grown fibers, Carbon **39** (2001) 2291–2301

3.243 F. L. Darkrim, P. Malbrunot, G. P. Tartaglia: Review of hydrogen storage adsorption in carbon nanotubes, Int. J. Hydrogen Energy **27** (2002) 193–202

3.244 C. Park, P. E. Anderson, C. D. Tan, R. Hidalgo, N. Rodriguez: Further studies of the interaction of hydrogen with graphite nanofibers, J. Phys. Chem. B **103** (1999) 10572–1058

3.245 C. C. Ahn, Y. Ye, B. V. Ratnakumar, C. Witham, R. C. Bowman, B. Fultz: Hydrogen adsorption measurements on graphite nanofibers, Appl. Phys. Lett. **73** (1998) 3378–3380

3.246 Q. Wang, J. K. Johnson: Computer simulations of hydrogen adsorption on graphite nanofibers, J. Phys. Chem. B **103** (1999) 277–281

3.247 M. Rzepka, P. Lamp, M. A. de la Casa-Lillo: Physisorption of hydrogen on microporous carbon and carbon nanotubes, J. Phys. Chem. B **102** (1998) 10894–10898

3.248 A. C. Dillon, K. M. Jones, T. A. Bekkedahl, C. H. Kiang, D. S. Bethune, M. J. Heben: Storage of hydrogen in single-walled carbon nanotubes, Nature **386** (1997) 377–379

3.249 Y. Ye, C. C. Ahn, C. Witham, R. C. Bowman, B. Fultz, J. Liu, A. G. Rinzler, D. Colbert, K. A. Smith, R. E. Smalley: Hydrogen adsorption and cohesive energy of single-walled carbon nanotubes, Appl. Phys. Lett. **74** (1999) 2307–2309

3.250 C. Liu, Y. Y. Fan, M. Liu, H. T. Cong, H. M. Cheng, M. S. Dresselhaus: Hydrogen storage in single-walled carbon nanotubes at room temperature, Science **286** (1999) 1127–1129

3.251 M. Hirscher, M. Becher, M. Haluska, U. Dettlaff-Weglikowska, A. Quintel, G. S. Duesberg, Y. M. Choi, P. Dwones, M. Hulman, S. Roth, I. Stepanek, P. Bernier: Hydrogen storage in sonicated carbon materials, Appl. Phys. A **72** (2001) 129–132

3.252 P. A. Gordon, R. B. Saeger: Molecular modeling of adsorptive energy storage: Hydrogen storage in single-walled carbon nanotubes, Ind. Eng. Chem. Res. **38** (1999) 4647–4655

3.253 P. Marinelli, R. Pellenq, J. Conard: H stocké dans les carbones un site légèrement métastable, matériaux, Tours 2002 AF-14-020

3.254 S. M. Lee, H. Y. Lee: Hydrogen storage in single-walled carbon nanotubes, Appl. Phys. Lett. **76** (2000) 2877–2879

3.255 S. M. Lee, K. S. Park, Y. C. Choi, Y. S. Park, J. M. Bok, D. J. Bae, K. S. Nahm, Y. G. Choi, C. S. Yu, N. Kim, T. Frauenheim, Y. H. Lee: Hydrogen adsorption in carbon nanotubes, Synth. Met. **113** (2000) 209–216

3.256 H. Cheng, G. P. Pez, A. C. Cooper: Mechanism of hydrogen sorption in single-walled carbon nanotubes, J. Am. Chem. Soc. **123** (2001) 5845–5846

3.257 A. Kusnetzova, D. B. Mawhinney, V. Naumenko, J. T. Yates, J. Liu, R. E. Smalley: Enhancement of adsorption inside of single-walled nanotubes: Opening the entry ports, Chem. Phys. Lett. **321** (2000) 292–296

3.258 G. E. Gadd, M. Blackford, S. Moricca, N. Webb, P. J. Evans, A. M. Smith, G. Jacobsen, S. Leung, A. Day, Q. Hua: The world's smallest gas cylinders?, Science **277** (1997) 933–936

3.259 Z. Mao, S. B. Sinnott: A computational study of molecular diffusion and dynamic flow through carbon nanotubes, J. Phys. Chem. B **104** (2000) 4618–4624

3.260 Z. Mao, S. B. Sinnott: Separation of organic molecular mixtures in carbon nanotubes and bundles: Molecular dynamics simulations, J. Phys. Chem. B **105** (2001) 6916–6924

3.261 C. Gu, G.-H. Gao, Y. X. Yu, T. Nitta: Simulation for separation of hydrogen and carbon monoxide by adsorption on single-walled carbon nanotubes, Fluid Phase Equilibria **194/197** (2002) 297–307

3.262 R. Q. Long, R. T. Yang: Carbon nanotubes as superior sorbent for dioxine removal, J. Am. Chem. Soc. **123** (2001) 2058–2059

3.263 Y. H. Li, S. Wang, A. Cao, D. Zhao, X. Zhang, C. Xu, Z. Luan, D. Ruan, J. Liang, D. Wu, B. Wei: Adsorption of fluoride from water by amorphous alumina supported on carbon nanotubes, Chem. Phys. Lett. **350** (2001) 412–416

3.264 Y. H. Li, S. Wang, J. Wei, X. Zhang, C. Xu, Z. Luan, D. Wu, B. Wei: Lead adsorption on carbon nanotubes, Chem. Phys. Lett. **357** (2002) 263–266

3.265 C. Park, E. S. Engel, A. Crowe, T. R. Gilbert, N. M. Rodriguez: Use of carbon nanofibers in the removal of organic solvents from water, Langmuir **16** (2000) 8050–8056

3.266 M. P. Mattson, R. C. Haddon, A. M. Rao: Molecular functionalization of carbon nanotubes and use as substrates for neuronal growth, J. Molec. Neurosci. **14** (2000) 175–182

3.267 J. J. Davis, M. L. H. Green, H. A. O. Hill, Y. C. Leung, P. J. Sadler, J. Sloan, A. V. Xavier, S. C. Tsang: The immobilization of proteins in carbon nanotubes, Inorg. Chim. Acta **272** (1998) 261–266

3.268 R. J. Chen, Y. Zhang, D. Wang, H. Dai: Noncovalent sidewall functionalization of single-walled carbon nanotubes for protein immobilization, J. Am. Chem. Soc. **123** (2001) 3838–3839

3.269 M. Shim, N. W. S. Kam, R. J. Chen, Y. Li, H. Dai: Functionalization of carbon nanotubes for biocompatibility and biomolecular recognition, Nanolett. **2** (2002) 285–288

3.270 C. Dwyer, M. Guthold, M. Falvo, S. Washburn, R. Superfine, D. Erie: DNA-functionalized single-walled carbon nanotubes, Nanotechnology **13** (2002) 601–604

3.271 H. Huang, S. Taylor, K. Fu, Y. Lin, D. Zhang, T. W. Hanks, A. M. Rao, Y. Sun: Attaching proteins to carbon nanotubes via diimide-activated amidation, Nanolett. **2** (2002) 311–314

3.272 C. V. Nguyen, L. Delzeit, A. M. Cassell, J. Li, J. Han, M. Meyyappan: Preparation of nucleic acid functionalized carbon nanotube arrays, Nanolett. **2** (2002) 1079–1081

3.273 B. R. Azamian, J. J. Davis, K. S. Coleman, C. B. Bagshaw, M. L. H. Green: Bioelectrochemical single-walled carbon nanotubes, J. Am. Chem. Soc. **124** (2002) 12664–12665

3.274 C. L. Xu, B. Q. Wei, R. Z. Ma, J. Liang, X. K. Ma, D. H. Wu: Fabrication of aluminum-carbon nanotube composites and their electrical properties, Carbon **37** (1999) 855–858

3.275 T. Kuzumaki, K. Miyazawa, H. Ichinose, K. Ito: Processing of carbon nanotube reinforced aluminum composite, J. Mater. Res. **13** (1998) 2445–2449

3.276 T. Kuzumaki, O. Ujiie, H. Ichinose, K. Ito: Mechanical characteristics and preparation of carbon nanotube fiber-reinforced Ti composite, Adv. Eng. Mater. **2** (2000) 416–418

3.277 S. R. Dong, J. P. Tu, X. B. Zhang: An investigation of the sliding wear behavior of Cu-matrix composite reinforced by carbon nanotubes, Mater. Sci. Eng. A **313** (2001) 83–87

3.278 Y. B. Li, Q. Ya, B. Q. Wei, J. Liang, D. H. Wu: Processing of a carbon nanotubes-Fe82P18 metallic glass composite, J. Mater. Sci. Lett. **17** (1998) 607–609

3.279 A. Peigney: Tougher ceramics with carbon nanotubes, Nature Mater. **2** (2003) 15–16

3.280 G. D. Zhan, J. D. Kuntz, J. Wan, A. K. Mukherjee: Single-wall carbon nanotubes as attractive toughening agents in alumina-based composites, Nature Mater. **2** (2003) 38–42

3.281 J. Sun, L. Gao, W. Li: Colloidal processing of carbon nanotube/alumina composites, Chem. Mater. **14** (2002) 5169–5172

3.282 V. G. Gavalas, R. Andrews, D. Bhattacharyya, L. G. Bachas: Carbon nanotube sol-gel composite materials, Nanolett. **1** (2001) 719–721

3.283 S. Rul, Ch. Laurent, A. Peigney, A. Rousset: Carbon nanotubes prepared in-situ in a cellular ceramic by the gelcasting-foam method, J. Eur. Ceram. Soc. **23** (2003) 1233–1241

3.284 R. Z. Ma, J. Wu, B. Q. Wei, J. Liang, D. H. Wu: Processing and properties of carbon nanotube/nano-SiC ceramic, J. Mater. Sci. **33** (1998) 5243–5246

3.285 C. Laurent, A. Peigney, O. Dumortier, A. Rousset: Carbon nanotubes-Fe-alumina nanocomposites. Part II: Microstructure and mechanical properties of the hot-pressed composites, J. Eur. Ceram. Soc. **18** (1998) 2005–2013

3.286 A. Peigney, C. Laurent, A. Rousset: Synthesis and characterization of alumina matrix nanocomposites containing carbon nanotubes, Key Eng. Mater. **132–136** (1997) 743–746

3.287 A. Peigney, C. Laurent, E. Flahaut, A. Rousset: Carbon nanotubes in novel ceramic matrix nanocomposites, Ceram. Intern. **26** (2000) 677–683

3.288 S. Rul: Synthèse de composites nanotubes de carbone-métal-oxyde. Ph.D. Thesis (Université Toulouse III, Toulouse 2002)

3.289 E. Flahaut, A. Peigney, C. Laurent, C. Marliere, F. Chastel, A. Rousset: Carbon nanotube-metal-oxide nanocomposites: Microstructure, electrical conductivity and mechanical properties, Acta Mater. **48** (2000) 3803–3812

3.290 R. R. Bacsa, C. Laurent, A. Peigney, W. S. Bacsa, T. Vaugien, A. Rousset: High specific surface area carbon nanotubes from catalytic chemical vapor deposition process, Chem. Phys. Lett. **323** (2000) 566–571

3.291 P. Coquay, A. Peigney, E. De Grave, R. E. Vandenberghe, C. Laurent: Carbon nanotubes by a CVD method. Part II: Formation of nanotubes from (Mg,Fe)O catalysts, J. Phys. Chem. B **106** (2002) 13199–13210

3.292 E. Flahaut, Ch. Laurent, A. Peigney: Double-walled carbon nanotubes in composite powders, J. Nanosci. Nanotech. **3** (2003) 151–158

3.293 A. Peigney, P. Coquay, E. Flahaut, R. E. Vandenberghe, E. De Grave, C. Laurent: A study of the formation of single- and double-walled carbon nanotubes by a CVD method, J Phys. Chem. B **105** (2001) 9699–9710

3.294 R. W. Siegel, S. K. Chang, B. J. Ash, J. Stone, P. M. Ajayan, R. W. Doremus, L. S. Schadler: Mechanical behavior of polymer and ceramic matrix nanocomposites, Scr. Mater. **44** (2001) 2061–2064

3.295 A. Peigney, E. Flahaut, C. Laurent, F. Chastel, A. Rousset: Aligned carbon nanotubes in ceramic-matrix nanocomposites prepared by high-temperature extrusion, Chem. Phys. Lett. **352** (2002) 20–25

3.296 S. L. Huang, M. R. Koblischka, K. Fossheim, T. W. Ebbesen, T. H. Johansen: Microstructure and flux distribution in both pure and carbon-nanotube-embedded $Bi_2Sr_2CaCu_2O_{8+\delta}$ superconductors, Physica C **311** (1999) 172–186

3.297 L. L. Yowell: Thermal management in ceramics: Synthesis and characterization of a zirconia-carbon nanotube composite, Proc Mater. Res. Soc. Symp. **633** (2002) A17.4.1–A17.4.6

3.298 A. Weidenkaff, S. G. Ebbinghaus, T. Lippert: $Ln_{1-x}A_xCoO_3$ ($Ln = Er, La$; $A = Ca, Sr$)/carbon nanotube composite materials applied for rechargeable Zn/Air batteries, Chem. Mater. **14** (2002) 1797–1805

3.299 D. S. Lim, J. W. An, H. J. Lee: Effect of carbon nanotube addition on the tribological behavior of carbon/carbon composites, Wear **252** (2002) 512–517

3.300 R. Andrews, D. Jacques, A. M. Rao, T. Rantell, F. Derbyshire, Y. Chen, J. Chen, R. C. Haddon: Nanotube composite carbon fibers, Appl. Phys. Lett. **75** (1999) 1329–1331

3.301 P. M. Ajayan, O. Stephan, C. Colliex, D. Trauth: Aligned carbon nanotube arrays formed by cutting a polymer resin-nanotube composite, Science **265** (1994) 1212–14

3.302 R. Haggenmueller, H. H. Gommans, A. G. Rinzler, J. E. Fischer, K. I. Winey: Aligned single-wall carbon nanotubes in composites by melt processing methods, Chem. Phys. Lett. **330** (2000) 219–225

3.303 L. S. Schadler, S. C. Giannaris, P. M. Ajayan: Load transfer in carbon nanotube epoxy composites, Appl. Phys. Lett. **73** (1998) 3842–3844

3.304 S. J. V. Frankland, A. Caglar, D. W. Brenner, M. Griebel: Molecular simulation of the influence of chemical cross-links on the shear strength of carbon nanotube-polymer interfaces, J. Phys. Chem. B **106** (2002) 3046–3048

3.305 H. D. Wagner: Nanotube-polymer adhesion: A mechanics approach, Chem. Phys. Lett. **361** (2002) 57–61

3.306 P. M. Ajayan, L. S. Schadler, C. Giannaris, A. Rubio: Single-walled carbon nanotube-polymer composites: Strength and weakness, Adv. Mater. **12** (2000) 750–753

3.307 X. Gong, J. Liu, S. Baskaran, R. D. Voise, J. S. Young: Surfactant-assisted processing of carbon nanotube/polymer composites, Chem. Mater. **12** (2000) 1049–1052

3.308 E. T. Thostenson, W. Z. Li, D. Z. Wang, Z. F. Ren, T. W. Chou: Carbon nanotube/carbon fiber hybrid multiscale composites, J. Appl. Phys. **91** (2002) 6034–6037

3.309 M. J. Biercuk, M. C. Llaguno, M. Radosavljevic, J. K. Hyun, A. T. Johnson, J. E. Fischer: Carbon nanotube composites for thermal management, Appl. Phys. Lett. **80** (2002) 2767–2769

3.310 S. Barrau, P. Demont, A. Peigney, C. Laurent, C. Lacabanne: DC and AC conductivity of carbon nanotube – Epoxy composites macromolecules, Macromol. **36** (2003) 5187–5194

3.311 J. Sandler, M. S. P. Shaffer, T. Prasse, W. Bauhofer, K. Schulte, A. H. Windle: Development of a dispersion process for carbon nanotubes in an epoxy matrix and the resulting electrical properties, Polymer **40** (1999) 5967–5971

3.312 Z. Jin, K. P. Pramoda, G. Xu, S. H. Goh: Dynamic mechanical behavior of melt-processed multi-walled carbon nanotube/poly(methyl methacrylate) composites, Chem. Phys. Lett. **337** (2001) 43–47

3.313 Z. Jin, K. P. Pramoda, S. H. Goh, G. Xu: Poly(vinylidene fluoride)-assisted melt-blending of multi-walled carbon nanotube/poly(methyl methacrylate) composites, Mat. Res. Bull. **37** (2002) 271–278

3.314 C. A. Cooper, D. Ravich, D. Lips, J. Mayer, H. D. Wagner: Distribution and alignment of carbon nanotubes and nanofibrils in a polymer matrix, Compos. Sci. Technol. **62** (2002) 1105–1112

3.315 J. M. Benoit, B. Corraze, S. Lefrant, W. J. Blau, P. Bernier, O. Chauvet: Transport properties of PMMA-carbon nanotubes composites, Synth. Met. **121** (2001) 1215–1216

3.316 J. M. Benoit, B. Corraze, O. Chauvet: Localization, Coulomb interactions, and electrical heating in single-wall carbon nanotubes/polymer composites, Phys. Rev. B: Cond. Matter. Mater. Phys. **65** (2002) 241405/1–241405/4

3.317 M. S. P. Shaffer, A. H. Windle: Fabrication and characterization of carbon nanotube/poly(vinyl alcohol) composites, Adv. Mater. **11** (1999) 937–941

3.318 L. Jin, C. Bower, O. Zhou: Alignment of carbon nanotubes in a polymer matrix by mechanical stretching, Appl. Phys. Lett. **73** (1998) 1197–1199

3.319 H. D. Wagner, O. Lourie, Y. Feldman, R. Tenne: Stress-induced fragmentation of multiwall carbon nanotubes in a polymer matrix, Appl. Phys. Lett. **72** (1998) 188–190

3.320 H. D. Wagner, O. Lourie, X. F. Zhou: Macrofragmentation and microfragmentation phenomena in composite materials, Composites Part A **30A** (1998) 59–66

3.321 J. R. Wood, Q. Zhao, H. D. Wagner: Orientation of carbon nanotubes in polymers and its detection by Raman spectroscopy, Composites Part A **32A** (2001) 391–399

3.322 Q. Zhao, J. R. Wood, H. D. Wagner: Using carbon nanotubes to detect polymer transitions, J. Polym. Sci., Part B: Polym. Phys. **39** (2001) 1492–1495

3.323 M. Cochet, W. K. Maser, A. M. Benito, M. A. Callejas, M. T. Martinesz, J. M. Benoit, J. Schreiber, O. Chauvet: Synthesis of a new polyaniline/nanotube composite: In-situ polymerisation and charge transfer through site-selective interaction, Chem. Commun. **16** (2001) 1450–1451

3.324 D. Qian, E. C. Dickey, R. Andrews, T. Rantell: Load transfer and deformation mechanisms in carbon nanotube-polystyrene composites, Appl. Phys. Lett. **76** (2000) 2868–2870

3.325 C. Wei, D. Srivastava, K. Cho: Thermal expansion and diffusion coefficients of carbon nanotube-polymer composites, Los Alamos Nat. Lab., Preprint Archive, Condensed Matter (archiv:cond-mat/0203349) (2002) 1–11

3.326 C. Pirlot, I. Willems, A. Fonseca, J. B. Nagy, J. Delhalle: Preparation and characterization of carbon nanotube/polyacrylonitrile composites, Adv. Eng. Mater. **4** (2002) 109–114

3.327 L. Cao, H. Chen, M. Wang, J. Sun, X. Zhang, F. Kong: Photoconductivity study of modified carbon nanotube/oxotitanium phthalocyanine composites, J. Phys. Chem. B **106** (2002) 8971–8975

3.328 I. Musa, M. Baxendale, G. A. J. Amaratunga, W. Eccleston: Properties of regular poly(3-octylthiophene)/multi-wall carbon nanotube composites, Synth. Met. **102** (1999) 1250

3.329 E. Kymakis, I. Alexandou, G. A. J. Amaratunga: Single-walled carbon nanotube-polymer composites: Electrical, optical and structural investigation, Synth. Met. **127** (2002) 59–62

3.330 K. Yoshino, H. Kajii, H. Araki, T. Sonoda, H. Take, S. Lee: Electrical and optical properties of conducting polymer-fullerene and conducting polymer-carbon nanotube composites, Fuller. Sci. Technol. **7** (1999) 695–711

3.331 S. A. Curran, P. M. Ajayan, W. J. Blau, D. L. Carroll, J. N. Coleman, A. B. Dalton, A. P. Davey, A. Drury, B. McCarthy, S. Maier, A. Strevens: A composite from poly(m-phenylenevinylene-co-2,5-dioctoxy-p-phenylenevinylene) and carbon nanotubes. A novel material for molecular optoelectronics, Adv. Mater. **10** (1998) 1091–1093

3.332 P. Fournet, D.F. O'Brien, J. N. Coleman, H. H. Horhold, W.J. Blau: A carbon nanotube composite as an electron transport layer for M3EH-PPV based light-emitting diodes, Synth. Met. **121** (2001) 1683–1684

3.333 H. S. Woo, R. Czerw, S. Webster, D. L. Carroll, J. Ballato, A. E. Strevens, D. O'Brien, W.J. Blau: Hole blocking in carbon nanotube-polymer composite organic light-emitting diodes based on poly (m-phenylene vinylene-co-2, 5-dioctoxy-p-phenylene vinylene), Appl. Phys. Lett. **77** (2000) 1393–1395

3.334 H. S. Woo, R. Czerw, S. Webster, D. L. Carroll, J. W. Park, J. H. Lee: Organic light emitting diodes fabricated with single wall carbon nanotubes dispersed in a hole conducting buffer: The role of carbon nanotubes in a hole conducting polymer, Synth. Met. **116** (2001) 369–372

3.335 H. Ago, K. Petritsch, M. S. P. Shaffer, A. H. Windle, R. H. Friend: Composites of carbon nanotubes and conjugated polymers for photovoltaic devices, Adv. Mater. **11** (1999) 1281–1285

3.336 B. Vigolo, A. Pénicaud, C. Coulon, C. Sauder, R. Pailler, C. Journet, P. Bernier, P. Poulin: Macroscopic fibers and ribbons of oriented carbon nanotubes, Science **290** (2000) 1331–1334

3.337 B. Vigolo, P. Poulin, M. Lucas, P. Launois, P. Bernier: Improved structure and properties of single-wall carbon nanotube spun fibers, Appl. Phys. Lett. **11** (2002) 1210–1212

3.338 P. Poulin, B. Vigolo, P. Launois: Films and fibers of oriented single wall nanotubes, Carbon **40** (2002) 1741–1749

3.339 K. Jiang, Q. Li, S. Fan: Spinning continuous carbon nanotube yarn, Nature **419** (2002) 801

3.340 P. Lambin, A. Fonseca, J. P. Vigneron, J. B'Nagy, A. A. Lucas: Structural and electronic properties of bent carbon nanotubes, Chem. Phys. Lett. **245** (1995) 85–89

3.341 L. Chico, V. H. Crespi, L. X. Benedict, S. G. Louie, M. L. Cohen: Pure carbon nanoscale devices: Nanotube heterojunctions, Phys. Rev. Lett. **76** (1996) 971–974

3.342 Z. Yao, H. W. C. Postma, L. Balents, C. Dekker: Carbon nanotube intramolecular junctions, Nature **402** (1999) 273–276

3.343 S. J. Tans, A. R. M. Verschueren, C. Dekker: Room temperature transistor based on single carbon nanotube, Nature **393** (1998) 49–52

3.344 R. Martel, T. Schmidt, H. R. Shea, T. Hertel, P. Avouris: Single and multi-wall carbon nanotube field effect transistors, Appl. Phys. Lett. **73** (1998) 2447–2449

3.345 V. Derycke, R. Martel, J. Appenzeller, P. Avouris: Carbon nanotube inter- and intramolecular logic gates, Nanolett. **1** (2001) 453–456

3.346 P. Kim, C. M. Lieber: Nanotube nanotweezers, Science **286** (1999) 2148–2150

3.347 P. G. Collins, M. S. Arnold, P. Avouris: Engineering carbon nanotubes using electrical breakdown, Science **292** (2001) 706–709

3.348 R. H. Baughman, C. Changxing, A. A. Zakhidov, Z. Iqbal, J. N. Barisci, G. M. Spinks, G. G. Wallace, A. Mazzoldi, D. de Rossi, A. G. Rinzler, O. Jaschinki S. Roth, M. Kertesz: Carbon nanotubes actuators, Science **284** (1999) 1340–1344

3.349 Y. Gao, Y. Bando: Carbon nanothermometer containing gallium, Nature **415** (2002) 599

3.350 C. Niu, E. K. Sichel, R. Hoch, D. Moy, H. Tennent: High power electro-chemical capacitors based on carbon nanotube electrodes, Appl. Phys. Lett. **70** (1997) 1480–1482

3.351 E. Frackowiak, F. Béguin: Electrochemical storage of energy in carbon nanotubes and nanostructured carbons, Carbon **40** (2002) 1775–1787

3.352 E. Frackowiak, K. Jurewicz, K. Szostak, S. Delpeux, F. Béguin: Nanotubular materials as electrodes for supercapacitors, Fuel Process. Technol. **77** (2002) 213–219

Nanowires

4. Nanowires

Nanowires are especially attractive for nanoscience studies as well as for nanotechnology applications. Nanowires, compared to other low dimensional systems, have two quantum confined directions while still leaving one unconfined direction for electrical conduction. This allows them to be used in applications which require electrical conduction, rather than tunneling transport. Because of their unique density of electronic states, nanowires in the limit of small diameters are expected to exhibit significantly different optical, electrical, and magnetic properties from their bulk 3-D crystalline counterparts. Increased surface area, very high density of electronic states and joint density of states near the energies of their van Hove singularities, enhanced exciton binding energy, diameter-dependent bandgap, and increased surface scattering for electrons and phonons are just some of the ways in which nanowires differ from their corresponding bulk materials. Yet the sizes of nanowires are typically large enough (>1 nm in the quantum confined direction) to have local crystal structures closely related to their parent materials, thereby allowing theoretical predictions about their properties to be made on the basis of an extensive literature relevant to their bulk properties.

Not only do nanowires exhibit many properties that are similar to, and others that are distinctly different from those of their bulk counterparts, nanowires have the advantage from an applications standpoint in that some of the materials parameters critical for certain properties can be independently controlled in nanowires but not in their bulk counterparts. Certain properties can also be enhanced nonlinearly in small diameter nanowires by exploiting the singular aspects of the 1-D electronic density of states. Furthermore, nanowires have been shown to provide a promising framework for applying the "bottom-up" approach [4.1] for the design of nanostructures for nanoscience investigations and for potential nanotechnology applications. Driven by: (1) these new research and development opportunities, (2) the smaller and smaller length scales now being used in the semiconductor, opto-electronics, and magnetics industries, and (3) the dramatic development of the biotechnology industry where the action is also at the nanoscale, the nanowire research field has developed with exceptional speed in the last few years. A review of the current status of nanowire research is therefore of significant broad interest at the present time. This review aims to focus on nanowire properties that differ from those of their parent crystalline bulk materials, with an eye toward possible applications that might emerge from the unique properties of nanowires and from future discoveries in this field.

4.1 **Synthesis** 100
  4.1.1 Template-Assisted Synthesis 100
  4.1.2 VLS Method for Nanowire Synthesis 105
  4.1.3 Other Synthesis Methods 107
  4.1.4 Hierarchical Arrangement and Superstructures of Nanowires 108

4.2 **Characterization and Physical Properties of Nanowires** 110
  4.2.1 Structural Characterization 110
  4.2.2 Transport Properties 115
  4.2.3 Optical Properties 126

4.3 **Applications** 131
  4.3.1 Electrical Applications 131
  4.3.2 Thermoelectric Applications 133
  4.3.3 Optical Applications 134
  4.3.4 Chemical and Biochemical Sensing Devices 137
  4.3.5 Magnetic Applications 137

4.4 **Concluding Remarks** 138

References 138

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

For quick reference, examples of typical nanowires that have been synthesized and studied are listed in Table 4.1. Also of use to the reader are review articles that focus on a comparison between nanowire and nanotube properties [4.2] and the many reviews that have been written about carbon nanotubes [4.3–5], which can be considered as a model one-dimensional system.

4.1 Synthesis

In this section we survey the most common synthetic approaches that have successfully afforded high quality nanowires of a large variety of materials (see Table 4.1). In Sect. 4.1.1, we discuss methods that make use of various templates with nanochannels to confine the nanowire growth to two dimensions. In Sect. 4.1.2, we present the synthesis of nanowires by the vapor-liquid-solid mechanism and its many variations. In Sect. 4.1.3, we present examples of other synthetic methods of general applicability. The last part of this section (Sect. 4.1.4) features several approaches that have been developed to organize nanowires into simple architectures.

4.1.1 Template-Assisted Synthesis

The template-assisted synthesis of nanowires is a conceptually simple and intuitive way to fabricate nanostructures [4.7–10]. These templates contain very small cylindrical pores or voids within the host material, and the empty spaces are filled with the chosen material, which adopts the pore morphology, to form nanowires. In this section, we describe the templates first and then describe strategies for filling them to make nanowires.

Template Synthesis

In template-assisted synthesis of nanostructures, the chemical stability and mechanical properties of the template, as well as the diameter, uniformity, and density of the pores are important characteristics to consider. Templates frequently used for nanowire synthesis include anodic alumina (Al_2O_3), nano-channel glass, ion track-etched polymers, and mica films.

Porous anodic alumina templates are produced by anodizing pure Al films in various acids [4.11–13]. Under carefully chosen anodization conditions, the resulting oxide film possesses a regular hexagonal array of parallel and nearly cylindrical channels, as shown in Fig. 4.1a. The self-organization of the pore structure in an anodic alumina template involves two coupled processes: pore formation with uniform diameters and pore ordering. The pores form with uniform diameters because of a delicate balance between electric-field-enhanced diffusion, which determines the growth rate of the alumina, and dissolution of the alumina into the acidic electrolyte [4.12]. The pores are believed to self-order because of mechanical stress at the aluminum-alumina interface due to volume expansion during the anodization. This stress produces a repulsive force between the pores, causing them to arrange in a hexagonal lattice [4.14]. Depending on the anodization conditions, the pore diameter can be systematically varied from < 10 nm up to 200 nm with a pore density in the range of 10^9–10^{11} pores/cm^2 [4.11, 12, 15, 16]. Many groups have shown that the pore size distribution and the pore ordering of the anodic alumina templates can be significantly improved by a two-step anodization technique [4.6, 17, 18] in which the aluminum oxide layer is dissolved after the first anodization in an acidic solution followed by a second anodization under the same conditions.

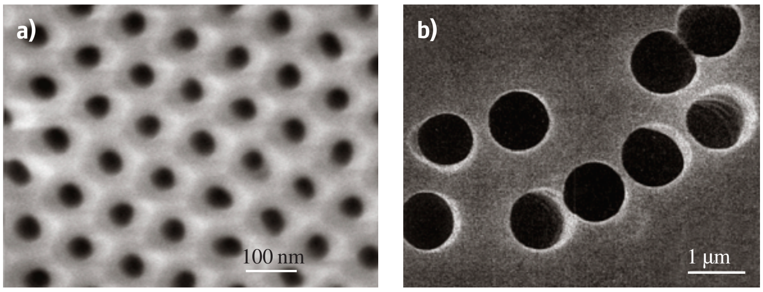

Fig. 4.1 **(a)** SEM images of the top surfaces of porous anodic alumina templates anodized with an average pore diameter of 44 nm [4.6]. **(b)** SEM image of the particle track-etched polycarbonate membrane, with a pore diameter of 1 µm [4.7]

Table 4.1 Selected syntheses of nanowires by material

Material	Growth Technique	Reference
Ag	DNA-template, redox	[4.23]
	template, pulsed ECD[a]	[4.24]
Au	template, ECD[a]	[4.25, 26]
Bi	stress-induced	[4.27]
	template, vapor-phase	[4.28]
	template, ECD[a]	[4.29], [4.30], [4.31]
	template, pressure-injection	[4.32], [4.16, 33]
Bi_2Te_3	template, dc ECD[a]	[4.34]
CdS	liquid-phase (surfactant), recrystallization	[4.35]
	template, ac ECD[a]	[4.36], [4.37]
CdSe	liquid-phase (surfactant), redox	[4.38]
	template, ac ECD[a]	[4.39], [4.40]
Cu	vapor deposition	[4.41]
	template, ECD[a]	[4.42]
Fe	template, ECD[c]	[4.15], [4.43]
	shadow deposition	[4.44]
GaN	template, CVD[c]	[4.45]
	VLS[b]	[4.46, 47]
GaAs	template, liquid/vapor OMCVD[d]	[4.48]
Ge	high-T, high-P liquid-phase, redox	[4.49]
	VLS[b]	[4.50]
	oxide-assisted	[4.51]
InAs	template, liquid/vapor OMCVD[d]	[4.48]
InP	VLS[b]	[4.52]
Mo	step decoration, ECD[a]+ redox	[4.53]
Ni	template, ECD[a]	[4.22], [4.31, 54]
PbSe	liquid phase	[4.55]
Pd	step decoration, ECD[a]	[4.56]
Se	liquid-phase, recrystallization	[4.57]
	template, pressure injection	[4.58]
Si	VLS[b]	[4.59]
	laser-ablation VLS[b]	[4.60]
	oxide-assisted	[4.61]
	low-T VLS[b]	[4.62]
Zn	template, vapor-phase	[4.63]
	template, ECD[a]	[4.64]
ZnO	VLS[b]	[4.65]
	template, ECD[a]	[4.64, 66]

[a] Electrochemical deposition

[b] Vapor-liquid-solid growth

[c] Chemical vapor deposition

[d] Organometallic chemical vapor deposition

Another type of porous template commonly used for nanowire synthesis is the template type fabricated by chemically etching particle tracks originating from ion bombardment [4.19], such as track-etched polycarbonate membranes (Fig. 4.1b) [4.7,20,21] and also mica films [4.22].

Other porous materials can be used as host templates for nanowire growth, as discussed by *Ozin* [4.8]. Nanochannel glass (NCG), for example, contains a regular hexagonal array of capillaries similar to the pore structure in anodic alumina with a packing density as high as 3×10^{10} pores/cm^2 [4.9]. Porous Vycor glass that contains an interconnected network of pores less than 10 nm was also employed for the early study of nanostructures [4.68]. Mesoporous molecular sieves [4.69], termed MCM-41, possess hexagonally packed pores with very small channel diameters that can be varied between 2 nm and 10 nm. Conducting organic filaments have been fabricated in the nanochannels of MCM-41 [4.70]. Recently the DNA molecule has also been used as a template for growing nanometer-sized wires [4.23].

Diblock copolymers, polymers that consist of two chain segments with different properties, have also been utilized as templates for nanowire growth. When the two segments are immiscible in each other, phase segregation occurs, and depending on their volume ratio, spheres, cylinders, and lamellae may self-assemble. To form self-assembled arrays of nanopores, copolymers composed of polystyrene and polymethylmethacrylate [P(S-b-MMA)] [4.71] were used. By applying an electric field while the copolymer was heated above the glass transition temperature of the two constituent polymers, the self-assembled cylinders of PMMA could be aligned with their main axis perpendicular to the film. Selective removal of the PMMA component afforded the preparation of 14-nm-diameter ordered pore arrays with a packing density of 1.9×10^{11} cm^{-3}.

Nanowire Template Assisted Growth by Pressure Injection

The pressure injection technique is often employed when fabricating highly crystalline nanowires from a low-melting point material and when using porous templates with robust mechanical strength. In the high-pressure injection method, the nanowires are formed by pressure injecting the desired material in liquid form into the evacuated pores of the template. Due to the heating and the pressurization processes, the templates used for the pressure injection method must be chemically stable and be able to maintain their structural integrity at high temperatures and high pressures. Anodic aluminum oxide films and nano-channel glass are two typical materials used as templates in conjunction with the pressure injection filling technique. Metal nanowires (Bi, In, Sn, and Al) and semiconductor nanowires (Se, Te, GaSb, and Bi_2Te_3) have been fabricated in anodic aluminum oxide templates using this method [4.32,58,67].

The pressure P required to overcome the surface tension for the liquid material to fill the pores with a diameter d_W is determined by the Washburn equation [4.72]:

$$d_W = -4\gamma\cos\theta/P\ , \tag{4.1}$$

where γ is the surface tension of the liquid, and θ is the contact angle between the liquid and the template. To reduce the required pressure and to maximize the filling factor, some surfactants are used to decrease the surface tension and the contact angle. For example, the introduction of Cu in the Bi melt can facilitate filling the

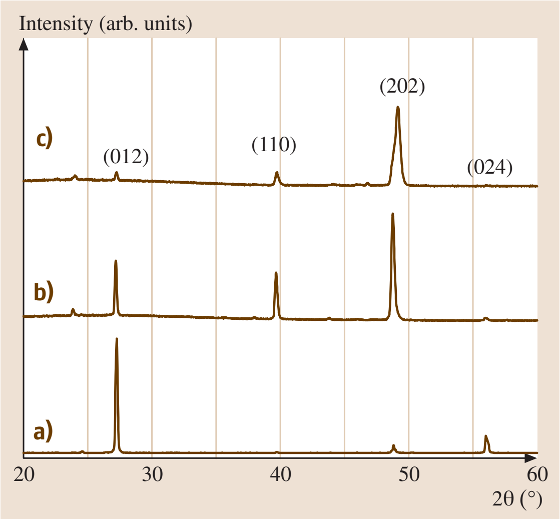

Fig. 4.2a–c XRD patterns of bismuth/anodic alumina nano-composites with average bismuth wire diameters of **(a)** 40 nm, **(b)** 52 nm, and **(c)** 95 nm [4.67]. The Miller indices corresponding to the lattice planes of bulk Bi are indicated above the individual peaks. The majority of the Bi nanowires are oriented along the $[10\bar{1}1]$ and $[01\bar{1}2]$ directions for $d_W \geq 60$ nm and $d_W \leq 50$ nm, respectively [4.16,67]. The existence of more than one dominant orientation in the 52-nm Bi nanowires is attributed to the transitional behavior of *intermediate*-diameter nanowires as the preferential growth orientation is shifted from $[10\bar{1}1]$ to $[01\bar{1}2]$ with decreasing d_W

pores in the anodic alumina template with liquid Bi and can increase the number of nanowires prepared [4.16]. But some of the surfactants might cause contamination problems and therefore should be avoided. Nanowires produced by the pressure injection technique usually possess high crystallinity and a preferred crystal orientation along the wire axis. For example, Fig. 4.2 shows the X-ray diffraction (XRD) patterns of Bi nanowire arrays of three different wire diameters with an injection pressure of $\sim 5{,}000$ psi [4.67], demonstrating that the major ($> 80\%$) crystal orientation of the wire axes in the 95-nm and 40-nm diameter Bi nanowire arrays are, respectively, normal to the (202) and (012) lattice planes. These are denoted by $[10\bar{1}1]$ and $[01\bar{1}2]$ when using a hexagonal unit cell, suggesting a wire-diameter-dependent crystal growth direction. On the other hand, 30 nm Bi nanowires produced using a much higher pressure of $> 20{,}000$ psi show a different crystal orientation of (001) along the wire axis [4.33], indicating that the preferred crystal orientation may also depend on the applied pressure, with the most dense packing direction along the wire axis for the highest applied pressure.

Electrochemical Deposition

The electrochemical deposition technique has attracted increasing attention as an alternative method for fabricating nanowires. Traditionally, electrochemistry has been used to grow thin films on conducting surfaces. Since electrochemical growth is usually controllable in the direction normal to the substrate surface, this method can be readily extended to fabricate 1-D or 0-D nanostructures, if the deposition is confined within the pores of an appropriate template. In the electrochemical methods, a thin conducting metal film is first coated on one side of the porous membrane to serve as the cathode for electroplating. The length of the deposited nanowires can be controlled by varying the duration of the electroplating process. This method has been used to synthesize a wide variety of nanowires e.g., metals (Bi [4.21, 29]; Co [4.73, 74]; Fe [4.15,75]; Cu [4.20,76]; Ni [4.22,73]; Ag [4.24,77]; Au [4.25,26]); conducting polymers [4.7,29]; superconductors (Pb [4.78]); semiconductors (CdS [4.37]); and even superlattice nanowires with A/B constituents (such as Cu/Co [4.20,76]) have been synthesized electrochemically (see Table 4.1).

In the electrochemical deposition process, the chosen template has to be chemically stable in the electrolyte during the electrolysis process. Cracks and defects in the templates are detrimental to the nanowire growth, since the deposition processes primarily occur in the more accessible cracks, leaving most of the nanopores unfilled. Particle track-etched mica films or polymer membranes are typical templates used in the simple dc electrolysis. To use anodic aluminum oxide films in the dc electrochemical deposition, the insulating barrier layer that separates the pores from the bottom aluminum substrate has to be removed, and a metal film is then evaporated onto the back of the template membrane [4.79]. Compound nanowire arrays, such as Bi_2Te_3, have been fabricated in alumina templates with a high filling factor using the dc electrochemical deposition [4.34]. Figures 4.3a and b, respectively, show the top view and cross-sectional SEM images of a Bi_2Te_3 nanowire array [4.34]. The light areas are associated with Bi_2Te_3 nanowires, the dark regions denote empty pores, and the surrounding gray matrix is alumina.

Surfactants are also used with electrochemical deposition when necessary. For example, when using templates derived from PMMA/PS diblock copolymers (see above), methanol is used as a surfactant is used to facilitate pore filling [4.71], thereby achieving $\sim 100\%$ filling factor.

It is also possible to employ an ac electrodeposition method in anodic alumina templates without the removal of the barrier layer, by utilizing the rectifying

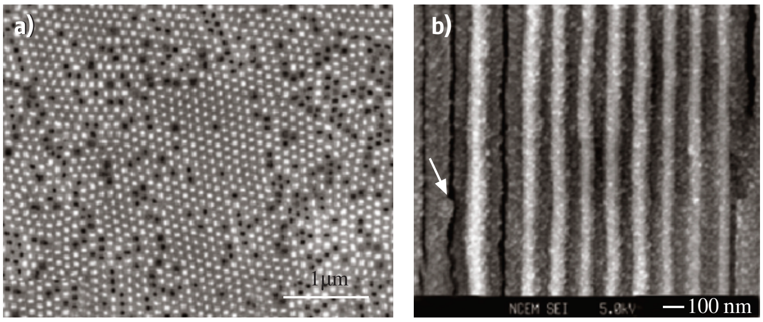

Fig. 4.3 **(a)** SEM image of a Bi_2Te_3 nanowire array in cross section showing a relatively high pore filling factor. **(b)** SEM image of a Bi_2Te_3 nanowire array composite along the wire axis [4.34]

properties of the oxide barrier. In ac electrochemical deposition, although the applied voltage is sinusoidal and symmetric, the current is greater during the cathodic half-cycles, making deposition dominant over the stripping, which occurs in the subsequent anodic half-cycles. Since no rectification occurs at defect sites, the deposition and stripping rates are equal, and no material is deposited. Hence, the difficulties associated with cracks are avoided. In this fashion, metals, such as Co [4.74] and Fe [4.15, 75], and semiconductors, such as CdS [4.37], have been deposited into the pores of anodic aluminum oxide templates without removing the barrier layer.

In contrast to nanowires synthesized by the pressure injection method, nanowires fabricated by the electrochemical process are usually polycrystalline, with no preferred crystal orientations, as observed by XRD studies. Some exceptions exist, however. For example, polycrystalline CdS nanowires, fabricated by an ac electrodeposition method in anodic alumina templates [4.37], possibly have a preferred wire growth orientation along the *c*-axis. In addition, *Xu* et al. have prepared a number of single-crystal II-VI semiconductor nanowires, including CdS, CdSe, and CdTe, by dc electrochemical deposition in anodic alumina templates with a nonaqueous electrolyte [4.36, 40]. Furthermore, single-crystal Pb nanowires can be formed by pulse electrodeposition under over-potential conditions, but no specific crystal orientation along the wire axis was observed [4.78]. The use of pulse currents is believed to be advantageous for the growth of crystalline wires because the metal ions in the solution can be regenerated between the electrical pulses and, therefore, uniform deposition conditions can be produced for each deposition pulse. Similarly, single crystal Ag nanowires were fabricated by pulsed electro-deposition [4.24].

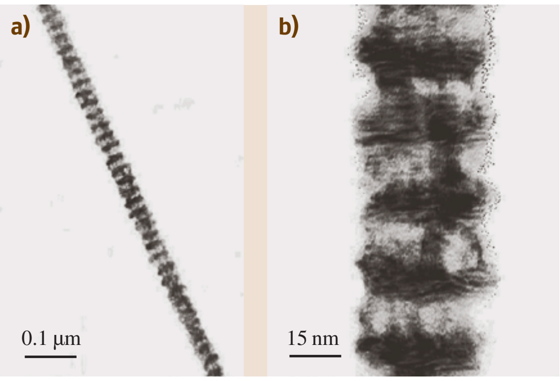

Fig. 4.4 **(a)** TEM image of a single Co(10 nm)/Cu(10 nm) multilayered nanowire. **(b)** A selected region of the sample at high magnification [4.76]

One advantage of the electrochemical deposition technique is the possibility of fabricating multilayered structures within nanowires. By varying the cathodic potentials in the electrolyte that contains two different kinds of ions, different metal layers can be controllably deposited. In this fashion, Co/Cu multilayered nanowires have been synthesized [4.20, 76]. Figure 4.4 shows TEM images of a single Co/Cu nanowire of about 40 nm in diameter [4.76]. The light bands represent Co-rich regions, and the dark bands represent Cu-rich layers. This electrodeposition method provides a low-cost approach to preparing multilayered 1-D nanostructures.

Vapor Deposition

Vapor deposition of nanowires includes physical vapor deposition (PVD) [4.28], chemical vapor deposition (CVD) [4.80], and metallorganic chemical vapor deposition (MOCVD) [4.48]. Like electrochemical deposition, vapor deposition is usually capable of preparing smaller-diameter (≤ 20 nm) nanowires than pressure injection methods, since it does not rely on the high pressure and surface tension involved to insert the material into the pores.

In the physical vapor deposition technique, the material to be filled is first heated to produce a vapor, which is then introduced through the pores of the template and cooled to solidify. Using an especially designed experimental setup [4.28], nearly single-crystal Bi nanowires in anodic aluminum templates with pore diameters as small as 7 nm have been synthesized, and these Bi nanowires were found to possess a preferred crystal growth orientation along the wire axis, similar to the Bi nanowires prepared by pressure injection [4.16, 28].

Compound materials that result from two reacting gases have also been prepared by the chemical vapor deposition (CVD) technique. For example, single-crystal GaN nanowires have been synthesized in anodic alumina templates through a gas reaction of Ga_2O vapor with a flowing ammonia atmosphere [4.45, 80]. A different liquid/gas phase approach has been used to prepare polycrystalline GaAs and InAs nanowires in a nanochannel glass array [4.48]. In this method, the nanochannels are filled with one liquid precursor (e.g., Me_3Ga or Et_3In) via a capillary effect, and the nanowires are formed within the template by reactions between the liquid precursor and the other gas reactant (e.g., AsH_3).

Nanotube Synthesis with Templates and as Templates

Recently carbon nanotubes, an important class of 1-D nanostructures, have been fabricated within the pores of anodic alumina templates using a chemical vapor deposition technique [4.81–84] to form highly ordered two-dimensional carbon nanotube arrays. A small amount of metal catalyst (e.g., Co) is first electrochemically deposited on the bottom of the pores. The templates are then placed in a furnace and heated to $\sim 700-800\,^{\circ}\mathrm{C}$ with a flowing gas consisting of a mixture of N_2 and acetylene (C_2H_2) or ethylene (C_2H_4). The hydrocarbon molecules are pyrolyzed to form carbon nanotubes in the pores of the template with the help of the metal catalysts. Well-aligned nanotube arrays have stimulated much interest because of their great potential in various applications, such as cold-cathode flat panel displays. Of particular interest is the use of zeolite templates with very narrow pores (< 1 nm diameter) that permit the growth of carbon nanotubes with diameters of 0.42 nm, having only 10 carbon atoms around the circumference [4.85].

The hollow cores of carbon nanotubes have also been used to synthesize a variety of nanowires of very small diameter [4.86]. These very small diameter nanowires have been extensively studied by high resolution TEM but have not yet been characterized to the same degree regarding their physical properties.

4.1.2 VLS Method for Nanowire Synthesis

Some of the recent successful syntheses of semiconductor nanowires are based on the so-called vapor-liquid-solid (VLS) mechanism of anisotropic crystal growth. This mechanism was first proposed for the growth of single crystal silicon whiskers 100 nm to hundreds of microns in diameter [4.87]. The proposed growth mechanism (see Fig. 4.5) involves the absorption of source material from the gas phase into a liquid droplet of catalyst (a molten particle of gold on a silicon substrate in the original work [4.87]). Upon supersaturation of the liquid alloy, a nucleation event generates a solid precipitate of the source material. This seed serves as a preferred site for further deposition of material at the interface of the liquid droplet, promoting the elongation of the seed into a nanowire or a whisker, and suppressing further nucleation events on the same catalyst. Since the liquid droplet catalyzes the incorporation of material from the gas source to the growing crystal, the deposit grows anisotropically as a whisker whose diameter is dictated by the diameter of the liquid alloy droplet. The nanowires thus obtained are of high purity, except for the end containing the solidified catalyst as an alloy particle (see Figs. 4.5, 4.6a). Real-time observations of the alloying, nucleation, and elongation steps in the growth of germanium nanowires from gold nanoclusters by the VLS method were recorded by in situ TEM [4.88].

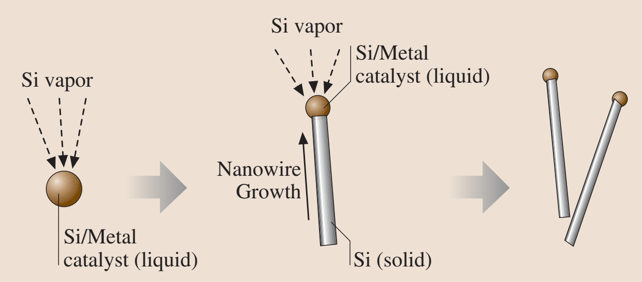

Fig. 4.5 Schematic diagram illustrating the growth of silicon nanowires by the VLS mechanism

Reduction of the average wire diameter to the nanometer scale requires the generation of nanosized catalyst droplets. Due to the balance between the liquid-vapor surface free energy and the free energy of condensation, however, the size of a liquid droplet, in equilibrium with its vapor, is usually limited to the micrometer range. This obstacle was overcome in recent years by several new methodologies: 1) Advances in the synthesis of metal nanoclusters have made monodispersed nanoparticles commercially available. These can be dispersed on a solid substrate in high dilution so that when the temperature is raised above the melting point, the liquid clusters do not aggregate [4.59]. 2) Alternatively, metal islands of nanoscale sizes can self-form when a strained thin layer is grown or heat treated on a non-epitaxial substrate [4.50]. 3) Laser-assisted catalytic VLS growth is a method used to generate nanowires under nonequilibrium conditions. By laser ablation of a target containing both the catalyst and the source materials, a plasma is generated from which catalyst nanoclusters nucleate as the plasma cools down. Single crystal nanowires grow as long as the particle remains liquid [4.60]. 4) Interestingly, by optimization of the material properties of the catalyst-nanowire system, conditions can be achieved for which nanocrystals nucleate in a liquid catalyst pool supersaturated with the nanowire material, migrate to the surface due to a large surface tension, and continue growing as nanowires perpendicular to the liquid surface [4.62]. In this case, supersaturated nanodroplets are sustained on the outer

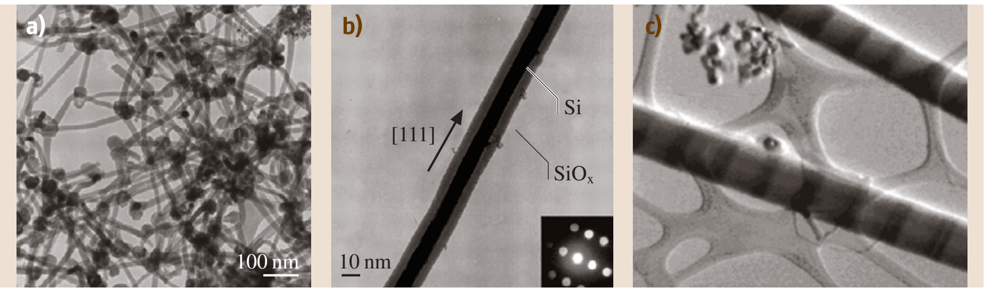

Fig. 4.6 **(a)** TEM images of Si nanowires produced after laser ablating a $Si_{0.9}Fe_{0.1}$ target. The dark spheres with a slightly larger diameter than the wires are solidified catalyst clusters [4.60]. **(b)** Diffraction contrast TEM image of a Si nanowire. The crystalline Si core appears darker than the amorphous oxide surface layer. The inset shows the convergent beam electron diffraction pattern recorded perpendicular to the wire axis, confirming the nanowire crystallinity [4.60]. **(c)** STEM image of $Si/Si_{1-x}Ge_x$ superlattice nanowires in the bright field mode. The scale bar is 500 nm [4.90]

end of the nanowire due to the low solubility of the nanowire material in the liquid [4.89].

A wide variety of elemental, binary, and compound semiconductor nanowires has been synthesized by the VLS method, and relatively good control over the nanowire diameter and diameter distribution has been achieved. Researchers are currently focusing attention on the controlled variation of the materials properties along the nanowire axis. To this context, researchers have modified the VLS synthesis apparatus to generate compositionally modulated nanowires. GaAs/GaP modulated nanowires have been synthesized by alternately ablating targets of the corresponding materials in the presence of gold nanoparticles [4.92]. p-Si/n-Si nanowires were grown by chemical vapor deposition from alternating gaseous mixtures containing the appropriate dopant [4.92]. $Si/Si_{1-x}Ge_x$ nanowires were grown by combining silicon from a gaseous source with germanium from a periodically ablated target (see Fig. 4.6c) [4.90]. Finally, using an ultra-high vacuum chamber and molecular beams, InAs/InP nanowires with atomically sharp interfaces were obtained [4.93]. These compositionally modulated nanowires are expected to exhibit exciting electronic, photonic, and thermoelectric properties.

Silicon and germanium nanowires grown by the VLS method consist of a crystalline core coated by a relatively thick amorphous oxide layer (2–3 nm) (see Fig. 4.6b). These layers are too thick to be the result of ambient oxidation, and it has been shown that these oxides play an important role in the nanowire growth process [4.61, 91]. Silicon oxides were found to serve as a special and highly selective catalyst, that significantly enhances the yield of Si nanowires, without the need for metal catalyst particles [4.61, 91, 94]. A similar yield enhancement was also found in the synthesis of Ge nanowires from the laser ablation of Ge powder mixed with GeO_2 [4.51]. The Si and Ge nanowires produced from these metal-free targets generally grow along the [112] crystal direction [4.95] and have the benefit that no catalyst clusters are found on either ends of the nanowires. Based on these observations

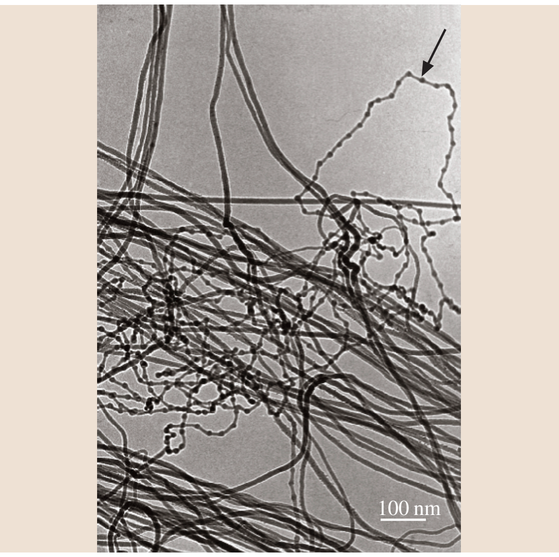

Fig. 4.7 TEM image showing the two major morphologies of Si nanowires prepared by the oxide assisted growth method [4.91]. Notice the absence of metal particles when compared to Fig. 4.6a. The *arrow* points at an oxide-linked chain of Si nanoparticles

and other TEM studies, an oxide-enhanced nanowire growth mechanism different from the classical VLS mechanism was proposed, in which no metal catalyst is required during the laser ablation-assisted synthesis [4.91]. It is postulated that the nanowire growth is dependent on the presence of SiO (or GeO) vapor. This decomposes in the nanowire tip region into Si (or Ge), which is incorporated into the crystalline phase, and SiO_2 (or GeO_2), which contributes to the outer coating. The initial nucleation events generate oxide-coated spherical nanocrystals. The [112] crystal faces have the fastest growth rate, and therefore the nanocrystals soon begin elongating along this direction to form one-dimensional structures. The Si_mO or Ge_mO ($m > 1$) layer on the nanowire tips may be at temperatures near their molten states, catalyzing the incorporation of gas molecules in a directional fashion [4.95]. Besides nanowires with smooth walls, a second morphology of chains of non-oriented nanocrystals linked by oxide necks is frequently observed (indicated by an arrow in Fig. 4.7). In addition, it was found by STM studies that about 1% of the wires consist of a regular array of two alternating segments, 10 nm and 5 nm in length, respectively [4.96]. The segments, whose junctions form an angle of 30°, are probably a result of alternating growth along different crystallographic orientations (see Sect. 4.2.1).

4.1.3 Other Synthesis Methods

In this section we review several other general procedures available for the synthesis of a variety of nanowires. We focus on "bottom-up" approaches, which afford many kinds of nanowires in large numbers and do not require highly sophisticated equipment (such as scanning microscopy or lithography based methods). We exclude cases for which the nanowires are not self sustained (such as in the case of atomic rows on the surface of crystals).

Gates et al. have demonstrated a solution-phase synthesis of nanowires with controllable diameters [4.57, 97], without the use of templates, catalysts, or surfactants. They make use of the anisotropy of the crystal structure of trigonal selenium and tellurium, which can be viewed as rows of 1-D helical atomic chains. They base their approach on the mass transfer of atoms during an aging step from a high free-energy solid phase (e.g., amorphous selenium) to a seed (e.g., trigonal selenium nanocrystal) that grows preferentially along one crystallographic axis. The lateral dimension of the seed, which dictates the diameter of the nanowire, can be controlled by the temperature of the nucleation step. Furthermore, Se/Te alloy nanowires were synthesized by this method, and Ag_2Se compound nanowires were obtained by treating selenium nanowires with $AgNO_3$ [4.98–100].

More often, however, the use of surfactants is necessary to promote the anisotropic 1-D growth of nanocrystals. Solution phase synthetic routes have been optimized to produce monodispersed quantum dots, i. e., zero-dimensional isotropic nanocrystals [4.101]. Surfactants are necessary in this case to stabilize the interfaces of the nanoparticles and retard oxidation and aggregation processes. Detailed studies on the effect of growth conditions revealed that they can be manipulated to induce a directional growth of the nanocrystals, usually generating nanorods (aspect ratio of ≈ 10), and in favorable cases, nanowires of high-aspect ratios. *Heath* and *LeGoues* [4.49] synthesized germanium nanowires by reducing a mixture of $GeCl_4$ and phenyl-$GeCl_3$ at high temperature and high pressure. The phenyl ligand was essential for the formation of high-aspect ratio nanowires [4.49]. In growing CdSe nanorods [4.38], Alivisatos et al. used a mixture of two surfactants, whose concentration ratio influenced the structure of the nanocrystal. It is believed that different surfactants have different affinities and different absorption rates for the different crystal faces of CdSe, thereby regulating the growth rate of these faces. A coordinating alkyl-diamine solvent was used to grow polycrystalline PbSe nanowires at low temperatures [4.55]. Here, the surfactant-induced directional growth is believed to occur, through the formation of organometallic complexes in which the bidentate ligand assumes the equatorial positions, thus hindering the ions from approaching each other in this plane. Additionally, the alkyl-amine molecules coat the external surface of the wire, preventing lateral growth. The aspect ratio of the wires increased as the temperature was lowered in the range $10 < T < 117\,°C$. Ethylenediamine was used to grow CdS nanowires and tetrapods by a solvo-thermal recrystallization process starting with CdS nanocrystals or amorphous particles [4.35]. While the coordinating solvent was crucial for the nanowire growth, the researchers did not clarify its role in the shape and phase control.

Stress-induced crystalline bismuth nanowires have been grown from sputtered films of layers of Bi and CrN. The nanowires presumably grow from defects and cleavage fractures in the film and are up to several millimeter in lengths with diameters ranging from 30 to 200 nm [4.27]. While the exploration of this technique has only begun, stress-induced unidirectional growth should be applicable to a variety of composite films.

Selective electrodeposition along the step edges in highly oriented pyrolytic graphite (HOPG) was used to obtain MoO_2 nanowires as shown in Fig. 4.8. The site-selectivity was achieved by applying a low overpotential to the electrochemical cell in which the HOPG served as cathode, thus minimizing the nucleation events on less favorable sites (i. e., plateaus). While these nanowires cannot be removed from the substrate, they can be reduced to metallic molybdenum nanowires, which can then be released as free-standing nanowires. Other metallic nanowires were also obtained by this method [4.53, 102]. In contrast to the template synthesis approaches described above, in this method the substrate defines only the position and orientation of the nanowire, not its diameter. In this context, other surface morphologies, such as self-assembled grooves in etched crystal planes, have been used to generate nanowire arrays via gas-phase shadow deposition (for example: Fe nanowires on (110)NaCl [4.44]). The cross-section of artificially prepared superlattice structures has also been used for site-selective electrodeposition of parallel and closely spaced nanowires [4.103]. Nanowires prepared on the above-mentioned substrates would have semicircular, rectangular, or other unconventional cross-sectional shapes.

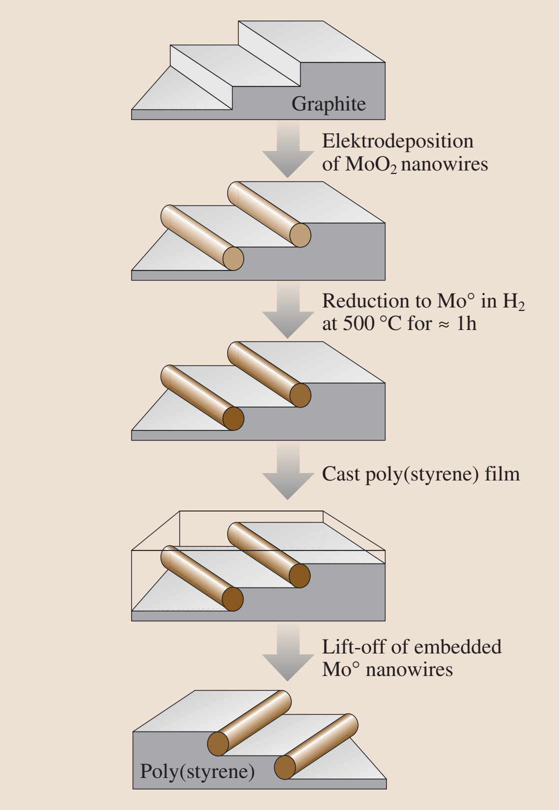
Graphite
Elektrodeposition of MoO_2 nanowires
Reduction to Mo° in H_2 at 500 °C for ≈ 1h
Cast poly(styrene) film
Lift-off of embedded Mo° nanowires
Poly(styrene)

Fig. 4.8 Schematic of the electrodeposition step edge decoration of HOPG (highly oriented pyrolytic graphite) for the synthesis of molybdenum nanowires [4.53, 102]

4.1.4 Hierarchical Arrangement and Superstructures of Nanowires

Ordering nanowires into useful structures is another challenge to address in order to harness the full potential of nanowires for applications. We will first review examples of nanowires having a nontrivial structure and then proceed to describe methods to create assemblies of nanowires of a predetermined structure.

We have already mentioned in Sect. 4.1.2 that the preparation of nanowires with a graded composition or with a superlattice structure along their main axis was demonstrated by controlling the gas-phase chemistry as a function of time during the growth of the nanowires by the VLS method. Control of the composition along the axial dimension was also demonstrated by a template-assisted method, for example by the consecutive electrochemical deposition of different metals in the pores of an alumina template [4.104]. Alternatively, the composition can be varied along the radial dimension of the nanowire, for example, by first growing a nanowire by the VLS method and then switching the synthesis conditions to grow a different material on the surface of the nanowire by CVD. This technique was demonstrated for the synthesis of Si/Ge and Ge/Si coaxial (or core-shell) nanowires [4.105], and it was shown that by a thermal annealing process the outer shell can be formed epitaxially on the inner core. A different approach was adopted by *Wang* et al. who generated a mixture of coaxial and biaxial SiC-SiO_x nanowires by the catalyst-free high-temperature reaction of amorphous silica and a carbon/graphite mixture [4.106].

A different category of nontrivial nanowires is that of nanowires having a nonlinear structure, resulting from multiple one-dimensional growth steps. Members of this category are tetrapods, which were mentioned in the context of liquid-phase synthesis (Sect. 4.1.3). In this process, a tetrahedral quantum-dot core is first grown, and then the conditions are modified to induce a one-dimensional growth of a nanowire from each one of the facets of the tetrahedron. A similar process

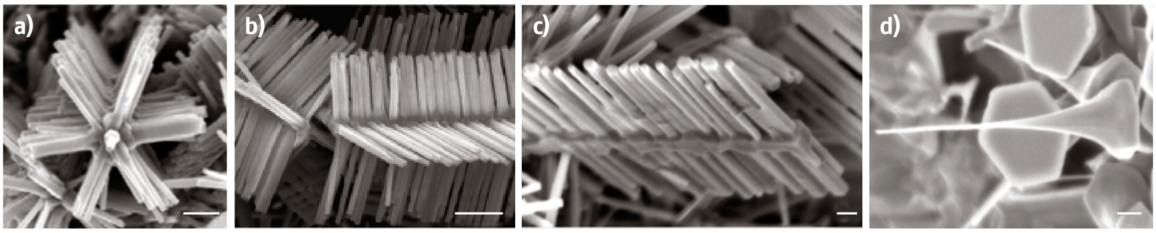

Fig. 4.9a–d SEM images of **(a)** 6-fold **(b)** 4-fold and **(c)** 2-fold symmetry nanobrushes made of an In_2O_3 core and ZnO nanowire brushes [4.107], and of **(d)** ZnO nanonails [4.108]

produced high-symmetry In_2O_3/ZnO hierarchical nanostructures. From a mixture of heat-treated In_2O_3, ZnO and graphite powders, faceted In_2O_3 nanowires were first obtained, on which oriented shorter ZnO nanowires crystallized [4.107]. Brush-like structures were obtained as a mixture of 11 structures of different symmetries. For example, two, four, or six rows of ZnO nanorods could be found on different core nanowires, depending on the crystallographic orientation of the main axis of the core nanowire, as shown in Fig. 4.9. Comb-like structures entirely made of ZnO were also reported [4.65].

Control of the position of a nanowire in the growth process is important for preparing devices or test structures containing nanowires, especially when involving a large array of nanowires. Post-synthesis methods to align and position nanowires include microfluidic channels [4.110], Langmuir–Blodgett assemblies [4.109], and electric-field assisted assembly [4.111]. The first method involves the orientation of the nanowires by the liquid flow direction when a nanowire solution is injected into a microfluidic channel assembly and by the interaction of the nanowires with the side walls of the channel. The second method involves the alignment of nanowires at a liquid-gas or liquid-liquid interface by the application of compressive forces on the interface (Fig. 4.10). The third technique is based on dielectrophoretic forces that pull polarizable nanowires toward regions of high field strength. The nanowires align between two isolated electrodes that are capacitatively coupled to a pair of buried electrodes biased with an AC voltage. Once a nanowire shorts the electrodes, the electric field is eliminated, preventing more nanowires from depositing. The above techniques have been successfully used to prepare electronic circuitry and optical devices out of nanowires (see Sects. 4.3.1 and 4.3.3). Alternatively, alignment and positioning of the nanowires can be specified and controlled during their growth by the proper design of the synthesis. For example, ZnO nanowires prepared by the VLS method were grown into an array in which both their position on the substrate and their growth direction and orientation were controlled [4.65]. The nanowire growth region was defined by patterning the gold film, which serves as a catalyst for the ZnO nanowire growth, employing soft-lithography, e-beam lithography, or photolithography. The orientation of the nanowires was achieved by selecting a substrate with a lattice structure matching that of the nanowire material to facilitate the epitaxial growth. These conditions result in an array of nanowire posts at predetermined positions, all vertically aligned with the same crystal growth orientation (see Fig. 4.11). A similar structure could be obtained by the template-mediated electrochemical synthesis of nanowires (see Sect. 4.1.1), particularly if anodic alumina with its parallel and ordered channels is used. The control over the

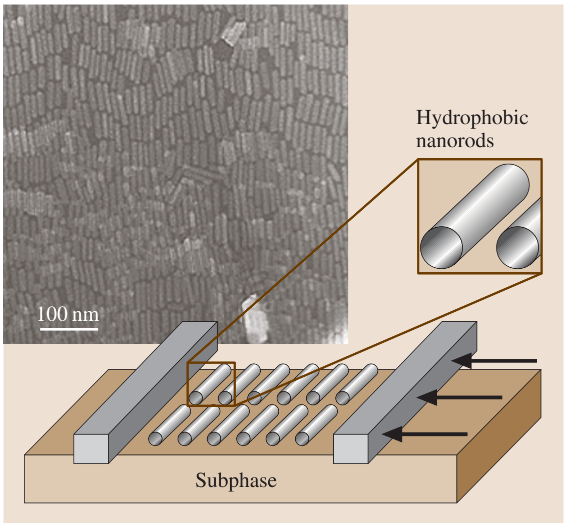

Fig. 4.10 A TEM image of a smectic phase of a $BaCrO_4$ nanorod film (*left inset*) achieved by the Langmuir–Blodgett technique, as depicted by the illustration [4.109]

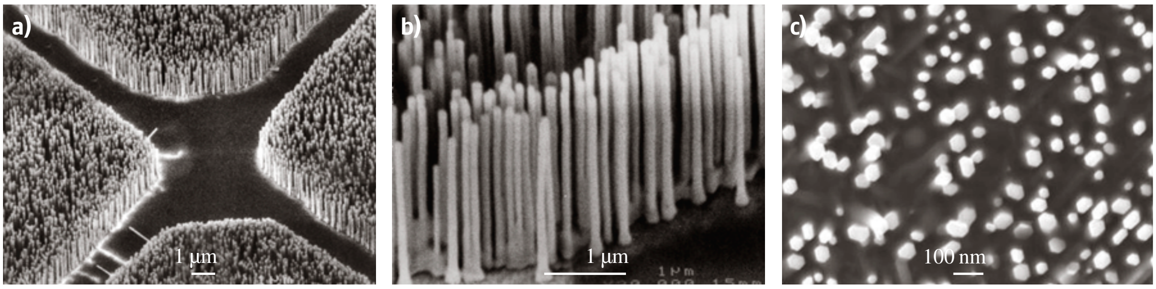

Fig. 4.11a–c SEM images of ZnO nanowire arrays grown on a sapphire substrate, where **(a)** shows patterned growth, **(b)** shows a higher resolution image of the parallel alignment of the nanowires, and **(c)** shows the faceted side-walls and the hexagonal cross section of the nanowires. For nanowire growth, the sapphire substrates were coated with a 1.0 to 3.5 nm thick patterned layer of Au as the catalyst, using a TEM grid as the shadow mask. These nanowires have been used for nanowire laser applications [4.115]

location of the nucleation of nanowires in the electrochemical deposition is determined by the pore positions and the back-electrode geometry. The pore positions can be precisely controlled by imprint lithography [4.112]. By growing the template on a patterned conductive substrate that serves as a back-electrode [4.113,114], different materials can be deposited in the pores at different regions of the template.

4.2 Characterization and Physical Properties of Nanowires

In this section we review the structure and properties of nanowires, and the interrelationship between the two. The discovery and investigation of nanostructures were galvanized by advances in various characterization and microscopy techniques that enable materials characterization to take place at smaller and smaller length scales, reaching down to individual atoms. For applications, characterization of the nanowire structural properties is especially important so that a reproducible relationship between their desired functionality and their geometrical and structural characteristics can be established. Due to the enhanced surface-to-volume ratio in nanowires, their properties may depend sensitively on their surface condition and geometrical configuration. Even nanowires made of the same material may possess dissimilar properties due to differences in their crystal phase, crystalline size, surface conditions, and aspect ratios, which depend on the synthesis methods and conditions used in their preparation.

4.2.1 Structural Characterization

Structural and geometric factors play an important role in determining the various attributes of nanowires, such as their electrical, optical and magnetic properties. Various novel tools, therefore, have been developed and employed to obtain this important structural information at the nanoscale. At the micron scale, optical techniques are extensively used for imaging structural features. Since the sizes of nanowires are usually comparable to or, in most cases, much smaller than the wavelength of visible light, traditional optical microscopy techniques are usually limited in characterizing the morphology and surface features of nanowires. Electron microscopy techniques, therefore, play a more dominant role at the nanoscale. Since electrons interact more strongly than photons, electron microscopy is particularly sensitive relative to X-rays for the analysis of tiny samples.

In this section we review and give examples of how scanning electron microscopy, transmission electron microscopy, scanning probe spectroscopies, and diffraction techniques are used to characterize the structure of nanowires. To provide the necessary basis for developing reliable structure-property relations, we apply multiple characterization tools to the same samples.

Scanning Electron Microscopy (SEM)

SEM usually produces images down to length scales of ~ 10 nm and provides valuable information regarding the structural arrangement, spatial distribution, wire density, and geometrical features of the nanowires. Examples of SEM micrographs shown in Figs. 4.1 and

4.3 indicate that structural features at the 10 nm to 10 μm length scales can be probed, providing information on the size, size distribution, shapes, spatial distributions, density, nanowire alignment, filling factors, granularity etc. As another example, Fig. 4.11a shows an SEM image of ZnO nanowire arrays grown on a sapphire substrate [4.115], which provides evidence for the nonuniform spatial distribution of the nanowires on the substrate. This distribution was attained by patterning the catalyst film to define high density growth regions and nanowire-free regions. Figure 4.11b, showing a higher magnification of the same system, indicates that these ZnO nanowires grow perpendicular to the substrate, are well aligned with approximately equal wire lengths, and have wire diameters in the range of $20 < d_W < 150$ nm. The SEM micrograph in Fig. 4.11c provides further information about the surface of the nanowires, showing it to be well-faceted and forming a hexagonal cross-section, indicative of nanowire growth along the ⟨0001⟩ direction. The uniformity of the nanowire size, its alignment perpendicular to the substrate, and its uniform growth direction, as suggested by the SEM data, are linked to the good epitaxial interface between the (0001) plane of the ZnO nanowire and the (110) plane of the sapphire substrate. (The crystal structures of ZnO and sapphire are essentially incommensurate, with the exception that the *a* axis of ZnO and the *c* axis of sapphire are related almost exactly by a factor of 4, with a mismatch of less than 0.08% at room temperature [4.115].) The well-faceted nature of these nanowires has important implications for their lasing action (see Sect. 4.3.2). Figure 4.12 shows an SEM image of GaN nanowires synthesized by a laser-assisted catalytic growth method [4.46], indicating a random spatial orientation of the nanowire axes and a wide diameter distribution for these nanowires, in contrast to the ZnO wires in Fig. 4.11 and to arrays of well-aligned nanowires prepared by template-assisted growth (see Fig. 4.3).

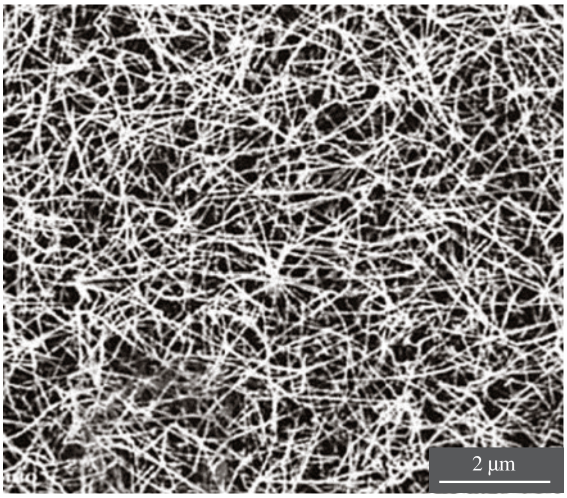

Fig. 4.12 SEM image of GaN nanowires in a mat arrangement synthesized by laser-assisted catalytic growth. The nanowires have diameters and lengths on the order of 10 nm and 10 μm, respectively [4.46]

Transmission Electron Microscopy (TEM)

TEM and high resolution transmission electron microscopy (HRTEM) are powerful imaging tools to study nanowires at the atomic scale, and they usually provide more detailed geometrical features than are seen in SEM images. TEM studies also yield information regarding the crystal structure, crystal quality, grain size, and crystal orientation of the nanowire axis. When operating in the diffraction mode, selected area electron diffraction (SAED) patterns can be made to determine the crystal structure of nanowires. As an example, the TEM images in Fig. 4.13 show four different morphologies for Si nanowires prepared by the laser ablation of a Si target [4.116]: (a) spring-shaped; (b) fishbone-shaped (indicated by solid arrow) and frog-egg-shaped (indicated by the hollow arrow), (c) pearl-shaped, while (d) shows the poly-sites of nanowire nucleation. The crystal quality of nanowires is revealed from high resolution TEM images with atomic resolution, along with selected area electron diffraction (SAED) patterns. For example, Fig. 4.14 shows a TEM image of one of the GaN nanowires from Fig. 4.12, indicating single crystallinity and showing (100) lattice planes, thus demonstrating the growth direction of the nanowire. This information is supplemented by the corresponding electron diffraction pattern in the upper right.

The high resolution of the TEM also allows for the investigation of the surface structure of the nanowires. In many cases the nanowires are sheathed with a native oxide layer or an amorphous oxide layer that forms during the growth process. This can be seen in Fig. 4.6b for silicon nanowires and in Fig. 4.15 for germanium nanowires [4.51], showing a mass-thickness contrast TEM image and a selected-area electron diffraction pattern of a Ge nanowire. The main TEM image shows that these Ge nanowires possess an amorphous GeO_2 sheath with a crystalline Ge core oriented in the [211] direction.

Dynamical processes of the surface layer of nanowires can be studied by using an in situ environ-

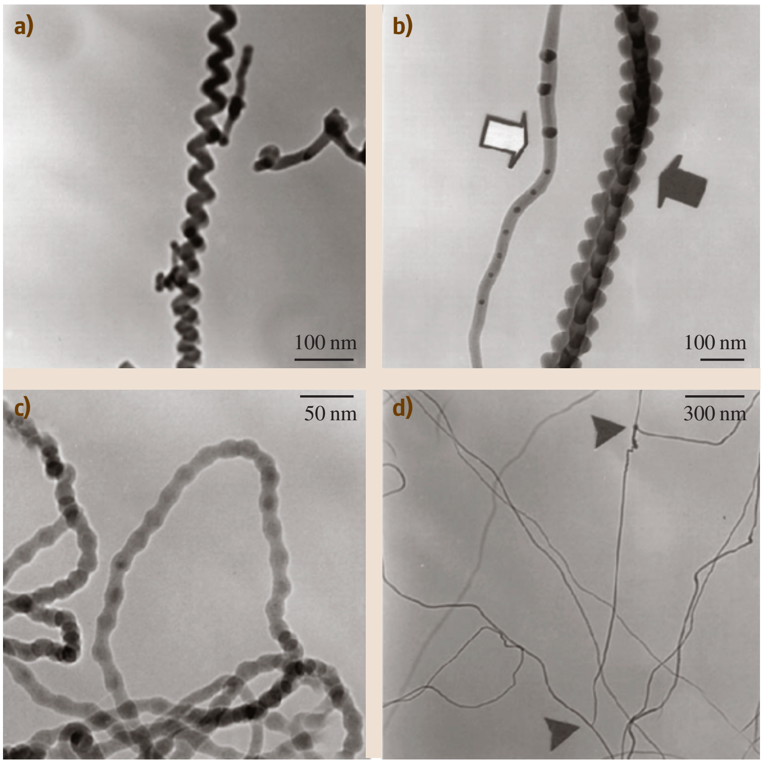

Fig. 4.13a–d TEM morphologies of four special forms of Si nanowires synthesized by the laser ablation of a Si powder target. **(a)** A spring-shaped Si nanowire; **(b)** fishbone-shaped (indicated by a *solid arrow*) and frog-egg-shaped (indicated by a *hollow arrow*) Si nanowires; and **(c)** pearl-shaped nanowires, while **(d)** shows poly-sites for the nucleation of silicon nanowires (indicated by *arrows*) [4.116]

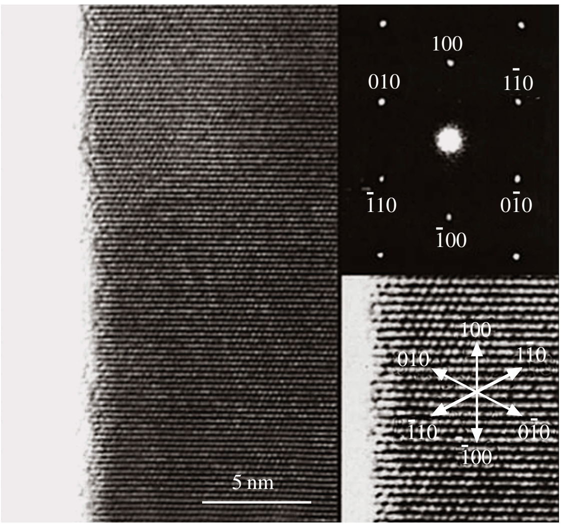

Fig. 4.14 Lattice resolved high resolution TEM image of one GaN nanowire (*left*) showing that (100) lattice planes are visible perpendicular to the wire axis. The electron diffraction pattern (*top right*) was recorded along the ⟨001⟩ zone axis. A lattice-resolved TEM image (*lower right*) highlights the continuity of the lattice up to the nanowire edge, where a thin native oxide layer is found. The directions of various crystallographic planes are indicated in the *lower right* figure [4.46]

mental TEM chamber, which allows TEM observations to be made while different gases are introduced or as the sample is heat treated at various temperatures, as illustrated in Fig. 4.16. The figure shows high resolution TEM images of a Bi nanowire with an oxide coating and the effect of a dynamic oxide removal process carried out within the environmental chamber of the TEM [4.117]. The amorphous bismuth-oxide layer

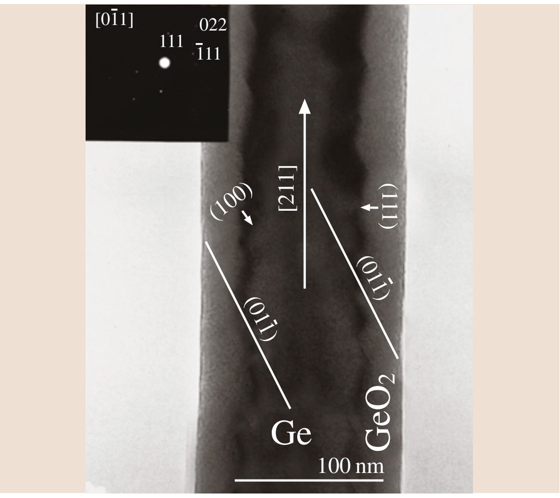

Fig. 4.15 A mass-thickness contrast TEM image of a Ge nanowire taken along the [0$\bar{1}$1] zone axis and a selected-area electron diffraction pattern (*upper left inset*) [4.51]. The Ge nanowires were synthesized by laser ablation of a mixture of Ge and GeO_2 powder. The core of the Ge nanowire is crystalline, while the surface GeO_2 is amorphous

coating the nanowire (Fig. 4.16a) is removed by exposure to hydrogen gas within the environmental chamber of the TEM, as indicated in Fig. 4.16b.

By coupling the powerful imaging capabilities of TEM with other characterization tools, such as an electron energy loss spectrometer (EELS) or an energy dispersive X-ray spectrometer (EDS) within the TEM instrument, additional properties of the nanowires can be probed with high spatial resolution. With the EELS technique, the energy and momentum of the incident and scattered electrons are measured in an inelastic electron scattering process to provide information on the energy and momentum of the excitations in the nanowire sample. Fig. 4.17 shows the dependence on nanowire diameter of the electron energy loss spectra of Bi nanowires. The spectra were taken from the center of the nanowire, and the shift in the energy of the peak position (Fig. 4.17) indicates the effect of the nanowire diameter on the plasmon frequency in the nanowires. The results show changes in the electronic structure of Bi nanowires as the wire diameter decreases [4.118]. Such changes in electronic structure as a function of nanowire diameter are also observed in their transport (Sect. 4.2.2) and optical (Sect. 4.2.3) properties and are related to quantum confinement effects.

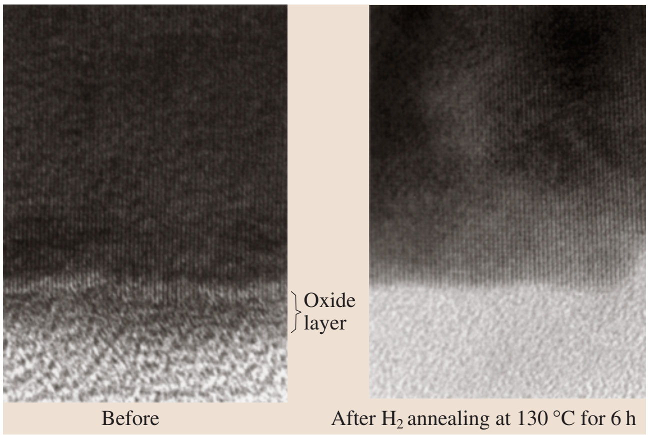

Fig. 4.16 High resolution transmission electron microscope (HRTEM) image of a Bi nanowire (*left*) before and (*right*) after annealing in hydrogen gas at 130 °C for 6 hours within the environmental chamber of the HRTEM instrument to remove the oxide surface layer [4.117]

EDS measures the energy and intensity distribution of X-rays generated by the impact of the electron beam on the surface of the sample. The elemental composition within the probed area can be determined to a high degree of precision. The technique was particularly useful for the compositional characterization of superlattice

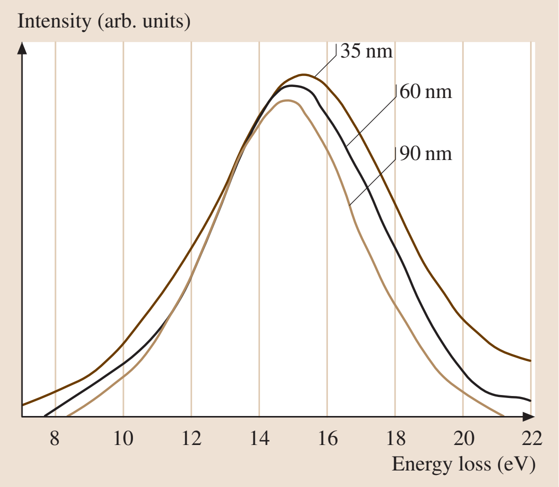

Fig. 4.17 Electron energy loss spectra (EELS) taken from the center of bismuth nanowires with diameters of 35, 60, and 90 nm. The shift in the volume plasmon peaks is due to the wire diameter effects on the electronic structure [4.118]

nanowires [4.90] and core-sheath nanowires [4.105] (see Sect. 4.1.2).

Scanning Tunneling Probes

Several scanning probe techniques [4.119], such as scanning tunneling microscopy (STM), electric field gradient microscopy (EFM) [4.16], magnetic field microscopy (MFM) [4.54], and scanning thermal microscopy (SThM) [4.120], combined with atomic force microscopy (AFM), have been employed to study the structural, electronic, magnetic, and thermal properties of nanowires. A scanning tunneling microscope can be employed to reveal both topographical structural information, such as that illustrated in Fig. 4.18, as well as information on the local electronic density of states of a nanowire, when used in the STS (scanning tunneling spectroscopy) mode. Figure 4.18 shows STM height images (taken in the constant current STM mode) of MoSe molecular wires deposited from a methanol or acetonitrile solution of $Li_2Mo_6Se_6$ on to Au substrates. The STM image of a single MoSe wire (Fig. 4.18a) exhibits a 0.45 nm lattice repeat distance in a MoSe molecular wire. When both STM and STS measurements are made on the same sample, the electronic and structural properties can be correlated, for example, as in the joint STM/STS studies on Si nanowires [4.96], showing alternating segments of a single nanowire identified with growth along [110] and [112] directions, and different I–V characteristics measured for the [110] segments as compared with the [112] segments.

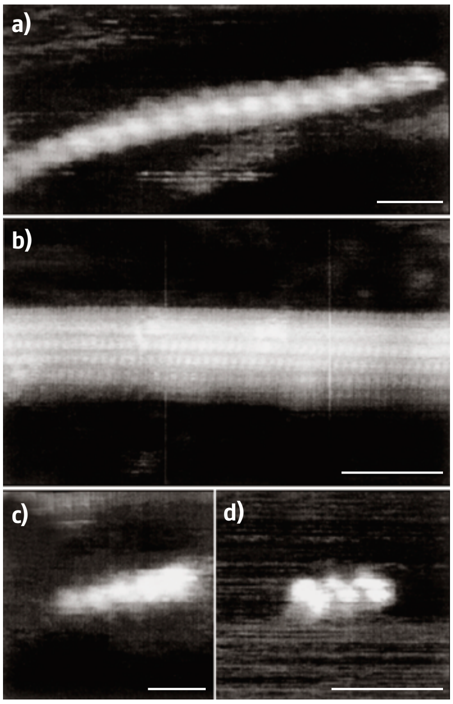

Fig. 4.18a–d STM height images, obtained in the constant current mode, of MoSe chains deposited on an Au(111) substrate. **(a)** A single chain image, and **(b)** a MoSe wire bundle. **(c)** and **(d)** Images of MoSe wire fragments containing 5 and 3 unit cells, respectively [4.119]. The scale bars are all 1 nm

Magnetic field microscopy (MFM) has been employed to study magnetic polarization of magnetic nanowires embedded in an insulating template, such as an anodic alumina template. For example, Fig. 4.19a shows the topographic image of an anodic alumina template filled with Ni nanowires, and Fig. 4.19b demonstrates the corresponding magnetic polarization of each nanowire in the template. This micrograph shows that a magnetic field microscopy probe can distinguish between spin-up and spin-down nanowires in the nanowire

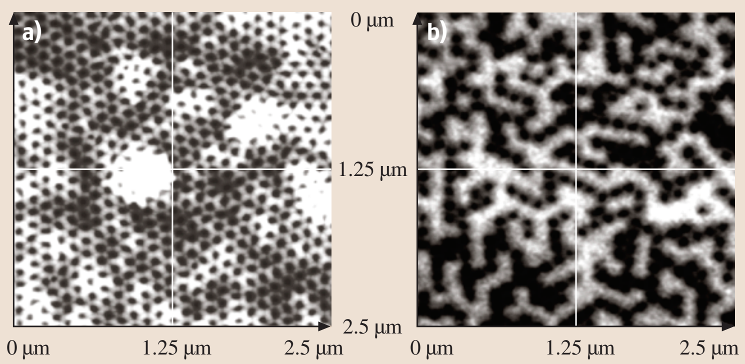

Fig. 4.19 **(a)** Topographic image of a highly ordered porous alumina template with a period of 100 nm filled with 35 nm diameter nickel nanowires. **(b)** The corresponding MFM (magnetic force microscope) image of the nano-magnet array, showing that the pillars are magnetized alternately "up" (*white*) and "down" (*black*) [4.54]

array, thereby providing a method for measuring interwire magnetic dipolar interactions [4.54].

X-Ray Analysis

Other characterization techniques commonly used to study the crystal structure and chemical composition of nanowires include X-ray diffraction and X-ray energy dispersion analysis (EDAX). The peak positions in the X-ray diffraction pattern can be used to determine the chemical composition and the crystal phase structure of the nanowires. For example, Fig. 4.2 shows that Bi nanowires have the same crystal structure and lattice constants as bulk bismuth. Both the X-ray diffraction pattern (XRD) for an array of aligned Bi nanowires (Fig. 4.2) and the SAED pattern for individual Bi nanowires [4.16] suggest that the nanowires have a common axis of crystal orientation.

As another example of an XRD pattern for an array of aligned nanowires, Fig. 4.20 shows the X-ray diffraction pattern of the ZnO nanowires displayed in Fig. 4.11. Only (00ℓ) diffraction peaks are observed for these aligned ZnO nanowires, indicating that their preferred growth direction is (001) along the wire axis. A similar preferred growth orientation was also observed for Bi nanowires produced by high pressure injection (see Sect. 4.1.1), while the vapor phase and low pressure injection filling techniques give preferred alignment along lower symmetry axes (see Fig. 4.2).

EDAX has been used to determine the chemical composition, stoichiometry of compound nanowires, or the impurity content in the nanowires. But the results from EDAX analysis should be interpreted carefully to avoid systematic errors.

4.2.2 Transport Properties

The study of nanowire electrical transport properties is important for nanowire characterization, electronic device applications, and the investigation of unusual transport phenomena arising from one-dimensional quantum effects. Important factors that determine the transport properties of nanowires include the wire diameter (important for both classical and quantum size effects), material composition, surface conditions, crystal quality, and the crystallographic orientation along the wire axis, which is important for materials with anisotropic materials parameters, such as the effective mass tensor, the Fermi surface, or the carrier mobility.

Electronic transport phenomena in low-dimensional systems can be roughly divided into two categories: ballistic and diffusive transport. Ballistic transport phenomena occur when electrons travel across the nanowire without any scattering. In this case, the conduction is mainly determined by the contacts between the nanowire and the external circuit, and the conductance is quantized into an integral number of universal conductance units $G_0 = 2e^2/h$ [4.121, 122]. Ballistic transport phenomena are usually observed in very short quantum wires, such as those produced by using mechanically controlled break junctions (MCBJ) [4.123, 124] where the electron mean free path is much longer than the wire length, and the conduction is a pure quantum phenomenon. To observe ballistic transport, the thermal energy must also obey the relation $k_\mathrm{B}T \ll \varepsilon_j - \varepsilon_{j-1}$, where $\varepsilon_j - \varepsilon_{j-1}$ is the energy separation between subband levels j and $j-1$. On the other hand, for nanowires with lengths much larger than the carrier mean free path, the electrons (or holes) undergo numerous scattering events when they travel along the wire. In this case, the transport is in the diffusive regime, and the conduction is dominated by carrier scattering within the wires due to phonons (lattice vibrations), boundary scattering, lattice and other structural defects, and impurity atoms.

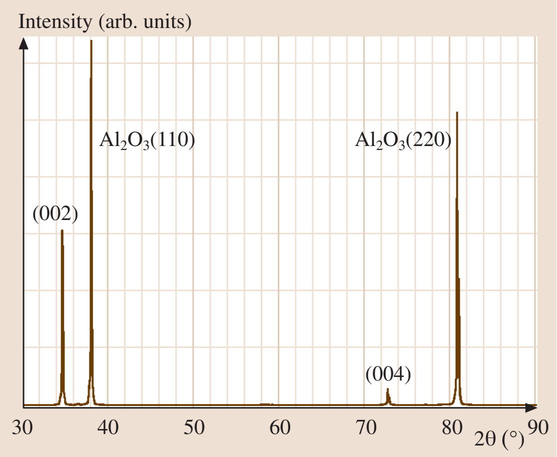

Fig. 4.20 X-ray diffraction pattern of aligned ZnO nanowires (see Fig. 4.11) grown on a sapphire substrate. Only $[00\ell]$ diffraction peaks are observed for the nanowires, owing to their well-oriented growth orientation. Strong diffraction peaks for the sapphire substrate are also found [4.115]

Conductance Quantization in Metallic Nanowires

The ballistic transport of 1-D systems has been extensively studied since the discovery of quantized conductance in 1-D systems in 1988 [4.121, 122]. The phenomena of conductance quantization occur when the diameter of the nanowire is comparable to the electron Fermi wavelength, which is on the order of 0.5 nm for most metals [4.125]. Most conductance quantization experiments up to the present were performed by joining and separating two metal electrodes. As the two metal electrodes are slowly separated, a nano-contact is formed before it breaks completely (see Fig. 4.21a), and conductance in integral multiple values of G_0 is observed through these nano-contacts. Fig. 4.21b shows the conductance histogram built with 18,000 contact breakage curves between two gold electrodes at room temperature [4.126], with the electrode separation up to ~ 1.8 nm. The conductance quantization behavior is found to be independent of the contact material and has been observed in various metals, such as Au [4.126], Ag, Na, Cu [4.127], and Hg [4.128]. For semimetals such as Bi, conductance quantization has also been observed for an electrode separation of as long as 100 nm at 4 K because of the long Fermi wavelength (~ 26 nm) [4.125], indicating that the conductance quantization may be due to the existence of well-defined quantum states localized at a constriction instead of resulting from the atom rearrangement as the electrodes separate. Since the conductance quantization is observed only in breaking contacts, or for very narrow and very short nanowires, most nanowires of practical interest (possessing lengths of several microns) lie in the diffusive transport regime, where the carrier scattering is significant and should be considered.

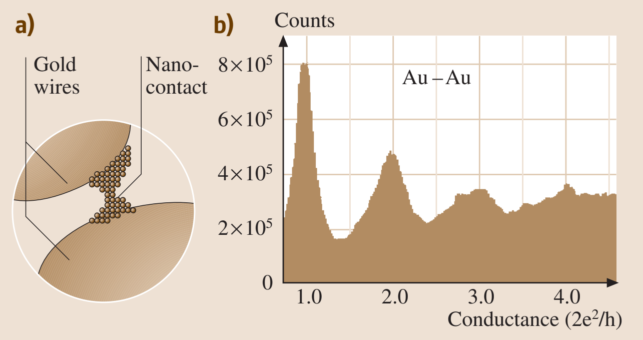

Fig. 4.21 **(a)** Schematic representation of the last stages of the contact breakage process [4.126]. **(b)** Histogram of conductance values built with 18,000 gold contact breakage experiments in air at room temperature, showing conductance peaks at integral values of G_0. In this experiment the gold electrodes approach and separate at 89,000 Å/s [4.126]

I–V Characterization of Semiconducting Nanowires

The electronic transport behavior of nanowires may be categorized based on the relative magnitude of three length scales: carrier mean free path ℓ_W, the de Broglie wavelength of electrons λ_e, and the wire diameter d_W. For wire diameters much larger than the carrier mean free path ($d_W \gg \ell_W$), the nanowires exhibit transport properties similar to bulk materials, which are independent of the wire diameter, since the scattering due to the wire boundary is negligible compared to other scattering mechanisms. For wire diameters comparable or smaller than the carrier mean free path ($d_W \sim \ell_W$ or $d_W < \ell_W$) but still much larger than the de Broglie wavelength of the electrons ($d_W \gg \lambda_e$), the transport in nanowires is in the classical finite size regime, in which the band structure of the nanowire is still similar to that of bulk while the scattering events at the wire boundary alter their transport behavior. For wire diameters comparable to the electronic wavelength $d_W \sim \lambda_e$, the electronic density of states is altered dramatically and quantum subbands are formed due to the quantum confinement effect at the wire boundary. In this regime, the transport properties are further influenced by the change in the band structure. Transport properties for nanowires in the classical finite size and quantum size regimes, therefore, are highly diameter-dependent.

Researchers have investigated the transport properties of various semiconducting nanowires and have demonstrated their potential for diverse electronic devices, such as for p-n diodes [4.129, 130], field effect transistors [4.129], memory cells, and switches [4.131] (see Sect. 4.3.1). The nanowires studied so far in this context have usually been made from conventional semiconducting materials, such as group IV and III-V compound semiconductors via the VLS growth method (see Sect. 4.1.2), and their nanowire properties have been compared to their well-established bulk properties. Interestingly, the physical principles for describing bulk semiconductor devices also hold for devices based on these semiconducting nanowires with wire diameters of tens of nanometers. For example, Fig. 4.22 shows the current–voltage (I–V) behavior of a 4-by-1 crossed p-Si/n-GaN junction array at room temperature [4.129]. The long horizontal wire in the figure is a p-Si nanowire (10–25 nm in diameter) and the four short vertical wires are n-GaN nanowires (10–30 nm in diameter). Each of the four nanoscale cross points independently forms a p-n junction with current rectification behavior, as shown by the I–V curves in Fig. 4.22, and the junction behavior (e.g., the turn-on voltage) can

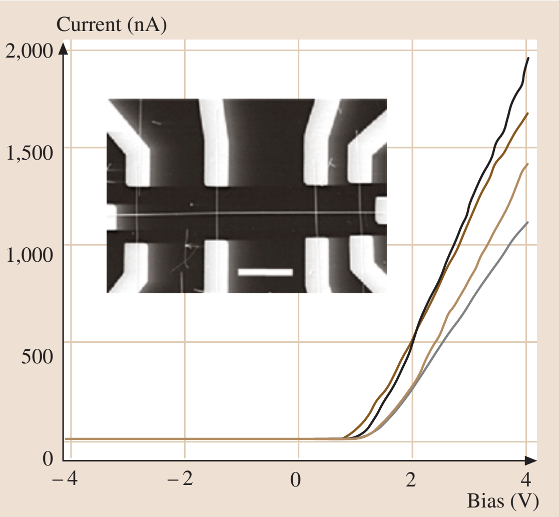

Fig. 4.22 I–V behavior for a 4(p) by 1(n) crossed p-Si/n-GaN junction array shown schematically in the inset. The four curves represent the I–V response for each of the four junctions, showing similar current rectifying characteristics in each case. The length scale bar between the two middle junctions is 2 μm [4.129]. The p-Si and n-GaN nanowires are 10–25 nm and 10–30 nm in diameter, respectively

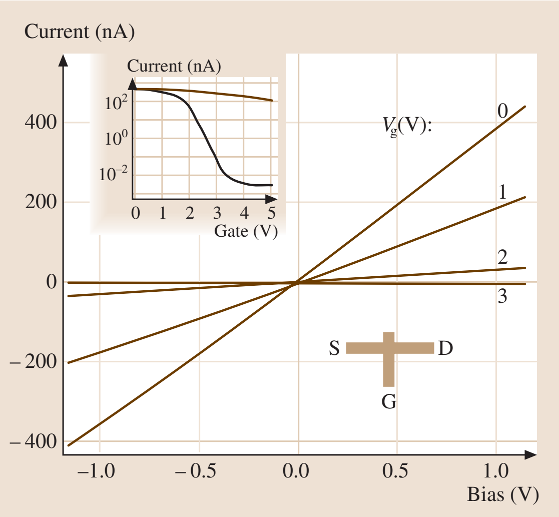

Fig. 4.23 Gate-dependent I–V characteristics of a crossed nanowire field-effect transistor (FET). The n-GaN nanowire is used as the nano-gate, with the gate voltage indicated (0, 1, 2, and 3 V). The inset shows the current vs. V_gate for a nanowire gate (*lower curve*) and for a global back-gate (*top curve*) when the bias voltage is set at 1 V [4.129]

be controlled by varying the oxide coating on these nanowires [4.129].

Huang et al. have demonstrated nanowire junction diodes with a high turn-on voltage (∼ 5 V) by increasing the oxide thickness at the junctions. The high turn-on voltage enables the use of the junction in a nanoscale FET as shown in Fig. 4.23 [4.129], where I–V data for a p-Si nanowire are presented, for which the n-GaN nanowire with a thick oxide coating is used as a nano-gate. By varying the nano-gate voltage, the conductance of the p-Si nanowire can be changed by more than a factor of 10^5 (lower curve in the inset), whereas the conductance changes by only a factor of ten when a global back-gate is used (top curve in the inset of Fig. 4.23). This behavior may be due to the thin gate dielectric between the crossed nanowires and the better control of the local carrier density through a nano-gate. Based on the gate-dependent I–V data of these p-Si nanowires, it is found that the mobility for holes in the p-Si nanowires may be higher than that for bulk p-Si, although further investigation is required for a complete understanding.

Because of the enhanced surface-to-volume ratio of nanowires, their transport behavior may be modified by changing their surface conditions. For example, researchers have found that by coating n-InP nanowires with a layer of redox molecules such as cobalt phthalocyanine, the conductance of the InP nanowires may change by orders of magnitude by altering the charge state of the redox molecules to provide bistable nanoscale switches [4.131]. The resistance (or conductance) of some nanowires (e.g., Pd nanowires) is also very sensitive to the presence of certain gases (e.g., H_2) [4.132, 133], and this property may be utilized for sensor applications to provide improved sensitivity compared to conventional sensors based on bulk material (see Sect. 4.3.4).

Although it remains unclear how the size may influence the transport properties and device performance of semiconducting nanowires, many of the larger diameter semiconducting nanowires are expected to be described by classical physics, since their quantization energies $\hbar^2/(2m_\mathrm{e}d_\mathrm{W}^2)$ are usually smaller than the thermal energy $k_\mathrm{B}T$. By comparing the quantization energy with the thermal energy, the critical wire diameter below which quantum confinement effects become significant is estimated to be 1 nm for Si nanowires at room temperature, which is much smaller than the size of many of the semiconducting nanowires that have been investigated so far. By using material systems with much smaller effective carrier masses m_e (such as bismuth), the critical diameter for which such quantum effects can be observed is increased, thereby facilitating the study of quantum confinement effects. It is for this reason that the bismuth nanowire system has been studied so extensively. Furthermore, since the crystal structure and lattice constants of bismuth nanowires are the same as for 3-D crystalline bismuth, it is possible to carry out detailed model calculations to guide and to interpret transport and optical experiments on bismuth nanowires. For these reasons, bismuth can be considered as a model system for studying 1-D effects in nanowires.

Temperature-Dependent Resistance Measurements

Although nanowires with electronic properties similar to their bulk counterparts are promising for constructing nano-devices based on well-established knowledge of their bulk counterparts, it is expected that quantum size effects in nanowires will likely be utilized to generate new phenomena absent in bulk materials and thus provide enhanced performance and novel functionality for certain applications. In this context, the transport properties of bismuth (Bi) nanowires have been extensively studied, both theoretically [4.134] and experimentally [4.28, 30, 67, 135–138] because of their promise for enhanced thermoelectric performance. Transport studies of ferromagnetic nanowire arrays, such as Ni or Fe, have also received much attention because of their potential for high-density magnetic storage applications.

The very small electron effective mass components and the long carrier mean free paths in Bi facilitate the study of quantum size effects in the transport properties of nanowires. Quantum size effects are expected to become significant in bismuth nanowires with diameters smaller than 50 nm [4.134], and the fabrication of crystalline nanowires of this diameter range is relatively easy.

Figure 4.24a shows the T dependence of the resistance $R(T)$ for Bi nanowires ($7 \leq d_W < 200$ nm) synthesized by vapor deposition and pressure injection [4.28], illustrating the quantum effects in their temperature-dependent resistance. In Fig. 4.24a, the $R(T)$ behavior of Bi nanowires is dramatically different from that of bulk Bi and is highly sensitive to the wire diameter. The $R(T)$ curves in Fig. 4.24a show a non-monotonic trend for larger-diameter (70 and 200 nm) nanowires, although $R(T)$ becomes monotonic with T for small-diameter (≤ 48 nm) nanowires. This dramatic change in the behavior of $R(T)$ as a function of d_W is attributed to a unique semimetal-semiconductor transition phenomena in Bi [4.139], induced by quantum size effects. Bi is a semimetal in bulk form, in which the T-point valence band overlaps with the L-point conduction band by 38 meV at 77 K. As the wire diameter decreases, the lowest conduction subband increases in energy, and the highest valence subband decreases in energy. Model calculations predict that the band overlap should vanish in Bi nanowires (with their wire axes along the trigonal direction) at a wire diameter ~ 50 nm [4.134].

The resistance of Bi nanowires is determined by two competing factors: the carrier density that increases with T, and the carrier mobility that decreases with T. The non-monotonic $R(T)$ for large-diameter Bi nanowires is due to a smaller carrier concentration variation at low temperature (≤ 100 K) in semimetals, so that the electrical resistance is dominated by the mobility factor in this temperature range. Based on the semiclassical transport model and the established band structure of Bi nanowires, the calculated $R(T)/R(300\,\mathrm{K})$ for 36-nm and 70-nm Bi nanowires is shown by the solid curves in Fig. 4.24c to illustrate different $R(T)$ trends for semiconducting and semimetallic nanowires, respectively [4.67]. The curves in Fig. 4.24c exhibit trends consistent with experimental results. The condition for the semimetal–semiconductor transition in Bi nanowires can be experimentally determined, as shown by the measured resistance ratio $R(10\,\mathrm{K})/R(100\,\mathrm{K})$ of Bi nanowires as a function of wire diameter [4.140] in Fig. 4.25. The maximum in the resistance ratio $R(10\,\mathrm{K})/R(100\,\mathrm{K})$ at $d_W \sim 48$ nm indicates the wire diameter for the transition of Bi nanowires from a semimetallic phase to a semiconducting phase. The semimetal–semiconductor transition and the semiconducting phase in Bi nanowires are examples of new transport phenomena, resulting from low dimensionality and absent in the bulk 3-D phase, which further increase the possible benefits from the properties of nanowires for desired applications (see Sect. 4.3.2).

It should be noted that good crystal quality is essential for observing the quantum size effect in nanowires, as shown by the $R(T)$ plots in Fig. 4.24a. For example, Fig. 4.24b shows the normalized $R(T)$ measurements of Bi nanowires with larger diameters (200 nm–2 μm) prepared by electrochemical deposition [4.30], and these nanowires possess monotonic $R(T)$ behaviors, quite different from those of the corresponding nanowire diameters shown in Fig. 4.24a. The absence of the resistance maximum in Fig. 4.24b is due to the lower crystalline quality for nanowires prepared by electrochemical deposition, which tends to produce polycrystalline nanowires with a much lower

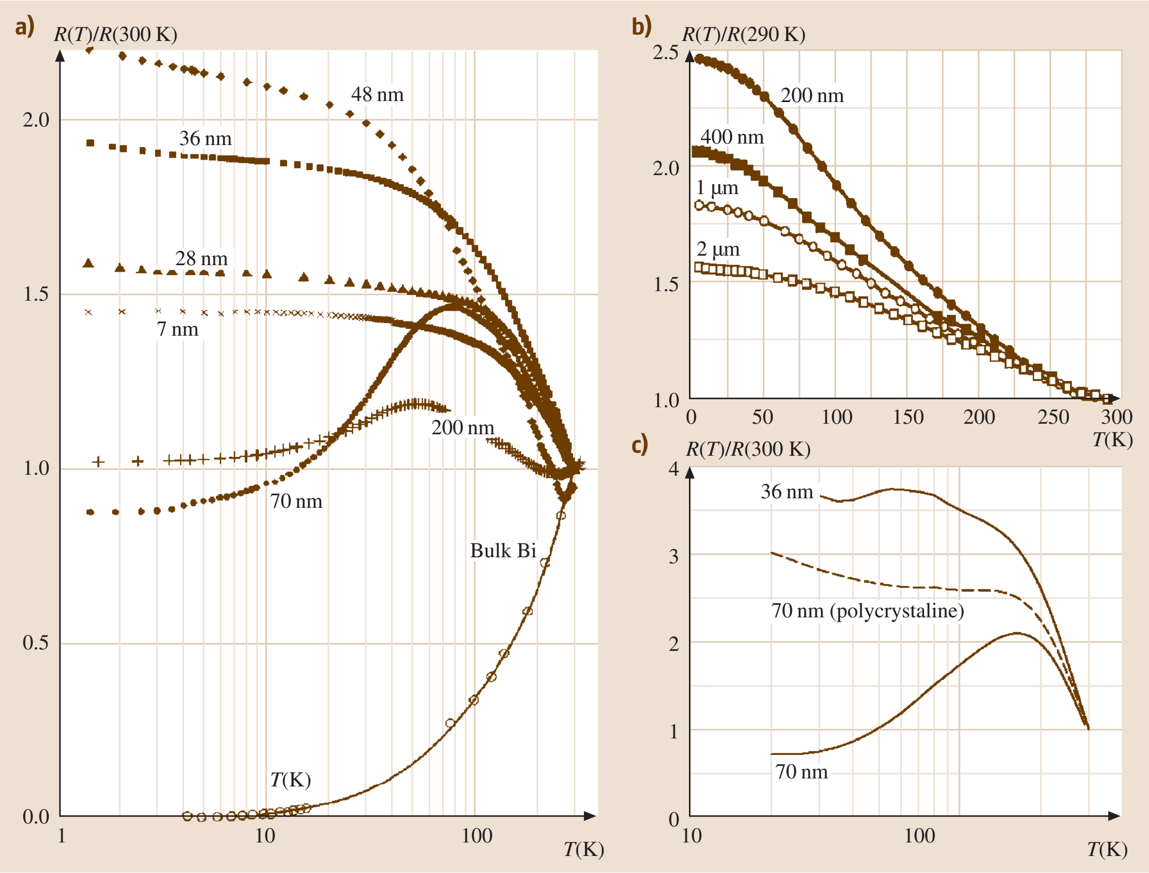

Fig. 4.24 **(a)** Measured temperature dependence of the resistance $R(T)$ normalized to the room temperature (300 K) resistance for bismuth nanowire arrays of various wire diameters d_W [4.28]. **(b)** $R(T)/R(290\,K)$ for bismuth wires of larger d_W and lower mobility [4.30]. **(c)** Calculated $R(T)/R(300\,K)$ of 36-nm and 70-nm bismuth nanowires. The *dashed curve* refers to a 70-nm polycrystalline wire with increased boundary scattering [4.139]

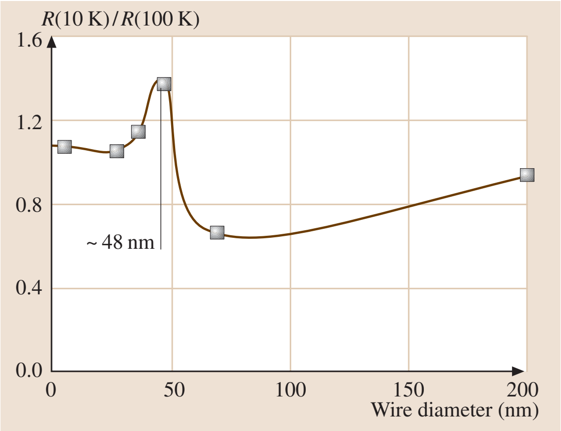

Fig. 4.25 Measured resistance ratio $R(10\,K)/R(100\,K)$ of Bi nanowires as a function of diameter. The peak indicates the transition from a semimetallic phase to a semiconducting phase as the wire diameter decreases [4.141]

carrier mobility. This monotonic $R(T)$ for semimetallic Bi nanowires at a higher defect level is also confirmed by theoretical calculations, as shown by the dashed curve in Fig. 4.24c for 70-nm wires with increased grain boundary scattering [4.139].

The theoretical model developed for Bi nanowires not only provides good agreement with experimental results, but it also plays an essential role in understanding the influence of the quantum size effect, the boundary scattering, and the crystal quality on their electrical properties. The transport model has also been generalized to predict the transport properties of Te-doped Bi nanowires [4.67], Sb nanowires [4.142], and BiSb alloy nanowires [4.143], and good agreement between experiment and theory has also been obtained for these cases. While the electronic density of states may be significantly altered due to quantum confinement effects, various scattering mechanisms related to the transport properties of nanowires can be accounted for by Matthiessen's rule.

For nanowires with diameters comparable to the phase-breaking length, their transport properties may be further influenced by localization effects. It has been predicted that in disordered systems, the extended electronic wave functions become localized near defect sites, resulting in the trapping of carriers and giving rise to a different transport behavior. Localization effects are also expected to be more pronounced as dimensionality and sample size are reduced. Localization effects on the transport properties of nanowire systems have been studied on Bi nanowires [4.144] and, more recently, on Zn nanowires [4.63]. Figure 4.26 shows the measured $R(T)/R(300\,\mathrm{K})$ of Zn nanowires fabricated by vapor deposition in porous silica or alumina [4.63]. While 15 nm Zn nanowires exhibit an $R(T)$ behavior with a T^1 dependence as expected for a metallic wire, the $R(T)$ of 9 nm and 4 nm Zn nanowires exhibits a temperature dependence of $T^{-1/2}$ at low temperature, consistent with 1-D localization theory. Thus, due to this localization effect, the use of nanowires with very small diameters for transport applications may be limited.

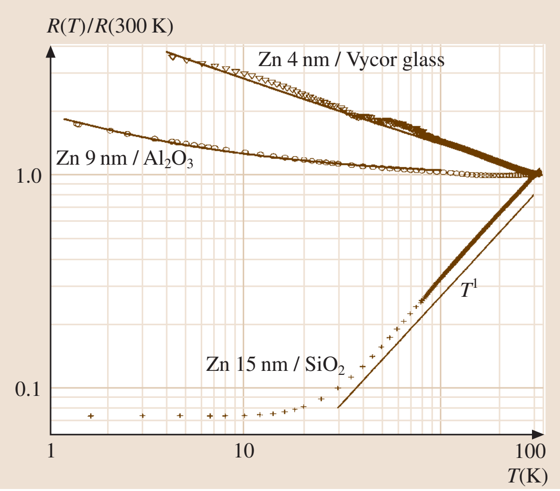

Fig. 4.26 Temperature dependence of the resistance of Zn nanowires synthesized by vapor deposition in various porous templates [4.63]. The data are given as points, the *full line* are fits to a T^1 law for 15 nm diameter Zn nanowires in an SiO_2 template, denoted by Zn/SiO_2. Fits to a combined T^1 and $T^{-1/2}$ law were made for the smaller nanowire diameter composite denoted by 9 nm Zn/Al_2O_3 and 4 nm Zn/Vycor glass

Magnetoresistance

Magnetoresistance (MR) measurements provide an informative technique for characterizing nanowires because these measurements yield a great deal of information about the electron scattering with wire boundaries, the effects of doping and annealing on scattering, and localization effects in the nanowires [4.137]. For example, at low fields the MR data show a quadratic dependence on the B field from which carrier mobility estimates can be made (see Fig. 4.27 at low B field).

Figure 4.27 shows the longitudinal magnetoresistance ($\boldsymbol{B}$ parallel to the wire axis) for 65 nm and 109 nm Bi nanowire samples (before thermal annealing) at 2 K. The MR maxima in Fig. 4.27a are due to the classical size effect, where the wire boundary scattering is reduced as the cyclotron radius becomes smaller than the wire radius in the high field limit, resulting in a decrease in the resistivity. This behavior is typical for the longitudinal MR of Bi nanowires in the diameter range of 45 nm to 200 nm [4.28, 136, 137, 145], and the peak position B_m moves to lower B field values as the wire diameter increases, as shown in Fig. 4.27c [4.145], in which B_m varies linearly with $1/d_\mathrm{W}$. The condition for the occurrance of B_m is approximately given by $B_\mathrm{m} \sim 2c\hbar k_\mathrm{F}/ed_\mathrm{W}$ where k_F is the wave vector at the Fermi energy. The peak position, B_m, is found to increase linearly with temperature in the range of 2 to 100 K, as shown in Fig. 4.27b [4.145]. As T is increased, phonon scattering becomes increasingly important, and therefore a higher magnetic field is required to reduce the resistivity associated with boundary scattering sufficiently to change the sign of the MR. Likewise, in-

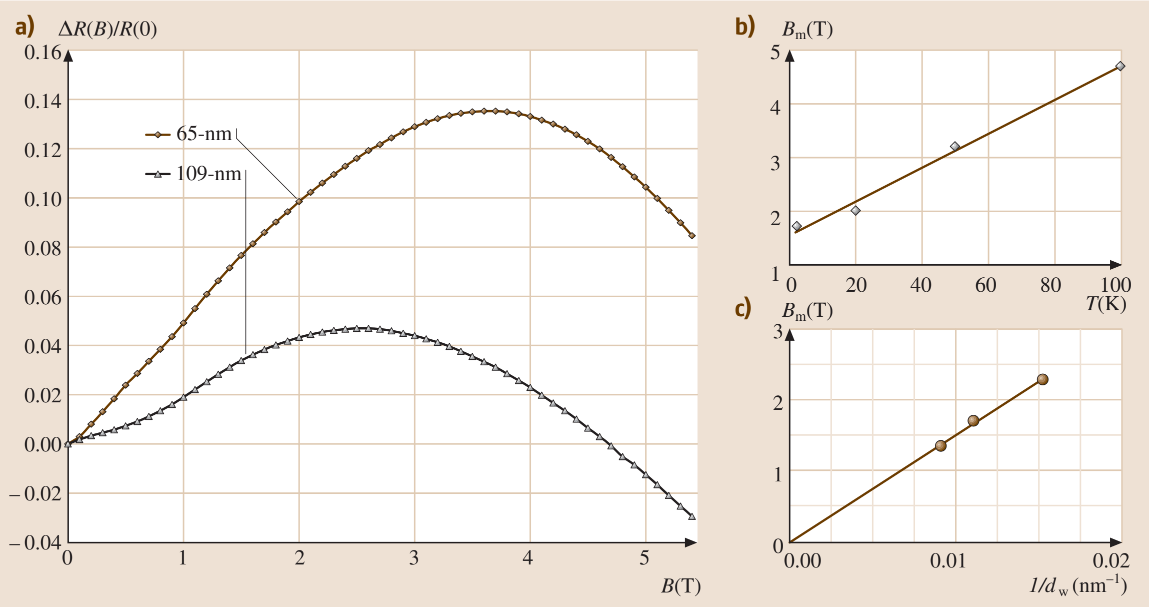

Fig. 4.27 **(a)** Longitudinal magnetoresistance, $\Delta R(B)/R(0)$, at 2 K as a function of B for Bi nanowire arrays with diameters 65 and 109 nm before thermal annealing. **(b)** The peak position B_m as a function of temperature for the 109 nm diameter Bi nanowire array after thermal annealing. **(c)** The peak position B_m of the longitudinal MR (after thermal annealing) at 2 K as a function of $1/d_W$, the reciprocal of the nanowire diameter [4.145]

creasing the grain boundary scattering is also expected to increase the value of B_m at a given T and wire diameter.

The presence of the peak in the longitudinal MR of nanowires requires a high crystal quality with long carrier mean free paths along the nanowire axis, so that most scattering events occur at the wire boundary instead of at a grain boundary, at impurity sites, or at defect sites within the nanowire. *Liu* et al. have investigated the MR of 400-nm Bi nanowires synthesized by electrochemical deposition [4.21], and no peak in the longitudinal MR is observed. The absence of a magnetoresistance peak may be attributed to a higher defect level in the nanowires produced electrochemically and to a large wire diameter, much longer than the carrier mean free path. The negative MR observed for the Bi nanowire arrays above B_m (see Fig. 4.27) shows that wire boundary scattering is a dominant scattering process for the longitudinal magnetoresistance, thereby establishing that the mean free path is larger than the wire diameter and that a ballistic transport behavior is indeed observed in the high field regime.

In addition to the longitudinal magnetoresistance measurements, transverse magnetoresistance measurements (**B** perpendicular to the wire axis) have also been performed on Bi nanowires array samples [4.28, 137, 145], where a monotonically increasing B^2 dependence over the entire range $0 \leq B \leq 5.5$ T is found for all Bi nanowires studied thus far. This is as expected, since the wire boundary scattering cannot be reduced by a magnetic field perpendicular to the wire axis. The transverse magnetoresistance is also found always to be larger than the longitudinal magnetoresistance in nanowire arrays.

By applying a magnetic field to nanowires at very low temperatures (≤ 5 K), one can induce a transition from a 1-D confined system at low magnetic fields to a 3-D confined system as the field strength increases, as shown in Fig. 4.28 for the longitudinal MR of Bi nanowire arrays of various nanowire diameters (28–70 nm) for $T < 5$ K [4.137]. In these curves, a subtle step-like feature is seen at low magnetic fields, which is found to depend only on the wire diameter, and is independent of temperature, the orientation of the magnetic field, and even on the nanowire material (e.g., in Sb nanowires [4.142]). The lack of a dependence of the magnetic field at which the step appears on temperature, field orientation, and material type indicates that

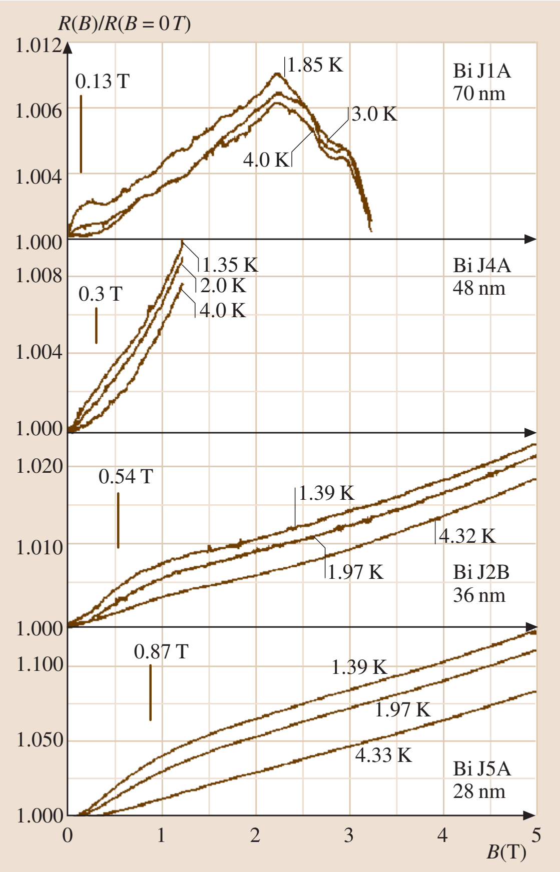

Fig. 4.28 Longitudinal magnetoresistance as a function of magnetic field for Bi nanowires of the diameters indicated. The vertical bars indicate the critical magnetic field B_c at which the magnetic length equals the nanowire diameter [4.137]

the phenomenon is related to the magnetic field length, $L_H = (\hbar/eB)^{1/2}$. The characteristic length L_H is the spatial extent of the wave function of electrons in the lowest Landau level, and L_H is independent of the carrier effective masses. Setting $L_H(B_c)$ equal to the diameter d_W of the nanowire defines a critical magnetic field strength, B_c, below which the wave function is confined by the nanowire boundary (the 1-D regime), and above which the wave function is confined by the magnetic field (the 3-D regime). The physical basis for this phenomenon is associated with confinement of a single magnetic flux quantum within the nanowire cross section [4.137]. This phenomenon, though independent of temperature, is observed for $T \leq 5$ K, since the phase breaking length has to be larger than the wire diameter. This calculated field strength, B_c, indicated in Fig. 4.28 by vertical lines for the appropriate nanowire diameters, provides a good fit to the step-like features in these MR curves.

The Shubnikov–de Haas (SdH) quantum oscillatory effect, which results from the passage of the quantized Landau levels through the Fermi energy as the field strength varies, should, in principle, provide the most direct measurement of the Fermi energy and carrier density. For example, *Heremans* et al. have demonstrated that SdH oscillations can be observed in Bi nanowire samples with diameters down to 200 nm [4.146] and have demonstrated that Te doping can be used to raise the Fermi energy in Bi nanowires. Such information on the Fermi energy is important because, for certain applications based on nanowires, it is necessary to place the Fermi energy near a subband edge where the density of states has a sharp feature. But due to the unusual 1-D geometry for nanowires, other characterization techniques commonly used in bulk materials to determine the Fermi energy and the carrier concentration (e.g., the Hall measurement) cannot be applied for nanowire systems. The observation of the SdH oscillatory effect requires very high crystal quality samples that allow carriers to execute a complete cyclotron orbit in the nanowire before they are scattered. For small nanowire diameters, large magnetic fields are required to produce cyclotron radii smaller than the wire radius. For some nanowire systems, all Landau levels may have passed through the Fermi level at such a high field strength, and in such a case, no oscillations can be observed. The localization effect may also prevent the observation of SdH oscillations for very small diameter (≤ 10 nm) nanowires. Observing SdH oscillations in highly doped samples (as may be required for certain applications) may be difficult because impurity scattering reduces the mean free path, requiring high $\boldsymbol{B}$ fields to satisfy the requirement that carriers complete a cyclotron orbit prior to scattering. Although SdH oscillations provide the most direct method of measuring the Fermi energy and carrier density of nanowire samples, this technique may not work, however, for smaller diameter nanowires, or for heavily doped nanowires.

Thermoelectric Properties

Nanowires are predicted to be promising for thermoelectric applications [4.134, 147], due to their novel band structure compared to their bulk counterparts and the expected reduction in thermal conductivity associated with enhanced boundary scattering (see Sect. 4.2.2). Due to the sharp density of states at the 1-D subband edges (where the van Hove singularities occur), nanowires are expected to exhibit enhanced Seebeck coefficients compared to their bulk counterparts. Since the Seebeck coefficient measurement is independent of the number of nanowires contributing to the signal, the measurements on nanowire arrays of uniform wire diameter are, in principle, as informative as single-wire measurements. The major challenge in measuring the Seebeck coefficient of nanowires lies in the design of tiny temperature probes to determine accurately the temperature difference across the nanowire. Figure 4.29a shows the schematic experimental setup for the Seebeck coefficient measurement of nanowire arrays [4.148], where two thermocouples are placed on both faces of a nanowire array and a heater is attached on one face of the array to generate a temperature gradient along the nanowire axis. Ideally the size of the thermocouples should be much smaller than the thickness of the nanowire array template (i. e., the nanowire length) to minimize error. However, due to the thinness of most templates ($\leq 50\,\mu m$) and the large size of commercially available thermocouples ($\sim 12\,\mu m$), the measured Seebeck coefficient values are usually underestimated.

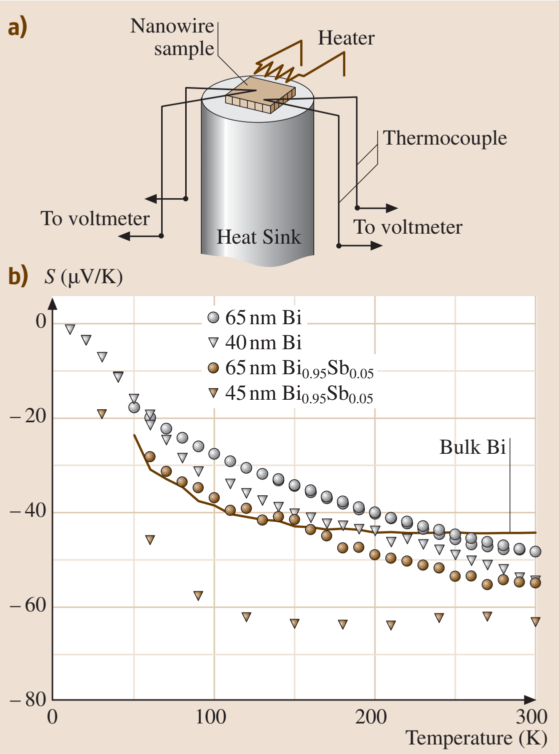

Fig. 4.29 **(a)** Experimental setup for the measurement of the Seebeck coefficient in nanowire arrays [4.148]. **(b)** Measured Seebeck coefficient as a function of temperature for Bi (○,▽) and $Bi_{0.95}Sb_{0.05}$ (●,▼) nanowires with different diameters. The *solid curve* denotes the Seebeck coefficient for bulk Bi [4.139]

The thermoelectric properties of Bi nanowire systems have been investigated extensively because of their potential as good thermoelectric materials. Figure 4.29b shows the measured Seebeck coefficient $S(T)$ as a function of temperature for nanowire arrays with diameters of 40 and 65 nm and different isoelectronic Sb alloy concentrations [4.139]; $S(T)$ results for bulk Bi are shown (*solid curve*) for comparison. Enhancement in the thermopower is observed in Fig. 4.29b as the wire diameter decreases and as the Sb content increases, which is attributed to the semimetal–semiconductor transition induced by quantum confinement and to Sb alloying effects in $Bi_{1-x}Sb_x$ nanowires. *Heremans* et al. have observed a substantial increase in the thermopower of Bi nanowires as the wire diameter further decreases, as shown in Fig. 4.30a for 15 nm Bi/silica and 9 nm Bi/alumina nanocomposites [4.63]. The enhancement is due to the sharp density of states near the Fermi energy in a 1-D system. Although the samples in Fig. 4.30a also possess very high electrical resistance ($\sim$ GΩ), the results for the 9 nm Bi/alumina samples show that the Seebeck coefficient can be enhanced by almost 1,000 times relative to bulk material. But for Bi nanowires with very small diameters (~ 4 nm), the localization effect becomes dominant, which compromises the thermopower enhancement. Therefore, for Bi nanowires, the optimal wire diameter range for the largest thermopower enhancement is found to be between 4 to 15 nm [4.63].

The effect of the nanowire diameter on the thermopower of nanowires has also been observed in Zn nanowires [4.63]. Figure 4.30b shows the Seebeck coefficient of 9 nm Zn/alumina and 4 nm Zn/Vycor glass nanocomposites, exhibiting enhanced thermopower as the wire diameter decreases. It is found that while 9 nm Zn nanowires still exhibit metallic behavior, the thermopower of 4 nm Zn nanowires shows a different temperature dependence, which may be due to the 1-D localization effect, although further investigation is required for a definitive identification of the conduction mechanism in such small nanowires.

Quantum Wire Superlattices

The studies on superlattice nanowires, which possess a periodic modulation in their materials composition

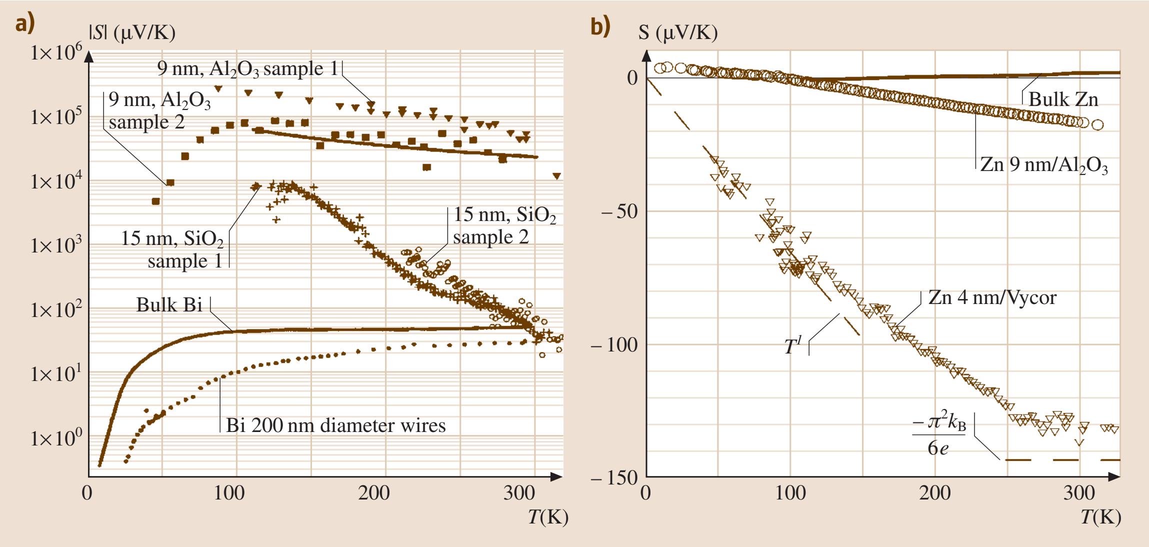

Fig. 4.30 **(a)** Absolute value of the Seebeck coefficient of two 15 nm Bi/silica samples, and two 9 nm Bi/alumina nanocomposite samples, in comparison to bulk Bi and 200 nm Bi nanowires in the pores of alumina templates [4.63]. The *full line* on top of the figure is a fit to a T^{-1} law. The Seebeck coefficient of the 9 nm Bi/alumina composite is positive; the rest are negative. **(b)** The Seebeck coefficient of 9 nm Zn/Al_2O_3 and 4 nm Zn/Vycor glass nanocomposite samples in comparison to bulk Zn [4.63]

along the wire axis, have attracted much attention recently because of their promise in such applications as thermoelectrics (see Sect. 4.3.2) [4.149], nanobarcodes (see Sect. 4.3.3) [4.90], nanolasers (see Sect. 4.3.3) [4.92], one-dimensional waveguides, and resonant tunneling diodes [4.93, 150]. Figure 4.31a shows a schematic structure of a superlattice nanowire consisting of interlaced quantum dots of two different materials, as denoted by A and B. Various techniques have been developed to synthesize superlattice nanowire structures with different interface conditions, as mentioned in Sect. 4.1.1 and Sect. 4.1.2.

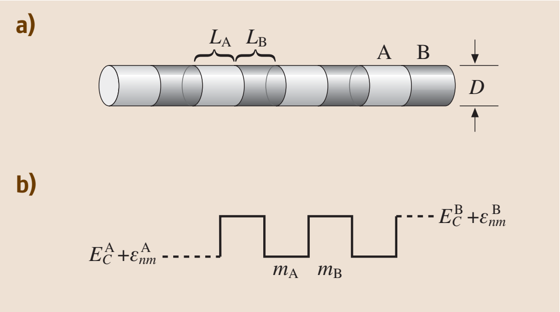

Fig. 4.31 **(a)** Schematic diagram of superlattice (segmented) nanowires consisting of interlaced nanodots A and B of the indicated length and wire diameter. **(b)** Schematic potential profile of the subbands in the superlattice nanowire [4.149]

In this superlattice (SL) nanowire structure, the electronic transport along the wire axis is made possible by the tunneling between adjacent quantum dots, while the uniqueness of each quantum dot and its 0-D characteristic behavior is maintained by the energy difference of the conduction or valence bands between quantum dots of different materials (see Fig. 4.31b), which provides some amount of quantum confinement. Recently *Björk* et al. have observed interesting nonlinear I–V characteristics with a negative differential resistance in one-dimensional heterogeneous structures made of InAs and InP, where InP serves as the potential barrier [4.93, 150]. The nonlinear I–V behavior is associated with the double barrier resonant tunneling in one-dimensional structures, demonstrating the capability of transport phenomena in superlattice nanowires via tunneling and the possibility of controlling the electronic band structure of the SL nanowires by carefully selecting the constituent materials. This kind of new structure is especially attractive for thermoelectric ap-

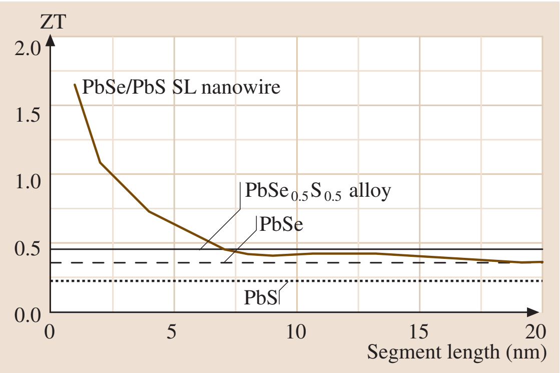

Fig. 4.32 Optimal ZT calculated as a function of segment length for 10-nm diameter PbSe/PbS nanowires at 77 K, where "optimal" refers to the placement of the Fermi level to optimize ZT. The optimal ZT for 10-nm diameter PbSe, PbS, and $PbSe_{0.5}S_{0.5}$ nanowires are 0.33, 0.22, and 0.48, respectively [4.141]

plications, because the interfaces between the nanodots can reduce the lattice thermal conductivity by blocking the phonon conduction along the wire axis, while electrical conduction may be sustained and even benefit from the unusual electronic band structures due to the periodic potential perturbation. For example, Fig. 4.32 shows the calculated dimensionless thermoelectric figure of merit $ZT = S^2\sigma T/\kappa$ (see Sect. 4.2.2) where κ is the total thermal conductivity (including both the lattice and electronic contribution) of 10-nm diameter PbS/PbSe superlattice nanowires as a function of the segment length. A higher thermoelectric performance than for $PbSe_{0.5}S_{0.5}$ alloy nanowires can be achieved for a 10 nm diameter superlattice nanowire with segment lengths ≤ 7 nm. But the localization effect, which may become important for very short segment lengths, may jeopardize this enhancement in the ZT of superlattice nanowires [4.141].

Thermal Conductivity of Nanowires

Experimental measurements of the temperature dependence of the thermal conductivity $\kappa(T)$ of individual suspended nanowires have been carried out to study the dependence of $\kappa(T)$ on the wire diameter. In this context, measurements have been made on nanowires down to only 22 nm in diameter [4.151]. Such measurements are very challenging and are now possible because of technological developments in fabricating and using nanometer size thermal scanning probes [4.120, 152, 153]. The experiments show that the thermal conductivity of small homogeneous nanowires may be more than one order of magnitude smaller than in the bulk, arising mainly from strong boundary scattering effects [4.154]. Phonon confinement effects may eventually become important at still smaller diameter nanowires (see Sect. 4.2.3). Measurements on mats of nanowires (see, for example, Fig. 4.12) do not generally give reliable results because the contact thermal resistance between adjacent nanowires tends to be high, which is in part due to the thin surface oxide coating that most nanowires have. This surface oxide coating may also be important for thermal conductivity measurements on individual suspended nanowires because of the relative importance of phonon scattering at the lateral walls of the nanowire.

The most extensive experimental thermal conductivity measurements have been done on Si nanowires [4.151] where $\kappa(T)$ measurements have been made on nanowires in the diameter range of $22 \leq d_W \leq 115$ nm. The results show a large decrease in the peak of $\kappa(T)$ associated with umklapp processes as d_W decreases, indicating a growing importance of boundary scattering and a corresponding decreasing importance of phonon–phonon scattering. At the smallest wire diameter of 22 nm, a linear $\kappa(T)$ dependence is found experimentally, consistent with a linear T dependence of the specific heat for a 1-D system and a temperature independent mean free path and velocity of sound.

Model calculations for $\kappa(T)$ based on a radiative heat transfer model have been carried out for Si nanowires [4.155]. These results show that the predicted $\kappa(T)$ behavior for Si nanowires is similar to that observed experimentally in the range of $37 \leq d_W \leq 115$ nm regarding both the functional form of $\kappa(T)$ and the magnitude of the relative decrease in the maximum thermal conductivity κ_{max} as a function of d_W. But the model calculations predict a substantially larger magnitude for $\kappa(T)$ (by 50% or more) than is observed experimentally. Furthermore, the model calculations (see Fig. 4.33) do not reproduce the experimentally observed linear T dependence for the 22 nm nanowires, but rather predict a 3-D behavior for both the density of states and the specific heat in 22 nm nanowires [4.155, 156].

Thermal conductance measurements on GaAs nanowires below 6 K show a power law dependence, but the T dependence becomes somewhat less pronounced below ~ 2.5 K [4.152]. This deviation from the power law temperature dependence led to a more detailed study of the quantum limit for the thermal conductance. To carry out these more detailed experiments, a mesoscopic phonon resonator and waveguide device were

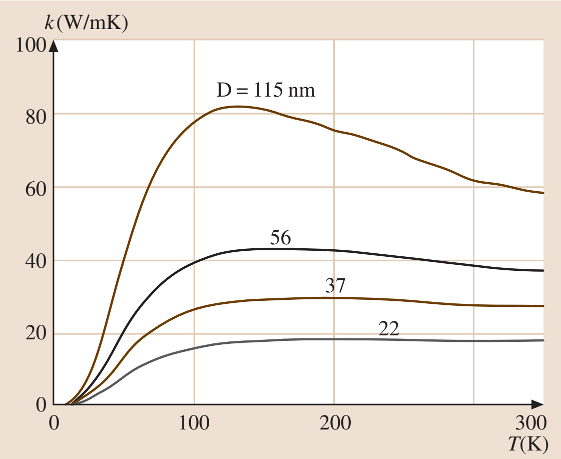

Fig. 4.33 Predicted thermal conductivity of Si nanowires of various diameters [4.155]

constructed that included four ~ 200 nm wide and 85 nm thick silicon nitride nanowire-like nano-constrictions (see Fig. 4.34a) to establish the quantized thermal conductance limit of $g_0 = \pi^2 k_B^2 T/3h$ (see Fig. 4.34b) for ballistic phonon transport [4.157, 158]. For temperatures above 0.8 K, the thermal conductance in Fig. 4.34b follows a T^3 law, but as T is further reduced, a transition to a linear T dependence is observed, consistent with a phonon mean free path of $\sim 1\,\mu$m, and a thermal conductance value approaching $16g_0$, corresponding to four massless phonon modes per channel and four channels in their phonon waveguide structure (see Fig. 4.34a). Ballistic phonon transport occurs when the thermal phonon wavelength (380 nm for the experimental structure) is somewhat greater than the width of the phonon waveguide at its constriction.

4.2.3 Optical Properties

Optical methods provide an easy and sensitive tool for measuring the electronic structure of nanowires since optical measurements require minimal sample preparation (for example, contacts are not required) and the measurements are sensitive to quantum effects. Optical spectra of 1-D systems, such as carbon nanotubes, often show intense features at specific energies near singularities in the joint density of states formed under strong quantum confinement conditions. A variety of optical techniques have shown that the properties of nanowires are different from those of their bulk counterparts, and this section of the review focuses on these differences.

Although optical properties have been shown to provide an extremely important tool for the characterization of nanowires, the interpretation of these measurements is not always straightforward. The wavelength of light used to probe the sample is usually smaller than the wire length but larger than the wire diameter. Hence, the probe light used in an optical measurement cannot be focused solely onto the wire, and the wire and the substrate on which the wire rests (or host material, if the wires are embedded in a template) are simultaneously probed. For measurements, such as photo-luminescence

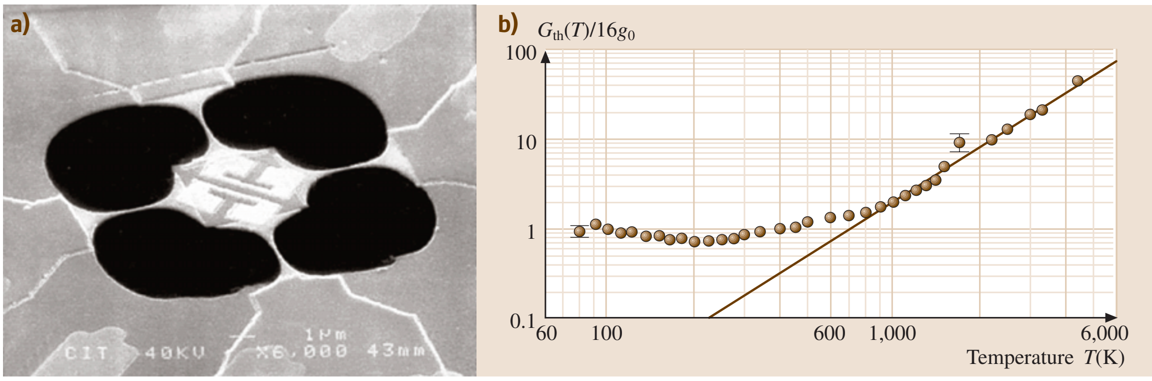

Fig. 4.34 **(a)** Suspended mesoscopic phonon device used to measure ballistic phonon transport. The device consists of an $4 \times 4\,\mu$m "phonon cavity" (*center*) connected to four Si_3N_4 membranes, 60 nm thick and less than 200 nm wide. The two bright "C" shaped objects on the phonon cavity are thin film heating and sensing Cr/Au resistors, whereas the *dark regions* are empty space. **(b)** Log–log plot of the temperature dependence of the thermal conductance G_0 of the structure in **(a)** normalized to $16g_0$ (see text) [4.158]

(PL), if the substrate does not luminescence or absorb in the frequency range of the measurements, PL directly measures the luminescence of the nanowires and the substrate can be ignored. In reflection and transmission measurements, however, even a nonabsorbing substrate can modify the measured spectra of nanowires.

In this section, we discuss the determination of the dielectric function for nanowires in the context of effective medium theories. We then discuss various optical techniques with appropriate examples that sensitively differentiate nanowire properties from those also found in the parent bulk material, giving particular emphasis to electronic quantum confinement effects. Finally, we review phonon confinement effects.

The Dielectric Function

In this subsection we review the use of effective medium theory as a method to handle the optical properties of nanowires whose diameters are typically smaller than the wavelength of light, noting that observable optical properties of materials can be related to the complex dielectric function [4.159, 160]. Effective medium theories [4.161, 162] can be applied to model the nanowire and substrate as one continuous composite with a single complex dielectric function $(\epsilon_1 + i\epsilon_2)$, where the real and imaginary parts of the dielectric function ϵ_1 and ϵ_2 are related to the index of refraction (n) and the absorption coefficient (K) by the relation $\epsilon_1 + i\epsilon_2 = (n + iK)^2$. Since photons at visible or infrared wavelengths "see" a dielectric function for the composite nanowire array/substrate system that is different from that of the nanowire itself, the optical transmission and reflection are different from what they would be if the light were focused only on the nanowire. One commonly observed consequence of effective medium theory is the shift in the plasma frequency in accordance with the percentage of nanowire material contained in the composite [4.163]. The plasma resonance occurs when $\epsilon_1(\omega)$ becomes zero, and the plasma frequency of the nanowire composite will shift to lower (higher) energies when the magnitude of the dielectric function of the host materials is larger (smaller) than that of the nanowire.

Although reflection and transmission measurements probe both the nanowire and the substrate, the optical properties of the nanowires can be independently determined. One technique for separating the dielectric function of the nanowires from the host is to use an effective medium theory in reverse. Since the dielectric function of the host material is often known, and since the dielectric function of the composite material can be measured by the standard method of using reflection and transmission measurements in combination with either the Kramer–Kronig relations or Maxwell's equations, the complex dielectric function of the nanowires can be deduced. This approach has been used successfully, for example, in determining the frequency dependence of the real and imaginary parts of the dielectric function $\epsilon_1(\omega)$ and $\epsilon_2(\omega)$ for a parallel array of bismuth nanowires filling the pores of an alumina template [4.164].

Optical Properties Characteristic of Nanowires

A wide range of optical techniques are available for the characterization of nanowires to distinguish their properties from those of their parent bulk materials. Some differences in properties are geometric, such as the small diameter size and the large length-to-diameter ratio (also called the aspect ratio), while other differences focus on quantum confinement issues.

Probably the most basic optical technique is to measure the reflection and/or transmission of a nanowire to determine the frequency-dependent real and imaginary parts of the dielectric function. This technique has been used, for example, to study the band gap and its temperature dependence in gallium nitride nanowires in the 10–50 nm range in comparison to bulk values [4.165]. The plasma frequency, free carrier density, and donor impurity concentration as a function of temperature were also determined from the infrared spectra, which is especially useful for nanowire research since Hall effect measurements cannot be made on nanowires.

Photo-luminescence (PL) or fluorescence spectroscopy is a common method to study nanowires. Emission techniques probe the nanowires directly, and the effect of the host material does not have to be considered. This characterization method has been used to study many properties of nanowires, such as the optical gap behavior, oxygen vacancies in ZnO nanowires [4.66], strain in Si nanowires [4.166], and quantum confinement effects in InP nanowires [4.92]. Figure 4.35 shows the photo-luminescence of InP nanowires as a function of wire diameter, thereby providing direct information on the effective band gap. As the wire diameter of an InP nanowire is decreased so that it becomes smaller than the bulk exciton diameter of 19 nm, quantum confinement effects set in, and the band gap is increased. This results in an increase in the PL peak energy because of the stronger electron–hole Coulomb binding energy within the quantum-confined nanowires as the wire radius gets smaller than the effective Bohr radius for the exciton for bulk InP. The smaller the effective mass, the larger the quantum con-

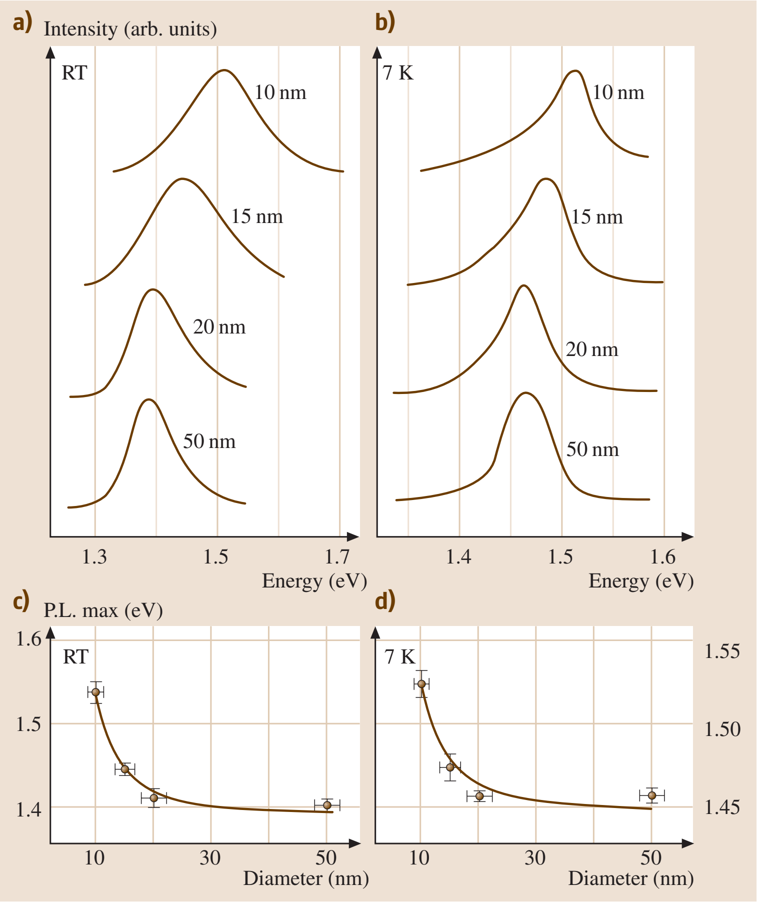

Fig. 4.35a–d Photo-luminescence of InP nanowires of varying diameters at 7 K ((**b**) and (**d**)) and room temperature ((**a**) and (**c**)), showing quantum confinement effects of the exciton for wire diameters less than 20 nm [4.92]

finement effects. When the shift in the peak energy as a function of nanowire diameter (Fig. 4.35a) is analyzed using an effective mass model, the reduced effective mass of the exciton is deduced to be $0.052m_0$, which agrees quite well with the literature value of $0.065m_0$ for bulk InP. Although the line widths of the PL peak for the small diameter nanowires (10 nm) are smaller at low temperature (7 K), the observation of strong quantum confinement and band gap tunability effects at room temperature are significant for photonics applications of nanowires (see Sect. 4.3.3).

The resolution of photo-luminescence (PL) optical imaging of a nanowire is, in general, limited by the wavelength of light. But when a sample is placed very close to the detector, the light is not given a chance to diffract, and so samples much smaller than the wavelength of light can be resolved. This technique is known as near-field scanning optical microscopy (NSOM) and has been used successfully [4.167] to image nanowires. For example, Fig. 4.36 shows the topographical (a) and PL (b) NSOM images of a single ZnO nanowire.

Magneto-optics can be used to measure the electronic band structure of nanowires. For example, magneto-optics in conjunction with photo-conductance has been proposed [4.168] as a tool to determine band parameters for nanowires, such as the Fermi energy, electron effective masses, and the number of subbands to be considered [4.168]. Since different nanowire subbands have different electrical transmission properties, the electrical conductivity changes when light is used to excite electrons to higher subbands, thereby providing a method for studying the electronic structure of nanowires optically. Magneto-optics can also be used to study the magnetic properties of nanowires in relation to bulk properties [4.44, 169]. For example, the surface magneto-optical Kerr effect has been used to measure

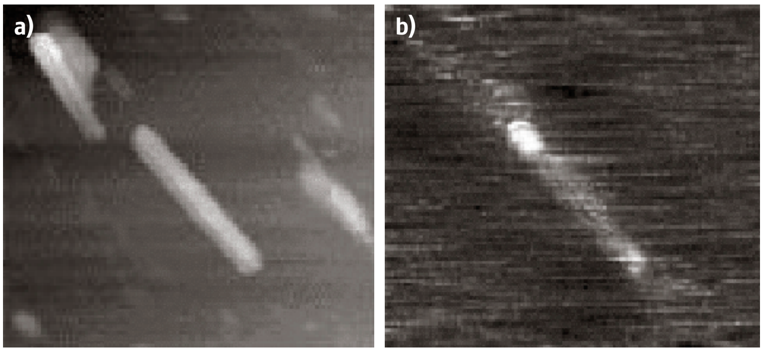

Fig. 4.36 **(a)** Topographical and **(b)** photoluminescence (PL) near-field scanning optical microscopy (NSOM) images of a single ZnO nanowire waveguide [4.167]

the dependence of the magnetic ordering temperature of Fe-Co alloy nanowires on the relative concentration of Fe and Co [4.169], and it was used to find that, unlike the case of bulk Fe-Co alloys, cobalt in nanowires inhibits magnetic ordering.

Nonlinear optical properties of nanowires have received particular attention since the nonlinear behavior is often enhanced over that in bulk materials and since the nonlinear effects can be utilized for many applications. One such study measured the second harmonic generation (SHG) and third harmonic generation (THG) in a single nanowire by using near-field optical microscopy [4.170]. ZnO nanowires were shown to have strong SHG and THG effects that are highly polarization-sensitive, and this polarization sensitivity can be explained on the basis of optical and geometrical considerations. Some components of the second harmonic polarization tensor are found to be enhanced in nanowires while others are suppressed as the wire diameter is decreased, and such effects could be of interest for device applications. The authors also showed that the second-order nonlinearities are mostly wavelength independent for $\lambda < 400\,\text{nm}$, which is in the transparent regime for ZnO, below the onset of band gap absorption; this observation is also of interest for device applications.

Reflectivity and transmission measurements have also been used to study the effects of quantum confinement and surface effects on the low energy indirect transition in bismuth nanowires [4.171]. *Black* et al. [4.171] investigated an intense and sharp absorption peak in bismuth nanowires, which is not observed in bulk bismuth. The energy position E_p of this strong absorption peak increases with decreasing diameter. But the rate of increase in energy with decreasing diameter $|\partial E_\text{p}/\partial d_\text{W}|$ is an order of magnitude less than that predicted for either a direct interband transition or for intersubband transitions in bismuth nanowires. On the other hand, the magnitude of $|\partial E_\text{p}/\partial d_\text{W}|$ agrees well with that predicted for an indirect L-point valence to T-point valence band transition (see Fig. 4.37). Since both the initial and final states for the indirect L-T point valence band transition downshift in energy as the wire diameter d_W is decreased, the shift in the absorption peak results from a *difference* between the effective masses and not from the actual value of either of the masses. Hence the diameter dependence of the absorption peak energy is an order of magnitude less for a valence to valence band indirect transition than for a direct interband L-point transition. Furthermore, the band-tracking effect for the indirect transition gives rise to a large value for the joint density of states, thus accounting for the high intensity of this feature. The enhanced coupling of this indirect transition to an applied optical field arises through the gradient of the dielectric function, which is large at the bismuth-air or bismuth-alumina interfaces. It should be noted that, in contrast to the surface effect for bulk samples, the whole nanowire contributes to the optical absorption due to the spatial variation in the dielectric function, since the penetration depth is larger than or comparable to the wire diameter. In addition, the intensity can be quite significant because abundant initial state electrons, final state holes, and appropriate phonons exist for making an indirect L-T point valence band transition at room temperature. Interestingly, the polarization dependence of this absorption peak is such that the strong absorption is present when the electric field is perpendicular to the wire axis but is absent when the electric field is parallel to the wire axis; this is contrary to a traditional polarizer, such as a carbon nanotube in which the optical E field is polarized by the nanotube itself to be aligned along the carbon nanotube axis. The observed polarization dependence for bismuth nanowires is consistent

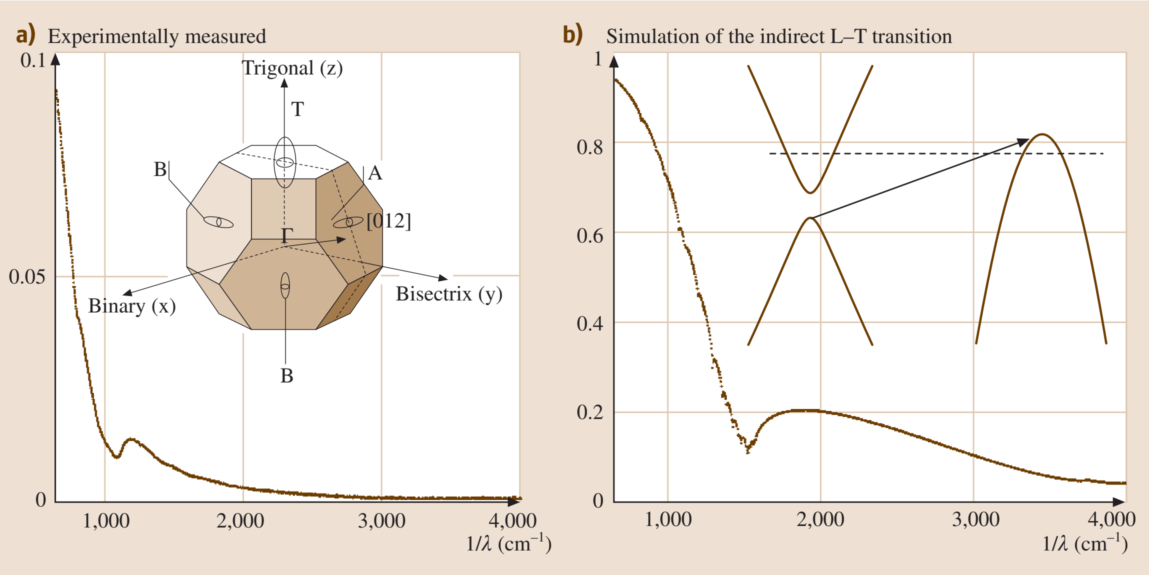

Fig. 4.37 **(a)** The measured optical transmission spectra as a function of wavenumber ($1/\lambda$) of a ~ 45 nm diameter bismuth nanowire array. **(b)** The simulated optical transmission spectrum resulting from an indirect transition of an L point electron to a T point valence subband state. The insert in **(a)** shows the bismuth Brillouin zone, and the location of the T-point hole and of the three L-point electron pockets, including the nondegenerate A, and the doubly degenerate B pockets. The insert in **(b)** shows the indirect L to T point electronic transition induced by a photon with an energy equal to the energy difference between the initial and final states minus the phonon energy (about $100\,\text{cm}^{-1}$) needed to satisfy conservation of energy in a Stokes process [4.171]

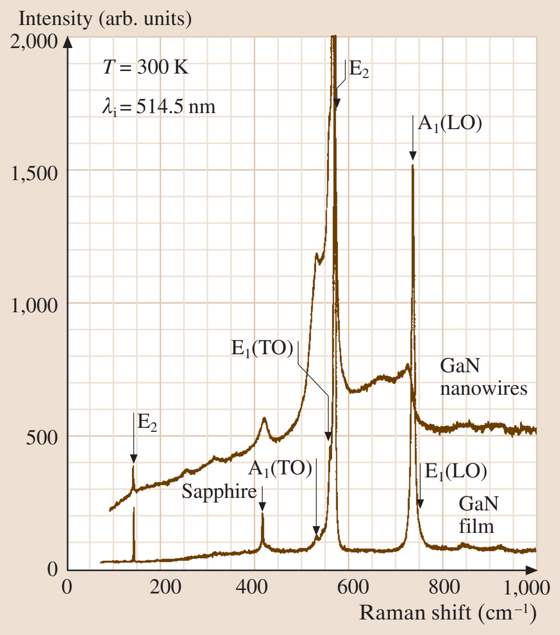

Fig. 4.38 Room temperature Raman-scattering spectra of GaN nanowires and of a 5 μm thick GaN epilayer film with green (514.5 nm) laser excitation. The Raman-scattering response was obtained by dividing the measured spectra by the Bose–Einstein thermal factor [4.172]

with a surface-induced effect that increases the coupling between the L-point and T-point bands throughout the full volume of the nanowire. Figure 4.37 shows the experimentally observed transmission spectrum in bismuth nanowires of ~ 45 nm diameter (a), and for comparison, the simulated optical transmission from an indirect transition in bismuth nanowires of ~ 45 nm diameter is also shown in (b). The indirect L to T point valence band transition mechanism [4.171] is also consistent with observations of the effect on the optical spectra of a decrease in the nanowire diameter and of n-type doping of bismuth nanowires with Te.

Phonon Confinement Effects

Phonons in nanowires are spatially confined by the nanowire cross-sectional area, crystalline boundaries and surface disorder. These finite size effects give rise to phonon confinement, causing an uncertainty in the phonon wave vector, which typically gives rise to a frequency shift and a lineshape broadening. Since zone center phonons tend to correspond to maxima in the phonon dispersion curves, the inclusion of contributions from a broader range of phonon wave vectors results in both a downshift in frequency and an asymmetric broadening of the Raman line that developes a low frequency tail. These phonon confinement effects have been theoretically predicted [4.173,174] and experimentally observed in GaN [4.172], as shown in Fig. 4.38 for GaN nanowires with diameters in the range 10–50 nm. The application of these theoretical models indicates that broadening effects should be noticeable as the wire diameter in GaN nanowires decreases to ~ 20 nm. When the wire diameter further decreases to ~ 10 nm, the frequency downshift and asymmetric Raman line broadening effects should become observable in the Raman spectra for the GaN nanowires but are not found in the corresponding spectra for bulk GaN.

The experimental spectra in Fig. 4.38 show the four $A_1 + E_1 + 2E_2$ modes expected from symmetry considerations for bulk GaN crystals. Two types of quantum confinement effects are observed. The first type is the observation of the downshift and asymmetric broadening effects discussed above. Observations of such downshifts and asymmetric broadening have also been recently reported in 7 nm diameter Si nanowires [4.175]. A second type of confinement effect found in Fig. 4.38 for GaN nanowires is the appearance of additional Raman features not found in the corresponding bulk spectra and associated with combination modes and a zone boundary mode. Resonant enhancement effects were also observed for the A_1(LO) phonon at 728 cm^{-1} (see Fig. 4.38) at higher laser excitation energies [4.172].

4.3 Applications

In the preceding sections we have reviewed many of the central characteristics that make nanowires in some cases similar and in some cases very different from their parent materials. We have also shown that some properties are diameter dependent and are therefore tunable during synthesis. Thus it is of great interest to find applications for nanowires that could benefit in unprecedented ways from both the unique and tunable properties of nanowires and the small size of these nanostructures, for use in the miniaturization of conventional devices. As the synthetic methods for the production of nanowires are maturing (Sect. 4.1), and nanowires can be made in reproducible and cost-effective ways, it is only a matter of time before applications will be explored seriously. This is a timely development, as the semiconductor industry will soon be reaching what seems to be its limit in feature-size reduction and approaching a classical-to-quantum size transition. At the same time the field of biotechnology is expanding through the availability of tremendous genome information and innovative screening assays. Since nanowires are the size of the shrinking electronic components and of cellular biomolecules, it is only natural for nanowires to be good candidates for applications in these fields. Commercialization of nanowire devices, however, will require reliable mass-production, effective assembly techniques, and quality-control methods.

In this section, we discuss applications of nanowires to electronics (Sect. 4.3.1), thermoelectrics (Sect. 4.3.2), optics (Sect. 4.3.3), chemical and biological sensing (Sect. 4.3.4), and magnetic media (Sect. 4.3.5).

4.3.1 Electrical Applications

The microelectronics industry continues to face technology (e.g., lithography) and economic challenges as the device feature size is decreased, especially below 100 nm. The self-assembly of nanowires might present a way to construct unconventional devices that do not rely on improvements in photo-lithography and, therefore, do not necessarily imply increasing fabrication costs. Devices made from nanowires have several advantages over those made by photolithography. A variety of approaches has been devised to organize nanowires via self-assembly (see Sect. 4.1.4), thus eliminating the need for the expensive lithographic techniques normally required to produce devices the size of typical nanowires, which we discuss earlier (see Sect. 4.1). In addition, unlike traditional silicon processing, different semiconductors can be simultaneously used in nanowire devices to produce diverse functionalities. Not only can wires of different materials be combined, but a single wire can be made of different materials. For example, junctions of GaAs and GaP show rectifying behavior [4.92], thus demonstrating that good electronic interfaces between two different semiconductors can be achieved in the synthesis of multicomponent nanowires. Transistors made from nanowires could also have advantages because of their unique morphology. For example, in bulk field effect transistors (FETs), the depletion layer formed below the source and drain region results in a source-drain capacitance that limits the operation speed. In nanowires, however, the conductor is surrounded by an oxide and thus the depletion layer cannot be formed. Depending on the device design, the source-drain capacitance in nanowires could be greatly minimized and possibly eliminated.

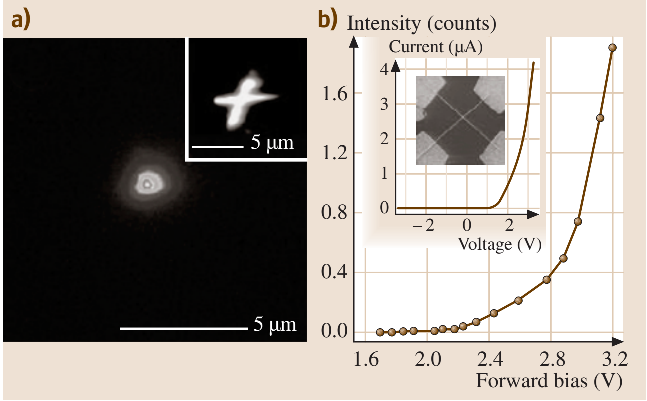

Fig. 4.39a,b Optoelectrical characterization of a crossed nanowire junction formed between 65-nm n-type and 68-nm p-type InP nanowires. **(a)** Electroluminescence (EL) image of the light emitted from a forward-biased nanowire p-n junction at 2.5 V. *Inset*, photoluminescence (PL) image of the junction. **(b)** EL intensity as a function of operation voltage. *Inset*, the SEM image and the I–V characteristics of the junction [4.52]. The scale bar in the *inset* is 5 µm

Device functionalities common in conventional semiconductor technologies, such the p-n junction diodes [4.129], field-effect transistors [4.131], logic gates [4.129], and light-emitting diodes [4.52, 92], have been recently demonstrated in nanowires, showing their promise as the building blocks for the construction of complex integrated circuits by employing the "bottom-up" paradigm. Several approaches have been investigated to form nanowire diodes (see Sect. 4.2.2). For example, Schottky diodes can be formed by contacting a GaN nanowire with Al electrodes [4.130]. Furthermore, p-n junction diodes can be formed at the crossing of two nanowires, such as the crossing of n and p-type InP nanowires doped by Te and Zn, respectively [4.52], or Si nanowires doped by phosphorus (n-type) and boron (p-type) [4.176]. In addition to the crossing of two distinctive nanowires, heterogeneous junctions have also been constructed inside a single wire, either along the wire axis in the form of a nanowire superlattice [4.92] or perpendicular to the wire axis by forming a core-shell structure of silicon and germanium [4.105]. These various nanowire junctions not only possess similar current rectifying properties (see Fig. 4.22) as expected for bulk semiconductor devices, but they also exhibit electro-luminescence (EL) that may be interesting for optoelectronic applications, as shown in Fig. 4.39 for the electroluminescence of a crossed junction of n and p-type InP nanowires [4.52] (see Sect. 4.3.3).

In addition to the two-terminal nanowire devices, such as the p-n junctions described above, it is found that the conductance of a semiconductor nanowire can be significantly modified by applying voltage at a third gate terminal, implying the utilization of nanowires as a field-effect transistor (FET). This gate terminal can either be the substrate [4.46, 47], a separate metal contact located close to the nanowire [4.177], or another nanowire with a thick oxide coating in the crossed nanowire junction configuration [4.129]. We discuss the operation principles of these nanowire-based FETs in Sect. 4.2.2. Various logic devices performing basic logic functions have been demonstrated using nanowire junctions [4.129], as shown in Fig. 4.40 for the OR and AND logic gates constructed from 2-by-1 and 1-by-3 nanowire p-n junctions, respectively. By functionalizing nanowires with redox active molecules to store charges, the nanowire FETs can exhibit bi-stable logic *on* or *off* states [4.131], which may be used for nonvolatile memory or as switches.

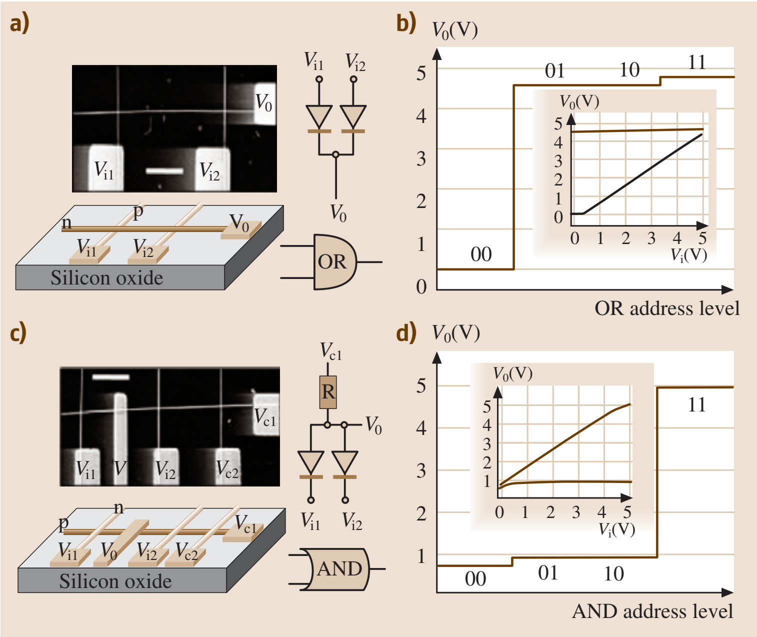

Fig. 4.40a–d Nanowire logic gates: **(a)** Schematic of logic OR gate constructed from a 2(p-Si) by 1(n-GaN) crossed nanowire junction. The *inset* shows the SEM image (*bar:* 1 μm) of an assembled OR gate and the symbolic electronic circuit. **(b)** The output voltage of the circuit in **(a)** versus the four possible logic address level inputs: (0,0); (0,1); (1,0); (1,1), where logic 0 input is 0 V and logic 1 is 5 V (same for below). **(c)** Schematic of logic AND gate constructed from a 1(p-Si) by 3(n-GaN) crossed nanowire junction. The *inset* shows the SEM image (*bar:* 1 μm) of an assembled AND gate and the symbolic electronic circuit. **(d)** The output voltage of the circuit in **(c)** versus the four possible logic address level inputs [4.129]

Nanowires have also been proposed for applications associated with electron field emission [4.178], such as flat panel displays, because of their small diameter and large curvature at the nanowire tip, which may reduce the threshold voltage for the electron emission [4.179]. In this connection, the demonstration of very high field emission currents from the sharp tip (∼ 10 nm radius) of a Si cone [4.178] and from carbon nanotubes [4.180] stimulates interest in this potential application opportunity for nanowires.

The concept of constructing electronic devices based on nanowires has already been demonstrated, and the next step for electronic applications would be to devise a feasible method for integration and mass production. We expect that in order to maintain the growing rate of device density and functionality in the existing electronics industry, new kinds of complementary electronic devices will emerge from this "bottom-up" scheme for nanowire electronics, different from what has been produced by the traditional "top-down" approach pursued by conventional electronics.

4.3.2 Thermoelectric Applications

One proposed application for nanowires is for thermoelectric cooling and for the conversion between thermal and electrical energy [4.156, 181]. The efficiency of a thermoelectric device is measured in terms of a dimensionless figure of merit ZT, where Z is defined as

$$Z = \frac{\sigma S^2}{\kappa}, \tag{4.2}$$

in which σ is the electrical conductivity, S is the Seebeck coefficient, κ is the thermal conductivity, and T is the temperature. In order to achieve a high ZT and therefore efficient thermoelectric performance, a high electrical conductivity, a high Seebeck coefficient, and a low thermal conductivity are required. In 3-D systems, the electronic contribution to κ is proportional to σ in accordance with the Wiedemann–Franz law, and normally materials with high S have a low σ. Hence an increase in the electrical conductivity (e.g., by electron donor doping) results in an adverse variation in both the Seebeck coefficient (decreasing) and the thermal conductivity (increasing). These two trade-offs set the upper limit for increasing ZT in bulk materials, with the maximum ZT remaining at ∼ 1 at room temperature for the 1960–1995 time frame.

The high electronic density of states in quantum-confined structures is proposed as a promising possibility to bypass the Seebeck/electrical conductivity trade-off and to control each thermoelectric-related variable independently, thereby allowing for an increased electrical conductivity, a relatively low thermal

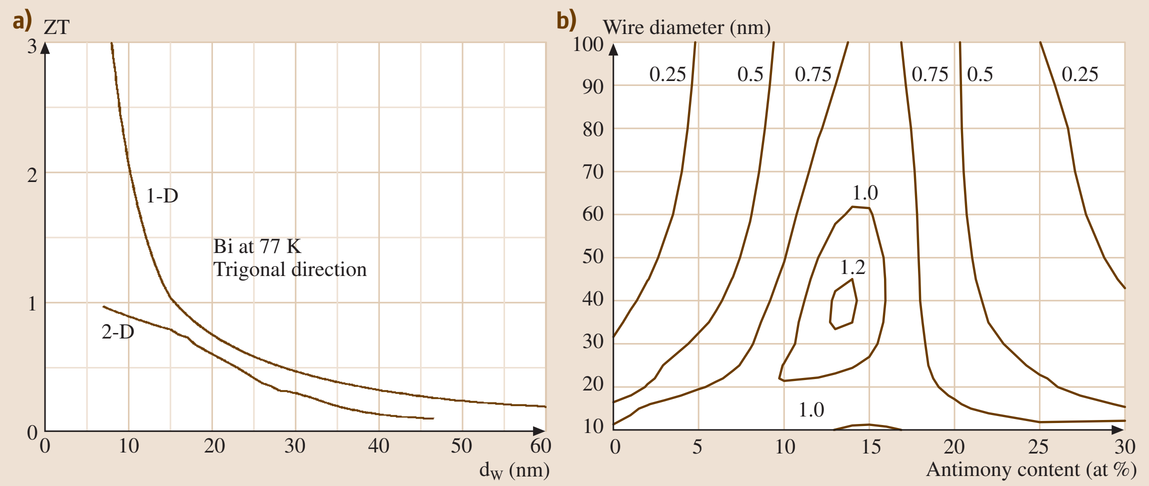

Fig. 4.41 **(a)** Calculated ZT of 1-D(nanowire) and 2-D(quantum well) bismuth systems at 77 K as a function of d_W, denoting the wire diameter or film thickness. The thermoelectric performance (i. e., ZT) is expected to improve greatly when the wire diameter is small enough so that the nanowire become a one-dimensional system. **(b)** Contour plot of optimal ZT values for p-type $Bi_{1-x}Sb_x$ nanowires vs. wire diameter and antimony concentration calculated at 77 K [4.183]

conductivity, and a large Seebeck coefficient simultaneously [4.182]. For example, Figs. 4.29b and 4.30a in Sect. 4.2.2 show an enhancement in S for bismuth an dbismuth-antimony nanowires as the wire diameter decreases. In addition to alleviating the undesired connections between σ, S, and the electronic contribution to the thermal conductivity, nanowires also have the advantage that the phonon contribution to the thermal conductivity is greatly reduced because of boundary scattering (see Sect. 4.2.2), thereby achieving a high ZT. Figure 4.41a shows the theoretical values for ZT vs. sample size for both bismuth thin films (2-D) and nanowires (1-D) in the quantum confined regime, exhibiting a rapidly increasing ZT as the quantum size effect becomes more and more important [4.182]. In addition, the quantum size effect in nanowires can be combined with other parameters to tailor the band structure and electronic transport behavior (e.g., Sb alloying in Bi) to further optimize ZT. For example, Fig. 4.41b shows the predicted ZT for p-type $Bi_{1-x}Sb_x$ alloy nanowires as a function of wire diameter and Sb content x [4.183]. The occurrence of a local ZT maxima in the vicinity of $x \sim 0.13$ and $d_W \sim 45$ nm is due to the coalesce of ten valence bands in the nanowire and the resulting unusual high density of states for holes, which is a phenomenon absent in bulk $Bi_{1-x}Sb_x$ alloys. For nanowires with very small diameters, it is speculated that localization effects will eventually limit the enhancement of ZT. But in bismuth nanowires, localization effects are not significant for wires with diameters larger than 9 nm [4.63]. In addition to 1-D nanowires, ZT values as high as ~ 2 have also been experimentally demonstrated in macroscopic samples containing PbSe quantum dots (0-D) [4.184] and stacked 2-D films [4.154].

Although the application of nanowires to thermoelectrics seems very promising, these materials are still in the research phase of the development cycle and quite far from being commercialized. One challenge for thermoelectric devices based on nanowires lies in finding a suitable host material that will not reduce ZT too much due to the unwanted heat conduction through the host material. The host material should, therefore, have a low thermal conductivity and occupy as low a volume percentage in the composite material as possible while still providing the quantum confinement and the support for the nanowires.

4.3.3 Optical Applications

Nanowires also hold promise for optical applications. One-dimensional systems exhibit a singularity in their joint density of states, allowing quantum effects in nanowires to be optically observable, sometimes, even at room temperature. Since the density of states of a nanowire in the quantum limit (small wire diameter) is highly localized in energy, the available states quickly fill up with electrons as the intensity of the incident light is increased. This filling up of the subbands, as well as other effects unique to low-dimensional materials, lead to strong optical nonlinearities in quantum wires. Quantum wires may thus yield optical switches with a lower switching energy and increased switching speed compared to currently available optical switches.

Light emission from nanowires can be achieved by photo-luminescence (PL) or electro-luminescence (EL), distinguished by whether the electronic excitation is achieved by optical illumination or by electrical stimulation across a p-n junction, respectively. PL is often used for optical properties characterization, as described in Sect. 4.2.3, but from the applications point of view, EL is a more convenient excitation method. Light emitting diodes (LEDs) have been achieved in junctions between a p-type and an n-type nanowire (Fig. 4.39) [4.52] and in superlattice nanowires with p-type and n-type segments [4.92]. The light emission was localized to the junction area and was polarized in the superlattice nanowire.

Light emission from quantum wire p-n junctions is especially interesting for laser applications, because quantum wires can form lasers with lower excitation thresholds compared to their bulk counterparts, and they also exhibit a decreased temperature sensitivity in their performance [4.185]. Furthermore, the emission wavelength can be tuned for a given material composition by only altering the geometry of the wire.

Lasing action has been reported in ZnO nanowires with wire diameters much smaller than the wavelength of the light emitted ($\lambda = 385$ nm) [4.115] (see Fig. 4.42). Since the edge and lateral surface of ZnO nanowires are faceted (see Sect. 4.2.1), they form optical cavities that sustain desired cavity modes. Compared to conventional semiconductor lasers, the exciton laser action employed in zinc oxide nanowire lasers exhibits a lower lasing threshold ($\sim 40\,\mathrm{kW/cm^2}$) than their 3-D counterparts ($\sim 300\,\mathrm{kW/cm^2}$). In order to utilize exciton confinement effects in the lasing action, the exciton binding energy (~ 60 meV in ZnO) must be greater than the thermal energy (~ 26 meV at 300 K). Decreasing the wire diameter increases the excitation binding energy and lowers the threshold for lasing. PL NSOM imaging confirmed the waveguiding properties of the anisotropic and well-faceted structure of ZnO nanowires, limiting the emission to the tips of the ZnO nanowires [4.167].

Laser action has been also observed in GaN nanowires [4.186]. Unlike ZnO, GaN has a small exciton binding energy, of only ~ 25 meV. Furthermore, since the wire radii used in this study (15–75 nm) [4.186] are larger than the Bohr radius of excitons in GaN (11 nm), the exciton binding energy is not expected to increase in these GaN wires and quantum confinement effects such as shown in Fig. 4.35 for InP are not expected. Some tunability of the center of the spectral intensity, however, was achieved by increasing the intensity of the pump power, causing a redshift in the laser emission, which is explained as a band gap renormalization as a result of the formation of an electron-hole plasma. Heating effects were excluded as the source of the spectral shift.

Nanowire photodetectors were also proposed. ZnO nanowires were found to display a strong photocurrent response to UV light irradiation [4.187]. The con-

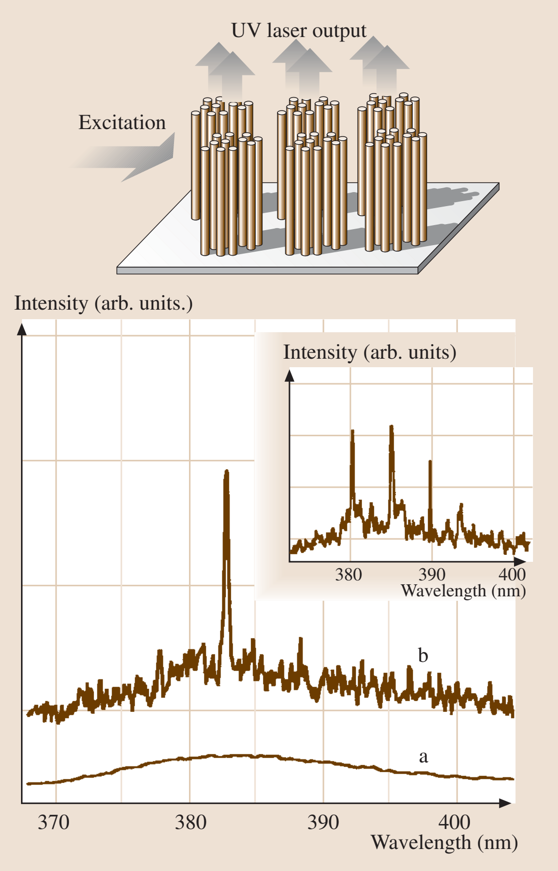

Fig. 4.42 A schematic of lasing in ZnO nanowires and the PL spectra of ZnO nanowires at two excitation intensities. One PL spectra is taken below the lasing threshold, and the other above it [4.115]

ductivity of the nanowire increased by four orders of magnitude compared to the dark state. The response of the nanowire was reversible and selective to photon energies above the band gap, suggesting that ZnO nanowires could be a good candidate for optoelectronic switches.

Nanowires have also been proposed for another type of optical switching. Light with its electric field normal to the wire axis excites a transverse free carrier resonance inside the wire, while light with its electric field parallel to the wire axis excites a longitudinal free carrier resonance inside the wire. Since nanowires are highly anisotropic, these two resonances occur at two different wavelengths and thus result in absorption peaks at two different energies. Gold nanowires dispersed in an aqueous solution align along the electric field when a DC voltage is applied. The energy of the absorption peak can be toggled between the transverse and longitudinal resonance energies by changing the alignment of the nanowires under polarized light illumination using an electric field [4.188, 189]. Thus, electro-optical modulation is achieved.

Nanowires may also be used as barcode tags for optical read out. Nanowires containing gold, silver, nickel, palladium, and platinum were fabricated [4.104] by electrochemical filling of porous anodic alumina, so that each nanowire consisted of segments of various metal constituents. Thus many types of nanowires can be made from a handful of materials and identified by the order of the metal segments along their main axis and the length of each segment. Barcode read out is possible by reflectance optical microscopy. The segment length is limited by the Rayleigh diffraction limit and not by synthesis limitations; thus it can be as small as 145 nm. Figure 4.43a shows an optical image of many Au-Ag-Au-Ag bar-coded wires, where the silver segments show higher reflectivity. Figure 4.43b is a backscattering mode FE-SEM image of a single nanowire, highlighting the composition and segment length variations along the nanowire.

Both the large surface area and the high conductivity along the length of nanowires are favorable for their use in inorganic-organic solar cells [4.190], which offer promise from a manufacture-ability and cost effective standpoint. In a hybrid nanocrystal-organic solar cell, the incident light forms bound electron–hole pairs (excitons) in both the inorganic nanocrystal and in the surrounding organic medium. These excitons diffuse to the inorganic–organic interface and disassociate to form an electron and a hole. Since conjugated polymers usually have poor electron mobilities, the inorganic phase is chosen to have a higher electron affinity than the organic phase so that the organic phase carries the holes and the semiconductor carries the electrons. The separated electrons and holes drift to the external electrodes through the inorganic and organic materials, respectively. But only those excitons formed within an exciton diffusion length from an interface can disassociate before recombining, and therefore the distance between the dissociation sites limits the efficiency of a solar cell. A solar cell prepared from a composite of CdSe nanorods inside poly (3-ethylthiophene) [4.190] yielded 6.9% monochromatic power efficiencies and 1.7% power conversion efficiencies under A.M. 1.5 illumination (equal to solar irradiance through 1.5 times the air mass of the earth at direct normal incidence). The nanorods provide a large surface area with good chemical bonding to the polymer for efficient charge transfer and exciton dissociation. Furthermore, they provide a good conduction path for the electrons to reach the electrode. Their enhanced absorption coefficient and their tunable band gap are also characteristics that can be used to enhance the energy conversion efficiency of solar cells.

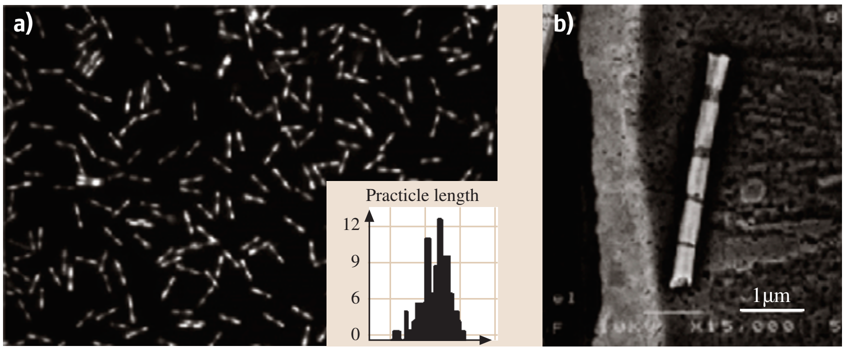

Fig. 4.43 **(a)** An optical image of many short Au-Ag-Au-Au bar-coded wires and **(b)** an FE-SEM image of an Au/Ag bar-coded wire with multiple strips of varying length. The insert in **(a)** shows a histogram of the particle lengths for 106 particles in this image [4.104]

4.3.4 Chemical and Biochemical Sensing Devices

Sensors for chemical and biochemical substances with nanowires as the sensing probe are a very attractive application area. Nanowire sensors will potentially be smaller, more sensitive, demand less power, and react faster than their macroscopic counterparts. Arrays of nanowire sensors could, in principle, achieve nanometer scale spatial resolution and therefore provide accurate real-time information regarding not only the concentration of a specific analyte but also its spatial distribution. Such arrays, for example, could be very useful for dynamic studies on the effects of chemical gradients on biological cells. The operation of sensors made with nanowires, nanotubes, or nano-contacts is based mostly on the reversible change in the conductance of the nanostructure upon absorption of the agent to be detected, but other detection methods, such as mechanical and optical detection, are conceptually plausible. The increased sensitivity and faster response time of nanowires are a result of the large surface-to-volume ratio and the small cross-section available for conduction channels. In the bulk, on the other hand, the abundance of charges can effectively shield external fields, and the abundance of material can afford many alternative conduction channels. Therefore a stronger chemical stimulus and longer response time are necessary to observe changes in the physical properties of a 3-D sensor in comparison to a nanowire.

Cui et al. placed silicon nanowires made by the VLS method (Sect. 4.1.2) between two metal electrodes and modified the silicon oxide coating of the wire by the addition of molecules sensitive to the analyte to be detected [4.191]. For example, a pH sensor was made by covalently linking an amine containing silane to the surface of the nanowire. Variations in the pH of the solution into which the nanowire was immersed caused protonation and deprotonation of the $-NH_2$ and the $-SiOH$ groups on the surface of the nanowire. The variation in surface charge densities regulates the conductance of the nanowire; due to the p-type characteristics of a silicon wire, the conductance increases with the addition of negative surface charge. The combined acid and base behavior of the surface groups results in an approximately linear dependence of the conductance on pH in the pH range 2 to 9, thus leading to a direct readout pH meter. This same type of approach was used for the detection of the binding of biomolecules, such as streptavidin using biotin-modified nanowires (see Fig. 4.44). This nanowire-based device has high sensitivity and could detect streptavidin binding down to a 10 pM (10^{-12} mole) concentration. The chemical detection devices were made in a field-effect transistor geometry, so that the back-gate potential could be used to regulate the conductance in conjunction with the chemical detection and to provide a real time direct read out [4.191]. The extension of this device to detect multiple analytes, using multiple nanowires each sensitized to a different analyte, could provide for fast, sensitive and in-situ screening procedures.

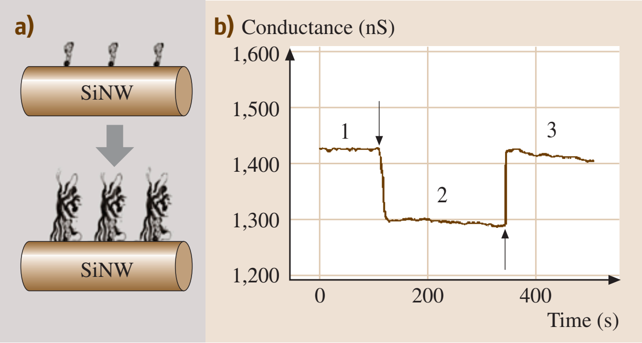

Fig. 4.44 **(a)** Streptavidin molecules bind to a silicon nanowire functionalized with biotin. The binding of streptavidin to biotin causes the nanowire to change its resistance. **(b)** The conductance of a biotin-modified silicon nanowire exposed to streptavidin in a buffer solution (regions 1 and 3) and with the introduction of a solution of antibiotin monoclonal antibody (region 2) [4.191]

Favier et al. used a similar approach, making a nanosensor for the detection of hydrogen out of an array of palladium nanowires between two metal contacts [4.56]. They demonstrated that nanogaps were present in their nanowire structure, and upon absorption of H_2 and formation of Pd hydride, the nanogap structure would close and improve the electrical contact, thereby increasing the conductance of the nanowire array. The response time of these sensors was 75 msec, and they could operate in the range of 0.5 to 5% H_2 before saturation occurred.

4.3.5 Magnetic Applications

It has been demonstrated that arrays of single domain magnetic nanowires can be prepared with controlled nanowire diameter and length, aligned along a common direction and arranged in a close-packed ordered array (see Sect. 4.1.1), and that the magnetic properties (coercivity, remanence, and dipolar magnetic interwire interaction) can be controlled to achieve a variety of magnetic applications [4.54, 71].

Magnetic information storage is the most interesting of these applications. The large nanowire aspect ratio (length/diameter) prevents the onset of the "superparamagnetic" limit at which the magnetization direction in the magnetic grains can be reversed by the thermal energy $k_B T$, thereby resulting in loss of recorded data in the magnetic recording medium. The magnetic energy in a grain can be increased by increasing either the volume or the anisotropy of the grain. If the volume is increased, the particle size increases, and thus, the resolution is decreased. For spherical magnetized grains, the superparamagnetic limit at room temperature is reached at 70 Gbit/in^2. In nanowires, the anisotropy is very large and yet the wire diameters are small, so the magnetostatic switching energy can easily be above the thermal energy while the spatial resolution is large. For magnetic data storage applications, a large aspect ratio is needed for the nanowires to maintain a high coercivity, and a sufficient separation between nanowires is needed to suppress inter-wire magnetic dipolar coupling (see Sect. 4.2.1). Thus nanowires can form stable and highly dense magnetic memory arrays with packing densities in excess of 10^{11} wires/cm^2.

The onset of superparamagnetism can be prevented in the single domain magnetic nanowire arrays that have already been fabricated either using porous alumina templates to make Ni nanowires with 35 nm diameters [4.54] or diblock copolymer templates [4.71] to make Co nanowires, with mean diameters of 14 nm and 100% filling of the template pores (see Sect. 4.1.1). Ordered magnetic nanowire arrays that have already been demonstrated offer the exciting promise for achieving systems permitting 10^{12} bits/in^2 data storage.

4.4 Concluding Remarks

A review has been presented of synthesis, characterization, and physical properties of nanowires, with particular reference to nanowire properties that differ from those of their bulk counterparts and to potential applications that might result from the special structure and properties of nanowires. We have shown in this chapter that the emerging field of nanowire research has developed rapidly over the past few years, driven by the development of a variety of complementary nanowire synthesis methods and of effective tools to measure both their structure and properties, as reviewed, respectively, in Sect. 4.1 and Sect. 4.2. At present, much of the progress is at the demonstration-of-concept level, with many gaps in knowledge remaining to be elucidated, theoretical models to be developed, and new nanowire systems to be explored. Having demonstrated that many of the most interesting discoveries to date relate to nanowire properties not present in their bulk material counterparts, we can expect future research emphasis to be focused increasingly on smaller diameter nanowires, where new unexplored physical phenomena related to quantum confinement effects are more likely to be found. We can also expect the development of applications to soon follow. Many promising applications are now at the early demonstration stage (see Sect. 4.3), but are moving ahead rapidly because of their promise for new functionality, not previously available, to the fields of electronics, opto-electronics, biotechnology, magnetics, and energy conversion and generation, among others. Many exciting challenges remain in advancing both the nanoscience and the nano-technological promise already demonstrated by the nanowire research described in this review.

References

4.1 R. P. Feynman: There's plenty of room at the bottom, Eng. Sci. (February 1960) 22

4.2 M. S. Dresselhaus, Y.-M. Lin, O. Rabin, A. Jorio, A. G. Souza Filho, M. A. Pimenta, R. Saito, G. G. Samsonidze, G. Dresselhaus: Nanowires and nanotubes, Mater. Sci. & Eng. C **1008** (2002) 1–12

4.3 R. Saito, G. Dresselhaus, M. S. Dresselhaus: *Physical Properties of Carbon Nanotubes* (Imperial College Press, London 1998)

4.4 M. S. Dresselhaus, G. Dresselhaus, P. Avouris (Eds.): *Carbon Nanotubes: Synthesis Structure, Properties and Applications* (Springer, Berlin, Heidelberg 2001)

4.5 R. C. Haddon: Special issue on carbon nanotubes, Accounts Chem. Res. **35** (2002) 997–1113

4.6 Y.-M. Lin, X. Sun, S. Cronin, Z. Zhang, J. Y. Ying, M. S. Dresselhaus: Fabrication and transport properties of Te-doped bismuth nanowire arrays, Molecular Electronics: MRS Symp. Proc., Boston 1999, ed. by S. T. Pantelides, M. A. Reed, J. Murday, A. Aviran (Materials Research Society Press, Pittsburgh 2000) **582** H10.3 (1–6)

4.7 C. R. Martin: Nanomaterials: A membrane-based synthetic approach, Science **266** (1994) 1961–1966

4.8 G. A. Ozin: Nanochemistry: synthesis in diminishing dimensions, Adv. Mater. **4** (1992) 612–649

4.9 R. J. Tonucci, B. L. Justus, A. J. Campillo, C. E. Ford: Nanochannel array glass, Science **258** (1992) 783–785

4.10 J. Y. Ying: Nanoporous systems and templates, Sci. Spectra **18** (1999) 56–63

4.11 J. W. Diggle, T. C. Downie, C. W. Goulding: Anodic oxide films on aluminum, Chem. Rev. **69** (1969) 365–405

4.12 J. P. O'Sullivan, G. C. Wood: The morphology and mechanism of formation of porous anodic films on aluminum, Proc. Roy. Soc. London A **317** (1970) 511–543

4.13 A. P. Li, F. Müller, A. Birner, K. Neilsch, U. Gösele: Hexagonal pore arrays with a 50–420 nm interpore distance formed by self-organization in anodic alumina, J. Appl. Phys. **84** (1998) 6023–6026

4.14 O. Jessensky, F. Müller, U. H. Gösele: Self-organized formation of hexagonal pore arrays in anaodic alumina, Appl. Phys. Lett. **72** (1998) 1173–1175

4.15 D. AlMawlawi, N. Coombs, M. Moskovits: Magnetic-properties of Fe deposited into anodic aluminum-oxide pores as a function of particle-size, J. Appl. Phys. **70** (1991) 4421–4425

4.16 Z. Zhang, D. Gekhtman, M. S. Dresselhaus, J. Y. Ying: Processing and characterization of single-crystalline ultrafine bismuth nanowires, Chem. Mater. **11** (1999) 1659–1665

4.17 F. Li, L. Zhang, R. M. Metzger: On the growth of highly ordered pores in anodized aluminum oxide, Chem. Mater. **10** (1998) 2470–2480

4.18 H. Masuda, M. Satoh: Fabrication of gold nanodot array using anodic porous alumina as an evaporation mask, Jpn. J. Appl. Phys. **35** (1996) L126–L129

4.19 E. Ferain, R. Legras: Track-etched membrane – dynamics of pore formation, Nucl. Instrum. Methods B **84** (1993) 331–336

4.20 A. Blondel, J. P. Meier, B. Doudin, J.-P. Ansermet: Giant magnetoresistance of nanowires of multilayers, Appl. Phys. Lett. **65** (1994) 3019–3021

4.21 K. Liu, C. L. Chien, P. C. Searson, Y. Z. Kui: Structural and magneto-transport properties of electrodeposited bismuth nanowires, Appl. Phys. Lett. **73** (1998) 1436–1438

4.22 L. Sun, P. C. Searson, C. L. Chien: Electrochemical deposition of nickel nanowire arrays in single-crystal mica films, Appl. Phys. Lett. **74** (1999) 2803–2805

4.23 E. Braun, Y. Eichen, U. Sivan, G. Ben-Yoseph: DNA-templated assembly and electrode attachment of a conducting silver wire, Nature **391** (1998) 775–778

4.24 G. Sauer, G. Brehm, S. Schneider, K. Nielsch, R. B. Wehrspohn, J. Choi, H. Hofmeister, U. Gösele: Highly ordered monocrystalline silver nanowire arrays, J. Appl. Phys. **91** (2002) 3243–3247

4.25 G. L. Hornyak, C. J. Patrissi, C. M. Martin: Fabrication, characterization, and optical properties of gold nanoparticle/porous alumina composites: the non-scattering Maxwell-Garnett limit, J. Phys. Chem. B **101** (1997) 1548–1555

4.26 X. Y. Zhang, L. D. Zhang, Y. Lei, L. X. Zhao, Y. Q. Mao: Fabrication and characterization of highly ordered Au nanowire arrays, J. Mater. Chem. **11** (2001) 1732–1734

4.27 Y.-T. Cheng, A. M. Weiner, C. A. Wong, M. P. Balogh, M. J. Lukitsch: Stress-induced growth of bismuth nanowires, Appl. Phys. Lett. **81** (2002) 3248–3250

4.28 J. Heremans, C. M. Thrush, Y.-M. Lin, S. Cronin, Z. Zhang, M. S. Dresselhaus, J. F. Mansfield: Bismuth nanowire arrays: synthesis and galvanomagnetic properties, Phys. Rev. B **61** (2000) 2921–2930

4.29 L. Piraux, S. Dubois, J. L. Duvail, A. Radulescu, S. Demoustier-Champagne, E. Ferain, R. Legras: Fabrication and properties of organic and metal nanocylinders in nanoporous membranes, J. Mater. Res. **14** (1999) 3042–3050

4.30 K. Hong, F. Y. Yang, K. Liu, D. H. Reich, P. C. Searson, C. L. Chien, F. F. Balakirev, G. S. Boebinger: Giant positive magnetoresistance of Bi nanowire arrays in high magnetic fields, J. Appl. Phys. **85** (1999) 6184–6186

4.31 A. J. Yin, J. Li, W. Jian, A. J. Bennett, J. M. Xu: Fabrication of highly ordered metallic nanowire arrays by electrodeposition, Appl. Phys. Lett. **79** (2001) 1039–1041

4.32 Z. Zhang, J. Y. Ying, M. S. Dresselhaus: Bismuth quantum-wire arrays fabricated by a vacuum melting and pressure injection process, J. Mater. Res. **13** (1998) 1745–1748

4.33 T. E. Huber, M. J. Graf, C. A. Foss Jr., P. Constant: Processing and characterization of high-conductance bismuth wire array composites, J. Mater. Res. **15** (2000) 1816–1821

4.34 M. S. Sander, A. L. Prieto, R. Gronsky, T. Sands, A. M. Stacy: Fabrication of high-density, high aspect ratio, large-area bismuth telluride nanowire arrays by electrodeposition into porous anodic alumina templates, Adv. Mater. **14** (2002) 665–667

4.35 M. Chen, Y. Xie, J. Lu, Y. J. Xiong, S. Y. Zhang, Y. T. Qian, X. M. Liu: Synthesis of rod-, twinrod-, and tetrapod-shaped CdS nanocrystals using a highly oriented solvothermal recrystallization technique, J. Mat. Chem. **12** (2002) 748–753

4.36 D. Xu, Y. Xu, D. Chen, G. Guo, L. Gui, Y. Tang: Preparation of CdS single-crystal nanowires by electrochemically induced deposition, Adv. Mater. **12** (2000) 520–522

4.37 D. Routkevitch, T. Bigioni, M. Moskovits, J. M. Xu: Electrochemical fabrication of CdS nanowire arrays in porous anodic aluminum oxide templates, J. Phys. Chem. **100** (1996) 14037–14047

4.38 L. Manna, E. C. Scher, A. P. Alivisatos: Synthesis of soluble and processable rod-, arrow-, teardrop-, and tetrapod-shaped CdSe nanocrystals, J. Am. Chem. Soc. **122** (2000) 12700–12706

4.39 D. Routkevitch, A. A. Tager, J. Haruyama, D. Al-Mawlawi, M. Moskovits, J. M. Xu: Nonlithographic nano-wire arrays: fabrication, physics, and device applications, IEEE Trans. Electron. Dev. **43** (1996) 1646–1658

4.40 D. S. Xu, D. P. Chen, Y. J. Xu, X. S. Shi, G. L. Guo, L. L. Gui, Y. Q. Tang: Preparation of II-VI group semiconductor nanowire arrays by dc electrochemical deposition in porous aluminum oxide templates, Pure Appl. Chem. **72** (2000) 127–135

4.41 R. Adelung, F. Ernst, A. Scott, M. Tabib-Azar, L. Kipp, M. Skibowski, S. Hollensteiner, E. Spiecker, W. Jäger, S. Gunst, A. Klein, W. Jägermann, V. Zaporojtchenko, F. Faupel: Self-assembled nanowire networks by deposition of copper onto layered-crystal surfaces, Adv. Mater. **14** (2002) 1056–1061

4.42 T. Gao, G. W. Meng, J. Zhang, Y. W. Wang, C. H. Liang, J. C. Fan, L. D. Zhang: Template synthesis of single-crystal Cu nanowire arrays by electrodeposition, Appl. Phys. A **73** (2001) 251–254

4.43 F. Li, R. M. Metzger: Activation volume of α-Fe particles in alumite films, J. Appl. Phys. **81** (1997) 3806–3808

4.44 A. Sugawara, T. Coyle, G. G. Hembree, M. R. Scheinfein: Self-organized Fe nanowire arrays prepared by shadow deposition on NaCl(110) templates, Appl. Phys. Lett. **70** (1997) 1043–1045

4.45 G. S. Cheng, L. D. Zhang, Y. Zhu, G. T. Fei, L. Li, C. M. Mo, Y. Q. Mao: Large-scale synthesis of single crystalline gallium nitride nanowires, Appl. Phys. Lett. **75** (1999) 2455–2457

4.46 Y. Huang, X. Duan, Y. Cui, C. M. Lieber: Gallium nitride nanowire nanodevices, Nano. Lett. **2** (2002) 101–104

4.47 Y. Cui, X. Duan, J. Hu, C. M. Lieber: Doping and electrical transport in silicon nanowires, J. Phys. Chem. B **104** (2000) 5213–5216

4.48 A. D. Berry, R. J. Tonucci, M. Fatemi: Fabrication of GaAs and InAs wires in nanochannel glass, Appl. Phys. Lett. **69** (1996) 2846–2848

4.49 J. R. Heath, F. K. LeGoues: A liquid solution synthesis of single-crystal germanium quantum wires, Chem. Phys. Lett. **208** (1993) 263–268

4.50 Y. Wu, P. Yang: Germanium nanowire growth via simple vapor transport, Chem. Mater. **12** (2000) 605–607

4.51 Y. F. Zhang, Y. H. Tang, N. Wang, C. S. Lee, I. Bello, S. T. Lee: Germanium nanowires sheathed with an oxide layer, Phys. Rev. B **61** (2000) 4518–4521

4.52 X. Duan, Y. Huang, Y. Cui, J. Wang, C. M. Lieber: Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices, Nature **409** (2001) 66–69

4.53 M. P. Zach, K. H. Ng, R. M. Penner: Molybdenum nanowires by electrodeposition, Science **290** (2000) 2120–2123

4.54 K. Nielsch, R. Wehrspohn, S. F. H. Kronmuller, J. Barthel, J. Kirschner, U. Gosele: Magnetic properties of 100 nm nickel nanowire arrays obtained from ordered porous alumina templates, MRS Symp. Proc. **636** (2001) D19 (1–6)

4.55 E. Lifshitz, M. Bashouti, V. Kloper, A. Kigel, M. S. Eisen, S. Berger: Synthesis and characterization of PbSe quantum wires, multipods, quantum rods, and cubes, Nano. Lett. **3** (2003) 857–862

4.56 F. Favier, E. C. Walter, M. P. Zach, T. Benter, R. M. Penner: Hydrogen sensors and switches from electrodeposited palladium mesowire arrays, Science **293** (2001) 2227–2231

4.57 B. Gates, B. Mayers, B. Cattle, Y. Xia: Synthesis and characterization of uniform nanowires of trigonal selenium, Adv. Funct. Mat. **12** (2002) 219–227

4.58 C. A. Huber, T. E. Huber, M. Sadoqi, J. A. Lubin, S. Manalis, C. B. Prater: Nanowire array composites, Science **263** (1994) 800–802

4.59 Y. Cui, L. J. Lauhon, M. S. Gudiksen, J. Wang, C. M. Lieber: Diameter-controlled synthesis of single crystal silicon nanowires, Appl. Phys. Lett. **78** (2001) 2214–2216

4.60 A. M. Morales, C. M. Lieber: A laser ablation method for the synthesis of crystalline semiconductor nanowires, Science **279** (1998) 208–211

4.61 N. Wang, Y. F. Zhang, Y. H. Tang, C. S. Lee, S. T. Lee: SiO_2-enhanced synthesis of Si nanowires by laser ablation, Appl. Phys. Lett. **73** (1998) 3902–3904

4.62 M. K. Sunkara, S. Sharma, R. Miranda, G. Lian, E. C. Dickey: Bulk synthesis of silicon nanowires using a low-temperature vapor-liquid-solid method, Appl. Phys. Lett. **79** (2001) 1546–1548

4.63 J. P. Heremans, C. M. Thrush, D. T. Morelli, M.-C. Wu: Thermoelectric power of bismuth nanocomposites, Phys. Rev. Lett. **88** (2002) 216801(1–4)

4.64 Y. Li, G. S. Cheng, L. D. Zhang: Fabrication of highly ordered ZnO nanowire arrays in anodic alumina membranes, J. Mater. Res. **15** (2000) 2305–2308

4.65 P. Yang, H. Yan, S. Mao, R. Russo, J. Johnson, R. Saykally, N. Morris, J. Pham, R. He, H.-J. Choi: Controlled growth of ZnO nanowires and their optical properties, Adv. Func. Mat. **12** (2002) 323–331

4.66 M. J. Zheng, L. D. Zhang, G. H. Li, W. Z. Chen: Fabrication and optical properties of large-scale uniform zinc oxide nanowire arrays by one-step electrochemical deposition technique, Chem. Phys. Lett. **363** (2002) 123–128

4.67 Y.-M. Lin, S. B. Cronin, J. Y. Ying, M. S. Dresselhaus, J. P. Heremans: Transport properties of Bi nanowire arrays, Appl. Phys. Lett. **76** (2000) 3944–3946

4.68 C. A. Huber, T. E. Huber: A novel microstructure: semiconductor-impregnated porous Vycor glass, J. Appl. Phys. **64** (1988) 6588–6590

4.69 J. S. Beck, J. C. Vartuli, W. J. Roth, M. E. Leonowicz, C. T. Kresge, K. D. Schmitt, C. T.-W. Chu, D. H. Olson, E. W. Sheppard, S. B. McCullen, J. B. Higgins, J. L. Schlenker: A new family of mesoporous molecular sieves prepared with liquid crystal templates, J. Am. Chem. Soc. **114** (1992) 10834–10843

4.70 C.-G. Wu, T. Bein: Conducting polyaniline filaments in a mesoporous channel host, Science **264** (1994) 1757–1759

4.71 T. Thurn-Albrecht, J. Schotter, G. A. Kästle, N. Emley, T. Shibauchi, L. Krusin-Elbaum, K. Guarini, C. T. Black, M. T. Tuominen, T. P. Russell: Ultrahigh-density nanowire arrays grown in self-assembled diblock copolymer templates, Science **290** (2000) 2126–2129

4.72 A. W. Adamson: *Physical Chemistry of Surfaces* (Wiley, New York 1982) p. 338

4.73 R. Ferré, K. Ounadjela, J. M. George, L. Piraux, S. Dubois: Magnetization processes in nickel and cobalt electrodeposited nanowires, Phys. Rev. B **56** (1997) 14066–14075

4.74 H. Zeng, M. Zheng, R. Skomski, D. J. Sellmyer, Y. Liu, L. Menon, S. Bandyopadhyay: Magnetic properties of self-assembled Co nanowires of varying length and diameter, J. Appl. Phys. **87** (2000) 4718–4720

4.75 Y. Peng, H.-L. Zhang, S.-L. Pan, H.-L. Li: Magnetic properties and magnetization reversal of α-Fe nanowires deposited on alumina film, J. Appl. Phys. **87** (2000) 7405–7408

4.76 L. Piraux, J. M. George, J. F. Despres, C. Leroy, E. Ferain, R. Legras, K. Ounadjela, A. Fert: Giant magnetoresistance in magnetic multilayered nanowires, Appl. Phys. Lett. **65** (1994) 2484–2486

4.77 S. Bhattacharrya, S. K. Saha, D. Chakravorty: Nanowire formation in a polymeric film, Appl. Phys. Lett. **76** (2000) 3896–3898

4.78 G. Yi, W. Schwarzacher: Single crystal superconductor nanowires by electrodeposition, Appl. Phys. Lett. **74** (1999) 1746–1748

4.79 D. AlMawlawi, C. Z. Liu, M. Moskovits: Nanowires formed in anodic oxide nanotemplates, J. Mater. Res. **9** (1994) 1014–1018

4.80 G. S. Cheng, L. D. Zhang, S. H. Chen, Y. Li, L. Li, X. G. Zhu, Y. Zhu, G. T. Fei, Y. Q. Mao: Ordered nanostructure of single-crystalline GaN nanowires in a honeycomb structure of anodic alumina, J. Mater. Res. **15** (2000) 347–350

4.81 D. N. Davydov, P. A. Sattari, D. AlMawlawi, A. Osika, T. L. Haslett, M. Moskovits: Field emitters based on porous aluminum oxide templates, J. Appl. Phys. **86** (1999) 3983–3987

4.82 J. Li, C. Papadopoulos, J. M. Xu, M. Moskovits: Highly-ordered carbon nanotube arrays for electronic applications, Appl. Phys. Lett. **75** (1999) 367–369

4.83 T. Iwasaki, T. Motoi, T. Den: Multiwalled carbon nanotubes growth in anodic alumina nanoholes, Appl. Phys. Lett. **75** (1999) 2044–2046

4.84 J. S. Suh, J. S. Lee: Highly ordered two-dimensional carbon nanotube arrays, Appl. Phys. Lett. **75** (1999) 2047–2049

4.85 Z. M. Li, Z. K. Tang, H. J. Liu, N. Wang, C. T. Chan, R. Saito, S. Okada, G. D. Li, J. S. Chen, N. Nagasawa, S. Tsuda: Polarized absorption spectra of single-walled 4 Å carbon nanotubes aligned in channels of an $AlPO_4$-5 single crystal, Phys. Rev. Lett. **87** (2001) 127401 (1–4)

4.86 J. Sloan, A. I. Kirkland, J. L. Hutchison, M. L. H. Green: Structural characterization of atomically regulated nanocrystals formed within single-walled carbon nanotubes using electron microscopy, Acc. Chem. Res. **35** (2002) 1054–1062

4.87 R. S. Wagner, W. C. Ellis: Vapor-liquid-solid mechanism of single crystal growth, Appl. Phys. Lett. **4** (1964) 89–90

4.88 Y. Wu, P. Yang: Direct observation of vapor-liquid-solid nanowire growth, J. Am. Chem. Soc. **123** (2001) 3165–3166

4.89 S. Sharma, M. K. Sunkara, R. Miranda, G. Lian, E. C. Dickey: A novel low temperature synthesis method for semiconductor nanowires Synthesis, Functional Properties and Applications of Nanostructures: Mat. Res. Soc. Symp. Proc., San Francisco 2001, ed. by H. W. Hahn, D. L. Feldheim, C. P. Kubiak, R. Tannenbaum, R. W. Siegel (Materials Research Society Press, Pittsburgh 2001) **676** Y1.6

4.90 Y. Wu, R. Fan, P. Yang: Block-by-block growth of single-crystalline Si/SiGe superlattice nanowires, Nano. Lett. **2** (2002) 83–86

4.91 N. Wang, Y. H. Tang, Y. F. Zhang, C. S. Lee, S. T. Lee: Nucleation and growth of Si nanowires from silicon oxide, Phys. Rev. B **58** (1998) R16024–R16026

4.92 M. S. Gudiksen, L. J. Lauhon, J. Wang, D. C. Smith, C. M. Lieber: Growth of nanowire superlattice structures for nanoscale photonics and electronics, Nature **415** (2002) 617–620

4.93 M. T. Björk, B. J. Ohlsson, T. Sass, A. I. Persson, C. Thelander, M. H. Magnusson, K. Deppert, L. R. Wallenberg, L. Samuelson: One-dimensional steeplechase for electrons realized, Nano. Lett. **2** (2002) 87–89

4.94 Y. F. Zhang, Y. H. Tang, N. Wang, C. S. Lee, I. Bello, S. T. Lee: One-dimensional growth mechanism of crystalline silicon nanowires, J. Cryst. Growth **197** (1999) 136–140

4.95 S. T. Lee, Y. F. Zhang, N. Wang, Y. H. Tang, I. Bello, C. S. Lee, Y. W. Chung: Semiconductor nanowires from oxides, J. Mater. Res. **14** (1999) 4503–4507

4.96 D. D. D. Ma, C. S. Lee, Y. Lifshitz, S. T. Lee: Periodic array of intramolecular junctions of silicon nanowires, Appl. Phys. Lett. **81** (2002) 3233–3235

4.97 B. Gates, Y. Yin, Y. Xia: A solution-phase approach to the synthesis of uniform nanowires of crystalline selenium with lateral dimensions in the range of 10 – 30 nm, J. Am. Chem. Soc. **122** (2000) 12582–12583

4.98 B. Mayers, B. Gates, Y. Yin, Y. Xia: Large-scale synthesis of monodisperse nanorods of Se/Te alloys through a homogeneous nucleation and solution growth process, Adv. Mat. **13** (2001) 1380–1384
4.99 B. Gates, Y. Wu, Y. Yin, P. Yang, Y. Xia: Single-crystalline nanowires of Ag_2Se can be synthesized by templating against nanowires of trigonal Se. J. Am. Chem. Soc., **123** (2001) 11500–11501
4.100 B. Gates, B. Mayers, Y. Wu, Y. Sun, B. Cattle, P. Yang, Y. Xia: Synthesis and characterization of crystalline Ag_2Se nanowires through a template-engaged reaction at room temperature, Adv. Funct. Mat. **12** (2002) 679–686
4.101 X. Peng, J. Wickham, A. P. Alivisatos: Kinetics of II-VI and III-V colloidal semiconductor nanocrystal growth: 'focusing' of size distributions, J. Am. Chem. Soc. **120** (1998) 5343–5344
4.102 M. P. Zach, K. Inazu, K. H. Ng, J. C. Hemminger, R. M. Penner: Synthesis of molybdenum nanowires with millimeter-scale lengths using electrochemical step edge decoration, Chem. Mater. **14** (2002) 3206–3216
4.103 N. A. Melosh, A. Boukai, F. Diana, B. Gerardot, A. Badolato, P. M. Petroff, J. R. Heath: Ultrahigh-density nanowire lattices and circuits, Science **300** (2003) 112–115
4.104 S. R. Nicewarner-Peña, R. G. Freeman, B. D. Reis, L. He, D. J. Peña, I. D. Walton, R. Cromer, C. D. Keating, M. J. Natan: Submicrometer metallic barcodes, Science **294** (2001) 137–141
4.105 L. J. Lauhon, M. S. Gudiksen, D. Wang, C. M. Lieber: Epitaxial core-shell and core-multishell nanowire heterostructures, Nature **420** (2002) 57–61
4.106 Z. L. Wang, Z. R. Dai, R. P. Gao, Z. G. Bai, J. L. Gole: Side-by-side silicon carbide-silica biaxial nanowires: Synthesis, structure, and mechanical properties, Appl. Phys. Lett. **77** (2000) 3349–3351
4.107 J. Y. Lao, J. G. Wen, Z. F. Ren: Hierarchical ZnO nanostructures, Nano. Lett. **2** (2002) 1287–1291
4.108 J. Y. Lao, J. Y. Huang, D. Z. Wang, Z. F. Ren: ZnO nanobridges and nanonails, Nano. Lett. **3** (2003) 235–238
4.109 P. Yang, F. Kim: Langmuir-Blodgett assembly of one-dimensional nanostructures, CHEMPHYSCHEM **3** (2002) 503–506
4.110 B. Messer, J. H. Song, P. Yang: Microchannel networks for nanowire patterning, J. Am. Chem. Soc. **122** (2000) 10232–10233
4.111 P. A. Smith, C. D. Nordquist, T. N. Jackson, T. S. Mayer, B. R. Martin, J. Mbindyo, T. E. Mallouk: Electric-field assisted assembly and alignment of metallic nanowires, Appl. Phys. Lett. **77** (2000) 1399–1401
4.112 H. Masuda, H. Yamada, M. Satoh, H. Asoh, M. Nakao, T. Tamamura: Highly ordered nanochannel-array architecture in anodic alumina, Appl. Phys. Lett. **71** (1997) 2770–2772
4.113 O. Rabin, P. R. Herz, S. B. Cronin, Y.-M. Lin, A. I. Akinwande, M. S. Dresselhaus: Nanofabrication using self-assembled alumina templates, Nonlithographic and Lithographic Methods for Nanofabrication: MRS Symposium Proceedings, Boston 2000, ed. by J. A. Rogers, A. Karim, L. Merhari, D. Norris, Y. Xia (Materials Research Society Press, Pittsburgh 2001) **636** D4.7 (1–6)
4.114 A. Rabin, P. R. Herz, Y.-M. Lin, A. I. Akinwande, S. B. Cronin, M. S. Dresselhaus: Formation of thick porous anodic alumina films and nanowire arrays on silicon wafers and glass, Adv. Funct. Mater. **13** (2003) 631–638
4.115 M. H. Huang, S. Mao, H. Feick, H. Yan, Y. Wu, H. Kind, E. Weber, R. Russo, P. Yang: Room-temperature ultraviolet nanowire nanolasers, Science **292** (2001) 1897–1899
4.116 Y. H. Tang, Y. F. Zhang, N. Wang, C. S. Lee, X. D. Han, I. Bello, S. T. Lee: Morphology of Si nanowires synthesized by high-temperature laser ablation, J. Appl. Phys. **85** (1999) 7981–7983
4.117 S. B. Cronin, Y.-M. Lin, O. Rabin, M. R. Black, G. Dresselhaus, M. S. Dresselhaus, P. L. Gai: Bismuth nanowires for potential applications in nanoscale electronics technology, Microsc. Microanal. **8** (2002) 58–63
4.118 M. S. Sander, R. Gronsky, Y.-M. Lin, M. S. Dresselhaus: Plasmon excitation modes in nanowire arrays, J. Appl. Phys. **89** (2001) 2733–2736
4.119 L. Venkataraman, C. M. Lieber: Molybdenum selenide molecular wires as one-dimensional conductors, Phys. Rev. Lett. **83** (1999) 5334–5337
4.120 A. Majumdar: Scanning thermal microscopy, Annual Rev. Mater. Sci. **29** (1999) 505–585
4.121 D. A. Wharam, T. J. Thornton, R. Newbury, M. Pepper, H. Ahmed, J. E. F. Frost, D. G. Hasko, D. C. Peacock, D. A. Ritchie, G. A. C. Jones: One-dimensional transport and the quantization of the ballistic resistance, J. Phys. C: Solid State Phys. **21** (1988) L209–L214
4.122 B. J. van Wees, H. van Houten, C. W. J. Beenakker, J. G. Williamson, L. P. Kouvenhoven, D. van der Marel, C. T. Foxon: Quantized conductance of point contacts in a two-dimensional electron gas, Phys. Rev. Lett. **60** (1988) 848–850
4.123 C. J. Muller, J. M. van Ruitenbeek, L. J. de Jongh: Conductance and supercurrent discontinuities in atomic-scale metallic constrictions of variable width, Phys. Rev. Lett. **69** (1992) 140–143
4.124 C. J. Muller, J. M. Krans, T. N. Todorov, M. A. Reed: Quantization effects in the conductance of metallic contacts at room temperature, Phys. Rev. B **53** (1996) 1022–1025
4.125 J. L. Costa-Krämer, N. Garcia, H. Olin: Conductance quantization in bismuth nanowires at 4 K, Phys. Rev. Lett. **78** (1997) 4990–4993
4.126 J. L. Costa-Krämer, N. Garcia, H. Olin: Conductance quantization histograms of gold nanowires at 4 K, Phys. Rev. B **55** (1997) 12910–12913
4.127 C. Z. Li, H. X. He, A. Bogozi, J. S. Bunch, N. J. Tao: Molecular detection based on conductance quan-

tization of nanowires, Appl. Phys. Lett. **76** (2000) 1333–1335

4.128 J. L. Costa-Krämer, N. Garcia, P. Garcia-Mochales, P. A. Serena, M. I. Marques, A. Correia: Conductance quantization in nanowires formed between micro and macroscopic metallic electrodes, Phys. Rev. B **55** (1997) 5416–5424

4.129 Y. Huang, X. Duan, Y. Cui, L. J. Lauhon, K.-H. Kim, C. Lieber: Logic gates and computation from assembled nanowire building blocks, Science **294** (2001) 1313–1317

4.130 J.-R. Kim, H. Oh, H. M. So, J.-J. Kim, J. Kim, C. J. Lee, S. C. Lyu: Schottky diodes based on a single GaN nanowire, Nanotechnology **13** (2002) 701–704

4.131 X. Duan, Y. Huang, C. M. Lieber: Nonvolatile memory and programmable logic from molecule-gated nanowires, Nano. Lett. **2** (2002) 487–490

4.132 E. C. Walter, R. M. Penner, H. Liu, K. H. Ng, M. P. Zach, F. Favier: Sensors from electrodeposited metal nanowires, Surf. Inter. Anal. **34** (2002) 409–412

4.133 E. C. Walter, K. H. Ng, M. P. Zach, R. M. Penner, F. Favier: Electronic devices from electrodeposited metal nanowires, Microelectron. Eng. **61–62** (2002) 555–561

4.134 Y.-M. Lin, X. Sun, M. S. Dresselhaus: Theoretical investigation of thermoelectric transport properties of cylindrical Bi nanowires, Phys. Rev. B **62** (2000) 4610–4623

4.135 K. Liu, C. L. Chien, P. C. Searson: Finite-size effects in bismuth nanowires, Phys. Rev. B **58** (1998) R14681–R14684

4.136 Z. Zhang, X. Sun, M. S. Dresselhaus, J. Y. Ying, J. Heremans: Magnetotransport investigations of ultrafine single-crystalline bismuth nanowire arrays, Appl. Phys. Lett. **73** (1998) 1589–1591

4.137 J. Heremans, C. M. Thrush, Z. Zhang, X. Sun, M. S. Dresselhaus, J. Y. Ying, D. T. Morelli: Magnetoresistance of bismuth nanowire arrays: A possible transition from one-dimensional to three dimensional localization, Phys. Rev. B **58** (1998) R10091–R10095

4.138 L. Sun, P. C. Searson, C. L. Chien: Finite-size effects in nickel nanowire arrays, Phys. Rev. B **61** (2000) R6463–R6466

4.139 Y.-M. Lin, O. Rabin, S. B. Cronin, J. Y. Ying, M. S. Dresselhaus: Semimetal-semiconductor transition in $Bi_{1-x}Sb_x$ alloy nanowires and their thermoelectric properties, Appl. Phys. Lett. **81** (2002) 2403–2405

4.140 Y.-M. Lin, S. B. Cronin, O. Rabin, J. Y. Ying, M. S. Dresselhaus: Transport properties and observation of semimetal-semiconductor transition in bi-based nanowires, Nanocrystalline Semiconductor Materials and Devices: MRS Symp. Proc., Boston 2002, ed. by J. M. Buriak, D. D. M. Wayner, L. Tsyeskov, F. Priolo, B. E. White, Boston 2003) **737** F3.14

4.141 Y.-M. Lin, M. S. Dresselhaus: Transport properties of superlattice nanowires and their potential for thermoelectric applications, Nanocrystalline Semiconductor Materials and Devices: MRS Symp. Proc., Boston 2002, ed. by J. M. Buriak, D. D. M. Wayner, L. Tsyeskov, F. Priolo, B. E. White (Materials Research Society Press, Boston 2003) **737** F8.18

4.142 J. Heremans, C. M. Thrush, Y.-M. Lin, S. B. Cronin, M. S. Dresselhaus: Transport properties of antimony nanowires, Phys. Rev. B **63** (2001) 085406(1–8)

4.143 Y.-M. Lin, S. B. Cronin, O. Rabin, J. Y. Ying, M. S. Dresselhaus: Transport properties of $Bi_{1-x}Sb_x$ alloy nanowires synthesized by pressure injection, Appl. Phys. Lett. **79** (2001) 677–679

4.144 D. E. Beutler, N. Giordano: Localization and electron–electron interaction effects in thin Bi wires and films, Phys. Rev. B **38** (1988) 8–19

4.145 Z. Zhang, X. Sun, M. S. Dresselhaus, J. Y. Ying, J. Heremans: Electronic transport properties of single crystal bismuth nanowire arrays, Phys. Rev. B **61** (2000) 4850–4861

4.146 J. Heremans, C. M. Thrush: Thermoelectric power of bismuth nanowires, Phys. Rev. B **59** (1999) 12579–12583

4.147 L. D. Hicks, M. S. Dresselhaus: Thermoelectric figure of merit of a one-dimensional conductor, Phys. Rev. B **47** (1993) 16631–16634

4.148 Y.-M. Lin, S. B. Cronin, O. Rabin, J. Heremans, M. S. Dresselhaus, J. Y. Ying: Transport properties of Bi-related nanowire systems, Anisotropic Nanoparticles: Synthesis, Characterization and Applications: MRS Symp. Proc., Boston 2000, ed. by S. Stranick, P. C. Searson, L. A. Lyon, C. Keating (Materials Research Society Press, Pittsburgh 2001) **635** C4301–C4306

4.149 Y.-M. Lin, M. S. Dresselhaus: Thermoelectric properties of superlattice nanowires, Phys. Rev. B **68** (2003) 075304 (1–14)

4.150 M. T. Björk, B. J. Ohlsson, C. Thelander, A. I. Persson, K. Deppert, L. R. Wallenberg, L. Samuelson: Nanowire resonant tunneling diodes, Appl. Phys. Lett. **81** (2002) 4458–4460

4.151 D. Li: Thermal transport in individual nanowires and nanotubes. Ph.D. Thesis (Univ. California, Berkeley 2002)

4.152 T. S. Tighe, J. M. Worlock, M. L. Roukes: Direct thermal conductance measurements on suspended monocrystalline nanostructures, Appl. Phys. Lett. **70** (1997) 2687–2689

4.153 S. T. Huxtable, A. R. Abramson, C.-L. Tien, A. Majumdar, C. LaBounty, X. Fan, G. Zeng, J. E. Bowers, A. Shakouri, E. T. Croke: Thermal conductivity of Si/SiGe and SiGe/SiGe superlattices, Appl. Phys. Lett. **80** (2002) 1737–1739

4.154 R. Venkatasubramanian, E. Siivola, T. Colpitts, B. O'Quinn: Thin-film thermoelectric devices with high room-temperature figures of merit, Nature **413** (2001) 597–602

4.155 C. Dames, G. Chen: Modeling the thermal conductivity of a SiGe segmented nanowire, 21st Int. Conf. Thermoelectrics: Proc. ICT Symposium, Long Beach 2002 (IEEE, Piscataway 2002) 317–320

4.156 G. Chen, M.S. Dresselhaus, G. Dresselhaus, J.-P. Fluerial, T. Caillat: Recent developments in thermoelectric materials, Int. Mater. Rev. **48** (2003) 45–66

4.157 K. Schwab, E.A. Henriksen, J.M. Worlock, M.L. Roukes: Measurement of the quantum of thermal conductance, Nature **404** (2000) 974–977

4.158 K. Schwab, J.L. Arlett, J.M. Worlock, M.L. Roukes: Thermal conductance through discrete quantum channels, Physica E **9** (2001) 60–68

4.159 M. Cardona: *Light scattering in solids* (Springer, Berlin, Heidelberg 1982)

4.160 P.Y. Yu, M. Cardona: *Fundamentals of Semiconductors* (Springer, Berlin, Heidelberg 1995) Chap. 7

4.161 J.C.M. Garnett: Colours in metal glasses, in metallic films, and in metallic solutions, Philos. Trans. Roy. Soc. London A **205** (1906) 237–288

4.162 D.E. Aspnes: Optical properties of thin films, Thin Solid Films **89** (1982) 249–262

4.163 U. Kreibig, L. Genzel: Optical absorption of small metallic particles, Surf. Sci. **156** (1985) 678–700

4.164 M.R. Black, Y.-M. Lin, S.B. Cronin, O. Rabin, M.S. Dresselhaus: Infrared absorption in bismuth nanowires resulting from quantum confinement, Phys. Rev. B **65** (2002) 195417(1–9)

4.165 M.W. Lee, H.Z. Twu, C.-C. Chen, C.-H. Chen: Optical characterization of wurtzite gallium nitride nanowires, Appl. Phys. Lett. **79** (2001) 3693–3695

4.166 D.M. Lyons, K.M. Ryan, M.A. Morris, J.D. Holmes: Tailoring the optical properties of silicon nanowire arrays through strain, Nano. Lett. **2** (2002) 811–816

4.167 J.C. Johnson, H. Yan, R.D. Schaller, L.H. Haber, R.J. Saykally, P. Yang: Single nanowire lasers, J. Phys. Chem. B **105** (2001) 11387–11390

4.168 S. Blom, L.Y. Gorelik, M. Jonson, R.I. Shekhter, A.G. Scherbakov, E.N. Bogachek, U. Landman: Magneto-optics of electronic transport in nanowires, Phys. Rev. B **58** (1998) 16305–16314

4.169 J.P. Pierce, E.W. Plummer, J. Shen: Ferromagnetism in cobalt-iron alloy nanowire arrays on W(110), Appl. Phys. Lett. **81** (2002) 1890–1892

4.170 J.C. Johnson, H. Yan, R.D. Schaller, P.B. Petersen, P. Yang, R.J. Saykally: Near-field imaging of nonlinear optical mixing in single zinc oxide nanowires, Nano. Lett. **2** (2002) 279–283

4.171 M.R. Black, Y.-M. Lin, S.B. Cronin, M.S. Dresselhaus: Using optical measurements to improve electronic models of bismuth nanowires, 21st Int. Conf. Thermoelectrics: Proc. ICT Symposium, Long Beach 2002 (IEEE, Piscataway 2002) 253–256

4.172 H.-L. Liu, C.-C. Chen, C.-T. Chia, C.-C. Yeh, C.-H. Chen, M.-Y. Yu, S. Keller, S.P. DenBaars: Infrared and Raman-scattering studies in single-crystalline GaN nanowires, Chem. Phys. Letts. **345** (2001) 245–251

4.173 H. Richter, Z.P. Wang, L. Ley: The one phonon Raman-spectrum in microcrystalline silicon, Solid State Commun. **39** (1981) 625–629

4.174 I.H. Campbell, P.M. Fauchet: The effects of microcrystal size and shape on the one phonon Raman-spectra of crystalline semiconductors, Solid State Commun. **58** (1986) 739–741

4.175 R. Gupta, Q. Xiong, C.K. Adu, U.J. Kim, P.C. Eklund and: Laser-induced Fanoresonance scattering in silicon nanowires, Nano. Lett. **3** (2003) 627–631

4.176 Y. Cui, C.M. Lieber: Functional nanoscale electronic devices assembled using silicon nanowire building blocks, Science **291** (2001) 851–853

4.177 S.-W. Chung, J.-Y. Yu, J.R. Heath: Silicon nanowire devices, Appl. Phys. Lett. **76** (2000) 2068–2070

4.178 M. Ding, H. Kim, A.I. Akinwande: Observation of valence band electron emission from n-type silicon field emitter arrays, Appl. Phys. Lett. **75** (1999) 823–825

4.179 F.C.K. Au, K.W. Wong, Y.H. Tang, Y.F. Zhang, I. Bello, S.T. Lee: Electron field emission from silicon nanowires, Appl. Phys. Lett. **75** (1999) 1700–1702

4.180 P.M. Ajayan, O.Z. Zhou: Applications of carbon nanotubes. In: *Carbon Nanotubes: Synthesis, Structure, Properties and Applications*, Vol. 80, ed. by M.S. Dresselhaus, G. Dresselhaus, P. Avouris (Springer, Berlin, Heidelberg 2001) pp. 391–425

4.181 G. Dresselhaus, M.S. Dresselhaus, Z. Zhang, X. Sun, J. Ying, G. Chen: Modeling thermoelectric behavior in Bi nano-wires, Seventeenth Int. Conf. Thermoelectrics: Proc. ICT'98, Nagoya 1998, ed. by K. Koumoto (Institute of Electrical and Electronics Engineers Inc, Piscataway 1998) 43–46

4.182 L.D. Hicks, M.S. Dresselhaus: The effect of quantum well structures on the thermoelectric figure of merit, Phys. Rev. B **47** (1993) 12727–12731

4.183 O. Rabin, Y.-M. Lin, M.S. Dresselhaus: Anomalously high thermoelectric figure of merit in $Bi_{1-x}Sb_x$ nanowires by carrier pocket alignment, Appl. Phys. Lett. **79** (2001) 81–83

4.184 T.C. Harman, P.J. Taylor, M.P. Walsh, B.E. LaForge: Quantum dot supertattice thermoelectric materials and devices, Science **297** (2002) 2229–2232

4.185 V. Dneprovskii, E. Zhukov, V. Karavanskii, V. Poborchii, I. Salamatini: Nonlinear optical properties of semiconductor quantum wires, Superlattices Microstruct. **23** (1998) 1217–1221

4.186 J.C. Johnson, H.-J. Choi, K.P. Knutsen, R.D. Schaller, P. Yang, R.J. Saykally: Single gallium nitride nanowire lasers, Nature Mater. **1** (2002) 106–110

4.187 H. Kind, H. Yan, B. Messer, M. Law, P. Yang: Nanowire ultraviolet photodetectors and optical switches, Adv. Mat. **14** (2002) 158–160

4.188 B. M. I. van der Zande, M. Böhmer, L. G. J. Fokkink, C. Schönenberger: Colloidal dispersions of gold rods: Synthesis and optical properties, Langmuir **16** (2000) 451–458

4.189 B. M. I. van der Zande, G. J. M. Koper, H. N. W. Lekkerkerker: Alignment of rod-shaped gold particles by electric fields, J. Phys. Chem. B **103** (1999) 5754–5760

4.190 W. U. Huynh, J. J. Dittmer, A. P. Alivisatos: Hybrid nanorod-polymer solar cells, Science **295** (2002) 2425–2427

4.191 Y. Cui, Q. Wei, H. Park, C. Lieber: Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species, Science **293** (2001) 1289–1292

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

5. Introduction to Micro/Nanofabrication

In this chapter, we discuss various *micro/nanofabrication* techniques used to manufacture structures in a wide range of dimensions (mm–nm). Starting with some of the most common microfabrication techniques (lithography, deposition, and etching), we present an array of *micromachining* and *MEMS* technologies that can be used to fabricate microstructures down to ~1 μm. These techniques have attained an adequate level of maturity to allow for a variety of MEMS-based commercial products (pressure sensors, accelerometers, gyroscopes, etc.). More recently, nanometer-sized structures have attracted an enormous amount of interest. This is mainly due to their unique electrical, magnetic, optical, thermal, and mechanical properties. These could lead to a variety of electronic, photonic, and sensing devices with a superior performance compared to their macro counterparts. Subsequent to our discussion on MEMS and micromachining, we present several important *nanofabrication* techniques currently under intense investigation. Although e-beam and other high resolution lithographies can be used to fabricate nanometer-sized structures, their serial nature and/or cost preclude their widespread application. This has forced investigators to explore alternative and potentially superior techniques such as *strain engineering*, *self-assembly*, and *nano-imprint* lithography. Among these, self-assembly is the most promising method, due to its low cost and the ability to produce nanostructures at different length scales.

5.1 Basic Microfabrication Techniques 148
  5.1.1 Lithography 148
  5.1.2 Thin Film Deposition and Doping ... 149
  5.1.3 Etching and Substrate Removal 153
  5.1.4 Substrate Bonding 157

5.2 MEMS Fabrication Techniques 159
  5.2.1 Bulk Micromachining 159
  5.2.2 Surface Micromachining 163
  5.2.3 High-Aspect-Ratio Micromachining 166

5.3 Nanofabrication Techniques 170
  5.3.1 E-Beam and Nano-Imprint Fabrication 171
  5.3.2 Epitaxy and Strain Engineering 172
  5.3.3 Scanned Probe Techniques 173
  5.3.4 Self-Assembly and Template Manufacturing 176

References 180

Recent innovations in the area of *micro/nanofabrication* have created a unique opportunity for manufacturing structures in the nanometer–millimeter range. The available six orders of magnitude dimensional span can be used to fabricate novel electronic, optical, magnetic, mechanical, and chemical/biological devices with applications ranging from sensors to computation and control. In this chapter, we will introduce major micro/nanofabrication techniques currently used to fabricate structures in the nanometer to several hundred micrometer range. We will focus mainly on the most important and widely used techniques and will not discuss specialized methods. After a brief introduction to basic microfabrication, we will discuss *MEMS*-fabrication techniques used to build microstructures down to about 1 μm in dimensions. Following this, we will discuss several major top-down and bottom-up nanofabrication methods that have shown great promise in manufacturing *nanostructures* (dimensions $< 1\,\mu m$).

5.1 Basic Microfabrication Techniques

Most micro/nanofabrication techniques have their roots in the standard fabrication methods developed for the semiconductor industry [5.1–3]. Therefore, a clear understanding of these techniques is necessary for anyone embarking on a research and development path in the micro/nano area. In this section, we will discuss major microfabrication techniques used most frequently in the manufacturing of micro/nanostructures. Some of these techniques such as thin-film deposition and etching are common between the micro/nano and VLSI microchip fabrication disciplines. However, several other techniques that are more specific to the micro/nanofabrication area will also be discussed in this section.

5.1.1 Lithography

Lithography is the technique used to transfer a computer generated pattern onto a substrate (silicon, glass, GaAs, etc.). This pattern is subsequently used to etch an underlying thin film (oxide, nitride, etc.) for various purposes (doping, etching, etc.). Although photolithography, i. e., the lithography using a UV light source, is by far the most widely used lithography technique in microelectronic fabrication, electron-beam (e-beam) and X-ray lithography are two other alternatives that have attracted considerable attention in the MEMS and nanofabrication areas. We will discuss photolithography in this section and postpone the discussion of e-beam and X-ray techniques to the subsequent sections dealing with MEMS and nanofabrication.

The starting point following the creation of the computer layout for a specific fabrication sequence is the generation of a photomask. This involves a sequence of photographic processes (using optical or e-beam pattern generators) that results in a glass plate having the desired pattern in the form of a thin (∼ 100 nm) chromium layer. Following the generation of photomask, the lithography process can proceed as shown in Fig. 5.1. This sequence demonstrates the pattern transfer onto a substrate coated with silicon dioxide. However, the same technique is applicable to other materials. After depositing the desired material on the substrate, the photolithography process starts with spin coating the substrate with a photoresist. This is a polymeric photosensitive material that can be spun onto the wafer in liquid form (usually an adhesion promoter such as hexamethyldisilazane HMDS is used prior to the application of the resist). The spinning speed and photoresist viscosity will determine the final resist thickness, which is typically between 0.5–2.5 µm. Two different kinds of photoresist are available: positive and negative. In the positive resist, the UV-exposed areas will be dissolved in the subsequent development stage, whereas in the negative photoresist, the exposed areas will remain intact after the development. Due to its better performance with regard to the process control in small geometries, the positive resist is the most extensively used photoresist in the VLSI processes. After spinning the photoresist on the wafer, the substrate is soft baked (5–30 min at 60–100 °C) in order to remove the solvents from the resist and improve the adhesion. Subsequently, the mask is aligned to the wafer and the photoresist is exposed to a UV source.

Depending on the separation between the mask and the wafer, three different exposure systems are available: 1) contact, 2) proximity, and 3) projection. Although contact printing gives better resolution compared to the proximity technique, the constant contact of the mask with photoresist reduces the process yield and can damage the mask. Projection printing uses a dual-lens optical system to project the mask image onto the wafer. Since only one die at a time can be exposed, this requires a step-and-repeat system to totally cover the wafer area. Projection printing is by far the most widely used system in microfabrication and can yield superior resolutions compared to contact and proximity methods. The exposure source for photolithography depends on the resolution. Above 0.25 µm minimum line width, high pressure mercury lamp is adequate (436 nm g-line and 365 nm i-line). However, between 0.25 and 0.13 µm, deep UV sources such as excimer lasers (248 nm KrF and 193 nm ArF) are required. Although there has been extensive competition for the below 0.13 µm regime (including e-beam and X-ray), extreme UV (EUV) with

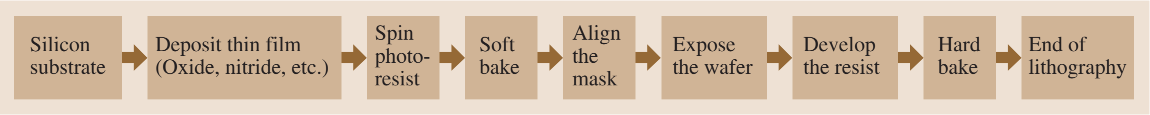

Fig. 5.1 Lithography process flow

a wavelength of 10–14 nm seems to be the preferred technique [5.4].

After exposure, the photoresist is developed in a process similar to the development of photographic film. The resist is subsequently hard baked (20–30 min at 120–180 °C) in order to further improve the adhesion. The hard bake step concludes the photolithography sequence by creating the desired pattern on the wafer. Next, the underlying thin film is etched, and the photoresist is stripped in acetone or other organic removal solvents. Figure 5.2 shows schematic drawing of photolithography steps with a positive photoresist.

5.1.2 Thin Film Deposition and Doping

Thin-film deposition and doping are used extensively in micro/nanofabrication technologies. Most of the fabricated micro/nanostructures contain materials other than that of the substrate, which are obtained by various deposition techniques, or by modification of the substrate. The following is a list of a few typical applications for the deposited and/or doped materials used in micro/nanofabrication that gives an idea of the required properties:

- Mechanical structure
- Electrical isolation
- Electrical connection
- Sensing or actuating
- Mask for etching and doping
- Support or mold during deposition of other materials (sacrificial materials)
- Passivation

Most of the deposited thin films have different properties than those of their corresponding "bulk" forms (for example, metals show higher resistivities as thin films). In addition, the techniques utilized to deposit these materials have a great impact on their final properties. For instance, internal stress (compressive or tensile) in a film is strongly process-dependent. Excessive stress may crack or detach the film from the substrate and therefore must be minimized, although it may also be useful for certain applications. Adhesion is another important issue that needs to be taken into account when depositing thin films. In some cases such as the deposition of noble metals (e.g., gold) an intermediate layer (chromium or titanium) may be needed to improve the adhesion. Finally, step coverage and conformality are two properties that can also influence the choice of deposition technique. Figure 5.3 illustrates these concepts.

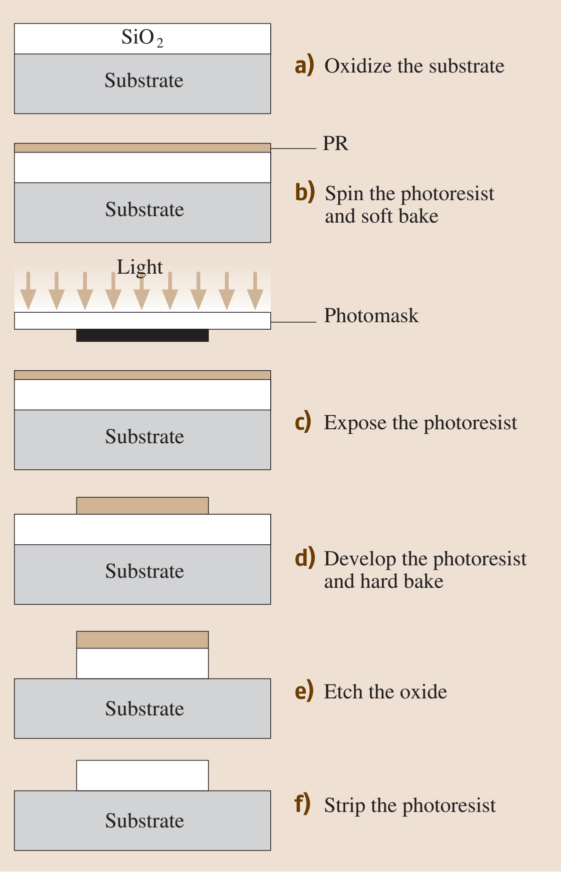

Fig. 5.2 Schematic drawing of the photolithographic steps with a positive photoresist (PR)

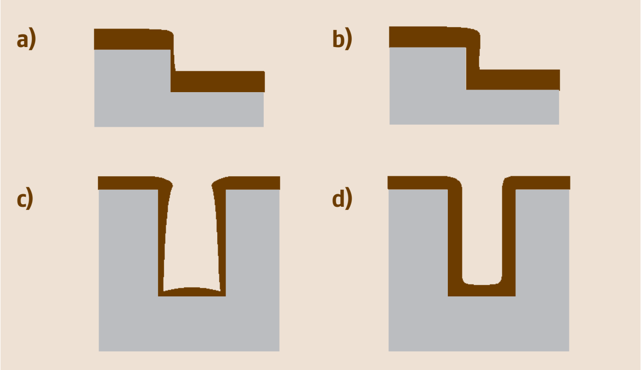

Fig. 5.3a–d Step coverage and conformality: **(a)** poor step coverage, **(b)** good step coverage, **(c)** nonconformal layer, and **(d)** conformal layer

Oxidation

Oxidation of silicon is a process used to obtain a thin film of SiO_2 of excellent quality (very low density of defects) and thickness homogeneity. Although it is not properly a deposition, the result is the same, i. e., a thin layer of a new material covering the surface is produced. The oxidation process is typically carried out at temperatures in the range of 900 °C to 1,200 °C in the presence of O_2 (dry oxidation) or H_2O (wet oxidation). The reactions for oxide formation are:

$$\mathrm{Si}_{\mathrm{(solid)}} + \mathrm{O}_{2\mathrm{(gas)}} \Rightarrow \mathrm{SiO}_{2\mathrm{(solid)}}$$

and

$$\mathrm{Si}_{\mathrm{(solid)}} + 2\mathrm{H}_2\mathrm{O}_{\mathrm{(steam)}} \Rightarrow \mathrm{SiO}_{2\mathrm{(solid)}} + 2\mathrm{H}_{2\mathrm{(gas)}} \,.$$

Although the rate of the oxide growth is higher for wet oxidation, it is achieved at the expense of a lower oxide quality (density). Since silicon atoms from the substrate participate in the reaction, the substrate is consumed as the oxide grows ($\sim 44\%$ of the total thickness lies above the line of the original silicon surface). The oxidation of silicon also occurs at room temperature, however, a layer of about 20 Å (native oxide) is enough to passivate the surface and prevent further oxidation. To grow thicker oxides, wafers are introduced into an electric resistance furnace such as the one represented in Fig. 5.4. Tens of wafers can be processed in a single batch in such equipment. By strictly controlling the timing, temperature, and gas flow entering the quartz tube, the desired thickness can be achieved with a high accuracy. Thicknesses ranging from a few tens of Angstroms to 2 µm can be obtained in reasonable times. Despite the good quality of the SiO_2 obtained by silicon oxidation (also called *thermal oxide*), the use of this process is often limited to the early stages of fabrication, since some of the materials added during the formation of structures may not withstand the high temperatures. The contamination of the furnace when the substrates have been previously in contact with certain etchants such as KOH, or when materials such as metals have been deposited also pose limitations in most of the cases.

Doping

The introduction of certain impurities in a semiconductor can change its electrical, chemical, and even mechanical properties. Typical impurities, or *dopants*, used in silicon include boron (to form p-type regions) and phosphorous or arsenic (to form n-type regions). Doping is the main process used in the microelectronic industry to fabricate major components such as diodes and transistors. In micro/nanofabrication technologies doping has additional applications such as the formation of piezoresistors for mechanical transducers and the creation of etch stop layers. Two different techniques are used to introduce the impurities into a semiconductor substrate: diffusion and ion implantation.

Diffusion is the process that became dominant in the initial years following the invention of the integrated circuit to form n- and p-type regions in the silicon. The diffusion of impurities into the silicon occurs only at high temperatures (above 800 °C). Furnaces used to carry out this process are similar to the oxidation ones. The dopants are introduced into the furnace's gaseous atmosphere from liquid or solid sources. Figure 5.5 illustrates the process of creating an n-type region by diffusion of phosphor from the surface into a p-type substrate. A masking material is previously deposited and patterned on the surface to define the areas to be doped. However, because diffusion is an isotropic process, the doped area will also extend underneath the mask. In microfabrication, diffusion is mainly used for the formation of very highly doped boron regions (p^{++}), which are usually used as an etch stop in bulk *micromachining*.

Ion implantation allows a more precise control of the dose (total amount of impurities introduced per area unit)

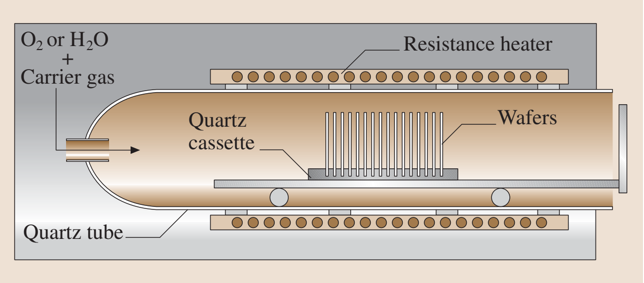

Fig. 5.4 Schematic representation of a typical oxidation furnace

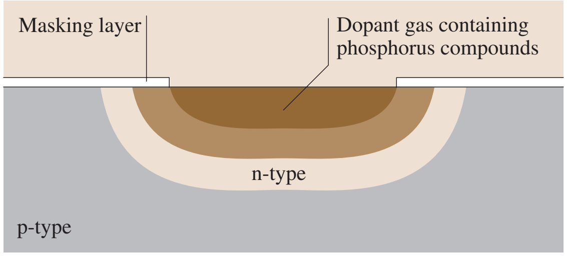

Fig. 5.5 Formation of an n-type region on a p-type silicon substrate by diffusion of phosphorous

and the impurity profile (concentration versus depth). In ion implantation, the impurities are ionized and accelerated toward the semiconductor surface. The penetration of impurities in the material follows a Gaussian distribution. After implantation, an annealing process is needed to activate the impurities and repair the damage in the crystal structure produced by the ion collisions. A *drive-in* process to redistribute the impurities performed in a standard furnace such as the ones used for oxidation or diffusion may be required as well.

Chemical Vapor Deposition and Epitaxy

As its name suggests, chemical vapor deposition (CVD) includes all the deposition techniques using the reaction of chemicals in a gas phase to form the deposited thin film. The energy needed for the chemical reaction to occur is usually supplied by maintaining the substrate at elevated temperatures. Other alternative energy sources such as plasma or optical excitation are also used, with the advantage of requiring a lower temperature at the substrate. The most common CVD processes in microfabrication are LPCVD (low pressure CVD) and PECVD (plasma enhanced CVD).

The LPCVD process is typically carried out in electrically heated tubes, similar to the oxidation tubes, equipped with pumping capabilities to achieve the needed low pressures (0.1 to 1.0 torr). A large number of wafers can be processed simultaneously, and the material is deposited in both sides of the wafers. The process temperatures depend on the material to be deposited, but generally are in the range of 550 to 900 °C. As in the oxidation, high temperatures and contamination issues can restrict the type of processes used prior to the LPCVD. Typical materials deposited by LPCVD include silicon oxide (e.g., $SiCl_2H_2 + 2N_2O \Rightarrow SiO_2 + 2N_2 + 2HCl$ at 900 °C), silicon nitride (e.g., $3SiH_4 + 4NH_3 \Rightarrow Si_3N_4 + 12H_2$ at 700–900 °C), and polysilicon (e.g., $SiH_4 \Rightarrow Si + 2H_2$ at 600 °C). Due to its faster etch rate in HF, in situ phosphorous-doped LPCVD oxide (phosphosilicate glass or PSG) is used extensively as the sacrificial layer in surface micromachining. Conformality in this process is excellent, even for very high-aspect-ratio structures. Mechanical properties of LPCVD materials are good compared to others such as PECVD, and LPCVD materials often used as structural materials in microfabricated devices. Stress in the deposited layers depends on the material, deposition conditions, and subsequent thermal history (e.g., post-deposition anneal). Typical values are: 100–300 MPa (compressive) for oxide, ∼ 1 GPa (tensile) for stoichiometric nitride, and ∼ 200–300 MPa (tensile) for polysilicon. The stress in nitride layers can be reduced to nearly zero by using a silicon-rich composition. Since the stress values can vary over a wide range, one has to measure and characterize the internal stress of deposited thin films for any specific equipment and deposition conditions.

The PECVD process is performed in plasma systems such as the one represented in Fig. 5.6. The use of RF energy to create highly reactive species in the plasma allows for the use of lower temperatures at the substrates (150 to 350 °C). Parallel-plate plasma reactors normally used in microfabrication can only process a limited number of wafers per batch. The wafers are positioned horizontally on top of the lower electrode, so only one side gets deposited. Typical materials deposited with PECVD include silicon oxide, nitride, and amorphous silicon. Conformality is good for low-aspect-ratio structures, but becomes very poor for deep trenches (20% of the surface thickness inside through-wafer holes with aspect ratio of ten). Stress depends on deposition parameters and can be either compressive or tensile. PECVD nitrides are typically non-stoichiometric (Si_xN_y) and are much less resistant to etchants in masking applications.

Another interesting type of CVD is epitaxial growth. In this process, a single crystalline material is grown as an extension of the crystal structure of the substrate. It is possible to grow dissimilar materials if the crystal structures are somehow similar (lattice-matched). Silicon-on-sapphire (SOS) substrates and some heterostructures are fabricated in this way. However, most

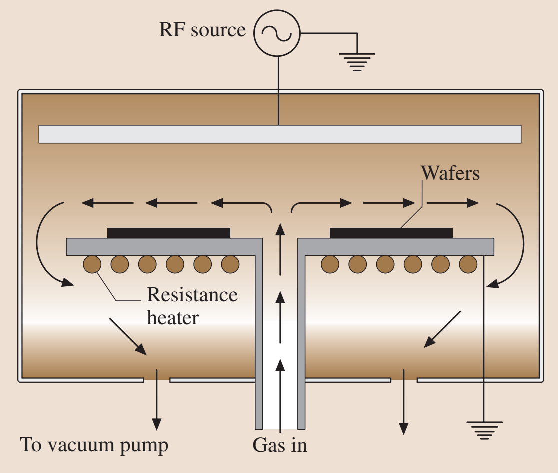

Fig. 5.6 Schematic representation of a typical PECVD system

common in microfabrication is the growth of silicon on another silicon substrate. Of particular interest for the formation of microstructures is the selective epitaxial growth. In this process the silicon crystal is allowed to grow only in windows patterned on a masking material. Many CVD techniques have been used to produce epitaxial growth. The most common for silicon is thermal chemical vapor deposition or *vapor phase epitaxy* (VPE). *Metallorganic chemical vapor deposition* (MOCVD) and *molecular beam epitaxy* (MBE) are the most common for growing high quality III-V compound layers with nearly atomic abrupt interfaces. The former uses vapors of organic compounds with group III atoms such as trimethylgallium [$Ga(CH_3)_3$] and group V hydrides such as AsH_3 in a CVD chamber with fast gas switching capabilities. The latter typically uses molecular beams from thermally evaporated elemental sources aiming at the substrate in an ultrahigh vacuum chamber. In this case, rapid on/off control of the beams is achieved by using shutters in front of the sources. Finally, it should be mentioned that many metals (molybdenum, tantalum, titanium, and tungsten) can also be deposited using LPCVD. These are attractive for their low resistivities and their ability to form silicides with silicon. Due to its application in the new interconnect technologies, copper CVD is an active area of research.

Physical Vapor Deposition (Evaporation and Sputtering)

In physical deposition systems the material to be deposited is transported from a source to the wafers, both being in the same chamber. Two physical principles are used to do so: evaporation and sputtering.

In evaporation, the source is placed in a small container with tapered walls, called a crucible, and heated up to a temperature at which evaporation occurs. Various techniques are utilized to reach the high temperatures needed, including the induction of high currents with coils wound around the crucible and the bombardment of the material surface with an electron beam (e-beam evaporators). This process is mainly used to deposit metals, although dielectrics can also be evaporated. In a typical system, the crucible is located at the bottom of a vacuum chamber, whereas the wafers are placed lining the dome-shaped ceiling of the chamber, Fig. 5.7. The main characteristic of this process is very poor step coverage, including shadow effects, as illustrated in Fig. 5.8. As will be explained in subsequent sections, some microfabrication techniques utilize these effects to pattern the deposited layer. One way to improve the step coverage is by rotating and/or heating the wafers during the deposition.

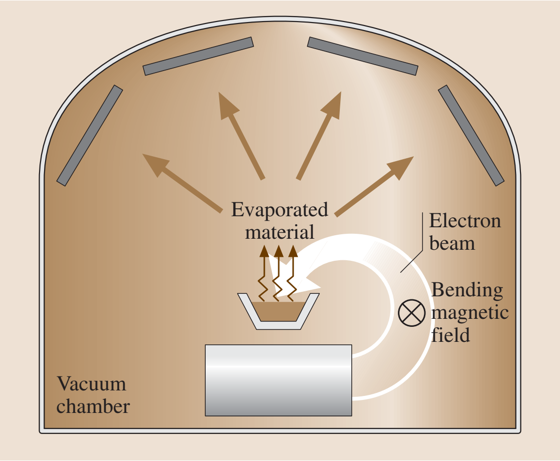

Fig. 5.7 Schematic representation of an electron-beam deposition system

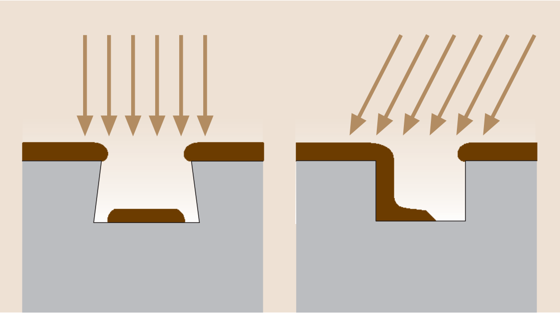

Fig. 5.8 Shadow effects observed in evaporated films. *Arrows* show the trajectory of the material atoms being deposited

In sputtering, a target of the material to be deposited is bombarded with high-energy inert ions (usually argon). The outcome of the bombardment is that individual atoms or clusters are removed from the surface and ejected toward the wafer. The physical nature of this process allows its use with virtually any existing material. Examples of interesting materials for microfabrication that are frequently sputtered include metals, dielectrics, alloys (such as shape memory alloys), and all kinds of compounds (for example, piezoelectric PZT). The inert ions bombarding the target are produced in DC or RF plasma. In a simple parallel-plate system, the top electrode is the target and the wafers are placed horizontally on top of the bottom electrode. In spite of its lower deposition rate, step coverage in sputtering is much better than in evaporation. Yet, the films obtained

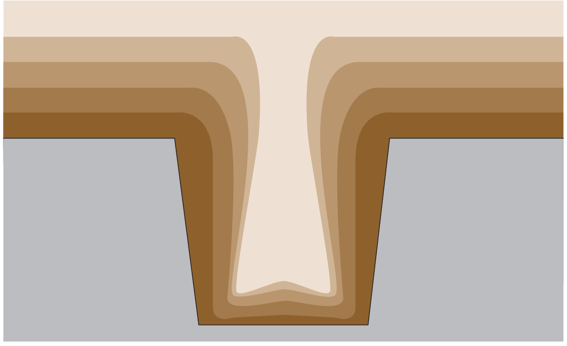

Fig. 5.9 Typical cross section evolution of a trench while being filled with sputter deposition

with this deposition process are nonconformal. Figure 5.9 illustrates successive sputtering profiles in a trench.

Both evaporation and sputtering systems are often capable of depositing more than one material simultaneously or sequentially. This capability is very useful in obtaining alloys and multilayers (e.g., multilayer magnetic recording heads are sputtered). For certain low reactivity metals such as Au and Pt, the previous deposition of a thin layer of another metal is needed to improve the adhesion. Ti and Cr are two frequently used adhesion promoters. Stress in evaporated or sputtered layers is typically tensile. The deposition rates are much higher than most CVD techniques. However, due to stress accumulation and cracking, a thickness beyond 2 μm is rarely deposited with these processes. For thicker depositions a technique described in the next section is sometimes used.

Electroplating

Electroplating (or electrodeposition) is a process typically used to obtain thick (tens of micrometers) metal structures. The sample to be electroplated is introduced in a solution containing a reducible form of the ion of the desired metal and is maintained at a negative potential (cathode) relative to a counter electrode (anode). The ions are reduced at the sample surface and the insoluble metal atoms are incorporated into the surface. As an example, copper electrodeposition is frequently done in copper sulfide-based solutions. The reaction taking place at the surface is $Cu^{2+} + 2e^- \rightarrow Cu_{(s)}$. Recommended current densities for electrodeposition processes are on the order of 5 to 100 mA/cm^2.

As can be deduced from the process mechanism, the surface to be electroplated has to be electrically conductive and preferably of the same material as the deposited one if a good adhesion is desired. In order to electrodeposit metals on top of an insulator (the most common case) a thin film of the same metal, called the seed layer, is previously deposited on the surface. Masking of the seed layer with a resist permits selective electroplatingat the patterned areas. Figure 5.10 illustrates a typical sequence of the steps required to obtain isolated metal structures.

5.1.3 Etching and Substrate Removal

Thin film and bulk substrate etching is another fabrication step that is of fundamental importance to both VLSI processes and micro/nanofabrication. In the VLSI area, various conducting and dielectric thin films deposited for passivation or masking purposes must be removed at some point or another. In micro/nanofabrication, in addition to thin film etching, very often the substrate (silicon, glass, GaAs, etc.) also needs to be removed in order to create various mechanical micro-/nanostructures (beams, plates, etc.). Two important figures of merit for any etching process are selectivity and directionality. Selectivity is the degree to which the etchant can differentiate between the masking layer and the layer to be

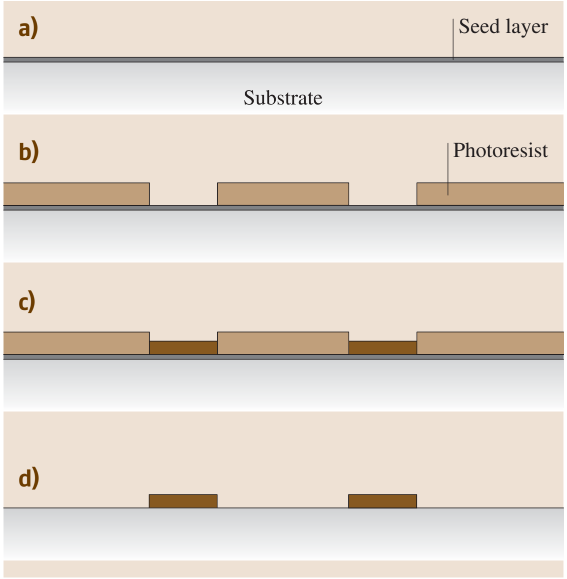

Fig. 5.10a–d Formation of isolated metal structures by electroplating through a mask: **(a)** seed layer deposition, **(b)** photoresist spinning and patterning, **(c)** electroplating, **(d)** photoresist and seed layer stripping

etched. Directionality has to do with the etch profile under the mask. In an isotropic etch, the etchant attacks the material in all directions at the same rate, creating a semicircular profile under the mask, Fig. 5.11a. In an anisotropic etch, the dissolution rate depends on specific directions, and one can obtain straight sidewalls or other noncircular profiles, Fig. 5.11b. One can also divide the various etching techniques into wet and dry categories. In this chapter, we will use this classification and discuss different wet etchants, followed by the dry etching techniques used most often in micro/nanofabrication.

Wet Etching

Historically, wet etching techniques preceded the dry ones. These still constitute an important group of etchants for micro/nanofabrication, in spite of their less frequent application in the VLSI processes. Wet etchants are by and large isotropic and show superior selectivity to the masking layer compared to various dry techniques. In addition, due to the lateral undercut, the minimum feature achievable with wet etchants is limited to $> 3\,\mu m$. Silicon dioxide is commonly etched in a dilute (6:1, 10:1, or 20:1 by volume) or buffered HF (BHF, HF + NH_4F) solutions (etch rate of $\sim 1{,}000$ Å/min in BHF). Photoresist and silicon nitride are the two most common masking materials for the wet oxide etch. The wet etchant for silicon nitride is hot (140–200 °C) phosphoric acid with silicon oxide as the masking material. Nitride wet etch is not very common (except for blanket etch), due to the masking difficulty and unrepeatable etch rates. Metals can be etched using various combinations of acids and base solutions. There are also many commercially available etchant formulations for aluminum, chromium, and gold that can be used easily. A comprehensive table of various metal etchants can be found in [5.5].

Anisotropic and isotropic wet etching of crystalline (silicon and gallium arsenide) and non-crystalline (glass) substrates are important topics in micro/nanofabrication [5.6–10]. In particular, the realization of the possibility of anisotropic wet etching of silicon is considered the beginning of the micromachining and MEMS discipline. Isotropic etching of silicon using HF/HNO_3/CH_3COOH (various different formulations have been used) dates back to the 1950s and is still frequently used to thin down the silicon wafer. The etch mechanism for this combination has been elucidated and is as follows: HNO_3 is used to oxidize the silicon, which is subsequently dissolved away in the HF. The acetic acid is used to prevent the dissociation of HNO_3 (the etch works as well without the acetic acid). For short etch times, silicon dioxide can be used as the masking material However, one needs to use silicon nitride if a longer etch time is desired. This etch also shows dopant selectivity with the etch rate dropping at lower doping concentrations ($< 10^{17}\,cm^{-3}$ n- or p-type). Although this effect can potentially be used as an etch stop mechanism in order to fabricate microstructures, the difficulty in masking has prevented a widespread application of this approach. Glass can also be isotropically etched using the HF/HNO_3 combination with the etch surfaces showing considerable roughness. This has been used extensively in fabricating microfluidic components (mainly channels). Although Cr/Au is usually used as the masking layer, long etch times require a more robust mask (bonded silicon has been used for this purpose).

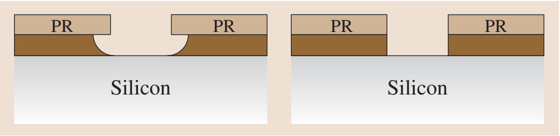

Fig. 5.11a,b Profile for isotropic **(a)** and anisotropic **(b)** etch through a photoresist (PR) mask

Silicon anisotropic wet etch constitutes an important technique in bulk micromachining. The three most important silicon etchants in this category are potassium hydroxide (KOH), ethylene diamine pyrochatechol (EDP), and tetramethyl ammonium hydroxide (TMAH). These are all anisotropic etchants that attack silicon along preferred crystallographic directions. In addition, they all show a marked reduction in the etch rate in heavily doped ($> 5\times10^{19}\,cm^{-3}$) boron ($p^{++}$) regions. The chemistry behind the action of these etchants is not yet very clear, but it seems that silicon atom oxidation at the surface and its reaction with hydroxyl ions (OH^-) are responsible for the formation of a soluble silicon complex ($SiO_2(OH)^{2-}$). The etch rate depends on the concentration and temperature and is usually around $1\,\mu m/min$ at temperatures of 85–115 °C. Common masking materials for anisotropic wet etchants are silicon dioxide and nitride, the latter being superior for longer etch times. The crystallographic plane which shows the slowest etch rate is the (111) plane. Although the lower atomic concentration along these planes have been speculated as the reason for this phenomenon, the evidence is inconclusive, and other factors must be included to account for this remarkable etch stop property. The anisotropic behavior of these etchants with respect to the (111) planes have been used extensively to create

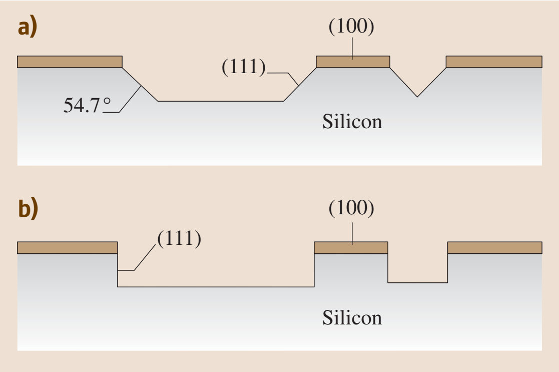

Fig. 5.12a,b Anisotropic etch profiles for: **(a)** (100) and **(b)** (110) silicon wafers

beams, membranes, and other mechanical and structural components. Figure 5.12 shows the typical cross sections of (100) and (110) silicon wafers etched with an anisotropic wet etchant. As can be seen, the (111) slow planes are exposed in both situations, one creating 54.7° sloped sidewalls in the (100) wafer and the other creating vertical sidewalls in the (110) wafer. Depending on the dimensions of the mask opening, a V-groove or a trapezoidal trench is formed in the (100) wafer. A large enough opening will allow the silicon to be etched all the way through the wafer, thus creating a thin dielectric membrane of the other side. It should be mentioned that exposed convex corners have a higher etch rate than the concave ones, resulting in an undercut that can be used to create dielectric (e.g., nitride) cantilever beams. Figure 5.13 shows a cantilever beam fabricated using the convex corner undercut on a (100) wafer.

The three above mentioned etchants show different directional and dopant selectivities. KOH has the best (111) selectivity (400/1) followed by TMAH and EDP. However, EDP has the highest selectivity with respect to the deep boron diffusion regions. Safety and CMOS compatibility are other important criteria for choosing a particular anisotropic etchant. Among the three mentioned etchants TMAH is the most benign, whereas EDP is extremely corrosive and carcinogenic. Silicon can be dissolved in TMAH in order to improve its selectivity with respect to aluminum. This property has made TMAH very appealing for post-CMOS micromachining for which aluminum lines have to be protected. Finally, it should be mentioned that one can modulate the etch rate using a reverse biased p–n junction (electrochemical etch stop). Figure 5.14 shows the setup commonly used to perform electrochemical etching. The silicon wafer under etch consists of an n-epi region on a p-type substrate. Upon the application of a reverse bias voltage to the structure (p-substrate is in contact with the solution and n-epi is protected using a water tight fixture), the p-substrate is etched away. When the n-epi regions are exposed to the solution, an oxide passivation layer is formed and etch is stopped. This technique can be used to fabricate single-crystalline silicon membranes for pressure sensors and other mechanical transducers.

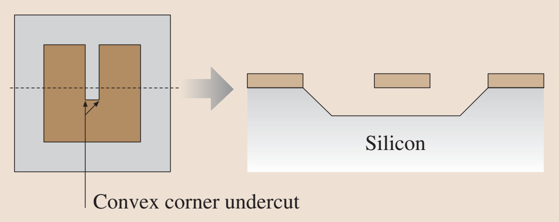

Fig. 5.13 Top view and cross section of a dielectric cantilever beam fabricated using convex corner undercut

Dry Etching

Most dry etching techniques are plasma-based. They have several advantages compared with wet etching. These include smaller undercut (allowing smaller lines to be patterned) and higher anisotropicity (allowing high-aspect-ratio vertical structures). However, the selectivity of dry etching techniques is lower than the wet etchants, and one must take into account the finite etch rate of the masking materials. The three basic dry etching techniques, namely, high-pressure plasma etching, reactive ion etching (RIE), and ion milling utilize different mechanisms to obtain directionality.

Ion milling is a purely physical process that utilizes accelerated inert ions (e.g., Ar^+) striking perpendic-

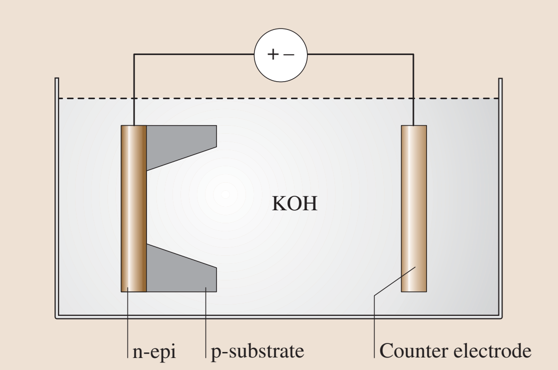

Fig. 5.14 Electrochemical etch setup

ular to the surface to remove the material (pressure $\sim 10^{-4}$–10^{-3} torr), Fig. 5.15a. The main characteristics of this technique are very low etch rates (on the order of a few nanometers per minute) and poor selectivity (close to 1:1 for most materials); hence it is generally used to etch very thin layers. In high-pressure (10^{-1}–5 torr) plasma etchers, highly reactive species are created that react with the material to be etched. The products of the reaction are volatile, so that they diffuse away and new material is exposed to the reactive species. Directionality can be achieved, if desired, with the sidewall passivation technique (Fig. 5.15b). In this technique, nonvolatile species produced in the chamber deposit and passivate the surfaces. The deposit can only be removed by physical collision with incident ions. Because the movement of the ions has a vertical directionality, the deposit is removed mainly at the horizontal surfaces, while the vertical walls remain passivated. In this fashion, the vertical etch rate becomes much higher than the lateral one.

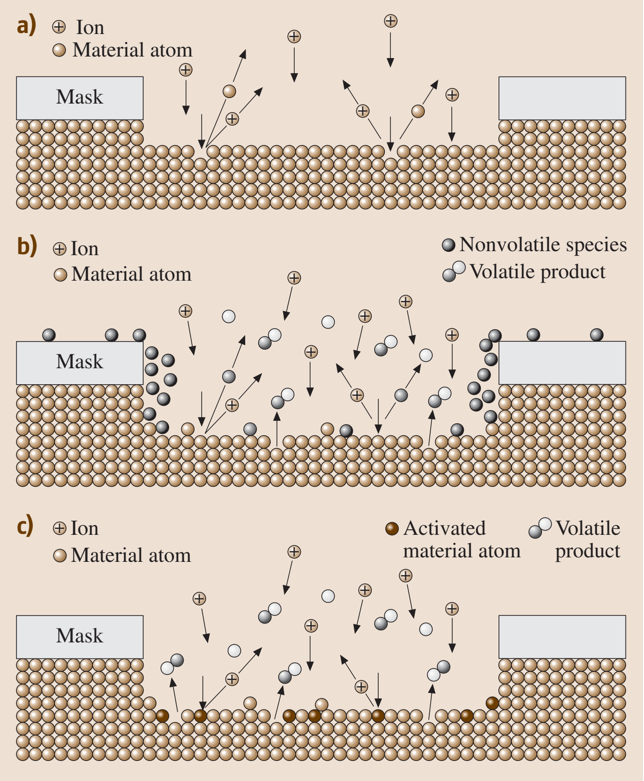

Fig. 5.15a–c Simplified representation of etching mechanisms for **(a)** ion milling, **(b)** high-pressure plasma etching, and **(c)** RIE

The RIE etching, also called ion-assisted etching, is a combination of physical and chemical processes. In this technique, the reactive species react with the material only when the surfaces are "activated" by the collision of incident ions from the plasma (e.g., by breaking bonds at the surface). As in the previous technique, the directionality of the ion's velocity produces many more collisions in the horizontal surfaces than in the walls, thus generating faster etching rates in the vertical direction (Fig. 5.15c). To further increase the etch anisotropy, sidewall passivation methods in some cases, are also used. An interesting case is the *deep reactive ion etching* (DRIE) technique, capable of achieving aspect ratios of 30:1 and silicon etching rates of 2 to 3 μm/min (through-wafer etch is possible). In this technique, the passivation deposition and etching steps are performed sequentially in a two-step cycle, as shown in Fig. 5.16. In commercial silicon DRIE etchers, SF_6/Ar is typically used for the etching step and a combination of Ar and a fluoropolymer ($n\mathrm{CF}_2$) for the passivation step. A Teflon-like polymer about 50 nm thick is deposited during the latter step, covering only the sidewalls (Ar^+ ion bombardment removes the Teflon on the horizontal surfaces). Due to the cyclic nature of this process, the sidewalls of the etched features show a periodic "wave-shape" roughness in the range of 50 to 400 nm.

Dry etching can also be performed in non-plasma equipment if the etching gases are reactive enough. The so-called vapor phase etching (VPE) processes can be carried out in a simple chamber with gas feeding and pumping capabilities. Two examples of VPE are xenon difluoride (XeF_2) etching of silicon and HF-vapor etching of silicon dioxide. Due to its isotropic nature, these processes are typically used for etching sacrificial layers and releasing structures while avoiding stiction problems (see Sects. 5.2.1 and 5.2.2).

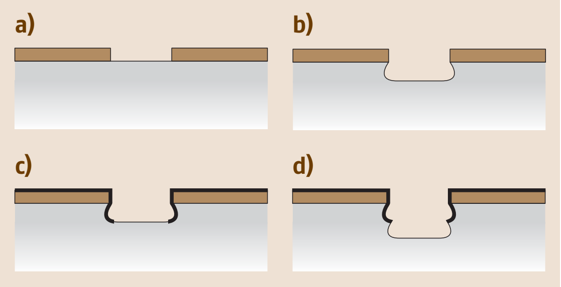

Fig. 5.16a–d DRIE cyclic process: **(a)** photoresist patterning, **(b)** etch step, **(c)** passivation step, and **(d)** etch step

Table 5.1 Typical dry etch chemistries

Si	CF_4/O_2, CF_2Cl_2, CF_3Cl, $SF_6/O_2/Cl_2$, $Cl_2/H_2/C_2F_6/CCl_4$, C_2ClF_5/O_2, Br_2, SiF_4/O_2, NF_3, ClF_3, CCl_4, CCl_3F_5, C_2ClF_5/SF_6, C_2F_6/CF_3Cl, CF_3Cl/Br_2
SiO_2	CF_4/H_2, C_2F_6, C_3F_8, CHF_3/O_2
Si_3N_4	$CF_4/O_2/H_2$, C_2F_6, C_3F_8, CHF_3
Organics	O_2, CF_4/O_2, SF_6/O_2
Al	BCl_3, BCl_3/Cl_2, $CCl_4/Cl_2/BCl_3$, $SiCl_4/Cl_2$
Silicides	CF_4/O_2, NF_3, SF_6/Cl_2, CF_4/Cl_2
Refractories	CF_4/O_2, NF_3/H_2, SF_6/O_2
GaAs	BCl_3/Ar, $Cl_2/O_2/H_2$, $CClH_3/H_2$ H_2, CH_4/H_2, $CCl_2F_2/O_2/Ar/He$
InP	CH_4/H_2, C_2H_6/H_2, Cl_2/Ar
Au	$C_2Cl_2F_4$, Cl_2, $CClF_3$

Most importantly, materials can be etched with the aforementioned techniques, and for each material a variety of chemistries is available. Table 5.1 lists some of the most common materials along with selected etch recipes [5.11]. For each chemistry, the etch rate, directionality, and selectivity with respect to the mask materials depend on parameters such as the flow/rates of the gases entering the chamber, the working pressure, and the RF power applied to the plasma.

5.1.4 Substrate Bonding

Substrate (wafer) bonding (silicon-silicon, silicon-glass, and glass-glass) is among the most important fabrication techniques in microsystem technology [5.12, 13]. It is frequently used to fabricate complex 3-D structures both as a functional unit and as part of the final microsystem package and encapsulation. The two most important bonding techniques are silicon-silicon fusion (or silicon direct bonding) and silicon-glass electrostatic (or anodic) bonding. In addition to these techniques, several other alternative methods that utilize an intermediate layer (eutectic, adhesive, and glass frit) have also been investigated. All these techniques can be used to bond the substrates at the wafer level. In this chapter, we will only discuss wafer-level techniques and will not treat the device level bonding methods (e.g., e-beam and laser welding).

Silicon Direct Bonding

Direct silicon, or fusion bonding, is used in the fabrication of micromechanical devices and silicon-on-insulator (SOI) substrates. Although it is mostly used to bond two silicon wafers with or without an oxide layer, it has also been used to bond different semiconductors such as GaAs and InP [5.13]. One main requirement for a successful bond is sufficient planarity (< 10 Å surface roughness and < 5 µm bow across a 4" wafer) and cleanliness of the surfaces. In addition, thermal expansion mismatch also needs to be considered if bonding of two dissimilar materials is contemplated. The bonding procedure is as follows: The silicon- or oxide-coated silicon wafers are first thoroughly cleaned. Subsequently, the surfaces are hydrated (activated) in HF or boiling nitric acid (RCA clean also works). This renders the surfaces hydrophilic by creating an abundance of hydroxyl ions. Then the substrates are brought together in close proximity (starting from the center to avoid the void formation). The close approximation of the bonding surfaces allows the short-range attractive van der Waals forces to bring the surfaces into intimate contact on an atomic scale. Following this step, a hydrogen bond between the two hydroxyl-coated silicon wafers bonds the substrates together. These steps can be performed at room temperatures; however, in order to increase the bond strength a high temperature (800–1,200 °C) anneal is usually required. A big advantage of silicon fusion bonding is the thermal matching of the substrates.

Anodic Bonding

Silicon-glass anodic (electrostatic) bonding is another major substrate joining technique that has been used extensively for microsensor packaging and device fabrication. The main advantage of this technique is its lower bonding temperature, which is around 300–400 °C. Figure 5.17 shows the bonding setup. A glass wafer (usually Pyrex 7740 for its thermal expansion match to silicon)

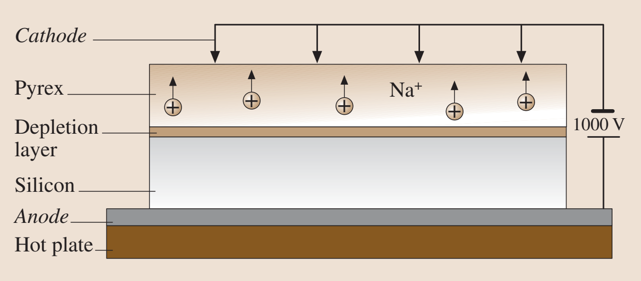

Fig. 5.17 Glass-to-silicon anodic bonding setup

is placed on top of a silicon wafer, and the sandwich is heated to 300–400 °C. Subsequently, a voltage of $\sim 1{,}000$ V is applied to the glass-silicon sandwich with the glass connected to the cathode. The bond starts immediately after the application of the voltage and spreads outward from the cathode contact point. The bond can be visually observed as a dark greyish front that expands throughout the wafer.

The bonding mechanism is as follows: During the heating period, glass sodium ions move toward the cathode and create a depletion layer at the silicon-glass interface. A strong electrostatic force is therefore created at the interface that pulls the substrates into intimate contact. The exact chemical reaction responsible for anodic bonding is not yet clear, but covalent silicon oxygen bond at the interface seems to be responsible for the bond. Silicon-to-silicon anodic bonding using sputtered or evaporated glass interlayer is also possible.

Bonding with Intermediate Layers

Various other wafer bonding techniques utilizing an intermediate layer have also been investigated [5.13]. Among the most important ones are adhesive, eutectic, and glass frit bonds. Adhesive bond using a polymer (e.g., polyimides, epoxies, thermoplast adhesives, and photoresists) in between the wafers has been used to join different wafer substrates [5.14]. Complete curing (in an oven or using dielectric heating) of the polymer before or during the bonding process prevents subsequent solvent outgas and void formation. Although reasonably high bonding strengths can be obtained, these bonds are non-hermetic and unstable over time.

In the eutectic bonding process, gold-coated silicon wafers are bonded together at temperatures greater than the silicon-gold eutectic point (363 °C, 2.85% silicon and 97.1% Au) [5.15]. This process can achieve high bonding strength and good stability at relatively low temperatures. For good bond uniformity, silicon dioxide must be removed from the silicon surface prior to the deposition of the gold. In addition, all organic contaminants on the gold surface must be removed (using UV light) prior to the bond. Pressure should also be applied in order to achieve a better contact. Although eutectic bond can be accomplished at low temperatures, achieving uniformity over large areas has proven to be challenging. More recently, localized silicon-gold eutectic bonding has been reported. In this technique, a silicon cap is bonded to another silicon wafer having a gold micro-heater (0.27 Å current for a bonding temperature of about 800 °C), Fig. 5.18 [5.16].

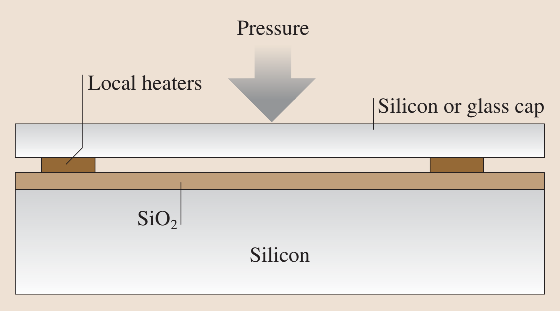

Fig. 5.18 Eutectic bonding with localized heating

Glass frit can also be used as an interlayer in substrate bonding. In this technique, first a thin layer of glass is deposited and pre-glazed. The glass-coated substrates are then brought into contact, and the sandwich is heated to above the glass melting temperature (typically < 600 °C). Like in the eutectic process, pressure must be applied for adequate contact [5.17].

Bond Characterization

Two major characterization techniques for substrate bonding evaluation are: 1) visual inspection and 2) blade test [5.12]. Once the bond has been completed, the bonding surfaces can be visually inspected for the presence of voids, cracks, and nonuniformity. The silicon-glass anodic bond interface does not require any particular imaging instrument and can be inspected with an unaided eye or under a light microscope. However, infrared imaging, X-ray photography, and ultrasonic methods are necessary for opaque substrates (mainly silicon-silicon). Although infrared transmission has some resolution limits due to its wavelength, it

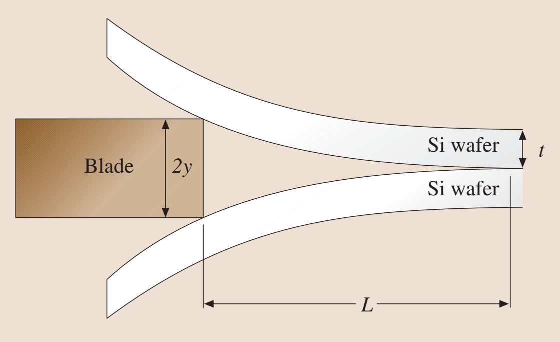

Fig. 5.19 Blade test technique for the measurement of bond strength

is the most common tool in bond inspection and offers a quality technique with a simple IR source and a CCD camera.

The blade test is often performed in order to estimate the specific surface energy (and hence bond strength), Fig. 5.19. In this test, a blade is inserted between the bonded substrates and a crack is induced. The length of the crack is then measured and the specific surface energy is calculated using the following equation [5.18]:

$$g = \frac{3Et^3y^2}{8L^4} ,$$

where g is the surface energy (erg/cm^2), E is the Young's modulus of substrate (dyne/cm^2), t is the thickness of the wafer, y is the half-thickness of the blade, and L is the crack length. The crack length is usually measured using the interference fringes parallel to the blade edge under an IR light source. A semi-quantitative technique for evaluation of the bond strength is the pull test. This technique is based on using a test fixture (or tensometer) for measuring the force required to pull two bonded substrates apart. The pull test is more effective on small samples and can be used to compare different bonding techniques without the need to accurately measure the surface energy.

5.2 MEMS Fabrication Techniques

In this section, we will discuss various important MEMS fabrication techniques commonly used to build various microdevices (microsensors and microactuators) [5.6–10]. The dimensional spectrum of the microstructures that can be fabricated using these techniques span from 1 mm to 1 µm. As mentioned in the introduction, we will mostly emphasize the more important techniques and will not discuss specialized methods.

5.2.1 Bulk Micromachining

Bulk micromachining is the oldest MEMS technology and hence probably one of the more mature ones [5.8, 9]. It is currently by far the most commercially successful one, helping manufacture devices such as pressure sensors and ink-jet print heads. Although there are many different variations, the basic concept behind bulk micromachining is selective removal of the substrate (silicon, glass, GaAs, etc.). This allows the creation of various micromechanical components such as beams, plates, and membranes that can be used to fabricate a variety of sensors and actuators. The most important microfabrication techniques used in bulk micromachining are wet and dry etch and substrate bonding. Although one can use different criteria in dividing the bulk micromachining techniques into separate categories, we will use a historical time line for this purpose. Starting with the more traditional wet etching techniques, we will proceed to the more recent ones that use deep RIE and wafer bonding.

Bulk Micromachining Using Wet Etch and Wafer Bonding

The use of anisotropic wet etchants to remove silicon can be regarded as the beginning of the micromachining era. Back side etch was used to create movable structures such as beams, membranes, and plates, Fig. 5.20. Initially, the etching was timed to create a specified thickness. However, this technique proved inadequate in creating thin structures (< 20 µm). Subsequent use of various etch stop techniques allowed the creation of thinner membranes in a more controlled fashion. As mentioned in Sect. 5.1.3, heavily doped boron regions and electrochemical bias can be used to drastically slow down the etch process and hence create controllable thickness microstructures. Figure 5.21a,b show the cross section of two piezoresistive pressure sensors fabricated using electrochemical and P^{++} etch stop techniques. The use of the P^{++} method requires the epitaxial growth of a lightly doped region on top of a P^{++} etch stop layer.

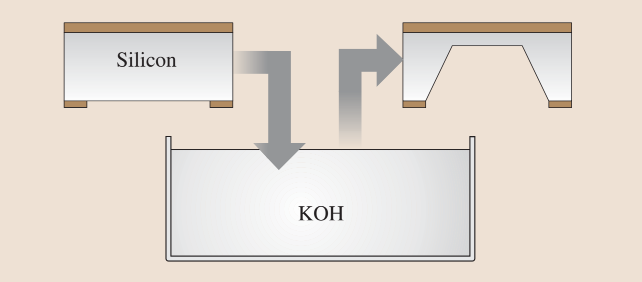

Fig. 5.20 Wet anisotropic silicon back side etch

This layer is subsequently used for the placement of piezoresistors. However, if no active component is required, one can simply use the P^{++} region to create a thin membrane, Fig. 5.21c.

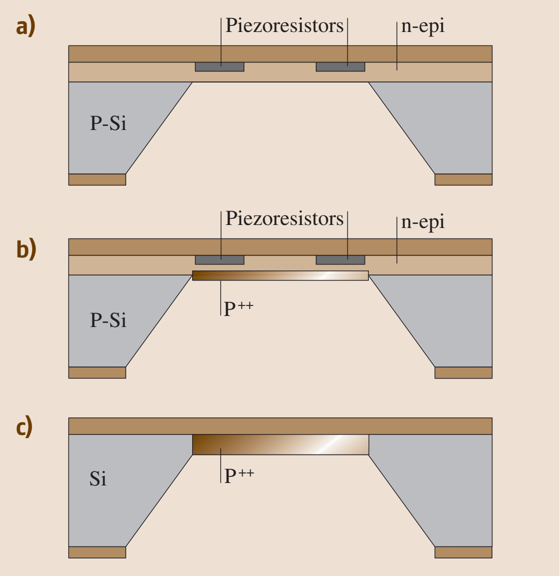

Fig. 5.21a–c Wet micromachining etch stop techniques: **(a)** electrochemical with n-epi on p substrate, **(b)** P^{++} etch stop with n-epi, and **(c)** P^{++} etch stop without n-epi

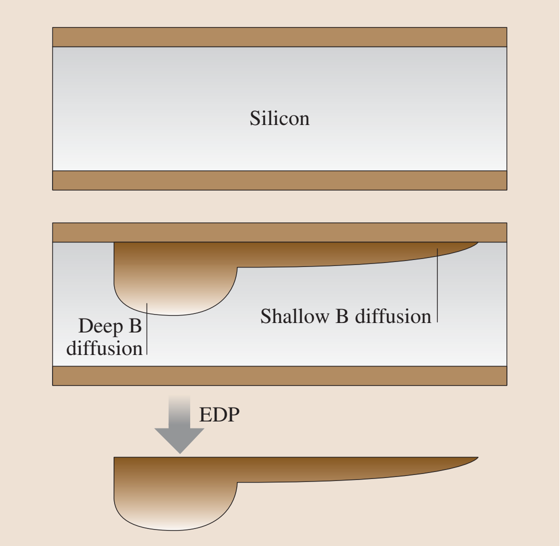

Fig. 5.22 Freestanding microstructure fabrication using deep and shallow boron diffusions and EDP release

The P^{++} etch stop technique can also be used to create isolated thin silicon structures through the dissolution of the entire lightly doped region [5.19]. This technique was successfully used to fabricate silicon recording and stimulate electrodes for biomedical applications. Figure 5.22 shows the cross section of such a process that relies on deep (15–20 μm) and shallow boron (2–5 μm) diffusion steps to create microelectrodes with flexible connecting ribbon cables. An extension of this process that uses a combination of P^{++} etch stop layers and silicon-glass anodic bonding has also been developed. This process is commonly known as the dissolved wafer process and has been used to fabricate a variety of microsensors and microactuators [5.20]. Figure 5.23 shows the cross section of this process. Figure 5.24 shows an SEM photograph of a micro-accelerometer fabricated using the dissolved wafer process.

It is also possible to merge wet bulk micromachining and microelectronics fabrication processes to build mi-

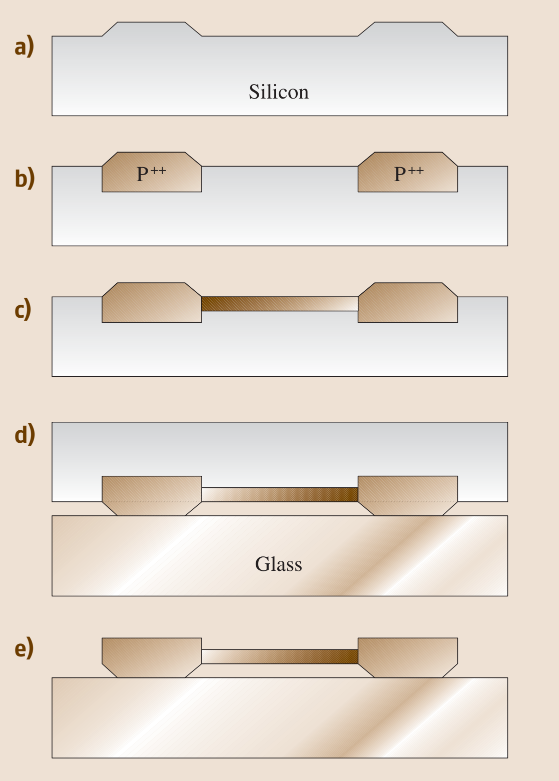

Fig. 5.23a–e Dissolved wafer process sequence, **(a)** KOH etch, **(b)** deep B diffusion, **(c)** shallow B diffusion, **(d)** silicon-glass anodic bond, and **(e)** release in EDP

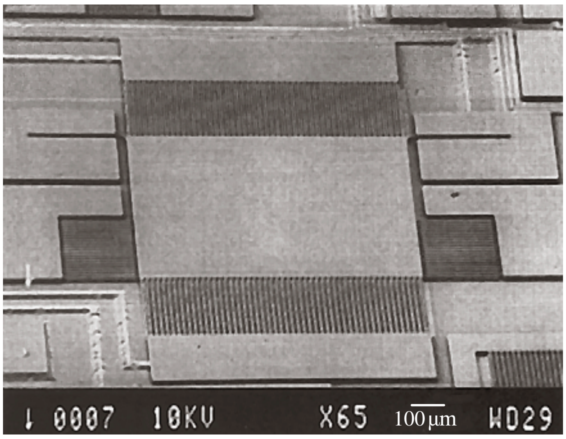

Fig. 5.24 SEM photograph of a micro-accelerometer fabricated using the dissolved wafer process [5.20]

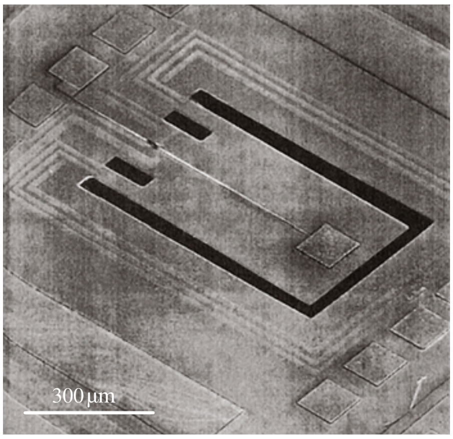

Fig. 5.26 Photograph of a post-CMOS processed cantilever beam resonator for chemical sensing [5.21]

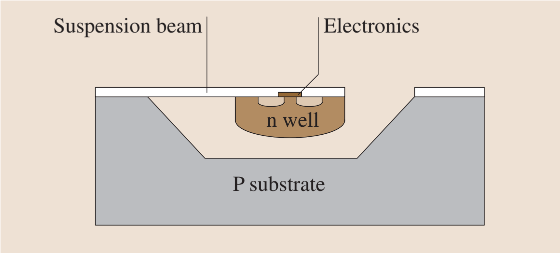

Fig. 5.25 Suspended island created on a prefabricated CMOS chip using front-side wet etch and electrochemical etch stop

cromechanical components on the same substrate as the integrated circuits (CMOS, Bipolar, or BiCMOS) [5.21]. This is very appealing since it allows the integration of interface and signal processing circuitry with MEMS structures on a single chip. However, important fabrication issues such as process compatibility and yield have to be carefully considered. Among the most popular techniques in this category is the post-processing of the CMOS integrated circuits by front side etching in TMAH solutions. As mentioned previously, silicon-rich TMAH does not attack aluminum and therefore can be used to undercut microstructures in an already processed CMOS chip. Figure 5.25 shows a schematic of such a process in which front side wet etch and electrochemical etch stop have been used to produce suspended beams. This technique has been used extensively to fabricate a variety of microsensors (e.g., humidity, gas, chemical, and pressure). Figure 5.26 shows a photograph of a post-CMOS processed chemical sensor.

Bulk Micromachining Using Dry Etch

Bulk silicon micromachining using dry etch is a very attractive alternative to the wet techniques described in the previous section. These techniques were developed in the mid-1990s, following the successful efforts to develop processes for anisotropic dry silicon etch. More recent advances in deep silicon RIE and the availability of SOI wafers with a thick top silicon layer have increased the application of these techniques. These techniques allow the fabrication of high-aspect-ratio vertical structures in isolation or along with on-chip electronics. Process compatibility with active microelectronics is less of a concern in dry methods, since many of them do not damage the circuit or its interconnect.

The most simple dry bulk micromachining technique relies on the front side undercut of microstructures using XeF_2 vapor phase etch [5.22]. As mentioned before, this, however, is an isotropic etch and, therefore, has a limited application. A combination of isotropic/anisotropic dry etch is more useful and can be used to create a variety of interesting structures. Two successful techniques using this combination are single-crystal reactive etching and metallization (SCREAM) [5.23] and post-CMOS dry release using aluminum/silicon dioxide laminate [5.24]. The first technique relies on the combination of isotropic/anisotropic dry etch to create single crystalline suspended structures. Figure 5.27 shows the cross section of the process. It starts with an anisotropic (Cl_2/BCl_3) silicon etch using an oxide mask (Fig. 5.27b). This is followed by a conformal PECVD oxide deposition (Fig. 5.27c). Subsequently, an anisotropic oxide etch is used to remove the oxide at the bottom of the

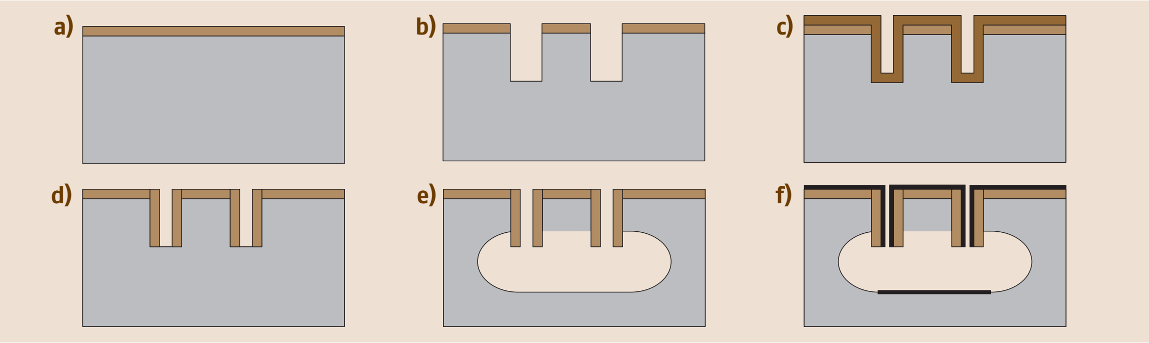

Fig. 5.27 Cross section of the SCREAM process

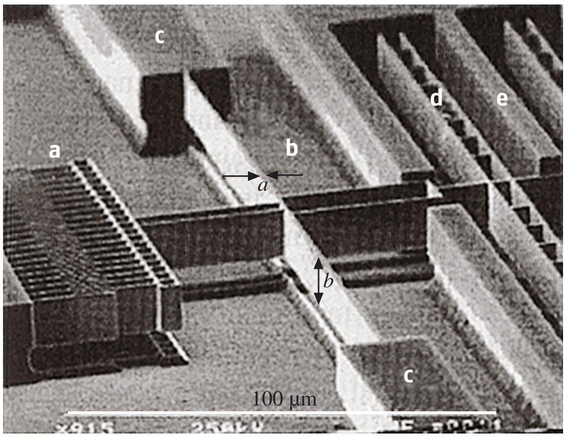

Fig. 5.28a–e SEM photograph of a structure fabricated using SCREAM process: **(a)** comb-drive actuator, **(b)** suspended spring, **(c)** spring support, **(d)** moving suspended capacitor plate, and **(e)** fixed capacitor plate [5.25]

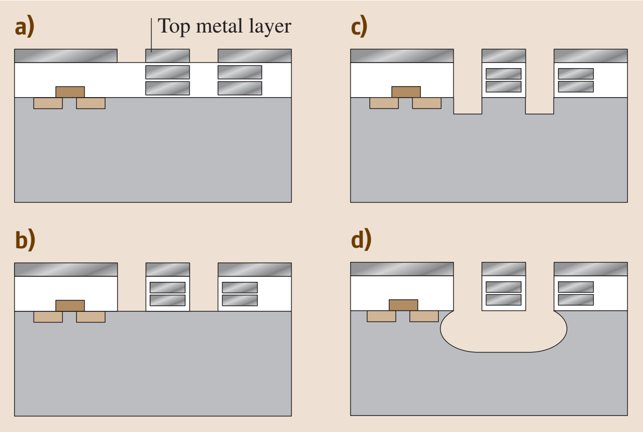

Fig. 5.29 Cross section of the process flow for the post-CMOS dry microstructure fabrication

trenches leaving the sidewall oxide intact (Fig. 5.27d). At this stage, an isotropic silicon etch (SF_6) is performed that results in undercut and release of the silicon structures (Fig. 5.27e). Finally, if electrostatic actuation is desired, a metal can be sputtered to cover the top and sidewall of the microstructure and the bottom of the cavity formed below it (Fig. 5.27f). Figure 5.28 shows an SEM photograph of a comb-drive actuator fabricated using SCREAM technology.

The second dry release technique relies on the masking capability of aluminum interconnect lines in a CMOS integrated circuit to create suspended microstructures. Figure 5.29 shows the cross section of this process. As can be seen, the third level Al of a prefabricated CMOS chip is used as a mask to anisotropically etch the underlying oxide layers all the way to the silicon (CHF_3/O_2), Fig. 5.29b. This is followed by an anisotropic silicon etch to create a recess in the silicon, which will be used in the final step to facilitate the undercut and release, Fig. 5.29c. Finally, an isotropic silicon etch is used to undercut and release the structures, Fig. 5.29d. Figure 5.30 shows an SEM photograph of a comb-drive actuator fabricated using this technology.

In addition to the methods described above, recent advancements in the development of deep reactive ion etching of silicon (DRIE, see Sect. 5.1.3) have created new opportunities for dry bulk micromachining techniques (see Sect. 5.2.3). One of the most important ones uses thick silicon SOI wafers that are commercially available in various top silicon thicknesses [5.26]. Figure 5.31 shows the cross section of a typical process using DRIE and SOI wafers. The top silicon layer is patterned and etched all the way to the buried oxide, Fig. 5.31b. The oxide is subsequently removed in HF, releasing suspended single crystalline microstructures, Fig. 5.31c. In a modification of this process, the substrate can also be removed from the back side, allowing

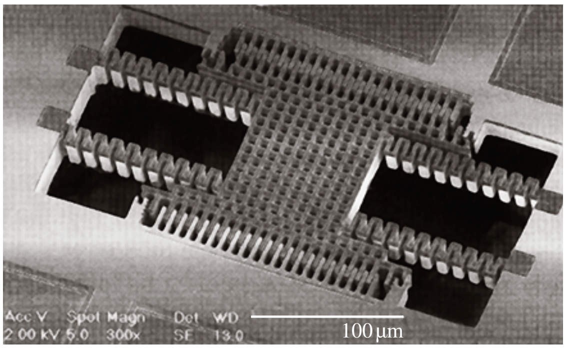

Fig. 5.30 SEM photograph of a comb-drive actuator fabricated using aluminum mask post-CMOS dry release [5.27]

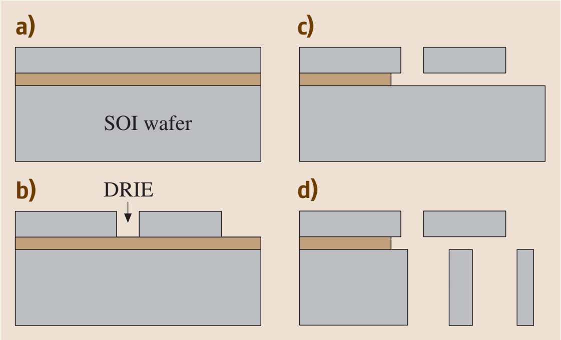

Fig. 5.31 DRIE processes using SOI wafers

easy access from both sides (this allows an easier release and prevents stiction), Fig. 5.31d.

5.2.2 Surface Micromachining

Surface micromachining is another important MEMS microfabrication technique that can be used to create movable microstructures on top of a silicon substrate [5.28]. This technique relies on the deposition of structural thin films on a sacrificial layer that is subsequently etched away, resulting in movable micromechanical structures (beams, membranes, plates, etc.). The main advantage of surface micromachining is that extremely small sizes can be obtained. In addition, it is relatively easy to integrate the micromachined structures with on-chip electronics for increased functionality. However, due to the increased surface nonplanarity with any additional layer, there is a limit to the number of layers that can be deposited. Although one of the earliest reported MEMS structures was a surface micromachined resonant gate transistor [5.29], material-related difficulties resulted in the termination of efforts in this area. In the mid-1980s, improvements in the field of thin film deposition rekindled interest in surface micromachining [5.30]. Later in the same decade, polysilicon surface micromachining was introduced, which opened the door to the fabrication of a variety of microsensors (accelerometers, gyroscopes, etc.) and microactuators (micromirrors, RF switches, etc.). In this section, we will concentrate on the key process steps involved in surface micromachining fabrication and the various materials used in the surface micromachining process. In addition, monolithic integration of CMOS with MEMS structures and 3-D surface micromachining are also discussed.

Basic Surface Micromachining Processes

The basic surface micromachining process is illustrated in Fig. 5.32. The process begins with a silicon substrate on top of which a sacrificial layer is grown and patterned, Fig. 5.32a. Subsequently, the structural material is deposited and patterned, Fig. 5.32b. As can be seen, the structural material is anchored to the substrate through the openings created in the sacrificial layer during the previous step. Finally, the sacrificial layer is removed, resulting in the release of the microstructures Fig. 5.32c. In wide structures, it is usually necessary to provide access holes in the structural layer for a fast sacrificial layer removal. It is also possible to seal microcavities cre-

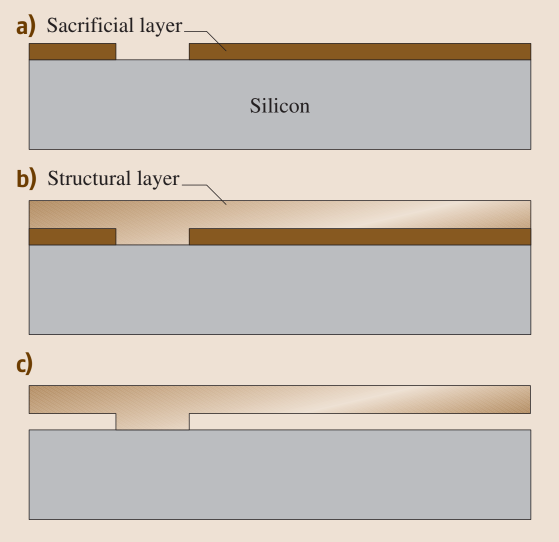

Fig. 5.32 Basic surface micromachining fabrication process

ated by the surface micromachining technique [5.8, 9]. This can be done at the wafer level and is a big advantage in applications such as pressure sensors that require a sealed cavity. Figure 5.33 shows two different techniques that can be used for this purpose. In the first, following the etching of the sacrificial layer, a LPCVD dielectric layer (oxide or nitride) is deposited to cover and seal the etch holes in the structural material, Fig. 5.33a. Since the LPCVD deposition is performed at reduced pressures, a sub-atmospheric pill-box microcavity can be created. In the second technique, also called reactive sealing, the polysilicon structural material is oxidized following the sacrificial layer removal, Fig. 5.33b. If access holes are small enough, the grown oxide can seal the cavity. Due to the consumption of the oxygen during the growth process, the cavity is sub-atmospheric.

The most common sacrificial and structural materials are phosphosilicate glass (PSG) and polysilicon, respectively (low temperature oxide, or LTO, is also frequently used as the sacrificial layer). However, there are several other sacrificial/structural combinations that have been used to create a variety of surface micromachined structures. Important design issues related to the choice of the sacrificial layer are: 1) quality (pinholes, etc.), 2) ease of deposition, 3) deposition rate, 4) deposition temperature, and 5) etch difficulty and selectivity (sacrificial layer etchant should not attack the structural layer). The particular choice of material for structural layers depends on the desired properties and the specific application. Several important requirements are: 1) ease of deposition, 2) deposition rate, 3) step coverage, 4) mechanical properties (internal stress, stress gradient, Young's modulus, fracture strength, and internal damping), 5) etch selectivity, 6) thermal budget and history, 7) electrical conductivity, and 8) optical reflectivity. Two examples from the commercially available surface micromachined devices illustrate various successful sacrificial/structural combinations. Texas Instruments's deformable mirror display (DMD) spatial light modulator uses aluminum as the structural material (good optical reflectivity) and photoresist as the sacrificial layer (easy dry etch and low processing temperatures, allowing easy post-IC integration with CMOS) [5.32], Fig. 5.34. In contrast, Analog Devices's microgyroscope uses polysilicon structural material and a PSG sacrificial layer, Fig. 5.35. Two recent additions to the collection of available structural layers are poly-silicon-germanium and poly-germanium [5.33, 34]. These are intended to be a substitute for polysilicon in applications where the high polysilicon deposition temperature (around 600 °C) is forbidding (e.g., CMOS integration).

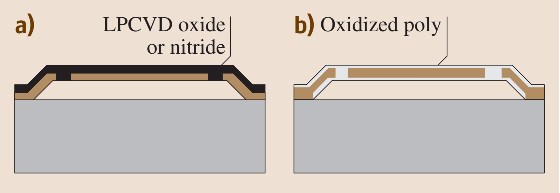

Fig. 5.33 Two sealing techniques for surface micromachined created cavities

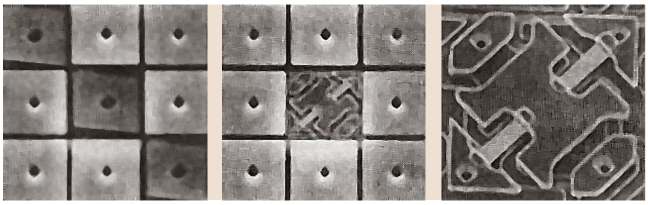

Fig. 5.34 SEM photographs of Texas Instrument micromirror array [5.28]

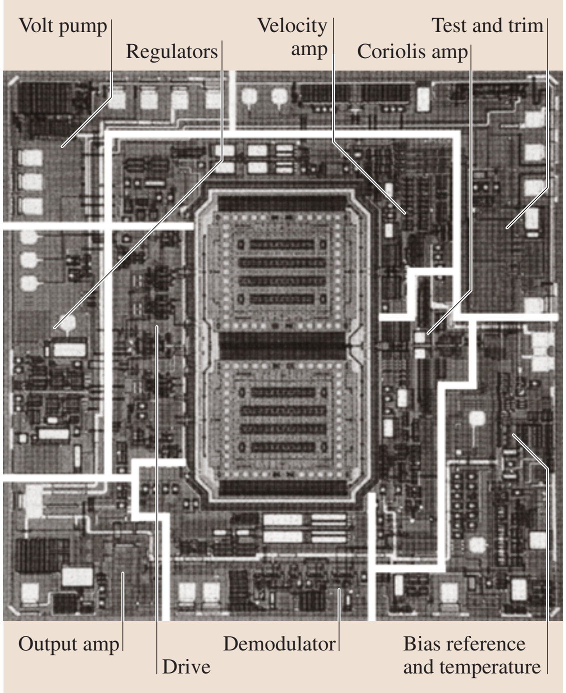

Fig. 5.35 SEM photograph of Analog Devices gyroscope [5.31]

Table 5.2 Several important surface micromachined sacrificial/structural combinations

System	Sacrificial layer	Structural layer	Structural layer etchant	Sacrificial layer etchant
1	PSG or LTO	Poly-Si	RIE	Wet or vapor HF
2	Photoresist, polyimide	Metals (Al, Ni, Co, Ni-Fe)	Various metal etchants	Organic solvents, plasma O_2
3	Poly-Si	Nitride	RIE	KOH
4	PSG or LTO	Poly-Ge	H_2O_2 or RCA1	Wet or vapor HF
5	PSG or LTO	Poly-Si-Ge	H_2O_2 or RCA1	Wet or vapor HF

Unlike LPCVD polysilicon, poly-germanium (poly-Ge) and poly-silicon-germanium (poly-$Si_{1-x}Ge_x$) can be deposited at temperatures as low as 350 °C (poly-Ge deposition temperature is usually lower than poly-SiGe). Table 5.2 summarizes important surface micromachined sacrificial/structural combinations.

An important consideration in the design and processing of surface micromachined structures is the issue of stiction [5.8, 9, 35, 36]. This can happen during the release step if a wet etchant is used to remove the sacrificial layer, or during the device lifetime. The reason for stiction during release is the surface tension of the liquid etchant, which can hold the microstructure down and cause stiction. This usually happens when the structure is compliant and does not possess enough spring constant to overcome the surface tension force of the rinsing liquid (i. e., water). There are several ways one can alleviate the release-related stiction problem. These include: 1) the use of dry or vapor phase etchant, 2) the use of solvents with lower surface tension, 3) geometrical modifications, 4) CO_2 critical drying, 5) freeze drying, and 6) *self-assembled monolayer* (SAM) or organic thin-film surface modification. The first technique prevents stiction by not using a wet etchant, although in the case of vapor phase release, the condensation possibility is existent and can cause some stiction. The second method uses rinsing solvents such as methanol with a lower surface tension than water. This is usually followed by a rapid evaporation of the solvent on a hot plate. However, this technique is not optimum, and many structures still stick. The third technique is geometrical, which provides dimples in the structural layer in order to reduce the contact surface area and hence reduce the attractive force. The fourth and fifth techniques rely on the phase change (in one case CO_2 and the other butyl-alcohol), which avoids the liquid phase altogether by directly going to the gas phase. The last technique uses self-assembled monolayers or organic thin films to coat the surfaces with a hydrophobic layer. The stiction that occurs during the operating lifetime of the device (in-use stiction) is due to the condensation of moisture on the surfaces, electrostatic charge accumulation, or direct chemical bonding. Surface passivation using self-assembled monolayers or organic thin films can be used to reduce the surface energy and reduce or eliminate the capillary forces and direct chemical bonding. These organic coatings also reduce electrostatic forces if a thin layer is applied directly to the semiconductor (without the intervening oxide layer). Commonly used organic coatings include fluorinated fatty acids (Texas Instruments aluminum micromirrors), silicone polymeric layers (Analog Devices accelerometers), and siloxane self-assembled monolayers.

Surface Micromachining Integration with Active Electronics

Integration of surface micromachined structures with on-chip circuitry can increase performance and simplify packaging. However, issues related to process compatibility and yield must be carefully considered. The two most common techniques are MEMS-first and MEMS-last techniques. In the MEMS-last technique, the integrated circuit is first fabricated and surface micromachined structures are subsequently built on top of the silicon wafer. An aluminum structural layer with a photoresist sacrificial layer is an attractive combination due to the low thermal budget of the process (Texeas Instruments micromirror array). However, in applications where mechanical properties of Al are not adequate, polysilicon structural material with LTO or PSG sacrificial layer must be used. Due to the rather high deposition temperature of polysilicon, this combination requires special attention with regard to the thermal budget. For example, aluminum metallization must be avoided and substituted with refractory metals such as tungsten. This can only be achieved at a greater process complexity and lower transistor performance.

MEMS-first technique alleviates these difficulties by fabricating the microstructures at the very beginning of the process. But if the microstructures are processed first, they have to be buried in a sealed trench to eliminate the interference of microstructures with subsequent CMOS processes. Figure 5.36 shows a cross section of a MEMS-first fabrication process developed at the Sandia National Laboratory [5.37]. The process starts with a shallow anisotropic etching of trenches in a silicon substrate to accommodate the height of the polysilicon structures fabricated later on. A silicon nitride layer is then deposited to provide isolation at the bottom of the trenches. Next, several layers of polysilicon and sacrificial oxide are deposited and patterned in a standard surface micromachining process. Subsequently, the trenches are completely filled with sacrificial oxide and the wafers are planarized with chemical-mechanical polishing (this avoids complication in the following lithographic steps). After an annealing step, the trenches are sealed with a nitride cap. At this point, a standard CMOS fabrication process is performed. At the end of the CMOS process, the nitride cap is etched and the buried structures released by etching the sacrificial oxide.

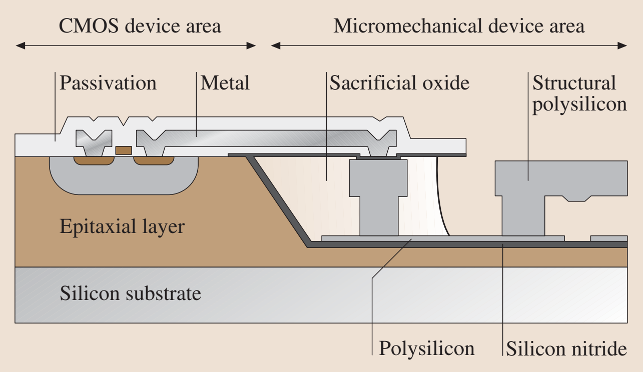

Fig. 5.36 Cross section of the Sandia MEMS-first integrated fabrication process

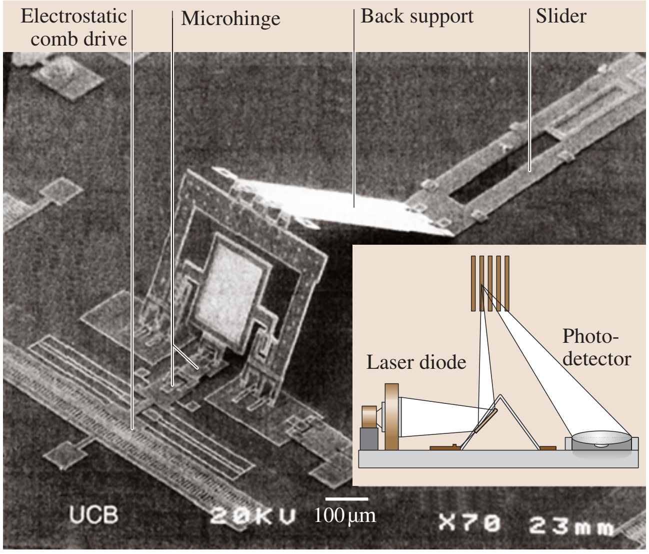

Fig. 5.37 Silicon pin-and-sample hinge scanner with 3-D surface micromachined structures [5.38]

3-D Microstructures Using Surface Micromachining

Three-dimensional surface microstructures can be fabricated using surface micromachining. The fabrication of hinges for the vertical assembly of MEMS was a major advance toward achieving 3-D microstructures [5.39]. Optical microsytems have greatly benefited from surface micromachined 3-D structures. These microstructures are used as passive or active components (micromirror, Fresnel lens, optical cavity, etc.) on a silicon optical bench (silicon microphotonics). An example is the Fresnel lens that has been surface micromachined in polysilicon and then erected using hinge structures and locked in place using micromachined tabs, thus liberating the structure from the horizontal plane of the wafer [5.38, 40]. Various microactuators (e.g., comb drive and vibromotors) have been used to move these structures out of the silicon plane and into position. Figure 5.37 shows an SEM photograph of a bar-code micro-scanner using a silicon optical microbench with 3-D surface micromachined structures.

5.2.3 High-Aspect-Ratio Micromachining

The bulk and surface micromachining technologies presented in the previous sections fulfill the requirements of a large group of applications. Certain applications, however, require the fabrication of high-aspect-ratio structures, which is not possible with the aforementioned technologies. In this section, we describe three technologies, LIGA, HEXSIL, and HARPSS, capable of producing structures with vertical dimensions much larger than the lateral dimensions by means of X-ray lithography (LIGA) and DRIE etching (HEXSIL and HARPSS).

LIGA

LIGA is a high-aspect-ratio micromachining process that relies on X-ray lithography and electroplating (in German: *LI*thographie *G*lvanoformung *A*bformung) [5.41, 42]. We already introduced the concept of the plating-through-mask technique in Sect. 5.1.2 (see Fig. 5.10). With standard UV photolithography and photoresists, the maximum thickness achievable is on the order of a few tens of microns and the resulting metal structures show tapered walls. LIGA is a technology based on the same plating-through-mask idea, but can be used to fabricate metal structures of thicknesses ranging from a few microns to a few millimeters with almost vertical sidewalls. This is achieved using X-ray lithography and special photoresists. Due to their short wavelength, X-rays are capable of penetrating a thick photoresist layer with no scattering and defining features with lateral dimensions down to 0.2 µm (aspect ratio > 100 : 1).

The photoresists used in LIGA should comply with certain requirements, including sensitivity to X-rays, resistance to electroplating chemicals, and good adhesion to the substrate. Based on such requirements Poly-(methylmethacrylate) (PMMA) is considered an optimal choice for the LIGA process. Application of the thick photoresist on top of the substrate can be performed by various techniques such as multiple spin coating, pre-cast PMMA sheets, and plasma polymerization coating. The mask structure and materials for X-ray lithography must also comply with certain requirements. The traditional masks based on

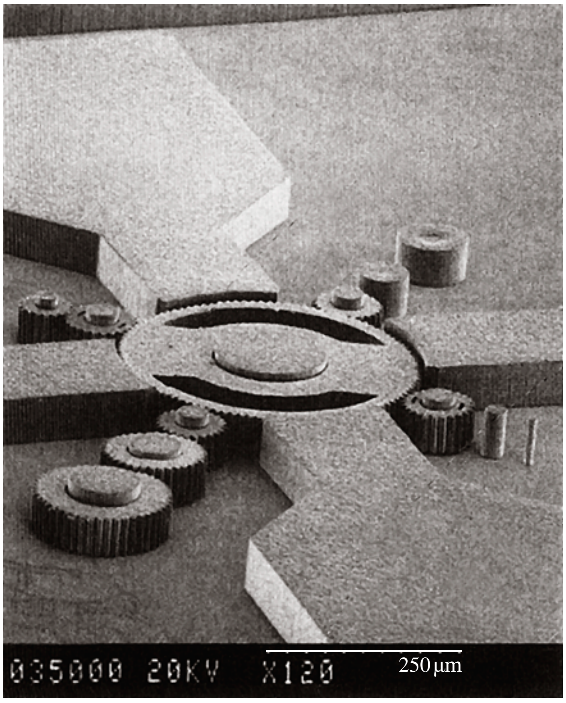

Fig. 5.38 SEM of assembled LIGA-fabricated nickel structures [5.41]

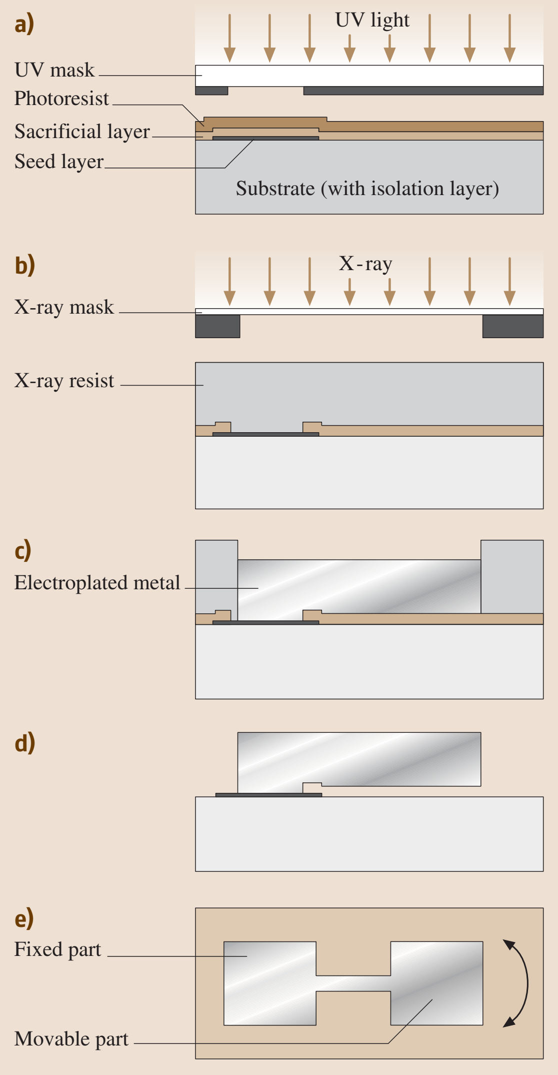

Fig. 5.39a–e Sacrificial LIGA process: **(a)** UV lithography for sacrificial layer patterning, **(b)** X-ray lithography, **(c)** electroplating, **(d)** structure releasing, **(e)** top view of the movable structure

glass plates with a patterned chrome thin layer are not suitable, because X-rays are not absorbed in the chromium layer and the glass plate is not transparent enough. Instead, X-ray lithography uses a silicon nitride mask with gold as the absorber material (typically formed by electroplating gold to a thickness of 10–20 μm). The nitride membrane is supported by a silicon frame, which can be fabricated using bulk micromachining techniques. Once the photoresist is exposed to the X-rays and developed, the process proceeds with the electroplating of the desired metal. Ni is the most common, although other metals and metallic compounds such as Cu, Au, NiFe, and NiW are also electroplated in LIGA processes. A good agitation of the plating solution is the key in obtaining a uniform and repeatable result during this step. A paddle plating cell, based on a windshield wiper-like device moving only 1 mm away from the substrate surface, provides extremely reproducible agitation. Figure 5.38 shows an SEM of a LIGA microstructure fabricated by electroplating nickel.

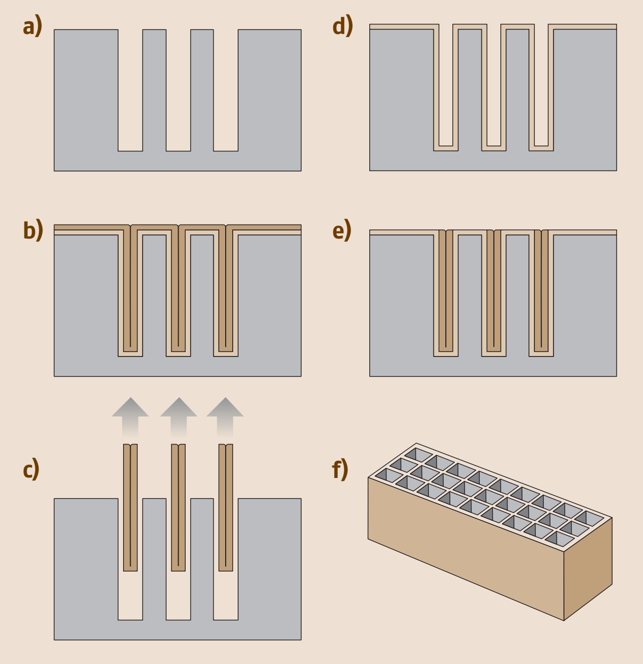

Fig. 5.40a–f HEXSIL process flow: **(a)** DRIE, **(b)** sacrificial layer deposition, **(c)** structural material deposition and trench filling, **(d)** etch structural layer from the surface, **(e)** etch sacrificial layer and pulling out of the structure, **(f)** example of a HEXSIL fabricated structure

Due to the high cost of the X-ray sources (synchrotron radiation), the LIGA technology was initially intended for the fabrication of molds that could be used many times in hot embossing or injection molding processes. However, it has been also used in many applications to directly form high-aspect-ratio metal structures on top of a substrate. A cheaper alternative to the LIGA process (with somewhat poorer qualities) called UV-LIGA, or "poor man's LIGA", has been proposed [5.43, 44]. This process uses SU-8 negative photoresists (available for spin coating at various thicknesses ranging from 1 to 500 μm) and standard contact lithography equipment. Using this technique, aspect ratios larger than 20:1 have been demonstrated. A major problem of this alternative is the removal of the SU-8 photoresist after plating. Various methods with different degrees of success have been proposed. These include: wet etching with special solvents, burning at high temperatures (600 °C), dry etching, use of a release layer, and high-pressure water jet etching.

A variation of the basic LIGA process, shown in Fig. 5.39, permits the fabrication of electrically isolated movable structures, and thus opens more possibilities for sensor and actuator design using this technology [5.45]. The so-called sacrificial LIGA (SLIGA) starts with the patterning of the seed layer. Subsequently, a sacrificial layer (e.g., titanium) is deposited and patterned. The process then proceeds as usual in standard LIGA until the last step when the sacrificial layer is removed. The electroplated structures that overlap with the sacrificial layer are released in this step.

HEXSIL

The second method for fabricating high-aspect-ratio structures, which is based on a template replication technology, is *HEXSIL* (*HEX*agonal honeycomb poly*SIL*icon) [5.46]. Figure 5.40 shows a simplified process flow. A high-aspect-ratio template is first formed in a silicon substrate using DRIE. Next, a sacrificial multilayer is deposited to allow the final release of the structures. The multilayer is composed of one or more PSG nonconformal layers to provide fast etch release (∼ 20 μm/min in 49% HF) alternated with conformal layers of either oxide or nitride to provide enough thickness for proper release of the structures. The total thickness of the sacrificial layer has to be larger than the shrinkage or elongation of the structures caused by the relaxation of the internal stress (compressive or tensile) during the release step. Otherwise, the structures will clamp themselves to the walls of the template and their retrieval will not be possible. Any material

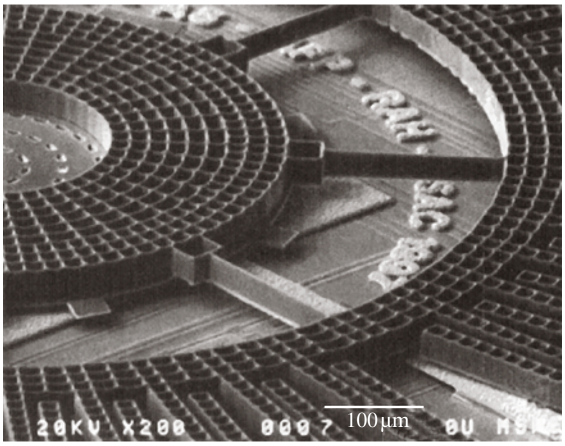

Fig. 5.41 SEM micrograph of an angular microactuator fabricated using HEXSIL [5.48]

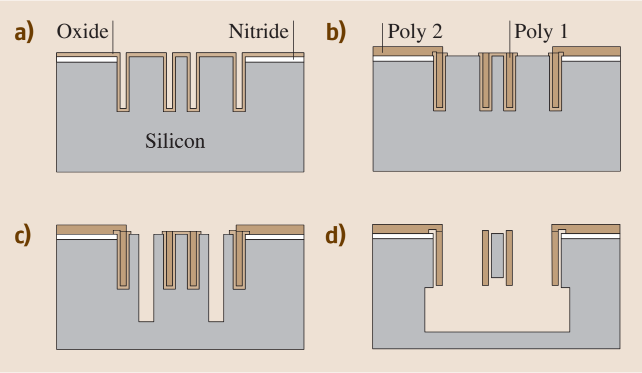

Fig. 5.42a–d HARPSS process flow. **(a)** Nitride deposition and patterning, DRIE etching and oxide deposition, **(b)** poly 1 deposition and etch back, oxide patterning and poly 2 deposition and patterning, **(c)** DRIE etching, **(d)** silicon isotropic etching

that can be conformally deposited and yet not damaged during the HF release step is suitable for the structural layer. Structures made of polysilicon, nitride, and electroless nickel [5.47] have been reported. Nickel can only be deposited in combination with polysilicon since a conductive surface is needed for the deposition to occur. After deposition of structural materials, a blanket etch (poly or nitride) or a mechanical lapping (nickel) is performed to remove the excess materials from the surface. Finally, a 49% HF with surfactant is used to dissolve the sacrificial layers. The process can be repeated many times using the same template, thus considerably lowering the fabrication costs. Figure 5.41 shows an SEM photograph of a microactuator fabricated using the HEXSIL process.

HARPSS

The *h*igh *a*spect *r*atio combined with *p*oly and *s*ingle-crystal *s*ilicon (*HARPSS*) technology is another technique capable of producing high-aspect-ratio, electrically isolated polycrystalline and single-crystal silicon microstructures with capacitive air gaps ranging from submicrometer to tens of micrometers [5.50]. The structures, tens to hundreds of micrometers thick, are defined by trenches etched with DRIE and filled with oxide and poly layers. The release of the microstructures is done at the end by means of a directional silicon etch followed by an isotropic etch. The small, vertical gaps and thick structures possible with this technology can be applied to the fabrication of a variety of MEMS devices, particularly inertial sensors [5.51] and RF beam resonators [5.52]. Figure 5.42 shows the process flow at the cross section of a single-crystal silicon beam resonator. The HARPSS process starts with the deposition and patterning of a silicon nitride layer that will be used to isolate the poly structure's connection pads from the substrate. High-aspect-ratio trenches (∼ 5 µm wide) are then etched into the substrate using a DRIE etcher. Next, a conformal oxide layer (LPCVD) is deposited. This layer has two functions: 1) to protect the structures during the dry etch release and 2) to define the submicrometer gap between silicon and polysilicon structures. Following the oxide deposition, the trenches are completely filled with LPCVD polysilicon. The polysilicon

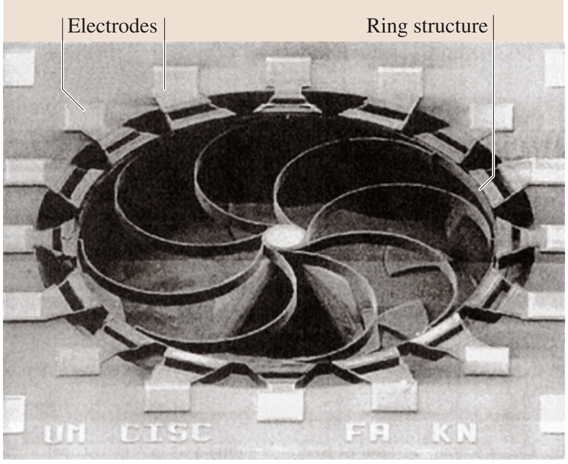

Fig. 5.43 SEM photograph of a micro-gyroscope fabricated using HARPSS process [5.49]

is etched back, and the underneath oxide is patterned to provide anchor points for the structures. A second layer of polysilicon is then deposited and patterned. Finally, the structures are released using a DRIE step followed by an isotropic silicon etch through a photoresist mask that exposes only the areas of silicon substrate surrounding the structures. It should be noted that single-crystal silicon structures are not protected at the bottom during the isotropic etch. This causes the single-crystal silicon structures to be etched vertically from the bottom, and thus be shorter than the polysilicon structures. Figure 5.43 shows an SEM photograph of a micro-gyroscope fabricated using the HARPSS process.

5.3 Nanofabrication Techniques

The microfabrication techniques discussed so far were mostly geared toward fabricating devices in the 1 mm to 1 µm dimensional range (submicron dimensions being possible in certain techniques such as *HARPSS* using a dielectric sacrificial layer). Recent years have witnessed a tremendous surge of interest in fabricating sub-micro (1 µm–100 nm) and nanostructures (1–100 nm range) [5.53]. This interest arises from both practical and fundamental view points. At the more scientific and fundamental level, nanostructures provide an interesting tool in studying electrical, magnetic, optical, thermal, and mechanical properties of matter at the nanometer scale. These include important quantum mechanical phenomena (e.g., conductance quantization, band-gap modification, coulomb blockade, etc.) arising from the confinement of charged carriers in structures such as quantum wells, wires, and dots, Fig. 5.44. On the practical side, nanostructures can provide significant improvements in the performance of electronic/optical devices and sensors. In the device area, investigators have been mostly interested in fabricating nanometer-sized transistors in anticipation of technical difficulties forecasted in extending Moore's law beyond 100 nm resolution. In addition, optical sources and detectors having nanometer-sized dimensions exhibit improved characteristics unachievable in larger devices (e.g., lower threshold current, improved dynamic behavior, and improved emission line-width in quantum dot lasers). These improvements create novel possibilities for next-generation computation and communication devices. In the sensors area, shrinking dimensions beyond conventional optical lithography can provide major improvements in sensitivity and selectivity.

One can broadly divide various nanofabrication techniques into top-down and bottom-up categories. The first approach starts with a bulk or thin film material and removes selective regions to fabricate nanostructures (similar to micromachining techniques). The second method relies on molecular recognition and self-assembly to fabricate nanostructures out of smaller building blocks (molecules, colloids, and clusters). The top-down approach is an offshoot of standard lithography and micromachining techniques. On the other hand, the bottom-up approach has more of a chemical engineering and material science flavor and relies on fundamentally different principles. In this chapter, we will discuss four major nano-fabrication techniques. These include: i) e-beam and *nano-imprint* fabrication, ii) *epitaxy* and *strain engineering*, iii) scanned probe techniques, and iv) *self-assembly* and *template manufacturing*.

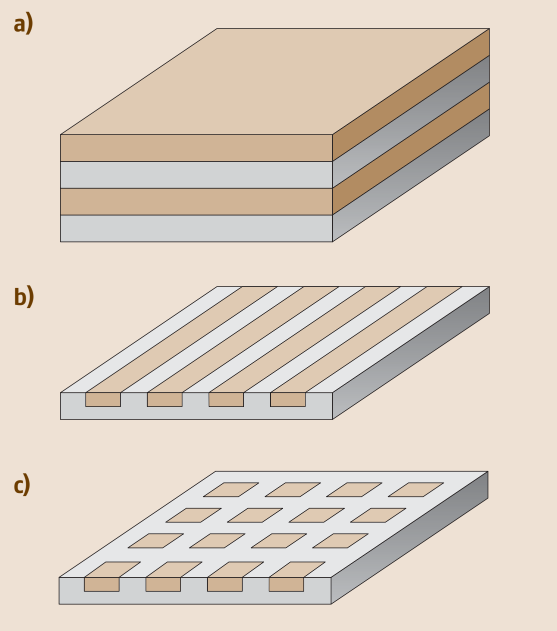

Fig. 5.44a–c Several important quantum confinement structures, **(a)** quantum well, **(b)** quantum wire, and **(c)** quantum dot

5.3.1 E-Beam and Nano-Imprint Fabrication

In previous sections, we discussed several important lithography techniques used commonly in MEMS and microfabrication. These include various forms of UV (regular, deep, and extreme) and X-ray lithographies. However, due to the lack of resolution (in the case of the UV), or the difficulty in manufacturing mask and radiation sources (X-ray), these techniques are not suitable for nanometer-scale fabrication. E-beam lithography is an attractive alternative technique for fabricating nanostructures [5.54]. It uses an electron beam to expose an electron-sensitive resist such as polymethyl methacrylate (PMMA) dissolved in trichlorobenzene (positive) or poly chloromethylstyrene (negative) [5.55]. The e-beam gun is usually part of a scanning electron microscope (SEM), although transmission electron microscopes (TEM) can also be used. Although electron wavelengths on the order of 1 Å can be easily achieved, electron scattering in the resist limits the attainable resolutions to > 10 nm. The beam control and pattern generation are achieved through a computer interface. E-beam lithography is serial and, hence, it has a low throughput. Although this is not a major concern in fabricating devices used in studying fundamental microphysics, it severely limits large scale nanofabrication. E-beam lithography, in conjunction with such processes as lift-off, etching, and electro-deposition, can be used to fabricate various nanostructures.

An interesting new technique that circumvents the serial and low throughput limitations of the e-beam lithography for fabricating nanostructures is the nano-imprint technology [5.56]. This technique uses an e-beam fabricated hard material master (or mold) to stamp and deform a polymeric resist. This is usually followed by a reactive ion etching step to transfer the stamped pattern to the substrate. This technique is economically superior, since a single stamp can be used repeatedly to fabricate a large number of nanostructures. Figure 5.45 shows a schematic illustration of nano-imprint fabrication. First, a hard material (e.g., silicon or SiO_2) stamp is created using e-beam lithography and reactive ion etching. Then, a resist-coated substrate is stamped, and, finally, an anisotropic RIE is performed to remove the resist residue in the stamped area. At this stage, the process is complete and one can either etch the substrate or, if metallic nanostructures are desired, evaporate the metal and perform a lift off. The resist used in nano-imprint technology can be thermal plastic, UV-curable, or thermal-curable polymer. For a thermal plastic resist (e.g., PMMA), the substrate is heated to above the glass transition temperature (T_g) of the polymer before stamping and is cooled to below T_g before the stamp is removed. Similarly, the UV and thermal-curable resists are fully cured before the stamp is separated. Nano-imprint technology resolution is limited by the mold and polymer strengths and can be as small as 10 nm. More recently, the nano-imprint technique has been used to stamp a silicon substrate in less than 250 ns using a XeCl excimer laser (308 nm) and a quartz mask (laser assisted direct imprint, LADI), Fig. 5.46 [5.57].

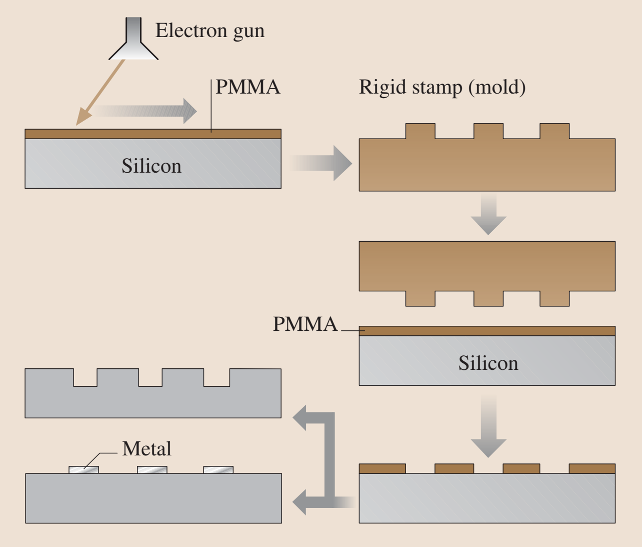

Fig. 5.45 Schematic illustration of nano-imprint fabrication

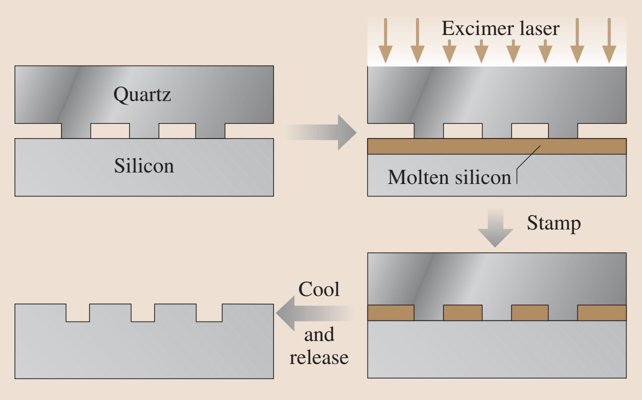

Fig. 5.46 Ultrafast silicon nano-imprinting using an excimer laser

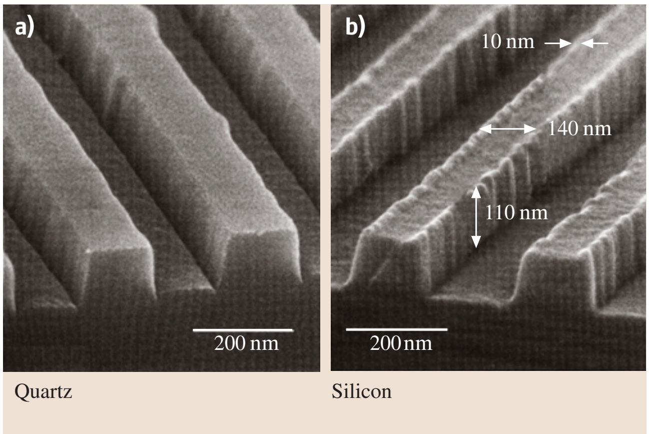

Fig. 5.47a,b SEM micrographs of **(a)** quartz mold and **(b)** imprinted silicon surface using LADI [5.57]

Figure 5.47 shows SEM micrographs of the quartz mold and imprinted silicon substrate with 140-nm lines using LADI.

5.3.2 Epitaxy and Strain Engineering

Atomic precision deposition techniques such as molecular beam epitaxy (MBE) and metallo-organic chemical vapor deposition (MOCVD) have proven to be effective tools in fabricating a variety of quantum confinement structures and devices (quantum well lasers, photodetectors, resonant tunneling diodes, etc.) [5.58–60]. Although quantum wells and superlattices are the structures that lend themselves most easily to these techniques (see Fig. 5.44a), quantum wires and dots have also been fabricated by adding subsequent steps such as etching and selective growth. Fabrication of quantum well and superlattice structures using epitaxial growth is a mature and well developed field and, therefore, will not be discussed in this chapter. Instead, we will concentrate on quantum wire and dot nanostructure fabrication using basic epitaxial techniques [5.61, 62].

Quantum Structure Nanofabrication Using Epitaxy on Patterned Substrates

There have been several different approaches to the fabrication of quantum wires and dots using epitaxial layers. The most straight forward technique involves e-beam lithography and etching of an epitaxial grown layer (e.g., InGaAs on GaAs substrate) [5.63]. However, due to the damage and/or contamination during lithography, this method is not very suitable for active device fabrication (e.g., quantum dot lasers). Several other methods involving regrowth of epitaxial layers over nonplanar surfaces such as step-edge, cleaved-edge, and patterned substrate have been used to fabricate quantum wires and dots without the need for lithography and etching of the quantum confined structure [5.62, 64]. These nonplanar surface templates can be fabricated in a variety of ways such as etching through a mask or cleavage along crystallographic planes. Subsequent epitaxial growth on top of these structures results in a set of planes with different growth rates depending on the geometry or surface diffusion and adsorption effects. These effects can significantly enhance or limit the growth rate on certain planes, resulting in lateral patterning and confinement of deposited epitaxial layers and formation of quantum wires (in V grooves) and dots (in inverted pyramids). Figure 5.48a shows a schematic cross section of an InGaAs quantum wire fabricated in a V-groove InP. As can be seen, the growth rate on the sidewalls is lower than that of the top and bottom surfaces. Therefore, the thicker InGaAs layer at the bottom of the V-groove forms a quantum wire confined from the sides by a thinner layer with a wider bandgap. Figure 5.48b shows a quantum wire formed using epitaxial growth over a dielectric pat-

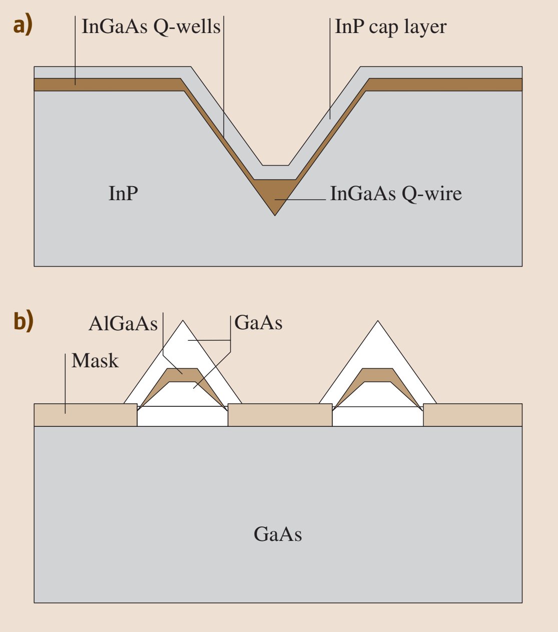

Fig. 5.48 **(a)** InGaAs quantum wire fabricated in a V-groove InP and **(b)** AlGaAs quantum wire fabricated through epitaxial growth on a masked GaAs substrate

terned planar substrate. In both of these techniques, it is relatively easy to create quantum wells. However, in order to create quantum wires and dots one still needs e-beam lithography to pattern the groove and window templates.

Quantum Structure Nanofabrication Using Strain-Induced Self-Assembly

A more recent technique for fabricating quantum wires and dots involves strain-induced *self-assembly* [5.62, 65]. The term *self-assembly* represents a process whereby a strained 2-D system reduces its energy by a transition into a 3-D morphology. The most commonly used material combination for this technique is the $In_xGa_{1-x}As/GaAs$ system, which offers a large lattice mismatch (7.2% between InAs and GaAs) [5.66,67], although recently Ge dots on Si substrate have also attracted considerable attention [5.68]. This method relies on lattice mismatch between an epitaxial grown layer and its substrate to form an array of quantum dots or wires. Figure 5.49 shows a schematic of the strain-induced self-assembly process. When the lattice constants of the substrate and the epitaxial layer differ considerably, only the first few deposited monolayers crystallize in the form of an epitaxial strained layer in which the lattice constants are equal. When a critical thickness is exceeded, a significant strain in the layer leads to the breakdown of this ordered structure and the spontaneous formation of randomly distributed islets of regular shape and similar size (usually < 30 nm in diameter). This mode of growth is usually referred to as the Stranski-Krastanow mode. The quantum dot size, separation, and height depend on the deposition parameters (i. e., total deposited material, growth rate, and temperature) and material combinations. As can be seen, this is a very convenient method to grow perfect crystalline nanostructures over a large area without any lithography and etching. One major drawback of this technique is the randomness of the quantum dot distribution. It should be mentioned that this technique can also be used to fabricate quantum wires by strain relaxation bunching at the step edges.

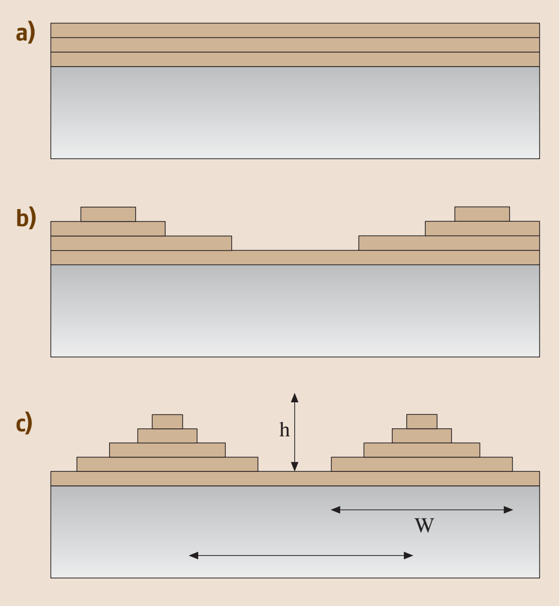

Fig. 5.49a–c Stranski-Krastanow growth mode, **(a)** 2-D wetting layer, **(b)** growth front roughening and breakup, and **(c)** coherent 3-D self-assembly

5.3.3 Scanned Probe Techniques

The invention of scanned probe microscopes in the 1980s revolutionized atomic-scale imaging and spectroscopy. In particular, scanning tunneling and atomic force microscopes (STM and AFM) have found widespread applications in physics, chemistry, material science, and biology. The possibility of atomic-scale manipulation, lithography, and nanomachining using such probes was considered from the beginning and has matured considerably over the past decade. In this section, after a brief introduction to scanned probe microscopes, we will discuss several important nano-lithography and machining techniques that have been used to create nanometer-sized structures.

The scanning probe microscopy (SPM) systems are capable of controlling the movement of an atomically sharp tip in close proximity to or in contact with a surface with subnanometer accuracy. Piezoelectric positioners are typically used to achieve such accuracy. High resolution images can be acquired by raster scanning the tips over a surface while simultaneously monitoring the interaction of the tip with the surface. In scanning tunneling microscope systems, a bias voltage is applied to the sample and the tip is positioned close enough to the surface, so that a tunneling current develops through the gap (Fig. 5.50a). Because this current is extremely sensitive to the distance between the tip and the surface, scanning the tip in the x-y plane while recording the tunnel current permits the mapping of the surface topography with resolution on the atomic scale. In a more common mode of operation, the amplified current signal is connected to the z-axis piezoelectric positioner through

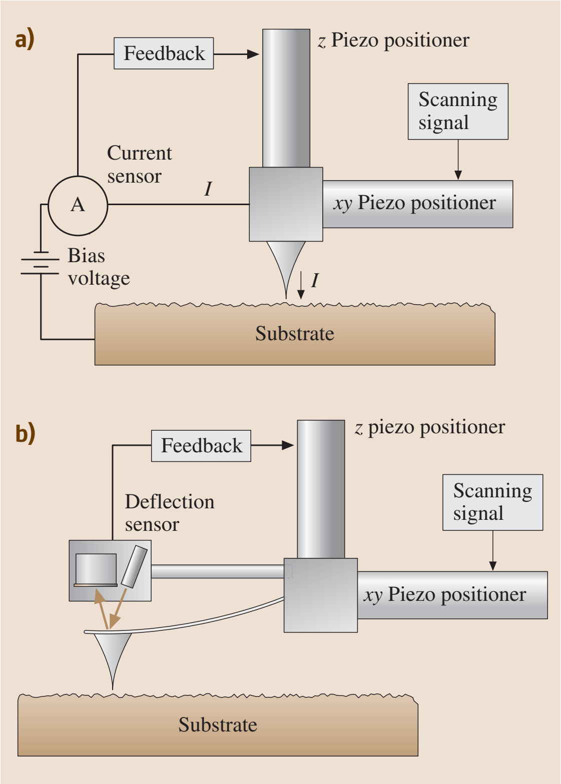

Fig. 5.50a,b Scanning probe systems: **(a)** STM and **(b)** AFM

a feedback loop, so that the current and, therefore, the distance are kept constant throughout the scanning. In this configuration the picture of the surface topography is obtained by recording the vertical position of the tip at each x-y position.

The STM system only works for conductive surfaces because of the need to establish a tunneling current. The atomic force microscopy was developed as an alternative for imaging either conducting or nonconducting surfaces. In AFM, the tip is attached to a flexible cantilever and is brought in contact with the surface (Fig. 5.50b). The force between the tip and the surface is detected by sensing the cantilever deflection. A topographic image of the surface is obtained by plotting the deflection as a function of the x-y position. In a more common mode of operation, a feedback loop is used to maintain a constant deflection, while the topographic information is obtained from the cantilever vertical displacement. Some scanning probe systems use a combination of the AFM and STM modes, i. e., the tip is mounted in a cantilever with an electrical connection, so both the surface forces and the tunneling currents can be controlled or monitored. STM systems can be operated in ultrahigh vacuum (UHV STM) or in air, whereas *AFM* systems are typically operated in air. When a scanning probe system is operated in air, water adsorbed onto the sample surface accumulates underneath the tip, forming a meniscus between the tip and the surface. This water meniscus plays an important role in some of the scanning probe techniques described below.

Scanning Probe Induced Oxidation

Nanometer-scale local oxidation of various materials can be achieved using scanning probes operated in air and biased at a sufficiently high voltage, Fig. 5.51. Tip bias of −2 to −10 V is normally used with writing speeds of 0.1–100 μm/s in an ambient humidity of 20–40 %. It is believed that the water meniscus formed at the contact point serves as an electrolyte such that the biased tip anodically oxidizes a small region of the surface [5.70]. The most common application of this principle is the oxidation of hydrogen-passivated silicon. A dip in HF solution is typically used to passivate the silicon surfaces with hydrogen atoms. Patterns of oxide "written" on a silicon surface can be used as a mask for wet or dry etching. Ten-nm line width patterns have been successfully transferred to a silicon substrate in this fashion [5.71]. Various metals have also been locally anodized using this approach such as alu-

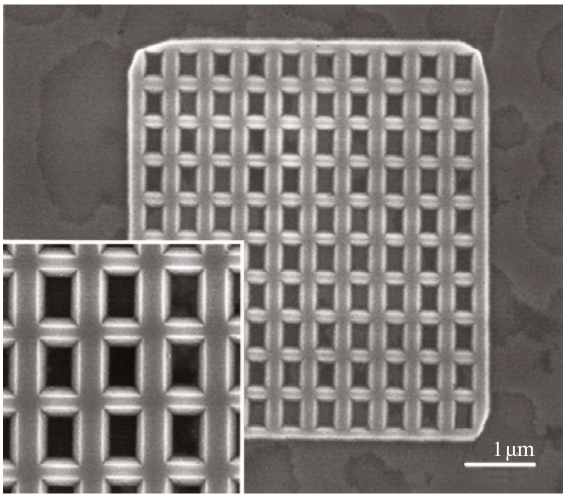

Fig. 5.51 SEM image of an inverted, truncated pyramid array fabricated on a silicon SOI wafer by SPM oxidation and subsequent etch in TMAH (pitch is 500 nm) [5.69]

minum or titanium [5.72]. An interesting variation of this process is the anodization of deposited amorphous silicon [5.73]. Amorphous silicon can be deposited at low temperature on top of many materials. The deposited silicon layer can be patterned and used as, for example, the gate of a 0.1 μm CMOS transistor [5.74], or it can be used as a mask to pattern an underlying film. The major drawback of this technique is poor reproducibility due to tip wear during the anodization. However, using AFM in noncontact mode has overcome this problem [5.70].

Scanning Probe Resist Exposure and Lithography

Electrons emitted from a biased SPM tip can be used to expose a resist the same way e-beam lithography does (Fig. 5.52) [5.74]. Various systems have been used for this lithographic technique. These include constant current STM, noncontact AFM, and AFM with constant tip-resist force and constant current. The systems using AFM cantilevers have the advantage of performing imaging and alignment tasks without exposing the resist. Resists well characterized for e-beam lithography (e.g., PMMA or SAL601) have been used with scanning probe lithography to achieve reliable sub-100-nm lithography. The procedure for this process is as follows. The wafers are cleaned and the native oxide (in the case of silicon or poly) is removed with a HF dip. Subsequently, 35–100 nm-thick resist is spin-coated on top of the surface. The exposure is done by moving the SPM tip over the surface while applying a bias voltage high enough to produce an emission of electrons from the tip (a few tens of volts). Development of the resist is performed in standard solutions following the exposure. Features below 50 nm in width have been achieved with this procedure.

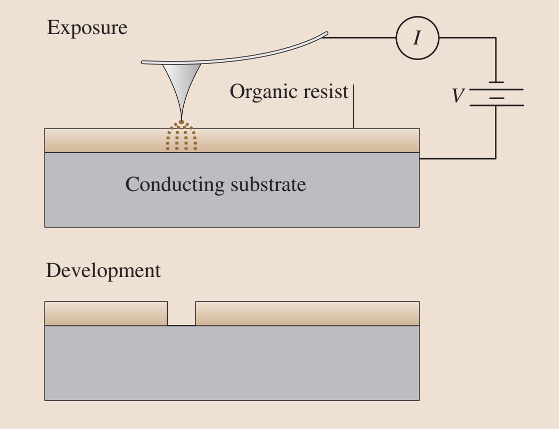

Fig. 5.52 Scanning probe lithography with organic resist

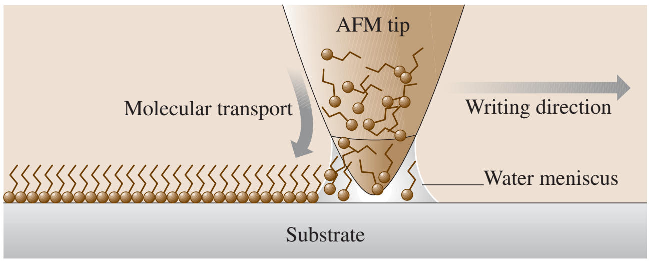

Fig. 5.53 Schematic representation of the dip-pen nanolithography working principle

Dip-Pen Nanolithography

In dip-pen nanolithography (DPN), the tip of an AFM operated in air is "inked" with a chemical of interest and brought into contact with a surface. The ink molecules flow from the tip onto the surface as with a fountain pen. The water meniscus that naturally forms between the tip and the surface enables the diffusion and transport of the molecules, as shown in Fig. 5.53. Inking can be done by dipping the tip in a solution containing a small concentration of the molecules followed by a drying step (e.g., blowing dry with compressed difluoroethane). Line widths down to 12 nm with spatial resolution of 5 nm have been demonstrated with this technique [5.75]. Species patterned with DPN include conducting polymers, gold, dendrimers, DNA, organic dyes, antibodies, and alkanethiols. Alkanethiols have also been used as an organic monolayer mask to etch a gold layer and subsequently etch the exposed silicon substrate.

Other Scanning Probe Nanofabrication Techniques

A great variety of nanofabrication techniques using scanning probe systems have been demonstrated. Some of these are proof of concept demonstrations, and their utility as a viable and repeatable fabrication process has yet to be evaluated. For example, a substrate can be mechanically machined using STM/AFM tips acting as plows or engraving tools [5.76]. This can be used to directly create structures in the substrate, although it is more commonly used to pattern a resist for a subsequent etch, liftoff, or electro-deposition step. Mechanical nanomachining with SPM probes can be facilitated by heating the tip above the glass transition of a polymeric substrate material. This approach has been applied to SPM-based high-density data storage in polycarbonate substrates [5.77].

Electric fields strong enough to induce the emission of atoms from the tip can be easily generated by applying voltage pulses above 3 V. This phenomenon has been used to transfer material from the tip to the surface and vice versa. Ten- to 20-nm mounds of metals such as Au, Ag, or Pt have been deposited or removed from a surface in this fashion [5.78]. The same approach has been used to extract single atoms from a semiconductor surface and re-deposit them elsewhere [5.79]. Manipulation of nanoparticles, molecules, and single atoms on top of a surface has also been achieved by simply pushing or sliding them with the SPM tip [5.80]. Metals can also be locally deposited by the STM chemical vapor deposition technique [5.81]. In this technique, a precursor organometallic gas is introduced in the STM chamber. A voltage pulse applied between the tip and the surface dissociates the precursor gas into a thin layer of metal. Local electro-chemical etch [5.82] and electrodeposition [5.83] are also possible using SPM systems. A droplet of suitable solution is first placed on the substrate. Then the STM tip is immersed into the droplet and a voltage is applied. In order to reduce Faradaic currents the tip is coated with wax such that only the very end is exposed to the solution. Sub-100-nm feature size has been achieved using this technique.

Using a single tip to serially produce the desired modification in a surface leads to very slow fabrication processes that are impractical for mass production. Many of the scanning probe techniques developed thus far, however, could also be performed by an array of tips, which would increase their throughput and make them more competitive compared with other parallel nanofabrication processes. This approach has been demonstrated for imaging, lithography [5.84], and data storage [5.85] using both 1-D and 2-D arrays of scanning probes. With the development of larger arrays with individual control of force, vertical position, and current advances, we might see these techniques become standard fabrication processes in the industry.

5.3.4 Self-Assembly and Template Manufacturing

Self-assembly is a nanofabrication technique that involves aggregation of colloidal nano-particles into the final desired structure [5.86]. This aggregation can be either spontaneous (entropic) and due to the thermodynamic minima (energy minimization) constraints, or chemical and due to the complementary binding of organic molecules and supramolecules (molecular self-assembly) [5.87]. Molecular self-assembly is one of the most important techniques used in biology for the development of complex functional structures. Since these techniques require that the target structures be thermodynamically stable, they tend to produce structures that are relatively defect-free and self-healing. Self-assembly is by no means limited to molecules or the nano-domain and can be carried out on just about any scale, making it a powerful bottom-up assembly and manufacturing method (multiscale ordering). Another attractive feature of this technique relates to the possibility of combining self-assembly properties of organic molecules with the electronic, magnetic, and photonic properties of inorganic components. *Template manufacturing* is another bottom-up technique that utilizes material deposition (electroplating, CVD, etc.) in nano-templates in order to fabricate nanostructures. The nano-templates used in this technique are usually prepared using self-assembly techniques. In the following sections, we will discuss various important self-assembly and template manufacturing techniques currently under heavy investigation.

Physical and Chemical Self-Assembly

The central theme behind the self-assembly process is spontaneous (physical) or chemical aggregation of colloidal nano-particles [5.88]. Spontaneous self-assembly exploits the tendency of mono-dispersed nano- or sub-micro colloidal spheres to organize into a face-centered cubic (FCC) lattice. The force driving this process is the desire of the system to achieve a thermodynamically stable state (minimum free energy). In addition to spontaneous thermal self-assembly, gravitational, convective, and electrohydrodynamic forces can also be used to induce aggregation into complex 3-D structures. Chemical self-assembly requires the attachment of a single molecular organic layer (*self-assembled monolayer*, or *SAM*) to the colloidal particles (organic or inorganic) and subsequent self-assembly of these components into a complex structure using molecular recognition and binding.

Physical Self-Assembly. This is an entropic-driven method that relies on the spontaneous organization of colloidal particles into a relatively stable structure through non-covalent interactions. For example, colloidal polystyrene spheres can be assembled into a 3-D structure on a substrate that is held vertically in the colloidal solution, Fig. 5.54 [5.89, 90]. Upon the

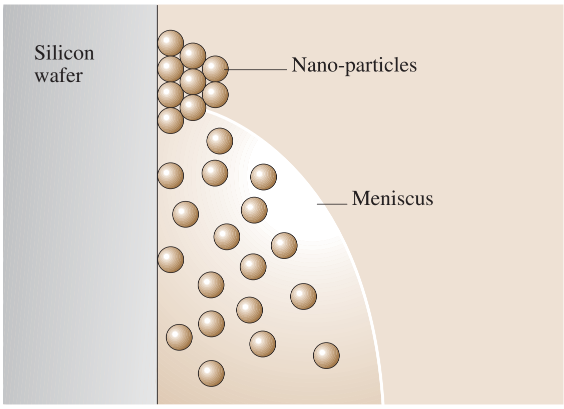

Fig. 5.54 Colloidal particle self-assembly onto solid substrates upon drying in vertical position

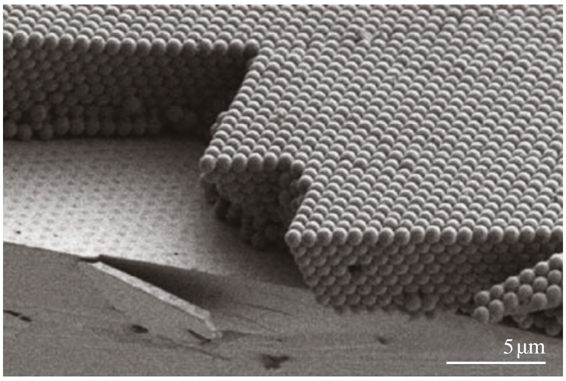

Fig. 5.55 Cross-sectional SEM image of a thin planar opal silica template (spheres 855 nm in diameter) assembled directly on a Si wafer [5.89]

evaporation of the solvent, the spheres aggregate into a hexagonal close packed (HCP) structure. The interstitial pore size and density are determined by the polymer sphere size. The polymer spheres can be etched into smaller sizes after forming the HCP arrays, thereby altering the template pore separations [5.91]. This technique can fabricate large patterned areas in a quick, simple, and cost-effective way. A classic example is the natural assembly of on-chip silicon photonic bandgap crystals [5.89], which are capable of reflecting the light arriving from any direction in a certain wavelength range [5.92]. In this method, a thin layer of silica colloidal spheres is assembled on a silicon substrate. This is achieved by placing a silicon wafer vertically in a vial containing an ethanolic suspension of silica spheres. A temperature gradient across the vial aids the flow of silica spheres. Figure 5.55 shows the cross-sectional SEM image of a thin planar opal template assembled directly on a Si wafer from 855 nm spheres. Once such a template is prepared, LPCVD can be used to fill the interstitial spaces with Si, so that the high refractive index of silicon provides the necessary bandgap.

One can also deposit colloidal particles into a patterned substrate (template-assisted self-assembly, TASA) [5.93, 94]. This method is based on the principle that when aqueous dispersion of colloidal particles is allowed to dewet from a solid surface that is already patterned, the colloidal particles are trapped by the recessed regions and assembled into aggregates of shapes and sizes determined by the geometric confinement provided by the template. The patterned arrays of templates can be fabricated using conventional contact-mode photolithography, which gives control over the shape and dimensions of the templates, thereby allowing the assembly of complex structures from colloidal particles. The cross-sectional view of a fluidic cell used in TASA is shown in Fig. 5.56. The fluidic cell has two parallel glass substrates to confine the aqueous dispersion of the colloidal particles. The surface of the bottom substrate is patterned with a 2-D array of templates. When the aqueous dispersion is allowed to slowly dewet across the cell, the capillary force exerted on the liquid pushes the colloidal spheres across the surface of the bottom substrate until they are physically trapped by the templates. If the concentration of the colloidal dispersion is high enough, the template will be filled by the maximum number of colloidal particles determined by the geometrical confinement. This method can be used to fabricate a variety of polygonal and polyhedral aggregates that are difficult to generate [5.95].

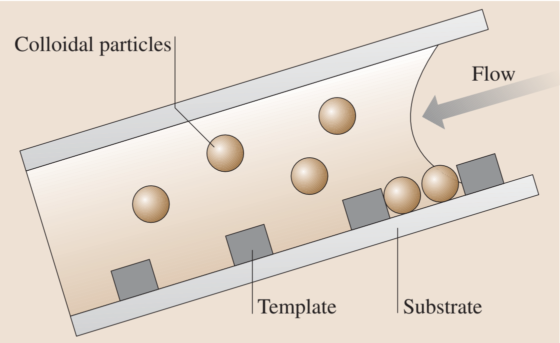

Fig. 5.56 A cross-sectional view of the fluidic cell used for template-assisted self-assembly

Chemical Self-Assembly. Organic and supramolecular SAMs play a critical role in colloidal particle self-assembly. SAMs are robust organic molecules that are chemically adsorbed onto solid substrates [5.96]. Most often, they have a hydrophilic (polar) head that can be bonded to various solid surfaces and a long, hydrophobic (nonpolar) tail that extends outward. SAMs are formed by the immersion of a substrate in a dilute solution of the molecule in an organic solvent. The resulting film is a dense organization of molecules arranged to expose the end group. The durability of a SAM is highly dependent on the effectiveness of the anchoring to the surface of the substrate. SAMs have been widely studied, because the end group can be functionalized to form precisely arranged molecular arrays for various applications ranging from simple, ultrathin insulators and lubricants to complex biological sensors. Chemical self-assembly uses organic or supramolecular SAMs as the binding and recognition sites for fabricating complex 3-D structures from colloidal nano-particles. Most commonly used organic monolayers include: 1) organosilicon compounds on glass and native surface oxide layer of silicon, 2) alkanethiols, dialkyl disulfides, and dialkyl sulfides on gold, 3) fatty acids on alumina and other metal oxides, and 4) DNA.

Octadecyltrichlorosilane (OTS) is the most common organosilane used in the formation of SAMs, mainly because of the fact that it is simple, readily available, and forms good, dense layers [5.97, 98]. Alkyltrichlorosilane monolayers can be prepared on clean silicon wafers whose surface is SiO_2 (with almost 5×10^{14} SiOHgroups/cm^2). Figure 5.57 shows the schematic representation of the formation of alkylsiloxane monolayers by adsorption of alkyltrichlorosilanes from solution onto Si/SiO_2 substrates. Since the silicon-chloride bond is susceptible to hydrolysis, a limited amount of water has to be present in the system in order to obtain good quality monolayers. Monolayers made of methyl- and vinyl-terminated alkylsilanes are autophobic to the hydrocarbon solution and hence emerge uniformly dry from the solution, whereas monolayers made of ester-terminated alkylsilanes emerge wet from the solution used in their formation. The disadvantage of this method is that if the alkyltrichlorosilane in the solvent adhering to the substrate is exposed to water, a cloudy film is deposited on the surface due to formation of a gel of polymeric siloxane.

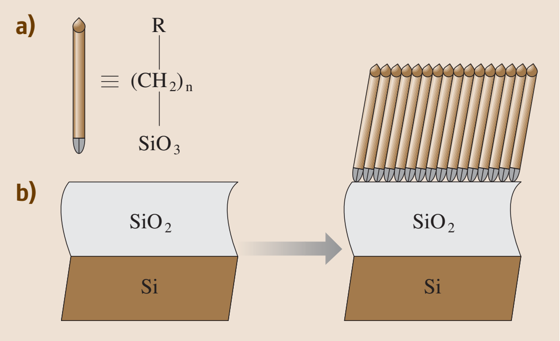

Fig. 5.57 **(a)** Alkylsiloxane formed from the adsorption of alkyltrichlorosilane on Si/SiO_2 substrates. **(b)** Schematic representation of the process

Another important organic SAM system is the alkanethiols ($X(CH_2)_nSH$, where X is the endgroup) on gold [5.96, 99–101]. A major advantage of using gold as the substrate material is that it does not have a stable oxide, and thus it can be handled in ambient conditions. When a fresh, clean, hydrophilic gold substrate is immersed (several minutes to several hours) into a dilute solution (10^{-3} M) of the organic sulfur compound (alkanethiols) in an inorganic solvent, a close-packed, oriented monolayer can be obtained. Sulfur is used as the head group, because of its strong interaction with gold substrate (44 kcal/mol), resulting in the formation of a close-packed, ordered monolayer. The end group of alkanethiol can be modified to render hydrophobic or hydrophilic properties to the adsorbed layer. Another method for depositing alkanethiol SAM is soft-lithography. This technique is based on inking a PDMS stamp with alkanethiol and its subsequent transfer to planar and nonplanar substrates. Alkanethiol functionalized surfaces (planar, nonplanar, spherical) can also be used to self-assemble a variety of intricate 3-D structures [5.102].

Carboxylic acid derivatives self-assemble on surfaces (e.g., glass, Al_2O_3, and Ag_2O) through an acid-base reaction, giving rise to monolayers of fatty acids [5.103]. The time required for the formation of a complete monolayer increases with decreased concentration. A higher concentration of carboxylic acid is required to form a monolayer on gold compared to Al_2O_3. This is due to differences in the affinity of the COOH groups (more affinity to Al_2O_3 and glass than gold) and also the surface concentration of the salt-forming oxides in the two substrates. In the case of amorphous metal oxide surfaces, the chemisorption of alkanoic acids is not unique. For example, on Ag_2O the carboxylate two oxygen atoms bind to the substrate in a nearly symmetrical manner, resulting in ordered monolayers with a chain tilt angle from the surface normally of 15° to 25°. But on CuO and Al_2O_3, the oxygen atoms bind themselves symmetrically and the chain tilt angle is close to 0°. The structure of the monolayers is thus a balance of the various interactions taking place in the polymer chains.

Deoxyribonucleic acid (DNA), the framework on which all life is built, can be used to self-assemble nanomaterials into useful macroscopic aggregates that display a number of desired physical properties [5.104]. DNA consists of two strands, which are coiled around each other to form a double helix. When the two strands are uncoiled, singular strands of nucleotides are left. These nucleotides consist of a sugar (pentose ring) a phosphate (PO_4), and a nitrogenous base. The order and architecture of these components are essential for the proper structure of a nucleotide. There are typically four nucleotides found in DNA: Adenine (A), Guanine (G), Cytosine (C), and Thyamine (T). A key property of the DNA structure is that the described nucleotides bind specifically to another nucleotide when arranged in the two-strand double helix (A to T and C to G). This specific bonding capability can be used to assemble nanophase material and nano-structures [5.105]. For example, nucleotide functionalized nano-gold particles have been assembled into complex 3-D structures by attaching DNA strands to the gold via an enabler or linker [5.106]. In a separate work, DNA was used to assemble nanoparticles into macroscopic materials. This method uses alkane dithiol as the linker molecule to connect the DNA template to the nanoparticle. The thiol groups at each end of the linker molecule covalently attach themselves to the colloidal particles to form aggregate structures [5.107].

Template Manufacturing

Template manufacturing refers to a set of techniques that can be used to fabricate 3-D organic or inorganic structures from a nano-template. These templates differ in material, pattern, feature size, overall template size, and periodicity. Although nano-templates can be fabricated using e-beam lithography, the serial nature of this technique prohibits its widespread application. Self-assembly is the preferred technique that can produce large-area nano-templates in a massively parallel fashion. Several nano-templates have been investigated for use in template manufacturing. These include polymer colloidal spheres, alumina membranes, and nuclear-track etched membranes. Colloidal spheres can be deposited in a regular 3-D array using the techniques described in the previous section (see Figs. 5.54–5.56). Porous aluminum oxide membranes can be fabricated by the anodic oxidation of aluminum [5.108]. The oxidized film consists of columnar arrays of hexagonal, close-packed pores with a separation comparable to the pore size. By controlling the electrolyte species, temperature, anodizing voltage and time, different pore sizes, density and height can be obtained. The pore size and depth can further be adjusted by etching the oxide in an appropriate acid. Templates of porous polycarbonate or mica membranes can be fabricated by nuclear-track etched membranes [5.109]. This technique is based on the passage of high-energy decay fragments from a radioactive source through a dielectric material. The particles leave behind chemically active damaged tracks that subsequently can be etched to create pores throughout the thickness of the membrane [5.110, 111]. Unlike the other methods, the pore separation and, hence, the pore density are independent of the pore size. The pore density is only determined by the irradiation process.

Following template fabrication, the interstitial spaces (in the case of colloidal spheres) or pores (in the case of alumina and polycarbonate membranes) in the template are filled with the desired material [5.91, 112]. This can be done using a variety of deposition techniques such as electroplating and CVD. The final structure can be a composite of nano-template and deposited mater-

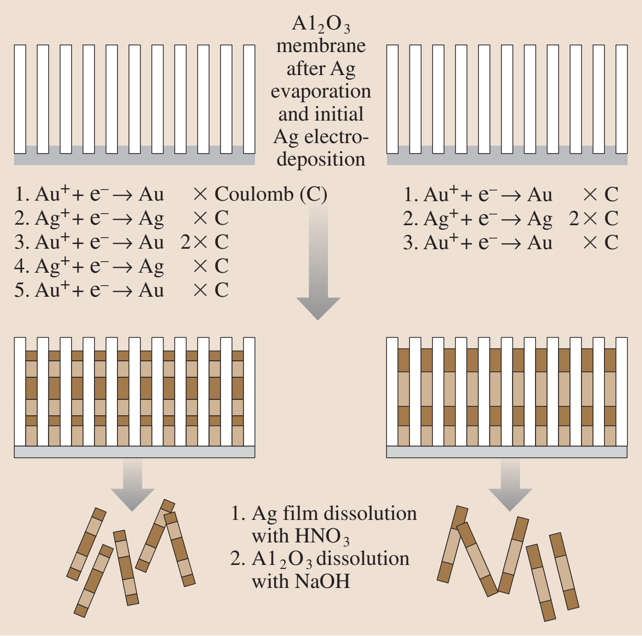
$A1_2O_3$ membrane after Ag evaporation and initial Ag electro-deposition
1. $Au^+ + e^- \rightarrow Au$ × Coulomb (C)
2. $Ag^+ + e^- \rightarrow Ag$ × C
3. $Au^+ + e^- \rightarrow Au$ 2× C
4. $Ag^+ + e^- \rightarrow Ag$ × C
5. $Au^+ + e^- \rightarrow Au$ × C
1. $Au^+ + e^- \rightarrow Au$ × C
2. $Ag^+ + e^- \rightarrow Ag$ 2× C
3. $Au^+ + e^- \rightarrow Au$ × C
1. Ag film dissolution with HNO_3
2. $A1_2O_3$ dissolution with NaOH

Fig. 5.58 Synthesis of nano-bar-code particles

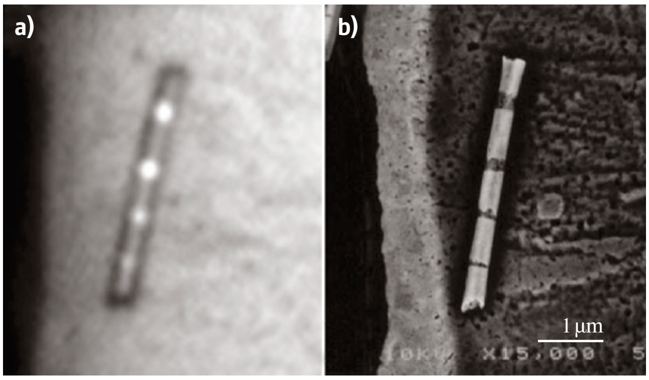

Fig. 5.59 **(a)** Optical and **(b)** FE-SEM images of Au-Ag multistripe particle 5.116

ial, or the template can be selectively etched, resulting in an air-filled 3-D complex structure. For example, nickel [5.111], iron [5.113], and cobalt [5.114] nanowires have been electrochemically grown into porous template matrices. Three-dimensional photonic crystals have been fabricated by electrochemical deposition of CdSe and silicon into polystyrene and silica colloidal assembly templates [5.89, 115]. An interesting example of template-assisted manufacturing is the synthesis of nanometer metallic bar codes [5.116]. These nano-bar codes are prepared by the electrochemical reduction of metallic ions into the pores of an aluminum oxide membrane followed by their release through etching of the template [5.117–119]. This procedure is schematically illustrated in Fig. 5.58. A back-side silver film is used as the working electrode for the reduction of metallic ions (silver and gold in this case) from solution. Up to seven different metallic segments as short as 10 nm and as long as several micrometers with 13 distinguishable stripes have been fabricated using this technique. Optical reflectivity is used to read out the striping pattern encoded in the metal particles [5.116]. Figure 5.59 shows optical and field emission scanning electron microscope images of a Au-Ag multistripe nano-bar code (Ag stripes ranging in length from 60 to 240 nm separated by Au segments of 550 nm). These coded nano-particles can be used in fluorescence and mass-spectrometry-based assays, enabling a wide variety of bioanalytical measurements.

References

5.1 S. A. Campbell: *The Science and Engineering of Microelectronic Fabrication* (Oxford Univ. Press, New York 2001)

5.2 C. J. Jaeger: *Introduction to Microelectronic Fabrication* (Prentice Hall, New Jersey 2002)

5.3 J. D. Plummer, M. D. Deal, P. B. Griffin: *Silicon VLSI Technology* (Prentice Hall, New Jersey 2000)

5.4 J. E. Bjorkholm: EUV lithography: the successor to optical lithography, Intel Technol. J. **2** (1998) 1–8

5.5 J. L. Vossen: *Thin Film Processes* (Academic Press, New York 1976)

5.6 M. Gad-el-Hak (Ed.): *The MEMS Handbook* (CRC Press, Boca Raton 2002)

5.7 T.-R. Hsu: *MEMS and Microsystems Design and Manufacture* (McGraw-Hill, New York 2002)

5.8 G. T. A. Kovacs: *Micromachined Transducers Sourcebook* (McGraw-Hill, New York 1998)

5.9 G. T. A. Kovacs, N. I. Maluf, K. A. Petersen: Bulk micromachining of silicon, Proc. IEEE **86**(8) (1998) 1536–1551

5.10 P. Rai-Choudhury (Ed.): *Handbook of Microlithography, Micromachining and Microfabrication* (SPIE, Bellingham 1997)

5.11 T. J. Cotler, M. E. Elta: Plasma-etch technology, IEEE Circuits & Devices Mag. **6** (1990) 38–43

5.12 U. Gosele, Q. Y. Tong: Semiconductor wafer bonding, Annu. Rev. Mater. Sci. **28** (1998) 215–241

5.13 Q. Y. Tong, U. Gosele: *Semiconductor Wafer Bonding: Science and Technology* (Wiley, New York 1999)

5.14 F. Niklaus, P. Enoksson, E. Kalveston, G. Stemme: Void-free full-wafer adhesive bonding, J. Micromech. Microeng. **11** (2000) 100–107

5.15 C. A. Harper: *Electronic Packaging and Interconnection Handbook* (McGraw-Hill, New York 2000)

5.16 Y. T. Cheng, L. Lin, K. Najafi: Localized silicon fusion and eutectic bonding for MEMS fabrication and packaging, J. Microelectromech. Syst. **9**(1) (2000) 3–8

5.17 W. H. Ko, J. T. Suminto, G. J. Yeh: Bonding techniques for microsensors. In: *Micromachining and Micropackaging for Transducers* (Elsevier, Amsterdam 1985)

5.18 W. P. Maszara, G. Goetz, A. Caviglia, J. B. McKitterick: Bonding of silicon wafers for silicon-on-insulator, J. Appl. Phys. **64**(10) (1988) 4943–4950

5.19 K. Najafi, K. D. Wise, T. Mochizuki: A high-yield IC-compatible multichannel recording array, IEEE Trans. Electron Devices **32** (1985) 1206–1211

5.20 A. Selvakumar, K. Najafi: A high-sensitivity *z* axis capacitive silicon microaccelerometer with a tor-tional suspension, J. Microelectromech. Syst. **7** (1998) 192–200

5.21 H. Baltes, O. Paul, O. Brand: Micromachined thermally based CMOS microsensors, Proc. IEEE **86**(8) (1998) 1660–1678

5.22 B. Eyre, K. S. J. Pister, W. Gekelman: Multi-axis microcoil sensors in standard CMOS, Proc. SPIE Conf. Micromachined Devices and Components, Austin 1995, 183–191

5.23 K. A. Shaw, Z. L. Zhang, N. C. MacDonnald: SCREAM: a single mask single-crystal silicon process for microelectromechanical structures, Proc. IEEE Workshop Micro Electro Mechanical Systems, Fort Lauderdale 1993, 155–160

5.24 G. K. Fedder, S. Santhanam, M. L. Reed, S. C. Eagle, D. F. Guillo, M. S. C. Lu, L. R. Carley: Laminated high-aspect-ratio microstructures in a conventional CMOS process, Proc. IEEE Workshop Micro Electro Mechanical Systems, San Diego 1996, 13–18

5.25 N. C. MacDonald: SCREAM MicroElectroMechanical Systems, Microelectron. Eng. **32** (1996) 51–55

5.26 B. P. Van Drieenhuizen, N. I. Maluf, I. E. Opris, G. T. A. Kovacs: Force-balanced accelerometer with mG resolution fabricated using silicon fusion bonding and deep reactive ion etching, Proc. Int. Conf. Solid-State Sensors and Actuators, Chicago 1997, 1229–1230

5.27 X. Huikai, L. Erdmann, Z. Xu, K. J. Gabriel, G. K. Fedder: Post-CMOS processing for high-aspect-ratio integrated silicon microstructures, J. Microelectromech. Syst. **11** (2002) 93–101

5.28 J. M. Bustillo, R. S. Muller: Surface micromachining for microelectromechanical systems, Proc. IEEE **86**(8) (1998) 1552–1574

5.29 H. C. Nathanson, W. E. Newell, R. A. Wickstrom, J. R. Davis: The resonant gate transistor, IEEE Trans. Electron Devices **14** (1967) 117–133

5.30 R. T. Howe, R. S. Muller: Polycrystalline silicon micromechanical beams, Proc. Electrochem. Soc. Spring Meeting, Montreal 1982, 184–185

5.31 J. A. Geen, S. J. Sherman, J. F. Chang, S. R. Lewis: Single-chip surface-micromachined integrated gyroscope with 50 degrees/hour root Allan variance, IEEE J. Solid-State Circuits **37** (2002) 1860–1866

5.32 P. F. Van Kessel, L. J. Hornbeck, R. E. Meier, M. R. Douglass: A MEMS-based projection display, Proc. IEEE **86**(8) (1998) 1687–1704

5.33 A. E. Franke, D. Bilic, D. T. Chang, P. T. Jones, R. T. Howe, G. C. Johnson: Post-CMOS integration of germanium microstructures, Proc. Micro Electro Mechanical Systems, Orlando 1999, 630–637

5.34 S. Sedky, P. Fiorini, M. Caymax, S. Loreti, K. Baert, L. Hermans, R. Mertens: Structural and mechanical properties of polycrystalline silicon germanium for micromachining applications, J. Microelectromech. Syst. **7** (1998) 365–372

5.35 N. Tas, T. Sonnenberg, H. Jansen, R. Legtenberg, M. Elwenspoek: Stiction in surface micromachining, J. Micromech. Microeng. **6** (1996) 385–397

5.36 R. Maboudian, R. T. Howe: Critical review: adhesion in surface micromechanical structures, J. Vac. Sci. Technol. B **15**(1) (1997) 1–20

5.37 J. H. Smith, S. Montague, J. J. Sniegowski, J. R. Murray, P. J. McWhorter: Embedded micromechanical devices for the monolithic integration of MEMS with CMOS, Proc. Int. Electron Devices Meeting, Washington 1995, 609–612

5.38 R. S. Muller, K. Y. Lau: Surface-micromachined microoptical elements and systems, Proc. IEEE **86**(8) (1998) 1705–1720

5.39 K. S. J. Pister, M. W. Judy, S. R. Burgett, R. S. Fearing: Microfabricated hinges: 1 mm vertical features with surface micromachining, Proc. 6th Int. Conf. Solid-State Sensors and Actuators, San Francisco 1991, 647–650

5.40 L. Y. Lin, S. S. Lee, M. C. Wu, K. S. J. Pister: Micromachined integrated optics for free space interconnection, Proc. IEEE MicroElectroMechanical Systems Workshop, Amsterdam 1995, 77–82

5.41 H. Guckel: High-aspect-ratio micromachining via deep X-ray lithography, Proc. IEEE **86**(8) (1998) 1586–1593

5.42 E. W. Becker, W. Ehrfeld, P. Hagmann, A. Maner, D. Munchmeyer: Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography, galvanoforming, and plastic molding (LIGA process), Microelectron. Eng. **4** (1986) 35–56

5.43 K. Y. Lee, N. LaBianca, S. A. Rishton, S. Zolgharnain, J. D. Gelorme, J. Shaw, T. H. P. Chang: Micromachining applications of a high resolution ultra-thick photoresist, J. Vac. Sci. & Technol. B **13** (1995) 3012–3016

5.44 K. Roberts, F. Williamson, G. Cibuzar, L. Thomas: The fabrication of an array of microcavities utilizing SU-8 photoresist as an alternative 'LIGA' technology, Proc. Thirteenth Biennial University/Government/Industry Microelectronics Symposium, Minnesota 1999 (IEEE, Piscataway 1999) 139–141

5.45 C. Burbaum, J. Mohr, P. Bley, W. Ehrfeld: Fabrication of capacitive acceleration sensors by the LIGA technique, Sensors & Actuators A **27** (1991) 559–563

5.46 C. G. Keller, R. T. Howe: Hexsil bimorphs for vertical actuation, Digest of Technical Papers 8th Int. Conf. Solid-State Sensors and Actuators and Eurosensors IX, Stockholm 1995, 99–102

5.47 C. G. Keller, R. T. Howe: Nickel-filled hexsil thermally actuated tweezers, Digest of Technical Papers 8th Int. Conf. Solid-State Sensors and Actuators and Eurosensors IX, Stockholm 1995, 376–379

5.48 D. A. Horsley, M. B. Cohn, A. Singh, R. Horowitz, A. P. Pisano: Design and fabrication of an angular microactuator for magnetic disk drives, J. Microelectromech. Syst. **7** (1998) 141–148

5.49 N. Yazdi, F. Ayazi, K. Najafi: Micromachined inertial sensors, Proc. IEEE **86** (1998) 1640–1659

5.50 F. Ayazi, K. Najafi: High aspect-ratio combined poly and single-crystal silicon (HARPSS) MEMS technology, J. Microelectromech. Syst. **9** (2000) 288–294

5.51 F. Ayazi, K. Najafi: A HARPSS polysilicon vibrating ring gyroscope, J. Microelectromech. Syst. **10** (2001) 169–179

5.52 Y. S. No, F. Ayazi: The HARPSS process for fabrication of nano-precision silicon electromechanical resonators, Proc. 2001 1st IEEE Conference on Nanotechnology, 2001, 489–494

5.53 G. Timp: *Nanotechnology* (Springer, New York 1998)

5.54 P. Rai-Choudhury (Ed.): *Handbook of Microlithography, Micromachining and Microfabrication* (SPIE, Bellingham 1997)

5.55 L. Ming, C. Bao-qin, Y. Tian-Chun, Q. He, X. Qiuxia: The sub-micron fabrication technology, Proc. 6th Int. Conf. Solid-State and Integrated-Circuit Technol. (IEEE, 2001) 452–455

5.56 S. Y. Chou: Nano-imprint lithography and lithographically induced self-assembly, MRS Bulletin **26** (2001) 512–517

5.57 S. Y. Chou, C. Keimel, J. Gu: Ultrafast and direct imprint of nanostructures in silicon, Nature **417** (2002) 835–837

5.58 M. A. Herman: *Molecular Beam Epitaxy: Fundamentals and Current Status* (Springer, New York 1996)

5.59 J. S. Frood, G. J. Davis, W. T. Tsang: *Chemical Beam Epitaxy and Related Techniques* (Wiley, New York 1997)

5.60 S. Mahajan, K. S. Sree Harsha: *Principles of Growth and Processing of Semiconductors* (McGraw-Hill, New York 1999)

5.61 S. Kim, M. Razegi: Advances in quantum dot structures. In: *Processing and Properties of Compound Semiconductors*, ed. by R. Willardson, H. S. Navawa (Academic Press, New York 2001)

5.62 D. Bimberg, M. Grundmann, N. N. Ledentsov: *Quantum Dot Heterostructures* (Wiley, New York 1999)

5.63 G. Seebohm, H. G. Craighead: Lithography and patterning for nanostructure fabrication. In: *Quantum Semiconductor Devices and Technologies*, ed. by T. P. Pearsall (Kluwer, Boston 2000)

5.64 E. Kapon: Lateral patterning of quantum well heterostructures by growth on nonplanar substrates. In: *Epitaxial Microstructures*, ed. by A. C. Gossard (Academic Press, New York 1994)

5.65 F. Guffarth, R. Heitz, A. Schliwa, O. Stier, N. N. Ledentsov, A. R. Kovsh, V. M. Ustinov, D. Bimberg: Strain engineering of self-organized InAs quantum dots, Phys. Rev. B **64** (2001) 085305(1)–085305(7)

5.66 M. Sugawara: *Self-Assembled InGaAs/GaAs Quantum Dots* (Academic Press, New York 1999)

5.67 B. C. Lee, S. D. Lin, C. P. Lee, H. M. Lee, J. C. Wu, K. W. Sun: Selective growth of single InAs quantum dots using strain engineering, Appl. Phys. Lett. **80** (2002) 326–328

5.68 K. Brunner: Si/Ge nanostructures, Rep. Prog. Phys. **65** (2002) 27–72

5.69 F. S. S. Chien, W. F. Hsieh, S. Gwo, A. E. Vladar, J. A. Dagata: Silicon nanostructures fabricated by scanning probe oxidation and tetra-methyl ammonium hydroxide etching, J. Appl. Phys. **91** (2002) 10044–10050

5.70 M. Calleja, J. Anguita, R. Garcia, K. Birkelund, F. Perez-Murano, J. A. Dagata: Nanometer-scale oxidation of silicon surfaces by dynamic force microscopy: reproducibility, kinetics and nanofabrication, Nanotechnology **10** (1999) 34–38

5.71 E. S. Snow, P. M. Campbell, F. K. Perkins: Nanofabrication with proximal probes, Proc. IEEE **85** (1997) 601–611

5.72 H. Sugimura, T. Uchida, N. Kitamura, H. Masuhara: Tip-induced anodization of titanium surfaces by scanning tunneling microscopy: a humidity effect on nanolithography, Appl. Phys. Lett. **63** (1993) 1288–1290

5.73 N. Kramer, J. Jorritsma, H. Birk, C. Schonenberger: Nanometer lithography on silicon and hydrogenated amorphous silicon with low energy electrons, J. Vac. Sci. & Technol. B **13** (1995) 805–811

5.74 H. T. Soh, K. W. Guarini, C. F. Quate: *Scanning Probe Lithography* (Kluwer, Boston 2001)

5.75 C. A. Mirkin: Dip-pen nanolithography: automated fabrication of custom multicomponent, sub-100 nanometer surface architectures, MRS Bulletin **26** (2001) 535–538

5.76 L. L. Sohn, R. L. Willett: Fabrication of nanostructures using atomic-force-microscope-based lithography, Appl. Phys. Lett. **67** (1995) 1552–1554

5.77 H. J. Mamin, B. D. Terris, L. S. Fan, S. Hoen, R. C. Barrett, D. Rugar: High-density data storage using proximal probe techniques, IBM J. Res. & Dev. **39** (1995) 681–699

5.78 K. Bessho, S. Hashimoto: Fabricating nanoscale structures on Au surface with scanning tunneling microscope, Appl. Phys. Lett. **65** (1994) 2142–2144

5.79 I. W. Lyo, P. Avouris: Field-induced nanometer-to atomic-scale manipulation of silicon surfaces with the STM, Science **253** (1991) 173–176

5.80 M. F. Crommie, C. P. Lutz, D. M. Eigler: Confinement of electrons to quantum corrals on a metal surface, Science **262** (1993) 218–220

5.81 A. de Lozanne: Pattern generation below 0.1 micron by localized chemical vapor deposition with the scanning tunneling microscope,, Japan. J. Appl. Physic **33** (1994) 7090–7093

5.82 L. A. Nagahara, T. Thundat, S. M. Lindsay: Nanolithography on semiconductor surfaces under an etching solution, Appl. Phys. Lett. **57** (1990) 270–272

5.83 T. Thundat, L. A. Nagahara, S. M. Lindsay: Scanning tunneling microscopy studies of semiconductor electrochemistry, J. Vac. Sci. & Technol. A **8** (1990) 539–543

5.84 S. C. Minne, S. R. Manalis, A. Atalar, C. F. Quate: Independent parallel lithography using the atomic force microscope, J. Vac. Sci. & Technol. B **14** (1996) 2456–2461

5.85 M. Lutwyche, C. Andreoli, G. Binnig, J. Brugger, U. Drechsler, W. Haeberle, H. Rohrer, H. Rothuizen, P. Vettiger: Microfabrication and parallel operation of 5 × 5 2D AFM cantilever arrays for data storage and imaging, Proc. MEMS '98 (1998), 8–11

5.86 G. M. Whitesides, B. Grzybowski: Self-assembly at all scales, Science **295** (2002) 2418–2421

5.87 P. Kazmaier, N. Chopra: Bridging size scales with self-assembling supramolecular materials, MRS Bulletin **25** (2000) 30–35

5.88 R. Plass, J. A. Last, N. C. Bartelt, G. L. Kellogg: Self-assembled domain patterns, Nature **412** (2001) 875

5.89 Y. A. Vlasov, X.-Z. Bo, J. G. Sturm, D. J. Norris: On-chip natural self-assembly of silicon photonic bandgap crystals, Nature **414** (2001) 289–293

5.90 C. Gigault, D.-K. Veress, J. R. Dutcher: Changes in the morphology of self-assembled polystyrene microsphere monolayers produced by annealing, J. Colloid Interface Sci. **243** (2001) 143–155

5.91 J. C. Hulteen, P. Van Duyne: Nanosphere lithography: a materials general fabrication process for periodic particle array surfaces, J. Vac. Sci. Technol. A **13** (1995) 1553–1558

5.92 J. D. Joannopoulos, P. R. Villeneuve, S. Fan: Photonic crystals: putting a new twist on light, Nature **386** (1997) 143–149

5.93 T. D. Clark, R. Ferrigno, J. Tien, K. E. Paul, G. M. Whitesides: Template-directed self-assembly of 10-μm-sized hexagonal plates, J. Am. Chem. Soc. **124** (2002) 5419–5426

5.94 S. A. Sapp, D. T. Mitchell, C. R. Martin: Using template-synthesized micro-and nanowires as building blocks for self-assembly of supramolecular architectures, Chem. Mater. **11** (1999) 1183–1185

5.95 Y. Yin, Y. Lu, B. Gates, Y. Xia: Template assisted self-assembly: a practical route to complex aggregates of monodispersed colloids with well-defined sizes, shapes and structures, J. Am. Chem. Soc. **123** (2001) 8718–8729

5.96 J. L. Wilbur, G. M. Whitesides: Self-assembly and self-assembles monolayers in micro and nanofabrication. In: *Nanotechnology*, ed. by G. Timp (Springer, New York 1999)

5.97 S. R. Wasserman, Y. T. Tao, G. M. Whitesides: Structure and reactivity of alkylsiloxane monolayers formed by reaction of alkyltrichlorosilanes on silicon substrates, Langmuir **5** (1989) 1074–1087

5.98 C. P. Tripp, M. L. Hair: An infrared study of the reaction of octadecyltrichlorosilane with silica, Langmuir **8** (1992) 1120–1126

5.99 D. R. Walt: Nanomaterials: top-to-bottom functional design, Nature **1** (2002) 17–18

5.100 J. Noh, T. Murase, K. Nakajima, H. Lee, M. Hara: Nanoscopic investigation of the self-assembly processes of dialkyl disulfides and dialkyl sulfides on Au(111), J. Phys. Chem. B **104** (2000) 7411–7416

5.101 M. Himmelhaus, F. Eisert, M. Buck, M. Grunze: Self-assembly of n-alkanethiol monolayers: a study by IR-visible sum frequency spectroscopy (SFG), J. Phys. Chem. **104** (2000) 576–584

5.102 A. K. Boal, F. Ilhan, J. E. DeRouchey, T. Thurn-Albrecht, T. P. Russell, V. M. Rotello: Self-assembly of nanoparticles into structures spherical and network aggregates, Nature **404** (2000) 746–748

5.103 A. Ulman: *An Introduction to Ultrathin Organic Films: From Langmuir–Blodgett to Self-Assembly* (Academic Press, New York 1991)

5.104 E. Winfree, F. Liu, L. A. Wenzler, N. C. Seeman: Design and self-assembly of two-dimensional DNA crystals, Nature **394** (1998) 539–544

5.105 J. H. Reif, T. H. LaBean, N. C. Seeman: Programmable assembly at the molecular scale: self-assembly of DNA lattices, Proc. 2001 IEEE Int. Conf. Robotics and Automation (IEEE, 2001) 966–971

5.106 A. P. Alivisatos, K. P. Johnsson, X. Peng, T. E. Wilson, C. J. Loweth, M. P. Bruchez Jr, P. G. Schultz: Organization of 'nanocrystal molecules' using DNA, Nature **382** (1996) 609–611

5.107 C. Y. Cao, R. Jin, C. A. Mirkin: Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection, Science **297** (2002) 1536–1540

5.108 H. Masuda, H. Yamada, M. Satoh, H. Asoh: Highly ordered nanochannel-array architecture in anodic alumina, Appl. Phys. Lett. **71** (1997) 2770–2772

5.109 R. L. Fleischer: *Nuclear Tracks in Solids: Principles and Applications* (Univ. of California Press, Berkeley 1976)

5.110 R. E. Packard, J. P. Pekola, P. B. Price, R. N. R. Spohr, K. H. Westmacott, Y. Q. Zhu: Manufacture observation and test of membranes with locatable single pores, Rev. Sci. Instrum. **57** (1986) 1654–1660

5.111 L. Sun, P. C. Searson, C. L. Chien: Electrochemical deposition of nickel nanowire arrays in single-crystal mica films, Appl. Phys. Lett. **74** (1999) 2803–2805

5.112 Y. Du, W. L. Cai, C. M. Mo, J. Chen, L. D. Zhang, X. G. Zhu: Preparation and photoluminescence of alumina membranes with ordered pore arrays, Appl. Phys. Lett. **74** (1999) 2951–2953

5.113 S. Yang, H. Zhu, D. Yu, Z. Jin, S. Tang, Y. Du: Preparation and magnetic property of Fe nanowire array, J. Magn. Mater. **222** (2000) 97–100

5.114 M. Sun, G. Zangari, R. M. Metzger: Cobalt island arrays with in-plane anisotropy electrodeposited in highly ordered alumina, IEEE Trans. Magn. **36** (2000) 3005–3008

5.115 P. V. Braun, P. Wiltzius: Electrochemically grown photonic crystals, Nature **402** (1999) 603–604

5.116 S. R. Nicewarner-Pena, R. G. Freeman, B. D. Reiss, L. He, D. J. Pena, I. D. Walton, R. Cromer, C. D. Keating, M. J. Natan: Submicrometer metallic barcodes, Science **294** (2001) 137–141

5.117 D. Almalawi, C. Z. Ziu, M. Moskovits: Nanowires formed in anodic oxide nanotemplates, J. Mater. Res. **9** (1996) 1014

5.118 J. C. Hulteen, C. R. Martin: A general template-based method for the preparation of nanomaterials, J. Mater. Chem. **7** (1997) 1075–1087

5.119 B. R. Martin, D. J. Dermody, B. D. Reiss, M. Fang, L. A. Lyon, M. J. Natan, T. E. Mallouk: Orthogonal self-assembly on colloidal gold-platinum nanorods, Adv. Mater. **11** (1999) 1021–1025

6. Stamping Techniques for Micro and Nanofabrication: Methods and Applications

This chapter highlights some recent advances in high resolution printing methods, in which a "stamp" forms a pattern of "ink" on the surface it contacts. It focuses on two approaches whose capabilities, level of development, and demonstrated applications indicate a strong potential for widespread use, especially in areas where conventional methods are unsuitable. The first of these, known as microcontact printing, uses a high resolution rubber stamp to print patterns of chemical inks, mainly those that lead to the formation of organic self-assembled monolayers (SAMs). These printed SAMs can be used either as resists in selective wet etching, or as templates in selective deposition to form structures of a variety of materials. The other approach, referred to as nanotransfer printing, uses similar high resolution stamps, but ones inked with solid thin film materials. In this case, SAMs, or other types of surface chemistries, bond these films to a substrate that the stamp contacts. The material transfer that results upon removal of the stamp forms a pattern in the geometry of the relief features, in a purely additive fashion. In addition to providing detailed descriptions of these micro/nanoprinting techniques, this chapter illustrates their use in some areas where these methods may provide attractive alternatives to more established lithographic methods.

6.1 **High Resolution Stamps** 186
6.2 **Microcontact Printing** 187
6.3 **Nanotransfer Printing** 190
6.4 **Applications** 193
6.4.1 Unconventional Electronic Systems 193
6.4.2 Lasers and Waveguide Structures 198
6.5 **Conclusions** 200
References 200

The demonstrator applications span fields as diverse as biotechnology (intravascular stents), fiber optics (tunable fiber devices), nanoanalytical chemistry (high resolution nuclear magnetic resonance), plastic electronics (paper-like displays), and integrated optics (distributed feedback lasers). The growing interest in nanoscience and nanotechnology motivates research and the development of new methods that can be used for nanofabricating the relevant test structures or devices. The attractive capabilities of the techniques described here, together with the interesting and subtle materials science, chemistry, and physics associated with them, make this a promising area for basic and applied study.

There is considerable interest in methods for building structures that have micron or nanometer dimensions. Historically, research and development in this area has been driven mainly by the needs of the microelectronics industry. The spectacularly successful techniques that have emerged from those efforts – photolithography, electron beam lithography, etc. – are extremely well suited to the tasks for which they were principally designed: forming structures of radiation sensitive materials (e.g., photoresists or electron beam resists) on ultra-flat glass or semiconductor surfaces. Significant challenges exist in adapting these methods for new emerging applications and areas of research that require patterning of unusual systems and materials (e.g., those in biotechnology, plastic electronics, etc.), structures with nanometer dimensions (i. e., below 50–100 nm), large areas in a single step (i. e., larger than a few square centimeters), or non-planar (i. e., rough or curved) surfaces. These established techniques also have the disadvantage of high capital and operational costs. As a result, some of the oldest and conceptually simplest forms of lithography – embossing, molding, stamping, writing, etc. – are now being re-examined for their potential to serve as the basis for nanofabrica-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

tion techniques that can avoid these limitations [6.1]. Considerable progress has been made in the last few years, mainly by combining these approaches or variants of them with new materials, chemistries, and processing techniques. This chapter highlights some recent advances in high resolution printing methods, in which a "stamp" forms a pattern of "ink" on a surface that it contacts. It focuses on approaches whose capabilities, level of development, and demonstrated applications indicate a strong potential for widespread use, especially in areas where conventional methods are unsuitable.

Contact printing involves the use of an element with surface relief (i. e., the stamp) for transferring material applied to its surface (i. e. the ink) to locations on a substrate that it contacts. The printing press, one of the earliest manufacturable implementations of this approach, was introduced by Gutenberg in the fifteenth century. The general approach has since been used almost exclusively for producing printed text or images with features that are one hundred microns or larger in their minimum dimension. The resolution is determined by the nature of the ink and its interaction with the stamp and/or substrate, the resolution of the stamp, and the processing conditions that are used for printing or converting the pattern of ink into a pattern of functional material. This chapter focuses on (1) printing techniques that are capable of micron and nanometer resolution, and (2) their use for fabricating key elements of active electronic or optical devices and subsystems. It begins with an overview of some methods for fabricating high resolution stamps and then illustrates two different ways that these stamps can be used to print patterns of functional materials. Applications that highlight the capabilities of these techniques and the performance of systems that are constructed with them are also presented.

6.1 High Resolution Stamps

The printing process can be separated into two parts: fabrication of the stamp and the use of this stamp to pattern features defined by the relief on its surface. These two processes are typically quite different, although it is possible in some cases to use patterns generated by a stamp to produce a replica of that stamp. The structure from which the stamp is derived, which is known as the "master", can be fabricated with any technique that is capable of producing well-defined structures of relief on a surface. This master can then be used directly as the stamp, or to produce stamps via molding or printing procedures. It is important to note that the technique for producing the master does not need to be fast or low in cost. It also does not need to possess many other characteristics that might be desirable for a given patterning task: It is used just once to produce a master, which is directly or indirectly used to fabricate stamps. Each one of these stamps can then be used many times for printing.

In a common approach for the high resolution techniques that are the focus of this chapter, an established lithographic technique, such as one of those developed for the microelectronics industry, defines the master. Figure 6.1 schematically illustrates typical processes. Here, photolithography patterns a thin layer of resist on a silicon wafer. Stamps are generated from this structure in one of two ways: By casting against this master, or by etching the substrate with the patterned resist as a mask. In the first approach, the master itself can be used multiple times to produce many stamps, typically using a light or heat-curable prepolymer. In the second, the etched substrate serves as the stamp. Additional stamps can be generated either by repeating the lithography and etching, or by using the original stamp to print replica stamps. For minimum lateral feature sizes greater than $\sim 1-2$ microns, contact or proximity mode photolithography with a mask produced by direct write photolithography represents a convenient method to fabricate the master. For features smaller than ~ 2 microns, several different techniques can be used [6.2], including: (1) projection mode photolithography [6.3], (2) direct write electron beam (or focused ion beam) lithography [6.4, 5], (3) scanning probe lithography [6.6–9] or (4) laser interference lithography [6.10]. The first approach requires a photomask generated by some other method, such as direct write photolithography or electron beam lithography. The reduction (typically 4×) provided by the projection optics relaxes the resolution requirements on the mask and enables features as small as ~ 90 nm when deep ultraviolet radiation and phase shifting masks are used. The costs for these systems are, however, very high and their availability for general research purposes is limited. The second method is flexible in the geometry of patterns that can be produced, and the writing sys-

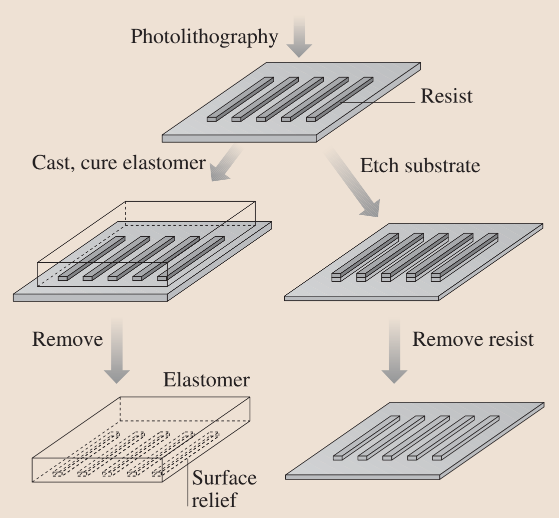

Fig. 6.1 Schematic illustration of two methods for producing high resolution stamps. The first step involves patterning a thin layer of some radiation sensitive material, known as the resist, on a flat substrate, such as a silicon wafer. It is convenient to use an established technique, such as photolithography or electron beam lithography, for this purpose. This structure, known as the master, is converted to a stamp either by etching or molding. In the first case, the resist acts as a mask for etching the underlying substrate. Removing the resist yields a stamp. This structure can be used directly as a stamp to print patterns, or to produce additional stamps. In the molding approach, a prepolymer is cast against the relief structure formed by the patterned resist on the substrate. Curing (thermally or optically) and then peeling the resulting polymer away from the substrate yields a stamp. In this approach, many stamps can be made with a single master, and each stamp can be used many times

tems are highly developed: 30–50 nm features can be achieved with commercial systems [6.11], and < 10 nm features are possible with research tools, as first demonstrated more than 25 years ago by *Broers* [6.12]. The main drawbacks of this method are that it is relatively slow and it is difficult to pattern large areas. Like projection mode photolithography, it can be expensive. The third method, scanning probe lithography, is quite powerful in principle, but the tools are not as well established as those for other approaches. This technique has atomic resolution, but its writing speed can be lower and the areas that can be patterned are smaller than electron beam systems. Interference lithography provides a powerful, low cost tool for generating periodic arrays of features with dimensions down to 100–200 nm; smaller sizes demand ultraviolet lasers, and patterns with aperiodic or non-regular features are difficult to produce.

6.2 Microcontact Printing

Microcontact printing (μCP) [6.13] is one of several soft lithographic techniques – replica molding, micromolding in capillaries, microtransfer molding, near-field conformal photolithography using an elastomeric phase-shifting mask, etc. – that have been developed as alternatives to established methods for micro- and nanofabrication [6.14–18]. μCP uses an elastomeric element (usually polydimethylsiloxane – PDMS) with high resolution features of relief as a stamp to print patterns of chemical inks. It was mainly developed for use with inks that form self-assembled monolayers (SAMs) of alkanethiolates on gold and silver. The procedure for carrying out μCP in these systems is remarkably simple: A stamp, inked with a solution of alkanethiol, is brought into contact with the surface of a substrate to transfer ink molecules to regions where the stamp and substrate contact. The resolution and effectiveness of μCP rely on conformal contact between the stamp and the surface of the substrate, rapid formation of highly ordered monolayers [6.19], and autophobicity of the SAM, which effectively blocks the reactive spreading of the ink across the surface [6.20]. It can pattern SAMs over relatively large areas (∼ up to 0.25 ft^2 have been demonstrated in prototype electronic devices) in a single impression [6.21]. The edge resolution of SAMs printed onto thermally evaporated gold films is on the order of 50 nm, as determined by lateral force microscopy [6.22]. Microcontact printing has been used with a range of different SAMs on various substrates [6.14]. Of these, alkanethiolates on gold, silver, and palladium [6.23] presently give the highest resolution. In many cases, the mechanical properties of the stamp limit the sizes of the smallest features that can be achieved: The most commonly used elastomer (Sylgard 184, Dow Corning) has a low modulus, which can lead to mechanical collapse or sagging for features of relief with aspect

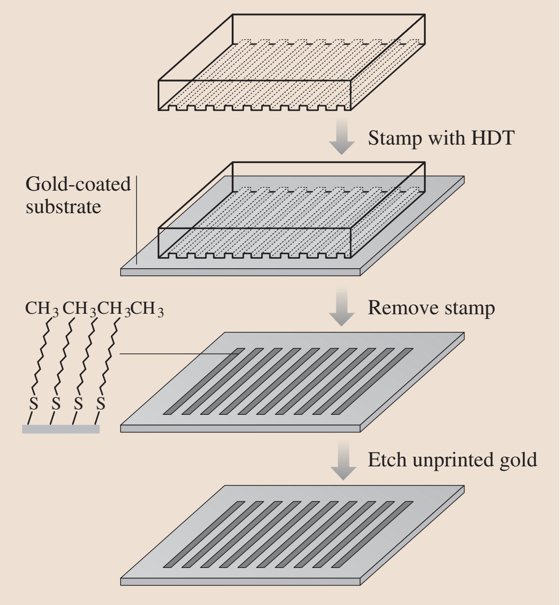

Fig. 6.2 Schematic illustration of microcontact printing. The first step involves inking a stamp with a solution of a material that is capable of forming a self-assembled monolayer (SAM) on a substrate that will be printed. In the case illustrated here, the ink is a millimolar concentration of hexadecanethiol (HDT) in ethanol. Directly applying the ink to the surface of the stamp with a pipette prepares the stamp for printing. Blowing the surface of the stamp dry and placing it on a substrate delivers the ink to areas where the stamp contacts the substrate. The substrate consists of a thin layer of Au on a flat support. Removing the stamp after a few seconds of contact leaves a patterned SAM of HDT on the surface of the Au film. The printed SAM can act as a resist for the aqueous-based wet etching of the exposed regions of the Au. The resulting pattern of conducting gold can be used to build devices of various types

ratios greater than ∼ 2 or less than ∼ 0.05. Stamps fabricated with high modulus elastomers avoid some of these problems [6.24, 25]. Conventional stamps are also susceptible to in-plane mechanical strains that can cause distortions in the printed patterns. Composite stamps that use thin elastomer layers on stiff supports are effective at minimizing this source of distortion [6.26]. Methods for printing that avoid direct mechanical manipulation of the stamp can reduce distortions with conventional and composite stamps [6.21]. This approach has proven effective in large area flexible circuit applications that require accurate multilevel registration.

The patterned SAM can be used either as a resist in selective wet etching or as a template in selective deposition to form structures of a variety of materials: metals, silicon, liquids, organic polymers, and even biological species. Figure 6.2 schematically illustrates the use of µCP and wet etching to pattern a thin film of Au. Figure 6.3 shows SEM images of nanostructures of gold (20 nm thick, thermally evaporated with a 2.5 nm layer of Ti as an adhesion promoter) and silver (∼ 100 nm thick formed by electroless deposition using commercially available plating baths) [6.27] that were fabricated using this approach. In the first and second examples, the masters for the stamps consisted of photoresist patterned on silicon wafers with projection and contact mode photolithography, respectively. Placing these masters in a desiccator for ∼ 1 h with a few drops of tridecafluoro-1,1,2,2-tetrahydrooctyl-1-trichlorosilane forms a silane monolayer on the exposed native oxide of the silicon. This monolayer prevents adhesion of the master to PDMS (Sylgard 184), which is cast and cured from a 10:1 mixture of prepolymer and curing agent. Placing a few drops of a ∼ 1 mM solution of hexadecanethiol (HDT) in ethanol on the surface of the stamps, and then blowing them dry with a stream of nitrogen prepares them for printing. Contacting the metal film for a few seconds with the stamp produces a patterned self-assembled monolayer (SAM) of HDT. An aqueous etchant (1 mM $K_4Fe(CN)_6$, 10 mM $K_3Fe(CN)_6$, and 0.1 M $Na_2S_2O_3$) removes the unprinted regions of the silver [6.28]. A similar solution (1 mM $K_4Fe(CN)_6$, 10 mM $K_3Fe(CN)_6$, 1.0 M KOH, and 0.1 M $Na_2S_2O_3$)

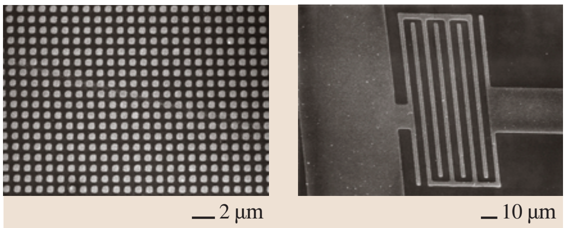

Fig. 6.3 Scanning electron micrographs of typical structures formed by microcontact printing a self-assembled monolayer ink of hexadecanethiol onto a thin metal film followed by etching of the unprinted areas of the film. The *left frame* shows an array of Au (20 nm-thick) dots with ∼ 500 nm diameters. The *right frame* shows a printed structure of Ag (100 nm thick) in the geometry of interdigitated source/drain electrodes for a transistor in a simple inverter circuit. The edge resolution of patterns that can be easily achieved with microcontact printing is ∼ 50–100 nm

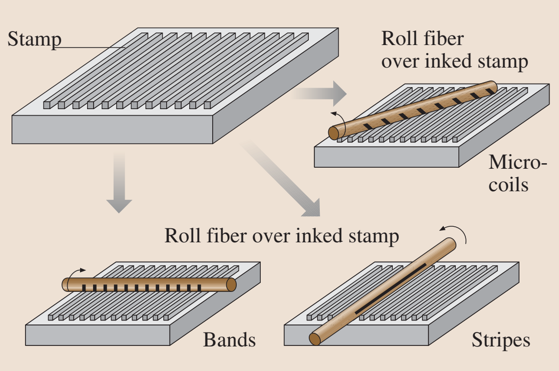

Fig. 6.4 Schematic illustration of a simple method to print lines on the surfaces of optical fibers. Rolling a fiber over the inked stamp prints a pattern onto the fiber surface. Depending on the orientation of the fiber axis with the line stamp illustrated here, it is possible in a single rotation of the fiber to produce a continuous microcoil, or arrays of bands or stripes

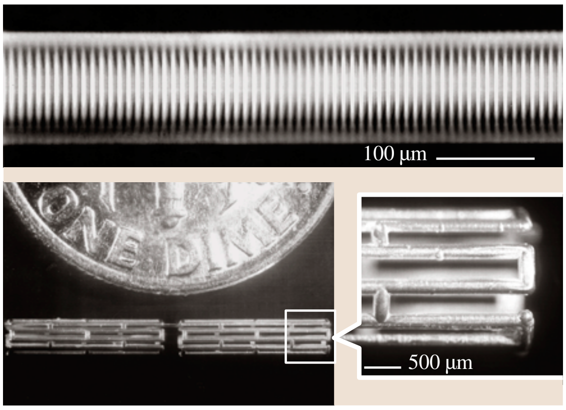

Fig. 6.5 Optical micrographs of some three-dimensional microstructures formed by microcontact printing on curved surfaces. The *top frame* shows an array of 3-micron lines of Au (20 nm)/Ti (1.5 nm) printed onto the surface of an optical fiber. This type of structure can be used as an integrated photomask for producing mode-coupling gratings in the core of the fiber. The *bottom frames* show a freestanding metallic microstructure formed by (1) microcontact printing and etching a thin (100 nm-thick) film of Ag on the surface of a glass microcapillary tube, (2) electroplating the Ag to increase its thickness (to tens of microns) and (3) etching away the glass microcapillary with concentrated hydrofluoric acid. The structure shown here has the geometry and mechanical properties of an intravascular stent, which is a biomedical device commonly used in balloon angioplasty

can be used for etching the bare gold [6.29]. The results of Fig. 6.3 show that the roughness on the edges of the patterns is ∼ 50–100 nm. The resolution is determined by the grain size of the metal films, the isotropic etching process, slight reactive spreading of the inks, and edge disorder in the patterned SAMs.

The structures of Fig. 6.3 were formed on the flat surfaces of silicon wafers (left image) and glass slides (right image). An attractive feature of µCP and certain other contact printing techniques is their ability to pattern features with high resolution on highly curved or rough surfaces [6.18, 30, 31]. This type of patterning task is difficult or impossible to accomplish with photolithography due to its limited depth of focus and the difficulty of casting uniform films of photoresist on non-flat surfaces. Figure 6.4 shows, as an example, a straightforward approach for high resolution printing on the highly curved surfaces of optical fibers. Here, simply rolling the fiber over an inked stamp prints a pattern on the entire outer surface of the fiber. Simple staging systems allow alignment of features to the fiber axis; they also ensure registration of the pattern from one side of the fiber to the other [6.16]. Figure 6.5 shows 3 micron-wide lines and spaces printed onto the surface of a single mode optical fiber (diameter 125 µm). The bottom frame shows a freestanding metallic structure with the geometry and mechanical properties of an intravascular stent, which is a biomedical device that is commonly used in balloon angioplasty procedures. In this latter case µCP followed by electroplating generated the Ag microstructure on a sacrificial glass cylinder that was subsequently etched away with concentrated hydrofluoric acid [6.32]. Other examples of microcontact printing on non-flat surfaces (i. e., low cost plastic sheets and optical ridge waveguides) appear in the applications section of this chapter.

6.3 Nanotransfer Printing

Nanotransfer printing (nTP) is a more recent high resolution printing technique, which uses surface chemistries as interfacial "glues" and "release" layers (rather than "inks", as in µCP) to control the transfer of solid material layers from relief features on a stamp to a substrate [6.33–35]. This approach is purely additive (i. e., material is only deposited in locations where it is needed), and it can generate complex patterns of single or multiple layers of materials with nanometer resolution over large areas in a single process step. It does not suffer from surface diffusion or edge disorder in the patterned inks of µCP, nor does it require post-printing etching or deposition steps to produce structures of functional materials. The method involves four components: (1) a stamp (rigid, flexible, or elastomeric) with relief features in the geometry of the desired pattern, (2) a method for depositing a thin layer of solid material onto the raised features of this stamp, (3) a means of bringing the stamp into intimate physical contact with a substrate, and (4) surface chemistries that prevent adhesion of the deposited material to the stamp and promote its strong adhesion to the substrate. nTP has been demonstrated with SAMs and other surface chemistries for printing onto flexible and rigid substrates with hard inorganic and soft polymer stamps. Figure 6.6 presents a set of procedures for using nTP to pattern a thin metal bilayer of Au/Ti with a surface transfer chemistry that relies on a dehydration reaction [6.33]. The process begins with fabrication of a suitable stamp. Elastomeric stamps can be built using the same casting and curing procedures described for µCP. Rigid stamps can be fabricated by (1) patterning resist (e.g., electron beam resist or photoresist) on a substrate (e.g., Si or GaAs), (2) etching the exposed regions of the substrate with an anisotropic reactive ion etch, and (3) removing the resist, as illustrated in Fig. 6.1. For both types of stamps, careful control of the lithography and the etching steps yields features of relief with nearly vertical or slightly re-entrant sidewalls. The stamps typically have depths of relief > 0.2 µm for patterning metal films with thicknesses < 50 nm.

Electron beam evaporation of Au (20 nm; 1 nm/s) and Ti (5 nm; 0.3 nm/s) generates uniform metal bilayers on the surfaces of the stamp. A vertical, collimated flux of metal from the source ensures uniform deposition only on the raised and recessed regions of relief. The gold adheres poorly to the surfaces of stamps made of GaAs, PDMS, glass, or Si. In the process of Fig. 6.6, a fluorinated silane monolayer acts to reduce further the adhesion when a Si stamp (with native oxide) is used. The Ti layer serves two purposes: (1) it promotes adhesion between the Au layer and the substrate after pattern

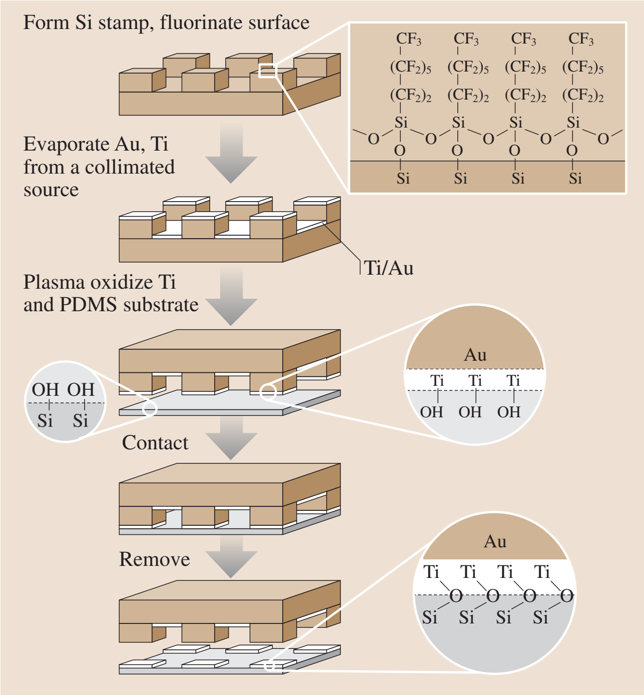
Form Si stamp, fluorinate surface
CF3 (CF2)5 (CF2)2 Si O O Si
Evaporate Au, Ti from a collimated source
Ti/Au
Plasma oxidize Ti and PDMS substrate
OH OH Si Si
Au Ti Ti Ti OH OH OH
Contact
Remove
Au Ti Ti Ti Ti O O O O Si Si Si Si

Fig. 6.6 Schematic illustration of nanotransfer printing procedures. Here, interfacial dehydration chemistries control the transfer of a thin metal film from a hard inorganic stamp to a conformable elastomeric substrate (thin film of polydimethylsiloxane (PDMS) on a plastic sheet). The process begins with the fabrication of a silicon stamp (by conventional lithography and etching) followed by surface functionalization of the native oxide with a fluorinated silane monolayer. This layer ensures poor adhesion between the stamp and a bilayer metal film (Au and Ti) deposited by electron beam evaporation. A collimated flux of metal oriented perpendicular to the surface of the stamp avoids deposition on the sidewalls of relief. Exposing the surface Ti layer to an oxygen plasma produces titanol groups. A similar exposure for the PDMS produces silanol groups. Contacting the metal-coated stamp to the PDMS results in a dehydration reaction that links the metal to the PDMS. Removing the stamp leaves a pattern of metal in the geometry of the relief features

transfer, and (2) it readily forms a ~ 3 nm oxide layer at ambient conditions, which provides a surface where the dehydration reaction can take place. Exposing the titanium oxide (TiO_x) surface to an oxygen plasma breaks bridging oxygen bonds, thus creating defect sites where water molecules can adsorb. The result is a titanium oxide surface with some fractional coverage of hydroxyl (−OH) groups (titanol).

In the case of Fig. 6.6, the substrate is a thin film of PDMS (10–50 µm thick) cast onto a sheet of poly(ethylene terephthalate) (PET; 175 µm thick). Exposing the PDMS to an oxygen plasma produces surface (−OH) groups (silanol). Placing the plasma oxidized, Au/Ti-coated stamp on top of these substrates leads to intimate, conformal contact between the raised regions of the stamp and the substrate, without the application of any external pressure. (The soft, conformable PDMS is important in this regard.) It is likely that a dehydration reaction takes place at the (−OH)-bearing interfaces during contact; this reaction results in permanent Ti−O−Si bonds that produce strong adhesion between the two surfaces. Peeling the substrate and stamp apart transfers the Au/Ti bilayer from the raised regions of the stamp (to which the metal has extremely poor adhesion) to the substrate. Complete pattern transfer from an elastomeric stamp to a thin elastomeric substrate occurs readily at room temperature in open air with contact times of less than 15 seconds. When a rigid stamp is employed, slight heating is needed to induce transfer. While the origin of this difference is unclear, it may reflect the comparatively poor contact when rigid stamps are used; similar differences are also observed in cold welding of gold films [6.36].

Figure 6.7 shows scanning electron micrographs of a pattern produced using a GaAs stamp generated by electron beam lithography and etching. The frames on the right show images of the metal-coated stamp before printing (top) and the transferred pattern (bottom). The resolution appears to be limited only by the resolution of the stamp itself, and perhaps by the grain size of the metal films. Although the accuracy in multilevel registration that is possible with nTP has not yet been quantified, its performance is likely similar to that of embossing techniques when rigid stamps are used [6.37].

A wide range of surface chemistries can be used for the transfer. SAMs are particularly attractive due to their chemical flexibility. Figure 6.8 illustrates the use of a thiol terminated SAM and nTP for forming patterns of Au on a silicon wafer [6.34]. Here, the vapor phase co-condensation of the methoxy groups of molecules of 3-mercaptopropyltrimethoxysilane (MPTMS) with the −OH terminated surface of the wafer produces a SAM of MPTMS with exposed thiol (−SH) groups. PDMS stamps can be prepared for printing on this surface by coating them with a thin film (~ 15 nm) of Au using conditions (thermal evaporation 1.0 nm/s; ~ 10^{-7} torr base pressure) that yield optically smooth, uniform films without the buckling that has been observed in the past with similar systems [6.38]. Nanocracking that sometimes occurs in the films deposited in this way can be reduced or eliminated by evaporating a small amount of Ti onto the PDMS before Au deposition and/or by exposing the PDMS surface briefly to an oxygen plasma. Bringing this coated stamp into contact with the MPTMS SAM leads to the formation of sulfur–gold bonds in the regions of contact. Removing the stamp after a few seconds efficiently transfers the gold from the raised regions of the stamp (Au does not adhere to the PDMS) to the substrate. Covalent bonding of the SAM glue to both the substrate and the gold leads to good adhesion of the printed patterns: They easily pass Scotch

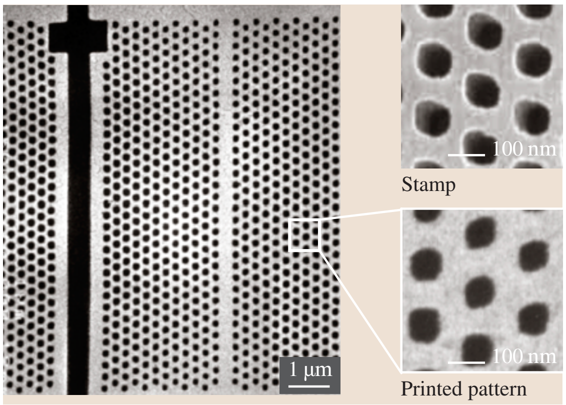

Fig. 6.7 Scanning electron micrograph (SEM) of a pattern produced by nanotransfer printing. The structure consists of a bilayer of Au (20 nm)/Ti (1 nm) (*white*) in the geometry of a photonic bandgap waveguide printed onto a thin layer of polydimethylsiloxane on a sheet of plastic (*black*). Electron beam lithography and etching of a GaAs wafer produced the stamp that was used in this case. The transfer chemistry relied on condensation reactions between titanol groups on the surface of the Ti and silanol groups on the surface of the PDMS. The *frames on the right* show SEMs of the Au/Ti-coated stamp (*top*) before printing and on the substrate (*bottom*) after printing. The electron beam lithography and etching used to fabricate the stamp limit the minimum feature size (~ 70 nm) and the edge resolution (~ 5–10 nm) of this pattern

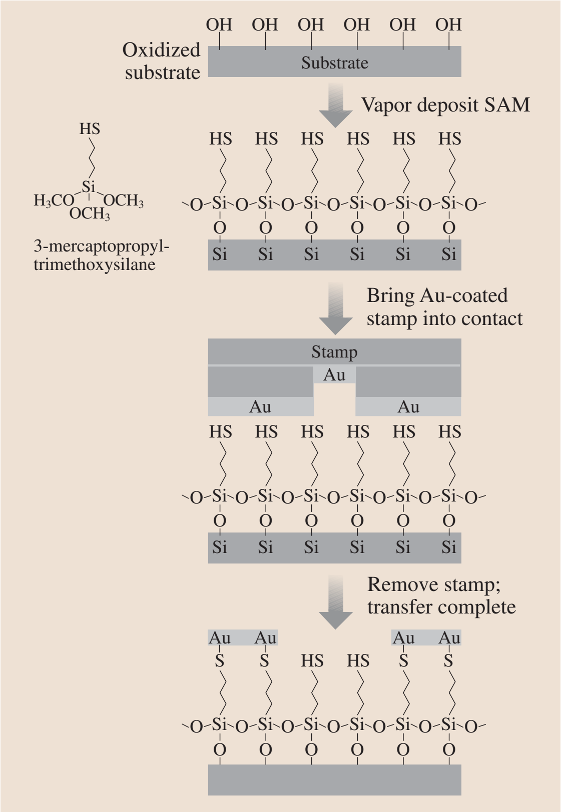

Fig. 6.8 Schematic illustration of steps for nanotransfer printing a pattern of a thin layer of Au onto a silicon wafer using a self-assembled monolayer (SAM) surface chemistry. Plasma oxidizing the surface of the wafer generates −OH groups. Solution or vapor phase exposure of the wafer to 3-mercaptopropyl-trimethoxysilane yields a SAM with exposed thiol groups. Contacting an Au-coated stamp to this surface produces thiol linkages that bond the gold to the substrate. Removing the stamp completes the transfer printing process

tape adhesion tests. Similar results can be obtained with other substrates containing surface −OH groups. For example, Au patterns can be printed onto ∼ 250 µm-thick sheets of poly(ethylene terephthalate) (PET) by first spin casting and curing (130 °C for 24 h) a thin film of an organosilsesquioxane on the PET. Exposing the cured film (∼ 1 µm thick) to an oxygen plasma and then to air produces the necessary surface (−OH) groups. Figure 6.9 shows some optical micrographs of typical printed patterns in this case [6.34].

Similar surface chemistries can guide transfer to other substrates. Alkanedithiols, for example, are useful for printing Au onto GaAs wafers [6.35]. Immersing these substrates (freshly etched with 37% HCl for ∼ 2 min to remove the surface oxide) in a 0.05 M solution of 1,8-octanedithiol in ethanol for 3 h produces a monolayer of dithiol on the surface. Although the chemistry of this system is not completely clear, it is generally believed that the thiol end groups bond chemically to the surface. Surface spectroscopy suggests the formation of Ga−S and As−S bonds. Contacting an Au-coated PDMS stamp with the treated substrate causes the exposed thiol endgroups to react with Au in the regions of contact. This reaction produces permanent Au−S bonds at the stamp/substrate interface (see insets in Fig. 6.2 for idealized chemical reaction schemes). Figure 6.10 illustrates a representative optical micrograph of an Au pattern generated by nTP on GaAs.

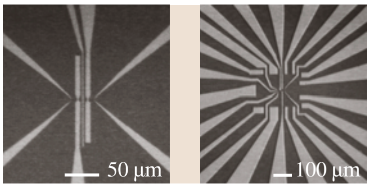

Fig. 6.9 Optical micrographs of patterns of Au (15 nm thick) formed on plastic (*left frame*) and silicon (*right frame*) substrates with nanotransfer printing. The transfer chemistries in both cases rely on self-assembled monolayers with exposed thiol groups. The minimum feature sizes and the edge resolution are both limited by the photolithography used to fabricate the stamps

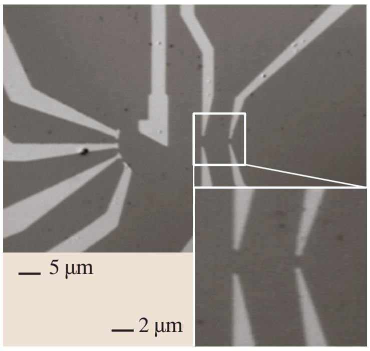

Fig. 6.10 Optical micrograph of a pattern of Au (15 nm thick) formed on a GaAs substrate using nanotransfer printing. The transfer chemistry in this case relies on a dithiol layer formed on the substrate. The minimum feature sizes and the edge resolution are both limited by the photolithography used to fabricate the stamps

6.4 Applications

Promising potential applications for the types of printing techniques described in the previous sections include those where the patterning requirements cannot be met with existing tools or those where the costs of these conventional methods are prohibitively high.We have focused our efforts partly on unusual electronic systems such as flexible plastic circuits and devices that rely on electrodes patterned on curved objects such as microcapillaries and optical fibers. We have also explored photonic systems such as distributed feedback structures for lasers and other integrated optical elements that demand sub-micron features.The sections below highlight several examples in each of these areas.

6.4.1 Unconventional Electronic Systems

A relatively new direction in electronics research seeks to establish low cost plastic materials, substrates, and printing techniques for large area flexible electronic devices, such as paper-like displays. These types of novel devices can complement those (e.g., high density memories, high speed microprocessors, etc.) that are well suited to existing inorganic (e.g., silicon) electronics technologies. High resolution patterning methods for defining the separation between the source and drain electrodes (i. e., channel length) of transistors in these plastic circuits are particularly important because this dimension determines current output and other important characteristics [6.39].

Although conventional patterning techniques, such as photolithography or electron beam lithography, have the required resolution, they are not appropriate because they are expensive and generally require multiple processing steps with resists, solvents, and developers, which can be difficult to use with organic active materials and plastic substrates. Microcontact and nanotransfer printing are both particularly well suited for this application. They can be combined and matched with other techniques, such as ink jet or screen printing, to form a complete system for patterning all layers in practical plastic electronic devices [6.40].

Figure 6.11 illustrates schematically a cross-sectional view of a typical organic transistor. The frame on the right shows the electrical switching characteristics of a device that uses source/drain electrodes of Au patterned by µCP, a dielectric layer of an organosilsesquioxane, a gate of indium tin oxide (ITO), and a PET substrate. The effective semiconductor mobility extracted from these data is comparable to those measured in devices that use the same semiconductor (pentacene in this case) with inorganic substrates and dielectrics, and gold source/drain electrodes defined by photolithography. Our recent work [6.1,27,41]with µCP in the area of plastic electronics demonstrates: (1) methods for using cylindrical "roller" stamps mounted on fixed axles for printing in a continuous reel-to-reel fashion, high resolution source/drain electrodes in ultrathin gold and silver deposited from solution at room temperature using electroless deposition, (2) techniques for performing registration and alignment of the printed features with other elements of a circuit over large areas, (3) strategies for achieving densities of defects that are as good as those observed with photolithography when the

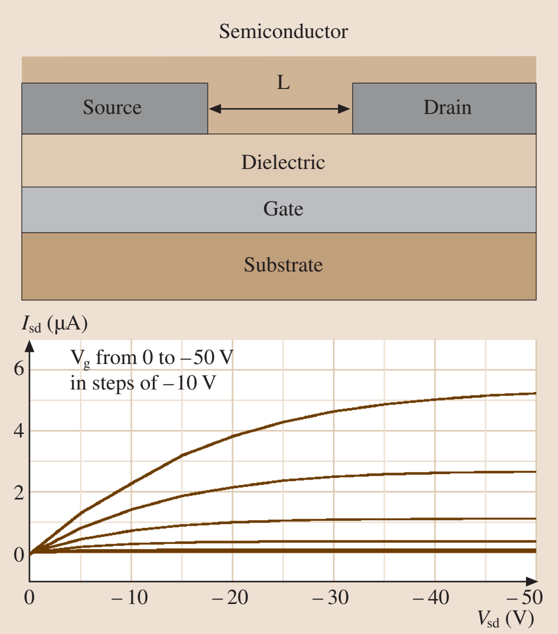

Fig. 6.11 Schematic cross-sectional view (*top*) and electrical performance (*bottom*) of an organic thin film transistor with microcontact printed source and drain electrodes. The structure consists of a substrate (PET), a gate electrode (indium tin oxide), a gate dielectric (spin cast layer of organosilsesquioxane), source and drain electrodes (20 nm Au and 1.5 nm Ti), and a layer of the organic semiconductor pentacene. The electrical properties of this device are comparable to or better than those that use pentacene with photolithographically defined source/drain electrodes and inorganic dielectrics, gates, and substrates

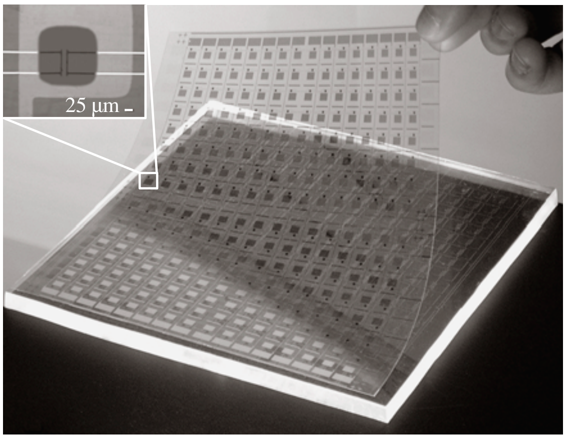

Fig. 6.12 Image of a flexible plastic active matrix backplane circuit whose finest features (transistor source/drain electrodes, and related interconnects) are patterned by microcontact printing. The circuit rests partly on the elastomeric stamp that was used for printing. The circuit consists of a square array of interconnected organic transistors, each of which acts locally as a voltage-controlled switch to control the color of an element in the display. The *inset* shows an optical micrograph of one of the transistors

patterning is performed outside of clean room facilities, (4) methods for removing the printed SAMs to allow good electrical contact of the electrodes with organic semiconductors deposited on top of them, and (5) materials and fabrication sequences that can efficiently exploit these printed electrodes for working organic TFTs in large scale circuits.

Figure 6.12 provides an image of a large area plastic circuit with critical features defined by µCP. This circuit is a flexible active matrix backplane for a display. It consists of a square array of interconnected transistors, each of which serves as a switching element that controls the color of a display pixel [6.21,42]. The transistors themselves have the layout illustrated in Fig. 6.11, and they use similar materials. The semiconductor in this image is blue (pentacene), the source/drain level is Au, the ITO appears green in the optical micrograph in the inset. Part of the circuit rests on the stamp that was used for µCP. The smallest features are the source and drain electrodes (∼ 15 µm lines), the interconnecting lines (∼ 15 µm lines), and the channel length of the transistor (∼ 15 µm). This circuit incorporates five layers of material patterned with good registration of the source/drain, gate, and semiconductor levels. The simple printing approach is illustrated in Fig. 6.13 [6.21]. Just before use, the surface of the stamp is cleaned using a conventional adhesive roller lint remover; this procedure removes dust from the stamp in a manner that does

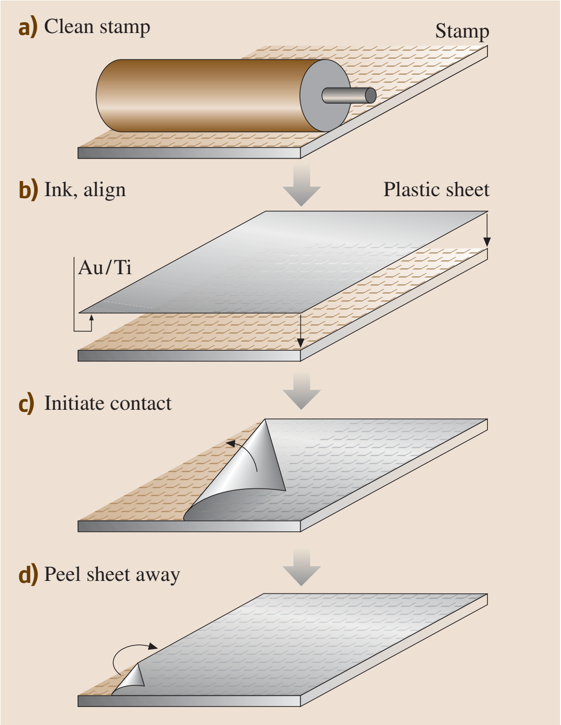

Fig. 6.13 Schematic illustration of fabrication steps for microcontact printing onto large areas of plastic sheets. The process begins with cleaning the stamp using a conventional adhesive roller lint remover. This procedure effectively removes dust particles. To minimize distortions, the stamp rests face up on a flat surface and is not manipulated directly during the printing. Alignment and registration are achieved with alignment marks on one side of the substrate and the stamp. By bending the plastic sheet, contact is initiated on one side of the stamp. The contact line is then allowed to progress gradually across the stamp. This approach avoids the formation of air bubbles that can frustrate good contact. After the substrate is in contact with the stamp for a few seconds, the plastic substrate is separated from the stamp by peeling it away, beginning in one corner. Good registration (maximum cumulative distortions less than 50 microns over an area of 0.25 square feet) and low density of defects can be achieved with this simple approach. It is also well suited for use with rigid composite stamps designed to even further reduce the level of distortion

not contaminate or damage its surface. Inking the stamp and placing it face up on a flat surface prepares it for printing. Matching the cross hair alignment marks on the corners of one edge of the stamp with those patterned in the ITO brings the substrate into registration with the stamp. During this alignment, features on the stamp are viewed directly through the semitransparent substrate. By bending the PET sheet, contact with the stamp is initiated on the edge of the substrate that contains the cross hair marks. Gradually unbending the sheet allows contact to progress across the rest of the surface. This printing procedure is attractive because it avoids distortions that can arise when directly manipulating the flexible rubber stamp. It also minimizes the number and size of trapped air pockets that can form between the stamp and substrate. Careful measurements performed after etching the unprinted areas of the gold show that over the entire 6"×6" area of the circuit, (1) the overall alignment accuracy for positioning the stamp relative to the substrate (i. e., the offset of the center of the distribution of registration errors) is $\sim 50-100\,\mu\text{m}$, even with the simple approach used here, and (2) the distortion in the positions of features in the source/drain level, when referenced to the gate level, can be as small as $\sim 50\,\mu\text{m}$ (i. e., the full width at half maximum of the distribution of registration errors). These distortions represent the cumulative effects of deformations in the stamp and distortions in the gate and column electrodes that may arise during the patterning and processing of the flexible PET sheet. The density of defects in the printed patterns is comparable to (or smaller than) that in resist patterned by contact mode photolithography when both procedures are performed outside of a cleanroom facility (i. e., when dust is the dominant source of defects).

Figure 6.14 shows an "exploded" view of a paper-like display that consists of a printed flexible plastic backplane circuit, like the one illustrated in Fig. 6.13, laminated against a thin layer of "electronic ink" [6.21, 43]. The electronic ink is composed of a monolayer of transparent polymer microcapsules that contain a suspension of charged white pigment particles suspended in a black liquid. The printed transistors in the backplane circuit act as local switches, which control electric fields that drive the pigments to the front or back of the display. When the particles flow to the front of a microcapsule, it appears white; when they flow to the back, it appears black. Figure 6.15 shows a working sheet of active matrix electronic paper that uses this design. This prototype display has several hundred pixels and an optical contrast that is both independent of the viewing angle and significantly better than newsprint. The device is ~ 1 mm thick, is mechanically flexible, and weighs $\sim 80\%$ less than a conventional liquid crystal display of similar size. Although these displays have only a relatively coarse resolution, all of the processing techniques, the μCP method, the materials, and the electronic inks are suitable for the large numbers of

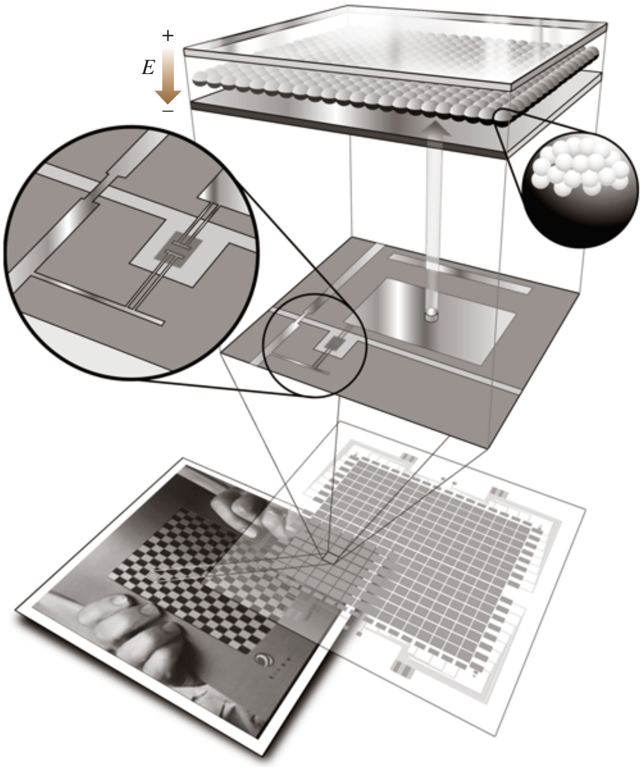

Fig. 6.14 Schematic exploded view of the components of a pixel in an electronic paper-like display (*bottom frame*) that uses a microcontact printed flexible active matrix backplane circuit (*illustration near the bottom frame*). The circuit is laminated against an unpatterned thin sheet of electronic ink (*top frame*) that consists of a monolayer of transparent polymer microcapsules (diameter ~ 100 microns). These capsules contain a heavily dyed black fluid and a suspension of charged white pigment particles (see *right inset*). When one of the transistors turns on, electric fields develop between an unpatterned transparent frontplane electrode (indium tin oxide) and a backplane electrode that connects to the transistor. Electrophoretic flow drives the pigment particles to the front or the back of the display, depending on the polarity of the field. This flow changes the color of the pixel, as viewed from the front of the display, from black to white or vice versa

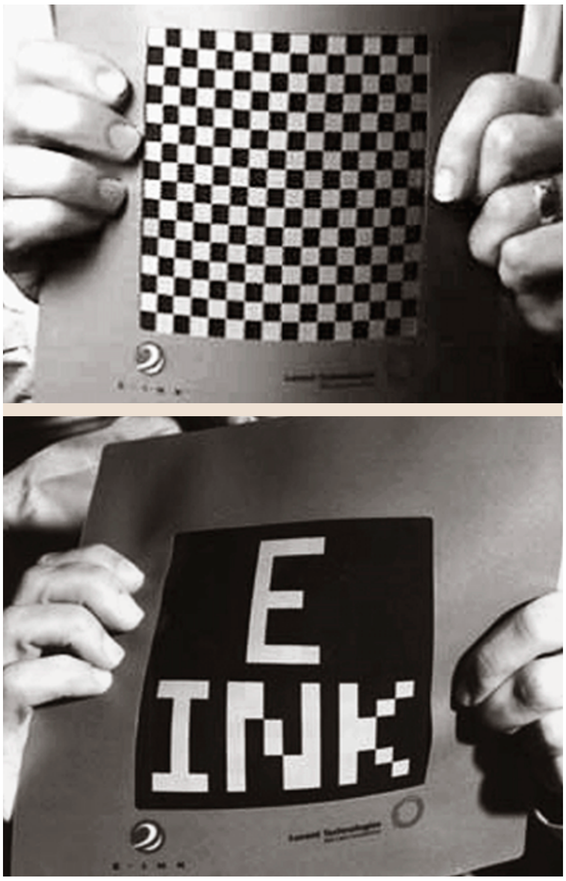

Fig. 6.15 Electronic paper-like display showing two different images. The device consists of several hundred pixels controlled by a flexible active matrix backplane circuit formed by microcontact printing. The relatively coarse resolution of the display is not limited by material properties or by the printing techniques. Instead, it is set by practical considerations for achieving high pixel yields in the relatively uncontrolled environment of the chemistry laboratory in which the circuits were fabricated

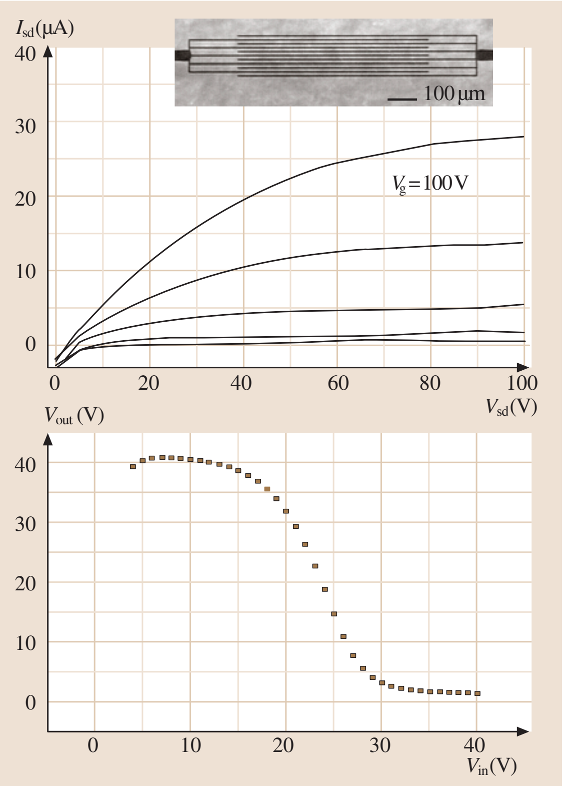

Fig. 6.16 The *top frame* shows current-voltage characteristics of an n-channel transistor formed with electrodes patterned by nanotransfer printing that are laminated against a substrate that supports an organic semiconductor, a gate dielectric, and a gate. The *inset* shows an optical micrograph of the interdigitated electrodes. The *bottom frame* on the right shows the transfer characteristics of a simple CMOS inverter circuit that uses this device and a similar one for the p-channel transistor

pixels required for high information content electronic newspapers and other systems.

Like μCP, nTP is well suited to forming high resolution source/drain electrodes for plastic electronics. nTP of Au/Ti features in the geometry of the drain and source level of organic transistors and with appropriate interconnects on a thin layer of PDMS on PET yields a substrate that can be used in an unusual but powerful way for building circuits: Soft, room temperature lamination of such a structure against a plastic substrate that supports the semiconductor, gate dielectric, and gate levels yields a high performance circuit embedded between two plastic sheets [6.33,44]. (Details of this lamination procedure are presented elsewhere.) The left frame of Fig. 6.16 shows the current-voltage characteristics of a laminated n-channel transistor that uses the organic semiconductor copper hexadecafluorophthalocyanine (n-type) and source/drain electrodes patterned with nTP. The inset shows an optical micrograph of the printed interdigitated source/drain electrodes of this device. The bottom frame of Fig. 6.16 shows the transfer characteristics of a laminated complementary organic inverter circuit whose electrodes and connecting lines are defined by nTP. The p-channel transistor in this circuit used pentacene for the semiconductor [6.33].

In addition to high resolution source/drain electrodes, it is possible to use nTP to form complex multilayer devices with electrical functionality on plastic substrates [6.34]. Figure 6.17 shows a metal/insulator/metal (MIM) structure of Au (50 nm), SiN$_x$ (100 nm; by plasma enhanced vapor deposition, PECVD), Ti (5 nm), and Au (50 nm) formed by transfer printing with a silicon stamp that is coated sequentially with these layers. In this case, a short reactive ion etch (with CF$_4$), after the second Au deposition, removes the SiN$_x$ from the sidewalls of the stamp. nTP transfers these layers in a patterned geometry to a substrate of Au(15 nm)/Ti(1 nm)-coated PDMS(50 µm)/PET(250 µm). Interfacial cold welding between the Au on the surfaces of the stamp and substrate bonds the multilayers to the substrate. Figure 6.3 illustrates the procedures, the structures (lateral dimensions of 250 µm × 250 µm, for ease of electrical probing), and their electrical characteristics. These MIM capacitors have performance similar to devices fabricated on silicon wafers by photolithography and lift-off. This example illustrates the ability of nTP to print patterns of materials whose growth conditions (high temperature SiN$_x$ by PECVD, in this case) prevent their direct deposition or processing on the substrate of interest (PET, in this case). The cold welding transfer approach has also been exploited in other ways for patterning components for plastic electronics [6.45, 46].

Another class of unusual electronic/optoelectronic devices relies on circuits or circuit elements on curved surfaces. This emerging area of research was stimulated primarily by the ability of µCP to print high resolution features on fibers and cylinders. Figure 6.18 shows a conducting microcoil printed with µCP on a microcapillary tube using the approach illustrated in Fig. 6.4. The coil serves as the excitation and detection element for high resolution proton nuclear magnetic resonance of nanoliter volumes of fluid that are housed in the bore of the microcapillary [6.47]. The high fill factor and other considerations lead to extremely high sensitivity with such printed coils. The bottom frame of Fig. 6.18 shows the spectrum of an ∼ 8 nL volume of ethylbenzene. The narrow lines demonstrate the high resolution that is possible with this approach. Similar coils can be used as magnets [6.48], springs [6.32], and electrical transformers [6.49]. Figure 6.19 shows an optical micrograph and the electrical measurements from a concentric cylindrical microtransformer that uses a microcoil printed on a microcapillary tube with a ferromagnetic wire threaded through its core. Inserting this

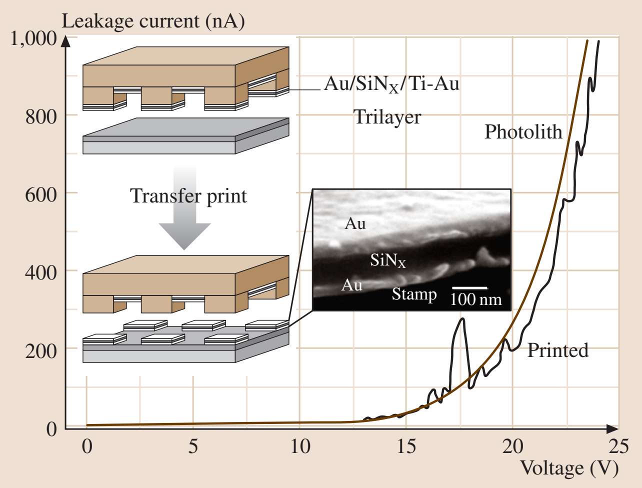

Fig. 6.17 Multilayer thin film capacitor structure printed in a single step onto a plastic substrate using the nanotransfer printing technique. A multilayer of Au/SiN$_x$/Ti/Au was first deposited onto a silicon stamp formed by photolithography and etching. Contacting this stamp to a substrate of Au/PDMS/PET forms a cold weld that bonds the exposed Au on the stamp to the Au coating on the substrate. Removing the stamp produces arrays of square (250 µm × 250 µm) metal/insulator/metal capacitors on the plastic support. The *black line* shows the measured current–voltage characteristics of one of these printed capacitors. The *brown line* corresponds to a similar structure formed on a rigid glass substrate using conventional photolithographic procedures. The characteristics are the same for these two cases. The slightly higher level of noise in the printed devices results, at least partly, from the difficulty in making good electrical contacts to structures on the flexible plastic substrate

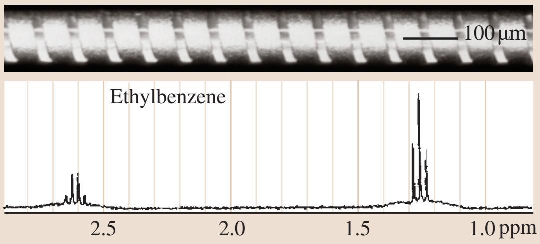

Fig. 6.18 The *top frame* shows an optical micrograph of a continuous conducting microcoil formed by microcontact printing onto a microcapillary tube. This type of printed microcoil is well suited for excitation and detection of nuclear magnetic resonance spectra from nanoliter volumes of fluid housed in the bore of the microcapillary. The *bottom frame* shows a spectra trace collected from an ∼ 8 nL volume of ethyl benzene using a structure similar to the one shown in the top frame

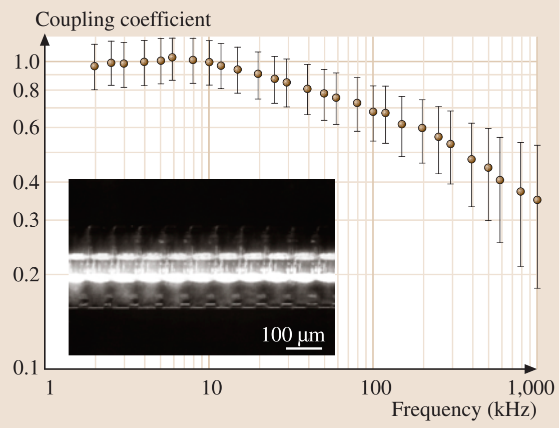

Fig. 6.19 The *inset* shows a concentric microtransformer formed using microcoils printed onto two different microcapillary tubes. The smaller of the tubes (outer diameter 135 microns) has a ferromagnetic wire threaded through its core. The larger one (outer diameter 350 microns) has the smaller tube threaded through its core. The resulting structure is a microtransformer that shows good coupling coefficients at frequencies up to ∼ 1 MHz. The graph shows its performance

structure into the core of a larger microcapillary that also supports a printed microcoil completes the transformer [6.49]. This type of device shows good coupling coefficients up to relatively high frequencies. Examples of other optoelectronic components appear in fiber optics where microfabricated on-fiber structures serve as integrated photomasks [6.16] and distributed thermal actuators [6.18].

6.4.2 Lasers and Waveguide Structures

In addition to integral components of unconventional electronic systems, useful structures for integrated optics can be built by using µCP and nTP to print sacrificial resist layers for etching glass waveguides. These printing techniques offer the most significant potential value for this area when they are used to pattern fea-

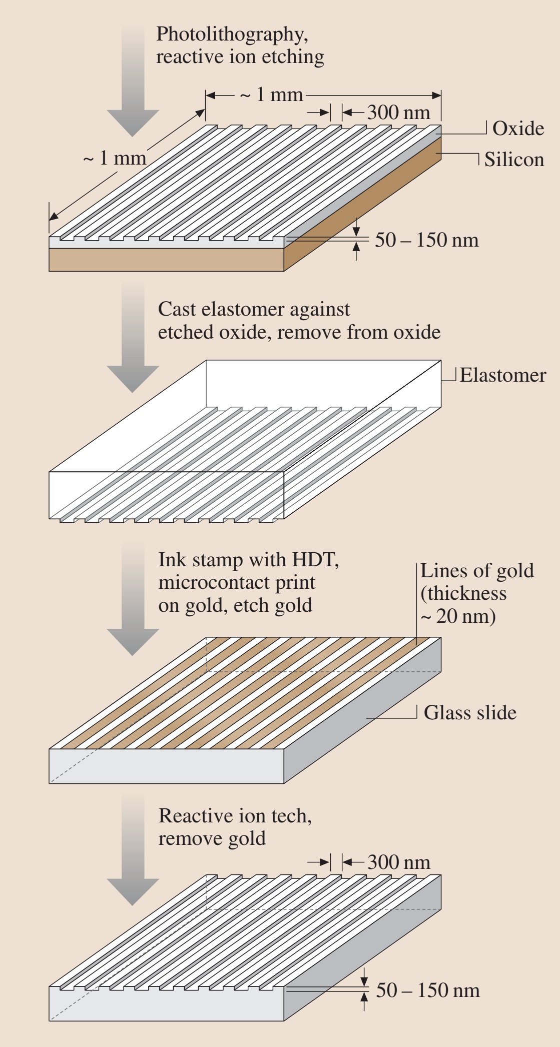
Photolithography, reactive ion etching
~ 1 mm
300 nm
Oxide
Silicon
~ 1 mm
50 – 150 nm
Cast elastomer against etched oxide, remove from oxide
Elastomer
Ink stamp with HDT, microcontact print on gold, etch gold
Lines of gold (thickness ~ 20 nm)
Glass slide
Reactive ion tech, remove gold
300 nm
50 – 150 nm

Fig. 6.20 ▶ Schematic illustration of the use of microcontact printing (µCP) for fabricating high resolution gratings that can be incorporated into distributed planar laser structures or other components for integrated optics. The geometries illustrated here are suitable for 3rd order distributed feedback (DFB) lasers that operate in the red

Fig. 6.21 ▶ Schematic illustrations and lasing spectra of plastic lasers that use microcontact printed resonators based on surface relief distributed Bragg reflectors (DBRs) and distributed feedback gratings (DFBs) on glass substrates. The grating periods are ∼ 600 nm in both cases. The lasers use thin film plastic gain media deposited onto the printed gratings. The laser shows emission in a narrow wavelength range, with a width that is limited by the resolution of the spectrometer used to characterize the output. In both cases, the emission profiles, the lasing thresholds and other characteristics of the devices are comparable to similar lasers that use resonators formed by high resolution projection mode photolithography

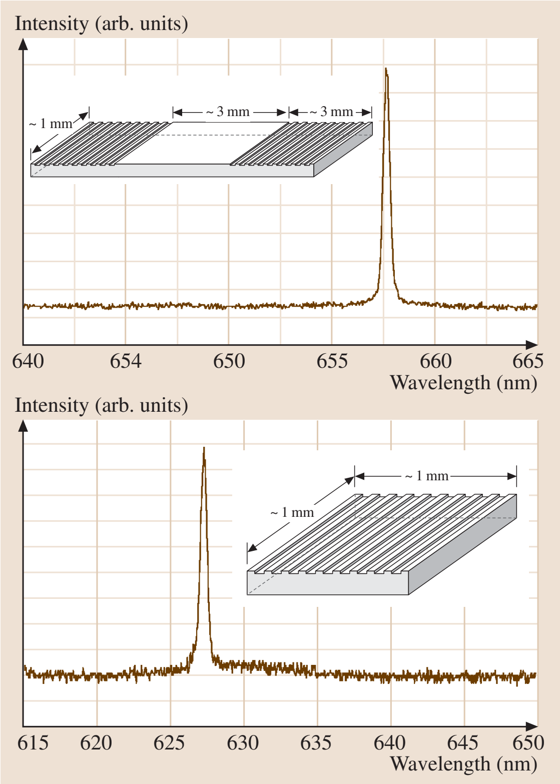

tures that are smaller than those that can be achieved with contact mode photolithography. Mode coupling gratings and distributed laser resonators are two such classes of structures. We have demonstrated µCP for forming distributed feedback (DFB) and distributed Bragg reflector (DBR) lasers that have narrow emission line widths [6.50]. This challenging fabrication

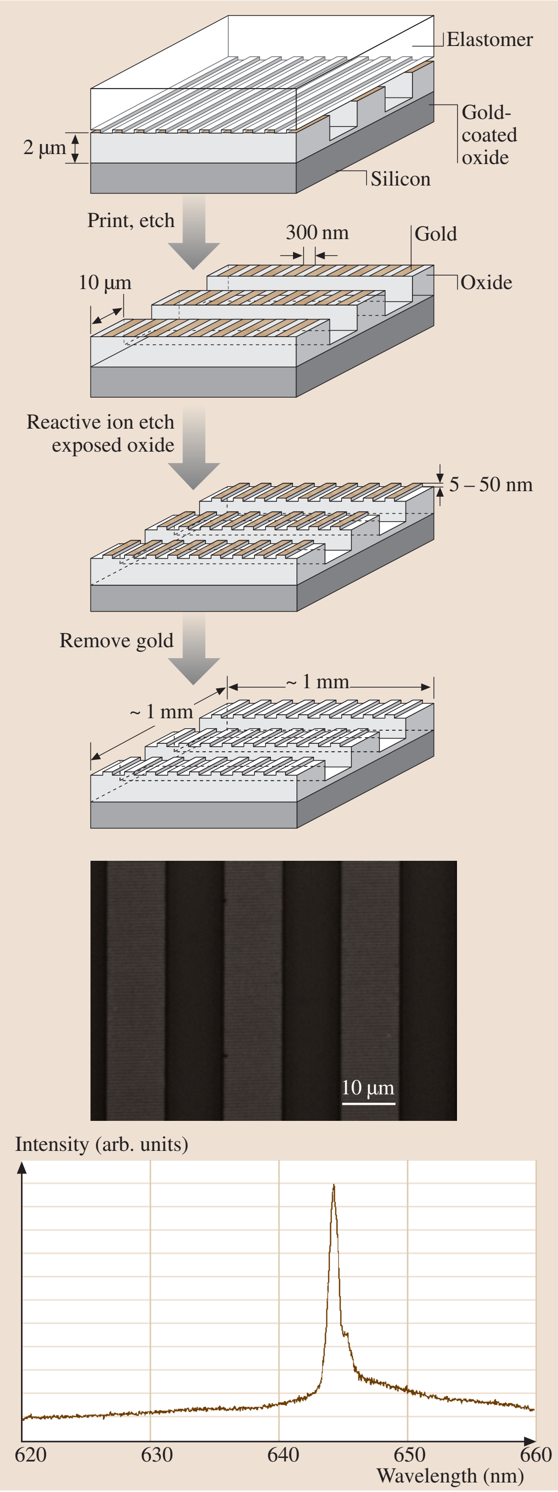
Elastomer
Gold-
coated
oxide
2 µm
Silicon
Print, etch
300 nm
Gold
10 µm
Oxide
Reactive ion etch
exposed oxide
5 – 50 nm
Remove gold
~ 1 mm
~ 1 mm
10 µm
Intensity (arb. units)
620
630
640
650
660
Wavelength (nm)

Fig. 6.22 The *top frames* gives a schematic illustration of steps for microcontact printing high resolution gratings directly onto the top surfaces of ridge waveguides. The printing defines a sacrificial etch mask of gold that is subsequently removed. Producing this type of structure with photolithography is difficult because of severe thickness nonuniformities that appear in photoresist spin cast on this type of non-planar substrate. The *upper bottom frame* shows a top view optical micrograph of printed gold lines on the ridge waveguides. The *lower bottom frame* shows the emission output of a plastic photopumped laser that uses the printed structure and a thin evaporated layer of gain media

demonstrates the suitability of µCP for building structures that have (1) feature sizes significantly less than one micron (∼ 300 nm), *and* (2) long-range spatial coherence (∼ 1 mm). The lasers employ optically pumped gain material deposited onto DFB or DBR resonators formed from periodic relief on a transparent substrate. The gain media confines light to the surface of the structure; its thickness is chosen to support a single transverse mode. To generate the required relief, lines of gold formed by µCP on a glass slide act as resists for reactive ion etching of the glass. Removing the gold leaves a periodic pattern of relief (600 nm period, 50 nm depth) on the surface of the glass (see Fig. 6.20). Figure 6.21 shows the performance of plastic lasers that use printed DFB and DBR resonators with gain media consisting of thin films of PBD doped with 1% by weight of coumarin 490 and DCMII, photopumped with 2 ns pulses from a nitrogen laser with intensities $> 5\,\mathrm{kW/cm^2}$ [6.50]. Multimode lasing at resolution-limited line widths was observed at wavelengths corresponding to the third harmonic of the gratings. These characteristics are similar to those observed in lasers that use resonators generated with photolithography and better than those that use imprinted polymers [6.51].

Contact printing not only provides a route to low cost equivalents of gratings fabricated with other approaches, but also allows the fabrication of structures that would be difficult or impossible to generate with photolithography. For example, µCP can be used to form DFB resonators directly on the top surfaces of ridge waveguides [6.52]. Figure 6.22 illustrates the procedures. The bottom left frame shows an optical micrograph of the printed gold lines. Sublimation of a ∼ 200 nm film of tris(8-hydroxyquinoline) aluminum (Alq) doped with 0.5–5.0 weight percent of the laser dye DCMII onto the resonators produces waveguide DFB lasers. The layer of gain material itself provides a planar waveguide with air and polymer as the cladding layers. The relief waveguide provides lateral confinement of the light. Photopumping these devices with the output of a pulsed nitrogen laser (∼ 2 ns, 337 nm) causes lasing due to Bragg reflections induced by the DFB structures on the top surfaces of the ridge waveguides. Some of the laser emission scatters out of the plane of the waveguide at an angle allowed by phase matching conditions. In this way, the grating also functions as an output coupler and offers a convenient way to characterize the laser emission. The bottom right frame of Fig. 6.22 shows the emission profile.

6.5 Conclusions

This chapter provides an overview of two contact printing techniques that are capable of micron and submicron resolution. It also illustrates some applications of these methods that may provide attractive alternatives to more established lithographic methods. The growing interest in nanoscience and technology makes crucial the development of new methods for fabricating the relevant test structures and devices. The simplicity of the techniques described here, together with the interesting and subtle materials science, chemistry, and physics associated with them, make this a promising area for basic and applied study.

References

6.1 C. A. Mirkin, J. A. Rogers: Emerging methods for micro- and nanofabrication, MRS Bull. **26** (2001) 506–507

6.2 H. I. Smith, H. G. Craighead: Nanofabrication, Phys. Today **43** (February 1990) 24–43

6.3 W. M. Moreau (Ed.): *Semiconductor Lithography: Principles and Materials* (Plenum, New York 1988)

6.4 S. Matsui, Y. Ochiai: Focused ion beam applications to solid state devices, Nanotechnology **7** (1996) 247–258

6.5 J. M. Gibson: Reading and writing with electron beams, Phys. Today **50** (1997) 56–61

6.6 L. L. Sohn, R. L. Willett: Fabrication of nanostructures using atomic-force microscope-based lithography, Appl. Phys. Lett. **67** (1995) 1552–1554

6.7 E. Betzig, K. Trautman: Near-Field optics – microscopy, spectroscopy, and surface modification beyond the diffraction limit, Science **257** (1992) 189–195

6.8 A. J. Bard, G. Denault, C. Lee, D. Mandler, D. O. Wipf: Scanning electrochemical microscopy: A new technique for the characterization and modification of surfaces, Acc. Chem. Res. **23** (1990) 357

6.9 J. A. Stroscio, D. M. Eigler: Atomic and molecular manipulation with the scanning tunneling microscope, Science **254** (1991) 1319–1326

6.10 J. Nole: Holographic lithography needs no mask, Laser Focus World **33** (1997) 209–212

6.11 A. N. Broers, A. C. F. Hoole, J. M. Ryan: Electron beam lithography – resolution limits, Microelectron. Eng. **32** (1996) 131–142

6.12 A. N. Broers, W. Molzen, J. Cuomo, N. Wittels: Electron-beam fabrication of 80 A metal structures, Appl. Phys. Lett. **29** (1976) 596

6.13 A. Kumar, G. M. Whitesides: Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol ink followed by chemical etching, Appl. Phys. Lett. **63** (1993) 2002–2004

6.14 Y. Xia, G. M. Whitesides: Soft lithography, Angew. Chem. Int. Ed. **37** (1998) 550–575

6.15 Y. Xia, J. A. Rogers, K. E. Paul, G. M. Whitesides: Unconventional methods for fabricating and patterning nanostructures, Chem. Rev. **99** (1999) 1823–1848

6.16 J. A. Rogers, R. J. Jackman, J. L. Wagener, A. M. Vengsarkar, G. M. Whitesides: Using microcontact printing to generate photomasks on the surface of optical fibers: A new method for producing in-fiber gratings, Appl. Phys. Lett. **70** (1997) 7–9

6.17 B. Michel, A. Bernard, A. Bietsch, E. Delamarche, M. Geissler, D. Juncker, H. Kind, J. P. Renault, H. Rothuizen, H. Schmid, P. Schmidt-Winkel, R. Stutz, H. Wolf: Printing meets lithography: Soft approaches to high-resolution printing, IBM J. Res. Dev. **45** (2001) 697–719

6.18 J. A. Rogers: Rubber stamping for plastic electronics and fiber optics, MRS Bull. **26** (2001) 530–534

6.19 N. B. Larsen, H. Biebuyck, E. Delamarche, B. Michel: Order in microcontact printed self-assembled monolayers, J. Am. Chem. Soc. **119** (1997) 3017–3026

6.20 H. A. Biebuyck, G. M. Whitesides: Self-organization of organic liquids on patterned self-assembled monolayers of alkanethiolates on gold, Langmuir **10** (1994) 2790–2793

6.21 J. A. Rogers, Z. Bao, K. Baldwin, A. Dodabalapur, B. Crone, V. R. Raju, V. Kuck, H. Katz, K. Amundson, J. Ewing, P. Drzaic: Paper-like electronic displays: Large area, rubber stamped plastic sheets of electronics and electrophoretic inks, Proc. Nat. Acad. Sci. **98** (2001) 4835–4840

6.22 J. L. Wilbur, H. A. Biebuyck, J. C. MacDonald, G. M. Whitesides: Scanning force microscopies can image patterned self-assembled monolayers, Langmuir **11** (1995) 825–831

6.23 J. C. Love, D. B. Wolfe, M. L. Chabinyc, K. E. Paul, G. M. Whitesides: Self-assembled monolayers of alkanethiolates on palladium are good etch resists, J. Am. Chem. Soc. **124** (2002) 1576–1577

6.24 H. Schmid, B. Michel: Siloxane polymers for high-resolution, high-accuracy soft lithography, Macromolecules **33** (2000) 3042–3049

6.25 K. Choi, J. A. Rogers: A photocurable poly(dimethylsiloxane) chemistry for soft lithography in the nanometer regime, J. Am. Chem. Soc. **125** (2003) 4060–4061

6.26 J. A. Rogers K. E. Paul, G. M. Whitesides: Quantifying distortions in soft lithography, J. Vac. Sci. Technol. B **16** (1998) 88–97

6.27 J. Tate, J. A. Rogers, C. D. W. Jones, W. Li, Z. Bao, D. W. Murphy, R. E. Slusher, A. Dodabalapur, H. E. Katz, A. J. Lovinger: Anodization and microcontact printing on electroless silver: Solution-based fabrication procedures for low voltage organic electronic systems, Langmuir **16** (2000) 6054–6060

6.28 Y. Xia, E. Kim, G. M. Whitesides: Microcontact printing of alkanethiols on silver and its application to microfabrication, J. Electrochem. Soc. **143** (1996) 1070–1079

6.29 Y. N. Xia, X. M. Zhao, E. Kim, G. M. Whitesides: A selective etching solution for use with patterned self-assembled monolayers of alkanethiolates on gold, Chem. Mater. **7** (1995) 2332–2337

6.30 R. J. Jackman, J. Wilbur, G. M. Whitesides: Fabrication of submicrometer features on curved substrates by microcontact printing, Science **269** (1995) 664–666

6.31 R. J. Jackman, S. T. Brittain, A. Adams, M. G. Prentiss, G. M. Whitesides: Design and fabrication of topologically complex, three-dimensional microstructures, Science **280** (1998) 2089–2091

6.32 J. A. Rogers, R. J. Jackman, G. M. Whitesides: Microcontact printing and electroplating on curved substrates: a new means for producing free-standing three-dimensional microstructures with possible applications ranging from micro-coil springs to coronary stents, Adv. Mater. **9** (1997) 475–477

6.33 Y.-L. Loo, R. W. Willett, K. Baldwin, J. A. Rogers: Additive, nanoscale patterning of metal films with a stamp and a surface chemistry mediated transfer process: Applications in plastic electronics, Appl. Phys. Lett. **81** (2002) 562–564

6.34 Y.-L. Loo, R. W. Willett, K. Baldwin, J. A. Rogers: Interfacial chemistries for nanoscale transfer printing, J. Am. Chem. Soc. **124** (2002) 7654–7655

6.35 Y.-L. Loo, J. W. P. Hsu, R. L. Willett, K. W. Baldwin, K. W. West, J. A. Rogers: High-resolution transfer printing on GaAs surfaces using alkane dithiol self-assembled monolayers, J. Vac. Sci. Technol. B. **20** (2002) 2853–2856

6.36 G. S. Ferguson, M. K. Chaudhury, G. B. Sigal, G. M. Whitesides: Contact adhesion of thin gold-films on elastomeric supports – cold welding under ambient conditions, Science **253** (1991) 776–778

6.37 W. Zhang, S. Y. Chou: Multilevel nanoimprint lithography with submicron alignment over 4 in Si wafers, Appl. Phys. Lett. **79** (2001) 845–847

6.38 N. Bowden, S. Brittain, A. G. Evans, J. W. Hutchinson, G. M. Whitesides: Spontaneous formation of ordered structures in thin films of metals supported on an elastomeric polymer, Nature **393** (1998) 146–149

6.39 J. A. Rogers, Z. Bao, A. Dodabalapur, A. Makhija: Organic smart pixels and complementary inverter circuits formed on plastic substrates by casting, printing and molding, IEEE Electron Dev. Lett. **21** (2000) 100–103

6.40 Z. Bao, J. A. Rogers, H. E. Katz: Printable organic and polymeric semiconducting materials and devices, J. Mater. Chem. **9** (1999) 1895–1904

6.41 J. A. Rogers, Z. Bao, A. Makhija: Non-photolithographic fabrication sequence suitable for reel-to-reel production of high performance organic transistors and circuits that incorporate them, Adv. Mater. **11** (1999) 741–745

6.42 P. Mach, S. Rodriguez, R. Nortrup, P. Wiltzius, J. A. Rogers: Active matrix displays that use printed organic transistors and polymer dispersed liquid crystals on flexible substrates, Appl. Phys. Lett. **78** (2001) 3592–3594

6.43 J. A. Rogers: Toward paperlike displays, Science **291** (2001) 1502–1503

6.44 Y.-L. Loo, T. Someya, K. W. Baldwin, P. Ho, Z. Bao, A. Dodabalapur, H. E. Katz, J. A. Rogers: Soft, conformable electrical contacts for organic transistors: High resolution circuits by lamination, Proc. Nat. Acad. Sci. USA **99** (2002) 10252–10256

6.45 C. Kim, P. E. Burrows, S. R. Forrest: Micropatterning of organic electronic devices by cold-welding, Science **288** (2000) 831–833

6.46 C. Kim, M. Shtein, S. R. Forrest: Nanolithography based on patterned metal transfer and its application to organic electronic devices, Appl. Phys. Lett. **80** (2002) 4051–4053

6.47 J. A. Rogers, R. J. Jackman G. M. Whitesides, D. L. Olson, J. V. Sweedler: Using microcontact printing to fabricate microcoils on capillaries for high resolution ^{1}H-NMR on nanoliter volumes, Appl. Phys. Lett. **70** (1997) 2464–2466

6.48 J. A. Rogers, R. J. Jackman, G. M. Whitesides: Constructing single and multiple helical microcoils and characterizing their performance as components of microinductors and microelectromagnets, J. Microelectromech. Syst. (JMEMS) **6** (1997) 184–192

6.49 R. J. Jackman, J. A. Rogers, G. M. Whitesides: Fabrication and characterization of a concentric, cylindrical microtransformer, IEEE Trans. Magn. **33** (1997) 2501–2503

6.50 J. A. Rogers, M. Meier, A. Dodabalapur: Using stamping and molding techniques to produce distributed feedback and Bragg reflector resonators for plastic lasers, Appl. Phys. Lett. **73** (1998) 1766–1768

6.51 M. Berggren, A. Dodabalapur, R. E. Slusher, A. Timko, O. Nalamasu: Organic solid-state lasers with imprinted gratings on plastic substrates, Appl. Phys. Lett. **72** (1998) 410–411

6.52 J. A. Rogers, M. Meier, A. Dodabalapur: Distributed feedback ridge waveguide lasers fabricated by nanoscale printing and molding on non-planar substrates, Appl. Phys. Lett. **74** (1999) 3257–3259

7. Materials Aspects of Micro- and Nanoelectromechanical Systems

Two of the more significant technological achievements during the last 20 years have been the development of MEMS and its new offshoot, NEMS. These developments were made possible in large measure by significant advancements in the materials and processes used in the fabrication of MEMS and NEMS devices. And while initial developments capitalized on a mature Si infrastructure, recent advances have used materials and processes not associated with IC fabrication, a trend that is likely to continue as new application areas are identified.

A well-rounded understanding of MEMS and NEMS requires a basic knowledge of the materials used to construct the devices, since material properties often govern device performance. An understanding of the materials used in MEMS and NEMS is really an understanding of material systems. Devices are rarely constructed of a single material, but rather a collection of materials, each providing a critical function and often working in conjunction with each other. It is from this perspective that the following chapter is constructed. A preview of the materials selected for inclusion is presented in Table 7.1. From this table it is easy to see that this chapter is not a summary of all materials used in MEMS and NEMS, as such a work would itself constitute a text of significant size. It does, however, present a selection of some of the more important material systems, including examples that illustrate the importance of viewing MEMS and NEMS in terms of material systems.

7.1 **Silicon** 203
  7.1.1 Single Crystal Silicon 204
  7.1.2 Polysilicon 205
  7.1.3 Porous Silicon 208
  7.1.4 Silicon Dioxide 208
  7.1.5 Silicon Nitride 209

7.2 **Germanium-Based Materials** 210
  7.2.1 Polycrystalline Ge 210
  7.2.2 Polycrystalline SiGe 210

7.3 **Metals** 211

7.4 **Harsh Environment Semiconductors** 212
  7.4.1 Silicon Carbide 212
  7.4.2 Diamond 215

7.5 **GaAs, InP, and Related III–V Materials** 217

7.6 **Ferroelectric Materials** 218

7.7 **Polymer Materials** 219
  7.7.1 Polyimide 219
  7.7.2 SU-8 220
  7.7.3 Parylene 220

7.8 **Future Trends** 220

References 221

7.1 Silicon

Use of silicon (Si) as a material for microfabricated sensors dates back to the 1950s when *C. S. Smith* published a paper describing the piezoresistive effect in germanium (Ge) and Si [7.1]. *Smith* found that the piezoresistive coefficients of Si were significantly higher than those associated with conventional metal strain gauges, a finding that led to the development of Si strain gauges in the early 1960s. Throughout the 1960s and early 1970s, various Si bulk micromachining techniques were developed, which ultimately led to the commercial production of piezoresistive Si pressure sensors during this period. The subsequent development of surface micromachining techniques for Si, along with the recognition that micromachined Si structures could potentially be integrated with Si IC devices, marked the advent of MEMS and positioned Si as the primary MEMS material.

Table 7.1 Distinguishing characteristics and application examples of selected materials for MEMS and NEMS

Material	Distinguishing Characteristics	Application Examples
Single crystal Silicon (Si)	High quality electronic material, selective anisotropic etching	Bulk micromachining, piezoresistive sensing
Polycrystalline Si (polysilicon)	Doped Si films on sacrificial layers	Surface micromachining, electrostatic actuation
Silicon Dioxide (SiO_2)	Insulating, etched by HF, compatible with polysilicon	Sacrificial layer in polysilicon surface micromachining, passivation layer for devices
Silicon Nitride (Si_3N_4, Si_xN_y)	Insulating, chemically resistant, mechanically durable	Isolation layer for electrostatic devices, membrane and bridge material
Polycrystalline Germanium (poly Ge), Polycrystalline Silicon-Germanium (poly Si-Ge)	Deposited at low temperatures	Integrated surface micromachined MEMS
Gold (Au), Aluminum (Al)	Conductive thin films, flexible deposition techniques	Innerconnect layers, masking layers, electromechanical switches
Nickel-Iron (NiFe)	Magnetic alloy	Magnetic actuation
Titanium-Nickel (TiNi)	Shape-memory alloy	Thermal actuation
Silicon Carbide (SiC) Diamond	Electrically and mechanically stable at high temperatures, chemically inert, high Young's modulus to density ratio	Harsh environment MEMS, high frequency MEMS/NEMS
Gallium Arsenide (GaAs), Indium Phosphide (InP), Indium Arsenide (InAs), and related materials	Wide bandgap, epitaxial growth on related ternary compounds	RF MEMS, optoelectronic devices, single crystal bulk and surface micromachining
Lead Zirconate Titanate (PZT)	Piezoelectric material	Mechanical sensors and actuators
Polyimide	Chemically resistant, high temperature polymer	Mechanically flexible MEMS, bioMEMS
SU-8	Thick, photo-definable resist	Micromolding, high-aspect-ratio structures
Parylene	Biocompatible polymer, deposited at room temperature by CVD	Protective coatings, molded polymer structures

7.1.1 Single Crystal Silicon

For MEMS applications, single crystal Si serves several key functions. It is one of the most versatile materials for bulk micromachining, due to the availability of anisotropic etching processes in conjunction with good mechanical properties. Having a Young's modulus of about 190 GPa, Si compares favorably, from a mechanical perspective, with steel, which has a Young's modulus of about 210 GPa. Favorable mechanical properties enable Si to be used as a material for membranes, beams, and other such structures. For surface micromachining applications, single crystal Si substrates are used primarily as mechanical platforms on which Si and non-Si device structures can be fabricated. Use of high-quality single crystal wafers enables the fabrication of integrated MEMS devices, at least for materials and processes that are compatible with Si ICs.

Bulk micromachining is a process whereby etching techniques are used in conjunction with etch masks and etch stops to selectively sculpt micromechanical structures from a bulk substrate. From the materials perspective, single crystal Si is a relatively easy material to bulk micromachine, due to the availability of anisotropic etchants, such as potassium hydroxide (KOH) and tetramethyl-aluminum hydroxide (TMAH), that attack the (100) and (110) Si crystal planes significantly faster than the (111) crystal planes. For example, the etching rate ratio of (100) to (111) planes in Si is about 400:1 for a typical KOH/water etch solution. Silicon dioxide (SiO_2), silicon nitride (Si_3N_4), and some metallic thin films (e.g., Cr, Au, etc.) provide good etch masks for most Si anisotropic etchants.

In contrast to anisotropic etching, isotropic etching exhibits no selectivity to the various crystal planes. From a processing perspective, isotropic etching of Si is commonly used for removal of work-damaged surfaces, creation of structures in single-crystal slices, and patterning single-crystal or polycrystalline films. The most commonly used isotropic Si etchants are mixtures of hydrofluoric (HF) and nitric (HNO_3) acid in water or acetic acid (CH_3COOH), with the etch rate dependent on the ratio of HF to HNO_3.

In terms of etch stops, boron-doped Si is effective for some liquid reagents. Boron-doped etch stops are often less than 10 μm thick, since the boron concentration in Si must exceed $7\times10^{19}\,\mathrm{cm}^3$ for the etch stop to be effective, and the doping is done by diffusion. It is possible to create a boron-doped etch stop below an undoped Si layer using ion implantation; however, the practical implant depth is limited to a few microns.

Well-established dry etching processes are available to pattern single crystal Si. The process spectrum ranges from physical mechanisms such as sputtering and ion milling to chemical mechanisms such as plasma etching. Reactive ion etching (RIE) is the most commonly used dry etching technique to etch Si. By combining both physical and chemical processes, RIE is a highly effective anisotropic Si etching technique that is independent of crystalline orientation. Fluorinated compounds such as CF_4, SF_6, and NF_3, or chlorinated compounds such as CCl_4 or Cl_2, sometimes mixed with He, O_2, or H_2, are commonly used in Si RIE. The RIE process is highly directional, which enables direct pattern transfer from an overlying masking material to the etched Si surface. SiO_2 thin films are often used as a masking material, owing to its chemical durability and ease in patterning. Process limitations (i. e., etch rates) restrict the etch depths of Si RIE to less than 10 microns. However, a process called deep reactive ion etching (DRIE) has extended the use of anisotropic dry etching to depths well beyond several hundred microns.

Using the aforementioned processes and techniques, a wide variety of microfabricated devices have been made from single crystal Si, such as piezoresistive pressure sensors, accelerometers, and mechanical resonators, to name a few. Using nearly the same approaches but on a smaller scale, top-down nanomachining techniques have been used to fabricate nanoelectromechanical devices from single crystal Si. Single crystal Si is particularly well-suited for nanofabrication, because high crystal quality substrates with very smooth surfaces are readily available. By coupling electron-beam (e-beam) lithographic techniques with conventional Si etching, device structures with submicron dimensions have been fabricated. Submicron, single crystal Si nanomechanical structures have been successfully micromachined from bulk Si wafers [7.2] and silicon-on-insulator (SOI) wafers [7.3]. In the former case, an isotropic Si etch was performed to release the device structures, whereas in the latter, the 50 nm to 200 nm structures were released by dissolving the underlying oxide layer in HF. An example of nanoelectromechanical beam structures fabricated from a single crystal Si substrate is shown in Fig. 7.1.

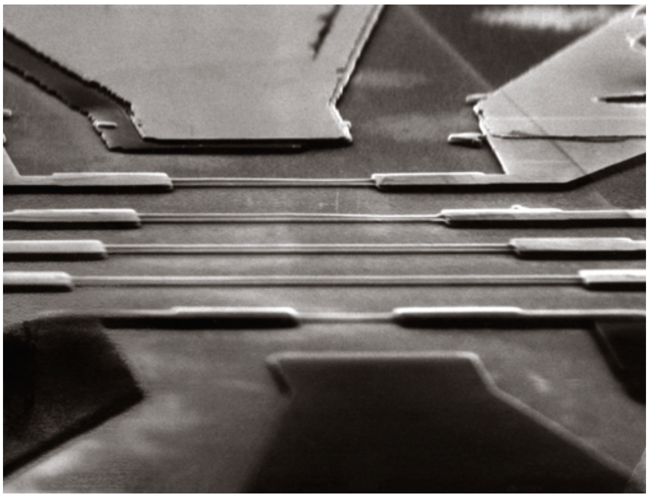
Fig. 7.1 A collection of Si nanoelectromechanical beam resonators fabricated from a single-crystal Si substrate (courtesy M. Roukes, Caltech)

7.1.2 Polysilicon

Surface micromachining is a process whereby a sequence of thin films, often of different materials, is deposited and selectively etched to form the desired micromechanical (or microelectromechanical) structure. In contrast to bulk micromachining, the substrate serves primarily as a platform to support the device. For Si-based surface micromachined MEMS, polycrystalline Si (polysilicon) is most often used as the structural material, silicon dioxide (SiO_2) as the sacrificial material, silicon nitride (Si_3N_4) for electrical isolation of device structures, and single crystal Si as the substrate. Like single crystal Si, polysilicon can be doped during or after film deposition. SiO_2 can be thermally grown or deposited on Si over a broad temperature range (e.g., 200 °C to 1,150 °C) to meet various process and material

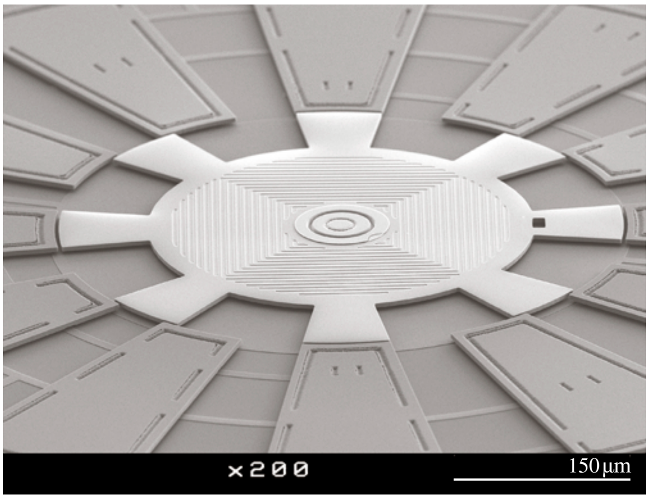

Fig. 7.2 SEM micrograph of a surface-micromachined polysilicon micromotor fabricated using a SiO_2 sacrificial layer

requirements. SiO_2 is readily dissolvable in hydrofluoric acid (HF), which does not etch polysilicon and, thus, can be used to dissolve SiO_2 sacrificial layers. Si_3N_4 is an insulating film that is highly resistant to oxide etchants. The polysilicon micromotor shown in Fig. 7.2 was surface micromachined using a process that included these materials.

For MEMS and IC applications, polysilicon films are commonly deposited using a process known as low-pressure chemical vapor deposition (LPCVD). The typical polysilicon LPCVD reactor is based on a hot wall, resistance-heated furnace. Typical processes are performed at temperatures ranging from 580 °C to 650 °C and pressures from 100 to 400 mtorr. The most commonly used source gas is silane (SiH_4).

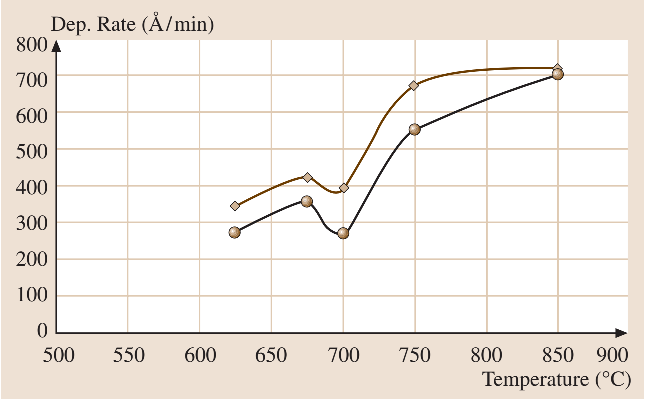

Fig. 7.3 Deposition rate versus substrate temperature for in situ boron-doped (◇) and undoped (●) polysilicon films grown by atmospheric pressure chemical vapor deposition [7.4]

Polysilicon thin films consist of a collection of small single crystal grains whose microstructure and orientation is a function of the deposition conditions [7.5]. For typical LPCVD processes (e.g., 200 mtorr), the amorphous-to-polycrystalline transition temperature is about 570 °C, with amorphous films deposited below the transition temperature. At 600 °C, the grains are small and equiaxed, while at 625 °C, the grains are large and columnar [7.5]. The crystal orientation is predominantly (110) Si for temperatures between 600 °C and 650 °C, while the (100) orientation is dominant for temperatures between 650 °C and 700 °C.

In terms of chemical resistance, polysilicon is very much like single crystal Si. The oxidation rate of undoped polysilicon is typically between that of (100)- and (111)-oriented single crystal Si. For temperatures below 1,000 °C, heavily phosphorus-doped polysilicon oxidizes at a rate significantly higher than undoped polysilicon.

The resistivity of polysilicon can be modified using the doping methods developed for single crystal Si. Diffusion is an effective method for doping polysilicon films, especially for heavy doping of thick films. Phosphorous, which is the most commonly used dopant in polysilicon MEMS, diffuses significantly faster in polysilicon than in single crystal Si, due primarily to enhanced diffusion rates along grain boundaries. The diffusivity in polysilicon thin films (i. e., small equiaxed grains) is about $1 \times 10^{12}\ cm^2/s$.

Ion implantation is also used to dope polysilicon films. A high-temperature annealing step is usually required to electrically activate the implanted dopants, as well as to repair implant-related damage in the polysilicon films. In general, the conductivity of implanted polysilicon films is not as high as films doped by diffusion.

In situ doping of polysilicon is performed by simply including a dopant gas, usually diborane (B_2H_6) or phosphine (PH_3), in the CVD process. The addition of dopants during the deposition process not only modifies the conductivity, but also affects the deposition rate of the polysilicon films. As shown in Fig. 7.3, the inclusion of boron generally increases the deposition rate of polysilicon relative to undoped films [7.4], while phosphorus (not shown) reduces the rate. In situ doping can be used to produce conductive films with uniform doping profiles without requiring the high-temperature steps usually associated with diffusion or ion implantation. Although commonly used to produce doped polysilicon for electrostatic devices, *Cao* et al. [7.6] have used in situ phosphorus-doped polysilicon films in piezoresistive strain gauges, achieving gauge factors as high as 15 for a single strip sensor.

The thermal conductivity of polysilicon is a strong function of its microstructure and, therefore, the conditions used during deposition [7.5]. For fine-grained films, the thermal conductivity is about 25% of the value of single crystal Si. For thick films with large grains, the thermal conductivity ranges between 50% and 85% of the single crystal value.

Like the electrical and thermal properties of polysilicon, the as-deposited residual stress in polysilicon films depends on microstructure. For films deposited under typical conditions (200 mtorr, 625 °C), the as-deposited polysilicon films have compressive residual stresses. The highest compressive stresses are found in amorphous Si films and polysilicon films with a strong, columnar (110) texture. For films with fine-grained microstructures, the stress tends to be tensile. Annealing can be used to reduce the compressive stress in as-deposited polysilicon films. For instance, compressive residual stresses on the order of 500 MPa can be reduced to less than 10 MPa by annealing the as-deposited films at 1,000 °C in a N_2 ambient [7.7, 8]. Recent advances in the area of rapid thermal annealing (RTA) indicate that RTA provides an effective method of stress reduction in polysilicon films. *Zhang* et al. [7.9] reported that a 10 s anneal at 1,100 °C was sufficient to completely relieve the stress in films that originally had a compressive stress of about 340 MPa. RTA is particularly attractive in situations where the process parameters require a low thermal budget.

As an alternative to high-temperature annealing, *Yang* et al. [7.10] have developed an approach that actually utilizes the residual stress characteristics of polysilicon deposited under various conditions to construct polysilicon multilayers that have the desired thickness and stress values. The multilayers are comprised of alternating tensile and compressive polysilicon layers that are deposited in a sequential manner. The tensile layers consist of fine-grained polysilicon grown at a temperature of 570 °C, while the compressive layers are made up of columnar polysilicon deposited at 615 °C. The overall stress in the composite film depends on the number of alternating layers and the thickness of each layer. With the proper set of parameters, a composite polysilicon multilayer can be deposited with near zero residual stress and no stress gradient. The process achieves stress reduction without high-temperature annealing, a considerable advantage for integrated MEMS processes.

Many device designs require thick polysilicon structural layers that are not readily achievable using conventional LPCVD polysilicon, due to process-related limitations. For these designs, epitaxial Si reactors can be used to grow polysilicon films. Unlike conventional LPCVD processes with deposition rates of less than 100 Å/min, epitaxial processes have deposition rates on the order of 1 micron/min [7.11]. The high deposition rates result from the much higher substrate temperatures (> 1,000 °C) and deposition pressures (> 50 torr) used in these processes. The polysilicon films are usually deposited on SiO_2 sacrificial layers to enable surface micromachining. An LPCVD polysilicon seed layer is sometimes used to control nucleation, grain size, and surface roughness. As with conventional polysilicon, the microstructure and residual stress of the epi-poly films, as they are known, is related to deposition conditions. Compressive films generally have a mixture of [110] and [311] grains [7.12, 13], while tensile films have a random mix of [110], [100], [111], and [311] grains [7.12]. The Young's modulus of epi-poly measured from micromachined test structures is comparable to LPCVD polysilicon [7.13]. Mechanical properties test structures [7.11–13], thermal actuators [7.11], electrostatically actuated accelerometers [7.11], and gryoscopes [7.14] have been fabricated from these films.

All of the aforementioned polysilicon deposition processes utilize substrate temperatures in excess of 570 °C either during film deposition, or in post-deposition processing steps. As a low temperature alternative to LPCVD polysilicon, *Abe* et al. [7.15] and *Honer* et al. [7.16] have developed sputtering processes for polysilicon. Early work [7.15] emphasized the ability to deposit very smooth (25 Å) polysilicon films on thermally oxidized wafers at reasonable deposition rates (191 Å/min) and with low residual compressive stresses. The process involved DC magnitron sputtering from a Si target using an Ar sputtering gas, a chamber pressure of 5 mtorr, and a power of 100 W. The authors reported that a post-deposition anneal at 700 °C in N_2 for two hours was needed to crystallize the deposited film and perhaps lower the stress. *Honer* et al. [7.16] sought to develop a polymer-friendly, Si-based surface micromachining process based on polysilicon sputtered onto polyimide and PSG sacrificial layers. To improve the conductivity of the micromachined Si structures, the sputtered Si films were sandwiched between two TiW cladding layers. The device structures on polyimide were released using oxygen plasma etching. The processing step with the highest temperature was, in fact, the polyimide cure at 350 °C. To test the robustness of the process, sputter-deposited Si microstructures were fabricated on substrates containing CMOS devices. As expected from thermal budget considerations, the authors reported no measurable degradation of device performance.

7.1.3 Porous Silicon

Porous Si is produced by room temperature electrochemical etching of Si in HF. If configured as an electrode in an HF-based electrochemical circuit, positive charge carriers (holes) at the Si surface facilitate the exchange of F atoms with H atoms, that terminate the Si surface. The exchange continues in the subsurface region, leading to the eventual removal of the fluorinated Si. The quality of the etched surface is related to the density of holes at the surface, which is controlled by the applied current density. For high current densities, the density of holes is high and the etched surface is smooth. For low current densities, the hole density is low and clustered in highly localized regions associated with surface defects. Surface defects become enlarged by etching, which leads to the formation of pores. Pore size and density are related to the type of Si used and the conditions of the electrochemical cell. Both single crystal and polycrystalline Si can be converted to porous Si.

The large surface-to-volume ratios make porous Si attractive for gaseous and liquid applications, including filter membranes and absorbing layers for chemical and mass sensing [7.17]. When single crystal substrates are used, the unetched porous layer remains single crystalline and is suitable for epitaxial Si growth. It has been shown that CVD coatings do not generally penetrate the porous regions, but rather overcoat the pores at the surface of the substrate [7.18]. The formation of localized Si-on-insulator structures is, therefore, possible by simply combining pore formation with epitaxial growth, followed by dry etching to create access holes to the porous region, and thermal oxidation of the underlying porous region. A third application uses porous Si as a sacrificial layer for polysilicon and single crystalline Si surface micromachining. As shown by *Lang* et al. [7.18], the process involves the electrical isolation of the solid structural Si layer by either pn-junction formation through selective doping, or use of electrically insulating thin films, since the formation of pores only occurs on electrically charged surfaces. A weak Si etchant will aggressively attack the porous regions with little damage to the structural Si layers and can be used to release the devices.

7.1.4 Silicon Dioxide

Silicon dioxide (SiO_2) is one of the most widely used materials in the fabrication of MEMS. In polysilicon surface micromachining, SiO_2 is used as a sacrificial material, since it can be easily dissolved using etchants that do not attack polysilicon. SiO_2 is widely used as etch mask for dry etching of thick polysilicon films, since it is chemically resistant to dry etching processes for polysilicon. SiO_2 films are also used as passivation layers on the surfaces of environmentally sensitive devices.

The most common processes used to produce SiO_2 films for polysilicon surface micromachining are thermal oxidation and LPCVD. Thermal oxidation of Si is performed at temperatures of 900 °C to 1,200 °C in the presence of oxygen or steam. Since thermal oxidation is a self-limiting process, the maximum practical film thickness that can be obtained is about 2 μm, which is sufficient for many sacrificial applications. As noted by its name, thermal oxidation of Si can only be performed on Si surfaces.

SiO_2 films can be deposited on a wide variety of substrate materials by LPCVD. In general, LPCVD provides a means for depositing thick (> 2 μm) SiO_2 films at temperatures much lower than thermal oxidation. Known as low-temperature oxides, or LTO for short, these films have a higher etch rate in HF than thermal oxides, which translates to significantly faster release times when LTO films are used as sacrificial layers. Phosphosilicate glass (PSG) can be formed using nearly the same deposition process as LTO by adding a phosphorus-containing gas to the precursor flows. PSG films are useful as sacrificial layers, since they generally have higher etching rates in HF than LTO films.

PSG and LTO films are deposited in hot-wall, low-pressure, fused silica furnaces in systems similar to those described previously for polysilicon. Precursor gases include SiH_4 as a Si source, O_2 as an oxygen source, and, in the case of PSG, PH_3 as a source of phosphorus. LTO and PSG films are typically deposited at temperatures of 425 °C to 450 °C and pressures ranging from 200 mtorr to 400 mtorr. The low deposition temperatures result in LTO and PSG films that are slightly less dense than thermal oxides, due to the incorporation of hydrogen in the films. LTO films can, however, be densified by an annealing step at high temperature (1,000 °C). The low density of LTO and PSG films is partially responsible for the increased etch rate in HF.

Thermal SiO_2 and LTO are electrical insulators used in numerous MEMS applications. The dielectric constants of thermal oxide and LTO are 3.9 and 4.3, respectively. The dielectric strength of thermal SiO_2 is 1.1×10^6 V/cm, and for LTO it is about 80% of that value [7.19]. The stress in thermal SiO_2 is compressive with a magnitude of about 300 MPa [7.19]. For LTO, however, the as-deposited residual stress is tensile, with a magnitude of about 100 MPa to 400 MPa [7.19]. The addition of phosphorous to LTO decreases the tensile residual stress to about 10 MPa for phosphorus concentrations of 8% [7.20]. As with polysilicon, the properties of LTO and PSG are dependent on processing conditions.

Plasma-enhanced chemical vapor deposition (PECVD) is another common method to produce oxides of silicon. Using a plasma to dissociate the gaseous precursors, the deposition temperatures needed to deposit PECVD oxide films is lower than for LPCVD films. For this reason, PECVD oxides are quite commonly used as masking, passivation, and protective layers, especially on devices that have been coated with metals.

Quartz is the crystalline form of SiO_2 and has interesting properties for MEMS. Quartz is optically transparent, piezoelectric, and electrically insulating. Like single crystal Si, quartz substrates are available as high quality, large area wafers that can be bulk micromachined using anisotropic etchants. A short review of the basics of quartz etching by *Danel* et al. [7.21] is recommended for those interested in the subject. Quartz has recently become a popular substrate material for microfluidic devices due to its optical, electronic, and chemical properties.

Another SiO_2-related material that has recently found uses in MEMS is spin-on-glass (SOG). SOG is a polymeric material with a viscosity suitable for spin coating. Two recent publications illustrate the potential for SOG in MEMS fabrication. In the first example, *Yasseen* et al. [7.22] detailed the development of SOG as a thick-film sacrificial molding material for thick polysilicon films. The authors reported a process to deposit, polish, and etch SOG films that were 20 microns thick. The thick SOG films were patterned into molds and filled with 10 micron-thick LPCVD polysilicon films, planarized by selective CMP, and subsequently dissolved in a wet etchant containing HCl, HF, and H_2O to reveal the patterned polysilicon structures. The cured SOG films were completely compatible with the polysilicon deposition process. In the second example, *Liu* et al. [7.23] fabricated high-aspect ratio channel plate microstructures from SOG. Electroplated nickel (Ni) was used as a molding material, with Ni channel plate molds fabricated using a conventional LIGA process. The Ni molds were then filled with SOG, and the sacrificial Ni molds were removed in a reverse electroplating process. In this case, the fabricated SOG structures (over 100 microns tall) were micromachined glass structures fabricated using a molding material more commonly used for structural components.

7.1.5 Silicon Nitride

Silicon nitride (Si_3N_4) is widely used in MEMS for electrical isolation, surface passivation, etch masking, and as a mechanical material. Two deposition methods are commonly used to deposit Si_3N_4 thin films: LPCVD and PECVD. PECVD silicon nitride is generally nonstoichiometric (sometimes denoted as $Si_xN_y:H$) and may contain significant concentrations of hydrogen. Use of PECVD silicon nitride in micromachining applications is somewhat limited because it has a high etch rate in HF (e.g., often higher than that of thermally grown SiO_2). However, PECVD offers the ability to deposit nearly stress-free silicon nitride films, an attractive property for encapsulation and packaging.

Unlike its PECVD counterpart, LPCVD Si_3N_4 is extremely resistant to chemical attack, making it the material of choice for many Si bulk and surface micromachining applications. LPCVD Si_3N_4 is commonly used as an insulating layer because it has a resistivity of $10^{16}\ \Omega \times$cm and field breakdown limit of 10^7 V/cm. LPCVD Si_3N_4 films are deposited in horizontal furnaces similar to those used for polysilicon deposition. Typical deposition temperatures and pressures range between 700 °C to 900 °C and 200 mtorr to 500 mtorr, respectively. The standard source gases are dichlorosilane (SiH_2Cl_2) and ammonia (NH_3). To produce stoichiometric Si_3N_4, a NH_3 to SiH_2Cl_2 ratio of 10:1 is commonly used. The microstructure of films deposited under these conditions is amorphous.

The residual stress in stoichiometric Si_3N_4 is large and tensile, with a magnitude of about 1 GPa. Such a large residual stress causes films thicker than a few thousand angstroms to crack. Nonetheless, thin stoichiometric Si_3N_4 films have been used as mechanical support structures and electrical insulating layers in piezoresistive pressure sensors [7.24]. To enable the use of Si_3N_4 films for applications that require micron-thick, durable, and chemically resistant membranes, Si_xN_y films can be deposited by LPCVD. These films, often referred to as Si-rich or low-stress nitride, are intentionally deposited with an excess of Si by simply decreasing the ratio of NH_3 to SiH_2Cl_2 during deposition. Nearly stress-free films can be deposited using a NH_3 to SiH_2Cl_2 ratio of 1 : 6, a deposition temperature of 850 °C, and a pressure of 500 mtorr [7.25]. The increase in Si content not only leads to a reduction in tensile stress, but also a decrease in the etch rate in HF. Such properties have enabled the development of fabrication techniques that would otherwise not be feasible with stoichiometric Si_3N_4. For example, low stress silicon nitride has been bulk micromachined using silicon as the sacrificial material. In this case, an anisotropic Si etchant (TMAH) was used to dissolve the sacrificial silicon. *French* et al. [7.26] used PSG as a sacrificial layer to surface micromachine low-stress nitride, capitalizing on the HF resistance of the nitride films.

7.2 Germanium-Based Materials

Like Si, Ge has a long history as a semiconductor device material, dating back to the development of the earliest transistors and semiconductor strain gauges. Issues related to the water solubility of germanium oxide, however, stymied the development of Ge for microelectronic devices. Nonetheless, there is a renewed interest in using Ge in micromachined devices due to the relatively low processing temperatures required to deposit the material.

7.2.1 Polycrystalline Ge

Thin polycrystalline Ge (poly-Ge) films can be deposited by LPCVD at temperatures as low as 325 °C on Si, Ge, and SiGe substrates [7.27]. Ge does not nucleate on SiO_2 surfaces, which prohibits the use of thermal oxides and LTO films as sacrificial layers, but enables the use of these films as sacrificial molds. Residual stress in poly-Ge films deposited on Si substrates can be reduced to nearly zero after short anneals at modest temperatures (30 s at 600 °C). Poly-Ge is essentially impervious to KOH, TMAH, and BOE, enabling the fabrication of Ge membranes on Si substrates [7.27]. The mechanical properties of poly-Ge are comparable to polysilicon, having a Young's modulus of 132 GPa and a fracture stress between 1.5 GPa and 3.0 GPa [7.28]. Mixtures of HNO_3, H_2O, and HCl and H_2O, H_2O_2, and HCl, as well as the RCA SC-1 cleaning solution isotropically etch Ge. Since these mixtures do not etch Si, SiO_2, Si_3N_4, and SiN, poly-Ge can be used as a sacrificial substrate layer in polysilicon surface micromachining. Using these techniques, devices such as poly-Ge-based thermistors and Si_3N_4 membrane-based pressure sensors, made using poly-Ge sacrificial layers, have been fabricated [7.27]. *Franke* et al. [7.28] found no performance degradation in Si CMOS devices following the fabrication of surface micromachined poly-Ge structures, thus demonstrating the potential for on-chip integration of Ge electromechanical devices with Si circuitry.

7.2.2 Polycrystalline SiGe

Like poly-Ge, polycrystalline SiGe (poly-SiGe) is a material that can be deposited at temperatures lower than those required for polysilicon. Deposition processes include LPCVD, APCVD, and RTCVD (rapid thermal CVD) using SiH_4 and GeH_4 as precursor gases. Deposition temperatures range from 450 °C for LPCVD [7.29] to 625 °C by rapid thermal CVD (RTCVD) [7.30]. In general, the deposition temperature is related to the concentration of Ge in the films, with higher Ge concentrations resulting in lower deposition temperatures. Like polysilicon, poly-SiGe can be doped with boron and phosphorus to modify its conductivity. In situ boron doping can be performed at temperatures as low as 450 °C [7.29]. *Sedky* et al. [7.30] showed that the deposition temperature of conductive films doped with boron could be further reduced to 400 °C if the Ge content was kept at or above 70%.

Unlike poly-Ge, poly-SiGe can be deposited on a number of sacrificial substrates, including SiO_2 [7.30], PSG [7.28], and poly-Ge [7.28]. For Ge rich films, a thin polysilicon seed layer is sometimes used on SiO_2 surfaces since Ge does not readily nucleate on oxide surfaces. Like many compound materials, variations in film composition can change the physical properties of the material. For instance, etching of poly-SiGe by H_2O_2 becomes significant for Ge concentrations over 70%. *Sedky* et al. [7.31] have shown that the microstructure, film conductivity, residual-stress, and residual stress gradient are related to the concentration of Ge in the material. With respect to residual stress, *Franke* et al. [7.29] produced in situ boron-doped films with residual compressive stresses as low as 10 MPa.

The poly-SiGe, poly-Ge material system is particularly attractive for surface micromachining since H_2O_2 can be used as a release agent. It has been reported that poly-Ge etches at a rate of 0.4 microns/min in H_2O_2, while poly-SiGe with Ge concentrations below 80% have no observable etch rate after 40 hrs [7.32]. The ability to use H_2O_2 as a sacrificial etchant makes the combination of poly-SiGe and poly-Ge extremely attractive for surface micromachining from the processing, safety, and materials compatibility points of view. Due to the conformal nature of LPCVD processing, poly-SiGe structural elements, such as gimbal-based microactuator structures, have been made by high-aspect ratio micromolding [7.32]. Capitalizing on the low deposition temperatures, an integrated MEMS fabrication process with Si ICs has been demonstrated [7.29]. In this process, CMOS structures are first fabricated on Si wafers. Poly-SiGe mechanical structures are then surface micromachined using a poly-Ge sacrificial layer. A significant advantage of this design is that the MEMS structure is positioned directly above the CMOS structure, thus reducing the parasitic capacitance and contact resistance characteristic of interconnects associated with side-by-side integration schemes. Use of H_2O_2 as the sacrificial etchant means that no special protective layers are required to protect the underlying CMOS layer during release. In addition to its utility as a material for integrated MEMS devices, poly-SiGe has been identified as a material well suited for micromachined thermopiles [7.33] due to its lower thermal conductivity relative to Si.

7.3 Metals

It can be argued that of all the material categories associated with MEMS, metals may be among the most enabling, since metallic thin films are used in many different capacities ranging from etch masks used in device fabrication to interconnects and structural elements in microsensors and microactuators. Metallic thin films can be deposited using a wide range of techniques, including evaporation, sputtering, CVD, and electroplating. Since a complete review of the metals used in MEMS is far beyond the scope of this chapter, the examples presented in this section were selected to represent a broad cross section of where metals have found uses in MEMS.

Aluminum (Al) and gold (Au) are among the most widely employed metals in microfabricated electronic and electromechanical devices, as a result of their use as innerconnect and packaging materials. In addition to these critical electrical functions, Al and Au are also desirable as electromechanical materials. One such example is the use of Au micromechanical switches for RF MEMS. For conventional RF applications, chip level switching is currently performed using FET- and PIN diode-based solid state devices fabricated from gallium arsenide (GaAs) substrates. Unfortunately, these devices suffer from insertion losses and poor electrical isolation. In an effort to develop replacements for GaAs-based solid state switches, *Hyman* et al. [7.34] reported the development of an electrostatically actuated, cantilever-based micromechanical switch fabricated on GaAs substrates. The device consisted of a silicon nitride-encased Au cantilever constructed on a sacrificial silicon dioxide layer. The silicon nitride and silicon dioxide layers were deposited by PECVD, and the Au beam was electroplated from a sodium sulfite solution inside a photoresist mold. A thin multilayer of Ti and Au was sputter deposited in the mold prior to electroplating. The trilayer cantilever structure was chosen to minimize the deleterious effects of thermal and process-related stress gradients in order to produce unbent and thermally stable beams. After deposition and pattering, the cantilevers were released in HF. The processing steps proved to be completely compatible with GaAs substrates. The released cantilevers demonstrated switching speeds better than 50 µs at 25 V with contact lifetimes exceeding 10^9 cycles.

In a second example from RF MEMS, *Chang* et al. [7.35] reported the fabrication of an Al-based micromachined switch as an alternative to GaAs FETs and PIN diodes. In contrast to the work by *Hyman* et al. [7.34], this switch utilizes the differences in the residual stresses in Al and Cr thin films to create bent cantilever switches that capitalize on the stress differences in the materials. Each switch is comprised of a series of linked bimorph cantilevers designed in such a way that the resulting structure bends significantly out of the plane of the wafer due to the stress differences in the bimorph. The switch is drawn closed by electrostatic attraction. The bimorph consists of metals that can easily be processed with GaAs wafers, thus making integration with GaAs devices possible. The released switches were relatively slow, at 10 ms, but an actuation voltage of only 26 V was needed to close the switch.

Thin-film metallic alloys that exhibit the shape-memory effect are of particular interest to the MEMS community for their potential in microactuators. The shape-memory effect relies on the reversible transformation from a ductile martensite phase to a stiff austenite phase in the material with the application of heat. The reversible phase change allows the shape-memory effect to be used as an actuation mechanism, since the material changes shape during the transition. It has been found that high forces and strains can be generated from shape-memory thin films at reasonable power inputs, thus enabling shape-memory actuation to be used in MEMS-based microfluidic devices, such as microvalves and micropumps. Titanium-nickel (TiNi) is among the most popular of the shape-memory alloys, owing to its high actuation work density ($50\,\mathrm{MJ/m^3}$) and large bandwidth (up to 0.1 kHz) [7.36]. TiNi is also attractive because conventional sputtering techniques can be employed to deposit thin films, as detailed in a recent report by *Shih* et al. [7.36]. In this study, TiNi films were deposited by co-sputtering elemental Ti and Ni targets, and a co-sputtering TiNi alloy and elemental Ti targets. It was reported that co-sputtering from TiNi and Ti targets produced better films due to process variations related to the roughening of the Ni target in the case of Ti and Ni co-sputtering. The TiNi/Ti co-sputtering process has been used to produce shape-memory material for a silicon spring-based microvalve [7.37].

Use of thin-film metal alloys in magnetic actuator systems is another example of the versatility of metallic materials in MEMS. Magnetic actuation in microdevices generally requires the magnetic layers to be relatively thick (tens to hundreds of microns) to generate magnetic fields of sufficient strength to generate the desired actuation. To this end, magnetic materials are often deposited by thick film methods, such as electroplating. The thickness of these layers exceeds what can feasibly be patterned by etching, so plating is often performed in microfabricated molds made from materials such as polymethylmethacrylate (PMMA). The PMMA mold thickness can exceed several hundred microns, so X-rays are used as the exposure source during the patterning steps. When necessary, a metallic thin film seed layer is deposited prior to plating. After plating, the mold is dissolved, which frees the metallic component. Known as LIGA (short for Lithography, Galvanoforming and Abformung), this process has been used to produce a wide variety of high-aspect ratio structures from plateable materials, such as nickel-iron (NiFe) magnetic alloys [7.38] and Ni [7.39].

In addition to elemental metals and simple compound alloys, more complex metallic alloys commonly used in commerical macroscopic applications are finding their way into MEMS applications. One such example is an alloy of titanium known as Ti-6Al-4V. Composed of 88% titanium, 6% aluminum and 4% vanadium, this alloy is widely used in commercial aviation due to its weight, strength, and temperature tolerance. *Pornsin-sirirak* et al. [7.40] have explored the use of this alloy in the manufacture of MEMS-based winged structures for micro aerial vehicles. The authors considered this alloy not only because of its weight and strength, but also because of its ductility and its etching rate at room temperature. The designs for the wing prototype were modelled after the wings of bats and various flying insects. For this application, Ti-alloy structures patterned from bulk (250 μm-thick) material by an $HF/HNO_3/H_2O$ etching solution were used, rather than thin films. Parylene-C (detailed in Sect. 7.7.3) was deposited on the patterned alloy to serve as the wing membrane. The miniature micromachined wings were integrated into a test setup, and several prototypes actually demonstrated short duration flight.

7.4 Harsh Environment Semiconductors

Silicon Carbide (SiC) has long been recognized as the leading semiconductor for use in high temperature and high power electronics and is currently receiving attention as a material for harsh environment MEMS.

7.4.1 Silicon Carbide

SiC is a polymorphic material that exists in cubic, hexagonal, and rhombehedral polytypes. The cubic polytype, called 3C-SiC, has an electronic bandgap of 2.3 eV, which is over twice that of Si. Numerous hexagonal and rhombehedral polytypes have been identified, the two most common being 4H-SiC and 6H-SiC. The electronic bandgap of 4H- and 6H-SiC is even higher than 3C-SiC – 2.9 and 3.2 eV, respectively. SiC films can be doped to create n-type and p-type material. The stiffness of SiC is quite large relative to Si, with a Young's modulus ranging from 300 GPa to 450 GPa, depending on the microstructure. SiC is not etched in any wet Si etchants or by XeF_2, a popular dry Si etchant used for releasing device structures [7.41]. SiC is a material that does not melt, but rather sublimes at temperatures in excess of 1,800 °C. Single crystal 4H- and 6H-SiC wafers are commercially available, but smaller in diameter (3 inches) and much more expensive than Si.

SiC thin films can be grown or deposited using a number of different techniques. For high-quality single crystal films, APCVD and LPCVD processes are most commonly employed. Homoepitaxial growth of 4H- and 6H-SiC yields high quality films suitable for microelectronic applications, but only on substrates of the same polytype. These processes usually employ dual precursors, such as SiH_4 and C_3H_8, and are performed at temperatures ranging from 1,500 °C to 1,700 °C. Epitaxial films with p-type or n-type conductivity can be grown using aluminum and boron for p-type films and nitrogen and phosphorus for n-type films. Nitrogen is so effective at modifying the conductivity of SiC that growth of undoped SiC is extremely challenging, because the concentrations of residual nitrogen in typical deposition systems are sufficient for n-type doping.

APCVD and LPCVD can also be used to deposit 3C-SiC on Si substrates. Heteroepitaxy is possible, despite a 20% lattice mismatch, because 3C-SiC and Si have similar lattice structures. The growth process involves two key steps. The first step, called carbonization, converts the near surface region of the Si substrate to 3C-SiC by simply exposing it to a hydrocarbon/hydrogen mixture at high substrate temperatures ($> 1,200$ °C). The carbonized layer forms a crystalline template on which a 3C-SiC film can be grown by adding a silicon-containing gas to the hydrogen/hydrocarbon mix. The lattice mismatch between Si and 3C-SiC results in the formation of crystalline defects in the 3C-SiC film, with the density being highest in the carbonization layer and decreasing with increasing thickness. The crystal quality of 3C-SiC films is nowhere near that of epitaxially grown 4H- and 6H-SiC films; however, the fact that 3C-SiC can be grown on Si substrates enables the use of Si bulk micromachining techniques to fabricate a host of 3C-SiC-based mechanical devices. These include microfabricated pressure sensors [7.42] and nanoelectromechanical resonant structures [7.43].

Polycrystalline SiC (poly-SiC) has proven to be a very versatile material for SiC MEMS. Unlike single crystal versions of SiC, poly-SiC can be deposited on a variety of substrate types, including such common surface micromachining materials as polysilicon, SiO_2, and Si_3N_4, using a much wider range of processes than epitaxial films. Commonly used deposition techniques include LPCVD [7.41, 44, 45] and APCVD [7.46, 47]. The deposition of poly-SiC requires much lower substrate temperatures than epitaxial films, ranging from roughly 700 °C to 1,200 °C. Amorphous SiC can be deposited at even lower temperatures (25 °C to 400 °C) by PECVD [7.48] and sputtering [7.49]. The microstructure of poly-SiC films is temperature, substrate, and process dependent. For amorphous substrates, such as SiO_2 and Si_3N_4, APCVD poly-SiC films deposited from silane and propane are oriented randomly with equiaxed grains [7.47], whereas for oriented substrates, such as polysilicon, the texture of the poly-SiC film matches that of the substrate [7.46]. By comparison, poly-SiC films deposited by LPCVD from dichlorosilane and acetylene are highly textured (111) films with a columnar microstructure [7.44], while films deposited from disilabutane have a distribution of orientations [7.41]. This variation suggests that device performance can be tailored by selecting the proper substrate and deposition conditions.

Direct bulk micromachining of SiC is very difficult due to its chemical inertness. Although conventional wet chemical techniques are not effective, several electrochemical etch processes have been demonstrated and used in the fabrication of 6H-SiC pressure sensors [7.50]. The etching processes are selective to the conductivity of the material, so dimensional control of the etched structures depends on the ability to form doped layers, which can only be formed by in situ or ion implantation processes, since solid source diffusion is not possible at reasonable processing temperatures. This constraint somewhat limits the geometrical complexity of the patterned structures, as compared with conventional plasma-based etching. To fabricate thick (hundreds of microns), 3-D, high-aspect ratio SiC structures, a molding technique has been developed [7.39]. The molds are fabricated from Si substrates using deep reactive ion etching and then filled with SiC using a combination of thin epitaxial and thick polycrystalline film CVD processes. The thin film process is used to protect the mold from pitting during the more aggressive mold filling SiC growth step. The mold filling process coats all surfaces of the mold with a SiC film as thick as the mold is deep. To release the SiC structure, the substrate is first mechanically polished to expose sections of the Si mold, then the substrate is immersed in a Si etchant to completely dissolve the mold. This process has been used to fabricate solid SiC fuel atomizers [7.39], and a variant has been used to fabricate SiC structures for micropower systems [7.51].

In addition to CVD processes, bulk micromachined SiC structures can be fabricated using sintered SiC powders. *Tanaka* et al. [7.52] describe a process in which SiC components, such as micro gas turbine engine rotors, can be fabricated from SiC powders using a micro-reaction-sintering process. The molds are microfabricated from Si using DRIE and filled with SiC and graphite powders mixed with a phenol resin. The molds are then reaction-sintered using a hot isostatic pressing technique. The SiC components are then released from the Si mold by wet chemical etching. The authors reported that the component shrinkage was less than 3%. The bending strength and Vickers hardness of the micro-reaction-sintered material was roughly 70 to 80% of commercially available reaction-sintered SiC, the difference being attributed to the presence of unreacted Si in the micro-scale components.

In a related process, *Liew* et al. [7.54] detail a technique to create silicon carbon nitride (SiCN) MEMS structures by molding injectable polymer precursors. Unlike the aforementioned processes, this technique uses SU-8 photoresists for the molds. SU-8, detailed later in this chapter, is a versatile photo-definable polymer in which thick films (hundreds of microns) can be patterned using conventional UV photolithographic techniques. After patterning, the molds are filled with the SiCN-containing polymer precursor, lightly polished, and then subjected to a multistep heat-treating process. During the thermal processing steps, the SU-8 mold decomposes, and the SiCN structure is released. The resulting SiCN structures retain many of the same properties of stoichiometric SiC.

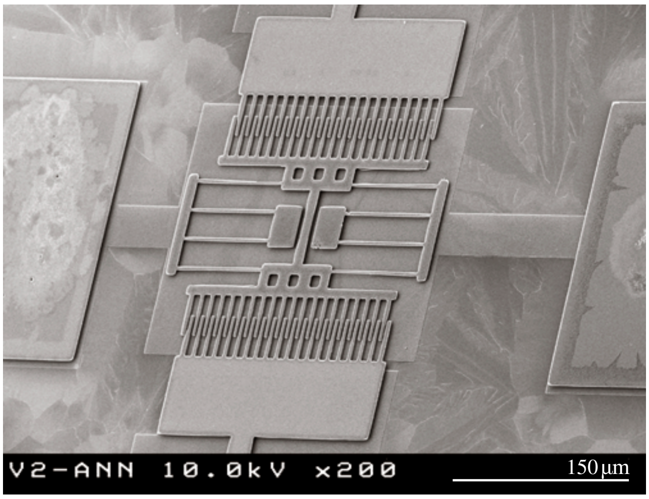

Fig. 7.4 SEM micrograph of a poly-SiC lateral resonant structure fabricated using a multilayer, micromolding-based micromachining process [7.53]

Although SiC cannot be etched using conventional wet etch techniques, thin SiC films can be patterned using conventional dry etching techniques. RIE processes using fluorinated compounds, such as CHF_3 and SF_6 combined with O_2, an inert gas, or H_2. The high oxygen content in these plasmas generally prohibits the use of photoresist as a masking material, therefore, hard masks made of Al, Ni, and ITO are often used. RIE-based SiC surface micromachining processes with polysilicon and SiO_2 sacrificial layers have been developed for single layer devices [7.55,56]. Multilayered structures are very difficult to fabricate by direct RIE because the etch rates of the sacrificial layers are much higher than the SiC structural layers, making critical dimensional control very difficult and RIE-based SiC multilayer processes impractical.

To address the issues related to RIE of multilayer SiC structures, a micromolding process for patterning SiC films on sacrificial layer substrates has been developed [7.57]. In essence, the micromolding technique is the thin film analog to the molding-based, bulk micromachining technique presented earlier. The micromolding process utilizes polysilicon and SiO_2 films as both molds and sacrificial substrate layers, with SiO_2 molds used with polysilicon sacrificial layers and vica versa. These films are deposited and patterned using conventional methods, thus leveraging the well characterized and highly selective processes developed for polysilicon MEMS. Poly-SiC films are simply deposited into the micromolds, and mechanical polishing is used to remove poly-SiC from the top of the molds. Appropriate etchants are then used to dissolve the molds and sacrificial layers. The micromolding method utilizes the differences in chemical properties of the three materials in this system in a way that bypasses the difficulties associated with chemical etching of SiC. This technique has been extended to multilayer processes that have been used to fabricate SiC micromotors [7.57] and the lateral resonant structure shown in Fig. 7.4 [7.53].

Yang et al. [7.43] have recently shown that the chemical inertness of SiC facilitates the fabrication of NEMS devices. In this work, the authors present a fabrication method to realize SiC mechanical resonators with submicron thickness and width dimensions. The resonators were fabricated from ∼ 260 nm-thick 3C-SiC films epitaxially grown on (100) Si wafers. The films were patterned into 150 nm-wide beams ranging in length from 2 to 8 µm. The beams were etched in a $NF_3/O_2/Ar$

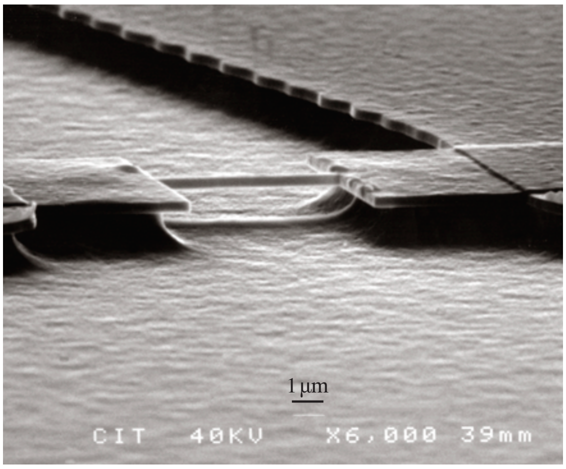

Fig. 7.5 SEM micrograph of a 3C-SiC nanomechanical beam resonator fabricated by electron-beam lithography and dry etching processes (courtesy M. Roukes, Caltech)

plasma using an evaporated Cr etch mask. After patterning, the beams were released by etching isotropically the underlying Si using a NF_3/Ar plasma. The inertness of the SiC film to the Si etchant enables the dry release of the nanomechanical beams. An example of a 3C-SiC nanomechanical beam is shown in Fig. 7.5.

7.4.2 Diamond

Diamond is commonly known as nature's hardest material, making it ideal for high wear environments. Diamond has a very large electronic bandgap (5.5 eV), which makes it attractive for high temperature electronics. Undoped diamond is a high quality insulator with a dielectric constant of 5.5; however it can be relatively easily doped with boron to create p-type conductivity. Diamond has a very high Young's modulus (1,035 GPa), making it suitable for high frequency micromachined resonators, and it is among nature's most chemically inert materials, making it well suited for harsh chemical environments.

Unlike SiC, fabrication of diamond MEMS is currently restricted to polycrystalline and amorphous material, since epitaxial-quality single-crystal diamond wafers are not yet commercially available. Polycrystalline diamond films can be deposited on Si and SiO_2 substrates, but the surfaces often must be seeded by diamond powders, or biased with a negative charge to initiate growth. In general, diamond nucleates much more readily on Si surfaces than on SiO_2 surfaces, an effect that has been used to selectively pattern diamond films into micromachined AFM cantilever probes using SiO_2 molding masks [7.58].

Bulk micromachining of diamond is very difficult given its extreme chemical inertness. Diamond structures have nevertheless been fabricated using bulk micromachined Si molds to pattern the structures [7.59]. The Si molds were fabricated using conventional micromachining techniques and filled with polycrystalline diamond deposited by hot filament chemical vapor deposition (HFCVD). The HFCVD process uses hydrogen as a carrier gas and methane as the carbon source. The process was performed at a substrate temperature of 850 °C to 900 °C and a pressure of 50 mtorr. The Si substrate was seeded prior to deposition using a diamond particle/ethanol solution. After deposition, the top surface of the structure was polished using a hot iron plate. After polishing, the Si mold was removed in a Si etchant, leaving behind the micromachined diamond structure. This process was used to produce high-aspect-ratio capillary channels for microfluidic applications [7.60] and components for diffractive optics, laser-to-fiber alignment, and power device cooling structures [7.61].

Due to the nucleation processes associated with diamond film growth, surface micromachining of polycrystalline diamond thin films requires modifications to conventional micromachining to facilitate film growth on sacrificial substrates. Conventional RIE methods are generally ineffective, so initial work focused on developing selective deposition techniques. One early method used selective seeding to form patterned templates for diamond nucleation. The selective seeding process employed the lithographic patterning of photoresist that contained diamond powders [7.62]. The diamond-loaded photoresist was deposited and patterned onto a Cr-coated Si wafer. During the onset of diamond growth, the patterned photoresist rapidly evaporates, leaving behind the diamond seed particles in the desired locations. A patterned diamond film is then selectively grown on these locations.

A second process utilizes selective deposition directly on sacrificial substrate layers. This process combines conventional diamond seeding with photolithographic patterning and etching to fabricate micromachined diamond structures on SiO_2 sacrificial layers [7.63]. The process can be performed in one of two ways. The first approach begins with the seeding of an oxidized Si wafer. The wafer is coated with photoresist and photolithographically patterned. Unmasked regions of the seeded SiO_2 film are then partially etched, forming a surface unfavorable for diamond growth. The photoresist is then removed, and a diamond film is deposited on the seeded regions. The second approach also begins with an oxidized Si wafer. The wafer is coated with photoresist, photolithographically patterned, and then seeded with diamond particles. The photoresist is removed, leaving behind a patterned seed layer suitable for selective growth. These techniques have been successfully used to fabricate cantilever beams and bridge structures.

A third method to surface micromachine polycrystalline diamond films follows the conventional approach of film deposition, dry etching, and release. The chemical inertness of diamond renders most conventional plasma chemistries useless; however, oxygen-based ion beam plasmas can be used to etch diamond thin films [7.64]. A simple surface micromachining process begins with the deposition of a polysilicon sacrificial layer on a Si_3N_4-coated Si wafer. The polysilicon layer is seeded using diamond slurry, and a diamond film is deposited by HFCVD. Since photoresist is not resistant to O_2 plasmas, an Al masking film is deposited and patterned. The diamond films are then etched in the O_2 ion beam plasma, and the structures are released by etching the polysilicon with KOH. This process has been used to create lateral resonant structures, but a significant stress gradient in the films rendered the devices nonoperational.

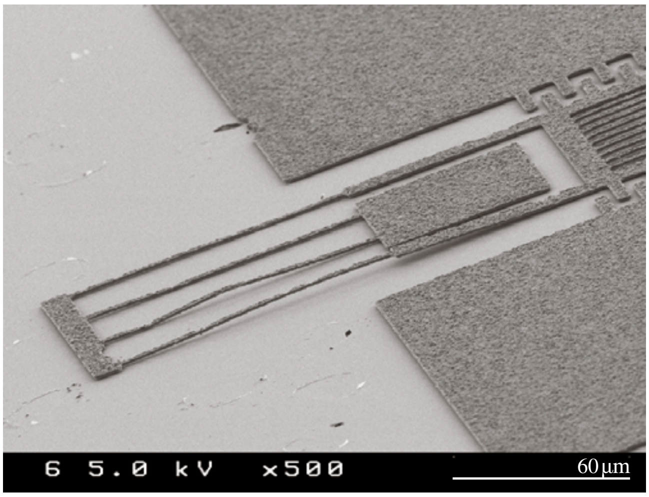

Fig. 7.6 SEM micrograph of the folded beam truss of a diamond lateral resonator. The diamond film was deposited using a seeding-based hot filament CVD process. The micrograph illustrates the challenges currently facing diamond MEMS, namely, roughened surfaces and residual stress gradients

In general, conventional HFCVD requires that the substrate be pretreated with a seeding layer prior to diamond film growth. However, a method called biased enhanced nucleation (BEN) has been developed that enables the growth of diamond on unseeded Si surfaces. *Wang* et al. [7.65] have shown that if Si substrates are masked with patterned SiO_2 films, selective diamond growth will occur primarily on the exposed Si surfaces, and a slight HF etch is sufficient to remove the adventitious diamond from the SiO_2 mask. This group was able to use this method to fabricate diamond micromotor rotors and stators on Si surfaces.

Diamond is a difficult but not impossible material to etch using RIE techniques. It is well-known that diamond will be etched in oxygen plasmas, but these plasmas are not generally well suited for device fabrication, because the etches tend to be isotropic. A recent development, however, suggests that RIE processes for diamond are close at hand. *Wang* et al. [7.65] describe a process to fabricate a vertically actuated, doubly clamped micromechanical diamond beam resonator using RIE. The process outlined here addresses two key issues related to diamond surface micromachining, namely, residual stress gradients in the diamond films and diamond patterning techniques. A microwave plasma CVD reactor was used to grow the diamond films on sacrificial SiO_2 layers pretreated with a nanocrystalline diamond powder, resulting in a uniform nucleation density at the diamond/SiO_2 interface. The diamond films were etched in a CF_4/O_2 plasma using Al as a hard mask. Reasonably straight sidewalls were created, with roughness attributable to the surface roughness of the faceted diamond film. An Au/Cr drive electrode beneath the sacrificial oxide remained covered throughout the diamond patterning steps and, thus, was undamaged during the diamond etching process.

In conjunction with recent advances in RIE and micromachining techniques, work is being performed to develop diamond deposition processes specifically for MEMS applications. Diamond films grown using conventional techniques, especially processes that require pre-growth seeding, tend to have high residual stress gradients and roughened surface morphologies as a result of the highly faceted, large grain polycrystalline films that are produced by these methods (see Fig. 7.6). The rough surface morphology degrades the patterning process, resulting in roughened sidewalls in etched structures and roughened surfaces of films deposited over these layers. Unlike polysilicon and SiC, a post-deposition polishing process is not technically feasible for diamond due to its extreme hardness. For the fabrication of multilayer diamond devices, methods to reduce the surface roughness of the as-deposited films are highly desirable. Along these lines, *Krauss* et al. [7.66] reported the development of an ultrananocrystalline diamond film that exhibits a much smoother surface morphology than comparable diamond films grown using conventional methods. Unlike conventional CVD diamond films that are grown using a mixture of H_2 and CH_4, the ultrananocrystalline diamond films are grown from mixtures of Ar, H_2, and C_{60} or Ar, H_2, and CH_4. Although the processes are still under development, *Krauss* et al. [7.66] have demonstrated the viability of the ultrananocrystalline diamond films as conformal coatings on Si surfaces, as well as their utility in several micromachining processes.

7.5 GaAs, InP, and Related III–V Materials

Gallium Arsenide (GaAs), Indium Phosphide (InP), and related III-V compounds have favorable piezoelectric and optoelectric properties, high piezoresistive constants, and wide electronic bandgaps relative to Si, making them attractive for various sensor and optoelectronic applications. Like Si, significant research in bulk crystal growth has led to the development of GaAs and InP substrates, which are commercially available as high quality, single crystal wafers. Unlike compound semiconductors, such as SiC, III-V materials can be deposited as ternary and quaternary alloys with lattice constants that closely match the binary compounds from which they are derived (i. e., $Al_xGa_{1-x}As$ and GaAs), thus permitting the fabrication of a wide variety of heterostructures that facilitate device performance.

Crystalline GaAs has a zinc blend crystal structure with an electronic bandgap of 1.4 eV, enabling GaAs electronic devices to function at temperatures as high as 350 °C [7.67]. High quality, single crystal wafers are commercially available, as are well developed metalorganic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) growth processes for epitaxial layers of GaAs and its alloys. GaAs does not outperform Si in terms of mechanical properties, however, its stiffness and fracture toughness are still suitable for micromechanical devices.

Micromachining of GaAs is relatively straightforward, since many of its lattice-matched ternary and quaternary alloys have sufficiently different chemical properties to allow their use as sacrificial layers [7.68]. For example, the most common ternary alloy for GaAs is $Al_xGa_{1-x}As$. For values of x less than or equal to 0.5, etchants containing mixtures of HF and H_2O will etch $Al_xGa_{1-x}As$ without attacking GaAs, while etchants containing NH_4OH and H_2O_2 attack GaAs isotropically, but do not etch $Al_xGa_{1-x}As$. Such selectivity enables the micromachining of GaAs wafers using lattice-matched etch stops and sacrificial layers. Devices fabricated using these methods include comb drive lateral resonant structures [7.68], pressure sensors [7.69, 70], thermopile sensors [7.71], Fabry–Perot detectors [7.72], and cantilever-based sensors and actuators [7.73, 74]. In addition, nanoelectromechanical devices, such as suspended micromechanical resonators [7.75] and tethered membranes [7.76], have been fabricated using these techniques. An example of a nanoelectromechanical beam structure fabricated from GaAs is shown in Fig. 7.7.

In addition to using epitaxial layers as etch stops, ion implantation methods can also be used to produce etch stops in GaAs layers. *Miao* et al. [7.77] describe a process that uses electrochemical etching to selectively remove n-type GaAs layers. The process relies on the creation of a highly resistive near surface GaAs layer on an n-type GaAs substrate by low dose nitrogen implantation in the MeV energy range. A pulsed electrochemical etch method using a H_2PtCl_6, H_3PO_4, H_2SO_4 platinum electrolytic solution at 40 °C with 17 V, 100 ms pulses is sufficient to selectively remove n-type GaAs at about 3 µm/min. Using this method, stress-free, tethered membranes could readily be fabricated from the high resistive GaAs layer. The high implant energies enable the fabrication of membranes several microns in thickness. Moreover, the authors demonstrated that if the GaAs wafer were etched in such a way as to create an undulating surface prior to ion implantation, corrugated membranes could be fabricated. These structures can sustain much higher deflection amplitudes than flat structures.

Micromachining of InP closely resembles the techniques used for GaAs. Many of the properties of InP are similar to GaAs in terms of crystal structure, mechanical stiffness, and hardness. However, the optical properties of InP make it particularly attractive for micro-optomechanical devices to be used in the 1.3 to 1.55 µm wavelength range [7.78]. Like GaAs, single crystal wafers of InP are readily available, and

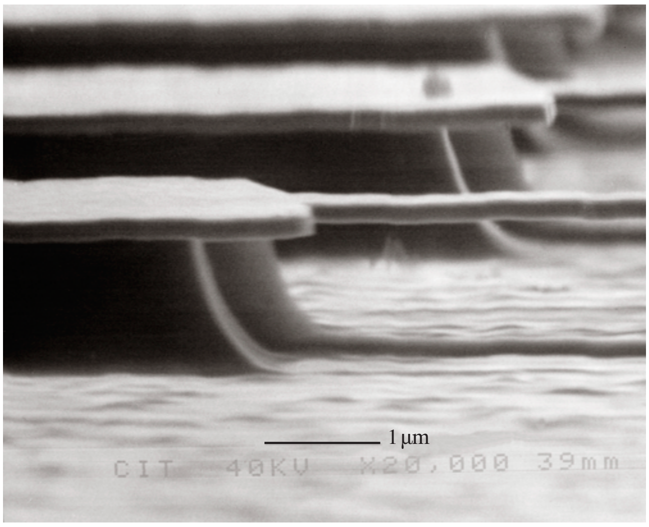

Fig. 7.7 SEM micrograph of a GaAs nanomechanical beam resonator fabricated by epitaxial growth, electron-beam lithography, and selective etching (courtesy of M. Roukes, Caltech)

ternary and quaternary lattice-matched alloys, such as InGaAs, InAlAs, InGaAsP, and InGaAlAs, can be used as either etch stop and/or sacrificial layers, depending on the etch chemistry [7.68]. For instance, InP structural layers deposited on $In_{0.53}Al_{0.47}As$ sacrificial layers can be released using etchants containing $C_6H_8O_7$, H_2O_2, and H_2O. In addition, InP films and substrates can be etched in solutions containing HCl and H_2O using $In_{0.53}Ga_{0.47}As$ films as etch stops. Using InP-based micromachining techniques, multi-air gap filters [7.79], bridge structures [7.78], and torsional membranes [7.72] have been fabricated from InP and its related alloys.

In addition to GaAs and InP, materials such as indium arsenide (InAs) can be micromachined into device structures. Despite a 7% lattice mismatch between InAs and (111) GaAs, high quality epitaxial layers can be grown on GaAs substrates. As described by *Yamaguchi* et al. [7.80], the surface Fermi level of InAs/GaAs structures is pinned in the conduction band, enabling the fabrication of very thin conductive membranes. In fact, the authors have successfully fabricated freestanding InAs structures that range in thickness from 30 to 300 nm. The thin InAs films were grown directly on GaAs substrates by MBE and etched using a solution containing H_2O, H_2O_2, and H_2SO_4. The structures, mainly doubly clamped cantilevers, were released by etching the GaAs substrate using a $H_2O/H_2O_2/NH_4OH$ solution.

7.6 Ferroelectric Materials

Piezoelectric materials play an important role in MEMS technology for sensing and mechanical actuation applications. In a piezoelectric material, mechanical stress produces a polarization, and, conversely, a voltage-induced polarization produces a mechanical stress. Many asymmetric materials, such as quartz, GaAs, and zinc oxide (ZnO), exhibit some piezoelectric behavior. Recent work in MEMS has focussed on the development of ferroelectric compounds, such as lead zirconate titanate, $Pb(Zr_xTi_{1-x})O_3$, or PZT for short, because such compounds have high piezoelectric constants that result in high mechanical transduction. It is relatively straightforward to fabricate a PZT structure on top of a thin freestanding structural layer (i. e., cantilever, diaphragm). Such a capability enables the piezoelectric material to be used in sensor applications or actuator applications for which piezoelectric materials are particularly well suited. Like Si, PZT films can be patterned using dry etch techniques based on chlorine chemistries, such as Cl_2/CCl_4, as well as ion beam milling using inert gases like Ar.

PZT has been successfully deposited in thin film form using co-sputtering, CVD, and sol-gel processing. Sol-gel processing is particularly attractive because the composition and homogeneity of the deposited material over large surface areas can be readily controlled. The sol-gel process outlined by *Lee* et al. [7.81] uses PZT solutions made from liquid precursors containing Pb, Ti, Zr, and O. The solution is deposited by spin coating on a Si wafer that has been coated with a $Pt/Ti/SiO_2$ thin film multilayer. The process is executed to produce a PZT film in layers, with each layer consisting of a spin-coated layer that is dried at 110 °C for 5 min and then heat treated at 600 °C for 20 min. After building up the PZT layer to the desired thickness, the multilayer was heated at 600 °C for up to 6 hrs. Prior to this anneal, a PbO top layer was deposited on the PZT surface. An Au/Cr electrode was then sputter-deposited on the surface of the piezoelectric stack. This process was used to fabricate a PZT-based force sensor. *Xu* et al. [7.82] describe a similar sol-gel process to produce 12 μm-thick, crack-free PZT films on Pt-coated Si wafers and 5 μm-thick films on insulating ZrO_2 layers to produce micromachined MHz-range, two-dimensional transducer arrays for acoustic imaging.

Thick film printing techniques for PZT have been developed to produce thick films in excess of 100 μm. Such thicknesses are desired for applications that require actuation forces that cannot be achieved with the much thinner sol-gel films. *Beeby* et al. [7.83] describe a thick film printing process whereby a PZT paste is made from a mixture of 95% PZT powder, 5% lead borosilicate powder, and an organic carrier. The paste was then printed through a stainless steel screen using a thick film printer. Printing was performed on an oxidized Si substrate that is capped with a Pt electrode. After printing, the paste was dried and then fired at 850 °C to 950 °C. Printing could be repeated to achieve the desired thickness. The top electrode consisted of an evaporated Al film. The authors found that it was possible to perform plasma-based processing on the printed substrates, but that the porous nature of the printed PZT films made them unsuitable for wet chemical processing.

7.7 Polymer Materials

Polyimides comprise an important class of durable polymers that are well suited for many of the techniques used in conventional MEMS processing. In general, polyimides can be acquired in bulk, or deposited as thin films by spin coating. Polyimides can be patterned using conventional dry etching techniques and processed at relatively high temperatures. These attributes make polyimides an attractive group of polymers for MEMS that require polymer structural and/or substrate layers, such as microfabricated biomedical devices, for which inertness and flexibility are important parameters.

7.7.1 Polyimide

Shearwood et al. [7.84] explored the use of polyimide as a robust mechanical material for microfabricated audio membranes. The authors fabricated 7 μm-thick, 8 mm-diameter membranes on GaAs substrates by bulk micromachining the GaAs substrate using a NH_3/H_2O_2 solution. They realized 100% yield and, despite a low Young's modulus (~ 3 GPa), observed flat membranes to within 1 nm after fabrication.

Jiang et al. [7.85] capitalized on the strength and flexibility of polyimide to fabricate a flexible sheer-stress sensor array based on Si sensors. The sensor array consisted of a collection of Si islands linked by two polyimide layers. Each Si sensor island was $250 \times 250\,\mu m^2$ in area and 80 μm in thickness. Al was used as an electrical innerconnect layer. The two polyimide layers served as highly flexible hinges, giving the sensor array the ability to be mounted on curved surfaces. The sensor array was successful in profiling the shear-stress distribution along the leading edge of a rounded delta wing.

The chemical and temperature durability of polyimides enables their use as a sacrificial layer for a number of commonly used materials, such as evaporated or sputter-deposited metals. *Memmi* et al. [7.86] developed a fabrication process for capacitive micromechanical ultransonic transducers using polyimide as a sacrificial layer. The authors showed that the polyimide could withstand the conditions used to deposit silicon monoxide by evaporation and silicon nitride by PECVD at 400 °C. Recent work by *Bagolini* et al. [7.87] has shown that polyimides can even be used as sacrificial layers for PECVD SiC.

In the area of microfabricated biomedical devices, polyimide is receiving attention as a substrate material for implantable devices, owing to its potential biocompatiblity and mechanical flexibility. *Stieglitz* [7.88] reported on the fabrication of multichannel microelectrodes on polyimide substrates. Instead of using polyimide sheets as starting substrates, Si carrier wafers coated with a 5 μm-thick polyimide film were used. Pt microelectrodes were then fabricated on these substrates using conventional techniques. Thin polyimide layers were deposited between various metal layers to serve as insulating layers. A capping polyimide layer was then deposited on the top of the substrates, and the entire polyimide/metal structure was peeled off the Si carrier wafers. Backside processing was then performed on the freestanding polyimide structures to create devices that have exposed electrodes on both surfaces. In a later paper, *Stieglitz* et al. [7.89] describe a variation of this process for neural prostheses.

7.7.2 SU-8

SU-8 is a negative-tone, epoxy-like photoresist that is receiving much attention for its versatility in MEMS processing. It is a high-aspect-ratio, UV-sensitive resist designed for applications requiring single-coat resists with thicknesses on the order of 500 microns [7.90]. SU-8 has favorable chemical properties that enable it to be used as a molding material for high-aspect-ratio electroplated structures (as an alternative to LIGA), and as a structural material for microfluidics [7.90]. In terms of mechanical properties, *Lorenz* et al. [7.91] reported that SU-8 has a modulus of elasticity of 4.02 GPa, which compares favorably with a commonly used polyamid (3.4 GPa).

In addition to the aforementioned conventional uses for SU-8, several interesting alternative uses are beginning to appear in the literature. *Conradie* et al. [7.92] have used SU-8 to trim the mass of silicon paddle oscillators as a means to adjust the resonant frequency of the beams. The trimming process involves the patterning of SU-8 posts on Si paddles. The process capitalizes on the relative chemical stability of the SU-8 resin, in conjunction with the relatively large masses that can be patterned using standard UV exposure processes.

SU-8 is also of interest as a bonding layer material for wafer bonding processes using patterned bonding layers. *Pan* et al. [7.93] compared several UV photo-definable polymeric materials and found that SU-8 exhibited the highest bonding strength (20.6 MPa) for layer thicknesses up to 100 μm.

7.7.3 Parylene

Parylene (poly-paraxylylene) is another emerging polymeric MEMS material due in large part to its biocompatibility. It is particularly attractive from the fabrication perspective, because it can be deposited by CVD at room temperature. Moreover, the deposition process is conformal, which enables parylene coatings to be applied to prefabricated structures, such as Si microneedles [7.94], low stress silicon nitride membrane particle filters [7.95], and micromachined polyimide/Au optical scanners [7.96]. In the first two cases, the parylene coating served to strengthen the microfabricated structures, while in the last case, it served to protect the structure from condensing water vapor.

In addition to its function as a protective coating, parylene can actually be micromachined into freestanding components. *Noh* et al. [7.97] demonstrated a method to create bulk micromachined parylene microcolumns for miniature gas chromatographs. The structure is fabricated using a micromolding technique by which Si molds are fabricated by DRIE and coated with parylene to form three sides of the microcolumn. A second wafer is coated with parylene, and the two are bonded together via a fusion bonding process. After bonding, the structure is released from the Si mold by KOH etching. In a second example, *Yao* et al. [7.98] describe a dry release process for parylene surface micromachining. In this process, sputtered Si is used as a sacrificial layer onto which a thick sacrificial photoresist is deposited. Parylene is then deposited on the photoresist and patterned into the desired structural shape. The release procedure is a two step process. First, the photoresist is dissolved in acetone. This results in the parylene structure sticking to the sputtered Si. Next, a dry BrF_3 etch is performed, which dissolves the Si and releases the parylene structures. Parylene beams that were 1 mm long and 4.5 μm thick were successfully fabricated using this technique.

7.8 Future Trends

The rapid expansion of MEMS in recent years is due in large part to the inclusion of new materials that have expanded the functionality of microfabricated devices beyond what is achievable in silicon. This trend will certainly continue as new application areas for micro- and nanofabricated devices are identified. Many of these applications will likely require both new materials and new processes to fabricate the micro- and nanomachined devices for these yet to be identified applications. Currently, conventional micromachining techniques employ a top-down approach that begins with either bulk substrates or thin films. Future MEMS and NEMS will likely incorporate materials that are created using a bottom-up approach. A significant challenge facing device design and fabrication engineers alike will be how to marry top-down and bottom-up approaches to create devices and systems that cannot be made using either process alone.

References

7.1 C. S. Smith: Piezoresistive effect in germanium and silicon, Phys. Rev. **94** (1954) 1–10

7.2 A. N. Cleland, M. L. Roukes: Fabrication of high frequency nanometer scale mechanical resonators from bulk Si crystals, Appl. Phys. Lett. **69** (1996) 2653–2655

7.3 D. W. Carr, H. G. Craighead: Fabrication of nanoelectromechanical systems in single crystal silicon using silicon on insulator substrates and electron beam lithography, J. Vac. Sci. Technol. B **15** (1997) 2760–2763

7.4 J. J. McMahon, J. M. Melzak, C. A. Zorman, J. Chung, M. Mehregany: Deposition and characterization of in situ boron doped polycristalline silicon films for microelectromechanical systems applications, Mater. Res. Soc. Proc. **605** (2000) 31–36

7.5 T. Kamins: *Polycrystalline Silicon for Integrated Circuits and Displays*, 2nd edn. (Kluwer, Boston 1998)

7.6 L. Cao, T. S. Kin, S. C. Mantell, D. Polla: Simulation and fabrication of piezoresistive membrane type MEMS strain sensors, Sens. Actuators **80** (2000) 273–279

7.7 H. Guckel, T. Randazzo, D. W. Burns: A simple technique for the determination of mechanical strain in thin films with application to polysilicon, J. Appl. Phys. **57** (1985) 1671–1675

7.8 R. T. Howe, R. S. Muller: Stress in polysilicon and amorphous silicon thin films, J. Appl. Phys. **54** (1983) 4674–4675

7.9 X. Zhang, T. Y. Zhang, M. Wong, Y. Zohar: Rapid thermal annealing of polysilicon thin films, J. Microelectromech. Syst. **7** (1998) 356–364

7.10 J. Yang, H. Kahn, A.-Q. He, S. M. Phillips, A. H. Heuer: A new technique for producing large-area as-deposited zero-stress LPCVD polysilicon films: The multipoly process, J. Microelectromech. Syst. **9** (2000) 485–494

7.11 P. Gennissen, M. Bartek, P. J. French, P. M. Sarro: Bipolar-compatible epitaxial poly for smart sensors: Stress minimization and applications, Sens. Actuators A **62** (1997) 636–645

7.12 P. Lange, M. Kirsten, W. Riethmuller, B. Wenk, G. Zwicker, J. R. Morante, F. Ericson, J. A. Schweitz: Thick polycrystalline silicon for surface-micromechanical applications: Deposition, structuring, and mechanical characterization, Sens. Actuators A **54** (1996) 674–678

7.13 S. Greek, F. Ericson, S. Johansson, M. Furtsch, A. Rump: Mechanical characterization of thick polysilicon films: Young's modulus and fracture strength evaluated with microstructures, **9** (1999) 245–251

7.14 K. Funk, H. Emmerich, A. Schilp, M. Offenberg, R. Neul, F. Larmer: A surface micromachined silicon gyroscope using a thick polysilicon layer, Proc. of the 12th Int. Conf. Microelectromech. Systems, Orlando 1999, ed. by K. J. Gabriel, K. Najafi (IEEE, Piscataway 1999) 57–60

7.15 T. Abe, M. L. Reed: Low strain sputtered polysilicon for micromechanical structures, Proc. 9th Int. Workshop Microelectromech. Systems, San Diego 1996, ed. by M. G. Allen, M. L. Reed (IEEE, Piscataway 1996) 258–262

7.16 K. Honer, G. T. A. Kovacs: Integration of sputtered silicon microstructures with pre-fabricated CMOS circuitry, Sens. Actuators A **91** (2001) 392–403

7.17 R. Anderson, R. S. Muller, C. W. Tobias: Porous polycrystalline silicon: A new material for MEMS, J. Microelectromech. Syst. **3** (1994) 10–18

7.18 W. Lang, P. Steiner, H. Sandmaier: Porous silicon: A novel material for microsystems, Sens. Actuators A **51** (1995) 31–36

7.19 S. K. Ghandhi: *VLSI Fabrication Principles – Silicon and Gallium Arsenide* (Wiley, New York 1983)

7.20 W. A. Pilskin: Comparison of properties of dielectric films deposited by various methods, J. Vac. Sci. Technol. **21** (1977) 1064–1081

7.21 J. S. Danel, F. Michel, G. Delapierre: Micromachining of quartz and its application to an acceleration sensor, Sens. Actuators A **21–23** (1990) 971–977

7.22 A. Yasseen, J. D. Cawley, M. Mehregany: Thick glass film technology for polysilicon surface micromachining, J. Microelectromech. Syst. **8** (1999) 172–179

7.23 R. Liu, M. J. Vasile, D. J. Beebe: The fabrication of nonplanar spin-on glass microstructures, J. Microelectromech. Syst. **8** (1999) 146–151

7.24 B. Folkmer, P. Steiner, W. Lang: Silicon nitride membrane sensors with monocrystalline transducers, Sens. Actuators A **51** (1995) 71–75

7.25 M. Sekimoto, H. Yoshihara, T. Ohkubo: Silicon nitride single-layer X-ray mask, J. Vac. Sci. Technol. **21** (1982) 1017–1021

7.26 P. J. French, P. M. Sarro, R. Mallee, E. J. M. Fakkeldij, R. F. Wolffenbuttel: Optimization of a low-stress silicon nitride process for surface micromachining applications, Sens. Actuators A **58** (1997) 149–157

7.27 B. Li, B. Xiong, L. Jiang, Y. Zohar, M. Wong: Germanium as a versatile material for low-temperature micromachining, J. Microelectromech. Syst. **8** (1999) 366–372

7.28 A. Franke, D. Bilic, D. T. Chang, P. T. Jones, T. J. King, R. T. Howe, C. G. Johnson: Post-CMOS integration of germanium microstructures, Proc. 12th Int. Conf. Microelectromech. Syst., Orlando 1999, ed. by K. J. Gabriel, K. Najafi (IEEE, Piscataway 1999) 630–637

7.29 A. E. Franke, Y. Jiao, M. T. Wu, T. J. King, R. T. Howe: Post-CMOS modular integration of poly-SiGe microstructures using poly-Ge sacrificial layers, Technical Digest – Solid State Sens. Actuator Workshop,

Hilton Head 2000, ed. by L. Bousse (Transducers Research Foundation, Cleveland Heights Ohio 2000) 18–21
7.30 S. Sedky, P. Fiorini, M. Caymax, S. Loreti, K. Baert, L. Hermans, R. Mertens: Structural and mechanical properties of polycrystalline silicon germanium for micromachining applications, J. Microelectromech. Syst. **7** (1998) 365–372
7.31 S. Sedky, A. Witvrouw, K. Baert: Poly SiGe, a promising material for MEMS monolithic integration with the driving electronics, Sens. Actuators A **97-98** (2002) 503–511
7.32 J. M. Heck, C. G. Keller, A. E. Franke, L. Muller, T.-J. King, R. T. Howe: High aspect ratio polysilicon-germanium microstructures, Proc. 10th Int. Conf. Solid State Sens. Actuators, Sendai 1999, ed. by M. Esashi, T. Morizumi (IEE of Japan, Tokyo 1999) 328–334
7.33 P. Van Gerwen, T. Slater, J. B. Chevrier, K. Baert, R. Mertens: Thin-film boron-doped polycrystalline silicon$_{70\%}$-germanium$_{30\%}$ for thermopiles, Sens. Actuators A **53** (1996) 325–329
7.34 D. Hyman, J. Lam, B. Warneke, A. Schmitz, T. Y. Hsu, J. Brown, J. Schaffner, A. Walson, R. Y. Loo, M. Mehregany, J. Lee: Surface micromachined RF MEMS switches on GaAs substrates, Int. J. Radio Frequency Microwave Commun. Eng. **9** (1999) 348–361
7.35 C. Chang, P. Chang: Innovative micromachined microwave switch with very low insertion loss, Sens. Actuators **79** (2000) 71–75
7.36 C. L. Shih, B. K. Lai, H. Kahn, S. M. Phillips, A. H. Heuer: A robust co-sputtering fabrication procedure for TiNi shape memory alloys for MEMS, J. Microelectromech. Syst. **10** (2001) 69–79
7.37 G. Hahm, H. Kahn, S. M. Phillips, A. H. Heuer: Fully microfabricated silicon spring biased shape memory actuated microvalve, Technical Digest – Solid State Sens. Actuator Workshop, Hilton Head Island 2000, ed. by L. Bousse (Transducers Research Foundation, Cleveland Heights Ohio 2000) 230–233
7.38 S. D. Leith, D. T. Schwartz: High-rate through-mold electrodeposition of thick (>200 micron) NiFe MEMS components with uniform composition, J. Microelectromech. Syst. **8** (1999) 384–392
7.39 N. Rajan, M. Mehregany, C. A. Zorman, S. Stefanescu, T. Kicher: Fabrication and testing of micromachined silicon carbide and nickel fuel atomizers for gas turbine engines, J. Microelectromech. Syst. **8** (1999) 251–257
7.40 T. Pornsin-sirirak, Y. C. Tai, H. Nassef, C. M. Ho: Titanium-alloy MEMS wing technology for a microaerial vehicle application, Sens. Actuators A **89** (2001) 95–103
7.41 C. R. Stoldt, C. Carraro, W. R. Ashurst, D. Gao, R. T. Howe, R. Maboudian: A low temperature CVD process for silicon carbide MEMS, Sens. Actuators A **97-98** (2002) 410–415
7.42 M. Eickhoff, H. Moller, G. Kroetz, J. von Berg, R. Ziermann: A high temperature pressure sensor prepared by selective deposition of cubic silicon carbide on SOI substrates, Sens. Actuators **74** (1999) 56–59
7.43 Y. T. Yang, K. L. Ekinci, X. M. H. Huang, L. M. Schiavone, M. L. Roukes, C. A. Zorman, M. Mehregany: Monocrystalline silicon carbide nanoelectromechanical systems, Appl. Phys. Lett. **78** (2001) 162–164
7.44 C. A. Zorman, S. Rajgolpal, X. A. Fu, R. Jezeski, J. Melzak, M. Mehregany: Deposition of polycrystalline 3C-SiC films on 100 mm-diameter (100) Si wafers in a large-volume LPCVD furnace, Electrochem. Solid State Lett. **5** (2002) G99–G101
7.45 I. Behrens, E. Peiner, A. S. Bakin, A. Schlachetzski: Micromachining of silicon carbide on silicon fabricated by low-pressure chemical vapor deposition, J. Micromech. Microeng. **12** (2002) 380–384
7.46 C. A. Zorman, S. Roy, C. H. Wu, A. J. Fleischman, M. Mehregany: Characterization of polycrystalline silicon carbide films grown by atmospheric pressure chemical vapor deposition on polycrystalline silicon, J. Mater. Res. **13** (1996) 406–412
7.47 C. H. Wu, C. H. Zorman, M. Mehregany: Growth of polycrystalline SiC films on SiO_2 and Si_3N_4 by APCVD, Thin Solid Films **355-356** (1999) 179–183
7.48 P. Sarro: Silicon carbide as a new MEMS technology, Sens. Actuators **82** (2000) 210–218
7.49 N. Ledermann, J. Baborowski, P. Muralt, N. Xantopoulos, J. M. Tellenbach: Sputtered silicon carbide thin films as protective coatings for MEMS applications, Surf. Coat. Technol. **125** (2000) 246–250
7.50 R. S. Okojie, A. A. Ned, A. D. Kurtz: Operation of a 6H-SiC pressure sensor at 500 °C, Sens. Actuators A **66** (1998) 200–204
7.51 K. Lohner, K. S. Chen, A. A. Ayon, M. S. Spearing: Microfabricated silicon carbide microengine structures, Mater. Res. Soc. Symp. Proc. **546** (1999) 85–90
7.52 S. Tanaka, S. Sugimoto, J.-F. Li, R. Watanabe, M. Esashi: Silicon carbide micro-reaction-sintering using micromachined silicon molds, J. Microelectromech. Syst. **10** (2001) 55–61
7.53 X. Song, S. Rajgolpal, J. M. Melzak, C. A. Zorman, M. Mehregany: Development of a multilayer SiC surface micromachining process with capabilities and design rules comparable with conventional polysilicon surface micromachining, Mater. Sci. Forum **389-393** (2001) 755–758
7.54 L. A. Liew, W. Zhang, V. M. Bright, A. Linan, M. L. Dunn, R. Raj: Fabrication of SiCN ceramic MEMS using injectable polymer-precursor technique, Sens. Actuators A **89** (2001) 64–70
7.55 A. J. Fleischman, S. Roy, C. A. Zorman, M. Mehregany: Polycrystalline silicon carbide for surface micromachining, Proc.9th Int. Workshop Microelectromech. Systems, San Diego 1996, ed. by M. G. Allen, M. L. Reid (IEEE, Piscataway 1996) 234–238

7.56 A. J. Fleischman, X. Wei, C. A. Zorman, M. Mehregany: Surface micromachining of polycrystalline SiC deposited on SiO_2 by APCVD, Mater. Sci. Forum **264-268** (1998) 885–888

7.57 A. Yasseen, C. H. Wu, C. A. Zorman, M. Mehregany: Fabrication and testing of surface micromachined polycrystalline SiC micromotors, Electron Device Lett. **21** (2000) 164–166

7.58 T. Shibata, Y. Kitamoto, K. Unno, E. Makino: Micromachining of diamond film for MEMS applications, J. Microelectromech. Syst. **9** (2000) 47–51

7.59 H. Bjorkman, P. Rangsten, P. Hollman, K. Hjort: Diamond replicas from microstructured silicon masters, Sens. Actuators **73** (1999) 24–29

7.60 P. Rangsten, H. Bjorkman, K. Hjort: Microfluidic components in diamond, Proc. 10th Int. Conf. Solid State Sens. Actuators, Sendai 1999, ed. by M. Esashi, T. Morizumi (IEE of Japan, Tokyo 1999) 190–193

7.61 H. Bjorkman, P. Rangsten, K. Hjort: Diamond microstructures for optical microelectromechanical systems, Sens. Actuators **78** (1999) 41–47

7.62 M. Aslam, D. Schulz: Technology of diamond microelectromechanical systems, Proc. 8th Int. Conf. Solid State Sens. Actuators, Stockholm 1995, ed. by I. Lundstrum (Foundation for Sensor and Actuator Technology, Stockholm 1995) 222–224

7.63 R. Ramesham: Fabrication of diamond microstructures for microelectromechanical systems (MEMS) by a surface micromachining process, Thin Solid Films **340** (1999) 1–6

7.64 Y. Yang, X. Wang, C. Ren, J. Xie, P. Lu, W. Wang: Diamond surface micromachining technology, Diamond Rel. Mater. **8** (1999) 1834–1837

7.65 X. D. Wang, G. D. Hong, J. Zhang, B. L. Lin, H. Q. Gong, W. Y. Wang: Precise patterning of diamond films for MEMS application, J. Mater. Processing Technol. **127** (2002) 230–233

7.66 A. R. Krauss, O. Auciello, D. M. Gruen, A. Jayatissa, A. Sumant, J. Tucek, D. C. Mancini, N. Moldovan, A. Erdemire, D. Ersoy, M. N. Gardos, H. G. Busmann, E. M. Meyer, M. Q. Ding: Ultrananocrystalline diamond thin films for MEMS and moving mechanical assembly devices, Diamond Rel. Mater. **10** (2001) 1952–1961

7.67 K. Hjort, J. Soderkvist, J.-A. Schweitz: Galium arsenide as a mechanical material, J. Micromech. Microeng. **4** (1994) 1–13

7.68 K. Hjort: Sacrificial etching of III-V compounds for micromechanical devices, J. Micromech. Microeng. **6** (1996) 370–365

7.69 K. Fobelets, R. Vounckx, G. Borghs: A GaAs pressure sensor based on resonant tunnelling diodes, J. Micromech. Microeng. **4** (1994) 123–128

7.70 A. Dehe, K. Fricke, H. L. Hartnagel: Infrared thermopile sensor based on AlGaAs-GaAs micromachining, Sens. Actuators A **46-47** (1995) 432–436

7.71 A. Dehe, K. Fricke, K. Mutamba, H. L. Hartnagel: A piezoresistive GaAs pressure sensor with GaAs/AlGaAs membrane technology, J. Micromech. Microeng. **5** (1995) 139–142

7.72 A. Dehe, J. Peerlings, J. Pfeiffer, R. Riemenschneider, A. Vogt, K. Streubel, H. Kunzel P. Meissner, H. L. Hartnagel: III-V compound semiconductor micromachined actuators for long resonator tunable Fabry–Perot detectors, Sens. Actuators A **68** (1998) 365–371

7.73 T. Lalinsky, S. Hascik, Z. Mozolova, E. Burian, M. Drzik: The improved performance of GaAs micromachined power sensor microsystem, Sens. Actuators **76** (1999) 241–246

7.74 T. Lalinsky, E. Burian, M. Drzik, S. Hascik, Z. Mozolova, J. Kuzmik, Z. Hatzopoulos: Performance of GaAs micromachined microactuator, Sens. Actuators **85** (2000) 365–370

7.75 H. X. Tang, X. M. H. Huang, M. L. Roukes, M. Bichler, W. Wegscheider: Two-dimensional electron-gas actuation and transduction for GaAs nanoelectromechanical systems, Appl. Phys. Lett. **81** (2002) 3879–3881

7.76 T. S. Tighe, J. M. Worlock, M. L. Roukes: Direct thermal conductance measurements on suspended monocrystalline nanostructures, Appl. Phys. Lett. **70** (1997) 2687–2689

7.77 J. Miao, B. L. Weiss, H. L. Hartnagel: Micromachining of three-dimensional GaAs membrane structures using high-energy nitrogen implantation, J. Micromech. Microeng. **13** (2003) 35–39

7.78 C. Seassal, J. L. Leclercq, P. Viktorovitch: Fabrication of InP-based freestanding microstructures by selective surface micromachining, J. Micromech. Microeng. **6** (1996) 261–265

7.79 J. Leclerq, R. P. Ribas, J. M. Karam, P. Viktorovitch: III-V micromachined devices for microsystems, Microelectron. J. **29** (1998) 613–619

7.80 H. Yamaguchi, R. Dreyfus, S. Miyashita, Y. Hirayama: Fabrication and elastic properties of InAs freestanding structures based on InAs/GaAs(111) A heteroepitaxial systems, Physica E **13** (2002) 1163–1167

7.81 J. B. Lee, J. English, C. H. Ahn, M. G. Allen: Planarization techniques for vertically integrated metallic MEMS on silicon foundry circuits, J. Micromech. Microeng. **7** (1997) 44–54

7.82 B. Xu, L. E. Cross, J. J. Bernstein: Ferroelectric and antiferroelectric films for microelectromechanical systems applications, Thin Solid Films **377-378** (2000) 712–718

7.83 S. P. Beeby, A. Blackburn, N. M. White: Processing of PZT piezoelectric thick films on silicon for microelectromechanical systems, J. Micromech. Microeng. **9** (1999) 218–229

7.84 C. Shearwood, M. A. Harradine, T. S. Birch, J. C. Stevens: Applications of polyimide membranes to MEMS technology, Microelectron. Eng. **30** (1996) 547–550

7.85 F. Jiang, G.B. Lee, Y.C. Tai, C.M. Ho: A flexible micromachine-based shear-stress sensor array and its application to separation-point detection, Sens. Actuators **79** (2000) 194–203

7.86 D. Memmi, V. Foglietti, E. Cianci, G. Caliano, M. Pappalardo: Fabrication of capacitive micromechanical ultrasonic transducers by low-temperature process, Sens. Actuators A **99** (2002) 85–91

7.87 A. Bagolini, L. Pakula, T. L. M. Scholtes, H. T. M. Pham, P. J. French, P. M. Sarro: Polyimide sacrificial layer and novel materials for post-processing surface micromachining, J. Micromech. Microeng. **12** (2002) 385–389

7.88 T. Stieglitz: Flexible biomedical microdevices with double-sided electrode arrangements for neural applications, Sens. Actuators A **90** (2001) 203–211

7.89 T. Stieglitz, G. Matthias: Flexible BioMEMS with electrode arrangements on front and back side as key component in neural prostheses and biohybrid systems, Sens. Actuators B **83** (2002) 8–14

7.90 H. Lorenz, M. Despont, N. Fahrni, J. Brugger, P. Vettiger, P. Renaud: High-aspect-ratio, ultrathick, negative-tone-near-UV photoresist and its applications in MEMS, Sens. Actuators A **64** (1998) 33–39

7.91 H. Lorenz, M. Despont, N. Fahrni, N. LaBianca, P. Renaud, P. Vettiger: SU-8: A low-cost negative resist for MEMS, J. Micromech. Microeng. **7** (1997) 121–124

7.92 E. H. Conradie, D. F. Moore: SU-8 thick photoresist processing as a functional material for MEMS applications, J. Micromech. Microeng. **12** (2002) 368–374

7.93 C. T. Pan, H. Yang, S. C. Shen, M. C. Chou, H. P. Chou: A low-temperature wafer bonding technique using patternable materials, J. Micromech. Microeng. **12** (2002) 611–615

7.94 P. A. Stupar, A. P. Pisano: Silicon parylene, and silicon/parylene micro-needles for strength and toughness, Technical Digest 11th Int. Conf. Solid State Sens. Actuators, Munich 2001, ed. by E. Obermeier (Springer, Berlin, Heidelberg 2001) 1368–1389

7.95 X. Yang, J. M. Yang, Y. C. Tai, C. M. Ho: Micromachined membrane particle filters, Sens. Actuators **73** (1999) 184–191

7.96 J. M. Zara, S.W. Smith: Optical scanner using a MEMS actuator, Sens. Actuators A **102** (2002) 176–184

7.97 H. S. Noh, P. J. Hesketh, G. C. Frye-Mason: Parylene gas chromatographic column for rapid thermal cycling, J. Microelectromech. Syst. **11** (2002) 718–725

7.98 T. J. Yao, X. Yang, Y. C. Tai: BrF_3 dry release technology for large freestanding parylene microstructures and electrostatic actuators, Sens. Actuators A **97–98** (2002) 771–775

8. MEMS/NEMS Devices and Applications

8.1 **MEMS Devices and Applications** 227
8.1.1 Pressure Sensor 227
8.1.2 Inertial Sensor 229
8.1.3 Optical MEMS 233
8.1.4 RF MEMS 239

8.2 **NEMS Devices and Applications** 246

8.3 **Current Challenges and Future Trends** 249

References 250

Microelectromechanical Systems (MEMS) have played key roles in many important areas, for example transportation, communication, automated manufacturing, environmental monitoring, health care, defense systems, and a wide range of consumer products. MEMS are inherently small, thus offering attractive characteristics such as reduced size, weight, and power dissipation and improved speed and precision compared to their macroscopic counterparts. Integrated circuits (IC) fabrication technology has been the primary enabling technology for MEMS besides a few special etching, bonding and assembly techniques. Microfabrication provides a powerful tool for batch processing and miniaturization of electromechanical devices and systems into a dimensional scale, which is not achievable by conventional machining techniques. As IC fabrication technology continues to scale toward deep sub-micron and nano-meter feature sizes, a variety of nanoelectromechanical systems (NEMS) can be envisioned in the foreseeable future. Nano-scale mechanical devices and systems integrated with nanoelectronics will open a vast number of new exploratory research areas in science and engineering. NEMS will most likely serve as an enabling technology merging engineering with the life sciences in ways that are not currently feasible with the micro-scale tools and technologies. MEMS has been applied to a wide range of fields. Over hundreds of micro-devices have been developed for specific applications. It is thus difficult to provide an overview covering every aspect of the topic. In this chapter, key aspects of MEMS technology and application impacts are illustrated through selecting a few demonstrative device examples, which consist of pressure sensors, inertial sensors, optical and wireless communication devices. Microstructure examples with dimensions on the order of sub-micron are presented with fabrication technologies for future NEMS applications. Although MEMS has experienced significant growth over the past decade, many challenges still remain. In broad terms, these challenges can be grouped into three general categories; (1) fabrication challenges, (2) packaging challenges, and (3) application challenges. Challenges in these areas will, in large measure, determine the commercial success of a particular MEMS device both in technical and economic terms. This chapter presents a brief discussion of some of these challenges as well as possible approaches to address them.

Microelectromechanical systems, generally referred to as MEMS, has had a history of research and development spanning a few decades. Besides the traditional microfabricated sensors and actuators, the field covers micromechanical components and systems integrated or micro-assembled with electronics on the same substrate or package, achieving high performance functional systems. These devices and systems have played key roles in many important areas, such as transportation, communication, automated manufacturing, environmental monitoring, health care, defense systems, and a wide range of consumer products. MEMS are inherently small, thus offering such attractive characteristics as reduced size, weight, and power dissipation, as well as improved speed and precision compared to their macroscopic counterparts. The development of MEMS requires appropriate fabrication technologies that enable the definition of small geometries, precise dimension

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

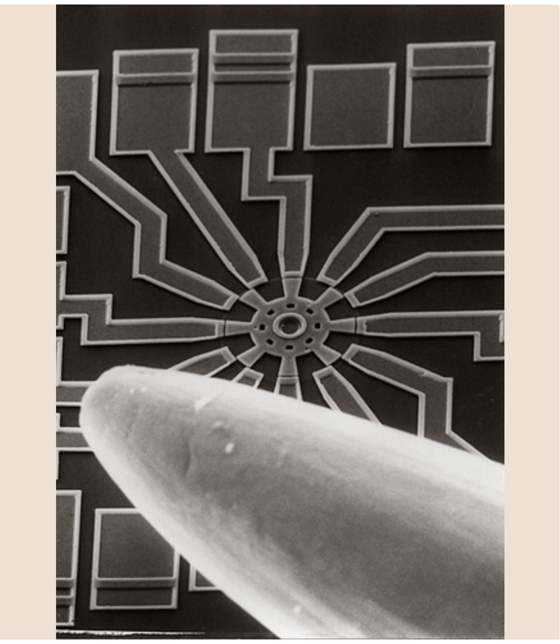

Fig. 8.1 SEM micrograph of a polysilicon microelectromechanical motor [8.1]

control, design flexibility, interfacing with microelectronics, repeatability, reliability, high yield, and low cost. Integrated circuits (IC) fabrication technology meets all of the above criteria and has been the primary enabling fabrication technology for MEMS, besides a few special etching, bonding and assembly techniques. Microfabrication provides a powerful tool for batch processing and miniaturization of electromechanical devices and systems into a dimensional scale, which is not achievable by conventional machining techniques. Most MEMS devices exhibit a length or width ranging from micrometers to several hundreds of micrometers with a thickness from submicrometer up to tens of micrometers, depending upon the fabrication technique employed. A physical displacement of a sensor or an actuator is typically on the same order of magnitude. Figure 8.1 shows an SEM micrograph of a microelectromechanical motor developed in the late 1980s [8.1]. Polycrystalline silicon (polysilicon) surface micromachining technology was used to fabricate the micromotor, achieving a diameter of 150 μm and a minimum vertical feature size on the order of a micrometer. A probe tip is also shown in the micrograph for a size comparison. This device example, and similar others [8.2], demonstrated what MEMS technology could accomplish in microscale machining at the time and served as a strong technology indicator for continued MEMS development. The field has expanded greatly in recent years along with rapid technology advances. Figure 8.2, for example, shows a photo of micro-gears fabricated in the mid-1990s using a five-level polysilicon surface micromachining technology [8.3]. This device represents one of the most advanced surface micromachining fabrication processes developed to date. One can imagine that a wide range of sophisticated microelectromechanical devices and systems can be realized through applying such technology in the future. As IC fabrication technology continues to scale toward deep submicron and nanometer feature sizes, a variety of nanoelectromechanical systems (NEMS) can be envisioned in the near future. Nanoscale mechanical devices and systems integrated with nanoelectronics will open a vast number of new exploratory research areas in science and engineering. NEMS will most likely serve as an enabling technology, merging engineering with the life sciences in ways that are not currently feasible with microscale tools and technologies.

This chapter will provide a general overview of MEMS and NEMS devices and their applications. MEMS technology has been applied to a wide range of fields. Over hundreds of micro-devices have been developed for specific applications. Therefore, it is difficult to provide an overview that covers every aspect of the topic. It is the authors' intent to illustrate key aspects of MEMS technology, and its impact on specific applications by selecting a few demonstrative device examples. For a wide-ranging discussion of nearly all types of micromachined sensors and actuators, books by *Kovacs* [8.4] and *Senturia* [8.5] are recommended.

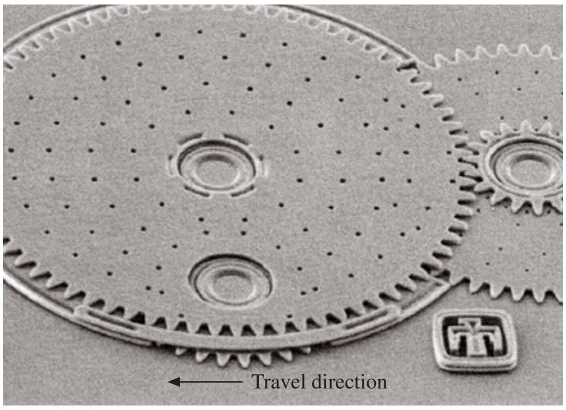

Fig. 8.2 SEM micrograph of polysilicon micro-gears [8.5]

8.1 MEMS Devices and Applications

MEMS devices have played key roles in many areas of development. Microfabricated sensors, actuators, and electronics are the most critical components in implementing a complete system for a specific function. Microsensors and actuators can be fabricated by various micromachining processing technologies. In this section, a number of selected MEMS devices are presented to illustrate the basic device operating principles, as well as demonstrate key aspects of microfabrication technology and its application impact.

8.1.1 Pressure Sensor

Pressure sensors are one of the early devices realized by silicon micromachining technologies and have become successful commercial products. The devices have been widely used in various industrial and biomedical applications. The sensors can be based on piezoelectric, piezoresistive, capacitive, and resonant sensing mechanisms. Silicon bulk and surface micromachining techniques have been used for sensor batch fabrication, thus achieving size miniaturization and low cost. Two types of pressure sensors – piezoresistive and capacitive – are described here for illustration purposes.

Piezoresistive Sensor

The piezoresistive effect in silicon has been widely used for implementing pressure sensors. A pressure-induced strain deforms the silicon band structure, thus changing the resistivity of the material. The piezoresistive effect is typically crystal-orientation dependent and is also affected by doping and temperature. A practical piezoresistive pressure sensor can be implemented by fabricating four sensing resistors along the edges of a thin silicon diaphragm, which acts as a mechanical amplifier to increase the stress and strain at the sensor site. The four sensing elements are connected in a bridge configuration with push-pull signals to increase the sensitivity. The measurable pressure range for such a sensor can be from 10^{-3} to 10^{6} torr, depending on the design. An example of a piezoresistive pressure sensor is shown in Fig. 8.3. The device consists of a silicon diaphragm suspended over a reference vacuum cavity to form an absolute pressure sensor. An external pressure applied over the diaphragm introduces a stress on the sensing resistors, resulting in a resistance value change corresponding to the pressure. The fabrication sequence is outlined as follows. The piezoresistors are typically first formed through a boron diffusion process, followed by a high temperature annealing step in order to achieve a resistance value on the order of a few kilo-ohms. The wafer is then passivated with a silicon dioxide layer, and contact windows are opened for metallization. At this point, the wafer is patterned on the backside, followed by a timed silicon wet etch to form the diaphragm, typically having a thickness around a few tens of micrometers. The diaphragm can have a length of several hundreds of micrometers. A second silicon wafer is then bonded to the device wafer in a vacuum to form a reference vacuum cavity, thus completing the fabrication process. The second wafer can also be further etched through to form an inlet port, implementing a gauge pressure sensor [8.6]. The piezoresistive sensors are simple to fabricate and can be readily interfaced with electronic systems. However, the resistors are temperature dependent and consume DC power. Long-term characteristic drift and resistor thermal noise ultimately limit the sensor resolution.

Capacitive Sensor

Capacitive pressure sensors are attractive because they are virtually temperature independent and consume zero DC power. The devices do not exhibit initial turn-on drift and are stable over time. Furthermore, CMOS microelectronic circuits can be readily interfaced with the sensors to provide advanced signal conditioning and processing, thus improving overall system performance. An example of a capacitive pressure sensor is shown in Fig. 8.4. The device consists of an edge-clamped silicon diaphragm suspended over a vacuum cavity. The diaphragm can be square or circular with a typical thickness of a few micrometers and a length or radius of a few hundred micrometers, respectively. The vacuum

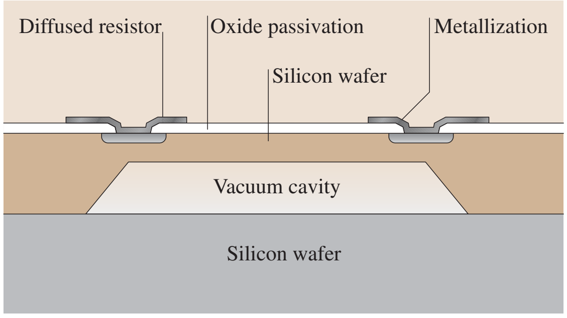

Fig. 8.3 Cross-sectional schematic of a piezoresistive pressure sensor

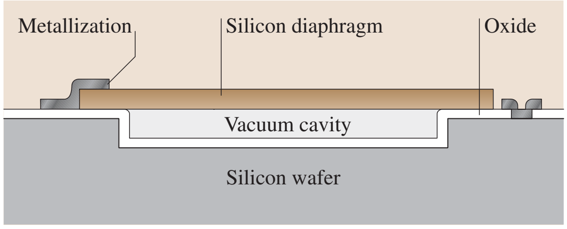

Fig. 8.4 Cross-sectional schematic of a capacitive pressure sensor

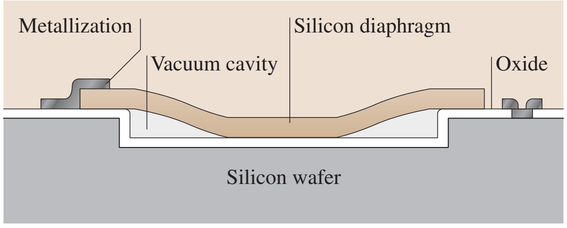

Fig. 8.5 Cross-sectional schematic of a touch-mode capacitive pressure sensor

cavity typically has a depth of a few micrometers. The diaphragm and substrate form a pressure dependent air-gap variable capacitor. An increased external pressure causes the diaphragm to deflect towards the substrate, resulting in an increase in the capacitance value. A simplified fabrication process can be outlined as follows. A silicon wafer is first patterned and etched to form the cavity. The wafer is then oxidized and bonded to a second silicon wafer with a heavily doped boron layer, which defines the diaphragm thickness, at the surface. The bonding process can be performed in a vacuum to realize the vacuum cavity. If the vacuum bonding is not performed at this stage, a low-pressure sealing process can be used to form the vacuum cavity after patterning the sensor diaphragm, provided that sealing channels are available. The silicon substrate above the boron layer is then removed through a wet etching process, followed by patterning to form the sensor diaphragm, which serves as the device top electrode. Contact pads are formed by metallization and patterning. This type of pressure sensor exhibits a nonlinear characteristic and a limited dynamic range. These phenomena, however, can be alleviated by applying an electrostatic-force-balanced feedback architecture. A common practice is to introduce another electrode above the sensing diaphragm through wafer bonding [8.7], forming two capacitors in series, with the diaphragm being the middle electrode. The capacitors are interfaced with electronic circuits, that convert the sensor capacitance value to an output voltage corresponding to the diaphragm position. This voltage is further processed to generate a feedback signal to the top electrode, thereby introducing an electrostatic pull up force to maintain the deflectable diaphragm at its nominal position. This negative feedback loop would substantially minimize the device nonlinearity and also extend the sensor dynamic range.

A capacitive pressure sensor achieving an inherent linear characteristic response and a wide dynamic range can be implemented by employing a touch mode architecture [8.8]. Figure 8.5 shows the cross-sectional view of a touch-mode pressure sensor. The device consists of an edge-clamped silicon diaphragm suspended over a vacuum cavity. The diaphragm deflects under an increasing external pressure and touches the substrate, causing a linear increase in the sensor capacitance value beyond the touch point pressure. Figure 8.6 shows a typical device characteristic curve. The touch point pressure can be designed by engineering the sensor geometric parameters, such as diaphragm size, thickness, cavity depth, etc., for various application requirements. The device can be fabricated using a process flow similar to the flow outlined for the basic capacitive pressure sensor. Figure 8.7 presents a photo of a fabricated touch-mode sensor employing a circular diaphragm with a diameter of 800 μm and a thickness of 5 μm suspended over a 2.5 μm vacuum cavity. The device achieves a touch point pressure of 8 psi and exhibits a linear capacitance range of 33 pF at 10 psi to 40 pF at 32 psi (absolute pressures).

The above processes use bulk silicon materials for machining and are usually referred to as bulk

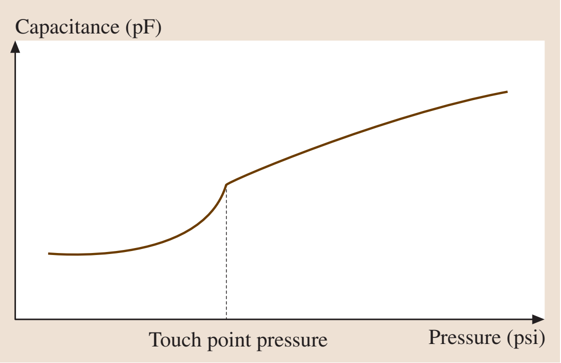

Fig. 8.6 Characteristic response of a touch-mode capacitive pressure

Springer Handbook of Nanotechnology
B. Bhushan • © Springer 2004

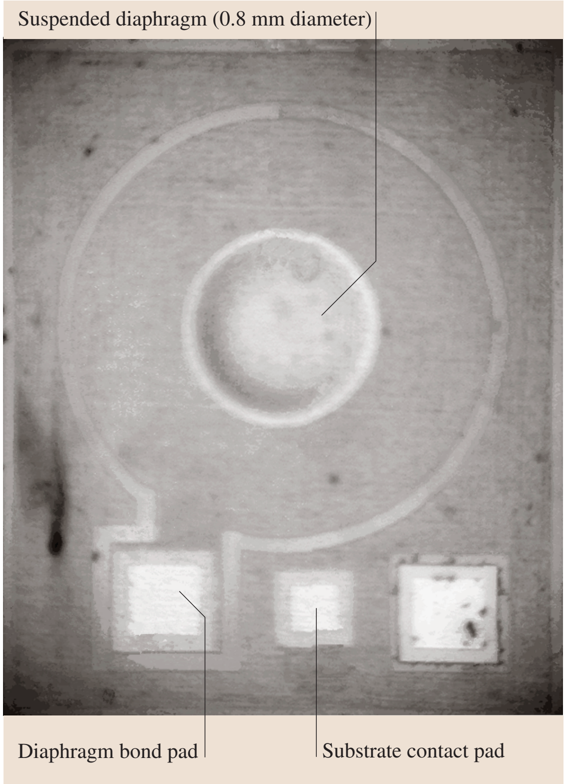

Fig. 8.7 Photo of a touch-mode capacitive pressure sensor [8.12]

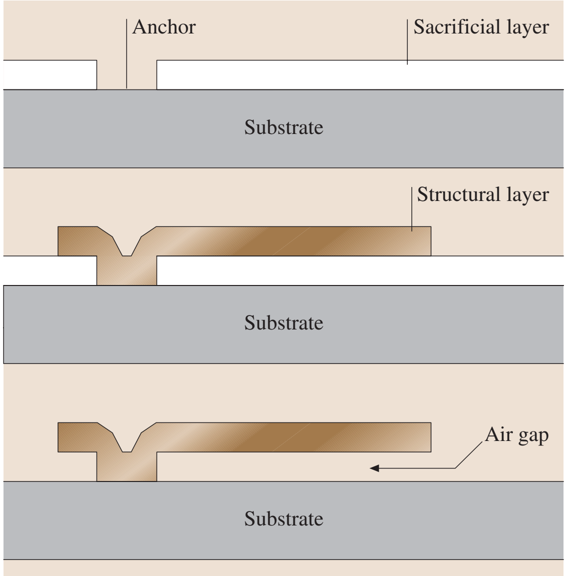

Fig. 8.8 Simplified fabrication sequence of surface micromachining technology

micromachining. The same devices can also be fabricated using so-called surface micromachining. Surface micromachining technology is attractive for integrating MEMS sensors with on-chip electronic circuits. As a result, advanced signal processing capabilities, such as data conversion, offset and noise cancellation, digital calibration, temperature compensation, etc., can be implemented adjacent to microsensors on the same substrate, providing a complete, high performance microsystem solution. The single chip approach also eliminates external wiring, which is critical for minimizing noise pickup and enhancing system performance. Surface micromachining, simply stated, is a method of fabricating MEMS through depositing, patterning, and etching a sequence of thin films with thicknesses on the order of a micrometer. Figure 8.8 illustrates a typical surface micromachining process flow [8.9]. The process starts by depositing a layer of sacrificial material, such as silicon dioxide, over a wafer, followed by anchor formation. A structural layer, typically a polysilicon film, is deposited and patterned. The underlying sacrificial layer is then removed to release the suspended microstructure and complete the fabrication sequence. The processing materials and steps are compatible with the standard integrated circuit process and, thus, can be readily incorporated as an add-on module to an IC process [8.9–11]. A similar surface micromachining technology has been developed to produce monolithic pressure sensor systems [8.12]. Figure 8.9 shows an SEM micrograph of an array of MEMS capacitive pressure sensors fabricated with BiCMOS electronics on the same substrate. Each sensor consists of a 0.8 μm-thick circular polysilicon membrane with a diameter on the order of 20 μm suspended over a 0.3 μm-deep vacuum cavity. The devices operate using the same principle as the sensor shown in Fig. 8.4. A close view of the sensor cross section is shown in Fig. 8.10, which shows the suspended membrane and underneath air gap. These sensors have demonstrated operations in pressure ranges up to 400 bar with an accuracy of 1.5%.

8.1.2 Inertial Sensor

Micromachined inertial sensors consist of accelerometers and gyroscopes. These devices are one of the important types of silicon-based MEMS sensors that have been successfully commercialized. MEMS ac-

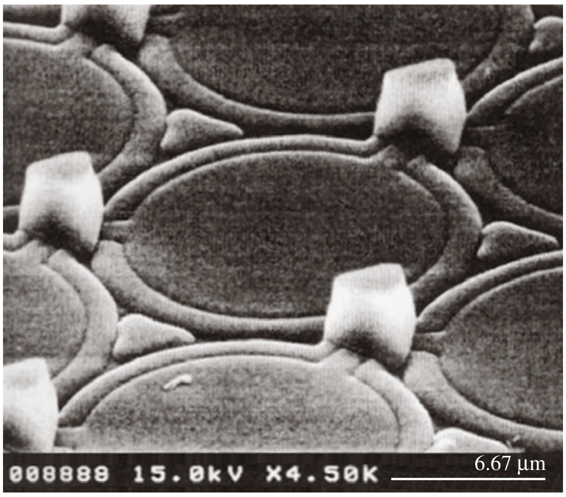

Fig. 8.9 SEM micrograph of polysilicon surface-micromachined capacitive pressure sensors [8.12]

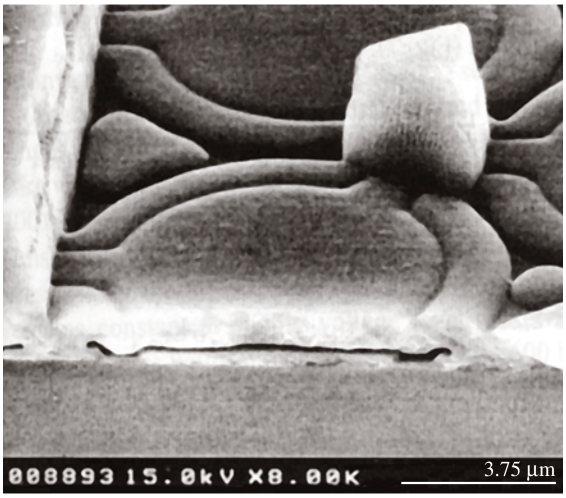

Fig. 8.10 SEM micrograph of a close-up view of a polysilicon surface-micromachined capacitive pressure sensor [8.12]

celerometers alone have the second largest sales volume after pressure sensors. Gyroscopes are expected to reach a comparable sales volume in the foreseeable future. Accelerometers have been used in a wide range of applications, including automotive application for safety systems, active suspension and stability control, biomedical application for activity monitoring, and numerous consumer products, such as head-mount displays, camcorders, three-dimensional mouse, etc. High sensitivity accelerometers are crucial for implementing self-contained navigation and guidance systems. A gyroscope is another type of inertial sensor that measures rate or angle of rotation. The devices can be used along with accelerometers to provide heading information in an inertial navigation system. Gyroscopes are also useful in such applications as automotive ride stabilization and rollover detection, camcorder stabilization, virtual reality, etc. Inertial sensors fabricated by micromachining technology can achieve reduced size, weight, and cost, all which are critical for consumer applications. More importantly, these sensors can be integrated with microelectronic circuits to achieve a functional microsystem with high performance.

Accelerometer

An accelerometer generally consists of a proof mass suspended by compliant mechanical suspensions anchored to a fixed frame. An external acceleration displaces the support frame relative to the proof mass. The displacement can result in an internal stress change in the suspension, which can be detected by piezoresistive sensors as a measure of the external acceleration. The displacement can also be detected as a capacitance change in capacitive accelerometers. Capacitive sensors are attractive for various applications because they exhibit high sensitivity and low temperature dependence, turn-on drift, power dissipation, and noise. The sensors can also be readily integrated with CMOS electronics to perform advanced signal processing for high system performance. Capacitive accelerometers may be divided into two categories: vertical and lateral sensors. Figure 8.11 shows sensor structures for the two versions. In a vertical device, the proof mass is suspended above the substrate electrode by a small gap typically on the order of a micrometer, forming a parallel-plate sense capacitance. The proof mass moves in the direction perpendicular to the substrate (z-axis) upon a vertical input acceleration, thus changing the gap and the capacitance value. The lateral accelerometer consists of a number of movable fingers attached to the proof mass, forming a sense capacitance with an array of fixed parallel fingers. The sensor proof mass moves in a plane parallel to the substrate when subjected to a lateral input acceleration, thus changing the overlap area of these fingers and, hence, the capacitance value. Figure 8.12 shows an SEM top view of a surface micromachined polysilicon z-axis accelerometer [8.13]. The device consists of a 400 µm × 400 µm proof mass with a thickness of 2 µm suspended above the substrate electrode by four folded beam suspensions with an air gap around 2 µm, thus achieving a sense capacitance of approximately 500 fF.

The visible holes are used to ensure complete removal of the sacrificial oxide underneath the proof mass at the end of the fabrication process. The sensor can be interfaced with a microelectronic charge amplifier, converting the capacitance value to an output voltage for further signal processing and analysis. Force feedback architecture can be applied to stabilize the proof mass position. The combs around the periphery of the proof mass can exert an electrostatic levitation force on the proof mass to achieve the position control, thus improving the system frequency response and linearity performance [8.13].

Surface micromachined accelerometers typically suffer from severe mechanical thermal vibration, commonly referred to as Brownian motion [8.14], due to the small proof mass, resulting in a high mechanical noise floor that ultimately limits the sensor resolution. Vacuum packaging can be employed to minimize this adverse effect, but with a penalty of increasing system complexity and cost. Accelerometers using large proof masses fabricated by bulk micromachining, or a combination of surface and bulk micromachining techniques are attractive for circumventing this problem. Figure 8.13 shows an SEM micrograph of an all-silicon *z*-axis accelerometer fabricated through a single silicon wafer by using a combined surface and bulk micromachining process to obtain a large proof mass with dimensions of approximately 2 mm × 1 mm × 450 µm [8.15]. The large mass suppresses the Brownian motion effect, achieving a high performance with a resolution on the order of several µg.

A surface micromachined lateral accelerometer developed by Analog Devices Inc. is shown in Fig. 8.14. The sensor consists of a center proof mass supported by folded beam suspensions with arrays of attached movable fingers, forming a sense capacitance with the fixed parallel fingers. The device is fabricated using a 6 µm-thick polysilicon structural layer with a small air gap on the order of a micrometer to increase the sensor capacitance value, thus improving the device resolution. Figure 8.15 shows a close-up view of the finger structure for a typical lateral accelerometer. Each movable finger forms differential capacitances with two adjacent fixed fingers. This sensing capacitance configuration is attractive for interfacing with differential electronic detection circuits to suppress common-mode noise and other undesirable signal coupling. Monolithic accelerometers with a three-axis sensing capability integrated with on-chip electronic detection circuits have been realized using surfacing micromachining and CMOS microelectronics fabrication technologies [8.16]. Figure 8.16 shows

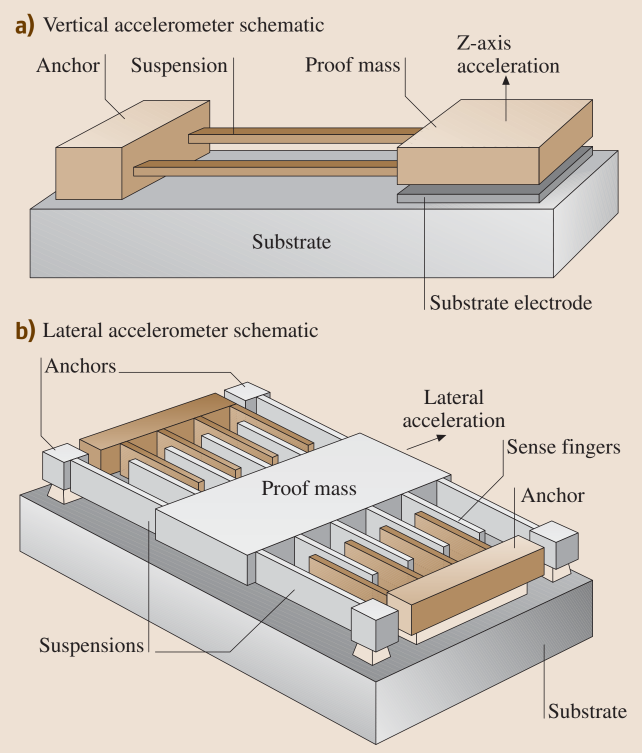

Fig. 8.11a,b Schematics of vertical **(a)** and lateral **(b)** accelerometers

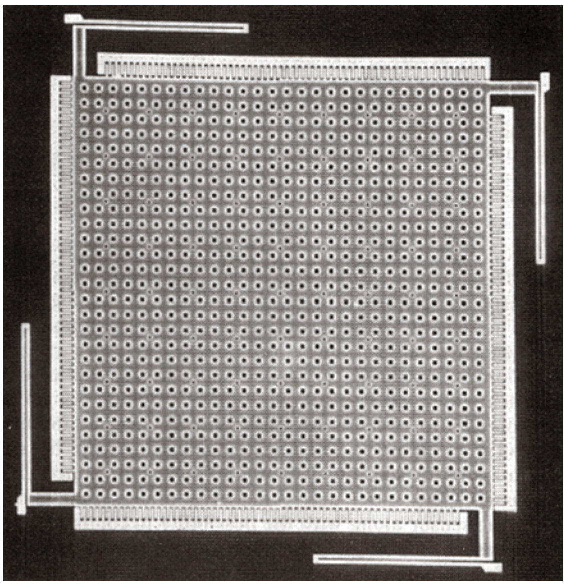

Fig. 8.12 SEM micrograph of a polysilicon surface-micromachined *z*-axis accelerometer [8.13]

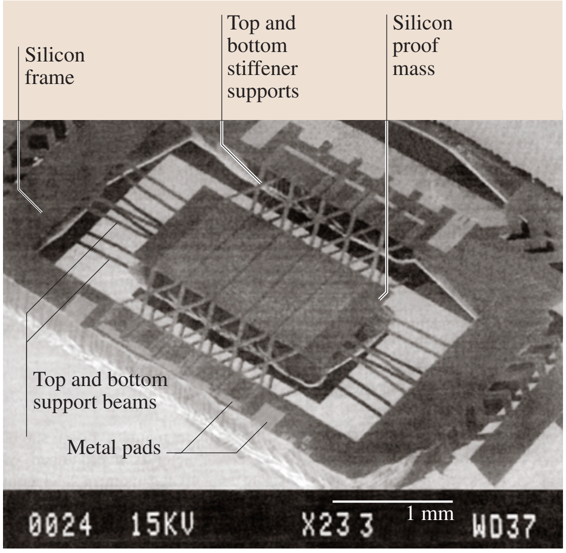

Fig. 8.13 SEM micrograph of a MEMS z-axis accelerometer fabricated using a combined surface and bulk micromachining technology [8.15]

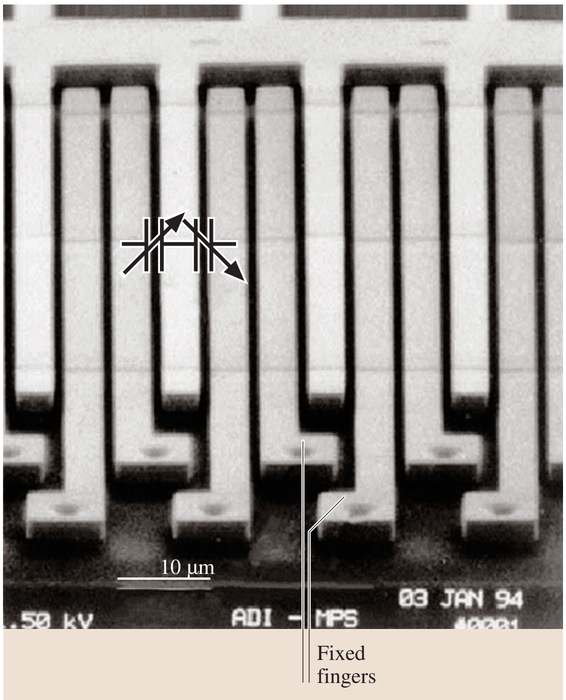

Fig. 8.15 SEM micrograph of a capacitive sensing finger structure

a photo of one of these microsystem chips, which has an area of 4 mm × 4 mm. One vertical accelerometer and two lateral accelerometers are placed at the chip center with corresponding detection electronics along the periphery. A z-axis reference device, which is not movable, is used with the vertical sensor for electronic interfacing. The prototype system achieves a sensing resolution on the order of 1 mG with a 100 Hz bandwidth along each axis. The level of performance is adequate for automobile safety activation systems, vehicle stability, active suspension control, and various consumer products.

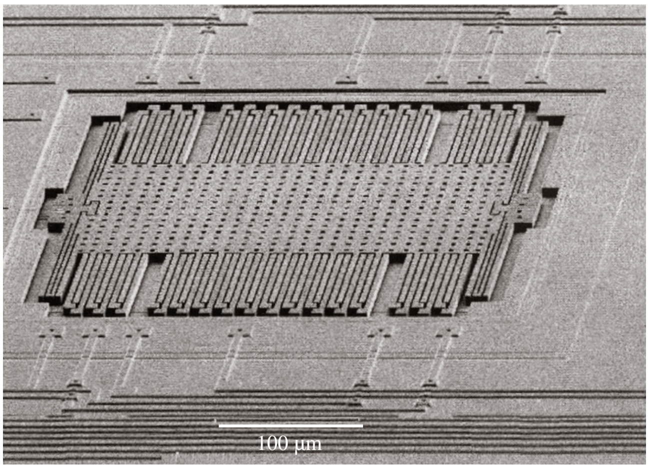

Fig. 8.14 SEM micrograph of a polysilicon surface-micromachined lateral accelerometer (courtesy of Analog Devices Inc.)

Gyroscope

Most micromachined gyroscopes employ vibrating mechanical elements to sense rotations. The sensors rely on the energy transfer between two vibration modes of a structure caused by Coriolis acceleration. Figure 8.17 presents a schematic of a z-axis vibratory rate gyroscope. The device consists of an oscillating mass electrostatically driven into resonance along the drive-mode axis using comb fingers. An angular rotation along the vertical axis (z-axis) introduces a Coriolis acceleration, which results in a structure deflection along the sense-mode axis, shown in the figure. The deflection changes the differential sense capacitance value, which can be detected as a measure of input angular rotation. A z-axis vibratory rate gyroscope operating on this principle is fabricated using surface micromachining technology and integrated together with electronic detection circuits, as illustrated in Fig. 8.18 [8.17]. The micromachined sensor is fabricated using polysilicon structural material with a thickness around 2 µm and occupies an area of 1 mm × 1 mm. The sensor achieves a resolution of approximately $1°/\mathrm{s}/\sqrt{\mathrm{Hz}}$ under a vacuum pressure around 50 mtorr.

A dual-axis gyroscope based on a rotational disk at its resonance can be used to sense angular rotation along two lateral axes (x-axis and y-axis). Figure 8.19 shows a device schematic demonstrating the operating principle. A rotor disk supported by four mechanical suspensions can be driven into angular resonance along the z-axis. An input angular rotation along the x-axis will generate a Coriolis acceleration, causing the disk to rotate along the y-axis and vice versa. This Coriolis acceleration-induced rotation will change the sensor capacitance values between the disk and the different sensing electrodes underneath. The capacitance change can be detected and processed by electronic interface circuits. Angular rotations along the two lateral axes can be measured simultaneously using this device architecture. Figure 8.20 shows a photo of a dual-axis gyroscope fabricated using a 2 µm-thick polysilicon surface micromachining technology [8.18]. As shown in the figure, curved electrostatic drive combs are positioned along the circumference of the rotor dick to drive it into resonance along the vertical axis. The gyroscope exhibits a low random walk of $1°/\sqrt{\mathrm{h}}$ under a vacuum pressure around 60 mtorr. With accelerometers and gyroscopes, each capable of three-axis sensing, a micromachine-based inertial measurement system providing a six-degree-of-freedom sensing capability can be realized. Figure 8.21 presents a photo of such a system containing a dual-axis gyroscope, a z-axis gyroscope, and a three-axis accelerometer chip integrated with microelectronic circuitry. Due to the precision in device layout and fabrication, the system can measure angular rotation and acceleration without the need to align individual sensors.

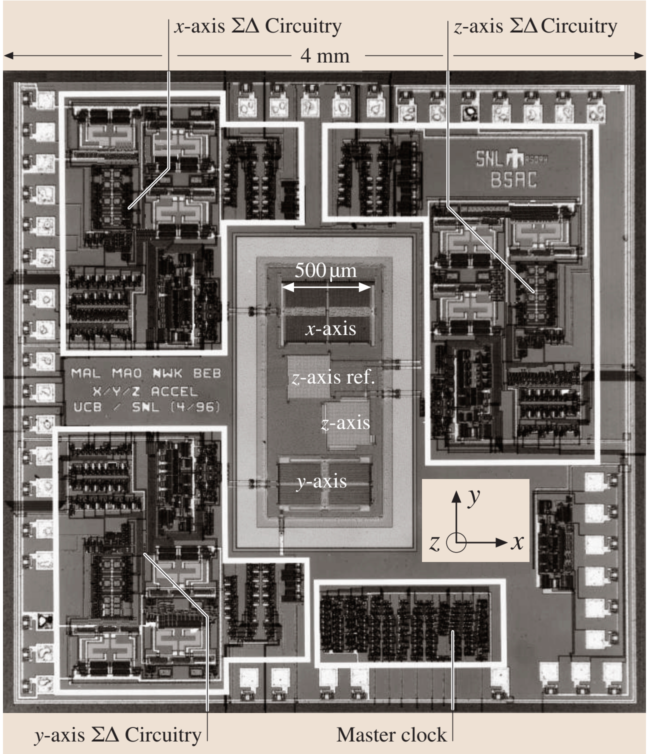

Fig. 8.16 Photo of a monolithic three-axis polysilicon surface-micromachined accelerometer with integrated interface and control electronics [8.16]

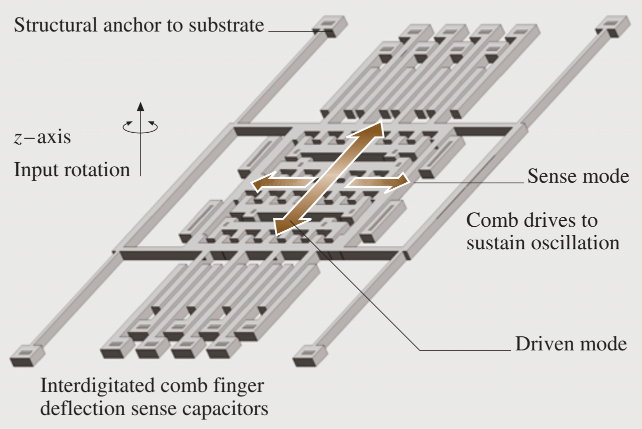
Structural anchor to substrate
z−axis
Input rotation
Sense mode
Comb drives to
sustain oscillation
Driven mode
Interdigitated comb finger
deflection sense capacitors

Fig. 8.17 Schematic of a vibratory rate gyroscope

8.1.3 Optical MEMS

Surface micromachining has served as a key enabling technology to realize microeletromechanical optical devices for various applications ranging from sophisticated visual information displays and fiber-optic telecommunication to bar-code reading. Most of the existing optical systems are implemented using conventional optical components, which suffer from bulky size, high cost, large power consumption, and poor efficiency and reliability issues. MEMS technology is promising for producing miniaturized, reliable, inexpensive

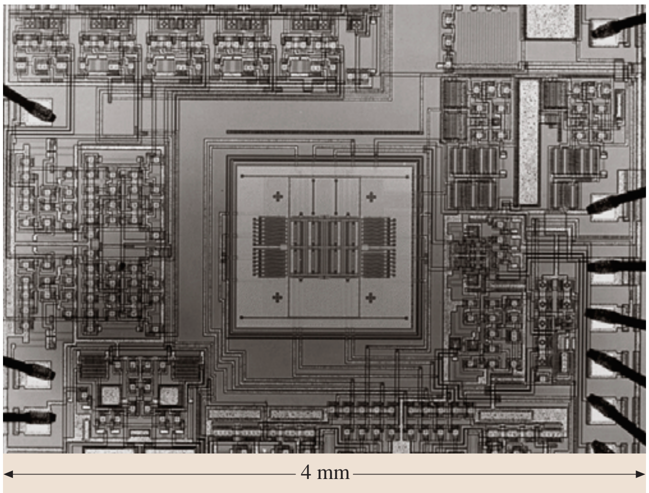

Fig. 8.18 Photo of a monolithic polysilicon surface-micromachined *z*-axis vibratory gyroscope with integrated interface and control electronics [8.17]

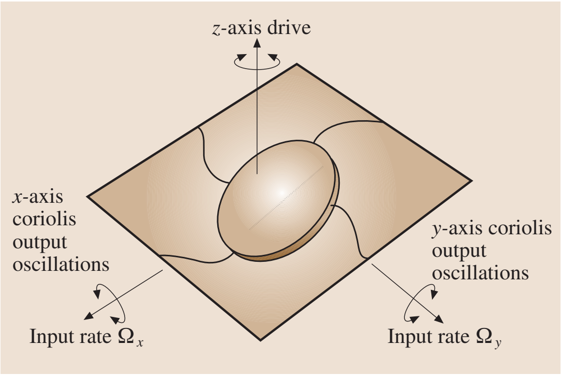
z-axis drive
x-axis coriolis output oscillations
y-axis coriolis output oscillations
Input rate Ω_x
Input rate Ω_y

Fig. 8.19 Schematic of a dual-axis gyroscope

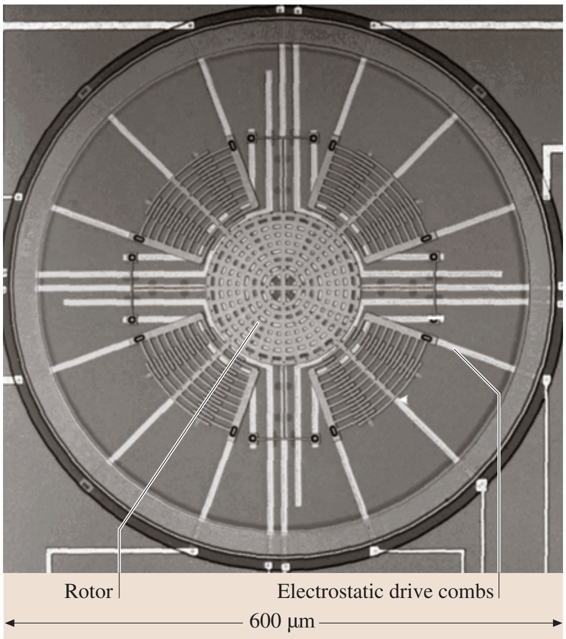
Rotor
Electrostatic drive combs
600 µm

Fig. 8.20 Photo of a polysilicon surface-micromachined dual-axis gyroscope [8.18]

Fig. 8.21 ▾ Photo of a surface-micromachined inertial measurement system with a six-degree sensing capability

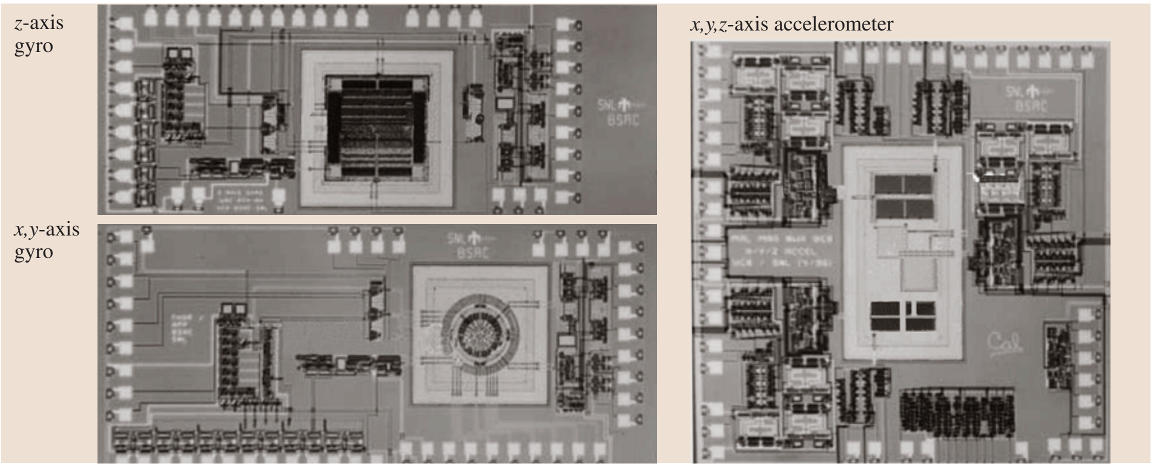
z-axis gyro
x,*y*-axis gyro
x,*y*,*z*-axis accelerometer

optical components to revolutionize conventional optical systems [8.19]. In this section of the chapter, a few selected MEMS optical devices will be presented to illustrate their impact on the fields of visual display, precision optical platform, and data switching for optical communication.

Visual Display

An early MEMS device successfully used for various display applications is the Texas Instruments Digital Micromirror Device (DMD). The DMD technology can achieve higher performance in terms of resolution, brightness, contrast ratio, and convergence than the conventional cathode ray tube and is critical for digital high definition television applications. A DMD consists of a large array of small mirrors with a typical area of 16 μm × 16 μm, as illustrated in Fig. 8.22. A probe tip is shown in the figure for a size comparison. Figure 8.23 shows an SEM micrograph of a close-up view of a DMD pixel array [8.20]. Each mirror is capable of rotating by ±10 degrees, corresponding to either the "on" or "off" position, due to an electrostatic actuation force. Light reflected from any on-mirrors passes through a projection lens and creates images on a large screen. Light from the remaining off-mirrors is reflected away from the projection lens to an absorber. The amount of time during each video frame that a mirror remains in the on-state determines shades of gray, from black for zero on-time to white for a hundred percent on-time. Color can be added by a color wheel or a three-DMD chip setup. The three-DMD chips are used for projecting red, green, and blue colors. Each DMD pixel consists of a mirror connected by a mirror support post to an underlying yoke. The yoke, in turn, is connected by torsion hinges to the support posts, as shown in Fig. 8.24 [8.21]. The support post and hinges are hidden under the mirror to avoid light diffraction and, thus, improve contrast ratio and optical efficiency. There are two gaps on the order of a micrometer, one between the mirror, the underlying hinges and the address electrodes, and a second between the coplanar address electrodes, the hinges and an underlying metal layer from the CMOS static random access memory (SRAM) structure. The yoke is tilted over the second gap by an electrostatic actuation force, thereby rotating the mirror plate. The SRAM determines which angle the mirror needs to be tilted by applying proper actuation voltages to the mirror and address electrodes. The DMD is fabricated using an aluminum-based surface micromachining technology. Three layers of aluminum thin film are deposited and patterned to form the mirror and its suspension system. Polymer material is used as the sacrificial layer and is removed by a plasma etch at the end of the process to freely release the micromirror structure. The micromachining process is compatible with standard CMOS fabrication, allowing the DMD to be monolithically integrated with a mature CMOS address circuit technology, thus achieving high yield and low cost. Figure 8.25 shows an SEM micrograph of a fabricated DMD pixel revealing its cross section after an ion milling. A close-up view of the yoke and hinge support under the mirror is shown in Fig. 8.26.

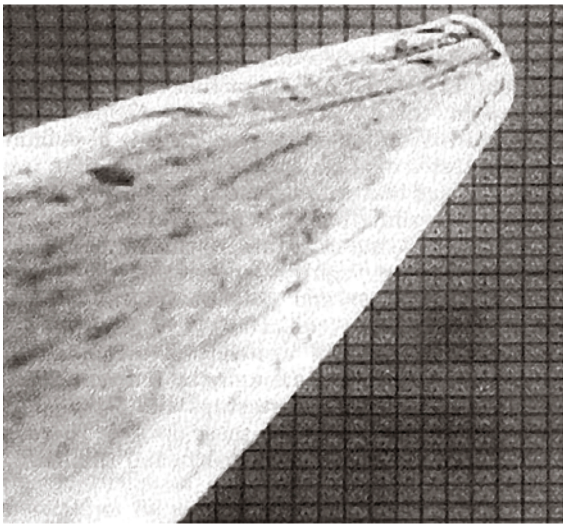

Fig. 8.22 Photo of a digital micromirror device (DMD) array (courtesy of Texas Instruments)

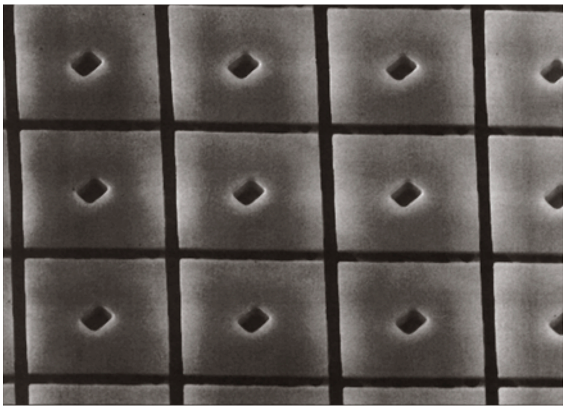

Fig. 8.23 SEM micrograph of a close-up view of a DMD pixel array [8.20]

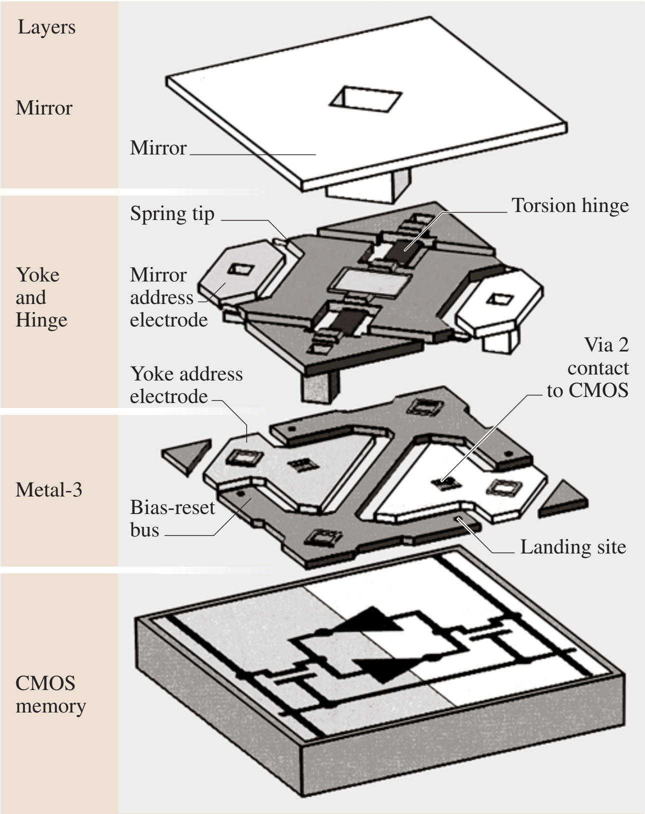

Fig. 8.24 Detailed structure layout of a DMD pixel [8.21]

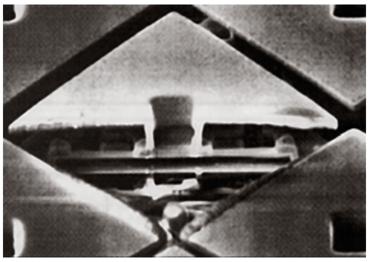

Fig. 8.25 SEM micrograph of a DMD pixel after removing half of the mirror plate using ion milling (courtesy of Texas Instruments)

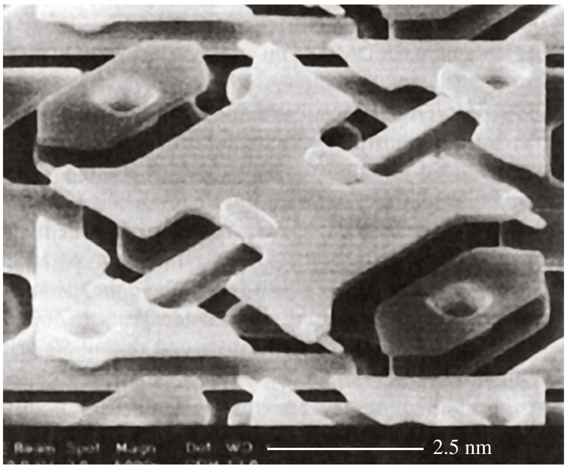

Fig. 8.26 SEM micrograph of a close view of a DMD yoke and hinges [8.21]

Precision Optical Platform

The growing optical communication and measurement industry requires low cost, high performance optoelectronic modules, such as laser-to-fiber couplers, scanners, interferometers, etc. A precision alignment and the ability to actuate optical components, such as mirrors, gratings, and lenses, with sufficient accuracy are critical for high performance optical applications. Conventional hybrid optical integration approaches, such as the silicon optical bench, suffers from a limited alignment tolerance of ±1 μm and also lacks component actuation capability [8.22, 23]. As a result, only simple optical systems can be constructed with no more than a few components, thus severely limiting the performance. Micromachining, however, provides a critical enabling technology, allowing movable optical components to be fabricated on a silicon substrate. High-precision component movement can be achieved through electrostatic actuation. By combining micromachined movable optical components with lasers, lenses, and fibers on the same substrate, an on-chip, complex, self-aligning optical system can be realized. Figure 8.27a shows an SEM micrograph of a surface micromachined, electrostatically actuated microreflector for laser-to-fiber coupling and external cavity laser applications [8.24]. The device consists of a polysilicon mirror plate hinged to a support beam. The mirror and the support, in turn, are hinged to a vibromotor-actuator slider. The micro-hinge technology [8.25] allows the joints to rotate out of the substrate plane to achieve large aspect ratios. Common-mode actuation of the sliders results in a translational motion, while differential slider motion produces an out-of-plane mirror rotation. These motions permit the microreflector to redirect an optical beam in a desirable location. Each of the two slides is actuated with an integrated microvibromotor, shown in Fig. 8.27b. The vibromotor consists of four electrostatic comb resonators with attached impact arms driving a slider through oblique

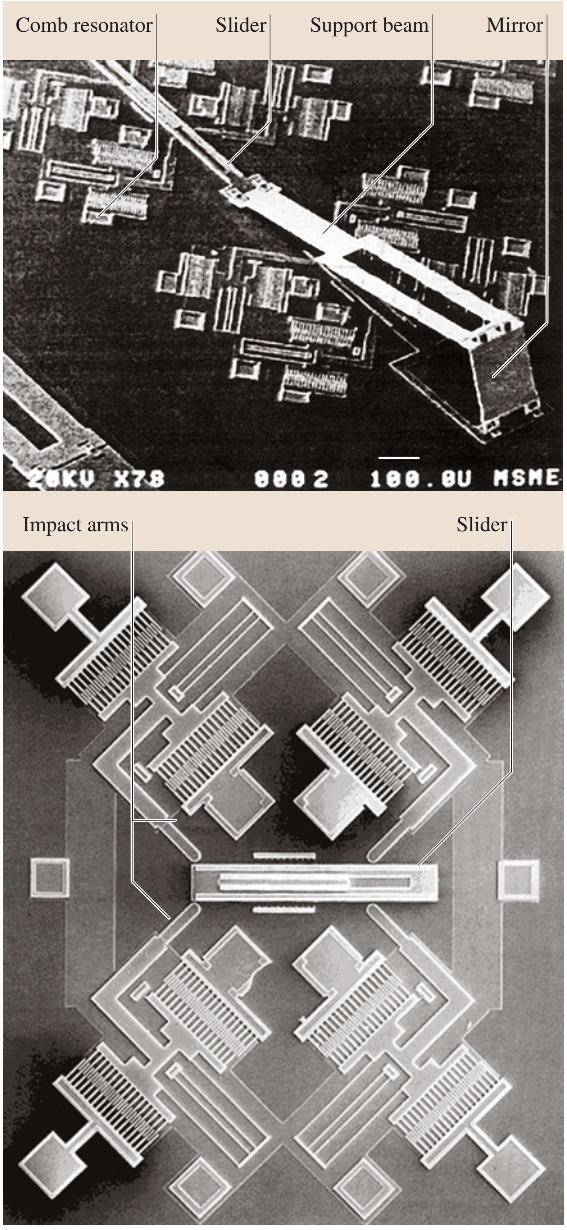

Fig. 8.27 **(a)** SEM micrograph of a surface-micromachined, electrostatically actuated microreflector; **(b)** SEM micrograph of a surface-micromachined vibromotor [8.24]

impact. The two opposing impacters are used for each travel direction to balance the forces. The resonator is a capacitively driven mass anchored to the substrate through a folded beam flexure. The flexure compliance determines the resonant frequency and travel range of the resonator. When the comb structures are driven at their resonant frequency (around 8 KHz), the slider exhibits a maximum velocity of over 1 mm/s. Characterization of the vibromotor also shows that a slider step resolution of less than 0.3 μm can be achieved [8.26], making it attractive for precision alignment of various optical components. The prototype microreflector can obtain an angular travel range over 90 degrees and a translational travel range of 60 μm. By using this device, beam steering, fiber coupling, and optical scanning have been demonstrated.

Optical Data Switching

High-speed communication infrastructures are highly desirable for transferring and processing real-time voice and video information. Optical fiber communication technology has been identified as the critical backbone to support such systems. A high performance optical data switching network, which routes various optical signals from sources to destinations, is one of the key building blocks for system implementation. At present, optical signal switching is performed by using hybrid optical-electronic-optical (O-E-O) switches. These devices first convert incoming light from input fibers to electrical signals and then route the electrical signals to the proper output ports after signal analyses. At the output ports, the electrical signals are converted back to streams of photons or optical signals for further transmission over the fibers. The O-E-O switches are expensive to build, integrate, and maintain. Furthermore, they consume a substantial amount of power and introduce additional latency. It is, therefore, highly desirable to develop an all-optical switching network in which optical signals can be routed without intermediate conversion into electrical form, thus minimizing power dissipation and system delay. While a number of approaches are being considered for building all-optical switches, MEMS technology is attractive because it can provide arrays of tiny movable mirrors that can redirect incoming beams from input fibers to corresponding output fibers. As described in the previous sections, these micromirrors can be batch fabricated using silicon micromachining technologies, achieving an integrated solution with the potential for low cost. A significant reduction in power dissipation is also expected.

Figure 8.28 shows an architecture of a two-dimensional micromirror array forming a switching matrix with rows of input fibers and columns of output fibers (or vice versa). An optical beam from an input fiber can be directed to an output fiber by activating the cor-

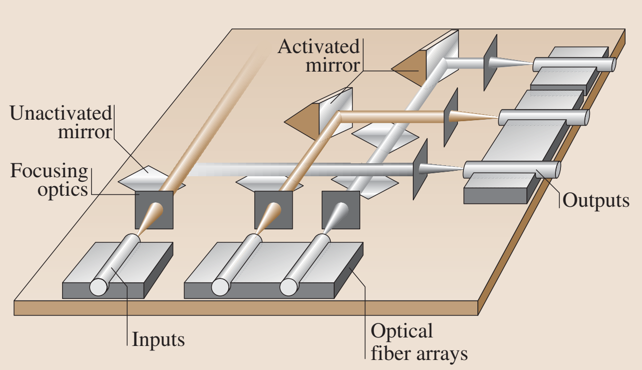

Fig. 8.28 Schematic of a two-dimensional micromirror-based fiber optic switching matrix

responding reflecting micromirror. Switches with eight inputs and eight outputs can be readily implemented using this technique, which can be further extended to a 64×64 matrix. The micromirrors are moved between two fixed stops by digital control, eliminating the need for precision motion control. Figure 8.29 presents an SEM micrograph of a simple 2×2 MEMS fiber optic switching network prototype for an illustration purpose [8.27]. The network includes a mirror chip passively integrated with a silicon submount, which contains optical fibers and ball lenses. The mirror chip consists of four surface-micromachined vertical torsion mirrors. The four mirrors are arranged so that in the "reflection" mode, the input beams are reflected by two 45-degree vertical torsion mirrors and coupled with the output fibers located on the same side of the chip. In the "transmission" mode, the vertical torsion mirrors are rotated out of the optical paths, allowing the input beams to be coupled with the opposing output fibers. Figure 8.30 shows an SEM micrograph of a polysilicon vertical torsion mirror. The device consists of a mirror plate attached to a vertical supporting frame by torsion beams and a vertical back electrode plate. The mirror plate is approximately 200 µm wide, 160 µm long, and 1.5 µm thick. The mirror surface is coated with a thin layer of gold to improve the optical reflectivity. The back plate is used to electrostatically actuate the mirror plate, so that the mirror can be rotated out of the optical path in the "transmission" mode. Surface micromachining with micro-hinge technology is used to realize the overall structure. The back electrode plate is integrated with a scratch drive actuator array [8.28] for self-assembly. The self-assembly approach is critical when multiple vertical torsion mirrors are used to implement more advanced functions.

A more sophisticated optical switching network with a large scaling potential can be implemented by using a three-dimensional (3-D) switching architecture, as shown in Fig. 8.31. The network consists of arrays of two-axis mirrors to steer optical beams from input fibers to output fibers. A precision analog closed-loop mirror position control is required to accurately direct a beam along two angles, so that one input fiber can be optically

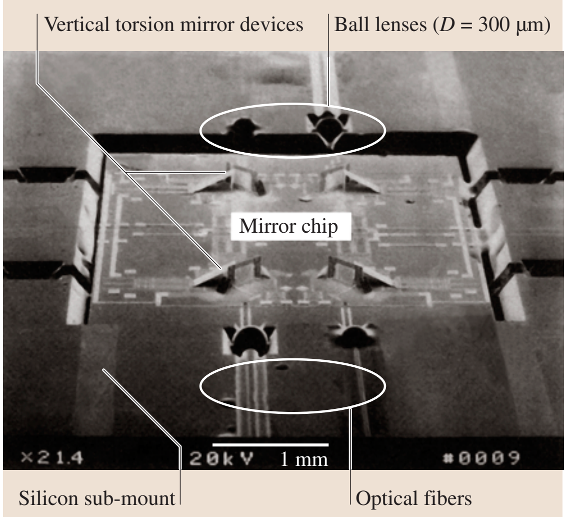

Fig. 8.29 SEM micrograph of a 2×2 MEMS fiber optic switching network [8.27]

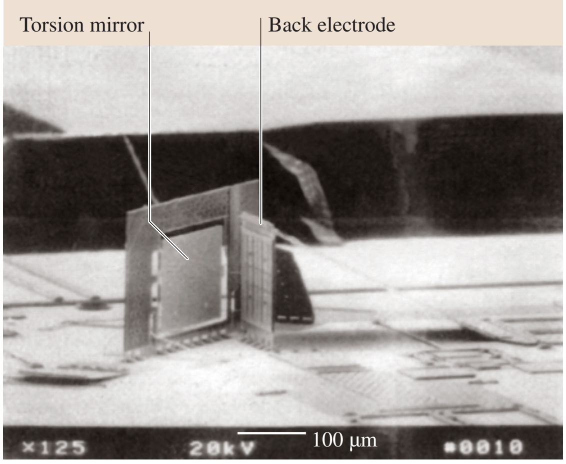

Fig. 8.30 SEM micrograph of a polysilicon surface-micromachined vertical torsion mirror [8.27]

connected to any output fiber. The optical length depends little on which sets of fibers are connected, thus achieving a more uniform switching characteristic, which is critical for implementing large scale networks. Two-axis mirrors are the crucial components for implementing the 3-D architecture. Figure 8.32 shows an SEM micrograph of a surface micromachined two-axis beam-steering mirror positioned by using self-assembly technique [8.29]. The self-assembly is accomplished during the final release step of the mirror processing sequence. Mechanical energy is stored in a special high-stress layer during the deposition, which is put on top of the four assembly arms. Immediately after the assembly arms are released, the tensile stress in this layer causes the arms to bend upward, pushing the mirror frame and lifting it above the silicon substrate. All mirrors used in the switching network can be fabricated simultaneously without any human intervention or external power supply.

8.1.4 RF MEMS

The increasing demand for wireless communication applications, such as cellular and cordless telephony, wireless data networks, two-way paging, global positioning systems, etc., feeds a growing interest in building miniaturized wireless transceivers with multi-standard capabilities. Such transceivers will greatly enhance the convenience and accessibility of various wireless services independent of geographic location. Miniaturizing current single-standard transceivers through a high level of integration is a critical step toward building transceivers that are compatible with multiple standards. Highly integrated transceivers will also result in reduced package complexity, power consumption, and cost. At present, most radio transceivers rely on a large number of discrete frequency-selection components, such as radio frequency (RF) and intermediate frequency (IF) band-pass filters, RF voltage-controlled oscillators (VCOs), quartz crystal oscillators, solid-state switches, etc., to perform the necessary analog signal processing. Figure 8.33 shows a schematic of a super-heterodyne radio architecture, in which discrete components are shaded. These off-chip devices occupy the majority of the system area, thus severely hindering transceiver miniaturization. MEMS technology, however, offers a potential solution to integrate these discrete components onto silicon substrates with microelectronics, achieving a size reduction of a few orders of magnitude. It is, therefore, expected to become a technology that will ultimately enable the miniaturization of radio transceivers for future wireless communications.

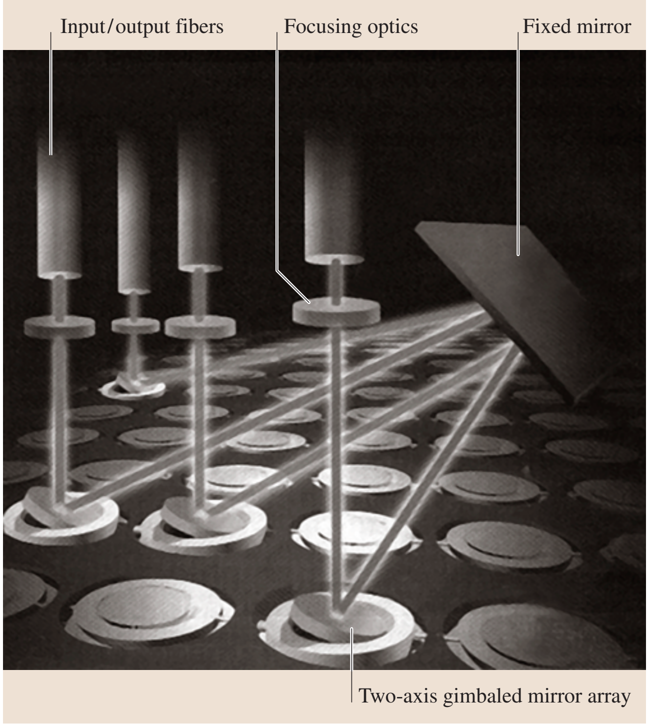

Fig. 8.31 Schematic of a three-dimensional micromirror-based fiber optic switching matrix

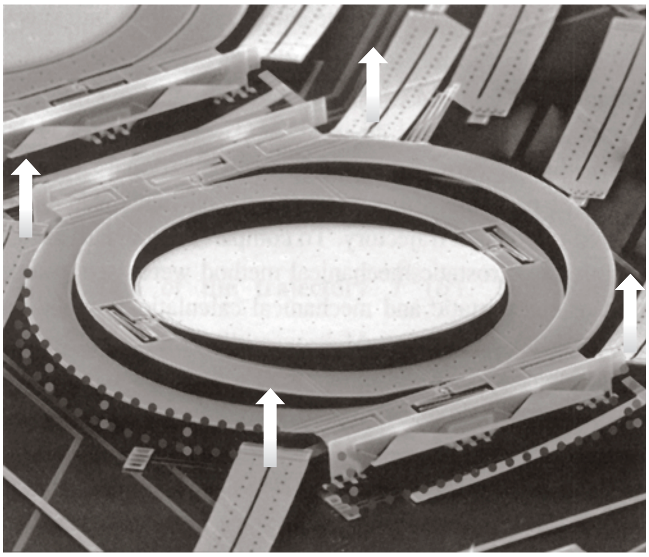

Fig. 8.32 SEM micrograph of a surface micromachined two-axis beam-steering micromirror positioned using self-assembly technique [8.29]

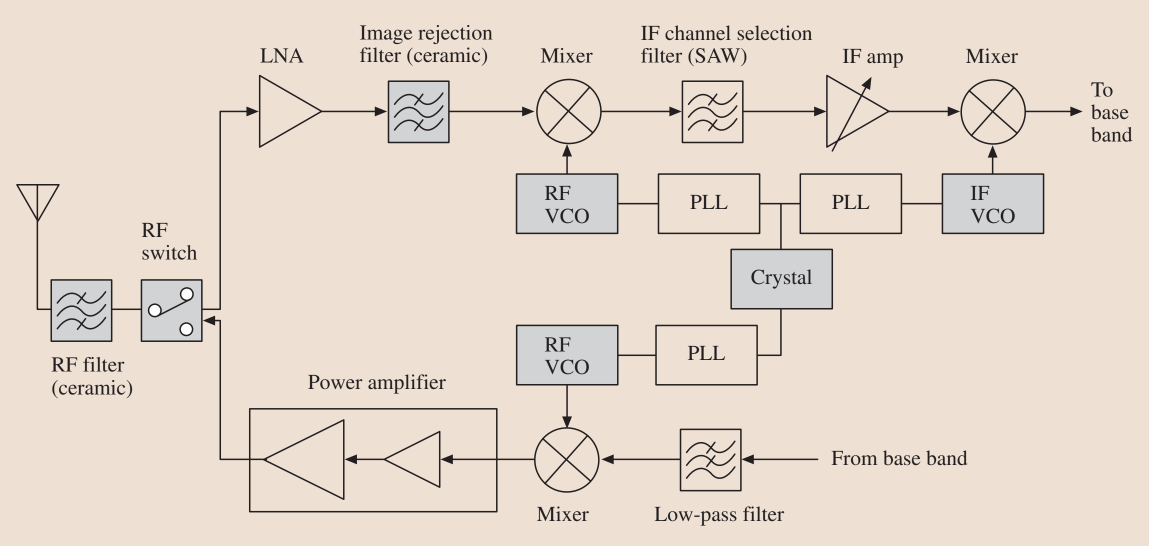
LNA
Image rejection filter (ceramic)
Mixer
IF channel selection filter (SAW)
IF amp
Mixer
To base band
RF VCO
PLL
PLL
IF VCO
Crystal
RF switch
RF filter (ceramic)
RF VCO
PLL
Power amplifier
From base band
Mixer
Low-pass filter

Fig. 8.33 Schematic of a super-heterodyne radio architecture

MEMS Variable Capacitors

Integrated high performance variable capacitors are critical for low noise VCOs, antenna tuning, tunable matching networks, etc. Capacitors with high quality factors (Q), large tuning range, and linear characteristics are crucial for achieving system performance requirements. On-chip silicon PN junction and MOS-based variable capacitors suffer from low quality factors (below 10 at 1 GHz), limited tuning range, and poor linearity, and are, therefore, inadequate for building high performance transceivers. MEMS technology has demonstrated monolithic variable capacitors, achieving stringent performance requirements. These devices typically rely on an electrostatic actuation method to vary the air gap between a set of parallel plates [8.30–33], or vary the capacitance area between a set of conductors [8.34], or mechanically displace a dielectric layer in an air-gap capacitor [8.35]. Improved tuning ranges have been achieved with various device configurations. Figure 8.34 shows SEM micrographs of an aluminum micromachined variable capacitor fabricated on a silicon substrate [8.30]. The device consists of a 200 μm × 200 μm aluminum plate with a thickness of

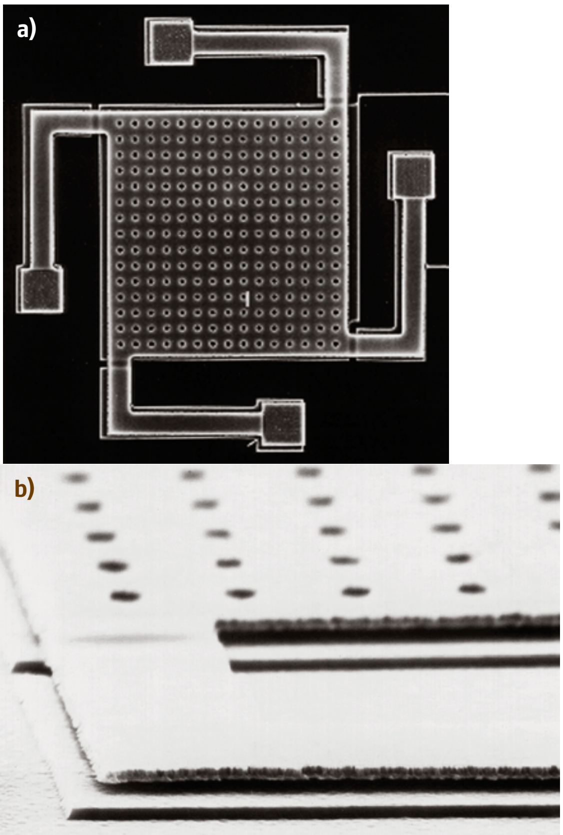

Fig. 8.34 **(a)** SEM micrograph of a top view of an aluminum surface-micromachined variable capacitor; **(b)** SEM micrograph of a cross-sectional view of the variable capacitor [8.30]

1 µm suspended above the bottom electrode by an air gap of 1.5 µm. Aluminum is selected as the structural material due to its low resistivity, which is critical for achieving a high quality factor at high frequencies. A DC voltage applied across the top and bottom electrodes introduces an electrostatic pull-down force, which pulls the top plate towards the bottom electrode, thus changing the device capacitance value. The capacitors are fabricated using aluminum-based surface micromachining technology. Sputtered aluminum is used for building the capacitor top and bottom electrodes. Photoresist serves as the sacrificial layer, which is then removed through an oxygen-based plasma dry etch to release the microstructure. The processing technology requires a low thermal budget, allowing the variable capacitors to be fabricated on top of wafers with completed electronic circuits without degrading the performance of active devices. Figure 8.35 presents an SEM micrograph of four MEMS tunable capacitors connected in parallel. This device achieves a nominal capacitance value of 2 pF and a tuning range of 15% with 3 V. A quality factor of 62 has been demonstrated at 1 GHz, which matches or exceeds that of discrete varactor diodes and is at least an order of magnitude larger than that of a typical junction capacitor implemented in a standard IC process.

MEMS tunable capacitors, based on the varying capacitance area between a set of conductors, have been demonstrated. Figure 8.36 shows an SEM micrograph of such a device [8.34]. The capacitor is composed of arrays of interdigitated electrodes, which can be electrostatically actuated to vary the electrode overlap area. A close-up view of the electrodes is shown in Fig. 8.37. The capacitor is fabricated using a silicon-on-insulator (SOI) substrate with a top silicon layer thickness of about 20 µm to obtain a high-aspect ratio for the electrodes, which is critical for achieving a large capacitance density and reduced tuning voltage. The silicon layer is etched to form the device structure, followed by the removal of the oxide underneath to release the capacitor. A thin aluminum layer is then sputtered over the capacitor to reduce the series resistive loss. The device exhibits a quality factor of 34 at 500 MHz and can be tuned between 2.48 pF and 5.19 pF with an actuation voltage under 5 V, corresponding to a tuning range of 100%. Figure 8.38 shows the variation of electrode overlap under different tuning voltages.

Tunable capacitors relying on a movable dielectric layer have been fabricated using MEMS technology. Figure 8.39 presents an SEM micrograph of a copper-based micromachined tunable capacitor [8.35]. The device consists of an array of copper top electrodes suspended above a bottom copper plate with an air gap of approximately 1 µm. A thin nitride layer is deposited, patterned, and suspended between the two copper layers by lateral mechanical spring suspensions after sacrificial release. A DC voltage applied across the copper layers introduces a lateral electrostatic pull-in force on the nitride, resulting in a movement that changes the overlapping area between each copper electrode and the bottom plate and, hence, the device capacitance. The

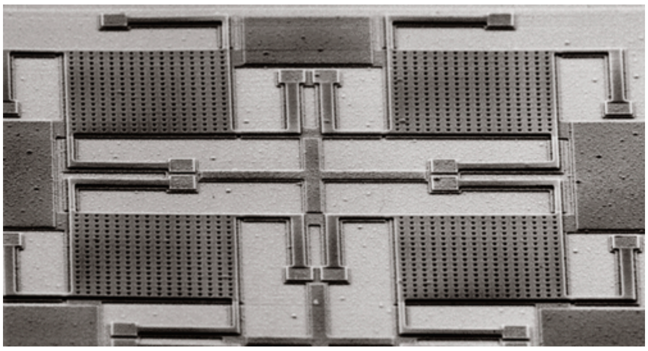

Fig. 8.35 SEM micrograph of four MEMS aluminum variable capacitors connected in parallel [8.30]

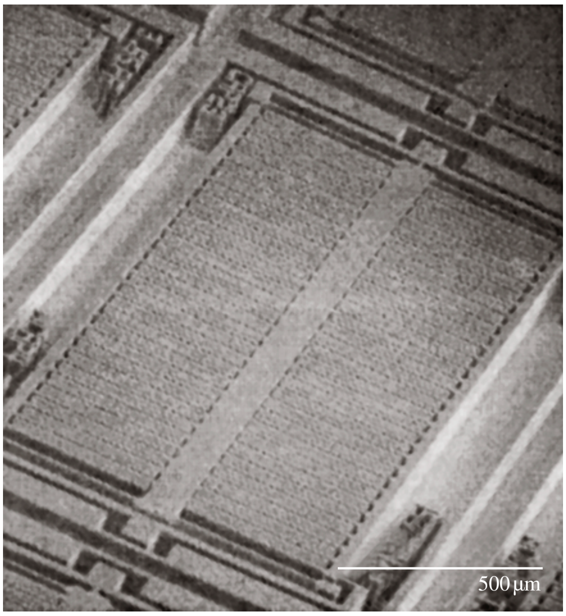

Fig. 8.36 SEM micrograph of a silicon tunable capacitor using a comb drive actuator [8.34]

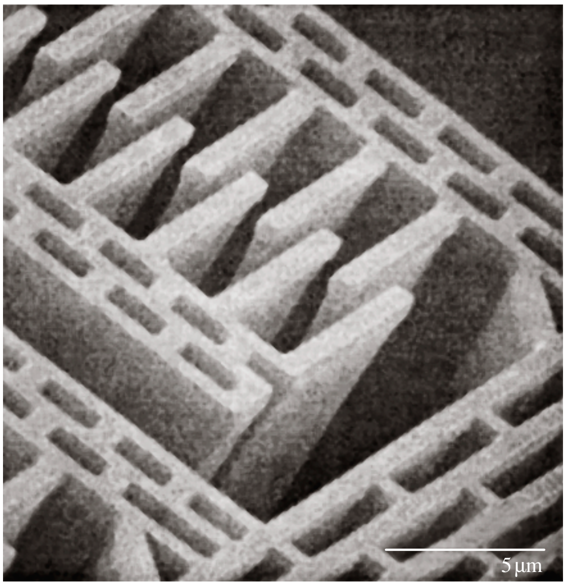

Fig. 8.37 SEM micrograph of a close view of tunable capacitor comb fingers [8.34]

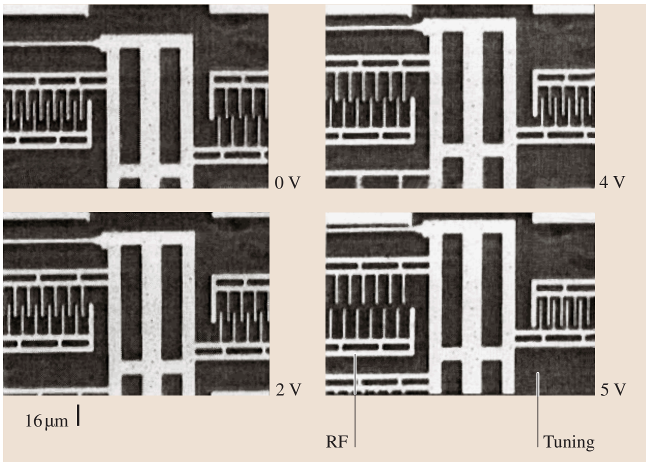

Fig. 8.38 Photos of comb fingers at different actuation voltages [8.34]

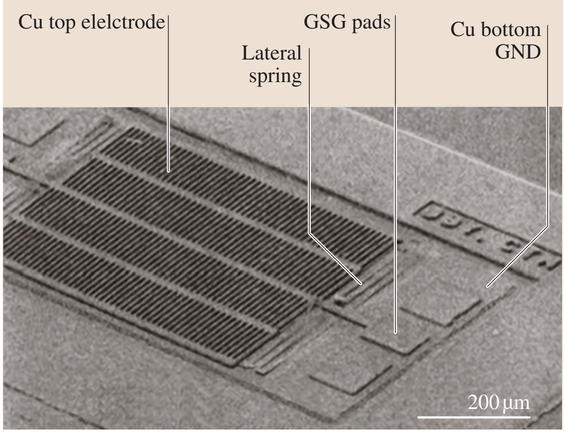

Fig. 8.39 SEM micrograph of a copper surface-micromachined tunable capacitor with a movable dielectric layer [8.35]

tunable capacitor achieves a quality factor of over 200 at 1 GHz with 1 pF capacitance, due to the highly conductive copper layers, and a tuning range around 8% with 10 V.

Micromachined Inductors

Integrated inductors with high quality factors are as critical as the tunable capacitors for high performance RF system implementation. They are the key components for building low noise oscillators, low loss matching networks, etc. Conventional on-chip spiral inductors suffer from limited quality factors of around 5 at 1 GHz, an order of magnitude lower than the required values from discrete counterparts. The poor performance is mainly due to substrate loss and metal resistive loss at high frequencies. Micromachining technology provides an attractive solution to minimizing these loss contributions, hence, enhancing the device quality factors. Figure 8.40 shows an SEM micrograph of a 3-D coil inductor fabricated on a silicon substrate [8.36]. The device consists of four-turn 5 µm-thick copper traces electroplated around an insulating Alumina core with a 650 µm × 500 µm cross section. Compared to spiral inductors, this geometry minimizes the coil area, which is in close proximity to the substrate, and, hence, the eddy-current loss, resulting in a maximized Q-factor and device self-resonant frequency. Copper is selected as the interconnect metal because of its low sheet resistance, critical for achieving a high Q-factor. The inductor achieves a 14 nH inductance value with a quality factor of 16 at 1 GHz. A single-turn 3-D device exhibits a Q-factor of 30 at 1 GHz, which matches the performance of discrete counterparts. The high-Q 3-D inductor and MEMS tunable capacitors, shown in Fig. 8.35, have been employed to implement a RF CMOS VCO achieving a low phase noise performance suitable for typical wireless communication applications, such as GMS cellular telephony [8.37].

Other 3-D inductor structures, such as the levitated spiral inductors, have been demonstrated using micromachining fabrication technology. Figure 8.41 shows an SEM micrograph of a levitated copper inductor,

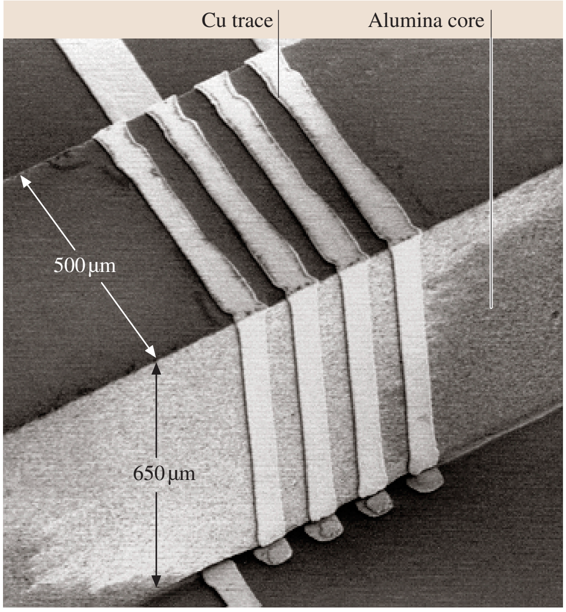

Fig. 8.40 SEM micrograph of a 3-D coil inductor fabricated on a silicon substrate [8.36]

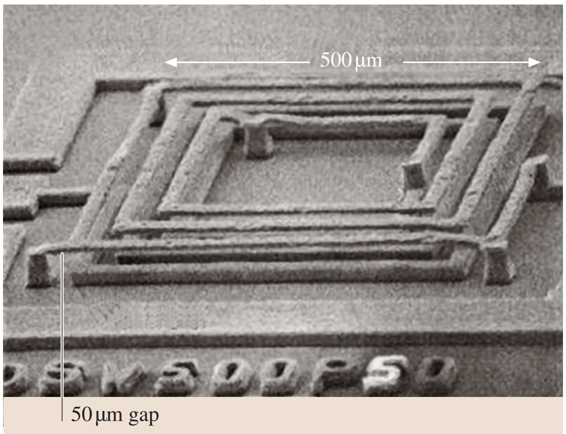

Fig. 8.41 SEM micrograph of a levitated spiral inductor fabricated on a glass substrate [8.38]

which is suspended above the substrate by supporting posts [8.38]. The levitated geometry can minimize the substrate loss, thus achieving an improved quality factor. The inductor shown in the figure achieves a 14 nH inductance value with a Q-factor of 38 at 1.8 GHz using a glass substrate.

A self-assembled out-of-plane coil has been fabricated using micromachining technology. The inductor winding traces are made of refractory metals with controlled built-in stress such that the traces can curl out of the substrate surface upon release and interlock into each other to form coil windings. Figure 8.42 shows an SEM micrograph of a self-assembled out-of-plane coil inductor [8.39]. A close-up view of an interlocking trace is shown in Fig. 8.43. Copper is plated on the interlocked traces to form highly conductive windings at the end of the processing sequence. The inductor shown in Fig. 8.42 achieves a quality factor around 40 at 1 GHz.

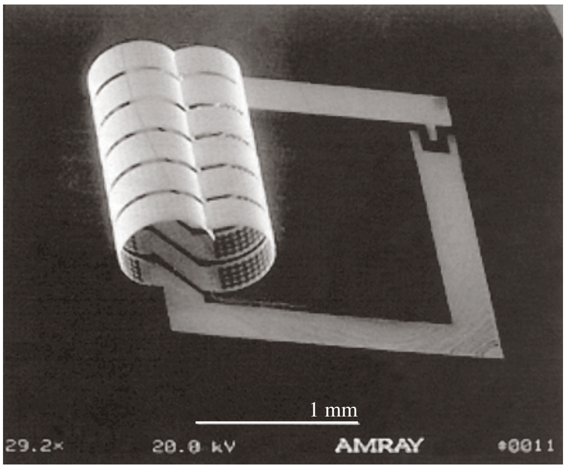

Fig. 8.42 SEM micrograph of a self-assembled out-of-plane coil inductor [8.39]

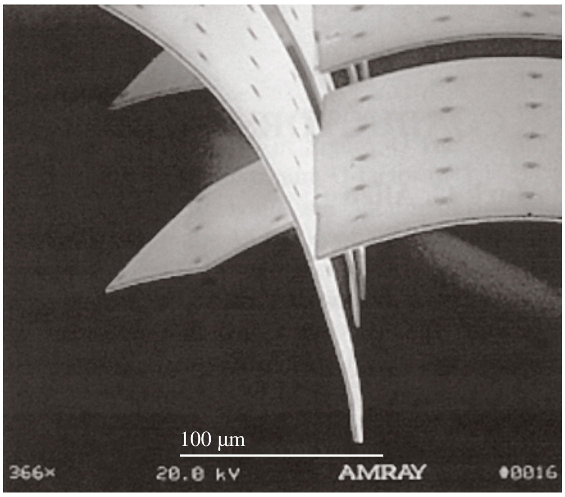

Fig. 8.43 SEM micrograph of an interlocking trace from a self-assembled out-of-plane oil inductor [8.39]

MEMS Switches

The microelectromechanical switch is another potentially attractive miniaturized component enabled by micromachining technologies. These switches offer superior electrical performance in terms of insertion loss, isolation, linearity, etc. and are intended to replace off-chip solid state counterparts, which provide switching between the receiver and transmitter signal paths. They are also critical for building phase shifters, tunable antennas, and filters. The MEMS switches can be broken into two categories: capacitive and metal-to-metal contact types. Figure 8.44 presents a cross-sectional schematic of a RF MEMS capacitive switch. The device consists of a conductive membrane, typically made of aluminum or gold alloy, suspended above a coplanar electrode by an air gap of a few micrometers. For RF or microwave applications, actual metal-to-metal contact is not necessary; rather, a step change in the plate-to-plate capacitance realizes the switching function. A thin silicon nitride layer with a thickness on the order of 1,000 Å is typically deposited above the bottom electrode. When the switch is in the on-state, the membrane is high, resulting in a small plate-to-plate capacitance and, hence, a minimum high-frequency signal coupling (high isolation) between the two electrodes. In the off-state (with a large enough applied DC voltage), the switch provides a large capacitance due to the thin dielectric layer, thus causing a strong signal coupling (low insertion loss). The capacitive switch consumes near zero power, which is attractive for low power portable applications. Switching cycles numbering in the millions for this type of device have been demonstrated. Figure 8.45 shows a top view photo of a fabricated MEMS capacitive switch [8.40]. Surface micromachining technology, using metal for the electrodes and polymer as the sacrificial layer, is used to fabricate the device. The switch can be actuated with a DC voltage on the order of 50 V and exhibits a low insertion loss of approximately −0.28 dB at 35 GHz and a high isolation of −35 dB at the same frequency.

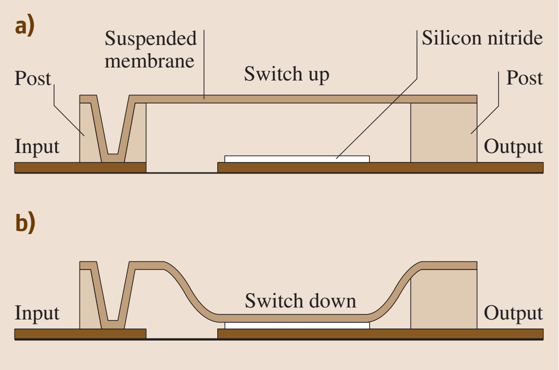

Fig. 8.44 Cross-sectional schematics of an RF MEMS capacitive switch

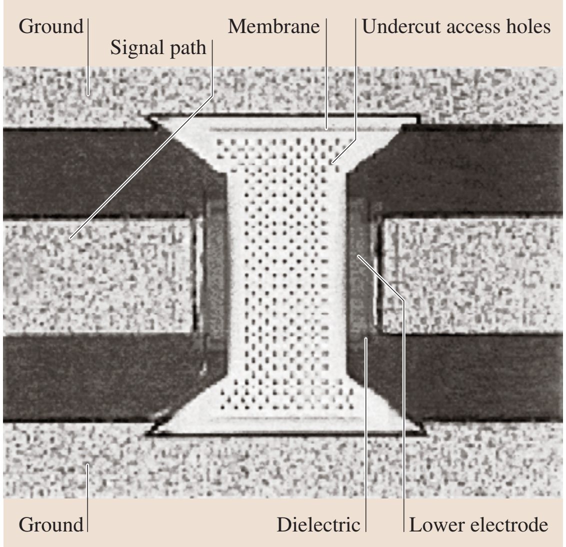

Fig. 8.45 Top view photo of a fabricated RF MEMS capacitive switch [8.40]

Metal-to-metal contact switches are important for interfacing large bandwidth signals, including DC. This type of device typically consists of a cantilever beam, or clamped-clamped bridge with a metallic contact pad positioned at the beam tip or underneath the bridge center. Through an electrostatic actuation, a contact can be formed between the suspended contact pad and an underlying electrode on the substrate [8.41–43]. Figure 8.46 shows a cross-sectional schematic of a metal-to-metal contact switch [8.43]. The top view of the fabricated device is presented in Fig. 8.47. The switch exhibits an actuation voltage of 30 V, a response time of 20 µs, and the mechanical strength to withstand 10^9 actuations. An isolation greater than 50 dB below 2 GHz and insertion loss less than 0.2 dB from DC through 40 GHz have been demonstrated.

MEMS Resonators

Microelectromechanical resonators based on polysilicon comb-drive structures, suspended beams, and center-pivoted disk configurations have been demonstrated for performing analog signal processing [8.44–48]. These micro-resonators can be excited into mechanical reso-

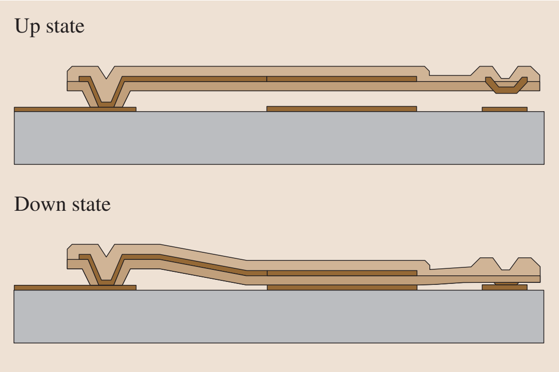

Fig. 8.46 Cross-sectional schematic of a metal-to-metal contact switch [8.43]

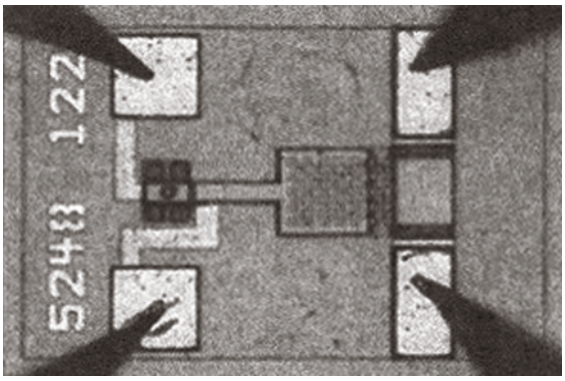

Fig. 8.47 Top view photo of a fabricated metal-to-metal contact switch [8.43]

nance through an electrostatic drive. The mechanical motion causes a change of device capacitance, resulting in an output electrical current when a proper DC bias voltage is used. The output current exhibits the same frequency as the mechanical resonance, thus achieving an electrical filtering function through the electro mechanical coupling. Micromachined polysilicon flexural-mode mechanical resonators have demonstrated a quality factor greater than 80,000 in a 50 µtorr vacuum [8.49]. This level of performance is comparable to a typical quartz crystal and is, therefore, attractive for implementing monolithic low noise and low drift reference signal sources. Figure 8.48 shows an SEM micrograph of a surface micromachined comb-drive resonator integrated with CMOS sustaining electronics on the same substrate to form a monolithic high-Q MEMS resonator-based oscillator [8.44]. The oscillator achieves an operating frequency of 16.5 KHz with a clean spectral purity. A chip area of approximately 420 µm × 230 µm is consumed for fabricating the overall system, representing a size reduction by orders of magnitude compared to conventional quartz crystal oscillators.

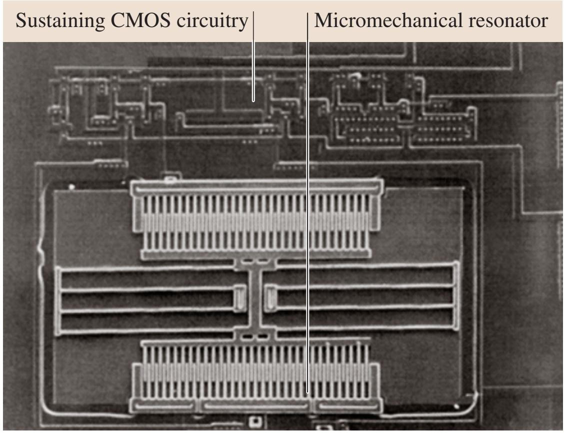

Fig. 8.48 SEM micrograph of a surface micromachined comb drive resonator integrated with CMOS sustaining electronics [8.44]

Micromachined high-Q resonators can be coupled to implement low-loss frequency selection filters. Figure 8.49 shows an SEM micrograph of a surface micromachined polysilicon two-resonator, spring-coupled bandpass micromechanical filter [8.46]. The filter consists of two silicon micromechanical clamped-clamped beam resonators, coupled mechanically by a soft spring, all suspended 0.1 µm above the substrate. Polysilicon strip lines underlie the central regions of each resonator and serve as capacitive transducer electrodes positioned

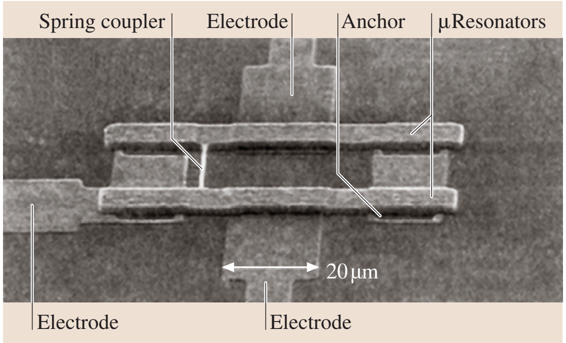

Fig. 8.49 SEM micrograph of a polysilicon surface micromachined two-resonator spring-coupled bandpass micromechanical filter [8.46]

to induce resonator vibration in a direction perpendicular to the substrate. Under normal operation, the device is excited capacitively by a signal voltage applied to the input electrode. The output is taken at the other end of the structure, also by capacitive transduction. The filter achieves a center frequency of 7.81 MHz, a bandwidth of 0.23%, and an insertion loss less than 2 dB. The achieved performance is attractive for implementing filters in the low MHz range.

To obtain a higher mechanical resonant frequency with low losses, a surface micromachined contour-mode disk resonator has been proposed, as shown in Fig. 8.50 [8.48]. The resonator consists of a polysilicon disk suspended 5,000 Å above the substrate with a single anchor at its center. Plated metal input electrodes surround the perimeter of the disk with a narrow separation of around 1,000 Å, which defines the capacitive, electromechanical transducer of the device. To operate the device, a DC bias voltage is applied to the structure with an AC input signal applied to the electrodes, resulting in a time varying electrostatic force acting radially on the disk. When the input signal matches the device resonant frequency, the resulting electrostatic force is amplified by the Q factor of the resonator, resulting in the expansion and contraction of the disk along its radius. This motion, in turn, produces a time varying output current at the same frequency, thus achieving the desirable filtering. The prototype resonator demonstrates an operating frequency of 156 MHz with a Q factor of 9,400 in vacuum. The increased resonant frequency is comparable to the first intermediate frequency used in a typical wireless transceiver design and is thus suitable for implementing IF bandpass filters. A higher operating frequency is needed in order to potentially replace the RF frequency selection filters with microelectromechanical resonators.

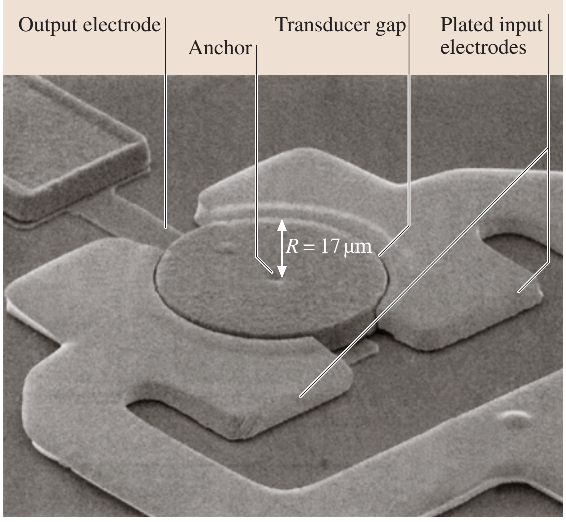

Fig. 8.50 SEM micrograph of a polysilicon surface-micromachined contour-mode disk resonator [8.48]

8.2 NEMS Devices and Applications

Unlike their microscale counterparts, nanoelectromechanical systems (NEMS) are made of electromechanical devices that have critical structural dimensions on the submicron scale. These devices are attractive for applications where structures of very small mass provide essential functionality, such as force sensors, chemical sensors, biological sensors, and ultrahigh frequency resonators, to name a few. NEMS fabrication processes can be categorized according to the two approaches used to create the structures. Top-down approaches utilize submicron lithographic techniques to fabricate device structures from bulk material, either thin films or thick substrates. Bottom-up approaches involve the fabrication of nanoscale devices in much the same way that nature constructs objects, by sequential assembly using atomic and molecular building blocks. While many contend that bottom-up approaches will eventually dominate the fabrication of NEMS, most of the functional NEMS devices are currently created utilizing top-down techniques that combine existing process technologies, such as submicron electron-beam lithography, conventional film growth, and chemical etching. Top-down approaches make integration with microscale packaging relatively straightforward since the only significant difference between the nanoscale and microscale processing steps is the method used to pattern the various features.

In large measure, NEMS has followed a developmental path similar to that of MEMS in that both have leveraged existing processing techniques. For instance, the electron-beam lithographic techniques used in top-down NEMS fabrication are the same techniques that have become standard in the fabrication of submicron ICs. Furthermore, the materials used in many of the first generation, top-down NEMS devices (Si, GaAs, Si_3N_4, SiC) were first used in ICs and then in MEMS. Like the first MEMS devices, the first generation NEMS structures consist of freestanding nanomechanical beams, paddle oscillators, and tethered plates made using simple bulk and single layer surface "nano" machining processes. As an introduction to this new and exciting class of devices, the following section reviews the development of some of these first generation NEMS devices.

Like Si MEMS, Si NEMS capitalizes on well developed processing techniques for Si and the availability of high quality substrates. *Cleland* et al. [8.50] reported a relatively simple process to fabricate nanomechanical clamped-clamped beams directly from single crystal (100) Si substrates. As illustrated in Fig. 8.51, the process begins with the thermal oxidation of a Si substrate (Fig. 8.51a). Large Ni contact pads were then fabricated using optical lithography and lift-off. A polymethyl methacrylate (PMMA) lift-off mold was then deposited and patterned, using electron-beam lithography, into the shape of nanomechanical beams (Fig. 8.51b). Ni was then deposited and patterned by lifting off the PMMA (Fig. 8.51c). Next, the underlying oxide film was patterned by RIE using the Ni film as an etch mask. After oxide etching, nanomechanical beams were patterned by etching the Si substrate using RIE, as shown in Fig. 8.51d. Following Si RIE, the Ni etch mask was removed, and the sidewalls of the Si nanomechanical beams were lightly oxidized in order to protect them during the release step (Fig. 8.51e). After performing an anisotropic SiO_2 etch to clear any oxide from the field areas, the Si beams were released using an isotropic Si RIE step, as shown in Fig. 8.51f. After release, the protective SiO_2 film was removed by wet etching in HF (Fig. 8.51g). Using this process, the authors reported the successful fabrication of nanomechanical Si beams with micron-scale lengths (~ 8 μm) and submicron widths (330 nm) and heights (800 nm).

The advent of silicon-on-insulator (SOI) substrates with high quality, submicron-thick silicon top layers enables the fabrication of nanomechanical Si beams with fewer processing steps than the aforementioned technique, since the buried oxide layer makes these device structures relatively easy to pattern and release. Additionally, the buried SiO_2 layer electrically isolates the beams from the substrate. *Carr* et al. [8.51] detail a process that uses SOI substrates to fabricate submicron clamped-clamped mechanical beams and

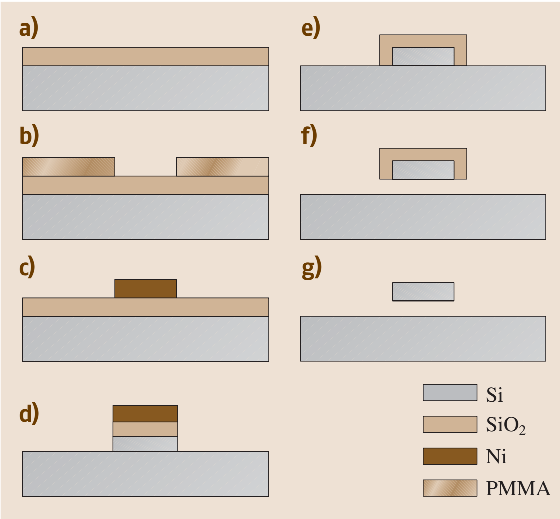

Fig. 8.51a–g Cross-sectional schematics of a process to fabricate nanomechanical beams directly from a silicon substrate: **(a)** after thermal oxidation; **(b)** after PMMA coating and electron beam lithography; **(c)** after Ni deposition and lift-off; **(d)** after anisotropic SiO_2 and Si etching; **(e)** after Ni etching, thermal oxidation, and anisotropic SiO_2 etching; **(f)** after release by isotropic Si etching; **(g)** after oxide removal

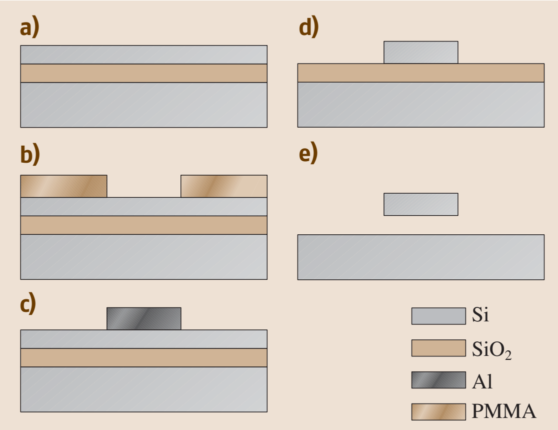

Fig. 8.52a–e Cross-sectional schematics of a process to fabricate nanomechanical structures using silicon-on-insulator substrates: **(a)** starting SOI substrate; **(b)** after coating with PMMA and electron beam lithography; **(c)** after Al deposition and lift-off; **(d)** after anisotropic Si etching; **(e)** after release by SiO_2 etching

suspended plates with submicron tethers. The process, presented in Fig. 8.52, begins with the deposition of PMMA on an SOI substrate. The SOI substrate has a top Si layer that is either 50 nm or 200 nm in thickness. The PMMA is patterned into a metal lift-off mask by electron beam lithography (Fig. 8.52b). An aluminum film is then deposited and patterned by lift-off into a Si etch mask, as shown in Fig. 8.52c. The nanomechanical beams are then patterned by Si RIE and released by etching the underlying SiO_2 in a buffered hydrofluoric acid solution, as shown in Fig. 8.52d and Fig. 8.52e, respectively. Using this process, nanomechanical beams that were 7 to 16 μm in length, 120 to 200 nm in width, and 50 or 200 nm in thickness were successfully fabricated.

Fabrication of NEMS structures is not limited to Si. In fact, III-V compounds, such as gallium arsenide (GaAs), make particularly good NEMS materials from a fabrication perspective because thin epitaxial GaAs films can be grown on lattice-matched materials that can be used as sacrificial release layers. A collection of clamped-clamped nanomechanical GaAs beams fabricated on lattice-matched sacrificial layers having micron-scale lengths and submicron widths and thicknesses is shown in Fig. 8.53. *Tighe* et al. [8.52] reported on the fabrication of GaAs plates suspended with nanomechanical tethers. The structures were made from single crystal GaAs films that were epitaxially grown on aluminum arsenide (AlAs) sacrificial layers. Ni etch masks were fabricated using electron beam lithography and lift-off, as described previously. The GaAs films were patterned into beams using a chemically assisted ion beam etching process and released using a highly selective AlAs etchant. In a second example, *Tang* et al. [8.53] capitalized on the ability to grow high quality GaAs layers on ternary compounds, such as $Al_xGa_{1-x}As$, to fabricate complex GaAs-based structures, such as submicron clamped-clamped beams from GaAs/AlGaAs quantum well heterostructures. As with the process described by *Tighe* et al. [8.52], this process exploits a lattice matched sacrificial layer, in this case $Al_{0.8}Ga_{0.2}As$, which can be selectively etched to release the heterostructure layers.

Several unique approaches have been developed to actuate and sense the motion of NEMS devices. Electrostatic actuation can be used to actuate beams [8.54], tethered meshes [8.55], and paddle oscillators [8.56]. *Sekaric* et al. [8.57] have recently shown that low power lasers can be used to drive paddle oscillators into self-oscillation by induced thermal effects on the structures.

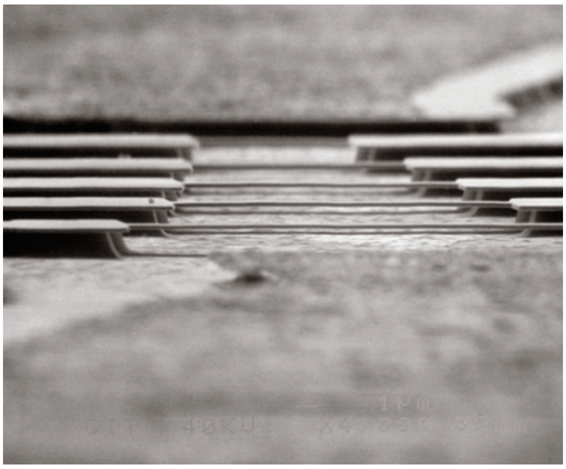

Fig. 8.53 SEM micrograph of a set of nanomechanical GaAs clamped-clamped beam resonators (courtesy of M. Roukes, Caltech)

In these examples, an optical detection scheme based on the modulation of incident laser light by a vibrating beam is used to detect the motion of the beams. *Cleland* et al. [8.50] describe a magnetomotive transduction technique that capitalizes on a time-varying Lorentz force created by an alternating current in the presence of a strong magnetic field. In this case, the nanomechanical beam is positioned in the magnetic field so that an AC current passing through the beam is transverse to the field lines. The resulting Lorentz force causes the beam to oscillate, which creates an electromotive force along the beam that can be detected as a voltage. Thus, in this method, the excitation and detection are performed electrically. In all of the above-mentioned cases, the measurements were performed in vacuum, presumably to minimize the effects of squeeze-film damping, as well as mass loading due to adsorbates from the environment.

For the most part, NEMS technology is still in the initial stage of development. Technological challenges related to fabrication and packaging will require innovative solutions before such devices make a significant commercial impact. Nevertheless, NEMS devices have already been used for precision measurements [8.58], enabling researchers to probe the properties of matter [8.59, 60] on a nanoscopic level. Whether or not NEMS makes an impact in the commercial sector, they will undoubtedly prove to be useful platforms for a host of experiments and scientific discoveries in fields ranging from physics to biology.

8.3 Current Challenges and Future Trends

Although the field of MEMS has experienced significant growth over the past decade, many challenges still remain. In broad terms, these challenges can be grouped into three general categories: (1) fabrication, (2) packaging, and (3) application. Challenges in these areas will, in large measure, determine the commercial success of a particular MEMS device, both in technical and economic terms. The following presents a brief discussion of some of these challenges, as well as possible approaches to address them.

In terms of fabrication, MEMS is currently dominated by planar processing techniques, which find their roots in silicon IC fabrication. The planar approach and the strong dependence on silicon worked well in the early years, since many of the processing tools and methodologies that are commonplace in IC fabrication could be directly utilized in the fabrication of MEMS devices. This approach lends itself to the integration of MEMS with silicon ICs. Therefore, it is still popular for various applications. However, modular process integration of micromachining with standard IC fabrication is not straightforward and represents a great challenge in terms of processing material compatibility, thermal budget requirements, etc. Furthermore, planar processing places significant geometric restrictions on device designs, especially for complex mechanical components requiring high-aspect ratio three-dimensional geometries, which are certain to increase as the application areas for MEMS continue to grow. Along the same lines, new applications will likely demand materials other than silicon that may not be compatible with the conventional microfabrication approach, posing a significant challenge if integration with silicon microelectronics is required. Micro-assembly technique can become an attractive solution to alleviate these issues. Multifunctional microsystems can be implemented by assembling various MEMS devices and electronic building blocks fabricated through disparate processing technologies. Microsystem on a common substrate will likely become the ultimate solution. Development of sophisticated modeling programs for device design and performance will become increasingly important as fabrication processes and device designs become more complex. In terms of NEMS, the most significant challenge is likely the integration of nano- and microfabrication techniques into a unified process, since NEMS devices are likely to consist of both nanoscale and microscale structures. Integration will be particularly challenging for nanoscale devices fabricated using a bottom-up approach, since no analog is found in microfabrication. Nevertheless, hybrid systems consisting of nanoscale and microscale components will become increasingly common as the field continues to expand.

Fabrication issues notwithstanding, packaging is and will continue to be a significant challenge to the implementation of MEMS. MEMS is unlike IC packaging, which benefits from a high degree of standardization. MEMS devices inherently require interaction with the environment, and since each application has in some way a unique environment, standardization of packaging becomes extremely difficult. This lack of standardization tends to drive up the costs associated with packaging, making MEMS less competitive with alternative approaches. In addition, packaging tends to negate the effects of miniaturization based upon microfabrication, especially for MEMS devices requiring protection from certain environmental conditions. Moreover, packaging can cause performance degradation of MEMS devices, especially in situations where the environment exerts mechanical stresses on the package, which, in turn, results in a long-term device performance drift. To address many of these issues, wafer-level packaging schemes that are customized to the device of interest will likely become more common. In essence, packaging of MEMS will move away from the conventional IC methods, which utilize independently manufactured packages, toward custom packages, which are created specifically for the device as part of the batch fabrication process.

Without question, the advancement of MEMS will open many new potential application areas to the technology. In most cases, MEMS will be one of several alternatives available for implementation. For cost-sensitive applications, the trade off between technical capabilities and cost will challenge those who desire to commercialize the technology. The biggest challenge to the field will be to identify application areas that are well suited for MEMS/NEMS technology and have no serious challengers. As MEMS technology moves away from the component level and towards microsystems solutions, it is likely that such application areas will come to the fore.

References

8.1 M. Mehregany, S. F. Bart, L. S. Tavrow, J. H. Lang, S. D. Senturia: Principles in design and microfabrication of variable-capacitance side-drive motors, J. Vac. Sci. Tech. A **8** (1990) 3614–3624

8.2 Y.-C. Tai, R. S. Muller: IC-processed electrostatic synchronous micromotors, Sens. Actuators **20** (1989) 49–55

8.3 M. S. Rodgers, J. J. Sniegowski: 5-level polysilicon surface micromachine technology: Application to complex mechanical systems, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island 1998 (Transducers Research Foundation, Cleveland 1998) 144–149

8.4 G. T. A. Kovacs: *Micromachined Transducer Sourcebook* (McGraw-Hill, Boston 1998)

8.5 S. D. Senturia: *Microsystem Design* (Kluwer, Dordrecht 1998)

8.6 J. E. Gragg, W.E. McCulley, W. B. Newton, C. E. Derrington: Compensation and calibration of a monolithic four terminal silicon pressure transducer, Technical Digest, IEEE Solid-State Sens. Actuators Workshop, Hilton Head Island 1984 (IEEE, Piscataway 1984) 21–27

8.7 Y. Wang, M. Esashi: A novel electrostatic servo capacitive vacuum sensor, Technical Digest, IEEE Int. Conf. Solid-State Sens. Actuators, Chicago 1997 (Transducers Conference, Chicago 1997) 1457–1460

8.8 W. H. Ko, Q. Wang: Touch mode capacitive pressure sensors, Sens. Actuators **75** (1999) 242–251

8.9 J. M. Bustillo, R.T. Howe, R.S. Muller: Surface micromachining for microelectromechanical systems, Proc. IEEE **86** (1998) 1552–1574

8.10 J. H. Smith, S. Montague, J. J. Sniegowski, J. R. Murray, P. J. McWhorter: Embedded micromechanical devices for the monolithic integration of MEMS with CMOS, Technical Digest, IEEE Int. Electron Devices Meeting, Washington D.C. 1993 (IEEE, Piscataway 1993) 609–612

8.11 T. A. Core, W. K. Tsang, S. J. Sherman: Fabrication technology for an integrated surface-micromachined sensor, Solid State Technol. **36**(10) (1993) 39–40, 42, 44, 46–47

8.12 H. Kapels, R. Aigner, C. Kolle: Monolithic surface-micromachined sensor system for high pressure applications, Technical Digest, Int. Conf. Solid-State Sens. Actuators, Munich 2001 (Transducers Conference, Chicago 2001) 56–59

8.13 C. Lu, M. Lemkin, B. E. Boser: A monolithic surface micromachined accelerometer with digital output, Technical Digest, IEEE Int.Solid-State Circuits Conferernce, San Francisco 1995 (IEEE, Piscataway 1995) 160–161

8.14 T. B. Gabrielson: Mechanical-thermal noise in micromachined acoustic and vibration sensors, IEEE Trans. Electron Devices **40** (1993) 903–909

8.15 N. Yazdi, K. Najafi: An all-silicon single-wafer fabrication technology for precision microaccelerometers, Technical Digest, IEEE Int. Conf. Solid-State Sens. Actuators, Chicago 1997 (Transducers Conference, Chicago 1997) 1181–1184

8.16 M. Lemkin, M. A. Ortiz, N. Wongkomet, B. E. Boser, J. H. Smith: A 3-axis surface micromachined $\Sigma\Delta$ accelerometer, Technical Digest, IEEE Int. Solid-State Circuits Conference, San Francisco 1997 (IEEE, Piscataway 1997) 202–203

8.17 W. A. Clark, R. T. Howe: Surface micromachined Z-axis vibratory rate gyroscope, Technical Digest, IEEE Solid-State Sens. Actuators Workshop, Hilton Head Island 1996 (Transducers Research Foundation, Inc., Cleveland 1996) 283–287

8.18 T. Juneau, A. P. Pisano: Micromachined dual input axis angular rate sensor, Technical Digest, IEEE Solid-State Sens. Actuators Workshop, Hilton Head Island 1996 (Transducers Research Foundation, Inc., Cleveland 1996) 299–302

8.19 R. S. Muller, K. Y. Lau: Surface-micromachined microoptical elements and systems, Proc. IEEE **86** (1998) 1705–1720

8.20 L. J. Hornbeck: Current status of the digital micromirror device (DMD) for projection television applications, Technical Digest, IEEE Int. Electron Devices Meeting, Washington D.C. 1993 (IEEE, Piscataway 1993) 381–384

8.21 P. F. Van Kessel, L. J. Hornbeck, R. E. Meier, M. R. Douglass: A MEMS-based projection display, Proc. IEEE **86** (1998) 1687–1704

8.22 M. S. Cohen, M. F. Cina, E. Bassous, M. M. Opyrsko, J. L. Speidell, F. J. Canora, M. J. DeFranza: Packaging of high density fiber/laser modules using passive alignment techniques, IEEE Trans. Comp., Hybrids, Manufact. Technol. **15** (1992) 944–954

8.23 M. J. Wale, C. Edge: Self-aligned flip-chip assembly of photonic devices with electrical and optical connections, IEEE Trans. Comp., Hybrids, Manufact. Technol. **13** (1990) 780–786

8.24 M. J. Daneman, N. C. Tien, O. Solgaard, K. Y. Lau, R. S. Muller: Linear vibromotor-actuated micromachined microreflector for integrated optical systems, Technical Digest, IEEE Solid-State Sens. Actuators Workshop, Hilton Head Island 1996 (Transducers Research Foundation, Inc., Cleveland 1996) 109–112

8.25 K. S. J. Pister, M. W. Judy, S. R. Burgett, R. S. Fearing: Microfabricated hinges, Sens. Actuators A **33** (1992) 249–256

8.26 O. Solgaard, M. Daneman, N. C. Tien, A. Friedberger, R. S. Muller, K. Y. Lau: Optoelectronic packaging using silicon surface-micromachined alignment mirrors, IEEE Photon. Technol. Lett. **7** (1995) 41–43

8.27 S. S. Lee, L. S. Huang, C. J. Kim, M. C. Wu: 2 × 2 MEMS fiber optic switches with silicon sub-mount for low-cost packaging, Technical Digest, IEEE Solid-State Sens. Actuators Workshop, Hilton Head Island 1998 (Transducers Research Foundation, Inc., Cleveland 1998) 281–284

8.28 T. Akiyama, H. Fujita: A quantitative analysis of scratch drive actuator using buckling motion, Technical Digest, 8th IEEE Int. MEMS Workshop, Amsterdam 1995 (IEEE, Piscataway 1995) 310–315

8.29 V. A. Aksyuk, M. E. Simon, F. Pardo, S. Arney: Optical MEMS design for telecommunications applications, Technical Digest, Solid-State Sensors and Actuators Workshop, Hilton Head Island 2002 (Transducers Research Foundation, Cleveland 2002) 1–6

8.30 D. J. Young, B. E. Boser: A micromachined variable capacitor for monolithic low-noise VCOs, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island 1996 (Transducers Research Foundation, Cleveland 1996) 86–89

8.31 A. Dec, K. Suyama: Micromachined electro-mechanically tunable capacitors and their applications to RF IC's, IEEE Trans. Microwave Theory Techniques **46** (1998) 2587–2596

8.32 Z. Li, N. C. Tien: A high tuning-ratio silicon-micromachined variable capacitor with low driving voltage, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island 2002 (Transducers Research Foundation, Inc., Cleveland 2002) 239–242

8.33 Z. Xiao, W. Peng, R. F. Wolffenbuttel, K. R. Farmer: Micromachined variable capacitor with wide tuning range, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island, South Carolina 2002 (Transducers Research Foundation, Inc., Cleveland 2002) 346–349

8.34 J. J. Yao, S. T. Park, J. DeNatale: High tuning-ratio MEMS-based tunable capacitors for RF communications applications, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island 1998 (Transducers Research Foundation, Inc., Cleveland 1998) 124–127

8.35 J. B. Yoon, C. T.-C. Nguyen: A high-Q tunable micromechanical capacitor with movable dielectric for RF applications, Technical Digest, IEEE Int. Electron Devices Meeting, San Francisco 2000 (IEEE, Piscataway 2000) 489–492

8.36 D. J. Young, V. Malba, J. J. Ou, A. F. Bernhardt, B. E. Boser: Monolithic high-performance three-dimensional coil inductors for wireless communication applications, Technical Digest, IEEE Int. Electron Devices Meeting, Washington D.C. 1997 (IEEE, Piscataway 1997) 67–70

8.37 D. J. Young, B. E. Boser, V. Malba, A. F. Bernhardt: A micromachined RF low phase noise voltage-controlled oscillator for wireless communication, Int. J. RF Microwave Computer-Aided Eng. **11** (2001) 285–300

8.38 J. B. Yoon, C. H. Han, E. Yoon, K. Lee, C. K. Kim: Monolithic High-Q Overhang Inductors Fabricated on Silicon and Glass Substrates, Technical Digest, IEEE Int. Electron Devices Meeting, Washington D.C. 1999 (IEEE, Piscataway 1999) 753–756

8.39 C. L. Chua, D. K. Fork, K. V. Schuylenbergh, J. P. Lu: Self-assembled out-of-plane high Q inductors, Technical Digest, Solid-State Sens. Actuator Workshop, Hilton Head Island 2002 (Transducers Research Foundation, Inc., Cleveland 2002) 372–373

8.40 C. L. Goldsmith, Z. Yao, S. Eshelman, D. Denniston: Performance of low-loss RF MEMS capacitive switches, IEEE Microwave Guided Wave Lett. **8** (1998) 269–271

8.41 J. J. Yao, M. F. Chang: A surface micromachined miniature switch for telecommunication applications with signal frequencies from DC up to 40 GHz, Technical Digest, 8th Int. Conf. Solid-State Sens. Actuators, Stockholm 1995 (Transducers Conference, Chicago 1995) 384–387

8.42 P. M. Zavracky, N. E. McGruer, R. H. Morrison, D. Potter: Microswitches and microrelays with a view toward microwave applications, Int. J. RF Microwave Computer-Aided Eng. **9** (1999) 338–347

8.43 D. Hyman, J. Lam, B. Warneke, A. Schmitz, T. Y. Hsu, J. Brown, J. Schaffner, A. Walston, R. Y. Loo, M. Mehregany, J. Lee: Surface-micromachined RF MEMs switches on GaAs substrates, Int. J. RF Microwave Computer-Aided Eng. **9** (1999) 348–361

8.44 C. T. C. Nguyen, R. T. Howe: CMOS microelectromechanical resonator oscillator, Technical Digest, IEEE Int. Electron Devices Meeting, Washington D.C. 1993 (IEEE, Piscataway 1993) 199–202

8.45 L. Lin, R. T. Howe, A. P. Pisano: Microelectromechanical filters for signal processing, IEEE J. Microelectromechan. Syst. **7** (1998) 286–294

8.46 F. D. Bannon III, J. R. Clark, C. T. C. Nguyen: High frequency micromechanical filter, IEEE J. Solid-State Circuits **35** (2000) 512–526

8.47 K. Wang, Y. Yu, A. C. Wong, C. T. C. Nguyen: VHF free-free beam high-Q micromechanical resonators, Technical Digest, 12th IEEE Int. Conf. Micro Electro Mechanical Systems, San Francisco 1999 (IEEE, Piscataway 1999) 453–458

8.48 J. R. Clark, W. T. Hsu, C. T. C. Nguyen: High-Q VHF micromechanical contour-mode disk resonators, Technical Digest, IEEE Int. Electron Devices Meeting, San Francisco 2000 (IEEE, Piscataway 2000) 493–496

8.49 C. T. C. Nguyen, R. T. Howe: Quality factor control for micromechanical resonator, Technical Digest, IEEE Int. Electron Devices Meeting, San Francisco 1992 (IEEE, Piscataway 1992) 505–508

8.50 A. N. Cleland, M. L. Roukes: Fabrication of high frequency nanometer scale mechanical resonators from bulk Si crystals, Appl. Phys. Lett. **69** (1996) 2653–2655

8.51 D. W. Carr, H. G. Craighead: Fabrication of nanoelectromechanical systems in single crystal silicon using silicon on insulator substrates and electron beam lithography, J. Vac. Sci. Technol. B **15** (1997) 2760–2763
8.52 T. S. Tighe, J. M. Worlock, M. L. Roukes: Direct thermal conductance measurements on suspended monocrystalline nanostructures, Appl. Phys. Lett. **70** (1997) 2687–2689
8.53 H. X. Tang, X. M. H. Huang, M. L. Roukes, M. Bichler, W. Wegsheider: Two-dimensional electron-gas actuation and transduction for GaAs nanoelectromechanical systems, Appl. Phys. Lett. **81** (2002) 3879–3881
8.54 D. W. Carr, S. Evoy, L. Sekaric, H. G. Craighead, J. M. Parpia: Measurement of mechanical resonance and losses in nanometer scale silicon wires, Appl. Phys. Lett. **75** (1999) 920–922
8.55 D. W. Carr, L. Sekaric, H. G. Craighead: Measurement of nanomechanical resonant structures in single-crystal silicon, J. Vac. Sci. Technol. B **16** (1998) 3821–3824
8.56 S. Evoy, D. W. Carr, L. Sekaric, A. Olkhovets, J. M. Parpia, H. G. Craighead: Nanofabrication and electrostatic operation of single-crystal silicon paddle oscillators, J. Appl. Phys. **86** (1999) 6072–6077
8.57 L. Sekaric, M. Zalalutdinov, S. W. Turner, A. T. Zehnder, J. M. Parpia, H. G. Craighead: Nanomechanical resonant structures as tunable passive modulators, Appl. Phys. Lett. **80** (2002) 3617–3619
8.58 A. N. Cleland, M. L. Roukes: A nanometre-scale mechanical electrometer, Nature **392** (1998) 160–162
8.59 K. Schwab, E. A. Henriksen, J. M. Worlock, M. L. Roukes: Measurement of the quantum of thermal conductance, Nature **404** (2000) 974–977
8.60 S. Evoy, A. Olkhovets, L. Sekaric, J. M. Parpia, H. G. Craighead, D. W. Carr: Temperature-dependent internal friction in silicon nanoelectromechanical systems, Appl. Phys. Lett. **77** (2000) 2397–2399

9. Microfluidics and Their Applications to Lab-on-a-Chip

Various microfluidic components and their characteristics, along with the demonstration of two recent achievements of lab-on-a-chip systems have been reviewed and discussed. Many microfluidic devices and components have been developed during the past few decades, as introduced earlier for various applications. The design and development of microfluidic devices still depend on the specific purposes of the devices (actuation or sensing) due to a wide variety of application areas, which encourages researchers to develop novel, purpose-specific microfluidic devices and systems. Microfluidics is the real multidisciplinary research field that requires basic knowledge in fluidics, micromachining, electromagnetics, materials, and chemistry for better applications.

Among the various application areas of microfluidics, one of the most important application areas is the lab-on-a-chip system. Lab-on-a-chip is becoming a revolutionary tool for many different applications in chemical and biological analyses due to its fascinating advantages (fast and low cost) over conventional chemical or biological laboratories. Furthermore, the simplicity of lab-on-a-chip systems will enable self-testing capability for patients or health consumers overcoming space limitation.

9.1 **Materials for Microfluidic Devices and Micro/Nano Fabrication Techniques** 254
  9.1.1 Silicon 254
  9.1.2 Glass 254
  9.1.3 Polymer 255

9.2 **Active Microfluidic Devices** 257
  9.2.1 Microvalves 258
  9.2.2 Micropumps 260

9.3 **Smart Passive Microfluidic Devices** 262
  9.3.1 Passive Microvalves 262
  9.3.2 Passive Micromixers 265
  9.3.3 Passive Microdispensers 266
  9.3.4 Microfluidic Multiplexer Integrated with Passive Microdispenser 267
  9.3.5 Passive Micropumps 269
  9.3.6 Advantages and Disadvantages of the Passive Microfluidic Approach 269

9.4 **Lab-on-a-Chip for Biochemical Analysis** 270
  9.4.1 Magnetic Micro/Nano Bead-Based Biochemical Detection System 270
  9.4.2 Disposable Smart Lab-on-a-Chip for Blood Analysis 273

References 276

Microfluidics covers the science of fluidic behaviors on the micro/nanoscales and the engineering of design, simulation, and fabrication of the fluidic devices for the transport, delivery, and handling of fluids on the order of microliters or smaller volumes. It is the backbone of the BioMEMS (Biological or Biomedical Microelectromechanical Systems) and lab-on-a-chip concept, as most biological analyses involve fluid transport and reaction. Biological or chemical reactions on the micro/nanoscale are usually rapid since small amounts of samples and reagents are used, which offers a quick and low cost analysis.

A fluidic volume of 1 nanoliter (1 nl) can be understood as the volume in a cube surrounded by $100\,\mu\text{m}$ in each direction. It is much smaller than the size of a grain of table salt! Microfluidic devices and systems handle sample fluids in this range for various applications, including inkjet printing, blood analysis, biochemical detection, chemical synthesis, drug screening/delivery, protein analysis, DNA sequencing, and so on.

Microfluidic systems consist of microfluidic platforms or devices for fluidic sampling, control, monitoring, transport, mixing, reaction, incubation, and analysis. To construct microfluidic systems, or lab-on-a-chips, microfluidic devices must be functionally integrated on a microfluidic platform using proper micro/nano fabrication techniques. In this chapter, the basics of microfluidicdevices and their applications to

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

lab-on-a-chip are briefly reviewed and summarized. Basic materials and fabrication techniques for microfluidic devices will be introducedfirst and various active and passive microfluidic components will be described. Then, their applications to lab-on-a-chip, or biochemical analysis, will be discussed.

9.1 Materials for Microfluidic Devices and Micro/Nano Fabrication Techniques

Various materials are being used for the fabrication of microfluidic devices and systems. Silicon is one of the most popular materials in micro/nano fabrication because its micromachining has been well established over decades. In general, the advantages to using silicon as a substrate or structural material include good mechanical properties, excellent chemical resistance, well characterized processing techniques, and the capability for integration of control/sensing circuitry as a semiconductor. Other materials such as glass, quartz, ceramics, metals, and polymers are also being used for substrates and structures in micro/nano fabrication, depending on the application. Among these materials, polymers or plastics have recently become one of the more promising materials for lab-on-a-chip applications, due to their excellent material properties for biochemical fluids and their low cost manufacturability. Main issues in the fabrication techniques of microfluidic devices and systems usually lie in forming microfluidic channels, key micro/nano structures of lab-on-a-chip. In this section, the basic micro/nano fabrication techniques for silicon, glass, and polymers are described.

9.1.1 Silicon

Microfluidic channels on silicon substrates are usually formed either by wet (chemical) etching or by dry (plasma) etching. Crystalline silicon has a preferential etch direction, depending on which crystalline plane is exposed to an etchant. Etch rate is slowest in ⟨111⟩ crystalline direction – approximately 100:1 etch rate anisotropy compared with ⟨100⟩:⟨111⟩ or ⟨110⟩:⟨111⟩. Potassium hydroxide (KOH), tetra-methyl ammonium hydroxide (TMAH), and ethylene diamine pyrocatechol (EDP) are commonly used silicon anisotropic etchants. In most cases, silicon dioxide (SiO_2) or silicon nitride (Si_3N_4) is used as a masking material during the etching process. Anisotropic etchants and basic etching mechanisms are summarized by *Ristic* et al. [9.1]. There is also an isotropic wet etching process available using a mixture of fluoric acid (HF), nitric acid (HNO_3), and acetic acid (CH_3COOH), so-called "HNA". HNA etches in all directions with almost the same etch rate regardless of crystalline directions. Figure 9.1 illustrates wet anisotropic and isotropic etching profiles.

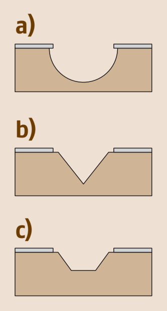

Fig. 9.1a–c Wet etching of silicon substrate for anisotropic and isotropic etching: **(a)** isotropic etching profile, **(b)** anisotropic etching profile (long term), and **(c)** anisotropic etching profile (short term)

Reactive ion etching (RIE) is also one of the most commonly used dry etching processes to generate microfluidic channels or deep trench structures on silicon substrate. In this dry etching technique, radio frequency (RF) energy is used to excite ions in a gas to an energetic state. The energized ions supply the necessary energy to generate physical and chemical reactions on the exposed area of the substrate, which starts the etching process. RIE can generate strong anisotropic, as well as isotropic profiles, depending on the gases used, the condition of plasma, and the applied power. Further information on reactive ion etching process, including deep reactive ion etching (DRIE), on silicon substrate can be found in the literature [9.2–5].

Many microfluidic devices were realized using silicon as a substrate material, including microvalves and micropumps, which are covered in the next section.

9.1.2 Glass

Glass substrate has been widely used for the fabrication of microfluidic systems and lab-on-a-chip due to its excellent optical transparency and ease of electro-osmotic flow (EOF). Chemical wet etching and thermal fusion bonding are the common fabrication techniques for glass substrate. Chemical wet etching and the bonding technique have also been widely reported [9.6, 7]. Most commonly used etchants are hydrofluoric acid (HF), buffered hydrofluoric acid, and a mixture of hydrofluoric acid, nitric acid, and deionized water (HF, HNO_3, H_2O). Gold with an adhesion layer of chrome is most often used as etch mask for wet etching of glass substrate. Since glass has no crystalline structure, only isotropic etch profiles are obtained as forming a hemispherical-shaped channel. Often the problem is that stresses within the surface layers of the glass cause preferential etching, and scratches created by polishing or handling errors cause spikes etched in the channels. Pre-etching is one method to release stress, which causes defects in channels after etching. Another way to improve the channel etching process is to anneal the glass wafers before etching. Annealing can be done nearly at glass transition temperature for at least a couple of hours. Figure 9.2 shows two examples of poorly etched channels without the pre-etching and annealing steps compared with an etched channel without defects. Other fabrication techniques for glass substrate include photoimageable glass, as *Dietrich* et al. [9.8] reported. Anisotropy is introduced in glass by making the glass photosensitive using lithium/aluminium/silicates composition.

One of the most successful examples of using glass as a substrate material in lab-on-a-chip applications is the capillary electrophoresis (CE) chip, fabricated using the glass etching and fusion bonding technique, since the most advantageous property of glass is an excellent optical transparency that is required for the most lab-on-a-chip using optical detection including capillary electrophoresis microchips [9.6, 7, 9].

9.1.3 Polymer

Among several substrates available for lab-on-a-chip, polymers, or plastics, have recently become one of the most popular and promising substrates due to their low cost, ease-of-fabrication, and favorable biochemical reliability and compatibility. Plastic substrates, such as polyimide, PMMA, PDMS, polyethylene or polycarbonate, offer a wide range of physical and chemical material parameters for the applications of lab-on-a-chip, generally at low cost using replication approaches. Polymers and plastics are promising materials in microfluidic and lab-on-a-chip applications because they can be used for mass production using casting, hot embossing, and injection molding techniques. This mass production capability allows the successful commercialization of lab-on-a-chip technology, including disposable lab-on-a-chip. While several fabrication methods have recently been developed, three fabrication techniques – casting, hot embossing, and injection molding – are major techniques of great interest. Figure 9.3 shows schematic illustrations of these polymer microfabrication techniques.

For polymer or plastic micro/nano fabrication, a mold master is essential for replication. Mold masters are fabricated using photolithography, silicon/glass bulk etching, and metal electroplating, depending on the application. Figure 9.4 summarizes mold masters from different fabrication techniques.

Photolithography, including UV-LIGA [9.10, 11] and LIGA [9.12], is used to fabricate mold masters for casting or soft lithography replication, while silicon-based and electroplated mold masters are used for hot embossing replication. For injection molding, electroplated metallic mold masters are preferable.

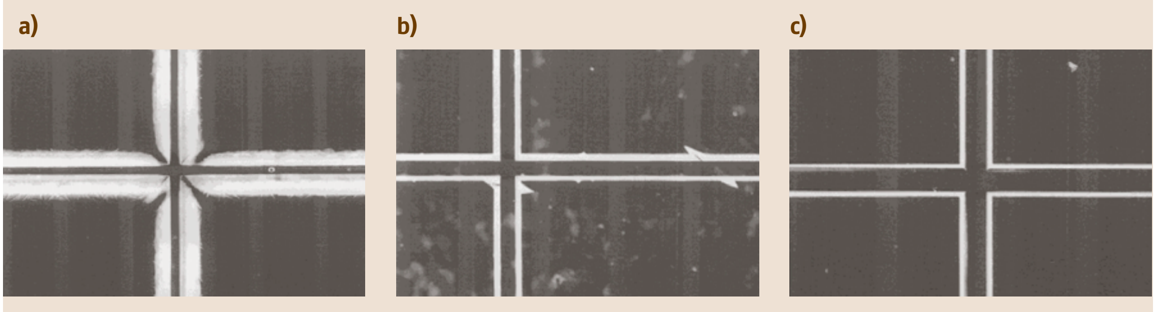

Fig. 9.2a–c Isotropically etched microfluidic channels on glass substrate: **(a)** poorly etched channel with large under etching, **(b)** poorly etched channel with spikes, and **(c)** well etched channel without any defects

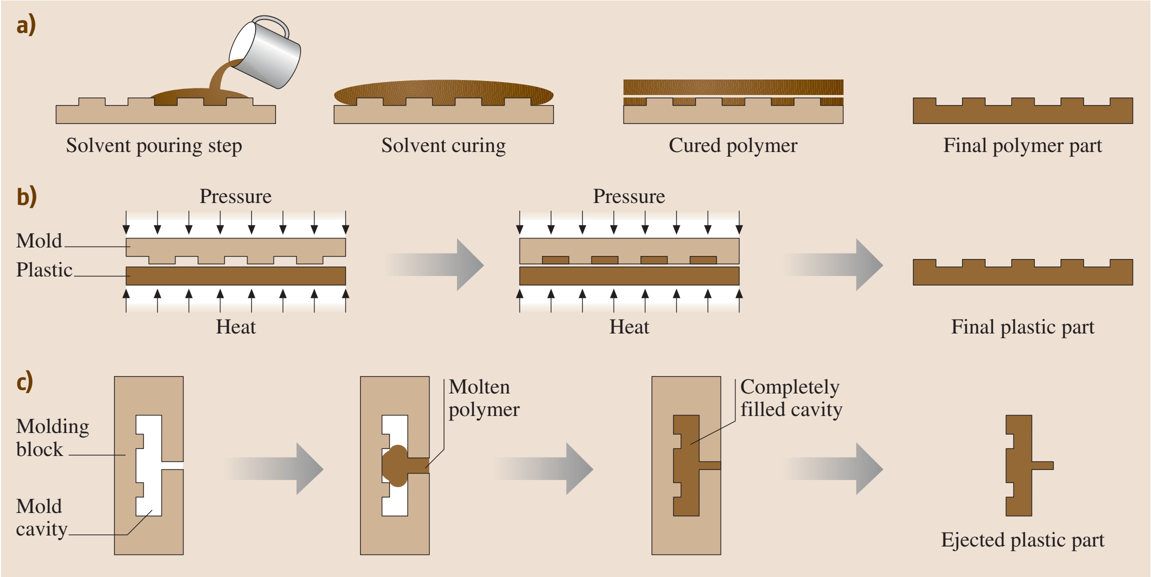
a)
Solvent pouring step
Solvent curing
Cured polymer
Final polymer part
b)
Pressure
Mold
Plastic
Heat
Pressure
Heat
Final plastic part
c)
Molding block
Mold cavity
Molten polymer
Completely filled cavity
Ejected plastic part

Fig. 9.3a–c Concept of polymer micro/nano fabrication techniques: **(a)** casting, **(b)** hot embossing, and **(c)** injection molding

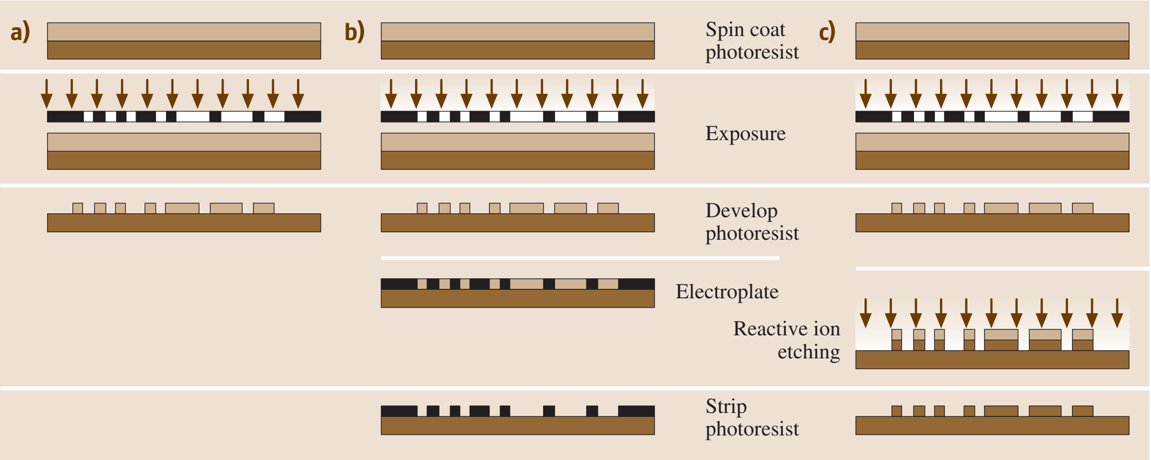
a)
b)
Spin coat photoresist
c)
Exposure
Develop photoresist
Electroplate
Reactive ion etching
Strip photoresist

Fig. 9.4a–c Mold masters in polymer/plastic fabrication: **(a)** photolithography-based mold master, **(b)** silicon-based mold master, and **(c)** mold master by electroplating

Casting or soft lithography [9.13, 14] usually offers flexible access to microfluidic structures using mostly poly(dimethylsiloxane) (PDMS) as a casting material. A mixture of the elastomer precursor and curing agent is poured over a master mold structure. After curing, the replicated elastomer is released from the mold master, having transferred a reverse structure of the mold master. Patterns of a few nm can be achieved using this technique.

While casting can be carried out in room temperature, hot embossing requires a slightly higher temperature – up to glass transition temperature of the plastic substrate to be replicated. The hot embossing technique has been developed by several research groups [9.15–17]. A mold master is placed in the chamber

Table 9.1 Overview of the different polymer micro/nano fabrication techniques

Fabrication type	Casting	Hot embossing	Injection molding
Investment	Low	Moderate	High
Manufacturability	Low	Moderate	High
Cycle time	8–10 hrs	1 hr	1 min
Polymer choices	Low	Moderate	Moderate
Mold replication	Good	Good	Good
Reusability of mold	No (photolithography-based molds)	Yes	Yes

of a hot embossing system with the plastic substrate, then heated plates press both the plastic substrate and the mold master, as illustrated in Fig. 9.3b. After a certain amount of time (typically 5–20 minutes, depending on the plastic substrate), the plates are cooled down to release the replicated plastic substrate. Hot embossing offers mass production of polymer microstructures, as its cycle time is less than an hour.

Injection molding is a technique to fabricate polymer microstructures at low cost and high volumes [9.18]. A micromachined mold master is placed in the molding block of the injection molding machine, as illustrated in Fig. 9.3c. The plastic in granular form is melted and then injected into the cavity of a closed mold block, where the mold master is located. The molten plastic continues to flow into and fill the mold cavity until the plastic cools down to a highly viscous melt, and a cooled plastic part is ejected. In order to ensure good flow properties during the injection molding process, thermoplastics with low or medium viscosity are desirable. So the filling of the mold cavity, and, subsequently, the microstructures depend on the viscosity of the plastic melt, injection speed, molding block temperature, and nozzle temperature of the injection unit. This technique allows very rapid replication and high volume mass production. Typical cycle time is several seconds for most applications. However, due to high shear force on the mold master inside of the mold cavity during injection molding, metallic mold masters are highly recommended. Polymethylmethacrylate (PMMA), polyethylene (PE), polystyrene (PS), polycarbonate (PC), and cyclic olefin copolymers (COC) are common polymer/plastic materials for both hot embossing and injection mold replication.

All of these polymer/plastic replication techniques are summarized in Table 9.1. Since each technique differs from the others, fabrication techniques and materials have to be selected according to the application.

9.2 Active Microfluidic Devices

Microfluidics devices are essential for the development of lab-on-a-chip or µTAS (Micro-Total Analysis Systems). A number of different microfluidic devices have been developed with basic structures analogous to the macroscale fluidic devices. Such devices include microfluidic valves, microfluidic pumps, microfluidic mixers, etc. The devices listed above have been developed both as active and passive devices. While passive microfluidic devices are generally easy to fabricate, they do not offer the same functional diversity that the active microfluidic devices provide. For example, passive microvalves based on surface tension effect [9.19] can operate a few times to hold the liquid, which is acceptable for a disposable format. Once the air-liquid interface passes over the valve mechanism, the operation characteristics of the valve will differ, due to the change of surface energy over the channel. Similarly, passive check valves [9.20–22] are dependent on the pressure of the fluid for operation. Since the active microvalves can be triggered on/off depending on an external signal regardless of the fluid system status, there have been considerable research efforts to develop active microfluidic devices. However, the active devices usually require its expense as desired functional and fabrication complexity.

This section reviews some of the active microfluidic devices, such as microvalves, micropumps, and active microfluidic mixers. Passive counterparts of these microfluidic devices will be discussed in the next section.

9.2.1 Microvalves

Active microvalves have been an area of intense research over the past decade, and a number of novel design and actuation schemes have been developed. This makes the categorization of active microvalves a confusing enterprise. Classification schemes for active microvalves include:

1. Fluidics handled – liquid/gas/liquid and gas
2. Materials used for the structure – silicon/polysilicon/glass/polymer
3. Actuation mechanisms – electrostatic/pneumatic/thermopneumatic, etc.
4. Physical actuating microstructures – membrane type, flap type, ball valve, etc.

All of the classification schemes listed above are valid, but the most commonly used method [9.23, 24] is the classification based on the actuation mechanisms.

In this section, the various microvalves are discussed in terms of their actuation mechanisms and their relevance to the valve mechanism, as well as fluid handled and special design criteria.

Pneumatically/Thermopneumatically Actuated Microvalves

Pneumatic actuation uses an external air line (or pneumatic source) to actuate a flexible diaphragm. Pneumatic actuation offers such attractive features as high force, high displacement, and rapid response time. Figure 9.5 illustrates a schematic concept of pneumatically actuated microvalves. *Schomburg* et al. [9.25] demonstrated pneumatically actuated microvalves.

Pneumatically actuated microvalves have also been demonstrated using polymeric substrate. *Hosokawa* et al. [9.26] demonstrated a pneumatically actuated three-way microvalve system using a PDMS platform. The microfluidic lines and the pneumatic lines are fabricated on separate layers.

Thermopneumatic actuation is typically performed by heating a fluid (usually gas) in a confined cavity, as illustrated in Fig. 9.6. The increase in temperature leads to a rise in pressure of the gas, and this pressure is used for deflecting a membrane for valve operation. Thermopneumatic actuation is an inherently slow technique but offers very high forces when compared to other techniques [9.23]. Thermopneumatically actuated microvalves have been realized by many researchers using various substrates and diaphragm materials [9.27–29].

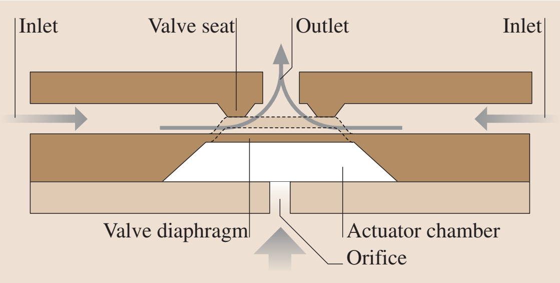

Fig. 9.5 Schematic concept of pneumatically actuated microvalve

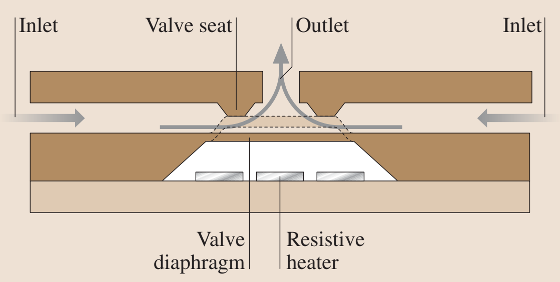

Fig. 9.6 Schematic concept of thermopneumatically actuated microvalve

Electrostatically Actuated Microvalves

Electrostatic actuation has been widely explored for a number of applications, including pressure sensors, comb drives, active mirror arrays, etc. Electrostatically actuated devices typically have a fairly simple structure and are easy to fabricate. A number of fabrication issues, such as stiction and release problems of membranes and valve flaps, need to be addressed for realizing practical electrostatic microvalves. *Sato* et al. [9.30] have developed a novel membrane design in which the deflection "propagates" through the membrane, rather than deforming it entirely. Figure 9.7 shows a schematic sketch of the actuation mechanism and valve design. The use of this "S-shaped" design allows them to have relatively large gaps across the two surfaces, as the electrostatic force must only be concentrated at the edges of the S-shape where the membrane is deflected. *Robertson* et al.

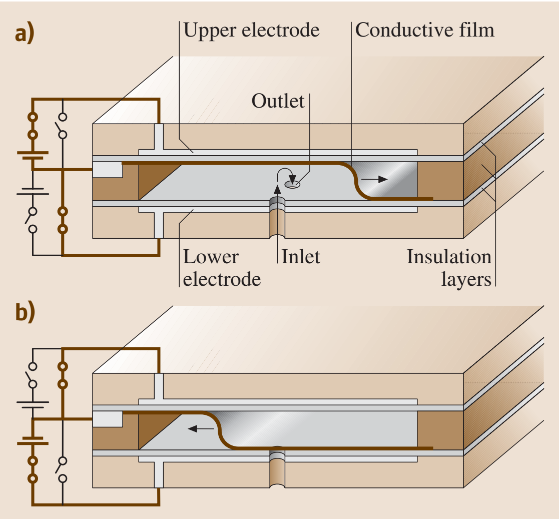

Fig. 9.7a,b Electrostatically actuated microvalve with S-Shaped film element: propagation of bend in the film as **(a)** open and **(b)** closed (adapted from [9.30]). Reproduced by permission of IOP Publishing Limited

[9.31] have developed an array of electrostatically actuated valves using a flap design (rather than a membrane) for sealing fluid flow. The demonstrated system is suitable for very low pressure gas control systems such as the ones needed in a clean room environment.

Wijngaart et al. [9.33] have developed a high-stroke, high pressure electrostatic actuator for valve applications. This reference provides a good overview of the theoretical design parameters used to design and analyze an electrostatically actuated microvalve.

Piezoelectrically Actuated Microvalve

Piezoelectric actuation schemes offer a significant advantage in terms of operating speed; they are typically the fastest actuation scheme at the expense of a reduced actuator stroke. Also, piezoelectric materials are more challenging to incorporate into fully integrated MEMS devices. *Watanabe* et al. [9.34] and *Stehr* et al. [9.35] demonstrated piezoelectric actuators for valve applications. A film of piezoelectric material is deposited on the movable membrane, and upon application of an electric potential, a small deformation occurs in the piezo film that is transmitted to the valve membrane.

Electromagnetically Actuated Microvalves

Electromagnetic actuators are typically capable of delivering high force and range of motion. A significant advantage of electromagnetic microvalves is that they are relatively insensitive to external interference. However, electromagnetic actuators involve a fairly complex fabrication process. Usually, a soft electromagnetic material like Ni-Fe (nickel-iron permalloy) is used as a membrane layer, and an external electromagnet is used to actuate this layer. *Sadler* et al. [9.32] have developed a microvalve using the electromagnetic actuation scheme shown in Fig. 9.8.

They have demonstrated a fully integrated magnetic actuator with magnetic interconnection vias to guide the magnetic flux. The valve seat design, also shown in Fig. 9.8, allows for very intimate contact between the Ni-Fe valve membrane and the valve seat, hence, achieving an ultralow leak rate when the valve is closed.

Jackson et al. [9.36] demonstrate an electromagnetic microvalve using "magnetic" PDMS as the membrane material. For this application, the PDMS pre-polymer is loaded with soft magnetic particles and then cured to form the valve membrane. The PDMS membrane

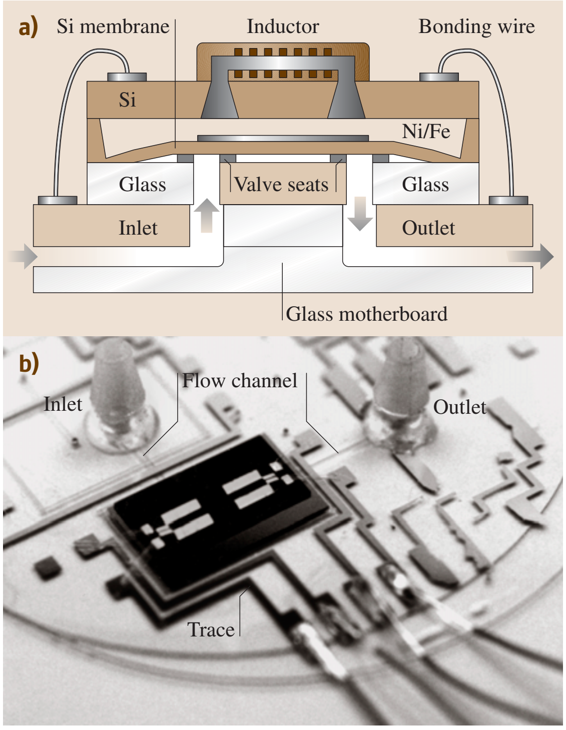

Fig. 9.8a,b Electromagnetically actuated microvalves: **(a)** schematic illustration and **(b)** photograph of the electromagnetically actuated microvalve as a part of lab-on-a-chip (after [9.32])

is then assembled over the valve body, and miniature electromagnets are used to actuate the membrane.

Other Microvalve Actuation Schemes

Microvalves have most commonly been implemented with one of the actuation schemes listed above. However, these are not the only actuation schemes that are used for microvalves. Some others include the use of SMA (shape memory alloys) [9.37], electrochemical actuation [9.38], etc. SMA actuation schemes offer the advantage of generating very large forces when the SMA material is heated to its original state. *Neagu* et al. [9.38] present an electrochemically actuated microvalve. In their device, an electrolysis reaction is used to generate oxygen in a confined chamber. This chamber is sealed by a deformable membrane that deflects due to increased pressure. The reported microvalve has relatively fast actuation time and can generate very high pressures. *Yoshida* et al. [9.39] present a novel approach to the microvalve design: a micro-electro-rheological valve (ER valve). An electro-rheological fluid is loaded into the microchannel, and, depending on the strength of an applied electric field, the viscosity of the ER fluid changes considerably. A higher viscosity is achieved when an electric field is applied perpendicular to the flow direction. This increased viscosity leads to a drop in the flow rate, allowing the ER fluid to act as a valve. Of course, this technique is limited to fluids that can exhibit such properties; nevertheless, it provides a novel idea to generate on-chip microvalves.

The ideal characteristics for a microvalve are listed by Kovacs [9.24]. However, of all the microvalves listed above, none can satisfy all the criteria. Thus, microvalve design, fabrication, and utility are highly application-specific and most microvalves try to generate the performance characteristics that are most useful for the intended application.

9.2.2 Micropumps

One of the most challenging tasks in developing a fully integrated microfluidic system has been the development of efficient and reliable micropumps. In macroscale, a number of pumping techniques exist such as peristaltic pumps, vacuum-driven pumps, venturi-effect pumps, etc. However, in microscale, most mechanical pumps rely on pressurizing the working fluid and forcing it to flow through the system. Practical vacuum pumps are not available in microscale. There are also some effects such as electro-osmotic pumping, that are only possible on the microscale. Consequently, electrokinetic driving mechanisms have also been widely studied for microfluidic pumping applications.

In the previous section, various microvalves that are used for microfluidic control were discussed. They were presented based on the actuation schemes that they employ. A similar classification can also be adopted for micropumps. However, rather than using the same classification scheme, the micropumps are categorized based on the type of microvalve mechanism used as part of the pumping mechanism. Broadly, the mechanical micropumps can be classified as check-valve controlled microvalves or diffuser pumps. Either mechanism can use various actuation schemes such as electrostatic, electromagnetic, piezoelectric, etc. Micropumps driven by direct electrical control form separate category and are discussed following the mechanical micropumps.

Micropumps Using Check Valve Design

Figure 9.9 shows a typical mechanical micropump with check valves. This pump consists of an inlet and an outlet check valve with a pumping chamber in between. A membrane is deflected upwards, and a low pressure zone is created in the pumping chamber. This forces the inlet check valve open, and fluid is sucked into the pumping chamber. As the membrane returns to its original state and continues to travel downwards, a positive pressure builds up, which seals the inlet valve while simultaneously opening the outlet valve. The fluid is then ejected and the pump is ready for another cycle.

A number of other techniques have been used to realize mechanical micropumps with check valves and a pumping chamber. *Jeon* et al. [9.40] present a micropump that uses PDMS flap valves to control the pumping mechanism, *Koch* et al. [9.41], *Cao* et al. [9.42], *Park* et al. [9.43], and *Koch* et al. [9.44] present piezoelectrically driven micropumps. *Xu* et al. [9.45] and *Makino*

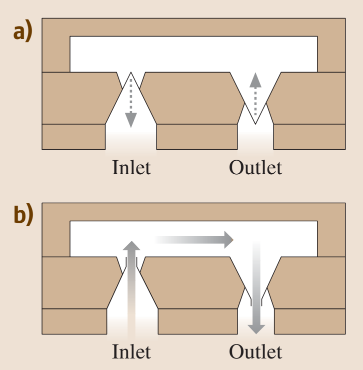

Fig. 9.9a,b Operation of micropump using check valve design: **(a)** check valves are closed and **(b)** check valves are open

et al. [9.46] present SMA driven micropumps. *Chou* et al. [9.47] present a novel rotary pump using a soft lithography approach. As explained in the active valve section, valves can be created on a PDMS layer using separate liquid and air layers. When a pressure is applied to the air lines, the membranes deflect to seal the fluidic path. *Chou* et al. [9.47] have implemented a series of such valves in a loop. When they are deflected in a set sequence, the liquid within the ring is pumped by peristaltic motion.

An interesting approach toward the development of a bidirectional micropump has been used by *Zengerle* et al. [9.48]. Their design has two flap valves at the inlet and outlet of the micropump, which work in the forward mode at low actuation frequencies and in the reverse direction at higher frequencies. *Zengerle* et al. [9.48] attribute the change in pumping direction to the phase shift between the response of the valves and the pressure difference that drives the fluid. *Carrozza* et al. [9.49] use a different approach to generate the check valves. Rather than using the conventional membrane or flap type valves, they use ball valves by employing a stereolithographic approach. The developed pump is actuated using a piezoelectric actuation scheme.

Diffuser Micropumps

The use of nozzle-diffuser sections, or pumps with fixed valves, or pumps with dynamic valves has been extensively researched. The basic principle of these pumps is based on the idea that the geometrical structure used as an inlet valve has a preferential flow direction toward the pumping chamber, and the outlet valve structure has a preferential flow direction away from the pumping chamber. An illustrative example of this concept is shown in Fig. 9.10. As can be readily seen, these pumps are designed to work with liquids only. When the pump is in suction mode, the flow in the inlet diffuser structure is primarily directed toward the pump, and a slight backflow occurs from the outlet diffuser (acting as a nozzle) section. When the pump is in pressure mode, the outlet diffuser section allows most of the flow out of the pump, whereas the inlet diffuser section allows a slight backflow. The net effect is that liquid pumping occurs from left to right.

Diffuser micropumps are simple and valveless structures that improve pumping reliability [9.50], but they cannot eliminate backflow problems.

These micropumps can also be described as flow rectifiers, analogous to diodes in electrical systems. *Forster* et al. [9.51] have used the Tesla valve geometry, instead of the diffuser section, and the reference presents a detailed discussion of the design parameters and operational characteristics of their fixed-valve micropumps.

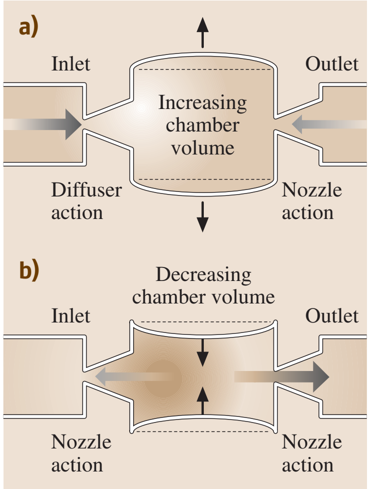

Fig. 9.10a,b Operating principle of a diffuser micropump: **(a)** suction mode and **(b)** pumping mode

Electric/Magnetic Field Driven Micropumps

An electric field can be used to directly pump liquids in microchannels using such techniques as electro-omosis (EO), electrohydrodynamic (EHD) pumping, magnetohydrodynamic (MHD) pumping, etc. These pumping techniques rely on creating an attractive force for some of the ions in the liquid, and the remaining liquid is dragged along to form a bulk flow.

When a liquid is introduced into a microchannel, a double-layer charge exists at the interface of the liquid and the microchannel wall. The magnitude of this charge is governed by the zeta potential of the channel-liquid pair. Figure 9.11 shows a schematic view of the electro-osmotic transport phenomenon.

As shown in Fig. 9.11, the channel wall is negatively charged, which attracts the positive ions in solution. When a strong electric field is applied along the length of the microchannel, the ions at the interface experience an attractive force toward the cathode. As the positive ions move toward the cathode, they exert a drag force on the bulk fluid, and net fluid transport occurs from the anode to the cathode. It is interesting to note that the flow profile of the liquid plug is significantly different from the pressure-driven flow. Unlike the parabolic flow profile of pressure-driven flow, electro-osmotic transport leads to an almost vertical flow profile. The electro-osmotic transport phenomenon is only effective across very narrow channels.

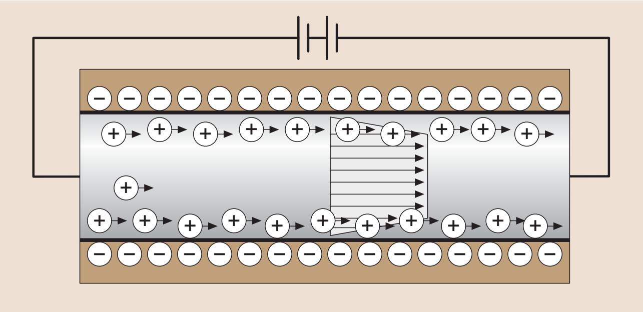

Fig. 9.11 Schematic sketch explaining the principle of electro-osmotic fluid transport

EHD can be broadly broken down into two subcategories: injection type and induction type. In injection type EHD, a strong electric field ($\sim 100\,\mathrm{kV/cm}$) is applied across a dielectric liquid. This induces charge formation in the liquid, and these induced (or injected) charges are then acted upon by the electrical field for pumping. Induction type EHD relies on generating a gradient/discontinuity in the conductivity and/or permitivity of the liquid. *Fuhr* et al. [9.52] explain various techniques that can be used to generate gradients for non-injection type EHD pumping.

MHD pumps rely on creating a Lorentz force on the liquid particles in the presence of an externally applied electric field. MHD has been demonstrated using both DC [9.53] and AC [9.54] excitations. MHD, EHD, and electro-osmotic transport share one feature in common that makes them very appealing for microfluidic systems; namely, none requires microvalves to regulate the flow. This makes these pumping techniques very reliable, as there is no concern of wear and tear of the microvalves, or any other moving parts of the micropump. However, it has been difficult to implement these actuation schemes on a complete microscale, owing to the high voltages/electromagnets, etc. required for these actuation schemes.

9.3 Smart Passive Microfluidic Devices

Passive microfluidics is a powerful technique for the rapidly evolving discipline of BioMEMS. It is a fluid control topology in which the physical configuration of the microfabricated system primarily determines the functional characteristics of the device/system. Typically, passive microfluidic devices do not require an external power source, and the control exerted by the devices is based, in part, on energy drawn from the working fluid, or based purely on surface effects, such as surface tension, selective hydrophobic/hydrophilic control, etc. Most passive microfluidic devices exploit various physical properties such as shape, contact angle, and flow characteristics to achieve the desired function. Passive microfluidic systems are usually easier to implement and allow for a simple microfluidic system with little or no control circuitry. A further list of advantages and disadvantages of passive microfluidic systems (or devices) is considered toward the end of this section.

Passive microfluidic devices can be categorized based on:

- Function – microvalves, micromixers, filters, reactors, etc.
- Fluidic medium – gas or liquid
- Application – biological, chemical, or other
- Substrate material – silicon, glass, polysilicon, polymer, or others

In this section, we will study various passive microfluidic devices that are categorized based on their function. Passive microfluidic devices include, but are not limited to, microvalves, micromixers, filters, dispensers, etc. [9.24].

9.3.1 Passive Microvalves

Passive microvalves have been a subject of great interest ever since the inception of the lab-on-a-chip concept. Microvalves are a key component of any microfluidic system and are essential for fluidic sequencing operations. Since most chemical and biochemical reactions require about five to six reaction steps, passive microvalves with limited functionality are ideally suited for such simple tasks. Passive microvalves can be broadly categorized as follows:

- Silicon/polysilicon or polymer-based check valves
- Passive valves based on surface tension effects
- Hydrogel-based biomimetic valves

Passive Check Valves

Shoji et al. [9.23] provide an excellent review of check-type passive microvalves. Some of the valves, shown in Fig. 9.12, illustrate the various techniques that can be used to fabricate check valves.

Figure 9.12a shows a microvalve fabricated using silicon bulk etching techniques. A through hole (pyramidal cavity) is etched through a silicon wafer that is sandwiched between two glass wafers. The normally closed valve is held in position by the spring effect of the silicon membrane. Upon applying pressure at the lower fluidic port, the membrane deflects upwards, allowing fluid flow through the check valve. The same working principle is employed by the microvalve shown in Fig. 9.12b. However, instead of using a membrane supported on all sides, a cantilever structure is used for the flap. This reduces the burst pressure, i.e., the minimum pressure required to open the microvalve. Figure 9.12c shows that the membrane structure can also be realized using a polysilicon layer deposited on a bulk-etched silicon wafer. Poly-silicon processes typically allow tighter control over dimensions and, consequently, offer more repeatable operating characteristics. Figure 9.12d shows the simplest type of check valve where the V-groove etched in a bulk silicon substrate acts as a check valve. However, the low contact area between the flaps of the microvalve leads to nontrivial leakage rates in the forward direction. Figures 9.12e and f show check valve designs that are realized using polymer/metal films, in addition to the traditional glass/silicon platform. This technique offers a significant advantage in terms of biocompatibility characteristic and controllable operating characteristics. The surface properties of polymers can be easily tailored using a wide variety of techniques such as plasma treatment and surface adsorption [9.55]. Thus, in applications for which the biocompatibility requirements are very stringent, it is preferable to have polymers as the fluid contacting material. Furthermore, polymer properties such as stiffness, can also be controlled in some cases based on the composition and/or processing conditions. Thus, it may be possible to fabricate microvalves with different burst pressures by using different processing conditions for the same polymer.

In addition to the passive check valves reviewed by *Shoji* et al. [9.23], other designs include check valves using composite titanium/polyimide membranes [9.20,21], polymeric membranes such as Mylar or KAPTON [9.56], or PDMS [9.40], and metallic membranes such as [9.22].

Terray et al. [9.57] present an interesting approach to fabricating ultra-small passive valve structures. They have demonstrated a technique to polymerize colloidal particles into linear structures using an optical trap to form microscale particulate valves.

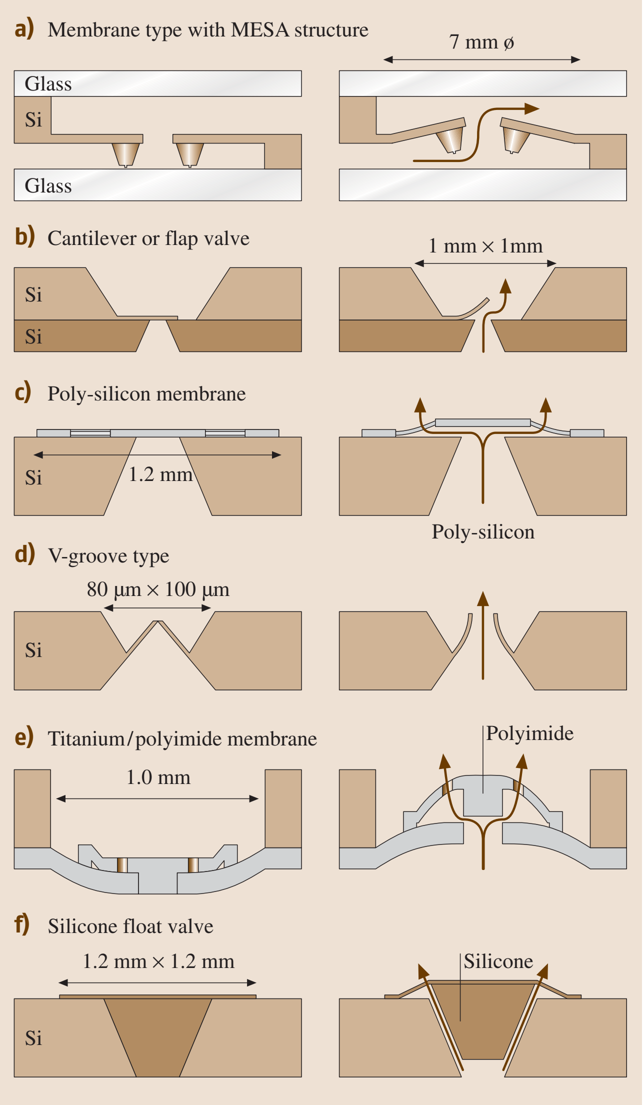

Fig. 9.12a–f Various types of passive microvalve designs: **(a)** membrane type with MESA structure; **(b)** cantilever or flap valve; **(c)** polysilicon membrane; **(d)** V-groove type; **(e)** titanium/polyimide membrane; and **(f)** silicone float valve. Adapted from [9.23]. Reproduced by permissions of IOP Publishing Limited

Passive Valves Based on Surface Tension Effects

The passive valves listed in the previous section use the forced motion of the membrane or flap to control the flow of fluids. These valves are prone to such problems as clogging and mechanical wear and tear. Passive valves based on surface tension effects, on the other hand, have no moving parts and control the fluid motion based on their physical structure and the surface property of the substrate. Figure 9.13 shows a schematic sketch of a passive microvalve on a hydrophobic substrate [9.19].

The Hagen–Poiseuille equation for laminar flow governs the pressure drops in microfluidic systems with laminar flow. For a rectangular channel with a high width to height ratio, the pressure drop is governed by the equation

$$\Delta P = \frac{12L\mu.Q}{wh^3}\,, \tag{9.1}$$

where L is the length of the microchannel, μ is the dynamic viscosity of the fluid, Q is the flow rate, and w and h are the width and height of the microchannel, respectively. Varying L or Q can control the pressure drop for a given set of w and h.

An abrupt change in the width of the channel causes a pressure drop at the point of restriction. For a hydrophobic channel material, an abrupt decrease in channel width causes a positive pressure drop:

$$\Delta P_2 = 2\sigma_1 \cos(\theta_c)\left[\left(\frac{1}{w_1}+\frac{1}{h_1}\right)-\left(\frac{1}{w_2}+\frac{1}{h_2}\right)\right]\,, \tag{9.2}$$

where σ_1 is the surface tension of the liquid, θ_c is the contact angle, and w_1, h_1 and w_2, h_2 are the width and height of the two sections before and during the restriction, respectively. Setting h as constant through the system, ΔP_2 can be varied by adjusting the ratio of w_1 and w_2. *Ahn* et al. [9.19] have proposed and implemented a novel structurally programmable microfluidic system (sPROMs) based on the passive microfluidic approach. In short, the sPROMs system consists of a network of microchannels with passive valves of the type shown in the passive valve section. If the pressure drop of the passive valves is set to be significantly higher than the pressure drop of the microchannel network, then the position of the liquid in the microchannel network can be controlled accurately. By applying sequentially higher pressure pulses, the liquid is forced to move from one passive valve to another. Thus, the movement of the fluid within the microfluidic channels is "programmed" using the physical structure of the microfluidic system, and this forms the basic idea behind sPROMs.

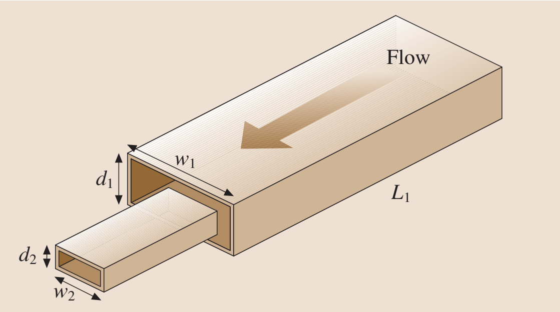

Fig. 9.13 Structure of a passive microvalve based on surface tension effects (adapted from [9.19])

The abrupt transition from a wide channel to a narrow channel can also be affected along the height of the microchannel. Furthermore, the passive valve geometry shown in Fig. 9.13 is not exclusive. *Puntambekar* et al. [9.58] have demonstrated different geometries of passive valves, as shown in Fig. 9.14a. As shown in Fig. 9.14b,

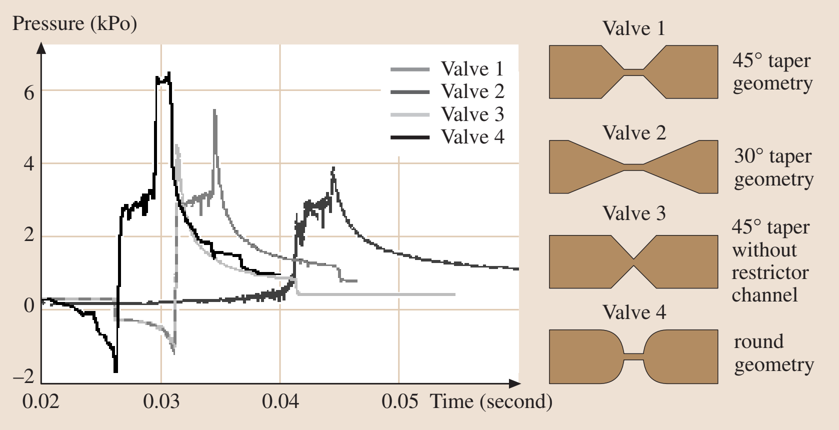

Fig. 9.14 Analysis of different geometries of passive valves. After [9.58], reproduced by permission of The Royal Society of Chemistry

the various geometries of the passive valves in Fig. 9.14a can act as effective passive valves without having an abrupt transition. This is important in order to avoid the dead volume that is commonly encountered across an abrupt step junction.

The use of surface tension to control the operation of passive valves is not limited to hydrophobic substrates. *Madou* et al. [9.59] demonstrate a capillarity-driven stop valve on a hydrophilic substrate. On a hydrophilic substrate, the fluid can easily "wick through" in the narrow region. However, at the abrupt transition to a larger channel section, the surface tension effects will not allow the fluid to leave the narrow channel. Thus, in this case, the fluid is held at the transition from the narrow capillary to the wide outlet channel.

Another mechanism to implement passive valves is the use of hydrophobic patches on a normally hydrophilic channel. *Handique* et al. [9.60] demonstrate the use of this technique to implement passive valves for a DNA analysis system. The fluid is sucked into the microfluidic channels via capillary suction force. The hydrophobic patch exerts a negative capillary pressure that stops the further flow of fluid. The use of hydrophobic patches as passive valves is reported by *Andersson* et al. [9.61].

Other Passive Microvalves

A novel approach to realizing passive valves that are responsive to their surrounding environment is demonstrated by *Low* et al. [9.62] and *Yu* et al. [9.63]. *Yu* et al. [9.63] have developed customized polymer "cocktails" that are polymerized in situ around prefabricated posts. The specialized polymer is selectively responsive to stimuli such as pH, temperature, electric fields, light, carbohydrates, and antigens.

Another interesting approach in developing passive valves has been adopted by *Foster* et al. [9.51]. They have developed the so-called no moving part (NMP) valves, which are based on a physical configuration that allows a higher flow rate along a direction compared to the reverse direction.

9.3.2 Passive Micromixers

The successful implementation of microfluidic systems for many lab-on-a-chip systems is partly owing to the significant reduction in volumes handled by such systems. This reduction in volume is made possible by the use of microfabricated features and channel dimensions ranging from a few µm to several hundred µm. Despite the advantage offered by the micron-sized channels, one of the significant challenges has been the implementation of effective microfluidic mixers on the microscale. Mixing on a macroscale is a turbulent flow regime process. However, on the microscale, because of the low Reynolds numbers as a result of the small channel dimensions, most flow streams are laminar in nature, which does not allow for efficient mixing. On the other hand, diffusion is an important factor in mixing because of the short diffusion lengths.

There have been numerous attempts to realize micromixers using both active and passive techniques. Active micromixers rely on creating localized turbulence to enhance the mixing process, whereas passive micromixers usually enhance the diffusion process. The diffusion process can be modeled by the following equation:

$$\tau = \frac{d^2}{D}\,, \tag{9.3}$$

where τ is the mixing time, d is the distance, and D is the diffusion coefficient. Equation (9.3) illustrates the diffusion-dominated mixing at the microscale. Because of the small diffusion lengths (d), the mixing times can be made very short. The simplest category of micromixers is illustrated in Fig. 9.15. In these mixers, creating a convoluted path increases the path length that the two fluids share, leading to higher diffusion and more complete mixing. However, these mixers only exhibit good mixing performance at low flow rates, in the range of a few µl/min.

Mitchell et al. [9.64] have demonstrated three-dimensional micromixers that can achieve better

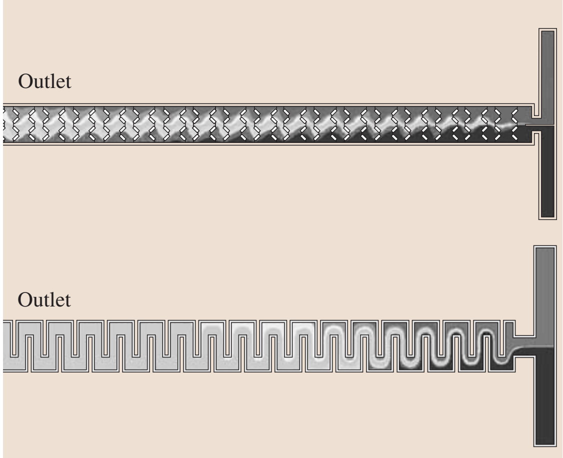

Fig. 9.15 Diffusion-enhanced mixers based on long, convoluted flow path

performance by alternately laminating the two fluid streams to be mixed. *Beebe* et al. [9.65] have created a chaotic mixer that has the convoluted channel along three dimensions. This micromixer works on the principle of forced advection resulting from repeated turns in three dimensions. Furthermore, at each turn, eddies are generated because of the difference in flow velocities along the inner and outer radii, which enhances the mixing.

Stroock et al. [9.68] demonstrate a passive micromixer that uses chaotic mixing by superimposing a transverse flow component on the axial flow. Ridges are fabricated at the bottom of microchannels. The flow resistance is lower along the ridges (peak/valley) and higher in the axial direction. This generates a helical flow pattern that is superimposed on the laminar flow. The demonstrated mixer shows good mixing performance over a wide range of flow velocities.

Hong et al. [9.66] have demonstrated a passive micromixer based on the Koanda effect. Their design uses the effects of diffusion mixing at low flow velocities; at high flow velocities, a convective component is added perpendicular to the flow direction, allowing for rapid mixing. This mixer shows excellent mixing performances across a wide range of flow rates because of the dual mixing effects. Figure 9.16 shows a schematic sketch of the mixer structure.

The mixer works on the principle of superimposing a parabolic flow profile in a direction perpendicular to the flow direction. The parabolic profile creates a Taylor dispersion pattern across the cross section of the flow path. The dispersion is directly proportional to the flow velocity, and higher flow rates generate more dispersion mixing.

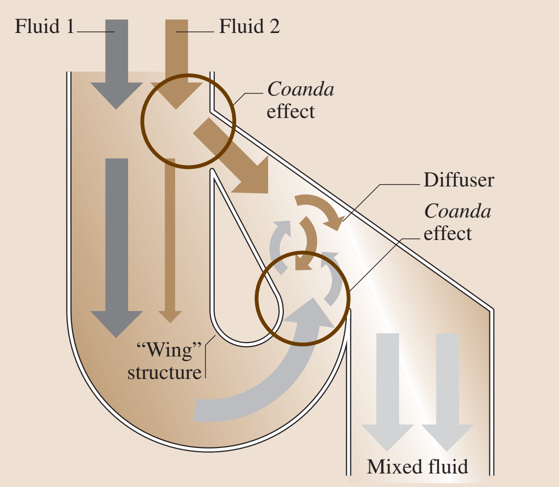

Fig. 9.16 Mixing unit design for the Coanda effect mixer. The actual mixer has mixing unit pairs in series (after [9.66])

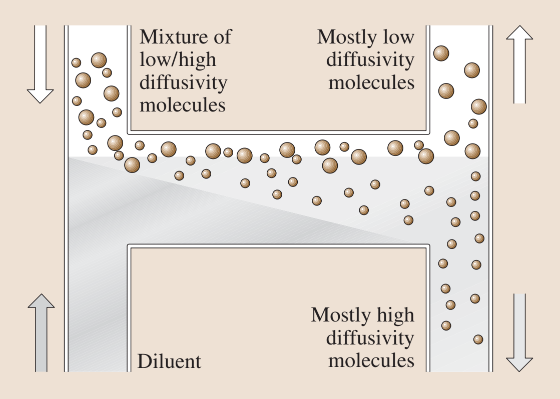
Mixture of low/high diffusivity molecules
Mostly low diffusivity molecules
Diluent
Mostly high diffusivity molecules

Fig. 9.17 Conceptual illustration of the H-Sensor. Adapted from [9.67]

Brody et al. [9.67] have used the laminar flow characteristics in a microchannel to develop a diffusion-based extractor. When two fluid streams, where fluid 1 is loaded with particles of different diffusivity and fluid 2 is a diluent, are forced to flow together in a microchannel, they form two laminar streams with little mixing. If the length that the two streams are in contact is carefully adjusted, only particles with high diffusivity (usually small molecules) can diffuse across into the diluent stream, as shown in Fig. 9.17. The same idea can be extended to a T-filter. *Weigl* et al. [9.69] have demonstrated a rapid diffusion immunoassay using the T-filter.

9.3.3 Passive Microdispensers

The principle of the structurally programmable microfluidic systems (sPROMs) was introduced earlier in the passive valve section. One of the key components of the sPROMs system is the microdispenser, which is designed to accurately and repeatedly dispense fluidic volumes in the micro- to nanoliter range. This would allow the dispensing of a controlled amount of the analyte into the system that could be used for further biochemical analysis. Figure 9.18 shows a schematic sketch of the microdispenser design [9.58].

The microdispenser works on the principle of graduated volume measurement. The fluid fills up the exact fixed volume of the reservoir, and the second passive valve at the other end of the microdispenser stops further

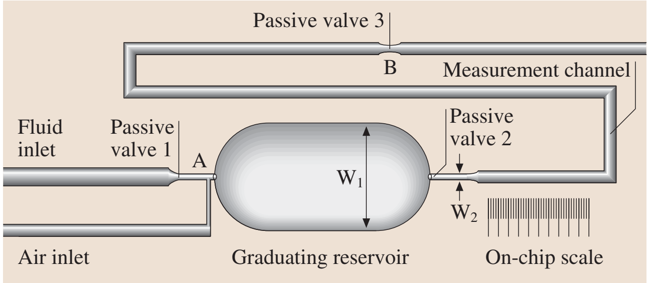

Fig. 9.18 Schematic sketch of the microdispenser. Adapted from [9.58], reproduced by permission of The Royal Society of Chemistry

motion. When the reservoir is filled with fluid, the fluidic actuation is stopped, and, simultaneously, pneumatic actuation from the air line causes a split in the fluid column at point A (Fig. 9.18). Thus, the accuracy of the reservoir decides the accuracy of the dispensed volume. Since the device is manufactured using UV-LIGA lithography techniques, highly accurate and reproducible volumes can be defined. The precisely measured volume of fluid is expelled to the right from the reservoir. The expelled fluid then starts to fill up the measuring channel. When the fluid reaches point B (Fig. 9.18), the third microvalve holds the fluid column.

Figure 9.19 shows an actual operation sequence of the microdispenser. Figure 9.19f shows that the dispensed volume is held by passive valve 3, and at this stage, the length (and hence volume) can be calculated using the on-chip scale. In experiments only, the region in the immediate vicinity of the scale was viewed using a stereomicroscope for measuring the length of the fluid column. The microdispenser demonstrated above is reported to have dispensing variation of less than 1% between multiple dispensing cycles.

9.3.4 Microfluidic Multiplexer Integrated with Passive Microdispenser

Ahn et al. [9.19] have demonstrated the sPROMs technology to be an innovative method of controlling liquid

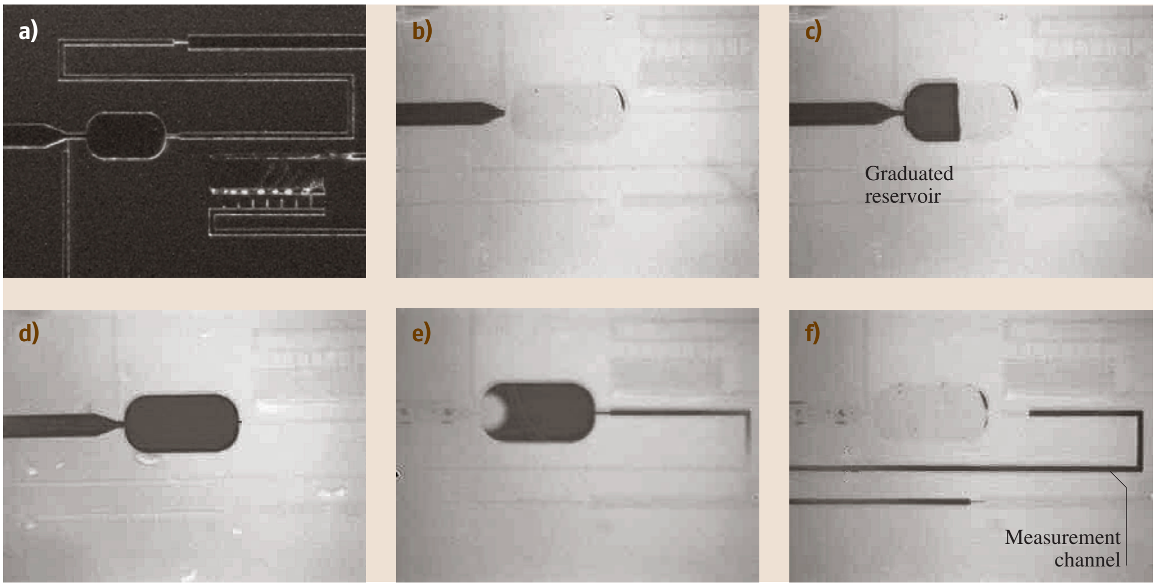

Fig. 9.19a–f Microphotographs of microdispenser sequence: **(a)** fabricated device; **(b)** fluid at reservoir inlet; **(c)** reservoir filling; **(d)** reservoir filled; **(e)** split in liquid column due to pneumatic actuation; and **(f)** fluid ejected to measurement channel and locked in by passive valve. After [9.58], reproduced by permission of The Royal Society of Chemistry

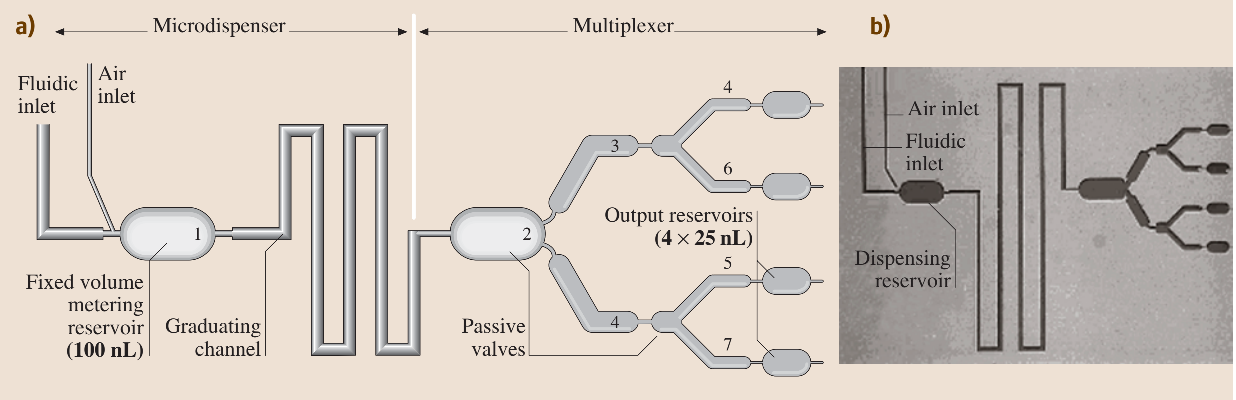

Fig. 9.20a,b Microfluidic multiplexer with integrated dispenser: **(a)** schematic sketch and **(b)** fabricated device filled with dye (after [9.58])

movement in a programmed fashion in a microfluidic network. By integrating this technique with the microdispensers, a more functionally useful microfluidic system can be realized. Figure 9.20a shows a schematic illustration of the microfluidic multiplexer with the integrated dispenser, and Fig. 9.20b shows an actual device fabricated using rapid prototyping techniques [9.70].

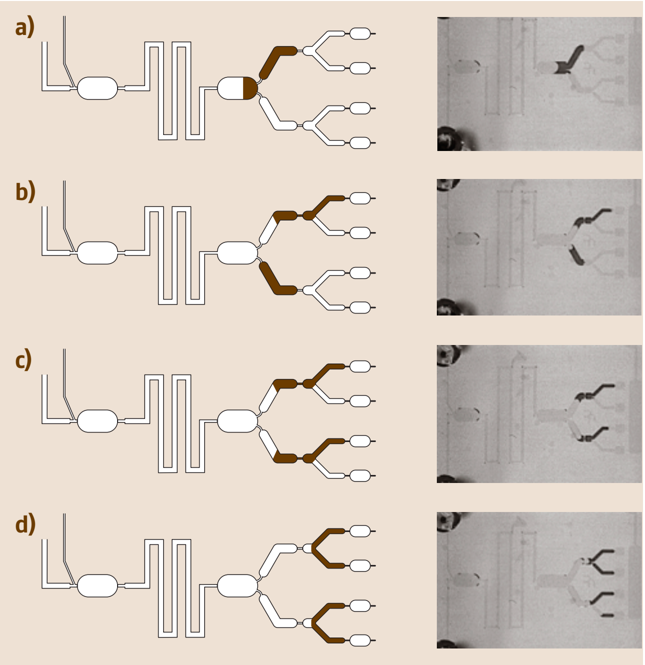

Fig. 9.21a–d Microphotographs showing operation of sequential multiplexer: **(a)** first level division; **(b)** second level division; **(c)** continued multiplexing; and **(d)** end of sequential multiplexing sequence (after [9.58])

The operation of the microdispenser has been explained earlier in this section. Briefly, the fluid is loaded in the fixed-volume metering reservoir via a syringe pump. The fluid is locked in the reservoir by the passive valve at the outlet of the reservoir. When a higher pressure is applied via the air inlet line, the liquid column is split and the fluid is dispensed into the graduating channel.

The microfluidic multiplexer is designed to have a programmed delivery sequence, as shown in Fig. 9.20a, where the numbers on each branch of the multiplexer indicates the filling sequence. This sequential filling is achieved by using different ratios of passive valves along the multiplexer section. For instance, at the first branching point, the passive valve at the upper branch offers less resistance than the passive valve at the beginning of the lower branch. Thus, the dispensed fluid will first fill the top branch. After filling the top branch, the liquid encounters another passive valve at the end of the top branch. The pressure needed to push beyond this valve is higher than the pressure needed to push liquid into the lower branch of the first split. The liquid will then fill the lower branch. This sequence of nonsymmetrical passive valves continues along all the branches of the multiplexer, as shown in Fig. 9.21.

The ability to sequentially divide and deliver liquid volumes was demonstrated for the first time using a passive microfluidic system. This approach has the potential to deliver very simple microfluidic control systems that are capable of a number of sequential microfluidic manipulation steps required in a biochemical analysis system.

9.3.5 Passive Micropumps

A passive system is defined inherently as one that does not require an external energy source. Thus, the term passive micropump might seem misnomer upon initial inspection. However, there have been some efforts dedicated to realizing a passive micropump that essentially does not draw energy from an external source, but stores the required actuation energy in some form and converts it to mechanical energy on demand.

Passive Micropump Based on Osmotic Pressure

Nagakura et al. [9.71] have demonstrated a mesoscale osmotic actuator that converts chemical energy to mechanical displacement. Osmosis is a well-known phenomenon by which liquid is transported across a semipermeable membrane to achieve a uniform concentration distribution across the membrane. If the membrane is flexible, such as the one used by *Nagakura* et al. [9.71], then the transfer of liquid would cause the membrane to deform and act as an actuator. The inherent drawback of using osmosis as an actuation mechanism is that it is a very slow process: Typical response times (on a macroscale) are on the order of several hours. However, osmotic transport scales favorably to the microscale, and it is expected that these devices will have response times on the order of several minutes, rather than hours. Based on this idea, *Nagakura* et al. [9.71] are developing a miniature insulin pump. *Su* et al. [9.72] have demonstrated a microscale osmotic actuator that is capable of developing pressures as high as 35 MPa. This is still a relatively unexplored realm in BioMEMS actuation, and it has good potential for applications such as sustained drug delivery.

Passive Pumping Based on Surface Tension

Glenn et al. have demonstrated pumping action using the difference between the surface tension pressure at the inlet and outlet of a microfluidic channel. In the simplest case, a small drop of a fluid is placed at one end of a straight microchannel, and a much larger drop of fluid is placed at the opposite end of the microchannel. The pressure within the small drop is significantly higher than the pressure within the large drop, due the difference in the surface tension effects across the two drops. Consequently, the liquid will flow from the small drop and add to the larger drop. The flow rate can be varied by changing various parameters such as the volume of the pumping drop, the surface free energy of the liquid, or the resistance of the microchannel, etc. This pumping scheme is very easy to realize and can be used for a wide variety of fluids.

Evaporation-Based Continuous Micropump

Effenhauser et al. [9.73] have demonstrated a continuous flow micropump based on a controlled evaporation approach. Their concept is based on the controlled evaporation of a liquid through a membrane into a gas reservoir. The reservoir contains a suitable adsorption agent that draws out the liquid vapors and maintains a low vapor pressure conducive to further evaporation. If the liquid being pumped is replenished from a reservoir, capillary forces will ensure that the fluid is continuously pumped through the microchannels as it evaporates on the other end into the adsorption reservoir. Though the pump suffers from inherent disadvantages such as strong temperature dependence and operation only in suction mode, it offers a very simple technique for fluidic transport.

9.3.6 Advantages and Disadvantages of the Passive Microfluidic Approach

This chapter has covered a number of different passive microfluidic devices and systems. Passive microfluidic devices have only recently been a subject of considerable research effort. One of the reasons for this interest is the long list of advantages that passive microfluidic devices offer. However, since most microfluidic devices are very application-specific (and even more so for passive microfluidic devices), the advantages are not to be considered universally applicable for all the devices/systems. Some of the advantages that are commonly found are:

- avoid the need for an "active" control system;
- usually very easy to fabricate;
- the passive microfluidic systems with no moving parts are inherently more reliable because of the lack of mechanical wear and tear;
- offer very repeatable performance once the underlying phenomena are well understood and characterized;
- highly suited for BioMEMS applications – can easily handle a limited number of microfluidic manipulation sequences;
- well suited for low cost, mass production approach;

- low cost offers the possibility of having disposable microfluidic systems for specific applications, such as working with blood;
- can offer other interesting possibilities, such as biomimetic response.

However, like all MEMS devices, passive microfluidic devices or systems are not the solution to the microfluidic handling problem. Usually they are very application-specific – they cannot be reconfigured for another application easily. Other disadvantages are listed below:

- suited for well understood, niche applications for which the fluidic sequencing steps are well decided,
- strongly dependent on variances in the fabrication process,
- usually not very suitable for a wide range of fluidic mediums.

9.4 Lab-on-a-Chip for Biochemical Analysis

Recent development in MEMS (microelectromechanical systems) has brought a new and revolutionary tool in biological or chemical applications: “lab-on-a-chip”. New terminology, such as micro total analysis systems and lab-on-a-chip, was introduced in the last decade, and several prototype systems have been reported.

The idea of lab-on-a-chip is basically to reduce biological or chemical laboratories to a microscale system, hand-held size or smaller. Lab-on-a-chip systems can be made out of silicon, glass, and polymeric materials, and the typical microfluidic channel dimensions are in the range of several tens to hundreds of μm. Liquid samples or reagents can be transported through the microchannels from reservoirs to reactors using electrokinetic, magnetic, or hydrodynamic pumping methods. Fluidic motions or biochemical reactions can also be monitored using various sensors, which are often used for biochemical detection of products.

There are many advantages to using lab-on-a-chip over conventional chemical or biological laboratories. One of the important advantages lies in its low cost. Many reagents and chemicals used in biological and chemical reactions are expensive, so the prospect of using very small amounts (in micro- to nanoliter range) of reagents and chemicals for an application is very appealing. Another advantage is that lab-on-a-chip requires very small amounts of reagents/chemicals (which enables rapid mixing and reaction) because biochemical reaction is mainly involved in the diffusion of two chemical or biological reagents, and microscale fluidics reduces diffusion time as it increases reaction probabilities. In practical terms, reaction products can be produced in a matter of seconds/minutes, whereas laboratory scale can take hours, or even days. In addition, lab-on-a-chip systems minimize harmful by-products since their volume is so small. Complex reactions of many reagents could happen on lab-on-a-chip that have ultimate potential in DNA analysis, biochemical warfare agent detection, biological cell/molecule sorting, blood analysis, drug screening/development, combinatorial chemistry, and protein analysis. In this section, two recent developments of microfluidic systems for lab-on-a-chip applications will be introduced: (a) magnetic micro/nano bead-based biochemical detection system and (b) disposable smart lab-on-a-chip for blood analysis.

9.4.1 Magnetic Micro/Nano Bead-Based Biochemical Detection System

In the past few years, a large number of microfluidic prototype devices and systems have been developed, specifically for biochemical warfare detection systems and portable diagnostic applications. The BioMEMS team at the University of Cincinnati has been working on the development of a remotely accessible generic microfluidic system for biochemical detection and biomedical analysis, based on the concepts of surface-mountable microfluidic motherboards, sandwich immunoassays, and electrochemical detection techniques [9.74, 75]. The limited goal of this work is to develop a generic MEMS-based microfluidic system and to apply the fluidic system to detect bio-molecules, such as specific proteins and/or antigens, in liquid samples. Figure 9.22 illustrates the schematic diagram of a generic microfluidic system for biochemical detection using a magnetic bead approach for both sampling and manipulating the target bio-molecules [9.76, 77].

The analytical concept is based on sandwich immunoassay and electrochemical detection [9.78], as illustrated in Fig. 9.23. Magnetic beads are used as both substrates of antibodies and carriers of target antigens.

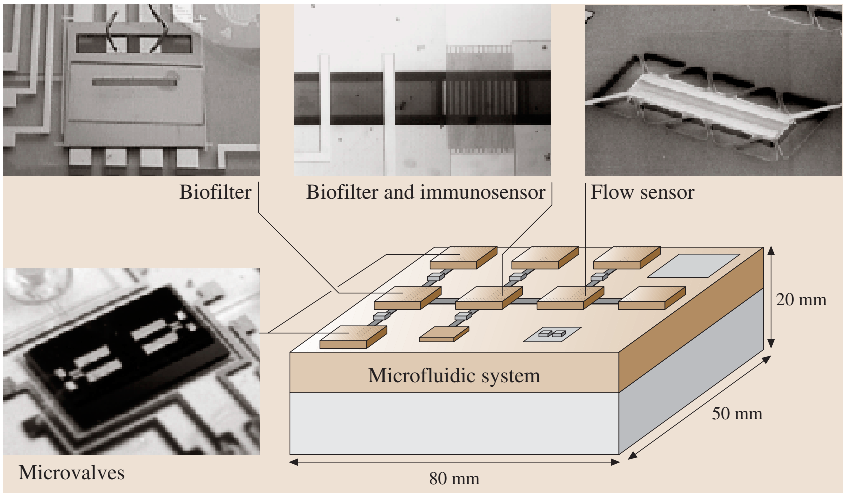

Fig. 9.22 Schematic diagram of a generic microfluidic system for biochemical detection (after [9.75])

A simple concept of magnetic bead-based bio-sampling with electromagnet for the case of sandwich immunoassay is shown in Fig. 9.24.

Antibody-coated beads are introduced on the electromagnet and separated by applying magnetic fields. While holding the antibody-coated beads, antigens are injected into the channel. Only target antigens are immobilized and, thus, separated on the magnetic bead surface due to antibody/antigen reaction. Other antigens get washed out with the flow. Next, enzyme-labeled secondary antibodies are introduced and incubated, along with the immobilized antigens. The chamber is then rinsed to remove all unbound secondary antibodies. Substrate solution, which will react with enzyme, is injected into the channel, and the electrochemical detection is performed. Finally, the magnetic beads are released to the waste chamber, and the bio-separator is ready for another immunoassay. Alkaline phosphatase (AP) and p-aminophenyl phosphate (PAPP) were chosen as enzyme and electrochemical substrate, respectively. Alkaline phosphatase makes PAPP turn into its electrochemical product, p-aminophenol (PAP). By applying potential, PAP gives electrons and turns into 4-quinoneimine (4QI), which is the oxidant form of PAP.

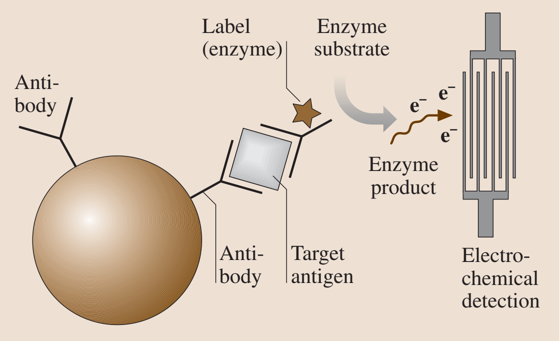

Fig. 9.23 Analytical concept based on sandwich immunoassay and electrochemical detection. After [9.77], reproduced by permission of The Royal Society of Chemistry

For the successful immunoassay, the biofilter [9.76] and the immunosensor were fabricated separately and integrated together. The integrated biofilter and immunosensor were surface-mounted using a fluoropolymer bonding technique [9.79] on a microfluidic motherboard, which contains microchannels fabricated by glass etching and glass-to-glass direct bonding technique. Each inlet and outlet were connected to sample reservoirs through custom-designed microvalves. Figure 9.25 shows the integrated microfluidic biochemical detection system for magnetic bead-based immunoassay.

After a fluidic sequencing test, full immunoassays were performed in the integrated microfluidic system to prove magnetic bead-based biochemical detection and sampling function. Magnetic beads (Dynabeads® M-280, Dynal Biotech Inc.) coated with biotinylated sheep anti-mouse immunoglobulin G (IgG) were injected into the reaction chamber and separated on the surface of the biofilter by applying magnetic fields. While holding the magnetic beads, antigen (mouse IgG) was injected into the chamber and incubated. Then

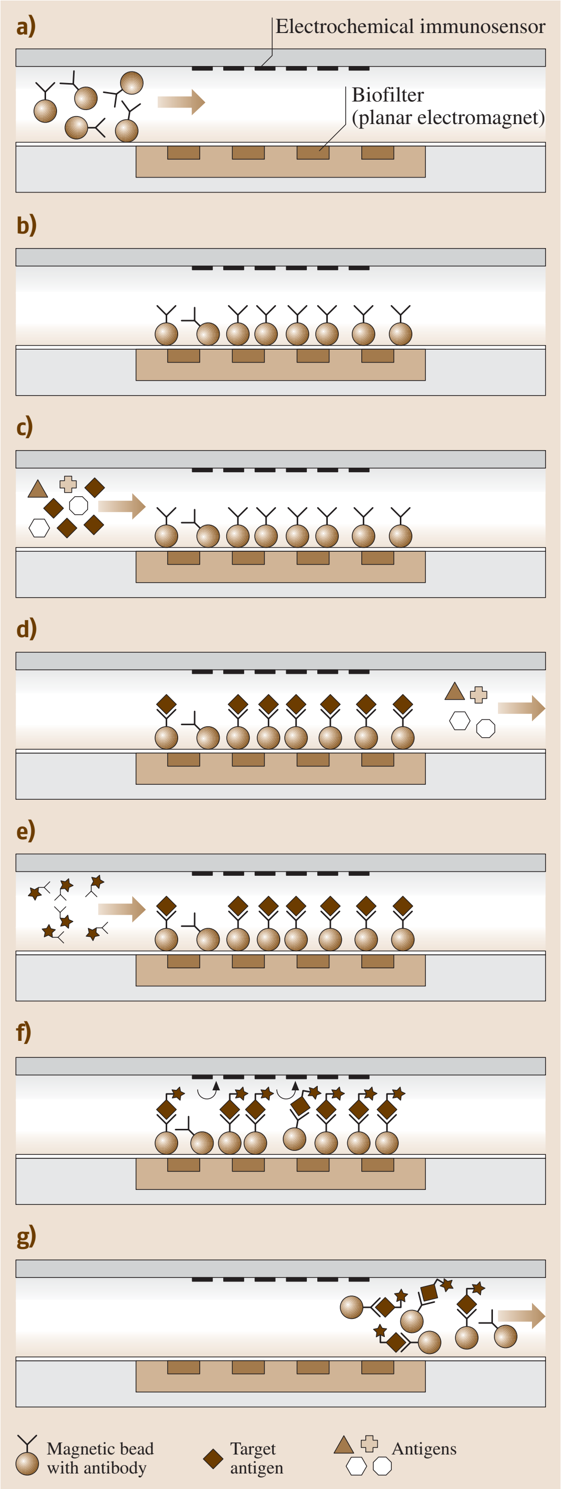

Fig. 9.24a–g Conceptual illustration of bio-sampling and immuno-assay procedure: **(a)** injection of magnetic beads; **(b)** separation and holding of beads; **(c)** flowing samples; **(d)** immobilization of target antigen; **(e)** flowing labeled antibody; **(f)** electrochemical detection; and **(g)** washing out magnetic beads and ready for another immunoassay (after [9.77], reproduced by permission of The Royal Society of Chemistry)

secondary antibody with label (rat anti-mouse IgG conjugated alkaline phosphatase) and electrochemical substrate (PAPP) to alkaline phosphatase was sequentially injected and incubated to ensure production of PAP. Electrochemical detection using an amperometric time-based detection method was performed during incubation. After detection, magnetic beads with all the reagents were washed away, and the system was ready for another immunoassay. This sequence was repeated for every new immunoassay. The flow rate was set to 20 μl/min in every step. After calibration of the electrochemical immunosensor, full immunoassays were performed following the sequence stated above for different antigen concentrations: 50, 75, 100, 250, and 500 ng/ml. Concentration of primary antibody-coated magnetic beads and conjugated secondary antibody was 1.02×10^7 beads/ml and 0.7 μg/ml, respectively. Immunoassay results for different antigen concentrations are shown in Fig. 9.26.

Immunoreactant consumed during one immunoassay was 10 μl (20 μl/min × 30 s), and total assay time was less than 20 min, including all incubation and detection steps.

The integrated microfluidic biochemical detection system has been successfully developed and fully tested for fast and low volume immunoassays using magnetic beads, which are used as both immobilization surfaces and bio-molecule carriers. Magnetic bead-based immunoassay, as a typical example of biochemical detection and analysis, has been performed on the integrated microfluidic biochemical analysis system that includes a surface-mounted biofilter and immunosensor on a glass microfluidic motherboard. Protein sampling capability has been demonstrated by capturing target antigens.

The methodology and system can also be applied to generic bio-molecule detection and analysis systems by replacing the antibody/antigen with appropriate bio receptors/reagents, such as DNA fragments or oligonucleotides, for the application to DNA analysis and/or high throughput protein analysis.

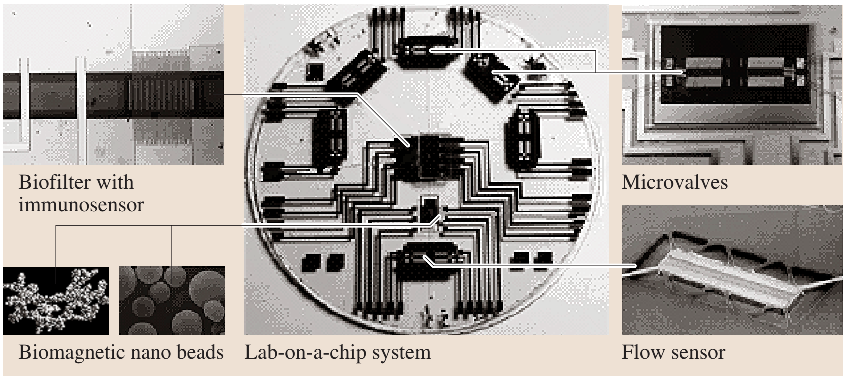

Fig. 9.25 Photograph of the fabricated lab-on-a-chip for magnetic bead-based immunoassay

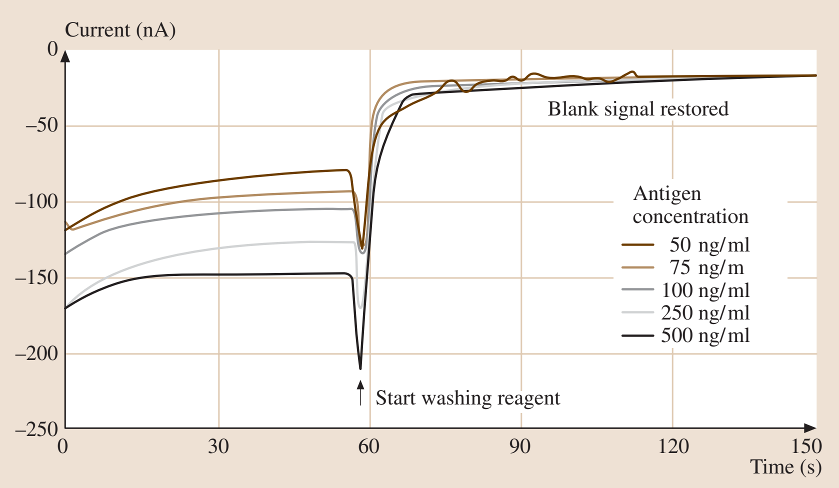

Fig. 9.26 Immunoassay results measured by amperometric time-based detection method (after [9.77], reproduced by permission of The Royal Society of Chemistry)

9.4.2 Disposable Smart Lab-on-a-Chip for Blood Analysis

One of several substrates available for bio-fluidic chips, plastics have recently become one of the most popular and promising substrates due to their low cost, ease of fabrication, and favorable biochemical reliability and compatibility. Plastic substrates, such as polyimide, PMMA, PDMS, polyethylene, or polycarbonate, offer a wide range of physical and chemical material parameters for the applications of bio-fluidic chips, generally at low cost using replication approaches. The disposable smart plastic biochip is composed of integrated modules of plastic fluidic chips for fluid regulation, chemical and biological sensors, and electronic controllers. As a demonstration vehicle, the biochip has the specific goal of detecting and identifying three metabolic parameters such as PO_2 (partial pressure of oxygen), lactate, and glucose from a blood sample. The schematic concept of the cartridge-type disposable lab-on-a-chip for blood analysis is illustrated in Fig. 9.27. The disposable lab-on-a-chip cartridge has been fabricated using plastic micro-injection molding and plastic-to-plastic direct bonding techniques. The biochip cartridge consists of a fixed volume microdispenser based on the sPROMs (structurally programmable microfluidic system) technique [9.70], an air-bursting, on-chip pressure source [9.80], and electrochemical biosensors [9.81].

A passive microfluidic dispenser measures exact amounts of sample to be analyzed, and then the air-bursting on-chip power source is detonated to push the graduated sample fluid from the dispenser reservoir. Upon air bursting, the graduated sample fluid travels through the microfluidic channel into sensing reservoirs, under which the biosensor array is located, as shown in Fig. 9.28.

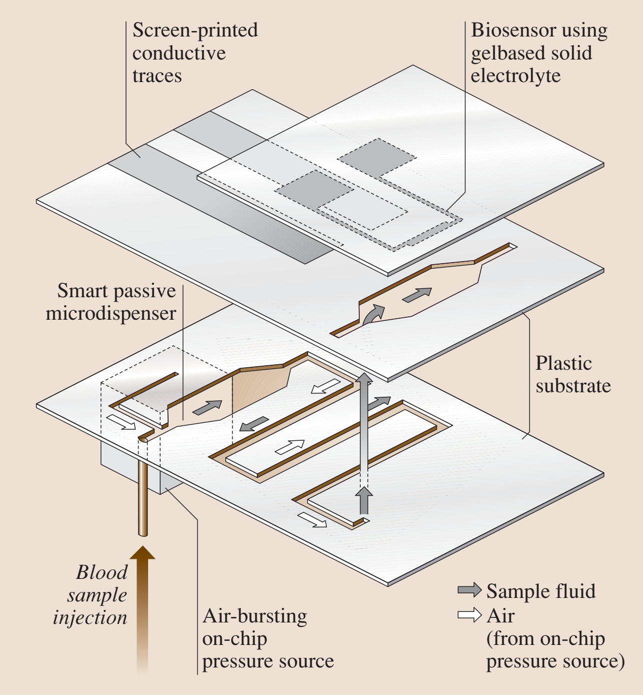
Screen-printed conductive traces
Biosensor using gelbased solid electrolyte
Smart passive microdispenser
Plastic substrate
Blood sample injection
Air-bursting on-chip pressure source
➡ Sample fluid
⇨ Air (from on-chip pressure source)

Fig. 9.27 Schematic illustration of smart and disposable plastic lab-on-a-chip by *Ahn* et al. [9.82]

An array of disposable biosensors consisting of an oxygen sensor and a glucose sensor has been fabricated using screen-printing technology [9.81]. Measurements from the developed biosensor array can be done based on tiny amounts of sample (as low as 100 nl). The glucose sensor is designed by adding two additional layers above the oxygen sensor (Fig. 9.29), which are glucose semipermeable membrane and glucose oxidase (GOD) layers.

The principle of the oxygen sensor is based on amperometric detection that when the oxygen diffusion profile from the sample to electrode surface is saturated, which means a constant oxygen gradient distribution profile is generated, the current is proportional only to the oxygen concentration in the solution. A silicone layer is spin-coated and utilized as an oxygen semipermeable membrane because of its high permeability and low signal-to-noise ratio. Water molecules pass through the silicone membrane and reconstitute the gel-based electrolyte so the Cl^- ions can move close to the anode to coalesce with Ag^+. The number of electrons in this reaction is counted by the measuring system.

For the glucose sensor, the additional layers – glucose semipermeable membrane (polyurethane) and the immobilized GOD (in a polyacrylamide gel) – allow direct conversion of the oxygen sensor to a glucose sensor. The glucose molecules will pass through the semipermeable layer and be oxidized immediately. The oxygen sensor will measure hydrogen peroxide, which is a by-product of glucose oxidation. The level of hydrogen peroxide is proportional to the glucose level in sample. The fabricated disposable plastic lab-on-a-chip cartridge was inserted into a hand-held type biochip analyzer for analysis of human blood sample, as shown in Fig. 9.30. The prototype biochip analyzer consists of biosensor detection circuitry, timing/sequence circuitry for the air-bursting, on-chip power source, and a display unit. The battery-operated biochip analyzer initiated the sensing sequence and displayed readings in one minute. Measured glucose and PO_2 levels in human blood sample are also shown in Fig. 9.31.

The development of the disposable smart microfluidic-based biochips is of immediate relevance to several patient-monitoring systems, specifically for point-of-care health monitors. Since the developed biochip is a low cost, plastic-based system, we envision a disposable application for monitoring clinically significant parameters such as PO_2, glucose, lactate, hematocrit, and pH. These health indicators provide an early warning system for the detection of patient status and can also serve as markers for disease and toxicity monitoring.

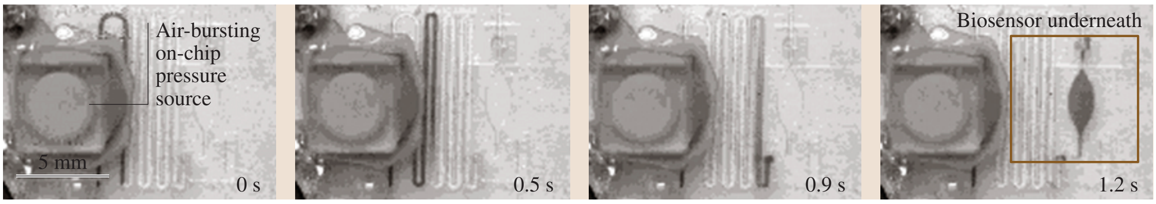

Fig. 9.28 Upon air-bursting, sample fluid travels through the microchannel into biosensor detection chamber. Biosensor was detached for clear micro-photographs. Dotted line indicates where biosensor is supposed to be attached (after [9.82])

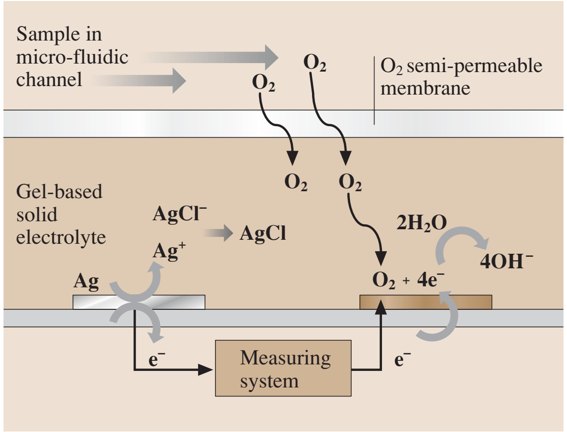
Sample in micro-fluidic channel
O₂
O₂
O₂ semi-permeable membrane
O₂
O₂
Gel-based solid electrolyte
AgCl⁻
AgCl
Ag⁺
2H₂O
4OH⁻
Ag
O₂ + 4e⁻
e⁻
Measuring system
e⁻

Fig. 9.29 Electrochemical and analytical principle of the developed disposable biosensor for partial oxygen concentration sensing (after, © 2002 IEEE [9.81])

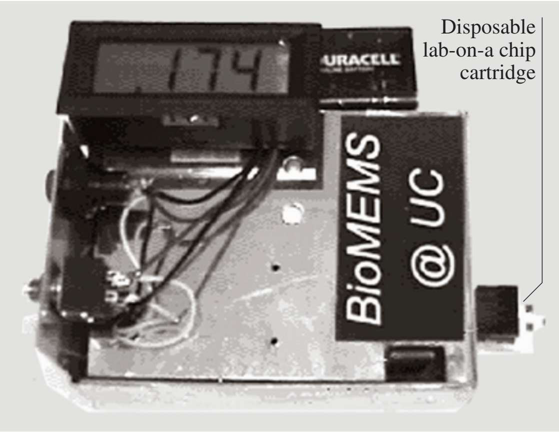
Disposable lab-on-a chip cartridge
URACELL
BioMEMS @ UC

Fig. 9.30 Hand-held type disposable smart lab-on-a-chip analyzer (after [9.82])

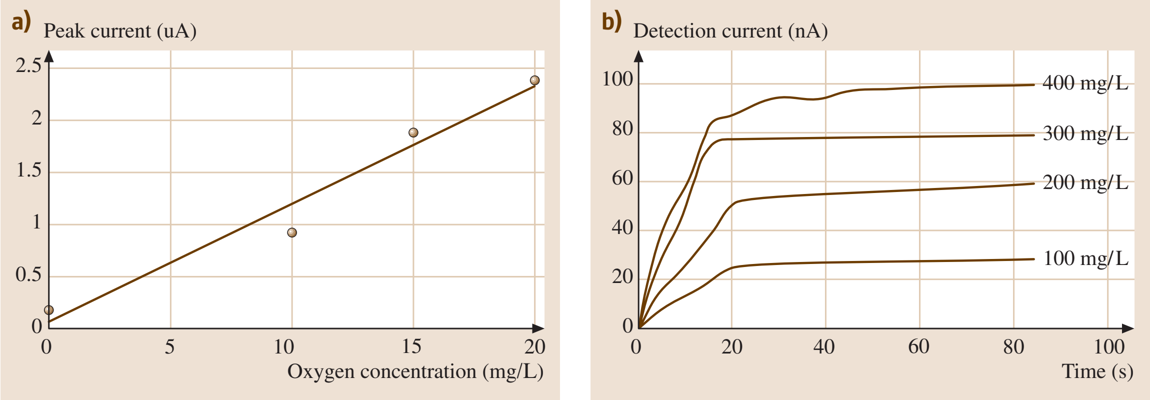

Fig. 9.31a,b Measurement results from the biochip cartridge and analyzer: **(a)** PO_2 calibration curve and **(b)**glucose level (after [9.81])

References

9.1 L. Ristic, H. Hughes, F. Shemansky: Bulk micromachining technology. In: *Sensor Technology and Devices*, ed. by L. Ristic (Artech House, Norwood 1994) pp. 49–93

9.2 M. Esashi, M. Takinami, Y. Wakabayashi, K. Minami: High-rate directional deep dry etching for bulk silicon micromachining, J. Micromech. Microeng. **5** (1995) 5–10

9.3 H. Jansen, M. de Boer, R. Legtenberg, M. Elwenspoek: The black silicon method: A universal method for determining the parameter setting of a fluorine-based reactive ion etcher in deep silicon trench etching with profile control, J. Micromech. Microeng. **5** (1995) 115–120

9.4 Z. L. Zhang, N. C. MacDonald: A RIE process for submicron silicon electromechanical structures, J. Micromech. Microeng. **2** (1992) 31–38

9.5 C. Linder, T. Tschan, N. F. de Rooij: Deep dry etching techniques as a new IC compatible tool for silicon micromachining, Technical Digest of the 6th International Conference on Solid-State Sensors and Actuators (Transducers '91), San Francisco June 24–27, 1991 (IEEE, Piscataway 1991) 524–527

9.6 D. J. Harrison, A. Manz, Z. Fan, H. Ludi, H. M. Widmer: Capillary electrophoresis and sample injection systems integrated on a planar glass chip, Anal. Chem. **64** (1992) 1926–1932

9.7 A. Manz, D. J. Harrison, E. M. J. Verpoorte, J. C. Fettinger, A. Paulus, H. Ludi, H. M. Widmer: Planar chips technology for miniaturization and integration of separation technique into monitoring systems, J. Chromatography **593** (1992) 253–258

9.8 T. R. Dietrich, W. Ehrfeld, M. Lacher, M. Krämer, B. Speit: Fabrication technologies for microsystems utilizing photoetchable glass, Microelectron. Eng. **30** (1996) 497–504

9.9 S. C. Jacobson, R. Hergenroder, L. B. Koutny, J. M. Ramsey: Open channel electrochromatography on a microchip, Anal. Chem **66** (1994) 2369–2373

9.10 M. G. Allen: Polyimide-based process for the fabrication of thick electroplated microstructures, Proc. of the 7th International Conference on Solid-State Sensors and Actuators (Transducers '93), Yokohama June 7–10, 1993 (IEE of Jpn., Tokyo 1993) 60–65

9.11 C. H. Ahn, M. G. Allen: A fully integrated surface micromachined magnetic microactuator with a multilevel meander magnetic core, IEEE/ASME J. Microelectromech. Syst. (MEMS) **2** (1993) 15–22

9.12 E. W. Becker, W. Ehrfeld, P. Hagmann, A. Maner, D. Munchmeyer: Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography, galvanofoming, and plastic moulding (LIGA Process), Microelectronic Eng. **4** (1986) 35–36

9.13 C. S. Effenhauser, G. J. M. Bruin, A. Paulus, M. Ehrat: Integrated capillary electrophoresis on flexible silicone microdevices: Analysis of DNA restriction fragments and detection of single DNA molecules on microchips, Anal. Chem. **69** (1997) 3451–3457

9.14 D. C. Duffy, J. C. McDonald, O. J. A. Schueller, G. M. Whitesides: Rapid prototyping of microfluidic systems in poly(dimethylsiloxane), Anal. Chem. **70** (1998) 4974–4984

9.15 L. Martynova, L. Locascio, M. Gaitan, G. W. Kramer, R. G. Christensen, W. A. MacCrehan: Fabrication of plastic microfluid channels by imprinting methods, Anal. Chem. **69** (1997) 4783–4789

9.16 H. Becker, W. Dietz, P. Dannberg: Microfluidic manifolds by polymer hot embossing for micro-TAS applications, Proc. of Micro-Total Analysis Systems '98, Banff, Canada 1998, ed. by D. J. Harrison, A. van den Berg (Kluwer, Dordrecht 1998) 253–256

9.17 H. Becker, U. Heim: Hot embossing as a method for the fabrication of polymer high aspect ratio structures, Sens. Actuators A **83** (2000) 130–135

9.18 R. M. McCormick, R. J. Nelson, M. G. Alonso-Amigo, J. Benvegnu, H. H. Hooper: Microchannel electrophoretic separations of DNA in injection-molded plastic substrates, Anal. Chem. **69** (1997) 2626–2630

9.19 C. H. Ahn, A. Puntambekar, S. M. Lee, H. J. Cho, C.-C. Hong: Structurally programmable microfluidic systems, Proc. of Micro-Total Analysis systems 2000, Enschede 2000, ed. by A. van den Berg, W. Olthius, P. Bergveld (Kluwer, Dordrecht 2000) 205–208

9.20 W. Schomburg, B. Scherrer: 3.5 mm thin valves in titanium membranes, J. Micromech. Microeng. **2** (1992) 184–186

9.21 A. Ilzofer, B. Ritter, C. Tsakmasis: Development of passive microvalves by the finite element method, J. Micromech. Microeng. **5** (1995) 226–230

9.22 B. Paul, T. Terharr: Comparison of two passive microvalve designs for microlamination architectures, J. Micromech. Microeng. **10** (2000) 15–20

9.23 S. Shoji, M. Esashi: Microflow device and systems, J. Micromech. Microeng. **4** (1994) 157–171

9.24 G. T. A. Kovacs: *Micromachined Transducers Sourcebook* (McGraw-Hill, New York 1998)

9.25 W. Schomburg, J. Fahrenberg, D. Maas, R. Rapp: Active valves and pumps for microfluidics, J. Micromech. Microeng. **3** (1993) 216–218

9.26 K. Hosokawa, R. Maeda: A pneumatically-actuated three way microvalve fabricated with polydimethylsiloxane using the membrane transfer technique, J. Micromech. Microeng. **10** (2000) 415–420

9.27 J. Fahrenberg, W. Bier, D. Maas, W. Menz, R. Ruprecht, W. Schomburg: A microvalve system fabricated by thermoplastic molding, J. Micromech. Microeng. **5** (1995) 169–171

9.28 C. Goll, W. Bacher, B. Bustgens, D. Maas, W. Menz, W. Schomburg: Microvalves with bistable buckled diaphragms, J. Micromech. Microeng **6** (1996) 77–79

9.29 A. Ruzzu, J. Fahrenberg, M. Heckele, T. Schaller: Multi-functional valve components fabricated by combination of LIGA process and high precision mechanical engineering, Microsystem Technol. **4** (1998) 128–131

9.30 K. Sato, M. Shikida: An electrostatically actuated gas valve with S-shaped film element, J. Micromech. Microeng. **4** (1994) 205–209

9.31 J. Robertson, K. D. Wise: A low pressure micromachined flow modulator, Sens. Actuators A **71** (1998) 98–106

9.32 D. J. Sadler, K. W. Oh, C. H. Ahn, S. Bhansali, H. T. Henderson: A new magnetically actuated microvalve for liquid and gas control applications, Proc. of the 6th International Conference on Solid-State Sensors and Actuators (Transducers '99), Sendai 1999 (IEE of Jpn., Tokyo 1999) 1812–1815

9.33 W. Wijngaart, H. Ask, P. Enoksson, G. Stemme: A high-stroke, high-pressure electrostatic actuator for valve applications, Sens. Actuators A **100** (2002) 264–271

9.34 T. Watanabe, H. Kuwano: A microvalve matrix using piezoelectric actuators, Microsystem Technol. **3** (1997) 107–111

9.35 M. Stehr, S. Messner, H. Sandmaier, R. Zangerle: The VAMP – A new device for handling liquids and gases, Sens. Actuators A **57** (1996) 153–157

9.36 W. Jackson, H. Tran, M. O'Brien, E. Rabinovich, G. Lopez: Rapid prototyping of active microfluidic components based on magnetically modified elastomeric materials, J. Vac. Sci. Technol. B **19** (2001) 596

9.37 M. Kohl, K. Skrobanek, S. Miyazaki: Development of stress-optimized shape memory microvalves, Sens. Actuators A **72** (1999) 243–250

9.38 C. Neagu, J. Gardeniers, M. Elwenspoek, J. Kelly: An electrochemical active valve, Electrochemica Acta **42** (1997) 3367–3373

9.39 K. Yoshida, M. Kikuchi, J. Park, S. Yokota: Fabrication of micro-electro-rheological valves (ER valves) by micromachining and experiments, Sens. Actuators A **95** (2002) 227–233

9.40 N. Jeon, D. Chiu, C. Wargo, H. Wu, I. Choi, J. Anderson, G. M. Whitesides: Design and fabrication of integrated passive valves and pumps for flexible polymer 3-dimensional microfluidic systems, Biomedical Microdevices **4** (2002) 117–121

9.41 M. Koch, N. Harris, A. Evans, N. White, A. Brunnschweiler: A novel micropump design with thick-film piezoelectric actuation, Meas. Sci. Technol. **8** (1997) 49–57

9.42 L. Cao, S. Mantell, D. Polla: Design and simulation of an implantable medical drug delivery system using microelectromechanical systems technology, Sens. Actuators A **94** (2001) 117–125

9.43 J. Park, K. Yoshida, S. Yokota: Resonantly driven piezoelectric micropump fabrication of a micropump having high power density, Mechatronics **9** (1999) 687–702

9.44 M. Koch, A. Evans, A. Brunnschweiler: The dynamic micropump driven with a screen printed PZT actuator, J. Micromech. Microeng. **8** (1998) 119–122

9.45 D. Xu, L. Wang, G. Ding, Y. Zhou, A. Yu, B. Cai: Characteristics and fabrication of NiTi/Si diaphragm micropump, Sens. Actuators A **93** (2001) 87–92

9.46 E. Makino, T. Mitsuya, T. Shibata: Fabrication of TiNi shape memory micropump, Sens. Actuators A **88** (2001) 256–262

9.47 H. Chou, M. Unger, S. Quake: A microfabricated rotary pump, Biomedical Microdevices **3** (2001) 323–330

9.48 R. Zengerle, J. Ulrich, S. Kulge, M. Richter, A. Richter: A bidirectional silicon micropump, Sens. Actuators A **50** (1995) 81–86

9.49 M. Carrozza, N. Croce, B. Magnani, P. Dario: A piezoelectric driven stereolithography-fabricated micropump, J. Micromech. Microeng. **5** (1995) 177–179

9.50 H. Andersson, W. Wijngaart, P. Nilsson, P. Enoksson, G. Stemme: A valveless diffuser micropump for microfluidic analytical systems, Sens. Actuators B **72** (2001) 259–265

9.51 F. Forster, R. Bardell, M. Afromowitz, N. Sharma: Design, fabrication and testing of fixed-valve micropumps, Proc. ASME Fluids Engineering Division, ASME Int. Mechanical Engineering Congress and Exposition **234** (1995) 39–44

9.52 G. Fuhr, T. Schnelle, B. Wagner: Travelling wave driven microfabricated electrohydrodynamic pumps for liquids, J. Micromech. Microeng. **4** (1994) 217–226

9.53 J. Jang, S. Lee: Theoretical and experimental study of MHD micropump, Sens. Actuators A **80** (2000) 84–89

9.54 A. Lemoff, A. P. Lee: An AC magnetohydrodynamic micropump, Sens. Actuators B **63** (2000) 178–185

9.55 J. DeSimone, G. York, J. McGrath: Synthesis and bulk, surface and microlithographic characterization of poly(1-butene sulfone)-g-poly(dimethylsiloxane), Macromolecules **24** (1994) 5330–5339

9.56 A. Wego, L. Pagel: A self-filling micropump based on PCB technology, Sens. Actuators A **88** (2001) 220–226

9.57 A. Terray, J. Oakey, D. Marr: Fabrication of colloidal structures for microfluidic applications, Appl. Phys. Lett. **81** (2002) 1555–1557

9.58 A. Puntambekar, J.-W. Choi, C. H. Ahn, S. Kim, V. B. Makhijani: Fixed-volume metering microdispenser module, Lab on a Chip **2** (2002) 213–218

9.59 M. Madou, Y. Lu, S. Lai, C. Koh, L. Lee, B. Wenner: A novel design on a CD disc for 2-point calibration measurement, Sens. Actuators A **91** (2001) 301–306

9.60 K. Handique, D. Burke, C. Mastrangelo, C. Burns: On-chip thermopneumatic pressure for discrete drop pumping, Anal. Chem. **73** (2001) 1831–1838

9.61 H. Andersson, W. Wijngaart, P. Griss, F. Niklaus, G. Stemme: Hydrophobic valves of plasma deposited octafluorocyclobutane in DRIE channels, Sens. Actuators B **75** (2001) 136–141

9.62 L. Low, S. Seetharaman, K. He, M. Madou: Microactuators toward microvalves for responsive controlled drug delivery, Sens. Actuators B **67** (2000) 149–160

9.63 Q. Yu, J. Bauer, J. Moore, D. Beebe: Responsive biomimetic hydrogel valve for microfluidics, Appl. Phys. Lett. **78** (2001) 2589–2591

9.64 M. Mitchell, V. Spikmans, F. Bessoth, A. Manz, A. de Mello: Towards organic synthesis in microfluidic devices: Multicomponent reactions for the construction of compound libraries, Proc. of Micro-Total Analysis systems 2000, Enschede 2000, ed. by A. van den Berg, W. Olthius, P. Bergveld (Kluwer, Dordrecht 2000) 463–465

9.65 D. Beebe, R. Adrian, M. Olsen, M. Stremler, H. Aref, B. Jo: Passive mixing in microchannels: Fabrication and flow experiments, Mecanique & Industries, **2** (2001) 343–348

9.66 C.-C. Hong, J.-W. Choi, C. H. Ahn: A novel in-plane passive micromixer using Coanda effect, Proc. of the μTAS 2001 Symposium, Monterey 2001, ed. by J. M. Ramsey, A. van den Berg (Kluwer, Dordrecht 2001) 31–33

9.67 J. Brody, P. Yager: Diffusion-based extraction in a microfabricated device, Sens. Actuators A **54** (1996) 704–708

9.68 A. Stroock, S. Dertinger, A. Ajdari, I. Mezic, H. Stone, G. M. Whitesides: Chaotic mixer for microchannels, Science **295** (2002) 647–651

9.69 B. Weigl, P. Yager: Microfluidic diffusion-based separation and detection, Science **283** (1999) 346–347

9.70 A. Puntambekar, J.-W. Choi, C. H. Ahn, S. Kim, S. Bayyuk, V. B. Makhijani: An air-driven fluidic multiplexer integrated with microdispensers, Proc. of the μTAS 2001 Symposium, Monterey 2001, ed. by J. M. Ramsey, A. van den Berg (Kluwer, Dordrecht 2001) 78–80

9.71 T. Nagakura, K. Ishihara, T. Furukawa, K. Masuda, T. Tsuda: Auto-regulated osmotic pump for insulin therapy by sensing glucose concentration without energy supply, Sens. Actuators B **34** (1996) 229–233

9.72 Y. Su, L. Lin, A. Pisano: Water-powered osmotic microactuator, Proc. of the 14th IEEE MEMS Workshop (MEMS 2001), Interlaken 2001 (IEEE, Piscataway 393–396

9.73 C. Effenhauser, H. Harttig, P. Kramer: An evaporation-based disposable micropump concept for continuous monitoring applications, Biomedical Devices **4** (2002) 27–32

9.74 C. H. Ahn, H. T. Henderson, W. R. Heineman, H. B. Halsall: Development of a generic microfluidic system for electrochemical immunoassay-based remote bio/chemical sensors, Proc. of Micro-Total Analysis Systems '98, Banff, Canada 1998, ed. by D. J. Harrison, A. van den Berg (Kluwer, Dordrecht 1998) 225–230

9.75 J.-W. Choi, K. W. Oh, A. Han, N. Okulan, C. A. Wijayawardhana, C. Lannes, S. Bhansali, K. T. Schlueter, W. R. Heineman, H. B. Halsall, J. H. Nevin, A. J. Helmicki, H. T. Henderson, C. H. Ahn: Development and characterization of microfluidic devices and systems for magnetic bead-based biochemical detection, Biomedical Microdevices **3** (2001) 191–200

9.76 J.-W. Choi, C. H. Ahn, S. Bhansali, H. T. Henderson: A new magnetic bead-based, filterless bio-separator with planar electromagnet surfaces for integrated bio-detection systems, Sens. Actuators B **68** (2000) 34–39

9.77 J.-W. Choi, K. W. Oh, J. H. Thomas, W. R. Heineman, H. B. Halsall, J. H. Nevin, A. J. Helmicki, H. T. Henderson, C. H. Ahn: An integrated microfluidic biochemical detection system for protein analysis with magnetic bead-based sampling capabilities, Lab on a Chip **2** (2002) 27–30

9.78 O. Niwa, Y. Xu, H. B. Halsall, W. R. Heineman: Small-volume voltammetric detection of 4-aminophenol with interdigitated array electrodes and its application to electrochemical enzyme immunoassay, Anal. Chem. **65** (1993) 1559–1563

9.79 K. W. Oh, A. Han, S. Bhansali, H. T. Henderson, C. H. Ahn: A low-temperature bonding technique using spin-on fluorocarbon polymers to assemble microsystems, J. Micromech. Microeng. **12** (2002) 187–191

9.80 C.-C. Hong, J.-W. Choi, C. H. Ahn: A disposable on-chip air detonator for driving fluids on point-of-care systems, Proc. of the 6th International Conference on Micro Total Analysis Systems (μ-TAS), Nara 2002, ed. by Y. Baba, S. Shoji, A. van den Berg (Kluwer, Dordrecht 2002)

9.81 C. Gao, J.-W. Choi, M. Dutta, S. Chilukuru, J. H. Nevin, J. Y. Lee, M. G. Bissell, C. H. Ahn: A fully integrated biosensor array for measurement of metabolic parameters in human blood, Proc. of the 2nd Second Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology, Madison May 2–4, 2002, ed. by A. Dittmar, D. Beebe (IEEE, New York 2002) 223–226

9.82 C. H. Ahn, J.-W. Choi, A. Puntambekar, C.-C. Hong, X. Zhu, C. Gao, R. Trichur, S. Chilukuru, M. Dutta, S. Murugesan, S. Kim, Y.-S. Sohn, J. H. Nevin, G. Beaucage, J.-B. Lee, J. Y. Lee, M. G. Bissell: Disposable biochip cartridge for clinical diagnostics toward point-of-care systems, Proc. of the 6th International Conference on Micro Total Analysis Systems (μ-TAS), Nara 2002, ed. by Y. Baba, S. Shoji, A. van den Berg (Kluwer, Dordrecht 2002) 187–189

10. Therapeutic Nanodevices

Therapeutic nanotechnology offers minimally invasive therapies with high densities of function concentrated in small volumes, features that may reduce patient morbidity and mortality. Unlike other areas of nanotechnology, novel physical properties associated with nanoscale dimensionality are not the raison d'etre of therapeutic nanotechnology, whereas the aggregation of multiple biochemical (or comparably precise) functions into controlled nanoarchitectures is. Multifunctionality is a hallmark of emerging nanotherapeutic devices, and multifunctionality can allow nanotherapeutic devices to perform multi-step work processes, with each functional component contributing to one or more nanodevice subroutine such that, in aggregate, subroutines sum to a cogent work process. Cannonical nanotherapeutic subroutines include tethering (targeting) to sites of disease, dispensing measured doses of drug (or bioactive compound), detection of residual disease after therapy and communication with an external clinician/operator. Emerging nanotherapeutics thus blur the boundaries between medical devices and traditional pharmaceuticals. Assembly of therapeutic nanodevices generally exploits either (bio)material self assembly properties or chemoselective bioconjugation techniques, or both. Given the complexity, composition, and the necessity for their tight chemical and structural definition inherent in the nature of nanotherapeutics, their cost of goods (COGs) might exceed that of (already expensive) biologics. Early therapeutic nanodevices will likely be applied to disease states which exhibit significant unmet patient need (cancer and cardiovascular disease), while application to other disease states well-served by conventional therapy may await perfection of nanotherapeutic design and assembly protocols.

10.1 **Definitions and Scope of Discussion** 280
10.1.1 Design Issues 281
10.1.2 Utility and Scope of Therapeutic Nanodevices 285

10.2 **Synthetic Approaches: "top-down" versus "bottom-up" Approaches for Nanotherapeutic Device Components** 285
10.2.1 Production of Nanoporous Membranes by Microfabrication Methods: A top-down Approach 285
10.2.2 Synthesis of Poly(amido) Amine (PAMAM) Dendrimers: A bottom-up Approach 286
10.2.3 The Limits of top-down and bottom-up Distinctions with Respect to Nanomaterials and Nanodevices 287

10.3 **Technological and Biological Opportunities** 288
10.3.1 Assembly Approaches 288
10.3.2 Targeting: Delimiting Nanotherapeutic Action in Three-Dimensional Space 296
10.3.3 Triggering: Delimiting Nanotherapeutic Action in Space and Time 298
10.3.4 Sensing Modalities 302
10.3.5 Imaging Using Nanotherapeutic Contrast Agents 304

10.4 **Applications for Nanotherapeutic Devices** 307
10.4.1 Nanotherapeutic Devices in Oncology 307
10.4.2 Cardiovascular Applications of Nanotherapeutics 310
10.4.3 Nanotherapeutics and Specific Host Immune Responses 311

10.5 **Concluding Remarks: Barriers to Practice and Prospects** 315
10.5.1 Complexity in Biology 315
10.5.2 Dissemination of Biological Information 315
10.5.3 Cultural Differences Between Technologists and Biologists 316

References 317

10.1 Definitions and Scope of Discussion

Nanotechnology is a field in rapid flux and development, as cursory examination of this volume shows, and definition of its meets and bounds, as well as identification of sub-disciplines embraced by it, can be elusive. The word means many things to many people, and aspects of multiple disciplines, from physics to information technology to biotechnology, legitimately fall into the intersection of the Venn diagram of disciplines that defines nanotechnology. The breadth of the field allows almost any interested party to contribute to it, but the same ambiguity can render the field diffuse and amorphous. If nanotechnology embraces everything, what then is it? The ambiguity fuels cognitive dissonance that can result in frustrating interactions between investigators and funders, authors and editors, and entrepreneurs and investors. Some consideration of the scope of the field therefore is useful.

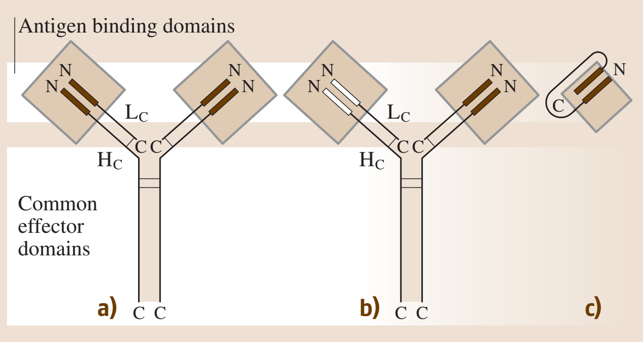

Fig. 10.1a–c Antibodies resemble purpose-built devices with distinct functional domains [10.1]. Native antibodies are composed of four polypeptide chains: two heavy chains (H_C) and two light chains (L_C), joined by interchain disulfide linkages (lines between H_C and L_C moieties). Amino and carboxy termini of individual polypeptide chains are indicated (by N and C). Antigen binding domains are responsible for specific antigen recognition, vary from antibody to antibody, and are indicated by the thicker lines. Common effector functions (F_C receptor binding, complement fixation, etc.) are delimited to domains of the antibodies that are constant from molecule to molecule. **(a)** A native IgG antibody is monospecific but bivalent in its antigen binding capacity. **(b)** An engineered, bispecific, bivalent antibody capable of recognizing two distinct antigens. **(c)** An engineered antibody fragment (single chain Fv or SCFv) that is monospecific and monovalent can recognize only one antigenic determinant and is engineered to lack common effector functions. This construct is translated as a single, continuous polypeptide chain (hence the name SCFv) because a peptide linker (indicated by the connecting line in the figure) is incorporated to connect the carboxy end of the H_C fragment and the amino end of the L_C fragment

To frame the discussion, we will define nanotechnology as the discipline that aims to satisfy desired objectives using materials and devices whose valuable properties are based on a specific nanometer-scale element of their structures. The field is unabashedly application-oriented, so its raison d'être is fulfillment of tasks of interest: technical information is important primarily to the extent that it bears on device design, function, or application.

The meaning of "therapeutic" is largely self-explanatory and refers here to intervention in *human* disease processes (although many of the approaches discussed are equally applicable to veterinary medicine). Our discussion will be confined primarily to therapeutics used in vivo, because such applications clearly benefit from the low invasiveness that ultra-small, but multipotent, nanotherapeutics potentially offer. It is debatable whether imaging, diagnostic, or sensing devices can be considered therapeutic in this context, though, as we will see, sensing/diagnostic functionalities are often inextricable elements of therapeutic nanodevices, and it is difficult to consider so-called smart nanotherapeutics without discussion of their sensing capabilities.

Our definition of nanotechnology projects several corollaries. First, it embraces macroscale structures whose useful properties derive from their nanoscale aspects. Second, the modifier "specific" (as in, "specific nanometer-scale elements") is intended to exclude materials whose utility derives solely from properties inherent in being finely divided (high surface-to-volume ratios, for instance), or other bulk chemical and physical properties. We made the exclusion based on our assessment that therapeutic nanodevices are more intriguing than nanomaterials per se (see below), though we will engage these attributes where they are germane to specific devices or therapeutic applications. Third, our definition implies that limited nanotechnology has been available since the 1970s in the form of biotechnology. Based on their nanoscale structures, individual biological macromolecules (such as proteins) often exhibit the coordinated, modular multifunctionality that is characteristic of purpose-built devices (Fig. 10.1). An analogous, but perhaps less persuasive, argument can be made that organic chemistry is an early form of nanotechnology. Compared to organic small molecules, protein functional capabilities and properties are generally more complex and extremely dependent on their conformation in three-dimensional space at nanometer scale. The nanotechnology sobriquet, therefore, may be more appropriate to biotechnology than organic chemistry.

Biological macromolecules rely on the deployment of specific chemical functionalities to specific relative distributions in space with nanometer (and greater) resolution for their function, so the inclusion of molecular engineering aspects of biotechnology practice under the nanotechnology rubric is legitimate, despite the discomfort it may cause traditionally trained engineers. As we will see, intervention in human disease often requires inclusion of biomolecules in therapeutic devices: frequently, no functional synthetic analogue of active proteins and nucleic acids is available.

As described above, this chapter focuses primarily on nanoscale therapeutic devices as opposed to therapeutic nanomaterials. Devices are integrated functional structures and not mixtures of materials. Devices exhibit desirable emergent properties inherent in their design: the properties emerge as the result of the spatial and/or temporal organization, and coordination and regulation of action of individual components. The organization of components in devices allows them to perform multistep, cogent work processes that can't be mimicked by simple admixtures of individual components. In fact, if device functions can be mimicked well by simple mixtures of components, the labor involved in configuring and constructing a nanoscale device is not warranted. Our device definition thus excludes nanomaterials used as drug formulation excipients (pharmacologically inert materials included in formulations that improve pharmacophore uptake, biodistribution, pharmacokinetic, handling, storage, or other properties), but embraces those same materials as integral components of drug delivery or other clinical devices.

10.1.1 Design Issues

The biotechnology industry historically has focused on production of individual soluble protein and nucleic acid molecules for pharmaceutical use, with only limited attention paid to functional supramolecular structures [10.2–7]. This bias toward free molecules flies in the face of the obvious importance of integrated supramolecular structures in biology and, to the casual observer, may seem an odd gap in attention and emphasis on the part of practicing biotechnologists. The bias toward single molecule, protein therapeutics, however, follows from the fact that biotechnology is an industrial activity, governed by market considerations. Of the myriad potential therapeutics that might be realized from biotechnology, single protein therapeutics are among the easiest to realize from both technical and regulatory perspectives and so warrant extensive industrial attention. This is changing, however, and more complex entities (actual supramolecular therapeutic devices) have and will appear with increasing frequency in the twenty-first century.

New top-down and bottom-up materials derived from micro/nanotechnology provide the opportunity to complement the traditional limits of biotechnology by providing scaffolds that can support higher level organization of multiple biomolecules to perform work activities they could not perform as free, soluble molecules. Such supramolecular structures have been called nanobiotechnological devices [10.8], nanobiological devices [10.2–7], or semi-synthetic nanodevices and figure prominently in therapeutic nanotechnology.

Incorporation of / Interaction with Biomolecules

In general, design of nanodevices is similar to design of other engineered structures, providing that the special properties of the materials (relating to their nanoscale aspects such as quantum, electrical, mechanical, biological properties, etc.), as well as their impact in therapy, are considered. Therapeutics can interact with patients on multiple levels, ranging from organismal to molecular, but it is reasonable to expect that most nanotherapeutics will interface with patients at the nanoscale at least to some extent [10.2–5, 9–14]. Typically, this means interaction between therapeutics and biological macromolecules, supramolecular structures and organelles, which, in turn, often dictates the incorporation of biological macromolecules (and other biostructures) into nanodevices [10.2, 5, 13–15]. Incorporating biological structures into (nanobiological) devices presents special challenges that do not occur in other aspects of engineering practice.

Unlike fully synthetic devices, semi-biological nanodevices must incorporate pre-fabricated biological components (or derivatives thereof), and therefore the intact nanodevices are seldom made entirely de novo. As a corollary, knowledge of properties of biological device components is often incomplete (as they were not made by human design), and therefore the range of activities inherent in any nanobiological device design may be much less obvious and less well-defined than it is for fully synthetic devices. Further complicating the issue, the activities of biological molecules are often multifaceted (many genes and proteins exhibit pleiotropic activities), and the full range of functionality of individual biological molecules in interactions with other biological systems (as in nanotherapeutics) is often not known. This makes design and prototyping of biological nanodevices an empirically intensive, iterative process [10.3–5, 14].

Biological macromolecules have properties, particularly those relating to their stability, that can limit their use in device contexts. In general, proteins, nucleic acids, lipids, and other biomolecules are more labile to physical insult than are synthetic materials. With the possible exceptions of topical agents or oral delivery and endosomal uptake of nanotherapeutics (both involving exposure to low pH), patients can tolerate conditions encountered by nanobiological therapeutics in vivo, and device lability in the face of *physical* insult is generally a major consideration only in ex vivo settings (relating to storage, sterilization, ex vivo cell culture, etc.). Living organisms remodel themselves constantly in response to stress, development, pathology, and external stimuli. For instance, epithelial tissues and blood components are constantly eliminated and regenerated, and bone and vasculature are continuously remodeled. The metabolic facilities responsible (circulating and tissue-bound proteases and other enzymes, various clearance organs, the immune system, etc.) can potentially process biological components of nanobiological therapeutic devices as well as endogenous materials, leading to partial or complete degradation of nanotherapeutic structure, function, or both. Furthermore, the host immune and wound responses protect the host against pathogenic organism incursions by mechanisms that involve sequestering and degrading the pathogens. Nanobiological therapeutics are subject to the actions of these host defense systems and to normal remodeling processes. As we will discuss, various strategies to stabilize biomolecules and structures in heterologous in vivo environments are applicable to nanobiological therapeutics [10.3, 16–20]. Conversely, instability of active biocomponents can offer a valuable and simple way to delimit the activity of nanotherapeutics containing biomolecules.

Nanotherapeutic Design Paradigms

Several early attempts to codify the canonical properties of ideal nanobiological devices, and therapeutic nanodevices in particular, have been made [10.5, 13, 15, 21] and are summarized in Table 10.1. In general, nanobiological devices contain biological components that retain their function in new (device) contexts. In other words, one must abstract enough of a functional biological unit from its native context to allow it to perform the function for which it was selected. If one wishes, for example, to appropriate the specific antigen recognition property of an antibody for a device function (say, in targeting, discussed later), it is not necessary to incorporate the entire 150,000 atomic mass unit (AMU) antibody, the bulk of which is devoted to functions other than antigen recognition (see Fig. 10.1 [10.1]), but it *is* critical to incorporate the approximately 20,000 AMU of the antibody essential for specific antibody–antigen

Table 10.1 Some ideal characteristics of nanodevices. (**A**) Characteristics of all nanobiological devices [10.2, 5, 13–15]. (**B**) Desirable characteristics of therapeutic platforms [10.2, 5, 13–15]

A
Biological molecules must retain function.
Device function is the result of the summed activities of device components.
The relative organization of device components drives device function.
Device functions can be unprecedented in the biological world.
B
Therapeutics should be minimally invasive.
Therapeutics should have the capacity to target sites of disease.
Therapeutics should be able to sense disease states in order to:
– report conditions at the disease site to clinicians.
– administer metered therapeutic interventions.
Therapeutic functions should be segregated into standardized modules.
Modules should be interchangeable to tune therapeutic function.

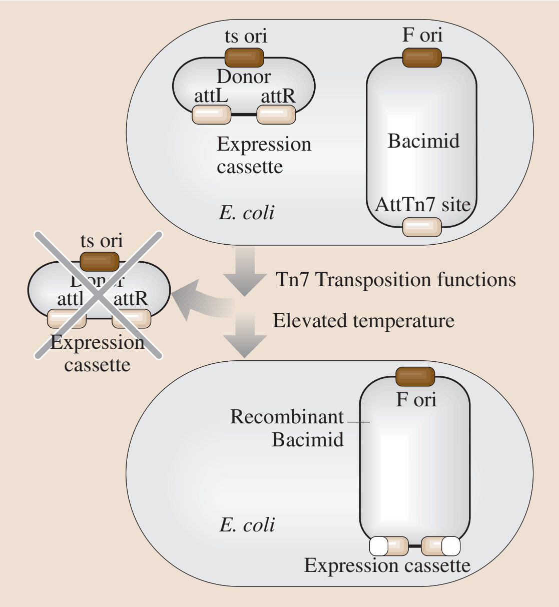

Fig. 10.2 The bacmid molecular cloning system is a molecular device designed to allow efficient production of recombinant insect viruses (baculovirus) in *Escherichia coli* [10.22–24]. Baculovirus is replicated in *E. coli* by the F plasmid origin of replication (F ori), and as such, is called a bacmid. The bacmid also includes an engineered transposable DNA element 7 (Tn7) attachment site isolated from the chromosome of an enteric bacteria (AttTn7). AttTn7 can receive Tn7 elements transposed from other cellular locations. A donor plasmid (donor) is replicated by a temperature-sensitive plasmid pSC101 origin of replication (ts ori). The donor also incorporates an expression cassette containing both the gene of interest for ultimate expression in insect cells and a selectable genetic marker operable in *E. coli*. The expression cassette is flanked by DNA sequences (attL and attR) that are recognized by the Tn7 transposition machinery. Tn7 transposition machinery resides elsewhere in the same *E. coli* cell. When donor plasmid is introduced into *E. coli* containing bacmid, Tn7 transposition machinery causes the physical relocation of expression cassettes from donor plasmid to bacmid. Unreacted donor plasmid is conveniently removed by elevating the incubation temperature, causing the ts pSC101 replicon to cease to function, in turn, causing the donor to be lost. If selection for the genetic markers within the expression cassette is applied at this point, the only *E. coli* that survive are those containing recombinant bacmid (i. e., those that have received the gene for insect cell expression by transposition from the donor). Recombinant bacmid are conveniently isolated from *E. coli* and introduced into insect cell culture, where expression of the gene of interest occurs

binding. Device function is the result of the summed and various activities of biological and synthetic device components, as well, though functional biological components generally exist in the context of higher order systems that support the organisms of which they are a part by the control the nanobiological device designer can exert on the relative organization of biological device components allows biomolecules abstracted from their native context and incorporated in nanobiotechnological devices to contribute to functions entirely different from those they performed in their organismal contexts. All of these features are illustrated in the bacmid, or Bac-to-Bac, system, a commercially available molecular cloning device ([10.22–24] and Fig. 10.2). This system configures prokaryotic genetic elements from multiple sources into a device for producing recombinant eukaryotic viruses, a function that is unprecedented in nature. The system is feasible because of the modularity of the genetic elements involved and because of the strict control of the relative arrangement of genetic elements allowed by recombinant DNA technology. Analogous devices based on bacterial and eukaryotic regulatory elements to pre-program the micro- and nanoscale architectural properties and physiological behavior of living things are now being realized [10.25, 26].

Bacmid provides an example of a nontherapeutic nanobiological device and illustrates some specific design approaches for building functional devices with biocomponents. Hypothetical properties of nanoscale devices specifically for therapeutic purposes have been codified (Table 10.1) and bear examination [10.13, 15].

In nanotherapeutic applications, devices should be noninvasive and target therapeutic payloads to sites of disease to maximize therapeutic benefit while minimizing undesired side effects. This of course implies the existence of therapeutic effector functions in these nanodevices, to give devices the ability to remediate a physiologically undesirable condition. Beyond that, several attributes relate to sensing of biomolecules, cells, or physical conditions (sensing disease itself, identification of residual disease, and, potentially, targeting capacity, responding to intrinsic or externally supplied triggers for payload release). Other properties relate to communication between device subunits (for instance, between sensor and effector domains of the device)

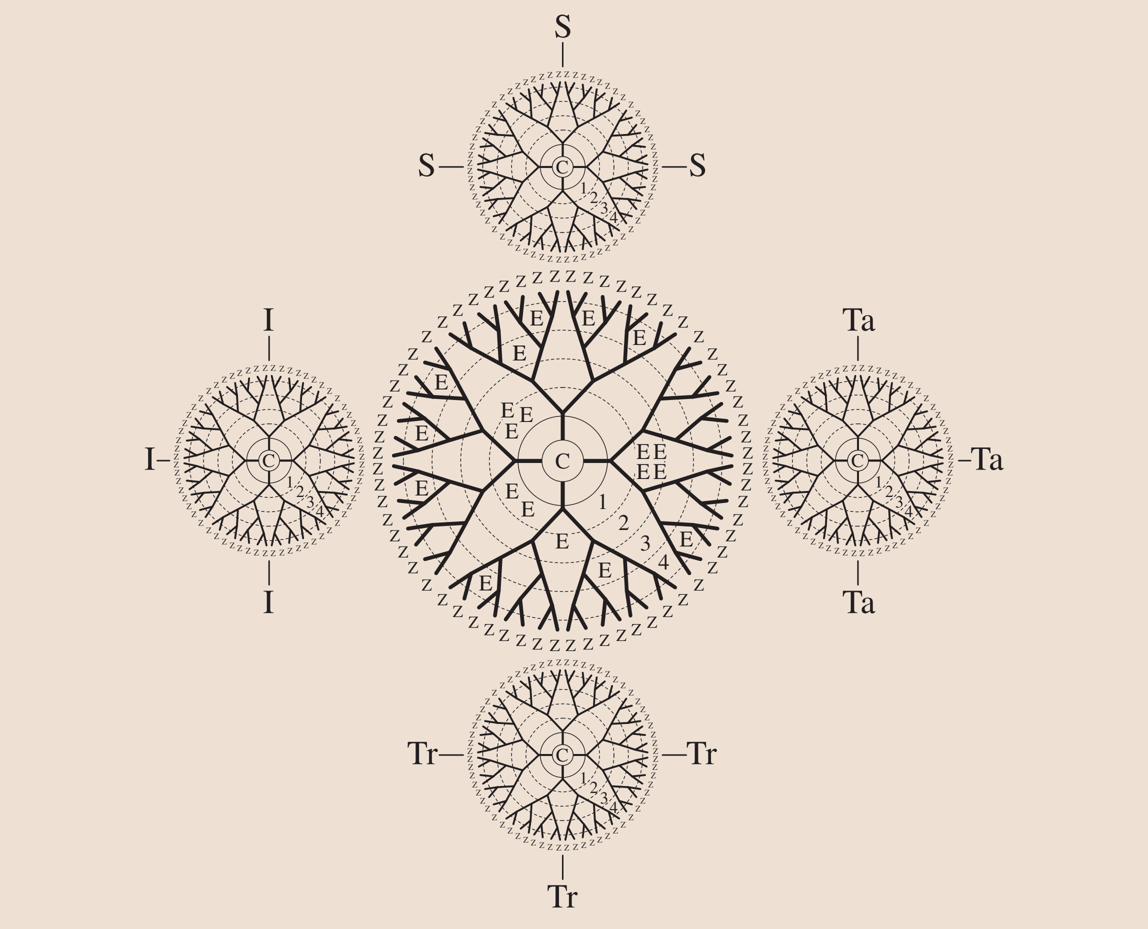

Fig. 10.3 A hypothetical, modular nanotherapeutic patterned after the dendrimer-based cluster agent for oncology of Baker [10.15]. As described in the text, each dendrimer subunit is grown from an initiator core (C), and the tunable surface groups of the dendrimers are represented by Z. Each dendrimer subunit has a specific, dedicated function in the device: the central dendrimer encapsulates small molecule therapeutics (E), whereas other functional components are segregated to other dendrimer components. These include biochemical targeting/tethering functions (Ta), therapeutic triggering functions to allow activation of prodrug portions of the device by an external operator (Tr), metal or other constituents for imaging (I), and sensing functions (S) to mediate intrinsically controlled activation/release of therapeutic. This design constitutes a therapeutic platform [10.13, 15] because of its modular design. The depicted device is only one possible configuration of an almost infinite number of analogous therapeutics that can be tuned to fit particular therapeutic needs by interchanging functional modules

or between the device and an external operator (external triggering and data documentation capability). With appropriate design, device functions can be modular [10.13, 15]. As discussed below, this approach allows construction of nanotherapeutic platforms (such as the dendrimer-based therapeutic of Fig. 10.3, dendrimer synthesis and assembly are discussed later), as opposed to single one-off devices that are capable of only one therapeutic task.

The vision of nanoscale therapeutic platforms arose from collaboration between the National Cancer Institute (NCI) and the National Aeronautics and Space Agency (NASA). NASA is concerned with minimal mass therapeutics: therapeutics, along with almost everything else used by astronauts, must be launched from Earth. NCI is interested in early detection of disease to improve prognosis. Since this requires screening a population predominantly of healthy patients, the screening mode must be minimally invasive. The considerations of both agencies might be met with ultra-small (micro- or nanoscale) multipotent therapeutic devices. Furthermore, the proposed therapeutic platforms should not only remediate undesired physiological conditions but also have the capacity to recognize them and report them. Extensive capability for molecular recognition and communication with external clinicians/operators is integral to the requirements of NASA and NCI. These capacities would allow drugs or other therapeutic interventions to be provided in a controlled fashion, to maximize benefit and minimize side effects, and is the essence of "smart" therapeutics (see the discussion of targeting and triggering below).

Though substantial progress has been made in device design and realization, no fully realized multipotent nanoscale therapeutic platform has yet been commercialized, but the therapeutic platform paradigm, in which devices are modular, with functional tasks segregated into individual modules, has potential to be extremely powerful. Classes of broadly similar devices could be tailored to specific disease states by interchanging modules (targeting modules, drugs dispensers, etc.) as appropriate to the disease state or therapeutic course. Thus, the hypothetical device represents a possible therapeutic platform composed of functional modules that could be for use in particular indications or in specific individuals.

10.1.2 Utility and Scope of Therapeutic Nanodevices

Therapeutic nanotechnology will be useful, of course, when the underlying biology of the disease states involved is amenable to intervention at the nanoscale. As we will discuss, while several disease states and physiological conditions (cancer, vaccination, cardiovascular disease, etc.) are particularly accessible to nanoscale interventions, some nanotechnological approaches may be applicable more broadly. Much as was the case with the introduction of recombinant protein therapeutics over the last 20 years, nanotherapeutics may present regulatory and pharmacoeconomic challenges related to their novelty and their cost of goods (COGs).

10.2 Synthetic Approaches: "top-down" versus "bottom-up" Approaches for Nanotherapeutic Device Components

Synthesis of nanomaterials is commonly thought of in terms of "top-down" or "bottom-up" processes. Top-down approaches begin with larger starting materials and, in a more or less controlled fashion (depending on the technique), remove material until the desired structure is achieved. Most microfabrication techniques for inorganic materials (lithography and milling techniques, etc.) fit this description. In contrast, bottom-up approaches begin with smaller subunits that are assembled, again with varying levels of control, depending on technique, into the final product. Key examples of materials made by bottom-up approaches include some inorganic structures. This includes "handmade" structures created by using direct atomic or molecular placement by force microscopy (discussed below), structures built using various deposition or growth methods, as well as all polymerization methods of synthesis. Thus, almost all biologic macromolecules and most biogenic structures, including mineralized biomaterials, are made by bottom-up methods. To make the distinction between top-down and bottom-up approaches more concrete, we will consider an example of each.

10.2.1 Production of Nanoporous Membranes by Microfabrication Methods: A top-down Approach

Lithography can be summarized by three basic steps: 1) Pattern Design (generation of masks), 2) Pattern Definition (exposure), and 3) Pattern Transfer (etching/liftoff). Optical lithography uses masks to form patterns on resist/substrate surfaces to produce features. The technique's power lies in its reproducibility and its capacity to manufacture via highly parallel processes. The key limitation of photolithography lies in the fact that resolution of features is diffraction-limited by the wavelength of light used. To address this limitation, short-wave radiation (i. e., X-rays with wavelengths of about 1nm wavelength) can be generated by synchrotron or other sources (from X-ray tubes, discharge plasma, or laser plasma), controlled, and focused for use in X-ray lithographic techniques [10.27]. The process is identical conceptually to optical lithography but requires special masks and resists amenable to the high-frequency radiation used. Combinations of filters and mirrors can produce resolutions in feature size of less than 100 nm, with fabrication throughputs congruent with those of other optical lithography processes. Limited access to synchrotron sources in turn limits the wide application of the method, however. The relatively long wavelengths used in conventional photolithography are generally unsuitable for formation of nanoscale features unless some clever technical expedient, like the use of a sacrificial layer (described below for generating nanoporous membranes) is employed.

As we will discuss in consideration of applications of nanotherapeutic devices, there are numerous potential applications for tunable nanopore membranes [10.16–18]. For certain applications, such as immunoisolation (see below), the distribution of pore sizes must be tight and nearly perfect. Until the late 1990s, this was an unattainable objective.

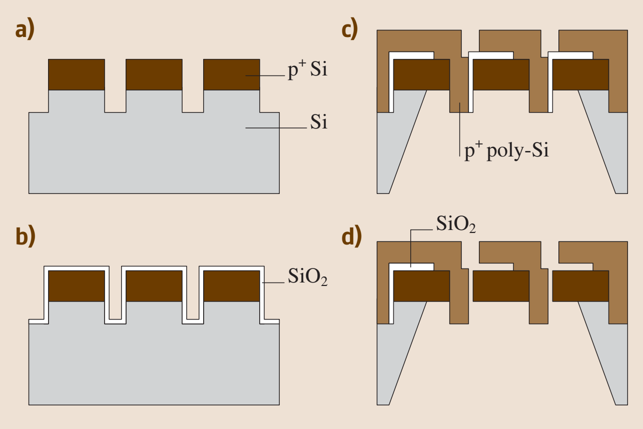

Fig. 10.4a–d Nanoporous silicon membranes are fabricated by a top-down approach [10.16–18]. **(a)** Pore exit holes are plasma-etched through p+ silicon (*dark brown*) and boronated silicon (*grey*). **(b)** Controlled depth silicon oxide (*white*) layer is grown on the structure. **(c)** A thick polysilicon layer (*light brown*) is deposited over the oxide, and pore exits are established by back-side plasma etch. **(d)** Pores are opened by removal of the sacrificial layer with hydrofluoric acid. The membranes themselves are macroscale objects that fit under the nanotechnology rubric because they derive their useful properties from controlled nanoscale architecture

The key technical innovation facilitating the microfabrication of highly defined nanoporous membranes was an approach featuring a sacrificial oxide layer sandwiched between two structural layers that ultimately is etched away to define the pore pathway and diameter [10.16–18]. The sacrificial layer (SiO_2) is sandwiched between a silicon wafer and a polysilicon layer: pore channel diameter is determined by the SiO_2 thickness. In the process (Fig. 10.4), the top of a silicon wafer is doped with boron to increase its mechanical robustness, and p+ silicon is overlaid. Pore exit holes are plasma etched through the p+ and doped material. In the pore dimension-determining step, SiO_2 is grown on the wafer by dry thermal oxidation, a process allowing control of oxide layer thickness to within 1 nm. This oxide constitutes the critical sacrificial layer. A thick polysilicon layer (again, boron-doped for mechanical strength) is then deposited over the oxide. Pore entry holes are etched by plasma etch, offset from the exit pores. Ultimately, the offset will require the diffusion pathway of the finished structure to pass through a “bottleneck” whose diameter is determined by the thickness of the sacrificial layer. The backside of the wafer is anisotropically etched to expose the doped layer of the wafer (now the bottom of the membrane structure). Pores are opened by removal of exposed sacrificial layer using concentrated hydrofluoric acid.

10.2.2 Synthesis of Poly(amido) Amine (PAMAM) Dendrimers: A bottom-up Approach

PAMAM dendrimers are remarkably defined synthetic molecules made using polymer chemistry (Fig. 10.5). Their unique structural attribute is their fractal geometry, and their unique physical property is their high monodispersity. Clever use of orthogonal conjugation strategies in their synthesis drives their monodispersity [10.28–30].

Orthogonal conjugation involves the mutually exclusive reactivity of the chemical specificities present in the reaction: when the orthogonal reactants are present in a vessel, and the reaction is carried at appropriate stoichiometry, only a monodisperse single product is formed. PAMAM dendrimers are among the most monodisperse synthetic materials available and, within limits, their sizes, surface chemistries, and shapes can be controlled at the synthetic level. To a greater or lesser extent, these desirable material properties of dendrimers are caused by an orthogonal synthetic strategy.

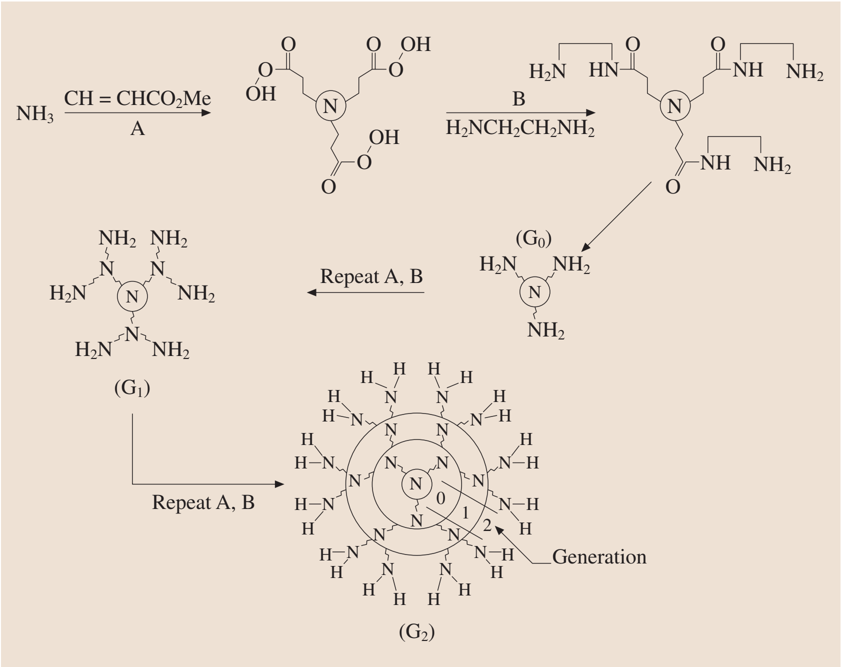

Fig. 10.5 PAMAM dendrimers are grown bottom-up using sequential orthogonal synthetic steps [10.28–30]. Dendrimers are grown from amine cores, in this case ammonia, by sequential addition of methyl acrylate, followed by an independent reaction addition reaction involving ethylene diamine. Together, these two half-steps constitute a full generation of dendrimer growth. PAMAM dendrimers are classified by their generation of growth as being G_0, G_1, G_2, etc., as indicated in the figure

PAMAM dendrimers are usually built from an initiator core (an amine) in sequential shells, called generations (i. e., G_0, G_1, G_2, etc.). Each generation comprises two synthetic half-steps, each of which is self-limiting in that the reagents performing each addition step are reactive only with the distinct chemical functionalities added to the growing dendrimer in the *previous* addition reaction. Reagent is added until the chemical functionalities from the previous step are fully consumed (Fig. 10.5). In the first synthetic half-step, methyl acrylate is added to the amine precusor, adding a carboxy-terminated functionality to all amine branch points. Polymer intermediates terminating in such carboxy functionalities are called "half-generation." In the second half-step, the generation is completed. Ethylene diamine is added, resulting in a branched amine-terminated adduct for each carboxylate group of the half-generation precursor. In general, before proceeding to downstream synthetic operations, products of individual reactions are purified so that only functionalities incorporated into growing dendrimer are available for subsequent reactions. Under ideal conditions, only a single chemical structure can result from each generation of growth, and with adequate purification between addition reactions; in the absence of limitations on the completeness of each reaction (such as steric limitations which become manifest in dendrimers of generation 6 and above), dendrimers of each generation are perfectly identical to each other.

10.2.3 The Limits of top-down and bottom-up Distinctions with Respect to Nanomaterials and Nanodevices

Some materials can be produced by alternate means, some of which are bottom-up, some of which are top-down approaches. For instance, carbon nanotubes can be synthesized in an arguably top-down approach from graphite sheets in an arc oven or can be grown bottom-up, by a metal-catalyzed polymerization method [10.31]. Additionally, not all finished materials can be classified as either top-down or bottom-up: synthetic protocols can contain both steps. For instance, while proteins are synthesized from lower molecular weight amino acid precursors by chemically or biologically mediated polymerization (bottom-up), they are often made as precursor molecules that are processed to a final product by chemical or enzymatic cleavage (top-down, see [10.32] for an overview). While the question of whether any given material is made top-down or bottom-up can be ambiguous, the synthetic provenance of multicomponent nanodevices can be even more so. In analogy to biogenic materials, synthetic polymers are bottom-up materials per se, but fabrication of raw polymeric materials into final device architectures often involves top-down steps [10.33]. As we will see, therapeutic nanodevices frequently contain both synthetic and biologic components, and strict top-down and bottom-up categorization of these devices is often not applicable.

10.3 Technological and Biological Opportunities

This section considers selected enablers for therapeutic nanodevices. Some are purely technological: nanomaterial self-assembly properties, bioconjugation methods, engineered polymers for conditional release of therapeutics, external triggering strategies, and so forth. Others relate to disease-state tissue or cell-specific biology that can be exploited by nanotherapeutics, such as the emerging vascular address system and intrinsic triggering approaches. In association with those applications, we will consider additional biological opportunities for nanoscale approaches specific to particular disease states.

10.3.1 Assembly Approaches

Assembly of components into devices is amenable to multiple approaches. In the case of devices comprising a single molecule or processed from a single crystal (some microfabricated structures, single polymers, or grafted polymeric structures) assembly may not be an issue. Integration of multiple, separately microfabricated components may sometimes be necessary (as in the immunoisolation capsule discussed below) and may sometimes drive the need for assembly, even for silicon devices. Furthermore, many therapeutic nanodevices contain multiple, chemically diverse components that must be assembled precisely to support their harmonious contribution to device function.

"One off" Nanostructures and Low Throughput Construction Methods

Direct-write technologies can obtain high (nanometer scale) resolution. For instance, electron-beam (e-beam) lithography is a technique requiring no mask, and that can yield resolutions on the order of tens of nanometers, depending on the resist materials used [10.34]. Resolution in e-beam lithography ultimately is limited by electron scattering in the resist and electron optics, and like most direct-write approaches, e-beam lithography is limited in its throughput. Parallel approaches involving simultaneous writing with up to 1,000 shaped e-beams are under development [10.34] and may mitigate limitations in manufacturing rate.

Force microscopy approaches utilize an ultrafine cantilever tip (typically with point diameters of 50 nm or less, Fig. 10.6) in contact with, or tapping, a surface or a stage. The technique can be used to image molecules, to analyze molecular biochemical properties (like ligand-receptor affinity [10.35]), or to manipulate materials at nanoscale. In the latter mode, force microscopy has been used to manipulate atoms to build individual nanostructures since the mid-1980s. This has led to construction of structures that are precise to atomic levels of resolution (Fig. 10.6), though the manufacturing throughput of "manual" placement of atoms by force microscopy is limited.

Dip-pen nanolithography (DPN) is a force microscopy methodology that can achieve high resolution features (features of 100 nm or less) in a single step. In DPN, the AFM tip is coated with molecules to be deployed on a surface, and the molecules are transferred from the AFM tip to the surface as the coated tip contacts it. DPN also can be used to functionalize surfaces with two or more constituents and is well suited for deployment of functional biomolecules on synthetic surfaces with nanoscale precision [10.36, 37]. DPN suffers the limitations of synthetic throughput typical of AFM construction strategies.

Much as multibeam strategies might improve throughput in e-beam lithography [10.34], multiple tandem probes may increase assembly throughput for construction methods that depend on force microscopy significantly, but probably not sufficiently to allow manufacture of bulk quantities of nanostructures, as will likely be needed for consumer nanotherapeutic devices. As standard of care evolves increasingly toward tailored courses of therapy [10.38] and individual therapeutics become increasingly multicapable and powerful, however, relatively low throughput synthesis/assembly methods may become more desirable. For the moment, though, ideal manufacturing approaches for nanotherapeutic devices resemble either industrial polymer chemistry, occurring in bulk, in convenient buffer

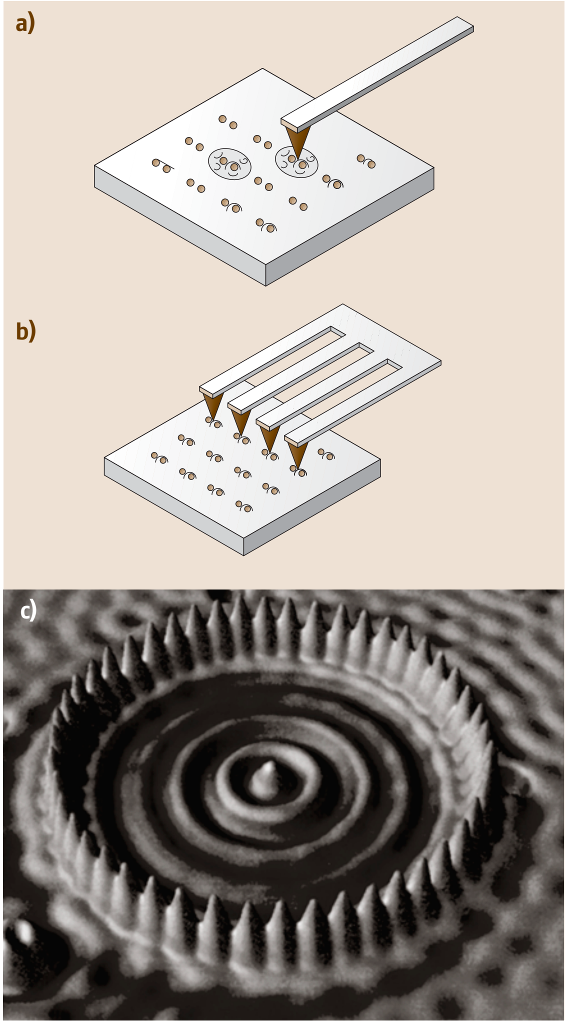

Fig. 10.6 **(a)** A schematic depiction of an atomic force microscope cantilever and tip interacting with materials on a surface. Tips typically have points of 50 nm or less in diameter [10.31, 39]. **(b)** Schematic of multiplexed AFM tips performing multiple operations in parallel [10.31, 39]. **(c)** AFM image of a quantum corral, a structure built using AFM manipulation of individual atoms (from the IBM Image Gallery)

systems, or in massively parallel industrial microfabrication approaches. In any case, therapy for a single patient may involve billions of billions of individual nanotherapeutic units, so each individual nanotherapeutic structure must require only minimal input from a human synthesis/manufacturing technician.

Self-Assembly of Nanostructures

Self-assembly has been long recognized as a potentially critical labor-saving approach to construction of nanostructures [10.40], and many organic and inorganic materials have self-assembly properties that can be exploited to build structures with controlled configurations. Self-assembly processes are driven by thermodynamic forces and generally result in structures that are not covalently linked. Intra/intermolecular forces driving assembly can be electrostatic or hydrophobic interactions, hydrogen bonds, and van der Waals interactions between and within subunits of the self-assembling structures and the assembly environment. Thus, final configurations are limited by the ability to "tune" the properties of the subunits and control the assembly environment to generate particular structures.

Self-Assembly of Carbon Nanostructures

Carbon nanotubes (Fig. 10.7) spontaneously assemble into higher order [10.31] structures (nanoropes) as the result of hydrophobic interactions between individual tubes. Multi-wall carbon nanotubes (MWCNT) are well-known structures that can be viewed as self-assembled, nested structures of nanotubes with tube diameters decreasing serially from the outermost to innermost tubes. The striking resemblance that MWCNT have to macroscale bearings has been noted and exploited [10.41]. MWCNT linear bearings can be actuated by application of mechanical force to the inner nanotubes of the MWCNT assembly. Actuation causes the assembly to undergo a reversible telescoping motion. Interestingly, these linear MWCNT bearings exhibit essentially no wear as the result of friction between bearing components.

C_{60} fullerenes and single wall carbon nanotubes (SWCNT) also spontaneously assemble (Fig. 10.7) into higher order nanostructures called "peapods" [10.42] in which fullerene molecules are encapsulated in nanotubes. The fullerenes of peapods modulate the local electronic properties of the SWCNT in which they are encapsulated and may allow tuning of carbon nanotube electrical properties. The potentially fine-level control of nanotube properties may prove useful in nanotube-containing electrical devices, particularly in cases wherein nanotubes are serving as molecular wires. In this capacity, carbon nanotubes have been incorporated into FETs (field effect transistors, discussed below under sensing architectures [10.43]), and other molecular electronic structures. Ultimately, these architectures may result in powerful, ultra-small computers to provide the intelligence of "smart," indwelling nanotherapeutic

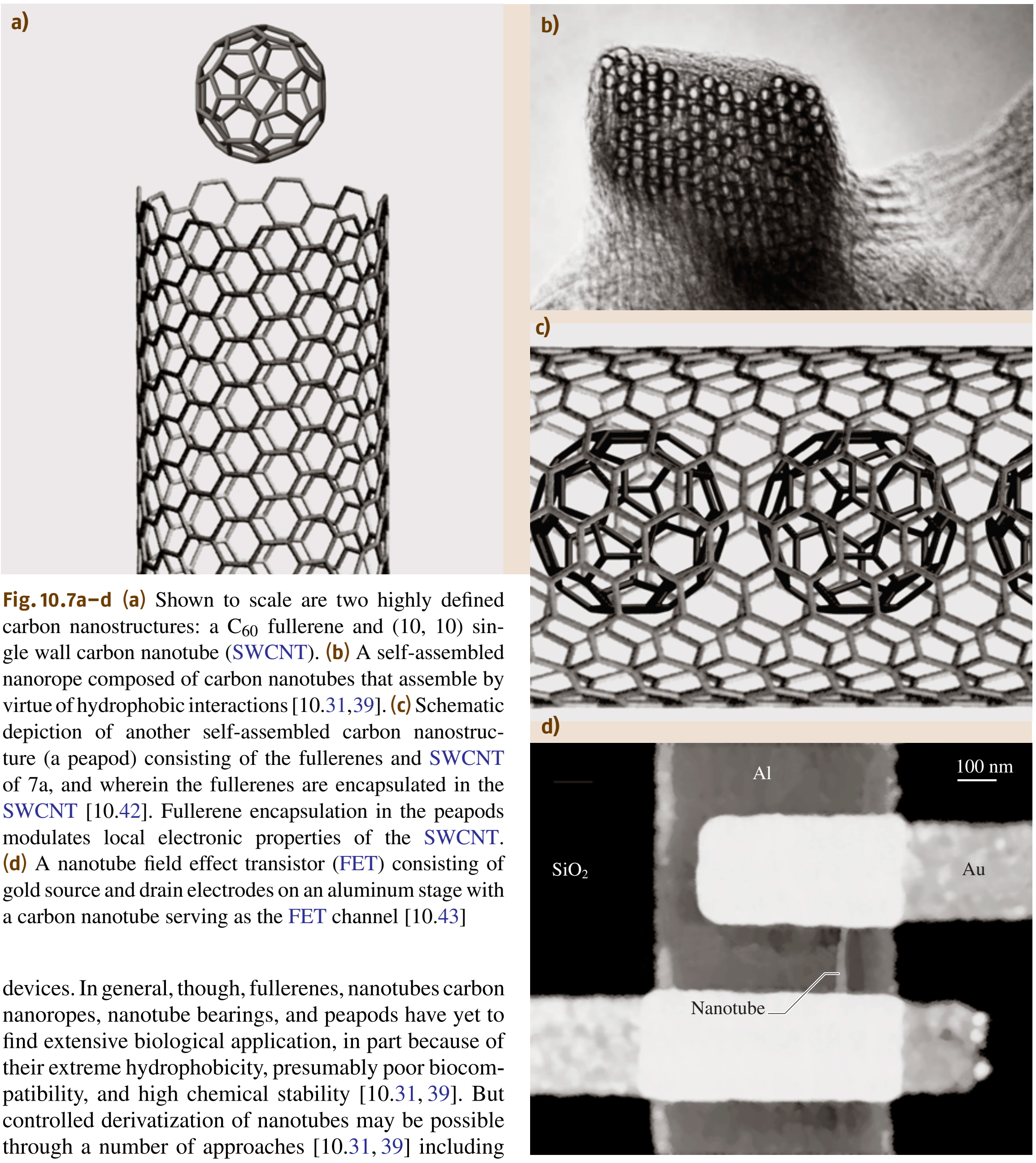

Fig. 10.7a–d **(a)** Shown to scale are two highly defined carbon nanostructures: a C_{60} fullerene and (10, 10) single wall carbon nanotube (SWCNT). **(b)** A self-assembled nanorope composed of carbon nanotubes that assemble by virtue of hydrophobic interactions [10.31,39]. **(c)** Schematic depiction of another self-assembled carbon nanostructure (a peapod) consisting of the fullerenes and SWCNT of 7a, and wherein the fullerenes are encapsulated in the SWCNT [10.42]. Fullerene encapsulation in the peapods modulates local electronic properties of the SWCNT. **(d)** A nanotube field effect transistor (FET) consisting of gold source and drain electrodes on an aluminum stage with a carbon nanotube serving as the FET channel [10.43]

devices. In general, though, fullerenes, nanotubes carbon nanoropes, nanotube bearings, and peapods have yet to find extensive biological application, in part because of their extreme hydrophobicity, presumably poor biocompatibility, and high chemical stability [10.31, 39]. But controlled derivatization of nanotubes may be possible through a number of approaches [10.31, 39] including controlled introduction of bond strain to render individual carbons of the tubes selectively chemically reactive (so-called mechanosynthesis). Carbon nanotubes and fullerenes, however, owe many of their remarkable properties (chemical stability, mechanical robustness, some electrical properties) to the fact that all the valences of the constituent carbon atoms (except those at the ends of nanotubes that are "open") are satisfied. Thus, to derivatize carbon nanostructures is to degrade them, a fact one must consider when the design purposes that drove incorporation of carbon structures into a therapeutic depends on the chemical perfection of the material.

Self-Assembly of Materials Made by Traditional Polymer Chemistry

In the realms of drug delivery and biomedical micro and nanodevices, the most familiar self-assembled structures are micelles [10.44–46]. These structures are formed from the association of block co-polymer subunits (Fig. 10.8), each individual subunit containing hydrophobic and hydrophilic domains. Micelles spontaneously form when the concentration of their subunits exceeds the critical micelle concentration (cmc) in a solvent in which one of the polymeric domains is immiscible (Fig. 10.8). The cmc is determined by the immiscible polymeric domain and can be adjusted by control of the chemistry and length of the immiscible domain, as well as by control of solvent conditions. Micelles formed at low concentrations from low-cmc polymers are stable at high dilution. Micelles formed from polymer monomers with high cmcs can dissociate upon dilution, a phenomenon that might be exploited to control release of therapeutic cargos. If desired, micelles can be stabilized by covalent cross-linking to generate shell-stabilized structures [10.44–47].

The size dispersity and other properties of micelles can be manipulated by control of solvent conditions, incorporation of excipients (to modulate polymer packing properties), temperature, and agitation. From the standpoint of size, reasonably monodisperse preparations (polydispersity of 1–5%) of nanoscale micellar structures can be prepared [10.44–47]. The immense versatility of industrial polymer chemistry allows micellar structures to be tuned chemically to suit the task at hand. They can be modified for targeting or to support higher order assembly properties. They can be made to imbibe therapeutic or other molecules for delivery and caused to dissociate or disgorge themselves of payloads at desired times or bodily sites under the influence of local physical/chemical conditions. The tunability of these and other properties at the level of monomeric polymer subunits as well as the level of assembled higher order structures make micelles potentially powerful nanoscale drug delivery and imaging vehicles.

Fractal materials, such as dendritic polymers, whose synthesis and structure were discussed previously, exhibit packing properties that can be exploited to assemble higher order aggregate structures called "tecto(dendrimers)" [10.48]. In fact, these self-assembly properties are being exploited in oncological nanotherapeutics as Fig. 10.3 shows [10.15, 49]. In principle, these self-assembling therapeutic complexes need not be pre-formed prior to administration. Individual functional modules of the therapeutic assembly might be administered sequentially, potentially to tailor therapies more precisely to individual patient responses.

Stoichiometric Control and Self-Assembly

As the preceding examples demonstrate, self-assembly approaches sometimes do not feature precise control of subunit identity and stoichiometry in the assembled complexes. This can be a limitation when the stoichiometry and relative arrangement of differentiable individual subunits is critical to device function. Stoichiometry is less an issue when the self-assembling components are identical and functionally fungible, as in the synthetic, peptidyl anti-infective illustrated in Fig. 10.9 [10.50,51].

In the anti-infective architecture, individual peptide components are flat, circular molecules. The planar character of the toroidal subunits is a consequence of the alternating chirality of alternating D-L amino acids (aas) in the primary sequence of the peptide rings. Alternating D and L aas is not possible in proteins made by ribosomal synthesis [10.32]. Ribosomes recognize and incorporate into nascent polypeptides only L amino acids, and so, as the result of aa chirality and bond strain, peptides made by ribosomes cannot be made flat, closed toroids like those of the peptidyl anti-infectives. Much as in α-helical domains of ribosomally synthesized proteins, however, the aa R-groups (which are of varying hydrophobic or hydrophilic chemical specificities [10.32]) are arranged in the plane of the closed D, L rings extending out from the center of the rings. Hydrogen bonds between individual rings govern self-assembly of the toroids into rod-like stacks, while the R-groups dominate interactions between multiple stacks of toroids and other macromolecules and structures (see below).

The planar toroidal subunits can be administered as monomers and self-assemble into multi-toroid rods at the desired site of action (in biological membranes). But the peptide toroids R-groups are chemically tuned so that the rod structures into which they spontaneously assemble intercalate preferentially in specific lipid bilayers (i. e., in pathogen vs. host membranes). Moreover, the assembled rods may undergo an additional level of self-assembly into multi-rod structures, spanning pathogen membranes [10.50, 51]. Whether as single rod or multiple rod assemblies, membrane intercalation by stacked toroids reduces the integrity of pathogen membranes selectively, and therefore particular toroid species exhibit selective toxicity to specific pathogens.

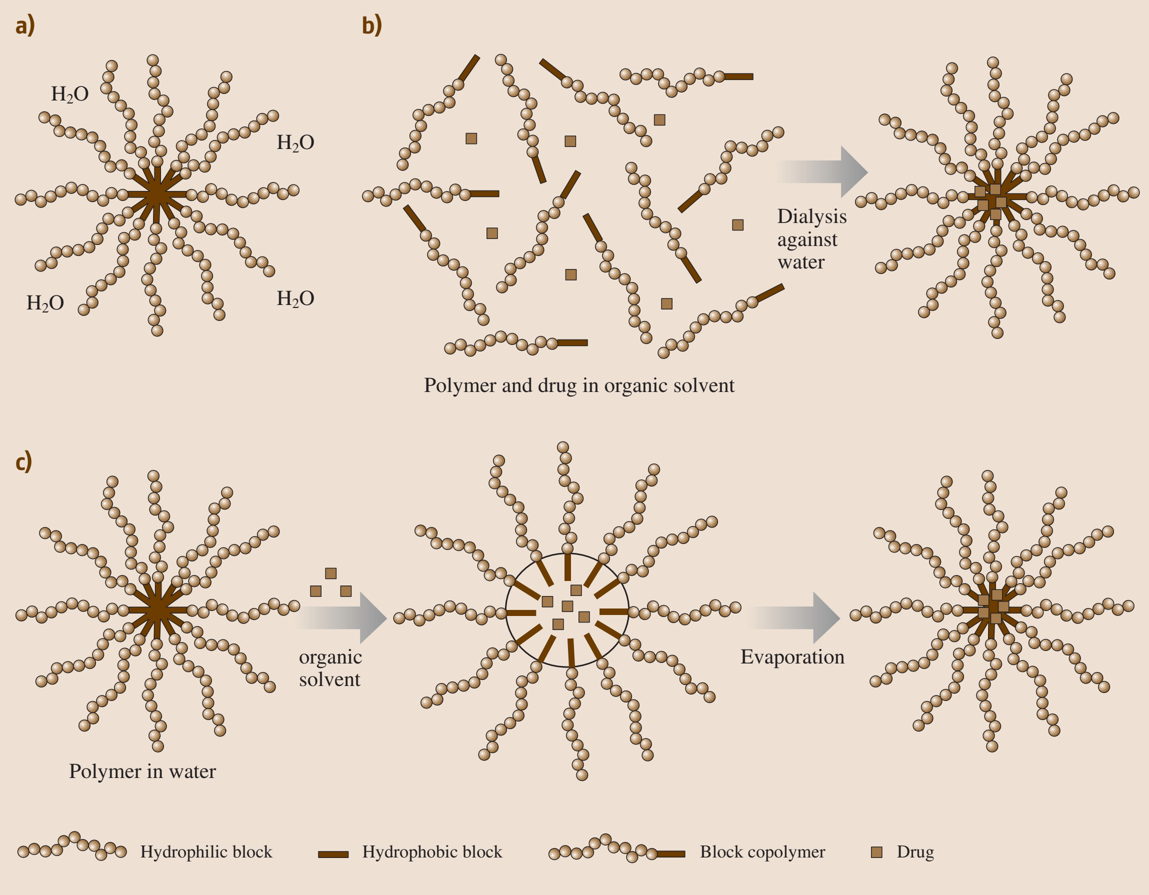
a)
H_2O
H_2O
H_2O
H_2O
b)
Dialysis against water
Polymer and drug in organic solvent
c)
organic solvent
Evaporation
Polymer in water
Hydrophilic block
Hydrophobic block
Block copolymer
Drug

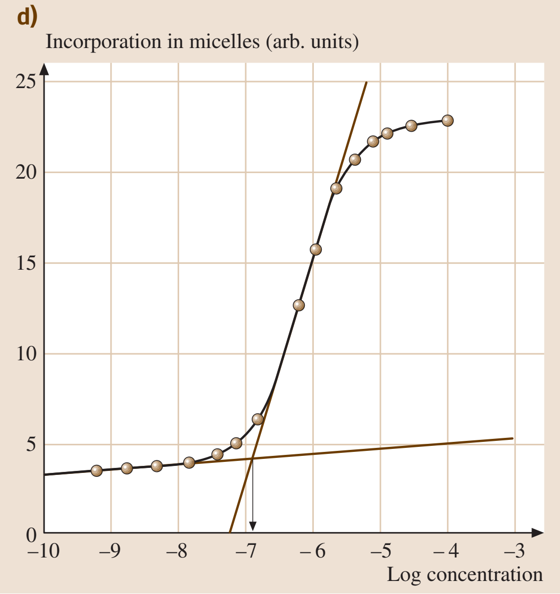

Fig. 10.8a–d Micellar drug delivery vehicles and their self-assembly from block copolymers [10.44–46]. **(a)** Morphology of a micelle in aqueous buffer. Hydrophobic and hydrophilic polymer blocks, copolymers containing the blocks, micelles generated from the block copolymers, and (hydrophobic) drugs for encapsulation in the micelles are indicated. **(b)** Micelle self-assembly and charging with drug occuring simultaneously when the drug-polymer formulation is transitioned from organic to aqueous solvent by dialysis. **(c)** Pre-formed micelles can be passively imbibed with drugs in organic solvent. Organic solvent is then removed by evaporation, resulting in compression of the (now) drug-bearing hydrophobic core of the micelle. **(d)** An illustration of concentration-driven micelle formation. At and above the critical micelle concentration (cmc), block copolymer monomers assemble into micelles, rather than exist as free block copolymer molecules. The *arrow* indicates the cmc for this system

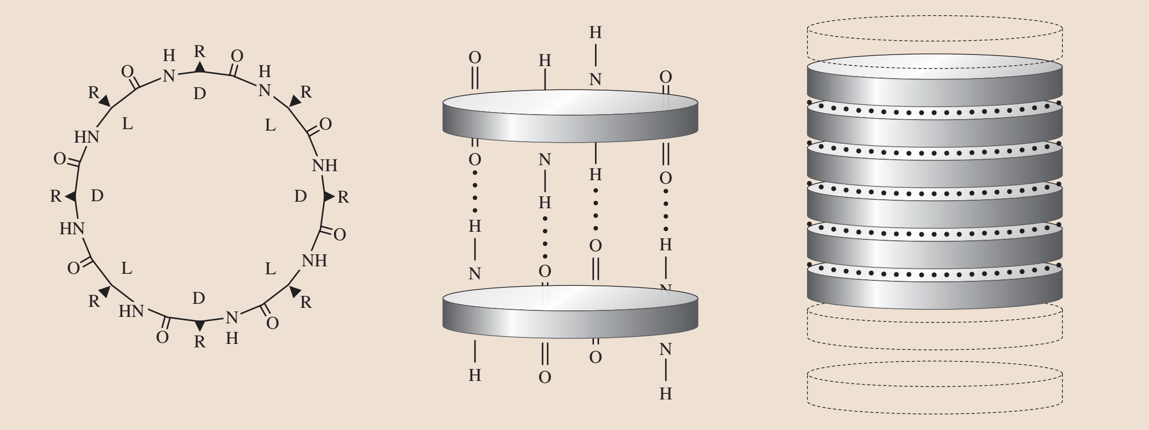

Fig. 10.9 A self-assembling peptide antibiotic nanostructure [10.50, 51]. Peptide linkages and the α-carbons and their pendant R groups are indicated. The synthetic peptide rings are planar as the result of the alternating chirality (D or L) of their amino acid (aa) constituents. R groups of aas radiate out from the center of the toroid structure. Individual toroids self-assemble (stack) as the result of hydrogen bonding interaction between amine and carboxy groups of the peptide backbones of adjacent toroids. The surface chemistry of multi-toroid stacks is tuned at the level of the aa sequence and, therefore, R group content of the synthetic peptide rings. The chemical properties of the stacked toroid surfaces allow them to intercalate into the membranes of pathogenic organisms with lethal consequences. The specific membrane preferences for intercalation of the compound are tuned by control of R group content of torroids

These toroidal, synthetic antibiotics, and other nanoscale antimicrobials represent critically needed, novel antibacterial agents. Resistance to traditional, microbially derived antibiotics often is tied to detoxifying functions associated with secondary metabolite synthesis; these detoxifying functions are essential for the viability of many antibiotic producing organisms (for instance, see [10.52]). The genes encoding such detoxifying functions are rapidly disseminated to other microorganisms, accounting for the rapid evolution of drug resistant organisms that has bedeviled antimicrobial chemotherapy for the last 25 years. Synthetic nanoscale antibiotics, like the peptide toroids [10.50, 51] and the N8N antimicrobial nanoemulsion [10.53], act by mechanisms entirely distinct from those of traditional secondary metabolite antibiotics, and no native detoxifying gene exists. Therefore, novel nanoscale antimicrobials may not be subject to the unfortunately rapid rise in resistant organisms associated with most secondary metabolite antibiotics, though this remains to be seen. As bacterial infection continues to re-emerge as a major cause of morbidity and mortality in the developed world, consequence of increasing antibiotic-resistant pathogens, novel nanoscale antibiotics will become more important.

Biomolecules in Therapeutic Nanodevices: Self-Assembly and Orthogonal Conjugation

Biological macromolecules undergo self-assembly at multiple levels, and like all instances of such construction, biological self-assembly processes are driven by thermodynamic forces. Some biomolecules undergo intramolecular self-assembly (as in protein folding from linear peptide sequences, Fig. 10.10). Higher order structures are, in turn, built by self-assembly of smaller self-assembled subunits (for instance, structures assembled by hybridization of multiple oligonucleotides, enzyme complexes, fluid mosaic membranes, ribosomes, organelles, cells, tissues, etc.).

Proteins are nonrandom copolymers of 20 chemically distinct amino acid (aa) subunits [10.32]. The precise order of aas (i. e., via interactions between aa side chains) drives the linear polypeptide chains to form specific secondary structures (the α helices and β sheet structures seen in Fig. 10.10). The secondary structures have their own preferences for association, which, in turn, leads to the formation of the tertiary and quartenary structures that constitute the folded protein structures. In its entirety, this process produces consistent structures that derive their biological functions from strict control of the deployment of chemical

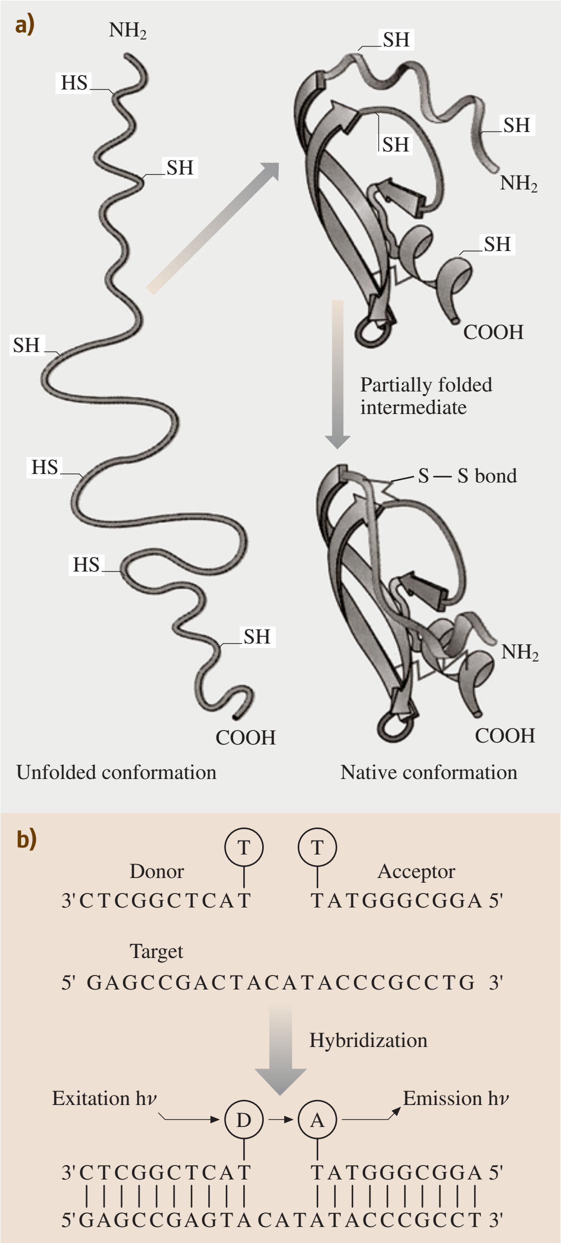

Fig. 10.10a,b Self-assembly and biological macromolecules **(a)** Linear peptide chains (with amino and carboxy ends, as well as sulfhydral groups of cysteine residues indicated) undergoe a multistep folding process that involves the formation of secondary structures (α-helices, indicated by heavy helical regions, and β-sheet regions, indicated by the heavy anti-parallel arrows) that themselves associate into a tertiary structure. Final conformation is stabilized by the formation of intrachain disulfide linkages involving cysteine thiol groups. **(b)** A fluorescence transfer device that depends on self-assembly of biomolecules. The device is composed of donor (D) and acceptor (A) molecules brought into close proximity (within a few angstroms) by the base pair hybridization of complementary oligonucleotides. When the structure is assembled, acceptor and donor are energetically coupled, and fluorescence transfer can occur [10.54–57]

specificities (the aa side chains) in three-dimensional space.

Biomolecules can be used to drive assembly of nanostructures, either as free molecules or conjugated to heterologous nanomaterials. For instance, three-dimensional nanostructures can be made by DNA hybridization [10.54–57]. Such DNA nanostructures can exhibit tightly controlled topographies but limited integrity in terms of geometry [10.57], due to the flexibility of DNA strands. Oligonucleotides, antibodies and other specific biological affinity reagents can also be used to assemble nanostructures (Fig. 10.10, see also targeting and triggering discussions below). Often the domains of biomolecules responsible for assembly and recognition are small, continuous, and discrete enough that they can be abstracted from their native context as modules and appended to other nanomaterials of interest to direct formation of controlled nanoscale architectures (for instance, see the SCFv antibody fragement of Fig. 10.1).

Several orthogonal bioconjugate approaches have arisen from the field of protein semisynthesis [10.20, 58, 59]. These protein synthetic chemistries allow site-specific conjugation of polypeptides to heterologous materials in bulk, as the result of conjugation between exclusively, mutually reactive electrophile-nucleophile pairs (analogous to dendrimer synthesis discussed above, see Table 10.2). They have been applied to the synthesis of multiple therapeutic nanodevices [10.2–5, 7, 14, 58, 59].

As described above, proteins are profoundly dependent on their three-dimensional shapes: chemical derivatization at critical aa sites can profoundly impact protein bioactivity. Because conjugation can be directed to pre-selected sites via orthogonal approaches, and since the sites of conjugation in the protein can be chosen because the proteins involved tolerate adducts at those positions, proteins coupled to nanomaterials by such orthogonal methodologies often retain their

	Electrophile		Nucleophile		Product
1.	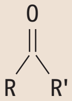	+	H_2N-O-R''	⟶	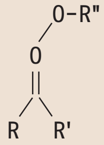
			Aminooxy		Oxime
2.	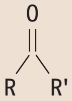	+	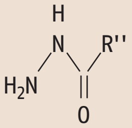	⟶	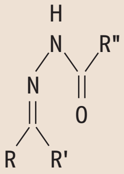
			Hydrazide		Hydrazone
3.	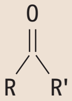	+	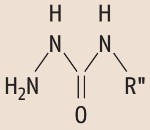	⟶	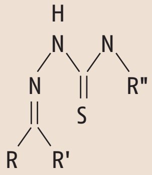
			Thiosemicarbazide		Thiosemicarbazone
4.	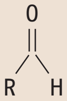	+	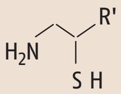	⟶	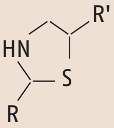
			β-Aminothiol		Thiazolidine
5.	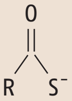	+	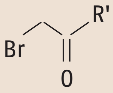	⟶	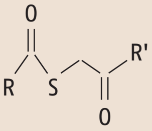
	Thiocarboxylate		α-Halocarbonyl		Thioester
6.	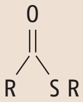	+	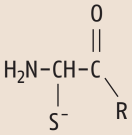 $H_2N-\underset{S^-}{CH}-\overset{O}{\overset{\Vert}{C}}-R$	⟶	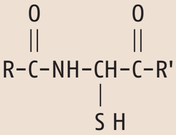 $R-\overset{O}{\overset{\Vert}{C}}-NH-\underset{SH}{CH}-\overset{O}{\overset{\Vert}{C}}-R'$
	Thioester		Cysteine		Amide

Table 10.2 Orthogonal conjugation chemistries originally derived for protein semi-synthesis but applicable to conjugation of proteins to nanomaterials [10.20, 58, 59] (R′ = H: Aldehyde, R = alkyl: Ketone)

biological activity. In contrast, protein bioactivity in conjugates generally is lost or profoundly impaired when proteins are coupled to nanomaterials using promiscuous chemistries. For instance, proteins such as cytokines (and other protein hormones) elicit their effects by interacting with a receptor, and a large fraction of their surface (20% or more) is involved in receptor binding, directly or indirectly. Promiscuous chemistries (1-ethyl-3-(3-diamethylaminopropyl) carbodiimide or EDC, conjugation, see [10.60]) used to conjugate cytokines to nanoparticles tend to inactivate hIL-3 and other cytokines whereas the same protein/particle bioconjugates retain bioactivity if judiciously chosen orthogonal conjugation strategies are used (Lee & Parthasarathy, unpublished). Proteins for which only a small portion of their surfaces contribute to the interesting portions of their bioactivities (from the standpoint of the nanodevice designer), such as some enzymes or intact antibodies (Fig. 10.1) may be somewhat less sensitive to promiscuity of the bioconjugate strategy used [10.20, 58, 59], but the benefits of orthogonal conjugation strategies can also apply to these protein bioconjugates [10.20, 58, 59]. The potential utility of orthogonal conjugation for incorporation of active biological structures into semi-synthetic nanodevices is becoming more fully recognized and cannot be overestimated [10.58, 59].

10.3.2 Targeting: Delimiting Nanotherapeutic Action in Three-Dimensional Space

Delivery of therapeutics to sites of action is a key strategy to enhance clinical benefit, particularly for drugs useful within only narrow windows of concentration because of their toxicity (i. e., drugs with narrow therapeutic windows). Diverse targeting approaches are available, ranging from methods exploiting differential extravasation limits of vasculature of different tissues (see the discussion of oncology below), sizes, and surface chemistry preferences for cellular uptake (see discussion of vaccines below); preferential partition of molecules and particles into specific tissues by virtue of their charges, sizes, surface chemistry, or extent of opsonization (see below); or the affinity of biological molecules decorating the nanodevice for counter-receptors on the cells or tissues of interest.

The Reticuloendothelial System and Clearance of Foreign Materials

Physical properties such as surface chemistry and particle size can drive targeting of nanomaterials (and presumably nanodevices containing them) to some tissues. For instance, the pharmacokinetic (P_k) and biodistribution (B_d) properties [10.32] of many drugs and nanomaterials are driven by their clearance in urine, which is in turn governed by the filtration preferences of the kidney. Most molecules making transit into urine have masses of less than 25 to 50 kilodaltons (kDa; 25–50 kDa particles corresponding *loosely* to effective diameters of about 5 nm or less) and are preferably positively charged; these parameters are routinely modulated to control clearance rates of administered drugs. Clearance of low molecular weight (nano)materials in urine can be suppressed by tuning their molecular weights and effective diameters, typically accomplished by chemical conjugation, to polymers such as poly(ethylene glycol) [10.19]. Polymer conjugation (pegylation) has been applied to many different materials and may provide some degree of charge shielding. Pegylation also increases effective molecular weights of small materials above the kidney exclusion limit, diverting them from rapid clearance in urine.

Coating foreign particles with serum proteins (opsonization [10.61]) is the first step in the clearance of foreign materials. Opsonized particles are recognized and taken up by tissue dendritic cells (DCs) and specific clearance organs. These tissues (thymus, liver, and spleen, constituting the organs of the reticuloendothelial system or RES) extract materials from circulation by both passive diffusion and active processes (receptor-mediated endocytosis). Charge-driven, receptor-mediated uptake of synthetic nanomaterials occurs in the RES and can result in partition of positively charged nanoparticles into the RES. For instance, PAMAM dendritic polymers exhibit high positive charge densities related to the large number of primary amines on their surfaces [10.28–30]. In experimental animals, biodistribution of unmodified PAMAM dendrimers is limited nearly exclusively to RES organs [10.62]. This unfavorable biodistribution can be modulated by "capping" the dendrimers (i. e., derivatizing the dendrimer to another chemical specificity, such as carboxy or hydroxyl functionalities [10.15, 62]).

Despite legitimate applications of targeting to the kidney and the RES (for instance in glomerular disease [10.63]), intrinsic targeting to clearance sites is of interest primarily as a technical problem that impedes therapeutic delivery to other sites. In such cases, numerous targeting strategies are available, some of which depend on synthetic nanomaterial properties (for instance, see discussion of the enhanced permeability and retention or EPR effect in the context of oncology, below) to minimize uptake of nanotherapeutic devices by clearance systems and maximize delivery to desired sites. Targeting via biological affinity reagents decorating the surfaces of therapeutic nanodevices may be the most direct approach.

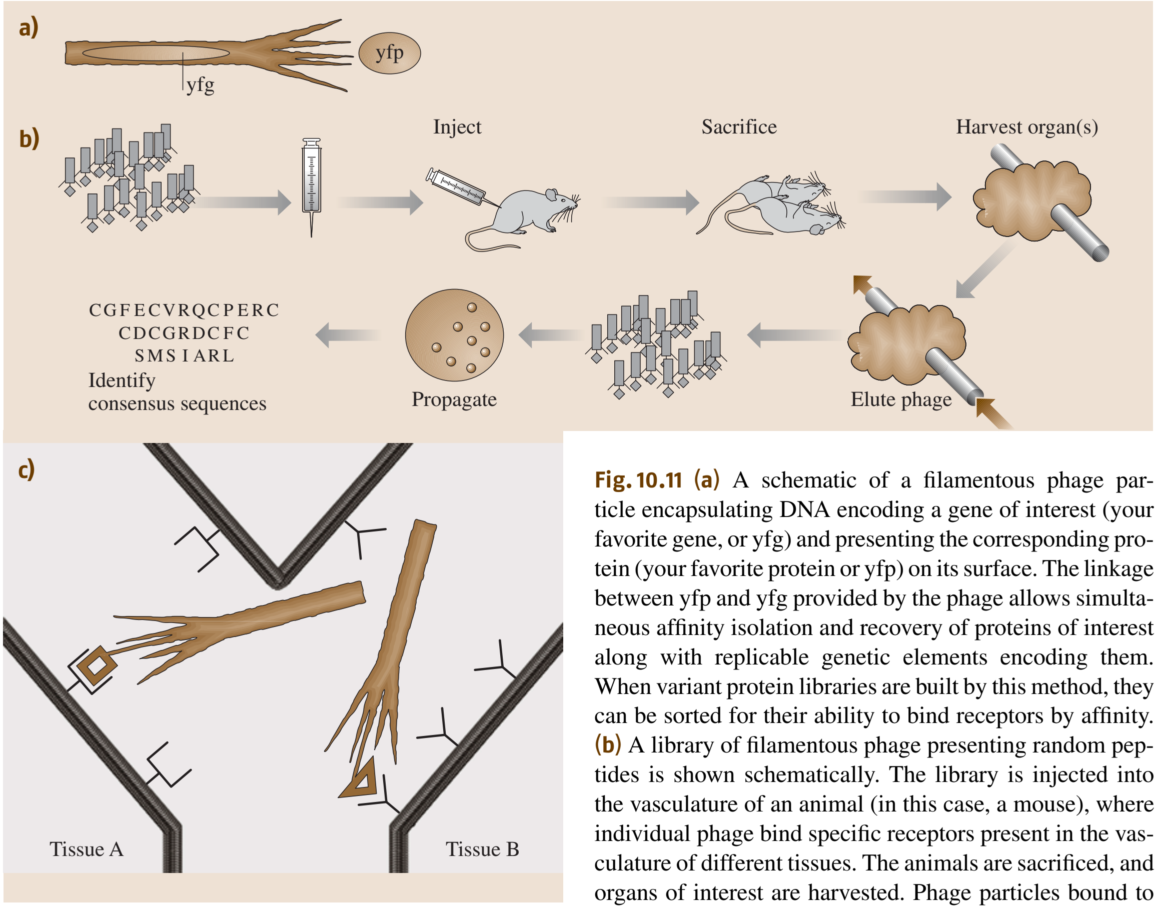

Fig. 10.11 **(a)** A schematic of a filamentous phage particle encapsulating DNA encoding a gene of interest (your favorite gene, or yfg) and presenting the corresponding protein (your favorite protein or yfp) on its surface. The linkage between yfp and yfg provided by the phage allows simultaneous affinity isolation and recovery of proteins of interest along with replicable genetic elements encoding them. When variant protein libraries are built by this method, they can be sorted for their ability to bind receptors by affinity. **(b)** A library of filamentous phage presenting random peptides is shown schematically. The library is injected into the vasculature of an animal (in this case, a mouse), where individual phage bind specific receptors present in the vasculature of different tissues. The animals are sacrificed, and organs of interest are harvested. Phage particles bound to receptors of vasculature of the organ at hand can be eluted by any of several methods (low pH elution is schematically shown here). *E. coli* is infected with the eluted phage particles and clonally propagated, allowing identification of the peptides encoded by the eluted phage, which mediated binding to the vascular receptors. Ultimately, this allows the identification of consensus peptide sequences that recognize specific receptors on the vasculature of the organ or tissue involved. These receptors constitute the molecular addresses for these organs, and the peptides isolated from phage eluted from them are biological affinity reagents that can be used to direct nanotherapeutic devices to vasculature of particular organs. **(c)** Specific recognition of vascular beds of two distinct tissues (A and B) via recognition of organ-specific vascular addresses (the Y and box-shaped receptors shown on vascular endothelium) by specific peptides (the *triangle* and the *square*) presented on the phage surface is depicted [10.64–66]

Nanotherapeutic Targeting Exploiting Biological Affinity Properties

Tissue-specific delivery by biological affinity requires the presence of tissue-specific surface features, most commonly proteins or glycoproteins (tissue-specific antigens). Historically the search for tissue-specific antigens for drug targeting, whether associated with tumors or other cells, organs or tissues, has been arduous and not entirely gratifying. The primary problems are specificity (few antigens are uniquely present in any single tissue), availability (some tissues may not have their own unique antigenic signature or marker), and therapeutic extravasation or directed migration in tissue spaces (markers in tissue may not be accessible from vasculature). An exciting recent development in biochemical targeting is the discovery of a vascular address system [10.64–66].

The vascular address system has been characterized by administering a peptide phage display to library animals, resecting individual organs, and extracting phage from the vasculature of the isolated organs (Fig. 10.11 [10.64–66]; see [10.67] for a discussion of display technology). Amazingly, phage isolated from different organs exhibited distinct consensus presented peptide sequences, indicating that the vasculature of individual organs presented unique cognate receptors, each bound by a different short (ten amino acids or fewer) consensus peptide sequence that had been affinity selected from the phage display library. Furthermore, the affinity-selected peptides have the capacity to tether nano- to microscale particles to the site of their cognate receptors (as illustrated by the binding phage particles presenting the peptides to specific vascular locations). Site-specific drug delivery using the vascular address system has already been demonstrated [10.68,69]: it has further been used to target an apoptotic (cytocidal) agent to the prostate and to direct destruction of the organ in an animal model [10.68].

Mapping of the vascular address system is currently underway [10.64] and holds the promise of specific delivery of therapeutic agents to vasculature of specific organs. It remains to be seen whether each organ has a single molecular marker constituting its address that is amenable to binding a single peptide sequence; organs may instead have unique constellations of antigenic markers. If so, specific targeting may be possible using multiple peptides, each peptide binding its cognate receptor on target organ vasculature very weakly. Peptides used in such a multivalent targeting strategy would be chosen to reflect the unique constellation of address markers present in the target tissue. Affinities of cognates for linear peptides are often very low [10.67], though their aggregate affinity may be substantially higher than that of any peptide-vascular address cognate alone. Under ideal conditions, the affinity of such a multipeptide, multicognate complex should be the equivalent of the products of the affinities of each constituent peptide for its individual constituent cognate. Such multivalent interaction avidities can be extremely high (and the corresponding effective affinity constants are also high) but seldom fully realize their theoretical maximums (Lee & Parthasarathy, unpublished; see [10.32] for a review of receptor biochemistry).

It should be noted that most of the vascular addresses identified to date deliver materials to the organ vasculature: extravasation and access of organ tissue spaces by nanotherapeutics remains a separate issue.

10.3.3 Triggering: Delimiting Nanotherapeutic Action in Space and Time

Controlled triggering of therapeutic action is the other side of the targeting coin. If the site and time of nanotherapeutic delivery cannot be adequately controlled, the site of therapeutic action can be delimited by spatially- or temporally-specific triggering. The triggering event might drive release of active therapeutic from a reservoir, or chemical or physical processing of drug materials from an inert to an active form (inert administrations that are converted to active form at a specific time or place drugs which are activated by a chemical reaction occurring their sites of action are called prodrugs, Fig. 10.12). Three major triggering strategies are widely used: external stimuli, intrinsic triggering, and secondary signaling (multicomponent systems). Triggering strategies require nanotherapeutic delivery devices to be sensitive to a controlled triggering event, or a spatially/temporally intrinsic triggering event mediated by the host. Obviously, the triggering event itself must be tolerable to the patient.

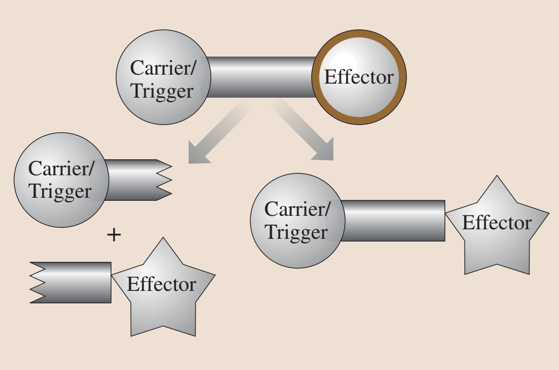

Fig. 10.12 One possible configuration of a prodrug, in which carrier/trigger and effector functions are separate functional domains of the therapeutic. The key feature of prodrugs is that they are therapeutically inert (as indicated by the colorless effector domain) until an activation event occurs (mediated here through the carrier/trigger domain, with effector activation indicated by its change to a star shape). Activation events can involve cleavage of inhibitory carrier/trigger domains from effectors. Other activation strategies involve a chemical change or shift in conformation of the effector, mediated through the carrier/trigger domain and in response to an environmental condition

Nanotherapeutic Triggering Using External Stimuli

External stimuli are provided by an external nanodevice operator/clinician, usually in the form of a site-specific energy input, typically light, ultrasound, or magnetic or electrical fields. Organic polymeric structures are very amenable to interaction with these energy sources. For instance, micellar structures can be reversibly dissociated with ultrasound, in which case they disgorge their contents or expose their internal spaces to the environment during ultrasound pulses (20 to 90 kilohertz range). The process has been used to control release of cytotoxin (doxorubicin) from micelles (Fig. 10.13), and short ultrasound transients might be used for pulsatile or intermittent exposure of patients to therapeutics [10.70].

Light is another popular external triggering modality. Bioactive materials can be covalently associated to a nanoscale-delivery vehicle by photo-labile linkages [10.15, 71, 72], or micelles can be constructed so that their permeability is altered as the result of exposure to light [10.72]. In the latter case, light input can cause photopolymerization resulting in micelle compaction that drives release of therapeutic cargo or photo-oxidation, which causes loss of micellar integrity to release encapsulated materials. The ability of light to penetrate dense tissues is a clear limitation to this approach. This concern can be accommodated by polymer systems responsive to wavelengths that penetrate tissue efficiently (usually in the red-infrared region of the spectrum), or by use of systems such as fiber optics to deliver light to deep tissues [10.15, 71, 72].

Externally applied magnetic fields can also be used to control nanotherapeutic activity. For instance, eddy currents induced by alternating magnetic fields can heat nanometallic particles and their immediate vicinity, an approach that has been successfully applied to control the bioactivity of individual biological macromolecules [10.73]. Colloidal gold particles are covalently conjugated to biomolecules (nucleic acid, or NA, duplexes or proteins), and alternating magnetic

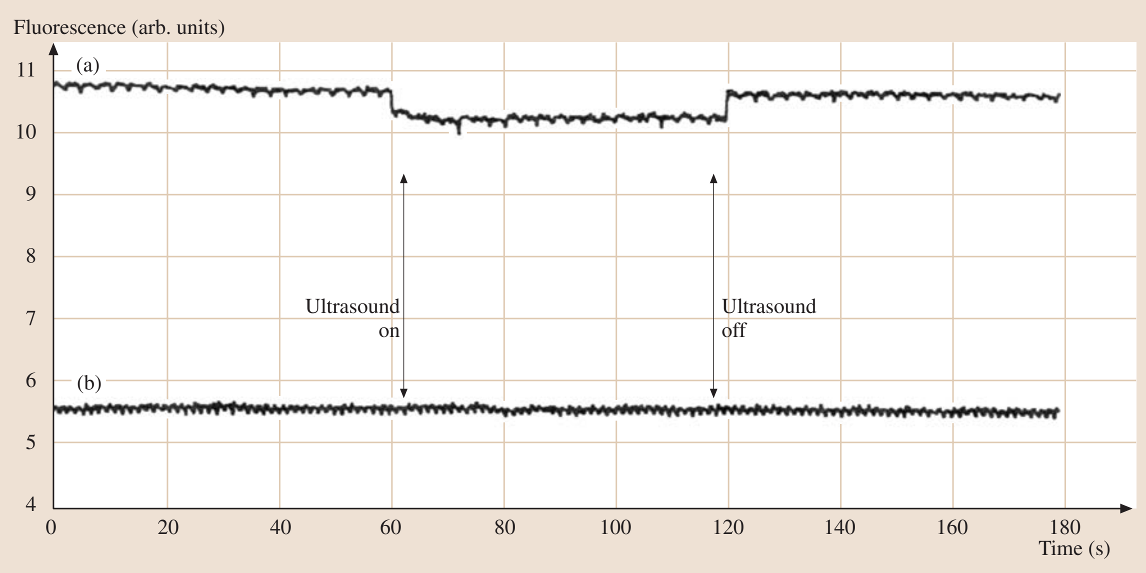

Fig. 10.13 Externally applied stimuli can trigger drug release: hydrophobic drug molecules are reversibly deployed from micelles in response to acoustic stimuli [10.70]. Intact pluronic micelles maintain the hydrophobic cytotoxin doxorubicin in a hydrophobic environment (the core of the micelle, as in Fig. 10.4). Doxorubicin fluorescence is quenched in aqueous environments, and hence changes in integrity of pluronic micelles carrying doxorubicin can be monitored by doxorubicin fluorescence. Here, an ultrasound pulse is applied to suspensions of such micelles or to doxorubicin in solution. Application of ultrasound triggers exposure of micelle-encapsulated doxorubicin to the aqueous solvent (i. e., drug release), as demonstrated by the reduction in fluorescence. After cessation of the acoustic pulse, doxorubicin is repackaged into the micelles, as evidenced by the increase in fluorescence after the ultrasound administration (*Line* a). Ultrasound has no impact on fluorescence of doxorubicin in solution, as expected (*Line* b)

fields are used to induce heating sufficient to cause dissociation of hybridized NA structures or denaturation of protein three-dimensional structures. Within certain parameters, the process is reversible and so allows the construction of semi-biological assembly or release switches. The modality has clear application to temporally specific triggered release, but the current inability to direct magnetic fields to pre-selected tissue locations may limit its use for spatially specific triggering.

Nanotherapeutic Triggering by Intrinsic Physiological Conditions

Intrinsic triggering is a prodrug strategy that depends on the conditions at the desired site of action to trigger the activity of the nanotherapeutic. Much as the three-dimensional conformation of, and therefore the activity of, proteins can be controlled in external triggering strategies [10.73], proteins can be engineered to make their conformations sensitive to intrinsic conditions at their desired sites of activity (see [10.3] for a discussion of protein engineering for use in nanobiotechnological devices). For instance, noncovalent complexes of a diphtheria toxin (DT) protein variant and a bispecific antibody (Fig. 10.1) have been engineered for specific toxicity to cells that can take up the complexes [10.75]. The bi-specific antibody recognizes both a cell-specific surface receptor as well as the DT in an inactive (nontoxic) configuration. Antibody binding to the receptor results in receptor-mediated endocytosis of the complex in target cells. As a result of the lower pH of the endosomal compartment (reaching pH 4–5 in endosomes, as opposed to the constant pH 7–8 in circulation) into which the bound complex is taken up, the DT variant protein undergoes a three-dimensional conformational shift to its toxic form. This conformation of DT is not recognized by the bispecific antibody, so the toxin dissociates from the complex and kills the cell. Hypoxia-triggered prodrugs have also been developed that are activated either by chemical reduction or by enzymatic activities induced in hypoxic tissues. Extreme hypoxia [10.76] is a unique feature of neoplastic tissue, so these strategies have clear application in oncology. Clinical manipulation of the extent of tumor oxygenation is possible [10.77] and may also be possible in other tissues. If so, redox state-dependent triggering approaches may have applications beyond cancer (Fig. 10.14a).

Enzyme-activated delivery (EAD [10.74], Fig.10.14) is another prodrug-like triggering strategy in which the properties of a nanoscale drug-delivery vehicle are altered at the site of action by an enzymatic activity endogenous to that site. Most typically, this involves a liposomal or micellar nanostructure from which designed pendent groups can be cleaved by metabolic enzymes (alkaline phosphatase, phospholipases, proteases, glycosidic enzymes, etc.) that are highly expressed at the site of therapy. Nanostructures are designed so that enzymatic cleavage of the pendent group causes a conformational or electrostatic change in the polymeric components of the delivery device, rendering the micellar structure fusogenic, leaky, or causing partial or complete dissociation of the structure [10.12] (Fig. 10.14c). This has the result of delivering or releasing therapeutic payloads at pre-selected sites.

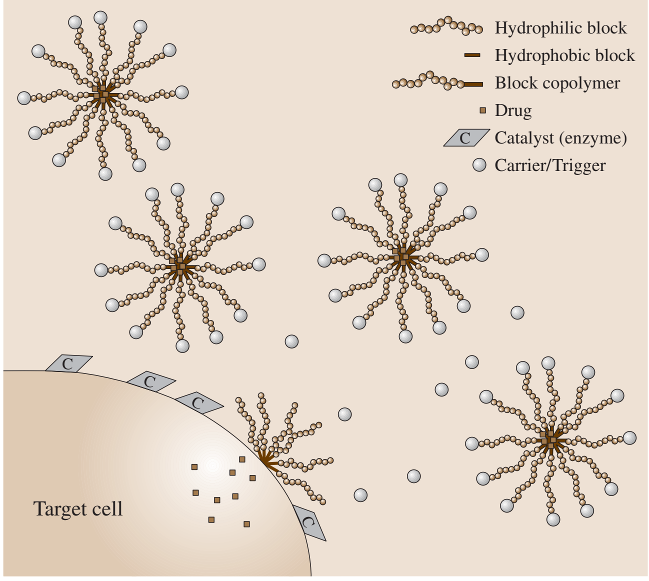

Fig. 10.14 Intrinsic conditions at desired sites of therapeutic action can trigger release of active drug molecules. Enzyme activated delivery (EAD) is an intrinsic triggering strategy that depends on chemical cleavage of the carrier/trigger to induce changes in the properties of the delivery vehicle to allow delivery of drug [10.74]. Here a surface functionality is cleaved from a micellar drug delivery vehicle, rendering the micelle fusogenic and allowing it to fuse with the cell membrane. As a result, the therapeutic payload of the micelle is delivered to the cytosol of target cells. Tuning of micellar surface functionalities and their linkages to the block copolymers constituting micelles theoretically can be used as target delivery to a variety of intracellular compartments

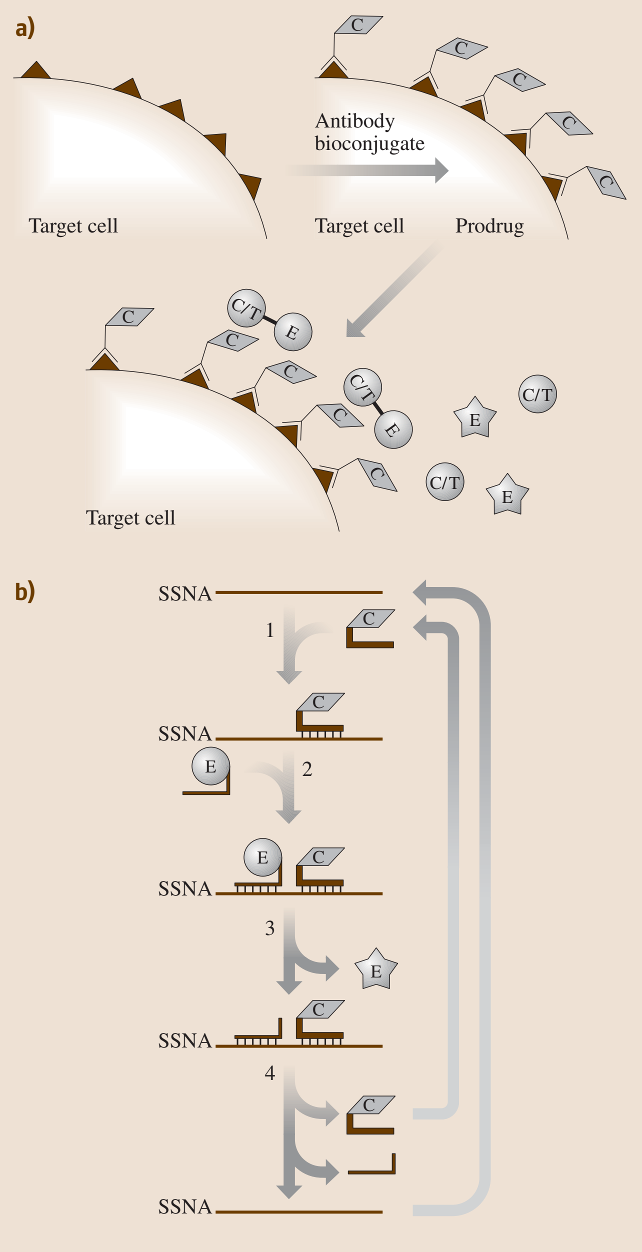

Fig. 10.15a,b Secondary biochemical signals can trigger release of active therapeutics. **(a)** Antibody directed enzyme-prodrug therapy (ADEPT) exploits enxyme-monoclonal antibody (mAB) bioconjugates to activate prodrugs [10.78]. Here, a mAB (like the native IgG of Fig. 10.1) is covalently liked to a catalyst (C in rhombus, usually an enzyme). The binding specificity of the monoclonal antibody causes the bioconjugate to bind to a desired target site (possibly by recognition of a vascular address, as shown in Fig. 10.11). The catalyst (enzyme) specificity of the bioconjugate is capable of triggering activation of a prodrug, much as in Fig. 10.12. The corresponding prodrug is then administered (represented as in Fig. 10.12). The enzyme localized at the desired target site of therapeutic action then activates the prodrug (represented here by cleavage of carrier/trigger moiety and change in shape of the effector to a star, as in Fig. 10.12). **(b)** Another secondary signaling triggering strategy [10.55]. Here, single stranded nucleic acid (SSNA) drives assembly of an activating complex for a prodrug. A catalyst (indicated again by C) is conjugated to an oligonucleotide that hybridizes to a single stranded nucleic acid (SSNA). A prodrug, containing an effector domain (E) and a second oligonucleotide, in this case homologous to a region of the same SSNA, and proximal to the site of hybridization of the first oligonucleotide, is also administered. Catalyst specificity is tuned to correspond to linkage between the prodrug effector and oligonucleotide. The catalyst in the assembled structure is sufficiently close to the prodrug to cleave the linkage between effector and oligonucleotide, releasing active drug (star). Potentially, prodrug and catalyst oligonucleotides might dissociate from the SSNA and the cycle could be repeated. Note the similarity to Fig. 10.10

Nanotherapeutic Triggering Using Secondary Signaling

Secondary signaling or multicomponent delivery systems are more complex delivery strategies (Fig. 10.15). These systems can feature site-specific and systemic delivery of one or more components. One class of such systems is the so-called ADEPTS devices (antibody directed enzyme-prodrug therapy, Fig. 10.15 [10.78]). ADEPTS systems feature affinity-based targeting of an enzymatic prodrug activator to a desired site, followed by systemic administration of a prodrug. ADEPTS is used primarily for oncology, and so the prodrug involved is generally one that can be activated to cytotoxicity. Since prodrug dosing is systemic, prodrug toxicity must be de minimus in the inactive form, and the activating catalyst must not be present at locations where cytotoxicity would be deleterious. To address this issue, the nanotherapeutic designer can chose to engineer an activation mechanism that has no physiological analogue in the host. In addition, tuned, high lability of the active drug can potentially delimit the site and extent of cytotoxic effect mediated by ADEPTS systems.

Other secondary signaling systems rely on DNA hybridization to assemble therapeutic components into catalytically active complexes at their sites of action [10.55]: in this system, the individual components lack therapeutic activity until they are brought within a few nanometers of each other by hybridization to a target single-stranded nucleic acid molecule (SSNA in Fig. 10.15, potentially a RNA species expressed in a tissue or cell type of interest). At this point, the catalytic moiety and the prodrug moieties are sufficiently close in space to allow catalytic cleavage of the active component from the prodrug. As depicted, this strategy makes no provision for delivery of drug and catalyst bioconjugates to the site of interest, nor for their transit across biological membranes. The system is similar to the DNA hybridization-driven fluorescence transfer system of Fig. 10.10.

Alternatively, secondary signaling systems can exploit competitive displacement of therapeutics from a carrier structure [10.79]. In this case, a non-covalent complex of engineered antibody and plasminogen activator (PA) is tethered to blood and fibrin clots by antibody affinity for fibrin. PA is released from the bound complex (to dissociate the clot) using bolus systemic administration of a nontoxic binding competitor for PA to the antibody complex. The strategy establishes a high local concentration of PA at clot sites, efficiently dissolving clots and potentially minimizing systemic side effects.

Layering Strategies for Fine Control of Nanotherapeutic Action

It should be clear that many of these approaches may be broadly applicable to trigger events other than drug release or to drive assembly or disassembly of therapeutic nanostructures in situ. It should also be clear that these approaches are often complementary, and that multiple approaches can be used in single nanotherapeutic devices (Fig. 10.3, Table 10.1 [10.11–13, 15, 71, 78, 80]). Layering targeting and triggering approaches tends to make devices more complex, but it also allows clinician/operators to intervene at multiple points in therapy, potentially leading to finer control of the therapeutic process and better clinical outcomes.

10.3.4 Sensing Modalities

The need for "smart" therapy is a key theme of therapeutic nanotechnology and pharmacology as a whole. Drugs with narrow therapeutic windows should be delivered only to their desired site of action and be pharmacologically active only when that activity is needed. These strategies can limit undesired secondary effects of therapy, some of which can be debilitating or life threatening, as discussed elsewhere in this chapter. One possible approach to this issue is the incorporation of sensing capability (specifically, the capacity to recognize appropriate contexts for therapeutic activity) into nanotherapeutic devices. Sensing capability may allow self-regulation of a therapeutic device, reporting to an external clinician/device operator (though this begins to touch on imaging applications, see below), or both. In the context of our discussion of therapeutic nano- and microscale devices, we will consider primarily electrical and electrochemical sensor systems, particularly microfabricated (Field Effect Transistor or FET, and cantilever) and conducting polymer sensors.

Sensor Systems

Sensing is predominantly a higher-order device functionality, depending on multiple device components, though one could argue that some targeting/triggering strategies, particularly targeting by bioaffinity and intrinsic triggering strategies, must, a priori, incorporate at least limited sensing capability. But biosensors, as they are typically considered, are multifunctional, multicomponent devices [10.81]. Usually a biosensor system is composed of signal transducer, sensor interface, biological detection (bioaffinity) agent, and an associated assay methodology, with each system component governed by its inherent operational considerations.

The transducer component determines the physical size and portability of the biosensor system. Signal transducers are moieties that are sensitive to a physical-chemical change in their environment and that undergo some detectable change in chemistry, structure, or state as the result of analyte (the thing to be sensed) recognition. Analytes for nanotherapeutic application could be biomolecules, like proteins, small molecules (organic or inorganic), ions (salts or hydrogen ions), or physical conditions (such as redox state, temperature). Interfaces are the sensor components that interact directly with the analyte. For sensor use in nanotherapeutic devices, immobilized or otherwise captured biological molecules (proteins, nucleic acids) often constitute the sensor interface. Whatever the chemical nature of the interface, it determines the selectivity, sensitivity, and stability of the sensing system and also is a dominant determinant of sensor operational limits. Assay methodology determines the need (or lack thereof) for analyte tracers, the number of analytical reagents, and the complexity and rapidity of the sensing process. Nanotherapeutics are of interest at least in part because they can be minimally invasive, low complexity, yet robust and accurate; convenient assay methods are therefore highly desirable.

Cantilever Biosensors

Micromechanical cantilevers (discussed above in conjunction with force microscopy, Fig. 10.6) transduce sensed events by mechanical means [10.82]. Both changes in the resonant frequency and deflection of cantilevers resulting from analyte binding or dissociation can be conveniently and sensitively detected. These changes in cantilever state can be conveniently detected by optical, capacitive, interferometry, or piezoresistive/piezoelectric methods, among others. Microfabricated cantilever dimensions range from micron to sub-micron range, with potential for further dimensional optimization (by carbon nanotubes appended to them, for instance [10.82]). They are operationally versatile and can be used in air, vacuum, or liquid, although they suffer some degradation in performance in liquid media. Like most micromachined structures, they can be batch fabricated and conveniently multiplexed.

Cantilevers used in atomic force microscopy (AFM, see also discussion of one-off nanostructures above) approaches can be used to study individual biomolecular interactions [10.35, 81]. In this approach, cantilever tips are derivatized with biomolecules (effectively, one member of a receptor/counter-receptor pair), and the tip-bound biomolecule is allowed to bind its counter-receptor (itself bound on a surface). Under non-equilibrium conditions, (i. e., conditions that result in thermodynamically irreversible changes in analyte molecular structures), the force required to disrupt single molecular interactions can be measured and related to classical biochemical parameters of receptor binding. The method has been applied to interactions between hormones and their receptors, sugars and lectins, as well as hybridizing DNA strands.

Cantilever systems sensitively detect changes in mass at their surfaces: changes as small as a mass density of $0.67\,\mathrm{ng/cm^{-3}}$ are theoretically detectable. This can allow detection of binding of extremely small objects to the cantilever and has been applied to detection and enumeration of prokaryotic and eukaryotic cells, as well as small numbers of macromolecules [10.82]. The incorporation of biological receptors or affinity reagents on the cantilever surface can drive specific binding events for particular sensing tasks. Microcantilevers are also highly sensitive to temperature, detecting changes as low as $10^{-5}\,\mathrm{K}$; they can also detect small changes in pH.

Field Effect Transistor Biosensors

Field effect transistor (FET) architectures are another sensing architecture that can be conveniently produced by micro-nanofabrication. FETs consist of a current source, a current drain, a conductive path (sensing channel) between them, and a sensing gate to which a bias can be applied. Analyte binding to the sensing channel induces a charge transfer resulting in a dipole between the surface and the underlying depletion region of the semiconductor: current that passes between the source and drain of a semiconductor FET is quite sensitive to the charge state and potential of the surface in the connecting channel region. Moving a standard silicon FET from depletion to strong inversion (i. e., shifting the surface potential by $>\sim 0.5\,\mathrm{eV}$) requires less than $\sim 10^{-7}\,\mathrm{C/cm^2}$ or $\sim 6\times10^{12}\,\mathrm{charges/cm^2}$, corresponding to transfer of $6.25\times10^{11}\,\mathrm{e/cm^2}$. With FETs of 2,000 square micrometers, detection of biological analytes in sub-nanomolar concentrations is easily feasible. Specificity for binding of macromolecular analytes of interest can be provided by deployment of biological affinity reagents in the FET sensing channel. Submicron FETs are routinely manufactured; use of carbon nanotubes in FETS will offer still greater miniaturization [10.43, 83].

Carbon nanotubes also have excellent mechanical properties and chemical stability in addition to potentially tunable electrical properties, making them highly desirable electrode/nanoelectrical materials for any number of nanoelectrical applications [10.84]. Biomolecules can be bound to carbon nanotubes, particularly in FET and nanoelectrode applications. Most biomolecules bound to carbon nanotubes are not covalently bound (as discussed above) and do not exhibit direct electrical communication with the nanotube, though redox enzymes bound to nanotubes and other conductive nanomaterials may [10.84, 85]. Flavin adenine dinucleotide (FAD) and flavoenyzme glucose oxidase (Gox) both display quasi-reversible one electron transfer when absorbed onto unannealed carbon nanotubes in glassy carbon electrodes. Gox so immobilized retains its substrate-specific (glucose) oxidative activity, leading to applications in sensing circulating glucose for diabetes and, perhaps, to a strategy of harvesting electrical power from metabolic energy.

Conducting Polymers and Sensor Biocompatibility

Biocompatibility of most metallic structures (as might be used in the bioelectrical sensors described) is limited at best; metal structures rapidly foul with serum proteins (i. e., become opsonized), undergo electrochemical degradation, or have other problematic properties. Polymer chemistry, however, has the capacity to tune composition to enhance biocompatibility properties and is commonly used to make synthetic surfaces more biologically tolerable. Electrically conductive polymers are potentially attractive in this context.

Polymeric materials are available with intrinsic conductivities comparable to that of metals (up to $1.5 \times 10^7\,(\Omega\,\mathrm{m})^{-1}$, which, by weight, is about twice that of copper [10.86]). Significant conductivity has been documented for a dozen or so polymers, including polyacetylene, polyparaphenylene, polypyrrole, and polyaniline, doped with various impurities. Careful control of doping can tune electrical conductivity properties over several orders of magnitude [10.87].

Conductive polymeric materials may be extremely well suited to biosensing applications. The polymeric materials themselves are often compatible with proteins and other biomolecules in solution. Furthermore, polymers are amenable to very simple assembly of sensor transducer-interface components by deposition of the polymer and trapped protein (sensor interface) directly on a metallic micro- or nanoelectrode surface [10.87,88]. Combined with a facile nanoelectrode array microfabrication method [10.89], simultaneous conducting polymer-interface protein deposition may offer an extremely simple way to fabricate multiplexed sensor arrays.

These electrochemical and electrophysical sensing modalities are of little use in and of themselves: they must communicate with either other device components (such as drug-dispensing effector components) or with external observers or operators. Sensor coupling in autonomously operating devices produced by microfabrication can be done directly, and the coupling linkages incorporated into the fabrication protocol. Coupling with biological device components can also be direct, as when conductive materials are conjugated to biomolecules, and can directly modulate their activity [10.73].

Sensor-device coupling can also be indirect, through electrochemically produced mediator molecules [10.85]. Sensors might be independent of nanotherapeutics and report conditions at the site of therapeutic action to an external operator, who would use any one of the external triggering strategies discussed above to engage therapy when and where appropriate. Communication/sensor interrogation might be accomplished most crudely by direct electrical wiring of sensors to an external observation station. Alternatively, if the event to be detected is transduced optically, as in colorimetric smart polymers that change optical properties in the presence of analyte [10.90], sensors might be interrogated by fiber optics. These modalities are most applicable to sensor arrays delivered to the site of interest by a catheter. The use of a catheter may be justified in some therapeutic applications, but it is an invasive procedure and is not optimal.

We have already seen multiple examples of prodrugs that are activated by cleavage of an inhibitory domain from the complex. One noninvasive approach to communication with external operators could exploit this phenomenon by detecting the cleaved fragment in bodily fluids. If the cleaved moiety cleared through urine, the extent of drug activation could be monitored noninvasively via urinalysis. There is no a priori need to connect the cleavage event to drug activation. For instance, an operator might administer a catalyst that cleaves a detectable material from a nanotherapeutic. If this secondary signaling moeity was independently targeted to the desired site of therapy, presence of the detectable cleaved product would provide information regarding the bodily location of the therapeutic nanodevice. The cleaved product would not be detectable unless the therapeutic and secondary signaling moiety co-localized at a single site. Other sophisticated communication involves ultrasound or electro-magnetic radiation to carry information. Application of these sorts of modalities currently occurs in in vivo imaging approaches.

10.3.5 Imaging Using Nanotherapeutic Contrast Agents

Imaging is a minimally invasive procedure that allows visualization of organs and tissues following the administration of a detectable moiety (contrast agent). The contrast agent is then exposed to some condition that interacts with the contrast agent so as to produce an emission or response detectable to an external monitoring device. Nano-sized particles (5–100 nm in diameter) have found application as contrast-enhancing agents for medical imaging modalities such as magnetic resonance imaging (MRI, our primary topic, reviewed in [10.91, 92]). MRI currently provides cross-sectional and volumetric images with high spatial resolution (< 1 mm) and is potentially applicable to many clinical purposes, though enhanced imaging capabilities are desirable. For instance, blood flow measurements of healthy and diseased arteries can be quantified better by the aid of improved contrast agents that highlight blood flow at the vessel wall. Similarly, some tissues (certain tumors) do not exhibit strong contrast within the MRI field and cannot be readily identified or characterized at present. It is of great interest, therefore, to develop nanoscale particles that provide enhanced contrast for many applications.

Magnetic Resonance Imaging (MRI): the Basics

Objects to be imaged are exposed to a strong magnetic field and a well-defined radio frequency pulse. The external magnetic field (B_0) serves to loosely align protons either with (lower energy level) or against (high energy level) the field, the difference between the two energy levels being proportional to B_0. Once the protons are separated into these two populations, a short multi-wavelength burst (or pulse) of radio frequency energy is applied. Any particular proton will absorb only the frequency that matches its particular energy (the Larmor frequency). This *resonance* absorption is followed by the excitation of protons from the low to high energy level and of equivalent protons moving from high to low energy levels. After the radio frequency pulse, protons rapidly return to their original equilibrium energy levels. This process is called relaxation and involves the release of absorbed energy. Once equilibrium is again established, another pulse can be applied.

Data is collected by positioning a receiver perpendicular to the transmitter: relaxation energy release induces a detectable, quantifiable signal (i. e., in amplitude, phase, and frequency) at the receiver coil. Since multiple protons in multiple chemical environments are involved, the signal at the receiver includes many frequencies. The received signal consists of multiple, superimposed signals (called the free induction decay or FID) signal, resulting from the relaxation of multiple, chemically distinct protons. FID is converted from the time domain to frequency domain (by Fourier transformation), within which individual proton types can be identified. Iterations of this procedure in two or three dimensions can create a high-resolution image of anatomical cross section or volume.

Intrinsic factors affecting image quality include the proton density of the tissues, local blood flow, and two relaxation time constants [10.93]: longitudinal relaxation time (T_1) and transverse relaxation time (T_2). Control of T_1 and T_2 relaxation effects are most critical for high-resolution MR images. T_1 relaxation measures energy transfer from an excited proton to its environment. In tissues, protons of fats and cholesterol molecules (relatively movement-constrained macromolecules) relax efficiently after a pulse and exhibit a short T_1 time. Water in solutions (a small molecule tumbling relatively freely in solution) has a much longer T_1 time. T_2 relaxation measures the duration of coherency between resonating protons after a pulse, prior to their return to equilibrium. Tightly packed, solid tissues with closely interacting hydrogen nuclei relax more quickly than loosely structured liquids, so tissues such as skeletal muscle have a short T_2s, while cerebrospinal fluid has a very long T_2. Intrinsic factors are manipulated through extrinsic factors, such as the external magnetic field strength, the specific pulse sequence, etc., allowing the collection of meaningful MR images.

Nanoparticle Contrast Agents

Signal intensity in tissue is influenced linearly by proton density, while changes in T_1 or T_2 result in exponential changes in signal intensity. T_1 and T_2, therefore, are manipulated to enhance imaging by administration of exogenous contrast-enhancing agents. MRI contrast agents are divided into paramagnetic, ferromagnetic, or superparamagnetic materials. Metal ion toxicity is an unfortunate consequence of physiologic administration of contrast agents but can be mitigated somewhat by complexation of the metals with organic molecules.

Paramagnetic metals used for enhanced MRI contrast (gadolinium, Gd, iron, Fe, chromium, Cr, and manganese, Mn) have permanent magnetic fields, though the magnetic moments of individual domains are unaligned [10.94]. Upon exposure to an external magnetic field, individual domain moments become aligned, generating a strong local field (up to 10^4 gauss [10.95]). Paramagnetic metal ions interact with water molecules, causing an enhanced relaxation of the water molecules via tumbling of the water-metal complex, dramatically decreasing the T_1 value for the water molecules and enhancing the proton signal [10.94]. Contrast enhancement by paramagnetics is thus due to the indirect effect the contrast agent has on water and its magnetic resonance properties.

Ferromagnetic and superparamagnetic materials both contain iron (Fe) clusters, which generate magnetic moments 10 to 1,000 times greater than do individual iron ions. Clusters greater than 30 nm in diameter are ferromagnetic, whereas smaller particles are superparamagnetic [10.93]. Ferromagnetic materials maintain their magnetic moment after the external field is removed, but superparamagnetic materials lose their magnetic field after the field is removed, as do paramagnetics. Both ferromagnetic and superparamagnetic substances minimize the proton signal by shortening T_2 [10.96], resulting in negative contrast (i. e., darkening of the image [10.97]).

Nanobiotechnological Contrast Agent Design

First generation contrast agents often contained a single metal ion/complex, whereas emerging agents incorporate nanoscale metal clusters, crystals, or aggregates, sometimes encapsulated within a synthetic or biopolymer matrix or shell [10.98–100]. These metal cluster agents improve contrast effects and, hence, output MR images profoundly. Furthermore, surface chemical groups (from the matrix or shell) can be derivatized to improve biocompatibility or allow targeting to a tissue or site of interest.

Typically particles are prepared from colloidal suspension where metallic cores are thoroughly mixed with the matrix material before being aggregated out of solution with a non-solvent. Dextran (a polymer of 1,6-β-D-glucose) is a typical matrix material used in commercial imaging reagents: Combidex is coated with 10,000 molecular weight dextran [10.99, 101, 102], Feridex has an incomplete, variable dextran coating [10.102], and Resovist is coated with carboxydextran [10.97]. Other polymeric materials are also used (oxidized starch [10.98, 100]), and matrix-less particles are also produced [10.103].

Nanoparticle contrast agents must be purified under tightly controlled conditions (generally by centrifugation or high pressure liquid chromatography [10.104]) and accurately characterized (for size dispersity by light scattering, chromatography, photon correlation, or electron microscopy [10.100, 104]) to assure the reproducibility of imaging agent production lots. Elemental analysis, X-ray powder diffraction, and Mössbauer spectroscopy have also been used to characterize metallic cores [10.103, 105]. Tight definition of the finished particles ensures accurate correlation between structural properties of imaged materials and prior in vitro and in vivo studies, allowing collection and interpretation of meaningful images. Metallic cores generally range from 4 to 20 nm in diameter [10.106], while coated particles can be up to 100 nm or more in diameter [10.105]. Compared to larger cores, however, nanoparticles less than 20 nm in diameter exhibit considerably longer blood half-life and improved T_1 and T_2 relaxivity effects [10.107].

As previously discussed, clearance of nanoparticles via the RES is a critical problem usually approached by surface modifications to mitigate nonspecific adsorption of proteins to the biomaterial surface (i. e., the opsonization of synthetic materials). Neutral, hydrophilic surfaces tend to adsorb less serum protein than hydrophobic or charged surfaces. Bisphosphonate and phosphorylcholine-derived thin film coatings have been applied to nanoparticles to stabilize iron oxide particles against pH, opsonization, and aggregation [10.108]. Such thin films do not fully eliminate protein adsorption, and dense layers or thick brushes of polysaccharides or hydrophilic polymers may more effectively avoid opsonization [10.109].

Other, as yet incompletely understood, biological factors influence the use of nanoparticles in vivo. For instance, a direct correlation between the circulating half-life of nanoparticles (i. e., their $T_{1/2}$s) and age of animals used has been reported [10.106]. The observed increase in $T_{1/2}$ may be correlated to age-related changes in phagocytic activity [10.106]. Local environments also influence particle stability: iron oxide particles are degraded at a pH of 4.5 or less [10.100], a condition sometimes attained in some intracellular vesicles.

First generation contrast agents were primarily blood-pool agents that moved freely through the entire vasculature. Targeting contrast-enhancing nanoparticles to sites of interest can reduce heavy metal toxicity associated with the commonly used agents by diminishing the dose required to obtain an acceptable image. Contrast agent targeting can also provide enhanced diagnostic information. For instance, nanoparticles that bind to molecular fibrin at a clot site on a vessel wall have been developed [10.110], potentially allowing differentiation between vulnerable and stable atherosclerotic plaques. Similarly prognostically valuable data regarding disruptions of the blood–brain barrier (BBB) have been visualized with MRI as the result of delivery of contrast-enhancing nanoparticles to the affected site [10.102].

10.4 Applications for Nanotherapeutic Devices

As discussed above, nanotherapeutic devices are novel, emerging therapeutics with properties not fully understood or predictable. Nanotherapeutics, therefore, must be justifiable on at least two levels. As we have seen, the nature of the therapeutic task and the state of current nanoscale-materials technology make the incorporation of biological macromolecules unavoidable for many nanotherapeutics. Proteins, for example, typically are substantially more expensive than small molecule therapeutics, and precise nanostructures containing proteins will be more costly still. Nanotherapeutics must justify their high COGs. Secondly, as new therapeutic modalities, nanotherapeutics may carry significantly larger risks than those associated with more conventional therapies. Expensive, novel moieties, such as nanobiotechnological therapeutic devices, are therefore most likely to be accepted for treatment of conditions that not only are accessible to intervention at the nanoscale but also for which existing therapeutic modalities have acknowledged shortcomings in patient morbidity or mortality. We have selected two disease states sufficiently grave and sufficiently unserved to warrant nanotherapy: cancer and cardiovascular disease. Modulation of immune responses and vaccination is our third application area.

10.4.1 Nanotherapeutic Devices in Oncology

The economic burden imposed by cancer is immense, measuring in the billions of dollars annually in the United States alone. Existing therapies such as surgical resection, radiotherapy, and chemotherapy have profoundly limited efficacy and frequently provide unfavorable outcomes as the result of catastrophic therapeutic side effects. Additionally, the biology, chemistry, and physics of cancer, in general, and solid tumors in particular, provide therapeutic avenues accessible only by nanoscale therapeutics. Oncology is thus an ideal arena for emerging nanotechnological therapies.

Tumor Architecture and Properties

Tumors as tissues are relatively chaotic structures exhibiting vast structural heterogeneity as a function of both time and space [10.9, 77, 111, 112]. In healthy tissues, vasculature resembles a regular mesh in which the mean distance of tissue spaces to the nearest vessel is tightly controlled and highly uniform. On the other hand, the vasculature of tumors resembles a percolation network containing regions experiencing vastly different levels of perfusion. Tumors of 1 mm^3 or larger typically contain measurably hypoxic domains, with pO2 values as much as two- to threefold lower than in normal tissue. High levels of hypoxia are characteristic of enhanced metastatic potential and tumor progression. Also due to insufficiency of perfusion, tumors frequently contain necrotic domains.

The average torturosity of vascular flow paths in tumors is also much greater than in healthy tissue, and transient thrombotic events lead to enhanced resistance to flow and ongoing vascular remodeling. Aside from its plasticity, tumor vasculature itself is highly irregular, may be incompletely lined with endothelial cells, and often exhibits significantly higher extravasation limits (the highest molecular weight of materials that can leave the vasculature and diffuse into the interstitial spaces) than normal vasculature. Tumor tissue is also poorly drained by lymphatics, so extravasated materials tend to remain in situ in tumors and are not cleared efficiently.

These biological phenomena all differentiate tumor tissue from normal tissue and can be exploited in therapy. For instance, the special vascular integrity and lymphatic drainage properties of tumors constitute the Enhanced Permeability and Retention effect (EPR) [10.11, 12, 80, 113]. EPR presents an obvious opportunity for intervention with nanoscale therapeutics (Fig. 10.3). Extravasation limits for normal tissues are variable, but structures larger than a few nanometers in diameter do not leave circulation efficiently in most tissues, whereas tumor vasculature frequently allows egress of materials in the tens to hundreds of nanometer range. Further, nanomaterials, once extravasated, are not cleared by lymphatic drainage. EPR provides tumor targeting that does not depend on biological affinity reagents: tumor tissue provides preferential depot sites for extravasated drug delivery devices. Targeting to a desired site of action is highly desirable for cytotoxic therapeutics, and EPR provides the basis for a growing class of polymer therapeutics [10.11, 12, 80, 113]. That said, EPR and targeting by biological affinity are not necessarily mutually exclusive. A number of antigens more or less specifically related to tumors are known (tumor associated antigens or TAAs), and as discussed above tumor-/organ-specific vascular addresses might be exploited for delivery of therapeutics [10.65, 66]. For instance, various biological reagents that recognize TAAs (i. e., antibody to carcinoembryonic antigen, transferrin) have been conjugated to nanoscale contrast agents to enhance contrast agent localization to tumoral sites [10.104].

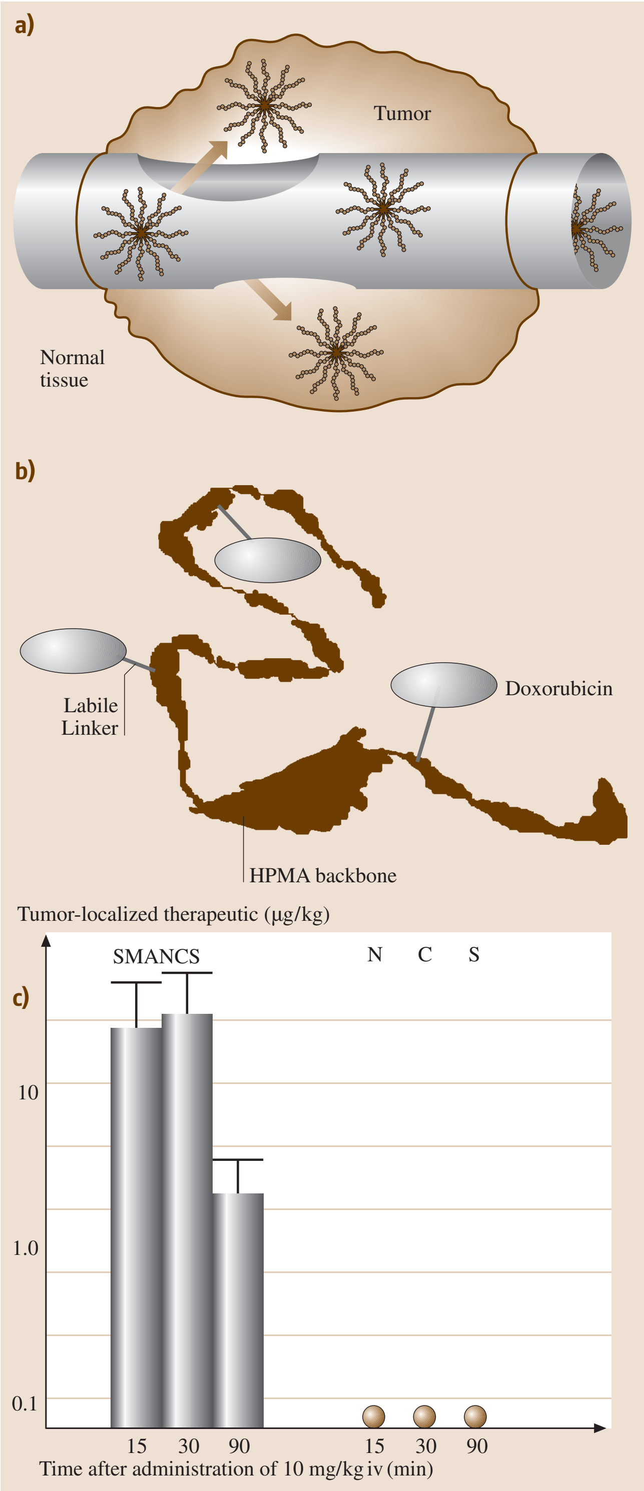

Fig. 10.16a–c Therapeutic nanodevices can be specifically delivered to tumors by virtue of their size [10.11, 12, 80, 113]. **(a)** An illustration of the relatively low integrity of tumor vasculature. Therapeutic nanodevices (in this case, micelles carrying cytotoxins) cannot leave (i. e., extravasate from) the vasculature of normal tissue, but the vasculature of tumors is sufficiently leaky to allow specific delivery to tumoral interstitial spaces. This is the basis of size-dependent targeting to tumors by the enhanced permeability and retention, or EPR effect. **(b)** A polymer nanotherapeutic (PK1, [10.80]) that exploits EPR effect to deliver a cytotoxin (doxorubicin) to tumors. The overall mass of the complex is tuned by control of the mass of the N-2-(hydroxypropyl) methacrylimide backbone, so as to be above extravasation limits for normal tissues but below both tumor extravasation limits and the size limit for excretion through the kidney. A final feature of the device that favors delimiting doxorubicin cytotoxicity to tumoral sites is the fact that the device is a prodrug. The cytotoxicity of doxorubicin is severely curtailed when the doxorubicin is covalently linked to the polymer therapeutic, though it manifests its full toxicity when it is released from the complex by cleavage of the labile (peptidic) linker. The labile linker contains a cleavage site recognized by a protease known to be over-expressed in the target tumor type. Thus, the device produces a depot of inactive doxorubicin in tumors that is activated preferentially to full toxicity by conditions prevalent in the tumor. **(c)** EPR can drive localization of appropriately sized therapeutics to tumoral sites [10.113]. The polymer therapeutic SMANCS accumulates rapidly in murine tumors, whereas the NCS, the parent therapeutic (not engineered for EPR delivery), does not

Tumor (EPR) Properties and Nanotherapeutics

PK1 (Fig. 10.16) is an early example of a growing and increasingly sophisticated class of polymer therapeutics [10.11, 12, 80, 113]. It is a nanoscale molecular therapeutic device wherein the therapeutic moiety (doxorubicin) is covalently linked to the polymeric backbone. As we will see, each component of the polymer therapeutic fulfills a discrete function, so the device rubric is warranted [10.6].

PK1 consists of a polymeric backbone (HPMA, hydroxyl polymethacrylamide) which, as discussed below, targets desired sites by virtue of its specific nanoscale size (by EPR effect) and a cytotoxin (doxorubicin), covalently linked together by a peptide. PK1 exhibits multiple design features that allow it to preferentially deliver doxorubicin to tumors. The size of the complex is tuned to be too large to extravasate in healthy tissue but still small enough for renal clearance. Effectively, tumors and the excretory system compete for PK1. This aspect of the relationship between therapeutic size and systemic tolerability is illustrated in Fig. 10.16. The therapeutic device has to be sufficiently large to have limited access to healthy tissues but must have relatively unimpaired access to tumors and the excretory system via the kidneys.

A second engineered feature of PK1 relates to tumor-specific release of cytotoxin. Doxorubicin is toxic as a free molecule but is nontoxic as a part of a polymer complex. In PK1, doxorubicin is linked covalently to the polymeric backbone by a peptide chosen specifically because it is the substrate for a highly expressed protease in the target tumor. Thus PK1 deposits cytotoxin to tumors in an inactive form that is processed to an active form preferentially by tumoral metabolic activity. In all, the design features of PK1 act to minimize systemic exposure to cytotoxin and augment tumoral exposure, thereby enhancing therapeutic benefit and diminishing undesired side effects.

Therapeutics need not be covalently linked complexes to be delivered by EPR. In some cases, bioactives can be engineered to interact with blood constituents non-covalently so as to bring the apparent size of the complex into a range that allows preferential tumor deposition. This is the mechanism of delivery for the oldest EPR-exploiting therapeutic in clinical use, SMANCS (S-Methacryl-neocarzinostatin) [10.113]. Neocarzinostatin (NCS) is a small (12 kDa, a few nm in diameter), cytotoxic protein that elicits its toxicity by generating intracellular superoxide radicals enzymatically, which in turn damage DNA. Unmodified NCS is cleared rapidly and exhibits dose-limiting bone-marrow toxicity. Both problems are mitigated by the site-specific conjugation of two small (about 2 kDa and less than 1 nm in length) styrene-comaleic acid polymer chains, resulting in SMANCS. The $T_{1/2}$ of SMANCS is about ten-fold higher than that of NCS, though SMANCS itself is too small to accumulate preferentially in tumors by EPR. This is mitigated by the fact that SMANCS (but not NCS) is efficiently bound by serum albumin. The resulting non-covalent complex is sufficiently large not to extravasate into normal tissue, though it still partitions efficiently into many tumors. The EPR-driven partitioning of the non-covalent SMANCS-albumin complexes into tumors suggests that other non-covalent complexes (such as the multi-dendrimer device of Fig. 10.3, were it pre-assembled prior to administration) might also target to tumors via EPR.

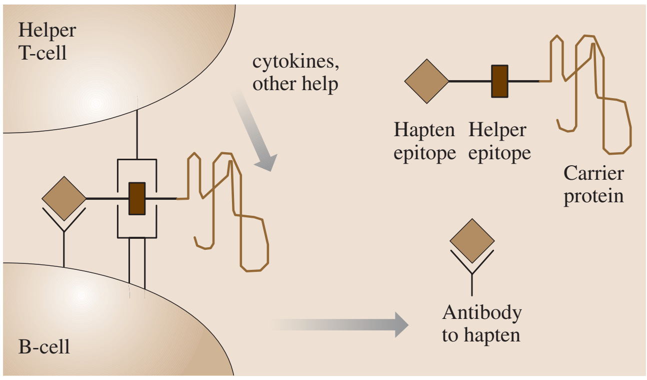

Fig. 10.17 Antibody responses [10.1] are efficiently triggered by antigens that contain both B-cell epitopes (diamond) and helper T-cell epitopes (rectangle). The B-cell epitope binds a receptor on the B-cell surface (the Y shape shown interacting with the B-cell epitope) that is essentially a membrane-bound form of the antibody species that the particular B-cell can produce. The peptidic T-helper epitope engages another receptor complex on the B-cell surface (class II MHC, the two part receptor schematically depicted as interacting with the T-helper epitope, see text). When presented in the context of the class II MHC complex of the B-cell, the T-helper epitope can be recognized by a cognate T-cell receptor on the helper T-cell surface (shown as a receptor on the T-cell surface interacting with the T-helper epitope/class II MHC complex). In response to recognition of its appropriately presented, cognate T-helper antigen, the helper T-cell provides multiple stimulatory signals that cause the B-cell to proliferate, secrete antibody, and differentiate. This is a classical hapten-carrier configuration of an immunogen [10.1], in which a synthetic B-cell epitope is the hapten (in our cases, a PAMAM dendrimer or SWCNT epitope), and the protein covalently linked to the hapten is the carrier, providing the T-helper epitope. Note that the antibody produced recognizes *only* the B-cell epitope and *not* the T-helper epitope. The T-helper epitope also need not be present for antibody recognition of carrier epitopes, nor is its presence necessary to maintain the ongoing antibody response once antibody production is triggered. Note that whether the T-helper epitope remains covalently linked to the B-cell epitope during antigen presentation (here, the intact carrier protein is shown linked to the B-cell epitope) is debatable. Proximity of the B-cell epitope/receptor complex and T-helper epitope/class II MHC complex on the B-cell surface may be sufficient to trigger a robust antibody response

Delivery of antitumor agents using EPR can be highly efficacious. As shown in Fig. 10.17, conjugation of therapeutics to produce controlled-size delivery complexes can enhance tumoral accumulation by multiple orders of magnitude relative to free therapeutics [10.11, 12, 80, 113]. In effect, EPR targeting broadens the therapeutic window (the range of concentration between the lowest effective and highest tolerable dose), allowing more intensive therapy. Increased therapeutic intensity allowed by EPR strategies also produces significantly enhanced clinical outcomes [10.11, 12, 80, 113]. The benefits of EPR strategies seem to be general in that a broad range of materials can be targeted to tumors and almost any particulate material falling within the appropriate size range will partition into tumor tissue via EPR. This includes not only linear and dendritic polymers but also higher-order, self-assembled structures such as micelles or shell cross-linked structures and inorganic materials such as ferromagnetic particles. In addition to EPR tumor targeting, each of these structures has intrinsic properties that might be exploited in therapeutic nanodevices (see the discussion of imaging above). Targeted contrast agents may be particularly important for early tumor detection.

10.4.2 Cardiovascular Applications of Nanotherapeutics

The cardiovascular system is one of the most dynamic and vital systems of the body. In principle, every cell in the body is accessible to the bloodstream via the vasculature. If controlled navigation of, and extravasation from, the vasculature can be accomplished, therefore, the potential of nanotherapeutic devices is virtually unlimited. But as mentioned throughout this chapter, opsonization and immune clearance, as well as targeting and triggering of therapeutics, remain key issues to address in order to facilitate wide application of nanotherapeutic devices to cardiovascular disease. In this section we provide a brief overview of categories and examples of nanotherapeutic devices used in blood-contacting applications.

Cardiovascular Tissue Engineering

A usable, immunologically compatible artery for coronary bypass is not always available, so synthetic or semi-synthetic artery substitutes are highly valuable. Tissue engineering approaches might mitigate the lack of acceptable homologous artery, as long as an acceptable artery substitute can be produced. Current tissue engineering approaches involve synthesis of three-dimensional, porous scaffolds that allow adhesion, growth, and proliferation of seeded cells to generate a functional vessel. Cellular organization and growth in synthetic scaffolds are often chaotic and random, and sufficient blood supply is seldom achieved within the scaffold. These conditions often result in death of the seeded cells and compromise the integrity of the vascular prosthesis.

Microfabrication techniques have allowed miniaturization of the tissue engineering scaffolds. These techniques allow the molecular-level control of cellular adhesion and propagation by control of surface topography, surface chemistry, and arrangement of morphogenic and proliferative signaling ligands and cytokines. Microfabrication schemes are being developed for the elucidation of parameters necessary for cellular attachment and orientation on well-defined silicon- and polymer-based surfaces. One particularly powerful application of controlled cellular growth is for the development of artificial capillaries to remediate vascular disease [10.114] (Moldovan, unpublished). Micro-machined silicon or polymer channels of the dimensions of capillaries have shown promise as effective conduits to direct endothelial cells to form tubes through which a fluid (blood) could eventually flow.

MEMS technology and nanoscale control of molecular events and interactions has also been applied to the development of cardiovascular sensors. For instance, blood cell counters have successfully detected red and white blood cells and show promise for quick, reliable blood cell analysis [10.115]. Additionally, force transducers have been developed to measure the contractile force of a single cardiac myocyte [10.116]. Basic research into cellular function and response to stimuli such as these will eventually support more efficacious and robust cardiovascular tissue engineering approaches.

Nanoparticulate Carriers for Therapy and Imaging

Two major cardiovascular applications of nanoparticles are targeted contrast agents for imaging (MRI, CT, X-ray, etc.) and targeted drug delivery vehicles to address vascular lesions (atherosclerotic plaques) and other vascular disease states. In either application, nanoparticle surfaces must be appropriately modified to avoid RES clearance and to drive interaction with the desired effector cells. Decades of research in the area of molecular and cell biology have provided researchers with many tools to make the appropriate protein component selection for nanotherapeutic contrast agents. For instance, knowing that the procoagulant protein tissue factor is expressed only at the surface of injured endothelial cells provides a means to differentiate healthy from diseased vasculature. It also provides a method to target disease sites. Similar sorts of information regarding disease and normal vasculature at specific tissue sites can be critically important for targeting specialized vascular beds, such as that which occurs at the BBB. Several groups have begun efforts to develop nanoparticles that selectively pass across the BBB to deliver chemotherapeutic or other drugs to the brain [10.117, 118], and this may improve outcomes from stroke.

With regard to MRI approaches, particles used for targeted contrast enhancement in specific tissues or vascular beds may be quite powerful for detection of unstable atherosclerotic plaques and quantitative analysis of blood flow. For instance, nonspecific accumulation of nanoparticles within human atherosclerotic plaques has been reported [10.119, 120] and may ultimately provide a way to distinguish vulnerable and stable plaques. Such data might be used to direct therapy, with the intent, as always, to improve therapeutic outcome. Furthermore, nanotherapeutic devices have been designed to include both contrast and drug delivery functions (Fig. 10.3). Single-platform, multifunctional particles for cardiovascular therapy represent a cutting edge in the cardiovascular field. These delivery vehicles are being designed and improved continual for longer circulation time, more efficient targeting [10.66, 121], degradability, and biocompatibility [10.122].

10.4.3 Nanotherapeutics and Specific Host Immune Responses

Immune responses directed to immunogens (molecules or structures capable of eliciting specific immune responses, also called antigens) are a pervasive aspect of the responses of vertebrates to exposure to foreign macromolecules, proteins, particles, and organisms (for overviews, see [10.1, 123, 124]). Undesired antibody responses, in particular, are the single most important hurdle for clinical use of recombinant proteins. On the other hand, undesired cellular immune responses are key mediators of frequently devastating disorders for patients, whether the responses are directed to self (as in autoimmune disease) or to foreign antigens (as in transplant rejection). As we shall see, host immune responses to nanotherapeutic devices are a critical determinant of their efficacy. Any strategy (including the biotechnological strategies) that allows controlled manipulation of immune responses either to augment desirable responses, as in vaccines, or to mitigate deleterious responses holds substantial potential benefit.

Basic Immunology

Specific immune responses are directed to individual macromolecules or assemblages of them and can be categorized into cellular responses, mediated by cytotoxic T-lymphocytes (CTLs), and humeral responses, mediated by soluble proteins (found in blood serum and other biological fluids) secreted by B-cells (antibodies). Regardless of the effector (CTLs or antibodies), specific responses are directed to individual molecular features of antigens called epitopes. The three types of epitopes are: cytotoxic T-cell (CTL) epitopes, helper T-cell epitopes, and B-cell epitopes. The first two types are peptides that must be presented on cell surfaces in the context of the class I (in the case of CTL epitopes for CTL responses) or class II (in the case of T-helper epitomes, involved in both antibody and CTL responses) major histocompatibility complex (MHC) molecules. B-cell epitopes are a chemically and structurally heterogeneous group of epitopes and include synthetic molecules as well as peptides. In addition to CTL and B-cell epitopes, which are recognized by immune effectors, T-helper epitopes are recognized by regulatory components of the immune system (helper T-cells). Vibrant CTL and antibody responses require immunogens containing T-helper epitopes along with either B-cell or CTL epitopes.

While epitope-presenting MHC class I and class II complexes must be present on cell surfaces to trigger immune responses, the machinery for antigen processing and loading of MHC class I or class II proteins for antigen presentation reside in distinct membrane-bound compartments in the cell [10.1]. For instance, processing for class I presentation (to trigger CTL responses) occurs through the cytosol, but charging of MHC class I complexes is accomplished in the endoplasmic reticulum. On the other hand, processing for class II presentation (to augment antibody and CTL responses), as well as the loading of class II, occur in endosomal compartments. Although these are membrane bounded compartments in the cytosol, topographically the interiors of endosomal compartments are part of the extracellular environment and are not inside the cell at all.

In the case of both class I and class II, the antigen presenting complexes are moved to the cell surface after charging. The moving of molecules between different cellular locations (called trafficking) is tightly regulated by and driven by specific cellular proteins, systems, and structures. Proper trafficking is essential to assure that all cellular components assume their correct position in the cellular organization, which in this case, is to assure antigens are directed to the compartments that will (most often) produce most protective immune responses. These complex antigen processing and presentation systems can be perturbed and exploited to offer novel opportunities in vaccine design.

Nanotherapeutic Vaccines: Eliciting Desired Host Immune Responses

Both therapeutic and prophylactic vaccines are naturals for nanobiological approaches. Therapeutics that modulate immune responses (antibody, CTL responses, or both), vaccines typically produce desired immunity to disease or pathogens. Ideally they are chemically well defined, safe, and induce long term (preferably, multiyear) protective immune responses [10.125]. Recent developments in immunobiology have opened opportunities for nanobiological vaccine design [10.126, 127].

Particulate Vaccines

The immune system, through dendritic cells (DCs), other antigen presenting cells (APCs), and immune effectors, performs ongoing surveillance for non-self antigens using the MHC antigen presentation system described above [10.1]. It has long been known that APCs sample serum for antigens that they process and present using the MHC class II pathway for antibody production [10.1]. Research of the last few years has shown that DCs, macrophages, and other APCs also take up particulate materials and process and present the constituent peptide epitopes via the MHC class I pathway to trigger specific CTL responses. Previous dogma held that only endogenous antigens (such as those produced in virally infected cells) could be presented through the class I MHC pathway. Both soluble and particulate antigens can be presented via the class I pathway, but particulate antigens are orders of magnitude more potent than the same immunogens formulated in soluble form [10.125–128].

APCs are not particularly fastidious regarding chemical composition of the particles they take up: latex beads [10.126, 128] as well as iron beads [10.126] have been used. Self-assembled nanoparticles made of lipidated epitope peptides also deliver orders of magnitude of better immunization than do soluble formulations of the same peptides [10.129]. The specific size for optimal CTL responses is not fully defined but clearly is in the nanometer to micrometer size range [10.125, 126, 128, 129]. One could imagine a systematic exploration of optimum size and compositional preferences for uptake by APCs using such chemically and morphologically tunable polymeric nanomaterials as dendrimers and tectodendrimers.

Nanobiological Design of Nanotherapeutic Vaccines

As mentioned earlier, antigen trafficking determines which specific MHC complex (class I or class II) epitopes are presented on and, therefore, which immune effectors are produced in the subsequent response. The process, while tightly regulated, is amenable to intervention using nanotherapeutic vaccines. Nanobiological design strategies (discussed above) can take advantage of the numerous proteins and peptides now known to mediate the trafficking of carried materials to specific sites. In the device context, these trafficking moieties might assure delivery of antigens to the appropriate cellular compartments to trigger desired class I- or class II-mediated responses.

Vaccine polypeptides, for instance, might be fused or formulated with pathogen polypeptides that will deliver them across epithelium from outside the host for use in mucosal vaccination [10.130]. Still other peptides are known to carry injected materials across the cell envelope into the cytosol [10.131]. Alternatively, vaccine entities taken up by receptor-mediated endocytosis (into endosomal vesicles, the traditional site of uptake for antigens destined for MHC class II presentation) could be shuttled into the cytosol (and hence to the class I MHC pathway) by the incorporation of peptide domains of bacterial toxins that mediate endosomal escape to cytoplasm [10.132]. Potentially, nanobiological therapeutic design strategies [10.5, 13, 15] can combine knowledge of polymer chemistry, nanomaterials, chemistry, cell, molecular, and immunobiology to generate highly effective nanotherapeutic vaccines.

Nanotechnology and Modulation of Immune Responses

On the other hand, protein immunogenicity represents a critical limitation to the therapeutic use of recombinant proteins, particularly when specific antibody responses to the therapeutics are engendered [10.32]. The consequences of antibody responses to proteins are variable, ranging from no effect to decidedly negative consequences to the host. Antibodies can neutralize therapeutic bioactivity, rendering drug entities inactive in repeated rounds of therapy, distort drug P_k and B_d properties, and, in the worst case, trigger autoimmune responses with cross-reactive host antigens (host antigens that share some identity with the immunizing epitopes).

Antibody Responses to Nanotherapeutic Devices

Antibody responses to protein components are a major challenge for nanobiological therapeutic devices in vivo. We, and others [10.6, 14, 21, 133–136], have demonstrated that protein components of bioconjugates to diverse nanomaterials can not only retain their intrinsic immunogenicity but, worse, can render the otherwise non-immunogenic synthetic nanomaterials to which they are linked capable of inducing antibody responses [10.6, 7, 14, 21, 126, 133–135]. The mechanism of immunogenicity seems to be haptenization, a well-known phenomenon in immunology. Haptenization occurs when antigens containing only B-cell epitopes are covalently linked to proteins containing T-helper epitopes, thereby creating hybrid antigens that are fully immunogenic with respect to antibody responses. The process is illustrated in Fig. 10.17.

As a result of the chemically repetitive structure of many synthetic nanomaterials, the antibodies directed to them have interesting properties. Most notably, antibodies generated to nanostructures are often cross-reactive to other chemically similar nanostructures. For example, antibodies raised using haptenized fullerenes also recognize carbon nanotubes [10.133], and antibodies raised to haptenized generation 0 (G_0) PAMAM dendrimers also recognize higher generation PAMAM dendrimers (G_1, G_2, G_3, etc. [10.14, 21, 133–136]. Presumably, cross-reactivity results from the presence of common B-cell epitopes in the immunogen used and in structurally similar cross-reactive antigens. Thus, patients immunized with any therapeutic containing one haptenized nanostructure could potentially raise antibodies that would recognize any nanotherapeutic containing a material structurally similar to the haptenized nanomaterial.

This is a potentially serious problem, in light of the documented ability of antibody responses to interfere with therapy by neutralizing drug or distorting P_k and B_d properties. As discussed before, molecular weights of serum antibodies are around 150,000 AMU with an effective diameter in excess of 10 nm [10.1, 123, 124]. Many of the desirable properties of nanotherapeutics derive from their precisely controlled sizes, and size-dependent properties like EPR targeting could be jeopardized if the therapeutic to be delivered were noncovalently complexed with antibodies directed to its synthetic components. Due to long-term persistence of memory immune responses, exposure to haptenized nanotherapeutics could endanger patient responsiveness to structurally similar materials for the duration of their lives. Obviously, immunogenicity is a critical challenge to the nanobiological therapeutic paradigm that has yet to be satisfactorily addressed.

Polymer Conjugation to Mitigate Immune Responses

There are multiple approaches to mitigating immunogenicity, some of which are applicable to nanobiotechnological therapeutic devices. Polymer conjugation, particularly conjugation of proteins to poly (ethylene glycol) (PEG, a flexible, biologically well-tolerated hydrophilic polymer) has been used to control immunogenicity of a broad class of proteins and synthetic molecules since the mid-1970s [10.19, 137]. Typically, pegylation involves covalent linkage (often by promiscuous conjugation chemistry) of variable numbers (usually two to six polymer molecules) of size-polydisperse (usually 3,000 to 10,000 AMU) linear or branched PEG chains to the potential immunogen. This approach results in polydisperse bioconjugates, which flies in the face of the molecular level-structural precision often considered essential for nanotechnology. Still, pegylation may be useful in some limited nanobiological applications.

The mechanism of molecule-specific immunosuppression remains obscure, despite almost 30 years of research. The flexibility of the PEG strands allows them to wrap around the molecules they are conjugated to, perhaps delimiting access of immune effectors and immune processing machinery to the potential immunogens. Pegylated molecules exhibit dramatically enhanced apparent diameters and vastly enhanced solubility in aqueous buffer. Both of these phenomena result in substantially diminished clearance of pegylated molecules by the RES and may account for the retention of in vivo bioactivity despite the apparent steric hindrance that may result from wrapping PEG moieties around their protein bioconjugate partners. Pegylation-related, bioconjugate-specific immunosuppressive phenomena cannot be fully accounted for by sterics, though. Immunological experiments in syngeneic mice show that the protein-specific immunosuppression associated with pegylated proteins is transmissible from one animal to another by surgically harvested splenocytes. This finding is more consistent with a mechanism of suppression involving immunologically induced tolerance of foreign antigens as opposed to one caused by simple steric considerations [10.137].

Synthetic Nanoporous Membranes for Immunoisolation

A second approach to avoiding immune responses to nanodevices or biocomponents is immunoisolation by encapsulation [10.16, 17]. In this strategy, potentially immunogenic structures (nanodevices, cells, macromolecules) are sequestered from high molecular weight, large host immune effectors, and immunosurveillance, while small molecules, such as glucose and small proteins, diffuse freely (Fig. 10.18). While conceptually simple, immunoisolation is technically challenging in that the number of pores must be high to maximize

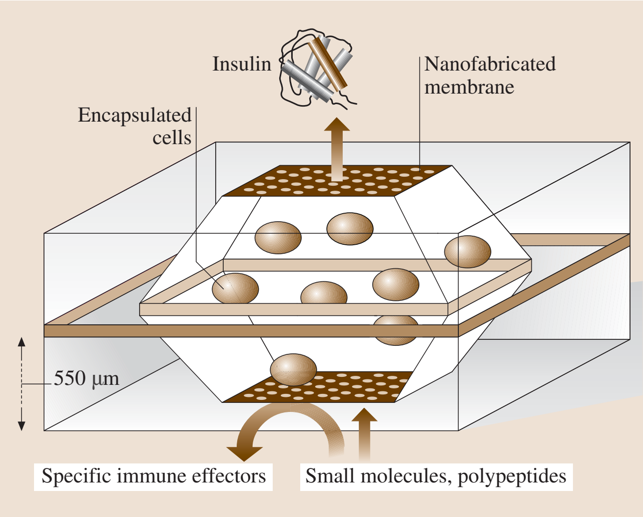

Fig. 10.18 Biological components can be protected from unfavorable environments by encapsulation [10.16–18]. Shown is an immunoisolation capsule containing heterologous (immunogenic) cells producing insulin (a four helical bundle protein, see text). The cells are protected from immune responses by membranes with 10 nm pores that allow free diffusion of small molecules (oxygen, nutrients) into the capsule and small proteins (insulin, shown here not to scale) out of it. However, immune effectors with diameters greater than pore size (immune cells like CTLs and APCs, antibodies, other effectors) don't access the capsule interior, protecting the xenograft from immune rejection. The structure integrates both micro- and nanoscale components (including the nanoporous membranes of Fig. 10.4) into a macroscale construct

flux, and the pores must be highly uniform in dimension.

Because of the exquisite sensitivity of the immune system, immunoisolation must be nearly perfect if it is to work at all. Particularly in the case of immunoisolation of cells for xenografts, if only a tiny fraction (as few as 1%) of the pores of an immunoisolation capsule are large enough to admit immune effectors (for instance, antibodies and complement components, exclusion of which requires pores of 50 nm or less), or the viability of the graft will be compromised [10.16, 17, 138]. Thus, the strategy requires fabrication of membranes with pores that have very tightly controlled dispersities in terms of their maximum diameters. Until recently, immunoisolation for xenografts has failed for lack of such membranes.

The technological innovations that made the synthesis of these high-quality nanoporous membranes were described above and in Fig. 10.4. The nanoporous membranes can be used to form immunoisolation capsules when either two such membranes are joined so that they bind an enclosed space or that a nanoporous membrane is used to enclose a compartment of some sort in a therapeutic structure. The enclosed spaces can be charges with functional nanodevices or, for that matter, living cells. Such immunoisolation capsules charged with heterologous pancreatic islet cells have demonstrated immense promise for regulating normoglycemia in diabetic animal models [10.16–18].

Biogenic, Nanoporous, Immunoisolation Membranes

The novel microfabrication strategy has made it possible to make these highly defined membranes in an industrial process, and they are applicable to numerous biomedical applications beyond immunoisolation [10.16, 17]. Still the fabrication process is rigorous, and an alternative biogenic method to derive selectively porous silicon capsules has been proposed [10.139, 140]. Diatoms are encapsulated in biogenic amorphous silicon shells (called frustules), which the living cells secrete. Frustules must necessarily be porous to allow the cells to take on nutrients and excrete wastes, and the porosity, pore dimensions, and other micro- and nanoscale morphological frustule features are tightly controlled at genetic and physiological levels. Using techniques well known to microbiologists (manipulation of culture conditions), geneticists (mutagenesis), and cell biologists (optically driven cell sorting in flow systems), it is theoretically possible to select and propagate diatoms whose frustule morphology, size, and pore characteristics fall within pre-selected limits [10.139, 140]. Frustules can be isolated intact from diatom cultures, so biogenic production of silicon membranes with porosity characteristics similar to the microfabricated membranes discussed above may be feasible. Though they have yet to be realized, immunoisolation capsules made from genetically/physiologically tuned frustules have been proposed [10.140]: one wonders what other biomineralized structures eventually might also be put to nanobiotechnological use.

10.5 Concluding Remarks: Barriers to Practice and Prospects

The effort to produce nanoscale therapeutic devices is clearly highly interdisciplinary. As we have seen, it touches on numerous established disciplines, encompassing elements of physiology, biotechnology, bioconjugate chemistry, electrical engineering, and materials science, to name just a few of the fields involved. Obviously, this broad sweep of knowledge is difficult for any one investigator to master fully. The breadth of the effort constitutes just one of the major barriers to entry in the field. Other challenges include the raw complexity of biology, the fashion in which biologists hold and distribute information, and cultural differences between engineers and biological scientists.

10.5.1 Complexity in Biology

Biology is characterized by particularity: nuances of biological systems are often unique to the system at hand and highly idiosyncratic. This follows from the fact that biological systems are not purpose-built, as are designed devices (although they can give that appearance, as seen in Fig. 10.1), but rather arose as the consequence of evolutionary processes. Evolution is a highly chaotic business, with the outcome of any given evolutionary process highly sensitive to initial conditions (populations of organisms subjected to selection, other organisms in the environment, biology of all organisms involved, resource availability, other environmental factors, etc.). Moreover, many of the variables affecting natural selection processes are nonstatic and change, as a function of time or space or both, *while* selection is exerted against a population of organisms. Conditions leading to one evolutionary adaptation or another are therefore seldom duplicated exactly, so that individual adaptive features are idiosyncratic, with elements relating not only to their biological functions but also to their evolutionary history. Individual systems and adaptive features thereof can be almost baroquely complex because the structures themselves arose under unique conditions, under unique selective pressures, and from unique initial biological systems. The extent of the complexity of biological systems is apparent even in individual macromolecular constituents of biological systems and can be made clear by comparison of synthetic nanostructures to biological nanostructures with similar dimensional aspects.

SWCNT (single wall carbon nanotubes) have a minimal diameter of about 1.3 nm, and many proteins likewise have diameters of few nms. SWCNTs are regular and homogeneous polymeric structures composed entirely of carbon atoms (Fig. 10.6). Proteins, on the other hand, are not polymers of one repeating subunit, but rather are nonrandom copolymers of 20 chemically distinct amino acids (aas). The order of the aas drives specific folding events that produce three-dimensional structures that, in turn, present specific aas and their side chains at specific positions in space (Fig. 10.10). This control of aa position in space drives protein activity: small changes in aa sequence can perturb structure and function profoundly. Therefore, though they fall within the same broad size regime, proteins are structurally and functionally much more complex than SWCNTs. The extent of biological complexity can be glimpsed when one considers that individual living things are ordered aggregates of multiple macromolecules and supramolecular structures (potentially, billions and billions of them), belonging to distinct chemical classes (lipids, proteins, nucleic acids, etc.), with each individual macromolecule being at least as complex in structure and function as the protein of Fig. 10.10. The nearly irreducible complexity of biological systems is a central fact of practice in the biological sciences and drives the way biologists gather and disseminate information. It is the key informant of the culture of biologists.

10.5.2 Dissemination of Biological Information

Biologists typically consider themselves primarily as scientists, as opposed to being engineers or technology developers. The product biologists generate, then, is information rather than devices or structures. Their interest in technological application of the knowledge they produce is typically secondary to their desire to develop the information itself. Additionally, as we have just seen, biology is a ferociously complex discipline.

As a practical matter, biological data itself is often much more rich (and often more ambiguous) than data from harder scientific endeavors. The data can be so very rich that biologists must make choices as to what data is relevant to a given phenomena and, therefore, what data they will publish. With absolutely no intent to conceal or mislead, biologists are often driven to publish only a small fraction of the data they gather.

Since they see themselves primarily as scientists, biologists tend to publish data they believe to be of broad scientific significance. Typically, the chosen information does not include data that might be critical to technology development, frequently making the biological literature an inadequate resource for engineers. The omitted information thus becomes lore. That is to say, the infomation is critical for the practice or development of a given technology but usually is not accessible to persons outside the field (as many nanotherapeutic device producers may be). This can be illustrated by recent experiences around phage display-mediated affinity maturation of four helical bundle proteins [10.141–143].

Four helical bundle proteins are loosely related small proteins consisting of four α helices arranged in a specific configuration (see the dynamite stick representation of insulin that occurs in Fig. 10.18). They include insulin, growth hormone, most cytokines, and various other molecules involved in cell to cell signaling. They have a wide range of therapeutically valuable properties that might be made even more desirable if their potencies could be enhanced. There is thus significant interest in variant four-helical bundle proteins with improved bioactivities.

Phage display [10.67] is a method that can be used to sort protein variant libraries for variants with enhanced affinity for a target receptor. This is usually accomplished by iterative rounds of affinity selection on the receptor followed by propagation of selected phage. As rounds of selection and propagation proceed, the mean affinity of the selected variants for the receptor increases (so-called affinity maturation). A phage display method is depicted in Fig. 10.11.

Affinity maturation by phage display can be used to identify variant proteins from libraries that have enhanced biological activities, providing that affinity for receptor is limiting in the overall activity of the parent protein [10.134]. Phage display affinity maturation has been successfully used to select some enhanced activity variants of four helical bundle proteins [10.144, 145] but is *not* applicable to engineering increased activity for *all* four helical bundle proteins [10.134]. The reason for the inapplicability of display methods to engineer high-activity human growth hormone (hGH) variants became apparent in a recent publication from Genentech [10.142].

In the case of hGH, the affinity of the parent molecule for receptor is in excess of that required to drive maximal biological responses [10.142]. This fact was inferred after a multiyear effort (beginning in the late 1980s and culminating in 1999) to derive more active hGH proteins, involving literally thousands of hGH variants. Many of the variants did indeed exhibit enhanced affinity for the hGH binding protein (receptor), but none exhibited significantly enhanced biological activity. Each data point (i. e., the biological activity of each individual variant) essentially constituted a failure of a technology to provide a desired result and was therefore seen, not unreasonably, as negative (and therefore uninteresting) data by the investigators. Consequently, the information was not broadly disseminated until a significant conclusion could be presented (i. e., that biological responses to hGH are not limited by ligand-receptor affinity). Thus, between the late 1980s and the 1999 publication, this information was known primarily and unambiguously only to the investigators involved, even though it would have been critical to any technologist planning to affinity mature hGH by phage display for enhanced activity [10.134].

10.5.3 Cultural Differences Between Technologists and Biologists

Biology is the realm of particularity, whereas more physical sciences are realms within which general rules can be derived and broadly applied. The vast complexity of biology usually drives biologists to be highly specialized, perhaps making a career from the study of particular macromolecules (specific enzymes, growth factors, genetic elements, etc.) or biological systems. Biologists are, therefore, neither trained nor encouraged by funding agencies to see themselves as fungible, and frequently they are very reluctant to step away from their primary research focuses into new areas. Doing so is a major undertaking, and a substantial risk, for most biologists.

Engineers, on the other hand, often see themselves as operating from first principles and are more (if not entirely) willing to step into new areas. After all, crossing disciplines is feasible for engineers working in the physical sciences, whereas it is much more difficult, not only for reasons of training but also because of the biological realities discussed above, for biologists. Also by reason of their training and experience, engineers often vastly underestimate the complexity of biological systems and sometimes exhibit what biologists perceive as naïveté in their approach to biological systems. Unfortunately, or fortunately, depending on your perspective, therapeutic nanodevices require both biological and engineering expertise. An often bruising debate as to how best to prepare students for practice in this extremely challenging field is currently playing out in biomedical engineering and other related academic departments across the country.

The traditional approach for training biomedical engineers has been to train conventional engineers with a smattering of biological knowledge. This "primarily engineers" model of biomedical engineering has been reasonably successful in multiple endeavors (orthopedics, imaging technologies, etc.), but it seems unlikely in the extreme that it will suffice for robust practice in nanotherapeutic devices. As this chapter shows, the role of biology is pivotal to success in the field, and sophisticated nanotherapeutics are virtually impossible to design without deep knowledge of the biology involved. A reasonable argument can be made that successful engineering of therapeutic nanostructures will demand intense focus on biology and a smattering of engineering knowledge (a "primarily biologists" approach, as opposed to the historically prevalent "primarily engineers" training model). This view is often strongly and passionately resisted in colleges of engineering, which are the usual homes of biomedical engineering and related efforts. The outcome of the debate is not finalized, but it is critically important: it will determine whether or not a ready cadre of researchers is prepared for a career in the area. If not, the field will progress much as it has, very slowly bringing the benefits of the research to patients. It seems apparent to us, and it also seems that experience to date shows, that a substantial biological component of training for the discipline is simply unavoidable if we are committed to realizing the vast potential of nanotechnology in human therapy.

References

10.1 C. A. Janeway, P. Travers, M. Walport, J. D. Capra: *Immunobiology* (Elsevier, London 1999)

10.2 S. C. Lee: Biotechnology for nanotechnology, Trends Biotechnol. **16** (1998) 239–240

10.3 S. C. Lee: Engineering the protein components of nanobiological devices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 67–74

10.4 S. C. Lee: How a molecular biologist can wind up organizing nanotechnology meetings. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998)

10.5 S. C. Lee: The nanobiological strategy for construction of nanodevices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 3–14

10.6 S. C. Lee: A biological nanodevice for drug delivery. In: *National Science and Technology Council. IWGN Workshop Report: Nanotechnology Research Directions. International Technology Research Institute, World Technology Division* (Kluwer, Baltimore 1999) pp. 91–92

10.7 S. C. Lee, R. Parthasarathy, K. Botwin: Protein-polymer conjugates: Synthesis of simple nanobiotechnological devices, Polymer Preprints **40** (1999) 449–450

10.8 L. Jelinski: Biologically related aspects of nanoparticles, nanostructured materials and nanodevices. In: *Nanostructure Science and Technology*, ed. by R. W. Siegel, E. Hu, M. C. Roco (Kluwer, Dordrecht 1999) pp. 113–130

10.9 J. Baish, Y. Gazit, D. Berk, M. Nozue, L. T. Baxter: Role of tumor vascular architecture in nutrient and drug delivery: An invasion percolation-based network model, Microvasc. Res. **51** (1996) 327–346

10.10 J. R. Baker Jr.: Therapeutic nanodevices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 173–183

10.11 R. Duncan: Drug targeting: Where are we now and where are we heading?, J. Drug Targeting **5** (1997) 1–4

10.12 R. Duncan, S. Gac-Breton, R. Keane, Y. N. Sat, R. Satchi, F. Searle: Polymer-drug conjugates, PDEPT and PELT: Basic principles for design and transfer from the laboratory to clinic, J. Cont. Release **74** (2001) 135–146

10.13 D. S. Goldin, C. A. Dahl, K. L. Olsen, L. H. Ostrach, R. D. Klausner: Biomedicine. The NASA-NCI collaboration on biomolecular sensors, Science **292** (2001) 443–444

10.14 S. C. Lee: Dendrimers in nanobiological devices. In: *Dendrimers and other Dendritic Polymers*, ed. by D. Tomalia, J. Frechet (Wiley, London 2001) pp. 548–557

10.15 J. R. Baker Jr., A. Quintana, L. Piehler, M. Banazak-Holl, D. Tomalia, E. Racka: The synthesis and testing of anti-cancer therapeutic nanodevices, Biomed. Microdevices **3** (2001) 61–69

10.16 T. Desai, D. Hansford, L. Kulinsky, A. Nashat, G. Rasi, J. Tu, Y. Wang, M. Zhang, M. Ferrari: Nanopore technology for biomedical applications, Biomed. Microdevices **21** (1999) 11–40

10.17 T. A. Desai, W. H. Chu, J. K. Tu, G. M. Beattie, A. Hayek, M. Ferrari: Microfabricated immunoisolating biocapsules, Biotechnol. Bioeng. **57** (1998) 118–120

10.18 M. Ferrari, W. H. Chu, T. A, Desai, J. Tu: *Microfabricated silicon biocapsule for immunoisolation of pancreatic islets*, Advanced Manufacturing Systems, CISM Courses and Lectures, Vol. 372, ed. by E. Kuljanic (Springer, Wien 1996) pp. 559–568

10.19 J. M. Harris, N. E. Martin, M. Modi: Pegylation: A novel process for modifying pharmacokinetics, Clin. Pharmacokin. **40** (2001) 539–551

10.20 S. B. H. Kent: Building proteins through chemistry: Total chemical synthesis of protein molecules by chemical ligation of unprotected protein segments. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 75–92

10.21 S. C. Lee, R. Parthasarathy, T. Duffin, K. Botwin, T. Beck, G. Lange, J. Zobel, D. Jansson, D. Kunneman, E. Rowold, C. F. Voliva: Antibodies to PAMAM dendrimers: Reagents for immune detection assembly and patterning of dendrimers. In: *Dendrimers and other Dendritic Polymers*, ed. by D. Tomalia, J. Frechet (Wiley, London 2001) pp. 559–566

10.22 S. C. Lee, M. S. Leusch, V. A. Luckow, P. Olins: Method of producing recombinant viruses in bacteria, (1993) US Patent 5,348,886

10.23 M. S. Leusch, S. C. Lee, P. O. Olins: A novel host-vector system for direct selection of recombinant baculoviruses (bacmids) in *Escherichia coli*, Gene. **160** (1995) 191–194

10.24 V. A. Luckow, S. C. Lee, G. F. Barry, P. O. Olins: Efficient generation of infectious recombinant baculoviruses by site-specific, transposon-mediated insertion of foreign DNA into a baculovirus genome propagated in *E. coli*, J. Virol. **67** (1993) 4566–4579

10.25 T. Gardner, C. R. Cantor, J. J. Collins: Construction of a genetic toggle switch in *Escherichia coli*, Nature **403** (2000) 339–342

10.26 J. Hasty, F. Isaacs, M. Dolnik, D. McMillen, J. J. Collins: Designer gene networks: Towards fundamental cellular control, Chaos **11** (2001) 107–220

10.27 E. Di Fabrisio, A. Nucara, M. Gentili, R. Cingolani: Design of a beamline for soft and deep lithography on third generation synchrotron radiation source, Rev. Sci. Instrum. **70** (1999) 1605–1613

10.28 G. M. Dykes: Dendrimers: A review of their appeal and applications, J. Chem. Tech. Biotech. **76** (2001) 903–918

10.29 R. Spindler: PAMAM starburst dendrimers: Designed nanoscopic reagents for biological applications. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 15–32

10.30 D. Tomalia, H. M. Brothers II: Regiospecific conjugation to dendritic polymers to produce nanodevices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 107–120

10.31 B. I. Yacobson, R. E. Smalley: Fullerene nanotubes: C1000000 and beyond, Amer. Scient. **85** (1997) 324–337

10.32 D. J. A. Crommelin, R. D. Sindelar: *Pharmaceutical Biotechnology* (Harwood Academic Publishers, Amsterdam 1997)

10.33 L. J. Lee: BioMEMS and micro-/nano-processing of polymers – An overview, Chin. J. Chem. Eng. **34** (2003) 25–46

10.34 T. R. Groves, D. Pickard, B. Rafferty, N. Crosland, D. Adam, G. Schubert: Maskless electron beam lithography: Propects, progress and challenges, Microelectron. Eng. **61** (2002) 285–293

10.35 M. Guthold, R. Superfine, R. Taylor: The rules are changing: Force measurements on single molecules and how they relate to bulk reaction kinetics and energies, Biomed. Microdevices **3** (2001) 9–18

10.36 L. M. Demers, D. S. Ginger, S.-J. Park, Z. Li, S.-W. Chung, C. A. Mirkin: Direct patterning of modified oligonucleotides on metals and insulatos by dip-pen nanolithography, Science **296** (2002) 1836–1838

10.37 K.-B. Lee, S.-J. Park, C. A. Mirkin, J. C. Smith, M. Mrksich: Protein nanoarrays generated by dip-pen nanolithography, Science **295** (2002) 1702–1705

10.38 M. Ferrari, J. Liu: The engineered course of treatment, Mech. Eng. **123** (2001) 44–47

10.39 H. W. Rohrs, R. S. Ruoff: The use of carbon nanotubes in hybrid nanometer-scale devices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 33–38

10.40 K. E. Drexler: *Engines of Creation: The Coming Era of Nanotechnology* (Anchor Books, New York 1986)

10.41 J. Cumings, A. Zetti: Low-friction nanoscale linear bearing realized from multiwall carbon nanotubes, Science **289** (2000) 602–604

10.42 D. J. Hornbaker, S.-J. Kahng, S. Mirsa, B. W. Smith, A. T. Johnson, E. J. Mele, D. E. Luzzi, A. Yazdoni: Mapping the one-dimensional electronic states of nanotube peapod structures, Science **295** (2002) 828–831

10.43 C. Dekker: Carbon nanotubes as molecular quantum wires, Phys. Today **28** (1999) 22–28

10.44 M.-C. Jones, J.-C. Leroux: Polymeric micelles-a new generation of colloidal drug carriers, Eur. J. Pharma. Biopharma. **48** (1999) 101–111

10.45 I. Uchegbu: Parenteral drug delivery: 1, Pharma. J. **263** (1999) 309–318

10.46 I. Uchegbu: Parenteral drug delivery: 2, Pharma. J. **263** (1999) 355–359

10.47 K. B. Thurmond II, H. Huang, K. L. Wooley: Stabilized micellar structures in nanodevices. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 39–43

10.48 S. Uppuluri, D. R. Swanson, L. T. Peihler, J. Li, G. Hagnauer, D. A. Tomalia: Core shell tecto(dendrimers). I. Synthesis and characterization of saturated shell models, Adv. Mater. **12** (2000) 796–800

10.49 A. K. Patri, I. J. Majoros, J. R. Baker Jr.: Dendritic polymer macromolecular carriers for drug delivery, Curr. Opin. Chem. Biol. **6** (2002) 466–471

10.50 S. Fernandez-Lopez, H.-S. Kim, E. C. Choi, M. Delgado, J. R. Granja, A. Khasanov, K. Kraehenbuehl, G. Long, D. A. Weinberger, K. M. Wilcoxen, M. Ghardiri: Antibacterial agents based on the cyclic D,L-alpha-peptide architecture, Nature **412** (2001) 452–455

10.51 A. Saghatelian, Y. Yokobayashi, K. Soltani, M. R. Ghadiri: A chiroselective peptide replicator, Nature **409** (2001) 777–778

10.52 J. Davies: Aminoglycoside-aminocyclitol antibiotics and their modifying enzymes. In: *Antibiotics in Laboratory Medicine*, ed. by V. Lorian (Williams and Wilkins, Baltimore 1984) pp. 474–489

10.53 T. Hamouda, A. Myc, B. Donovan, A. Y. Shih, J. D. Reuter, J. R. Baker Jr.: A novel surfactant nanoemulsion with a unique non-irritant topical antimicrobial activity against bacteria, enveloped viruses and fungi, Microbiol. Res. **156** (2001) 1–7

10.54 M. J. Heller: Utilization of synthetic DNA for molecular electronic and photonic-based device applications. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 59–66

10.55 Z. Ma, S. Taylor: Nucleic acid triggered catalytic drug release, Proc. Nat. Acad. Sci. USA **97** (2000) 11159–11163

10.56 R. C. Merkle: Biotechnology as a route to nanotechnology, Trends Biotechnol. **17** (1999) 271–274

10.57 N. C. Seeman, J. Chen, Z. Zhang, B. Lu, H. Qiu, T.-J. Fu, Y. Wang, X. Li, J. Qi, F. Liu, L. A. Wenzler, S. Du, J. E. Mueller, H. Wang, C. Mao, W. Sun, Z. Shen, M. H. Wong, R. Sha: A bottom-up approach to nanotechnology using DNA. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 45–58

10.58 G. Lemieux, C. Bertozzi: Chemoselective ligation reactions with proteins, oligosaccharides and cells, Trends Biotechnol. **16** (1998) 506–512

10.59 R. Offord, K. Rose: Multicomponent synthetic constructs. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology, Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 93–105

10.60 G. T. Hermanson: *Bioconjugate Chemistry* (Academic, San Diego 1996)

10.61 S. S. Davis: Biomedical applications of nanotechnology-implications for drug targeting and gene therapy, Trends Biotechnol. **15** (1997) 217–224

10.62 J. C. Roberts, M. K. Bhalgat, R. T. Zera: Preliminary biological evaluation of polyamidoamine (PAMAM) starburst dendrimers, J. Biomed. Mater. Res. **30** (1996) 53–65

10.63 N. S. Nahman, T. Drost, U. Bhatt, T. Sferra, A. Johnson, P. Gamboa, G. Hinkle, A. Haynam, V. Bergdall, C. Hickey, J. D. Bonagura, L. Brannon-Pappas, J. Ellison, A. Mansfield, S. Shiwe, N. Shen: Biodegradable microparticles for in vivo glomerular targeting: Implications for gene therapy of glomerular disease, Biomed. Microdevices **4** (2002) 189–196

10.64 W. Arap, M. Kolonin, M. Trepel, J. Lahdenranta, M. Cardo-Vila, R. Giordano, P. J. Mintz, P. Ardelt, V. Yao, C. Vidal, L. Chen, A. Flamm, H. Valtanen, L. M. Weavind, M. E. Hicks, R. Pollock, G. H. Botz, C. D. Bucana, E. Koivunen, D. Cahil, P. Troncosco, K. A. Baggerly, R. D. Pentz, K.-A. Do, C. Logothetis, R. Pasqualini: Steps towards mapping the human vasculature by phage display, Nature Med. **8** (2002) 121–127

10.65 M. Kolonin, R. Pasqualini, W. Arap: Molecular addresses in blood vessels as targets for therapy, Curr. Opin. Chem. Biol. **5** (2001) 308–313

10.66 E. Ruoslahti: Special delivery of drugs by targeting to tissue-specific receptors in vasculature, Pharmaceutical News **7** (2000) 35–40

10.67 B. K. Kay, J. Winter, J. McCofferty: *Phage Display of Peptides and Proteins* (Academic, San Diego 1996)

10.68 W. Arap, W. Haedicke, M. Bernasconi, R. Kain, D. Rajotte, S. Krajewski, M. Ellerby, R. Pasqualini, E. Ruoslahti: Targeting the prostate for destruction through a vascular address, Proc. Nat. Acad. Sci. USA **99** (2002) 1527–1531

10.69 M. Essier, E. Ruoslahti: Molecular specialization of breast vasculature: A breast homing phage displayed peptide binds to aminopeptidase P in breast vasculature, Proc. Nat. Acad. Sci. USA **99** (2002) 2252–2257

10.70 G. A. Husseini, G. D. Myrup, W. G. Pitt, D. Christensen, N. Y. Rapoport: Factors affecting acoustically triggered release of drugs from polymeric micelles, J. Cont. Release **69** (2000) 43–52

10.71 A. Quintana, E. Raczka, L. Piehler, I. Lee, A. Myc, I. Majoros, A. K. Patri, T. Thomas, J. Mule, J. R. Baker Jr.: Design and function of a dendrimer-based therapeutic nanodevice targeted to tumor cells through the folate receptor, Pharma. Res. **19** (2002) 1310–1316

10.72 P. Shum, J.-M. Kim, D. H. Thompson: Phototriggering of liposomal delivery systems, Adv. Drug Deliv. Rev. **53** (2001) 273–284

10.73 K. Hamad-Schifferli, J. J. Schwartz, A. T. Santos, S. Zhang, J. M. Jacobson: Remote electronic control of DNA hybridization through inductive coupling to an attached metal nanocrystal antenna, Nature **415** (2002) 152–155

10.74 P. Meers: Enzyme-activated targeting of liposomes, Adv. Drug Deliv. Rev. **53** (2001) 265–272

10.75 V. Raso, M. Brown, J. McGrath: Intracellular triggering with low pH-triggered bispecific antibodies, J. Biol. Chem. **272** (1997) 27623–27628

10.76 W. A. Denny: The role of hypoxia activated prodrugs in cancer therapy, The Lancet Oncol. **1** (2000) 25–29

10.77 P. Vaupel, D. K. Kelleher, O. Thews: Modulation of tumor oxygenation, Int. J. Radiation Oncology Biol. Phys. **42** (1998) 843–848

10.78 P. D. Senter, C. J. Springer: Selective activation of anticancer prodrugs by monoclonal antibody-enzyme conjugates, Adv. Drug Deliv. Rev. **53** (2001) 247–264

10.79 H. Wang, H. Song, V. C. Yang: A recombinant prodrug type approach for triggered delivery of streptokinase, J. Cont. Release **59** (1999) 119–122

10.80 R. Duncan: Polymer therapeutics for tumor specific delivery, Chem. Industry **7** (1997) 262–264

10.81 K. Rogers: Principles of affinity-based biosensors, Mol. Biotechnol. **14** (2000) 109–129

10.82 R. Raiteri, M. Grattarola, H.-J. Butt, P. Skladl: Micromechanical cantilever-based biosensors, Sensors Actuat. B **79** (2001) 115–126

10.83 R. Martel, T. Schmidt, H. R. Shea, T. Hertel, P. Avouris: Single and multiwall carbon nanotube field-effect transistors, Appl. Phys. Lett. **73** (1998) 2447–2449

10.84 A. Guiseppi-Elie, C. Lei, R. H. Baughman: Direct electron transfer of glucose oxidase on carbon nanotubes, Nanotechnology **13** (2002) 559–564

10.85 C. N. Campbell: How far are we from detecting single bioconjugation events?. In: *Biological Molecules in Nanotechnology: The Convergence of Biotechnology Polymer Chemistry and Materials Science*, ed. by S. C. Lee, L. Savage (IBC, Southborough 1998) pp. 163–171

10.86 W. D. Callister Jr.: *Material Science and Engineering: an Introduction* (Wiley, New York 1997)

10.87 M. Gerard, A. Chaubey, B. D. Malhotra: Applications of conducting polymers to biosensors, Biosens. Bioelectron. **17** (2002) 345–349

10.88 P. N. Bartlett, Y. Astier: Microelectrochemical enzyme transistors, Chem. Commun. **2** (2000) 105–112

10.89 I. Kleps, A. Angelscu, R. Valisco, D. Dascalu: New micro and nanoelectrode arrays for biomedical applications, Biomed. Microdevices **3** (2001) 29–33

10.90 J. Song, Q. Cheng, S. Zhu, R. C. Stevens: "smart" materials for biosensing devices: cell-mimicking supramolecular assemblies and colorometric detection of pathogenic agents, Biomed. Microdevices **4** (2002) 213–222

10.91 M. Brown, R. Semelka: *MRI: Basic Principles and Applications* (Wiley, New York 1999)

10.92 A. Elster, S. Handel, A. Goldman: *Magnetic Resonance Imaging: A Reference Guide and Atlas* (Lippincott, Philadelphia 1997)

10.93 H. Paajanen, M. Kormano: Contrast agents in magnetic resonance imaging. In: *Radiographic Contrast Agents*, ed. by J. Skucas (Aspen, Rockville 1989) pp. 377–406

10.94 L. Thunus, R. Lejeune: Overview of transition metal and lanthanide complexes as diagnostic tools, Coord. Chem. Rev. **184** (1999) 125–155

10.95 M. Mendoca-Dias, E. Gaggelli, P. Lauterbur: Paramagnetic contrast agents in nuclear magnetic resonance medical imaging, Sem. Nuclear Med. **13** (1983) 364–376

10.96 M. Ollsen, B. Persson, L. Salford: Ferromagnetic particles as contrast agent in T2 NMR imaging, Magn. Reson. Imag. **4** (1986) 437–440

10.97 D. Kehagias, A. Gouliamos, V. Smyrniotis, L. Vlahos: Diagnostic efficacy and safety of MRI of the liver with superparamagnetic iron oxide particles (SH U 555 A), J. Magn. Reson. Im. **14** (2001) 595–601

10.98 C. Nolte-Ernsting, G. Adam, A. Bucker, S. Berges, A. Bjornerud, R. Gunther: Abdominal MR angiography performed using blood pool contrast agents, Am. J. Roentgenol. **171** (1998) 107–113

10.99 F. Rety, O. Clement, N. Siauve, C.-A. Cuenod, F. Carnot, M. Sich, A. Buisine, G. Frija: MR lymphography using iron oxide nanoparticles in rats: pharmacokinetics in the lymphatic system after intravenous injection, J. Magn. Reson. Im. **12** (2000) 734–739

10.100 T. Skotland, P. Sontum, I. Oulie: In vitro stability analyses as a model for metabolism of ferromagnetic particles (Clariscan^{TM}), a contrast agent for magnetic resonance imaging, J. Pharm. Biomed. Anal. **28** (2002) 323–329

10.101 D. Hogemann, L. Josephson, R. Weissleder, J. Basilion: Improvement of MRI probes to allow efficient detection of gene expression, Bioconjugate Chem. **11** (2000) 941–946

10.102 L. L. Muldoon, M. A. Pagel, R. A. Kroll, S. Roman-Goldstein, R. S. Jones, E. A. Neuwelt: A physiological barrier distal to the anatomic blood-brain barrier in a model of transvascular delivery, AJNR Am. J. Neuroradial. **20** (1999) 217–222

10.103 G. Biddlecombe, Y. Gun'ko, J. Kelly, S. Pillai, J. Coey, M. Ventatesan, A. Douvalis: Preparation of magnetic nanoparticles and their assemblies using a new Fe(II) alkoxide precursor, J. Mater. Chem. **11** (2001) 2937–2939

10.104 L. Tiefenauer, G. Kuhne, R. Andres: Antibody-magnetite nanoparticles: in vitro characterization of a potential tumor-specific contrast agent for magnetic resonance imaging, Bioconjugate Chem. **4** (1993) 347–352

10.105 C. Liu, Z. Zhang: Size-dependent superparamagnetic properties of Mn spinel ferrite nanoparticles synthesized from reverse micelles, Chem. Mater. **13** (2001) 2092–2096

10.106 J. Schnorr, M. Taupitz, S. Wagner, H. Pilgrimm, J. Hansel, B. Hamm: Age-related blood half-life of particulate contrast materials: Experimental results with a USPIO in rats, J. Magn. Reson. Im. **12** (2000) 740–744

10.107 R. Weissleder, G. Elizondo, J. Wittenberg, C. Rabito, H. Bengele, L. Josephson: Ultrasmall superparamagnetic iron oxide: Characterization of a new class of contrast agents for MR imaging, Radiology **175** (1990) 489–493

10.108 D. Portet, B. Denizot, E. Rump, J.-J. Lejeune, P. Jallet: Nonpolymeric coatings of iron oxide colloids for biological use as magnetic resonance imaging contrast agents, J. Coll. Inter. Sci. **238** (2001) 37–42

10.109 M. Ruegsegger, R. Marchant: Reduced protein adsorption and platelet adhesion by controlled variation of oligomaltose surfactant polymer coatings, J. Biomed. Mater. Res. **56** (2001) 159–167

10.110 S. Flack, S. Fischer, M. Scott, R. Fuhrhop, J. Allen, M. McLean, P. Winter, G. Sicard, P. Gaffney, S. Wickline, G. Lanza: Novel MRI contrast agent for molecular fibrin, Circulation **104** (2001) 1280–1285

10.111 D. F. Baban, L. W. Seymour: Control of tumor vascular permeability, Adv. Drug Deliv. Rev. **34** (1998) 109–119

10.112 R. K. Jain: Delivery of molecular and cellular medicine to tumors, J. Cont. Release **53** (1998) 49–67

10.113 H. Maeda, T. Sawa, T. Konno: Mechanism of tumor-targeted delivery of macromolecular drugs, including the EPR effect in solid tumor and clinical overview of the prototype polymeric drug SMANCS, J. Cont. Release **74** (2001) 47–61

10.114 J. Borenstein, H. Terai, K. King, E. Weinberg, M. Kaazempour-Mofrad, J. Vacanti: Microfabrication technology for vascularized tissue engineering, Biomed. Microdevices **4** (2002) 167–175

10.115 D. Satake, H. Ebi, N. Oku, K. Matsuda, H. Takao, M. Ashiki, M. Ishida: A sensor for blood cell counters using MEMS technology, Sensors Actuat. B – Chem. **83** (2002) 77–81

10.116 G. Lin, R. Palmer, K. Pister, K. Roos: Miniature heart cell force transducer system implemented in MEMS technology, IEEE Trans. Biomed. Engin. **48** (2001) 996–1006

10.117 J. Kreuter, D. Shamenkov, V. Petrov, P. Ramge, K. Cychutek, C. Koch-Brandt, R. Alyautdin: Apolipoprotein-mediated transport of nanoparticle-bound drugs across the blood-brain barrier, J. Drug Target **10** (2002) 317–325

10.118 P. Lockman, R. Mumper, M. Khan, D. Allen: Nanoparticle technology for drug delivery across the blood-brain barrier, Drug Devel. Indust. Pharm. **28** (2002) 1–13

10.119 H. Quick, J. Debatin, M. Ladd: MR imaging of the vessel wall, Eur. Radiol. **12** (2002) 889–900

10.120 S. Schmitz, M. Taupitz, S. Wagner, K.-J. Wolf, D. Beyersdorff, D. Hamm: Magnetic resonance imaging of atherosclerotic plaques using supermagnetic iron oxide particles, J. Magn. Reson. Im. **14** (2001) 355–361

10.121 D. Ranney: Biomimetic transport and rational drug delivery, Biochem. Pharm. **59** (2000) 105–114

10.122 K. Soppimath, T. Aminabhavi, A. Kulkarni, W. Rudzinski: Biodegradable polymeric nanoparticles as drug delivery devices, J. Cont. Release **70** (2001) 1–20

10.123 F. Breitling, S. Dubel: *Recombinant Antibodies* (Wiley, London 1998)

10.124 E. Harlow, D. Lane: *Antibodies: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor 1988)

10.125 S. Raychanduri, K. L. Rock: Fully mobilizing host defense: Building better vaccines, Nature Biotech. **16** (1998) 1025–1031

10.126 M. Kovasovics-Bankowski, K. Clark, B. Benacerraf, K. L. Rock: Efficient major histocompatibility complex class I presentation of exogenous antigen upon phagocytosis by macrophages, Proc. Nat. Acad. Sci. USA **90** (1993) 4942–4946

10.127 K. Rock, S. Gamble, L. Rothstein: Presentation of exogenous antigen with class I major histocompatibility complex molecules, Science **249** (1990) 918–921

10.128 C. V. Harding, R. Song: Phagocytic processing of exogenous particulate antigens by macrophages for presentation by class I MHC molecules, J. Immunol. **152** (1994) 4925–4933

10.129 C. Oseroff, A. Sette, P. Wentworth, E. Celis, A. Maewal, C. Dahlberg, J. Fikes, R. T. Kubo, R. W. Chestnut, H. M. Grey, J. Alexander: Pools of lapidated HTL-CTL constructs prime for multiple HBV and HCV CTL epitope responses, Vaccine **16** (1998) 823–833

10.130 R. Mrsny, A. L. Daughtery, M. Mckee, D. Fitzgerald: Bacterial toxins as tools for mucosal vaccination, Drug Disc. Today **7** (2000) 247–257

10.131 S. Fawell, J. Seery, Y. Daikh, C. Moore, L. L. Chen, B. Pepinsky, J. Barsoum: Tat-mediated delivery of heterologous proteins to cells, Proc. Nat. Acad. Sci. USA **91** (1994) 664–668

10.132 T. J. Golentz, K. Klimpel, S. Leppla, J. M. Keith, J. A. Berzofsky: Delivery of antigens to the MHC class I pathway using bacterial toxins, Hum. Immunol. **54** (1997) 129–136

10.133 B.-X. Chen, S. R. Wilson, M. Das, D. J. Coughlin, B. F. Erlanger: Antigenicity of fullerenes: antibodies specific for fullerenes and their characteristics, Proc. Nat. Acad. Sci. USA **95** (1998) 10809–10813

10.134 S. C. Lee, R. Ibdah, C. D. van Valkenburgh, E. Rowold, A. Donelly, A. Abegg, J. Klover, S. Merlin, J. McKearn: Phage display mutagenesis of the chimeric dual cytokine receptor agonist myelopoietin, Leukemia **15** (2001) 1277–1285

10.135 S. C. Lee, R. Parthasarathy, K. Botwin, D. Kunneman, E. Rowold, G. Lange, J. Zobel, T. Beck, T. Miller, C. F. Voliva: Humeral immune responses to polymeric nanomaterials. In: *Functional Condensation Polymers*, ed. by C. Carraher, G. Swift (Kluwer, New York 2002) pp. 31–41

10.136 S. C. Lee, R. Parthasarathy, T. Duffin, K. Botwin, T. Beck, G. Lange, J. Zobel, D. Kunneman, E. Rowold, C. F. Voliva: Recognition properties of antibodies to PAMAM dendrimers and their use in immune detection of dendrimers, Biomed. Microdevices **3** (2001) 51–57

10.137 A. H. Sehon: Suppression of antibody responses by conjugates of antigens and monomethoxypoly(ethylene glycol). In: *Poly(ethylene glycol) Chemistry*, ed. by J. M. Harris (Plenum, New York 1992) pp. 139–151

10.138 R. P. Lanza, S. J. Sullivan, W. L. Chick: Perspectives in diabetes. Islet transplantation with immunoisolation, Diabetes **41** (1992) 1503–1510

10.139 R. Gordon: Computer controlled evolution of diatoms: Design for a compustat, Nova Hedwigia **112** (1996) 215–219

10.140 J. Parkinson, R. Gordon: Beyond micromachining: The potential of diatoms, Trends Biotechnol. **17** (1999) 190–196

10.141 S. C. Lee: Antibody responses to nanomaterials. In: *Nanospace 2001: Exploring Interdisciplinary Frontiers*, ed. by T. Nicodemus (Institute for Advanced Interdisciplinary Research, Houston 2002)

10.142 K. Pearce, B. Cunningham, G. Fuh, T. Teeri, J. A. Wells: Growth hormone affinity for its receptor surpasses the requirements for cellular activity, Biochem. **38** (1999) 81–89

10.143 T. L. Ciardelli: Reengineering growth factors through the looking glass, Nat. Biotechnol. **14** (1996) 1652

10.144 P. J. Buchli, Z. Wu, T. L. Ciardelli: Functional display of interleukin-2 on filamentous phage, Arch. Biochem. Biophys. **339** (1997) 79–84

10.145 I. Saggio, I. Gloaguen, R. Laufer: Functional phage display of cilliary neurotropic factor, Gene **152** (1995) 35–39

Part B

Part B Scanning Probe Microscopy

11 **Scanning Probe Microscopy – Principle of Operation, Instrumentation, and Probes**
Bharat Bhushan, Columbus, USA
Othmar Marti, Ulm, Germany

12 **Probes in Scanning Microscopies**
Jason H. Hafner, Houston, USA

13 **Noncontact Atomic Force Microscopy and Its Related Topics**
Seizo Morita, Suita-Citiy, Osaka, Japan
Franz J. Giessibl, Augsburg, Germany
Yasuhiro Sugawara, Suita, Japan
Hirotaka Hosoi, Sapporo, Japan
Koichi Mukasa, Sapporo, Japan
Akira Sasahara, Kanagawa, Japan
Hiroshi Onishi, Kanagawa, Japan

14 **Low Temperature Scanning Probe Microscopy**
Markus Morgenstern, Hamburg, Germany
Alexander Schwarz, Hamburg, Germany
Udo D. Schwarz, New Haven, USA

15 **Dynamic Force Microscopy**
A. Schirmeisen, Münster, Germany
B. Anczykowski, Münster, Germany
Harald Fuchs, Münster, Germany

16 **Molecular Recognition Force Microscopy**
Peter Hinterdorfer, Linz, Austria

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

11. Scanning Probe Microscopy – Principle of Operation, Instrumentation, and Probes

Scanning Pro

Since the introduction of the STM in 1981 and AFM in 1985, many variations of probe based microscopies, referred to as SPMs, have been developed. While the pure imaging capabilities of SPM techniques is dominated by the application of these methods at their early development stages, the physics of probe–sample interactions and the quantitative analyses of tribological, electronic, magnetic, biological, and chemical surfaces have now become of increasing interest. Nanoscale science and technology are strongly driven by SPMs which allow investigation and manipulation of surfaces down to the atomic scale. With growing understanding of the underlying interaction mechanisms, SPMs have found applications in many fields outside basic research fields. In addition, various derivatives of all these methods have been developed for special applications, some of them targeted far beyond microscopy.

This chapter presents an overview of STM and AFM and various probes (tips) used in these instruments, followed by details on AFM instrumentation and analyses.

11.1 Scanning Tunneling Microscope 327
11.1.1 Binnig et al.'s Design.................... 327
11.1.2 Commercial STMs.......................... 328
11.1.3 STM Probe Construction 330

11.2 Atomic Force Microscope 331
11.2.1 Binnig et al.'s Design.................... 333
11.2.2 Commercial AFM........................... 333
11.2.3 AFM Probe Construction 338
11.2.4 Friction Measurement Methods...... 342
11.2.5 Normal Force and Friction Force Calibrations of Cantilever Beams...................... 346

11.3 AFM Instrumentation and Analyses 347
11.3.1 The Mechanics of Cantilevers 347
11.3.2 Instrumentation and Analyses of Detection Systems for Cantilever Deflections 350
11.3.3 Combinations for 3-D-Force Measurements 358
11.3.4 Scanning and Control Systems 359

References 364

The Scanning Tunneling Microscope (STM) developed by *Dr. Gerd Binnig* and his colleagues in 1981 at the IBM Zurich Research Laboratory, Rueschlikon, Switzerland, is the first instrument capable of directly obtaining three-dimensional (3-D) images of solid surfaces with atomic resolution [11.1]. *Binnig* and *Rohrer* received a Nobel Prize in Physics in 1986 for their discovery. STMs can only be used to study surfaces which are electrically conductive to some degree. Based on their design of the STM, in 1985, *Binnig* et al. developed an Atomic Force Microscope (AFM) to measure ultrasmall forces (less than $1\,\mu\text{N}$) present between the AFM tip surface and the sample surface [11.2] (also see [11.3]). AFMs can be used for measurement of all engineering surfaces which may be either electrically conductive or insulating. The AFM has become a popular surface profiler for topographic and normal force measurements on the micro to nanoscale [11.4]. AFMs modified in order to measure both normal and lateral forces, are called Lateral Force Microscopes (LFMs) or Friction Force Microscopes (FFMs) [11.5–11]. FFMs have been further modified to measure lateral forces in two orthogonal directions [11.12–16]. A number of researchers have continued to improve the AFM/FFM designs and used them to measure adhesion and friction of solid and liquid surfaces on micro- and nanoscales [11.4, 17–30]. AFMs have been used for scratching, wear, and measurement of elastic/plastic mechanical properties (such as indentation hardness and the modulus of elasticity) [11.4,10,11,21,23,26–29,31–36]. AFMs have been used for manipulation of individual atoms of Xenon [11.37], molecules [11.38], silicon surfaces [11.39], and polymer surfaces [11.40]. STMs have been used for formation of nanofeatures by localized heating or by inducing chemical reactions under the STM tip [11.41–43] and nanomachining [11.44]. AFMs

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Table 11.1 Comparison of Various Conventional Microscopes with SPMs

	Optical	**SEM/TEM**	**Confocal**	**SPM**
Magnification	10^3	10^7	10^4	10^9
Instrument Price (U.S. \$)	\$10 k	\$250 k	\$30 k	\$100 k
Technology Age	200 yrs	40 yrs	20 yrs	20 yrs
Applications	Ubiquitous	Science and technology	New and unfolding	Cutting edge
Market 1993	\$800 M	\$400 M	\$80 M	\$100 M
Growth Rate	10%	10%	30%	70%

have been used for nanofabrication [11.4, 10, 45–47] and nanomachining [11.48].

STMs and AFMs are used at extreme magnifications ranging from 10^3 to 10^9 in x, y and z directions for imaging macro to atomic dimensions with high resolution information and for spectroscopy. These instruments can be used in any environment such as ambient air [11.2, 49], various gases [11.17], liquid [11.50–52], vacuum [11.1, 53], at low temperatures (lower than about 100 K) [11.54–58] and high temperatures [11.59, 60]. Imaging in liquid allows the study of live biological samples and it also eliminates water capillary forces present in ambient air present at the tip–sample interface. Low temperature (liquid helium temperatures) imaging is useful for the study of biological and organic materials and the study of low-temperature phenomena such as superconductivity or charge-density waves. Low-temperature operation is also advantageous for high-sensitivity force mapping due to the reduction in thermal vibration. They also have been used to image liquids such as liquid crystals and lubricant molecules on graphite surfaces [11.61–64]. While the pure imaging capabilities of SPM techniques dominated the application of these methods at their early development stages, the physics and chemistry of probe–sample interactions and the quantitative analyses of tribological, electronic, magnetic, biological, and chemical surfaces have now become of increasing interest. Nanoscale science and technology are strongly driven by SPMs which allow investigation and manipulation of surfaces down to the atomic scale. With growing understanding of the underlying interaction mechanisms, SPMs have found applications in many fields outside basic research fields. In addition, various derivatives of all these methods have been developed for special applications, some of them targeting far beyond microscopy.

Families of instruments based on STMs and AFMs, called Scanning Probe Microscopes (SPMs), have been developed for various applications of scientific and industrial interest. These include – STM, AFM, FFM (or LFM), scanning electrostatic force microscopy (SEFM) [11.65, 66], scanning force acoustic microscopy (SFAM) (or atomic force acoustic microscopy (AFAM)) [11.21, 22, 36, 67–69], scanning magnetic microscopy (SMM) (or magnetic force microscopy (MFM)) [11.70–73], scanning near field optical microscopy (SNOM) [11.74–77], scanning thermal microscopy (SThM) [11.78–80], scanning electrochemical microscopy (SEcM) [11.81], scanning Kelvin Probe microscopy (SKPM) [11.82–86], scanning chemical potential microscopy (SCPM) [11.79], scanning ion conductance microscopy (SICM) [11.87, 88], and scanning capacitance microscopy (SCM) [11.82, 89–91]. Families of instruments which measure forces (e.g., AFM, FFM, SEFM, SFAM, and SMM) are also referred to as scanning force microscopy (SFM). Although these instruments offer atomic resolution and are ideal for basic research, they are used for cutting edge industrial applications which do not require atomic resolution. Commercial production of SPMs started with the STM in 1987 and the AFM in 1989 by Digital Instruments Inc. For comparisons of SPMs with other microscopes, see Table 11.1 (Veeco Instruments, Inc.). Numbers of these instruments are equally divided between the U.S., Japan and Europe, with the following industry/university and Government labs. splits: 50/50, 70/30, and 30/70, respectively. It is clear that research and industrial applications of SPMs are rapidly expanding.

11.1 Scanning Tunneling Microscope

The principle of electron tunneling was proposed by *Giaever* [11.92]. He envisioned that if a potential difference is applied to two metals separated by a thin insulating film, a current will flow because of the ability of electrons to penetrate a potential barrier. To be able to measure a tunneling current, the two metals must be spaced no more than 10 nm apart. *Binnig* et al. [11.1] introduced vacuum tunneling combined with lateral scanning. The vacuum provides the ideal barrier for tunneling. The lateral scanning allows one to image surfaces with exquisite resolution, lateral-less than 1 nm and vertical-less than 0.1 nm, sufficient to define the position of single atoms. The very high vertical resolution of the STM is obtained because the tunnel current varies exponentially with the distance between the two electrodes, that is, the metal tip and the scanned surface. Typically, tunneling current decreases by a factor of 2 as the separation is increased by 0.2 nm. Very high lateral resolution depends upon sharp tips. *Binnig* et al. overcame two key obstacles for damping external vibrations and for moving the tunneling probe in close proximity to the sample. Their instrument is called the scanning tunneling microscope (STM). Today's STMs can be used in the ambient environment for atomic-scale image of surfaces. Excellent reviews on this subject are presented by *Hansma* and *Tersoff* [11.93], *Sarid* and *Elings* [11.94], *Durig* et al. [11.95]; *Frommer* [11.96], *Güntherodt* and *Wiesendanger* [11.97], *Wiesendanger* and *Güntherodt* [11.98], *Bonnell* [11.99], *Marti* and *Amrein* [11.100], *Stroscio* and *Kaiser* [11.101], and *Güntherodt* et al. [11.102].

The principle of the STM is straightforward. A sharp metal tip (one electrode of the tunnel junction) is brought close enough (0.3–1 nm) to the surface to be investigated (the second electrode) that, at a convenient operating voltage (10 mV–1 V), the tunneling current varies from 0.2 to 10 nA which is measurable. The tip is scanned over a surface at a distance of 0.3–1 nm, while the tunneling current between it and the surface is measured. The STM can be operated in either the constant current mode or the constant height mode, Fig. 11.1. The left-hand column of Fig. 11.1 shows the basic constant current mode of operation. A feedback network changes the height of the tip z to keep the current constant. The displacement of the tip given by the voltage applied to the piezoelectric drives then yields a topographic map of the surface. Alternatively, in the constant height mode, a metal tip can be scanned across a surface at nearly constant height and constant voltage while the current is monitored, as shown in the right-hand column of Fig. 11.1. In this case, the feedback network responds only rapidly enough to keep the average current constant. A current mode is generally used for atomic-scale images. This mode is not practical for rough surfaces. A three-dimensional picture $[z(x, y)]$ of a surface consists of multiple scans $[z(x)]$ displayed laterally from each other in the y direction. It should be noted that if different atomic species are present in a sample, the different atomic species within a sample may produce different tunneling currents for a given bias voltage. Thus the height data may not be a direct representation of the topography of the surface of the sample.

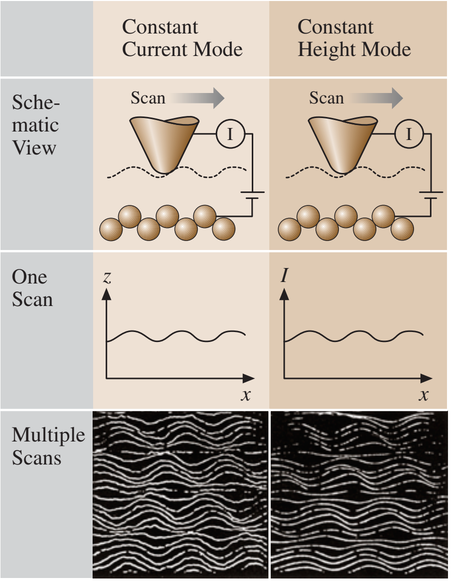

Fig. 11.1 STM can be operated in either the constant-current or the constant-height mode. The images are of graphite in air [11.93]

11.1.1 Binnig et al.'s Design

Figure 11.2 shows a schematic of one of Binnig and Rohrer's designs for operation in ultrahigh vacuum [11.1, 103]. The metal tip was fixed to rectangular piezodrives P_x, P_y, and P_z made out of commercial piezoceramic material for scanning. The sample is mounted via either superconducting magnetic levitation or a two-stage spring system to achieve a stability of the gap width of about 0.02 nm. The tunnel current J_T is a sensitive function of the gap width d that is $J_\mathrm{T} \propto V_\mathrm{T} \exp(-A\phi^{1/2}d)$, where V_T is the bias voltage, ϕ is the average barrier height (work function) and the constant $A = 1.025\,\mathrm{eV}^{-1/2}\,\mathrm{Å}^{-1}$. With a work function of a few eV, J_T changes by an order of magnitude for

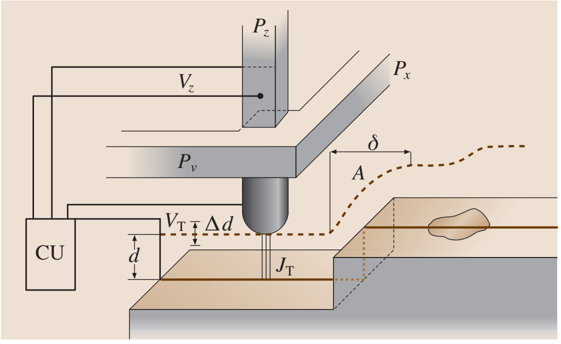

Fig. 11.2 Principle of operation of the STM made by *Binnig* and *Rohrer* [11.103]

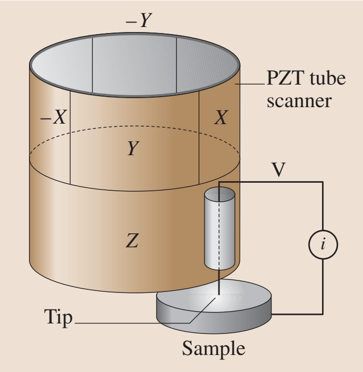

Fig. 11.3 Principle of operation of a commercial STM, a sharp tip attached to a piezoelectric tube scanner is scanned on a sample

every angstrom change of d. If the current is kept constant to within, for example, 2%, then the gap d remains constant to within 1 pm. For operation in the constant current mode, the control unit CU applies a voltage V_z to the piezo P_z such that J_T remains constant when scanning the tip with P_y and P_x over the surface. At constant work function ϕ, $V_z(V_x, V_y)$ yields the roughness of the surface $z(x, y)$ directly, as illustrated at a surface step at A. Smearing the step, δ (lateral resolution) is on the order of $(R)^{1/2}$, where R is the radius of the curvature of the tip. Thus, a lateral resolution of about 2 nm requires tip radii on the order of 10 nm. A 1-mm-diameter solid rod ground at one end at roughly 90° yields overall tip radii of only a few hundred nm, but with closest protrusion of rather sharp microtips on the relatively dull end yields a lateral resolution of about 2 nm. In-situ sharpening of the tips by gently touching the surface brings the resolution down to the 1-nm range; by applying high fields (on the order of 10^8 V/cm) during, for example, half an hour, resolutions considerably below 1 nm could be reached. Most experiments were done with tungsten wires either ground or etched to a radius typically in the range of 0.1–10 µm. In some cases, in-situ processing of the tips was done for further reduction of tip radii.

11.1.2 Commercial STMs

There are a number of commercial STMs available on the market. Digital Instruments, Inc. located in Santa Barbara, CA introduced the first commercial STM, the Nanoscope I, in 1987. In a recent Nanoscope IV STM for operation in ambient air, the sample is held in position while a piezoelectric crystal in the form of a cylindrical tube (referred to as PZT tube scanner) scans the sharp metallic probe over the surface in a raster pattern while sensing and outputting the tunneling current to the control station (Fig. 11.3). The digital signal processor (DSP) calculates the desired separation of the tip from the sample by sensing the tunneling current flowing between the sample and the tip. The bias voltage applied between the sample and the tip encourages the tunneling current to flow. The DSP completes the digital feedback loop by outputting the desired voltage to the piezoelectric tube. The STM operates in both the "constant height" and "constant current" modes depending on a parameter selection in the control panel. In the constant current mode, the feedback gains are set high, the tunneling tip closely tracks the sample surface, and the variation in the tip height required to maintain constant tunneling current is measured by the change in the voltage applied to the piezo tube. In the constant height mode, the feedback gains are set low, the tip remains at a nearly constant height as it sweeps over the sample surface, and the tunneling current is imaged.

Physically, the Nanoscope STM consists of three main parts: the head which houses the piezoelectric tube scanner for three dimensional motion of the tip and the preamplifier circuit (FET input amplifier) mounted on top of the head for the tunneling current, the base on which the sample is mounted, and the base support, which supports the base and head [11.4]. The base accommodates samples up to 10 mm by 20 mm and 10 mm in thickness. Scan sizes available for the STM are 0.7 µm (for atomic resolution), 12 µm, 75 µm and 125 µm square.

The scanning head controls the three dimensional motion of the tip. The removable head consists of a piezo tube scanner, about 12.7 mm in diameter, mounted into an invar shell used to minimize vertical thermal drifts because of a good thermal match between the piezo tube and the Invar. The piezo tube has separate electrodes for X, Y and Z which are driven by separate drive circuits.

The electrode configuration (Fig. 11.3) provides x and y motions which are perpendicular to each other, minimizes horizontal and vertical coupling, and provides good sensitivity. The vertical motion of the tube is controlled by the Z electrode which is driven by the feedback loop. The x and y scanning motions are each controlled by two electrodes which are driven by voltages of the same magnitude, but opposite signs. These electrodes are called $-Y$, $-X$, $+Y$, and $+X$. Applying complimentary voltages allows a short, stiff tube to provide a good scan range without large voltages. The motion of the tip due to external vibrations is proportional to the square of the ratio of vibration frequency to the resonant frequency of the tube. Therefore, to minimize the tip vibrations, the resonant frequencies of the tube are high at about 60 kHz in the vertical direction and about 40 kHz in the horizontal direction. The tip holder is a stainless steel tube with a 300 µm inner diameter for 250 µm diameter tips, mounted in ceramic in order to keep the mass on the end of the tube low. The tip is mounted either on the front edge of the tube (to keep mounting mass low and resonant frequency high) (Fig. 11.3) or the center of the tube for large range scanners, namely 75 and 125 µm (to preserve the symmetry of the scanning). This commercial STM accepts any tip with a 250 µm diameter shaft. The piezotube requires X-Y calibration which is carried out by imaging an appropriate calibration standard. Cleaved graphite is used for the small-scan length head while two dimensional grids (a gold plated ruling) can be used for longer range heads.

The Invar base holds the sample in position, supports the head, and provides coarse X-Y motion for the sample. A spring-steel sample clip with two thumb screws holds the sample in place. An x-y translation stage built into the base allows the sample to be repositioned under the tip. Three precision screws arranged in a triangular pattern support the head and provide coarse and fine adjustment of the tip height. The base support consists of the base support ring and the motor housing. The stepper motor enclosed in the motor housing allows the tip to be engaged and withdrawn from the surface automatically.

Samples to be imaged with the STM must be conductive enough to allow a few nanoamperes of current to flow from the bias voltage source to the area to be scanned. In many cases, nonconductive samples can be coated with a thin layer of a conductive material to facilitate imaging. The bias voltage and the tunneling current depend on the sample. Usually they are set at a standard value for engagement and fine tuned to enhance the quality of the image. The scan size depends on the sample and the features of interest. A maximum scan rate of 122 Hz can be used. The maximum scan rate is usually related to the scan size. Scan rates above 10 Hz are used for small scans (typically 60 Hz for atomic-scale imaging with a 0.7 µm scanner). The scan rate should be lowered for large scans, especially if the sample surfaces are rough or contain large steps. Moving the tip quickly along the sample surface at high scan rates with large scan sizes will usually lead to a tip crash. Essentially, the scan rate should be inversely proportional to the scan size (typically 2–4 Hz for 1 µm, 0.5–1 Hz for 12 µm, and 0.2 Hz for 125 µm scan sizes). Scan rate in length/time, is equal to scan length divided by the scan rate in Hz. For example, for 10 µm × 10 µm scan size scanned at 0.5 Hz, the scan rate is 10 µm/s. Typically, 256 × 256 data formats are most commonly used. The lateral resolution at larger scans is approximately equal to scan length divided by 256.

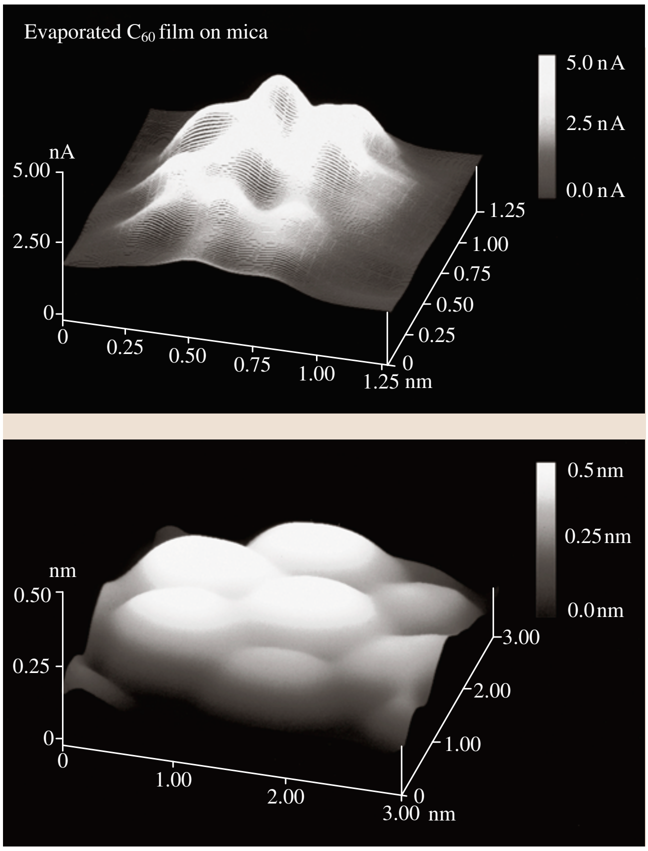

Fig. 11.4 STM images of evaporated C$_{60}$ film on a gold-coated freshly-cleaved mica using a mechanically sheared Pt-Ir (80-20) tip in constant height mode [11.104]

Figure 11.4 shows an example of STM images of an evaporated C_{60} film on a gold-coated freshly-cleaved mica taken at room temperature and ambient pressure [11.104]. Images with atomic resolution at two scan sizes are obtained. Next we describe STM designs which are available for special applications.

Electrochemical STM

The electrochemical STM is used to perform and monitor the electrochemical reactions inside the STM. It includes a microscope base with an integral potentiostat, a short head with a 0.7 µm scan range and a differential preamp and the software required to operate the potentiostat and display the result of the electrochemical reaction.

Standalone STM

The stand alone STMs are available to scan large samples which rest directly on the sample. From Digital instruments, it is available in 12 and 75 µm scan ranges. It is similar to the standard STM except the sample base has been eliminated.

11.1.3 STM Probe Construction

The STM probe should have a cantilever integrated with a sharp metal tip with a low aspect ratio (tip length/tip shank) to minimize flexural vibrations. Ideally, the tip should be atomically sharp, but in practice most tip preparation methods produce a tip with a rather ragged profile that consists of several asperities with the one closest to the surface responsible for tunneling. STM cantilevers with sharp tips are typically fabricated from metal wires of tungsten (W), platinum-iridium (Pt-Ir), or gold (Au) and sharpened by grinding, cutting with a wire cutter or razor blade, field emission/evaporator, ion milling, fracture, or electrochemical polishing/etching [11.105, 106]. The two most commonly used tips are made from either a Pt-Ir (80/20) alloy or tungsten wire. Iridium is used to provide stiffness. The Pt-Ir tips are generally mechanically formed and are readily available. The tungsten tips are etched from tungsten wire with an electrochemical process, for example by using 1 molar KOH solution with a platinum electrode in a electrochemical cell at about 30 V. In general, Pt-Ir tips provide better atomic resolution than tungsten tips, probably due to the lower reactivity of Pt. But tungsten tips are more uniformly shaped and may perform better on samples with steeply sloped features. The tungsten wire diameter used for the cantilever is typically 250 µm with the radius of curvature ranging from 20 to 100 nm and a cone angle ranging from 10 to 60° (Fig. 11.5). The wire can be bent in an *L* shape, if so required, for use in the instrument. For calculations of normal spring constant and natural frequency of round cantilevers, see *Sarid* and *Elings* [11.94].

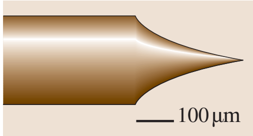

Fig. 11.5 Schematic of a typical tungsten cantilever with a sharp tip produced by electrochemical etching

For imaging of deep trenches, high-aspect-ratio, controlled geometry (CG) Pt-Ir probes are commercially available (Fig. 11.6). These probes are electrochemically etched from Pt-Ir (80/20) wire and polished to a specific shape which is consistent from tip to tip. Probes have a full cone angle of approximately 15°, and a tip radius of less than 50 nm. For imaging of very deep trenches (> 0.25 µm) and nanofeatures, focused ion beam (FIB) milled CG probes with an extremely sharp tip (radius < 5 nm) are used. For electrochemistry, Pt/Ir probes are coated with a nonconducting film (not shown in the figure). These probes are available from Materials Analytical Services, Raleigh, North Carolina.

Pt alloy and W tips are very sharp and have high resolution, but are fragile and sometimes break when contacting a surface. Diamond tips have been used by *Kaneko* and *Oguchi* [11.107]. The diamond tip made conductive by boron ion implantation is found to be chip resistant.

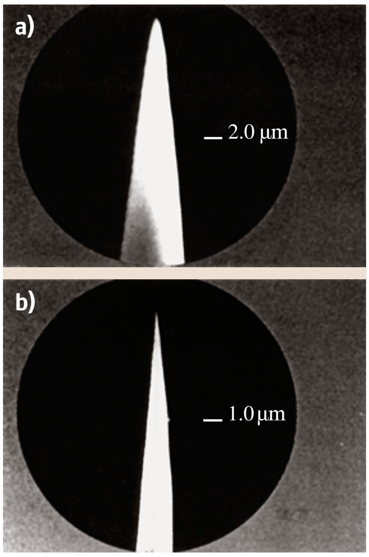

Fig. 11.6a,b Schematics of **(a)** CG Pt-Ir probe, and **(b)** CG Pt-Ir FIB milled probe

11.2 Atomic Force Microscope

Like the STM, the AFM relies on a scanning technique to produce very high resolution, 3-D images of sample surfaces. The AFM measures ultrasmall forces (less than 1 nN) present between the AFM tip surface and a sample surface. These small forces are measured by measuring the motion of a very flexible cantilever beam having an ultrasmall mass. While STMs require that the surface to be measured be electrically conductive, AFMs are capable of investigating surfaces of both conductors and insulators on an atomic scale if suitable techniques for measurement of cantilever motion are used. In the operation of a high resolution AFM, the sample is generally scanned instead of the tip as with the STM, because the AFM measures the relative displacement between the cantilever surface and reference surface and any cantilever movement would add vibrations. For measurements of large samples, AFMs are available where the tip is scanned and the sample is stationary. As long as the AFM is operated in the so-called contact mode, little if any vibration is introduced.

The AFM combines the principles of the STM and the stylus profiler (Fig. 11.7). In an AFM, the force between the sample and tip is detected, rather than the tunneling current, to sense the proximity of the tip to the sample. The AFM can be used either in a static or dynamic mode. In the static mode, also referred to as repulsive mode or contact mode [11.2], a sharp tip at the end of a cantilever is brought in contact with a sample surface. During initial contact, the atoms at the end of the tip experience a very weak repulsive force due to electronic orbital overlap with the atoms in the sample surface. The force acting on the tip causes a cantilever deflection which is measured by tunneling, capacitive, or optical detectors. The deflection can be measured to within 0.02 nm, so for typical cantilever spring constant of 10 N/m a force as low as 0.2 nN (corresponding normal pressure ~ 200 MPa for a Si_3N_4 tip with radius of about 50 nm against single-crystal silicon) can be detected. (To put these number in perspective, individual atoms and human hair are typically a fraction of a nanometer and about 75 μm in diameter, respectively, and a drop of water and an eyelash have a mass of about 10 μN and 100 nN, respectively.) In the dynamic mode of operation for the AFM, also referred to as attractive force imaging or noncontact imaging mode, the tip is brought in close proximity (within a few nm) to, and not in contact with the sample. The cantilever is deliberately vibrated either in amplitude modulation (AM) mode [11.65] or frequency modulation (FM) mode [11.65, 94, 108, 109]. Very weak van der Waals attractive forces are present at the tip–sample interface. Although in this technique, the normal pressure exerted at the interface is zero (desirable to avoid any surface deformation), it is slow, and is difficult to use, and is rarely used outside research environments. In the two modes, surface topography is measured by laterally scanning the sample under the tip while simultaneously measuring the separation-dependent force or force gradient (derivative) between the tip and the surface (Fig. 11.7). In the contact (static) mode, the interaction force between tip and sample is measured by measuring the cantilever deflection. In the noncontact (or dynamic) mode, the force gradient is obtained by vibrating the cantilever and measuring the shift of resonant frequency of the cantilever. To obtain topographic information, the interaction force is either recorded directly, or used as a control parameter for a feedback circuit that maintains the force or force derivative at a constant value. With an AFM operated in the contact mode, topographic images with a vertical resolution of less than 0.1 nm (as low as 0.01 nm) and a lateral resolution of about 0.2 nm have been obtained [11.3, 50, 110–114]. With a 0.01 nm displacement sensitivity, 10 nN to 1 pN forces are measurable. These forces are comparable to the forces associated with chemical bonding, e.g., 0.1 μN for an ionic bond and 10 pN for a hydrogen bond [11.2]. For further reading, see [11.94–96, 100, 102, 115–119].

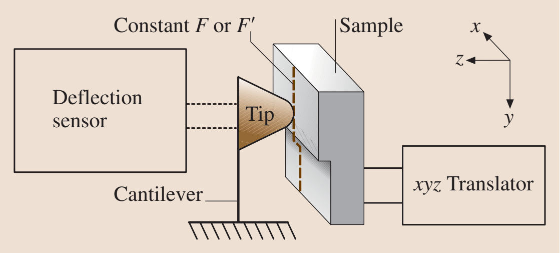

Fig. 11.7 Principle of operation of the AFM. Sample mounted on a piezoelectric tube scanner is scanned against a short tip and the cantilever deflection is measured, mostly, using a laser deflection technique. Force (contact mode) or force gradient (noncontact mode) is measured during scanning

Lateral forces being applied at the tip during scanning in the contact mode, affect roughness measurements [11.120]. To minimize the effects of friction and other lateral forces in the topography measurements in the contact-mode, and to measure the topography of soft surfaces, AFMs can be operated in the so called tapping mode or force modulation mode [11.32, 121].

The STM is ideal for atomic-scale imaging. To obtain atomic resolution with the AFM, the spring constant of the cantilever should be weaker than the equivalent spring between atoms. For example, the vibration frequencies ω of atoms bound in a molecule or in a crystalline solid are typically 10^{13} Hz or higher. Combining this with the mass of the atoms m, on the order of 10^{-25} kg, gives an interatomic spring constants k, given by $\omega^2 m$, on the order of 10 N/m [11.115]. (For comparison, the spring constant of a piece of household aluminum foil that is 4 mm long and 1 mm wide is about 1 N/m.) Therefore, a cantilever beam with a spring constant of about 1 N/m or lower is desirable. Tips have to be as sharp as possible. Tips with a radius ranging from 5 to 50 nm are commonly available.

Atomic resolution cannot be achieved with these tips at the normal load in the nN range. Atomic structures at these loads have been obtained from lattice imaging or by imaging of the crystal periodicity. Reported data show either perfectly ordered periodic atomic structures or defects on a larger lateral scale, but no well-defined, laterally resolved atomic-scale defects like those seen in images routinely obtained with a STM. Interatomic forces with one or several atoms in contact are 20–40 or 50–100 pN, respectively. Thus, atomic resolution with an AFM is only possible with a sharp tip on a flexible cantilever at a net repulsive force of 100 pN or lower [11.122]. Upon increasing the force from 10 pN, *Ohnesorge* and *Binnig* [11.122] observed that monoatomic steplines were slowly wiped away and a perfectly ordered structure was left. This observation explains why mostly defect-free atomic resolution has been observed with AFM. Note that for atomic-resolution measurements, the cantilever should not be too soft to avoid jumps. Further note that measurements in the noncontact imaging mode may be desirable for imaging with atomic resolution.

The key component in an AFM is the sensor for measuring the force on the tip due to its interaction with the sample. A cantilever (with a sharp tip) with extremely low spring constant is required for high vertical and lateral resolutions at small forces (0.1 nN or lower) but at the same time a high resonant frequency is desirable (about 10 to 100 kHz) in order to minimize the sensitivity to vibration noise from the building which is near 100 Hz. This requires a spring with extremely low vertical spring constant (typically 0.05 to 1 N/m) as well as low mass (on the order of 1 ng). Today, the most advanced AFM cantilevers are microfabricated from silicon or silicon nitride using photolithographic techniques. (For further details on cantilevers, see a later section). Typical lateral dimensions are on the order of 100 µm, with the thicknesses on the order of 1 µm. The force on the tip due to its interaction with the sample is sensed by detecting the deflection of the compliant lever with a known spring constant. This cantilever deflection (displacement smaller than 0.1 nm) has been measured by detecting a tunneling current similar to that used in the STM in the pioneering work of *Binnig* et al. [11.2] and later used by *Giessibl* et al. [11.56], by capacitance detection [11.123, 124], piezoresistive detection [11.125, 126], and by four optical techniques namely (1) by optical interferometry [11.5, 6, 127, 128] with the use of optical fibers [11.57, 129] (2) by optical polarization detection [11.72, 130], (3) by laser diode feedback [11.131] and (4) by optical (laser) beam deflection [11.7, 8, 53, 111, 112]. Schematics of the four more commonly used detection systems are shown in Fig. 11.8. The tunneling method originally used by *Binnig* et al. [11.2] in the first version of the AFM, uses a second tip to monitor the deflection of the cantilever with its force sensing tip. Tunneling is rather sensitive to contaminants and the interaction between the tunneling tip and the rear side of the cantilever can become comparable to the interaction between the tip and sample. Tunneling is rarely used and is mentioned first for historical purposes. *Giessibl* et al. [11.56] have used it for a low temperature AFM/STM design. In contrast to tunneling, other deflection sensors are far

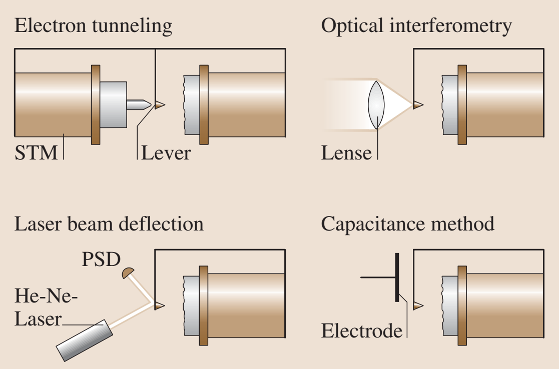

Fig. 11.8 Schematics of the four more commonly used detection systems for measurement of cantilever deflection. In each set up, the sample mounted on piezoelectric body is shown on the right, the cantilever in the middle, and the corresponding deflection sensor on the left [11.118]

away from the cantilever at distances of microns to tens of mm. The optical techniques are believed to be more sensitive, reliable and easily implemented detection methods than others [11.94, 118]. The optical beam deflection method has the largest working distance, is insensitive to distance changes and is capable of measuring angular changes (friction forces), therefore, it is most commonly used in the commercial SPMs.

Almost all SPMs use piezo translators to scan the sample, or alternatively, to scan the tip. An electric field applied across a piezoelectric material causes a change in the crystal structure, with expansion in some directions and contraction in others. A net change in volume also occurs [11.132]. The first STM used a piezo tripod for scanning [11.1]. The piezo tripod is one way to generate three-dimensional movement of a tip attached to its center. However, the tripod needs to be fairly large (~ 50 mm) to get a suitable range. Its size and asymmetric shape makes it susceptible to thermal drift. The tube scanners are widely used in AFMs [11.133]. These provide ample scanning range with a small size. Control electronics systems for AFMs can use either analog or digital feedback. Digital feedback circuits are better suited for ultra low noise operation.

Images from the AFMs need to be processed. An ideal AFM is a noise free device that images a sample with perfect tips of known shape and has a perfectly linear scanning piezo. In reality, scanning devices are affected by distortions and these distortions must be corrected for. The distortions can be linear and nonlinear. Linear distortions mainly result from imperfections in the machining of the piezo translators causing cross talk between the Z-piezo to the X- and Y-piezos, and vice versa. Nonlinear distortions mainly result because of the presence of a hysteresis loop in piezoelectric ceramics. These may also result if the scan frequency approaches the upper frequency limit of the X- and Y-drive amplifiers or the upper frequency limit of the feedback loop (z-component). In addition, electronic noise may be present in the system. The noise is removed by digital filtering in real space [11.134] or in the spatial frequency domain (Fourier space) [11.135].

Processed data consists of many tens of thousand of points per plane (or data set). The output of the first STM and AFM images were recorded on an X-Y chart recorder, with z-value plotted against the tip position in the fast scan direction. Chart recorders have slow response so computers are used to display the data. The data are displayed as a wire mesh display or gray scale display (with at least 64 shades of gray).

11.2.1 Binnig et al.'s Design

In the first AFM design developed by *Binnig* et al. [11.2], AFM images were obtained by measurement of the force on a sharp tip created by the proximity to the surface of the sample mounted on a 3-D piezoelectric scanner. Tunneling current between STM tip and the backside of the cantilever beam with attached tip was measured to obtain the normal force. This force was kept at a constant level with a feedback mechanism. The STM tip was also mounted on a piezoelectric element to maintain the tunneling current at a constant level.

11.2.2 Commercial AFM

A review of early designs of AFMs is presented by *Bhushan* [11.4]. There are a number of commercial AFMs available on the market. Major manufacturers of AFMs for use in an ambient environment are: Digital Instruments Inc., a subsidiary of Veeco Instruments, Inc., Santa Barbara, California; Topometrix Corp., a subsidiary of Veeco Instruments, Inc., Santa Clara, California; other subsidiaries of Veeco Instruments Inc., Woodbury, New York; Molecular Imaging Corp., Phoenix, Arizona; Quesant Instrument Corp., Agoura Hills, California; Nanoscience Instruments Inc., Phoenix, Arizona; Seiko Instruments, Japan; and Olympus, Japan. AFM/STMs for use in UHV environments are manufactured by Omicron Vakuumphysik GMBH, Taunusstein, Germany.

We describe here two commercial AFMs – small sample and large sample AFMs – for operation in the contact mode, produced by Digital Instruments, Inc., Santa Barbara, CA, with scanning lengths ranging from about 0.7 μm (for atomic resolution) to about 125 μm [11.9, 111, 114, 136]. The original design of these AFMs comes from *Meyer* and *Amer* [11.53]. Basically the AFM scans the sample in a raster pattern while outputting the cantilever deflection error signal to the control station. The cantilever deflection (or the force) is measured using laser deflection technique (Fig. 11.9). The DSP in the workstation controls the z position of the piezo based on the cantilever deflection error signal. The AFM operates in both the "constant height" and "constant force" modes. The DSP always adjusts the height of the sample under the tip based on the cantilever deflection error signal, but if the feedback gains are low the piezo remains at a nearly "constant height" and the cantilever deflection data is collected. With the high gains the piezo height changes to keep the cantilever deflection nearly constant (therefore the force is

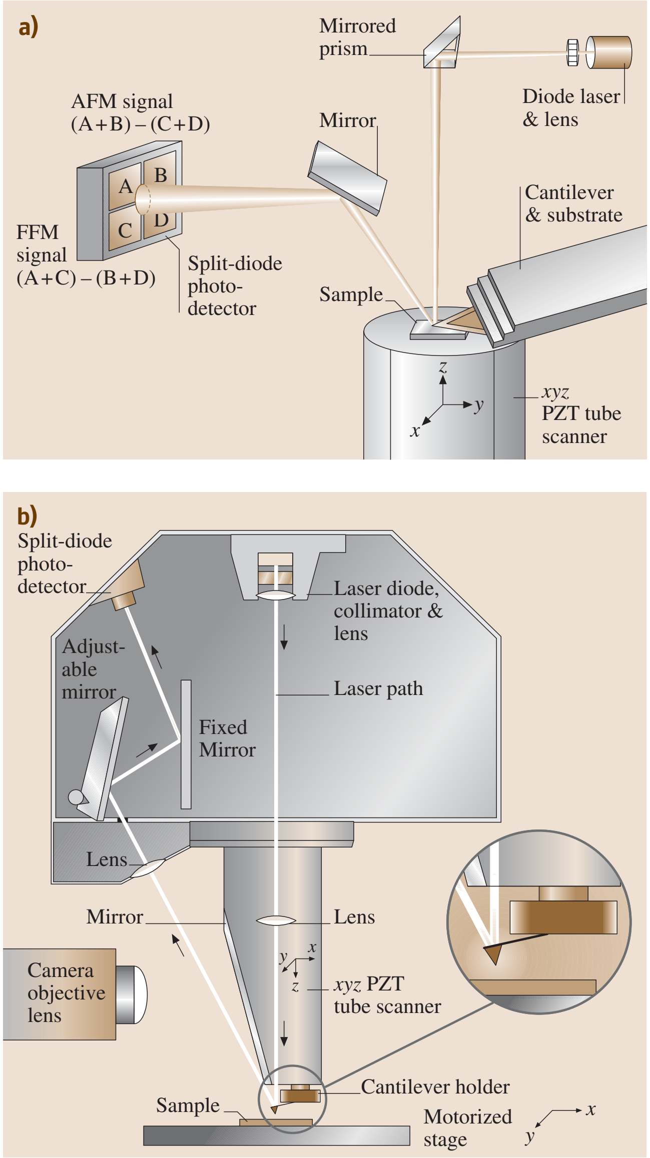
a)
Mirrored prism
Diode laser & lens
AFM signal (A+B) – (C+D)
Mirror
A
B
C
D
Cantilever & substrate
FFM signal (A+C) – (B+D)
Split-diode photo-detector
Sample
z
y
x
xyz PZT tube scanner
b)
Split-diode photo-detector
Laser diode, collimator & lens
Adjust-able mirror
Laser path
Fixed Mirror
Lens
Mirror
Lens
Camera objective lens
y
x
z
xyz PZT tube scanner
Sample
Cantilever holder
Motorized stage
x
y

Fig. 11.9a,b Principles of operation of **(a)** a commercial small sample AFM/FFM, and **(b)** a large sample AFM/FFM

constant) and the change in piezo height is collected by the system.

To further describe the principle of operation of the commercial small sample AFM shown in Fig. 11.9a, the sample, generally no larger than 10 mm × 10 mm, is mounted on a PZT tube scanner which consists of separate electrodes to scan precisely the sample in the x-y plane in a raster pattern and to move the sample in the vertical (z) direction. A sharp tip at the free end of a flexible cantilever is brought in contact with the sample. Features on the sample surface cause the cantilever to deflect in the vertical and lateral directions as the sample moves under the tip. A laser beam from a diode laser (5 mW max peak output at 670 nm) is directed by a prism onto the back of a cantilever near its free end, tilted downward at about 10° with respect to the horizontal plane. The reflected beam from the vertex of the cantilever is directed through a mirror onto a quad photodetector (split photodetector with four quadrants) (commonly called position-sensitive detector or PSD, produced by Silicon Detector Corp., Camarillo, California). The differential signal from the top and bottom photodiodes provides the AFM signal which is a sensitive measure of the cantilever vertical deflection. Topographic features of the sample cause the tip to deflect in the vertical direction as the sample is scanned under the tip. This tip deflection will change the direction of the reflected laser beam, changing the intensity difference between the top and bottom sets of photodetectors (AFM signal). In the AFM operating mode called the height mode, for topographic imaging or for any other operation in which the applied normal force is to be kept a constant, a feedback circuit is used to modulate the voltage applied to the PZT scanner to adjust the height of the PZT, so that the cantilever vertical deflection (given by the intensity difference between the top and bottom detector) will remain constant during scanning. The PZT height variation is thus a direct measure of the surface roughness of the sample.

In a large sample AFM, both force sensors using optical deflection methods and scanning unit are mounted on the microscope head (Fig. 11.9b). Because of vibrations added by cantilever movement, lateral resolution of this design is somewhat poorer than the design in Fig. 11.9a in which the sample is scanned instead of the cantilever beam. The advantage of the large sample AFM is that large samples can be measured readily.

Most AFMs can be used for topography measurements in the so-called tapping mode (intermittent contact mode), also referred to as dynamic force microscopy. In the tapping mode, during scanning over the surface, the cantilever/tip assembly is sinusoidally vibrated by a piezo mounted above it, and the oscillating tip slightly taps the surface at the resonant frequency of the cantilever (70–400 Hz) with a constant (20–100 nm) oscillating amplitude introduced in the vertical direction with a feedback loop keeping the average normal force constant (Fig. 11.10). The oscillating amplitude is kept

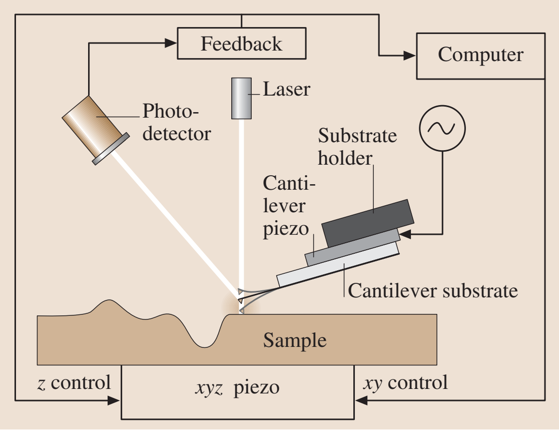

Fig. 11.10 Schematic of tapping mode used for surface roughness measurements

large enough so that the tip does not get stuck to the sample because of adhesive attractions. The tapping mode is used in topography measurements to minimize effects of friction and other lateral forces to measure the topography of soft surfaces.

Topographic measurements are made at any scanning angle. At a first instance, scanning angle may not appear to be an important parameter. However, the friction force between the tip and the sample will affect the topographic measurements in a parallel scan (scanning along the long axis of the cantilever). Therefore a perpendicular scan may be more desirable. Generally, one picks a scanning angle which gives the same topographic data in both directions; this angle may be slightly different than that for the perpendicular scan.

For measurement of the friction force being applied at the tip surface during sliding, the left hand and right hand sets of quadrants of the photodetector are used. In the so-called friction mode, the sample is scanned back and forth in a direction orthogonal to the long axis of the cantilever beam. A friction force between the sample and the tip will produce a twisting of the cantilever. As a result, the laser beam will be reflected out of the plane defined by the incident beam and the beam reflected vertically from an untwisted cantilever. This produces an intensity difference of the laser beam received in the left hand and right hand sets of quadrants of the photodetector. The intensity difference between the two sets of detectors (FFM signal) is directly related to the degree of twisting and hence to the magnitude of the friction force. This method provides three-dimensional maps of the friction force. One problem associated with this method is that any misalignment between the laser beam and the photodetector axis would introduce error in the measurement. However, by following the procedures developed by *Ruan* and *Bhushan* [11.136], in which the average FFM signal for the sample scanned in two opposite directions is subtracted from the friction profiles of each of the two scans, the misalignment effect is eliminated. By following the friction force calibration procedures developed by *Ruan* and *Bhushan* [11.136], voltages corresponding to friction forces can be converted to force units. The coefficient of friction is obtained from the slope of friction force data measured as a function of normal loads typically ranging from 10 to 150 nN. This approach eliminates any contributions due to the adhesive forces [11.10]. For calculation of the coefficient of friction based on a single point measurement, friction force should be divided by the sum of applied normal load and intrinsic adhesive force. Furthermore, it should be pointed out that for a single asperity contact, the coefficient of friction is not independent of load. For further details, refer to a later section.

The tip is scanned in such a way that its trajectory on the sample forms a triangular pattern (Fig. 11.11). Scanning speeds in the fast and slow scan directions depend on the scan area and scan frequency. Scan sizes ranging from less than 1 nm × 1 nm to 125 μm × 125 μm and scan rates from less than 0.5 to 122 Hz typically can be used. Higher scan rates are used for smaller scan lengths. For example, scan rates in the fast and slow scan directions for an area of 10 μm × 10 μm scanned at 0.5 Hz are 10 μm/s and 20 nm/s, respectively.

We now describe the construction of a small sample AFM in more detail. It consists of three main parts: the optical head which senses the cantilever deflection, a PZT tube scanner which controls the scanning motion of the sample mounted on its one end, and the base

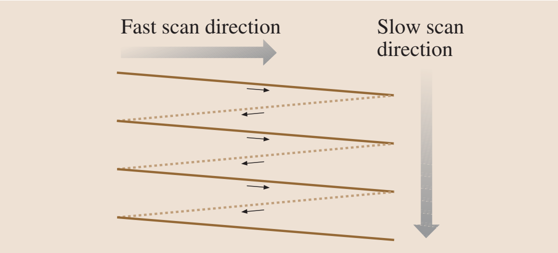

Fig. 11.11 Schematic of triangular pattern trajectory of the AFM tip as the sample is scanned in two dimensions. During imaging, data are recorded only during scans along the solid scan lines

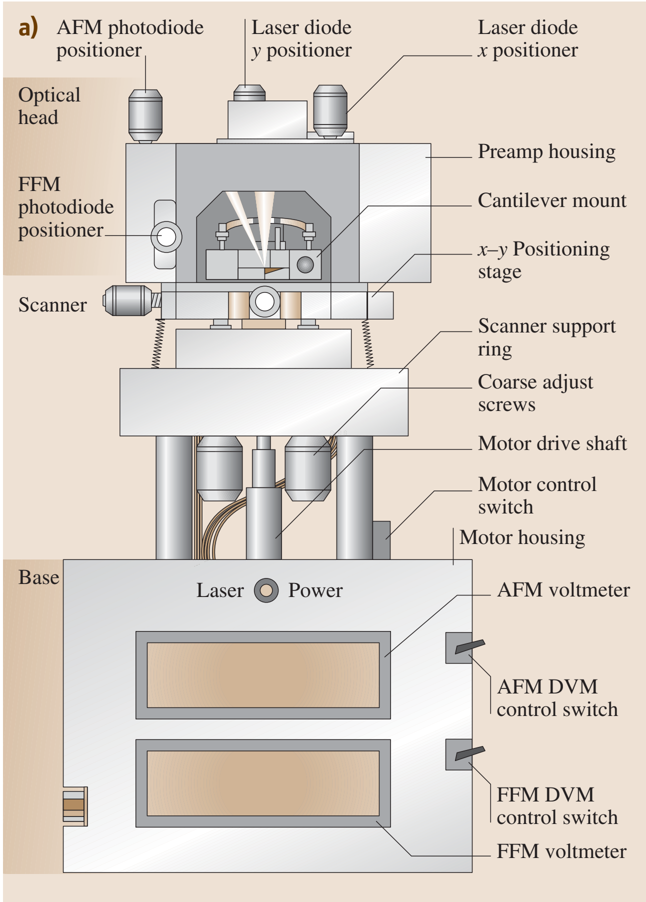
a)
AFM photodiode positioner
Laser diode y positioner
Laser diode x positioner
Optical head
Preamp housing
Cantilever mount
FFM photodiode positioner
x–y Positioning stage
Scanner
Scanner support ring
Coarse adjust screws
Motor drive shaft
Motor control switch
Motor housing
Base
Laser
Power
AFM voltmeter
AFM DVM control switch
FFM DVM control switch
FFM voltmeter

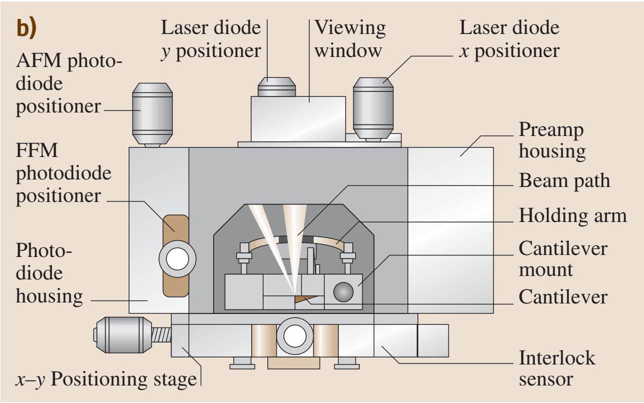
b)
Laser diode y positioner
Viewing window
Laser diode x positioner
AFM photo-diode positioner
Preamp housing
FFM photodiode positioner
Beam path
Holding arm
Photo-diode housing
Cantilever mount
Cantilever
x–y Positioning stage
Interlock sensor

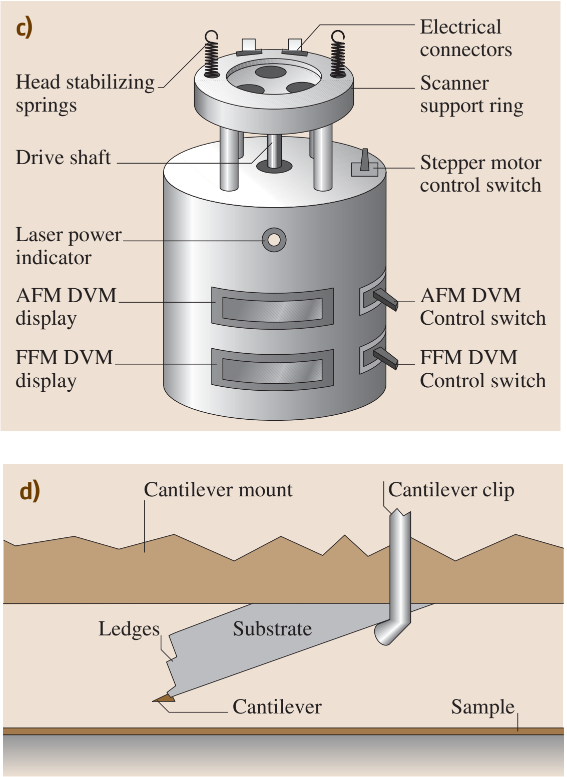
c)
Electrical connectors
Head stabilizing springs
Scanner support ring
Drive shaft
Stepper motor control switch
Laser power indicator
AFM DVM display
AFM DVM Control switch
FFM DVM display
FFM DVM Control switch
d)
Cantilever mount
Cantilever clip
Ledges
Substrate
Cantilever
Sample

Fig. 11.12a–d Schematics of a commercial AFM/FFM made by Digital Instruments Inc. **(a)** front view, **(b)** optical head, **(c)** base, and **(d)** cantilever substrate mounted on cantilever mount (not to scale)

which supports the scanner and head and includes circuits for the deflection signal (Fig. 11.12a). The AFM connects directly to a control system. The optical head consists of laser diode stage, photodiode stage preamp board, cantilever mount and its holding arm, and deflection beam reflecting mirror (Fig. 11.12b). The laser diode stage is a tilt stage used to adjust the position of the laser beam relative to the cantilever. It consists of the laser diode, collimator, focusing lens, baseplate, and the *X* and *Y* laser diode positioners. The positioners are used to place the laser spot on the end of the cantilever. The photodiode stage is an adjustable stage used to position the photodiode elements relative to the reflected laser beam. It consists of the split photodiode, the base plate, and the photodiode positioners. The deflection beam reflecting mirror is mounted on the upper left in the interior of the head which reflects the deflected beam toward the photodiode. The cantilever mount is a metal (for operation in air) or glass (for operation in water) block which holds the cantilever firmly at the proper angle (Fig. 11.12d). Next, the tube scanner consists of an Invar cylinder holding a single tube made of piezoelectric crystal which provides the necessary three-dimensional motion to the sample. Mounted on top of the tube is a magnetic cap on which the steel sample puck is placed. The tube is rigidly held at one end with the sample mounted on the other end of the tube. The scanner also contains three fine-pitched screws which form the mount for the optical head. The optical head rests on the tips of the screws which are used to adjust the position of the head relative to the sample. The scanner fits into the scanner support ring mounted on the base of the microscope (Fig. 11.12c). The stepper motor is controlled manually with the switch on the upper surface of the base and automatically by the computer during the tip engage and tip-withdraw processes.

The scan sizes available for these instruments are 0.7 µm, 12 µm and 125 µm. The scan rate must be decreased as the scan size is increased. A maximum scan rate of 122 Hz can be used. Scan rates of about 60 Hz should be used for small scan lengths (0.7 µm). Scan rates of 0.5 to 2.5 Hz should be used for large scans on samples with tall features. High scan rates help reduce drift, but they can only be used on flat samples with small scan sizes. Scan rate, or scanning speed in length/time in the fast scan direction, is equal to twice the scan length times the scan rate in Hz, and in the slow direction, it is equal to scan length times the scan rate in Hz divided by number of data points in the transverse direction. For example, for 10 µm × 10 µm scan size scanned at 0.5 Hz, the scan rates in the fast and slow scan directions are 10 µm/s and 20 nm/s, respectively. Normally 256 × 256 data points are taken for each image. The lateral resolution at larger scans is approximately equal to scan length divided by 256. The piezo tube requires x-y calibration which is carried out by imaging an appropriate calibration standard. Cleaved graphite is used for small scan heads while two-dimensional grids (a gold plating ruling) can be used for longer range heads.

Examples of AFM images of freshly-cleaved highly-oriented pyrolytic (HOP) graphite and mica surfaces are shown in Fig. 11.13 [11.50, 110, 114]. Images with near atomic resolution are obtained.

The force calibration mode is used to study interactions between the cantilever and the sample surface. In the force calibration mode, the X and Y voltages applied to the piezo tube are held at zero and a sawtooth voltage is applied to the Z electrode of the piezo tube, Fig. 11.14a. The force measurement starts with the sample far away and the cantilever in its rest position. As a result of the applied voltage, the sample is moved up and down relative to the stationary cantilever tip. As the

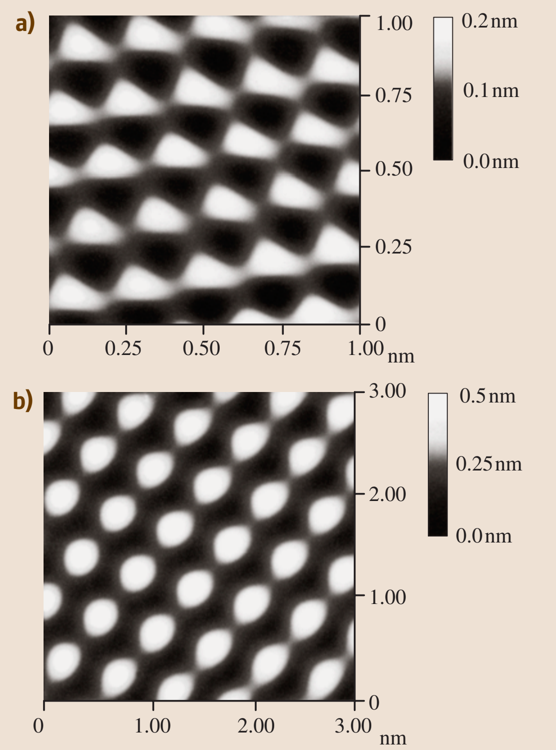

Fig. 11.13a,b Typical AFM images of freshly-cleaved **(a)** highly-oriented pyrolytic graphite and **(b)** mica surfaces taken using a square pyramidal Si_3N_4 tip

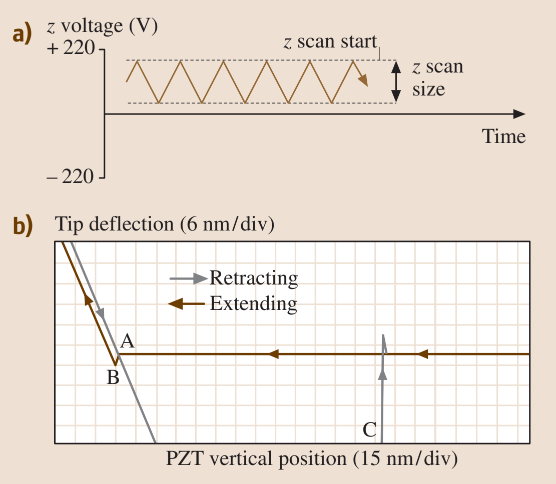

Fig. 11.14 **(a)** Force calibration Z waveform, and **(b)** a typical force-distance curve for a tip in contact with a sample. Contact occurs at point B; tip breaks free of adhesive forces at point C as the samples moves away from the tip

piezo moves the sample up and down, the cantilever deflection signal from the photodiode is monitored. The force–distance curve, a plot of the cantilever tip deflection signal as a function of the voltage applied to the piezo tube, is obtained. Figure 11.14b shows a typical force–distance curve showing the various features of the curve. The arrow heads reveal the direction of piezo travel. As the piezo extends, it approaches the tip, which is at this point in free air and hence shows no deflection. This is indicated by the flat portion of the curve. As the tip approaches the sample within a few nanometers (point A), an attractive force exists between the atoms of the tip surface and the atoms of the sample surface. The tip is pulled towards the sample and contact occurs at point B on the graph. From this point on, the tip is in contact with the surface and as the piezo further extends, the tip gets further deflected. This is represented by the sloped portion of the curve. As the piezo retracts, the tip goes beyond the zero deflection (flat) line because of attractive forces (van der Waals forces and long range meniscus forces), into the adhesive regime. At point C in the graph, the tip snaps free of the adhesive forces, and is again in free air. The horizontal distance between point B and C along the retrace line gives the distance moved by the tip in the adhesive regime. This distance multiplied by the stiffness of the cantilever gives the adhesive force. Incidentally, the horizontal shift between the loading and unloading curves results from the hysteresis in the PZT tube [11.4].

Multimode Capabilities

The multimode AFM can be used for topography measurements in the contact mode and tapping mode, described earlier, and for measurements of lateral (friction) force, electric force gradients and magnetic force gradients.

The multimode AFM, used with a grounded conducting tip, can measure electric field gradients by oscillating the tip near its resonant frequency. When the lever encounters a force gradient from the electric field, the effective spring constant of the cantilever is altered, changing its resonant frequency. Depending on which side of the resonance curve is chosen, the oscillation amplitude of the cantilever increases or decreases due to the shift in the resonant frequency. By recording the amplitude of the cantilever, an image revealing the strength of the electric field gradient is obtained.

In the magnetic force microscope (MFM) used with a magnetically-coated tip, static cantilever deflection is detected which occurs when a magnetic field exerts a force on the tip and the MFM images of magnetic materials can be produced. MFM sensitivity can be enhanced by oscillating the cantilever near its resonant frequency. When the tip encounters a magnetic force gradient, the effective spring constant, and hence the resonant frequency, is shifted. By driving the cantilever above or below the resonant frequency, the oscillation amplitude varies as the resonance shifts. An image of magnetic field gradients is obtained by recording the oscillation amplitude as the tip is scanned over the sample.

Topographic information is separated from the electric field gradients and magnetic field images by using a so-called lift mode. Measurements in lift mode are taken in two passes over each scan line. On the first pass, topographical information is recorded in the standard tapping mode where the oscillating cantilever lightly taps the surface. On the second pass, the tip is lifted to a user-selected separation (typically 20–200 nm) between the tip and local surface topography. By using the stored topographical data instead of the standard feedback, the separation remains constant without sensing the surface. At this height, cantilever amplitudes are sensitive to electric field force gradients or relatively weak but long-range magnetic forces without being influenced by topographic features. Two-pass measurements are taken for every scan line, producing separate topographic and magnetic force images.

Electrochemical AFM

This option allows one to perform electrochemical reactions on the AFM. It includes a potentiostat, a fluid cell with a transparent cantilever holder and electrodes, and the software required to operate the potentiostat and display the results of the electrochemical reaction.

11.2.3 AFM Probe Construction

Various probes (cantilevers and tips) are used for AFM studies. The cantilever stylus used in the AFM should meet the following criteria: (1) low normal spring constant (stiffness), (2) a high resonant frequency, (3) a high quality factor of the cantilever Q, (4) high lateral spring constant (stiffness), (5) short cantilever length, (6) incorporation of components (such as mirror) for deflection sensing, and (7) a sharp protruding tip [11.137]. In order to register a measurable deflection with small forces, the cantilever must flex with a relativly low force (on the order of few nN) requiring vertical spring constants of 10^{-2} to 10^{2} N/m for atomic resolution in the contact profiling mode. The data rate or imaging rate in the AFM is limited by the mechanical resonant frequency

Table 11.2 Relevant properties of materials used for cantilevers

Property	**Young's Modulus (E) (GPa)**	**Density (ρg) (kg/m^3)**	**Microhardness (GPa)**	**Speed of sound ($\sqrt{E/\rho}$) (m/s)**
Diamond	900–1,050	3,515	78.4–102	17,000
Si_3N_4	310	3,180	19.6	9,900
Si	130–188	2,330	9–10	8,200
W	350	19,310	3.2	4,250
Ir	530	–	~ 3	5,300

of the cantilever. To achieve a large imaging bandwidth, the AFM cantilever should have a resonant frequency greater than about 10 kHz (30–100 kHz is preferable) in order to make the cantilever the least sensitive part of the system. Fast imaging rates are not just a matter of convenience, since the effects of thermal drifts are more pronounced with slow-scanning speeds. The combined requirements of a low spring constant and a high resonant frequency is met by reducing the mass of the cantilever. The quality factor Q $(= \omega_R/(c/m)$, where ω_R is the resonant frequency of the damped oscillator and c is the damping constant and m is the mass of the oscillator) should have a high value for some applications. For example, resonance curve detection is a sensitive modulation technique for measuring small force gradients in noncontact imaging. Increasing the Q increases the sensitivity of the measurements. Mechanical Q values of 100–1,000 are typical. In contact modes, the Q is of less importance. High lateral spring constant in the cantilever is desirable to reduce the effect of lateral forces in the AFM as frictional forces can cause appreciable lateral bending of the cantilever. Lateral bending results in error in the topography measurements. For friction measurements, cantilevers with less lateral rigidity is preferred. A sharp protruding tip must be formed at the end of the cantilever to provide a well defined interaction with sample over a small area. The tip radius should be much smaller than the radii of corrugations in the sample in order for these to be measured accurately. The lateral spring constant depends critically on the tip length. Additionally, the tip should be centered at the free end.

In the past, cantilevers have been cut by hand from thin metal foils or formed from fine wires. Tips for these cantilevers were prepared by attaching diamond fragments to the ends of the cantilevers by hand, or in the case of wire cantilevers, electrochemically etching the wire to a sharp point. Several cantilever geometries for wire cantilevers have been used. The simplest geometry is the L-shaped cantilever, usually made by bending a wire at a 90° angle. Other geometries include single-V and double-V geometries with a sharp tip attached at the apex of V, and double-X configuration with a sharp tip attached at the intersection [11.31, 138]. These cantilevers can be constructed with high vertical spring constants. For example, double-cross cantilever with an effective spring constant of 250 N/m was used by *Burnham* and *Colton* [11.31]. The small size and low mass needed in the AFM make hand fabrication of the cantilever a difficult process with poor reproducibility. Conventional microfabrication techniques are ideal for constructing planar thin-film structures which have submicron lateral dimensions. The triangular (V-shaped) cantilevers have improved (higher) lateral spring constant in comparison to rectangular cantilevers. In terms of spring constants, the triangular cantilevers are approximately equivalent to two rectangular cantilevers in parallel [11.137]. Although the macroscopic radius of a photolithographically patterned corner is seldom much less than about 50 nm, microscopic asperities on the etched surface provide tips with near atomic dimensions.

Cantilevers have been used from a whole range of materials. Most commonly are cantilevers made of Si_3N_4, Si, and diamond. Young's modulus and the density are the material parameters which determine the resonant frequency, besides the geometry. Table 11.2 shows the relevant properties and the speed of sound, indicative of the resonant frequency for a given shape. Hardness is important to judge the durability of the cantilevers, and is also listed in the table. Materials used for STM cantilevers are also included.

Silicon nitride cantilevers are less expensive than those made of other materials. They are very rugged and well suited to imaging in almost all environments. They are especially compatible to organic and biological materials. Microfabricated silicon nitride triangular beams with integrated square pyramidal tips made of plasma-enhanced chemical vapor deposition (PECVD) are most commonly used [11.137]. Four cantilevers with different

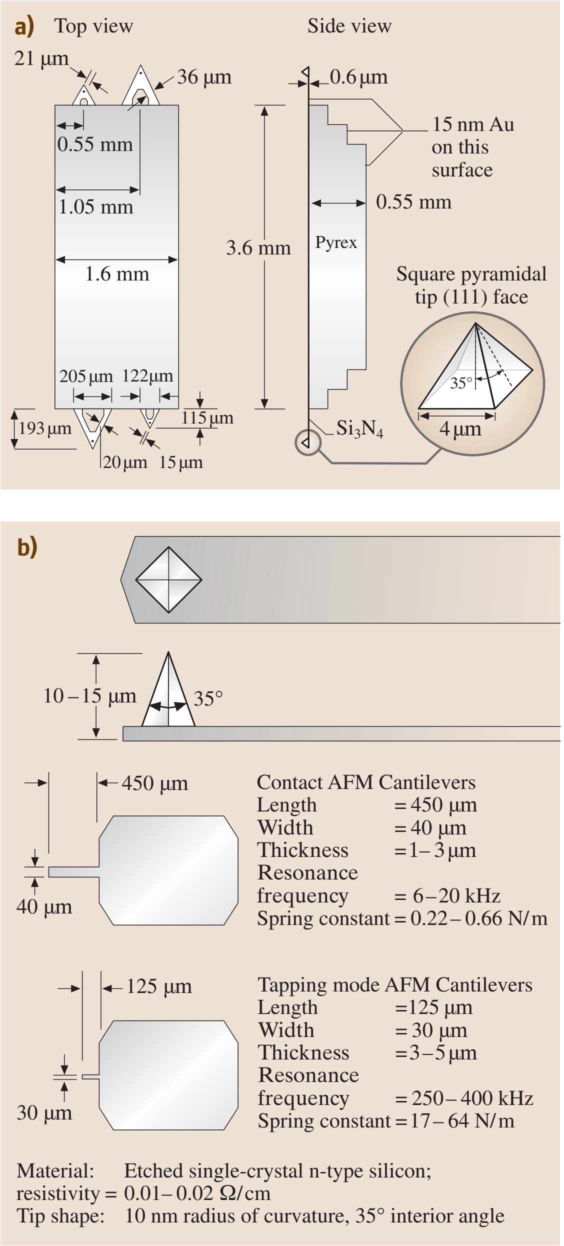
a)
Top view
Side view
21 µm
36 µm
0.6 µm
0.55 mm
15 nm Au on this surface
1.05 mm
0.55 mm
3.6 mm
Pyrex
1.6 mm
Square pyramidal tip (111) face
35°
205 µm
122 µm
193 µm
115 µm
Si3N4
4 µm
20 µm
15 µm
b)
10 – 15 µm
35°
450 µm
40 µm
Contact AFM Cantilevers
Length = 450 µm
Width = 40 µm
Thickness = 1 – 3 µm
Resonance frequency = 6 – 20 kHz
Spring constant = 0.22 – 0.66 N/m
125 µm
30 µm
Tapping mode AFM Cantilevers
Length = 125 µm
Width = 30 µm
Thickness = 3 – 5 µm
Resonance frequency = 250 – 400 kHz
Spring constant = 17 – 64 N/m
Material: Etched single-crystal n-type silicon; resistivity = 0.01 – 0.02 Ω/cm
Tip shape: 10 nm radius of curvature, 35° interior angle

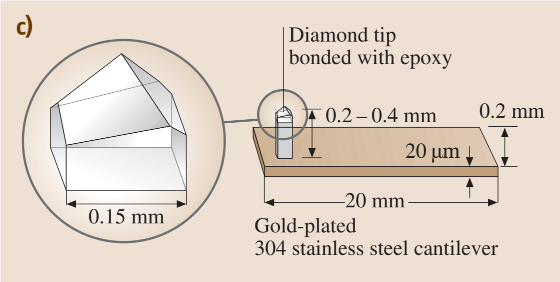
c)
Diamond tip bonded with epoxy
0.2 – 0.4 mm
0.2 mm
20 µm
0.15 mm
20 mm
Gold-plated 304 stainless steel cantilever

Fig. 11.15a–c Schematics of **(a)** triangular cantilever beam with square pyramidal tips made of PECVD Si_3N_4, **(b)** rectangular cantilever beams with square pyramidal tips made of etched single-crystal silicon, and **(c)** rectangular cantilever stainless steel beam with three-sided pyramidal natural diamond tip

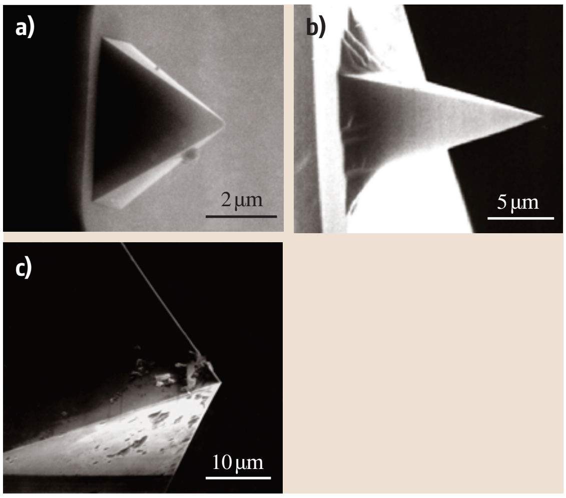
a)
2 µm
b)
5 µm
c)
10 µm

Fig. 11.16a–c SEM micrographs of a square-pyramidal PECVD Si_3N_4 tip **(a)**, a square pyramidal etched single-crystal silicon tip **(b)**, and a three-sided pyramidal natural diamond tip **(c)**

sizes and spring constants on each cantilever substrate made of boron silicate glass (Pyrex), marketed by Digital Instruments, are shown in Figs. 11.15a and 11.16. Two pairs of the cantilevers on each substrate measure about 115 and 193 µm from the substrate to the apex of the triangular cantilever with base widths of 122 and 205 µm, respectively. Both cantilever legs, with the same thickness (0.6 µm) of all the cantilevers, are available with wide and narrow legs. Only one cantilever is selected and used from each substrate. Calculated spring constant and measured natural frequencies for each of the configurations are listed in Table 11.3. The most commonly used cantilever beam is the 115-µm long, wide-legged cantilever (vertical spring constant = 0.58 N/m). Cantilevers with smaller spring constants should be used on softer samples. The pyramidal tips are highly symmetric with its end having a radius of about 20–50 nm. The

Table 11.3 Measured vertical spring constants and natural frequencies of triangular (V-shaped) cantilevers made of PECVD Si_3N_4 (Data provided by Digital Instruments, Inc.)

Cantilever dimension	Spring constant (k_z) (N/m)	Natural frequency (ω_0) (kHz)
115-µm long, narrow leg	0.38	40
115-µm long, wide leg	0.58	40
193-µm long, narrow leg	0.06	13–22
193-µm long, wide leg	0.12	13–22

Table 11.4 Vertical (k_z), lateral (k_y), and torsional ($k_{y\mathrm{T}}$) spring constants of rectangular cantilevers made of Si (IBM) and PECVD Si_3N_4 (Veeco Instruments, Inc.)

Dimensions/stiffness	Si cantilever	Si_3N_4 cantilever
Length (L) (µm)	100	100
Width (b) (µm)	10	20
Thickness (h) (µm)	1	0.6
Tip length (ℓ) (µm)	5	3
k_z (N/m)	0.4	0.15
k_y (N/m)	40	175
$k_{y\mathrm{T}}$ (N/m)	120	116
ω_0 (kHz)	~ 90	~ 65

Note: $k_z = Ebh^3/4L^3$, $k_y = Eb^3h/4\ell^3$, $k_{y\mathrm{T}} = Gbh^3/3L\ell^2$, and $\omega_0 = [k_z/(m_\mathrm{c}+0.24bhL\rho)]^{1/2}$, where E is Young's modulus, G is the modulus of rigidity [$= E/2(1+\nu)$, ν is Poisson's ratio], ρ is the mass density of the cantilever, and m_c is the concentrated mass of the tip (~ 4 ng) [11.94]. For Si, $E = 130$ GPa, $\rho g = 2{,}300\,\mathrm{kg/m^3}$, and $\nu = 0.3$. For Si_3N_4, $E = 150$ GPa, $\rho g = 3{,}100\,\mathrm{kg/m^3}$, and $\nu = 0.3$

tip side walls have a slope of 35 deg and the length of the edges of the tip at the cantilever base is about 4 µm.

An alternative to silicon nitride cantilevers with integrated tips are microfabricated single-crystal silicon cantilevers with integrated tips. Si tips are sharper than Si_3N_4 tips because they are directly formed by the anisotropic etch in single-crystal Si rather than using an etch pit as a mask for deposited materials [11.139]. Etched single-crystal *n*-type silicon rectangular cantilevers with square pyramidal tips with a lower radius of less than 10 nm for contact and tapping mode (tapping-mode etched silicon probe or TESP) AFMs are commercially available from Digital Instruments and Nanosensors GmbH, Aidlingen, Germany, Figs. 11.15b and 11.16. Spring constants and resonant frequencies are also presented in the Fig. 11.15b.

Commercial triangular Si_3N_4 cantilevers have a typical width-thickness ratio of 10 to 30 which results in 100 to 1000 times stiffer spring constants in the lateral direction compared to the normal direction. Therefore these cantilevers are not well suited for torsion. For friction measurements, the torsional spring constant should be minimized in order to be sensitive to the lateral forces. Rather long cantilevers with small thickness and large tip length are most suitable. Rectangular beams have lower torsional spring constants in comparison to the triangular (V-shaped) cantilevers. Table 11.4 lists the spring constants (with full length of the beam used) in three directions of the typical rectangular beams. We note that lateral and torsional spring constants are about two orders of magnitude larger than the normal spring constants. A cantilever beam required for the tapping mode is quite stiff and may not be sensitive enough for friction measurements. *Meyer* et al. [11.140] used a specially designed rectangular silicon cantilever with length = 200 µm, width = 21 µm, thickness = 0.4 µm, tip length = 12.5 µm, and shear modulus = 50 GPa, giving a normal spring constant of 0.007 N/m and torsional spring constant of 0.72 N/m which gives a lateral force sensitivity of 10 pN and an angle of resolution of 10^{-7} rad. With this particular geometry, sensitivity to lateral forces could be improved by about a factor of 100 compared with commercial V-shaped Si_3N_4 or rectangular Si or Si_3N_4 cantilevers used by *Meyer* and *Amer* [11.8] with torsional spring constant of ~ 100 N/m. *Ruan* and *Bhushan* [11.136] and *Bhushan* and *Ruan* [11.9] used 115-µm long, wide-legged V-shaped cantilevers made of Si_3N_4 for friction measurements.

For scratching, wear, and indentation studies, single-crystal natural diamond tips ground to the shape of a three-sided pyramid with an apex angle of either 60° or 80° whose point is sharpened to a radius of about 100 nm are commonly used [11.4, 10] (Figs. 11.15c and 11.16). The tips are bonded with conductive epoxy to a gold-plated 304 stainless steel spring sheet (length = 20 mm, width = 0.2 mm, thickness = 20 to 60 µm) which acts as a cantilever. The free length of the spring is varied to change the beam stiffness. The normal spring con-

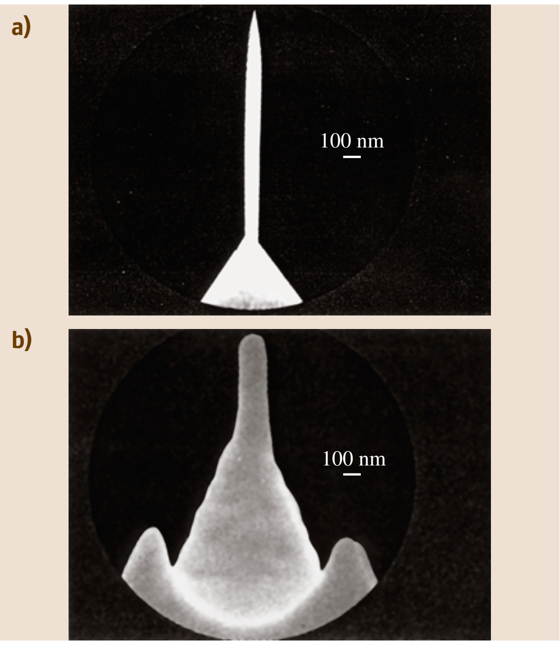

Fig. 11.17a,b Schematics of **(a)** HART Si_3N_4 probe, and **(b)** FIB milled Si_3N_4 probe

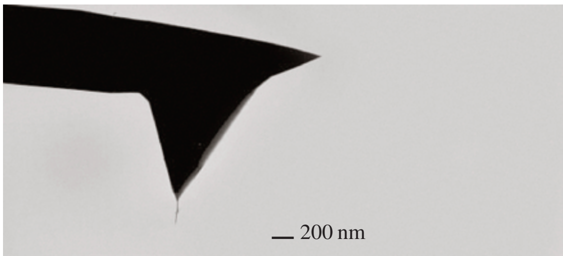

Fig. 11.18 SEM micrograph of a multi-walled carbon nanotube (MWNT) tip physically attached on the single-crystal silicon, square-pyramidal tip (Courtesy Piezomax Technologies, Inc.)

stant of the beam ranges from about 5 to 600 N/m for a 20 µm thick beam. The tips are produced by R-DEC Co., Tsukuba, Japan.

For imaging within trenches by an AFM, high aspect ratio tips are used. Examples of the two probes are shown in Fig. 11.17. The high-aspect ratio tip (Hart) probes are produced by starting with a conventional Si_3N_4 pyramidal probe. Through a combination of focused ion beam (FIB) and high resolution scanning electron microscopy (SEM) techniques, a thin filament is grown at the apex of the pyramid. The probe filament is approximately 1 µm long and 0.1 µm in diameter. It tapers to an extremely sharp point (radius better than the resolution of most SEMs). The long thin shape and sharp radius make it ideal for imaging within "vias" of microstructures and trenches (> 0.25 µm). Because of flexing of the probe, it is unsuitable for imaging structures at the atomic level since the flexing of the probe can create image artifacts. For atomic-scale imaging, a FIB-milled probe is used which is relatively stiff yet allows for closely spaced topography. These probes start out as conventional Si_3N_4 pyramidal probes but the pyramid is FIB milled until a small cone shape is formed which has a high aspect ratio with 0.2–0.3 µm in length. The milled probes allow nanostructure resolution without sacrificing rigidity. These types of probes are manufactured by various manufacturers including Materials Analytical Services, Raleigh, North Carolina.

Carbon nanotube tips having small diameter and high aspect ratio are used for high resolution imaging of surfaces and of deep trenches, in the tapping mode or noncontact mode. Single-walled carbon nanotubes (SWNT) are microscopic graphitic cylinders that are 0.7 to 3 nm in diameter and up to many microns in length. Larger structures called multi-walled carbon nanotubes (MWNT) consist of nested, concentrically arranged SWNT and have diameters ranging from 3 to 50 nm. MWNT carbon nanotube AFM tips are produced by manual assembly [11.141], chemical vapor deposition (CVD) synthesis, and hybrid fabrication process [11.142]. Figure 11.18 shows TEM micrograph of a carbon nanotube tip, ProbeMax^{TM}, commercially produced by mechanical assembly by Piezomax Technologies, Inc., Middleton, Wisconsin. For production of these tips, MWNT nanotubes are produced by carbon arc. They are physically attached on the single-crystal silicon, square-pyramidal tips in the SEM using a manipulator and the SEM stage to control the nanotubes and the tip independently. Once the nanotube is attached to the tip, it is usually too long to image with. It is shortened by using an AFM and applying voltage between the tip and the sample. Nanotube tips are also commercially produced by CVD synthesis by NanoDevices, Santa Barbara, California.

11.2.4 Friction Measurement Methods

Based on the work by *Ruan* and *Bhushan* [11.136], the two methods for friction measurements are now described in more detail. (Also see [11.8].) A scanning angle is defined as the angle relative to the y-axis in Fig. 11.19a. This is also the long axis of the cantilever.

A zero degree scanning angle corresponds to the sample scanning in the y direction, and a 90 degree scanning angle corresponds to the sample scanning perpendicular to this axis in the xy plane (along x axis). If the scanning direction is in both the y and $-y$ directions, we call this "parallel scan". Similarly, a "perpendicular scan" means the scanning direction is in the x and $-x$ directions. The sample traveling direction for each of these two methods is illustrated in Fig. 11.19b.

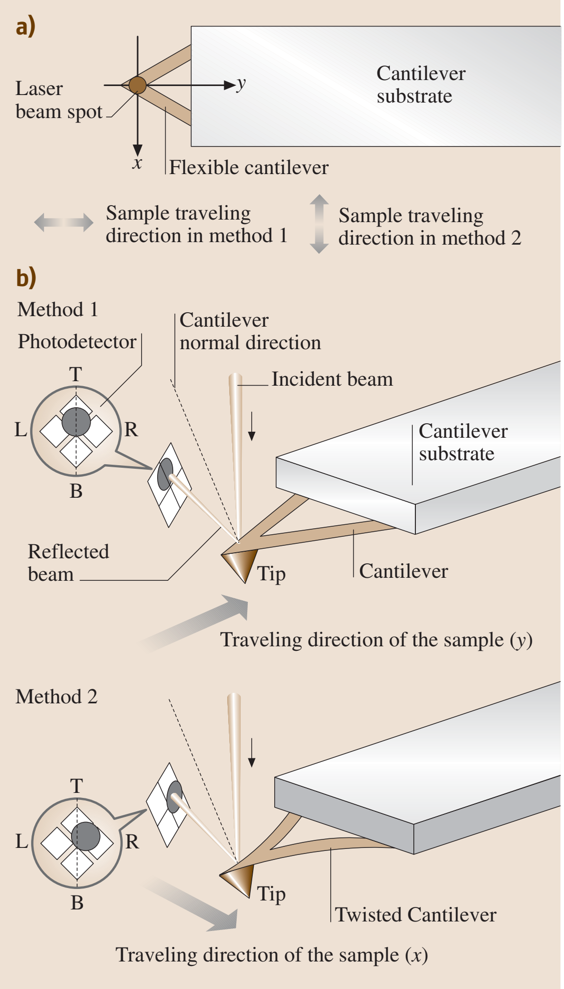
a)
Laser beam spot
y
x
Cantilever substrate
Flexible cantilever
Sample traveling direction in method 1
Sample traveling direction in method 2
b)
Method 1
Photodetector
Cantilever normal direction
Incident beam
T
L
R
B
Cantilever substrate
Reflected beam
Tip
Cantilever
Traveling direction of the sample (y)
Method 2
T
L
R
B
Tip
Twisted Cantilever
Traveling direction of the sample (x)

Fig. 11.19 **(a)** Schematic defining the x- and y-directions relative to the cantilever, and showing the sample traveling direction in two different measurement methods discussed in the text, **(b)** schematic of deformation of the tip and cantilever shown as a result of sliding in the x- and y-directions. A twist is introduced to the cantilever if the scanning is in the x-direction (**(b)**, *lower part*) [11.136]

In method 1 (using "height" mode with parallel scans) in addition to topographic imaging, it is also possible to measure friction force when the scanning direction of the sample is parallel to the y direction (parallel scan). If there were no friction force between the tip and the moving sample, the topographic feature would be the only factor which cause the cantilever to be deflected vertically. However, friction force does exist on all contact surfaces where one object is moving relative to another. The friction force between the sample and the tip will also cause a cantilever deflection. We assume that the normal force between the sample and the tip is W_0 when the sample is stationary (W_0 is typically in the range of 10 nN to 200 nN), and the friction force between the sample and the tip is W_f as the sample scans against the tip. The direction of friction force (W_f) is reversed as the scanning direction of the sample is reversed from positive (y) to negative ($-y$) directions ($\overrightarrow{W}_{f(y)} = -\overrightarrow{W}_{f(-y)}$).

When the vertical cantilever deflection is set at a constant level, it is the total force (normal force and friction force) applied to the cantilever that keeps the cantilever deflection at this level. Since the friction force is in opposite directions as the traveling direction of the sample is reversed, the normal force will have to be adjusted accordingly when the sample reverses its traveling direction, so that the total deflection of the cantilever will remain the same. We can calculate the difference of the normal force between the two traveling directions for a given friction force W_f. First, by means of a constant deflection, the total moment applied to the cantilever is constant. If we take the reference point to be the point where the cantilever joins the cantilever holder (substrate), point P in Fig. 11.20, we have the following relationship:

$$\begin{aligned}(W_0 - \Delta W_1)L + W_f \ell \\ = (W_0 + \Delta W_2)L - W_f \ell \end{aligned} \tag{11.1}$$

or

$$(\Delta W_1 + \Delta W_2)L = 2 W_f \ell \,. \tag{11.2}$$

Thus

$$W_f = (\Delta W_1 + \Delta W_2)L/(2\ell) \,, \tag{11.3}$$

where ΔW_1 and ΔW_2 are the absolute value of the changes of normal force when the sample is traveling in $-y$ and y directions, respectively, as shown in Fig. 11.20; L is the length of the cantilever; ℓ is vertical distance between the end of the tip and point P. The coefficient of

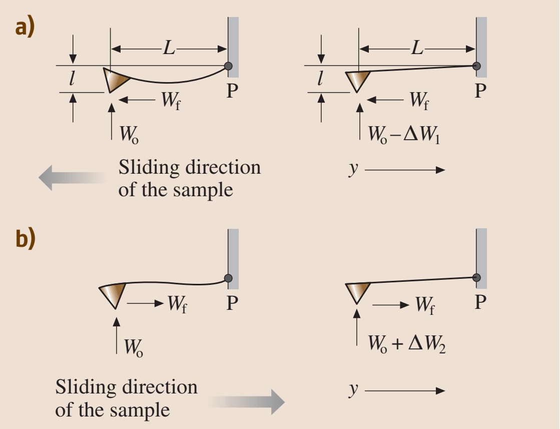

Fig. 11.20 **(a)** Schematic showing an additional bending of the cantilever – due to friction force when the sample is scanned in the y- or $-y$-direction (*left*). **(b)** This effect will be canceled by adjusting the piezo height by a feedback circuit (*right*) [11.136]

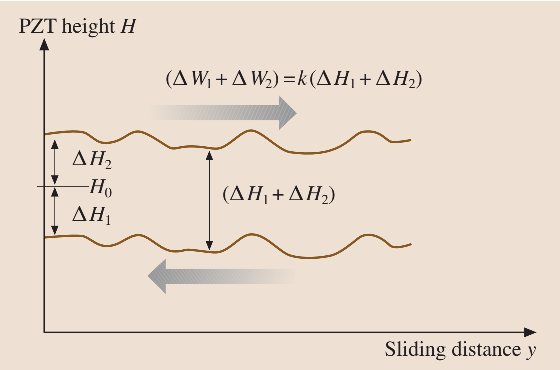
PZT height H
$(\Delta W_1+\Delta W_2)=k(\Delta H_1+\Delta H_2)$
ΔH_2
H_0
ΔH_1
$(\Delta H_1+\Delta H_2)$
Sliding distance y

Fig. 11.21 Schematic illustration of the height difference of the piezoelectric tube scanner as the sample is scanned in y and $-y$ directions

friction (μ) between the tip and the sample is then given as

$$\mu=\frac{W_f}{W_0}=\left(\frac{(\Delta W_1+\Delta W_2)}{W_0}\right)\left(\frac{L}{2\ell}\right). \tag{11.4}$$

In all circumstances, there are adhesive and interatomic attractive forces between the cantilever tip and the sample. The adhesive force can be due to water from the capillary condensation and other contaminants present at the surface which form meniscus bridges [11.4, 143, 144] and the interatomic attractive force includes van der Waals attraction [11.18]. If these forces (and indentation effect as well, which is usually small for rigid samples) can be neglected, the normal force W_0 is then equal to the initial cantilever deflection H_0 multiplied by the spring constant of the cantilever. $(\Delta W_1+\Delta W_2)$ can be measured by multiplying the same spring constant by the height difference of the piezo tube between the two traveling directions (y and $-y$ directions) of the sample. This height difference is denoted as $(\Delta H_1+\Delta H_2)$, shown schematically in Fig. 11.21. Thus, (11.4) can be rewritten as

$$\mu=\frac{W_f}{W_0}=\left(\frac{(\Delta H_1+\Delta H_2)}{H_0}\right)\left(\frac{L}{2\ell}\right). \tag{11.5}$$

Since the piezo tube vertical position is affected by the surface topographic profile of the sample in addition to the friction force being applied at the tip, this difference has to be taken point by point at the same location on the sample surface as shown in Fig. 11.21. Subtraction of point by point measurements may introduce errors, particularly for rough samples. We will come back to this point later. In addition, precise measurement of L and ℓ (which should include the cantilever angle) are also required.

If the adhesive forces between the tip and the sample are large enough that it can not be neglected, one should include it in the calculation. However, there could be a large uncertainty in determining this force, thus an uncertainty in using (11.5). An alternative approach is to make the measurements at different normal loads and to use $\Delta(H_0)$ and $\Delta(\Delta H_1+\Delta H_2)$ from the measurements in (11.5). Another comment on (11.5) is that, since only the ratio between $(\Delta H_1+\Delta H_2)$ and H_0 comes into this equation, the piezo tube vertical position H_0 and its position difference $(\Delta H_1+\Delta H_2)$ can be in the units of volts as long as the vertical traveling distance of the piezo tube and the voltage applied to it has a linear relationship. However, if there is a large nonlinearity between the piezo tube traveling distance and the applied voltage, this nonlinearity must be included in the calculation.

It should also be pointed out that (11.4) and (11.5) are derived under the assumption that the friction force W_f is the same for the two scanning directions of the sample. This is an approximation since the normal force is slightly different for the two scans and there may also be a directionality effect in friction. However, this difference is much smaller than W_0 itself. We can ignore the second order correction.

Method 2 ("aux" mode with perpendicular scan) to measure friction was suggested by *Meyer* and *Amer* [11.8]. The sample is scanned perpendicular to the long axis of the cantilever beam (i. e., to scan along the x or $-x$ direction in Fig. 11.19a) and the output of the horizontal two quadrants of the photodiode-detector is measured. In this arrangement, as the sample moves under the tip, the friction force will cause the cantilever to twist. Therefore the light intensity between the left and right (L and R in Fig. 11.19b, right) detectors will be different. The differential signal between the left and right detectors is denoted as FFM signal $[(\mathrm{L}-\mathrm{R})/(\mathrm{L}+\mathrm{R})]$. This signal can be related to the degree of twisting, hence to the magnitude of friction force. Again, because of a possible error in determining normal force due to the presence of an adhesive force at the tip–sample interface, the slope of the friction data (FFM signal vs. normal load) needs to be taken for an accurate value of the coefficient of friction.

While friction force contributes to the FFM signal, friction force may not be the only contributing factor in commercial FFM instruments (for example, NanoScope IV). One can notice this fact by simply engaging the cantilever tip with the sample. Before engaging, the left and right detectors can be balanced by adjusting the position of the detectors so that the intensity difference between these two detectors is zero (FFM signal is zero). Once the tip is engaged with the sample, this signal is no longer zero even if the sample is not moving in the xy plane with no friction force applied. This would be a detrimental effect. It has to be understood and eliminated from the data acquisition before any quantitative measurement of friction force becomes possible.

One of the fundamental reasons for this observation is the following. The detectors may not have been properly aligned with respect to the laser beam. To be precise, the vertical axis of the detector assembly (the line joining T-B in Fig. 11.22) is not in the plane defined by the incident laser beam and the beam reflected from an untwisted cantilever (we call this plane the "beam plane"). When the cantilever vertical deflection changes due to a change of applied normal force (without having the sample scanned in the xy plane), the laser beam will be reflected up and down and form a projected trajectory on the detector. (Note that this trajectory is in the defined beam plane.) If this trajectory is not coincident with the vertical axis of the detector, the laser beam will not evenly bisect the left and right quadrants of the detectors, even under the condition of no torsional motion of the cantilever, see Fig. 11.22. Thus when the laser beam is reflected up and down due a change of the normal force, the intensity difference between the left and right detectors will also change. In other words, the FFM signal will change as the normal force applied to the tip is changed, even if the tip is not experiencing any friction force. This (FFM) signal is unrelated to friction force or to the actual twisting of the cantilever. We will call this part of FFM signal "$\mathrm{FFM_F}$", and the part which is truly related to friction force "$\mathrm{FFM_T}$".

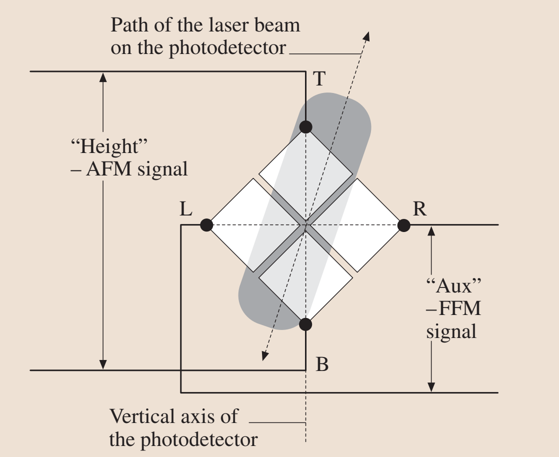

Fig. 11.22 The trajectory of the laser beam on the photodetectors in as the cantilever is vertically deflected (with no torsional motion) for a misaligned photodetector with respect to the laser beam. For a change of normal force (vertical deflection of the cantilever), the laser beam is projected at a different position on the detector. Due to a misalignment, the projected trajectory of the laser beam on the detector is not parallel with the detector vertical axis (the line joint T-B) [11.136]

The $\mathrm{FFM_F}$ signal can be eliminated. One way of doing this is as follows. First the sample is scanned in both x and $-x$ directions and the FFM signal for scans in each direction is recorded. Since friction force reverses its directions when the scanning direction is reversed from x to $-x$ direction, the $\mathrm{FFM_T}$ signal will have opposite signs as the scanning direction of the sample is reversed ($\mathrm{FFM_T}(x) = -\mathrm{FFM_T}(-x)$). Hence the $\mathrm{FFM_T}$ signal will be canceled out if we take the sum of the FFM signals for the two scans. The average value of the two scans will be related to $\mathrm{FFM_F}$ due to the misalignment,

$$\mathrm{FFM}(x)+\mathrm{FFM}(-x)=2\mathrm{FFM_F}\,. \tag{11.6}$$

This value can therefore be subtracted from the original FFM signals of each of these two scans to obtain the true FFM signal ($\mathrm{FFM_T}$). Or, alternately, by taking the difference of the two FFM signals, one directly gets the $\mathrm{FFM_T}$ value

$$\mathrm{FFM}(x)-\mathrm{FFM}(-x)=\mathrm{FFM_T}(x)-\mathrm{FFM_T}(-x)$$
$$=2\mathrm{FFM_T}(x)\,. \quad (11.7)$$

Ruan and *Bhushan* [11.136] have shown that the error signal ($\mathrm{FFM_F}$) can be very large compared to the friction signal $\mathrm{FFM_T}$, thus correction is required.

Now we compare the two methods. The method of using "height" mode and parallel scan (method 1) is very simple to use. Technically, this method can provide 3-D friction profiles and the corresponding topographic profiles. However, there are some problems with this method. Under most circumstances, the piezo scanner displays a hysteresis when the traveling direction of the sample is reversed. Therefore the measured surface topographic profiles will be shifted relative to each other along the y-axis for the two opposite (y and $-y$) scans. This would make it difficult to measure the local height difference of the piezo tube for the two scans. However, the average height difference between the two scans and hence the average friction can still be measured. The measurement of average friction can serve as an internal means of friction force calibration. Method 2 is a more desirable approach. The subtraction of $\mathrm{FFM_F}$ signal from FFM for the two scans does not introduce error to local friction force data. An ideal approach in using this method would be to add the average value of the two profiles in order to get the error component ($\mathrm{FFM_F}$) and then subtract this component from either profiles to get true friction profiles in either directions. By making measurements at various loads, we can get the average value of the coefficient of friction which then can be used to convert the friction profile to the coefficient of friction profile. Thus any directionality and local variations in friction can be easily measured. In this method, since topography data are not affected by friction, accurate topography data can be measured simultaneously with friction data and a better localized relationship between the two can be established.

11.2.5 Normal Force and Friction Force Calibrations of Cantilever Beams

Based on *Ruan* and *Bhushan* [11.136], we now discuss normal force and friction force calibrations. In order to calculate the absolute value of normal and friction forces in Newtons using the measured AFM and $\mathrm{FFM_T}$ voltage signals, it is necessary to first have an accurate value of the spring constant of the cantilever (k_c). The spring constant can be calculated using the geometry and the physical properties of the cantilever material [11.8,94,137]. However, the properties of the PECVD $\mathrm{Si_3N_4}$ (used in fabricating cantilevers) could be different from those of the bulk material. For example, by using an ultrasonic measurement, we found the Young's modulus of the cantilever beam to be about 238 ± 18 GPa which is less than that of bulk $\mathrm{Si_3N_4}$ (310 GPa). Furthermore the thickness of the beam is nonuniform and difficult to measure precisely. Since the stiffness of a beam goes as the cube of thickness, minor errors in precise measurements of thickness can introduce substantial stiffness errors. Thus one should experimentally measure the spring constant of the cantilever. *Cleveland* et al. [11.145] measured the normal spring constant by measuring resonant frequencies of the beams.

For normal spring constant measurement, *Ruan* and *Bhushan* [11.136] used a stainless steel spring sheet of known stiffness (width = 1.35 mm, thickness = 15 μm, free hanging length = 5.2 mm). One end of the spring was attached to the sample holder and the other end was made to contact with the cantilever tip during the measurement, see Fig. 11.23. They measured the piezo traveling distance for a given cantilever deflection. For a rigid sample (such as diamond), the piezo traveling distance Z_t (measured from the point where the tip touches the sample) should equal the cantilever deflection. To keep the cantilever deflection at the same level using a flexible spring sheet, the new piezo traveling distance $Z_{t'}$ would be different from Z_t. The difference between $Z_{t'}$ and Z_t corresponds to the deflection of the spring

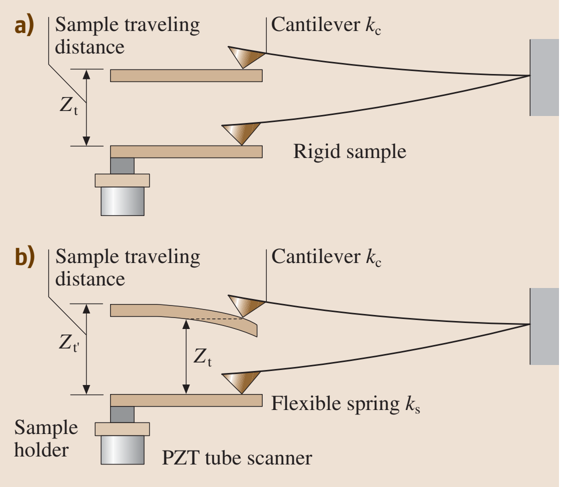

Fig. 11.23a,b Illustration showing the deflection of cantilever as it is pushed by **(a)** a rigid sample or by **(b)** a flexible spring sheet [11.136]

sheet. If the spring constant of the spring sheet is k_s, the spring constant of the cantilever k_c can be calculated by

$$(Z_{\mathrm{t}'} - Z_\mathrm{t})k_\mathrm{s} = Z_\mathrm{t} k_\mathrm{c}$$

or

$$k_\mathrm{c} = k_\mathrm{s}(Z_{\mathrm{t}'} - Z_\mathrm{t})/Z_\mathrm{t} \,. \tag{11.8}$$

The spring constant of the spring sheet (k_s) used in this study is calculated to be 1.54 N/m. For a wide-legged cantilever used in our study (length = 115 μm, base width = 122 μm, leg width = 21 μm, and thickness = 0.6 μm), k_c was measured to be 0.40 N/m instead of 0.58 N/m reported by its manufacturer – Digital Instruments Inc. To relate photodiode detector output to the cantilever deflection in nm, they used the same rigid sample to push against the AFM tip. Since for a rigid sample the cantilever vertical deflection equals the sample traveling distance measured from the point where the tip touches the sample, the photodiode output as the tip is pushed by the sample can be converted directly to cantilever deflection. For these measurements, they found the conversion factor to be 20 nm/V.

The normal force applied to the tip can be calculated by multiplying the cantilever vertical deflection by the cantilever spring constant for samples which have very small adhesive force with the tip. If the adhesive force between the sample and the tip is large, it should be included in the normal force calculation. This is particularly important in atomic-scale force measurement because in this region, the typical normal force that is measured is in the range of a few hundreds of nN to a few mN. The adhesive force could be comparable to the applied force.

The conversion of friction signal (from $\mathrm{FFM_T}$ to friction force) is not as straightforward. For example, one can calculate the degree of twisting for a given friction force using the geometry and the physical properties of the cantilever [11.53, 144]. One would need the information on the detectors such as the quantum efficiency of the detector, the laser power, the instrument's gain, etc. in order to be able convert the signal into the degree of twisting. Generally speaking, this procedure can not be accomplished without having some detailed information about the instrument. This information is not usually provided by the manufactures. Even if this information is readily available, error may still occur in using this approach because there will always be variations as a result of the instrumental set up. For example, it has been noticed that the measured $\mathrm{FFM_T}$ signal could be different for the same sample when different AFM microscopes of the same kind are used. The essence is that, one can not calibrate the instrument experimentally using this calculation. *O'Shea* et al. [11.144] did perform a calibration procedure in which the torsional signal was measured as the sample is displaced a known distance laterally while ensuring that the tip does not slide over the surface. However, it is difficult to verify that tip sliding does not occur.

Apparently, a new method of calibration is required. There is a more direct and simpler way of doing this. The first method described (method 1) to measure friction can directly provide an absolute value of the coefficient of friction. It can therefore be used just as an internal means of calibration for the data obtained using method 2. Or for a polished sample which introduces least error in friction measurement using method 1, method 1 can be used to obtain calibration for friction force for method 2. Then this calibration can be used for measurement on all samples using method 2. In method 1, the length of the cantilever required can be measured using an optical microscope; the length of the tip can be measured using a scanning electron microscope. The relative angle between the cantilever and the horizontal sample surface can be measured directly. Thus the coefficient of friction can be measured with few unknown parameters. The friction force can then be calculated by multiplying the coefficient of friction by the normal load. The $\mathrm{FFM_T}$ signal obtained using method 2 can then be converted into friction force. For their instrument, they found the conversion to be 8.6 nN/V.

11.3 AFM Instrumentation and Analyses

The performance of AFMs and the quality of AFM images greatly depend on the instrument available and the probes (cantilever and tips) in use. This section describes the mechanics of cantilevers, instrumentation and analysis of force detection systems for cantilever deflections, and scanning and control systems.

11.3.1 The Mechanics of Cantilevers

Stiffness and Resonances of Lumped Mass Systems

Any one of the building blocks of an AFM, be it the body of the microscope itself or the force measuring cantilevers, are mechanical resonators. These resonances can be excited either by the surroundings or by the rapid movement of the tip or the sample. To avoid problems due to building or air induced oscillations, it is of paramount importance to optimize the design of AFMs for high resonant frequencies. This usually means decreasing the size of the microscope [11.146]. By using cube-like or sphere-like structures for the microscope, one can considerably increase the lowest eigenfrequency. The fundamental natural frequency, ω_0, of any spring is given by

$$\omega_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m_{\text{eff}}}}\,, \tag{11.9}$$

where k is the spring constant (stiffness) in the normal direction and m_{eff} is the effective mass. The spring constant k of a cantilever beam with uniform cross section (Fig. 11.24) is given by [11.147]

$$k = \frac{3EI}{L^3}\,, \tag{11.10}$$

where E is the Young's modulus of the material, L is the length of the beam and I is the moment of inertia of the cross section. For a rectangular cross section with a width b (perpendicular to the deflection) and a height h one obtains an expression for I

$$I = \frac{bh^3}{12}\,. \tag{11.11}$$

Combining (11.9), (11.10) and (11.11) we get an expression for ω_0

$$\omega_0 = \sqrt{\frac{Ebh^3}{4L^3 m_{\text{eff}}}}\,. \tag{11.12}$$

The effective mass can be calculated using Raleigh's method. The general formula using Raleigh's method for the kinetic energy T of a bar is

$$T = \frac{1}{2}\int_0^L \frac{m}{L}\left(\frac{\partial z(x)}{\partial t}\right)^2 \mathrm{d}x\,. \tag{11.13}$$

For the case of a uniform beam with a constant cross section and length L one obtains for the deflection $z(x) = z_{\max}\left[1-(3x/2L)+(x^3/2L^3)\right]$. Inserting $z_{\max}$ into (11.13) and solving the integral gives

$$T = \frac{1}{2}\int_0^L \frac{m}{L}\left[\frac{\partial z_{\max}(x)}{\partial t}\left(1-\frac{3x}{2L}\right)+\left(\frac{x^3}{L^3}\right)\right]^2 \mathrm{d}x$$
$$= \frac{1}{2} m_{\text{eff}}(z_{\max} t)^2\,,$$

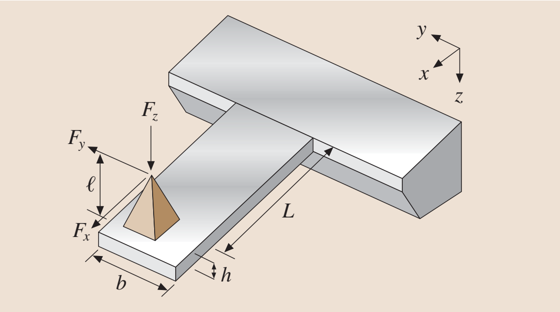

Fig. 11.24 A typical AFM cantilever with length L, width b, and height h. The height of the tip is ℓ. The material is characterized by Young's modulus E, the shear modulus G and a mass density ρ. Normal (F_z), axial (F_x), and lateral (F_y) forces exist at the end of the tip

which gives

$$m_{\text{eff}} = \frac{9}{20}m\,. \tag{11.14}$$

Substituting (11.14) into (11.12) and noting that $m = \rho Lbh$, where ρ is the mass density, one obtains the following expression

$$\omega_0 = \left(\frac{\sqrt{5}}{3}\sqrt{\frac{\mathrm{E}}{\rho}}\right)\frac{\mathrm{h}}{L^2}\,. \tag{11.15}$$

It is evident from (11.15), that one way to increase the natural frequency is to choose a material with a high ratio E/ρ; see Table 11.2 for typical values of $\sqrt{E/\rho}$ of various commonly used materials. Another way to increase the lowest eigenfrequency is also evident in (11.15). By optimizing the ratio h/L^2, one can increase the resonant frequency. However it does not help to make the length of the structure smaller than the width or height. Their roles will just be interchanged. Hence the optimum structure is a cube. This leads to the design rule, that long, thin structures like sheet metal should be avoided. For a given resonant frequency, the quality factor Q should be as low as possible. This means that an inelastic medium such as rubber should be in contact with the structure to convert kinetic energy into heat.

Stiffness and Resonances of Cantilevers

Cantilevers are mechanical devices specially shaped to measure tiny forces. The analysis given in the previous section is applicable. However, to understand better the intricacies of force detection systems we will discuss the example of a cantilever beam with uniform cross section, Fig. 11.24. The bending of a beam due to a normal load on the beam is governed by the Euler equation [11.147]

$$M = EI\,(x)\,\frac{\mathrm{d}^2 z}{\mathrm{d}x^2}\,, \tag{11.16}$$

where M is the bending moment acting on the beam cross section. $I(x)$ the moment of inertia of the cross section with respect to the neutral axis defined by

$$I(x) = \int\limits_z \int\limits_y z^2\,\mathrm{d}y\,\mathrm{d}z\,. \tag{11.17}$$

For a normal force F_z acting at the tip,

$$M(x) = (L - x)\,F_z \tag{11.18}$$

since the moment must vanish at the endpoint of the cantilever. Integrating (11.16) for a normal force F_z acting at the tip and observing that EI is a constant for beams with a uniform cross section, one gets

$$z(x) = \frac{L^3}{6EI}\left(\frac{x}{L}\right)^2\left(3 - \frac{x}{L}\right)F_z\,. \tag{11.19}$$

The slope of the of the beam is

$$z'\,(x) = \frac{Lx}{2EI}\left(2 - \frac{x}{L}\right)F_z\,. \tag{11.20}$$

From (11.19) and (11.20), at the end of the cantilever, i. e. for $x = L$, for a rectangular beam, and by using an expression for I in (11.11), one gets,

$$z(L) = \frac{4}{Eb}\left(\frac{L}{h}\right)^3 F_z\,, \tag{11.21}$$

$$z'(L) = \frac{3}{2}\left(\frac{z}{L}\right)\,. \tag{11.22}$$

Now the stiffness in the normal (z) direction, k_z, is

$$k_z = \frac{F_z}{z(L)} = \frac{Eb}{4}\left(\frac{h}{L}\right)^3\,. \tag{11.23}$$

and a change in angular orientation of the end of cantilever beam is

$$\Delta\alpha = \frac{3}{2}\frac{z}{L} = \frac{6}{Ebh}\left(\frac{L}{h}\right)^2 F_z\,. \tag{11.24}$$

Now we ask what will, to first order, happen if we apply a lateral force F_y to the end of the tip (Fig. 11.24). The cantilever will bend sideways and it will twist. The stiffness in the lateral (y) direction, k_y, can be calculated with (11.23) by exchanging b and h

$$k_y = \frac{Eh}{4}\left(\frac{b}{L}\right)^3\,. \tag{11.25}$$

Therefore the bending stiffness in lateral direction is larger than the stiffness for bending in the normal direction by $(b/h)^2$. The twisting or torsion on the other hand is more complicated to handle. For a wide, thin cantilever ($b \gg h$) we obtain torsional stiffness along y axis, k_{yT}

$$k_{yT} = \frac{Gbh^3}{3L\ell^2}\,, \tag{11.26}$$

where G is the modulus of rigidity ($= E/2(1+\nu)$, where ν is the Poisson's ratio). The ratio of the torsional stiffness to the lateral bending stiffness is

$$\frac{k_{yT}}{k_y} = \frac{1}{2}\left(\frac{\ell b}{hL}\right)^2\,, \tag{11.27}$$

where we assume $\nu = 0.333$. We see that thin, wide cantilevers with long tips favor torsion while cantilevers with square cross sections and short tips favor bending. Finally we calculate the ratio between the torsional stiffness and the normal bending stiffness,

$$\frac{k_{yT}}{k_z} = 2\left(\frac{L}{\ell}\right)^2\,. \tag{11.28}$$

Equations (11.26) to (11.28) hold in the case where the cantilever tip is exactly in the middle axis of the cantilever. Triangular cantilevers and cantilevers with tips not on the middle axis can be dealt with by finite element methods.

The third possible deflection mode is the one from the force on the end of the tip along the cantilever axis, F_x (Fig. 11.24). The bending moment at the free end of the cantilever is equal to the $F_x\ell$. This leads to the following modification of (11.18) for the case of forces F_z and F_x

$$M\,(x) = (L - x)\,F_z + F_x\ell\,. \tag{11.29}$$

Integration of (11.16) now leads to

$$z(x) = \frac{1}{2EI}\left[Lx^2\left(1 - \frac{x}{3L}\right)F_z + \ell x^2 F_x\right] \tag{11.30}$$

and

$$z'\,(x) = \frac{1}{EI}\left[\frac{Lx}{2}\left(2 - \frac{x}{L}\right)F_z + \ell x F_x\right]\,. \tag{11.31}$$

Evaluating (11.30) and (11.31) at the end of the cantilever, we get the deflection and the tilt

$$z\,(L) = \frac{L^2}{EI}\left(\frac{L}{3}F_z - \frac{\ell}{2}F_x\right)\,,$$

$$z'\,(L) = \frac{L}{EI}\left(\frac{L}{2}F_z + \ell F_x\right)\,. \tag{11.32}$$

From these equations, one gets

$$F_z = \frac{12EI}{L^3}\left[z(L) - \frac{Lz'(L)}{2}\right],$$
$$F_x = \frac{2EI}{\ell L^2}\left[2Lz'(L) - 3z(L)\right]. \tag{11.33}$$

A second class of interesting properties of cantilevers is their resonance behavior. For cantilever beams one can calculate the resonant frequencies [11.147, 148]

$$\omega_n^{\text{free}} = \frac{\lambda_n^2}{2\sqrt{3}}\frac{h}{L^2}\sqrt{\frac{E}{\rho}} \tag{11.34}$$

with $\lambda_0 = (0.596864\ldots)\pi$, $\lambda_1 = (1.494175\ldots)\pi$, $\lambda_n \to (n+1/2)\pi$. The subscript n represents the order of the frequency, e.g., fundamental, second mode, and the nth mode.

A similar equation to (11.34) holds for cantilevers in rigid contact with the surface. Since there is an additional restriction on the movement of the cantilever, namely the location of its end point, the resonant frequency increases. Only the λ_n's terms change to [11.148]

$$\lambda_0' = (1.2498763\ldots)\pi, \quad \lambda_1' = (2.2499997\ldots)\pi,$$
$$\lambda_n' \to (n+1/4)\pi. \tag{11.35}$$

The ratio of the fundamental resonant frequency in contact to the fundamental resonant frequency not in contact is 4.3851.

For the torsional mode we can calculate the resonant frequencies as

$$\omega_0^{\text{tors}} = 2\pi\frac{h}{Lb}\sqrt{\frac{G}{\rho}}. \tag{11.36}$$

For cantilevers in rigid contact with the surface, we obtain the expression for the fundamental resonant frequency [11.148]

$$\omega_0^{\text{tors, contact}} = \frac{\omega_0^{\text{tors}}}{\sqrt{1+3(2L/b)^2}}. \tag{11.37}$$

The amplitude of the thermally induced vibration can be calculated from the resonant frequency using

$$\Delta z_{\text{therm}} = \sqrt{\frac{k_{\text{B}}T}{k}}, \tag{11.38}$$

where k_{B} is Boltzmann's constant and T is the absolute temperature. Since AFM cantilevers are resonant structures, sometimes with rather high Q, the thermal noise is not evenly distributed as (11.38) suggests. The spectral noise density below the peak of the response curve is [11.148]

$$z_0 = \sqrt{\frac{4k_{\text{B}}T}{k\omega_0 Q}} \qquad (\text{in m}/\sqrt{\text{Hz}}), \tag{11.39}$$

where Q is the quality factor of the cantilever, described earlier.

11.3.2 Instrumentation and Analyses of Detection Systems for Cantilever Deflections

A summary of selected detection systems was provided in Fig. 11.8. Here we discuss in detail pros and cons of various systems.

Optical Interferometer Detection Systems

Soon after the first papers on the AFM [11.2] appeared, which used a tunneling sensor, an instrument based on an interferometer was published [11.149]. The sensitivity of the interferometer depends on the wavelength of the light employed in the apparatus. Figure 11.25 shows the principle of such an interferometeric design. The light incident from the left is focused by a lens on the cantilever. The reflected light is collimated by the same lens and interferes with the light reflected at the flat. To separate the reflected light from the incident light a $\lambda/4$ plate converts the linear polarized incident light to circular polarization. The reflected light is made linear polarized again by the $\lambda/4$-plate, but with a polarization orthogonal to that of the incident light. The polarizing

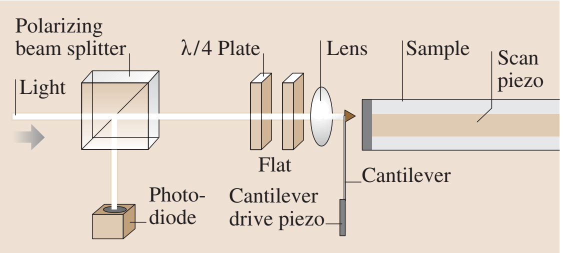

Fig. 11.25 Principle of an interferometric AFM. The light of the laser light source is polarized by the polarizing beam splitter and focused on the back of the cantilever. The light passes twice through a quarter wave plate and is hence orthogonally polarized to the incident light. The second arm of the interferometer is formed by the flat. The interference pattern is modulated by the oscillating cantilever

beam splitter then deflects the reflected light to the photo diode.

Homodyne Interferometer. To improve the signal to noise ratio of the interferometer the cantilever is driven by a piezo near its resonant frequency. The amplitude Δz of the cantilever as a function of driving frequency Ω is

$$\Delta z\,(\Omega) = \Delta z_0 \frac{\Omega_0^2}{\sqrt{\left(\Omega^2-\Omega_0^2\right)^2+\frac{\Omega^2\Omega_0^2}{Q^2}}}\,, \tag{11.40}$$

where Δz_0 is the constant drive amplitude and Ω_0 the resonant frequency of the cantilever. The resonant frequency of the cantilever is given by the effective potential

$$\Omega_0 = \sqrt{\left(k+\frac{\partial^2 U}{\partial z^2}\right)\frac{1}{m_{\text{eff}}}}\,, \tag{11.41}$$

where U is the interaction potential between the tip and the sample. Equation (11.41) shows that an attractive potential decreases Ω_0. The change in Ω_0 in turn results in a change of the Δz (see (11.40)). The movement of the cantilever changes the path difference in the interferometer. The light reflected from the cantilever with the amplitude $A_{\ell,0}$ and the reference light with the amplitude $A_{r,0}$ interfere on the detector. The detected intensity $I(t) = [A_\ell(t)+A_r(t)]^2$ consists of two constant terms and a fluctuating term

$$2A_\ell\,(t)\,A_r\,(t) = A_{\ell,0}A_{r,0}\sin\left[\omega t+\frac{4\pi\delta}{\lambda}+\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\right]\sin\,(\omega t)\;. \tag{11.42}$$

Here ω is the frequency of the light, λ is the wavelength of the light, δ is the path difference in the interferometer, and Δz is the instantaneous amplitude of the cantilever, given according to (11.40) and (11.41) as a function of Ω, k, and U. The time average of (11.42) then becomes

$$\begin{aligned}\langle 2A_\ell\,(t)\,A_r\,(t)\rangle_T &\propto \cos\left[\frac{4\pi\delta}{\lambda}+\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\right]\\ &\approx \cos\left(\frac{4\pi\delta}{\lambda}\right)-\sin\left[\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\right]\\ &\approx \cos\left(\frac{4\pi\delta}{\lambda}\right)-\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\;.\end{aligned} \tag{11.43}$$

Here all small quantities have been omitted and functions with small arguments have been linearized. The amplitude of Δz can be recovered with a lock-in technique. However, (11.43) shows that the measured amplitude is also a function of the path difference δ in the interferometer. Hence this path difference δ must be very stable. The best sensitivity is obtained when $\sin(4\delta/\lambda) \approx 0$.

Heterodyne Interferometer. This influence is not present in the heterodyne detection scheme shown in Fig. 11.26. Light incident from the left with a frequency ω is split in a reference path (upper path in Fig. 11.26) and a measurement path. Light in the measurement path is shifted in frequency to $\omega_1 = \omega + \Delta\omega$ and focused on the cantilever. The cantilever oscillates at the frequency Ω, as in the homodyne detection scheme. The reflected light $A_\ell(t)$ is collimated by the same lens and interferes on the photo diode with the reference light $A_r(t)$. The fluctuating term of the intensity is given by

$$\begin{aligned}2A_\ell\,(t)\,A_r\,(t) = {}& A_{\ell,0}A_{r,0}\sin\left[(\omega+\Delta\omega)\,t+\frac{4\pi\delta}{\lambda}\right.\\ &\left.+\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\right]\sin\,(\omega t)\;,\end{aligned} \tag{11.44}$$

where the variables are defined as in (11.42). Setting the path difference $\sin(4\pi\delta/\lambda) \approx 0$ and taking the time average, omitting small quantities and linearizing functions with small arguments we get

$$\begin{aligned}&\langle 2A_\ell\,(t)\,A_r\,(t)\rangle_T\\ &\propto \cos\left[\Delta\omega t+\frac{4\pi\delta}{\lambda}+\frac{4\pi\Delta z}{\lambda}\sin\,(\Omega t)\right]\end{aligned}$$

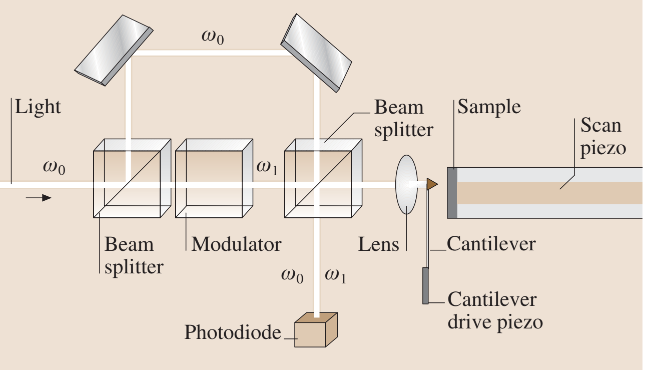

Fig. 11.26 Principle of a heterodyne interferometric AFM. Light with the frequency ω_0 is split into a reference path (upper path) and a measurement path. The light in the measurement path is frequency shifted to ω_1 by an acousto-optical modulator (or an electro-optical modulator) The light reflected from the oscillating cantilever interferes with the reference beam on the detector

$$
\begin{aligned}
&= \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\cos\left[\frac{4\pi\Delta z}{\lambda}\sin(\Omega t)\right]\\
&\quad - \sin\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\sin\left[\frac{4\pi\Delta z}{\lambda}\sin(\Omega t)\right]\\
&\approx \cos\left(\frac{4\pi\delta}{\lambda}\right) - \sin\left[\frac{4\pi\Delta z}{\lambda}\sin(\Omega t)\right]\\
&\approx \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\left[1 - \frac{8\pi^2\Delta z^2}{\lambda^2}\sin(\Omega t)\right]\\
&\quad - \frac{4\pi\Delta z}{\lambda}\sin\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\sin(\Omega t)\\
&= \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right) - \frac{8\pi^2\Delta z^2}{\lambda^2}\cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\\
&\quad \times \sin(\Omega t) - \frac{4\pi\Delta z}{\lambda}\sin\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\sin(\Omega t)\\
&= \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right) - \frac{4\pi^2\Delta z^2}{\lambda^2}\cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\\
&\quad + \frac{4\pi^2\Delta z^2}{\lambda^2}\cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\cos(2\Omega t)\\
&\quad - \frac{4\pi\Delta z}{\lambda}\sin\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\sin(\Omega t)\\
&= \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\left(1 - \frac{4\pi^2\Delta z^2}{\lambda^2}\right)\\
&\quad + \frac{2\pi^2\Delta z^2}{\lambda^2}\left\{\cos\left[(\Delta\omega + 2\Omega)\,t + \frac{4\pi\delta}{\lambda}\right]\right.\\
&\quad \left. + \cos\left[(\Delta\omega - 2\Omega)\,t + \frac{4\pi\delta}{\lambda}\right]\right\}\\
&\quad + \frac{2\pi\Delta z}{\lambda}\left\{\cos\left[(\Delta\omega + \Omega)\,t + \frac{4\pi\delta}{\lambda}\right]\right.\\
&\quad \left. + \cos\left[(\Delta\omega - \Omega)\,t + \frac{4\pi\delta}{\lambda}\right]\right\} .
\end{aligned}
\tag{11.45}
$$

Multiplying electronically the components oscillating at $\Delta\omega$ and $\Delta\omega + \Omega$ and rejecting any product except the one oscillating at Ω we obtain

$$
\begin{aligned}
A &= \frac{2\Delta z}{\lambda}\left(1 - \frac{4\pi^2\Delta z^2}{\lambda^2}\right)\cos\left[(\Delta\omega + 2\Omega)\,t + \frac{4\pi\delta}{\lambda}\right]\\
&\quad \times \cos\left(\Delta\omega t + \frac{4\pi\delta}{\lambda}\right)\\
&= \frac{\Delta z}{\lambda}\left(1 - \frac{4\pi^2\Delta z^2}{\lambda^2}\right)\left\{\cos\left[(2\Delta\omega + \Omega)\,t + \frac{8\pi\delta}{\lambda}\right]\right.\\
&\quad \left. + \cos(\Omega t)\right\}\\
&\approx \frac{\pi\Delta z}{\lambda}\cos(\Omega t) .
\end{aligned}
\tag{11.46}
$$

Unlike in the homodyne detection scheme the recovered signal is independent from the path difference δ of the interferometer. Furthermore a lock-in amplifier with the reference set $\sin(\Delta\omega t)$ can measure the path difference δ independent of the cantilever oscillation. If necessary, a feedback circuit can keep $\delta = 0$.

Fiber-optical Interferometer. The fiber-optical interferometer [11.129] is one of the simplest interferometers to build and use. Its principle is sketched in Fig. 11.27. The light of a laser is fed into an optical fiber. Laser diodes with integrated fiber pigtails are convenient light sources. The light is split in a fiber-optic beam splitter into two fibers. One fiber is terminated by index matching oil to avoid any reflections back into the fiber. The end of the other fiber is brought close to the cantilever in the AFM. The emerging light is partially reflected back into the fiber by the cantilever. Most of the light, however, is lost. This is not a big problem since only 4% of the light is reflected at the end of the fiber, at the glass-air interface. The two reflected light waves interfere with each other. The product is guided back into the fiber coupler and again split into two parts. One half is analyzed by the photodiode. The other half is fed back into the laser. Communications grade laser diodes are sufficiently resistant against feedback to be operated in this environment. They have, however, a bad coherence length, which in this case does not matter, since the optical path difference is in any case no larger than 5 μm. Again the end of the fiber has to be positioned on a piezo drive to set the distance between the fiber and the cantilever to $\lambda(n + 1/4)$.

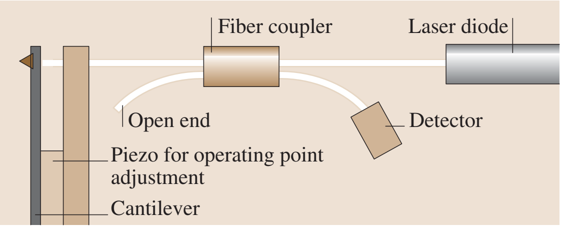

Fig. 11.27 A typical setup for a fiber optic interferometer readout

Nomarski-Interferometer. Another solution to minimize the optical path difference is to use the Nomarski interferometer [11.130]. Figure 11.28 shows a schematic of the microscope. The light of a laser is focused on the cantilever by lens. A birefringent crystal (for instance calcite) between the cantilever and the lens with its optical axis 45° off the polarization direction of the light

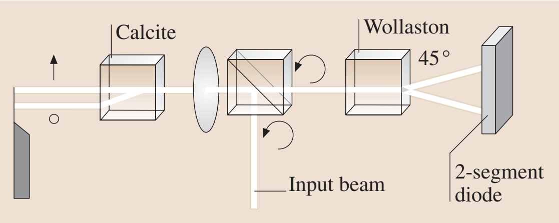

Fig. 11.28 Principle of Nomarski AFM. The circular polarized input beam is deflected to the left by a non-polarizing beam splitter. The light is focused onto a cantilever. The calcite crystal between the lens and the cantilever splits the circular polarized light into two spatially separated beams with orthogonal polarizations. The two light beams reflected from the lever are superimposed by the calcite crystal and collected by the lens. The resulting beam is again circular polarized. A Wollaston prism produces two interfering beams with $\pi/2$ phase shift between them. The minimal path difference accounts for the excellent stability of this microscope

splits the light beam into two paths, offset by a distance given by the length of the crystal. Birefringent crystals have varying indexes of refraction. In calcite, one crystal axis has a lower index than the other two. This means, that certain light rays will propagate at a different speed through the crystal than the others. By choosing a correct polarization, one can select the ordinary ray, the extraordinary ray or one can get any distribution of the intensity amongst those two rays. A detailed description of birefringence can be found in textbooks (e.g., [11.150]). A calcite crystal deflects the extraordinary ray at an angle of 6° within the crystal. By choosing a suitable length of the calcite crystal, any separation can be set.

The focus of one light ray is positioned near the free end of the cantilever while the other is placed close to the clamped end. Both arms of the interferometer pass through the same space, except for the distance between the calcite crystal and the lever. The closer the calcite crystal is placed to the lever, the less influence disturbances like air currents have.

Sarid [11.116] has given values for the sensitivity of the different interferometeric detection systems. Table 11.5 presents a summary of his results.

Optical Lever

The most common cantilever deflection detection system is the optical lever [11.53, 111]. This method, depicted in Fig. 11.29, employs the same technique as light beam deflection galvanometers. A fairly well collimated light beam is reflected off a mirror and projected to a receiving target. Any change in the angular position of the mirror will change the position, where the light ray hits the target. Galvanometers use optical path lengths of several meters and scales projected to the target wall as a read-out help.

For the AFM using the optical lever method a photodiode segmented into two (or four) closely spaced devices detects the orientation of the end of the cantilever. Initially, the light ray is set to hit the photodiodes in the middle of the two sub-diodes. Any deflection of

Table 11.5 Noise in Interferometers. F is the finesse of the cavity in the homodyne interferometer, P_i the incident power, P_d is the power on the detector, η is the sensitivity of the photodetector and RIN is the relative intensity noise of the laser. P_R and P_S are the power in the reference and sample beam in the heterodyne interferometer. P is the power in the Nomarski interferometer, $\delta\theta$ is the phase difference between the reference and the probe beam in the Nomarski interferometer. B is the bandwidth, e is the electron charge, λ is the wavelength of the laser, k the cantilever stiffness, ω_0 is the resonant frequency of the cantilever, Q is the quality factor of the cantilever, T is the temperature, and δi is the variation of current i

	Homodyne interferometer, fiber optic interferometer	**Heterodyne interferometer**	**Nomarski interferometer**
Laser noise $\left\langle \delta \mathrm{i}^2 \right\rangle_\mathrm{L}$	$\frac{1}{4}\eta^2 F^2 P_i^2$ RIN	$\eta^2 \left(P_R^2 + P_S^2\right)$ RIN	$\frac{1}{16}\eta^2 P^2 \delta\theta$
Thermal noise $\left\langle \delta \mathrm{i}^2 \right\rangle_\mathrm{T}$	$\frac{16\pi^2}{\lambda^2}\eta^2 F^2 P_i^2 \frac{4k_\mathrm{B} TBQ}{\omega_0 k}$	$\frac{4\pi^2}{\lambda^2}\eta^2 P_d^2 \frac{4k_\mathrm{B} TBQ}{\omega_0 k}$	$\frac{\pi^2}{\lambda^2}\eta^2 P^2 \frac{4k_\mathrm{B} TBQ}{\omega_0 k}$
Shot Noise $\left\langle \delta \mathrm{i}^2 \right\rangle_\mathrm{S}$	$4e\eta P_\mathrm{d} B$	$2e\eta\,(P_\mathrm{R} + P_\mathrm{S})\,B$	$\frac{1}{2}e\eta PB$

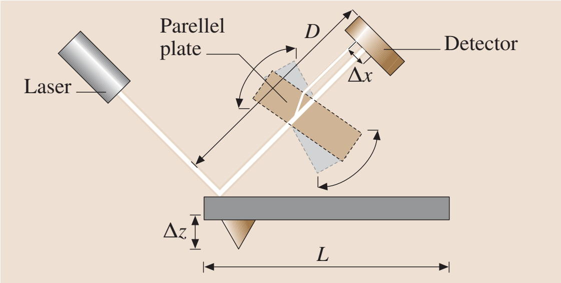

Fig. 11.29 The setup of optical lever detection microscope

the cantilever will cause an imbalance of the number of photons reaching the two halves. Hence the electrical currents in the photodiodes will be unbalanced too. The difference signal is further amplified and is the input signal to the feedback loop. Unlike the interferometeric AFMs, where often a modulation technique is necessary to get a sufficient signal to noise ratio, most AFMs employing the optical lever method are operated in a static mode. AFMs based on the optical lever method are universally used. It is the simplest method for constructing an optical readout and it can be confined in volumes smaller than 5 cm on the side.

The optical lever detection system is a simple yet elegant way to detect normal and lateral force signals simultaneously [11.7, 8, 53, 111]. It has the additional advantage that it is the fact that it is a remote detection system.

Implementations. Light from a laser diode or from a super luminescent diode is focused on the end of the cantilever. The reflected light is directed onto a quadrant diode that measures the direction of the light beam. A Gaussian light beam far from its waist is characterized by an opening angle β. The deflection of the light beam by the cantilever surface tilted by an angle α is 2α. The intensity on the detector then shifts to the side by the product of 2α and the separation between the detector and the cantilever. The readout electronics calculates the difference of the photocurrents. The photocurrents in turn are proportional to the intensity incident on the diode.

The output signal is hence proportional to the change in intensity on the segments

$$I_{\text{sig}} \propto 4\frac{\alpha}{\beta} I_{\text{tot}} \,. \qquad (11.47)$$

For the sake of simplicity, we assume that the light beam is of uniform intensity with its cross section increasing proportional to the distance between the cantilever and the quadrant detector. The movement of the center of the light beam is then given by

$$\Delta x_{\text{Det}} = \Delta z \frac{D}{L} \,. \qquad (11.48)$$

The photocurrent generated in a photodiode is proportional to the number of incoming photons hitting it. If the light beam contains a total number of N_0 photons then the change in difference current becomes

$$\Delta\left(I_{\text{R}} - I_{\text{L}}\right) = \Delta I = \text{const}\, \Delta z\, D\, N_0 \,. \qquad (11.49)$$

Combining (11.48) and (11.49), one obtains that the difference current ΔI is independent of the separation of the quadrant detector and the cantilever. This relation is true, if the light spot is smaller than the quadrant detector. If it is greater, the difference current ΔI becomes smaller with increasing distance. In reality the light beam has a Gaussian intensity profile. For small movements Δx (compared to the diameter of the light spot at the quadrant detector), (11.49) still holds. Larger movements Δx, however, will introduce a nonlinear response. If the AFM is operated in a constant force mode, only small movements Δx of the light spot will occur. The feedback loop will cancel out all other movements.

The scanning of a sample with an AFM can twist the microfabricated cantilevers because of lateral forces [11.5, 7, 8] and affect the images [11.120]. When the tip is subjected to lateral forces, it will twist the cantilever and the light beam reflected from the end of the cantilever will be deflected perpendicular to the ordinary deflection direction. For many investigations this influence of lateral forces is unwanted. The design of the triangular cantilevers stems from the desire, to minimize the torsion effects. However, lateral forces open up a new dimension in force measurements. They allow, for instance, a distinction of two materials because of the different friction coefficient, or the determination of adhesion energies. To measure lateral forces the original optical lever AFM has to be modified. The only modification compared with Fig. 11.29 is the use of a quadrant detector photodiode instead of a two-segment photodiode and the necessary readout electronics, see Fig. 11.9a. The electronics calculates the following signals:

$$\begin{aligned} U_{\text{Normal Force}} &= \alpha \left[\left(I_{\text{Upper Left}} + I_{\text{Upper Right}}\right)\right. \\ &\quad \left. - \left(I_{\text{Lower Left}} + I_{\text{Lower Right}}\right)\right] , \\ U_{\text{Lateral Force}} &= \beta \left[\left(I_{\text{Upper Left}} + I_{\text{Lower Left}}\right)\right. \\ &\quad \left. - \left(I_{\text{Upper Right}} + I_{\text{Lower Right}}\right)\right] . \end{aligned} \qquad (11.50)$$

The calculation of the lateral force as a function of the deflection angle does not have a simple solution for cross-sections other than circles. An approximate formula for the angle of twist for rectangular beams is [11.151]

$$\theta = \frac{M_t L}{\beta G b^3 h}\,, \tag{11.51}$$

where $M_t = F_y \ell$ is the external twisting moment due to lateral force, F_y, and β a constant determined by the value of h/b. For the equation to hold, h has to be larger than b.

Inserting the values for a typical microfabricated cantilever with integrated tips

$$\begin{aligned} b &= 6\times 10^{-7}\,\mathrm{m}\,,\\ h &= 10^{-5}\,\mathrm{m}\,,\\ L &= 10^{-4}\,\mathrm{m}\,,\\ \ell &= 3.3\times 10^{-6}\,\mathrm{m}\,,\\ G &= 5\times 10^{10}\,\mathrm{Pa}\,,\\ \beta &= 0.333 \end{aligned} \tag{11.52}$$

into (11.51) we obtain the relation

$$F_y = 1.1\times 10^{-4}\,\mathrm{N}\times\theta\,. \tag{11.53}$$

Typical lateral forces are of order 10^{-10} N.

Sensitivity. The sensitivity of this setup has been calculated in various papers [11.116, 148, 152]. Assuming a Gaussian beam the resulting output signal as a function of the deflection angle is dispersion like. Equation (11.47) shows that the sensitivity can be increased by increasing the intensity of the light beam I_{tot} or by decreasing the divergence of the laser beam. The upper bound of the intensity of the light I_{tot} is given by saturation effects on the photodiode. If we decrease the divergence of a laser beam we automatically increase the beam waist. If the beam waist becomes larger than the width of the cantilever we start to get diffraction. Diffraction sets a lower bound on the divergence angle. Hence one can calculate the optimal beam waist w_{opt} and the optimal divergence angle β [11.148, 152]

$$\begin{aligned} w_{\mathrm{opt}} &\approx 0.36b\,,\\ \theta_{\mathrm{opt}} &\approx 0.89\frac{\lambda}{b}\,. \end{aligned} \tag{11.54}$$

The optimal sensitivity of the optical lever then becomes

$$\varepsilon\,[\mathrm{mW/rad}] = 1.8\frac{b}{\lambda} I_{\mathrm{tot}}\ [\mathrm{mW}]\ . \tag{11.55}$$

The angular sensitivity optical lever can be measured by introducing a parallel plate into the beam. A tilt of the parallel plate results in a displacement of the beam, mimicking an angular deflection.

Additional noise sources can be considered. Of little importance is the quantum mechanical uncertainty of the position [11.148, 152], which is, for typical cantilevers at room temperature

$$\Delta z = \sqrt{\frac{\hbar}{2m\omega_0}} = 0.05\,\mathrm{fm}\,, \tag{11.56}$$

where $\hbar$ is the Planck constant ($= 6.626\times 10^{-34}$ J s). At very low temperatures and for high frequency cantilevers this could become the dominant noise source. A second noise source is the shot noise of the light. The shot noise is related to the particle number. We can calculate the number of photons incident on the detector

$$n = \frac{I\tau}{\hbar\omega} = \frac{I\lambda}{2\pi B\hbar c} = 1.8\times 10^{9}\,\frac{I[\mathrm{W}]}{B[\mathrm{Hz}]}\,, \tag{11.57}$$

where I is the intensity of the light, τ the measurement time, $B = 1/\tau$ the bandwidth, and c the speed of light. The shot noise is proportional to the square root of the number of particles. Equating the shot noise signal with the signal resulting for the deflection of the cantilever one obtains

$$\Delta z_{\mathrm{shot}} = 68\frac{L}{w}\sqrt{\frac{B\,[\mathrm{kHz}]}{I\,[\mathrm{mW}]}}\,[\mathrm{fm}]\ , \tag{11.58}$$

where w is the diameter of the focal spot. Typical AFM setups have a shot noise of 2 pm. The thermal noise can be calculated from the equipartition principle. The amplitude at the resonant frequency is

$$\Delta z_{\mathrm{therm}} = 129\sqrt{\frac{B}{k\,[\mathrm{N/m}]\,\omega_0 Q}}\,[\mathrm{pm}]\,. \tag{11.59}$$

A typical value is 16 pm. Upon touching the surface, the cantilever increases its resonant frequency by a factor of 4.39. This results in a new thermal noise amplitude of 3.2 pm for the cantilever in contact with the sample.

Piezoresistive Detection

Implementation. An alternative detection system which is not as widely used as the optical detection schemes are piezoresistive cantilevers [11.125, 126, 132]. These cantilevers are based on the fact that the resistivity of certain materials, in particular of Si, changes with the applied stress. Figure 11.30 shows a typical implementation of a piezo-resistive cantilever. Four

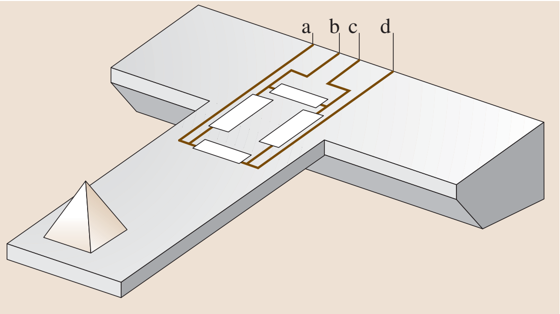

Fig. 11.30 A typical setup for a piezoresistive readout

resistances are integrated on the chip, forming a Wheatstone bridge. Two of the resistors are in unstrained parts of the cantilever, the other two are measuring the bending at the point of the maximal deflection. For instance when an AC voltage is applied between terminals a and c one can measure the detuning of the bridge between terminals b and d. With such a connection the output signal varies only due to bending, but not due to changing of the ambient temperature and thus the coefficient of the piezoresistance.

Sensitivity. The resistance change is [11.126]

$$\frac{\Delta R}{R_0} = \Pi\delta\,, \tag{11.60}$$

where Π is the tensor element of the piezo-resistive coefficients, δ the mechanical stress tensor element and R_0 the equilibrium resistance. For a single resistor they separate the mechanical stress and the tensor element in longitudinal and transverse components

$$\frac{\Delta R}{R_0} = \Pi_{\mathrm{t}}\delta_{\mathrm{t}} + \Pi_{\mathrm{l}}\delta_{\mathrm{l}}\,. \tag{11.61}$$

The maximum value of the stress components are $\Pi_{\mathrm{t}} = -64.0\times10^{-11}\,\mathrm{m}^2/\mathrm{N}$ and $\Pi_{\mathrm{l}} = -71.4\times10^{-11}\,\mathrm{m}^2/\mathrm{N}$ for a resistor oriented along the (110) direction in silicon [11.126]. In the resistor arrangement of Fig. 11.30 two of the resistors are subject to the longitudinal piezoresistive effect and two of them are subject to the transversal piezo-resistive effect. The sensitivity of that setup is about four times that of a single resistor, with the advantage that temperature effects cancel to first order. The resistance change is then calculated as

$$\frac{\Delta R}{R_0} = \Pi\frac{3Eh}{2L^2}\Delta z = \Pi\frac{6L}{bh^2}F_z\,, \tag{11.62}$$

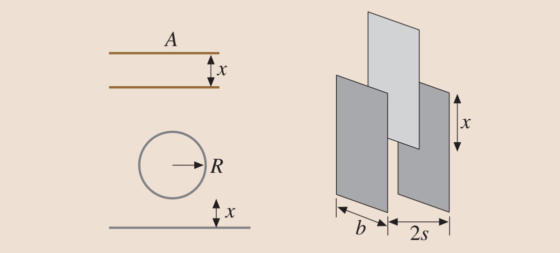

Fig. 11.31 Three possible arrangements of a capacitive readout. The *upper left* shows the cross section through a parallel plate capacitor. The *lower left* shows the geometry of sphere versus plane. The *right side* shows the more complicated, but linear capacitive readout

where $\Pi = 67.7\times10^{-11}\,\mathrm{m}^2/\mathrm{N}$ is the averaged piezoresistive coefficient. Plugging in typical values for the dimensions (Fig. 11.24) ($L = 100\,\mu\mathrm{m}$, $b = 10\,\mu\mathrm{m}$, $h = 1\,\mu\mathrm{m}$) one obtains

$$\frac{\Delta R}{R_0} = \frac{4\times10^{-5}}{\mathrm{nN}}F_z\,. \tag{11.63}$$

The sensitivity can be tailored by optimizing the dimensions of the cantilever.

Capacitance Detection

The capacitance of an arrangement of conductors depends on the geometry. Generally speaking, the capacitance increases for decreasing separations. Two parallel plates form a simple capacitor (see Fig. 11.31, upper left), with the capacitance

$$C = \frac{\varepsilon\varepsilon_0 A}{x}\,, \tag{11.64}$$

where A is the area of the plates, assumed equal, and x is the separation. Alternatively one can consider a sphere versus an infinite plane (see Fig. 11.31, lower left). Here the capacitance is [11.116]

$$C = 4\pi\varepsilon_0 R\sum_{n=2}^{\infty}\frac{\sinh(\alpha)}{\sinh(n\alpha)} \tag{11.65}$$

where R is the radius of the sphere, and α is defined by

$$\alpha = \ln\left(1 + \frac{z}{R} + \sqrt{\frac{z^2}{R^2} + 2\frac{z}{R}}\right)\,. \tag{11.66}$$

One has to keep in mind that capacitance of a parallel plate capacitor is a nonlinear function of the separation. Using a voltage divider one can circumvent this problem.

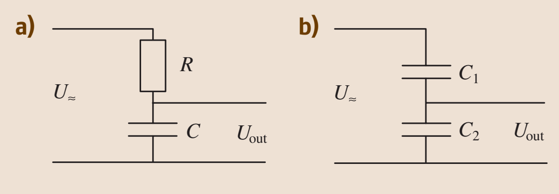

Fig. 11.32a,b Measuring the capacitance. The *left side*, **(a)**, shows a low pass filter, the *right side*, **(b)**, shows a capacitive divider. C (*left*) or C_2 (*right*) are the capacitances under test

Figure 11.32a shows a low pass filter. The output voltage is given by

$$U_{\text{out}} = U_{\approx}\frac{\frac{1}{j\omega C}}{R+\frac{1}{j\omega C}} = U_{\approx}\frac{1}{j\omega CR+1} \cong \frac{U_{\approx}}{j\omega CR} . \qquad (11.67)$$

Here C is given by (11.64), ω is the excitation frequency and j is the imaginary unit. The approximate relation in the end is true when $\omega CR \gg 1$. This is equivalent to the statement that C is fed by a current source, since R must be large in this setup. Plugging (11.64) into (11.67) and neglecting the phase information one obtains

$$U_{\text{out}} = \frac{U_{\approx}x}{\omega R\varepsilon\varepsilon_0 A} , \qquad (11.68)$$

which is linear in the displacement x.

Figure 11.32b shows a capacitive divider. Again the output voltage U_{out} is given by

$$U_{\text{out}} = U_{\approx}\frac{C_1}{C_2+C_1} = U_{\approx}\frac{C_1}{\frac{\varepsilon\varepsilon_0 A}{x}+C_1} . \qquad (11.69)$$

If there is a stray capacitance C_{s} then (11.69) is modified as

$$U_{\text{out}} = U_{\approx}\frac{C_1}{\frac{\varepsilon\varepsilon_0 A}{x}+C_{\text{s}}+C_1} . \qquad (11.70)$$

Provided $C_{\text{s}}+C_1 \ll C_2$ one has a system which is linear in x. The driving voltage $U_{\approx}$ has to be large (more than 100 V) to have the output voltage in the range of 1 V. The linearity of the readout depends on the capacitance C_1 (Fig. 11.33).

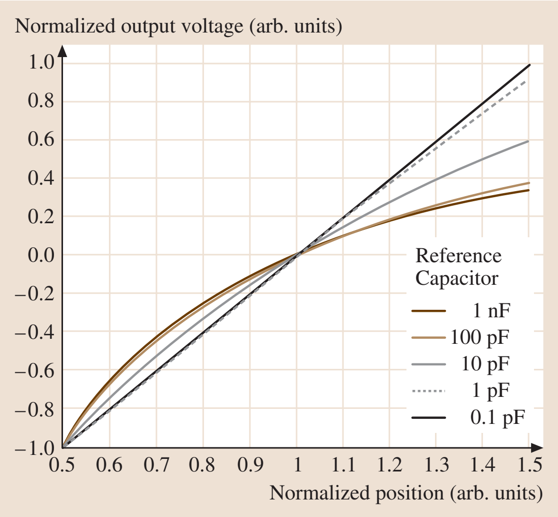

Fig. 11.33 Linearity of the capacitance readout as a function of the reference capacitor

Another idea is to keep the distance constant and to change the relative overlap of the plates (see Fig. 11.31, right side). The capacitance of the moving center plate versus the stationary outer plates becomes

$$C = C_{\text{s}} + 2\frac{\varepsilon\varepsilon_0 bx}{s} , \qquad (11.71)$$

where the variables are defined in Fig. 11.31. The stray capacitance comprises all effects, including the capacitance of the fringe fields. When the length x is comparable to the width b of the plates one can safely assume that the stray capacitance is constant and independent of x. The main disadvantage of this setup is that it is not as easily incorporated in a microfabricated device as the others.

Sensitivity. The capacitance itself is not a measure of the sensitivity, but its derivative is indicative of the signals one can expect. Using the situation described in Fig. 11.31, upper left, and in (11.64) one obtains for the parallel plate capacitor

$$\frac{\mathrm{d}C}{\mathrm{d}x} = -\frac{\varepsilon\varepsilon_0 A}{x^2} . \qquad (11.72)$$

Assuming a plate area A of 20 µm by 40 µm and a separation of 1 µm one obtains a capacitance of 31 fF (neglecting stray capacitance and the capacitance of the connection leads) and a $\mathrm{d}C/\mathrm{d}x$ of $3.1\times10^{-8}\,\text{F/m} = 31\,\text{fF/µm}$. Hence it is of paramount importance to maximize the area between the two contacts and to minimize the distance x. The latter however is far from being trivial. One has to go to the limits of microfabrication to achieve a decent sensitivity.

If the capacitance is measured by the circuit shown in Fig. 11.32 one obtains for the sensitivity

$$\frac{\mathrm{d}U_{\mathrm{out}}}{U_{\approx}} = \frac{\mathrm{d}x}{\omega R \varepsilon \varepsilon_0 A} . \tag{11.73}$$

Using the same value for A as above, setting the reference frequency to 100 kHz, and selecting $R = 1\,\mathrm{G}\Omega$, we get the relative change of the output voltage U_{out} to

$$\frac{\mathrm{d}U_{\mathrm{out}}}{U_{\approx}} = \frac{22.5 \times 10^{-6}}{\text{Å}} \times \mathrm{d}x . \tag{11.74}$$

A driving voltage of 45 V then translates to a sensitivity of 1 mV/Å. A problem in this setup is the stray capacitances. They are in parallel to the original capacitance and decrease the sensitivity considerably.

Alternatively one could build an oscillator with this capacitance and measure the frequency. RC-oscillators typically have an oscillation frequency of

$$f_{\mathrm{res}} \propto \frac{1}{RC} = \frac{x}{R \varepsilon \varepsilon_0 A} . \tag{11.75}$$

Again the resistance R must be of the order of $1\,\mathrm{G}\Omega$, when stray capacitances C_{s} are neglected. However C_{s} is of the order of 1 pF. Therefore one gets $R = 10\,\mathrm{M}\Omega$. Using these values the sensitivity becomes

$$\mathrm{d}f_{\mathrm{res}} = \frac{C\,\mathrm{d}x}{R\,(C + C_{\mathrm{s}})^2\,x} \approx \frac{0.1\,\mathrm{Hz}}{\text{Å}}\,\mathrm{d}x . \tag{11.76}$$

The bad thing is that the stray capacitances have made the signal nonlinear again. The linearized setup in Fig. 11.31 has a sensitivity of

$$\frac{\mathrm{d}C}{\mathrm{d}x} = 2\frac{\varepsilon \varepsilon_0 b}{s} . \tag{11.77}$$

Substituting typical values, $b = 10\,\mu\mathrm{m}$, $s = 1\,\mu\mathrm{m}$ one gets $\mathrm{d}C/\mathrm{d}x = 1.8 \times 10^{-10}\,\mathrm{F/m}$. It is noteworthy that the sensitivity remains constant for scaled devices.

Implementations. The readout of the capacitance can be done in different ways [11.123, 124]. All include an alternating current or voltage with frequencies in the 100 kHz to the 100 MHz range. One possibility is to build a tuned circuit with the capacitance of the cantilever determining the frequency. The resonance frequency of a high quality Q tuned circuit is

$$\omega_0 = (LC)^{-1/2} . \tag{11.78}$$

where L is the inductance of the circuit. The capacitance C includes not only the sensor capacitance but also the capacitance of the leads. The precision of a frequency measurement is mainly determined by the ratio of L and C

$$Q = \left(\frac{L}{C}\right)^{1/2} \frac{1}{R} . \tag{11.79}$$

Here R symbolizes the losses in the circuit. The higher the quality the more precise the frequency measurement. For instance a frequency of 100 MHz and a capacitance of 1 pF gives an inductance of 250 μH. The quality becomes then 2.5×10^8. This value is an upper limit, since losses are usually too high.

Using a value of $\mathrm{d}C/\mathrm{d}x = 31\,\mathrm{fF}/\mu\mathrm{m}$ one gets $\Delta C/\text{Å} = 3.1\,\mathrm{aF}/\text{Å}$. With a capacitance of 1 pF one gets

$$\frac{\Delta\omega}{\omega} = \frac{1}{2}\frac{\Delta C}{C} ,$$

$$\Delta\omega = 100\,\mathrm{MHz} \times \frac{1}{2}\frac{3.1\,\mathrm{aF}}{1\,\mathrm{pF}} = 155\,\mathrm{Hz} . \tag{11.80}$$

This is the frequency shift for 1 Å deflection. The calculation shows, that this is a measurable quantity. The quality also indicates that there is no physical reason why this scheme should not work.

11.3.3 Combinations for 3-D-Force Measurements

Three dimensional force measurements are essential if one wants to know all the details of the interaction between the tip and the cantilever. The straightforward attempt to measure three forces is complicated, since force sensors such as interferometers or capacitive sensors need a minimal detection volume, which often is too large. The second problem is that the force-sensing tip has to be held by some means. This implies that one of the three Cartesian axes is stiffer than the others.

However by the combination of different sensors one can achieve this goal. Straight cantilevers are employed for these measurements, because they can be handled analytically. The key observation is, that the optical lever method does not determine the position of the end of the cantilever. It measures the orientation. In the previous sections, one has always made use of the fact, that for a force along one of the orthogonal symmetry directions at the end of the cantilever (normal force, lateral force, force along the cantilever beam axis) there is a one to one correspondence of the tilt angle and the deflection. The problem is, that the force along the cantilever beam axis and the normal force create a deflection in the same direction. Hence what is called the normal force component is actually a mixture of two forces. The deflection of the cantilever is the third quantity, which is not considered in most of the AFMs. A fiber optic interferometer in parallel to the optical lever measures the deflection. Three measured quantities then allow the separation of the three orthonormal force directions, as is evident from (11.27) and (11.33) [11.12–16].

Alternatively one can put the fast scanning direction along the axis of the cantilever. Forward and backward scans then exert opposite forces F_x. If the piezo movement is linearized, both force components in AFM based on the optical lever detection can be determined. In this case, the normal force is simply the average of the forces in the forward and backward direction. The force, F_x, is the difference of the forces measured in forward and backward direction.

11.3.4 Scanning and Control Systems

Almost all SPMs use piezo translators to scan the tip or the sample. Even the first STM [11.1, 103] and some of the predecessor instruments [11.153, 154] used them. Other materials or setups for nano-positioning have been proposed, but were not successful [11.155, 156].

Piezo Tubes

A popular solution is tube scanners (Fig. 11.34). They are now widely used in SPMs due to their simplicity and their small size [11.133, 157]. The outer electrode is segmented in four equal sectors of 90 degrees. Opposite sectors are driven by signals of the same magnitude, but opposite sign. This gives, through bending, a two-dimensional movement on, approximately, a sphere. The inner electrode is normally driven by the z signal. It is possible, however, to use only the outer electrodes for scanning and for the z-movement. The main drawback of applying the z-signal to the outer electrodes is, that the applied voltage is the sum of both the x- or y-movement and the z-movement. Hence a larger scan size effectively reduces the available range for the z-control.

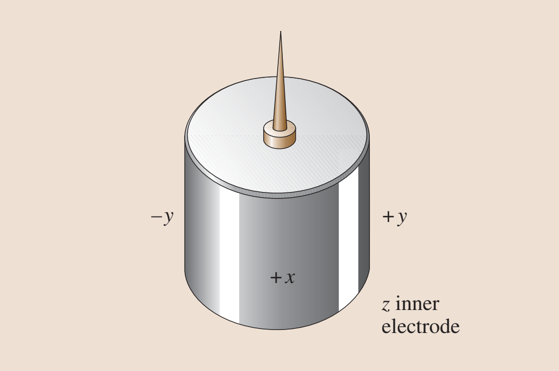

Fig. 11.34 Schematic drawing of a piezoelectric tube scanner. The piezo ceramic is molded into a tube form. The outer electrode is separated into four segments and connected to the scanning voltages. The z-voltage is applied to the inner electrode

Piezo Effect

An electric field applied across a piezoelectric material causes a change in the crystal structure, with expansion in some directions and contraction in others. Also, a net volume change occurs [11.132]. Many SPMs use the transverse piezo electric effect, where the applied electric field $\boldsymbol{E}$ is perpendicular to the expansion/contraction direction.

$$\Delta L = L\,(\boldsymbol{E}\cdot\boldsymbol{n})\,d_{31} = L\frac{V}{t}d_{31}\,, \qquad (11.81)$$

where d_{31} is the transverse piezoelectric constant, V is the applied voltage, t is the thickness of the piezo slab or the distance between the electrodes where the voltage is applied, L is the free length of the piezo slab, and $\boldsymbol{n}$ is the direction of polarization. Piezo translators based on the transverse piezoelectric effect have a wide range of sensitivities, limited mainly by mechanical stability and breakdown voltage.

Scan Range

The calculation of the scanning range of a piezotube is difficult [11.157–159]. The bending of the tube depends on the electric fields and the nonuniform strain induced. A finite element calculation where the piezo tube was divided into 218 identical elements was used [11.158] to calculate the deflection. On each node the mechanical stress, stiffness, strain and piezoelectric stress was calculated when a voltage was applied on one electrode. The results were found to be linear on the first iteration and higher-order corrections were very small even for large electrode voltages. It was found that to first order the x- and z-movement of the tube could be reasonably well approximated by assuming that the piezo tube is a segment of a torus. Using this model one obtains

$$\mathrm{d}x = (V_+ - V_-)\,|d_{31}|\,\frac{L^2}{2t\,\mathrm{d}}\,, \qquad (11.82)$$

$$\mathrm{d}z = (V_+ + V_- - 2V_z)\,|d_{31}|\,\frac{L}{2t}\,, \qquad (11.83)$$

where $|d_{31}|$ is the coefficient of the transversal piezoelectric effect, L is tube's free length, t is tube's wall thickness, d is tube's diameter, V_+ is voltage on positive outer electrode while V_- is voltage of the opposite quadrant negative electrode, and V_z is voltage of the inner electrode.

The cantilever or sample mounted on the piezotube has an additional lateral movement because the point of measurement is not in the end plane of the piezotube. The additional lateral displacement of the end of the tip is $\ell \sin\varphi \approx \ell\varphi$, where ℓ is the tip length and φ is the deflection angle of the end surface. Assuming that the sample or cantilever are always perpendicular to the end of the walls of the tube and calculating with the torus model one gets for the angle

$$\varphi = \frac{L}{R} = \frac{2\,\mathrm{d}x}{L}\,, \tag{11.84}$$

where R is the radius of curvature the piezo tube. Using the result of (11.84) one obtains for the additional x-movement

$$\begin{aligned}\mathrm{d}x_{\text{add}} &= \ell\varphi = \frac{2\,\mathrm{d}x\ell}{L}\\ &= (V_+ - V_-)\,|d_{31}|\,\frac{\ell L}{td}\end{aligned} \tag{11.85}$$

and for the additional z-movement due to the x-movement

$$\begin{aligned}\mathrm{d}z_{\text{add}} &= \ell - \ell\cos\varphi = \frac{\ell\varphi^2}{2} = \frac{2\ell\,(\mathrm{d}x)^2}{L^2}\\ &= (V_+ - V_-)^2\,|d_{31}|^2\,\frac{\ell L^2}{2t^2d^2}\,.\end{aligned} \tag{11.86}$$

Carr [11.158] assumed for his finite element calculations that the top of the tube was completely free to move and, as a consequence, the top surface was distorted, leading to a deflection angle about half that of the geometrical model. Depending on the attachment of the sample or the cantilever this distortion may be smaller, leading to a deflection angle in-between that of the geometrical model and the one of the finite element calculation.

Nonlinearities and Creep

Piezo materials with a high conversion ratio, i. e. a large d_{31} or small electrode separations, with large scanning ranges are hampered by substantial hysteresis resulting in a deviation from linearity by more than 10%. The sensitivity of the piezo ceramic material (mechanical displacement divided by driving voltage) decreases with reduced scanning range, whereas the hysteresis is reduced. A careful selection of the material for the piezo scanners, the design of the scanners, and of the operating conditions is necessary to get optimum performance.

Passive Linearization: Calculation. The analysis of images affected by piezo nonlinearities [11.160–163] shows that the dominant term is

$$x = AV + BV^2\,, \tag{11.87}$$

where x is the excursion of the piezo, V the applied voltage and A and B two coefficients describing the sensitivity of the material. Equation (11.87) holds for scanning from $V = 0$ to large V. For the reverse direction the equation becomes

$$x = \tilde{A}V - \tilde{B}\,(V - V_{\max})^2\,, \tag{11.88}$$

where $\tilde{A}$ and $\tilde{B}$ are the coefficients for the back scan and $V_{\max}$ is the applied voltage at the turning point. Both equations demonstrate that the true x-travel is small at the beginning of the scan and becomes larger towards the end. Therefore images are stretched at the beginning and compressed at the end.

Similar equations hold for the slow scan direction. The coefficients, however, are different. The combined action causes a greatly distorted image. This distortion can be calculated. The data acquisition systems record the signal as a function of V. However the data is measured as a function of x. Therefore we have to distribute the x-values evenly across the image this can be done by inverting an approximation of (11.87). First we write

$$x = AV\left(1 - \frac{B}{A}V\right)\,. \tag{11.89}$$

For $B \ll A$ we can approximate

$$V = \frac{x}{A}\,. \tag{11.90}$$

We now substitute (11.90) into the nonlinear term of (11.89). This gives

$$\begin{aligned}x &= AV\left(1 + \frac{Bx}{A^2}\right)\,,\\ V &= \frac{x}{A}\,\frac{1}{(1 + Bx/A^2)} \approx \frac{x}{A}\left(1 - \frac{Bx}{A^2}\right)\,.\end{aligned} \tag{11.91}$$

Hence an equation of the type

$$\begin{aligned}&x_{\text{true}} = x\,(\alpha - \beta x/x_{\max})\\ &\text{with}\quad 1 = \alpha - \beta\end{aligned} \tag{11.92}$$

takes out the distortion of an image. α and β are dependent on the scan range, the scan speed and on the scan history and have to be determined with exactly the same settings as for the measurement. $x_{\max}$ is the maximal scanning range. The condition for α and β guarantees that the image is transformed onto itself.

Similar equations as the empirical one shown above (11.92) can be derived by analyzing the movements of domain walls in piezo ceramics.

Passive Linearization: Measuring the Position. An alternative strategy is to measure the position of the piezo translators. Several possibilities exist.

1. The interferometers described above can be used to measure the elongation of the piezo elongation. The fiber optic interferometer is especially easy to implement. The coherence length of the laser only limits the measurement range. However the signal is of a periodic nature. Hence a direct use of the signal in a feedback circuit for the position is not possible. However as a measurement tool and, especially, as a calibration tool the interferometer is without competition. The wavelength of the light, for instance in a HeNe laser, is so well defined that the precision of the other components determines the error of the calibration or measurement.
2. The movement of the light spot on the quadrant detector can be used to measure the position of a piezo [11.164]. The output current changes by $0.5\,\mathrm{A/cm} \times P(\mathrm{W})/R(\mathrm{cm})$. Typical values ($P = 1\,\mathrm{mW}$, $R = 0.001\,\mathrm{cm}$) give $0.5\,\mathrm{A/cm}$. The noise limit is typically $0.15\,\mathrm{nm} \times \sqrt{\Delta f(\mathrm{Hz})/H(\mathrm{W/cm^2})}$. Again this means that the laser beam above would have a 0.1 nm noise limitation for a bandwidth of 21 Hz. The advantage of this method is that, in principle, one can linearize two axes with only one detector.
3. A knife-edge blocking part of a light beam incident on a photodiode can be used to measure the position of the piezo. This technique, commonly used in optical shear force detection [11.75, 165], has a sensitivity of better than 0.1 nm.
4. The capacitive detection [11.166, 167] of the cantilever deflection can be applied to the measurement of the piezo elongation. Equations (11.64) to (11.79) apply to the problem. This technique is used in some commercial instruments. The difficulties lie in the avoidance of fringe effects at the borders of the two plates. While conceptually simple, one needs the latest technology in surface preparation to get a decent linearity. The electronic circuits used for the readout are often proprietary.
5. Linear Variable Differential Transformers (LVDT) are a convenient means to measure positions down to 1 nm. They can be used together with a solid state joint setup, as often used for large scan range stages. Unlike the capacitive detection there are few difficulties in implementation. The sensors and the detection circuits LVDTs are available commercially.
6. A popular measurement technique is the use of strain gauges. They are especially sensitive when mounted on a solid state joint where the curvature is maximal. The resolution depends mainly on the induced curvature. A precision of 1 nm is attainable. The signals are low – a Wheatstone bridge is needed for the readout.

Active Linearization. Active linearization is done with feedback systems. Sensors need to be monotonic. Hence all the systems described above, with the exception of the interferometers are suitable. The most common solutions include the strain gauge approach, the capacitance measurement or the LVDT, which are all electronic solutions. Optical detection systems have the disadvantage that the intensity enters into the calibration.

Alternative Scanning Systems

The first STMs were based on piezo tripods [11.1]. The piezo tripod (Fig. 11.35) is an intuitive way to generate the three dimensional movement of a tip attached to its center. However, to get a suitable stability and scanning range, the tripod needs to be fairly large (about 50 mm). Some instruments use piezo stacks instead of monolithic piezoactuators. They are arranged in the tripod arrangement. Piezo stacks are thin layers of piezoactive materials glued together to form a device with up to 200 μm of actuation range. Preloading with a suitable metal casing reduces the nonlinearity.

If one tries to construct a homebuilt scanning system, the use of linearized scanning tables is recommended. They are built around solid state joints and actuated by piezo stacks. The joints guarantee that the movement is parallel with little deviation from the predefined scanning plane. Due to the construction it is easy to add measurement devices such as capacitive sensors, LVDTs

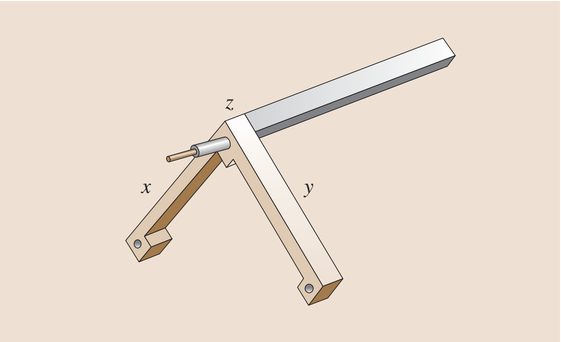

Fig. 11.35 An alternative type of piezo scanners: the tripod

or strain gauges which are essential for a closed loop linearization. Two-dimensional tables can be bought from several manufacturers. They have a linearity of better than 0.1% and a noise level of 10^{-4} to 10^{-5} of the maximal scanning range.

Control Systems

Basics. The electronics and software play an important role in the optimal performance of an SPM. Control electronics and software are supplied with commercial SPMs. Control electronic systems can use either analog or digital feedback. While digital feedback offers greater flexibility and the ease of configuration, analog feedback circuits might be better suited for ultralow noise operation. We will describe here the basic setups for AFMs.

Figure 11.36 shows a block schematic of a typical AFM feedback loop. The signal from the force transducer is fed into the feedback loop consisting mainly of a subtraction stage to get an error signal and an integrator. The gain of the integrator (high gain corresponds to short integration times) is set as high as possible without generating more than 1% overshoot. High gain minimizes the error margin of the current and forces the tip to follow the contours of constant density of states as well as possible. This operating mode is known as Constant Force Mode. A high voltage amplifier amplifies the outputs of the integrator. As AFMs using piezotubes usually require ±150 V at the output, the output of the integrator needs to be amplified by a high voltage amplifier.

In order to scan the sample, additional voltages at high tension are required to drive the piezo. For example, with a tube scanner, four scanning voltages are required, namely $+V_x$, $-V_x$, $+V_y$ and $-V_y$. The x- and y-scanning voltages are generated in a scan generator (analog or computer controlled). Both voltages are input to the two respective power amplifiers. Two inverting amplifiers generate the input voltages for the other two power amplifiers. The topography of the sample surface is determined by recording the input-voltage to the high voltage amplifier for the z-channel as a function of x and y (Constant Force Mode).

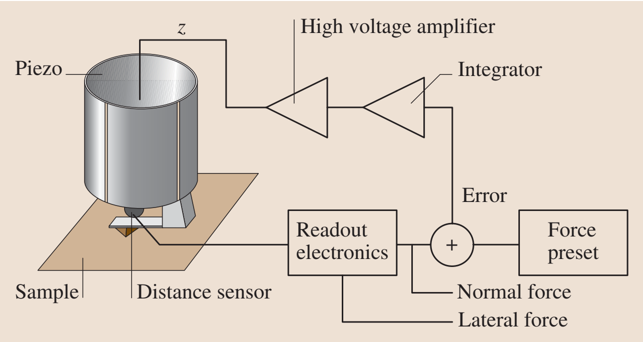

Fig. 11.36 Block schematic of the feedback control loop of an AFM

Another operating mode is the Variable Force Mode. The gain in the feedback loop is lowered and the scanning speed increased such that the force on the cantilever is no longer constant. Here the force is recorded as a function of x and y.

Force Spectroscopy. Four modes of spectroscopic imaging are in common use with force microscopes: measuring lateral forces, $\partial F/\partial z$, $\partial F/\partial x$ spatially resolved, and measuring force versus distance curves. Lateral forces can be measured by detecting the deflection of a cantilever in a direction orthogonal to the normal direction. The optical lever deflection method most easily does this. Lateral force measurements give indications of adhesion forces between the tip and the sample.

$\partial F/\partial z$ measurements probe the local elasticity of the sample surface. In many cases the measured quantity originates from a volume of a few cubic nanometers. The $\partial F/\partial z$ or local stiffness signal is proportional to Young's modulus, as far as one can define this quantity. Local stiffness is measured by vibrating the cantilever by a small amount in the z-direction. The expected signal for very stiff samples is zero: for very soft samples one gets, independent of the stiffness, also a constant signal. This signal is again zero for the optical lever deflection and equal to the driving amplitude for interferometeric measurements. The best sensitivity is obtained when the compliance of the cantilever matches the stiffness of the sample.

A third spectroscopic quantity is the lateral stiffness. It is measured by applying a small modulation in the x-direction on the cantilever. The signal is again optimal when the lateral compliance of the cantilever matches the lateral stiffness of the sample. The lateral stiffness is, in turn, related to the shear modulus of the sample.

Detailed information on the interaction of the tip and the sample can be gained by measuring force versus distance curves. It is necessary to have cantilevers with high enough compliance to avoid instabilities due to the attractive forces on the sample.

Using the Control Electronics as a Two-Dimensional Measurement Tool. Usually the control electronics of an AFM is used to control the x- and y-piezo signals while several data acquisition channels record the position dependent signals. The control electronics can be used in another way: it can be viewed as a two-dimensional function generator. What is normally the x- and y-signal can be used to control two independent variables of an experiment. The control logic of the AFM then ensures that the available parameter space is systematically probed at equally spaced points. An example is friction force curves measured along a line across a step on graphite.

Figure 11.37 shows the connections. The z-piezo is connected as usual, like the x-piezo. However the y-output is used to command the desired input parameter. The offset of the y-channel determines the position of the tip on the sample surface, together with the x-channel.

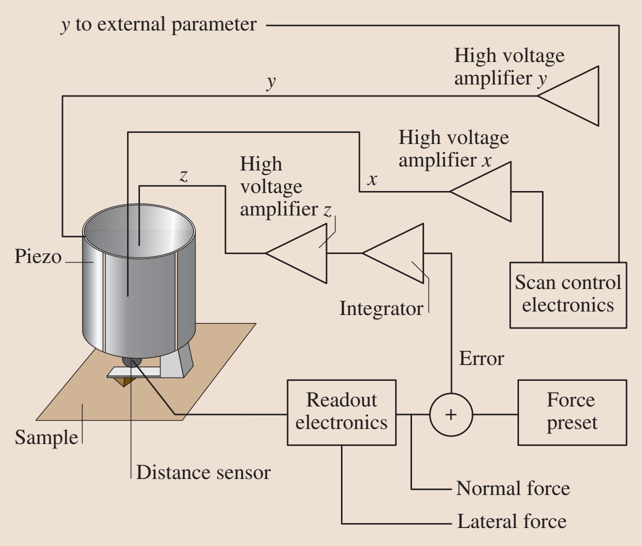

Fig. 11.37 Wiring of an AFM to measure friction force curves along a line

Some Imaging Processing Methods

The visualization and interpretation of images from AFMs is intimately connected to the processing of these images. An ideal AFM is a noise-free device that images a sample with perfect tips of known shape and has perfect linear scanning piezos. In reality, AFMs are not that ideal. The scanning device in AFMs is affected by distortions. The distortions are both linear and nonlinear. Linear distortions mainly result from imperfections in the machining of the piezotranslators causing crosstalk from the Z-piezo to the X- and Y-piezos, and vice versa. Among the linear distortions, there are two kinds which are very important. First, scanning piezos invariably have different sensitivities along the different scan axes due to the variation of the piezo material and uneven sizes of the electrode areas. Second, the same reasons might cause the scanning axes not to be orthogonal. Furthermore, the plane in which the piezoscanner moves for constant height z is hardly ever coincident with the sample plane. Hence, a linear ramp is added to the sample data. This ramp is especially bothersome when the height z is displayed as an intensity map.

The nonlinear distortions are harder to deal with. They can affect AFM data for a variety of reasons. First, piezoelectric ceramics do have a hysteresis loop, much like ferromagnetic materials. The deviations of piezoceramic materials from linearity increase with increasing amplitude of the driving voltage. The mechanical position for one voltage depends on the previously applied voltages to the piezo. Hence, to get the best positional accuracy, one should always approach a point on the sample from the same direction. Another type of nonlinear distortion of the images occurs when the scan frequency approaches the upper frequency limit of the X- and Y-drive amplifiers or the upper frequency limit of the feedback loop (z-component). This distortion, due to feedback loop, can only be minimized by reducing the scan frequency. On the other hand, there is a simple way to reduce distortions due to the X- and Y-piezo drive amplifiers. To keep the system as simple as possible, one normally uses a triangular waveform for driving the scanning piezos. However, triangular waves contain frequency components as multiples of the scan frequency. If the cutoff frequency of the X- and Y-drive electronics or of the feedback loop is too close to the scanning frequency (two or three times the scanning frequency), the triangular drive voltage is rounded off at the turning points. This rounding error causes, first, a distortion of the scan linearity and, second, through phase lags, the projection of part of the backward scan onto the forward scan. This type of distortion can be minimized by carefully selecting the scanning frequency and by using driving voltages for the X- and Y-piezos with waveforms like trapezoidal waves, which are closer to a sine wave. The values measured for X, Y, or Z piezos are affected by noise. The origin of this noise can be either electronic, some disturbances, or a property of the sample surface due to adsorbates. In addition to this incoherent noise, interference with main and other equipment nearby might be present. Depending on the type of noise, one can filter it in the real space or in Fourier space. The most important part of image processing is to visualize the measured data. Typical AFM data sets can consist of many thousands to over a million points per plane. There may be more than one image plane present.

The AFM data represents a topography in various data spaces.

Most commercial data acquisition systems use implicitly some kind of data processing. Since the original data is commonly subject to slopes on the surface, most programs use some kind of slope correction. The least disturbing way is to substrate a plane $z(x, y) = Ax + By + C$ from the data. The coefficients are determined by fitting $z(x, y)$ to the data. Another operation is to subtract a second order function such as $z(x, y) = Ax^2 + By^2 + Cxy + Dx + Ey + F$. Again, the parameters are determined with a fit. This function is appropriate for almost planar data, where the nonlinearity of the piezos caused such a distortion.

In the image processing software from Digital Instruments, up to three operations are performed on the raw data. First, a zero-order flatten is applied. The flatten operation is used to eliminate image bow in the slow scan direction (caused by a physical bow in the instrument itself), slope in the slow scan direction, bands in the image (caused by differences in the scan height from one scan line to the next). The flattening operation takes each scan line and subtracts the average value of the height along each scan line from each point in that scan line. This brings each scan line to the same height. Next, a first-order plane-fit is applied in the fast scan direction. The plane-fit operation is used to eliminate bow and slope in the fast scan direction. The plane-fit operation calculated a best-fit plane for the image and subtracts it from the image. This plane has a constant non-zero slope in the fast scan direction. In some cases, a higher-order polynomial "plane" may be required. Depending upon the quality of the raw data, the flattening operation and/or the planefit operation may not be required at all.

References

11.1 G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel: Surface studies by scanning tunneling microscopy, Phys. Rev. Lett. **49** (1982) 57–61

11.2 G. Binnig, C. F. Quate, Ch. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

11.3 G. Binnig, Ch. Gerber, E. Stoll, T. R. Albrecht, C. F. Quate: Atomic resolution with atomic force microscope, Europhys. Lett. **3** (1987) 1281–1286

11.4 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

11.5 C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang: Atomic-scale friction of a tungsten tip on a graphite surface, Phys. Rev. Lett. **59** (1987) 1942–1945

11.6 R. Erlandsson, G. M. McClelland, C. M. Mate, S. Chiang: Atomic force microscopy using optical interferometry, J. Vac. Sci. Technol. A **6** (1988) 266–270

11.7 O. Marti, J. Colchero, J. Mlynek: Combined scanning force and friction microscopy of mica, Nanotechnol. **1** (1990) 141–144

11.8 G. Meyer, N. M. Amer: Simultaneous measurement of lateral and normal forces with an optical-beam-deflection atomic force microscope, Appl. Phys. Lett. **57** (1990) 2089–2091

11.9 B. Bhushan, J. Ruan: Atomic-scale friction measurements using friction force microscopy: Part II – Application to magnetic media, ASME J. Tribol. **116** (1994) 389–396

11.10 B. Bhushan, V. N. Koinkar, J. Ruan: Microtribology of magnetic media, Proc. Inst. Mech. Eng., Part J: J. Eng. Tribol. **208** (1994) 17–29

11.11 B. Bhushan, J. N. Israelachvili, U. Landman: Nanotribology: Friction, wear, and lubrication at the atomic scale, Nature **374** (1995) 607–616

11.12 S. Fujisawa, M. Ohta, T. Konishi, Y. Sugawara, S. Morita: Difference between the forces measured by an optical lever deflection and by an optical interferometer in an atomic force microscope, Rev. Sci. Instrum. **65** (1994) 644–647

11.13 S. Fujisawa, E. Kishi, Y. Sugawara, S. Morita: Fluctuation in 2-dimensional stick-slip phenomenon observed with 2-dimensional frictional force microscope, Jpn. J. Appl. Phys. **33** (1994) 3752–3755

11.14 S. Grafstrom, J. Ackermann, T. Hagen, R. Neumann, O. Probst: Analysis of lateral force effects on the topography in scanning force microscopy, J. Vac. Sci. Technol. B **12** (1994) 1559–1564

11.15 R. M. Overney, H. Takano, M. Fujihira, W. Paulus, H. Ringsdorf: Anisotropy in friction and molecular stick-slip motion, Phys. Rev. Lett. **72** (1994) 3546–3549

11.16 R. J. Warmack, X. Y. Zheng, T. Thundat, D. P. Allison: Friction effects in the deflection of atomic force microscope cantilevers, Rev. Sci. Instrum. **65** (1994) 394–399

11.17 N. A. Burnham, D. D. Domiguez, R. L. Mowery, R. J. Colton: Probing the surface forces of monolayer films with an atomic force microscope, Phys. Rev. Lett. **64** (1990) 1931–1934

11.18 N. A. Burham, R. J. Colton, H. M. Pollock: Interpretation issues in force microscopy, J. Vac. Sci. Technol. A **9** (1991) 2548–2556

11.19 C. D. Frisbie, L. F. Rozsnyai, A. Noy, M. S. Wrighton, C. M. Lieber: Functional group imaging by chemical force microscopy, Science **265** (1994) 2071–2074

11.20 V. N. Koinkar, B. Bhushan: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy, J. Vac. Sci. Technol. A **14** (1996) 2378–2391

11.21 V. Scherer, B. Bhushan, U. Rabe, W. Arnold: Local elasticity and lubrication measurements using atomic force and friction force microscopy at ultrasonic frequencies, IEEE Trans. Mag. **33** (1997) 4077–4079

11.22 V. Scherer, W. Arnold, B. Bhushan: Lateral force microscopy using acoustic friction force microscopy, Surf. Interface Anal. **27** (1999) 578–587

11.23 B. Bhushan, S. Sundararajan: Micro/Nanoscale friction and wear mechanisms of thin films using atomic force and friction force microscopy, Acta Mater. **46** (1998) 3793–3804

11.24 U. Krotil, T. Stifter, H. Waschipky, K. Weishaupt, S. Hild, O. Marti: Pulse force mode: A new method for the investigation of surface properties, Surf. Interface Anal. **27** (1999) 336–340

11.25 B. Bhushan, C. Dandavate: Thin-film friction and adhesion studies using atomic force microscopy, J. Appl. Phys. **87** (2000) 1201–1210

11.26 B. Bhushan: *Micro/Nanotribology and its Applications* (Kluwer, Dordrecht 1997)

11.27 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

11.28 B. Bhushan: *Modern Tribology Handbook Vol. 1: Principles of Tribology* (CRC, Boca Raton 2001)

11.29 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

11.30 M. Reinstaedtler, U. Rabe, V. Scherer, U. Hartmann, A. Goldade, B. Bhushan, W. Arnold: On the nanoscale measurement of friction using atomic force microscope cantilever torsional resonances, Appl. Phys. Lett. **82** (2003) 2604–2606

11.31 N. A. Burnham, R. J. Colton: Measuring the nanomechanical properties and surface forces of materials using an atomic force microscope, J. Vac. Sci. Technol. A **7** (1989) 2906–2913

11.32 P. Maivald, H. J. Butt, S. A. C. Gould, C. B. Prater, B. Drake, J. A. Gurley, V. B. Elings, P. K. Hansma: Using force modulation to image surface elasticities with the atomic force microscope, Nanotechnol. **2** (1991) 103–106

11.33 B. Bhushan, A. V. Kulkarni, W. Bonin, J. T. Wyrobek: Nano/Picoindentation measurements using capacitive transducer in atomic force microscopy, Philos. Mag. A **74** (1996) 1117–1128

11.34 B. Bhushan, V. N. Koinkar: Nanoindentation hardness measurements using atomic force microscopy, Appl. Phys. Lett. **75** (1994) 5741–5746

11.35 D. DeVecchio, B. Bhushan: Localized surface elasticity measurements using an atomic force microscope, Rev. Sci. Instrum. **68** (1997) 4498–4505

11.36 S. Amelio, A. V. Goldade, U. Rabe, V. Scherer, B. Bhushan, W. Arnold: Measurements of mechanical properties of ultra-thin diamond-like carbon coatings using atomic force acoustic microscopy, Thin Solid Films **392** (2001) 75–84

11.37 D. M. Eigler, E. K. Schweizer: Positioning single atoms with a scanning tunnelling microscope, Nature **344** (1990) 524–528

11.38 A. L. Weisenhorn, J. E. MacDougall, J. A. C. Gould, S. D. Cox, W. S. Wise, J. Massie, P. Maivald, V. B. Elings, G. D. Stucky, P. K. Hansma: Imaging and manipulating of molecules on a zeolite surface with an atomic force microscope, Science **247** (1990) 1330–1333

11.39 I. W. Lyo, Ph. Avouris: Field-induced nanometer-to-atomic-scale manipulation of silicon surfaces with the STM, Science **253** (1991) 173–176

11.40 O. M. Leung, M. C. Goh: Orientation ordering of polymers by atomic force microscope tip-surface interactions, Science **225** (1992) 64–66

11.41 D. W. Abraham, H. J. Mamin, E. Ganz, J. Clark: Surface modification with the scanning tunneling microscope, IBM J. Res. Dev. **30** (1986) 492–499

11.42 R. M. Silver, E. E. Ehrichs, A. L. de Lozanne: Direct writing of submicron metallic features with a scanning tunnelling microscope, Appl. Phys. Lett. **51** (1987) 247–249

11.43 A. Kobayashi, F. Grey, R. S. Williams, M. Ano: Formation of nanometer-scale grooves in silicon with a scanning tunneling microscope, Science **259** (1993) 1724–1726

11.44 B. Parkinson: Layer-by-layer nanometer scale etching of two-dimensional substrates using the scanning tunneling microscopy, J. Am. Chem. Soc. **112** (1990) 7498–7502

11.45 A. Majumdar, P. I. Oden, J. P. Carrejo, L. A. Nagahara, J. J. Graham, J. Alexander: Nanometer-scale lithography using the atomic force microscope, Appl. Phys. Lett. **61** (1992) 2293–2295

11.46 B. Bhushan: Micro/Nanotribology and its applications to magnetic storage devices and MEMS, Tribol. Int. **28** (1995) 85–96

11.47 L. Tsau, D. Wang, K. L. Wang: Nanometer scale patterning of silicon(100) surface by an atomic force microscope operating in air, Appl. Phys. Lett. **64** (1994) 2133–2135

11.48 E. Delawski, B. A. Parkinson: Layer-by-layer etching of two-dimensional metal chalcogenides with the atomic force microscope, J. Am. Chem. Soc. **114** (1992) 1661–1667

11.49 B. Bhushan, G. S. Blackman: Atomic force microscopy of magnetic rigid disks and sliders and its applications to tribology, ASME J. Tribol. **113** (1991) 452–458

11.50 O. Marti, B. Drake, P. K. Hansma: Atomic force microscopy of liquid-covered surfaces: atomic resolution images, Appl. Phys. Lett. **51** (1987) 484–486

11.51 B. Drake, C. B. Prater, A. L. Weisenhorn, S. A. C. Gould, T. R. Albrecht, C. F. Quate, D. S. Cannell, H. G. Hansma, P. K. Hansma: Imaging crystals, polymers and processes in water with the atomic force microscope, Science **243** (1989) 1586–1589

11.52 M. Binggeli, R. Christoph, H. E. Hintermann, J. Colchero, O. Marti: Friction force measurements on potential controlled graphite in an electrolytic environment, Nanotechnol. **4** (1993) 59–63

11.53 G. Meyer, N. M. Amer: Novel optical approach to atomic force microscopy, Appl. Phys. Lett. **53** (1988) 1045–1047

11.54 J. H. Coombs, J. B. Pethica: Properties of vacuum tunneling currents: Anomalous barrier heights, IBM J. Res. Dev. **30** (1986) 455–459

11.55 M. D. Kirk, T. Albrecht, C. F. Quate: Low-temperature atomic force microscopy, Rev. Sci. Instrum. **59** (1988) 833–835

11.56 F. J. Giessibl, Ch. Gerber, G. Binnig: A low-temperature atomic force/scanning tunneling microscope for ultrahigh vacuum, J. Vac. Sci. Technol. B **9** (1991) 984–988

11.57 T. R. Albrecht, P. Grutter, D. Rugar, D. P. E. Smith: Low temperature force microscope with all-fiber interferometer, Ultramicroscopy **42–44** (1992) 1638–1646

11.58 H. J. Hug, A. Moser, Th. Jung, O. Fritz, A. Wadas, I. Parashikor, H. J. Güntherodt: Low temperature magnetic force microscopy, Rev. Sci. Instrum. **64** (1993) 2920–2925

11.59 C. Basire, D. A. Ivanov: Evolution of the lamellar structure during crystallization of a semicrystalline-amorphous polymer blend: Time-resolved hot-stage SPM study, Phys. Rev. Lett. **85** (2000) 5587–5590

11.60 H. Liu, B. Bhushan: Investigation of nanotribological properties of self-assembled monolayers with alkyl and biphenyl spacer chains, Ultramicroscopy **91** (2002) 185–202

11.61 J. Foster, J. Frommer: Imaging of liquid crystal using a tunneling microscope, Nature **333** (1988) 542–547

11.62 D. Smith, H. Horber, C. Gerber, G. Binnig: Smectic liquid crystal monolayers on graphite observed by scanning tunneling microscopy, Science **245** (1989) 43–45

11.63 D. Smith, J. Horber, G. Binnig, H. Nejoh: Structure, registry and imaging mechanism of alkylcyanobiphenyl molecules by tunnelling microscopy, Nature **344** (1990) 641–644

11.64 Y. Andoh, S. Oguchi, R. Kaneko, T. Miyamoto: Evaluation of very thin lubricant films, J. Phys. D **25** (1992) A71–A75

11.65 Y. Martin, C. C. Williams, H. K. Wickramasinghe: Atomic force microscope-force mapping and profiling on a sub 100-A scale, J. Appl. Phys. **61** (1987) 4723–4729

11.66 J. E. Stern, B. D. Terris, H. J. Mamin, D. Rugar: Deposition and imaging of localized charge on insulator surfaces using a force microscope, Appl. Phys. Lett. **53** (1988) 2717–2719

11.67 K. Yamanaka, H. Ogisco, O. Kolosov: Ultrasonic force microscopy for nanometer resolution subsurface imaging, Appl. Phys. Lett. **64** (1994) 178–180

11.68 K. Yamanaka, E. Tomita: Lateral force modulation atomic force microscope for selective imaging of friction forces, Jpn. J. Appl. Phys. **34** (1995) 2879–2882

11.69 U. Rabe, K. Janser, W. Arnold: Vibrations of free and surface-coupled atomic force microscope: Theory and experiment, Rev. Sci. Instrum. **67** (1996) 3281–3293

11.70 Y. Martin, H. K. Wickramasinghe: Magnetic imaging by force microscopy with 1000 Å resolution, Appl. Phys. Lett. **50** (1987) 1455–1457

11.71 D. Rugar, H. J. Mamin, P. Guethner, S. E. Lambert, J. E. Stern, I. McFadyen, T. Yogi: Magnetic force microscopy – General principles and application to longitudinal recording media, J. Appl. Phys. **63** (1990) 1169–1183

11.72 C. Schoenenberger, S. F. Alvarado: Understanding magnetic force microscopy, Z. Phys. B **80** (1990) 373–383

11.73 U. Hartmann: Magnetic force microscopy, Annu. Rev. Mater. Sci. **29** (1999) 53–87

11.74 D. W. Pohl, W. Denk, M. Lanz: Optical stethoscopy-image recording with resolution lambda/20, Appl. Phys. Lett. **44** (1984) 651–653

11.75 E. Betzig, J. K. Troutman, T. D. Harris, J. S. Weiner, R. L. Kostelak: Breaking the diffraction barrier – optical microscopy on a nanometric scale, Science **251** (1991) 1468–1470

11.76 E. Betzig, P. L. Finn, J. S. Weiner: Combined shear force and near-field scanning optical microscopy, Appl. Phys. Lett. **60** (1992) 2484

11.77 P. F. Barbara, D. M. Adams, D. B. O'Connor: Characterization of organic thin film materials with near-field scanning optical microscopy (NSOM), Annu. Rev. Mater. Sci. **29** (1999) 433–469

11.78 C. C. Williams, H. K. Wickramasinghe: Scanning thermal profiler, Appl. Phys. Lett. **49** (1986) 1587–1589

11.79 C. C. Williams, H. K. Wickramasinghe: Microscopy of chemical-potential variations on an atomic scale, Nature **344** (1990) 317–319

11.80 A. Majumdar: Scanning thermal microscopy, Annu. Rev. Mater. Sci. **29** (1999) 505–585

11.81 O. E. Husser, D. H. Craston, A. J. Bard: Scanning electrochemical microscopy – high resolution deposition and etching of materials, J. Electrochem. Soc. **136** (1989) 3222–3229

11.82 Y. Martin, D. W. Abraham, H. K. Wickramasinghe: High-resolution capacitance measurement and potentiometry by force microscopy, Appl. Phys. Lett. **52** (1988) 1103–1105

11.83 M. Nonnenmacher, M. P. O'Boyle, H. K. Wickramasinghe: Kelvin probe force microscopy, Appl. Phys. Lett. **58** (1991) 2921–2923

11.84 J. M. R. Weaver, D. W. Abraham: High resolution atomic force microscopy potentiometry, J. Vac. Sci. Technol. B **9** (1991) 1559–1561

11.85 D. DeVecchio, B. Bhushan: Use of a nanoscale Kelvin probe for detecting wear precursors, Rev. Sci. Instrum. **69** (1998) 3618–3624

11.86 B. Bhushan, A. V. Goldade: Measurements and analysis of surface potential change during wear of single-crystal silicon (100) at ultralow loads using Kelvin probe microscopy, Appl. Surf. Sci. **157** (2000) 373–381

11.87 P. K. Hansma, B. Drake, O. Marti, S. A. C. Gould, C. B. Prater: The scanning ion-conductance microscope, Science **243** (1989) 641–643

11.88 C. B. Prater, P. K. Hansma, M. Tortonese, C. F. Quate: Improved scanning ion-conductance microscope using microfabricated probes, Rev. Sci. Instrum. **62** (1991) 2634–2638

11.89 J. Matey, J. Blanc: Scanning capacitance microscopy, J. Appl. Phys. **57** (1985) 1437–1444

11.90 C. C. Williams: Two-dimensional dopant profiling by scanning capacitance microscopy, Annu. Rev. Mater. Sci. **29** (1999) 471–504

11.91 D. T. Lee, J. P. Pelz, B. Bhushan: Instrumentation for direct, low frequency scanning capacitance microscopy, and analysis of position dependent stray capacitance, Rev. Sci. Instrum. **73** (2002) 3523–3533

11.92 I. Giaever: Energy gap in superconductors measured by electron tunneling, Phys. Rev. Lett. **5** (1960) 147–148

11.93 P. K. Hansma, J. Tersoff: Scanning tunneling microscopy, J. Appl. Phys. **61** (1987) R1–R23

11.94 D. Sarid, V. Elings: Review of scanning force microscopy, J. Vac. Sci. Technol. B **9** (1991) 431–437

11.95 U. Durig, O. Zuger, A. Stalder: Interaction force detection in scanning probe microscopy: Methods and applications, J. Appl. Phys. **72** (1992) 1778–1797

11.96 J. Frommer: Scanning tunneling microscopy and atomic force microscopy in organic chemistry, Angew. Chem. Int. Ed. Engl. **31** (1992) 1298–1328

11.97 H. J. Güntherodt, R. Wiesendanger (Eds.): *Scanning Tunneling Microscopy I: General Principles and Applications to Clean and Adsorbate-Covered Surfaces* (Springer, Berlin, Heidelberg 1992)

11.98 R. Wiesendanger, H. J. Güntherodt (Eds.): *Scanning Tunneling Microscopy, II: Further Applications and Related Scanning Techniques* (Springer, Berlin, Heidelberg 1992)

11.99 D. A. Bonnell (Ed.): *Scanning Tunneling Microscopy and Spectroscopy – Theory, Techniques, and Applications* (VCH, New York 1993)

11.100 O. Marti, M. Amrein (Eds.): *STM and SFM in Biology* (Academic, San Diego 1993)

11.101 J. A. Stroscio, W. J. Kaiser (Eds.): *Scanning Tunneling Microscopy* (Academic, Boston 1993)

11.102 H. J. Güntherodt, D. Anselmetti, E. Meyer (Eds.): *Forces in Scanning Probe Methods* (Kluwer, Dordrecht 1995)

11.103 G. Binnig, H. Rohrer: Scanning tunnelling microscopy, Surf. Sci. **126** (1983) 236–244

11.104 B. Bhushan, J. Ruan, B. K. Gupta: A scanning tunnelling microscopy study of fullerene films, J. Phys. D **26** (1993) 1319–1322

11.105 R. L. Nicolaides, W. E. Yong, W. F. Packard, H. A. Zhou: Scanning tunneling microscope tip structures, J. Vac. Sci. Technol. A **6** (1988) 445–447

11.106 J. P. Ibe, P. P. Bey, S. L. Brandon, R. A. Brizzolara, N. A. Burnham, D. P. DiLella, K. P. Lee, C. R. K. Marrian, R. J. Colton: On the electrochemical etching of tips for scanning tunneling microscopy, J. Vac. Sci. Technol. A **8** (1990) 3570–3575

11.107 R. Kaneko, S. Oguchi: Ion-implanted diamond tip for a scanning tunneling microscope, Jpn. J. Appl. Phys. **28** (1990) 1854–1855

11.108 F. J. Giessibl: Atomic resolution of the silicon(111)-(7×7) surface by atomic force microscopy, Science **267** (1995) 68–71

11.109 B. Anczykowski, D. Krueger, K. L. Babcock, H. Fuchs: Basic properties of dynamic force spectroscopy with the scanning force microscope in experiment and simulation, Ultramicroscopy **66** (1996) 251–259

11.110 T. R. Albrecht and C. F. Quate: Atomic resolution imaging of a nonconductor by atomic force microscopy, J. Appl. Phys. **62** (1987) 2599–2602

11.111 S. Alexander, L. Hellemans, O. Marti, J. Schneir, V. Elings, P. K. Hansma: An atomic-resolution atomic-force microscope implemented using an optical lever, J. Appl. Phys. **65** (1989) 164–167

11.112 G. Meyer, N. M. Amer: Optical-beam-deflection atomic force microscopy: The NaCl(001) surface, Appl. Phys. Lett. **56** (1990) 2100–2101

11.113 A. L. Weisenhorn, M. Egger, F. Ohnesorge, S. A. C. Gould, S. P. Heyn, H. G. Hansma, R. L. Sinsheimer, H. E. Gaub, P. K. Hansma: Molecular resolution images of Langmuir–Blodgett films and DNA by atomic force microscopy, Langmuir **7** (1991) 8–12

11.114 J. Ruan, B. Bhushan: Atomic-scale and microscale friction of graphite and diamond using friction force microscopy, J. Appl. Phys. **76** (1994) 5022–5035

11.115 D. Rugar, P. K. Hansma: Atomic force microscopy, Phys. Today **43** (1990) 23–30

11.116 D. Sarid: *Scanning Force Microscopy* (Oxford Univ. Press, Oxford 1991)

11.117 G. Binnig: Force microscopy, Ultramicroscopy **42–44** (1992) 7–15

11.118 E. Meyer: Atomic force microscopy, Surf. Sci. **41** (1992) 3–49

11.119 H. K. Wickramasinghe: Progress in scanning probe microscopy, Acta Mater. **48** (2000) 347–358

11.120 A. J. den Boef: The influence of lateral forces in scanning force microscopy, Rev. Sci. Instrum. **62** (1991) 88–92

11.121 M. Radmacher, R. W. Tillman, M. Fritz, H. E. Gaub: From molecules to cells: Imaging soft samples with the atomic force microscope, Science **257** (1992) 1900–1905

11.122 F. Ohnesorge, G. Binnig: True atomic resolution by atomic force microscopy through repulsive and attractive forces, Science **260** (1993) 1451–1456

11.123 G. Neubauer, S. R. Coben, G. M. McClelland, D. Horne, C. M. Mate: Force microscopy with a bidirectional capacitance sensor, Rev. Sci. Instrum. **61** (1990) 2296–2308

11.124 T. Goddenhenrich, H. Lemke, U. Hartmann, C. Heiden: Force microscope with capacitive displacement detection, J. Vac. Sci. Technol. A **8** (1990) 383–387

11.125 U. Stahl, C. W. Yuan, A. L. Delozanne, M. Tortonese: Atomic force microscope using piezoresistive cantilevers and combined with a scanning electron microscope, Appl. Phys. Lett. **65** (1994) 2878–2880

11.126 R. Kassing, E. Oesterschulze: Sensors for scanning probe microscopy. In: *Micro/Nanotribology and Its Applications*, ed. by B. Bhushan (Kluwer, Dordrecht 1997) pp. 35–54

11.127 C. M. Mate: Atomic-force-microscope study of polymer lubricants on silicon surfaces, Phys. Rev. Lett. **68** (1992) 3323–3326

11.128 S. P. Jarvis, A. Oral, T. P. Weihs, J. B. Pethica: A novel force microscope and point contact probe, Rev. Sci. Instrum. **64** (1993) 3515–3520

11.129 D. Rugar, H. J. Mamin, P. Guethner: Improved fiber-optical interferometer for atomic force microscopy, Appl. Phys. Lett. **55** (1989) 2588–2590

11.130 C. Schoenenberger, S. F. Alvarado: A differential interferometer for force microscopy, Rev. Sci. Instrum. **60** (1989) 3131–3135

11.131 D. Sarid, D. Iams, V. Weissenberger, L. S. Bell: Compact scanning-force microscope using laser diode, Opt. Lett. **13** (1988) 1057–1059

11.132 N. W. Ashcroft, N. D. Mermin: *Solid State Physics* (Holt Reinhart and Winston, New York 1976)

11.133 G. Binnig, D. P. E. Smith: Single-tube three-dimensional scanner for scanning tunneling microscopy, Rev. Sci. Instrum. **57** (1986) 1688

11.134 S. I. Park, C. F. Quate: Digital filtering of STM images, J. Appl. Phys. **62** (1987) 312

11.135 J. W. Cooley, J. W. Tukey: An algorithm for machine calculation of complex fourier series, Math. Computation **19** (1965) 297

11.136 J. Ruan, B. Bhushan: Atomic-scale friction measurements using friction force microscopy: Part I – General principles and new measurement techniques, ASME J. Tribol. **116** (1994) 378–388

11.137 T. R. Albrecht, S. Akamine, T. E. Carver, C. F. Quate: Microfabrication of cantilever styli for the atomic force microscope, J. Vac. Sci. Technol. A **8** (1990) 3386–3396

11.138 O. Marti, S. Gould, P. K. Hansma: Control electronics for atomic force microscopy, Rev. Sci. Instrum. **59** (1988) 836–839

11.139 O. Wolter, T. Bayer, J. Greschner: Micromachined silicon sensors for scanning force microscopy, J. Vac. Sci. Technol. B **9** (1991) 1353–1357

11.140 E. Meyer, R. Overney, R. Luthi, D. Brodbeck: Friction force microscopy of mixed Langmuir–Blodgett films, Thin Solid Films **220** (1992) 132–137

11.141 H. J. Dai, J. H. Hafner, A. G. Rinzler, D. T. Colbert, R. E. Smalley: Nanotubes as nanoprobes in scanning probe microscopy, Nature **384** (1996) 147–150

11.142 J. H. Hafner, C. L. Cheung, A. T. Woolley, C. M. Lieber: Structural and functional imaging with carbon nanotube AFM probes, Prog. Biophys. Mol. Biol. **77** (2001) 73–110

11.143 G. S. Blackman, C. M. Mate, M. R. Philpott: Interaction forces of a sharp tungsten tip with molecular films on silicon surface, Phys. Rev. Lett. **65** (1990) 2270–2273

11.144 S. J. O'Shea, M. E. Welland, T. Rayment: Atomic force microscope study of boundary layer lubrication, Appl. Phys. Lett. **61** (1992) 2240–2242

11.145 J. P. Cleveland, S. Manne, D. Bocek, P. K. Hansma: A nondestructive method for determining the spring constant of cantilevers for scanning force microscopy, Rev. Sci. Instrum. **64** (1993) 403–405

11.146 D. W. Pohl: Some design criteria in STM, IBM J. Res. Dev. **30** (1986) 417

11.147 W. T. Thomson, M. D. Dahleh: *Theory of Vibration with Applications*, 5th edn. (Prentice Hall, Upper Saddle River 1998)

11.148 J. Colchero: Reibungskraftmikroskopie. Ph.D. Thesis (University of Konstanz, Konstanz 1993)

11.149 G. M. McClelland, R. Erlandsson, S. Chiang: Atomic force microscopy: General principles and a new implementation. In: *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 6B, ed. by D. O. Thompson, D. E. Chimenti (Plenum, New York 1987) pp. 1307–1314

11.150 Y. R. Shen: *The Principles of Nonlinear Optics* (Wiley, New York 1984)

11.151 T. Baumeister, S. L. Marks: *Standard Handbook for Mechanical Engineers*, 7th edn. (McGraw-Hill, New York 1967)

11.152 J. Colchero, O. Marti, H. Bielefeldt, J. Mlynek: Scanning force and friction microscopy, Phys. Stat. Sol. **131** (1991) 73–75

11.153 R. Young, J. Ward, F. Scire: Observation of metal-vacuum-metal tunneling, field emission, and the transition region, Phys. Rev. Lett. **27** (1971) 922

11.154 R. Young, J. Ward, F. Scire: The topographiner: An instrument for measuring surface microtopography, Rev. Sci. Instrum. **43** (1972) 999

11.155 C. Gerber, O. Marti: Magnetostrictive positioner, IBM Tech. Disclosure Bull. **27** (1985) 6373

11.156 R. Garcìa Cantù, M. A. Huerta Garnica: Long-scan imaging by STM, J. Vac. Sci. Technol. A **8** (1990) 354
11.157 C. J. Chen: In situ testing and calibration of tube piezoelectric scanners, Ultramicroscopy **42–44** (1992) 1653–1658
11.158 R. G. Carr: J. Microscopy **152** (1988) 379
11.159 C. J. Chen: Electromechanical deflections of piezoelectric tubes with quartered electrodes, Appl. Phys. Lett. **60** (1992) 132
11.160 N. Libioulle, A. Ronda, M. Taborelli, J. M. Gilles: Deformations and nonlinearity in scanning tunneling microscope images, J. Vac. Sci. Technol. B **9** (1991) 655–658
11.161 E. P. Stoll: Restoration of STM images distorted by time-dependent piezo driver aftereffects, Ultramicroscopy **42–44** (1991) 1585–1589
11.162 R. Durselen, U. Grunewald, W. Preuss: Calibration and applications of a high precision piezo scanner for nanometrology, Scanning **17** (1995) 91–96
11.163 J. Fu: In situ testing and calibrating of Z-piezo of an atomic force microscope, Rev. Sci. Instrum. **66** (1995) 3785–3788
11.164 R. C. Barrett, C. F. Quate: Optical scan-correction system applied to atomic force microscopy, Rev. Sci. Instrum. **62** (1991) 1393
11.165 R. Toledo-Crow, P. C. Yang, Y. Chen, M. Vaez-Iravani: Near-field differential scanning optical microscope with atomic force regulation, Appl. Phys. Lett. **60** (1992) 2957–2959
11.166 J. E. Griffith, G. L. Miller, C. A. Green: A scanning tunneling microscope with a capacitance-based position monitor, J. Vac. Sci. Technol. B **8** (1990) 2023–2027
11.167 A. E. Holman, C. D. Laman, P. M. L. O. Scholte, W. C. Heerens, F. Tuinstra: A calibrated scanning tunneling microscope equipped with capacitive sensors, Rev. Sci. Instrum. **67** (1996) 2274–2280

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Probes in Sca

12. Probes in Scanning Microscopies

12.1 **Atomic Force Microscopy** 372
  12.1.1 Principles of Operation 372
  12.1.2 Standard Probe Tips..................... 373
  12.1.3 Probe Tip Performance 374
  12.1.4 Oxide-Sharpened Tips.................. 375
  12.1.5 FIB tips 376
  12.1.6 EBD tips 376
  12.1.7 Carbon Nanotube Tips.................. 376

12.2 **Scanning Tunneling Microscopy** 382
  12.2.1 Mechanically Cut STM Tips............. 382
  12.2.2 Electrochemically Etched STM Tips 383

References 383

Scanning probe microscopy (SPM) provides nanometer-scale mapping of numerous sample properties in essentially any environment. This unique combination of high resolution and broad applicability has lead to the application of SPM to many areas of science and technology, especially those interested in the structure and properties of materials at the nanometer scale. SPM images are generated through measurements of a tip-sample interaction. A well-characterized tip is the key element to data interpretation and is typically the limiting factor.

Commercially available atomic force microscopy (AFM) tips, integrated with force sensing cantilevers, are microfabricated from silicon and silicon nitride by lithographic and anisotropic etching techniques. The performance of these tips can be characterized by imaging nanometer-scale standards of known dimension, and the resolution is found to roughly correspond to the tip radius of curvature, the tip aspect ratio, and the sample height. Although silicon and silicon nitride tips have a somewhat large radius of curvature, low aspect ratio, and limited lifetime due to wear, the widespread use of AFM today is due in large part to the broad availability of these tips. In some special cases, small asperities on the tip can provide resolution much higher than the tip radius of curvature for low-Z samples such as crystal surfaces and ordered protein arrays.

Several strategies have been developed to improve AFM tip performance. Oxide sharpening improves tip sharpness and enhances tip asperities. For high-aspect-ratio samples such as integrated circuits, silicon AFM tips can be modified by focused ion beam (FIB) milling. FIB tips reach three-degree cone angles over lengths of several microns and can be fabricated at arbitrary angles. Other high resolution and high-aspect-ratio tips are produced by electron beam deposition (EBD) in which a carbon spike is deposited onto the tip apex from the background gases in an electron microscope. Finally, carbon nanotubes have been employed as AFM tips. Their nanometer-scale diameter, long length, high stiffness, and elastic buckling properties make carbon nanotubes possibly the ultimate tip material for AFM. Nanotubes can be manually attached to silicon or silicon nitride AFM tips or "grown" onto tips by chemical vapor deposition (CVD), which should soon make them widely available. In scanning tunneling microscopy (STM), the electron tunneling signal decays exponentially with tip-sample separation, so that in principle only the last few atoms contribute to the signal. STM tips are, therefore, not as sensitive to the nanoscale tip geometry and can be made by simple mechanical cutting or electrochemical etching of metal wires. In choosing tip materials, one prefers hard, stiff metals that will not oxidize or corrode in the imaging environment.

In scanning probe microscopy (SPM), an image is created by raster scanning a sharp probe tip over a sample and measuring some highly localized tip-sample interaction as a function of position. SPMs are based on several interactions, the major types including scanning tunneling microscopy (STM), which measures an electronic tunneling current; atomic force microscopy (AFM), which measures force interactions; and near-field scan-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

ning optical microscopy (NSOM), which measures local optical properties by exploiting near-field effects (Fig. 12.1). These methods allow the characterization of many properties (structural, mechanical, electronic, optical) on essentially any material (metals, semiconductors, insulators, biomolecules) and in essentially any environment (vacuum, liquid, or ambient air conditions). The unique combination of nanoscale resolution, previously the domain of electron microscopy, *and broad applicability* has led to the proliferation of SPM into virtually all areas of nanometer-scale science and technology.

Several enabling technologies have been developed for SPM, or borrowed from other techniques. Piezoelectric tube scanners allow accurate, sub-angstrom positioning of the tip or sample in three dimensions. Optical deflection systems and microfabricated cantilevers can detect forces in AFM down to the picoNewton range. Sensitive electronics can measure STM currents less than 1 picoamp. High transmission fiber optics and sensitive photodetectors can manipulate and detect small optical signals of NSOM. Environmental control has been developed to allow SPM imaging in UHV, cryogenic temperatures, at elevated temperatures, and in fluids. Vibration and drift have been controlled such that a probe tip can be held over a single molecule for hours of observation. Microfabrication techniques have been developed for the mass production of probe tips, making SPMs commercially available and allowing the development of many new SPM modes and combinations with other characterization methods. However, of all this SPM development over the past 20 years, what has received the least attention is perhaps the most important aspect: the probe tip.

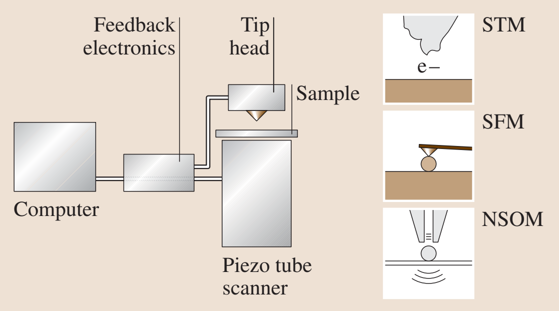

Fig. 12.1 A schematic of the components of a scanning probe microscope and the three types of signals observed: STM senses electron tunneling currents, AFM measures forces, and NSOM measures near-field optical properties via a sub-wavelength aperture

Interactions measured in SPMs occur at the tip-sample interface, which can range in size from a single atom to tens of nanometers. The size, shape, surface chemistry, electronic and mechanical properties of the tip apex will directly influence the data signal and the interpretation of the image. Clearly, the better characterized the tip the more useful the image information. In this chapter, the fabrication and performance of AFM and STM probes will be described.

12.1 Atomic Force Microscopy

AFM is the most widely used form of SPM, since it requires neither an electrically conductive sample, as in STM, nor an optically transparent sample or substrate, as in most NSOMs. Basic AFM modes measure the topography of a sample with the only requirement being that the sample is deposited on a flat surface and rigid enough to withstand imaging. Since AFM can measure a variety of forces, including van der Waals forces, electrostatic forces, magnetic forces, adhesion forces and friction forces, specialized modes of AFM can characterize the electrical, mechanical, and chemical properties of a sample in addition to its topography.

12.1.1 Principles of Operation

In AFM, a probe tip is integrated with a microfabricated force-sensing cantilever. A variety of silicon and silicon nitride cantilevers are commercially available with micron-scale dimensions, spring constants ranging from 0.01 to 100 N/m, and resonant frequencies ranging from 5 kHz to over 300 kHz. The cantilever deflection is detected by optical beam deflection, as illustrated in Fig. 12.2. A laser beam bounces off the back of the cantilever and is centered on a split photodiode. Cantilever deflections areproportional to the difference signal $V_A - V_B$. Sub-angstrom deflections can be deflected and, therefore, forces down to tens of picoNewtons can be measured. A more recently developed method of cantilever deflection measurement is through a piezoelectric layer on the cantilever that registers a voltage upon deflection [12.1].

A piezoelectric scanner rasters the sample under the tip while the forces are measured through deflections of the cantilever. To achieve more controlled imaging con-

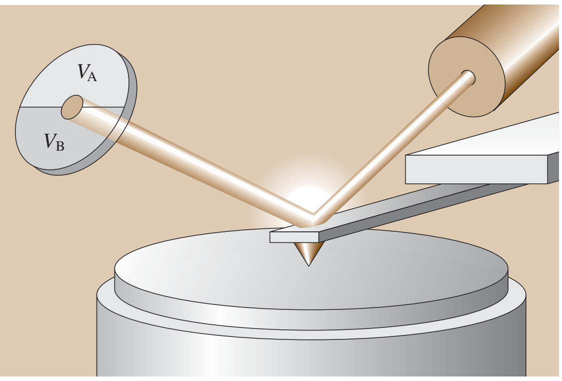

Fig. 12.2 An illustration of the optical beam deflection system that detects cantilever motion in the AFM. The voltage signal $V_A - V_B$ is proportional to the deflection

ditions, a feedback loop monitors the tip-sample force and adjusts the sample Z-position to hold the force constant. The topographic image of the sample is then taken from the sample Z-position data. The mode described is called contact mode, in which the tip is deflected by the sample due to repulsive forces, or "contact". It is generally only used for flat samples that can withstand lateral forces during scanning. To minimize lateral forces and sample damage, two AC modes have been developed. In these, the cantilever is driven into AC oscillation near its resonant frequency (tens to hundreds of kHz) with amplitudes of 5 to tens of s. When the tip approaches the sample, the oscillation is damped, and the reduced amplitude is the feedback signal, rather than the DC deflection. Again, topography is taken from the varying Z-position of the sample required to keep the tip oscillation amplitude constant. The two AC modes differ only in the nature of the interaction. In intermittent contact mode, also called tapping mode, the tip contacts the sample on each cycle, so the amplitude is reduced by ionic repulsion as in contact mode. In non-contact mode, long-range van der Waals forces reduce the amplitude by effectively shifting the spring constant experienced by the tip and changing its resonant frequency.

12.1.2 Standard Probe Tips

In early AFM work, cantilevers were made by hand from thin metal foils or small metal wires. Tips were created by gluing diamond fragments to the foil cantilevers or electrochemically etching the wires to a sharp point. Since these methods were labor intensive and not highly reproducible, they were not amenable to large-scale production. To address this problem, and the need for smaller cantilevers with higher resonant frequencies, batch fabrication techniques were developed (see Fig. 12.3). Building on existing methods to batch fabricate Si_3N_4 cantilevers, *Albrecht* et al. [12.2] etched an array of small square openings in an SiO_2 mask layer over a (100) silicon surface. The exposed square (100) regions were etched with KOH, an anisotropic etchant that terminates at the (111) planes, thus creating pyramidal etch pits in the silicon surface. The etch pit mask was then removed and another was applied to define the cantilever shapes with the pyramidal etch pits at the end. The Si wafer was then coated with a low stress Si_3N_4 layer by LPCVD. The Si_3N_4 fills the etch pit, using it as a mold to create a pyramidal tip. The silicon was later removed by etching to free the cantilevers and tips. Further steps resulting in the attachment of the cantilever to a macroscopic piece of glass are not described here. The resulting pyramidal tips were highly symmetric and had a tip radius of less than 30 nm, as determined by scanning electron microscopy (SEM). This procedure has likely not changed significantly, since commercially available Si_3N_4 tips are still specified to have a curvature radius of 30 nm.

Wolter et al. [12.3] developed methods to batch fabricate single-crystal Si cantilevers with integrated tips. Microfabricated Si cantilevers were first prepared using previously described methods, and a small mask was formed at the end of the cantilever. The Si around the mask was etched by KOH, so that the mask was under cut. This resulted in a pyramidal silicon tip under the mask, which was then removed. Again, this partial description of the full procedure only describes tip fabrication. With some refinements the silicon tips were made

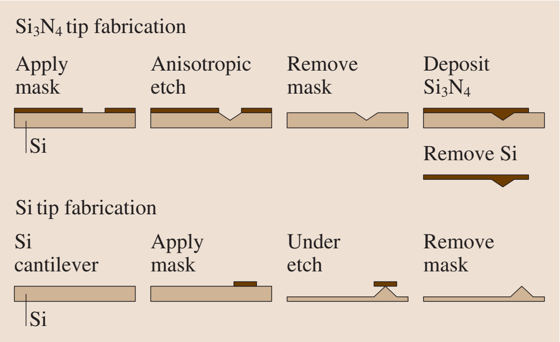

Fig. 12.3 A schematic overview of the fabrication of Si and Si_3N_4 tip fabrication as described in the text

in high yield with curvature radii of less than 10 nm. Si tips are sharper than Si_3N_4 tips, because they are directly formed by the anisotropic etch in single-crystal Si, rather than using an etch pit as a mask for deposited material. Commercially available silicon probes are made by similar refined techniques and provide a curvature typical radius of < 10 nm.

12.1.3 Probe Tip Performance

In atomic force microscopy the question of resolution can be a rather complicated issue. As an initial approximation, resolution is often considered strictly in geometrical terms that assume rigid tip-sample contact. The topographical image of a feature is broadened or narrowed by the size of the probe tip, so the resolution is approximately the width of the tip. Therefore, the resolution of AFM with standard commercially available tips is on the order of 5 to 10 nm. *Bustamante* and *Keller* [12.4] carried the geometrical model further by drawing an analogy to resolution in optical systems. Consider two sharp spikes separated by a distance d to be point objects imaged by AFM (see Fig. 12.4). Assume the tip has a parabolic shape with an end radius R. The tip-broadened image of these spikes will appear as inverted parabolas. There will be a small depression between the images of depth Δz. The two spikes are considered "resolved" if Δz is larger than the instrumental noise in the z direction. Defined in this manner, the resolution d, the minimum separation at which the spikes are resolved, is

$$d = 2\sqrt{2R\,(\Delta z)}\,, \tag{12.1}$$

where one must enter a minimal detectable depression for the instrument (Δz) to determine the resolution. So for a silicon tip with radius 5 nm and a minimum detectable Δz of 0.5 nm, the resolution is about 4.5 nm. However, the above model assumes the spikes are of equal height. *Bustamante* and *Keller* [12.4] went on to point out that if the height of the spikes is not equal, the resolution will be affected. Assuming a height difference of Δh, the resolution becomes:

$$d = \sqrt{2R}\left(\sqrt{\Delta z} + \sqrt{\Delta z + \Delta h}\right)\,. \tag{12.2}$$

For a pair of spikes with a 2 nm height difference, the resolution drops to 7.2 nm for a 5 nm tip and 0.5 nm minimum detectable Δz. While geometrical considerations are a good starting point for defining resolution, they ignore factors such as the possible compression and deformation of the tip and sample. *Vesenka* et al. [12.5] confirmed a similar geometrical resolution model by imaging monodisperse gold nanoparticles with tips characterized by transmission electron microscopy (TEM).

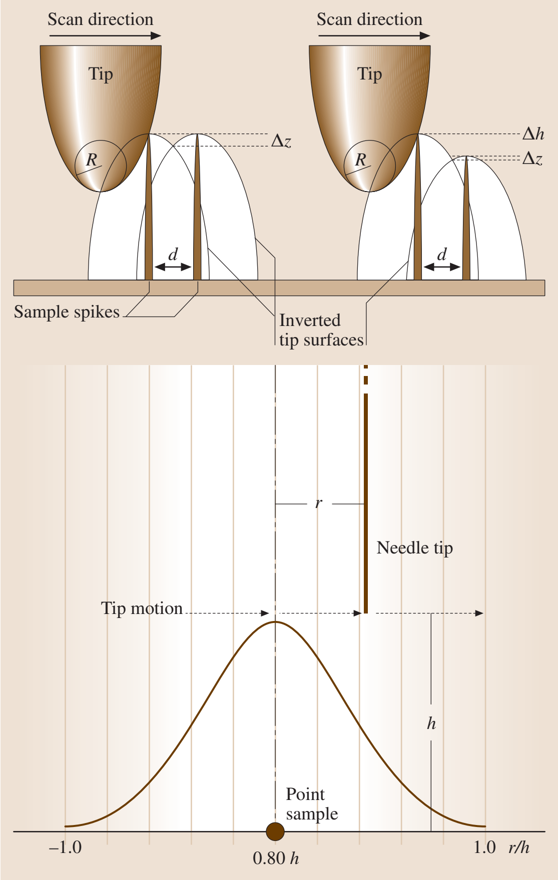

Fig. 12.4 The factors that determine AFM imaging resolution in contact mode (*top*) and noncontact mode (*bottom*), adapted from [12.4]

Noncontact AFM contrast is generated by long-range interactions such as van der Waals forces, so resolution will not simply be determined by geometry because the tip and sample are not in rigid contact. *Bustamante* and *Keller* [12.4] have derived an expression for the resolution in noncontact AFM for an idealized, infinitely thin "line" tip and a point particle as the sample (Fig. 12.4). Noncontact AFM is sensitive to the gradient of long-range forces, so the van der Waals force gradient was calculated as a function of position for the tip at height h above the surface. If the resolution d is defined as the full width at half maximum of this curve, the resolution is:

$$d = 0.8h\ . \tag{12.3}$$

This shows that even for an ideal geometry, the resolution is fundamentally limited in noncontact mode by the tip-sample separation. Under UHV conditions, the tip-sample separation can be made very small, so atomic resolution is possible on flat, crystalline surfaces. Under ambient conditions, however, the separation must be larger to keep the tip from being trapped in the ambient water layer on the surface. This larger separation can lead to a point where further improvements in tip sharpness do not improve resolution. It has been found that imaging 5 nm gold nanoparticles in noncontact mode with carbon nanotube tips of 2 nm diameter leads to particle widths of 12 nm, larger than the 7 nm width one would expect assuming rigid contact [12.8]. However, in tapping mode operation, the geometrical definition of resolution is relevant, since the tip and sample come into rigid contact. When imaging 5 nm gold particles with 2 nm carbon nanotube tips in tapping mode, the expected 7 nm particle width is obtained [12.9].

The above descriptions of AFM resolution cannot explain the sub-nanometer resolution achieved on crystal surfaces [12.10] and ordered arrays of biomolecules [12.11] in contact mode with commercially available probe tips. Such tips have nominal radii of curvature ranging from 5 nm to 30 nm, an order of magnitude larger than the resolution achieved. A detailed model to explain the high resolution on ordered membrane proteins has been put forth by [12.6]. In this model, the larger part of the silicon nitride tip apex balances the tip-sample interaction through electrostatic forces, while a very small tip asperity interacts with the sample to provide contrast (see Fig. 12.5). This model is supported by measurements at varying salt concentrations to vary the electrostatic interaction strength and the observation of defects in the ordered samples. However, the existence of such asperities has never been confirmed by independent electron microscopy images of the tip. Another model, considered especially applicable to atomic resolution on crystal surfaces, assumes the tip is in contact with a region of the sample much larger than the resolution observed, and that force components matching the periodicity of the sample are transmitted to the tip, resulting in an "averaged" image of the periodic lattice. Regardless of the mechanism, the structures determined are accurate and make this a highly valuable method for membrane proteins. However, this level of resolution should not be expected for most biological systems.

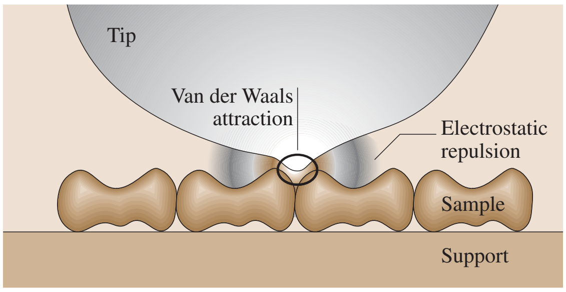

Fig. 12.5 A tip model to explain the high resolution obtained on ordered samples in contact mode, from [12.6]

12.1.4 Oxide-Sharpened Tips

Both Si and Si_3N_4 tips with increased aspect ratio and reduced tip radius can be fabricated through oxide sharpening of the tip. If a pyramidal or cone-shaped silicon tip is thermally oxidized to SiO_2 at low temperature ($< 1{,}050\,°C$), Si-SiO_2 stress formation reduces the oxidation rate at regions of high curvature. The result is a sharper, higher-aspect-ratio cone of silicon at the high curvature tip apex inside the outer pyramidal layer of SiO_2 (see Fig. 12.6). Etching the SiO_2 layer with HF then leaves tips with aspect ratios up to 10:1 and radii down to 1 nm [12.7], although 5–10 nm is the nominal specification for most commercially available tips. This oxide sharpening technique can also be applied to Si_3N_4 tips by oxidizing the silicon etch pits that are used as molds. As with tip fabrication, oxide sharpening is not quite as effective for Si_3N_4. Si_3N_4 tips were reported to have an

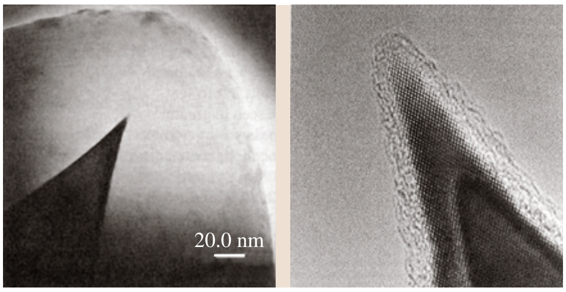

Fig. 12.6 Oxide sharpening of silicon tips. The *left image* shows a sharpened core of silicon in an outer layer of SiO_2. The *right image* is a higher magnification view of such a tip after the SiO_2 is removed. Adapted from [12.7]

11 nm radius of curvature [12.12], while commercially available oxide-sharpened Si_3N_4 tips have a nominal radius of < 20 nm.

12.1.5 FIB tips

A common AFM application in integrated circuit manufacture and MEMs is to image structures with very steep sidewalls such as trenches. To accurately image these features, one must consider the micron-scale tip structure, rather than the nanometer-scale structure of the tip apex. Since tip fabrication processes rely on anisotropic etchants, the cone half-angles of pyramidal tips are approximately 20 degrees. Images of deep trenches taken with such tips display slanted sidewalls and may not reach the bottom of the trench due to the tip broadening effects. To image such samples more faithfully, high-aspect-ratio tips are fabricated by focused ion beam (FIB) machining a Si tip to produce a sharp spike at the tip apex. Commercially available FIB tips have half cone angles of < 3 degrees over lengths of several microns, yielding aspect ratios of approximately 10:1. The radius of curvature at the tip end is similar to that of the tip before the FIB machining. Another consideration for high-aspect-ratio tips is the tip tilt. To ensure that the pyramidal tip is the lowest part of the tip-cantilever assembly, most AFM designs tilt the cantilever about 15 degrees from parallel. Therefore, even an ideal "line tip" will not give an accurate image of high steep sidewalls, but will produce an image that depends on the scan angle. Due to the versatility of the FIB machining, tips are available with the spikes at an angle to compensate for this effect.

12.1.6 EBD tips

Another method of producing high-aspect-ratio tips for AFM is called electron beam deposition (EBD). First developed for STM tips [12.13, 14], EBD tips were introduced for AFM by focusing an SEM onto the apex of a pyramidal tip arranged so that it pointed along the electron beam axis (see Fig. 12.7). Carbon material was deposited by the dissociation of background gases in the SEM vacuum chamber. *Schiffmann* [12.15] systematically studied the following parameters and how they affected EBD tip geometry:

Deposition time :	0.5 to 8 min
Beam current :	3–300 pA
Beam energy :	1–30 keV
Working distance :	8–48 mm .

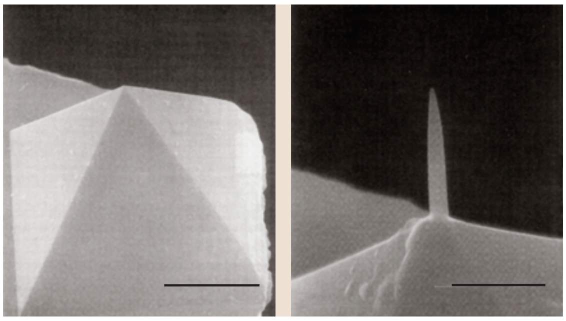

Fig. 12.7 A pyramidal tip before (*left*, 2-μm-scale bar) and after (*right*, 1-μm-scale bar) electron beam deposition, adapted from [12.13]

EBD tips were cylindrical with end radii of 20–40 nm, lengths of 1 to 5 μm, and diameters of 100 to 200 nm. Like FIB tips, EBD tips were found to achieve improved imaging of steep features. By controlling the position of the focused beam, the tip geometry can be further controlled. Tips were fabricated with lengths over 5 μm and aspect ratios greater than 100:1, yet these were too fragile to use as a tip in AFM [12.13].

12.1.7 Carbon Nanotube Tips

Carbon nanotubes are microscopic graphitic cylinders that are nanometers in diameter, yet many microns in length. Single-walled carbon nanotubes (SWNT) consist of single sp^2 hybridized carbon sheets rolled into seamless tubes and have diameters ranging from 0.7 to 3 nm.

Carbon Nanotube Structure

Larger structures called multiwalled carbon nanotubes (MWNT) consist of nested, concentrically arranged SWNT and have diameters ranging from 3 to 50 nm. Figure 12.8 shows a model of nanotube structure, as well as TEM images of a SWNT and a MWNT. The small diameter and high aspect ratio of carbon nanotubes suggests their application as high resolution, high-aspect-ratio AFM probes.

Carbon Nanotube Mechanical Properties

Carbon nanotubes possess exceptional mechanical properties that impact their use as probes. Their lateral stiffness can be approximated from that of a solid elastic rod:

$$k_{\text{lat}} = \frac{3\pi Y r^4}{4l^3} , \qquad (12.4)$$

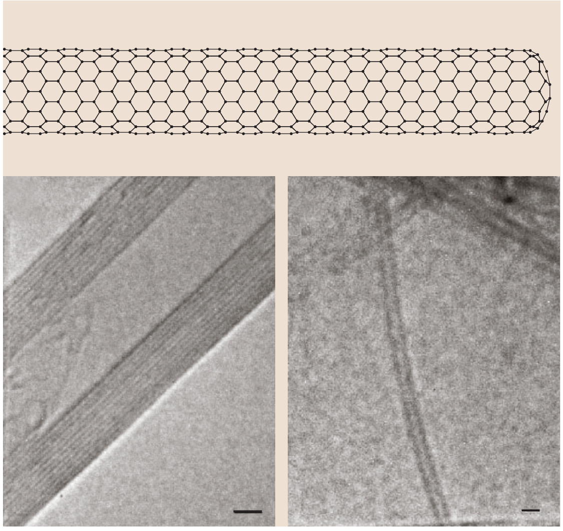

Fig. 12.8 The structure of carbon nanotubes, including TEM images of a MWNT (*left*) and a SWNT (*right*), from [12.16]

where the spring constant k_{lat} represents the restoring force per unit lateral displacement, r is the radius, l is the length, and Y is the Young's modulus (also called the elastic modulus) of the material. For the small diameters and extreme aspect ratios of carbon nanotube tips, the thermal vibrations of the probe tip at room temperature can become sufficient to degrade image resolution. These thermal vibrations can be approximated by equating $\frac{1}{2}k_{\text{B}}T$ of thermal energy to the energy of an oscillating nanotube:

$$\frac{1}{2}k_{\text{B}}T = \frac{1}{2}k_{\text{lat}}a^2 \,, \tag{12.5}$$

where k_{B} is Boltzmann's constant, T is the temperature, and a is the vibration amplitude. Substituting for k_{lat} from (12.4) yields:

$$a = \sqrt{\frac{4k_{\text{B}}Tl^3}{3\pi Yr^4}} \,. \tag{12.6}$$

The strong dependence on radius and length reveals that one must carefully control the tip geometry at this size scale. Equation (12.6) implies that the stiffer the material, i. e., the higher its Young's modulus, the smaller the thermal vibrations and the longer and thinner a tip can be. The Young's moduli of carbon nanotubes have been determined by measurements of the thermal vibration amplitude by TEM [12.18, 19] and by directly measuring the forces required to deflect a pinned carbon nanotube in an AFM [12.20]. These experiments revealed that the Young's modulus of carbon nanotubes is 1–2 TPa, in agreement with theoretical predictions [12.21]. This makes carbon nanotubes the stiffest known material and, therefore, the best for fabricating thin, high-aspect-ratio tips. A more detailed and accurate derivation of the thermal vibration amplitudes was derived for the Young's modulus measurements [12.18, 19].

Carbon nanotubes elastically buckle under large loads, rather than fracture or plastically deform like most materials. Nanotubes were first observed in the buckled state by transmission electron microscopy [12.17], as shown in Fig. 12.9. The first experimental evidence that nanotube buckling is elastic came from the application of nanotubes as probe tips [12.22], described in detail below. A more direct experimental observation of elastic buckling was obtained by deflecting nanotubes pinned to a low friction surface with an AFM tip [12.20]. Both reports found that the buckling force could be approximated with the macroscopic Euler buckling formula for an elastic column:

$$F_{\text{Euler}} = \frac{\pi^3 Yr^4}{4l^2} \,. \tag{12.7}$$

The buckling force puts another constraint on the tip length: If the nanotube is too long the buckling force

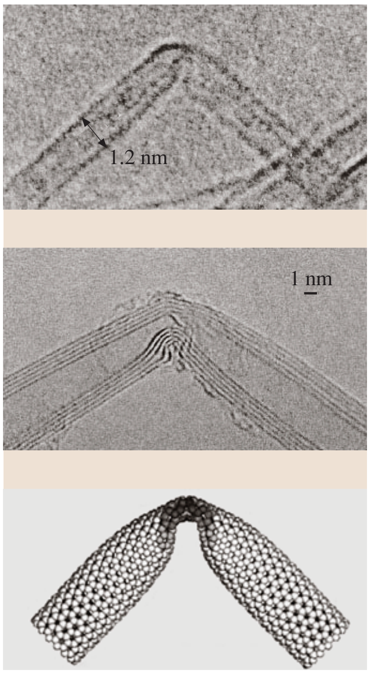

Fig. 12.9 TEM images and a model of a buckled nanotube, adapted from [12.17]

will be too low for stable imaging. The elastic buckling property of carbon nanotubes has significant implications for their use as AFM probes. If a large force is applied to the tip inadvertently, or if the tip encounters a large step in sample height, the nanotube can buckle to the side, then snap back without degraded imaging resolution when the force is removed, making these tips highly robust. No other tip material displays this buckling characteristic.

Manually Assembled Nanotube Probes

The first carbon nanotube AFM probes [12.22] were fabricated by techniques developed for assembling single-nanotube field emission tips [12.23]. This process, illustrated in Fig. 12.10, used purified MWNT material synthesized by the carbon arc procedure. The raw material, which must contain at least a few percent of long nanotubes ($> 10\,\mu\text{m}$) by weight, purified by oxidation to approximately 1% of its original mass. A torn edge of the purified material was attached to a micromanipulator by carbon tape and viewed under a high power optical microscope. Individual nanotubes and nanotube bundles were visible as filaments under dark field illumination. A commercially available AFM tip was attached to another micromanipulator opposing the nanotube material. Glue was applied to the tip apex from high vacuum carbon tape supporting the nanotube material. Nanotubes were then manually attached to the tip apex by micromanipulation. As assembled, MWNT tips were often too long for imaging due to thermal vibrations and low buckling forces described in Sect. 12.1.7. Nanotubes tips were shortened by applying 10 V pulses to the tip while it was near a sputtered niobium surface. This process etched ∼ 100 nm lengths of nanotube per pulse.

The manually assembled MWNT tips demonstrated several important nanotube tip properties [12.22]. First, the high aspect ratio of the MWNT tips allowed the accurate imaging of trenches in silicon with steep sidewalls, similar to FIB and EBD tips. Second, elastic buckling was observed indirectly through force curves (see Fig. 12.11). Note that as the tip taps the sample, the amplitude drops to zero and a DC deflection is observed, because the nanotube is unbuckled and is essentially rigid. As the tip moves closer, the force on the nanotube eventually exceeds the buckling force. The nanotube buckles, allowing the vibration amplitude to partially recover, and the deflection remains constant. Numeric tip trajectory simulations could only reproduce these force curves if elastic buckling was included in the nanotube response. Finally, the nanotube tips were highly robust. Even after "tip crashes" or hundreds of controlled buckling cycles, the tip retained its resolution and high aspect ratio.

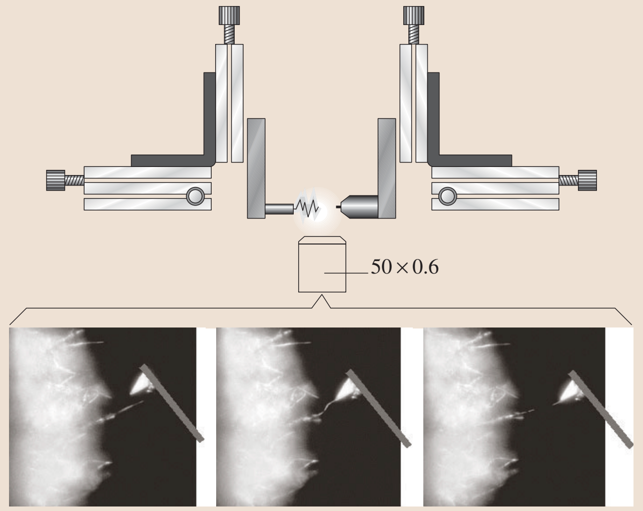

Fig. 12.10 A schematic drawing of the setup for manual assembly of carbon nanotube tips (*top*) and optical microscopy images of the assembly process (the cantilever was drawn in for clarity)

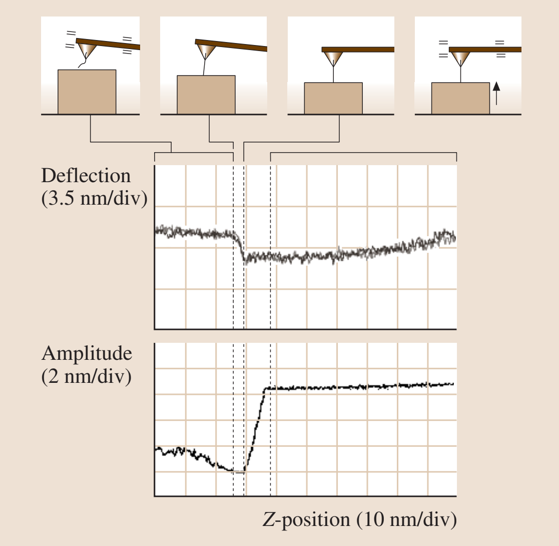

Fig. 12.11 Nanotube tip buckling. *Top* diagrams correspond to labeled regions of the force curves. As the nanotube tip buckles, the deflection remains constant and the amplitude increases, from [12.16]

Manual assembly of carbon nanotube probe tips is straightforward, but has several limitations. It is labor intensive and not amenable to mass production. Although MWNT tips have been made commercially available by this method, they are about ten times more expensive than silicon probes. The manual assembly method has also been carried out in an SEM, rather than an optical microscope [12.24]. This eliminates the need for pulse-etching, since short nanotubes can be attached to the tip, and the "glue" can be applied by EBD. But this is still not the key to mass production, since nanotube tips are made individually. MWNT tips provided a modest improvement in resolution on biological samples, but typical MWNT radii are similar to that of silicon tips, so they cannot provide the ultimate resolution possible with a SWNT tip. SWNT bundles can be attached to silicon probes by manual assembly. Pulse etching at times produces very high resolution tips that likely result from exposing a small number of nanotubes from the bundle, but this is not reproducible [12.25]. Even if a sample could be prepared that consisted of individual SWNT for manual assembly, such nanotubes would not be easily visible by optical microscopy or SEM.

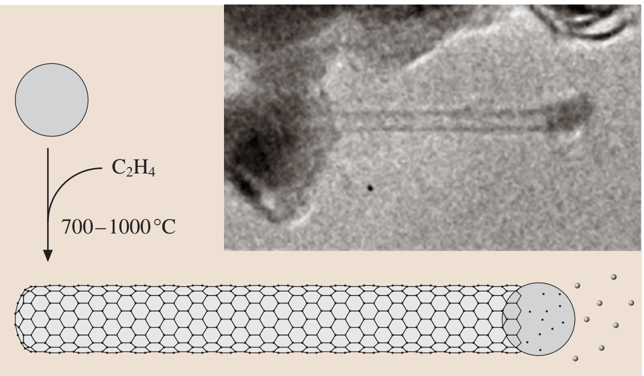

Fig. 12.12 CVD nanotube synthesis. Ethylene reacts with a nanometer-scale iron catalyst particle at high temperature to form a carbon nanotube. The *inset* in the upper right is a TEM image showing a catalyst particle at the end of a nanotube, from [12.16]

CVD Nanotube Probe Synthesis

The problems of manual assembly of nanotube probes discussed above can largely be solved by directly growing nanotubes onto AFM tips by metal-catalyzed chemical vapor deposition (CVD). The key features of the nanotube CVD process are illustrated in Fig. 12.12. Nanometer-scale metal catalyst particles are heated in a gas mixture containing hydrocarbon or CO. The gas molecules dissociate on the metal surface, and carbon is adsorbed into the catalyst particle. When this carbon precipitates, it nucleates a nanotube of similar diameter to the catalyst particle. Therefore, CVD allows control over nanotube size and structure, including the production of SWNTs [12.26] with radii as low as 3.5 Angstrom [12.27].

Several key issues must be addressed to grow nanotube AFM tips by CVD: (1) the alignment of the nanotubes at the tip, (2) the number of nanotubes that grow at the tip, and (3) the length of the nanotube tip. *Li* et al. [12.28] found that nanotubes grow perpendicular to a porous surface containing embedded catalyst. This approach was exploited to fabricate nanotube tips by CVD [12.29] with the proper alignment, as illustrated in Fig. 12.13. A flattened area of approximately 1–5 µm^2 was created on Si tips by scanning in contact mode at high load (1 µN) on a hard, synthetic diamond surface. The tip was then anodized in HF to create 100 nm-diameter pores in this flat surface [12.30]. It is important to only anodize the last 20–40 µm of the cantilever, which includes the tip, so that the rest

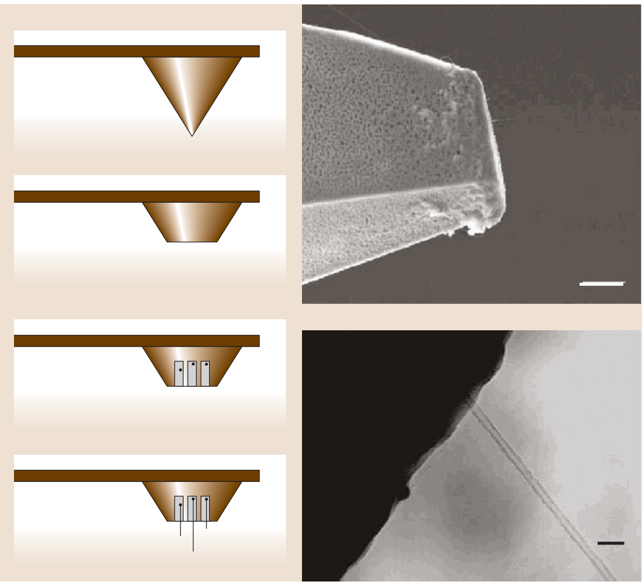

Fig. 12.13 Pore-growth CVD nanotube tip fabrication. The *left panel*, from top to bottom, shows the steps described in the text. The *upper right* is an SEM image of such a tip with a small nanotube protruding from the pores (scale bar is 1 µm). The *lower right* is a TEM of a nanotube protruding from the pores (scale bar is 20 nm), from [12.16]

of the cantilever is still reflective for use in the AFM. This was achieved by anodizing the tip in a small drop of HF under the view of an optical microscope. Next, iron was electrochemically deposited into the pores to form catalyst particles [12.31]. Tips prepared in this way were heated in low concentrations of ethylene at 800 °C, which is known to favor the growth of thin nanotubes [12.26]. When imaged by SEM, nanotubes were found to grow perpendicular to the surface from the pores as desired (Fig. 12.13). TEM revealed that the nanotubes were thin, individual, multiwalled nanotubes with typical radii ranging from 3–5 nm. If nanotubes did not grow in an acceptable orientation, the carbon could be removed by oxidation, and then CVD repeated to grow new nanotube tips.

These "pore-growth" CVD nanotube tips were typically several microns in length – too long for imaging – and were pulse-etched to a usable length of < 500 nm. The tips exhibited elastic buckling behavior and were very robust in imaging. In addition, the thin, individual nanotube tips enabled improved resolution [12.29] on isolated proteins. The pore-growth method demonstrated the potential of CVD to simplify the fabrication of nanotube tips, although there were still limitations. In particular, the porous layer was difficult to prepare and rather fragile.

An alternative approach for CVD fabrication of nanotube tips involves direct growth of SWNTs on the surface of a pyramidal AFM tip [12.32, 33]. In this "surface-growth" approach, an alumina/iron/molybdenum-powdered catalyst known to produce SWNT [12.26] was dispersed in ethanol at 1 mg/mL. Silicon tips were dipped in this solution and allowed to dry, leaving a sparse layer of ~ 100 nm catalyst clusters on the tip. When CVD conditions were applied, single-walled nanotubes grew along the silicon tip surface. At a pyramid edge, nanotubes can either bend to align with the edge, or protrude from the surface. If the energy required to bend the tube and follow the edge is less than the attractive nanotube-surface energy, then the nanotube will follow the pyramid edge to the apex. Therefore, nanotubes were effectively steered toward the tip apex by the pyramid edges. At the apex, the nanotube protruded from the tip, since the energetic cost of bending around the sharp silicon tip was too high. The high aspect ratio at the oxide-sharpened silicon tip apex was critical for good nanotube alignment. A schematic of this approach is shown in Fig. 12.14. Evidence for this model came from SEM investigations that show that a very high yield of tips contains nanotubes only at the apex, with very few protruding elsewhere from the pyramid. TEM analysis demonstrated that the tips typically consist of small SWNT bundles that are formed by nanotubes coming together from different edges of the pyramid to join at the apex, supporting the surface growth model described above (Fig. 12.14). The "surface growth" nanotube tips exhibit a high aspect ratio and high resolution imaging, as well as elastic buckling.

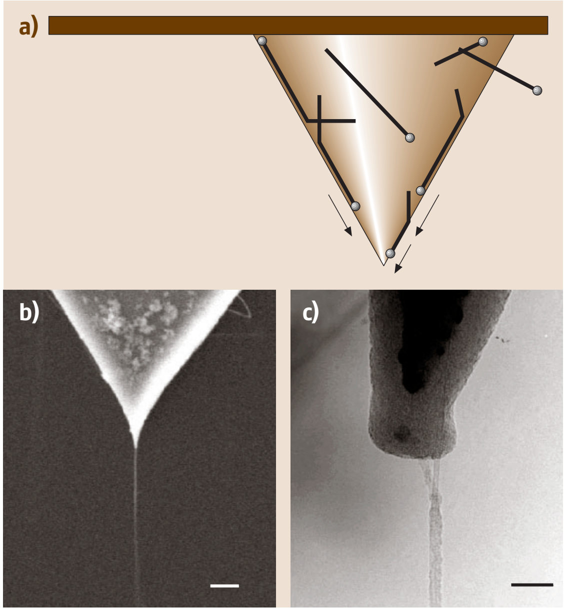

Fig. 12.14a–c Surface-growth nanotube tip fabrication. **(a)** Schematic represents the surface growth process in which nanotubes growing on the pyramidal tip are guided to the tip apex. **(b)**,**(c)** Images show **(b)** SEM (200-nm-scale bar) and **(c)** TEM (20-nm-scale bar) images of a surface growth tip, from [12.16]

The surface growth method has been expanded to include wafer-scale production of nanotube tips with yields of over 90% [12.34], yet one obstacle remains to the mass production of nanotube probe tips. Nanotubes protruding from the tip are several microns long, and since they are so thin, they must be etched to less than 100 nm. While the pulse-etching step is fairly reproducible, it must be carried out on nanotube tips in a serial fashion, so surface growth does not yet represent a true method of batch nanotube tip fabrication.

Hybrid Nanotube Tip Fabrication: Pick-up Tips

Another method of creating nanotube tips is something of a hybrid between assembly and CVD. The motivation was to create AFM probes that have an *individual* SWNT at the tip to achieve the ultimate imaging resolution. In order to synthesize isolated SWNT, they must be nucleated at sites separated farther than their typical length. The alumina-supported catalyst contains a high density of catalyst particles per 100 nm cluster, so nanotube bundles cannot be avoided. To fabricate completely isolated nanotubes, isolated catalyst particles were formed by dipping a silicon wafer in an isopropyl alcohol solution of $Fe(NO_3)_3$. This effectively left a submonolayer of iron on the wafer, so that when it was heated in a CVD furnace, the iron became mobile and aggregated to form isolated iron particles. During CVD conditions, these particles nucleated and grew SWNTs. By controlling the reaction time, the SWNT lengths were kept shorter than their typical separation, so that the nanotubes never had a chance to form bundles. AFM analysis of these samples revealed 1–3 nm-diameter SWNT and un-nucleated particles on the surface (Fig. 12.15). However, there were tall objects that were difficult to image at a density of about 1 per 50 μm^2. SEM analysis at an oblique angle demonstrated that these were SWNTs that had grown perpendicular to the surface (Fig. 12.15).

In the "pick-up tip" method, these isolated SWNT substrates were imaged by AFM with silicon tips in air [12.9]. When the tip encountered a vertical SWNT, the oscillation amplitude was damped, so the AFM pulled the sample away from the tip. This pulled the SWNT into contact with the tip along its length, so that it became attached to the tip. This assembly process happened automatically when imaging in tapping mode – no special tip manipulation was required. When imaging a wafer with the density shown in Fig. 12.15, one nanotube was attached per 8 µm × 8 µm scan at 512×512 and 2 Hz, so a nanotube tip could be made in about 5 min. Since the as-formed SWNT tip continued to image, there was usually no evidence in the topographic image that a nanotube had been attached. However, the pick-up event was identified when the Z-voltage suddenly stepped to larger tip-

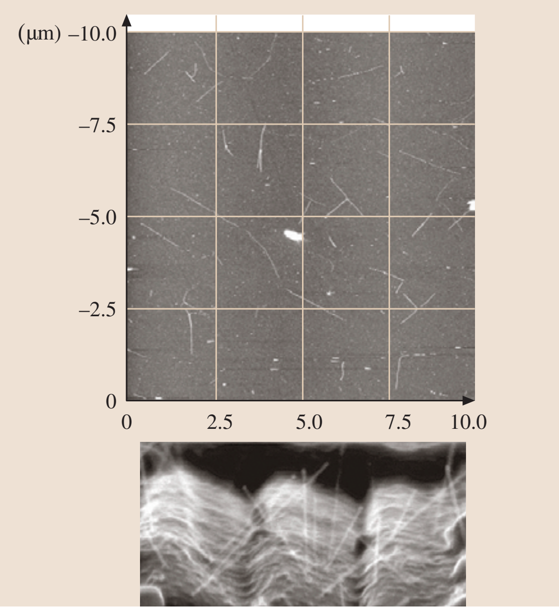

Fig. 12.15 Atomic force microscopy image (*top*) of a wafer with isolated nanotubes synthesized by CVD. The SEM view provides evidence that some of these nanotubes are arranged vertically

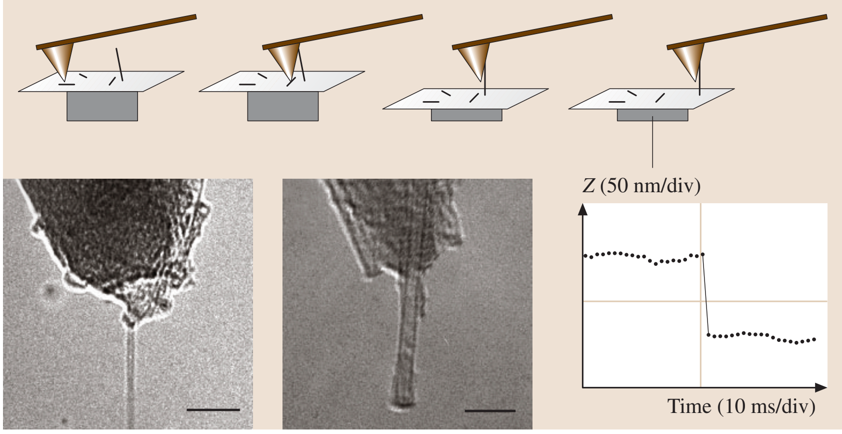

Fig. 12.16 Pick-up tip assembly of nanotube probes. The *top* illustrates the nanotube pick-up process that occurs while imaging vertical nanotubes in an AFM, including a trace of the Z-position during a pick-up event. The *lower left* TEM images show single nanotubes (diameters 0.9 nm and 2.8 nm) on an AFM tip fabricated by this method, adapted from [12.9]

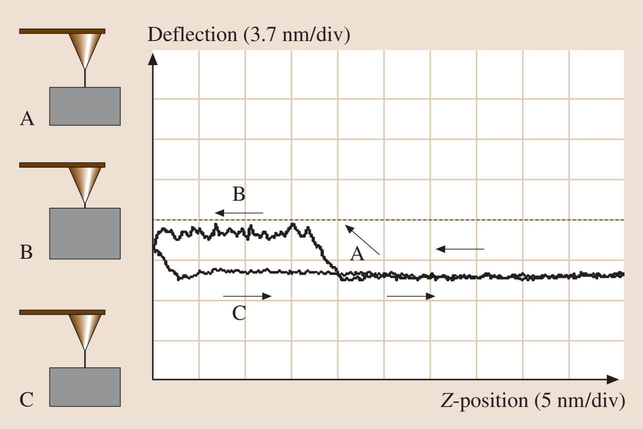

Fig. 12.17 The process by which nanotube tips can be shortened in AFM force curves. The hysteresis in the deflection trace (*bottom*) reveals that ∼ 17 nm were removed

sample separation due to the effective increase in tip length.

Individual SWNT tips must be quite short, typically less than 50 nm in length, for reasons outlined above. Pulse-etching, which removes 50–100 nm of nanotube length at a time, lacked the necessary precision for shortening pick-up tips, so the tips were shortened through force curves. A pick-up tip force curve is shown in Fig. 12.17. When the tip first interacted with the sample, the amplitude decreased to zero, and further approach generated a small deflection that ultimately saturated. However, this saturation was not due to buckling. Note that the amplitude did not recover, as in the buckling curves. This leveling was due to the nanotube sliding on the pyramidal tip, which was confirmed by the hysteresis in the amplitude and deflection curves. If the force curve was repeated, the tip showed no deflection or amplitude drop until further down, because the tip was essentially shorter.

Pick-up SWNT tips achieve the highest resolution of all nanotube tips, since they are always individuals rather than bundles. In tapping mode, they produce images of 5 nm gold particles that have a *full width* of ∼ 7 nm, the expected geometrical resolution for a 2 nm cylindrical probe [12.9]. Although the pick-up method is serial in nature, it may still be the key to the mass production of nanotube tips. Note that the original nanotube tip length can be measured electronically from the size of the Z-piezo step. The shortening can be electronically controlled through the hysteresis in the force curves. Therefore, the entire procedure (including tip exchange) can be automated by computer.

12.2 Scanning Tunneling Microscopy

Scanning tunneling microscopy (STM) was the original scanning probe microscopy and generally produces the highest resolution images, routinely achieving atomic resolution on flat, conductive surfaces. In STM, the probe tip consists of a sharpened metal wire that is held 0.3 to 1 nm from the sample. A potential difference of 0.1 V to 1 V between the tip and sample leads to tunneling currents on the order of 0.1 to 1 nA. As in AFM, a piezo-scanner rasters the sample under the tip, and the Z-position is adjusted to hold the tunneling current constant. The Z-position data represents the "topography", or in this case the surface, of constant electron density. As with other SPMs, the tip properties and performance greatly depend on the experiment being carried out. Although it is nearly impossible to prepare a tip with a known atomic structure, a number of factors are known to affect tip performance, and several preparation methods have been developed that produce good tips.

The nature of the sample being investigated and the scanning environment will affect the choice of the tip material and how the tip is fabricated. Factors to consider are mechanical properties – a hard material that will resist damage during tip-sample contact is desired. Chemical properties should also be considered – formation of oxides, or other insulating contaminants will affect tip performance. Tungsten is a common tip material because it is very hard and will resist damage, but its use is limited to ultrahigh vacuum (UHV) conditions, since it readily oxidizes. For imaging under ambient conditions an inert tip material such as platinum or gold is preferred. Platinum is typically alloyed with iridium to increase its stiffness.

12.2.1 Mechanically Cut STM Tips

STM tips can be fabricated by simple mechanical procedures such as grinding or cutting metal wires. Such tips are not formed with highly reproducible shapes and have a large opening angle and a large radius of curvature in the range of 0.1 to 1 μm (see Fig. 12.18a). They are not useful for imaging samples with surface roughness above a few nanometers. However, on atomically flat samples, mechanically cut tips can achieve atomic resolution due to the nature of the tunneling signal, which drops exponentially with tip-sample separation. Since mechanically cut tips contain many small asperities on the larger tip structure, atomic resolution is easily achieved as long as one atom of the tip is just a few angstroms lower than all of the others.

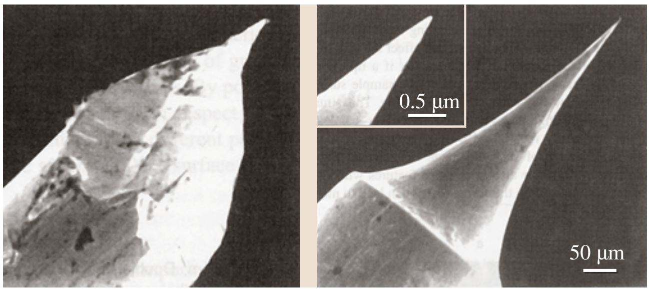

Fig. 12.18 A mechanically cut STM tip (*left*) and an electrochemically etched STM tip (*right*), from [12.35]

12.2.2 Electrochemically Etched STM Tips

For samples with more than a few nanometers of surface roughness, the tip structure in the nanometer-size range becomes an issue. Electrochemical etching can provide tips with reproducible and desirable shapes and sizes (Fig. 12.18), although the exact atomic structure of the tip apex is still not well controlled. The parameters of electrochemical etching depend greatly on the tip material and the desired tip shape. The following is an entirely general description. A fine metal wire (0.1–1 mm diameter) of the tip material is immersed in an appropriate electrochemical etchant solution. A voltage bias of 1–10 V is applied between the tip and a counterelectrode such that the tip is etched. Due to the enhanced etch rate at the electrolyte-air interface, a neck is formed in the wire. This neck is eventually etched thin enough so that it cannot support the weight of the part of the wire suspended in the solution, and it breaks to form a sharp tip. The widely varying parameters and methods will be not be covered in detail here, but many recipes are found in the literature for common tip materials [12.36–39].

References

12.1 R. Linnemann, T. Gotszalk, I. W. Rangelow, P. Dumania, E. Oesterschulze: Atomic force microscopy and lateral force microscopy using piezoresistive cantilevers, J. Vac. Sci. Technol. B **14**(2) (1996) 856–860

12.2 T. R. Albrecht, S. Akamine, T. E. Carver, C. F. Quate: Microfabrication of cantilever styli for the atomic force microscope, J. Vac. Sci. Technol. A **8**(4) (1990) 3386–3396

12.3 O. Wolter, T. Bayer, J. Greschner: Micromachined silicon sensors for scanning force microscopy, J. Vac. Sci. Technol. B **9**(2) (1991) 1353–1357

12.4 C. Bustamante, D. Keller: Scanning force microscopy in biology, Phys. Today **48**(12) (1995) 32–38

12.5 J. Vesenka, S. Manne, R. Giberson, T. Marsh, E. Henderson: Colloidal gold particles as an incompressible atomic force microscope imaging standard for assessing the compressibility of biomolecules, Biophys. J. **65** (1993) 992–997

12.6 D. J. Muller, D. Fotiadis, S. Scheuring, S. A. Muller, A. Engel: Electrostatically balanced subnanometer imaging of biological specimens by atomic force microscope, Biophys. J. **76**(2) (1999) 1101–1111

12.7 R. B. Marcus, T. S. Ravi, T. Gmitter, K. Chin, D. Liu, W. J. Orvis, D. R. Ciarlo, C. E. Hunt, J. Trujillo: Formation of silicon tips with <1 nm radius, Appl. Phys. Lett. **56**(3) (1990) 236–238

12.8 J. H. Hafner, C. L. Cheung, C. M. Lieber: unpublished results (2001)

12.9 J. H. Hafner, C. L. Cheung, T. H. Oosterkamp, C. M. Lieber: High-yield assembly of individual single-walled carbon nanotube tips for scanning probe microscopies, J. Phys. Chem. B **105**(4) (2001) 743–746

12.10 F. Ohnesorge, G. Binnig: True atomic resolution by atomic force microscopy through repulsive and attractive forces, Science **260** (1993) 1451–1456

12.11 D. J. Muller, D. Fotiadis, A. Engel: Mapping flexible protein domains at subnanometer resolution with the atomic force microscope, FEBS Lett. **430**(1–2 Special Issue SI) (1998) 105–111

12.12 S. Akamine, R. C. Barrett, C. F. Quate: Improved atomic force microscope images using microcantilevers with sharp tips, Appl. Phys. Lett. **57**(3) (1990) 316–318

12.13 D.J. Keller, C. Chih-Chung: Imaging steep, high structures by scanning force microscopy with electron beam deposited tips, Surf. Sci. **268** (1992) 333–339

12.14 T. Ichihashi, S. Matsui: In situ observation on electron beam induced chemical vapor deposition by transmission electron microscopy, J. Vac. Sci. Technol. B **6**(6) (1988) 1869–1872

12.15 K. I. Schiffmann: Investigation of fabrication parameters for the electron-beam-induced deposition of contamination tips used in atomic force microscopy, Nanotechnology **4** (1993) 163–169

12.16 J. H. Hafner, C. L. Cheung, A. T. Woolley, C. M. Lieber: Structural and functional imaging with carbon nanotube AFM probes, Prog. Biophys. Mol. Biol. **77**(1) (2001) 73–110

12.17 S. Iijima, C. Brabec, A. Maiti, J. Bernholc: Structural flexibility of carbon nanotubes, J. Chem. Phys. **104**(5) (1996) 2089–2092

12.18 M. M. J. Treacy, T. W. Ebbesen, J. M. Gibson: Exceptionally high Young's modulus observed for individual carbon nanotubes, Nature **381** (1996) 678–680

12.19 A. Krishnan, E. Dujardin, T. W. Ebbesen, P. N. Yianilos, M. M. J. Treacy: Young's modulus of single-walled nanotubes, Phys. Rev. B **58**(20) (1998) 14013–14019

12.20 E. W. Wong, P. E. Sheehan, C. M. Lieber: Nanobeam mechanics – elasticity, strength, and toughness of nanorods and nanotubes, Science **277**(5334) (1997) 1971–1975

12.21 J. P. Lu: Elastic properties of carbon nanotubes and nanoropes, Phys. Rev. Lett. **79**(7) (1997) 1297–1300

12.22 H. J. Dai, J. H. Hafner, A. G. Rinzler, D. T. Colbert, R. E. Smalley: Nanotubes as nanoprobes in scanning probe microscopy, Nature **384**(6605) (1996) 147–150

12.23 A. G. Rinzler, Y. H. Hafner, P. Nikolaev, L. Lou, S. G. Kim, D. Tomanek, D. T. Colbert, R. E. Smalley: Unraveling nanotubes: Field emission from atomic wire, Science **269** (1995) 1550

12.24 H. Nishijima, S. Kamo, S. Akita, Y. Nakayama, K. I. Hohmura, S. H. Yoshimura, K. Takeyasu: Carbon-nanotube tips for scanning probe microscopy: Preparation by a controlled process and observation of deoxyribonucleic acid, Appl. Phys. Lett. **74**(26) (1999) 4061–4063

12.25 S. S. Wong, A. T. Woolley, T. W. Odom, J. L. Huang, P. Kim, D. V. Vezenov , C. M. Lieber: Single-walled carbon nanotube probes for high-resolution nanostructure imaging, Appl. Phys. Lett. **73**(23) (1998) 3465–3467

12.26 J. H. Hafner, M. J. Bronikowski, B. R. Azamian, P. Nikolaev, A. G. Rinzler, D. T. Colbert, K. A. Smith, R. E. Smalley: Catalytic growth of single-wall carbon nanotubes from metal particles, Chem. Phys. Lett. **296**(1–2) (1998) 195–202

12.27 P. Nikolaev, M. J. Bronikowski, R. K. Bradley, F. Rohmund, D. T. Colbert, K. A. Smith, R. E. Smalley: Gas-phase catalytic growth of single-walled carbon nanotubes from carbon monoxide, Chem. Phys. Lett. **313**(1–2) (1999) 91–97

12.28 W. Z. Li, S. S. Xie, L. X. Qian, B. H. Chang, B. S. Zou, W. Y. Zhou, R. A. Zhao, G. Wang: Large-scale synthesis of aligned carbon nanotubes, Science **274**(5293) (1996) 1701–1703

12.29 J. H. Hafner, C. L. Cheung, C. M. Lieber: Growth of nanotubes for probe microscopy tips, Nature **398**(6730) (1999) 761–762

12.30 V. Lehmann: The physics of macroporous silicon formation, Thin Solid Films **255** (1995) 1–4

12.31 F. Ronkel, J. W. Schultze, R. Arensfischer: Electrical contact to porous silicon by electrodeposition of iron, Thin Solid Films **276**(1–2) (1996) 40–43

12.32 J. H. Hafner, C. L. Cheung, C. M. Lieber: Direct growth of single-walled carbon nanotube scanning probe microscopy tips, J. Am. Chem. Soc. **121**(41) (1999) 9750–9751

12.33 E. B. Cooper, S. R. Manalis, H. Fang, H. Dai, K. Matsumoto, S. C. Minne, T. Hunt, C. F. Quate: Terabit-per-square-inch data storage with the atomic force microscope, Appl. Phys. Lett. **75**(22) (1999) 3566–3568

12.34 E. Yenilmez, Q. Wang, R. J. Chen, D. Wang, H. Dai: Wafer scale production of carbon nanotube scanning probe tips for atomic force microscopy, Appl. Phys. Lett. **80**(12) (2002) 2225–2227

12.35 A. Stemmer, A. Hefti, U. Aebi, A. Engel: Scanning tunneling and transmission electron microscopy on identical areas of biological specimens, Ultramicroscopy **30**(3) (1989) 263

12.36 R. Nicolaides, L. Yong, W. E. Packard, W. F. Zhou, H. A. Blackstead, K. K. Chin, J. D. Dow, J. K. Furdyna, M. H. Wei, R. C. Jaklevic, W. J. Kaiser, A. R. Pelton, M. V. Zeller, J. J. Bellina: Scanning tunneling microscope tip structures, J. Vac. Sci. Technol. A **6**(2) (1988) 445–447

12.37 J. P. Ibe, P. P. Bey, S. L. Brandow, R. A. Brizzolara, N. A. Burnham, D. P. DiLella, K. P. Lee, C. R. K. Marrian, R. J. Colton: On the electrochemical etching of tips for scanning tunneling microscopy, J. Vac. Sci. Technol. A **8** (1990) 3570–3575

12.38 L. Libioulle, Y. Houbion, J.-M. Gilles: Very sharp platinum tips for scanning tunneling microscopy, Rev. Sci. Instrum. **66**(1) (1995) 97–100

12.39 A. J. Nam, A. Teren, T. A. Lusby, A. J. Melmed: Benign making of sharp tips for STM and FIM: Pt, Ir, Au, Pd, and Rh, J. Vac. Sci. Technol. B **13**(4) (1995) 1556–1559

13. Noncontact Atomic Force Microscopy and Its Related Topics

The scanning probe microscopy (SPM) such as the STM and the NC-AFM is the basic technology for the nanotechnology and also for the future bottom-up process. In Sect. 13.1, the principles of AFM such as operating modes and the frequency modulation method of the NC-AFM are fully explained. Then, in Sect. 13.2, applications of NC-AFM to semiconductors that make clear its potentials such as spatial resolution and functions are introduced. Next, in Sect. 13.3, applications of NC-AFM to insulators such as alkali halides, fluorides and transition metal oxides are introduced. At last, in Sect. 13.4, applications of NC-AFM to molecules such as carboxylate ($RCOO^-$) with $R = H$, CH_3, $C(CH_3)_3$ and CF_3 are introduced. Thus, the NC-AFM can observe atoms and molecules on various kinds of surfaces such as semiconductor, insulator and metal oxide with atomic/molecular resolutions. These sections are essential to understand the status of the art and the future possibility of NC-AFM that is the second generation of atom/molecule technology.

13.1 **Principles of Noncontact Atomic Force Microscope (NC-AFM)** 386
13.1.1 Imaging Signal in AFM 386
13.1.2 Experimental Measurement and Noise 387
13.1.3 Static AFM Operating Mode 387
13.1.4 Dynamic AFM Operating Mode 388
13.1.5 The Four Additional Challenges Faced by AFM 388
13.1.6 Frequency-Modulation AFM (FM-AFM) 389
13.1.7 Relation Between Frequency Shift and Forces 390
13.1.8 Noise in Frequency-Modulation AFM – Generic Calculation 391
13.1.9 Conclusion 391

13.2 **Applications to Semiconductors** 391
13.2.1 Si(111)7×7 Surface 392
13.2.2 Si(100)2×1 and Si(100)2×1:H Monohydride Surfaces 393
13.2.3 Metal-Deposited Si Surface 395

13.3 **Applications to Insulators** 397
13.3.1 Alkali Halides, Fluorides, and Metal Oxides 397
13.3.2 Atomically Resolved Imaging of a NiO(001) Surface 402
13.3.3 Atomically Resolved Imaging Using Noncoated and Fe-Coated Si Tips 402

13.4 **Applications to Molecules** 404
13.4.1 Why Molecules and What Molecules? 404
13.4.2 Mechanism of Molecular Imaging 404
13.4.3 Perspectives 407

References 407

The scanning tunneling microscope (STM) is an atomic tool based on an electric method that measures the tunneling current between a conductive tip and a conductive surface. It can electrically observe individual atoms/molecules. It can characterize or analyze the electronic nature around surface atoms/molecules. In addition, it can manipulate individual atoms/molecules. Hence, the STM is the first generation of atom/molecule technology. On the other hand, the atomic force microscope (AFM) is a unique atomic tool based on a mechanical method that can deal with even the insulator surface. Since the invention of noncontactAFM (NC-AFM) in 1995, the NC-AFM and the NC-AFM-based method have rapidly developed into a powerful surface tool on atomic/molecular scales, because the NC-AFM has the following characteristics: (1) it has true atomic resolution, (2) it can measure atomic force (so-called atomic force spectroscopy), (3) it can observe even insulators, and (4) it can measure mechanical responses such as elastic deformation. Thus, the NC-AFM is the second generation of atom/molecule technology. The scanning probe microscopy (SPM) such as the STM and the NC-AFM is the basic technology for nanotechnology and also for the future bottom-up process.

In Sect. 13.1, principles of NC-AFM will be fully introduced. Then, in Sect. 13.2, applications to semi-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

conductors will be presented. Next, in Sect. 13.3, applications to insulators will be described. And, in Sect. 13.4, applications to molecules will be introduced. These sections are essential to understanding the status of the art and the future possibility of the NC-AFM.

13.1 Principles of Noncontact Atomic Force Microscope (NC-AFM)

The atomic force microscope (AFM), invented by *Binnig* [13.1] and introduced in 1986 by *Binnig*, *Quate*, and *Gerber* [13.2], is an offspring of the scanning tunneling microscope (STM) [13.3]. The STM is covered in books and review articles, e.g., [13.4–9]. Early on in the development of STM it became evident that relatively strong forces act between a tip in close proximity to a sample. It was found that these forces could be put to good use in the atomic force microscope (AFM). Detailed information about the noncontact AFM can be found in [13.10–12].

13.1.1 Imaging Signal in AFM

Figure 13.1 shows a sharp tip close to a sample. The potential energy between the tip and sample V_{ts} causes a z component of the tip-sample force $F_{\mathrm{ts}} = -\partial V_{\mathrm{ts}}/\partial z$. Depending on the mode of operation, the AFM uses F_{ts} or some entity derived from F_{ts} as the imaging signal.

Unlike the tunneling current, which has a very strong distance dependence, F_{ts} has long- and short-range contributions. We can classify the contributions by their range and strength. In vacuum, there are van der Waals, electrostatic, and magnetic forces with long-range (up to 100 nm) and short-range chemical forces (fractions of a nm).

The van der Waals interaction is caused by fluctuations in the electric dipole moment of atoms and their mutual polarization. For a spherical tip with radius R next to a flat surface (z is the distance between the plane connecting the centers of the surface atoms and the center of the closest tip atom), the van der Waals potential is given by [13.13]:

$$V_{\mathrm{vdW}} = -\frac{A_{\mathrm{H}}}{6z}\,. \tag{13.1}$$

The "Hamaker constant", A_{H}, depends on the type of materials (atomic polarizability and density) of the tip and sample and is on the order of 1 eV for most solids [13.13].

When the tip and sample are both conductive and have an electrostatic potential difference, $U \neq 0$, electrostatic forces are important. For a spherical tip with radius R, the force is given by [13.14]:

$$F_{\mathrm{electrostatic}} = -\frac{\pi\varepsilon_0 R U^2}{z}\,. \tag{13.2}$$

Chemical forces are more complicated. Empirical model potentials for chemical bonds are the Morse potential (see e.g., [13.13]).

$$V_{\mathrm{Morse}} = -E_{\mathrm{bond}}\left(2\,\mathrm{e}^{-\kappa(z-\sigma)} - \mathrm{e}^{-2\kappa(z-\sigma)}\right) \tag{13.3}$$

and the Lennard-Jones potential [13.13]:

$$V_{\mathrm{Lennard\text{-}Jones}} = -E_{\mathrm{bond}}\left(2\frac{\sigma^6}{z^6} - \frac{\sigma^{12}}{z^{12}}\right)\,. \tag{13.4}$$

These potentials describe a chemical bond with bonding energy E_{bond} and equilibrium distance σ. The Morse potential has an additional parameter – a decay length κ.

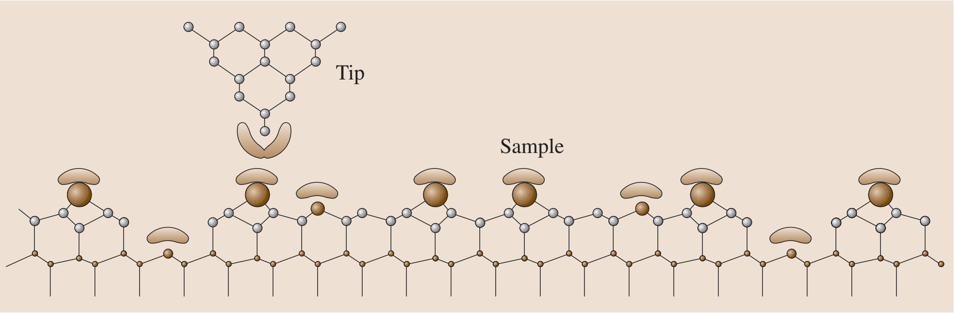

Fig. 13.1 Schematic view of an AFM tip close to a sample

13.1.2 Experimental Measurement and Noise

Forces between the tip and sample are typically measured by recording the deflection of a cantilever beam that has a tip mounted to its end (see Fig. 13.2). Today's microfabricated silicon cantilevers were first created by the *Calvin F. Quate* group [13.15–17] and *Wolter* et al. at IBM [13.18].

The cantilever is characterized by its spring constant k, eigenfrequency f_0, and quality factor Q.

For a rectangular cantilever with dimensions w, t, and L (see Fig. 13.2), the spring constant k is given by [13.6]:

$$k = \frac{E_Y w t^3}{4L^3} , \tag{13.5}$$

where E_Y is Young's modulus. The eigenfrequency f_0 is given by [13.6]:

$$f_0 = 0.162 \frac{t}{L^2} \sqrt{E_Y/\rho} , \tag{13.6}$$

where ρ is the mass density of the cantilever material. The Q-factor depends on the damping mechanisms present in the cantilever. For micromachined cantilevers operated in air, Q is typically a few hundred but can reach hundreds of thousands in vacuum.

In the first AFM, the deflection of the cantilever was measured with an STM – the backside of the cantilever was metalized, and a tunneling tip was brought close to it to measure the deflection [13.2]. Today's designs use optical (interferometer, beam-bounce) or electrical methods (piezoresistive, piezoelectric) for measuring the cantilever deflection. A discussion of the various techniques can be found in [13.19], descriptions of piezoresistive detection schemes are found in [13.17, 20], and piezoelectric methods are explained in [13.21–24].

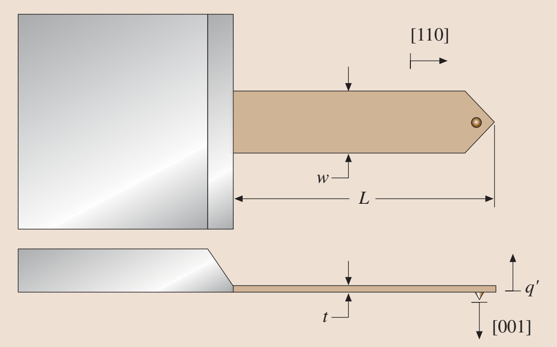

Fig. 13.2 Top view and side view of a microfabricated silicon cantilever (schematic)

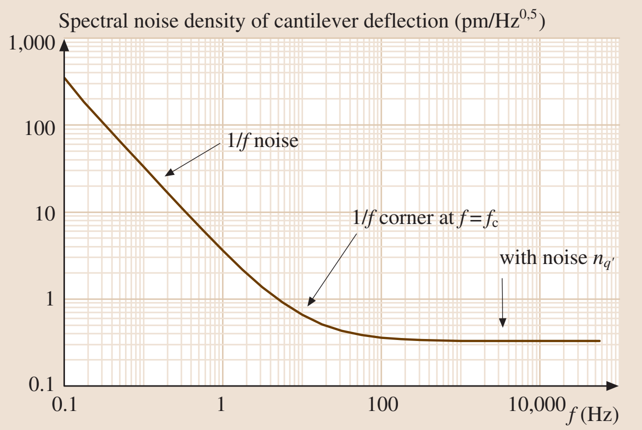

Fig. 13.3 Schematic view of $1/f$ noise apparent in force detectors. Static AFMs operate in a frequency range from 0.01 Hz to a few hundred Hz, while dynamic AFMs operate at frequencies around 10 kHz to a few hundred kHz. The noise of the cantilever deflection sensor is characterized by the $1/f$ corner frequency f_c and the constant deflection noise density $n_{q'}$ for the frequency range where white noise is dominating

The quality of the cantilever deflection measurement can be expressed in a schematic plot of the deflection noise density versus frequency, as in Fig. 13.3.

The noise density has a $1/f$ dependence for low frequency and merges into a constant noise density ("white noise") above the "$1/f$ corner frequency".

13.1.3 Static AFM Operating Mode

In the static mode of operation, the force translates into a deflection $q' = F_{ts}/k$ of the cantilever, yielding images as maps $z(x, y, F_{ts} = \text{const.})$. The noise level of the force measurement is then given by the cantilever's spring constant k times the noise level of the deflection measurement. In this respect, a small value for k increases force sensitivity. On the other hand, instabilities are more likely to occur with soft cantilevers (see Sect. 13.1.5). Because the deflection of the cantilever should be significantly larger than the deformation of the tip and sample, the cantilever should be much softer than the bonds between the bulk atoms in the tip and sample. Inter-atomic force constants in solids are in a range from 10 N/m to about 100 N/m – in biological samples, they can be as small as 0.1 N/m. Thus, typical values for k in the static mode are 0.01–5 N/m.

Even though it has been demonstrated that atomic resolution is possible with static AFM, the method can only be applied in certain cases. The detrimental effects of $1/f$ noise can be limited by working at low temperatures [13.25], where the coefficients of thermal expansion are very small, or by building the AFM of a material with a low thermal expansion coefficient [13.26]. The long-range attractive forces have to be cancelled by immersing the tip and sample in a liquid [13.26], or by partly compensating the attractive force by pulling at the cantilever after jump-to-contact has occurred [13.27]. *Jarvis* et al. have cancelled the long-range attractive force with an electromagnetic force applied to the cantilever [13.28]. Even with these restrictions, static AFM does not produce atomic resolution on reactive surfaces like silicon, as the chemical bonding of the AFM tip and sample pose an unsurmountable problem [13.29, 30].

13.1.4 Dynamic AFM Operating Mode

In the dynamic operation modes, the cantilever is deliberately vibrated. There are two basic methods of dynamic operation: amplitude modulation (AM) and frequency modulation (FM) operation. In AM-AFM [13.31], the actuator is driven by a fixed amplitude A_{drive} at a fixed frequency f_{drive}, where f_{drive} is close to f_0. When the tip approaches the sample, elastic and inelastic interactions cause a change in both the amplitude and the phase (relative to the driving signal) of the cantilever. These changes are used as the feedback signal. While the AM mode was initially used in a noncontact mode, it was later implemented very successfully at a closer distance range in ambient conditions involving repulsive tip-sample interactions.

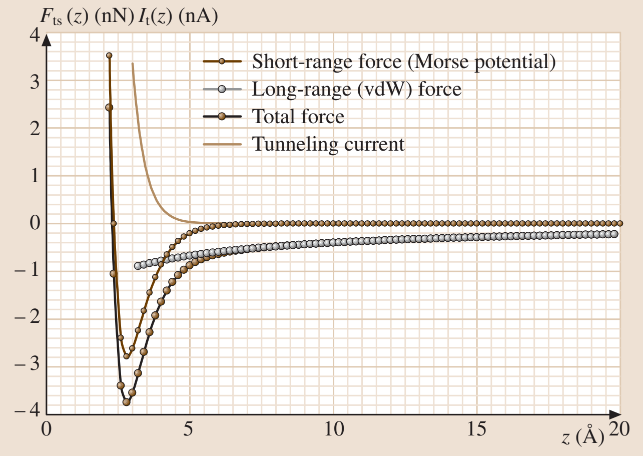

Fig. 13.4 Plot of tunneling current I_t and force F_{ts} (typical values) as a function of distance z between front atom and surface atom layer

The change in amplitude in AM mode does not occur instantaneously with a change in the tip-sample interaction, but on a time scale of $\tau_{\text{AM}} \approx 2Q/f_0$, and the AM mode is slow with high-Q cantilevers. However, the use of high Q-factors reduces noise. *Albrecht* et al. found a way to combine the benefits of high Q and high speed by introducing the frequency modulation (FM) mode [13.32], where the change in the eigenfrequency settles on a time scale of $\tau_{\text{FM}} \approx 1/f_0$.

Using the FM mode, the resolution was improved dramatically, and finally atomic resolution [13.33, 34] was obtained by reducing the tip-sample distance and working in vacuum. For atomic studies in vacuum, the FM mode (see Sect. 13.1.6) is now the preferred AFM technique. However, atomic resolution in vacuum can also be obtained with the AM mode, as demonstrated by *Erlandsson* et al. [13.35].

13.1.5 The Four Additional Challenges Faced by AFM

Some of the inherent AFM challenges are apparent by comparing the tunneling current and tip-sample force as a function of distance (Fig. 13.4).

The tunneling current is a monotonic function of the tip-sample distance and has a very sharp distance dependence. In contrast, the tip-sample force has long- and short-range components and is not monotonic.

Jump-to-Contact and Other Instabilities

If the tip is mounted on a soft cantilever, the initially attractive tip-sample forces can cause a sudden "jump-to-contact" when the tip approaches the sample. This instability occurs in the quasi-static mode if [13.36, 37]

$$k < \max\left(-\frac{\partial^2 V_{\text{ts}}}{\partial z^2}\right) = k_{\text{ts}}^{\max} . \qquad (13.7)$$

Jump-to-contact can be avoided even for soft cantilevers by oscillating it at a large enough amplitude A [13.38]:

$$kA > \max(-F_{\text{ts}}) . \qquad (13.8)$$

If hysteresis occurs in the $F_{\text{ts}}(z)$-relation, the energy ΔE_{ts} needs to be supplied to the cantilever for each oscillation cycle. If this energy loss is large compared to the intrinsic energy loss of the cantilever, amplitude control can become difficult. An additional approximate criterion for k and A is then

$$\frac{kA^2}{2} \geq \frac{\Delta E_{\text{ts}} Q}{2\pi} . \qquad (13.9)$$

Contribution of Long-Range Forces

The force between tip and sample is composed of many contributions: electrostatic, magnetic, van der Waals, and chemical forces in vacuum. All of these force types except for the chemical forces have strong long-range components that conceal the atomic-force components. For imaging by AFM with atomic resolution, it is desirable to filter out the long-range force contributions and only measure the force components that vary at the atomic scale. While there is no way to discriminate between long- and short-range forces in static AFM, it is possible to enhance the short-range contributions in dynamic AFM through the proper choice of the oscillation amplitude A of the cantilever.

Noise in the Imaging Signal

Measuring the cantilever deflection is subject to noise, especially at low frequencies ($1/f$ noise). In static AFM, this noise is particularly problematic because of the approximate $1/f$ dependence. In dynamic AFM, the low-frequency noise is easily discriminated when using a bandpass filter with a center frequency around f_0.

Non-monotonic Imaging Signal

The tip-sample force is not monotonic. In general, the force is attractive for large distances, and upon decreasing the distance between tip and sample, the force turns repulsive (see Fig. 13.4). Stable feedback is only possible on a monotonic subbranch of the force curve.

Frequency-modulation AFM (FM-AFM) helps to overcome challenges. The non-monotonic imaging signal in AFM is a remaining complication for FM-AFM.

13.1.6 Frequency-Modulation AFM (FM-AFM)

In FM-AFM, a cantilever with eigenfrequency f_0 and spring constant k is subject to controlled positive feedback such that it oscillates with a constant amplitude A [13.32], as shown in Fig. 13.5.

Experimental Setup

The deflection signal is phase shifted, routed through an automatic gain control circuit and fed back to the actuator. The frequency f is a function of f_0, its quality factor Q, and the phase shift ϕ between the mechanical excitation generated at the actuator and the deflection of the cantilever. If $\phi = \pi/2$, the loop oscillates at $f = f_0$.

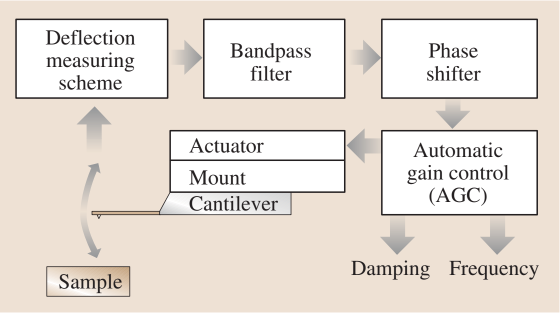

Fig. 13.5 Block diagram of a frequency-modulation force sensor

Three physical observables can be recorded: 1) a change in resonance frequency, Δf, 2) the control signal of the automatic gain control unit as a measure of the tip-sample energy dissipation and 3) an average tunneling current (for conducting cantilevers and tips).

Applications

FM-AFM was introduced by *Albrecht* and coworkers in magnetic force microscopy [13.32]. The noise level and imaging speed were enhanced significantly compared to amplitude modulation techniques. Achieving atomic resolution on the Si(111)-(7×7) surface has been an important step in the development of the STM [13.39], and in 1994, this surface was imaged by AFM with true atomic resolution for the first time [13.33] (see Fig. 13.6).

The initial parameters that provided *true atomic resolution* (see caption of Fig. 13.6) were found empirically. Surprisingly, the amplitude necessary for obtaining good

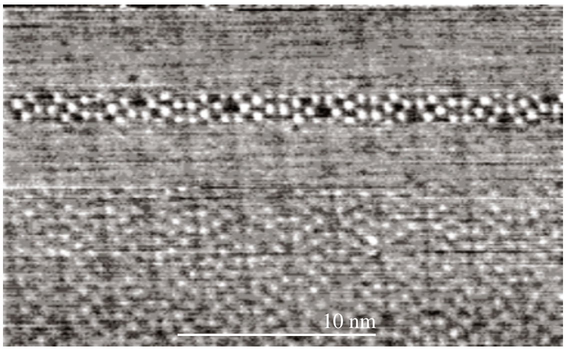

Fig. 13.6 First AFM image of the Si(111)-(7×7) surface. Parameters $k = 17\,\text{Nm}$, $f_0 = 114\,\text{kHz}$, $Q = 28{,}000$, $A = 34\,\text{nm}$, $\Delta f = -70\,\text{Hz}$, $V_{\text{t}} = 0\,\text{V}$

results was very large compared to atomic dimensions. It turned out later that the amplitudes had to be that large in order to fulfill the stability criteria listed in Sect. 13.1.5. Cantilevers with $k \approx 2{,}000\,\mathrm{N/m}$ can be operated with amplitudes in the Å-range [13.24].

13.1.7 Relation Between Frequency Shift and Forces

The cantilever (spring constant k, effective mass m^*) is a macroscopic object and its motion can be described by classical mechanics. Figure 13.7 shows the deflection $q'(t)$ of the tip of the cantilever: It oscillates with an amplitude A at a distance $q(t)$ to a sample.

Generic Calculation

The Hamiltonian of the cantilever is:

$$H = \frac{p^2}{2m^*} + \frac{kq'^2}{2} + V_{\mathrm{ts}}(q)\,, \tag{13.10}$$

where $p = m^*\,\mathrm{d}q'/\mathrm{d}t$. The unperturbed motion is given by:

$$q'(t) = A\cos(2\square f_0 t)\,, \tag{13.11}$$

and the frequency is:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m^*}}\,. \tag{13.12}$$

If the force gradient $k_{\mathrm{ts}} = -\partial F_{\mathrm{ts}}/\partial z = \partial^2 V_{\mathrm{ts}}/\partial z^2$ is constant during the oscillation cycle, the calculation of the frequency shift is trivial:

$$\Delta f = \frac{f_0}{2k} k_{\mathrm{ts}}\,. \tag{13.13}$$

However, in classic FM-AFM, k_{ts} varies in orders of magnitude during one oscillation cycle, and a perturbation approach, as shown below, has to be employed for the calculation of the frequency shift.

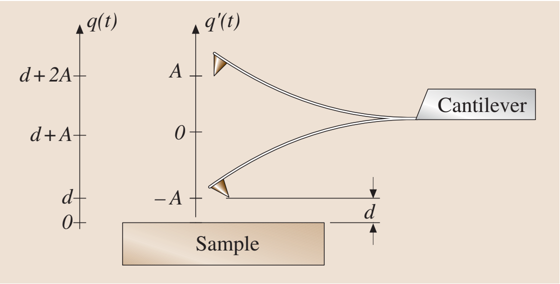

Fig. 13.7 Schematic view of an oscillating cantilever and definition of geometric terms

Hamilton–Jacobi Method

The first derivation of the frequency shift in FM-AFM was achieved in 1997 [13.38] using canonical perturbation theory [13.40]. The result of this calculation is:

$$\begin{aligned}\Delta f &= -\frac{f_0}{kA^2}\langle F_{\mathrm{ts}}q'\rangle \\ &= -\frac{f_0}{kA^2}\int_0^{1/f_0} F_{\mathrm{ts}}(d+A+q'(t))q'(t)\,\mathrm{d}t\,.\end{aligned} \tag{13.14}$$

The applicability of first-order perturbation theory is justified, because in FM-AFM, E is typically in the range of several keVs, while V_{ts} is on the order of a few eVs. *Dürig* [13.41] has found a generalized algorithm that even allows the reconstruction of the tip-sample potential if not only the frequency shift, but the higher harmonics of the cantilever oscillation are known.

A Descriptive Expression for Frequency Shifts as a Function of the Tip-Sample Forces

With integration by parts, the complicated expression (13.14) transforms into a very simple expression that resembles (13.13) [13.42].

$$\Delta f = \frac{f_0}{2k}\int_{-A}^{A} k_{\mathrm{ts}}(z-q')\frac{\sqrt{A^2-q'^2}}{\frac{\pi}{2}kA^2}\,\mathrm{d}q'\,. \tag{13.15}$$

This expression is closely related to (13.13): The constant k_{ts} is replaced by a weighted average, where the weight function $w(q', A)$ is a semicircle with radius A divided by the area of the semicircle $\pi A^2/2$ (see Fig. 13.8). For $A \to 0$, $w(q', A)$ is a representation of Dirac's delta function and the trivial zero-amplitude result of (13.13) is immediately recovered. The frequency shift results from a convolution between the tip-sample force gradient and the weight function. This convolution can easily be reversed with a linear transformation, and the tip-sample force can be recovered from the frequency shift versus distance curve [13.42].

The dependence of the frequency shift with amplitude confirms an empirical conjecture: Small amplitudes increase the sensitivity to short-range forces! Adjusting the amplitude in FM-AFM resembles tuning an optical spectrometer to a passing wavelength. When short-range interactions are to be probed, the amplitude should be in the range of the short-range forces. While using amplitudes in the Å-range has been elusive with conventional cantilevers, because of the instability problems described in Sect. 13.1.5, cantilevers with a stiffness on the order of $1{,}000\,\mathrm{N/m}$ like the ones introduced in [13.23] are well suited for small-amplitude operation.

13.1.8 Noise in Frequency-Modulation AFM – Generic Calculation

The vertical noise in FM-AFM is given by the ratio between the noise in the imaging signal and the slope of the imaging signal with respect to z:

$$\delta z = \frac{\delta \Delta f}{\left|\frac{\partial \Delta f}{\partial z}\right|} . \qquad (13.16)$$

Figure 13.9 shows a typical frequency shift versus distance curve. Because the distance between the tip and sample is measured indirectly through the frequency shift, it is clearly evident from Fig. 13.9 that the noise in the frequency measurement $\delta\Delta f$ translates into vertical noise δz and is given by the ratio between $\delta\Delta f$ and the slope of the frequency shift curve $\Delta f(z)$ (13.16). Low vertical noise is obtained for a low-noise frequency measurement and a steep slope of the frequency shift curve.

The frequency noise $\delta\Delta f$ is typically inversely proportional to the cantilever amplitude A [13.32, 43]. The derivative of the frequency shift versus distance is constant for $A \ll \lambda$, where λ is the range of the tip-sample interaction and proportional to $A^{-1.5}$ for $A \gg \lambda$ [13.38]. Thus, minimal noise occurs if [13.44]

$$A_{\text{optimal}} \approx \lambda , \qquad (13.17)$$

for chemical forces, $\lambda \approx 1$ Å. However, for stability reasons (Sect. 13.1.5), extremely stiff cantilevers are needed for small amplitude operation. The excellent noise performance of the stiff cantilever and small amplitude technique has been verified experimentally [13.24].

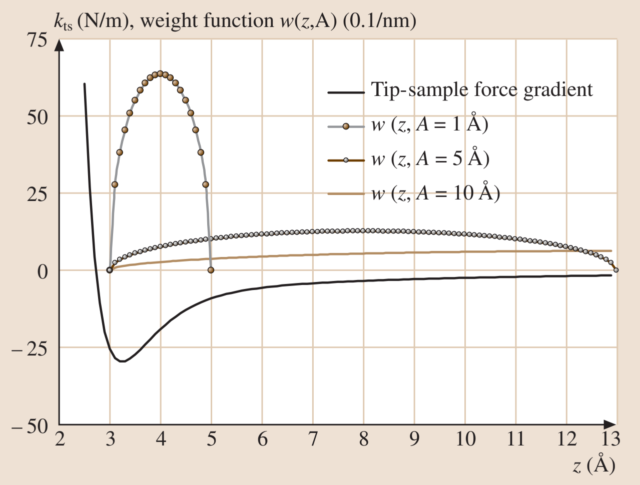

Fig. 13.8 Tip-sample force gradient k_{ts} and weight function for the calculation of the frequency shift

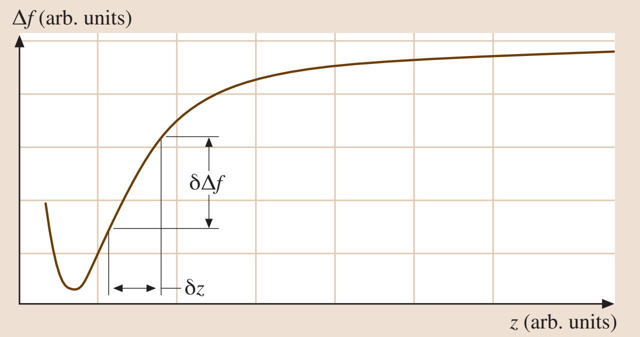

Fig. 13.9 Plot of frequency shift Δf as a function of tip-sample distance z. The noise in the tip-sample distance measurement is given by the noise of the frequency measurement $\delta\Delta f$ divided by the slope of the frequency shift curve

13.1.9 Conclusion

Dynamic force microscopy, in particular frequency-modulation atomic force microscopy, has matured into a viable technique that allows true atomic resolution of conducting and insulating surfaces and spectroscopic measurements on individual atoms [13.10, 45]. Even true atomic resolution in lateral force microscopy is now possible [13.46]. Challenges remain in the chemical composition and structural arrangement of the AFM tip.

13.2 Applications to Semiconductors

For the first time, corner holes and adatoms on Si(111)7×7 surface have been observed at a very local area by *Giessibl* using the pure noncontact AFM in ultrahigh vacuum (UHV) [13.33]. It became the breakthrough for the true atomic resolution imaging on a well-defined clean surface using noncontact AFM. Then, Si(111)7×7 [13.34, 35, 45, 47], InP(110) [13.48], and Si(100)2×1 [13.34] surfaces were successively resolved with true atomic resolution. Furthermore, thermally induced motion of atoms or atomic-scale point defects on InP(110) surface have been observed at room temperature [13.48]. In this section, we will describe typical results of atomically resolved noncontact AFM imaging of semiconductor surfaces.

13.2.1 Si(111)7×7 Surface

Figure 13.10 shows the atomic resolution images of the Si(111)7×7 surface [13.49]. Here, Fig. 13.10a (TYPE I) was obtained using the Si tip without dangling, which is covered with an inert oxide layer. Fig. 13.10b (TYPE II) was obtained using the Si tip with dangling bond on which the Si atoms were deposited, due to the mechanical soft contact between the tip and the Si surface. The variable frequency shift mode was used. We can see not only adatoms and corner holes, but also missing adatoms described by the dimer-adatom-stacking fault (DAS) model. We can see that the image contrast in Fig. 13.10b is clearly stronger than that in Fig. 13.10a.

Interestingly, by using the Si tip with dangling bond, we observed contrast between inequivalent halves and between inequivalent adatoms of the 7×7 unit cell. Namely, as shown in Fig. 13.11a, the faulted halves surrounded with a solid line show brighter than the unfaulted halves surrounded with a broken line. Here, the positions of the faulted and unfaulted halves were decided from the step direction. From the cross-sectional profile along the long diagonal of the 7×7 unit cell in Fig. 13.11b, the heights of the corner adatoms are slightly higher than those of the adjacent center adatoms in the faulted and unfaulted halves of the unit cell. The measured corrugations are in the following decreasing order:

$$\text{Co-F} > \text{Ce-F} > \text{Co-U} > \text{Ce-U}\,,$$

where Co-F and Ce-F indicate the corner and center adatoms in faulted halves, and Co-U and Ce-U indicate the corner and center adatoms in unfaulted halves, respectively. As averaging over several units, the corrugation height differences are estimated to be 0.25 Å, 0.15 Å, and 0.05 Å for Co-F, Ce-F, and Co-U, respectively, referring to Ce-U. This tendency, that the heights of the corner adatoms are higher than those of the center adatoms, is consistent with the experimental results using a silicon tip [13.47], although they could not determine faulted and unfaulted halves of the unit cell in the measured AFM images. However, this tendency is completely contrary to the experimental results using a tungsten tip [13.35]. This difference may originate from the difference in the materials of the tip, which seems to affect the interaction between the tip and the reactive sample surface. Another possibility is that the tip is in contact with the surface during the small fraction of the oscillating cycle in their experiments [13.35].

We consider that contrast between inequivalent adatoms is not caused by tip artifacts for the following reasons: 1) Each adatom, corner hole and defect were clearly observed; 2) the apparent heights of adatoms are the same whether they are located adjacent to defects or not; 3) the same contrast in several images for the different tips has been observed.

It should be noted that the corrugation amplitude of adatoms ∼ 1.4 Å in Fig. 13.11b is higher than that of 0.8–1.0 Å obtained with the STM, although the depth of corner holes obtained with noncontact AFM is almost the same as that observed with STM. Besides, in noncontact mode AFM images, the corrugation ampli-

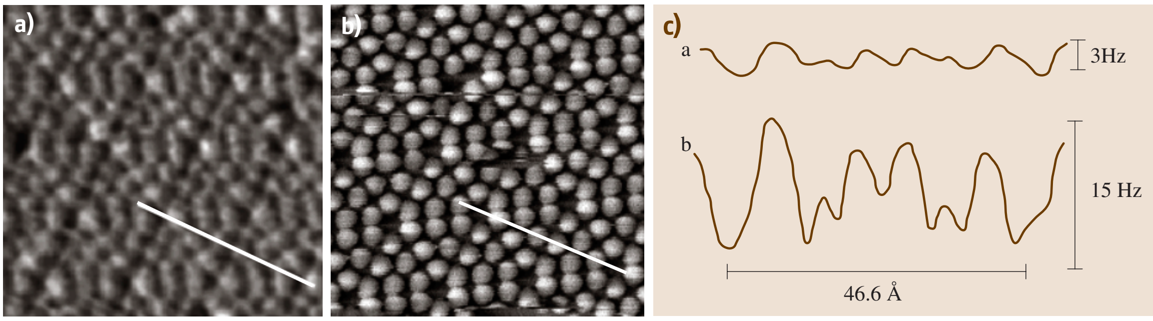

Fig. 13.10a–c Noncontact mode AFM images of Si(111) 7×7 reconstructed surface obtained using the Si tips **(a)** without and **(b)** with dangling bond. The scan area is 99 Å×99 Å. **(c)** The cross sectional profiles along long diagonal of the 7×7 unit cell indicated by *white lines* in **(a)** and **(b)**

tude of adatoms was frequently larger than the depth of the corner holes. The origin of such a large corrugation of adatoms may be due to the effect of the chemical interaction, but that is yet unclear.

The atom positions, surface energies, dynamic properties, and chemical reactivities on the Si(111)7 × 7 reconstructed surface have been extensively investigated theoretically and experimentally. From these investigations, the possible origins of the contrast between inequivalent adatoms in AFM images are the following: the true atomic heights that correspond to the adatom core positions, the stiffness (spring constant) of inter-atomic bonding with the adatoms that corresponds to the frequencies of the surface mode, the amount of charge of adatom, and the chemical reactivity of adatoms. Table 13.1 summarizes the decreasing orders of the inequivalent adatoms for individual property. From Table 13.1, we can see that the calculated adatom heights and the stiffness of inter-atomic bonding cannot explain the AFM data, while the amount of charge of adatom and the chemical reactivity of adatoms can explain our data. The contrast due to the amount of charge of adatom means that the AFM image originated from the difference of the vdW, or the electrostatic physical interactions between the tip and the valence electrons at the adatoms. The contrast due to the chemical reactivity of adatoms means that the AFM image originated from the difference of covalent bonding chemical interaction between the atoms at the tip apex and the dangling bond of adatoms. Thus, we can see there are two possible interactions that explain the strong contrast between inequivalent adatoms of the 7 × 7 unit cell observed using the Si tip with dangling bond.

A weak contrast image in Fig. 13.10a is due to vdW and/or electrostatic force interactions. On the other hand, strong contrast images in Figs. 13.10b and 13.11a are due to a covalent bonding formation between the AFM tip with Si atoms and Si adatoms. These results signify the capability of the noncontact mode AFM to image the variation in chemical reactivity of Si adatoms. In the future, by controlling an atomic species at the tip apex, the study of chemical reactivity on an atomic scale will be possible using the noncontact AFM.

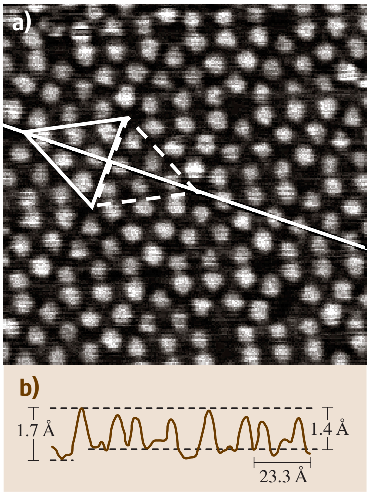

Fig. 13.11 **(a)** Noncontact mode AFM image with contrast of unequivalent adatoms and **(b)** a cross-sectional profile indicated by *white line*. The halves of the 7 × 7 unit cell surrounded with *solid line* and *broken line* correspond to faulted and unfaulted halves, respectively. The scan area is 89 Å × 89 Å

13.2.2 Si(100)2×1 and Si(100)2×1:H Monohydride Surfaces

In order to investigate the imaging mechanism of the noncontact AFM, a comparative study between a reactive surface and an insensitive surface using the same tip is very useful. Si(100)2 × 1:H monohydride surface is a Si(100)2 × 1 reconstructed surface that is terminated by a hydrogen atom. It does not reconstruct as metal is deposited on the semiconductor surface. The surface structure hardly changes. Thus, Si(100)2 × 1:H monohydride surface is one of most useful surfaces for a model system to investigate the imaging mechanism, experimentally and theoretically. Furthermore, whether the interaction between a very small atom such as hy-

Table 13.1 Comparison between the adatom heights observed in an AFM image and the variety of properties for inequivalent adatoms

	Decreasing order	Agreement
AFM image	Co−F > Ce−F > Co−U > Ce−U	–
Calculated height	Co−F > Co−U > Ce−F > Ce−U	×
Stiffness of inter-atomic bonding	Ce−U > Co−U > Ce−F > Co−F	×
Amount of charge of adatom	Co−F > Ce−F > Co−U > Ce−U	○
Calculated chemical reactivity	Faulted > Unfaulted	○
Experimental chemical reactivity	Co−F > Ce−F > Co−U > Ce−U	○

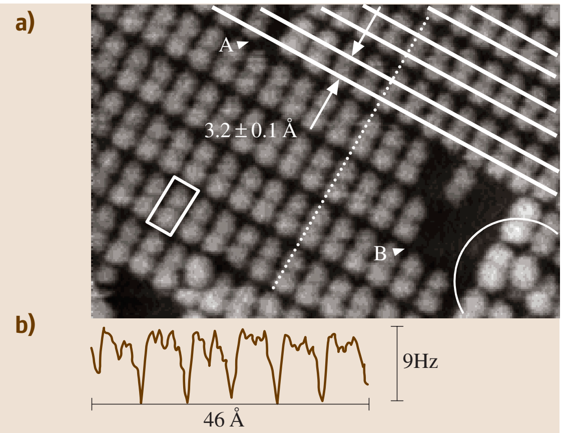

Fig. 13.12 **(a)** Noncontact AFM image of a Si(001)2 × 1 reconstructed surface. The scan area was 69 × 46 Å. One 2 × 1 unit cell is outlined with a *box*. *White rows* are superimposed to show the bright spots arrangement. The distance between bright spots on dimer row is 3.2 ± 0.1 Å. Into *white arc*, the alternative bright spots are shown. **(b)** Cross-sectional profile indicated by a *white dotted line*

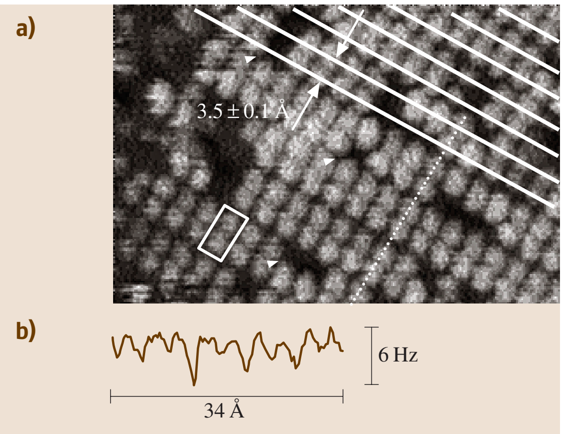

Fig. 13.13 **(a)** Noncontact AFM image of Si(001)2 × 1-H surface. The scan area was 69 × 46 Å. One 2 × 1 unit cell is outlined with a *box*. *White rows* are superimposed to show the bright spots arrangement. The distance between bright spots on dimer row is 3.5 ± 0.1 Å. **(b)** Cross-sectional profile indicated by a *white dotted line*

drogen and a tip apex is observable with noncontact AFM is interesting. Here, we show noncontact AFM images measured on a Si(100)2 × 1 reconstructed surface with a dangling bond and a Si(100)2 × 1:H monohydride surface with a dangling bond that is terminated by a hydrogen atom [13.50].

Figure 13.12a shows the atomic resolution image of a Si(100)2 × 1 reconstructed surface. The paired bright spots arranged like rows with a 2 × 1 symmetry was observed with clear contrast. Missing pairs of bright spots were observed, as indicated by arrows. Further, pairs of bright spots are shown by the white dashed arc and appear to be the stabilize-buckled asymmetric dimer structure. The distance between paired bright spots is 3.2 ± 0.1 Å.

Figure 13.13a shows the atomic resolution image of a Si(100)2 × 1:H monohydride surface. Paired bright spots arranged like rows were observed. Missing paired bright spots, as well as those paired in rows and single bright spots were observed, as indicated by arrows. The distance between paired bright spots is 3.5 ± 0.1 Å. This distance is 0.2 Å larger than that of a Si(100)2 × 1 reconstructed surface. It was found that the distance between bright spots increases in size due to the hydrogen termination.

The bright spots in Fig. 13.12 do not merely image the silicon atom site, because the distance between bright spots forming the dimer structure of Fig. 13.12a (3.2 ± 0.1 Å) is larger than the distance between silicon atoms of every dimer structure model (2.9 Å is the maximum distance between upper silicons on asymmetric dimer structures). This seems to be due to the contribution to imaging of the chemical bonding interaction between the dangling bond out of the silicon tip apex and the dangling bond on the Si(100)2 × 1 reconstructed surface. The chemical bonding interaction works with strong direction dependence between the dangling bond out of the silicon dimer structure on the Si(100)2 × 1 reconstructed surface and the dangling bond out of the silicon tip apex, and the dimer structure is obtained with a larger distance than between silicons on the surface.

The bright spots of Fig. 13.13 seem to be located at hydrogen atom sites on a Si(100)2 × 1:H monohydride surface, because the distance between bright spots forming the dimer structures (3.5 ± 0.1 Å) agrees approximately with the distance between the hydrogen, i. e., 3.52 Å. Thus, noncontact AFM atomically resolves the individual hydrogen atoms on the topmost layer. On this surface, the dangling bond is terminated by a hydrogen atom, and the hydrogen atom on the topmost layer does not have a chemical reactivity. Therefore, the interaction between the hydrogen atom on the top most layer and the silicon tip apex does not contribute to the chemical bonding interaction with strong direction dependence as in a silicon surface, and the bright spots of a noncontact AFM image correspond to the hydrogen atom sites on the topmost layer.

13.2.3 Metal-Deposited Si Surface

In this section, we will introduce the comparative study of force interaction between a Si tip and a metal deposited Si surface and between a metal adsorbed Si tip and a metal-deposited Si surface [13.51, 52]. As for the metal-deposited Si surface, a Si(111)$\sqrt{3}\times\sqrt{3}$-Ag (hereafter referred to as $\sqrt{3}$-Ag) surface was used.

For the $\sqrt{3}$-Ag surface, the honeycomb-chained trimer (HCT) model has been accepted as the appropriate model. As shown in Fig. 13.14, this structure contains Si trimer at the second layer, 0.75 Å below the Ag trimer at the topmost layer. The topmost Ag atoms and the lower Si atoms form covalent bonds. The inter-atomic distances between the nearest neighbor Ag atoms forming Ag trimer and between lower Si atoms forming Si trimer are 3.43 Å and 2.31 Å, respectively. The apexes of Si trimers and Ag trimers face the $[11\bar{2}]$ direction and the direction tilted a little to the $[\bar{1}12]$ direction, respectively.

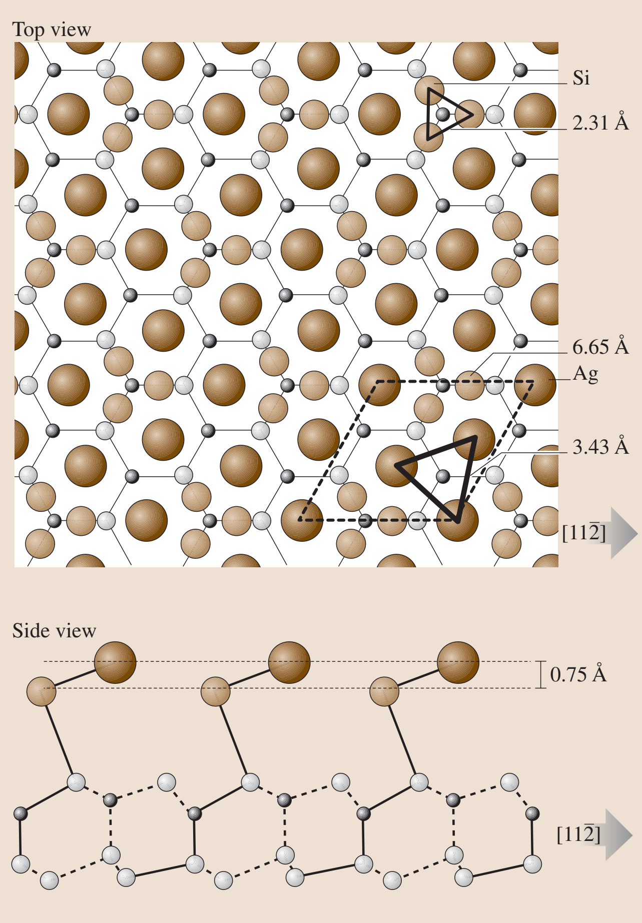

Fig. 13.14 HCT model for the structure of the Si(111)$\sqrt{3}\times\sqrt{3}$-Ag surface. *Black closed circle*, *gray closed circle*, *open circle*, and *closed circle* with *horizontal line* indicate Ag atom at the topmost layer, Si atom at the second layer, Si atom at the third layer, and Si atom at the fourth layer, respectively. *Rhombus* indicates $\sqrt{3}\times\sqrt{3}$ unit cell. *Thick*, *large solid triangle* indicates an Ag trimer. *Thin*, *small, solid triangle* indicates a Si trimer

In Fig. 13.15, we show the noncontact AFM images measured using a normal Si tip at the frequency shift of (a) −37 Hz, (b) −43 Hz, and (c) −51 Hz, respectively. These frequency shifts correspond to the tip-sample distance of about 0–3 Å. We defined the zero position of tip-sample distance, i. e., the contact point, where the vibration amplitude began to decrease. The rhombus indicates the $\sqrt{3}\times\sqrt{3}$ unit cell. When the tip approached the surface, the contrast of the noncontact AFM images became strong and the pattern remarkably changed. That is, by moving the tip toward the sample surface, the hexagonal pattern, the trefoil-like pattern composed of three dark lines, and the triangle pattern can be observed sequentially. In Fig. 13.15a, the distance between the bright spots is 3.9 ± 0.2 Å. In Fig. 13.15c, the distance

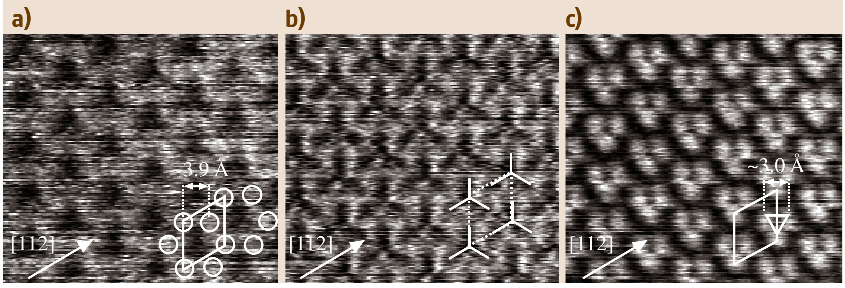

Fig. 13.15a–c Noncontact AFM images obtained at the frequency shifts of **(a)** −37 Hz, **(b)** −43 Hz, and **(c)** −51 Hz on a Si(111)$\sqrt{3}\times\sqrt{3}$-Ag surface. This distance dependence was obtained with the Si tip. Scan area is 38 Å × 34 Å. A *rhombus* indicates $\sqrt{3}\times\sqrt{3}$ unit cell

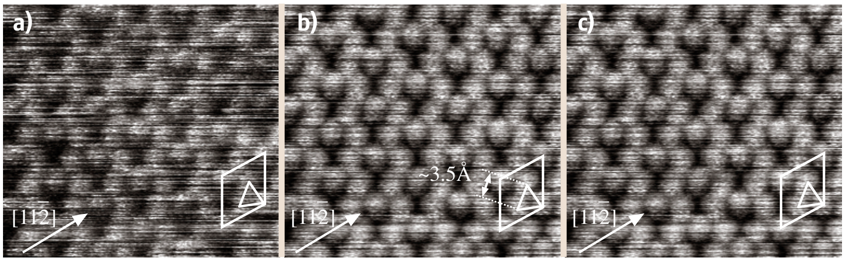

Fig. 13.16a–c Noncontact AFM images obtained at the frequency shifts of **(a)** -4.4 Hz, **(b)** -6.9 Hz, and **(c)** -9.4 Hz on a Si(111)$\sqrt{3}\times\sqrt{3}$-Ag surface. This distance dependence was obtained with the Ag adsorbed tip. Scan area is 38 Å × 34 Å

between the bright spots is 3.0 ± 0.2 Å, and the direction of the apex of all triangles composed of three bright spots is $[11\bar{2}]$.

In Fig. 13.16, we show the noncontact AFM images measured by using an Ag adsorbed tip at the frequency shift of (a) -4.4 Hz, (b) -6.9 Hz, and (c) -9.4 Hz, respectively. The tip-sample distances Z are roughly estimated to be $Z = 1.9$ Å, 0.6 Å, and ~ 0 Å (in noncontact region), respectively. When the tip approached the surface, the pattern of the noncontact AFM images didn't change, although the contrasts became clearer. A triangle pattern can be observed. The distance between the bright spots is 3.5 ± 0.2 Å. The direction of the apex of all triangles composed of three bright spots is tilted a little from the $[\bar{1}\bar{1}2]$ direction.

Thus, noncontact AFM images measured on a Si(111)$\sqrt{3}\times\sqrt{3}$-Ag surface showed two types of distance dependence of the image patterns, depending on the atom species on the tip apex.

By using the normal Si tip with dangling bond, as in Fig. 13.15a, the measured distance between the bright spot of 3.9 ± 0.2 Å agrees with the distance of 3.84 Å between the center of the Ag trimer in the HCT model within the experimental error. Furthermore, the hexagonal pattern composed of six bright spots also agrees with the honeycomb structure of the Ag trimer in the HCT model. So the most appropriate site corresponding to the bright spots in Fig. 13.15a is the site of the center of the Ag trimers. In Fig. 13.15c, the measured distance of 3.0 ± 0.2 Å between the bright spots forming the triangle pattern does not agree with either the distance between the Si trimer of 2.31 Å or the distance between the Ag trimer of 3.43 Å in the HCT model, while the direction of the apex of triangles composed of three bright spots agrees with the $[11\bar{2}]$ direction of the apex of the Si trimer in the HCT model. So the most appropriate site corresponding to the bright spots in Fig. 13.15c is the intermediate site between Si atoms and Ag atoms. On the other hand, by using the Ag adsorbed tip, the measured distance of 3.5 ± 0.2 Å between the bright spots in Fig. 13.16 agrees with the distance of 3.43 Å between the nearest neighbor Ag atoms forming the Ag trimer at the topmost layer in the HCT model within the experimental error. In addition, the direction of the apex of the triangles composed of three bright spots also agrees with the direction of the apex of the Ag trimer, i. e., tilted $[\bar{1}\bar{1}2]$, in the HCT model. So the most appropriate site corresponding to the bright spots in Fig. 13.16 is the site of individual Ag atoms forming the Ag trimer at the topmost layer.

It should be noted that by using the noncontact AFM with an Ag adsorbed tip the individual Ag atom on the $\sqrt{3}$-Ag surface could be resolved in real space for the first time, although it could not be resolved by using the noncontact AFM with a Si tip. So far, the $\sqrt{3}$-Ag surface has been observed by the scanning tunneling microscope (STM) with atomic resolution. However, the STM can measure the local charge density of state near Fermi level on the surface. From the first principle calculation, it was proven that the unoccupied surface state is densely distributed around the center of the Ag trimer. As a result, bright contrast is obtained at the center of the Ag trimer with the STM.

Finally, we consider the origin of the atomic resolution imaging of the individual Ag atoms on the $\sqrt{3}$-Ag surface. Here, we discuss the difference of the force interactions between the Si and the Ag adsorbed tips. As shown in Fig. 13.17a, by using the Si tip, there is the dangling bond out of the topmost Si atom on the Si tip apex. As a result, the force interaction is dominated by the physical bonding interaction such as Coulomb force far from the surface and by chemical bonding interaction very close to the surface. In other words, if the reactive Si tip with dangling bond approaches the surface, at the distance far from the surface, the Coulomb force acts between the electron localized at the dangling bond out of the topmost Si atom on the tip apex and the positive charge distributed around the center of the Ag trimer. At the distance very close to the surface, the chemical bonding interaction will occur due to the onset of the or-

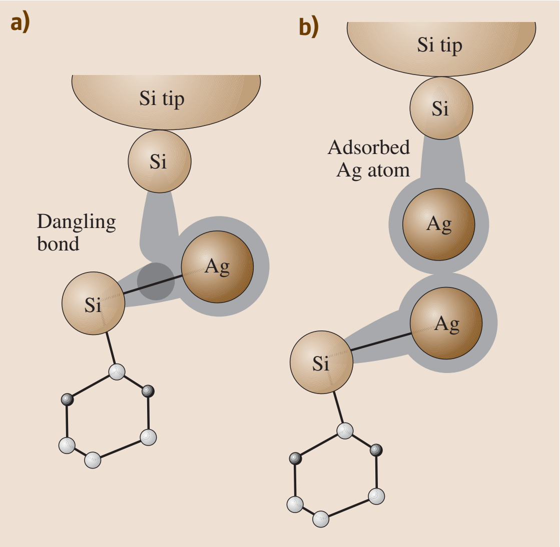

Fig. 13.17a,b Schematic illustration of **(a)** the Si atom with dangling bond and **(b)** the Ag adsorbed tip above the Si-Ag covalent bond on Si(111)$\sqrt{3}\times\sqrt{3}$-Ag surface

bital hybridization between the dangling bond out of the topmost Si atom on the Si tip apex and the Si-Ag covalent bond on the surface. Hence, the individual Ag atoms will not be resolved, the image pattern will change depending on the tip-sample distance. On the other hand, as shown in Fig. 13.17b, by using an Ag adsorbed tip, the dangling bond localized out of the topmost Si atom on the Si tip apex is terminated by the adsorbed Ag atom. As a result, even at very close tip-sample distance, the force interaction is dominated by the physical bonding interaction such as the vdW force interaction. If the Ag adsorbed tip approaches the surface, the vdW force acts between the Ag atom on the tip apex and Ag or Si atom on the surface. Ag atoms at the topmost layer on the $\sqrt{3}$-Ag surface are located higher than the Si atoms at the lower layer. Hence, the individual Ag atoms (or nearly true topography) will be resolved, and the image pattern will not change even at very close tip-sample distance. It should be emphasized that there is a possibility to identify or recognize atom species on a sample surface using the noncontact AFM if we can control an atomic species at the tip apex.

13.3 Applications to Insulators

Insulators such as alkali halides, fluorides, and metal oxides are key materials in many applications, including optics, microelectronics, catalysis, and so on. Surface properties are important in these technologies, but they are usually poorly understood. This is due to their low conductivity, which makes it difficult to investigate them using electron- and ion-based measurement techniques such as low-energy electron diffraction, ion scattering spectroscopy, and scanning tunneling microscopy (STM). Surface imaging by noncontact atomic force microscopy (NC-AFM) does not require the sample to exhibit high conductivity, because NC-AFM detects a force between the tip on the cantilever and the surface of the sample. Since the first report of atomically resolved NC-AFM on a Si(111) 7×7 surface [13.33], several groups have succeeded in obtaining “true” atomic resolution images of insulators, including defects, and it has been shown that NC-AFM is a powerful new tool for the atomic-scale surface investigation of insulators.

In this section, we will describe typical results of atomically resolved NC-AFM imaging of insulators such as alkali halides, fluorides, and metal oxides. For alkali halides and fluorides, we will focus on the contrast formation, which is the most important issue for interpreting atomically resolved images of binary compounds on the basis of experimental and theoretical results. For the metal oxides, typical examples of atomically resolved imaging will be shown. Also, simultaneous imaging using STM and NC-AFM will be described, and the difference between the STM and NC-AFM images will be demonstrated. Finally, we will describe the results obtained when we imaged an antiferromagnetic NiO(001) surface with a ferromagnetic Fe-coated tip in order to explore the possibility of detecting short-range magnetic interaction using the NC-AFM.

13.3.1 Alkali Halides, Fluorides, and Metal Oxides

The surfaces of alkali halides were the first insulating materials to be imaged by NC-AFM with true atomic resolution [13.53]. Up until now, there have been reports on atomically resolved images of (001) cleaved surfaces for single crystal NaF, RbBr, LiF, KI, NaCl [13.54], KBr [13.55], and thin films of NaCl(001) on Cu(111) [13.56]. In this section, we describe the contrast formation of alkali halides surface on the basis of experimental and theoretical results.

Alkali Halides

In experiments of alkali halides, the symmetry of the observed topographic images indicates that the protrusions exhibit only one type of ions, either positively or negatively charged ions. This leads to the conclusion that the atomic contrast is dominantly caused by electrostatic interactions between a charged atom at the tip apex and the surface ions, i. e., the long-range forces between the macroscopic tip and the sample such as van der Waals force is modulated by an alternating short-range electrostatic interaction of the surface ions. Theoretical work employing the atomistic simulation technique has revealed the mechanism for contrast formation on an ionic surface [13.57]. A significant part of the contrast is due to the displacement of ions in the force field, not only enhancing the atomic corrugations, but also contributing to the electrostatic potential by forming dipoles at the surface. The experimentally observed atomic corrugation height is determined by the interplay of the long- and short-range forces. In the case of NaCl, it has been experimentally demonstrated that a more blunt tip produces a larger corrugation when the tip-sample distance is shorter [13.54]. This result shows that the increased long-range forces induced by a blunt tip allow for more stable imaging closer to the surface. The stronger electrostatic short-range interaction and the larger ion displacement produce a more pronounced atomic corrugation.

At steps and kinks of an NaCl thin film on Cu(111), the amplitude of the atomic corrugations has been observed to increase by up to a factor of two, as large as that of atomically flat terraces [13.56]. This increase in the contrast is due to the low coordination of the step-edge and the kinked ions. The low coordination of the ions results in an enhancement of the electrostatic potential over the site and an increase in the displacement induced by the interaction with the tip. These results demonstrate that the short-range interaction dominates the atomic contrast and leads to a considerable displacement of the surface ions from the ideal position.

Theoretical study predicts that the image contrast depends on the chemical species of the tip apex. *Bennewitz* et al. [13.56] have performed the calculations using an MgO tip terminated by oxygen, an Mg ion, and an OH. The magnitude of the atomic contrast for the Mg-terminated tip shows a slight increase in comparison with an oxygen-terminated tip. The atomic contrast with the oxygen-terminated tip is dominated by the attractive electrostatic interaction between the oxygen on the tip apex and the Na ion, but the Mg-terminated tip attractively interacts with the Cl ion. Also, it has been demonstrated that the magnitude of the contrast for an OH-terminated tip is almost an order of magnitude less than for the oxygen- and Mg-terminated tips.

In order to explore the imaging mechanism for the alkali halides, the heterogeneous surface of NaCl crystal containing $(OCN)^-$ molecules as impurities has been observed by NC-AFM [13.58]. In this case, the atomic periodicity of the NaCl lattice is imaged. However, high-resolution investigations of impurities are limited by the convolution between the tip apex and the molecules at the surface. *Bennewitz* et al. [13.59] reported on the heterogeneous surfaces of alkali halide crystals. In this experiment, which was designed to observe the chemically nonhomogeneous alkali halide surface, a mixed crystal composing of 60% KCl and 40% KBr was used as the sample. In this crystal, which is a so-called solid solution, the Cl and Br ions are mixed randomly. The image of the cleaved $KCl_{0.6}Br_{0.4}(001)$ surface indicates that only one type of ion is imaged as protrusions as if it were a pure alkali halide crystal. However, there is a significant difference in comparison with the surface of a pure crystal. The amplitude of atomic corrugation varies strongly between the positions of the ions imaged as depressions. This variation in the corrugations corresponds to the constituents of the crystal, i. e., the Cl and Br ions, and it is concluded that the tip apex is negatively charged. Also, it is demonstrated that a comparison between the distribution of the deep depressions and the composition of the mixed crystal assigns the deep depressions to Br ions. Although an evaluation of the effect of the surface buckling induced by a random distribution of anions is needed in order to understand the contrast mechanism, this result demonstrates that NC-AFM measurement exhibits chemical sensitivity on an atomic scale. For applications to alkali halide surfaces, real space imaging with NC-AFM presents the possibility for chemical discrimination at the surface.

Fluorides

Fluorides are important materials for the progress of atomic-scale resolution NC-AFM imaging on insulators. There are reports in the literature of surface images for single crystal BaF_2, SrF_2 [13.60], and CaF_2 [13.61–63]. Surfaces of fluorite-type crystals are prepared by cleaving along the (111) planes. Their structure is more complex than the structure of alkali halides with the NaCl structure. The complexity is of great interest for atomic-resolution imaging using an NC-AFM and also for theoretical predictions of the interpretation of atomic-scale contrast information.

The first atomically resolved images of a $CaF_2(111)$ surface have been obtained in topographic mode [13.62] in which the tip-sample distance is regulated by maintaining a constant frequency shift. In this imaging mode, the surface ions mostly appear as spherical caps. Similar measurements have been performed on the (111) surfaces of SrF_2 and BaF_2.

Barth et al. [13.64] have found that the $CaF_2(111)$ surface images obtained using the constant height mode, in which the frequency shift is recorded with a very low loop gain, can be categorized into two contrast patterns. In the first, the ions appear as triangles, and in the other, they have the appearance of circles, similar to the contrast obtained in a topographic image. These differences are revealed by comparing the cross sections along the [121] direction. Theoretical studies demonstrated that these two different contrast patterns could be explained as a result of imaging with tips of different polarity [13.64–66]. When imaging with a positively charged (cation-terminated) tip, the triangular pattern appears. In this case, the contrast is dominated by the strong short-range electrostatic attraction between the positive tip and the negative F ions. The cross section along the [121] direction of the triangular image shows two maxima: one is a larger peak over the F ions located in the topmost layer, and the other is a smaller peak at the position of the F ions in the third layer. The minima appear at the position of the Ca ions in the second layer. When imaging with a negatively charged (anion-terminated) tip, the spherical image contrast appears, and the main periodicity is created by the Ca ions in the second layer between the topmost and the third F ion layers. In the cross section along the [121] direction, the large maxima correspond to the Ca sites, because of the strong attraction of the negative tip, and the minima appear at the sites of maximum repulsion over the F ions in the first layer. At a position between two F ions, there are smaller maxima. This reflects the weaker repulsion over the third layer F ion sites compared to the protruding topmost F ion sites and a slower decay in the contrast on one side of the Ca ions. These results demonstrate that the image contrast of a $CaF_2(111)$ surface depends on the state of the tip apex.

The triangular pattern obtained with a positively charged tip appears at relatively large tip-sample distance, as shown in Fig. 13.18a. On the cross section along the [121] direction, experimental results and theoretical studies both demonstrate the large peak and small shoulder characteristic for the triangular pattern image (Fig. 13.18d). When the tip approaches the surface more closely, the triangular pattern of the experimental images is more vivid (Fig. 13.18b), just as predicted in the theoretical works. On the cross section along the [121] direction, as the tip approaches, the amplitude of the shoulder increases until it is equal to that of the main peak, and this feature gives rise to the honeycomb pattern image, as shown Fig. 13.18c. These results demonstrate that there is a more complex process of image contrast formation. Also, from detailed theoretical analysis of the electrostatic potential [13.67], it is suggested that the change in displacement of the ions due to the proximity of the tip plays an important role in the formation of the image contrast. Such a drastic change in image contrast, depending on both the polarity of the terminated tip atom and the tip-sample distance, is inherent in

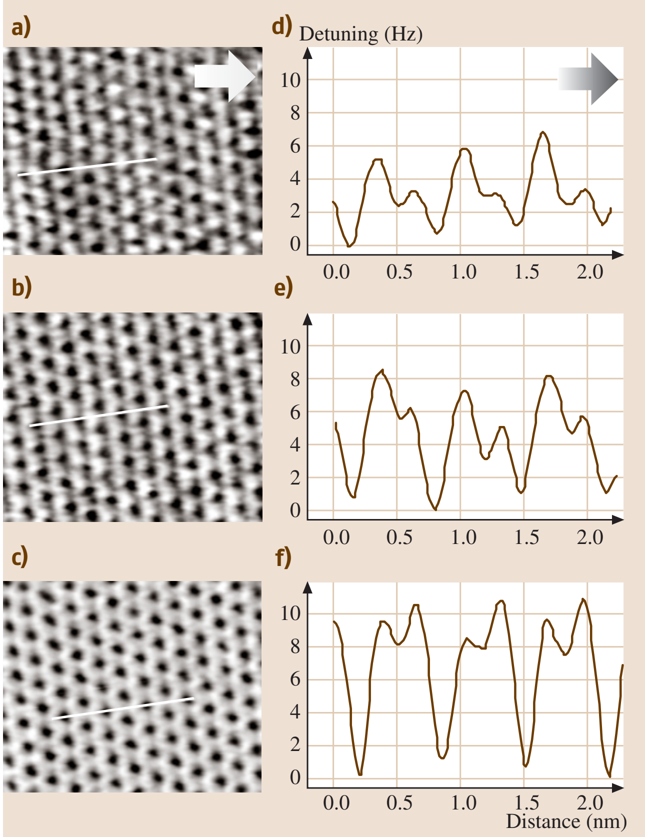

Fig. 13.18 **(a)–(c)** $CaF_2(111)$ surface images obtained by using the constant height mode. From **(a)** to **(c)** the frequency shift was lowered. The *white lines* represent the positions of the cross section. **(d)–(f)** The cross section extracted from the Fourier-filtered images of **(a)–(c)**. The *white* and *black arrows* represent the scanning direction. The images and the cross sections are from [13.64]

the fluoride (111) surface, and this image contrast feature cannot be seen on the (001) surface of alkali halides with a simple crystal structure.

These results have demonstrated that there is good agreement between experiment and theory. However, the results of careful experiments show another feature: The cross sections in the forward and backward scan directions do not yield identical results [13.64]. Also, it is shown that the structures of cross sections taken along the three equivalent [121] directions are different. It is thought that this can be attributed to the asymmetry of the nano-cluster at the tip apex, which leads to different interactions in the equivalent directions. A better understanding of the asymmetric image contrast may require a more complicated modeling of the tip structure. However, the most desirable solution would be the development of the suitable techniques for well-defined tip preparation. In fact, it should be mentioned that perfect tips on an atomic scale can occasionally be obtained. These tips do yield identical results in forward and backward scanning, and cross sections in the three equivalent directions taken with this tip are almost identical [13.68].

The fluoride (111) surface is an excellent standard surface for calibrating tips on an atomic scale. The polarity of the tip-terminated atom can be determined from the image contrast pattern (spherical or triangular pattern). The irregularities in the tip structure can be detected, since the surface structure is highly symmetric. Therefore, once such a tip has been prepared, it can be used as a calibrated tip for imaging unknown surfaces.

The states of the tip apex play an important role in interpreting NC-AFM images of alkali halide and fluoride surfaces. It is expected that the achievement of a good correlation between experimental and theoretical studies will help advance surface imaging for insulators by NC-AFM.

Metal Oxides

Most of the metal oxides that have attracted strong interest for their technological importance are insulative. Therefore, in the case of atomically resolved imaging of metal oxide surfaces by STM, efforts to increase the conductivity of the sample are needed, for example, the introduction of defects of anions or cations, doping with other atoms, and surface observations during heating of the sample. However, in principle, NC-AFM provides the possibility of observing nonconductive metal oxides without these efforts. In cases where the conductivity of the metal oxides is high enough for a tunneling current to flow, it should be noted that most of the surface images obtained by NC-AFM and STM are not identical.

Since the first report of atomically resolved images on a $TiO_2(110)$ surface with oxygen point defects [13.69], they have also been reported on $TiO_2(100)$ [13.70–72], $SnO_2(110)$ [13.73], NiO(001) [13.74, 75], $SrTiO_3(100)$ [13.76], and $CeO_2(111)$ [13.77] surfaces. Also, *Barth* et al. [13.78] have succeeded in obtaining atomically resolved NC-AFM images of a clean α-$Al_2O_3(0001)$ surface, which is impossible to investigate using STM. In this section, we describe typical results of the imaging of metal oxides by NC-AFM.

The α-$Al_2O_3(0001)$ surface exists in several ordered phases that can be reversibly transformed into each other by thermal treatments and oxygen exposure. It is known that the high temperature phase has a large $(\sqrt{31}\times\sqrt{31})R\pm 9°$ unit cell. However, the details of the atomic structure of this surface have not been revealed, and two models are proposed from the results of low-energy electron diffraction and X-ray diffraction studies. Recently, *Barth* et al. [13.78] have directly observed this reconstructed α-$Al_2O_3(0001)$ surface by NC-AFM. They confirmed that the dominant contrast of the low magnification image corresponds to a rhombic grid representing a unit cell of $(\sqrt{31}\times\sqrt{31})R+9°$, as shown in Fig. 13.19a. Also, more details of atomic structures are determined from the higher magnification image (Fig. 13.19b), which was taken at a reduced tip-sample distance. In this atomically resolved image, it is revealed that each side of the rhombus is intersected by 10 atomic rows, and that a hexagonal arrangement of atoms exists in the center of the rhombi (Fig. 13.19c). This feature agrees with the proposed surface structure that predicts order in the center of hexagonal surface domains and disorder at the domain boundaries. Their result is an excellent demonstration of the capabilities of the NC-AFM as a powerful tool for the atomic-scale surface investigation of insulators.

In addition, *Barth* et al. [13.78] have performed observations of the $Al_2O_3(0001)$ surface when it is exposed to water and hydrogen. It is suggested that the region of atomic disorder has an extremely high reactivity. Such investigations of the surface structure of adsorbed insulator substrates can only be provided by NC-AFM. The NC-AFM imaging of metal growth on oxides [13.79] has also been reported. A description of the adsorbed molecular arrangement on a TiO_2 surface can be found in the next section.

The atomic structure of $SrTiO_3(100)$-$(\sqrt{5}\times\sqrt{5})$ $R26.6°$ surface, as well as that of $Al_2O_3(0001)$-$(\sqrt{31}\times\sqrt{31})R\pm 9°$ can be determined on the basis of

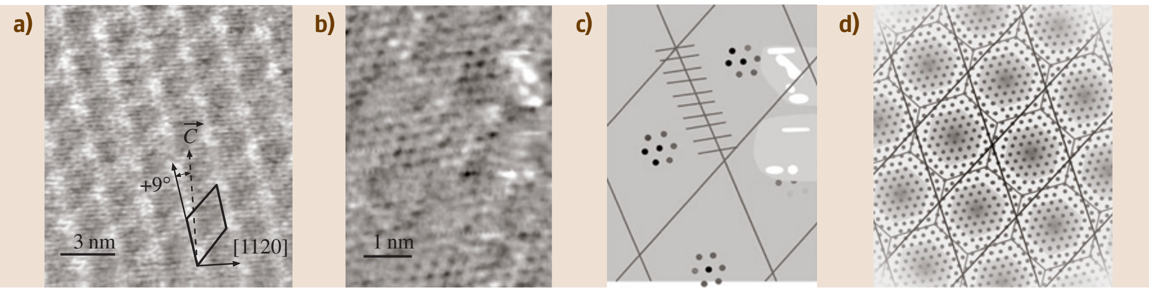

Fig. 13.19 **(a)** Image of the high temperature, reconstructed clean α-Al_2O_3 surface obtained by using the constant height mode. The *rhombus* represents the unit cell of the $(\sqrt{31}\times\sqrt{31})R+9°$ reconstructed surface. **(b)** Higher magnification image of **(a)**. Imaging was performed at a reduced tip-sample distance. **(c)** Schematic representation of the indicating regions of hexagonal order in the center of reconstructed rhombi. **(d)** Superposition of the hexagonal domain with reconstruction rhombi found by NC-AFM imaging. Atoms in *gray shaded regions* are well-ordered. The images and the schematic representations are from [13.78]

the results of NC-AFM imaging [13.76]. $SrTiO_3$ is one of the perovskite oxides and its (100) surface exhibits the many different kinds of reconstructed structures. In the case of the $(\sqrt{5}\times\sqrt{5})R26.6°$ reconstruction, the oxygen vacancy-Ti^{3+}-oxygen model (where the terminated surface is TiO_2 and the observed spots are related to oxygen vacancies) was proposed from the results of STM imaging. As shown in Fig. 13.20, *Kubo* et al. [13.76] have performed measurements using both STM and NC-AFM and have found that the size of the bright spots, as observed by NC-AFM, is always smaller than that for STM measurement and that the dark spots, which are not observed by STM, are arranged along the [001] and [010] directions in the NC-AFM image. A theoretical simulation of the NC-AFM image using first-principles calculations shows that the bright and dark spots correspond to Sr and oxygen atoms, respectively. It has been proposed that the structural model of the reconstructed surface consists of an ordered Sr adatom located at the oxygen fourfold site on the TiO_2-terminated layer (Fig. 13.20c).

Because STM images are related to the spatial distribution of the wave functions near the Fermi level, atoms, which do not have a local density of states near the Fermi level, are generally invisible even on the conductive materials. On the other hand, the NC-AFM image reflects the strength of the tip-sample interaction force originating from chemical, electrostatic interactions, and so on. Therefore, even STM and NC-AFM images obtained using an identical tip and sample will generally differ.

Simultaneous imaging of a $TiO_2(110)$ surface with STM and NC-AFM shows a typical result that differs in terms of the atoms that are visualized [13.72]. For simultaneous imaging, the tip-sample distance is controlled by keeping a constant frequency shift, and the tunneling current between the oscillating cantilever and the surface is measured. The two images show different atomic contrasts, i. e., the atomic corrugation of the

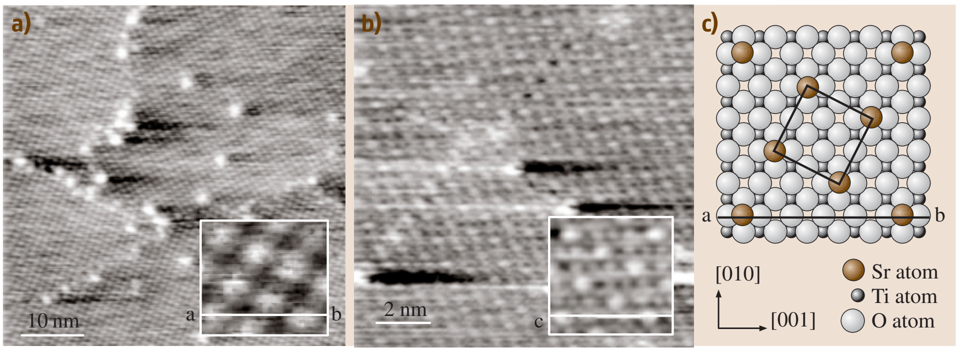

Fig. 13.20 **(a)** STM and **(b)** NC-AFM images of a $SrTiO_3(100)$ surface. **(c)** A proposed model of the $SrTiO_3(100)$-$(\sqrt{5}\times\sqrt{5})R26.6°$ surface reconstruction. The images and the schematic representations are from [13.76]

STM image is in antiphase in the (1×1) and in phase in the (1×2) with respect to the NC-AFM image. In the (1×1), the STM image shows that the dangling bond states at the tip apex overlap with the dangling bonds of the $3d$ states that are protruding from the Ti atom, while the NC-AFM primarily imaged the uppermost oxygen atom. The results of $SrTiO_3$ and TiO_2 surface imaging show that the information obtained by STM and NC-AFM differ, and that the simultaneous imaging of a metal oxide surface enables the investigation of a more detailed surface structure.

13.3.2 Atomically Resolved Imaging of a NiO(001) Surface

The transition metal oxides such as NiO, CoO, and FeO feature the simultaneous existence of an energy gap and unpaired electrons, which gives rise to magnetic properties. Such magnetic insulators are widely used for the exchange biasing in spin valve devices. The NC-AFM enables the direct surface imaging of magnetic insulators on an atomic scale. The forces detected by NC-AFM originate from several kinds of interaction between the surface and the tip, including magnetic interactions in some cases. Theoretical studies predict that short-range magnetic interactions such as the exchange interaction should enable the NC-AFM to image magnetic moments on an atomic scale. In this section, we will describe an imaging of the antiferromagnetic NiO(001) surface using a ferromagnetic Fe-coated tip in order to explore the detection of the short-range magnetic interaction. Also, theoretical studies of the exchange force interaction between a magnetic tip and a sample will be described.

Theoretical Studies of the Exchange Force

The NC-AFM measures various kinds of interaction between tip and sample such as van der Waals, and electrostatic and short-range chemical interactions. In the case of a magnetic tip and sample, the interaction detected by the NC-AFM includes the long-range magnetic dipole interaction and the short-range magnetic interaction. The short-range magnetic interaction between the atoms at the tip apex and the sample surface is governed by a combination of the Coulomb interaction and the Pauli exclusion principle and may be included among the short-range chemical interactions. The magnetic interaction energy depends on the electron spin state of the atoms, and the energy difference in spin alignments (parallel or anti-parallel) is referred to as the exchange interaction energy. Therefore, if the short-range magnetic interaction can be detected, an atomically resolved NC-AFM image, in which the short-range interaction contributes to the atomic protrusions, should reveal the local energy difference in the spin alignment.

In the past, extensive theoretical studies on the short-range magnetic interaction between a ferromagnetic tip and a ferromagnetic sample have been performed by a simple calculation [13.80], a tight-binding approximation [13.81], and first-principles calculations [13.82]. In the calculations performed by *Nakamura* et al. [13.82], three-atomic-layer Fe(001) films, which are separated by a vacuum gap, are used as a model for the tip and sample. The exchange force (F_{ex}) is defined as the difference of the forces (F_P, F_{AP}) in each spin configuration (parallel: P, anti-parallel: AP) of the tip and sample. The tip-sample distance dependency of the exchange force demonstrates that the amplitude of the exchange force is on the order of 10^{-10} N or above and is large enough to be detected by NC-AFM. Also, the site dependency of the exchange force indicates that the discrimination of the exchange force enables the direct imaging of the magnetic moments on an atomic scale. Recently, the interaction between a spin-polarized H atom and a Ni atom on a NiO(001) surface has been investigated theoretically [13.83]. *Foster* and *Shluger* demonstrated that the difference in magnitude of the exchange interaction between opposite spin Ni ions in a NiO surface could be sufficient to be measured in a low temperature NC-AFM experiment. Also, in order to investigate the effect of surface relaxation, they calculated the interaction of different metal tips with the MgO surface used as a model system. The results indicated that the interaction over the oxygen site is stronger than that over the Mg site, and it was pointed out that the tips should be coated in metals that interact weakly with the surface for the detection of the exchange interaction between the atom at the tip apex and the Ni atom on the NiO surface. However, we must take into account another possibility, that is, that a metal atom at the ferromagnetic tip apex may interact with a Ni atom on the second layer through a magnetic interaction mediated by the electrons in an oxygen atom on the surface.

13.3.3 Atomically Resolved Imaging Using Noncoated and Fe-Coated Si Tips

It is thought that the magnetic insulator NiO is one of the best candidates for detecting short-range magnetic interactions such as the exchange interaction, because an experimental configuration of a ferromagnetic tip and an antiferromagnetic sample minimizes the long-range magnetic dipole interaction that might hinder the detection of the short-range interaction.

Below the Néel temperature of 525 K, NiO is antiferromagnetic and has a rhombohedral structure contracted along the $\langle 111\rangle$ axes. The direction of the spin on nickel atoms in the same $\{111\}$ plane is parallel, a so-called ferromagnetic sheet. The direction of spin in the adjacent $\{111\}$ planes is anti-parallel. As a result, the spins at the Ni atoms on the (001) surface are arranged in a checkerboard pattern, i. e., the spin alignment along the $\langle 111\rangle$ direction is antiferromagnetic in order and that along the $\langle 1\,{-1}\,0\rangle$ direction is ferromagnetic in order.

Figure 13.21a shows an atomically resolved image of a NiO(001) surface with a ferromagnetic Fe-coated tip [13.84]. The bright protrusions correspond to atoms spaced about 0.42 nm apart, consistent with the expected periodic arrangement of the NiO(001) surface. The corrugation amplitude is typically 30 pm, which is comparable to the value previously reported [13.74,75], as shown in Fig. 13.21b. Also, because of the symmetry of the image, it was concluded that only one type of atom is visualized by NC-AFM. However, from this image, we think that it is difficult to distinguish which of the atoms are observed, oxygen or nickel. The theoretical works indicate that a metal tip interacts strongly with the oxygen atoms on the MgO(001) surface [13.83]. From this result, it is presumed that the bright protrusions correspond to the oxygen atoms. However, it is still questionable which of the atoms are visible with a Fe-coated tip.

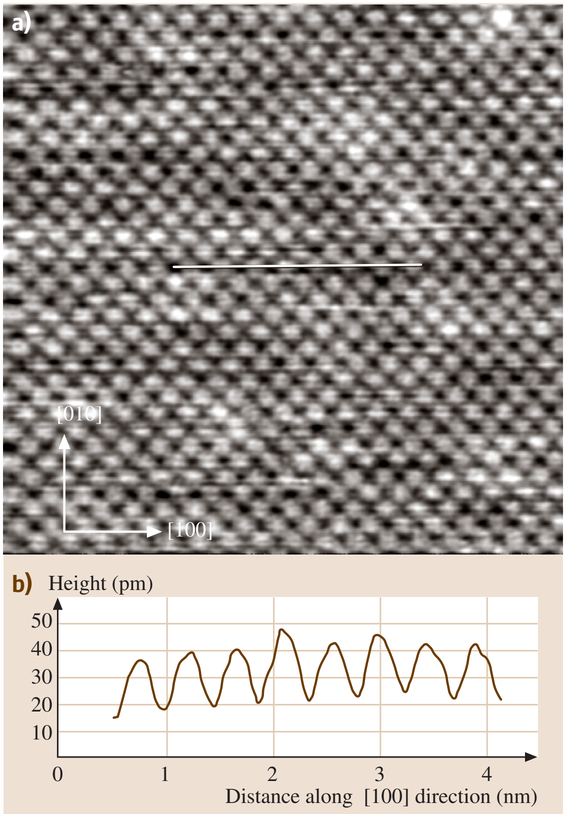

Fig. 13.21 **(a)** Atomically resolved image obtained with an Fe-coated tip. **(b)** shows the cross sections of the middle part in **(a)**. Their corrugations are about 30 pm

If the short-range magnetic interaction is included in the interactions detected by NC-AFM, the corrugation amplitude of the atoms should depend on the direction of the spin at the atom-site. The results of a comparison of the corrugation amplitudes of the same species of atoms with opposite spin can reveal the difference in the short-range magnetic interaction, because the nonmagnetic interactions peculiar to a particular atomic species should be identical. From the results of first-principles calculations [13.82], the contribution of the short-range magnetic interaction to the measured corrugation amplitude is expected to be about a few percent of the total interaction. Discrimination of such small perturbations is therefore needed. In order to reduce the noise, the corrugation amplitude was added on the basis of the periodicity of the NC-AFM image. In addition, the topographical asymmetry, which is the index characterizing the difference in atomic corrugation amplitude, has been defined [13.85]. The result shows that the value of the topographical asymmetry calculated from the image obtained with an Fe-coated Si tip depends on the direction of summing of the corrugation amplitude, and that the dependency corresponds to the antiferromagnetic spin ordering of the NiO(001) surface. With a paramagnetic Si tip, no significant indication of a directional dependency of the topographical asymmetry can be seen. Therefore, it is believed that the results obtained using a Fe-coated tip support the conclusion that the dependency of the topographical asymmetry originates in the short-range magnetic interaction.

The measurements presented here demonstrate the possibility of imaging magnetic structures on an atomic scale by NC-AFM. However, the origin of the detected short-range magnetic interaction is under discussion. Therefore, further experiments such as the measurement of force-distance curves and a theoretical study of the interaction between the ferromagnetic tip and an antiferromagnetic sample are needed.

13.4 Applications to Molecules

In the future, it is expected that electronic, chemical, and medical devices will be downsized to a nanoscale. To achieve this, visualizing and assembling individual molecular components is of fundamental importance. Topographic imaging of nonconductive materials is a challenge for atomic force microscopy (AFM). Nanometer-sized domains of surfactants terminated with different functional groups have been identified by lateral force microscopy (LFM) [13.88] and chemical force microscopy (CFM) [13.89] as extensions of AFM. At a higher resolution, a periodic array of molecules, Langmuir–Blodgett films [13.90] for example, was recognized by AFM. However, it remains difficult to visualize an isolated molecule, molecule vacancy, or the boundary of different periodic domains with a microscope with the tip in contact.

13.4.1 Why Molecules and What Molecules?

The access to individual molecules has not been a trivial task even for noncontact atomic force microscopy (NC-AFM). The force pulling the tip into the surface is less sensitive to the gap width (r), especially when chemically stable molecules cover the surface. The attractive potential between two stable molecules is shallow and exhibits r^{-6} decay [13.13].

High-resolution topography of formate ($HCOO^-$) [13.91] was first reported in 1997 as a molecular adsorbate. The number of imaged molecules is now increasing because of the technological importance of molecular interfaces. To date, the following studies on molecular topography have been published: C_{60} [13.86, 92], DNAs [13.87, 93], adenine and thymine [13.94], alkanethiols [13.94, 95], a perylene-derivative (PTCDA) [13.96], a metal porphyrin (Cu-TBPP) [13.97], glycine sulfate [13.98], polypropylene [13.99], vinylidene fluoride [13.100], and a series of carboxylates ($RCOO^-$) [13.101–107]. Two of these are presented in Figs. 13.22 and 13.23 to demonstrate the current stage of achievement. The proceedings of the annual NC-AFM conference represent a convenient opportunity for us to update the list of molecules imaged.

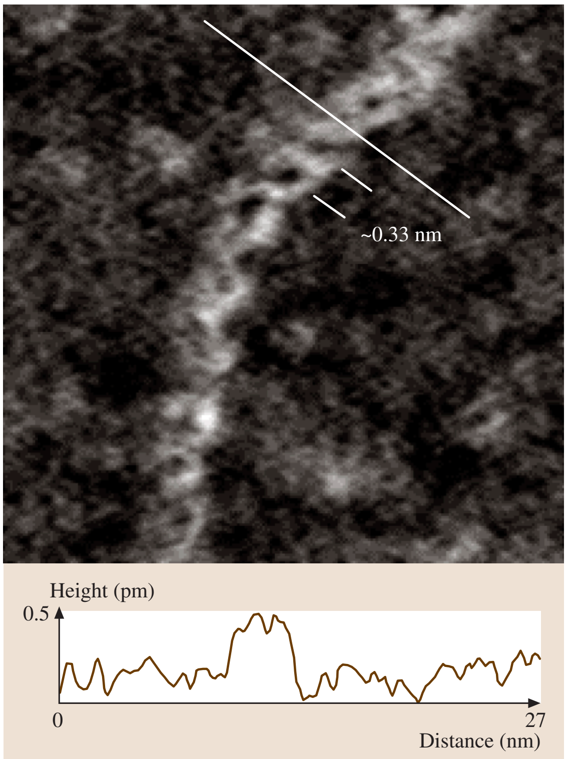

Fig. 13.23 The constant frequency-shift topography of a DNA helix on a mica surface based on [13.87]. Image size: 43 × 43 nm^2. The image revealed features with a spacing of 3.3 nm, consistent with the helix turn of B-DNA

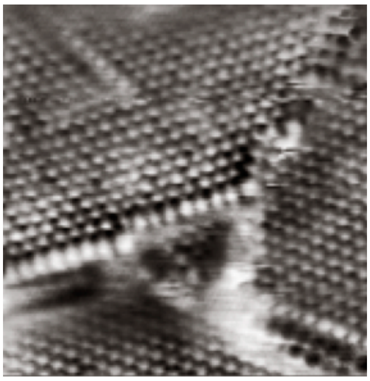

Fig. 13.22 The constant frequency-shift topography of domain boundaries on a C_{60} multilayered film deposited on a Si(111) surface based on [13.86]. Image size: 35 × 35 nm^2

13.4.2 Mechanism of Molecular Imaging

A systematic study on carboxylate ($RCOO^-$) with R = H, CH_3, $C(CH_3)_3$, $C \equiv CH$, and CF_3 revealed that the van der Waals force is responsible for the molecule-dependent microscope topography despite its long-range (r^{-6}) nature. Carboxylates adsorbed on the (110) sur-

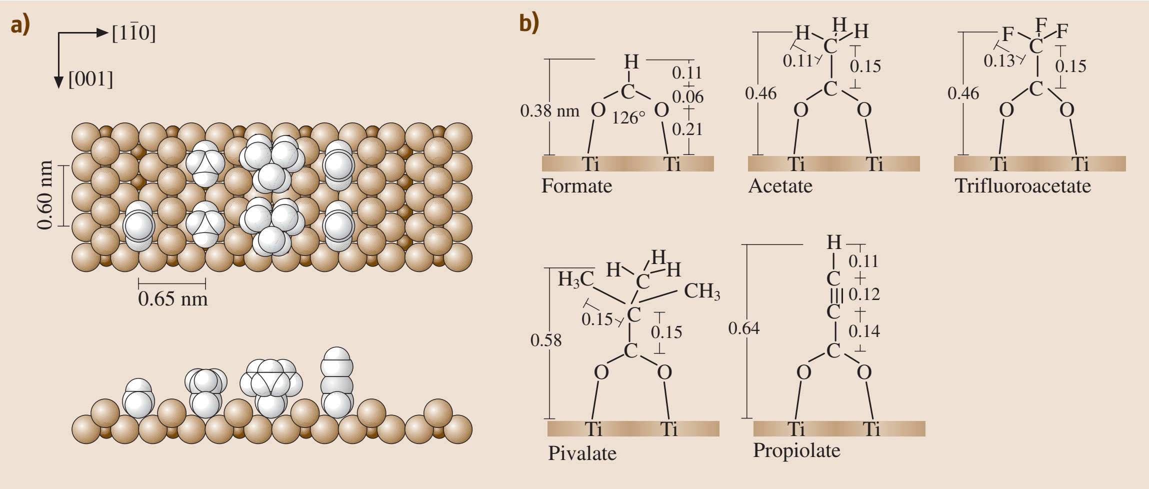
a)
[1$\bar{1}$0]
[001]
0.60 nm
0.65 nm
b)
H
C
126°
0.38 nm
O
O
0.11
0.06
0.21
Ti
Ti
Formate
H H H
C
0.11
0.15
0.46
C
O
O
Ti
Ti
Acetate
F F F
C
0.13
0.15
0.46
C
O
O
Ti
Ti
Trifluoroacetate
H$_3$C
H H H
C
CH$_3$
0.15
C
0.15
0.58
C
O
O
Ti
Ti
Pivalate
H
0.11
C
0.12
C
0.14
0.64
C
O
O
Ti
Ti
Propiolate

Fig. 13.24a,b The carboxylates and TiO_2 substrate. **(a)** Top and side view ball model. *Small shaded* and *large shaded balls* represent Ti and O atoms in the substrate. Protons yielded in the dissociation reaction are not shown. **(b)** Atom geometry of formate, acetate, pivalate, propiolate, and trifluoroacetate adsorbed on the $TiO_2(110)$ surface. The O-Ti distance and $O-C-O$ angle of the formate were determined in the quantitative analysis of photoelectron diffraction [13.109]

face of rutile TiO_2 have been extensively studied as a prototype of organic materials interfaced with an inorganic metal oxide [13.108]. A carboxylic acid molecule (RCOOH) dissociates on this surface to a carboxylate ($RCOO^-$) and a proton (H^+) at room temperature, as illustrated in Fig. 13.24. The pair of negatively charged oxygen atoms in $RCOO^-$ coordinate two positively charged Ti atoms at the surface. The adsorbed carboxylates create a long-range ordered monolayer. The lateral distances of the adsorbates in the ordered monolayer are regulated at 0.65 and 0.59 nm along the [1$\bar{1}$0] and [001] directions. By scanning a mixed monolayer containing different carboxylates, the microscope topography of the terminal groups can be quantitatively compared while minimizing tip-dependent artifacts.

Figure 13.25 presents the observed constant frequency-shift topography of four carboxylates terminated by different alkyl groups. On the formate-covered surface of panel (a) individual formates (R = H) were resolved as protrusions of a uniform brightness. The dark

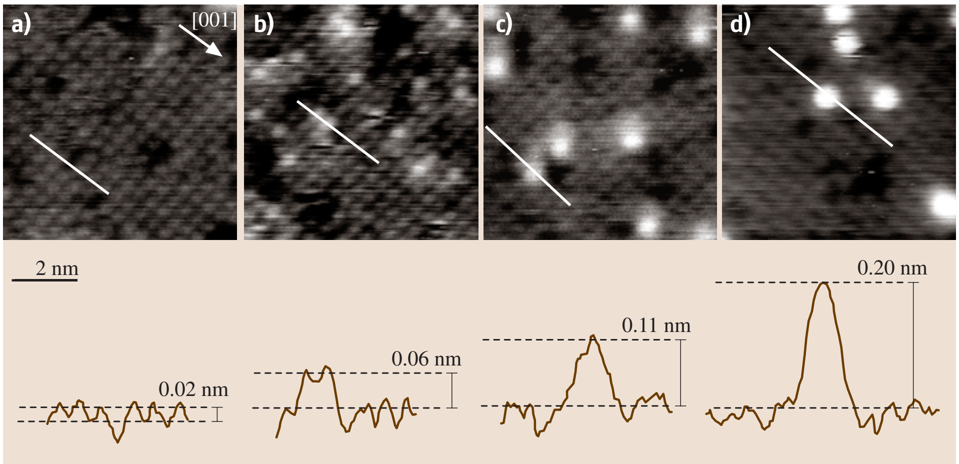
a)
[001]
b)
c)
d)
2 nm
0.02 nm
0.06 nm
0.11 nm
0.20 nm

Fig. 13.25a–d The constant frequency-shift topography of carboxylate monolayers prepared on the $TiO_2(110)$ surface based on [13.102, 104, 106]. Image size: $10 \times 10\,nm^2$. **(a)** Pure formate monolayer; **(b)** formate-acetate mixed layer; **(c)** formate-pivalate mixed layer; **(d)** formate-propiolate mixed layer. Cross sections determined on the *lines* are shown in the lower panel

holes represent unoccupied surface sites. The cross section in the lower panel shows that the accuracy of the height measurement was 0.01 nm or better. Brighter particles appeared in the image when the formate monolayer was exposed to acetic acid (CH_3COOH), as shown in panel (b). Some formates were exchanged with acetates ($R = CH_3$) impinging from the gas phase [13.110]. Because the number of brighter spots increased with exposure time to acetic acid, the brighter particle was assigned to the acetate [13.102]. Twenty-nine acetates and 188 formates were identified in the topography. An isolated acetate and its surrounding formates exhibited an image height difference of 0.06 nm. Pivalate is terminated by bulky $R = C(CH_3)_3$. Nine bright pivalates were surrounded by formates of ordinary brightness in the image of panel (c) [13.104]. The image height difference of an isolated pivalate over the formates was 0.11 nm. Propiolate with $C{\equiv}CH$ is a needle-like adsorbate of single-atom diameter. That molecule exhibited in panel (d) a microscope topography 0.20 nm higher than that of the formate [13.106].

The image topography of formate, acetate, pivalate, and propiolate followed the order of the size of the alkyl groups. Their physical topography can be assumed on the C−C and C−H bond lengths in the corresponding RCOOH molecules in the gas phase [13.111] and is illustrated in Fig. 13.24. The top hydrogen atom of the formate is located 0.38 nm above the surface plane containing the Ti atom pair, while three equivalent hydrogen atoms of the acetate are more elevated at 0.46 nm. The uppermost H atoms in the pivalate are raised by 0.58 nm relative to the Ti plane. The H atom terminating the triple-bonded carbon chain in the propiolate is at 0.64 nm. Figure 13.26 summarizes the observed image heights relative to the formate as a function of the physical height of the topmost H atoms given in the model. The straight line fitted the four observations [13.104]. When the horizontal axis was scaled with other properties (molecular weight, the number of atoms in a molecule, or the number of electrons in valence states), the correlation became poor.

On the other hand, if the tip apex traced the contour of a molecule composed of hard-sphere atoms, the image topography would reproduce the physical topography in a one-to-one ratio, as shown by the broken line in Fig. 13.26. However, the slope of the fitted line was 0.7. A slope of less than unity is interpreted with the long-range nature of the tip-molecule force. The observable frequency shift reflects the sum of the forces between the tip apex and individual molecules. When the tip passes above a tall molecule embedded in short molecules, it is pulled up to compensate for the increased force originating from the tall molecule. Forces between the lifted tip and the short molecules are reduced due to the increased tip-surface distance. Feedback regulation pushes down the probe to restore the lost forces.

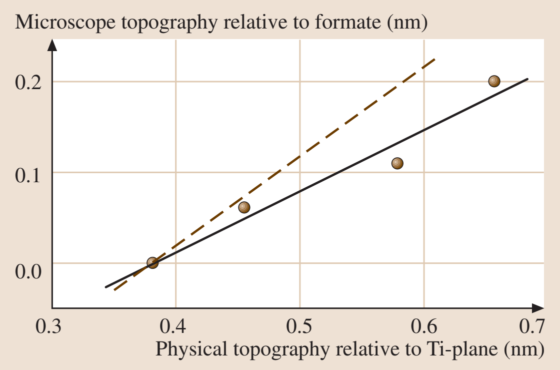

Fig. 13.26 The constant frequency-shift topography of the alkyl-substituted carboxylates as a function of their physical topography given in the model of Fig. 13.3 based on [13.104]

This picture predicts that microscope topography is sensitive to the lateral distribution of the molecules, and that was in fact the case. Two-dimensionally clustered acetates exhibited enhanced image height over an isolated acetate [13.102]. The tip-molecule force, therefore, remained non-zero at distances over the lateral separation of the carboxylates on this surface (0.59−0.65 nm). Chemical bond interactions cannot be important across such a wide tip-molecule gap, whereas atom-scale images of Si(111)-(7×7) are interpreted with the fractional formation of tip-surface chemical bonds [13.24, 45, 49]. Instead, the attractive component of the van der Waals force is probable for the observed molecule-dependent topography. The absence of the tip-surface chemical bond is reasonable on the carboxylate-covered surface terminated with stable C−H bonds.

The attractive component of the van der Waals force contains electrostatic terms caused by permanent-dipole/permanent-dipole coupling, permanent-dipole/induced-dipole coupling, and induced-dipole/induced-dipole coupling (dispersion force). The four carboxylates examined are equivalent in terms of their permanent electric dipole, because the alkyl groups are nonpolar. The image contrast of one carboxylate relative to another is thus ascribed to the dispersion force and/or the force created by the coupling between the permanent dipole on the tip and the induced dipole on the molecule. If we further assume that the Si tip used exhibits the smallest permanent dipole, the dispersion force remains dominant to create the NC-AFM topography dependent on the nonpolar groups of atoms. A numerical simulation on this assumption [13.106] successfully reproduced the propiolate topography of Fig. 13.25d. A calculation that does not include quantum chemical treatments is expected to work, unless the tip approaches the surface too closely, or the molecule possesses a dangling bond.

In addition to the contribution of the dispersion force, the permanent dipole moment of molecules may perturb the microscope topography through electrostatic coupling with the tip. Its possible role was demonstrated by imaging a fluorine-substituted acetate. The strongly polarized C−F bonds were expected to perturb the electrostatic field over the molecule. The constant frequency-shift topography of acetate ($R = CH_3$) and trifluoroacetate ($R = CF_3$) was indeed sensitive to the fluorine substitution. The acetate was observed 0.05 nm higher than the trifluoroacetate [13.103], although the F atoms in the trifluoroacetate, as well as the H atoms in the acetate were lifted by 0.46 nm from the surface plane, as illustrated in Fig. 13.24.

13.4.3 Perspectives

Experimental results summarized in this section prove the feasibility of NC-AFM in identifying individual molecules. A systematic study on the constant frequency-shift topography of carboxylates with $R = H$, CH_3, $C(CH_3)_3$, $C{\equiv}CH$, and CF_3 has revealed the mechanism behind the high-resolution imaging of the chemically stable molecules. The dispersion force is primarily responsible for the molecule-dependent topography. The permanent dipole moment of the imaged molecule, if it exists, perturbs the topography through the electrostatic coupling with the tip. A tiny calculation containing empirical force fields works in simulating the microscope topography.

These results make us optimistic about analyzing physical and chemical properties of nanoscale supramolecular assemblies constructed on a solid surface. If the accuracy of topographic measurement is developed by one more order of magnitude, which is not an unrealistic concept, it may be possible to identify structural isomers, chiral isomers, and conformational isomers of a molecule. Kelvin probe force microscopy (KPFM), an extension of NC-AFM, provides a nanoscale analysis of molecular electronic properties [13.99, 100]. Force spectroscopy with chemically modified tips seems promising for the detection of a selected chemical force. Operation in a liquid atmosphere [13.112] is demanded to observe biochemical materials in the raw.

References

13.1 G. Binnig: Atomic force microscope and method for imaging surfaces with atomic resolution, US Patent 4,724,318 (1986)
13.2 G. Binnig, C. F. Quate, C. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933
13.3 G. Binnig, H. Rohrer, C. Gerber, E. Weibel: Surface studies by scanning tunneling microscopy, Phys. Rev. Lett. **49** (1982) 57–61
13.4 G. Binnig, H. Rohrer: The scanning tunneling microscope, Sci. Am. **253** (1985) 50–56
13.5 G. Binnig, H. Rohrer: In touch with atoms, Rev. Mod. Phys. **71** (1999) S324–S320
13.6 C. J. Chen: *Introduction to Scanning Tunneling Microscopy* (Oxford Univ. Press, Oxford 1993)
13.7 H.-J. Güntherodt, R. Wiesendanger (Eds.): *Scanning Tunneling Microscopy I-III* (Springer, Berlin, Heidelberg 1991)
13.8 J. A. Stroscio, W. J. Kaiser (Eds.): *Scanning Tunneling Microscopy* (Academic, Boston 1993)
13.9 R. Wiesendanger: *Scanning Probe Microscopy and Spectroscopy: Methods and Applications* (Cambridge Univ. Press, Cambridge 1994)
13.10 S. Morita, R. Wiesendanger, E. Meyer (Eds.): *Noncontact Atomic Force Microscopy* (Springer, Berlin, Heidelberg 2002)
13.11 R. Garcia, R. Perez: Dynamic atomic force microscopy methods, Surf. Sci. Rep. **47** (2002) 197–301
13.12 F. J. Giessibl: Advances in atomic force microscopy, Rev. Mod. Phys. **75** (2003) 949–983
13.13 J. Israelachvili: *Intermolecular and Surface Forces*, 2nd edn. (Academic, London 1992)
13.14 L. Olsson, N. Lin, V. Yakimov, R. Erlandsson: A method for *in situ* characterization of tip shape in ac-mode atomic force microscopy using electrostatic interaction, J. Appl. Phys. **84** (1998) 4060–4064
13.15 S. Akamine, R. C. Barrett, C. F. Quate: Improved atomic force microscopy images using cantilevers with sharp tips, Appl. Phys. Lett. **57** (1990) 316–318
13.16 T. R. Albrecht, S. Akamine, T. E. Carver, C. F. Quate: Microfabrication of cantilever styli for the atomic force microscope, J. Vac. Sci. Technol. A **8** (1990) 3386–3396

13.17 M. Tortonese, R. C. Barrett, C. Quate: Atomic resolution with an atomic force microscope using piezoresistive detection, Appl. Phys. Lett. **62** (1993) 834–836

13.18 O. Wolter, T. Bayer, J. Greschner: Micromachined silicon sensors for scanning force microscopy, J. Vac. Sci. Technol. **9** (1991) 1353–1357

13.19 D. Sarid: *Scanning Force Microscopy*, 2nd edn. (Oxford Univ. Press, New York 1994)

13.20 F. J. Giessibl, B. M. Trafas: Piezoresistive cantilevers utilized for scanning tunneling and scanning force microscope in ultrahigh vacuum, Rev. Sci. Instrum. **65** (1994) 1923–1929

13.21 P. Güthner, U. C. Fischer, K. Dransfeld: Scanning near-field acoustic microscopy, Appl. Phys. B **48** (1989) 89–92

13.22 K. Karrai, R. D. Grober: Piezoelectric tip-sample distance control for near field optical microscopes, Appl. Phys. Lett. **66** (1995) 1842–1844

13.23 F. J. Giessibl: High-speed force sensor for force microscopy and profilometry utilizing a quartz tuning fork, Appl. Phys. Lett. **73** (1998) 3956–3958

13.24 F. J. Giessibl, S. Hembacher, H. Bielefeldt, J. Mannhart: Subatomic features on the silicon (111)-(7 × 7) surface observed by atomic force microscopy, Science **289** (2000) 422–425

13.25 F. Giessibl, C. Gerber, G. Binnig: A low-temperature atomic force/scanning tunneling microscope for ultrahigh vacuum, J. Vac. Sci. Technol., B **9** (1991) 984–988

13.26 F. Ohnesorge, G. Binnig: True atomic resolution by atomic force microscopy through repulsive and attractive forces, Science **260** (1993) 1451–1456

13.27 F. J. Giessibl, G. Binnig: True atomic resolution on KBr with a low-temperature atomic force microscope in ultrahigh vacuum, Ultramicroscopy **42–44** (1992) 281–286

13.28 S. P. Jarvis, H. Yamada, H. Tokumoto, J. B. Pethica: Direct mechanical measurement of interatomic potentials, Nature **384** (1996) 247–249

13.29 L. Howald, R. Lüthi, E. Meyer, P. Guthner, H.-J. Güntherodt: Scanning force microscopy on the Si(111)7 × 7 surface reconstruction, Z. Phys. B **93** (1994) 267–268

13.30 L. Howald, R. Lüthi, E. Meyer, H.-J. Güntherodt: Atomic-force microscopy on the Si(111)7 × 7 surface, Phys. Rev. B **51** (1995) 5484–5487

13.31 Y. Martin, C. C. Williams, H. K. Wickramasinghe: Atomic force microscope – force mapping and profiling on a sub 100 Å scale, J. Appl. Phys. **61** (1987) 4723–4729

13.32 T. R. Albrecht, P. Grutter, H. K. Horne, D. Rugar: Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity, J. Appl. Phys. **69** (1991) 668–673

13.33 F. J. Giessibl: Atomic resolution of the silicon (111)-(7 × 7) surface by atomic force microscopy, Science **267** (1995) 68–71

13.34 S. Kitamura, M. Iwatsuki: Observation of silicon surfaces using ultrahigh-vacuum noncontact atomic force microscopy, Jpn. J. Appl. Phys. **35** (1995) L668–L671

13.35 R. Erlandsson, L. Olsson, P. Martensson: Inequivalent atoms and imaging mechanisms in ac-mode atomic-force microscopy of Si(111)7 × 7, Phys. Rev. B **54** (1996) R8309–R8312

13.36 N. Burnham, R. J. Colton: Measuring the nanomechanical and surface forces of materials using an atomic force microscope, J. Vac. Sci. Technol. A **7** (1989) 2906–2913

13.37 D. Tabor, R. H. S. Winterton: Direct measurement of normal and retarded van der Waals forces, Proc. R. Soc. London A **312** (1969) 435

13.38 F. J. Giessibl: Forces and frequency shifts in atomic resolution dynamic force microscopy, Phys. Rev. B **56** (1997) 16010–16015

13.39 G. Binnig, H. Rohrer, C. Gerber, E. Weibel: 7 × 7 reconstruction on Si(111) resolved in real space, Phys. Rev. Lett. **50** (1983) 120–123

13.40 H. Goldstein: *Classical Mechanics* (Addison Wesley, Reading 1980)

13.41 U. Dürig: Interaction sensing in dynamic force microscopy, New J. Phys. **2** (2000) 5.1–5.12

13.42 F. J. Giessibl: A direct method to calculate tip-sample forces from frequency shifts in frequency-modulation atomic force microscopy, Appl. Phys. Lett. **78** (2001) 123–125

13.43 U. Dürig, H. P. Steinauer, N. Blanc: Dynamic force microscopy by means of the phase-controlled oscillator method, J. Appl. Phys. **82** (1997) 3641–3651

13.44 F. J. Giessibl, H. Bielefeldt, S. Hembacher, J. Mannhart: Calculation of the optimal imaging parameters for frequency modulation atomic force microscopy, Appl. Surf. Sci. **140** (1999) 352–357

13.45 M. A. Lantz, H. J. Hug, R. Hoffmann, P. J. A. van Schendel, P. Kappenberger, S. Martin, A. Baratoff, H.-J. Güntherodt: Quantitative measurement of short-range chemical bonding forces, Science **291** (2001) 2580–2583

13.46 F. J. Giessibl, M. Herz, J. Mannhart: Friction traced to the single atom, Proc. Nat. Acad. Sci. USA **99** (2002) 12006–12010

13.47 N. Nakagiri, M. Suzuki, K. Oguchi, H. Sugimura: Site discrimination of adatoms in Si(111)-7 × 7 by noncontact atomic force microscopy, Surf. Sci. Lett. **373** (1997) L329–L332

13.48 Y. Sugawara, M. Ohta, H. Ueyama, S. Morita: Defect motion on an InP(110) surface observed with noncontact atomic force microscopy, Science **270** (1995) 1646–1648

13.49 T. Uchihashi, Y. Sugawara, T. Tsukamoto, M. Ohta, S. Morita: Role of a covalent bonding interaction in noncontact-mode atomic-force microscopy on Si(111)7 × 7, Phys. Rev. B **56** (1997) 9834–9840

13.50 K. Yokoyama, T. Ochi, A. Yoshimoto, Y. Sugawara, S. Morita: Atomic resolution imaging on

$Si(100)2\times1$ and $Si(100)2\times1$-H surfaces using a noncontact atomic force microscope, Jpn. J. Appl. Phys. **39** (2000) L113–L115

13.51 Y. Sugawara, T. Minobe, S. Orisaka, T. Uchihashi, T. Tsukamoto, S. Morita: Non-contact AFM images measured on $Si(111)\sqrt{3}\times\sqrt{3}$-Ag and Ag(111) surfaces, Surf. Interface Anal. **27** (1999) 456–461

13.52 K. Yokoyama, T. Ochi, Y. Sugawara, S. Morita: Atomically resolved Ag imaging on $Si(111)\sqrt{3}\times\sqrt{3}$-Ag surface with noncontact atomic force microscope, Phys. Rev. Lett. **83** (1999) 5023–5026

13.53 M. Bammerlin, R. Lüthi, E. Meyer, A. Baratoff, J. Lü, M. Guggisberg, Ch. Gerber, L. Howald, H.-J. Güntherodt: True atomic resolution on the surface of an insulator via ultrahigh vacuum dynamic force microscopy, Probe Microsc. **1** (1997) 3–7

13.54 M. Bammerlin, R. Lüthi, E. Meyer, A. Baratoff, J. Lü, M. Guggisberg, C. Loppacher, Ch. Gerber, H.-J. Güntherodt: Dynamic SFM with true atomic resolution on alkali halide surfaces, Appl. Phys. A **66** (1998) S293–S294

13.55 R. Hoffmann, M. A. Lantz, H. J. Hug, P. J. A. van Schendel, P. Kappenberger, S. Martin, A. Baratoff, H.-J. Güntherodt: Atomic resolution imaging and force versus distance measurements on KBr(001) using low temperature scanning force microscopy, Appl. Surf. Sci. **188** (2002) 238–244

13.56 R. Bennewitz, A. S. Foster, L. N. Kantorovich, M. Bammerlin, Ch. Loppacher, S. Schär, M. Guggisberg, E. Meyer, A. L. Shluger: Atomically resolved edges and kinks of NaCl islands on Cu(111): experiment and theory, Phys. Rev. B **62** (2000) 2074–2084

13.57 A. I. Livshits, A. L. Shluger, A. L. Rohl, A. S. Foster: Model of noncontact scanning force microscopy on ionic surfaces, Phys. Rev. B **59** (1999) 2436–2448

13.58 R. Bennewitz, M. Bammerlin, Ch. Loppacher, M. Guggisberg, L. Eng, E. Meyer, H.-J. Güntherodt, C. P. An, F. Luty: Molecular impurities at the NaCl(100) surface observed by scanning force microscopy, Rad. Effects Defects Solids **150** (1998) 321–326

13.59 R. Bennewitz, O. Pfeiffer, S. Schär, V. Barwich, E. Meyer, L. N. Kantorovich: Atomic corrugation in NC-AFM of alkali halides, Appl. Surf. Sci. **188** (2002) 232–237

13.60 C. Barth, M. Reichling: Resolving ions and vacancies at step edges on insulating surfaces, Surf. Sci. **470** (2000) L99–L103

13.61 R. Bennewitz, M. Reichling, E. Matthias: Force microscopy of cleaved and electron-irradiated $CaF_2(111)$ surfaces in ultra-high vacuum, Surf. Sci. **387** (1997) 69–77

13.62 M. Reichling, C. Barth: Scanning force imaging of atomic size defects on the $CaF_2(111)$ surface, Phys. Rev. Lett. **83** (1999) 768–771

13.63 M. Reichling, M. Huisinga, S. Gogoll, C. Barth: Degradation of the $CaF_2(111)$ surface by air exposure, Surf. Sci. **439** (1999) 181–190

13.64 C. Barth, A. S. Foster, M. Reichling, A. L. Shluger: Contrast formation in atomic resolution scanning force microscopy of $CaF_2(111)$: experiment and theory, J. Phys. Condens. Matter **13** (2001) 2061–2079

13.65 A. S. Foster, C. Barth, A. L. Shluger, M. Reichling: Unambiguous interpretation of atomically resolved force microscopy images of an insulator, Phys. Rev. Lett. **86** (2001) 2373–2376

13.66 A. S. Foster, A. L. Rohl, A. L. Shluger: Imaging problems on insulators: what can be learnt from NC-AFM modeling on CaF_2?, Appl. Phys. A **72** (2001) S31–S34

13.67 A. S. Foster, A. L. Shluger, R. M. Nieminen: Quantitative modeling in scanning force microscopy on insulators, Appl. Surf. Sci. **188** (2002) 306–318

13.68 M. Reichling, C. Barth: Atomically resoluiton imaging on fluorides. In: *Noncontact Atomic Force Microscopy*, ed. by S. Morita, R. Wiesendanger, E. Meyer (Springer, Berlin, Heidelberg 2002)

13.69 K. Fukui, H. Ohnishi, Y. Iwasawa: Atom-resolved image of the $TiO_2(110)$ surface by noncontact atomic force microscopy, Phys. Rev. Lett. **79** (1997) 4202–4205

13.70 H. Raza, C. L. Pang, S. A. Haycock, G. Thornton: Non-contact atomic force microscopy imaging of $TiO_2(100)$ surfaces, Appl. Surf. Sci. **140** (1999) 271–275

13.71 C. L. Pang, H. Raza, S. A. Haycock, G. Thornton: Imaging reconstructed $TiO_2(100)$ surfaces with noncontact atomic force microscopy, Appl. Surf. Sci. **157** (2000) 223–238

13.72 M. Ashino, T. Uchihashi, K. Yokoyama, Y. Sugawara, S. Morita, M. Ishikawa: STM and atomic-resolution noncontact AFM of an oxygen-deficient $TiO_2(110)$ surface, Phys. Rev. B **61** (2000) 13955–13959

13.73 C. L. Pang, S. A. Haycock, H. Raza, P. J. Møller, G. Thornton: Structures of the 4×1 and 1×2 reconstructions of $SnO_2(110)$, Phys. Rev. B **62** (2000) R7775–R7778

13.74 H. Hosoi, K. Sueoka, K. Hayakawa, K. Mukasa: Atomic resolved imaging of cleaved NiO(100) surfaces by NC-AFM, Appl. Surf. Sci. **157** (2000) 218–221

13.75 W. Allers, S. Langkat, R. Wiesendanger: Dynamic low-temperature scanning force microscopy on nickel oxide (001), Appl. Phys. A **72** (2001) S27–S30

13.76 T. Kubo, H. Nozoye: Surface structure of $SrTiO_3(100)$-$(\sqrt{5}\times\sqrt{5})-R26.6^\circ$, Phys. Rev. Lett. **86** (2001) 1801–1804

13.77 K. Fukui, Y. Namai, Y. Iwasawa: Imaging of surface oxygen atoms and their defect structures on $CeO_2(111)$ by noncontact atomic force microscopy, Appl. Surf. Sci. **188** (2002) 252–256

13.78 C. Barth, M. Reichling: Imaging the atomic arrangements on the high-temperature reconstructed α-Al_2O_3 surface, Nature **414** (2001) 54–57

13.79 C. L. Pang, H. Raza, S. A. Haycock, G. Thornton: Growth of copper and palladium α-Al_2O_3(0001), Surf. Sci. **460** (2000) L510–L514

13.80 K. Mukasa, H. Hasegawa, Y. Tazuke, K. Sueoka, M. Sasaki, K. Hayakawa: Exchange interaction between magnetic moments of ferromagnetic sample and tip: Possibility of atomic-resolution images of exchange interactions using exchange force microscopy, Jpn. J. Appl. Phys. **33** (1994) 2692–2695

13.81 H. Ness, F. Gautier: Theoretical study of the interaction between a magnetic nanotip and a magnetic surface, Phys. Rev. B **52** (1995) 7352–7362

13.82 K. Nakamura, H. Hasegawa, T. Ohuchi, K. Sueoka, K. Hayakawa, K. Mukasa: First-principles calculation of the exchange interaction and the exchange force between magnetic Fe films, Phys. Rev. B **56** (1997) 3218–3221

13.83 A. S. Foster, A. L. Shluger: Spin-contrast in noncontact SFM on oxide surfaces: theoretical modeling of NiO(001) surface, Surf. Sci. **490** (2001) 211–219

13.84 H. Hosoi, K. Kimura, K. Sueoka, K. Hayakawa, K. Mukasa: Non-contact atomic force microscopy of an antiferromagnetic NiO(100) surface using a ferromagnetic tip, Appl. Phys. A **72** (2001) S23–S26

13.85 H. Hosoi, K. Sueoka, K. Hayakawa, K. Mukasa: Atomically resolved imaging of a NiO(001) surface. In: *Noncontact Atomic Force Microscopy*, ed. by S. Morita, R. Wiesendanger, E. Meyer (Springer, Berlin, Heidelberg 2002) pp. 125–134

13.86 K. Kobayashi, H. Yamada, T. Horiuchi, K. Matsushige: Structures and electrical properties of fullerene thin films on Si(111)-7 × 7 surface investigated by noncontact atomic force microscopy, Jpn. J. Appl. Phys. **39** (2000) 3821–3829

13.87 T. Uchihashi, M. Tanigawa, M. Ashino, Y. Sugawara, K. Yokoyama, S. Morita, M. Ishikawa: Identification of B-form DNA in an ultrahigh vacuum by noncontact-mode atomic force microscopy, Langmuir **16** (2000) 1349–1353

13.88 R. M. Overney, E. Meyer, J. Frommer, D. Brodbeck, R. Lüthi, L. Howald, H.-J. Güntherodt, M. Fujihira, H. Takano, Y. Gotoh: Friction measurements on phase-separated thin films with a modified atomic force microscope, Nature **359** (1992) 133–135

13.89 D. Frisbie, L. F. Rozsnyai, A. Noy, M. S. Wrighton, C. M. Lieber: Functional group imaging by chemical force microscopy, Science **265** (1994) 2071–2074

13.90 E. Meyer, L. Howald, R. M. Overney, H. Heinzelmann, J. Frommer, H.-J. Güntherodt, T. Wagner, H. Schier, S. Roth: Molecular-resolution images of Langmuir–Blodgett films using atomic force microscopy, Nature **349** (1992) 398–400

13.91 K. Fukui, H. Onishi, Y. Iwasawa: Imaging of individual formate ions adsorbed on TiO_2(110) surface by non-contact atomic force microscopy, Chem. Phys. Lett. **280** (1997) 296–301

13.92 K. Kobayashi, H. Yamada, T. Horiuchi, K. Matsushige: Investigations of $C_6$0 molecules deposited on Si(111) by noncontact atomic force microscopy, Appl. Surf. Sci. **140** (1999) 281–286

13.93 Y. Maeda, T. Matsumoto, T. Kawai: Observation of single- and double-strand DNA using non-contact atomic force microscopy, Appl. Surf. Sci. **140** (1999) 400–405

13.94 T. Uchihashi, T. Ishida, M. Komiyama, M. Ashino, Y. Sugawara, W. Mizutani, K. Yokoyama, S. Morita, H. Tokumoto, M. Ishikawa: High-resolution imaging of organic monolayers using noncontact AFM, Appl. Surf. Sci **157** (2000) 244–250

13.95 T. Fukuma, K. Kobayashi, T. Horiuchi, H. Yamada, K. Matsushige: Alkanethiol self-assembled monolayers on Au(111) surfaces investigated by non-contact AFM, Appl. Phys. A **72** (2001) S109–S112

13.96 B. Gotsmann, C. Schmidt, C. Seidel, H. Fuchs: Molecular resolution of an organic monolayer by dynamic AFM, Europ. Phys. J. B **4** (1998) 267–268

13.97 Ch. Loppacher, M. Bammerlin, M. Guggisberg, E. Meyer, H.-J. Güntherodt, R. Lüthi, R. Schlittler, J. K. Gimzewski: Forces with submolecular resolution between the probing tip and Cu-TBPP molecules on Cu(100) observed with a combined AFM/STM, Appl. Phys. A **72** (2001) S105–S108

13.98 L. M. Eng, M. Bammerlin, Ch. Loppacher, M. Guggisberg, R. Bennewitz, R. Lüthi, E. Meyer, H.-J. Güntherodt: Surface morphology, chemical contrast, and ferroelectric domains in TGS bulk single crystals differentiated with UHV non-contact force microscopy, Appl. Surf. Sci. **140** (1999) 253–258

13.99 S. Kitamura, K. Suzuki, M. Iwatsuki: High resolution imaging of contact potential difference using a novel ultrahigh vacuum non-contact atomic force microscope technique, Appl. Surf. Sci. **140** (1999) 265–270

13.100 H. Yamada, T. Fukuma, K. Umeda, K. Kobayashi, K. Matsushige: Local structures and electrical properties of organic molecular films investigated by non-contact atomic force microscopy, Appl. Surf. Sci **188** (2000) 391–398

13.101 K. Fukui, Y. Iwasawa: Fluctuation of acetate ions in the (2 × 1)-acetate overlayer on TiO_2(110)-(1 × 1) observed by noncontact atomic force microscopy, Surf. Sci. **464** (2000) L719–L726

13.102 A. Sasahara, H. Uetsuka, H. Onishi: Single-molecule analysis by non-contact atomic force microscopy, J. Phys. Chem. B **105** (2001) 1–4

13.103 A. Sasahara, H. Uetsuka, H. Onishi: NC-AFM topography of HCOO and CH_3COO molecules co-adsorbed on TiO_2(110), Appl. Phys. A **72** (2001) S101–S103

13.104 A. Sasahara, H. Uetsuka, H. Onishi: Image topography of alkyl-substituted carboxylates observed by noncontact atomic force microscopy, Surf. Sci. **481** (2001) L437–L442

13.105 A. Sasahara, H. Uetsuka, H. Onishi: Noncontact atomic force microscope topography dependent on

permanent dipole of individual molecules, Phys. Rev. B **64** (2001) 121406(R)
13.106 A. Sasahara, H. Uetsuka, T. Ishibashi, H. Onishi: A needle-like organic molecule imaged by noncontact atomic force microscopy, Appl. Surf. Sci. **188** (2002) 265–271
13.107 H. Onishi, A. Sasahara, H. Uetsuka, T. Ishibashi: Molecule-dependent topography determined by noncontact atomic force microscopy: Carboxylates on $TiO_2(110)$, Appl. Surf. Sci. **188** (2002) 257–264
13.108 H. Onishi: Carboxylates adsorbed on $TiO_2(110)$. In: *Chemistry of Nano-molecular Systems*, ed. by T. Nakamura (Springer, Berlin, Heidelberg 2002) pp. 75–89
13.109 S. Thevuthasan, G. S. Herman, Y. J. Kim, S. A. Chambers, C. H. F. Peden, Z. Wang, R. X. Ynzunza, E. D. Tober, J. Morais, C. S. Fadley: The structure of formate on $TiO_2(110)$ by scanned-energy and scanned-angle photoelectron diffraction, Surf. Sci. **401** (1998) 261–268
13.110 H. Uetsuka, A. Sasahara, A. Yamakata, H. Onishi: Microscopic identification of a bimolecular reaction intermediate, J. Phys. Chem. B **106** (2002) 11549–11552
13.111 D. R. Lide: *Handbook of Chemistry and Physics*, 81st edn. (CRC, Boca Raton 2000)
13.112 K. Kobayashi, H. Yamada, K. Matsushige: Dynamic force microscopy using FM detection in various environments, Appl. Surf. Sci. **188** (2002) 430–434

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

14. Low Temperature Scanning Probe Microscopy

Low Tempera

This chapter is dedicated to scanning probe microscopy, one of the most important techniques in nanotechnology. In general, scanning probe techniques allow the measurement of physical properties down to the nanometer scale. Some techniques, such as the scanning tunneling microscope and the scanning force microscope even go down to the atomic scale. The properties that are accessible are various. Most importantly, one can image the arrangement of atoms on conducting surfaces by scanning tunneling microscopy and on insulating substrates by scanning force microscopy. But also the arrangement of electrons (scanning tunneling spectroscopy), the force interaction between different atoms (scanning force spectroscopy), magnetic domains (magnetic force microscopy), the local capacitance (scanning capacitance microscopy), the local temperature (scanning thermo microscopy), and local light-induced excitations (scanning near-field microscopy) can be measured with high spatial resolution. In addition, some techniques even allow the manipulation of atomic configurations.

Probably the most important advantage of the low-temperature operation of scanning probe techniques is that they lead to a significantly better signal-to-noise ratio than measuring at room temperature. This is why many researchers work below 100 K. However, there are also physical reasons to use low-temperature equipment. For example, the manipulation of atoms or scanning tunneling spectroscopy with high energy resolution can only be realized at low temperatures. Moreover, some physical effects such as superconductivity or the Kondo effect are restricted to low temperatures. Here, we describe the design criteria of low-temperature scanning probe equipment and summarize some of the most spectacular results achieved since the invention of the method about 20 years ago. We first focus on the scanning tunneling microscope, giving examples of atomic manipulation and the analysis of electronic properties in different material arrangements. Afterwards, we describe results obtained by scanning force microscopy, showing atomic-scale imaging on insulators, as well as force spectroscopy analysis. Finally, the magnetic force microscope, which images domain patterns in ferromagnets and vortex patterns in superconductors, is discussed. Although this list is far from complete, we feel that it gives an adequate impression of the fascinating possibilities of low-temperature scanning probe instruments.

In this chapter low temperatures are defined as lower than about 100 K and are normally achieved by cooling with liquid nitrogen or liquid helium. Applications in which SPMs are operated close to 0 °C are not covered in this chapter.

14.1 **Microscope Operation at Low Temperatures** 414
  14.1.1 Drift 414
  14.1.2 Noise 415
  14.1.3 Stability 415
  14.1.4 Piezo Relaxation and Hysteresis 415

14.2 **Instrumentation** 415
  14.2.1 A Simple Design for a Variable Temperature STM 416
  14.2.2 A Low Temperature SFM Based on a Bath Cryostat 417

14.3 **Scanning Tunneling Microscopy and Spectroscopy** 419
  14.3.1 Atomic Manipulation 419
  14.3.2 Imaging Atomic Motion 420
  14.3.3 Detecting Light from Single Atoms and Molecules 421
  14.3.4 High Resolution Spectroscopy 422
  14.3.5 Imaging Electronic Wave Functions 427
  14.3.6 Imaging Spin Polarization: Nanomagnetism 431

14.4 **Scanning Force Microscopy and Spectroscopy** 433
  14.4.1 Atomic-Scale Imaging 434
  14.4.2 Force Spectroscopy 436
  14.4.3 Electrostatic Force Microscopy 438
  14.4.4 Magnetic Force Microscopy 439

References 442

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

More than two decades ago, the first design of an experimental setup was presented where a sharp tip was systematically scanned over a sample surface in order to obtain local information on the tip-sample interaction down to the atomic scale. This original instrument used the tunneling current between a conducting tip and a conducting sample as a feedback signal and was named *scanning tunneling microscope* accordingly [14.1]. Soon after this historic breakthrough, it became widely recognized that virtually any type of tip-sample interaction can be used to obtain local information on the sample by applying the same general principle, provided that the selected interaction was reasonably short-ranged. Thus, a whole variety of new methods has been introduced, which are denoted collectively as *scanning probe methods*. An overview is given by *Wiesendanger* [14.2].

The various methods, especially the above mentioned scanning tunneling microscopy (STM) and scanning force microscopy (SFM) – which is often further classified into subdisciplines such as the topography-reflecting atomic force microscopy (AFM), the magnetic force microscopy (MFM), or the electrostatic force microscopy (EFM) – have been established as standard methods for surface characterization on the nanometer scale. The reason is that they feature extremely high resolution (often down to the atomic scale for STM and AFM), despite a principally simple, compact, and comparatively inexpensive design.

A side effect of the simple working principle and the compact design of many scanning probe microscopes (SPMs) is that they can be adapted to different environments such as air, all kinds of gaseous atmospheres, liquids, or vacuum with reasonable effort. Another advantage is their ability to work within a wide temperature range. A microscope operation at higher temperatures is chosen to study surface diffusion, surface reactivity, surface reconstructions that only manifest at elevated temperatures, high-temperature phase transitions, or to simulate conditions as they occur, e.g., in engines, catalytic converters, or reactors. Ultimately, the upper limit for the operation of an SPM is determined by the stability of the sample, but thermal drift, which limits the ability to move the tip in a controlled manner over the sample, as well as the depolarization temperature of the piezoelectric positioning elements might further restrict successful measurements.

On the other hand, low-temperature (LT) application of SPMs is much more widespread than operation at high temperatures. Essentially five reasons make researchers adapt their experimental setups to low-temperature compatibility. These are: (1) the reduced thermal drift, (2) lower noise levels, (3) enhanced stability of tip and sample, (4) the reduction in piezo hysteresis/creep, and (5) probably the most obvious, the fact that many physical effects are restricted to low temperature. Reasons (1) to (4) only apply unconditionally if the whole microscope body is kept at low temperature (typically in or attached to a bath cryostat, see Sect. 14.2). Setups in which only the sample is cooled may show considerably less favorable operating characteristics. As a result of (1) to (4), ultrahigh resolution and long-term stability can be achieved on a level that significantly exceeds what can be accomplished at room temperature even under the most favorable circumstances. Typical examples for (5) are superconductivity [14.3] and the Kondo effect [14.4].

14.1 Microscope Operation at Low Temperatures

Nevertheless, before we devote ourselves to a small overview of experimental LTSPM work, we will take a closer look at the specifics of microscope operation at low temperatures, including a discussion of the corresponding instrumentation.

14.1.1 Drift

Thermal drift originates from thermally activated movements of the individual atoms, which are reflected by the thermal expansion coefficient. At room temperature, typical values for solids are on the order of $(1{-}50)\times10^{-6}\,\mathrm{K}^{-1}$. If the temperature could be kept precisely constant, any thermal drift would vanish, regardless of the absolute temperature of the system. The close coupling of the microscope to a large temperature bath that keeps a constant temperature ensures a significant reduction in thermal drift and allows for distortion-free, long-term measurements. Microscopes that are efficiently attached to sufficiently large bath cryostats, therefore, show a one- to two-order-of-magnitude increase in thermal stability compared with non-stabilized setups operated at room temperature.

A second effect also helps suppress thermally induced drift of the probing tip relative to a specific location on the sample surface. The thermal expansion coefficients are at liquid helium temperature two or more orders of magnitude smaller than at room temperature. Consequently, the thermal drift during low-temperature operation decreases accordingly.

For some specific scanning probe methods, there may be additional ways a change in temperature can affect the quality of the data. In *frequency-modulation SFM* (FM-SFM), for example, the measurement principle relies on the accurate determination of the eigenfrequency of the cantilever, which is determined by its spring constant and its effective mass. However, the spring constant changes with temperature due to both thermal expansion (i. e., the resulting change in the cantilever dimensions) and the variation of Young's modulus with temperature. Assuming drift rates of about 2 mK/min, as is typical for room temperature measurements, this effect might have a significant influence on the obtained data.

14.1.2 Noise

The theoretically achievable resolution in SPMs often increases with decreasing temperature due to a decrease in thermally induced noise. An example is the thermal noise in SFMs, which is proportional to the square root of the temperature [14.5, 6]. Lowering the temperature from $T = 300\,\mathrm{K}$ to $T = 10\,\mathrm{K}$ thus results in a reduction of the thermal frequency noise of more than a factor of five. Graphite, e.g., has been imaged atomically resolved only at low temperatures due to its extremely low corrugation, which was below the room temperature noise level [14.7, 8].

Another, even more striking, example is the spectroscopic resolution in *scanning tunneling spectroscopy* (STS). It depends linearly on the temperature [14.2] and is consequently reduced even more at LT than the thermal noise in AFM. This provides the opportunity to study structures or physical effects not accessible at room temperature such as spin and Landau levels in semiconductors [14.9].

Finally, it might be worth mentioning that the enhanced stiffness of most materials at low temperatures (increased Young's modulus) leads to a reduced coupling to external noise. Even though this effect is considered small [14.6], it should not be ignored.

14.1.3 Stability

There are two major stability issues that considerably improve at low temperature. First, low temperatures close to the temperature of liquid helium inhibit most of the thermally activated diffusion processes. As a consequence, the sample surfaces show a significantly increased long-term stability, since defect motion or adatom diffusion is massively suppressed. Most strikingly, even single xenon atoms deposited on suitable substrates can be successfully imaged [14.10, 11], or even manipulated [14.12]. In the same way, the low temperatures also stabilize the atomic configuration at the tip end by preventing sudden jumps of the most loosely bound foremost tip atom(s). Second, the large cryostat that usually surrounds the microscope acts as an effective cryo pump. Thus samples can be kept clean for several weeks, which is a multiple of the corresponding time at room temperature (about 3–4 h).

14.1.4 Piezo Relaxation and Hysteresis

The last important benefit from low-temperature operation of SPMs is that artifacts given by the response of the piezoelectric scanners are substantially reduced. After applying a voltage ramp to one electrode of a piezoelectric scanner, its immediate initial deflection, l_0, is followed by a much slower relaxation, Δl, with a logarithmic time dependence. This effect, known as piezo relaxation or "creep", diminishes substantially at low temperatures, typically by a factor of ten or more. As a consequence, piezo nonlinearities and piezo hysteresis decrease accordingly. Additional information is given by *Hug* et al. [14.13].

14.2 Instrumentation

The two main design criteria for all vacuum-based scanning probe microscope systems are (1) to provide an efficient decoupling of the microscope from the vacuum system and other sources of external vibrations, and (2) to avoid most internal noise sources through a high mechanical rigidity of the microscope body itself. In vacuum systems designed for low-temperature applications, a significant degree of complexity is added, since, on the one hand, close thermal contact of the SPM and cryogen is necessary to ensure the (approximately) drift-free conditions described above, while, on the other hand, a good vibration isolation (both from the outside world, as well as from the boiling or flowing cryogen) has to be maintained.

Plenty of microscope designs have been presented in the last ten to 15 years, predominantly in the field of STM. Due to the variety of the different approaches, we will, somewhat arbitrarily, give two examples on different levels of complexity that might serve as illustrative model designs.

14.2.1 A Simple Design for a Variable Temperature STM

A simple design for a variable temperature STM system is presented in Fig. 14.1 (Similar systems are also offered by Omicron (Germany) or Jeol (Japan)). It should give an impression of what the minimum requirements are, if samples are to be investigated successfully at low temperatures. It features a single ultrahigh vacuum (UHV) chamber that houses the microscope in its center. The general idea to keep the setup simple is that only the sample is cooled by means of a flow cryostat that ends in the small liquid nitrogen (LN) reservoir. This reservoir is connected to the sample holder with copper braids. The role of the copper braids is to thermally attach the LN reservoir to the sample located on the sample holder in an effective manner, while vibrations due to the flow of the cryogen should be blocked as much as possible. In this way, a sample temperature of about 100 K is reached. Alternatively, with liquid helium

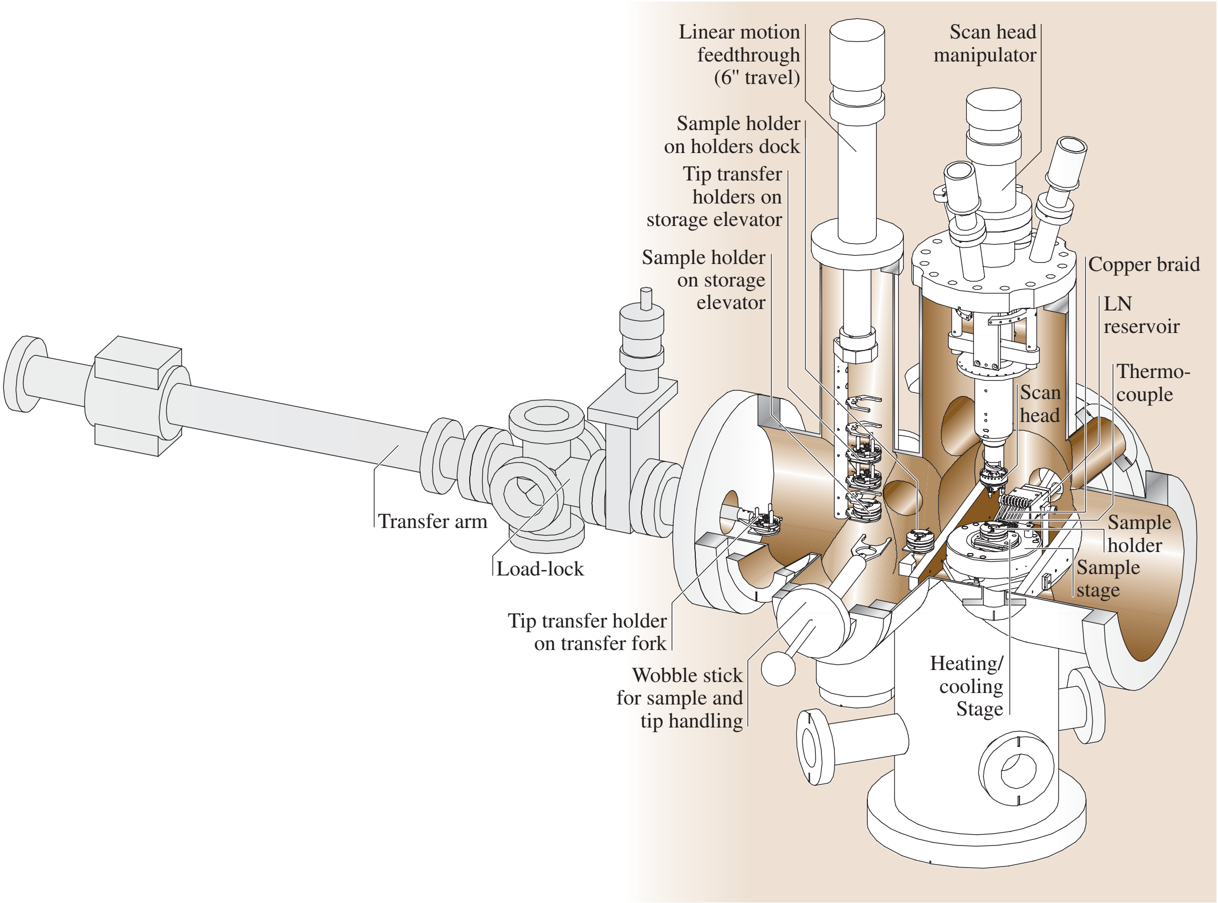
Linear motion
feedthrough
(6" travel)
Scan head
manipulator
Sample holder
on holders dock
Tip transfer
holders on
storage elevator
Sample holder
on storage
elevator
Copper braid
LN
reservoir
Thermo-
couple
Scan
head
Transfer arm
Load-lock
Sample
holder
Sample
stage
Tip transfer holder
on transfer fork
Wobble stick
for sample and
tip handling
Heating/
cooling
Stage

Fig. 14.1 One chamber UHV system with variable temperature STM based on a flow cryostat design. (Courtesy of RHK Technology, USA)

operation, a base temperature of below 30 K can be achieved, while a heater that is integrated into the sample stage enables high-temperature operation up to 1,000 K.

A typical experiment would run as follows: First, the sample is brought into the system by placing it in the so-called *load-lock*. This small part of the chamber can be separated from the rest of the system by a valve, so that the main part of the system can remain under vacuum at all times (i. e., even if the load-lock is opened for introducing the sample). After vacuum is re-established, the sample is transferred to the main chamber using the transfer arm. A linear motion feedthrough enables the storage of sample holders or, alternatively, specialized holders that carry replacement tips for the STM. Extending the transfer arm further, the sample can be placed on the sample stage and subsequently cooled down to the desired temperature. The scan head, which carries the STM tip, is then lowered with the scan head manipulator onto the sample holder (see Fig. 14.2). The special design of the scan head (see [14.14] for details) allows not only a flexible positioning of the tip on any desired location on the sample surface, but also compensates to a certain degree for thermal drift that inevitably occurs in such a design due to temperature gradients.

In fact, thermal drift is often much more prominent in LT-SPM designs, where only the sample is cooled, than in room temperature designs. Therefore, to fully benefit from the high stability conditions described in the introduction, it is mandatory to keep the whole microscope at the exact same temperature. This is mostly realized by using bath cryostats, which add a certain degree of complexity.

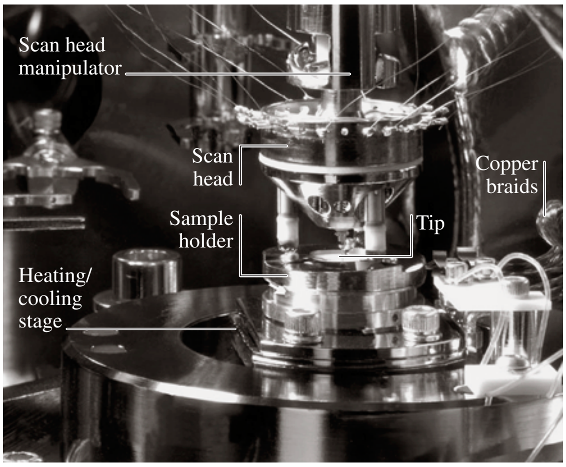
Scan head manipulator
Scan head
Copper braids
Sample holder
Tip
Heating/ cooling stage

Fig. 14.2 Photograph of the STM located inside the system sketched in Fig. 14.1. After the scan head has been lowered onto the sample holder, it is fully decoupled from the scan head manipulator and can be moved laterally using the three piezo legs it stands on. (Courtesy of RHK Technology, USA)

14.2.2 A Low Temperature SFM Based on a Bath Cryostat

As an example for an LT-SPM setup based on a bath cryostat, let us take a closer look at the LT-SFM system sketched in Fig. 14.3, which has been used to acquire the images on graphite, xenon, NiO, and InAs presented in Sect. 14.4. The force microscope is

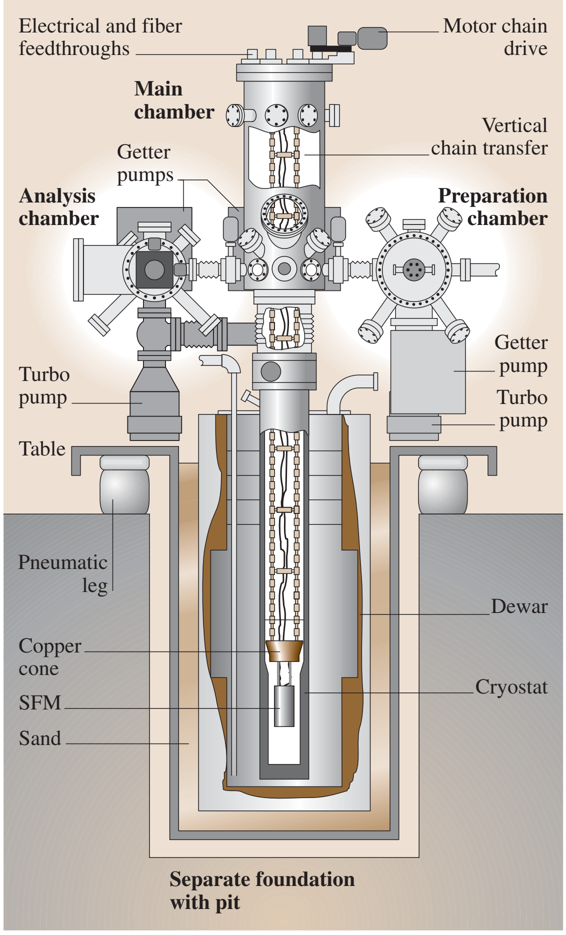
Electrical and fiber feedthroughs
Motor chain drive
Main chamber
Vertical chain transfer
Getter pumps
Analysis chamber
Preparation chamber
Getter pump
Turbo pump
Turbo pump
Table
Pneumatic leg
Dewar
Copper cone
Cryostat
SFM
Sand
Separate foundation with pit

Fig. 14.3 Three-chamber UHV and bath cryostat system for scanning force microscopy, front view

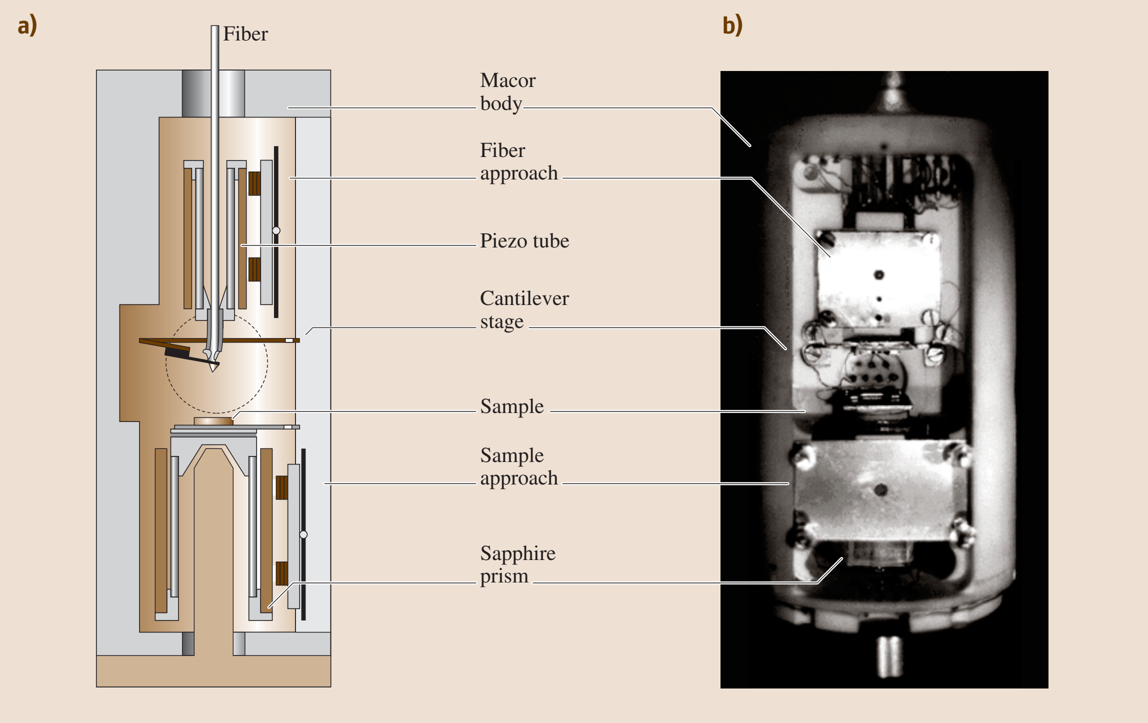

Fig. 14.4a,b The scanning force microscope incorporated into the system presented in Fig. 14.3. **(a)** Section along plane of symmetry. **(b)** Photo from the front

built into a UHV system that comprises three vacuum chambers: one for cantilever and sample preparation, which also serves as a transfer chamber, one for analysis purposes, and a main chamber that houses the microscope. A specially designed vertical transfer mechanism based on a double chain allows the lowering of the microscope into a UHV compatible bath cryostat attached underneath the main chamber. To damp the system, it is mounted on a table carried by pneumatic damping legs, which, in turn, stand on a separate foundation to decouple it from building vibrations. The cryostat and dewar are separated from the rest of the UHV system by a bellow. In addition, the dewar is surrounded by sand for acoustic isolation.

In this design, tip and sample are exchanged at room temperature in the main chamber. After the transfer into the cryostat, the SFM can be cooled by either liquid nitrogen or liquid helium, reaching temperatures down to 10 K. An all-fiber interferometer as the detection mechanism for the cantilever deflection ensures high resolution, while simultaneously allowing the construction of a comparatively small, rigid, and symmetric microscope.

Figure 14.4 highlights the layout of the SFM body itself. Along with the careful choice of materials, the symmetric design eliminates most of the problems with drift inside the microscope encountered when cooling or warming it up. The microscope body has an overall cylindrical shape with a 13 cm height and a 6 cm diameter and exact mirror symmetry along the cantilever axis. The main body is made of a single block of macor, a machinable glass ceramic, which ensures a rigid and stable design. For most of the metallic parts titanium was used, which has a temperature coefficient similar to macor. The controlled but stable accomplishment of movements, such as coarse approach and lateral positioning in other microscope designs, is a difficult task at low temperatures. The present design uses a special type of piezo motor that moves a sapphire prism (see the "fiber approach" and the "sample approach" signs in Fig. 14.3). It is described in detail in [14.15]. More information regarding this design is given in [14.16].

14.3 Scanning Tunneling Microscopy and Spectroscopy

In this section, we review some of the most important results achieved by LTSTM. After summarizing the results with emphasis on the necessity of LT equipment, we turn to details of the different experiments and the physical meaning of the obtained results.

As described in Sect. 14.1, the LT equipment has basically three advantages for scanning tunneling microscopy (STM) and spectroscopy (STS). First, the instruments are much more stable with respect to thermal drift and the coupling to external noise, allowing the establishment of new functionalities of the instrument. In particular, the LTSTM has been used to move atoms on a surface [14.12], cut molecules in pieces [14.17], reform bonds [14.18], and, consequently, establish new structures on the nanometer scale. Also, the detection of light resulting from tunneling into a particular atom [14.19] and the visualization of thermally induced atomic movements [14.20] partly require LT instrumentation.

Second, the spectroscopic resolution in STS depends linearly on temperature and is, therefore, considerably reduced at LT. This provides the opportunity to study physical effects unaccessible at room temperature. Obvious examples are the resolution of spin and Landau levels in semiconductors [14.9], or the investigation of lifetime broadening effects of particular electronic states on the nanometer scale [14.21]. More spectacular, electronic wave functions have been imaged for the first time in real space using an LTSTM [14.22], and vibrational levels giving rise to additional inelastic tunneling have been detected [14.23] and localized within particular molecules [14.24].

Third and most obvious, many physical effects, in particular, effects guided by electronic correlations, are restricted to low temperature. Typical examples are superconductivity [14.3], the Kondo effect [14.4], and many of the electron phases found in semiconductors [14.25]. Here, the LTSTM provides the possibility to study electronic effects on a local scale, and intensive work has been done in this field, the most elaborate with respect to high temperature superconductivity [14.26].

14.3.1 Atomic Manipulation

Although manipulation of surfaces on the atomic scale can be achieved at room temperature [14.27], only the use of LTSTMs allow the placement of individual atoms at desired atomic positions [14.28].

The usual technique to manipulate atoms is to increase the current above a certain atom, which reduces the tip-atom distance, then to move the tip with the atom to a desired position, and finally to reduce the current again in order to decouple atom and tip. The first demonstration of this technique was performed by Eigler and Schweizer (1990), who used Xe atoms on a Ni(110) surface to write the three letters “IBM” (their employer) on the atomic scale (Fig. 14.5a). Nowadays, many laboratories are able to move different kinds of atoms and molecules on different surfaces with high precision. An example featuring CO molecules on Cu(110) is shown in Fig. 14.5b–g. Basic modes of controlled motion, pushing, pulling, and sliding of the molecules have been established that depend on the tunneling current, i. e., the distance and the particular molecule-substrate combination [14.29]. It is believed that the electric field between tip and molecule is the strongest force moving the molecules, but other mechanisms such as electromigration caused by the high current density [14.28] or modifications of the surface potential due to the presence of the tip [14.30] have been put forth as important for some of the manipulation modes.

Meanwhile, other types of manipulation on the atomic scale have been developed. Some of them require inelastic tunneling into vibrational or rotational modes of molecules or atoms. They lead to controlled desorption [14.31], diffusion [14.32], pick-up of molecules by the tip [14.18], or rotation of individual entities [14.33, 34]. Also, dissociation of molecules by voltage pulses [14.17], conformational changes induced by dramatic change of tip-molecule distance [14.35], and association of pieces into larger molecules by reducing their lateral distance [14.18] have been shown. Fig. 14.5h–m shows the production of biphenyl from two iodobenzene molecules [14.36]. The iodine is abstracted by voltage pulses (Fig. 14.5i,j), then the iodine is moved to the terrace by the pulling mode (Fig. 14.5k,l), and finally the two phenyl parts are slid along the step edge until they are close enough to react (Fig. 14.5m). The chemical identification of the components is not deduced straightforwardly from STM images and partly requires detailed calculations of their apparent shape.

Low temperatures are not always required in these experiments, but they increase the reproducibility because of the higher stability of the instrument, as discussed in Sect. 14.1. Moreover, rotation or diffusion of entities could be excited at higher temperatures, making the intentionally produced configurations unstable.

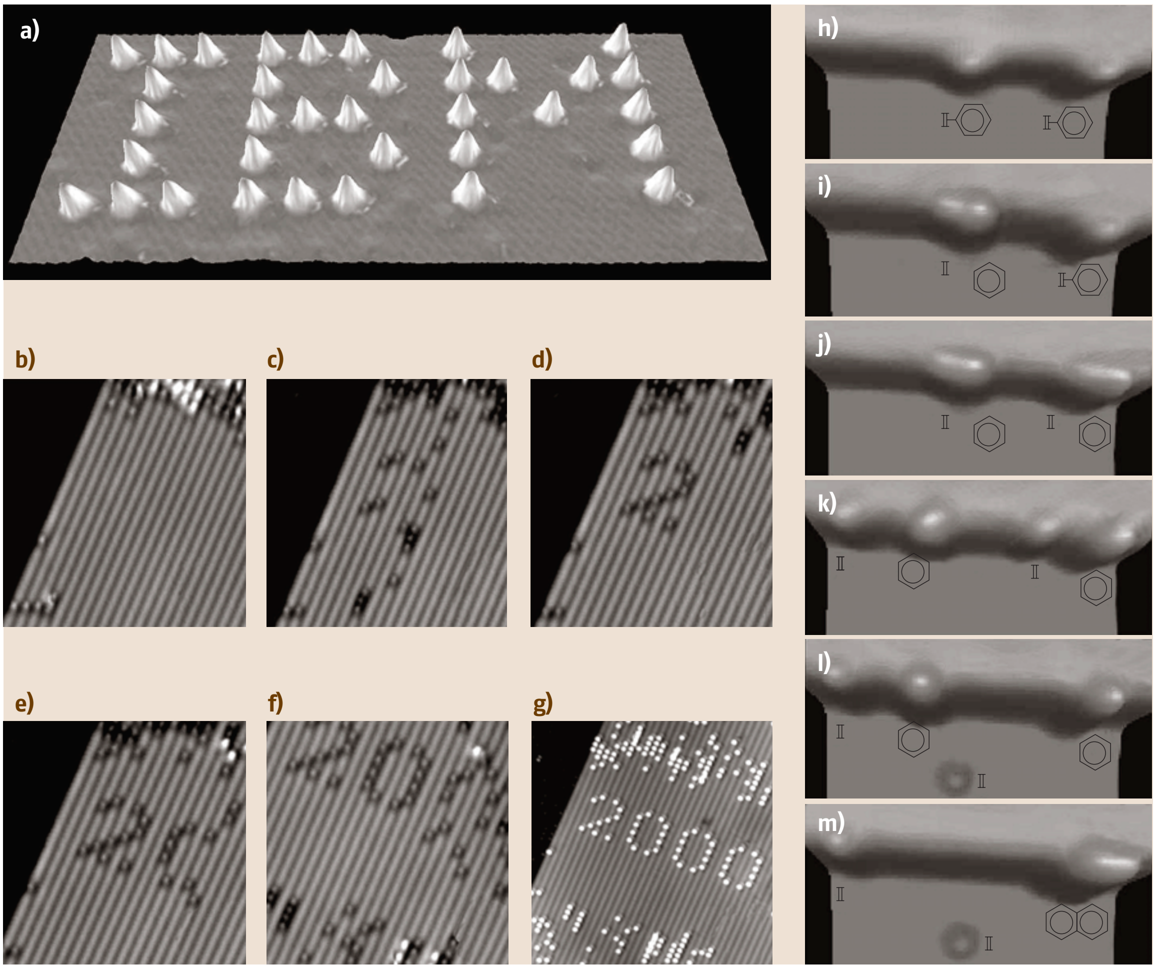

Fig. 14.5 **(a)** STM image of single Xe atoms positioned on a Ni(110) surface in order to realize the letters IBM on the atomic scale (courtesy of Eigler, IBM); **(b)–(f)** STM images recorded after different positioning processes of CO molecules on a Cu(110) surface; **(g)** final artwork greeting the new millennium on the atomic scale (**(b)–(g)** courtesy of Meyer, Berlin). **(h)–(m)** Synthesis of biphenyl from two iodobenzene molecules on Cu(111): First, iodine is abstracted from both molecules **(i)**,**(j)**, then the iodine between the two phenyl groups is removed from the step **(k)**, and finally one of the phenyls is slid along the Cu-step **(l)** until it reacts with the other phenyl **(m)**; the line drawings symbolize the actual status of the molecules (**(h)–(m)** courtesy of Hla and Rieder, Ohio)

14.3.2 Imaging Atomic Motion

Since individual manipulation processes last seconds to minutes, they probably cannot be used to manufacture large and repetitive structures. A possibility to construct such structures is self-assembled growth. It partly relies on the temperature dependence of different diffusion processes on the surface. A detailed knowledge of the diffusion parameters is required, which can be deduced from sequences of STM images measured at temperatures close to the onset of the process of interest [14.37]. Since many diffusion processes set on at LT, LT are partly required [14.20]. Consecutive images of so-called HtBDC molecules on Cu(110) recorded at $T = 194\,\mathrm{K}$ are shown in Fig. 14.6a–c [14.38]. As indicated by the arrows, the position of the molecules changes with time,

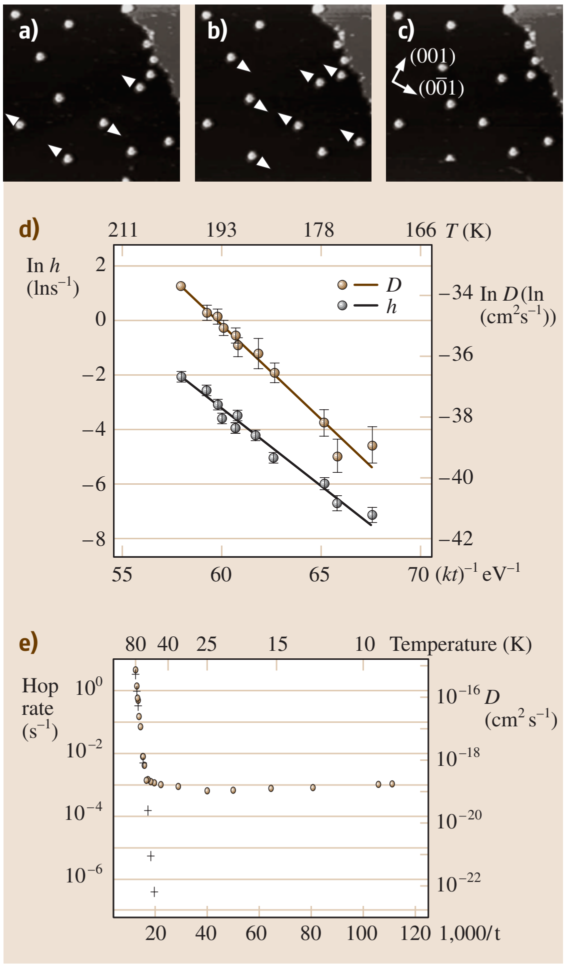

Fig. 14.6 **(a)–(c)** Consecutive STM images of hexa-tert-butyl decacyclene molecules on Cu(110) imaged at $T = 194$ K; *arrows* indicate the direction of motion of the molecules between two images. **(d)** Arrhenius plot of the hopping rate h determined from images like **(a)–(c)** as a function of inverse temperature (*grey symbols*); the *brown symbols* show the corresponding diffusion constant D; *lines* are fit results revealing an energy barrier of 570 meV for molecular diffusion (**(a)–(d)** courtesy of M. Schuhnack and F. Besenbacher, Aarhus). **(e)** Arrhenius plot for D (*crosses*) and H (*circles*) on Cu(001). The constant hopping rate of H below 65 K indicates a non-thermal diffusion process, probably tunneling (courtesy of Ho, Irvine)

implying diffusion. Diffusion parameters are obtained from Arrhenius plots of the determined hopping rate h, as shown in Fig. 14.6d. Of course, one must make sure that the diffusion process is not influenced by the presence of the tip, since it is known from manipulation experiments that the presence of the tip can move a molecule. However, particularly at low tunneling voltages these conditions can be fulfilled.

Besides the determination of diffusion parameters, studies of the diffusion of individual molecules showed the importance of mutual interactions in diffusion, which can lead to concerted motion of several molecules [14.20], or, very interestingly, the influence of quantum tunneling [14.39]. The latter is deduced from the Arrhenius plot of hopping rates of H and D on Cu(001), as shown in Fig. 14.6e. The hopping rate of H levels off at about 65 K, while the hopping rate of the heavier D atom goes down to nearly zero, as expected from thermally induced hopping.

Other diffusion processes such as the movement of surface vacancies [14.40] or of bulk interstitials close to the surface [14.41], and the Brownian motion of vacancy islands [14.42] have also been displayed.

14.3.3 Detecting Light from Single Atoms and Molecules

It had already been realized in 1988 that STM experiments are accompanied by light emission [14.43]. The fact that molecular resolution in the light intensity was achieved at LT (Fig. 14.7a,b) [14.19] raised the hope of performing quasi-optical experiments on the molecular scale. Meanwhile, it is clear that the basic emission process observed on metals is the decay of a local plasmon induced in the area around the tip by inelastic tunneling processes [14.44, 45]. Thus, the molecular resolution is basically a change in the plasmon environment, largely given by the increased height of the tip with respect to the surface above the molecule [14.46]. However, the electron can, in principle, also decay via single-particle excitations. Indeed, signatures of single-particle levels are observed. Figure 14.7c shows light spectra measured at different tunneling voltage V above a nearly complete Na monolayer on Cu(111) [14.47]. The plasmon mode peak energy (arrow) is found, as usual, to be proportional to V, but an additional peak that does not move with V appears at 1.6 eV (p). Plotting the light intensity as a function of photon energy and V (Fig. 14.7d) clearly shows that this additional peak is fixed in photon energy and corresponds to the separation of quantum well levels of the Na $(E_n - E_m)$.

Light has also been detected from semiconductors [14.48], including heterostructures [14.49]. There, the light is mostly caused by single-particle relaxation

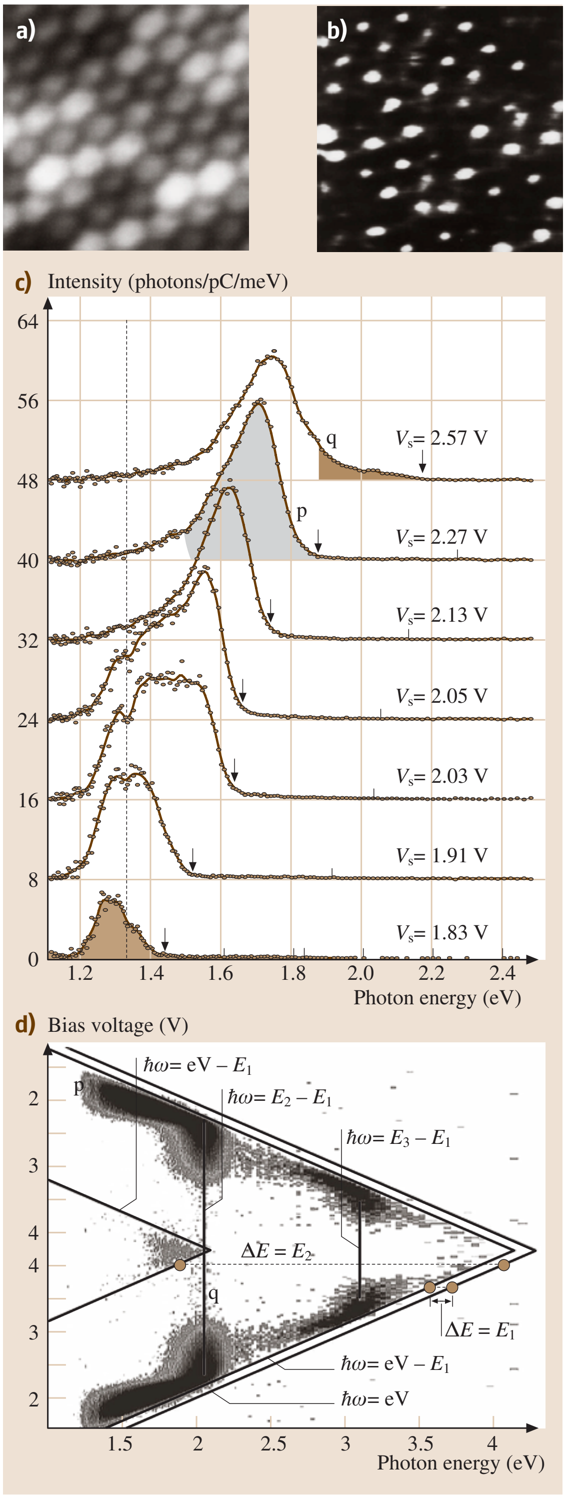

Fig. 14.7 **(a)** STM image of C_{60} molecules on Au(110) imaged at $T = 50$ K. **(b)** STM-induced photon intensity map of the same area; all photons from 1.5 eV to 2.8 eV contribute to the image, tunneling voltage $V = -2.8$ V (**(a)**,**(b)** courtesy of Berndt, Kiel). **(c)** STM-induced photon spectrum measured on 0.6 monolayer of Na on Cu(111) at different tunneling voltages as indicated. Besides the shifting plasmon mode marked by an *arrow*, an energetically constant part named p is recognizable. **(d)** Greyscale map of photon intensity as a function of tunnelling voltage and photon energy measured on 2.0 monolayer Na on Cu(111). The energetically constant photons are identified with intersubband transitions of the Na quantum well, as marked by $E_n - E_m$ (**(c)**,**(d)** courtesy of Hoffmann, Hamburg)

of the injected electrons, allowing a very local source of electron injection.

14.3.4 High Resolution Spectroscopy

One of the most important modes of LTSTM is STS. It detects the differential conductivity dI/dV as a function of the applied voltage V and the position (x, y). The dI/dV-signal is basically proportional to the local density of states (LDOS) of the sample, the sum over squared single-particle wave functions Ψ_i [14.2]

$$\frac{dI}{dV}(V, x, y) \propto \text{LDOS}(E, x, y) = \sum_{\Delta E} |\Psi_i(E, x, y)|^2 \,, \qquad (14.1)$$

where ΔE is the energy resolution of the experiment. In simple terms, each state corresponds to a tunneling channel if it is located between the Fermi levels (E_F) of the tip and the sample. Thus, all states located in this energy interval contribute to I, while $dI/dV(V)$ detects only the states at the energy E corresponding to V. The local intensity of each channel depends further on the LDOS of the state at the corresponding surface position and its decay length into vacuum. For s-like tip states, *Tersoff* and *Hamann* have shown that it is simply proportional to the LDOS at the position of the tip [14.50]. Therefore, as long as the decay length is spatially constant, one measures the LDOS at the surface (14.1). Note that the contributing states are not only surface states, but also bulk states. However, surface states usually dominate if present. *Chen* has shown that higher orbital tip states lead to the so-called derivation rule [14.51]: p_z-type tip states detect d(LDOS)/dz, d_z^2-states detect d^2(LDOS)/dz^2, and so on. As long as the decay into vacuum is exponential and spatially constant, this leads only to an additional factor in $\mathrm{d}I/\mathrm{d}V$. Thus, it is still the LDOS that is measured (14.1). The requirement of a spatially constant decay is usually fulfilled on larger length scales, but not on the atomic scale [14.51]. There, states located close to the atoms show a stronger decay into vacuum than the less localized states in the interstitial region. This effect can lead to corrugations that are larger than the real LDOS corrugations [14.52].

The voltage dependence of $\mathrm{d}I/\mathrm{d}V$ is sensitive to a changing decay length with V, which increases with V. Additionally, $\mathrm{d}I/\mathrm{d}V(V)$-curves might be influenced by possible structures in the DOS of the tip, which also contributes to the number of tunneling channels [14.53]. However, these structures can usually be identified, and only tips free of characteristic DOS structures are used for quantitative experiments.

Importantly, the energy resolution ΔE is largely determined by temperature. It is defined as the smallest energy distance of two δ-peaks in the LDOS that can still be resolved as two individual peaks in $\mathrm{d}I/\mathrm{d}V(V)$-curves and is $\Delta E = 3.3\,\mathrm{kT}$ [14.2]. The temperature dependence is nicely demonstrated in Fig. 14.8, where the tunneling gap of the superconductor Nb is measured at different temperatures [14.54]. The peaks at the rim of the gap get wider at temperatures well below the critical temperature of the superconductor ($T_\mathrm{c} = 9.2\,\mathrm{K}$).

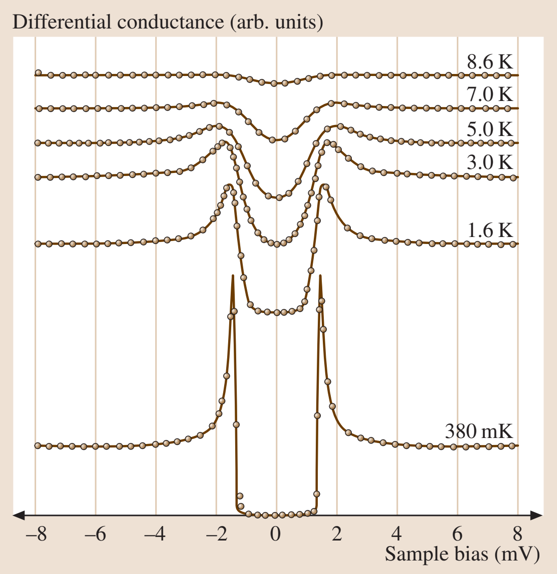

Fig. 14.8 Differential conductivity curve $\mathrm{d}I/\mathrm{d}V(V)$ measured on a Au surface by a Nb tip (*symbols*). Different temperatures are indicated; the lines are fits according to the superconducting gap of Nb folded with the temperature broadened Fermi distribution of the Au (courtesy of Pan, Houston)

Lifetime Broadening

Besides ΔE, intrinsic properties of the sample lead to a broadening of spectroscopic features. Basically, the finite lifetime of the electron or hole in the corresponding state broadens its energetic width. Any kind of interaction such as electron-electron interaction can be responsible. Lifetime broadening has usually been measured by photoemission spectroscopy (PES), but it turned out that lifetimes of surface states on noble metal surfaces determined by STS (Fig. 14.9a,b) are up to a factor of three larger than the ones measured by PES [14.55]. The reason is probably that defects broaden the PES spectrum. Defects are unavoidable in a spatially integrating technique such as PES, thus, STS has the advantage of choosing a particularly clean area for lifetime measurements. The STS results can be successfully compared to theory, highlighting the dominating influence of intraband transitions for the surface state lifetime on Au(111) and Cu(111), at least close to the onset of the surface band [14.21].

With respect to band electrons, the analysis of the width of the band onset in $\mathrm{d}I/\mathrm{d}V(V)$-curves has the disadvantage of being restricted to the onset energy. Another method circumvents this problem by measuring the decay of standing electron waves scattered from a step edge as a function of energy [14.56]. Figure 14.9c,d shows the resulting oscillating $\mathrm{d}I/\mathrm{d}V$-signal measured for two different energies. To deduce the coherence length L_Φ, which is inversely proportional to the lifetime τ_Φ, one has to consider that the finite ΔE in the experiment also leads to a decay of the standing wave away from the step edge. The dotted fit line using $L_\Phi = \infty$ indicates this effect and, more importantly, shows a discrepancy with the measured curve. Only including a finite coherence length of 6.2 nm results in good agreement, which, in turn, determines L_Φ and thus τ_Φ, as displayed in Fig. 14.9c. The found $1/E^2$-dependence of τ_Φ points to a dominating influence of electron-electron interactions at higher energies in the surface band.

Landau and Spin Levels

Moreover, the increased energy resolution at LT allows the resolution of electronic states that are not resolvable at RT. For example, Landau and spin quantization appearing in magnetic field B have been probed on InAs(110) [14.9, 57]. The corresponding quantization

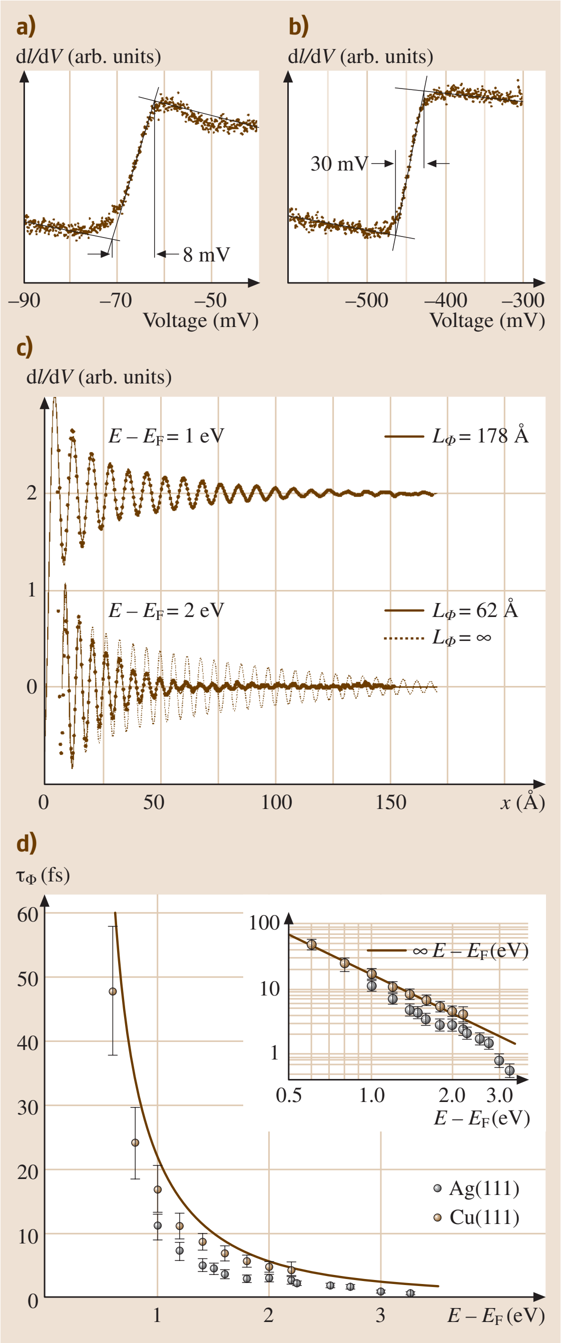

Fig. 14.9 **(a)**,**(b)** spatially averaged $\mathrm{d}I/\mathrm{d}V(V)$-curves of Ag(111) and Cu(111); both surfaces exhibit a surface state with parabolic dispersion starting at -65 meV and -430 meV, respectively. The lines are drawn to determine the energetic width of the onset of these surface bands (**(a)**,**(b)** courtesy of Berndt, Kiel); **(c)** $\mathrm{d}I/\mathrm{d}V$-intensity as a function of position away from a step edge of Cu(111) measured at the voltages $(E - E_\mathrm{F})$, as indicated (*points*); the lines are fits assuming standing electron waves with a phase coherence length L_Φ as marked; **(d)** resulting phase coherence time as a function of energy for Ag(111) and Cu(111). *Inset* shows the same data on a double logarithmic scale evidencing the E^{-2}-dependence (*line*) (**(c)**,**(d)** courtesy of Brune, Lausanne)

energies are given by $E_\mathrm{Landau} = \hbar eB/m_\mathrm{eff}$ and $E_\mathrm{spin} = g\mu B$. Thus InAs is a good choice, since it exhibits a low effective mass $m_\mathrm{eff}/m_\mathrm{e} = 0.023$ and a high g-factor of 14 in the bulk conduction band. The values in metals are $m_\mathrm{eff}/m_\mathrm{e} \approx 1$ and $g \approx 2$, resulting in energy splittings of only 1.25 meV and 1.2 meV at $B = 10$ T. This is obviously lower than the typical lifetime broadenings discussed in the previous section and also close to $\Delta E = 1.1$ meV achievable at $T = 4$ K.

Fortunately, the electron density in doped semiconductors is much lower, and thus the lifetime increases significantly. Figure 14.10a shows a set of spectroscopy curves obtained on InAs(110) in different magnetic fields [14.9]. Above E_F, oscillations with increasing intensity and energy distance are observed. They show the separation expected from Landau quantization. In turn, they can be used to deduce m_eff from the peak separation (Fig. 14.10b). An increase of m_eff with increasing E has been found as expected from theory. Also, at high fields spin quantization is observed (Fig. 14.10c). It is larger than expected from the bare g-factor due to contributions from exchange enhancement [14.58]. A detailed discussion of the peaks revealed that they belong to the so-called tip-induced quantum dot resulting from the work function difference between tip and sample.

Vibrational Levels

As discussed with respect to light emission in STM, inelastic tunneling processes contribute to the tunneling current. The coupling of electronic states to vibrational levels is one source of inelastic tunneling [14.23]. It provides additional channels contributing to $\mathrm{d}I/\mathrm{d}V(V)$ with final states at energies different from V. The final energy is simply shifted by the energy of the vibrational level. If only discrete vibrational energy levels couple to a smooth electronic DOS, one expects a peak in $\mathrm{d}^2I/\mathrm{d}V^2$ at the vibrational energy. This situation appears for molecules on noble metal surfaces. As usual, the isotope effect on the vibrational energy can be used

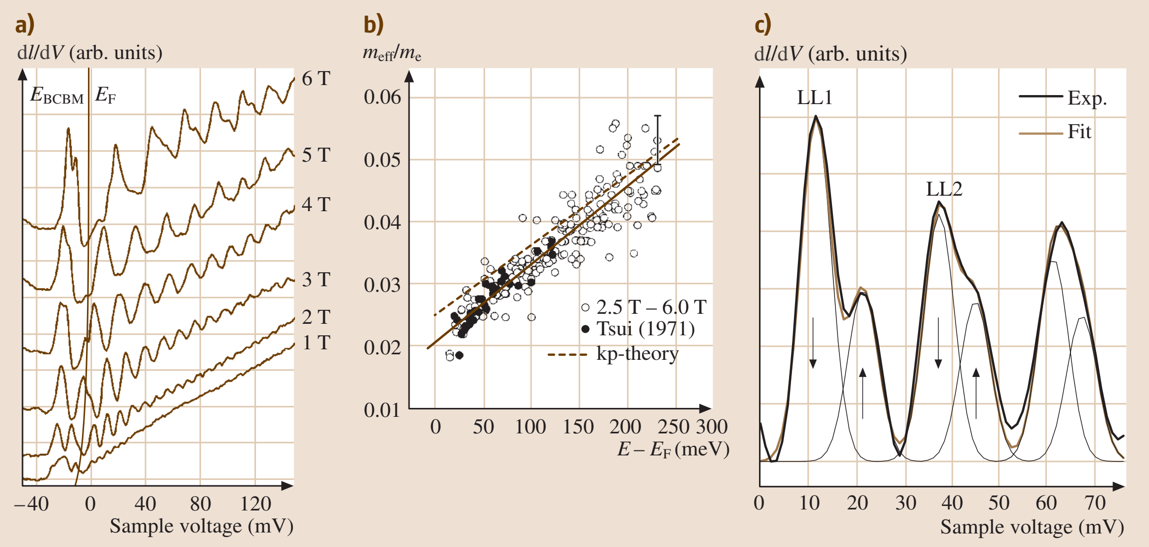

Fig. 14.10 **(a)** dI/dV-curves of n-InAs(110) at different magnetic fields as indicated; E_{BCBM} marks the bulk conduction band minimum; oscillations above E_{BCBM} are caused by Landau quantization; the double peaks at $B = 6\,\mathrm{T}$ are caused by spin quantization. **(b)** Effective mass data deduced from the distance of adjacent Landau peaks ΔE according to $\Delta E = h eB/m_{\mathrm{eff}}$ (*open symbols*); *filled symbols* are data from planar tunnel junctions (Tsui), the *solid line* is a mean-sqare fit of the data and the *dashed line* is the expected effective mass of InAs according to kp-theory. **(c)** Magnification of a dI/dV-curve at $B = 6\,\mathrm{T}$ exhibiting spin splitting; the Gaussian curves marked by *arrows* are the fitted spin levels

to verify the vibrational origin of the peak. First indications of vibrational levels have been found for H_2O and D_2O on TiO_2 [14.59], and completely convincing work has been performed for C_2H_2 and C_2D_2 on Cu(001) [14.23] (Fig. 14.11a). The technique can be used to identify individual molecules on the surface by their characteristic vibrational levels. In particular, surface reactions, as described in Fig. 14.5h–m, can be directly verified. Moreover, the orientation of complexes with respect to the surface can be determined to a certain extent, since the vibrational excitation depends on the position of the tunneling current within the molecule. Finally, the excitation of certain molecular levels can induce such corresponding motions as hopping [14.32], rotation [14.34] (Fig. 14.11b–e), or desorption [14.31], leading to additional possibilities for manipulation on the atomic scale.

Kondo Resonance

A rather intricate interaction effect is the Kondo effect. It results from a second order scattering process between itinerate states and a localized state [14.60]. The two states exchange some degree of freedom back and forth, leading to a divergence of the scattering probability at the Fermi level of the itinerate states. Due to the divergence, the effect strongly modifies sample properties. For example, it leads to an unexpected increase in resistance with decreasing temperature for metals containing magnetic impurities [14.4]. Here, the exchanged degree of freedom is the spin. A spectroscopic signature of the Kondo effect is a narrow peak in the DOS at the Fermi level disappearing above a characteristic temperature (Kondo temperature). STS provides the opportunity to study this effect on the local scale [14.61, 62].

Figure 14.12a–d shows an example of Co clusters deposited on a carbon nanotube [14.63]. While a small dip at the Fermi level, which is probably caused by curvature influences on the π-orbitals, is observed without Co (Fig. 14.12b) [14.64], a strong peak is found around a Co cluster deposited on top of the tube (Fig. 14.12a, arrow). The peak is slightly shifted with respect to $V = 0\,\mathrm{mV}$ due to the so-called Fano resonance [14.65], which results from interference of the tunneling processes into the localized Co-level and the itinerant nanotube levels. The resonance disappears within a several nanometer distance from the cluster, as shown in Fig. 14.12d.

The Kondo effect has also been detected for different magnetic atoms deposited on noble metal sur-

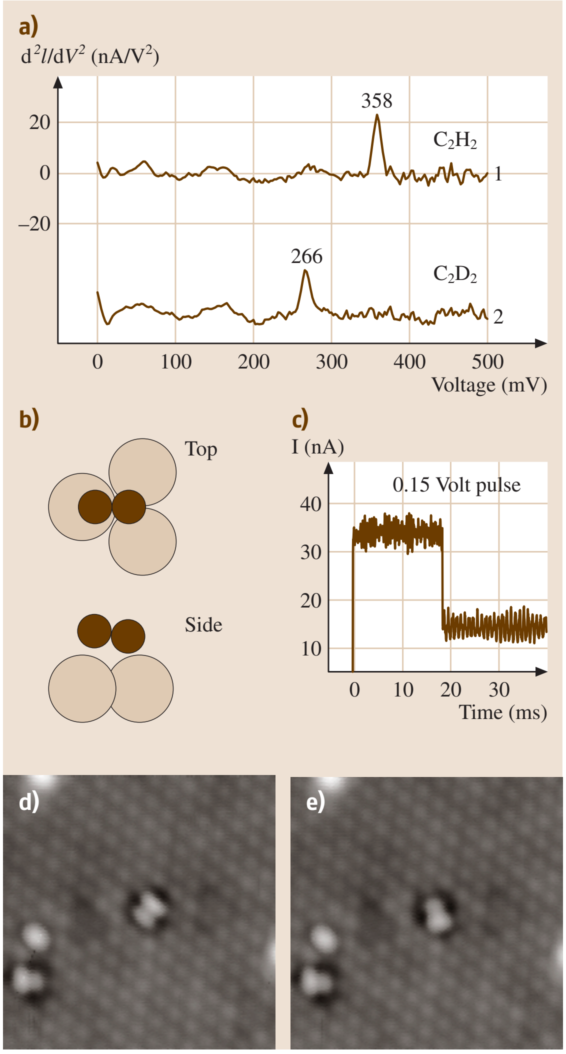

Fig. 14.11 **(a)** d^2I/dV^2-curves taken above a C_2H_2 and a C_2D_2 molecule on Cu(100); the peaks correspond to the C-H, respectively, C-D stretch mode energy of the molecule. **(b)** Sketch of O_2 molecule on Pt(111). **(c)** Tunnelling current above an O_2 molecule on Pt(111) during a voltage pulse of 0.15 V; the jump in current indicates rotation of the molecule. **(d)**, **(e)** STM image of an O_2 molecule on Pt(111) ($V = 0.05$ V) prior and after rotation induced by a voltage pulse to 0.15 V (**(a)**–**(e)** courtesy of Ho, Irvine)

faces [14.61,62]. There, it disappears at about 1 nm away from the magnetic impurity, and the effect of the Fano resonance is more pronounced, contributing to dips in $dI/dV(V)$-curves instead of peaks.

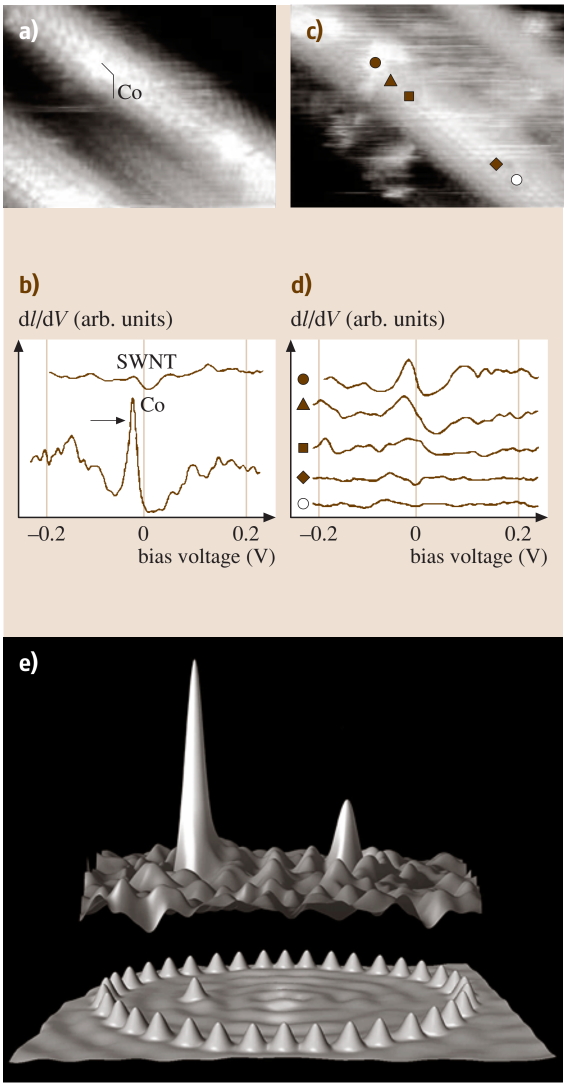

Fig. 14.12 **(a)** STM image of a Co cluster on a single-wall carbon nanotube (SWNT). **(b)** dI/dV-curves taken directly above the Co cluster (Co) and far away from the Co cluster (SWNT); the *arrow* marks the Kondo peak. **(c)** STM image of another Co cluster on a SWNT with *symbols* marking the positions where the dI/dV-curves displayed in **(d)** are taken. **(d)** dI/dV-curves taken at the positions marked in **(c)** (**(a)**–**(d)** courtesy of Lieber, Cambridge). **(e)** *Lower part:* STM image of a quantum corral of elliptic shape made from Co atoms on Cu(111); one Co atom is placed in one of the focii of the ellipse. *Upper part:* map of the strength of the Kondo signal in the corral; note that there is also a Kondo signal in the focus, which is not covered by a Co atom (**(e)** courtesy of Eigler, Almaden)

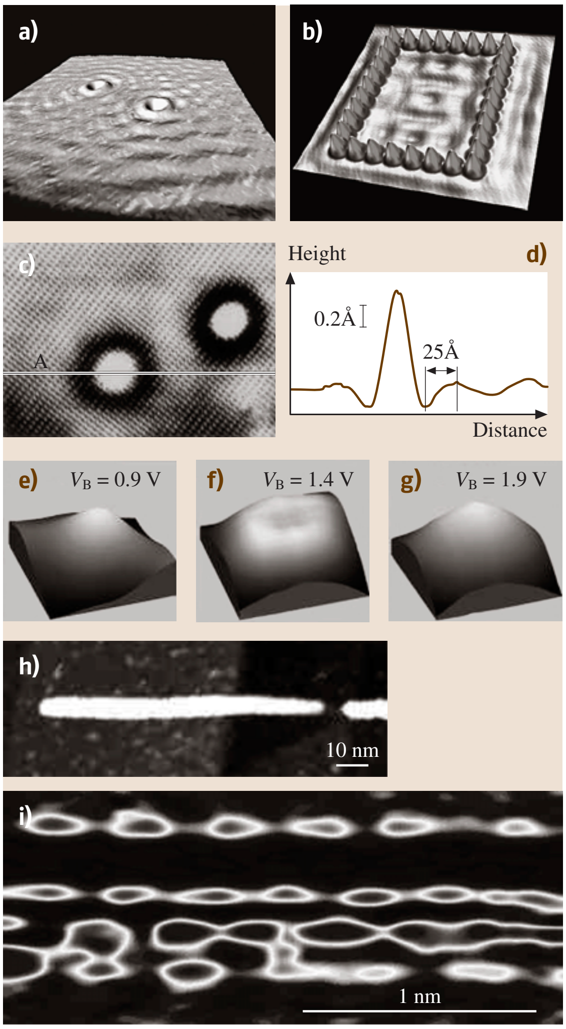

Fig. 14.13 **(a)** Low voltage STM image of Cu(111), including two defect atoms; the waves are electronic Bloch waves scattered at the defects. **(b)** Low voltage STM image of a rectangular quantum corral made from single atoms on Cu(111); the pattern inside the corral is the confined state of the corral close to E_F; (**(a)**,**(b)** courtesy of Eigler, Almaden). **(c)** STM image of GaAs(110) around a Si donor, $V = -2.5$ V; the line scan along A shown in **(d)** exhibits an additional oscillation around the donor caused by a standing Bloch wave; the grid like pattern corresponds to the atomic corrugation of the Bloch wave (**(c)**,**(d)** courtesy of van Kempen, Nijmegen). **(e)**–**(g)** STM images of an InAs/ZnSe-core/shell-nanocluster at different V. The image is measured in the so-called constant-height mode, i. e., the images display the tunneling current at constant height above the surface; the hill in **(e)** corresponds to the s-state of the cluster, the ring in **(f)** to the degenerate p_x- and p_y-state and the hill in **(g)** to the p_z-state (**(e)**–**(g)** courtesy of Millo, Jerusalem). **(h)** STM-image of a short-cut carbon nanotube. **(i)** Colour plot of the dI/dV intensity inside the short-cut nanotube as a function of position and tunneling voltage; four wavy patterns of different wavelength are visible in the voltage range from -0.1 to 0.15 V (**(h)**,**(i)** courtesy of Dekker, Delft)

A fascinating experiment has been performed by *Manoharan* et al. [14.66], who used manipulation to form an elliptic cage for the surface states of Cu(111) (Fig. 14.12e, bottom). This cage was constructed to have a quantized level at E_F. Then, a cobalt atom was placed in one focus of the elliptic cage, producing a Kondo resonance. Surprisingly, the same resonance reappeared in the opposite focus, but not away from the focus (Fig. 14.12e, top). This shows amazingly that complex local effects such as the Kondo resonance can be guided to remote points.

Orbital scattering as a source of the Kondo resonance has also been found around a defect on Cr(001) [14.67]. Here, it is believed that itinerate sp-levels scatter at a localized d-level to produce the Kondo peak.

14.3.5 Imaging Electronic Wave Functions

Since STS measures the sum of squared wave functions (14.1), it is an obvious task to measure the local appearance of the most simple wave functions in solids, namely, the Bloch waves.

Bloch Waves

The atomically periodic part of the Bloch wave is always measured if atomic resolution is achieved (inset of Fig. 14.14a). However, the long-range, wavy part requires the presence of scatterers. The electron wave impinges on the scatterer and gets reflected, leading to self-interference. In other words, the phase of the Bloch wave gets fixed by the scatterer.

Such self-interference patterns were first found on Graphite(0001) [14.68] and later on noble metal surfaces, where adsorbates or step edges scatter the surface states (Fig. 14.13a) [14.22]. Fourier transforms of the real-space images reveal the k-space distribution of the corresponding states [14.69], which may include additional contributions besides the surface state [14.70]. Using particular geometries as the so-called quantum corrals to form a cage for the electron wave, the scatter-

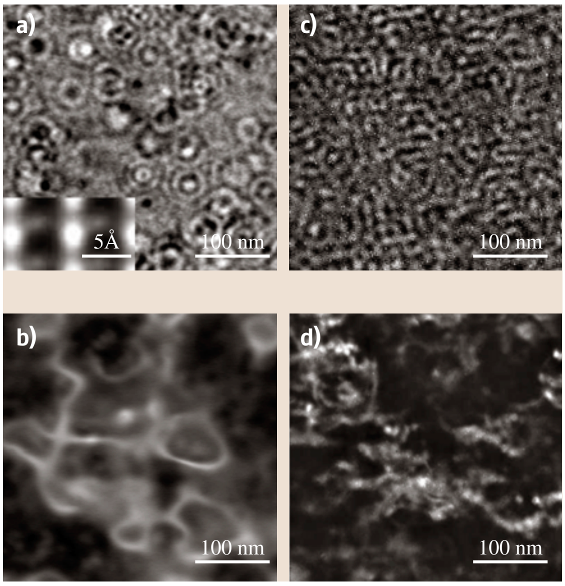

Fig. 14.14 **(a)** dI/dV-image of InAs(110) at $V = 50\,\mathrm{mV}$, $B = 0\,\mathrm{T}$; circular wave patterns corresponding to standing Bloch waves around each S donor are visible; inset shows a magnification revealing the atomically periodic part of the Bloch wave. **(b)** Same as **(a)**, but at $B = 6\,\mathrm{T}$; the stripe structures are drift states. **(c)** dI/dV-image of a 2-D electron system on InAs(110) induced by the deposition of Fe, $B = 0\,\mathrm{T}$. **(d)** Same as **(c)** but at $B = 6\,\mathrm{T}$; note that the contrast in **(a)** is increased by a factor of ten with respect to **(b)–(d)**

ing state can be rather complex (Fig. 14.13b). Anyway, it can usually be reproduced by simple calculations involving single-particle states [14.71].

Bloch waves in semiconductors scattered at charged dopants (Fig. 14.13c,d) [14.72], Bloch states confined in semiconductor quantum dots (Fig. 14.13e–g) [14.73], and Bloch waves confined in short-cut carbon nanotubes (Fig. 14.13h,i) [14.74, 75] have been visualized.

Drift States

More complex wave functions result from interactions. A nice playground to study such interactions are doped semiconductors. The reduced electron density with respect to metals increases the importance of electron interactions with potential disorder and other electrons. Applying a magnetic field quenches the kinetic energy, further enhancing the importance of interactions. A dramatic effect can be observed on InAs(110), where 3-D bulk states are displayed. While the usual scattering states around individual dopants are observed at $B = 0\,\mathrm{T}$ (Fig. 14.14a) [14.76], stripe structures are found in high magnetic field (Fig. 14.14b) [14.77]. They run along equipotential lines of the disorder potential. This can be understood by recalling that the electron tries to move in a cyclotron orbit, which is accelerated and decelerated in electrostatic potential, leading to a drift motion along an equipotential line [14.78].

The same effect has been found in 2-D electron systems (2-DES) of the same substrate, where the scattering states at $B = 0\,\mathrm{T}$ are, however, found to be more complex (Fig. 14.14c) [14.79]. The reason is the tendency of a 2-DES to exhibit closed scattering paths [14.80]. Consequently, the self-interference does not result from scattering at individual scatterers, but from complicated self-interference paths involving many scatterers. But drift states are also observed in the 2-DES at high magnetic fields (Fig. 14.14d) [14.81].

Charge Density Waves

Another interaction modifying the LDOS is the electron-phonon interaction. Phonons scatter electrons between different Fermi points. If the wave vectors connecting Fermi points exhibit a preferential orientation, a so-called Peierls instability occurs [14.82]. The corresponding phonon energy goes to zero, the atoms are slightly displaced with the periodicity of the corresponding wave vector, and a charge density wave (CDW) with the same periodicity appears. Essentially, the CDW increases the overlap of the electronic states with the phonon by phase fixing with respect to the atomic lattice. The Peierls transition naturally occurs in 1-D systems, where only two Fermi points are present, and, hence, preferential orientation is pathological. It can also occur in 2-D systems if large areas of the Fermi line run in parallel.

STS studies of CDWs are numerous (e.g., [14.83, 84]). Examples of a 1-D CDW on a quasi 1-D-like bulk material and of a 2-D CDW are shown in Fig. 14.15a–d and Fig. 14.15e, respectively [14.85, 86]. In contrast to usual scattering states where LDOS corrugations are only found close to the scatterer, the corrugations of CDWs are continous across the surface. Heating the substrate toward the transition temperature leads to a melting of the CDW lattice, as shown in Fig. 14.15f–h.

CDWs have also been found on monolayers of adsorbates such as a monolayer of Pb on Ge(111) [14.87]. These authors performed a nice temperature-dependent study revealing that the CDW is nucleated by scattering states around defects as one might expect [14.88]. 1-D systems have also been prepared on surfaces show-

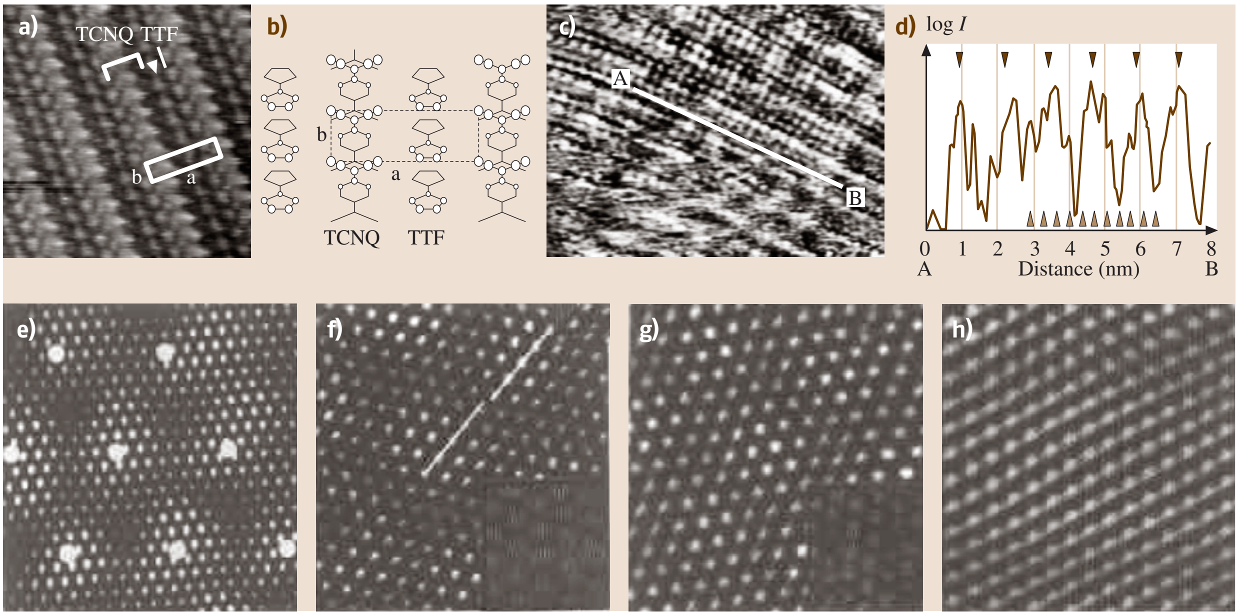

Fig. 14.15 **(a)** STM image of the ab-plane of the organic quasi 1-D conductor TTF-TCNQ, $T = 300\,\text{K}$; while the TCNQ chains are conducting, the TTF chains are insulating. **(b)** Stick and ball model of the ab-plane of TTF-TCNQ. **(c)** STM image taken at $T = 61\,\text{K}$, the additional modulation due to the Peierls-transition is visible in the profile along line A shown in **(d)**; the *open triangles* mark the atomic periodicity and the *filled triangles* the expected CDW periodicity (**(a)–(d)** courtesy of Kageshima, Kanagawa). **(e)–(h)** Low voltage STM-images of the two-dimensional CDW-system 1 T-TaS_2 at $T = 242\,\text{K}$ **(e)**, 298 K **(f)**, 349 K **(g)**, 357 K **(h)**. A long-range, hexagonal modulation is visible besides the atomic spots; its periodicity is highlighted by *big white dots* in **(e)**; the additional modulation obviously weakens with increasing T, but is still apparent in **(f)** and **(g)**, as evidenced in the lower magnification images in the insets (**(e)–(h)** courtesy of Lieber, Cambridge)

ing Peierls transitions [14.89,90]. Finally, the energy gap occurring at the transition has been studied by measuring $\text{d}I/\text{d}V(V)$-curves [14.91].

Superconductors

An intriguing effect resulting from electron-phonon interaction is superconductivity. Here, the attractive part of the electron-phonon interaction leads to the coupling of electronic states with opposite wave vector and mostly opposite spin [14.92]. Since the resulting Cooper pairs are bosons, they can condense at LT, forming a coherent many-particle phase, which can carry current without resistance. Interestingly, defect scattering does not influence the condensate if the coupling along the Fermi surface is homogeneous (s-wave superconductor). The reason is that the symmetry of the scattering of the two components of a Cooper pair effectively leads to a scattering from one Cooper pair state to another without affecting the condensate. This is different if the scatterer is magnetic, since the different spin components of the pair are scattered differently, leading to an effective pair breaking visible as a single-particle excitation within the superconducting gap. On a local scale, the effect has first been demonstrated by putting Mn, Gd, and Ag atoms on a Nb(110) surface [14.93]. While the nonmagnetic Ag does not modify the gap shown in Fig. 14.16a, it is modified in an asymmetric fashion close to Mn or Gd adsorbates, as shown in Fig. 14.16b. The asymmetry of the additional intensity is caused by the breaking of the particle-hole symmetry due to the exchange interaction between the localized Mn state and the itinerate Nb states.

Another important local effect is caused by the relatively large coherence length of the condensate. At a material interface, the condensate wave function cannot stop abruptly, but laps into the surrounding material (proximity effect). Consequently, a superconducting gap can be measured in areas of non-superconducting material. Several studies have shown this effect on the

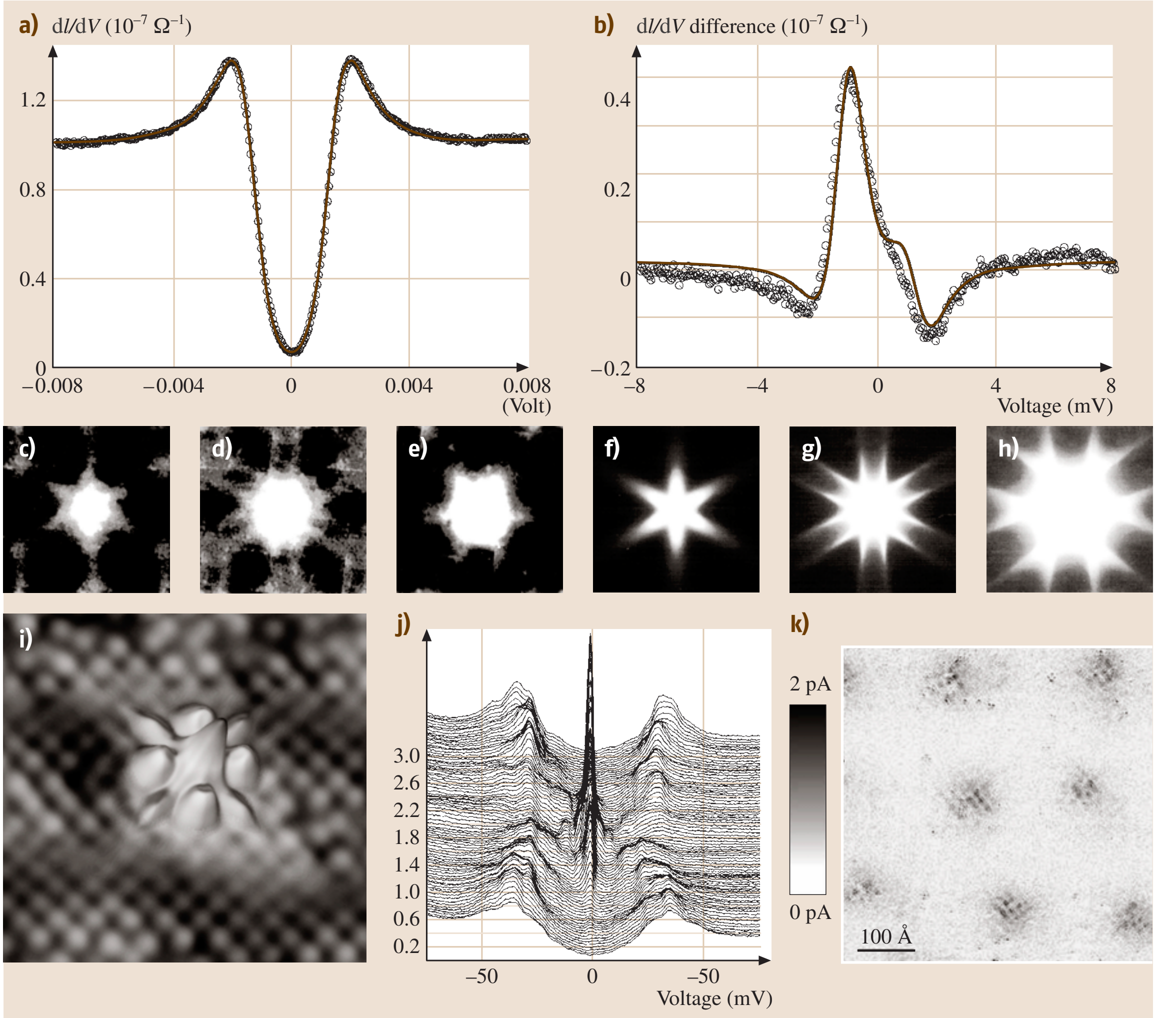

Fig. 14.16 **(a)** $\mathrm{d}I/\mathrm{d}V$-curve of Nb(110) at $T = 3.8\,\mathrm{K}$ (*symbols*) in comparison with a BCS-fit of the superconducting gap of Nb (*line*). **(b)** Difference between the $\mathrm{d}I/\mathrm{d}V$-curve taken directly above a Mn-atom on Nb(110) and the $\mathrm{d}I/\mathrm{d}V$-curve taken above the clean Nb(110) (*symbols*) in comparison with a fit using the Bogulubov-de Gennes equations (*line*) (**(a)**,**(b)** courtesy of Eigler, Almaden). **(c)**–**(e)** $\mathrm{d}I/\mathrm{d}V$-images of a vortex core in the type-II superconductor 2H-NbSe$_2$ at 0 mV **(c)**, 0.24 mV **(d)**, and 0.48 mV **(e)** (**(c)**–**(e)** courtesy of H. F. Hess). **(f)**–**(h)** Corresponding calculated LDOS images within the Eilenberger framework (**(f)**–**(h)** courtesy of Machida, Okayama). **(i)** Overlap of an STM image at $V = -100\,\mathrm{mV}$ (backround 2-D image) and a $\mathrm{d}I/\mathrm{d}V$-image at $V = 0\,\mathrm{mV}$ (overlapped 3-D image) of optimally doped $\mathrm{Bi_2Sr_2CaCu_2O_{8+\delta}}$ containing 0.6% Zn impurities. The STM image shows the atomic structure of the cleavage plane, while the $\mathrm{d}I/\mathrm{d}V$-image shows a bound state within the superconducting gap, which is located around a single Zn impurity. The fourfold symmetry of the bound state reflects the d-like symmetry of the superconducting pairing function; **(j)** $\mathrm{d}I/\mathrm{d}V$-curves taken at different positions across the Zn impurity; the bound state close to 0 mV is visible close to the Zn atom; **(k)** LDOS in the vortex core of slightly overdoped $\mathrm{Bi_2Sr_2CaCu_2O_{8+\delta}}$, $B = 5\,\mathrm{T}$; the $\mathrm{d}I/\mathrm{d}V$-image taken at $B = 5\,\mathrm{T}$ is integrated over $V = 1{-}12\,\mathrm{mV}$, and the corresponding $\mathrm{d}I/\mathrm{d}V$-image at $B = 0\,\mathrm{T}$ is subtracted to highlight the LDOS induced by the magnetic field. The checkerboard pattern within the seven vortex cores exhibits a periodicity, which is fourfold with respect to the atomic lattice shown in **(i)** and is thus assumed to be a CDW (**(i)**–**(k)** courtesy of S. Davis, Cornell and S. Uchida, Tokyo)

local scale using metals and doped semiconductors as surrounding materials [14.94, 95].

While the classical type-I superconductors are ideal diamagnets, the so-called type-II superconductors can contain magnetic flux. The flux forms vortices, each containing one flux quantum. These vortices are accompanied by a disappearance of the superconducting gap and, therefore, can be probed by STS [14.96]. LDOS maps measured inside the gap lead to bright features in the area of the vortex core. Importantly, the length scale of these features is different from the length scale of the magnetic flux due to the difference between London's penetration depth and coherence length. Thus, STS probes another property of the vortex than the usual magnetic imaging techniques (see Sect. 14.4.4). Surprisingly, first measurements of the vortices on $NbSe_2$ revealed vortices shaped as a sixfold star [14.97] (Fig. 14.16c). With increasing voltage inside the gap, the orientation of the star rotates by 30° (Fig. 14.16d,e). The shape of these stars could finally be reproduced by theory, assuming an anisotropic pairing of electrons in the superconductor (Fig. 14.16f–h) [14.98]. Additionally, bound states inside the vortex core, which result from confinement by the surrounding superconducting material, are found [14.97]. Further experiments investigated the arrangement of the vortex lattice, including transitions between hexagonal and quadratic lattices [14.99], the influence of pinning centers [14.100], and the vortex motion induced by current [14.101].

An important topic is still the understanding of high temperature superconductivity (HTCS). An almost accepted property of HTCS is its d-wave pairing symmetry. In contrast to s-wave superconductors, scattering can lead to pair breaking, since the Cooper pair density vanishes in certain directions. Indeed, scattering states (bound states in the gap) around nonmagnetic Zn impurities have been observed in $Bi_2Sr_2CaCu_2O_{8+\delta}$ (BSCCO) (Fig. 14.16i,j) [14.26]. They reveal a d-like symmetry, but not the one expected from simple Cooper pair scattering. Other effects such as magnetic polarization in the environment probably have to be taken into account [14.102]. Moreover, it has been found that magnetic Ni impurities exhibit a weaker scattering structure than Zn impurities [14.103]. Thus, BSCCO shows exactly the opposite behavior of the Nb discussed above (Fig. 14.16a,b). An interesting topic is the importance of inhomogeneities in HTCS materials. Evidence for inhomogeneities has indeed been found in underdoped materials, where puddles of the superconducting phase are shown to be embedded in non-superconducting areas [14.104].

Of course, vortices have also been investigated in HTCS materials [14.105]. Bound states are found, but at energies that are in disagreement with simple models, assuming a BCS-like d-wave superconductor [14.106, 107]. Theory predicts, instead, that the bound states are magnetic-field-induced spin-density waves stressing the competition between anti-ferromagnetic order and superconductivity in HTCS materials [14.108]. Since the spin density wave is accompanied by a charge density wave of half wavelength, it can be probed by STS [14.109]. Indeed, a checkerboard pattern of the right periodicity has been found in and around vortex cores in BSCCO (Fig. 14.16k). It exceeds the width of an individual vortex core, implying that the superconducting coherence length is different from the anti-ferromagnetic one.

Complex Systems (Manganites)

Complex phase diagrams are not restricted to HTCS materials (cuprates). They exist with similar complexity for other doped oxides such as manganites. Only a few studies of these materials have been performed by STS, mainly showing the inhomogeneous evolution of metallic and insulating phases [14.110, 111]. Similarities to the granular case of an underdoped HTCS material are obvious. Since inhomogeneities seem to be crucial in many of these materials, a local method such as STS might continue to be important for the understanding of their complex properties.

14.3.6 Imaging Spin Polarization: Nanomagnetism

Conventional STS couples to the LDOS, i. e., the charge distribution of the electronic states. Since electrons also have spin, it is desirable to also probe the spin distribution of the states. This can be achieved if the tunneling tip is covered by a ferromagnetic material [14.112]. The coating acts as a spin filter or, more precisely, the tunneling current depends on the relative angle α_{ij} between the spins of the tip and the sample according to $\cos(\alpha_{ij})$. In ferromagnets, the spins mostly have one preferential orientation along the so-called easy axis, i. e., a particular tip is not sensitive to spin orientations of the sample perpendicular to the spin orientation of the tip. Different tips have to be prepared to detect different spin orientations of the sample. Moreover, the magnetic stray field of the tip can perturb the spin orientation of the sample. To avoid this, a technique

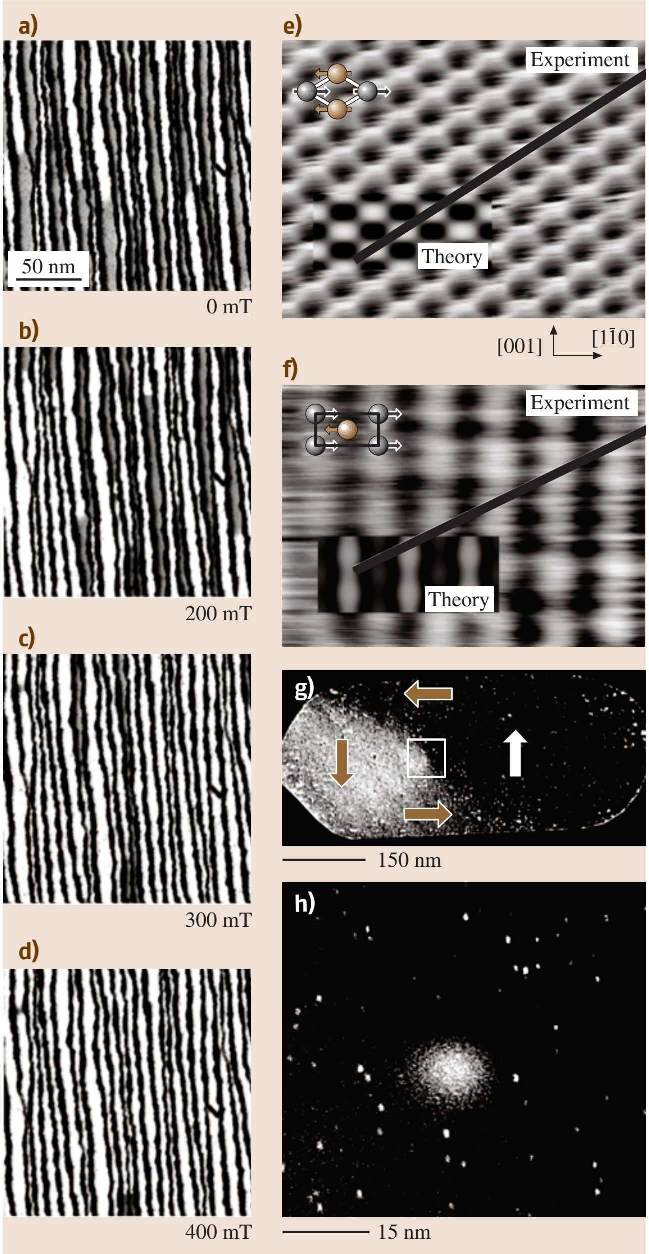

Fig. 14.17 (a)–(d) Spin-polarized STM images of 1.65 monolayer of Fe deposited on a stepped W(110) surface measured at different *B*-fields, as indicated. Double layer and monolayer Fe stripes are formed on the W substrate; only the double layer stripes exhibit magnetic contrast with an out-of-plane sensitive tip as used here. *White* and *grey areas* correspond to different domains. Note that more white areas appear with increasing field. **(e)** STM image of an antiferromagnetic Mn monolayer on W(110). **(f)** Spin-polarized STM-image of the same surface (in-plane tip). The insets in **(e)** and **(f)** show the calculated STM and spin-polarized STM images, respectively, and the stick and ball models symbolize the atomic and the magnetic unit cell (**(a)–(f)** courtesy of M. Bode, Hamburg). **(g)** Spin-polarized STM image of a 6-nm-high Fe island on W(110) (in-plane tip). Four different areas are identified as four different domains with domain orientations as indicated by the *arrows*. **(h)** Spin-polarized STM image of the central area of an island; the size of the area is indicated by the rectangle in **(g)**; the measurement is performed with an out-of-plane sensitive tip showing that the magnetization turns out-of-plane in the center of the island

using anti-ferromagnetic Cr as a tip coating material has been developed [14.113]. This avoids stray fields, but still provides a preferential spin orientation of the few atoms at the tip apex that dominate the tunneling current. Depending on the thickness of the Cr coating, spin orientations perpendicular or parallel to the sample surface are prepared.

So far, the described technique has been used to image the evolution of magnetic domains with increasing *B*-field(Fig. 14.17a–d) [14.114], the anti-ferromagnetic order of a Mn monolayer on W(110) (Fig. 14.17e,f) [14.115], and the out-of-plane orientation predicted for a magnetic vortex core as it exists in the center of a Fe island exhibiting four domains in the flux closure configuration (Fig. 14.17g,h) [14.116].

Besides the obvious strong impact on nanomagnetism, the technique might also be used to investigate other electronic phases such as the proposed spin density wave around a HTCS vortex core.

14.4 Scanning Force Microscopy and Spectroscopy

SFSThe examples discussed in the previous section show the wide variety of physical questions that have been tackled with the help of LTSTM. Here, we turn to the other prominent scanning probe method that is applied at low temperatures, namely, SFM, which gives complementary information on sample properties on the atomic scale.

The ability to sensitively detect *forces* with spatial resolution down to the atomic scale is of great interest, since force is one of the most fundamental quantities in physics. Mechanical force probes usually consist of a cantilever with a tip at its free end that is brought close to the sample surface. The cantilever can be mounted parallel or perpendicular to the surface (general aspects of force probe designs are described in Chapt. 12). Basically, two methods exist to detect forces with cantilever-based probes – the *static* and the *dynamic* mode (see Chapt. 11). They can be used to generate a laterally resolved image (*microscopy* mode) or determine its distance dependence (*spectroscopy* mode). One can argue about the terminology, since spectroscopy is usually related to energies and not to distance dependences. Nevertheless, we will use it throughout the text, because it avoids lenghty paraphrases and is established in this sense throughout the literature.

In the static mode, a force that acts on the tip bends the cantilever. By measuring its deflection Δz the tip-sample force F_{ts} can be directly calculated with Hooke's law: $F_{\mathrm{ts}} = c_z \cdot \Delta z$, where c_z denotes the spring constant of the cantilever. In the various dynamic modes, the cantilever is oscillated with amplitude A at or near its eigenfrequency f_0, but in some applications also off resonance. At ambient pressures or in liquids, amplitude modulation (AM-SFM) is used to detect amplitude changes or the phase shift between driving force and cantilever oscillation. In vacuum, the frequency shift Δf of the cantilever due to a tip-sample interaction is measured by frequency modulation technique (FM-SFM). The nomenclature is not standardized. Terms like tapping mode or intermittent contact mode are used instead of AM-SFM, and NC-AFM (noncontact atomic force microscopy) or DFM (dynamic force microscopy) instead of FM-SFM or FM-AFM. However, all these modes are *dynamic*, i. e., they involve an oscillating cantilever and can be used in the noncontact, as well as in the contact regime. Therefore, we believe that the best and most consistent way is to distinguish them by their different detection schemes. Converting the measured quantity (amplitude, phase, or frequency shift) into a physically meaningful quantity, e.g., the tip-sample interaction force F_{ts} or the force gradient $\partial F_{\mathrm{ts}}/\partial z$, is not always straightforward and requires an analysis of the equation of motion of the oscillating tip (see Chaps. 15, 13).

Whatever method is used, the resolution of a cantilever-based force detection is fundamentally limited by its intrinsic *thermomechanical* noise. If the cantilever is in thermal equilibrium at a temperature T, the equipartition theorem predicts a thermally induced *root mean square* (rms) motion of the cantilever in z direction of $z_{\mathrm{rms}} = (k_{\mathrm{B}}T/c_{\mathrm{eff}})^{1/2}$, where k_{B} is the Boltzmann constant and $c_{\mathrm{eff}} = c_z + \partial F_{\mathrm{ts}}/\partial z$. Note that usually $\mathrm{d}F_{\mathrm{ts}}/\mathrm{d}z \gg c_z$ in contact and $\mathrm{d}F_{\mathrm{ts}}/\mathrm{d}z < c_z$ in noncontact. Evidently, this fundamentally limits the force resolution in the static mode, particularly if operated in noncontact. Of course, the same is true for the different dynamic modes, because the thermal energy $k_{\mathrm{B}}T$ excites the eigenfrequency f_0 of the cantilever. Thermal noise is *white* noise, i. e., its spectral density is flat. However, if the cantilever transfer function is taken into account, one can see that the thermal energy mainly excites f_0. This explains the term "thermo" in thermomechanical noise, but what is the "mechanical" part?

A more detailed analysis reveals that the thermally induced cantilever motion is given by

$$z_{\mathrm{rms}} = \sqrt{\frac{2k_{\mathrm{B}}TB}{\pi c_z f_0 Q}}\,, \qquad (14.2)$$

where B is the measurement bandwidth and Q is the quality factor of the cantilever. Analog expressions can be obtained for all quantities measured in dynamic modes, because the deflection noise translates, e.g., into frequency noise [14.5]. Note that f_0 and c_z are correlated with each other via $2\pi f_0 = (c_z/m_{\mathrm{eff}})^{1/2}$, where the effective mass m_{eff} depends on the geometry, density, and elasticity of the material. The Q-factor of the cantilever is related to the external damping of the cantilever motion in a medium and on the intrinsic damping within the material. This is the "mechanical" part of the fundamental cantilever noise.

It is possible to operate a low temperature force microscope directly immersed in the cryogen [14.117, 118] or in the cooling gas [14.119], whereby the cooling is simple and very effective. However, it is evident from (14.2) that the smallest fundamental noise is achievable in vacuum, where the Q-factors are more than 100 times larger than in air, and at low temperatures.

The best force resolution up to now, which is better than 1×10^{-18} N/Hz$^{1/2}$, has been achieved by *Mamin* et al. [14.120] in vacuum at a temperature below 300 mK. Due to the reduced thermal noise and the lower thermal drift, which results in a higher stability of the tip-sample gap and a better signal-to-noise ratio, the highest resolution is possible at low temperatures in ultrahigh vacuum with FM-SFM. A vertical rms-noise below 2 pm [14.121, 122] and a force resolution below 1 aN [14.120] have been reported.

Besides the reduced noise, the application of force detection at low temperatures is motivated by the increased stability and the possibility to observe phenomena, which appear below a certain critical temperature T_c, as outlined at page 414. The experiments, which have been performed at low temperatures up until now, were motivated by at least one of these reasons and can be roughly divided into four groups: (i) atomic-scale imaging, (ii) force spectroscopy, (iii) investigation of quantum phenomena by measuring electrostatic forces, and (iv) utilizing magnetic probes to study ferromagnets, superconductors, and single spins. In the following, we describe some exemplary results.

14.4.1 Atomic-Scale Imaging

In a simplified picture, the dimensions of the tip end and its distance to the surface limit the lateral resolution of force microscopy, since it is a near-field technique. Consequently, atomic resolution requires a stable single atom at the tip apex that has to be brought within a distance of some tenths of a nanometer to an atomically flat surface. Both conditions can be fulfilled by FM-AFM, and nowadays *true* atomic resolution is routinely obtained (see Chapt. 13). However, the above statement says nothing about the nature and the underlying processes of imaging with atomic resolution. Si(111)-(7 × 7) was the first surface on which true atomic resolution was achieved [14.123], and several studies have been performed at low temperatures on this well-known material [14.124–126]. First principle simulations performed on semiconductors with a silicon tip revealed that *chemical* interactions, i. e., a significant charge redistribution between the dangling bonds of tip and sample, dominate the atomic-scale contrast [14.127–129]. On V-III semiconductors, it was found that only one atomic species, the group V atoms, is imaged as protrusions with a silicon tip [14.128, 129]. Furthermore, these simulations revealed that the sample, as well as the tip atoms are noticeably displaced from their equilibrium position due to the interaction forces. At low temperatures, both aspects could be observed with silicon tips on indium arsenide [14.121, 130].

Chemical Sensitivity of Force Microscopy

The (110) surface of the III-V semiconductor indium arsenide exhibits both atomic species in the top layer (see Fig. 14.18a). Therefore, this sample is well suited to study the chemical sensitivity of force microscopy [14.121]. In Fig. 14.18b, the usually observed atomic-scale contrast on InAs(110) is displayed. As predicted, the arsenic atoms, which are shifted by 80 pm above the indium layer due to the (1 × 1) relaxation, are imaged as protrusions. While this general appearance was similar for most tips, two other distinctively different contrasts were also observed: a second protru-

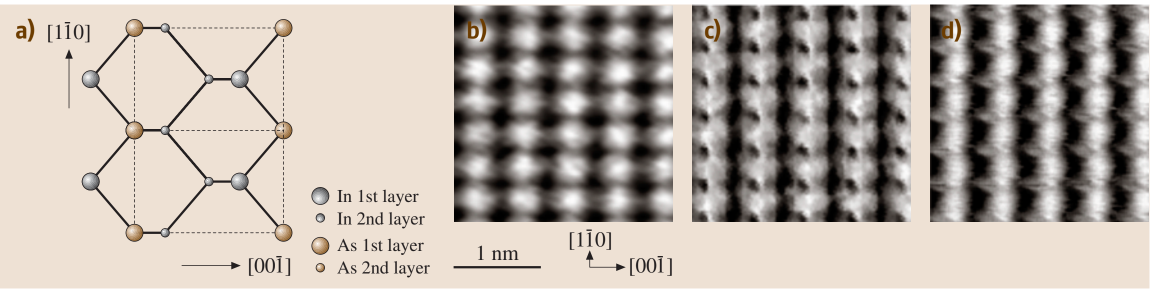

Fig. 14.18a–d The structure of InAs(110) as seen from top **(a)** and three FM-AFM images of this surface obtained with different tips at 14 K **(b)–(d)**. In **(b)**, only the arsenic atoms are imaged as protrusions, as predicted for a silicon tip. The two features in **(c)** and **(d)** corresponds to the zigzag arrangement of the indium and arsenic atoms. Since force microscopy is sensitive to short-range chemical forces, the appearance of the indium atoms can be associated with a chemically different tip apex

sion (c) and a sharp depression (d). The arrangement of these two features corresponds well to the zigzag configuration of the indium and arsenic atoms along the [1$\bar{1}$0]-direction. A sound explanation would be as follows: The contrast usually obtained with one feature per surface unit cell corresponds to a silicon-terminated tip, as predicted by simulations. A different atomic species at the tip apex, however, can result in a very different charge redistribution. Since the atomic-scale contrast is due to a chemical interaction, the two other contrasts would then correspond to a tip that has been accidentally contaminated with sample material (arsenic or indium-terminated tip apex). Nevertheless, this explanation has not yet been verified by simulations for this material.

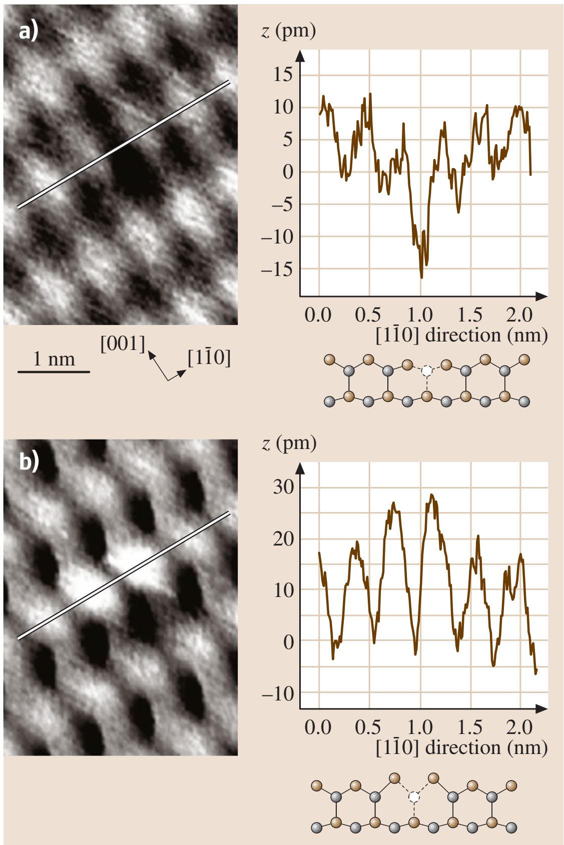

Fig. 14.19a,b Two FM-AFM images of the identical indium-site point defect (presumably an indium vacancy) recorded at 14 K. If the tip is relatively far away, the theoretically predicted inward relaxation of two arsenic atoms adjacent to an indium vacancy is visible **(a)**. At a closer tip-sample distance **(b)**, the two arsenic atoms are pulled farther toward the tip compared to the other arsenic atoms, since they have only two instead of three bonds

Tip-Induced Atomic Relaxation

Schwarz et al. [14.121] were able to directly visualize the predicted tip-induced relaxation during atomic-scale imaging near a point defect. Figure 14.19 shows two FM-AFM images of the same point defect recorded with different constant frequency shifts on InAs(110), i. e., the tip was closer to the surface in (b) compared to (a). The arsenic atoms are imaged as protrusions with the used silicon tip. From the symmetry of the defect, an indium-site defect can be inferred, since the distance-dependent contrast is consistent with what has to be expected for an indium vacancy. This expectation is based on calculations performed for the similar III-V semiconductor GaP(110), where the two surface gallium atoms around a P-vacancy were found to relax downward [14.131]. This corresponds to the situation in Fig. 14.19a, where the tip is relatively far away and an inward relaxation of the two arsenic atoms is observed. The considerably larger attractive force in Fig. 14.19b, however, pulls the two arsenic atoms toward the tip. All other arsenic atoms are also pulled, but they are less displaced, because they have three bonds to the bulk, while the two arsenic atoms in the neighborhood of an indium vacancy have only two bonds. This direct experimental proof of the presence of tip-induced relaxations is also relevant for STM measurements, because the tip-sample distances are similar during atomic resolution imaging. Moreover, the result demonstrates that FM-AFM can probe elastic properties on an atomic level.

Imaging of Weakly Interacting van der Waals Surfaces

For weakly interacting van der Waals surfaces like the (0001) surface of graphite, much smaller atomic corrugation amplitudes are expected compared to strongly interacting surfaces such as semiconductors. In graphite, a layered material, the carbon atoms are covalently bonded and arranged in a honeycomb structure within the (0001) plane. Individual layers of (0001) planes ($ABA\ldots$ stacking) stick together only by weak van der Waals forces. Three distinctive sites exist on the surface: carbon atoms with (A-type) and without (B-type) a neighbor in the next graphite layer and the *hollow site* (H-site) in the hexagon center. Neither static contact force microscopy nor STM can resolve the three different sites. For both methods, the contrast is well understood and exhibits protrusions with a sixfold symmetry and a periodicity of 246 pm. In high resolution FM-AFM images acquired at low temperatures, a large maximum and two different minima have been resolved (see Fig. 14.20a). A simulation using the Lennard-Jones potential, in which the attractive part is given by the short-range interatomic van der Waals force, reproduced these three features very well. By comparison with the experimental data (cf. the solid and the dotted line in the section of Fig. 14.20a), the large maximum could be assigned to the hollow site, while the two different minima represent A- and B-type carbon atoms [14.132]. Note that the image contrast is inverted with respect the arrangement of the atoms, i. e., the minima correspond to the position of the carbon atoms. The carbon-carbon distance of only 142 pm is the smallest interatomic distance that has been resolved with FM-AFM so far. Moreover, the weak tip-sample interaction results in typical corrugation amplitudes around 10 pm peak-to-peak, i. e., just above the noise level [14.7].

While experiments on graphite basically gain advantages from the increased stability and signal-to-noise ratio at low temperatures, solid xenon (melting temperature $T_{\mathrm{m}} = 161$ K) can only be observed at sufficiently low temperatures [14.8]. In addition, xenon is a pure van der Waals crystal, and since it is an insulator, FM-AFM is the only real space method available today that allows the study of solid xenon on the atomic scale.

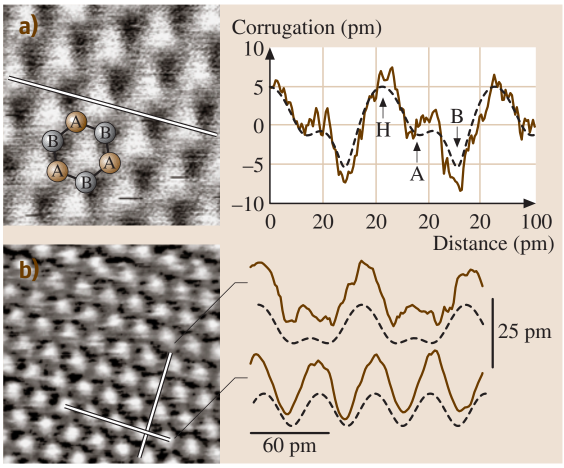

Fig. 14.20a,b FM-AFM images of graphite(0001) **(a)** and Xe(111) **(b)** recorded at 22 K. On the right side, line sections taken from the experimental data (*solid lines*) are compared to simulations (*dotted lines*). A- and B-type carbon atoms, as well as the hollow site (H-site) on graphite can be distinguished, but are imaged with inverted contrast, i. e., the carbon sites are displayed as minima. Such an inversion does not occur on Xe(111)

Allers et al. [14.8] adsorbed a well-ordered xenon film on cold graphite(0001) ($T < 55$ K) and subsequently studied it at 22 K by FM-AFM (see Fig. 14.20b). The six-fold symmetry and the distance between the protrusions correspond well with the nearest neighbor distance in the closed, packed (111) plane of bulk xenon, which crystallizes in the face-centered cubic structure. A comparison between experiment and simulation confirmed that the protrusions correspond to the position of the xenon atoms [14.133]. However, the simulated corrugation amplitudes do not fit as well as for graphite (see sections in Fig. 14.20b). A possible reason is that tip-induced relaxations, which were not considered in the simulations, are more important for this pure van der Waals crystal xenon than they are for graphite, because in-plane graphite exhibits strong covalent bonds. Nevertheless, the results demonstrated that a weakly bonded van der Waals crystal could be imaged nondestructively on the atomic scale for the first time.

14.4.2 Force Spectroscopy

A wealth of information about the nature of the tip-sample interaction can be obtained by measuring its distance dependence. This is usually done by recording the measured quantity (deflection, frequency shift, amplitude change, phase shift) and applying an appropriate voltage ramp to the z-electrode of the scanner piezo, while the z-feedback is switched off. According to (14.2), low temperatures and high Q-factors (vacuum) considerably increase the force resolution. In the static mode, long-range forces and contact forces can be examined. Force measurements at small tip-sample distances are inhibited by the *jump-to-contact* phenomenon: If the force gradient $\partial F_{\mathrm{ts}}/\partial z$ becomes larger than the spring constant c_z, the cantilever cannot resist the attractive tip-sample forces and the tip snaps onto the surface. Sufficiently large spring constants prevent this effect, but reduce the force resolution. In the dynamic modes, the jump-to-contact can be avoided due to the additional restoring force ($c_z A$) at the lower turnaround point. The highest sensitivity can be achieved in vacuum by using the FM technique, i. e., by recording $\Delta f(z)$-curves. An alternative FM spectroscopy method, the recording of $\Delta f(A)$-curves, has been suggested by *Hölscher* et al. [14.134]. Note that if the amplitude is much larger than the characteristic decay length of the tip-sample force, the frequency shift cannot simply be converted into force gradients by using $\partial F_{\mathrm{ts}}/\partial z = 2c_z \cdot \Delta f / f_0$ [14.135]. Several methods have been published to convert $\Delta f(z)$ data into the tip-sample potential $V_{\mathrm{ts}}(z)$ and tip-sample force $F_{\mathrm{ts}}(z)$ (see, e.g., [14.136, 137]).

Measurement of Interatomic Forces at Specific Atomic Sites

FM force spectroscopy has been successfully used to measure and determine quantitatively the short-range chemical force between the foremost tip atom and specific surface atoms [14.109, 138, 139]. Figure 14.21 displays an example for the quantitative determination of the short-range force. Figure 14.21a shows two $\Delta f(z)$-curves measured with a silicon tip above a corner hole and above an adatom. Their position is indicated by arrows in the inset, which displays the atomically resolved Si(111)-(7×7) surface. The two curves differ from each other only for small tip-sample distances, because the long-range forces do not contribute to the atomic-scale contrast. The low, thermally induced lateral drift and the high stability at low temperatures were required to precisely address the two specific sites. To extract the short-range force, the long-range van der Waals and/or electrostatic forces can be subtracted from the total force. The grey curve in Fig. 14.21b has been reconstructed from the $\Delta f(z)$-curve recorded above an adatom and represents the total force. After removing the long-range contribution from the data, the much steeper black line is obtained, which corresponds to the short-range force between the adatom and the atom at the tip apex. The measured maximum attractive force (−2.1 nN) agrees well with first principle calculations (−2.25 nN).

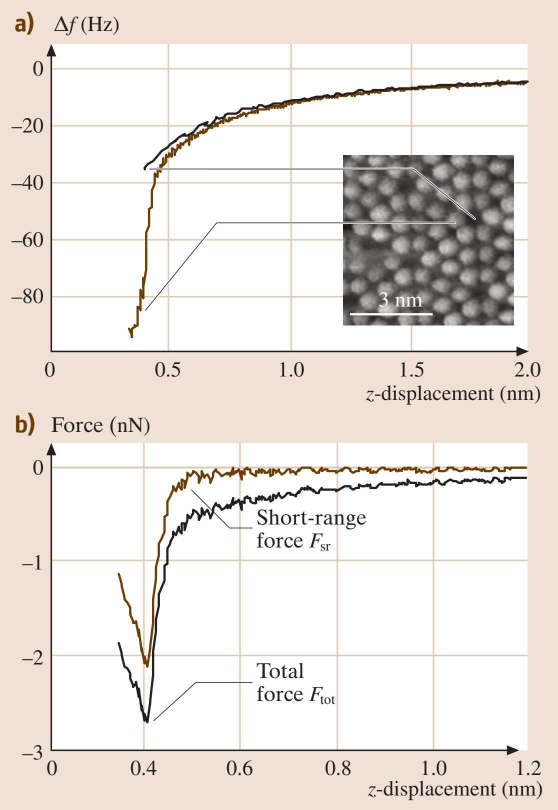

Fig. 14.21a,b FM force spectroscopy on specific atomic sites at 7.2 K. In **(a)**, an FM-SFM image of the Si(111)-(7×7) surface is displayed together with two $\Delta f(z)$-curves, which have been recorded at the positions indicated by the *arrows*, i. e., above the cornerhole (*brown*) and above an adatom (*black*). In **(b)**, the total force above an adatom (*brown line*) has been recovered from the $\Delta f(z)$-curve. After subtraction of the long-range part, the short-range force can be determined (*black line*) (courtesy of H. J. Hug; cf. [14.138])

Three-dimensional Force Field Spectroscopy

Further progress with the FM technique has been made by *Hölscher* et al. [14.140]. They acquired a complete 3-D force field on NiO(001) with atomic resolution (*3-D force field spectroscopy*). In Fig. 14.22a, the atomically resolved FM-AFM image of NiO(001) is shown together with the used coordinate system and the tip to illustrate the measurement principle. NiO(001) crystallizes in the rock salt structure. The distance between the protrusions corresponds to the lattice constant of 417 pm, i. e., only one type of atom (most likely the oxygen) is imaged as protrusion. In an area of 1 nm × 1 nm, 32 × 32 individual $\Delta f(z)$-curves have been recorded at every (x, y) image point and converted into $F_{\mathrm{ts}}(z)$-curves. The $\Delta f(x, y, z)$ data set is thereby converted into the 3-D force field $F_{\mathrm{ts}}(x, y, z)$. Figure 14.22b, where a specific x-z-plane is displayed, demonstrates that atomic resolution is achieved. It represents a 2-D cut $F_{\mathrm{ts}}(x, y = \mathrm{const}, z)$ along the [100]-direction (corresponding to the shaded slice marked in Fig. 14.22a). Since a large number of curves have been recorded, *Langkat* et al. [14.139] could evaluate the whole data set by standard statistical means to extract the long- and short-range forces. A possible future application of 3-D force field spectroscopy could be to map the short-range forces of complex molecules with functionalized tips in order to locally resolve their chemical reactivity.

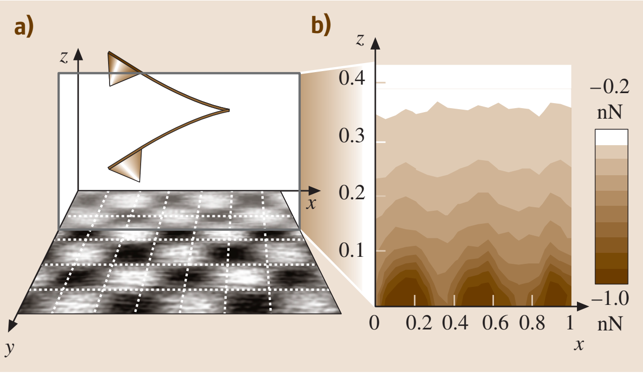

Fig. 14.22a,b Principle of the 3-D force filed spectroscopy method **(a)** and a 2-D cut through the 3-D force field$F_{\mathrm{ts}}(x,y,z)$ recorded at 14 K **(b)**. At all 32×32 image points of the 1 nm × 1 nm large scan area on NiO(001), a $\Delta f(z)$-curve has been recorded. The obtained $\Delta f(x, y, z)$ data set is then converted into the 3-D tip-sample force field $F_{\mathrm{ts}}(x, y, z)$. The *shaded slice* $F_{\mathrm{ts}}(x, y = \mathrm{const}, z)$ in **(a)** corresponds to a cut along the [100]-direction and demonstrates that atomic resolution has been obtained, because the distance between the protrusions corresponds well to the lattice constant of nickel oxide

Noncontact Friction

Another approach to achieve small tip-sample distances in combination with high force sensitivity is to use soft springs in a perpendicular configuration. The much higher cantilever stiffness along the cantilever axis prevents the jump-to-contact, but the lateral resolution is limited by the magnitude of the oscillation amplitude. However, with such a setup at low temperatures, *Stipe* et al. [14.141] measured the distance dependence of the very small force due to noncontact friction between tip and sample in vacuum. The effect was attributed to electric charges, which are moved parallel to the surface by the oscillating tip. Since the topography was not recorded in situ, the influence of contaminants or surface steps remained unknown.

14.4.3 Electrostatic Force Microscopy

Electrostatic forces are readily detectable by a force microscope, because tip and sample can be regarded as two electrodes of a capacitor. If they are electrically connected via their back sides and have different work functions, electrons will flow between tip and sample until their Fermi levels are equalized. As a result, an electric field and, consequently, an attractive electrostatic force exists between them at zero bias. This *contact potential difference* can be balanced by applying an appropriate bias voltage. It has been demonstrated that individual doping atoms in semiconducting materials can be detected by electrostatic interactions due to the local variation of the surface potential around them [14.142, 143].

Detection of Edge Channels in the Quantum Hall Regime

At low temperatures, electrostatic force microscopy has been used to measure the electrostatic potential in the quantum Hall regime of a *two-dimensional electron gas* (2-DEG) buried in epitaxially grown GaAs/AlGaAs heterostructures [14.144–147]. In the 2-DEG, electrons can move freely in the x-y-plane, but they cannot move in z-direction. Electrical transport properties of a 2-DEG are very different compared to normal metallic conduction. Particularly, the Hall resistance $R_{\mathrm{H}} = h/ne^2$ (where h represents Planck's constant, e is the electron charge, and $n = 1, 2, \ldots$) is quantized in the quantum Hall regime, i. e., at sufficiently low temperatures ($T < 4\,\mathrm{K}$) and high magnetic fields (up to 20 T). Under these conditions, theoretical calculations predict the existence of *edge channels* in a Hall bar. A Hall bar is a strip conductor that is contacted in a specific way to allow longitudinal and transversal transport measurements in a perpendicular magnetic field. The current is not evenly distributed over the cross section of the bar, but passes mainly along rather thin paths close to the edges. This prediction has been verified by measuring profiles of the electrostatic potential across a Hall bar in different perpendicular external magnetic fields [14.144–146].

Figure 14.23a shows the experimental setup used to observe these edge channels on top of a Hall bar with a force microscope. The tip is positioned above the surface of a Hall bar under which the 2-DEG is buried. The direction of the magnetic field is oriented perpendicular to the 2-DEG. Note that although the 2-DEG is located several tens of nanometers below the surface, its influence on the electrostatic surface potential can be detected. In Fig. 14.23b, the results of scans perpendicular to the Hall bar are plotted against the magnitude of the external magnetic field. The value of the electrostatic potential is grey-coded in arbitrary units. In certain field ranges, the potential changes linearly across the Hall bar, while in other field ranges the potential drop is confined to the edges of the Hall bar. The predicted edge channels can explain this behavior. The periodicity of the phenomenon is related to the filling factor ν, i. e., the number of Landau levels that are filled with electrons (see also Sect. 14.3.4). Its value depends on $1/B$

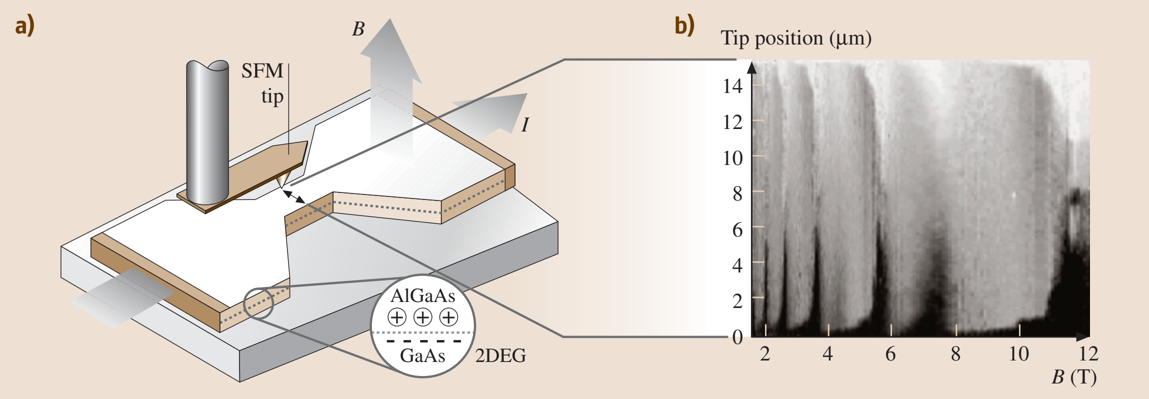

Fig. 14.23a,b Configuration of the Hall bar within a low temperature ($T < 1$ K) force microscope **(a)** and profiles (y-axis) at different magnetic field (x-axis) of the electrostatic potential across a 14-μm-wide Hall bar in the quantum Hall regime **(b)**. The external magnetic field is oriented perpendicular to the 2-DEG, which is buried below the surface. *Bright* and *dark regions* reflect the characteristic changes of the electrostatic potential across the Hall bar at different magnetic fields and can be explained by the existence of the theoretically predicted edge channels (courtesy of E. Ahlswede; cf. [14.146])

and is proportional to the electron concentration n_e in the 2-DEG ($\nu = n_e h/eB$, where h represents Planck's constant and e the electron charge).

14.4.4 Magnetic Force Microscopy

To detect magnetostatic tip-sample interactions with magnetic force microscopy (MFM), a ferromagnetic probe has to be used. Such probes are readily prepared by evaporating a thin magnetic layer, e.g., 10 nm iron, onto the tip. Due to the in-plane shape anisotropy of thin films, the magnetization of such tips lies predominantly along the tip axis, i. e., perpendicular to the surface. Since magnetostatic interactions are long-range, they can be separated from the topography by scanning in a certain constant height (typically around 20 nm) above the surface, where the z-component of the sample stray field is probed (see Fig. 14.24a). Therefore, MFM is always operated in noncontact. The signal from the cantilever is directly recorded while the z-feedback is switched off. MFM can be operated in the static mode or in the dynamic modes (AM-MFM at ambient pressures and FM-MFM in vacuum). A lateral resolution below 50 nm can be routinely obtained.

Observation of Domain Patterns

MFM is widely used to visualize domain patterns of ferromagnetic materials. At low temperatures, *Moloni* et al. [14.148] observed the domain structure of magnetite below its Verwey transition temperature ($T_\mathrm{V} = 122$ K), but most of the work concentrated on thin films of $\mathrm{La}_{1-x}\mathrm{Ca}_x\mathrm{MnO}_3$ [14.149–151]. Below T_V, the conductivity decreases by two orders of magnitude and a small structural distortion is observed. The domain structure of this mixed valence manganite is of great interest, because its resistivity strongly depends on the external magnetic field, i. e., it exhibits a large colossal magneto resistive effect. To investigate the field dependence of domain patterns in ambient conditions, electromagnets have to be used. They can cause severe thermal drift problems due to Joule-heating of the coils by large currents. Field on the order of 100 mT can be achieved. In contrast, much larger fields (more than 10 T) can be rather easily produced by implementing a superconducting magnet in low temperature setups. With such a design, *Liebmann* et al. [14.151] recorded the domain structure along the major hysteresis loop of $\mathrm{La}_{0.7}\mathrm{Ca}_{0.3}\mathrm{MnO}_3$ epitaxially grown on LaAlO_3 (see Fig. 14.24b–f). The film geometry (thickness is 100 nm) favors an in-plane magnetization, but the lattice mismatch with the substrate induces an out-of-plane anisotropy. Thereby, an irregular pattern of strip domains appears in zero field. If the external magnetic field is increased, the domains with anti-parallel orientation shrink and finally disappear in saturation (see Fig. 14.24b,c). The residual contrast in saturation (d) reflects topographic features. If the field is decreased after saturation (see Fig. 14.24e,f), cylindrical domains first nucleate and then start to grow. At zero field, the maze-type domain pattern has been evolved again.

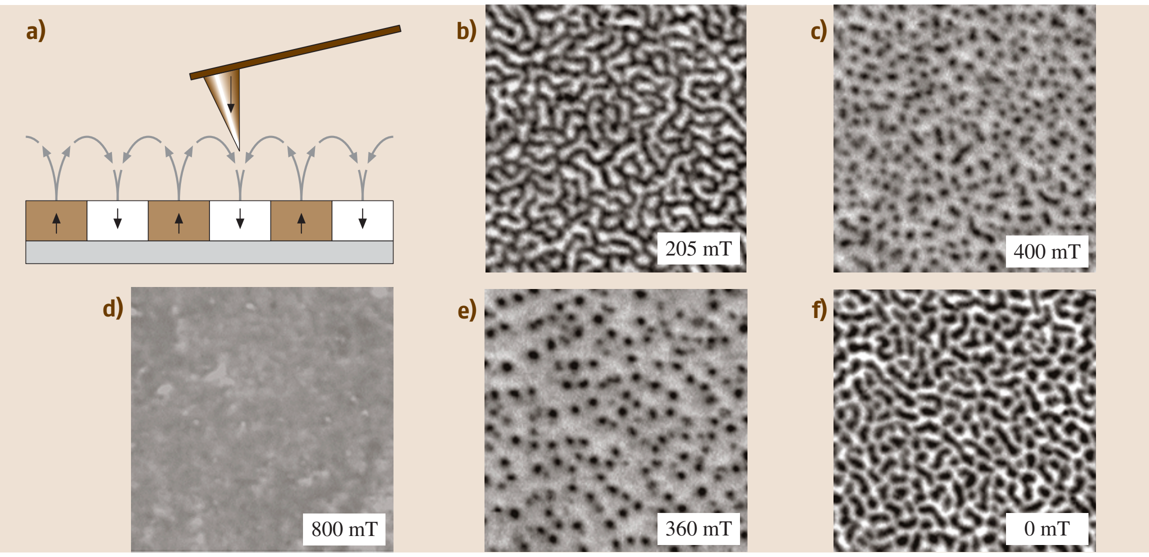

Fig. 14.24a–f Principle of MFM operation **(a)** and field-dependent domain structure of a ferromagnetic thin film **(b)–(f)** recorded at 5.2 K with FM-MFM. All images were recorded on the same 4 µm × 4 µm large scan area. The $La_{0.7}Ca_{0.3}MnO_3/LaAlO_3$ system exhibits a substrate-induced out-of-plane anisotropy. *Bright* and *dark areas* are visible and correspond to attractive and repulsive magnetostatic interactions, respectively. The series shows how the domain pattern evolves along the major hysteresis loop from, i. e., zero field to saturation at 600 mT and back to zero field

Such data sets can be used to analyze domain nucleation and the domain growth mode. Moreover, due to the negligible drift, domain structure and surface morphology can be directly compared, because every MFM can be used as a regular topography-imaging force microscope.

Detection of Individual Vortices in Superconductors

Numerous low temperature MFM experiments have been performed on superconductors [14.152–159]. Some basic features of superconductors have been mentioned already in Sect. 14.3.5. The main difference of STM/STS compared to MFM is the high sensitivity to the electronic properties of the surface. Therefore, careful sample preparation is a prerequisite. This is not so important for MFM experiments, since the tip is scanned at a certain distance above the surface.

Superconductors can be divided into two classes with respect to their behavior in an external magnetic field. For type-I superconductors, any magnetic flux is entirely excluded below their critical temperature T_c (Meissner effect), while for type-II superconductors, cylindrical inclusions (*vortices*) of normal material exist in a superconducting matrix (*vortex* state). The radius of the vortex *core*, where the Cooper pair density decreases to zero, is on the order of the coherence length ξ. Since the superconducting gap vanishes in the core, they can be detected by STS (see Sect. 14.3.5). Additionally, each vortex contains one magnetic quantum flux $\Phi = h/2e$ (where h represents Planck's constant and e the electron charge). Circular supercurrents around the core screen the magnetic field associated with a vortex; their radius is given by the London penetration depth λ of the material. This magnetic field of the vortices can be detected by MFM. Investigations have been performed on the two most popular copper oxide high-T_c superconductors, $YBa_2Cu_3O_7$ [14.152, 153, 155] and $Bi_2Sr_2CaCu_2O_8$ [14.153, 159], on the only elemental conventional type-II superconductor Nb [14.156, 157], and on the layered compound crystal $NbSe_2$ [14.154, 156].

Most often, vortices have been generated by cooling the sample from the normal state to below T_c in an external magnetic field. After such a *field cooling* procedure, the most energetically favorable vortex arrangement is a regular triangular Abrikosov lattice. *Volodin* et al. [14.154] were able to observe such an Abrikosov lattice on $NbSe_2$. The intervortex distance d is related to the external field during B cool down

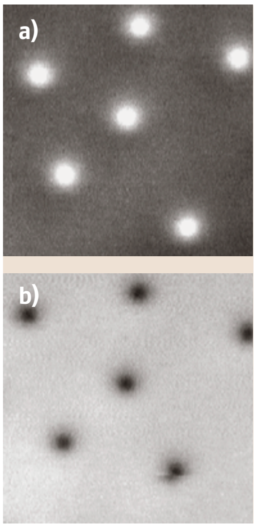

Fig. 14.25a,b Two $5\,\mu\text{m} \times 5\,\mu\text{m}$ FM-MFM images of vortices in a niobium thin film after field-cooling at 0.5 mT **(a)** and -0.5 mT **(b)**, respectively. Since the external magnetic field was parallel in **(a)** and antiparallel in **(b)** with respect to the tip magnetization, the vortices exhibit reversed contrast. Strong pinning dominates the position of the vortices, since they appear at identical locations in **(a)** and **(b)** and are not arranged in a regular Abrikosov lattice (courtesy of P. Grütter; cf. [14.157])

via $d = (4/3)^{1/4}(\Phi/B)^{1/2}$. Another way to introduce vortices into a type-II superconductor is vortex penetration from the edge by applying a magnetic field at temperatures below T_c. According to the Bean model, a vortex density gradient exists under such conditions within the superconducting material. *Pi* et al. [14.159] slowly increased the external magnetic field until the vortex front approaching from the edge reached the scanning area.

If the vortex configuration is dominated by the *pinning* of vortices at randomly distributed structural defects, no Abrikosov lattice emerges. The influence of pinning centers can be studied easily by MFM, because every MFM can be used to scan the topography in its AFM mode. This has been done for natural growth defects by *Moser* et al. [14.155] on $YBa_2Cu_3O_7$ and for $YBa_2Cu_3O_7$ and niobium thin films, respectively, by *Volodin* et al. [14.158]. *Roseman* et al. [14.160] investigated the formation of vortices in the presence of an artificial structure on niobium films, while *Pi* et al. [14.159] produced columnar defects by heavy ion bombardment in a $Bi_2Sr_2CaCu_2O_8$ single crystal to study the strong pinning at these defects.

Figure 14.25 demonstrates that MFM is sensitive to the polarity of vortices. In Fig. 14.25a, six vortices have been produced in a niobium film by field cooling in $+0.5$ mT. External magnetic field and tip magnetization are parallel, and, therefore, the tip-vortex interaction is attractive (bright contrast). To remove the vortices, the niobium was heated above T_c (≈ 9 K). Thereafter, vortices of opposite polarity were produced by field cooling in -0.5 mT, which appear dark in Fig. 14.25b. The vortices are probably bound to strong pinning sites, because the vortex positions are identical in both images of Fig. 14.25. By imaging the vortices at different scanning heights, *Roseman* et al. [14.157] tried to extract values for the London penetration depth from the scan height dependence of their profiles. While good qualitative agreement with theoretical predictions has been found, the absolute values do not agree with published literature values. The disagreement was attributed to the convolution between tip and vortex stray field. Better values might be obtained with calibrated tips.

Toward Single Spin Detection

So far, only collective magnetic phenomena like ferromagnetic domains have been observed via magnetostatic tip-sample interactions detected by MFM. However, magnetic ordering exists due to the exchange interaction between the electron spins of neighboring atoms in a solid. The most energetically favorable situation can be either ferromagnetic (parallel orientation) or anti-ferromagnetic (anti-parallel orientation) ordering. It has been predicted that the exchange force between an individual spin of a magnetically ordered sample and the spin of the foremost atom of a magnetic tip can be detected at sufficiently small tip-sample distances [14.161, 162].

The experimental realization, however, is very difficult, because the exchange force is about a factor of ten weaker and of even shorter range than the chemical interactions that are responsible for the atomic-scale contrast. FM-AFM experiments with a ferromagnetic tip have been performed on the anti-ferromagnetic NiO(001) surface at room temperature [14.163] and with a considerable better signal-to-noise ratio at low temperatures [14.122]. Although it was possible to achieve atomic resolution, a periodic contrast that could be attributed to the anti-ferromagnetically ordered spins of the nickel atoms could not be observed.

Even more ambitious is the proposed detection of individual nuclear spins by magnetic resonance force microscopy (MRFM) using a magnetic tip [14.164, 165]. The interest in this technique is driven by the possibility of reaching subsurface true 3-D imaging with atomic resolution and chemical specificity. The idea of MRFM is to combine aspects of force microscopy (atomic resolution capability) with nuclear magnetic resonance (3-D imaging and elemental selectivity). Low temperatures are required, because the forces are extremely small.

Up until now, no individual spins have been detected by MRFM. However, electron spin resonance [14.166, 167], nuclear magnetic resonance [14.168], and ferromagnetic resonance [14.169] experiments of spin ensembles have been performed with micrometer resolution. On the way toward single spin detection, the design of ultrasensitive cantilevers made considerable progress, and the detection of forces below $1\times10^{-18}\,\mathrm{N}$ has been achieved [14.120].

References

14.1 G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel: Surface studies by scanning tunneling microscopy, Phys. Rev. Lett. **49** (1982) 57–61

14.2 R. Wiesendanger: *Scanning Probe Microscopy and Spectroscopy* (Cambridge Univ. Press, Cambridge 1994)

14.3 M. Tinkham: *Introduction to Superconductivity* (McGraw-Hill, New York 1996)

14.4 J. Kondo: Theory of dilute magnetic alloys, Solid State Phys. **23** (1969) 183–281

14.5 T. R. Albrecht, P. Grütter, H. K. Horne, D. Rugar: Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity, J. Appl. Phys. **69** (1991) 668–673

14.6 F. J. Giessibl, H. Bielefeld, S. Hembacher, J. Mannhart: Calculation of the optimal imaging parameters for frequency modulation atomic force microscopy, Appl. Surf. Sci. **140** (1999) 352–357

14.7 W. Allers, A. Schwarz, U. D. Schwarz, R. Wiesendanger: Dynamic scanning force microscopy at low temperatures on a van der Waals surface: graphite(0001), Appl. Surf. Sci. **140** (1999) 247–252

14.8 W. Allers, A. Schwarz, U. D. Schwarz, R. Wiesendanger: Dynamic scanning force microscopy at low temperatures on a noble-gas crystal: atomic resolution on the xenon(111) surface, Europhys. Lett. **48** (1999) 276–279

14.9 M. Morgenstern, D. Haude, V. Gudmundsson, C. Wittneven, R. Dombrowski, R. Wiesendanger: Origin of Landau oscillations observed in scanning tunneling spectroscopy on *n*-InAs(110), Phys. Rev. B **62** (2000) 7257–7263

14.10 D. M. Eigler, P. S. Weiss, E. K. Schweizer, N. D. Lang: Imaging Xe with a low-temperature scanning tunneling microscope, Phys. Rev. Lett. **66** (1991) 1189–1192

14.11 P. S. Weiss, D. M. Eigler: Site dependence of the apparent shape of a molecule in scanning tunneling micoscope images: Benzene on Pt{111}, Phys. Rev. Lett. **71** (1992) 3139–3142

14.12 D. M. Eigler, E. K. Schweizer: Positioning single atoms with a scanning tunneling microscope, Nature **344** (1990) 524–526

14.13 H. Hug, B. Stiefel, P. J. A. van Schendel, A. Moser, S. Martin, H.-J. Güntherodt: A low temperature ultrahigh vacuum scanning force microscope, Rev. Sci. Instrum. **70** (1999) 3627–3640

14.14 S. Behler, M. K. Rose, D. F. Ogletree, F. Salmeron: Method to characterize the vibrational response of a beetle type scanning tunneling microscope, Rev. Sci. Instrum. **68** (1997) 124–128

14.15 C. Wittneven, R. Dombrowski, S. H. Pan, R. Wiesendanger: A low-temperature ultrahigh-vacuum scanning tunneling microscope with rotatable magnetic field, Rev. Sci. Instrum. **68** (1997) 3806–3810

14.16 W. Allers, A. Schwarz, U. D. Schwarz, R. Wiesendanger: A scanning force microscope with atomic resolution in ultrahigh vacuum and at low temperatures, Rev. Sci. Instrum. **69** (1998) 221–225

14.17 G. Dujardin, R. E. Walkup, Ph. Avouris: Dissociation of individual molecules with electrons from the tip of a scanning tunneling microscope, Science **255** (1992) 1232–1235

14.18 H. J. Lee, W. Ho: Single-bond formation and characterization with a scanning tunneling microscope, Science **286** (1999) 1719–1722

14.19 R. Berndt, R. Gaisch, J. K. Gimzewski, B. Reihl, R. R. Schlittler, W. D. Schneider, M. Tschudy: Photon emission at molecular resolution induced by a scanning tunneling microscope, Science **262** (1993) 1425–1427

14.20 B. G. Briner, M. Doering, H. P. Rust, A. M. Bradshaw: Microscopic diffusion enhanced by adsorbate interaction, Science **278** (1997) 257–260

14.21 J. Kliewer, R. Berndt, E. V. Chulkov, V. M. Silkin, P. M. Echenique, S. Crampin: Dimensionality effects in the lifetime of surface states, Science **288** (2000) 1399–1401

14.22 M. F. Crommie, C. P. Lutz, D. M. Eigler: Imaging standing waves in a two-dimensional electron gas, Nature **363** (1993) 524–527

14.23 B. C. Stipe, M. A. Rezaei, W. Ho: Single-molecule vibrational spectroscopy and microscopy, Science **280** (1998) 1732–1735

14.24 H. J. Lee, W. Ho: Structural determination by single-molecule vibrational spectroscopy and microscopy: Contrast between copper and iron carbonyls, Phys. Rev. B **61** (2000) R16347–R16350

14.25 C. W. J. Beenakker, H. van Houten: Quantum transport in semiconductor nanostructures, Solid State Phys. **44** (1991) 1–228

14.26 S. H. Pan, E. W. Hudson, K. M. Lang, H. Eisaki, S. Uchida, J. C. Davis: Imaging the effects of individual zinc impurity atoms on superconductivity in $Bi_2Sr_2CaCu_2O_{8+\delta}$, Nature **403** (2000) 746–750

14.27 R.S. Becker, J.A. Golovchenko, B.S. Swartzentruber: Atomic-scale surface modifications using a tunneling microscope, Nature **325** (1987) 419–42

14.28 J.A. Stroscio, D.M. Eigler: Atomic and molecular manipulation with the scanning tunneling microscope, Science **254** (1991) 1319–1326

14.29 L. Bartels, G. Meyer, K.H. Rieder: Basic steps of lateral manipulation of single atoms and diatomic clusters with a scanning tunneling microscope, Phys. Rev. Lett. **79** (1997) 697–700

14.30 J.J. Schulz, R. Koch, K.H. Rieder: New mechanism for single atom manipulation, Phys. Rev. Lett. **84** (2000) 4597–4600

14.31 T.C. Shen, C. Wang, G.C. Abeln, J.R. Tucker, J.W. Lyding, Ph. Avouris, R.E. Walkup: Atomic-scale desorption through electronic and vibrational excitation mechanisms, Science **268** (1995) 1590–1592

14.32 T. Komeda, Y. Kim, M. Kawai, B.N.J. Persson, H. Ueba: Lateral hopping of molecules induced by excitations of internal vibration mode, Science **295** (2002) 2055–2058

14.33 Y.W. Mo: Reversible rotation of antimony dimers on the silicon(001) surface with a scanning tunneling microscope, Science **261** (1993) 886–888

14.34 B.C. Stipe, M.A. Rezaei, W. Ho: Inducing and viewing the rotational motion of a single molecule, Science **279** (1998) 1907–1909

14.35 F. Moresco, G. Meyer, K.H. Rieder, H. Tang, A. Gourdon, C. Joachim: Conformational changes of single molecules by scanning tunneling microscopy manipulation: a route to molecular switching, Phys. Rev. Lett. **86** (2001) 672–675

14.36 S.W. Hla, L. Bartels, G. Meyer, K.H. Rieder: Inducing all steps of a chemical reaction with the scanning tunneling microscope tip: Towards single molecule engineering, Phys. Rev. Lett. **85** (2000) 2777–2780

14.37 E. Ganz, S.K. Theiss, I.S. Hwang, J. Golovchenko: Direct measurement of diffusion by hot tunneling microscopy: Activations energy, anisotropy, and long jumps, Phys. Rev. Lett. **68** (1992) 1567–1570

14.38 M. Schuhnack, T.R. Linderoth, F. Rosei, E. Laegsgaard, I. Stensgaard, F. Besenbacher: Long jumps in the surface diffusion of large molecules, Phys. Rev. Lett. **88** (2002) 156102, 1–4

14.39 L.J. Lauhon, W. Ho: Direct observation of the quantum tunneling of single hydrogen atoms with a scanning tunneling microscope, Phys. Rev. Lett. **85** (2000) 4566–4569

14.40 N. Kitamura, M. Lagally, M.B. Webb: Real-time observation of vacancy diffusion on Si(001)-(2 × 1) by scanning tunneling microscopy, Phys. Rev. Lett. **71** (1993) 2082–2085

14.41 M. Morgenstern, T. Michely, G. Comsa: Onset of interstitial diffusion determined by scanning tunneling microscopy, Phys. Rev. Lett. **79** (1997) 1305–1308

14.42 K. Morgenstern, G. Rosenfeld, B. Poelsema, G. Comsa: Brownian motion of vacancy islands on Ag(111), Phys. Rev. Lett. **74** (1995) 2058–2061

14.43 B. Reihl, J.H. Coombs, J.K. Gimzewski: Local inverse photoemission with the scanning tunneling microscope, Surf. Sci. **211–212** (1989) 156–164

14.44 R. Berndt, J.K. Gimzewski, P. Johansson: Inelastic tunneling excitation of tip-induced plasmon modes on noble-metal surfaces, Phys. Rev. Lett. **67** (1991) 3796–3799

14.45 P. Johansson, R. Monreal, P. Apell: Theory for light emission from a scanning tunneling microscope, Phys. Rev. B **42** (1990) 9210–9213

14.46 J. Aizpurua, G. Hoffmann, S.P. Apell, R. Berndt: Electromagnetic coupling on an atomic scale, Phys. Rev. Lett. **89** (2002) 156803, 1–4

14.47 G. Hoffmann, J. Kliewer, R. Berndt: Luminescence from metallic quantum wells in a scanning tunneling microscope, Phys. Rev. Lett. **78** (2001) 176803, 1–4

14.48 A. Downes, M.E. Welland: Photon emission from Si(111)-(7 × 7) induced by scanning tunneling microscopy: atomic scale and material contrast, Phys. Rev. Lett. **81** (1998) 1857–1860

14.49 M. Kemerink, K. Sauthoff, P.M. Koenraad, J.W. Geritsen, H. van Kempen, J.H. Wolter: Optical detection of ballistic electrons injected by a scanning-tunneling microscope, Phys. Rev. Lett. **86** (2001) 2404–2407

14.50 J. Tersoff, D.R. Hamann: Theory and application for the scanning tunneling microscope, Phys. Rev. Lett. **50** (1983) 1998–2001

14.51 C.J. Chen: *Introduction to Scanning Tunneling Microscopy* (Oxford Univ. Press, Oxford 1993)

14.52 J. Winterlin, J. Wiechers, H. Brune, T. Gritsch, H. Hofer, R.J. Behm: Atomic-resolution imaging of close-packed metal surfaces by scanning tunneling microscopy, Phys. Rev. Lett. **62** (1989) 59–62

14.53 A.L. Vazquez de Parga, O.S. Hernan, R. Miranda, A. Levy Yeyati, N. Mingo, A. Martin-Rodero, F. Flores: Electron resonances in sharp tips and their role in tunneling spectroscopy, Phys. Rev. Lett. **80** (1998) 357–360

14.54 S.H. Pan, E.W. Hudson, J.C. Davis: Vacuum tunneling of superconducting quasiparticles from atomically sharp scanning tunneling microscope tips, Appl. Phys. Lett. **73** (1998) 2992–2994

14.55 J.T. Li, W.D. Schneider, R. Berndt, O.R. Bryant, S. Crampin: Surface-state lifetime measured by scanning tunneling spectroscopy, Phys. Rev. Lett. **81** (1998) 4464–4467

14.56 L. Bürgi, O. Jeandupeux, H. Brune, K. Kern: Probing hot-electron dynamics with a cold scanning tunneling microscope, Phys. Rev. Lett. **82** (1999) 4516–4519

14.57 J. W. G. Wildoer, C. J. P. M. Harmans, H. van Kempen: Observation of Landau levels at the InAs(110) surface by scanning tunneling spectroscopy, Phys. Rev. B **55** (1997) R16013–R16016

14.58 M. Morgenstern, V. Gudmundsson, C. Wittneven, R. Dombrowski, R. Wiesendanger: Nonlocality of the exchange interaction probed by scanning tunneling spectroscopy, Phys. Rev. B **63** (2001) 201301(R), 1–4

14.59 M. V. Grishin, F. I. Dalidchik, S. A. Kovalevskii, N. N. Kolchenko, B. R. Shub: Isotope effect in the vibrational spectra of water measured in experiments with a scanning tunneling microscope, JETP Lett. **66** (1997) 37–40

14.60 A. Hewson: *From the Kondo Effect to Heavy Fermions* (Cambridge Univ. Press, Cambridge 1993)

14.61 V. Madhavan, W. Chen, T. Jamneala, M. F. Crommie, N. S. Wingreen: Tunneling into a single magnetic atom: Spectroscopic evidence of the Kondo resonance, Science **280** (1998) 567–569

14.62 J. Li, W. D. Schneider, R. Berndt, B. Delley: Kondo scattering observed at a single magnetic impurity, Phys. Rev. Lett. **80** (1998) 2893–2896

14.63 T. W. Odom, J. L. Huang, C. L. Cheung, C. M. Lieber: Magnetic clusters on single-walled carbon nanotubes: the Kondo effect in a one-dimensional host, Science **290** (2000) 1549–1552

14.64 M. Ouyang, J. L. Huang, C. L. Cheung, C. M. Lieber: Energy gaps in metallic single-walled carbon nanotubes, Science **292** (2001) 702–705

14.65 U. Fano: Effects of configuration interaction on intensities and phase shifts, Phys. Rev. **124** (1961) 1866–1878

14.66 H. C. Manoharan, C. P. Lutz, D. M. Eigler: Quantum mirages formed by coherent projection of electronic structure, Nature **403** (2000) 512–515

14.67 O. Y. Kolesnychenko, R. de Kort, M. I. Katsnelson, A. I. Lichtenstein, H. van Kempen: Real-space observation of an orbital Kondo resonance on the Cr(001) surface, Nature **415** (2002) 507–509

14.68 H. A. Mizes, J. S. Foster: Long-range electronic perturbations caused by defects using scanning tunneling microscopy, Science **244** (1989) 559–562

14.69 P. T. Sprunger, L. Petersen, E. W. Plummer, E. Laegsgaard, F. Besenbacher: Giant Friedel oscillations on beryllium (0001) surface, Science **275** (1997) 1764–1767

14.70 P. Hofmann, B. G. Briner, M. Doering, H. P. Rust, E. W. Plummer, A. M. Bradshaw: Anisotropic two-dimensional Friedel oscillations, Phys. Rev. Lett. **79** (1997) 265–268

14.71 E. J. Heller, M. F. Crommie, C. P. Lutz, D. M. Eigler: Scattering and adsorption of surface electron waves in quantum corrals, Nature **369** (1994) 464–466

14.72 M. C. M. M. van der Wielen, A. J. A. van Roij, H. van Kempen: Direct observation of Friedel oscillations around incorporated Si_{Ga} dopants in GaAs by low-temperature scanning tunneling microscopy, Phys. Rev. Lett. **76** (1996) 1075–1078

14.73 O. Millo, D. Katz, Y. W. Cao, U. Banin: Imaging and spectroscopy of artificial-atom states in core/shell nanocrystal quantum dots, Phys. Rev. Lett. **86** (2001) 5751–5754

14.74 L. C. Venema, J. W. G. Wildoer, J. W. Janssen, S. J. Tans, L. J. T. Tuinstra, L. P. Kouwenhoven, C. Dekker: Imaging electron wave functions of quantized energy levels in carbon nanotubes, Nature **283** (1999) 52–55

14.75 S. G. Lemay, J. W. Jannsen, M. van den Hout, M. Mooij, M. J. Bronikowski, P. A. Willis, R. E. Smalley, L. P. Kouwenhoven, C. Dekker: Two-dimensional imaging of electronic wavefunctions in carbon nanotubes, Nature **412** (2001) 617–620

14.76 C. Wittneven, R. Dombrowski, M. Morgenstern, R. Wiesendanger: Scattering states of ionized dopants probed by low temperature scanning tunneling spectroscopy, Phys. Rev. Lett. **81** (1998) 5616–5619

14.77 D. Haude, M. Morgenstern, I. Meinel, R. Wiesendanger: Local density of states of a three-dimensional conductor in the extreme quantum limit, Phys. Rev. Lett. **86** (2001) 1582–1585

14.78 R. Joynt, R. E. Prange: Conditions for the quantum Hall effect, Phys. Rev. B **29** (1984) 3303–3317

14.79 M. Morgenstern, J. Klijn, C. Meyer, M. Getzlaff, R. Adelung, R. A. Römer, K. Rossnagel, L. Kipp, M. Skibowski, R. Wiesendanger: Direct comparison between potential landscape and local density of states in a disordered two-dimensional electron system, Phys. Rev. Lett. **89** (2002) 136806, 1–4

14.80 E. Abrahams, P. W. Anderson, D. C. Licciardello, T. V. Ramakrishnan: Scaling theory of localization: absence of quantum diffusion in two dimensions, Phys. Rev. Lett. **42** (1979) 673–676

14.81 M. Morgenstern, J. Klijn, R. Wiesendanger: Real space observation of drift states in a two-dimensional electron system at high magnetic fields, Phys. Rev. Lett. **90** (2003) 056804, 1–4

14.82 R. E. Peierls: *Quantum Theory of Solids* (Clarendon, Oxford 1955)

14.83 C. G. Slough, W. W. McNairy, R. V. Coleman, B. Drake, P. K. Hansma: Charge-density waves studied with the use of a scanning tunneling microscope, Phys. Rev. B **34** (1986) 994–1005

14.84 X. L. Wu, C. M. Lieber: Hexagonal domain-like charge-density wave of TaS_2 determined by scanning tunneling microscopy, Science **243** (1989) 1703–1705

14.85 T. Nishiguchi, M. Kageshima, N. Ara-Kato, A. Kawazu: Behaviour of charge density waves in a one-dimensional organic conductor visualized by scanning tunneling microscopy, Phys. Rev. Lett. **81** (1998) 3187–3190

14.86 X. L. Wu, C. M. Lieber: Direct observation of growth and melting of the hexagonal-domain charge-density-wave phase in 1T-TaS_2 by scanning tunneling microscopy, Phys. Rev. Lett. **64** (1990) 1150–1153

14.87 J. M. Carpinelli, H. H. Weitering, E. W. Plummer, R. Stumpf: Direct observation of a surface charge density wave, Nature **381** (1996) 398–400

14.88 H. H. Weitering, J. M. Carpinelli, A. V. Melechenko, J. Zhang, M. Bartkowiak, E. W. Plummer: Defect-mediated condensation of a charge density wave, Science **285** (1999) 2107–2110

14.89 H. W. Yeom, S. Takeda, E. Rotenberg, I. Matsuda, K. Horikoshi, J. Schäfer, C. M. Lee, S. D. Kevan, T. Ohta, T. Nagao, S. Hasegawa: Instability and charge density wave of metallic quantum chains on a silicon surface, Phys. Rev. Lett. **82** (1999) 4898–4901

14.90 K. Swamy, A. Menzel, R. Beer, E. Bertel: Charge-density waves in self-assembled halogen-bridged metal chains, Phys. Rev. Lett. **86** (2001) 1299–1302

14.91 J. J. Kim, W. Yamaguchi, T. Hasegawa, K. Kitazawa: Observation of Mott localization gap using low temperature scanning tunneling spectroscopy in commensurate 1T-TaSe_2, Phys. Rev. Lett. **73** (1994) 2103–2106

14.92 J. Bardeen, L. N. Cooper, J. R. Schrieffer: Theory of superconductivity, Phys. Rev. **108** (1957) 1175–1204

14.93 A. Yazdani, B. A. Jones, C. P. Lutz, M. F. Crommie, D. M. Eigler: Probing the local effects of magnetic impurities on superconductivity, Science **275** (1997) 1767–1770

14.94 S. H. Tessmer, M. B. Tarlie, D. J. van Harlingen, D. L. Maslov, P. M. Goldbart: Probing the superconducting proximity effect in NbSe_2 by scanning tunneling micrsocopy, Phys. Rev. Lett **77** (1996) 924–927

14.95 K. Inoue, H. Takayanagi: Local tunneling spectroscopy of Nb/InAs/Nb superconducting proximity system with a scanning tunneling microscope, Phys. Rev. B **43** (1991) 6214–6215

14.96 H. F. Hess, R. B. Robinson, R. C. Dynes, J. M. Valles, J. V. Waszczak: Scanning-tunneling-microscope observation of the Abrikosov flux lattice and the density of states near and inside a fluxoid, Phys. Rev. Lett. **62** (1989) 214–217

14.97 H. F. Hess, R. B. Robinson, J. V. Waszczak: Vortex-core structure observed with a scanning tunneling microscope, Phys. Rev. Lett. **64** (1990) 2711–2714

14.98 N. Hayashi, M. Ichioka, K. Machida: Star-shaped local density of states around vortices in a type-II superconductor, Phys. Rev. Lett. **77** (1996) 4074–4077

14.99 H. Sakata, M. Oosawa, K. Matsuba, N. Nishida: Imaging of vortex lattice transition in YNi_2B_2C by scanning tunneling spectroscopy, Phys. Rev. Lett. **84** (2000) 1583–1586

14.100 S. Behler, S. H. Pan, P. Jess, A. Baratoff, H.-J. Güntherodt, F. Levy, G. Wirth, J. Wiesner: Vortex pinning in ion-irrediated NbSe_2 studied by scanning tunneling microscopy, Phys. Rev. Lett. **72** (1994) 1750–1753

14.101 R. Berthe, U. Hartmann, C. Heiden: Influence of a transport current on the Abrikosov flux lattice observed with a low-temperature scanning tunneling microscope, Ultramicroscopy **42–44** (1992) 696–698

14.102 A. Polkovnikov, S. Sachdev, M. Vojta: Impurity in a d-wave superconductor: Kondo effect and STM spectra, Phys. Rev. Lett. **86** (2001) 296–299

14.103 E. W. Hudson, K. M. Lang, V. Madhavan, S. H. Pan, S. Uchida, J. C. Davis: Interplay of magnetism and high-T_c superconductivity at individual Ni impurity atoms in $Bi_2Sr_2CaCu_2O_{8+\delta}$, Nature **411** (2001) 920–924

14.104 K. M. Lang, V. Madhavan, J. E. Hoffman, E. W. Hudson, H. Eisaki, S. Uchida, J. C. Davis: Imaging the granular structure of high-T_c superconductivity in underdoped $Bi_2Sr_2CaCu_2O_{8+\delta}$, Nature **415** (2002) 412–416

14.105 I. Maggio-Aprile, C. Renner, E. Erb, E. Walker, Ø. Fischer: Direct vortex lattice imaging and tunneling spectroscopy of flux lines on $YBa_2Cu_3O_{7-\delta}$, Phys. Rev. Lett. **75** (1995) 2754–2757

14.106 C. Renner, B. Revaz, K. Kadowaki, I. Maggio-Aprile, Ø. Fischer: Observation of the low temperature pseudogap in the vortex cores of $Bi_2Sr_2CaCu_2O_{8+\delta}$, Phys. Rev. Lett. **80** (1998) 3606–3609

14.107 S. H. Pan, E. W. Hudson, A. K. Gupta, K. W. Ng, H. Eisaki, S. Uchida, J. C. Davis: STM studies of the electronic structure of vortex cores in $Bi_2Sr_2CaCu_2O_{8+\delta}$, Phys. Rev. Lett. **85** (2000) 1536–1539

14.108 D. P. Arovas, A. J. Berlinsky, C. Kallin, S. C. Zhang: Superconducting vortex with antiferromagnetic core, Phys. Rev. Lett. **79** (1997) 2871–2874

14.109 J. E. Hoffmann, E. W. Hudson, K. M. Lang, V. Madhavan, H. Eisaki, S. Uchida, J. C. Davis: A four unit cell periodic pattern of quasi-particle states surrounding vortex cores in $Bi_2Sr_2CaCu_2O_{8+\delta}$, Science **295** (2002) 466–469

14.110 M. Fäth, S. Freisem, A. A. Menovsky, Y. Tomioka, J. Aaarts, J. A. Mydosh: Spatially inhomogeneous metal–insulator transition in doped manganites, Science **285** (1999) 1540–1542

14.111 C. Renner, G. Aeppli, B. G. Kim, Y. A. Soh, S. W. Cheong: Atomic-scale images of charge ordering in a mixed-valence manganite, Nature **416** (2000) 518–521

14.112 M. Bode, M. Getzlaff, R. Wiesendanger: Spin-polarized vacuum tunneling into the exchange-split surface state of Gd(0001), Phys. Rev. Lett. **81** (1998) 4256–4259

14.113 A. Kubetzka, M. Bode, O. Pietzsch, R. Wiesendanger: Spin-polarized scanning tunneling microscopy with antiferromagnetic probe tips, Phys. Rev. Lett. **88** (2002) 057201, 1–4

14.114 O. Pietzsch, A. Kubetzka, M. Bode, R. Wiesendanger: Observation of magnetic hysteresis at the nanometer scale by spin-polarized scanning tunneling spectroscopy, Science **292** (2001) 2053–2056

14.115 S. Heinze, M. Bode, A. Kubetzka, O. Pietzsch, X. Xie, S. Blügel, R. Wiesendanger: Real-space imaging of two-dimensional antiferromagnetism on the atomic scale, Science **288** (2000) 1805–1808

14.116 A. Wachowiak, J. Wiebe, M. Bode, O. Pietzsch, M. Morgenstern, R. Wiesendanger: Internal spin-structure of magnetic vortex cores observed by spin-polarized scanning tunneling microscopy, Science **298** (2002) 577–580

14.117 M. D. Kirk, T. R. Albrecht, C. F. Quate: Low-temperature atomic force microscopy, Rev. Sci. Instrum. **59** (1988) 833–835

14.118 D. Pelekhov, J. Becker, J. G. Nunes: Atomic force microscope for operation in high magnetic fields at milliKelvin temperatures, Rev. Sci. Instrum. **70** (1999) 114–120

14.119 J. Mou, Y. Jie, Z. Shao: An optical detection low temperature atomic force microscope at ambient pressure for biological research, Rev. Sci. Instrum. **64** (1993) 1483–1488

14.120 H. J. Mamin, D. Rugar: Sub-attoNewton force detection at milliKelvin temperatures, Appl. Phys. Lett. **79** (2001) 3358–3360

14.121 A. Schwarz, W. Allers, U. D. Schwarz, R. Wiesendanger: Dynamic mode scanning force microscopy of *n*-InAs(110)-(1 × 1) at low temperatures, Phys. Rev. B **61** (2000) 2837–2845

14.122 W. Allers, S. Langkat, R. Wiesendanger: Dynamic low-temperature scanning force microscopy on nickel oxide(001), Appl. Phys. A **72** (2001) S27–S30

14.123 F. J. Giessibl: Atomic resolution of the silicon(111)-(7 × 7) surface by atomic force microscopy, Science **267** (1995) 68–71

14.124 M. A. Lantz, H. J. Hug, P. J. A. van Schendel, R. Hoffmann, S. Martin, A. Baratoff, A. Abdurixit, H.-J. Güntherodt: Low temperature scanning force microscopy of the Si(111)-(7 × 7) surface, Phys. Rev. Lett. **84** (2000) 2642–2465

14.125 K. Suzuki, H. Iwatsuki, S. Kitamura, C. B. Mooney: Development of low temperature ultrahigh vacuum force microscope/scanning tunneling microscope, Jpn. J. Appl. Phys. **39** (2000) 3750–3752

14.126 N. Suehira, Y. Sugawara, S. Morita: Artifact and fact of Si(111)-(7 × 7) surface images observed with a low temperature noncontact atomic force microscope (LT-NC-AFM), Jpn. J. Appl. Phys. **40** (2001) 292–294

14.127 R. Peréz, M. C. Payne, I. Štich, K. Terakura: Role of covalent tip-surface interactions in noncontact atomic force microscopy on reactive surfaces, Phys. Rev. Lett. **78** (1997) 678–681

14.128 S. H. Ke, T. Uda, R. Pérez, I. Štich, K. Terakura: First principles investigation of tip-surface interaction on GaAs(110): Implication for atomic force and tunneling microscopies, Phys. Rev. B **60** (1999) 11631–11638

14.129 J. Tobik, I. Štich, R. Peréz, K. Terakura: Simulation of tip-surface interactions in atomic force microscopy of an InP(110) surface with a Si tip, Phys. Rev. B **60** (1999) 11639–11644

14.130 A. Schwarz, W. Allers, U. D. Schwarz, R. Wiesendanger: Simultaneous imaging of the In and As sublattice on InAs(110)-(1 × 1) with dynamic scanning force microscopy, Appl. Surf. Sci. **140** (1999) 293–297

14.131 G. Schwarz, A. Kley, J. Neugebauer, M. Scheffler: Electronic and structural properties of vacancies on and below the GaP(110) surface, Phys. Rev. B **58** (1998) 1392–1499

14.132 H. Hölscher, W. Allers, U. D. Schwarz, A. Schwarz, R. Wiesendanger: Interpretation of 'true atomic resolution' images of graphite (0001) in noncontact atomic force microscopy, Phys. Rev. B **62** (2000) 6967–6970

14.133 H. Hölscher, W. Allers, U. D. Schwarz, A. Schwarz, R. Wiesendanger: Simulation of NC-AFM images of xenon(111), Appl. Phys. A **72** (2001) S35–S38

14.134 H. Hölscher, W. Allers, U. D. Schwarz, A. Schwarz, R. Wiesendanger: Determination of tip-sample interaction potentials by dynamic force spectroscopy, Phys. Rev. Lett. **83** (1999) 4780–4783

14.135 H. Hölscher, U. D. Schwarz, R. Wiesendanger: Calculation of the frequency shift in dynamic force microscopy, Appl. Surf. Sci. **140** (1999) 344–351

14.136 B. Gotsman, B. Anczykowski, C. Seidel, H. Fuchs: Determination of tip-sample interaction forces from measured dynamic force spectroscopy curves, Appl. Surf. Sci. **140** (1999) 314–319

14.137 U. Dürig: Extracting interaction forces and complementary observables in dynamic probe microscopy, Appl. Phys. Lett. **76** (2000) 1203–1205

14.138 M. A. Lantz, H. J. Hug, R. Hoffmann, P. J. A. van Schendel, P. Kappenberger, S. Martin, A. Baratoff, H.-J. Güntherodt: Quantitative measurement of short-range chemical bonding forces, Science **291** (2001) 2580–2583

14.139 S. M. Langkat, H. Hölscher, A. Schwarz, R. Wiesendanger: Determination of site specific forces between an iron coated tip and the NiO(001) surface by force field spectroscopy, Surf. Sci. (2002) in press

14.140 H. Hölscher, S. M. Langkat, A. Schwarz, R. Wiesendanger: Measurement of three-dimensional force fields with atomic resolution using dynamic force spectroscopy, Appl. Phys. Lett. (2002) in press

14.141 B. C. Stipe, H. J. Mamin, T. D. Stowe, T. W. Kenny, D. Rugar: Noncontact friction and force fluctuations between closely spaced bodies, Phys. Rev. Lett. **87** (2001)

14.142 C. Sommerhalter, T. W. Matthes, T. Glatzel, A. Jäger-Waldau, M. C. Lux-Steiner: High-sensitivity quantitative Kelvin probe microscopy by noncontact ultra-high-vacuum atomic force microscopy, Appl. Phys. Lett. **75** (1999) 286–288

14.143 A. Schwarz, W. Allers, U. D. Schwarz, R. Wiesendanger: Dynamic mode scanning force microscopy of *n*-InAs(110)-(1×1) at low temperatures, Phys. Rev. B **62** (2000) 13617–13622

14.144 K. L. McCormick, M. T. Woodside, M. Huang, M. Wu, P. L. McEuen, C. Duruoz, J. S. Harris: Scanned potential microscopy of edge and bulk currents in the quantum Hall regime, Phys. Rev. B **59** (1999) 4656–4657

14.145 P. Weitz, E. Ahlswede, J. Weis, K. v. Klitzing, K. Eberl: Hall-potential investigations under quantum Hall conditions using scanning force microscopy, Physica E **6** (2000) 247–250

14.146 E. Ahlswede, P. Weitz, J. Weis, K. v. Klitzing, K. Eberl: Hall potential profiles in the quantum Hall regime measured by a scanning force microscope, Physica B **298** (2001) 562–566

14.147 M.T. Woodside, C. Vale, P. L. McEuen, C. Kadow, K. D. Maranowski, A. C. Gossard: Imaging interedge-state scattering centers in the quantum Hall regime, Phys. Rev. B **64** (2001) 041310-1–041310-4

14.148 K. Moloni, B. M. Moskowitz, E. D. Dahlberg: Domain structures in single crystal magnetite below the Verwey transition as observed with a low-temperature magnetic force microscope, Geophys. Res. Lett. **23** (1996) 2851–2854

14.149 Q. Lu, C. C. Chen, A. de Lozanne: Observation of magnetic domain behavior in colossal magnetoresistive materials with a magnetic force microscope, Science **276** (1997) 2006–2008

14.150 G. Xiao, J. H. Ross, A. Parasiris, K. D. D. Rathnayaka, D. G. Naugle: Low-temperature MFM studies of CMR manganites, Physica C **341–348** (2000) 769–770

14.151 M. Liebmann, U. Kaiser, A. Schwarz, R. Wiesendanger, U. H. Pi, T. W. Noh, Z. G. Khim, D. W. Kim: Domain nucleation and growth of $La_{07}Ca_{0.3}MnO_{3-\delta}$/$LaAlO_3$ films studied by low temperature MFM, J. Appl. Phys. **93** (2003) 8319–8321

14.152 A. Moser, H. J. Hug, I. Parashikov, B. Stiefel, O. Fritz, H. Thomas, A. Baratoff, H. J. Güntherodt, P. Chaudhari: Observation of single vortices condensed into a vortex-glass phase by magnetic force microscopy, Phys. Rev. Lett. **74** (1995) 1847–1850

14.153 C.W. Yuan, Z. Zheng, A. L. de Lozanne, M. Tortonese, D. A. Rudman, J. N. Eckstein: Vortex images in thin films of $YBa_2Cu_3O_{7-x}$ and $Bi_2Sr_2Ca_1Cu_2O_{8-x}$ obtained by low-temperature magnetic force microscopy, J. Vac. Sci. Technol. B **14** (1996) 1210–1213

14.154 A. Volodin, K. Temst, C. van Haesendonck, Y. Bruynseraede: Observation of the Abrikosov vortex lattice in $NbSe_2$ with magnetic force microscopy, Appl. Phys. Lett. **73** (1998) 1134–1136

14.155 A. Moser, H.J. Hug, B. Stiefel, H.J. Güntherodt: Low temperature magnetic force microscopy on $YBa_2Cu_3O_{7-\delta}$ thin films, J. Magn. Magn. Mater. **190** (1998) 114–123

14.156 A. Volodin, K. Temst, C. van Haesendonck, Y. Bruynseraede: Imaging of vortices in conventional superconductors by magnetic force microscopy images, Physica C **332** (2000) 156–159

14.157 M. Roseman, P. Grütter: Estimating the magnetic penetration depth using constant-height magnetic force microscopy images of vortices, New J. Phys. **3** (2001) 24.1–24.8

14.158 A. Volodin, K. Temst, C. van Haesendonck, Y. Bruynseraede, M. I. Montero, I. K. Schuller: Magnetic force microscopy of vortices in thin niobium films: Correlation between the vortex distribution and the thickness-dependent film morphology, Europhys. Lett. **58** (2002) 582–588

14.159 U. H. Pi, T. W. Noh, Z. G. Khim, U. Kaiser, M. Liebmann, A. Schwarz, R. Wiesendanger: Vortex dynamics in $Bi_2Sr_2CaCu_2O_8$ single crystal with low density columnar defects studied by magnetic force microscopy, J. Low Temp. Phys. **131** (2003) 993–1002

14.160 M. Roseman, P. Grütter, A. Badia, V. Metlushko: Flux lattice imaging of a patterned niobium thin film, J. Appl. Phys. **89** (2001) 6787–6789

14.161 K. Nakamura, H. Hasegawa, T. Oguchi, K. Sueoka, K. Hayakawa, K. Mukasa: First-principles calculation of the exchange interaction and the exchange force between magnetic Fe films, Phys. Rev. B **56** (1997) 3218–3221

14.162 A. S. Foster, A. L. Shluger: Spin-contrast in noncontact AFM on oxide surfaces: Theoretical modeling of NiO(001) surface, Surf. Sci. **490** (2001) 211–219

14.163 H. Hoisoi, M. Kimura, K. Hayakawa, K. Sueoka, K. Mukasa: Non-contact atomic force microscopy of an antiferromagnetic NiO(100) surface using a ferromagnetic tip, Appl. Phys. A **72** (2001) S23–S26

14.164 J. A. Sidles, J. L. Garbini, G. P. Drobny: The theory of oscillator-coupled magnetic resonance with potential applications to molecular imaging, Rev. Sci. Instrum. **63** (1992) 3881–3899

14.165 J. A. Sidles, J. L. Garbini, K. J. Bruland, D. Rugar, O. Züger, S. Hoen, C. S. Yannoni: Magnetic resonance force microscopy, Rev. Mod. Phys. **67** (1995) 249–265

14.166 D. Rugar, C. S. Yannoni, J. A. Sidles: Mechanical detection of magnetic resonance, Nature **360** (1992) 563–566

14.167 K. Wago, D. Botkin, O. Züger, R. Kendrick, C. S. Yannoni, D. Rugar: Force-detected electron spin resonance: Adiabatic inversion, nutation and spin echo, Phys. Rev. B **57** (1998) 1108–1114

14.168 D. Rugar, O. Züger, S. Hoen, C. S. Yannoni, H. M. Vieth, R. D. Kendrick: Force detection of nuclear magnetic resonance, Science **264** (1994) 1560–1563

14.169 Z. Zhang, P. C. Hammel, P. E. Wigen: Observation of ferromagnetic resonance in a microscopic sample using magnetic resonance force microscopy, Appl. Phys. Lett. **68** (1996) 2005–2007

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

15. Dynamic Force Microscopy

15.1 **Motivation: Measurement of a Single Atomic Bond** 450

15.2 **Harmonic Oscillator: A Model System for Dynamic AFM** 454

15.3 **Dynamic AFM Operational Modes** 455
- 15.3.1 Amplitude-Modulation/Tapping-Mode AFMs 456
- 15.3.2 Self-Excitation Modes 461

15.4 ***Q*-Control** 464

15.5 **Dissipation Processes Measured with Dynamic AFM** 468

15.6 **Conclusion** 471

References 471

This chapter presents an introduction to the concept of the dynamic operational mode of the atomic force microscope (dynamic AFM). While the static, or contact mode AFM is a widespread technique to obtain nanometer resolution images on a wide variety of surfaces, true atomic resolution imaging is routinely observed only in the dynamic mode. We will explain the jump-to-contact phenomenon encountered in static AFM and present the dynamic operational mode as a solution to overcome this effect. The dynamic force microscope is modeled as a harmonic oscillator to gain a basic understanding of the underlying physics in this mode.

Dynamic AFM comprises a whole family of operational modes. A systematic overview of the different modes typically encountered in force microscopy is presented, and special care is taken to explain the distinct features of each mode. Two modes of operation dominate the application of dynamic AFM. First, the amplitude modulation mode (also called tapping mode) is shown to exhibit an instability, which separates the purely attractive force interaction regime from the attractive–repulsive regime. Second, the self-excitation mode is derived and its experimental realization is outlined. While the first is primarily used for imaging in air and liquid, the second dominates imaging in UHV (ultrahigh vacuum) for atomic resolution imaging. In particular, we explain the influence of different forces on spectroscopy curves obtained in dynamic force microscopy. A quantitative link between the measurement values and the interaction forces is established.

Force microscopy in air suffers from small quality factors of the force sensor (i. e., the cantilever beam), which are shown to limit the achievable resolution. Also, the above mentioned instability in amplitude modulation mode often hinders imaging of soft and fragile samples. A combination of the amplitude modulation with the self-excitation mode is shown to increase the quality, or *Q*-factor, and extend the regime of stable operation, making the so-called *Q*-control module a valuable tool. Apart from the advantages of dynamic force microscopy as a nondestructive, high-resolution imaging method, it can also be used to obtain information about energy dissipation phenomena at the nanometer scale. This measurement channel can provide crucial information on electric and magnetic surface properties. Even atomic resolution imaging has been obtained in the dissipation mode. Therefore, in the last section, the quantitative relation between the experimental measurement values and the dissipated power is derived.

15.1 Motivation: Measurement of a Single Atomic Bond

The direct measurement of the force interaction between two distinct molecules has been the motivation for scientists for many years now. The fundamental forces responsible for the solid state of matter can be directly investigated, ultimately between defined single molecules. But it has not been until very recently that the chemical forces could be quantitatively measured for a single atomic bond [15.1]. How can we reliably measure forces that may be as small as one billionth of 1 N? How can we identify one single pair of atoms as the source of the force interaction?

The same mechanical principle that is used to measure the gravitational force exerted by your body weight (e.g., with the scale in your bathroom) can be employed to measure the forces between single atoms. A spring with a defined elasticity is compressed by an arbitrary force (e.g., your weight). The compression Δz of the spring (with spring constant k) is a direct measure of the force F exerted, which in the regime of elastic deformation obeys Hooke's law:

$$F = k\Delta z \,. \tag{15.1}$$

The only difference with regard to your bathroom scale is the sensitivity of the measurement. Typically springs with a stiffness of 0.1 N/m to 10 N/m are used, which will be deflected by 0.1 nm to 10 nm upon application of an interatomic force of some nN. Experimentally, a laser deflection technique is used to measure the movement of the spring. The spring is a bendable cantilever microfabricated from silicon wafers. If a sufficiently sharp tip, usually directly attached to the cantilever, is approached toward a surface within some nanometers, we can measure the interacting forces through changes in the deflected laser beam. This is a static measurement; hence it is called *static AFM*. Alternatively, the cantilever can be excited to vibrate at its resonant frequency. Under the influence of tip-sample forces the resonant frequency (and consequently also amplitude and phase) of the cantilever will change and serve as measurement parameters. This is called the *dynamic AFM*. Due to the multitude of possible operational modes, expressions like noncontact mode, intermittent-contact mode, tapping mode, FM-mode, AM-mode, self-excitation, constant-excitation, or constant-amplitude mode AFM are found in the literature, which will be systematically categorized in the following paragraphs.

In fact, the first AFMs were operated in the dynamic mode. In 1986, *Binnig*, *Quate* and *Gerber* presented the concept of the atomic force microscope [15.2]. The deflection of the cantilever with the tip was measured with sub-angstrom precision by an additional scanning tunneling microscope (STM). While the cantilever was externally oscillated close to its resonant frequency, the amplitude and phase of the oscillation were measured. If the tip is approached toward the surface, the oscillation parameters, amplitude and phase, are influenced by the tip-surface interaction, and can, therefore, be used as feedback channels. A certain set-point, for example, the amplitude, is given, and the feedback loop will adjust the tip-sample distance such that the amplitude remains constant. The controller parameter is recorded as a function of the lateral position of the tip with respect to the sample, and the scanned image essentially represents the surface topography.

What then is the difference between the static and the dynamic mode of operation for the AFM? The static deflection AFM directly gives the interaction force between tip and sample using (15.1). In the dynamic mode, we find that the resonant frequency, amplitude, and phase of the oscillation change as a consequence of the interaction forces (and also dissipative processes, as discussed in the last section).

In order to get a basic understanding of the underlying physics, it is instructive to consider a very simplified case. Assume that the vibration amplitude is small compared to the range of force interaction. Since van der Waals forces range over typical distances of 10 nm, the vibration amplitude should be less than 1 nm. Furthermore, we require that the force gradient $\partial F_{\text{ts}}/\partial z$ does not vary significantly over one oscillation cycle.

We can view the AFM setup as a coupling of two springs (see Fig. 15.1). Whereas the cantilever is represented by a spring with spring constant k, the force interaction between tip and surface can be modeled by

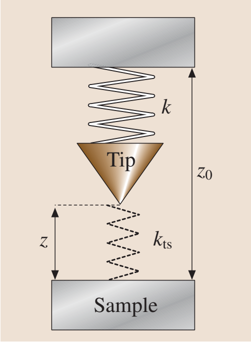

Fig. 15.1 Model of the AFM tip while experiencing tip-sample forces. The tip is attached to a cantilever with spring constant k, and the force interaction is modeled by a spring with a stiffness k_{ts} equal to the force gradient. Note that the force interaction spring is not constant, but depends on the tip-sample distance z

a second spring. The derivative of the force with respect to tip-sample distance is the force gradient and represents the spring constant k_{ts} of the interaction spring. This spring constant k_{ts} is constant only with respect to one oscillation cycle, but varies with the average tip-sample distance as the probe is approached to the sample. The two springs are coupled in series. Therefore, we can write for the total spring constant of the AFM system:

$$k_{total} = k + k_{ts} = k - \frac{\partial F_{ts}}{\partial z} \,. \tag{15.2}$$

From the simple harmonic oscillator (neglecting any damping effects) we find that the resonant frequency ω of the system is shifted by $\Delta\omega$ from the free resonant frequency ω_0 due to the force interaction:

$$\begin{aligned}\omega^2 &= (\omega_0 + \Delta\omega)^2 \\ &= k_{total}/m^* = \left(k + \frac{\partial F_{ts}}{\partial z}\right)/m^* \,.\end{aligned} \tag{15.3}$$

Here m^* represents the effective mass of the cantilever. A detailed analysis of how m^* is related to the geometry and total mass of the cantilever can be found in the literature [15.3]. In the approximation that $\Delta\omega$ is much smaller than ω_0, we can write:

$$\frac{\Delta\omega}{\omega_0} \cong -\frac{1}{2k}\frac{\partial F_{ts}}{\partial z} \,. \tag{15.4}$$

Therefore, we find that the frequency shift of the cantilever resonance is proportional to the force gradient of the tip-sample interaction.

Although the above consideration is based on a very simplified model, it shows qualitatively that in dynamic force microscopy we will find that the oscillation frequency depends on the force gradient, while static force microscopy measures the force itself. In principle, we can calculate the force curve from the force gradient and vice versa (neglecting a constant offset). It seems, therefore, that the two methods are equivalent, and our choice will depend on whether we can measure the beam deflection or the frequency shift with better precision at the cost of technical effort.

However, we have neglected one important issue for the operation of the AFM thus far: the mechanical stability of the measurement. In static AFM, the tip is slowly approached toward the surface. The force between the tip and the surface will always be counteracted by the restoring force of the cantilever. In Fig. 15.2, you can see a typical force-distance curve. Upon approach of the tip toward the sample, the negative attractive forces, representing, e.g., van der Waals or chemical interaction forces, increase until a maximum is reached. This turnaround point is due to the onset of repulsive forces caused by coulomb repulsion, which will start to dominate upon further approach. The spring constant of the cantilever is represented by the slope of the straight line. The position of the z-transducer (typically a piezo element), which moves the probe, is at the intersection of the line with the horizontal axis. The position of the tip, shifted from the probe's base due to the lever bending, can be found at the intersection of the cantilever line with the force curve. Hence, the total force is zero, i. e., the cantilever is in its equilibrium position (note that the spring constant line here shows attractive forces, although in reality the forces are repulsive, i. e., pulling back the tip from the surface). As soon as the position A in Fig. 15.2 is reached, we find two possible intersection points, and upon further approach there are even three force equilibrium points. However, between point A to B the tip is at a local energy minimum and, therefore, will still follow the force curve. But at

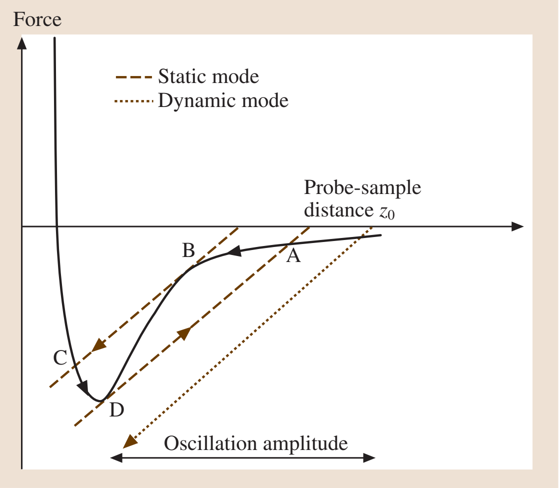

Fig. 15.2 Force-distance curve of a typical tip-sample interaction. In static mode AFM, the tip would follow the force curve until point B is reached. If the slope of the force curve becomes larger than the spring constant of the cantilever (*dashed line*), the tip will suddenly jump to position C. Upon retraction, a different path will be followed along D and A again. In dynamic AFM, the cantilever oscillates with amplitude A. Although the equilibrium position of the oscillation is far from the surface, the tip will experience the maximum attractive force at point D during some parts of the oscillation cycle. However, the total force is always pointing away from the surface, therefore, avoiding an instability

point B, when the adhesion force upon further approach would become larger than the spring restoring force, the tip will suddenly jump to point C. We can then probe the predominantly repulsive force interaction by further reducing the tip-sample distance. Then, while retracting the tip, we will pass point C, because the tip is still in a local energy minimum. At position D the tip will jump suddenly to point A again, since the restoring force now exceeds the adhesion. From Fig. 15.2 we can see that the sudden instability will happen at exactly the point where the slope of the adhesion force exceeds the slope of the spring constant. Therefore, if the negative force gradient of the tip-sample interaction will at any point exceed the spring constant, a mechanical instability occurs. Mathematically speaking, we demand that for a stable measurement:

$$\left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z} > k \quad \text{for all points } z\,. \tag{15.5}$$

The phenomenon of mechanical instability is often referred to as the "jump-to-contact".

Looking at Fig. 15.2, we realize that large parts of the force curve cannot be measured if the jump-to-contact phenomenon occurs. We will not be able to measure the point at which the attractive forces reach their maximum, representing the temporary chemical bonding of the tip and the surface atoms. Secondly, the sudden instability, the jump-to-contact, will often cause the tip to change the very last tip or surface atoms. A smooth, careful approach needed to measure the full force curve does not seem feasible. Our goal of measuring the chemical interaction forces of two single molecules may become impossible.

There are several solutions to the jump-to-contact problem: On the one hand, we can simply choose a sufficiently stiff spring, so that (15.5) is fulfilled at all points of the force curve. On the other hand, we can resort to a trick to enhance the counteracting force of the cantilever: We can oscillate the cantilever with large amplitude, thereby making it virtually stiffer at the point of strong force interaction.

Consider the first solution, which seems simpler at first glance. Chemical bonding forces extend over a distance range of about 0.1 nm. Typical binding energies of a couple of eV will lead to adhesion forces on the order of some nN. Force gradients will, therefore, reach values of some 10 N/m. A spring for stable force measurements will have to be as stiff as 100 N/m to ensure that no instability occurs (a safety factor of ten seems to be a minimum requirement, since usually one cannot be sure a priori that only one atom will dominate the interaction). In order to measure the nN interaction force, a static cantilever deflection of 0.01 nm has to be detected. With standard beam deflection AFM setups this becomes a challenging task.

This problem was solved [15.4, 5] using an in situ optical interferometer measuring the beam deflection at liquid nitrogen temperature in a UHV environment. In order to ensure that the force gradients are smaller than the lever spring constant (50 N/m), the tips were fabricated to terminate in only three atoms, therefore, minimizing the total force interaction. The only tool known to be able to engineer SPM tips down to atomic dimensions is the field ion microscope (FIM). This technique not only allows imaging of the tip apex with atomic precision, but also can be used to manipulate the tip atoms by field evaporation [15.6], as shown in Fig. 15.3. Atomic interaction forces were measured with sub-nanonewton precision, revealing force curves of only a few atoms interacting without mechanical hysteresis. However, the technical effort to achieve this type of measurement is considerable, and most researchers today have resorted to the second solution.

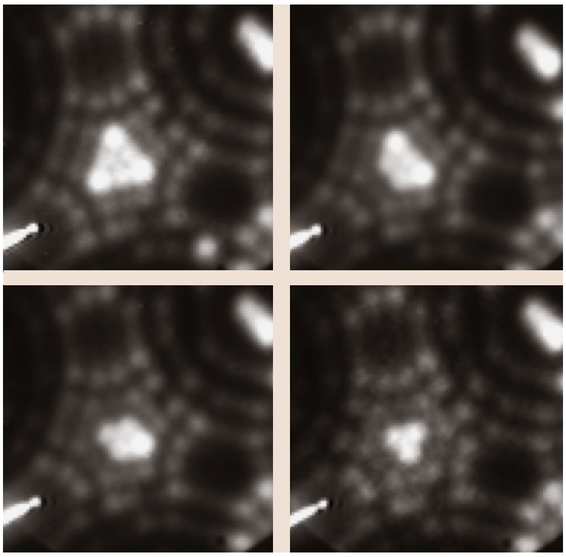

Fig. 15.3 Manipulation of the apex atoms of an AFM tip using field ion microscopy (FIM). Images are obtained at a tip bias of 4.5 kV. The last six atoms of the tip can be inspected in this example. Field evaporation to remove single atoms is performed by increasing the bias voltage for a short time to 5.2 kV. Each of the outer three atoms can be consecutively removed, eventually leaving a trimer tip apex

The alternative solution can be visualized in Fig. 15.2. The straight, dashed line now represents the force values of the oscillating cantilever, with amplitude *A* assuming Hooke's law is valid. This is the tensile force of the cantilever spring pulling the tip away from the sample. The restoring force of the cantilever is at all points stronger than the adhesion force. For example, the total force at point D is still pointing away from the sample, although the spring has the same stiffness as before. Mathematically speaking, the measurement is stable as long as the cantilever spring force $F_{\text{cb}} = kA$ is larger than the attractive tip-sample force F_{ts} [15.7]. In the static mode we would already experience an instability at that point. However, in the dynamic mode, the spring is pre-loaded with a force stronger than the attractive tip-sample force. The equilibrium point of the oscillation is still far away from the point of closest contact of the tip and surface atoms. The total force curve can be probed by varying the equilibrium point of the oscillation, i. e., by adjusting the *z*-piezo.

The diagram also shows that the oscillation amplitude has to be quite large if fairly soft cantilevers are to be used. With lever spring constants of 10 N/m, the amplitude must be at least 1 nm to ensure forces of 1 nN can be reliably measured. In practical applications, amplitudes of 10–100 nm are used to stay on the safe side. This means that the oscillation amplitude is much larger than the force interaction range. The above simplification, that the force gradient remains constant within one oscillation cycle, does not hold anymore. Measurement stability is gained at the cost of a simple quantitative analysis of the experiments. In fact, dynamic AFM was first used to obtain atomic resolution images of clean surfaces [15.8], and it took another six years [15.1] before quantitative measurements of single bond forces were obtained.

The technical realization of dynamic mode AFMs is based on the same key components as a DC-AFM setup. The most common principle is the method of laser deflection sensing (see, e.g., Fig. 15.4). A laser beam is focused on the back side of a microfabricated cantilever. The reflected laser spot is detected with a positional sensitive diode (PSD). This photodiode is sectioned into two parts that are read-out separately (usually even a four-quadrant diode is used to detect torsional movements of the cantilever for lateral friction measurements). With the cantilever at equilibrium, the spot is adjusted such that the two sections show the same intensity. If the cantilever bends up or down, the spot moves, and the difference signal between the upper and lower sections is a measure of the bending.

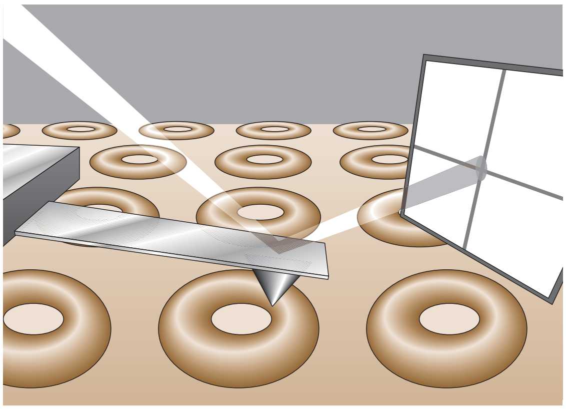

Fig. 15.4 Representation of an AFM setup with the laser beam deflection method. Cantilever and tip are microfabricated from silicon wafers. A laser beam is deflected from the back side of the cantilever and again focused on a photosensitive diode. The diode is segmented into four quadrants, which allows the measurement of vertical and torsional bending of the cantilever (artwork from *Jörg Heimel* rendered with POV-Ray 3.0)

In order to enhance sensitivity, several groups have adopted an interferometer system to measure the cantilever deflection. A thorough comparison of different measurement methods with analysis of sensitivity and noise level is given by [15.3].

The cantilever is mounted on a device that allows oscillation of the beam. Often a piezo-element serves this purpose. The reflected laser beam is analyzed for oscillation amplitude, frequency, and phase difference between excitation and vibration. Depending on the mode of operation, a feedback mechanism will adjust oscillation parameters and/or tip-sample distance during the scanning. The setup can be operated in air, UHV, and even in fluids. This allows measurement of a wide range of surface properties from atomic resolution imaging [15.8] up to studying biological processes in liquid [15.9, 10].

15.2 Harmonic Oscillator: A Model System for Dynamic AFM

The oscillating cantilever has three degrees of freedom: the amplitude, the frequency, and the phase difference between excitation and oscillation. Let us consider the damped driven harmonic oscillator. The cantilever is mounted on a piezoelectric element that is oscillating with amplitude A_d at frequency ω:

$$z_\mathrm{d}(t) = A_\mathrm{d}\cos(\omega t)\ . \qquad (15.6)$$

We assume that the cantilever spring obeys Hooke's law. Secondly, we introduce a friction force that is proportional to the speed of the cantilever motion, whereas α denotes the damping coefficient (Amontons's law). With Newton's first law we find for the oscillating system the following equation of motion for the position $z(t)$ of the cantilever tip (see also Fig. 15.1):

$$m\ddot{z}(t) = -\alpha\dot{z}(t) - kz(t) - kz_\mathrm{d}(t)\ . \qquad (15.7)$$

We define $\omega_0^2 = k/m^*$, which turns out to be the resonant frequency of the free (undamped, i. e., $\alpha = 0$) oscillating beam. We further define the dimensionless quality factor $Q = m^*\omega_0/\alpha$, antiproportional to the damping coefficient. The quality factor describes the number of oscillation cycles, after which the damped oscillation amplitude decays to $1/\mathrm{e}$ of the initial amplitude with no external excitation ($A_\mathrm{d} = 0$). After some basic math, this results in the following differential equation:

$$\ddot{z}(t) + \frac{\omega_0}{Q}\dot{z}(t) + \omega_0^2 z(t) = A_\mathrm{d}\omega_0^2\cos(\omega t)\ . \qquad (15.8)$$

The solution is a linear combination of two regimes [15.11]. Starting from rest and switching on the piezo-excitation at $t = 0$, the amplitude will increase from zero to the final magnitude and reach a steady state, where amplitude, phase, and frequency of the oscillation stay constant over time. The steady-state solution $z_1(t)$ is reached after $2Q$ oscillation cycles and follows the external excitation with amplitude A_0 and phase difference φ:

$$z_1(t) = A_0\cos(\omega t + \varphi)\ . \qquad (15.9)$$

The oscillation amplitude in the transient regime during the first $2Q$ cycles follows:

$$z_2(t) = A_\mathrm{t}\,\mathrm{e}^{-\omega_0 t/2Q}\sin(\omega_0 t + \varphi_\mathrm{t})\ . \qquad (15.10)$$

We emphasize the important fact that the exponential term causes $z_2(t)$ to diminish exponentially with time constant τ:

$$\tau = 2Q/\omega_0\ . \qquad (15.11)$$

In vacuum conditions, only the internal dissipation due to bending of the cantilever is present, and Q reaches values of 10,000 at typical resonant frequencies of 100,000 Hz. This results in a relatively long transient regime of $\tau \cong 30$ ms, which limits the possible operational modes for dynamic AFM (detailed analysis by [15.11]). Changes in the measured amplitude, which reflect a change of atomic forces, will have a time lag of 30 ms, which is very slow considering one wants to scan a 200 × 200 point image within a few minutes. In air, however, viscous damping due to air friction dominates and Q goes down to less than 1,000, resulting in a time constant below the millisecond level. This response time is fast enough to use the amplitude as a measurement parameter.

If we evaluate the steady state solution $z_1(t)$ in the differential equation, we find the following well-known

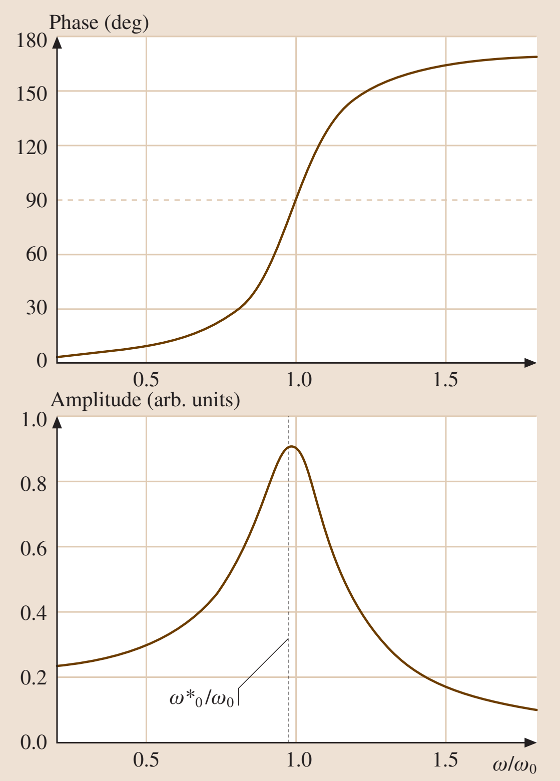

Fig. 15.5 Amplitude and phase versus excitation frequency curves for the damped harmonic oscillator with a quality factor of $Q = 4$

solution for amplitude and phase of the oscillation as a function of the excitation frequency ω:

$$A_0 = \frac{A_\mathrm{d} Q \omega_0^2}{\sqrt{\omega^2 \omega_0^2 + Q^2 \left(\omega_0^2 - \omega^2\right)^2}} \tag{15.12}$$

$$\varphi = \arctan\left(\frac{\omega\omega_0}{Q\left(\omega_0^2 - \omega^2\right)}\right) . \tag{15.13}$$

Amplitude and phase diagrams are depicted in Fig. 15.5. As can be seen from (15.12), the amplitude will reach its maximum at a frequency different from ω_0, if Q has a finite value. The damping term of the harmonic oscillator causes the resonant frequency to shift from ω_0 to ω_0^*:

$$\omega_0^* = \omega_0 \sqrt{1 - \frac{1}{2Q^2}} . \tag{15.14}$$

The shift is negligible for Q-factors of 100 and above, which is the case for most applications in vacuum or air. However, for measurements in liquids, Q can be smaller than 10 and ω_0 differs significantly from ω_0^*. As we will discuss later, it is also possible to enhance Q by using a special excitation method.

In the case that the excitation frequency is equal to the resonant frequency of the undamped cantilever, $\omega = \omega_0$, we find the useful relation:

$$A_0 = Q A_\mathrm{d} \quad \text{for } \omega = \omega_0 . \tag{15.15}$$

Since $\omega_0^* \approx \omega_0$ for most cases, we find that (15.15) holds true for exciting the cantilever at its resonance. From a similar argument, the phase becomes approximately 90 deg for the resonance case. We also see that in order to reach vibration amplitudes of some 10 nm, the excitation only has to be as small as 1 pm for typical cantilevers operated in vacuum.

So far, we have not considered an additional force term, describing the interaction between the probing tip and the sample. For typical, large vibration amplitudes of 10–100 nm no general solution for this analytical problem has been found yet. The cantilever tip will experience a whole range of force interactions during one single oscillation cycle, rather than one defined tip-sample force. Only in the special case of a self-excited cantilever oscillation has the problem been solved semi-analytically, as we will see later.

15.3 Dynamic AFM Operational Modes

While the quantitative interpretation of force curves in contact AFM is straightforward using (15.1), we explained in the previous paragraphs that its application to assess short-range attractive interatomic forces is rather limited. The dynamic mode of operation seems to open a viable direction toward achieving this task. However interpretation of the measurements generally appears to be more difficult. Different operational modes are employed in dynamic AFM, and the following paragraphs are intended to distinguish these modes and categorize them in a systematic way.

The oscillation trajectory of a dynamically driven cantilever is determined by three parameters: the amplitude, the phase, and the frequency. Tip-sample interactions can influence all three parameters, in the following, termed the internal parameters. The oscillation is driven externally, with excitation amplitude A_d and excitation frequency ω. These variables will be referred to as the external parameters. The external parameters are set by the experimentalist, whereas the internal parameters are measured and contain the crucial information about the force interaction. In scanning probe applications, it is common to control the probe-surface distance z_0 in order to keep an internal parameter constant (i. e., the tunneling current in STM or the beam deflection in contact AFM), which represents a certain tip-sample interaction. In z-spectroscopy mode, the distance is varied in a certain range, and the change of the internal parameters are measured as a fingerprint of the tip-sample interactions.

In dynamic AFM the situation is rather complex. Any of the internal parameters can be used for feedback of the tip-sample distance z_0. However, we already realized that, in general, the tip-sample forces could only be fully assessed by measuring all three parameters. This makes it difficult to obtain images where the distance z_0 is representative of the surface at one force set-point. A solution to this problem is to establish additional feedback loops, which keep the internal parameters constant by adjusting the external variables.

In the simplest setup, the excitation frequency is set to a predefined value, and the excitation amplitude remains constant by a feedback loop. This is called the AM mode (amplitude modulation), or tapping mode (TM). As stated before, in principle, any of the internal parameters can be used for feedback to the tip-sample distance – in AM mode the amplitude signal is used. A certain amplitude (smaller than the free oscillation amplitude) at a frequency close to the resonance of the cantilever is chosen, the tip is approached toward the surface under investigation, and the approach is stopped as soon as the set-point amplitude is reached. The oscillation-phase is usually recorded during the scan, however, the shift of the resonant frequency of the cantilever cannot be directly accessed, since this degree of freedom is blocked by the external excitation at a fixed frequency. It turns out that this mode is simple to operate from a technical perspective, but quantitative information about the tip-sample interaction forces has so far not been reliably extracted from AM-mode AFM. Still, it is one of the most commonly used modes in dynamic AFM operated in air, and even in liquid. The strength of this mode is the qualitative imaging of a large variety of surfaces.

It is interesting to discuss the AM mode in the extreme situation that the external excitation frequency is much lower than the resonant frequency [15.12,13]. This results in a quasi-static measurement, although a dynamic oscillation force is applied, and, therefore, this mode can be viewed as a hybrid between static and dynamic AFM. Unfortunately, it has the drawbacks of the static mode, namely, that stiff spring constants must be used and, therefore, the sensitivity of the deflection measurement must be very good, typically employing a high resolution interferometer. Still, it has the advantage of the static measurement in quantitative interpretation, since in the regime of small amplitudes (< 0.1 nm) a direct interpretation of the experiments is possible. In particular, the force gradient at tip-sample distance z_0 is given by the change of the amplitude A and the phase angle φ:

$$\left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z_0} = k\left(1-\frac{A_0}{A}\cos\varphi\right) . \tag{15.16}$$

In effect, the modulated AFM, in contrast to the purely static AFM, can enhance sensitivity due to the use of lock-in techniques, which allows the measurement of the amplitude and phase of the oscillation signal with high precision.

As stated before, the internal parameters can be fed back to the external excitation variables. One of the most useful applications in this direction is the self-excitation system. Here the resonant frequency of the cantilever is detected and selected again as the excitation frequency. In a typical setup, this is done with a phase shift of 90 deg by feeding back the detector signal to the excitation piezo, i. e., the cantilever is always excited in resonance. Influences of the tip-sample interaction forces on the resonant frequency do not change the two other parameters of the oscillation, i. e., amplitude and phase; only the oscillation frequency is shifted. Therefore, it is sufficient to measure the frequency shift between the free oscillation and the oscillation with tip-sample interaction. Since the phase remains at a fixed value, the oscillating system is much better defined than before, and the degrees of freedom for the oscillation are reduced. To even reduce the last degree of freedom, the oscillation amplitude, another feedback loop can be established to keep the oscillation amplitude A constant by varying the excitation amplitude A_{d}. Now, all internal parameters have a fixed relation to the external excitation variables, the system is well-defined, and all parameters can be assessed during the measurement. As it turns out, this mode is the only dynamic mode in which a quantitative relation between tip-sample forces and the change of the resonant frequency can be established.

In the following section we want to discuss the two most popular operational modes, tapping mode and self-excitation mode, in more detail.

15.3.1 Amplitude-Modulation/Tapping-Mode AFMs

In tapping mode, or AM-AFM, the cantilever is excited externally at a constant frequency close to its resonance. Oscillation amplitude and phase during approach of tip and sample serve as the experimental observation channels. Figure 15.6 shows a diagram of a typical tapping-mode AFM setup. The oscillation amplitude and the phase (not shown in diagram) detected with the photodiode are analyzed with a lock-in amplifier. The amplitude is compared to the set-point, and the difference or error signal is used to adjust the z-piezo, i. e., the probe-sample distance. The external modulation unit supplies the signal for the excitation piezo, and, at the same time, the oscillation signal serves as the reference for the lock-in amplifier.

During one oscillation cycle with amplitudes of 10–100 nm, the tip-sample interaction will range over a wide distribution of forces, including attractive, as well as repulsive forces. We will, therefore, measure a convolution of the force-distance curve with the oscillation trajectory. This complicates the interpretation of AM-AFM measurements appreciably.

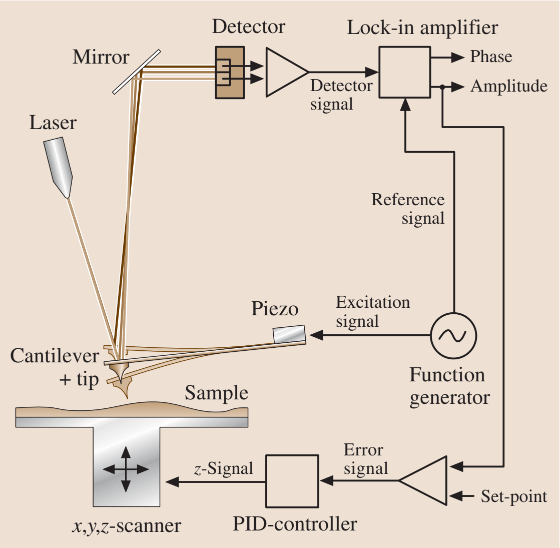

Fig. 15.6 Setup of a dynamic force microscope operated in the AM or tapping mode. A laser beam is deflected by the back side of the cantilever, and the deflection is detected by a split photodiode. The excitation frequency is chosen externally with a modulation unit, which drives the excitation piezo. A lock-in amplifier analyzes phase and amplitude of the cantilever oscillation. The amplitude is used as the feedback signal for the probe-sample distance control

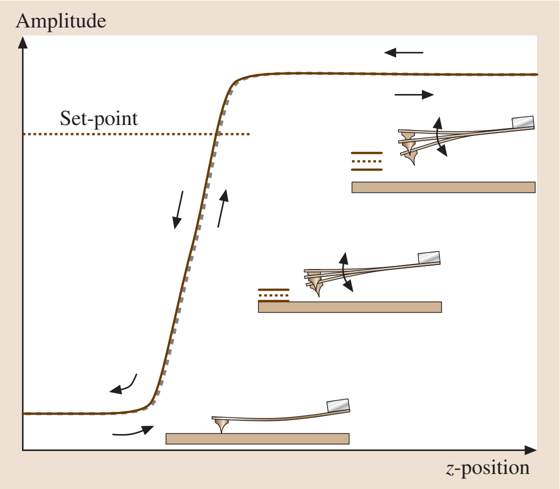

Fig. 15.7 Simplified model showing the oscillation amplitude in tapping-mode AFM for various probe-sample distances

At the same time, the resonant frequency of the cantilever will change due to the appearing force gradients, as could already be seen in the simplified model from (15.4). If the cantilever is excited exactly at its resonant frequency before, it will be excited off-resonance after interaction forces are encountered. This, in turn, changes amplitude and phase (15.12, 15.13), which serve as the measurement signals. Consequently, a different amplitude will cause a change in the encountered effective force. We can see already from this simple *gedanken*-experiment that the interpretation of force curves will be highly complicated. In fact, there is no quantitative theory for AM-AFM available, which allows the experimentalist to unambiguously convert the experimental data to a force-distance relationship.

The qualitative behavior for amplitude versus z_0-position curves is depicted in Fig. 15.7. At large distances, where the forces between tip and sample are negligible, the cantilever oscillates with its free oscillation amplitude. Upon approach of the probe toward the surface the interaction forces cause the amplitude to change, resulting typically in an amplitude getting smaller with continuously decreased tip-sample distance. This is expected, since the force-distance curve will eventually reach the repulsive part and the tip is hindered from indenting further into the sample, resulting in smaller oscillation amplitudes.

However, in order to gain some qualitative insight into the complex relationship between forces and oscillation parameters, we resort to numerical simulations. *Anczykowski* et al. [15.14, 15] have calculated the oscillation trajectory of the cantilever under the influence of a given force model. Van der Waals interactions were considered the only effective, attractive forces, and the total interaction resembled a Lennard-Jones-type potential. Mechanical relaxations of the tip and sample surface were treated in the limits of continuum theory with the numerical MYD/BHW [15.16, 17] approach, which allows the simulations to be compared to corresponding experiments. Figure 15.8 shows the force-distance curves for different tip radii underlying the dynamic AFM simulations.

The cantilever trajectory was analyzed by solving the differential equation (15.7) extended by the force-distance relations from Fig. 15.8 using the numerical Verlet algorithm [15.18, 19]. The results of the simulation for the amplitude and phase of the tip oscillation as a function of z-position of the probe are presented in Fig. 15.9. One has to keep in mind that the z-position

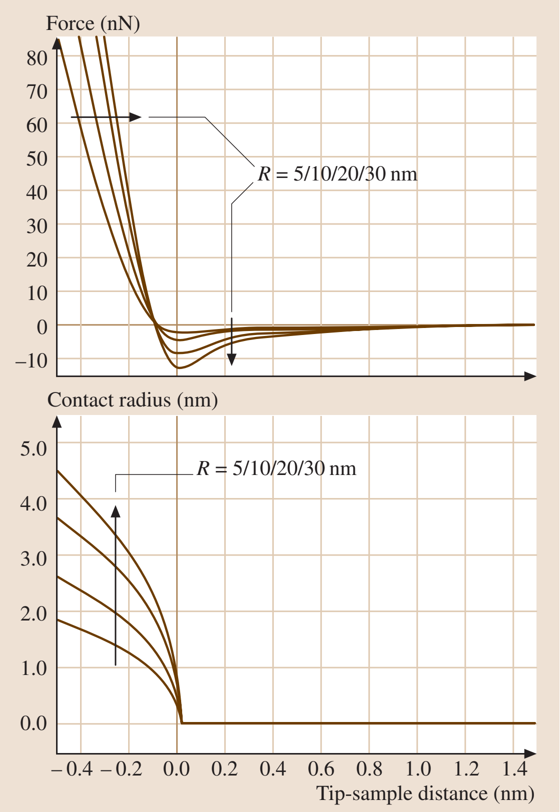

Fig. 15.8 Force curves and corresponding contact radius calculated with the MYD/BHW-model as a function of tip radius for a Si–Si contact. These force curves are used for the tapping-mode AFM simulations

of the probe is not equivalent to the real tip-sample distance at equilibrium position, since the cantilever might bend statically due to the interaction forces. The behavior of the cantilever can be subdivided into three different regimes. We distinguish the cases in which the beam is oscillated below its resonant frequency ω_0, exactly at ω_0, and above ω_0. In the following, we will refer to ω_0 as the resonant frequency, although the correct resonant frequency is ω_0^* if taking into account the finite Q-value.

Clearly, Fig. 15.9 exhibits more features than were anticipated from the initial, simple arguments. Amplitude and phase seem to change rather abruptly at certain points when the z_0-position is decreased. Besides, the amplitude or phase-distance curves don't resemble the force-distance curves from Fig. 15.8 in a simple, direct manner. Additionally, we find a hysteresis between approach and retraction.

As an example, let us start by discussing the discontinuous features in the AFM spectroscopy curves of the first case, where the excitation frequency is smaller than ω_0. Consider the oscillation amplitude as a function of excitation frequency in Fig. 15.5 in conjunction with a typical force curve, as depicted in Fig. 15.8. Upon approach of probe and sample, attractive forces will lower the effective resonant frequency of the oscillator. Therefore, the excitation frequency will now be closer to the resonant frequency, causing the vibration amplitude to increase. This, in turn, reduces the tip-sample distance, which again gives rise to a stronger attractive force. The system becomes unstable until the point $z_0 = d_{\text{app}}$ is reached, where repulsive forces stop the self-enhancing instability. This can be clearly observed in Fig. 15.9. Large parts of the force-distance curve cannot be measured due to this instability.

In the second case, where the excitation equals the free resonant frequency, only a small discontinuity is observed upon reduction of the z-position. Here, a shift of the resonant frequency toward smaller values, induced by the attractive force interaction, will reduce the oscillation amplitude. The distance between tip and sample is, therefore, reduced as well, and the self-amplifying effect with the sudden instability does not occur as long as repulsive forces are not encountered. However, at closer tip-sample distances, repulsive forces will cause the resonant frequency to shift again toward higher values, increasing the amplitude with decreasing tip-sample distance. Therefore, a self-enhancing instability will also occur in this case, but at the crossover from purely attractive forces to the regime where repulsive forces occur. Correspondingly, a small kink in the amplitude curve can be observed in Fig. 15.9. An even clearer indication of this effect is manifested by the sudden change in the phase signal at d_{app}.

In the last case, with $\omega > \omega_0$, the effect of amplitude reduction due to the resonant frequency shift is even larger. Again, we find no instability in the amplitude signal during approach in the attractive force regime. However, as soon as the repulsive force regime is reached, the instability occurs due to the induced positive frequency shift. Consequently, a large jump in the phase curve from values smaller than 90 deg to values larger than 90 deg is observed. The small change in the amplitude curve is not resolved in the simulated curves in Fig. 15.9, however, it can be clearly seen in the experimental curves in Fig. 15.10.

Figure 15.10 depicts the corresponding experimental amplitude and phase curves. The measurements were performed in air with a Si cantilever approaching

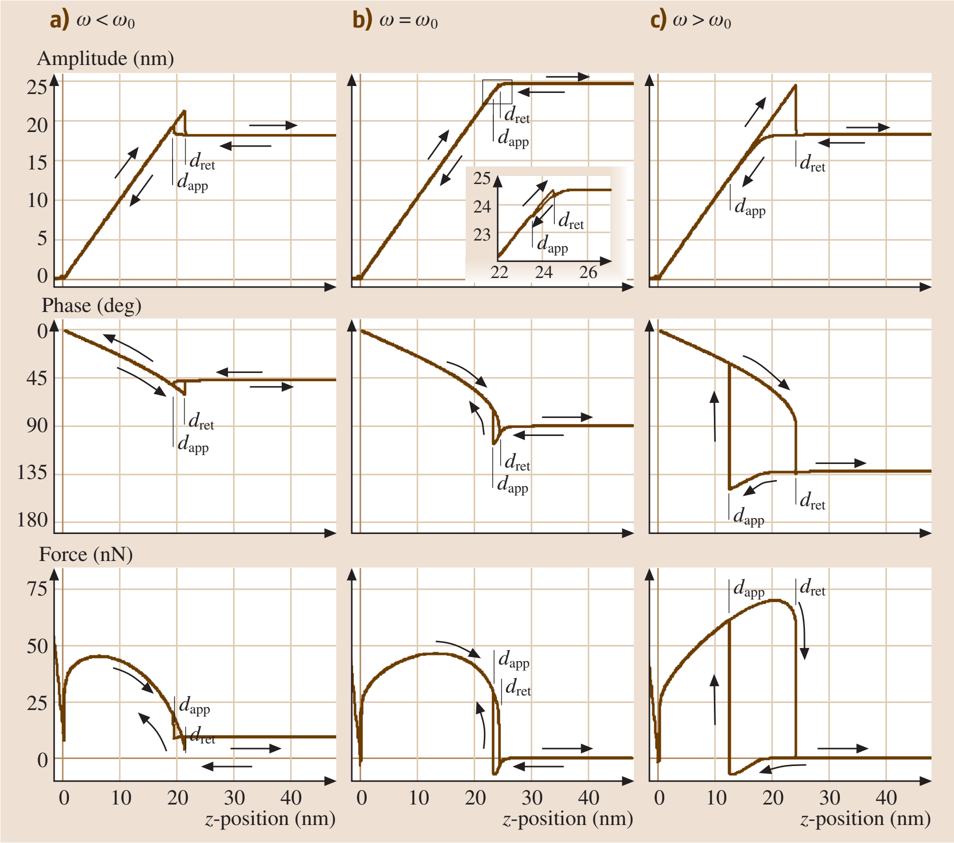

Fig. 15.9a–c Amplitude and phase diagrams with excitation frequency **(a)** below, **(b)** exactly at, and **(c)** above the resonant frequency for tapping-mode AFM from the numerical simulations. Additionally, the bottom diagrams show the interaction forces at the point of closest tip–sample distance, i. e., the lower turnaround point of the oscillation

a Si wafer, with a cantilever resonant frequency of 299.95 kHz. Qualitatively, all prominent features of the simulated curves can also be found in the experimental data sets. Hence, the above model seems to capture the important factors necessary for an appropriate description of the experimental situation.

But what is the reason for this unexpected behavior? We have to turn to the numerical simulations again, where we have access to all physical parameters, in order to understand the underlying processes. The lower part of Fig. 15.9 also shows the interaction force between the tip and the sample at the point of closest approach, i. e., the sample-sided turnaround point of the oscillation. We see that exactly at the points of the discontinuities the total interaction force changes from the net-attractive regime to the attractive-repulsive regime, also termed the intermittent contact regime. The term net-attractive is used to emphasize that the total force is attractive, despite the fact that some minor contributions might still originate from repulsive forces. As soon as a minimum distance is reached, the tip also starts to experience repulsive forces, which completely changes the oscillation behavior. In other words, the dynamic system switches between two oscillatory states.

Directly related to this fact is the second phenomenon: the hysteresis effect. We find separate curves for the approach of the probe toward the surface and the retraction. This seems to be somewhat counterintuitive, since the tip is constantly approaching and retracting from the surface, and the average values of amplitude and phase should be independent of the direction of the average tip-sample distance movement. A hysteresis between approach and retraction within one oscillation due to dissipative processes should directly influence amplitude and phase. However, no dissipation models were included in the simulation. In this case, the hysteresis in Fig. 15.9 is due to the fact that the oscillation jumps into different modes, the system exhibits bistability. This effect is often observed

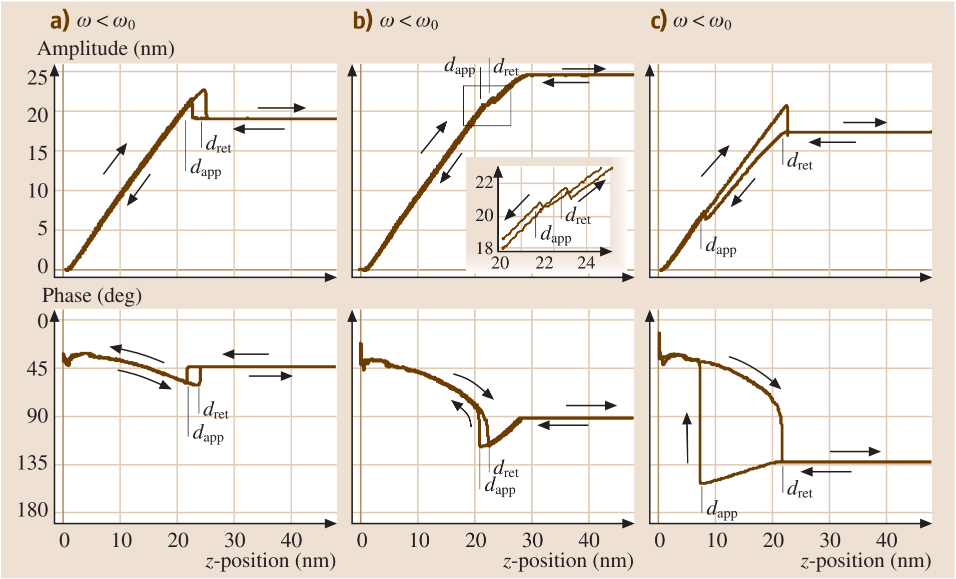

Fig. 15.10a–c Amplitude and phase diagrams with excitation frequency **(a)** below, **(b)** exactly at, and **(c)** above the resonant frequency for tapping-mode AFM from experiments with a Si cantilever on a Si wafer in air

in oscillators under the influence of nonlinear forces [15.20].

For the interpretation of these effects it is helpful to look at Fig. 15.11, which shows the behavior of the simulated tip vibration and force during one oscillation cycle over time. The data is recorded at the same z-position at a point where hysteresis is observed, while a) was taken during the approach and b) during the retraction. Excitation is in resonance, representing case 2, where the amplitude shows a small hysteresis. Also note that the amplitude is almost exactly the same in a) and b). We see that the oscillation at the same z-position exhibits two different modes: While in a) the experienced force is net-attractive, in b) the tip is exposed to attractive and repulsive interactions. Experimental and simulated data show that the change between the net-attractive and intermittent contact mode takes place at different z-positions (d_{app} and d_{ret}) for approach and retraction. Between d_{app} and d_{ret} the system is in a bistable mode. Depending on the history of the measurement, e.g., whether the position d_{app} during the approach (or d_{ret} during retraction) has been reached, the system flips to the other oscillation mode. While the amplitude might not be influenced strongly (case 2), the phase is a clear indicator of the mode switch. On the other hand, if point d_{app} is never reached during the approach, the system will stay in the net-attractive regime and no hysteresis is observed, i. e., the system remains stable.

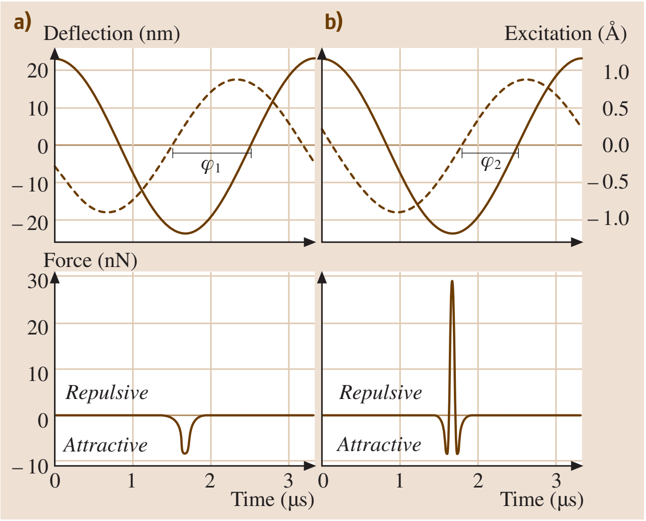

Fig. 15.11a,b Simulation of the tapping-mode cantilever oscillation in the **(a)** net-attractive and **(b)** the intermittent contact regime. The *dashed line* represents the excitation amplitude and the *solid line* is the oscillation amplitude

In conclusion, we find that although a qualitative interpretation of the interaction forces is possible, the AM-nor tapping-mode AFM is not suitable to gain direct quantitative knowledge of tip-sample force interactions. However, it is a very useful tool for imaging nanometer-sized structures in a wide variety of setups, in air or even in liquid. We find that two distinct modes exist for the externally excited oscillation – the net-attractive and the intermittent contact mode – which describe what kind of forces the tip-sample interaction is governed by. The phase can be used as an indicator of which mode the system is running in.

In particular, it can be easily seen that if the free resonant frequency of the cantilever is higher than the excitation frequency, the system cannot stay in the net-attractive regime due to a self-enhancing instability. Since in many applications involving soft and delicate biological samples strong repulsive forces should be avoided, the tapping-mode AFM should be operated at frequencies equal to or above the free resonant frequency [15.21]. Even then, statistical changes of tip-sample forces during the scan might induce a sudden jump into the intermittent contact mode, and the previously explained hysteresis will tend to keep the system in this mode. It is, therefore, of high importance to tune the oscillation parameters in such a way that the AFM stays in the net-attractive regime [15.22]. A concept that achieves this task is the Q-control system, which will be discussed in some detail in the forthcoming paragraphs.

A last word concerning the overlap of simulation and experimental data: While the qualitative agreement down to the detailed shape of hysteresis and instabilities is rather striking, we still find some quantitative discrepancies between the positions of the instabilities d_{app} and d_{ret}. This is probably due to the simplified force model, which only takes into account van der Waals and repulsive forces. Especially at ambient conditions, an omnipresent water meniscus between tip and sample will give rise to much stronger attractive and also dissipative forces than considered in the model. A very interesting feature is that the simulated phase curves in the intermittent contact regime tend to have a steeper slope in the simulation than in the experiments (see also Fig. 15.12). We will later show that this effect is a fingerprint of an effect that had not been included in the above simulation at all: dissipative processes during the oscillation, giving rise to an additional loss of oscillation energy.

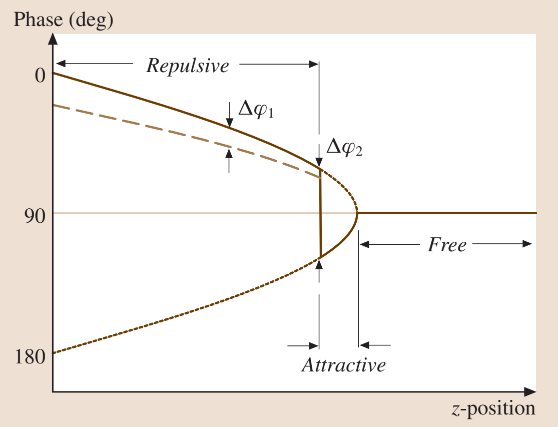

Fig. 15.12 Phase shift in tapping mode as a function of tip-sample distance

15.3.2 Self-Excitation Modes

Despite the wide range of technical applications of the AM or tapping mode of dynamic AFM, it has been found unsuitable for measurements in an environment extremely useful for scientific research: vacuum or ultrahigh vacuum (UHV) with pressures reaching 1×10^{-10} mbar. The STM has already shown how much insight can be gained from some highly defined experiments under those conditions. Consider (15.11) from the above section. The time constant τ for the amplitude to adjust to a different tip-sample force scales with $1/Q$. In vacuum applications, Q of the cantilever is on the order of 10,000, which means that τ is in the range of some 10 ms. This is clearly too long for a scan of at least (100×100) data points. The temperature-induced drift of the sample will render useful interpretation of images an impossible task. However, the resonant frequency of the system will react instantaneously to tip-sample forces. This has led *Albrecht* et al. [15.11] to use a modified excitation scheme.

The system is always oscillated at its resonant frequency. This is achieved by feeding back the oscillation signal from the cantilever into the excitation piezoelement. Figure 15.13 pictures the method in a block diagram. The signal from the PSD is phase shifted by 90 deg (and, therefore, always exciting in resonance) and used as the excitation signal of the cantilever. An additional feedback loop adjusts the excitation amplitude in such a way that the oscillation amplitude remains constant. This ensures that the tip-sample distance is not influenced by changes in the oscillation amplitude. The only degree of freedom of the oscillation system that can react to the tip-sample forces is the change of the resonant frequency. This shift of the frequency is detected and used as the set-point signal for surface scans. Therefore, this mode is also called the FM (frequency modulated) mode.

Let us take a look at the sensitivity of the dynamic AFM. If electronic noise, laser noise, and thermal drift

© Springer 2004

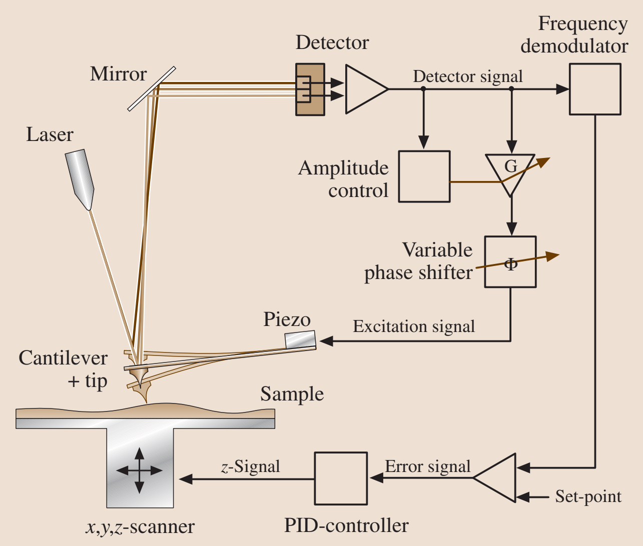

Fig. 15.13 Dynamic AFM operated in the self-excitation mode, where the oscillation signal is directly feedbacked to the excitation piezo. The detector signal is amplified with the variable gain G and phase shifted by phase ϕ. The frequency demodulator detects the frequency shift due to tip-sample interactions, which serves as the control signal for the probe-sample distance

can be neglected, the main noise contribution will come from thermal excitation of the cantilever. A detailed analysis of a dynamic system yields for the minimum detectable force gradient the following relation [15.11]:

$$\left.\frac{\partial F}{\partial z}\right|_{\mathrm{MIN}} = \sqrt{\frac{4k_{\mathrm{B}}TB \cdot k}{\omega_0 Q \langle z_{\mathrm{osc}}^2 \rangle}} \,. \tag{15.17}$$

Here B is the bandwidth of the measurement, T the temperature, and $\langle z_{\mathrm{osc}}^2 \rangle$ is the mean-square amplitude of the oscillation. Please note that this sensitivity limit was deliberately calculated for the FM mode. A similar analysis of the AM mode, however, yields virtually the same result [15.23]. We find that the minimum detectable force gradient, i. e., the measurement sensitivity, is inversely proportional to the square root of the Q-factor of the cantilever. This means that it should be possible to achieve very high-resolution imaging in vacuum conditions. In contrast, AM- or tapping-mode AFM cannot be usefully pursued with large Q. Only the FM mode makes it possible to take practical advantage of (15.17).

A breakthrough in high-resolution AFM application was the atomic resolution imaging of the Si(111)-(7×7) surface reconstruction by Giessibl [15.8] under UHV conditions. Today, atomic resolution imaging has become a standard feature guaranteed by industrially produced dynamic AFM systems. While STM has already proven to be an indispensable tool to gain detailed insight into surface structures of conductors with atomic resolution, the dynamic AFM has opened up the avenue into investigating nonconductive surfaces with equal precision (for example, on aluminium oxide by *Barth* et al. [15.24]).

However, we are concerned with measuring atomic force potentials of a single pair of molecules. Clearly, FM-mode AFM will allow us to identify single atoms, and with sufficient care we will be able to ensure that only one atom from the tip contributes to the total force interaction. Can we, therefore, fill in the last bit of information and find a quantitative relation between the oscillation parameters and the force?

Gotsmann et al. [15.25] investigated this relation by employing a numerical approach. During each oscillation cycle the tip experiences a whole range of forces. For each step during the approach the differential equation for the whole oscillation loop (including also the feedback system) was evaluated and the relation between force and frequency shift revealed. It was possible to determine the quantitative interaction forces of a metallic contact of nanometer dimensions.

However, it was shown that there also exists an analytical relationship if some approximations are accepted [15.7, 26, 27]. Here, we will follow the route indicated by *Dürig* [15.27], although different alternative ways have also proven successful. Consider the tip oscillation trajectory reaching over a large part of the force gradient curve in Fig. 15.2. We model the tip-sample interaction as a spring constant of stiffness $k_{\mathrm{ts}}(z) = \partial F/\partial z|_{z_0}$, as in Fig. 15.1. For small oscillation amplitudes, we already found that the frequency shift is proportional to the force gradient in (15.4). For large amplitudes, we can calculate an effective force gradient k_{eff} as a convolution of the force and the fraction of time the tip spends between the positions x and $x + \mathrm{d}x$:

$$k_{\mathrm{eff}}(z) = \frac{2}{\pi A^2} \int_{z}^{z+2A} F(x) g\left(\frac{x-z}{A} - 1\right) \mathrm{d}x$$
$$\text{with } g(u) = -\frac{u}{\sqrt{1-u^2}} \,. \tag{15.18}$$

In the approximation that the vibration amplitude is much larger than the range of the tip-sample forces, the above equation can be simplified to:

$$k_{\text{eff}}(z) = \frac{\sqrt{2}}{\pi} A^{3/2} \int_z^\infty \frac{F(x)}{\sqrt{x-z}}\, \mathrm{d}x\,. \tag{15.19}$$

This effective force gradient can now be used in (15.4), the relation between frequency shift and force gradient. We find:

$$\Delta f = \frac{f_0}{\sqrt{2}\pi k A^{3/2}} \int_z^\infty \frac{F(x)}{\sqrt{x-z}}\, \mathrm{d}x\,. \tag{15.20}$$

If we separate the integral from other parameters, we can define:

$$\Delta f = \frac{f_0}{kA^{3/2}}\gamma(z)$$

$$\text{with } \gamma(z) = \frac{1}{\sqrt{2}\pi} \int_z^\infty \frac{F(x)}{\sqrt{x-z}}\, \mathrm{d}x\,. \tag{15.21}$$

This means we can define $\gamma(z)$, which is only dependent on the shape of the force curve $F(z)$ but independent of the external parameters of the oscillation. The function $\gamma(z)$ is also referred to as the "normalized frequency shift" [15.7], a very useful parameter, which allows us to compare measurements independent of resonant frequency, amplitude and spring constant of the cantilever.

The dependence of the frequency shift on the vibration amplitude is an especially useful relation, since this parameter can be easily varied during one experiment. A nice example is depicted in Fig. 15.14, where frequency shift curves for different amplitudes were found to coincide very well in the $\gamma(z)$ diagrams [15.28].

This relationship has been nicely exploited for the calibration of the vibration amplitude by *Guggisberg* [15.29], which is a problem often encountered in dynamic AFM operation and worthwhile discussing. One approaches tip and sample and takes frequency shift curves versus distance, which show a reproducible shape. Then, the z-feedback is permanently turned off, and several curves with different amplitudes are taken. The amplitudes are typically chosen by adjusting the amplitude set-point in volts. One has to take care that drift in the z-direction is negligible. An analysis of the corresponding $\gamma(z)$-curves will show the same curves (as in Fig. 15.14), but the curves will be shifted in the horizontal axis. These shifts correspond to the change in amplitude, allowing one to correlate the voltage values with the z-distances.

For the oft-encountered force contributions from electrostatic, van der Waals, and chemical binding forces the frequency shift has been calculated from the force laws. In the approximation that the tip radius R is larger than the tip-sample distance z, an electrostatic potential V will yield a normalized frequency shift of (adapted from [15.30]):

$$\gamma(z) = \frac{\pi\varepsilon_0 R V^2}{\sqrt{2}} z^{-1/2}\,. \tag{15.22}$$

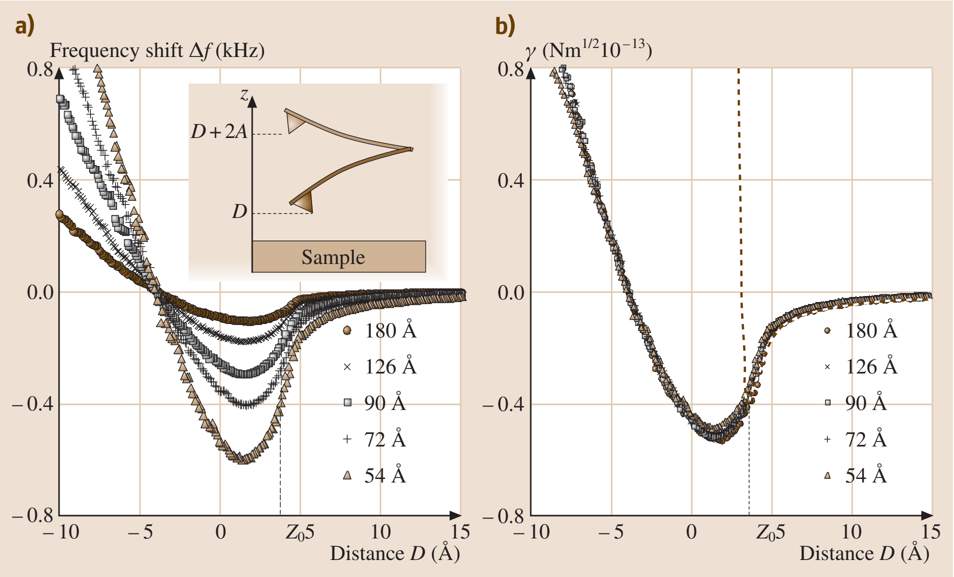

Fig. 15.14 **(a)** Frequency-shift curves for different oscillation amplitudes for a silicon tip on a graphite surface in UHV, **(b)** γ-curves calculated from the Δf-curves in **(a)**. (After [15.28] with permission, Copyright (2000) by The American Physical Society)

For van der Waals forces with Hamaker constant H and with R larger than z, we find accordingly:

$$\gamma(z) = \frac{HR}{12\sqrt{2}} z^{-3/2} \,. \tag{15.23}$$

Finally, short-range chemical forces represented by the well-known Morse potential (with the parameters binding energy U_0, decay length λ, and equilibrium distance z_{equ}) yield:

$$\gamma(z) = \frac{U_0\sqrt{2}}{\sqrt{\pi\lambda}} \exp\left(-\frac{(z - z_{\text{equ}})}{\lambda}\right) \,. \tag{15.24}$$

These equations allow the experimentalist to directly interpret the spectroscopic measurements. For example, the contributions of the electrostatic and van der Waals forces can be easily distinguished by their slope in a log-log plot (for an example, see [15.30]).

Alternatively, if the force law is not known beforehand, the experimentalist wants to analyze the experimental frequency-shift data curves and extract the force or energy potential curves. We, therefore, have to invert the integral in (15.21) to find the tip-sample interaction potential V_{ts} from the $\gamma(z)$-curves [15.27]:

$$V_{\text{ts}}(z) = \sqrt{2} \int_z^\infty \frac{\gamma(x)}{\sqrt{x-z}} \, dx \,. \tag{15.25}$$

Using this method, quantitative force curves were extracted from Δf-spectroscopy measurements on different, atomically resolved sites of the Si(111)-(7×7) reconstruction [15.1]. Comparison to theoretical MD simulations showed good quantitative agreement with theory and confirmed the assumption that force interactions were governed by a single atom at the tip apex. Our initially formulated goal seems to be achieved: With FM-AFM we have found a powerful method that allows us to measure the chemical bond formation of single molecules. The last uncertainty, the exact shape and identity of the tip apex atom, can possibly be resolved by employing the FIM technique to characterize the tip surface in combination with FM-AFM.

All the above equations are strictly valid only in the approximation that the oscillation amplitudes are much larger than the distance range of the encountered forces. For amplitudes of 10 nm and long-range forces like electrostatic interactions this approximation may not be valid. An iterative approach has been presented by *Dürig* [15.31] to overcome this problem. The interaction force is calculated from the frequency shift curves [gradient of (15.25)]. This force curve again is used to calculate the frequency shift with the exact relation from (15.18). The difference between the experimental Δf-curve and the reconstructed Δf-curve can now be used to calculate a first-order correction term to the interaction force. This procedure can be iteratively followed until sufficient agreement is found.

In this context it is worthwhile to point out a slightly different dynamic AFM method. While in the typical FM-AFM setup the oscillation amplitude is controlled to stay constant by a dedicated feedback circuit, one could simply keep the excitation amplitude constant (this has been termed *CA = constant amplitude*, as opposed to the *CE = constant excitation* mode). It is expected that this mode is more gentle to the surface, because any dissipative interaction will reduce the amplitude and therefore increase the tip-sample distance. The tip is prevented from deeply indenting the surface. This mode has been employed to image soft biological molecules like DNA or thiols in UHV [15.32]. However, quantitative interpretation of the obtained frequency spectra is more complicated, since the amplitude and tip-sample distance are altered during the measurement. Until now, we have always associated the self-excitation scheme with vacuum applications. Although it is difficult to operate the FM-AFM in constant amplitude mode in air, since large dissipative effects make it difficult to ensure a constant amplitude, it is indeed possible to use the constant-excitation FM-AFM in air or even in liquid. Still, only a few applications of FM-AFM under ambient or liquid conditions have been reported so far. In fact, a low budget construction set (employing a tuning-fork force sensor) for a CE-mode dynamic AFM setup has been published on the Internet (*http://www.sxm4.uni-muenster.de*).

15.4 *Q*-Control

We have already discussed the virtues of a high Q value for high sensitivity measurements: The minimum detectable force gradient was inversely proportional to the square root of Q. In vacuum, Q mainly represents the internal dissipation of the cantilever during oscillation, an internal damping factor. Little damping is obtained by using high quality cantilevers, which are cut (or etched) from defect-free, single-crystal silicon wafers. Under ambient or liquid conditions, the quality factor is dominated by dissipative interactions between the cantilever and the surrounding medium, and Q values can be as low as 100 for air or even 5 in liquid. Still, we ask if it is somehow possible to compensate for the damping effect by exciting the cantilever in a sophisticated way?

It turns out that the shape of the resonance curves in Fig. 15.15 can be influenced toward higher (or lower) Q values by an amplitude feedback loop. In principle, there are several mechanisms to couple the amplitude signal back to the cantilever, by the photothermal effect [15.33] or capacitive forces [15.34]. Figure 15.16 shows a method in which the amplitude feedback is mediated directly by the excitation piezo [15.35]. This has the advantage that no additional mechanical setups are necessary.

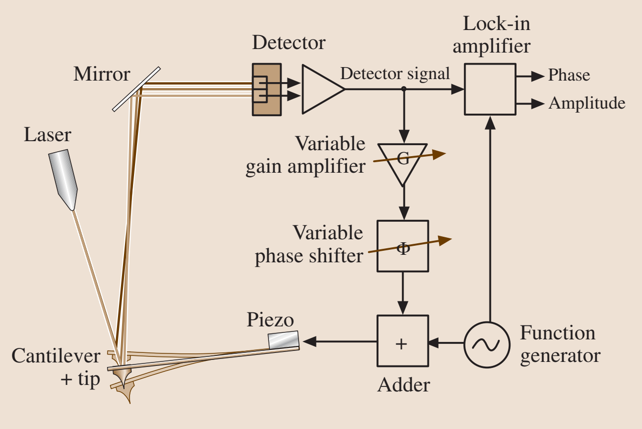

Fig. 15.16 Schematic diagram of operating a Q-Control feedback circuit with an externally driven dynamic AFM. The tapping mode setup is in effect extended by an additional feedback loop

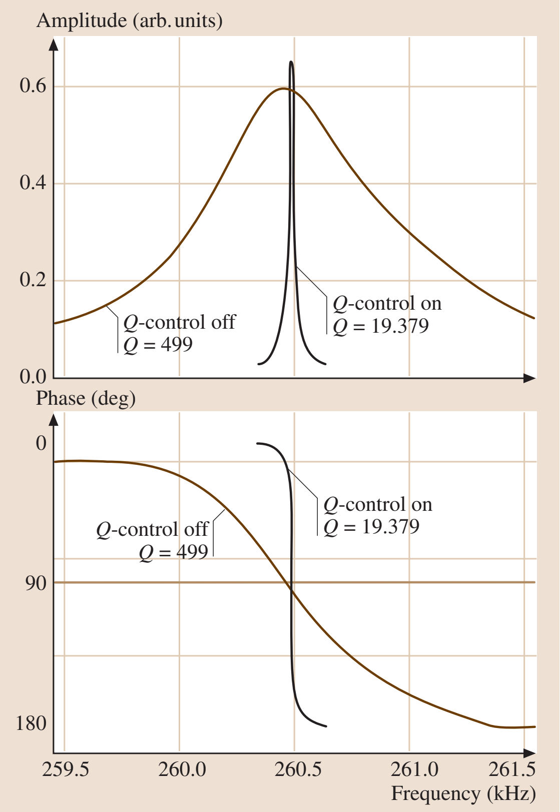

Fig. 15.15 Amplitude and phase diagrams measured in air with a Si cantilever far away from the sample. The quality factor can be increased from 450 to 20,000 by using the Q-Control feedback method

The working principle of the feedback loop can be understood by analyzing the equation of motion of the modified dynamic system:

$$m^*\ddot{z}(t) + \alpha\dot{z}(t) + kz(t) - F_{\mathrm{ts}}\,(z_0 + z(t)) = F_{\mathrm{ext}}\cos(\omega t) + G\,\mathrm{e}^{\mathrm{i}\phi} z(t)\,. \tag{15.26}$$

This ansatz takes into account the feedback of the detector signal through a phase shifter, amplifier, and adder as an additional force, which is linked to the cantilever deflection $z(t)$ through the gain G and the phase shift $\mathrm{e}^{\mathrm{i}\phi}$. To a good approximation, we assume that the oscillation can be described by a harmonic trajectory. With a phase

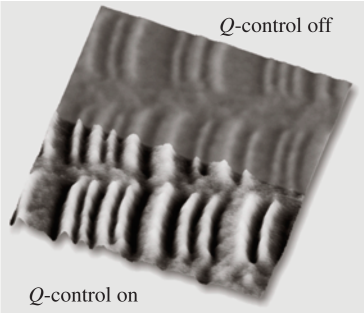

Fig. 15.17 Enhancement of the contrast in the phase channel due to Q-Control on a magnetic hard disk measured with a magnetic tip in tapping-mode AFM in air. Scan size 5 × 5 µm, phase range 10 deg (www.nanoanalytics.com)

shift of $\phi = \pm\pi/2$ we find:

$$e^{\pm i\pi/2} z(t) = \pm \frac{1}{\omega} \dot{z}(t)\,. \tag{15.27}$$

This means that the additional feedback force signal $G e^{i\phi} z(t)$ is proportional to the velocity of the cantilever, just like the damping term in the equation of motion. We can define an effective damping constant α_{eff}, which combines the two terms:

$$\begin{aligned} &m^* \ddot{z}(t) + \alpha_{\text{eff}} \dot{z}(t) + kz(t) - F_{\text{ts}}(z_0 + z(t)) \\ &\quad = F_{\text{ext}} \cos(\omega t) \\ &\quad \text{with } \alpha_{\text{eff}} = \alpha \mp \frac{1}{\omega} G \quad \text{for } \phi = \pm \frac{\pi}{2}\,. \end{aligned} \tag{15.28}$$

Equation (15.28) shows that the dampening of the oscillator can be enhanced or weakened by the choice of G, $\phi = +\pi/2$ or $\phi = -\pi/2$, respectively. The feedback loop, therefore, allows us to vary the effective quality factor $Q_{\text{eff}} = m\omega_0/\alpha_{\text{eff}}$ of the complete dynamic system. Hence, we term this system Q-Control. Figure 15.15 shows experimental data on the effect of Q-Control on the amplitude and phase as a function of the external excitation frequency [15.35]. While $Q = 449$, without Q-Control the quality factor of the system operated in air is enhanced by a factor of more than 40 by applying the feedback loop.

The effect of improved image contrast is demonstrated in Fig. 15.17. Here, a computer hard disk was analyzed with a magnetic tip in tapping mode, where the magnetic contrast is observed in the phase image. The upper part shows the recorded magnetic data structures in standard mode, whereas in the lower part of the image Q-Control feedback was activated, giving rise to an improved signal, i. e., magnetic contrast. A more detailed analysis of measurements on a magnetic tape shows that the signal amplitude (upper diagrams in Fig. 15.18), i. e.,

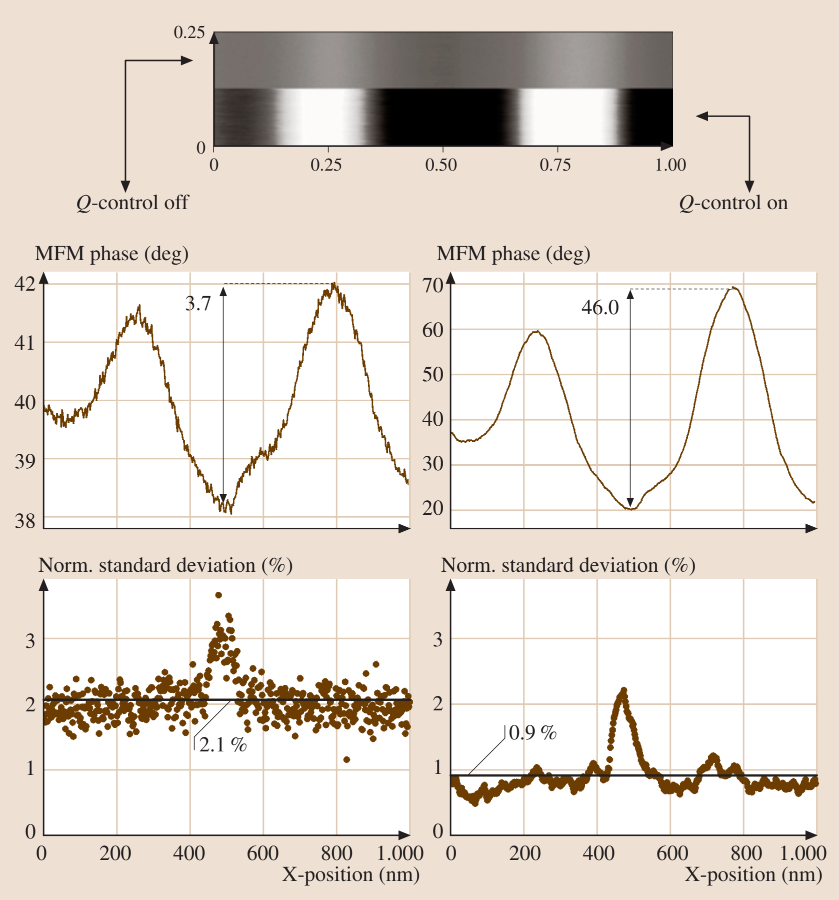

Fig. 15.18 Signal-to-noise analysis with a magnetic tip in tapping-mode AFM on a magnetic tape sample with Q-Control

the image contrast, was increased by a factor of 12.4 by the Q-Control feedback. The lower image shows a noise analysis of the signal, indicating an increase of the signal-to-noise ratio by a factor of 2.3.

Note that the diagrams represent measurements in air with an AFM operated in AM mode. Only then can we make a distinction between excitation and vibration frequency, since in the FM mode these two frequencies are equal by definition. Although the relation between sensitivity and Q-factor in (15.17) is the same for AM and FM mode, it must be critically investigated to see whether the enhanced quality factor by Q-Control can be inserted in the equation for FM-mode AFM. In vacuum applications, Q is already very high, which makes it oblique to operate an additional Q-Control module.

As stated before, we can also use Q-Control to enhance the dampening in the oscillating system. This would decrease the sensitivity of the system. But on the other hand, the response time of the amplitude change is decreased as well. For tapping-mode applications, where high speed scanning is the goal, Q-Control will be able to reduce the relaxation time [15.36].

A large quality factor Q does not only have the virtue of increasing the force sensitivity of the instrument. It also has the advantage of increasing the parameter space of stable AFM operation in tapping- or AM-mode AFM. Consider the resonance curve of Fig. 15.5. When approaching the tip toward the surface there are two competing mechanisms: On the one hand, we bring the tip closer to the sample, which results in an increase in attractive forces (see Fig. 15.2). On the other hand, for the case $\omega > \omega_0$, the resonant frequency of the cantilever is shifted toward smaller values due to the attractive forces, which causes the amplitude to become smaller, preventing a tip-sample contact. This is the desirable regime, where stable operation of the AFM is possible in the net-attractive regime. But as explained before, below a certain tip-sample separation d_{app}, the system switches suddenly into intermittent contact mode, where surface modifications are likely due to the onset of strong repulsive forces. The steeper the amplitude curve the larger the regime of stable, net-attractive AFM operation. Looking at Fig. 15.5 we find that the slope of the amplitude curve is governed by the quality factor Q. A high Q, therefore, facilitates stable operation of the AM-AFM in the net-attractive regime.

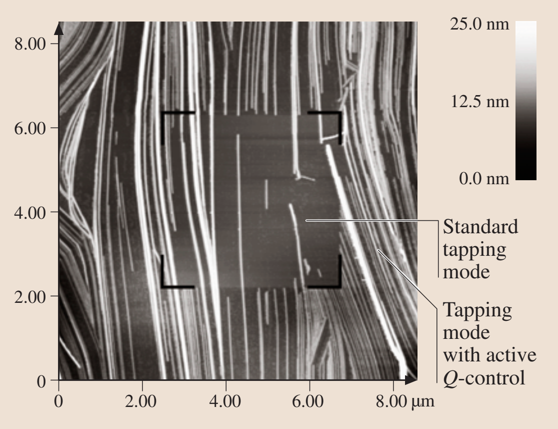

Fig. 15.19 Imaging of a delicate organic surface with Q-Control. Sample was a Langmuir-Blodgett film (ethyl-2,3-dihydroxyoctadecanoate) on a mica substrate. The topographical image clearly shows that the highly sensitive sample surface can only be imaged non-destructively with active Q-Control, whereas the periodic repulsive contact with the probe in standard operation without Q-Control leads to a significant modification or destruction of the surface structure. (Data courtesy of *Lifeng Chi* and coworkers, Westfälische Wilhelms-Universität, Münster, Germany)

An example can be found in Fig. 15.19. Here, a surface scan of an ultrathin organic film is acquired in tapping mode under ambient conditions. First, the inner square is scanned without the Q enhancement, and then a wider surface area was scanned with applied Q-Control. The high quality factor provides a larger parameter space for operating the AFM in the net-attractive regime, allowing good resolution of the delicate organic surface structure. Without the Q-Control the surface structures are deformed and even destroyed due to the strong repulsive tip-sample interactions [15.37–39]. This also allowed imaging of DNA structures without predominantly depressing the soft material during imaging. It was then possible to observe a DNA diameter close to the theoretical value with the Q-Control feedback [15.40].

In conclusion, we have shown that by applying an additional feedback circuit to the dynamic AFM system it is possible to influence the quality factor Q of the oscillator system. High-resolution, high-speed, or low-force scanning is then possible.

15.5 Dissipation Processes Measured with Dynamic AFM

Dynamic AFM methods have proven their great potential for imaging surface structures at the nanoscale, and we have also discussed methods that allow the assessment of forces between distinct single molecules. However, there is another physical mechanism that can be analyzed with the dynamic mode and has been mentioned in some previous paragraphs: energy dissipation. In Fig. 15.12, we have already shown an example, where the phase signal in tapping mode cannot be explained by conservative forces alone; dissipative processes must also play a role. In constant-amplitude FM mode, where the quantitative interpretation of experiments has proven to be less difficult, an intuitive distinction between conservative and dissipative tip-sample interaction is possible. While we have shown the correlation between forces and frequency shifts of the oscillating system, we have neglected one experimental input channel. The excitation amplitude, which is necessary to keep the oscillation amplitude constant, is a direct indication of the energy dissipated during one oscillation cycle. *Dürig* [15.41] has shown that in self-excitation mode (with an excitation-oscillation phase difference of 90 deg), conservative and dissipative interactions can be strictly separated. Part of this energy is dissipated in the cantilever itself, another part is due to external viscous forces in the surrounding medium. But more interestingly, some energy is dissipated at the tip-sample junction. This is the focus of the following paragraphs.

In contrast to conservative forces acting at the tip-sample junction, which at least in vacuum can be understood in terms of van der Waals, electrostatic, and chemical interactions, the dissipative processes are poorly understood. *Stowe* et al. [15.42] have shown that if a voltage potential is applied between tip and sample, charges are induced in the sample surface, which will follow the tip motion (here the oscillation is parallel to the surface). Due to the finite resistance of the sample material, energy will be dissipated during the charge movement. This effect has been exploited to image the doping level of semiconductors. Energy dissipation has also been observed in imaging magnetic materials. *Liu* et al. [15.43] found that energy dissipation due to magnetic interactions was enhanced at the boundaries of magnetic domains, which was attributed to domain wall oscillations. But also in the absence of external electromagnetic fields, energy dissipation was observed in close proximity of tip and sample, within 1 nm. Clearly, mechanical surface relaxations must give rise to energy losses. One could model the AFM tip as a small hammer, hitting the surface at high frequency, possibly resulting in phonon excitations. From a contiuum mechanics point of view, we assume that the mechanical relaxation of the surface is not only governed by elastic responses. Viscoelastic effects of soft surfaces will also render a significant contribution to energy dissipation. The whole area of phase imaging in tapping mode is concerned with those effects [15.44–47].

In the atomistic view, the last tip atom can be envisaged to change position while yielding to the tip-sample force field. A strictly reversible change of position would not result in a loss of energy. Still, it has been pointed out by *Sasaki* et al. [15.48] that a change in atom position would result in a change in the force interaction itself. Therefore, it is possible that the tip atom changes position at different tip-surface distances during approach and retraction, effectively causing an atomic-scale hysteresis to develop. In fact, *Hoffmann* et al. [15.13] have measured short-range energy dissipation for a tungsten tip on silicon in UHV. A theoretical model explaining this effect on the basis of a two-energy-state system was developed. However, a clear understanding of the underlying physical mechanism is still lacking.

Nonetheless, the dissipation channel has been used to image surfaces with atomic resolution [15.49]. Instead of feedbacking the distance on the frequency shift, the excitation amplitude in FM mode has been used as the control signal. The Si(111)-(7×7) reconstruction was successfully imaged in this mode. The step edges of monoatomic NaCl islands on single-crystalline copper have also rendered atomic resolution contrast in the dissipation channel [15.50]. The dissipation processes discussed so far are mostly in the configuration in which the tip is oscillated perpendicular to the surface. Friction is usually referred to as the energy loss due to lateral movement of solid bodies in contact. It is interesting to note in this context that *Israelachivili* [15.51] has pointed out a quantitative relationship between lateral and vertical (with respect to the surface) dissipation. He states that the hysteresis in vertical force-distance curves should equal the energy loss in lateral friction. An experimental confirmation of this conjecture at the molecular level is still missing.

Physical interpretation of energy dissipation processes at the atomic scale seems to be a daunting task at this point. Notwithstanding, we can find a quantitative relation between the energy loss per oscillation cycle and the experimental parameters in dynamic AFM, as will be shown in the following section.

In static AFM it was found that permanent changes of the sample surface by indentations can cause a hysteresis between approach and retraction (e.g., [15.5]). The area between the approach and retraction curves in a force-distance diagram represents the lost or dissipated energy caused by the irreversible change of the surface structure. In dynamic-mode AFM, the oscillation parameters like amplitude, frequency, and phase must contain the information about the dissipated energy per cycle. So far, we have resorted to a treatment of the equation of motion of the cantilever vibration in order to find a quantitative correlation between forces and the experimental parameters. For the dissipation it is useful to treat the system from the energy conservation point of view.

Assuming a dynamic system is in equilibrium, the average energy input must equal the average energy output or dissipation. Applying this rule to an AFM running in a dynamic mode means that the average power fed into the cantilever oscillation by an external driver, denoted $\bar{P}_{\text{in}}$, must equal the average power dissipated by the motion of the cantilever beam $\bar{P}_0$ and by tip-sample interaction $\bar{P}_{\text{tip}}$:

$$\bar{P}_{\text{in}} = \bar{P}_0 + \bar{P}_{\text{tip}} \,. \tag{15.29}$$

The term $\bar{P}_{\text{tip}}$ is what we are interested in, since it gives us a direct physical quantity to characterize the tip-sample interaction. Therefore, we have first to calculate and then measure the two other terms in (15.29) in order to determine the power dissipated when the tip periodically probes the sample surface. This requires an appropriate rheological model to describe the dynamic system. Although there are investigations in which the complete flexural motion of the cantilever beam has been considered [15.52], a simplified model, comprising a spring and two dashpots (Fig. 15.20), represents a good approximation in this case [15.53].

The spring, characterized by the constant k according to Hooke's law, represents the only channel through which power P_{in} can be delivered to the oscillating tip $z(t)$ by the external driver $z_{\text{d}}(t)$. Therefore, the instantaneous power fed into the dynamic system is equal to the force exerted by the driver times the velocity of the driver (the force that is necessary to move the base side of the dashpot can be neglected, since this power is directly dissipated and, therefore, does not contribute to the power delivered to the oscillating tip):

$$P_{\text{in}}(t) = F_{\text{d}}(t)\dot{z}_{\text{d}}(t) = k\,[z(t) - z_{\text{d}}(t)]\,\dot{z}_{\text{d}}(t) \,. \tag{15.30}$$

Assuming a sinusoidal steady state response and that the base of the cantilever is driven sinusoidally (see (15.6)) with amplitude A_{d} and frequency ω, the deflection from equilibrium of the end of the cantilever follows (15.9), where A and $0 \le \varphi \le \pi$ are the oscillation amplitude and phase shift, respectively. This allows us to calculate the average power input per oscillation cycle by integrating (15.30) over one period $T = 2\pi/\omega$:

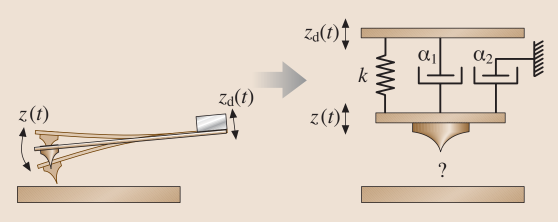

Fig. 15.20 Rheological model applied to describe the dynamic AFM system, comprising the oscillating cantilever and tip interacting with the sample surface. The movement of the cantilever base and the tip is denoted as $z_{\text{d}}(t)$ and $z(t)$, respectively. The cantilever is characterized by the spring constant k and the damping constant α. In a first approach, damping is broken into two pieces, α_1 and α_2. First, intrinsic damping caused by the movement of the cantilever's tip relative to its base. Second, damping related to the movement of the cantilever body in a surrounding medium, e.g., air damping

$$\bar{P}_{\text{in}} = \frac{1}{T}\int_0^T P_{\text{in}}(t)\;\text{d}t = \frac{1}{2}k\omega A_{\text{d}}A\sin\varphi \,. \tag{15.31}$$

This contains the familiar result that the maximum power is delivered to an oscillator when the response is 90 deg out of phase with the drive.

The simplified rheological model, as it is depicted in Fig. 15.20, exhibits two major contributions to the damping term $\bar{P}_0$. Both are related to the motion of the cantilever body and assumed to be well modeled by viscous damping with coefficients α_1 and α_2. The dominant damping mechanism in UHV conditions is intrinsic damping caused by the deflection of the cantilever beam, i. e., the motion of the tip relative to the cantilever base. Therefore, the instantaneous power dissipated by such a mechanism is given by

$$P_{01}(t) = |F_{01}(t)\dot{z}(t)| = |\alpha_1\,[\dot{z}(t) - \dot{z}_{\text{d}}(t)]\,\dot{z}(t)| \,. \tag{15.32}$$

Note that the absolute value has to be calculated, since all dissipated power is "lost" and, therefore, cannot be returned to the dynamic system.

However, when running an AFM in ambient conditions an additional damping mechanism has to be considered. Damping due to the motion of the cantilever body in the surrounding medium, e.g., air damping, is in most cases the dominant effect. The corresponding instantaneous power dissipation is given by

$$P_{02}(t) = |F_{02}(t)\dot{z}(t)| = \alpha_2 \dot{z}^2(t)\;. \tag{15.33}$$

In order to calculate the average power dissipation, (15.32) and (15.33) have to be integrated over one complete oscillation cycle. This yields

$$\begin{aligned}\bar{P}_{01} &= \frac{1}{T}\int_0^T P_{01}(t)\,\mathrm{d}t\\ &= \frac{1}{\pi}\alpha_1\omega^2 A\left[(A - A_\mathrm{d}\cos\varphi)\right.\\ &\quad\times\arcsin\left(\frac{A - A_\mathrm{d}\cos\varphi}{\sqrt{A^2 + A_\mathrm{d}^2 - 2AA_\mathrm{d}\cos\varphi}}\right)\\ &\quad\left.+A_\mathrm{d}\sin\varphi\right]\;,\end{aligned} \tag{15.34}$$

and

$$P_{02}(t) = \frac{1}{T}\int_0^T P_{02}(t)\,\mathrm{d}t = \frac{1}{2}\alpha_2\omega^2 A^2\;. \tag{15.35}$$

Considering the fact that commonly used cantilevers exhibit a quality factor of at least several hundreds (in UHV several tens of thousands), we can assume that the oscillation amplitude is significantly larger than the drive amplitude when the dynamic system is driven at or near its resonance frequency: $A \gg A_\mathrm{d}$. Therefore, (15.34) can be simplified in first order approximation to an expression similar to (15.35). Combining the two equations yields the total average power dissipated by the oscillating cantilever

$$\bar{P}_0 = \frac{1}{2}\alpha\omega^2 A^2 \quad \text{with } \alpha = \alpha_1 + \alpha_2\;, \tag{15.36}$$

where α denotes the overall effective damping constant.

We can now solve (15.29) for the power dissipation localized to the small interaction volume of the probing tip with the sample surface, represented by the question mark in Fig. 15.20. Furthermore, by expressing the damping constant α in terms of experimentally accessible quantities such as the spring constant k, the quality factor Q, and the natural resonant frequency ω_0 of the free oscillating cantilever, $\alpha = \frac{k}{Q\omega_0}$, we obtain:

$$\begin{aligned}\bar{P}_\mathrm{tip} &= \bar{P}_\mathrm{in} - \bar{P}_0\\ &= \frac{1}{2}\frac{k\omega}{Q}\left(Q_\mathrm{cant}A_\mathrm{d}A\sin\varphi - A^2\frac{\omega}{\omega_0}\right)\;.\end{aligned} \tag{15.37}$$

Note that so far no assumptions have been made regarding how the AFM is operated, except that the motion of the oscillating cantilever has to remain sinusoidal to a good approximation. Therefore, (15.37) is applicable to a variety of different dynamic AFM modes.

For example, in FM-mode AFM the oscillation frequency ω changes due to tip-sample interaction, while at the same time the oscillation amplitude A is kept constant by adjusting the drive amplitude A_d. By measuring these quantities, one can apply (15.37) to determine the average power dissipation related to tip-sample interaction. In spectroscopy applications, usually $A_\mathrm{d}(d)$ is not measured directly, but a signal $G(d)$ proportional to $A_\mathrm{d}(d)$ is acquired representing the gain factor applied to the excitation piezo. With the help of (15.15) we can write:

$$A_\mathrm{d}(d) = \frac{A_0 G(d)}{QG_0}\;, \tag{15.38}$$

while A_0 and G_0 are the amplitude and gain at large tip-sample distances, where the tip-sample interactions are negligible.

Now let us consider the tapping-mode, or AM-AFM. In this case, the cantilever is driven at a fixed frequency and with constant drive amplitude, while the oscillation amplitude and phase shift may change when the probing tip interacts with the sample surface. Assuming that the oscillation frequency is chosen to be ω_0, (15.37) can be further simplified by again employing (15.15) for the free oscillation amplitude A_0. This yields

$$\bar{P}_\mathrm{tip} = \frac{1}{2}\frac{k\omega_0}{Q}\left(A_0A\sin\varphi - A^2\right)\;. \tag{15.39}$$

Equation (15.39) implies that if the oscillation amplitude A is kept constant by a feedback loop, like it is commonly done in tapping mode, simultaneously acquired phase data can be interpreted in terms of energy dissipation [15.45, 47, 54, 55]. When analyzing such phase images [15.56–58], one has also to consider the fact that the phase may also change due to the transition from net-attractive ($\varphi > 90°$) to intermittent contact ($\varphi < 90°$) interaction between the tip and the sample [15.15, 35, 59, 60]. For example, consider the phase shift in tapping mode as a function of z-position, Fig. 15.12. If phase measurements are performed close to the point where the oscillation switches from the net-attractive to the intermittent contact regime, a large contrast in the phase channel is observed. However, this contrast is not due to dissipative processes. Only a variation of the phase signal within the intermittent contact regime will give information of the tip-sample dissipative processes.

An example of a dissipation measurement is depicted in Fig. 15.21. The surface of a polymer blend was imaged in air, simultaneously acquiring the topography and dissipation. The dissipation on the softer polyurethane-matrix is significantly larger than on the embedded, mechanically stiffer polypropylene particles.

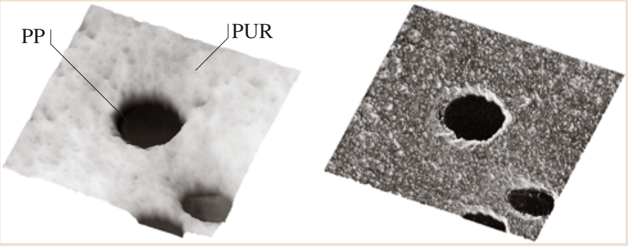

Topography
x,y,z-range: 5 µm × 5 µm × 546 nm

Dissipation
Data range: 3.0 pW or 257 eV

Fig. 15.21 Topography and phase image in tapping-mode AFM of a polymer blend composed of polypropylene (PP) particles embedded in a polyurethane (PUR) matrix. The dissipation image shows a strong contrast between the harder PP (little dissipation, *dark*) to the softer PUR (large dissipation, *bright*) surface

15.6 Conclusion

Dynamic force microscopy is a powerful tool that is capable of imaging surfaces with atomic precision. It also allows us to look at surface dynamics, and it can operate in vacuum, air, or even in liquid. However, the oscillating cantilever system introduces a level of complexity that disallows a straightforward interpretation of acquired images. An exception is the self-excitation mode, where tip-sample forces can be successfully extracted from spectroscopic experiments. However, not only conservative forces can be investigated with dynamic AFM; energy dissipation also influences the cantilever oscillation and can, therefore, serve as a new information channel.

Open questions still exist concerning the exact geometric and chemical identity of the probing tip, which significantly influences the imaging and spectroscopic results. Using predefined tips like single-walled nanotubes or atomic resolution techniques like field ion microscopy to image the tip itself are possible approaches to addressing this issue. Furthermore, little is known about the dissipative processes at the scale where only a few atoms between tip and sample interact. It is also desirable to learn more about the interpretation of images acquired in tapping mode.

References

15.1 M.A. Lantz, H.J. Hug, R. Hoffmann, P.J.A. van Schendel, P. Kappenberger, S. Martin, A. Baratoff, H.-J. Güntherodt: Quantitative measurement of short-range chemical bonding forces, Science **291** (2001) 2580–2583

15.2 G. Binnig, C.F. Quate, Ch. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

15.3 O. Marti: AFM instrumentation and tips. In: *Handbook of Micro/Nanotribology*, 2nd edn., ed. by B. Bhushan (CRC, Boca Raton 1999) pp. 81–144

15.4 G. Cross, A. Schirmeisen, A. Stalder, P. Grütter, M. Tschudy, U. Dürig: Adhesion interaction between atoically defined tip and sample, Phys. Rev. Lett. **80** (1998) 4685–4688

15.5 A. Schirmeisen, G. Cross, A. Stalder, P. Grütter, U. Dürig: Metallic adhesion and tunneling at the atomic scale, New J. Phys. **2** (2000) 29.1–29.10

15.6 A. Schirmeisen: Metallic adhesion and tunneling at the atomic scale. Ph.D. Thesis (McGill University, Montréal 1999) pp. 29–38

15.7 F.J. Giessibl: Forces and frequency shifts in atomic-resolution dynamic-force microscopy, Phys. Rev. B **56** (1997) 16010–16015

15.8 F.J. Giessibl: Atomic resolution of the silicon (111)-(7×7) surface by atomic force microscopy, Science **267** (1995) 68–71

15.9 M. Bezanilaa, B. Drake, E. Nudler, M. Kashlev, P. K. Hansma, H. G. Hansma: Motion and enzymatic degradation of DNA in the atomic force microscope, Biophys. J. **67** (1994) 2454–2459

15.10 Y. Jiao, D. I. Cherny, G. Heim, T. M. Jovin, T. E. Schäffer: Dynamic interactions of p53 with DNA in solution by time-lapse atomic force microscopy, J. Mol. Biol. **314** (2001) 233–243

15.11 T. R. Albrecht, P. Grütter, D. Horne, D. Rugar: Frequency modulation detection using high-Q cantilevers for enhanced force microscopy sensitivity, J. Appl. Phys. **69** (1991) 668–673

15.12 S. P. Jarvis, M. A. Lantz, U. Dürig, H. Tokumoto: Off resonance AC mode force spectroscopy and imaging with an atomic force microscope, Appl. Surf. Sci. **140** (1999) 309–313

15.13 P. M. Hoffmann, S. Jeffery, J. B. Pethica, H.Ö. Özer, A. Oral: Energy dissipation in atomic force microscopy and atomic loss processes, Phys. Rev. Lett. **87** (2001) 265502–265505

15.14 B. Anczykowski, D. Krüger, H. Fuchs: Cantilever dynamics in quasinoncontact force microscopy: Spectroscopic aspects, Phys. Rev. B **53** (1996) 15485–15488

15.15 B. Anczykowski, D. Krüger, K. L. Babcock, H. Fuchs: Basic properties of dynamic force spectroscopy with the scanning force microscope in experiment and simulation, Ultramicroscopy **66** (1996) 251–259

15.16 V. M. Muller, V. S. Yushchenko, B. V. Derjaguin: On the influence of molecular forces on the deformation of an elastic sphere and its sticking to a rigid plane, J. Coll. Interf. Sci. **77** (1980) 91–101

15.17 B. D. Hughes, L. R. White: 'Soft' contact problems in linear elasticity, Quart. J. Mech. Appl. Math. **32** (1979) 445–471

15.18 L. Verlet: Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules, Phys. Rev. **159** (1967) 98–103

15.19 L. Verlet: Computer "experiments" on classical fluids. II. Equilibrium correlation functions, Phys. Rev. **165** (1968) 201–214

15.20 P. Gleyzes, P. K. Kuo, A. C. Boccara: Bistable behavior of a vibrating tip near a solid surface, Appl. Phys. Lett. **58** (1991) 2989–2991

15.21 A. San Paulo, R. García: High-resolution imaging of antibodies by tapping-mode atomic force microscopy: Attractive and repulsive tip-sample interaction regimes, Biophys. J. **78** (2000) 1599–1605

15.22 D. Krüger, B. Anczykowski, H. Fuchs: Physical properties of dynamic force microscopies in contact and noncontact operation, Ann. Phys. **6** (1997) 341–363

15.23 Y. Martin, C. C. Williams, H. K. Wickramasinghe: Atomic force microscope – force mapping and profiling on a sub 100-Å scale, J. Appl. Phys. **61** (1987) 4723–4729

15.24 C. Barth, M. Reichling: Imaging the atomic arrangement on the high-temperature reconstructed α-Al$_2$O$_3$(0001) surface, Nature **414** (2001) 54–57

15.25 B. Gotsmann, H. Fuchs: Dynamic force spectroscopy of conservative and dissipative forces in an Al-Au(111) tip-sample system, Phys. Rev. Lett. **86** (2001) 2597–2600

15.26 H. Hölscher, W. Allers, U. D. Schwarz, A. Schwarz, R. Wiesendanger: Determination of tip-sample interaction potentials by dynamic force spectroscopy, Phys. Rev. Lett. **83** (1999) 4780–4783

15.27 U. Dürig: Relations between interaction force and frequency shift in large-amplitude dynamic force microscopy, Appl. Phys. Lett. **75** (1999) 433–435

15.28 H. Hölscher, A. Schwarz, W. Allers, U. D. Schwarz, R. Wiesendanger: Quantitative analysis of dynamic-force-spectroscopy data on graphite(0001) in the contact and noncontact regime, Phys. Rev. B **61** (2000) 12678–12681

15.29 M. Guggisberg: Lokale Messung von atomaren Kräften. Ph.D. Thesis (University Basel, Basel 2000) pp. 9–11

15.30 M. Guggisberg, M. Bammerlin, E. Meyer, H.-J. Güntherodt: Separation of interactions by noncontact force microscopy, Phys. Rev. B **61** (2000) 11151–11155

15.31 U. Dürig: Extracting interaction forces and complementary observables in dynamic probe microscopy, Appl. Phys. Lett. **76** (2000) 1203–1205

15.32 T. Uchihasi, T. Ishida, M. Komiyama, M. Ashino, Y. Sugawara, W. Mizutani, K. Yokoyama, S. Morita, H. Tokumoto, M. Ishikawa: High-resolution imaging of organic monolayers using noncontact AFM, Appl. Surf. Sci. **157** (2000) 244–250

15.33 J. Mertz, O. Marti, J. Mlynek: Regulation of a microcantilever response by force feedback, Appl. Phys. Lett. **62** (1993) 2344–2346

15.34 D. Rugar, P. Grütter: Mechanical parametric amplification and thermomechanical noise squeezing, Phys. Rev. Lett. **67** (1991) 699–702

15.35 B. Anczykowski , J. P. Cleveland, D. Krüger, V. B. Elings, H. Fuchs: Analysis of the interaction mechanisms in dynamic mode SFM by means of experimental data and computer simulation, Appl. Phys. A **66** (1998) 885–889

15.36 T. Sulchek, G. G. Yaralioglu, C. F. Quate, S. C. Minne: Characterization and optimisation of scan speed for tapping-mode atomic force microscopy, Rev. Sci. Instrum. **73** (2002) 2928–2936

15.37 L. F. Chi, S. Jacobi, B. Anczykowski, M. Overs, H.-J. Schäfer, H. Fuchs: Supermolecular periodic structures in monolayers, Adv. Mater. **12** (2000) 25–30

15.38 S. Gao, L. F. Chi, S. Lenhert, B. Anczykowski, C. Niemeyer, M. Adler, H. Fuchs: High-quality mapping of DNA-protein complexes by dynamic scanning force microscopy, ChemPhysChem **6** (2001) 384–388

15.39 B. Zou, M. Wang, D. Qiu, X. Zhang, L. F. Chi, H. Fuchs: Confined supramolecular nanostructures of mesogen-bearing amphiphiles, Chem. Commun. **9** (2002) 1008–1009

15.40 B. Pignataro, L. F. Chi, S. Gao, B. Anczykowski, C. Niemeyer, M. Adler, H. Fuchs: Dynamic scanning force microscopy study of self-assembled DNA-protein nanostructures, Appl. Phys. A **74** (2002) 447–452
15.41 U. Dürig: Interaction sensing in dynamic force microscopy, New J. Phys. **2** (2000) 5.1–5.2
15.42 T. D. Stowe, T. W. Kenny, D. J. Thomson, D. Rugar: Silicon dopant imaging by dissipation force microscopy, Appl. Phys. Lett. **75** (1999) 2785–2787
15.43 Y. Liu, P. Grütter: Magnetic dissipation force microscopy studies of magnetic materials, J. Appl. Phys. **83** (1998) 7333–7338
15.44 J. Tamayo, R. García: Effects of elastic and inelastic interactions on phase contrast images in tapping-mode scanning force microscopy, Appl. Phys. Lett. **71** (1997) 2394–2396
15.45 J. P. Cleveland, B. Anczykowski, A. E. Schmid, V. B. Elings: Energy dissipation in tapping-mode atomic force microscopy, Appl. Phys. Lett. **72** (1998) 2613–2615
15.46 R. García, J. Tamayo, M. Calleja, F. García: Phase contrast in tapping-mode scanning force microscopy, Appl. Phys. A **66** (1998) 309–312
15.47 B. Anczykowski, B. Gotsmann, H. Fuchs, J. P. Cleveland, V. B. Elings: How to measure energy dissipation in dynamic mode atomic force microscopy, Appl. Surf. Sci. **140** (1999) 376–382
15.48 N. Sasaki, M. Tsukada: Effect of microscopic nonconservative process on noncontact atomic force microscopy, Jpn. J. Appl. Phys. **39** (2000) L1334–L1337
15.49 R. Lüthi, E. Meyer, M. Bammerlin, A. Baratoff, L. Howald, C. Gerber, H.-J Güntherodt: Ultrahigh vacuum atomic force microscopy: true atomic resolution, Surf. Rev. Lett. **4** (1997) 1025–1029
15.50 R. Bennewitz, A. S. Foster, L. N. Kantorovich, M. Bammerlin, Ch. Loppacher, S. Schär, M. Guggisberg, E. Meyer, A. L. Shluger: Atomically resolved edges and kinks of NaCl islands on Cu(111): experiment and theory, Phys. Rev. B **62** (2000) 2074–2084
15.51 J. Israelachvili: *Intermolecular and Surface Forces* (Academic, London 1992)
15.52 U. Rabe, J. Turner, W. Arnold: Analysis of the high-frequency response of atomic force microscope cantilevers, Appl. Phys. A **66** (1998) 277–282
15.53 T. R. Rodríguez, R. García: Tip motion in amplitude modulation (tapping-mode) atomic-force microscopy: Comparison between continuous and point-mass models, Appl. Phys. Lett. **80** (2002) 1646–1648
15.54 J. Tamayo, R. García: Relationship between phase shift and energy dissipation in tapping-mode scanning force microscopy, Appl. Phys. Lett. **73** (1998) 2926–2928
15.55 R. García, J. Tamayo, A. San Paulo: Phase contrast and surface energy hysteresis in tapping mode scanning force microsopy, Surf. Interface Anal. **27** (1999) 312–316
15.56 S. N. Magonov, V. B. Elings, M. H. Whangbo: Phase imaging and stiffness in tapping-mode atomic force microscopy, Surf. Sci. **375** (1997) L385–L391
15.57 J. P. Pickering, G. J. Vancso: Apparent contrast reversal in tapping mode atomic force microscope images on films of polystyrene-b-polyisoprene-b-polystyrene, Polymer Bull. **40** (1998) 549–554
15.58 X. Chen, S. L. McGurk, M. C. Davies, C. J. Roberts, K. M. Shakesheff, S. J. B. Tendler, P. M. Williams, J. Davies, A. C. Dwakes, A. Domb: Chemical and morphological analysis of surface enrichment in a biodegradable polymer blend by phase-detection imaging atomic force microscopy, Macromolecules **31** (1998) 2278–2283
15.59 A. Kühle, A. H. Sørensen, J. Bohr: Role of attractive forces in tapping tip force microscopy, J. Appl. Phys. **81** (1997) 6562–6569
15.60 A. Kühle, A. H. Sørensen, J. B. Zandbergen, J. Bohr: Contrast artifacts in tapping tip atomic force microscopy, Appl. Phys. A **66** (1998) 329–332

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

16. Molecular Recognition Force Microscopy

Atomic force microscopy (AFM), developed in the late eighties to explore atomic details on hard material surfaces, has evolved to an imaging method capable of achieving fine structural details on biological samples. Its particular advantage in biology is that the measurements can be carried out in aqueous and physiological environment, which opens the possibility to study the dynamics of biological processes in vivo. The additional potential of the AFM to measure ultra-low forces at high lateral resolution has paved the way for measuring inter- and intra-molecular forces of bio-molecules on the single molecule level. Molecular recognition studies using AFM open the possibility to detect specific ligand–receptor interaction forces and to observe molecular recognition of a single ligand–receptor pair. Applications include biotin–avidin, antibody–antigen, NTA nitrilotriacetate–hexahistidine 6, and cellular proteins, either isolated or in cell membranes.

The general strategy is to bind ligands to AFM tips and receptors to probe surfaces (or vice versa), respectively. In a force–distance cycle, the tip is first approached towards the surface whereupon a single receptor–ligand complex is formed, due to the specific ligand receptor recognition. During subsequent tip–surface retraction a temporarily increasing force is exerted to the ligand–receptor connection thus reducing its lifetime until the interaction bond breaks at a critical force (unbinding force). Such experiments allow for estimation of affinity, rate constants, and structural data of the binding pocket. Comparing them with values obtained from ensemble-average techniques and binding energies is of particular interest. The dependences of unbinding force on the rate of load increase exerted to the receptor–ligand bond reveal details of the molecular dynamics of the recognition process and energy landscapes. Similar experimental strategies were also used for studying intra-molecular force properties of polymers and unfolding–refolding kinetics of filamentous proteins. Recognition imaging, developed by combing dynamic force microscopy with force spectroscopy, allows for localization of receptor sites on surfaces with nanometer positional accuracy.

16.1 **Ligand Tip Chemistry** 476

16.2 **Fixation of Receptors to Probe Surfaces** .. 478

16.3 **Single-Molecule Recognition Force Detection** 479

16.4 **Principles of Molecular Recognition Force Spectroscopy** 482

16.5 **Recognition Force Spectroscopy: From Isolated Molecules to Biological Membranes** 484
- 16.5.1 Forces, Energies, and Kinetic Rates. 484
- 16.5.2 Complex Bonds and Energy Landscapes 486
- 16.5.3 Live Cells and Membranes 489

16.6 **Recognition Imaging** 489

16.7 **Concluding Remarks** 491

References 492

Macromolecular interactions play a key role in cellular regulation and biological function. Molecular recognition is one of the most important regulatory elements, as it is often the initiating step in reaction pathways and cascades. Since the immune system has evolved recognition molecules that are both highly specific and versatile, antibody-antigen interaction is often used as a paradigm for molecular recognition. However, molecular recognition also plays fundamental roles in the regulation of gene expression, cellular metabolism, and drug design.

Molecular recognition studies emphasize specific biological interactions between receptors and their cog-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

nitive ligands. Despite a growing literature on the structure and function of receptor-ligand complexes, it is still not possible to predict reaction kinetics or energetics for any given complex formation, even when the structures are known. Additional insights, in particular about the molecular dynamics within the complex during the association and dissociation process, are needed. The high-end strategy is to probe quantitatively the science that underlies a specific biological recognition in which the chemical structures can be defined at atomic resolution, and genetic or chemical means are available to vary the structure of the interacting partners.

Receptor-ligand complexes are usually formed by a few tens of parallel weak interactions of contacting chemical groups in complementary determining regions supported by framework residues providing structurally conserved scaffolding. Both the complementary determining regions and the framework have a considerable amount of plasticity and flexibility resulting in conformational changes upon association and dissociation (induced fit). In addition to what is known about structure, energies, and kinetic constants, information about conformational movements is required for the understanding of the recognition process. It is likely that insight into the temporal and spatial action of the many weak interactions, in particular the cooparativity of bond formation, is the key to understanding receptor-ligand recognition.

For this, experiments on the single-molecule level at time scales typical for receptor-ligand complex formation and dissociation appear to be required. The potential of the atomic force microscope (AFM) [16.1] to measure ultralow forces at high lateral resolution has paved the way for single-molecule recognition force microscopy studies. The particular advantage of AFM in biology is that the measurements can be carried out in aqueous and physiological environments, which opens the possibility for studying biological processes in vivo. The methodology for investigating the molecular dynamics of receptor-ligand interactions described in this chapter, *Molecular Recognition Force Microscopy* (MRFM) [16.2–4], was developed from scanning probe microscopy (SPM) [16.1]. A force is exerted on a receptor-ligand complex, and the dissociation process is followed over time. Dynamic aspects of recognition are addressed in force spectroscopy (FS) experiments, in which distinct force-time profiles are applied to give insight into the changes in conformations and states during receptor-ligand dissociation. It will be shown that MRFM is a versatile tool to explore kinetic and structural details of receptor-ligand recognition.

16.1 Ligand Tip Chemistry

In MRFM experiments, the binding of ligands on AFM tips to surface-bound receptors (or vice versa) is studied by applying a force to the receptor-ligand complex that reduces its lifetime until the bond breaks at a measurable unbinding force. This requires a careful AFM tip sensor design, including tight attachment of the ligands to the tip surface. In the first pioneering demonstrations of single-molecule recognition force measurements [16.2, 3], strong physical adsorption of bovine serum albumin (BSA) was used to directly coat the tip [16.3] or, alternatively, a glass bead glued to it [16.2]. This physisorbed protein layer may then serve as a functional matrix for the biochemical modification with chemically active ligands (Fig. 16.1). In spite of the large number of probe molecules on the tip ($10^3-10^4/\mu m^2$), the low fraction of properly oriented molecules, or internal blocks of most reactive sites (see Fig. 16.1), allowed the measurement of single receptor-ligand unbinding forces. Nevertheless, parallel breakage of multiple bonds was predominately observed with this configuration.

For measuring interactions between isolated receptor-ligand pairs, strictly defined conditions need to be fulfilled. Covalently coupling ligands to gold-coated tip surfaces via freely accessible SH-groups guarantees a sufficiently stable attachment, because these bonds are about ten times stronger than typical ligand-receptor interactions [16.5]. This chemistry has been used to detect the forces between complementary DNA strands [16.6] and single nucleotides [16.7]. Self-assembled monolayers of dithio-bis(succinimidylundecanoate) were formed to enable covalent coupling of biomolecules via amines [16.8], and were used to study the binding strength between cell adhesion proteoglycans [16.9] and between biotin-directed IgG antibodies and biotin [16.10]. A vectorial orientation of Fab molecules on gold tips was achieved by a firsthand site-directed chemical binding via their SH-groups [16.11], without the need of additional linkers. For this, antibodies were

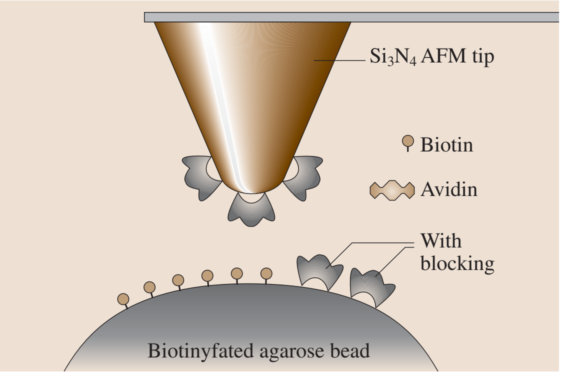

Fig. 16.1 Avidin-functionalized AFM tip. A dense layer of biotinylated BSA was adsorbed to the tip and subsequently saturated with avidin. The biotinylated agarose bead opposing the tip also contained a high surface density of reactive sites, which were, therefore, partly blocked with avidin to achieve single-molecule binding events (after [16.3])

digested with papain and subsequently purified to generate Fab fragments with freely accessible SH-groups in the hinge region.

Gold surfaces provide a unique and selective affinity for thiols, although the adhesion strength of the formed bond is weaker than that of other covalent bonds [16.5]. Since all commercially available AFM tips are etched of silicon nitride or silicon oxide material, deposition of a gold layer onto the tip surface is required prior to using this chemistry. Therefore, designing a sensor with covalent attachments of the biomolecules to the silicon surface may be more straightforward. Amine-functionalization procedures, a strategy widely used in surface bio-chemistry, were applied using ethanolamine [16.4, 12] and various silanization methods [16.13–15] as a first step in thoroughly developed surface anchoring protocols suitable for single molecule experiments. Since the amine density achieved in these procedures often determines the number of ligands on the tip that can specifically bind to the receptors on the surface, it has to be sufficiently low to guarantee single molecule detection [16.4, 12]. One molecule per tip apex ($\sim 5-20$ nm) corresponds to a macroscopic density of about 500 molecules per μm^2. A most striking example of a single ligand molecule tip was realized by gluing a single nanotube to the cantilever, thus minimizing the contact between tip and surface with a chemically well-defined and extremely small tip apex [16.16]. The ligand molecule was reacted to an activated chemical site at the open end of the nanotube.

In a number of laboratories, a distensible and flexible linker was used to space the ligand molecule from the tip surface (e.g., [16.4, 14]) (Fig. 16.2). At a given low number of spacer molecules per tip, the ligand can freely orient and diffuse within a certain volume provided by the length of the tether to achieve unconstrained binding to its receptor. The unbinding process occurs with little torque, and the ligand molecule escapes the danger of being squeezed between the tip and the surface. It also opens the possibility of site-directed coupling for a defined orientation of the ligand relative to the receptor at receptor-ligand unbinding. As a crosslinking element, poly(ethylene glycol) (PEG), a water soluble, nontoxic polymer with a wide range of applications in surface technology and clinical research, was often used [16.17]. PEG is known to prevent surface adsorption of proteins and lipid structures and therefore appears ideally suited for this purpose. Gluteraldehyde [16.13] and DNA [16.6] were also successfully applied in recognition force studies. Crosslinker lengths, ideally arriving at a good compromise between high tip molecule mobility and narrow lateral resolution of the target recognition site, varied from 2 to 100 nm.

For coupling to the tip surface and to the ligands, respectively, the crosslinker is mostly equipped with two different functional ends, e.g., an N-hydroxy-

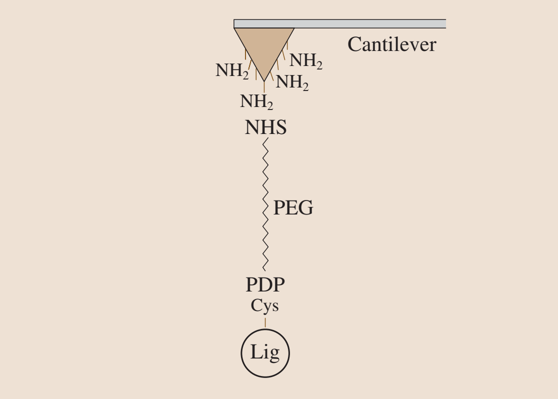

Fig. 16.2 Linkage of ligands to AFM tips. Ligands were covalently coupled to AFM tips via a heterobifunctional polyethylene glycol (PEG) derivative of 8 nm length. Silicon tips were first functionalized with ethanolamine ($NH_2-C_2H_4OH\cdot HCl$). Then, the NHS (N-hydroxysuccinimide) end of the PEG linker was covalently bound to amines on the tip surface before ligands were attached to the PDP (Pyridoyldithiopropionate) end via a free thiol or cysteine

succinimidyl (NHS) residue on the one end is reactive to amines on the tip and a 2-pyridyldithiopropionyl (PDP) [16.18] or a vinyl sulfon [16.19] residue on the other end, can be covalently bound to thiols of ligands (Fig. 16.2). This sulfur chemistry is highly advantageous, since it is very reactive and renders site-directed coupling possible. However, free thiols are hardly available on native ligands and must, therefore, be generated.

Different strategies were used to achieve this goal. Lysins of protein-ligands were derived from the short heterobifunctional linker N-succinnimidyl-3-(S-acethylthio)propionate (SATP) and subsequent deprotection with NH_2OH led to reactive SH groups [16.18]. Since it is very delicate to react distinct lysins with this method, the coupling to the crosslinker is often not specifically site directed. Several protocols are commercially available (Pierce, Rockford, IL) to generate active antibody fragments with free cysteines. Half-antibodies are produced by cleaving the two disulfides in the central region of the heavy chain using 2-mercaptoethylamine HCl [16.20] and Fab fragments are generated from digestion using papain [16.11]. The most elegant methods are to mutate a cysteine into the primary sequence of proteins and to append a thiol to the end of synthesized DNA strands [16.6], because they allow for a defined, sequence-specific coupling of the ligand to the crosslinker.

A nice alternative for a most common, strong but noncovalent site-directed anchor on the spacer has been introduced recently. The binding strength of the NTA (nitrilotriacetate) - His_6 (histidine 6) system, routinely used on chromatographic and biosensor matrices for the binding of recombinant proteins to which a His_6 tag is appended to the primary sequence, was found to be significantly larger than typical values of other ligand-receptor systems [16.21–23]. Therefore, a crosslinker containing an NTA residue is ideally suited for coupling a recombinant ligand carrying a His_6 in its sequence to the AFM tip. This general, site-directed, and oriented coupling strategy also allows rigid and fast control of the specificity of ligand-receptor recognition by using Ni^{++} as a molecular switch of the NTA-His_6 bond.

16.2 Fixation of Receptors to Probe Surfaces

For the recognition by ligands on the AFM tip, receptors should be tightly attached to probe surfaces. Loose receptor fixation could lead to a pull-off of the receptor from the surface by the ligand on the tip, which would consequently block ligand-receptor recognition and obscure the recognition force experiments.

Freshly cleaved muscovite mica is a perfectly pure and atomically flat surface and, therefore, ideally suited for MRFM studies. Moreover, the strong negative charge of mica accomplishes very tight electrostatic binding of certain types of biomolecules. Some receptor proteins such as lysozyme [16.20] or avidin [16.24] strongly adhere to mica due to the strong positive charge of these highly basic proteins at $pH < 8$. In this case, it is safe to purely adsorb the receptors from the solution, since the unspecific attachment to the surface is sufficiently strong for recognition force experiments. Nucleic acids are firmly bound to mica via Zn^{2+}, Ni^{2+}, or Mg^{2+} bridges [16.25]. Electrostatic interaction has also been used to adsorb the strongly acidic sarcoplasmic domain of the Ca^{2+} release channel via Ca^{2+}-bridges to probe the cytoplasmic surface [16.26]. Similarly, protein crystals and bacterial layers have been deposited onto mica in defined orientations with an appropriate choice of buffer conditions [16.27, 28].

In many cases, however, one cannot rely on electrostatic binding toward surfaces. Ideally, water-soluble receptors like globular antigenic proteins or extracellular protein chimeras are then anchored covalently. When glass, silicon, or mica is used as the probe surface, exactly the same surface chemistry described for the silicon AFM tips applies (see above). The number of reactive SiOH groups of the relatively chemically inert mica can be optionally increased by water plasma treatment [16.29]. Frequently, crosslinkers were also used on probe surfaces to provide the receptors with motional freedom, too [16.4]. Protein binding to photo-crosslinkers thereby enables rigid temporal control of the reaction [16.30].

A major limitation of the silicon chemistry is the impossibility of reaching a high surface density of functional sites ($< 1{,}000/\mu m^2$). By comparison, a monolayer of streptavidin would consist of 60,000 molecules per μm^2 and a phospholipid monolayer would consist of 1.7×10^6 molecules per μm^2. The latter density is also achieved by chemisorption of alkanethiols to gold. Tightly bound, ordered, self-assembled monolayers are formed on ultraflat gold surfaces and display excellent probes for AFM [16.10]. Alkanethiols can carry reactive groups at the free end that allow for covalent attachment of biomolecules [16.10, 31] (Fig. 16.3).

A recently developed anchoring strategy to gold uses dithio-phospholipids consisting of a propyldithio group at their hydrophobic end and either a phosphocholine head group (as host lipid) or a phosphethanolamine head group [16.32]. The latter head group is chemically reactive and was derived from a long chain biotin for the molecular recognition of streptavidin molecules in an initial study [16.32]. This phospholipid layer closely mimicks a cell surface and has been optimized by nature to afford little nonspecific adsorption. Additionally, it can be spread as an insoluble monolayer at an air-water interface. Thereby, the ratio of functionalized thio-lipids in host thio-lipids accurately defines the surface density of bio-reactive sites in the monolayer. Subsequent transfer onto gold substrates leads to covalent and, therefore, tight attachment of the monolayer, and thus it is suited for recognition force studies.

Immobilization of cells mainly depends on the type of cell. Cells with adherent growth are readily usable for MRFM, whereas cells that grow in solution without contact to a surface have to be adsorbed. Various protocols for a tight cell anchoring are available. The easiest way is to grow the cells directly on glass or other surfaces in their cell culture medium [16.33]. Various adhesive coatings like Cell-Tak [16.34], gelatin, and poly-lysin increase the strength of adhesion and/or display appropriate surfaces to immobilize spherical cells. Other hydrophic surfaces like gold or carbon are suitable matrices as well [16.35]. Covalent binding of cells to surfaces can be accomplished by using PEG crosslinkers as the one described for the tip chemistry, since they react with free thiols on the cell surface [16.34]. Alternatively, PEG crosslinkers carrying a fatty acid penetrate into the interior of the cell membrane, which guarantees a sufficiently strong fixation without interference with membrane proteins [16.34].

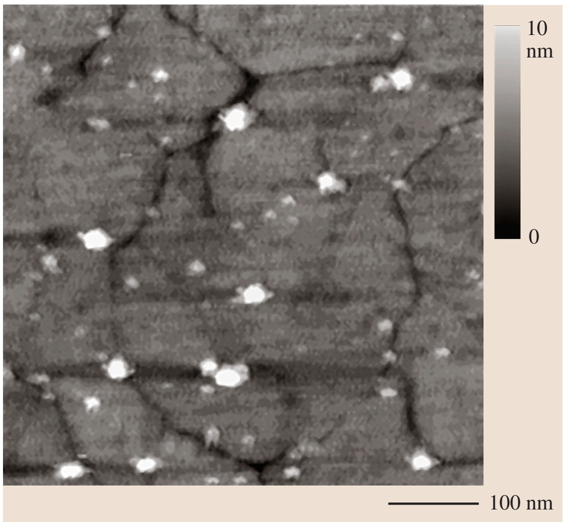

Fig. 16.3 AFM image of hisRNAP molecules specifically bound to nickel-NTA domains on a functionalized gold surface. Alkanethiols terminated with ethylene glycol groups to resist unspecific protein adsorption served as host matrices and were doped with 10% nickel-NTA alkenthiols. The sample was prepared to achieve full monolayer coverage. Ten individual hisRNAP molecules can be clearly seen bound to the surface. The more abundant, smaller, lower features are NTA islands with no bound molecules. The underlying morphology of the gold can also be distinguished (after [16.31])

16.3 Single-Molecule Recognition Force Detection

Quantification of recognition forces goes back to ensemble techniques such as shear flow detachment measurements (SFD) [16.36] and the surface force apparatus (SFA) [16.37]. In SFD, receptors are fixed to a surface to which beads or cells containing ligands attach via their specific bonds. A fluid shear stress is applied to the surface-bound particles by a flow of given velocity that disrupts the ligand-receptor-bonds at a critical force. However, the unbinding force per single bond can only be estimated, because the calculation of the net force on the particle is complicated by the stress distribution and the number of bonds per particle is assumed from the geometry of the contact area.

SFA directly measures the vertical forces between two macroscopic surfaces containing receptors and ligands, respectively, using a spring as a force prober and interferometry for detection. Although SFA does not have a great sensitivity in measuring absolute forces and requires assumptions to be made for down-scaling to

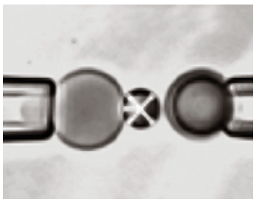

Fig. 16.4 Experimental setup of the biomembrane force probe (BFP). The spring in the BFP is a pressurized membrane capsule. Its spring constant is set by the membrane tension, which is controlled by micropipette suction. The BFP tip, a glass microbead of 1–2 μm diameter, was chemically glued to the membrane. Operated on the stage of an inverted microscope, the BFP (on the *left*) was kept stationary and the microbead test surface (on the *right*) was translated to/from contact with the BFP tip by precision piezo control (after [16.38])

single bonds, its z-resolution is highly accurate (better than 1 Å) over a long range. Adhesion and compression forces, as well as rapid transient effects can be followed in real time.

The bio force probe (BFP) technique (for a recent paper see [16.38]) (Fig. 16.4) uses soft membranes as the force transducer, rather than a mechanical spring. A single cell or giant vesicle is held by suction at the tip of a glass micropipette and its spring constants are adjusted by the aspiration pressure and can be varied over several orders of magnitude. Interaction forces with opposing cells, vesicles, or functionalized beads are obtained from the micromechanical analysis of the deformations of the membranes and/or via interference contrast microscopy. A force range from 0.1 pN to 1 nN is accessible with high force sensitivity down to the single-bond level.

In optical tweezers (OT) small particles are manipulated by optical traps [16.39]. Three-dimensional light intensity gradients of a focused laser beam are used to pull or push the particles against another one. Particle displacements can be followed with nanometer resolution. Movements of single molecular motors [16.40] and force-extension profiles of single extensible molecules such as titin [16.41] and DNA [16.42] were measured. Defined, force-controlled twisting of DNA using rotating, magnetically manipulated particles gave even further insights into DNA viscoelastic properties [16.43].

The atomic force microscope (AFM) is the force measuring method with the smallest-sized force sensor and therefore achieves the highest lateral resolution. Radii of commercially available AFM tips vary between 2 and 50 nm. In contrast, the sensing particles in SFD, BFP, and OT are in the 1 to 10 μm range, and the surfaces used in SFA exceed millimeter extensions. The small apex of the AFM tip allows for the probing of single biomolecules. Originally designed to explore the topography of rigorous surfaces with atomic resolution [16.1] in area scans, the AFM was used in biology to image the secondary structures of nucleotides [16.25], as well as polar lipid head groups and subunits of proteins embedded in membranes in liquid environments [16.27, 28].

The additional potential of the AFM to detect ultralow forces has been successfully applied to the detection of interaction forces of single receptor-ligand pairs [16.2–4]. These forces are measured in so called force-distance cycles using a ligand-carrying tip mounted to a cantilever and a probe surface with receptors firmly attached. A force-distance cycle, where a distensible tether was used to couple ligands to tips and receptors to surfaces is shown in Fig. 16.5 (from [16.4]). At a fixed lateral position the tip approaches the probe surface and is subsequently retracted, and the cantilever deflection Δx is measured in dependence of the tip-surface separation Δz. Cantilever deflections are detected by reflecting a laser beam off the back of the cantilever onto a split photodiode. Bending of the cantilever causes a change in the position where the laser beam strikes the photodiode and, therefore, a change in the output voltage of the photodiode.

The force F acting on the cantilever directly relates to the cantilever deflection Δx according to Hook's law $F = k\Delta x$, where k is the cantilever spring constant. During tip-surface approach (trace, and line 1 to 5), the cantilever deflection remains zero far away from the surface (1 to 4), because there is no detectable tip-surface interaction. Upon tip-surface contact (4), the cantilever bends upward (4 to 5), consistent with a repulsive force that increases linearly with z the harder the tip is pushed into the surface. Subsequent tip-surface retraction (retrace, 5 to 7) first leads to relaxation of the cantilever bending and the repulsive force drops back down to zero (5 to 4).

When ligand-receptor binding has occurred, the cantilever bends downward upon retraction, reflecting an attractive force (retrace, 4 to 7) that increases with increasing tip-surface separation. Since the cognitive molecules were tethered to the surfaces via flexible and distensible crosslinkers, the shape of the attractive force-distance profile is, in contrast to the repulsive force-distance profile of the contact region

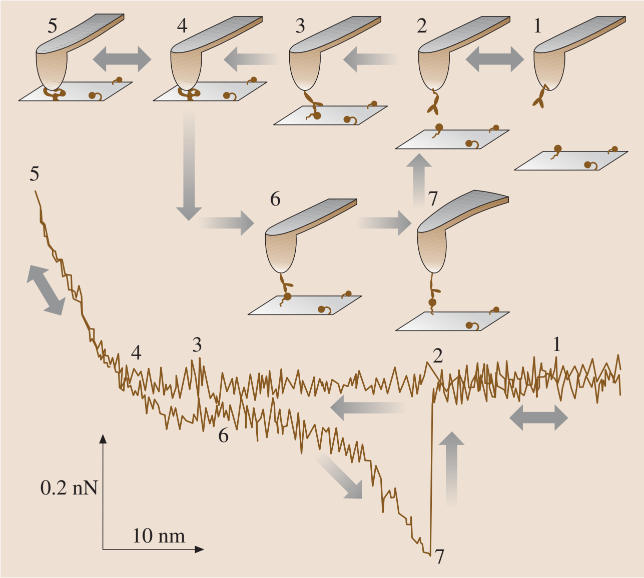

Fig. 16.5 Single-molecule recognition event detected with AFM. Raw data from a force-distance cycle with 100 nm z-amplitude at 1 Hz sweep frequency measured in PBS. Binding of the antibody on the tip to the antigen on the surface during approach (trace points 1 to 5) physically connects tip to probe. This causes a distinct force signal of distinct shape (points 6 to 7) during tip retraction, reflecting extension of the distensible crosslinker-antibody-antigen connection. The force increases until unbinding occurs at an unbinding force of 268 pN (points 7 to 2). (After [16.4])

(4 to 7), nonlinear. Its shape is determined by the elastic properties of the flexible PEG crosslinker [16.17,44] and shows parabolic-like characteristics, which reflects the increase of the spring constant of the crosslinker during extension. The force attractive profile contains the features of a single-molecule recognition event.

The physical connection between the tip and surface sustains the increasing force until the ligand-receptor complex dissociates at a certain critical force (unbinding force, 7), and the cantilever finally jumps back to the resting position (7 to 2). The quantitative force measure of the unbinding force f of a single receptor-ligand pair is directly given by the force at the moment of unbinding (7).

If the ligand on the tip does not form a specific bond with one of the receptors on the probe surface, the recognition event is missing and retrace looks like trace. Also, the specificity of the ligand-receptor binding is usually demonstrated in block experiments. Free ligands are injected in solution so as to block receptor sites on the surfaces. As a further consequence, recognition signals completely disappear, because the receptor sites on the surface are blocked by the ligand of the solution and, thus, prevent recognition by the ligand on the tip.

The force resolution of the AFM is limited by the thermal noise of the cantilever that is determined by its spring constant. According to the equipartition theorem, the cantilever has on average a thermal energy of $0.5\,k_{\mathrm{B}}T$, therefore, $0.5\,k\Delta z^2 = 0.5\,k_{\mathrm{B}}T$, where k_{B} is Boltzmann's constant and T the absolute temperature. Since $F = k\Delta x$, the force sensitivity is given by $\Delta F = (k_{\mathrm{B}}Tk)^{1/2}$ and is, therefore, better the softer the cantilever. The smallest force that can be detected with commercially available cantilevers is in the few picoNewton range. For thinner levers with low spring constants, the force resolution decreases by an order of magnitude [16.45]. Due to their smaller mass, they have a higher resonance frequency (30–100 kHz in liquid). When deflection is recorded at a typical bandwidth of 10 kHz, only a fraction of the thermal noise is included and fast movements of the tip can still be detected [16.45]. An alternative route for improving the force resolution was found by electronically increasing the apparent damping constant, which results in the reduction of thermal noise [16.46].

Besides the detection of intermolecular forces between receptors and ligands, the AFM also shows great potential in measuring intramolecular force profiles of single molecules. The molecule is clamped in between the tip and the probe surface and its visco-elastic properties are studied in force-distance cycles. The detailed force extension profile of the recognition event shown in Fig. 16.5 contains, for example, the elastic features of the crosslinker (PEG in this particular case) [16.44] with which the ligand is tethered to the tip of the AFM.

16.4 Principles of Molecular Recognition Force Spectroscopy

Specific interactions underlying molecular recognition are a distinct class of highly complementary, non-covalent bonds between biological macromolecules composed typically of a few tens of weak interactions such as electrostatic, polar, van der Waals, hydrophobic, and hydrogen bonds. The energy of each single bond is just slightly higher than the thermal energy $k_{\mathrm{B}}T$. Due to the power law dependence of each weak interaction potential and the directionality of the hydrogen bonds, the attractive forces between the receptors and their cognitive ligands are extremely short-range. Therefore, a close geometrical and chemical fit within the binding pocket is a prerequisite for molecular recognition.

Conformational flexibility of at least one of the two binding partners is required for their structural adaptation. The weak interaction bonds are believed to be formed in a spatially and temporarily correlated fashion, thus creating a strong and highly specific molecular interaction. Recognition binding sites are structurally and chemically unique and selective for distinct receptor/ligand combinations. The overall affinity or binding energy of a receptor-ligand bond of a few tens of $k_{\mathrm{B}}T$ is typically sufficient to effect chemical or physical changes of the binding partners, which often initiate biochemical reaction cascades or cellular metabolic pathways.

The binding energy E_{B}, given by the free energy difference between the bound and the free state, is the common parameter to describe the strength of a bond and can be determined in ensemble average calorimetric experiments. E_{B} determines the ratio of bound complexes, [RL], to the product of free reactants, [R] [L], at equilibrium in solution and is related to the equilibrium dissociation constant K_{D}, the affinity parameter, according to $E_{\mathrm{B}}\alpha - \log(K_{\mathrm{D}})$. At thermodynamic equilibrium, the number of complexes that form per unit of time, k_{on} [R] [L], equals the number of complexes that dissociate per time unit, k_{off} [RL]. The empirical kinetic rate constants, on-rate k_{on} and off-rate k_{off}, are related to the equilibrium dissociation constant K_{D} through $K_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}}$.

In order to get an estimate for the interaction forces f from binding energies E_{B}, the dimension of the binding pocket may be used as the characteristic length dimension l; thus $f = E_{\mathrm{B}}/l$. Typical values of $E_{\mathrm{B}} = 20\,k_{\mathrm{B}}T$ and $l = 0.5\,\mathrm{nm}$ yields $f \sim 170\,\mathrm{pN}$ as an order of magnitude feeling for the strength of a single receptor-ligand bond. Classical mechanics describes the force required to separate interacting molecules as the gradient in energy along the interaction potential. The complexed molecules would, therefore, dissociate when force exceeds the steepest gradient in energy. However, activation barriers, temperature, time scales, and the detailed characteristics of the energy landscape, all known to be essential for the understanding of receptor-ligand dissociation, are not considered in this purely mechanical picture.

Ligand-receptor binding is generally a reversible reaction. Viewed on the single-molecule level, the average lifetime of a ligand-receptor bond, $\tau(0)$, is given by the inverse of the kinetic off-rate, k_{off}, so $\tau(0) = k_{\mathrm{off}}^{-1}$. Therefore, ligands will dissociate from receptors without any force applied to the bond at times larger than $\tau(0)$ (covering a range from milliseconds to days for different receptor-ligand combinations). In contrast, if molecules are pulled faster than $\tau(0)$, the bond will resist and require a force for detachment. The size of the unbinding force may vary over more than one order of magnitude, and even exceed the adiabatic limit given by the steepest energy gradient of the interaction potential if bond breakage occurs faster than diffusive relaxation (nanosecond range for biomolecules in viscous aqueous medium) and friction effects become dominant [16.47]. Therefore, unbinding forces do not resemble unitary values, and the dynamics of the experiment defines the physics underlying the unbinding process.

At the millisecond to second time scale of AFM experiments, thermal impulses govern the unbinding process. In the thermal activation model, the lifetime of a complex in solution is described by a Boltzmann ansatz, $\tau(0) = \tau_{\mathrm{osc}} \exp(E_{\mathrm{b}}/k_{\mathrm{B}}T)$ [16.48], where τ_{osc} is the inverse of the natural oscillation frequency and E_{b} the energy barrier for dissociation. Hence, due to the thermal energy there is a finite probability of overcoming the energy barrier E_{b}, which leads to the separation of the receptor-ligand complex. The mean complex lifetime in solution is, therefore, determined by the ratio of E_{b} over the thermal energy.

A force acting on a binding complex deforms the interaction energy landscape and lowers the activation energy barrier, very similar to the mode of operation of an enzyme (Fig. 16.6). Thus the input of thermal energy by forces acting in this time regime reduces the bond lifetime. The lifetime $\tau(f)$ of a bond loaded with a constant force f is given as $\tau(f) = \tau_{\mathrm{osc}} \exp[(E_{\mathrm{b}} - x_{\beta} f)/k_{\mathrm{B}}T]$ [16.48], x_{β} being interpreted as the distance of the energy barrier E_{b} from the energy minimum along the direction of the

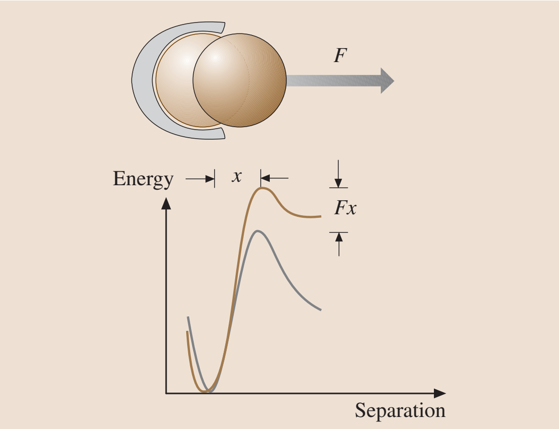

Fig. 16.6 Dissociation over a single sharp energy barrier. Under a constant force, the barrier is linearly decreased by the applied force F, giving rise to a characteristic length scale x that is interpreted as the distance of the energy barrier from the energy minimum along the projection of the force (after [16.50])

applied force. In this model, the energy barrier decreases linearly with the force applied. A detailed physical basis for Bell's theory has been derived in [16.49], where the strength of weak non-covalent bonds was studied in liquid using Kramer's theory for reaction kinetics in liquids under the influence of force.

The lifetime $\tau(f)$ under constant force f compares to the lifetime at zero force, $\tau(0)$, according to $\tau(f) = \tau(0)\exp(-x_\beta f/k_\mathrm{B}T)$ [16.4] for a single sharp energy barrier for which a mono-exponential dependence is characteristic. Using AFM, receptor-ligand unbinding is commonly measured in force-distance cycles (cf. previous chapter), during which the applied force to the complex does not remain constant. Rather, it increases in a complex fashion at a nonlinear rate determined by the pulling velocity, the spring constant of the cantilever, and the force-distance profile of the complexed biomolecules. The main contribution of the thermal activation, however, comes from the part of the force curve, which is close to unbinding. Therefore, an effective force increase or loading rate r can be deduced from $r = \mathrm{d}f/\mathrm{d}t =$ pulling velocity times effective spring constant at the end of the force curve just before unbinding occurs. The approximation of a constant loading rate using force-distance cycles with AFM is sound with a properly defined effective loading rate [16.51].

The dependence of the unbinding force on a linear loading rate in the thermally activated regime was first derived in [16.49] and further described in [16.50]. Force-induced dissociation of receptor-ligand complexes using AFM or BFP can be regarded as an irreversible process, because the two binding partners are further separated after dissociation has occurred. Therefore, rebinding and back-reactions are negligible. The unbinding process itself is of a stochastic nature, and the likelihood of bond survival is expressed in the master equation as time-dependent probability $N(t)$ to be in the bound state under a linearly increasing force $f = rt$, thus $\mathrm{d}N(t) = -k_\mathrm{off}(rt)N(t)$ [16.50].

Rewriting Bell's formula using $\tau(f) = k_\mathrm{off}^{-1}(f)$ relates the kinetic off-rate at a given force f, $k_\mathrm{off}(f)$, to the off-rate at zero force, $k_\mathrm{off}(0)$: $k_\mathrm{off}(f) = k_\mathrm{off}(0)\exp(l_\mathrm{r} f/k_\mathrm{B}T)$. The master equation combined with Bell's formula results in a spectrum of Gaussian-like distributions of unbinding forces (frequency of occurrence versus force) parameterized by the loading rate r [16.50]. Thus the maximum of each force distribution, $f^*(r)$, reflects the most probable force of unbinding for the respective loading rate r. f^* is related to r through $f^* = f_\beta \log_\mathrm{e}(r k_\mathrm{off}^{-1}/f_\beta)$, where the slope is governed by f_β, a force scale set by the ratio of the thermal energy $k_\mathrm{B}T$ to the length scale x_β, which marks the thermally averaged position of the energy barrier along the direction of the force [16.49, 50]. Apparently, the unbinding force f^* scales linearly with the logarithm of the loading rate. For a single barrier, this would give rise to a simple, linear dependence of the force versus log loading rate. In cases where more barriers are involved along the escape path, the curve will follow a sequence of linear regimes, each of which marks a particular barrier [16.38]. Hence, transition from one regime to the other is associated with an abrupt change of slope determined by the characteristic barrier length scale.

In force spectroscopy experiments, the dynamics of pulling on specific receptor-ligand bonds is varied, which leads to detailed structural and kinetic information of the bond breakage, as discussed above and shown in the following chapter on explicit examples. Length scales of energy barriers are obtained from the slope of the spectroscopy plot (force versus loading rate) and their relative heights may be gained from the force shifts [16.38]. Extrapolation to zero forces yields the kinetic off-rate constant for the dissociation of the complex in solution [16.52].

16.5 Recognition Force Spectroscopy: From Isolated Molecules to Biological Membranes

In the early pioneering AFM studies receptor-ligand bond strengths were only measured at one loading rate. Therefore, only a single point in the spectrum of forces, that depend on the dynamics of the experiment, was determined. Avidin–biotin is often regarded as the prototype of receptor-ligand interaction due to its enormously high affinity ($K_\mathrm{D} = 10^{-13}\,\mathrm{M}$) and long bond lifetime ($\tau(0) = 80\,\mathrm{days}$). Hence, it may not be too surprising that the first realizations of single-molecule recognition force detections were made with biotin and its cognitive receptors, streptavidin [16.2], the succinilated form of avidin, and avidin [16.3]. Unbinding forces of 250–300 pN and 160 pN were found for streptavidin and avidin, respectively. It was also shown [16.53] that the forces vary with the spring constant of the cantilever using the same pulling velocity, which is consistent with the loading rate dependency of the unbinding force discussed above. The recognition forces of various biotin analogs with avidin and streptavidin [16.53], as well as with site-directed streptavidin mutants [16.54] were related to their energies and found to be correlated with the equilibrium binding enthalpy [16.53] and the enthalpic activation barrier [16.54] but independent of free energy parameters. It was suggested that internal energies of the bond breakage, rather than entropic changes were probed by the force measurements [16.54].

16.5.1 Forces, Energies, and Kinetic Rates

In first measurements of inter- and intramolecular forces of the DNA double helix [16.6], complementary single-strand DNA oligonucleotides covalently attached to the AFM tip and the probe surface, respectively, were probed. The multimodal force distributions obtained (three distinct, separated force peaks) were associated with interchain interactions of single pairs of 12, 16, and 20 base sequence length. A long strain single-strand DNA with cohesive ends was sandwiched in between tip and substrate and its intramolecular force pattern revealed elastic properties comparable to polymeric molecules [16.6]. Single molecule interactions between all possible 16 combinations of the four bases were also studied [16.7] and yielded specific binding only when complemetary bases were present on the tip and the probe surface, indicating that AFM is capable of following Watson-Crick base pairing.

Anibody-antigen interactions are of key importance for the function of the immune system. Their affinities are known to vary over orders of magnitudes. For single-molecule recognition AFM studies, the molecules were surface-coupled via flexible and distensible crosslinkers [16.4, 10, 13, 14] to provide them with sufficient motional freedom, so that problems of misorientation and steric hindrance that can obscure specific recognition are avoided. In [16.4, 12], the tips were functionalized with a low antibody density, so that on the average only a single antibody on the tip end could access the antigens on the surface. Hence, isolated single molecular antibody-antigen complexes could be examined. It was observed that the interaction sites of the two Fab fragments of the antibody (the antibody consists of two antigen-active Fab-fragments and one Fc-portion) are able to bind simultaneously and independently with equal binding probability. Single antibody-antigen binding events were also studied with tip-bound antigens [16.10] using an antibody biosensor probe surface [16.13], or single chain antibody-fragments attached to gold via freely accessible sulfhidryl groups [16.14]. In the latter study, the unbinding forces of two mutant single-chain antibodies, varying by an order of magnitude in their affinity, were compared. The mutant of ten times lower affinity (consistent with 12% lower free energy due to the logarithmic dependence of the affinity on free energy) showed 20% lower unbinding force.

All these studies describe detection of intermolecular forces between single receptors and their cognitive ligands. However, the AFM can also be used to characterize single-molecule intramolecular forces, e.g., interactions responsible for stabilizing the conformations and tertiary structures of polymers and proteins. The force-extension profile (cf. also stretching of PEG in Fig. 16.5) of polysaccharide dextran revealed multiple elastic regimes dominated by entropic effects, internal bond angle twistings, and conformational changes [16.55, 56]. The complex quantities of elastic and viscous force contributions upon stretching a single dextran molecule were resolved by resonant energy tracking using dynamic force spectroscopy [16.57]. Titin, a fibrillose muscle protein, contains a large number of repetitive IgG domains that shows a characteristic sawtooth pattern in the force profile, reflecting their sequential unfolding (Fig. 16.7) [16.58]. Unfolding and refolding kinetics and energetics of proteins

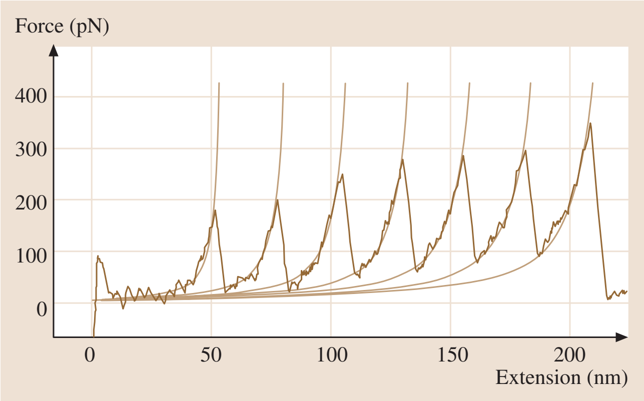

Fig. 16.7 Force extension profile of a single titin molecule. The periodic sawtooth-like pattern reflects domain unfolding in the peptide chain consistent with its modular construction and can be explained as a stepwise increase of the molecule's length upon each unfolding event (after [16.58])

were studied in pull-hold-release cycles [16.59] by carefully handling the molecule between tip and surface. These studies yield crucial information on the relation between structure and function of single molecules.

Besides the investigation of forces, single-molecule recognition force microscopy studies allow for the estimation of association and dissociation rates [16.4, 12, 22, 52, 60, 61], energies [16.38], and structural parameters of the binding pocket [16.4, 12, 15, 60, 61]. For single molecule studies, quantification of the on-rate constant k_{on} for the association of the ligand on the tip to a receptor on the surface requires determination of the interaction time $t_{0.5}$ needed for half-maximal probability of binding. This value is obtained from an experiment in which the encounter duration of the ligand-coated tip on the receptor-containing surface is varied until a range of the unbinding activity from zero to saturation is achieved [16.60]. With the knowledge of the effective ligand concentration c_{eff} on the tip available for receptor interaction, k_{on} is given by $k_{\text{on}} = t_{0.5}^{-1} c_{\text{eff}}^{-1}$. The effective concentration c_{eff} is described by the effective volume V_{eff} the tip-tethered ligand diffuses about the tip, which yields $c_{\text{eff}} = N_{\text{A}}^{-1} V_{\text{eff}}^{-1}$, where N_{A} is the Avogadro's number. V_{eff} is essentially a half-sphere with a radius of the effective tether length, and the kinetic on-rate can, therefore, be calculated from $k_{\text{on}} = t_{0.5}^{-1} N_{\text{A}} 2\pi/3r^3$. However, since k_{on} critically depends on the tether length r, only order of magnitude estimates are gained [16.60].

Additional information about the unbinding process is contained in the unbinding force distributions. Not only the maximum f^*, but also its intrinsic width σ increases with increasing loading rate [16.22, 38]. Apparently, at slower loading rates the systems adjust closer to equilibrium, which leads to smaller values of both the force f^* and its variation σ. This is another indication that dynamics is an important issue when studying recognition forces. An effective lifetime $\tau(f)$ of the bond under an applied force f was estimated by the time the cantilever spends in the force window spanned by the standard deviation σ of the f distribution [16.4]. The time the force increases from $f^* - \sigma$ to $f^* + \sigma$ is thus given by $\tau(f) \approx 2\sigma/\mathrm{d}f/\mathrm{d}t$ [16.4]. In an example of the Ni^{2+}-chelating ligand/receptor interaction nitrotrilacetate/hexa-histidin [16.22], the lifetime $\tau(f)$ decreased with increasing pulling force f from 17 ms at 150 pN to 2.5 ms at 194 pN. The data were fitted with Bell's formula, confirming the theoretically predicted exponential lifetime-force relation for the reduction of the lifetime $\tau(f)$ by the applied force f and, thus yielding a lifetime at zero force, $\tau_0 = 15$ s. Direct measurements of lifetimes are only possible by using a force clamp, where a constant adjustable force is applied to the complex via a feedback loop and the time duration of bond survival is detected. This configuration has been developed recently to study the force dependence of the unfolding probability of the filament muscle protein titin [16.62].

Theory predicts (cf. previous chapter) that the unbinding force of specific and reversible ligand-receptor bonds is dependent on the rate of the increasing force [16.47, 49, 63] during force-distance cycles. Indeed, in experiments unbinding forces were found to assume not a unitary value, but rather were dependent on both the pulling velocity [16.52, 60] and the cantilever spring constant [16.2]. The theoretical findings were confirmed by experimental studies and revealed a logarithmic dependence of the unbinding force on the loading rate [16.15, 22, 38, 52, 60, 61], consistent with the exponential lifetime-force relation described above. On a half-logarithmic plot, a single linear slope is characteristic for a single energy barrier probed in the thermally activated regime [16.38]. The same relation was also found for unfolding forces of proteins [16.55]. The slope of the force spectroscopy curve contains the length scale of the energy barrier, which may be connected to the length dimension of the interaction binding pocket [16.60]. The kinetic off-rate k_{off} at zero force is gained from the extrapolation to zero force [16.52, 60]. Combining k_{off} with k_{on} leads to a value for the equilibrium dissociation constant K_{D}, according to $K_{\mathrm{D}} = k_{\mathrm{off}}/k_{\mathrm{on}}$.

A simple correlation between unbinding force and solution kinetics was found for nine different single-chain Fv fragments constructed from point mutations of three unrelated anti-fluorescein antibodies [16.61]. These molecules provide a good model system because they are monovalent and follow a 1:1 binding model without allosteric effects, and the ligand (fluoerscein) is inert and rigid. For each mutant, the affinity K_{D}, the kinetic rate constants k_{on} and k_{off}, and the temperature dependence of k_{off} yielding the activation barrier E_{b}, were determined in solution. In a series of single-molecule AFM recognition events, the loading rate (r) dependence of the unbinding forces (f) was measured. The maxima of the force distributions f^* scaled linearly with $\log r$. Therefore, unbinding was dictated by a single energy barrier along the reaction path (Fig. 16.8). Since the measured unbinding forces of all mutants precisely correlated with k_{off}, the unbinding path followed and the transition state crossed cannot be too dissimilar in spontaneous and forced unbinding. Interestingly, the distance of the energy barrier along the unbinding pathway determined from the force spectroscopy curve was found to be proportional to the barrier height and, thus, most likely includes elastic stretching of the antibody construct during the unbinding process.

16.5.2 Complex Bonds and Energy Landscapes

More complex force spectra were found from unbinding force measurements between complementary DNA strands of different lengths [16.15]. DNA duplexes with 10, 20, and 30 base pairs, were investigated to reveal unbinding forces from 20 to 50 pN of single duplexes, depending on the loading rate and sequence length. The unbinding force of the respective duplexes scaled

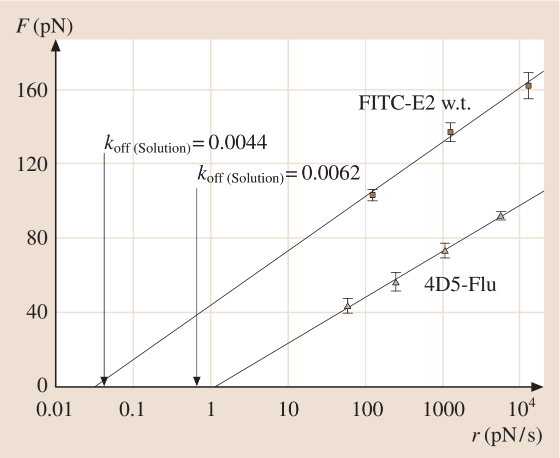

Fig. 16.8 Loading rate dependence of the unbinding force dependence of two anti-fluoresceins. For both FITC-E2 w.t. and 4D5-Flu, a strictly mono-logarithmic dependence was found in the range accessed, indicating that only a single energy barrier was probed. The same energy barrier dominates dissociation without forces applied, because extrapolation to zero force matches kinetic off-rates determined in solution (indicated by *arrow*) (after [16.61])

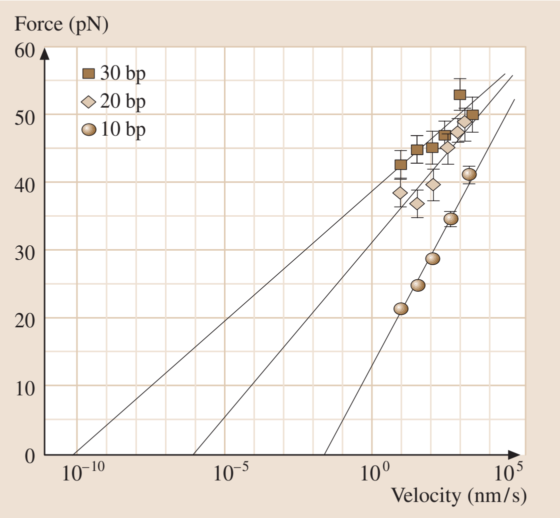

Fig. 16.9 Dependence of the unbinding force between DNA single-strand duplexes on the retract velocity. Besides the mono-logarithmic behavior, the unbinding force scales with the length of the strands, increasing from the 10- to 20- to 30-base-pair duplexes (after [16.15])

with the logarithm of the loading rate, however, each of them with a different slope (Fig. 16.9). The distance of the energy barrier along the separation path and the logarithm of the kinetic off-rate, both obtained from the force spectroscopy plot as parameters describing the energy landscape, were found to be proportional to the number of DNA complexes per duplex. This defined scaling leads to a model of fully cooperative bonds in a series for the unbinding of the base pairs in the duplex, described by one distinct length and time scale for each duplex length, where each single bond contributes a defined increment [16.15]. A temperature study revealed that entropic contributions also play a role in the unbinding between complementary DNA strands [16.64].

Energy landscapes of receptor-ligand bonds in biology can be rugged terrains with more than one prominent energy barrier. The inner barriers are undetectable in solution assays, but can become predominant for certain time scales in force-induced unbinding. Therefore, force spectroscopy appears to be a proper tool to discover hidden activation barriers and explore the energy landscape in greater detail. In initial AFM studies [16.53, 54], it was concluded that the interaction strength of biotin-streptavidin is in the range of 200–300 pN and the strength of biotin-avidin is somewhat lower (~ 150 pN). However, using BFP, the loading rate was varied over eight orders of magnitude (from 10^{-2} to 10^6 pN/s) and gained a detailed picture of the force spectroscopy curve with forces increasing from 5 to 200 pN (Fig. 16.10) [16.38].

Distinct linear regimes that demonstrate the thermally activated nature of the bond breakage are visible, abrupt changes in slope imply a number of sharp barriers along the dissociation pathway. Above 85 pN, there is a common high strength regime for both biotin-streptavidin and biotin-avidin with a slope of $f_\beta = 34$ pN, locating a barrier deep in the binding pocket at $x_\beta = 0.12$ nm. Below 85 pN, the slopes for biotin-streptavidin and biotin-avidin are different. $f_\beta = 8$ pN for biotin-strepavidin marks a barrier at $x_\beta = 0.5$ nm, whereas the steeper slope $f_\beta = 13–14$ pN between 38 and 85 pN in the biotin-avidin spectrum indicates that its next barrier is located at 0.3 nm. Below 11 pN, the biotin-avidin spectrum exhibits a very low strength regime with a slope of $f_\beta = 1.4$ pN that maps to $x_\beta = 3$ nm. In addition to the map of barrier locations, the logarithmic intercepts found by extrapolation of each linear regime to zero force also yield estimates of the energy differences between the activation barriers (Fig. 16.10).

A meticulous picture of the biotin-avidin bond rupture process with amino acid resolution was obtained

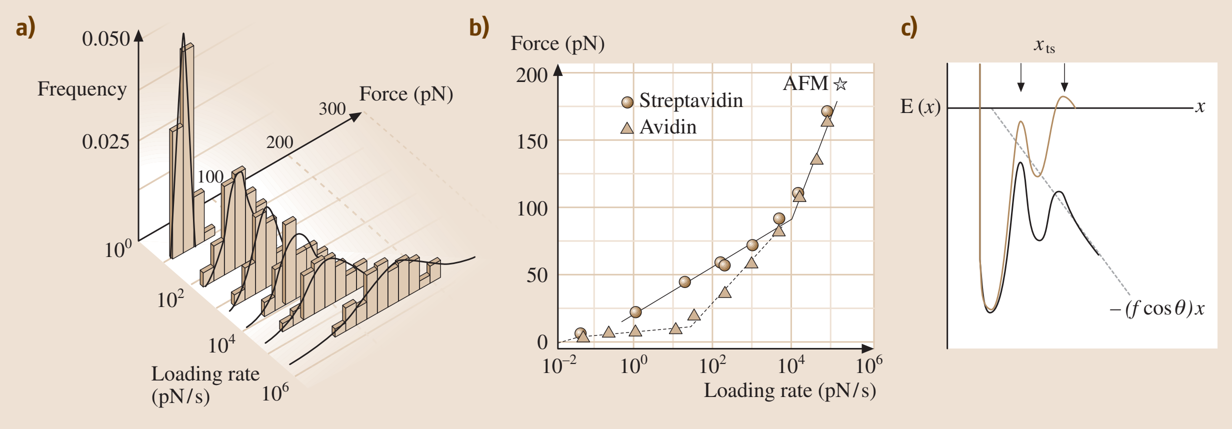

Fig. 16.10a–c Unbinding force distributions and energy landscape of a complex molecular bond. **(a)** Force histograms of single biotin-streptavidin bonds recorded at different loading rates. The shift in peak location and the increase in width with increasing loading rate is clearly demonstrated. **(b)** Dynamic force spectra for biotin-streptavidin (*circles*) and biotin-avidin (*triangles*). The slopes of the linear regimes mark distinct activation barriers along the direction of force. **(c)** Conceptual energy landscape traversed along a reaction coordinate under force under an angle θ to the molecular coordinate. The external force f adds a mechanical potential that tilts the energy landscape and lowers the barriers. The inner barrier starts to dominate when the outer has fallen below it due to the applied force (after [16.38])

with molecular dynamics simulation (MDS) [16.47,63]. Although the time scales of MDS for AFM pulling experiments are orders of magnitudes faster, forces in laboratory speeds were approximated. By reconstructing the interaction potential, the experimentally found transition states were readily identified. Simulations of antibody-antigen unbinding, however, revealed a large heterogeneity of enforced unbinding pathways and a correspondingly large flexibility of the binding pocket region, which exhibited significant induced fit motions [16.65].

The velocity dependence of rupture forces and adhesion probability of the receptor/ligand system P-selectin/P-selectin glycoprotein ligand-1 (PSGL-1) and their relation to kinetic and thermodynamic parameters obtained in ensemble average assays were analyzed in detail in [16.52]. The adhesive interactions of this system play a central immunogenic role in directing the PSGLT-1-containing leukocytes out of the blood stream and into sites of inflammation on cell layers via P-selectin. The high molecular elasticity of the P-selectin/PSGL-1 complex was determined from the force extension profile gained in force-distance cycles and described by the freely jointed chain, a polymer model that yields the Contour length (overall maximum extension), the persistence length (effective segment dimension), and the molecular spring constant. Rupture forces of single complex unbinding increased logarithmically with the loading rate (spring constant times pulling velocity), as expected from theory, and yielded $k_{\text{off}} = 0.02\,\text{s}^{-1}$ and $l = 0.25\,\text{nm}$.

Another characteristic of the P-selectin/PSGL-1 relevant to its biological function was found by investigating the adhesion probability, defined as the ratio between the force-distance cycles showing an adhesion event and the total number of cycles, on the pulling velocity. Counterintuitively and in contrast to experiments with antibody-antigen [16.4], cell adhesion proteins [16.9] and cadherins [16.60], the adhesion probability increased with increasing velocities [16.52]. Since the adhesion probability approaches 1.0 at fast velocities, P-selectin/PSGL-1 complex formation must occur as soon as the tip reaches the surface, thus reflecting a fast kinetic on-rate. Upon the increase in force during pulling, the complexes only dissociate under a measurable force when pulled at fast velocities. No force will be detected at low velocities because the complex has already dissociated by the pure thermal energy and/or a force beneath the noise limit. A quantitative analysis revealed a fast kinetic off-rate of $k_{\text{off}} = 15 \pm 2\,\text{s}^{-1}$. The fast kinetic rates and high chain-like elasticity of the PSGL-1 system guarantee a rapid exchange of the leukocytes on endothelial cell surfaces and therefore provide the basis of the physiological leukocyte *rolling* process [16.52].

PSGL-1 recognition to another member of the selectin family, L-selectin, was studied using BFP [16.66]. The loading rate was varied over three orders of magnitude, thereby gaining rupture forces from a few to 200 pN. Plotted on a logarithmic scale of loading rates, the rupture forces reveal two prominent energy barriers along the unbinding pathway. Strengths above 75 pN arise from rapid detachment ($< 0.01\,\text{s}$) impeded by an inner barrier that requires Ca^{2+}, whereas strengths below 75 pN occur under slow detachment ($> 0.01\,\text{s}$) impeded by an outer barrier, most likely constituted by an array of weak hydrogen bonds. It was speculated that a complex hierarchy of inner and outer barriers is significant in selectin-mediated function [16.66].

Single molecule recognition was also used to investigate the mechanism of the cadherin-mediated adhesion process of vascular endothelial cells. A monolayer of these cells constitutes the major barrier of the body that separates the blood compartment from the extracellular space of tissues. Homophilic cadherin adhesion acting within the cell junctions allows fast dynamic cellular remodeling to change the barrier properties of the cell layers. Recordings of force-distance cycles showed specific recognition events between tip- and surface-bound chimeric VE-cadherin dimers. Measuring the binding activity in dependence of the free Ca^{2+} concentration [16.60] revealed an apparent K_{D} of 1.15 mM with a Hill coefficient of $n_{\text{H}} = 5.04$, indicating high cooperativity and steep dependency. It might be of physiological relevance that a local drop of free Ca^{2+} in the narrow intercellular cleft weakens intercellular adhesion and is, therefore, involved in facilitating cellular remodeling.

The rather low trans-interaction force of a single bond between two cadherin strand dimers was found to be amplified by complexes of cumulative binding strength, visible as distinct force quanta in the adhesion force distributions. These associates were formed by time-controlled lateral and trans-oligomerization, suggesting that cadherins from opposing cells associate to form complexes and, thus, increase the intercellular binding strength. The rather low values for the kinetic rates and the adhesive binding activity determined as described above, make cadherins ideal candidates for adhesion regulation by cytoskeletal tethering, thereby setting the kinetic rate of lateral self-association [16.60].

16.5.3 Live Cells and Membranes

So far, there have been only a few attempts to apply recognition force microscopy to living cells. Recently, *Lehenkari* et al. [16.67] measured the binding forces between integrin receptors in intact cells and Arg-Gly-Asp (RGD) amino acid sequence containing extracellular matrix protein ligands. Ligands, which had a higher predicted affinity for the receptor, gave greater unbinding forces, and the amino acid sequence/pH/divalent cation dependency of the receptor-ligand interaction examined by AFM was similar to the one observed under bulk measurement conditions using other techniques. In another study [16.68], the AFM was used to detect the adhesive strength between concavalin A (con A) coupled to an AFM tip and Con A receptors on the surface of NIH3T3 fibroblast cells. Crosslinking of receptors led to an increase in adhesion that could be attributed to enhanced cooperativity among adhesion complexes, which has been proposed as a physiological mechanism for modulating cell adhesion.

The sidedness and accessability of protein epitopes of the Na^{2+}/D-glucose co-transporter in intact brush border membrane vesicles was sensed with an AFM tip containing an antibody directed against an a sequence close to the glucose binding site [16.35]. Orientations and conformations of the transporter molecule were found to change during glucose binding and transmembrane transport, as evident from the observation of three distinct states discernable in interaction strength and binding activity. Force spectroscopy experiments between leukocyte function-associated antigen-1 (LFA-1) on the tip and intercellular adhesion molecule-1 (ICAM-1) expressed on cell surfaces [16.69] revealed a bilinear behavior in the force vs. log loading rate curve, marking a steep inner activation barrier and a wide outer activation barrier. Since addition of Mg^{2+}, a co-factor that stabilizes the LFA-1/ICAM-1 interaction, elevated the unbinding force of the complex in the slow loading regime and EDTA suppressed the inner barrier, it was suggested that the equilibrium dissociation constant is regulated by the energetics of the outer barrier, whereas the inner barrier determines the ability to resist an external force.

Forces about ten times higher (3 nN) were achieved when whole cells were used to modify the tip instead of single molecules, as shown in measurements of inter-cell force detections between trophoplasts and uterine pithelium cells [16.70]. In a similar configuration but with defined control of the interaction, de-adhesion forces at the resolution of individual cell adhesion molecules in cell membranes were detected and the expression of a gene was quantitatively linked to the function of its product in cell adhesion [16.71].

16.6 Recognition Imaging

Besides studying ligand-receptor recognition processes, the identification and localization of receptor binding sites on bio-surfaces such as cells and membranes are of particular interest. For this, high resolution imaging must by combined with force detection, so that binding sites can be assigned to structures.

A method for imaging functional groups using chemical force microscopy was developed in [16.72]. Both tip and surface were chemically modified, and the adhesive interactions between tip and surface directly correlated with friction images of the patterned sample surfaces. Simultaneous information for topography and recognition forces on a µm-sized streptavidin pattern was obtained by lateral force mapping on a µm-sized streptavidin pattern with a biotinylated AFM tip. An approach-retract cycle was performed in every pixel of the image, and the information for topography was gained from the contact region, whereas a recognition image was constructed from the recognition forces of the force curve [16.72]. This method was also used to map functional receptors on living cells [16.73] and differentiate red blood cells of different blood groups within a mixed population using AFM tips conjugated with a blood group A specific lectin [16.74].

A first identification and localization of single-molecule antigenic sites was realized by laterally scanning an AFM tip containing antibodies in one dimension across a surface containing a low density of antigens during recording force-distance cycles. The binding probability was determined in dependence of the lateral position, resulting in Gaussian-like profiles [16.4, 12]. The position of the antigen could be determined from the position of the maximum, yielding a positional accuracy of 1.5 nm. The width of the binding profile of 6 nm reflects the dynamical reach of the antibody on the tip, provided by the crosslinker used for coupling the antibody to the tip, and can be considered the lateral resolution obtained by this technique [16.4]. With a similar configuration, height and adhesion force images were simultaneously obtained with resolution approaching the single-molecule level as well [16.75].

The strategies of force mapping, however, lack high lateral resolution [16.72] and/or are much slower in data acquisition [16.4,12,75] than topography imaging, since the frequency of the retract-approach cycles performed in every pixel is limited by hydrodynamic forces in the aqueous solution. In addition, obtaining the force image requires the ligand to be disrupted from the receptor in each retract-approach cycle. For this, the z-amplitude of the retract-approach cycle must be at least 50 nm and, therefore, the ligand on the tip is without access to the receptor on the surface for most of the experiment. The development of shorter cantilevers [16.45], however, increases the speed for force mapping, because the hydrodynamic forces are significantly reduced and the resonance frequency is higher than that of commercially available cantilevers. The short cantilevers were also recently applied to follow the association and dissociation of individual chaperonin proteins GroES to GroEL in real time using dynamic force microscopy topography imaging [16.76].

An imaging method for the mapping of antigenic sites on surfaces was recently developed [16.20] by combining molecular recognition force spectroscopy [16.4] with dynamic force microscopy (DFM) [16.25, 77]. In DFM, the AFM tip is oscillated across a surface and the amplitude reduction arising from tip-surface interactions is held constant by a feedback loop that lifts or lowers the tip according to the detected amplitude signal. Since the tip only intermittently contacts the surface, this technique provides very gentle tip-surface interactions, and the specific interaction of the antibody on the tip with the antigen on the surface can be used to localize antigenic sites for recording recognition images. The principles of this method are described in detail below and in Fig. 16.11. A magnetically coated tip is oscillated by an alternating magnetic field at an ampli-

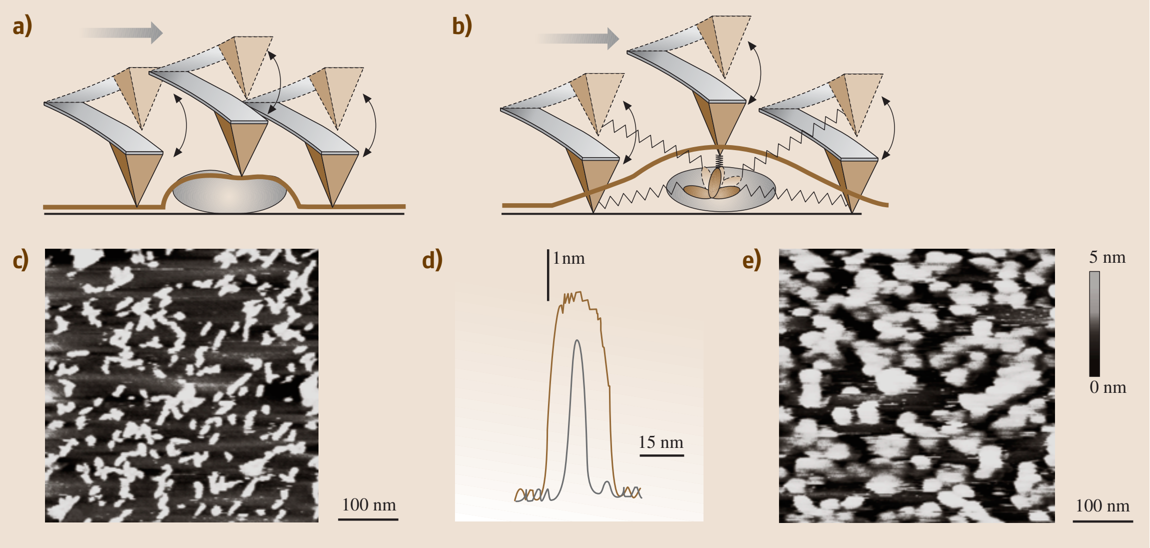

Fig. 16.11a–e Topography and recognition image. **(a)** AFM tip-lysozyme interaction during topography imaging. The *brown line* indicates the height profile obtained from a single lysozyme molecule with the AFM using a bare tip. **(b)** AFM tip-lysozyme interaction during recognition imaging. Half-antibodies are bound to the AFM tip via a flexible tether (*jagged line*) for the recognition of lysozyme on the surface. Imaging results in a height profile, as indicated (*brown line*). **(c)** Topography image. Single lysozyme molecules can be clearly resolved. Sometimes small lysozyme aggregates are observed. Image size was 500 nm. False color bar for heights from 0 (*dark*) to 5 (*bright*) nm. **(d)** Height profiles. Cross-sectional profiles of single lysozyme molecules obtained from the topography (*black line*) and the recognition (*brown line*) image. **(e)** Recognition image. The *bright dots* represent recognition profiles of single lysozyme molecules. Imaging was performed using an AFM tip carrying one half-antibody with access to the antigens on the surface. Conditions were exactly the same as in **(c)**. (After [16.20])

tude of 5 nm while being scanned along the surface. Since the tether has a length of 8 nm, the antibody on the tip always has a chance of recognition when passing an antigenic site, which increases the binding probability enormously. Since the oscillation frequency is more than 100 times faster than typical frequencies in conventional force mapping, the data acquisition rate is much higher.

Half-antibodies were used to provide a monovalent ligand on the tip. The antigen, lysozyme, was tightly adsorbed to mica in conditions that yielded a low surface coverage (for details see [16.20]). A topographic image of this preparation was first recorded in buffer using a bare AFM tip as a control (Fig. 16.11c). Single lysozyme molecules were clearly resolved (Fig. 16.11c). A cross-sectional analysis (Fig. 16.11d, profile in black) reveals that the molecules appear 8 to 12 nm in diameter and 2.0 to 2.5 nm in height. Imaging with a half-antibody tethered to the tip in conditions identical to those used to obtain the topographical image gave strikingly different images (Fig. 16.11e). They differ significantly both in height and diameter compared to the topography image (Fig. 16.11c). Cross-sectional analysis (Fig. 16.11d, trace in red) reveals a height of 3.0 to 3.5 nm and a diameter of 20 to 25 nm. Therefore, profiles obtained from the recognition image appear at least 1 nm higher and 10 nm broader than profiles from the topography image.

The antibody-antigen recognition process during imaging is depicted in Fig. 16.11b. Approaching the antigen in a lateral scan from the left, the antibody on the tip binds to the antigen about 10 nm before the tip end is above the antigenic site (Fig. 16.11b, left tip), due to the flexible tethering provided by the crosslinker. In the bound state, the z-oscillation of the cantilever is further reduced by the attractive force of the crosslinker-antibody-antigen connection, which is acting as a nonlinear spring. Since the AFM detects the z-projection of the force, the amount of the attractive force measured increases when the tip moves further to the right and reaches its maximum just above the position of the antigenic site (Fig. 16.11b, tip in middle). At lateral distances comparable to the length of the antibody-crosslinker connection, the antibody on the tip dissociates from the antibody on the surface and the attractive force goes to zero.

The diameter of cross-sectional profiles obtained from the recognition image (Fig. 16.11d, brown trace) corresponds to about twice the length of the crosslinker (6 nm) plus antibody (6 nm). Increased heights detected in comparison to profiles of the topography image (Fig. 16.11d, black trace) reflect the amplitude reduction owing to antibody-antigen recognition. The maxima of the cross-sectional profiles of the recognition image, as shown in Fig. 16.11d (brown trace), indicate the position of the antigenic site. The accuracy of maximum determination, which in turn reflects the positional accuracy of determining the position of the antigenic site, was 3 nm. The specific nature of the antibody-antigen interaction was tested by injecting free antibody into the liquid cell, so as to block the antigenic sites on the surface, and subsequent images showed a reduction in apparent height.

With this methodology, topography and recognition images can be obtained at the same time and distinct receptor sites in the recognition image can be assigned to structures from the topography image. The described methodology [16.20] is applicable to any ligand and, therefore, it should prove possible to recognize many types of proteins or protein layers and carry out epitope mapping on a nanometer scale on membranes and cells.

16.7 Concluding Remarks

Atomic force microscopy has evolved into the imaging method that yields the greatest structural details on live, biological samples like proteins, nucleotides, membranes and cells in their native, aqueous environment at ambient conditions. Due to its high lateral resolution and sensitive force detection capability, the exciting possibility of measuring inter- and intramolecular forces of biomolecules on the single-molecule level has become possible. The proof-of-principle stage of the pioneering experiments has already developed into a high-end analysis method for exploring kinetic and structural details of interactions underlying protein folding and molecular recognition by taking the advantages of single molecule studies. The information obtained from force spectroscopy includes physical parameters unachievable by other methods and opens new perspectives in exploring the regulation of the dynamics of biological processes such as molecular recognition. With ongoing instrumental developments to increase sensitivity and speed, exciting new fields may also open up in drug and cell screening technologies.

References

16.1 G. Binnig, C. F. Quate, Ch. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

16.2 G. U. Lee, D. A. Kidwell, R. J. Colton: Sensing discrete streptavidin-biotin interactions with atomic force microscopy, Langmuir **10** (1994) 354–357

16.3 E. L. Florin, V. T. Moy, H. E. Gaub: Adhesion forces between individual ligand receptor pairs, Science **264** (1994) 415–417

16.4 P. Hinterdorfer, W. Baumgartner, H. J. Gruber, K. Schilcher, H. Schindler: Detection and localization of individual antibody-antigen recognition events by atomic force microscopy, Proc. Nat. Acad. Sci. USA **93** (1996) 3477–3481

16.5 M. Grandbois, W. Dettmann, M. Benoit, H. E. Gaub: How strong is a covalent bond?, Science **283** (1999) 1727–1730

16.6 G. U. Lee, A. C. Chrisey, J. C. Colton: Direct measurement of the forces between complementary strands of DNA, Science **266** (1994) 771–773

16.7 T. Boland, B. D. Ratner: Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy, Proc. Nat. Acad. Sci. USA **92** (1995) 5297–5301

16.8 P. Wagner, M. Hegner, P. Kernen, F. Zaugg, G. Semenza: Covalent immobilization of native biomolecules onto Au(111) via N-hydroxysuccinimide ester functionalized self assembled monolayers for scanning probe microscopy, Biophys. J **70** (1996) 2052–2066

16.9 U. Dammer, O. Popescu, P. Wagner, D. Anselmetti, H.-J. Güntherodt, G. M. Misevic: Binding strength between cell adhesion proteoglycans measured by atomic force microscopy, Science **267** (1995) 1173–1175

16.10 U. Dammer, M. Hegner, D. Anselmetti, P. Wagner, M. Dreier, W. Huber, H.-J. Güntherodt: Specific antigen/antibody interactions measured by force microscopy, Biophys. J. **70** (1996) 2437–2441

16.11 Y. Harada, M. Kuroda, A. Ishida: Specific and quantized antibody-antigen interaction by atomic force microscopy, Langmuir **16** (2000) 708–715

16.12 P. Hinterdorfer, K. Schilcher, W. Baumgartner, H. J. Gruber, H. Schindler: A mechanistic study of the dissociation of individual antibody-antigen pairs by atomic force microscopy, Nanobiology **4** (1998) 39–50

16.13 S. Allen, X. Chen, J. Davies, M. C. Davies, A. C. Dawkes, J. C. Edwards, C. J. Roberts, J. Sefton, S. J. B. Tendler, P. M. Williams: Spatial mapping of specific molecular recognition sites by atomic force microscopy, Biochemistry **36** (1997) 7457–7463

16.14 R. Ros, F. Schwesinger, D. Anselmetti, M. Kubon, R. Schäfer, A. Plückthun, L. Tiefenauer: Antigen binding forces of individually addressed single-chain Fv antibody molecules, Proc. Nat. Acad. Sci. USA **95** (1998) 7402–7405

16.15 T. Strunz, K. Oroszlan, R. Schäfer, H.-J. Güntherodt: Dynamic force spectroscopy of single DNA molecules, Proc. Nat. Acad. Sci. USA **96** (1999) 11 277–11 282

16.16 S. S. Wong, E. Joselevich, A. T. Woolley, C. L. Cheung, C. M. Lieber: Covalently functionalyzed nanotubes as nanometre-sized probes in chemistry and biology, Nature **394** (1998) 52–55

16.17 P. Hinterdorfer, F. Kienberger, A. Raab, H. J. Gruber, W. Baumgartner, G. Kada, C. Riener, S. Wielert-Badt, C. Borken, H. Schindler: Poly-(ethylene glycol): an ideal spacer for molecular recognition force microscopy/spectroscopy, Single Mol. **1** (2000) 99–103

16.18 Th. Haselgrübler, A. Amerstorfer, H. Schindler, H. J. Gruber: Synthesis and applications of a new poly(ethylene glycol) derivative for the crosslinking of amines with thiols, Bioconj. Chem. **6** (1995) 242–248

16.19 C. K. Riener, G. Kada, C. Borken, F. Kienberger, P. Hinterdorfer, H. Schindler, G. J. Schütz, T. Schmidt, C. D. Hahn, H. J. Gruber: Bioconjugation for biospecific detection of single molecules in atomic force microscopy (AFM) and in single dye tracing (SDT), Recent Res. Devel. Bioconj. Chem. **1** (2002) 133–149

16.20 A. Raab, W. Han, D. Badt, S. J. Smith-Gill, S. M. Lindsay, H. Schindler, P. Hinterdorfer: Antibody recognition imaging by force microscopy, Nature Biotech. **17** (1999) 902–905

16.21 M. Conti, G. Falini, B. Samori: How strong is the coordination bond between a histidine tag and Ni-Nitriloacetate? An experiment of mechanochemistry on single molecules, Angew. Chem. **112** (2000) 221–224

16.22 F. Kienberger, G. Kada, H. J. Gruber, V. Ph. Pastushenko, C. Riener, M. Trieb, H.-G. Knaus, H. Schindler, P. Hinterdorfer: Recognition force spectroscopy studies of the NTA-His6 bond, Single Mol. **1** (2000) 59–65

16.23 L. Schmitt, M. Ludwig, H. E. Gaub, R. Tampe: A metal-chelating microscopy tip as a new toolbox for single-molecule experiments by atomic force microscopy, Biophys. J. **78** (2000) 3275–3285

16.24 C. Yuan, A. Chen, P. Kolb, V. T. Moy: Energy landscape of avidin-biotin complexes measured by atomic force microscopy, Biochemistry **39** (2000) 10 219–10 223

16.25 W. Han, S. M. Lindsay, M. Dlakic, R. E. Harrington: Kinked DNA, Nature **386** (1997) 563

16.26 G. Kada, L. Blaney, L. H. Jeyakumar, F. Kienberger, V. Ph. Pastushenko, S. Fleischer, H. Schindler,

F. A. Lai, P. Hinterdorfer: Recognition force microscopy/spectroscopy of ion channels: applications to the skeletal muscle Ca^{2+} release channel (RYR1), Ultramicroscopy **86** (2001) 129–137

16.27 D. J. Müller, W. Baumeister, A. Engel: Controlled unzipping of a bacterial surface layer atomic force microscopy, Proc. Nat. Acad. Sci. USA **96** (1999) 13170–13174

16.28 F. Oesterhelt, D. Oesterhelt, M. Pfeiffer, A. Engle, H. E. Gaub, D. J. Müller: Unfolding pathways of individual bacteriorhodopsins, Science **288** (2000) 143–146

16.29 E. Kiss, C.-G. Gölander: Chemical derivatization of muscovite mica surfaces, Coll. Surf. **49** (1990) 335–342

16.30 S. Karrasch, M. Dolder, F. Schabert, J. Ramsden, A. Engel: Covalent binding of biological samples to solid supports for scanning probe microscopy in buffer solution, Biophys. J. **65** (1993) 2437–2446

16.31 N. H. Thomson, B. L. Smith, N. Almquist, L. Schmitt, M. Kashlev, E. T. Kool, P. K. Hansma: Oriented, active *escherichia coli* RNA polymerase: an atomic force microscopy study, Biophys. J. **76** (1999) 1024–1033

16.32 G. Kada, C. K. Riener, P. Hinterdorfer, F. Kienberger, C. M. Stroh, H. J. Gruber: Dithio-phospholipids for biospecific immobilization of proteins on gold surfaces, Single Mol. **3** (2002) 119–125

16.33 C. Le Grimellec, E. Lesniewska, M. C. Giocondi, E. Finot, V. Vie, J. P. Goudonnet: Imaging of the surface of living cells by low-force contact-mode atomic force microscopy, Biophys. J. **75**(2) (1998) 695–703

16.34 K. Schilcher, P. Hinterdorfer, H. J. Gruber, H. Schindler: A non-invasive method for the tight anchoring of cells for scanning force microscopy, Cell. Biol. Int. **21** (1997) 769–778

16.35 S. Wielert-Badt, P. Hinterdorfer, H. J. Gruber, J.-T. Lin, D. Badt, H. Schindler, R. K.-H. Kinne: Single molecule recognition of protein binding epitopes in brush border membranes by force microscopy, Biophys. J. **82** (2002) 2767–2774

16.36 P. Bongrand, C. Capo, J.-L. Mege, A.-M. Benoliel: Use of hydrodynamic flows to study cell adhesion. In: *Physical Basis of Cell Adhesion*, ed. by P. Bongrand (CRC Press, Boca Raton 1988) pp. 125–156

16.37 J. N. Israelachvili: *Intermolecular and Surface Forces*, 2nd edn. (Academic Press, London 1991)

16.38 R. Merkel, P. Nassoy, A. Leung, K. Ritchie, E. Evans: Energy landscapes of receptor–ligand bonds explored with dynamic force spectroscopy, Nature **397** (1999) 50–53

16.39 A. Askin: Optical trapping and manipulation of neutral particles using lasers, Proc. Nat. Acad. Sci. USA **94** (1997) 4853–4860

16.40 K. Svoboda, C. F. Schmidt, B. J. Schnapp, S. M. Block: Direct observation of kinesin stepping by optical trapping interferometry, Nature **365** (1993) 721–727

16.41 M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, C. Bustamante: Folding-unfolding transitions in single titin molecules characterized with laser tweezers, Science **276** (1997) 1112–1216

16.42 S. Smith, Y. Cui, C. Bustamante: Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules, Science **271** (1996) 795–799

16.43 T. R. Strick, J. F. Allemend, D. Bensimon, A. Bensimon, V. Croquette: The elasticity of a single supercoiled DNA molecule, Biophys. J. **271** (1996) 1835–1837

16.44 F. Kienberger, V. Ph. Pastushenko, G. Kada, H. J. Gruber, C. Riener, H. Schindler, P. Hinterdorfer: Static and dynamical properties of single poly(ethylene glycol) molecules investigated by force spectroscopy, Single Mol. **1** (2000) 123–128

16.45 B. V. Viani, T. E. Schäffer, A. Chand, M. Rief, H. E. Gaub, P. K. Hansma: Small cantilevers for force spectroscopy of single molecules, J. Appl. Phys. **86** (1999) 2258–2262

16.46 S. Liang, D. Medich, D. M. Czajkowsky, S. Sheng, J.-Y. Yuan, Z. Shao: Thermal noise reduction of mechanical oscillators by actively controlled external dissipative forces, Ultramicroscopy **84** (2000) 119–125

16.47 H. Grubmüller, B. Heymann, P. Tavan: Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force, Science **271** (1996) 997–999

16.48 G. I. Bell: Models for the specific adhesion of cells to cells, Science **200** (1978) 618–627

16.49 E. Evans, K. Ritchie: Dynamic strength of molecular adhesion bonds, Biophys. J. **72** (1997) 1541–1555

16.50 T. Strunz, K. Oroszlan, I. Schumakovitch. H.-J. Güntherodt, M. Hegner: Model energy landscapes and the force-induced dissociation of ligand-receptor bonds, Biophys. J. **79** (2000) 1206–1212

16.51 E. Evans, K. Ritchie: Strength of weak bondconnecting flexible polymer chains, Biophys. J. **76** (1999) 2439–2447

16.52 J. Fritz, A. G. Katopidis, F. Kolbinger, D. Anselmetti: Force-mediated kinetics of single P-selectin/ligand complexes observed by atomic force microscopy, Proc. Nat. Acad. Sci. USA **95** (1998) 12 283–12 288

16.53 V. T. Moy, E.-L. Florin, H. E. Gaub: Adhesive forces between ligand and receptor measured by AFM, Science **266** (1994) 257–259

16.54 A. Chilkoti, T. Boland, B. Ratner, P. S. Stayton: The relationship between ligand-binding thermodynamics and protein-ligand interaction forces measured by atomic force microscopy, Biophys. J. **69** (1995) 2125–2130

16.55 M. Rief, F. Oesterhelt, B. Heymann, H. E. Gaub: Single molecule force spectroscopy on polysaccharides by atomic force microscopy, Science **275** (1997) 1295–1297
16.56 P. E. Marzsalek, A. F. Oberhauser, Y.-P. Pang, J. M. Fernandez: Polysaccharide elasticity governed by chair-boat transitions of the glucopyranose ring, Nature **396** (1998) 661–664
16.57 A. D. L. Humphries, J. Tamayo, M. J. Miles: Active quality factor control in liquids for force spectroscopy, Langmuir **16** (2000) 7891–7894
16.58 M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, H. E. Gaub: Reversible unfolding of individual titin immunoglobulin domains by AFM, Science **276** (1997) 1109–1112
16.59 A. F. Oberhauser, P. E. Marzsalek, H. P. Erickson, J. M. Fernandez: The molecular elasticity of the extracellular matrix tenascin, Nature **393** (1998) 181–185
16.60 W. Baumgartner, P. Hinterdorfer, W. Ness, A. Raab, D. Vestweber, H. Schindler, D. Drenckhahn: Cadherin interaction probed by atomic force microscopy, Proc. Nat. Acad. Sci. USA **97** (2000) 4005–4010
16.61 F. Schwesinger, R. Ros, T. Strunz, D. Anselmetti, H.-J. Güntherodt, A. Honegger, L. Jermutus, L. Tiefenauer, A. Plückthun: Unbinding forces of single antibody-antigen complexes correlate with their thermal dissociation rates, Proc. Nat. Acad. Sci. USA **29** (2000) 9972–9977
16.62 A. F. Oberhauser, P. K. Hansma, M. Carrion-Vazquez, J. M. Fernandez: Stepwise unfolding of titin under force-clamp atomic force microscopy, Proc. Nat. Acad. Sci. USA **98** (2001) 468–472
16.63 S. Izraelev, S. Stepaniants, M. Balsera, Y. Oono, K. Schulten: Molecular dynamics study of unbinding of the avidin-biotin complex, Biophys. J. **72** (1997) 1568–1581
16.64 I. Schumakovitch, W. Grange, T. Strunz, P. Bertoncini, H.-J. Güntherodt, M. Hegner: Temperature dependence of unbinding forces between complementary DNA strands, Biophys. J. **82** (2002) 517–521
16.65 B. Heymann, H. Grubmüller: Molecular dynamics force probe simulations of antibody/antigen unbinding: entropic control and non additivity of unbinding forces, Biophys. J. **81** (2001) 1295–1313
16.66 E. Evans, E. Leung, D. Hammer, S. Simon: Chemically distinct transition states govern rapid dissociation of single L-selectin bonds under force, Proc. Nat. Acad. Sci. USA **98** (2001) 3784–3789
16.67 P. P. Lehenkari, M. A. Horton: Single integrin molecule adhesion forces in intact cells measured by atomic force microscopy, Biochem. Biophys. Res. Com. **259** (1999) 645–650
16.68 A. Chen, V. T. Moy: Cross-linking of cell surface receptors enhances cooperativity of molecular adhesion, Biophys. J. **78** (2000) 2814–2820
16.69 X. Zhang, E. Woijcikiewicz, V. T. Moy: Force spectroscopy of the leukocyte function-associated antigen-1/intercellular adhesion molecule-1 interaction, Biophys. J. **83** (2002) 2270–2279
16.70 M. Thie, R. Rospel, W. Dettmann, M. Benoit, M. Ludwig, H. E. Gaub, H. W. Denker: Interactions between trophoblasts and uterine epithelium: monitoring of adhesive forces, Hum. Reprod. **13** (1998) 3211–3219
16.71 M. Benoit, D. Gabriel, G. Gerisch, H. E. Gaub: Discrete interactions in cell adhesion measured by single molecule force spectroscopy, Nature Cell Biol. **2** (2000) 313–317
16.72 M. Ludwig, W. Dettmann, H. E. Gaub: Atomic force microscopy imaging contrast based on molecular recognition, Biophys. J. **72** (1997) 445–448
16.73 P. P. Lehenkari, G. T. Charras, G. T. Nykänen, M. A. Horton: Adapting force microscopy for cell biology, Ultramicroscopy **82** (2000) 289–295
16.74 M. Grandbois, M. Beyer, M. Rief, H. Clausen-Schaumann, H. E. Gaub: Affinity imaging of red blood cells using an atomic force microscope, J. Histochem. Cytochem. **48** (2000) 719–724
16.75 O. H. Willemsen, M. M. E. Snel, K. O. van der Werf, B. G. de Grooth, J. Greve, P. Hinterdorfer, H. J. Gruber, H. Schindler, Y. van Kyook, C. G. Figdor: Simultaneous height and adhesion imaging of antibody antigen interactions by atomic force microscopy, Biophys. J. **57** (1998) 2220–2228
16.76 B. V. Viani, L. I. Pietrasanta, J. B. Thompson, A. Chand, I. C. Gebeshuber, J. H. Kindt, M. Richter, H. G. Hansma, P. K. Hansma: Probing protein-protein interactions in real time, Nature Struct. Biol. **7** (2000) 644–647
16.77 W. Han, S. M. Lindsay, T. Jing: A magnetically driven oscillating probe microscope for operation in liquid, Appl. Phys. Lett. **69** (1996) 1–3

Part C

Nanotribo

Part C Nanotribology and Nanomechanics

17 Micro/Nanotribology and Materials Characterization Studies Using Scanning Probe Microscopy
Bharat Bhushan, Columbus, USA

18 Surface Forces and Nanorheology of Molecularly Thin Films
Marina Ruths, Åbo, Finland
Alan D. Berman, Los Angeles, USA
Jacob N. Israelachvili, Santa Barbara, USA

19 Scanning Probe Studies of Nanoscale Adhesion Between Solids in the Presence of Liquids and Monolayer Films
Robert W. Carpick, Madison, USA
James D. Batteas, Gaithersburg, USA

20 Friction and Wear on the Atomic Scale
Enrico Gnecco, Basel, Switzerland
Roland Bennewitz, Montreal, Canada
Oliver Pfeiffer, Basel, Switzerland
Anisoara Socoliuc, Basel, Switzerland
Ernst Meyer, Basel, Switzerland

21 Nanoscale Mechanical Properties – Measuring Techniques and Applications
Andrzej J. Kulik, Lausanne, Switzerland
András Kis, Lausanne, Switzerland
Gérard Gremaud, Lausanne, Switzerland
Stefan Hengsberger, Fribourg, Switzerland
Philippe K. Zysset, Wien, Austria
Lásló Forró, Lausanne, Switzerland

22 Nanomechanical Properties of Solid Surfaces and Thin Films
Adrian B. Mann, Piscataway, USA

23 Atomistic Computer Simulations of Nanotribology
Martin H. Müser, London, Canada
Mark O. Robbins, Baltimore, USA

24 Mechanics of Biological Nanotechnology
Rob Phillips, Pasadena, USA
Prashant K. Purohit, Pasadena, USA
Jané Kondev, Waltham, USA

25 Mechanical Properties of Nanostructures
Bharat Bhushan, Columbus, USA

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

17. Micro/Nanotribology and Materials Characterization Studies Using Scanning Probe Microscopy

A sharp AFM/FFM tip sliding on a surface simulates just one asperity contact. However, asperities come in all shapes and sizes. The effect of radius of a single asperity (tip) on the friction/adhesion performance can be studied using tips of different radii. AFM/FFM are used to study the various tribological phenomena, which include surface/roughness, adhesion, friction, scratching, wear, indentation, detection of material transfer, and boundary lubrication.

Directionality in the friction is observed on both micro- and macroscales and results from the surface roughness and surface preparation. Microscale friction is generally found to be smaller than the macrofriction, as there is less plowing contribution in microscale measurements. The mechanism of material removal on the microscale is studied. Evolution of wear has also been studied using AFM. Wear is found to be initiated at nanoscratches. For a sliding interface requiring near-zero friction and wear, contact stresses should be below the hardness of the softer material to minimize plastic deformation, and surfaces should be free of nanoscratches. Wear precursors can be detected at early stages of wear by using surface potential measurements. Detection of material transfer on a nanoscale is possible with AFM. In situ surface characterization of local deformation of materials and thin coatings can be carried out using a tensile stage inside an AFM.

Boundary lubrication studies can be conducted using AFM. Chemically bonded lubricant films and self-assembled monolayers are superior in friction and wear resistance. For chemically bonded lubricant films, the adsorption of water, the formation of meniscus and its change during sliding, viscosity, and surface properties play an important role on the friction, adhesion, and durability of these films. For SAMs, their friction mechanism is explained by a so-called "molecular spring" model.

17.1 **Description of AFM/FFM and Various Measurement Techniques** 499
17.1.1 Surface Roughness and Friction Force Measurements ... 500
17.1.2 Adhesion Measurements 502
17.1.3 Scratching, Wear and Fabrication/Machining............ 503
17.1.4 Surface Potential Measurements 503
17.1.5 In Situ Characterization of Local Deformation Studies 504
17.1.6 Nanoindentation Measurements 504
17.1.7 Localized Surface Elasticity and Viscoelasticity Mapping 505
17.1.8 Boundary Lubrication Measurements 507

17.2 **Friction and Adhesion** 507
17.2.1 Atomic-Scale Friction.................. 507
17.2.2 Microscale Friction....................... 507
17.2.3 Directionality Effect on Microfriction 511
17.2.4 Velocity Dependence on Microfriction 513
17.2.5 Effect of Tip Radii and Humidity on Adhesion and Friction 515
17.2.6 Scale Dependence on Friction 518

17.3 **Scratching, Wear, Local Deformation, and Fabrication/Machining**................... 518
17.3.1 Nanoscale Wear 518
17.3.2 Microscale Scratching................... 519
17.3.3 Microscale Wear 520
17.3.4 In Situ Characterization of Local Deformation 524
17.3.5 Nanofabrication/Nanomachining ... 526

17.4 **Indentation** 526
17.4.1 Picoindentation 526
17.4.2 Nanoscale Indentation................. 527
17.4.3 Localized Surface Elasticity and Viscoelasticity Mapping 528

17.5 **Boundary Lubrication** 530
17.5.1 Perfluoropolyether Lubricants........ 530
17.5.2 Self-Assembled Monolayers........... 536
17.5.3 Liquid Film Thickness Measurements 537

17.6 **Closure** 538

References 539

The mechanisms and dynamics of the interactions of two contacting solids during relative motion ranging from atomic- to microscale need to be understood in order to develop a fundamental understanding of adhesion, friction, wear, indentation, and lubrication processes. At most solid-solid interfaces of technological relevance contact occurs at many asperities. Consequently, the importance of investigating single asperity contacts in studies of the fundamental micro/nanomechanical and micro/nanotribological properties of surfaces and interfaces has long been recognized. The recent emergence and proliferation of proximal probes, including scanning probe microscopies (the scanning tunneling microscope and the atomic force microscope), the surface force apparatus, and computational techniques for simulating tip-surface interactions and interfacial properties, have allowed systematic investigations of interfacial problems with high resolution, as well as ways and means of modifying and manipulating nanoscale structures. These advances have led to the appearance of the new field of micro/nanotribology, which pertains to experimental and theoretical investigations of interfacial processes on scales ranging from the atomic- and molecular- to the microscale, occurring during adhesion, friction, scratching, wear, indentation, and thin-film lubrication at sliding surfaces [17.1–12].

Micro/nanotribological studies are needed to develop a fundamental understanding of interfacial phenomena on a small scale and study interfacial phenomena in micro/nanostructures used in magnetic storage systems, micro/nanoelectromechanical systems (MEMS/NEMS), and other applications [17.3, 7–9, 13, 14]. Friction and wear of lightly loaded micro/nanocomponents are highly dependent on the surface interactions (few atomic layers). These structures are generally lubricated with molecularly thin films. Micro/nanotribological studies are also valuable in understanding interfacial phenomena in macrostructures and provide a bridge between science and engineering.

The surface force apparatus (SFA), the scanning tunneling microscopes (STM), atomic force and friction force microscopes (AFM and FFM) are widely used in micro/nanotribological studies. Typical operating parameters are compared in Table 17.1. The SFA was developed in 1968 and is commonly employed to study both static and dynamic properties of molecularly thin films sandwiched between two molecularly smooth surfaces. The STM, developed in 1981, allows imaging of electrically conducting surfaces with atomic resolution and has been used for the imaging of clean surfaces, as well as of lubricant molecules. The introduction of the atomic force microscope in 1985 provided a method for measuring ultra-small forces between a probe tip and an engineering (electrically conducting or insulating) surface and has been used for topographical measurements of surfaces on the nanoscale, as well as for adhesion and electrostatic force measurements. Subsequent modifications of the AFM led to the development of the friction

Table 17.1 Comparison of typical operating parameters in SFA, STM, and AFM/FFM used for micro/nanotribological studies

Operating parameter	SFA	STM[a]	AFM/FFM
Radius of mating surface/tip	~ 10 mm	5–100 nm	5–100 nm
Radius of contact area	10–40 μm	N/A	0.05–0.5 nm
Normal load	10–100 mN	N/A	< 0.1 nN – 500 nN
Sliding velocity	0.001–100 μm/s	0.02–200 μm/s (scan size ~ 1 nm ×1 nm to 125 μm ×125 μm; scan rate < 1–122 Hz)	0.02–200 μm/s (scan size ~ 1 nm ×1 nm to 125 μm ×125 μm; scan rate < 1–122 Hz)
Sample limitations	Typically atomically smooth, optically transparent mica; opaque ceramic, smooth surfaces can also be used	Electrically conducting samples	None

[a] Can only be used for atomic-scale imaging

force microscope (FFM), designed for atomic-scale and microscale studies of friction. This instrument measures forces in the scanning direction. The AFM is also used in investigations of scratching, wear, indentation, detection of transfer of material, boundary lubrication, and fabrication and machining. Meanwhile, significant progress in understanding the fundamental nature of bonding and interactions in materials, combined with advances in computer-based modeling and simulation methods, have allowed theoretical studies of complex interfacial phenomena with high resolution in space and time. Such simulations provide insights into atomic-scale energetics, structure, dynamics, thermodynamics, transport and rheological aspects of tribological processes.

The nature of interactions between two surfaces brought close together and those between two surfaces in contact as they are separated have been studied experimentally with the surface force apparatus. This has led to a basic understanding of the normal forces between surfaces and the way in which these are modified by the presence of a thin liquid or a polymer film. The frictional properties of such systems have been studied by moving the surfaces laterally, and such experiments have provided insights into the molecular-scale operation of lubricants such as thin liquid or polymer films. Complementary to these studies are those in which the AFM or FFM is used to provide a model asperity in contact with a solid or lubricated surface, Fig. 17.1. These experiments have demonstrated that the relationship between friction and surface roughness is not always simple or obvious. AFM studies have also revealed much about the nanoscale nature of intimate contact during wear and indentation.

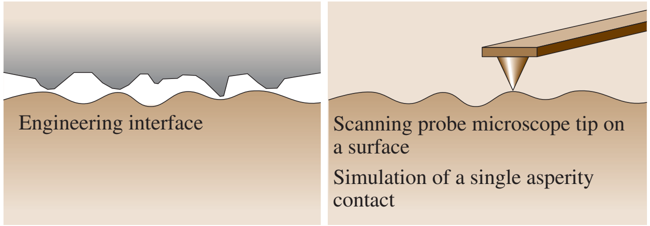
Engineering interface
Scanning probe microscope tip on a surface
Simulation of a single asperity contact

Fig. 17.1 Schematics of an engineering interface and scanning probe microscope tip in contact with an engineering interface

In this chapter, we present a review of significant aspects of micro/nanotribological studies conducted using AFM/FFM.

17.1 Description of AFM/FFM and Various Measurement Techniques

An atomic force microscope (AFM) was developed by *Binnig* et al. in 1985. It is capable of investigating surfaces of scientific and engineering interest on an atomic scale [17.15, 16]. The AFM relies on a scanning technique to produce very high resolution, 3-D images of sample surfaces. It measures ultrasmall forces (less than 1 nN) present between the AFM tip surface mounted on a flexible cantilever beam and a sample surface. These small forces are determined by measuring the motion of a very flexible cantilever beam with an ultra-small mass by a variety of measurement techniques including optical deflection, optical interference, capacitance, and tunneling current. The deflection can be measured to within 0.02 nm, so for a typical cantilever spring constant of 10 N/m, a force as low as 0.2 nN can be detected. To put these numbers in perspective, individual atoms and human hair are typically a fraction of a nanometer and about 75 μm in diameter, respectively, and a drop of water and an eyelash have a mass of about 10 μN and 100 nN, respectively. In the operation of high resolution AFM, the sample is generally scanned, rather than the tip, because any cantilever movement would add vibrations. AFMs are available for measurement of large samples where the tip is scanned and the sample is stationary. To obtain atomic resolution with the AFM, the spring constant of the cantilever should be weaker than the equivalent spring between atoms. A cantilever beam with a spring constant of about 1 N/m or lower is desirable. For high lateral resolution, tips should be as sharp as possible. Tips with a radius ranging from 10–100 nm are commonly available.

A modification to AFM, providing a sensor to measure the lateral force, led to the development of the friction force microscope (FFM), or the lateral force microscope (LFM), designed for atomic-scale and microscale studies of friction [17.3–5, 7, 8, 11, 17–25] and lubrication [17.26–30]. This instrument measures lateral or friction forces (in the plane of sample surface and in the scanning direction). By using a standard or a sharp diamond tip mounted on a stiff cantilever beam, AFM is also used in investigations of scratching and wear [17.6, 9, 11, 22, 31–33], indentation [17.6, 9, 11, 22, 34, 35], and fabrication/machining [17.4, 11, 22].

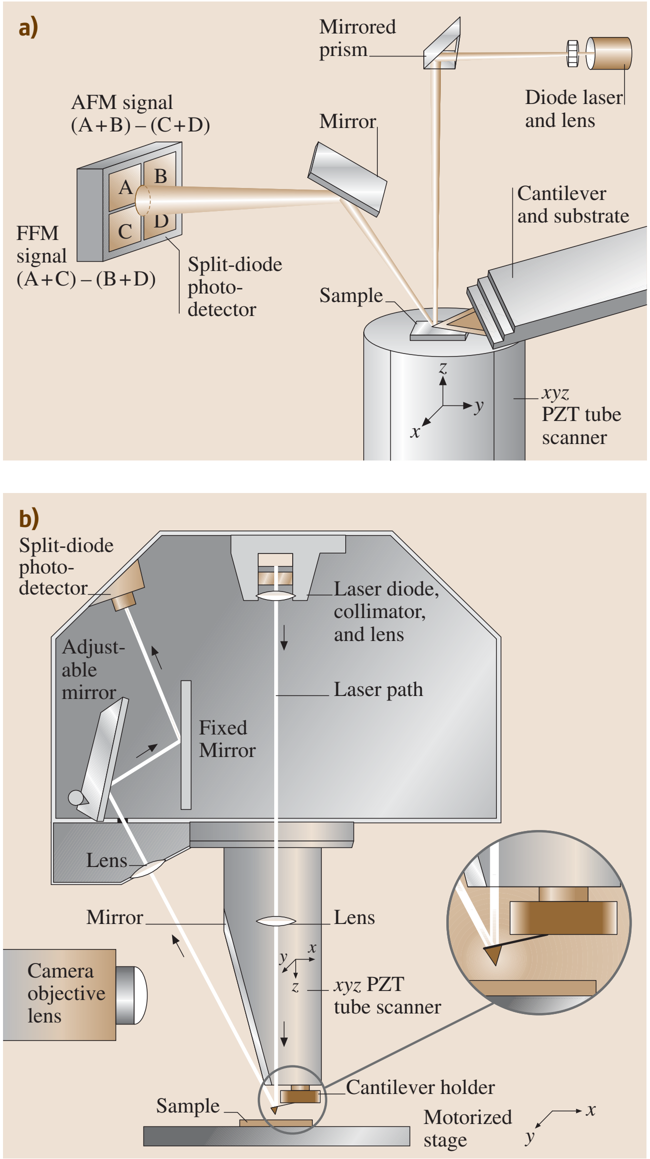
a)
Mirrored prism
Diode laser and lens
AFM signal (A+B) – (C+D)
Mirror
A
B
C
D
FFM signal (A+C) – (B+D)
Split-diode photo-detector
Cantilever and substrate
Sample
z
y
x
xyz PZT tube scanner
b)
Split-diode photo-detector
Laser diode, collimator, and lens
Adjust-able mirror
Laser path
Fixed Mirror
Lens
Mirror
Lens
y
x
z
xyz PZT tube scanner
Camera objective lens
Cantilever holder
Sample
Motorized stage
x
y

Fig. 17.2a,b Schematics **(a)** of a commercial small sample atomic force microscope/friction force microscope (AFM/FFM) and **(b)** a large sample AFM/FFM

Surface roughness, including atomic-scale imaging, is routinely measured using the AFM, Fig. 17.2. Adhesion, friction, wear, and boundary lubrication at the interface between two solids with and without liquid films are studied using AFM and FFM. Nanomechanical properties are also measured using an AFM. In situ surface characterization of local deformation of materials and thin coatings has been carried out by imaging the sample surfaces using an AFM during tensile deformation using a tensile stage.

17.1.1 Surface Roughness and Friction Force Measurements

Simultaneous measurements of surface roughness and friction force can be made with a commercial AFM/FFM. These instruments are available for the measurement of small samples and large samples. In a small sample AFM, shown in Fig. 17.2a, the sample, generally no larger than 10 mm × 10 mm, is mounted on a PZT tube scanner that consists of separate electrodes to scan precisely the sample in the x-y plane in a raster pattern and move it in the vertical (z) direction. A sharp tip at the free end of a flexible cantilever is brought in contact with the sample. Normal and frictional forces being applied at the tip-sample interface are measured using a laser beam deflection technique. A laser beam from a diode laser is directed by a prism onto the back of a cantilever near its free end, which is tilted downward at about 10° with respect to the horizontal plane. The reflected beam from the vertex of the cantilever is directed through a mirror onto a quad photodetector (split photodetector with four quadrants). The differential signal from the top and bottom photodiodes provides the AFM signal, which is a sensitive measure of the cantilever vertical deflection. Topographic features of the sample cause the tip to deflect in the vertical direction as the sample is scanned under the tip. This tip deflection will change the direction of the reflected laser beam, changing the intensity difference between the top and bottom sets of photodetectors (AFM signal). In the AFM operating mode, called the height mode, for topographic imaging, or for any other operation in which the applied normal force is to be kept a constant, a feedback circuit is used to modulate the voltage applied to the PZT scanner to adjust the height of the PZT, so that the cantilever vertical deflection (given by the intensity difference between the top and bottom detector) will remain constant during scanning. The PZT height variation is thus a direct measure of the surface roughness of the sample.

In a large sample AFM both force sensors using optical deflection method and scanning unit are mounted on the microscope head, Fig. 17.2b. Because of vibrations added by cantilever movement, lateral resolution of this design is somewhat poorer than the design in Fig. 17.2a in which the sample is scanned instead of cantilever beam. The advantage of the large sample AFM is that large samples can be measured readily.

Most AFMs can be used for topography measurements in the so-called tapping mode (intermittent contact mode), also referred to as dynamic force microscopy. In the tapping mode, during scanning over the surface, the cantilever/tip assembly is sinusoidally vibrated by a piezo mounted above it, and the oscillating tip slightly taps the surface at the resonant frequency of the cantilever (70–400 Hz) with a constant (20–100 nm) oscillating amplitude introduced in the vertical direction with a feedback loop keeping the average normal force constant, Fig. 17.3. The oscillating amplitude is kept large enough so that the tip does not get stuck to the sample because of adhesive attractions. The tapping mode is used in topography measurements to minimize the effects of friction and other lateral forces and to measure topography of soft surfaces.

For measurement of friction force being applied at the tip surface during sliding, left-hand and right-hand sets of quadrants of the photodetector are used. In the so-called friction mode, the sample is scanned back and forth in a direction orthogonal to the long axis of the cantilever beam. A friction force between the sample and the tip will produce a twisting of the cantilever. As a result, the laser beam will be reflected out of the plane defined by the incident beam and the beam reflected vertically from an untwisted cantilever. This produces an intensity difference of the laser beam received in the left-hand and right-hand sets of quadrants of the photodetector. The intensity difference between the two sets of detectors (FFM signal) is directly related to the degree of twisting and, hence, to the magnitude of the friction force. One problem associated with this method is that any misalignment between the laser beam and the photodetector axis would introduce error in the measurement. However, by following the procedures developed by *Ruan* and *Bhushan* [17.19], in which the average FFM signal for the sample scanned in two opposite directions is subtracted from the friction profiles of each of the two scans, the misalignment effect is eliminated. This method provides three-dimensional maps of friction force. By following the friction force calibration procedures developed by *Ruan* and *Bhushan* [17.19], voltages corresponding to friction forces can be converted to force units. The coefficient of friction is obtained from the slope of friction force data measured as a function of normal loads typically ranging from 10–150 nN. This approach eliminates any contributions due to the adhesive forces [17.22]. For calculation of the coefficient of friction based on a single point measurement, friction force should be divided by the sum of applied normal load and intrinsic adhesive force. Furthermore, it should be pointed out that for a single asperity contact, the coefficient of friction is not independent of load. This will be discussed in greater detail later in the chapter.

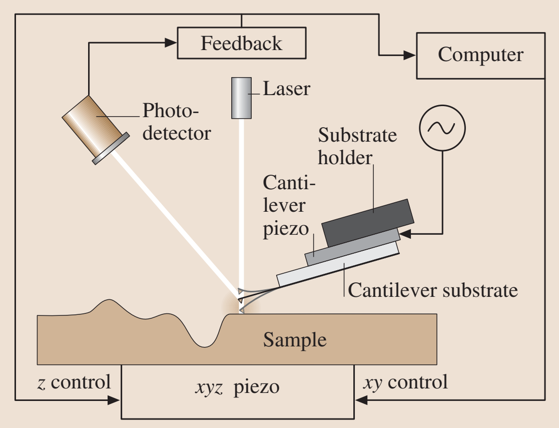

Fig. 17.3 Schematic of tapping mode used for surface topography measurements

Topographic measurements in the contact mode are typically made using a sharp, microfabricated square-pyramidal Si_3N_4 tip with a radius of 30–50 nm on a triangular cantilever beam (Fig. 17.4a) with normal stiffness on the order of 0.5 N/m at a normal load of about 10 nN, and friction measurements are carried out in the load range of 10–150 nN. Topography measurements in the tapping mode utilize a stiff cantilever with high resonant frequency; typically, a square-pyramidal, etched, single-crystal silicon tip with a tip radius of 5–10 nm mounted on a stiff rectangular silicon cantilever beam (Fig. 17.4a) with a normal stiffness on the order of 50 N/m is used. Carbon nanotube tips with small diameters (few nm) and high aspect ratios attached on the single-crystal silicon, square pyramidal tips are used for high resolution imaging of surfaces and deep trenches in the tapping mode or noncontact mode (Fig. 17.4b). To study the effect of the radius of a single asperity (tip) on adhesion and friction, microspheres of silica with radii ranging from about 4–15 µm are attached with epoxy at the ends of tips of Si_3N_4 cantilever beams. Optical micrographs of two of the microspheres mounted at the ends of triangular cantilever beams are shown in Fig. 17.4c.

The tip is scanned in such a way that its trajectory on the sample forms a triangular pattern, Fig. 17.5.

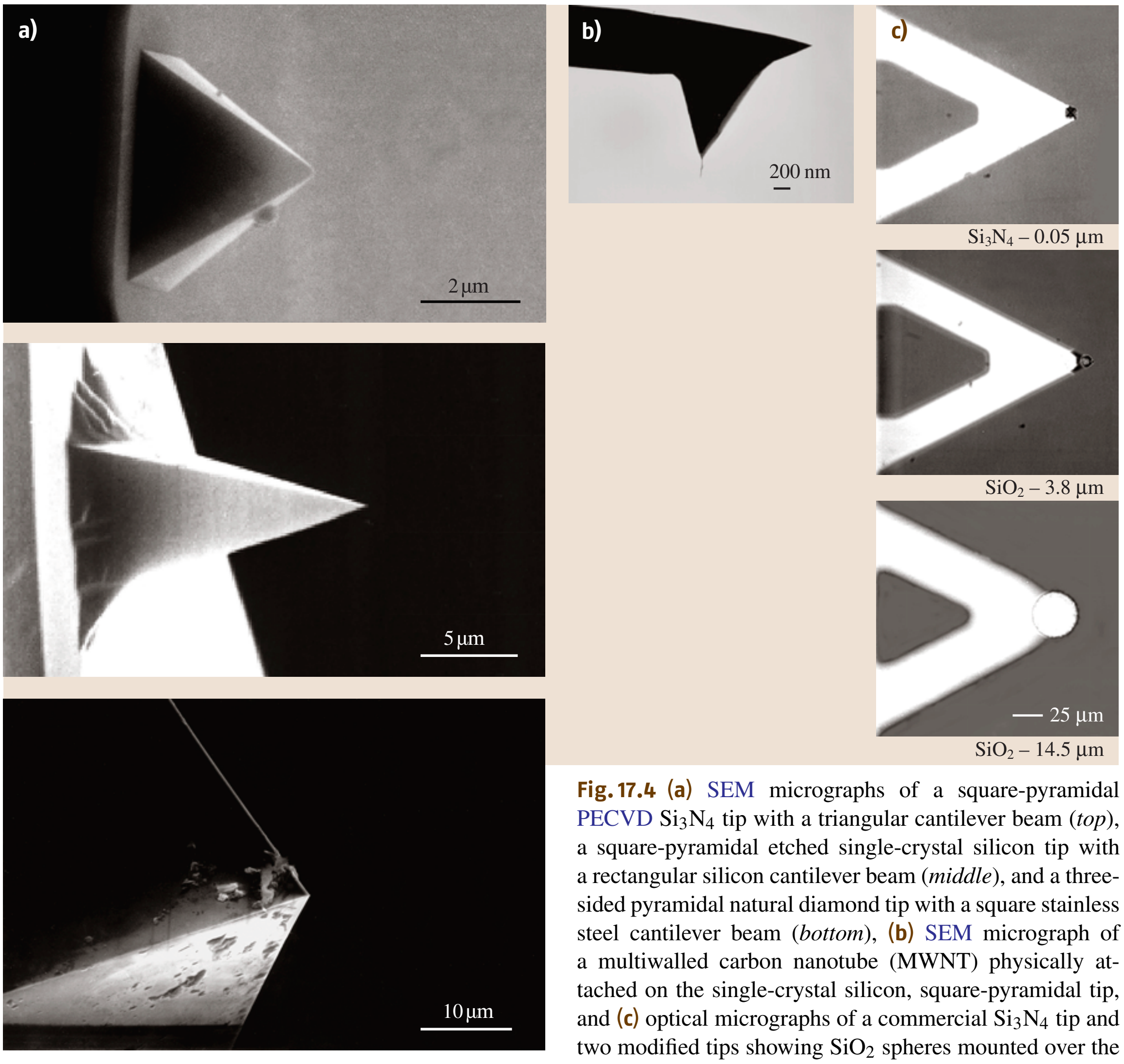

Fig. 17.4 **(a)** SEM micrographs of a square-pyramidal PECVD Si_3N_4 tip with a triangular cantilever beam (*top*), a square-pyramidal etched single-crystal silicon tip with a rectangular silicon cantilever beam (*middle*), and a three-sided pyramidal natural diamond tip with a square stainless steel cantilever beam (*bottom*), **(b)** SEM micrograph of a multiwalled carbon nanotube (MWNT) physically attached on the single-crystal silicon, square-pyramidal tip, and **(c)** optical micrographs of a commercial Si_3N_4 tip and two modified tips showing SiO_2 spheres mounted over the sharp tip and at the end of the triangular Si_3N_4 cantilever beams. (Radii of the tips are given in the figure)

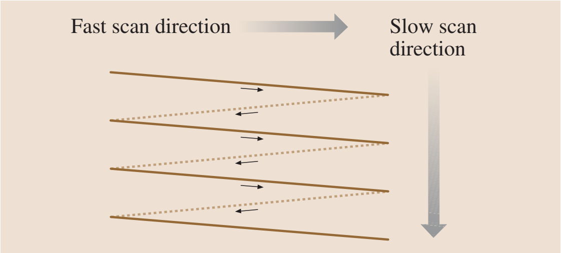

Fig. 17.5 Schematic of triangular pattern trajectory of the tip as the sample (or the tip) is scanned in two dimensions. During scanning, data are recorded only during scans along the solid scan lines

Scanning speeds in the fast and slow scan directions depend on the scan area and scan frequency. Scan sizes ranging from less than 1 nm × 1 nm to 125 µm × 125 µm and scan rates from less than 0.5 to 122 Hz typically can be used. Higher scan rates are used for smaller scan lengths. For example, scan rates in the fast and slow scan directions for an area of 10 µm × 10 µm scanned at 0.5 Hz are 10 µm/s and 20 nm/s, respectively.

17.1.2 Adhesion Measurements

Adhesive force measurements are performed in the so-called force calibration mode. In this mode, force-distance curves are obtained, for an example, see Fig. 17.6. The horizontal axis gives the distance the piezo (and hence the sample) travels, and the vertical axis gives the tip deflection. As the piezo extends, it approaches the tip, which is at this point in free air and hence shows no deflection. This is indicated by the flat portion of the curve. As the tip approaches within a few nanometers of the sample (point A), an attractive force exists between the atoms of the tip surface and the atoms of the sample surface. The tip is pulled toward the sample and contact occurs at point B on the graph. From this point on, the tip is in contact with the surface, and as the piezo extends farther, the tip gets deflected farther. This is represented by the sloped portion of the curve. As the piezo retracts, the tip goes beyond the zero deflection (flat) line, due to attractive forces (van der Waals forces and long-range meniscus forces), into the adhesive regime. At point C in the graph, the tip snaps free of the adhesive forces and is again in free air. The horizontal distance between points B and C along the retrace line gives the distance the tip moved in the adhesive regime. This distance multiplied by the stiffness of the cantilever gives the adhesive force. Incidentally, the horizontal shift between the loading and unloading curves results from the hysteresis in the PZT tube [17.4, 36].

17.1.3 Scratching, Wear and Fabrication/Machining

For microscale scratching, microscale wear, nanofabrication/nanomachining, and nanoindentation hardness measurements, an extremely hard tip is required. A three-sided pyramidal single-crystal natural diamond tip with an apex angle of 80° and a radius of about 100 nm mounted on a stainless steel cantilever beam with normal stiffness of about 25 N/m is used at relatively higher loads (1–150 μN), Fig. 17.4a. For scratching and wear studies, the sample is generally scanned in a direction orthogonal to the long axis of the cantilever beam (typically at a rate of 0.5 Hz), so that friction can be measured during scratching and wear. The tip is mounted on the cantilever such that one of its edges is orthogonal to the long axis of the beam. Therefore, wear during scanning along the beam axis is higher (about two to three times) than during scanning orthogonal to the beam axis. For wear studies, an area on the order of 2 μm × 2 μm is scanned at various normal loads (ranging from 1 to 100 μN) for a selected number of cycles [17.4, 22].

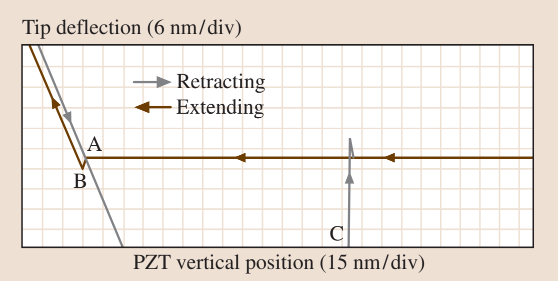

Fig. 17.6 Typical force-distance curve for a contact between Si_3N_4 tip and single-crystal silicon surface in measurements made in the ambient environment. Contact between the tip and silicon occurs at point B; tip breaks free of adhesive forces at point C as the sample moves away from the tip

Scratching can also be performed at ramped loads and the coefficient of friction can be measured during scratching [17.33]. A linear increase in the normal load approximated by a large number of normal load increments of small magnitude is applied using a software interface (lithography module in Nanoscope III) that allows the user to generate controlled movement of the tip with respect to the sample. The friction signal is tapped out of the AFM and is recorded on a computer. A scratch length on the order of 25 μm and a velocity on the order of 0.5 μm/s are used and the number of loading steps is usually 50.

Nanofabrication/nanomachining is conducted by scratching the sample surface with a diamond tip at specified locations and scratching angles. The normal load used for scratching (writing) is on the order of 1–100 μN with a writing speed on the order of 0.1–200 μm/s [17.4, 5, 11, 22, 37].

17.1.4 Surface Potential Measurements

To detect wear precursors and study the early stages of localized wear, the multimode AFM can be used to measure the potential difference between the tip and the sample by applying a DC bias potential and an oscillating (AC) potential to a conducting tip over a grounded substrate in a Kelvin probe microscopy, or so-called "nano-Kelvin probe" technique [17.38–40].

Mapping of the surface potential is made in the so-called "lift mode" (Fig. 17.7). These measurements are made simultaneously with the topography scan in the

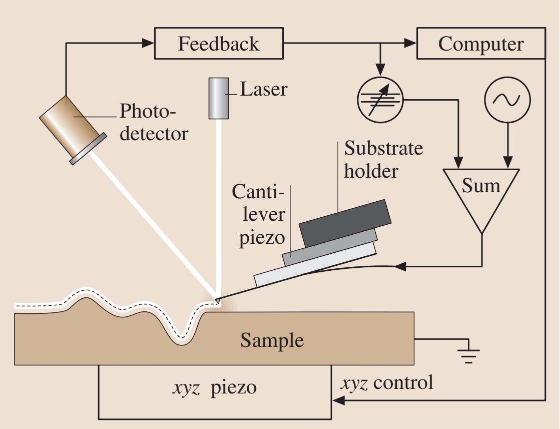

Fig. 17.7 Schematic of lift mode used to make surface potential measurement. The topography is collected in tapping mode in the primary scan. The cantilever piezo is deactivated. Using topography information of the primary scan, the cantilever is scanned across the surface at a constant height above the sample. An oscillating voltage at the resonant frequency is applied to the tip and a feedback loop adjusts the DC bias of the tip to maintain the cantilever amplitude at zero. The output of the feedback loop is recorded by the computer and becomes the surface potential map [17.4]

tapping mode using an electrically conducting (nickel-coated single-crystal silicon) tip. After each line of the topography scan is completed, the feedback loop controlling the vertical piezo is turned off, and the tip is lifted from the surface and traced over the same topography at a constant distance of 100 nm. During the lift mode, a DC bias potential and an oscillating potential (3–7 V) is applied to the tip. The frequency of oscillation is chosen to be equal to the resonant frequency of the cantilever (∼ 80 kHz). When a DC bias potential equal to the negative value of surface potential of the sample (on the order of ±2 V) is applied to the tip, it does not vibrate. During scanning, a difference between the DC bias potential applied to the tip and the potential of the surface will create DC electric fields that interact with the oscillating charges (as a result of the AC potential), causing the cantilever to oscillate at its resonant frequency, as in tapping mode. However, a feedback loop is used to adjust the DC bias on the tip to exactly nullify the electric field and, thus, the vibrations of the cantilever. The required bias voltage follows the localized potential of the surface. The surface potential was obtained by reversing the sign of the bias potential provided by the electronics [17.39, 40]. Surface and subsurface changes of structure and/or chemistry can cause changes in the measured potential of a surface. Thus, the mapping of the surface potential after sliding can be used for detecting wear precursors and studying the early stages of localized wear.

17.1.5 In Situ Characterization of Local Deformation Studies

In situ characterization of local deformation of materials can be carried out by performing tensile, bending, or compression experiments inside an AFM and by observing nanoscale changes during the deformation experiment [17.6]. In these experiments, small deformation stages are used to deform the samples inside an AFM. In tensile testing of the polymeric films carried out by *Bobji* and *Bhushan* [17.41, 42], a tensile stage was used (Fig. 17.8). The stage with a left-right combination lead screw (which helps to move the slider in the opposite direction) was used to stretch the sample and minimize the movement of the scanning area, which was kept close to the center of the tensile specimen. One end of the sample was mounted on the slider via a force sensor to monitor the tensile load. The samples were stretched for various strains using a stepper motor, and the same control area at different strains was imaged. In order to better locate the control area for imaging, a set of four markers was created at the corners of a 30 μm × 30 μm square at the center of the sample by scratching the sample with a sharp silicon tip. The scratching depth was controlled such that it did not affect cracking behavior of the coating. A minimum displacement of 1.6 μm could be obtained. This corresponded to a strain increment of 8×10^{-3}% for a sample length of 38 mm. The maximum travel was about 100 mm. The resolution of the force sensor was 10 mN with a capacity of 45 N. During stretching, a stress-strain curve was obtained during the experiment to study any correlation between the degree of plastic strain and propensity of cracks.

17.1.6 Nanoindentation Measurements

For nanoindentation hardness measurements, the scan size is set to zero, and a normal load is then applied to make the indents using the diamond tip. During this procedure, the tip is continuously pressed against the sample surface for about two seconds at various indentation loads. The sample surface is scanned before and after the scratching, wear, or indentation to obtain the

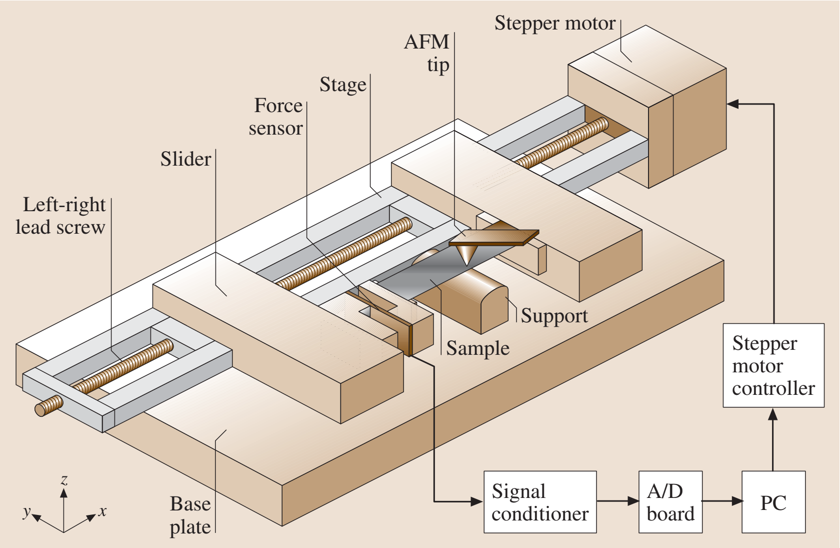

Fig. 17.8 Schematic of the tensile stage to conduct in situ tensile testing of the polymeric films in AFM [17.42]

initial and final surface topographies at a low normal load of about 0.3 µN using the same diamond tip. An area larger than the indentation region is scanned to observe the indentation marks. Nanohardness is calculated by dividing the indentation load by the projected residual area of the indents [17.34].

Direct imaging of the indent allows one to quantify piling up of ductile material around the indenter. However, it becomes difficult to identify the boundary of the indentation mark with great accuracy. This makes the direct measurement of contact area somewhat inaccurate. A technique with the dual capability of depth-sensing, as well as in situ imaging, which is most appropriate in nanomechanical property studies, is used for accurate measurement of hardness with shallow depths [17.4, 35]. This indentation system is used to make load-displacement measurement and, subsequently, carry out in situ imaging of the indent, if required. The indentation system consists of a three-plate transducer with electrostatic actuation hardware used for direct application of normal load and a capacitive sensor used for measurement of vertical displacement. The AFM head is replaced with this transducer assembly while the specimen is mounted on the PZT scanner, which remains stationary during indentation experiments. Indent area and, consequently, hardness value can be obtained from the load-displacement data. The Young's modulus of elasticity is obtained from the slope of the unloading curve.

17.1.7 Localized Surface Elasticity and Viscoelasticity Mapping

Indentation experiments provide a single-point measurement of the Young's modulus of elasticity calculated from the slope of the indentation curve during unloading. Localized surface elasticity maps can be obtained using dynamic force microscopy in which an oscillating tip is scanned over the sample surface in contact under steady and oscillating load. Lower frequency operation mode in the kHz range such as force modulation microscopy [17.43–45] and pulsed force microscopy [17.46] are well suited for soft samples such as polymers. However, if the tip-sample contact stiffness becomes significantly higher than the cantilever stiffness, the sensitivity of these techniques strongly decreases. In this case, sensitivity of the measurement of stiff materials can be improved by using high-frequency operation modes in the MHz range such as acoustic (ultrasonic) force microscopy [17.47]. We only describe here the force modulation technique.

In the force modulation technique, the oscillation is applied to the cantilever substrate with a cantilever piezo (bimorph), Fig. 17.9a. For measurements, an etched silicon tip is first brought in contact with a sample under a static load of 50–300 nN. In addition to the static load applied by the sample piezo, a small oscillating (modulating) load is applied by a bimorph, generally at a frequency (about 8 kHz) far below that of the natural resonance of the cantilever (70–400 kHz). When the

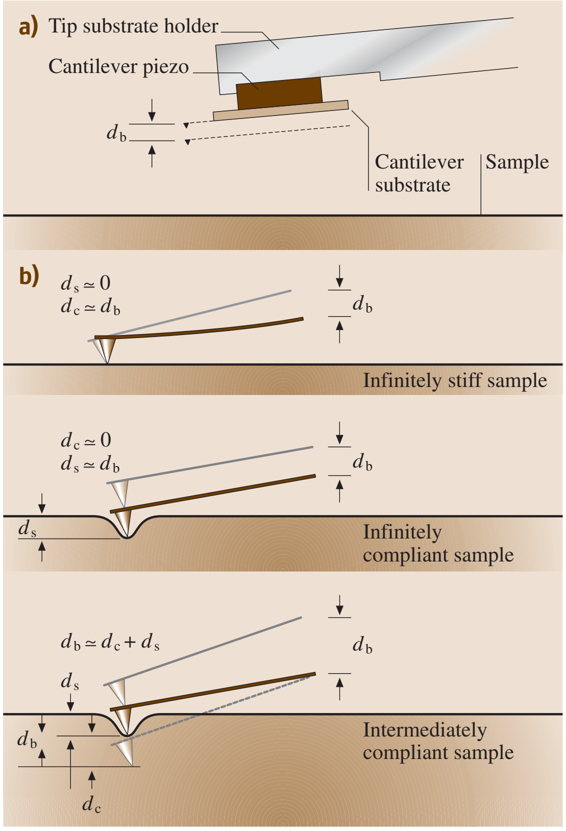
a)
Tip substrate holder
Cantilever piezo
d_b
Cantilever substrate
Sample
b)
$d_s \simeq 0$
$d_c \simeq d_b$
d_b
Infinitely stiff sample
$d_c \simeq 0$
$d_s \simeq d_b$
d_b
d_s
Infinitely compliant sample
$d_b \simeq d_c + d_s$
d_s
d_b
d_b
d_c
Intermediately compliant sample

Fig. 17.9a,b Schematics **(a)** of the bimorph assembly and **(b)** of the motion of the cantilever and tip as a result of the oscillations of the bimorph for an infinitely stiff sample, an infinitely compliant sample, and an intermediate sample. The *thin line* represents the cantilever at the top cycle, and the *thick line* corresponds to the bottom of the cycle. The *dashed line* represents the position of the tip if the sample was not present or was infinitely compliant. d_c, d_s and d_b are the oscillating (AC) deflection amplitude of the cantilever, penetration depth, and oscillating (AC) amplitude of the bimorph, respectively [17.44]

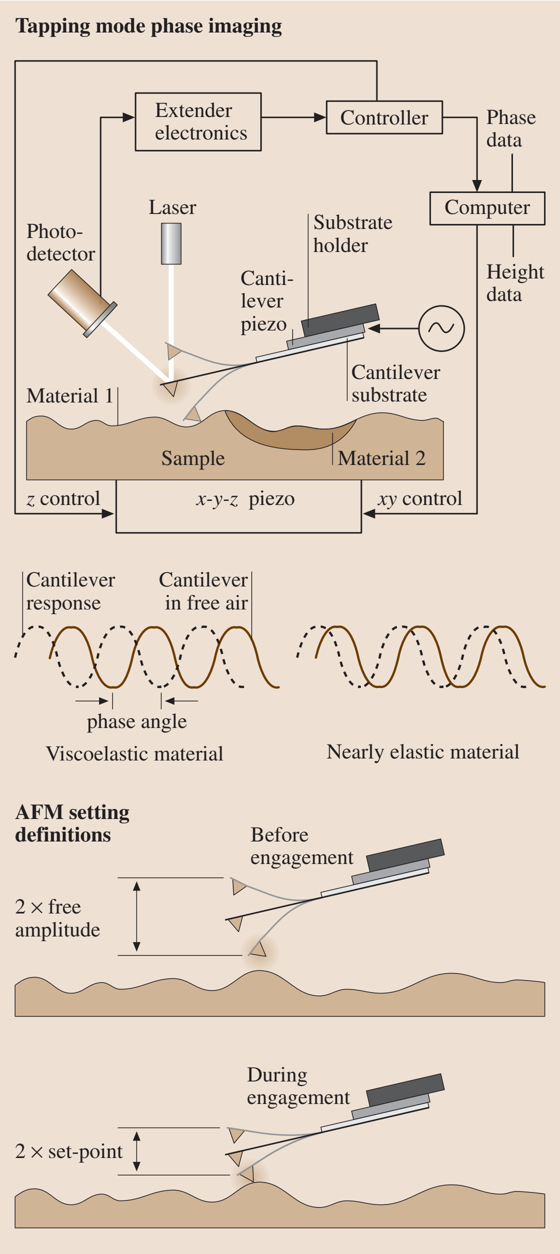
Tapping mode phase imaging
Extender electronics
Controller
Phase data
Computer
Height data
Photo-detector
Laser
Substrate holder
Canti-lever piezo
Cantilever substrate
Material 1
Sample
Material 2
z control
x-y-z piezo
xy control
Cantilever response
Cantilever in free air
phase angle
Viscoelastic material
Nearly elastic material
AFM setting definitions
Before engagement
2 × free amplitude
During engagement
2 × set-point

Fig. 17.10 Schematic of tapping mode used to obtain height and phase data and definitions of tapping amplitude and set point. During scanning the cantilever is vibrated at its resonant frequency and the sample x-y-z piezo is adjusted by feedback control in the z-direction to maintain a constant set point. The computer records height and phase angle (which is a function of the viscoelastic properties of the sample materials) data

tip is brought into contact with the sample, the surface resists the oscillations of the tip, and the cantilever deflects. Under the same applied load, a stiff area on the sample would deform less than a soft one; i. e., stiffer areas cause greater deflection amplitudes of the cantilever, Fig. 17.9b. The variations in the deflection amplitudes provide a measure of the relative stiffness of the surface. Contact analyses can be used to obtain a quantitative measure of localized elasticity of soft surfaces [17.44]. The elasticity data are collected simultaneously with the surface height data using a so-called negative lift mode technique. In this mode, each scan line of each topography image (obtained in tapping mode) is replaced with the tapping action disabled and the tip lowered into steady contact with the surface.

Another form of dynamic force microscopy, phase contrast microscopy, is used to detect the contrast in viscoelastic (viscous energy dissipation) properties of the different materials across the surface [17.48–52]. In this technique, both deflection amplitude and phase angle contrasts are measured, which are the measures of the relative stiffness and viscoelastic properties, respectively. Tapping mode is utilized. For tapping mode during scanning, the cantilever/tip assembly is sinusoidally vibrated at its resonant frequency, and the sample X-Y-Z piezo is adjusted using feedback control in the z-direction to maintain a constant set point. As shown in Fig. 17.10, the cantilever/tip assembly is vibrated at some amplitude, here referred to as the tapping amplitude, by a cantilever piezo before the tip engages the sample. The tip engages the sample at some set point, which may be thought of as the amplitude of the cantilever as influenced by contact with the sample. A lower set point gives a reduced amplitude and closer mean tip-to-sample distance. This is also depicted in Fig. 17.10. The feedback signal to the z-direction sample piezo (to keep the set point constant) is a measure of topography. This height data is recorded on a computer. The extender electronics is used to measure the phase angle lag between the cantilever piezo drive signal and the cantilever response during sample engagement. As illustrated in Fig. 17.10, the phase angle lag is (at least partially) a function of the viscoelastic properties of the sample material. A range of tapping amplitudes and set points can be used for measurements. A commercially etched single-crystal silicon tip (DI TESP) used for tapping mode, with a radius of 5–10 nm, a stiffness of 20–100 N/m, and a natural frequency of 350–400 kHz is normally used. Scanning is normally set to a rate of 1 Hz along the fast axis.

17.1.8 Boundary Lubrication Measurements

To study nanoscale boundary lubrication, adhesive forces are measured in the force calibration mode, as previously described. The adhesive forces are also calculated from the horizontal intercept of friction versus normal load curves at a zero value of friction force. For friction measurements, the samples are typically scanned using an Si_3N_4 tip over an area of $2\times2\,\mu m$ at the normal load ranging from 5 to 130 nN. The samples are generally scanned with a scan rate of 0.5 Hz resulting in a scanning speed of $2\,\mu m/s$. Velocity effects on friction are studied by changing the scan frequency from 0.1 to 60 Hz while the scan size is maintained at $2\times2\,\mu m$, which allows the velocity to vary from 0.4 to $240\,\mu m/s$. To study the durability properties, the friction force and coefficient of friction are monitored during scanning at a normal load of 70 nN and a scanning speed of $0.8\,\mu m/s$ for a desired number of cycles [17.27, 28, 30].

17.2 Friction and Adhesion

17.2.1 Atomic-Scale Friction

To study friction mechanisms on an atomic scale, a well characterized, freshly cleaved surface of highly oriented pyrolytic graphite (HOPG) has been used by *Mate* et al. [17.17] and *Ruan* and *Bhushan* [17.20].

The atomic-scale friction force of HOPG exhibited the same periodicity as that of the corresponding topography (Fig. 17.11a), but the peaks in friction and those in topography are displaced relative to each other (Fig. 17.11b). A Fourier expansion of the interatomic potential was used to calculate the conservative interatomic forces between atoms of the FFM tip and those of the graphite surface. Maxima in the interatomic forces in the normal and lateral directions do not occur at the same location, which explains the observed shift between the peaks in the lateral force and those in the corresponding topography. Furthermore, the observed local variations in friction force were explained by variation in the intrinsic lateral force between the sample and the FFM tip resulting from an atomic-scale stick-slip process [17.17, 20].

17.2.2 Microscale Friction

Local variations in the microscale friction of cleaved graphite have been observed (Fig. 17.12). These arise from structural changes that occur during the cleaving process [17.21]. The cleaved HOPG surface is largely atomically smooth, but exhibits line-shaped regions in which the coefficient of friction is more than an order of magnitude larger. Transmission electron microscopy indicates that the line-shaped regions consist of graphite planes of different orientation, as well as of amorphous

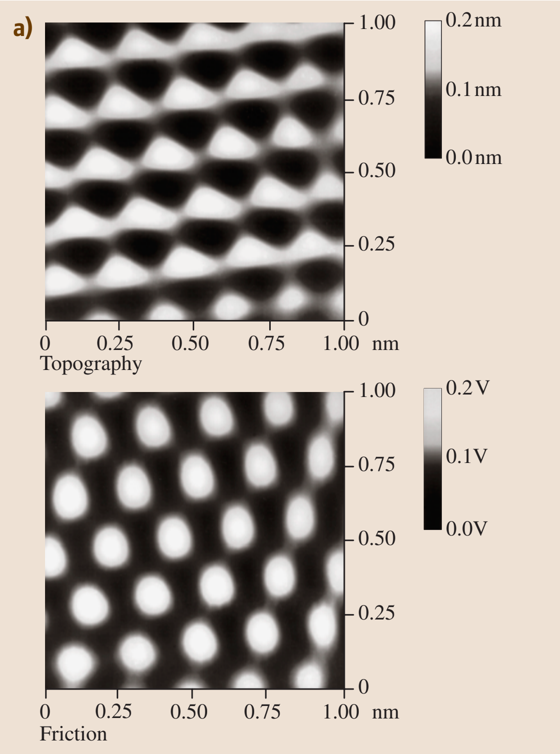

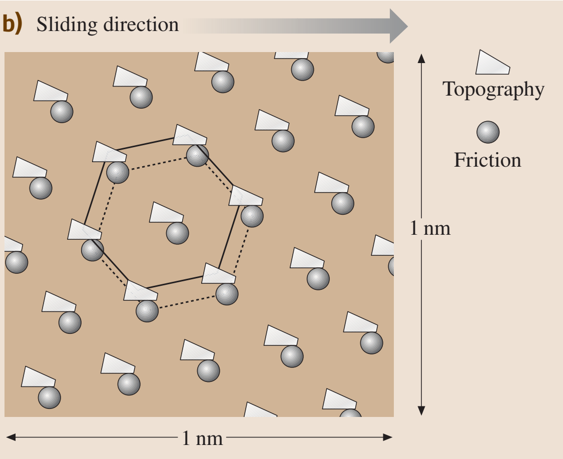

Fig. 17.11 **(a)** Gray-scale plots of surface topography and friction force maps of a 1 nm × 1 nm area of freshly cleaved HOPG showing the atomic-scale variation of topography and friction and **(b)** schematic of superimposed topography and friction maps from **(a)**; the symbols correspond to maxima. Note the spatial shift between the two plots [17.20]

carbon. Differences in friction have also been observed for multiphase ceramic materials [17.53]. Figure 17.13 shows the surface roughness and friction force maps of Al_2O_3-TiC (70–30 wt%). TiC grains have a Knoop hardness of about 2,800 kg/mm^2. Therefore, they do not polish as much and result in a slightly higher elevation (about 2–3 nm higher than that of Al_2O_3 grains). TiC grains exhibit higher friction force than Al_2O_3 grains. The coefficients of friction of TiC and Al_2O_3 grains are 0.034 and 0.026, respectively, and the coefficient of friction of Al_2O_3-TiC composite is 0.03. Local variation in friction force also arises from the scratches present on the Al_2O_3-TiC surface. *Meyer* et al. [17.54] also used FFM to measure structural variations of organic mono- and multilayer films. All of these measurements suggest that the FFM can be used for structural mapping of the

Fig. 17.12 **(a)** Surface roughness and **(b)** friction force maps at a normal load of 42 nN of freshly cleaved HOPG surface against an Si_3N_4 FFM tip. Friction in the *line-shaped region* is over an order of magnitude larger than the *smooth areas* [17.20]

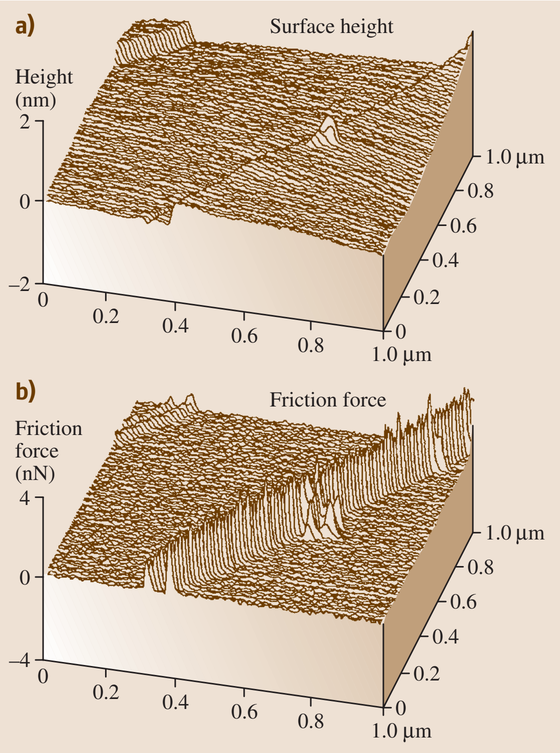

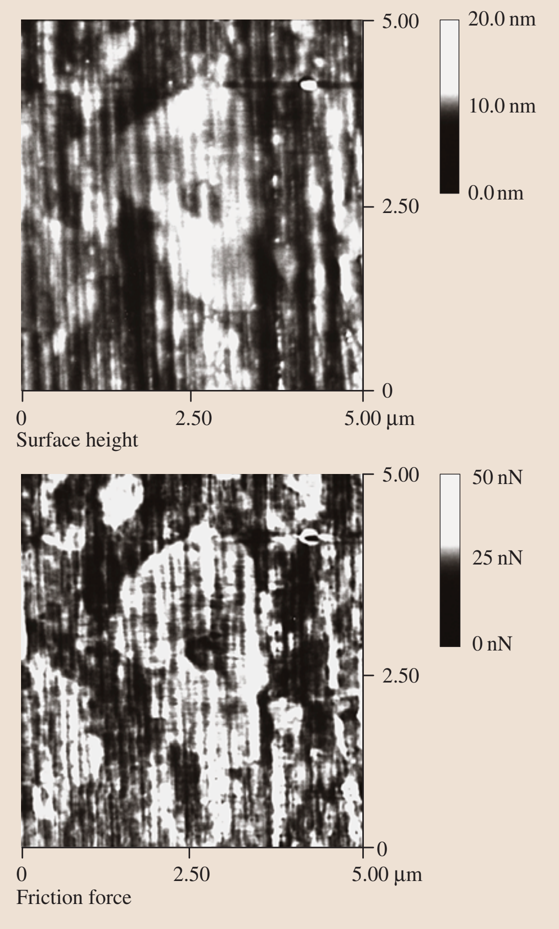

Fig. 17.13 Gray-scale surface roughness ($\sigma = 0.80$ nm) and friction force maps (mean = 7.0 nN, $\sigma = 0.90$ nN) for Al_2O_3-TiC (70–30 wt%) at a normal load of 138 nN [17.53]

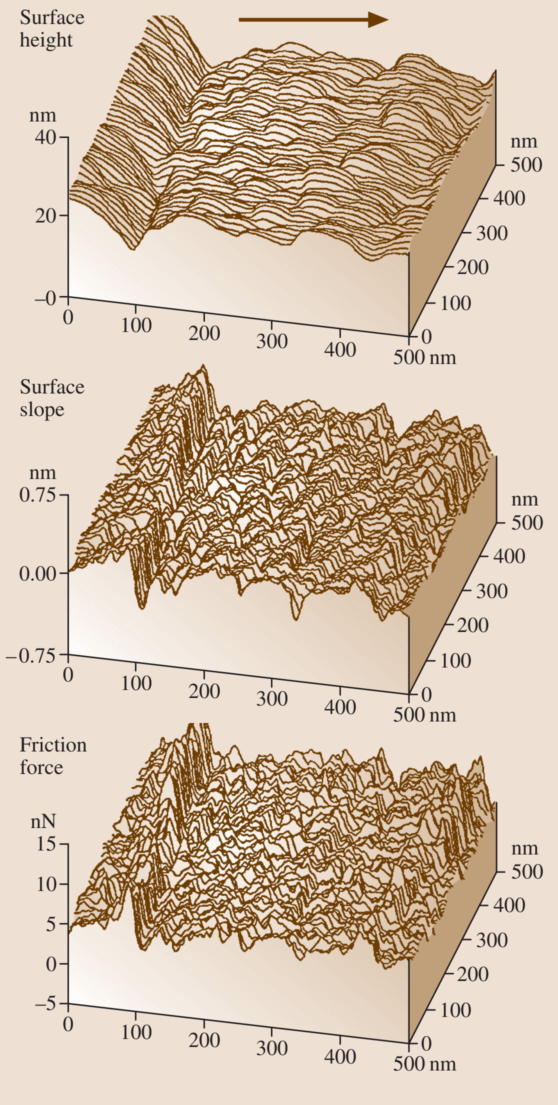

Fig. 17.14 Surface roughness map ($\sigma = 4.4$ nm); surface slope map taken in the sample sliding direction, the horizontal axis, (mean = 0.023, $\sigma = 0.197$); and friction force map (mean = 6.2 nN, $\sigma = 2.1$ nN) for a lubricated thin-film magnetic rigid disk for a normal load of 160 nN [17.22]

surfaces. FFM measurements can be used to map chemical variations, as indicated by the use of the FFM with a modified probe tip to map the spatial arrangement of chemical functional groups in mixed organic monolayer films [17.55]. Here, sample regions that had stronger interactions with the functionalized probe tip exhibited larger friction.

Local variations in the microscale friction of nominally rough, homogeneous surfaces can be significant and are seen to depend on the local surface slope, rather than the surface height distribution, Fig. 17.14. This dependence was first reported by *Bhushan* and *Ruan* [17.18], *Bhushan* et al. [17.22], and *Bhushan* [17.37] and later discussed in more detail in [17.56] and [17.57]. In order to show elegantly any correlation between local values of friction and surface roughness, surface roughness and friction force maps of a gold-coated ruling with somewhat rectangular grids and a silicon grid with square pits were obtained, see Fig. 17.15 [17.57]. Figures 17.14 and 17.15 show the surface roughness map, the slopes of the roughness map taken along the sliding direction (surface slope map), and the friction force map for various samples. There is a strong correlation between the surface slopes and friction forces. For example, in Fig. 17.15, friction force is high locally at the edge of the grids and pits with a positive slope and is low at the edges with a negative slope.

We now examine the mechanism of microscale friction, which may explain the resemblance between the slope of surface roughness maps and the corresponding friction force maps [17.4, 5, 18, 20–22, 56, 57]. There are three dominant mechanisms of friction: adhesive, adhesive and roughness (ratchet), and plowing [17.10, 57]. First, we may assume these to be additive. The adhesive mechanism cannot explain the local variation in friction. Next, we consider the ratchet mechanism. We consider a small tip sliding over an asperity making an angle θ with the horizontal plane, Fig. 17.16. The normal force W (normal to the general surface) applied by the tip to the sample surface is constant. The friction force F on the sample would be a constant for a smooth surface if the friction mechanism does not change. For a rough surface, as shown in Fig. 17.16, if the adhesive mechanism does not change during sliding, the local value of coefficient of friction remains constant,

$$\mu_0 = S/N\,, \tag{17.1}$$

Fig. 17.15a,b Surface roughness map, surface slope map taken in the sample sliding direction (the horizontal axis), and friction force map for **(a)** a gold-coated ruling (with somewhat rectangular grids with a pitch of 1 μm and a ruling step height of about 70 nm) at a normal load of 25 nN and **(b)** a silicon grid (with 5 μm square pits of 180-nm depth and a pitch of 10 μm) [17.57]

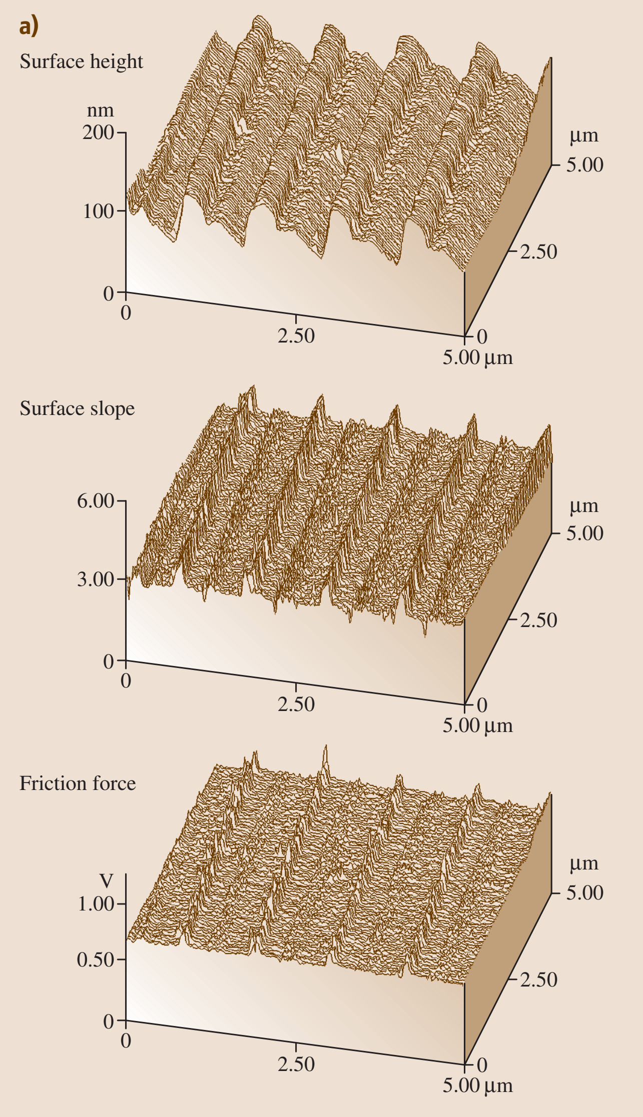

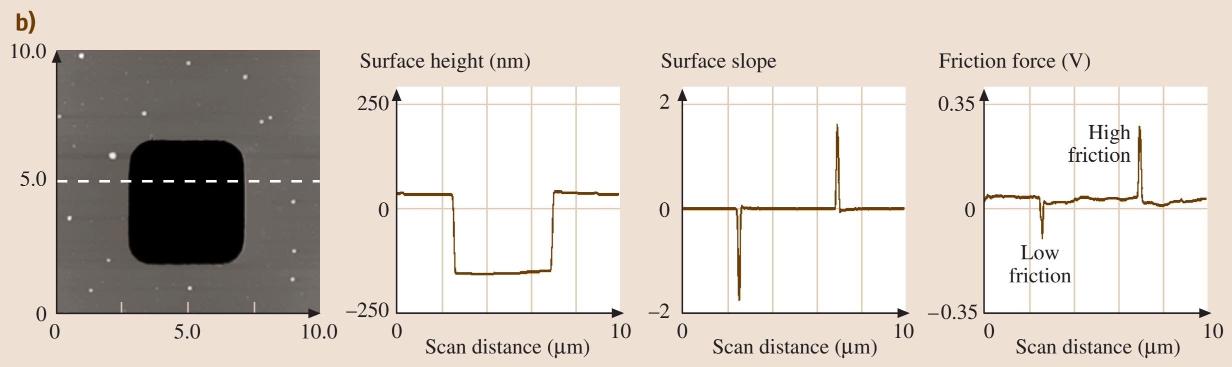

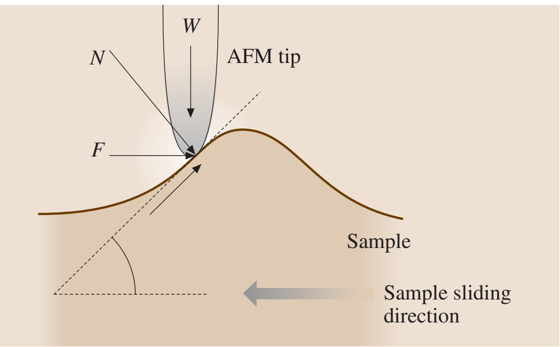

Fig. 17.16 Schematic illustration showing the effect of an asperity (making an angle θ with the horizontal plane) on the surface in contact with the tip on local friction in the presence of adhesive friction mechanism. W and F are the normal and friction forces, respectively, and S and N are the force components along and perpendicular to the local surface of the sample at the contact point, respectively

where S is the local friction force and N is the local normal force. However, the friction and normal forces are measured with respect to global horizontal and normal axes, respectively. The measured local coefficient of friction μ_1 in the ascending part is

$$\begin{aligned}\mu_1 &= F/W\\ &= (\mu_0+\tan\theta)\,/\,(1-\mu_0\tan\theta) \sim \mu_0+\tan\theta\,,\\ &\text{for small } \mu_0\tan\theta\,, \end{aligned} \tag{17.2}$$

indicating that in ascending part of the asperity one may simply add the friction force and the asperity slope to one another. Similarly, on the right-hand side (descending part) of the asperity,

$$\begin{aligned}\mu_2 &= (\mu_0-\tan\theta)\,/\,(1+\mu_0\tan\theta) \sim \mu_0-\tan\theta\,,\\ &\text{for small } \mu_0\tan\theta\,. \end{aligned} \tag{17.3}$$

For a symmetrical asperity, the average coefficient of friction experienced by the FFM tip traveling across the whole asperity is

$$\begin{aligned}\mu_{\text{ave}} &= (\mu_1+\mu_2)\,/2\\ &= \mu_0\left(1+\tan^2\theta\right)/\left(1-\mu_0^2\tan^2\theta\right)\\ &\sim \mu_0\left(1+\tan^2\theta\right)\,,\\ &\text{for small } \mu_0\tan\theta\,. \end{aligned} \tag{17.4}$$

Finally, we consider the plowing component of friction with tip sliding in either direction, which is [17.10, 58]

$$\mu_{\text{p}} \sim \tan\theta\,. \tag{17.5}$$

Because we notice little damage of the sample surface in FFM measurements, the contribution by plowing is expected to be small and the ratchet mechanism is believed to be the dominant mechanism for the local variations in the friction force map. With the tip sliding over the leading (ascending) edge of an asperity, the surface slope is positive; it is negative during sliding over the trailing (descending) edge of an asperity. Thus, measured friction is high at the leading edge of asperities and low at the trailing edge. In addition to the slope effect, the collision of the tip when encountering an asperity with a positive slope produces additional torsion of the cantilever beam, leading to higher measured friction force. When encountering an asperity with the same negative slope, however, there is no collision effect and hence no effect on torsion. This effect also contributes to the difference in friction forces when the tip scans up and down on the same topography feature. The ratchet mechanism and the collision effects thus explain semiquantitatively the correlation between the slopes of the roughness maps and friction force maps observed in Figs. 17.14 and 17.15. We note that in the ratchet mechanism, the FFM tip is assumed to be small compared to the size of asperities. This is valid since the typical radius of curvature of the tips is about 10–50 nm. The radii of curvature of samples, the asperities measured here (the asperities that produce most of the friction variation) are found to typically be about 100–200 nm, which is larger than that of the FFM tip [17.59]. It is important to note that the measured local values of friction and normal forces are measured with respect to global (and not local) horizontal and vertical axes, which are believed to be relevant in applications.

17.2.3 Directionality Effect on Microfriction

During friction measurements, the friction force data from both the forward (trace) and backward (retrace) scans are useful in understanding the origins of the observed friction forces. Magnitudes of material-induced effects are independent of the scanning direction, whereas topography-induced effects are different between forward and backward scanning directions. Since the sign of the friction force changes as the scanning direction is reversed (because of the reversal of torque applied to the end of the tip), addition of the friction force data of the forward and backward scan eliminates the material-induced effects, while topography-induced effects still remain. Subtraction of the data between forward and backward scans does not eliminate either effect, see Fig. 17.17 [17.57].

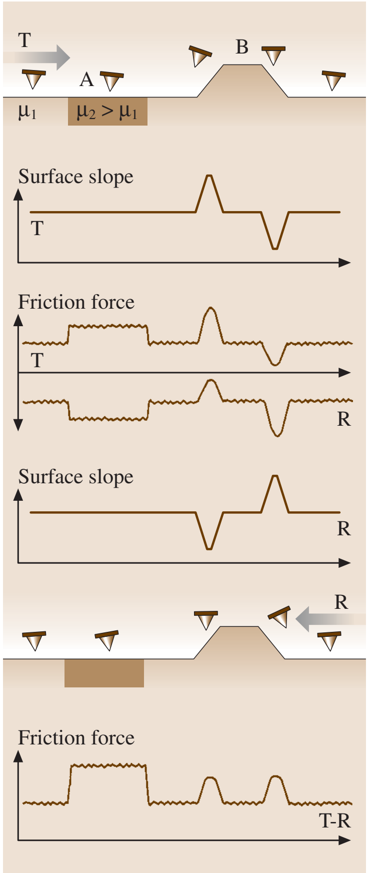

Fig. 17.17 Schematic of friction forces expected when a tip traverses a sample that is composed of different materials and sharp changes in topography. A schematic of surface slope is also shown (T = Trace, R = Retrace)

Owing to the reversal of the sign of the retrace (R) friction force with respect to the trace (T) data, the friction force variations due to topography are in the same direction (peaks in trace correspond to peaks in retrace). However, the magnitudes of the peaks in trace and retrace at a given location are different. An increase in the friction force experienced by the tip when scanning up a sharp change in topography is more than the decrease in the friction force experienced when scanning down the same topography change, partly due to the collision effects discussed earlier. Asperities on engineering surfaces are asymmetrical, which also affect the magnitude of friction force in the two directions. In addition, asymmetry in tip shape may have an effect on the directionality effect of friction. We will later note that the magnitude of surface slopes are virtually identical. Therefore, the tip shape asymmetry should not have much effect.

Figure 17.18 shows surface height and friction force data for gold ruler and a silicon grid in the trace and retrace directions. Subtraction of two friction data yields a residual peak, because of the differences in the magnitudes of friction forces in the two directions. This effect is observed at all locations of significant changes in topography.

In order to facilitate comparison of directionality effect on friction, it is important to take into account the sign change of the surface slope and friction force in

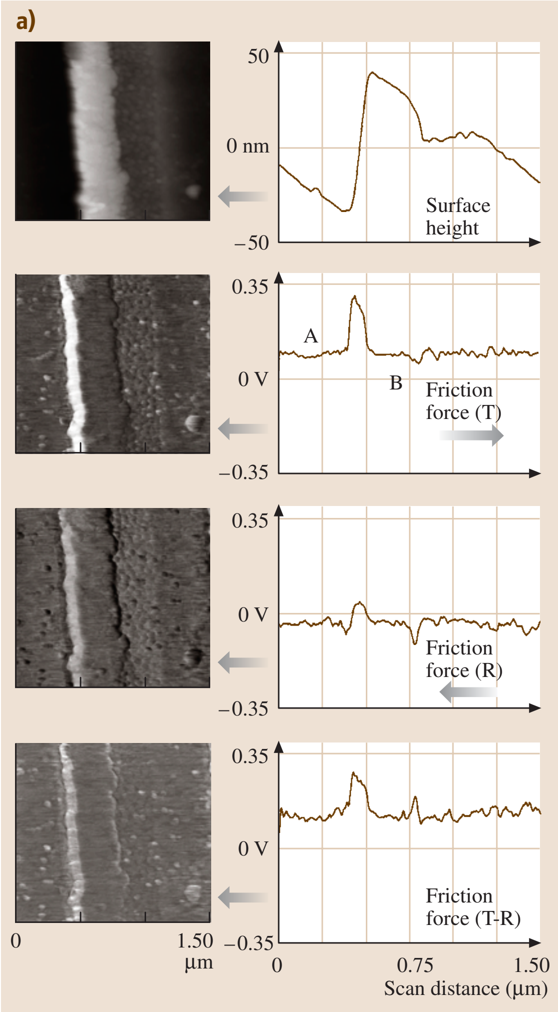

Fig. 17.18 **(a)** Gray-scale images and two-dimensional profiles of surface height and friction forces across a single ruling of the gold-coated ruling and **(b)** (*see next page*) two-dimensional profiles of surface height and friction forces across a silicon grid pit. Friction force data in trace (T) and retrace (R) directions and substrated force data are presented

the trace and retrace directions. Figure 17.19 shows surface height, surface slope, and friction force data for the two samples in the trace and retrace directions. The correlation between surface slope and friction forces is clear. The third column in the figures shows retrace slope and friction data with an inverted sign (−retrace). Now we can compare trace data with −retrace data. It is clear that the friction experienced by the tip is dependent upon the scanning direction because of surface topography. In addition to the effect of topographical changes discussed earlier, during surface-finishing processes, material can be transferred preferentially onto one side of the asperities, which also causes asymmetry and direction dependence. Reduction in local variations and in directionality of friction properties requires careful optimization of surface roughness distributions and of surface-finishing processes.

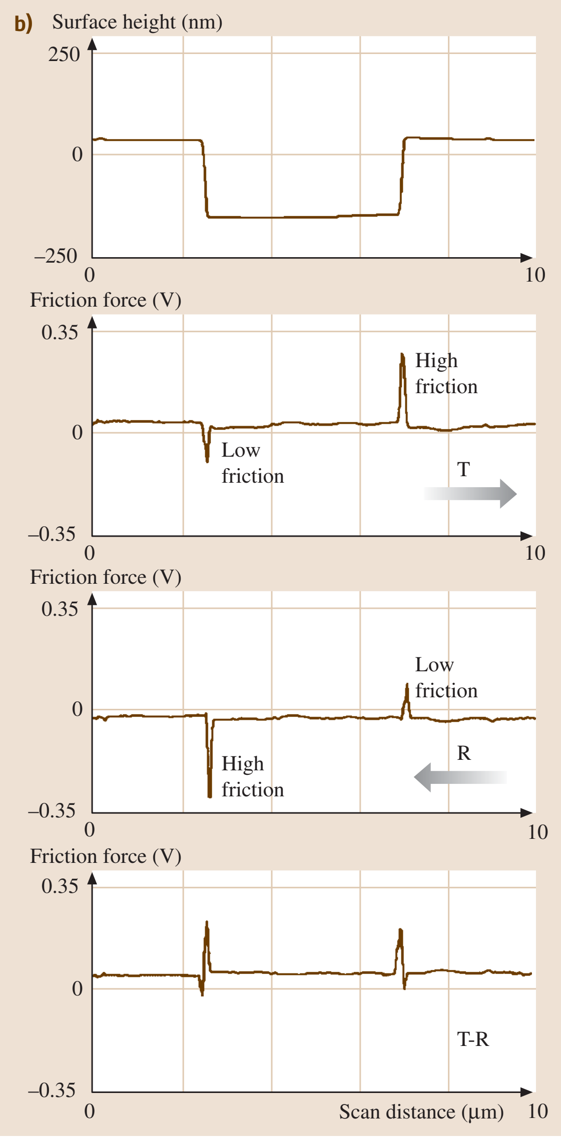

Fig. 17.18 b

The directionality as a result of surface asperities effect will also be manifested in macroscopic friction data, i. e., the coefficient of friction may be different in one sliding direction than in the other direction. Asymmetrical shape of asperities accentuates this effect. The frictional directionality can also exist in materials with particles with a preferred orientation. The directionality effect in friction on a macroscale is observed in some magnetic tapes. In a macroscale test, a 12.7-mm wide polymeric magnetic tape was wrapped over an aluminum drum and slid in a reciprocating motion with a normal load of 0.5 N and a sliding speed of about 60 mm/s [17.3]. The coefficient of friction as a function of sliding distance in either direction is shown in Fig. 17.20. We note that the coefficient of friction on a macroscale for this tape varies in different directions. Directionality in friction is sometimes observed on the macroscale; on the microscale this is the norm [17.4, 13]. On the macroscale, the effect of surface asperities is normally averaged out over a large number of contacting asperities.

17.2.4 Velocity Dependence on Microfriction

AFM/FFM experiments are generally conducted at relative velocities as high as a few tens of µm/s. To simulate applications, it is of interest to conduct friction experiments at higher velocities. High velocities can be achieved by mounting either the sample or the cantilever beam on a shear wave transducer (ultrasonic transducer) to produce surface oscillations at MHz frequencies [17.24, 25, 60]. The velocities on the order of a few mm/s can thus be achieved. The effect of in-plane and out-of-plane sample vibration amplitude on the coefficient of friction is shown in Fig. 17.21. Vibration of a sample at ultrasonic frequencies (> 20 kHz) can substantially reduce the coefficient of friction, known as ultrasonic lubrication or sonolubrication. When the surface is vibrated in-plane, classical hydrodynamic lubrication develops hydrodynamic pressure, which supports the tip and reduces friction. When the surface is vibrated out-of-plane, a lift-off caused by the squeeze-film lubrication (a form of hydrodynamic lubrication), reduces friction.

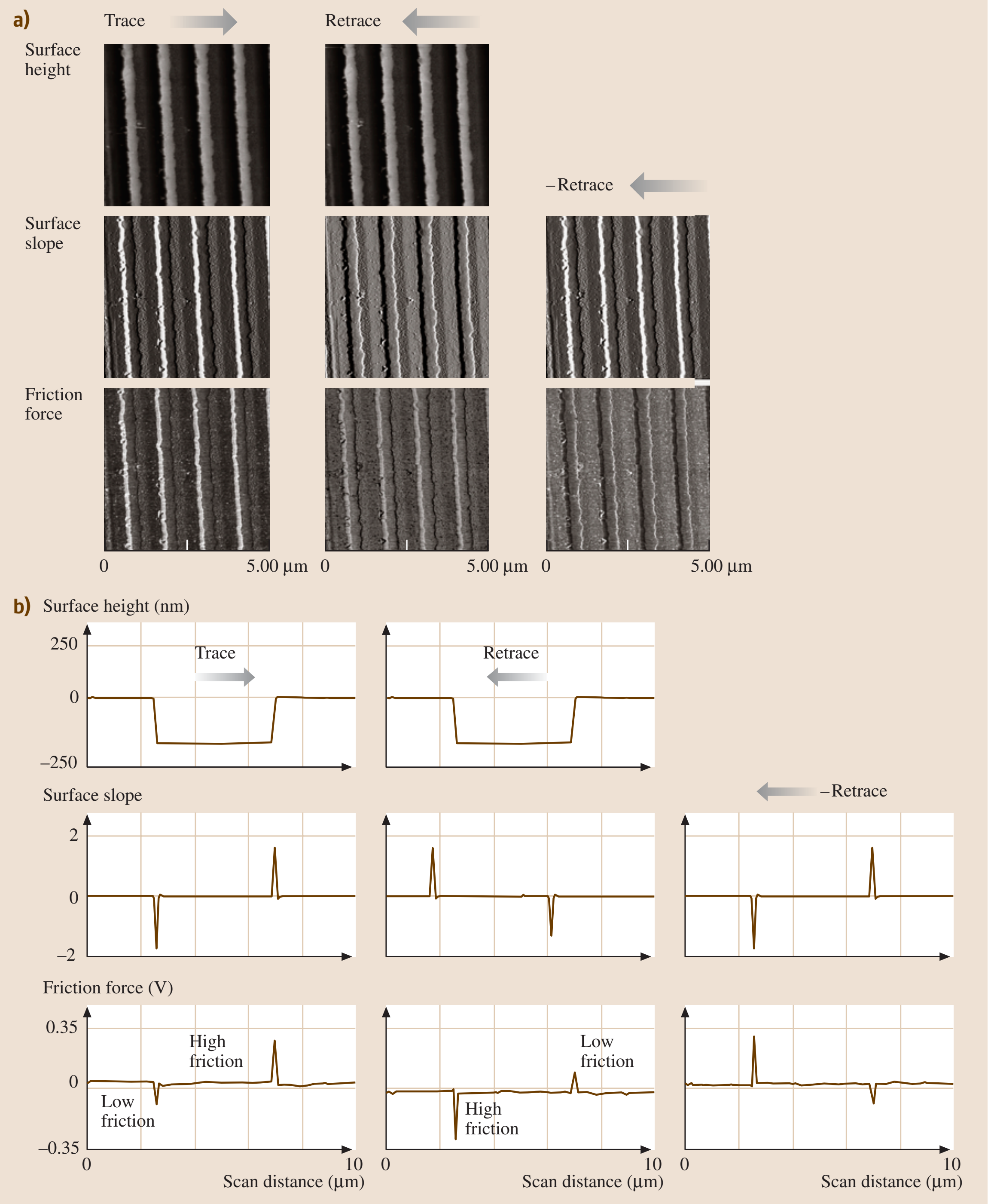

Fig. 17.19 **(a)** Gray-scale images of surface heights, surface slopes, and friction forces for scans across a gold-coated ruling, and **(b)** two-dimensional profiles of surface heights, surface slopes, and friction forces for scans across the silicon grid pit. *Arrows* indicate the tip sliding direction [17.57]

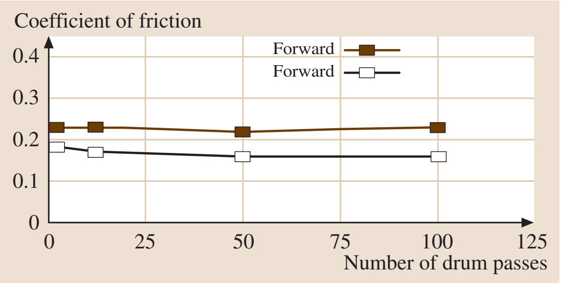

Fig. 17.20 Coefficient of macroscale friction as a function of drum passes for a polymeric magnetic tape sliding over an aluminum drum in a reciprocating mode in both directions. Normal load = 0.5 N over 12.7-mm wide tape, sliding speed = 60 mm/s [17.37]

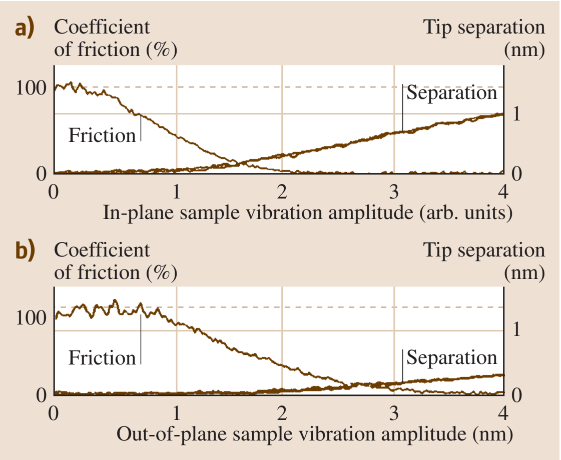

Fig. 17.21a,b Reduction of coefficient of friction measured at a normal load of 100 nN and average tip separation as a function of surface amplitude on a single-crystal silicon subjected to **(a)** in-plane and **(b)** out-of-plane vibrations at about 1 MHz against a silicon nitride tip [17.24]

17.2.5 Effect of Tip Radii and Humidity on Adhesion and Friction

The tip radius and relative humidity affect adhesion and friction for dry and lubricated surfaces [17.23, 36].

Experimental Observations

Figure 17.22 shows the variation of single-point adhesive-force measurements as a function of tip radius on a Si(100) sample for several humidities. The adhesive force data are also plotted as a function of relative humidity for several tip radii. The general trend at humidities up to the ambient is that a 50-nm radius Si_3N_4 tip exhibits a slightly lower adhesive force compared to the other microtips of larger radii; in the latter case, values are similar. Thus, for the microtips there is no appreciable variation in adhesive force with tip radius at a given humidity up to the ambient. The adhesive force increases as relative humidity increases for all tips.

Sources of adhesive force between a tip and a sample surface are van der Waals attraction and meniscus formation [17.10, 58]. Relative magnitudes of the forces from the two sources are dependent upon various factors, including the distance between the tip and the sample surface, their surface roughness, their hydrophobicity, and relative humidity [17.61]. For most rough surfaces, meniscus contribution dominates at moderate to high humidities, which arise from capillary condensation of water vapor from the environment. If enough liquid is present to form a meniscus bridge, the meniscus force should increase with an increase in tip radius (proportional to tip radius for a spherical tip). In addition, an increase in tip radius results in increased contact area leading to higher values of van der Waals forces. However, if nanoasperities on the tip and the sample are considered, then the number of contacting and near-contacting asperities forming meniscus bridges increases with an increase of humidity leading to an increase in meniscus forces. These explain the trends observed in Fig. 17.22. From the data, the tip radius has little effect on the adhesive forces at low humidities, but increases with tip radius at high humidity. Adhesive force also increases with an increase in humidity for all tips. This observation suggests that thickness of the liquid film at low humidities is insufficient to form continuous meniscus bridges to affect adhesive forces in the case of all tips.

Figure 17.22 also shows the variation in coefficient of friction as a function of tip radius at a given humidity and as a function of relative humidity for a given tip radius for Si(100). It can be observed that for 0% RH, the coefficient of friction is about the same for the tip radii except for the largest tip, which shows a higher value. At all other humidities, the trend consistently shows that the coefficient of friction increases with tip radius. An increase in friction with tip radius at low to moderate humidities arises from increased contact area (higher van der Waals forces) and higher values of shear forces required for larger contact area. At high humidities, similar to adhesive force data, an increase with tip radius occurs due to both contact area and meniscus effects. Although AFM/FFM measurements are able to measure the combined effect of the contribution of van der Waals

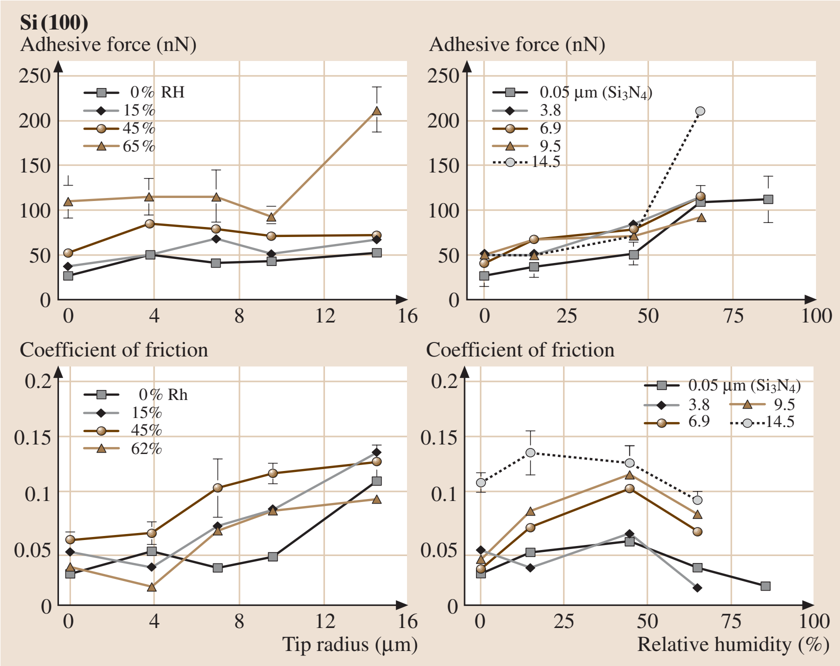

Fig. 17.22 Adhesive force and coefficient of friction as a function of tip radius at several humidities and as a function of relative humidity (RH) at several tip radii on Si(100) [17.23]

and meniscus forces toward friction force or adhesive force, it is difficult to measure their individual contributions separately. It can be seen that for all tips, the coefficient of friction increases with humidity to about ambient, beyond which it starts to decrease. The initial increase in the coefficient of friction with humidity arises from the fact that the thickness of the water film increases with an increase in the humidity, which results in a larger number of nanoasperities forming meniscus bridges and leads to higher friction (larger shear force). The same trend is expected with the microtips beyond 65% RH. This is attributed to the fact that at higher humidities, the adsorbed water film on the surface acts as a lubricant between the two surfaces. Thus the interface is changed at higher humidities, resulting in lower shear strength and, hence, lower friction force and coefficient of friction.

Adhesion and Friction Force Expressions for a Single-Asperity Contact

We now obtain the expressions for the adhesive force and coefficient of friction for a single-asperity contact with a meniscus formed at the interface. For a spherical asperity of radius R in contact with a flat and smooth surface with the composite modulus of elasticity E^* and with a concave meniscus, the attractive meniscus force (adhesive force) W_{ad} is given as [17.10, 58]

$$W_{\mathrm{ad}} = 2\pi R \gamma_1 (\cos\theta_1 + \cos\theta_2)\ . \tag{17.6}$$

For an elastic contact for both extrinsic (W) and intrinsic (W_{ad}) normal load, the friction force is given as

$$F_{\mathrm{e}} = \pi\tau \left(\frac{3\,(W + W_{\mathrm{ad}})\,R}{4E^*}\right)^{2/3}\ , \tag{17.7}$$

where γ_1 is the surface tension of the liquid in air, θ_1 and θ_2 are the contact angles between the liquid and the two surfaces, W is the external load, and τ is the average shear strength of the contacts. (Surface energy effects are not considered here.) Note that adhesive force increases linearly with an increase in the tip radius, and the friction force increases with an increase in tip radius as $R^{2/3}$ and with normal load as $(W + W_{\mathrm{ad}})^{2/3}$. The experimental data in support of $W^{2/3}$ dependence on the friction force can be found in various references (see, for example, [17.62]). The coefficient of friction μ_{e} is obtained from (17.7) as

$$\begin{aligned}\mu_{\mathrm{e}} &= \frac{F_{\mathrm{e}}}{(W + W_{\mathrm{ad}})} \\ &= \pi\tau \left(\frac{3R}{4E^*}\right)^{2/3} \frac{1}{(W + W_{\mathrm{ad}})^{1/3}}\ .\end{aligned} \tag{17.8}$$

In the plastic contact regime [17.58], the coefficient of friction μ_p is obtained as

$$\mu_\mathrm{p} = \frac{F_\mathrm{p}}{(W + W_\mathrm{ad})} = \frac{\tau}{H_\mathrm{s}} \,, \qquad (17.9)$$

where H_s is the hardness of the softer material. Note that in the plastic contact regime, the coefficient of friction is independent of external load, adhesive contributions, and surface geometry.

In comparison, for multiple-asperity contacts in the elastic contact regime, the total adhesive force W_ad is the summation of adhesive forces at n individual contacts,

$$W_\mathrm{ad} = \sum_{i=1}^{n} (W_\mathrm{ad})_i \quad \text{and}$$

$$\mu_\mathrm{e} \approx \frac{3.2\tau}{E^* \left(\sigma_\mathrm{p}/R_\mathrm{p}\right)^{1/2} + W_\mathrm{ad}/W} \,, \qquad (17.10)$$

where σ_p and R_p are the standard deviation of summit heights and average summit radius, respectively. Note that the coefficient of friction depends upon the surface roughness. In the plastic contact regime, the expression for μ_p in (17.9) does not change.

The source of the adhesive force, in a wet contact in the AFM experiments being performed in an ambient environment, includes mainly attractive meniscus force due to capillary condensation of water vapor from the environment. The meniscus force for a single contact increases with an increase in tip radius. A sharp AFM tip in contact with a smooth surface at low loads (on the order of a few nN) for most materials can be simulated as a single-asperity contact. At higher loads, for rough and soft surfaces, multiple contacts would occur. Furthermore, at low loads (nN range) for most materials, the local deformation would be primarily elastic. Assuming that shear strength of contacts does not change, the adhesive force for smooth and hard surfaces at low normal load (on the order of few nN) (for a single-asperity contact in the elastic contact regime) would increase with an increase in tip radius, and the coefficient of friction would decrease with an increase in total normal load as $(W + W_\mathrm{ad})^{-1/3}$ and would increase with an increase of tip radius as $R^{2/3}$. In this case, the Amontons law of friction, which states that the coefficient of friction is independent of normal load and is independent of apparent area of contact, does not hold. For a single-asperity plastic contact and multiple-asperity plastic contacts, neither the normal load nor tip radius come into play in the calculation of the coefficient of friction. In the case of multiple-asperity contacts, the number of contacts increase with an increase of normal load. Therefore, adhesive force increases with an increase in load.

In the data presented earlier in this section, the effect of tip radius and humidity on the adhesive forces and coefficient of friction is investigated for experiments with Si(100) surface at loads in the range of 10–100 nN. The multiple-asperity elastic contact regime is relevant for this study. An increase in humidity generally results in an increase in the number of meniscus bridges that would increase the adhesive force. As suggested earlier, that increase in humidity may also decrease the shear strength of contacts. A combination of an increase in adhesive

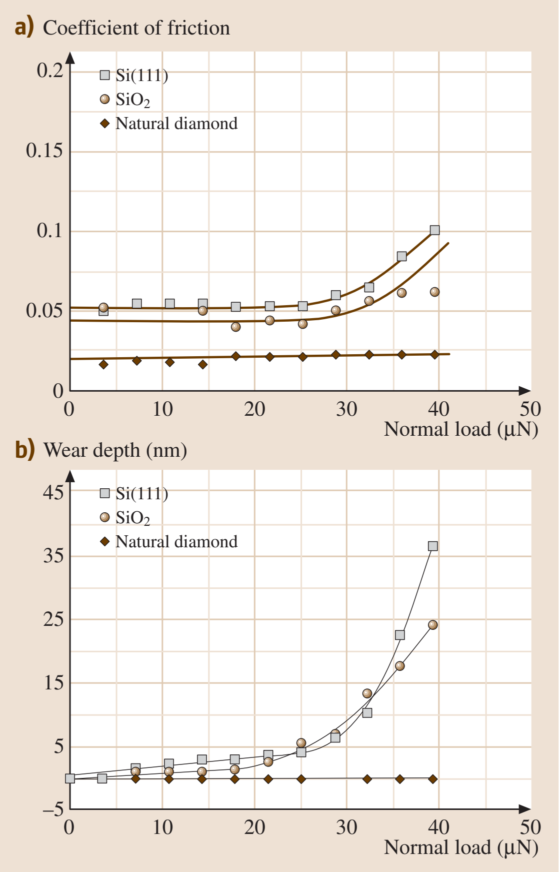

Fig. 17.23 **(a)** Coefficient of friction as a function of normal load and **(b)** corresponding wear depth as a function of normal load for silicon, SiO_2 coating, and natural diamond. Inflections in the curves for silicon and SiO_2 correspond to the contact stresses equal to the hardnesses of these materials [17.63]

Table 17.2 Surface roughness (standard deviation of surface heights or σ) and coefficients of friction on nano- and macroscales of various samples in air

Material	σ (nm)	Coefficient of nanoscale friction versus Si_3N_4 tip[a]	Coefficient of macroscale friction versus Si_3N_4 ball[b]
Graphite (HOPG)	0.09	0.006	0.1
Natural diamond	2.3	0.04	0.2
Si(100)	0.14	0.07	0.4

[a] Tip radius of about 50 nm in the load range of 10–150 nN (2.5–6.1 GPa) and a scanning speed of 0.5 nm/s and scan area of 1 nm × 1 nm for HOPG and a scanning speed of 4 µm/s and scan area of 1 µm × 1 µm for diamond and Si(100)

[b] Ball radius of 3 mm at a normal load of 1 N (0.6 GPa) and average sliding speed of 0.8 mm/s

force and a decrease in shear strength would affect the coefficient of friction. An increase in tip radius would increase the meniscus force (adhesive force). A substantial increase in the tip radius may also increase interatomic forces. These effects influence the coefficient of friction with an increase in the tip radius.

17.2.6 Scale Dependence on Friction

Table 17.2 shows the coefficient of friction measured for various materials on nano- and macroscales [17.19]. It is clearly observed that friction values are scale-dependent. The values on the nano/microscale are much lower than those on the macroscale. There can be the following four, and possibly more, differences in the operating conditions responsible for these differences. First, the contact stresses at AFM conditions, in spite of small tip radii, generally do not exceed the sample hardness that minimizes plastic deformation. Average contact stresses in macrocontacts are generally lower than in AFM contacts, however, a large number of asperities come into contact that go through some plastic deformation. Second, when measured for the small contact areas and very low loads used in microscale studies, indentation hardness is higher than at the macroscale, as will be discussed later [17.11, 34, 35]. Lack of plastic deformation and improved mechanical properties reduce the degree of wear and friction. Third, the small apparent area of contact reduces the number of particles trapped at the interface, and thus minimizes the third-body plowing contribution to the friction force [17.11]. As a fourth and final difference, we have seen in the previous section that coefficient of friction increases with an increase in the AFM tip radius. AFM data are taken with a sharp tip, whereas asperities coming in contact in macroscale tests range from nanoasperities to much larger asperities, which may be responsible for larger values of friction force on the macroscale.

To demonstrate the load dependence on the coefficient of friction, stiff cantilevers were used to conduct friction experiments at high loads, see Fig. 17.23 [17.63]. At higher loads (with contact stresses exceeding the hardness of the softer material), as anticipated, the coefficient of friction for microscale measurements increases toward values comparable to those obtained from macroscale measurements, and surface damage also increases. At high loads, plowing is a dominant contributor to the friction force. Based on these results, Amontons rule of friction, which states that the coefficient of friction is independent of apparent area and normal load, does not hold for microscale measurements. These findings also suggest that microcomponents sliding under lightly loaded conditions should experience ultralow friction and near-zero wear [17.11].

17.3 Scratching, Wear, Local Deformation, and Fabrication/Machining

17.3.1 Nanoscale Wear

Bhushan and *Ruan* [17.18] conducted nanoscale wear tests on polymeric magnetic tapes using conventional silicon nitride tips at two different loads of 10 and 100 nN (Fig. 17.24). For a low normal load of 10 nN, measurements were made twice. There was no discernible difference between consecutive measurements for this load. However, as the load was increased from 10 nN to 100 nN, topographical changes were ob-

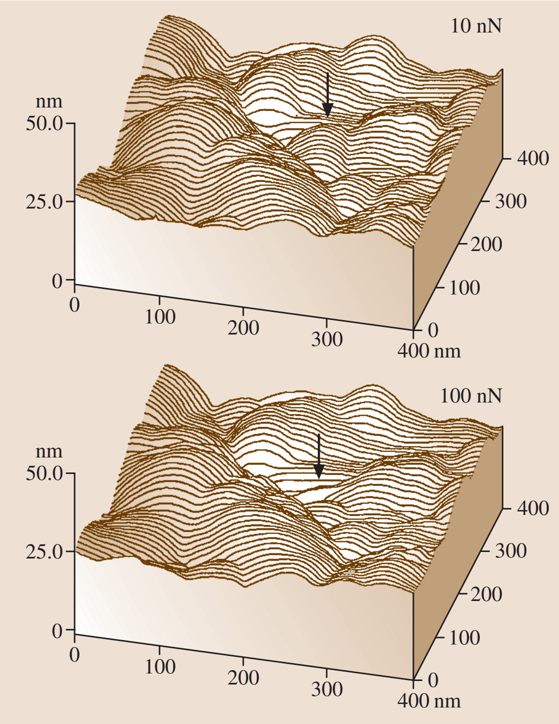

Fig. 17.24 Surface roughness maps of a polymeric magnetic tape at the applied normal load of 10 nN and 100 nN. Location of the change in surface topography as a result of nanowear is indicated by *arrows* [17.18]

served during subsequent scanning at a normal load of 10 nN; material was pushed in the sliding direction of the AFM tip relative to the sample. The material movement is believed to occur as a result of plastic deformation of the tape surface. Thus, deformation and movement of the soft materials on a nanoscale can be observed.

17.3.2 Microscale Scratching

The AFM can be used to investigate how surface materials can be moved or removed on micro- to nanoscales, for example, in scratching and wear [17.4] (where these things are undesirable), and nanofabrication/nanomachining (where they are desirable). Figure 17.25a shows microscratches made on Si(111) at various loads and scanning velocity of 2 µm/s after 10 cycles [17.22]. As expected, the scratch depth increases linearly with load. Such microscratching measurements can be used to study failure mechanisms on the microscale and evaluate the mechanical integrity (scratch resistance) of ultrathin films at low loads.

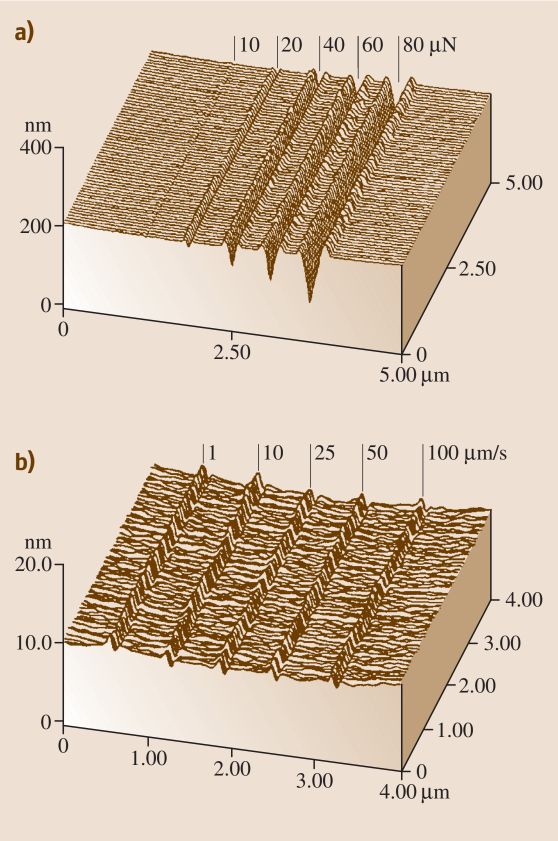

Fig. 17.25a,b Surface plots of **(a)** Si(111) scratched for 10 cycles at various loads and a scanning velocity of 2 µm/s. Note that x and y axes are in µm and z axis is in nm. **(b)** Si(100) scratched in one unidirectional scan cycle at a normal force of 80 µN and different scanning velocities

To study the effect of scanning velocity, unidirectional scratches, 5 µm in length, were generated at scanning velocities ranging from 1–100 µm/s at various normal loads ranging from 40–140 µN. There is no effect of scanning velocity obtained at a given normal load. (For representative scratch profiles at 80 µN, see Fig. 17.25b.) This may be because of a small effect of frictional heating with the change in scanning velocity used here. Furthermore, for a small change in interface temperature, there is a large underlying volume to dissipate the heat generated during scratching.

Scratching can be performed under ramped loading to determine the scratch resistance of materials and

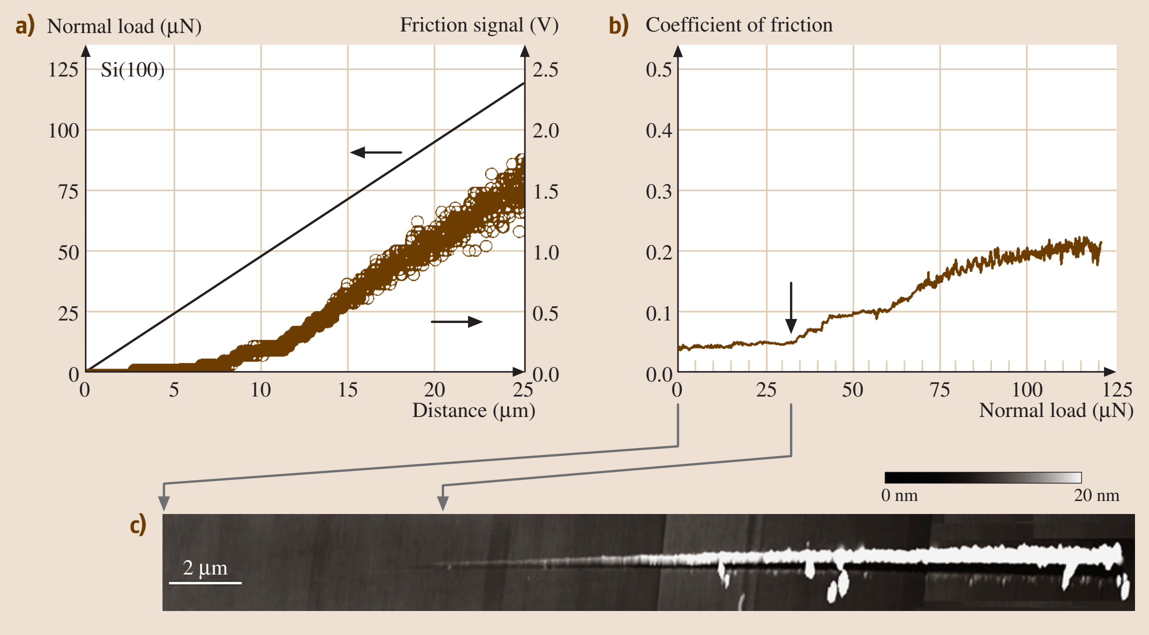

Fig. 17.26 (a) Applied normal load and friction signal measured during the microscratch experiment on Si(100) as a function of scratch distance, **(b)** friction data plotted in the form of coefficient of friction as a function of normal load, and **(c)** AFM surface height image of scratch obtained in tapping mode [17.33]

coatings. The coefficient of friction is measured during scratching, and the load at which the coefficient of friction increases rapidly is known as the "critical load", which is a measure of scratch resistance. In addition, post-scratch imaging can be performed in situ with the AFM in tapping mode to study failure mechanisms. Figure 17.26 shows data from a scratch test on Si(100) with a scratch length of 25 μm and a scratching velocity of 0.5 μm/s. At the beginning of the scratch, the coefficient of friction is 0.04, which indicates a typical value for silicon. At about 35 μN (indicated by the arrow in the figure), there is a sharp increase in the coefficient of friction, which indicates the critical load. Beyond the critical load, the coefficient of friction continues to increase steadily. In the post-scratch image, we note that at the critical load, a clear groove starts to form. This implies that Si(100) was damaged by plowing at the critical load, associated with the plastic flow of the material. At and after the critical load, small and uniform debris is observed and the amount of debris increases with increasing normal load. *Sundararajan* and *Bhushan* [17.33] have also used this technique to measure the scratch resistance of diamond-like carbon coatings ranging in thickness from 3.5–20 nm.

17.3.3 Microscale Wear

By scanning the sample in two dimensions with the AFM, wear scars are generated on the surface. Figure 17.27 shows the effect of normal load on wear depth. We note that wear depth is very small below 20 μN of normal load [17.64, 65]. A normal load of 20 μN corresponds to contact stresses comparable

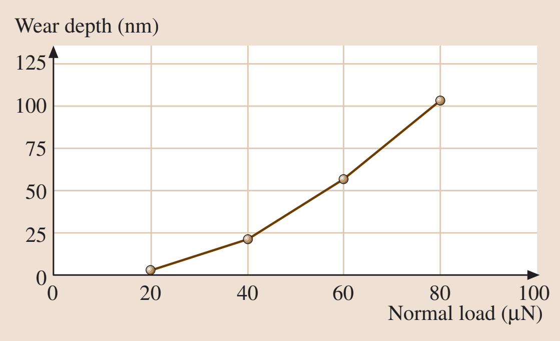

Fig. 17.27 Wear depth as a function of normal load for Si(100) after one cycle [17.65]

to the hardness of the silicon. Elastic deformation at loads below 20 µN is primarily responsible for low wear [17.63].

A typical wear mark, of the size 2 µm × 2 µm, generated at a normal load of 40 µN for one scan cycle and imaged using AFM with scan size of 4 µm × 4 µm at 300 nN load is shown in Fig. 17.28a [17.64]. The inverted map of wear marks shown in Fig. 17.28b indicates the uniform material removal at the bottom of the wear mark. An AFM image of the wear mark shows small debris at the edges, swiped during AFM scanning. Thus, the debris is loose (not sticky) and can be removed during the AFM scanning.

Next, we examine the mechanism of material removal on a microscale in AFM wear experiments [17.23, 64, 65]. Figure 17.29 shows a secondary electron image of the wear mark and associated wear particles. The specimen used for the scanning electron microscope (SEM) was not scanned with the AFM after initial wear, in order to retain wear debris in the wear region. Wear debris is clearly observed. In the SEM micrographs, the wear debris appear to be agglomerated because of high surface energy of the fine particles. Particles appear to be a mixture of rounded and so-called cutting type (feather-like or ribbon-like material). *Zhao* and *Bhushan* [17.65] reported an increase in the number and size of cutting-type particles with the normal load. The presence of cutting-type particles indicates that the material is removed primarily by plastic deformation.

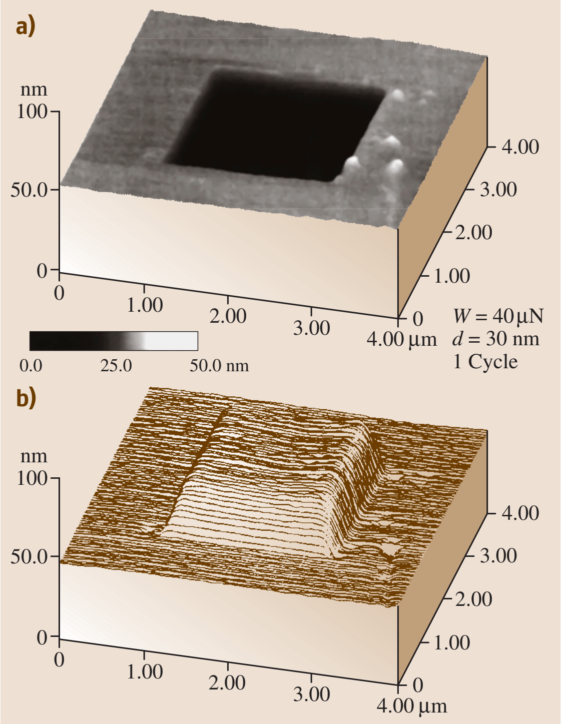

Fig. 17.28 **(a)** Typical gray-scale and **(b)** inverted AFM images of wear mark created using a diamond tip at a normal load of 40 µN and one scan cycle on Si(100) surface

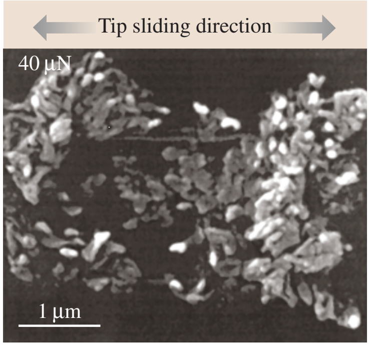

Fig. 17.29 Secondary electron image of wear mark and debris for Si(100) produced at a normal load of 40 µN and one scan cycle

To better understand the material removal mechanisms, transmission electron microscopy (TEM) has been used. The TEM micrograph of the worn region and associated diffraction pattern are shown in Fig. 17.30a,b. The bend contours are observed to pass through the wear mark in the micrograph. The bend contours around and inside the wear mark are indicative of a strain field, which, in the absence of applied stresses, can be interpreted as plastic deformation and/or elastic residual stresses. Often, localized plastic deformation during loading would lead to residual stresses during unloading; therefore, bend contours reflect a mix of elastic and plastic strains. The wear debris is observed outside the wear mark. The enlarged view of the wear debris in Fig. 17.30c shows that much of the debris is ribbon-like, indicating that material is removed by a cutting process via plastic deformation, which is consistent with the SEM observations. The diffraction pattern from inside the wear mark is similar to that of virgin silicon, showing no evidence of any phase transformation (amorphization) during wear. A selected area diffraction pattern of the wear debris shows some diffuse rings, which indicates the existence of amorphous

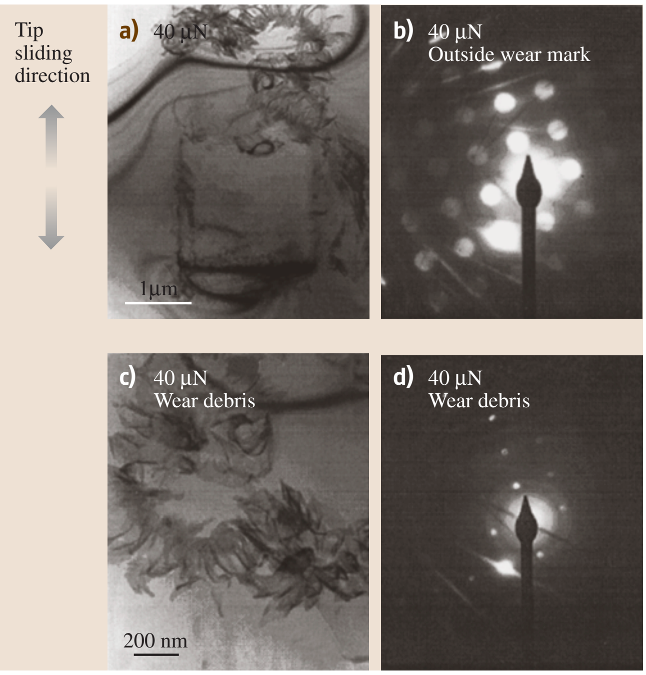

Fig. 17.30a–d Bright field TEM micrographs (*left*) and diffraction patterns (*right*) of wear mark **(a)**,**(b)** and wear debris **(c)**,**(d)** in Si(100) produced at a normal load of 40 µN and one scan cycle. Bend contours around and inside wear mark are observed

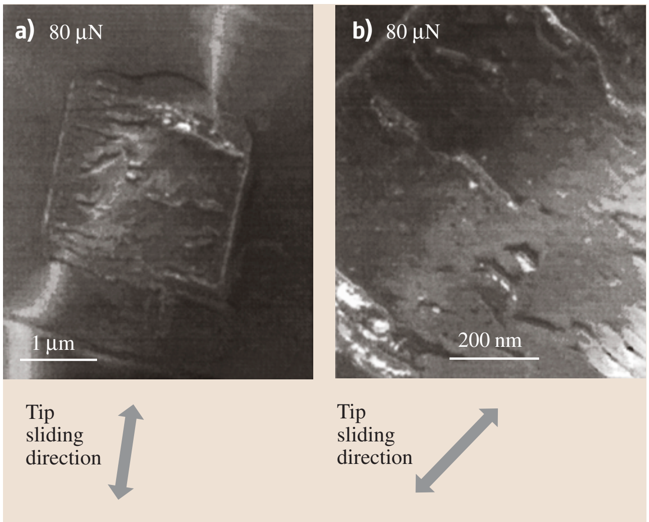

Fig. 17.31 **(a)** Bright field and **(b)** weak beam TEM micrographs of wear mark in Si(100) produced at a normal load of 80 µN and one scan cycle showing bend contours and dislocations [17.65]

material in the wear debris, confirmed as silicon oxide products from chemical analysis. It is known that plastic deformation occurs by generation and propagation of dislocations. No dislocation activity or cracking was observed at 40 µN. However, dislocation arrays could be observed at 80 µN. Figure 17.31 shows the TEM micrographs of the worn region at 80 µN; for better observation of the worn surface, wear debris was moved out of the wear mark using AFM with a large area scan at 300 nN after the wear test. The existence of dislocation arrays confirms that material removal occurs by plastic deformation. This corroborates the observations made in scratch tests at ramped load in the previous section. It is concluded that the material on microscale at high loads is removed by plastic deformation with a small contribution from elastic fracture [17.65].

To understand wear mechanisms, evolution of wear can be studied using AFM. Figure 17.32 shows evolution of wear marks of a DLC-coated disk sample. The data illustrate how the microwear profile for a load of 20 µN develops as a function of the number of scanning cycles [17.22]. Wear is not uniform, but is initiated at the nanoscratches. Surface defects (with high surface energy) present at nanoscratches act as initiation sites for wear. Coating deposition also may not be uniform on and near nanoscratches, which may lead to coating delamination. Thus, scratch-free surfaces will be relatively resistant to wear.

Wear precursors (precursors to measurable wear) can be studied by making surface potential measurements [17.38–40]. The contact potential difference, or simply surface potential between two surfaces, depends on a variety of parameters such as electronic work function, adsorption, and oxide layers. The surface potential map of an interface gives a measure of changes in the work function, which is sensitive to both physical and chemical conditions of the surfaces, including structural and chemical changes. Before material is actually removed in a wear process, the surface experiences stresses that result in surface and subsurface changes of structure and/or chemistry. These can cause changes in the measured potential of a surface. An AFM tip allows mapping of surface potential with nanoscale resolution. Surface height and change in surface potential maps of a polished single-crystal aluminum (100) sample abraded using a diamond tip at loads of 1 µN and 9 µN are shown in Fig. 17.33a. (Note that the sign of the change in surface potential is reversed here from that in [17.38].) It is evident that both abraded regions show a large potential contrast (∼ 0.17 V) with respect

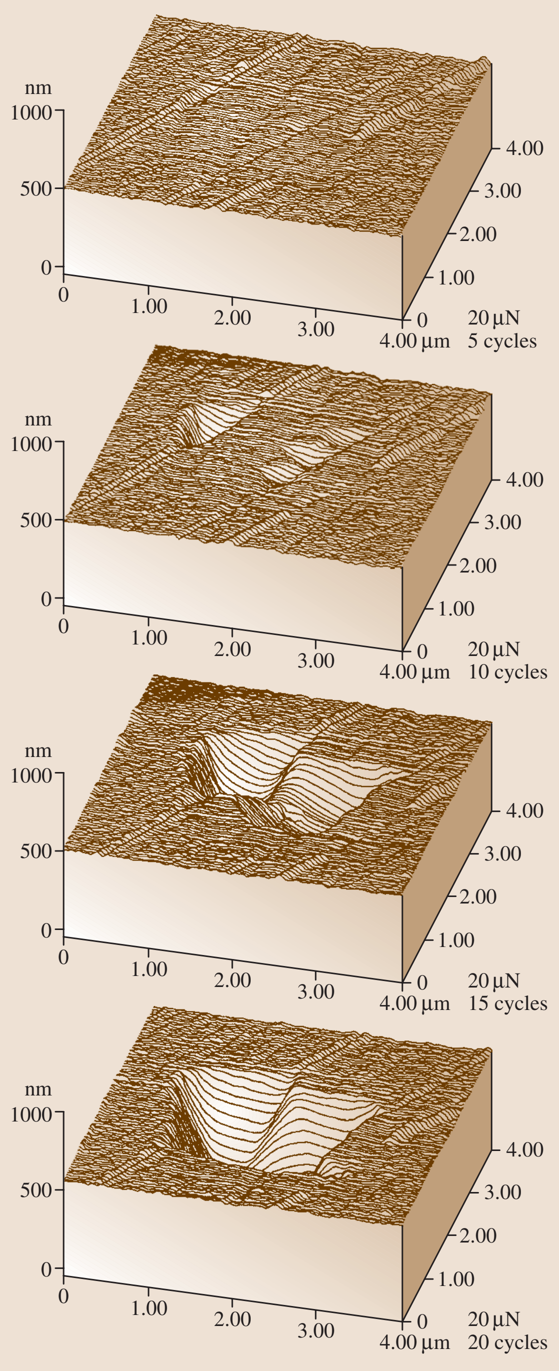

Fig. 17.32 Surface plots of diamond-like carbon-coated thin-film disk showing the worn region; the normal load and number of test cycles are indicated [17.22]

to the non-abraded area. The black region in the lower right-hand part of the topography scan shows a step that was created during the polishing phase. There is no potential contrast between the high region and the low region of the sample, indicating that the technique is independent of surface height. Figure 17.33b shows a close-up scan of the upper (low load) wear region in Fig. 17.33a. Notice that while there is no detectable change in the surface topography, there is, nonetheless, a large change in the potential of the surface in the worn region. Indeed, the wear mark of Fig. 17.33b might not have been visible at all in the topography map were it not for the noted absence of wear debris generated nearby and then swept off during the low load scan. Thus, even in the case of zero wear (no mea-

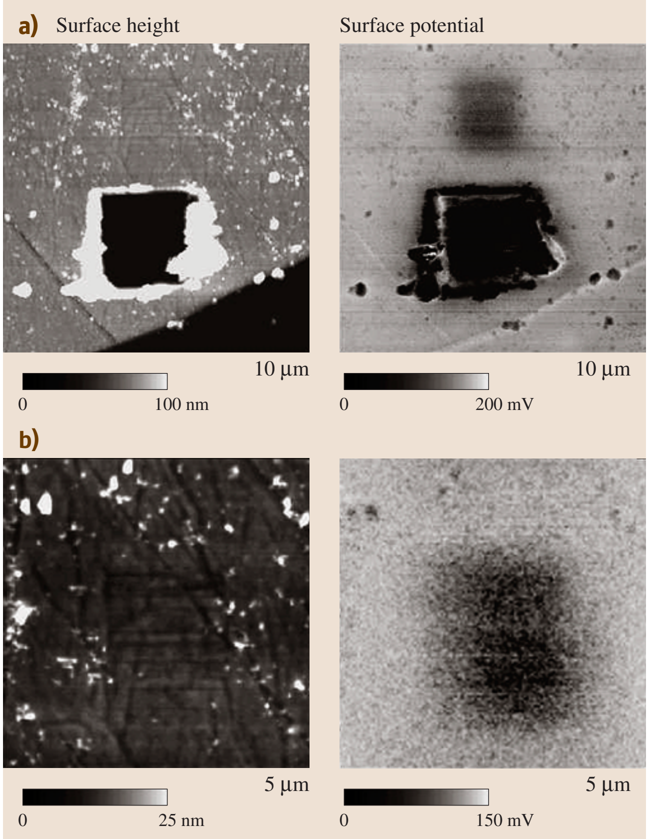

Fig. 17.33 (a) Surface height and change in surface potential maps of wear regions generated at 1 µN (*top*) and 9 µN (*bottom*) on a single-crystal aluminum sample showing bright contrast in the surface potential map on the worn regions. **(b)** Close up of upper (low load) wear region [17.38]

surable deformation of the surface using AFM), there can be a significant change in the surface potential inside the wear mark, which is useful for the study of wear precursors. It is believed that the removal of the thin contaminant layer including the natural oxide layer gives rise to the initial change in surface potential. The structural changes, which precede generation of wear debris and/or measurable wear scars, occur under ultralow loads in the top few nanometers of the sample and are primarily responsible for the subsequent changes in surface potential.

17.3.4 In Situ Characterization of Local Deformation

In situ surface characterization of local deformation of materials and thin films is carried out using a tensile stage inside an AFM. Failure mechanisms of polymeric thin films were studied by *Bobji* and *Bhushan* [17.41, 42]. The specimens were strained at a rate of 4×10^{-3}% per second, and AFM images were captured at different strains up to about 10% to monitor generation and propagation of cracks and deformation bands.

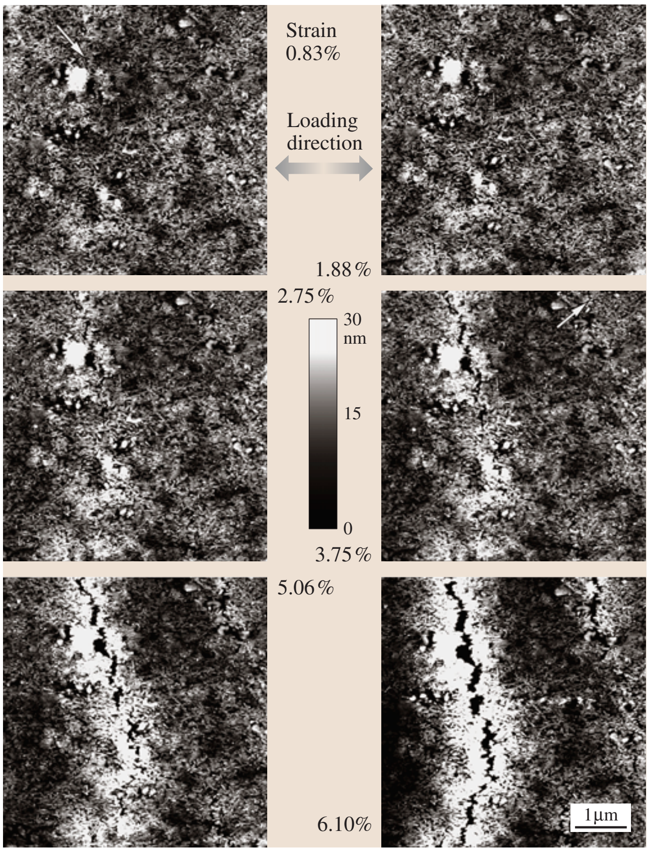

Fig. 17.34 Topographical images of the MP magnetic tape at different strains [17.41]

Bobji and *Bhushan* [17.41, 42] studied three magnetic tapes with thickness ranging from 7 to 8.5 µm. One of these was with acicular-shaped metal particle (MP) coating, and the other two with metal-evaporated (ME) coating and with and without a thin diamond-like carbon (DLC) overcoat, both on a polymeric substrate and all with particulate back coating [17.13]. They also studied the polyethylene terephthalate (PET) substrate with 6 µm thickness. They reported that cracking of the coatings started at about 1% strain for all tapes, much before the substrate starts to yield at about 2% strain. Figure 17.34 shows the topographical images of the MP tape at different strains. At 0.83% strain, a crack can be seen originating at the marked point. As the tape is further stretched along the direction, as shown in Fig. 17.34, the crack propagates along the shorter boundary of the ellipsoidal particle. However, the general direction of the crack propagation remains perpendicular to the direction of the stretching. The length, width, and depth of the cracks increase with strain, and at the same time, newer cracks keep on nucleating and propagating with reduced crack spacing. At 3.75% strain, another crack can be seen nucleating. This crack continues to grow parallel to the first. When the tape is unloaded after stretching up to a strain of about 2%, i. e., within the elastic limit of the substrate, the cracks rejoin perfectly and it is impossible to determine the difference from the unstrained tape.

Figure 17.35 shows topographical images of the three magnetic tapes and the PET substrate after being strained to 3.75%, which is well beyond the elastic limit of the substrate. MP tape develops short and numerous cracks perpendicular to the direction of loading. In tapes with metallic coating, the cracks extend throughout the tape width. In ME tape with DLC coating, there is a bulge in the coating around the primary cracks that are initiated when the substrate is still elastic, like crack A in the figure. The white band on the right-hand side of the figure is the bulge of another crack. The secondary cracks like B and C are generated at higher strains and are straighter compared to the primary cracks. In ME tape, which has a Co-O film on PET substrate, with a thickness ratio of 0.03, both with and without DLC coating, no difference is observed in the rate of growth between primary and secondary cracks. The failure is cohesive with no

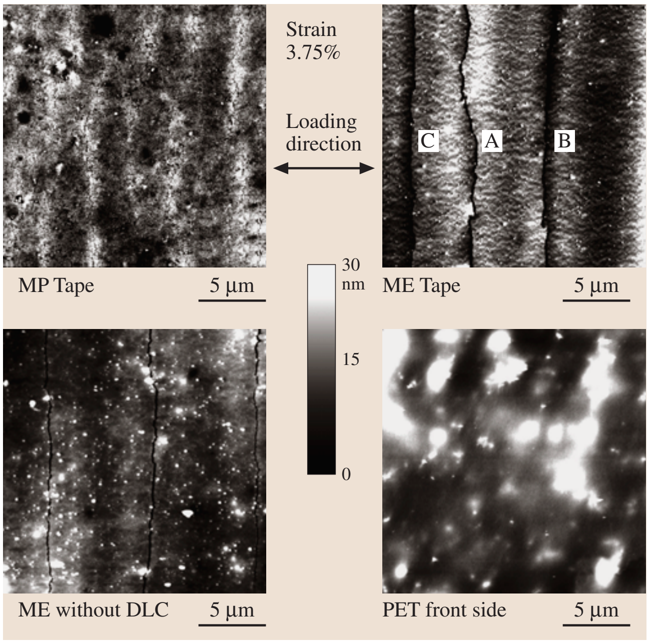

Fig. 17.35 Comparison of crack morphologies at 3.75% strain in three magnetic tapes and PET substrate. Cracks B and C, nucleated at higher strains, are more linear than crack A [17.42]

bulging of the coating. This seems to suggest that the DLC coating has residual stresses that relax when the coating cracks, causing delamination. Since the stresses are already relaxed, the secondary crack does not result in delamination. The presence of the residual stress is confirmed by the fact that a freestanding ME tape curls up (in a cylindrical form with its axis perpendicular to the tape length) with a radius of curvature of about 6 mm; the ME tape without the DLC does not curl. The magnetic coating side of PET substrate is much smoother at smaller scan lengths. However, in 20 µm scans it has a lot of bulging out, which appears as white spots in the figure. These spots change shape even while scanning the samples in tapping mode at very low contact forces.

The variation of average crack width and average crack spacing with strain is plotted in Fig. 17.36. The crack width is measured at a spot along a given crack over a distance of 1 µm in the 5 µm scan image at different strains. The crack spacing is obtained by averaging the inter-crack distance measured in five separate 50 µm scans at each strain. It can be seen that the cracks nucleate at a strain of about 0.7–1.0%, well within the elastic limit of the substrate. There is a definite change in the slope of the load-displacement curve at the strain where cracks nucleate, and the slope after that is closer to the slope of the elastic portion of the substrate. This would mean that most of the load is supported by the substrate once the coating fails by cracking.

In situ surface characterization of unstretched and stretched films has been used to measure Poisson's ratio of polymeric thin films [17.66]. Uniaxial tension is applied by the tensile stage. Surface height profiles obtained from the AFM images of unstretched and stretched samples are used to monitor simultaneously the changes in displacements of the polymer films in the longitudinal and lateral directions.

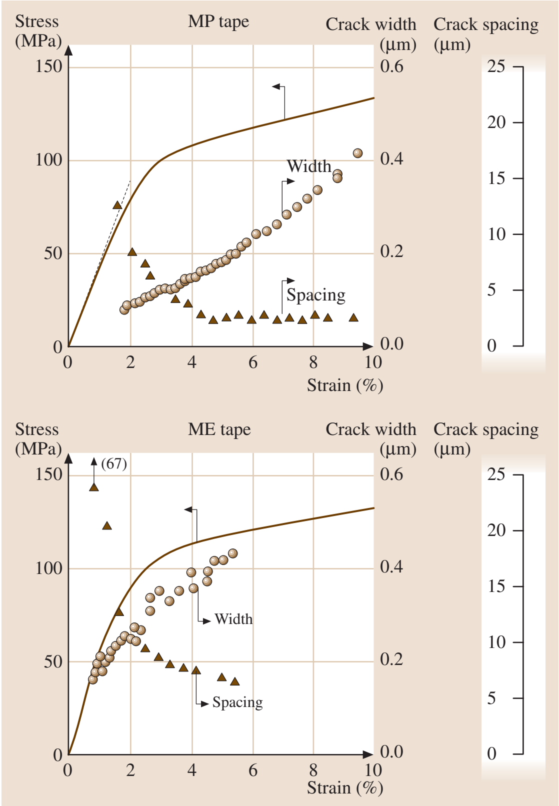

Fig. 17.36 Variation of stress, crack width, and crack spacing with strain in three magnetic tapes and PET substrate [17.41]

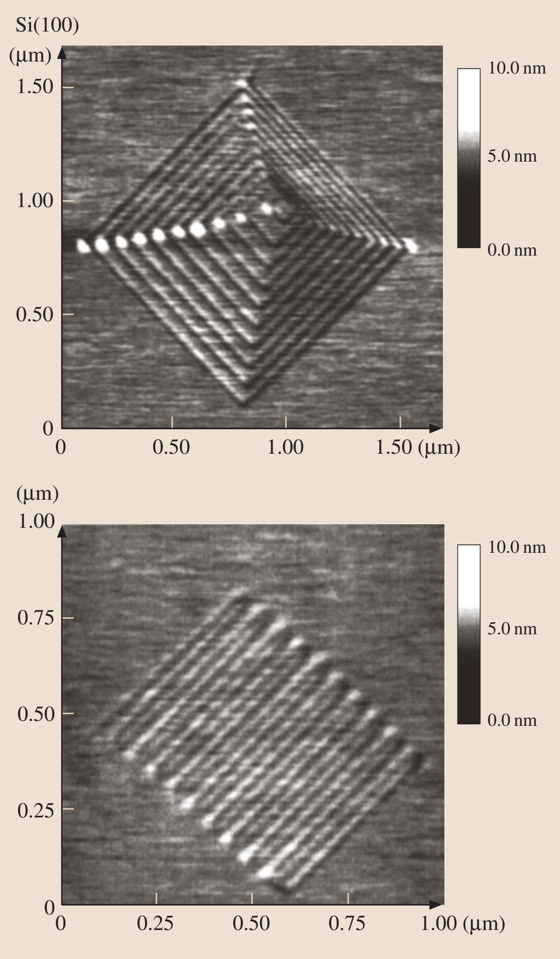

Fig. 17.37 (a) Trim and **(b)** spiral patterns generated by scratching a Si(100) surface using a diamond tip at a normal load of 15 µN and writing speed of 0.5 µm/s

17.3.5 Nanofabrication/Nanomachining

An AFM can be used for nanofabrication/nanomachining by extending the microscale scratching operation [17.4, 11,22,37]. Figure 17.37 shows two examples of nanofabrication. The patterns were created on a single-crystal silicon (100) wafer by scratching the sample surface with a diamond tip at specified locations and scratching angles. Each line is scribed manually at a normal load of 15 µN and a writing speed of 0.5 µm/s. The separation between lines is about 50 nm, and the variation in line width is due to the tip asymmetry. Nanofabrication parameters – normal load, scanning speed, and tip geometry – can be controlled precisely to control the depth and length of the devices.

Nanofabrication using mechanical scratching has several advantages over other techniques. Better control over the applied normal load, scan size, and scanning speed can be used for nanofabrication of devices. Using the technique, nanofabrication can be performed on any engineering surface. Use of chemical etching or reactions is not required, and this dry nanofabrication process can be used where use of chemicals and electric field is prohibited. One disadvantage of this technique is the formation of debris during scratching. At light loads, debris formation is not a problem compared to high-load scratching. However, debris can be removed easily from the scan area at light loads during scanning.

17.4 Indentation

Mechanical properties such as hardness and Young's modulus of elasticity can be determined on micro- to picoscales using the AFM [17.18,22,31,34] and a depth-sensing indentation system used in conjunction with an AFM [17.35,67–69].

17.4.1 Picoindentation

Indentability on the scale of subnanometers of soft samples can be studied in the force calibration mode (Fig. 17.6) by monitoring the slope of cantilever deflection as a function of sample traveling distance after the tip is engaged and the sample is pushed against the tip. For a rigid sample, cantilever deflection equals the sample traveling distance, but the former quantity is smaller if the tip indents the sample. In an example of a polymeric magnetic tape shown in Fig. 17.38, the line in the left portion of the figure is curved with a slope of less than 1 shortly after the sample touches the tip, which suggests that the tip has indented the sample [17.18]. Later, the slope is equal to 1, suggesting that the tip no longer indents the sample. This observation indicates that the tape surface is soft locally (polymer-rich) but hard (as a result of magnetic particles) underneath. Since

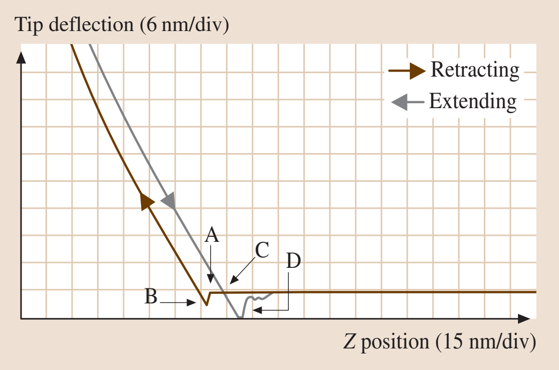

Fig. 17.38 Tip deflection (normal load) as a function of the Z (separation distance) curve for a polymeric magnetic tape [17.18]

the curves in extending and retracting modes are identical, the indentation is elastic up to a maximum load of about 22 nN used in the measurements.

Detection of transfer of material on a nanoscale is possible with the AFM. Indentation of C_{60}-rich fullerene films with an AFM tip has been shown [17.70] to result in the transfer of fullerene molecules to the AFM tip, as indicated by discontinuities in the cantilever deflection as a function of sample traveling distance in subsequent indentation studies.

17.4.2 Nanoscale Indentation

The indentation hardness of surface films with an indentation depth as small as about 1 nm can be measured using an AFM [17.11, 34, 35]. Figure 17.39 shows the gray-scale plots of indentation marks made on Si(111) at normal loads of 60, 65, 70, and 100 µN. Triangular indents can be clearly observed with very shallow depths. At a normal load of 60 µN, indents are observed and the depth of penetration is about 1 nm. As the normal load is increased, the indents become clearer and indentation depth increases. For the case of hardness measurements at shallow depths on the same order as variations in surface roughness, it is desirable to subtract the original (unindented) map from the indent map for accurate measurement of the indentation size and depth [17.22].

To make accurate measurements of hardness at shallow depths, a depth-sensing indentation system is used [17.35]. Figure 17.40 shows the load-displacement curves at different peak loads for Si(100). Loading/unloading curves often exhibit sharp discontinuities, particularly at high loads. Discontinuities, referred to as pop-ins, during the initial loading part of the curve mark a sharp transition from pure elastic loading to a plastic deformation of the specimen surface, thus correspond-

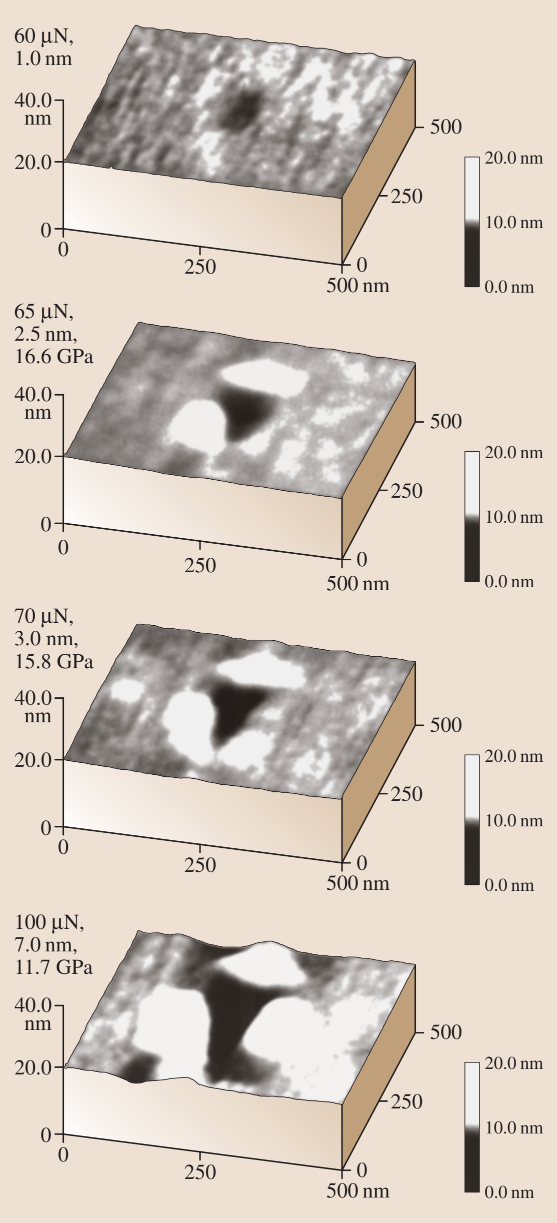
60 µN,
1.0 nm
40.0
nm
20.0
0
0
250
500 nm
0
250
500
20.0 nm
10.0 nm
0.0 nm
65 µN,
2.5 nm,
16.6 GPa
40.0
nm
20.0
0
0
250
500 nm
0
250
500
20.0 nm
10.0 nm
0.0 nm
70 µN,
3.0 nm,
15.8 GPa
40.0
nm
20.0
0
0
250
500 nm
0
250
500
20.0 nm
10.0 nm
0.0 nm
100 µN,
7.0 nm,
11.7 GPa
40.0
nm
20.0
0
0
250
500 nm
0
250
500
20.0 nm
10.0 nm
0.0 nm

Fig. 17.39 Gray-scale plots of indentation marks on the Si(111) sample at various indentation loads. Loads, indentation depths, and hardness values are listed in the figure [17.34]

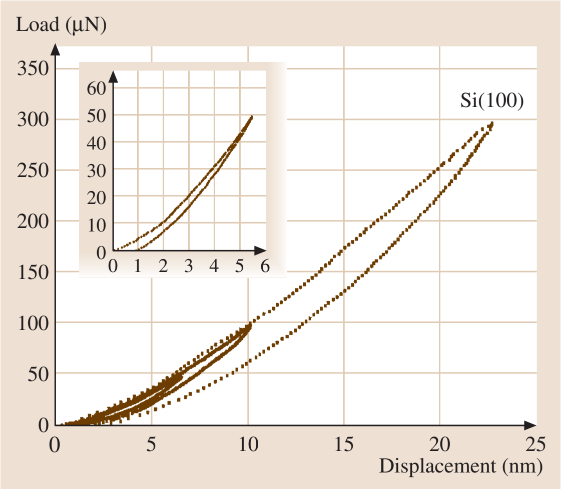

Fig. 17.40 Load-displacement curves at various peak loads for Si(100) [17.35]

ing to an initial yield point. The sharp discontinuities in the unloading part of the curves are believed to be due to the formation of lateral cracks that form at the base of the median crack, which results in the surface of the specimen being thrust upward. Load-displacement data at residual depths as low as about 1 nm can be obtained, and the indentation hardness of surface films has been measured for Si(100) [17.35, 67–69]. The hardness of silicon on a nanoscale is found to be higher than on a microscale (Fig. 17.41). Microhardness has also been reported to be higher than on the millimeter scale by several investigators. The data reported to date show that hardness exhibits size effect. According to the strain gradient plasticity theory advanced by *Fleck* et al. [17.71], large strain gradients inherent in small indentations lead to the accumulation of geometrically necessary dislocations for strain compatibility reasons that cause enhanced hardening. In addition, the decrease in hardness with an increase in indentation depth can possibly be rationalized on the basis that as the volume of deformed material increases, there is a higher probability of encountering material defects.

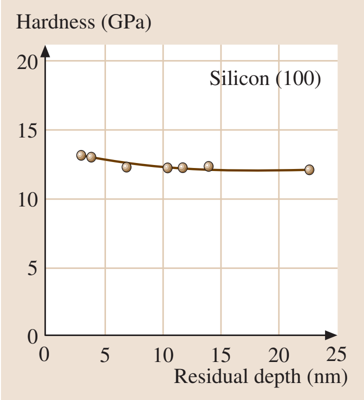

Fig. 17.41 Indentation hardness as a function of residual indentation depth for Si(100) [17.35]

Bhushan and *Koinkar* [17.31] have used AFM measurements to show that ion implantation of silicon surfaces increases their hardness and thus their wear resistance. Formation of surface alloy films with improved mechanical properties by ion implantation is of growing technological importance as a means of improving the mechanical properties of materials. Hardness of 20 nm-thick diamond-like carbon films has been measured by *Kulkarni* and *Bhushan* [17.69].

The creep and strain-rate effects (viscoelastic effects) of ceramics can be studied using a depth-sensing indentation system. *Bhushan* et al. [17.35] and *Kulkarni* and *Bhushan* [17.67–69] have reported that ceramics exhibit significant plasticity and creep on a nanoscale. Figure 17.42a shows the load-displacement curves for single-crystal silicon at various peak loads held at 180 s. To demonstrate the creep effects, the load-displacement curves for a 500 µN peak load held at 0 and 30 s are also shown as an inset. Note that significant creep occurs at room temperature. Nanoindenter experiments conducted by *Li* et al. [17.72] exhibited significant creep only at high temperatures (greater than or equal to 0.25 times the melting point of silicon). The mechanism of dislocation glide plasticity is believed to dominate the indentation creep process on the macroscale. To study the strain-rate sensitivity of silicon, data at two different (constant) rates of loading are presented in Fig. 17.42b. Note that a change in the loading rate by a factor of about five results in a significant change in the load-displacement data. The viscoelastic effects observed here for silicon at ambient temperature could arise from size effects mentioned earlier. Most likely, creep and strain rate experiments are being conducted on the hydrated films present on the silicon surface in ambient environment, and these films are expected to be viscoelastic.

17.4.3 Localized Surface Elasticity and Viscoelasticity Mapping

The Young's modulus of elasticity is calculated from the slope of the indentation curve during unloading. However, these measurements provide a single-point measurement. By using the force modulation technique, it is possible to get localized elasticity maps of soft and

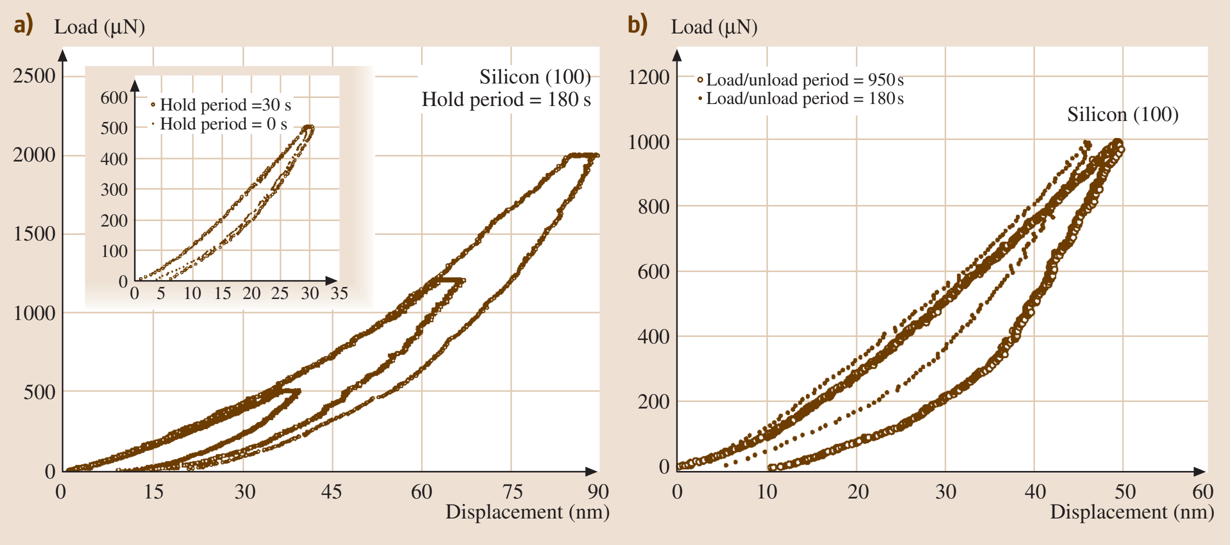

Fig. 17.42 **(a)** Creep behavior and **(b)** strain-rate sensitivity of Si(100) [17.35]

compliant materials with penetration depths of less than 100 nm. This technique has been successfully used for polymeric magnetic tapes, which consist of magnetic and nonmagnetic ceramic particles in a polymeric matrix. Elasticity maps of a tape can be used to identify relative distribution of hard magnetic and nonmagnetic ceramic particles on the tape surface, which has an effect on friction and stiction at the head-tape interface [17.13]. Figure 17.43 shows surface height and elasticity maps on a polymeric magnetic tape [17.44]. The elasticity image reveals sharp variations in surface elasticity due to the composite nature of the film. As can be clearly seen, regions of high elasticity do not always correspond to high or low topography. Based on a Hertzian elastic-contact analysis, the static indentation depth of these samples during the force modulation scan is estimated to be about 1 nm. We conclude that the contrast seen is influenced most strongly by material properties in the top few nanometers, independent of the composite structure beneath the surface layer.

By using phase contrast microscopy, it is possible to get phase contrast maps or the contrast in viscoelastic properties of near surface regions. This technique has been used successfully for polymeric films and magnetic tapes that consist of ceramic particles in a polymeric matrix [17.51, 52]. Figure 17.44 shows topography, phase angle, and friction force images for polyethylene terephthalate (PET) film. Three particles, which show up as approximately circular, lightly shaded (high points) regions, are seen in the topography image. In the phase angle images, the particles show up dark against the film background. A combination of high tapping amplitude and low set point was found to maximize contrast in phase images. This result may be explained by the degree to which the tip is able to penetrate and thus deform the sample. A high tapping amplitude and low set point should maximize the force with which the tip strikes the sample. This should give more penetration. So the mater-

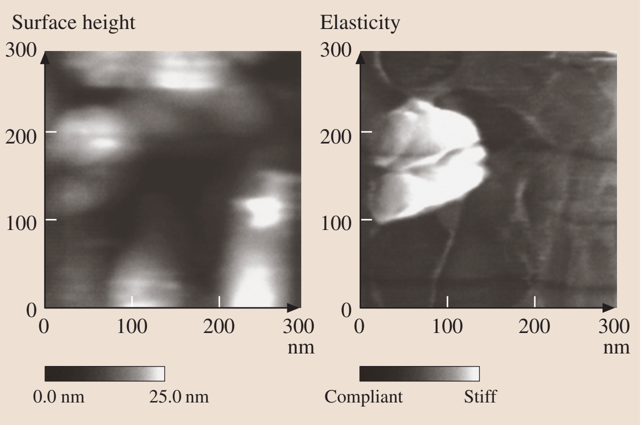

Fig. 17.43 Surface height and elasticity maps on a polymeric magnetic tape ($\sigma = 6.72$ nm and P-V = 31.7 nm; σ and P-V refer to standard deviation of surface heights and peak-to-valley distance, respectively). The gray scale on the elasticity map is arbitrary [17.44]

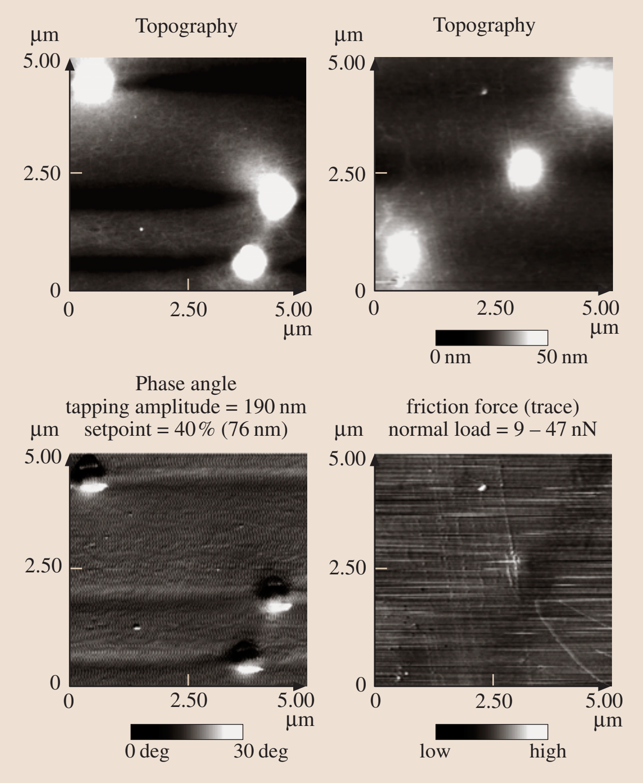

Fig. 17.44 Tapping mode height and phase angle images acquired using a TESP tip at a tapping amplitude of 190 nm and a setpoint of 40% (74 nm), and contact mode height and friction force (trace) images using an MESP tip with a normal load between 9 and 47 nN for PET film with embedded ceramic particles [17.51]

ial or viscoelastic contribution, as opposed to the surface force and adhesion hysteresis contributions, should be more dominant for this tapping condition. The phase angle image shown in Fig. 17.44 was acquired using tapping mode with a relatively high tapping amplitude of 190 nm and a relatively low set point of 40% (of the tapping amplitude) to emphasize viscoelastic properties. Very little correlation is found between the phase angle and friction force images. Friction force images were obtained using an etched single-crystal silicon tip (DI MESP) with a radius of 20–50 nm, a stiffness of 1–5 N/m, and a natural frequency of 60–70 kHz. A lack of correlation may be a further indication that something other than adhesion, which should correlate to friction force, causes the phase angle contrast; this may imply that viscoelastic properties may be dominant. These results also indicate that phase angle imaging yields information that cannot be obtained using other, more conventional AFM modes.

17.5 Boundary Lubrication

17.5.1 Perfluoropolyether Lubricants

The classical approach to lubrication uses freely supported multimolecular layers of liquid lubricants [17.10, 13, 58, 73]. The liquid lubricants are sometimes chemically bonded to improve their wear resistance [17.10, 13, 58]. Partially chemically bonded, molecularly thick perfluoropolyether (PFPE) films are used for lubrication of magnetic storage media [17.13]. These are considered potential candidate lubricants for micro/nanoelectromechanical systems (MEMS/NEMS). Molecularly thick PFPEs are well suited because of the following properties: low surface tension and low contact angle, which allow easy spreading on surfaces and provide hydrophobic properties; chemical and thermal stability, which minimize degradation under use; low vapor pressure, which provides low out-gassing; high adhesion to substrate via organic functional bonds; and good lubricity, which reduces contact surface wear.

For boundary lubrication studies, friction, adhesion, and durability experiments have been performed on virgin Si(100) surfaces and silicon surfaces lubricated with two PFPE lubricants – Z-15 (with $-CF_3$ nonpolar end groups) and Z-DOL (with $-OH$ polar end groups) [17.27, 28, 30]. Z-DOL film was thermally bonded at 150 °C for 30 minutes and an unbonded fraction was removed by a solvent (BW) [17.13]. The thicknesses of Z-15 and Z-DOL films were 2.8 nm and 2.3 nm, respectively.

The adhesive forces of Si(100), Z-15, and Z-DOL (BW) measured by force calibration plot and friction force versus normal load plot are summarized in Fig. 17.45. The results measured by these two methods are in good agreement. Figure 17.45 shows that the presence of mobile Z-15 lubricant film increases the adhesive force as compared to that of Si(100) by meniscus formation. The presence of solid phase Z-DOL (BW) film reduces the adhesive force as compared to that of

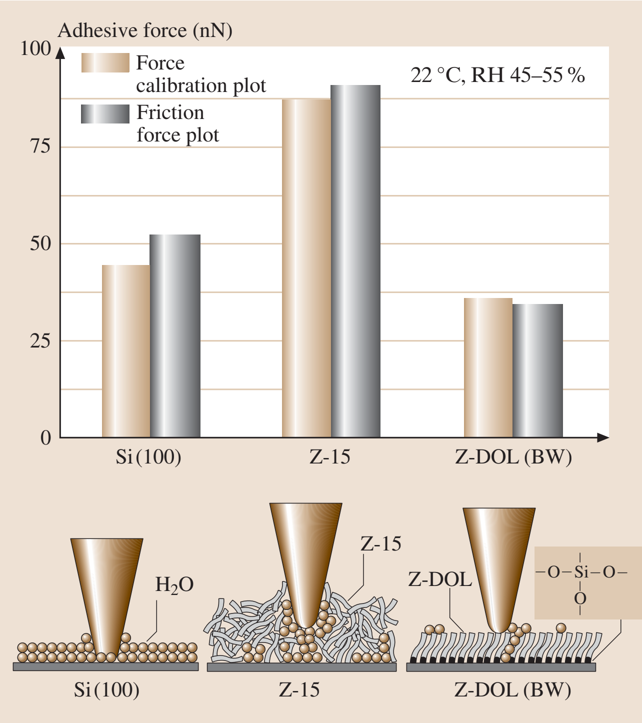

Fig. 17.45 Summary of the adhesive forces of Si(100) and Z-15 and Z-DOL (BW) films measured by force calibration plots and friction force versus normal load plots in ambient air. The schematic (*bottom*) showing the effect of meniscus formed between the AFM tip and surface sample on the adhesive and friction forces [17.30]

Si(100) because of the absence of mobile liquid. The schematic (bottom) in Fig. 17.45 shows relative size and sources of meniscus. It is well-known that the native oxide layer (SiO_2) on the top of a Si(100) wafer exhibits hydrophilic properties, and some water molecules can be adsorbed on this surface. The condensed water will form meniscus as the tip approaches the sample surface. The larger adhesive force in Z-15 is not only caused by the Z-15 meniscus; the nonpolarized Z-15 liquid does not have good wettability and strong bonding with Si(100). In the ambient environment, the condensed water molecules from the environment will permeate through the liquid Z-15 lubricant film and compete with the lubricant molecules present on the substrate. The interaction of the liquid lubricant with the substrate is weakened, and a boundary layer of the liquid lubricant forms puddles [17.27, 28]. This de-wetting allows water molecules to be adsorbed on the Si(100) surface as aggregates along with Z-15 molecules. And both of them can form meniscus while the tip approaches the surface. Thus, the de-wetting of liquid Z-15 film results in higher adhesive force and poorer lubrication performance. In addition, as the Z-15 film is pretty soft compared to the solid Si(100) surface, and penetration of the tip in the film occurs while pushing the tip down. This leads the large area of the tip involved to form the meniscus at the tip-liquid (mixture of Z-15 and water) interface. It should also be noted that Z-15 has a higher viscosity compared to water and, therefore, Z-15 film provides higher resistance to motion and coefficient of friction. In the case of Z-DOL (BW) film, both of the active groups of Z-DOL molecules are mostly bonded on Si(100) substrate. Thus, the Z-DOL (BW) film has low free surface energy and cannot be displaced readily by water molecules or readily adsorb water molecules. Therefore, the use of Z-DOL (BW) can reduce the adhesive force.

To study the velocity effect on friction and adhesion, the variation of friction force, adhesive force, and coefficient of friction of Si(100), Z-15 and Z-DOL (BW) as a function of velocity are summarized in Fig. 17.46. It indicates that for silicon wafer the friction force decreases logarithmically with increasing velocity. For Z-15, the friction force decreases with increasing velocity up to 10 μm/s, after which it remains almost constant. The velocity has very small effect on the friction force of Z-DOL (BW): It reduced slightly only at very high velocity. Figure 17.46 also indicates that the adhesive force of Si(100) increases when the velocity is higher than 10 μm/s. The adhesive force of Z-15 is reduced dramatically with a velocity increase up to 20 μm/s, after which it is reduced slightly. And the adhesive force of Z-DOL (BW) also decreases at high velocity. In the testing range of velocity, only the coefficient of friction of Si(100) decreases with velocity, but the coefficients of friction of Z-15 and Z-DOL (BW) almost remain constant. This implies that the friction mechanisms of Z-15 and Z-DOL (BW) do not change with the variation of velocity.

The mechanisms of the effect of velocity on the adhesion and friction are explained based on schematics shown in Fig. 17.46 (right). For Si(100), tribochemical reaction plays a major role. Although at high velocity the meniscus is broken and does not have enough time to rebuild, the contact stresses and high velocity lead to tribochemical reactions of Si(100) wafer (which has native oxide (SiO_2)) and a Si_3N_4 tip with water molecules and form $Si(OH)_4$. The $Si(OH)_4$ is removed and continuously replenished during sliding. The $Si(OH)_4$ layer between the tip and Si(100) surface is known to be of low shear strength and causes a decrease in friction force and

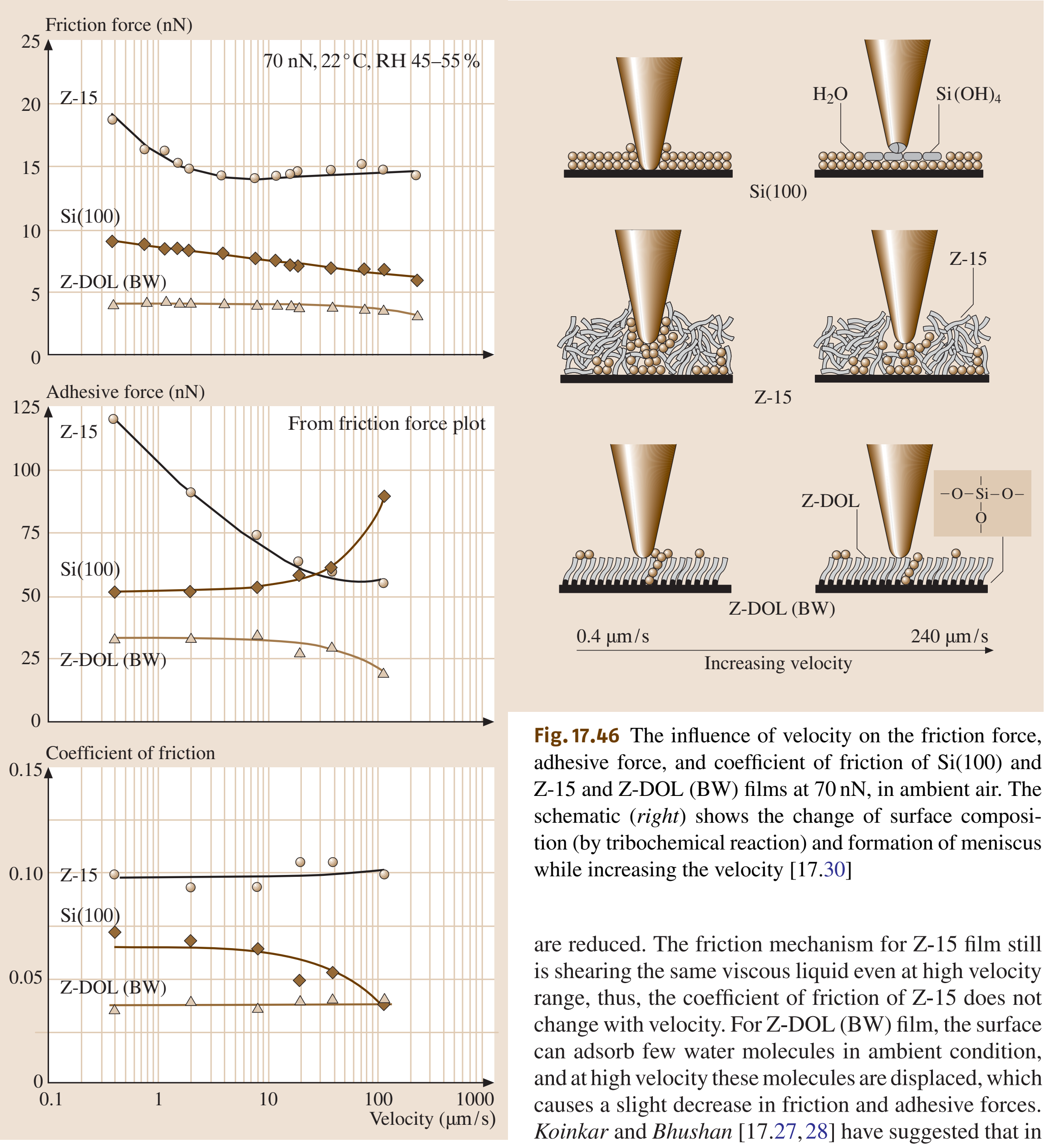

Fig. 17.46 The influence of velocity on the friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, in ambient air. The schematic (*right*) shows the change of surface composition (by tribochemical reaction) and formation of meniscus while increasing the velocity [17.30]

coefficient of friction [17.10,58]. The chemical bonds of Si−OH between the tip and Si(100) surface induce large adhesive force. For Z-15 film, at high velocity the meniscus formed by condensed water and Z-15 molecules is broken and does not have enough time to rebuild. Therefore, the adhesive force and, consequently, friction force are reduced. The friction mechanism for Z-15 film still is shearing the same viscous liquid even at high velocity range, thus, the coefficient of friction of Z-15 does not change with velocity. For Z-DOL (BW) film, the surface can adsorb few water molecules in ambient condition, and at high velocity these molecules are displaced, which causes a slight decrease in friction and adhesive forces. *Koinkar* and *Bhushan* [17.27,28] have suggested that in the case of samples with mobile films such as condensed water and Z-15 films, alignment of liquid molecules (shear thinning) is responsible for the drop in friction force with an increase in scanning velocity. This could be another reason for the decrease in friction force for Si(100) and Z-15 film with velocity in this study.

To study the relative humidity effect on friction and adhesion, the variations of friction force, adhesive force,

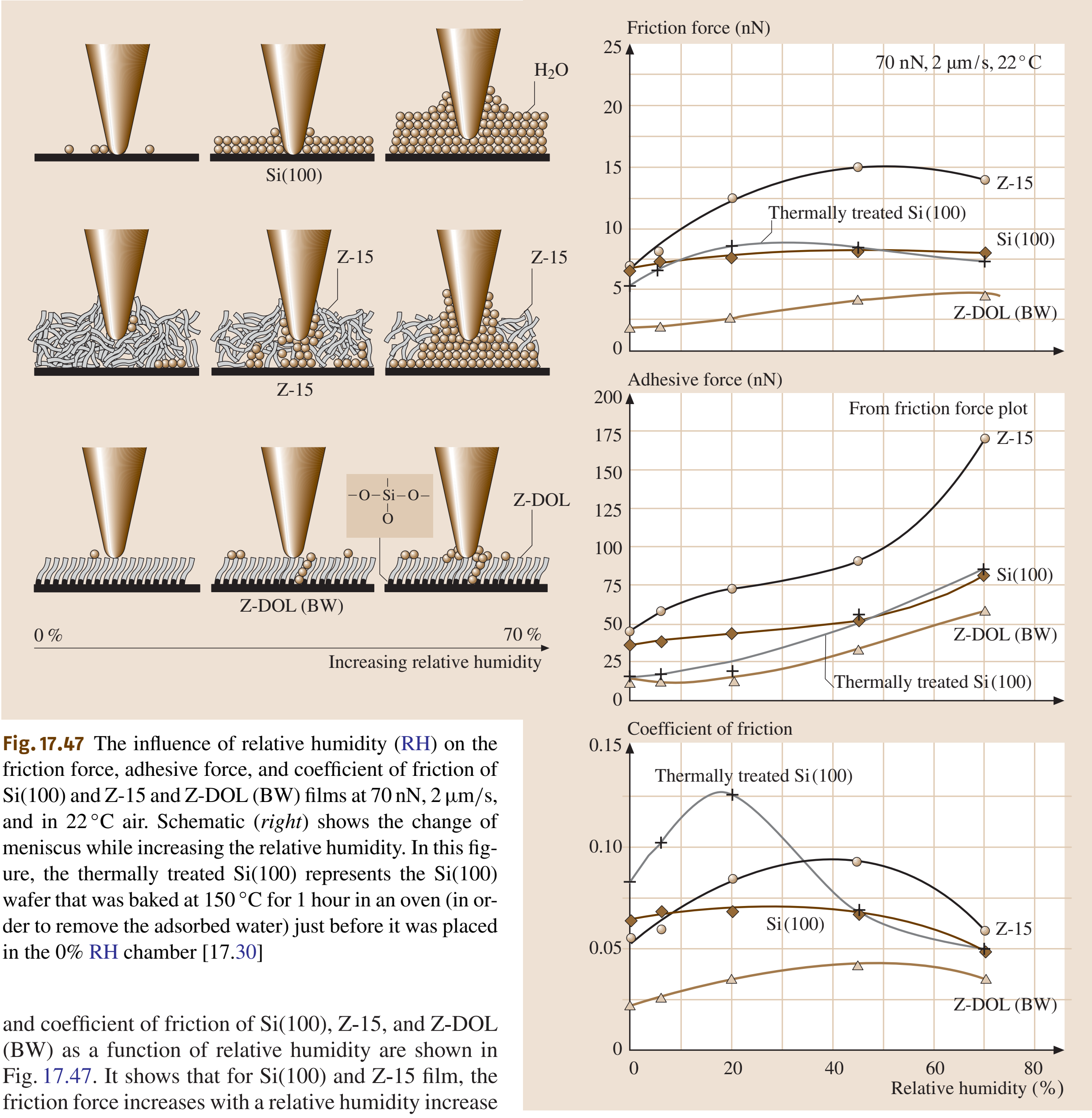

Fig. 17.47 The influence of relative humidity (RH) on the friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, 2 µm/s, and in 22 °C air. Schematic (*right*) shows the change of meniscus while increasing the relative humidity. In this figure, the thermally treated Si(100) represents the Si(100) wafer that was baked at 150 °C for 1 hour in an oven (in order to remove the adsorbed water) just before it was placed in the 0% RH chamber [17.30]

and coefficient of friction of Si(100), Z-15, and Z-DOL (BW) as a function of relative humidity are shown in Fig. 17.47. It shows that for Si(100) and Z-15 film, the friction force increases with a relative humidity increase of up to 45%, and then slightly decreases with a further increase in the relative humidity. Z-DOL (BW) has a smaller friction force than Si(100) and Z-15 in the whole testing range. And its friction force shows a relative apparent increase when the relative humidity is higher than 45%. For Si(100), Z-15, and Z-DOL (BW), their adhesive forces increase with relative humidity, and their coefficients of friction increase with relative humidity up to 45%, after which they decrease with further increasing of the relative humidity. It is also observed that the humidity effect on Si(100) really depends on the history of the Si(100) sample. As the surface of Si(100) wafer readily adsorbs water in air, without any pre-treatment the Si(100) used in our study almost

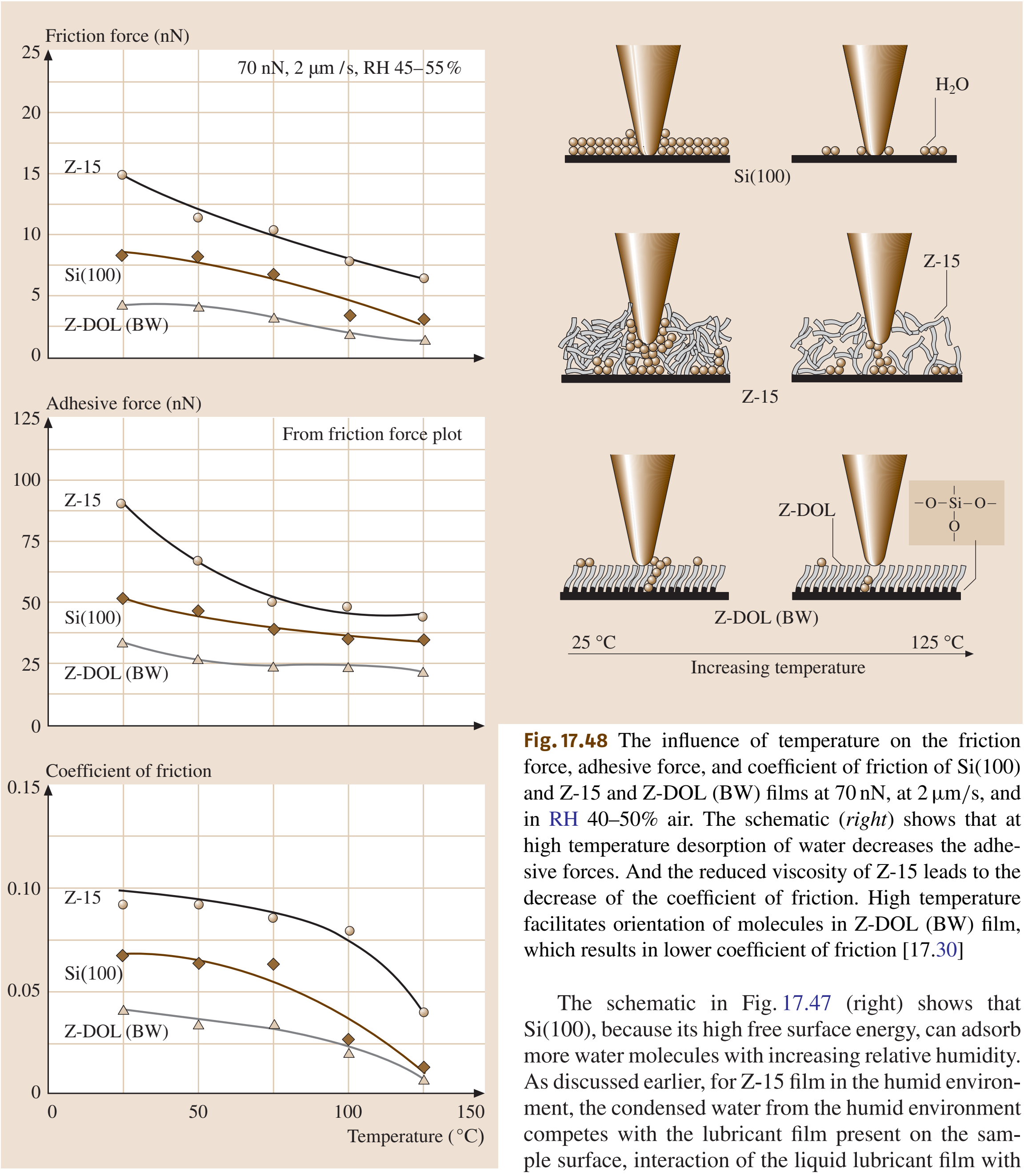

Fig. 17.48 The influence of temperature on the friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, at 2 μm/s, and in RH 40–50% air. The schematic (*right*) shows that at high temperature desorption of water decreases the adhesive forces. And the reduced viscosity of Z-15 leads to the decrease of the coefficient of friction. High temperature facilitates orientation of molecules in Z-DOL (BW) film, which results in lower coefficient of friction [17.30]

reaches to its saturate stage of adsorbed water and is responsible for less effect during increasing relative humidity. However, once the Si(100) wafer was thermally treated by baking at 150 °C for 1 hour, a bigger effect was observed.

The schematic in Fig. 17.47 (right) shows that Si(100), because its high free surface energy, can adsorb more water molecules with increasing relative humidity. As discussed earlier, for Z-15 film in the humid environment, the condensed water from the humid environment competes with the lubricant film present on the sample surface, interaction of the liquid lubricant film with the silicon substrate is weakened, and a boundary layer of the liquid lubricant forms puddles. This dewetting allows water molecules to be adsorbed on the Si(100) substrate mixed with Z-15 molecules [17.27, 28]. Obviously, more water molecules can be adsorbed on Z-15 surface while increasing relative humidity. The more adsorbed water molecules in the case of Si(100), along with lubricant molecules in Z-15 film case, form bigger water meniscus which leads to an increase of friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 with humidity. But at a very high humidity of 70%, large quantities of adsorbed water can form a water layer that separates the tip and sample surface and act as a kind of lubricant, which causes a decrease in the friction force and coefficient of friction. For Z-DOL (BW) film, because of its hydrophobic surface properties, water molecules can be adsorbed at a humidity higher than 45% and causes an increase in the adhesive and friction forces.

To study the temperature effect on friction and adhesion, the variations of friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL (BW) as a function of temperature are summarized in Fig. 17.48. It shows that the increasing temperature causes a decrease in friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL (BW). The schematic (right) in Fig. 17.48 indicates that at high temperature, desorption of water leads to the decrease of friction force, adhesive forces, and coefficient of friction for all of the samples. For Z-15 film, the reduction of viscosity at high temperature also contributes to the decrease of friction force and coefficient of friction. In the case of Z-DOL (BW) film, molecules are more easily oriented at high temperature, which may be partly responsible for the low friction force and coefficient of friction.

To study the durability of lubricant films at the nanoscale, the friction of Si(100), Z-15, and Z-DOL (BW) as a function of the number of scanning cycles is shown in Fig. 17.49. As observed earlier, friction force and coefficient of friction of Z-15 is higher than that of Si(100) with the lowest values for Z-DOL (BW). During cycling, friction force and coefficient of friction of Si(100) show a slight decrease during the initial few cycles then remain constant. This is related to the removal of the top adsorbed layer. In the case of Z-15 film, the friction force and coefficient of friction show an increase during the initial few cycles, and then approach higher and stable values. This is believed to be caused by the attachment of the Z-15 molecules to the tip. The molecular interaction between these attached molecules to the tip and molecules on the film surface is responsible for an increase in friction. But after several scans, this molecular interaction reaches the equilibrium, and after that, friction force and coefficient of friction remain constant. In the case of Z-DOL (BW) film, the friction force and coefficient of friction start out low and remain

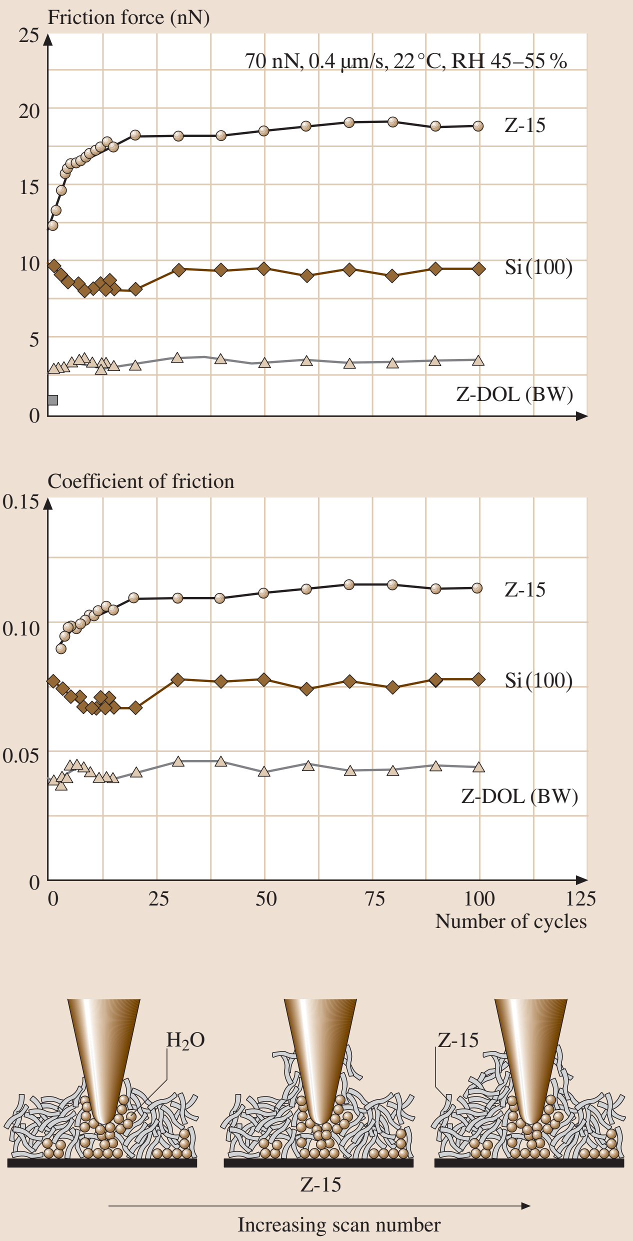

Fig. 17.49 Friction force and coefficient of friction versus number of sliding cycles for Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, 0.8 μm/s, and in ambient air. Schematic (*bottom*) shows that some liquid Z-15 molecules can be attached to the tip. The molecular interaction between the attached molecules to the tip with the Z-15 molecules in the film results in an increase of the friction force with multiple scanning [17.30]

low during the entire test for 100 cycles, suggesting that Z-DOL (BW) molecules do not get attached or displaced as readily as Z-15 molecules.

As a brief summary, the influence of velocity, relative humidity, and temperature on the friction force of mobile Z-15 film is presented in Fig. 17.50. The changing trends are also addressed in this figure.

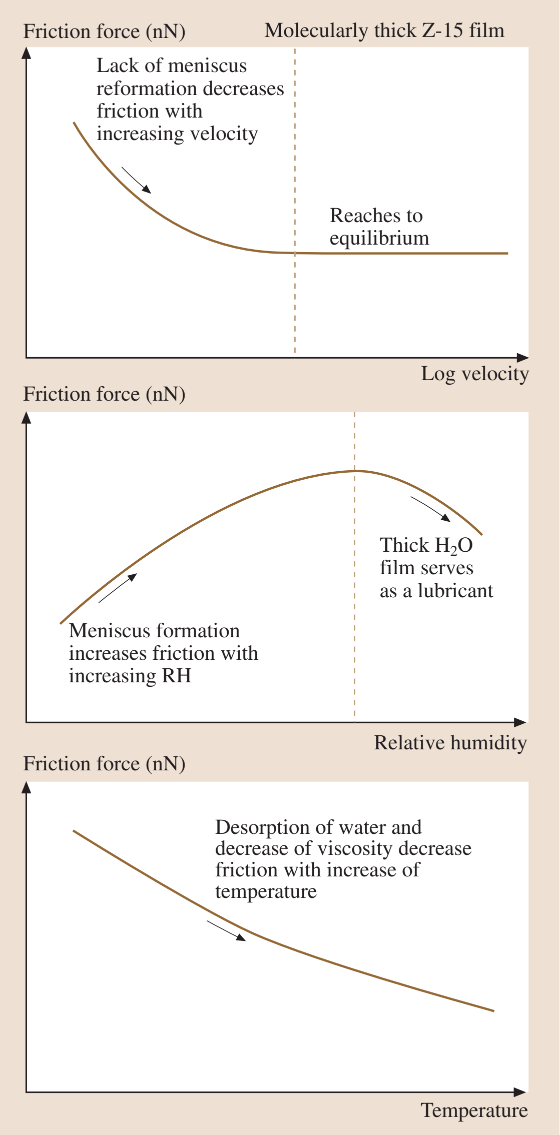

Fig. 17.50 Schematic shows the change of friction force of molecularly thick Z-15 films with log velocity, relative humidity, and temperature. The changing trends are also addressed in this figure [17.30]

17.5.2 Self-Assembled Monolayers

For lubrication of MEMS/NEMS, another effective approach involves the deposition of organized and dense molecular layers of long-chain molecules. Two common methods to produce monolayers and thin films are the Langmuir–Blodgett (L-B) deposition and self-assembled monolayers (SAMs) by chemical grafting of molecules. L-B films are physically bonded to the substrate by weak van der Waals attraction, while SAMs are chemically bonded via covalent bonds to the substrate. Because of the choice of chain length and terminal linking group that SAMs offer, they hold great promise for boundary lubrication of MEMS/NEMS. A number of studies have been conducted to study tribological properties of various SAMs [17.26, 29, 74–76]. It has been reported that SAMs with high-compliance long carbon chains exhibit low friction; chain compliance is desirable for low friction. Based on [17.29], the friction mechanism of SAMs is explained by a so-called "molecular spring" model (Fig. 17.51). According to this model, the chemically adsorbed self-assembled molecules on a substrate are just like assembled molecular springs anchored to the substrate. An asperity sliding on the surface of SAMs is like a tip sliding on the top of "molecular springs or brush". The molecular spring assembly has compliant features and can experience orientation and compression under load. The orientation of the molecular springs or brush under normal load reduces the shearing force at the interface, which, in turn, reduces the friction force.

The SAMs with high-compliance long carbon chains also exhibit the best wear resistance [17.29, 75]. In wear experiments, the wear depth as a function of normal load curves shows critical normal loads (Fig. 17.52). Below the critical normal load SAMs undergo orien-

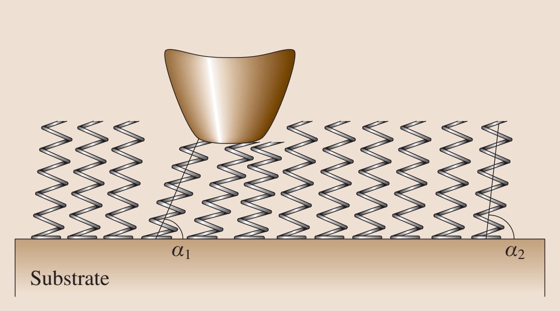

Fig. 17.51 Molecular spring model of SAMs. In this figure, $\alpha_1 < \alpha_2$, which is caused by the further orientation under the normal load applied by an asperity tip [17.29]

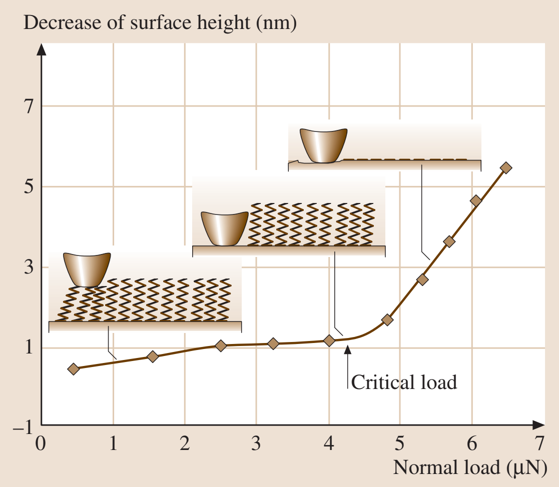

Fig. 17.52 Illustration of the wear mechanism of SAMs with increasing normal load [17.75]

tation, at the critical load SAMs wear away from the substrate due to weak interface bond strengths, while above the critical normal load severe wear takes place on the substrate.

17.5.3 Liquid Film Thickness Measurements

Liquid film thickness measurement of thin lubricant films (on the order of 10 nm or thicker) with nanometer lateral resolution can be made with the AFM [17.4, 59]. The lubricant thickness is obtained by measuring the force on the tip as it approaches, contacts, and pushes through the liquid film and ultimately contacts the substrate. The distance between the sharp snap-in (owing to the formation of a liquid meniscus between the film and the tip) at the liquid surface and the hard repulsion at the substrate surface is a measure of the liquid film thickness.

Lubricant film thickness mapping of ultrathin films (on the order of 2 nm) can be obtained using friction force microscopy [17.27] and adhesive force mapping [17.36]. Figure 17.53 shows gray-scale plots of the surface topography and friction force obtained simultaneously for unbonded Demnum-type PFPE lubricant film on silicon. The friction force plot shows well distinguished low and high friction regions roughly corresponding to high and low regions in surface topography (thick and thin lubricant regions). A uniformly lubricated sample does not show such a variation in the friction. Friction force imaging can thus be used to measure the lubricant uniformity on the sample surface, which cannot be identified by surface topography alone. Figure 17.54 shows the gray-scale plots of the adhesive force distribution for silicon samples coated uniformly and nonuniformly with Z-DOL type PFPE lubricant. It can be clearly seen that a region exists that has an adhesive force distinctly different from the other region for the nonuniformly coated sample. This implies that the liquid film thickness is nonuniform, giving rise to a difference in the meniscus forces.

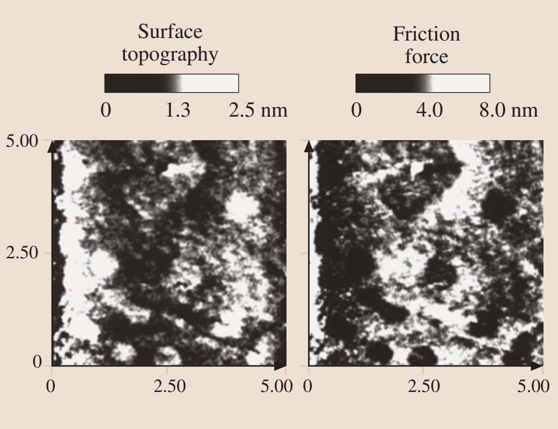

Fig. 17.53 Gray-scale plots of the surface topography and friction force obtained simultaneously for unbonded Demnum-type perfluoropolyether lubricant film on silicon [17.27]

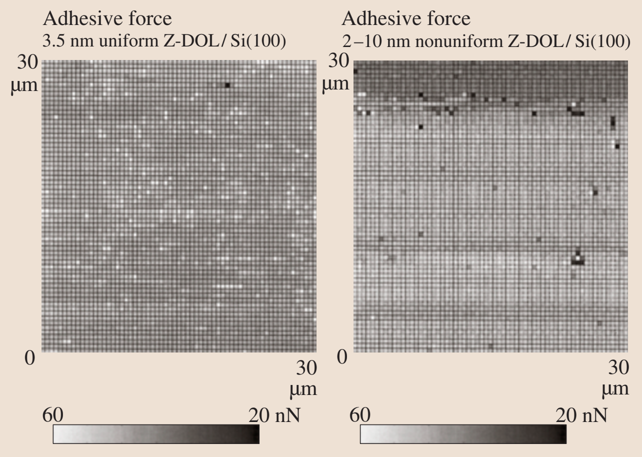

Fig. 17.54 Gray-scale plots of the adhesive force distribution of a uniformly coated, 3.5-nm thick unbonded Z-DOL film on silicon and 3- to 10-nm thick unbonded Z-DOL film on silicon that was deliberately coated nonuniformly by vibrating the sample during the coating process [17.36]

17.6 Closure

At most solid-solid interfaces of technological relevance, contact occurs at many asperities. A sharp AFM/FFM tip sliding on a surface simulates just one such contact. However, asperities come in all shapes and sizes. The effect of radius of a single asperity (tip) on the friction/adhesion performance can be studied using tips of different radii. AFM/FFM are used to study various tribological phenomena, including surface roughness, adhesion, friction, scratching, wear, indentation, detection of material transfer, and boundary lubrication. Measurement of atomic-scale friction of a freshly cleaved, highly oriented pyrolytic graphite exhibits the same periodicity as that of the corresponding topography. However, the peaks in friction and those in the corresponding topography are displaced relative to one another. Variations in atomic-scale friction and the observed displacement can be explained by the variation in interatomic forces in the normal and lateral directions. Local variations in microscale friction occur and are found to correspond to the local slopes, suggesting that a ratchet mechanism and collision effects are responsible for this variation. Directionality in the friction is observed on both micro- and macroscales, resulting from the surface roughness and surface preparation. Anisotropy in surface roughness accentuates this effect. Microscale friction is generally found to be smaller than the macrofriction, as there is less plowing contribution in microscale measurements. Microscale friction is load-dependent, and friction values increase with an increase in the normal load, approaching the macrofriction at contact stresses higher than the hardness of the softer material. The tip radius also affects adhesion and friction.

Mechanism of material removal on the microscale is studied. Wear rate for single-crystal silicon is negligible below 20 μN and is much higher and remains approximately constant at higher loads. Elastic deformation at low loads is responsible for negligible wear. Most of the wear debris is loose. SEM and TEM studies of the wear region suggest that the material on the microscale is removed by plastic deformation with a small contribution from elastic fracture; this observation corroborates the scratch data. Evolution of wear has also been studied using AFM. Wear is found to be initiated at nanoscratches. For a sliding interface requiring near-zero friction and wear, contact stresses should be below the hardness of the softer material to minimize plastic deformation, and surfaces should be free of nanoscratches. Furthermore, wear precursors can be detected at early stages of wear by using surface potential measurements. It is found that even in the case of zero wear (no measurable deformation of the surface using AFM), there can be a significant change in the surface potential inside the wear mark, which is useful for study of wear precursors. Detection of material transfer on a nanoscale is possible with AFM.

In situ surface characterization of local deformation of materials and thin coatings can be carried out using a tensile stage inside an AFM. An AFM can also be used for nanofabrication/nanomachining.

By using the force modulation technique, localized surface elasticity maps of composite materials with penetrating depths of less than 100 nm can be obtained. It is possible using phase contrast microscopy to get phase contrast maps or the contrast in viscoelastic properties of near surface regions. Scratching and indentation on nanoscales are powerful ways to screen for adhesion and resistance to deformation of ultrathin films. Modified AFM can be used to obtain load-displacement curves and for measurement of nanoindentation hardness and Young's modulus of elasticity, with depth of indentation as low as 1 nm. Hardness of ceramics on the nanoscale is found to be higher than that on the microscale. Ceramics exhibit significant plasticity and creep on a nanoscale.

Boundary lubrication studies and measurement of lubricant-film thickness with a lateral resolution on a nanoscale can be conducted using AFM. Chemically bonded lubricant films and self-assembled monolayers are superior in friction and wear resistance. For chemically bonded lubricant films, the adsorption of water, the formation of meniscus, and its change during sliding, viscosity, and surface properties play an important role on the friction, adhesion, and durability of these films. For SAMs, their friction mechanism is explained by a so-called "molecular spring" model. The films with high-compliance long carbon chains exhibit low friction and wear.

Investigations of wear, scratching, and indentation on the nanoscale using the AFM can provide insights into failure mechanisms of materials. Coefficients of friction, wear rates, and mechanical properties such as hardness have been found to be different on the nanoscale than on the macroscale; generally, coefficients of friction and wear rates on micro- and nanoscales are smaller, whereas hardness is greater. Therefore, micro/nanotribological studies may help define the regimes for ultralow friction and near-zero wear. These studies also provide insight into the atomic origins of adhesion, friction, wear, and lubrication mechanisms.

References

17.1 I. L. Singer, H. M. Pollock: *Fundamentals of Friction: Macroscopic and Microscopic Processes*, Nato Sci. Ser. E, Vol. 220 (Kluwer, Dordrecht 1992)

17.2 B. N. J. Persson, E. Tosatti: *Physics of Sliding Friction*, Nato Sci. Ser. E, Vol. 311 (Kluwer, Dordrecht 1996)

17.3 B. Bhushan: *Micro/Nanotribology and Its Applications*, Nato Sci. Ser. E, Vol. 330 (Kluwer, Dordrecht 1997)

17.4 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

17.5 B. Bhushan: Nanoscale tribophysics and tribomechanics, Wear **225–229** (1999) 465–492

17.6 B. Bhushan: Wear and mechanical characterisation on micro- to picoscales using AFM, Int. Mat. Rev. **44** (1999) 105–117

17.7 B. Bhushan: *Modern Tribology Handbook, Vol. 1: Principles of Tribology* (CRC, Boca Raton 2001)

17.8 B. Bhushan: *Fundamentals of tribology and bridging the gap between the macro- and micro/nanoscales*, NATO Sci. Ser. E, Vol. 10 (Kluwer, Dordrecht 2001)

17.9 B. Bhushan: Nano- to microscale wear and mechanical characterization studies using scanning probe microscopy, Wear **251** (2001) 1105–1123

17.10 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

17.11 B. Bhushan, J. N. Israelachvili, U. Landman: Nanotribology: Friction, wear and lubrication at the atomic scale, Nature **374** (1995) 607–616

17.12 H. J. Guntherodt, D. Anselmetti, E. Meyer: *Forces in Scanning Probe Methods*, Nato Sci. Ser. E, Vol. 286 (Kluwer, Dordrecht 1995)

17.13 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)

17.14 B. Bhushan: *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)

17.15 G. Binnig, C. F. Quate, Ch. Gerber: Atomic force microscopy, Phys. Rev. Lett. **56** (1986) 930–933

17.16 G. Binnig, Ch. Gerber, E. Stoll, T. R. Albrecht, C. F. Quate: Atomic resolution with atomic force microscope, Europhys. Lett. **3** (1987) 1281–1286

17.17 C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang: Atomic-scale friction of a tungsten tip on a graphite surface, Phys. Rev. Lett. **59** (1987) 1942–1945

17.18 B. Bhushan, J. Ruan: Atomic-scale friction measurements using friction force microscopy: part II – application to magnetic media, ASME J. Tribol. **116** (1994) 389–396

17.19 J. Ruan, B. Bhushan: Atomic-scale friction measurements using friction force microscopy: part I – general principles and new measurement techniques, ASME J. Tribol. **116** (1994) 378–388

17.20 J. Ruan, B. Bhushan: Atomic-scale and microscale friction of graphite and diamond using friction force microscopy, J. Appl. Phys. **76** (1994) 5022–5035

17.21 J. Ruan, B. Bhushan: Frictional behavior of highly oriented pyrolytic graphite, J. Appl. Phys. **76** (1994) 8117–8120

17.22 B. Bhushan, V. N. Koinkar, J. Ruan: Microtribology of magnetic media, Proc. Inst. Mech. Eng., Part J: J. Eng. Tribol. **208** (1994) 17–29

17.23 B. Bhushan, S. Sundararajan: Micro/nanoscale friction and wear mechanisms of thin films using atomic force and friction force microscopy, Acta Mater. **46** (1998) 3793–3804

17.24 V. Scherer, W. Arnold, B. Bhushan: Active friction control using ultrasonic vibration. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 463–469

17.25 V. Scherer, W. Arnold, B. Bhushan: Lateral force microscopy using acoustic friction force microscopy, Surf. Interface Anal. **27** (1999) 578–587

17.26 B. Bhushan, A. V. Kulkarni, V. N. Koinkar, M. Boehm, L. Odoni, C. Martelet, M. Belin: Microtribological characterization of self-assembled and Langmuir–Blodgett monolayers by atomic and friction force microscopy, Langmuir **11** (1995) 3189–3198

17.27 V. N. Koinkar, B. Bhushan: Micro/nanoscale studies of boundary layers of liquid lubricants for magnetic disks, J. Appl. Phys. **79** (1996) 8071–8075

17.28 V. N. Koinkar, B. Bhushan: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy, J. Vac. Sci. Technol. A. **14** (1996) 2378–2391

17.29 B. Bhushan, H. Liu: Nanotribological properties and mechanisms of alkylthiol and biphenyl thiol self-assembled monolayers studied by AFM, Phys. Rev. B **63** (2001) 245412-1–245412-11

17.30 H. Liu, B. Bhushan: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM, Ultramicroscopy **97** (2003) 321–340

17.31 B. Bhushan, V. N. Koinkar: Tribological studies of silicon for magnetic recording applications, J. Appl. Phys. **75** (1994) 5741–5746

17.32 V. N. Koinkar, B. Bhushan: Microtribological properties of hard amorphous carbon protective coatings for thin film magnetic disks and heads, Proc. Inst. Mech. Eng. Part J: J. Eng. Tribol. **211** (1997) 365–372

17.33 S. Sundararajan, B. Bhushan: Development of a continuous microscratch technique in an atomic force microscope and its application to study scratch resistance of ultra-thin hard amorphous carbon coatings, J. Mater. Res. **16** (2001) 75–84

17.34 B. Bhushan, V. N. Koinkar: Nanoindentation hardness measurements using atomic force microscopy, Appl. Phys. Lett. **64** (1994) 1653–1655

17.35 B. Bhushan, A. V. Kulkarni, W. Bonin, J. T. Wyrobek: Nano/picoindentation measurement using a capacitance transducer system in atomic force microscopy, Philos. Mag. **74** (1996) 1117–1128

17.36 B. Bhushan, C. Dandavate: Thin-film friction and adhesion studies using atomic force microscopy, J. Appl. Phys. **87** (2000) 1201–1210

17.37 B. Bhushan: Micro/nanotribology and its applications to magnetic storage devices and MEMS, Tribol. Int. **28** (1995) 85–95

17.38 D. DeVecchio, B. Bhushan: Use of a nanoscale Kelvin probe for detecting wear precursors, Rev. Sci. Instrum. **69** (1998) 3618–3624

17.39 B. Bhushan, A.V. Goldade: Measurements and analysis of surface potential change during wear of single crystal silicon (100) at ultralow loads using Kelvin probe microscopy, Appl. Surf. Sci **157** (2000) 373–381

17.40 B. Bhushan, A.V. Goldade: Kelvin probe microscopy measurements of surface potential change under wear at low loads, Wear **244** (2000) 104–117

17.41 M.S. Bobji, B. Bhushan: Atomic force microscopic study of the micro-cracking of magnetic thin films under tension, Scripta Mater. **44** (2001) 37–42

17.42 M.S. Bobji, B. Bhushan: In-situ microscopic surface characterization studies of polymeric thin films during tensile deformation using atomic force microscopy, J. Mater. Res. **16** (2001) 844–855

17.43 P. Maivald, H.J. Butt, S.A.C. Gould, C.B. Prater, B. Drake, J.A. Gurley, V.B. Elings, P.K. Hansma: Using force modulation to image surface elasticities with the atomic force microscope, Nanotechnol. **2** (1991) 103–106

17.44 D. DeVecchio, B. Bhushan: Localized surface elasticity measurements using an atomic force microscope, Rev. Sci. Instrum. **68** (1997) 4498–4505

17.45 V. Scherer, B. Bhushan, U. Rabe, W. Arnold: Local elasticity and lubrication measurements using atomic force and friction force microscopy at ultrasonic frequencies, IEEE Trans. Mag. **33** (1997) 4077–4079

17.46 H. U. Krotil, T. Stifter, H. Waschipky, K. Weishaupt, S. Hild, O. Marti: Pulse force mode: A new method for the investigation of surface properties, Surf. Interface Anal. **27** (1999) 336–340

17.47 S. Amelio, A.V. Goldade, U. Rabe, V. Scherer, B. Bhushan, W. Arnold: Measurements of elastic properties of ultra-thin diamond-like carbon coatings using atomic force acoustic microscopy, Thin Solid Films **392** (2001) 75–84

17.48 B. Anczykowski, D. Kruger, K. L. Babcock, H. Fuchs: Basic properties of dynamic force microscopy with the scanning force microscope in experiment and simulation, Ultramicroscopy **66** (1996) 251–259

17.49 J. Tamayo, R. Garcia: Deformation, contact time, and phase contrast in tapping mode scanning force microscopy, Longmuir **12** (1996) 4430–4435

17.50 R. Garcia, J. Tamayo, M. Calleja, F. Garcia: Phase contrast in tapping-mode scanning force microscopy, Appl. Phys. A **66** (1998) 309–312

17.51 W.W. Scott, B. Bhushan: Use of phase imaging in atomic force microscopy for measurement of viscoelastic contrast in polymer nanocomposites and molecularly-thick lubricant films, Ultramicroscopy **97** (2003) 151–169

17.52 B. Bhushan, J. Qi: Phase contrast imaging of nanocomposites and molecularly- thick lubricant films in magnetic media, Nanotechnol. **14** (2003) 886–895

17.53 V. N. Koinkar, B. Bhushan: Microtribological studies of Al_2O_3-TiC, polycrystalline and single-crystal Mn-Zn ferrite and SiC head slider materials, Wear **202** (1996) 110–122

17.54 E. Meyer, R. Overney, R. Luthi, D. Brodbeck, L. Howald, J. Frommer, H.J. Guntherodt, O. Wolter, M. Fujihira, T. Takano, Y. Gotoh: Friction force microscopy of mixed Langmuir–Blodgett films, Thin Solid Films **220** (1992) 132–137

17.55 C. D. Frisbie, L. F. Rozsnyai, A. Noy, M.S. Wrighton, C.M. Lieber: Functional group imaging by chemical force microscopy, Science **265** (1994) 2071–2074

17.56 V.N. Koinkar, B. Bhushan: Effect of scan size and surface roughness on microscale friction measurements, J. Appl. Phys. **81** (1997) 2472–2479

17.57 S. Sundararajan, B. Bhushan: Topography-induced contributions to friction forces measured using an atomic force/friction force microscope, J. Appl. Phys. **88** (2000) 4825–4831

17.58 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

17.59 B. Bhushan, G.S. Blackman: Atomic force microscopy of magnetic rigid disks and sliders and its applications to tribology, ASME J. Tribol. **113** (1991) 452–458

17.60 M. Reinstaedtler, U. Rabe, V. Scherer, U. Hartmann, A. Goldade, B. Bhushan, W. Arnold: On the nanoscale measurement of friction using atomic-force microscope cantilever torsional resonances, Appl. Phys. Lett. **82** (2003) 2604–2606

17.61 T. Stifter, O. Marti, B. Bhushan: Theoretical investigation of the distance dependence of capillary and van der Waals forces in scanning probe microscopy, Phys. Rev. B **62** (2000) 13667–13673

17.62 U. D. Schwarz, O. Zwoerner, P. Koester, R. Wiesendanger: Friction force spectroscopy in the low-load regime with well-defined tips. In: *Micro/Nanotribology and Its Applications*, ed. by B. Bhushan (Kluwer, Dordrecht 1997) pp. 233–238

17.63 B. Bhushan, A. V. Kulkarni: Effect of normal load on microscale friction measurements, Thin Solid Films **278** (1996) 49–56; Errata: **293** 333

17.64 V. N. Koinkar, B. Bhushan: Scanning and transmission electron microscopies of single-crystal silicon microworn/machined using atomic force microscopy, J. Mater. Res. **12** (1997) 3219–3224

17.65 X. Zhao, B. Bhushan: Material removal mechanism of single-crystal silicon on nanoscale and at ultralow loads, Wear **223** (1998) 66–78

17.66 B. Bhushan, P. S. Mokashi, T. Ma: A new technique to measure Poisson's ratio of ultrathin polymeric films using atomic force microscopy, Rev. Sci. Instrumen. **74** (2003) 1043–1047

17.67 A. V. Kulkarni, B. Bhushan: Nanoscale mechanical property measurements using modified atomic force microscopy, Thin Solid Films **290–291** (1996) 206–210

17.68 A. V. Kulkarni, B. Bhushan: Nano/picoindentation measurements on single-crystal aluminum using modified atomic force microscopy, Mater. Lett. **29** (1996) 221–227

17.69 A. V. Kulkarni, B. Bhushan: Nanoindentation measurement of amorphous carbon coatings, J. Mater. Res. **12** (1997) 2707–2714

17.70 J. Ruan, B. Bhushan: Nanoindentation studies of fullerene films using atomic force microscopy, J. Mater. Res. **8** (1993) 3019–3022

17.71 N. A. Fleck, G. M. Muller, M. F. Ashby, J. W. Hutchinson: Strain gradient plasticity: Theory and experiment, Acta Metall. Mater. **42** (1994) 475–487

17.72 W. B. Li, J. L. Henshall, R. M. Hooper, K. E. Easterling: The mechanism of indentation creep, Acta Metall. Mater. **39** (1991) 3099–3110

17.73 F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids*, Vol. 1 (Clarendon, Oxford 1950)

17.74 H. Liu, B. Bhushan: Investigation of the adhesion, friction, and wear properties of biphenyl thiol self-assembled monolayers by atomic force microscopy, J. Vac. Sci. Technol. A **19** (2001) 1234–1240

17.75 H. Liu, B. Bhushan: Investigation of nanotribological properties of self-assembled monolayers with alkyl and biphenyl spacer chains, Ultramicroscopy **91** (2002) 185–202

17.76 B. Bhushan: Self-assembled monolayers for controlling hydrophobicity and/or friction and wear. In: *Modern Tribology Handbook, Vol. 2: Materials, Coatings, and Industrial Applications*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 909–929

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

18. Surface Forces and Nanorheology of Molecularly Thin Films

In this chapter, we describe the static and dynamic normal forces that occur between surfaces in vacuum or liquids and the different modes of friction that can be observed between (i) bare surfaces in contact (dry or interfacial friction), (ii) surfaces separated by a thin liquid film (lubricated friction), and (iii) surfaces coated with organic monolayers (boundary friction).

Experimental methods suitable for measuring normal surface forces, adhesion and friction (lateral or shear) forces of different magnitude at the molecular level are described. We explain the molecular origin of van der Waals, electrostatic, solvation and polymer mediated interactions, and basic models for the contact mechanics of adhesive and nonadhesive elastically deforming bodies. The effects of interaction forces, molecular shape, surface structure and roughness on adhesion and friction are discussed.

Simple models for the contributions of the adhesion force and external load to interfacial friction are illustrated with experimental data on both unlubricated and lubricated systems, as measured with the surface forces apparatus. We discuss rate-dependent adhesion (adhesion hysteresis) and how this is related to friction. Some examples of the transition from wearless friction to friction with wear are shown.

Lubrication in different lubricant thickness regimes is described together with explanations of nanorheological concepts. The occurrence of and transitions between smooth and stick–slip sliding in various types of dry (unlubricated and solid boundary lubricated) and liquid lubricated systems are discussed based on recent experimental results and models for stick–slip involving memory distance and dilatency.

18.1 **Introduction: Types of Surface Forces** 544

18.2 **Methods Used to Study Surface Forces** 546
  18.2.1 Force Laws 546
  18.2.2 Adhesion Forces 547
  18.2.3 The SFA and AFM 547
  18.2.4 Some Other Force-Measuring Techniques 549

18.3 **Normal Forces Between Dry (Unlubricated) Surfaces** 550
  18.3.1 Van der Waals Forces in Vacuum and Inert Vapors 550
  18.3.2 Charge Exchange Interactions 552
  18.3.3 Sintering and Cold Welding 553

18.4 **Normal Forces Between Surfaces in Liquids** 554
  18.4.1 Van der Waals Forces in Liquids 554
  18.4.2 Electrostatic and Ion Correlation Forces 554
  18.4.3 Solvation and Structural Forces 557
  18.4.4 Hydration and Hydrophobic Forces 559
  18.4.5 Polymer-Mediated Forces 561
  18.4.6 Thermal Fluctuation Forces 563

18.5 **Adhesion and Capillary Forces** 564
  18.5.1 Capillary Forces 564
  18.5.2 Adhesion Mechanics 566
  18.5.3 Effects of Surface Structure, Roughness, and Lattice Mismatch 566
  18.5.4 Nonequilibrium and Rate-Dependent Interactions: Adhesion Hysteresis 567

18.6 **Introduction: Different Modes of Friction and the Limits of Continuum Models** 569

18.7 **Relationship Between Adhesion and Friction Between Dry (Unlubricated and Solid Boundary Lubricated) Surfaces** 571
  18.7.1 Amontons' Law and Deviations from It Due to Adhesion: The Cobblestone Model 571
  18.7.2 Adhesion Force and Load Contribution to Interfacial Friction 572
  18.7.3 Examples of Experimentally Observed Friction of Dry Surfaces... 576
  18.7.4 Transition from Interfacial to Normal Friction with Wear........ 579

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

18.8 **Liquid Lubricated Surfaces** 580
18.8.1 Viscous Forces and Friction of Thick Films: Continuum Regime 580
18.8.2 Friction of Intermediate Thickness Films 582
18.8.3 Boundary Lubrication of Molecularly Thin Films: Nanorheology 584
18.9 **Role of Molecular Shape and Surface Structure in Friction** 591
References 594

18.1 Introduction: Types of Surface Forces

In this chapter, we discuss the most important types of surface forces and the relevant equations for the force and friction laws. Several different attractive and repulsive forces operate between surfaces and particles. Some forces occur in vacuum, for example, attractive van der Waals and repulsive hard-core interactions. Other types of forces can arise only when the interacting surfaces are separated by another condensed phase, which is usually a liquid. The most common types of surface forces and their main characteristics are listed in Table 18.1.

In *vacuum*, the two main long-range interactions are the attractive van der Waals and electrostatic (Coulomb) forces. At smaller surface separations (corresponding to molecular contact at surface separations of $D \approx 0.2\,\text{nm}$), additional attractive interactions can be found such as covalent or metallic bonding forces. These attractive forces are stabilized by the hard-core repulsion. Together they determine the surface and interfacial energies of planar surfaces, as well as the strengths of materials and adhesive junctions. Adhesion forces are often strong enough to elastically or plastically deform bodies or particles when they come into contact.

In *vapors* (e.g., atmospheric air containing water and organic molecules), solid surfaces in or close to contact will generally have a surface layer of chemisorbed or physisorbed molecules, or a capillary condensed liquid bridge between them. A surface layer usually causes the adhesion to decrease, but in the case of capillary condensation, the additional Laplace pressure or attractive "capillary" force may make the adhesion between the surfaces stronger than in inert gas or vacuum.

When totally immersed in a *liquid*, the force between particles or surfaces is completely modified from that in vacuum or air (vapor). The van der Waals attraction is generally reduced, but other forces can now arise that can qualitatively change both the range and even the sign of the interaction. The attractive force in such a system can be either stronger or weaker than in the absence of the intervening liquid. For example, the overall attraction can be stronger in the case of two hydrophobic surfaces separated by water, but weaker for two hydrophilic surfaces. Depending on the different forces that may be operating simultaneously in solution, the overall force law is not generally monotonically attractive even at long range; it can be repulsive, or the force can change sign at some finite surface separation. In such cases, the potential energy minimum, which determines the adhesion force or energy, occurs not at true molecular contact between the surfaces, but at some small distance farther out.

The forces between surfaces in a liquid medium can be particularly complex at *short range*, i. e., at surface separations below a few nanometers or 4–10 molecular diameters. This is partly because with increasing confinement, a liquid ceases to behave as a structureless continuum with bulk properties; instead, the size and shape of its molecules begin to determine the overall interaction. In addition, the surfaces themselves can no longer be treated as inert and structureless walls (i. e., mathematically flat) and their physical and chemical properties at the atomic scale must now be taken into account. The force laws will then depend on whether the surfaces are amorphous or crystalline (and whether the lattices of crystalline surfaces are matched or not), rough or smooth, rigid or soft (fluid-like), and hydrophobic or hydrophilic.

It is also important to distinguish between *static* (i. e., equilibrium) interactions and *dynamic* (i. e., nonequilibrium) forces such as viscous and friction forces. For example, certain liquid films confined between two contacting surfaces may take a surprisingly long time to equilibrate, as may the surfaces themselves, so that the short-range and adhesion forces appear to be time-dependent, resulting in "aging" effects.

Table 18.1 Types of surface forces in vacuum vs. in liquid (colloidal forces)

Type of force	Subclasses or alternative names	Main characteristics
	Attractive forces	
van der Waals	Debye induced dipole force (v & s) London dispersion force (v & s) Casimir force (v & s)	Ubiquitous, occurs both in vacuum and in liquids
Electrostatic	Ionic bond (v) Coulombic force (v & s) Hydrogen bond (v) Charge-exchange interaction (v & s) Acid–base interaction (s) “Harpooning” interaction (v)	Strong, long-range, arises in polar solvents; requires surface charging or charge-separation mechanism
Ion correlation	van der Waals force of polarizable ions (s)	Requires mobile charges on surfaces in a polar solvent
Quantum mechanical	Covalent bond (v) Metallic bond (v) Exchange interaction (v)	Strong, short-range, responsible for contact binding of crystalline surfaces
Solvation	Oscillatory force (s) Depletion force (s)	Mainly entropic in origin, the oscillatory force alternates between attraction and repulsion
Hydrophobic	Attractive hydration force (s)	Strong, apparently long-range; origin not yet understood
Specific binding	“Lock-and-key” or complementary binding (v & s) Receptor–ligand interaction (s) Antibody–antigen interaction (s)	Subtle combination of different non-covalent forces giving rise to highly specific binding; main recognition mechanism of biological systems
	Repulsive forces	
van der Waals	van der Waals disjoining pressure (s)	Arises only between dissimilar bodies interacting in a medium
Electrostatic	Coulombic force (v & s)	Arises only for certain constrained surface charge distributions
Quantum mechanical	Hard-core or steric repulsion (v) Born repulsion (v)	Short-range, stabilizing attractive covalent and ionic binding forces, effectively determine molecular size and shape
Solvation	Oscillatory solvation force (s) Structural force (s) Hydration force (s)	Monotonically repulsive forces, believed to arise when solvent molecules bind strongly to surfaces
Entropic	Osmotic repulsion (s) Double-layer force (s) Thermal fluctuation force (s) Steric polymer repulsion (s) Undulation force (s) Protrusion force (s)	Due to confinement of molecular or ionic species; requires mechanism that keeps trapped species between the surfaces
	Dynamic interactions	
Non-equilibrium	Hydrodynamic forces (s) Viscous forces (s) Friction forces (v & s) Lubrication forces (s)	Energy-dissipating forces occurring during relative motion of surfaces or bodies

Note: (v) applies only to interactions in vacuum, (s) applies only to interactions in solution (or to surfaces separated by a liquid), and (v & s) applies to interactions occurring both in vacuum and in solution.

18.2 Methods Used to Study Surface Forces

18.2.1 Force Laws

The full force law $F(D)$ between two surfaces, i. e., the force F as a function of surface separation D, can be measured in a number of ways [18.2–6]. The simplest is to move the base of a spring by a known amount, ΔD_0. Figure 18.1 illustrates this method when applied to the interaction of two magnets. However, the method is applicable also at the microscopic or molecular level, and it forms the basis of all direct force-measuring apparatuses such as the surface forces apparatus (SFA; [18.3, 7]) and the atomic force microscope (AFM; [18.8–10]). If there is a detectable force between the surfaces, this will cause the force-measuring spring to deflect by ΔD_s, while the surface separation changes by ΔD. These three displacements are related by

$$\Delta D_s = \Delta D_0 - \Delta D\,. \tag{18.1}$$

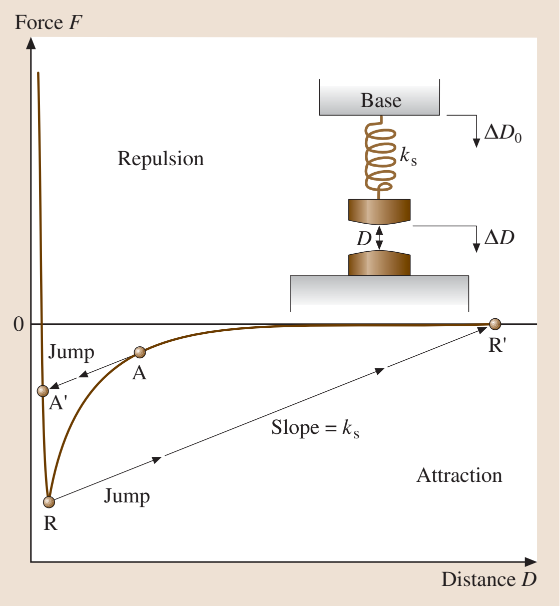

Fig. 18.1 Schematic attractive force law between two macroscopic objects such as two magnets, or between two microscopic objects such as the van der Waals force between a metal tip and a surface. On lowering the base supporting the spring, the latter will expand or contract such that at any equilibrium separation D the attractive force balances the elastic spring restoring force. If the gradient of the attractive force $\mathrm{d}F/\mathrm{d}D$ exceeds the gradient of the spring's restoring force (defined by the spring constant k_s), the upper surface will jump from A into contact at A′ (A for "advancing"). On separating the surfaces by raising the base, the two surfaces will jump apart from R to R′ (R for "receding"). The distance R−R′ multiplied by k_s gives the adhesion force, i. e., the value of F at the point R (after [18.1] with permission)

The difference in force, ΔF, between the initial and final separations is given by

$$\Delta F = k_s \Delta D_s\,, \tag{18.2}$$

where k_s is the spring constant. The equations above provide the basis for measurements of the force difference between any two surface separations. For example, if a force-measuring apparatus with a known k_s can measure D (and thus ΔD), ΔD_0, and ΔD_s, the force difference ΔF can be measured between a large initial or reference separation D, where the force is zero ($F = 0$), and another separation $D - \Delta D$. By working one's way in increasing increments of $\Delta D = \Delta D_0 - \Delta D_s$, the full force law $F(D)$ can be constructed over any desired distance regime.

In order to measure an equilibrium force law, it is essential to establish that the two surfaces have stopped moving before the displacements are measured. When displacements are measured while two surfaces are still in relative motion, one also measures a viscous or frictional contribution to the total force. Such dynamic force measurements have enabled the viscosities of liquids near surfaces and in thin films to be accurately determined [18.11–13].

In practice, it is difficult to measure the forces between two perfectly flat surfaces, because of the stringent requirement of perfect alignment for making reliable measurements at distances of a few tenths of a nanometer. It is far easier to measure the forces between curved surfaces, e.g., two spheres, a sphere and a flat surface, or two crossed cylinders. Furthermore, the force $F(D)$ measured between two curved surfaces can be directly related to the energy per unit area $E(D)$ between two flat surfaces at the same separation, D, by the so-called Derjaguin approximation [18.14]:

$$E(D) = \frac{F(D)}{2\pi R}\,, \tag{18.3}$$

where R is the radius of the sphere (for a sphere and a flat surface) or the radii of the cylinders (for two crossed cylinders).

18.2.2 Adhesion Forces

The most direct way to measure the adhesion of two solid surfaces (such as two spheres or a sphere on a flat) is to suspend one of them on a spring and measure the adhesion or "pull-off" force needed to separate the two bodies from its deflection. If k_s is the stiffness of the force-measuring spring and ΔD the distance the two surfaces jump apart when they separate, then the adhesion force F_s is given by

$$F_\mathrm{s} = F_\mathrm{max} = k_\mathrm{s}\Delta D\,, \tag{18.4}$$

where we note that in liquids, the maximum or minimum in the force may occur at some nonzero surface separation (see Fig. 18.7). From F_s and a known surface geometry, and assuming that the surfaces were everywhere in molecular contact, one may also calculate the surface or interfacial energy γ. For an elastically deformable sphere of radius R on a flat surface, or for two crossed cylinders of radius R, we have [18.3, 15]

$$\gamma = \frac{F_\mathrm{s}}{3\pi R}\,, \tag{18.5}$$

while for two spheres of radii R_1 and R_2

$$\gamma = \frac{F_\mathrm{s}}{3\pi}\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\,, \tag{18.6}$$

where γ is in units of J m^{-2} (see Sect. 18.5.2).

18.2.3 The SFA and AFM

In a typical force-measuring experiment, at least two of the above displacement parameters – ΔD_0, ΔD, and ΔD_s – are directly or indirectly measured, and from these the third displacement and the resulting force law $F(D)$ are deduced using (18.1) and (18.2) together with a measured value of k_s. For example, in SFA experiments, ΔD_0 is changed by expanding or contracting a piezoelectric crystal by a known amount or by moving the base of the spring with sensitive motor-driven mechanical stages. The resulting change in surface separation ΔD is measured optically, and the spring deflection ΔD_s can then be obtained according to (18.1). In AFM experiments, ΔD_0 and ΔD_s are measured using a combination of piezoelectric, optical, capacitance or magnetic techniques, from which the change in surface separation ΔD is deduced. Once a force law is established, the geometry of the two surfaces (e.g., their radii) must also be known before the results can be compared with theory or with other experiments.

The SFA (Fig. 18.2) is used for measurements of adhesion and force laws between two curved molecularly smooth surfaces immersed in liquids or controlled vapors [18.3, 7, 16]. The surface separation is measured by multiple beam interferometry with an accuracy of ± 0.1 nm. From the shape of the interference fringes one also obtains the radius of the surfaces, R, and any surface deformation that arises during an interaction [18.17–19]. The resolution in the lateral direction is about 1 μm. The surface separation can be independently controlled to within 0.1 nm, and the force sensitivity is about 10^{-8} N. For a typical surface radius of $R \approx 1$ cm, γ values can be measured to an accuracy of about 10^{-3} mJ m^{-2}.

Several different materials have been used to form the surfaces in the SFA, including mica [18.20, 21], silica [18.22], sapphire [18.23], and polymer sheets [18.24]. These materials can also be used as supporting substrates in experiments on the forces between adsorbed or chemically bound polymer layers [18.13, 25–30], surfactant and lipid monolayers and bilayers [18.31–34], and metal and metal oxide layers [18.35–42]. The range of liquids and vapors that can be used is almost endless, and have thus far included aqueous solutions, organic liquids and solvents, polymer melts, various petroleum oils and lubricant liquids, and liquid crystals.

Friction attachments for the SFA [18.43–47] allow for the two surfaces to be sheared laterally past each other at varying sliding speeds or oscillating frequencies, while simultaneously measuring both the transverse (frictional or shear) force and the normal force (load) between them. The ranges of friction forces and sliding speeds that can be studied with such methods are currently 10^{-7}–10^{-1} N and 10^{-13}–10^{-2} m s^{-1}, respectively [18.48]. The externally applied load, L, can be varied continuously, and both positive and negative loads can be applied. The distance between the surfaces, D, their true molecular contact area, their elastic (or viscoelastic or elastohydrodynamic) deformation, and their lateral motion can all be monitored simultaneously by recording the moving interference fringe pattern.

In the atomic force microscope (Fig. 18.3), the force is measured by monitoring the deflection of a soft cantilever supporting a sub-microscopic tip ($R \approx 10$–200 nm) as this is interacting with a flat, macroscopic surface [18.8, 49, 50]. The measurements can be done in a vapor or liquid. The normal (bending) spring stiffness of the cantilever can be as small as 0.01 N m^{-1}, allowing measurements of normal forces as small as 1 pN (10^{-12} N), which corresponds to the bond strength

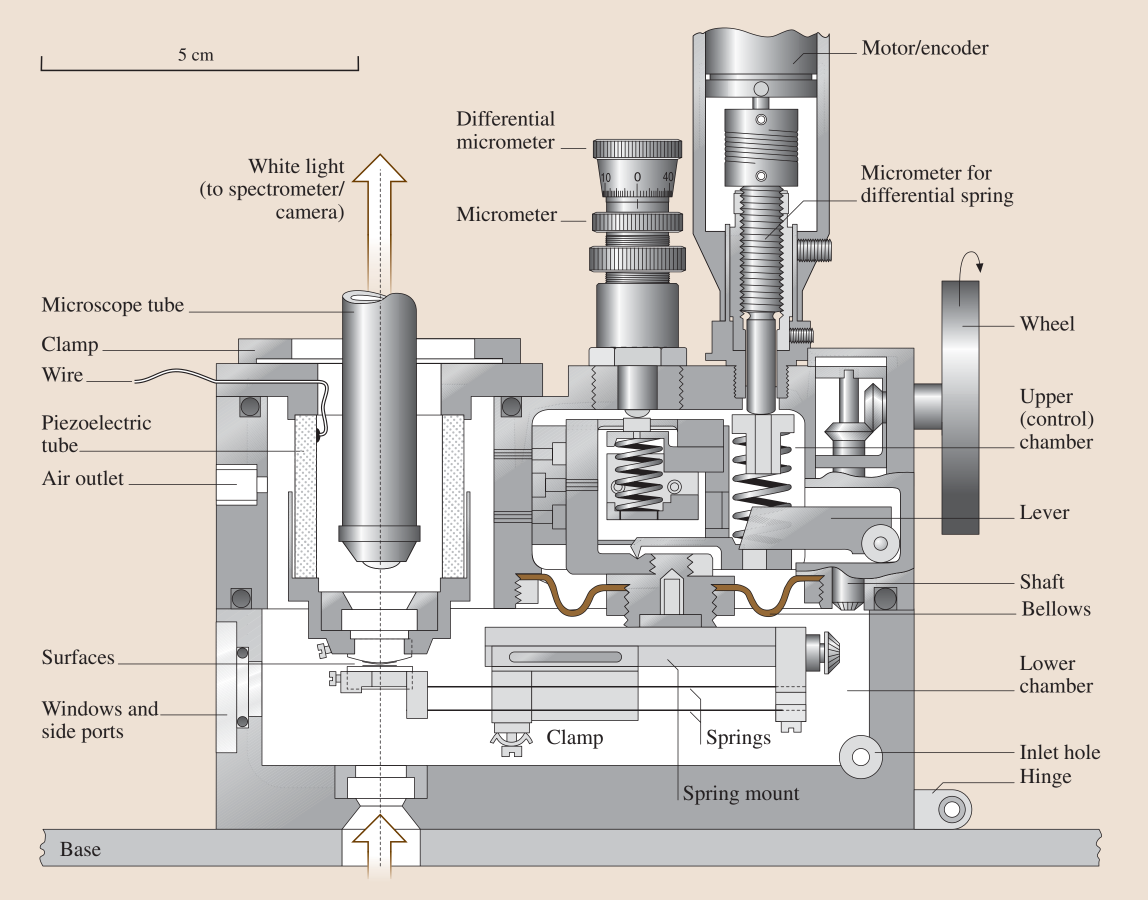
5 cm
White light
(to spectrometer/
camera)
Differential
micrometer
Micrometer
Motor/encoder
Micrometer for
differential spring
Wheel
Upper
(control)
chamber
Lever
Shaft
Bellows
Lower
chamber
Inlet hole
Hinge
Microscope tube
Clamp
Wire
Piezoelectric
tube
Air outlet
Surfaces
Windows and
side ports
Clamp
Springs
Spring mount
Base
10
0
40

Fig. 18.2 A surface forces apparatus (SFA) where the intermolecular forces between two macroscopic, cylindrical surfaces of local radius R can be directly measured as a function of surface separation over a large distance regime from tenths of a nanometer to micrometers. Local or transient surface deformations can be detected optically. Various attachments for moving one surface laterally with respect to the other have been developed for friction measurements in different regimes of sliding velocity and sliding distance (after [18.16] with permission)

of single molecules [18.51, 52]. Distances can be inferred with an accuracy of about 1 nm, and changes in distance can be measured to about 0.1 nm. Since the contact area is small, different interaction regimes can be resolved on samples with a heterogeneous composition on lateral scales of about ten nanometers. Height differences and the roughness of the sample can be measured directly from the cantilever deflection or, alternatively, by using a feedback system to raise or lower the sample so that the deflection (the normal force) is kept constant during a scan over the area of interest. Weak interaction forces and larger (microscopic) interaction areas can be investigated by replacing the tip with a micrometer-sized sphere to form a "colloidal probe" [18.9].

The atomic force microscope can also be used for friction measurements (lateral force microscopy, LFM) by monitoring the torsion of the cantilever as the sample is scanned in the direction perpendicular to the long axis of the cantilever [18.10, 50, 53, 54]. Typically, the stiffness of the cantilever to lateral bending is much larger than to bending in the normal direction and to torsion, so that these signals are decoupled and height and friction can be detected simultaneously. The torsional spring constant can be as low as $0.1\,\mathrm{N\,m^{-1}}$, giving a lateral (friction) force sensitivity of 10^{-11} N.

Rapid technical developments have facilitated the calibrations of the normal [18.55, 56] and lateral spring constants [18.54, 57], as well as in situ measurements of

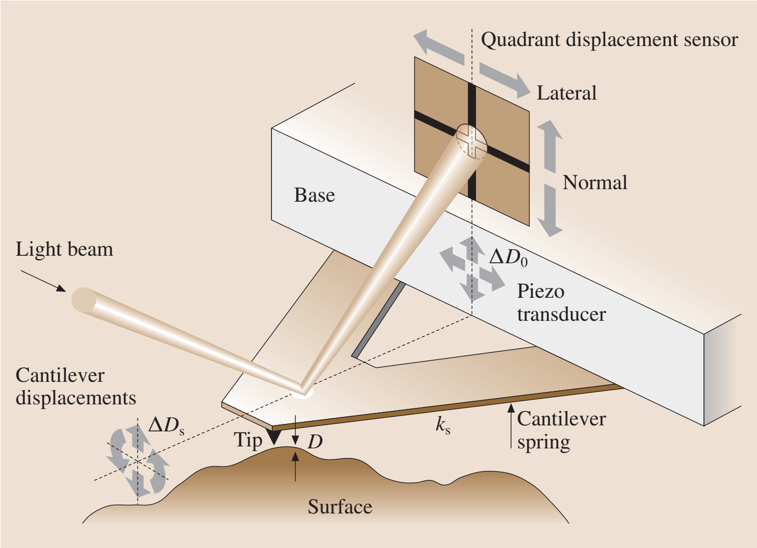

Fig. 18.3 Schematic drawing of an atomic force microscope (AFM) tip supported on a triangular cantilever interacting with an arbitrary solid surface. The normal force and topology are measured by monitoring the calibrated deflection of the cantilever as the tip is moved across the surface by means of a piezoelectric transducer. Various designs have been developed that move either the sample or the cantilever during the scan. Friction forces can be measured from the torsion of the cantilever when the scanning is in the direction perpendicular to its long axis (after [18.62] with permission)

the macroscopic tip radius [18.58]. Cantilevers of different shapes with a large range of spring constants, tip radii, and surface treatments (inorganic or organic coatings) are commercially available. The flat surface, and also the particle in the colloidal probe technique, can be any material of interest. However, remaining difficulties with this technique are that the distance between the tip and the substrate, D, and the deformations of the tip and sample are not directly measurable. Another important difference between the AFM/LFM and SFA techniques is the different size of the contact area, and the related observation that even when a cantilever with a very low spring constant is used in the AFM, the pressure in the contact zone is typically much higher than in the SFA. Hydrodynamic effects in liquids also affect the measurements of normal forces differently on certain time scales [18.59–61].

18.2.4 Some Other Force-Measuring Techniques

A large number of other techniques are available for the measurements of the normal forces between solid or fluid surfaces (see [18.5, 62]). The techniques discussed in this section are not used for lateral (friction) force measurements, but are commonly used to study normal forces, particularly in biological systems.

Micropipette aspiration is used to measure the forces between cells or vesicles, or between a cell or vesicle and another surface [18.63–65]. The cell or vesicle is held by suction at the tip of a glass micropipette and deforms elastically in response to the net interactions with another surface and to the applied suction. The shape of the deformed surface (cell membrane) is measured and used to deduce the force between the surfaces and the membrane tension [18.64]. The membrane tension, and thus the stiffness of the cell or vesicle, is regulated by applying different hydrostatic pressures. Forces can be measured in the range of 0.1 pN to 1 nN, and the distance resolution is a few nanometers. The interactions between a colloidal particle and another surface can be studied by attaching the particle to the cell membrane [18.66].

In the osmotic stress technique, pressures are measured between colloidal particles in aqueous solution, membranes or bilayers, or other ordered colloidal structures (viruses, DNA). The separation between the particle surfaces and the magnitude of membrane undulations are measured by X-ray or neutron scattering techniques. This is combined with a measurement of the osmotic pressure of the solution [18.67–70]. The technique has been used to measure repulsive forces, such as DLVO interactions, steric forces, and hydration forces [18.71]. The sensitivity in pressure is $0.1\,\mathrm{mN\,m^{-2}}$, and distances can be resolved to 0.1 nm.

The optical tweezers technique is based on the trapping of dielectric particles at the center of a focused laser beam by restoring forces arising from radiation pressure and light intensity gradients [18.72, 73]. The forces experienced by particles as they are moved toward or away from one another can be measured with a sensitivity in the pN range. Small biological molecules are typically attached to a larger bead of a material with suitable refractive properties. Recent development allows determinations of position with nanometer resolution [18.74], which makes this technique useful for studying the forces during extension of single molecules.

In total internal reflection microscopy (TIRM), the potential energy between a micrometer-sized colloidal particle and flat surface in aqueous solution is deduced from the average equilibrium height of the particle above the surface, measured from the intensity of scattered light. The average height ($D \approx 10-100\,\mathrm{nm}$) results from a balance of gravitational force, radiation pressure from a laser beam focused at the particle from below, and intermolecular forces [18.75]. The technique is particularly suitable for measuring weak forces (sensitivity ca. $10^{-14}\,\mathrm{N}$), but is more difficult to use for systems with strong interactions. A related technique is reflection interference contrast microscopy (RICM), where optical interference is used to also monitor changes in the shape of the approaching colloidal particle or vesicle [18.76].

An estimate of bond strengths can be obtained from the hydrodynamic shear force exerted by a fluid on particles or cells attached to a substrate [18.77, 78]. At a critical force, the bonds are broken and the particle or cell will be detached and move with the velocity of the fluid. This method requires knowledge of the contact area and the flow velocity profile of the fluid. Furthermore, a uniform stress distribution in the contact area is generally assumed. At low bond density, this technique can be used to determine the strength of single bonds (1 pN).

18.3 Normal Forces Between Dry (Unlubricated) Surfaces

18.3.1 Van der Waals Forces in Vacuum and Inert Vapors

Forces between macroscopic bodies (such as colloidal particles) across vacuum arise from interactions between the constituent atoms or molecules of each body across the gap separating them. These intermolecular interactions are electromagnetic forces between permanent or induced dipoles (van der Waals forces), and between ions (electrostatic forces). In this section, we describe the van der Waals forces, which occur between all atoms and molecules and between all macroscopic bodies (see [18.3]).

The interaction between two permanent dipoles with a fixed relative orientation can be attractive or repulsive. For the specific case of two freely rotating permanent dipoles in a liquid or vapor (orientational or Keesom interaction), and for a permanent dipole and an induced dipole in an atom or polar or nonpolar molecule (induction or Debye interaction), the interaction is on average always attractive. The third type of van der Waals interaction, the fluctuation or London dispersion interaction, arises from instantaneous polarization of one nonpolar or polar molecule due to fluctuations in the charge distribution of a neighboring nonpolar or polar molecule (Fig. 18.4a). Correlation between these fluctuating induced dipole moments gives an attraction that is present between any two molecules or surfaces across vacuum. At very small separations, the interaction will ultimately be repulsive as the electron clouds of atoms and molecules begin to overlap. The total interaction is thus a combination of a short-range repulsion and a relatively long-range attraction.

Except for in highly polar materials such as water, London dispersion interactions give the largest contribution (70–100 %) to the van der Waals attraction. The interaction energy of the van der Waals force between atoms or molecules depends on the separation r as

$$E(D) = \frac{-C_{\mathrm{vdW}}}{r^6}\,, \tag{18.7}$$

where the constant C_{vdW} depends on the dipole moments and polarizabilities of the molecules. At large separations ($> 10\,\mathrm{nm}$), the London interaction is weakened by a randomizing effect caused by the rapid fluctuations. That is, the induced temporary dipole moment of one molecule may have changed during the time needed for the transmission of the electromagnetic wave (photon) generated by its fluctuating charge density to another molecule and the return of the photon generated by the induced fluctuation in this second molecule. This phenomenon is called retardation and causes the interaction energy to decay as r^{-7} at large separations [18.79].

Dispersion interactions are to a first approximation additive, and their contribution to the interaction energy between two macroscopic bodies (such as colloidal particles) across vacuum can be found by summing the pair-wise interactions [18.80]. The interaction is generally described in terms of the Hamaker constant, A_H. Another approach is to treat the interacting bodies and an intervening medium as continuous phases and determine the strength of the interaction from bulk dielectric properties of the materials [18.81, 82]. Unlike the pair-wise summation, this method takes into account the screening of the interactions between molecules inside the bodies by the molecules closer to the surfaces and the effects of the intervening medium. For the interaction between material 1 and material 3 across material 2, the non-retarded Hamaker constant given by the Lifshitz theory

Table 18.2 Van der Waals interaction energy and force between macroscopic bodies of different geometries

Geometry of bodies with surfaces *D* apart ($D \ll R$)			**van der Waals interaction**	
			Energy, *E*	**Force, *F***
Two atoms or small molecules	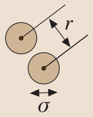	$r \geqslant \sigma$	$\frac{-C_{vdW}}{r^6}$	$\frac{-6C_{vdW}}{r^7}$
Two flat surfaces (per unit area)	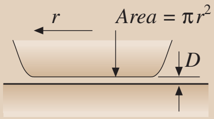	$r \gg D$	$\frac{-A_H}{12\pi D^2}$	$\frac{-A_H}{6\pi D^3}$
Two spheres or macromolecules of radii R_1 and R_2	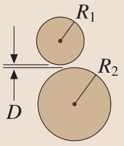	$R_1, R_2 \gg D$	$\frac{-A_H}{6D}\left(\frac{R_1R_2}{R_1+R_2}\right)$	$\frac{-A_H}{6D^2}\left(\frac{R_1R_2}{R_1+R_2}\right)$
Sphere or macromolecule of radius R near a flat surface	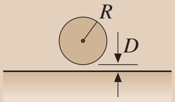	$R \gg D$	$\frac{-A_H R}{6D}$	$\frac{-A_H R}{6D^2}$
Two parallel cylinders or rods of radii R_1 and R_2 (per unit length)	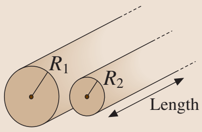	$R_1, R_2 \gg D$	$\frac{-A_H}{12\sqrt{2}\,D^{3/2}}\left(\frac{R_1R_2}{R_1+R_2}\right)^{1/2}$	$\frac{-A_H}{8\sqrt{2}\,D^{5/2}}\left(\frac{R_1R_2}{R_1+R_2}\right)^{1/2}$
Cylinder of radius R near a flat surface (per unit length)	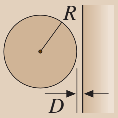	$R \gg D$	$\frac{-A_H\sqrt{R}}{12\sqrt{2}\,D^{3/2}}$	$\frac{-A_H\sqrt{R}}{8\sqrt{2}\,D^{5/2}}$
Two cylinders or filaments of radii R_1 and R_2 crossed at 90°	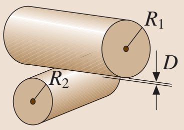	$R_1, R_2 \gg D$	$\frac{-A_H\sqrt{R_1R_2}}{6D}$	$\frac{-A_H\sqrt{R_1R_2}}{6D^2}$

A negative force (A_H positive) implies attraction, a positive force means repulsion (A_H negative)
(after [18.62] with permission)

is approximately [18.3]:

$$
\begin{aligned}
A_{\mathrm{H},123} &= A_{\mathrm{H},\nu=0} + A_{\mathrm{H},\nu>0} \\
&\approx \tfrac{3}{4} k_{\mathrm{B}} T \left(\tfrac{\varepsilon_1-\varepsilon_2}{\varepsilon_1+\varepsilon_2}\right)\left(\tfrac{\varepsilon_3-\varepsilon_2}{\varepsilon_3+\varepsilon_2}\right) \\
&+ \tfrac{3h\nu_{\mathrm{e}}}{8\sqrt{2}} \frac{(n_1^2-n_2^2)(n_3^2-n_2^2)}{\sqrt{(n_1^2+n_2^2)}\sqrt{(n_3^2+n_2^2)}\left(\sqrt{(n_1^2+n_2^2)}+\sqrt{(n_3^2+n_2^2)}\right)}\,, \qquad (18.8)
\end{aligned}
$$

where the first term ($\nu = 0$) represents the permanent dipole and dipole–induced dipole interactions and the second ($\nu > 0$) the London (dispersion) interaction. ε_i and n_i are the static dielectric constants and refractive indexes of the materials, respectively. ν_{e} is the frequency of the lowest electron transition (around $3 \times 10^{15}\,\mathrm{s}^{-1}$). Either one of the materials 1, 2, or 3 in (18.8) can be vacuum or air ($\varepsilon = n = 1$). A_{H} is typically 10^{-20}–10^{-19} J (the higher values are found for metals) for interactions between solids and liquids across vacuum or air.

The interaction energy between two macroscopic bodies is dependent on the geometry and is always attractive between two bodies of the same material [A_{H} positive, see (18.8)]. The van der Waals interaction energy and force laws ($F = -\mathrm{d}E(D)/\mathrm{d}D$) for some common geometries are given in Table 18.2. Because of the retardation effect, the equations in Table 18.2 will lead to an overestimation of the dispersion force at large separations. It is, however, apparent that the interaction energy between macroscopic bodies decays more slowly with separation (i. e., has a longer range) than between two molecules.

For inert nonpolar surfaces, e.g., consisting of hydrocarbons or van der Waals solids and liquids, the Lifshitz theory has been found to apply even at molecular contact, where it can be used to predict the surface energies (surface tensions) of such solids and liquids. For example, for hydrocarbon surfaces, $A_{\mathrm{H}} = 5 \times 10^{-20}$ J. Inserting this value into the equation for two flat surfaces (Table 18.2) and using a "cut-off" distance of $D_0 \approx 0.165$ nm as an effective separation when the surfaces are in contact [18.3], we obtain for the surface energy γ (which is defined as half the interaction energy)

$$\gamma = \frac{E}{2} = \frac{A_{\mathrm{H}}}{24\pi D_0^2} \approx 24\,\mathrm{mJ\,m^{-2}}\,, \qquad (18.9)$$

a value that is typical for hydrocarbon solids and liquids [18.83].

If the adhesion force is measured between a spherical surface of radius $R = 1$ cm and a flat surface using an SFA, we expect the adhesion force to be (see Table 18.2) $F = A_{\mathrm{H}}R/(6D_0^2) = 4\pi R\gamma \approx 3.0$ mN. Using a spring constant of $k_{\mathrm{s}} = 100\,\mathrm{N\,m^{-1}}$, such an adhesive force will cause the two surfaces to jump apart by $\Delta D = F/k_{\mathrm{s}} = 30\,\mu\mathrm{m}$, which can be accurately measured. (For elastic bodies that deform in adhesive contact, R changes during the interaction and the measured adhesion force is 25% lower, see Sect. 18.5.2). Surface energies of solids can thus be directly measured with the SFA and, in principle, with the AFM if the contact geometry can be quantified. The measured values are in good agreement with calculated values based on the known surface energies γ of the materials, and for nonpolar low-energy solids they are well accounted for by the Lifshitz theory [18.3].

18.3.2 Charge Exchange Interactions

Electrostatic interactions are present between ions (Coulomb interactions), between ions and permanent dipoles, and between ions and nonpolar molecules in which a charge induces a dipole moment. The interaction energy between ions or between a charge and a fixed permanent dipole can be attractive or repulsive. For an induced dipole or a freely rotating permanent dipole in vacuum or air, the interaction energy with a charge is always attractive.

Spontaneous charge transfer may occur between two dissimilar materials in contact [18.84]. The phenomenon is especially prominent in contacts between a metal, for example, mercury, and a material with low conductivity, but is also observed, for example, between two different polymer layers. During separation, rolling or sliding of one body over the other, the surfaces experience both charge transition from one surface to the other and charge transfer (conductance) along each surface (Fig. 18.4b). The latter process is typically slower, and, as a result, charges remain on the surfaces as they are separated in vacuum or dry nitrogen gas. The charging gives rise to a strong adhesion with adhesion energies of over 1,000 $\mathrm{mJ\,m^{-2}}$, similar to fracture or cohesion energies of the solid bodies themselves [18.84, 85]. Upon separating the surfaces farther apart, a strong, long-range electrostatic attraction is observed. The charging can be decreased through discharges across the gap between the surfaces (which requires a high charging) or through conducting in the solids. It has been suggested that charge exchange interactions are particularly important in rolling friction between dry surfaces (which can simplistically be thought of as an adhesion–separation process), where the distance dependence of forces acting normally to the surfaces plays a larger role than in sliding friction. Recent experiments on the sliding friction between metal–insulator surfaces indicate that stick–slip would be accompanied by charge transfer events [18.86].

Photo-induced charge transfer, or harpooning, involves the transfer of an electron between an atom in a molecular beam or at a solid surface (typically alkali or transition metal) to an atom or molecule in a gas (typically halides) to form a negatively charged molecular ion in a highly excited vibrational state. This transfer process can occur at atomic distances of 0.5–0.7 nm, which is far from molecular contact. The formed molecular ion is attracted to the surface and chemisorbs onto it. Photo-induced charge transfer processes also occur in the photosynthesis in green plants and in photoelectrochemical cells (solar cells) at the junction between two semiconductors or between a semiconductor and an electrolyte solution [18.87].

18.3.3 Sintering and Cold Welding

When macroscopic particles in a powder or in a suspension come into molecular contact, they can bond together to form a network or solid body with very different density and shear strength compared to the powder (a typical example is porcelain). The rate of bonding is dependent on the surface energy (causing a stress at the edge of the contact) and the atomic mobility (diffusion rate) of the contacting materials. To increase the diffusion rate, objects formed from powders are heated to about one-half of the melting temperature of the components in a process called sintering, which can be done in different atmospheres or in a liquid.

In the sintering process, the surface energy of the system is lowered due to the reduction of total surface area (Fig. 18.4c). In metal and ceramic systems, the most important mechanism is solid-state diffusion, initially surface diffusion. As the surface area decreases and the grain boundaries increase at the contacts, grain boundary diffusion and diffusion through the crystal lattice become more important. The grain boundaries will eventually migrate, so that larger particles are formed (coarsening). Mass can also be transferred through evaporation and condensation, and through viscous and plastic flow. In liquid-phase sintering, the materials can melt, which increases the mass transport. Amorphous materials like polymers and glasses do not have real grain boundaries and sinter by viscous flow [18.88].

Some of these mechanisms (surface diffusion and evaporation–condensation) reduce the surface area and increase the grain size (coarsening) without densification, in contrast to bulk transport mechanisms like grain boundary diffusion and plastic and viscous flow. As the material becomes denser, elongated pores collapse to form smaller, spherical pores with a lower surface energy. Models for sintering typically consider the size and growth rate of the grain boundary (the “neck”) formed between two spherical particles. At a high stage of densification, the sintering stress σ at the curved neck between two particles is given by [18.88]

$$\sigma = \frac{2\gamma_{\mathrm{SS}}}{G} + \frac{2\gamma_{\mathrm{SV}}}{r_{\mathrm{p}}} , \tag{18.10}$$

where γ_{SS} is the solid–solid grain boundary energy, γ_{SV} is the solid–vapor surface energy, G is the grain size, and r_{p} is the radius of the pore.

A related phenomenon is cold welding, which is the spontaneous formation of strong junctions between clean (unoxidized) metal surfaces with a mutual solubility when they are brought in contact with or without an applied pressure. The plastic deformations accompanying the formation and breaking of such contacts on a molecular scale during motion of one surface normally (see Fig. 18.10c,d) or laterally (shearing) with respect to the other have been studied both experimentally [18.89] and theoretically [18.90–95]. The breaking of a cold welded contact is generally associated with damage or deformation of the surface structure.

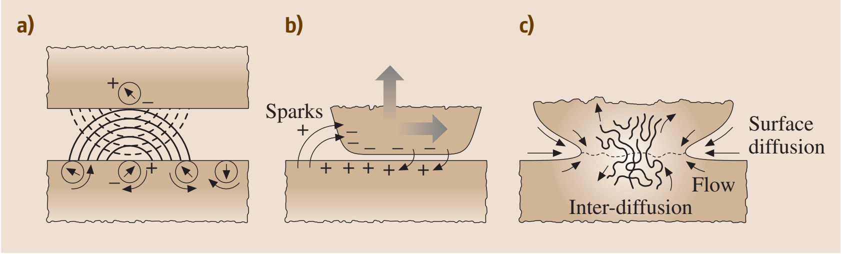

Fig. 18.4a–c Schematic representation of **(a)** van der Waals interaction (dipole–induced dipole interaction), **(b)** charge exchange, which acts to increase adhesion and friction forces, and **(c)** sintering between two surfaces

18.4 Normal Forces Between Surfaces in Liquids

18.4.1 Van der Waals Forces in Liquids

The dispersion interaction in a medium will be significantly lower than in vacuum, since the attractive interaction between two solute molecules in a medium (solvent) involves displacement and reorientation of the nearest-neighbor solvent molecules. Even though the surrounding medium may change the dipole moment and polarizability from that in vacuum, the interaction between two identical molecules remains attractive in a binary mixture. The extension of the interactions to the case of two macroscopic bodies is the same as described in Sect. 18.3.1. Typically, the Hamaker constants for interactions in a medium are an order of magnitude lower than in vacuum. Between macroscopic surfaces in liquids, van der Waals forces become important at distances below 10–15 nm and may at these distances start to dominate interactions of different origin that have been observed at larger separations.

Figure 18.5 shows the measured van der Waals forces between two crossed cylindrical mica surfaces in water and various salt solutions. Good agreement is obtained between experiment and theory. At larger surface separations, above about 5 nm, the measured forces fall off more rapidly than as D^{-2}. This retardation effect (see Sect. 18.3.1) is also predicted by Lifshitz theory and is due to the time needed for propagation of the induced dipole moments over large distances.

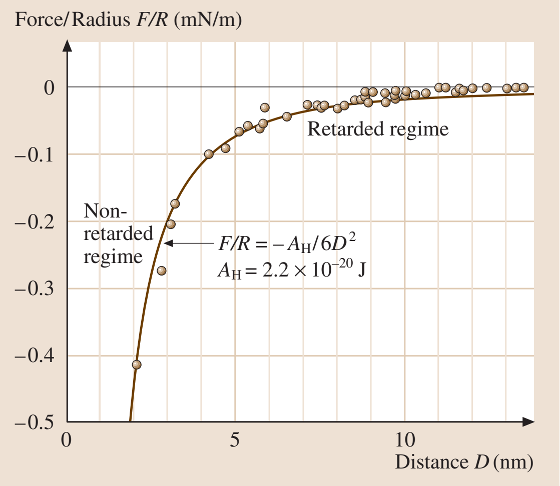

Fig. 18.5 Attractive van der Waals force F between two curved mica surfaces of radius $R \approx 1$ cm measured in water and various aqueous electrolyte solutions. The electrostatic interaction has been subtracted from the total measured force. The measured non-retarded Hamaker constant is $A_H = 2.2 \times 10^{-20}$ J. Retardation effects are apparent at distances larger than 5 nm, as expected theoretically. (After [18.3]. Copyright 1991, with permission from Elsevier Science)

From Fig. 18.5, we may conclude that at separations above about 2 nm, or 8 molecular diameters of water, the *continuum* Lifshitz theory is valid. This would mean that water films as thin as 2 nm may be expected to have bulk-like properties, at least as far as their interaction forces are concerned. Similar results have been obtained with other liquids, where in general continuum properties are manifested, both as regards their interactions and other properties such as viscosity, at a film thickness larger than 5 or 10 molecular diameters. In the absence of a solvent (in vacuum), the agreement of measured van der Waals forces with the continuum Lifshitz theory is generally good at all separations down to molecular contact ($D = D_0$).

Van der Waals interactions in a system of three or more different materials (see (18.8)) can be attractive or repulsive, depending on their dielectric properties. Numerous experimental studies show the attractive van der Waals forces in various systems [18.3], and also repulsive van der Waals forces have been measured directly [18.96]. A practical consequence of the repulsive interaction obtained across a medium with intermediate dielectric properties is that the van der Waals forces will give rise to preferential, nonspecific adsorption of molecules with an intermediate dielectric constant. This is commonly seen as adsorption of vapors or solutes to a solid surface. It is also possible to diminish the attractive interaction between dispersed colloidal particles by adsorption of a thin layer of material with dielectric properties close to those of the surrounding medium (matching of refractive index), or by adsorption of a polymer that gives a steric repulsive force that keeps the particles separated at a distance where the magnitude of the van der Waals attraction is negligible. Thermal motion will then keep the particles dispersed.

18.4.2 Electrostatic and Ion Correlation Forces

Most surfaces in contact with a highly polar liquid (such as water) acquire a surface charge, either by dissociation of ions from the surface into the solution or by preferential adsorption of certain ions from the solution.

The surface charge is balanced by a layer of oppositely charged ions (counterions) in the solution at some small distance from the surface (see [18.3]). In dilute solution, this distance is the Debye length, κ^{-1}, which is purely a property of the electrolyte *solution*. The Debye length falls with increasing ionic strength (i. e., with the molar concentration M_i and valency z_i) of the ions in solution:

$$\kappa^{-1} = \left(\frac{\varepsilon \varepsilon_0 k_B T}{e^2 N_A \sum_i z_i^2 M_i} \right)^{1/2} , \tag{18.11}$$

where e is the electronic charge. For example, for 1:1 electrolytes at 25 °C, $\kappa^{-1} = 0.304\,\text{nm}/\sqrt{M_{1:1}}$, where M_i is given in M ($\text{mol}\,\text{dm}^{-3}$). κ^{-1} is thus ca. 10 nm in a 1 mM NaCl solution and 0.3 nm in a 1 M solution. In totally pure water at pH 7, where $M_i = 10^{-7}$ M, κ^{-1} is 960 nm, or about 1 μm. The Debye length also relates the surface charge density σ of a surface to the electrostatic surface potential ψ_0 via the Grahame equation, which for 1:1 electrolytes can be expressed as:

$$\sigma = \sqrt{8\varepsilon\varepsilon_0 k_B T}\,\sinh\left(e\psi_0/2k_B T\right) \times \sqrt{M_{1:1}} . \tag{18.12}$$

Since the Debye length is a measure of the thickness of the diffuse atmosphere of counterions near a charged surface, it also determines the range of the electrostatic “double-layer” interaction between two charged surfaces. The electrostatic double-layer interaction is an entropic effect that arises upon decreasing the thickness of the liquid film containing the dissolved ions. Because of the attractive force between the dissolved ions and opposite charges on the surfaces, the ions stay between the surfaces, but an osmotic repulsion arises as their concentration increases. The long-range electrostatic interaction energy at large separations (weak overlap) between two similarly charged molecules or surfaces is typically repulsive and is roughly an exponentially decaying function of D:

$$E(D) \approx +C_{ES}\,e^{-\kappa D} , \tag{18.13}$$

where C_{ES} is a constant that depends on the geometry of the interacting surfaces, on their surface charge density, and the solution conditions (Table 18.3). We see that the Debye length is the decay length of the interaction energy between two surfaces (and of the mean potential away from one surface). C_{ES} can be determined by solving the so-called Poisson–Boltzmann equation or by using other theories [18.97, 98]. The equations in Table 18.3 are expressed in terms of a constant, Z, defined as

$$Z = 64\pi\varepsilon\varepsilon_0 (k_B T/e)^2 \tanh^2\left[ze\psi_0/(4k_B T)\right] , \tag{18.14}$$

which depends only on the properties of the *surfaces*.

The above approximate expressions are accurate only for surface separations larger than about one Debye length. At smaller separations one must use numerical solutions of the Poisson–Boltzmann equation to obtain the exact interaction potential, for which there are no simple expressions. In the limit of small D, it can be shown that the interaction energy depends on whether the surfaces remain at constant potential ψ_0 (as assumed in the above equations) or at constant charge σ (when the repulsion exceeds that predicted by the above equations), or somewhere between these limits. In the “constant charge limit” the total *number* of counterions in the compressed film does not change as D is decreased, whereas at constant potential, the *concentration* of counterions is constant.The limiting pressure (or force per unit area) at constant charge is the osmotic pressure of the confined ions:

$$\begin{aligned} F &= k_B T \times \text{ion number density} \\ &= 2\sigma k_B T/(zeD), \quad \text{for } D \ll \kappa^{-1} . \end{aligned} \tag{18.15}$$

That is, as $D \to 0$ the double-layer pressure at constant surface charge becomes infinitely repulsive and independent of the salt concentration (at constant potential the force instead becomes a constant at small D). However, at small separations, the van der Waals attraction (which goes as D^{-2} between two spheres or as D^{-3} between two planar surfaces, see Table 18.2) wins out over the double-layer repulsion, unless some other short-range interaction becomes dominant (see Sect. 18.4.4). This is the theoretical prediction that forms the basis of the so-called Derjaguin–Landau–Verwey–Overbeek (DLVO) theory [18.97, 99], illustrated in Fig. 18.6.

Because of the different distance dependence of the van der Waals and electrostatic interactions, the total force law, as described by the DLVO theory, can show several minima and maxima. Typically, the depth of the outer (secondary) minimum is a few $k_B T$, enough to cause reversible flocculation of particles from an aqueous dispersion. If the force barrier between the secondary and primary minimum is lowered, for example, by increasing the electrolyte concentration, particles can be irreversibly coagulated in the primary minimum. In practice, other forces (described

Table 18.3 Electrical "double-layer" interaction energy $E(D)$ and force ($F = -\mathrm{d}E/\mathrm{d}D$) between macroscopic bodies

Geometry of bodies with surfaces D apart ($D \ll R$)			**Electric "double-layer" interaction**	
			Energy, E	**Force, F**
Two ions or small molecules	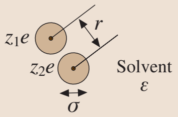 z_1e r z_2e Solvent ε σ	$r \gg \sigma$	$\frac{+z_1z_2e^2}{4\pi\varepsilon\varepsilon_0 r}\frac{e^{-\kappa(r-\sigma)}}{(1+\kappa\sigma)}$	$\frac{+z_1z_2e^2}{4\pi\varepsilon\varepsilon_0 r^2}\frac{(1+\kappa r)}{(1+\kappa\sigma)}e^{-\kappa(r-\sigma)}$
Two flat surfaces (per unit area)	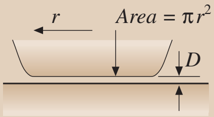	$r \gg D$	$(\kappa/2\pi)\,Z\mathrm{e}^{-\kappa D}$	$\left(\kappa^2/2\pi\right)Z\mathrm{e}^{-\kappa D}$
Two spheres or macromolecules of radii R_1 and R_2	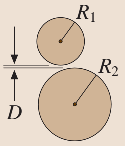	$R_1, R_2 \gg D$	$\left(\frac{R_1R_2}{R_1+R_2}\right)Z\mathrm{e}^{-\kappa D}$	$\kappa\left(\frac{R_1R_2}{R_1+R_2}\right)Z\mathrm{e}^{-\kappa D}$
Sphere or macro-molecule of radius R near a flat surface	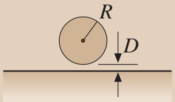	$R \gg D$	$RZ\mathrm{e}^{-\kappa D}$	$\kappa RZ\mathrm{e}^{-\kappa D}$
Two parallel cylinders or rods of radii R_1 and R_2 (per unit length)	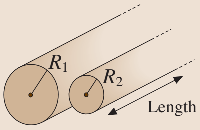	$R_1, R_2 \gg D$	$\frac{\kappa^{1/2}}{\sqrt{2\pi}}\left(\frac{R_1R_2}{R_1+R_2}\right)^{1/2}Z\mathrm{e}^{-\kappa D}$	$\frac{\kappa^{3/2}}{\sqrt{2\pi}}\left(\frac{R_1R_2}{R_1+R_2}\right)^{1/2}Z\mathrm{e}^{-\kappa D}$
Cylinder of radius R near a flat surface (per unit length)	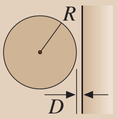	$R \gg D$	$\kappa^{1/2}\sqrt{\frac{R}{2\pi}}Z\mathrm{e}^{-\kappa D}$	$\kappa^{3/2}\sqrt{\frac{R}{2\pi}}Z\mathrm{e}^{-\kappa D}$
Two cylinders or filaments of radii R_1 and R_2 crossed at 90°	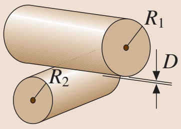	$R_1, R_2 \gg D$	$\sqrt{R_1R_2}Z\mathrm{e}^{-\kappa D}$	$\kappa\sqrt{R_1R_2}Z\mathrm{e}^{-\kappa D}$

The interaction energy and force for bodies of different geometries is based on the Poisson–Boltzmann equation (a continuum, mean-field theory). Equation (18.14) gives the interaction constant Z (in terms of surface potential ψ_0) for the interaction between similarly charged (ionized) surfaces in aqueous solutions of monovalent electrolyte. It can also be expressed in terms of the surface charge density σ by applying the Grahame equation (18.12) (after [18.62] with permission)

in the following sections) often appear at very small separations, so that the full force law between two surfaces or colloidal particles in solution can be more complex than might be expected from the DLVO theory.

There are situations when the double-layer interaction can be attractive at short range even between surfaces of similar charge, especially in systems with charge regulation due to dissociation of chargeable groups on the surfaces [18.100]; ion condensa-

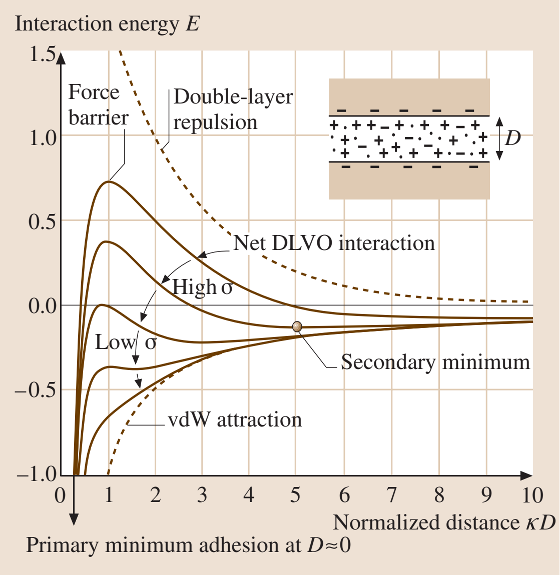

Fig. 18.6 Schematic plots of the DLVO interaction potential energy E between two flat, charged surfaces (or, according to the Derjaguin approximation, (18.3), the force F between two curved surfaces) as a function of the surface separation normalized by the Debye length, κ^{-1}. The van der Waals attraction (inverse power-law dependence on D) together with the repulsive electrostatic "double-layer" force (roughly exponential) at different surface charge σ (or potential, see (18.12)) determine the net interaction potential in aqueous electrolyte solution (after [18.62] with permission)

tion [18.101], which may lower the effective surface charge density in systems containing di-and trivalent counterions; or ion correlation, which is an additional van der Waals-like attraction due to mobile and therefore highly polarizable counterions located at the surface [18.102]. The ion correlation (or charge fluctuation) force becomes significant at separations below 4 nm and increases with the surface charge density σ and the valency z of the counterions. Computer simulations have shown that at high charge density and monovalent counterions, the ion correlation force can reduce the effective double-layer repulsion by 10–15 %. With divalent counterions, the ion correlation force was found to exceed the double-layer repulsion and the total force then became attractive at a separation below 2 nm even in dilute electrolyte solution [18.103]. Experimentally, such short-range attractive forces have been found in charged bilayer systems [18.104, 105].

18.4.3 Solvation and Structural Forces

When a liquid is confined within a restricted space, for example, a very thin film between two surfaces, it ceases to behave as a structureless continuum. At small surface separations (below about 10 molecular diameters), the van der Waals force between two surfaces or even two solute molecules in a liquid (solvent) is no longer a smoothly varying attraction. Instead, there arises an additional "solvation" force that generally oscillates between attraction and repulsion with distance, with a periodicity equal to some mean dimension σ of the liquid molecules [18.106]. Figure 18.7a shows the force law between two smooth mica surfaces across the hydrocarbon liquid tetradecane, whose inert, chain-like molecules have a width of $\sigma \approx 0.4$ nm.

The short-range oscillatory force law is related to the "density distribution function" and "potential of mean force" characteristic of intermolecular interactions in liquids. These forces arise from the confining effects two surfaces have on liquid molecules, forcing them to order into quasi-discrete layers. Such layers are energetically or entropically favored and correspond to the minima in the free energy, whereas fractional layers are disfavored (energy maxima). This effect is quite general and arises in all simple liquids when they are confined between two smooth, rigid surfaces, both flat and curved. Oscillatory forces do not require any attractive liquid–liquid or liquid–wall interaction, only two hard walls confining molecules whose shape is not too irregular and that are free to exchange with molecules in a bulk liquid reservoir. In the absence of any attractive pressure between the molecules, the bulk liquid density could be maintained by an external hydrostatic pressure – in real liquids attractive van der Waals forces play the role of such an external pressure.

Oscillatory forces are now well understood theoretically, at least for simple liquids, and a number of theoretical studies and computer simulations of various confined liquids (including water) that interact via some form of Lennard-Jones potential have invariably led to an oscillatory solvation force at surface separations below a few molecular diameters [18.107–114]. In a first approximation, the oscillatory force law may be described by an exponentially decaying cosine function of the form

$$E \approx E_0 \cos(2\pi D/\sigma)\,\mathrm{e}^{-D/\sigma} , \quad (18.16)$$

where both theory and experiments show that the oscillatory period and the characteristic decay length of the envelope are close to σ.

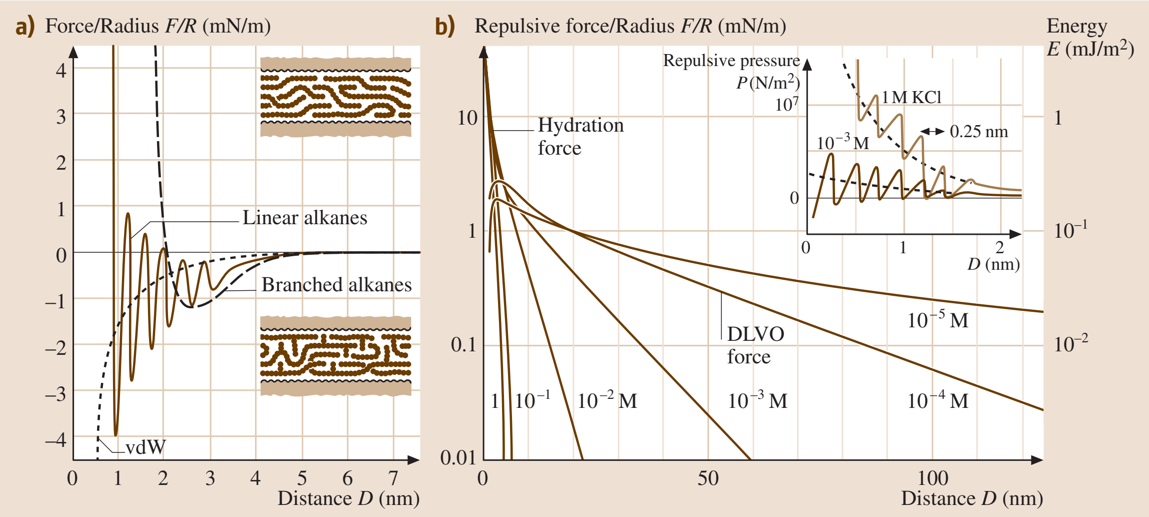

Fig. 18.7 (a) *Solid curve:* Forces measured between two mica surfaces across saturated linear chain alkanes such as n-tetradecane and n-hexadecane [18.115, 116]. The 0.4 nm periodicity of the oscillations indicates that the molecules are preferentially oriented parallel to the surfaces, as shown schematically in the *upper insert*. The theoretical continuum van der Waals attraction is shown as a *dotted curve*. *Dashed curve:* Smooth, non-oscillatory force law exhibited by irregularly shaped alkanes (such as 2-methyloctadecane) that cannot order into well-defined layers (*lower insert*) [18.116, 117]. Similar non-oscillatory forces are also observed between "rough" surfaces, even when these interact across a saturated linear chain liquid. This is because the irregularly shaped surfaces (rather than the liquid) now prevent the liquid molecules from ordering in the gap. **(b)** Forces measured between charged mica surfaces in KCl solutions of varying concentrations [18.20]. In dilute solutions (10^{-5} and 10^{-4} M), the measured forces are excellently described by the DLVO theory, based on exact solutions to the nonlinear Poisson–Boltzmann equation for the electrostatic forces and the Lifshitz theory for the van der Waals forces (using a Hamaker constant of $A_H = 2.2 \times 10^{-20}$ J). At higher concentrations, as more hydrated K^+ cations adsorb onto the negatively charged surfaces, an additional hydration force appears superimposed on the DLVO interaction at distances below 3–4 nm. This force has both an oscillatory and a monotonic component. *Insert:* Short-range hydration forces between mica surfaces shown as pressure versus distance. The lower and upper curves show surfaces 40% and 95% saturated with K^+ ions. At larger separations, the forces are in good agreement with the DLVO theory. (After [18.3]. Copyright 1991, with permission from Elsevier Science)

Once the solvation zones of the two surfaces overlap, the mean liquid density in the gap is no longer the same as in the bulk liquid. Since the van der Waals interaction depends on the optical properties of the liquid, which in turn depends on the density, the van der Waals and the oscillatory solvation forces are not strictly additive. It is more correct to think of the solvation force as *the* van der Waals force at small separations with the molecular properties and density variations of the medium taken into account. It is also important to appreciate that solvation forces do not arise simply because liquid molecules tend to structure into semi-ordered layers at surfaces. They arise because of the disruption or *change* of this ordering during the approach of a second surface. The two effects are related; the greater the tendency toward structuring at an isolated surface the greater the solvation force between two such surfaces, but there is a real distinction between the two phenomena that should be borne in mind.

Oscillatory forces lead to different adhesion values depending on the energy minimum from which two surfaces are being separated. For an interaction energy described by (18.16), "quantized" adhesion energies will be E_0 at $D = 0$ (primary minimum), E_0/e at $D = \sigma$, E_0/e^2 at $D = 2\sigma$, etc. E_0 can be thought of as a depletion force (see Sect. 18.4.5) that is approximately given by the osmotic limit $E_0 \approx -k_B T/\sigma^2$, which can exceed the contribution to the adhesion energy in contact from the van der Waals forces (at $D_0 \approx 0.15–0.20$ nm, as discussed in Sect. 18.3.1, keeping in mind that the

Lifshitz theory fails to describe the force law at *intermediate* distances). Such multivalued adhesion forces have been observed in a number of systems, including the interactions of fibers.

Measurements of oscillatory forces between different surfaces across both aqueous and nonaqueous liquids have revealed their richness of properties [18.118–121], for example, their great sensitivity to the shape and rigidity of the solvent molecules, to the presence of other components, and to the structure of the confining surfaces (see Sects. 18.5.3 and 18.9). In particular, the oscillations can be smeared out if the molecules are irregularly shaped (e.g., branched) and therefore unable to pack into ordered layers, or when the interacting surfaces are rough or fluid-like (see Sect. 18.4.6).

It is easy to understand how oscillatory forces arise between two flat, plane parallel surfaces. Between two curved surfaces, e.g., two spheres, one might imagine the molecular ordering and oscillatory forces to be smeared out in the same way that they are smeared out between two randomly rough surfaces (see Sect. 18.5.3); however, this is not the case. Ordering can occur so long as the curvature or roughness is itself regular or uniform, i. e., not random. This is due to the Derjaguin approximation (18.3). If the energy between two flat surfaces is given by a decaying oscillatory function (for example, a cosine function as in (18.16)), then the force (and energy) between two curved surfaces will also be an oscillatory function of distance with some phase shift. Likewise, two surfaces with regularly curved regions will also retain their oscillatory force profile, albeit modified, as long as the corrugations are truly regular, i. e., periodic. On the other hand, surface roughness, even on the nanometer scale, can smear out oscillations if the roughness is random and the confined molecules are smaller than the size of the surface asperities [18.122, 123]. If an organic liquid contains small amounts of water, the expected oscillatory force can be replaced by a strongly attractive capillary force (see Sect. 18.5.1).

18.4.4 Hydration and Hydrophobic Forces

The forces occurring in water and electrolyte solutions are more complex than those occurring in nonpolar liquids. According to continuum theories, the attractive van der Waals force is always expected to win over the repulsive electrostatic “double-layer” force at small surface separations (Fig. 18.6). However, certain surfaces (usually oxide or hydroxide surfaces such as clays or silica) swell spontaneously or repel each other in aqueous solution, even at high salt concentrations. Yet in all these systems one would expect the surfaces or particles to remain in strong adhesive contact or coagulate in a primary minimum if the only forces operating were DLVO forces.

There are many other aqueous systems in which the DLVO theory fails and where there is an additional short-range force that is not oscillatory but monotonic. Between hydrophilic surfaces this force is exponentially repulsive and is commonly referred to as the *hydration*, or *structural*, force. The origin and nature of this force has long been controversial, especially in the colloidal and biological literature. Repulsive hydration forces are believed to arise from strongly hydrogen-bonding surface groups, such as hydrated ions or hydroxyl (−OH) groups, which modify the hydrogen-bonding network of liquid water adjacent to them. Because this network is quite extensive in range [18.124], the resulting interaction force is also of relatively long range.

Repulsive hydration forces were first extensively studied between clay surfaces [18.125]. More recently, they have been measured in detail between mica and silica surfaces [18.20–22], where they have been found to decay exponentially with decay lengths of about 1 nm. Their effective range is about 3 to 5 nm, which is about twice the range of the oscillatory solvation force in water. Empirically, the hydration repulsion between two hydrophilic surfaces appears to follow the simple equation

$$E = E_0\,\mathrm{e}^{-D/\lambda_0}\,, \tag{18.17}$$

where $\lambda_0 \approx 0.6{-}1.1\,\mathrm{nm}$ for 1:1 electrolytes and $E_0 = 3{-}30\,\mathrm{mJ\,m^{-2}}$ depending on the hydration (hydrophilicity) of the surfaces, higher E_0 values generally being associated with lower λ_0 values.

The interactions between molecularly smooth mica surfaces in dilute electrolyte solutions obey the DLVO theory (Fig. 18.7b). However, at higher salt concentrations, specific to each electrolyte, hydrated cations bind to the negatively charged surfaces and give rise to a repulsive hydration force [18.20, 21]. This is believed to be due to the energy needed to dehydrate the bound cations, which presumably retain some of their water of hydration on binding. This conclusion was arrived at after noting that the strength and range of the hydration forces increase with the known hydration numbers of the electrolyte cations in the order: $\mathrm{Mg^{2+} > Ca^{2+} > Li^{+} \sim Na^{+} > K^{+} > Cs^{+}}$. Similar trends are observed with other negatively charged colloidal surfaces.

While the hydration force between two mica surfaces is overall repulsive below a distance of 4 nm, it is not always monotonic below about 1.5 nm but exhibits oscillations of mean periodicity of 0.25 ± 0.03 nm, roughly equal to the diameter of the water molecule. This is shown in the insert in Fig. 18.7b, where we may note that the first three minima at $D = 0$, 0.28, and 0.56 nm occur at negative energies, a result that rationalizes observations on certain colloidal systems. For example, clay platelets such as montmorillonite often repel each other increasingly strongly as they come closer together, but they are also known to stack into stable aggregates with water interlayers of typical thickness 0.25 and 0.55 nm between them [18.126, 127], suggestive of a turnabout in the force law from a monotonic repulsion to discretized attraction. In chemistry we would refer to such structures as stable hydrates of fixed stoichiometry, whereas in physics we may think of them as experiencing an oscillatory force.

Both surface force and clay swelling experiments have shown that hydration forces can be modified or "regulated" by exchanging ions of different hydration on surfaces, an effect that has important practical applications in controlling the stability of colloidal dispersions. It has long been known that colloidal particles can be precipitated (coagulated or flocculated) by increasing the electrolyte concentration, an effect that was traditionally attributed to the reduced screening of the electrostatic double-layer repulsion between the particles due to the reduced Debye length. However, there are many examples where colloids are stabilized at high salt concentrations, not at low concentrations. This effect is now recognized as being due to the increased hydration repulsion experienced by certain surfaces when they bind highly hydrated ions at higher salt concentrations. Hydration regulation of adhesion and interparticle forces is an important practical method for controlling various processes such as clay swelling [18.126, 127], ceramic processing and rheology [18.128, 129], material fracture [18.128], and colloidal particle and bubble coalescence [18.130].

Water appears to be unique in having a solvation (hydration) force that exhibits both a monotonic and an oscillatory component. Between hydrophilic surfaces the monotonic component is repulsive (Fig. 18.7b), but between hydrophobic surfaces it is attractive and the final adhesion is much greater than expected from the Lifshitz theory.

A hydrophobic surface is one that is inert to water in the sense that it cannot bind to water molecules via ionic or hydrogen bonds. Hydrocarbons and fluorocarbons are hydrophobic, as is air, and the strongly attractive hydrophobic force has many important manifestations and consequences such as the low solubility or miscibility of water and oil molecules, micellization, protein folding, strong adhesion and rapid coagulation of hydrophobic surfaces, non-wetting of water on hydrophobic surfaces, and hydrophobic particle attachment to rising air bubbles (the basic principle of froth flotation).

In recent years, there has been a steady accumulation of experimental data on the force laws between various hydrophobic surfaces in aqueous solution [18.131–143]. These studies have found that the hydrophobic force law between two macroscopic surfaces is of surprisingly long range, decaying exponentially with a characteristic decay length of 1 to 2 nm in the separation range of 0 to 10 nm, and then more gradually further out. The hydrophobic force can be far stronger than the van der Waals attraction, especially between hydrocarbon surfaces in water, for which the Hamaker constant is quite small. The magnitude of the hydrophobic attraction has been found to decrease with the decreasing hydrophobicity (increasing hydrophilicity) of lecithin lipid bilayer surfaces [18.31] and silanated surfaces [18.139], whereas examples of the opposite trend have been shown for some Langmuir–Blodgett-deposited monolayers [18.144].

For two surfaces in water the purely hydrophobic interaction energy (ignoring DLVO and oscillatory forces) in the range of 0 to 10 nm is given by

$$E = -2\gamma \, \mathrm{e}^{-D/\lambda_0} \, , \qquad (18.18)$$

where typically $\lambda_0 = 1–2$ nm, and $\gamma = 10–50\,\mathrm{mJ\,m^{-2}}$. The higher value corresponds to the interfacial energy of a pure hydrocarbon–water interface.

At a separation below 10 nm, the hydrophobic force appears to be insensitive or only weakly sensitive to changes in the type and concentration of electrolyte ions in the solution. The absence of a "screening" effect by ions attests to the non-electrostatic origin of this interaction. In contrast, some experiments have shown that at separations greater than 10 nm, the attraction does depend on the intervening electrolyte, and that in dilute solutions, or solutions containing divalent ions, it can continue to exceed the van der Waals attraction out to separations of 80 nm [18.136, 145].

The long-range nature of the hydrophobic interaction has a number of important consequences. It accounts for the rapid coagulation of hydrophobic particles in water and may also account for the rapid folding of proteins. It also explains the ease with which water films rupture on hydrophobic surfaces. In this case, the van der Waals force across the water film is repulsive and therefore favors wetting, but this is more than offset by the attractive hydrophobic interaction acting between the two hydrophobic phases across water. Hydrophobic forces are increasingly being implicated in the adhesion and fusion of biological membranes and cells. It is known that both osmotic and electric field stresses enhance membrane fusion, an effect that may be due to the concomitant increase in the hydrophobic area exposed between two adjacent surfaces.

From the previous discussion we can infer that hydration and hydrophobic forces are not of a simple nature. These interactions are probably the most important, yet the least understood of all the forces in aqueous solutions. The unusual properties of water and the nature of the surfaces (including their homogeneity and stability) appear to be equally important. Some particle surfaces can have their hydration forces regulated, for example, by ion exchange. Others appear to be intrinsically hydrophilic (e.g., silica) and cannot be coagulated by changing the ionic condition, but can be rendered hydrophobic by chemically modifying their surface groups. For example, on heating silica to above 600 °C, two adjacent surface silanol (−OH) groups release a water molecule and form a hydrophobic siloxane (−O−) group, whence the repulsive hydration force changes into an attractive hydrophobic force.

How do these exponentially decaying repulsive or attractive forces arise? Theoretical work and computer simulations [18.109, 111, 146, 147] suggest that the solvation forces in water should be purely oscillatory, whereas other theoretical studies [18.148–153] suggest a monotonically exponential repulsion or attraction, possibly superimposed on an oscillatory force. The latter is consistent with experimental findings, as shown in the inset to Fig. 18.7b, where it appears that the oscillatory force is simply additive with the monotonic hydration and DLVO forces, suggesting that these arise from essentially different mechanisms.

It is probable that the short-range hydration force between all smooth, rigid, or crystalline surfaces (e.g., mineral surfaces such as mica) has an oscillatory component. This may or may not be superimposed on a monotonic force due to image interactions [18.150] and/or structural or hydrogen bonding interactions [18.148, 149].

Like the repulsive hydration force, the origin of the hydrophobic force is still unknown. *Luzar* et al. [18.152] carried out a Monte Carlo simulation of the interaction between two hydrophobic surfaces across water at separations below 1.5 nm. They obtained a decaying oscillatory force superimposed on a monotonically attractive curve. In more recent work [18.154, 155], it has been suggested that hydrophobic surfaces generate a depleted region of water around them, and that a long-range attractive force due to depletion arises between two such surfaces.

It is questionable whether the hydration or hydrophobic force should be viewed as an ordinary type of solvation or structural force that is reflecting the packing of water molecules. The energy (or entropy) associated with the hydrogen bonding network, which extends over a much larger region of space than the molecular correlations, is probably at the root of the long-range interactions of water. The situation in water appears to be governed by much more than the molecular packing effects that dominate the interactions in simpler liquids.

18.4.5 Polymer-Mediated Forces

Polymers or macromolecules are chain-like molecules consisting of many identical segments (monomers or repeating units) held together by covalent bonds. The size of a polymer coil in solution or in the melt is determined by a balance between van der Waals attraction (and hydrogen bonding, if present) between polymer segments, and the entropy of mixing, which causes the polymer coil to expand. In polymer melts above the glass transition temperature, and at certain conditions in solution, the attraction between polymer segments is exactly balanced by the entropy effect. The polymer solution will then behave virtually ideally, and the density distribution of segments in the coil is Gaussian. This is called the theta (θ) condition, and it occurs at the theta or Flory temperature for a particular combination of polymer and solvent or solvent mixture. At lower temperatures (in a poor or bad solvent), the polymer–polymer interactions dominate over the entropic, and the coil will shrink or precipitate. At higher temperatures (good solvent conditions), the polymer coil will be expanded.

High molecular weight polymers form large coils, which significantly affect the properties of a solution even when the total mass of polymer is very low. The radius of the polymer coil is proportional to the segment length, a, and the number of segments, n. At theta conditions, the hydrodynamic radius of the polymer coil (the root-mean-square separation of the ends of one polymer chain) is theoretically given by $R_\mathrm{h} = a\,n^{1/2}$, and the unperturbed radius of gyration (the average root-mean-square distance of a segment from the center of mass of the molecule) is $R_\mathrm{g} = a\,(n/6)^{1/2}$. In a good solvent the perturbed size of the polymer coil, the Flory radius R_F, is proportional to $n^{3/5}$.

Polymers interact with surfaces mainly through van der Waals and electrostatic interactions. The physisorption of polymers containing only one type of segment is reversible and highly dynamic, but the rate of exchange of adsorbed chains with free chains in the solution is low, since the polymer remains bound to the surface as long as one segment along the chain is adsorbed. The adsorption energy per segment is on the order of $k_B T$. In a good solvent, the conformation of a polymer on a surface is very different from the coil conformation in bulk solution. Polymers adsorb in "trains", separated by "loops" extending into solution and dangling "tails" (the ends of the chain). Compared to adsorption at lower temperatures, good solvent conditions favor more of the polymer chain being in the solvent, where it can attain its optimum conformation. As a result, the extension of the polymer is longer, even though the total amount of adsorbed polymer is lower. In a good solvent, the polymer chains can also be effectively repelled from a surface, if the loss in conformational entropy close to the surface is not compensated for by a gain in enthalpy from adsorption of segments. In this case, there will be a layer of solution (thickness $\approx R_g$) close to the surfaces that is depleted of polymer.

The interaction forces between two surfaces across a polymer solution will depend on whether the polymer is adsorbing onto the surfaces or is repelled from them, and also on whether the interaction occurs at "true" or "restricted" thermodynamic equilibrium. At true or full equilibrium, the polymer between the surfaces can equilibrate (exchange) with polymer in the bulk solution at all surface separations. Some theories [18.156, 157] predict that at full equilibrium, the polymer chains would move from the confined gap into the bulk solution where they could attain entropically more favorable conformations, and that a monotonic attraction at all distances would result from bridging and depletion interactions (which will be discussed below). Other theories suggest that the interaction at small separations would be ultimately repulsive, since some polymer chains would remain in the gap due to their attractive interactions with many sites on the surface (enthalpic) – more sites would be available to the remaining polymer chains if some others desorbed and diffused out from the gap [18.64, 158, 159].

At restricted equilibrium, the polymer is kinetically trapped, and the adsorbed amount is thus constant as the surfaces are brought toward each other, but the chains can still rearrange on the surfaces and in the gap. Experimentally, the true equilibrium situation is very difficult to attain, and most experiments are done at restricted equilibrium conditions. Even the equilibration of conformations assumed in theoretical models for restricted equilibrium conditions can be so slow that this condition is difficult to reach.

In systems of adsorbing polymer, bridging of chains from one surface to the other can give rise to a long-range attraction, since the bridging chains would gain conformational entropy if the surfaces were closer together. In poor solvents, both bridging and intersegment interactions contribute to an attraction [18.26]. However, regardless of solvent and equilibrium conditions, a strong repulsion due to the osmotic interactions is seen as small surface separations in systems of adsorbing polymers at restricted equilibrium.

In systems containing high concentrations of non-adsorbing polymer, the difference in solute concentration in the bulk and between the surfaces at separations smaller than the approximate polymer coil diameter ($2R_g$, i. e., when the polymer has been squeezed out from the gap between the surfaces) may give rise to an attractive osmotic force ("depletion attraction") [18.160–164]. In addition, if the polymer coils become initially compressed as the surfaces approach each other, this can give rise to a repulsion ("depletion stabilization") at large separations [18.163]. For a system of two cylindrical surfaces or radius R, the maximum depletion force, F_{dep}, is expected to occur when the surfaces are in contact and is given by multiplying the depletion (osmotic) pressure, $P_{dep} = \rho k_B T$, by the contact area πr^2, where r is given by the chord theorem: $r^2 = (2R - R_g)R_g \approx 2RR_g$ [18.3]:

$$F_{dep}/R = -2\pi R_g \rho k_B T \;, \qquad (18.19)$$

where ρ is the number density of the polymer in the bulk solution.

If a part of the polymer (typically an end group) is different from the rest of the chain, this part may preferentially adsorb to the surface. End-adsorbed polymer is attached to the surface at only one point, and the extension of the chain is dependent on the grafting density, i. e., the average distance, s, between adsorbed end groups on the surface (Fig. 18.8). One distinguishes between different regions of increasing overlap of the chains (stretching) called pancake, mushroom, and brush regimes [18.165]. In the mushroom regime, where the coverage is sufficiently low so that there is no overlap between neighboring chains, the thickness of the adsorbed layer is proportional to $n^{1/2}$ (i. e., to R_g) at theta conditions and to $n^{3/5}$ in a good solvent.

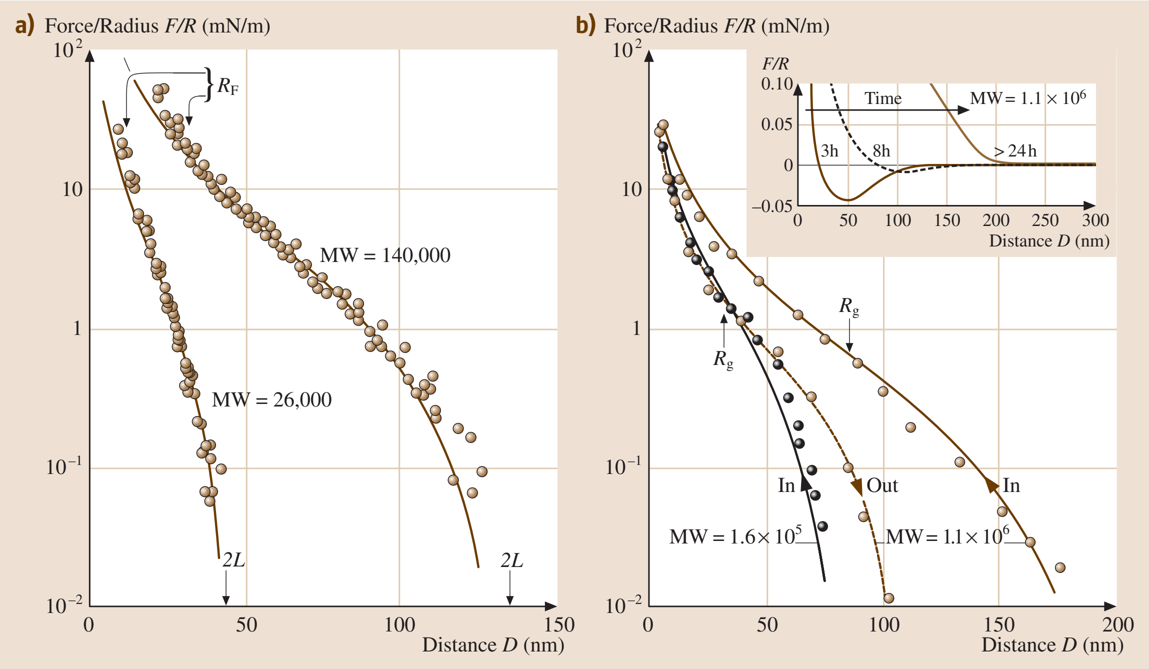

Fig. 18.8a,b Experimentally determined forces in systems of two interacting polymer brushes: **(a)** Polystyrene brush layers grafted via an adsorbing chain-end group onto mica surfaces in toluene (a good solvent for polystyrene). *Left curve:* $\mathrm{MW} = 26{,}000\,\mathrm{g/mol}$, $R_\mathrm{F} = 12\,\mathrm{nm}$. *Right curve*: $\mathrm{MW} = 140{,}000\,\mathrm{g/mol}$, $R_\mathrm{F} = 32\,\mathrm{nm}$. Both force curves were reversible on approach and separation. The *solid curves* are theoretical fits using the Alexander–de Gennes theory with the following measured parameters: spacing between attachments sites: $s = 8.5\,\mathrm{nm}$, brush thickness: $L = 22.5\,\mathrm{nm}$ and 65 nm, respectively. (Adapted from [18.171]). **(b)** Polyethylene oxide layers physisorbed onto mica from $150\,\mu\mathrm{g/ml}$ solution in aqueous 0.1 M KNO_3 (a good solvent for polyethylene oxide). *Main figure:* Equilibrium forces at full coverage after $\sim 16\,\mathrm{h}$ adsorption time. *Left curve:* $\mathrm{MW} = 160{,}000\,\mathrm{g/mol}$, $R_\mathrm{g} = 32\,\mathrm{nm}$. *Right curve:* $\mathrm{MW} = 1{,}100{,}000\,\mathrm{g/mol}$, $R_\mathrm{g} = 86\,\mathrm{nm}$. Note the hysteresis (irreversibility) on approach and separation for this *physisorbed* layer, in contrast to the absence of hysteresis with *grafted* chains in case **(a)**. The *solid curves* are based on a modified form of the Alexander–de Gennes theory. *Inset* in **(b)**: Evolution of the forces with the time allowed for the higher MW polymer to adsorb from solution. Note the gradual reduction in the attractive bridging component. (Adapted from [18.172–174]. After [18.3]. Copyright 1991, with permission from Elsevier Science)

Several models [18.165–170] have been developed for the extension and interactions between two brushes (strongly stretched grafted chains). They are based on a balance between osmotic pressure within the brush layers (uncompressed and compressed) and elastic energy of the chains and differ mainly in the assumptions of the segment density profile, which can be a step function or parabolic. At high coverage (in the brush regime),where the chains will avoid overlapping each other, the thickness of the layer is proportional to n.

Experimental work on both monodisperse [18.27, 28, 171, 175] and polydisperse [18.30] systems at different solvent conditions has confirmed the expected range and magnitude of the repulsive interactions resulting from compression of densely packed grafted layers.

18.4.6 Thermal Fluctuation Forces

If a surface is not rigid but very soft or even fluid-like, this can act to smear out any oscillatory solvation force. This is because the thermal fluctuations of such interfaces make them dynamically "rough" at any instant, even though they may be perfectly smooth on a time average. The types of surfaces that fall into this category are fluid-like amphiphilic surfaces of micelles, bilayers, emulsions, soap films, etc., but also solid colloidal particle surfaces that are coated with surfactant monolayers, as occurs in lubricating oils, paints, toners, etc.

These thermal fluctuation forces (also called entropic or steric forces) are usually short-range and repulsive and are very effective at stabilizing the attractive van der Waals forces at some small but finite separation. This can reduce the adhesion energy or force by up to three orders of magnitude. It is mainly for this reason that fluid-like micelles and bilayers, biological membranes, emulsion droplets, or gas bubbles adhere to each other only very weakly.

Because of their short range it was, and still is, commonly believed that these forces arise from water ordering or "structuring" effects at surfaces, and that they reflect some unique or characteristic property of water. However, it is now known that these repulsive forces also exist in other liquids. Moreover, they appear to become stronger with increasing temperature, which is unlikely if the force would originate from molecular ordering effects at surfaces. Recent experiments, theory, and computer simulations [18.176–178] have shown that these repulsive forces have an entropic origin arising from the osmotic repulsion between exposed thermally mobile surface groups once these overlap in a liquid. These phenomena include undulating and peristaltic forces between membranes or bilayers, and, on the molecular scale, protrusion and head group overlap forces where the interactions also are influenced by hydration forces.

18.5 Adhesion and Capillary Forces

18.5.1 Capillary Forces

When considering the adhesion of two solid surfaces or particles in air or in a liquid, it is easy to overlook or underestimate the important role of capillary forces, i. e., forces arising from the Laplace pressure of curved menisci formed by condensation of a liquid between and around two adhering surfaces (Fig. 18.9).

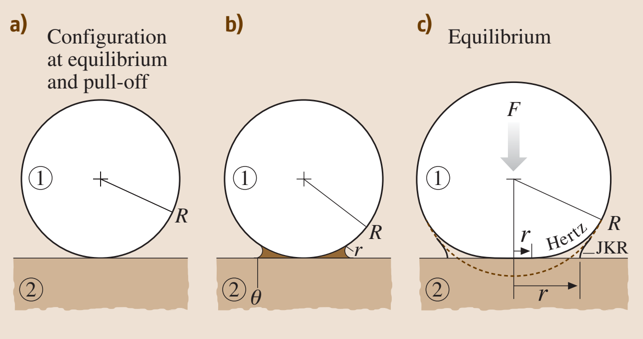

Fig. 18.9a–c Adhesion and capillary forces: **(a)** a non-deforming sphere on a rigid, flat surface in an inert atmosphere and **(b)** in a vapor that can "capillary condense" around the contact zone. At equilibrium, the concave radius, r, of the liquid meniscus is given by the Kelvin equation. For a concave meniscus to form, the contact angle θ has to be less than 90°. In the case of hydrophobic surfaces surrounded by water, a vapor cavity can form between the surfaces. As long as the surfaces are perfectly smooth, the contribution of the meniscus to the adhesion force is independent of r. (After [18.1] with permission.) **(c)** Elastically deformable sphere on a rigid flat surface in the absence (Hertz) and presence (JKR) of adhesion [**(a)** and **(c)** after [18.3]. Copyright 1991, with permission from Elsevier Science]

The adhesion force between a non-deformable spherical particle of radius R and a flat surface in an inert atmosphere (Fig. 18.9a) is

$$F_\mathrm{s} = 4\pi R \gamma_\mathrm{SV}\ . \tag{18.20}$$

But in an atmosphere containing a condensable vapor, the expression above is replaced by

$$F_\mathrm{s} = 4\pi R(\gamma_\mathrm{LV}\cos\theta + \gamma_\mathrm{SL})\ , \tag{18.21}$$

where the first term is due to the Laplace pressure of the meniscus and the second is due to the direct adhesion of the two contacting solids within the liquid. Note that the above equation does not contain the radius of curvature, r, of the liquid meniscus (see Fig. 18.9b). This is because for smaller r the Laplace pressure γ_LV/r increases, but the area over which it acts decreases by the same amount, so the two effects cancel out. Experiments with inert liquids, such as hydrocarbons, condensing between two mica surfaces indicate that (18.21) is valid for values of r as small as 1–2 nm, corresponding to vapor pressures as low as 40% of saturation [18.121, 179, 180]. Capillary condensation also occurs in binary liquid systems, e.g., when water dissolved in hydrocarbon liquids condense around two contacting hydrophilic surfaces or when a vapor cavity forms in water around two hydrophobic surfaces. In the case of water condensing from vapor or from oil, it also appears that the bulk value of γ_LV is applicable for meniscus radii as small as 2 nm.

The capillary condensation of liquids, especially water, from vapor can have additional effects on the

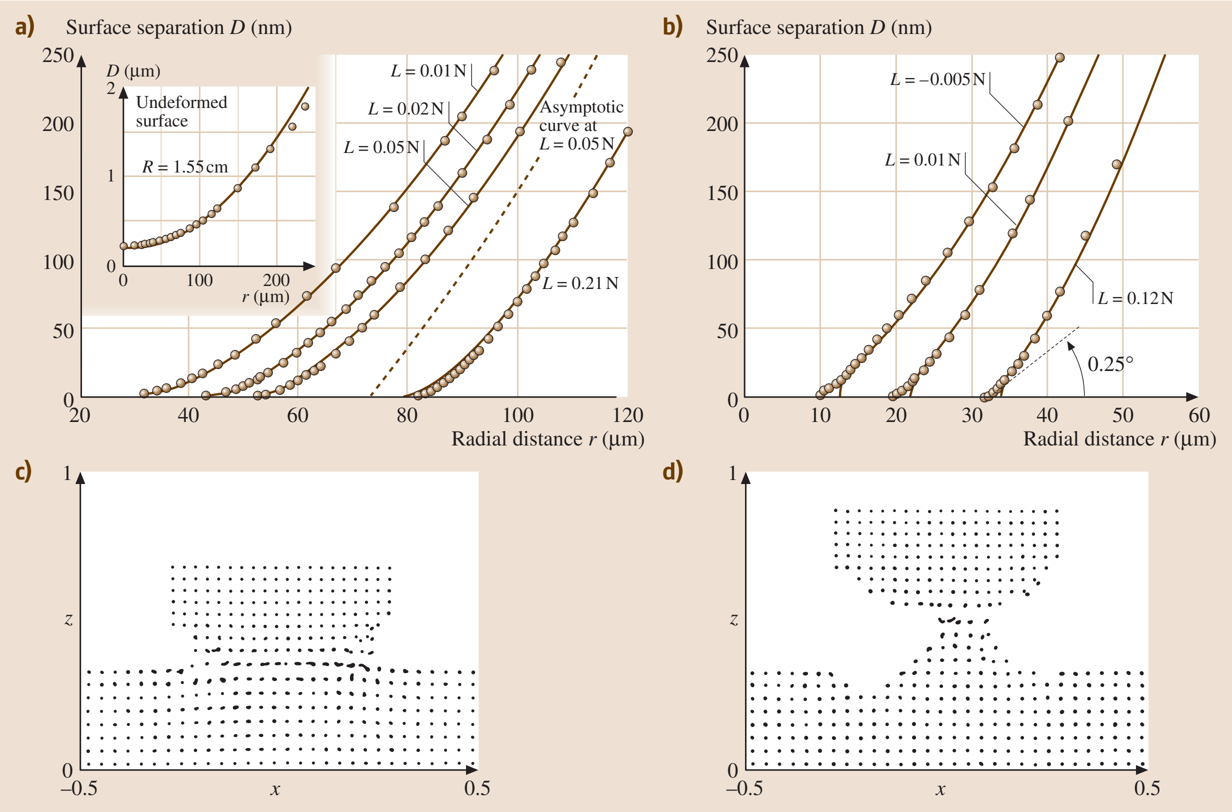

Fig. 18.10a–d Experimental and computer simulation data on contact mechanics for ideal Hertz and JKR contacts. **(a)** Measured profiles of surfaces in nonadhesive contact (*circles*) compared with Hertz profiles (*continuous curves*). The system was mica surfaces in a concentrated KCl solution in which they do not adhere. When not in contact, the surface shape is accurately described by a sphere of radius $R = 1.55$ cm (*inset*). The applied loads were 0.01, 0.02, 0.05, and 0.21 N. The last profile was measured in a different region of the surfaces where the local radius of curvature was 1.45 cm. The Hertz profiles correspond to central displacements of $\delta = 66.5$, 124, 173, and 441 nm. The *dashed line* shows the shape of the undeformed sphere corresponding to the curve at a load of 0.05 N; it fits the experimental points at larger distances (not shown). **(b)** Surface profiles measured with adhesive contact (mica surfaces adhering in dry nitrogen gas) at applied loads of -0.005, 0.01, and 0.12 N. The continuous lines are JKR profiles obtained by adjusting the central displacement in each case to get the best fit to points at larger distances. The values are $\delta = -4.2$, 75.6, and 256 nm. Note that the scales of this figure exaggerate the apparent angle at the junction of the surfaces. This angle, which is insensitive to load, is only about 0.25°. **(c)** and **(d)** Molecular dynamics simulation illustrating the formation of a connective neck between an Ni tip (topmost eight layers) and an Au substrate. The figures show the atomic configuration in a slice through the system at indentation **(c)** and during separation **(d)**. Note the crystalline structure of the neck. Distances are given in units of x and z, where $x = 1$ and $z = 1$ correspond to 6.12 nm. [**(a)** and **(b)** after [18.181]. Copyright 1987, with permission from Elsevier Science. **(c)** and **(d)** after [18.91], with kind permission from Kluwer Academic Publishers]

physical state of the contact zone. For example, if the surfaces contain ions, these will diffuse and build up within the liquid bridge, thereby changing the chemical composition of the contact zone, as well as influencing the adhesion. In the case of surfaces covered with surfactant or polymer molecules (amphiphilic surfaces), the molecules can overturn on exposure to humid air, so that the surface nonpolar groups become replaced by polar groups, which renders the surfaces hydrophilic. When two such surfaces come into contact, water will condense around the contact zone and the adhesion force will also be affected – generally increasing well above the value expected for inert hydrophobic surfaces. It is apparent that the adhesion in vapor or a solvent is often largely determined by capillary forces arising from the condensation of liquid that may be present only in very small quantities, e.g., 10–20 % of saturation in the vapor, or 20 ppm in the solvent.

18.5.2 Adhesion Mechanics

Two bodies in contact deform as a result of surface forces and/or applied normal forces. For the simplest case of two interacting elastic spheres (a model that is easily extended to an elastic sphere interacting with an undeformable surface (or vice versa)) and in the absence of attractive surface forces, the vertical central displacement (compression) was derived by *Hertz* [18.182]. In this model, the displacement and the contact area are equal to zero when no external force (load) is applied, i. e., at the points of contact and of separation.

In systems where attractive surface forces are present between the surfaces, the deformations are more complicated. Modern theories of the adhesion mechanics of two contacting solid surfaces are based on the Johnson–Kendall–Roberts (JKR) theory [18.15, 183], or on the Derjaguin–Muller–Toporov (DMT) theory [18.184–186]. The JKR theory is applicable to easily deformable, large bodies with high surface energy, whereas the DMT theory better describes very small and hard bodies with low surface energy [18.187].

In the JKR theory, two spheres of radii R_1 and R_2, bulk elastic modulus K, and surface energy γ will flatten due to attractive surface forces when in contact at no external load. The contact area will increase under an external load or normal force F, such that at mechanical equilibrium the radius of the contact area, r, is given by

$$r^3 = \frac{R}{K}\left[F + 6\pi R\gamma + \sqrt{12\pi R\gamma F + (6\pi R\gamma)^2}\right] , \tag{18.22}$$

where $R = R_1 R_2/(R_1 + R_2)$. In the absence of surface energy, γ, equation (18.22) is reduced to the expression for the radius of the contact area in the Hertz model. Another important result of the JKR theory gives the adhesion force or "pull-off" force:

$$F_S = -3\pi R\gamma_S , \tag{18.23}$$

where the surface energy, γ_S, is defined as $W = 2\gamma_S$, where W is the reversible work of adhesion. Note that according to the JKR theory a finite elastic modulus, K, while having an effect on the load–area curve, has no effect on the adhesion force, an interesting and unexpected result that has nevertheless been verified experimentally [18.15, 181, 188, 189].

Equation (18.22) and (18.23) provide the framework for analyzing results of adhesion measurements of contacting solids, known as contact mechanics [18.183, 190], and for studying the effects of surface conditions and time on adhesion energy hysteresis (see Sect. 18.5.4).

18.5.3 Effects of Surface Structure, Roughness, and Lattice Mismatch

In a contact between two rough surfaces, the real area of contact varies with the applied load in a different manner than between smooth surfaces [18.191, 192]. For a Hertzian contact, it has been shown that the contact area for rough surfaces increases approximately linearly with the applied normal force (load), F, instead of as $F^{2/3}$ for smooth surfaces. In systems with attractive surface forces, there is a competition between this attraction and repulsive forces arising from compression of high asperities. As a result, the adhesion in such systems can be very low, especially if the surfaces are not easily deformed [18.193–195]. The opposite is possible for soft (viscoelastic) surfaces where the real (molecular) contact area might be larger than for two perfectly smooth surfaces.

Adhesion forces may also vary depending on the commensurability of the crystallographic lattices of the interacting surfaces. *McGuiggan* and *Israelachvili* [18.196] measured the adhesion between two mica surfaces as a function of the orientation (twist angle) of their surface lattices. The forces were measured in air, water, and an aqueous salt solution where oscillatory structural forces were present. In humid air, the adhesion was found to be relatively independent of the twist angle θ due to the adsorption of a 0.4-nm-thick amorphous layer of organics and water at the interface. In contrast, in water, sharp adhesion peaks (energy minima) occurred at $\theta = 0°, \pm 60°, \pm 120°$ and 180°, corresponding to the "coincidence" angles of the surface lattices (Fig. 18.11). As little as ±1° away from these peaks, the energy decreased by 50%. In aqueous KCl solution, due to potassium ion adsorption the water between the surfaces becomes ordered, resulting in an oscillatory force profile where the adhesive minima occur at discrete separations of about 0.25 nm, corresponding to integral numbers of water layers. The whole interaction potential was now found to depend on the orientation of

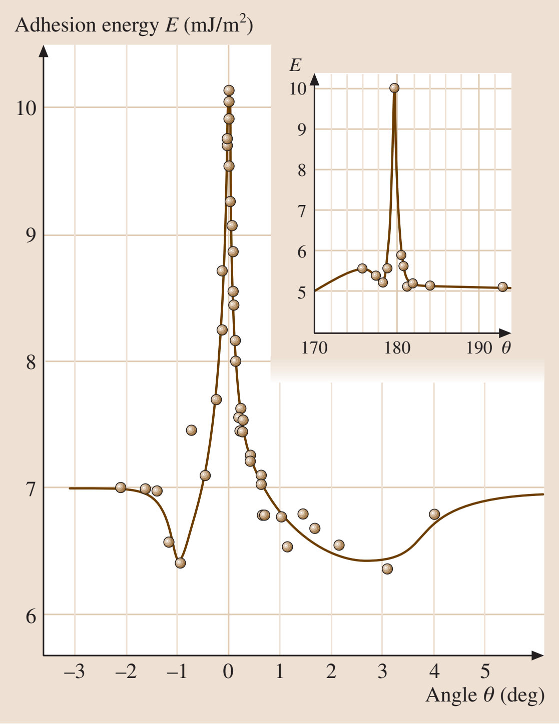

Fig. 18.11 Adhesion energy for two mica surfaces in contact in water (in the primary minimum of an oscillatory force curve) as a function of the mismatch angle θ about $\theta = 0°$ and 180° between the mica surface lattices. (After [18.197] with permission)

the surface lattices, and the effect extended at least four molecular layers.

It has also been appreciated that the structure of the confining surfaces is just as important as the nature of the liquid for determining the solvation forces [18.90, 122, 123, 198–202]. Between two surfaces that are completely flat but "unstructured", the liquid molecules will order into layers, but there will be no lateral ordering within the layers. In other words, there will be positional ordering normal but not parallel to the surfaces. If the surfaces have a crystalline (periodic) lattice, this may induce ordering parallel to the surfaces, as well, and the oscillatory force then also depends on the structure of the surface lattices. Further, if the two lattices have different dimensions ("mismatched" or "incommensurate" lattices), or if the lattices are similar but are not in register relative to each other, the oscillatory force law is further modified [18.196, 203] and the tribological properties of the film are also influenced, as discussed in Sect. 18.9 [18.203, 204].

As shown by the experiments, these effects can alter the magnitude of the adhesive minima found at a given separation within the last one or two nanometers of a thin film by a factor of two. The force barriers (maxima) may also depend on orientation. This could be even more important than the effects on the minima. A high barrier could prevent two surfaces from coming closer together into a much deeper adhesive well. Thus the maxima can effectively contribute to determining not only the final separation of two surfaces, but also their final adhesion. Such considerations should be particularly important for determining the thickness and strength of intergranular spaces in ceramics, the adhesion forces between colloidal particles in concentrated electrolyte solution, and the forces between two surfaces in a crack containing capillary condensed water.

For surfaces that are *randomly* rough, oscillatory forces become smoothed out and disappear altogether, to be replaced by a purely monotonic solvation force [18.116, 122, 123]. This occurs even if the liquid molecules themselves are perfectly capable of ordering into layers. The situation of *symmetric* liquid molecules confined between *rough* surfaces is therefore not unlike that of *asymmetric* molecules between *smooth* surfaces (see Sect. 18.4.3 and Fig. 18.7a). To summarize, for there to be an oscillatory solvation force, the liquid molecules must be able to be correlated over a reasonably long range. This requires that both the liquid molecules and the surfaces have a high degree of order or symmetry. If either is missing, so will the oscillations. A roughness of only a few tenths of a nanometer is often sufficient to eliminate any oscillatory component of the force law.

18.5.4 Nonequilibrium and Rate-Dependent Interactions: Adhesion Hysteresis

Under ideal conditions the adhesion energy is a well-defined thermodynamic quantity. It is normally denoted by E or W (the work of adhesion) or γ (the surface tension, where $W = 2\gamma$) and gives the reversible work done on bringing two surfaces together or the work needed to separate two surfaces from contact. Under ideal, equilibrium conditions these two quantities are the same, but under most realistic conditions they are not: The work needed to separate two surfaces is always greater than that originally gained by bringing them together. An understanding of the molecular mechanisms underlying this phenomenon is essential for understanding many adhesion phenomena, energy dissipation during loading–unloading cycles, contact angle hysteresis, and

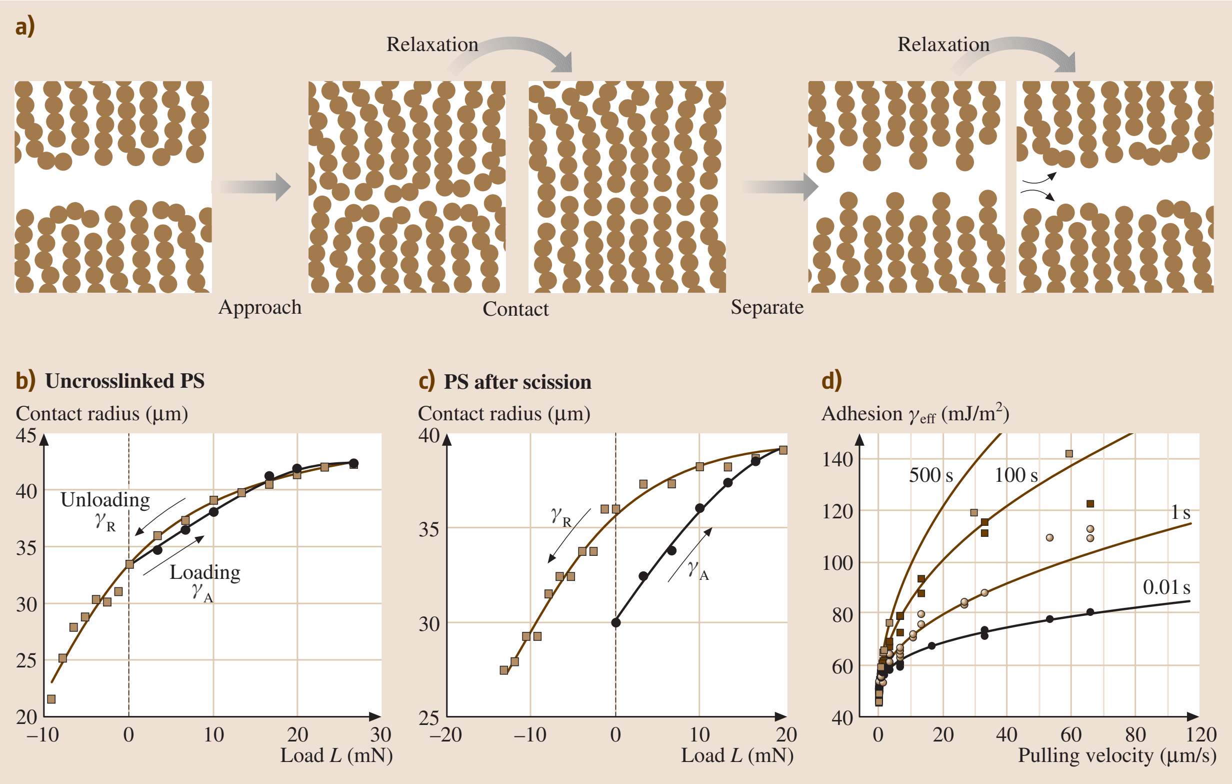

Fig. 18.12 **(a)** Schematic representation of interpenetrating chains. **(b)** and **(c)**: JKR plots (contact radius r as a function of applied load L) showing small adhesion hysteresis for uncrosslinked polystyrene and larger adhesion hysteresis after chain scission at the surfaces after 18 h irradiation with ultraviolet light in an oxygen atmosphere. The adhesion hysteresis continues to increase with the irradiation time. **(d)** Rate-dependent adhesion of CTAB surfactant monolayers. The *solid curves* [18.205] are fits to experimental data on CTAB adhesion after different contact times [18.206] using an approximate analytical solution for a JKR model, including crack tip dissipation. Due to the limited range of validity of the approximation, the fits rely on the low effective adhesion energy part of the experimental data only. From the fits one can determine the thermodynamic adhesion energy, the characteristic dissipation velocity, and the intrinsic dissipation exponent of the model. [**(a)** after [18.207]. Copyright 1993 American Chemical Society. **(b)** and **(c)** after [18.208]. Copyright 2002 American Association for the Advancement of Science. **(d)** after [18.205]. Copyright 2000 American Chemical Society]

the molecular mechanisms associated with many frictional processes.

It is wrong to think that hysteresis arises because of some imperfection in the system such as rough or chemically heterogeneous surfaces, or because the supporting material is viscoelastic. Adhesion hysteresis can arise even between perfectly smooth and chemically homogenous surfaces supported by perfectly elastic materials. It can be responsible for such phenomena as rolling friction and elastoplastic adhesive contacts [18.190, 209–212] during loading–unloading and adhesion–decohesion cycles.

Adhesion hysteresis may be thought of as being due to mechanical effects such as instabilities, or chemical effects such as interdiffusion, interdigitation, molecular reorientations and exchange processes occurring at an interface after contact, as illustrated in Fig. 18.12. Such processes induce roughness and chemical heterogeneity even though initially (and after separation and re-equilibration) both surfaces are perfectly smooth and chemically homogeneous. In general, if the energy change, or work done, on separating two surfaces from adhesive contact is not fully recoverable on bringing the two surfaces back into contact again, the adhesion hysteresis may be expressed as

$$\underset{\text{Receding}}{W_\mathrm{R}} > \underset{\text{Advancing}}{W_\mathrm{A}}$$

or

$$\Delta W = (W_\mathrm{R} - W_\mathrm{A}) > 0 \,, \tag{18.24}$$

where W_R and W_A are the adhesion or surface energies for receding (separating) and advancing (approaching) two solid surfaces, respectively.

Hysteresis effects are also commonly observed in wetting/dewetting phenomena [18.213]. For example, when a liquid spreads and then retracts from a surface the advancing contact angle θ_A is generally larger than the receding angle θ_R. Since the contact angle, θ, is related to the liquid–vapor surface tension, γ_L, and the solid–liquid adhesion energy, W, by the Dupré equation,

$$(1 + \cos\theta)\gamma_\mathrm{L} = W \,, \tag{18.25}$$

we see that *wetting hysteresis* or *contact angle hysteresis* ($\theta_\mathrm{A} > \theta_\mathrm{R}$) actually implies adhesion hysteresis, $W_\mathrm{R} > W_\mathrm{A}$, as given by (18.24).

Energy dissipating processes such as adhesion and contact angle hysteresis arise because of practical constraints of the *finite time* of measurements and the *finite elasticity* of materials. This prevents many loading–unloading or approach–separation cycles from being thermodynamically reversible, even though if carried out infinitely slowly they would be. By thermodynamically irreversible one simply means that one cannot go through the approach–separation cycle via a continuous series of equilibrium states, because some of these are connected via spontaneous – and therefore thermodynamically irreversible – instabilities or transitions where energy is liberated and therefore "lost" via heat or phonon release [18.214]. This is an area of much current interest and activity, especially regarding the fundamental molecular origins of adhesion and friction in polymer and surfactant systems, and the relationships between them [18.190, 205, 206, 208, 210, 215–218].

18.6 Introduction: Different Modes of Friction and the Limits of Continuum Models

Most frictional processes occur with the sliding surfaces becoming damaged in one form or another [18.209]. This may be referred to as "normal" friction. In the case of brittle materials, the damaged surfaces slide past each other while separated by relatively large, micron-sized wear particles. With more ductile surfaces, the damage remains localized to nanometer-sized, plastically deformed asperities. Some features of the friction between damaged surfaces will be described in Sect. 18.7.4.

There are also situations in which sliding can occur between two perfectly smooth, undamaged surfaces. This may be referred to as "interfacial" sliding or

Fig. 18.13 Stribeck curve: an empirical curve giving the trend generally observed in the friction forces or friction coefficients as a function of sliding velocity, the bulk viscosity of the lubricating fluid, and the applied load. The three friction/lubrication regimes are known as the boundary lubrication regime (see Sect. 18.7), the intermediate or mixed lubrication regime (Sect. 18.8.2), and thick film or elastohydrodynamic (EHD) lubrication regime (Sect. 18.8.1). The film thicknesses believed to correspond to each of these regimes are also shown. For thick films, the "friction" force is purely viscous, e.g., Couette flow at low shear rates, but may become complicated at higher shear rates where EHD deformations of surfaces can occur during sliding. (After [18.1] with permission)

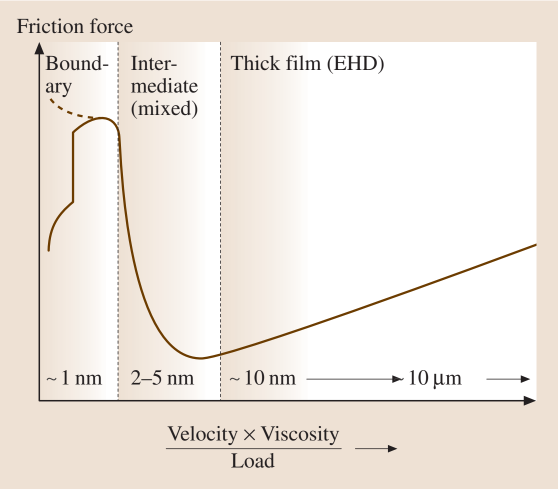

Table 18.4 The three main tribological regimes characterizing the changing properties of liquids subjected to an increasing confinement between two solid surfaces[a]

Regime	Conditions for getting into this regime	Static/equilibrium properties[b]	Dynamic properties[c]
Bulk	• Thick films (> 10 molecular diameters, $\gg R_\mathrm{g}$ for polymers) • Low or zero loads • High shear rates	Bulk (continuum) properties: • Bulk liquid density • No long-range order	Bulk (continuum) properties: • Newtonian viscosity • Fast relaxation times • No glass temperature • No yield point • Elastohydrodynamic lubrication
Intermediate mixed	• Intermediately thick films (4–10 molecular diameters, $\sim R_\mathrm{g}$ for polymers) • Low loads or pressure	Modified fluid properties include: • Modified positional and orientational order[a] • Medium to long-range molecular correlations • Highly entangled states	Modified rheological properties include: • Non-Newtonian flow • Glassy states • Long relaxation times • Mixed lubrication
Boundary	• Molecularly thin films (< 4 molecular diameters) • High loads or pressure • Low shear rates • Smooth surfaces or asperities	Onset of non-fluidlike properties: • Liquid-like to solid-like phase transitions • Appearance of new liquid-crystalline states • Epitaxially induced long-range ordering	Onset of tribological properties: • No flow until yield point or critical shear stress reached • Solid-like film behavior characterized by defect diffusion, dislocation motion, shear melting • Boundary lubrication

Based on work by *Granick* [18.219], *Hu* and *Granick* [18.220], and others [18.38, 207, 221] on the dynamic properties of short chain molecules such as alkanes and polymer melts confined between surfaces

[a] Confinement can lead to an increased or decreased order in a film, depending both on the surface lattice structure and the geometry of the confining cavity

[b] In each regime both the static and dynamic properties change. The static properties include the film density, the density distribution function, the potential of mean force, and various positional and orientational order parameters

[c] Dynamic properties include viscosity, viscoelastic constants, and tribological yield points such as the friction coefficient and critical shear stress

"boundary" friction and is the focus of the following sections. The term "boundary lubrication" is more commonly used to denote the friction of surfaces that contain a thin protective lubricating layer such as a surfactant monolayer, but here we shall use the term more broadly to include any molecularly thin solid, liquid, surfactant, or polymer film.

Experiments have shown that as a liquid film becomes progressively thinner, its physical properties change, at first quantitatively and then qualitatively [18.44, 47, 219, 220, 222, 223]. The quantitative changes are manifested by an increased viscosity, non-Newtonian flow behavior, and the replacement of normal melting by a glass transition, but the film remains recognizable as a liquid (Fig. 18.13). In tribology, this regime is commonly known as the "mixed lubrication" regime, where the rheological properties of a film are intermediate between the bulk and boundary properties. One may also refer to it as the "intermediate" regime (Table 18.4).

For even thinner films, the changes in behavior are more dramatic, resulting in a qualitative change in properties. Thus first-order phase transitions can now occur to solid or liquid-crystalline phases [18.46,201,207,221, 224–227], whose properties can no longer be characterized even qualitatively in terms of bulk or continuum liquid properties such as viscosity. These films now exhibit yield points (characteristic of fracture in solids) and their molecular diffusion and relaxation times can be ten orders of magnitude longer than in the bulk liquid or even in films that are just slightly thicker. The three friction regimes are summarized in Table 18.4.

18.7 Relationship Between Adhesion and Friction Between Dry (Unlubricated and Solid Boundary Lubricated) Surfaces

18.7.1 Amontons' Law and Deviations from It Due to Adhesion: The Cobblestone Model

Early theories and mechanisms for the dependence of friction on the applied normal force or load, L, were developed by *da Vinci*, *Amontons*, *Coulomb* and *Euler* [18.228]. For the macroscopic objects investigated, the friction was found to be directly proportional to the load, with no dependence on the contact area. This is described by the so-called Amontons' law:

$$F = \mu L \, , \tag{18.26}$$

where F is the shear or friction force and μ is a constant defined as the coefficient of friction. This friction law has a broad range of applicability and is still the principal means of quantitatively describing the friction between surfaces. However, particularly in the case of adhering surfaces, Amontons' law does not adequately describe the friction behavior with load, because of the finite friction force measured at zero and even negative applied loads.

When a lateral force, or shear stress, is applied to two surfaces in adhesive contact, the surfaces initially remain "pinned" to each other until some critical shear force is reached. At this point, the surfaces begin to slide past each other either smoothly or in jerks. The frictional force needed to initiate sliding from rest is known as the *static* friction force, denoted by F_s, while the force needed to maintain smooth sliding is referred to as the *kinetic* or *dynamic* friction force, denoted by F_k. In general, $F_s > F_k$. Two sliding surfaces may also move in regular jerks, known as stick–slip sliding, which is discussed in more detail in Sect. 18.8.3. Such friction forces cannot be described by models used for thick films that are viscous (see Sect. 18.8.1) and, therefore, shear as soon as the smallest shear force is applied.

Experimentally, it has been found that during both smooth and stick–slip sliding the local geometry of the contact zone remains largely unchanged from the static geometry. In an adhesive contact, the contact area as a function of load is thus generally well described by the JKR equation, (18.22). The friction force between two molecularly smooth surfaces sliding in adhesive contact is not simply proportional to the applied load, L, as might be expected from Amontons' law. There is an additional adhesion contribution that is proportional to the area of contact, A. Thus, in general, the interfacial friction force of dry, unlubricated surfaces sliding smoothly past each other in adhesive contact is given by

$$F = F_k = S_c A + \mu L \, , \tag{18.27}$$

where S_c is the "critical shear stress" (assumed to be constant), $A = \pi r^2$ is the contact area of radius r given by (18.22), and μ is the coefficient of friction. For low loads we have:

$$\begin{aligned} F &= S_c A = S_c \pi r^2 \\ &= S_c \pi \left[\frac{R}{K} \left(L + 6\pi R\gamma + \sqrt{12\pi R\gamma L + (6\pi R\gamma)^2} \right) \right]^{2/3} ; \end{aligned} \tag{18.28}$$

whereas for high loads (or high μ), (18.27) reduces to Amontons' law: $F = \mu L$. Depending on whether the friction force in (18.27) is dominated by the first or second term, one may refer to the friction as *adhesion-controlled* or *load-controlled*, respectively.

The following friction model, first proposed by *Tabor* [18.229] and developed further by *Sutcliffe* et al. [18.230], *McClelland* [18.231], and *Homola* et al. [18.45], has been quite successful at explaining the interfacial and boundary friction of two solid crystalline surfaces sliding past each other in the absence of wear. The surfaces may be unlubricated, or they may be separated by a monolayer or more of some boundary lubricant or liquid molecules. In this model, the values of the critical shear stress S_c, and coefficient of friction μ, of (18.27) are calculated in terms of the energy needed to overcome the attractive intermolecular forces and compressive externally applied load as one surface is raised and then slid across the molecular-sized asperities of the other.

This model (variously referred to as the *interlocking asperity model*, *Coulomb friction*, or the *cobblestone model*) is similar to pushing a cart over a road of cobblestones where the cartwheels (which represent the molecules of the upper surface or film) must be made to roll over the cobblestones (representing the molecules of the lower surface) before the cart can move. In the case of the cart, the downward force of gravity replaces the attractive intermolecular forces between two material surfaces. When at rest, the cartwheels find grooves between the cobblestones where they sit in potential energy minima, and so the cart is at some stable mechanical equilibrium. A certain lateral force (the "push") is required to raise the cartwheels against the force of gravity in order to initiate motion. Motion will continue as long as the cart is pushed, and rapidly stops once it is no longer pushed. Energy is dissipated by the liberation of heat (phonons, acoustic waves, etc.) every time a wheel hits the next cobblestone. The cobblestone model is not unlike the *Coulomb* and *interlocking asperity* models of friction [18.228] except that it is being applied at the molecular level and for a situation where the external load is augmented by attractive intermolecular forces.

There are thus two contributions to the force pulling two surfaces together: the externally applied load or pressure, and the (internal) attractive intermolecular forces that determine the adhesion between the two surfaces. Each of these contributions affects the friction force in a different way, which we will discuss in more detail below.

18.7.2 Adhesion Force and Load Contribution to Interfacial Friction

Adhesion Force Contribution

Consider the case of two surfaces sliding past each other, as shown in Fig. 18.14. When the two surfaces are initially in adhesive contact, the surface molecules will adjust themselves to fit snugly together [18.233], in an analogous manner to the self-positioning of the cartwheels on the cobblestone road. A small tangential force applied to one surface will therefore not result in the sliding of that surface relative to the other. The attractive van der Waals forces between the surfaces must first be overcome by having the surfaces separate by a small amount. To initiate motion, let the separation between the two surfaces increase by a small amount ΔD, while the lateral distance moved is $\Delta\sigma$. These two values will be related via the geometry of the two surface lattices. The energy put into the system by the force F acting over a lateral distance $\Delta\sigma$ is

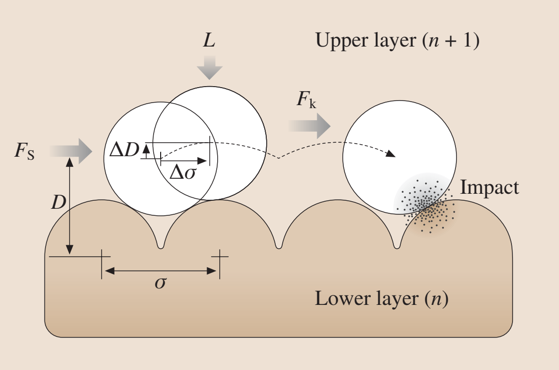

Fig. 18.14 Schematic illustration of how one molecularly smooth surface moves over another when a lateral force F is applied (the "cobblestone model"). As the upper surface moves laterally by some fraction of the lattice dimension $\Delta\sigma$, it must also move up by some fraction of an atomic or molecular dimension ΔD before it can slide across the lower surface. On impact, some fraction ε of the kinetic energy is "transmitted" to the lower surface, the rest being "reflected" back to the colliding molecule (upper surface). (After [18.232] with permission)

$$\text{Input energy: } F\times\Delta\sigma\,. \tag{18.29}$$

This energy may be equated with the change in interfacial or surface energy associated with separating the surfaces by ΔD, i.e., from the equilibrium separation $D = D_0$ to $D = (D_0 + \Delta D)$. Since $\gamma \propto D^{-2}$ for two flat surfaces, the energy cost may be approximated by:

$$\text{Surface energy change}\times\text{area:}$$
$$2\gamma A\left[1-D_0^2/(D_0+\Delta D)^2\right] \approx 4\gamma A(\Delta D/D_0)\,, \tag{18.30}$$

where γ is the surface energy, A the contact area, and D_0 the surface separation at equilibrium. During steady state sliding (kinetic friction), not all of this energy will be "lost" or absorbed by the lattice every time the surface molecules move by one lattice spacing: Some fraction will be reflected during each impact of the "cartwheel" molecules [18.231]. Assuming that a fraction ε of the above surface energy is "lost" every time the surfaces move across the characteristic length $\Delta\sigma$ (Fig. 18.14), we obtain after equating (18.29) and (18.30)

$$S_c = \frac{F}{A} = \frac{4\gamma\varepsilon\Delta D}{D_0\Delta\sigma} \,. \tag{18.31}$$

For a typical hydrocarbon or a van der Waals surface, $\gamma \approx 25\,\mathrm{mJ\,m^{-2}}$. Other typical values would be: $\Delta D \approx 0.05\,\mathrm{nm}$, $D_0 \approx 0.2\,\mathrm{nm}$, $\Delta\sigma \approx 0.1\,\mathrm{nm}$, and $\varepsilon \approx 0.1$–0.5. Using the above parameters, (18.31) predicts $S_c \approx (2.5$–$12.5)\times 10^7\,\mathrm{N\,m^{-2}}$ for van der Waals surfaces. This range of values compares very well with typical experimental values of $2\times 10^7\,\mathrm{N\,m^{-2}}$ for hydrocarbon or mica surfaces sliding in air (see Fig. 18.16) or separated by one molecular layer of cyclohexane [18.45].

The above model suggests that all interfaces, whether dry or lubricated, dilate just before they shear or slip. This is a small but important effect: The dilation provides the crucial extra space needed for the molecules to slide across each other or flow. This dilation has been computed by *Thompson* and *Robbins* [18.201] and *Zaloj* et al. [18.234] and measured by *Dhinojwala* et al. [18.235].

This model may be extended, at least semi-quantitatively, to lubricated sliding, where a thin liquid film is present between the surfaces. With an increase in the number of liquid layers between the surfaces, D_0 increases while ΔD decreases, hence the lower the friction force. This is precisely what is observed, but with more than one liquid layer between two surfaces the situation becomes too complex to analyze analytically (actually, even with one or no interfacial layers, the calculation of the fraction of energy dissipated per molecular collision ε is not a simple matter). Furthermore, even in systems as simple as linear alkanes, interdigitation and interdiffusion have been found to contribute strongly to the properties of the system [18.114, 236]. Sophisticated modeling based on computer simulations is now required, as described in the following section.

Relation Between Boundary Friction and Adhesion Energy Hysteresis

While the above equations suggest that there is a direct correlation between friction and adhesion, this is not the case. The correlation is really between friction and adhesion hysteresis, described in Sect. 18.5.4. In the case of friction, this subtle point is hidden in the factor ε, which is a measure of the amount of energy absorbed (dissipated, transferred, or "lost") by the lower surface when it is impacted by a molecule from the upper surface. If $\varepsilon = 0$, all the energy is reflected, and there will be no kinetic friction force nor any adhesion hysteresis, but the absolute magnitude of the adhesion force or energy will remain finite and unchanged. This is illustrated in Figs. 18.17 and 18.19.

The following simple model shows how adhesion hysteresis and friction may be quantitatively related. Let $\Delta\gamma = \gamma_R - \gamma_A$ be the adhesion energy hysteresis per unit area, as measured during a typical loading–unloading cycle (see Figs. 18.17a and 18.19c,d). Now consider the same two surfaces sliding past each other and assume that frictional energy dissipation occurs through the same mechanism as adhesion energy dissipation, and that both occur over the same characteristic molecular length scale σ. Thus, when the two surfaces (of contact area $A = \pi r^2$) move a distance σ, equating the frictional energy $(F\times\sigma)$ to the dissipated adhesion energy $(A\times\Delta\gamma)$, we obtain

$$\text{Friction force: } F = \frac{A\times\Delta\gamma}{\sigma} = \frac{\pi r^2}{\sigma}(\gamma_R - \gamma_A) \,, \tag{18.32}$$

$$\text{or}\quad \text{Critical shear stress: } S_c = F/A = \Delta\gamma/\sigma \,, \tag{18.33}$$

which is the desired expression and has been found to give order of magnitude agreement between measured friction forces and adhesion energy hysteresis [18.207]. If we equate (18.33) with (18.31), since $4\Delta D/(D_0\Delta\sigma) \approx 1/\sigma$, we obtain the intuitive relation

$$\varepsilon = \frac{\Delta\gamma}{\gamma} \,. \tag{18.34}$$

External Load Contribution to Interfacial Friction

When there is no interfacial adhesion, S_c is zero. Thus, in the absence of any adhesive forces between two surfaces, the only "attractive" force that needs to be overcome for sliding to occur is the externally applied load or pressure.

For a preliminary discussion of this question, it is instructive to compare the magnitudes of the *externally* applied pressure to the *internal* van der Waals pressure between two smooth surfaces. The internal van der Waals pressure between two flat surfaces is given (see Table 18.2) by $P = A_H/6\pi D_0^3 \approx 1\,\mathrm{GPa}$ (10^4 atm), using a typical Hamaker constant of $A_H = 10^{-19}\,\mathrm{J}$, and assuming $D_0 \approx 2\,\mathrm{nm}$ for the equilibrium interatomic spacing. This implies that we should not expect the externally applied load to affect the interfacial friction force F, as defined by (18.27), until the externally applied pressure L/A begins to exceed $\sim 100\,\mathrm{MPa}$ (10^3 atm). This is in agreement with experimental data [18.237] where the effect of load became dominant at pressures in excess of 10^3 atm.

For a more general semiquantitative analysis, again consider the cobblestone model used to derive (18.31), but now include an additional contribution to the surface energy change of (18.30) due to the work done against the external load or pressure, $L\Delta D = P_{\mathrm{ext}}A\Delta D$ (this is equivalent to the work done against gravity in the case of a cart being pushed over cobblestones). Thus:

$$S_{\mathrm{c}} = \frac{F}{A} = \frac{4\gamma\varepsilon\Delta D}{D_0\Delta\sigma} + \frac{P_{\mathrm{ext}}\varepsilon\Delta D}{\Delta\sigma}\,, \qquad (18.35)$$

which gives the more general relation

$$S_{\mathrm{c}} = F/A = C_1 + C_2 P_{\mathrm{ext}}\,, \qquad (18.36)$$

where $P_{\mathrm{ext}} = L/A$ and C_1 and C_2 are constants characteristic of the surfaces and sliding conditions. The constant $C_1 = 4\gamma\varepsilon\Delta D/(D_0\Delta\sigma)$ depends on the mutual adhesion of the two surfaces, while both C_1 and $C_2 = \varepsilon\Delta D/\Delta\sigma$ depend on the topography or atomic bumpiness of the surface groups (Fig. 18.14). The smoother the surface groups the smaller the ratio $\Delta D/\Delta\sigma$ and hence the lower the value of C_2. In addition, both C_1 and C_2 depend on ε (the fraction of energy dissipated per collision), which depends on the relative masses of the shearing molecules, the sliding velocity, the temperature, and the characteristic molecular relaxation processes of the surfaces. This is by far the most difficult parameter to compute, and yet it is the most important since it represents the energy transfer mechanism in any friction process, and since ε can vary between 1 and 0, it determines whether a particular friction force will be large or close to zero. Molecular simulations offer the best way to understand and predict the magnitude of ε, but the complex multi-body nature of the problem makes simple conclusions difficult to draw [18.239–241]. Some of the basic physics of the energy transfer and dissipation of the molecular collisions can be drawn from simplified models such as a 1-D three-body system [18.214].

Finally, the above equation may also be expressed in terms of the friction force F:

$$F = S_{\mathrm{c}}A = C_1 A + C_2 L\,. \qquad (18.37)$$

Equations similar to (18.36) and (18.37) were previously derived by *Derjaguin* [18.242, 243] and by *Briscoe*

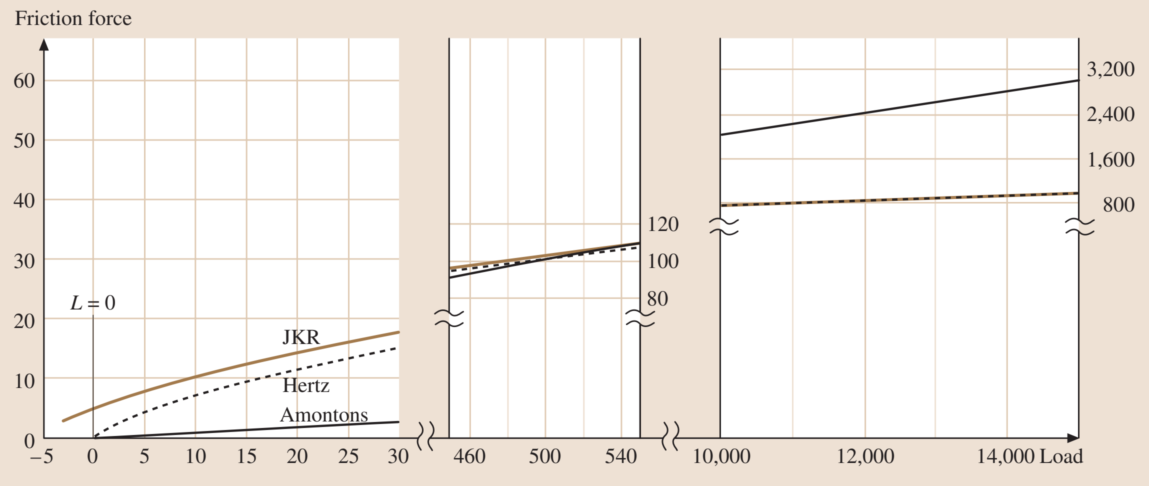

Fig. 18.15 Friction as a function of load for smooth surfaces. At low loads, the friction is dominated by the C_1A term of (18.37). The adhesion contribution (JKR curve) is most prominent near zero load where the Hertzian and Amontons' contributions to the friction are minimal. As the load increases, the adhesion contribution becomes smaller as the JKR and Hertz curves converge. In this range of loads, the linear C_2L contribution surpasses the area contribution to the friction. At much higher loads the explicit load dependence of the friction dominates the interactions, and the observed behavior approaches Amontons' law. It is interesting to note that for smooth surfaces the pressure over the contact area does not increase as rapidly as the load. This is because as the load is increased, the surfaces deform to increase the surface area and thus moderate the contact pressure. (After [18.238] with permission of Kluwer Academic Publishers)

and *Evans* [18.244], where the constants C_1 and C_2 were interpreted somewhat differently than in this model.

In the absence of any attractive interfacial force, we have $C_1 \approx 0$, and the second term in (18.36) and (18.37) should dominate (Fig. 18.15). Such situations typically arise when surfaces repel each other across the lubricating liquid film. In such cases, the total frictional force should be low and should increase *linearly* with the external load according to

$$F = C_2 L \,. \tag{18.38}$$

An example of such lubricated sliding occurs when two mica surfaces slide in water or in salt solution (see Fig. 18.20), where the short-range "hydration" forces between the surfaces are repulsive. Thus, for sliding in 0.5 M KCl it was found that $C_2 = 0.015$ [18.245]. Another case where repulsive surfaces eliminate the adhesive contribution to friction is for polymer chains attached to surfaces at one end and swollen by a good solvent [18.175]. For this class of systems, $C_2 < 0.001$ for a finite range of polymer layer compressions (normal loads, L). The low friction between the surfaces in this regime is attributed to the entropic repulsion between the opposing brush layers with a minimum of entanglement between the two layers. However, with higher normal loads, the brush layers become compressed and begin to entangle, which results in higher friction (see [18.246]).

It is important to note that (18.38) has exactly the same form as Amontons' Law

$$F = \mu L \,, \tag{18.39}$$

where μ is the coefficient of friction.

At the molecular level a thermodynamic analog of the Coulomb or cobblestone models (see Sect. 18.7.1) based on the "contact value theorem" [18.3, 238, 245] can explain why $F \propto L$ also holds at the microscopic or molecular level. In this analysis we consider the surface molecular groups as being momentarily compressed and decompressed as the surfaces move along. Under irreversible conditions, which always occur when a cycle is completed in a finite amount of time, the energy "lost" in the compression–decompression cycle is dissipated as heat. For two non-adhering surfaces, the stabilizing pressure P_i acting locally between any two elemental contact points i of the surfaces may be expressed by the contact value theorem [18.3]:

$$P_i = \rho_i k_B T = k_B T / V_i \,, \tag{18.40}$$

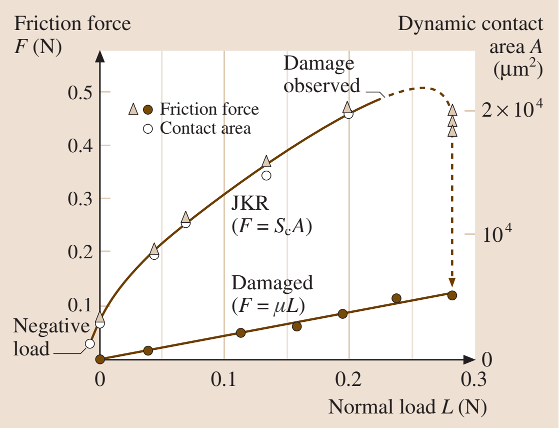

Fig. 18.16 Friction force F and contact area A vs. load L for two mica surfaces sliding in adhesive contact in dry air. The contact area is well described by the JKR theory, (18.22), even during sliding, and the friction force is found to be directly proportional to this area, (18.28). The *vertical dashed line* and *arrow* show the transition from interfacial to normal friction with the onset of wear (*lower curve*). Sliding velocity is 0.2 μm s^{-1}. (After [18.45] with permission, copyright 1989 American Society of Mechanical Engineers)

where $\rho_i = V_i^{-1}$ is the local number density (per unit volume) or activity of the interacting entities, be they molecules, atoms, ions or the electron clouds of atoms. This equation is essentially the osmotic or entropic pressure of a gas of confined molecules. As one surface moves across the other, local regions become compressed and decompressed by a volume ΔV_i. The work done per cycle can be written as $\varepsilon P_i \Delta V_i$, where ε ($\varepsilon \le 1$) is the fraction of energy per cycle "lost" as heat, as defined earlier. The energy balance shows that for each compression–decompression cycle, the dissipated energy is related to the friction force by

$$F_i x_i = \varepsilon P_i \Delta V_i \,, \tag{18.41}$$

where x_i is the lateral distance moved per cycle, which can be the distance between asperities or the distance between surface lattice sites. The pressure at each contact junction can be expressed in terms of the local normal load L_i and local area of contact A_i as $P_i = L_i / A_i$. The volume change over a cycle can thus be expressed as $\Delta V_i = A_i z_i$, where z_i is the vertical distance of confinement. Inserting these into (18.41), we get

$$F_i = \varepsilon L_i (z_i / x_i) \,, \tag{18.42}$$

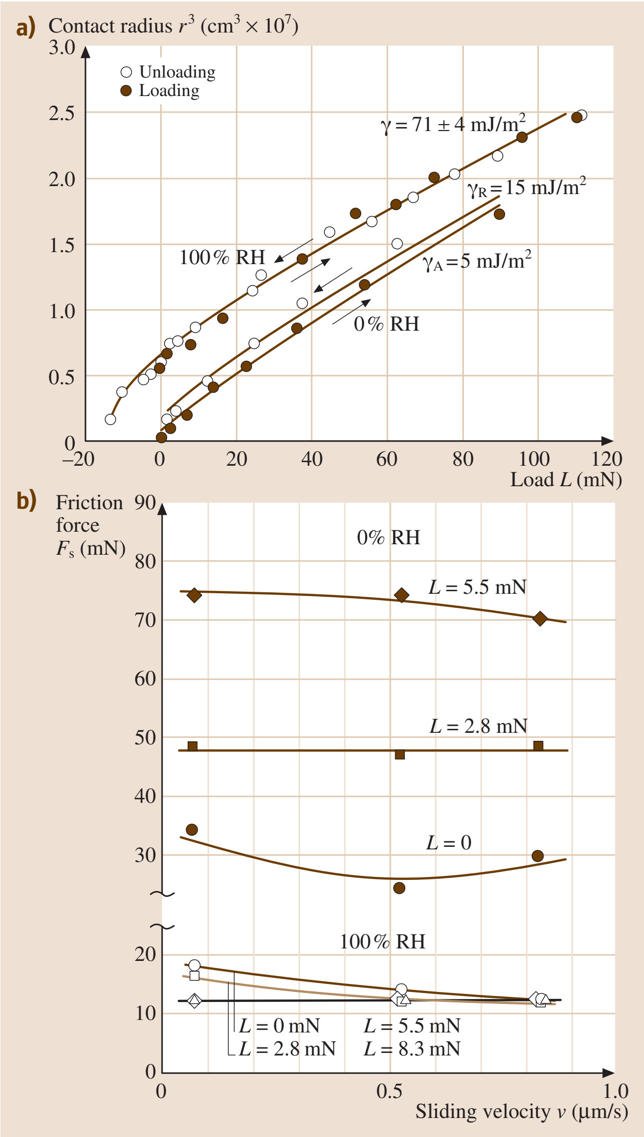

Fig. 18.17 **(a)** Contact radius r versus externally applied load L for loading and unloading of two hydrophilic silica surfaces exposed to dry and humid atmospheres. Note that while the adhesion is higher in humid air, the *hysteresis* in the adhesion is higher in dry air. **(b)** Effect of velocity on the static friction force F_s for hydrophobic (heat-treated electron beam evaporated) silica in dry and humid air. The effects of humidity, load, and sliding velocity on the friction forces, as well as the stick–slip friction of the hydrophobic surfaces, are qualitatively consistent with a "friction" phase diagram representation as in Fig. 18.28. (After [18.41]. Copyright 1994, with permission from Elsevier Science)

which is independent of the local contact area A_i. The total friction force is thus

$$F = \sum F_i = \sum \varepsilon L_i (z_i/x_i) = \varepsilon \langle z_i/x_i \rangle \sum L_i = \mu L \,, \qquad (18.43)$$

where it is assumed that on average the local values of L_i and P_i are independent of the local "slope" z_i/x_i. Therefore, the friction coefficient μ is a function only of the average surface topography and the sliding velocity, but is independent of the local (real) or macroscopic (apparent) contact areas.

While this analysis explains non-adhering surfaces, there is still an additional explicit contact area contribution for the case of adhering surfaces, as in (18.37). The distinction between the two cases arises because the initial assumption of the contact value theorem, (18.40), is incomplete for adhering systems. A more appropriate starting equation would reflect the full intermolecular interaction potential, including the attractive interactions, in addition to the purely repulsive contributions of (18.40), much as the van der Waals equation of state modifies the ideal gas law.

18.7.3 Examples of Experimentally Observed Friction of Dry Surfaces

Numerous model systems have been studied with a surface forces apparatus (SFA) modified for friction experiments (see Sect. 18.2.3). The apparatus allows for control of load (normal force) and sliding speed, and simultaneous measurement of surface separation, surface shape, true (molecular) area of contact between smooth surfaces, and friction forces. A variety of both unlubricated and solid and liquid lubricated surfaces have been studied both as smooth single-asperity contacts and after they have been roughened by shear-induced damage.

Figure 18.16 shows the contact area, A, and friction force, F, both plotted against the applied load, L, in an experiment in which two molecularly smooth surfaces of mica in adhesive contact were slid past each other in an atmosphere of dry nitrogen gas. This is an example of the low load, adhesion controlled limit, which is excellently described by (18.28). In a number of different experiments, S_c was measured to be $2.5 \times 10^7\,\mathrm{N\,m^{-2}}$ and to be independent of the sliding velocity [18.45, 247]. Note that there is a friction force even at negative loads, where the surfaces are still sliding in adhesive contact.

Figure 18.17 shows the correlation between adhesion hysteresis and friction for two surfaces consisting of silica films deposited on mica substrates [18.41].

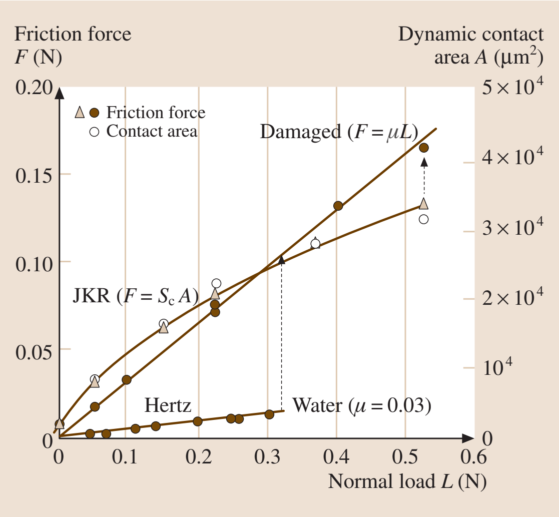

Fig. 18.18 Sliding of mica surfaces, each coated with a 2.5-nm-thick monolayer of calcium stearate surfactant, in the absence of damage (obeying JKR type boundary friction) and in the presence of damage (obeying Amontons' type normal friction). Note that both for this system and for the bare mica in Figs. 18.16 and 18.20, the friction force obeys Amontons' law with a friction coefficient of $\mu \approx 0.3$ after damage occurs. At much higher applied loads, the undamaged surfaces also follow Amontons' type sliding, but for a different reason: The dependence on adhesion becomes smaller. *Lower line:* Interfacial sliding with a monolayer of water between the mica surfaces, shown for comparison. (After [18.247]. Copyright 1990, with permission from Elsevier Science)

The friction between undamaged hydrophobic silica surfaces showed stick–slip both at dry conditions and at 100% relative humidity. Similar to the mica surfaces in Figs. 18.16, 18.18, and 18.20, the friction of damaged silica surfaces obeyed Amontons' law with a friction coefficient of 0.25 both at dry conditions and at 55% relative humidity.

The high friction force of unlubricated sliding can often be reduced by treating the solid surface with a boundary layer of some other solid material that exhibits lower friction such as a surfactant monolayer, or by ensuring that during sliding, a thin liquid film remains between the surfaces (as will be discussed in Sect. 18.8). The effectiveness of a solid boundary lubricant layer on reducing the forces of friction is illustrated in Fig. 18.18. Comparing this with the friction of the unlubricated/untreated surfaces (Fig. 18.16) shows that the critical shear stress has been reduced by a factor of about ten: from 2.5×10^7 to $3.5\times10^6\,\mathrm{N\,m^{-2}}$. At much higher applied loads or pressures, the friction force is proportional to the load, rather than the area of contact [18.237], as expected from (18.27).

Yamada and *Israelachvili* [18.248] studied the adhesion and friction of fluorocarbon surfaces (surfactant-coated boundary lubricant layers), which were compared to those of hydrocarbon surfaces. They concluded that well-ordered fluorocarbon surfaces have a high friction, in spite of their lower adhesion energy (in agreement with previous findings). The low friction coefficient of PTFE (Teflon) must, therefore, be due to some effect other than a low adhesion. For example, the softness of PTFE, which allows material to flow at the interface, which thus behaves like a fluid lubricant. On a related issue, *Luengo* et al. [18.249] found that C_{60} surfaces also exhibited low adhesion but high friction. In both cases the high friction appears to arise from the bulky surface groups – fluorocarbon compared to hydrocarbon groups in the former, large fullerene spheres in the latter. Apparently, the fact that C_{60} molecules rotate *in their lattice* does not make them a good lubricant: The molecules of the opposing surface must still climb over them in order to slide, and this requires energy that is independent of whether the surface molecules are fixed or freely rotating. Larger particles such as ~ 25 nm-sized nanoparticles (also known as "inorganic fullerenes") do appear to produce low friction by behaving like molecular ball bearings, but the potential of this promising new class of solid lubricant has still to be explored [18.250].

Figure 18.19 illustrates the relationship between adhesion hysteresis and friction for surfactant-coated surfaces under different conditions. This effect, however, is much more general and has been shown to hold for other surfaces as well [18.41, 232, 251].

Direct comparisons between absolute adhesion energies and friction forces show little correlation. In some cases, higher adhesion energies for the same system under different conditions correspond with lower friction forces. For example, for hydrophilic silica surfaces (Fig. 18.17) it was found that with increasing relative humidity the adhesion energy *increases*, but the adhesion energy hysteresis measured in a loading–unloading cycle *decreases*, as does the friction force [18.41]. For hydrophobic silica surfaces under dry conditions, the friction at load $L = 5.5\,\mathrm{mN}$ was $F = 75\,\mathrm{mN}$. For the same sample, the adhesion energy hysteresis was $\Delta\gamma = 10\,\mathrm{mJ\,m^{-2}}$, with a contact area of $A \approx 10^{-8}\,\mathrm{m^2}$ at the same load. Assuming a value for the characteristic distance σ on the order of one lattice spacing, $\sigma \approx 1$ nm, and inserting these values into (18.32), the friction force is predicted to be $F \approx 100\,\text{mN}$ for the kinetic friction force, which is close to the measured value of 75 mN. Alternatively, we may conclude that the dissipation factor is $\varepsilon = 0.75$, i. e., that almost all the energy is dissipated as heat at each molecular collision.

A liquid lubricant film (Sect. 18.8.3) is usually much more effective at lowering the friction of two surfaces than a solid boundary lubricant layer. However, to successfully use a liquid lubricant, it must "wet" the surfaces, that is, it should have a high affinity for the surfaces, so that not all the liquid molecules become squeezed out when the surfaces come close together, even under a large compressive load. Another important requirement is that the liquid film remains a liquid under tribological conditions, i. e., that it does not epitaxially solidify between the surfaces.

Effective lubrication usually requires that the lubricant be injected between the surfaces, but in some cases the liquid can be made to condense from the vapor. This is illustrated in Fig. 18.20 for two untreated mica surfaces sliding with a thin layer of water between them. A monomolecular film of water (of thickness 0.25 nm per surface) has reduced S_c from its value for dry surfaces (Fig. 18.16) by a factor of more than 30, which

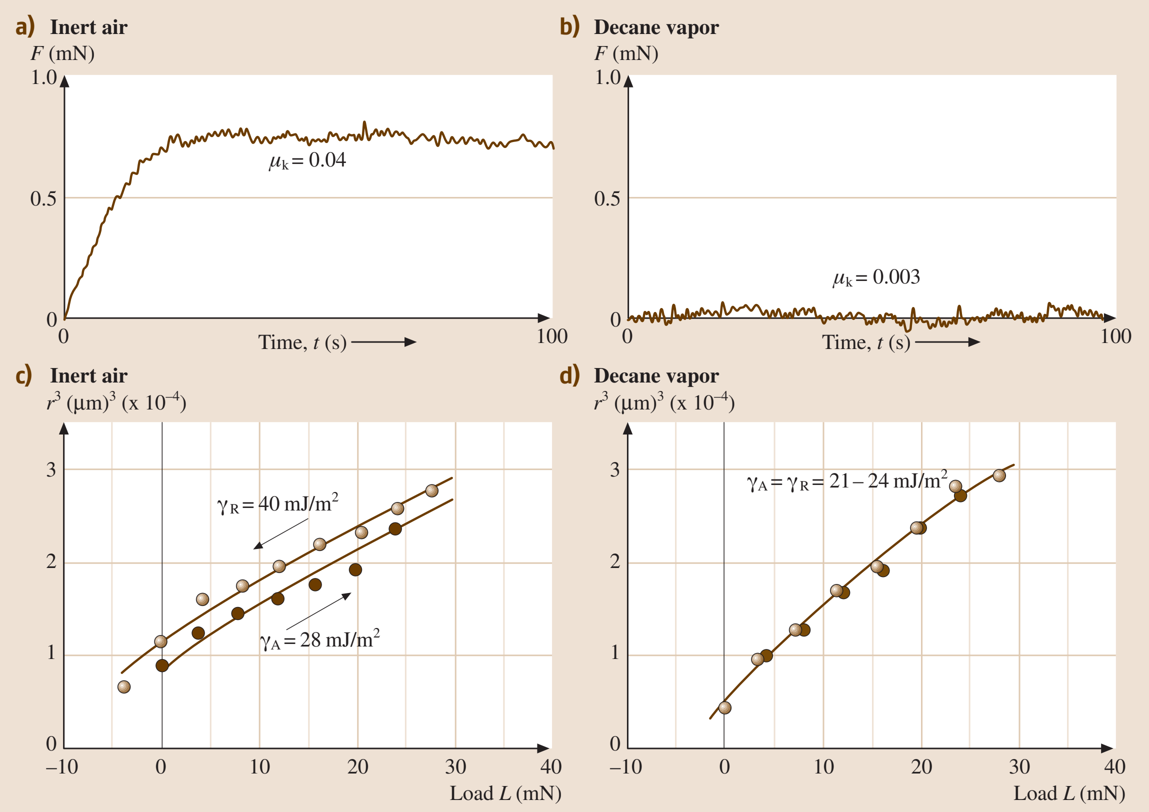

Fig. 18.19a–d *Top:* Friction traces for two fluid-like calcium alkylbenzene sulfonate monolayer-coated surfaces at 25 °C showing that the friction force is much higher between dry monolayers **(a)** than between monolayers whose fluidity has been enhanced by hydrocarbon penetration from vapor **(b)**. *Bottom:* Contact radius vs. load (r^3 vs. L) data measured for the same two surfaces as above and fitted to the JKR equation (18.22), shown by the *solid curves*. For dry monolayers **(c)** the adhesion energy on unloading ($\gamma_R = 40\,\text{mJ}\,\text{m}^{-2}$) is greater than that on loading ($\gamma_R = 28\,\text{mJ}\,\text{m}^{-2}$), indicative of an adhesion energy hysteresis of $\Delta\gamma = \gamma_R - \gamma_A = 12\,\text{mJ}\,\text{m}^{-2}$. For monolayers exposed to saturated decane vapor **(d)** their adhesion hysteresis is zero ($\gamma_A = \gamma_R$), and both the loading and unloading data are well fitted by the thermodynamic value of the surface energy of fluid hydrocarbon chains, $\gamma = 24\,\text{mJ}\,\text{m}^{-2}$. (After [18.207] with permission. Copyright 1993 American Chemical Society)

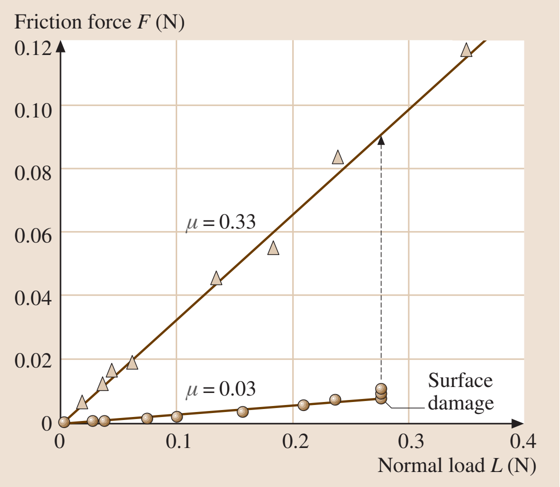

Fig. 18.20 Two mica surfaces sliding past each other while immersed in a 0.01 M KCl salt solution (nonadhesive conditions). The water film is molecularly thin, 0.25 to 0.5 nm thick, and the interfacial friction force is very low: $S_c \approx 5 \times 10^5\,\mathrm{N\,m^{-2}}$, $\mu \approx 0.03$ (before damage occurs). After the surfaces have become damaged, the friction coefficient is ca. 0.3. (After [18.247]. Copyright 1990, with permission from Elsevier Science)

may be compared with the factor of ten attained with the boundary lubricant layer (of thickness 2.5 nm per surface) in Fig. 18.18. Water appears to have unusual lubricating properties and usually gives wearless friction with no stick–slip.

The effectiveness of a water film only 0.25 nm thick to lower the friction force by more than an order of magnitude is attributed to the "hydrophilicity" of the mica surface (mica is "wetted" by water) and to the existence of a strongly repulsive short-range hydration force between such surfaces in aqueous solutions, which effectively removes the adhesion-controlled contribution to the friction force [18.245]. It is also interesting that a 0.25-nm-thick water film between two mica surfaces is sufficient to bring the coefficient of friction down to 0.01–0.02, a value that corresponds to the unusually low friction of ice. Clearly, a single monolayer of water can be a very good lubricant – much better than most other monomolecular liquid films – for reasons that will be discussed in Sect. 18.9.

Dry polymer layers (Fig. 18.21) typically show a high initial static friction ("stiction") as sliding commences from rest in adhesive contact. The development of the friction force after a change in sliding direction, a gradual transition from stick–slip to smooth sliding, is shown in Fig. 18.21. A correlation between adhesion hysteresis and friction similar to the one observed for silica surfaces in Fig. 18.17 can be seen for dry polymer layers below their glass transition temperature. As shown in Fig. 18.12b,c, the adhesion hysteresis for polystyrene surfaces can be increased by irradiation to induce scission of chains, and it has been found that the steady state friction force (kinetic friction) shows a similar increase with irradiation time [18.208].

Figure 18.22 shows an example of a computer simulation of the sliding of two unlubricated silicon surfaces (modeled as a tip sliding over a planar surface) [18.91]. The sliding proceeds through a series of stick–slip events, and information on the friction force and the local order of the initially crystalline surfaces can be obtained. Similar studies for cold welding systems [18.91] have demonstrated the occurrence of shear or friction damage within the sliding surface (tip) as the lowest layer of it adheres to the bottom surface.

18.7.4 Transition from Interfacial to Normal Friction with Wear

Frictional damage can have many causes such as adhesive tearing at high loads or overheating at high sliding speeds. Once damage occurs, there is a transition from "interfacial" to "normal" or load-controlled friction as the surfaces become forced apart by the torn-out as-

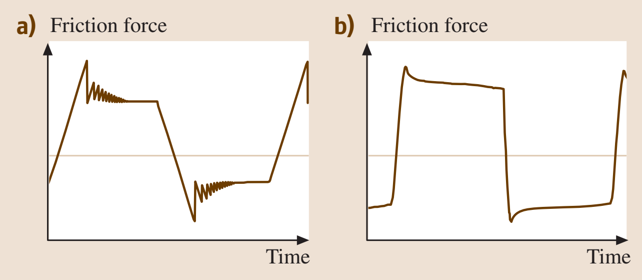

Fig. 18.21a,b Typical friction traces showing how the friction force varies with the sliding time for two symmetric, glassy polymer films at dry conditions. Qualitative features that are common to both polystyrene and polyvinyl benzyl chloride: **(a)** Decaying stick–slip motion is observed until smooth sliding is attained if the motion continues for a sufficiently long distance. **(b)** Smooth sliding observed at sufficiently high speeds. Similar observations have been made by *Berthoud* et al. [18.252] in measurements on polymethyl methacrylate. (After [18.208] with permission. Copyright 2002 American Association for the Advancement of Science)

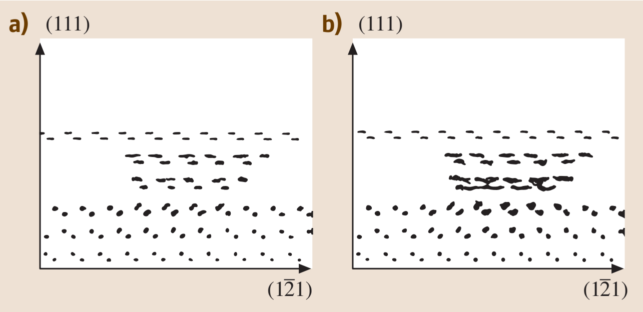

Fig. 18.22a,b Computer simulation of the sliding of two contacting Si surfaces (a tip and a flat surface). Shown are particle trajectories in a constant-force simulation, $F_{z,\text{external}} = -2.15 \times 10^{-8}$ N, viewed along the (10$\bar{1}$) direction just before **(a)** and after **(b)** a stick–slip event for a large, initally ordered, dynamic tip. (After [18.91] with permission of Kluwer Academic Publishers)

perities (wear particles). For low loads, the friction changes from obeying $F = S_c A$ to obeying Amontons' law, $F = \mu L$, as shown in Figs. 18.16 and 18.18, and sliding now proceeds smoothly with the surfaces separated by a 10–100 nm forest of wear debris (in this case, mica flakes). The wear particles keep the surfaces apart over an area that is much greater than their size, so that even one submicroscopic particle or asperity can cause a significant reduction in the area of contact and, therefore, in the friction [18.247]. For this type of frictional sliding, one can no longer talk of the molecular contact area of the two surfaces, although the macroscopic or "apparent" area is still a useful parameter.

One remarkable feature of the transition from interfacial to normal friction of brittle surfaces is that while the strength of interfacial friction, as reflected in the values of S_c, is very dependent on the type of surface and on the liquid film between the surfaces, this is not the case once the transition to normal friction has occurred. At the onset of damage, the material properties of the underlying substrates control the friction. In Figs. 18.16, 18.18, and 18.20 the friction for the damaged surfaces is that of any damaged mica–mica system, $\mu \approx 0.3$, *independent of the initial surface coatings or liquid films between the surfaces*. A similar friction coefficient was found for damaged silica surfaces [18.41].

In order to practically modify the frictional behavior of such brittle materials, it is important to use coatings that will both alter the interfacial tribological character and remain intact and protect the surfaces from damage during sliding. *Berman* et al. [18.253] found that the friction of a strongly bound octadecyl phosphonic acid monolayer on alumina surfaces was higher than for untreated, undamaged α-alumina surfaces, but the bare surfaces easily became damaged upon sliding, resulting in an ultimately higher friction system with greater wear rates than the more robust monolayer-coated surfaces.

Clearly, the mechanism and factors that determine *normal* friction are quite different from those that govern *interfacial* friction (Sects. 18.7.1–18.7.2). This effect is not general and may only apply to brittle materials. For example, the friction of ductile surfaces is totally different and involves the continuous plastic deformation of contacting surface asperities during sliding, rather than the rolling of two surfaces on hard wear particles [18.209]. Furthermore, in the case of ductile surfaces, water and other surface-active components do have an effect on the friction coefficients under "normal" sliding conditions.

18.8 Liquid Lubricated Surfaces

18.8.1 Viscous Forces and Friction of Thick Films: Continuum Regime

Experimentally, it is usually difficult to unambiguously establish which type of sliding mode is occurring, but an empirical criterion, based on the Stribeck curve (Fig. 18.13), is often used as an indicator. This curve shows how the friction force or the coefficient of friction is expected to vary with sliding speed, depending on which type of friction regime is operating. For thick liquid lubricant films whose behavior can be described by bulk (continuum) properties, the friction forces are essentially the hydrodynamic or viscous drag forces. For example, for two plane parallel surfaces of area A separated by a distance D and moving laterally relative to each other with velocity v, if the intervening liquid is *Newtonian*, i. e., if its viscosity η is independent of the shear rate, the frictional force experienced by the surfaces is given by

$$F = \frac{\eta A v}{D}\,, \tag{18.44}$$

where the shear rate $\dot{\gamma}$ is defined by

$$\dot{\gamma} = \frac{v}{D} . \qquad (18.45)$$

At higher shear rates, two additional effects often come into play. First, certain properties of liquids may change at high $\dot{\gamma}$ values. In particular, the effective viscosity may become non-Newtonian, one form given by

$$\eta = \dot{\gamma}^n , \qquad (18.46)$$

where $n = 0$ (i. e., $\eta_{\text{eff}} = \text{constant}$) for Newtonian fluids, $n > 0$ for shear thickening (dilatent) fluids, and $n < 0$ for shear thinning (pseudoplastic) fluids (the latter become less viscous, i. e., flow more easily, with increasing shear rate). An additional effect on η can arise from the higher local stresses (pressures) experienced by the liquid film as $\dot{\gamma}$ increases. Since the viscosity is generally also sensitive to the pressure (usually increasing with P), this effect also acts to increase η_{eff} and thus the friction force.

A second effect that occurs at high shear rates is surface deformation, arising from the large hydrodynamic forces acting on the sliding surfaces. For example, Fig. 18.23 shows how two surfaces deform elastically when the sliding speed increases to a high value. These deformations alter the hydrodynamic friction forces, and this type of friction is often referred to as *elastohydrodynamic lubrication* (EHD or EHL), as mentioned in Table 18.4.

How thin can a liquid film be before its dynamic, e.g., viscous flow, behavior ceases to be described by bulk properties and continuum models? Concerning the static properties, we have already seen in Sect. 18.4.3 that films composed of simple liquids display continuum behavior down to thicknesses of 4–10 molecular diameters. Similar effects have been found to apply to the dynamic properties, such as the viscosity, of simple liquids in thin films. Concerning viscosity measurements, a number of dynamic techniques were recently developed [18.11–13, 43, 254, 255] for directly measuring the viscosity as a function of film thickness and shear rate across very thin liquid films between two surfaces. By comparing the results with theoretical predictions of fluid flow in thin films, one can determine the effective positions of the shear planes and the onset of non-Newtonian behavior in very thin films.

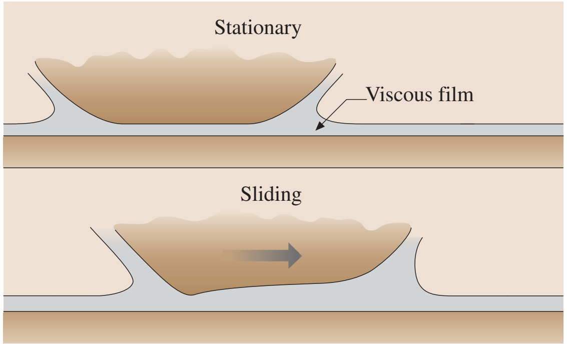

Fig. 18.23 *Top:* Stationary surfaces (one more deformable and one rigid) separated by a thick liquid film. *Bottom:* Elastohydrodynamic deformation of the upper surface during sliding. (After [18.1] with permission)

The results show that for simple liquids, including linear chain molecules such as alkanes, the viscosity in thin films is the same, within 10%, as the bulk even for films as thin as 10 molecular diameters (or segment widths) [18.11–13,254,255]. This implies that the shear plane is effectively located within one molecular diameter of the solid–liquid interface, and these conclusions were found to remain valid even at the highest shear rates studied (of $\sim 2 \times 10^5\ \text{s}^{-1}$). With water between two mica or silica surfaces [18.22, 254–256] this has been found to be the case (to within ±10%) down to surface separations as small as 2 nm, implying that the shear planes must also be within a few tenths of a nanometer of the solid–liquid interfaces. These results appear to be independent of the existence of electrostatic "double-layer" or "hydration" forces. For the case of the simple liquid toluene confined between surfaces with adsorbed layers of C_{60} molecules, this type of viscosity measurement has shown that the traditional no-slip assumption for flow at a solid interface does not always hold [18.257]. The C_{60} layer at the mica–toluene interface results in a "full slip" boundary, which dramatically lowers the viscous drag or effective viscosity for regular Couette or Poiseuille flow.

With polymeric liquids (polymer melts) such as polydimethylsiloxanes (PDMS) and polybutadienes (PBD), or with polystyrene (PS) adsorbed onto surfaces from solution, the far-field viscosity is again equal to the bulk value, but with the non-slip plane (hydrodynamic layer thickness) being located at $D = 1{-}2R_g$ away from each surface [18.11, 47], or at $D = L$ or less for polymer brush layers of thickness L per surface [18.13, 258]. In contrast, the same technique was used to show that for non-adsorbing polymers in solution, there is actually a depletion layer of nearly pure solvent that exists at the surfaces that affects the confined solution flow properties [18.256]. These effects are observed from near contact to surface separations in excess of 200 nm.

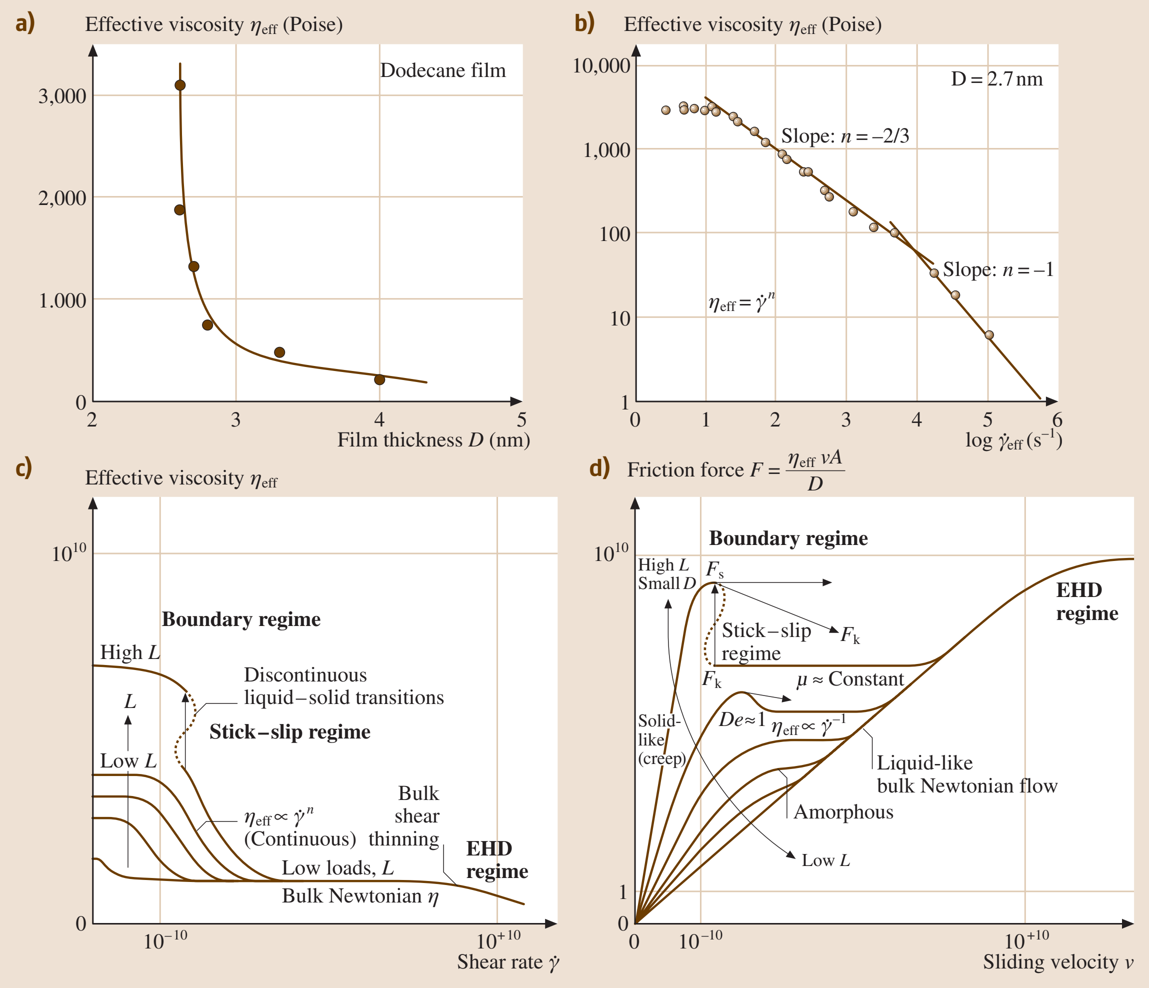

Further experiments with surfaces closer than a few molecular diameters ($D < 2$–4 nm for simple liquids, or $D < 2$–$4R_g$ for polymer fluids) indicate that large deviations occur for thinner films, described below. One important conclusion from these studies is, therefore, that the dynamic properties of simple liquids, including water, near an *isolated* surface are similar to those of the bulk liquid *already within the first layer of molecules adjacent to the surface*, only changing when another surface approaches the first. In other words, the viscosity and position of the shear plane near a surface are not simply a property of that surface, but of how far that surface is from another surface. The reason for this is because when two surfaces are close together, the constraining effects on the liquid molecules between them are much more severe than when there is only one surface. Another obvious consequence of the above is that one should not make measurements on a single, isolated solid–liquid interface and then draw conclusions about the state of the liquid or its interactions in a thin film *between* two surfaces.

18.8.2 Friction of Intermediate Thickness Films

For liquid films in the thickness range between 4 and 10 molecular diameters, the properties can be significantly different from those of bulk films. Still, the fluids remain recognizable as fluids; in other words, they do not undergo a phase transition into a solid

◀ **Fig. 18.24a–d** Typical rheological behavior of liquid films in the mixed lubrication regime. **(a)** Increase in effective viscosity of dodecane film between two mica surfaces with decreasing film thickness. At distances larger than 4–5 nm, the effective viscosity η_{eff} approaches the bulk value η_{bulk} and does not depend on the shear rate $\dot{\gamma}$. (After [18.219]. Copyright 1991 American Association for the Advancement of Science.) **(b)** Non-Newtonian variation of η_{eff} with shear rate of a 2.7-nm-thick dodecane film at a net normal pressure of 0.12 MPa and at 28 °C. The effective viscosity decays as a power law, as in (18.46). In this example, $n = 0$ at the lowest $\dot{\gamma}$ and changes to $n = -2/3$ and -1 at higher $\dot{\gamma}$. For films of bulk thickness, dodecane is a low viscosity Newtonian fluid ($n = 0$). **(c)** Proposed general friction map of effective viscosity η_{eff} (arbitrary units) as a function of effective shear rate $\dot{\gamma}$ (arbitrary units) on logarithmic scales. Three main classes of behavior emerge: (i) Thick films: elastohydrodynamic sliding. At $L = 0$, approximating bulk conditions, η_{eff} is independent of shear rate except when shear thinning might occur at sufficiently large $\dot{\gamma}$. (ii) Boundary layer films, intermediate regime. A Newtonian regime is again observed [$\eta_{\mathrm{eff}} = $ constant, $n = 0$ in (18.46)] at low loads and low shear rates, but η_{eff} is much higher than the bulk value. As the shear rate $\dot{\gamma}$ increases beyond $\dot{\gamma}_{\mathrm{min}}$, the effective viscosity starts to drop with a power-law dependence on the shear rate [see panel **(b)**], with n in the range $-1/2$ to -1 most commonly observed. As the shear rate $\dot{\gamma}$ increases still more, beyond $\dot{\gamma}_{\mathrm{max}}$, a second Newtonian plateau is encountered. (iii) Boundary layer films, high load. The η_{eff} continues to grow with load and to be Newtonian provided that the shear rate is sufficiently low. Transition to sliding at high velocity is discontinuous ($n < -1$) and usually of the stick–slip variety. **(d)** Proposed friction map of friction force as a function of sliding velocity in various tribological regimes. With increasing load, Newtonian flow in the elastohydrodynamic (EHD) regimes crosses into the boundary regime of lubrication. Note that even EHD lubrication changes, at the highest velocities, to limiting shear stress response. At the highest loads (L) and smallest film thickness (D), the friction force goes through a maximum (the static friction, F_{s}), followed by a regime where the friction coefficient (μ) is roughly constant with increasing velocity (meaning that the kinetic friction, F_{k}, is roughly constant). Non-Newtonian shear thinning is observed at somewhat smaller load and larger film thickness; the friction force passes through a maximum at the point where $De = 1$. De, the Deborah number, is the point at which the applied shear rate exceeds the natural relaxation time of the boundary layer film. The velocity axis from 10^{-10} to 10^{10} (arbitrary units) indicates a large span. (Panels **(b)–(d)** after [18.259]. Copyright 1996, with permission from Elsevier Science)

or liquid-crystalline phase. This regime has recently been studied by *Granick* et al. [18.44, 219, 220, 222, 223], who used a friction attachment [18.43, 44] to the SFA where a sinusoidal input signal to a piezoelectric device makes the two surfaces slide back and forth laterally past each other at small amplitudes. This method provides information on the real and imaginary parts (elastic and dissipative components, respectively) of the shear modulus of thin films at different shear rates and film thickness. *Granick* [18.219] and *Hu* et al. [18.223] found that films of simple liquids become non-Newtonian in the 2.5–5 nm regime (about 10 molecular diameters, see Fig. 18.24). Polymer melts become non-Newtonian at much larger film thicknesses, depending on their molecular weight [18.47].

Klein and *Kumacheva* [18.46, 226, 260] studied the interaction forces and friction of small quasi-spherical liquid molecules such as cyclohexane between molecularly smooth mica surfaces. They concluded that surface epitaxial effects can cause the liquid film to already solidify at 6 molecular diameters, resulting in a sudden (discontinuous) jump to high friction at low shear rates. Such dynamic first-order transitions, however, may depend on the shear rate.

A generalized friction map (Fig. 18.24c,d) has been proposed by *Luengo* et al. [18.259] that illustrates the changes in η_{eff} from bulk Newtonian behavior ($n = 0$, $\eta_{\mathrm{eff}} = \eta_{\mathrm{bulk}}$) through the transition regime where n reaches a minimum of -1 with decreasing shear rate to the solid-like creep regime at very low $\dot{\gamma}$ where n returns to 0. A number of results from experimental, theoretical, and computer simulation work have shown values of n from $-1/2$ to -1 for this transition regime for a variety of systems and assumptions [18.219, 220, 239, 261–267].

The intermediate regime appears to extend over a narrow range of film thickness, from about 4 to 10 molecular diameters or polymer radii of gyration. Thinner films begin to adopt "boundary" or "interfacial" friction properties (described below, see also Table 18.5). Note that the intermediate regime is actually a very narrow one when defined in terms of film thickness, for example, varying from about $D = 2$ to 4 nm for hexadecane films [18.219].

A fluid's effective viscosity η_{eff} in the intermediate regime is usually higher than in the bulk, but η_{eff} usu-

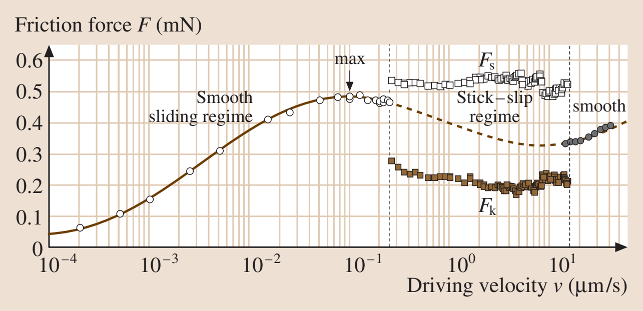

Fig. 18.25 Transition from smooth sliding to "inverted" stick–slip and to a second smooth sliding regime with increasing driving velocity during shear of two adsorbed surfactant monolayers in aqueous solution at a load of $L = 4.5$ mN and $T = 20\,°$C. The smooth sliding (*open circles*) to inverted stick–slip (*squares*) transition occurs at $v_c \sim 0.3$ μm/s. Prior to the transition, the kinetic stress levels off at after a logarithmic dependence on velocity. The quasi-smooth regime persists up to the transition at v_c. At high driving velocities (*filled circles*), a new transition to a smooth sliding regime is observed between 14 and 17 μm/s. (After [18.268] with permission)

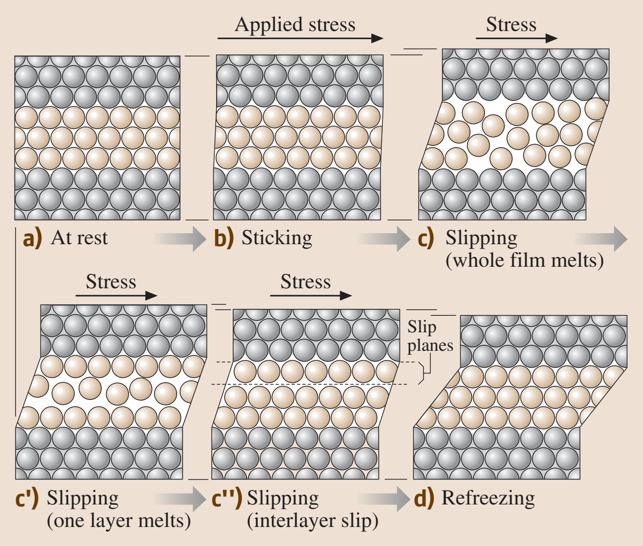
Applied stress
Stress
a) At rest
b) Sticking
c) Slipping (whole film melts)
Stress
Stress
Slip planes
c') Slipping (one layer melts)
c'') Slipping (interlayer slip)
d) Refreezing

Fig. 18.26a–d Idealized schematic illustration of molecular rearrangements occurring in a molecularly thin film of spherical or simple chain molecules between two solid surfaces during shear. Depending on the system, a number of different molecular configurations within the film are possible during slipping and sliding, shown here as stages **(c)**: total disorder as the whole film melts; **(c′)**: partial disorder; and **(c″)**: order persists even during sliding with slip occurring at a single slip plane either within the film or at the walls. A dilation is predicted in the direction normal to the surfaces. (After [18.224] with permission)

ally *decreases* with increasing sliding velocity, v (known as *shear thinning*). When two surfaces slide in the intermediate regime, the motion tends to thicken the film (dilatancy). This sends the system into the bulk EHL regime where, as indicated by (18.44), the friction force now *increases* with velocity. This initial decrease, followed by an increase, in the frictional forces of many lubricant systems is the basis for the empirical Stribeck curve of Fig. 18.13. In the transition from bulk to boundary behavior there is first a quantitative change in the material properties (viscosity and elasticity), which can be continuous, to discontinuous qualitative changes that result in yield stresses and non-liquidlike behavior.

The rest of this section is devoted to friction in the boundary regime. Boundary friction may be thought of as applying to the case where a lubricant film is present, but where this film is of molecular dimensions – a few molecular layers or less.

18.8.3 Boundary Lubrication of Molecularly Thin Films: Nanorheology

When a liquid is confined between two surfaces or within any narrow space whose dimensions are less than 4 to 10 molecular diameters, both the static (equilibrium) and dynamic properties of the liquid, such as its compressibility and viscosity, can no longer be described even qualitatively in terms of the bulk properties. The molecules confined within such molecularly thin films become ordered into layers ("out-of-plane" ordering), and within each layer they can also have lateral order ("in-plane" ordering). Such films may be thought of as behaving more like a liquid crystal or a solid than a liquid.

As described in Sect. 18.4.3, the measured *normal* forces between two solid surfaces across molecularly thin films exhibit exponentially decaying oscillations, varying between attraction and repulsion with a periodicity equal to some molecular dimension of the solvent molecules. Thus most liquid films can sustain a finite normal stress, and the adhesion force between two surfaces across such films is "quantized", depending on the thickness (or number of liquid layers) between the surfaces. The structuring of molecules in thin films and the oscillatory forces it gives rise to are now reasonably well understood, both experimentally and theoretically, at least for simple liquids.

Work has also recently been done on the dynamic, e.g., viscous or shear, forces associated with molecularly thin films. Both experiments [18.38, 46, 203, 221, 226,227,269,270] and theory [18.200,201,261,271] indicate that even when two surfaces are in steady state sliding they still prefer to remain in one of their stable potential energy minima, i.e., a sheared film of liquid can retain its basic layered structure. Thus even during motion the film does not become totally liquid-like. Indeed, if there is some "in-plane" ordering within a film, it will exhibit a yield-point before it begins to flow. Such films can therefore sustain a finite shear stress, in addition to a finite normal stress. The value of the yield stress depends on the number of layers comprising the film and represents another "quantized" property of molecularly thin films [18.200].

The dynamic properties of a liquid film undergoing shear are very complex. Depending on whether the film is more liquid-like or solid-like, the motion will be smooth or of the stick–slip type. During sliding, transitions can occur between n layers and $(n-1)$ or $(n+1)$ layers (see Fig. 18.27). The details of the motion depend critically on the externally applied load, the temperature, the sliding velocity, the twist angle between the two surface lattices, and the sliding direction relative to the lattices.

Smooth and Stick–Slip Sliding

Recent advances in friction-measuring techniques have enabled the interfacial friction of molecularly thin films to be measured with great accuracy. Some of these advances have involved the surface forces apparatus technique [18.38, 45, 47, 221, 247, 269, 270], while others have involved the atomic force microscope [18.10, 53, 231, 272]. In addition, computer simulations [18.90,200,201,271,273,274] have become sufficiently sophisticated to enable fairly complex tribological systems to be studied. All these advances are necessary if one is to probe such subtle effects as smooth or stick–slip friction, transient and memory effects, and ultralow friction mechanisms at the molecular level.

The theoretical models presented in this section will be concerned with a situation commonly observed experimentally: Stick–slip occurs between a static state with high friction and a low-friction kinetic state, and a transition from this sliding regime to smooth sliding can be induced by an increase in velocity. Experimental data on various systems showing this behavior are shown in Figs. 18.27, 18.30b, and 18.31a. Recent studies on adhesive systems have revealed the possibility of other dynamic responses such as "inverted" stick–slip between two kinetic states of higher and lower friction and with a transition from smooth sliding to stick–slip with increasing velocity, as shown in Fig. 18.25 [18.268].

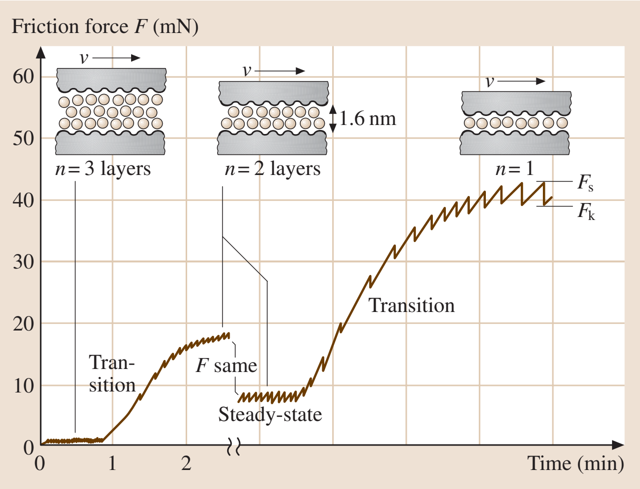

Fig. 18.27 Measured change in friction during interlayer transitions of the silicone liquid octamethylcyclotetrasiloxane (OMCTS), an inert liquid whose quasispherical molecules have a diameter of 0.8 nm. In this system, the shear stress $S_c = F/A$ was found to be constant as long as the number of layers, n, remained constant. Qualitatively similar results have been obtained with other quasi-spherical molecules such as cyclohexane [18.269]. The shear stresses are only weakly dependent on the sliding velocity v. However, for sliding velocities above some critical value, v_c, the stick–slip disappears and sliding proceeds smoothly at the kinetic value. (After [18.221] with permission)

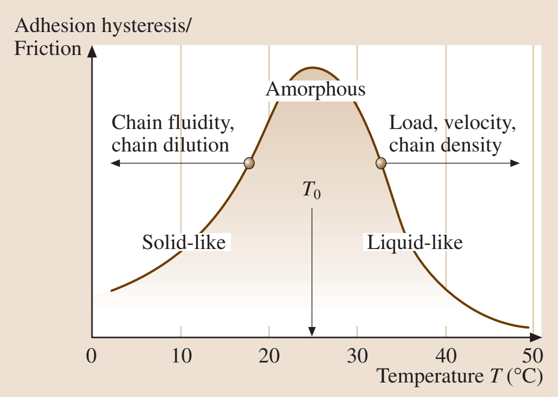

Fig. 18.28 Schematic friction phase diagram representing the trends observed in the boundary friction of a variety of different surfactant monolayers. The characteristic bell-shaped curve also correlates with the adhesion energy hysteresis of the monolayers. The *arrows* indicate the direction in which the whole curve is dragged when the load, velocity, etc., is increased. (After [18.232] with permission)

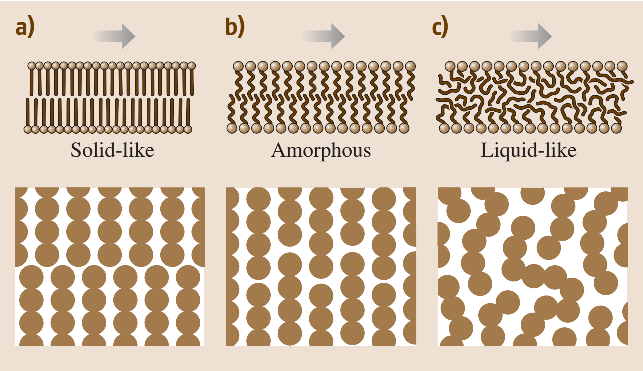

Fig. 18.29a–c Different dynamic phase states of boundary monolayers during adhesive contact and/or frictional sliding. Solid-like **(a)** and liquid-like monolayers **(c)** exhibit low adhesion hysteresis and friction. Increasing the temperature generally shifts a system from the left to the right. Changing the load, sliding velocity, or other experimental conditions can also change the dynamic phase state of surface layers, as shown in Fig. 18.28. (After [18.232] with permission)

With the added insights provided by computer simulations, a number of distinct molecular processes have been identified during smooth and stick–slip sliding in model systems for the more familiar static-to-kinetic stick–slip and transition from stick–slip to smooth sliding. These are shown schematically in Fig. 18.26 for the case of spherical liquid molecules between two solid crystalline surfaces. The following regimes may be identified:

Surfaces at rest (Fig. 18.26a): Even with no externally applied load, solvent–surface epitaxial interactions can cause the liquid molecules in the film to attain a solid-like state. Thus at rest the surfaces are stuck to each other through the film.

Sticking regime (frozen, solid-like film) (Fig. 18.26b): A progressively increasing lateral shear stress is applied. The solid-like film responds elastically with a small lateral displacement and a small increase or "dilatancy" in film thickness (less than a lattice spacing or molecular dimension, σ). In this regime the film retains its frozen, solid-like state: All the strains are elastic and reversible, and the surfaces remain effectively stuck to each other. However, slow creep may occur over long time periods.

Slipping and sliding regimes (molten, liquid-like film) (Fig. 18.26c,c′,c″): When the applied shear stress or force has reached a certain critical value, the *static* friction force, F_s, the film suddenly melts (known as "shear melting") or rearranges to allow for wall-slip or slip within the film to occur at which point the two surfaces begin to slip rapidly past each other. If the applied stress is kept at a high value, the upper surface will continue to slide indefinitely.

Refreezing regime (resolidification of film) (Fig. 18.26d): In many practical cases, the rapid slip of the upper surface relieves some of the applied force, which eventually falls below another critical value, the *kinetic* friction force F_k, at which point the film resolidifies and the whole stick–slip cycle is repeated. On the other hand, if the slip rate is smaller than the rate at which the external stress is applied, the surfaces will continue to slide smoothly in the kinetic state and there will be no more stick–slip. The critical velocity at which stick–slip disappears is discussed in more detail in Sect. 18.8.3.

Experiments with linear chain (alkane) molecules show that the film thickness remains quantized during sliding, so that the structure of such films is probably more like that of a nematic liquid crystal where the liquid molecules have become shear aligned in some direction, enabling shear motion to occur while retaining some order within the film. Experiments on the friction of two molecularly smooth mica surfaces separated by three molecular layers of the liquid OMCTS (Fig. 18.27) show how the friction increases to higher values in a quantized way when the number of layers falls from $n = 3$ to $n = 2$ and then to $n = 1$.

Computer simulations for simple spherical molecules [18.201] further indicate that during the slip the film thickness is roughly 15% higher than at rest (i. e., the film density falls), and that the order parameter within the film drops from 0.85 to about 0.25. Such dilatancy has been investigated both experimentally [18.235] and in further computer simulations [18.234]. The changes in thickness and in order parameter are consistent with a disorganized state for the whole film during the slip [18.271], as illustrated schematically in Fig. 18.26c. At this stage, we can only speculate on other possible configurations of molecules in the slipping and sliding regimes. This probably depends on the shapes of the molecules (e.g., whether spherical or linear or branched), on the atomic structure of the surfaces, on the sliding velocity, etc. [18.275]. Figure 18.26c,c′,c″ shows three possible sliding modes wherein the shearing film either totally melts, or where the molecules retain their layered structure and where slip occurs between two or more layers. Other sliding modes, for example, involving the movement of dislocations or disclinations are also possible, and it is unlikely that one single mechanism applies in all cases.

Both friction and adhesion hysteresis vary nonlinearly with temperature, often peaking at some

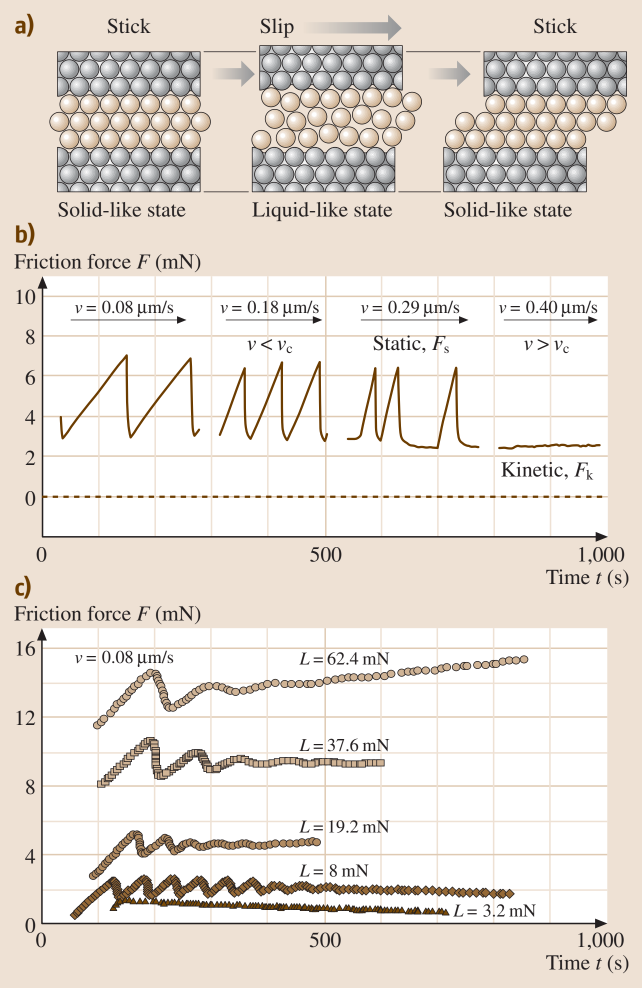

Fig. 18.30 **(a)** "Phase transitions" model of stick–slip where a thin liquid film alternately freezes and melts as it shears, shown here for 22 spherical molecules confined between two solid crystalline surfaces. In contrast to the velocity-dependent friction model, the intrinsic friction force is assumed to change abruptly (at the transitions), rather than smooth or continuously. The resulting stick–slip is also different, for example, the peaks are sharper and the stick–slip disappears above some critical velocity v_c. Note that while the slip displacement is here shown to be only two lattice spacings, in most practical situations it is much larger, and that freezing and melting transitions at surfaces or in thin films may not be the same as freezing or melting transitions between the bulk solid and liquid phases. **(b)** Exact reproduction of a chart-recorder trace of the friction force for hexadecane between two untreated mica surfaces at increasing sliding velocity v, plotted as a function of time. In general, with increasing sliding speed, the stick–slip spikes increase in frequency and decrease in magnitude. As the critical sliding velocity v_c is approached, the spikes become erratic, eventually disappearing altogether at $v = v_c$. At higher sliding velocities the sliding continues smoothly in the kinetic state. Such friction traces are fairly typical for simple liquid lubricants and dry boundary lubricant systems (see Fig. 18.31a) and may be referred to as the "conventional" type of static–kinetic friction (in contrast to Fig. 18.25). Experimental conditions: contact area $A = 4 \times 10^{-9}$ m^2, load $L = 10$ mN, film thickness $D = 0.4$–0.8 nm, $v = 0.08$–0.4 µm s^{-1}, $v_c \approx 0.3$ µm s^{-1}, atmosphere: dry N$_2$ gas, $T = 18\,°$C. [**(a)** and **(b)** after [18.276] with permission. Copyright 1993 American Chemical Society.] **(c)** Friction response of a thin squalane (a branched hydrocarbon) film at different loads and a constant sliding velocity $v = 0.08$ µm s^{-1}, slightly above the critical velocity for this system at low loads. Initally, with increasing load, the stick–slip amplitude and the mean friction force decrease with sliding time or sliding distance. However, at high loads or pressures, the mean friction force increases with time, and the stick–slip takes on a more symmetrical, sinusoidal shape. At all loads investigated, the stick–slip component gradually decayed as the friction proceeded towards smooth sliding. (*Gourdon* and *Israelachvili*, unpublished data)

particular temperature, T_0. The temperature dependence of these forces can, therefore, be represented on a friction phase diagram such as the one shown in Fig. 18.28. Experiments have shown that T_0, and the whole bell-shaped curve, are shifted along the temperature-axis (as well as in the vertical direction) in a systematic way when the load, sliding velocity, etc., are varied. These shifts also appear to be highly correlated with one another, for example, an increase in temperature producing effects that are similar to *decreasing* the sliding speed or load.

Such effects are also commonly observed in other energy dissipating phenomena such as polymer viscoelasticity [18.277], and it is likely that a similar physical mechanism is at the heart all of such phenomena. A possible molecular process underlying the energy dissipation of chain molecules during boundary layer sliding is illustrated in Fig. 18.29, which shows the three main dynamic phase states of boundary monolayers.

In contrast to the characteristic relaxation time associated with fluid lubricants, it has been established

a) Solid surfactant

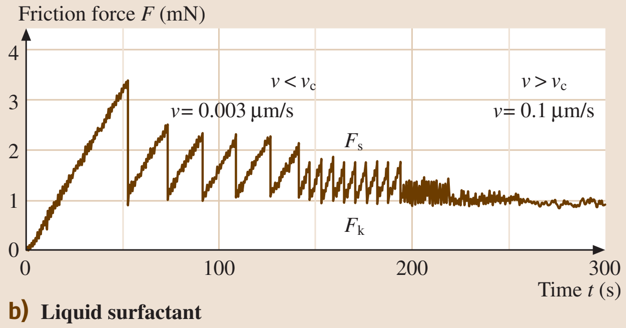

b) Liquid surfactant

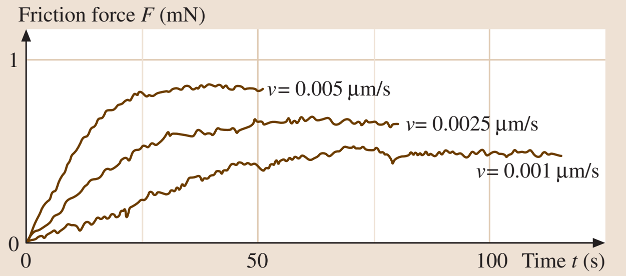

Fig. 18.31 **(a)** Exact reproduction of chart-recorder trace for the friction of closely packed surfactant monolayers (L-α-dimirystoyl-phosphatidyl-ethanolamine, DMPE) on mica (dry boundary friction) showing qualitatively similar behavior to that obtained with a liquid hexadecane film (Fig. 18.30b). In this case, $L = 0$, $v_c \approx 0.1\,\mu\mathrm{m\,s}^{-1}$, atmosphere: dry N_2 gas, $T = 25\,°\mathrm{C}$. **(b)** Sliding typical of liquid-like monolayers, here shown for calcium alkylbenzene sulfonate in dry N_2 gas at $T = 25\,°\mathrm{C}$ and $L = 0$. (After [18.207] with permission. Copyright 1993 American Chemical Society)

that for unlubricated (dry, solid, rough) surfaces, there is a characteristic memory distance that must be exceeded before the system looses all memory of its initial state (original surface topography). The underlying mechanism for a characteristic distance was first used to successfully explain rock mechanics and earthquake faults [18.278] and, more recently, the tribological behavior of unlubricated surfaces of ceramics, paper and elastomeric polymers [18.252, 279]. Recent experiments [18.275, 280, 281] suggest that fluid lubricants composed of complex branched-chained or polymer molecules may also have characteristic distances (in addition to characteristic relaxation times) associated with their tribological behavior – the characteristic distance being the total sliding distance that must be exceeded before the system reaches its steady-state tribological conditions (see Sect. 18.8.3).

Abrupt vs. Continuous Transitions Between Smooth and Stick–Slip Sliding

An understanding of stick–slip is of great practical importance in tribology [18.282], since these spikes are the major cause of damage and wear of moving parts. Stick–slip motion is a very common phenomenon and is also the cause of sound generation (the sound of a violin string, a squeaking door, or the chatter of machinery), sensory perception (taste texture and feel), earthquakes, granular flow, nonuniform fluid flow such as the spurting flow of polymeric liquids, etc. In the previous section, the stick–slip motion arising from freezing–melting transitions in thin interfacial films was described. There are other mechanisms that can give rise to stick–slip friction, which will now be considered. However, before proceeding with this, it is important to clarify exactly what one is measuring during a friction experiment.

Most tribological systems and experiments can be described in terms of an equivalent mechanical circuit with certain characteristics. The friction force F_0, which is generated at the surfaces, is generally measured as F at some other place in the setup. The mechanical coupling between the two may be described in terms of a simple elastic stiffness or compliance, K, or as more complex nonelastic coefficients, depending on the system. The distinction between F and F_0 is important because in almost all practical cases, the applied, measured, or detected force, F, is *not* the same as the "real" or "intrinsic" friction force, F_0, generated at the surfaces. F and F_0 are coupled in a way that depends on the mechanical construction of the system, for example, the axle of a car wheel that connects it to the engine. This coupling can be modeled as an elastic spring of stiffness K and mass m. This is the simplest type of mechanical coupling and is also the same as in SFA and AFM type experiments. More complicated real systems can be reduced to a system of springs and dashpots, as described by *Peachey* et al. [18.283] and *Luengo* et al. [18.47].

We now consider four different models of stick–slip friction, where the mechanical couplings are assumed to be of the simple elastic spring type. The first three mechanisms may be considered traditional or classical mechanisms or models [18.282], the fourth is essentially the same as the freezing–melting phase-transition model described in Sect. 18.8.3.

Rough Surfaces or Surface Topology Model. Rapid slips can occur whenever an asperity on one surface goes over the top of an asperity on the other surface. The extent of the slip will depend on asperity heights and slopes, on the speed of sliding, and on the elastic compliance of the surfaces and the moving stage. As in all cases of stick–slip motion, the driving velocity v may be constant, but the resulting motion at the surfaces v_0 will display large slips. This type of stick–slip has been described by *Rabinowicz* [18.282]. It will not be of much concern here since it is essentially a noise-type fluctuation, resulting from surface imperfections rather than from the intrinsic interaction between two surfaces. Actually, at the atomic level, the regular atomic-scale corrugations of surfaces can lead to periodic stick–slip motion of the type shown here. This is what is sometimes measured by AFM tips [18.10, 53, 231, 272].

Distance-Dependent or Creep Model. Another theory of stick–slip, observed in solid-on-solid sliding, is one that involves a characteristic *distance*, but also a characteristic time, τ_s, this being the characteristic time required for two asperities to increase their adhesion strength after coming into contact. Originally proposed by *Rabinowicz* [18.282, 284], this model suggests that two rough macroscopic surfaces adhere through their microscopic asperities of characteristic length. During shearing, each surface must first creep this distance (the size of the contacting junctions) after which the surfaces continue to slide, but with a lower (kinetic) friction force than the original (static) value. The reason for the decrease in the friction force is that even though, on average, new asperity junctions should form as rapidly as the old ones break, the time-dependent adhesion and friction of the new ones will be lower than the old ones.

The friction force, therefore, remains high during the creep stage of the slip. But once the surfaces have moved the characteristic distance, the friction rapidly drops to the kinetic value. For any system where the kinetic friction is less than the static force (or one that has a negative slope over some part of its F_0 versus v_0 curve) will exhibit regular stick–slip sliding motion for certain values of K, m, and driving velocity, v.

This type of friction has been observed in a variety of dry (unlubricated) systems such as paper-on-paper [18.285, 286] and steel-on-steel [18.284, 287, 288]. This model is also used extensively in geologic systems to analyze rock-on-rock sliding [18.289, 290]. While originally described for adhering macroscopic asperity junctions, the distance-dependent model may also apply to molecularly smooth surfaces. For example, for polymer lubricant films, the characteristic length would now be the chain–chain entanglement length, which could be much larger in a confined geometry than in the bulk.

Velocity-Dependent Friction Model. In contrast to the two friction models mentioned above, which apply mainly to unlubricated, solid-on-solid contacts, the stick–slip of surfaces with thin liquid films between them is better described by other mechanisms. The velocity-dependent friction model is the most studied mechanism of stick–slip and, until recently, was considered to be the only cause of intrinsic stick–slip. If a friction force decreases with increasing sliding velocity, as occurs with boundary films exhibiting shear thinning, the force (F_s) needed to initiate motion will be higher than the force (F_k) needed to maintain motion.

A decreasing intrinsic friction force F_0 with sliding velocity v_0 results in the sliding surface or stage moving in a periodic fashion, where during each cycle rapid acceleration is followed by rapid deceleration. So long as the drive continues to move at a fixed velocity v, the surfaces will continue to move in a periodic fashion punctuated by abrupt stops and starts whose frequency and amplitude depend not only on the function $F_0(v_0)$, but also on the stiffness K and mass m of the moving stage, and on the starting conditions at $t = 0$.

More precisely, the motion of the sliding surface or stage can be determined by solving the following differential equation:

$$\begin{aligned} &m\ddot{x} = (F_0 - F) = F_0 - (x_0 - x)K \\ &\text{or} \quad m\ddot{x} + (x_0 - x)K - F_0 = 0\,, \end{aligned} \tag{18.47}$$

where $F_0 = F_0(x_0, v_0, t)$ is the intrinsic or "real" friction force at the shearing surfaces, F is the force on the spring (the externally applied or measured force), and $(F_0 - F)$ is the force on the stage. To fully solve (18.47), one must also know the initial (starting) conditions at $t = 0$, and the driving or steady-state conditions at finite t. Commonly, the driving condition is: $x = 0$ for $t < 0$ and $x = vt$ for $t > 0$, where $v =$ constant. In other systems, the appropriate driving condition may be $F =$ constant.

Various, mainly phenomenological, forms for $F_0 = F_0(x_0, v_0, t)$ have been proposed to explain various kinds of stick–slip phenomena. These models generally assume a particular functional form for the friction as a function of velocity only, $F_0 = F_0(v_0)$, and they may also contain a number of mechanically coupled elements comprising the stage [18.291, 292]. One version is a two-state model characterized by two friction forces, F_s and F_k, which is a simplified version of the phase transitions model described below. More complicated versions can have a rich F–v spectrum, as proposed by *Persson* [18.293]. Unless the experimental data is very detailed and extensive, these models cannot generally distinguish between different types of mechanisms.

Neither do they address the basic question of the *origin* of the friction force, since this is assumed to begin with.

Experimental data has been used to calculate the friction force as a function of velocity *within* an individual stick–slip cycle [18.294]. For a macroscopic granular material confined between solid surfaces, the data shows a velocity-weakening friction force during the first half of the slip. However, the data also shows a hysteresis loop in the friction–velocity plot, with a different behavior in the deceleration half of the slip phase. Similar results were observed for a 1–2 nm liquid lubricant film between mica surfaces [18.275]. These results indicate that a purely velocity-dependent friction law is insufficient to describe such systems, and an additional element such as the *state* of the confined material must be considered.

Phase Transitions Model. In recent molecular dynamics computer simulations it has been found that thin interfacial films undergo first-order phase transitions between solid-like and liquid-like states during sliding [18.201, 273]. It has been suggested that this is responsible for the observed stick–slip behavior of simple isotropic liquids between two solid crystalline surfaces. With this interpretation, stick–slip is seen to arise because of the abrupt change in the flow properties of a film at a transition [18.224,225,261], rather than the gradual or continuous change, as occurs in the previous example. Other computer simulations indicate that it is the stick–slip that induces a disorder ("shear-melting") in the film, not the other way around [18.295].

An interpretation of the well-known phenomenon of decreasing coefficient of friction with increasing sliding velocity has been proposed by *Thompson* and *Robbins* [18.201] based on their computer simulation. This postulates that it is not the friction that changes with sliding speed v, but rather the time various parts of the system spend in the sticking and sliding modes. In other words, at any instant during sliding, the friction at any local region is always F_s or F_k, corresponding to the "static" or "kinetic" values. The measured frictional force, however, is the sum of all these discrete values averaged over the whole contact area. Since as v increases each local region spends more time in the sliding regime (F_k) and less in the sticking regime (F_s), the overall friction coefficient falls. One may note that this interpretation reverses the traditional way that stick–slip has been explained. Rather than invoking a decreasing friction with velocity to explain stick–slip, it is now the more fundamental stick–slip phenomenon that is producing the apparent decrease in the friction force with increasing sliding velocity. This approach has been studied analytically by *Carlson* and *Batista* [18.296], with a comprehensive rate- and state- dependent friction force law. This model includes an analytic description of the freezing–melting transitions of a film, resulting in a friction force that is a function of sliding velocity in a natural way. This model predicts a full range of stick–slip behavior observed experimentally.

An example of the rate- and state-dependent model is observed when shearing thin films of OMCTS between mica surfaces [18.297, 298]. In this case, the static friction between the surfaces is dependent on the time that the surfaces are at rest with respect to each other, while the intrinsic kinetic friction $F_{k,0}$ is relatively constant over the range of velocities. At slow driving velocities, the system responds with stick–slip sliding with the surfaces reaching maximum static friction before each slip event, and the amplitude of the stick–slip, $F_s - F_k$, is relatively constant. As the driving velocity increases, the static friction decreases as the time at relative rest becomes shorter with respect to the characteristic time of the lubricant film. As the static friction decreases with increasing drive velocity, it eventually equals the intrinsic kinetic friction $F_{k,0}$, which defines the critical velocity v_c, above which the surfaces slide smoothly without the jerky stick–slip motion.

The above classifications of stick–slip are not exclusive, and molecular mechanisms of real systems may exhibit aspects of different models simultaneously. They do, however, provide a convenient classification of existing models and indicate which experimental parameters should be varied to test the different models.

Critical Velocity for Stick–Slip. For any given set of conditions, stick–slip disappears above some critical sliding velocity v_c, above which the motion continues smoothly in the liquid-like or kinetic state. The critical velocity is well described by two simple equations. Both are based on the phase transition model, and both include some parameter associated with the inertia of the measuring instrument. The first equation is based on both experiments and simple theoretical modeling [18.276]:

$$v_c \approx \frac{(F_s - F_k)}{5K\tau_0}\,, \tag{18.48}$$

where τ_0 is the *characteristic nucleation time* or freezing time of the film. For example, inserting the following typically measured values for a ~1-nm-thick hexadecane film between mica: $(F_s - F_k) \approx 5\,\text{mN}$, spring constant $K \approx 500\,\text{N}\,\text{m}^{-1}$, and nucleation time [18.276]

$\tau_0 \approx 5\,\mathrm{s}$, we obtain $v_c \approx 0.4\,\mu\mathrm{m\,s^{-1}}$, which is close to typically measured values (Fig. 18.30b).

The second equation is based on computer simulations [18.273]:

$$v_c \approx 0.1\sqrt{\frac{F_s\sigma}{m}}\,, \tag{18.49}$$

where σ is a molecular dimension and m is the mass of the stage. Again, inserting typical experimental values into this equation, viz., $m \approx 20\,\mathrm{g}$, $\sigma \approx 0.5\,\mathrm{nm}$, and $(F_s - F_k) \approx 5\,\mathrm{mN}$ as before, we obtain $v_c \approx 0.3\,\mu\mathrm{m\,s^{-1}}$, which is also close to measured values.

Stick–slip also disappears above some critical temperature T_c, which is not the same as the melting temperature of the bulk fluid. Certain correlations have been found between v_c and T_c and between various other tribological parameters that appear to be consistent with the principle of time–temperature superposition (see Sect. 18.8.3), similar to that occurring in viscoelastic polymer fluids [18.277, 299, 300].

Recent work on the coupling between the mechanical resonances of the sliding system and molecular-scale relaxations [18.234, 301–303] has resulted in a better understanding of a phenomenon previously noted in various engineering applications: the vibrating of one of the sliding surfaces perpendicularly to the sliding direction can lead to a significant reduction of the friction. At certain oscillation amplitudes and a frequency higher than the molecular-scale relaxation frequency, stick–slip friction can be eliminated and replaced by an ultralow kinetic friction state.

18.9 Role of Molecular Shape and Surface Structure in Friction

The above scenario is already quite complicated, and yet this is the situation for the simplest type of experimental system. The factors that appear to determine the critical velocity v_c depend on the type of liquid between the surfaces (as well as on the surface lattice structure). Small spherical molecules such as cyclohexane and OMCTS have been found to have very high v_c, which indicates that these molecules can rearrange relatively quickly in thin films. Chain molecules and especially branched chain molecules have been found to have much lower v_c, which is to be expected, and such liquids tend to slide smoothly or with erratic stick–slip [18.275], rather than in a stick–slip fashion (see Table 18.5). With highly asymmetric molecules, such as multiply branched isoparaffins and polymer melts, no regular spikes or stick–slip behavior occurs at any speed, since these molecules can never order themselves sufficiently to solidify. Examples of such liquids are some perfluoropolyethers and polydimethylsiloxanes (PDMS).

Recent computer simulations [18.274, 304, 305] of the structure, interaction forces, and tribological behavior of chain molecules between two shearing surfaces indicate that both linear *and* singly or doubly branched chain molecules order between two flat surfaces by aligning into discrete layers parallel to the surfaces. However, in the case of the weakly branched molecules, the expected oscillatory forces do not appear because of a complex cancellation of entropic and enthalpic contributions to the interaction free energy, which results in a monotonically smooth interaction, exhibiting a weak energy minimum rather than the oscillatory force profile that is characteristic of linear molecules. During sliding, however, these molecules can be induced to further align, which can result in a transition from smooth to stick–slip sliding.

Table 18.5 shows the trends observed with some organic and polymeric liquid between smooth mica surfaces. Also listed are the bulk viscosities of the liquids. From the data of Table 18.5 it appears that there is a direct correlation between the shapes of molecules and their coefficient of friction or effectiveness as lubricants (at least at low shear rates). Small spherical or chain molecules have high friction with stick–slip, because they can pack into ordered solid-like layers. In contrast, longer chained and irregularly shaped molecules remain in an entangled, disordered, fluid-like state even in very thin films, and these give low friction and smoother sliding. It is probably for this reason that irregularly shaped branched chain molecules are usually better lubricants. It is interesting to note that the friction coefficient generally decreases as the bulk viscosity of the liquids *increases*. This unexpected trend occurs because the factors that are conducive to low friction are generally conducive to high viscosity. Thus molecules with side groups such as branched alkanes and polymer melts usually have higher bulk viscosities than their linear homologues for obvious reasons. However, in thin films the linear molecules have higher shear stresses, because of their ability to become ordered. The only exception to the

Table 18.5 Effect of molecular shape and short-range forces on tribological properties[a]

Liquid	Short-range force	Type of friction	Friction coefficient	Bulk liquid viscosity (cP)
Organic (water-free)				
Cyclohexane	Oscillatory	Quantized stick–slip	$\gg 1$	0.6
OMCTS[b]	Oscillatory	Quantized stick–slip	$\gg 1$	2.3
Octane	Oscillatory	Quantized stick–slip	1.5	0.5
Tetradecane	Oscillatory $\leftrightarrow$ smooth	stick–slip $\leftrightarrow$ smooth	1.0	2.3
Octadecane (branched)	Oscillatory $\leftrightarrow$ smooth	stick–slip $\leftrightarrow$ smooth	0.3	5.5
PDMS[b] ($M = 3{,}700\,\mathrm{g\,mol^{-1}}$, melt)	Oscillatory $\leftrightarrow$ smooth	Smooth	0.4	50
PBD[b] ($M = 3{,}500\,\mathrm{g\,mol^{-1}}$, branched)	Smooth	Smooth	0.03	800
Water				
Water (KCl solution)	Smooth	Smooth	0.01–0.03	1.0

[a] For molecularly thin liquid films between two shearing mica surfaces at 20 °C

[b] OMCTS: Octamethylcyclotetrasiloxane, PDMS: Polydimethylsiloxane, PBD: Polybutadiene

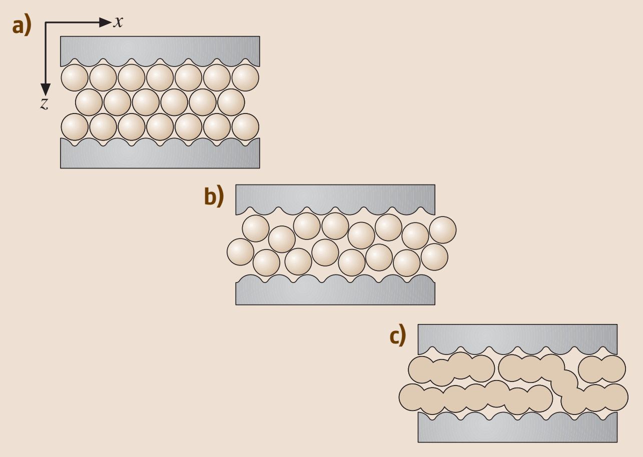

Fig. 18.32a–c Schematic view of interfacial film composed of spherical molecules under a compressive pressure between two solid crystalline surfaces. **(a)** If the two surface lattices are free to move in the x-y-z directions, so as to attain the lowest energy state, they could equilibrate at values of x, y, and z, which induce the trapped molecules to become "epitaxially" ordered into a "solid-like" film. **(b)** Similar view of trapped molecules between two solid surfaces that are not free to adjust their positions, for example, as occurs in capillary pores or in brittle cracks. **(c)** Similar to **(a)**, but with chain molecules replacing the spherical molecules in the gap. These may not be able to order as easily as do spherical molecules even if x, y, and z can adjust, resulting in a situation that is more akin to **(b)**. (After [18.276] with permission. Copyright 1993 American Chemical Society)

above correlations is water, which has been found to exhibit both low viscosity *and* low friction (see Fig. 18.20, and Sect. 18.7.3). In addition, the presence of water can drastically lower the friction and eliminate the stick–slip of hydrocarbon liquids when the sliding surfaces are hydrophilic.

Fig. 18.33a–f Schematic representation of the film under shear. **(a)** The lubricant molecules are just confined, but not oriented in any particular direction. Because of the need to shear, the film dilates **(b)**. The molecules disentangle **(c)** and get oriented in a certain direction related to the shear direction **(d)**. **(e)** Slowly evolving domains grow inside the contact region. These macroscopic domains are responsible for the long relaxation times. **(f)** At the steady-state, a continuous gradient of confinement time and molecular order is established in the contact region, which is different for molecules adsorbed on the upper and lower surfaces. Molecules entering into the contact are not oriented or ordered. The required sliding distance to modify their state defines a characteristic distance. Molecules leaving the contact region need some (short) characteristic time to regain their bulk, unconfined configuration. (After [18.281] with permission. Copyright 2000 American Chemical Society)

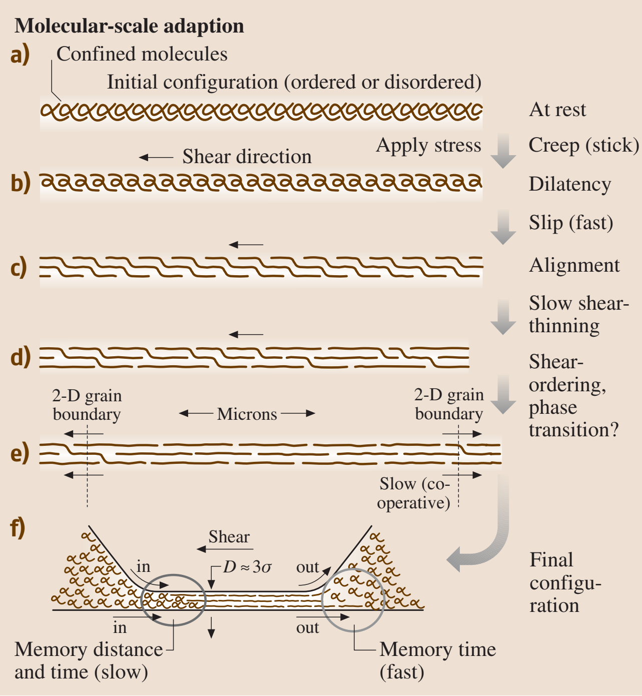

If an "effective" viscosity, η_{eff}, were to be calculated for the liquids of Table 18.5, the values would be many orders of magnitude higher than those of the bulk liquids. This can be demonstrated by the following simple calculation based on the usual equation for Couette flow (see (18.44)):

$$\eta_{\text{eff}} = F_{\text{k}} D / A v\,, \tag{18.50}$$

where F_{k} is the kinetic friction force, D is the film thickness, A the contact area, and v the sliding velocity. Using typical values for experiments with hexadecane [18.276]: $F_{\text{k}} = 5\,\text{mN}$, $D = 1\,\text{nm}$, $A = 3\times10^{-9}\,\text{m}^2$, and $v = 1\,\mu\text{m}\,\text{s}^{-1}$, one gets $\eta_{\text{eff}} \approx 2{,}000\,\text{Ns}\,\text{m}^{-2}$, or 20,000 Poise, which is $\sim 10^6$ times higher than the bulk viscosity, η_{bulk}, of the liquid. It is instructive to consider that this very high effective viscosity nevertheless still produces a low friction force or friction coefficient μ of about 0.25. It is interesting to speculate that if a 1 nm film were to exhibit bulk viscous behavior, the friction coefficient under the same sliding conditions would be as low as 0.000001. While such a low value has never been reported for any tribological system, one may consider it a theoretical lower limit that could conceivably be attained under certain experimental conditions.

Various studies [18.44, 219, 220, 222] have shown that confinement and load generally increase the effective viscosity and/or relaxation times of molecules, suggestive of an increased glassiness or solid-like behavior (Figs. 18.32 and 18.33). This is in marked contrast to studies of liquids in small confining capillaries where the opposite effects have been observed [18.306, 307]. The reason for this is probably because the two modes of confinement are different. In the former case (confinement of molecules between two structured solid surfaces), there is generally little opposition to any lateral or vertical displacement of the two surface lattices relative to each other. This means that the two lattices can shift in the x-y-z plane (Fig. 18.32a) to accommodate the trapped molecules in the most crystallographically commensurate or epitaxial way, which would favor an ordered, solid-like state. In contrast, the walls of capillaries are rigid and cannot easily move or adjust to accommodate the confined molecules (Fig. 18.32b), which will therefore be forced into a more disordered, liquid-like state (unless the capillary wall geometry and lattice are *exactly* commensurate with the liquid molecules, as occurs in certain zeolites [18.307]).

Experiments have demonstrated the effects of surface lattice mismatch on the friction between surfaces [18.203, 204, 308]. Similar to the effects of lattice mismatch on adhesion (Fig. 18.11), the static friction of a confined liquid film is maximum when the lattices of the confining surfaces are aligned. For OMCTS confined between mica surfaces [18.204] the static friction was found to vary by more than a factor of 4, while for bare mica surfaces the variation was by a factor of 3.5 [18.308]. In contrast to the sharp variations in adhesion energy over small twist angles, the variations in friction as a function of twist angle were much more broad both in magnitude and angular spread. Similar variations in friction as a function of twist or misfit angles have also been observed in computer simulations [18.309].

Robbins et al. [18.274] computed the friction forces of two clean crystalline surfaces as a function of the angle between their surface lattices. They found that for all non-zero angles (finite "twist" angles) the friction forces fell to zero due to incommensurability effects. They further found that submonolayer amounts of organic or other impurities trapped between two incommensurate surfaces can generate a finite friction force. They, therefore, concluded that any finite friction force measured between incommensurate surfaces is probably due to such "third body" effects.

With rough surfaces, i. e., those that have *random* protrusions rather than being periodically structured, we expect a smearing out of the correlated intermolecular interactions that are involved in film freezing and melting (and in phase transitions in general). This should effectively eliminate the highly regular stick–slip and may also affect the location of the slipping planes [18.123, 252, 279]. The stick–slip friction of "real" surfaces, which are generally rough, may, therefore, be quite different from those of perfectly smooth surfaces composed of the same material. We should note, however, that even between rough surfaces, most of the contacts occur between the tips of microscopic asperities, which may be smooth over their microscopic contact area.

References

18.1 J. N. Israelachvili: Surface Forces and Microrheology of Molecularly Thin Liquid Films. In: *Handbook of Micro/Nanotribology*, ed. by B. Bhushan (CRC Press, Boca Raton 1995) pp. 267–319
18.2 K. B. Lodge: Techniques for the measurement of forces between solids, Adv. Colloid Interface Sci. **19** (1983) 27–73
18.3 J. N. Israelachvili: *Intermolecular and Surface Forces*, 2nd edn. (Academic Press, London 1991)
18.4 P. F. Luckham, B. A. de L. Costello: Recent developments in the measurement of interparticle forces, Adv. Colloid Interface Sci. **44** (1993) 183–240
18.5 P. M. Claesson, T. Ederth, V. Bergeron, M. W. Rutland: Techniques for measuring surface forces, Adv. Colloid Interface Sci. **67** (1996) 119–183
18.6 V. S. J. Craig: An historical review of surface force measurement techniques, Colloids Surf. A **129–130** (1997) 75–94
18.7 J. N. Israelachvili, G. E. Adams: Measurements of forces between two mica surfaces in aqueous electrolyte solutions in the range 0–100 nm, J. Chem. Soc. Faraday Trans. I **74** (1978) 975–1001
18.8 G. Binnig, C. F. Quate, C. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933
18.9 W. A. Ducker, T. J. Senden, R. M. Pashley: Direct measurement of colloidal forces using an atomic force microscope, Nature **353** (1991) 239–241
18.10 E. Meyer, R. M. Overney, K. Dransfeld, T. Gyalog: *Nanoscience: Friction and Rheology on the Nanometer Scale* (World Scientific, Singapore 1998)
18.11 J. N. Israelachvili: Measurements of the viscosity of thin fluid films between two surfaces with and without adsorbed polymers, Colloid Polym. Sci. **264** (1986) 1060–1065
18.12 J. P. Montfort, G. Hadziioannou: Equilibrium and dynamic behavior of thin films of a perfluorinated polyether, J. Chem. Phys. **88** (1988) 7187–7196
18.13 A. Dhinojwala, S. Granick: Surface forces in the tapping mode: Solvent permeability and hydrodynamic thickness of adsorbed polymer brushes, Macromolecules **30** (1997) 1079–1085
18.14 B. V. Derjaguin: Untersuchungen über die Reibung und Adhäsion, IV. Theorie des Anhaftens kleiner Teilchen, Kolloid Z. **69** (1934) 155–164
18.15 K. L. Johnson, K. Kendall, A. D. Roberts: Surface energy and the contact of elastic solids, Proc. R. Soc. London A **324** (1971) 301–313
18.16 J. N. Israelachvili, P. M. McGuiggan: Adhesion and short-range forces between surfaces. Part 1: New apparatus for surface force measurements, J. Mater. Res. **5** (1990) 2223–2231
18.17 J. N. Israelachvili: Thin film studies using multiple-beam interferometry, J. Colloid Interface Sci. **44** (1973) 259–272
18.18 Y. L. Chen, T. Kuhl, J. Israelachvili: Mechanism of cavitation damage in thin liquid films: Collapse damage vs. inception damage, Wear **153** (1992) 31–51
18.19 M. Heuberger, G. Luengo, J. Israelachvili: Topographic information from multiple beam interferometry in the surface forces apparatus, Langmuir **13** (1997) 3839–3848
18.20 R. M. Pashley: DLVO and hydration forces between mica surfaces in Li^+, Na^+, K^+, and Cs^+ electrolyte solutions: A correlation of double-layer and hydration forces with surface cation exchange properties, J. Colloid Interface Sci. **83** (1981) 531–546

18.21 R. M. Pashley: Hydration forces between mica surfaces in electrolyte solution, Adv. Colloid Interface Sci. **16** (1982) 57–62

18.22 R. G. Horn, D. T. Smith, W. Haller: Surface forces and viscosity of water measured between silica sheets, Chem. Phys. Lett. **162** (1989) 404–408

18.23 R. G. Horn, D. R. Clarke, M. T. Clarkson: Direct measurements of surface forces between sapphire crystals in aqueous solutions, J. Mater. Res. **3** (1988) 413–416

18.24 W. W. Merrill, A. V. Pocius, B. V. Thakker, M. Tirrell: Direct measurement of molecular level adhesion forces between biaxially oriented solid polymer films, Langmuir **7** (1991) 1975–1980

18.25 J. Klein: Forces between mica surfaces bearing adsorbed macromolecules in liquid media, J. Chem. Soc. Faraday Trans. I **79** (1983) 99–118

18.26 S. S. Patel, M. Tirrell: Measurement of forces between surfaces in polymer fluids, Annu. Rev. Phys. Chem. **40** (1989) 597–635

18.27 H. Watanabe, M. Tirrell: Measurements of forces in symmetric and asymmetric interactions between diblock copolymer layers adsorbed on mica, Macromolecules **26** (1993) 6455–6466

18.28 T. L. Kuhl, D.E. Leckband, D. D. Lasic, J. N. Israelachvili: Modulation and modeling of interaction forces between lipid bilayers exposing terminally grafted polymer chains. In: *Stealth Liposomes*, ed. by D. Lasic, F. Martin (CRC Press, Boca Raton 1995) pp. 73–91

18.29 J. Klein: Shear, friction, and lubrication forces between polymer-bearing surfaces, Annu. Rev. Mater. Sci. **26** (1996) 581–612

18.30 M. Ruths, D. Johannsmann, J. Rühe, W. Knoll: Repulsive forces and relaxation on compression of entangled, polydisperse polystyrene brushes, Macromolecules **33** (2000) 3860–3870

18.31 C. A. Helm, J. N. Israelachvili, P. M. McGuiggan: Molecular mechanisms and forces involved in the adhesion and fusion of amphiphilic bilayers, Science **246** (1989) 919–922

18.32 Y. L. Chen, C. A. Helm, J. N. Israelachvili: Molecular mechanisms associated with adhesion and contact angle hysteresis of monolayer surfaces, J. Phys. Chem. **95** (1991) 10736–10747

18.33 D. E. Leckband, J. N. Israelachvili, F.-J. Schmitt, W. Knoll: Long-range attraction and molecular rearrangements in receptor-ligand interactions, Science **255** (1992) 1419–1421

18.34 J. Peanasky, H. M. Schneider, S. Granick, C. R. Kessel: Self-assembled monolayers on mica for experiments utilizing the surface forces apparatus, Langmuir **11** (1995) 953–962

18.35 C. J. Coakley, D. Tabor: Direct measurement of van der Waals forces between solids in air, J. Phys. D **11** (1978) L77–L82

18.36 J. L. Parker, H. K. Christenson: Measurements of the forces between a metal surface and mica across liquids, J. Chem. Phys. **88** (1988) 8013–8014

18.37 C. P. Smith, M. Maeda, L. Atanasoska, H. S. White, D. J. McClure: Ultrathin platinum films on mica and the measurement of forces at the platinum/water interface, J. Phys. Chem. **92** (1988) 199–205

18.38 S. J. Hirz, A. M. Homola, G. Hadziioannou, C. W. Frank: Effect of substrate on shearing properties of ultrathin polymer films, Langmuir **8** (1992) 328–333

18.39 J. M. Levins, T. K. Vanderlick: Reduction of the roughness of silver films by the controlled application of surface forces, J. Phys. Chem. **96** (1992) 10405–10411

18.40 S. Steinberg, W. Ducker, G. Vigil, C. Hyukjin, C. Frank, M. Z. Tseng, D. R. Clarke, J. N. Israelachvili: Van der Waals epitaxial growth of α-alumina nanocrystals on mica, Science **260** (1993) 656–659

18.41 G. Vigil, Z. Xu, S. Steinberg, J. Israelachvili: Interactions of silica surfaces, J. Colloid Interface Sci. **165** (1994) 367–385

18.42 M. Ruths, M. Heuberger, V. Scheumann, J. Hu, W. Knoll: Confinement-induced film thickness transitions in liquid crystals between two alkanethiol monolayers on gold, Langmuir **17** (2001) 6213–6219

18.43 J. Van Alsten, S. Granick: Molecular tribometry of ultrathin liquid films, Phys. Rev. Lett. **61** (1988) 2570–2573

18.44 J. Van Alsten, S. Granick: Shear rheology in a confined geometry: Polysiloxane melts, Macromolecules **23** (1990) 4856–4862

18.45 A. M. Homola, J. N. Israelachvili, M. L. Gee, P. M. McGuiggan: Measurements of and relation between the adhesion and friction of two surfaces separated by molecularly thin liquid films, J. Tribol. **111** (1989) 675–682

18.46 J. Klein, E. Kumacheva: Simple liquids confined to molecularly thin layers. I. Confinement-induced liquid-to-solid phase transitions, J. Chem. Phys. **108** (1998) 6996–7009

18.47 G. Luengo, F.-J. Schmitt, R. Hill, J. Israelachvili: Thin film rheology and tribology of confined polymer melts: Contrasts with bulk properties, Macromolecules **30** (1997) 2482–2494

18.48 E. Kumacheva: Interfacial friction measurements in surface force apparatus, Prog. Surf. Sci. **58** (1998) 75–120

18.49 A. L. Weisenhorn, P. K. Hansma, T. R. Albrecht, C. F. Quate: Forces in atomic force microscopy in air and water, Appl. Phys. Lett. **54** (1989) 2651–2653

18.50 G. Meyer, N. M. Amer: Simultaneous measurement of lateral and normal forces with an optical-beam-deflection atomic force microscope, Appl. Phys. Lett. **57** (1990) 2089–2091

18.51 E. L. Florin, V. T. Moy, H. E. Gaub: Adhesion forces between individual ligand–receptor pairs, Science **264** (1994) 415–417

18.52 G. U. Lee, D. A. Kidwell, R. J. Colton: Sensing discrete streptavidin–biotin interactions with atomic force microscopy, Langmuir **10** (1994) 354–357

18.53 C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang: Atomic-scale friction of a tungsten tip on a graphite surface, Phys. Rev. Lett. **59** (1987) 1942–1945

18.54 R. W. Carpick, M. Salmeron: Scratching the surface: Fundamental investigations of tribology with atomic force microscopy, Chem. Rev. **97** (1997) 1163–1194

18.55 J. P. Cleveland, S. Manne, D. Bocek, P. K. Hansma: A nondestructive method for determining the spring constant of cantilevers for scanning force microscopy, Rev. Sci. Instrum. **64** (1993) 403–405

18.56 J. E. Sader, J. W. M. Chon, P. Mulvaney: Calibration of rectangular atomic force microscope cantilevers, Rev. Sci. Instrum. **70** (1999) 3967–3969

18.57 Y. Liu, T. Wu, D. F. Evans: Lateral force microscopy study on the shear properties of self-assembled monolayers of dialkylammonium surfactant on mica, Langmuir **10** (1994) 2241–2245

18.58 C. Neto, V. S. J. Craig: Colloid probe characterization: Radius and roughness determination, Langmuir **17** (2001) 2097–2099

18.59 R. G. Horn, J. N. Israelachvili: Molecular organization and viscosity of a thin film of molten polymer between two surfaces as probed by force measurements, Macromolecules **21** (1988) 2836–2841

18.60 R. G. Horn, S. J. Hirz, G. Hadziioannou, C. W. Frank, J. M. Catala: A reevaluation of forces measured across thin polymer films: Nonequilibrium and pinning effects, J. Chem. Phys. **90** (1989) 6767–6774

18.61 O. I. Vinogradova, H.-J. Butt, G. E. Yakubov, F. Feuillebois: Dynamic effects on force measurements. I. Viscous drag on the atomic force microscope cantilever, Rev. Sci. Instrum. **72** (2001) 2330–2339

18.62 D. Leckband, J. Israelachvili: Intermolecular forces in biology, Quart. Rev. Biophys. **34** (2001) 105–267

18.63 E. Evans, D. Needham: Physical properties of surfactant bilayer membranes: Thermal transitions, elasticity, rigidity, cohesion, and colloidal interactions, J. Phys. Chem. **91** (1987) 4219–4228

18.64 E. Evans, D. Needham: Attraction between lipid bilayer membranes in concentrated solutions of nonadsorbing polymers: Comparison of mean-field theory with measurements of adhesion energy, Macromolecules **21** (1988) 1822–1831

18.65 S. E. Chesla, P. Selvaraj, C. Zhu: Measuring two-dimensional receptor-ligand binding kinetics by micropipette, Biophys. J. **75** (1998) 1553–1572

18.66 E. Evans, K. Ritchie, R. Merkel: Sensitive force technique to probe molecular adhesion and structural linkages at biological interfaces, Biophys. J. **68** (1995) 2580–2587

18.67 D. M. LeNeveu, R. P. Rand, V. A. Parsegian: Measurements of forces between lecithin bilayers, Nature **259** (1976) 601–603

18.68 A. Homola, A. A. Robertson: A compression method for measuring forces between colloidal particles, J. Colloid Interface Sci. **54** (1976) 286–297

18.69 V. A. Parsegian, N. Fuller, R. P. Rand: Measured work of deformation and repulsion of lecithin bilayers, Proc. Nat. Acad. Sci. USA **76** (1979) 2750–2754

18.70 R. P. Rand, V. A. Parsegian: Hydration forces between phospholipid bilayers, Biochim. Biophys. Acta **988** (1989) 351–376

18.71 S. Leikin, V. A. Parsegian, D. C. Rau, R. P. Rand: Hydration forces, Annu. Rev. Phys. Chem. **44** (1993) 369–395

18.72 S. Chu, J. E. Bjorkholm, A. Ashkin, A. Cable: Experimental observation of optically trapped atoms, Phys. Rev. Lett. **57** (1986) 314–317

18.73 A. Ashkin: Optical trapping and manipulation of neutral particles using lasers, Proc. Nat. Acad. Sci. USA **94** (1997) 4853–4860

18.74 K. Visscher, S. P. Gross, S. M. Block: Construction of multiple-beam optical traps with nanometer-resolution positioning sensing, IEEE J. Sel. Top. Quantum Electron. **2** (1996) 1066–1076

18.75 D. C. Prieve, N. A. Frej: Total internal reflection microscopy: A quantitative tool for the measurement of colloidal forces, Langmuir **6** (1990) 396–403

18.76 J. Rädler, E. Sackmann: On the measurement of weak repulsive and frictional colloidal forces by reflection interference contrast microscopy, Langmuir **8** (1992) 848–853

18.77 P. Bongrand, C. Capo, J.-L. Mege, A.-M. Benoliel: Use of hydrodynamic flows to study cell adhesion. In: *Physical Basis of Cell–Cell Adhesion*, ed. by P. Bongrand (CRC Press, Boca Raton 1988) pp. 125–156

18.78 G. Kaplanski, C. Farnarier, O. Tissot, A. Pierres, A.-M. Benoliel, A.-C. Alessi, S. Kaplanski, P. Bongrand: Granulocyte–endothelium initial adhesion: Analysis of transient binding events mediated by E-selectin in a laminar shear flow, Biophys. J. **64** (1993) 1922–1933

18.79 H. B. G. Casimir, D. Polder: The influence of retardation on the London–van der Waals forces, Phys. Rev. **73** (1948) 360–372

18.80 H. C. Hamaker: The London–van der Waals attraction between spherical particles, Physica **4** (1937) 1058–1072

18.81 E. M. Lifshitz: The theory of molecular attraction forces between solid bodies, Sov. Phys. JETP (English translation) **2** (1956) 73–83

18.82 I. E. Dzyaloshinskii, E. M. Lifshitz, L. P. Pitaevskii: The general theory of van der Waals forces, Adv. Phys. **10** (1961) 165–209

18.83 H. W. Fox, W. A. Zisman: The spreading of liquids on low-energy surfaces. III. Hydrocarbon surfaces, J. Colloid Sci. **7** (1952) 428–442

18.84 B. V. Derjaguin, V. P. Smilga: Electrostatic component of the rolling friction force moment, Wear **7** (1964) 270–281

18.85 R. G. Horn, D. T. Smith: Contact electrification and adhesion between dissimilar materials, Science **256** (1992) 362–364

18.86 R. Budakian, S. J. Putterman: Correlation between charge transfer and stick–slip friction at a metal–insulator interface, Phys. Rev. Lett. **85** (2000) 1000–1003

18.87 M. Grätzel: Photoelectrochemical cells, Nature **414** (2001) 338–344

18.88 R. M. German: *Sintering Theory and Practice* (Wiley, New York 1996)

18.89 R. Budakian, S. J. Putterman: Time scales for cold welding and the origins of stick–slip friction, Phys. Rev. B **65** (2002) 235429/1–5

18.90 U. Landman, W. D. Luedtke, N. A. Burnham, R. J. Colton: Atomistic mechanisms and dynamics of adhesion, nanoindentation, and fracture, Science **248** (1990) 454–461

18.91 U. Landman, W. D. Luedtke, E. M. Ringer: Molecular dynamics simulations of adhesive contact formation and friction, NATO Science Ser. E **220** (1992) 463–510

18.92 W. D. Luedtke, U. Landman: Solid and liquid junctions, Comput. Mater. Sci. **1** (1992) 1–24

18.93 U. Landman, W. D. Luedtke: Interfacial junctions and cavitation, MRS Bull. **18** (1993) 36–44

18.94 B. Bhushan, J. N. Israelachvili, U. Landman: Nanotribology: Friction, wear and lubrication at the atomic scale, Nature **374** (1995) 607–616

18.95 M. R. Sørensen, K. W. Jacobsen, P. Stoltze: Simulations of atomic-scale sliding friction, Phys. Rev. B **53** (1996) 2101–2113

18.96 A. Meurk, P. F. Luckham, L. Bergström: Direct measurement of repulsive and attractive van der Waals forces between inorganic materials, Langmuir **13** (1997) 3896–3899

18.97 E. J. W. Verwey, J. T. G. Overbeek: *Theory of the Stability of Lyophobic Colloids*, 1st edn. (Elsevier, Amsterdam 1948)

18.98 D. Y. C. Chan, R. M. Pashley, L. R. White: A simple algorithm for the calculation of the electrostatic repulsion between identical charged surfaces in electrolyte, J. Colloid Interface Sci. **77** (1980) 283–285

18.99 B. Derjaguin, L. Landau: Theory of the stability of strongly charged lyophobic sols and of the adhesion of strongly charged particles in solutions of electrolytes, Acta Physicochim. URSS **14** (1941) 633–662

18.100 D. Chan, T. W. Healy, L. R. White: Electrical double layer interactions under regulation by surface ionization equilibriums – dissimilar amphoteric surfaces, J. Chem. Soc. Faraday Trans. 1 **72** (1976) 2844–2865

18.101 G. S. Manning: Limiting laws and counterion condensation in polyelectrolyte solutions. I. Colligative properties, J. Chem. Phys. **51** (1969) 924–933

18.102 L. Guldbrand, V. Jönsson, H. Wennerström, P. Linse: Electrical double-layer forces: A Monte Carlo study, J. Chem. Phys. **80** (1984) 2221–2228

18.103 H. Wennerström, B. Jönsson, P. Linse: The cell model for polyelectrolyte systems. Exact statistical mechanical relations, Monte Carlo simulations, and the Poisson–Boltzmann approximation, J. Chem. Phys. **76** (1982) 4665–4670

18.104 J. Marra: Effects of counterion specificity on the interactions between quaternary ammonium surfactants in monolayers and bilayers, J. Phys. Chem. **90** (1986) 2145–2150

18.105 J. Marra: Direct measurement of the interaction between phosphatidylglycerol bilayers in aqueous-electrolyte solutions, Biophys. J. **50** (1986) 815–825

18.106 R. G. Horn, J. N. Israelachvili: Direct measurement of structural forces between two surfaces in a nonpolar liquid, J. Chem. Phys. **75** (1981) 1400–1411

18.107 I. K. Snook, W. van Megen: Solvation forces in simple dense fluids I, J. Chem. Phys. **72** (1980) 2907–2913

18.108 W. J. van Megen, I. K. Snook: Solvation forces in simple dense fluids II. Effect of chemical potential, J. Chem. Phys. **74** (1981) 1409–1411

18.109 R. Kjellander, S. Marcelja: Perturbation of hydrogen bonding in water near polar surfaces, Chem. Phys. Lett. **120** (1985) 393–396

18.110 P. Tarazona, L. Vicente: A model for the density oscillations in liquids between solid walls, Mol. Phys. **56** (1985) 557–572

18.111 D. Henderson, M. Lozada-Cassou: A simple theory for the force between spheres immersed in a fluid, J. Colloid Interface Sci. **114** (1986) 180–183

18.112 J. E. Curry, J. H. Cushman: Structure in confined fluids: Phase separation of binary simple liquid mixtures, Tribol. Lett. **4** (1998) 129–136

18.113 M. Schoen, T. Gruhn, D. J. Diestler: Solvation forces in thin films confined between macroscopically curved substrates, J. Chem. Phys. **109** (1998) 301–311

18.114 F. Porcheron, B. Rousseau, M. Schoen, A. H. Fuchs: Structure and solvation forces in confined alkane films, Phys. Chem. Chem. Phys. **3** (2001) 1155–1159

18.115 H. K. Christenson, D. W. R. Gruen, R. G. Horn, J. N. Israelachvili: Structuring in liquid alkanes between solid surfaces: Force measurements and mean-field theory, J. Chem. Phys. **87** (1987) 1834–1841

18.116 M. L. Gee, J. N. Israelachvili: Interactions of surfactant monolayers across hydrocarbon liquids, J. Chem. Soc. Faraday Trans. **86** (1990) 4049–4058

18.117 J. N. Israelachvili, S. J. Kott, M. L. Gee, T. A. Witten: Forces between mica surfaces across hydrocarbon liquids: Effects of branching and polydispersity, Macromolecules **22** (1989) 4247–4253

18.118 H. K. Christenson: Forces between solid surfaces in a binary mixture of non-polar liquids, Chem. Phys. Lett. **118** (1985) 455–458

18.119 H. K. Christenson, R. G. Horn: Solvation forces measured in non-aqueous liquids, Chem. Scr. **25** (1985) 37–41

18.120 J. Israelachvili: Solvation forces and liquid structure, as probed by direct force measurements, Acc. Chem. Res. **20** (1987) 415–421

18.121 H. K. Christenson: Non-DLVO forces between surfaces – solvation, hydration and capillary effects, J. Disp. Sci. Technol. **9** (1988) 171–206

18.122 L. J. D. Frink, F. van Swol: Solvation forces between rough surfaces, J. Chem. Phys. **108** (1998) 5588–5598

18.123 J. Gao, W. D. Luedtke, U. Landman: Structures, solvation forces and shear of molecular films in a rough nano-confinement, Tribol. Lett. **9** (2000) 3–13

18.124 H. E. Stanley, J. Teixeira: Interpretation of the unusual behavior of H_2O and D_2O at low temperatures: Tests of a percolation model, J. Chem. Phys. **73** (1980) 3404–3422

18.125 H. van Olphen: *An Introduction to Clay Colloid Chemistry*, 2nd edn. (Wiley, New York 1977) Chap. 10

18.126 U. Del Pennino, E. Mazzega, S. Valeri, A. Alietti, M. F. Brigatti, L. Poppi: Interlayer water and swelling properties of monoionic montmorillonites, J. Colloid Interface Sci. **84** (1981) 301–309

18.127 B. E. Viani, P. F. Low, C. B. Roth: Direct measurement of the relation between interlayer force and interlayer distance in the swelling of montmorillonite, J. Colloid Interface Sci. **96** (1983) 229–244

18.128 R. G. Horn: Surface forces and their action in ceramic materials, J. Am. Ceram. Soc. **73** (1990) 1117–1135

18.129 B. V. Velamakanni, J. C. Chang, F. F. Lange, D. S. Pearson: New method for efficient colloidal particle packing via modulation of repulsive lubricating hydration forces, Langmuir **6** (1990) 1323–1325

18.130 R. R. Lessard, S. A. Zieminski: Bubble coalescence and gas transfer in aqueous electrolytic solutions, Ind. Eng. Chem. Fundam. **10** (1971) 260–269

18.131 J. Israelachvili, R. Pashley: The hydrophobic interaction is long range, decaying exponentially with distance, Nature **300** (1982) 341–342

18.132 R. M. Pashley, P. M. McGuiggan, B. W. Ninham, D. F. Evans: Attractive forces between uncharged hydrophobic surfaces: Direct measurements in aqueous solutions, Science **229** (1985) 1088–1089

18.133 P. M. Claesson, C. E. Blom, P. C. Herder, B. W. Ninham: Interactions between water-stable hydrophobic Langmuir–Blodgett monolayers on mica, J. Colloid Interface Sci. **114** (1986) 234–242

18.134 Ya. I. Rabinovich, B. V. Derjaguin: Interaction of hydrophobized filaments in aqueous electrolyte solutions, Colloids Surf. **30** (1988) 243–251

18.135 J. L. Parker, D. L. Cho, P. M. Claesson: Plasma modification of mica: Forces between fluorocarbon surfaces in water and a nonpolar liquid, J. Phys. Chem. **93** (1989) 6121–6125

18.136 H. K. Christenson, J. Fang, B. W. Ninham, J. L. Parker: Effect of divalent electrolyte on the hydrophobic attraction, J. Phys. Chem. **94** (1990) 8004–8006

18.137 K. Kurihara, S. Kato, T. Kunitake: Very strong long range attractive forces between stable hydrophobic monolayers of a polymerized ammonium surfactant, Chem. Lett. (1990) 1555–1558

18.138 Y. H. Tsao, D. F. Evans, H. Wennerstrom: Long-range attractive force between hydrophobic surfaces observed by atomic force microscopy, Science **262** (1993) 547–550

18.139 Ya. I. Rabinovich, R.-H. Yoon: Use of atomic force microscope for the measurements of hydrophobic forces between silanated silica plate and glass sphere, Langmuir **10** (1994) 1903–1909

18.140 V. S. J. Craig, B. W. Ninham, R. M. Pashley: Study of the long-range hydrophobic attraction in concentrated salt solutions and its implications for electrostatic models, Langmuir **14** (1998) 3326–3332

18.141 P. Kékicheff, O. Spalla: Long-range electrostatic attraction between similar, charge-neutral walls, Phys. Rev. Lett. **75** (1995) 1851–1854

18.142 H. K. Christenson, P. M. Claesson: Direct measurements of the force between hydrophobic surfaces in water, Adv. Colloid Interface Sci. **91** (2001) 391–436

18.143 P. Attard: Nanobubbles and the hydrophobic attraction, Adv. Colloid Interface Sci. **104** (2003) 75–91

18.144 M. Hato: Attractive forces between surfaces of controlled “hydrophobicity” across water: A possible range of “hydrophobic interactions” between macroscopic hydrophobic surfaces across water, J. Phys. Chem. **100** (1996) 18530–18538

18.145 H. K. Christenson, P. M. Claesson, J. Berg, P. C. Herder: Forces between fluorocarbon surfactant monolayers: Salt effects on the hydrophobic interaction, J. Phys. Chem. **93** (1989) 1472–1478

18.146 N. I. Christou, J. S. Whitehouse, D. Nicholson, N. G. Parsonage: A Monte Carlo study of fluid water in contact with structureless walls, Faraday Symp. Chem. Soc. **16** (1981) 139–149

18.147 B. Jönsson: Monte Carlo simulations of liquid water between two rigid walls, Chem. Phys. Lett. **82** (1981) 520–525

18.148 S. Marcelja, D. J. Mitchell, B. W. Ninham, M. J. Sculley: Role of solvent structure in solution theory, J. Chem. Soc. Faraday Trans. II **73** (1977) 630–648

18.149 D. W. R. Gruen, S. Marcelja: Spatially varying polarization in water: A model for the electric double layer and the hydration force, J. Chem. Soc. Faraday Trans. 2 **79** (1983) 225–242

18.150 B. Jönsson, H. Wennerström: Image-charge forces in phospholipid bilayer systems, J. Chem. Soc. Faraday Trans. 2 **79** (1983) 19–35

18.151 D. Schiby, E. Ruckenstein: The role of the polarization layers in hydration forces, Chem. Phys. Lett. **95** (1983) 435–438

18.152 A. Luzar, D. Bratko, L. Blum: Monte Carlo simulation of hydrophobic interaction, J. Chem. Phys. **86** (1987) 2955–2959

18.153 P. Attard, M.T. Batchelor: A mechanism for the hydration force demonstrated in a model system, Chem. Phys. Lett. **149** (1988) 206–211

18.154 K. Leung, A. Luzar: Dynamics of capillary evaporation. II. Free energy barriers, J. Chem. Phys. **113** (2000) 5845–5852

18.155 D. Bratko, R. A. Curtis, H. W. Blanch, J. M. Prausnitz: Interaction between hydrophobic surfaces with metastable intervening liquid, J. Chem. Phys. **115** (2001) 3873–3877

18.156 P. G. de Gennes: Polymers at an interface. 2. Interaction between two plates carrying adsorbed polymer layers, Macromolecules **15** (1982) 492–500

18.157 J. M. H. M. Scheutjens, G. J. Fleer: Interaction between two adsorbed polymer layers, Macromolecules **18** (1985) 1882–1900

18.158 E. A. Evans: Force between surfaces that confine a polymer solution: Derivation from self-consistent field theories, Macromolecules **22** (1989) 2277–2286

18.159 H. J. Ploehn: Compression of polymer interphases, Macromolecules **27** (1994) 1627–1636

18.160 S. Asakura, F. Oosawa: Interaction between particles suspended in solutions of macromolecules, J. Polym. Sci. **33** (1958) 183–192

18.161 J. F. Joanny, L. Leibler, P. G. de Gennes: Effects of polymer solutions on colloid stability, J. Polym. Sci. Polym. Phys. **17** (1979) 1073–1084

18.162 B. Vincent, P. F. Luckham, F. A. Waite: The effect of free polymer on the stability of sterically stabilized dispersions, J. Colloid Interface Sci. **73** (1980) 508–521

18.163 R. I. Feigin, D. H. Napper: Stabilization of colloids by free polymer, J. Colloid Interface Sci. **74** (1980) 567–571

18.164 P. G. de Gennes: Polymer solutions near an interface. 1. Adsorption and depletion layers, Macromolecules **14** (1981) 1637–1644

18.165 P. G. de Gennes: Polymers at an interface; a simplified view, Adv. Colloid Interface Sci. **27** (1987) 189–209

18.166 S. Alexander: Adsorption of chain molecules with a polar head. A scaling description, J. Phys. (France) **38** (1977) 983–987

18.167 P. G. de Gennes: Conformations of polymers attached to an interface, Macromolecules **13** (1980) 1069–1075

18.168 S. T. Milner, T. A. Witten, M. E. Cates: Theory of the grafted polymer brush, Macromolecules **21** (1988) 2610–2619

18.169 S. T. Milner, T. A. Witten, M. E. Cates: Effects of polydispersity in the end-grafted polymer brush, Macromolecules **22** (1989) 853–861

18.170 E. B. Zhulina, O. V. Borisov, V. A. Priamitsyn: Theory of steric stabilization of colloid dispersions by grafted polymers, J. Colloid Interface Sci. **137** (1990) 495–511

18.171 H. J. Taunton, C. Toprakcioglu, L. J. Fetters, J. Klein: Interactions between surfaces bearing end-adsorbed chains in a good solvent, Macromolecules **23** (1990) 571–580

18.172 J. Klein, P. Luckham: Forces between two adsorbed poly(ethylene oxide) layers immersed in a good aqueous solvent, Nature **300** (1982) 429–431

18.173 J. Klein, P. F. Luckham: Long-range attractive forces between two mica surfaces in an aqueous polymer solution, Nature **308** (1984) 836–837

18.174 P. F. Luckham, J. Klein: Forces between mica surfaces bearing adsorbed homopolymers in good solvents, J. Chem. Soc. Faraday Trans. **86** (1990) 1363–1368

18.175 J. Klein, E. Kumacheva, D. Mahalu, D. Perahia, L. J. Fetters: Reduction of frictional forces between solid surfaces bearing polymer brushes, Nature **370** (1994) 634–636

18.176 J. N. Israelachvili, H. Wennerström: Hydration or steric forces between amphiphilic surfaces?, Langmuir **6** (1990) 873–876

18.177 J. N. Israelachvili, H. Wennerström: Entropic forces between amphiphilic surfaces in liquids, J. Phys. Chem. **96** (1992) 520–531

18.178 M. K. Granfeldt, S. J. Miklavic: A simulation study of flexible zwitterionic monolayers. Interlayer interaction and headgroup conformation, J. Phys. Chem. **95** (1991) 6351–6360

18.179 L. R. Fisher, J. N. Israelachvili: Direct measurements of the effect of meniscus forces on adhesion: A study of the applicability of macroscopic thermodynamics to microscopic liquid interfaces, Colloids Surf. **3** (1981) 303–319

18.180 H. K. Christenson: Adhesion between surfaces in unsaturated vapors – a reexamination of the influence of meniscus curvature and surface forces, J. Colloid Interface Sci. **121** (1988) 170–178

18.181 R. G. Horn, J. N. Israelachvili, F. Pribac: Measurement of the deformation and adhesion of solids in contact, J. Colloid Interface Sci. **115** (1987) 480–492

18.182 H. Hertz: Über die Berührung fester elastischer Körper, J. Reine Angew. Math. **92** (1881) 156–171

18.183 H. M. Pollock, D. Maugis, M. Barquins: The force of adhesion between solid surfaces in contact, Appl. Phys. Lett. **33** (1978) 798–799

18.184 B. V. Derjaguin, V. M. Muller, Yu. P. Toporov: Effect of contact deformations on the adhesion of particles, J. Colloid Interface Sci. **53** (1975) 314–326

18.185 V. M. Muller, V. S. Yushchenko, B. V. Derjaguin: On the influence of molecular forces on the deforma-

tion of an elastic sphere and its sticking to a rigid plane, J. Colloid Interface Sci. **77** (1980) 91–101
18.186 V. M. Muller, B. V. Derjaguin, Y. P. Toporov: On 2 methods of calculation of the force of sticking of an elastic sphere to a rigid plane, Colloids Surf. **7** (1983) 251–259
18.187 D. Maugis: Adhesion of spheres: The JKR–DMT transition using a Dugdale model, J. Colloid Interface Sci. **150** (1992) 243–269
18.188 V. Mangipudi, M. Tirrell, A. V. Pocius: Direct measurement of molecular level adhesion between poly(ethylene terephthalate) and polyethylene films: Determination of surface and interfacial energies, J. Adh. Sci. Technol. **8** (1994) 1251–1270
18.189 H. K. Christenson: Surface deformations in direct force measurements, Langmuir **12** (1996) 1404–1405
18.190 M. Barquins, D. Maugis: Fracture mechanics and the adherence of viscoelastic bodies, J. Phys. D: Appl Phys. **11** (1978) 1989–2023
18.191 J. A. Greenwood, J. B. P. Williamson: Contact of nominally flat surfaces, Proc. R. Soc. London A **295** (1966) 300–319
18.192 B. N. J. Persson: Elastoplastic contact between randomly rough surfaces, Phys. Rev. Lett. **87** (2001) 116101/1–4
18.193 K. N. G. Fuller, D. Tabor: The effect of surface roughness on the adhesion of elastic solids, Proc. R. Soc. London A **345** (1975) 327–342
18.194 D. Maugis: On the contact and adhesion of rough surfaces, J. Adh. Sci. Technol. **10** (1996) 161–175
18.195 B. N. J. Persson, E. Tosatti: The effect of surface roughness on the adhesion of elastic solids, J. Chem. Phys. **115** (2001) 5597–5610
18.196 P. M. McGuiggan, J. N. Israelachvili: Adhesion and short-range forces between surfaces. Part II: Effects of surface lattice mismatch, J. Mater. Res. **5** (1990) 2232–2243
18.197 P. M. McGuiggan, J. Israelachvili: Measurements of the effects of angular lattice mismatch on the adhesion energy between two mica surfaces in water, Mat. Res. Soc. Symp. Proc. **138** (1989) 349–360
18.198 C. L. Rhykerd, Jr., M. Schoen, D. J. Diestler, J. H. Cushman: Epitaxy in simple classical fluids in micropores and near-solid surfaces, Nature **330** (1987) 461–463
18.199 M. Schoen, D. J. Diestler, J. H. Cushman: Fluids in micropores. I. Structure of a simple classical fluid in a slit-pore, J. Chem. Phys. **87** (1987) 5464–5476
18.200 M. Schoen, C. L. Rhykerd, Jr., D. J. Diestler, J. H. Cushman: Shear forces in molecularly thin films, Science **245** (1989) 1223–1225
18.201 P. A. Thompson, M. O. Robbins: Origin of stick–slip motion in boundary lubrication, Science **250** (1990) 792–794
18.202 K. K. Han, J. H. Cushman, D. J. Diestler: Grand canonical Monte Carlo simulations of a Stockmayer fluid in a slit micropore, Mol. Phys. **79** (1993) 537–545
18.203 M. Ruths, S. Granick: Influence of alignment of crystalline confining surfaces on static forces and shear in a liquid crystal, 4′-*n*-pentyl-4-cyanobiphenyl (5CB), Langmuir **16** (2000) 8368–8376
18.204 A. D. Berman: Dynamics of molecules at surfaces. Ph.D. Thesis (Univ. of California, Santa Barbara 1996)
18.205 E. Barthel, S. Roux: Velocity dependent adherence: An analytical approach for the JKR and DMT models, Langmuir **16** (2000) 8134–8138
18.206 M. Ruths, S. Granick: Rate-dependent adhesion between polymer and surfactant monolayers on elastic substrates, Langmuir **14** (1998) 1804–1814
18.207 H. Yoshizawa, Y. L. Chen, J. Israelachvili: Fundamental mechanisms of interfacial friction. 1: Relation between adhesion and friction, J. Phys. Chem. **97** (1993) 4128–4140
18.208 N. Maeda, N. Chen, M. Tirrell, J. N. Israelachvili: Adhesion and friction mechanisms of polymer-on-polymer surfaces, Science **297** (2002) 379–382
18.209 F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids* (Clarendon, London 1971)
18.210 J. A. Greenwood, K. L. Johnson: The mechanics of adhesion of viscoelastic solids, Phil. Mag. A **43** (1981) 697–711
18.211 D. Maugis: Subcritical crack growth, surface energy, fracture toughness, stick–slip and embrittlement, J. Mater. Sci. **20** (1985) 3041–3073
18.212 F. Michel, M. E. R. Shanahan: Kinetics of the JKR experiment, C. R. Acad. Sci. II (Paris) **310** (1990) 17–20
18.213 C. A. Miller, P. Neogi: *Interfacial Phenomena: Equilibrium and Dynamic Effects* (Dekker, New York 1985)
18.214 J. Israelachvili, A. Berman: Irreversibility, energy dissipation, and time effects in intermolecular and surface interactions, Israel J. Chem. **35** (1995) 85–91
18.215 A. N. Gent, A. J. Kinloch: Adhesion of viscoelastic materials to rigid substrates. III. Energy criterion for failure, J. Polym. Sci. A-2 **9** (1971) 659–668
18.216 A. N. Gent: Adhesion and strength of viscoelastic solids. Is there a relationship between adhesion and bulk properties?, Langmuir **12** (1996) 4492–4496
18.217 H. R. Brown: The adhesion between polymers, Annu. Rev. Mater. Sci. **21** (1991) 463–489
18.218 M. Deruelle, M. Tirrell, Y. Marciano, H. Hervet, L. Léger: Adhesion energy between polymer networks and solid surfaces modified by polymer attachment, Faraday Discuss. **98** (1995) 55–65
18.219 S. Granick: Motions and relaxations of confined liquids, Science **253** (1991) 1374–1379
18.220 H. W. Hu, S. Granick: Viscoelastic dynamics of confined polymer melts, Science **258** (1992) 1339–1342
18.221 M. L. Gee, P. M. McGuiggan, J. N. Israelachvili, A. M. Homola: Liquid to solidlike transitions of

molecularly thin films under shear, J. Chem. Phys. **93** (1990) 1895–1906
18.222 J. Van Alsten, S. Granick: The origin of static friction in ultrathin liquid films, Langmuir **6** (1990) 876–880
18.223 H.-W. Hu, G. A. Carson, S. Granick: Relaxation time of confined liquids under shear, Phys. Rev. Lett. **66** (1991) 2758–2761
18.224 J. Israelachvili, M. Gee, P. McGuiggan, P. Thompson, M. Robbins: Melting–freezing transitions in molecularly thin liquid films during shear, Fall Meeting of the Mater. Res. Soc., Boston, MA. 1990, ed. by J. M. Drake, J. Klafter, R. Kopelman (MRS Publications, Pittsburgh, PA1990) 3–6
18.225 J. Israelachvili, P. McGuiggan, M. Gee, A. Homola, M. Robbins, P. Thompson: Liquid dynamics in molecularly thin films, J. Phys.: Condens. Matter **2** (1990) SA89–SA98
18.226 J. Klein, E. Kumacheva: Confinement-induced phase transitions in simple liquids, Science **269** (1995) 816–819
18.227 A. L. Demirel, S. Granick: Glasslike transition of a confined simple fluid, Phys. Rev. Lett. **77** (1996) 2261–2264
18.228 D. Dowson: *History of Tribology*, 2nd edn. (Professional Engineering Publishing, London 1998)
18.229 D. Tabor: The role of surface and intermolecular forces in thin film lubrication, Tribol. Ser. **7** (1982) 651–682
18.230 M. J. Sutcliffe, S. R. Taylor, A. Cameron: Molecular asperity theory of boundary friction, Wear **51** (1978) 181–192
18.231 G. M. McClelland: Friction at weakly interacting interfaces. In: *Adhesion and Friction*, ed. by M. Grunze, H. J. Kreuzer (Springer, Berlin, Heidelberg 1989) pp. 1–16
18.232 J. N. Israelachvili, Y.-L. Chen, H. Yoshizawa: Relationship between adhesion and friction forces, J. Adh. Sci. Technol. **8** (1994) 1231–1249
18.233 D. H. Buckley: The metal-to-metal interface and its effect on adhesion and friction, J. Colloid Interface Sci. **58** (1977) 36–53
18.234 V. Zaloj, M. Urbakh, J. Klafter: Modifying friction by manipulating normal response to lateral motion, Phys. Rev. Lett. **82** (1999) 4823–4826
18.235 A. Dhinojwala, S. C. Bae, S. Granick: Shear-induced dilation of confined liquid films, Tribol. Lett. **9** (2000) 55–62
18.236 L. M. Qian, G. Luengo, E. Perez: Thermally activated lubrication with alkanes: The effect of chain length, Europhys. Lett. **61** (2003) 268–274
18.237 B. J. Briscoe, D. C. B. Evans, D. Tabor: The influence of contact pressure and saponification on the sliding behavior of steric acid monolayers, J. Colloid Interface Sci. **61** (1977) 9–13
18.238 A. Berman, J. Israelachvili: Control and minimization of friction via surface modification, NATO ASI Ser. E: Appl. Sci. **330** (1997) 317–329
18.239 M. Urbakh, L. Daikhin, J. Klafter: Dynamics of confined liquids under shear, Phys. Rev. E **51** (1995) 2137–2141
18.240 M. G. Rozman, M. Urbakh, J. Klafter: Origin of stick-slip motion in a driven two-wave potential, Phys. Rev. E **54** (1996) 6485–6494
18.241 M. G. Rozman, M. Urbakh, J. Klafter: Stick–slip dynamics as a probe of frictional forces, Europhys. Lett. **39** (1997) 183–188
18.242 B. V. Derjaguin: Molekulartheorie der äußeren Reibung, Z. Physik **88** (1934) 661–675
18.243 B. V. Derjaguin: Mechanical properties of the boundary lubrication layer, Wear **128** (1988) 19–27
18.244 B. J. Briscoe, D. C. B. Evans: The shear properties of Langmuir–Blodgett layers, Proc. R. Soc. London A **380** (1982) 389–407
18.245 A. Berman, C. Drummond, J. Israelachvili: Amontons' law at the molecular level, Tribol. Lett. **4** (1998) 95–101
18.246 P. F. Luckham, S. Manimaaran: Investigating adsorbed polymer layer behaviour using dynamic surface force apparatuses – a review, Adv. Colloid Interface Sci. **73** (1997) 1–46
18.247 A. M. Homola, J. N. Israelachvili, P. M. McGuiggan, M. L. Gee: Fundamental experimental studies in tribology: The transition from "interfacial" friction of undamaged molecularly smooth surfaces to "normal" friction with wear, Wear **136** (1990) 65–83
18.248 S. Yamada, J. Israelachvili: Friction and adhesion hysteresis of fluorocarbon surfactant monolayer-coated surfaces measured with the surface forces apparatus, J. Phys. Chem. B **102** (1998) 234–244
18.249 G. Luengo, S. E. Campbell, V. I. Srdanov, F. Wudl, J. N. Israelachvili: Direct measurement of the adhesion and friction of smooth C_{60} surfaces, Chem. Mater. **9** (1997) 1166–1171
18.250 L. Rapoport, Y. Bilik, Y. Feldman, M. Homyonfer, S. R. Cohen, R. Tenne: Hollow nanoparticles of WS_2 as potential solid-state lubricants, Nature **387** (1997) 791–793
18.251 J. Israelachvili, Y.-L. Chen, H. Yoshizawa: Relationship between adhesion and friction forces. In: *Fundamentals of Adhesion and Interfaces*, ed. by D. S. Rimai, L. P. DeMejo, K. L. Mittal (VSP, Utrecht, The Netherlands 1995) pp. 261–279
18.252 P. Berthoud, T. Baumberger, C. G'Sell, J. M. Hiver: Physical analysis of the state- and rate-dependent friction law: Static friction, Phys. Rev. B **59** (1999) 14313–14327
18.253 A. Berman, S. Steinberg, S. Campbell, A. Ulman, J. Israelachvili: Controlled microtribology of a metal oxide surface, Tribol. Lett. **4** (1998) 43–48

18.254 D. Y. C. Chan, R. G. Horn: The drainage of thin liquid films between solid surfaces, J. Chem. Phys. **83** (1985) 5311–5324

18.255 J. N. Israelachvili, S. J. Kott: Shear properties and structure of simple liquids in molecularly thin films: The transition from bulk (continuum) to molecular behavior with decreasing film thickness, J. Colloid Interface Sci. **129** (1989) 461–467

18.256 T. L. Kuhl, A. D. Berman, S. W. Hui, J. N. Israelachvili: Part 1: Direct measurement of depletion attraction and thin film viscosity between lipid bilayers in aqueous polyethylene glycol solutions, Macromolecules **31** (1998) 8250–8257

18.257 S. E. Campbell, G. Luengo, V. I. Srdanov, F. Wudl, J. N. Israelachvili: Very low viscosity at the solid–liquid interface induced by adsorbed C_{60} monolayers, Nature **382** (1996) 520–522

18.258 J. Klein, Y. Kamiyama, H. Yoshizawa, J. N. Israelachvili, G. H. Fredrickson, P. Pincus, L. J. Fetters: Lubrication forces between surfaces bearing polymer brushes, Macromolecules **26** (1993) 5552–5560

18.259 G. Luengo, J. Israelachvili, A. Dhinojwala, S. Granick: Generalized effects in confined fluids: New friction map for boundary lubrication, Wear **200** (1996) 328–335

18.260 E. Kumacheva, J. Klein: Simple liquids confined to molecularly thin layers. II. Shear and frictional behavior of solidified films, J. Chem. Phys. **108** (1998) 7010–7022

18.261 P. A. Thompson, G. S. Grest, M. O. Robbins: Phase transitions and universal dynamics in confined films, Phys. Rev. Lett. **68** (1992) 3448–3451

18.262 Y. Rabin, I. Hersht: Thin liquid layers in shear: Non-Newtonian effects, Physica A **200** (1993) 708–712

18.263 P. A. Thompson, M. O. Robbins, G. S. Grest: Structure and shear response in nanometer-thick films, Israel J. Chem. **35** (1995) 93–106

18.264 M. Urbakh, L. Daikhin, J. Klafter: Sheared liquids in the nanoscale range, J. Chem. Phys. **103** (1995) 10707–10713

18.265 A. Subbotin, A. Semenov, E. Manias, G. Hadziioannou, G. ten Brinke: Rheology of confined polymer melts under shear flow: Strong adsorption limit, Macromolecules **28** (1995) 1511–1515

18.266 A. Subbotin, A. Semenov, E. Manias, G. Hadziioannou, G. ten Brinke: Rheology of confined polymer melts under shear flow: Weak adsorption limit, Macromolecules **28** (1995) 3901–3903

18.267 J. Huh, A. Balazs: Behavior of confined telechelic chains under shear, J. Chem. Phys. **113** (2000) 2025–2031

18.268 C. Drummond, J. Elezgaray, P. Richetti: Behavior of adhesive boundary lubricated surfaces under shear: A new dynamic transition, Europhys. Lett. **58** (2002) 503–509

18.269 J. N. Israelachvili, P. M. McGuiggan, A. M. Homola: Dynamic properties of molecularly thin liquid films, Science **240** (1988) 189–191

18.270 A. M. Homola, H. V. Nguyen, G. Hadziioannou: Influence of monomer architecture on the shear properties of molecularly thin polymer melts, J. Chem. Phys. **94** (1991) 2346–2351

18.271 M. Schoen, S. Hess, D. J. Diestler: Rheological properties of confined thin films, Phys. Rev. E **52** (1995) 2587–2602

18.272 G. M. McClelland, S. R. Cohen: *Chemistry and Physics of Solid Surfaces VIII* (Springer, Berlin, Heidelberg 1990) pp. 419–445

18.273 M. O. Robbins, P. A. Thompson: Critical velocity of stick–slip motion, Science **253** (1991) 916

18.274 G. He, M. Müser, M. Robbins: Adsorbed layers and the origin of static friction, Science **284** (1999) 1650–1652

18.275 C. Drummond, J. Israelachvili: Dynamic phase transitions in confined lubricant fluids under shear, Phys. Rev. E. **63** (2001) 041506/1–11

18.276 H. Yoshizawa, J. Israelachvili: Fundamental mechanisms of interfacial friction. 2: Stick–slip friction of spherical and chain molecules, J. Phys. Chem. **97** (1993) 11300–11313

18.277 J. D. Ferry: *Viscoelastic Properties of Polymers*, 3rd edn. (Wiley, New York 1980)

18.278 A. Ruina: Slip instability and state variable friction laws, J. Geophys. Res. **88** (1983) 10359–10370

18.279 T. Baumberger, P. Berthoud, C. Caroli: Physical analysis of the state- and rate-dependent friction law. II. Dynamic friction, Phys. Rev. B **60** (1999) 3928–3939

18.280 J. Israelachvili, S. Giasson, T. Kuhl, C. Drummond, A. Berman, G. Luengo, J.-M. Pan, M. Heuberger, W. Ducker, N. Alcantar: Some fundamental differences in the adhesion and friction of rough versus smooth surfaces, Tribol. Ser. **38** (2000) 3–12

18.281 C. Drummond, J. Israelachvili: Dynamic behavior of confined branched hydrocarbon lubricant fluids under shear, Macromolecules **33** (2000) 4910–4920

18.282 E. Rabinowicz: *Friction and Wear of Materials*, 2nd edn. (Wiley, New York 1995) Chap. 4

18.283 J. Peachey, J. Van Alsten, S. Granick: Design of an apparatus to measure the shear response of ultrathin liquid films, Rev. Sci. Instrum. **62** (1991) 463–473

18.284 E. Rabinowicz: The intrinsic variables affecting the stick–slip process, Proc. Phys. Soc. **71** (1958) 668–675

18.285 T. Baumberger, F. Heslot, B. Perrin: Crossover from creep to inertial motion in friction dynamics, Nature **367** (1994) 544–546

18.286 F. Heslot, T. Baumberger, B. Perrin, B. Caroli, C. Caroli: Creep, stick–slip, and dry-friction dynamics: Experiments and a heuristic model, Phys. Rev. E **49** (1994) 4973–4988

18.287 J. Sampson, F. Morgan, D. Reed, M. Muskat: Friction behavior during the slip portion of the stick–slip process, J. Appl. Phys. **14** (1943) 689–700
18.288 F. Heymann, E. Rabinowicz, B. Rightmire: Friction apparatus for very low-speed sliding studies, Rev. Sci. Instrum. **26** (1954) 56–58
18.289 J. H. Dieterich: Time-dependent friction and the mechanics of stick–slip, Pure Appl. Geophys. **116** (1978) 790–806
18.290 J. H. Dieterich: Modeling of rock friction. 1. Experimental results and constitutive equations, J. Geophys. Res. **84** (1979) 2162–2168
18.291 G. A. Tomlinson: A molecular theory of friction, Phil. Mag. **7** (1929) 905–939
18.292 J. M. Carlson, J. S. Langer: Mechanical model of an earthquake fault, Phys. Rev. A **40** (1989) 6470–6484
18.293 B. N. J. Persson: Theory of friction: The role of elasticity in boundary lubrication, Phys. Rev. B **50** (1994) 4771–4786
18.294 S. Nasuno, A. Kudrolli, J. P. Gollub: Friction in granular layers: Hysteresis and precursors, Phys. Rev. Lett. **79** (1997) 949–952
18.295 P. Bordarier, M. Schoen, A. Fuchs: Stick–slip phase transitions in confined solidlike films from an equilibrium perspective, Phys. Rev. E **57** (1998) 1621–1635
18.296 J. M. Carlson, A. A. Batista: Constitutive relation for the friction between lubricated surfaces, Phys. Rev. E **53** (1996) 4153–4165
18.297 A. D. Berman, W. A. Ducker, J. N. Israelachvili: Origin and characterization of different stick–slip friction mechanisms, Langmuir **12** (1996) 4559–4563
18.298 A. D. Berman, W. A. Ducker, J. N. Israelachvili: Experimental and theoretical investigations of stick–slip friction mechanisms, NATO ASI Ser. E: Appl. Sci. **311** (1996) 51–67
18.299 K G. McLaren, D. Tabor: Viscoelastic properties and the friction of solids. Friction of polymers and influence of speed and temperature, Nature **197** (1963) 856–858
18.300 K. A. Grosch: Viscoelastic properties and friction of solids. Relation between friction and viscoelastic properties of rubber, Nature **197** (1963) 858–859
18.301 J. Gao, W. D. Luedtke, U. Landman: Friction control in thin-film lubrication, J. Phys. Chem. B **102** (1998) 5033–5037
18.302 M. Heuberger, C. Drummond, J. Israelachvili: Coupling of normal and transverse motions during frictional sliding, J. Phys. Chem. B **102** (1998) 5038–5041
18.303 L. Bureau, T. Baumberger, C. Caroli: Shear response of a frictional interface to a normal load modulation, Phys. Rev. E **62** (2000) 6810–6820
18.304 J. P. Gao, W. D. Luedtke, U. Landman: Structure and solvation forces in confined films: Linear and branched alkanes, J. Chem. Phys. **106** (1997) 4309–4318
18.305 J. Gao, W. D. Luedtke, U. Landman: Layering transitions and dynamics of confined liquid films, Phys. Rev. Lett. **79** (1997) 705–708
18.306 J. Warnock, D. D. Awschalom, M. W. Shafer: Orientational behavior of molecular liquids in restricted geometries, Phys. Rev. B **34** (1986) 475–478
18.307 D. D. Awschalom, J. Warnock: Supercooled liquids and solids in porous glass, Phys. Rev. B **35** (1987) 6779–6785
18.308 M. Hirano, K. Shinjo, R. Kaneko, Y. Murata: Anisotropy of frictional forces in muscovite mica, Phys. Rev. Lett. **67** (1991) 2642–2645
18.309 T. Gyalog, H. Thomas: Friction between atomically flat surfaces, Europhys. Lett. **37** (1997) 195–200

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

19. Scanning Probe Studies of Nanoscale Adhesion Between Solids in the Presence of Liquids and Monolayer Films

Adhesion between solids is a ubiquitous phenomenon whose importance is magnified at the micrometer and nanometer scales, where the surface-to-volume ratio diverges as we approach the single atom.

Numerous techniques for measuring adhesion at the atomic scale have been developed. Yet significant limitations exist. Instrumental improvements and reliable quantification are still needed.

Recent studies have highlighted the unique and important effect of liquid capillaries, particularly water, at the nanometer scale. The results demonstrate that macroscopic considerations of classic meniscus theory must be, at the very least, corrected to take into account new scaling and geometric relationships unique to the nanometer scale. More generally, a molecular-scale description of wetting and capillary condensation as it applies to nanometer-scale interfaces is clearly desirable, but remains an important challenge.

The measurement of adhesion between self-assembled monolayers has proven to be a reliable means for probing the influence of surface chemistry and local environment on adhesion. To date, however, few systems have been investigated quantitatively in detail. The molecular origins of adhesion down to the single bond level remain to be fully investigated. The most recent studies illustrate that although new information about adhesion in these systems has been revealed, further enhancements of current techniques, as well as the development of new methodologies, coupled with accurate theoretical modeling, are required to adequately tackle these complex measurements.

19.1 **The Importance of Adhesion at the Nanoscale** 605
19.2 **Techniques for Measuring Adhesion** 606
19.3 **Calibration of Forces, Displacements, and Tips** 610
19.3.1 Force Calibration 610
19.3.2 Probe Tip Characterization 611
19.3.3 Displacement Calibration 612
19.4 **The Effect of Liquid Capillaries on Adhesion** 612
19.4.1 Theoretical Background 612
19.4.2 Experimental and Theoretical Studies of Capillary Formation with Scanning Probes 614
19.4.3 Future Directions 618
19.5 **Self-Assembled Monolayers** 618
19.5.1 Adhesion at SAM Interfaces 618
19.5.2 Chemical Force Microscopy: General Methodology 619
19.5.3 Adhesion at SAM-Modified Surfaces in Liquids 620
19.5.4 Impact of Intra- and Inter-Chain Interactions on Adhesion 621
19.5.5 Adhesion at the Single-Bond Level. 622
19.5.6 Future Directions 623
19.6 **Concluding Remarks** 624
References 624

19.1 The Importance of Adhesion at the Nanoscale

The mutual attraction and bonding of two surfaces, which can occur with or without an intervening medium, is a commonly observed phenomenon with far-reaching manifestations and applications in society. The adherence between a raindrop and a window pane, the climbing of a gecko up a vine, the sticking of multiple adhesive note pads to a professor's wall, the force required to separate hook-and-eye (Velcro$^{\mathrm{TM}}$) strips, the building of a sand castle, and the book page turned by a wetted finger are all examples in which adhesion is important. Within the complexity of these examples and others lies a central theme: The mechanical forces between a pair of materials can be fundamentally affected not just by the macroscopic or microscopic structure of the surfaces, but also by the interatomic and intermolecular forces between them.

Adhesion and intermolecular forces have been studied for many years dating back to ancient times [19.1], and active research continues today of topics as broad as insect and reptile locomotion [19.2], interactions between cells in the body [19.3], and the design of self-healing composites [19.4], to name a few. While adhesion is clearly of interest for a wide range of macroscopic applications, its importance becomes dominant at the micrometer and nanometer scales. This is primarily due to the dramatic increase in the surface-to-volume ratio of materials at these scales, an effect that renders friction and interfacial wear critical phenomena [19.5]. For example, the dominating effect of adhesion at this scale has affected the development of microelectromechanical systems (MEMS), in which interfacial forces can prevent devices from functioning properly, since the small flexible parts often emerge from the fabrication line stuck together. Studies of adhesion in MEMS are ongoing, and commercially deployed MEMS devices rely upon sophisticated surface treatments to reduce and control adhesion [19.6–8].

Detailed control of adhesion at the molecular level will be essential for the proposed design of the smaller nanoelectromechanical systems (NEMS). Much has been written about the possibilities for nanoscale machines, sensors, and actuators, etc. It is crucial to understand that a molecule is *all* surface, and, therefore, molecular and nanoscale devices cannot be properly designed, implemented, or characterized unless an understanding of atomic-scale adhesion is thoroughly presented, particularly if these devices are to involve any moving parts that will come into and out of contact with each other. Studying adhesion at the nanoscale is important for other reasons. The protruding asperities in most MEMS materials often have nanoscale dimensions, and, therefore, a complete understanding of adhesion in MEMS requires the investigation of the adhesive properties of the individual nanoscale asperities. In addition, the experimental study of adhesion at the nanoscale is required for the development of detailed atomic-scale models of adhesion. Such progress requires close collaboration between experiment and theory.

There has been significant progress in the experimental study of adhesion at the nanometer scale using scanning probe methods, but numerous challenges exist. A discussion of solid-solid adhesion without an intervening medium is provided in part B and C of this book. This chapter focuses on how adhesion is affected by the ubiquitous presence of water, and how it can be controlled through the application of self-assembled monolayer (SAM) coatings, again in the presence of a liquid medium. In addition, we discuss specific instrumental challenges that are inherent to adhesion measurements. We do not delve into the realm of atomic-scale modeling of adhesion, nor do we discuss the role of more complex coatings such as polymer brushes and blends. Rather, we will focus on critically evaluating the relevant experimental techniques and critically reviewing recent results from studies of water and SAM films, which are perhaps the two most commonly encountered media in nanoscale adhesion applications.

19.2 Techniques for Measuring Adhesion

The experimental study of adhesion at the nanoscale has experienced two renaisances in this century. The first occurred with the development of the surface forces apparatus (SFA) [19.9, 10], and the second with the later development of the AFM [19.11] and other related scanning force techniques.

SFA experiments have contributed profoundly to our understanding of adhesion. The SFA consists of a pair of atomically smooth surfaces, usually mica sheets, mounted on crossed cylinders that can be pressed together to form a controlled circular contact. The applied load, normal displacement, surface separation, contact area, and shear force (if applied) can all be controlled and/or measured [19.9, 12–14]. The SFA can be operated in air, a controlled environment, or under liquid conditions. The surfaces are often treated to attach molecules whose behavior under confinement can be studied. Alternately, the behavior of a confined fluid layer can be observed. The surface separation can be measured and controlled in the Ångstrom (Å) regime. The lateral resolution is limited to the range of several tens of micrometers. However, the true contact area between the interfaces can be directly measured, which is a key advantage since it allows an adhesive force to be converted to a force (or energy) per unit area (or per molecule), thus separating geometrical contributions to adhesion from chemical contributions. In this chapter, we will refer to relevant results in the context of the scanning probe experiments. As a very brief summary, some of the most important results pertaining to adhesion include: the observation of capillary effects on adhesion [19.15, 16], the presence of hysteresis in adhesion due to pressure-induced restructuring of the interface [19.17], and numerous studies of how interfacial chemistry (hydrophilicity, surface charge, specific chemical groups, etc.) affect adhesive forces [19.1, 18, 19].

AFM instrumentation is described in part B of this book. Key ways in which the AFM differs from the SFA are: (1) the lateral contact size is nm, not µm, due to the fact that the tip is usually < 100 nm in radius, and the contact area at low applied loads will be a fraction of this radius [19.20]; (2) the force resolution is as good as 10^{-10} N or better, as opposed to $\sim 10^{-6}$ N with the SFA; (3) the contact area is not directly observable, although it may be derived through other means; (4) the actual separation between the tip and sample are not directly observable, which is a key disadvantage; (5) the relative separation between the sample and the *cantilever* (not the tip) is controlled in the 0.1 nm range or better; (6) the measurement bandwidth is typically in the kHz regime, but can extend into the MHz regime depending on the data acquisition technique; (7) the operating environment includes ultrahigh vacuum (UHV) and cryogenic to elevated temperatures; (8) there is virtually no limit to the range of materials that can be probed by the AFM; and (9) half of the interface (the tip) is essentially unknown or uncontrolled, whereas with the SFA, both surfaces are well-defined, which is another key challenge for AFM that has yet to be consistently addressed.

The general setup of the AFM is as follows. A small sharp tip (with a radius typically between 10–100 nm) is attached to the end of a compliant cantilever (see Fig. 19.1). The tip is brought into close proximity with a sample surface in a fashion that is identical to STM tip–sample approach mechanisms. Forces acting between the AFM tip and the sample will result in deflections of the cantilever (Fig. 19.2). The cantilever bends vertically (i. e., toward the sample) in response to attractive and/or repulsive forces acting on the tip. The deflection of the cantilever from its equilibrium position is proportional to the normal load applied to the tip by the cantilever. Lateral forces result in a twisting of the cantilever from its equilibrium position. These measurements can be performed in a variety of environments: ambient air, controlled atmosphere, liquids [19.21], or UHV [19.22–24].

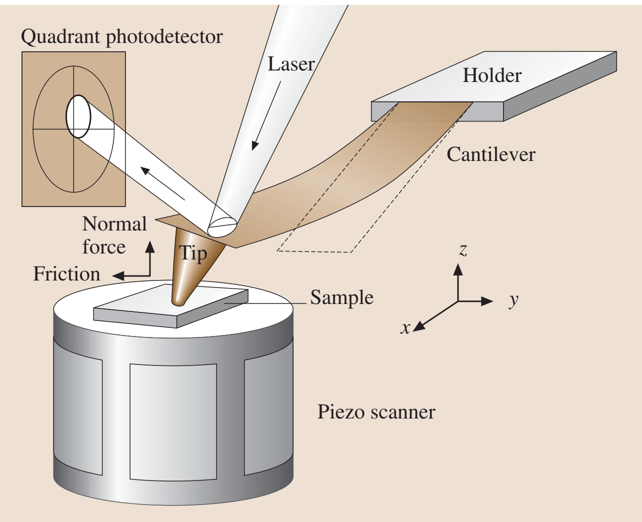

Fig. 19.1 Diagram of the AFM setup for the optical beam deflection method. The tip is in contact with a sample surface. A laser beam is focused on the back of the cantilever and reflects into a four-quadrant photodetector. Normal forces deflect the cantilever up or down, lateral forces twist the cantilever left and right. These deflections are simultaneously and independently measured by monitoring the deflection of the reflected laser beam

The simplest way (perhaps deceptively so) to measure adhesion with the AFM is through so-called force–displacement plots (also referred to in the literature as force curves and force–distance curves or plots). A force-displacement plot displays the vertical cantilever bending (i. e., the cantilever deflection, which can be calibrated to give the tip–sample interaction force) vs. lever–sample displacement. This displacement is measured between the sample and the rigidly held back end of the cantilever (as opposed to the front end with the tip, which will bend in response to interaction forces). This displacement is altered by varying the vertical position of the piezo tube, which, depending on the type of AFM, displaces the tip or the sample in the direction normal to the sample. Referring to Fig. 19.2a, the stages of acquisition of the force curve are as follows: (a) The lever and sample are initially far apart and no forces act. (b) As the lever is brought close to the sample, the tip senses attractive forces that cause the tip end of the lever to bend downward, thus signifying a negative (attractive) force. These forces may be of electrostatic, van der

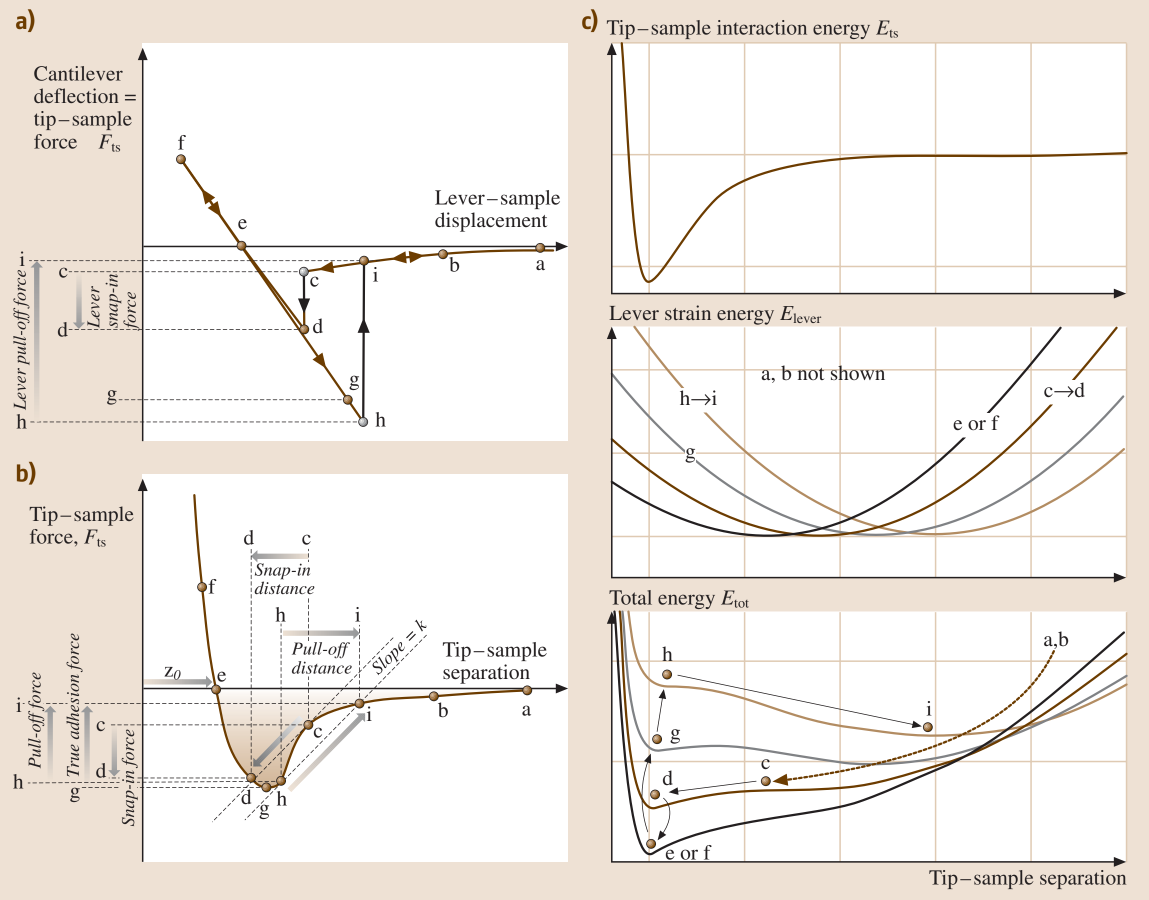

Fig. 19.2 **(a)** A "force-displacement" curve displays the vertical cantilever deflection vs. lever–sample displacement. This displacement is measured between the sample and the rigidly held back end of the cantilever (as opposed to the front end with the tip that will bend in response to interaction forces). **(b)** The true tip–sample interaction as a function of tip–sample separation. **(c)** Tip–sample interaction energy, lever strain energy, and total energy as a function of tip–sample separation

Waals, or other origin. (c) At this point, the attractive force gradient (slope) of the tip-sample interaction force exceeds the spring constant k of the lever, and this instability causes the tip to snap into contact (d) with the sample. The lever-sample displacement can continue to be reduced and eventually crosses the force axis (e), which corresponds to zero externally applied load. As this tip is in repulsive contact with the sample, the front end of the lever is pushed further and further upward, and the force corresponds to the externally applied load. (f) The displacement direction is reversed at a point chosen by the user. As the lever-sample distance is reduced, the force becomes negative. Adhesion between the tip and sample maintains the contact, although there is now a net negative (tensile) load. Eventually the tip passes through the point of maximum adhesion (g), which is not distinguishable on this plot (see Fig. 19.2b). Finally, an unstable point is reached at which the cantilever stiffness exceeds the force gradient of the tip-sample interaction, and the tip snaps out of contact with the sample (h–i). The resulting force relaxation is usually called the pull-off force. Note that the forces and distances are not drawn to scale, and, in particular, the attractive part of the interaction is exaggerated beyond that which often occurs for inert, neutral surfaces.

The force-displacement curve can be measured at a single point, or a series of measurements can be carried out over an area of interest. The so-called adhesion mapping techniques allow for spatially resolved adhesion measurements to be correlated with other sample properties such as friction, chemical termination, and other types of material heterogeneity.

In order to properly derive adhesion measurements from the AFM, it is critical to understand the nature of the mechanical instability of both the snap-in and pull-off processes. The fundamental point to comprehend is that *the pull-off force is* not *the adhesive bond strength*. In other words, *it is not a direct measure of the actual adhesive forces that were acting between the tip and sample in the absence of applied load*. This important point is often overlooked or misunderstood, and so we discuss it here in some detail.

A cartoon of the interaction force (F_{ts}) between the tip and sample is sketched (not to any particular scale) as a function of the true tip-sample separation in Fig. 19.2b. Superimposed on this is the force-distance relation of the cantilever: a straight line with slope k (N/m), the cantilever stiffness (diagonal dashed lines). Points of instability are shown as gray dots, labels, and arrows, as opposed to stable points, shown as black dots, labels, and arrows. Snap-in occurs when the attractive force gradient $\mathrm{d}F_{\text{ts}}/\mathrm{d}z$ just exceeds k, in other words, when the dashed line is tangent to the tip-sample force curve, as shown for point c. This instability is a direct consequence of Newton's 2^{nd} law and is explained further below. Similarly, the pull-off occurs during retraction when k finally just exceeds $\mathrm{d}F_{\text{ts}}/\mathrm{d}z$, as shown for point h, where once again, the diagonal line is just tangent to the tip-sample force curve. By definition, since k is finite, this point *cannot* correspond to the point of maximum attractive force (or adhesive force), g. Thus, the force at which pull-off occurs is *not* precisely equal to the true adhesion force. The significance of this deviation is discussed below.

This can also be illustrated from an energetic perspective. In Fig. 19.2c, we show separately the tip-sample interaction energy, E_{ts}, as a function of tip-sample separation (top), followed by the quadratic elastic strain energy, E_{lever}, of the cantilever (middle) and the sum, E_{total}, of the two (bottom). As the lever-sample displacement is varied, the elastic energy curve is shifted to the appropriate position, which, for this example, is represented on the tip-sample separation axis. Far away from the sample, the system resides in a deep minimum, points a and b (not shown). As the tip-sample separation is reduced, this minimum becomes more and more shallow (evolving along the dashed arrow). At the snap-in point c, the minimum is eliminated, since the attractive energy of the tip-sample interaction has overwhelmed the energy minimum of the cantilever. Mathematically this is described as an inflection point, where:

$$\frac{\mathrm{d}^2 E_{\text{tot}}}{\mathrm{d}z^2} = 0 \, .$$

Now,

$$E_{\text{tot}} = E_{\text{ts}} + E_{\text{lever}} = E_{\text{ts}} + \tfrac{1}{2}k(z - z_0)^2 \qquad (19.1)$$

$\therefore$ snap-in occurs when

$$\frac{\mathrm{d}^2 E_{\text{ts}}}{\mathrm{d}z^2} = -\frac{\mathrm{d}F_{\text{ts}}}{\mathrm{d}z} = -k \, .$$

This explains why the instability is described by the line of slope k being tangent to the tip-sample force curve, as shown in Fig. 19.2b. The system now rapidly finds the new minimum by following the gray arrow to point d, where a mechanical contact between the tip and sample is now formed.

Note that once the contact is formed, the tip and sample elastically deform, and so, strictly speaking, the tip-sample potential will be distinct from the potential shown before contact occurs. For simplicity, we have left out this change and instead have drawn the potential as a single-valued function. However, such changes in the energy landscape should be considered in a complete description of the problem [19.25].

Nevertheless, it will also hold true that upon retracting the tip from the surface, the stable minimum seen at e, f, and then g becomes more shallow, and eventually disappears as it is overwhelmed by the strain energy of the cantilever. Again, an inflection point is created when

$$\frac{\mathrm{d}^2 E_{\text{attr}}}{\mathrm{d}z^2} = -k \, , \qquad (19.2)$$

in other words, when again

$$\frac{\mathrm{d}F_{\text{ts}}}{\mathrm{d}z} = k \, . \qquad (19.3)$$

This second instability occurs at point h, the pull-off point. The system then follows the gray arrow to point i, whereby the tip is now out of contact with the sample.

If the spring is sufficiently compliant (low k), or the potential sufficiently stiff (large $\mathrm{d}F/\mathrm{d}z$ shortly past the minimum at g), then the pull-off force *does* nearly correspond to the attractive force minimum F_{min}. In other words, point h would occur very close to point g. But if the cantilever is somewhat stiff, or if the potential is rather compliant (which may be the case for organic, polymeric, biological, or liquid systems), then the pull-off force may differ substantially from F_{min}. For such cases, this distinction between pull-off force and true adhesion is important, and the limitations imposed by the AFM's intrinsic instabilities become apparent. Of course, if a cantilever is used whose stiffness exceeds the attractive force gradient at all points, one will avoid the instabilities. But since it is the deflection of the cantilever that is used to sense the force, one would have to trade force sensitivity for stability, which is usually an unsuitable compromise.

When a compliant spring or holder is used to manipulate a probe, as is the case for AFM, the technique is generally referred to as being "load-controlled", since the load can be prescribed, but the actual displacement of the probe with respect to the sample cannot (as illustrated by the jump in displacement that occurs during snap-in or pull-off).

In contrast, "displacement-controlled" techniques avoid this instability by effectively eliminating the compliance of the spring or holder. This has been carried out for decades in the mechanical testing community, and over the past 10 years, displacement-controlled scanning probe, have been developed. This is accomplished by displacing the tip by applying force to the tip itself. *Pethica* et al. [19.26,27] use a magnetic coating on the tip and external coils to apply forces to the tip. They refer to the instrument as a "force-controlled microscope". *Houston* et al. [19.28,29] control the force electrostatically and refer to the instrument as an "interfacial force microscope" (IFM). *Lieber* et al. use a variation on Pethica's method, whereby a magnetic coil is used to apply a force to the cantilever [19.30], and adapted it to work in solution.

An example of an adhesion measurement with the IFM is shown in [19.31]. which shows that the instrument is able to measure the entire interaction force curve without instabilities. Thus, this provides a *direct* measurement of the minimum interaction force (as well as forces at all other tip-sample separations) and is a more reliable measure of adhesion. The disadvantage with these techniques is primarily one of inconvenience: The probes require extra manufacturing steps and control electronics. However, given the importance of adhesion in nanoscale science and technology, the extra information gained makes these techniques clearly worth the effort.

Pull-off instabilities can occur even in a displacement-controlled experiment. In this case, the instability is an "intrinsic" instability, whereby the adhesive force gradient competes with the stiffness of the *contact itself* [19.32, 33]. Adhesive materials with low stiffness, such as polymers, may show this behavior.

Finally, the pull-off force may show a time dependence that arises from intrinsic viscoelasticity of the tip or sample materials [19.34], or kinetic effects due to adsorption or reaction of materials at the interface [19.35]. These regimes remain relatively unexplored, but certainly, to compare adhesion measurements between labs, the velocity of approach and retraction, as well as the time in contact ought to be reported for any published experimental results.

A quantitative and reliable examination of adhesion, therefore, requires careful consideration of the mechanics of the contact and the cantilever. For an AFM experiment in which instabilities occur, one can conclude that the pull-off force is a good measure of adhesion only if

1. the materials are fully elastic with little or no viscoelastic character,
2. the interface is chemically stable,
3. the cantilever stiffness is sufficiently low compared with the adhesive force gradient, and
4. the contact stiffness is sufficiently high compared with the adhesive force gradient. Otherwise, a more thorough investigation is required.

19.3 Calibration of Forces, Displacements, and Tips

Whether the forces are to be measured with load-controlled or displacement-controlled techniques, measurements cannot be compared between laboratories unless the forces are properly calibrated. Unfortunately, this can be a rather involved task, and adoption of standards has yet to become widespread and robust. Here we provide a summary of the pertinent issues with references to other work for further reading; these issues are also discussed in part B and C of this book.

19.3.1 Force Calibration

Commercially available AFM cantilevers often come in two forms: V-shaped and rectangular. Silicon and silicon nitride are the typical materials, and reflective coatings are often applied to enhance the reflectivity of the laser beam. The normal force constant of a monolithic rectangular cantilever beam is well-defined [19.36], but it relies on knowledge of all lever dimensions and its modulus. For a rectangular beam of length L, width w, thickness t, Young's modulus E, the stiffness k is given by:

$$k = \frac{Ewt^3}{4L^3} \,. \tag{19.4}$$

The cubic dependence on thickness is particularly problematic since in microfabrication processes t is usually determined by an etching process that is not precisely controlled, and is, therefore, difficult and cumbersome to measure experimentally. Variations in E can also occur depending on the type of cantilever, particularly for silicon nitride cantilevers, although the dependence on E is clearly not as critical. If a bulk value for E is used, the thickness can be determined by measuring the resonance frequency of the cantilever [19.37]. However, this method only works for uncoated monolithic cantilevers, since the metal coatings applied to enhance reflectivity, or for other purposes, will also alter substantially the spring constant [19.38]. In addition, formulae for V-shaped levers are substantially more complicated. Uncoated, single-crystal silicon cantilevers are perhaps the only ones for which the force constant can be reliably determined by using (19.4) and the resonance frequency [19.39]. Otherwise, an experimental, preferably in situ calibration method is strongly encouraged.

Experimental methods to calibrate the force constant of AFM cantilevers have been discussed extensively [19.37, 40–50]. It has been shown repeatedly that the manufacturer's quoted spring constants can be in error by large factors and should simply not be used in any quantitative research effort. We will not delve into the details of these calibration methods, but we do make note of one particularly recent method proposed by *Sader* et al. [19.51] that appears to be reliable and simple to carry out for rectangular levers. It relies on measuring the resonance frequency and the quality factor of the cantilever in air. Use of the hydrodynamic function relates the damping of air to the quality factor, and the dependence of E and t is eliminated from the resulting formula for the force constant.

19.3.2 Probe Tip Characterization

A problem of quite a different nature is that the geometry of the contact formed between the AFM tip and sample surface is not defined if the tip shape and composition is not known. This issue is of crucial importance since one is trying to understand the properties of an interface, and the tip is half of that interface.

The adhesion force between the tip and sample is a meaningless quantity in the absence of any knowledge of the tip shape [19.20]. The only use for such measurements is in cases where the *same* tip is used to compare different samples or different conditions, and verification by cyclic repetition of the experiments is carried out to ensure that the tip itself did not change during the experiment.

There are several in situ methods to characterize the tip shape. A topographic AFM image is actually a convolution of the tip and sample geometry [19.52]. Separation of the tip and sample contributions by contact imaging of known, or at least sharp, sample features allows determination of the tip shape to a significant extent [19.53–64]. Ex-situ tip imaging by transmission electron microscopy has also been performed [19.61, 65]. Some of these measurements have revealed that a *majority* of microfabricated cantilevers possess double tips and other unsuitable tip structures [19.56, 60, 64]. This convincingly demonstrates that tip characterization is absolutely necessary for useful nanotribological measurements with AFM. Thin film coatings applied to microfabricated levers can provide robust, smooth, and even conductive coatings [19.65–67]. Further work in this direction would be useful so as to provide a wider array of dependable tip structures and materials.

In addition to the shape of the tip, its chemical composition is equally important, but is also challenging to determine or control. *Xiao* et al. have shown that the AFM tip is readily chemically modified when scanned in contact with various materials [19.68], even tips that have been coated with self-assembled monolayers in order to control their chemistry. They recommend "running in" the tip with a standard sample to give reproducible results. The stresses that take place in a nano-contact can be very large [19.5], and so modification of both the chemistry and structure of the tip is important to consider.

One class of experiments for which the tip shape and chemistry is not as critical is when a molecule or nanostructure is tethered to the end of the tip and specific interactions are probed [19.69–73]. Another alternative method is to use colloid probes [19.74,75], whereby colloidal particles are attached to the cantilever on top of (or in place of) the tip. This method requires a unique calibration procedure [19.76], but provides particles whose structure and chemistry can be measured and perhaps controlled prior to attachment to the lever.

19.3.3 Displacement Calibration

Proper signal and spatial calibration also requires knowing the sensitivity of the piezoelectric scanning elements. This can involve complications due to instrumental drift [19.77] and inherent piezoelectric effects, namely nonlinearity, hysteresis, creep, and variations of sensitivity with applied voltage [19.78–82]. Caution must be exercised when determining and relying upon these parameters. Techniques such as laser interferometry [19.83], scanning sloped samples [19.45, 78], scanning known surface step heights [19.84], or the use of pre-calibrated piezoelectrics [19.85] can facilitate piezo calibration.

19.4 The Effect of Liquid Capillaries on Adhesion

The adsorption of water and other liquids onto surfaces, and their subsequent behavior at interfaces, continue to be a vibrant area of research. The importance of liquid-solid interface behavior is massive, encompassing topics as broad as paints, textiles, lubricants, geology, and environmental chemistry, and covering all corners of biology.

19.4.1 Theoretical Background

The ability to measure forces at the nanoscale using scanning probe microscopy has generated much interest in these fields. For any force measurement carried out in ambient laboratory conditions, one must consider the possibility of a capillary neck forming between the tip and sample. The study of such necks may, in turn, provide insight into the behavior of the liquid, which is discussed in other chapters in this book. In order to provide background for these emerging areas, we consider here the fundamental mechanical and chemical aspects of adhesion in the presence of a liquid meniscus. By way of introduction, *Israelachvili* [19.17] and *de Gennes* [19.87] provide rigorous coverage of the terminology, physics, and chemistry of liquid films and their wetting properties.

Water readily adsorbs at many surfaces. At a crack or sharp corner it can condense to form a meniscus if its contact angle is small enough. The small gap between an AFM tip and a surface is, therefore, an ideal occasion for such condensation. Early on, it was realized that liquid condensation plays a significant role in tip-sample interactions [19.88, 89].

The classic theory of capillary condensation starts by considering the thermodynamics of capillary formation. The general geometry of the capillary is shown in Fig. 19.3, adapted from [19.86], with hydrophobic surfaces sketched on the left half ($r_1 > 0$) and hydrophilic surfaces on the right half ($r_1 < 0$). The tip is idealized as a sphere of radius R. The water contact angles with the sample and tip respectively are θ_1 and θ_2. D represents the separation of the tip and sample. The angle ϕ is referred to as the "filling angle". The pressure difference, or Laplace pressure, across a curved interface is given by the Young–Laplace equation [19.17, 86]:

$$\Delta p = \gamma \left(\frac{1}{r_1} + \frac{1}{r_2} \right) , \qquad (19.5)$$

where r_1 and r_2 are defined in Fig. 19.3 and γ is the surface tension of the liquid. The resulting force is attractive if $\Delta p < 0$. Note that $r_1 > 0$ and $r_2 < 0$. If the capillary formation is isothermal, then one can derive the Kelvin equation:

$$RT \ln \frac{p}{p_0} = \gamma V \left(\frac{1}{r_1} + \frac{1}{r_2} \right) , \qquad (19.6)$$

where p_0 is the saturation pressure of the liquid, V is the molar volume of the liquid, T is the temperature, and R the molar gas constant. The ratio p/p_0 simply corresponds to the relative vapor pressure of the liquid,

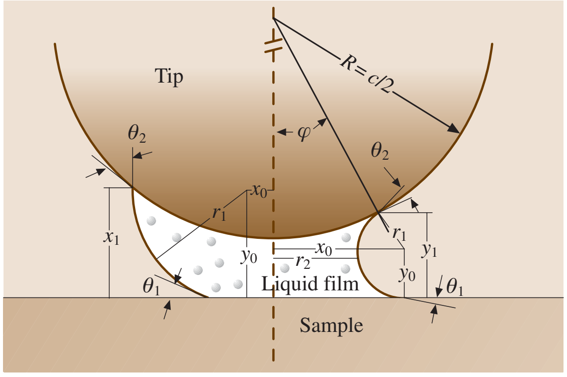

Fig. 19.3 The AFM tip considered as a sphere at a distance D from the sample. The liquid film in between may form a concave (*right*) or convex (*left*) shaped meniscus in the plane shown (adapted after [19.86])

which in the case of water is just the relative humidity (RH). This is often rewritten in terms of the Kelvin radius r_k:

$$r_k^{-1} = \left(\frac{1}{r_1} + \frac{1}{r_2}\right) = \frac{R_\mathrm{M}T}{\gamma V} \ln \frac{p}{p_0} \,. \tag{19.7}$$

The Kelvin radius varies logarithmically with the partial pressure of the liquid and is by definition less than zero (negative values correspond to convex curvature). For water at 20 °C, $\gamma V/RT = 0.54\,\mathrm{nm}$. A graph of r_k vs. p/p_0 is shown in Fig. 19.4. Of particular note is the fact that starting from $p/p_0 = 0.35$ and higher, we find $|r_\mathrm{k}| < 2\,\mathrm{nm}$. The application of this classic theory to capillaries with such small curvature radii is certainly a question to be considered carefully, and we will address it below.

A simple equation for the maximum attractive capillary force F_c between the tip and sample that is commonly used is:

$$F_\mathrm{c} = -4\pi\gamma R\cos\theta \,. \tag{19.8}$$

This equation is calculated by considering the Laplace pressure only. The adhesion force is predicted to be independent of RH. Notice that an immediate problem arises, which is that it predicts a finite force even at 0% RH, where there can be no capillary formation. Derivations of (19.8) are presented in several other publications [19.17, 86, 90, 91]. Significantly, this equation and its use with AFM experiments contain several assumptions, *many of which may not be valid for an AFM experiment*:

1. $|r_1| \ll R$, which is equivalent to saying ϕ is small,
2. $|r_1| \ll |r_2|$,
3. $\theta_1 = \theta_2$,
4. $D \ll |r_1|, |r_2|$,
5. the tip is shaped like a perfect sphere,
6. the effect of solid-solid adhesion is negligible with respect to the meniscus force,
7. the meniscus cross sections are perfect, circular arcs,
8. the force from the Laplace pressure dominates the force due to the resolved surface tension of the meniscus,
9. the surface tension γ is independent of the meniscus size,
10. the meniscus volume remains constant as the tip is retracted,
11. the maximum force of attraction is equal to the pull-off force,
12. the tip and sample are perfectly rigid.

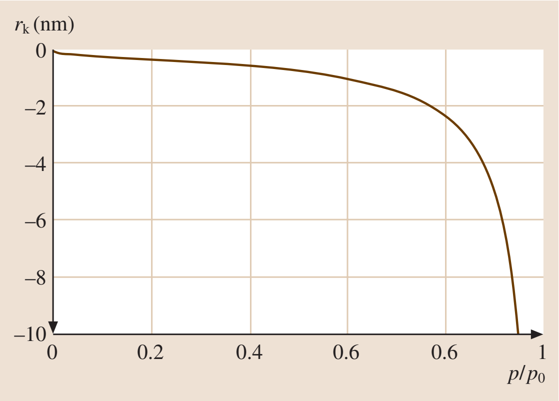

Fig. 19.4 The Kelvin radius of water at 20 °C plotted as a function of the relative humidity

Assumption 1 may not be true, since the tip radius may indeed be small and comparable to the meniscus curvature radii. Assumption 2 may not be true since the small tip geometry may lead the two meniscus curvature radii to be similar. Assumption 3 is by no means true if the tip and sample are made of different materials. Assumption 4 may not be true since both the separation of the tip and sample, as well as the radii r_1 and r_2 may be in the nanometer range. Assumption 5 may be slightly or grossly in error and is a particularly dubious assumption in the absence of tip characterization. In addition, for large menisci, the capillary will grow beyond the end of the tip and start climbing up its shank, which may be pyramidal or conical in shape. Assumption 6 may also be inaccurate if van der Waals or other adhesive forces are significant (this is discussed further in the next section). Assumption 7 is not correct [19.90], and the regimes where it is a reasonable approximation require close scrutiny. Assumption 8 will be inaccurate at high relative vapor pressures [19.17, 91]. The nature of assumptions 1 to 8 and corrections to the theory to account for their violation are presented in the work of *Orr* and *Scriven* [19.90]. This theory remains within the bounds of the classical picture of capillary formation.

Assumption 9 concerns an important scientific question that has not been fully resolved and represents the possible violation of the classical framework of molecular effects at the nanoscale. SFA measurements by *Israelachvili* have indicated that for cyclohexane and other inert organic liquids, γ remains nearly equal to its macroscopic value even for Kelvin radii that, remarkably, correspond to one or two molecular diameters. However, for water, the adhesion force comes to within

10% of the bulk prediction only when $p/p_0 > 0.9$, which corresponds to a Kelvin radius greater than ~ 5 nm in magnitude [19.16]. This deviation from macroscopic thermodynamic predictions alone calls the use of (19.8) into serious question for AFM measurements. Rather, the exploration of this deviation at the molecular scale presents a unique opportunity for scanning probe measurements.

Assumptions 10 and 11 are not assumptions of (19.8) itself, but rather are assumptions that are often used when applying (19.8) to AFM measurements. Equation (19.8) simply gives the maximum force of attraction between the tip and sample. As discussed above, an AFM does *not* measure this quantity. Rather, it measures the force at which an instability occurs. If a capillary has formed between the tip and sample, then the force as a function of distance can be calculated. Calculating this force requires making one of two assumptions: Either the volume of the capillary is conserved (i. e., the rate of displacement is large with respect to the adsorption or desorption kinetics of the liquid), or the Kelvin radius is conserved (i. e., the rate of displacement is slow with respect to the adsorption or desorption kinetics of the liquid, and so the capillary remains in equilibrium). The constant volume assumption 10 has been used in every paper we have reviewed. *Israelachvili*, however, pointed out the difference between these two approaches in his book [19.17] and left the solution of the problem as an exercise to the reader! As with the problem of scale-dependent surface tension mentioned above, the kinetics of capillary formation and dissolution is a relatively unexplored problem and is therefore worthy of further investigation. A recent study of the humidity dependence of friction as a function of sliding speed is an example where this issue is raised [19.92].

Once an assumption about how the meniscus changes with displacement has been made, one still needs to consider the nature of the instability in order to relate the AFM pull-off measurement to the capillary's properties. As stated above and shown in Fig. 19.2, a low lever stiffness k or a strongly varying adhesive force will lead to a pull-off force that is nearly equal to the adhesive force, and so assumption 11 would be valid. However, if k is sufficiently large, or the capillary stiffness sufficiently weak, this assumption will fail. As we shall see below, experimental efforts to investigate this point, and theoretical work to predict it, have yet to be carried out.

Finally, assumption 12, if violated, requires a substantially more complex analysis. The question has been addressed independently by *Maugis* [19.93] and *Fogden* and *White* [19.94]. Both papers provide a nondimensional parameter that allows one to determine the severity of the effect. In the limit of small tips, stiff materials, large (in magnitude) Kelvin radii, and low surface tensions, the effect of elastic deformation is negligible. However, for relatively compliant materials, large tips, and small Kelvin radii, the meniscus can appreciably deform the contact in the immediate vicinity of the meniscus. This can substantially alter the mechanics of adhesion, as well as significantly affect the stresses. The dependence on the Kelvin radius is particularly critical. This effect may be of particular concern with soft materials like polymers or biological specimens. According to *Maugis*, the problem is analogous to the adhesive contact problem for solids, discussed by *Johnson* et al. [19.95] and studied further in many papers since [19.25, 96, 97].

19.4.2 Experimental and Theoretical Studies of Capillary Formation with Scanning Probes

There have been several experimental and theoretical investigations of how pull-off forces are affected by liquid capillaries. These studies have mostly focused on the effect of relative humidity. We do not present an exhaustive review here, but rather summarize a few key results that highlight the important trends observed and the outstanding questions that remain.

Early on, it was realized that water capillary formation occurred readily if at least one of the two surfaces in contact was hydrophilic. Higher adhesion will lead to higher contact forces and therefore larger elastic contact areas, and this can degrade the lateral spatial resolution of the AFM, as observed by *Thundat* et al. [19.98]. Furthermore, several observations have confirmed the expected result that capillary formation was readily prevented for hydrophobic surfaces. For example, *Binggeli* and *Mate* [19.88, 99] showed that tungsten tips in contact with clean silicon wafers with a hydrophilic native oxide exhibited strong adhesion and long pull-off lengths, whereas surfaces treated with a hydrophobic perfluoropolyether (such as Z-dol) showed pull-off forces reduced by a factor of 2 to 3. These results were confirmed by *Bhushan* and *Sundararajan* [19.100] who investigated the pull-off force between silicon nitride tips and Si(100) with and without Z-dol coatings for a range of relative humidities. As another example, MoO_3 films, likely hydrophilic, showed a sixfold increase in pull-off force measured with silicon tips from 0 to 50% RH [19.101].

While these examples illustrate some basic trends, the need for more finely resolved measurements as

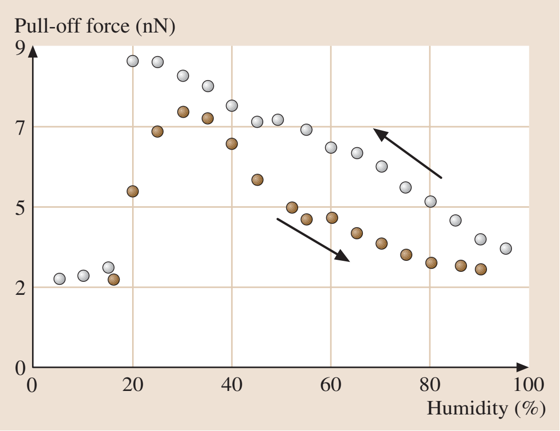

Fig. 19.5 Pull-off force between a silicon nitride tip and the muscovite mica surface as a function of RH. Two sets of data are shown: increasing (*grey circles*) and decreasing (*brown circles*) humidity (after [19.35])

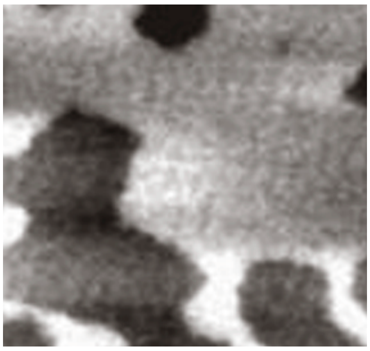

Fig. 19.6 Scanning polarization for microscopy image of water structures on the muscovite mica surface. A degree of polygonal shape to the boundaries is seen. The signal represents two distinct phases of water that are present as a molecular film below 45% RH (after [19.35])

a function of humidity has been addressed only quite recently. For example, *Xu* et al. measured adhesion between a silicon nitride tip and (hydrophilic) muscovite mica in ∼ 5% RH increments. The result is shown in Fig. 19.5. There are three identifiable regions: constant adhesion at low RH (< 20%), increasing adhesion, and then decreasing adhesion. Some hysteresis is seen between experiments conducted with increasing and decreasing RH, but the overall trend is preserved. *Xu* et al. correlated their measurements with detailed studies of the growth of molecular water films on the mica surface, which they could image directly using scanning polarization force microscopy (Fig. 19.6). They proposed that below 20%, capillary condensation does not occur, and, indeed, they saw no evidence of a water film at these low humidities. Above 20% RH, a strongly bound molecular water layer was formed on the bare mica surface (Fig. 19.6). In the presence of the tip, a capillary meniscus condensed. Above 40% RH, the pull-off force decreased. Recognizing one of the limitations of (19.8), *Xu* et al. attribute this to the violation of assumption 5, listed above. They argue that for a pyramidal AFM tip, r_1 and r_2 become comparable in magnitude (and remain opposite in sign) once the capillary reaches the shank of the tip, leading to a near cancellation of the Laplace pressure given in (19.5). This argument is certainly plausible, although a rigorous proof is not provided, and the other limitations of (19.8) are not discussed in relation to this issue. Nevertheless, the correlation between the onset of adhesion increase and the formation of the molecular water film, as seen directly in their dramatic images, is an extremely convincing case where the classical assumptions of (19.8) must be modified to account for the molecular structure of the water film.

Further considerations of the limitations to the (19.8) were measured, discussed, and modeled in a detailed paper by *Xiao* and *Qian* [19.91]. Adhesion measurements, collected in large numbers for good statistics, were carried out with the same silicon nitride tip on two different surfaces: hydrophilic SiO_2 and a hydrophobic layer of *n*-octadecyltrimethoxysilane (OTE) on SiO_2. Contact angle measurements to confirm the assertion of hydrophilicity were not presented, however, the OTE surface was confirmed to be hydrophobic with a water contact angle of 108°. The results are shown in Fig. 19.7. The hydrophobic surface shows no dependence on RH, whereas the hydrophilic bare SiO_2 surface

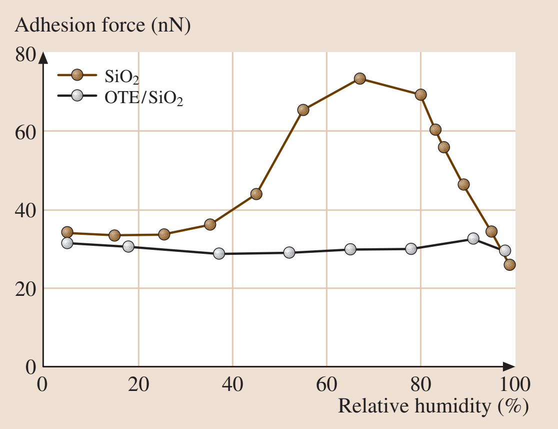

Fig. 19.7 Pull-off force as a function of RH for adhesion between a silicon nitride tip and SiO_2 (*brown circles*) and OTE/SiO_2 (*grey circles*) samples (after [19.91])

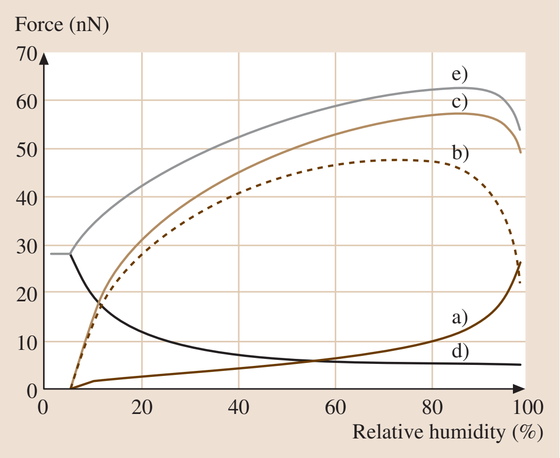

Fig. 19.8 Contributions to the adhesion force as a function of humidity: a) resolved surface tension force; b) Laplace pressure force; c) total capillary force (Laplace + surface tension); d) van der Waals force; e) total adhesion force. (After [19.91])

shows three regimes similar to the result of *Xu* et al.: constant pull-off force (>∼ 30% RH), increasing pull-off force (30%–70% RH), and then decreasing pull-off force. *Xiao* and *Qian* discuss the limitations of (19.8) in substantial detail. In particular, they consider the violation of assumptions 1 through 8 listed above. This includes a treatment of van der Waals adhesion, taking into consideration the effect of electrostatic screening of this force by the water itself. Equations for this more general case are presented. With all these aspects taken into account, they are only able to fit the model to their data qualitatively (Fig. 19.8). At low RH (< 10%), the van der Waals force dominates and the adhesion is initially constant. At intermediate values, the Laplace pressure contribution increases and then begins to saturate. The contribution from the resolved surface tension becomes significant at high RH (above ∼ 80%), and the contribution from the Laplace pressure begins to drop strongly around this same point. Qualitatively, this reproduces their results (and the aforementioned ones) by producing regimes of constant, increasing, and then decreasing adhesion. However, the humidities at which the transitions occur, and the relative changes in adhesion, do not match the data. Somewhat better agreement at high RH (> 70%) is found by considering alternate (blunt) tip shapes. Nevertheless, the most significant discrepancy occurs for low RH, where the extent of the constant adhesion force is underestimated by the classical theory. The authors attribute this to the failure of the classical continuum theory to properly describe the properties of a molecular-scale meniscus. Another interesting point of this study is that at low RH, the adhesion is very similar for both samples, a fact that the authors attribute to the dominance of the van der Waals force for both samples (which is largely determined by the substrate and not affected significantly by the OTE film).

Slightly more recently, *He* et al. have studied capillary forces for a variety of tip-sample pairs [19.102]. Hydrophilic tips (silicon and silicon nitride with no surface treatment) and hydrophobic tips (coated with *n*-octadecyltricholorsilane) were used. The hydrophobic character of the tip was asserted based on a water contact angle measurement of 105.5°, presumably taken on

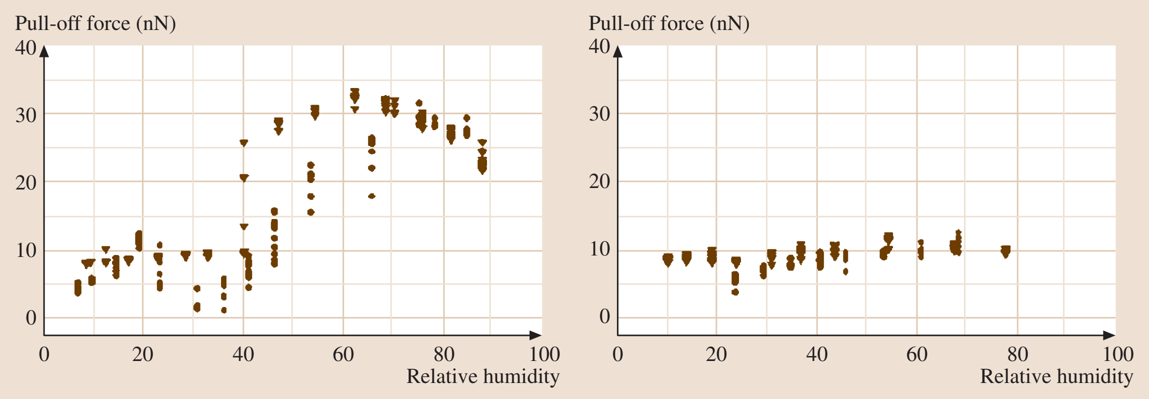

Fig. 19.9 **(a)** Pull-off force vs. RH measured between a hydrophilic tip and a flat silicon sample. *Circles*: measured when increasing RH. *Triangles*: measured when decreasing RH. **(b)** Pull-off force vs. RH measured between a sharp SFM tip coated with OTS and a flat silicon sample. The pull-off force is independent of humidity. (After [19.102])

a different region of the cantilever chip. Solvent-cleaned silicon samples and calcium fluoride films were used as hydrophilic substrates. As with the measurements of Xiao and Qian, contact angle measurements for the hydrophilic samples and tips were not presented. Results for hydrophobic and hydrophilic tips are shown in Fig. 19.9.

These results, which were carried out independently and without knowledge of Xiao and Qian's work, show impressive agreement. This is particularly interesting given that here the tip was varied from hydrophilic to hydrophobic (while the sample stayed hydrophilic), whereas in Xiao and Qian's work, the tip presumably remained hydrophilic and the sample was varied from hydrophilic to hydrophobic. For the hydrophobic tip, the pull-off force remains constant, indicating once again that capillary formation was suppressed. However, for the hydrophilic tip, three regimes of adhesion are found (Fig. 19.9). Similar results are found when a hydrophilic glass microsphere is used as a tip. The authors refer to these three regimes as the van der Waals regime, the mixed van der Waals–capillary regime, and the capillary regime. In agreement with Xiao and Qian's assessment, the authors propose that at low RH (<~ 35% in this case), the formation of the water meniscus is suppressed and the adhesion is dominated by solid-solid (presumably van der Waals) interactions. They propose, based on the work of *de Gennes* on the theory of spreading [19.87], that a minimum precursor film thickness is required to form the meniscus. The authors also present a calculation of the adhesion force when assumptions 1 and 2 are relaxed.

Note that the results discussed above consistently indicate that only one hydrophilic surface is required to form the water film. This means that should meniscus formation be undesirable in a nanoscale application (as is the case for MEMS), then hydrophobic treatment of *both* contacting surfaces is likely required.

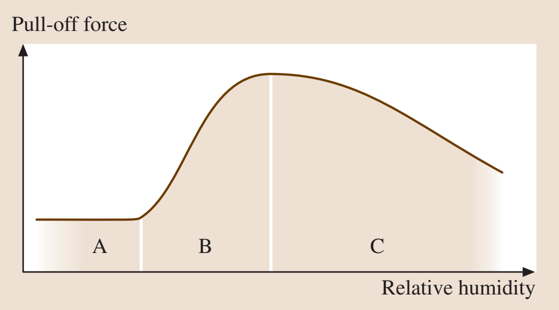

Fig. 19.10 Figure illustrating the distinct regimes of the pull-off force as a function of RH. Regimes A, B, and C are referred to as the van der Waals regime, mixed van der Waals–capillary regime, and capillary regime, respectively. (After [19.102])

Practical guidance for discerning the dependence of capillary and van der Waals force is provided by *Stifter* et al. [19.86], who have produced a detailed theoretical model of the meniscus as a function of contact angles, tip size, relative humidity, separation, and surface tension (to represent liquids other than water). An example is shown in Fig. 19.11. The model relaxes assumptions

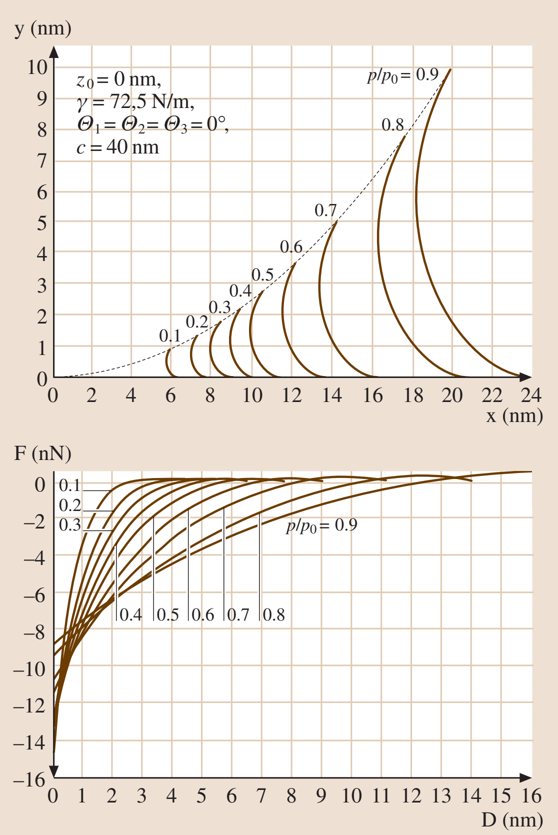

Fig. 19.11 *Top*: geometric shape of the meniscus for a range of RH values. *Bottom*: the meniscus force between the tip and the surface due to the Laplace pressure. The *dashed line* represents a cantilever with force constant 1 N/m. This relatively stiff value is still less stiff than the capillary force curve and suggests that the pull-off force would be very close to the true adhesion force. Of course, summation of the van der Waals and other contributions are necessary to make a full comparison of this sort. (After [19.86])

1 through 4 and 6. These studies show regimes where either van der Waals force or Laplace pressure forces dominate and provide guidance for actual values of the tip geometry and contact angles that minimize the capillary forces. It is not clear how much of an effect the resolved surface tension would have, as that is not included in this particular calculation. Another practical result is that for most of the plots of force vs. displacement, the "stiffness" (gradient) of the curve rarely drops below 1 N/m. This indicates that, within the assumptions of this paper, using cantilevers with force constants of 0.1 N/m will give pull-off forces that are reasonably accurate measurements of the adhesion force. This is consistent with the results of *He* et al., who found no dependence of their measurements when using cantilevers whose nominal force constants ranged from 0.032–0.5 N/m.

19.4.3 Future Directions

The results and modeling so far indicate the possibility of two trends. Hydrophobic surfaces will exhibit little dependence of adhesion on RH, whereas if one surface is hydrophilic, three regimes of behavior occur: constant solid-solid adhesion at low RH, increasing adhesion at intermediate RH, and decreasing adhesion at the highest RH. Some insight into the physical mechanisms behind these regimes has been presented, but it would be desirable to pursue further work in this area.

Perhaps most critically needed is an atomic-scale picture of the menisci and water films present under low partial pressures. This could address the question of why adhesion is initially independent of RH, and when the meniscus itself would start to form. Further theoretical developments that address the assumptions laid out above would help to clarify the picture. It is also important to extend these studies beyond simply the case of water, as the properties of other liquids are also of great interest and could be compared to previous experiments with the SFA and other tools.

Studies that clarify the kinetics of meniscus formation are also needed. There should be a noticeable transition in behavior once the rate of displacement becomes comparable to the appropriate kinetic rates, and this could provide valuable information about these kinetic processes at the nanometer scale.

Finally, as mentioned before, there continues to be a gap in reproducibility and comparability between laboratories that will only be bridged when standard techniques for tip characterization and force and displacement calibration are addressed. Efforts that take these considerations into account are worthy of support.

19.5 Self-Assembled Monolayers

When thinking about adhesion and the related phenomenon of friction, it is important to realize that the interfaces of real surfaces in contact are rarely atomically smooth. Surfaces that appear smooth on the macroscopic scale, upon closer inspection are found to consist of nanometer-scaled asperities (typically on the order of 10 nanometers) whose intentional or accidental interactions ultimately control adhesion, friction, and wear at contacts [19.6, 7, 103–105]. The size of these asperities becomes particularly important when one considers that the true contact area between interfaces for the load distribution is localized through these asperity-asperity interactions in wich extremely high pressures can be produced, sharply increasing local stress fields that can cause materials to yield and shear as they encounter each other during sliding and intermittent contact. In addition to load distribution at nanoscale asperity-asperity contacts, their size will influence surface wetting and adhesion due to capillary forces localized at the contacts [19.7, 104, 105]. The structure of applied lubricant films at such asperities will be highly dependent upon asperity curvature, and defects in lubricant film structure may form more readily here than on atomically flat surfaces.

19.5.1 Adhesion at SAM Interfaces

The minimization of adhesion at such asperity-asperity contacts is a critical issue in MEMS devices [19.6, 7, 104, 105]. In fact, the intentional introduction of surface roughness (on the order of 10 nm RMS) can be employed to lead to reduced stiction during post–processing feature release. These same asperities, however, must later resist wear during controlled or accidental contact during device operation. Thus, the specific details of adhesion and energy dissipation at such asperity-asperity contacts are required for the rational design of such systems.

To function as a protective lubricant layer in such systems, self-assembled monolayers (SAMs) of alkylsilane and fluorosilane compounds with chain lengths ranging from C_{10} to C_{18} have been useful in the reduction of friction and adhesion in MEMS [19.7, 104–106]. Such direct applications of SAMs as lubricant films, combined with the ease of sample preparation and the ability to generate model surfaces with well-defined film chemistry and structure, have made nanotribological and adhesion studies of SAMs a rich area of research [19.5]. Many of these nanotribological studies have used AFM to examine either SAMs of alkylsilane films on atomically smooth Si wafers, glass, or mica surfaces. Alternatively, many researchers have examined alkanethiol films on atomically smooth Au(111) surfaces. Using this approach, many molecular level details, such as the influence of film chemistry and molecular organization on friction, and the adhesion and wear of SAMs, can be obtained. Developing a clear understanding of the details of adhesion at SAM-modified surfaces allows for the complex links between surface chemistry, adhesion, and friction to be understood at the molecular level. In this section, we overview AFM studies of adhesion on SAM-modified surfaces.

19.5.2 Chemical Force Microscopy: General Methodology

In order to probe adhesion between chemically modified surfaces using AFM, the probe tips and sample surfaces are typically modified via self-assembly of monolayers using organosilanes on such surfaces as mica, glass, or oxidized Si or Si_3N_4 (the latter two being the typical materials of which AFM tips are made), or formed from thiols on Au-coated AFM tips and surfaces (Fig. 19.12) [19.107, 108]. While this has been shown to be a facile method for the modification of AFM tips for chemical force measurements, it should be noted that the details of the packing densities of the monolayers formed on the AFM tips are generally not known. This lack of detail regarding molecular overlayers on AFM probe tips can be a problem that requires careful consideration when using such chemically modified tips for the quantitative determination of adhesion forces and molecular interactions, as the number density of species in the tip sample contact is related to the measured adhesion.

Details of the environment in which the adhesion measurements are carried out are also an important consideration. Under ambient environmental lab conditions, surfaces are contaminated with organic compounds from the air, as well as with a layer of condensed water vapor, which varies with humidity. The condensed water layer can form a contact meniscus between the tip and sample, introducing a capillary force into the measured adhesion [19.109]. The presence of this capillary force can overwhelm the details of the adhesion from the SAM-terminated surfaces to be probed. To eliminate capillary forces, many studies are performed under liquid environments or ultrahigh vacuum. Under liquids, the nature of the solvent will of course impact the measured adhesion for a given pair of interacting surfaces, as solvent exclusion plays an important role. Also, in the case of water, the pH and ionic strength of the water environment can influence the measured adhesion in the presence of any surface-bound charges.

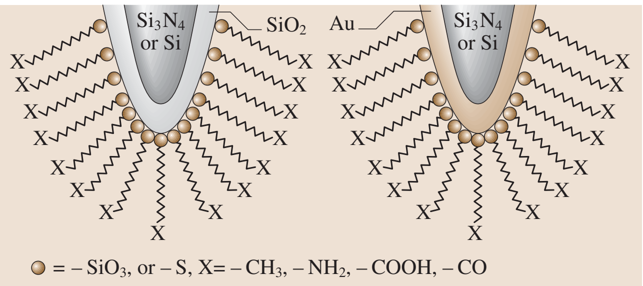

Fig. 19.12 The modification of AFM tips for chemical force microscopy is frequently carried out by chemical functionalization by alkylsilane monolayers on the oxidized surfaces of Si- or Si_3N_4-based tips, or by alkylthiol monolayers assembled on Au-coated (∼ 50 nm) AFM tips. The Au tips also typically have a Cr binding layer (∼ 5 nm) placed on before Au coating

In quantifying adhesion energies from AFM measurements, the contact mechanics model developed by Johnson, Kendall, and Roberts (JKR model) [19.20, 95] is often employed in the analysis of the adhesion data acquired by force-distance spectroscopy, whereby the number of interacting species (and consequently the average "unit" interaction force or energy) can be derived from the estimated contact area and the average molecular packing density. Using the JKR model, the force of adhesion (AFM pull-off force) is related to the work of adhesion, W_{adh}, and the reduced radius, R, of the tip-surface contact:

$$F_{adh} = -\frac{3}{2}\pi R W_{adh} . \tag{19.9}$$

The work of adhesion is a combination of the tip-surface (γ_{ts}), tip-solvent (γ_{tl}) and surface-solvent (γ_{sl}) interfacial energies ($W_{adh} = \gamma_{sl} + \gamma_{tl} - \gamma_{ts}$), and for tip-surface combinations that have the same chemical composition, the surface energy may be estimated directly from the adhesion measurement as W_{adh}. The effective contact radius at separation, r_s, from the JKR model is given as:

$$r_\mathrm{s} = \left(\frac{3\pi W_\mathrm{adh} R^2}{2K}\right)^{1/3} , \qquad (19.10)$$

where K is the reduced elastic moduli of the tip and surface. Using the contact area at separation and the assumed packing density of the molecules at the surfaces in contact, an estimate of the adhesion force or interaction energy on a per molecule basis can be made.

The accuracy of the interfacial energies and per molecule values obtained using this approach, however, must be carefully considered, due to the accumulation of error carried by the imprecise knowledge of the contact, including the tip radius, molecular packing densities of the modified surfaces, as well as the associated elastic properties of the contact at the monolayer level. As the details of the elastic properties of self-assembled monolayers are generally not known, the elastic properties of the contacts are typically assumed to be dominated by those of the underlying substrate, and the bulk values of the surface and/or tip materials (Au, Si, mica, Si_3N_4, SiO_2) are often employed in these calculations. Moreover, as mentioned above, if the molecular packing densities of the monolayers being evaluated are not known (as is the case with a typical AFM tip), then estimations must be used. For contact areas at pull-off approaching 1 nm^2, if the error in packing density is as much as 1 molecule per nm^2, this can lead to an error as high as 25% for a typical alkanethiol, or more in the reported per molecule adhesion force. Such details should be taken into consideration when describing quantitative measurements.

Table 19.1 List of various interactions evaluated by atomic force microscopy adhesion measurements

Chemical Contacts	References
$-CH_3/-CH_3$	[19.110–119]
$-CH_3/-COOH$	[19.110, 114, 116, 118]
$-CH_3/-CONH_2$	[19.118]
$-CH_3/-NH_2$	[19.118]
$-CH_3/-OH$	[19.118]
$-CF_3/-CF_3$	[19.113]
$-CF_3/-CH_3$	[19.113]
$-OH/-OH$	[19.30, 115, 117–121]
$-OH/NH_2$	[19.118]
$-OH/-COOH$	[19.117, 118]
$-OH/-CONH_2$	[19.118]
$-COOH/-COOH$	[19.30, 111, 114, 115, 117, 118, 122]
$-CONH_2/-CONH_2$	[19.118]
$-CONH_2/-COOH$	[19.118]
$-NH_2/-NH_2$	[19.116–118, 123]
$-CH_3/-NH_2$	[19.118, 123]
$-NH_2/-COOH$	[19.118]
$-NH_2/-CONH_2$	[19.118]
$-SO_3H/-SO_3H$	[19.123]
$-CH_3/-SO_3H$	[19.123]
$-NH_2/-SO_3H$	[19.123]
$-PO_3H_2/-PO_3H_2$	[19.116]
$-SH/-SH$	[19.119]

An alternative approach to the measure of adhesive interactions based on Poisson statistics has been promoted by *Beebe* et al. for the statistical evaluation of single bond forces without a priori knowledge of the tip-surface contact details involved [19.119, 121, 122]. A principal limitation of this approach, however, is that a completely homogeneous chemical system is assumed, i. e., that there is only one type of discrete interaction present that gives rise to the observed adhesion. Unfortunately, for many solution phase systems a number of different interactions are typically operative, including energetic exchange with and reorganization of solvent molecules, depending on the solution conditions. These issues have never been thoroughly addressed in any molecular-level measurement of adhesion.

19.5.3 Adhesion at SAM-Modified Surfaces in Liquids

A number of researchers have used AFM to probe interfacial adhesion for a variety of different chemical systems. Most notably, *Lieber* et al. promoted the use of chemically modified substrates and AFM tips to study selective molecular interactions [19.73, 107, 110, 114, 117, 124]. A number of other researchers have adopted a similar methodology, leading to the measurement of adhesion forces and interfacial energies for the interactions of a variety of molecular functional groups in various environments, although to date few specific types of interactions have been thoroughly investigated (Table 19.1) [19.30, 107, 110–115, 120, 122, 125–129].

As described above, measurements in air suffer from difficulties of water vapor, thus, the adhesion of many SAM-terminated surfaces have been evaluated under liquids, and it is these systems that we shall focus on here. The value of the adhesion is of course modified depending on the solvent. For example, the interaction of methyl-terminated interfaces is much stronger in water than in nonpolar solvents, in agreement with the general concept that upon separation the generation of hydrophobic interfaces in a polar solvent is highly energetically unfavorable. The impact of solvent on adhesion can be addressed basically as a variation on the Hamaker constant.

When ionizable end groups are studied in water, the details of the adhesion measurements and results become more complicated, as the chemical nature of the surfaces are now dependent on the pH and ionic strength of the solution. In these circumstances, multiple interactions, including ionic, van der Waals and double-layer forces, come into play simultaneously. Under these conditions, the general form for the JKR adhesion force may be modified to include the following additional forces:

$$F_{\text{adh}} = -\frac{3}{2}\pi R W_{\text{adh}} - \frac{AR}{6D} + 6\pi R W_{\text{dl}} \,, \tag{19.11}$$

where the second term now includes the attractive van der Waals component and the third term is the repulsive double layer.

Several groups have utilized the ability of AFM to function as a local probe of ionicity and to carry out local force titration measurements on several functional groups, including $-$COOH [19.30, 114–118, 122], $-NH_2$ [19.116–118, 123], $-SO_3H$ [19.123] and $-PO_3H_2$ [19.116, 130] (see Table 19.1). The force-distance curves can give a general sense of the local chemical state based on whether the approach curve is attractive or repulsive. In force titrations, the adhesion is measured as a function of pH, and the change in force is dependent upon the equilibrium mixture of charged species within the tip-samples contact at the time of measurement. Peaks in the adhesion force vs. pH allow for the determination of the local pK values for the ionized species, which are shown to be dependent on ionic strength. A force titration of a diprotic acid (11-thioundecyl-1-phosphonic acid) is shown in Fig. 19.13. Shifts in the pK values from those measured in free solution have been ascribed to a variety of factors, including buildup of excess surface charge and solvation effects of the surface-bound ionic species. The localization of charge and change migration are effects that have yet to be fully explored in these systems, however.

19.5.4 Impact of Intra- and Inter-Chain Interactions on Adhesion

Several studies of the effects of chain length on friction and adhesion have found that, in general, adhesion decreases with increasing chain length, due to increased stability from lateral chain–chain van der Waals interactions within the film, which increases the overall film stiffness. The increased stiffness of the films acts to reduce the effective contact area that develops under compression, consequently reducing adhesion and friction [19.126, 127, 131]. In alkyl chain-based monolayers, the end-group orientation, however, is also dependent on chain length. This results in an "odd-even" effect on the measured adhesion/friction of the monolayer. With $-CH_3$-terminated films, the methyl group orientation differs for odd and even length molecules. This impacts the orientation of the methyl group net dipole and, hence, the local surface free energy [19.73]. In circumstances where interchain hydrogen bonding can occur, additional film stability can be introduced, also yielding an odd–even effect, as observed by *Houston* et al. [19.111, 125].

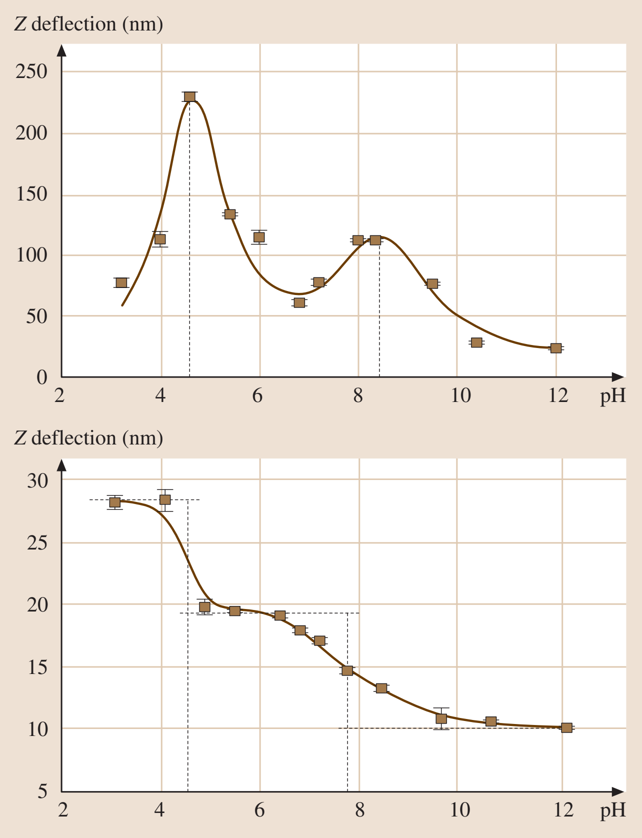

Fig. 19.13a,b Adhesion vs. pH for 11-thioundecyl-1-phosphonic acid in buffer illustrating the impact of pH and ionic strength, **(a)** 10^{-4} M and **(b)** 10^{-1} M on the measured adhesion. The peaks in adhesion provide local measures of the surface pK (adapted after [19.130])

Studies by AFM of alkanethiol films on Au surfaces have shown that, depending upon the tip size, under varying loads, the tip can readily penetrate the SAM film, displacing the thiol layer from the tip-sample contact [19.132–134]. Upon reduction of the force, the tip again moves out of the SAM film and the surface structure is returned to its original condition. Of key importance in such studies is the mechanism by which the film is displaced. Recently, this has been modeled by *Harrison* using molecular dynamics simulations [19.135]. These studies have clearly demonstrated that in the initial stages of film compression and penetration by an asperity, gauche defects within the typically all-trans configurations of molecules in the SAM layers appear and propagate. The introduction of such defects is the catalyst for the weakening of the chain-chain lateral interactions that help stabilize and maintain film integrity. As this is lost, the asperity can rapidly penetrate the film, and alkanethiol displacement can occur either via chain collapse or bond scission from the surface. *Salmeron* et al. demonstrated that the prevalence of molecular displacement versus film compression depends heavily on the AFM tip size, with sharp tips readily penetrating and displacing surface-bound thiols, while large tips spread the load over more molecules and induce compression over displacement (Fig. 19.14).

The inherent stability of the film structure has also been confirmed by sum-frequency generation spectroscopic studies that have indicated that without the presence of lateral chain-chain interactions, gauche defects appear within the chain structure that reduce overall order and lateral interactions within films [19.136]. The ability to form these defects through poor film order will consequently increase adhesion, friction, and wear of the film. This same spectroscopic study further demonstrated that the appearance of gauche defects can be induced by controlling the local environment, i. e., with the presence of water, causing chains to collapse back upon themselves.

As molecular structure in SAMs moves away from ideally organized layers, chain-chain interactions between the contacting surfaces can also result in entanglements. A recent study of the adhesive interactions of Au(111) surfaces modified with dialkylsulfides $[\mathrm{CH_3(CH_2)}_n\mathrm{-S-(CH_2)_9CH_3};\ n = 9, 11, 13, 15, 17]$ with varying chain arm lengths probed the combined effect of chain length, solvent, and inter-surface chain entanglement on friction and adhesion using simultaneously modified surfaces and AFM tips [19.137]. This study found that chain-chain interpenetration produced the reverse dependence on chain length on the measured adhesion mentioned above. This work points to the need to examine the nature of inter-surface chain entanglements in nanoscaled systems. Such entanglements should be more prevalent in asperity-asperity contacts where ideal film structure will not be feasible.

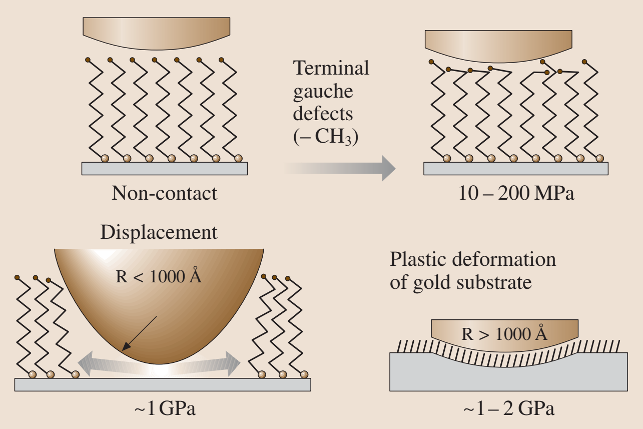

Fig. 19.14 Schematic illustrating the compression of a model lubricant layer under an AMF tip. At low pressures, the gauche defects can form at the tip–sample junction, but the molecules remain in place. Depending on tip size and load, displacement (for sharp tips) or trapping of monolayer molecules (for blunt tips) can occur (adapted after [19.5])

19.5.5 Adhesion at the Single-Bond Level

The ability to resolve the discrete components of interactions is highly desirable. There have been few reports of the direct observation of discrete force components observed with the separation of an AFM tip from a surface. *Beebe* et al., have utilized a statistical method (as described above) for the direct determination of single-bond forces for a variety of interactions [19.119, 121, 122], including biological systems such as biotin-avidin [19.138, 139]. The first report of quantized force measurements was described by *Hoh* et al., which lead to the estimation of single hydrogen bonding forces from studies of glass surfaces in water to be on the order of 10 pN [19.140]. The use of AFM for the study of the energetics of true single chemical bond cleavage has also received little attention. One previous report described discrete covalent bond scission using AFM. In that case, it was proposed that the jumps in the observed pull-off curves were due to sequential scission of chemical bonds contained in a large,

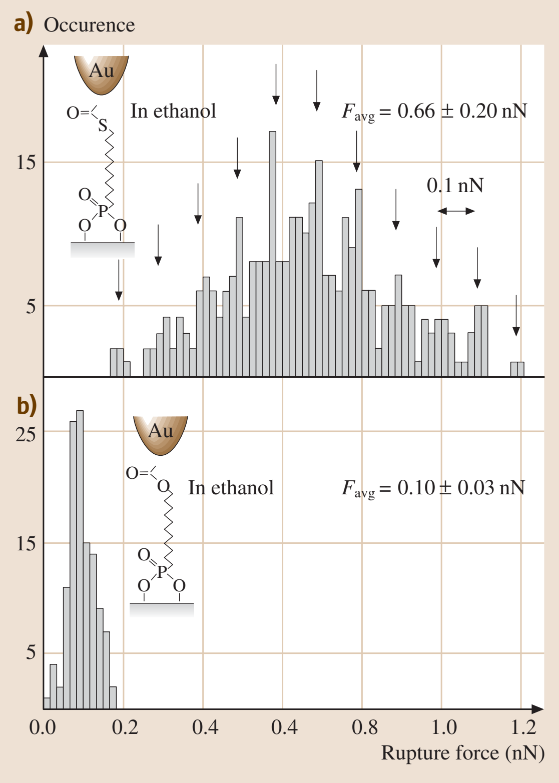

Fig. 19.15 Pull-off force distribution for Au atom abstraction in ethanol. The 100 pN peaks in the distribution represent quantized numbers of Au atoms being abstracted from the sample surface (adapted after [19.129])

multifunctional, polymeric species as it progressively detached from the substrate [19.141]. However, the identification of the relevant chemical bonds involved at each stage was largely based on the known (gas phase) bond strengths of the potentially active functional groups such that solvation effects on the bond energies were ignored – a simplification that profoundly affects the estimated energetics, as energetic exchange with the solvent must be included. More recently, the measurement of discrete bond scission was reported by *Frisbie* et al. for Au-thiol complexes (Fig. 19.15). Here the details of Au atom abstraction were reported with a quantized value of 100 pN (estimated at ~ 10 kJ/mol, based on an assumed bond rupture length of 1 Å) [19.129]. These studies have demonstrated the feasibility of probing local single bond energetics and have suggested some general requirements for the measurement of adhesion quantization in SAM layers, including the need for a significant negative tip-surface interfacial energy coupled with minimal solvent surface tension [19.129, 142].

19.5.6 Future Directions

Extending AFM adhesion measurements to reactive systems in which chemical bonds can form between the tip and surface affords an expansion of chemical reaction dynamics to solution-based chemistries, whereby the energetic details of single reaction events, previously only accessible for gas phase scattering experiments, may be obtained. Studies of such complex heterogeneous systems will open the door to the evaluation of the energetic pathways of solution-based chemistries for any system in which the appropriate functionalization of surfaces can be exploited. Adhesion has already been demonstrated as a reasonable local probe of surface reaction kinetics, whereby the local changes in the chemical forces may be followed as a function of time during surface chemical reactions [19.143, 144]. In addition to advances in measurements of reactive systems by AFM, complete insight into the operative molecular mechanisms can only be gained when combined with detailed theory that takes into account not only the specific types of interactions present between the surfaces, but also the requisite energetic exchange with local solvent molecules. Advances in computer technologies and in computational theory have made this realistic [19.128].

To advance the area of adhesion measurements at the molecular level, energetic barriers for specific interactions also need to be evaluated, with attention given to the nature of the molecular interactions being probed, *including* the details of energetic exchange with the solvent surrounding the interacting molecules. Studies by AFM of molecular interactions within sharply confined geometries (i. e., ~ 1 nm^2 contact area) provide an opportunity to evaluate such contributions with molecular detail. Here again, *Lieber* et al. have been advancing the approach of chemically functionalizing carbon nanotubes to reduce both the type and number of specific interacting species [19.73]. This approach may hold some promise for probing well-defined, specific chemical interactions and/or reactions, as long as the nanotubes can be sufficiently stabilized against buckling [19.135, 145, 146].

In addition to the modification of probe geometries to improve localization of interactions for adhesion measurements, one of the principle difficulties in performing AFM measurements of adhesion at SAM surfaces is the unavoidable snap to contact. This makes the details of the long-range interaction potential almost completely inaccessible. To address this issue, *Lieber* et al. have developed a modified AFM system in which the cantilever is magnetized, providing an additional feedback mechanism to aid in avoiding snap-in [19.30]. When used, this approach provides a smooth approach and retract curve. Similarly, the capacitive coupling feedback mechanism of the interfacial force microscope (IFM) also affords measurements of this transition without snap-in [19.125].

19.6 Concluding Remarks

Scanning probes are powerful tools for determining the fundamental molecular basis of adhesion. Continuum models of adhesion and capillary condensations are useful, but careful attention must be paid to their limits and assumptions. Further progress in these areas requires detailed analysis of the structure and chemistry of both the tip and sample, as well as their environment (solvent, humidity, etc.).

The ability to determine the effects of molecular-scale water menisci or single bond energetics are truly spectacular accomplishments that continue to inspire researchers worldwide to pursue these measurements. With attention paid to previous literature and consideration of issues raised in this chapter, many more discoveries are sure to be made.

References

19.1 I. L. Singer, H. M. Pollock (Eds.): *Fundamentals of Friction* (Kluwer, Dordrecht 1992) p. 351

19.2 K. Autumn, Y. A. Liang, S. T. Hsieh, W. Zesch et al.: Adhesive force of a single gecko foot-hair, Nature **405** (2000) 681–685

19.3 A. L. Baldwin, G. Thurston: Mechanics of endothelial cell architecture and vascular permeability, Critical Rev. Biomed. Eng. **29** (2001) 247–278

19.4 S. R. White, N. R. Sottos, P. H. Guebelle, J. S. Moore et al.: Autonomic healing of polymer composites, Nature **409** (2001) 794–797

19.5 R. W. Carpick, M. Salmeron: Scratching the surface: Fundamental investigations of tribology with atomic force microscopy, Chem. Rev. **97** (1997) 1163–1194

19.6 R. Maboudian: Adhesion and friction issues associated with reliable operation of MEMS, MRS Bull. **23** (1998) 47–51

19.7 R. Maboudian, W. R. Ashurst, C. Caffaro: Tribological challenges in micromechanical systems, Tribol. Lett. **12** (2002) 95–100

19.8 R. Maboudian, R. T. Howe: Critical review: Adhesion in surface micromechanical structures, J. Vac. Sci. Technol. **15** (1997) 1

19.9 J. N. Israelachvili: Thin film studies using multiple-beam interferometry, J. Colloid Interface Sci. **44** (1973) 259–272

19.10 J. N. Israelachvili, D. Tabor: The measurement of van der Waals dispersion forces in the range 1.5 to 130 nm, Proc. R. Soc. London A **331** (1972) 19–38

19.11 G. Binnig, C. F. Quate, C. Gerber: Invention of AFM, Phys. Rev. Lett. **56** (1986) 930

19.12 P. Frantz, N. Agraït, M. Salmeron: SFA w/ capacitance, Langmuir **12** (1996) 3289–3294

19.13 J. N. Israelachvili, P. M. McGuiggan, A. M. Homola: Dynamic properties of molecularly thin liquid films, Science **240** (1988) 189–191

19.14 J. Peachey, J. Van Alsten, S. Granick: Design of an apparatus to measure the shear response of ultrathin liquid, Rev. Sci. Instrum. **62** (1991) 463–473

19.15 L. R. Fisher, J. N. Israelachvili: Direct measurement of the effect of meniscus forces on adhesion: A study of the applicability of macroscopic thermodynamics to microscopic liquid interfaces, Colloids Surf. **3** (1981) 303–319

19.16 L. R. Fisher, J. N. Israelachvili: Experimental studies on the applicability of the Kelvin equation to highly curved concave menisci, J. Colloid Interface Sci. **80** (1981) 528–541

19.17 J. N. Israelachvili: *Intermolecular and Surface Forces* (Academic Press, London 1992)

19.18 S. Granick: Molecular tribology of fluids. In: *Fundamentals of Friction*, ed. by I. L. Singer, H. M. Pollock (Kluwer, Dordrecht 1992) p. 387

19.19 G. Reiter, A. L. Demirel, J. Peanasky, L. L. Cai et al.: Stick to slip transition and adhesion of lubricated surfaces in moving contact, J. Chem. Phys. **101** (1994) 2606

19.20 K. L. Johnson: *Contact Mechanics* (Cambridge Univ. Press, Cambridge 1987)

19.21 O. Marti, B. Drake, P. K. Hansma: Atomic force microscopy of liquid-covered surfaces: Atomic resolution images, Appl. Phys. Lett. **51** (1987) 484–486

19.22 G. J. Germann, S. R. Cohen, G. Neubauer, G. M. McClelland et al.: Atomic scale friction of a diamond tip on diamond (100) and (111) surfaces, J. Appl. Phys. **73** (1993) 163

19.23 L. Howald, E. Meyer, R. Lüthi, H. Haefke et al.: UHV AFM design, Appl. Phys. Lett. **63** (1993) 117

19.24 M. Kageshima, H. Yamada, K. Nakayama, H. Sakama et al.: Development of an ultrahigh vacuum atomic force microscope for investigations of semiconductor surfaces, J. Vac. Sci. Technol. B **11** (1993) 1987–1991

19.25 J. A. Greenwood: Adhesion of elastic spheres, Proc. R. Soc. London A **453** (1997) 1277–1297

19.26 S. P. Jarvis, A. Oral, T. P. Weihs, J. B. Pethica: A novel force microscope and point contact probe, Rev. Sci. Instrum. **64** (1993) 3515

19.27 S. P. Jarvis, H. Yamada, S.-I. Yamamoto, H. Tokumoto: A new force controlled atomic force microscope for use in ultrahigh vacuum, Rev. Sci. Instrum. **67** (1996) 2281

19.28 S. A. Joyce, J. E. Houston: A new force sensor incorporating force-feedback control for interfacial force microscopy, Rev. Sci. Instrum. **62** (1991) 710–715

19.29 S. A. Joyce, J. E. Houston, T. A. Michalske: Differentiation of topographical and chemical structures using an interfacial force microscope, Appl. Phys. Lett. **60** (1992) 1175

19.30 P. D. Ashby, L. W. Chen, C. M. Lieber: Probing intermolecular forces and potentials with magnetic feedback chemical force microscopy, J. Am. Chem. Soc. **122** (2000) 9467–9472

19.31 H. I. Kim, V. Boiadjiev, J. E. Houston, X.-Y. Zhu, J. D. Kiely: Tribological properties of self-assembled monolayers on Au, SiO_x and Si surfaces, Tribol. Lett. **10**(1–2) (2001) 97–101

19.32 J. S. Nelson, B. W. Dodson, P. A. Taylor: Adhesive avalanche in covalently bonded materials, Phys. Rev. B **45** (1992) 4439–4444

19.33 P. A. Taylor, J. S. Nelson, B. W. Dodson: Adhesion between atomically flat metallic surfaces, Phys. Rev. B **44** (1991) 5834–5841

19.34 K. R. Shull: Contact mechanics and the adhesion of soft solids, Mater. Sci. Eng. R: Rep. **R 36** (2002) 1–45

19.35 L. Xu, A. Lio, J. Hu, D. F. Ogletree et al.: Wetting and capillary phenomena of water on mica, J. Phys. Chem. B **102** (1998) 540–548

19.36 S. P. Timoshenko, J. N. Goodier: *Theory of Elasticity* (McGraw Hill, New York 1987)

19.37 J. P. Cleveland, S. Manne, D. Bocek, P. K. Hansma: A nondestructive method for determining the spring constant of cantilevers for scanning force microscopy, Rev. Sci. Instrum. **64** (1993) 403

19.38 J. E. Sader: Parallel beam approximation for V-shaped atomic force microscope cantilevers, Rev. Sci. Instrum. **66** (1995) 4583

19.39 M. Tortonese, M. Kirk: Characterization of application specific probes for SPMs, Proc. SPIE **3009** (1997) 53–60

19.40 T. R. Albrecht, S. Akamine, T. E. Carver, C. F. Quate: Microfabrication of cantilever styli for the atomic force microscope, J. Vac. Sci. Technol. A **8** (1990) 3386–3396

19.41 H.-J. Butt, P. Siedle, K. Seifert, K. Fendler et al.: Scan speed limit in atomic force microscopy, J. Microsc. **169** (1993) 75–84

19.42 Y. Q. Li, N. J. Tao, J. Pan, A. A. Garcia et al.: Direct measurement of interaction forces between colloidal particles using the scanning force microscope, Langmuir **9** (1993) 637–641

19.43 R. Lüthi, E. Meyer, H. Haefke, L. Howald et al.: Nanotribology: An UHV-SFM study on thin films of C_{60} and AgBr, Surf. Sci. **338** (1995) 247–260

19.44 J. M. Neumeister, W. A. Ducker: Lateral, normal, and longitudinal spring constants of atomic force microscopy cantilevers, Rev. Sci. Instrum. **65** (1994) 2527–2531

19.45 D. F. Ogletree, R. W. Carpick, M. Salmeron: Calibration of frictional forces in atomic force microscopy, Rev. Sci. Instrum. **67** (1996) 3298–3306

19.46 J. A. Ruan, B. Bhushan: Atomic-scale friction measurements using friction force microscopy: Part I – general principles and new measurement techniques, Trans. ASME J. Tribol. **116** (1994) 378

19.47 J. E. Sader, I. Larson, P. Mulvaney, L. R. White: Method for the calibration of atomic force microscope cantilevers, Rev. Sci. Instrum. **66** (1995) 3789

19.48 U. D. Schwarz, P. Koster, R. Wiesendanger: Quantitative analysis of lateral force microscopy experiments, Rev. Sci. Instrum. **67** (1996) 2560

19.49 T. J. Senden, W. A. Ducker: Experimental determination of spring constants in atomic force microscopy, Langmuir **10** (1994) 1003

19.50 A. Torii, M. Sasaki, K. Hane, S. Okuma: A method for determining the spring constant of cantilevers for atomic force microscopy, Meas. Sci. Technol. **7** (1996) 179–184

19.51 J. E. Sader, J. W. M. Chon, P. Mulvaney: Calibration of rectangular atomic force microscope cantilevers, Rev. Sci. Instrum. **70** (1999) 3967–3969

19.52 J. S. Villarrubia: Morphological estimation of tip geometry for scanned probe microscopy, Surf. Sci. **321** (1994) 287–300

19.53 F. Atamny, A. Baiker: Direct imaging of the tip shape by AFM, Surf. Sci. **323** (1995) L314

19.54 R. W. Carpick, N. Agraït, D. F. Ogletree, M. Salmeron: Measurement of interfacial shear (friction) with an ultrahigh vacuum atomic force microscope, J. Vac. Sci. Technol. B **14** (1996) 1289–1295

19.55 R. Dixson, J. Schneir, T. McWaid, N. Sullivan et al.: Toward accurate linewidth metrology using atomic force microscopy and tip characterization, Proc. SPIE **2725** (1996) 589–607

19.56 L. S. Dongmo, J. S. Villarrubia, S. N. Jones, T. B. Renegar et al.: Experimental test of blind tip reconstruction for scanning probe microscopy, Ultramicroscopy **85** (2000) 141–153

19.57 K. F. Jarausch, T. J. Stark, P. E. Russell: Silicon structures for in situ characterization of atomic force microscope probe geometry, J. Vac. Sci. Technol. B **14** (1996) 3425–3430

19.58 P. Markiewicz, M. C. Goh: Atomic force microscope tip deconvolution using calibration arrays, Rev. Sci. Instrum. **66** (1995) 3186–3190

19.59 C. Odin, J. P. Aimé, Z. El Kaakour, T. Bouhacina: Tip's finite size effects on atomic force microscopy in the contact mode: Simple geometrical considerations for rapid estimation of apex radius and tip angle based on the study of polystyrene latex balls, Surf. Sci. **317** (1994) 321–340

19.60 S. S. Sheiko, M. Moller, E. M. C. M. Reuvekamp, H. W. Zandbergen: Evaluation of the probing profile of scanning force microscopy tips, Ultramicroscopy **53** (1994) 371–380

19.61 P. Siedle, H.-J. Butt, E. Bamberg, D. N. Wang et al.: Determining the form of atomic force microscope tips. In: *X-Ray Optics and Microanalysis*, ed. by P. B. Kenway, P. J. Duke, G. W. Lorimer et al. (Iop, Manchester 1992) pp. 361–364

19.62 J. S. Villarrubia: Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation, J. Res. Nat. Inst. Stand. Technol. **102** (1997) 425–454

19.63 K. L. Westra, D. J. Thomson: Atomic force microscope tip radius needed for accurate imaging of thin film surfaces, J. Vac. Sci. Technol. B **12** (1994) 3176–3181

19.64 S. Xu, M. F. Arnsdorf: Calibration of the scanning (atomic) force microscope with gold particles, J. Microsc. **3** (1994) 199–210

19.65 U. D. Schwarz, O. Zwörner, P. Köster, R. Wiesendanger: Friction force spectroscopy in the low-load regime with well-defined tips. In: *Micro/Nanotribology and Its Applications*, ed. by B. Bhushan (Kluwer, Dordrecht 1997)

19.66 P. Niedermann, W. Hanni, N. Blanc, R. Christoph et al.: Chemical vapor deposition diamond for tips in nanoprobe experiments, J. Vac. Sci. Technol. A **14** (1995) 1233–1236

19.67 S. J. O'Shea, R. N. Atta, M. E. Welland: Characterization of tips for conducting atomic force microscopy, Rev. Sci. Instrum. **66** (1995) 2508–2512

19.68 L. M. Qian, X. D. Xiao, S. Z. Wen: Tip in situ chemical modification and its effects on tribological measurements, Langmuir **16** (2000) 662–670

19.69 E. L. Florin, V. T. Moy, H. E. Gaub: Adhesion forces between individual ligand-receptor pairs, Science **264** (1994) 415–417

19.70 G. U. Lee, L. A. Chrisey, R. J. Colton: Direct measurement of the forces between complementary strands of DNA, Science **266** (1994) 771–773

19.71 V. T. Moy, E. L. Florin, H. E. Gaub: Intermolecular forces and energies between ligands and receptors, Science **266** (1994) 257–259

19.72 O. H. Willemsen, M. M. E. Snel, K. O. van der Werf, B. G. de Grooth et al.: Simultaneous height and adhesion imaging of antibody-antigen interactions by atomic force microscopy, Biophys. J. **75** (1998) 2220–2228

19.73 S.-S. Wong, H. Takano, M. D. Porter: Mapping orientation differences of terminal functional groups by friction force microscopy, Anal. Chem. **70** (1998) 5209–5212

19.74 H. J. Butt: Measuring electrostatic, van der Waals, and hydration forces in electrolyte solutions with an atomic force microscope, Biophys. J. **60** (1991) 1438–1444

19.75 W. A. Ducker, T. J. Senden, R. M. Pashley: Direct measurement of colloidal forces using an atomic force microscope, Nature **353** (1991) 239–241

19.76 V. S. J. Craig, C. Neto: In situ calibration of colloid probe cantilevers in force microscopy: Hydrodynamic drag on a sphere approaching a wall, Langmuir **17** (2001) 6018–6022

19.77 R. Staub, D. Alliata, C. Nicolini: Drift elimination in the calibration of scanning probe microscopes, Rev. Sci. Instrum. **66** (1995) 2513–2516

19.78 J. Fu: In situ testing and calibrating of z-piezo of an atomic force microscope, Rev. Sci. Instrum. **66** (1995) 3785–3788

19.79 J. Garnaes, L. Nielsen, K. Dirscherl, J. F. Jorgensen et al.: Two-dimensional nanometer-scale calibration based on one-dimensional gratings, Appl. Phys. A **66** (1998) 831–835

19.80 S. M. Hues, C. F. Draper, K. P. Lee, R. J. Colton: Effect of PZT and PMN actuator hysteresis and creep on nanoindentation measurements using force microscopy, Rev. Sci. Instrum. **65** (1994) 1561

19.81 J. F. Jorgensen, K. Carneiro, L. L. Madsen, K. Conradsen: Hysteresis correction of scanning tunneling microscope images, J. Vac. Sci. Technol. B **12** (1994) 1702–1704

19.82 J. F. Jorgensen, L. L. Madsen, J. Garnaes, K. Carneiro et al.: Calibration, drift elimination, and molecular structure analysis, J. Vac. Sci. Technol. B **12** (1994) 1698–1701

19.83 M. Jaschke, H. J. Butt: Height calibration of optical lever atomic force microscopes by simple laser interferometry, Rev. Sci. Instrum. **66** (1995) 1258

19.84 L. A. Nagahara, K. Hashimoto, A. Fujishima, D. Snowden-Ifft et al.: Mica etch pits as a height calibration source for atomic force microscopy, J. Vac. Sci. Technol. B **12** (1993) 1694–1697

19.85 H. M. Brodowsky, U.-C. Boehnke, F. Kremer: Wide range standard for scanning probe microscopy height calibration, Rev. Sci. Instrum. **67** (1996) 4198–4200

19.86 T. Stifter, O. Marti, B. Bhushan: Theoretical investigation of the distance dependence of capillary and van der Waals forces in scanning force microscopy, Phys. Rev. B **62** (2000) 13667–13673

19.87 P. G. de Gennes: Wetting: Statistics and dynamics, Rev. Mod. Phys. **57** (1985) 827–863

19.88 M. Binggeli, C. M. Mate: Influence of capillary condensation of water on nanotribology studied by force microscopy, Appl. Phys. Lett. **65** (1994) 415

19.89 Y. Sugawara, M. Ohta, T. Konishi, S. Morita et al.: Effects of humidity and tip radius on the adhesive force measured with atomic force microscopy, Wear **168** (1993) 13–16

19.90 F. M. Orr, L. E. Scriven, A. P. Rivas: Pendular rings between solids: Meniscus properties and capillary force, J. Fluid Mech. **67** (1975) 723–742

19.91 X. Xiao, Q. Linmao: Investigation of humidity-dependent capillary force, Langmuir **16** (2000) 8153–8158

19.92 E. Riedo, F. Levy, H. Brune: Kinetics of capillary condensation in nanoscopic sliding friction, Phys. Rev. Lett. **88** (2002) 185505/1–4

19.93 D. Maugis, B. Gauthiermanuel: JKR-DMT transition in the presence of a liquid meniscus, J. Adhes. Sci. Technol. **8** (1994) 1311–1322

19.94 A. Fogden, L. R. White: Contact elasticity in the presence of capillary condensation. 1. The nonadhesive Hertz problem, J. Colloid Interface Sci. **138** (1990) 414

19.95 K. L. Johnson, K. Kendall, A. D. Roberts: Surface energy and the contact of elastic solids, Proc. R. Soc. London A **324** (1971) 301–313

19.96 K. Johnson, J. Greenwood: An adhesion map for the contact of elastic spheres, J. Colloid Interface Sci. **192** (1997) 326–333

19.97 K. L. Johnson: Adhesion and friction between a smooth elastic asperity and a plane surface, Proc. R. Soc. London A **453** (1997) 163

19.98 T. Thundat, X. Y. Zheng, G. Y. Chen, R. J. Warmack: Role of relative humidity in atomic force microscopy imaging, Surf. Sci. **294** (1993) 939–943

19.99 M. Binggeli, R. Christoph, H.-E. Hintermann: Observation of controlled, electrochemically induced friction force modulations in the nano-Newton range, Tribol. Lett. **1** (1995) 13

19.100 B. Bhushan, S. Sundararajan: Micro/nanoscale friction and wear mechanisms of thin films using atomic force and friction force microscopy, Acta Mater. **46** (1998) 3793–3804

19.101 W. Gulbinski, D. Pailharey, T. Suszko, Y. Mathey: Study of the influence of adsorbed water on AFM friction measurements on molybdenum trioxide thin films, Surf. Sci. **475** (2001) 149–158

19.102 M. He, A. S. Blum, D. E. Aston, C. Buenviaje et al.: Critical phenomena of water bridges in nanoasperity contacts, J. Chem. Phys. **114** (2001) 1355–1360

19.103 J. A. Greenwood, J. B. P. Williamson: Contact of nominally flat surfaces, Proc. R. Soc. London A **295** (1966) 300

19.104 K. Komvopoulous: Surface engineering and microtribology for microelectromechanical systems, Wear **200** (1996) 305–327

19.105 R. Maboudian, W. R. Ashurst, C. Carraro: Self-assembled monolayers as anti-stiction coatings for MEMS: Characteristics and recent developments, Sens. Actuators A **82** (2000) 219

19.106 R. Maboudian: Surface processes in MEMS technology, Surf. Sci. Rep. **30** (1998) 207–270

19.107 A. Noy, D. V. Vezenov, C. M. Lieber: Chemical force microscopy, Annu. Rev. Mater. Sci. **27** (1997) 381–421

19.108 H. Takano, J. R. Kenseth, S.-S. Wong, J. C. O'Brien et al.: Chemical and biochemical analysis using scanning force microscopy, Chem. Rev. **99** (1999) 2845

19.109 D. L. Sedin, K. L. Rowlen: Adhesion forces measured by atomic force microscopy in humid air, Anal. Chem. **72** (2000) 2183–2189

19.110 C. D. Frisbie, L. F. Rozsnyai, A. Noy, M. S. Wrighton et al.: Functional group imaging by chemical force microscopy, Science **265** (1994) 2071–2074

19.111 H. I. Kim, J. E. Houston: Separating mechanical and chemical contributions to molecular-level friction, J. Am. Chem. Soc. **122** (2000) 12045–12046

19.112 T. Nakagawa, K. Ogawa, T. Kurumizawa: Discriminating molecular length of chemically adsorbed molecules using an atomic force microscope having a tip covered with sensor molecules (an atomic force microscope having chemical sensing function), Jpn. J. Appl. Phys. **32** (1993) 294–296

19.113 T. Nakagawa, K. Ogawa, T. Kurumizawa: Atomic force microscope for chemical sensing, J. Vac. Sci. Techol. **B 12** (1994) 2215–2218

19.114 A. Noy, C. D. Frisbie, L. F. Rozsnyai, M. S. Wrighton et al.: Chemical force microscopy: Exploiting chemically-modified tips to quantify adhesion, friction, and functional group distributions in molecular assemblies, J. Am. Chem. Soc. **117** (1995) 7943–7951

19.115 S. K. Sinniah, A. B. Steel, C. J. Miller, J. E. Reutt-Robey: Solvent exclusion and chemical contrast in scanning force microscopy, J. Am. Chem. Soc. **118** (1996) 8925–8931

19.116 E. W. v.d. Vegte, G. Hadziioannou: Acid-base properties and the chemical imaging of surface-bound functional groups with scanning force microscopy, J. Phys. Chem. **B 101** (1997) 9563–9569

19.117 D. V. Vezenov, A. Noy, L. F. Rozsnyai, C. M. Lieber: Force titrations and ionization state sensitive imaging of functional group distributions in molecular assemblies, J. Am. Chem. Soc. **119** (1997) 2006–2015

19.118 E. W. v.d. Vegte, G. Hadziioannou: Scanning force microscopy with chemical specificity: An extensive study of chemically specific tip-surface interactions and the chemical imaging of surface functional groups, Langmuir **13** (1997) 4357–4368

19.119 L. A. Wenzler, G. L. Moyes, L. G. Olson, J. M. Harris et al.: Single-molecule bond rupture force analysis of interactions between AFM tips and substrates modified with organosilanes, Anal. Chem. **69** (1997) 2855–2861
19.120 T. Ito, M. Namba, P. Buhlmann, Y. Umezawa: Modification of silicon nitride tips with trichlorosilane self-assembled monolayers (SAMs) for chemical force microscopy, Langmuir **13** (1997) 4323–4332
19.121 L. A. Wenzler, G. L. Moyes, G. N. Raikar, R. L. Hansen et al.: Measurements of single-molecule bond rupture forces between self-assembled monolayers of organosilanes with the atomic force microscope, Langmuir **13** (1997) 3761–3768
19.122 T. Han, J. M. Williams, T. P. Beebe: Chemical bonds studied with functionalized atomic force microscopy tips, Anal. Chim. Acta **307** (1995) 365–376
19.123 V. Tsukruk, V. N. Bliznyuk: Adhesive and friction forces between chemically modified silicon and silicon nitride surfaces, Langmuir **14** (1998) 446–455
19.124 D.V. Vezenov, A.V. Zhuk, G.M. Whitesides, C.M. Likeber: Chemical force spectroscopy in heterogeneous systems: Intermolecular interactions involving epoxy polymer, mixed monolayers, and polar solvents, J. Am. Chem. Soc. **124** (2002) 10578–10588
19.125 J. E. Houston, H. I. Kim: Adhesion, friction, and mechanical properties of functionalized alkanethiol self-assembled monolayers, Accounts Chem. Res. **35** (2002) 547–553
19.126 H. I. Kim, M. Graupe, O. Oloba, T. Koini et al.: Molecularly specific studies of the frictional properties of monolayer films: A systematic comparison of CF_3-, $(CH3)_2(CH)$-, and CH_3- terminated films, Langmuir **15** (1999) 3179–3185
19.127 S. Lee, Y.S. Shon, R. Colorado, R.L. Guenard et al.: The influence of packing densities and surface order on the frictional properties of alkanethiol self-assembled monolayers (SAMs) on gold: A comparison of SAMs derived from normal and spiroalkanedithiols, Langmuir **16** (2000) 2220–2224
19.128 Y. Leng, S. Jiang: Dynamic simulations of adhesion and friction in chemical force microscopy, J. Am. Chem. Soc. **124** (2002) 11764–11770
19.129 H. Skulason, C. D. Frisbie: Detection of discrete interactiuons upon rupture of Au microcontacts to self-assembled monolayers terminated with $-S(CO)CH_3$ or $-SH$, J. Am. Chem. Soc. **122** (2000) 9750–9760
19.130 J. Zhang, J. Kirkham, C. Robinson, M. L. Wallwork et al.: Determination of the ionization state of 11-thioundecyl-1-phosphonic acid in self-assembled monolayers by chemical force microscopy, Anal. Chem. **72** (2000) 1973–1978
19.131 A. Lio, D. H. Charych, M. Salmeron: Comparative atomic force microscopy study of the chain length dependence of frictional properties of alkanethiols on gold and alkylsilanes on mica, J. Phys. Chem. **B 101** (1997) 3800
19.132 A. Lio, C. Morant, D. F. Ogletree, M. Salmeron: Atomic force microscopy study of the pressure-dependent structural and frictional properties of n-alkanethiols on gold, J. Phys. Chem. B **101** (1997) 4767–4773
19.133 G.-Y. Liu, M. Salmeron: Reversible displacement of chemisorbed n-alkane thiol molecules on Au(111) surface: An atomic force microscopy study, Langmuir **10** (1994) 367–370
19.134 X. Xiao, J. Hu, D. H. Charych, M. Salmeron: Chain length dependence of the frictional properties of alkylsilane molecules self-assembled on mica studied by atomic force microscopy, Langmuir **12** (1996) 235
19.135 A. B. Tutein, S. J. Stuart, J. A. Harrison: Indentation analysis of linear-chain hydrocarbon monolayers anchored to diamond, J. Phys. Chem. **B 103** (1999) 11357
19.136 R. L. Pizzolatto, Y. J. Yang, L. K. Wolf, M. C. Messmer: Conformational aspects of model chromatographic surfaces studied by sum-frequency generation, Anal. Chim. Acta **397** (1999) 81
19.137 E.W. v.d. Vegte, A. Subbotin, G. Hadziioannou: Nanotribological properties of unsymmetrical n-dialkyl sulfide monolayers on gold: Effect of chain length on adhesion, friction and imaging, Langmuir **16** (2000) 3249–3256
19.138 Y.-S. Lo, N. D. Huefner, W. S. Chan, F. Stevebs et al.: Specific interactions between biotin and avidin studies by atomic force microscopy using the poisson statistical analysis method, Langmuir **15** (1999) 1373–1382
19.139 Y.-S. Lo, J. Simons, T. P. Beebe, Jr.: Temperature dependence of the biotin–avidin bond rupture force studied by atomic force microscopy, J. Phys. Chem. **B 106** (2002) 9847–9857
19.140 J. H. Hoh, J. P. Cleavland, C. B. Prater, J.-P. Revel et al.: Quantitized adhesion detected with the atomic force microscope, J. Am. Chem. Soc. **114** (1992) 4917–4918
19.141 M. Grandbois, M. Beyer, M. Rief, H. Clausen-Schaumann et al.: How strong is a covalent bond?, Science **283** (1999) 1727–1730
19.142 H. Skulason, C. D. Frisbie: Contact mechanics modeling of pull-off measurements: Effect of solvent, probe radius, and chemical binding probability on the detection of single-bond rupture forces by atomic force microscopy, Anal. Chem. **74** (2002) 3096–3104
19.143 H. Schonherr, V. Chechik, C.J.M. Stirling, G.J. Vancso: Monitoring surface reactions at an AFM tip: An approach to following reaction kinetics in self-assembled monolayers on the nanometer scale, J. Am. Chem. Soc. **122** (2000) 3679–3687

19.144 M. P. L. Werts, E. W. v.d. Vegte, G. Hadziioannou: Surface chemical reactions probed with scanning force microscopy, Langmuir **13** (1997) 4939–4942

19.145 S. S. Wong, E. Joselevich, A. T. Woolley, C. Chin Li et al.: Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology, Nature **394** (1998) 52–55

19.146 S. S. Wong, A. T. Woolley, E. Joselevich, C. L. Cheung et al.: Covalently-functionalized single-walled carbon nanotube tips for chemical force microscopy, J. Am. Chem. Soc. **120** (1998) 8557–8558

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

20. Friction and Wear on the Atomic Scale

Friction is an old subject of research: the empirical da Vinci–Amontons laws are common knowledge. Macroscopic experiments systematically performed by the school of *Bowden* and *Tabor* have revealed that macroscopic friction can be related to the collective action of small asperities. During the last 15 years, experiments performed with the atomic force microscope gave new insight into the physics of single asperities sliding over surfaces. This development, together with complementary experiments by means of surface force apparatus and quartz microbalance, established the new field of nanotribology. At the same time, increasing computing power allowed for the simulation of the processes in sliding contacts consisting of several hundred atoms. It became clear that atomic processes cannot be neglected in the interpretation of nanotribology experiments. Experiments on even well-defined surfaces directly revealed atomic structures in friction forces. This chapter will describe friction force microscopy experiments that reveal, more or less directly, atomic processes in the sliding contact.

We will begin by introducing friction force microscopy, including the calibration of cantilever force sensors and special aspects of the ultra-high vacuum environment. The empirical Tomlinson model largely describes atomic stick-slip results and is therefore presented in detail. We review experimental results regarding atomic friction. These include thermal activation, velocity dependence, as well as temperature dependence. The geometry of the contact plays a crucial role in the interpretation of experimental results, as we will demonstrate, for example, for the calculation of the lateral contact stiffness. The onset of wear on atomic scale has recently come into the scope of experimental studies and is described here. In order to compare the respective results, we present molecular dynamics simulations that are directly related to atomic friction experiments. We close the chapter with a discussion of dissipation measurements performed in noncontact force microscopy, which may become an important complementary tool for the study of mechanical dissipation in nanoscopic devices.

20.1 **Friction Force Microscopy in Ultra-High Vacuum** 632
20.1.1 Friction Force Microscopy 632
20.1.2 Force Calibration 632
20.1.3 The Ultra-high Vacuum Environment 635
20.1.4 A Typical Microscope in UHV 635

20.2 **The Tomlinson Model** 636
20.2.1 One-dimensional Tomlinson Model 636
20.2.2 Two-dimensional Tomlinson Model 637
20.2.3 Friction Between Atomically Flat Surfaces ... 637

20.3 **Friction Experiments on Atomic Scale** 638
20.3.1 Anisotropy of Friction 642

20.4 **Thermal Effects on Atomic Friction** 642
20.4.1 The Tomlinson Model at Finite Temperature 642
20.4.2 Velocity Dependence of Friction 644
20.4.3 Temperature Dependence of Friction 645

20.5 **Geometry Effects in Nanocontacts** 646
20.5.1 Continuum Mechanics of Single Asperities 646
20.5.2 Load Dependence of Friction 647
20.5.3 Estimation of the Contact Area 647

20.6 **Wear on the Atomic Scale** 649
20.6.1 Abrasive Wear on the Atomic Scale . 649
20.6.2 Wear Contribution to Friction 650

20.7 **Molecular Dynamics Simulations of Atomic Friction and Wear** 651
20.7.1 Molecular Dynamics Simulation of Friction Processes 651
20.7.2 Molecular Dynamics Simulations of Abrasive Wear 652

20.8 **Energy Dissipation in Noncontact Atomic Force Microscopy** .. 654

20.9 **Conclusion** 656

References 657

20.1 Friction Force Microscopy in Ultra-High Vacuum

The *friction force microscope* (FFM, also called lateral force microscope, LFM) exploits the interaction of a sharp tip sliding on a surface to quantify dissipative processes down to the atomic scale (Fig. 20.1).

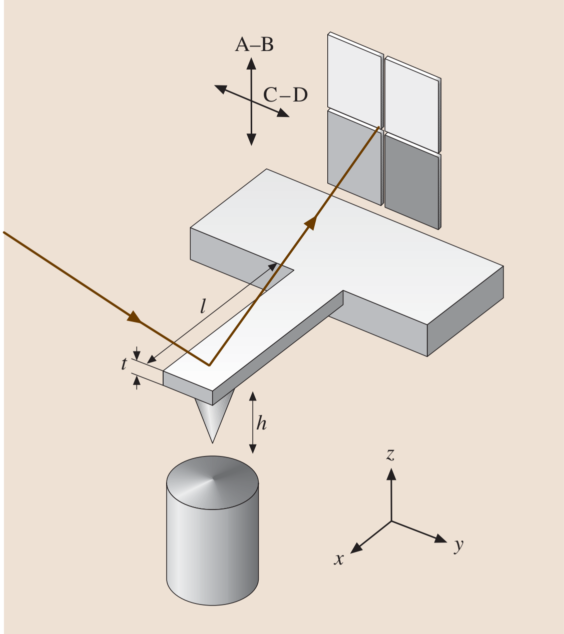

Fig. 20.1 Schematic diagram of a beam-deflection friction force microscope

20.1.1 Friction Force Microscopy

The relative motion of tip and surface is realized by a *scanner* formed by piezoelectric elements, which moves the surface perpendicularly to the tip with a certain periodicity. The scanner can be also extended or retracted in order to vary the normal force, F_N, which is applied on the surface. This force is responsible for the deflection of the *cantilever* that supports the tip. If the normal force F_N increases while scanning due to the local slope of the surface, the scanner is retracted by a feedback loop. On the other hand, if F_N decreases, the surface is brought closer to the tip by extending the scanner. In such way, the surface topography can be determined line by line from the vertical displacement of the scanner. An accurate control of such vertical movement is made possible by a light beam, which is reflected from the rear of the lever into a photodetector. When the bending of the cantilever changes, the light spot on the detector moves up or down and causes a variation of the photocurrent that corresponds to the normal force F_N to be controlled.

Usually the relative sliding of tip and surface is also accompanied by *friction*. A lateral force, F_L, with the opposite direction of the scan velocity, v, hinders the motion of the tip. This force provokes the torsion of the cantilever, and it can be observed with the topography if the photodetector can reveal not only the normal deflection but also the lateral movement of the lever while scanning. In practice this is realized by four-quadrants photodetectors, as shown in Fig. 20.1. We should notice that friction forces also cause the lateral bending of the cantilever, but this effect is negligible if the thickness of the lever is much less than the width.

The FFM was first used by *Mate* et al. in 1987 to reveal friction with atomic features [20.1], i. e. just one year after *Binnig*, *Quate*, and *Gerber* introduced the atomic force microscope [20.2]. In their experiment, Mate used a tungsten wire and a slightly different technique to detect lateral forces (non-fiber interferometry). The optical beam deflection was introduced later by *Marti* et al., and *Meyer* et al. [20.3, 4]. Other methods to measure the forces between tip and surface are given by capacitance detection [20.5], dual fiber interferometry [20.6], and piezolevers [20.7]. In the first method, two plates close to the cantilever reveal capacitance while scanning. The second technique uses two optical fibers to detect the cantilever deflection along two orthogonal directions angled 45° with respect to the surface normal. Finally, in the third method, cantilevers with two Wheatstone bridges at their base reveal normal and lateral forces, which are respectively proportional to the sum and the difference of both bridge signals.

20.1.2 Force Calibration

The force calibration is relatively simple if rectangular cantilevers are used. Due to possible discrepancies with the geometric values provided by manufacturers, one should use optical and electron microscopes to determine the width, thickness, and length of cantilever, w, t, l, the tip height, h, and the position of the tip with respect to the cantilever. The thickness of the cantilever can also be determined from the resonance frequency of the lever, f_0, using the relation [20.8]:

$$t = \frac{2\sqrt{12}\pi}{1.875^2}\sqrt{\frac{\rho}{E}}f_0 l^2 , \tag{20.1}$$

In (20.1) ρ is the density of the cantilever and E is its Young's modulus. The normal spring constant, c_N, and the lateral spring constant, c_L, of the lever are given by

$$c_N = \frac{Ewt^3}{4l^3} , \quad c_L = \frac{Gwt^3}{3h^2 l} , \tag{20.2}$$

where G is the shear modulus. Figure 20.2 shows some SEM images of rectangular silicon cantilevers used for FFM. In the case of silicon, $\rho = 2.33 \times 10^3$ kg/m^3, $E = 1.69 \times 10^{11}$ N/m^2 and $G = 0.5 \times 10^{11}$N/m^2. Thus, for the cantilever in Fig. 20.2, $c_N = 1.9$ N/m and $c_L = 675$ N/m.

The next step in force calibration consists of measuring the sensitivity of the photodetector, S_z (nm/V). For beam-deflection FFMs, the sensitivity S_z can be determined by force vs. distance curves measured on hard surfaces (e.g., Al_2O_3), where elastic deformations are negligible and the vertical movement of the scanner equals the deflection of the cantilever. A typical relation between the difference of the vertical signals on the four-quadrant detector, V_N, and the distance from the surface, z, is sketched in Fig. 20.3. When the tip is approached, no signal is revealed until the tip jumps into contact at $z = z_1$. Further extension or retraction of the scanner results in an elastic behavior until the tip jumps again out of contact at a distance $z_2 > z_1$. The slope of the elastic part of the curve gives the required sensitivity, S_z.

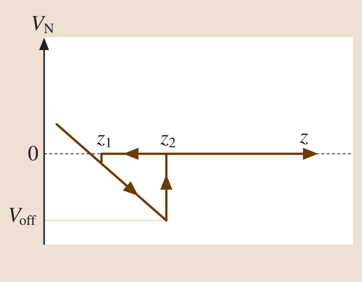

Fig. 20.3 Sketch of a typical force vs. distance curve

The normal and lateral forces are related to the voltage V_N, and the difference between the horizontal signals, V_L, as follows:

$$F_N = c_N S_z V_N , \quad F_L = \frac{3}{2} c_L \frac{h}{l} S_z V_L , \tag{20.3}$$

in which it is assumed that the light beam is positioned above the probing tip.

The normal spring constant c_N can also be calibrated with alternative methods. *Cleveland* et al. [20.10] attached tungsten spheres to the tip, which changes the resonance frequency f_0 according to the formula

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{c_N}{M + m^*}} . \tag{20.4}$$

M is the mass of the added object, and m^* is an effective mass of the cantilever, which depends on its geometry [20.10]. The spring constant can be extrapolated from the frequency shifts corresponding to the different masses attached.

As an alternative, *Hutter* et al. observed that the spring constant c_N can be related to the area of the power spectrum of the thermal fluctuations of the cantilever, P [20.11]. The correct relation is $c_N = 4k_B T/3P$, where $k_B = 1.38 \times 10^{-3}$ J/K is the Boltzmann's constant and T is the temperature [20.12].

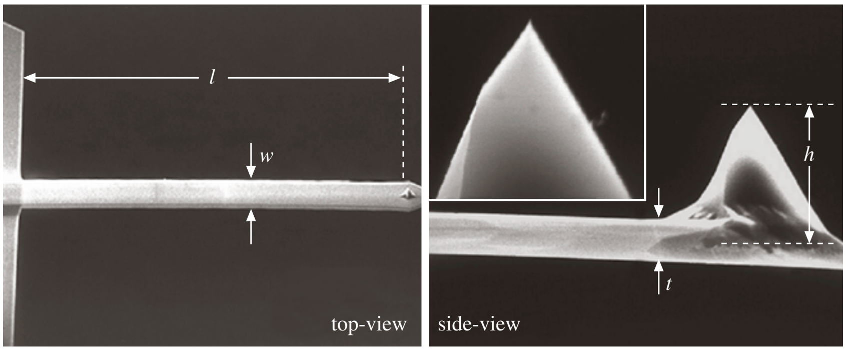

Fig. 20.2 SEM images of a rectangular cantilever. The relevant dimensions are $l = 445$ µm, $w = 43$ µm, $t = 4.5$ µm, $h = 14.75$ µm. Note that h is given by the sum of the tip height and half of the cantilever thickness. (After [20.9])

Cantilevers with different shape require finite element analysis, although in few cases analytical formulas can be derived. For V-shaped cantilevers, *Neumeister* et al. found the following approximation for the lateral spring constant c_{L} [20.13]:

$$c_{\mathrm{L}}=\frac{Et^3}{3(1+\nu)h^2}\times\left(\frac{1}{\tan\alpha}\ln\frac{w}{d\sin\alpha}+\frac{L\cos\alpha}{w}-\frac{3\sin 2\alpha}{8}\right)^{-1},\qquad(20.5)$$

where the geometrical quantities L, w, α, d, t, and h are defined in Fig. 20.4. The expression for the normal constant is more complex and can be found in the cited reference.

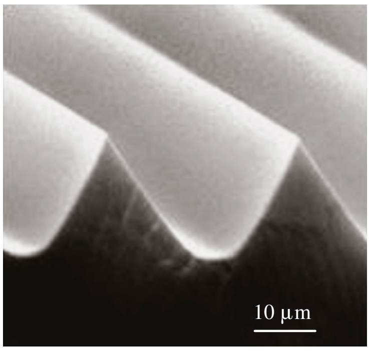

Fig. 20.5 Silicon grating formed by alternated faces angled ±55° (courtesy Silicon-MDT Ltd., POB 50, 103305, Moscow, Russia)

Surfaces with well-defined profiles provide an alternative in situ calibration of lateral forces [20.14]. We present a slightly modified version of the method [20.15]. Figure 20.5 shows a commercial grating formed by alternated faces with opposite inclinations with respect to the scan direction. When the tip slides on the inclined planes, the normal force, F_{N}, and the lateral force, F_{L}, with respect to the surface, are different from the two components $F_{\perp}$ and $F_{\parallel}$, which are separated by the photodiode (see Fig. 20.6a).

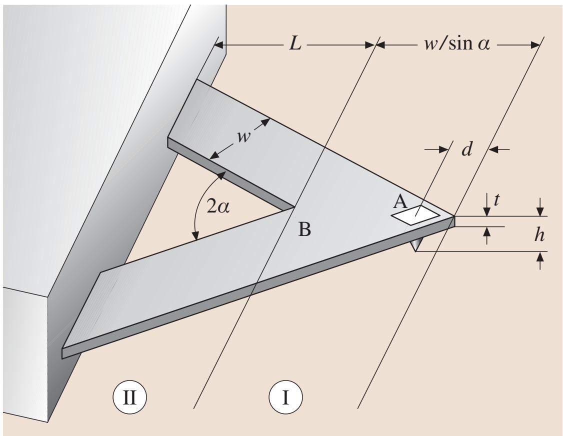

Fig. 20.4 Geometry of a V-shaped cantilever. (After [20.13])

If the linear relation $F_{\mathrm{L}}=\mu F_{\mathrm{N}}$ holds (see Sect. 20.5), the component $F_{\parallel}$ can be expressed in terms of $F_{\perp}$:

$$F_{\parallel}=\frac{\mu+\tan\theta}{1-\mu\tan\theta}F_{\perp}\,.\qquad(20.6)$$

The component $F_{\perp}$ is kept constant by the feedback loop. The sum and the difference of $F_{\parallel}$ on the two planes (1) and (2) are given by

$$F_{+}\equiv F_{\parallel}^{(1)}+F_{\parallel}^{(2)}=\frac{2\mu\left(1+\tan^2\theta\right)}{1-\mu^2\tan^2\theta}F_{\perp}$$

$$F_{-}\equiv F_{\parallel}^{(1)}-F_{\parallel}^{(2)}=\frac{2\left(1+\mu^2\right)\tan\theta}{1-\mu^2\tan^2\theta}F_{\perp}\,.\qquad(20.7)$$

The values of F_{+} and F_{-} (in volts) can be measured by scanning the profile back and forth (Fig. 20.6b). If F_{+} and F_{-} are recorded with different values of $F_{\perp}$, one can determine the conversion ratio between volts and nanonewtons as well as the coefficient of friction, μ.

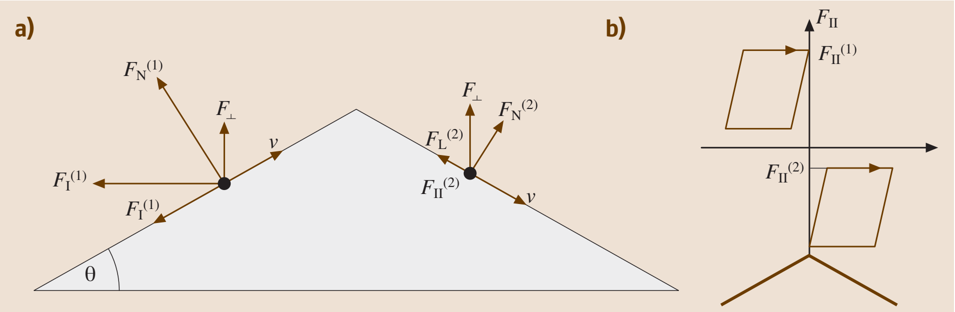

Fig. 20.6 **(a)** Forces acting on a FFM tip sliding on the grating in Fig. 20.5; **(b)** Friction loops acquired on the two faces

An accurate analysis of the errors in lateral force calibration was given by *Schwarz* et al., who revealed the important role of the cantilever's oscillations induced by the feedback loop and of the so-called pull-off force (Sect. 20.5) in friction measurements, aside from the geometrical setting of cantilevers and laser beams [20.16].

For some applications, an adequate estimation of the radius of curvature of the tip, R, is important (Sect. 20.5.2). This quantity can be evaluated with a scanning electron microscope. Otherwise, well-defined structures as step sites [20.17, 18] or whiskers [20.19] can be imaged. The image of these high-aspect ratio structures is a convolution with the tip structure. A deconvolution algorithm that allows for the extraction of the probing tip's radius of curvature was suggested by *Villarrubia* [20.21].

20.1.3 The Ultra-high Vacuum Environment

Atomic friction studies require well-defined surfaces and – whenever possible – tips. For the surfaces, the established methods of surface science can be employed working in ultra-high vacuum (UHV). Ionic crystals such as NaCl have become standard materials for friction force microscopy on atomic scale. Atomically clean and flat surfaces can be prepared by cleavage in UHV. The crystal has to be heated to about 150 °C for 1 h in order to remove charging after the cleavage process. Metal surfaces can be cleaned and flattened by cycles of sputtering with argon ions and annealing. Even surfaces prepared in air or liquids like self-assembled molecular monolayers can be transferred into the vacuum and studied after careful heating procedures, which remove water layers.

Tip preparation in UHV is a more difficult subject. Most force sensors for friction studies have silicon nitride or pure silicon tips. Tips can be cleaned and oxide layers removed by sputtering with argon ions. But the sharpness of the tip is normally reduced by sputtering. As an alternative, tips can be etched in fluoric acid directly before transfer to the UHV. The significance of tip preparation is limited by the fact that the chemical and geometrical structure of the tip can undergo significant changes when sliding over the surface.

The implementation of the friction force microscope into UHV requires some additional efforts. First of all, only materials with a low vapor pressure can be used, thereby excluding most plastics and lubricants. Beam-deflection type force microscopes employ either a light source in the vacuum chamber or an optical fiber guiding the light into the chamber. The positioning of the light beam onto the cantilever and of the reflected beam onto the position-sensitive detector is performed by motorized mirrors [20.20] or by moving light source or detector [20.22]. Furthermore, a motorized sample approach has to be realized.

The quality force sensor's electrical signal can seriously deteriorate on the way out of the vacuum chamber. Low noise and high bandwidth can be preserved by implementing a pre-amplifier into the vacuum. Again the choice of materials for print and devices is limited by the condition of low vapor pressure. Stronger heating of the electrical circuitry in vacuum, therefore, must be considered.

20.1.4 A Typical Microscope in UHV

A typical AFM for UHV application is shown in Fig. 20.7. The housing (1) contains the light source and a set of lenses that focus the light onto the cantilever. Alternatively, the light can be guided via an optical fiber into the vacuum. By using light emitting diodes with low coherency you avoid disturbing interference effects often found in instruments equipped with lasers as light source. A plane mirror fixed on the spherical rotor of a first stepping motor (2) can be rotated

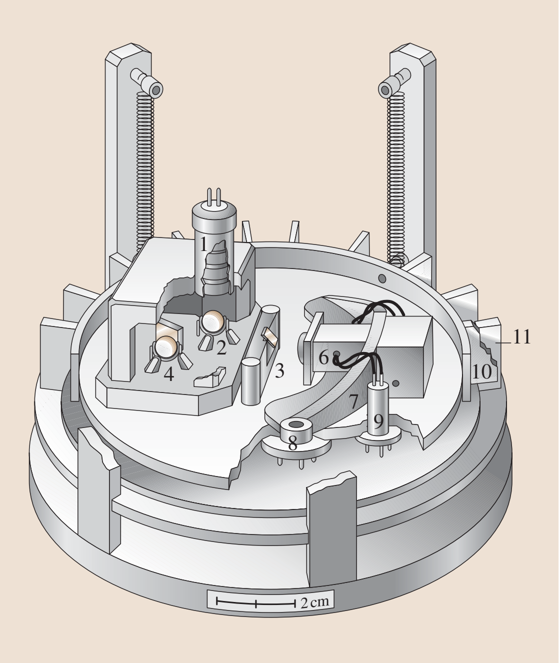

Fig. 20.7 Schematic view of the UHV-AVM realized at the University of Basel. (After [20.20])

around a vertical and horizontal axis to guide the light beam onto the rear of the cantilever, mounted on a removable carrier plate (3). The light is reflected off the cantilever toward a second motorized mirror (4) that guides the beam to the center of the quadrant photodiode (5), where the light is then converted into four photo-currents. Four pre-amplifiers in close vicinity to the photodiode allow low-noise measurements with 3 MHz bandwidth.

The two motors with spherical rotors, used to realign the light path after exchange of the cantilever, work as *inertial stepping motors*: the sphere is resting on three piezoelectric legs that can be moved by a small amount tangentially to the sphere. Each step of the motor consists of a slow forward motion of two legs followed by an abrupt jump backwards. During the slow forward motion the sphere follows the legs due to friction, whereas it cannot follow the sudden jump due to its inertia. A series of such tiny steps rotates the sphere macroscopically.

The sample, also on an exchangeable carrier plate, is mounted on the end of a tube scanner (6), which can move the sample in three dimensions over several micrometers. The whole scanning head (7) is the slider of a third inertial stepping motor for coarse positioning of the sample. It rests with its flat and polished bottom on three supports. Two of them are symmetrically placed piezoelectric legs (8), whereas the third central support is passive. The slider (7) can be translated in two dimensions and rotated about a vertical axis by several millimeters (rotation is achieved by antiparallel operation of the two legs). The slider is held down by two magnets, close to the active supports, and its travel is limited by two fixed posts (9) that also serve as cable attachments. The whole platform is suspended by four springs. A ring of radial copper lamellae (10), floating between a ring of permanent magnets (11) on the base flange, acts as efficient eddy current damping.

20.2 The Tomlinson Model

In Sect. 20.3, we will show that the FFM can reveal friction forces down to the atomic scale, which are characterized by a typical sawtooth pattern. This phenomenon can be seen as a consequence of a *stick-slip* mechanism, first discussed by *Tomlinson* in 1929 [20.23].

20.2.1 One-dimensional Tomlinson Model

In the Tomlinson model the motion of the tip is influenced by both the interaction with the atomic lattice of the surface and the elastic deformations of the cantilever. The shape of the tip-surface potential, $V(\boldsymbol{r})$, depends on several factors, such as the chemical composition of the materials in contact and the atomic arrangement at the tip end. For the sake of simplicity, we will start the analysis in the one-dimensional case considering a sinusoidal profile with the periodicity of the atomic lattice, a, and a peak-to-peak amplitude E_0. In Sect. 20.5, we will show how the elasticity of the cantilever and of the contact area are described in a unique framework by introducing an effective lateral spring constant, k_{eff}. If the cantilever moves with a constant velocity v along the x direction, the total energy of the system is

$$E_{\text{tot}}(x,t) = -\frac{E_0}{2}\cos\frac{2\pi x}{a} + \frac{1}{2}k_{\text{eff}}(vt-x)^2 \,. \quad (20.8)$$

Figure 20.8 shows the energy profile $E_{\text{tot}}(x,t)$ at two different instants. When $t=0$ the tip is localized in the absolute minimum of E_{tot}. This minimum increases with time due to the cantilever motion, until the tip position becomes unstable when $t=t^*$.

At a given time t the position of the tip can be determined by equating to zero the first derivative of $E_{\text{tot}}(x,t)$ with respect to x:

$$\frac{\partial E_{\text{tot}}}{\partial x} = \frac{\pi E_0}{a}\sin\frac{2\pi x}{a} - k_{\text{eff}}(vt-x) = 0 \,. \quad (20.9)$$

The critical position x^* corresponding to $t=t^*$ is determined by equating to zero the second derivative

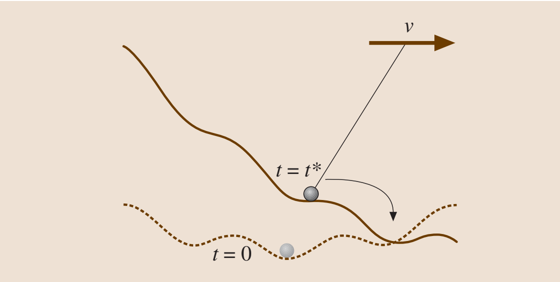

Fig. 20.8 Energy profile experienced by the FFM tip (*black circle*) at $t=0$ (*dotted line*) and $t=t^*$ (*continuous line*)

$\partial^2 E_{\text{tot}}(x, t)/\partial x^2$, which gives

$$x^* = \frac{a}{4} \arccos\left(-\frac{1}{\gamma}\right), \quad \gamma = \frac{2\pi^2 E_0}{k_{\text{eff}} a^2} . \tag{20.10}$$

The coefficient γ compares the strength of interaction between tip and surface with the stiffness of the system. When $t = t^*$ the tip suddenly *jumps* into the next minimum of the potential profile. The lateral force $F^* = k_{\text{eff}}(vt - x^*)$, which induces the jump, can be evaluated from (20.9) and (20.10):

$$F^* = \frac{k_{\text{eff}}\, a}{2\pi} \sqrt{\gamma^2 - 1} . \tag{20.11}$$

Thus the stick-slip is observed only if $\gamma > 1$, i. e. when the system is not too stiff or the tip–surface interaction is strong enough. Figure 20.9 shows the lateral force F_L as a function of the cantilever position, X. When the cantilever is moved rightward, the lower part of the curve in Fig. 20.9 is obtained. If, at a certain point, the cantilever's direction of motion is suddenly inverted, the force has the profile in the upper part of the curve. The area of the *friction loop* obtained by scanning back and forth gives the total energy dissipated.

20.2.2 Two-dimensional Tomlinson Model

In two dimensions, the energy of our system is given by

$$E_{\text{tot}}(\boldsymbol{r}, t) = U(\boldsymbol{r}) + \frac{k_{\text{eff}}}{2} (\boldsymbol{v}t - \boldsymbol{r})^2 , \tag{20.12}$$

where $\boldsymbol{r} \equiv (x, y)$ and $\boldsymbol{v}$ is arbitrarily oriented on the surface (note that $\boldsymbol{v} \neq \mathrm{d}\boldsymbol{r}/\mathrm{d}t$!). Figure 20.10 shows the total energy corresponding to a periodic potential of the form

$$U(x, y, t) = -\frac{E_0}{2}\left(\cos\frac{2\pi x}{a} + \cos\frac{2\pi y}{a}\right) + E_1 \cos\frac{2\pi x}{a} \cos\frac{2\pi y}{a} . \tag{20.13}$$

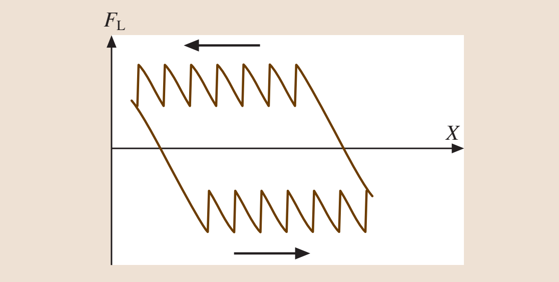

Fig. 20.9 Friction loop obtained by scanning back and forth in the 1-D Tomlinson model. The effective spring constant k_{eff} is the slope of the sticking part of the loop (if $\gamma \gg 1$)

The equilibrium condition becomes

$$\nabla E_{\text{tot}}(\boldsymbol{r}, t) = \nabla U(\boldsymbol{r}) + k_{\text{eff}}(\boldsymbol{r} - \boldsymbol{v}t) = 0 . \tag{20.14}$$

The stability of the equilibrium can be discussed introducing the Hessian matrix

$$H = \begin{pmatrix} \dfrac{\partial^2 U}{\partial x^2} + k_{\text{eff}} & \dfrac{\partial^2 U}{\partial x \partial y} \\ \dfrac{\partial^2 U}{\partial y \partial x} & \dfrac{\partial^2 U}{\partial y^2} + k_{\text{eff}} \end{pmatrix} . \tag{20.15}$$

When both eigenvalues $\lambda_{1,2}$ of the Hessian are positive, the position of the tip is stable. Figure 20.11 shows such regions for a potential of the form (20.13). The tip follows the cantilever adiabatically as long as it remains in the (++)-region. When the tip is dragged to the border of the region, it suddenly jumps into the next (++)-region. A comparison between a theoretical friction map deduced from the 2-D Tomlinson model and an experimental map acquired by UHV-FFM is given in the next section.

20.2.3 Friction Between Atomically Flat Surfaces

So far we have implicitly assumed that the tip is terminated by only one atom. It is also instructive to consider the case of a periodic surface sliding on another periodic surface. In the Frenkel–Kontorova–Tomlinson (FKT) model, the atoms of one surface are harmonically coupled with their nearest neighbors. We will restrict ourselves to the case of quadratic symmetries, with lattice constants a_1 and a_2 for the upper and lower surface, respectively (Fig. 20.12). In such context the role of *commensurability* is essential. It is well known that any real number z can be represented as a continued

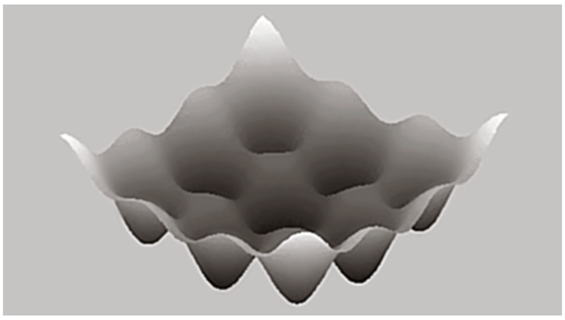

Fig. 20.10 Energy landscape experienced by the FFM tip in 2-D

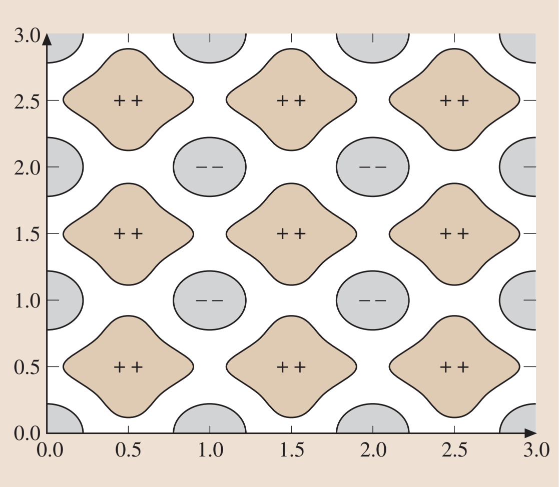

Fig. 20.11 Regions on the tip plane labeled according to the signs of the eigenvalues of the Hessian matrix. (After [20.24])

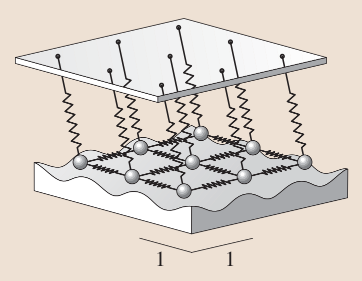

Fig. 20.12 The FKT model in 2-D. (After [20.25])

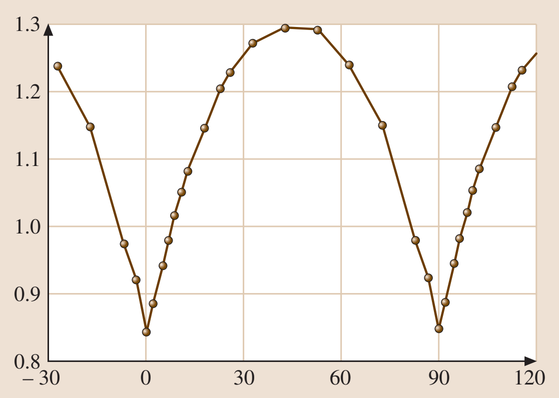

Fig. 20.13 Friction as a function of the sliding angle φ in the 2-D FKT model. (After [20.25])

fraction:

$$z = N_0 + \frac{1}{N_1 + \frac{1}{N_2 + \ldots}} \,. \tag{20.16}$$

The sequence that converges most slowly is obtained when all $N_i = 1$, which corresponds to the *golden mean* $\bar{z} = (\sqrt{5} - 1)/2$. In 1-D *Weiss* and *Elmer* predicted that friction should decrease with decreasing commensurability, the minimum of friction being reached when $a_1/a_2 = \bar{z}$ [20.26].

In 2-D *Gyalog* and *Thomas* studied the case $a_1 = a_2$, with a misalignment between the two lattices given by an angle θ [20.25]. When the sliding direction changes, friction also varies from a minimum value corresponding to the sliding angle $\varphi = \theta/2$ to a maximum value, which is reached when $\varphi = \theta/2 + \pi/4$ (Fig. 20.13). The misfit angle θ is related with commensurability. Since the misfit angles giving rise to commensurate structure form a dense subset, the dependence of friction on θ should be discontinuous. The numerical simulations performed by Gyalog are in agreement with this conclusion.

20.3 Friction Experiments on Atomic Scale

Figure 20.14 shows the first friction map observed by Mate on atomic scale. The periodicity of the lateral force is the same as of the atomic lattice of graphite. The series of friction loops in Fig. 20.15 reveals the stick-slip effect discussed in the previous section. The applied loads are in the range of tens of μN. According to the continuum models discussed in Sect. 20.5 these values correspond to contact diameters of 100 nm. A possible explanation for the atomic features observed at such high loads is that graphite flakes might have been detached from the surface and adhered on the tip [20.27]. Another explanation is that the contact between tip and surface consisted of few nm-scale asperities and the corrugation was not entirely averaged out while sliding. The load dependence of friction found by Mate is rather linear, with a small friction coefficient $\mu = 0.01$ (Fig. 20.16).

The UHV environment reduces the influence of contaminants on the surface and leads to more precise and reproducible results. *Meyer* et al. [20.28] obtained

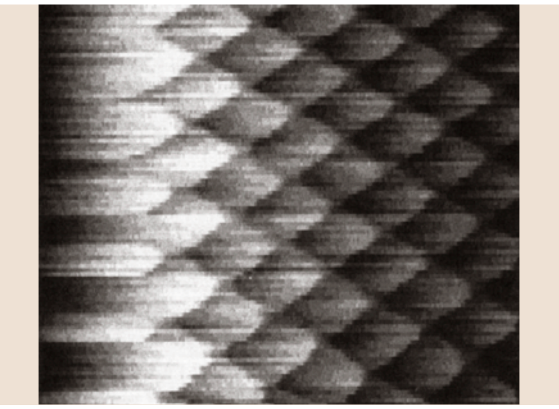

Fig. 20.14 First atomic friction map acquired on graphite with a normal force $F_N = 56\,\mu N$. Frame size: 2 nm. (After [20.1])

a series of interesting results on ionic crystals using the UHV-FFM apparatus described in Sect. 20.1.4. In Fig. 20.17 a friction map recorded on KBr(100) is compared with a theoretical map, which was obtained with the 2-D-Tomlinson model. The periodicity $a = 0.47$ nm corresponds to the spacing between equally charged ions. No individual defects were observed. One possible reason is that the contact realized by the FFM tip is always formed by many atoms, which superimpose and average their effects. Molecular dynamics (MD) calculations (Sect. 20.7) show that even a single-atom contact may cause rather large stresses in the sample, which lead to the motion of defects far away from the contact area. In a picturesque frame, we can say that "defects behave like dolphins that swim away in front of an ocean cruiser" [20.28].

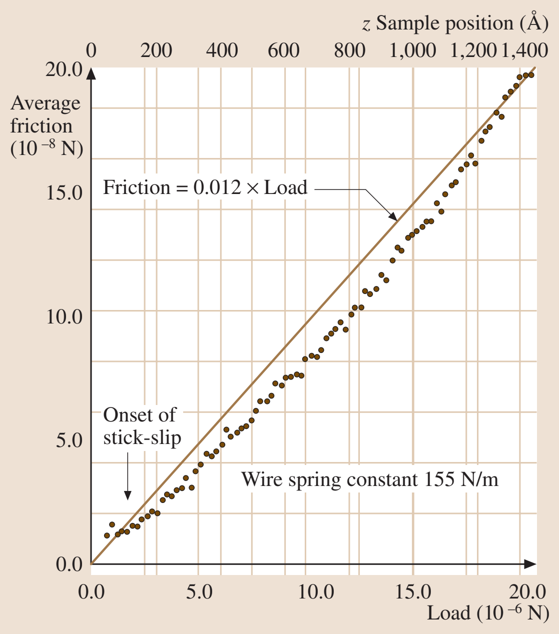

Fig. 20.16 Load dependence of friction on graphite. (After [20.1])

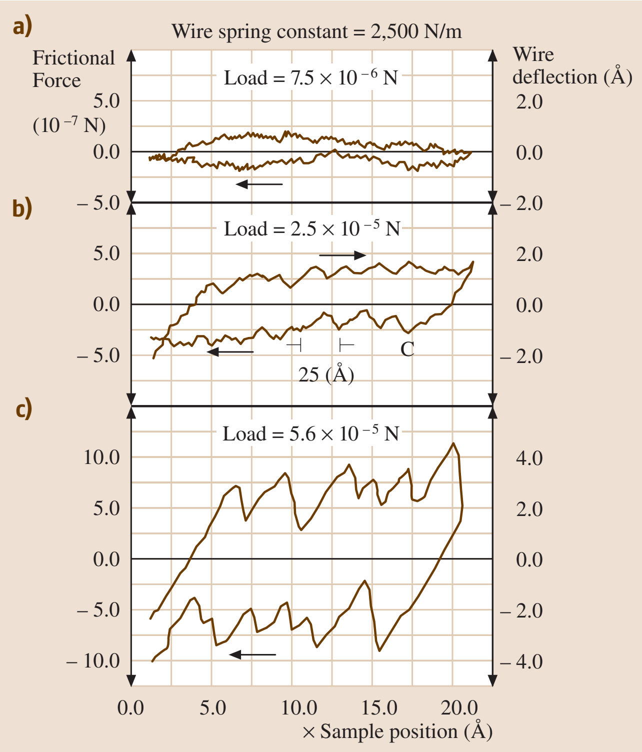

Fig. 20.15a–c Friction loops on graphite acquired with **(a)** $F_N = 7.5\,\mu N$, **(b)** $24\,\mu N$ and **(c)** $75\,\mu N$. (After [20.1])

Lüthi et al. [20.29] detected atomic-scale friction even on the reconstructed Si(111)7×7 surface. But uncoated Si-tips and tips coated with Pt, Au, Ag, Cr, Pt/C damaged the sample irreversibly, and the observation of atomic features was achieved only after coating the tips with polytetrafluoroethylene (PTFE), which has lubricant properties and does not react with the dangling bonds of Si(111)7×7 (Fig. 20.18).

Recently friction could be resolved on atomic scale even on metallic surfaces in UHV [20.30]. In Fig. 20.19a a reproducible stick-slip process on Cu(111) is shown.

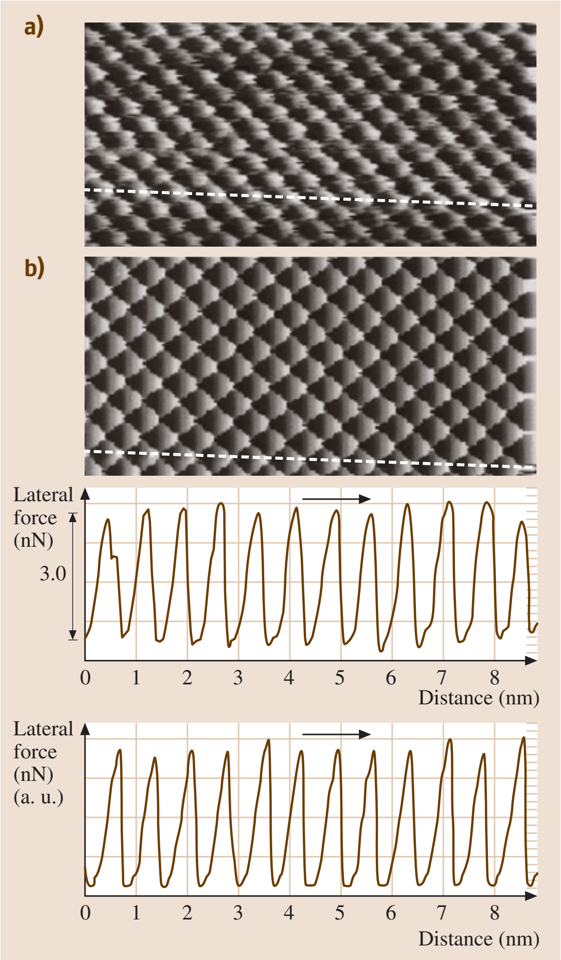

Fig. 20.17 **(a)** Measured and **(b)** theoretical friction map on KBr(100). (After [20.31])

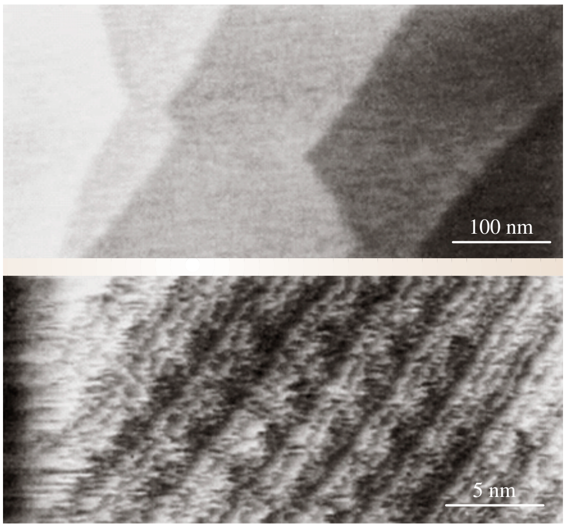

Fig. 20.18 **(a)** Topography and **(b)** friction image of Si(111)7×7 measured with a PTFE coated Si-tip. (After [20.29])

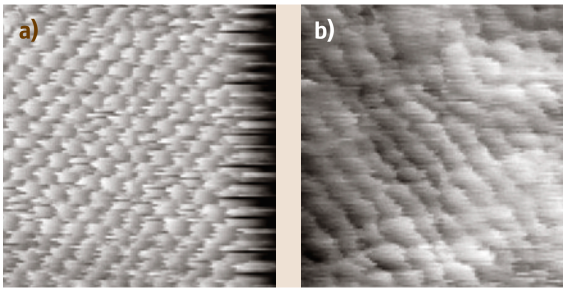

Fig. 20.19a,b Friction images of **(a)** Cu(111) and **(b)** Cu(100). Frame size: 3 nm. (After [20.34])

Sliding on the (100) surface of copper produced irregular patterns, although atomic features were recognized even in this case (Fig. 20.19b). Molecular dynamics suggests that wear should occur more easily on the Cu(100) surface than on the closed packed Cu(111) (Sect. 20.7). This conclusion was achieved by adopting copper tips in computer simulations. The assumption that the FFM tip used in the experiments was covered by copper is supported by current measurements performed at the same time.

Atomic stick-slip on diamond was observed by *Germann* et al. with an apposite diamond tip prepared by chemical vapor deposition [20.32] and, a few years later, by *van der Oetelaar* et al. [20.33] with standard silicon tips. The values of friction have huge discrepancies corresponding to the presence or absence of hydrogen on the surface.

Fujisawa et al. [20.35] measured friction on mica and on MoS_2 with a 2-D-FFM apparatus, which could also reveal forces perpendicular to the scan direction. The features in Fig. 20.20 correspond to a zig-zag walk of the tip, which is predicted by the 2-D-Tomlinson model [20.36]. The 2-D stick-slip on NaF was detected with normal forces below 14 nN, whereas loads up to 10 µN could be applied on layered materials. The contact between tip and NaF was thus formed by one or a few atoms. The zig-zag walk on mica was also observed by *Kawakatsu* et al. using an original 2-D-FFM with two laser beams and two quadrant photodetectors [20.37].

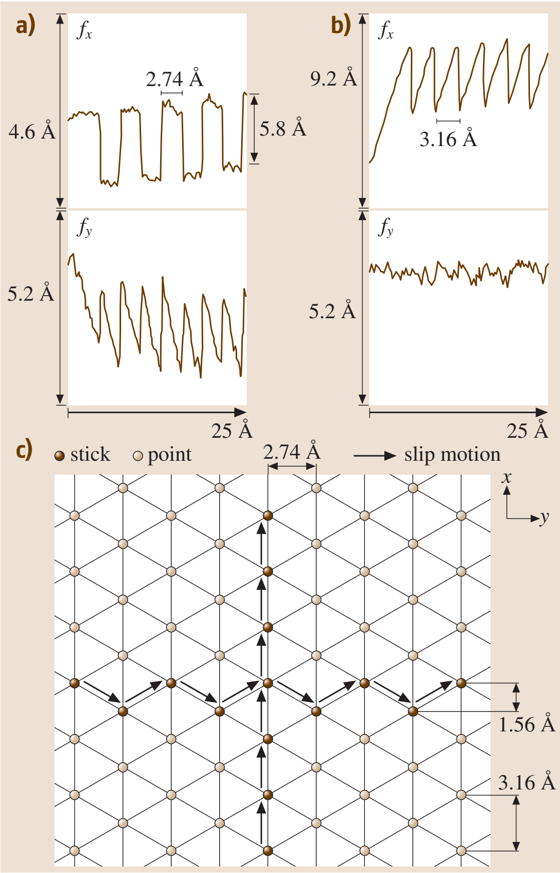

Fig. 20.20 **(a)** Friction force on MoS_2 acquired by scanning along the cantilever and **(b)** across the cantilever. **(c)** Motion of the tip on the sample. (After [20.35])

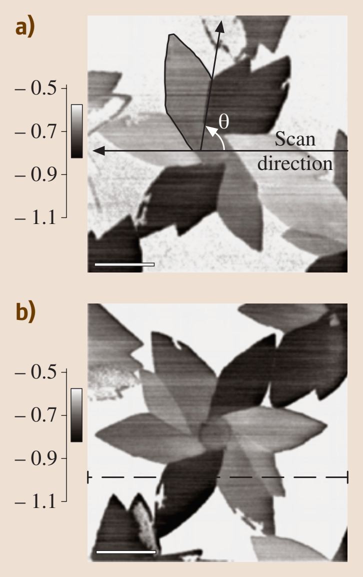
a)
– 0.5
– 0.7
– 0.9
– 1.1
θ
Scan
direction
b)
– 0.5
– 0.7
– 0.9
– 1.1

Fig. 20.21 Friction images of a thiolipid monolayer on a mica surface. (After [20.38])

Fig. 20.22 Sequence of topography images of C_{60} islands on NaCl(100). (After [20.39])

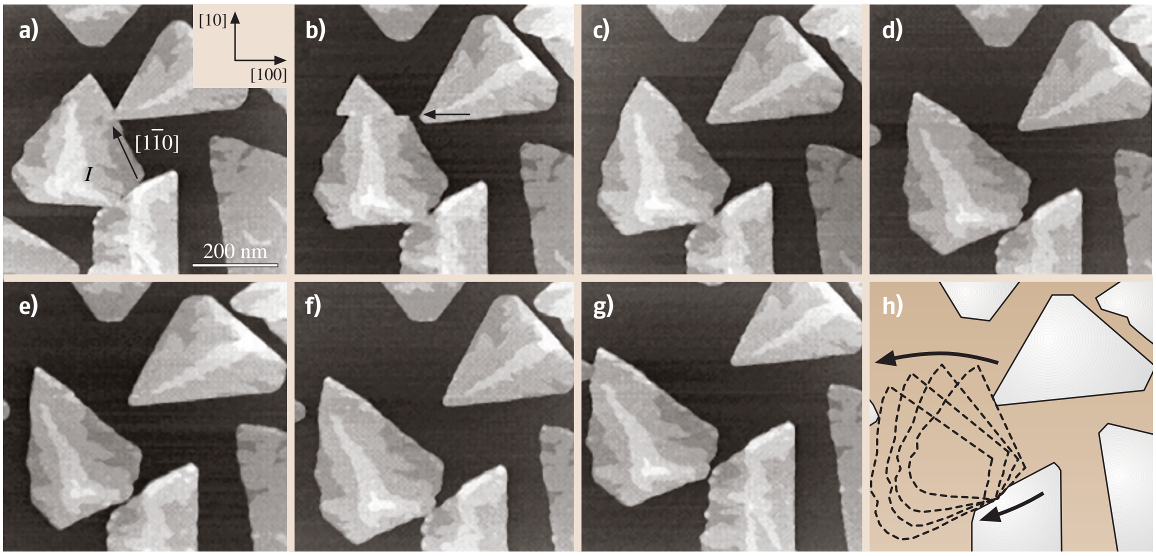
a)
[10]
[100]
[1$\bar{1}$0]
I
200 nm
b)
c)
d)
e)
f)
g)
h)

20.3.1 Anisotropy of Friction

The importance of the misfit angle in the reciprocal sliding of two flat surfaces was first observed experimentally by *Hirano* et al. in the contact of two mica sheets [20.41]. Friction increased when the two surfaces formed commensurate structures, in agreement with the discussion in Sect. 20.2.3. In more recent measurements with a monocrystalline tungsten tip on Si(001), *Hirano* et al. observed *superlubricity* in the incommensurate case [20.42].

Overney et al. [20.43] studied the effects of friction anisotropy on a bilayer lipid film. In such case different molecular alignments resulted in significant variation of friction. Other measurements of friction anisotropy on stearic acid single crystals were reported by *Takano* and *Fujihira* [20.44].

Liley et al. [20.38] observed flower-shaped islands of a lipid monolayer on mica, which consisted of domains with different molecular orientation (Fig. 20.21). The angular dependence of friction reflects the tilt direction of the alkyl chains of the monolayer, which was revealed by other techniques.

Lüthi et al. [20.39] used the FFM tip to move C_{60} islands, which slided on sodium chloride in UHV without disruption (Fig. 20.22). In this experiment friction was found to be independent of the sliding direction. This was not the case in other experiments performed by *Sheehan* and *Lieber*, who observed that the misfit angle is relevant when MoO_3 islands are dragged on the MoS_2 surface [20.45]. In these experiments, sliding was possible only along low index directions. The weak orientation dependence found by *Lüthi* et al. [20.39] is probably due to the large mismatch of C_{60} on NaCl.

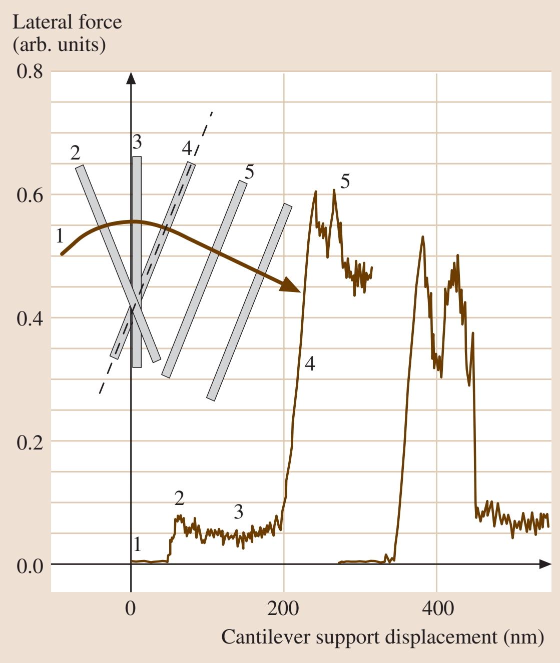

Fig. 20.23 Friction force experienced as a carbon nanotube is rotated into (*left trace*) and out of (*right trace*) commensurate contact. (After [20.40])

A recent example of friction anisotropy is related to carbon nanotubes. *Falvo* et al. [20.40] manipulated nanotubes on graphite using a FFM tip (Fig. 20.23). A dramatic increase of the lateral force was found in the directions corresponding to commensurate contact. At the same time, the nanotube motion changed from sliding/rotating to stick-roll.

20.4 Thermal Effects on Atomic Friction

Although the Tomlinson model gives a good interpretation of the basic mechanism of the atomic stick-slip discussed in Sect. 20.2, it cannot explain some minor features observed in the atomic friction. For example, Fig. 20.24 shows a friction loop acquired on NaCl(100). The peaks in the sawtooth profile have different heights, which is in contrast to the result in Fig. 20.9. Another effect is observed if the scan velocity v is varied: the mean friction force increases with the logarithm of v (Fig. 20.25). This effect cannot be interpreted within the mechanical approach in Sect. 20.2 without further assumptions.

20.4.1 The Tomlinson Model at Finite Temperature

Let us focus again on the energy profile discussed in Sect. 20.2.1. For sake of simplicity, we will assume that $\gamma \gg 1$. At a given time $t < t^*$, the tip jump is prevented by the energy barrier $\Delta E = E(x_{\max}, t) - E(x_{\min}, t)$, where $x_{\max}$ corresponds to the first maximum observed in the energy profile and $x_{\min}$ is the actual position of the tip (Fig. 20.26). The quantity ΔE decreases with time or, equivalently, with the frictional force F_L until it vanishes when $F_L = F^*$ (Fig. 20.27). Close to the critical

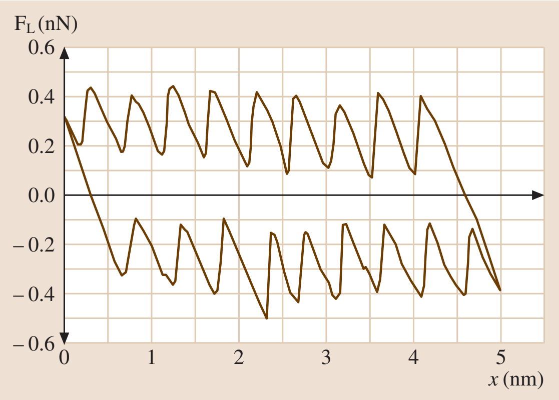

Fig. 20.24 Friction loop on NaCl(100). (After [20.46])

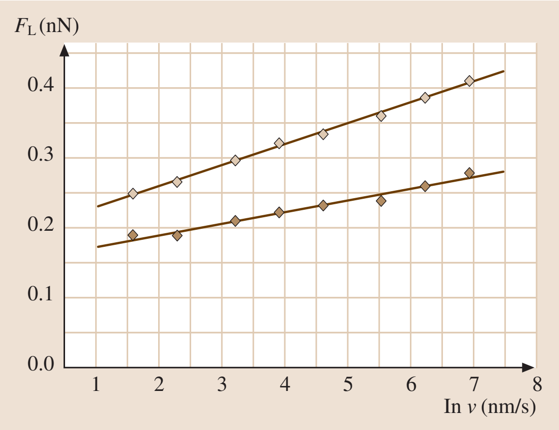

Fig. 20.25 Mean friction force vs. scanning velocity on NaCl(100) at $F_N = 0.44$nN (+) and $F_N = 0.65$nN (×). (After [20.46])

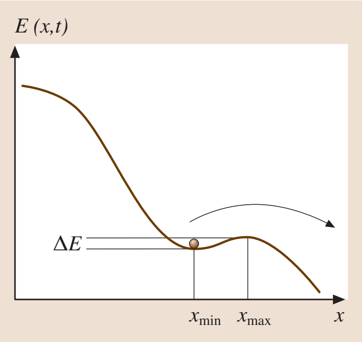

Fig. 20.26 Energy barrier that hinders the tip jump in the Tomlinson model

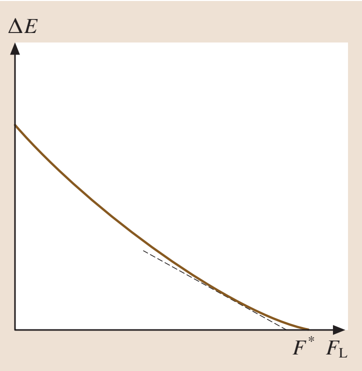

Fig. 20.27 Energy barrier ΔE as a function of the lateral force F_L. The *dashed line* close to the critical value corresponds to the linear approximation (20.17)

point the energy barrier can be written approximately as

$$\Delta E = \lambda(\tilde{F} - F_L) , \qquad (20.17)$$

where $\tilde{F}$ is close to the critical value $F^* = \pi E_0/a$.

At finite temperature T, the lateral force required to induce a jump is lower than F^*. To estimate the most probable value of F_L at this point, we first consider the probability p that the tip does *not* jump. The probability p changes with time t according to the master equation

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = -f_0 \exp\left(-\frac{\Delta E(t)}{k_B T}\right) p(t) , \qquad (20.18)$$

where f_0 is a characteristic frequency of the system. The physical meaning of this frequency is discussed in Sect. 20.4.2. We should note that the probability of a reverse jump is neglected, since in this case the energy barrier to overcome is much higher than ΔE. If time is replaced by the corresponding lateral force, the master equation becomes

$$\frac{\mathrm{d}p(F_L)}{\mathrm{d}F_L} = -f_0 \exp\left(-\frac{\Delta E(F_L)}{k_B T}\right)\left(\frac{\mathrm{d}F_L}{\mathrm{d}t}\right)^{-1} p(F_L) . \qquad (20.19)$$

At this point, we substitute

$$\frac{\mathrm{d}F_L}{\mathrm{d}t} = \frac{\mathrm{d}F_L}{\mathrm{d}X}\frac{\mathrm{d}X}{\mathrm{d}t} = k_{\text{eff}} v \qquad (20.20)$$

and use the approximation (20.17). The maximum probability transition condition $\mathrm{d}^2 p(F)/\mathrm{d}F^2 = 0$ then yields

$$F_L(v) = F^* - \frac{k_B T}{\lambda} \ln \frac{v_c}{v} \qquad (20.21)$$

with

$$v_c = \frac{f_0 k_B T}{k_{\text{eff}} \lambda} . \qquad (20.22)$$

Thus the lateral force depends logarithmically on the sliding velocity, as observed experimentally. But the approximation (20.17) does not hold when the tip jump occurs very close to the critical point $x = x^*$, which is the case at high velocities. In this instance, the factor $(\mathrm{d}F_\mathrm{L}/\mathrm{d}t)^{-1}$ in (20.19) is small and, consequently, the probability $p(t)$ does not change significantly until it suddenly approaches 1 when $t \to t^*$. Thus friction is constant at high velocities, in agreement with the classical Coulomb's law of friction [20.28].

Sang et al. [20.48] observed that the energy barrier close to the critical point is better approximated by a relation like

$$\Delta E = \mu (F^* - F_\mathrm{L})^{3/2} \,. \tag{20.23}$$

The same analysis performed using the approximation (20.23) instead of (20.17) leads to the expression [20.49]

$$\frac{\mu (F^* - F_\mathrm{L})^{3/2}}{k_\mathrm{B} T} = \ln \frac{v_\mathrm{c}}{v} - \ln \sqrt{1 - \frac{F^*}{F_\mathrm{L}}} \,, \tag{20.24}$$

where the critical velocity v_c is now

$$v_\mathrm{c} = \frac{\pi \sqrt{2}}{2} \frac{f_0 k_\mathrm{B} T}{k_\mathrm{eff} a} \,. \tag{20.25}$$

The velocity v_c discriminates between two different regimes. If $v \ll v_\mathrm{c}$ the second logarithm in (20.24) can be neglected, which leads to the logarithmic dependence

$$F_\mathrm{L}(v) = F^* - \left(\frac{k_\mathrm{B} T}{\mu} \right)^{2/3} \left(\ln \frac{v_\mathrm{c}}{v} \right)^{2/3} \,. \tag{20.26}$$

In the opposite case, $v \gg v_\mathrm{c}$, the term on the left in (20.23) is negligible and

$$F_\mathrm{L}(v) = F^* \left(1 - \frac{v_\mathrm{c}}{v} \right)^2 \,. \tag{20.27}$$

In such a case, the lateral force F_L tends to F^*, as expected.

20.4.2 Velocity Dependence of Friction

The velocity dependence of friction was studied by FFM only recently. *Zwörner* et al. observed that friction between silicon tips and diamond, graphite or amorphous carbon is constant with scan velocities of few µm/s [20.50]. Friction decreased when v is reduced below 1 µm/s. In their experiment on lipid films on mica (Sect. 20.3.1), *Gourdon* et al. [20.47] explored a range of velocities from 0.01 to 50 µm/s and found a critical velocity $v_\mathrm{c} = 3.5$ µm/s, which discriminates between an increasing friction and a constant friction regime (Fig. 20.28). Although these results were not explained with thermal activation, we argue that the previous theoretical discussion gives the correct interpretative key. A clear observation of a logarithmic dependence of friction on the micrometer scale was reported by *Bouhacina* et al., who studied friction on triethoxysilane molecules and polymers grafted on silica with sliding velocity up to $v = 300$ µm/s [20.51]. The result was explained with a thermally activated Eyring model, which does not differ significantly from the model discussed in the previous subsection [20.52, 53].

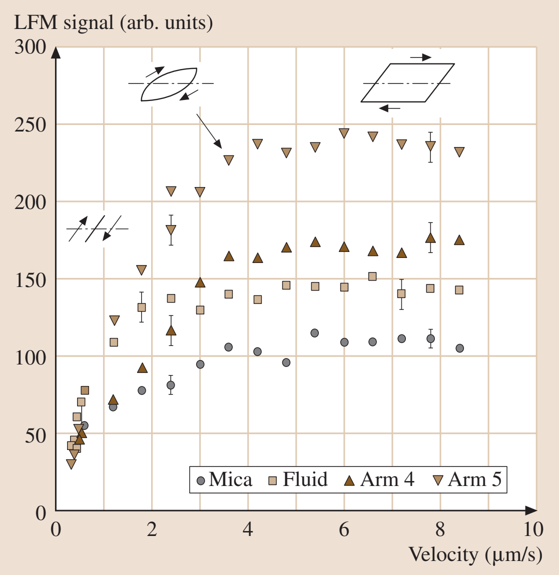

Fig. 20.28 Velocity dependence of friction on mica and on lipid films with different orientation (arms 4 and 5) and in a fluid phase. (After [20.47])

The first measurements on the atomic scale were performed by *Bennewitz* et al. on copper and sodium chloride [20.30, 46]; in both cases a logarithmic dependence of friction was revealed up to $v < 1$ µm/s (Fig. 20.25), in agreement with (20.21). Higher values of velocities were not explored, due to the limited range of the scan frequencies applicable by FFM on atomic scale. The same limitation does not allow a clear distinction between (20.21) and (20.26) in the interpretation of the experimental results.

At this point we would like to discuss the physical meaning of the characteristic frequency f_0. With lattice constants a of a few angstroms and effective spring

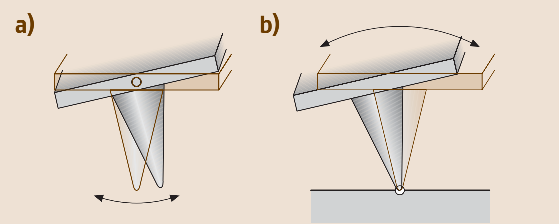

Fig. 20.29a,b Torsional modes of cantilever oscillations **(a)** when the tip is free and **(b)** when the tip is in contact with a surface. (After [20.54])

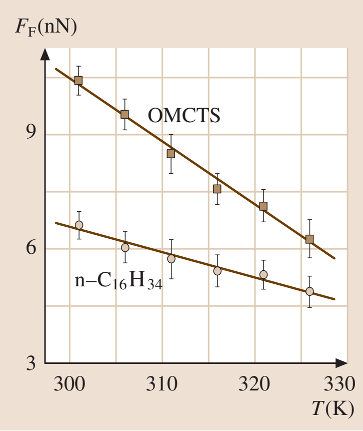

Fig. 20.31 Temperature dependence of friction on n-hexadecane and octamethylcyclotetrasiloxane. (After [20.56])

constants k_{eff} of about 1 N/m, which is typical in FFM experiments, (20.25) gives values of few hundreds of kHz for f_0. This is the characteristic range in which the torsional eigenfrequencies of the cantilevers are located in both contact and noncontact modes (Fig. 20.29). Future work might clarify whether or not f_0 has to be identified with these frequencies.

To conclude this paragraph, we should emphasize that the increase of friction with increasing velocity is ultimately related to the materials and the environment in which the measurements are realized. In humid environment, *Riedo* et al. observed that the surface wettability has an important role [20.55]. Friction *decreases* with increasing velocity on hydrophilic surfaces, and the rate of such decrease depends drastically on humidity. On partially hydrophobic surfaces, a logarithmic increase is found again (Fig. 20.30). These results were interpreted considering the thermally activated nucleation of water bridges between tip and sample asperities, as discussed in the cited reference.

20.4.3 Temperature Dependence of Friction

Thus far we have used thermal activation to explain the velocity dependence of friction. The same mechanism also predicts that friction should change with temperature. Master equation (20.18) shows that the probability of a tip jump is reduced at low temperatures T until it vanishes when $T = 0$. Within this limit case, thermal ac-

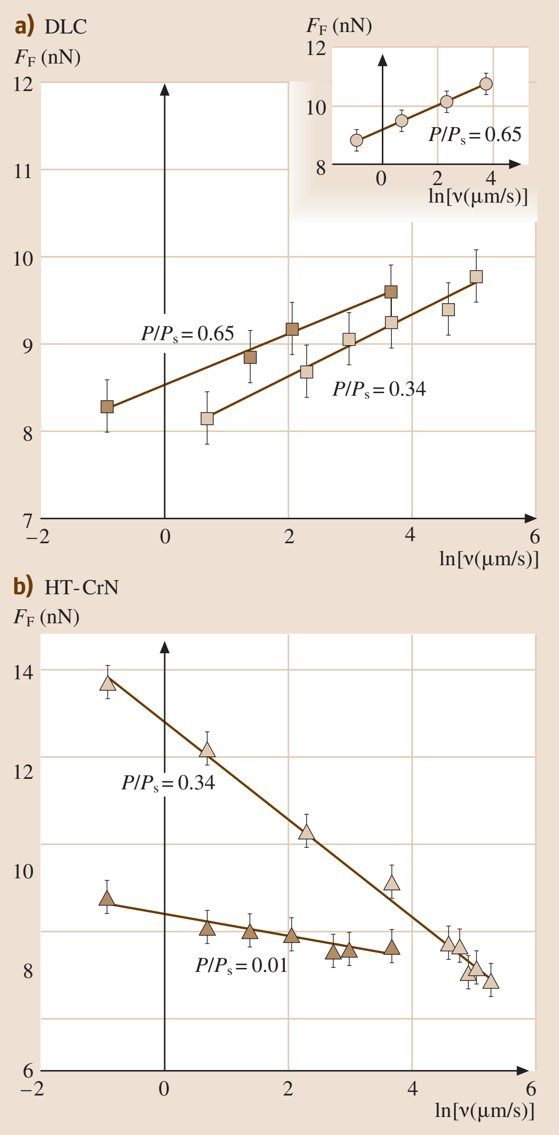

Fig. 20.30a,b Friction vs. sliding velocity **(a)** on hydrophobic surfaces and **(b)** on hydrophilic surfaces (After [20.55])

tivation is excluded, and the lateral force F_L is equal to F^*, independently of the scanning velocity v.

To our knowledge stick-slip processes at low temperatures have not been reported. A significant increase of the mean friction with decreasing temperature was recently measured by *He* et al. [20.56] by FFM (Fig. 20.31). Neglecting the logarithmic contributions (20.21) and (20.26) predict $(F^* - F_L) \sim T$ and $(F^* - F_L) \sim T^{2/3}$ respectively for the temperature dependence of friction. Although He et al. applied a linear fit to their data, the range of 30 K they considered is again not large enough to prove that a $T^{2/3}$ fit would be preferable.

20.5 Geometry Effects in Nanocontacts

Friction is ultimately related to the real shape of the contact between the sliding surfaces. On macroscopic scale the contact between two bodies is studied within continuum mechanics, which are all based on the elasticity theory developed by Hertz in 19th century. Various FFM experiments showed that continuum mechanics is still valid down to contact areas a few nanometers in size. Only when contact is reduced to few atoms does the continuum frame become unsuitable, and other approaches like molecular dynamics are necessary. This section will deal with continuum mechanics theory; molecular dynamics will be discussed in Sect. 20.7.

20.5.1 Continuum Mechanics of Single Asperities

The lateral force F_L between two surfaces in reciprocal motion depends on the size of the real area of contact, A, which can be a few orders of magnitude smaller than the apparent area of contact. The simplest assumption is that friction is proportional to A; the proportionality factor is called *shear strength* σ [20.57]:

$$F_L = \sigma A \,. \tag{20.28}$$

In case of plastic deformation, the asperities are compressed until the pressure, p, equals a certain yield value, p^*. The resulting contact area is thus $A = F_N/p^*$, and the well-known Amontons' law is obtained: $F_L = \mu F_N$, where $\mu = \sigma/p^*$ is the *coefficient of friction*. The same idea can be extended to contacts formed by many asperities, and it leads again to the Amontons' law. The simplicity of this analysis explains why most of the friction processes were related to plastic deformation for a long time. Such mechanism, however, should provoke a quick disruption of the surfaces, which is not observed in practice.

Elastic deformation can be easily studied in the case of a sphere of radius R pressed against a flat surface. In such case the contact area is

$$A(F_N) = \pi \left(\frac{R}{K}\right)^{2/3} F_N^{2/3} \,, \tag{20.29}$$

where $K = 3E^*/4$ and E^* is an effective Young's modulus, related to the Young's moduli, E_1 and E_2, and the Poisson's numbers, ν_1 and ν_2, of sphere and plane, by the following relation [20.58]:

$$\frac{1}{E^*} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \,. \tag{20.30}$$

The result $A \propto F_N^{2/3}$ is in contrast with the Amontons' law. But a linear relation between F_L and F_N can be obtained for contacts formed by several asperities in particular cases. For example, the area of contact between a flat surface and a set of asperities with an exponential height distribution and the same radius of curvature R depends linearly on the normal force F_N [20.59]. The same conclusion holds approximately even for a Gaussian height distribution. But the hypothesis that the radii of curvature are the same for all asperities is not realistic. A general model was recently proposed by *Persson*, who derived analytically the proportionality between contact area and load for a large variety of elastoplastic contacts formed by surfaces with arbitrary roughness [20.60]. The discussion is not straightforward and goes beyond the purposes of this section.

Further effects are observed if adhesive forces between the asperities are taken into account. If the range of action of these forces is smaller than the elastic deformation, (20.29) is extended to the Johnson–Kendall–Roberts (JKR) relation

$$A(F_N) = \pi \left(\frac{R}{K}\right)^{2/3} \times \left(F_N + 3\pi\gamma R + \sqrt{6\pi\gamma R F_N + (3\pi\gamma R)^2}\right)^{2/3} , \tag{20.31}$$

where γ is the surface tension of sphere and plane [20.61]. The real contact area at zero load is finite and the sphere can be detached only by pulling it away with a certain force. This is also true in the opposite case, in which the range of action of adhesive forces is larger than the elastic deformation. In such a case, the relation between contact area and load has the simple form

$$A(F_\mathrm{N}) = \pi \left(\frac{R}{K}\right)^{2/3} (F_\mathrm{N} - F_\mathrm{off})^{2/3} , \tag{20.32}$$

where F_off is the negative load required to break the contact. The Hertz-plus-offset relation (20.32) can be derived from the Derjaguin–Muller–Toporov (DMT) model [20.63]. To discriminate between the JKR or DMT models *Tabor* introduced a nondimensional parameter

$$\Phi = \left(\frac{9R\gamma^2}{4K^2 z_0^3}\right)^{1/3} , \tag{20.33}$$

where z_0 is the equilibrium distance in contact. The JKR model can be applied if $\Phi > 5$; the DMT model holds when $\Phi < 0.1$ [20.64]. For intermediate values of Φ, the Maugis–Dugdale model [20.65] could reasonably explain experimental results (Sect. 20.5.3).

20.5.2 Load Dependence of Friction

The FFM tip represents a single asperity sliding on a surface. The previous discussion suggests a nonlinear dependence of friction on the applied load, provided that continuum mechanics is applicable. Schwarz et al. observed the Hertz-plus-offset relation (20.32) on graphite, diamond, amorphous carbon and C_{60} in argon atmosphere (Fig. 20.32). In their measurements they used well-defined spherical tips with radii of curvature of tens of nanometers, obtained by contaminating silicon tips with amorphous carbon in a transmission electron microscope. In order to compare the tribological behavior of different materials, *Schwarz* et al. suggested the introduction of an effective coefficient of friction, $\tilde{C}$, which is independent of the tip curvature [20.62].

Meyer et al., *Carpick* et al., and *Polaczyc* et al. performed friction measurements in UHV in agreement with JKR theory [20.17, 66, 67]. In these experiments different materials were considered, i. e. ionic crystals, mica and metals. In order to correlate lateral and normal forces with improved statistics, Meyer et al. applied an original 2-D-histogram technique (Fig. 20.33). Carpick et al. extended the JKR relation (20.32) to include nonspherical tips. In the case of axisymmetric tip profile $z \propto r^{2n}$ ($n > 1$), it can be proven analytically that the increase of friction becomes less pronounced with increasing n (Fig. 20.34).

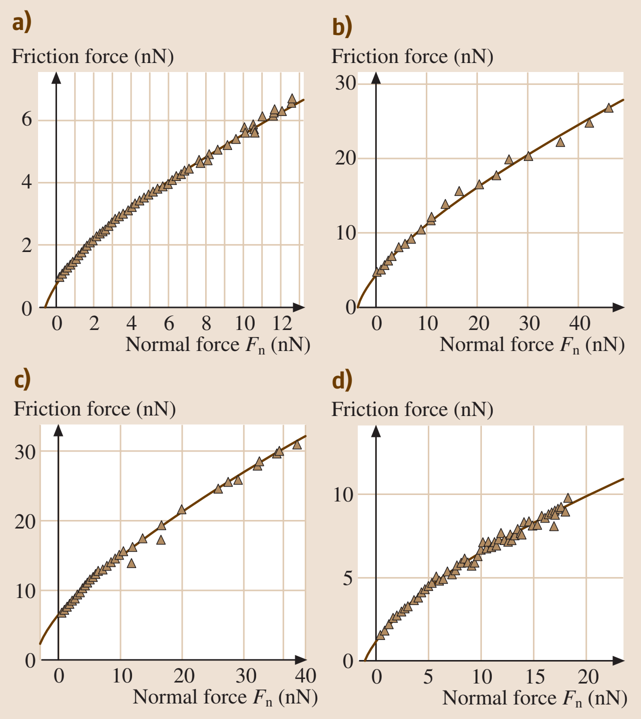

Fig. 20.32a–d Friction vs. load curve on amorphous carbon in argon atmosphere. Curves **(a)**–**(d)** refer to tips with different radii of curvature. (After [20.62])

20.5.3 Estimation of the Contact Area

In contrast to other tribological instruments, such as the surface force apparatus [20.68], the contact area cannot be directly measured by FFM. Indirect methods are provided by contact stiffness measurements. The contact between the FFM tip and the sample can be modeled by a series of two springs (Fig. 20.35). The effective constant k^z_eff of the series is given by

$$\frac{1}{k^z_\mathrm{eff}} = \frac{1}{k^z_\mathrm{contact}} + \frac{1}{c_\mathrm{N}} , \tag{20.34}$$

where c_N is the normal spring constant of the cantilever and k^z_contact is the normal stiffness of the contact. This quantity is related to the radius of the contact area a, by the simple relation

$$k^z_\mathrm{contact} = 2aE^* , \tag{20.35}$$

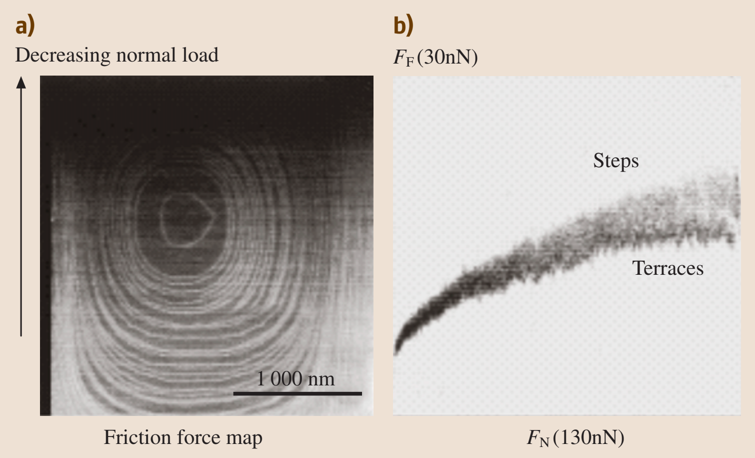

Fig. 20.33 **(a)** Friction force map on NaCl(100). The load is decreased from 140 to 0 nN (jump-off point) during imaging. **(b)** 2-D-hystogram of **(a)**. (After [20.17])

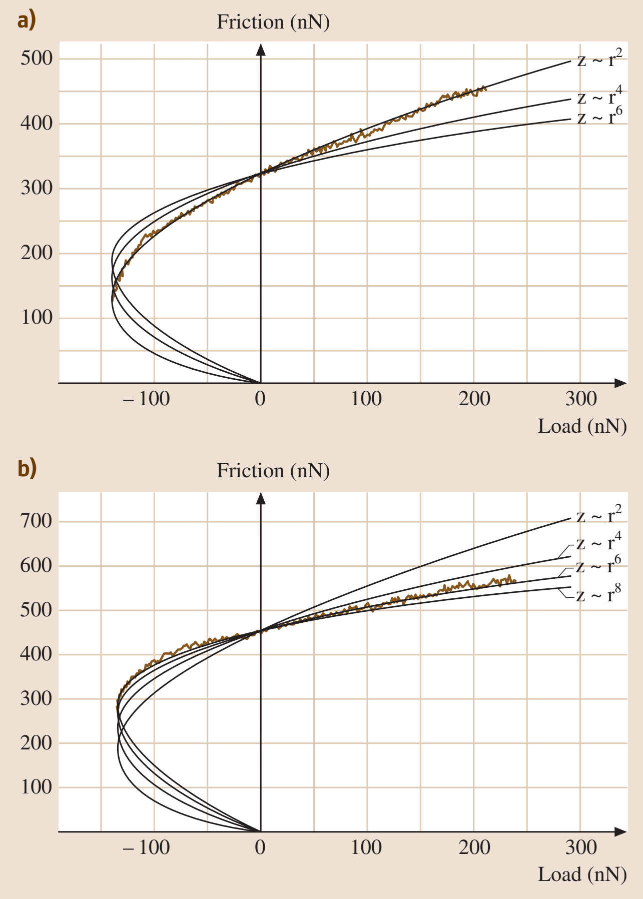

Fig. 20.34a,b Friction vs. load curves **(a)** for a spherical tip and **(b)** for a blunted tip. The *solid curves* are determined with the JKR theory. (After [20.66])

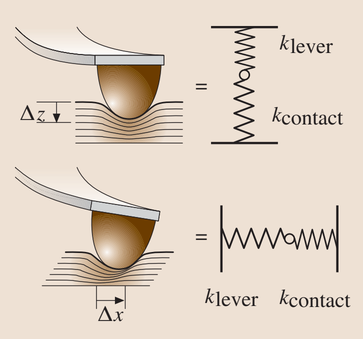

Fig. 20.35 Sketch of normal and lateral stiffness of the contact between tip and surface. (After [20.70])

where E^* is the effective Young's modulus previously introduced [20.69]. Typical values of k^z_{contact} are order of magnitude larger than c_N, however, and a practical application of (20.34) is not possible.

Carpick et al. suggested an alternative method independently [20.70, 71]. According to various models, the *lateral* contact stiffness of the contact between a sphere and a flat surface is [20.72]

$$k^x_{\text{contact}} = 8aG^* , \tag{20.36}$$

where the effective shear stress G^* is defined by

$$\frac{1}{G^*} = \frac{2-\nu_1^2}{G_1} + \frac{2-\nu_2^2}{G_2} \tag{20.37}$$

(G_1, G_2 are the shear moduli of sphere and plane). The contact between the FFM tip and the sample can be modeled again by a series of springs (Fig. 20.35). The effective constant k^x_{eff} of the series is given by

$$\frac{1}{k^x_{\text{eff}}} = \frac{1}{k^x_{\text{contact}}} + \frac{1}{k^x_{\text{tip}}} + \frac{1}{c_L} , \tag{20.38}$$

where c_L is the lateral spring constant of the cantilever and k^x_{contact} is the lateral stiffness of the contact. As suggested by Lantz, (20.38) includes also the lateral stiffness of the tip, k^x_{tip}, which can be comparable to the lateral spring constant. The effective spring constant k^x_{eff} is simply given by the slope $\mathrm{d}F_L/\mathrm{d}x$ of the friction loops (Sect. 20.2.1). Once k^x_{contact} is determined, the contact radius a is easily estimated by (20.36).

The lateral stiffness method was applied to contacts between silicon nitride and muscovite mica in air and between $NbSe_2$ and graphite in UHV. Both the spring constant k^x_{eff} and the lateral force F_L dependence on the load F_N were explained within the same models (JKR and Maugis–Dugdale, respectively), which confirms that friction is proportional to the contact area in the applied range of loads (up to $F_N = 40$ nN in both experiments).

Enachescu et al. estimated the contact area by measuring the contact conductance on diamond as a function of the applied load [20.73, 74]. Their experimental data were fitted with the DMT model, which was also used to explain the dependence of friction vs. load. Since the contact conductance is proportional to the contact area, the validity of the hypothesis (20.28) was confirmed again.

20.6 Wear on the Atomic Scale

If the normal force F_N applied with the FFM exceeds a critical value, which depends on the tip shape and on the material under investigation, the surface topography is permanently modified. In some cases wear is exploited to create patterns with well-defined shape. Here we will focus on the mechanisms acting on the nanometer scale, where recent experiments have proven the unique capability of the FFM in both scratching and imaging surfaces down to the atomic scale.

20.6.1 Abrasive Wear on the Atomic Scale

In the case of ionic crystals, *Lüthi* et al. observed the appearance of wear at very low loads, i. e. $F_N = 3$ nN [20.31]. Atomically resolved images of the damage produced by scratching the FFM tip areas on potassium bromide were obtained very recently by *Gnecco* et al. [20.75]. In Fig. 20.36 a small mound grown up at the end of a groove on KBr(100) is shown under different magnifications. The groove was created a few minutes before imaging by repeated scanning with the normal force $F_N = 21$ nN. The image shows a lateral force map acquired with a load of about 1 nN; no atomic features were observed in the corresponding topography signal. Figure 20.36a,b shows that the debris extracted from the groove recrystallized with the same atomic arrangement of the undamaged surface, which suggests that the wear process occurred like an epitaxial growth assisted by the microscope tip.

Although it is not straightforward to understand how wear is initiated and how the tip transports the debris, important indications are given by the profile of the lateral force F_L recorded while scratching. Figure 20.37 shows some friction loops acquired when the tip was scanned laterally on 5×5 nm^2 areas. The mean lateral force multiplied by the scanned length gives the total energy dissipated in the process. The result of the tip movement are the pits in Fig. 20.38a. Thanks to the pseudo-atomic

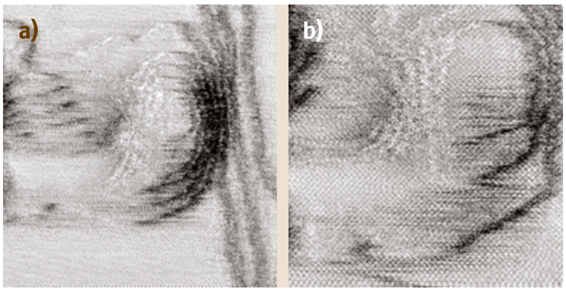

Fig. 20.36a,b Lateral force images acquired at the end of a groove scratched 256 times with a normal force $F_N = 21$nN. Frame sizes: **(a)** 39 nm, **(b)** 25 nm

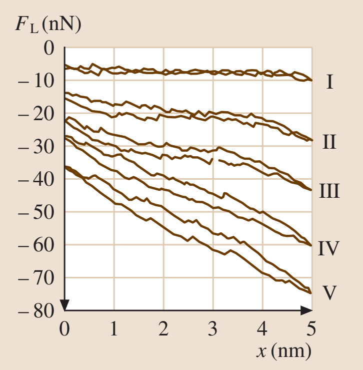

Fig. 20.37 Friction loops acquired while scratching the KBr surface on 5 nm long lines with different loads $F_N = 5.7$ to 22.8 nN. (After [20.75])

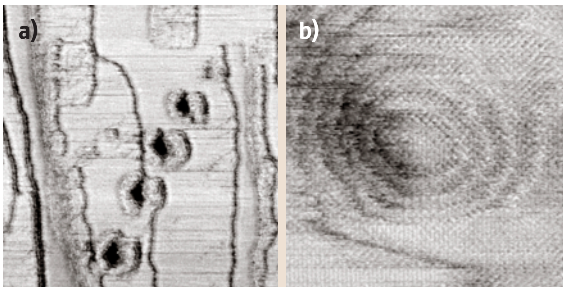

Fig. 20.38 **(a)** Lateral force images of the pits produced with $F_N = 5.7$ to 22.8 nN. Frame size: 150 nm; **(b)** Detailed image of the fourth pit from the top with psuedo-atomic resolution. Frame size: 20 nm

resolution obtained by FFM (Fig. 20.38b), the number of removed atoms could be determined from the lateral force images, which allowed estimating that 70% of the dissipated energy went into wear-less friction [20.75]. Figures 20.37 and 20.38 clearly show that the damage increases with increasing load. On the other hand, changing the scan velocity v between 25 and 100 nm/s did not produce any significant variation in the wear process.

A different kind of wear was observed on layered materials. *Kopta* et al. [20.76] could remove layers from a muscovite mica surface by scratching with the normal force $F_N = 230$ nN (Fig. 20.39a). Fourier filtered images acquired on very small areas revealed the different periodicities of the underlying layers, which reflect the complex structure of the muscovite mica (Fig. 20.39b,c).

20.6.2 Wear Contribution to Friction

In Fig. 20.40 the mean lateral force detected while scratching the KBr(100) surface with a fixed load $F_N = 11$ nN is shown. A rather continuous increase of "friction" with the number of scratches N is observed, which can be approximated with the following exponential law:

$$F_L = F_0\,\mathrm{e}^{-N/N_0} + F_\infty\left(1 - \mathrm{e}^{-N/N_0}\right) . \qquad (20.39)$$

Equation (20.39) is easily interpreted assuming that friction is proportional to the contact area $A(N)$, and that time evolution of $A(N)$ can be described by

$$\frac{\mathrm{d}A}{\mathrm{d}N} = \frac{A_\infty - A(N)}{N_0} , \qquad (20.40)$$

A_∞ being the limit area in which the applied load can be balanced without scratching.

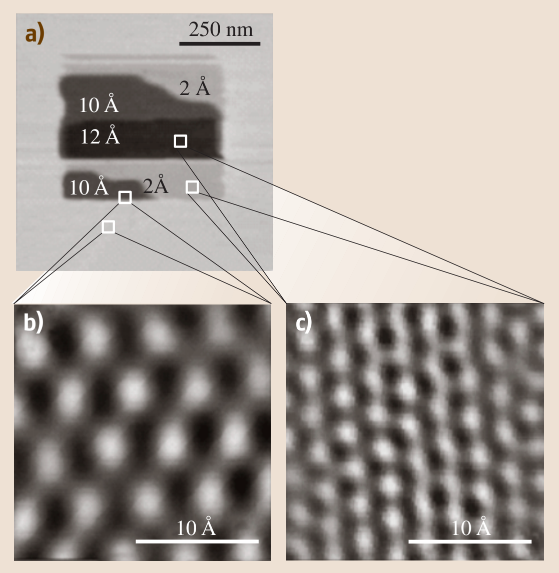

Fig. 20.39 **(a)** Topography image of an area scratched on muscovite mica with $F_N = 230$ nN; **(b)**,**(c)** Fourier filtered images of different regions. (After [20.76])

To interpret their experiment on mica, Kopta et al. assumed that wear is initiated by atomic defects. When the defects accumulate beyond a critical concentration, they grow to form the scars shown in Fig. 20.39. Such process was once again related to thermal activation. The number of defects created in the contact area $A(F_N)$ is

$$N_{\mathrm{def}}(F_N) = t_{\mathrm{res}} n_0 A(F_N) f_0 \exp\left(-\frac{\Delta E}{k_B T}\right) , \qquad (20.41)$$

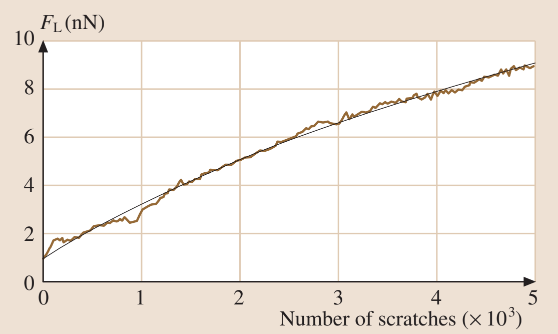

Fig. 20.40 Mean value of the lateral force during repeated scratching with $F_N = 11$ nN on a 500 nm line. (After [20.75])

where t_{res} is the residence time of the tip, n_0 is the surface density of atoms, and f_0 is the attempt frequency to overcome the energy barrier ΔE to break a Si−O bond, which depends on the applied load. When the defect density reaches a critical value, a hole is nucleated. The friction force during the creation of a hole could also be estimated with thermal activation by Kopta et al., who derived the formula

$$F_{\mathrm{L}} = c(F_{\mathrm{N}} - F_{\mathrm{off}})^{2/3} + \gamma F_{\mathrm{N}}^{2/3} \exp\left(B_0 F_{\mathrm{N}}^{2/3}\right) . \tag{20.42}$$

The first term on the right gives the wearless dependence of friction in the Hertz-plus-offset model (Sect. 20.5.1); the second term is the contribution of the defect production. The agreement between (20.42) and the experiment can be observed in Fig. 20.41.

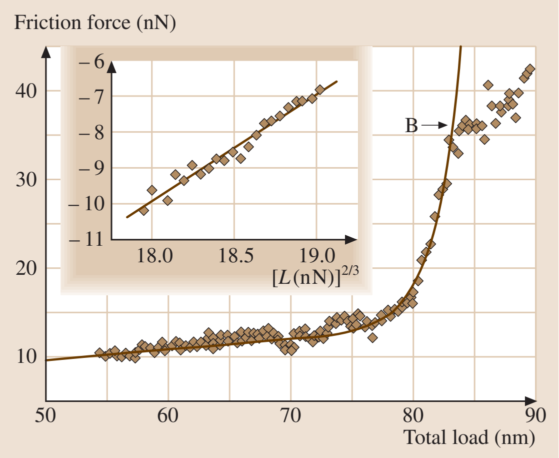

Fig. 20.41 Friction vs. load curve during the creation of a hole in the muscovite mica. (After [20.76])

20.7 Molecular Dynamics Simulations of Atomic Friction and Wear

Section 20.5 has introduced how small sliding contacts can be modeled by continuum mechanics. This modeling has several limitations. The first, most obvious, is that continuum mechanics cannot account for atomic-scale processes like atomic stick-slip. While this limit can be overcome by semiclassical descriptions like the Tomlinson model, a definite limit is the determination of contact stiffness for contacts with a radius of a few nanometers. Interpretation of experimental results with the methods introduced in Sect. 20.5.3 regularly yields contact radii of atomic or even smaller size, in clear contradiction to the minimal contact size given by adhesion forces. Such macroscopic quantities as shear modulus or pressure fail to describe the mechanical behavior of these contacts. A microscopic modeling including the atomic structure of the contact is therefore required. This is usually done by a *molecular dynamics* (MD) simulation of the contact. In such simulations, the sliding contact is set up by boundaries of fixed atoms in relative motion and the atoms of the contact, which are allowed to relax their positions according to interactions between each pair of atoms. The methods of computer simulations in tribology are discussed elsewhere in this book. In Sect. 20.7.1 we will discuss such simulations, which can be directly compared to experimental results showing atomic friction processes. The major outcome of the simulations beyond the inclusion of the atomic structure is the importance of displacement of atoms for a correct prediction of forces. In the following Sect. 20.7.2, we will present simulation studies including wear of tip or surface.

20.7.1 Molecular Dynamics Simulation of Friction Processes

The first experiments that exhibited atomic friction features were performed on layered materials, often graphite. A theoretical study of forces between an atomically sharp diamond tip and the graphite surface has been reported by *Tang* et al. [20.77]. The authors found a significant distance dependence of forces, in which the strongest contrast appeared at different distances for normal and lateral forces due to the strong displacement of surface atoms. The order of magnitude found in this study was one nanonewton, much less than in most experimental reports, which indicated that in such experiments a contact area of far larger dimension was realized. Tang et al. already determined that the distance dependence of forces could even change the symmetry appearance of the observed lateral forces. The experimental situation has also been studied in numerical simulations using simplified one-atom potential for the tip–surface interaction but including the spring potential of the probing force sensor [20.36]. The motivation for these studies was the observation of a hexagonal pattern in the friction force, while the surface atoms of graphite are ordered in a honeycomb structure. The simulations revealed how the jump path of the tip under lateral force is dependent on the force constant of the probing force sensor.

Surfaces of ionic crystals have become model systems for studies in atomic friction. Atomic stick-slip behavior has been observed by several groups with a lateral force modulation of the order of 1 nN. Pioneering work in atomistic simulation of sliding contacts has been done by *Landman* et al.. The first ionic system studied was a CaF_2 tip sliding over a $CaF_2(111)$ surface [20.78]. In MD calculation with controlled temperature, the tip was first approached toward the surface up to the point at which an attractive normal force of −3 nN acted on the tip. Then the tip was laterally moved, and the lateral force determined. An oscillation with the atomic periodicity of the surface and with an amplitude decreasing from 8 nN was found. Inspection of the atomic positions revealed a wear process by shear cleavage of the tip. Such transfer of atoms between tip and surface plays a crucial role in atomic friction studies as was shown by *Shluger* et al. [20.79]. These authors simulated a MgO tip scanning laterally over a LiF(100) surface. Initially an irregular oscillation of the system's energy is found together with transfer of atoms between surface and tip. After a while, the tip apex structure is changed by adsorption of Li and F ions in such a way that nondestructive sliding with perfectly regular oscillation of the energy with the periodicity of the surface could be observed. The authors called this effect self-lubrication and speculate that generally a dynamic self-organization of the surface material on the tip might promote the observation of periodic forces. In a less costly approach of a molecular mechanics study, in which the forces are calculated for each fixed tip–sample configuration, *Tang* et al. produced lateral and normal force maps for a diamond tip over a NaCl(100) surface including such defects as vacancies and a step [20.80]. In accordance with the studies mentioned before, they found that significant atomic force contrast can be expected for tip–sample distances below 0.35 nm, while distances below 0.15 nm result in destructive forces. For the idealized conditions of scanning at constant height in this regime, the authors predict that even atomic-sized defects could be imaged. Experimentally, these conditions cannot be stabilized in static modes, which so far have been used in lateral force measurements. But dynamic modes of force microscopy have proven atomic resolution of defects when entering the distance regime between 0.2 and 0.4 nm [20.81]. Recent experimental progress in atomic friction studies of surfaces of ionic crystals include the velocity dependence of lateral forces and atomic-scale wear processes. Such phenomena are not yet accessible by MD studies: the experimentally realized scanning time scale is too far from the atomic relaxation time scales that govern MD simulations. Furthermore, the number of freely transferable atoms that can be included in a simulation is simply limited by meaningful calculation time.

Landman et al. also simulated a system of high reactivity, namely a silicon tip sliding over a silicon surface [20.82]. A clear stick-slip variation of the lateral force was observed for this situation. Strong atom displacements created an interstitial atom under the influence of the tip, which was annealed as the tip moved on. Permanent damage was predicted, however, in the case that the tip enters the repulsive force regime. Although the simulated Si(111) surface is experimentally not accessible, it should be mentioned that on the Si(111)7×7 reconstructed surface the tip had to be passivated by a Teflon layer before nondestructive contact-mode measurements became possible (Sect. 20.3). It is worth noting that the simulations for the Cu(111) surface revealed a linear relation between contact area and mean lateral force, similar to classical macroscopic laws.

For metallic sliding over metallic surfaces, wear processes are predicted by several MD studies, which will be discussed in the following section. For a (111)-terminated copper tip over the Cu(111) surface, however, *Sørensen* et al. found that nondestructive sliding is possible while the lateral force exhibits the sawtooth-like shape characteristic of atomic stick-slip (Fig. 20.42). In contrast, a Cu(100) surface would be disordered by a sliding contact (Fig. 20.43). This difference between the (100) and (111) plane as well as the absolute lateral forces has been confirmed experimentally (Sect. 20.3).

20.7.2 Molecular Dynamics Simulations of Abrasive Wear

The long time scales characteristic of the wear processes and the large amount of material involved make any attempt to simulate these mechanisms on a computer a tremendous challenge. Despite these premises, MD can provide useful insight on the mechanism of removal and deposition of single atoms operated by the FFM tip, which is a kind of information not directly observable in the experiments. Complex processes like

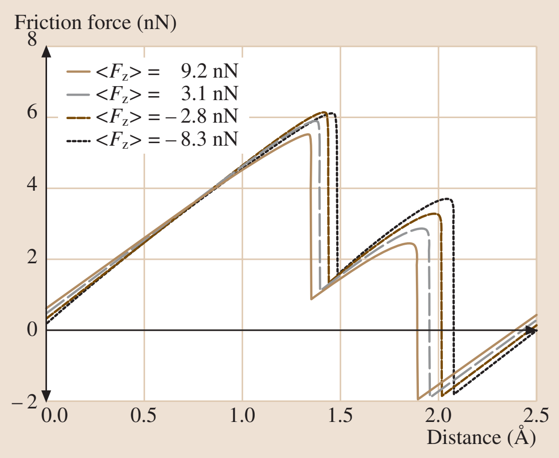

Fig. 20.42 Lateral force acting on a Cu(111) tip in matching contact with a Cu(111) substrate as a function of the sliding distance at different loads. (After [20.83])

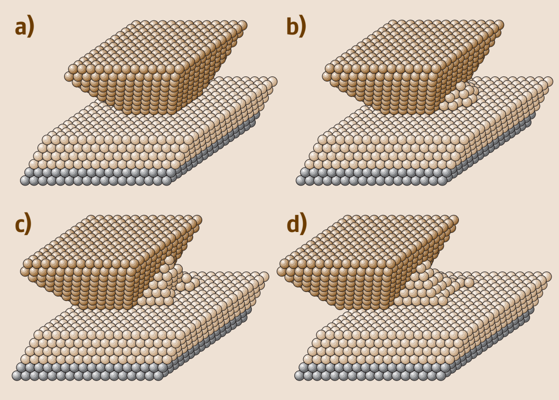

Fig. 20.43a–d Snapshot of a Cu(100) tip on a Cu(100) substrate during sliding. **(a)** Starting configuration; **(b)**–**(d)** snapshots after 2, 4, and 6 slips. (After [20.83])

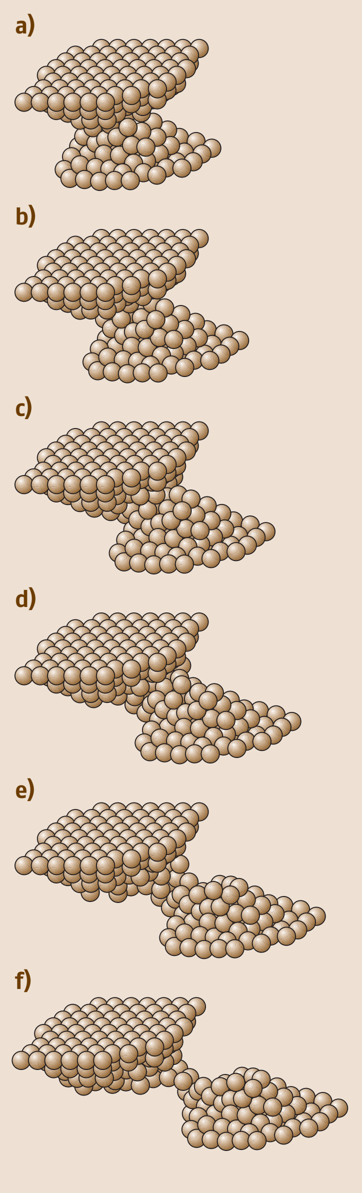

Fig. 20.44a–f Snapshot of a Cu(100) neck during shearing starting from the configuration **(a)**. The upper substrate has been displaced 4.2 Å between subsequent pictures. (After [20.83])

abrasive wear and nanolithography can be investigated only within approximate classical mechanics.

The observation made by *Livshits* and *Shluger* that the FFM tip undergoes a process of self-lubrication when scanning ionic surfaces (Sect. 20.7.1) proves that friction and wear are strictly related phenomena. In their MD simulations on copper, *Sørensen* et al. considered not only ordered (111) and (100) terminated tips but even amorphous structures obtained by "heating" the tip at high temperatures [20.83]. The lateral motion of the neck thus formed revealed a stick-slip behavior due to combined sliding and stretching, and also ruptures, accompanied by deposition of debris on the surface (Fig. 20.44).

To our knowledge, only a few examples of abrasive wear simulations on the atomic scale have been reported. *Buldum* and *Ciraci*, for instance, studied nanoindentation and sliding of a sharp Ni(111) tip on Cu(110) and of a blunt Ni(001) tip on Cu(100) [20.84]. In the case of the sharp tip quasiperiodic variations of the lateral force were observed, due to stick-slip involving phase-transition. One layer of the asperity was deformed to match the substrate during the first slip and then two asperity layers merged into one through structural transition during the second slip. In the case of the blunt tip, the stick-slip was less regular.

Different results have been reported in which the tip is harder than the underlying sample. *Komanduri* et al. considered an infinitely hard Ni indenter scratching single crystal aluminum at extremely low depths (Fig. 20.45) [20.85]. In this case a linear relation between friction and load was found, with a high coefficient of friction $\mu = 0.6$, independent of the scratch depth. Nanolithography simulations were recently performed by *Fang* et al. [20.86], who investigated the role of the displacement of the FFM tip along the direction of slow motion between a scan line and the next one. They found a certain correlation with FFM experiments on silicon films coated with aluminum.

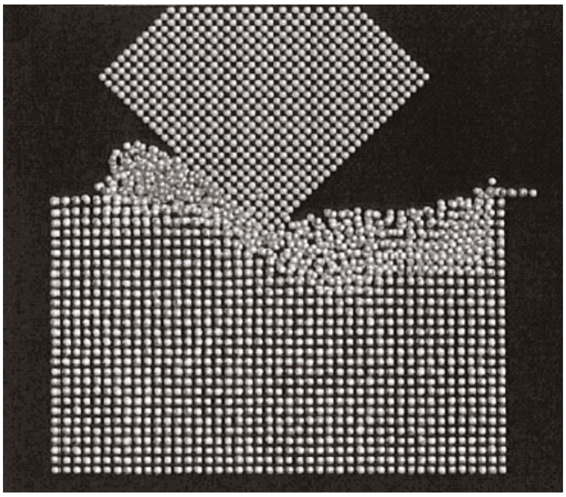

Fig. 20.45 MD simulation of a scratch realized with an infinitely hard tool. (After [20.85])

20.8 Energy Dissipation in Noncontact Atomic Force Microscopy

Historically, the measurement of energy dissipation induced by tip–sample interaction has been the domain of friction-force microscopy, where the sharp AFM-tip is sliding in gentle contact over a sample. The origins of dissipation in friction are related to phonon excitation, electronic excitation, and irreversible changes of the surface. In a typical stick-slip experiment, the energy dissipated in a single atomic slip event is of the order of 1 eV.

But the lateral resolution of force microscopy in the contact mode is limited by a minimum contact area containing several atoms due to adhesion between tip and sample.

This problem has been overcome in noncontact dynamic force microscopy. In the dynamic mode, the tip is oscillating with a constant amplitude A of typically 1–20 nm at the eigenfrequency f of the cantilever, which shifts by Δf due to interaction forces between tip and sample. This technique has been described in detail in Part B of this book.

Dissipation also occurs in noncontact mode of force microscopy, where the atomic structure of tip and sample are reliably preserved. In this dynamic mode, the damping of the cantilever oscillation can be deduced from the excitation amplitude A_{exc} required to maintain the constant tip oscillation amplitude on resonance.

Compared to friction force microscopy, the interpretation of noncontact-AFM (nc-AFM) experiments is complicated due to the perpendicular oscillation of the tip, typically with an amplitude that is large compared to the minimum tip–sample separation. Another problem is to relate the measured damping of the cantilever to the different origins of dissipation.

In all dynamic force microscopy measurements a power dissipation P_0 caused by internal friction of the freely oscillating cantilever is observed, which is proportional to the eigenfrequency ω_0 and to the square of the amplitude A and inverse proportional to the known Q-value of the cantilever. When the tip–sample distance is reduced, the tip interacts with the sample and therefore an additional damping of the oscillation is encountered. This extra dissipation P_{ts} caused by tip–sample interaction can be calculated from the excitation signal A_{exc} [20.87].

The observed energy losses per oscillation cycle (100 meV) [20.88] are roughly similar to the 1 eV energy loss in the contact slip-process. When estimating the contact area in the contact-mode to a few atoms, the energy dissipation per atom that can be associated with a bond being broken and reformed is also around 100 meV.

The idea to relate additional damping of the tip oscillation to dissipative tip–sample interaction has recently attracted much attention [20.89]. The origin of this additional dissipation are manifold: one may distinguish between apparent energy dissipation (for example from an inharmonic cantilever-motion, artifacts from the phase-controller, or slow fluctuations round the steady-state solution [20.89, 90]), velocity dependent dissipation (for example electric- and magnetic-field mediated Joule dissipation [20.91, 92]) and hysteresis-related dissipation (due to atomic instabilities [20.93, 94] or hysteresis due to adhesion [20.95]).

By recording Δf and A_{exc} simultaneously during a typical AFM-experiment, forces and dissipation can be measured. Many experiments show true atomic contrast in topography (controlled by Δf) and in the dissipation-signal A_{exc} [20.96]; however, the origin of the atomic energy dissipation process is still not completely resolved.

To prove that the observed atomic-scale variation of the damping is indeed due to atomic-scale energy dissipation and not an artefact of the distance feedback, *Loppacher* et al. [20.88] carried out an nc-AFM experiment on Si(111)-7×7 at constant height, i. e. with stopped distance feedback. Frequency-shift and dissipation exhibit an atomic-scale contrast, demonstrating a true atomic-scale variation of force and dissipation.

A strong atomic-scale dissipation contrast at step edges has been demonstrated in a few experiments (NaCl on Cu [20.81]; or measurements on KBr [20.97]). In Fig. 20.46 ultrathin NaCl islands grown on Cu(111) are shown. As it is shown in Fig. 20.46a, the island edges have a higher contrast compared to the NaCl-terrace and show an atomically resolved corrugation. The strongly enhanced contrast of the step edges and kink sites could be attributed to a slower decay of the electric field and to easier relaxation of theses positions of the ions at these locations. The dissipation image shown in Fig. 20.46b was recorded simultaneously. To establish a direct spatial correspondence between the excitation and the topography-signal, the match between topography and A_{exc} on many images has been studied. Sometimes topography and A_{exc} are in phase, sometimes they are shifted a little bit, and sometimes A_{exc} is at a minimum when topography is at a maximum. The local contrast formation thus depends strongly on the atomic tip structure. In fact, the strong dependence of the dissipation contrast on the atomic state of the tip apex is impressively confirmed by the tip change observed in the experimental images shown in Fig. 20.46b. The dissipation contrast is seriously enhanced, while the topography contrast remains nearly unchanged. Clearly, the dissipation depends strongly on the state of the tip and exhibits more short-range character than the frequency shift.

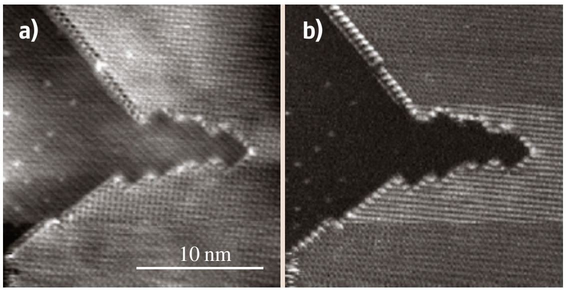

Fig. 20.46 **(a)** Topography and **(b)** A_{exc} images of a NaCl island on Cu(111). The tip changes after 1/4 of the scan, thereby changing the contrast in the topography and enhancing the contrast in A_{exc}. After 2/3 of the scan, the contrast from the *lower part* of the image is reproduced, indicating that the tip change was reversible. (After [20.81])

More directly related to friction measurements, where the tip is sliding in contact over the sample, are nc-AFM experiments, where the tip is oscillating parallel to the surface. *Stowe* et al. [20.98] oriented cantilever beams with in-plane tips perpendicular to the surface, so that the tip motion was approximately parallel to the surface. The noncontact damping of the lever was used to measure localized electrical Joule dissipation. They were able to image the dopant density for n- and p-type silicon samples with 150 nm spatial resolution. An U_{ts}^2 dependence on the tip–sample voltage for the dissipation was found, as proposed by *Denk* and *Pohl* [20.91] for electric-field Joule-dissipation.

Stipe et al. [20.99] measured the noncontact friction between a Au(111)-surface and a gold-coated tip with the same setup. They observed the same U_{ts}^2 dependence of the Bias-voltage and a distance dependence that follows a power law $1/d^n$, where n is between 1.3 and 3 [20.99, 100]. Even when the external Bias-voltage is zero, a substantial electric-field is present. The presence of inhomogeneous tip–sample electric fields is difficult to avoid, even under the best experimental conditions. Although this dissipation is electrical in origin, the detailed mechanism is not totally clear. The most straightforward mechanism is to assume that inhomogeneous fields emanating from the tip and the sample induce surface charges in the nearby metallic sample. When the tip moves, currents are induced, causing Ohmic dissipation [20.91, 98]. But in metals with good electrical conductivity, Ohmic dissipation is insufficient to account for the observed effect [20.101]. Thus the tip–sample electric field must have an additional effect, such as driving the motion of adsorbates and surface defects.

When exciting the torsional oscillation of commercial, rectangular AFM cantilevers, the tip is oscillating approximately parallel to the surface. In this mode, lateral forces acting on the tip at step edges and near impurities could be measured quantitatively [20.54]. An enhanced energy-dissipation was observed at the impurities as well. When the tip is further approached toward the sample, contact formation transforms the nearly free torsional oscillation of the cantilever into a different mode with the tip–sample contact acting as a hinge. When this contact is formed, a rapid increase of the power required to maintain a constant tip oscillation amplitude and a positive frequency shift are found. The onset of the simultaneously recorded damping and positive frequency shift are sharp and essentially coincide. It is assumed that these changes indicate the formation of a tip–sample contact. Two recent studies [20.102, 103] report on the use of the torsional eigenmode for measuring the elastic properties of the tip–sample contact, where the tip is in contact with the sample and the shear stiffness depends on the normal load.

Kawagishi et al. [20.104] scanned with lateral amplitudes in the order of 10 pm to 3 nm; their imaging technique gave contrast between graphite terraces, silicon and silicon dioxide, graphite and mica. Torsional self-excitation showed nanometric features of self-assembled monolayer islands due to different lateral dissipation.

Recently *Giessibl* et al. [20.105] established true atomic resolution of both conservative and dissipative forces by lateral force microscopy. The interaction between a single-tip atom oscillated parallel to an Si(111)-7×7 surface is measured. A dissipation energy of up to 4 eV per oscillation cycle is found, which is explained by a plucking action of one atom on to the other as described by *Tomlinson* in 1929 [20.23].

A detailed review of dissipation phenomena in noncontact force microscopy has been given by *Hug* [20.106].

20.9 Conclusion

Over the last 15 years, two instrumental developments have stimulated scientific activities in the field of nanotribology. On the one hand, the invention and development of friction force microscopy allows us to quantitatively study friction of a single asperity. As we have discussed in this chapter, atomic processes are observed by forces on the order of 1 nN, i. e. forces related to single chemical bonds. On the other hand, the enormous increase of achievable computing power provides the basis for molecular dynamics simulations of systems containing several hundred atoms. Therewithin, the development of the atomic structure in a sliding contact can be analyzed and forces predicted.

The most prominent observation of atomic friction is a stick-slip behavior with the periodicity of the surface atomic lattice. Semiclassical models can explain the experimental findings including the velocity dependence, which is a consequence of thermal activation of slip events. Classical continuum mechanics can also describe the load dependence of friction in contacts with an extension of several ten nanometers. But when trying to apply continuum mechanics to contacts formed at just ten atoms, obviously wrong numbers result, such as, for example, for the contact radius. Only a comparison with atomistic simulations can provide a full meaningful picture of the physical parameters of such sliding contacts. These simulations predict a close connection between wear and friction, in particular a transfer of atoms between surface and tip, which in some case can even lower friction in a process of self-lubrication.

First experiments have succeeded in studying the onset of wear with atomic resolution. The research in microscopic wear processes will certainly grow in importance as nanostructures are produced and their mechanical properties exploited. Simulation of such processes involving transfer of thousands of atoms will become feasible with further increase of computing power. Another perspective of nanotribology is the expansion of atomic friction experiments toward surfaces with a well-defined roughness. In general, the problem of bridging the gap between single asperity experiments on well-defined surfaces and macroscopic friction should be approached, both experimentally and in modeling.

© Springer 2004

References

20.1 C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang: Atomic-scale friction of a tungsten tip on a graphite surface, Phys. Rev. Lett. **59** (1987) 1942–1945

20.2 G. Binnig, C. F. Quate, Ch. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

20.3 O. Marti, J. Colchero, J. Mlynek: Combined scanning force and friction microscopy of mica, Nanotechnology **1** (1990) 141–144

20.4 G. Meyer, N. Amer: Simultaneous measurement of lateral and normal forces with an optical-beam-deflection atomic force microscope, Appl. Phys. Lett. **57** (1990) 2089–2091

20.5 G. Neubauer, S. R. Cohen, G. M. McClelland, D. E. Horn, C. M. Mate: Force microscopy with a bidirectional capacitance sensor, Rev. Sci. Instrum. **61** (1990) 2296–2308

20.6 G. M. McClelland, J. N. Glosli: Friction at the atomic scale. In: *NATO ASI Series E*, Vol. 220, ed. by L. Singer, H. M. Pollock (Kluwer, Dordrecht 1992) pp. 405–425

20.7 R. Linnemann, T. Gotszalk, I. W. Rangelow, P. Dumania, E. Oesterschulze: Atomic force microscopy and lateral force microscopy using piezoresistive cantilevers, J. Vac. Sci. Technol. B **14** (1996) 856–860

20.8 M. Nonnenmacher, J. Greschner, O. Wolter, R. Kassing: Scanning force microscopy with micromachined silicon sensors, J. Vac. Sci. Technol. B **9** (1991) 1358–1362

20.9 R. Lüthi: Untersuchungen zur Nanotribologie und zur Auflösungsgrenze im Ultrahochvakuum mittels Rasterkraftmikroskopie. Ph.D. Thesis (Univ. of Basel, Basel 1996)

20.10 J. Cleveland, S. Manne, D. Bocek, P. K. Hansma: A nondestructive method for determining the spring constant of cantilevers for scanning force microscopy, Rev. Sci. Instrum. **64** (1993) 403–405

20.11 J. L. Hutter, J. Bechhoefer: Calibration of atomic-force microscope tips, Rev. Sci. Instrum. **64** (1993) 1868–1873

20.12 H. J. Butt, M. Jaschke: Calculation of thermal noise in atomic-force microscopy, Nanotechnology **6** (1995) 1–7

20.13 J. M. Neumeister, W. A. Ducker: Lateral, normal, and longitudinal spring constants of atomic-force microscopy cantilevers, Rev. Sci. Instrum. **65** (1994) 2527–2531

20.14 D. F. Ogletree, R. W. Carpick, M. Salmeron: Calibration of frictional forces in atomic force microscopy, Rev. Sci. Instrum. **67** (1996) 3298–3306

20.15 E. Gnecco: AFM study of friction phenomena on the nanometer scale. Ph.D. Thesis (Univ. of Genova, Genova 2001)

20.16 U. D. Schwarz, P. Köster, R. Wiesendanger: Quantitative analysis of lateral force microscopy experiments, Rev. Sci. Instrum. **67** (1996) 2560–2567

20.17 E. Meyer, R. Lüthi, L. Howald, M. Bammerlin, M. Guggisberg, H.-J. Güntherodt: Site-specific friction force spectroscopy, J. Vac. Sci. Technol. B **14** (1996) 1285–1288

20.18 S. S. Sheiko, M. Möller, E. M. C. M. Reuvekamp, H. W. Zandberger: Calibration and evaluation of scanning-force-microscopy probes, Phys. Rev B **48** (1993) 5675

20.19 F. Atamny, A. Baiker: Direct imaging of the tip shape by AFM, Surf. Sci. **323** (1995) L314

20.20 L. Howald, E. Meyer, R. Lüthi, H. Haefke, R. Overney, H. Rudin, H.-J. Güntherodt: Multifunctional probe microscope for facile operation in ultrahigh vacuum, Appl. Phys. Lett. **63** (1993) 117–119

20.21 J. S. Villarrubia: Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation, J. Res. Natl. Inst. Stand. Technol. **102** (1997) 425–454

20.22 Q. Dai, R. Vollmer, R. W. Carpick, D. F. Ogletree, M. Salmeron: A variable temperature ultrahigh vacuum atomic force microscope, Rev. Sci. Instrum. **66** (1995) 5266–5271

20.23 G. A. Tomlinson: A molecular theory of friction, Philos. Mag. Ser. **7** (1929) 905

20.24 T. Gyalog, M. Bammerlin, R. Lüthi, E. Meyer, H. Thomas: Mechanism of atomic friction, Europhys. Lett. **31** (1995) 269–274

20.25 T. Gyalog, H. Thomas: Friction between atomically flat surfaces, Europhys. Lett. **37** (1997) 195–200

20.26 M. Weiss, F. J. Elmer: Dry friction in the Frenkel–Kontorova–Tomlinson model: Static properties, Phys. Rev. B **53** (1996) 7539–7549

20.27 J. B. Pethica: Comment on "Interatomic forces in scanning tunneling microscopy: Giant corrugations of the graphite surface", Phys. Rev. Lett. **57** (1986) 3235

20.28 E. Meyer, R. M. Overney, K. Dransfeld, T. Gyalog: *Nanoscience, Friction and Rheology on the Nanometer Scale* (World Scientific, Singapore 1998)

20.29 L. Howald, R. Lüthi, E. Meyer, H.-J. Güntherodt: Atomic-force microscopy on the Si(111)7 × 7 surface, Phys. Rev. B **51** (1995) 5484–5487

20.30 R. Bennewitz, T. Gyalog, M. Guggisberg, M. Bammerlin, E. Meyer, H.-J. Güntherodt: Atomic-scale stick-slip processes on Cu(111), Phys. Rev. B **60** (1999) R11301–R11304

20.31 R. Lüthi, E. Meyer, M. Bammerlin, L. Howald, H. Haefke, T. Lehmann, C. Loppacher, H.-J. Güntherodt, T. Gyalog, H. Thomas: Friction on the atomic scale: An ultrahigh vacuum atomic force

microscopy study on ionic crystals, J. Vac. Sci. Technol. B **14** (1996) 1280–1284
20.32 G. J. Germann, S. R. Cohen, G. Neubauer, G. M. McClelland, H. Seki: Atomic scale friction of a diamond tip on diamond (100) and (111) surfaces, J. Appl. Phys. **73** (1993) 163–167
20.33 R. J. A.van den Oetelaar, C. F. J. Flipse: Atomic-scale friction on diamond(111) studied by ultra-high vacuum atomic force microscopy, Surf. Sci. **384** (1997) L828–L835
20.34 R. Bennewitz, E. Gnecco, T. Gyalog, E. Meyer: Atomic friction studies on well-defined surfaces, Tribol. Lett. **10** (2001) 51–56
20.35 S. Fujisawa, E. Kishi, Y. Sugawara, S. Morita: Atomic-scale friction observed with a two-dimensional frictional-force microscope, Phys. Rev. B **51** (1995) 7849–7857
20.36 N. Sasaki, M. Kobayashi, M. Tsukada: Atomic-scale friction image of graphite in atomic-force microscopy, Phys. Rev. B **54** (1996) 2138–2149
20.37 H. Kawakatsu, T. Saito: Scanning force microscopy with two optical levers for detection of deformations of the cantilever, J. Vac. Sci. Technol. B **14** (1996) 872–876
20.38 M. Liley, D. Gourdon, D. Stamou, U. Meseth, T. M. Fischer, C. Lautz, H. Stahlberg, H. Vogel, N. A. Burnham, C. Duschl: Friction anisotropy and asymmetry of a compliant monolayer induced by a small molecular tilt, Science **280** (1998) 273–275
20.39 R. Lüthi, E. Meyer, H. Haefke, L. Howald, W. Gutmannsbauer, H.-J. Güntherodt: Sled-type motion on the nanometer scale: Determination of dissipation and cohesive energies of C_{60}, Science **266** (1994) 1979–1981
20.40 M. R. Falvo, J. Steele, R. M. Taylor, R. Superfine: Evidence of commensurate contact and rolling motion: AFM manipulation studies of carbon nanotubes on HOPG, Tribol. Lett. **9** (2000) 73–76
20.41 M. Hirano, K. Shinjo, R. Kaneko, Y. Murata: Anisotropy of frictional forces in muscovite mica, Phys. Rev. Lett. **67** (1991) 2642–2645
20.42 M. Hirano, K. Shinjo, R. Kaneko, Y. Murata: Observation of superlubricity by scanning tunneling microscopy, Phys. Rev. Lett. **78** (1997) 1448–1451
20.43 R. M. Overney, H. Takano, M. Fujihira, W. Paulus, H. Ringsdorf: Anisotropy in friction and molecular stick-slip motion, Phys. Rev. Lett. **72** (1994) 3546–3549
20.44 H. Takano, M. Fujihira: Study of molecular scale friction on stearic acid crystals by friction force microscopy, J. Vac. Sci. Technol. B **14** (1996) 1272–1275
20.45 P. E. Sheehan, C. M. Lieber: Nanotribology and nanofabrication of MoO_3 structures by atomic force microscopy, Science **272** (1996) 1158–1161
20.46 E. Gnecco, R. Bennewitz, T. Gyalog, Ch. Loppacher, M. Bammerlin, E. Meyer, H.-J. Güntherodt: Velocity dependence of atomic friction, Phys. Rev. Lett. **84** (2000) 1172–1175
20.47 D. Gourdon, N. A. Burnham, A. Kulik, E. Dupas, F. Oulevey, G. Gremaud, D. Stamou, M. Liley, Z. Dienes, H. Vogel, C. Duschl: The dependence of friction anisotropies on the molecular organization of LB films as observed by AFM, Tribol. Lett. **3** (1997) 317–324
20.48 Y. Sang, M. Dubé, M. Grant: Thermal effects on atomic friction, Phys. Rev. Lett. **87** (2001) 174301
20.49 E. Riedo, E. Gnecco, R. Bennewitz, E. Meyer, H. Brune: Interaction potential and hopping dynamics governing sliding friction, Phys. Rev. Lett. **91** (2003) 084502
20.50 O. Zwörner, H. Hölscher, U. D. Schwarz, R. Wiesendanger: The velocity dependence of frictional forces in point-contact friction, Appl. Phys. A **66** (1998) 263–267
20.51 T. Bouhacina, J. P. Aimé, S. Gauthier, D. Michel, V. Heroguez: Tribological behavior of a polymer grafted on silanized silica probed with a nanotip, Phys. Rev. B **56** (1997) 7694–7703
20.52 H. J. Eyring: The activated complex in chemical reactions, J. Chem. Phys. **3** (1937) 107
20.53 J. N. Glosli, G. M. McClelland: Molecular dynamics study of sliding friction of ordered organic monolayers, Phys. Rev. Lett. **70** (1993) 1960–1963
20.54 O. Pfeiffer, R. Bennewitz, A. Baratoff, E. Meyer, P. Grütter: Lateral-force measurements in dynamic force microscopy, Phys. Rev. B **65** (2002) 161403
20.55 E. Riedo, F. Lévy, H. Brune: Kinetics of capillary condensation in nanoscopic sliding friction, Phys. Rev. Lett. **88** (2002) 185505
20.56 M. He, A. S. Blum, G. Overney, R. M. Overney: Effect of interfacial liquid structuring on the coherence length in nanolubrucation, Phys. Rev. Lett. **88** (2002) 154302
20.57 F. P. Bowden, F. P. Tabor: *The Friction and Lubrication of Solids* (Oxford Univ. Press, Oxford 1950)
20.58 L. D. Landau, E. M. Lifshitz: *Introduction to Theoretical Physics* (Nauka, Moscow 1998) Vol. 7
20.59 J. A. Greenwood, J. B. P. Williamson: Contact of nominally flat surfaces, Proc. R. Soc. Lond. A **295** (1966) 300
20.60 B. N. J. Persson: Elastoplastic contact between randomly rough surfaces, Phys. Rev. Lett. **87** (2001) 116101
20.61 K. L. Johnson, K. Kendall, A. D. Roberts: Surface energy and contact of elastic solids, Proc. R. Soc. Lond. A **324** (1971) 301
20.62 U. D. Schwarz, O. Zwörner, P. Köster, R. Wiesendanger: Quantitative analysis of the frictional properties of solid materials at low loads, Phys. Rev. B **56** (1997) 6987–6996
20.63 B. V. Derjaguin, V. M. Muller, Y. P. Toporov: Effect of contact deformations on adhesion of particles, J. Colloid Interface Sci. **53** (1975) 314–326
20.64 D. Tabor: Surface forces and surface interactions, J. Colloid Interface Sci. **58** (1977) 2–13

20.65 D. Maugis: Adhesion of spheres: the JKR-DMT transition using a Dugdale model, J. Colloid Interface Sci. **150** (1992) 243–269

20.66 R.W. Carpick, N. Agraït, D.F. Ogletree, M. Salmeron: Measurement of interfacial shear (friction) with an ultrahigh vacuum atomic force microscope, J. Vac. Sci. Technol. B **14** (1996) 1289–1295

20.67 C. Polaczyk, T. Schneider, J. Schöfer, E. Santner: Microtribological behavior of Au(001) studied by AFM/FFM, Surf. Sci. **402** (1998) 454–458

20.68 J.N. Israelachvili, D. Tabor: Measurement of van der Waals dispersion forces in range 1.5 to 130 nm, Proc. R. Soc. Lond. A **331** (1972) 19

20.69 S.P. Jarvis, A. Oral, T.P. Weihs, J.B. Pethica: A novel force microscope and point-contact probe, Rev. Sci. Instrum. **64** (1993) 3515–3520

20.70 R.W. Carpick, D.F. Ogletree, M. Salmeron: Lateral stiffness: A new nanomechanical measurement for the determination of shear strengths with friction force microscopy, Appl. Phys. Lett. **70** (1997) 1548–1550

20.71 M.A. Lantz, S.J. O'Shea, M.E. Welland, K.L. Johnson: Atomic-force-microscope study of contact area and friction on $NbSe_2$, Phys. Rev. B **55** (1997) 10776–10785

20.72 K.L. Johnson: *Contact Mechanics* (Cambridge Univ. Press, Cambridge 1985)

20.73 M. Enachescu, R.J.A. van den Oetelaar, R.W. Carpick, D.F. Ogletree, C.F.J. Flipse, M. Salmeron: Atomic force microscopy study of an ideally hard contact: the diamond(111)/tungsten carbide interface, Phys. Rev. Lett. **81** (1998) 1877–1880

20.74 M. Enachescu, R.J.A. van den Oetelaar, R.W. Carpick, D.F. Ogletree, C.F.J. Flipse, M. Salmeron: Observation of proportionality between friction and contact area at the nanometer scale, Tribol. Lett. **7** (1999) 73–78

20.75 E. Gnecco, R. Bennewitz, E. Meyer: Abrasive wear on the atomic scale, Phys. Rev. Lett. **88** (2002) 215501

20.76 S. Kopta, M. Salmeron: The atomic scale origin of wear on mica and its contribution to friction, J. Chem. Phys. **113** (2000) 8249–8252

20.77 H. Tang, C. Joachim, J. Devillers: Interpretation of AFM images – the graphite surface with a diamond tip, Surf. Sci. **291** (1993) 439–450

20.78 U. Landman, W.D. Luedtke, E.M. Ringer: Atomistic mechanisms of adhesive contact formation and interfacial processes, Wear **153** (1992) 3–30

20.79 A.I. Livshits, A.L. Shluger: Self-lubrication in scanning-force-microscope image formation on ionic surfaces, Phys. Rev. B **56** (1997) 12482–12489

20.80 H. Tang, X. Bouju, C. Joachim, C. Girard, J. Devillers: Theoretical study of the atomic-force-microscopy imaging process on the NaCl(100) surface, J. Chem. Phys. **108** (1998) 359–367

20.81 R. Bennewitz, A.S. Foster, L.N. Kantorovich, M. Bammerlin, Ch. Loppacher, S. Schär, M. Guggisberg, E. Meyer, A.L. Shluger: Atomically resolved edges and kinks of NaCl islands on Cu(111): Experiment and theory, Phys. Rev. B **62** (2000) 2074–2084

20.82 U. Landman, W.D. Luetke, M.W. Ribarsky: Structural and dynamical consequences of interactions in interfacial systems, J. Vac. Sci. Technol. A **7** (1989) 2829–2839

20.83 M.R. Sørensen, K.W. Jacobsen, P. Stoltze: Simulations of atomic-scale sliding friction, Phys. Rev. B **53** (1996) 2101–2113

20.84 A. Buldum, C. Ciraci: Contact, nanoindentation, and sliding friction, Phys. Rev. B **57** (1998) 2468–2476

20.85 R. Komanduri, N. Chandrasekaran, L.M. Raff: Molecular dynamics simulation of atomic-scale friction, Phys. Rev. B **61** (2000) 14007–14019

20.86 T.H. Fang, C.I. Weng, J.G. Chang: Molecular dynamics simulation of nanolithography process using atomic force microscopy, Surf. Sci. **501** (2002) 138–147

20.87 B. Gotsmann, C. Seidel, B. Anczykowski, H. Fuchs: Conservative and dissipative tip-sample interaction forces probed with dynamic AFM, Phys. Rev. B **60** (1999) 11051–11061

20.88 C. Loppacher, R. Bennewitz, O. Pfeiffer, M. Guggisberg, M. Bammerlin, S. Schär, V. Barwich, A. Baratoff, E. Meyer: Experimental aspects of dissipation force microscopy, Phys. Rev. B **62** (2000) 13674–13679

20.89 M. Gauthier, M. Tsukada: Theory of noncontact dissipation force microscopy, Phys. Rev. B **60** (1999) 11716–11722

20.90 J.P. Aimé, R. Boisgard, L. Nony, G. Couturier: Nonlinear dynamic behavior of an oscillating tip-microlever system and contrast at the atomic scale, Phys. Rev. Lett. **82** (1999) 3388–3391

20.91 W. Denk, D.W. Pohl: Local electrical dissipation imaged by scanning force microscopy, Appl. Phys. Lett. **59** (1991) 2171–2173

20.92 S. Hirsekorn, U. Rabe, A. Boub, W. Arnold: On the contrast in eddy current microscopy using atomic force microscopes, Surf. Interface Anal. **27** (1999) 474–481

20.93 U. Dürig: Forces in scanning probe methods. In: *NATO ASI*, Ser. E, Vol. 286, ed. by H.J. Güntherodt, D. Anselmetti, E. Meyer (Kluwer, Dordrecht 1995) pp. 353–366

20.94 N. Sasaki, M. Tsukada: Effect of microscopic nonconservative process on noncontact atomic force microscopy, Jpn. J. of Appl. Phys. **39** (2000) L1334–L1337

20.95 B. Gotsmann, H. Fuchs: The measurement of hysteretic forces by dynamic AFM, Appl. Phys. A **72** (2001) 55–58

20.96 M. Guggisberg, M. Bammerlin, A. Baratoff, R. Lüthi, C. Loppacher, F.M. Battiston, J. Lü, R. Bennewitz,

E. Meyer, H. J. Güntherodt: Dynamic force microscopy across steps on the Si(111)-(7 × 7) surface, Surf. Sci. **461** (2000) 255–265
20.97 R. Bennewitz, S. Schär, V. Barwich, O. Pfeiffer, E. Meyer, F. Krok, B. Such, J. Kolodzej, M. Szymonski: Atomic-resolution images of radiation damage in KBr, Surf. Sci. **474** (2001) 197–202
20.98 T. D. Stowe, T. W. Kenny, J. Thomson, D. Rugar: Silicon dopant imaging by dissipation force microscopy, Appl. Phys. Lett **75** (1999) 2785–2787
20.99 B. C. Stipe, H. J. Mamin, T. D. Stowe, T. W. Kenny, D. Rugar: Noncontact friction and force fluctuations between closely spaced bodies, Phys. Rev. Lett **87** (2001) 96801
20.100 B. Gotsmann, H. Fuchs: Dynamic force spectroscopy of conservative and dissipative forces in an Al-Au(111) tip-sample system, Phys. Rev. Lett. **86** (2001) 2597–2600
20.101 B. N. J. Persson, A. I. Volokitin: Comment on "Brownian motion of microscopic solids under the action of fluctuating electromagnetic fields", Phys. Rev. Lett. **84** (2000) 3504
20.102 K. Yamanaka, A. Noguchi, T. Tsuji, T. Koike, T. Goto: Quantitative material characterization by ultrasonic AFM, Surf. Interface Anal. **27** (1999) 600–606
20.103 T. Drobek, R. W. Stark, W. M. Heckl: Determination of shear stiffness based on thermal noise analysis in atomic force microscopy: Passive overtone microscopy, Phys. Rev. B **64** (2001) 045401
20.104 T. Kawagishi, A. Kato, Y. Hoshi, H. Kawakatsu: Mapping of lateral vibration of the tip in atomic force microscopy at the torsional resonance of the cantilever, Ultramicroscopy **91** (2002) 37–48
20.105 F. J. Giessibl, M. Herz, J. Mannhart: Friction traced to the single atom, PNAS **99** (2002) 12006–12010
20.106 H.-J. Hug, A. Baratoff: Measurement of dissipation induced by tip-sample interactions. In: *Noncontact Atomic Force Microscopy*, ed. by S. Morita, R. Wiesendanger, E. Meyer (Springer, Berlin, Heidelberg 2002) p. 395

21. Nanoscale Mechanical Properties – Measuring Techniques and Applications

The first part of this chapter describes local (at the scale of nanometers) measurements of mechanical properties. It includes detailed state-of-the-art presentation and in-depth analysis of experimental techniques, results, and interpretations.

After a short introduction, the second part describes local mechanical spectroscopy using coupled Atomic Force Microscopy and ultrasound. This technique allows us to map quickly not only spatial distribution of the elasticity but anelastic properties as well. At one point in the sample, semi-quantitative measurements can be made as a function of the temperature. On the nanometer scale, results have close similitudes to bulk measurements and interpretable differences. Local elasticity and damping were measured during phase transition of polymer samples and shape-memory alloys.

The third part describes the "nano-Swiss cheese" method of measuring the elastic properties of such tubular nanometer size objects as carbon nanotubes and microtubules. It is probably the only experiment in which properties of single-wall nanotube ropes were measured as a function of the rope diameter. We extended this idea to biological objects, microtubules, and successfully solved major experimental difficulties. We not only measured the temperature dependency of microtubule modulus in pseudo-physiological conditions but also estimated shear modulus using the same microtubule with several lengths of suspended segments.

The fourth section demonstrates the scanning nanoindentation technique as applied to human bone tissue. This instrument allows performing topography scans and indentation tests using the identical tip. The available surface scan allows a high positioning precision of the indenter tip on the structure of interest. For very inhomogeneous samples, such as bone tissue, this tool provides a probe to detect local variations of the mechanical properties. The indentation test supplies quantitative parameters like elastic modulus and hardness on the submicron level. Local mechanical properties of compact and trabecular bone lamellae were tested under both dry and pseudo-physiological conditions.

Finally, last part is given to a discussion of future prospects and conclusions.

21.1 **Local Mechanical Spectroscopy by Contact AFM** ... 662
  21.1.1 The Variable-Temperature SLAM (T-SLAM) ... 663
  21.1.2 Example One: Local Mechanical Spectroscopy of Polymers ... 664
  21.1.3 Example Two: Local Mechanical Spectroscopy of NiTi ... 665

21.2 **Static Methods – Mesoscopic Samples** ... 667
  21.2.1 Carbon Nanotubes – Introduction to Basic Morphologies and Production Methods ... 667
  21.2.2 Measurements of the Mechanical Properties of Carbon Nanotubes by SPM ... 668
  21.2.3 Microtubules and Their Elastic Properties ... 673

21.3 **Scanning Nanoindentation: An Application to Bone Tissue** ... 674
  21.3.1 Scanning Nanoindentation ... 674
  21.3.2 Application of Scanning Nanoindentation ... 674
  21.3.3 Example: Study of Mechanical Properties of Bone Lamellae Using SN ... 675
  21.3.4 Conclusion ... 681

21.4 **Conclusions and Perspectives** ... 682

References ... 682

Experiments measuring the mechanical properties at nanoscale are key to understanding mesoscopic materials and inhomogeneous materials.

The most prominent technique for such local measurements implements an Atomic Force Microscopy in static mode to obtain force–distance curves (f-d). Ideally this provides the force applied to the tip and the tip–sample distance allowing the local reduced Young's modulus to be determined. In practice, however, these values are not measured directly. Instead, the applied force is offset by adhesive and capillary forces. Furthermore, the tip–sample distance must be obtained by subtraction of the sample (Z-piezo) displacement and the deflection of the cantilever, neither of which are calibrated on most commercial instruments. Usually only the voltage applied to nonlinear and hysteretic Z-piezo and the uncalibrated cantilever deflection is available. These available data must be tediously calibrated and converted to real displacements and forces. The calibration procedure given by *Radmacher* [21.1] works well but only when applied to very compliant materials such as biological or polymer samples. An additional difficulty is that since one end of the cantilever is fixed, neither the orientation of the tip nor its position on the sample surface remains constant during f-d curve acquisition. Movement of the tip on the sample surface, therefore occurs not only vertically but also along the cantilever axis. As a result, applied contact mechanics may not be valid. Finally, uncertainty about the exact tip shape and radius complicate the procedure even further. Instrumental challenges are so significant that applying static AFM to obtain absolute values of the elastic modulus using commercial instruments is difficult at best, especially on stiff surfaces. And this, of course, is the motivation for developing of other techniques.

Although absolute values of mechanical properties are difficult to measure, relative measurements or maps (images of physical properties) are still very interesting. Using acoustic vibrations of the AFM sample surface one can access local elastic and anelastic properties. The first part of this chapter describes these techniques and applications.

Studies of mesoscopic specimens are even more challenging: it is difficult and time consuming to locate a sample, specially properly positioned. In the second part of the chapter, we used 'nano Swiss cheese' technique to get insight into the axial and shear modulus of several species of carbon nanotubes. The method was further extended to biological microtubules, where The experiment was performed in liquid and as a function of the temperature.

Reliable mechanical measurements can be obtained by combining the best of two worlds, AFM and nanoindenter. Classical nanoindenters use optical tip positioning, which are inadequate in the nanoworld. Scanning nanoindenter uses the same diamond tip for imaging (with microNewton forces) and measuring. Its application to human bones in liquid and 37 °C is described in the third part of this chapter.

One should underline the importance of AFM linearity and stability for such experiments: only good designs may limit frustration and experiment time of courageous PhD students, and lot of improvements are still lacking.

21.1 Local Mechanical Spectroscopy by Contact AFM

Mechanical properties of solids (elasticity, anelasticity, plasticity) are generally measured on macroscopic samples. But many phenomena in materials science require measurements of mechanical properties at the surface of a material or at the interfaces between thin layers deposited on a surface. High spatial resolution is also important, for example in the cases of multiphased materials or composite materials, phase transitions, lattice softening in shape-memory alloys, precipitation in light alloys, and glass transition of amorphous materials.

In the case of multiphased materials, such as nanomaterials, composites, alloys, or polymer blends, the location of the dissipative mechanism in one phase must be determined either through modeling or by separately studying each phase, when possible, without changing its behavior. The latter is only possible in a limited number of cases due to the interactions between the different phases within a material. To give an example, the global behavior of a composite is mostly driven by the stress transfer properties between the reinforcement and the matrix, which are controlled by the local mechanical properties in the interface region and in particular the dynamics of the structural defects in this area. It is obviously impossible to prepare a sample only composed of interface regions. Therefore a method for locally studying the dynamics of the structural defects will thus provide important steps in the understanding and the improvement of such materials.

Different techniques based on Scanning Probe Microscopies (SPM) have been developed to probe the elastic and anelastic properties of surfaces, interfaces, or phases of inhomogeneous materials at the micrometer and the nanometer scales. Scanning Acoustic Microscopy (SAM), first developped in the mid-1970s, allows one to study the materials properties at the micrometer scale.

Among the different ways explored to study local mechanical properties of materials, several groups have recently used techniques based on Scanning Force Microscopy (SFM) [21.2]. For most of these, the focus has been placed on "elasticity," using the so-called Force Modulation Mode (FMM) at low frequencies [21.3]. Force modulation mode generally uses a large amplitude (greater than 10 nm), low frequency (some kHz) vibration of the sample underneath the SFM tip. The component of the tip motion at the excitation frequency and the tip mean position are simultaneously recorded, giving several images of the sample surface. In particular, the in-phase and out-of-phase components of force modulation mode at room temperatures have been interpreted in terms of stiffness ("elasticity") and damping ("viscoelasticity") [21.4, 5]. But, it has been recently shown by *Mazeran* et al. [21.6] that the contrast of force modulation mode is dominated by friction properties, giving only little information on the elasticity. Consequently, some care has to be taken in the interpretation of these low-frequency studies. A way to suppress the influence of friction on the contrast is to use smaller amplitudes (some Å) at higher frequencies [21.7]. Scanning Local-Acceleration Microscopy (SLAM) implements this idea [21.8]: SLAM is a modification of contact-mode scanning force microscopy. Its principle is to vibrate the sample at a frequency just above the resonance of the tip–sample system. In this case, the inertia of the tip prevents it from completely following the imposed high frequency displacement, inducing nonnegligible forces and giving rise to elastic deformation of the sample. Contact stiffness is obtained from the measure of the residual displacement of the tip. Mapping the contact stiffness at different temperatures with SLAM [21.9] has allowed local mechanical spectroscopy. Some other techniques also use high frequencies but with different approaches to image elasticity at room temperature [21.7, 10, 11]. Each of these high-frequency techniques is capable of mapping properties such as stiffness or adhesion at constant temperature with a very high lateral resolution.

21.1.1 The Variable-Temperature SLAM (T-SLAM)

The technique described here combines the lateral resolution of scanning force microscopy with the physical information available from temperature ramps. It is the variable-temperature SLAM (T-SLAM) [21.12, 13]. Ramping the temperature of the sample during a SLAM measurement and acquiring both the amplitude and the phase of the tip's motion allows one to obtain local mechanical spectroscopy data. A simple model enables the interpretation of the measurement in terms of material properties.

T-SLAM is based on SLAM [21.8, 14], using a variable temperature sample holder [21.9, 15]. An ultrasonic transducer is placed beneath the sample in a commercial scanning force microscope (Fig. 21.1) and excited by means of a function generator. The ultrasound is

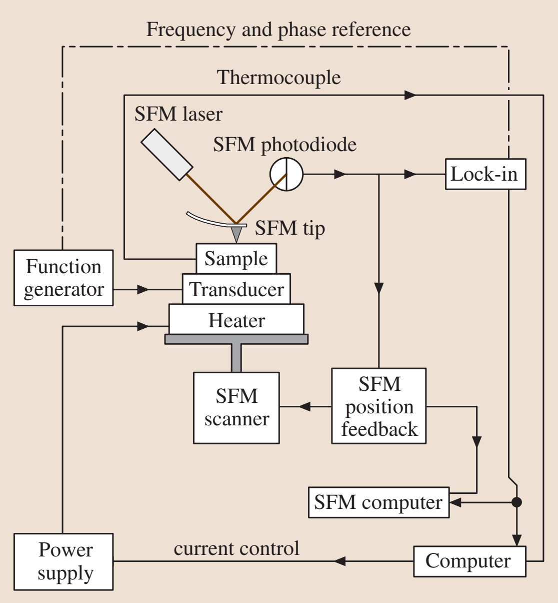

Fig. 21.1 *Schematic diagram of T-SLAM*. Ultrasound is generated with a transducer connected to a function generator and placed beneath the sample. The motion of the scanning force microscope tip is detected optically. The detection signal is fed to a lock-in amplifier, which extracts amplitude and phase relative to the transducer's motion. The temperature is controlled with a small resistive heater and measured with a thermocouple. The rest of the microscope head, to which the tip, the laser and the photodiode are attached, is not represented in this diagram

transmitted through the sample, forcing periodic local deformation of the sample's surface underneath the tip. The motion of the scanning force microscopy tip is detected optically by laser beam deflection from the backside of the cantilever. The detection signal is then fed to a lock-in amplifier, which extracts the tip's amplitude (related to elasticity) and phase (related to viscoelasticity) relative to the transducer's motion. The transducer's typical frequency is 825 kHz. This installation can operate in two ways, either by mapping the amplitude and phase of the signal as a function of position at fixed temperature or by recording the amplitude and phase as a function of temperature at a "fixed" location. The first method is known as "SLAM imaging," where the output signal of the lock-in is fed into an auxiliary data acquisition channel of the microscope. An extra computer is used to store local mechanical spectroscopy data as a function of the temperature at a fixed location and to control temperature. The heat is produced with a small resistive heater below the transducer, and the temperature is measured with a thermocouple. The sample must be prepared with a sufficiently low surface roughness in order to avoid artifacts in the measurements due to the sample's topography. Due to the geometry of the contact, SLAM measures only a small volume near the surface. Viewing SLAM as a very fast indentation measurement, the probed volume is approximately a half-sphere with a radius of $10a$, where a is the contact radius between tip and sample [21.16]. The typical value of a is some nm. So even if the sample is a thick film, T-SLAM gives access to the near-surface mechanical properties that may differ from bulk.

The mechanical properties of the deformed region are obtained from the measure of the residual displacement of the tip. The tip vibration amplitude d_1 is related to the contact stiffness, proportional to the dynamic elastic modulus, while the phase lag φ between the tip motion and the surface motions is related to the internal friction (energy dissipation inside the deformed volume). Mapping d_1 and φ at different temperatures with T-SLAM allows the study of the homogeneity of mechanical relaxations or of phase transitions. By recording d_1 and φ as a function of temperature at a fixed location, local mechanical spectroscopy can be performed.

When used just above the first resonance of the tip–sample system, the SLAM system can be described and analyzed using a point mass rheological model [21.9, 12], which allows one to obtain the equation relating the damping (loss factor η') to the measured parameters

$$\eta' = \left(\frac{1}{2}\frac{k_\mathrm{e}}{k_\mathrm{c} - m\omega^2}\right)\frac{\sin\varphi}{d_1/z_1}\,, \tag{21.1}$$

where ω is the measurement frequency, k_c is a parameter related to the elastic modulus of the AFM cantilever and k_e to the elastic modulus of the sample, z_1 is the transducer vibration amplitude, and $m = k_\mathrm{c}/\omega_\mathrm{c}^2$ is the equivalent point mass of the tip, where ω_c is the free resonant frequency of the cantilever.

The principal limitation of (21.1) comes from the model assumption that the cantilever is a point mass restricted to vertical displacements. Due to the possible lateral displacements of the tip [21.6] or the existence of other vibrational modes of the cantilever, which are not described by the point mass model, (21.1) is only reliable for the measurement of the damping just above the first resonance of the tip–sample system. In order to obtain a quantitative measurement of this quantity in a larger frequency spectrum, it is necessary to develop a more realistic model of the cantilever interacting with the sample surface such as, for example, the model presented by *Dupas* et al. [21.17], in which the cantilever is modeled as a beam.

21.1.2 Example One: Local Mechanical Spectroscopy of Polymers

Figure 21.2 shows local mechanical spectra [21.13, 18] obtained in bulk technical PVC (including plasticizer and pigments and taken off-the-shelf) as a function of temperature (Fig. 21.2a) and a Differential Scanning Calorimetry (DSC) measurement of the same sample (Fig. 21.2b). The mechanical measurement (Fig. 21.2a) displays the amplitude of vibration of the SLAM tip (thin line) and its phase lag (thick line) as a function of temperature. Four temperature domains can be identified. At lower temperatures (1), the first domain shows a phase lag peak associated with a decrease of amplitude. The second temperature (2) domain corresponds to a zone where the vibration amplitude increases, without any variation in the phase lag. The third temperature domain (3) is characterized by a large phase lag peak and a significant decrease of vibration amplitude. The last temperature domain (4) shows a slow increase in the phase lag without variation of the vibration amplitude.

The calorimetry curves (b) display the heat flow as a function of temperature for the same material but at a larger size scale. For clarity, the same temperature do-

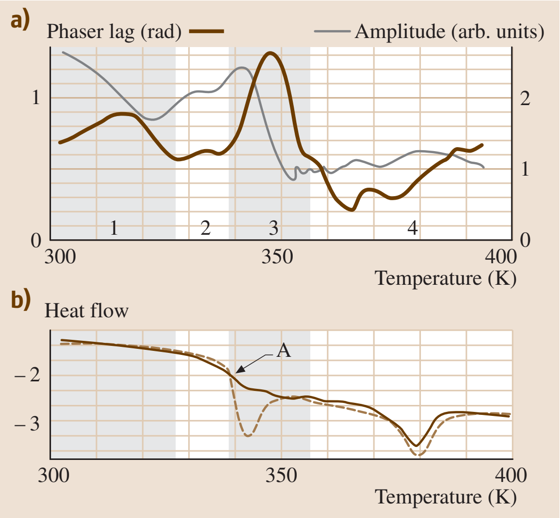

Fig. 21.2 **(a)** *Local mechanical spectroscopy spectra as a function of temperature of a technical PVC*: The vibration amplitude is displayed as a *thin line* and the phase lag as a *thick line*. Four temperature domains can be distinguished: From left to right, a small "vibration amplitude" decrease is associated with a first "phase lag" peak (1), in the next domain, "vibration amplitude" increases (2), then a large decrease of "vibration amplitude" is associated with a large "phase lag" peak (3) and finally "phase lag" increases slowly (4). **(b)** *DSC measurements of the same sample*. For clarity, the temperature domains observed on Fig. 21.5a have been reported. The graph displays the first (*dashed line*) and the second (*solid line*) heating. The glass transition can easily be recognized around 340 K (labeled *A*), slightly below the temperature range of domain 3. An irreversible endothermic relaxation takes place in the same temperature range (only visible on the first run). A reversible endothermic event occurs around 380 K

mains have been reported on the graph. The dashed line displays the first heating, where both reversible and irreversible events are present; the solid line shows the second heating, where only reversible events are still present. On the solid line, an endothermic event (labeled A) can be observed near the border between temperature domains 2 and 3. An irreversible endothermic relaxation is superimposed with (A) on the first heating. A reversible endothermic peak is observed around 380 K, in domain 4. The endothermic event (A) has the characteristics of a glass transition: the specific heat goes from one value to another without a peak [21.19]. The glass transition temperature of this PVC is, therefore, approximately 340 K. The irreversible relaxation occurring in the same temperature range is certainly associated with physical aging or structural relaxation. The reversible peak at higher temperature is associated with the melting of the small crystalline volume fraction.

These local mechanical spectroscopy results are in good qualitative agreement with macroscopic global measurements [21.13, 18]. They yield the same information as the macroscopic global measurements, but from a much smaller volume, and allow as a consequence the location of the different mechanisms. Both measurements (local and global) show a peak in phase for the primary and secondary relaxations, connected with a large decrease of the stiffness for the primary relaxation and a much smaller one for the secondary relaxation. Plasticity induces an increase of the phase lag in both cases. The temperatures observed for the relaxation related to the glass transition compare well with the calorimetric data. Based on all these examples, there is no doubt that the amplitude and phase lag of T-SLAM are functions of the stiffness and damping. In this regard, T-SLAM provides an extension of the global method, allowing location of dissipative phenomena and study of the spatial homogeneity of phase transitions or of relaxations. This will bring new insight into the field of inhomogeneous or composite materials. In addition, T-SLAM measures only a small volume near the surface. So even if the sample is a thick film, T-SLAM gives access to the near-surface mechanical properties, which may differ from bulk properties. This will bring interesting perspectives to the still debated field of surface dynamics.

21.1.3 Example Two: Local Mechanical Spectroscopy of NiTi

Near-stoichiometric NiTi alloys exhibit a martensitic phase transformation between a low-temperature monoclinic phase, called martensite, and a high temperature cubic phase with B2 structure, called austenite (1). This transformation is responsible for the shape memory and pseudo-elastic effects in deformed NiTi alloys. Upon transformation to the martensitic phase, an intermediate rhombohedral (R) phase can be formed [21.20, 21]. Although *Bataillard* [21.21] tends to demonstrate that the R phase is finely dispersed inside the material, a controversy remains over the spatial scale at which the decomposition of the transformation occurs. Another puzzling question concerns the transformation itself. Optical microscopy observation suggests that the transformation occurs very suddenly inside an austenite grain. This has led to the concept of "military transformation." The width of the globally measured transformation would then be a sum of different narrow contributions coming from different places inside the sample. This image is, however, not universally accepted.

A measurement inside one single grain of a polycrystal would be a way to address these questions. Both the spatial scale of the R phase distribution and the "military" character of the transformation will have an effect on the result of such a measurement. This is the reason for which local mechanical spectroscopy measurements of the martensitic phase transition of NiTi by T-SLAM have been performed [21.12, 22]. They are presented in Fig. 21.3. The global transformation behavior of NiTi is defined by calorimetry spectra (Fig. 21.3b, temperature scanning rate 10 K/min).

The vibration amplitude ("elasticity") and the phase lag ("internal friction") are shown in the top and bottom curves, respectively. The presented data is incomplete due to experimental limitations. A phase lag peak can be observed upon heating at approximately 370 K, associated with a change in the vibration amplitude level. This event can be correlated with the phase transformation from martensite to austenite observed by calorimetry (Fig. 21.3b). Upon cooling, no event is observed around 370 K: the vibration amplitude is stable and the phase lag curve does not exhibit any peak. But, an intense phase lag peak can be observed upon cooling near 330 K, associated with a restoration of the vibration amplitude level characteristic of the martensite. This event correlates with the phase transformation from austenite to martensite observed by calorimetry. A change in the vibration amplitude spectrum is correlated with each event on the phase lag spectrum. Two main features should be noted. First, the transformation peaks observed with the SLAM method are narrower than those observed with the (global) calorimetry measurement. Second, the measured temperatures for the local peak are located on the high temperature side of the peaks measured by calorimetry. Details of the two transformation peaks are displayed in Figs. 21.3c and 21.3d. The reverse transformation (Fig. 21.3c) is characterized by a change in the vibration amplitude level and a phase lag peak. This peak (A1) may have a shoulder on the low temperature

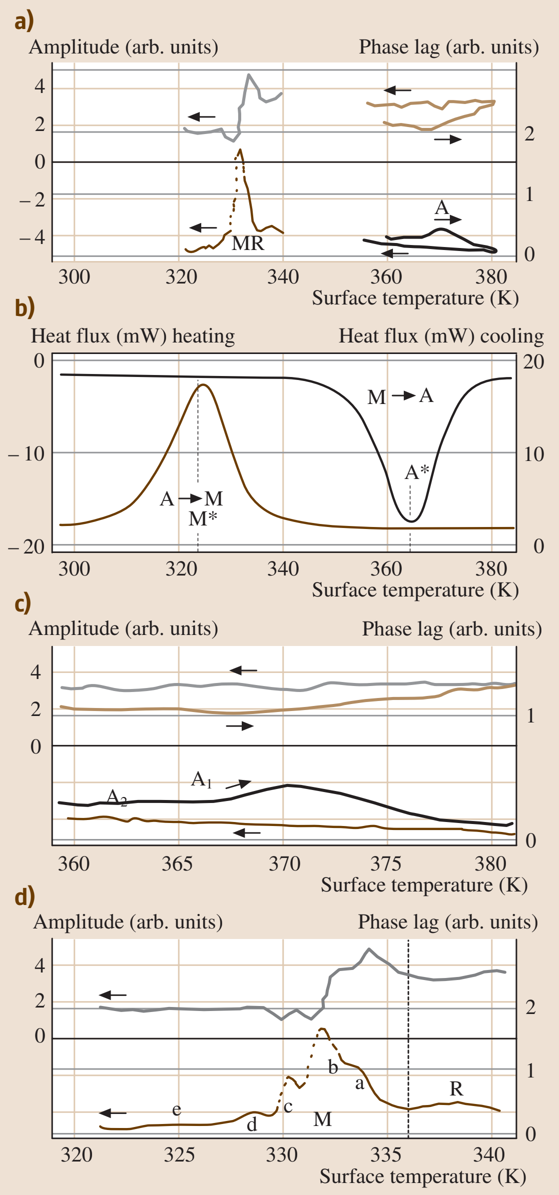

Fig. 21.3 **(a)–(b)** *Comparison of the local mechanical spectroscopy measurements of* NiTi *with calorimetric experiment*. **(a)** Both the reverse and direct transformation are associated with a phase lag peak and a modulus variation. **(b)** Calorimetric measurement of the same sample. **(c)–(d)** *Zoom on the reverse* **(c)** *and direct* **(d)** *transformation's temperature ranges*. The peak (A1) may exhibit a shoulder on the low temperature side (A2). The direct transformation is characterized by a recovery of the original vibration amplitude and a complex phase lag spectrum, formed of two main maxima, noted R and M. Peaks a,b,c,d,e are a substructure of the M peak

side (A2). A small decrease of the vibration amplitude may precede this increase. But, the magnitude of this softening and the very low intensity of peak A2 are too small to exclude experimental artifacts. The direct transformation (Fig. 21.3d) displays a recovery of the original vibration amplitude and a complicated phase lag maximum.

The correlation of the transition temperatures measured by calorimetry and the local mechanical spectroscopy lead to the conclusion that the peaks observed in phase lag and the change in vibration amplitude originate from the martensitic phase transition. At the scale of the observation, the martensite transforms into austenite around 370 K (A1 peak) and the austenite transforms into martensite close to 332 K (M peak), with the formation of the rhombohedral R phase at 337 K (R peak). The peaks A2 and e could be linked to the presence of two different types of martensite as already observed elsewhere [21.20, 21]. It is striking to note that all the events observed in macroscopic experiments seem to be reproducible at the scale of observation, namely less than $10^{-3}\,\mu\text{m}^3$. In particular, this would confirm that the R phase precipitates are very finely distributed in the austenite matrix as observed by Bataillard [21.21]. Otherwise the R peak could not be observed reproducibly using such a technique. The local measurement differs from global ones in at least two aspects: first, the width of the transformation temperature range is smaller; second, the M peak exhibits a substructure.

Global internal friction measurements on these materials have given spectra with peaks having a similar breadth as the calorimetry peaks [21.23], whereas the peaks measured locally are narrower. This is easy to understand if the martensitic transformation occurs inhomogeneously inside the sample. Grains tend to transform at different temperatures depending on their stress state. As the local measurement has sufficient spatial resolution to probe a single grain inside the material, it is logical that the transformation occurs over a narrower temperature range than in a global experiment probing a large number of grains.

The substructure in the M peak could have three origins: (i) the technique is sensitive to the mechanical relaxation inside newly formed plates, such as stress relaxation by twin motion; (ii) the probed volume contains several martensite plates that grow one by one, each with its own, distinct "elementary" peak (a, b, etc.); (iii) the analyzed region of the sample surface changes due to thermal drifts, therefore probing the transformation of several growing martensite plates, leading to multiple "elementary" peaks.

21.2 Static Methods – Mesoscopic Samples

21.2.1 Carbon Nanotubes – Introduction to Basic Morphologies and Production Methods

Carbon nanotubes (CNTs) are the newest form of carbon, found in 1991 by *Iijima* [21.24]. Because of their remarkable properties, this discovery has opened whole new fields of study in physics, chemistry, and material science. They possess a unique combination of small size (diameters ranging from ∼ 1 to 50 nm with lengths up to ∼ 10 μm), low density (similar to that of graphite), high stiffness, high strength, and a broad range of electronic properties from metallic to p- and n-doped semiconducting. Their potential field of application is immense and includes reinforcing elements in high strength composites, electron sources in field emission displays and small X-ray sources, ultra-sharp and resistant AFM tips with high aspect ratios, gas sensors, and components of future, nanoscale electronics. In addition, they represent a widely used system for studying fundamental physical phenomena on the mesoscopic scale. Following advances in manufacturing and processing they are likely to be integral to many devices we use in our everyday life.

From the structural point of view, carbon nanotubes can be thought of as rolled up single sheets of graphite, *graphene*. They can be divided in two distinct groups. The first, multiwalled carbon nanotubes (MWNT), exhibit a Russian doll-like structure of nested, concentric tubes, Fig. 21.4a and b. They were also the first CNT to be discovered experimentally. The interlayer spacing can range from 0.342 to 0.375 nm, depending on the diameter and number of shells comprising the tube [21.25]. For comparison, the interlayer spacing in graphite is 0.335 nm, suggesting relatively weak interaction between individual shells. This fact has been corroborated by studies of mechanical and electronic properties of CNTs.

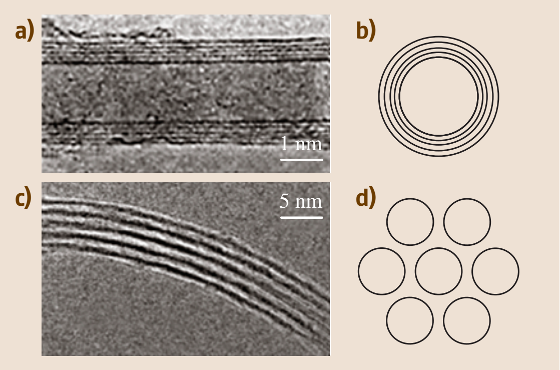

Fig. 21.4a–d TEM images and schematic drawings of cross sections for different morphologies of carbon nanotubes. **(a)** and **(b)** Multiwalled carbon nanotubes, consisting of concentric, nested tubes (MWNT). **(c)** and **(d)** ropes of single-walled carbon nanotubes, consisting of single carbon nanotubes bundled up in ropes and held together by van der Waals interaction

The second type of carbon nanotubes is in the basic form of a rolled-up graphitic sheet: a single-walled CNT (SWNT). During the production, their diameter distribution is relatively narrow so they often bundle in the form of crystalline "ropes" [21.26], Fig. 21.4c and d, in which the single tubes are held together by van der Waals interaction.

There are several distinct classes of production methods. The earliest is based on the cooling of carbon plasma. When voltage is applied between two graphitic electrodes in an inert atmosphere, they gradually evaporate and form plasma. On cooling, soot containing multiwalled nanotubes is formed [21.24]. If the anode is filled with catalysts such as, for example, cobalt of iron particles, SWNT ropes form. Another way of producing CNTs from carbon plasma is by laser ablation of a graphitic target [21.26]. It is considered that, in general, these methods produce CNTs of higher quality albeit in very small quantities and without the possibility of scale-up to industrial production. Other methods are based on chemical vapor deposition (CVD), a catalytic decomposition of various hydrocarbons, e.g. methane or acetylene mixed with nitrogen or hydrogen in the presence of catalysts [21.27]. This method offers the possibility of controlling the growth of nanotubes by patterning the catalyst [21.28] and is therefore more suitable for producing nanoscale structures with integrated CNTs. This method is also capable of producing CNTs in industrial quantities. Their main disadvantage is the higher concentration of defects, which as a consequence diminish their mechanical properties.

21.2.2 Measurements of the Mechanical Properties of Carbon Nanotubes by SPM

Mechanical measurements on CNTs performed with the AFM have confirmed theoretical expectations [21.29] of their superior mechanical properties. They involve measurements of deformations under controlled forces by bending immobilized carbon nanotubes either in the lateral [21.30] or normal direction [21.31] and also tensile stretching of CNTs fixed on their two ends to AFM tips observed in an electron microscope [21.32].

It is not obvious that continuum mechanics and its concepts like Young's, shear moduli, and tensile strength should work on the nanoscale. In order to apply them, one should also define the "thickness" of the nanotube's walls, a graphene sheet, in the frame of the continuous beam approximation. Most scientists working in this field are using the value 0.34 nm, close to the interlayer separation in graphite as the thickness of a nanotube. When comparing different results, however, one has to bear in mind that to convert relatively precise force–displacement measurements into macroscopic quantities like Young's or shear modulus, one has to introduce various geometrical factors, including diameter and length. Even a small imprecision in their determination is very unforgiving because the diameter d enters into equations for beam deformation as d^4 and length l as l^3, leading to large uncertainty in final results.

The first quantitative measurement of the Young's modulus of carbon nanotubes was reported by *Wong* et al. [21.30] in which they laterally bent MWNTs and SiC nanorods (similar in dimensions to MWNTs) deposited on flat surfaces. MWNTs were first randomly dispersed on a flat surface of MoS_2 single crystals that were used because of their low friction coefficient and exceedingly flat surface. Friction between the tubes and substrate was further reduced by performing the measurements in water. Tubes were then pinned on one side to this substrate by a deposition of an array of square pads through a shadow mask, Fig. 21.5a–c. AFM was used to locate and characterize the dimensions of protruding tubes. The beam was deformed laterally by the AFM tip, until at a certain deformation the tip would pass over the tube, allowing the tube to snap back to its relaxed position. During measurements, the force–distance curves were acquired at different positions along the chosen beam, Fig. 21.5d–e. Maximum

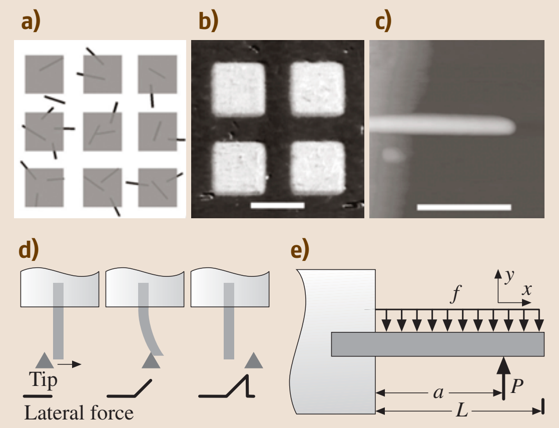

Fig. 21.5a–e Schematic of the experiment performed by *Wong* et al. [21.30] **(a)** nanotubes were dispersed on a substrate and pinned down by a deposition of SiO pads. **(b)** Optical micrograph showing the pads (*light*) and MoS_2 substrate (*dark*). The scale bar is 8 μm long. **(c)** AFM image of a SiC nanorod protruding from the pad. The scale bar is 500 nm long. The same experimental setup was used for elastically deforming MWNTs. **(d)** The tip shown as a triangle moves in the direction of the arrow. The lateral force is indicated by the trace at the bottom. Before the tip contacts the beam, the lateral force remains constant and equal to the friction force. During the bending a linear increase in the lateral force with deflection is measured. After the tip has passed over the beam the lateral force drops to its initial value and the beam snaps back to its equilibrium position. **(e)** Schematic of a pinned beam with a free end. The beam of length L is subjected to a point load P at $x = a$ and friction force f (abstracted with permission from [21.30] © American Association for the Advancement of Science)

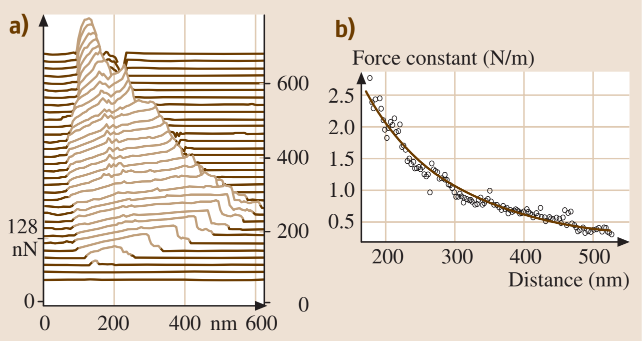

Fig. 21.6 **(a)** A series of lateral force distance curves acquired by *Wong* et al. [21.30] for different positions along a MWNT. **(b)** The lateral spring constant as a function of position on the beam. The curve is a fit to (21.3) (abstracted with permission from [21.30] © American Association for the Advancement of Science)

deflection of the nanobeam can be controlled to a certain degree by the applied normal load, and in this way tube breaking can be avoided or achieved in a controlled manner. The applied lateral load P in terms of lateral displacement y at the position x along the beam is given by the equation:

$$P(x, y) = 3EI\frac{y}{x^3} + \frac{f}{8}\left(x - 4L - 6\frac{L^2}{x}\right) , \qquad (21.2)$$

where E is the Young's modulus of the beam, I the second moment of the cross section, equal to $\pi d^4/64$ for a solid cylinder of diameter d and f the unknown friction force, presumably small due to the experimental design. The lateral force, Fig. 21.6a, is known only up to a factor of proportionality because the AFM lever's lateral force constant wasn't calibrated for these measurements. This uncertainty and the effect of friction were eliminated by calculating the nanobeam lateral force constant:

$$\frac{\mathrm{d}P}{\mathrm{d}y} \equiv k = \frac{3\pi d^4}{64x^3}E \qquad (21.3)$$

shown in Fig. 21.6b.

The mean value for the Young's modulus of MWNTs was $E = 1.3 \pm 0.6$ TPa, similar to that of diamond ($E = 1.2$ TPa). For larger deformations, discontinuities in bending curves were also observed, attributed to elastic buckling of nanotubes [21.33].

In another series of experiments, *Salvetat* et al. measured the Young's modulus of isolated SWNTs and SWNT ropes [21.31], MWNTs produced using different methods [21.34], and the shear modulus of SWNT ropes [21.35]. The experimental setup that enabled them to perform measurements on such a wide range of CNT morphologies involved measuring the vertical deflection of nanotubes bridging holes in a porous membrane.

In their measurement method, they suspended CNTs in ethanol and deposited them on the surface of a well-polished alumina Al_2O_3 ultrafiltration membrane. Tubes adhered to the surface due to the van der Waals interaction occasionally spanning holes, Fig. 21.7a. After a suitable nanotube had been found, a series of AFM images was taken under different loads, in which every image would thus correspond to the surface (and the tube) under a given normal load. Extracted linescans across the tube revealed the vertical deformation, Fig. 21.7b. For the range of applied normal loads, the deflection of a thin, long nanotube at the midpoint δ as

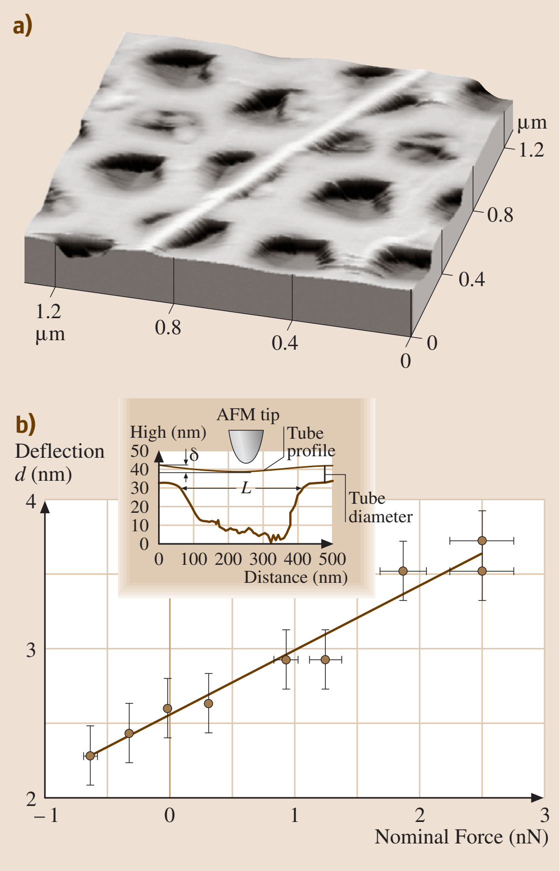

Fig. 21.7 **(a)** An AFM image of a CNT lying on a porous alumina [21.35] filter. **(b)** Measured dependence of vertical deflection on the applied nominal force. The inset shows a comparison between linescans taken on the tube and over a hole [21.34] (abstracted with permission **(a)** from [21.35] © 1999 American Physical Society, **(b)** from [21.34] © 1999 Wiley)

a function of the normal force F can be fitted using the clamped beam formula [21.36]:

$$\delta = \frac{FL^3}{192EI}\,, \qquad (21.4)$$

where L is the suspended length.

The fitted line does not pass through the origin because the force acting on the nanotube is not equal to the nominal force alone; it contains a constant term arising from the attractive force between the AFM tip and the tube. The tube's deflection should also contain a term corresponding to the interaction between the tube and the substrate, but this is generally regarded as negligible.

This variable load imaging technique is advantageous for obtaining quantitative information as one is assured that the AFM tip is in the desired location when deforming the tube. Equation (21.4) is valid only if the tubes adhere well to the substrate, confirmed by the fact that the images reveal no displacement of the parts of the tube in contact with the membrane.

Using this technique, Young's modulus of 1 TPa was found for SWNTs. Values for the MWNTs show a strong dependence on the amount of disorder in the graphitic layers: an average value of $E = 870$ GPa was found for the arc-discharge grown tubes, while the catalytically grown MWNTs, known to include a high concentration of deffects, can have a Young's modulus as low as 12 GPa.

For the deflection of SWNT ropes, an additional term in the bending formula has to be taken into account because of the influence of shearing between the tubes comprising the rope. Single CNTs are held together in the tube only by weak, van der Waals interactions. Ropes, therefore, behave as an assembly of individual tubes rather than as a thick beam. The deflection can be modeled as a sum of deflections due to bending and shearing [21.36]:

$$\delta = \delta_{\text{bending}} + \delta_{\text{shearing}} = \frac{FL^3}{192EI} + f_s \frac{FL}{4GA} = \frac{FL^3}{192E_{\text{bending}}I}\,, \qquad (21.5)$$

where f_s is a shape factor, equal to 10/9 for a cylinder, G the shear modulus and A the area of the beam's cross section. E_{bending} is the effective, bending modulus, equal to Young's modulus in the case in which the influence of shearing can be neglected (for thin, long ropes).

The Young's and the shear modulus can thus be extrapolated by measuring the E_{bending} of an ensemble of ropes with different diameter to length ratios: for thin ropes one obtains the value of the Young's modulus, while for the thick ones, the E_{bending} approaches the value of G on the order of 10 GPa (see Fig. 21.8).

Walters et al. pinned ropes of single-wall nanotubes beneath metal pads on an oxidized silicon surface, then released them by wet etching, Fig. 21.9a. The SWNT rope was deflected in the lateral direction using an AFM tip, Fig. 21.9b. As the suspended length is on the order of µm, the SWNT rope can be modeled as an elastic string stretched between the pads. Upon deformation all of the strain goes into stretching. In the simple case of a tube lying perpendicular to the trench and the AFM tip deforming the tube in the middle, the force F exerted on

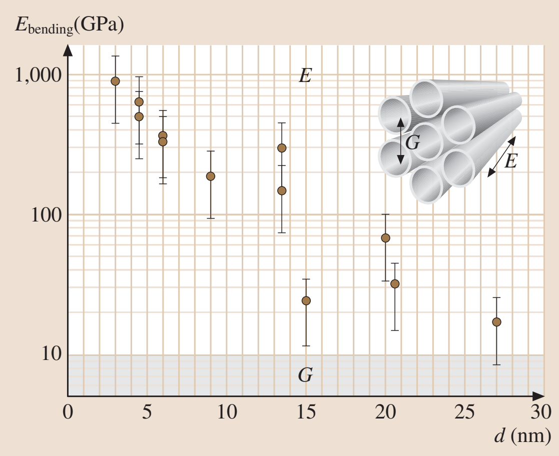

Fig. 21.8 Values of the shear modulus for 12 SWNT ropes of different diameters. The measured E_{bending} of thin ropes corresponds to E, while for thick ropes one obtains the value of shear modulus G (abstracted with permission from [21.35] © 1999 American Physical Society)

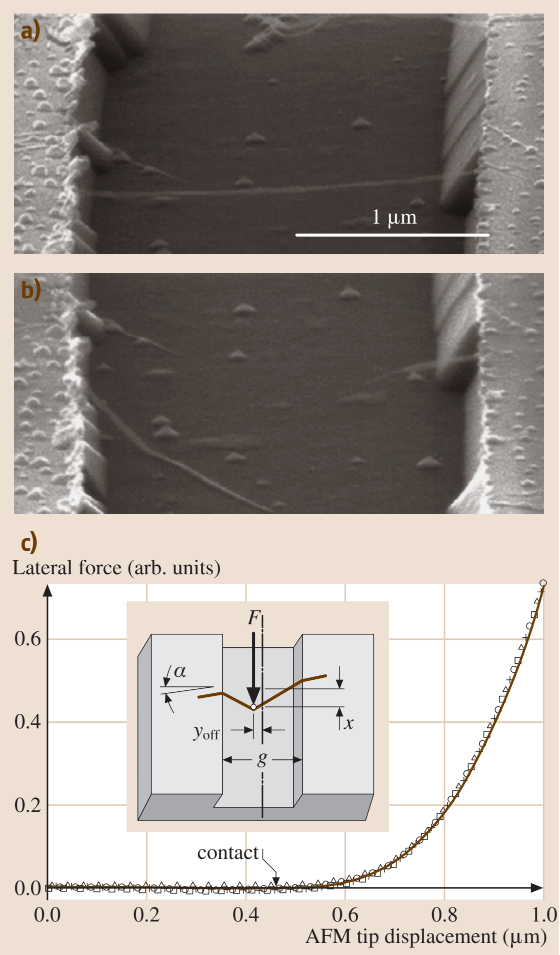

Fig. 21.9 **(a)** SEM image of a SWNT rope suspended over trench in silicon before and **(b)** after being deformed past its elastic limit. **(c)** Lateral force on a single-wall nanotube rope as a function of AFM tip displacement (abstracted with permission from [21.37] © American Institute of Physics)

the tube by the AFM tip is given by the expression:

$$F = 2T\sin\theta = 2T\frac{2x}{L} \approx \frac{8kx^3}{L_0}\,, \qquad (21.6)$$

where T is the tension in the string, L_0 its equilibrium length, k the spring constant and x the lateral deflection in the middle. Using this setup, they deformed SNWT ropes to the maximal strain of $5.8 \pm 0.9\,\%$ and determined a lower bound of 45 ± 7 GPa on the tensile strength, assuming a value of 1.2 TPa for the Young's modulus.

Kim et al. used a setup in which the SWNT rope was embedded in metallic electrodes deposited on a silicon substrate coated with poly(methylmethacrylate) (PMMA). In their experiment the tube can also be modeled as an elastic string. Using an AFM tip, they vertically deformed the rope and obtained an estimate of $E = 0.4$ TPa for the Young's modulus of an SWNT rope.

Finally, the first direct measurements of the elastic properties of CNTs that haven't relied on the beam or stretching string setup involved deforming MWNTs [21.38] and SWNT ropes [21.32] under axial strain. This was achieved by identifying and attaching opposite ends of MWNTs or SWNT ropes to two AFM tips, all inside a SEM. The AFM tips were integrated with different cantilevers, one rigid with a spring constant above 20 N/m and the other compliant with a spring constant of 0.1 N/m, Fig. 21.10. The rigid lever was then driven using a linear piezomotor. On the other end, the compliant lever bent under the applied tensile load. The deflection of the compliant cantilever – corresponding to the force applied on the nanotube – and the strain of the nanotube were simultaneously measured. The force F is calculated as $F = kd$ where k is the spring constant of the flexible AFM lever and d its displacement in the vertical direction. The strain of the nanotube is $\delta L/L$, Fig. 21.11. From the stress–strain curves obtained in this fashion, Fig. 21.11b, Young's moduli ranging from 270–950 GPa were found. Exami-

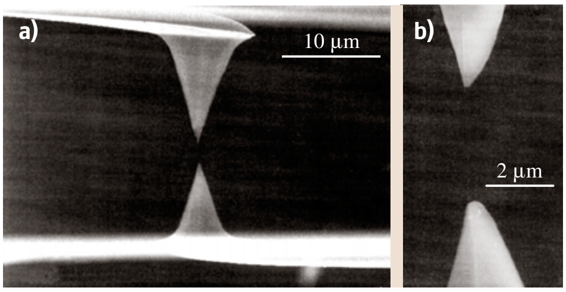

Fig. 21.10 **(a)** A SEM image of a MWNT mounted between two opposing AFM tips. **(b)** A close-up of the region indicated by a rectangle in **(a)** (abstracted with permission from [21.38] © 1999 American Association for the Advancement of Science)

nations of the same broken tubes inside a TEM revealed that nanotubes break with the "sword in sheath" mechanism, where only the outer layer appears to have carried the load. After it breaks, pullout of inner shells follows. An average bending strength of 14 GPa and axial strengths up to 63 GPa were found.

Firm attachment of nanotubes to AFM tips was ensured by a deposition of carbonaceous material induced by the electron beam concentrated in the contact area [21.40].

All these measurements of the elastic properties of carbon nanotubes are summarized in Table 21.1.

Before comparing them, it should be noted that the absolute values of these mechanical constants have relatively large uncertainties because of the huge influence of the precision of determining tube diameters and lengths on the final result. Also, the cited values represent mean values of results obtained on several tubes, with the exceptions of the lowest value given for catalytically grown MWNTs by Salvetat et al. [21.31], a single value for an individual SWNT rope from *Kim* et al. [21.39], and the lowest and highest values from *Yu* et al. [21.38]. One also has to bear in mind that concepts

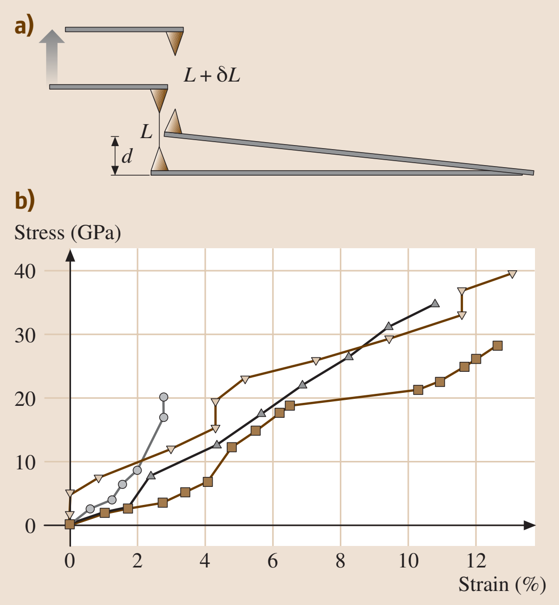

Fig. 21.11 **(a)** The principle of the experiment performed by *Yu* et al. As the rigid cantilever is driven upward, the lower, compliant cantilever bends by the amount d and the nanotube is stretched by δL. As a result, the nanotube is strained by $\delta L/L$ under the action of force $F = kd$, where k is the elastic constant of the lower AFM lever. **(b)** Plot of stress vs. strain curves for different individual MWNTs (abstracted with permission from [21.38] © 1999 American Association for the Advancement of Science)

Table 21.1 Summary of the mechanical properties of carbon nanotubes measured using SPM methods

Young's modulus *E* (GPa)	Shear modulus *G* (GPa)	Tensile strength *σ* (GPa)	Nanotube type	Deformation method	Reference
1,300±600	–	–	MWNTs arc grown	Lateral bending	[21.30]
1,000±600	–	–	SWNTs	Normal bending	[21.31]
1,000±600	∼ 1 GPa		SWNT ropes	Normal bending	[21.35]
870±400	–	–	MWNT arc-grown	Normal bending	[21.31]
12±6	–	–	MWNT catalytic	Normal bending	[21.31]
–	–	45±7	SWNT rope	Lateral bending	[21.37]
400	–	–	SWNT rope	Normal bending	[21.39]
1,020	–	30	SWNT ropes	Tensile loading	[21.32]
270–950	–	11–63	MWNTs arc-grown	Tensile loading	[21.38]

like Young's and shear moduli and tensile strength were introduced for describing macroscopic and continuous structures. Their application for describing mesoscopic objects like nanotubes, therefore, has its limitations.

The methods presented above are all in their nature "single-molecule" methods, in the sense that they measure properties of individual objects. A result of a single experiment, therefore, represents properties of a particular object and differs from case to case because of, for example, defects coming from production and purification, or for more prosaic reasons such as experimental errors. In order to perform comparisons, therefore, it is more practical to deal with values averaged for multiple tubes. The average values for the Young's modulus of high quality tubes are, within the experimental error, all on the order of 1 TPa, close to that of diamond (1.2 TPa), while the tensile strength can be 30 times higher than that of steel (1.9 GPa). Catalytically grown MWNTs with the Young's modulus that can be as low as 12±6 GPa are definitely disappointing and show that the production method plays an important role in the quality of carbon nanotubes from the point of view of their mechanical properties.

Future improvements in large scale production and processing are therefore necessary before applying them as the ultimate reinforcement fiber. Even so, the progress in measuring the mechanical properties of CNTs will continue to be closely related to, and often motivate, the progress in nanoscale manipulation in general.

21.2.3 Microtubules and Their Elastic Properties

Microtubules are a vital biological nanostructure, similar in dimension and shape to carbon nanotubes. In fact, the first name given to carbon nanotubes by their discoverer Iijima was "microtubules of graphitic carbon." From the structural point of view, they are much more complicated, as is common for biological structures. They self-assemble in buffers maintained at the physiological temperature of 37 °C out of protein subunits, α and β-tubulin, each having a molecular weight of 40 kDa. These protein subunits are bound laterally into protofilaments, which in turn are arranged in a shape of a hollow cylinder with an external diameter of 25 nm and an internal diameter of 15 nm. Microtubules are a remarkable material: inside living cells, their length incessantly fluctuates; they can even completely disassemble and consume energy in the form of GTP.

Together with actin and intermediate filaments, microtubules constitute the cellular cytoskeleton. In addition, they perform various unique vital functions: they act as tracks along which molecular motors move, they help pull apart chromosomes during cell division and form bundles that propel sperm and some bacteria. All these roles are determined by their structure and mechanical properties. Yet, after more than a decade, there is still a large discrepancy in the values of their Young's modulus reported in the literature. Several methods have been applied, such as bending or buckling microtubules using optical tweezers [21.42], hydrodynamic flow [21.43], thermally induced vibrations or shape fluctuations [21.44], buckling in vesicles [21.45], or squashing with an AFM tip [21.46]. These methods yielded results ranging from 1 MPa [21.46] to several GPa [21.42].

Since microtubules are geometrically similar to nanotubes, *Kis* et al. [21.41] used the suspended tube configuration. Microtubules were deposited on porous membranes, and AFM images were acquired under different nominal normal loads. All the measurements were performed in liquid and at controlled temperatures, in order to prevent the degradation of proteins, so the substrate had to be functionalized in order to ensure good adhesion between the tubes and the support. They used

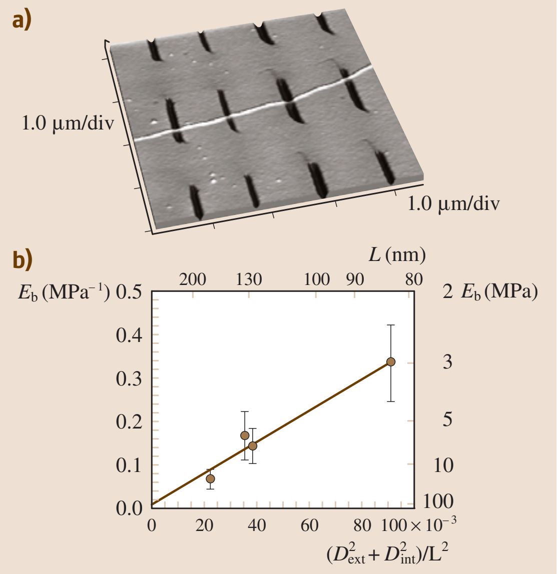

Fig. 21.12 **(a)** A pseudo 3-D rendering of a single microtubule suspended over four different-sized holes in PMMA. **(b)** From the variation of the $E_{bending}$ with varying length, the shear and the Young's modulus have been determined for the microtubule displayed in **(a)** (abtracted with permission from [21.41] © 2002 American Physical Society)

silicon with a layer of PMMA as a support. Slits were prepared in PMMA using electron-beam lithography, providing the possibility to measure the elastic response of individual microtubules lying over four different-sized holes ranging from 80–200 nm across and 400 nm deep, Fig. 21.12.

Clearly the results are dependent on the hole diameter, which could only result from a shear component within the microtubules. Simplifying (21.5), the deformation of microtubules can be modeled as:

$$\frac{1}{E_{\text{bending}}} = \frac{1}{E} + \frac{10}{3}\frac{D_{\text{ext}}^2 + D_{\text{int}}^2}{L^2}\frac{1}{G}\,, \tag{21.7}$$

where and D_{int} represent the external and internal MT's diameter. From the plot of $1/E_{\text{bending}}$ vs. the $\left(D_{\text{ext}}^2 + D_{\text{int}}^2\right)/L^2$ the shear modulus can be calculated from the slope of the fit, and the inverse Young's modulus will be equal to the intercept on the y-axis. In this way the shear modulus of $G = 1.4 \pm 0.3\,\text{GPa}$ and a lower limit of $E = 150\,\text{MPa}$ were simultaneously obtained from a measurement on an individual microtubule.

This anisotropy comes from the fact that microtubules are built of monomers that are strongly bound in the longitudinal direction, along single protofilaments. The link between neighboring protofilaments is much weaker. Microtubules, therefore, have to be considered as anisotropic beams, having an E_{bending} that depends on the scale on which they are deformed [21.41]. This is in all respects analogous to the situation in SWNT ropes that are built of stiff individual SWNTs, held together in bundles by a weak van der Waals interaction. This conclusion may explain the large discrepancy in the value of Young's moduli reported in the literature over the past ten years using inadequate modeling.

As in the case of carbon nanotubes, measuring the mechanical properties of microtubules can provide valuable insight into their structure and provide deeper understanding about the functioning of these remarkable structures.

21.3 Scanning Nanoindentation: An Application to Bone Tissue

In the following section we discuss nanoindentation as a tool to determine nanomechanical properties. This method was developed in the early 1980s evolving from traditional Vicker hardness testing devices. The latter is based on the concept that a pyramidal tip is loaded into a material applying a known force. After the test, the size of the remaining imprint is measured under an optical microscope. The ratio of the employed force and the imprint area after load removal was defined as hardness that represents a mean pressure the material can resist. Unfortunately this mechanical parameter is a complex combination of elastic and postyield properties and can't be easily explained on the level of continuum mechanics. This point raised important concerns for determining elastic constants such as the Young modulus from an indentation test. Further improvements on the transducer sensitivity were necessary to provide continuous acquisition of the employed load and the resulting indentation depth. Nanoindentation represents the state of the art of this development allowing mechanical tests on the nanometer scale.

21.3.1 Scanning Nanoindentation

Based on this continuous force–displacement data, an elastic modulus of volumes in the submicron regime can be quantified. To investigate features down to the micrometer regime, classical nanoindentation tools are combined with an optical microscope for positioning the indenter tip on the region of interest. But to defeat the limits of optical resolution, indentation and scanning probe microscopy were combined using the same tip to allow nanoscale control of the indenting tip position [21.47]. This instrument allows for scanning the material topography in an AFM mode and performing nanoindentation tests employing the same tip. The mechanical tests are restricted to the scanned area (100 µm × 100 µm maximum) where the indenter can be positioned with a precision of better than 100 nm. The surface roughness can be measured, which is helpful for choosing the area to be indented. Furthermore the available in situ scan of the indented region can provide information about the piling-up or sink-in behavior of the material. *Kulkarni* and *Bhushan* [21.48,49] introduced this device and demonstrated that measurements on aluminum and silicon were similar to results of non-scanning indentation systems.

21.3.2 Application of Scanning Nanoindentation

The development of the scanning nanoindentation (SN) technique has led to a variety of studies, primarily characterizing thin layers. Applications range from the mechanical characterization of corrosion-free film apposition on single-crystalline iron [21.50] to indentation and microscratch tests on Fe-N/Ti-N multilayers [21.51]. One author included nanoindentation to discuss two different electrochemical deposition techniques of thin Ni-P layers on pure Ni [21.52]. *Rar* et al. [21.53] reported studies on the growth of thin gold layer on native oxides of silicium while other investigations focussed on wear-resistant TiB$_2$(N) coatings [21.54].

Studies of heterogeneous materials clearly demonstrate the advantage of the available surface scanning. *Malkow* and *Bull* [21.55] investigated the elastic/plastic behaviour of carbon-nitride films deposited on silicon. Therefore they determined the load-on and load-off hardness, the latter being accessible by scanning the remaining indent impression. *Shima* et al. [21.56] studied silicon oxynitride films on pure silicon and demonstrated hardness as a function of the employed deposition temperature. *Göken* et al. [21.57] used the high positioning precision to characterize individual lamellae of a TiAl alloy that consists of a two-phase structure and also to study the mechanical properties of nanometer-size precipitates of nickel-based super-alloys [21.58]. Performing in situ electrochemical treatments of iron single crystals, *Seo* et al. [21.59] used SN to study the variation of the remaining imprint shape with time.

In the field of biomechanics, work was done on bone tissue in wild-type and gene-mutated zebrafish. The characterization of the residual indentations in AFM mode supported the statement that gene mutation can change bone brittleness [21.60]. Other investigations focussed on the demineralising effects of soft drinks on tooth enamel by studying changes of elastic properties and surface topography [21.61]. *Habelitz* et al. [21.62] and *Marshall* et al. [21.63] characterized the junction between human tooth enamel and the mechanically different dentin.

Hengsberger et al. [21.64] took advantage of the nanometer positioning capability of the SN to investigate the elastic and plastic properties of individual human bone lamellae. Related work is presented in greater detail below to demonstrate the use and benefits of the SN technique at this example.

21.3.3 Example: Study of Mechanical Properties of Bone Lamellae Using SN

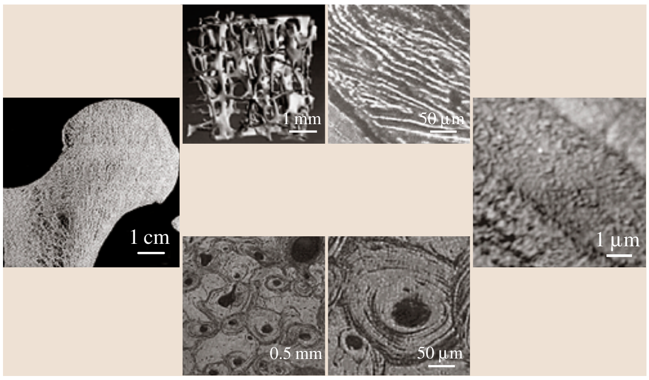

Fig. 21.13 Hierarchy of the human femoral neck (from left to the right) the far left figure shows a cut through a frontal plane of the femoral neck. The outer shell is constituted of compact bone while the inside is made of the spongier trabecular bone. The left pair of images shows the trabecular structure (*top*) and a transverse cut of the compact shell that shows vascular canals (*bottom*).
The right pair of images shows packets of trabecular bone lamellae (*top*) and a single osteon (*bottom*), a vascular canal surrounded by concentric lamellae. Packets of trabecular bone lamellae and osteons represent a structural unit (BSU). Note the alternation of bright (*thick*) and dark (*thin*) lamellae.
The figure far right shows three bone lamellae, structures that are similar for trabecular and compact bone. The bone matrix within the lamellae are mainly constituted of collagen fibers and hydroxyapatite crystals

Before we discuss nanomechanical properties of bone lamellae, it is useful to know the structure on the macroscopic level. Figure 21.13 presents the hierarchy of human bone tissue at the example of the femoral neck. The outer shell is constituted of cortical bone while the porous trabecular bone structure gives inner support. On the next lower level, both bone types show structural units (BSU, after [21.65]) constituted of some tens of lamellae. For compact bone the lamellae have a concentric organization while for trabecular bone these lamellae are parallel. The nomenclature "BSU" was motivated by its underlying cellular process. The formation of such an individual structural unit occurs within a single cellular process. The optical contrast shows two alternating types of lamellae, thin lamellae that appear dark, and thick lamellae that appear bright. These bone lamellae show widths ranging from 1 to 3 µm for thin and 2 to 4 µm for thick lamellae [21.64]. On the next lower level the bone matrix is constituted of a complex collagen and mineral structure. It is still not entirely understood what type of variations in the collagen/mineral meshwork are responsible for the lamellar structure.

The mechanical properties on these different levels of bone hierarchy are increasingly well understood. But, little is known about how the macroscopic mechanical properties of the whole bone relate to its nanomechanical properties. The possibility that factors such as fracture risk might be better understood with an analysis of small bone volumes motivates the application of nanoindentation in this field.

A set of nanoindentation studies is available (mainly employing the device combined with an optical microscope) that presents the intrinsic mechanical properties of bone tissue. Among recent work, differences were reported between donors, anatomical sites, BSU, and thin and thick lamellae [21.64, 66–69]. Due to technical constraints of this sensitive nanomechanical device, the majority of these studies present dehydrated or dried tissue properties measured at ambient temperature [21.70–73]. But, removal of the water content may lead to anisotropic shrinking of the matrix that creates microcracks and alters the mechanical properties of the bone constituents. For an accurate characterization of the in vivo properties, the nanomechanical tests should therefore be done under physiological conditions.

There have been a few attempts to characterize the in vivo bone properties [21.66, 68, 74]. This was achieved by studying fresh bone samples that are kept moist with a thin layer of liquid (less than a hundred microns) on the surface or with subsurface water irrigation. But, local evaporation of the thin liquid layer may have led to indents on areas that were partially dried during the test. One possible solution to such local drying might be to conduct measurements in which the indentation tip and tissue sample are both fully immersed in a liquid cell and simultaneously heated to body temperature.

Practical limitations on temperature stability and inaccuracy in contact force detection due to liquid on the surface make such measurements extraordinarily difficult, however. For statistically powerful studies it seems unavoidable to dry the samples.

The objective of this study was to use SN to determine the effect of drying on the stiffness of single bone lamellae. The goal was to determine a conversion factor that allows dry tissue properties to be recalculated into their in vivo properties.

For this purpose, we measured the identical set of lamellae selected from human trabecular and compact bone at first under physiological and then under dry conditions.

Experimental Setup and Technical Features:

Figure 21.14 demonstrates an optical picture and a sketch of the combined AFM and nanoindenter device (Hysitron Inc. Minneapolis, MN). The transducer consists of a three-plate capacitor on whose central plate a Berkovich (three-sided pyramid) diamond tip is mounted. The transducer provides a contact force feedback between the tip and the sample surface. The sample is mounted on a piezoelectric scanner that allows moving in the x,y and z-directions. During x,y-surface scanning, the piezoelectric scanner keeps this feedback signal constant by correcting the z-height. The movement of the piezoelectric scanner, therefore, describes the negative surface of constant contact force, commonly called an AFM-scan. For the liquid cell tests, the sample was glued in a plexiglass cup for the addition of a several millimeter high liquid layer. A commercially available liquid cell tip (Hysitron Inc. Minneapolis, MN) was used, which contains an additional shaft of approximately 300 µm diameter and 5 mm length to protect the transducer from the fluid. A small layer of latex was placed between the sample holder and the magnetic stick of the piezoelectric scanner to protect the latter from liquid. The additional shaft of the indenter tip and the latex layer

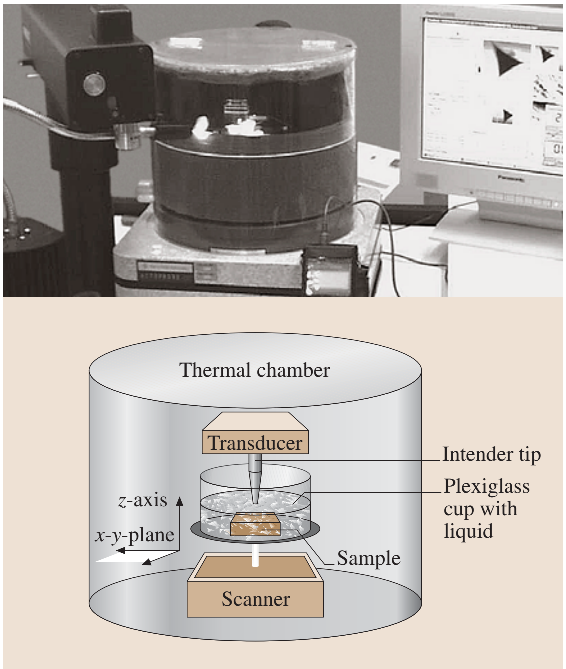

Fig. 21.14 Optical picture (*top*) and sketch (*bottom*) of the scanning nanoindentation device for measurements under physiological conditions. For this purpose the sample was installed in a plexiglass cup for addition of liquid. The entire instrument was heated in a custom-made thermal chamber

represent elastic components that increased the machine compliance ($C_m = 7\,\mathrm{nm/mN}$ instead of $3\,\mathrm{nm/mN}$). This variable corrects the deformation of all the machine components while indentation data are recorded and can be determined from the tip shape calibration curve (Hysitron, Minneapolis). The nanoindentation device was installed in a custom-made thermal chamber to allow sample heating to 37 °C. Note that the increased temperature and humidity also changed the value of the electrostatic force constant (EFC). The latter corrects the force due to the springs that support the central plate on which the indenter tip is mounted. The EFC can be calibrated doing out-of-contact indents and varying this value until a zero-line in the force–displacement curve is measured.

Based on an image acquired in AFM mode, the indentation area can be selected with a high spatial resolution ($< 0.1\,\mu\mathrm{m}$). Figure 21.15 shows the force–displacement data of a typical nanoindentation curve. In the first step, the tip is pressed into the material, resulting in indistinguishable elastic and plastic deformation. Then the tip is held at maximum force resulting in creep of the material under the tip. In a third step, unloading is done that leads to elastic recovery of the material. Typically each indent requires between 15 seconds to several minutes, representing a compromise between a desired quasistatic strain rate and the thermal drift of the instrument (possibly below 0.1 nm/second). This device is load-controlled, and linearly increasing and decreasing loading protocols were therefore applied. The loading rate corresponds to

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{P_{\max}}{T} \tag{21.8}$$

with $P_{\max}$ as the maximum load and T as the (un)loading time.

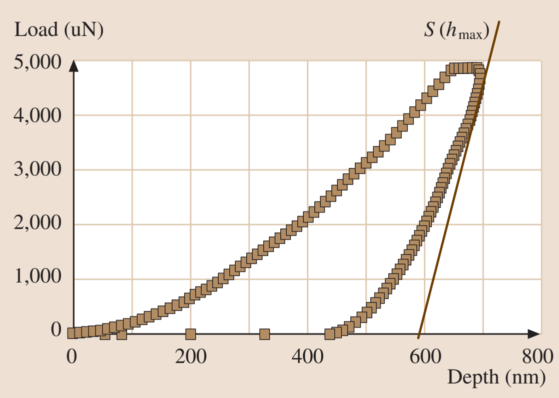

Fig. 21.15 Force–displacement curve obtained during a nanoindentation test that consists of three parts (see text). Hardness is determined at maximum load. The elastic indentation modulus is calculated with the slope at the point of initial unloading $S(h_{\max})$

The elastic constants of the sample were determined using the unloading part (step 3) of the nanoindentation curve. Based on the analytical work by *Sneddon* [21.75], *Oliver* and *Pharr* [21.76] derived the following equation for an indenter of revolution pressed into an isotropic elastic material:

$$S\left(h_{\max}\right) = \frac{\mathrm{d}P}{\mathrm{d}h}\left(h_{\max}\right) = \frac{2}{\sqrt{\pi}} E_{\mathrm{r}} \sqrt{A_{\mathrm{c}}\left(h_{\max}\right)}\,. \tag{21.9}$$

Pharr et al. [21.77] showed that this equation is a good approximation for a Berkovich indenter tip. P represents the applied load, $S\left(h_{\max}\right)$ is the derivative of the unloading curve at the point of initial unloading $h_{\max}$. This variable is determined by fitting typically 40% to 95% of the unloading curve to avoid the influence of viscoelastic effects at initial unloading when a singularity in the strain rate occurs. $A_{\mathrm{c}}(h)$ is the contact area over which the material and the indenter are in instantaneous contact. The latter function has to be calibrated by performing indents with increasing depth in a standard material, typically fused silica. The latter has a known reduced modulus of $E_{\mathrm{r}} = 69.9\,\mathrm{GPa}$ that allows calibration of the contact area $A_{\mathrm{c}}(h)$ using the measured contact stiffness $S(h)$ and (21.9). The reduced modulus E_{r} combines the deformation of the material and the diamond tip as follows [21.78]:

$$\frac{1}{E_{\mathrm{r}}} = \frac{1-\nu_{\mathrm{specimen}}^2}{E_{\mathrm{specimen}}} + \frac{1-\nu_{\mathrm{tip}}^2}{E_{\mathrm{tip}}}\,. \tag{21.10}$$

We use the indentation modulus

$$E_{\mathrm{ind}} = \left(\frac{1}{E_{\mathrm{r}}} - \frac{1-\nu_{\mathrm{tip}}^2}{E_{\mathrm{tip}}}\right)^{-1} \tag{21.11}$$

that can be calculated with the reduced modulus and the elastic properties of the diamond indenter tip $\nu_{\mathrm{tip}} = 0.07$ and $E_{\mathrm{tip}} = 1{,}140\,\mathrm{GPa}$. This variable combines with

$$E_{\mathrm{ind}} = \frac{E_{\mathrm{specimen}}}{1-\nu_{\mathrm{specimen}}^2} \tag{21.12}$$

the local Young's modulus E_{specimen} and Poisson ratio ν_{specimen} of the specimen.

It is important to note that this theory assumes an isotropic material. For an anisotropic material the indentation modulus represents some average of the elastic properties in all directions. The latter strongly depends on the geometry of the indenter. For indents with a blunt indenter (such as Berkovich with an opening angle of 143°) nonnegligible deformations will occur in the plane perpendicular to the loading direction. In this context it is useful to present an approximation of the volume ganged by the indentation test.

The stress field generated by the indentation process is heterogeneous and leads to plastic deformation and damage in the vicinity of the tip. Using Hertz's theory [21.78], the spatial dependence of the stress components during indentation can be estimated by considering the elastic contact of a spherical indenter with a semi-infinite half space. In the direction of loading (z-axis), the stress component below the indenter decreases according to

$$\frac{\sigma_{zz}}{p_0} = -\left(1 + \frac{z^2}{a^2}\right) , \tag{21.13}$$

where p_0 indicates the maximum pressure below the indenter and a the contact radius.

In a horizontal plane at the surface ($z = 0$), the radial and circumferential components of the stress field next to the contact area obey

$$\frac{\sigma_{rr}}{p_0} = -\frac{\sigma_{\theta\theta}}{p_0} = \frac{(1-2\nu)\,a^2}{3r^2} , \tag{21.14}$$

where θ and r are the cylindrical coordinates of the periphery of the indenter and ν is the Poisson ratio. For $\nu = 0.3$ the stress field components reach their 10%-boundary defined by $\sigma_{zz} = -0.1 p_0$ and $\sigma_{rr} = -\sigma_{\theta\theta} = 0.1 p_0$ at a depth of $z \cong 3a$ or in lateral direction at $r \cong 2/\sqrt{3}a$. For a Berkovich tip, the ratio between maximum indentation depth and contact radius is approximately $a \cong 3h_{\max}$. The mechanical properties measured by nanoindentation, therefore, correspond to a semi-ellipsoidal volume extending to about nine times the employed indentation depth ($z \cong 3a \cong 9h_{\max}$) in the vertical direction (z) and about seven times this same depth ($2r \cong 4/\sqrt{3}a \cong 7h_{\max}$) in the radial direction (r).

The load-on hardness is determined with

$$H = \frac{P\,(h_{\max})}{A_{\mathrm{c}}\,(h_{\max})} \tag{21.15}$$

as the ratio of the maximum load and the ("on load") contact area at maximum load. This is different from Vicker hardness in which the contact area is characterized by the remaining imprint after load removal. Differences between nanoindentation and Vicker, therefore, occur for materials with nonnegligible elastic recovery. After unloading, such materials may expose an imprint smaller than the area of contact at maximum load. Since the SN tool can be used to image the remaining imprint after unloading, hardness values using both definitions may thus be compared.

For fused silica the literature provides a hardness range between 8.3 to 9.5 GPa while an intercomparison of SN-users showed an average of 8.96 GPa (Surface, Hückelhoven). Conveniently, the hardness value also represents a possibility to check the area function of the tip based on the reduced modulus of the calibration material.

Tests of Bone Tissue Under Pseudo-physiological Conditions

Samples containing trabecular and compact bone lamellae were dissected from the medial part of the femoral neck of an 86-year-old female. After embedding in PMMA, the samples were polished with successive grades of carbide paper and finished with 0.05 μm alumina solution. Polishing represents an important preparation step since the here-employed theory assumes an infinitely flat surface. The mean surface roughness of the indented area should, therefore, be far below the employed indentation depth. Unfortunately no objective criteria that determines the maximum allowable surface roughness as a function of the indentation depth has been formulated so far.

Thin and thick lamellae of trabecular and compact bone were first characterized under physiological conditions, i. e. fully wet and at 37 °C. Then the specimens were dried for 24 h at 50 °C and identical tests were carried out again but under dry conditions.

In both cases 16 indentation tests were performed to maximum depths of 100 nm and 500 nm. Each test consisted of 10 s loading, 10 s holding, and 10 s unloading. The maximum allowable thermal drift was set 0.1 nm/s. To avoid proximity effects of neighbouring indents, an adjacent testing area in the identical lamellae was chosen after changing the testing conditions.

The tests under fully wet conditions at body temperature were found to be very sensitive with respect to thermal stability. The nanoindentation device was heated for several days to reach stable thermal conditions of the instrument components before the sample could be installed. The electrostatic force constant (EFC) was checked daily before beginning data acquisition. The approach of the indenter tip in the liquid environment was performed employing a contact force of 7 μN. This

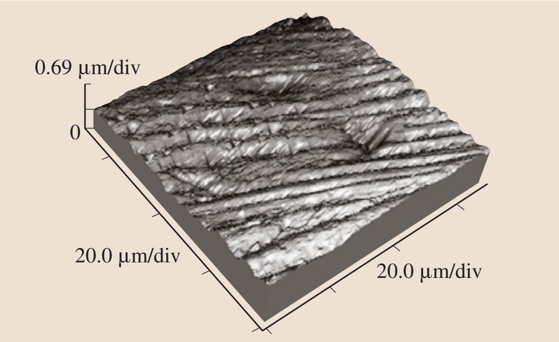

Fig. 21.16 Surface topography of a trabecular bone structural unit that shows the lamellar structure. Thick (*bright*) lamellae correspond to the tops in the topography. Note, the two holes are not remaining indents but are ellipsoidal lacunae that embed bone cells

value is offset by approximately 1 µN per mm of water penetration, however, due to Archimedes force acting on the fluid cell tip. Such additional effects as water surface tension may also have a repulsive force on the tip.

AFM-scans as in Fig. 21.16, which shows the topography of trabecular bone lamellae under dry conditions, allowed the two lamella types to be identified. Thick lamellae correspond to the tops (bright contrast) and thin lamellae to the valleys (dark contrast) where the surface relief results from preferential polishing.

Indentation Modulus

Trabecular bone showed under wet conditions a mean indentation modulus of 12.5 ± 4.0 GPa when the data of both lamella types and indentation depths were combined. Under dry conditions the results showed a mean of 19.6 ± 2.6 GPa, an increase of 57%. Lamellae of compact bone increased their mean stiffness by 76.5% from 14.9 ± 4.5 GPa to 26.4 ± 3.8 GPa.

These determined conversion factors between dried and fully wet tissue properties should be compared with other studies. *Rho* [21.74] has reported an increase of indentation modulus of bovine compact bone by 15.8%. Our study demonstrates a change of +76.5% for human compact bone. This high discrepancy may be attributed to different preparation and testing protocols. Rho tested the bone samples at ambient temperature while kept moist by a thin film of deionized water. In our study the tissue properties were determined under fully wet conditions and at body temperature. Furthermore, Rho dried for 14 days at ambient temperature while in our study the drying process occurred during 24 h at 50 °C. These points may explain the different relative change of mechanical properties we detected.

Figure 21.17 presents the indentation moduli (combining both depths) normalized with respect to their initial wet values under physiological conditions. It is interesting that the increase of this elastic parameter after drying was significantly ($p < 0.00001$) higher for thin than for thick lamellae. The relative change of stiffness due to drying was +44% for thick lamellae and +109% for thin lamellae of compact bone. For trabecular bone the corresponding values were +37% and +78% for thick and thin lamellae, respectively.

Table 21.2 shows the results for both indentation depths. The differences are likely related to the volume sampled during the indentation, a semi-ellipsoid following the approximations of (21.13) and (21.14) with 0.7 µm diameter and 0.9 µm height for 100 nm indents (and 3.5 µm × 4.5 µm for 500 nm indents). Given the typical lamellae dimensions of 1–4 µm, the shallow indentation measurements, therefore, represent properties of single lamellae whereas the deeper indents include neighboring lamellae. It is also worth noting that thin lamellae showed a greater effect of drying when only the shallow indents were considered.

This important result should be discussed in the frame of published models that address the phenomenon of bone lamellation. *Marotti* [21.79] proposed that bone lamellae are the result of alternating changes of the collagen fiber density. According to his SEM-results the density of collagen is higher in thin lamellae than in thick lamellae. Collagen fibers are long chains of proteins and contain adsorption sites for polar water

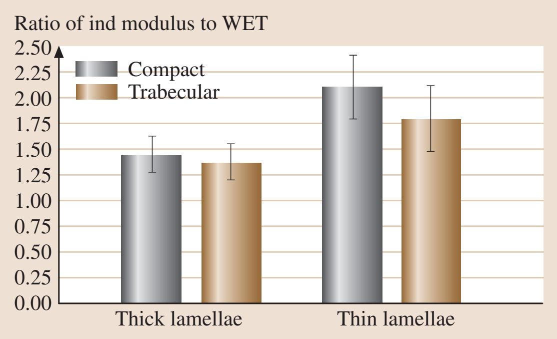

Fig. 21.17 Relative increase of indentation modulus for compact and trabecular bone lamellae after drying. The results are normalized with respect to their initial value under wet conditions. Note that thin lamellae are more affected by drying for both types of bone

Table 21.2 Absolute and relative changes of the indentation modulus with respect to the initial values under wet conditions. Note the greater increase for the thin lamellae after drying. Indents to 100 nm depth represent properties of single lamellae while 500 nm indents are also influenced by neighboring lamellae

Indentation modulus	Lamella	Ind. depth (nm)	Wet (GPa)	Dry (GPa)	Rel. change (%)
Trabecular bone	Thin	100	11.6 ± 4.1	20.5 ± 4.3	+76
		500	9.8 ± 1.9	18.1 ± 3.2	+84
	Thick	100	12.4±5	19.1 ± 1.11	+54
		500	15.6 ± 1.3	19.0 ± 1.1	+22
Compact bone	Thin	100	10.4 ± 0.2	23.4 ± 3.6	+124
		500	13.9 ± 2.0	27.1 ± 3.1	+95
	Thick	100	19.9 ± 4.8	27.9 ± 4.6	+40
		500	15.5 ± 2.6	23.0 ± 0.8	+49

molecules. A higher density of the collagen fibers in the thinner lamellae results in a higher water binding capacity that may explain the higher relative influence of drying.

Giraud-Guille [21.80], on the other hand, proposed a nested-arc structure for bone lamellae with smooth orientation changes between adjacent collagen fibers. According to this model indents on thick lamellae load into the longitudinal direction of the fibers, whereas the load is perpendicular to the long axis for thin lamellae. Removal of the liquid phase leads to packaging of the collagen fibers, an effect that is intuitively anisotropic and that may explain why the effect of drying was different for both lamella types. In addition, drying leads to microfracture initiation within thick lamellae [21.64], possibly explaining the diminished effect of drying on these structure.

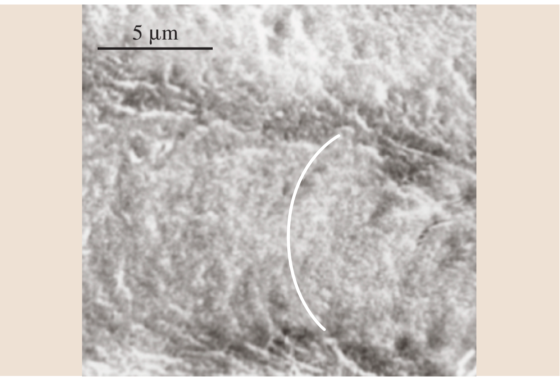

Fig. 21.18 SEM-scan of bone lamellae adjacent to the spot characterized by nanoindentation. This SEM-scan gives support for the bone lamellation model that is based on a smooth orientation change of collagen fibers [21.80]

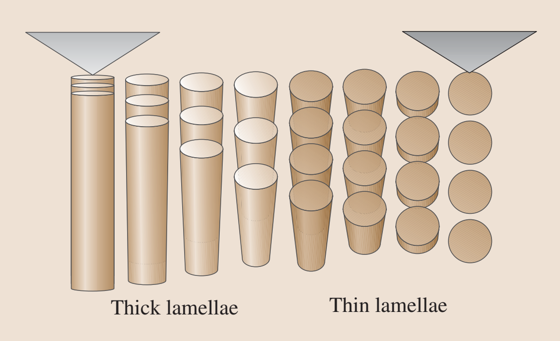

Fig. 21.19 Bone lamellation model proposed by this study. This model represents a combination of a smooth orientation change and density variations of the collagen fibers. Note also the spots where the indentations were done. For thin lamellae a greater change of mechanical properties was detected after drying

Figure 21.18 presents a SEM-image from a bone sample used in this study. This scan confirms the nested-arc structure Giraud-Guille observed with TEM.

Our results are, therefore, in agreement with both the Marotti and the Giraud-Guille models. This implies a model that combines smooth orientation changes of the collagen fibers with density changes. Such an architecture is sketched in Fig. 21.19. Note also the spots where the indentations were done.

Hardness

Compact bone revealed a mean hardness of 0.46 GPa under wet conditions that increased by 74% to 0.80 GPa after drying. Due to drying trabecular bone lamellae showed a mean hardness change of 76% from 0.41 up to 0.72 GPa.

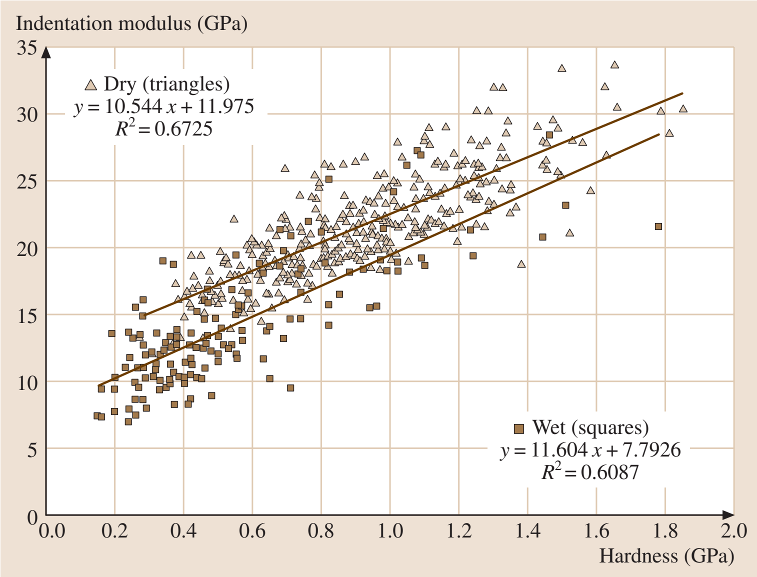

Fig. 21.20 Indentation modulus and hardness under dry and wet conditions using raw data from *Hengsberger* et al. [21.64]. Drying does not change the correlation coefficient and changes slightly but significantly the slope of the regression curve

For compact bone the effect of drying was again significantly ($p = 0.0006$) greater for thin lamellae (+108%) than for thick lamellae (+44%). Similar results were found for trabecular bone lamellae with a +99% increase for thin and a +56% increase for thick lamellae.

The results for hardness, therefore, appear to similar trends as the indentation modulus. This motivates us to discuss a further point, the correlation between indentation modulus and hardness. These mechanical parameters show an empirically proven relationship [21.81] that may change as a result of drying. For this purpose we related the data of the indentation modulus of Hengsberger et al. [21.64] with the corresponding unpublished hardness results. The correlation between hardness and indentation modulus is similar under wet ($R^2 = 0.61$) and under dry ($R^2 = 0.67$) conditions (Fig. 21.20). The slopes of the linear regression, however, show close but significantly different values ($p < 0.0001$) for the wet and dry samples. Hardness and indentation modulus show a similar but significantly different relative shift as a result of drying.

21.3.4 Conclusion

Modern scanning nanoindentation has clearly solved two problems: indentations in the nanometer scale can be performed with high force and displacement resolution, and nanometer lateral control of the indentation position allows the characterization of small structures of heterogeneous materials. Interfaces in composites, local density or composition variations of chemical products, or biological structures can then be investigated using the SN tool.

Based on recent theoretical progress [21.82], the nanoindentation technique can also be used to characterize anisotropic materials by indenting in different directions. Other parameters can also be determined. For instance, creep behavior is accessible when the indentation load is held constant at a maximum load. Hysteresis of the force–displacement curve represents the energy dissipated during an indentation providing information about post-elastic behavior of a sample. Such dynamical variables as viscoelastic properties are also accessible by including sinusoidal oscillations of different frequencies to the loading history.

Future interests of the nanotechnology community may direct the work towards tests on lower levels of structural organization like the molecular or atomic level. This would require strong improvements of the transducer sensitivity and of the indenter tip machining. New concepts to reduce the thermal drifting properties would also be necessary for the next generation of this device.

21.4 Conclusions and Perspectives

We have demonstrated that AFM is an ideal tool for investigating variations in local properties of bulk materials, like bone, and also for performing physical measurements on individual nanometer scale objects such as, for example, carbon nanotubes and protein polymers. In the latter case, the technique provides previously inaccessible quantities in living matter such as, for instance, the shear modulus. It should be emphasized that shear is omnipresent since biostructures are "composite" materials, with strong anisotropic interactions between their constituents. In the case of proteins, we have explored the mechanical properties as a function of temperature, as well. Interactions can vary remarkably even over a range of a few tens of degrees, providing deeper insight in the functioning of these structures. One task, which seems very difficult for the time being, is the frequency dependent response of nano and biostructures. In the latter case it is limited to a few kilohertz, mainly because of the surrounding liquid, but for a meaningful analysis one requires several decades in frequency.

Regarding the scanning probe tips themselves, improvements are necessary to provide better resolution, longer lifetimes, and easy functionalization for sensing different chemical environments. This might be achieved by carbon nanotube tips (Fig. 21.21), which have a Young's modulus of 1 TPa, a very well-defined and sharp tip structure, and pentagon "defects" at the apex providing a site for functionalization. It is generally agreed upon by the scanning probe community that carbon nanotubes will open new avenues for the study of living matter.

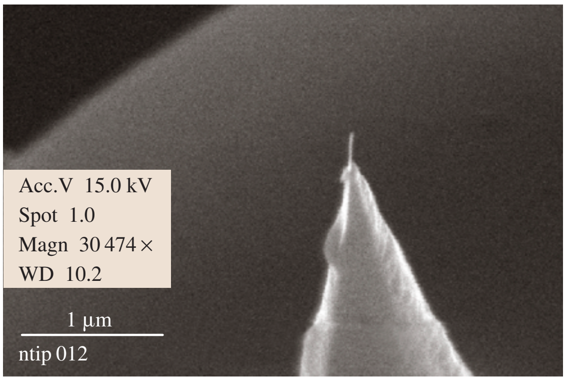

Fig. 21.21 SEM picture of an AFM tip with attached multiwall Carbon nanotube

Finally the development of the photonic force microscope in Heidelberg has allowed imaging of bio-structures with unprecedented three-dimensional resolution, even including features that are otherwise inaccessible to AFM tips. In the future, this instrument will certainly provide radical new insight in biological functioning.

References

- 21.1 M. Radmacher: Measuring the elastic properties of biological samples with the AFM, IEEE Eng. Med. Biol. Mag. **16**(2) (1997) 47–57
- 21.2 G. Binnig, C. F. Quate, C. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933
- 21.3 P. Maivald, H. J. Butt, S. A. C. Gould, C. B. Prater, B. Drake, J. A. Gurley, V. B. Elings, P. K. Hansma: Using force modulation to image surface elasticities with the atomic force microscope, Nanotechnology **2** (1991) 103–106
- 21.4 T. Kajiyama, K. Takata, I. Ohki, S.-R. Ge, J.-S. Yoon, A. Takahara: Imaging of dynamic viscoelstic properties of a phase-separated polymer by forced oscillation atomic force microscopy, Macromolecules **27** (1994) 7932–7934
- 21.5 B. Nysten, R. Legras, J.-L. Costa: Atomic force microscopy imaging of viscoelastic properties in toughened polypropylene resins, J. Appl. Phys. **78** (1995) 5953–5958
- 21.6 P.-E. Mazeran, J.-L. Loubet: Normal and lateral modulation with a scanning force microscope, an analysis: Implication in quantitative elastic and friction imaging, Tribol. Lett. **3** (1997) 125–129
- 21.7 B. Cretin, F. Sthal: Scanning microdeformation microscopy, Appl. Phys. Lett. **62** (1993) 829–831
- 21.8 N. A. Burnham, A. J. Kulik, G. Gremaud, P.-J. Gallo, F. Oulevey: Scanning local-acceleration microscopy, J. Vac. Sci. Technol. B **14** (1996) 794–799
- 21.9 F. Oulevey, N. A. Burnham, A. J. Kulik, G. Gremaud, W. Benoit: Mechanical properties studied at the nanoscale using scanning local-acceleration microscopy (SLAM), J. Phys. **IV** (1996) C8-731–734
- 21.10 U. Rabe, W. Arnold: Acoustic microscopy by atomic force microscopy, Appl.Phys. Lett **64** (1994) 1493–1495
- 21.11 O. Kolosov, K. Yamanaka: Nonlinear detection of ultrasonic vibrations in an atomic force microscope, Jpn. J. Appl. Phys. **32** (1993) 22–25

21.12 F. Oulevey: Cartographie et spectrométrie des propriétés mécaniques à l'échelle nanométrique par microscopie acoustique en champ proche. Ph.D. Thesis (EPFL, Lausanne 1999)

21.13 F. Oulevey, G. Gremaud, A. Semoroz, A.J. Kulik, N.A. Burnham, E. Dupas, D. Gourdon: Local mechanical spectroscopy with nanometer-scale lateral resolution, Rev. Sci. Instrum. **69** (1998) 2085–2094

21.14 N.A. Burnham, G. Gremaud, A.J. Kulik, P.-J. Gallo, F. Oulevey: Materials' properties measurements: Choosing the optimal scanning probe microscope configuration, J. Vac. Sci. Technol. B **14** (1996) 1308–1312

21.15 F. Oulevey, G. Gremaud, A.J. Kulik, B. Guisolan: Simple low-drift heating stage for scanning probe microscopes, Rev. Sci. Instrum. **70** (1999) 1889–1890

21.16 H.M. Pollock: Nanoindentation. In: *Friction lubrication and wear technology*, AMS Handbook, Vol. 18 (AMS, Ohio 1992) p. 419

21.17 E. Dupas, G. Gremaud, A. Kulik, J.-L. Loubet: High-frequency mechanical spectroscopy with an atomic force microscope, Rev. Sci. Instrum. **72**(10) (2001) 3891–3897

21.18 F. Oulevey, N.A. Burnham, G. Gremaud, A.J. Kulik, H.M. Pollock, A. Hammiche, M. Reading, M. Song, D.J. Hourston: Dynamic mechanical analysis at the submicron scale, Polymer **41** (2000) 3087–3092

21.19 W.Wm. Wendlandt: *Thermal Analysis in Chemical Analysis*, Vol. 19, 3rd edn. (Wiley, New York 1996) p. 360

21.20 D. Mari, D.C. Dunand: NiTi and NiTi-TiC composites: Part I. Transformation and thermal cycling behavior, Metall. Mater. Trans. A **26A** (1995) 2833–2847

21.21 L. Bataillard: Transformation martensitique multiple dans un alliage à mémoire de forme Ni-Ti. Ph.D. Thesis (EPF, Lausanne 1996)

21.22 F. Oulevey, G. Gremaud, D. Mari, A.J. Kulik, N.A. Burnham, W. Benoit: Martensitic transformation of NiTi studied at the nanometer scale by local mechanical spectroscopy, Scripta Mater. **42** (2000) 31–36

21.23 D. Mari, L. Bataillard, D.C. Dunand, R. Gotthardt: Martensitic transformation of NiTi and NiTi-TiC composites, J. Phys. (France) **IV** (1995) C8–659–664

21.24 S. Iijima: Helical microtubules of graphitic carbon, Nature **354** (1991) 56–58

21.25 C.-H. Kiang, M. Endo, P.M. Ajayan, G. Dresselhaus, M.S. Dresselhaus: Size effects in carbon nanotubes, Phys. Rev. Lett. **81** (1998) 1869–1872

21.26 A. Thess, R. Lee, P. Nikolaev, H. Dai, P. Petit, J. Robert, C. Xu, Y.H. Lee, S.G. Kim, A.G. Rinzler, Daniel T. Colbert, G.U. Scuseria, D. Tománek, J.E. Fischer, R.E. Smalley: Crystalline ropes of metallic carbon nanotubes, Science **273** (1996) 483–487

21.27 W.Z. Li, S.S. Xie, L.X. Qian, B.H. Chang, B.S. Zou, W.Y. Zhou, R.A. Zhao, G. Wang: Large-scale synthesis of aligned carbon nanotubes, Science **274** (1996) 1701–1703

21.28 J. Kong, H.T. Soh, A.M. Cassell, C.F. Quate, H.J. Dai: Synthesis of individual single-walled carbon nanotubes on patterned silicon wafers, Nature **395** (1998) 878–881

21.29 J.P. Lu: Elastic properties of carbon nanotubes and nanoropes, Phys. Rev. Lett. **79** (1997) 1297–1300

21.30 E.W. Wong, P.E. Sheehan, C.M. Lieber: Nanobeam mechanics: Elasticity, strength and toughness of nanorods and nanotubes, Science **277** (1997) 1971–1975

21.31 J.-P. Salvetat, J.M. Bonard, N.H. Thomson, A.J. Kulik, L. Forró, W. Benoit, L. Zuppiroli: Mechanical properties of carbon nanotubes, Appl. Phys. A **69** (1999) 255–260

21.32 M.-F. Yu, B.S. Files, S. Arepalli, R.S. Ruoff: Tensile loading of ropes of single wall carbon nanotubes and their mechanical properties, Phys. Rev. Lett. **84** (2000) 5552–5555

21.33 S. Iijima, C. Brabec, A. Maiti, J. Bernholc: Structural flexibility of carbon nanotubes, J. Chem. Phys. **104** (1996) 2089–2092

21.34 J.-P. Salvetat, A.J. Kulik, J.-M. Bonard, G.A.D. Briggs, T. Stöckli, K. Méténier, S. Bonnamy, F. Béguin, N.A. Burnham, L. Forró: Elastic modulus of ordered and disordered multiwalled carbon nanotubes, Adv. Mater. **11** (1999) 161–165

21.35 J.-P. Salvetat, G.A.D. Briggs, J.-M. Bonard, R.R. Bacsa, A.J. Kulik, T. Stöckli, N. Burnham, L. Forró: Elastic and shear moduli of single-walled carbon nanotube ropes, Phys. Rev. Lett. **82** (1999) 944–947

21.36 J.M. Gere, S.P. Timoshenko: *Mechanics of Materials* (PWS-Kent, Boston 1990)

21.37 D.A. Walters, L.M. Ericson, M.J. Casavant, J. Liu, D.T. Colbert, K.A. Smith, R.E. Smalley: Elastic strain of freely suspended single-wall carbon nanotube ropes, Appl. Phys. Lett. **74** (1999) 3803–3805

21.38 M.-F. Yu, O. Lourie, M.J. Dyer, K. Moloni, T.F. Kelly, R.S. Ruoff: Strength and breaking mechanism of multiwalled carbon nanotubes under tensile load, Science **287** (2000) 637–640

21.39 G.-T. Kim, G. Gu, U. Waizmann, S. Roth: Simple method to prepare individual suspended nanofibers, Appl. Phys. Lett. **80** (2002) 1815–1817

21.40 T. Fujii, M. Suzuki, M. Miyashita, M. Yamaguchi, T. Onuki, H. Nakamura, T. Matsubara, H. Yamada, K. Nakayama: Micropattern measurement with an atomic force microscope, J. Vac. Sci. Technol. B **9** (1991) 666–669

21.41 A.S.K. Kis, B. Babic, A.J. Kulik, W. Benoît, G.A.D. Briggs, C. Schönenberger, S. Catsicas, L. Forró: Nanomechanics of microtubules, Phys. Rev. Lett. **89** (2002) 248101

21.42 M. Kurachi, M. Hoshi, H. Tashiro: Buckling of a single microtubule by optical trapping forces: Direct measurement of microtubule rigidity, Cell. Motil. Cyt. **30** (1995) 221–228

21.43 R. B. Dye, S. P. Fink, R. C. Williams: Taxol-induced flexibillity of microtubules and its reversal by Map-2 and Tau, J. Biol. Chem. **268** (1993) 6847–6850

21.44 F. Gittes, B. Mickey, J. Nettleton, J. Howard: Flexual rigidity of microtubules and actin filaments measured from thermal fluctuations in shape, J. Cell Biol. **120** (1993) 923–934

21.45 M. Elbaum, D. K. Fygenson, A. Libchaber: Buckling microtubules in vesicles, Phys. Rev. Lett. **76** (1996) 4078–4081

21.46 A. Vinckier, C. Dumortier, Y. Engelborghs, L. Hellemans: Dynamical and mechanical study of immobilized microtubules with atomic force microscopy, J. Vac. Sci. Technol. B **14** (1996) 1427–1431

21.47 Hysitron, Incorporated 2010 East Hennepin Avenue Minneapolis, MN 55413,

21.48 A. V. Kulkarni, B. Bhushan: Nanoscale mechanical property measurements using modified atomic force microscopy, Thin Solid Films **290-291** (1996) 206–210

21.49 A. V. Kulkarni, B. Bhushan: Nano/picoindentation measurements on single-crystal aluminum using modified atomic force microscopy, Mater. Lett. **29** (1996) 221–227

21.50 M. Chiba, M. Seo: Effects of dichromate treatment on mechanical properties of passivated single crystal iron (100) and (110) surfaces, Corros. Sci. **44** (2002) 2379–2391

21.51 X. C. Lu, B. Shi, L. K. Y. Li, J. Luo, X. Chang, Z. Tian, J. I. Mou: Nanoindentation and microtribological behavior of Fe-N/Ti-N multilayers with different thickness of Ti-N layers, Wear **251** (2001) 1144–1149

21.52 P. Peeters, Gvd. Horn, T. Daenen, A. Kurowski, G. Staikov: Properties of electroless and electroplated Ni-P and its application in microgalvanics, Electrochim. Acta **47** (2001) 161–169

21.53 A. Rar, J. N. Zhou, W. J. Liu, J. A. Barnard, A. Bennett, S. C. Street: Dendrimer-mediated growth of very flat ultrathin Au films, Appl. Surf. Sci. **175-176** (2001) 134–139

21.54 R. D. Ott, C. Ruby, F. Huang, M. L. Weaver, J. A. Barnard: Nanotribology and surface chemistry of reactively sputtered Ti-B-N hard coatings, Thin Solid Films **377-378** (2000) 602–606

21.55 T. Malkow, S. J. Bull: Hardness measurements on thin IBAD CN_x films – a comparative study, Surf. Coat. Technol. **137** (2001) 197–204

21.56 Y. Shima, H. Hasuyama, T. Kondoh, Y. Imaoka, T. Watari, K. Baba, R. Hatada: Mechanical properties of silicon oxynitride thin films prepared by low energy ion beam assisted deposition, Nucl. Instrum. Methods Phys. Res., Sect. B **148** (1999) 599–603

21.57 M. Göken, M. Kempf, W. D. Nix: Hardness and modulus of the lamellar microstructure in PST-TiAl studied by nanoindentations and AFM, Acta Mater. **49** (2001) 903–911

21.58 M. Göken, M. Kempf: Microstructural properties of superalloys investigated by nanoindentations in an atomic force microscope, Acta Mater. **47** (1999) 1043–1052

21.59 M. Seo, M. Chiba, K. Suzuki: Nano-mechano-electrochemistry of the iron (100) surface in solution, J. Electroanal. Chem. **473** (1999) 49–53

21.60 Y. Zhang, F. Z. Cui, X. M. Wang, Q. L. Feng, X. D. Zhu: Mechanical properties of skeletal bone in gene-mutated *stöpsel*^{dtl28d} and wild-type zebrafish (*Danio rerio*) measured by atomic force microscopy-based nanoindentation, Bone **30** (2002) 541–546

21.61 M. Finke, J. A. Hughes, D. M. Parker, K. D. Jandt: Mechanical properties of in situ demineralised human enamel measured by AFM nanoindentation, Surf. Sci. **491** (2001) 456–467

21.62 S. Habelitz, S. J. Marshall, G. W. Marshall, J. R. Balooch, M. Balooch: The functional width of the dentino-enamel junction determined by AFM-based nanoscratching, J. Struct. Biol. **135** (2001) 294–301

21.63 G. W. Marshall Jr., M. Balooch, R. R. Gallagher, S. A. Gansky, S. J. Marshall: Mechanical properties of the dentinoenamel junction: AFM studies of nanohardness, elastic modulus, and fracture, J. Biomed. Mater. Res. **54** (2001) 87–95

21.64 S. Hengsberger, A. Kulik: Nanoindentation discriminates the elastic properties of individual human bone lamellae under dry and physiological conditions, Bone **30** (2002) 178–184

21.65 E. F. Eriksen, D. W. Axelrod, F. M. Melsen: *Bone Histomorphometry*, 1st edn. (Raven, New York 1994)

21.66 C. E. Hoffler, K. E. Moore, K. Kozloff, P. K. Zysset, M. B. Brown, S. A. Goldstein: Heterogeneity of bone lamellar-level elastic moduli, Bone **26** (2000) 603–609

21.67 C. E. Hoffler, K. E. Moore, K. Kozloff, P. K. Zysset, S. A. Goldstein: Age, gender, and bone lamellae elastic moduli, J. Orth. Res. **18** (2000) 432–437

21.68 P. K. Zysset, X. E. Guo, C. E. Hoffler, K. E. Moore, S. A. Goldstein: Elastic modulus and hardness of cortical and trabecular bone lamellae measured by nanoindentation in the human femur, J. Biomech. **32** (1999) 1005–1012

21.69 J. Y. Rho, P. Zioupos, J. D. Currey, G. M. Pharr: Variations in the individual thick lamellar properties within osteons by nanoindentation, Bone **25** (1999) 295–300

21.70 J. Y. Rho, M. E. Roy, T. Y. Tsui, G. M. Pharr: Elastic properties of microstructural components of human bone tissue as measured by nanoindentation, J. Biomed. Mater. Res. **45** (1999) 48–54

21.71 S. Hengsberger, A. Kulik, P. Zysset: A combined atomic force microscopy and nanoindentation technique to investigate the elastic properties of bone structural units, Europ. Cells Mater. **1** (2001) 12–16
21.72 C. H. Turner, J. Y. Rho, Y. Takano, T. Y. Tsui, G. M. Pharr: The elastic properties of trabecular and cortical bone tissues are similar: Results from two microscopic measurement techniques, J. Biomech. **32** (1999) 437–441
21.73 M. E. Roy, J. Y. Rho, T. Y. Tsui, N. S. Evans, G. M. Pharr: Mechanical and morphological variation of the human lumbar vertebral cortical and trabecular bone, J. Biomed. Mater. Res. **44** (1999) 191–199
21.74 J. Y. Rho, G. M. Pharr: Effects of drying on the mechanical properties of bovine femur measured by nanoindentation, J. Mater. Sci.: Mater. Med. **10** (1999) 485–488
21.75 I. N. Sneddon: The relation between load and penetration in the axisymmetric Boussinesq problem for a punch of arbitrary profile, Int. J. Eng. Sci. **3** (1965) 47–57
21.76 W. C. Oliver, G. M. Pharr: An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments, Mat. Res. Soc. **7/6** (1992) 1564–1583
21.77 G. M. Pharr, W. C. Oliver, F. R. Brotzen: On the generality of the relationship among contact stiffness, contact area, and elastic modulus during indentation, J. Mater. Res. **7/3** (1992) 613–617
21.78 K. L. Johnson: *Contact Mechanics*, 1st edn. (Cambridge Univ. Press, Cambridge 1985) pp. 84–106
21.79 G. Marotti: A new theory of bone lamellation, Calcif Tissue Int. **53** (1993) 47–56
21.80 M. M. Giraud-Guille: Plywood structures in nature, Curr. Op. Solid State Mater. Sci. **3** (1998) 221–227
21.81 G. P. Evans, J. C. Behiri, J. D. Currey, W. Bonfield: Microhardness and Young's modulus in cortical bone exhibiting a wide range of mineral volume fractions, and in a bone analogue, J. Mater. Sci.: Mater. Med. **1** (1990) 38–43
21.82 J. G. Swadener, G. M. Pharr: Indentation of elastically anisotropic half-spaces by cones and parabolae of revolution, Philos. Mag. A **81** (2001) 447–466

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

22. Nanomechanical Properties of Solid Surfaces and Thin Films

Nanomechan

Instrumentation for the testing of mechanical properties on the submicron scale has developed enormously in recent years. This has enabled the mechanical behavior of surfaces, thin films, and coatings to be studied with unprecedented accuracy. In this chapter, the various techniques available for studying nanomechanical properties are reviewed with particular emphasis on nanoindentation. The standard methods for analyzing the raw data obtained using these techniques are described, along with the main sources of error. These include residual stresses, environmental effects, elastic anisotropy, and substrate effects. The methods that have been developed for extracting thin-film mechanical properties from the often convoluted mix of film and substrate properties measured by nanoindentation are discussed. Interpreting the data is frequently difficult, as residual stresses can modify the contact geometry and, hence, invalidate the standard analysis routines. Work hardening in the deformed region can also result in variations in mechanical behavior with indentation depth. A further unavoidable complication stems from the ratio of film to substrate mechanical properties and the depth of indentation in comparison to film thickness. Even very shallow indentations may be influenced by substrate properties if the film is hard and very elastic but the substrate is compliant. Under these circumstances nonstandard methods of analysis must be used. For multilayered systems many different mechanisms affect the nanomechanical behavior, including Orowan strengthening, Hall–Petch behavior, image force effects, coherency and thermal stresses, and composition modulation.

The application of nanoindentation to the study of phase transformations in semiconductors, fracture in brittle materials, and mechanical properties in biological materials are described. Recent developments such as the testing of viscoelasticity using nanoindentation methods are likely to be particularly important in future studies of polymers and biological materials. The importance of using a range of complementary methods such as electron microscopy, in situ AFM imaging, acoustic monitoring, and electrical contact measurements is emphasized. These are especially important on the nanoscale because so many different physical and chemical processes can affect the measured mechanical properties.

22.1 **Instrumentation** 688
  22.1.1 AFM and Scanning Probe Microscopy 688
  22.1.2 Nanoindentation 689
  22.1.3 Adaptations of Nanoindentation 690
  22.1.4 Complimentary Techniques 691
  22.1.5 Bulge Tests 691
  22.1.6 Acoustic Methods 692
  22.1.7 Imaging Methods 693

22.2 **Data Analysis** 694
  22.2.1 Elastic Contacts 694
  22.2.2 Indentation of Ideal Plastic Materials 694
  22.2.3 Adhesive Contacts 695
  22.2.4 Indenter Geometry 696
  22.2.5 Analyzing Load/Displacement Curves 696
  22.2.6 Modifications to the Analysis 699
  22.2.7 Alternative Methods of Analysis 700
  22.2.8 Measuring Contact Stiffness 701
  22.2.9 Measuring Viscoelasticity 702

22.3 **Modes of Deformation** 702
  22.3.1 Defect Nucleation 702
  22.3.2 Variations with Depth 704
  22.3.3 Anisotropic Materials 704
  22.3.4 Fracture and Delamination 704
  22.3.5 Phase Transformations 705

22.4 **Thin Films and Multilayers** 707
  22.4.1 Thin Films 707
  22.4.2 Multilayers 709

22.5 **Developing Areas** 711

References 712

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

When two bodies come into contact their surfaces experience the first and usually largest mechanical loads. Hence, characterizing and understanding the mechanical properties of surfaces is of paramount importance in a wide range of engineering applications. Obvious examples of where surface mechanical properties are important are in wear-resistant coatings on reciprocating surfaces and hard coatings for machine tool bits. This chapter details the current methods for measuring the mechanical properties of surfaces and highlights some of the key experimental results that have been obtained.

The experimental technique that is highlighted in this chapter is nanoindentation. This is for the simple reason that it is now recognized as the preferred method for testing thin film and surface mechanical properties. Despite this recognition, there are still many pitfalls for the unwary researcher when performing nanoindentation tests. The commercial instruments that are currently available all have attractive, user-friendly software, which makes the performance and analysis of nanoindentation tests easy. Hidden within the software, however, are a myriad of assumptions regarding the tests that are being performed and the material that is being examined. Unless the researcher is aware of these, there is a real danger that the results obtained will say more about the analysis routines than they do about the material being tested.

22.1 Instrumentation

The instruments used to examine nanomechanical properties of surfaces and thin films can be split into those based on point probes and those complimentary methods that can be used separately or in conjunction with point probes. The complimentary methods include a wide variety of techniques ranging from optical tests such as micro-Raman spectroscopy to high-energy diffraction studies using X-rays, neutrons, or electrons to mechanical tests such as bulge or blister testing.

Point-probe methods have developed from two historically different methodologies, namely, scanning probe microscopy [22.1] and microindentation [22.2]. The two converge at a length scale between 10–1,000 nm. Point-probe mechanical tests in this range are often referred to as nanoindentation.

22.1.1 AFM and Scanning Probe Microscopy

Atomic force microscopy (AFM) and other scanning probe microscopies are covered in detail elsewhere in this volume, but it is worth briefly highlighting the main features in order to demonstrate the similarities to nanoindentation. There are now a myriad of different variants on the basic scanning probe microscope. All use piezoelectric stacks to move either a probe tip or the sample with subnanometer precision in the lateral and vertical planes. The probe itself can be as simple as a tungsten wire electrochemically polished to give a single atom at the tip, or as complex as an AFM tip that is bio-active with, for instance, antigens attached. A range of scanning probes have been developed with the intention of measuring specific physical properties such as magnetism and heat capacity.

To measure mechanical properties with an AFM, the standard configuration is a hard probe tip (such as silicon nitride or diamond) mounted on a cantilever (see Fig. 22.1). The elastic deflection of the cantilever is monitored either directly or via a feedback mechanism

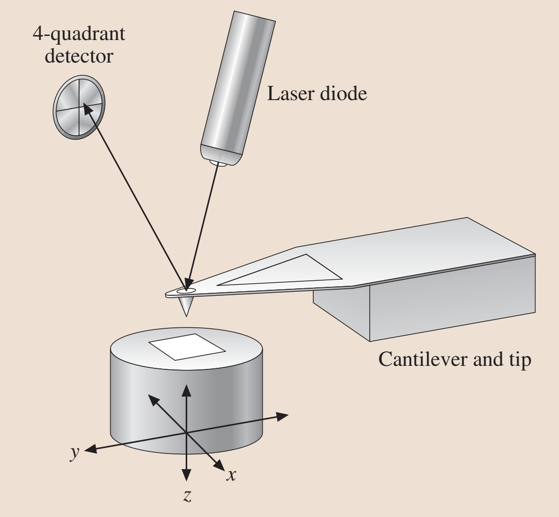

Fig. 22.1 Diagram of a commercial AFM. The AFM tip is mounted on a compliant cantilever, and a laser light is reflected off the back of the cantilever onto a position-sensitive detector (4-quadrant detector). Any movements of the cantilever beam cause a deflection of the laser light that the detector senses. The sample is moved using piezoelectric stack, and forces are calculated from the cantilever's stiffness and the measured deflection

to measure the forces acting on the probe. In general, the forces experienced by the probe tip split into attractive or repulsive forces. As the tip approaches the surface, it experiences intermolecular forces that are attractive, although they can be repulsive under certain circumstances [22.3]. Once in contact with the surface the tip usually experiences a combination of attractive intermolecular forces and repulsive elastic forces. Two schools of thought exist regarding the attractive forces when the tip is in contact with the surface. The first is often referred to as the DMT or Bradley model. It holds that attractive forces only act outside the region of contact [22.4–7]. The second theory, usually called the JKR model, assumes that all the forces experienced by the tip, whether attractive or repulsive, act in the region of contact [22.8]. Most real nanoscale contacts lie somewhere between these two theoretical extremes.

22.1.2 Nanoindentation

The fundamental difference between AFM and nanoindentation is that during a nanoindentation experiment an external load is applied to the indenter tip. This load enables the tip to be pushed into the sample, creating a nanoscale impression on the surface, otherwise referred to as a nanoindentation or nanoindent.

Conventional indentation or microindentation tests involve pushing a hard tip of known geometry into the sample surface using a fixed peak load. The area of indentation that is created is then measured, and the mechanical properties of the sample, in particular its hardness, is calculated from the peak load and the indentation area. Various types of indentation testing are used in measuring hardness, including Rockwell, Vickers, and Knoop tests. The geometries and definitions of hardness used in these tests are shown by Fig. 22.2.

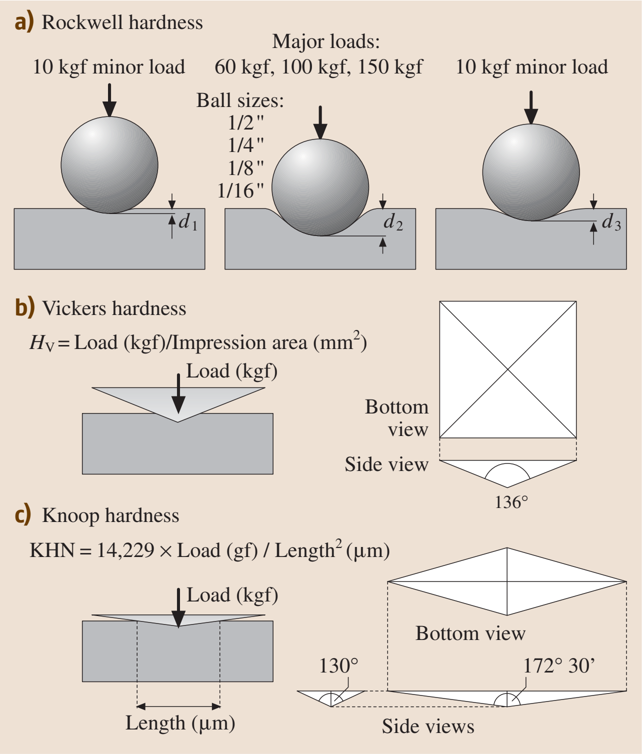

Fig. 22.2 **(a)** The standard Rockwell hardness test involves pushing a ball into the sample with a minor load, recording the depth, d_1, then applying a major load and recording the depth, d_2, then returning to the lower load and recording the depth, d_3. Using the depths, the hardness is calculated. **(b)** Vickers hardness testing uses a four-sided pyramid pushed into the sample with a known load. The area of the resulting indentation is measured optically and the hardness calculated as the load is divided by area. **(c)** Knoop indentation uses the same definition of hardness as the Vickers test, load divided by area, but the indenter geometry has one long diagonal and one short diagonal

When indentations are performed on the nanoscale there is a basic problem in measuring the size of the indents. Standard optical techniques cannot easily be used to image anything smaller than a micron, while electron microscopy is simply impractical due to the time involved in finding and imaging small indents. To overcome these difficulties, nanoindentation methods have been developed that continuously record the load, displacement, time, and contact stiffness throughout the indentation process. This type of continuously recording indentation testing was originally developed in the former Soviet Union [22.9–13] as an extension of microindentation tests. It was applied to nanoscale indentation testing in the early 1980s [22.14, 15], hence, giving rise to the field of nanoindentation testing.

In general, nanoindentation instruments include a loading system that may be electrostatic, electromagnetic, or mechanical, along with a displacement measuring system that may be capacitive or optical. Schematics of several commercial nanoindentation instruments are shown in Fig. 22.3a–c.

Among the many advantages of nanoindentation over conventional microindentation testing is the ability to measure the elastic, as well as the plastic properties of the test sample. The elastic modulus is obtained from the contact stiffness (S) using the following equation that

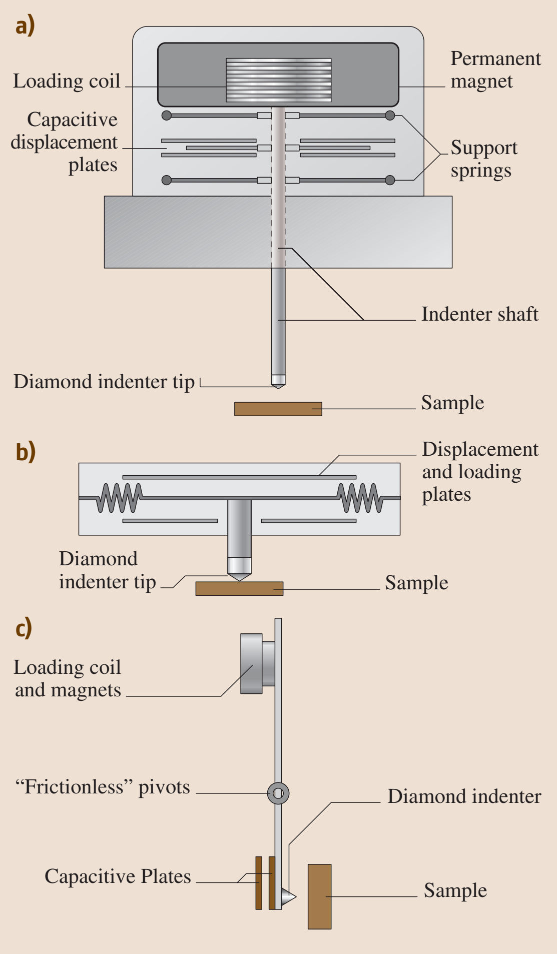

Fig. 22.3a–c Schematics of three commercial nanoindentation devices made by **(a)** MTS Nanoinstruments, Oak Ridge, Tennessee, **(b)** Hysitron Inc., Minneapolis, Minnesota, **(c)** Micro Materials Limited, Wrexham, UK. Instruments **(a)** and **(c)** use electromagnetic loading, while **(b)** uses electrostatic loads

appears to be valid for all elastic contacts [22.16, 17]:

$$S = \frac{2}{\sqrt{\pi}} E_r \sqrt{A} \,. \qquad (22.1)$$

A is the contact area and E_r is the reduced modulus of the tip and sample as given by:

$$\frac{1}{E_r} = \frac{\left(1-\nu_t^2\right)}{E_t} + \frac{\left(1-\nu_s^2\right)}{E_s} \,, \qquad (22.2)$$

where E_t, ν_t and E_s, ν_s are the elastic modulus and Poissons ratio of the tip and sample, respectively.

22.1.3 Adaptations of Nanoindentation

Several adaptations to the basic nanoindentation setup have been used to obtain additional information about the processes that occur during nanoindentation testing, for example, in situ measurements of acoustic emissions and contact resistance. Environmental control has also been used to examine the effects of temperature and surface chemistry on the mechanical behavior of nanocontacts. In general, it is fair to say that the more information that can be obtained and the greater the control over the experimental parameters the easier it will be to understand the nanoindentation results. Load-displacement curves provide a lot of information, but they are only part of the story.

During nanoindentation testing discontinuities are frequently seen in the load–displacement curve. These are often called "pop-ins" or "pop-outs", depending on their direction. These sudden changes in the indenter displacement, at a constant load (see Fig. 22.4), can be caused by a wide range of events, including fracture, delamination, dislocation multiplication, or nucleation and phase transformations. To help distinguish between the various sources of discontinuities, acoustic transducers have been placed either in contact with the sample or immediately behind the indenter tip. For example, the results of nanoindentation tests that monitor acoustic emissions have shown that the phase transformations seen in silicon during nanoindentation are not the sudden events that they would appear to be from the load–displacement curve. There is no acoustic emission associated with the pop-out seen in the unloading curve of silicon [22.18]. An acoustic emission would be expected if there were a very rapid phase transformation causing a sudden change in volume. Fracture and delamination of films, however, give very strong acoustic signals [22.19],

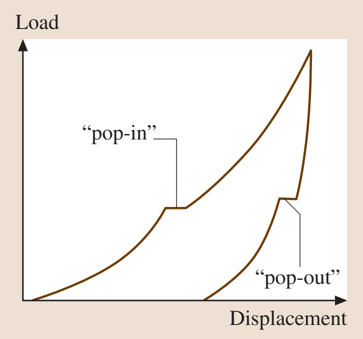

Fig. 22.4 Sketch of a load/displacement curve showing a pop-in and a pop-out

but the exact form of the signal appears to be more closely related to the sample geometry than to the event [22.20].

Additional information about the nature of the deformed region under the nanoindentation can be obtained by performing in situ measurements of contact resistance. The basic setup for this type of testing is shown in Fig. 22.5. An electrically conductive tip is needed to study contact resistance. Consequently, a conventional diamond tip is of limited use. Elastic, hard, and metallically conductive materials such as vanadium carbide can be used as substitutes for diamond [22.21, 22], or a thin conductive film (e.g., Ag) can be deposited on the diamond's surface (such a film is easily transferred to the indented surface so great care must be taken if multiple indents are performed). Measurements of contact resistance have been most useful for examining phase transformations in semiconductors [22.21, 22] and the dielectric breakdown of oxide films under mechanical loading [22.23].

One factor that is all too frequently neglected during nanoindentation testing is the effect of the experimental environment. Two obvious ways in which the environment can affect the results of nanoindentation tests are increases in temperature, which give elevated creep rates, and condensation of water vapor, which modifies the tip-sample interactions. Both of these environmental effects have been shown to significantly affect the measured mechanical properties and the modes of deformation that occur during nanoindentation [22.24–27]. Other environmental effects, for instance, those due to photoplasticity or hydrogen ion absorption, are also possible, but they are generally less troublesome than temperature fluctuations and variations in atmospheric humidity.

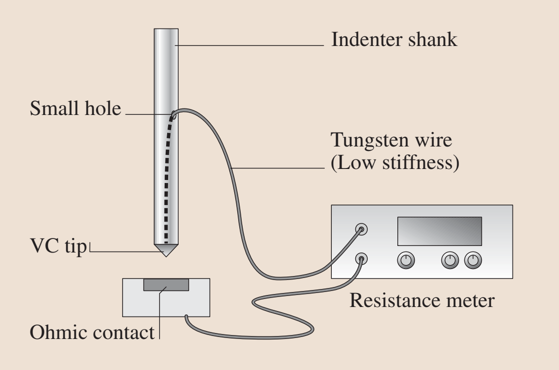

Fig. 22.5 Schematic of the basic setup for making contact resistance measurements during nanoindentation testing

22.1.4 Complimentary Techniques

Nanoindentation testing is probably the most important technique for characterizing the mechanical properties of thin films and surfaces, but there are many alternative or additional techniques that can be used. One of the most important alternative methods for measuring the mechanical properties of thin films uses bulge or blister testing [22.28]. Bulge tests are performed on thin films mounted on supporting substrates. A small area of the substrate is removed to give a window of unsupported film. A pressure is then applied to one side of the window causing it to bulge. By measuring the height of the bulge, the stress-strain curve and the residual stress are obtained. The basic configuration for bulge testing is shown in Fig. 22.6.

22.1.5 Bulge Tests

The original bulge tests used circular windows because they are easier to analyze mathematically, but now square and rectangular windows have become common [22.29]. These geometries tend to be easier to fabricate. Unfortunately, there are several sources of errors in bulge testing that can potentially lead to large errors in the measured mechanical properties. These errors at one time led to the belief that multilayer films can show a "super modulus" effect, where the elastic modulus of the multilayer is several times that of its constituent

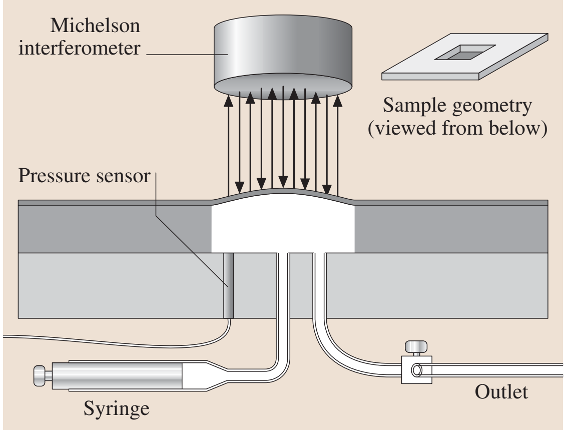

Fig. 22.6 Schematic of the basic setup for bulge testing. The sample is prepared so that it is a thin membrane, and then a pressure is applied to the back of the membrane to make it bulge upwards. The height of the bulge is measured using an interferometer

layers [22.30]. It is now accepted that any enhancement to the elastic modulus in multilayer films is small, on the order of 15% [22.31]. The main sources of error stem from compressive stresses in the film (tensile stress is not a problem), small variations in the dimensions of the window, and uncertainty in the exact height of the bulge. Despite these difficulties, one advantage of bulge testing over nanoindentation testing is that the stress state is biaxial, so that only properties in the plane of the film are measured. In contrast, nanoindentation testing measures a combination of in-plane and out-of-plane properties.

22.1.6 Acoustic Methods

Acoustic and ultrasonic techniques have been used for many years to study the elastic properties of materials. Essentially, these techniques take advantage of the fact that the velocity of sound in a material is dependent on the inter-atomic or inter-molecular forces in the material. These, of course, are directly related to the material's elastic constants. In fact, any nonlinearity of inter-atomic forces enables slight variations in acoustic signals to be used as a measure of residual stress.

An acoustic method ideally suited to studying surfaces is scanning acoustic microscopy (SAM) [22.32]. There are also several other techniques that have been used to study surface films and multilayers, but we will first consider SAM in detail. In a SAM, a lens made of sapphire is used to bring acoustic waves to focus via a coupling fluid on the surface. A small piezoelectric transducer at the top of the lens generates the acoustic signal. The same transducer can be used to detect the signal when the SAM is used in reflection mode. The use of a transducer as both generator and detector, a common imaging mode, necessitates the use of a pulsed rather than a continuous acoustic signal. Continuous waves can be used if phase changes are used to build up the image. The transducer lens generates two types of acoustic waves in the material: longitudinal and shear. The ability of a solid to sustain both types of wave (liquids can only sustain longitudinal waves) gives rise to a third type of acoustic wave called a Rayleigh, or surface, wave. These waves are generated as a result of superposition of the shear and longitudinal waves with a common phase velocity. The stresses and displacements associated with a Rayleigh wave are only of significance to a depth of ≈ 0.6 Rayleigh wavelengths below the solid surface. Hence, using SAM to examine Rayleigh waves in a material is a true surface characterization technique.

Using a SAM in reflection mode gives an image where the contrast is directly related to the Rayleigh wave velocity, which is in turn a function of the material's elastic constants. The resolution of the image depends on the frequency of the transducer used, i. e., for a 2 GHz signal a resolution better than 1 µm is achievable. The contrast in the image results from the interference of two different waves in the coupling fluid. Rayleigh waves that are excited in the surface "leak" into the coupling fluid and interfere with the acoustic signal that is directly reflected back from the surface. It is usually assumed that the properties of the coupling fluid are well characterized. The interference of the two waves gives a characteristic $V(z)$ curve, as illustrated by Fig. 22.7, where z is the separation between the lens and the surface. Analyzing the periodicity of the $V(z)$ curves provides information on the Rayleigh wave velocity. As with other acoustic waves, the Rayleigh velocity is related to the elastic constants of the material. When using the SAM for a material's characterization, the lens is usually held in a fixed position on the surface. By using a lens designed specifically to give a line-focus beam, rather than the standard spherical lens, it is possible to use SAM to look at anisotropy in the wave velocity [22.33] and hence in elastic properties by producing waves with a specific direction.

One advantage of using SAM in conjunction with nanoindentation to characterize a surface is that the measurements obtained with the two methods have a slightly different dependence on the test material's elastic properties, E_s and ν_s (the elastic modulus and Poisson's ratio). As a result, it is possible to use SAM

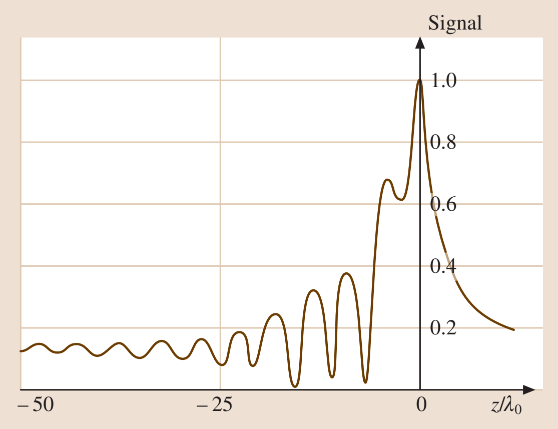

Fig. 22.7 A typical $V(z)$ curve obtained with a SAM when testing fused silica

and nanoindentation combined to find both E_s and ν_s, as illustrated by Fig. 22.8 [22.34]. This is not possible when using only one of the techniques alone.

In addition to measuring surface properties, SAM has been used to study thin films on a surface. However, the Rayleigh wave velocity can be dependent on a complex mix of the film and substrate properties. Other acoustic methods have been utilized to study freestanding films. A freestanding film can be regarded as a plate, and, therefore, it is possible to excite Lamb waves in the film. Using a pulsed laser to generate the waves and a heterodyne interferometer to detect the arrival of the Lamb wave, it is possible to measure the flexural modulus of the film [22.35]. This has been successfully demonstrated for multilayer films with a total thickness $< 10\,\mu\text{m}$. In the plate configuration, due to the nonlinearity of elastic properties, it is also possible to measure stress. This has been demonstrated for horizontally polarized shear waves in plates [22.36], but thin plates require very high frequency transducers or laser sources.

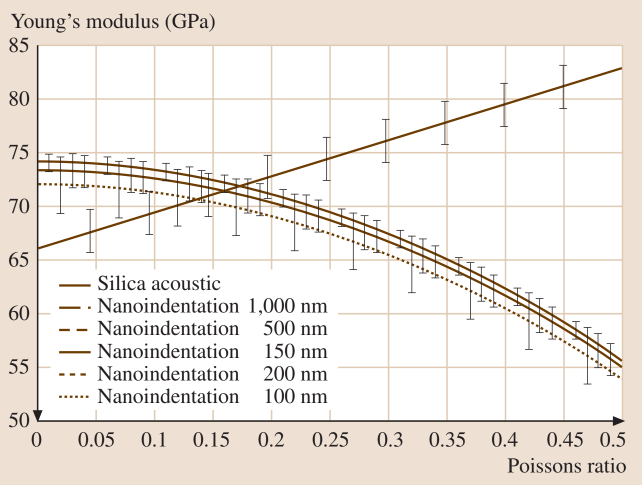

Fig. 22.8 Because SAM and nanoindentation have different dependencies on Young's modulus, E, and Poissons ratio, ν, it is possible to use the two techniques in combination to find E and ν [22.34]. On the graph, the intersection of the curves gives E and ν

22.1.7 Imaging Methods

When measuring the mechanical properties of a surface or thin film using nanoindentation, it is not always easy to visualize what is happening. In many instances there is a risk that the mechanical data can be completely misinterpreted if the geometry of the test is not as expected. To expedite the correct interpretation of the mechanical data, it is generally worthwhile to use optical, electron, or atomic force microscopy to image the nanoindentations. Obviously, optical techniques are only of use for larger indentations, but they will often reveal the presence of median or lateral cracks [22.37]. Electron microscopy and AFM, however, can be used to examine even the smallest nanoindentations. The principle problem with these microscopy techniques is the difficulty in finding the nanoindentations. It is usually necessary to make large, "marker" indentations in the vicinity of the nanoindentations to be examined in order to find them [22.38].

It is possible to see features such as extrusions with a scanning electron microscope (SEM) [22.39], as well as pile-up and sink-in around the nanoindents, though AFM is generally better for this. Transmission electron microscopy (TEM) is useful for examining what has happened subsurface, for instance, the indentation induced dislocations in a metal [22.40] or the phases present under a nanoindent in silicon [22.21]. However, with TEM there is the added difficulty of sample preparation and the associated risk of observing artifacts. Recently, there has been considerable interest in the use of focused ion beams to cut cross sections through nanoindents [22.41]. When used in conjunction with SEM or TEM this provides an excellent means to see what has happened in the subsurface region.

One other technique that has proved to be useful in studying nanoindents is micro-Raman spectroscopy. This involves using a microscope to focus a laser on the sample surface. The same microscope is also used to collect the scattered laser light, which is then fed into a spectroscope. The Raman peaks in the spectrum provide information on the bonding present in a material, while small shifts in the wave number of the peaks can be used as a measure of strain. Micro-Raman has proven to be particularly useful for examining the phases present around nanoindentations in silicon [22.42].

22.2 Data Analysis

The analysis of nanoindentation data is far from simple. This is mostly due to the lack of effective models that are able to combine elastic and plastic deformation under a contact. However, provided certain precautions are taken, the models for perfectly elastic deformation and ideal plastic materials can be used in the analysis of nanoindentation data. For this reason, it is worth briefly reviewing the models for perfect contacts.

22.2.1 Elastic Contacts

The theoretical modeling of elastic contacts can be traced back many years, at least to the late nineteenth century and the work of *Hertz* (1882) [22.43] and *Boussinesq* (1885) [22.44]. These models, which are still widely used today, consider two axisymmetric curved surfaces in contact over an elliptical region (see Fig. 22.9). The contact region is taken to be small in comparison to the radius of curvature of the contacting surfaces, which are treated as elastic half-spaces. For an elastic sphere, radius R, in contact with a flat, elastic half-space, the contact region will be circular and the Hertz model gives the following relationships:

$$a=\sqrt[3]{\frac{3PR}{4E_\mathrm{r}}}\,, \tag{22.3}$$

$$\delta=\sqrt[3]{\frac{9P^2}{16RE_\mathrm{r}^2}}\,, \tag{22.4}$$

$$P_0=\sqrt[3]{\frac{6PE_\mathrm{r}^2}{\pi^3R^2}}\,, \tag{22.5}$$

where a is the radius of the contact region, E_r is given by (22.2), δ is the displacement of the sphere into the surface, P is the applied load and P_0 is the maximum pressure under the contact (in this case at the center of the contact).

The work of *Hertz* and *Boussinesq* was extended by *Love* [22.45, 46] and later by *Sneddon* [22.47], who simplified the analysis using Hankel transforms. *Love* showed how *Boussinesq*'s model could be used for a flat-ended cylinder and a conical indenter, while *Sneddon* produced a generalized relationship for any rigid axisymmetric punch pushed into an elastic half-space. *Sneddon* applied his new analysis to punches of various shapes and derived the following relationships between the applied load, P, and displacement, δ, into the elastic half-space for, respectively, a flat-ended cylinder, a cone of semi-vertical angle ϕ, and a parabola of revolution where $a^2=2k\delta$:

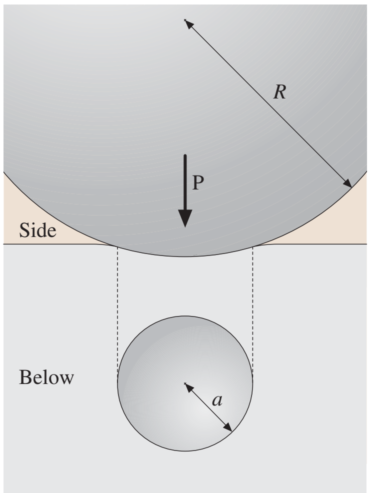

Fig. 22.9 Hertzian contact of a sphere, radius R, on a semi-infinite, flat surface. The contact in this case is a circular region of radius a

$$P=\frac{4\mu a\delta}{1-\nu}\,, \tag{22.6}$$

$$P=\frac{4\mu\cot\phi}{\pi\,(1-\nu)}\delta^2\,, \tag{22.7}$$

$$P=\frac{8\mu}{3\,(1-\nu)}\left(2k\delta^3\right)^{1/2}\,, \tag{22.8}$$

where μ and ν are the shear modulus and Poisson's ratio of the elastic half-space, respectively.

The key point to note about (22.6), (22.7), and (22.8) is that they all have the same basic form, namely:

$$P=\alpha\delta^m\,, \tag{22.9}$$

where α and m are constants for each geometry.

Equation (22.9) and the relationships developed by Hertz and his successors, (22.3–22.8), form the foundation for much of the current nanoindentation data analysis routines.

22.2.2 Indentation of Ideal Plastic Materials

Plastic deformation during indentation testing is not easy to model. However, the indentation response of ideal plastic metals was considered by *Tabor* in his classic text, "The Hardness of Metals" [22.48]. An ideal plastic material (or more accurately an ideal elastic-plastic material) has a linear stress-strain curve until it reaches its elastic limit and then yields plastically at a yield stress,Y_0, that remains constant even after deformation has commenced. In a 2-D problem, the yielding occurs because the Huber-Mises [22.49] criterion has been reached. In other words, the maximum shear stress acting on the material is around $1.15Y_0/2$.

First, we consider a 2-D flat punch pushed into an ideal plastic material. By using the method of slip lines it is found that the mean pressure, P_m, across the end of the punch is related to the yield stress by:

$$P_\mathrm{m} = 3Y_0 . \qquad (22.10)$$

If the *Tresca* criterion [22.50] is used, then P_m is closer to $2.6Y_0$. In general, for both 2-D and three-dimensional punches pushed into ideal plastic materials, full plasticity across the entire contact region can be expected when $P_\mathrm{m} = 2.6$ to $3.0Y_0$. However, significant deviations from this range can be seen if, for instance, the material undergoes work-hardening during indentation, or the material is a ceramic, or there is friction between the indenter and the surface.

The apparently straightforward relationship between P_m and Y_0 makes the mean pressure a very useful quantity to measure. In fact, P_m is very similar to the Vickers hardness, H_V, of a material:

$$H_\mathrm{V} = 0.927P_\mathrm{m} . \qquad (22.11)$$

During nanoindentation testing it is the convention to take the mean pressure as the nanohardness. Thus, the “nanohardness”, H, is defined as the peak load, P, applied during a nanoindentation divided by the projected area, A, of the nanoindentation in the plane of the surface, hence:

$$H = \frac{P}{A} . \qquad (22.12)$$

22.2.3 Adhesive Contacts

During microindentation testing and even most nanoindentation testing the effects of intermolecular and surface forces can be neglected. Very small nanoindentations, however, can be influenced by the effects of intermolecular forces between the sample and the tip. These adhesive effects are most readily seen when testing soft polymers, but there is some evidence that forces between the tip and sample may be important in even relatively strong materials [22.51, 52].

Contact adhesion is usually described by either the JKR or DMT model, as discussed earlier in this chapter. Both the models consider totally elastic spherical contacts under the influence of attractive surface forces. The JKR model considers the surface forces in terms of the associated surface energy, whereas the DMT model considers the effects of adding van der Waals forces to the Hertzian contact model. The differences between the two models are illustrated by Fig. 22.10.

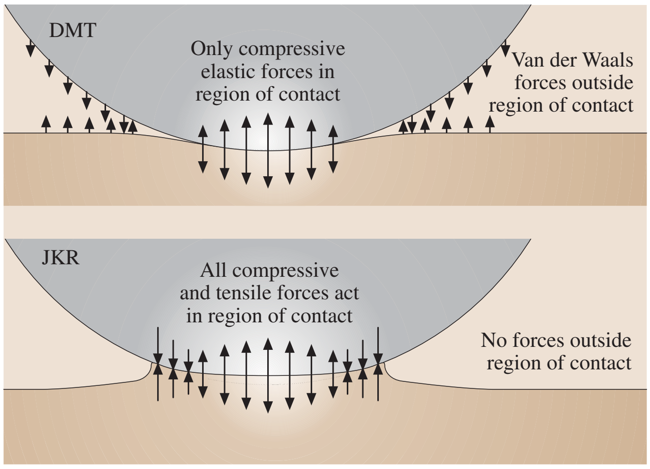
DMT
Only compressive elastic forces in region of contact
Van der Waals forces outside region of contact
JKR
All compressive and tensile forces act in region of contact
No forces outside region of contact

Fig. 22.10 The contact geometry for the DMT and JKR models for adhesive contact. Both models are based on the Hertzian model. In the DMT, model van der Waals forces outside the region of contact introduce an additional load in the Hertz model. But for the JKR model, it is assumed that tensile, as well as compressive stresses can be sustained within the region of contact

For nanoindentation tests conducted in air the condensation of water vapor at the tip-sample interface usually determines the size of the adhesive force acting during unloading. The effects of water vapor on a single nanoasperity contact have been studied using force-controlled AFM techniques [22.53] and, more recently, nanoindentation methods [22.26]. Unsurprisingly, it has also been found that water vapor can affect the deformation of surfaces during nanoindentation testing [22.27].

In addition to water vapor, other surface adsorbates can cause dramatic changes in the nanoscale mechanical behavior. For instance, oxygen on a clean metal surface can cause an increase in the apparent strength of the metal [22.54]. These effects are likely to be related to, firstly, changes in the surface and intermolecular forces acting between the tip and the sample and, secondly, changes in the mechanical stability of surface nanoasperities and ledges. Adsorbates can help stabilize atomic-scale variations in surface morphology, thereby making defect generation at the surface more difficult.

22.2.4 Indenter Geometry

All of the indenter geometries considered up to this point have been axisymmetric, largely because they are easier to deal with theoretically. Unfortunately, fabricating axisymmetric nanoindentation tips is extremely difficult, because shaping a hard tip on the scale of a few nanometers is virtually impossible. Despite these problems, there has been considerable effort put into the use of spherical nanoindentation tips [22.55]. This clearly demonstrates that the spherical geometry can be useful at larger indentation depths.

Because of the problems associated with creating axisymmetric nanoindentation tips, pyramidal indenter geometries have now become standard during nanoindentation testing. The most common geometries are the three-sided Berkovich pyramid and cube-corner (see Fig. 22.11). The Berkovich pyramid is based on the four-sided Vickers pyramid, the opposite sides of which make an 136° angle. For both the Vickers and Berkovich pyramids the cross-sectional area of the pyramid's base, A, is related to the pyramid's height, D, by:

$$A = 24.5D^2 \,. \tag{22.13}$$

The cube-corner geometry is now widely used for making very small nanoindentations, because it is much sharper than the Berkovich pyramid. This makes it easier to initiate plastic deformation at very light loads, but great care should be taken when using the cube-corner geometry. Sharp cube-corners can wear down quickly and become blunt, hence the cross-sectional area as a function of depth can change over the course of several indentations. There is also a potential problem with the standard analysis routines [22.56], which were developed for much blunter geometries and are based on the elastic contact models outlined earlier. The elastic contact models all assume the displacement into the surface is small compared to the tip radius. For the cube-corner geometry this is probably only the case for nanoindentations that are no more than a few nanometers deep.

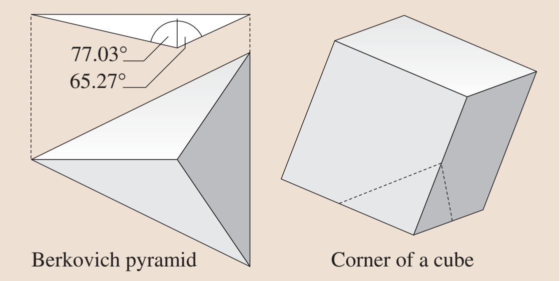

Fig. 22.11 The ideal geometry for the three-sided Berkovich pyramid and cube corner tips

22.2.5 Analyzing Load/Displacement Curves

The load/displacement curves obtained during nanoindentation testing are deceptively simple. Most newcomers to the area will see the curves as being somewhat akin to the stress/strain curves obtained during tensile testing. There is also a real temptation just to use the values of hardness, H, and elastic modulus, E, obtained from standard analysis software packages as the "true" values. This may be the case in many instances, but for very shallow nanoindents and tests on thin films the geometry of the contact can differ significantly from the geometry assumed in the analysis routines. Consequently, experimentalists should think very carefully about the test itself before concluding that the values of H and E are correct.

The basic shape of a load/displacement curve can reveal a great deal about the type of material being tested. Figure 22.12 shows some examples of ideal curves for materials with different elastic moduli and yield stresses. Discontinuities in the load/displacement curve can also provide information on such processes as fracture, dislocation nucleation, and phase transformations. Initially,

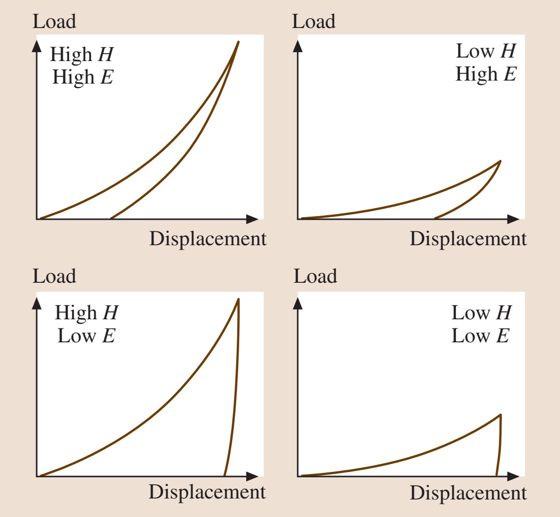

Fig. 22.12 Examples of load/displacement curves for idealized materials with a range of hardness and elastic properties

though, we will consider ideal situations such as those illustrated by Fig. 22.12.

The loading section of the load/displacement curve approximates a parabola [22.60] whose width depends on a combination of the material's elastic and plastic properties. The unloading curve, however, has been shown to follow a more general relationship [22.56] of the form:

$$P = \alpha\,(\delta - \delta_i)^m \;, \qquad (22.14)$$

where δ is the total displacement and δ_i is the intercept of the unloading curve with the displacement axis shown in Fig. 22.13.

Equation (22.14) is essentially the same as (22.9) but with the origin displaced. Since (22.9) is obtained by considering purely elastic deformation, it follows that the unloading curve is exhibiting purely elastic behavior. Since the shape of the unloading curve is determined by the elastic recovery of the indented region, it is not entirely surprising that its shape resembles that found for purely elastic deformation. What is fortuitous is that the elastic analysis used for an elastic half-space seems to be valid for a surface where there is a plastically formed indentation crater present under the contact. However, the validity of this analysis may only hold when the crater is relatively shallow and the geometry of the surface does not differ significantly from that of a flat, elastic half-space. For nanoindentations with a Berkovich pyramid, this is generally the case.

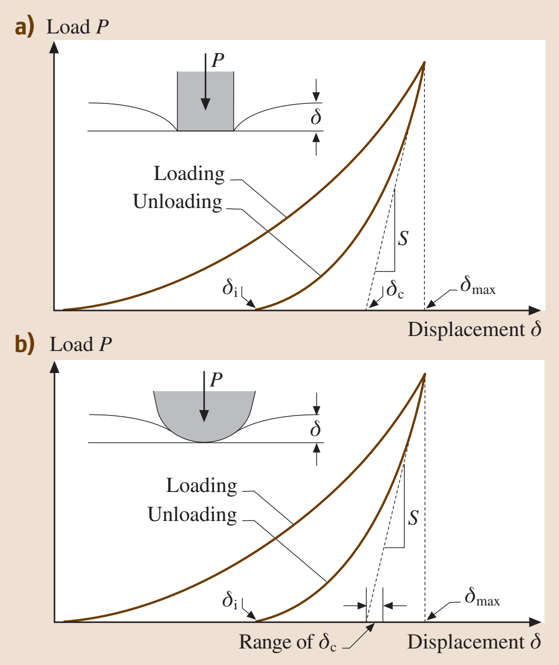

Fig. 22.13a,b Analysis of the load/displacement curve gives the contact stiffness, S, and the contact depth, δ_c. These can then be used to find the hardness, H, and elastic, or Young's modulus, E. **(a)** The first method of analysis [22.57–59] assumed the unloading curve could be approximated by a flat punch on an elastic half-space. **(b)** A more refined analysis [22.56] uses a paraboloid on an elastic half-space

Before *Oliver* and *Pharr* [22.56] proposed their now standard method for analyzing nanoindentation data, the analysis had been based on the observation that the initial part of the unloading curve is almost linear. A linear unloading curve, equivalent to $m = 1$ in (22.14), is expected when a flat punch is used on an elastic half-space. The flat punch approximation for the unloading curve was used in [22.57–59] to analyze nanoindentation data. When Oliver and Pharr looked at a range of materials they found m was typically larger than 1, and that $m = 1.5$, or a paraboloid, was a better approximation than a flat punch. Oliver and Pharr used (22.1) and (22.12) to obtain the values for a material's elastic modulus and hardness. Equation (22.1) relates the contact stiffness during the initial part of the unloading curve (see Fig. 22.13) to the reduced elastic modulus and the contact area at the peak load. Equation (22.12) gives the hardness as the peak load divided by the contact area. It is immediately obvious that the key to measuring the mechanical properties of a material is knowing the contact area at the peak load. This is the single most important factor in analyzing nanoindentation data. Most mistakes in the analysis come from incorrect assumptions about the contact area.

To find the contact area, a function relating the contact area, A_c, to the contact depth, δ_c, is needed. For a perfect Berkovich pyramid this would be the same as (22.13). But since making a perfect nanoindenter tip is impossible, an expanded equation is used:

$$A_c\,(\delta_c) = 24.5\delta_c^2 + \sum_{j=1}^{7} C_j \sqrt[2^j]{\delta_c} \;, \qquad (22.15)$$

where C_j are calibration constants of the tip.

There is a crucial step in the analysis before A_c can be calculated, namely, finding δ_c. The contact depth is not the same as the indentation depth, because the surface around the indentation will be elastically deflected during loading, as illustrated by Fig. 22.14. Sneddon's analysis [22.47] provides a way to calculate the deflection of the surface at the edge of an axisymmetric

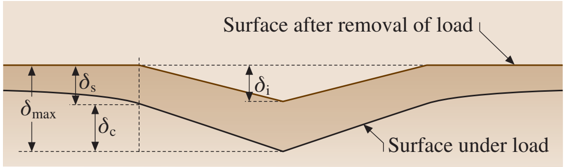

Fig. 22.14 Profile of surface under load and unloaded showing how δ_c compares to δ_i and δ_{max}

contact. Subtracting the deflection from the total indentation depth at peak load gives the contact depth. For a paraboloid, as used by *Oliver* and *Pharr* [22.56] in their analysis, the elastic deflection at the edge of the contact is given by:

$$\delta_s = \varepsilon \frac{P}{S} = 0.75 \frac{P}{S} \,, \tag{22.16}$$

where S is the contact stiffness and P the peak load. The constant ε is 0.75 for a paraboloid, but ranges between 0.72 (conic indenter) and 1 (flat punch). Figure 22.15 shows how the contact depth depends on the value of ε. The contact depth at the peak load is, therefore:

$$\delta_c = \delta - \delta_s \,. \tag{22.17}$$

Using the load/displacement data from the unloading curve and (22.1), (22.2), (22.12), (22.14–22.17), the hardness and reduced elastic modulus for the test sample can be calculated. To find the elastic modulus of the sample, E_s, it is also necessary to know Poisson's ratio, ν_s, for the sample, as well as the elastic modulus, E_t, and Poisson's ratio, ν_t, of the indenter tip. For diamond these are 1,141 GPa and 0.07, respectively.

There also remains the issue of calibrating the tip shape, or finding the values for C_j in (22.15). Knowing the exact expansion of $A_c(\delta_c)$ is vital if the values for E_s and H are to be accurate. Several methods for calibrating the tip shape have been used, including imaging the tip with an electron microscope, measuring the size of nanoindentations using SEM or TEM of negative replicas, and using scanning probes to examine either the tip itself or the nanoindentations made with the tip. There are strengths and weaknesses to each of these methods. In general, however, the accuracy and usefulness of the methods depends largely on how patient and rigorous the experimentalist is in performing the calibration.

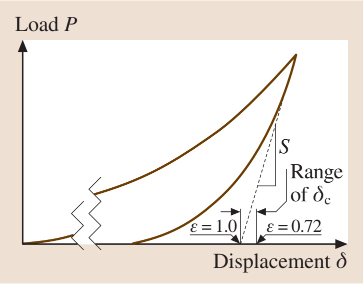

Fig. 22.15 Load-displacement curve showing how δ_c varies with ε

Because of the experimental difficulties and time involved in calibrating the tip shape by these methods, *Oliver* and *Pharr* [22.56] developed a method for calibration based on standard specimens. With a standard specimen that is mechanically isotropic and has a known E and H that does not vary with indentation depth, it should be possible to perform nanoindentations to a range of depths, and then use the analysis routines in reverse to deduce the tip area function, $A_c(\delta_c)$. In other words, if you perform a nanoindentation test, you can find the contact stiffness, S, at the peak load, P, and the contact depth, δ_c, from the unloading curve. Then if you know E a priori, (22.1) can be used to calculate the contact area, A, and, hence, you have a value for A_c at a depth δ_c. Repeating this procedure for a range of depths will give a numerical version of the function $A_c(\delta_c)$. Then, it is simply a case of fitting (22.15) to the numerical data. If the hardness, H, is known and not a function of depth, and the calibration specimen was fully plastic during testing, then essentially the same approach could be used but based on (22.12). Situations where a constant H is used to calibrate the tip are extremely rare.

In addition to the tip shape function, the machine compliance must be calibrated. Basic Newtonian mechanics tells us that for the tip to be pushed into a surface the tip must be pushing off of another body. During nanoindentation testing the other body is the machine frame. As a result, during a nanoindentation test it is not just the sample, but the machine frame that is being loaded. Consequently, a very small elastic deformation of the machine frame contributes to the total stiffness obtained from the unloading curve. The machine frame is usually very stiff, $> 10^6$ N/m, so the effect is only important at relatively large loads.

To calibrate the machine frame stiffness or compliance, large nanoindentations are made in a soft material such as aluminum with a known, isotropic elastic modulus. For very deep nanoindentations made with a Berkovich pyramid, the contact area, $A_c(\delta_c)$, can be reasonably approximated to $24.5\delta_c^2$, thus (22.1) can be used to find the expected contact stiffness for the material. Any difference between the expected value of S

and the value measured from the unloading curve will be due to the compliance of the machine frame. Performing a number of deep nanoindentations enables an accurate value for the machine frame compliance to be obtained.

Currently, because of its ready availability and predictable mechanical properties, the most popular calibration material is fused silica ($E = 72\,\text{GPa}$, $\nu = 0.17$), though aluminum is still used occasionally.

22.2.6 Modifications to the Analysis

Since the development of the analysis routines in the early 1990s, it has become apparent that the standard analysis of nanoindentation data is not applicable in all situations, usually because errors occur in the calculated contact depth or contact area. *Pharr* et al. [22.61–64] have used finite element modeling (FEM) to help understand and overcome the limitations of the standard analysis. Two important sources of errors have been identified in this way. The first is residual stress at the sample surface. The second is the change in the shape of nanoindents after elastic recovery.

The effect of residual stresses at a surface on the indentation properties has been the subject of debate for many years [22.65–67]. The perceived effect was that compressive stresses increased hardness, while tensile stresses decreased hardness. Using FEM it is possible to model a pointed nanoindenter being pushed into a model material that is in residual tension or compression. An FEM model of nanoindentation into aluminum alloy 8009 [22.61] has confirmed earlier experimental observations [22.68] indicating that the contact area calculated from the unloading curve is incorrect if there are residual stresses. In the FEM model of an aluminum alloy the mechanical behavior of the material is modeled using a stress-strain curve, which resembles that of an elastic-perfectly-plastic metal with a flow stress of 425.6 MPa. Yielding starts at 353.1 MPa and includes a small amount of work hardening. The FEM model was used to find the contact area directly and using the simulated unloading curve in conjunction with Oliver and Pharr's method. The results as a function of residual stress are illustrated in Fig. 22.16. Note that the differences between the two measured contact areas lead to miscalculations of E and H.

Errors in the calculated contact area stem from incorrect assumptions about the pile-up and sink-in at the edge of the contact, as illustrated by Fig. 22.17. The Oliver and Pharr analysis assumes the geometry of the sample surface is the same as that given by *Sneddon* [22.47] in his analytical model for the indentation of elastic surfaces. Clearly, for materials where there is significant plastic deformation, it is possible that there will be large deviations from the surface geometry found using Sneddon's elastic model. In reality, the error in the contact area depends on how much the geometry of the test sample surface differs from that of the calibration material (typically fused silica). It is possible that a test sample, even without a residual stress, will have a different surface geometry and, hence, contact area at a given depth, when compared to the calibration material. This is often seen for thin films on a substrate (e.g., *Tsui* et al. [22.69, 70]). Residual stresses increase the likelihood that the contact area calculated using Oliver and Pharr's method will be incorrect.

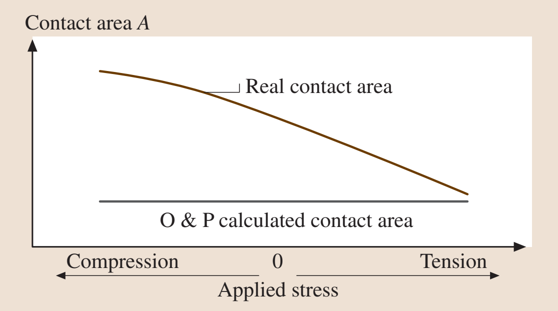

Fig. 22.16 When a surface is in a state of stress there is a significant difference between the contact area calculated using the Oliver and Pharr method and the actual contact area [22.61]. For an aluminum alloy this can lead to significant errors in the calculated hardness and elastic modulus

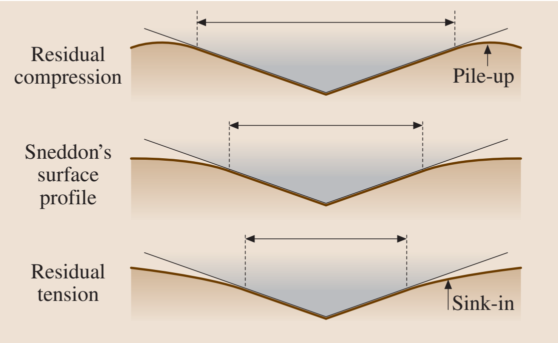

Fig. 22.17 Pile-up and sink-in are affected by residual stresses, and, hence, errors are introduced into standard Oliver and Pharr analysis

The issue of sink-in and pile-up is always a factor in nanoindentation testing. However, there is still no effective way to deal with these phenomena other than reverting to imaging of the indentations to identify the true contact area. Even this is difficult, as the edge of an indentation is not easy to identify using AFM or electron microscopy. One approach that has been used [22.71] with some success is measuring the ratio E_r^2/H, rather than E_r and H separately. Because E_r is proportional to $1/\sqrt{A}$ and H is proportional to $1/A$, E_r^2/H should be independent of A and, hence, unaffected by pile-up or sink-in. While this does not provide quantitative values for mechanical properties, it does provide a way to identify any variations in mechanical properties with indentation depth or between similar samples with different residual stresses.

Another source of error in the Oliver and Pharr analysis is due to incorrect assumptions about the nanoindentation geometry after unloading [22.63]. Once again, this is due to differences between the test sample and the calibration material. The exact shape of an unloaded nanoindentation on a material exhibiting elastic recovery is not simply an impression of the tip shape; rather, there is some elastic recovery of the nanoindentation sides giving them a slightly convex shape (see Fig. 22.18). The shape actually depends on Poisson's ratio, so the standard Oliver and Pharr analysis will only be valid for a material where $\nu = 0.17$, the value for fused silica, assuming it is used for the calibration.

To deal with the variations in the recovered nanoindentation shape, it has been suggested [22.63] that a modified nanoindenter geometry with a slightly concave side be used in the analysis (see Fig. 22.18). This requires a modification to (22.1):

$$S = \gamma 2E_r\sqrt{\frac{A}{\pi}}\,, \tag{22.18}$$

where γ is a correction term dependent on the tip geometry. For a Berkovich pyramid the best value is:

$$\gamma = \frac{\frac{\pi}{4} + 0.15483073 \cot \Phi \left(\frac{(1-2\nu_s)}{4(1-\nu_s)}\right)}{\left[\frac{\pi}{2} - 0.83119312 \cot \Phi \left(\frac{(1-2\nu_s)}{4(1-\nu_s)}\right)\right]^2}\,, \tag{22.19}$$

where $\Phi = 70.32°$. For a cube corner the correction can be even larger and γ is given by:

$$\gamma = 1 + \left(\frac{(1-2\nu_s)}{4(1-\nu_s)\tan\Phi}\right)\,, \tag{22.20}$$

where $\Phi = 42.28°$. Figure 22.19 shows how the modified contact area varies with depth for a real diamond Berkovich pyramid.

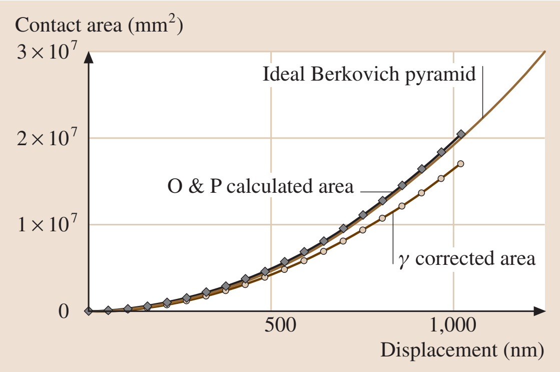

Fig. 22.19 For a real Berkovich tip the γ corrected area [22.63] is less at a given depth than the area calculated using the Oliver and Pharr method

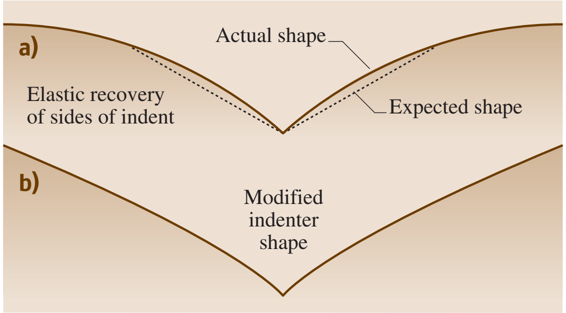

Fig. 22.18 **(a)** *Hay* et al. [22.63] found from experiments and FEM simulations that the actual shape of an indentation after unloading is not as expected. **(b)** They introduced a γ term to correct for this effect. This assumes the indenter has slightly concave sides

The validity of the γ-modified geometry is questionable from the perspective of contact mechanics since it relies on assuming an incorrect geometry for the nanoindenter tip to correct for an error in the geometry of the nanoindentation impression. The values for E and H obtained using the γ-modification are, however, good and can be significantly different to the values obtained with the standard Oliver and Pharr analysis.

22.2.7 Alternative Methods of Analysis

All of the preceding discussion on the analysis of nanoindentation curves has focused on the unloading curve, virtually ignoring the loading curve data. This is for the simple reason that the unloading curve can in many cases be regarded as purely elastic, whereas the shape of the loading curve is determined by a complex mix of elastic and plastic properties.

It is clear that there is substantially more data in the loading curve if it can be extracted. *Page* et al. [22.60, 72] have explored the possibility of curve fitting to the loading data using a combination of elastic and plastic properties. By a combination of analysis and empirical fitting to experimental data, it was suggested that the loading curve is of the following form:

$$P = E\left(\psi\sqrt{\frac{H}{E}}+\phi\sqrt{\frac{E}{H}}\right)^{-2}\delta^2\,, \qquad (22.21)$$

where ψ and ϕ are determined experimentally to be 0.930 and 0.194, respectively. For homogenous samples this equation gives a linear relationship between P and δ^2. Coatings, thin film systems, and samples that strain-harden can give significant deviations from linearity. Analysis of the loading curve has yet to gain popularity as a standard method for examining nanoindentation data, but it should certainly be regarded as a prime area for further investigation.

Another alternative method of analysis is based on the work involved in making an indentation. In essence, the nanoindentation curve is a plot of force against distance indicating integration under the loading curve will give the total work of indentation, or the sum of the elastic strain energy and the plastic work of indentation. Integrating under the unloading curve should give only the elastic strain energy. Thus, the work involved in both elastic and plastic deformation during nanoindentation can be found. *Cheng* and *Cheng* [22.73] combined measurements of the work of indentation with a dimensional analysis that deals with the effects of scaling in a material that work-hardens to estimate H/E_r. They subsequently evaluated H and E using the Oliver and Pharr approach to find the contact area.

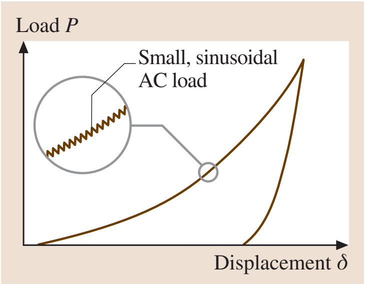

Fig. 22.20 A small AC load can be added to the DC load. This enables the contact stiffness, S, to be calculated throughout the indentation cycle

22.2.8 Measuring Contact Stiffness

As discussed earlier, it is possible to add a small AC load on top of the DC load used during nanoindentation testing, providing a way to measure the contact stiffness throughout the entire loading and unloading cycle [22.74, 75]. The AC load is typically at a frequency of ≈ 60 Hz and creates a dynamic system, with the sample acting as a spring with stiffness S (the contact stiffness), and the nanoindentation system acting as a series of springs and dampers. Figure 22.20 illustrates how the small AC load is added to the DC load. Figure 22.21 shows how the resulting dynamic system can be modeled. An analysis of the dynamic system gives the following relationships for S based on the amplitude of the AC displacement oscillation and the phase difference between the AC load and displacement signals:

$$\left|\frac{P_{\text{os}}}{\delta(\omega)}\right| = \sqrt{\left[\left(S^{-1}+C_{\text{f}}\right)^{-1}+K_{\text{s}}-m\omega^2\right]^2+\omega^2D^2}\,, \qquad (22.22)$$

$$\tan(\chi) = \frac{\omega D}{\left(S^{-1}+C_{\text{f}}\right)^{-1}+K_{\text{s}}-m\omega^2}\,, \qquad (22.23)$$

where C_f is the load frame compliance (the reciprocal of the load frame stiffness), K_s is the stiffness of the support springs (typically in the region of 50–100 N/m), D is the damping coefficient, P_os is the magnitude of the load oscillation, $\delta(\omega)$ is the magnitude of the displacement oscillation, ω is the oscillation frequency, m is the mass of the indenter, and χ is the phase angle between the force and the displacement.

In order to find S using either (22.22) or (22.23), it is necessary to calibrate the dynamic response of the system when the tip is not in contact with a sample ($S^{-1} = 0$). This calibration combined with the standard

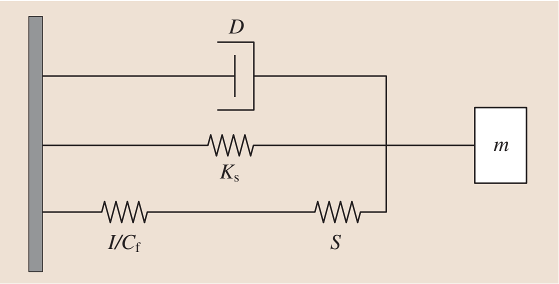

Fig. 22.21 The dynamic model used in the analysis of the AC response of a nanoindentation device

DC calibrations will provide the values for all of the constants in the two equations. All that needs to be measured in order to obtain S is either $\delta(\omega)$ or χ, both of which are measured by the lock-in amplifier used to generate the AC signal. Since the S obtained is the same as the S in (22.1), it follows that the Oliver and Pharr analysis can be applied to obtain E_r and H throughout the entire nanoindentation cycle.

The dynamic analysis detailed here was developed for the MTS Nanoindenter™ (Oakridge, Tennessee), but a similar analysis has been applied to other commercial instruments such as the Hysitron Triboscope™ (Minneapolis, Minnesota) [22.76]. For all instruments, an AC oscillation is used in addition to the DC voltage, and a dynamic model is used to analyze the response.

22.2.9 Measuring Viscoelasticity

Using an AC oscillation in addition to the DC load introduces the possibility of measuring viscoelastic properties during nanoindentation testing. This has recently been the subject of considerable interest with researchers looking at the loss modulus, storage modulus, and loss tangent of various polymeric materials [22.25, 77]. Recording the displacement response to the AC force oscillation enables the complex modulus (including the loss and storage modulus) to be found. If the modulus is complex, it is clear from (22.1) that the stiffness also becomes complex. In fact, the stiffness will have two components: S', the component in phase with the AC force and S'', the component out of phase with the AC force.

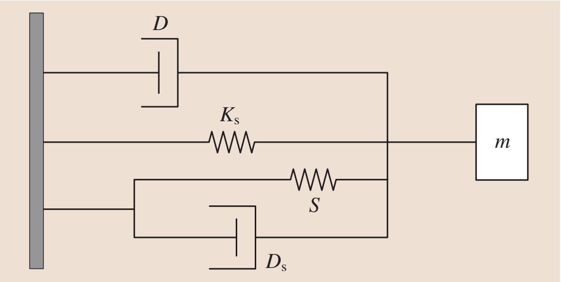

Fig. 22.22 The simplified dynamic model used when the sample is viscoelastic. It is assumed that the load frame compliance is negligible

The dynamic model illustrated in Fig. 22.21 is no longer appropriate for this situation, as the contact on the test sample also includes a damping term, shown in Fig. 22.22. Equations (22.22) and (22.23) must also be revised. Neglecting the load frame compliance, C_f, which in most real situations is negligible, (22.22) and (22.23) when the sample damping, D_s, is included become:

$$\left|\frac{P_\mathrm{os}}{\delta(\omega)}\right| = \sqrt{\left\{S+K_\mathrm{s}-m\omega^2\right\}^2+\omega^2\left(D+D_\mathrm{s}\right)^2}\,, \tag{22.24}$$

$$\tan(\chi) = \frac{\omega\left(D+D_\mathrm{s}\right)}{S+K_\mathrm{s}-m\omega^2}\,. \tag{22.25}$$

In order to find the loss modulus and storage modulus, (22.1) is used to relate S' (storage component) and S'' (loss component) to the complex modulus.

This method for measuring viscoelastic properties using nanoindentation has now been proven in principal, but has still only been applied to a very small range of polymers and remains an area of future growth.

22.3 Modes of Deformation

As described earlier, the analysis of nanoindentation data is based firmly on the results of elastic continuum mechanics. In reality, this idealized, purely elastic situation rarely occurs. For very shallow contacts on metals with thin surface films such as oxides, carbon layers, or organic layers [22.78, 79], the contact can initially be very similar to that modeled by Hertz and, later, Sneddon. It is very important to realize that this in itself does not constitute proof that the contact is purely elastic, because in many cases a small number of defects are present. These may be preexisting defects that move in the strain-field generated beneath the contact. Alternatively, defects can be generated either when the contact is first made or during the initial loading [22.52, 80]. When defects such as short lengths of dislocation are present the curves may still appear to be elastic even though inelastic processes like dislocation glide and cross-slip are taking place.

22.3.1 Defect Nucleation

Nucleation of defects during nanoindentation testing has been the subject of many experimental [22.81, 82] and theoretical studies [22.83, 84]. This is probably because nanoindentation is seen as a way to deform a small, defect-free volume of material to its elastic limit and beyond in a highly controlled geometry. There are, unfortunately, problems in comparing experimental results with theoretical predictions, largely because the kinetic processes involved in defect nucleation are difficult to model. Simulations conducted at 0 K do not permit kinetic processes, and molecular dynamics simulations are too fast (nanoseconds or picoseconds). Real nanoindentation experiments take place at ≈ 293 K and last for seconds or even minutes.

Kinetic effects appear in many forms, for instance, during the initial contact between the indenter tip and the surface when defects can be generated by the combined action of the impact velocity and surface forces [22.51]. A second example of a kinetic effect occurs during hold cycles at large loads when what appears to be an elastic contact can suddenly exhibit a large discontinuity in the displacement data [22.80]. Figure 22.23 shows how these kinetic effects can affect the nanoindentation data and the apparent yield point load.

During the initial formation of a contact, the deformation of surface asperities [22.51] and ledges [22.85] can create either point defects or short lengths of dislocation line. During the subsequent loading, the defects can help in the nucleation and multiplication of dislocations. The large strains present in the region surrounding the contact, coupled with the existence of defects generated on contact, can result in the extremely rapid multiplication of dislocations and, hence, pronounced discontinuities in the load-displacement curve. It is important to realize that the discontinuities are due to the rapid multiplication of dislocations, which may or may not occur at the same time that the first dislocation is nucleated. Dislocations may have been present for some time with the discontinuity only occurring when the existing defects are configured appropriately, as a Frank-Read source, for instance. Even under large strains, the time taken for a dislocation source to form from preexisting defects may be long. It is, therefore, not surprising that large discontinuities can be seen during hold cycles or unloading.

The generation of defects at the surface and the initiation of yielding is a complex process that is extremely dependent on surface asperities and surface forces. These, in turn, are closely related to the surface chemistry. It is not only the magnitude of surface forces, but also their range in comparison to the height of surface asperities that determines whether defects are generated on contact. Small changes in the surface chemistry or the velocity of the indenter tip when it first contacts the surface, can cause a transition from a situation in which defects are generated on contact to one where the contact is purely elastic [22.52].

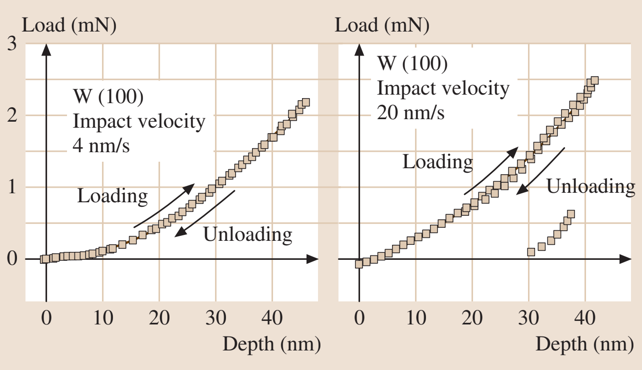

Fig. 22.23a,b Load-displacement curves for W(100) showing how changes in the impact velocity can cause a transition from perfectly elastic behavior to yielding during unloading

When the generation of defects during the initial contact is avoided and the deformed region under the contact is truly defect free, then the yielding of the sample should occur at the yield stress of a perfect crystal lattice. The load at which plastic deformation commences under these circumstances becomes very reproducible [22.86]. Unfortunately, nanoindenter tips on the near-atomic scale are not perfectly smooth or axisymmetric. As a result, accurately measuring the yield stress is very difficult. In fact, a slight rotation in the plane of the surface of either the sample or the tip can give a substantial change in the observed yield point load. Coating the surface in a cushioning self-assembled monolayer [22.87] can alleviate some of these variations, but it also introduces a large uncertainty in the contact area. Surface oxide layers, which may be several nanometers thick, have also been found to enhance the elastic behavior seen for very shallow nanoindentations on metallic surfaces [22.78]. Removal of the oxide has been shown to alleviate the initial elastic response.

While nanoindentation testing is ideal for examining the mechanical properties of defect-free volumes and looking at the generation of defects in perfect crystal lattices, it should be clear from the preceding discussion that great care must be taken in examining how the surface properties and the loading rate affect the results, particularly when comparisons are being made to theoretical models for defect generation.

22.3.2 Variations with Depth

Ideal elastic-plastic behavior, as described by *Tabor* [22.48], can be seen during indentation testing, provided the sample has been work-hardened so that the flow stress is a constant. However, it is often the case that the mechanical properties appear to change as the load (or depth) is increased. This apparent change can be a result of several processes, including work-hardening during the test. This is a particularly important effect for soft metals like copper. These metals usually have a high hardness at shallow depths, but it decreases asymptotically with increasing indentation depth to a hardness value that may be less than half that observed at shallow depths. This type of behavior is due to the increasing density of geometrically necessary dislocations at shallow depths [22.88]. Hence the effects of work-hardening are most pronounced at shallow depths. For hard materials the effect is less obvious.

Work-hardening is one of the factors that contribute to the so-called indentation size effect (ISE), whereby at shallow indentation depths the material appears to be harder. The ISE has been widely observed during microindentation testing, with at least part of the effect appearing to result from the increased difficulty in optically measuring the area of an indentation when it is small. During nanoindentation testing the ISE can also be observed, but it is often due to the tip area function, $A_c(\delta_c)$, being incorrectly calibrated. However, there are physical reasons other than work-hardening for expecting an increase in mechanical strength in small volumes. As described in the previous section, small volumes of crystalline materials can have either no defects or only a small number of defects present, making plastic yielding more difficult. Also, because of dislocation line tension, the shear stress required to make a dislocation bow out increases as the radius of the bow decreases. Thus, the shear stress needed to make a dislocation bow out in a small volume is greater than it is in a large volume. These physical reasons for small volumes appearing stronger than large volumes are particularly important in thin film systems, as will be discussed later. Note, however, that these physical reasons for increased hardness do not apply for an amorphous material such as fused silica, which partially explains its value as a calibration material.

22.3.3 Anisotropic Materials

The analysis methods detailed earlier were concerned primarily with the interpretation of data from nanoindentations in isotropic materials where the elastic modulus is assumed to be either independent of direction or a polycrystalline average of a material's elastic constants. Many crystalline materials exhibit considerable anisotropy in their elastic constants, hence, these analysis techniques may not always be appropriate. The theoretical problem of a rigid indenter pressed into an elastic, anisotropic half-space has been considered by *Vlassak* and *Nix* [22.89]. Their aim was to identify the feasibility of interpreting data from a depth-sensing indentation apparatus for samples with elastic constants that are anisotropic. Nanoindentation experiments [22.90] have shown the validity of the elastic analysis for crystalline zinc, copper, and beta-brass. The observed indentation modulus for zinc, as predicted, varied by as much as a factor of 2 between different orientations. The variations in the observed hardness values for the same materials were smaller, with a maximum variation with orientation of 20% detected in zinc. While these variations are clearly detectable with nanoindentation techniques, the variations are small in comparison to the actual anisotropy of the test material's elastic properties. This is because the indentation modulus is a weighted average of the stiffness in all directions.

At this time the effects of anisotropy on the hardness measured using nanoindentation have not been fully explored. For materials with many active slip planes it is likely that the small anisotropy observed by *Vlassak* and *Nix* is correct once plastic flow has been initiated. It is possible, however, that for defect-free crystalline specimens with a limited number of active slip planes that very shallow nanoindentations may show a much larger anisotropy in the observed hardness and initial yield point load.

22.3.4 Fracture and Delamination

Indentation testing has been widely used to study fracture in brittle materials [22.91], but the lower loads and smaller deformation regions of nanoindentation tests make it harder to initiate cracks and, hence, less useful as a way to evaluate fracture toughness. To overcome these problems the cube corner geometry, which generates larger shear stresses than the Berkovich pyramid, has been used with nanoindentation testing to study fracture [22.92]. These studies have had mixed success, because the cube corner geometry blunts very quickly when used on hard materials. In many cases, brittle materials are very hard.

Depth sensing indentation is better suited to studying delamination of thin films. Recent work extends the research conducted by *Marshall* et al. [22.93, 94], who examined the deformation of residually stressed films by indentation. A schematic of their analysis is given by Fig. 22.24. Their indentations were several microns deep, but the basic analysis is valid for nanoindentations. The analysis has been extended to multilayers [22.95], which is important since it enables a quantitative assessment of adhesion energy when an additional stressed film has been deposited on top of the film and substrate of interest. The additional film limits the plastic deformation of the film of interest and also applies extra stress that aids in the delamination. After indentation, the area of the delaminated film is measured optically or with an AFM to assess the extent of the delamination. This measurement, coupled with the load-displacement data, enables quantitative assessment of the adhesion energy to be made for metals [22.96] and polymers [22.97].

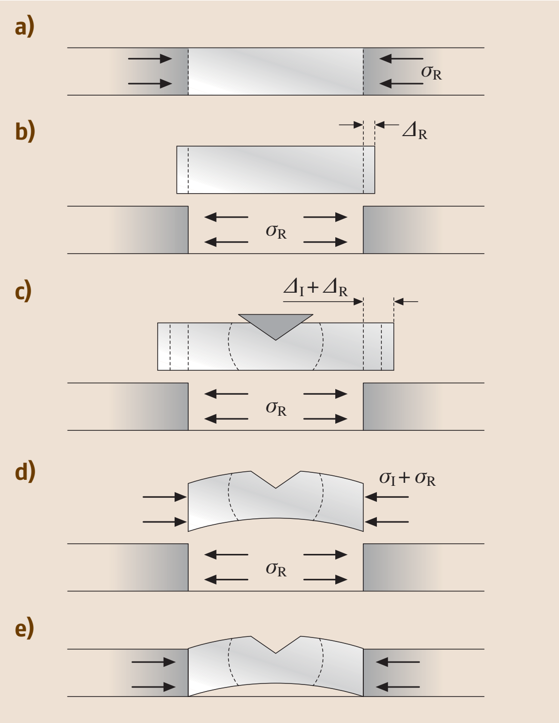

Fig. 22.24a–e To model delamination *Kriese* et al. [22.95] adapted the model developed by *Marshall* and *Evans* [22.93]. The model considers a segment of removed stressed film that is allowed to expand and then indented, thereby expanding it further. Replacing the segment in its original position requires an additional stress, and the segment bulges upwards

22.3.5 Phase Transformations

The pressure applied to the surface of a material during indentation testing can be very high. Equation (22.10) indicates that the pressure during plastic yielding is about three times the yield stress. For many materials, high hydrostatic pressures can cause phase transformations, and provided the transformation pressure is less than the pressure required to cause plastic yielding, it is possible during indentation testing to induce a phase transformation. This was first reported for silicon [22.98], but it has also been speculated [22.99] that many other materials may show the same effects. Most studies still focus on silicon because of its enormous technological importance, although there is some evidence that germanium also undergoes a phase transformation during the nanoindentation testing [22.100].

Recent results [22.21, 22, 41, 101, 102] indicate that there are actually multiple phase transformations during the nanoindentation of silicon. TEM of nanoindenta-

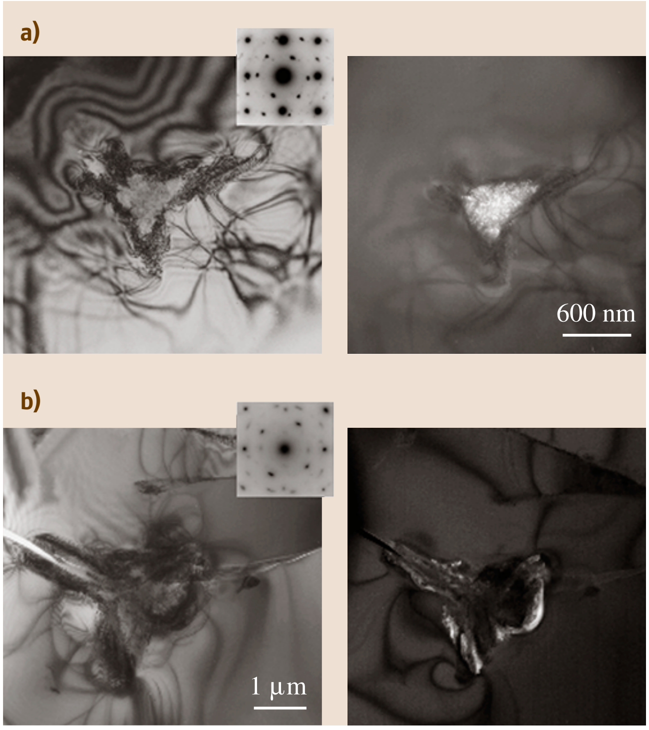

Fig. 22.25a,b Bright-field and dark-field TEM of **(a)** small and **(b)** large nanoindentations in Si. In small nanoindents the metastable phase BC-8 is seen in the center, but for large nanoindents BC-8 is confined to the edge of the indent, while the center is amorphous

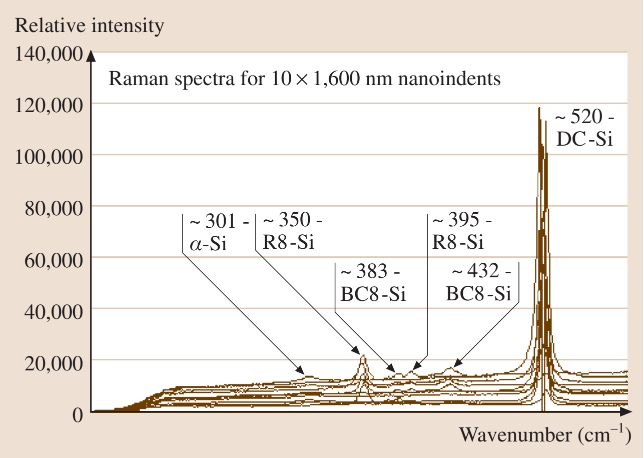

Fig. 22.26 Micro-Raman generally shows the BC-8 and R-8 Si phases that are at the edge of the nanoindents, but the amorphous phase in the center is not easy to detect, as it is often subsurface and the Raman peak is broad

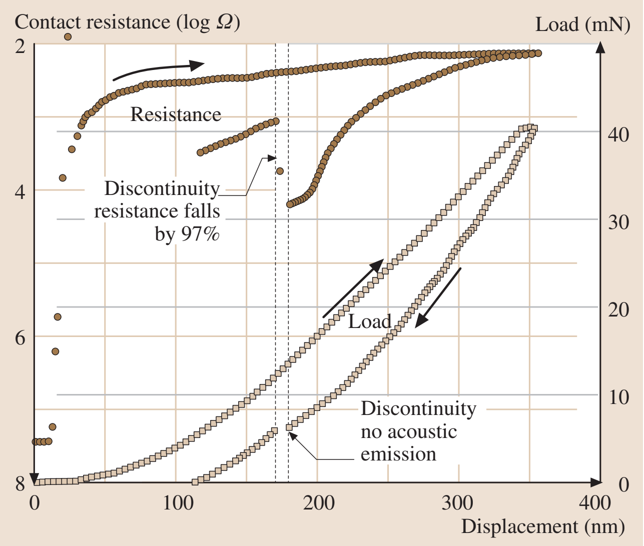

Fig. 22.27 Nanoindentation curves for deep indents on Si show a discontinuity during unloading and simultaneously a large drop in contact resistance

tions in diamond cubic silicon have shown the presence of amorphous-Si and the body-centered cubic BC-8 phase (see Fig. 22.25). Micro-Raman spectroscopy has indicated the presence of a further phase, the rhombhedral R-8 (see Fig. 22.26). For many nanoindentations on silicon there is a characteristic discontinuity in the unloading curve (see Fig. 22.27), which seems to correlate with a phase transformation. The exact sequence in which the phases form is still highly controversial with some [22.42], suggesting that the sequence during loading and unloading is:

Increasing load →
Diamond cubic Si → β-Sn Si

← Decreasing load
BC-8 Si and R-8 Si ← β-Sn Si

Other groups [22.21,22] suggest that the above sequence is only valid for shallow nanoindentations that do not exhibit an unloading discontinuity. For large nanoindentations that show an unloading discontinuity, they suggest the sequence will be:

Increasing load →
Diamond cubic Si → β-Sn Si

← Decreasing load
α-Si ← BC-8 Si and R-8 Si ← β-Sn Si

The disagreement is over the origin of the unloading discontinuity. *Mann* et al. [22.21] suggest it is due to the formation of amorphous silicon, while *Gogotsi* et al. [22.42] believe it is the β-Sn Si to BC-8 or R-8 transformation. Mann et al. argue that the high contact resistance before the discontinuity and the low contact resistance afterwards rule out the discontinuity being the metallic β-Sn Si transforming to the more resistive BC-8 or R-8. The counterargument is that amorphous Si is only seen with micro-Raman spectroscopy when the unloading is very rapid or there is a large nanoindentation with no unloading discontinuity. The importance of unloading rate and cracking in determining the phases present are further complications. The controversy will remain until in situ characterization of the phases present is undertaken.

22.4 Thin Films and Multilayers

In almost all real applications, surfaces are coated with thin films. These may be intentionally added such as hard carbide coatings on a tool bit, or they may simply be native films such as an oxide layer. It is also likely that there will be adsorbed films of water and organic contaminants that can range from a single molecule in thickness up to several nanometers. All of these films, whether native or intentionally placed on the surface, will affect the surface's mechanical behavior on the nanoscale. Adsorbates can have a significant impact on the surface forces [22.3] and, hence, the geometry and stability of asperity contacts. Oxide films can have dramatically different mechanical properties to the bulk and will also modify the surface forces. Some of the effects of native films have been detailed in the earlier sections on dislocation nucleation and adhesive contacts.

The importance of thin films in enhancing the mechanical behavior of surfaces is illustrated by the abundance of publications on thin film mechanical properties (see for instance *Nix* [22.88] or *Cammarata* [22.31] or *Was* and *Foecke* [22.103]). In the following sections, the mechanical properties of films intentionally deposited on the surface will be discussed.

22.4.1 Thin Films

Measuring the mechanical properties of a single thin surface film has always been difficult. Any measurement performed on the whole sample will inevitably be dominated by the bulk substrate. Nanoindentation, since it looks at the mechanical properties of a very small region close to the surface, offers a possible solution to the problem of measuring thin film mechanical properties. However, there are certain inherent problems in using nanoindentation testing to examine the properties of thin films. The problems stem in part from the presence of an interface between the film and substrate. The quality of the interface can be affected by many variables, resulting in a range of effects on the apparent elastic and plastic properties of the film. In particular, when the deformation region around the indent approaches the interface, the indentation curve may exhibit features due to the thin film, the bulk, the interface, or a combination of all three. As a direct consequence of these complications, models for thin-film behavior must attempt to take into account not only the properties of the film and substrate, but also the interface between them.

If, initially, the effect of the interface is neglected, it is possible to divide thin-coated systems into a number of categories that depend on the values of E (elastic modulus) and Y (the yield stress) of the film and substrate. These categories are typically [22.104, 105]:

1. coatings with high E and high Y, substrates with high E and high Y;
2. coatings with high E and high Y, substrates with high or low E and low Y;
3. coatings with high or low E and low Y, substrates with high E and high Y;
4. coatings with high or low E and low Y, substrates with high or low E and low Y.

The reasons for splitting thin film systems into these different categories have been amply demonstrated experimentally by *Whitehead* and *Page* [22.104, 105] and theoretically by *Fabes* et al. [22.106]. Essentially, hard, elastic materials (high E and Y) will possess smaller plastic zones than soft, inelastic (low E and Y) materials. Thus, when different combinations of materials are used as film and substrate, the overall plastic zone will differ significantly. In some cases, the plasticity is confined to the film, and in other cases, it is in both the film and substrate, as shown by Fig. 22.28. If the standard nanoindentation analysis routines are to be used, it is essential that the plastic zone and the elastic strain field are

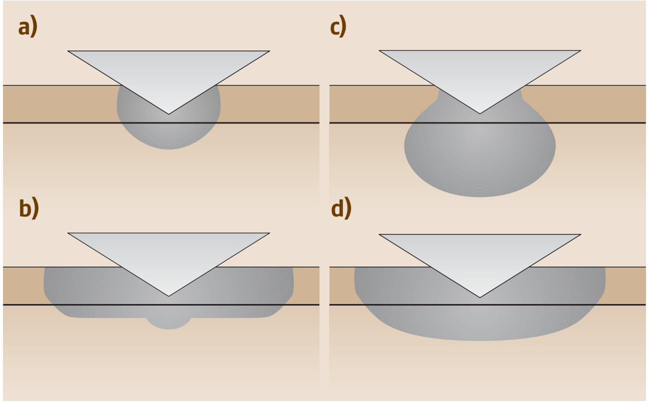

Fig. 22.28a–d Variations in the plastic zone for indents on films and substrates of different properties. **(a)** Film and substrate have high E and Y, **(b)** film has a high E and Y, substrate has a high or low E and low Y, **(c)** film has a high or low E and a low Y, and substrate has a high E and Y, **(d)** film has a high or low E and a low Y, and substrate has a high or low E and a low Y

both confined to the film and do not reach the substrate. Clearly, this is difficult to achieve unless extremely shallow nanoindentations are used. There is an often quoted 10% rule, that says nanoindents in a film must have a depth of less than 10% of the film's thickness if only the film properties are to be measured. This has no real validity [22.107]. There are film/substrate combinations for which 10% is very conservative, while for other combinations even 5% may be too deep. The effect of the substrate for different combinations of film and substrate properties has been studied using FEM [22.108], which has shown that the maximum nanoindentation depth to measure film only properties decreases in moving from soft on hard to hard on soft combinations. For a very soft film on a hard substrate, nanoindentations of 50% of the film thickness are alright, but this drops to $< 10\%$ for a hard film on a soft substrate. For a very strong film on a soft substrate, the surface film behaves like an elastic membrane or a bending plate.

Theoretical analysis of thin-film mechanical behavior is difficult. One theoretical approach that has been adopted uses the volumes of plastically deformed material in the film and substrate to predict the overall hardness of the system. However, it should be noted that this method is only really appropriate for soft coatings and indentation depths below the thickness of the coating (see cases c and d of Fig. 22.28), otherwise the behavior will be closer to that detailed later and shown by Fig. 22.29.

The technique of combining the mechanical properties of the film and substrate to evaluate the overall hardness of the system is generally referred to as the rule of mixtures. It stems from work by *Burnett* et al. [22.109–111] and *Sargent* [22.112], who derived a weighted average to relate the "composite" hardness (H) to the volumes of plastically deformed material in the film (V_f) and substrate (V_s) and their respective values of hardness, H_f and H_s. Thus,

$$H = \frac{H_f V_f + H_s V_s}{V_{total}} , \tag{22.26}$$

where V_{total} is $V_f + V_s$.

Equation (22.26) was further developed by *Burnett* and *Page* [22.109] to take into account the indentation size effect. They replaced H_s with $K\delta_c^{n-2}$, where K and n are experimentally determined constants dependent on the indenter and sample, and δ_c is the contact depth. This expression is derived directly from Meyer's law for spherical indentations, which gives the relationship $P = Kd^n$ between load, P, and the indentation dimension, d. Burnett and Page also employed a further refinement to enable the theory to fit experimental results from a specific sample, ion-implanted silicon. This particular modification essentially took into account the different sizes of the plastic zones in the two materials by multiplying H_s by a dimensionless factor (V_s/V_{total}). While this seems to be a sensible approach, it is mostly empirical, and the physical justification for using this particular factor is not entirely clear. Later, *Burnett* and *Rickerby* [22.110, 111] took this idea further and tried to generalize the equations to take into account all of the possible scenarios. Thus, the following equations were suggested:

$$H = \frac{H_f \left(\Omega^3\right) V_f + H_s V_s}{V_{total}} , \tag{22.27}$$

$$H = \frac{H_f V_f + H_s \left(\Omega^3\right) V_s}{V_{total}} . \tag{22.28}$$

The first of these, (22.27), deals with the case of a soft film on a hard substrate, and the second, (22.28), with a hard film on a soft substrate. The Ω term expresses the variation of the total plastic zone from the ideal hemispherical shape. This was taken still further by *Bull* and *Rickerby* [22.113], who derived an approximation for Ω based on the film and substrate zone radii being related to their respective hardness and elastic modulii [22.114, 115]. Hence:

$$\Omega = (E_f H_s / E_s H_f)^l , \tag{22.29}$$

where l is determined empirically. E_f and H_f and E_s and H_s are the elastic modulus and hardness of the film and substrate, respectively.

Experimental data [22.116] indicate that the effect of the substrate on the elastic modulus of the film can be quite different than the effect on hardness, due to

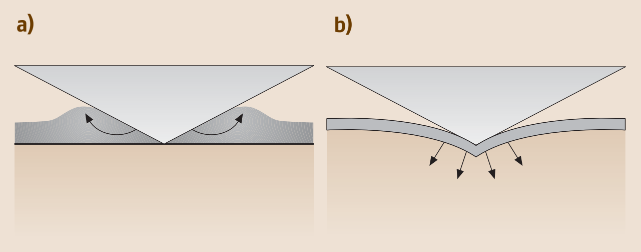

Fig. 22.29a,b Two different modes of deformation during nanoindentation of films. In **(a)** materials move upwards and outwards, while in **(b)** the film acts like a membrane and the substrate deforms

the zones of the elastic and plastic strain fields being different sizes.

Chechechin et al. [22.117] have recently studied the behavior of Al_2O_3 films of various thicknesses on different substrates. Their results indicate that many of the models correctly predict the transition between the properties of the film and those of the substrate, but do not always fit the observed hardness against depth curves. This group have also studied the pop-in behavior of Al_2O_3 films [22.118] and have attempted to model the range of loads and depths at which they occur via a Weibull-type distribution, as utilized in fracture analysis.

A point raised by Burnett and Rickerby should be emphasized. They state that there are two very distinct modes of deformation during nanoindentation testing. The first, referred to as *Tabor*'s [22.48] model for low Y/E materials, involves the buildup of material at the side of the indenter through movement of material at slip lines. The second, for materials with large Y/E does not result in surface pile-up. The displaced material is then accommodated by radial displacements [22.115]. The point is that a thin, strong, and well-bonded surface film can cause a substrate that would normally deform by Tabor's method to behave more like a material with high Y/E (see Fig. 22.29). It should be noted that this only applies as long as the film does not fail.

In recent theoretical and experimental studies the importance of material pile-up and sink-in has been investigated extensively. As discussed in an earlier section, pile-up can be increased by residual compressive stresses, but even in the absence of residual stresses pile-up can introduce a significant error in the calculated contact area. This is most pronounced in materials that do not work-harden [22.62]. For these materials using the Oliver and Pharr method fails to account for the pile-up and results in a large error in the values for E and H. For thin films *Tsui* et al. have used a focused ion beam to section through Knoop indentations in both soft films on hard substrates [22.69] and hard films on soft substrates [22.70]. The soft films, as expected, exhibit pile-up, while the hard film acts more like a membrane and the indentation exhibits sink-in with most of the plasticity in the substrate. Thus, there are three clearly identifiable factors affecting the pile-up and sink-in around nanoindents during testing of thin films:

1. Residual stresses
2. Degree of work-hardening
3. Ratio of film and substrate mechanical properties.

The bonding or adhesion between the film and substrate could also be added to this list. And it should not be forgotten that the depth of the nanoindentation relative to the film thickness also affects pile-up. For a very deep nanoindentation into a thin, soft film on a hard substrate pile-up is reduced, due to the combined constraints on the film of the tip and substrate [22.119]. Due to all of these complications, using nanoindentation to study thin film mechanical properties is fraught with danger. Many unprepared researchers have misguidedly taken the values of E and H obtained during nanoindentation testing to be absolute values only to find out later that the values contain significant errors.

Many of the problems associated with nanoindentation testing are related to incorrectly calculating the contact area, A. The *Joslin* and *Oliver* method [22.71] is one way that A can be removed from the calculations. This approach has been used with some success to look at strained epitaxial II/VI semiconductor films [22.120], but there is evidence that the lattice mismatch in these films can cause dramatic changes in the mechanical properties of the films [22.121]. This may be due to image forces and the film/substrate interface acting as a barrier to dislocation motion. Recently, it has been shown that using films and substrates with known matching elastic moduli, it is possible to use the assumption of constant elastic modulus with depth to evaluate H [22.122]. In effect, this is using (22.1) to evaluate A from the contact stiffness data, and then substituting the value for A into (22.12). The value of E is measured independently, for instance, using acoustic techniques.

22.4.2 Multilayers

Multilayered materials with individual layers that are a micron or less in thickness, sometimes referred to as superlattices, can exhibit substantial enhancements in hardness or strength. This should be distinguished from the super modulus effect discussed earlier, which has been shown to be largely an artifact. The enhancements in hardness can be as much as 100% when compared to the value expected from the rule of mixtures, which is essentially a weighted average of the hardness for the constituents of the two layers [22.123]. Table 22.1 shows how the properties of isostructural multilayers can show a substantial increase in hardness over that for fully interdiffused layers. The table also shows how there can be a substantial enhancement in hardness for non-isostructural multilayers compared to the values for the same materials when they are homogeneous.

Table 22.1 Results for some experimental studies of multilayer hardness

Study	Multilayer	Maximum hardness and multilayer repeat length	Reference hardness value	Range of hardness values for multilayers
Isostructural Knoop hardness [22.124]	Cu/Ni	524 at 11.6 nm	284 (interdiffused)	295–524
Non-isostructural Nanoindentation [22.125]	Mo/NbN	33 GPa at 2 nm	NbN – 17 GPa Mo – 2.7 GPa Wo – 7 GPa	12–33 GPa
	W/NbN	29 GPa at 3 nm	(individual layer materials)	23–29 GPa

There are many factors that contribute to enhanced hardness in multilayers. These can be summarized as [22.103]:

1. Hall-Petch behavior
2. Orowan strengthening
3. Image effects
4. Coherency and thermal stresses
5. Composition modulation

Hall-Petch behavior is related to dislocations piling-up at grain boundaries. (Note that pile-up is used to describe two distinct effects: One is material building up at the side of an indentation, the other is an accumulation of dislocations on a slip-plane.) The dislocation pile-up at grain boundaries impedes the motion of dislocations. For materials with a fine grain structure there are many grain boundaries, and, hence, dislocations find it hard to move. In polycrystalline multilayers, it is often the case that the size of the grains within a layer scales with the layer thickness so that reducing the layer thickness reduces the grain size. Thus, the Hall-Petch relationship (below) should be applicable to polycrystalline multilayer films with the grain size, d_g, replaced by the layer thickness.

$$Y = Y_0 + k_{HP}\, d_g^{-0.5} \,, \tag{22.30}$$

where Y is the enhanced yield stress, Y_0 is the yield stress for a single crystal, and k_{HP} is a constant.

There is an ongoing argument about whether Hall-Petch behavior really takes place in nanostructured multilayers. The basic model assumes many dislocations are present in the pile-up, but such large dislocation pile-ups are not seen in small grains [22.126] and are unlikely to be present in multilayers. As a direct consequence, studies have found a range of values, between 0 and -1, for the exponent in (22.30), rather than the -0.5 predicted for Hall-Petch behavior.

Orowan strengthening is due to dislocations in layered materials being effectively pinned at the interfaces. As a result, the dislocations are forced to bow out along the layers. In narrow films, dislocations are pinned at both the top and bottom interfaces of a layer and bow out parallel to the plane of the interface [22.127, 128]. Forcing a dislocation to bow out in a layered material requires an increase in the applied shear stress beyond that required to bow out a dislocation in a homogeneous sample. This additional shear stress would be expected to increase as the film thickness is reduced.

Image effects were suggested by *Koehler* [22.129] as a possible source of enhanced yield stress in multilayered materials. If two metals, A and B, are used to make a laminate and one of them, A, has a high dislocation line energy, but the other, B, has a low dislocation line energy, then there will be an increased resistance to dislocation motion due to image forces. However, if the individual layers are thick enough that there may be a dislocation source present within the layer, then dislocations could pile-up at the interface. This will create a local stress concentration point and the enhancement to the strength will be very limited. If the layers are thin enough that there will be no dislocation source present, the enhanced mechanical strength may be substantial. In Koehler's model only nearest neighbor layers were taken to contribute to the image forces. However, this was extended to include more layers [22.130] without substantial changes in the results. The consequence on image effects of reducing the thickness of the individual layers in a multilayer is that it prevents dislocation sources from being active within the layer.

For many multilayer systems there is an increase in strength as the bilayer repeat length is reduced, but there is often a critical repeat length (e.g., 3 nm for the W/NbN multilayer of Table 22.1) below which the strength falls. One explanation for the fall in strength involves the effects of coherency and thermal stresses on dislocation energy. Unlike image effects where the energy of dislocations are a maximum or minimum in the center of layers, the energy maxima and minima are at the interfaces for coherency stresses. Combining the effects of varying moduli and coherency stresses shows that the dependence of strength on layer thickness has a peak near the repeat period where coherency strains begin decreasing [22.131].

Another source of deviations in behavior at very small repeat periods is the imperfect nature of interfaces. With the exception of atomically perfect epitaxial films, interfaces are generally not atomically flat and there is some interdiffusion. For the Cu/Ni film of Table 22.1, the effects of interdiffusion on hardness were examined [22.124] by annealing the multilayers. The results were in agreement with a model by *Krzanowski* [22.132] that predicted the variations in hardness would be proportional to the amplitude of the composition modulation.

It is interesting to note that the explanations for enhanced mechanical properties in multilayered materials are all based on dislocation mechanisms. So it would seem natural to assume that multilayered materials that do not contain dislocations will show no enhanced hardness over their rule of mixtures values. This has been verified by studies on amorphous metal multilayers [22.133], which shows that the hardness of the multilayers, firstly, lies between that of the two individual materials and, secondly, has almost no variation with repeat period.

22.5 Developing Areas

Over the past 20 to 30 years, the driving force for studying nanomechanical behavior of surfaces and thin films has been largely, though not exclusively, the microelectronics industry. The importance of electronics to the modern world is only likely to grow in the foreseeable future, but other technological areas may overtake microelectronics as the driving force for research, including the broad fields of biomaterials and nanotechnology. In several places in this chapter, a number of developing areas have been mentioned. These include nanoscale measurements of viscoelasticity and the study of environmental effects (temperature and surface chemistry) on nanomechanical properties. Both of these topics will be vital in the study of biological systems and, as a result, will be increasingly important from a research point of view.

We still have a relatively rudimentary understanding of the nanomechanics of complex biological systems such as bone cells (osteoblasts and osteoclasts) and skin cells (fibroblasts), or even, for that matter, simpler biological structures such as dental enamel. For example, Fig. 22.30 shows how the mechanical properties of dental enamel can vary within a single tooth [22.134]. But this is still a relatively large-scale measure of mechanical behavior. The prismatic structure of enamel means that there are variations in mechanical properties on a range of scales from millimeters down to nanometers.

In terms of data analysis, there remains much to be done. If an analysis method can be developed that deals with the problems of pile-up and sink-in, the utility of nanoindentation testing will be greatly enhanced.

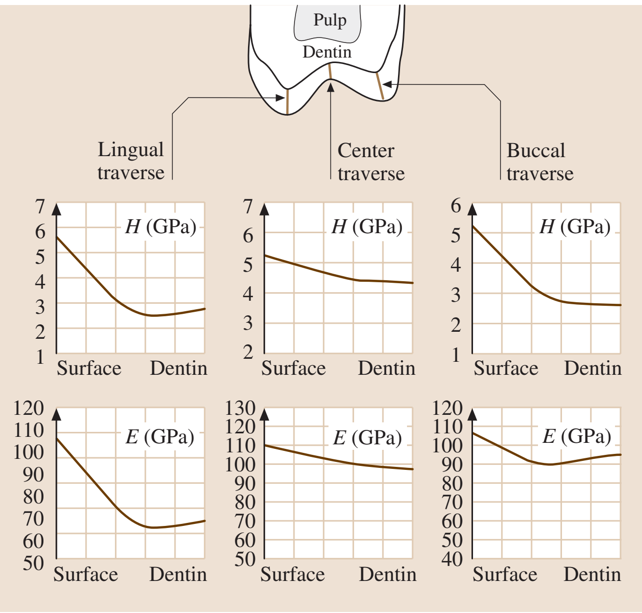

Fig. 22.30 Variations in E and H across human dental enamel. The sample is an upper 2nd molar cut in cross section from the lingual to the buccal side. Nanoindentations are performed across the surface to examine how the mechanical properties vary

References

22.1 G. Binnig, C. F. Quate, C. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

22.2 P. J. Blau, B. R. Lawn (Eds.): *Microindentation Techniques in Materials Science and Engineering* (ASTM, Pennsylvannia 1986)

22.3 J. N. Israelachvili: *Intermolecular and Surface Forces* (Academic, London 1992)

22.4 R. S. Bradley: The cohesive force between solid surfaces and the surface energy of solids, Philos. Mag. **13** (1932) 853–862

22.5 B. V. Derjaguin, V. M. Muller, Yu. P. Toporov: Effect of contact deformations on the adhesion of particles, J. Coll. Interface Sci. **53** (1975) 314–326

22.6 V. M. Muller, V. S. Yuschenko, B. V. Derjaguin: On the influence of molecular forces on the deformation of an elastic sphere and its sticking to a rigid plane, J. Coll. Interface Sci. **77** (1980) 91–101

22.7 V. M. Muller, B. V. Derjaguin, Yu. P. Toporov: On two methods of calculation of the force of sticking of an elastic sphere to a rigid plane, Coll. Surf. **7** (1983) 251–259

22.8 K. L. Johnson, K. Kendal, A. D. Roberts: Surface energy and the contact of elastic solids, Proc. R. Soc. A **324** (1971) 301–320

22.9 A. P. Ternovskii, V. P. Alekhin, M. Kh. Shorshorov, M. M. Khrushchov, V. N. Skvortsov: Zavod Lab. **39** (1973) 1242

22.10 S. I. Bulychev, V. P. Alekhin, M. Kh. Shorshorov, A. P. Ternovskii, G. D. Shnyrev: Determining Young's modulus from the indentor penetration diagram, Zavod Lab. **41** (1975) 1137

22.11 S. I. Bulychev, V. P. Alekhin, M. Kh. Shorshorov, A. P. Ternovskii: Mechanical properties of materials studied from kinetic diagrams of load versus depth of impression during microimpression, Prob. Prochn. **9** (1976) 79

22.12 S. I. Bulychev, V. P. Alekhin: Zavod Lab. **53** (1987) 76

22.13 M. Kh. Shorshorov, S. I. Bulychev, V. P. Alekhin: Sov. Phys. Doklady **26** (1982) 769

22.14 J. B. Pethica: Microhardness tests with penetration depths less than ion implanted layer thickness. In: *Ion Implantation into Metals*, ed. by V. Ashworth, W. Grant, R. Procter (Pergamon, Oxford 1982) p. 147

22.15 D. Newey, M. A. Wilkens, H. M. Pollock: An ultra-low-load penetration hardness tester, J. Phys. E: Sci. Instrum. **15** (1982) 119

22.16 D. Kendall, D. Tabor: An ultrasonic study of the area of contact between stationary and sliding surfaces, Proc. R. Soc. A **323** (1971) 321–340

22.17 G. M. Pharr, W. C. Oliver, F. R. Brotzen: On the generality of the relationship among contact stiffness, contact area and elastic-modulus during indentation, J. Mater. Res. **7** (1992) 613

22.18 T. P. Weihs, C. W. Lawrence, B. Derby, C. B. Scruby, J. B. Pethica: Acoustic emissions during indentation tests, MRS Symp. Proc. **239** (1992) 361–366

22.19 D. F. Bahr, J. W. Hoehn, N. R. Moody, W. W. Gerberich: Adhesion and acoustic emission analysis of failures in nitride films with 14 metal interlayer, Acta Mater. **45** (1997) 5163

22.20 D. F. Bahr, W. W. Gerberich: Relationships between acoustic emission signals and physical phemomena during indentation, J. Mat. Res. **13** (1998) 1065

22.21 A. B. Mann, D. van Heerden, J. B. Pethica, T. P. Weihs: Size-dependent phase transformations during point-loading of silicon, J. Mater. Res. **15** (2000) 1754

22.22 A. B. Mann, D. van Heerden, J. B. Pethica, P. Bowes, T. P. Weihs: Contact resistance and phase transformations during nanoindentation of silicon, Philos. Mag. A **82** (2002) 1921

22.23 S. Jeffery, C. J. Sofield, J. B. Pethica: The influence of mechanical stress on the dielectric breakdown field strength of SiO_2 films, Appl. Phys. Lett. **73** (1998) 172

22.24 B. N. Lucas, W. C. Oliver: Indentation power-law creep of high-purity indium, Metall. Trans. A **30** (1999) 601

22.25 S. A. Syed Asif: Time dependent micro deformation of materials. Ph.D. Thesis (Oxford Univ., Oxford 1997)

22.26 S. A. Syed Asif, R. J. Colton, K. J. Wahl: Nanoscale Surface Mechanical Property Measurements: Force Modulation Techniques Applied to Nanoindentation. In: *Interfacial Properties on the Submicron Scale*, ed. by J. Frommer, R. Overney (ACS Books, Whashington 2000)

22.27 A. B. Mann, J. B. Pethica: Nanoindentation studies in a liquid environment, Langmuir **12** (1996) 4583

22.28 J. W. Beams: Mechanical properties of thin films of gold and silver. In: *Structure and Properties of Thin Films*, ed. by C. A. Neugebauer, J. B. Newkirk, D. A. Vermilyea (Wiley, New York 1959) pp. 183–192

22.29 J. J. Vlassak, W. D. Nix: A new bulge test technique for the determination of Youngs modulus and Poissons ratio of thin-films, J. Mater. Res. **7** (1992) 3242

22.30 W. M. C. Yang, T. Tsakalakos, J. E. Hilliard: Enhanced elastic modulus in composition modulated gold-nickel and copper-palladium foils, J. Appl. Phys. **48** (1977) 876

22.31 R. C. Cammarata: Mechanical properties of nanocomposite thin-films, Thin Solid Films **240** (1994) 82

22.32 G. A. D. Briggs: *Acoustic Microscopy* (Clarendon, Oxford 1992)

22.33 J. Kushibiki, N. Chubachi: Material characterization by line-focus-beam acoustic microscope, IEEE Trans. Sonics Ultrasonics **32** (1985) 189–212

22.34 M. J. Bamber, K. E. Cooke, A. B. Mann, B. Derby: Accurate determination of Young's modulus and Poisson's ratio of thin films by a combination of acoustic microscopy and nanoindentation, Thin Solid Films **398** (2001) 299–305
22.35 S. E. Bobbin, R. C. Cammarata, J. W. Wagner: Determination of the flexural modulus of thin-films from measurement of the 1st arrival of the symmetrical Lamb wave, Appl. Phys. Lett. **59** (1991) 1544–1546
22.36 R. B. King, C. M. Fortunko: Determination of inplane residual-stress state in plates using horizontally polarized shear waves, J. Appl. Phys. **54** (1983) 3027–3035
22.37 R. F. Cook, G. M. Pharr: Direct observation and analysis of indentation cracking in glasses and ceramics, J. Am. Ceram. Soc. **73** (1990) 787–817
22.38 T. F. Page, W. C. Oliver, C. J. McHargue: The deformation-behavior of ceramic crystals subjected to very low load (nano)indentations, J. Mater. Res. **7** (1992) 450–473
22.39 G. M. Pharr, W. C. Oliver, D. S. Harding: New evidence for a pressure-induced phase-transformation during the indentation of silicon, J. Mater. Res. **6** (1991) 1129–1130
22.40 C. F. Robertson, M. C. Fivel: The study of submicron indent-induced plastic deformation, J. Mater. Res. **14** (1999) 2251–2258
22.41 J. E. Bradby, J. S. Williams, J. Wong-Leung, M. V. Swain, P. Munroe: Transmission electron microscopy observation of deformation microstructure under spherical indentation in silicon, Appl. Phys. Lett. **77** (2000) 3749–3751
22.42 Y. G. Gogotsi, V. Domnich, S. N. Dub, A. Kailer, K. G. Nickel: Cyclic nanoindentation and Raman microspectroscopy study of phase transformations in semiconductors, J. Mater. Res. **15** (2000) 871–879
22.43 H. Hertz: Über die Berührung fester elastischer Körper, J. reine angew. Math. **92** (1882) 156–171
22.44 J. Boussinesq: *Application des potentiels à l'étude de l'équilibre et du mouvement des solides élastiques* (Blanchard, Paris 1885) Reprint (1996)
22.45 A. E. H. Love: The stress produced in a semi-infinite solid by pressure on part of the boundary, Philos. Trans. R. Soc. **228** (1929) 377–420
22.46 A. E. H. Love: Boussinesq's problem for a rigid cone, Quarter. J. Math. **10** (1939) 161
22.47 I. N. Sneddon: The relationship between load and penetration in the axisymmetric Boussinesq problem for a punch of arbitrary profile, Int. J. Eng. Sci. **3** (1965) 47–57
22.48 D. Tabor: *Hardness of Metals* (Oxford Univ. Press, Oxford 1951)
22.49 R. von Mises: Mechanik der festen Körper in plastisch deformablen Zustand, Goettinger Nachr. Math.-Phys. **K1** (1913) 582–592
22.50 H. Tresca: Sur l'ecoulement des corps solids soumis s fortes pression, Compt. Rend. **59** (1864) 754
22.51 A. B. Mann, J. B. Pethica: The role of atomic-size asperities in the mechanical deformation of nanocontacts, Appl. Phys. Lett. **69** (1996) 907–909
22.52 A. B. Mann, J. B. Pethica: The effect of tip momentum on the contact stiffness and yielding during nanoindentation testing, Philos. Mag. A **79** (1999) 577–592
22.53 S. P. Jarvis: Atomic force microscopy and tip-surface interactions. Ph.D. Thesis (Oxford Univ., Oxford 1993)
22.54 J. B. Pethica, D. Tabor: Contact of characterised metal surfaces at very low loads: Deformation and adhesion, Surf. Sci. **89** (1979) 182
22.55 J. S. Field, M. V. Swain: Determining the mechanical-properties of small volumes of materials from submicrometer spherical indentations, J. Mater. Res. **10** (1995) 101–112
22.56 W. C. Oliver, G. M. Pharr: An improved technique for determining hardness and elastic-modulus using load and displacement sensing indentation experiments, J. Mater. Res. **7** (1992) 1564–1583
22.57 J. L. Loubet, J. M. Georges, O. Marchesini, G. Meille: Vickers indentation curves of magnesium oxide (MgO), Mech. Eng. **105** (1983) 91–92
22.58 J. L. Loubet, J. M. Georges, O. Marchesini, G. Meille: Vickers indentation curves of magnesium oxide (MgO), J. Tribol. Trans. ASME **106** (1984) 43–48
22.59 M. F. Doerner, W. D. Nix: A method for interpreting the data from depth sensing indentation experiments, J. Mater. Res. **1** (1986) 601–609
22.60 S. V. Hainsworth, H. W. Chandler, T. F. Page: Analysis of nanoindentation load-displacement loading curves, J. Mater. Res. **11** (1996) 1987–1995
22.61 A. Bolshakov, W. C. Oliver, G. M. Pharr: Influences of stress on the measurement of mechanical properties using nanoindentation. 2. Finite element simulations, J. Mater. Res. **11** (1996) 760–768
22.62 A. Bolshakov, G. M. Pharr: Influences of pileup on the measurement of mechanical properties by load and depth sensing instruments, J. Mater. Res. **13** (1998) 1049–1058
22.63 J. C. Hay, A. Bolshakov, G. M. Pharr: A critical examination of the fundamental relations used in the analysis of nanoindentation data, J. Mater. Res. **14** (1999) 2296–2305
22.64 G. M. Pharr, T. Y. Tsui, A. Bolshakov, W. C. Oliver: Effects of residual-stress on the measurement of hardness and elastic-modulus using nanoindentation, MRS Symp. Proc. **338** (1994) 127–134
22.65 T. R. Simes, S. G. Mellor, D. A. Hills: A note on the influence of residual-stress on measured hardness, J. Strain Anal. Eng. Des. **19** (1984) 135–137
22.66 W. R. Lafontaine, B. Yost, C. Y. Li: Effect of residual-stress and adhesion on the hardness of copper-films deposited on silicon, J. Mater. Res. **5** (1990) 776–783
22.67 W. R. Lafontaine, C. A. Paszkiet, M. A. Korhonen, C. Y. Li: Residual stress measurements of thin

aluminum metallizations by continuous indentation and X-ray stress measurement techniques, J. Mater. Res. **6** (1991) 2084–2090
22.68 T. Y. Tsui, W. C. Oliver, G. M. Pharr: Influences of stress on the measurement of mechanical properties using nanoindentation. 1. Experimental studies in an aluminum alloy, J. Mater. Res. **11** (1996) 752–759
22.69 T. Y. Tsui, J. Vlassak, W. D. Nix: Indentation plastic displacement field: Part I. The case of soft films on hard substrates, J. Mater. Res. **14** (1999) 2196–2203
22.70 T. Y. Tsui, J. Vlassak, W. D. Nix: Indentation plastic displacement field: Part II. The case of hard films on soft substrates, J. Mater. Res. **14** (1999) 2204–2209
22.71 D. L. Joslin, W. C. Oliver: A new method for analyzing data from continuous depth-sensing microindentation tests, J. Mater. Res. **5** (1990) 123–126
22.72 M. R. McGurk, T. F. Page: Using the P-delta(2) analysis to deconvolute the nanoindentation response of hard-coated systems, J. Mater. Res. **14** (1999) 2283–2295
22.73 Y. T. Cheng, C. M. Cheng: Relationships between hardness, elastic modulus, and the work of indentation, Appl. Phys. Lett. **73** (1998) 614–616
22.74 J. B. Pethica, W. C. Oliver: Mechanical properties of nanometer volumes of material: Use of the elastic response of small area indentations, MRS Symp. Proc. **130** (1989) 13–23
22.75 W. C. Oliver, J. B. Pethica: Method for continuous determination of the elastic stiffness of contact between two bodies, United States Patent Number 4,848,141, (1989)
22.76 S. A. S. Asif, K. J. Wahl, R. J. Colton: Nanoindentation and contact stiffness measurement using force modulation with a capacitive load-displacement transducer, Rev. Sci. Instrum. **70** (1999) 2408–2413
22.77 J. L. Loubet, W. C. Oliver, B. N. Lucas: Measurement of the loss tangent of low-density polyethylene with a nanoindentation technique, J. Mater. Res. **15** (2000) 1195–1198
22.78 W. W. Gerberich, J. C. Nelson, E. T. Lilleodden, P. Anderson, J. T. Wyrobek: Indentation induced dislocation nucleation: The initial yield point, Acta Mater. **44** (1996) 3585–3598
22.79 J. D. Kiely, J. E. Houston: Nanomechanical properties of Au(111), (001), and (110) surfaces, Phys. Rev. B **57** (1998) 12588–12594
22.80 D. F. Bahr, D. E. Wilson, D. A. Crowson: Energy considerations regarding yield points during indentation, J. Mater. Res. **14** (1999) 2269–2275
22.81 D. E. Kramer, K. B. Yoder, W. W. Gerberich: Surface constrained plasticity: Oxide rupture and the yield point process, Philos. Mag. A **81** (2001) 2033–2058
22.82 S. G. Corcoran, R. J. Colton, E. T. Lilleodden, W. W. Gerberich: Anomalous plastic deformation at surfaces: Nanoindentation of gold single crystals, Phys. Rev. B **55** (1997) 16057–16060
22.83 E. B. Tadmor, R. Miller, R. Phillips, M. Ortiz: Nanoindentation and incipient plasticity, J. Mater. Res. **14** (1999) 2233–2250
22.84 J. A. Zimmerman, C. L. Kelchner, P. A. Klein, J. C. Hamilton, S. M. Foiles: Surface step effects on nanoindentation, Phys. Rev. Lett. **87** (2001) article 165507 (1–4)
22.85 J. D. Kiely, R. Q. Hwang, J. E. Houston: Effect of surface steps on the plastic threshold in nanoindentation, Phys. Rev. Lett. **81** (1998) 4424–4427
22.86 A. B. Mann, P. C. Searson, J. B. Pethica, T. P. Weihs: The relationship between near-surface mechanical properties, loading rate and surface chemistry, Mater. Res. Soc. Symp. Proc. **505** (1998) 307–318
22.87 R. C. Thomas, J. E. Houston, T. A. Michalske, R. M. Crooks: The mechanical response of gold substrates passivated by self-assembling monolayer films, Science **259** (1993) 1883–1885
22.88 W. D. Nix: Elastic and plastic properties of thin films on substrates: Nanoindentation techniques, Mater. Sci. Eng. A **234** (1997) 37–44
22.89 J. J. Vlassak, W. D. Nix: Indentation modulus of elastically anisotropic half-spaces, Philos. Mag. A **67** (1993) 1045–1056
22.90 J. J. Vlassak, W. D. Nix: Measuring the elastic properties of anisotropic materials by means of indentation experiments, J. Mech. Phys. Solids **42** (1994) 1223–1245
22.91 B. R. Lawn: *Fracture of Brittle Solids* (Cambridge Univ. Press, Cambridge 1993)
22.92 G. M. Pharr: Measurement of mechanical properties by ultra-low load indentation, Mater. Sci. Eng. A **253** (1998) 151–159
22.93 D. B. Marshall, A. G. Evans: Measurement of adherence of residually stressed thin-films by indentation. 1. Mechanics of interface delamination, J. Appl. Phys. **56** (1984) 2632–2638
22.94 C. Rossington, A. G. Evans, D. B. Marshall, B. T. Khuriyakub: Measurement of adherence of residually stressed thin-films by indentation. 2. Experiments with ZnO/Si, J. Appl. Phys. **56** (1984) 2639–2644
22.95 M. D. Kriese, W. W. Gerberich, N. R. Moody: Quantitative adhesion measures of multilayer films: Part I. Indentation mechanics, J. Mater. Res. **14** (1999) 3007–3018
22.96 M. D. Kriese, W. W. Gerberich, N. R. Moody: Quantitative adhesion measures of multilayer films: Part II. Indentation of W/Cu, W/W, Cr/W, J. Mater. Res. **14** (1999) 3019–3026
22.97 M. Li, C. B. Carter, M. A. Hillmyer, W. W. Gerberich: Adhesion of polymer-inorganic interfaces by nanoindentation, J. Mater. Res. **16** (2001) 3378–3388
22.98 D. R. Clarke, M. C. Kroll, P. D. Kirchner, R. F. Cook, B. J. Hockey: Amorphization and conductivity of

silicon and germanium induced by indentation, Phys. Rev. Lett. **60** (1988) 2156–2159
22.99 J.J. Gilman: Insulator-metal transitions at microindentation, J. Mater. Res. **7** (1992) 535–538
22.100 G. M. Pharr, W. C. Oliver, R. F. Cook, P. D. Kirchner, M. C. Kroll, T. R. Dinger, D. R. Clarke: Electrical-resistance of metallic contacts on silicon and germanium during indentation, J. Mater. Res. **7** (1992) 961–972
22.101 A. Kailer, Y. G. Gogotsi, K. G. Nickel: Phase transformations of silicon caused by contact loading, J. Appl. Phys. **81** (1997) 3057–3063
22.102 J. E. Bradby, J. S. Williams, J. Wong-Leung, M. V. Swain, P. Munroe: Mechanical deformation in silicon by micro-indentation, J. Mater. Res. **16** (2000) 1500–1507
22.103 G. S. Was, T. Foecke: Deformation and fracture in microlaminates, Thin Solid Films **286** (1996) 1–31
22.104 A. J. Whitehead, T. F. Page: Nanoindentation studies of thin-film coated systems, Thin Solid Films **220** (1992) 277–283
22.105 A. J. Whitehead, T. F. Page: Nanoindentation studies of thin-coated systems, NATO ASI Ser. E **233** (1993) 481–488
22.106 B. D. Fabes, W. C. Oliver, R. A. McKee, F. J. Walker: The determination of film hardness from the composite response of film and substrate to nanometer scale indentations, J. Mater. Res. **7** (1992) 3056–3064
22.107 T. F. Page, S. V. Hainsworth: Using nanoindentation techniques for the characterization of coated systems – a critique, Surface Coat. Technol. **61** (1993) 201–208
22.108 X. Chen, J. J. Vlassak: Numerical study on the measurement of thin film mechanical properties by means of nanoindentation, J. Mater. Res. **16** (2001) 2974–2982
22.109 P. J. Burnett, T. F. Page: Surface softening in silicon by ion-implantation, J. Mater. Sci. **19** (1984) 845–860
22.110 P. J. Burnett, D. S. Rickerby: The mechanical-properties of wear resistant coatings. 1. Modeling of hardness behavior, Thin Solid Films **148** (1987) 41–50
22.111 P. J. Burnett, D. S. Rickerby: The mechanical-properties of wear resistant coatings. 2. Experimental studies and interpretation of hardness, Thin Solid Films **148** (1987) 51–65
22.112 P. M. Sargent: A better way to present results from a least-squares fit to experimental-data – an example from microhardness testing, J. Test. Eval. **14** (1986) 122–127
22.113 S. J. Bull, D. S. Rickerby: Evaluation of coatings, Brit. Ceram. Trans. J. **88** (1989) 177–183
22.114 B. R. Lawn, A. G. Evans, D. B. Marshall: Elastic/plastic indentation damage in ceramics: The median/radial crack system, J. Am. Ceram. Soc. **63** (1980) 574–581
22.115 R. Hill: *The Mathematical Theory of Plasticity* (Clarendon, Oxford 1950)
22.116 W. C. Oliver, C. J. McHargue, S. J. Zinkle: Thin-film characterization using a mechanical-properties microprobe, Thin Solid Films **153** (1987) 185–196
22.117 N. G. Chechechin, J. Bottiger, J. P. Krog: Nanoindentation of amorphous aluminum oxide films. 1. Influence of the substrate on the plastic properties, Thin Solid Films **261** (1995) 219–227
22.118 N. G. Chechechin, J. Bottiger, J. P. Krog: Nanoindentation of amorphous aluminum oxide films. 2. Critical parameters for the breakthrough and a membrane effect in thin hard films on soft substrates, Thin Solid Films **261** (1995) 228–235
22.119 D. E. Kramer, A. A. Volinsky, N. R. Moody, W. W. Gerberich: Substrate effects on indentation plastic zone development in thin soft films, J. Mater. Res. **16** (2001) 3150–3157
22.120 A. B. Mann: Nanomechanical measurements: Surface and environmental effects. Ph.D. Thesis (Oxford Univ., Oxford 1995)
22.121 A. B. Mann, J. B. Pethica, W. D. Nix, S. Tomiya: Nanoindentation of epitaxial films: A study of pop-in events, Mater. Res. Soc. Symp. Proc. **356** (1995) 271–276
22.122 R. Saha, W. D. Nix: Effects of the substrate on the determination of thin film mechanical properties by nanoindentation, Acta Mater. **50** (2002) 23–38
22.123 S. A. Barnett: Deposition and mechanical properties of superlattice thin films. In: *Physics of Thin Films*, ed. by M. H. Francombe, J. L. Vossen (Academic, New York 1993)
22.124 R. R. Oberle, R. C. Cammarata: Dependence of hardness on modulation amplitude in electrodeposited Cu-Ni compositionally modulated thin-films, Scripta Metall. **32** (1995) 583–588
22.125 A. Madan, Y. Y. Wang, S. A. Barnett, C. Engstrom, H. Ljungcrantz, L. Hultman, M. Grimsditch: Enhanced mechanical hardness in epitaxial non-isostructural Mo/NbN and W/NbN superlattices, J. Appl. Phys. **84** (1998) 776–785
22.126 R. Venkatraman, J. C. Bravman: Separation of film thickness and grain-boundary strengthening effects in Al thin-films on Si, J. Mater. Res. **7** (1992) 2040–2048
22.127 J. D. Embury, J. P. Hirth: On dislocation storage and the mechanical response of fine-scale microstructures, Acta Mater. **42** (1994) 2051–2056
22.128 D. J. Srolovitz, S. M. Yalisove, J. C. Bilello: Design of multiscalar metallic multilayer composites for high-strength, high toughness, and low CTE

mismatch, Metall. Trans. A **26** (1995) 1805–1813
22.129 J. S. Koehler: Attempt to design a strong solid, Phys. Rev. B **2** (1970) 547–551
22.130 S. V. Kamat, J. P. Hirth, B. Carnahan: Image forces on screw dislocations in multilayer structures, Scripta Metall. **21** (1987) 1587–1592
22.131 M. Shinn, L. Hultman, S. A. Barnett: Growth, structure, and microhardness of epitaxial TiN/NbN superlattices, J. Mater. Res. **7** (1992) 901–911
22.132 J. E. Krzanowski: The effect of composition profile on the strength of metallic multilayer structures, Scripta Metall. **25** (1991) 1465–1470
22.133 J. B. Vella, R. C. Cammarata, T. P. Weihs, C. L. Chien, A. B. Mann, H. Kung: Nanoindentation study of amorphous metal multilayered thin films, MRS Symp. Proc. **594** (2000) 25–29
22.134 J. L. Cuy, A. B. Mann, K. J. Livi, M. F. Teaford, T. P. Weihs: Nanoindentation mapping of the mechanical properties of human molar tooth enamel, Arch. Oral Biol. **47** (2002) 281–291

23. Atomistic Computer Simulations of Nanotribology

Molecular dynamics (MD) and related simulation techniques are powerful tools for improving our understanding of nanotribology. In simulations, materials properties and boundary conditions can be varied at will, and the resulting changes in both macroscopic variables and the dynamics of individual atoms can be observed. This allows one to study systematically the effects of many different factors on friction and wear at the nano-scale. Some examples that are considered in this chapter are (i) the symmetry of contacting surfaces (disordered vs. crystalline and with or without common periods), (ii) surface elasticity, (iii) surface curvature or topology, (iv) interfacial adhesion, and (v) lubricant and/or contaminant molecules present at the interface. Results from simulations and experiments on isolated nanoscale contacts often contradict our experience from macroscopic systems. Kinetic friction coefficients can be orders of magnitude smaller than those observed in macroscopic experiments. Detailed calculations even suggest that there should be no static friction between most pairs of clean, chemically passivated surfaces unless the load is large enough to produce wear. Simulations that test a series of possible mechanisms for static friction are described. Geometrical interlocking can produce static friction in contacts containing only a few atoms, such as an atomic force microscope tip. Larger contacts only exhibit static friction when there is wear, or when the surfaces are separated by a glassy contaminant layer that locks them together. Most surfaces are coated by such films and they are shown to yield friction forces that agree with both nanoscale and macroscopic experiments.

23.1 **Molecular Dynamics** ... 718
23.1.1 Model Potentials ... 719
23.1.2 Maintaining Constant Temperature. 720
23.1.3 Imposing Load and Shear ... 721
23.1.4 The Time-Scale and Length-Scale Gaps ... 721
23.1.5 A Summary of Possible Traps ... 722

23.2 **Friction Mechanisms at the Atomic Scale**. 723
23.2.1 Geometric Interlocking ... 723
23.2.2 Elastic Instabilities ... 724
23.2.3 Role of Dimensionality and Disorder 727
23.2.4 Elastic Instabilities vs. Wear in Atomistic Models ... 727
23.2.5 Hydrodynamic Lubrication and Its Confinement-Induced Breakdown . 729
23.2.6 Submonolayer Films ... 731

23.3 **Stick-Slip Dynamics** ... 732

23.4 **Conclusions** ... 734

References ... 735

One of the main barriers to the design of nanodevices with moving parts is our limited understanding of nanotribology; friction, lubrication and wear at the nanometer scale. New experimental methods such as the atomic force microscope (AFM) [23.1,2], quartz-crystal microbalance (QCM) [23.3], and surface forces apparatus (SFA) [23.4, 5] have revolutionalized our ability to measure friction in systems with atomic scale dimensions. However, the results are often qualitatively different than in macroscopic systems. For example, solids may slide more easily than fluids [23.3, 6] and fluids may behave like solids [23.4, 7]. These results have also revealed that there are gaps in our understanding of the molecular underpinnings of macroscopic tribology. There is a growing recognition that the behavior of nanometer-thick interfacial layers is crucial to tribology in systems of all scales, which has led to a fruitful convergence of researchers from many disciplines.

Understanding friction is an interesting and challenging scientific task, owing to the many factors that come into play: elastic and plastic properties of the two solids in relative sliding motion, surface roughness in multi-asperity contacts and tip geometry in single asperity contacts, geometric correlation between the solids in a microcontact, the amount of lubricant between the two solids, the physical nature of the lubricants, chemical

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

properties and reactions, adhesion, thermal fluctuations and aging, heat conductivity, etc.

Computer simulations have played an increasingly important role in improving our understanding of tribological phenomena. Unlike experiments, they allow us to perfectly control the geometry, the chemical composition, the sliding conditions, etc. and thus make it possible to explore the effect of each variable on friction, lubrication, and wear. Moreover, theorists have no other general approach to analyze processes like friction and wear. There is no known principle like minimization of free energy that determines the steady state of nonequilibrium systems. Even if there were, simulations would be needed to address the complex systems of interest, just as in many equilibrium problems.

Two different types of friction are commonly defined. Static friction, F_s, is the lateral force that must be applied to initiate sliding of one object over another, while kinetic friction, F_k, is the lateral force that must be applied to maintain sliding. At the macroscopic scale both forces usually obey Amontons's laws, which state that friction is proportional to the load pushing the surfaces together and is independent of area. However, there are obvious exceptions, such as tape, where friction occurs at zero load.

Despite significant experimental, theoretical, and computational efforts, no consensus has yet been reached on the relevant atomistic origins of friction between two solid bodies. Strictly speaking, static friction is not a force but a threshold. Its presence implies that the objects have locked together into a local energy minimum that must be overcome by the external force. This threshold force is observed between all the objects around us, yet its molecular origins have remained baffling. How does almost every pair of surfaces manage to lock together, and why does the friction usually obey Amontons's laws? Why does F_k not vanish linearly at small sliding velocities as it would if sliding occurred adiabatically? Different mechanisms have been suggested, such as geometric interlocking, elastic instabilities, interlocking mediated by so-called third bodies, plastic deformation, and crack propagation, to name a few. All these mechanisms may well be applicable, although each one under different circumstances.

In the following, we present a selected overview of the major results from the growing simulation literature. The emphasis is on providing a coherent picture of the key issues in the field, rather than a historical or a complete review. A much more detailed review of tribological computer simulations, including the sliding of adsorbed layers on surfaces, stick-slip motion, tribochemistry, cutting, high-velocity friction between metals, etc. has been given recently by *Robbins* et al. [23.8], and a review of the theoretical foundations of nanotribological issues is given in *Müser* et al. [23.9]. Reviews emphasizing the importance of surface chemistry and the role of self-assembled monolayers have been provided by *Harrison* et al. [23.10, 11].

23.1 Molecular Dynamics

The simulations described in this chapter all use an approach called classical molecular dynamics (MD), which is described extensively in a number of review articles and books [23.12, 13]. The basic outline of the method is straightforward. One begins by defining the interaction potentials. These produce forces on the individual particles whose dynamics will be followed, typically atoms or molecules. Next, the geometry and boundary conditions are specified, and initial coordinates and velocities are given to each particle. Then, the equations of motion for the particles are integrated numerically, stepping forward in time by discrete steps of size Δt. Quantities such as forces, velocities, work, heat flow, and correlation functions are calculated as a function of time to determine their steady-state values and dynamic fluctuations. The relation between changes in these quantities and the motion of individual molecules is also explored.

When designing a simulation, care must be taken to choose interaction potentials and ensembles that capture the essential physics that is to be addressed. The potentials may be as simple as ideal spring constants for studies of general phenomena, or as complex as electronic density-functional calculations in quantitative simulations. The ensemble can also be tailored to the problem of interest. Solving Newton's equations yields a constant energy and volume, or microcanonical, ensemble. By adding terms in the equation of motion that simulate heat baths or pistons, simulations can be done at constant temperature, pressure, lateral force, or velocity. Constant chemical potential can also be maintained by adding or removing particles using Monte Carlo methods or explicit particle baths.

The amount of computational effort typically grows linearly with both the number of particles, N, and the number of time-steps, M. The prefactor increases rapidly with the complexity of the interactions, and substantial ingenuity is required to achieve linear scaling with N for long-range interactions or density-functional approaches. Complex interactions also lead to a wide range of characteristic frequencies such as fast bond-stretching and slow bond-bending modes. Unfortunately, the time step Δt must be small compared to the period of the slowest mode ($\lessapprox 5\%$). This means that many time steps are needed before one obtains information about the slow modes.

The maximum feasible simulation size has increased continuously with advances in computer technology, but remains relatively limited. The product of N times M in the largest simulations described below is about 10^{12}. A cubic region of 10^6 atoms would have a side of about 50 nm. Such linear dimensions allow reasonable models of an atomic force microscope tip, the boundary lubricant in a surface force apparatus, or an individual asperity contact on a rough surface. However, 10^6 time steps is only about 10 ns, which is much shorter than experimental measurement times. This requires intelligent choices in the problems that are attacked and how results are extrapolated to experiments. It also limits sliding velocities to relatively high values, typically meters per second or above.

A number of efforts are underway to increase the accessible time scale, but the problem remains unsolved. Current algorithms attempt to follow the deterministic equations of motion, usually with the Verlet or predictor-corrector algorithms [23.12]. One alternative approach is to make stochastic steps. This would be a nonequilibrium generalization of the Monte Carlo approach that is commonly used in equilibrium systems. The difficulty is that there is no general principle for determining the appropriate probability distribution of steps in a nonequilibrium system.

In the following subsection, we describe some of the potentials that are commonly used and the situations where they are appropriate. The next two subsections describe methods for maintaining constant temperature and constant load. The section concludes with discussions of the challenges associated with the wide range of relevant time and length scales, and common pitfalls in simulations.

23.1.1 Model Potentials

A wide range of potentials has been employed in studies of tribology. Many of the studies described in the next section use simple ideal springs and sine-wave potentials. The Lennard-Jones potential gives a more realistic representation of typical interatomic interactions and is also commonly used in studies of general behavior. In order to model specific materials, more detail must be built into the potential. Simulations of metals frequently use the embedded atom method, while studies of hydrocarbons use potentials that include bond-stretching, bending, torsional forces, and even chemical reactivity. In this section, we give a brief definition of the most commonly used models. The reader may consult the original literature for more detail.

The Lennard-Jones (LJ) potential is a two-body potential that is commonly used for interactions between atoms or molecules with closed electron shells. It is applied not only to the interaction between noble gases, but also to the interaction between different segments on polymers. In the latter case, one LJ particle may reflect a single atom on the chain (explicit atom model), a CH_2 segment (united atom model), or even a Kuhn's segment consisting of several CH_2 units (coarse-grained model). United atom models [23.14] have been shown by *Paul* et al. [23.15] to successfully reproduce explicit atom results for polymer melts. Significant progress has been made recently to map back from the coarse-grained models to more detailed models [23.16, 17].

The 12-6 LJ potential has the form

$$U(r_{ij}) = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] , \qquad (23.1)$$

where r_{ij} is the distance between particles i and j, ϵ is the LJ interaction energy, and σ is the LJ diameter. The exponents 12 and 6 used above are very common, but other values may be chosen depending on the system of interest. Many of the simulation results discussed in subsequent sections are expressed in units derived from ϵ, σ, and a characteristic mass of the particles. For example, the standard LJ time unit is defined as $t_{\mathrm{LJ}} = \sqrt{m\sigma^2/\epsilon}$ and would typically correspond to a few picoseconds. A convenient time step is $\Delta t = 0.005 t_{\mathrm{LJ}}$ for a LJ liquid or solid at external pressures and temperatures that are not too large.

Most realistic atomic potentials cannot be expressed as two-body potentials. For example, bond angle potentials in a polymer are effectively three-body interactions and torsional potentials correspond to four-body interactions. Classical models of these interactions [23.18, 19] assume that a polymer is chemically inert and interactions between different molecules are modeled by two-body potentials. In the most sophisticated models, bond-stretching, bending, and torsional energies depend on the position of neighboring molecules and bonds are allowed to rearrange [23.20, 21]. Such generalized model potentials are needed to model friction-induced chemical interactions.

For the interaction between metals, a different approach has proven fruitful. The embedded atom method (EAM) developed by *Daw* et al. [23.22] includes a contribution in the potential energy associated with the cost of "embedding" an atom in the local electron density ρ_i produced by surrounding atoms. The total potential energy U is approximated by

$$U = \sum_i \tilde{F}_i(\rho_i) + \sum_i \sum_{j<i} \phi_{ij}(r_{ij}) \; , \tag{23.2}$$

where $\tilde{F}_i$ is the embedding energy, whose functional form depends on the particular metal. The pair potential $\phi_{ij}(r_{ij})$ is a doubly screened short-range potential reflecting core-core repulsion. The computational cost of the EAM is not substantially greater than pair potential calculations, because the density ρ_i is approximated by a sum of independent atomic densities. When compared to simple two-body potentials such as Lennard–Jones or Morse potentials, the EAM has been particularly successful in reproducing experimental vacancy formation energies and surface energies, even though the potential parameters were only adjusted to reproduce bulk properties. This feature makes the EAM an important tool in tribological applications, where surfaces and defects play a major role.

23.1.2 Maintaining Constant Temperature

An important issue for tribological simulations is temperature regulation. The work done when two walls slide past each other is ultimately converted into random thermal motion. The temperature of the system would increase indefinitely if there was no way for this heat to flow out of the system. In an experiment, heat flows away from the sliding interface into the surrounding solid. In simulations, the effect of the surrounding solid must be mimicked by coupling the particles to a heat bath.

Techniques for maintaining constant temperature T in equilibrium systems are well developed. Equipartition guarantees that the average kinetic energy of each particle along each Cartesian coordinate is $k_\mathrm{B}T/2$, where k_B is Boltzmann's constant. To thermostat the system, the equations of motion are modified, so that the average kinetic energy stays at this equilibrium value. This approach assumes that T is not too small compared to the Debye temperature, so that quantum statistics are not important. The applicability of classical MD decreases at lower T.

One class of approaches removes or adds kinetic energy to the system by multiplying the velocities of all particles by the same global factor. In the simplest version, velocity rescaling, the factor is chosen to keep the kinetic energy exactly constant at each time step. However, in a true constant temperature ensemble there are fluctuations in the kinetic energy. Improvements such as the Berendsen and Nosé–Hoover methods [23.23] add equations of motion that gradually scale the velocities to maintain the correct average kinetic energy over a longer time scale.

Another approach is to couple each atom to its own local thermostat [23.24, 25]. The exchange of energy with the outside world is modeled by a Langevin equation that includes a damping coefficient γ and a random force $f_i(t)$ on each atom i. The equations of motion for the α component of the position $x_{i\alpha}$ become:

$$m_i \frac{\mathrm{d}^2 x_{i\alpha}}{\mathrm{d}t^2} = -\frac{\partial}{\partial x_{i\alpha}} U - m_i \gamma \frac{\mathrm{d}x_{i\alpha}}{\mathrm{d}t} + f_{i\alpha}(t) \; , \tag{23.3}$$

where U is the total potential energy and m_i is the mass of the atom. To produce the appropriate temperature, the forces must be completely random, have zero mean, and have a second moment given by

$$\langle \delta f_{i\alpha}(t)^2 \rangle = 2 k_\mathrm{B} T m_i \gamma / \Delta t \; . \tag{23.4}$$

The damping coefficient γ must be large enough that energy can be removed from the atoms without substantial temperature increases. However, it should be small enough that the trajectories of individual particles are not perturbed too strongly.

The first issue in nonequilibrium simulations is what temperature means. Near equilibrium hydrodynamic theories define a local temperature in terms of the equilibrium equipartition formula and the kinetic energy relative to the local rest frame [23.26]. In d dimensions, the definition is

$$k_\mathrm{B} T = \frac{1}{dN} \sum_i m_i \left[\frac{\mathrm{d}\boldsymbol{x}_i}{\mathrm{d}t} - \langle \boldsymbol{v}(\boldsymbol{x}) \rangle \right]^2 \; , \tag{23.5}$$

where the sum is over all N particles and $\langle \boldsymbol{v}(\boldsymbol{x}) \rangle$ is the mean velocity in a region around $\boldsymbol{x}$. As long as the change in mean velocity with position is sufficiently slow, $\langle \boldsymbol{v}(\boldsymbol{x}) \rangle$ is well-defined and this definition of temperature is on solid theoretical ground.

When the mean velocity difference between neighboring molecules becomes comparable to the random thermal velocities, temperature is not well-defined. An important consequence is that different strategies for defining and controlling temperature give very different structural order and shear forces [23.27–29]. In addition, the distribution of velocities may become non-Gaussian, and different directions, α, may have different effective temperatures. Care should be taken in drawing conclusions from simulations in this extreme regime. Fortunately, the above condition typically implies that the velocities of neighboring atoms differ by something approaching 10% of the speed of sound. This is generally higher than any experimental velocity and would certainly lead to heat buildup and melting at the interface.

In order to mimic experiments, the thermostat is often applied only to those atoms that are at the outer boundary of the simulation cell. This models the flow of heat into surrounding material that is not included explicitly in the simulation. The resulting temperature profile is peaked at the shearing interface (e.g., *Bowden* et al. [23.30], *Khare* et al. [23.31]). In some cases the temperature rise may lead to undesirable changes in the structure and dynamics, even at the lowest velocity, that can be treated in the available simulation time. In this case, a weak thermostat applied throughout the system may maintain the correct temperature and yield the dynamics that would be observed in longer simulations at lower velocities. The safest course is to couple the thermostat only to those velocity components that are perpendicular to the mean flow.

There may be a marginal advantage to local Langevin methods in nonequilibrium simulations, because they remove heat only from atoms that are too hot. Global methods like Nosé–Hoover remove heat everywhere. This can leave high temperatures in the region where heat is generated, while other regions are at an artificially low temperature.

23.1.3 Imposing Load and Shear

The magnitude of the friction that is measured in an experiment or simulation may be strongly influenced by the way in which the normal load and tangential motion are imposed [23.32]. Experiments almost always impose a constant normal load. The mechanical system applying shear can usually be described as a spring attached to a stage moving at controlled velocity. The effective spring constant includes the compliance of all elements of the loading device, including the material on either side of the interface. Very compliant springs apply a nearly constant force, while very stiff springs apply a nearly constant velocity.

In simulations, it is easiest to treat the boundary of the system as rigid and to translate atoms in this region at a constant height and tangential velocity. However, this does not allow atoms to move around or up and over atoms on the opposing surface. Even the atomic-scale roughness of a crystalline surface can lead to order of magnitude variations in normal and tangential force with lateral position when sliding is imposed in this way [23.33]. The difference between constant separation and pressure simulations of thin films can be arbitrarily large, since they produce different power law relations between viscosity and sliding velocity [23.34].

One way of minimizing the difference between constant separation and pressure ensembles is to include a large portion of the elastic solids that bound the interface. This extra compliance allows the surfaces to slide at a more uniform normal and lateral force. However, the extra atoms greatly increase the computational effort.

To simulate the usual experimental ensemble more directly, one can add equations of motion that describe the position of the boundary [23.34–36], much as equations are added to maintain constant temperature. The boundary is given an effective mass that moves in response to the net force from interactions with mobile atoms and from the external pressure and lateral spring. The mass of the wall should be large enough that its dynamics are slower than those of individual atoms, but not too much slower or the response time will be long compared to the simulation time. The spring constant should also be chosen to produce an appropriate response time and ensemble.

23.1.4 The Time-Scale and Length-Scale Gaps

One important motivation of tribological computer simulations is to find the relevant atomistic processes that occur in a single, well-defined sliding contact such as in the SFA or in the AFM. While it has become computationally feasible to study atomistically the core of a nanometer-scale AFM contact, the sliding speeds, v_{sl}, employed in the simulations are still about ten decades larger than in experiments. The use of clever new algorithms such as kinetic Monte Carlo might prove fruitful to overcome this time-scale gap in a brute-force way. However, such an approach is not necessarily required in order to make a meaningful comparison between experiment and simulation. Instead, one can try to find out what the relevant time scales are and what parameters define them, e.g., by studying a phenomenological description of an embedded system under shear. In such an analysis, *Porto* et al. [23.37] found that the motion of the top plate/tip is not sensitive to the characteristic atomic frequencies ω_{mic} – as long as there is a reasonable time-scale separation between ω_{mic} and the characteristic frequency of the tip Ω_{tip}. The experimentally relevant time scale Ω_{tip} is given by $\sqrt{K/M}$, where K is a restoring spring constant (defined by the tip's elasticity and/or the tip-substrate interactions) and M is the tip's inertia, eventually including the cantilever's mass. The experimental value of M will be about 20 decades larger than in the simulation, so values of the relevant parameter $v_{\text{sl}}\Omega_{\text{tip}}$ end up being similar in simulation and experiment. However, there is controvery over whether the tip can be treated as a mass point, or whether long-range elastic deformations in the tip are important [23.35], [23.38, Chap. 9.9].

Another way to overcome the gap in time scales is to alter interaction strengths, temperature, or other variables in order to accelerate the processes of interest. One such example is the cold welding study of *Müser* [23.40], where two dissimilar solids were placed on top of each other in vacuum. It would have been by no means desirable to base the calculations on interfacial interactions that correspond to typical experimental values, because the resulting time scales would be too long. By increasing the magnitude of the mixing energy, one drives the mixing process toward the same ground state on simulation time scales. This simple trick makes it possible to indirectly bridge several orders of magnitude in time scale between experiment and simulation.

Some open issues appear to require the explicit treatment of many different length scales. For example, the distribution of contacts and total contact area depend on both atomic-scale interactions at the surface and long-range elastic deformation of the bounding solid. Of particular interest is whether different contacting asperities can move independently, or whether those within an elastic coherence length l_{e} move in lock-step. Estimates of l_{e} range from the size of a single asperity [23.41] to many times the linear dimension of most objects of interest [23.42]. The correct value is of great importance to the question of whether clean, corrugated (macroscopic) surfaces have the potential to show elastic pinning and, hence, wearless friction. In order to address such issues, it will be necessary to do simulations that incorporate both atomistic and finite element methods (e.g., *Abraham* et al. [23.43]).

23.1.5 A Summary of Possible Traps

There are many possible traps that may render a well-done simulation useless or applicable only to highly idealized situations. In some aspects more care has to be taken than in equilibrium simulations, in other aspects one gets away with less care. We hope that this section helps to create some sensitivity as to what "ingredients" are relevant. For example, if we calculate the static friction force between two solids separated by a boundary lubricant in a MD simulation, the results will be fairly insensitive to the thermostat and to whether or not the thermostat is applied to the boundary lubricant. One should also avoid moves in which a "stuck" atom magically disappears from the contact through a grand canonical-type Monte Carlo move. The results depend more strongly on the thermostat if the lubricant is thick and the heating effects are of interest. In such a case, only the outermost atoms in the confining solids should be thermostated.

The typical situation in experiment is constant load, *not* constant separation. In some cases, there is something like a correspondence between these two "ensembles", but often there are *qualitative* differences between them, in particular, when strongly irreversible processes take place such as the production of wear particles [23.40]. Contacting experimental surfaces are also likely to have different geometry and chemistry – or at least different crystallographic orientations. However, most simulations consider the artificial case of contact

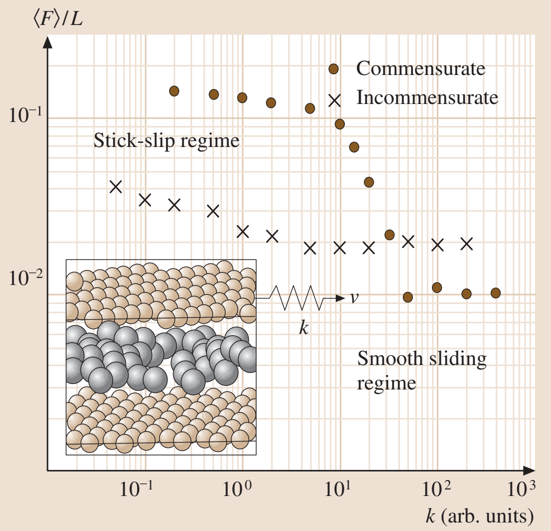

Fig. 23.1 Kinetic friction force F_{k} divided by load L as a function of spring constant k for commensurate and incommensurate walls lubricated by a quarter monolayer of LJ atoms. A schematic of the simulation is shown as well. The data shown was produced at a normal pressure of 0.4 GPa and a velocity of about 1 m/s. (After [23.39])

between identical, aligned crystals. This is the simplest example of a commensurate system, where the two surfaces share a common spatial period. As discussed below, the tribological behavior of commensurate systems is dramatically different than the more realistic case of incommensurate surfaces – even in the presence of a few layers of lubricant. One example of this is shown in Fig. 23.1 [23.39].

Another issue is the curvature of the surfaces. Real contacts typically take place between two curved surfaces or a curved and a flat surface. The pressure profiles differ from those in "flat" contacts, and that discrepancy can be relevant in certain circumstances. If we want to incorporate curvature effects, care should be taken that the resulting pressure profiles resemble a realistic profile, as was done by *Wenning* et al. [23.44].

23.2 Friction Mechanisms at the Atomic Scale

As noted in the introduction, static friction implies that the contacting surfaces have locked into a local free energy minimum. There have been many suggestions for how this can happen between nearly every pair of surfaces, whether macroscopic or nanoscale. The purpose of this section is to review some of these mechanisms and discuss their relevance to nanotribological experiments.

One obvious way that surfaces can interlock is through plastic deformation that brings the surfaces into conformity. A classic example is called plowing friction [23.30]. In this case a sharp feature on one surface penetrates into the opposing surface, causing permanent plastic deformation. The static friction represents the force needed to plow a groove through the surface. This type of motion has been observed in simulations [23.8, 9, 45]. and experiments (Sect. 20.6.1) using AFM probes with hard tips Because it leads to rapid wear, it is only desirable when one wants to remove material by abrasion or cutting.

Many nanotribological studies show friction without measurable wear, and this is the desirable mode of operation in most applications. For this reason, the following subsections focus on mechanisms that lock surfaces together without plastic deformation. We begin by discussing contact between rigid surfaces (Sect. 23.2.1), and then explore the ability of elastic deformations at the atomic scale (Sect. 23.2.2) or between contacts (Sect. 23.2.3) to lock surfaces together. We then describe one of the simulations that shows how unlikely these mechanisms are to yield static friction without wear unless the contact contains only a few tens of atoms (Sect. 23.2.4). The final two subsections consider the role of contaminants or "third bodies" that are present in the contacts between nearly all surfaces. First, changes in the behavior of lubricants with film thickness are discussed. Then molecularly thin films are shown to produce static and kinetic friction that is consistent with both nanoscale experiments and macroscopic observations.

23.2.1 Geometric Interlocking

Early theories of friction were based on the purely geometric argument that friction is caused by the interlocking of impenetrable and rigid surface asperities [23.30, 46]. The idea (Fig. 23.2) is that the top solid must be lifted up a typical slope, $\tan\alpha$, determined by roughness on the bottom surface. If there is no microscopic friction between the surfaces, then the minimum force to initiate sliding is

$$F_s = \mu_s L \tag{23.6}$$

with $\mu_s = \tan\alpha$. This result satisfies Amontons's laws with a constant coefficient of friction $\mu_s \equiv F_s/L = \tan\alpha$. In 1737, Bélidor obtained a typical experimental value of $\mu_s \approx 0.35$ by modeling rough surfaces as spherical asperities arranged to form commensurate crystalline walls [23.46]. However, asperities on real surfaces do not match as well as envisioned in these models or as sketched in Fig. 23.2. In most cases the average force vanishes because for every asperity or atom going up a ramp, there is another going down.

The sum of forces between crystalline surfaces is considered explicitly in the next subsection, and shown

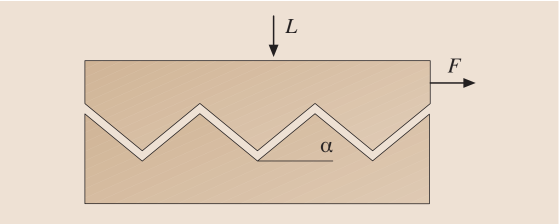

Fig. 23.2 Sketch of two surfaces with interlocking asperities. The top surface experiences a normal load L and a lateral force F, which attempts to pull the top surface up the slope $\tan\alpha$. The bottom wall is fixed

to vanish except in the unusual case where the surfaces have the same periodicity. In the case were one or both surfaces are random, simple analytic arguments show that μ_s will fall as $\mathcal{L}^{-d_{\mathrm{int}}/2}$ when the linear size $\mathcal{L}$ of the contact goes to infinity [23.47], provided that the surface roughnesses are small and the solids are completely rigid. Here, d_{int} is the dimension of the contact and the pressure normal to the interface is assumed to be kept constant. This result is in contradiction with one of Amontons's laws, namely, the independence of F_s on the area of contact A. Nevertheless, the linear relation between F_s and L at fixed A is maintained if one assumes that the surfaces are not completely impenetrable but repel via an energy penalty that increases exponentially fast with increasing overlap of the two solids [23.47]. The resulting theoretical prediction that $F_s \propto L/\sqrt{A}$ between rigid, flat, disordered walls also holds reasonably well for more general interactions between the solids, as tested explicitly by atomistic computer simulations.

Most real surfaces are rough and disordered. The minimal model that includes the effect of roughness consists of a single asperity contact, for instance the curved tip of an AFM [23.2] on a flat substrate. The tip is usually assumed to be disordered. Moreover, the contact mechanics of a nanoscale tip and a substrate in vacuum can best be described by the so-called Derjaguin–Müller–Toporov (DMT) limit [23.48]. DMT is similar to Hertzian contact mechanics, but it includes an additional adhesive off-set load L_{adh} that is incorporated into the effective load $L_{\mathrm{eff}} = L + L_{\mathrm{adh}}$. The area of contact is then given by $A \propto L_{\mathrm{eff}}^{2/3}$, Substituting this proportionality into the relation $F_s \propto L_{\mathrm{eff}}/\sqrt{A}$ discussed above, *Wenning* et al. [23.44] found

$$F_s \propto L^{2/3} \,. \tag{23.7}$$

This relationship is indeed observed experimentally in ultrahigh vacuum (UHV) [23.49, 50], as well as in atomistic computer simulations [23.44]. Of course, when the contacts are commensurate, we can expect the effective friction coefficient $\mu_{\mathrm{eff}} = F_s/L_{\mathrm{eff}}$ to be independent of L_{eff}, while for amorphous systems and in particular incommensurate systems, μ_{eff} decreases with increasing L_{eff}. The scaling with load in Fig. 23.3 confirms this expectation.

As noted above, the discussion thus far only deals with atomic/geometric interlocking. From Fig. 23.2 and the discussion above, one can understand how this interlocking leads to static friction, which is the threshold force to initiate sliding. However, kinetic friction $F_k(v)$ may nevertheless vanish in the limit of zero sliding velocity v, since in principle the energy that is gained when the system slides "downhill" can be stored and used later to make the slider go "uphill" again. Such a process would require that we can drive the system adiabatically. In the following, we will discuss atomistic processes (instabilities) that make it impossible to drive the system adiabatically. The presence of instabilities automatically leads to finite kinetic and static friction even in the absence of geometric interlocking.

23.2.2 Elastic Instabilities

In principle, elastic deflections of atoms can be large enough to prevent the interfacial forces from averaging to zero. Two simple ball and spring models are useful in illustrating how such motion can affect static and kinetic friction in principle as well as in understanding the results of detailed simulations or laboratory experiments. Both models consider two clean, flat, crystalline surfaces in direct contact (Fig. 23.4a). The bottom solid is assumed to be rigid, so that it can be treated as a fixed periodic substrate potential acting on the

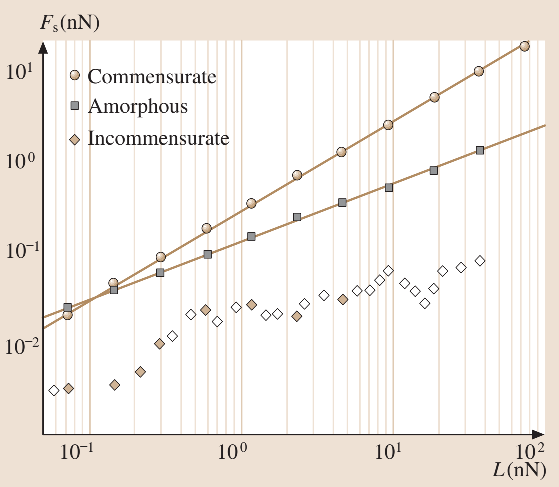

Fig. 23.3 Static friction force F_s vs. normal load L for a commensurate tip, an incommensurate tip, and an amorphous tip. In all three cases, the radius of curvature was $R_c = 70\,\text{Å}$ and contacts were nonadhesive. Straight lines are fits according to $F_s \propto L^{\beta}$ with the results $\beta = 0.97 \pm 0.005$ (commensurate) and $\beta = 0.63 \pm 0.01$ (amorphous). [After [23.44] plus new data from L. *Wenning* (*open diamonds*)]

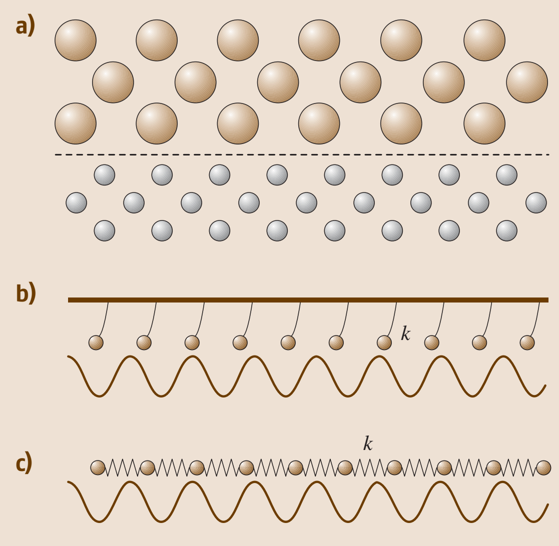

Fig. 23.4 **(a)** Two ideal, flat crystals making contact at the plane indicated by the *dashed line*. The nearest-neighbor spacings in the bottom and top walls are a and b, respectively. The Prandtl–Tomlinson model **(b)** and Frenkel–Kontorova model **(c)** replace the bottom surface by a periodic potential. The former model keeps elastic forces between atoms on the top surface and the center of mass of the top wall, and the latter includes springs of stiffness k between neighbors in the top wall. (After [23.56])

top solid. In order to make the problem analytically tractable, only the bottom layer of atoms from the top solid is retained, and the interactions within the top wall are simplified. In the Prandtl–Tomlinson model (Fig. 23.4b), the atoms are coupled to the center of mass of the top wall by springs of stiffness k, and coupling between neighboring atoms is ignored [23.51–53]. In the Frenkel–Kontorova model (Fig. 23.4c), the atoms are coupled to nearest-neighbors by springs, and the coupling to the atoms above is ignored [23.54, 55]. Due to their simplicity, these models arise in a number of different problems and a great deal is known about their properties.

Many features of the Prandtl–Tomlinson and Frenkel–Kontorova models can be understood from their 1-D versions. One important parameter is the ratio η between the lattice constants of the two surfaces $\eta \equiv b/a$. The other is the strength of the periodic potential from the substrate relative to the spring stiffness k that represents interactions within the top solid. If the substrate potential has a single Fourier component, then the periodic force can be written as

$$f(x) = -f_1 \sin\left(\frac{2\pi}{a}x\right) . \tag{23.8}$$

The relative strength of the potential and springs can be characterized by the dimensionless constant $\lambda \equiv 2\pi f_1/ka$.

In the limit of infinitely strong springs ($\lambda \to 0$), both models represent rigid solids. The atoms of the top wall are confined to lattice sites $x_l^0 = x_{\mathrm{CM}} + lb$, where the integer l labels successive atoms and x_{CM} represents a rigid translation of the entire wall. The total lateral or friction force is given by summing (23.8):

$$F = -f_1 \sum_{l=1}^{N} \sin\left[\frac{2\pi}{a}(lb + x_{\mathrm{CM}})\right] , \tag{23.9}$$

where N is the number of atoms in the top wall. In the special case of equal lattice constants ($\eta = b/a = 1$), the forces on all atoms add in phase and $F = -Nf_1 \sin(2\pi x_{\mathrm{CM}}/a)$. The maximum of this restraining force gives the static friction $F_{\mathrm{s}} = Nf_1$.

As noted above, the case of equal lattice constants is artificial and for real surfaces η is most likely to be an irrational number. Such surfaces are called incommensurate, while surfaces with a rational value of η are commensurate. When η is irrational, atoms on the top surface sample all phases of the periodic force with equal probability and the net force (23.9) vanishes exactly. When η is a rational number, it can be expressed as p/q, where p and q are integers with no common factors. In this case, atoms only sample q different phases. The net force from (23.9) still vanishes, because the force is a pure sine wave and the phases are equally spaced. However, the static friction is finite if the potential has higher harmonics. A Fourier component with wave vector $q2\pi/a$ and magnitude f_q contributes Nf_q to F_{s}. Studies of surface potentials [23.57] show that f_q drops exponentially with increasing q and thus imply that F_{s} will only be significant for small q.

As the springs become weaker, the top wall deforms more easily into a configuration that lowers the potential energy. The Prandtl–Tomlinson model is the simplest case to consider, because each atom can be treated as an independent oscillator within the upper surface. The equations of motion for the position x_l of the l^{th} atom can be written as

$$m\ddot{x}_l = -\gamma \dot{x}_l - f_1 \sin\left(\frac{2\pi}{a}x_l\right) - k(x_l - x_l^0) , \tag{23.10}$$

where m is the atomic mass and x_l^0 is the position of the lattice site. Here γ is a phenomenological damping coefficient, like that in a Langevin thermostat, that allows work done on the atom to be dissipated as heat. It represents the coupling to external degrees of freedom such as lattice vibrations in the solids.

In any steady state of the system, the total friction can be determined either from the sum of the forces exerted by the springs on the top wall, or from the sum of the periodic potentials acting on the atoms (23.9). If the time average of these forces differed, there would be a net force on the atoms and a steady acceleration. The static friction is related to the force in metastable states of the system where $\ddot{x}_l = \dot{x}_l = 0$. This requires that spring and substrate forces cancel for each l,

$$k(x_l - x_l^0) = -f_1 \sin\left(\frac{2\pi}{a} x_l\right) . \qquad (23.11)$$

As shown graphically in Fig. 23.5a, there is only one solution for weak interfacial potentials and stiff solids ($\lambda < 1$). In this limit, the behavior is essentially the same as for infinitely rigid solids. There is static friction for $\eta = 1$, but not for incommensurate cases. Even though incommensurate potentials displace atoms from lattice sites, there are exactly as many displaced to the right as to the left, and the force sums to zero.

A new type of behavior sets in when λ exceeds unity. The interfacial potential is now strong enough compared to the springs that multiple metastable states are possible. These states must satisfy both (23.11) and the condition that the second derivative of the potential energy is positive: $1 + \lambda \cos(2\pi x_l/a) > 0$. The number of metastable solutions increases as λ increases. As illustrated in Fig. 23.5b, once an atom is in a given metastable minimum it is trapped there until the center of mass moves far enough away that the second derivative of the potential vanishes and the minimum becomes unstable. The atom then pops forward very rapidly to the nearest remaining metastable state. Such motion, where a system is stuck for a long time and then suddenly jumps forward by an atomic spacing, is called atomic-scale stick-slip motion. During the pops, atoms reach large velocities and hence dissipate a large amount of energy into the lattice vibrations. This makes it possible to have a finite kinetic friction that does not vanish linearly with sliding velocity even when the surfaces are incommensurate [23.8, 9].

A similar analysis can be done for the 1-D Frenkel–Kontorova model [23.55, 58, 59]. The main difference is that the static friction and ground state depend strongly on η. For any given irrational value of η, there is a threshold potential strength λ_c. For weaker potentials, the static friction vanishes. For stronger potentials, elastic metastability produces a finite static friction. The transition to the onset of static friction was termed a breaking of analyticity by *Aubry* [23.59] and is often called the Aubry transition. The metastable states for $\lambda > \lambda_c$ take the form of locally commensurate regions that are separated by domain walls where the two crystals are out of phase. Within the locally commensurate regions the ratio of the periods is a rational number p/q that is close to η. The range of η where locking occurs grows with increasing potential strength (λ) until it spans all values. At this point there is an infinite number of different metastable ground states that form a fascinating "Devil's staircase" as η varies [23.55, 59].

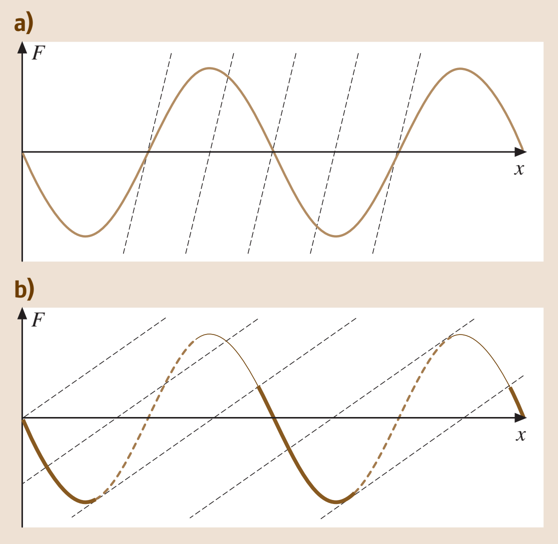

Fig. 23.5a,b Graphical solution for metastability of an atom in the Prandtl–Tomlinson model for **(a)** $\lambda = 0.5$ and **(b)** $\lambda = 3$. The *straight dashed lines* show the force from the spring, $k(x_l - x_l^0)$, at different x_l^0, and the *curved lines* show the periodic substrate potential. For $\lambda < 1$, there is a single intersection of the *dashed lines* with the substrate potential for each value of x_l^0, and thus a single metastable state. For $\lambda > 1$, there are multiple intersections with the substrate potential. *Dotted portions* of the potential curve indicate unstable maxima, and *solid regions* indicate metastable solutions. Fast pops with concurrent energy dissipation cannot be avoided for $\lambda > 1$ even when the slider moves slowly. (After [23.56])

Although elastic instabilities can, in principle, induce finite static friction at the atomic scale, this appears to be rare. There have been a significant number of computer simulations showing that the *wearless* static friction between crystalline surfaces is extremely small, except in the artificial case of identical and perfectly aligned walls [23.8, 10]. Studies of iron on iron [23.60], a blunt pyramidal diamond tip on a metal surface [23.61], a blunt pyramidal copper tip on incommensurate copper substrates [23.62], interlayer sliding in multiwalled carbon nanotubes [23.63], and Lennard-Jonesium on Lennard-Jonesium [23.40] all confirm the simple prediction that lateral forces cancel out. These simulations have in common that the solids were treated truly three-dimensionally and that the atoms (interacting via microscopic interaction potentials) were allowed to relax, thus making long-range elastic deformations possible. Significant lateral forces were only observed in combination with strongly irreversible processes such as production of wear debris, plastic deformation, material mixing, or cold welding and not due to geometric interlocking or elastic instabilities alone.

23.2.3 Role of Dimensionality and Disorder

The previous section considered whether local atomic displacements could lead to static friction. Long-range deformations between individual contacts could also allow surfaces to lock together [23.41, 42] Although we are concerned with the friction between two 3-D solid bodies, it is instructive to analyze the role of dimensionality in determining the effect of disorder. To do this, let us consider a d_{obj}-dimensional elastic solid, for example, a generalization of the elastic Frenkel–Kontorova chain to higher dimensions. We may safely assume the free elastic solid to be mechanically stable, meaning that the tensor of the elastic constants is positive definite. The dimension of the interface between the slider and disordered substrate is denoted by d_{int}.

In such a situation, there will be a competition between the random substrate-slider interactions and the elastic coupling within the solids. An important question to ask is how the interactions change when we change the scale of the system, for example, how strong are the random and the elastic interactions on a length scale $2\mathcal{L}$ if we know their respective strengths on a scale $\mathcal{L}$. Here $\mathcal{L}$ gives the linear dimension of the solids in all directions, that is to say, parallel and normal to the interface. As discussed above, the random forces between substrate and slider add incoherently and thus their sum scales as the square root of the interface area $\sim \mathcal{L}^{d_{\text{int}}/2}$. The elastic forces, on the other hand, scale as $\mathcal{L}^{d_{\text{obj}}-2}$. A linear chain can be more easily compressed if we replace one spring with two springs coupled in series. In two dimensions, springs are not only coupled in series but also in parallel, so that the elastic coupling remains invariant to a "block transformation". Each additional dimension strengthens the effect of "parallel" coupling.

In the thermodynamic limit $\mathcal{L} \to \infty$, the effect of disorder will always dominate the elastic interactions or vice versa, except in the special case where

$$\mathcal{L}^{d_{\text{int}}/2} \propto \mathcal{L}^{d_{\text{obj}}-2} \,. \tag{23.12}$$

For $\lim_{\mathcal{L}\to\infty} \mathcal{L}^{d_{\text{int}}/2}/\mathcal{L}^{d_{\text{obj}}-2} \gg 1$, the random interactions will dominate, and hence pinning via elastic instabilities cannot be avoided. This disorder-induced elastic pinning is similar to the pinning of compliant, ordered systems, as discussed above within the context of the Prandtl–Tomlinson and the Frenkel–Kontorova models. For $\lim_{\mathcal{L}\to\infty} \mathcal{L}^{d_{\text{int}}/2}/\mathcal{L}^{d_{\text{obj}}-2} \ll 1$, the long-range elastic forces dominate the long-range random forces. The slider's motion can only be opposed by elastic instabilities if the elastic coupling is sufficiently weak at finite $\mathcal{L}$ in order to make local pinning possible, again akin to the case of $\lambda > 1$ in the Prandtl–Tomlinson model.

The so-called *marginal* dimension, in which both contributions scale with the same exponent, $d_{\text{int}}/2 = d_{\text{obj}} - 2$, occurs in the important case of 3-D solid bodies with 2-D surfaces. In the marginal dimension, the friction force can stay finite, however, detailed analysis [23.42, 64] shows the friction force per unit area, and hence the friction coefficient, to be exponentially small. Even for objects with a dimensionality d_{mar} less than their marginal dimension $d_{\text{mar}} = 2 + d_{\text{int}}/2$ (see (23.12)), friction forces may turn out to be small. One example is an experimental quartz crystal microbalance study [23.65] of solid and liquid krypton films on disordered gold surfaces for which the pinning forces were undetectably small. For adsorbed layers, which satisfy $d_{\text{obj}} = d_{\text{int}}$, the marginal dimension would be $d_{\text{mar}} = 4$.

23.2.4 Elastic Instabilities vs. Wear in Atomistic Models

As mentioned in Sect. 23.2.2, there is no published computer simulation study known to us in which the solids were treated truly atomistically in three dimensions and where the interfacial interactions lead to elastic pinning according to Prandtl–Tomlinson or Frenkel–Kontorova. In this context, we want to discuss exemplarily computer simulations of a copper tip on copper by *Sørensen* et al. [23.62]. We consider their study to be rather generic and also well-done from a technical point of view for a variety of reasons: The effects of commensurability and finite contact area were examined in detail, the simulations were done at constant normal load instead of constant separation, the interaction potentials were realistic, and initial and boundary conditions were chosen such that the tip had the possibility to deform. Snapshots of one of their sliding simulations are shown in Fig. 23.6. Note that one must keep in mind that the simulation only addresses the tribological properties in UHV.

Flat, clean tips with (111) or (100) surfaces were brought into contact with corresponding crystalline substrates [23.62]. The two exterior layers of tip and surface were treated as rigid units, and the dynamics of the remaining mobile layers was followed. Interatomic forces and energies were calculated using semi-empirical potentials derived from effective medium theory [23.66]. At finite temperatures, the outer mobile layers of both tip and surface were coupled to a Langevin thermostat. Zero temperature simulations gave similar results.

In the commensurate Cu(111) case, *Sørensen* et al. [23.62] observed atomic-scale stick-slip motion of the tip like that described above for the Tomlinson model. The trajectory was of zigzag form, which could be related to jumps of the tip's surface between fcc and hcp positions. Detailed analysis of the slips showed that they occurred via a dislocation mechanism. Dislocations were nucleated at the corner of the interface, and then moved rapidly through the contact region. Adhesion led to a large static friction at zero load: The static friction per unit area, or critical yield stress, dropped from 3.0 GPa to 2.3 GPa as T increased from 0 to 300 K. The kinetic friction increased linearly with load with a surprisingly small differential friction coefficient $\tilde{\mu}_\mathrm{k} \equiv \partial F_\mathrm{k}/\partial L \approx 0.03$. In the load regime investigated, $\tilde{\mu}_\mathrm{k}$ was independent of temperature and load. No velocity dependence was detectable up to sliding velocities of $v = 5\,\mathrm{m/s}$. At higher velocities, the friction decreased. Even though the interactions between the surfaces are identical to those within the surfaces, no wear was observed. This was attributed to the fact that (111) surfaces are the preferred slip planes in fcc metals.

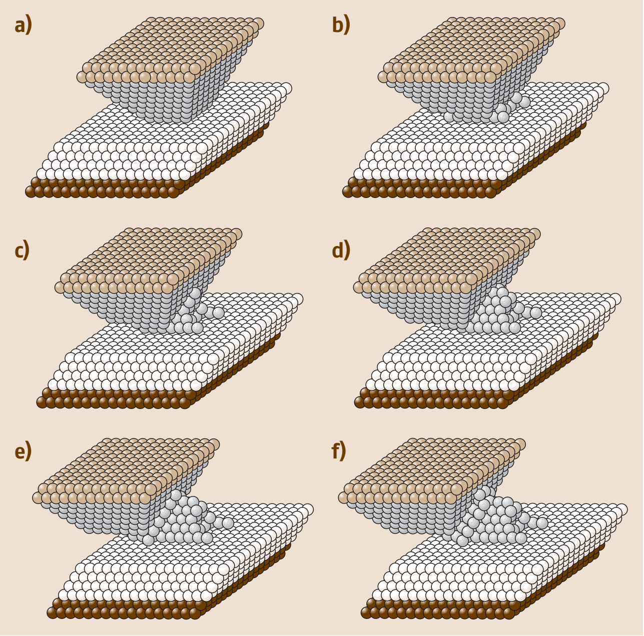

Fig. 23.6 Snapshots showing the evolution of a Cu(100) tip on a Cu(100) substrate during sliding to the left. (After [23.62])

Adhesive wear was observed between a commensurate (100) tip and substrate (Fig. 23.6). Sliding in the (011) direction at either constant height or constant load led to inter-plane sliding between (111) planes inside the tip. As shown in Fig. 23.6, this plastic deformation led to wear of the tip, which left a trail of atoms in its wake. The total energy was an increasing function of sliding distance due to the extra surface area. The constant evolution of the tip kept the motion from being periodic, but the sawtoothed variation of force with displacement that is characteristic of atomic-scale stick-slip was still observed.

Sørensen et al. [23.62] also examined the effect of incommensurability. An incommensurate Cu(111) system was obtained by rotating the tip by 16.1° about the axis perpendicular to the substrate. For a small tip (5 × 5 atoms), they observed an Aubry transition from smooth sliding with no static friction at low loads to atomic-scale stick-slip motion at larger loads. Further increases in load led to sliding within the tip and plastic deformation. Finite systems are never truly incommensurate, and pinning was found to occur at the corners of the contact, suggesting it was a finite-size effect. Larger tips (19 × 19) slid without static friction at all loads. Similar behavior was observed for incommensurate Cu(100) systems.

These results confirm the conclusions of *Hirano* et al. [23.60] that even bare metal surfaces of the same material will not exhibit static friction if the surfaces are incommensurate. They also indicate that contact areas as small as a few hundred atoms are large enough to exhibit this effect. Similar results were obtained for Lennard Jonesium [23.6, 40, 67]: The interbulk interactions between unlike atoms (each solid is initially composed of one Lennard-Jones species) have to be distinctly larger than the intrabulk interactions between identical atoms in order to induce elastic instability. However, as soon as the interactions between unlike atoms are larger than those between like atoms, there is a mixing instability leading to a cold welded junction.

Lancon [23.69] reported that there is the possibility for elastic instabilities under carefully designed conditions between two incommensurate gold surfaces. The importance of the interfacial interactions with respect to the intrabulk interactions was increased through the application of a pressure normal to the interface. At 4 GPa, an elastic Aubry transition was observed concurrently with finite static friction. The normal pressures employed in those simulations were well above the yield strength of gold. However, plastic flow at the contact's circumference could be avoided by the use of flat surfaces that filled out the whole simulation cell parallel to the interface. Additional examples of elastic instabilities may be found for very compliant solids, but the majority of simple solids will not pin elastically on an atomic scale. Note, however, that cold welding may still occur on long time scales due to diffusion. On such time scales, the interface geometry may rearrange to produce locally commensurate regions that lower the interfacial free energy and allow geometrical pinning.

23.2.5 Hydrodynamic Lubrication and Its Confinement-Induced Breakdown

So far, we have been concerned with dry friction. However, most surfaces are covered by dirt and/or lubricants. Hydrodynamics and elasto-hydrodynamics have been very successful in describing lubrication by micron-thick films [23.70]. These continuum theories begin to break down as atomic structure becomes important. Experiments and simulations reveal a sequence of dramatic changes in the static and dynamic properties of fluid films as their thickness decreases from microns down to molecular scales. These changes have important implications for the function of boundary lubricants and the origins of static friction.

Let us consider two flat, lubricated walls separated by a distance d and sliding with a relative parallel velocity v. Let us further assume that the lubricant's viscosity μ in the center of the contact equals that of the bulk fluid phase. The shear stress must be uniform in steady state, because any imbalance would lead to accelerations. Since the velocity gradient times the viscosity μ gives the stress in the fluid, the kinetic friction per unit area is

$$F_{\mathrm{k}} = \frac{v\mu}{d+\mathcal{S}}\,, \tag{23.13}$$

where we introduce the so-called slip length $\mathcal{S}$. It represents the distance into the wall at which the velocity gradient would extrapolate to zero; see Fig. 23.7f for a graphical illustration of $\mathcal{S}$. The calculation of slip length from velocity profiles has some ambiguity, and we refer to work by *Bocquet* et al. [23.71] for the least ambiguous resolution to the definition of the slip length known to us.

Macroscopic experiments are generally well-described by a "no-slip" boundary condition (BC), where $\mathcal{S} = 0$ and the tangential component of the fluid velocity equals that of the solid at the surface. The one prominent exception is contact-line motion, where an interface between two fluids moves along a solid surface. This motion would require an infinite force in hydrodynamic theory, unless slip occurred near the contact line.

As d decreases, any deviation from $\mathcal{S} = 0$ becomes more evident in (23.13). Studies of confined films show

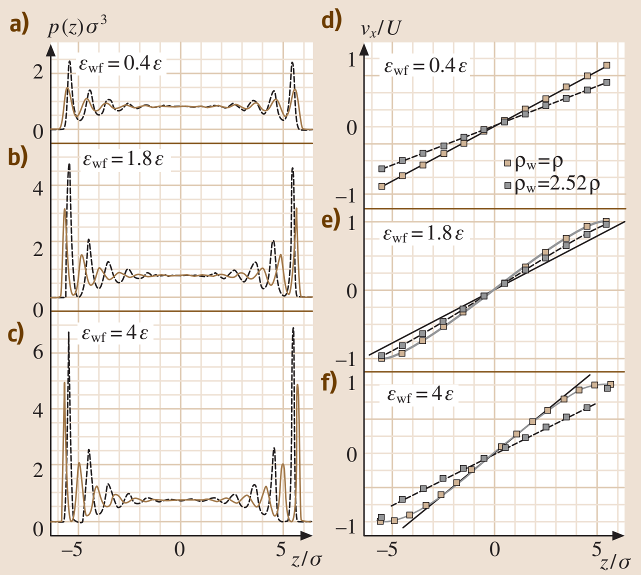

Fig. 23.7a–f Panels **(a)–(c)** show density as a function of position z relative to two walls whose atoms are centered at $z/\sigma = \pm 6.4$. Values of ϵ_{wf} are indicated. *Solid lines* are for equal wall and fluid densities ($\rho = 0.81\sigma^{-3}$), and *dotted lines* are for $\rho_w = 2.52\rho$. *Squares* in panels **(d)–(f)** show the average velocity in each layer as a function of z, and *solid* and *dotted lines* are fits through these values for low and high density walls, respectively. For this data, the walls were moved in opposite directions at speed $U = 1\sigma/t_{\mathrm{LJ}}$. The *dashed line* in **(e)** represents the flow expected from hydrodynamics with a no-slip BC ($\mathcal{S} = 0$). As shown in **(f)**, the slope (*dashed line*) far from the walls is used to define $\mathcal{S}$. (From [23.8] after [23.68])

that slip may occur ($\mathcal{S} > 0$) or layers may stick onto the solid ($\mathcal{S} < 0$) [23.68, 71–73]. These deviations from the no-slip BC have been correlated with structural changes in the fluid near the walls. The first observation was that layering occurs in planes parallel to the walls. It is induced by the sharp cutoff in fluid density at the wall and the pair correlation function $g(r)$ between fluid atoms. An initial fluid layer forms at the preferred wall-fluid spacing. Additional fluid molecules tend to lie in a second layer, at the preferred fluid-fluid spacing. This layer induces a third, and so on. Layering can alter the BC and the viscosity of the confined fluid. As suggested by the contrasting behavior of commensurate and incommensurate solids, the BC depends sensitively on the degree to which the walls imprint their in-plane density modulations into the lubricant.

Some of the trends that are observed in the degree and extent of layering are illustrated in Fig. 23.7 [23.68, 72]. The fluid density is plotted as a function of the distance between walls for a model that is also considered in almost all of the other studies of flow BCs described below. The fluid consists of spherical molecules interacting through a Lennard-Jones potential (23.1) with characteristic energy and length scales ϵ and σ. They are confined by crystalline walls containing discrete atoms. In this case, the walls were planar (001) surfaces of an fcc crystal. Wall and fluid atoms also interact with a Lennard-Jones potential, but with a different binding energy ϵ_{wf}.

The net adsorption potential from the walls can be increased by raising ϵ_{wf}, or by increasing the density of the walls ρ_w, so that more wall atoms attract the fluid. Figure 23.7 shows that both increases lead to increases in the height of the first density peak. The height also increases with the pressure in the fluid [23.73], since that forces atoms into steeper regions of the adsorption potential. The height of subsequent density peaks decreases smoothly with distance from the wall, and only four or five well-defined layers are seen near each wall in Fig. 23.7. The rate at which the density oscillations decay is determined by the decay length of structure in the bulk pair-correlation function of the fluid. Since all panels of Fig. 23.7 have the same conditions in the bulk, the decay rate is the same. The adsorption potential only determines the initial height of the peaks.

As mentioned above, solid surfaces also induce density modulations within the plane of the layers [23.7]. These modulations can be quantified by the 2-D static structure factor at the shortest reciprocal lattice vector Q of the substrate [23.35, 68]. When normalized by the number of atoms in the layer, N_l, this becomes an intensive variable that would correspond to the Debye–Waller factor in a crystal. The maximum possible value, $S(Q)/N_l = 1$, corresponds to fixing all atoms exactly at crystalline lattice sites. In a strongly ordered case such as $\rho_w = \rho$ in Fig. 23.7c, the small oscillations about lattice sites in the first layer only decrease $S(Q)/N_l$ to 0.71. This is well above the value of 0.6 that is typical of bulk 3-D solids at their melting point [23.74] and indicates that the first layer has crystallized onto the wall. The values of $S(Q)/N_l$ in the second and third layers are 0.31 and 0.07, respectively, and atoms in these layers exhibit typical fluid diffusion.

There is some correlation between the factors that produce strong layering and those that produce strong in-plane modulations. For example, chain molecules have several conflicting length scales that tend to frustrate both layering and in-plane order [23.36, 75, 76]. However, the dependence of in-plane order on the density of substrate atoms is more complicated than for layering. When $\rho_w = \rho$, the fluid atoms can naturally sit on the sites of a commensurate lattice, and $S(Q)$ is large. When the substrate density ρ_w is increased by a factor of 2.52, the fluid atoms no longer fit easily into the corrugation potential. The degree of induced in-plane order drops sharply, although the layering becomes stronger (Fig. 23.7). Sufficiently strong adsorption potentials may eventually lead to crystalline order in the first layers and stronger layering. However, this may actually increase slip.

Thompson et al. [23.68] found that all of their results for $\mathcal{S}$ collapsed onto a universal curve when plotted against the structure factor $S(Q)/N_l$. When one or more layers crystallized onto the wall, the same collapse could be applied as long as the effective wall position was shifted by a layer and the Q for the outer wall layer was used. The success of this collapse at small $S(Q)/N_l$ can be understood from a perturbation theory [23.6, 77] developed to quantitatively explain the kinetic friction on adsorbed monolayers [23.3].

One of the surprising features of Fig. 23.7 is that the viscosity remains the same even in regions near the wall where there is pronounced layering. Any change in viscosity would produce a change in the velocity gradient, since the stress is constant. However, the flow profiles in panel (d) remain linear throughout the cell. The profile for the dense walls in panel (f) is linear up to the last layer, which has crystallized onto the wall. From Fig. 23.7 it is apparent that local density variations by at least a factor of seven can be accommodated without a viscosity change.

Bitsanis et al. [23.78] examined the change in viscosity with film thickness. They found that results for film thicknesses $h > 4\sigma$ could be fit using the bulk viscosity for the average density. However, as h decreased below 4σ, the viscosity diverged much more rapidly than any model based on bulk viscosity could explain. These observations were consistent with surface force apparatus (SFA) experiments on nanometer thick films of a wide variety of small molecules [23.4, 5]. Layering in molecularly thin films gave rise to oscillations in the energy, normal force, and effective viscosity as the film thickness decreased. Most films of one to three molecular layers exhibited a yield stress characteristic of solid-like behavior, even though the molecules form a simple Newtonian fluid in the bulk. Deviations from the no-slip condition might cause μ_{eff} to differ from the bulk viscosity by an order of magnitude. However, they could not explain the observed changes of μ_{eff} by more than five orders of magnitude, or the even more dramatic changes by 10 to 12 orders of magnitude in the characteristic viscoelastic relaxation time determined from the shear rate dependence of μ_{eff} [23.5].

More recent experimental and theoretical studies confirm that most fluids exhibit a glass transition when confined to channels less than 3 to 10 molecular diameters [23.33, 79]. *Thompson* et al. [23.34, 36] considered changes in flow with film thickness in a wide range of model systems. As in experiment, they found dramatic divergences in the viscosity and relaxation time that suggested confinement induced a glass transition. Later work by *Baljon* et al. [23.33, 80] indicates that the same glass transition is produced by changes in thickness, pressure, or temperature. In particular, the shear-rate-dependent viscoelastic response of systems of all sizes, pressures, and temperatures can be collapsed onto the same universal curve above the glass transition [23.33].

23.2.6 Submonolayer Films

Physisorbed molecules such as short hydrocarbon chains can be expected to sit on any surface exposed to atmospheric conditions. Even in ultrahigh vacuum, special surface treatments are needed to remove strongly physisorbed species from surfaces. Given the generic observation of glassy behavior in thin films, *Robbins* et al. [23.81] suggested that these submonolayer contaminants could provide a general explanation for the universal observation of static friction. Subsequent simulations have confirmed that physisorbed molecules qualitatively alter tribological behavior. For example, they lead to static friction between two incommensurate walls [23.67, 82] or two disordered walls [23.47] that would slide without friction if they were clean. The theoretical explanation of this effect can be understood from a simple analytical model considered by *Müser* et al. [23.47]. In order to make two surfaces pin, a nonvanishing fraction of the density modulations must share the same period and lock into phase. This happens automatically for commensurate surfaces, but not for incommensurate surfaces unless they are unrealistically compliant. However, a submonolayer of atoms or molecules that does not form strong covalent bonds with the confining solids can simultaneously lock to the density modulations of two incommensurate walls and hence give rise to a finite friction coefficient.

Simulation results for the friction due to physisorbed layers are consistent with a large number of experimental observations. While Amontons's laws state that the friction is independent of the macroscopic area, it is well established that the friction is proportional to the real area of molecular contact A_{real} [23.30, 83, 84]. Experiments where both the friction and molecular area can be measured [23.4, 85–89] show that the local shear stress τ_{s} needed to initiate sliding rises linearly with the normal pressure P:

$$\tau_{\text{s}} = \tau_0 + \alpha P \; . \tag{23.14}$$

This gives a macroscopic friction force $F_{\text{s}} = \tau_0 A_{\text{real}} + \mu L$, where the total load is just $L = PA_{\text{real}}$. This friction law is actually more general than Amontons's laws, reducing to it at large loads and small areas, but explaining why the friction between adhesive surfaces (like tape) does not vanish at zero load.

For a wide range of parameters [23.82, 90], the friction calculated for physisorbed monolayers follows (23.14) up to the gigapascal pressures that are present in real contacts. Moreover, the results are remarkably insensitive to parameters that are not controlled in experiments such as the crystallographic orientation of the solids, the thickness of the adsorbed layer, and the direction of sliding. The only exceptions occur for the unlikely case of commensurate walls [23.90, 91], as discussed further below. Subsequent studies of kinetic friction [23.39, 92] explain both the close relation between experimental values of static friction and the slow variation with velocity. In particular, the logarithmic rate dependence seen in τ for many systems [23.83] can be explained through thermally activated hopping of physisorbed molecules [23.92].

Müser et al. [23.47] have constructed a simple analytical theory that captures many of the results and conclusions from the above simulations. At the pressures of interest, the atoms act like hard spheres. In order to slide, the bounding surfaces must lift up over these spheres against the normal pressure. The situation is analogous to the simple cartoon in Fig. 23.2, but with atomic-scale ramps, and the required stress increases linearly with pressure for the same reason. The offset, τ_0, is related to adhesive interactions between the two surfaces that increase the effective load [23.90]. Unlike the original model of geometrical interlocking, the atomic-scale ramps rearrange during sliding. The dissipation during these rearrangements is responsible for kinetic friction [23.47, 92], which, as we now discuss, is very different for commensurate and incommensurate walls.

The characteristic motion of individual boundary lubricant atoms depends on the relative displacement for commensurate walls, because the distribution of metastable states depends on the relative alignment of the two walls [23.90, 92]. For example, no "pops" of the lubricant atoms are observed when the two commensurate solids are aligned directly above each other. This behavior is in contrast with that for incommensurate surfaces for which the distribution of metastable states does not change as the slider is translated with respect to the substrate. Stressing the analogy to the Prandtl–Tomlinson model, static and kinetic friction are related for incommensurate, boundary-lubricated solids, while they are not necessarily related for commensurate, boundary-lubricated solids. Indeed, commensurate, boundary-lubricated surfaces exhibit a dramatic decrease in friction at the transition from stick-slip motion to smooth sliding, while this dramatic decrease is absent for incommensurate, boundary-lubricated walls [23.39].

It has been noted recently, that dimensionality not only plays an important role for elastic objects, but also for boundary lubricants [23.39]. It was predicted that kinetic friction would be Stokes-like (i. e., vanish linearly with velocity) between incommensurate surfaces, if it were possible to confine lubricant particles into trenches of (sub)nanometer depth and width, so that the motion of the lubricant embedded between the surfaces becomes 1-D. At the same time, however, static friction would remain finite.

The above results suggest that adsorbed molecules and other "third-bodies" may prove key to understanding macroscopic friction measurements. It will be interesting to extend these studies to more realistic molecular potentials and to rough surfaces. To date, realistic potentials have only been used for "between-sorbed" particles between commensurate surfaces. This work is reviewed elsewhere [23.8, 10, 11].

23.3 Stick-Slip Dynamics

The dynamics of sliding systems can be very complex and depend on many factors, including the types of metastable states in the system, the times needed to transform between states, and the mechanical properties of the device that imposes the stress. At high rates or stresses, systems usually slide smoothly. At low rates, the motion often becomes intermittent, with the system alternately sticking and slipping forward [23.30, 32]. Everyday examples of such stick-slip motion include the squeak of hinges and the music of violins.

The alternation between stuck and sliding states of the system reflects changes in the way energy is stored. While the system is stuck, elastic energy is pumped into the system by the driving device. When the system slips, this elastic energy is released into kinetic energy and eventually dissipated as heat. The system then sticks once more, begins to store elastic energy, and the process continues. Both elastic and kinetic energy can be stored in all the mechanical elements that drive the system. The whole coupled assembly must be included in any analysis of the dynamics.

The simplest type of intermittent motion is the atomic-scale stick-slip that occurs in the multistable regime ($\lambda > 1$) of the Prandtl–Tomlinson model. Energy is stored in the springs while atoms are trapped in a metastable state, and converted to kinetic energy as they pop to the next metastable state. This phenomenon is quite general and has been observed in many simulations of wearless friction [23.8, 10], as well as in the motion of atomic force microscope tips [23.1, 2]. In these cases, motion involves a simple ratcheting over the surface potential through a regular series of hops between neighboring metastable states. The slip distance is determined entirely by the periodicity of the surface potential.

Many examples of stick-slip involve a rather different type of motion that can lead to intermittency and chaos [23.93–97]. Instead of jumping between neighboring metastable states, the system slips for very long distances before sticking. For example, *Gee* et al. [23.4] and *Yoshizawa* et al. [23.98] observed slip distances of many microns in their studies of confined films with the

SFA. This distance is much larger than any characteristic periodicity in the potential and varied with velocity, load, and the mass and stiffness of the SFA. The fact that the SFA does not stick after moving by a lattice constant indicates that sliding has changed the state of the system in some manner, so that it can continue sliding even at forces less than the yield stress.

One simple property that depends on past history is the amount of stored kinetic energy. This can carry a system over potential energy barriers even when the stress is below the yield stress. The simplest example is the underdamped Prandtl–Tomlinson model, which has been thoroughly studied in the mathematically equivalent case of an underdamped Josephson junction [23.99]. One finds a hysteretic response function, where static and moving steady-states coexist over a range of forces between F_{min} and the static friction F_{s}. There is a minimum stable steady-state velocity v_{min} corresponding to F_{min}. At lower velocities, the only steady state is linearly unstable because $\partial v/\partial F < 0$ – pulling harder slows the system [23.30, 32]. If the top wall of the Tomlinson model is pulled at an average velocity less than v_{min} by a sufficiently compliant system, it will exhibit large-scale stick-slip motion.

Experimental systems typically have many other internal degrees of freedom that may depend on past history. Macroscopic models normally assume that a single unspecified state variable is relevant and add dynamical rules for this variable to create a phenomenological rate-state model [23.94, 100]. This approach has captured experimental behavior from a wide range of systems, including surface force apparatus experiments [23.38, 101], paper sliding on paper [23.93], and earthquakes [23.102]. However, in most cases, there is no microscopic model for the state variable and its time dependence.

One set of systems where microscopic pictures for the state variable have been considered are confined films in SFA experiments. Confined films have structural degrees of freedom that can change during sliding, and this provides an alternative mechanism for stick-slip motion [23.35]. Some of these structural changes are illustrated in Fig. 23.8, which shows stick-slip motion of a two-layer film of simple spherical molecules. The bounding walls were held together by a constant normal load. A lateral force was applied to the top wall through a spring k attached to a stage that moved with fixed velocity v in the x direction. The equilibrium configuration of the film at $v = 0$ is a commensurate crystal that resists shear. Thus at small times, the center of mass of the top wall remains pinned at $X = 0$. The force grows linearly with time, $F = kv$, as the stage advances ahead of the wall. When F exceeds F_{s}, the wall slips forward. The force drops rapidly because the slip velocity $\dot{X}$ is much greater than v. When the force drops sufficiently, the

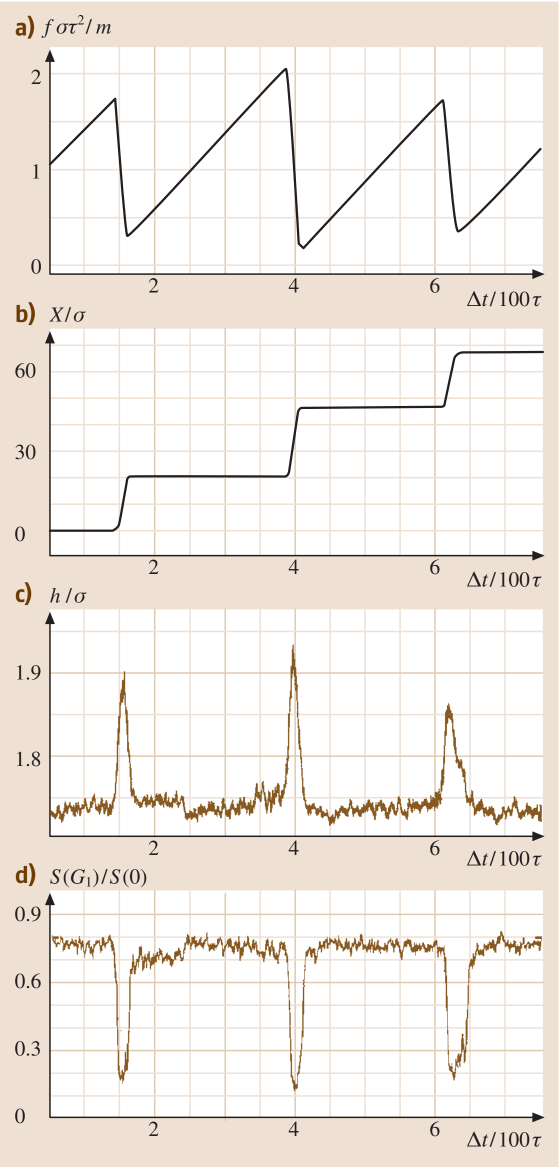

Fig. 23.8a–d Time dependence of **(a)** force per unit area f, **(b)** lateral position of the top wall X, **(c)** wall spacing h, and **(d)** degree of crystallinity $S(G_1)/S(0)$ as a function of time during stick-slip motion of walls confining two layers of spherical molecules. (After [23.35])

film recrystallizes, the wall stops, and the force begins to rise once more.

One structural change that occurs during each slip event is dilation by about 10% (Fig. 23.8c). *Dhinojwala* et al. [23.103] have recently confirmed that dilation occurs during slip in SFA experiments. The increased volume makes it easier for atoms to slide past each other and is part of the reason the sliding friction is lower than F_s. The system may be able to keep sliding in this dilated state as long as it takes more time for the volume to contract than for the wall to advance by a lattice constant. Dilation of this type plays a crucial role in the yield, flow, and stick-slip dynamics of granular media [23.104–106]. In some simulations, the dilation is localized at a single plane – either a wall/film interface or a plane within the film [23.33, 107]. In some cases, there is also in-plane ordering of the film to enable it to slide more easily over the wall. This ordering remains after sliding stops and provides a mechanism for the long-term memory seen in some experiments [23.4, 98, 108].

The degree of crystallinity throughout the film may also change during sliding. Deviations from an ideal crystalline structure can be quantified by the Debye–Waller factor $S(G_1)/S(0)$ (Fig. 23.8d), where G_1 is one of the shortest reciprocal lattice vectors and $S(0) = N$ is the total number of atoms in the film. When the system is stuck, $S(G_1)/S(0)$ has a large value that is characteristic of a 3-D crystal. During each slip event, $S(G_1)/S(0)$ drops dramatically. The minimum values are characteristic of simple fluids that would show a no-slip boundary condition (Sect. 23.2.5). The atoms also exhibit rapid diffusion that is characteristic of a fluid. The periodic melting and freezing transitions that occur during stick-slip are induced by shear, not by the negligible changes in temperature. Shear-melting transitions at constant temperature have been observed in both theoretical and experimental studies of bulk colloidal systems [23.29, 109]. While the above simulations of confined films used a fixed number of particles, *Lupowski* et al. [23.110] found equivalent results at fixed chemical potential.

Very similar behavior has been observed in simulations of sand [23.104], chain molecules [23.33], incommensurate or amorphous walls [23.35], and minimal models with single molecules between surfaces [23.95–97]. These systems transform between glassy and fluid states during stick-slip motion. As in equilibrium, the structural differences between glass and fluid states are small. However, there are strong changes in the self-diffusion and other dynamic properties when the film goes from the static glassy to sliding fluid state. *Buldum* and *Ciraci* [23.111] found stick-slip motion due to periodic structural transformations in the bottom layers of a pyramidal Ni(111) tip sliding on an incommensurate Cu(110) surface.

23.4 Conclusions

The discussion of computer simulations in this chapter aimed to build a general atomistic understanding of tribological processes rather than to elucidate a specific sliding system. The advantage of studying simplified models, as opposed to chemically realistic systems, is that system properties can be varied at will. This allows the role of adhesion, surface curvature, contaminant layers and other factors to be isolated and explored. For example, to explore the effects of adhesion one can compare results from a simulation that includes the long-range attractive interactions between atoms with those from a simulation where all attractive interactions between surfaces are eliminated, leaving all other parameters unchanged. Care should be taken to avoid simplifications that are known to yield atypical results. In particular, the use of commensurate, perfectly aligned, crystalline surfaces has been shown to produce qualitatively different results, even in the presence of a lubricating film! Surface curvature must also be included in a way that properly models stress distributions from the bounding solids.

One central question is how contacting surfaces lock together to produce a static friction force. This is particularly perplexing in the context of many nanotribological systems where friction occurs without significant plastic deformation or wear. Many theories for wearless solid friction assume that surfaces deform elastically to conform to each other, and that atoms advance in a series of elastic instabilities during sliding. However, the simulations discussed above show that simple solids should hardly ever show elastic instabilities on an atomic scale without concurrent plastic flow or material transfer. Despite extensive efforts to design an atomistic system whose static friction mechanism is entirely elastic, no such model could be found as long as the ingredients were: three-dimensional incommensurate solids, flat or smoothly curved contacts, and two-body interactions between all atoms. Under these conditions the relative lateral motion is only opposed by a viscous-like force that goes to zero for small sliding velocities. When interfacial interactions are increased in an attempt to produce elastic instabilities, plastic flow is also observed.

Wearless friction between clean solid surfaces is only found in simulations of very small contacts. The geometric interlocking in such contacts is particularly relevant to AFM experiments. When only a few atoms on the tip make contact, the potential from the substrate does not average to zero. Static friction can then occur without elastic instabilities within the tip. The Prandtl–Tomlinson model can be applied to these systems with the cantilever stiffness replacing the internal stiffness of the material in the tip. As the diameter of the contact increases, due to blunting of the tip or increased load, the friction coefficient produced by geometric interlocking decreases. In the specific case of non-adhesive contacts, the friction between an amorphous tip and a clean crystalline substrate scales as $F \propto L^{2/3}$ and μ vanishes as $L^{-1/3}$. Thus geometric interlocking is not expected to yield significant friction when the contact area contains more than $\sim 10^3$ atoms.

Most contacting solids are separated by contaminant layers, or third bodies, that introduce many additional structural degrees of freedom. The intervening atoms can rearrange into a local free energy minimum for any geometry of the bounding solids without paying the large elastic energy cost that limits rearrangements within these solids. Simulations that include these layers reproduce many details of SFA experiments. They also provide an explanation for Amontons's laws, and exceptions to them, in larger contacts.

Many important problems remain to be addressed by simulations. Of particular interest are the roles of realistic, random surface roughness and subsurface plastic deformation. Studies of these phenomena will require new multiscale approaches that treat long-range elastic and plastic deformations in the solids. Other important problems will require extensions in the time available to simulations. Examples include cold-welding through inter-surface diffusion, wear, and fatigue. Addressing these problems will be a continuing challenge to the computer modeling community.

References

23.1 C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang: Atomic-scale friction of a tungsten tip on a graphite surface, Phys. Rev. Lett. **59** (1987) 1942–1945

23.2 R. W. Carpick, M. Salmeron: Scratching the surface: Fundamental investigations of tribology with atomic force microscopy, Chem. Rev. **97** (1997) 1163–1194

23.3 J. Krim: Friction at the atomic scale, Sci. Am. **275**(4) (1996) 74–80

23.4 M. L. Gee, P. M. McGuiggan, J. N. Israelachvili, A. M. Homola: Liquid to solid transitions of molecularly thin films under shear, J. Chem. Phys. **93** (1990) 1895–1906

23.5 S. Granick: Motions and relaxations of confined liquids, Science **253** (1992) 1374–1379

23.6 M. Cieplak, E. D. Smith, M. O. Robbins: Molecular origins of friction: The force on adsorbed layers, Science **265** (1994) 1209–1212

23.7 M. Schoen, C. L. Rhykerd, D. J. Diestler, J. H. Cushman: Fluids in micropores: I. Structure of a simple classical fluid in a slit-pore, J. Chem. Phys. **87** (1987) 5464–5476

23.8 M. O. Robbins, M. H. Müser: Computer simulations of friction, lubrication and wear. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 717–765

23.9 M. H. Müser, M. Urbakh, M. O. Robbins: Statistical mechanics of static and low-velocity kinetic friction, Adv. Chem. Phys. **126** (2003) 187–272

23.10 J. A. Harrison, S. J. Stuart, D. W. Brenner: Atomic-scale simulation of tribological and related phenomena. In: *Handbook of Micro/Nanotribology*, ed. by B. Bhushan (CRC, Boca Raton 1999) pp. 525–594

23.11 J. A. Harrison, G. T. Gao, R. J. Harrison, G. M. Chateauneuf, P. T. Mikulski: The friction of model self-assembled monolayers. In: *Encyclopedia of Nanoscience and Nanotechnology*, ed. by H. S. Nalwa (American Scientific, Los Angeles 2004)

23.12 M. P. Allen, D. J. Tildesley: *Computer Simulation of Liquids* (Clarendon, Oxford 1987)

23.13 D. Frenkel, B. Smit: *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego 1996)

23.14 J. P. Ryckaert, A. Bellemans: Molecular dynamics of liquid alkanes, Faraday Discuss. **66** (1978) 95–106

23.15 W. Paul, D. Y. Yoon, G. D. Smith: An optimized united atom model for simulations of polymethylene melts, J. Chem. Phys. **103** (1995) 1702–1709

23.16 W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, O. Hahn: Simulation of polymer melts. I. Coarse graining procedure for polycarbonates, Acta Polym. **49** (1998) 61–74

23.17 W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, O. Hahn: Simulation of polymer melts. II. From coarse grained models back to atomistic description, Acta Polym. **49** (1998) 75–79
23.18 P. Flory: *Statistical Mechanics of Chain Molecules* (Hanser, München 1988)
23.19 K. Binder: *Monte Carlo and Molecular Dynamics Simulations in Polymer Science* (Oxford, New York 1995)
23.20 D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, B. Ni, S. B. Sinnot: Second generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons, J. Phys. C **14** (2002) 783–802
23.21 S. J. Stuart, A. B. Tutein, J. A. Harrison: A reactive potential for hydrocarbons with intermolecular interactions, J. Chem. Phys. **112** (2000) 6472–6486
23.22 M. S. Daw, M. I. Baskes: Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals, Phys. Rev. B **29** (1984) 6443–6453
23.23 S. Nosé: Constant temperature molecular dynamics methods, Prog. Theor. Phys. Supp. **103** (1991) 1–46
23.24 T. Schneider, E. Stoll: Molecular dynamics study of a three-dimensional one-component model for distortive phase transitions, Phys. Rev. B **17** (1978) 1302–1322
23.25 G. S. Grest, K. Kremer: Molecular dynamics simulations for polymers in the presence of a heat bath, Phys. Rev. A **33** (1986) 3628–3631
23.26 S. S. Sarman, D. J. Evans, P. T. Cummings: Recent developments in non-Newtonian molecular dynamics, Phys. Rep. **305** (1998) 1–92
23.27 D. J. Evans, G. P. Morriss: Shear thickening and turbulence in simple fluids, Phys. Rev. Lett. **56** (1986) 2172–2175
23.28 W. Loose, G. Ciccotti: Temperature and temperature control in nonequilibrium-molecular-dynamics simulations of the shear flow of dense liquids, Phys. Rev. A **45** (1992) 3859–3866
23.29 M. J. Stevens, M. O. Robbins: Simulations of shear-induced melting and ordering, Phys. Rev. E **48** (1993) 3778–3792
23.30 F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids* (Clarendon, Oxford 1986)
23.31 R. Khare, J. J. de Pablo, A. Yethiraj: Rheology of confined polymer melts, Macromolecules **29** (1996) 7910–7918
23.32 E. Rabinowicz: *Friction and Wear of Materials* (Wiley, New York 1965)
23.33 M. O. Robbins, A. R. C. Baljon: Response of thin oligomer films to steady and transient shear. In: *Microstructure and Microtribology of Polymer Surfaces*, ed. by V. V. Tsukruk, K. J. Wahl (American Chemical Society, Washington 2000) pp. 91–117
23.34 P. A. Thompson, G. S. Grest, M. O. Robbins: Phase transitions and universal dynamics in confined films, Phys. Rev. Lett. **68** (1992) 3448–3451
23.35 P. A. Thompson, M. O. Robbins: Origin of stick-slip motion in boundary lubrication, Science **250** (1990) 792–794
23.36 P. A. Thompson, M. O. Robbins, G. S. Grest: Structure and shear response in nanometer-thick films, Israel J. Chem. **35** (1995) 93–106
23.37 M. Porto, V. Zaloj, M. Urbakh, J. Klafter: Macroscopic versus microscopic description of friction: From Tomlinson model to shearons, Tribol. Lett. **9** (2000) 45–54
23.38 B. N. J. Persson: *Sliding Friction: Physical Principles and Applications* (Springer, Berlin 1998)
23.39 M. H. Müser: Nature of mechanical instabilities and their effect on kinetic friction, Phys. Rev. Lett. **89** (2002) 224301: 1–4
23.40 M. H. Müser: Dry friction between flat surfaces: Wearless multistable elasticity vs. material transfer and plastic deformation, Tribol. Lett. **10** (2001) 15–22
23.41 J. B. Sokoloff: Static friction between elastic solids due to random asperities, Phys. Rev. Lett. **86** (2001) 3312–3315
23.42 B. N. J. Persson, E. Tosatti: Theory of friction: Elastic coherence length and earthquake dynamics, Solid State Commun. **109** (1999) 739–744
23.43 F. F. Abraham, N. Bernstein, J. Q. Broughton, D. Hess: Dynamic fracture of silicon: Concurrent simulation of quantum electrons, classical atoms, and the continuum solid, MRS Bull. **25** (2000) 27–32
23.44 L. Wenning, M. H. Müser: Friction laws for elastic nano-scale contacts, Europhys. Lett. **54** (2001) 693–699
23.45 U. Landman, W. D. Luedtke, N. A. Burnham, R. J. Colton: Atomistic mechanisms and dynamics of adhesion, nanoindentation, and fracture, Science **248** (1990) 454–461
23.46 D. Dowson: *History of Tribology* (Longman, New York 1979)
23.47 M. H. Müser, L. Wenning, M. O. Robbins: Simple microscopic theory of Amontons's laws for static friction, Phys. Rev. Lett. **86** (2001) 1295–1298
23.48 M. Enachescu, R. J. A. van den Oetelaar, R. W. Carpick, D. F. Ogletree, C. J. F. Flipse, M. Salmeron: An AFM study of an ideally hard contact: The diamond (111)/tungsten carbide interface, Phys. Rev. Lett. **81** (1998) 1877–1880
23.49 C. A. J. Putman, M. Igarashi, R. Kaneko: Single-asperity friction in friction force microscope: The composite model, Appl. Phys. Lett. **66** (1995) 3221–3223
23.50 P. Schwarz, O. Zwörner, P. Köster, R. Wiesendanger: Quantitative analysis of the frictional properties of solid materials at low loads. II. Mica and germanium sulfide, Phys. Rev. B **56** (1997) 6987–7000
23.51 L. Prandtl: Ein Gedankenmodell zur kinetischen Theorie der festen Körper, Z. Angew. Math. Mech. **8** (1928) 85–106

23.52 G. A. Tomlinson: A molecular theory of friction, Philos. Mag. Series **7** (1929) 905–939

23.53 G. M. McClelland, S. R. Cohen: Tribology at the atomic scale. In: *Chemistry and Physics of Solid Surfaces*, Vol. VII, ed. by R. Vanselow, R. Rowe (Springer, Berlin, Heidelberg 1990) pp. 419–445

23.54 Y. I. Frenkel, T. Kontorova: On the theory of plastic deformation and twinning, Zh. Eksp. Teor. Fiz. **8** (1938) 1340

23.55 O. M. Braun, Y. S. Kivshar: Nonlinear dynamics of the Frenkel–Kontorova model, Phys. Rep. **306** (1998) 1–108

23.56 M. O. Robbins: Jamming, friction, and unsteady rheology. In: *Jamming and Rheology: Constrained Dynamics on Microscopic and Macroscopic Scales*, ed. by A. J. Liu, S. R. Nagel (Taylor and Francis, London 2000)

23.57 L. W. Bruch, M. W. Cole, E. Zaremba: *Physical Adsorption: Forces and Phenomena* (Oxford, New York 1997)

23.58 F. C. Frank, J. H. van der Merwe: One-dimensional dislocations. I. Static theory, Proc. R. Soc. A **198** (1949) 205–225

23.59 S. Aubry: The new concept of transitions by breaking of analyticity in a crystallographic model. In: *Solitons and Condensed Matter Physics*, ed. by A. R. Bishop, T. Schneider (Springer, Berlin, Heidelberg 1979) pp. 264–290

23.60 M. Hirano, K. Shinjo: Atomistic locking and friction, Phys. Rev. B **41** (1990) 11837–11851

23.61 J. Belak, I. F. Stowers: The indentation and scraping of a metal surface: A molecular dynamics study. In: *Fundamentals of Friction: Macroscopic and Microscopic Processes*, ed. by I. L. Singer, H. M. Pollock (Kluwer, Dordrecht 1992) pp. 511–520

23.62 M. R. Sørensen, K. W. Jacobsen, P. Stoltze: Simulations of atomic-scale sliding friction, Phys Rev. B **53** (1996) 2101–2113

23.63 A. N. Kolmogorov, V. H. Crespi: Smoothest bearings: Interlayer sliding in multiwalled carbon nanotubes, Phys. Rev. Lett. **85** (2000) 4727–4730

23.64 A. Volmer, T. Natterman: Towards a statistical theory of solid dry friction, Z. Phys. B **104** (1997) 363–371

23.65 C. Mak, J. Krim: Quartz-crystal microbalance studies of disorder induced lubrication, Faraday Discuss. **107** (1997) 389–397

23.66 K. W. Jacobsen, J. K. Norskov, M. J. Puska: Interatomic interactions in the effective-medium theory, Phys. Rev. B **35** (1987) 7423–7442

23.67 M. H. Müser, M. O. Robbins: Conditions for static friction between flat crystalline surfaces, Phys. Rev. B **61** (2000) 2335–2342

23.68 P. A. Thompson, M. O. Robbins: Shear flow near solids: Epitaxial order and flow boundary conditions, Phys. Rev. A **41** (1990) 6830–6837

23.69 F. Lancon: Aubry transition in a real material: Prediction for its existence in an incommensurate gold/gold interface, Eur. Phys. Lett. **57** (2002) 74–79

23.70 D. Dowson, G. R. Higginson: *Elastohydrodynamic Lubrication* (Pergamon, Oxford 1968)

23.71 L. Bocquet, J.-L. Barrat: Hydrodynamic boundary conditions and correlation functions of confined fluids, Phys. Rev. Lett. **70** (1993) 2726–2729

23.72 P. A. Thompson, M. O. Robbins: Simulations of contact-line motion: Slip and the dynamic contact angle, Phys. Rev. Lett. **63** (1989) 766–769

23.73 J.-L. Barrat, L. Bocquet: Large slip effect at a nonwetting fluid-solid interface, Phys. Rev. Lett. **82** (1999) 4671–4674

23.74 M. J. Stevens, M. O. Robbins: Melting of Yukawa systems: A test of phenomenological melting criteria, J. Chem. Phys. **98** (1993) 2319–2324

23.75 J. Gao, W. D. Luedtke, U. Landman: Structure and solvation forces in confined films: Linear and branched alkanes, J. Chem. Phys. **106** (1997) 4309–4318

23.76 A. Koike, M. Yoneya: Chain length effects on frictional behavior of confined ultrathin films of linear alkanes under shear, J. Phys. Chem. B **102** (1998) 3669–3675

23.77 E. D. Smith, M. Cieplak, M. O. Robbins: The friction on adsorbed monolayers, Phys. Rev. B **54** (1996) 8252–8260

23.78 I. Bitsanis, S. A. Somers, H. T. Davis, M. Tirrell: Microscopic dynamics of flow in molecularly narrow pores, J. Chem. Phys. **93** (1990) 3427–3431

23.79 A. L. Demirel, S. Granick: Glasslike transition of a confined simple fluid, Phys. Rev. Lett. **77** (1996) 2261–2264

23.80 A. R. C. Baljon, M. O. Robbins: Energy dissipation during rupture of adhesive bonds, Science **271** (1996) 482–484

23.81 M. O. Robbins, E. D. Smith: Connecting molecular-scale and macroscopic tribology, Langmuir **12** (1996) 4543–4547

23.82 G. He, M. H. Müser, M. O. Robbins: Adsorbed layers and the origin of static friction, Science **284** (1999) 1650–1652

23.83 J. H. Dieterich, B. D. Kilgore: Direct observation of frictional contacts: New insights for state-dependent properties, Pure Appl. Geophys. **143** (1994) 238–302

23.84 P. Berthoud, T. Baumberger: Shear stiffness of a solid-solid multicontact interface, Proc. R. Soc. Lond. A **454** (1998) 1615–1634

23.85 A. L. Demirel, S. Granick: Transition from static to kinetic friction in a model lubricated system, J. Chem. Phys. **109** (1998) 6889–6897

23.86 I. L. Singer: Solid lubrication processes. In: *Fundamentals of Friction: Macroscopic and Microscopic Processes* (Elsevier, Amsterdam 1992) pp. 237–261

23.87 B. J. Briscoe: Friction and wear of organic solids and the adhesion model of friction, Philos. Mag. A **43** (1981) 511–527

23.88 B. J. Briscoe, A. C. Smith: The interfacial shear strength of molybdenum disulphide and graphite films, Trans. ASLE **25** (1982) 349–354

23.89 A. Berman, C. Drummond, J. N. Israelachvili: Amontons's law at the molecular level, Tribol. Lett. **4** (1998) 95–101

23.90 G. He, M. O. Robbins: Simulations of the static friction due to adsorbed molecules, Phys. Rev. B **64** (2001) 035413

23.91 J. Gao, X. C. Zeng, D. J. Diestler: Nonlinear effects of physisorption on static friction, J. Chem. Phys. **113** (2000) 11293–11296

23.92 G. He, M. O. Robbins: Simulations of the kinetic friction due to adsorbed molecules, Tribol. Lett. **10** (2001) 7–14

23.93 F. Heslot, T. Baumberger, B. Perrin, B. Caroli, C. Caroli: Creep, stick-slip, and dry-friction dynamics: Experiments and a heuristic model, Phys. Rev. E **49** (1994) 4973–4988

23.94 A. Ruina: Slip instability and state variable friction laws, J. Geophys. Res. **88** (1983) 10359–10370

23.95 M. G. Rozman, M. Urbakh, J. Klafter: Stick-slip motion and force fluctuations in a driven two-wave potential, Phys. Rev. Lett. **77** (1996) 683–686

23.96 M. G. Rozman, M. Urbakh, J. Klafter: Stick-slip dynamics as a probe of frictional forces, Europhys. Lett. **39** (1997) 183–188

23.97 M. G. Rozman, M. Urbakh, J. Klafter, F. J. Elmer: Atomic scale friction and different phases of motion of embedded systems, J. Phys. Chem. B **102** (1998) 7924–7930

23.98 H. Yoshizawa, J. N. Israelachvili: Fundamental mechanisms of interfacial friction. 1. Stick-slip friction of spherical and chain molecules, J. Phys. Chem. **97** (1993) 11300–11313

23.99 D. E. McCumber: Effect of ac impedance on the dc voltage-current characteristics of superconductor weak-link junctions, J. App. Phys. **39**(7) (1968) 3113–3118

23.100 J. H. Dieterich: Modeling of rock friction. 2. Simulation of pre-seismic slip, J. Geophys. Res. **84** (1979) 2169–2175

23.101 A. A. Batista, J. M. Carlson: Bifurcations from steady sliding to stick slip in boundary lubrication, Phys. Rev. E **57** (1998) 4986–4996

23.102 J. H. Dieterich, B. D. Kilgore: Implications of fault constitutive properties for earthquake prediction, Proc. Natl. Acad. Sci. USA **93** (1996) 3787–3794

23.103 A. Dhinojwala, S. C. Bae, S. Granick: Shear-induced dilation of confined liquid films, Tribol. Lett. **9** (2000) 55–62

23.104 P. A. Thompson, G. S. Grest: Granular flow: Friction and the dilatancy transition, Phys. Rev. Lett. **67** (1991) 1751–1754

23.105 H. M. Jaeger, S. R. Nagel, R. P. Behringer: Granular solids, liquids, and gases, Rev. Mod. Phys. **68** (1996) 1259–1273

23.106 S. Nasuno, A. Kudrolli, J. P. Gollub: Friction in granular layers: Hysteresis and precursors, Phys. Rev. Lett. **79** (1997) 949–952

23.107 A. R. C. Baljon, M. O. Robbins: Stick-slip motion, transient behavior, and memory in confined films. In: *Micro/Nanotribology and Its Applications*, ed. by B. Bhushan (Kluwer, Dordrecht 1997) pp. 533–553

23.108 A. L. Demirel, S. Granick: Friction fluctuations and friction memory in stick-slip motion, Phys. Rev. Lett. **77** (1996) 4330–4333

23.109 B. J. Ackerson, J. B. Hayter, N. A. Clark, L. Cotter: Neutron scattering from charge stabilized suspensions undergoing shear, J. Chem. Phys. **84** (1986) 2344–2349

23.110 M. Lupowski, F. van Swol: Ultrathin films under shear, J. Chem. Phys. **95** (1991) 1995–1998

23.111 A. Buldum, S. Ciraci: Interplay between stick-slip motion and structural phase transitions in dry sliding friction, Phys. Rev. B **55** (1997) 12892–12895

24. Mechanics of Biological Nanotechnology

Mechanics of

One of the most compelling areas to be touched by nanotechnology is biological science. Indeed, we will argue that there is a fascinating interplay between these two subjects, with biology as a key beneficiary of advances in nanotechnology as a result of a new generation of single molecule experiments that complement traditional assays involving statistical assemblages of molecules. This interplay runs in both directions, with nanotechnology continually receiving inspiration from biology itself. The goal of this chapter is to highlight some representative examples of the exchange between biology and nanotechnology and to illustrate the role of nanomechanics in this field and how mechanical models have arisen in response to the emergence of this new field. Primary attention will be given to the particular example of the processes that attend the life cycle of bacterial viruses. Viruses feature many of the key lessons of biological nanotechnology, including self assembly, as evidenced in the spontaneous formation of the protein shell (capsid) within which the viral genome is packaged, and a motor-mediated biological process, namely, the packaging of DNA in this capsid by a molecular motor that pushes the DNA into the capsid. We argue that these processes in viruses are a compelling real-world example of nature's nanotechnology and reveal the nanomechanical challenges that will continue to be confronted at the nanotechnology-biology interface.

24.1 Science at the Biology–Nanotechnology Interface 740
24.1.1 Biological Nanotechnology 740
24.1.2 Self-Assembly as Biological Nanotechnology 740
24.1.3 Molecular Motors as Biological Nanotechnology 740
24.1.4 Molecular Channels and Pumps as Biological Nanotechnology 741
24.1.5 Biologically Inspired Nanotechnology.......................... 742
24.1.6 Nanotechnology and Single Molecule Assays in Biology 743
24.1.7 The Challenge of Modeling the Bio-Nano Interface................ 744

24.2 Scales at the Bio-Nano Interface........... 746
24.2.1 Spatial Scales and Structures 747
24.2.2 Temporal Scales and Processes....... 749
24.2.3 Force and Energy Scales: The Interplay of Deterministic and Thermal Forces 750

24.3 Modeling at the Nano-Bio Interface....... 752
24.3.1 Tension Between Universality and Specificity 752
24.3.2 Atomic-Level Analysis of Biological Systems 753
24.3.3 Continuum Analysis of Biological Systems 753

24.4 Nature's Nanotechnology Revealed: Viruses as a Case Study......................... 755

24.5 Concluding Remarks 760

References 761

Nanotechnology is the seat of a broad variety of interdisciplinary activity with applications as diverse as optoelectronics, microfluidics, and medicine. However, one of the richest interdisciplinary areas only now beginning to be harvested is the interface between the biological sciences and nanotechnology. The aim of the present chapter is threefold. First, we wish to illustrate the synergy that exists between biological nanotechnology (e.g., molecular motors, transmembrane pumps, etc.) and biologically inspired nanotechnology (e.g., synthetic proteins, artificial viruses aimed at delivering drugs, biofunctionalized cantilevers, etc.). Secondly, through a series of order-of-magnitude estimates and an associated examination of the units for describing spatial dimensions, temporal processes, forces at the nanoscale, and the energy budget in nanoscale systems, we aim to build an intuitive sense of the workings of nanotechnology in the biological set-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

ting. Finally, through several specific case studies, we hope to illustrate some of the challenges faced in modeling the mechanical processes of biological nanotechnology and show how such challenges have been met thus far and how they might be met in the future.

24.1 Science at the Biology–Nanotechnology Interface

24.1.1 Biological Nanotechnology

Though the innovations leading to the adoption of the expression "nanotechnology" are indeed impressive, the perspective of this chapter is that one need only look inward to the way that our muscles move, to how we digest and synthesize molecules of dazzling complexity, to the way in which we think the thoughts that permit us to fill the shelves of libraries with scientific journals to realize that the greatest nanotechnology of all is that which is revealed in the living world. That is, one of the central thrusts is the idea that the microscopic workings of life offer an inspiring vision of nanotechnology. Clearly, the diversity of the examples of "nanotechnology" seen in the living world can (and do) fill the pages of learned texts. A guided tour of the machinery of the cell can be found in *Alberts* et al. [24.1], while a vision of the cell as an assemblage of "protein machines" has been argued for by *Alberts* [24.2]. The ambition of the present section is to provide several cursory examples of the nanotechnological marvels that power the living realm.

24.1.2 Self-Assembly as Biological Nanotechnology

One of the most intriguing nanotechnological tricks of the living world is the central role played by self-assembly in such systems. Whether we consider the spontaneous assembly of viruses, either in test tubes or in the interior of an infected cell, or the fusion of one membrane-bound region to another through vesicle fusion, spontaneous formation of functional "materials" is a key part of the biological repertoire. To be more concrete, we note that self-assembly in biological systems takes place in a number of different guises. First, the self-assembly of linear assemblies is a key part of the cytoskeletal assembly process with G-actin associating to form actin filaments and, similarly, tubulin monomers joining to form microtubules. This is also the process that is used by certain bacteria such as *Listeria* for locomotion. Both of these examples are described more fully in [24.3]. This same type of process is taken to the next level of sophistication in simple viruses such as tobacco mosaic virus that involve not only the assembly of protein monomers, but also the genetic message in the form of long RNA molecules. A second broad class of self-assembly processes is associated with the formation of containers such as liposomes or viral capsids. In the case of liposomes, lipid molecules such as phosphatidylcholine spontaneously organize in a way that sequesters the hydrophobic tails from the surrounding water. Similarly, protein subunits spontaneously assemble to form viral capsids [24.4]. These structures play a variety of roles in the biological realm, from serving as containers for different macromolecules to providing for concentration gradients of small ions that are maintained and utilized by molecular motors such as ATPsynthase [24.1].

Beyond the simplistic description of self-assembly advanced above, a second key feature of biological self-assembly must also be considered. In particular, a variety of self-assembly processes in the biological realm are templated (or coded). As is well-known, proteins are a hugely versatile class of molecules all based upon the same fundamental building blocks. Interestingly, the enormous diversity of protein action is founded upon 20 distinct amino acid building blocks, and the template for assembling a given protein is carried in the form of messenger RNA, which is then read and used as the basis for protein synthesis by the ribosome. One of the most exciting developments in modern nanotechnology is the attempt to exploit templated self-assembly processes for the purposes of creating new materials. An example in the context of protein-based materials can be found in the article of *van Hest* and *Tirrell* [24.5], where the machinery of the cell has been tricked into incorporating artificial amino acids into synthetic proteins.

24.1.3 Molecular Motors as Biological Nanotechnology

One of the hallmarks of life is change and motion. At the cellular level, such motion is effected through a dizzying variety of mechanisms, most of which when viewed at the molecular level are seen to be the result of the action of molecular motors [24.3]. For example, muscle contraction reflects the coherent action of huge numbers of myosin motor molecules as they march along actin filaments. Similarly, the motion of certain bacteria can be traced to the rotation of a rotary motor embedded in the cell wall, which is attached to filaments known as flagella [24.6]. In a similar vein, just as our modern society is replete with examples of systems aimed at allowing for communication and transport between widely separated geographic locations, so, too, has the living world had to answer these same challenges. As will be shown in the section of this chapter aimed at making estimates of various scales and processes in nanotechnology, one of the key mechanisms for communication and transport is diffusion. That is, a chemical concentration at one place can make its presence known at a distance x with a characteristic time, $t_{\text{diffusive}} \approx x^2/2D$, where D is the diffusion constant. However, as will be shown later in the chapter, there are many cellular processes that cannot wait as long as $t_{\text{diffusion}}$. As a result, a host of molecular motors and an associated transport system (the elements of the cytoskeleton) permit active transport. For example, kinesin is a motor molecule that transports material along long, relatively rigid polymeric assemblies known as microtubules [24.1].

Even more incredible are rotary machines like those associated with ATPsynthase and bacterial flagella. ATPsynthase is a membrane-embedded machine that rotates and, in so doing, synthesizes new ATP molecules (adenosine triphosphate, energy currency of the cell) [24.1, 7–9]. Similarly, the bacterial flagellar motor rotates with the result that the attached bacterial flagellum rotates and thereby induces motion of the cell [24.3, 6]. The exquisite details of the construction of these devices are themselves breathtaking. The rotary motors in bacteria are constructed from several components that are much like a rotor and stator and perform periodic motions by deriving energy from the flow of protons across a membrane [24.6]. These rotary motors are very powerful, as evidenced by F_1-ATPsynthase, which can generate a torque large enough to rotate a molecule of actin 100 times its own length [24.10]. In addition, bacterial motors have another layer of sophistication in that stimuli from the external environment, which are sensed by the bacteria through the pores on its membrane, can change the direction of rotation of the motor in a process known as chemotaxis [24.3]. Feedback and signal transmission are implemented in engineering devices by means of complex circuitry, whereas in the living world these functions are often accomplished by means of chemistry and conformational changes in large molecules.

All of these molecular machines involve a rich interplay between chemistry, thermodynamics, and mechanics. From a structural perspective, most molecular motors are proteins with different subunits performing different functions [24.11]. Often there is some region within the motor where chemical energy is derived from the hydrolysis of adenosine triphosphate (ATP). This chemical energy is then converted to mechanical energy through a conformational change in the protein. These examples serve to call the reader's attention to the importance and variety of motor molecules found throughout the living world and which almost any sensible definition of nanotechnology would have to include as particulary sophisticated examples.

24.1.4 Molecular Channels and Pumps as Biological Nanotechnology

From a structural perspective, one of the most intriguing features of cellular systems is their division into a number of separate membrane-bound compartments. We have already touched upon this compartmentalization in the context of self-assembly, and note that the role of membranes and the proteins bound within them is described clearly in *Alberts* et al. [24.1]. The presence of such compartments, often marked by large concentration gradients with respect to the surrounding medium, hints at another nanotechnological wonder of the living world, namely, the presence of a wide range of transmembrane channels that mediate the exchange of material between these different compartments. Certain passive versions of these channels are gated by various mechanisms such as the arrival of signaling molecules or tension in the membrane within which they are found [24.12–14]. Once the channel is in the open state, ions pass through passively, by diffusion. Active versions of ion channels that transfer ions such as Na^+ and Ca^{2+} are similarly critical for the functioning of a cell, both in the case of unicellular and multicellular organisms. For example, in the case of Na^+ ions, typical concentrations within the cell can be as much as a factor of 10–20 less than those in the extracellular milieu. Such concentration gradients imply the need for sophisticated active "devices" which can do work against such gradients. One of the most remarkable of machines of this type is the Na^+–K^+ pump [24.1]. This machine is powered by hydrolysis of ATP (i. e., the consumption of ATP fuel), and it can pump ions up a potential gradient. In particular, this pump hydrolyzes ATP and pumps Na^+ ions *out* of the cell against a very steep concentration gradient while pumping K^+ ions *in* again against a steep gradient.

The point of this brief discussion has been to illustrate the first of several perspectives that we will bring to bear on the question of the biology-nanotechnology interface. Thus far, we have noted that nature is replete with examples of macromolecules and macromolecular assemblies that perform nanotechnological tasks and in this capacity serve as examples of biological nanotechnology. Next, we wish to examine the ways in which biological phenomena can inspire nanotechnology itself.

24.1.5 Biologically Inspired Nanotechnology

As noted in the beginning of this chapter, biological systems are nanotechnologically relevant for several reasons. First, as shown above, the living world is full of examples of nanotechnological devices. However, a second key point is that nanotechnology has been driven and inspired by the example of biological systems and the need (for example, in medicine) to influence biological systems at the scale of a single cell. In addition, preliminary steps have been taken to harness the nanotechnology of biological systems and use it to perform useful functions.

One compelling example of a proof-of-principle, biologically inspired device emerged from work aimed at exploring the function of ATPsynthase, the rotary device already described above. Fluorescently labelled filamentous proteins from the cytoskeleton, known as actin, were attached to the putative rotary component of the F_1 subunit. The point of this exercise is that such filaments are observable using light microscopy. It was then observed that the long actin filament rotated like a propeller when the ATPsynthase performed ATP hydrolysis [24.15].

A second example, also involving ATPsynthase, is suggested by a set of beautiful experiments done by *Racker* and others [24.16, 17]. The idea is illustrated schematically in Fig. 24.1. Two different protein machines are "reconstituted" in an artificial membrane-bound region known as a liposome. One of these devices is known as bacteriorhodopsin and has the capacity to pump hydrogen ions when it is exposed to light. Since both the bacteriorhodopsin and ATPsynthase are embedded in the same membrane, the ATPsynthase can then exploit the light-induced proton gradient to perform ATP synthesis. Again, these experiments were undertaken not for their role as possible devices, but rather to probe the nature of various molecular machines. Nevertheless, we view them as a provocative demonstration of both the manipulation and use of such machines in artificial environments. As such, they provide an inspiring vision of the possibilities for biologically inspired nanotechnology.

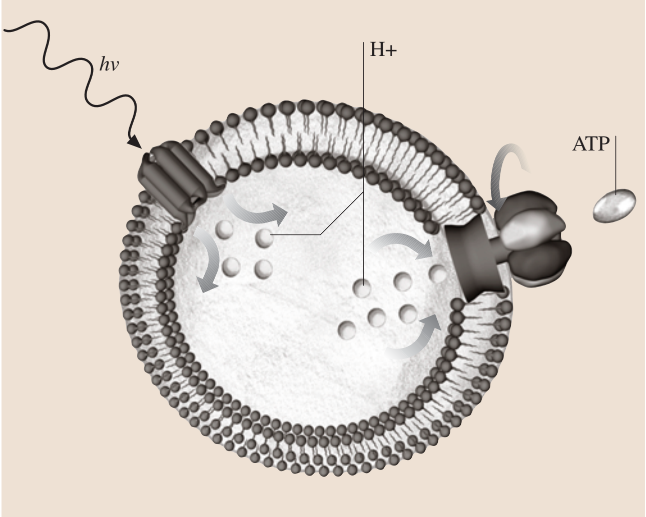

Fig. 24.1 Schematic of the experiment in which bacteriorhodopsin (*top of figure*) and ATPsynthase (*right of figure*) are artificially reconstituted in a liposome, and they act in unison to produce ATP molecules

Another fascinating example of biologically inspired nanotechnology is that of biofunctionalized cantilevers (see Fig. 24.2). A typical example is provided by the experiments of *Wu* et al. [24.18], who have demonstrated the use of a biofunctionalized cantilever as a scheme for detecting small concentrations of biologically interesting molecules such as prostrate-specific antigen and single-stranded DNA. One surface of the cantilever is coated with an antibody, and then it is placed in environments containing different concentrations of the antigen. The key ideas from a mechanical perspective are: i) the difference in surface energy between the top and bottom surfaces of the cantilever induces spontaneous bending and ii) the binding of molecules of interest to target molecules initially present on the surface leads to surface energy differences and bending that can be detected by optical means. In this way the concentration of the molecules of interest can be measured. The specificity and sensitivity of the method makes it viable for use in the laboratory, as well as for commercial purposes. We return to this example in our discussion of the modeling challenges posed by problems at the interface between biology and nanotechnology.

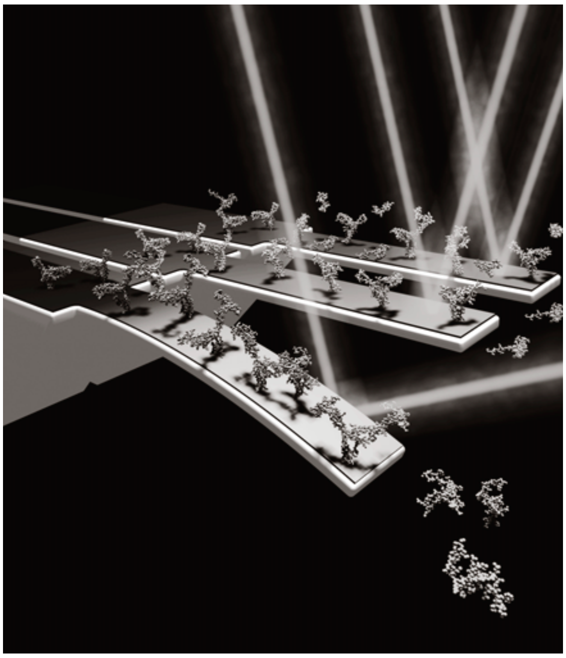

Fig. 24.2 Schematic of the use of biofunctionalized cantilevers as a tool for detecting molecules of biological interest (figure courtesy of Arun Majumdar, Berkeley)

24.1.6 Nanotechnology and Single Molecule Assays in Biology

One of the key refrains of the nanotechnological era is Feynman's quip that "there is plenty of room at the bottom" [24.19]. The benefits of miniaturization are evident at every turn in applications ranging from our cars to our computers. Associated with the development of the inspiring new techniques made possible by nanotechnology has been the emergence of a host of scientific opportunities. One of the arenas to benefit from these new techniques is biology. As scientists and engineers have taken the plunge to the nanotechnological "bottom" foreseen by Feynman, opportunities have constantly arisen to manipulate biological systems in ways that were heretofore unimagined, culminating in a new era of single molecule biology. Single molecule experiments have, in fact, presented us with a view of Feynman's "room at the bottom" as being filled with very complicated machines whose functioning makes life possible.

Single molecule assays complement statistical/collective studies involving a large number of molecular actors by revealing the prominent role of fluctuations at the sub-cellular scale. For example, a photospectrometer measures the optical response of a huge collection of molecules, whereas optical tweezers pull on a single molecule of DNA and enable us to follow the changes in conformation or the breakage of bonds. In fact, experimental methods are so advanced that it is now possible to manipulate a single molecule even as we watch it on a screen as it jiggles around in different conformations. In what follows, we give several examples of how nanotechnology has reached out to help create single molecule biology and in the process has led to the advent of new quantitative opportunities to investigate biological systems.

Atomic-Force Microscopy

One of the tools that has revolutionized nanotechnology, in general, and single molecule biology, in particular, is the atomic-force microscope. The AFM has helped create the field of single molecule force spectroscopy [24.20]. We note that mechanics has a long tradition of using force-extension data (much like the electrical engineer uses current-voltage data) to probe the inner workings of various materials. It is now possible to apply forces of known magnitude on a macromolecule and study how it deforms under the force. This furnishes structural information and provides insights into the energy landscape the molecule needs to navigate as it undergoes force-induced conformational changes. The energy of deformation associated with such molecules is primarily determined by weak forces such as hydrogen bonds, van der Waals contacts, and hydration effects. On a more philosophical note, these experiments force us to think in terms of forces and not energy, complementing the traditional views held in molecular biology, and they can lead to many new insights about the relation between structure and function in proteins, polynucleotides, and other macromolecular entities.

The AFM has been used in a wide variety of single molecule experiments on many of the key classes of molecules found in the living world, including nucleic acids, proteins, and carbohydrates. One fascinating example is the use of atomic force microscopy to examine the mechanical properties of the muscle protein titin [24.21]. The experiment is illustrated schematically in Fig. 24.3, which shows that there is a series of force-extension signatures (increasing force and extension followed by a precipitous load drop) that correspond to the unbinding of the individual domains that make up this protein.

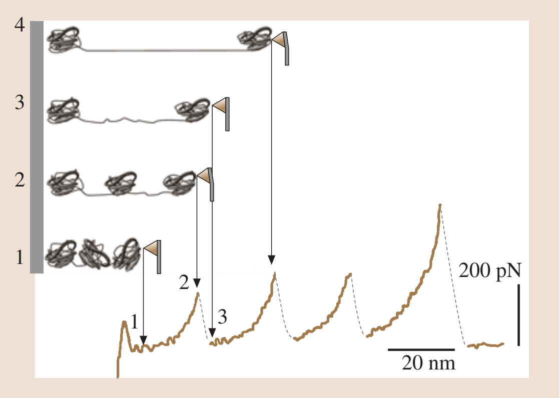

Fig. 24.3 Schematic illustrating the way that the AFM has been harnessed as a nanotechnological analogue of the Instron machine for the measurement of the force-extension properties of single molecules. Four snapshots in the life history of a globular protein subject to loading are shown, as well as the measured force-extension curve. Each sawtooth corresponds to the unfolding of a single protein domain. (Figure adapted from [24.22])

Optical Tweezers

Another instrument that has been used with great success in the realm of biophysics for the purposes of performing nanomechanical measurements on macromolecules and their assemblies is the optical tweezer [24.23, 24]. While the AFM is relatively stiff and applies large forces (on the order of 100–1,000 pN), optical tweezers are compliant and can measure smaller forces (on the order of 0.1–10 pN). Some of the most interesting experiments performed with optical tweezers concern the functioning of molecular motors, which by themselves are marvels of nanotechnology. For example, *Svoboda* et al. [24.25] attached kinesin to an optically trapped bead and observed its movement along a microtubule. An example of the type of data to emerge from such experiments is shown in Fig. 24.4. One of the conclusions to emerge from such experiments is that kinesin can exert forces on the order of 5–7 pN before it stalls. Such experiments also permit an examination of the effect of changing the concentration of ATP on the functioning of kinesin [24.26]. Similar experiments have been performed on RNA polymerase as it advances along DNA to deduce not only the stall force, but also its velocity as a function of the constraining force [24.27]. Such measurements provide a mechano-chemical basis of biological function and go a long way in revealing connections between chemical kinetics and mechanical processes at the molecular level.

As noted in the abstract, one of the most fascinating examples of the biology-nanotechnology interface is that of bacterial viruses, known as bacteriophages. The life cycle of a large class of bacteriophages is characterized by self-assembly processes that lead to the formation of the protein shell of the virus followed by active packaging of the viral DNA within this shell by a motor. The structure of this so-called "viral portal motor" has recently been solved using X-ray crystallography [24.28]. In a recent experiment, *Smith* et al. used optical tweezers to study the characteristics of the DNA packaging process of the ϕ-29 bacteriophage [24.29]. One of the conclusions of this experiment is that the motor has to act against an increasing resistive force as more and more of the DNA is packed inside the capsid. From a quantitative perspective, this experiment yields the force and the rate of packing as a function of the fraction of the genome packed. It is also important to note that bacteriophages are not only an obscure subject of quiet enquiry, but are also the basis of a huge range of cloning products (see, for example, the lambda ZAP vectors of Stratagene) used for doing experiments with recombinant DNA, and more generally, viruses are being explored as the basis of gene therapy.

Our discussion thus far has been aimed at providing a rough overview of the vast landscape that sits at the interface between biology and nanotechnology. It is hoped that the few representative case studies set forth above suffice to illustrate our basic thesis, namely, that biological nanotechnology represents nanotechnology at its best.

24.1.7 The Challenge of Modeling the Bio-Nano Interface

As highlighted in the previous two subsections, there have been huge advances at the interface between nanotechnology and biology. We have argued that there are two distinct representations of the interface between biology and nanotechnology, and each has its own associated set of modeling challenges. The argument of the present discussion is that another key part of the infrastructure that must attend these developments is that associated with the modeling of these systems. One of the intriguing ways in which modeling at the biology-nanotechnology interface is assuming greater importance is that with increasing regularity, experimental data on biological systems is of a quantitative character. As a result, the models that are put forth to greet these experiments must similarly be of a quantitative character [24.30].

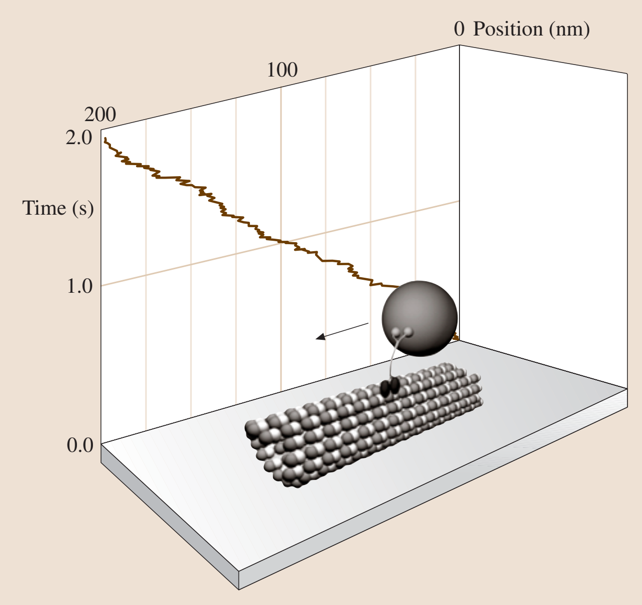

Fig. 24.4 Schematic illustrating the use of optical tweezers to measure the speed of the molecular motor kinesin in its journey along a microtubule

As an example of the type of modeling challenges that must be faced in contemplating the types of problems described above, we return to the example of biofunctionalized cantilevers as a problem in nanomechanics [24.18, 31]. The basic physics behind the use of biofunctionalized cantilevers as sensors is a competition between elastic bending energy and the surface free energy difference between the upper and lower faces of the cantilever. The face with the lower free energy per unit area tends to increase its area by bending the cantilever. The amount of bending, on the other hand, is limited by the elastic energy cost. The utility of this device derives from the fact that the difference in the surface free energy is affected by specific binding of target molecules to probe molecules that are initially deposited on one side of the cantilever.

To provide a quantitative model of the biofunctionalized cantilever, we construct an energy functional that takes into account the elastic energy of beam bending and the surface energy. Both contributions to the total energy can be written as functionals of $u(x)$, the deflection of the cantilever. Note that in the case in which the two surfaces have the same free energy per unit area, $\gamma_{\text{up}} = \gamma_{\text{down}}$, the equilibrium configuration corresponds to $u(x) = 0$. The case of interest here is that in which the two surfaces have different energies.

We recall that the energy associated with beam bending is of the form [24.32]

$$E_{\text{bend}}\left[u(x)\right] = \frac{EI}{2}\int_0^L \left[u''(x)\right]^2 \mathrm{d}x\,, \tag{24.1}$$

where E is Young's modulus, L is the length of the beam, and $I = wt^3/12$ is the areal moment of inertia; t and w are the beam thickness and width. The main approximation we have made in writing (24.1) is that the cross section of the beam remains unchanged by the bending process. The surface contribution to the total energy is associated with the changes of the areas of the upper and lower surfaces by virtue of the beam deforming. In particular, we have

$$\begin{aligned} E_{\text{surf}}[u(x)] = {} & \gamma_{\text{up}} w \left[L - \frac{t}{2}\int_0^L u''(x)\,\mathrm{d}x\right] \\ & + \gamma_{\text{down}} w \left[L + \frac{t}{2}\int_0^L u''(x)\,\mathrm{d}x\right], \end{aligned} \tag{24.2}$$

where the terms in parentheses are the arc lengths along the top and bottom surface of the beam, respectively. The physical content of this functional is the idea that if $u'' > 0$ (concave upward) then the upper surface area will shrink while the lower surface area will increase. Neglecting uninteresting constant terms, the total energy functional is

$$\begin{aligned} E_{\text{tot}}[u(x)] = {} & \frac{EI}{2}\int_0^L \left[u''(x)\right]^2 \mathrm{d}x \\ & + \frac{tw}{2}(\gamma_{\text{down}} - \gamma_{\text{up}})\int_0^L u''(x)\,\mathrm{d}x\,. \end{aligned} \tag{24.3}$$

Our goal is to find the displacement profile $u(x)$ that yields the minimum value for the total energy. The general mathematical framework for effecting this minimization is the calculus of variations. For the functional in (24.3), a more direct route to the result can be obtained by completing the square. Namely, we note that the E_{tot} involves a term quadratic in $u''(x)$ and a second term that is linear in the same function. As a result, the total energy can be rewritten in the form

$$E_{\text{tot}}[u(x)] = \frac{EI}{2}\int_0^L \mathrm{d}x \left[u''(x) + \frac{tw\Delta\gamma}{2EI}\right]^2 , \qquad (24.4)$$

where we have introduced $\Delta\gamma = \gamma_{\text{down}} - \gamma_{\text{up}}$, and we neglect an uninteresting constant term. Clearly the $u(x)$ that minimizes the energy is one for which the integrand vanishes everywhere, or

$$u''(x) + \frac{tw\Delta\gamma}{2EI} = 0 . \qquad (24.5)$$

At this point, we are left with a standard differential equation whose solution, given the boundary conditions $u(0) = 0$ and $u'(0) = 0$, is

$$u(x) = -\frac{tw}{4EI}\Delta\gamma x^2 . \qquad (24.6)$$

The physical meaning of this solution is hinted at by observing that if $\gamma_{\text{down}} > \gamma_{\text{up}}$ then the beam curves downward with the result that the area of the lower surface is reduced.

Measurements of *Wu* et al. [24.31] indicate an upward cantilever deflection when single stranded DNA (ssDNA), up to 20 nucleotides in length, hybridizes with complementary strands of ssDNA, which were deposited initially so as to functionalize the beam. This effect might be attributed to the difference in elastic properties of ssDNA and its double stranded counterpart. Namely, under physiological conditions, ssDNA has a persistence length equal to two nucleotides, while dsDNA is much stiffer (due to hydrogen bonding and base stacking interactions between the two strands) and has a persistence length of 150 nucleotides. Therefore, deposition of the flexible ssDNA molecules initially leads to bending of the cantilever downward due to entropic repulsion between the ssDNA chains. After hybridization, rigid dsDNA strands are formed, there is no longer any entropy to be gained by increasing the area of the top surface, and the beam bends back upwards. Remarkably, *Wu* et al. demonstrate that their biofunctionalized cantilever is sensitive to ssDNA differing in length by a single nucleotide!

To gain further insight into the physics of the biofunctionalized cantilever, it is instructive to examine some quantitative aspects of the experiment using (24.6). Namely, *Wu* et al. find a deflection of $u(L) = 12\,\text{nm}$ when a 20-nucleotide strand of ssDNA hybridizes with a 20-nucleotide-long complementary target. Using the quoted numbers for the cantilever, $E = 180\,\text{GPa}$, $L = 200\,\mu\text{m}$, $t = 0.5\,\mu\text{m}$, $w = 20\,\mu\text{m}$, leads to $\Delta\gamma = 4.5\,\text{fJ}/\mu\text{m}^2$. From this result we can estimate the change in surface free energy due to beam bending,

$$\Delta E_{\text{surf}} \approx wL\Delta\gamma . \qquad (24.7)$$

We find $\Delta E_{\text{surf}} = -18\,\text{pJ}$, which, given the quoted areal chain density of $6\times 10^{12}\,\text{cm}^{-2}$, leads to a decrease in free energy of 75 pN nm, or $18\,k_{\text{B}}T$, per chain. This is comparable to the entropy of a 20 nt ssDNA, which is on the order of $10\,k_{\text{B}}$, lending support to the idea that entropic repulsion between ssDNA strands is implicated in cantilever bending.

While this example gives a feel for the way in which quantitative models have been put forth to respond to biologically inspired nanotechnologies, the remainder of the chapter will emphasize attempts to construct nanomechanical models of biological nanotechnology itself. We turn first to a discussion of the various scales that arise in thinking about the biology-nanotechnology interface and then conclude with several modeling examples.

24.2 Scales at the Bio–Nano Interface

Every scientific discipline has a preferred set of units that lends itself to building intuition concerning the system at hand. For example, an astronomer thinks of distances between stars in light years, not kilometers. Though most of us have an intuitive sense of the meaning of a kilometer, by the time we add more than six zeros, all intuition is lost. At terrestrial scales, we talk of distances between cities in terms of the flying time or the driving time between them and usually not in terms of hundreds of miles. Similar choices must be faced in the biological setting. For example, a biologist might characterize the complexity of an organism by the size (in kilo basepairs) of the genome and not by the organism's physical size. The aim of the present section is to highlight some of the key scales and units that reveal themselves at the interface between biology and nanotechnology. Indeed, we go further and assert that as yet we are still in the process of searching for the most suitable units to characterize the biology-nanotechnology interface. Although our attempt to determine such units and scales might involve seemingly complicated interconversions, such as measuring distances in terms of time (via diffusion), or measuring concentrations in terms of distances, we hold that the approximate *numerical* characterization of the scales of interest is of crucial importance to the endeavor of considering nanomechanics at the biology-nanotechnology interface. In particular, the right choice of units can assist us in building intuition about these systems. Our goal in this section is to emphasize the scales in length, time, force, energy, and power that are relevant in contemplating the nanomechanics of biological systems.

24.2.1 Spatial Scales and Structures

We begin with a discussion of the length scales that arise in contemplating nanomechanics at the biology-nanotechnology interface. In this case, the prefix *nano* leads justifiably to a consideration of the nanometer as one possible choice as the fundamental unit of length. However, to prepare ourselves for the question of how best to describe the dimensions of the spatial structures of interest here, it is important to consider the *hierarchy* of length scales that arise in the nano-bio arena. After examining this hierarchy of structures, we reformulate these length scales in terms of the volumes of these structures measured in units of the volume of a typical bacterial cell, and conclude the present section with a discussion of the way that chemical concentrations can also be interpreted as determining a length scale.

As noted above, a first step in developing intuition concerning the spatial scales found at the biology-nanotechnology interface is through reference to the hierarchy of scales and structures that arise in this arena. The shortest distance in this hierarchy of scales that will interest us is that associated with the size of individual atoms. We recall that the size of a hydrogen atom is roughly 0.1 nm. The scale characterizing the linkage between atoms is that of typical bond lengths that range from roughly 0.1–0.3 nm. A step further in this hierarchy brings us to the basic building blocks of the biological world such as amino acids, nucleotides, individual sugars, lipid molecules, etc. A typical length scale that characterizes these building blocks is the nanometer itself. For example, we have already made reference to the importance of lipid-bilayer membranes in bounding different regions of the cell. Phosphatidylcholine is one of the molecular building blocks of many such membranes. It has a polar head group and a hydrocarbon tail with an overall dimension of 2–3 nm. Similarly, the dimensions of single amino acids and nucleotides are on the order of 1 nm as well.

As is well-known, individual nucleotides are assembled to form nucleic acids such as DNA, amino acids are assembled to form proteins, sugars combine to form polysaccharides, and lipids self-assemble to form membranes. One way of estimating the size of the resulting molecules is by taking the scale of the individual units and scaling up with the number of such units. The various molecular actors of relevance to the present discussion can also be characterized by a length, scale known as the persistence length, which gives a rough description of the length over which the molecule behaves as a stiff rod. For example, the persistence length of DNA is ≈ 50 nm, while that of the cytoskeletal filaments is in the range of 15 μm for actin and 6 mm for microtubules [24.11]. We note that in our later discussions of viruses as a profound example of biological nanotechnology, the ratio of the persistence length of DNA to the size of the viral capsid (the container within which the DNA is packaged) will serve as a measure of the energetic cost of packing the DNA within the capsid and will signal the need for molecular motors to take active part in the packaging process.

The next scale above that of the various macromolecular building blocks are assemblies of such molecules in the form of various molecular machines which are some of the most compelling examples of nature's nanotechnology. The ability to begin formulating mechanistic models of such machines is founded upon key advances in both X-ray crystallography and cryo-electron microscopy [24.33, 34]. Examples of such machines and their associated dimensions include: the machine responsible for making the message carried by DNA readable by the protein synthesis machinery, namely, RNA polymerase (≈ 15 nm) [24.35], the machine that produces ATP, the energy currency of the cell, namely, ATPsynthase (≈ 10 nm) [24.9], and the machine that carries out protein synthesis, namely, the ribosome (≈ 25 nm) [24.36]. A second class of assemblies of particular relevance to the present article are viruses, representative examples of which include lambda phage (≈ 27 nm), tobacco mosaic virus (≈ 250 nm in length), and the HIV virus (≈ 110–125 nm) [24.37].

From the standpoint of cellular function, the next level of structural organization is associated with the various organelles within the cell, structures such as the cell nucleus (3–10 μm), the mitochondria that serve as the power plant of the cell (≈ 1,000 nm), the Golgi apparatus wherein modifications are made to newly synthesized macromolecular components (≈ 1,000 nm), and so on [24.1]. These organelles should be thought of as factories in which many molecular motors of the type described earlier do their job simultaneously. At larger scales yet, and constituting a higher level of overall organization, life as nanotechnology is revealed as cells themselves, the fundamental unit of life that is self-replicating and self-sustaining. We will make special reference to one particular bacterial cell, namely, *E. coli* with typical dimensions of 1 µm. This should be contrasted with a typical eucaryotic cell such as yeast, which has linear dimensions roughly a factor of 10 larger than the *E. coli* cell.

We note that the various biological structures described above should be seen with reference to the sorts of man-made structures to which they are interfaced. For example, earlier in the chapter we mentioned the importance of optical tweezers as a means of communicating forces to macromolecules and their assemblies. Typical dimensions for the optical beads used in optical tweezers are on the scale of 500–1,000 nm. We note that though such dimensions are characteristic of organelles, they are much larger than the individual molecules they are used to study. A second way in which individual macromolecules are communicated with is through small tips such as those found on an AFM. In this case, the size of the tip can be understood through reference to its radius of curvature, which is typically of the order of 50 nm.

Our discussion thus far has centered on the use of a single characteristic length to describe the spatial extent of biological structures. In our quest to develop intuition about the typical spatial scales found at the biology-nanotechnology interface we note that a second way to evaluate such scales is through reference to the typical *volumes* of the various structures of interest. As noted in the introduction to this section, we claim that a key way to develop intuition is by making sure to use appropriate and revealing units. For the present purposes, we argue that one useful unit of volume is that of an *E. coli* bacterium, which if idealized as a cylinder with diameter 1 µm and height 2 µm, has a volume of $\approx 1.5 \times 10^9$ nm^3. In particular, we will measure the sizes of the various entities described above in terms of how many of them can fit into a single *E. coli* bacterium. Our estimates of the comparative sizes of various structures of interest are given in Table 24.1. Note that we do not claim that this is literally how many of each of these entities are found in an *E. coli* bacterium, but rather we seek to give an impression of the relative volumes of some of the different structures that arise when contemplating biological nanotechnology.

Table 24.1 Sizes of entities in comparison to the size of *E. coli*

Entity	Size (nm)	# fitting in an *E. coli*
Amino acid	1	5.0×10^9
Nucleotide	1	5.0×10^9
Monosaccharide	1	5.0×10^9
Phosphatidyl choline	2–3	2.0×10^8
Proteins	5–6	3.0×10^7
Ribosome	20	1.0×10^6
Mitochondria	500–1,000	3–4
T4 phage	100	5,000

The idea of counting the number of molecules that fit into a given volume is actually quite standard. This is exactly what chemists do when they invoke the notion of concentration: measuring the strength of an acid or base using pH amounts to expressing the concentration of H^+ ions in solution. Similarly, when we refer to the molarity of a given solution, it is a statement of how many copies of a given molecule will be found per liter of solution. These ideas are pertinent for our examination of the scales that arise in contemplating biological nanotechnology. As an example, we consider the action of molecular pumps such as that which maintains the concentration gradient in Ca^{2+} ions between the cellular interior and the extracellular medium. Note that this is described in detail in chapter 12 of *Alberts* et al. [24.1]. The Ca^{2+} pump is responsible for insuring that the intercellular concentration of Ca^{2+} ions is 10^{-7} smaller than that in the cellular exterior. One explanation for this low a concentration of Ca^{2+} in the cell is the role played by these ions in signaling other activities. The low concentration might serve as a scheme for increasing the signal-to-noise ratio. The reported intracellular concentration of such ions is roughly 10^{-7} mM. It is interesting to ask how many Ca^{2+} ions this corresponds to in a eucaryotic cell. The volume of a eucaryotic cell measuring 20 µm in diameter is about 4×10^{-12} liters, which translates into ≈ 250 Ca^{2+} ions in the cytoplasm. For the case of *E. coli*, which has a volume approximately 1,000 smaller, this concentration would correspond to one ion for every four cells. This number gives us far greater insight than the standard molar (M = mol/liter) method of expressing concentrations. In a similar vein, we argue that distances between individual ions/molecules in solution are perhaps a better way of thinking about pH and molarity when using them in the context of biological nanosystems. To this end, Fig. 24.5 shows a plot of pH and molarity versus average distance d in nm between individual ions/molecules. Note that, in addition, we have also translated these concentrations into number of copies per *E. coli* cell and the time for an ion to diffuse over the average separation distance. For the time estimate we make use of the typical diffusion constant for small ions in water at 25 °C, $D = 2{,}000\,\mu\text{m}^2/\text{s}$.

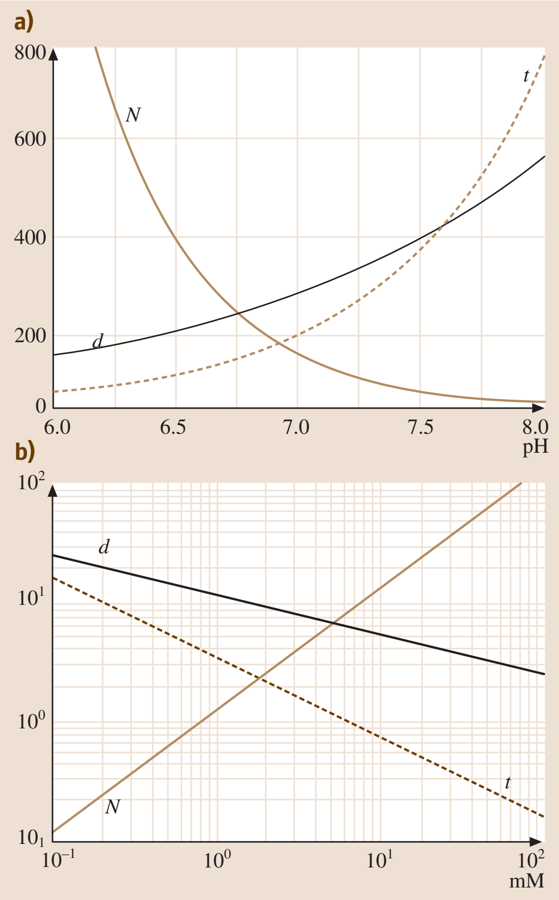

Fig. 24.5 **(a)** Representations of pH in terms of average distance between ions (d in nm), the number of ions per one *E. coli* cell volume (N), and the typical time for an ion to diffuse over a distance equal to their average separation (t in 100 ns). **(b)** Concentration unit millimolar (mM) represented in terms of d (nm), N (millions of molecules), and t (10 ns)

We have seen that the world of biological nanotechnology utilizes length scales that span a rather wide range, from fractions of a nanometer to tens of microns. This provides a challenge to modeling, whereby methods with atomic-scale resolution need to be combined in a consistent and seamless way with coarse-grained continuum descriptions to provide a complete picture. An example of such modeling will be provided in later sections when we take up the mechanical aspect of the viral life cycle.

24.2.2 Temporal Scales and Processes

We have seen that a study of spatial scales is a first step in our quest to understand the units that are suited to characterizing the biology-nanotechnology interface. We note that our understanding of spatial structures both in the living world, as well as those used in intriguing man-made technologies such as microelectronics have been built around important advances such as electron microscopy, X-ray crystallography, nuclear magnetic resonance, etc. that have revolutionized structural biology and materials science alike. On the other hand, one of the key challenges that remain in really appreciating the structure-properties paradigm, which is as central to biology as it is to materials science, is the need to acknowledge the dynamical evolution of these structures.

We note that just as there is a hierarchy of spatial scales that are important to consider in describing biological nanotechnology, so too is there a hierarchy of *temporal processes* that also demand a careful consideration of relevant units. An impressive representation of the temporal hierarchy that must be faced in contemplating biological systems is given by *Chan* and *Dill* [24.38] in their Fig. 5. They organize their temporal hierarchy according to factors of 1,000, starting at the femtosecond time scale and ending with time scales representative of the cell cycle itself. One of the elementary dynamical processes undergone by all of the molecular actors of the living world is thermal vibration. Vibrations of atoms occur at the time scales of 10^{-15} to 10^{-12} s. As will be noted later, from a modeling perspective, this depressing fact manifests itself in the necessity of using time steps of the same order when integrating the equations of motion describing molecular dynamics. The reason that this is unfortunate is that almost everything of dynamical interest occurs at time scales much longer than the femtosecond time scale characteristic of molecular jiggling, thus making molecular dynamics investigations computationally very expensive.

We follow *Chan* and *Dill* [24.38] in examining how successive thousandfold increases in our temporal resolution smears out features such as vibrational motion and brings into focus more interesting processes characteristic of macromolecular function. Indeed, the side chains of amino acids rotate with a characteristic period on the order of 10^{-9} s. Another thousandfold increase in time scale begins to bring key biological processes such as polymerization into view. Just as a length scale on the order of 10–100 nm is perhaps most characteristic of the length scales of biological nanotechnology, the time scale that is most relevant for considering the processes associated with biological nanotechnology is something between a microsecond and a millisecond. To drive home this point, we reconsider the machines described in the previous section from the length scale perspective, but now with an eye toward how fast they perform their function. Several of the molecular machines considered in the previous section mediate polymerization reactions. For example, RNA polymerase produces messenger RNA molecules by the repeated addition of nucleotides to a chain of ever increasing length. Such messenger RNA molecules serve in turn as the template for synthesis of new proteins by the ribosome, which reads the message contained in the RNA and adds individual amino acids onto the protein. From the temporal perspective of the present section, it is of interest to examine the rate at which new monomers are added to these polymers – roughly 10 per s in the case of RNA polymerase [24.39], and 2 per s in the case of the ribosome in eucaryotic cells [24.1] (bacterial ribosomes are 10 times faster). Once proteins are synthesized, in order to assume their full cellular responsibilities, they must fold into their native state, a process that occurs on time scales on the order of a millisecond.

In addition to translational machines like RNA polymerase and the ribosome, there are a host of intriguing rotary machines such as ATP synthase and the flagellar rotary motor. The flagellar motor, which is the means of locomotion for several bacteria, rotates at about 100 rpm [24.6]. ATP synthase makes use of a proton gradient across a lipid bilayer to provide rotation at a rate of roughly 6,000 rpm [24.41], approaching that of turbojet engines.

Thus far, our discussion has emphasized the rate of a variety of active processes of importance to biological nanotechnology. It is of interest to contrast these scales with those pertaining to diffusion. We note that in many instances in the biological setting the time scale of interest is that determined by the time it takes for diffusive communication of two spatially separated regions. In particular, as noted earlier in the chapter, the operative time scale is given by $t_{\text{diffusive}} = x^2/2D$, where D is the diffusion constant and x is the distance over which the diffusion has taken place.

To get a sense for all of the time scales described in this section, Table 24.2 presents a variety of results considered above in tabular form, but now with all times measured in units of the time it takes for ATPsynthase to make one rotation (≈ 10 ms). Our reason for adopting ATPsynthase rotation as defining a unit of time is that it exhibits motions associated with one of the most important of life's processes, namely, the synthesis of new ATP molecules. In addition, our later discussion of units associated with both energy and power will once again appeal to the central role played by ATP in biological nanotechnology.

24.2.3 Force and Energy Scales: The Interplay of Deterministic and Thermal Forces

Another set of scales of great relevance and importance to the nano-bio interface concerns the nature of the forces that act in this setting. Thus far, we have examined the spatial extent and the cycle time of a variety of examples of biological nanotechnology. As a next step in our examination of scales, we consider the forces

Table 24.2 Time scales of various events in biological nanotechnology measured in units of the average time taken for ATPsynthase to make a single rotation

Process	Time scale (units of ATPase rotation)	Ref.
1/(frequency of amino acid addition) – eukaryotic ribosome	50	[24.1]
1/(frequency of monomer addition) – RNA polymerase	10	[24.39]
Time between motion reset of *E. coli*	10	[24.6]
Kinesin step	1	[24.26]
Diffusion time for protein (ion) to cross an *E. coli*	0.5 (0.02)	[24.11]
Vibrational period of nanocantilever	10^{-3}–10^{-5}	[24.40]

and energies associated with these structures. Two outstanding sources for developing a feel for the relevant numbers are the books of *Wrigglesworth* [24.42] and *Smil* [24.43]. Perhaps the most compelling feature when thinking about forces at these scales is the interplay between deterministic and thermal forces. To substantiate this claim, we note that at room temperature the fundamental energetic quantity is $k_B T \approx 4.1$ pN nm. The reason this number is of interest to the current endeavor is the realization that many of the molecular motors that have thus far been investigated act with forces on the piconewton scale over distances of nanometers.

The fact that the piconewton is the relevant unit of force can be gleaned from a simple estimate. Namely, consider a typical skeletal muscle in the human arm. The cross-sectional size of the muscle is of order 3 cm. The muscle consists of cylindrical rods of protein called myofibrils, which are roughly 2 μm in diameter, while the myofibrils themselves are made of strands of actin and myosin filaments, which total some 60 nm in diameter [24.1]. This gives 10^{12} myosin filaments per cross-sectional area of the muscle. As each myosin filament over the length of a single sarcomere (the contractile unit of myofibrils, some 2.5 μm long) contains some 300 myosin heads, lifting a 30 kg load corresponds to a force of 1 pN per myosin head. This is certainly an underestimate, since not all myosin heads are attached to the actin filament at the same time. Since our estimate leaves out many details, we might wonder how it compares to the measured forces exerted by molecular motors. Sophisticated experiments with optical tweezers have revealed that actin-myosin motors stall at a force of around 5 pN [24.44]; RNA polymerase stalls at a force of about 20 pN [24.39]; the portal motor of the ϕ-29 bacteriophage exerts forces up to 50 pN [24.29].

As noted earlier, RNA polymerase is a molecular motor that moves along DNA while transcribing genes into messenger RNA. The DNA itself is an elastic object that is deformed by forces exerted on it by various proteins. At forces less than 0.1 pN its response is that of an entropic spring with a stretch modulus of 0.1 pN, while at forces exceeding 10 pN its stretch modulus is determined by hydrogen bonding of the base pairs and is roughly 1,000 pN [24.45]. All the above mentioned data reinforces the argument that piconewton is the relevant unit of force in the nanomechanical world of the cell.

The concept of "stress" is closely related to that of force. Stress is a continuum mechanical concept of force per unit area, and it has been used with great success in solid and fluid mechanics. We extend the idea of stress to the nanomechanical level to see what numbers we arrive at. From the data above we can deduce that a single myosin fiber sustains a stress of about 10^{-2} pN/nm^2, which is the same as 10^{-2} MPa. Migratory animal cells such as fibroblasts, which are responsible for scavenging and destroying undesirable products in tissues, can generate a maximum stress on their substrate on the order of 3.0×10^{-2} MPa [24.46]. DNA can sustain stresses in the excess of 20 pN/nm^2, or 20 MPa, at which point an interesting structural transition accompanied by an overall increase of its contour length is observed [24.45]. Engineering materials such as steel and aluminium, on the other hand, can sustain stresses of about 100 MPa. This goes to show that nanotechnology in the context of biological systems is built from rather soft materials.

It is of interest to translate our intuition concerning piconewton forces into corresponding energetic terms. The kinesin motor advances 8 nm in each step against forces as high as 5 pN [24.25]. This translates to a work done on the order of 40 pN nm. A myosin motor suffers displacements of about 15 nm with forces in the piconewton range. ATP hydrolysis (to ADP) releases energy (at pH 7 and room temperature) of about 50 pN nm [24.1]. When a titin molecule is pulled, it unfolds under forces of 30–300 pN, causing discrete expansions of 10–30 nm [24.22], implying energies in the range of 300–9,000 pN nm. Experimental data provides a compelling argument in favor of thinking of the pN nm as a unit of energy for nanomechanics. However, the observation that $k_B T = 4.1$ pN nm at room temperature gives important insight, since it reveals that thermal forces and entropic effects play a competing role in biological nanomechanics. This provides nature with unique design challenges, whereby molecular motors that can perform useful work must do so in the presence of strong thermal fluctuations for the normal functioning of the cell. Operation of motors in such a noisy environment is governed by laws that are probabilistic in nature. This is manifest in single molecule experiments that observe motors stalling and sometimes reversing direction.

Energy conversion is crucial to any developmental or evolutionary process. The steam engine powered the industrial revolution. However, long before thinking beings with man-made machines founded the industrial revolution, power generation had already become a central part of life's nanotechnology. Indeed, the development of ATP-synthase is one of the cornerstones for the evolution of higher life-forms. An inevitable concomitant of evolution is the necessity for faster and more efficient operations. The industrial revolution led to the emergence of bigger and faster modes of transport; biological evolution led to the emergence of complex and intelligent organisms. Invariably, a machine or an organism is limited in its abilities by the speed at which it can convert one form of energy (usually chemical) to other forms of energy (usually mechanical). This is why a study of their "power plants" becomes important.

Power plants (or engines) are usually characterized by their force-velocity curves and compared using their power-to-weight (P/W) ratio. For example, the myosin motor has a P/W ratio of 2×10^4 W/kg, the bacterial flagellar motor stands at 100 W/kg, an internal combustion engine is at about 300 W/kg, and a turbojet engine stands at 3,000 W/kg [24.47]. These figures tell us that linear motors like myosin and kinesin are extremely powerful machines.

24.3 Modeling at the Nano-Bio Interface

We have already provided a number of different views of the biology-nanotechnology interface, all of which reveal the insights that can emerge from model building. Indeed, one of the key thrusts of this entire chapter is the view that as the type of data that emerges concerning biological systems becomes increasingly quantitative, it must be responded to with models that are also quantitative [24.30]. The plan of this section is to show how atomistic and continuum analyses each offer insights into problems of nanotechnological significance, but under some circumstances, both are found wanting and it is only through a synthesis of both types of models that certain problems will surrender. The plan of this section is to examine the advantages and difficulties associated with adopting both atomistic and continuum perspectives and then to hint at the benefits of seeking mixed representations.

24.3.1 Tension Between Universality and Specificity

One of the key insights concerning model building in nanomechanics, whether we are talking about the nanoscale tribological questions pertinent to magnetic recording or the operations of molecular motors, is that such questions live in the no-man's-land between traditional continuum analysis at one extreme and all-atom approaches such as molecular dynamics on the other. Indeed, there is much discussion about the breakdown of continuum mechanics in modeling the mechanics of systems at the nanoscale. This dichotomy between continuum theories, which treat matter as continuously distributed, and atomic-level models, which explicitly acknowledge the graininess of matter, can be restated in a different (and perhaps more enlightening) way. In particular, it is possible to see atomistic and continuum theories as offering complementary views of the same underlying physics. Continuum models are suitable for characterizing those features of a system that can be thought of as averages over the underlying microscopic fluctuations. By way of contrast, atomistic models reveal the details that a continuum model will never capture, and in particular, they shed light on the specificity of the problem at hand.

The perspective adopted here is that continuum models and atomistic models each reveal important features of a given problem. For example, in contemplating the competition between fracture and plasticity at crack tips, a continuum analysis provides critical insights into the nature of the elastic fields surrounding a defect such as a crack. These fields adopt a fundamental and universal form at large scales with all detailed material features buried in simple material parameters. By way of contrast, the precise details of the dissipative processes occurring at a crack tip (in particular, the competition between bond breaking with the creation of new free surface and dislocation nucleation) require detailed atomic-level descriptions of the energetics of bond stretching and breaking. In the biological setting, similar remarks can be made. In certain instances, the description of biological polymers as random coils suffices and yields insights into features such as the mean size of the polymer chain as a function of its length. On the other hand, if our objective is mechanistic understanding of processes such as how phosphorylation of a particular protein induces conformational change, this is an intrinsically atomic-level question. The language we invoke to describe this dichotomy is the use of the terms *universality* and *specificity*, where, as described above, insights of a universal character refer to those features of systems that are generic, while specificity refers to the features of systems that depend upon precise details such as whether or not a particular molecule is bound at a particular site. This fundamental tension between atomistic and continuum perspectives is elaborated in *Phillips* et al. [24.48].

24.3.2 Atomic-Level Analysis of Biological Systems

As already described in the introduction, one of the intriguing roles of nanotechnology in the biological setting is that it has brought the Instron technology of traditional solid mechanics to the nanoscale and has permitted the investigation of the force-extension characteristics of nanoscale systems (macromolecules and their assemblies, in particular). As noted earlier, mechanical force spectroscopy [24.20] is emerging as a profound tool for exploring the connection between structure, force, and chemistry, in much the same way that conventional stress-strain tests provided insights into the connection between structure and properties of conventional materials. Figure 24.3 gives one such example in the case of the muscle protein titin. Similar insights have been obtained through systematic examination of the force-extension properties of DNA [24.49].

The objective of this section is to call the reader's attention to the types of modeling that can be done from an all-atom perspective. What exactly is meant by a model in this setting? We begin by noting that for the purposes of the general discussion given here, the same basic ideas are present whether one is modeling tribological processes such as the sliding of adjacent surfaces, or attempting to examine the operation of a protein machine such as ATPsynthase. The set of degrees of freedom considered by the atomic-scale modeler is the full set of atomic positions $(\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots \boldsymbol{R}_N)$, which we also refer to as $\{\boldsymbol{R}_i\}$. For most purposes, one proceeds through reference to a *classical* potential energy function, $E_{\text{tot}}(\{\boldsymbol{R}_i\})$, which is a rule that assigns an energy for every configuration $\{\boldsymbol{R}_i\}$. While it would be most appealing to be able to perform a full quantum mechanical analysis, such calculations are computationally prohibitive. Given the potential energy, the forces on each and every atom can be computed where, for example, the force in the αth Cartesian direction on the ith atom is given by

$$F_{i,\alpha} = -\frac{\partial E_{\text{tot}}}{\partial R_{i,\alpha}} \,. \qquad (24.8)$$

For those interested in finding the energy minimizers, such forces can be used in conjunction with methods such as the conjugate gradient method or the Newton–Raphson method. Alternatively, many questions are of a dynamic character, and in these cases, Newton's equations of motion are integrated, thus permitting an investigation of the temporal evolution of the system of interest. We note that we have neglected to discuss subtleties of how one maintains the system at constant temperature, and we leave such subtleties to the curious reader, who can learn more about them in *Frenkel* and *Smit* [24.50].

To give a flavor of where such calculations can lead we note that the same force-extension characteristics already shown in Fig. 24.3 have been computed in a molecular dynamics simulation [24.51]. One of the insights to emerge from these calculations was the particular dynamical pathway, namely, the breaking of a particular collection of hydrogen bonds during the rupture process of each of the immunoglobulin domains. We further note that in the case of titin there has been an especially pleasing synergy between the atomic-scale calculations and the corresponding force-extension measurements. In particular, in response to the suggestion that it was a particular set of hydrogen bonds that were impugned in the rupture process, mutated versions of the titin protein were created in which the number of such hydrogen bonds was changed with the result that the rupture force was changed according to expectation [24.52]. This can be seen as a primitive example of the ultimate goal of tailoring new materials (both biological and otherwise) through appropriate computer modeling.

24.3.3 Continuum Analysis of Biological Systems

We have already noted that there are many appealing features to strictly continuum analyses of material systems. In particular, models based upon continuum mechanics result in a mathematical formulation that permits us to uncork the traditional tools associated with partial differential equations and functional analysis. In keeping with our argument that it is the role of models to serve our intuition, the ability to write down continuum models raises the possibility of obtaining analytic solutions to problems of interest, an eventuality that is nearly impossible once the all-atom framework has been adopted.

Macromolecules as Elastic Rods

Whether we contemplate the information carrying nucleic acids, the workhorse proteins, or energy storing sugars, ultimately, the molecular business of the living world is dominated by long chain molecules. For the model builder, such molecules suggest two complementary perspectives, each of which contains a part of the whole truth. On the one hand, polymer physicists have gained huge insights [24.53] by thinking of long chain molecules as random walks. Stated simply, the key virtue of the random-walk description of long chain molecules is that it reflects the overwhelming importance of entropy in governing the geometric conformations that can be adopted by macromolecules. The other side of the same coin considers long chain molecules as elastic rods with a stiffness that governs their propensity for bending. We will take up this perspective in great detail in the final section of this chapter when we consider the energetics of DNA packaging in viruses.

For our present purposes, we examine one case study in treating macromolecules as elastic rods. The example of interest here is chosen in part because it reflects another fascinating aspect of biological nanotechnology, namely, regulation and control. It is well-known that genes are switched on and off as they are needed. These topics are described beautifully in the work of *Ptashne* [24.54]. The basic idea can be elucidated through the example of a particular set of genes in *E. coli*: the *lac* operon. There are a set of enzymes exploited by this bacterial cell when it needs to digest the sugar lactose. The gene that codes for these enzymes is only turned on when lactose is present and certain other sugars are absent. The nanotechnological solution to the problem of regulating this gene has been solved by nature through the presence of a molecule known as the *lac* repressor, which binds onto the DNA in the region of this *lac* gene and prevents the gene from being expressed. More specifically, the *lac* repressor binds onto two different regions of the DNA molecule simultaneously, forming a loop in the intervening region, and thus rendering that region inoperative for transcription.

We have belabored the difficulties that attend the all-atom simulation of most problems of nanotechnological interest – system sizes are too small and simulated times are too short. On the other hand, clearly it is of great interest to probe the atomic-level dynamics of the way that molecules such as *lac* repressor interact with DNA and thereby serve as gene regulators. A compromise position adopted by *Balaeff* et al. [24.55] was to use elasticity theory to model the structure and energetics of DNA, which then served effectively as a boundary condition for their all-atom calculation of the properties of the *lac* repressor/DNA complex. The simulation cell is shown in Fig. 24.6, which also shows the looped region of DNA that serves as a boundary condition for the atomic-level model of the complex.

Membranes as Elastic Media

One way to classify biological structures is along the lines of their dimensionality. In the previous discussion, we examined the sense in which many of the key macromolecules of the living world can be thought of as one-dimensional rods. The next level in this dimensional hierarchy is to consider the various membranes that compartmentalize the cell and which can be examined from the perspective of two-dimensional elasticity. Just as there are a huge number of modeling questions to be posed concerning the structure and function of long chain molecules such as nucleic acids and proteins, there is a similar list of questions that attend the presence of a host of different membranes throughout the cell. To name a few, we first remind the reader that such membranes are full of various proteins that serve in a variety of different capacities, some of which depend upon the mechanical state of the membrane itself. In a different vein, there has been great interest in examining the factors that give rise to the equilibrium shapes of cells [24.56, 57]. In both instances, the logic associated

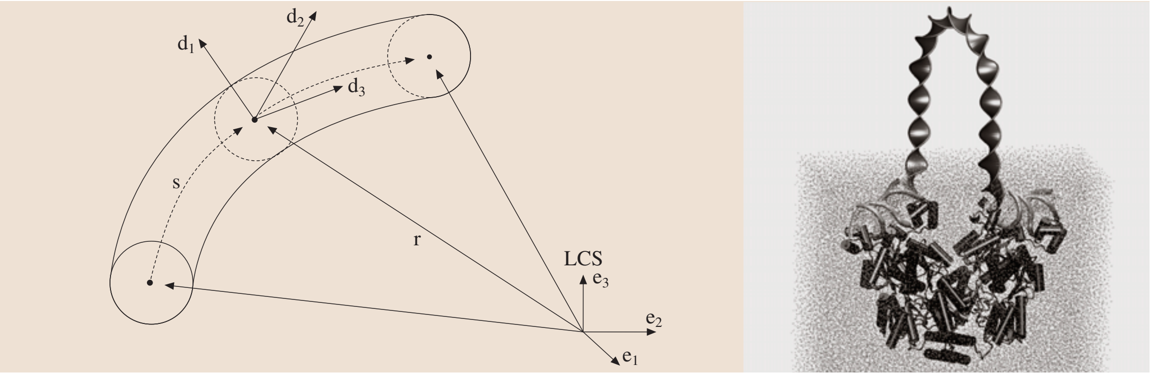

Fig. 24.6 Illustration of the way that the properties of a particular DNA-protein binding interaction were modeled by virtue of an elastic treatment of the looped region of the DNA [24.48]

with building models of these phenomena centers on the construction of an elastic Hamiltonian that captures the energetics of deformations expressed in terms of surface area, curvature, and variations in thickness. Also, just as there has been great progress in measuring the properties of single proteins and nucleic acids, there has been considerable progress in examining the force response of membranes as well [24.58].

The prospect of bringing the tools of traditional continuum theory to bear on problems of biological significance is indeed a daunting one. As noted earlier, to our way of thinking one of the biggest challenges posed by biological problems is the dependence of these systems on detailed molecular structures and dynamics – what we have earlier characterized as "specificity", one of the hallmarks of biological action.

24.4 Nature's Nanotechnology Revealed: Viruses as a Case Study

Over the course of this chapter, we have presented different facets of the relationship between biology and nanotechnology and the modeling challenges they encompass. For example, we have noted that two key aspects of nature's nanotechnologies are the exploitation of self-assembly processes for the construction of molecular machines and the role of active processes mediated by molecular motors. We have also argued for the synergistic role of nanotechnology in producing new methods for the experimental analysis of biological systems with examples ranging from new molecular dyes to the use of optical tweezers. Finally, we have described some of the modeling challenges posed by contemplating the biology-nanotechnology interface and the sorts of coarse-grained models that have arisen to meet these challenges. In this final section, we present a discussion of viruses as a case study at the confluence of these different themes.

Though the importance of viruses from a health perspective are well-known even to the casual observer, they are similarly important both from the technological perspective (as will be explained below) and as a compelling and profound example of nature's nanotechnology par excellence. To appreciate the sense in which viruses serve as a compelling example of biological nanotechnology, we begin by reviewing the nature of the viral life cycle with special reference to the case of bacterial viruses (i. e., viruses that infect bacteria), known as bacteriophages. For concreteness, we consider the life cycle of a bacteriophage such as the famed lambda phage that infects *E. coli*. The life cycle of such viruses is shown schematically in Fig. 24.7. Upon an encounter with the *E. coli* host, the virus attaches to a receptor (protein) embedded in the bacterial membrane and ejects its DNA into the host cell, leaving an empty capsid as refuse from the process. As an aside, it is worth noting that experiments such as the famed Hershey–Chase experiment used tagged DNA and tagged proteins on viruses to settle the question of whether proteins or nucleic acids are the carriers of genetic information. The outcome of these experiments was the conclusion that DNA is the genetic material. For the purposes of the present discussion, the other interesting outcome of the Hershey–Chase experiment is that it provides insights into the mechanistic process associated with delivery of the viral genome.

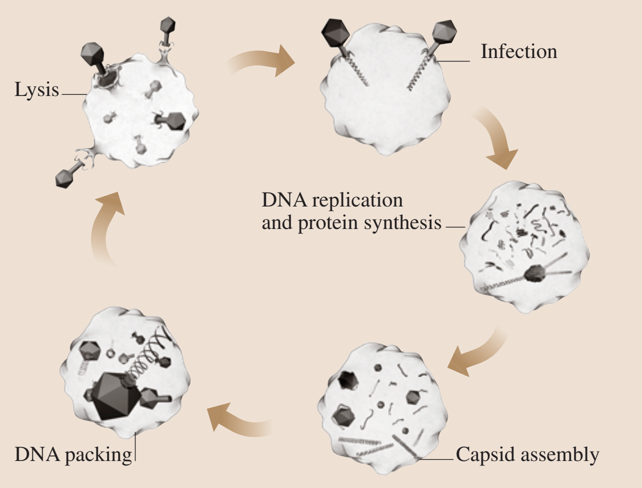

Fig. 24.7 Schematic representation of the viral life cycle illustrating the various nanotechnological actions in this cycle. During the infection stage, the viral genome is delivered to the victimized bacterial cell. During the process labeled DNA replication and protein synthesis, the machinery of the host cell is enlisted to produce renegade DNA and proteins to be used in the assembly of new viral particles. The capsid assembly process involves the attachment of the various protein building blocks that make up the capsid. DNA packing refers to the active packaging of DNA into the newly formed capsids by a molecular motor. Finally, once the assembly process is complete, the infectious phage particles are released from the victimized cell to repeat their infectious act elsewhere

Once the viral genome has been delivered to the host cell (we now oversimplify), the replication machinery of *E. coli* is hijacked to do the virus' bidding. In particular, the genes coded for in the viral DNA are expressed, and the proteins needed to make copies of the phage particle (i. e., the ingredients for an eventual self-assembly process) are created. Interestingly, part of the gene products associated with this process are the components of the molecular motor, which is responsible for packaging the replicated viral DNA into the new protein capsids that will eventually become the next generation of viruses. Indeed, once assembled, this motor takes the replicated viral DNA and packs it into the viral capsid. Recall our insistence that another of the important themes presented by biological examples of nanotechnology is the huge role played by active processes that reflect mechanochemical coupling. Once the packaging process is completed, and the remaining parts of the self-assembly process have been effected (such as the attachment of the viral sheath and legs whereby the phage attaches to host cells), enzymes are released that breakdown the cell wall of the infected cell with the ultimate result that what started as a single bacteriophage has in less than an hour become on the order of 100 new phage particles ready to infect new cells.

As noted above, a second sense in which viruses are deserving of case study status in the current chapter is the role viruses such as lambda phage play in biotechnology itself. For example, from the standpoint of both cloning and the construction of genomic libraries, the use of viruses is commonplace. From a more speculative perspective, viruses are also drawing increasing attention from the standpoint of gene therapy (as a way of delivering DNA to specific locations) and, more generally, as small-scale containers. To substantiate these assertions, we briefly consider the construction of genomic libraries and the role played by viruses in these manipulations. We pose the following question: Given that the length of the human genome is on the order of 10^9 base pairs, how can one organize and store all of this genetic information for the purposes of experiments such as the sequencing of the human genome? One answer to this challenge is the use of bacteriophage to deliver phage DNA to *E. coli* cells, but with the subtlety that the delivered DNA fragments have ligated within them fragments of the human genome whose lengths are on the order of 10 kb (kb = kilo base-pairs). The particulars of this procedure involve first cutting the DNA into fragments of roughly 10 kb in length using a class of enzymes known as restriction enzymes. The result of this operation is that the original genome is now separated into a random collection of fragments. These fragments are then mixed with the original lambda phage DNA that has also been cut at a single site such that the genomic fragments and the lambda fragments are complementary. Using a second enzyme known as ligase, the genomic DNA is joined to the original lambda phage fragments so that the resulting DNA resembles the original lambda DNA, but now with a 10 kb fragment inserted in the middle. These cloned lambda DNA molecules are then packaged into the lambda phage using a packaging reaction (see the website www.stratagene.com for an example of these products). The resulting lambda phage, now fully packed with cloned DNA, are used to infect *E. coli* cells, and the cloned DNA, once in the bacterial cell, circularizes into a DNA fragment known as a plasmid. The plasmid is then passed from one generation of *E. coli* to the next. This is the so-called prophage pathway in which, unlike the lytic pathway shown in Fig. 24.7, the bacteriophage is latent and does not destroy the cell [24.1]. Note that different *E. coli* cells are infected with viruses containing different cloned fragments. As a result, the collection of all such *E. coli* cells constitute a library of the original genomic DNA.

We round out our introductory discussion on viruses with a discussion of the compelling recent experiments that have been performed to investigate the problem of viral packing and, similarly, how ideas like those described in this chapter can be used to model these processes. *Smith* et al. [24.29] used optical tweezers to measure the force applied by the packaging motor of the ϕ-29 bacteriophage during the DNA packaging process. In particular, they measured the force and rate of packaging as a function of the amount of DNA packed into the viral head with the result that as more DNA is packed, the resistive force due to the packed DNA increases, and the packing rate is reduced.

As noted above, the viral problem is interesting not only because it exemplifies many of the features of biological nanotechnology introduced throughout this chapter, but also illustrates the way in which model building has arisen in response to experimental insights. The various competing energies that are implicated in the DNA packing process have been described by *Riemer* and *Bloomfield* [24.59]. The energetics of viral packing is characterized by a number of different factors, including: i) the entropic-spring effect that causes the DNA in solution to adopt a more spread out configuration than that in the viral capsid, ii) the energetics of elastic bending, which results from inducing curvature in the DNA on a scale that is considerably smaller than the persistence length of $\xi_p \approx 50$ nm, and iii) those factors related to the presence of charge both on the DNA itself and in the surrounding solution. As shown by *Riemer* and *Bloomfield* [24.59], the entropic contribution is smaller by a factor of 10 or more relative to the bending energies and those mediated by the charges on the DNA and the surrounding solution, and hence we make no further reference to it. As a result, just like in earlier work [24.60, 61], we examine the interplay of elastic and interaction forces, though we neglect surface terms originating from DNA-capsid and DNA-solvent interactions.

We note that the viral packing process involves DNA segments with lengths on the order of 10 µm and takes place on the time scale of minutes. As a result, from a modeling perspective, it is clear that such problems are clearly out of reach of conventional molecular dynamics. As a result, we exploit a continuum description of the DNA packing process with the proviso that such models will ultimately need to be refined to account for the sequence dependence of the elasticity of DNA.

A mathematical description of the energetics of viral packing must account for two competing factors, namely, the energy cost to bend the DNA and place it in the capsid and the repulsive interaction between adjacent DNA segments that are too close together. The structural picture of the packaged DNA is inspired by experiments that indicate that the DNA is packed in concentric rings from large radii inwards [24.62–64]. The bending energy cost of accumulating hoops within the capsid is given by

$$E_{\mathrm{el}} = \pi \xi_{\mathrm{p}} k_{\mathrm{B}} T \sum_i \frac{N(R_i)}{R_i} , \qquad (24.9)$$

where $N(R_i)$ is the number of hoops that are packed at the radius R_i [24.59]. The basic idea of this expression is that we are adding up the elastic energy on a hoop by hoop basis, with each hoop penalized by the usual energy cost to bend an elastic beam into a circular arc. The presence of $N(R_i)$ reflects the fact that because of the shape of the capsid, as the radius gets smaller the DNA can pack higher up into the capsid, thus increasing the number of allowed hoops. To make analytic progress with the expression for the stored elastic energy given above, we convert it into an integral of the form

$$E_{\mathrm{el}} = \frac{\pi \xi_{\mathrm{p}} k_{\mathrm{B}} T}{\sqrt{3} d_{\mathrm{s}}/2} \int_R^{R_{\mathrm{out}}} \frac{N(R')}{R'} \, \mathrm{d}R' . \qquad (24.10)$$

The summation $\sum_i$ has been replaced by an integral $\int_R^{R_{\mathrm{out}}} \mathrm{d}R'/(\sqrt{3}d_{\mathrm{s}}/2)$, where the integration bounds are the inner and outer radii of the inner spool and $\sqrt{3}d_{\mathrm{s}}/2$ is the horizontal spacing between adjacent strands of the DNA. The geometrical factor $\sqrt{3}/2$ owes its presence to the hexagonal packing of the DNA strands. R is the radius of the innermost stack of hoops, R_{out} is the radius of the capsid, $N(R')$ is the number of hoops at radius R', d_{s} is the spacing between adjacent hoops, and ξ_{p} is the persistence length of DNA. We note that this expression is a manifestation of the first term in the energy functional in (24.3), used earlier in the context of the biofunctionalized cantilever and now specialized to the geometry of a partially filled capsid. Just as the energy functional in the cantilever context reflected a competition between different contributions to the total energy (in that case, elastic bending energy and surface terms), the DNA packing problem reflects a similar competition, this time between the elastic bending and the interactions between nearby DNA segments by virtue of the charge along the DNA. In particular, the interaction energy between the DNA hoops packed in the viral capsid scales with the length and is given by

$$E_{\mathrm{int}} = F_0 L (c^2 + c d_{\mathrm{s}}) \exp(-d_{\mathrm{s}}/c) , \qquad (24.11)$$

where c and F_0 are constants for a given solvent and L is the length of the packed DNA. This form of the interaction energy was shown to be very robust for a variety of solvents containing monovalent and divalent cations in a series of experiments by *Rau* et al. [24.65, 66] and *Parsegian* et al. [24.67]. The forces arising from this kind of an energy are purely repulsive in character. Interestingly, in this setting, rather than using temperature as a control parameter, it is much more common to change the circumstances in such experiments by tuning concentrations of various chemical constituents. In particular, when trivalent or tetravalent cations are present in solution there is an effective attractive interaction between DNA segments. In that case, the forces are repulsive only when d_{s} is smaller than a certain critical separation d_0 (and attractive otherwise) and the form of the interaction energy changes to

$$E_{\mathrm{int}} = F_0 L (c^2 + c d_{\mathrm{s}}) \exp\left(\frac{d_0 - d_{\mathrm{s}}}{c}\right) . \qquad (24.12)$$

Once both the elastic and interaction contributions to the energy have been reckoned, we are in a position to compute the total energy. If we recognize that the length L of the packed DNA is given by

$$L = \frac{2}{\sqrt{3} d_{\mathrm{s}}} \int_R^{R_{\mathrm{out}}} 2\pi R' N(R') \, \mathrm{d}R' , \qquad (24.13)$$

then the total energy for the packed DNA, in terms of L and d_s, is given by

$$E(L, d_\mathrm{s}) = \frac{\pi \xi_\mathrm{p} k_\mathrm{B} T}{\sqrt{3} d_\mathrm{s}/2} \int_R^{R_\mathrm{out}} \frac{N(R')}{R'} \,\mathrm{d}R' \tag{24.14}$$

$$+ F_0 (c^2 + c d_\mathrm{s}) \exp(-d_\mathrm{s}/c) \frac{2}{\sqrt{3} d_\mathrm{s}} \int_R^{R_\mathrm{out}} 2\pi R' N(R') \,\mathrm{d}R' \,.$$

The spacing at a given length L is determined by requiring that the packed DNA be in a minimum energy configuration, which is equivalent to asking that $\partial E / \partial d_\mathrm{s} = 0$. Hence, for a given geometry and length packed, one obtains d_s and thereby calculates the energy. Thus the energy is known as a function of the length packed, and to compute the force as a function of length of DNA packed one need only differentiate this energy with respect to the length packed. The result of this procedure is an expression for the force of the form

$$F(L) = F_0 (c^2 + c d_\mathrm{s}) \exp\left(-\frac{d_\mathrm{s}}{c}\right) + \frac{\xi_P k_\mathrm{B} T}{2R^2} \,. \tag{24.15}$$

This is a generic expression valid for a capsid of any shape. The effect of the geometry is captured in the variation of d_s as a function of the length packed, as well as through the inner packing radius R. The resulting force-packing curves are shown in Fig. 24.8. In particular, this figure shows the packing force as a function of a fraction of the genome packed for a number of different solutions, with the different curves revealing the large role played by positive counterions in dictating the overall energetics of these processes.

On the basis of calculations like those given above, *Kindt* et al. [24.61] estimate that the pressure inside the capsid of the phage is on the order of 35 atmospheres. A crude estimate by *Smith* et al. [24.29] gives a figure close to 60 atmospheres. These numbers are intriguing in their own right, but more importantly they demonstrate the promise that proteins and other biological materials hold as candidates for engineering materials. To further explore the structural integrity of viruses in their role as pressurized protein shells, we now turn to an examination of the rupture stress of viral capsids.

Throughout this chapter, one of our main arguments has been the idea that whether it is our ambition to model conventional materials or their biological counterparts, modeling at the nanoscale implies challenges that can not be met either by purely continuum ideas or by traditional all-atom thinking. Indeed, one set of powerful methods is built around the attempt to borrow those features of continuum and atomistic thinking that are most robust, while rejecting those that are inapplicable at the atomic scale. As an example of this type of thinking, we examine another question drawn from the viruses-as-nanotechnology setting. In particular, since viruses have been considered a means of transporting material other than the genetic material of the virus itself, it is of interest to consider the maximum internal pressure viral capsids can sustain without rupture, as this will determine how much material can be safely packaged. To that end, we use continuum mechanics to estimate the stresses within the capsid walls. These stresses are then mapped onto atomic-level forces by appealing to the details of the protein structure of the monomers making up the capsid and a knowledge of the forces that link them. By relating the continuum and atomistic calculations, we then determine the maximum sustainable internal pressure.

We imagine the capsid to be a hollow sphere loaded by a pressure p_i from inside and a pressure p_o from outside. The inner and outer radii are R_i and R_o, respectively. As a representative example, bacteriophage

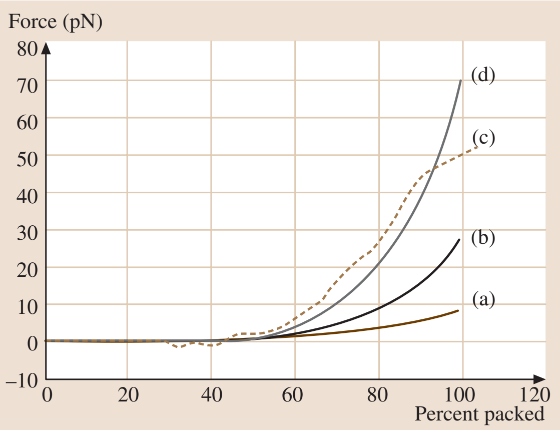

Fig. 24.8 Force as a function of the percent packed for a cylindrical capsid under purely repulsive solvent conditions. The dimensions of the capsid and the length of genome packed were chosen to represent the ϕ29 phage: $R_\mathrm{out} = 210$ Å, $z = 540$ Å, and $L = 6.6\,\mu$m. Curve (c) shows the experimental results of *Smith* et al. [24.29], while theoretical curves (a), (b), and (d) are given by (24.15), with $F_0 = 10 \times 55{,}000$, $40 \times 55{,}000$, and $100 \times 55{,}000\,\mathrm{pN/nm^2}$, respectively

GA is characterized geometrically by $R_\text{i} = 123\,\text{Å}$ and $R_\text{o} = 145\,\text{Å}$. Evaluation of a number of different capsid types suggests that treating capsids as though they have a mean thickness of roughly 15 Å suffices for the level of modeling being considered here. For the purposes of computing the internal stresses within the capsid, we begin with a statement of equilibrium from continuum mechanics that requires that at every point in the capsid

$$\nabla \cdot \sigma = \mathbf{0}\,, \tag{24.16}$$

where σ is the stress tensor comprising three normal stresses and three shear stresses. For a problem with spherical symmetry, like that considered here, the stresses reduce to a radial stress σ_R and a circumferential stress σ_T. Solution of the equilibrium equations results in stresses of the form [24.32]

$$\sigma_R = \frac{C}{r^3} + D, \quad \sigma_T = -\frac{C}{2r^3} + D\,. \tag{24.17}$$

By using the boundary conditions $\sigma_R|_{r=R_\text{i}} = -p_\text{i}$ and $\sigma_R|_{r=R_\text{o}} = -p_\text{o}$, the constants C and D can be determined with the result

$$\sigma_R = \frac{p_\text{o} R_\text{o}^3 \left(r^3 - R_\text{i}^3\right)}{r^3 \left(R_\text{i}^3 - R_\text{o}^3\right)} + \frac{p_\text{i} R_\text{i}^3 \left(R_\text{o}^3 - r^3\right)}{r^3 \left(R_\text{i}^3 - R_\text{o}^3\right)}\,, \tag{24.18}$$

$$\sigma_T = \frac{p_\text{o} R_\text{o}^3 \left(2r^3 + R_\text{i}^3\right)}{2r^3 \left(R_\text{i}^3 - R_\text{o}^3\right)} - \frac{p_\text{i} R_\text{i}^3 \left(2r^3 + R_\text{o}^3\right)}{2r^3 \left(R_\text{i}^3 - R_\text{o}^3\right)}\,. \tag{24.19}$$

The stress σ_T is our primary concern, since it acts so as to tear the sphere apart. By looking at the expressions above, we can see that this stress is maximum at $r = R_\text{i}$ and the maximum value is given by

$$\sigma_T^\text{max} = \frac{3 p_\text{o} R_\text{o}^3 - p_\text{i} \left(2R_\text{i}^3 + R_\text{o}^3\right)}{2\left(R_\text{i}^3 - R_\text{o}^3\right)}\,. \tag{24.20}$$

We note that elasticity theory in and of itself is unable to comment on σ_T^max, since this is effectively a material parameter that characterizes the contacts between the various protein monomers that make up the capsid. As a result, we first examine how the rupture strength depends upon capsid dimensions in abstract terms and then turn to a concrete estimate of σ_T^max itself from several complementary perspectives. If (24.20) is rewritten using $p_\text{i} = p_\text{o} + \Delta p$ and $R_\text{o} = R_\text{i} + \Delta R$, and, further, it is realized that for typical capsid dimensions $\Delta R / R_\text{i} \ll 1$, it can be shown by rearranging (24.20) and by considering the case in which σ_T^max is much larger than p_o, the maximum sustainable internal pressure is given by

$$p_\text{i}^\text{max} = \frac{2\sigma_T^\text{max} \Delta R}{R_\text{i}}\,, \tag{24.21}$$

where p_i^max is really the quantity of interest, namely, the maximum sustainable internal pressure.

To make concrete progress to the point where we can actually estimate the rupture stress in atm, we need to consider capsids for which the structure is known and for which, at least approximately, the bonds between the monomers making up the capsid are understood. We note that in the language of fracture mechanics, what we seek is a cohesive surface model that provides a measure of the energy of interaction between two surfaces as a function of their separation [24.68]. There are a number of different ways to go about estimating the effective interaction between the monomers making up the capsid, one of which is by appealing to atomic-level calculations like those made by *Reddy* et al. [24.69]. The Viper website has systematized such information for a number of capsids, and one of the avenues we take to estimate σ_T^max is to appeal directly to their calculations. To that end, we assume that the energy of interaction per unit area as a function of separation x between adjacent monomers making up the capsid can be written in the form

$$E(x) = V_0 \left[\frac{1}{4}\left(\frac{x^*}{x}\right)^{12} - \frac{1}{2}\left(\frac{x^*}{x}\right)^{6}\right]. \tag{24.22}$$

The motivation for this functional form is the idea that the energy of interaction between adjacent monomeric units making up the capsid is the result of van der Waals contacts, and hence our cohesive surface law has inherited the properties of the underlying atomic force fields. To proceed to an estimate of σ_T^max itself, we must determine the parameters in the cohesive model described above. To that end, we note that *Reddy* et al. [24.69] have computed the association energies of various inequivalent contacts throughout a number of different icosahedral capsids. Their calculations result in a roughly constant value of $\approx -45\,\text{cal/mole}\,\text{Å}^2$ as the association energy, which in the language of our cohesive potential results in $V_0 \approx 12.5\,\text{pN/Å}$. This may be seen by noting that the association energy is given by $E(x^*) = -V_0/4$.

Once we have fixed x^*, which amounts to choosing a particular value for the equilibrium separation between two monomers, the material parameter σ_T^max is obtained by evaluating $\partial E(x)/\partial x$ at a value of x corresponding to the point of inflection ($\partial^2 E(x)/\partial x^2 = 0$) in the cohesive surface function. For the cohesive surface function used above, this results in a maximum stress of the form

$$\sigma_T^{\max} = \frac{7^{7/6}\times 18}{13^{13/6}}\,\frac{V_0}{x^*}\,. \tag{24.23}$$

With $\sigma_T^{\max}$ in hand, the maximum sustainable pressure is obtained from (24.21), and the results are shown in Fig. 24.9. These estimates suggest that the pressures within capsids as a result of packed DNA while large, are still smaller than our estimated rupture stresses. To more completely examine this question we have undertaken finite element elasticity calculations to examine the stresses in capsids exhibiting irregularities in both shape and thickness. In addition, further atomistic analysis of $\sigma_T^{\max}$ is needed with special reference to its dependence on distance between protein units. It would also be of interest to examine mutant versions of the monomeric units making up the capsid to see the implications of such mutations for $\sigma_T^{\max}$.

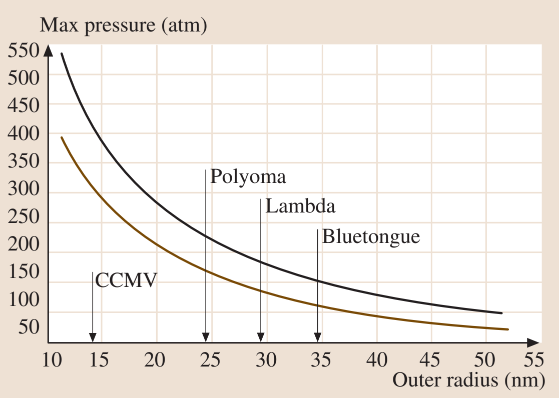

Fig. 24.9 Rupture pressure as a function of capsid radius for $x^* = 3\,\text{Å}$ (*upper curve*) and $x^* = 4\,\text{Å}$ (*lower curve*). The width of the capsid walls was set to $\Delta R = 1.5\,\text{nm}$, while $V_0 = 180\,\text{cal/molÅ}^2$

24.5 Concluding Remarks

One of the most compelling areas to be touched by nanotechnology is biological science. Indeed, we have argued that there is a fascinating interplay between these two subjects, with biology as a key beneficiary of advances in nanotechnology as a result of a new generation of single molecule experiments that complement traditional assays involving statistical assemblages of molecules. This interplay runs in both directions with nanotechnology continually receiving inspiration from biology itself. The goal of this chapter has been to highlight some representative examples of the interplay between biology and nanotechnology and to illustrate the role of nanomechanics in this field and how mechanical models have arisen in response to the emergence of this new field. Primary attention has been given to the particular example of the processes that attend the life cycle of bacterial viruses. Viruses feature many of the key lessons of biological nanotechnology, including self-assembly, as evidenced in the spontaneous formation of the protein shell (capsid) within which the viral genome is packaged and a motor-mediated biological process, namely, the packaging of DNA in this capsid by a molecular motor that pushes the DNA into the capsid. We argue that these processes in viruses are a compelling real-world example of nature's nanotechnology and reveal the nanomechanical challenges that will continue to be confronted at the nanotechnology-biology interface.

Though this chapter represents something of a departure from the rest of the chapters in this volume, it advances the view that biological nanotechnology serves as an inspiring vision of what nanotechnology can offer and reflects the authors' views that biologically inspired nanotechnology will play an ever increasing role as Feynman's view that "there is plenty of room at the bottom" continues to play out. Moreover, developing simple models of nature's nanotechnology can provide important insights into viable strategies for making machines at the nanoscale.

We have argued that the advance of single molecule technology presents both scientific and technological possibilities. Both classes of questions imply significant modeling demands, just as earlier advances in other settings, such as in materials for microelectronics applications, did. We claim that these modeling challenges are perhaps more acute as a result of the vicious chemical and structural nonhomogeneity of biological systems. As a result, we see an ever increasing role for modeling methods that aim to keep atomic-level specificity where needed, while rejecting such resolution elsewhere.

References

24.1 B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter: *Essential Cell Biology* (Garland, New York 1997) Chap. 12

24.2 B. Alberts: The cell as a collection of protein machines: Preparing the next generation of molecular biologists, Cell **92** (1998) 291

24.3 D. Bray: *Cell Movements From Molecules to Motility* (Garland, New York 2001)

24.4 S. J. Flint, L. W. Enquist, R. M. Krug, V. R. Racaniello, A. M. Skalka: *Principles of Virology* (ASM, Washington, DC 2000)

24.5 J. C. M. van Hest, D. A. Tirrell: Protein-based materials, toward a new level of structural control, Chem. Commun. **19** (2001) 1807–1904

24.6 H. Berg: Motile behavior of bacteria, Phys. Today **53** (2000) 24

24.7 J. E. Walker: ATP synthesis by rotary catalysis, Angew. Chem. Int. Edn. **37** (1998) 2308

24.8 P. D. Boyer: The ATP synthase – A splendid molecular machine, Annu. Rev. Biochem. **66** (1997) 717

24.9 M. Yoshida, E. Muneyuki, T. Hisabori: ATP synthase – A marvelous rotary engine of the cell, Nature Rev. Mol. Cell Bio. **2** (2001) 669

24.10 R. Yasuda, H. Noji, K. Kinosita, M. Yoshida: F_1-ATPase is a highly efficient molecular motor that rotates with discrete 120° steps, Cell **93** (1998) 1117

24.11 J. Howard: *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland 2001)

24.12 B. Hille: *Ion Channels of Excitable Membranes* (Sinauer, Sunderland 2001)

24.13 S. Sukharev, S. R. Durell, H. R. Guy: Structural Models of the MscL gating mechanism, Biophys. J. **81** (2001) 917

24.14 S. Sukharev, M. Betanzos, C.-S. Chiang, H. R. Guy: The gating mechanism of the large mechanosensitive channel MscL, Nature **409** (2001) 720

24.15 K. Kinosita, R. Yasuda, H. Noji: F_1-ATPase: A highly efficient rotary ATP machine, Essays Biochem. **35** (2000) 3

24.16 E. Racker, W. Stoeckenius: Reconstitution of purple membrane vesicles catalyzing light-driven proton uptake and adenosine triphosphate formation, J. Biol. Chem. **249** (1974) 662

24.17 G. Groth, J. E. Walker: ATP synthase from bovine heart mitochondria: reconstitution into unilamellar phospholipid vesicles of the pure enzyme in a functional state, Biochem. J. **318** (1996) 351

24.18 G. Wu, R. H. Datar, K. M. Hansen, T. Thundat, R. J. Cote, A. Majumdar: Bioassay of prostrate-specific antigen (PSA) using microcantilevers, Nature Biotech. **19** (2001) 856

24.19 R. P. Feynmann: There's plenty of room at the bottom, Eng. Sci. **23** (1960) 22–36, and www.zyvex.com/nanotech/feynman.html (1959)

24.20 E. Evans: Probing the relation between force-lifetime and chemistry in single molecular bonds, Annu. Rev. Biophys. Biomol. Struct. **30** (2001) 105

24.21 T. E. Fisher, P. E. Marszalek, J. M. Fernandez: Stretching single molecules into novel conformations using the atomic force microscope, Nature Struct. Bio. **7** (2000) 719

24.22 M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, H. Gaub: Reversible unfolding of individual titin immunoglobulin domains by AFM, Science **276** (1997) 1109

24.23 K. Svoboda, S. M. Block: Biological applications of optical forces, Annu. Rev. Biophys. Biomol. Struct. **23** (1994) 247

24.24 C. Bustamante, J. C. Macosko, G. J. L. Wuite: Grabbing the cat by the tail: Manipulating molecules one by one, Nature Rev. Mol. Cell Bio. **1** (2000) 130

24.25 K. Svoboda, S. M. Block: Force and velocity measured for single kinesin molecules, Cell **77** (1994) 773

24.26 M. J. Schnitzer, S. M. Block: Kinesin hydrolyses one ATP per 8-nm step, Nature **388** (1997) 386

24.27 B. Maier, T. R. Strick, V. Croquette, D. Bensimon: Study of DNA motors by single molecule micromanipulation, Single Mol. **1** (2000) 145

24.28 A. A. Simpson, Y. Tao, P. G. Leiman, M. O. Badasso, Y. He, P. J. Jardine, N. H. Olson, M. C. Morais, S. Grimes, D. L. Anderson, T. S. Baker, M. G. Rossmann: Structure of the bacteriophage ϕ-29 DNA packaging motor, Nature **408** (2000) 745

24.29 D. E. Smith, S. J. Tans, S. B. Smith, S. Grimes, D. L. Anderson, C. Bustamante: The bacteriophage ϕ-29 portal motor can package DNA against a large internal force, Nature **413** (2001) 748

24.30 D. Bray: Reasoning for results, Nature **412** (2001) 863

24.31 G. Wu, H. Ji, K. Hansen, T. Thundat, R. Datar, R. Cote, M. F. Hagan, A. K. Charkraborty, A. Majumdar: Origin of nanomechanical cantilever motion generated from biomolecular interactions, Proc. Nat. Acad. Sci. **98** (2001) 1560

24.32 L. D. Landau, E. M. Lifshitz: *Theory of Elasticity* (Pergamon, Oxford 1986)

24.33 D. H. Bamford, R. J. C. Gilbert, J. M. Grimes, D. I. Stuart: Macromolecular assemblies: greater than their parts, Curr. Op. Struc. Biology **11** (2001) 107

24.34 J. Frank: Single-particle imaging of macromolecules by cryo-electron microscopy, Annu. Rev. Biophys. Biomol. Struct. **31** (2002) 303

24.35 S. A. Darst: Bacterial RNA polymerase, Curr. Op. Struc. Biology **11** (2001) 155

24.36 N. Ban, P. Nissen, J. Hansen, P. B. Moore, T. A. Steitz: The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution, Science **289** (2000) 905

24.37 T. S. Baker, N. H. Olson, S. D. Fuller: Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs, Microbiol. Mol. Bio. Rev. **63** (1999) 862

24.38 H.-S. Chan, K. A. Dill: The protein folding problem, Phys. Today **46** (February 1993) 24

24.39 M. D. Wang, M. J. Schnitzer, H. Yin, R. Landick, J. Gelles, S. M. Block: Force and velocity measured for single molecules of RNA polymerase, Science **282** (1998) 902

24.40 M. L. Roukes: Nanoelectromechanical Systems, Technical Digest of the 2000 Solid-State Sensor and Actuator Workshop (2000)

24.41 K. Adachi, R. Yasuda, H. Noji, H. Itoh, Y. Harada, M. Yoshida, K. Kinosita: Stepping rotation of F_1-ATPase visualized through angle-resolved single-fluorophore imaging, Proc. Nat. Acad. Sci. **97** (2000) 7243

24.42 J. Wrigglesworth: *Energy and Life* (Taylor and Francis, London 1997)

24.43 V. Smil: *Energies* (MIT, Cambridge 1999)

24.44 J. T. Finer, R. M. Simmons, J. A. Spudich: Single myosin molecule mechanics: Piconewton forces and nanometre steps, Nature **368** (1994) 113

24.45 S. B. Smith, Y. Cui, C. Bustamante: Overstretching B-DNA: The elastic response of individual double-stranded and single-sranded DNA molecules, Science **271** (1996) 795

24.46 S. Munevar, Y. Wang, M. Dembo: Traction force microscopy of migrating normal and H-ras transformed 3T3 fibroblasts, Biophys. J. **80** (2001) 1744

24.47 L. Mahadevan, P. Matsudaira: Motility powered by supramolecular springs and ratchets, Science **288** (2000) 95

24.48 R. Phillips, M. Dittrich, K. Schulten: Quasicontinuum representations of atomic-scale mechanics: From proteins to dislocations, Ann. Rev. Mat. Sci. **32** (2002) 219

24.49 T. Strick, J.-F. Allemand, V. Croquette, D. Bensimon: Twisting and stretching single DNA molecules, Prog. Biophys. Mol. Bio. **74** (2000) 115

24.50 D. Frenkel, B. Smit: *Understanding Molecular Simulation* (Academic, San Diego 1996)

24.51 H. Lu, B. Isralewitz, A. Krammer, V. Vogel, K. Schulten: Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation, Biophys. J. **75** (1998) 662

24.52 M. Carrion-Vazquez, A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, J. M. Fernandez: Mechanical design of proteins studies by single-molecule force spectroscopy and protein engineering, Prog. Biophys. Mol. Bio. **74** (2000) 63

24.53 A. Y. Grosberg, A. R. Khokhlov: *Giant Molecules* (Academic, San Diego 1997)

24.54 M. Ptashne, A. Gann: *Genes and Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor 2002)

24.55 A. Balaeff, L. Mahadevan, K. Schulten: Elastic rod model of a DNA loop in the Lac Operon, Phys. Rev. Lett. **83** (1999) 4900

24.56 U. Seifert: Configurations of fluid membranes and vesicles, Adv. Phys. **46** (1997) 13

24.57 D. Boal: *Mechanics of the Cell* (Cambridge Univ. Press, Cambridge 2002)

24.58 D. Leckband, J. Israelachvili: Intermolecular forces in biology, Q. Rev. Biophys. **34** (2001) 105

24.59 S. C. Riemer, V. A. Bloomfield: Packaging of DNA in bacteriophage heads: Some considerations on energetics, Biopolymers **17** (1978) 785

24.60 T. Odijk: Hexagonally packed DNA within bacteriophage T7 stabilized by curvature stress, Biophys. J. **75** (1998) 1223

24.61 J. Kindt, S. Tzlil, A. Ben-Shaul, W. Gelbart: DNA packaging and ejection forces in bacteriophage, Proc. Nat. Acad. Sci. **98** (2001) 13671

24.62 K. E. Richards, R. C. Williams, R. Calendar: Mode of DNA packing within bacteriophage heads, J. Mol. Bio. **78** (1973) 255

24.63 M. E. Cerritelli, N. Cheng, A. H. Rosenberg, C. E. McPherson, F. P. Booy, A. C. Steven: Encapsidated conformation of bacteriophage T7 DNA, Cell **91** (1997) 271

24.64 N. H. Olson, M. Gingery, F. A. Eiserling, T. S. Baker: The structure of isomeric capsids of bacteriophage T4, Virology **279** (2001) 385

24.65 D. C. Rau, B. Lee, V. A. Parsegian: Measurement of the repulsive force between polyelectrolyte molecules in ionic solution: Hydration forces between parallel DNA double helices, Proc. Natl. Acad. Sci. **81** (1984) 2621

24.66 D. C. Rau, V. A. Parsegian: Direct measurement of the intermolecular forces between counterion-condensed DNA double helices, Biophys. J. **61** (1992) 246

24.67 V. A. Parsegian, R. P. Rand, N. L. Fuller, D. C. Rau: Osmotic stress for the direct measurement of intermolecular forces, Meth. Enzymol. **127** (1986) 400

24.68 R. Phillips: *Crystals, Defects and Microstructures* (Cambridge Univ. Press, Cambridge 2001)

24.69 V. S. Reddy, H. A. Giesing, R. T. Morton, A. Kumar, C. B. Post, C. L. Brooks, J. E. Johnson: Energetics of quasiequivalence: Computational analysis of protein-protein interactions in icosahedral viruses, Biophys. J. **74** (1998) 546
The parameters can be found at http://www.scripps.edu/pub/olson-web/gmm/autodock/ad305/Using_AutoDock_305.a.html

25. Mechanical Properties of Nanostructures

Knowledge of the mechanical properties of nanostructures is necessary for designing realistic MEMS/NEMS devices. Microelectromechanical systems (MEMS) refer to microscopic devices that have a characteristic length of less than 1 mm but more than 1 μm and combine electrical and mechanical components. Nanoelectromechanical systems (NEMS) refer to nanoscopic devices that have a characteristic length of less than 1 μm and combine electrical and mechanical components.

Elastic and inelastic properties are needed to predict deformation from an applied load in the elastic and inelastic regimes, respectively. The strength property is needed to predict the allowable operating limit. Some of the properties of interest are hardness, elastic modulus, bending strength, fracture toughness, and fatigue strength.

Structural integrity is of paramount importance in all devices. Load applied during the use of devices can result in component failure. Cracks can develop and propagate under tensile stresses, leading to failure. Atomic force microscopy and nanoindenters can be used satisfactorily to evaluate the mechanical properties of micro/nanoscale structures for use in MEMS/NEMS.

The most commonly used materials are single-crystal silicon and silicon-based materials, e.g., SiO_2 and polysilicon films deposited by low-pressure chemical vapor deposition. An early study showed silicon to be a mechanically resilient material in addition to its favorable electronic properties. Single-crystal SiC deposited on large-area silicon substrates is used for high-temperature micro/nanosensors and actuators. Amorphous alloys can be formed on both metal and silicon substrates by sputtering and plating techniques, providing more flexibility in surface-integration. Electroless deposited Ni-P amorphous thin films have been used to construct microdevices, especially using the so-called LIGA techniques. Micro/nanodevices need conductors to provide power, as well as electrical/magnetic signals to make them functional. Electroplated gold films have found wide applications in electronic devices because of their ability to make thin films and process simply. Use of SiC, Ni-P, and Au films, together with silicon and silicon-based materials, opens up new design opportunities for MEMS/NEMS devices.

This chapter presents a review of mechanical property measurements on the nanoscale of various materials of interest and stress and deformation analyses of nanostructures.

25.1 Experimental Techniques for Measurement of Mechanical Properties of Nanostructures 765
25.1.1 Indentation and Scratch Tests Using Micro/Nanoindenters 765
25.1.2 Bending Tests of Nanostructures Using an AFM 765
25.1.3 Bending Tests Using a Nanoindenter 769

25.2 Experimental Results and Discussion 770
25.2.1 Indentation and Scratch Tests of Various Materials Using Micro/Nanoindenters 770
25.2.2 Bending Tests of Nanobeams Using an AFM 773
25.2.3 Bending Tests of Microbeams Using a Nanoindenter 777

25.3 Finite Element Analysis of Nanostructures with Roughness and Scratches 778
25.3.1 Stress Distribution in a Smooth Nanobeam 779
25.3.2 Effect of Roughness in the Longitudinal Direction 781
25.3.3 Effect of Roughness in the Transverse Direction and Scratches 781
25.3.4 Effect on Stresses and Displacements for Materials That Are Elastic, Elastic-Plastic, or Elastic-Perfectly Plastic 784

25.4 Closure 785

References 786

Microelectromechanical systems (MEMS) refer to microscopic devices that have a characteristic length of less than 1 mm but more than 1 µm and combine electrical and mechanical components. Nanoelectromechanical systems (NEMS) refer to nanoscopic devices that have a characteristic length of less than 1 µm and combine electrical and mechanical components. To put the dimensions in perspective, individual atoms are typically fraction of a nanometer in diameter, DNA molecules are about 2.5 nm wide, biological cells are in the range of thousands of nm in diameter, and human hair is about 75 µm in diameter. The mass of a micromachined silicon structure can be as low as 1 nN, and NEMS can be built with a mass as low as 10^{-20} N with cross sections of about 10 nm. In comparison, the mass of a drop of water is about 10 µN and the mass of an eyelash is about 100 nN. The acronym MEMS originated in the United States. The term commonly used in Europe and Japan is micro/nanodevices, which is used in a much broader sense. MEMS/NEMS terms are also now used in a broad sense. A micro/nanosystem, a term commonly used in Europe, is referred to as an intelligent miniaturized system comprising sensing, processing, and/or actuating functions.

A wide variety of MEMS, including Si-based devices, chemical and biological sensors and actuators, and miniature non-silicon structures (e.g., devices made from plastics or ceramics) have been fabricated with dimensions in the range of a couple to a few thousand microns (see e.g., [25.1–9]). A variety of NEMS have also been produced [25.10–15]. Two of the largest "killer" industrial applications of MEMS are accelerometers (about 85 million units in 2002) and digital micromirror devices (about $ 400 million in revenues in 2002). BIOMEMS and BIONEMS are increasingly used in commercial applications. The largest applications of BIOMEMS include silicon-based disposable blood pressure sensor chips for blood pressure monitoring (about 20 million units in 2002) and a variety of biosensors.

Structural integrity is of paramount importance in all devices. Load applied during the use of devices can result in component failure. Cracks can develop and propagate under tensile stresses, leading to failure [25.16, 17]. Friction/stiction and wear limit the lifetimes and compromise the performance and reliability of the devices involving relative motion [25.4, 18]. Most MEMS/NEMS applications demand extreme reliability. Stress and deformation analyses are carried out for an optimal design. MEMS/NEMS designers require mechanical properties on the nanoscale. Mechanical properties include elastic, inelastic (plastic, fracture, or viscoelastic), and strength. Elastic and inelastic properties are needed to predict deformation from an applied load in the elastic and inelastic regimes, respectively. The strength property is needed to predict the allowable operating limit. Some of the properties of interest are hardness, elastic modulus, bending strength (fracture stress), fracture toughness, and fatigue strength. Micro/nanostructures have some surface topography and local scratches dependent upon the manufacturing process. Surface roughness and local scratches may compromise the reliability of the devices and their effect needs to be studied.

Most mechanical properties are size-dependent [25.11, 16, 17]. Several researchers have measured mechanical properties of silicon and silicon-based milli- to microscale structures using tensile tests and bending tests [25.19–28], resonant structure tests for measurement of elastic properties [25.29], fracture toughness tests [25.20, 22, 30–34], and fatigue tests [25.32, 35, 36]. Most recently, a few researchers have measured mechanical properties of nanoscale structures using atomic force microscopy (AFM) [25.37, 38] and nanoindentation [25.39, 40]. For stress and deformation analyses of simple geometries and boundary condition, analytical models can be used. For analysis of complex geometries, numerical models are needed. The conventional finite element method (FEM) can be used down to a few tens of nanometers, although its applicability is questionable at the nanoscale. FEM has been used for simulation and prediction of residual stresses and strains induced in MEMS devices during fabrication [25.41], to perform fault analysis in order to study MEMS faulty behavior [25.42], to compute mechanical strain resulting from the doping of silicon [25.43], and to analyze micromechanical experimental data [25.22, 44, 45] and nanomechanical experimental data [25.38]. FEM analysis of nanostructures has been carried out to analyze the effect of types of surface roughness and scratches on stresses in nanostructures [25.46, 47].

The most commonly used materials for MEMS/NEMS are single-crystal silicon and silicon-based materials (e.g., SiO_2 and polysilicon films deposited by low pressure chemical vapor deposition (LPCVD) process) [25.11]. An early study showed silicon to be a mechanically resilient material in addition to its favorable electronic properties [25.48]. Single-crystal 3C-SiC (cubic or β-SiC) films deposited by atmospheric pressure chemical vapor deposition (APCVD) process on large-area silicon substrates are produced for high-temperature micro/nanosensor and actuator applications [25.49–51]. Amorphous alloys can be formed on both metal and silicon substrates by sputtering and plating techniques, providing more flexibility in surface-integration. Electroless deposited Ni-P amorphous thin films have been used to construct microdevices, especially using the so-called LIGA techniques [25.11, 40]. Micro/nanodevices need conductors to provide power, as well as electrical/magnetic signals to make them functional. Electroplated gold films have found wide applications in electronic devices, because of their ability to make thin films and process simplicity [25.40]. Use of SiC, Ni-P, and Au films together with silicon and silicon-based materials open up new design opportunities for MEMS/NEMS devices.

This chapter presents a review of mechanical property measurements on the nanoscale of various materials of interest, and stress and deformation analyses of nanostructures.

25.1 Experimental Techniques for Measurement of Mechanical Properties of Nanostructures

25.1.1 Indentation and Scratch Tests Using Micro/Nanoindenters

A nanoindenter is commonly used to measure hardness, elastic modulus, and fracture toughness and to perform micro/nanoscratch studies to get a measure of scratch/wear resistance of materials [25.11, 52].

Hardness and Elastic Modulus

The nanoindenter monitors and records the dynamic load and displacement of the three-sided pyramidal diamond (Berkovich) indenter during indentation with a force resolution of about 75 nN and displacement resolution of about 0.1 nm. Hardness and elastic modulus are calculated from the load-displacement data [25.11]. The peak indentation load depends on the mechanical properties of the specimen; a harder material requires higher load for a reasonable indentation depth.

Fracture Toughness

The indentation technique for fracture toughness measurement on the microscale is based on the measurement of the lengths of median-radial cracks produced by indentation. A Vickers indenter (a four-sided diamond pyramid) is used in a microhardness test. A load on the order of 0.5 N is typically used in making the Vickers indentations. The indentation impressions are examined using an optical microscope with Nomarski interference contrast to measure the length of median-radial cracks, c. The fracture toughness (K_{IC}) is calculated by the following relation [25.53]:

$$K_{\mathrm{IC}} = \alpha \left(\frac{E}{H}\right)^{1/2} \left(\frac{P}{c^{3/2}}\right) , \tag{25.1}$$

where α is an empirical constant depending on the geometry of the indenter, H and E are hardness and elastic moduli, and P is the peak indentation load. For Vickers indenters, α has been empirically found based on experimental data and is equal to 0.016 [25.11]. Both E and H values are obtained from the nanoindentation data. The crack length is measured from the center of the indent to the end of the crack using an optical microscope. For one indent, all crack lengths are measured. The crack length c is obtained from the average values of several indents.

Scratch Resistance

In micro/nanoscratch studies, a conical diamond indenter with a tip radius of about 1 µm and an included angle of 60° is drawn over the sample surface, and the load is ramped up until substantial damage occurs [25.11]. The coefficient of friction is monitored during scratching. In order to obtain scratch depths during scratching, the surface profile of the sample surface is first obtained by translating the sample at a low load of about 0.2 mN, which is insufficient to damage a hard sample surface. The 500-µm-long scratches are made by translating the sample while ramping the loads on the conical tip over different loads depending on the material hardness. The actual depth during scratching is obtained by subtracting the initial profile from the scratch depth measured during scratching. In order to measure the scratch depth after the scratch, the scratched surface is profiled at a low load of 0.2 mN and is subtracted from the actual surface profile before scratching.

25.1.2 Bending Tests of Nanostructures Using an AFM

Quasi-static bending tests of fixed nanobeam arrays are carried out using an AFM [25.37, 38, 54]. A three-sided pyramidal diamond tip (with a radius of about 200 nm)

Springer Handbook of Nanotechnology
B. Bhushan • © Springer 2004

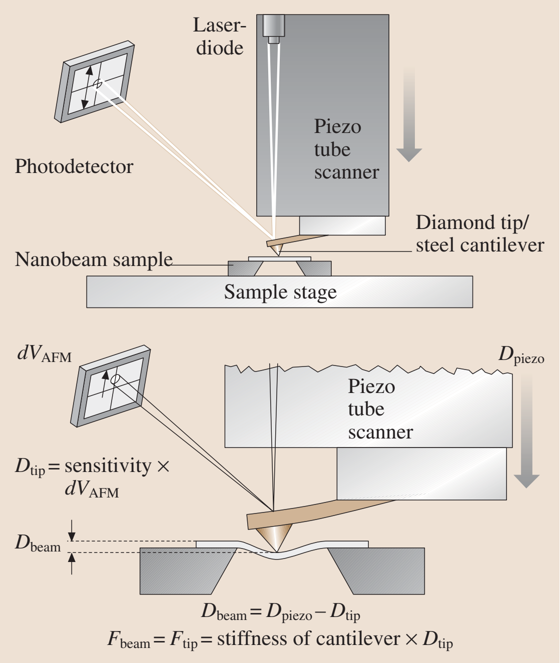

Fig. 25.1 Schematic showing the details of a nanoscale bending test. The AFM tip is brought to the center of the nanobeam and the piezo is extended over a known distance. By measuring the tip displacement, a load-displacement curve of the nanobeam can be obtained [25.38]

mounted on a rectangular stainless steel cantilever is used for the bending tests. The beam stiffness is selected based on the desired load range. The stiffness of the cantilever beams for application of normal load up to 100 µN is about 150–200 N/m.

The wafer with nanobeam array is fixed onto a flat sample chuck using double-stick tape [25.38]. For the bending test, the tip is brought over the nanobeam array with the help of the sample stage of the AFM and a built-in high magnification optical microscope (Fig. 25.1). For the fine positioning of the tip over a chosen beam, the array is scanned in contact mode at a contact load of about 2–4 µN, which results in negligible damage to the sample. After scanning, the tip is located at one end of a chosen beam. To position the tip at the center of the beam span, the tip is moved to the other end of the beam by giving the X-piezo an offset voltage. The value of this offset is determined after several such attempts have been made in order to minimize effects of piezo drift. Half of this offset is then applied to the X-piezo after the tip is positioned at one end of the beam, which usually results in the tip being moved to the center of the span. Once the tip is positioned over the center of the beam span, the tip is held stationary without scanning, and the Z-piezo is extended by a known distance, typically about 2.5 µm, at a rate of 10 nm/s, as shown in Fig. 25.1. During this time, the vertical deflection signal ($\mathrm{d}V_{\mathrm{AFM}}$), which is proportional to the deflection of the cantilever (D_{tip}), is monitored. The displacement of the piezo is equal to the sum of the displacements of the cantilever and nanobeam. Hence, the displacement of the nanobeam (D_{beam}) under the point of load can be determined as

$$D_{\mathrm{beam}} = D_{\mathrm{piezo}} - D_{\mathrm{tip}} \,. \tag{25.2}$$

The load (F_{beam}) on the nanobeam is the same as the load on the tip/cantilever (F_{tip}) and is given by

$$F_{\mathrm{beam}} = F_{\mathrm{tip}} = D_{\mathrm{tip}} \times k \,, \tag{25.3}$$

where k is the stiffness of the tip/cantilever. In this manner, a load displacement curve for each nanobeam can be obtained.

The photodetector sensitivity of the cantilever needs to be calibrated to obtain D_{tip} in nm. For this calibration, the tip is pushed against a smooth diamond sample by moving the Z-piezo over a known distance. For the hard diamond material, the actual deflection of the tip can be assumed to be the same as the Z-piezo travel (D_{piezo}), and the photodetector sensitivity (S) for the cantilever setup is determined as

$$S = D_{\mathrm{piezo}}/\mathrm{d}V_{\mathrm{AFM}}\ \mathrm{nm/V} \,. \tag{25.4}$$

In the measurements, D_{tip} is given as $\mathrm{d}V_{\mathrm{AFM}} \times S$.

Since a sharp tip would result in an undesirable large local indentation, *Sundararajan* and *Bhushan* [25.38] used a worn (blunt) diamond tip. Indentation experiments using this tip on a silicon substrate yielded a residual depth of less than 8 nm at a maximum load of 120 µN, which is negligible compared to displacements of the beams (several hundred nm). Hence, we can assume that negligible local indentation or damage is created during the bending process of the beams, and that the displacement calculated from (25.2) is entirely that of the beam structure.

Elastic Modulus and Bending Strength

Elastic modulus and bending strength (fracture stress) of the beams can be estimated by equations based on the assumption that the beams follow linear elastic theory of an isotropic material. This is probably valid since the beams have high length-to-width (ℓ/w) and length-to-thickness (ℓ/t) ratios and also since the length direction is along the principal stress direction during the test. For a fixed elastic beam loaded at the center of the span, the elastic modulus is expressed as

$$E = \frac{\ell^3}{192I} m , \tag{25.5}$$

where ℓ is the beam length, I is the area moment of inertia for the beam cross section, and m is the slope of the load-displacement curve during bending [25.55]. The area moment of inertia for a beam with a trapezoidal cross section is calculated from the following equation:

$$I = \frac{w_1^2 + 4w_1w_2 + w_2^2}{36(w_1 + w_2)} t^3 , \tag{25.6}$$

where w_1 and w_2 are the upper and lower widths, respectively, and t is the thickness of the beam. According to linear elastic theory, for a centrally loaded beam, the moment diagram is shown in Fig. 25.2. The maximum moments are generated at the ends (negative moment) and under the loading point (positive moment), as shown in Fig. 25.2. The bending stresses generated in the beam are proportional to the moments and are compressive or tensile about the neutral axis (line of zero stress). The maximum tensile stress (σ_b, which is the bending strength or fracture stress) is produced on the top surface at both the ends and is given by [25.55]

$$\sigma_b = \frac{F_{\max} \ell e_1}{8I} , \tag{25.7}$$

where $F_{\max}$ is the applied load at failure and e_1 is the distance of the top surface from the neutral plane of the beam cross section and is given by [25.55]

$$e_1 = \frac{t(w_1 + 2w_2)}{3(w_1 + w_2)} . \tag{25.8}$$

Although the moment value at the center of the beam is the same as at the ends, the tensile stresses at the center (generated on the bottom surface) are less than those generated at the ends, as per (25.7), because the distance from the neutral axis to the bottom surface is less than e_1. This is because of the trapezoidal cross section of the beam, which results in the neutral axis being closer to the bottom surface than the top (Fig. 25.2).

In the preceding analysis, the beams were assumed to have fixed ends. However, in the nanobeams used by *Sundararajan* and *Bhushan* [25.38], the underside of the beams was pinned over some distance on either side of the span. Hence, a finite element model of the beams was created to see if the difference in the boundary conditions affected the stresses and displacements of the beams. It was found that the difference in the stresses was less than 1%. This indicates that the boundary conditions near the ends of the actual beams are not that different from that of the fixed ends. Therefore, the bending strength values can be calculated from (25.7).

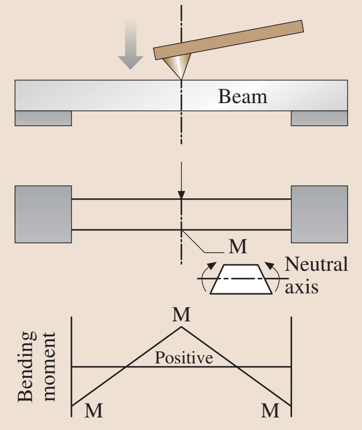

Fig. 25.2 A schematic of the bending moments generated in the beam during a quasi-static bending experiment with the load at the center of the span. The maximum moments occur under the load and at the fixed ends. Due to the trapezoidal cross section, the maximum tensile bending stresses occur at the top surfaces at the fixed ends

Fracture Toughness

Fracture toughness is another important parameter for brittle materials such as silicon. In the case of the nanobeam arrays, these are not best suited for fracture toughness measurements, because they do not possess regions of uniform stress during bending. *Sundararajan* and *Bhushan* [25.38] developed a methodology, and its steps are outlined schematically in Fig. 25.3a. First, a crack of known geometry is introduced in the region of maximum tensile bending stress, i. e., on the top surface near the ends of the beam. This is achieved by generating a scratch at high normal load across the width (w_1) of the beam using a sharp diamond tip (radius < 100 nm). A typical scratch thus generated is shown in Fig. 25.3b. By bending the beam as shown, a stress concentration will be formed under the scratch. This will lead to failure of the beam under the scratch once a critical load (fracture load) is attained. The fracture load and relevant dimensions of the scratch are input into the FEM model, which is used to generate the fracture stress plots. Figure 25.3c shows an FEM simulation of one such experiment, which reveals that the maximum stress does occur under the scratch.

If we assume that the scratch tip acts as a crack tip, a bending stress will tend to open the crack in Mode I. In this case, the stress field around the crack tip can be described by the stress intensity parameter K_I (for Mode I) for linear elastic materials [25.56]. In particular, the stresses corresponding to the bending stresses are

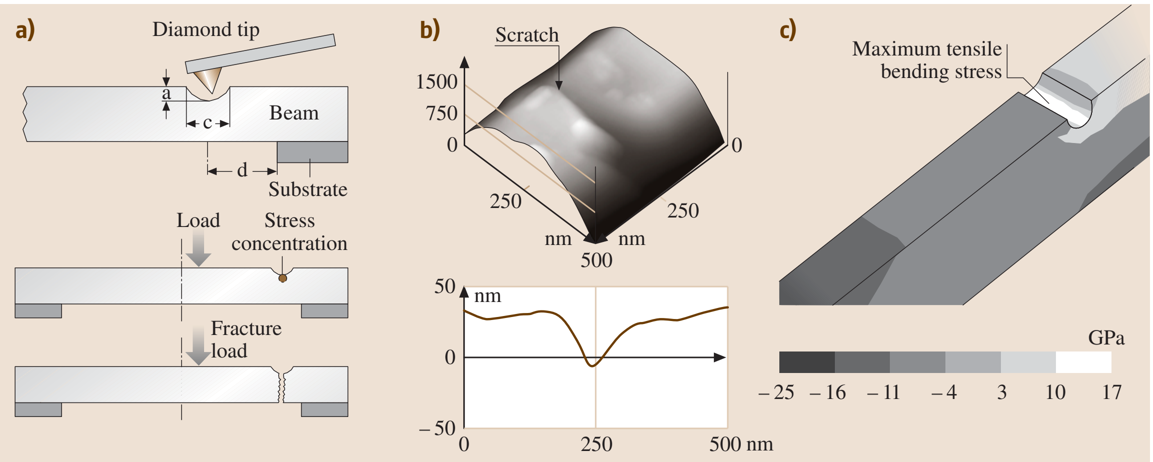

Fig. 25.3 **(a)** Schematic of technique to generate a defect (crack) of known dimensions in order to estimate fracture toughness. A diamond tip is used to generate a scratch across the width of the beam. When the beam is loaded as shown, a stress concentration is formed at the bottom of the scratch. The fracture load is then used to evaluate the stresses using FEM. **(b)** AFM 3-D image and 2-D profile of a typical scratch. **(c)** Finite element model results verifying that the maximum bending stress occurs at the bottom of the scratch [25.38]

described by

$$\sigma = \frac{K_\mathrm{I}}{\sqrt{2\pi r}} \cos\left(\frac{\theta}{2}\right)\left[1+\sin\left(\frac{\theta}{2}\right)\sin\left(\frac{3\theta}{2}\right)\right] \qquad (25.9)$$

for every point $p(r,\theta)$ around the crack tip, as shown in Fig. 25.4. If we substitute the fracture stress (σ_f) to the left hand side of (25.9), then the K_I value can be substituted with its critical value, which is the fracture toughness K_IC. Now, the fracture stress can be determined for the point ($r = 0$, $\theta = 0$), i. e., right under the crack tip as explained above. However, we cannot substitute $r = 0$ in (25.9). The alternative is to substitute a value for r, which is as close to zero as possible. For silicon, a reasonable number is the distance between neighboring atoms in the (111) plane, the plane along which silicon exhibits the lowest fracture energy. This value was calculated from silicon unit cell dimensions of 0.5431 nm [25.57] to be 0.4 nm (half of the face diagonal). This assumes that Si displays no plastic zone around the crack tip, which is reasonable since in tension, silicon is not known to display much plastic deformation at room temperature. *Sundararajan* and *Bhushan* [25.38] used values $r = 0.4–1.6$ nm (i. e., distances up to four times the distance between the nearest neighboring atoms) to estimate the fracture toughness for both Si and $\mathrm{SiO_2}$ according to the following equation:

$$K_\mathrm{IC} = \sigma_\mathrm{f}\sqrt{2\pi r} \qquad r = 0.4–1.6\,\mathrm{nm}\,. \qquad (25.10)$$

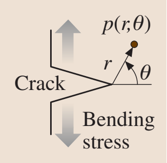

Fig. 25.4 Schematic of crack tip and coordinate systems used in (25.9) to describe a stress field around the crack tip in terms of the stress intensity parameter, K_I [25.38]

Fatigue Strength

In addition to the properties mentioned so far that can be evaluated from quasi-static bending tests, the fatigue properties of nanostructures are also of interest. This is especially true for MEMS/NEMS involving vibrating structures such as oscillators and comb drives [25.58] and hinges in digital micromirror devices [25.59]. To study the fatigue properties of the nanobeams, *Sundararajan* and *Bhushan* [25.38] applied monotonic cyclic stresses using an AFM, Fig. 25.5a. Similar to the bending test, the diamond tip is first positioned at the center of the beam span. In order to ensure that the tip is always in contact with the beam (as opposed to impacting it), the piezo is first extended by a distance D_1, which ensures a minimum stress on the beam. After this extension, a cyclic displacement of amplitude, D_2, is applied continuously until failure of the beam occurs.

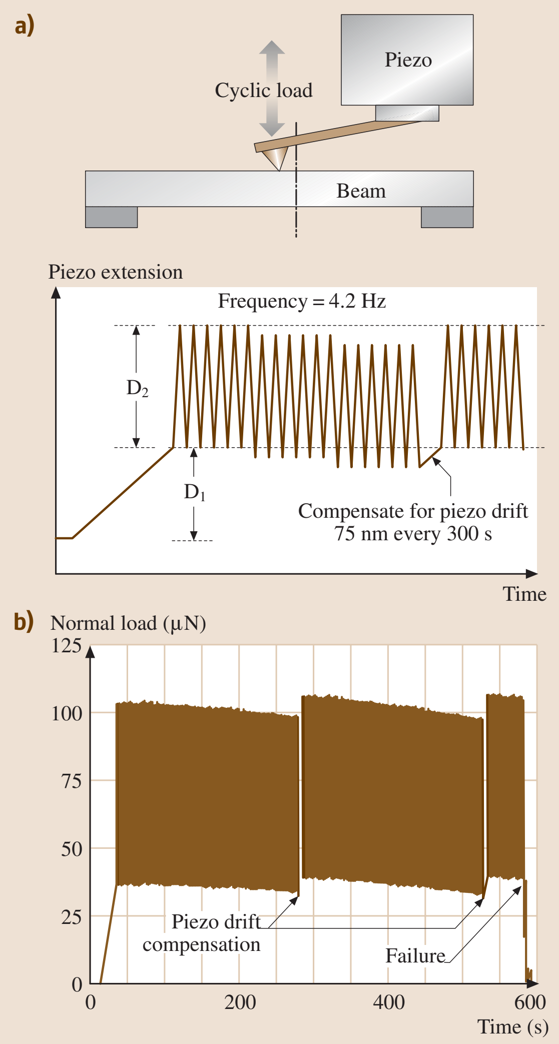

Fig. 25.5 **(a)** Schematic showing the details of the technique to study fatigue behavior of the nanobeams. The diamond tip is located at the middle of the span, and a cyclic load at 4.2 Hz is applied to the beam by forcing the piezo to move in the pattern shown. An extension is made every 300 s to compensate for the piezo drift to ensure that the load on the beam is kept fairly constant. **(b)** Data from a fatigue experiment on a nanobeam until failure. The normal load is computed from the raw vertical deflection signal. The compensations for piezo drift keep the load fairly constant [25.38]

This results in the application of a cyclic load to the beam. The maximum frequency of the cyclic load that could be attained using the AFM by *Sundararajan* and *Bhushan* [25.38] was 4.2 Hz. The vertical deflection signal of the tip is monitored throughout the experiment. The signal follows the pattern of the piezo input up to failure, which is indicated by a sudden drop in the signal. During initial runs, piezo drift was observed that caused the piezo to gradually move away from the beam (i. e., to retract), resulting in a continuous decrease in the applied normal load. In order to compensate for this, the piezo is given a finite extension of 75 nm every 300 s, as shown in Fig. 25.5a. This results in keeping the applied loads fairly constant. The normal load variation (calculated from the vertical deflection signal) from a fatigue test is shown in Fig. 25.5b. The cyclic stress amplitudes (corresponding to D_2) and fatigue lives are recorded for every sample tested. Values for D_1 are set such that minimum stress levels are about 20% of the bending strengths.

25.1.3 Bending Tests Using a Nanoindenter

Quasi-static bending tests of micro/nanostructures are also carried out using a nanoindenter [25.39, 40]. The advantage of the nanoindenter is that loads up to about 500 mN, higher than that in AFM (up to about 100 μN), can be used for structures requiring high loads for experiments. To avoid the indenter tip pushing into the specimen, a blunt tip is used in the bending and fatigue tests. *Li* et al. [25.40] used a diamond conical indenter with a radius of 1 μm and an included angle of 60°. The load position used was at the center of the span for the bridge beams and 10 μm off from the free end of the cantilever beams. An optical microscope with a magnification of 1,500× or an in situ AFM is used to locate the loading position. Then the specimen is moved under the indenter location with a resolution of about 200 nm in the longitudinal direction and less than 100 nm in the lateral direction.

Using the analysis presented earlier, elastic modulus and bending strength of the beams can be obtained from the load-displacement curves [25.40]. For fatigue tests, an oscillating load is applied and contact stiffness is measured during the tests. A significant drop in the contact stiffness during the test is a measure of the number of cycles to failure [25.39].

25.2 Experimental Results and Discussion

25.2.1 Indentation and Scratch Tests of Various Materials Using Micro/Nanoindenters

Studies have been conducted on five different materials: undoped single-crystal Si(100), undoped polysilicon film, SiO_2 film, SiC film, electroless deposited Ni-11.5 wt% P amorphous film, and electroplated Au film [25.40, 50, 51]. A 3-µm thick polysilicon film was deposited by a low pressure chemical vapor deposition (LPCVD) process on a Si(100) substrate. The 1-µm thick SiO_2 film was deposited by a plasma enhanced chemical vapor deposition (PECVD) process on a Si(111) substrate. A 3-µm thick 3C-SiC film was epitaxially grown using an atmospheric pressure chemical vapor deposition (APCVD) process on a Si(100) substrate. A 12-µm thick Ni-P film was electroless plated on a 0.8-mm-thick Al-4.5 wt% Mg alloy substrate. A 3-µm-thick Au film was electroplated on a Si(100) substrate.

Hardness and Elastic Modulus

Hardness and elastic modulus measurements are made using a nanoindenter [25.40]. The hardness and elastic modulus values of various materials at a peak indentation depth of 50 nm are summarized in Fig. 25.6 and Table 25.1. The SiC film exhibits the highest hardness of about 25 GPa and an elastic modulus of about 395 GPa among the samples examined, followed by the undoped Si(100), undoped polysilicon film, SiO_2 film, Ni-P film, and Au film. The hardness and elastic modulus data of the undoped Si(100) and undoped polysilicon film are comparable. For the metal alloy films, the Ni-P film exhibits higher hardness and elastic modulus than the Au film.

Table 25.1 Hardness, elastic modulus, fracture toughness, and critical load results of the bulk single-crystal Si(100) and thin films of undoped polysilicon, SiO_2, SiC, Ni-P, and Au

Samples	**Hardness (GPa)**	**Elastic modulus (GPa)**	**Fracture toughness ($MPa\,m^{1/2}$)**	**Critical load (mN)**
Undoped Si(100)	12	165	0.75	11
Undoped polysilicon film	12	167	1.11	11
SiO_2 film	9.5	144	0.58 (Bulk)	9.5
SiC film	24.5	395	0.78	14
Ni-P film	6.5	130		0.4 (Plowing)
Au film	4	72		0.4 (Plowing)

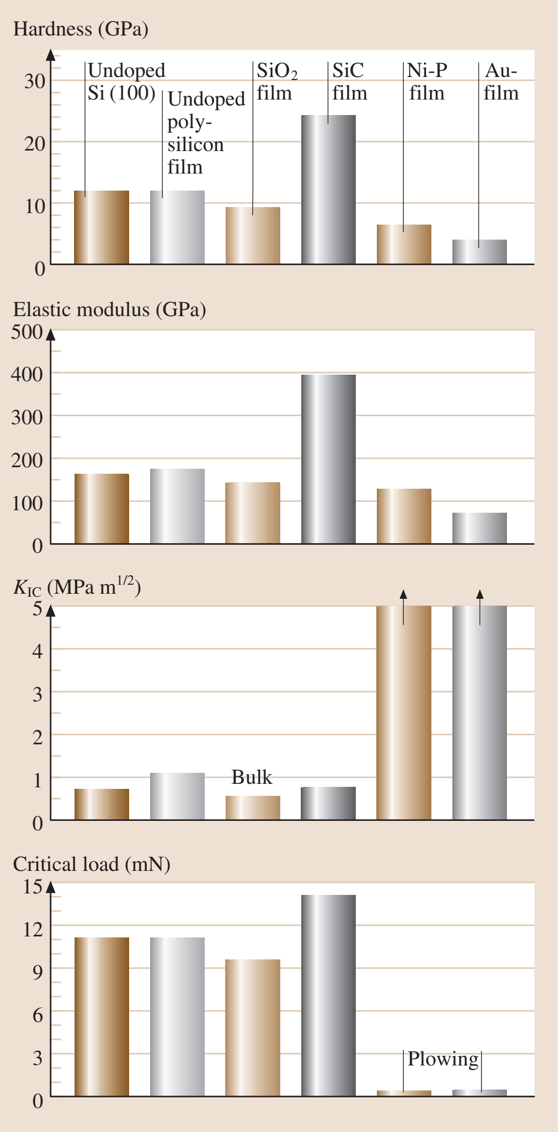

Fig. 25.6 Bar chart summarizing the hardness, elastic modulus, fracture toughness, and critical load (from scratch tests) results of the bulk undoped single-crystal Si(100) and thin films of undoped polysilicon, SiO_2, SiC, Ni-P, and Au [25.40]

Fracture Toughness

The optical images of Vickers indentations made using a microindenter at a normal load of 0.5 N held for 15 s on the undoped Si(100), undoped polysilicon film, and SiC film are shown in Fig. 25.7 [25.51]. The SiC film exhibits the smallest indentation mark, followed by the undoped polysilicon film and undoped Si(100). These Vickers indentation depths are smaller than one-third of the film thickness. Thus, the influence of the substrate on the fracture toughness of the films can be ignored. In addition to the indentation marks, radial cracks are observed emanating from the indentation corners. The SiC film shows the longest radial crack length, followed by the undoped Si(100) and undoped polysilicon film. The radial cracks for the undoped Si(100) are straight, whereas those for the SiC and undoped polysilicon film are not straight but go in a zigzag manner. The fracture toughness (K_{IC}) is calculated using (25.1).

The fracture toughness values of all samples are summarized in Fig. 25.6 and Table 25.1. The SiO_2 film used in this study is about 1 μm thick, which is not thick enough for a fracture toughness measurement. The fracture toughness values of bulk silica are listed instead for a reference. The Ni-P and Au films exhibit very high fracture toughness values that cannot be measured by indentation methods. For other samples, the undoped polysilicon film has the highest value, followed by the undoped Si(100), SiC film, and SiO_2 film. For the undoped polysilicon film, the grain boundaries can stop the radial cracks and change the propagation directions of the radial cracks, making the propagation of these cracks more difficult. Values of fracture toughness for the undoped Si(100) and SiC film are comparable. Since the undoped Si(100) and SiC film are single crystal, no grain boundaries are present to stop the radial cracks and change the propagation directions of the radial cracks. This is why the SiC film shows a lower fracture toughness value than the bulk polycrystal SiC materials of 3.6 MPa m$^{1/2}$ [25.60].

Scratch Resistance

The scratch resistance of various materials has been studied using a nanoindenter by *Li* et al. [25.40]. Figure 25.8 compares the coefficient of friction and scratch depth profiles as a function of increasing normal load and optical images of three regions over scratches: at the beginning of the scratch (indicated by A on the friction profile), at the point of initiation of damage at which the coefficient of friction increases to a high value or increases abruptly (indicated by B on the friction profile), and towards the end of the scratch (indicated by C on the friction profile) for all samples. Note that the ramp loads for the Ni-P and Au range are from 0.2 to 5 mN, whereas the ramp loads for other samples are from 0.2 to 20 mN. All samples exhibit a continuous increase in the coefficient of friction with increasing normal load from the beginning of the scratch. The continuous increase in the coefficient of friction during scratching is attributed to the increasing plowing of the sample by the tip with increasing normal load, as shown in the SEM images in Fig. 25.8. The abrupt increase in the coefficient of friction is associated with catastrophic failure, as well as significant plowing of the tip into the sample. Before the critical load, the coefficient of friction of the undoped polysilicon, SiC, and SiO_2 films increased at a slower rate and was smoother than that of the other samples. The undoped Si(100) exhibits some bursts in the friction profiles before the critical load. At the critical load, the SiC and undoped polysilicon films exhibit a small increase in the coefficient of friction, whereas the undoped Si(100) and undoped polysilicon film exhibit a sudden increase in the coefficient of friction. The Ni-P and Au films show a continuous increase in the coefficient of friction, indicating the behavior of a ductile metal. The bursts in the friction profile might result from the plastic deformation and material pile-up in front of the scratch tip. The Au film exhibits a higher coefficient of friction than the Ni-P film, because it has lower hardness and elastic modulus values than the Ni-P film.

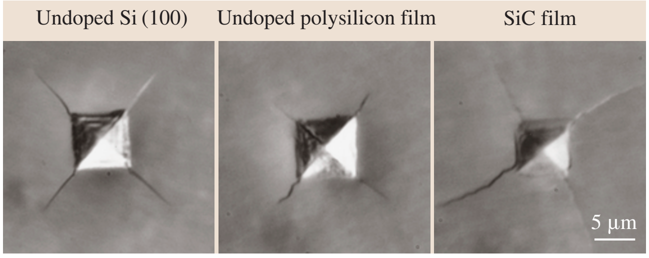

Fig. 25.7 Optical images of Vickers indentations made at a normal load of 0.5 N held for 15 s on the undoped Si(100), undoped polysilicon film, and SiC film [25.51]

The SEM images show that below the critical loads the undoped Si(100) and undoped polysilicon film were damaged by plowing, associated with the plastic flow of the material and formation of debris on the sides of the scratch. For the SiC and SiO_2 films, in region A, a plowing scratch track was found without any debris on the side of the scratch, which is probably responsible for the smoother curve and slower increase in the

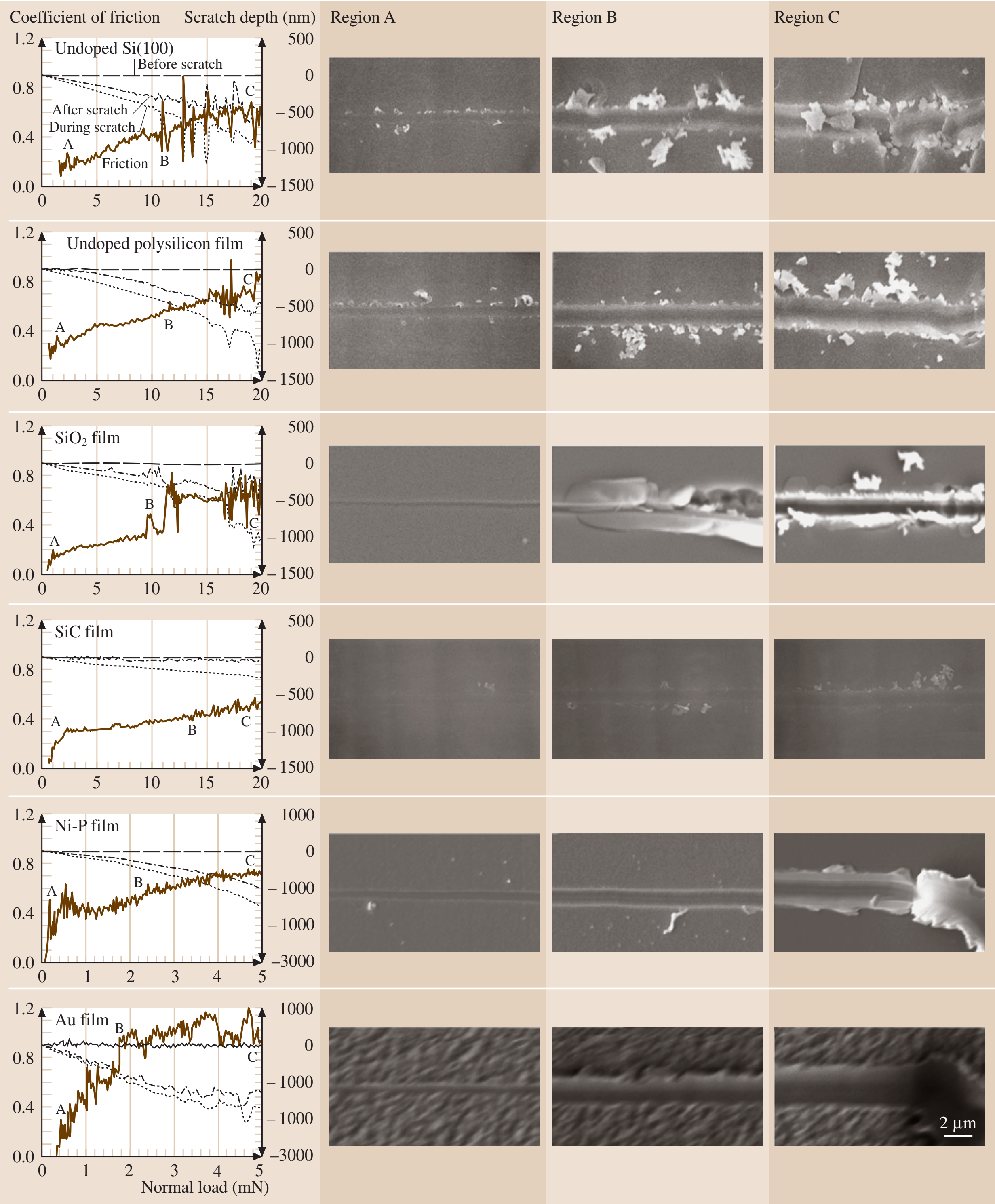
Coefficient of friction
Scratch depth (nm)
Region A
Region B
Region C
Undoped Si(100)
Before scratch
After scratch
During scratch
Friction
A
B
C
Undoped polysilicon film
A
B
C
SiO₂ film
A
B
C
SiC film
A
B
C
Ni-P film
A
B
C
Au film
A
B
C
1.2
0.8
0.4
0.0
500
0
– 500
– 1000
– 1500
1000
–3000
0
5
10
15
20
1
2
3
4
5
Normal load (mN)
2 µm

Fig. 25.8 Coefficient of friction and scratch depth profiles as a function of increasing normal load and optical images of three regions over scratches: at the beginning of the scratch (indicated by A on the friction profile), at the point of initiation of damage at which the coefficient of friction increases to a high value or increases abruptly (indicated by B on the friction profile), and towards the end of the scratch (indicated by C on the friction profile) for all samples [25.40]

coefficient of friction before the critical load. After the critical load, for the SiO_2 film, delamination of the film from the substrate occurred, followed by cracking along the scratch track. For the SiC film, only several small debris particles were found without any cracks on the side of the scratch, which is responsible for the small increase in the coefficient of friction at the critical load. For the undoped Si(100), cracks were found on the side of the scratch right from the critical load and up, which is probably responsible for the big bursts in the friction profile. For the undoped polysilicon film, no cracks were found on the side of the scratch at the critical load. This might result from grain boundaries, which can stop the propagation of cracks. At the end of the scratch, some of the surface material was torn away and cracks were found on the side of the scratch in the undoped Si(100). A couple of small cracks were found in the undoped polysilicon and SiO_2 films. No crack was found in the SiC film. Even at the end of the scratch, less debris was found in the SiC film. A curly chip was found at the end of the scratch in both Ni-P and Au films. This is a typical characteristic of ductile metal alloys. The Ni-P and Au films were damaged by plowing right from the beginning of the scratch with material pile-up at the side of the scratch.

The scratch depth profiles obtained during and after the scratch on all samples with respect to initial profile, after the cylindrical curvature is removed, are plotted in Fig. 25.8. Reduction in scratch depth is observed after scratching as opposed to during scratching. This reduction in scratch depth is attributed to an elastic recovery after removal of the normal load. The scratch depth after scratching indicates the final depth, which reflects the extent of permanent damage and plowing of the tip into the sample surface and is probably more relevant for visualizing the damage that can occur in real applications. For the undoped Si(100), undoped polysilicon film, and SiO_2 film, there is a large scatter in the scratch depth data after the critical loads, which is associated with the generation of cracks, material removal, and debris. The scratch depth profile is smooth for the SiC film. It is noted that the SiC film exhibits the lowest scratch depth among the samples examined. The scratch depths of the undoped Si(100), undoped polysilicon film, and SiO_2 film are comparable. The Ni-P and Au films exhibit much larger scratch depths than other samples. The scratch depth of the Ni-P film is smaller than that of the Au film.

The critical loads estimated from friction profiles for all samples are compared in Fig. 25.6 and Table 25.1. The SiC film exhibits the highest critical load of about 14 mN, as compared to other samples. The undoped Si(100) and undoped polysilicon film show comparable critical load of about 11 mN, whereas the SiO_2 film shows a low critical load of about 9.5 mN. The Ni-P and Au films were damaged by plowing right from the beginning of the scratch.

25.2.2 Bending Tests of Nanobeams Using an AFM

Bending tests have been performed on Si and SiO_2 nanobeam arrays [25.38, 54]. The single-crystal silicon bridge nanobeams were fabricated by bulk micromachining incorporating enhanced-field anodization using an AFM [25.37]. The Si nanobeams are oriented along the [110] direction in the (001) plane. Subsequent thermal oxidation of the beams results in the formation of SiO_2 beams. The cross section of the nanobeams is trapezoidal owing to the anisotropic wet etching process. SEM micrographs of Si and SiO_2 nanobeam arrays and a schematic of the shape of a typical nanobeam are shown in Fig. 25.9. The actual widths and thicknesses of nanobeams were measured using an AFM in tapping mode prior to tests using a standard Si tapping mode tip (tip radius < 10 nm). Surface roughness measurements of the nanobeam surfaces in tapping mode yielded a σ of 0.7 ± 0.2 nm and peak-to-valley (P–V) distance of 4 ± 1.2 nm for Si and a σ of 0.8 ± 0.3 nm and a P–V distance of 3.1 ± 0.8 nm for SiO_2. Prior to testing, the Si nanobeams were cleaned by immersing them in a "piranha etch" solution (3 : 1 solution by volume of 98% sulphuric acid and 30% hydrogen peroxide) for 600 s to remove any organic contaminants.

Bending Strength

Figure 25.10 shows typical load-displacement curves for Si and SiO_2 beams that were bent to failure [25.38, 54]. The upper width (w_1) of the beams is indicated in the figure. Also indicated are the elastic modulus values obtained from the slope of the load-displacement curve

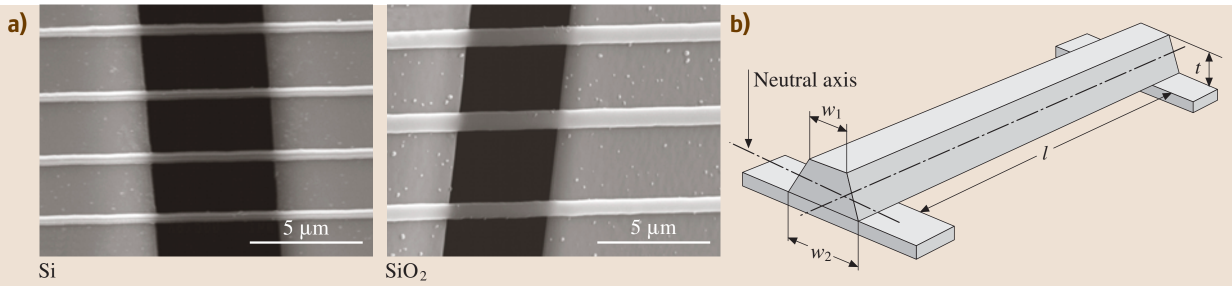

Fig. 25.9 **(a)** SEM micrographs of nanobeam arrays, and **(b)** a schematic of the shape of a typical nanobeam. The trapezoidal cross section is due to the anisotropic wet etching during the fabrication [25.54]

(25.5). All the beams tested showed linear elastic behavior followed by abrupt failure, which is suggestive of brittle fracture. Figure 25.11 shows the scatter in the values of elastic modulus obtained for both Si and SiO$_2$ along with the average values (± standard deviation). The scatter in the values may be due to differences in the orientation of the beams with respect to the trench and the loading point being a little off center with respect to the beam span. The average values are a little higher than the bulk values (169 GPa for Si[110] and 73 GPa for SiO$_2$ in Table 25.2). However, the values of E obtained from (25.5) have an error of about 20% due to the uncertainties in beam dimensions and spring constant of the tip/cantilever (which affects the measured load). Hence the elastic modulus values on the nanoscale can be considered comparable to bulk values.

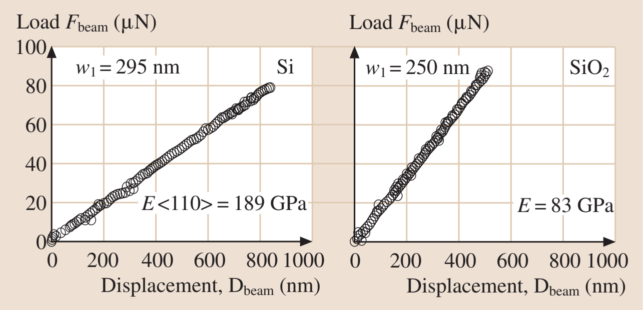

Fig. 25.10 Typical load-displacement curves of silicon and SiO$_2$ nanobeams. The curves are linear until sudden failure, indicative of brittle fracture of the beams. The elastic modulus (E) values calculated from the curves are shown. The dimensions of the Si beam were $w_1 = 295$ nm, $w_2 = 484$ nm, and $t = 255$ nm, while those of the SiO$_2$ beam were $w_1 = 250$ nm, $w_2 = 560$ nm, and $t = 425$ nm [25.54]

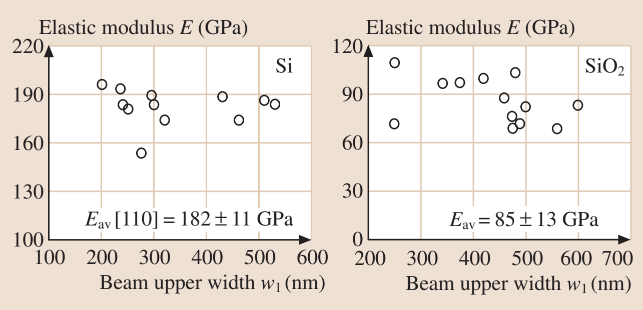

Fig. 25.11 Elastic modulus values measured for Si and SiO$_2$. The average values are shown. These are comparable to bulk values, which demonstrates that elastic modulus shows no specimen size dependence [25.38]

Most of the beams when loaded quasi-statically at the center of the span broke at the ends, as shown in Fig. 25.12a, which is consistent with the fact that maximum tensile stresses occur on the top surfaces near the ends. (See FEM stress distribution results in Fig. 25.12b.) Figure 25.13 shows the values of bending strength obtained for different beams. There appears to be no trend in bending strength with the upper width (w_1) of the beams. The large scatter is expected for the strength of brittle materials, since they are dependent on preexisting flaw population in the material and, hence, are statistical in nature. The Weibull distribution, a statistical analysis, can be used to describe the scatter in the bending strength values. The means of the Weibull distributions were found to be 17.9 GPa and 7.6 GPa for Si and SiO$_2$, respectively. Previously reported numbers of strengths range from 1 to 6 GPa for silicon [25.19, 20, 22–25, 27, 45, 61, 62] and about 1 GPa for SiO$_2$ [25.34] microscale specimens. This clearly indicates that bending strength shows a specimen size dependence. Strength of brittle materials is dependent on preexisting flaws in the material. Since for nanoscale specimens the volume is smaller than for micro- and macroscale specimens, the flaw population will be smaller as well, resulting in higher values of strength.

Table 25.2 Summary of measured parameters from quasi-static bending tests

Sample	Elastic modulus E (GPa)		Bending strength σ_b (GPa)		Fracture toughness K_{IC} (MPa $\sqrt{m}$)		
	Measured	Bulk value	Measured	Reported (microscale)	Estimated	Reported (microscale)	Bulk value
Si	182±11	169^1	18±3	$< 10^3$	1.67±0.4	$0.6-1.65^5$	0.9^6
SiO_2	85±13	73^2	7.6±2	$< 2^4$	0.60±0.2	$0.5-0.9^4$	–

[1] Si(110), [25.63] [2] [25.64] [3] [25.19, 20, 22–25, 45, 61, 62] [4] [25.34] [5] [25.30–33] [6] [25.57]

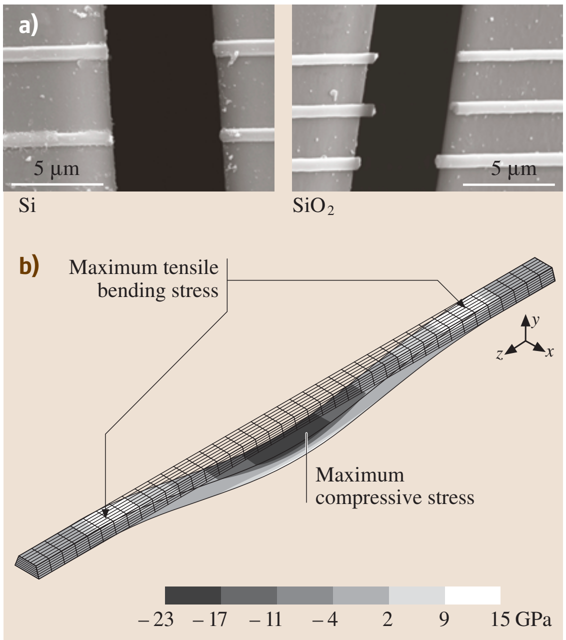

Fig. 25.12 **(a)** SEM micrographs of nanobeams that failed during quasi-static bending experiments. The beams failed at or near the ends, which is the location of maximum tensile bending stress [25.54], and **(b)** bending stress distribution for silicon nanobeam, indicating that the maximum tensile stresses occur on the top surfaces near the fixed ends

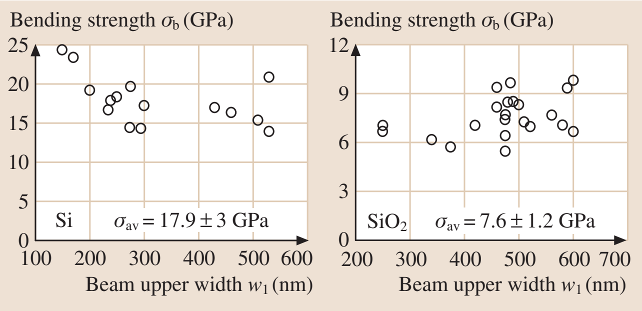

Fig. 25.13 Bending strength values obtained from bending experiments. Average values are indicated. These values are much higher than values reported for microscale specimens, indicating that bending strength shows a specimen size effect [25.38]

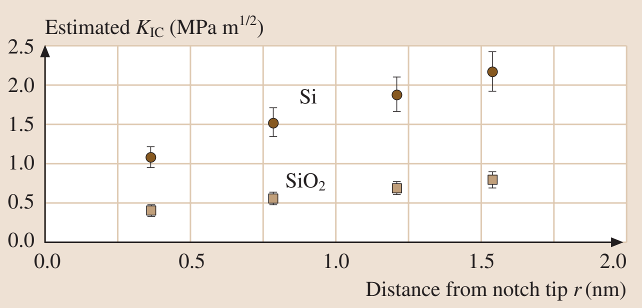

Fig. 25.14 Fracture toughness (K_{IC}) values for increasing values of r corresponding to distance between neighboring atoms in {111} planes of silicon (0.4 nm). Hence r values between 0.4 and 1.6 nm are chosen. The K_{IC} values thus estimated are comparable to values reported by others for both Si and SiO_2 [25.38]

Fracture Toughness

Estimates of fracture toughness calculated using (25.10) for Si and SiO_2 are shown in Fig. 25.14 [25.38]. The results show that the K_{IC} estimate for Si is about 1–2 MPa $\sqrt{m}$, whereas for SiO_2 the estimate is about 0.5–0.9 MPa $\sqrt{m}$. These values are comparable to values reported by others on larger specimens for Si [25.30–33] and SiO_2 [25.34]. The high values obtained for Si could be due to the fact that the scratches, despite being quite sharp, still have a finite radius of about 100 nm. The bulk value for silicon is about 0.9 MPa $\sqrt{m}$ (Table 25.2). Fracture toughness is considered a material property and is believed to be independent of specimen size. The values obtained in this study, given its limitations, appear to show that fracture toughness is comparable, if not a little higher on the nanoscale.

Fatigue Strength

Fatigue strength measurements of Si nanobeams have been carried out by *Sundararajan* and *Bhushan* [25.38] using an AFM and *Li* and *Bhushan* [25.39] using a nanoindenter. Various stress levels were applied to nanobeams by *Sundararajan* and *Bhushan* [25.38]. The minimum stress was 3.5 GPa for Si beams and 2.2 GPa for SiO_2 beams. The frequency of applied load was 4.2 Hz. In general, the fatigue life decreased with increasing mean stress as well as increasing stress amplitude. When the stress amplitude was less than 15% of the bending strength, the fatigue life was greater than 30,000 cycles for both Si and SiO_2. However, the mean stress had to be less than 30% of the bending strength for a life of greater than 30,000 for Si, whereas even at a mean stress of 43% of the bending strength, SiO_2 beams showed a life greater than 30,000. During fatigue, the beams broke under the loading point or at the ends when loaded at the center of the span. This was different from the quasi-static bending tests, where the beams broke at the ends almost every time. This could be due to the fact that the stress levels under the load and at the ends are not that different and fatigue crack propagation could occur at either location. Figure 25.15 shows a nanoscale *S*-*N* curve with bending stress (*S*) as a function of fatigue in cycles (*N*) with an apparent endurance life at lower stress. This study clearly demonstrates that fatigue properties of nanoscale specimens can be studied.

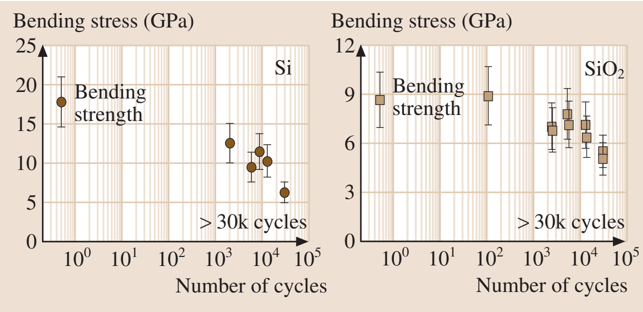

Fig. 25.15 Fatigue test data showing applied bending stress as a function of number of cycles. A single load-unload sequence is considered 1 cycle. The bending strength data points are, therefore, associated with $\frac{1}{2}$ cycle, since failure occurs upon loading [25.38]

SEM Observations of Fracture Surfaces

Figure 25.16 shows SEM images of the fracture surfaces of nanobeams broken during quasi-static bending, as well as fatigue [25.38]. In the quasi-static cases, the maximum tensile stresses occur on the top surface, so it is reasonable to assume that fracture initiated at or

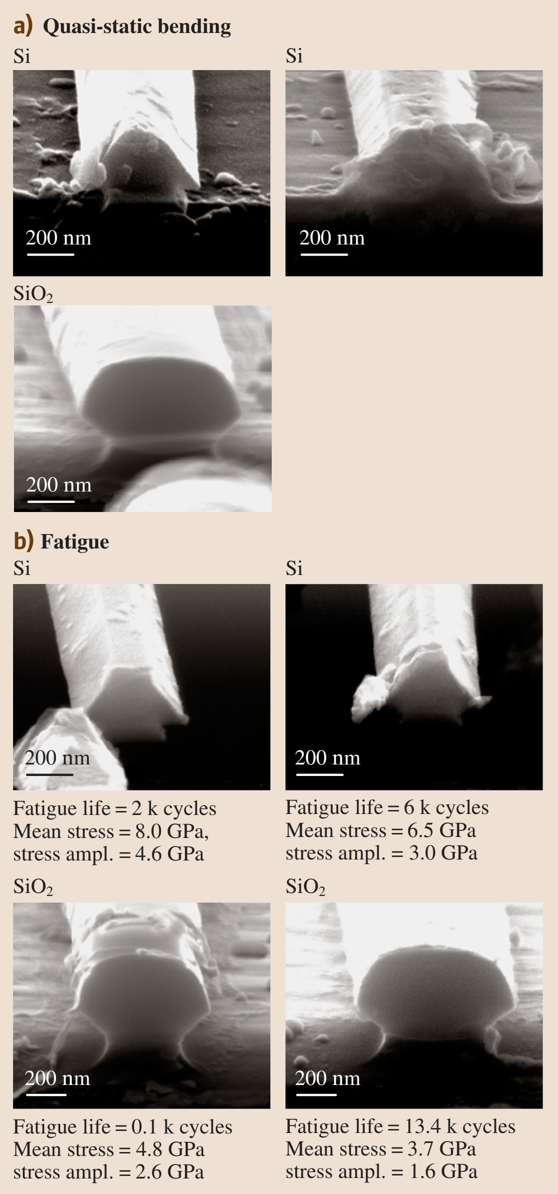
a) Quasi-static bending
Si
200 nm
Si
200 nm
SiO_2
200 nm
b) Fatigue
Si
200 nm
Fatigue life = 2 k cycles
Mean stress = 8.0 GPa,
stress ampl. = 4.6 GPa
Si
200 nm
Fatigue life = 6 k cycles
Mean stress = 6.5 GPa
stress ampl. = 3.0 GPa
SiO_2
200 nm
Fatigue life = 0.1 k cycles
Mean stress = 4.8 GPa
stress ampl. = 2.6 GPa
SiO_2
200 nm
Fatigue life = 13.4 k cycles
Mean stress = 3.7 GPa
stress ampl. = 1.6 GPa

Fig. 25.16a,b SEM micrographs of fracture surfaces of silicon and SiO_2 beams subjected to **(a)** quasi-static bending and **(b)** fatigue [25.38] near the top surface and propagated downward. The fracture surfaces of the beams suggest a cleavage type of fracture. Silicon beam surfaces show various ledges or facets, which is typical for crystalline brittle materials. Silicon usually fractures along the (111) plane, due to this plane having the lowest surface energy to overcome by a propagating crack. However, failure has also been known to occur along the (110) planes in microscale specimens, despite the higher energy required compared to the (111) planes [25.45]. The plane normal to the beam direction in these samples is the (110) plane, while (111) planes will be oriented at 35° from the (110) plane. The presence of facets and irregularities on the silicon surface in Fig. 25.16a suggest that it is a combination of these two types of fractures that has occurred. Since the stress levels are very high for these specimens, it is reasonable to assume that crack propagating forces will be high enough to result in (110) type failures.

In contrast, the silicon fracture surfaces under fatigue, shown in Fig. 25.16b, appear very smooth without facets or irregularities. This is suggestive of low energy fracture, i. e., of (111) type fracture. We do not see evidence of fatigue crack propagation in the form of steps or striations on the fracture surface. We believe that for the stress levels applied in these fatigue experiments, failure in silicon occurred via cleavage associated with "static fatigue" type of failures.

SiO_2 shows very smooth fracture surfaces for both quasi-static bending and fatigue. This is in contrast to the hackled surface one might expect for the brittle failure of an amorphous material on the macroscale. However, in larger scale fracture surfaces for such materials, the region near the crack initiation usually appears smooth or mirror-like. Since the fracture surface here is so small and very near the crack initiation site, it is not unreasonable to see such a smooth surface for SiO_2 on this scale. There appears to be no difference between the fracture surfaces obtained by quasi-static bending and fatigue for SiO_2.

Summary of Mechanical Properties Measured Using Quasi-Static Bending Tests

Table 25.2 summarizes the various properties measured via quasi-static bending in this study [25.38]. Also shown are bulk values of the parameters, along with values reported on larger scale specimens by other researchers. Elastic modulus and fracture toughness values appear to be comparable to bulk values and show no dependence on specimen size. However, bending strength shows a clear specimen size dependence with nanoscale numbers being twice as large as numbers reported for larger-scale specimens.

25.2.3 Bending Tests of Microbeams Using a Nanoindenter

Bending tests have been performed on Ni-P and Au microbeams [25.40]. The Ni-P cantilever microbeams were fabricated by focused ion beam machining technique. The dimensions were $10 \times 12 \times 50\,\mu m^3$. Notches with a depth of 3 μm and a tip radius of 0.25 μm were introduced in the microbeams to facilitate failure at a lower

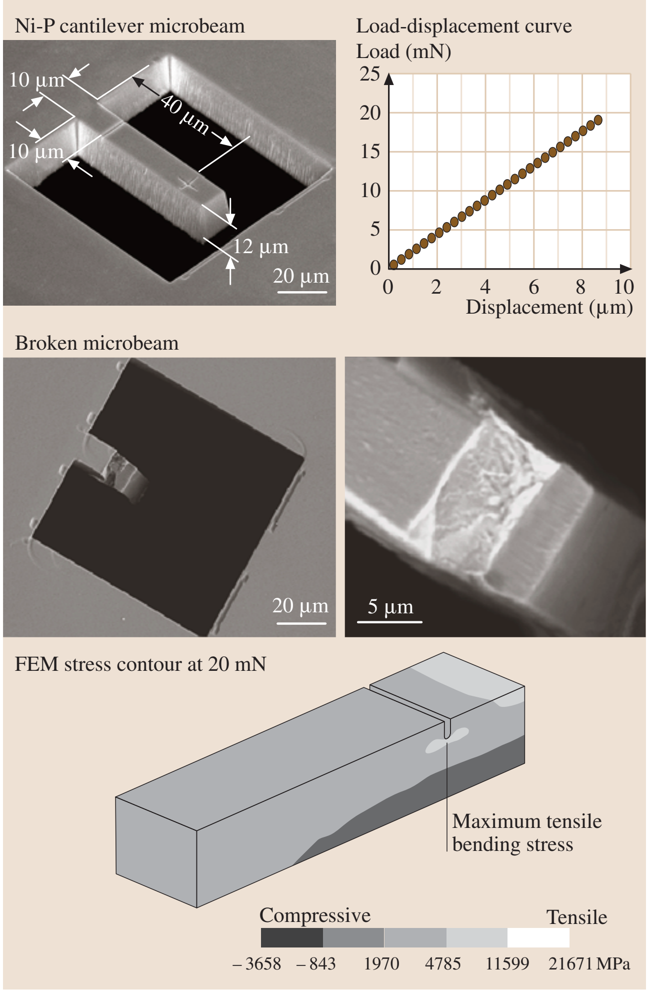

Fig. 25.17 SEM micrographs of the new and broken beams, load-displacement curve, and FEM stress contour for the notched Ni-P cantilever microbeam [25.40]

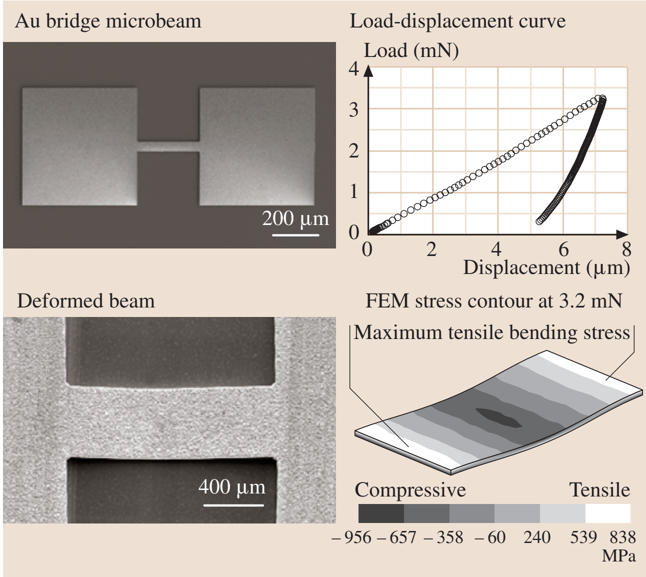

Fig. 25.18 SEM micrographs of the new and deformed beams, load-displacement curve, and FEM stress contour for the Au bridge microbeam [25.40]

load in the bending tests. The Au bridge microbeams were fabricated by electroplating technique.

Figure 25.17 shows the SEM images, load displacement curve, and FEM stress contour for the notched Ni-P cantilever microbeam that was bent to failure [25.40]. The distance between the loading position and the fixed end is 40 µm. The 3-µm-deep notch is 10 µm from the fixed end. The notched beam showed linear behavior followed by abrupt failure. The FEM stress contour shows that there is higher stress concentration at the notch tip. The maximum tensile stress σ_m at the notch tip can be analyzed using Griffith fracture theory as follows [25.53]:

$$\sigma_\mathrm{m} \approx 2\sigma_\mathrm{o}\left(\frac{c}{\rho}\right)^{1/2} , \tag{25.11}$$

where σ_o is the average applied tensile stress on the beam, c is the crack length, and ρ is the crack tip radius. Therefore, elastic-plastic deformation will first occur locally at the end of the notch tip, followed by abrupt fracture failure after the σ_m reaches the ultimate tensile strength of Ni-P, even though the rest of the beam is still in the elastic regime. The SEM image of the fracture surface shows that the fracture started right from the notch tip with plastic deformation characteristic. This indicates that although local plastic deformation occurred at the notch tip area, the whole beam failed catastrophically. The present study shows that FEM simulation can predict well the stress concentration and helps in understanding the failure mechanism of the notched beams.

Figure 25.18 shows the SEM images, load-displacement curve and FEM stress contour for the Au bridge microbeam that was deformed by the indenter [25.40]. The recession gap between the beam and substrate is about 7 µm, which is not large enough to break the beam at the load applied. From the load-displacement curve, we note that the beam experienced elastic-plastic deformation. The FEM stress contour shows that the maximum tensile stress is located at the fixed ends, whereas the minimum compressive stress is located around the center of the beam. The SEM image shows that the beam has been permanently deformed. No crack was found on the beam surface. The present study shows a possibility for mechanically forming the Au film into the shape as needed. This may help in designing/fabricating functionally complex smart micro/nanodevices that need conductors for power supply and input/output signals.

25.3 Finite Element Analysis of Nanostructures with Roughness and Scratches

Micro/nanostructures have some surface topography and local scratches dependent upon the manufacturing process. Surface roughness and local scratches may compromise the reliability of the devices and their effect needs to be studied. Finite element modeling is used to perform parametric analysis to study the effect of surface roughness and scratches in different well-defined forms on tensile stresses that are responsible for crack propagation [25.46, 47]. The analysis has been carried out on trapezoidal beams supported at the bottom whose data (on Si and SiO_2 nanobeams) have been presented earlier.

The finite element analysis has been carried out using the static analysis of ANSYS 5.7, which calculates the deflections and stresses produced by the applied loading. The type of element selected for the present study was the SOLID95, which allows the use of different shapes without much loss of accuracy. This element is 3-D with 20 nodes, each node having three degrees of freedom that imply translation in the x, y, and z directions. Each nanobeam cross section is divided into six elements along the width and the thickness and 40 elements along the length. SOLID95 has plasticity, creep, stress stiffening, large deflection, and large strain capabilities. The large displacement analysis is used for large loads. The mesh is kept finer near the asperities and the scratches in order to take into account variation in the bending stresses. The beam materials studied are made of single-crystal silicon (110) and SiO_2 films whose data have been presented earlier. Based on bending experiments presented earlier, the beam materials can be assumed to be linearly elastic isotropic materials. Young's modulus of elasticity (E) and Poisson's ratio (ν) for Si and SiO_2 are 169 GPa [25.63] and 0.28 [25.57], and 73 GPa [25.64] and 0.17 [25.64], respectively. A sample nanobeam of silicon was chosen for performing most of the analysis, as silicon is the most widely used MEMS/NEMS material. The cross section of the fabricated beams used in the experiment is trapezoidal and supported at the bottom, Fig. 25.9. The following dimensions are used: $w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, and $\ell = 6$ µm. In the boundary conditions, the displacements are constrained in all directions on the bottom surface for 1 µm from each end. A point load applied at the center of the beam is simulated with the load being applied at three closely located central nodes on the beam used. It has been observed from the experimental results that the Si nanobeam breaks at around 80 µN. Therefore, in this analysis, a nominal load of 70 µN is selected. At this load, deformations are large and a large displacement option is used.

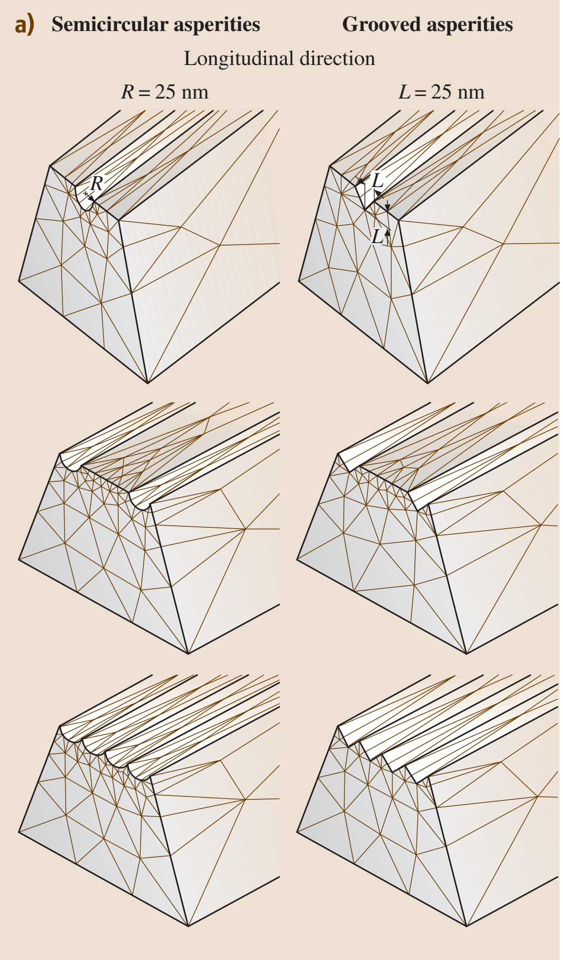

Fig. 25.19 **(a)** Plots showing the geometries of modeled roughness – semicircular and grooved asperities along the nanobeam length with defined geometrical parameters. **(b)** (*See on next page.*) Schematic showing semicircular asperities and scratches in the transverse direction, followed by the illustration of the mesh created on the beam with fine mesh near the asperities and scratches. Also shown are the semicircular asperities and scratches at different pitch values

To study the effect of surface roughness and scratches on the maximum bending stresses the following cases were studied. First, the semicircular and grooved asperities in the longitudinal direction with defined geometrical parameters are analyzed, Fig. 25.19a. Next, semicircular asperities and scratches placed along the transverse direction at a distance c from the end and separated by pitch p from each other are analyzed, Fig. 25.19b. Lastly, the beam material is assumed to be either purely elastic, elastic-plastic, or elastic-perfectly plastic. In the following section, we begin with the stress distribution in smooth nanobeams followed by the effect of surface roughness in the longitudinal and transverse directions and scratches in the transverse direction.

25.3.1 Stress Distribution in a Smooth Nanobeam

Figure 25.20 shows the stress and vertical displacement contours for a nanobeam supported at the bottom and loaded at the center [25.46, 47]. As expected, the maximum tensile stress occurs at the ends, while the maximum compressive stress occurs under the load at

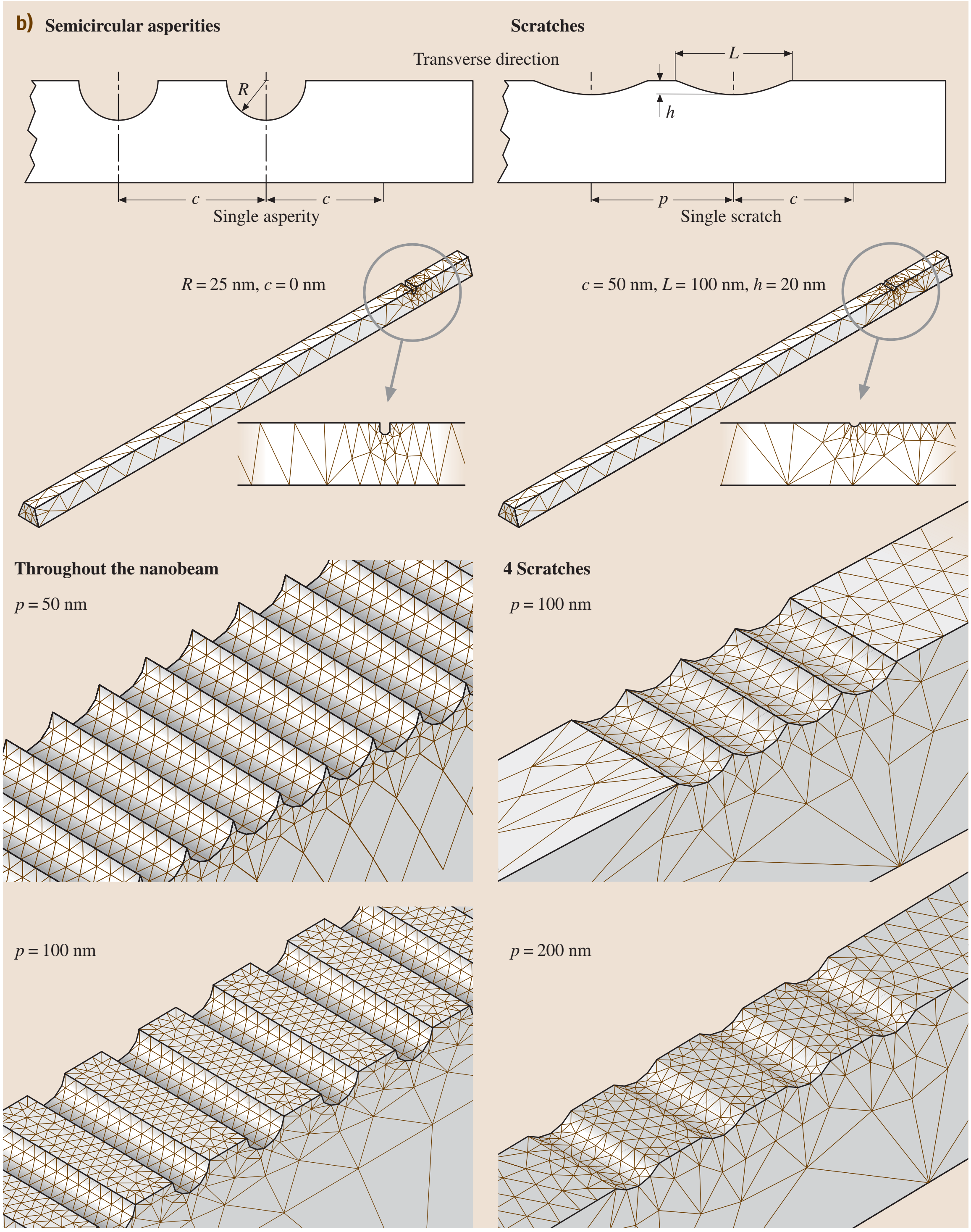
b)
Semicircular asperities
Scratches
Transverse direction
R
c
c
Single asperity
L
h
p
c
Single scratch
$R = 25$ nm, $c = 0$ nm
$c = 50$ nm, $L = 100$ nm, $h = 20$ nm
Throughout the nanobeam
$p = 50$ nm
4 Scratches
$p = 100$ nm
$p = 100$ nm
$p = 200$ nm

Fig. 25.19b

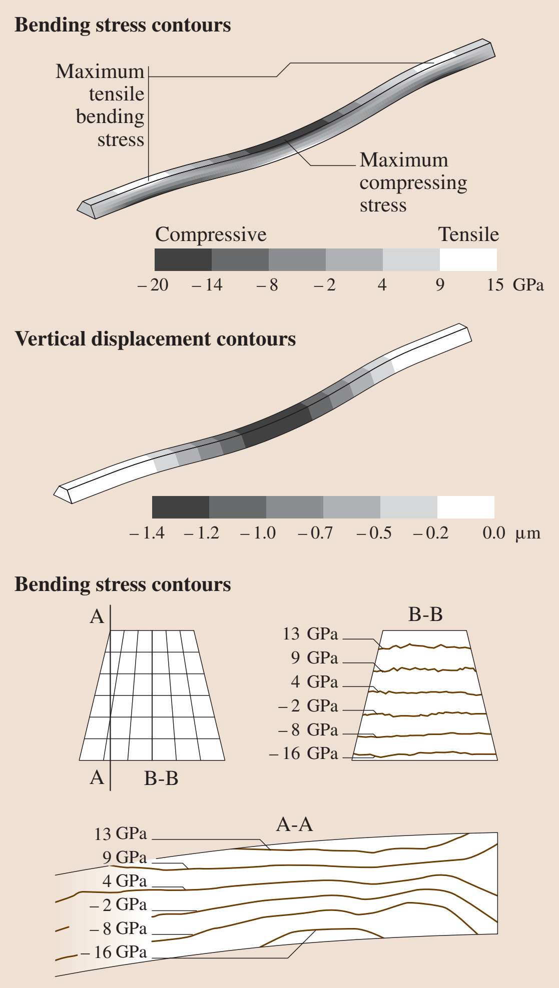

Fig. 25.20 Bending stress contours, vertical displacement contours, and bending stress contours after loading trapezoidal Si nanobeam ($w_1 = 200\,\text{nm}$, $w_2 = 370\,\text{nm}$, $t = 255\,\text{nm}$, $\ell = 6\,\mu\text{m}$, $E = 169\,\text{GPa}$, $\nu = 0.28$) at $70\,\mu\text{N}$ load [25.47]

the center. Stress contours obtained at a section of the beam from the front and side are also shown. In the beam cross section, the stresses remain constant at a given vertical distance from one side to another and change with a change in vertical location. This can be explained by the fact that the bending moment is constant at a particular cross section, so the stress is only dependent on the distance from the neutral axis. However, in cross section A-A the high tensile and compressive stresses are localized near the end of the beam at top and bottom, respectively, whereas the lower values are spread out away from the ends. High value of tensile stresses occurs near the ends because of high bending moment.

25.3.2 Effect of Roughness in the Longitudinal Direction

The roughness in the form of semicircular and grooved asperities in the longitudinal direction on the maximum bending stresses are analyzed [25.47]. The radius R and depth L are kept fixed at 25 nm, while the number of asperities is varied and their effect is observed on the maximum bending stresses. Figure 25.21 shows the variation of maximum bending stresses as a function of asperity shape and the number of asperities. The maximum bending stresses increase as the asperity number increases for both semicircular and grooved asperities. This can be attributed to the fact that as the asperity number increases, the moment of inertia decreases for that cross section. Also, the distance from the neutral axis increases because the neutral axis shifts downwards. Both these factors lead to the increase in the maximum bending stresses, and this effect is more pronounced in the case of semicircular asperity as it exhibits a higher value of maximum bending stress than that in grooved asperity. Figure 25.21 shows the stress contours obtained at a section of the beam from the front side for both cases when we have a single semicircular asperity and when four adjacent semicircular asperities are present. Trends are similar to those observed earlier for a smooth nanobeam (Fig. 25.20).

25.3.3 Effect of Roughness in the Transverse Direction and Scratches

We analyze semicircular asperities when placed along the transverse direction followed by the effect of scratches on the maximum bending stresses in varying numbers and pitch [25.47]. In the analysis of semicircular transverse asperities, three cases were considered, which included a single asperity and asperities throughout the nanobeam surface separated by pitch equal to 50 nm and 100 nm. In all of these cases, c value was kept equal to 0 nm. Figure 25.22 shows that the value of maximum tensile stress is 42 GPa, which is much larger than the maximum tensile stress value with no asperity of 16 GPa, or when the semicircular asperity is present in the longitudinal direction. It is also observed that the maximum tensile stress does not vary with the number of asperities or the pitch, while the maximum compres-

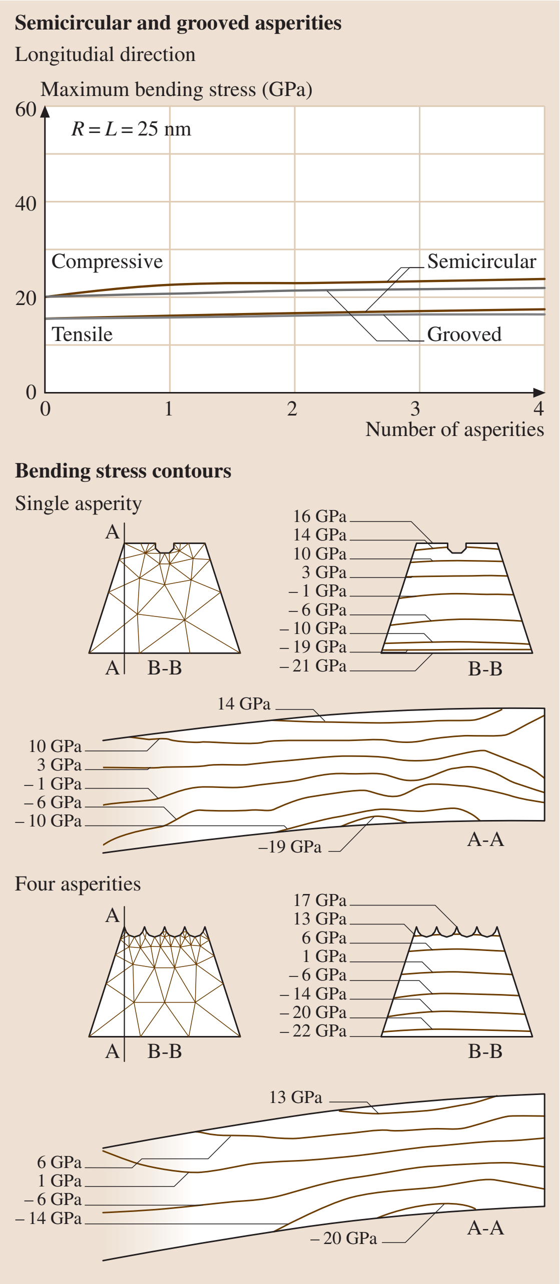

Fig. 25.21 Effect of longitudinal semicircular and grooved asperities in different numbers on maximum bending stresses after loading trapezoidal Si nanobeams ($w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, $\ell = 6$ µm, $E = 169$ GPa, $\nu = 0.28$, load = 70 µN). Bending stress contours obtained in the beam with semicircular single asperity and four adjacent asperities of $R = 25$ nm [25.47]

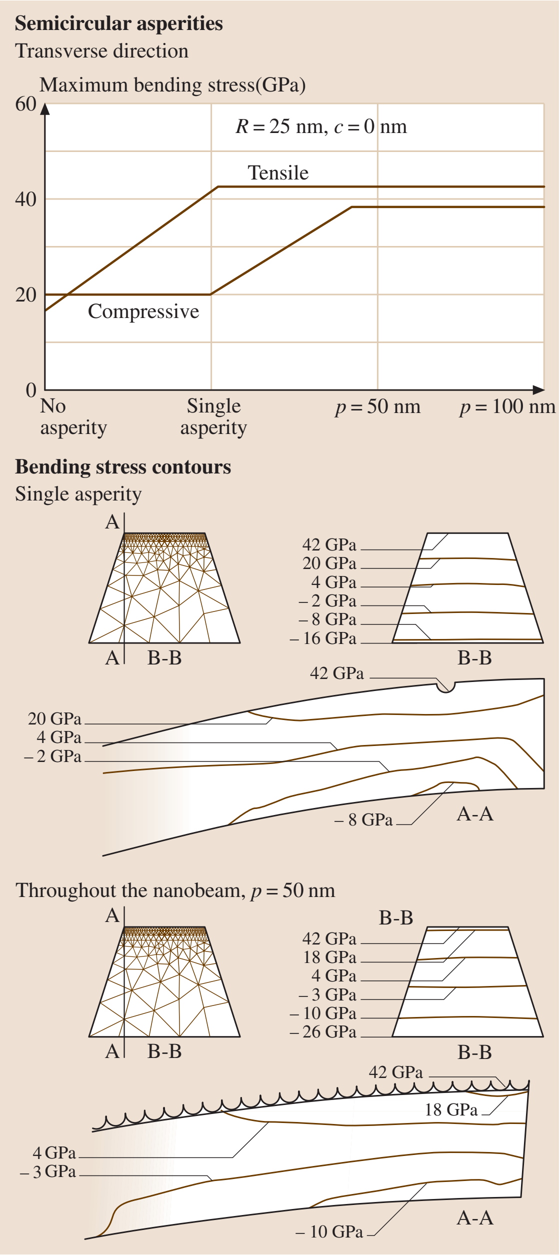

Fig. 25.22 Effect of transverse semicircular asperities located at different pitch values on the maximum bending stresses after loading trapezoidal Si nanobeams ($w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, $\ell = 6$ µm, $E = 169$ GPa, $\nu = 0.28$, load = 70 µN). Bending stress contours obtained in the beam with semicircular single asperity and semicircular asperities throughout the nanobeam surface at $p = 50$ nm [25.47]

sive stress does increase dramatically for the asperities present throughout the beam surface from its value when a single asperity is present. Maximum tensile stress occurs at the ends, and an increase in p does not add any asperities at the ends, whereas asperities are added in the central region where compressive stresses are maximum. The semicircular asperities present at the center cause the local perturbation in the stress distribution at the center of the asperity where load is being applied, leading to a high value of maximum compressive stress [25.65]. Figure 25.22 also shows the stress contours obtained at a section of the beam from the front and side for both cases when there is a single semicircular asperity and when asperities are present throughout the beam surface at a pitch equal to 50 nm. Trends are similar to those observed earlier for a smooth nanobeam (Fig. 25.20).

In the study pertaining to scratches, the number of scratches are varied along with the variation in the pitch. Furthermore, the load is applied at the center of the beam and at the center of the scratch near the end, as for all the cases. In all of these, c value was kept equal to 50 nm and L value was equal to 100 nm with h value being 20 nm. Figure 25.23 shows that the value of maximum tensile stress remains almost the same with the number of scratches for both types of loading – that is, when load is applied at the center of the beam and at the center of the scratch near the end. This is because the maximum tensile stress occurs at the beam ends no matter where the load gets applied. But the presence of scratch does increase the maximum tensile stress compared to its value for a smooth nanobeam, although the number of scratches no longer matter as the maximum tensile stress occurring at the nanobeam end is unaffected by the presence of more scratches beyond the first scratch in the direction toward the center. The value of the tensile stress is much lower when the load is applied at the center of scratch, and it can be explained as follows. The negative bending moment at the end near the applied load decreases with load offset after two-thirds of the length of the beam [25.66]. Since this negative bending moment is responsible for tensile stresses, their behavior with the load offset is the same as the negative bending moment. Also, the value of maximum compressive stress when load is applied at the center of the nanobeam remains almost the same as the center geometry is unchanged due to the number of scratches and, hence, the maximum compressive stress occurring below the load at the center is same. On the other hand, when the load is applied at the center of the scratch we observe that the maximum compressive stress in-

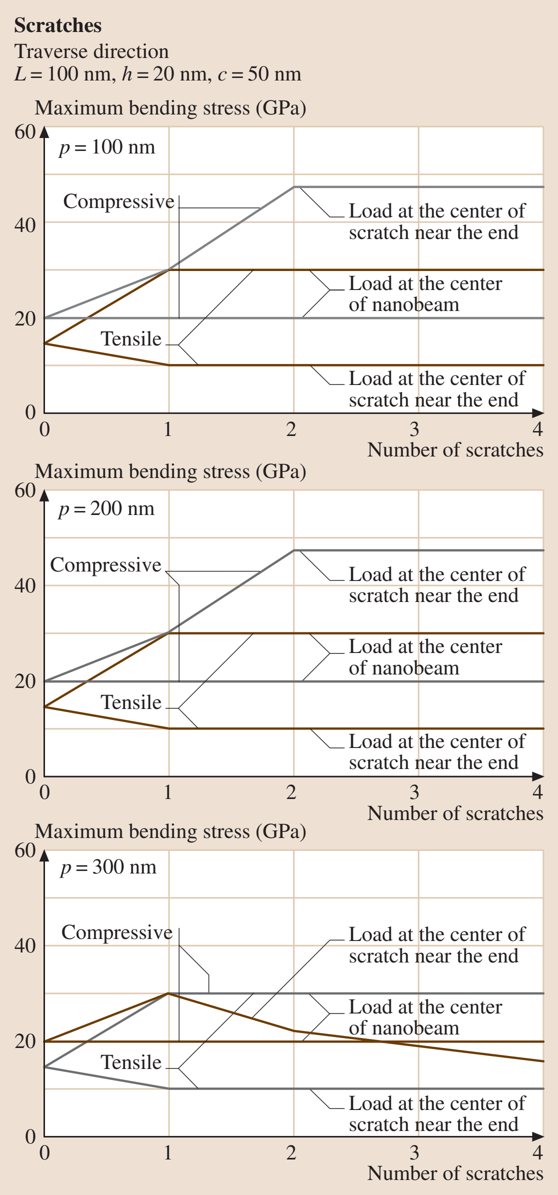

Fig. 25.23 Effect of number of scratches along with the variation in the pitch on the maximum bending stresses after loading trapezoidal Si nanobeams ($w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, $\ell = 6\,\mu$m, $E = 169$ GPa, $\nu = 0.28$, load = 70 μN). Also shown is the effect of load when applied at the center of the beam and at the center of the scratch near the end [25.47]

creases dramatically because the local perturbation in the stress distribution at the center of the scratch where load is being applied leads to a high value of maximum compressive stress [25.65]. It increases further with the number of scratches and then levels off. This can be attributed to the fact that when there is another scratch present close to the scratch near the end, the stress concentration is more, as the effect of local perturbation in the stress distribution is more significant. However, this effect is insignificant when more than two scratches are present.

Now we address the effect of pitch on the maximum compressive stress when the load is applied at the center of the scratch near the end. When the pitch is up to a value of 200 nm the maximum compressive stress increases with the number of scratches, as discussed earlier. On the other hand, when the pitch value goes beyond 225 nm this effect is reversed. This is because the presence of another scratch no longer affects the local perturbation in the stress distribution at the scratch near the end. Instead, more scratches at a fair distance distribute the maximum compressive stress at the scratch near the end and the stress starts going down. Such observations of maximum bending stresses can help in identifying the number of asperities and scratches allowed separated by an optimum distance from each other.

25.3.4 Effect on Stresses and Displacements for Materials That Are Elastic, Elastic-Plastic, or Elastic-Perfectly Plastic

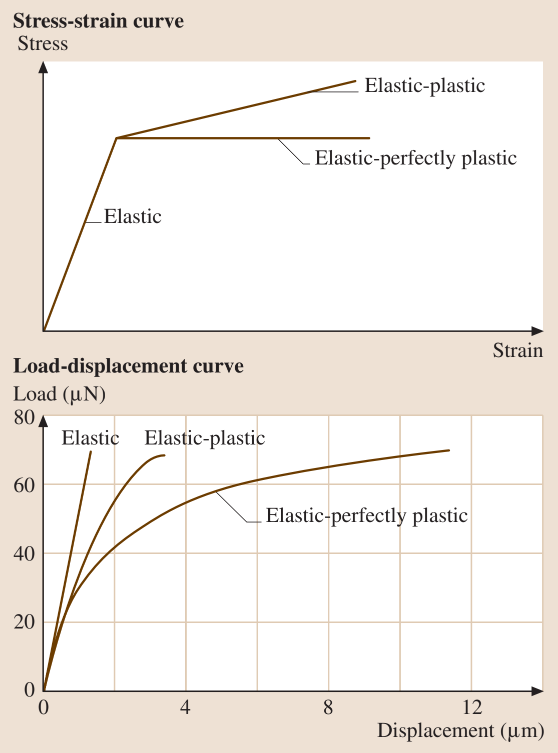

Fig. 25.24 Schematic representation of stress-strain curves and load-displacement curves for material when it is elastic, elastic-plastic, or elastic-perfectly plastic for a Si nanobeam ($w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, $\ell = 6$ μm, $E = 169$ GPa, tangent modulus in plastic range $= 0.5$ E, $\nu = 0.28$) [25.47]

This section deals with the beam modeled as elastic, elastic-plastic, and elastic-perfectly plastic to observe the variation in the stresses and displacements from an elastic model used so far [25.47]. Figure 25.24 shows the typical stress-strain curves for the three types of deformation regimes and their corresponding load-displacement curves obtained from the model of an Si nanobeam that are found to exhibit the same trends.

Table 25.3 shows the comparison of maximum von Mises stress and maximum displacements for

Table 25.3 Stresses and displacements for materials that are elastic, elastic-plastic, or elastic-perfectly plastic (load $= 70$ μN, $w_1 = 200$ nm, $w_2 = 370$ nm, $t = 255$ nm, $\ell = 6$ μm, $R = 25$ nm, $E = 169$ GPa, tangent modulus in plastic range $= 0.5E$, $\nu = 0.28$)

	Elastic		Elastic-plastic		Elastic-perfectly plastic	
	Smooth nanobeam	**Single semicircular longitudinal asperity**	**Smooth nanobeam**	**Single semicircular longitudinal asperity**	**Smooth nanobeam**	**Single semicircular longitudinal asperity**
Max. von Mises stress (GPa)	18.2	19.3	13.5	15.2	7.8	9.1
Max. displacement (μm)	1.34	1.40	3.35	3.65	11.5	12.3

both smooth nanobeam and nanobeam with a defined roughness, which is single semicircular longitudinal asperity of R value equal to 25 nm for the three different models. It is observed that the maximum value of stress is obtained at a given load for elastic material, whereas the displacement is maximum for elastic-perfectly plastic material. Also, the pattern that the maximum bending stress value increases for a rough nanobeam still holds true in the other models as well.

25.4 Closure

Mechanical properties of nanostructures are necessary in designing realistic MEMS/NEMS devices. Most mechanical properties are scale-dependent. Micro/nanomechanical properties, hardness, elastic modulus, and scratch resistance of bulk materials of undoped single-crystal silicon (Si) and thin films of undoped polysilicon, SiO_2, SiC, Ni-P, and Au are presented. It is found that the SiC film exhibits higher hardness, elastic modulus, and scratch resistance as compared to other materials.

Bending tests have been performed on the Si and SiO_2 nanobeams and Ni-P and Au microbeams using an AFM and a depth-sensing nanoindenter, respectively. The bending tests were used to evaluate elastic modulus, bending strength (fracture stress), fracture toughness (K_{IC}), and fatigue strength of the beam materials. The Si and SiO_2 nanobeams exhibited elastic linear response with sudden brittle fracture. The notched Ni-P beam showed linear deformation behavior followed by abrupt failure. The Au beam showed elastic-plastic deformation behavior. Elastic modulus values of 182±11 GPa for Si(110) and 85±3 GPa for SiO_2 were obtained, which are comparable to bulk values. Bending strength values of 18±3 GPa for Si and 7.6±2 GPa for SiO_2 were obtained, which are twice as large as values reported on larger scale specimens. This indicates that bending strength shows a specimen size dependence. Fracture toughness value estimates obtained were $1.67 \pm 0.4\,\mathrm{MPa}\sqrt{\mathrm{m}}$ for Si and $0.60 \pm 0.2\,\mathrm{MPa}\sqrt{\mathrm{m}}$ for SiO_2, which are also comparable to values obtained on larger specimens. At stress amplitudes less than 15% of their bending strength and at mean stresses of less than 30% of the bending strength, Si and SiO_2 displayed an apparent endurance life of greater than 30,000 cycles. SEM observations of the fracture surfaces revealed a cleavage type of fracture for both materials when subjected to bending, as well as fatigue. The AFM and nanoindenters used in this study can be satisfactorily used to evaluate the mechanical properties of micro/nanoscale structures for use in MEMS/NEMS.

FEM simulations are used to predict the stress and deformation in nanostructures. The FEM has been used to analyze the effect of the type of surface roughness and scratches on stresses and deformation of nanostructures. We find that roughness affects the maximum bending stresses. The maximum bending stresses increase as the asperity number increases for both semicircular and grooved asperities in the longitudinal direction. When the semicircular asperity is present in the transverse direction the maximum tensile stress is much larger than the maximum tensile stress value with no asperity or when the semicircular asperity is present in the longitudinal direction. This observation suggests that the asperity in the transverse direction is more detrimental. The presence of scratches increases the maximum tensile stress. The maximum tensile stress remains almost the same with the number of scratches for two types of loading, that is, when load is applied at the center of the beam or at the center of the scratch near the end, although the value of the tensile stress is much lower when the load is applied at the center of the scratch. This means that the load applied at the ends is less damaging. This analysis shows that FEM simulations can be useful to designers to develop the most suitable geometry for nanostructures.

References

25.1 R. S. Muller, R. T. Howe, S. D. Senturia, R. L. Smith, R. M. White: *Microsensors* (IEEE, New York 1990)

25.2 I. Fujimasa: *Micromachines: A New Era in Mechanical Engineering* (Oxford Univ. Press, Oxford 1996)

25.3 W. S. Trimmer (Ed.): *Micromachines and MEMS, Classic and Seminal Papers to 1990* (IEEE, New York 1997)

25.4 B. Bhushan: *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)

25.5 G. T. A. Kovacs: *Micromachined Transducers Sourcebook* (WCB McGraw-Hill, Boston 1998)

25.6 S. D. Senturia: *Microsystem Design* (Kluwer, Boston 2001)

25.7 M. Gad-el-Hak: *The MEMS Handbook* (CRC, Boca Raton 2002)

25.8 T. R. Hsu: *MEMS and Microsystems* (McGraw-Hill, Boston 2002)

25.9 M. Madou: *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd edn. (CRC, Boca Raton 2002)

25.10 K. E. Drexler: *Nanosystems: Molecular Machinery, Manufacturing and Computation* (Wiley, New York 1992)

25.11 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

25.12 G. Timp (Ed.): *Nanotechnology* (Springer, New York 1999)

25.13 E. A. Rietman: *Molecular Engineering of Nanosystems* (Springer, New York 2001)

25.14 H. S. Nalwa (Ed.): *Nanostructures Materials and Nanotechnology* (Academic, San Diego 2002)

25.15 W. A. Goddard, D. W. Brenner, S. E. Lyshevski, G. J. Iafrate: *Handbook of Nanoscience, Engineering, and Technology* (CRC, Boca Raton 2003)

25.16 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

25.17 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

25.18 B. Bhushan: Macro- and microtribology of MEMS materials. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 1515–1548

25.19 S. Johansson, J. A. Schweitz, L. Tenerz, J. Tiren: Fracture testing of silicon microelements in-situ in a scanning electron microscope, J. Appl. Phys. **63** (1988) 4799–4803

25.20 F. Ericson, J. A. Schweitz: Micromechanical fracture strength of silicon, J. Appl. Phys. **68** (1990) 5840–5844

25.21 E. Obermeier: Mechanical and thermophysical properties of thin film materials for MEMS: Techniques and devices, Micromechan. Struct. Mater. Res. Symp. Proc. **444** (1996) 39–57

25.22 C. J. Wilson, A. Ormeggi, M. Narbutovskih: Fracture testing of silicon microcantilever beams, J. Appl. Phys. **79** (1995) 2386–2393

25.23 W. N. Sharpe, Jr., B. Yuan, R. L. Edwards: A new technique for measuring the mechanical properties of thin films, J. Microelectromech. Syst. **6** (1997) 193–199

25.24 K. Sato, T. Yoshioka, T. Anso, M. Shikida, T. Kawabata: Tensile testing of silicon film having different crystallographic orientations carried out on a silicon chip, Sens. Actuators A **70** (1998) 148–152

25.25 S. Greek, F. Ericson, S. Johansson, M. Furtsch, A. Rump: Mechanical characterization of thick polysilicon films: Young's modulus and fracture strength evaluated with microstructures, J. Micromech. Microeng. **9** (1999) 245–251

25.26 D. A. LaVan, T. E. Buchheit: Strength of polysilicon for MEMS devices, Proc. SPIE **3880** (1999) 40–44

25.27 E. Mazza, J. Dual: Mechanical behavior of a μm-sized single crystal silicon structure with sharp notches, J. Mech. Phys. Solids **47** (1999) 1795–1821

25.28 T. Yi, C. J. Kim: Measurement of mechanical properties for MEMS materials, Meas. Sci. Technol. **10** (1999) 706–716

25.29 H. Kahn, M. A. Huff, A. H. Heuer: Heating effects on the Young's modulus of films sputtered onto micromachined resonators, Microelectromech. Struct. Mater. Res. Symp. Proc. **518** (1998) 33–38

25.30 S. Johansson, F. Ericson, J. A. Schweitz: Influence of surface-coatings on elasticity, residual-stresses, and fracture properties of silicon microelements, J. Appl. Phys. **65** (1989) 122–128

25.31 R. Ballarini, R. L. Mullen, Y. Yin, H. Kahn, S. Stemmer, A. H. Heuer: The fracture toughness of polysilicon microdevices: A first report, J. Mater. Res. **12** (1997) 915–922

25.32 H. Kahn, R. Ballarini, R. L. Mullen, A. H. Heuer: Electrostatically actuated failure of microfabricated polysilicon fracture mechanics specimens, Proc. R. Soc. London A **455** (1999) 3807–3823

25.33 A. M. Fitzgerald, R. H. Dauskardt, T. W. Kenny: Fracture toughness and crack growth phenomena of plasma-etched single crystal silicon, Sens. Actuators A **83** (2000) 194–199

25.34 T. Tsuchiya, A. Inoue, J. Sakata: Tensile testing of insulating thin films: Humidity effect on tensile strength of SiO_2 films, Sens. Actuators A **82** (2000) 286–290

25.35 J. A. Connally, S. B. Brown: Micromechanical fatigue testing, Exp. Mech. **33** (1993) 81–90

25.36 K. Komai, K. Minoshima, S. Inoue: Fracture and fatigue behavior of single-crystal silicon microelements and nanoscopic AFM damage evaluation, Microsyst. Technol. **5** (1998) 30–37

25.37 T. Namazu, Y. Isono, T. Tanaka: Evaluation of size effect on mechanical properties of single-crystal silicon by nanoscale bending test using AFM, J. Microelectromech. Syst. **9** (2000) 450–459

25.38 S. Sundararajan, B. Bhushan: Development of AFM-based techniques to measure mechanical properties of nanoscale structures, Sens. Actuators A **101** (2002) 338–351

25.39 X. Li, B. Bhushan: Fatigue studies of nanoscale structures for MEMS/NEMS applications using nanoindentation techniques, Surf. Coat. Technol. **163–164** (2003) 521–526

25.40 X. Li, B. Bhushan, K. Takashima, C.W. Baek, Y.K. Kim: Mechanical characterization of micro/nanoscale structures for MEMS/NEMS applications using nanoindentation techniques, Ultramicroscopy **97** (2003) 481–494

25.41 T. Hsu, N. Sun: Residual stresses/strains analysis of MEMS, Proc. of the Int. Conf. on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators, ed. by M. Laudon, B. Romanowicz (Computational Publications, Cambridge 1998) 82–87

25.42 A. Kolpekwar, C. Kellen, R.D. (Shawn) Blanton: Fault model generation for MEMS, Proc. of the Int. Conf. on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators, ed. by M. Laudon, B. Romanowicz (Computational Publications, Cambridge 1998) 111–116

25.43 H.A. Rueda, M.E. Law: Modeling of strain in boron-doped silicon cantilevers, Proc. of the Int. Conf. on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators, ed. by M. Laudon, B. Romanowicz (Computational Publications, Cambridge 1998) 94–99

25.44 M. Heinzelmann, M. Petzold: FEM analysis of microbeam bending experiments using ultra-micro indentation, Comput. Mater. Sci. **3** (1994) 169–176

25.45 C.J. Wilson, P.A. Beck: Fracture testing of bulk silicon microcantilever beams subjected to a side load, J. Microelectromech. Syst. **5** (1996) 142–150

25.46 B. Bhushan, G.B. Agrawal: Stress analysis of nanostructures using a finite element method, Nanotechnology **13** (2002) 515–523

25.47 B. Bhushan, G.B. Agrawal: Finite element analysis of nanostructures with roughness and scratches, Ultramicroscopy **97** (2003) 495–501

25.48 K.E. Petersen: Silicon as a mechanical material, Proc. IEEE **70** (1982) 420–457

25.49 B. Bhushan, S. Sundararajan, X. Li, C.A. Zorman, M. Mehregany: Micro/nanotribological studies of single-crystal silicon and polysilicon and SiC films for use in MEMS devices. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 407–430

25.50 S. Sundararajan, B. Bhushan: Micro/nanotribological studies of polysilicon and SiC films for MEMS applications, Wear **217** (1998) 251–261

25.51 X. Li, B. Bhushan: Micro/nanomechanical characterization of ceramic films for microdevices, Thin Solid Films **340** (1999) 210–217

25.52 B. Bhushan, X. Li: Nanomechanical characterization of solid surfaces and thin films, Int. Mater. Rev. **48** (2003) 125–164

25.53 B.R. Lawn, A.G. Evans, D.B. Marshall: Elastic/plastic indentation damage in ceramics: The median/radial system, J. Am. Ceram. Soc. **63** (1980) 574

25.54 S. Sundararajan, B. Bhushan, T. Namazu, Y. Isono: Mechanical property measurements of nanoscale structures using an atomic force microscope, Ultramicroscopy **91** (2002) 111–118

25.55 R.J. Roark: *Formulas for Stress and Strain*, 6th edn. (McGraw-Hill, New York 1989)

25.56 R.W. Hertzberg: *Deformation and Fracture Mechanics of Engineering Materials*, 3rd edn. (Wiley, New York 1989) pp. 277–278

25.57 Anonymous: Properties of silicon, EMIS Datareviews Series No. 4 (1988)

25.58 C.T.C. Nguyen, R.T. Howe: An integrated CMOS micromechanical resonator high-Q oscillator, IEEE J. Solid State Circ. **34** (1999) 440–455

25.59 L.J. Hornbeck: A digital light processing$^{\text{TM}}$ update – status and future applications, Proc. SPIE **3634** (1999) 158–170

25.60 M. Tanaka: Fracture toughness and crack morphology in indentation fracture of brittle materials, J. Mater. Sci. **31** (1996) 749

25.61 T. Tsuchiya, O. Tabata, J. Sakata, Y. Taga: Specimen size effect on tensile strength of surface-micromachined polycrystalline silicon thin films, J. Microelectromech. Syst. **7** (1998) 106–113

25.62 T. Yi, L. Li, C.J. Kim: Microscale material testing of single crystalline silicon: Process effects on surface morphology and tensile strength, Sens. Actuators A **83** (2000) 172–178

25.63 B. Bhushan, S. Venkatesan: Mechanical and tribological properties of silicon for micromechanical applications: A review, Adv. Info. Storage Syst. **5** (1993) 211–239

25.64 B. Bhushan, B.K. Gupta: *Handbook of Tribology: Materials, Coatings, and Surface Treatments* (McGraw-Hill, New York 1991)

25.65 S.P. Timoshenko, J.N. Goodier: *Theory of Elasticity*, 3rd edn. (McGraw-Hill, New York 1970)

25.66 J.E. Shigley, L.D. Mitchell: *Mechanical Engineering Design*, 4th edn. (McGraw-Hill, New York 1993)

Part D

Molecular

Part D Molecularly Thick Films for Lubrication

26 **Nanotribology of Ultrathin and Hard Amorphous Carbon Films**
Bharat Bhushan, Columbus, USA

27 **Self-Assembled Monolayers for Controlling Adhesion, Friction and Wear**
Bharat Bhushan, Columbus, USA
Huiwen Liu, Columbus, USA

28 **Nanoscale Boundary Lubrication Studies**
Bharat Bhushan, Columbus, USA
Huiwen Liu, Columbus, USA

29 **Kinetics and Energetics in Nanolubrication**
Rene M. Overney, Seattle, USA
George W. Tyndall, San Jose, USA
Jane Frommer, San Jose, USA

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

26. Nanotribology of Ultrathin and Hard Amorphous Carbon Films

Diamond material and its smooth coatings are used for very low wear and relatively low friction. Major limitations of the true diamond coatings are that they need to be deposited at high temperatures, can only be deposited on selected substrates, and require surface finishing. Hard amorphous carbon, commonly known as diamond-like carbon or DLC coatings, exhibit mechanical, thermal, and optical properties close to that of diamond. These can be deposited with a large range of thicknesses by using a variety of deposition processes on a variety of substrates at or near room temperature. The coatings reproduce substrate topography, avoiding the need of post-finishing. Friction and wear properties of some DLC coatings can be very attractive for tribological applications. The largest industrial application of these coatings is in magnetic storage devices.

The prevailing atomic arrangement in the DLC coatings is amorphous or quasi-amorphous with small diamond, graphite, and other unidentifiable micro- or nanocrystallites. Most DLC coatings, except those produced by filtered cathodic arc, contain from a few to about 50 at% hydrogen. Sometimes hydrogen is deliberately incorporated in the sputtered and ion plated coatings to tailor their properties.

EELS and Raman spectroscopies can be successfully used for chemical characterization of amorphous carbon coatings. The prevailing atomic arrangement in the DLC coatings is amorphous or quasi-amorphous with small diamond (sp^3), graphite (sp^2) and other unidentifiable micro- or nanocrystallites. Most DLC coatings except those produced by filtered cathodic arc contain from a few to about 50 at% hydrogen. Sometimes hydrogen is deliberately incorporated in the sputtered and ion plated coatings to tailor their properties.

Amorphous carbon coatings deposited by various techniques exhibit different mechanical and tribological properties. The nanoindenter can be successfully used for measurement of hardness, elastic modulus, fracture toughness, and fatigue life. Microscratch and microwear experiments can be performed using either a nanoindenter or an AFM. Thin coatings deposited by filtered cathodic arc, ion beam, and ECR-CVD hold a promise for tribological applications. Coatings as thin as 5 nm or even thinner provide wear protection. Microscratch, microwear, and accelerated wear testing, if simulated properly, can be successfully used to screen coating candidates for industrial applications. In the examples shown in this chapter, trends observed in the microscratch, microwear, and accelerated macrofriction wear tests are similar to that found in functional tests.

In this chapter, the state-of-the-art of recent developments in the chemical, mechanical, and tribological characterization of ultrathin amorphous carbon coatings is presented.

26.1 Description of Commonly Used Deposition Techniques 795
26.1.1 Filtered Cathodic Arc Deposition Technique 798
26.1.2 Ion Beam Deposition Technique 798
26.1.3 Electron Cyclotron Resonance Chemical Vapor Deposition Technique 799
26.1.4 Sputtering Deposition Technique ... 799
26.1.5 Plasma-Enhanced Chemical Vapor Deposition Technique 799

26.2 Chemical Characterization and Effect of Deposition Conditions on Chemical Characteristics and Physical Properties ... 800
26.2.1 EELS and Raman Spectroscopy 800
26.2.2 Hydrogen Concentrations 804
26.2.3 Physical Properties 804
26.2.4 Summary 805

26.3 Micromechanical and Tribological Characterizations of Coatings Deposited by Various Techniques 805
26.3.1 Micromechanical Characterization . 805
26.3.2 Microscratch and Microwear Studies 813
26.3.3 Macroscale Tribological Characterization 822
26.3.4 Coating Continuity Analysis 826

References 827

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Carbon exists in both crystalline and amorphous forms and exhibits both metallic and nonmetallic characteristics [26.1–3]. Crystalline carbon includes graphite, diamond, and a family of fullerenes, Fig. 26.1. The graphite and diamond are infinite periodic network solids with a planar structure, whereas the fullerenes are a molecular form of pure carbon with a finite network with a nonplanar structure. Graphite has a hexagonal, layered structure with weak interlayer bonding forces and exhibits excellent lubrication properties. The graphite crystal may be visualized as infinite parallel layers of hexagons stacked 0.34 nm apart with a 0.1415-nm interatomic distance between the carbon atoms in the basal plane. The atoms lying in the basal planes are trigonally coordinated and closely packed with strong σ (covalent) bonds to its three carbon neighbors using the hybrid sp^2 orbitals. The fourth electron lies in a p_z orbital lying normal to the σ bonding plane and forms a weak π bond by overlapping side to side with a p_z orbital of an adjacent atom to which carbon is attached by a σ bond. The layers (basal planes) themselves are relatively far apart and the forces that bond them are weak van der Waals forces. These layers can align themselves parallel to the direction of the relative motion and slide over one another with relative ease, thus providing low friction. Strong interatomic bonding and packing in each layer is thought to help reduce wear. The operating environment has a significant influence on lubrication, i. e., low friction and low wear, properties of graphite. It lubricates better in a humid environment than a dry one, resulting from adsorption of water vapor and other gases from the environment, which further weakens the interlayer bonding forces and results in easy shear and transfer of the crystallite platelets to the mating surface. Thus, transfer plays an important role in controlling friction and wear. Graphite oxidizes at high operating temperatures and can be used up to about 430 °C.

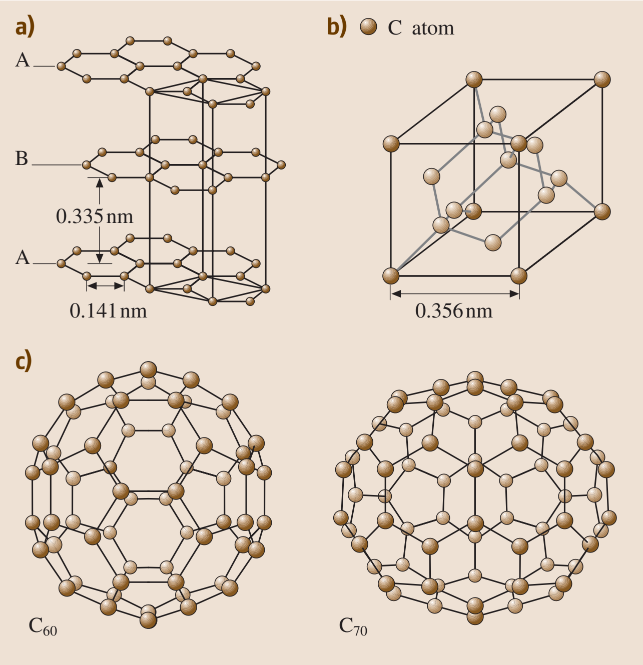

Fig. 26.1a–c The structure of three known forms of crystalline carbon **(a)** hexagonal structure of graphite, **(b)** modified face-centered cubic (fcc) structure, two interpenetrating fcc lattices displaced by one-quarter of the cube diagonal, of diamond (each atom is bonded to four others that form the corners of the pyramidal structure called tetrahedron), and **(c)** the structure of the two most common forms of fullerenes-soccer ball C_{60} and rugby ball C_{70} molecules

One of the fullerene molecules is C_{60}, commonly known as Buckyball. Since the C_{60} molecules are very stable and do not require additional atoms to satisfy chemical bonding requirements, they are expected to have low adhesion to the mating surface and low surface energy. Since C_{60} molecules with a perfect spherical symmetry are weakly bonded to other molecules, C_{60} clusters get detached readily, similar to other layered lattice structures, and either get transferred to the mating surface by mechanical compaction, or are present as loose wear particles that may roll like tiny ball bearings in a sliding contact, resulting in low friction and wear. The wear particles are expected to be harder than as-deposited C_{60} molecules, because of their phase transformation at high asperity contact pressures present in a sliding interface. The low surface energy, spherical shape of C_{60} molecules, weak intermolecular bonding, and high load bearing capacity offer potential for various mechanical and tribological applications. The sublimed C_{60} coatings and fullerene particles as an additive to mineral oils and greases have been reported to be good solid lubricants comparable to graphite and MoS_2 [26.4–6].

Diamond crystallizes in the modified face centered cubic (*fcc*) structure with an interactomic distance of 0.154 nm. The diamond cubic lattice consists of two interpenetrating fcc lattices displaced by one-quarter of the cube diagonal. Each carbon atom is tetrahedrally coordinated, making strong σ (covalent) bonds to its four carbon neighbors using the hybrid sp^3 atomic orbitals, which accounts for its highest hardness (80–104 GPa) and thermal conductivity (900–2,100 W/mK, on the order of five times that of copper) of any known solid, and a high electrical resistivity, optical transmission, and a large optical band gap. It is relatively chemically inert, and it exhibits poor adhesion with other solids with consequent low friction and wear. Its high thermal conductivity allows dissipation of frictional heat during sliding and protects the interface, and the dangling carbon bonds on the surface react with the environment to form hydrocarbons that act as good lubrication films. These are some of the reasons for low friction and wear of the diamond. Diamond and its coatings find many industrial applications: tribological applications (low friction and wear), optical applications (exceptional optical transmission, high abrasion resistance), and thermal management or heat sink applications (high thermal conductivity). The diamond can be used to high temperatures, and it starts to graphitize at about 1,000 °C in ambient air and at about 1,400 °C in vacuum. Diamond is an attractive material for cutting tools, as an abrasive for grinding wheels and lapping compounds, and other extreme wear applications.

The natural diamond, particularly in large sizes, is very expensive and its coatings, a low cost alternative, are attractive. The true diamond coatings are deposited by chemical vapor deposition (CVD) processes at high substrate temperatures (on the order of 800 °C). They adhere best on a silicon substrate and require an interlayer for other substrates. A major roadblock to the widespread use of true diamond films in tribological, optical, and thermal management applications is the surface roughness. Growth of the diamond phase on a non-diamond substrate is initiated by nucleation either at randomly seeded sites or at thermally favored sites, due to statistical thermal fluctuation at the substrate surface. Based on growth temperature and pressure conditions, favored crystal orientations dominate the competitive growth process. As a result, the grown films are polycrystalline in nature with relatively large grain size (> 1 μm) and terminate in very rough surfaces with RMS roughnesses ranging from a few tenths of a micron to tens of microns. Techniques for polishing these films have been developed. It has been reported that the laser polished films exhibit friction and wear properties almost comparable to that of bulk polished diamond [26.7, 8].

Amorphous carbon has no long-range order, and the short-range order of carbon atoms can have one or more of three bonding configurations: – sp^3 (diamond), sp^2 (graphite), or sp^1 (with two electrons forming strong

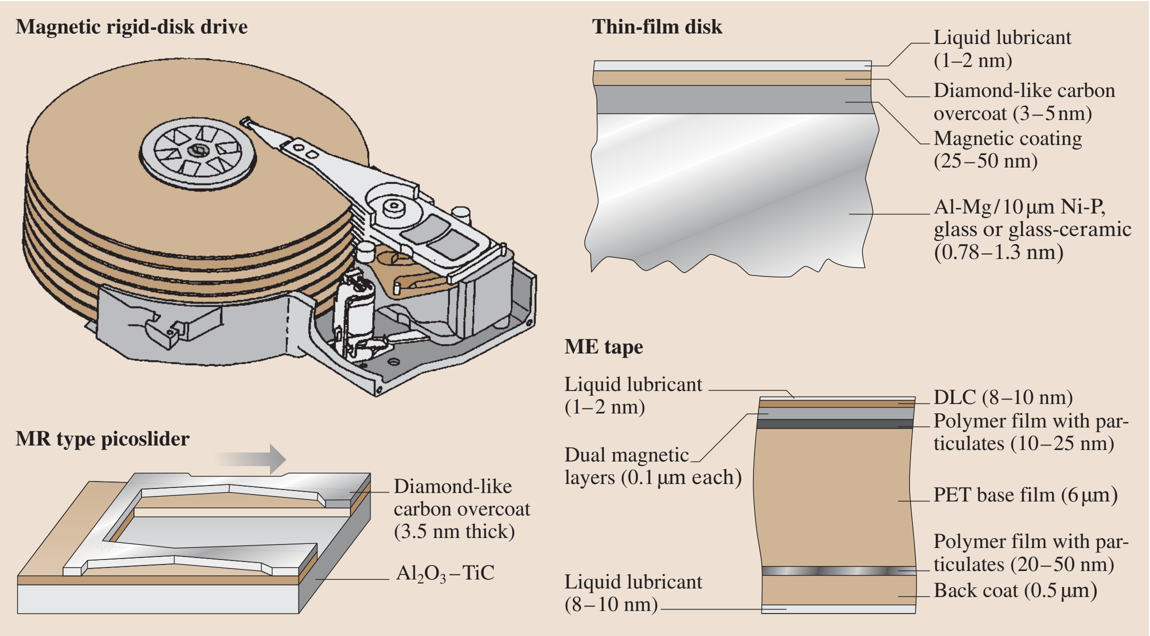
Magnetic rigid-disk drive
Thin-film disk
Liquid lubricant (1–2 nm)
Diamond-like carbon overcoat (3–5 nm)
Magnetic coating (25–50 nm)
Al-Mg/10 μm Ni-P, glass or glass-ceramic (0.78–1.3 nm)
ME tape
Liquid lubricant (1–2 nm)
Dual magnetic layers (0.1 μm each)
Liquid lubricant (8–10 nm)
DLC (8–10 nm)
Polymer film with particulates (10–25 nm)
PET base film (6 μm)
Polymer film with particulates (20–50 nm)
Back coat (0.5 μm)
MR type picoslider
Diamond-like carbon overcoat (3.5 nm thick)
Al_2O_3–TiC

Fig. 26.2 Schematic of a magnetic rigid-disk drive and MR type picoslider and cross-sectional schematics of a magnetic thin-film rigid disk and a metal evaporated (ME) tape

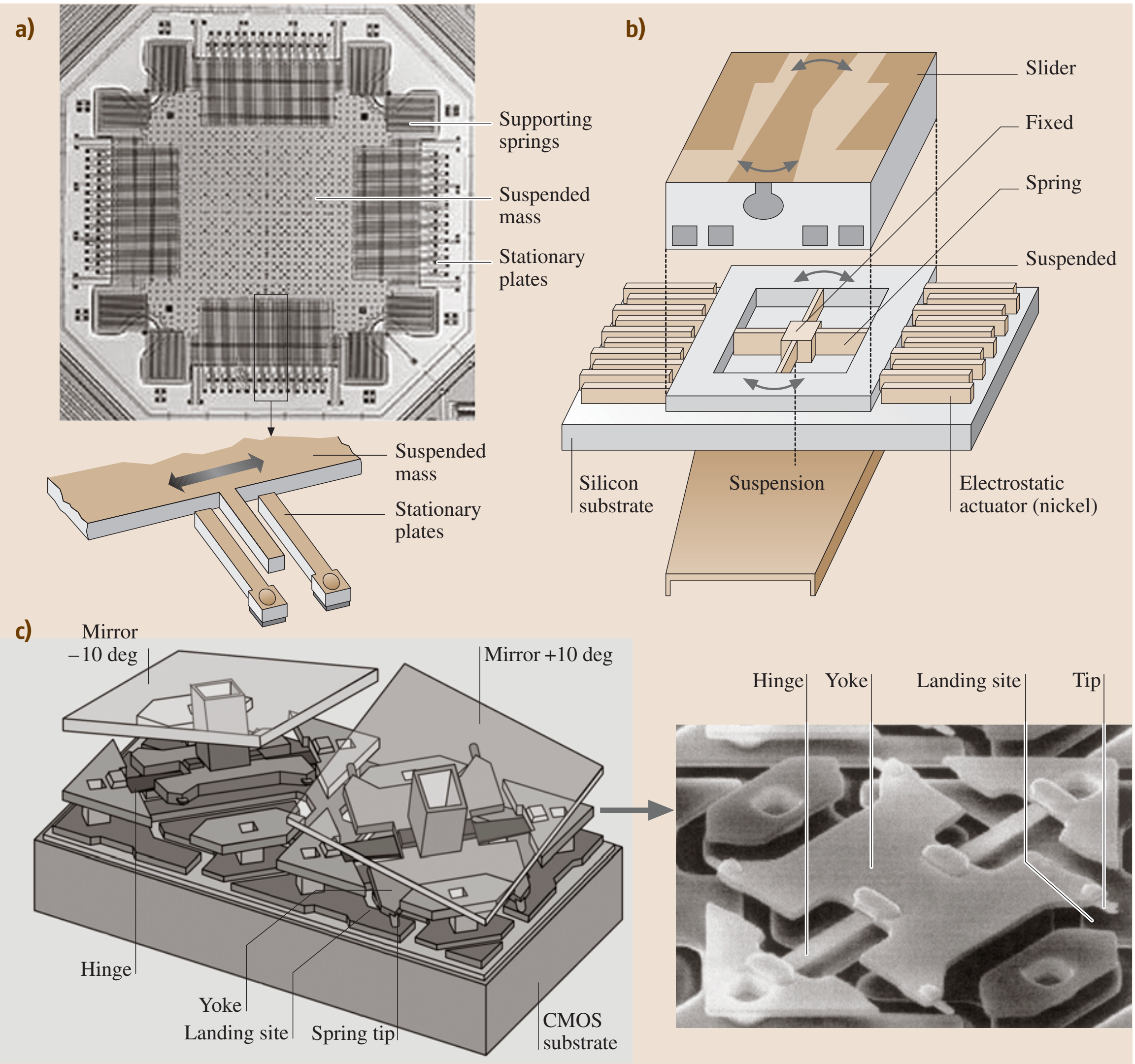
a)
Supporting springs
Suspended mass
Stationary plates
Suspended mass
Stationary plates
b)
Slider
Fixed
Spring
Suspended
Silicon substrate
Suspension
Electrostatic actuator (nickel)
c)
Mirror −10 deg
Mirror +10 deg
Hinge
Yoke
Landing site
Spring tip
CMOS substrate
Hinge
Yoke
Landing site
Tip

Fig. 26.3a–c Schematics of **(a)** a capacitive-type silicon accelerometer for automotive sensory applications, **(b)** digital micrometer devices for high-projection displays, and **(c)** polysilicon rotary microactuator for magnetic disk drives

σ bonds, and the remaining two electrons left in orthogonal p_y and p_z orbitals to form weak π bonds). Short-range order controls the properties of amorphous materials and coatings. Hard amorphous carbon (a-C) coatings, commonly known as diamond-like carbon or DLC (implying high hardness) coatings, are a class of coatings that are mostly metastable amorphous materials, but include a micro- or nanocrystalline phase. The coatings are random networks of covalently bonded carbon in hybridized tetragonal (sp^3) and trigonal (sp^2) local coordination with some of the bonds terminated by hydrogen. These coatings have been successfully deposited by a variety of vacuum deposition techniques on a variety of substrates at or near room temperature. These coatings generally reproduce substrate topography and do not require any post-finishing. However, these coatings mostly adhere best on silicon substrates. The best adhesion is obtained on substrates that form carbides, e.g., Si, Fe, and Ti. Based on depth profile analyses using Auger and XPS of DLC coatings deposited on silicon substrates, it has been reported that a substantial amount of silicon carbide (on the order of 5–10 nm in thickness) is present at the carbon-silicon interface for the coatings with good adhesion and high hardness (e.g., [26.9]). For good adhesion of DLC coatings to other substrates, in most cases, an interlayer of silicon is required except for cathodic arc deposited coatings.

There is significant interest in DLC coatings because of their unique combination of desirable properties. These properties include high hardness and wear resistance, chemical inertness to both acids and alkalis, lack of magnetic response, and an optical band gap ranging from zero to a few eV, depending upon the deposition conditions. These are used in a wide range of applications, including tribological, optical, electronic, and biomedical applications [26.1, 10, 11]. The high hardness, good friction and wear properties, versatility in deposition and substrates, and no requirements of post-finishing make them very attractive for tribological applications. Two primary examples include overcoats for magnetic media (thin-film disks and ME tapes) and MR type magnetic heads for magnetic storage devices, Fig. 26.2 [26.12–20] and the emerging field of microelectromechanical systems, Fig. 26.3 [26.21–24]. The largest industrial application of the family of amorphous carbon coatings, typically deposited by DC/RF magnetron sputtering, plasma-enhanced chemical vapor deposition, or ion beam deposition techniques, is in magnetic storage devices. These are employed to protect against wear and corrosion, magnetic coatings on thin-film rigid disks and metal evaporated tapes, and the thin-film head structure of a read/write disk head (Fig. 26.2). To maintain low physical spacing between the magnetic element of a read/write head and the magnetic layer of a media, thicknesses ranging from 3 to 10 nm are employed. Mechanical properties affect friction wear and therefore need to be optimized. In 1998, Gillette introduced Mach 3 razor blades with ultrathin DLC coatings, which has the potential of becoming a very large industrial application. DLC coatings are also used in other commercial applications such as glass windows of supermarket laser barcode scanners and sunglasses. These coatings are actively pursued in microelectromechanical systems (MEMS) components [26.23].

In this chapter, a state-of-the-art review of recent developments in the chemical, mechanical, and tribological characterization of ultrathin amorphous carbon coatings is presented. An overview of the most commonly used deposition techniques is presented, and followed by typical chemical and mechanical characterization data and typical tribological data both from coupon level testing and functional testing.

26.1 Description of Commonly Used Deposition Techniques

The first hard amorphous carbon coatings were deposited by a beam of carbon ions produced in an argon plasma on room-temperature substrates, as reported by *Aisenberg* and *Chabot* [26.25]. Subsequent confirmation by *Spencer* et al. [26.26] led to the explosive growth of this field. Following the first work, several alternative techniques have been developed. The amorphous carbon coatings have been prepared by a variety of deposition techniques and precursors, including evaporation, DC, RF or ion beam sputtering, RF or DC plasma-enhanced chemical vapor deposition (PECVD), electron cyclotron resonance chemical vapor deposition (ECR-CVD), direct ion beam deposition, pulsed laser vaporization and vacuum arc, from a variety of carbon-bearing solids or gaseous source materials [26.1, 27]. Coatings with both graphitic and diamond-like properties have been produced. Evaporation and ion plating techniques have been used to produce coatings with graphitic properties (low hardness, high electrical conductivity, very low friction, etc.), and all techniques have been used to produce coatings with diamond-like properties.

The structure and properties of a coating are dependent upon the deposition technique and parameters. High-energy surface bombardment has been used to produce harder and denser coatings. It is reported that sp^3/sp^2 fractions are in the decreasing order for cathodic arc deposition, pulsed laser vaporization, direct ion beam deposition, plasma-enhanced chemical vapor deposition, ion beam sputtering, and DC/RF sputtering [26.12,28,29]. A common feature to these techniques is that the deposition is energetic, i. e., carbon species arrive with an energy significantly greater than that represented by the substrate temperature. The resultant coatings are amorphous in structure, with hydrogen content up to 50%, and display a high degree of sp^3 character. From the results of previous investigations, it has been proposed that deposition of sp^3-bonded carbon requires that the depositing species have kinetic energies on the order of 100 eV or higher, well above

Table 26.1 Summary of the most commonly used deposition techniques and the kinetic energy of depositing species and deposition rates

Deposition technique	Process	Kinetic energy (eV)	Deposition rate (nm/s)
Filtered cathodic arc (FCA)	Energetic carbon ions produced by a vacuum arc discharge between a graphite cathode and grounded anode	100–2,500	0.1–1
Direct ion beam (IB)	Carbon ions produced from methane gas in an ion source and accelerated toward a substrate	50–500	0.1–1
Plasma-enhanced chemical vapor deposition (PECVD)	Hydrocarbon species, produced by plasma decomposition of hydrocarbon gases (e.g., acetylene), are accelerated toward a DC-biased substrate	1–30	1–10
Electron cyclotron resonance plasma chemical vapor deposition (ECR-CVD)	Hydrocarbon ions, produced by plasma decomposition of ethylene gas in the presence of a plasma in electron cyclotron resonance condition, are accelerated toward a RF-biased substrate	1–50	1–10
DC/RF sputtering	Sputtering of graphite target by argon ion plasma	1–10	1–10

those obtained in thermal processes like evaporation (0–0.1 eV). The species must then be quenched into the metastable configuration via rapid energy removal. Excess energy, such as that provided by substrate heating, is detrimental to the achievement of a high sp^3 fraction. In general, a high fraction of the sp^3-bonded carbon atoms in an amorphous network results in a higher hardness [26.29–36]. The mechanical and tribological properties of a carbon coating depend on the sp^3/sp^2-bonded carbon ratio, the amount of hydrogen in the coating, and adhesion of the coating to the substrate, which are influenced by the precursor material, kinetic energy of the carbon species prior to deposition, deposition rate, substrate temperature, substrate biasing, and the substrate itself [26.29, 33, 35, 37–46]. The kinetic energies and deposition rates involved in selected deposition processes used in the deposition of DLC coatings are compared in Table 26.1 [26.1, 28].

In the studies by *Gupta* and *Bhushan* [26.12, 47], *Li* and *Bhushan* [26.48, 49], and *Sundararajan* and *Bhushan* [26.50], DLC coatings ranging typically in thickness from 3.5 nm to 20 nm were deposited on single-crystal silicon, magnetic Ni-Zn ferrite, and Al_2O_3-TiC substrates (surface roughness ≈ 1–3 nm RMS) by filtered cathodic arc (FCA) deposition, (direct) ion beam deposition (IBD), electron cyclotron resonance chemical vapor deposition (ECR-CVD), plasma-enhanced chemical vapor deposition (PECVD), and DC/RF planar, magnetron sputtering (SP) deposition techniques [26.51]. In this chapter, we will limit the presentation of data of coatings deposited by FCA, IBD, ECR-CVD, and SP deposition techniques.

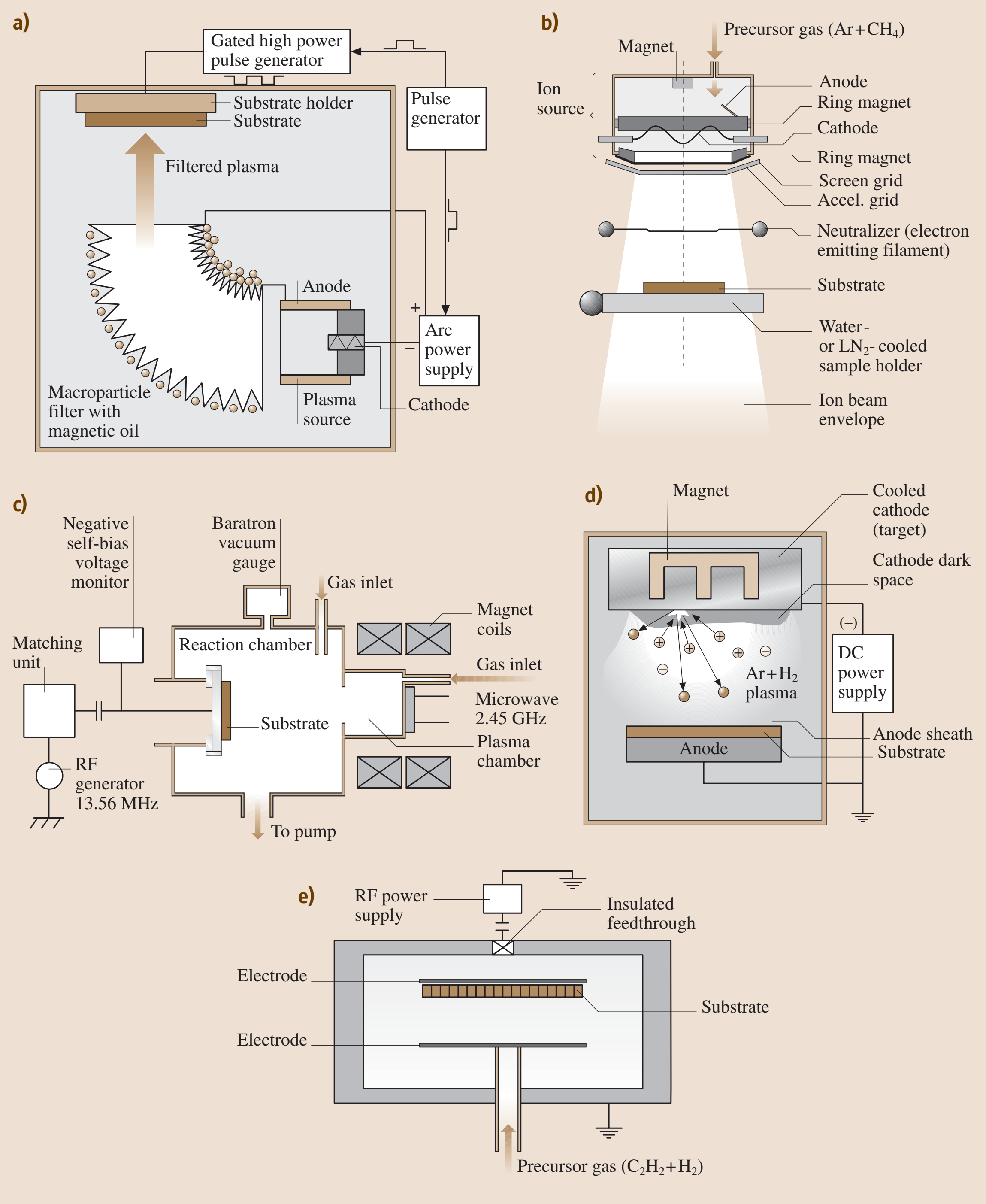
a)
Gated high power pulse generator
Substrate holder
Substrate
Pulse generator
Filtered plasma
Anode
+
−
Arc power supply
Macroparticle filter with magnetic oil
Plasma source
Cathode
b)
Magnet
Precursor gas (Ar+CH$_4$)
Ion source
Anode
Ring magnet
Cathode
Ring magnet
Screen grid
Accel. grid
Neutralizer (electron emitting filament)
Substrate
Water- or LN$_2$- cooled sample holder
Ion beam envelope
c)
Negative self-bias voltage monitor
Baratron vacuum gauge
Gas inlet
Matching unit
Reaction chamber
Magnet coils
Gas inlet
Microwave 2.45 GHz
Substrate
Plasma chamber
RF generator 13.56 MHz
To pump
d)
Magnet
Cooled cathode (target)
Cathode dark space
(−)
DC power supply
Ar+H$_2$ plasma
Anode
Anode sheath
Substrate
e)
RF power supply
Insulated feedthrough
Electrode
Substrate
Electrode
Precursor gas (C$_2$H$_2$+H$_2$)

Fig. 26.4a–e Schematic diagrams of deposition by **(a)** filtered cathodic arc deposition, **(b)** ion beam deposition, **(c)** electron cyclotron resonance chemical vapor deposition (ECR-CVD), **(d)** DC planar magnetron sputtering, and **(e)** plasma-enhanced chemical vapor deposition (PECVD)

26.1.1 Filtered Cathodic Arc Deposition Technique

In the filtered cathodic arc deposition of carbon coating [26.29, 52–59], a vacuum arc plasma source is used to form carbon film. In the FCA technique used by *Bhushan* et al. (e.g., [26.12]), energetic carbon ions are produced by a vacuum arc discharge between a planar graphite cathode and grounded anode, Fig. 26.4a. The cathode is a 6-mm-diameter high-density graphite disk mounted on a water-cooled copper block. The arc is driven at an arc current of 200 A, arc duration of 5 ms, and arc repetition rate of 1 Hz. The plasma beam is guided by a magnetic field that transports current between the electrodes to form tiny, rapidly moving spots on the cathode surface. The source is coupled to a 90° bent magnetic filter to remove the macroparticles produced concurrently with the plasma in the cathode spots. The ion current density at the substrate is in the range of 10–50 $\mathrm{mA/cm^2}$. The base pressure is less than 10^{-4} Pa. Compared with electron beam evaporation with auxiliary discharge, much higher plasma density is achieved with the aid of powerful arc discharge. In this process, the cathodic material suffers a complicated transition from the solid phase to an expanding, nonequilibrium plasma via liquid and dense, equilibrium non-ideal plasma phases [26.58]. The carbon ions in the vacuum arc plasma have a direct kinetic energy of 20–30 eV. The high voltage pulses are applied to the substrate mounted on a water-cooled sample holder, and ions are accelerated through the sheath and arrive at the substrate with an additional energy given by the potential difference between the plasma and the substrate. The substrate holder is pulse biased to a negative voltage up to −2 kV with a pulse duration of 1 μs. The negative biasing of −2 kV corresponds to 2 keV kinetic energy of the carbon ions. The use of a pulse bias instead of a DC bias has advantages of applying a much higher voltage and building a surface potential on a nonconducting film. The energy of the ions is varied during the deposition. For the first 10% of the deposition the substrates are pulsed biased to −2 keV with a pulse duty cycle of 25%, i. e., for 25% of the time the energy is 2 keV, for the remaining 75% it is 20 eV, which is the "natural" energy of carbon ions in a vacuum discharge. For the last 90% of the deposition the pulsed bias voltage is reduced to −200 eV with a pulse bias duty cycle of 25%, i. e., the energy is 200 eV for 25% and 20 eV for 75% of the deposition. The high energy at the beginning leads to a good intermixing and adhesion of the films, whereas the lower energy at the later stage leads to hard films. Under the conditions described, the deposition rate at the substrate is about 0.1 nm/s, which is slow. Compared with most gaseous plasma, the cathodic arc plasma is nearly fully ionized, and the ionized carbon atoms have high kinetic energy of carbon ions which help achieving a high fraction of sp^3-bonded carbon ions, which in turn result in a high hardness and higher interfacial adhesion. *Cuomo* et al. [26.42] have reported that based on electron energy loss spectroscopy (EELS) analysis, the sp^3-bonded carbon fraction of cathodic arc coating is 83% compared to 38% for the ion beam sputtered carbon. These coatings are reported to be *nonhydrogenated*.

This technique does not require an adhesion underlayer for non-silicon substrates. However, adhesion of the DLC coatings on the electrically insulating substrate is poor, as negative pulse biasing forms an electrical sheath that accelerates depositing ions to the substrate and enhances the adhesion of the coating to the substrate with associated ion implantation. It is difficult to build potential on an insulating substrate, and lack of biasing results in poor adhesion.

26.1.2 Ion Beam Deposition Technique

In the direct ion beam deposition of carbon coating [26.60–64], as used by *Bhushan* et al. (e.g., [26.12]), the carbon coating is deposited from an accelerated carbon ion beam. The sample is pre-cleaned by ion etching. For the case of non-silicon substrates, a 2–3-nm-thick amorphous silicon adhesion layer is deposited by ion beam sputtering using an ion beam of a mixture of methane and argon at 200 V. For the carbon deposition, the chamber is pumped to about 10^{-4} Pa, and methane gas is fed through the cylindrical ion source and is ionized by energetic electrons produced by a hot-wire filament, Fig. 26.4b. Ionized species then pass through a grid with a bias voltage of about 50 eV, where they gain a high acceleration energy and reach a hot-wire filament, emitting thermionic electrons that neutralize the incoming ions. The discharging of ions is important when insulating ceramics are used as substrates. The species are then deposited on a water-cooled substrate. Operating conditions are adjusted to give an ion beam with an acceleration energy of about 200 eV and a current density of about 1 $\mathrm{mA/cm^2}$. At this operating condition, the deposition rate is about 0.1 nm/s, which is slow. Incidentally, tough and soft coatings are deposited at a high acceleration energy of about 400 eV and at a deposition rate of about 1 nm/s. The ion beam deposited carbon coatings are reported to be hydrogenated (30–40 at% hydrogen).

26.1.3 Electron Cyclotron Resonance Chemical Vapor Deposition Technique

ECR plasma's lack of electrodes and ability to create high densities of charged and excited species at low pressures ($\leq 10^{-4}$ Torr) make it an attractive processing discharge in the coating depositions [26.65]. In the ECR-CVD deposition process of carbon coating described by *Suzuki* and *Okada* [26.66] and used by *Li* and *Bhushan* [26.48, 49] and *Sundararajan* and *Bhushan* [26.50], microwave power is generated by a magnetron operating in continuous mode at a frequency of 2.45 GHz, Fig. 26.4c. The plasma chamber functions as a microwave cavity resonator. The magnetic coils arranged around the plasma chamber generate a magnetic field of 875 G, necessary for electron cyclotron resonance condition. The substrate is placed on a stage that is connected capacitively to a 13.56 MHz RF generator. The process gas is introduced into the plasma chamber and the hydrocarbon ions generated are accelerated by a negative self-bias voltage, which is generated by applying RF power to the substrate. Both the substrate stage and plasma chamber are water-cooled. The process gas used is 100% ethylene and its flow rate is held constant at 100 sccm. The microwave power is 100–900 W. The RF power is 30–120 W. The pressure during deposition is kept close to the optimum value of 5.5×10^{-3} Torr. Before the deposition, the substrates are cleaned using Ar ions generated in the ECR plasma chamber.

26.1.4 Sputtering Deposition Technique

In DC planar magnetron sputtered carbon coating [26.13, 33, 37, 40, 67–71], the carbon coating is deposited by the sputtering of graphite target with Ar ion plasma. In the glow discharge, positive ions from the plasma strike the target with sufficient energy to dislodge the atoms by momentum transfer, which are intercepted by the substrate. First, an about 5-nm-thick amorphous silicon adhesion layer is deposited by sputtering if the deposition is to be carried out on a non-silicon surface. In the process used by *Bhushan* et al. (e.g., [26.12]), the coating is deposited by the sputtering of a graphite target with Ar ion plasma at 300 W power for a 200-mm-diameter target and pressure of about 0.5 Pa (6 mTorr), Fig. 26.4d. Plasma is generated by applying a DC potential between the substrate and a target. *Bhushan* et al. [26.35] reported that sputtered carbon coating contains about 35 at% hydrogen. Hydrogen comes from the hydrocarbon contaminants present in the deposition chamber. To produce hydrogenated carbon coating with a larger concentration of hydrogen, deposition is carried out in Ar and hydrogen plasma.

26.1.5 Plasma-Enhanced Chemical Vapor Deposition Technique

In the RF-PECVD deposition of carbon coating, as used by *Bhushan* et al. (e.g., [26.12]), carbon coating is deposited by adsorption of most free radicals of hydrocarbon to the substrate and chemical bonding to other atoms on the surface. The hydrocarbon species are produced by the RF plasma decomposition of hydrocarbon precusors such as acetylene (C_2H_2), Fig. 26.4e [26.27, 69, 72–75]. Instead of requiring thermal energy in thermal CVD, the energetic electrons in the plasma (at pressures ranging from 1 to 5×10^2 Pa, typically less than 10 Pa) can activate almost any reaction among the gases in the glow discharge at relatively low substrate temperatures ranging from 100 to 600 °C (typically less than 300 °C). To deposit the coating on non-silicon substrates, an about 4-nm-thick amorphous silicon adhesion layer, used to improve adhesion, is first deposited under similar conditions from a gas mixture of 1% silane in argon [26.76]. In the process used by Bhushan and coworkers [26.12], the plasma is sustained in a parallel-plate geometry by a capacitive discharge at 13.56 MHz, at a surface power density on the order of 100 mW/cm^2. The deposition is performed at a flow rate on the order of 6 sccm and a pressure on the order of 4 Pa (30 mTorr) on a cathode-mounted substrate maintained at a substrate temperature of 180 °C. The cathode bias is held fixed at about −120 V with an external DC power supply attached to the substrate (powered electrode). The carbon coatings deposited by PECVD usually contain hydrogen up to 50% [26.35, 77].

26.2 Chemical Characterization and Effect of Deposition Conditions on Chemical Characteristics and Physical Properties

The chemical structure and properties of amorphous carbon coatings are a function of deposition conditions. It is important to understand the relationship of the chemical structure of amorphous carbon coatings to the properties in order to define useful deposition parameters. Amorphous carbon films are metastable phases formed when carbon particles are condensed on a substrate. The prevailing atomic arrangement in the DLC coatings is amorphous or quasi-amorphous with small diamond (sp^3), graphite (sp^2), and other unidentifiable micro- or nanocrystallites. The coating is dependent upon the deposition process and its conditions contain varying amounts of sp^3/sp^2 ratio and hydrogen. The sp^3/sp^2 ratio of DLC coatings ranges typically from 50% to close to 100% with an increase in hardness with the sp^3/sp^2 ratio. Most DLC coatings, except those produced by filtered cathodic arc, contain from a few to about 50 at% hydrogen. Sometimes hydrogen and nitrogen are deliberately added to produce hydrogenated (a-C:H) and nitrogenated amorphous carbon (a-C:N) coatings, respectively. Hydrogen helps to stabilize sp^3 sites (most of the carbon atoms attached to hydrogen have a tetrahedral structure); therefore, the sp^3/sp^2 ratio for hydrogenated carbon is higher [26.30]. Optimum sp^3/sp^2 in a random covalent network composed of sp^3 and sp^2 carbon sites (N_{sp^2} and N_{sp^3}) and hydrogen is [26.30]:

$$\frac{N_{sp^3}}{N_{sp^2}} = \frac{6X_H - 1}{8 - 13X_H}\,, \tag{26.1}$$

where X_H is the atom fraction of hydrogen. The hydrogenated carbon has a larger optical band gap, higher electrical resistivity (semiconductor), and a lower optical absorption or high optical transmission. The hydrogenated coatings have a lower density, probably because of the reduction of cross-linking due to hydrogen incorporation. However, hardness decreases with an increase of the hydrogen, even though the proportion of sp^3 sites increases (that is, as the local bonding environment becomes more diamond-like) [26.78, 79]. It is speculated that the high hydrogen content introduces frequent terminations in the otherwise strong 3-D network, and hydrogen increases the soft polymeric component of the structure more than it enhances the cross-linking sp^3 fraction.

A number of investigations have been performed to identify the microstructure of amorphous carbon films using a variety of techniques such as Raman spectroscopy, EELS, nuclear magnetic resonance, optical measurements, transmission electron microscopy, and X-ray photoelectron spectroscopy [26.33]. The structure of diamond-like amorphous carbon is amorphous or quasi-amorphous with small graphitic (sp^2) and tetrahedrally coordinated (sp^3) and other unidentifiable nanocrystallites (typically on the order of a couple nm, randomly oriented) [26.33, 80, 81]. These studies indicate that the chemical structure and physical properties of the coatings are quite variable, depending on the deposition techniques and film growth conditions. It is clear that both sp^2 and sp^3-bonded atomic sites are incorporated in diamond-like amorphous carbon coatings and that the physical and chemical properties of the coatings depend strongly on their chemical bonding and microstructure. Systematic studies have been conducted to carry out chemical characterization and investigate how the physical and chemical properties of amorphous carbon coatings vary as a function of deposition parameters (e.g., [26.33, 35, 40]). EELS and Raman spectroscopy are commonly used to characterize the chemical bonding and microstructure. Hydrogen concentration of the coatings is obtained by means of forward recoil spectrometry (FRS). A variety of physical properties of the coatings relevant to tribological performance are measured.

To present the typical data obtained for characterization of typical amorphous carbon coatings and their relationships to physical properties, we present data on several sputtered coatings, RF-PECVD amorphous carbon and microwave-PECVD (MPECVD) diamond coatings [26.33,35,40]. The sputtered coatings were DC magnetron sputtered at a chamber pressure of 10 mTorr under sputtering power densities of 0.1 and 2.1 W/cm^2 in a pure Ar plasma, labeled as W1 and W2, respectively. They were prepared at a power density of 2.1 W/cm^2 with various hydrogen fractions of 0.5, 1, 3, 5, 7 and 10% of Ar/H, the gas mixtures labeled as H1, H2, H3, H4, H5, and H6, respectively.

26.2.1 EELS and Raman Spectroscopy

EELS and Raman spectra of four sputtered (W1, W2, H1, and H3) and one PECVD carbon samples were obtained. Figure 26.5 shows the EELS spectra of these carbon coatings. EELS spectra for bulk diamond and polycrystalline graphite in an energy range up to 50 eV are also shown in Fig. 26.5. One prominent peak is seen at 35 eV in diamond, while two peaks are seen at 27 eV and 6.5 eV in graphite, which are called $(\pi+\sigma)$ and (π) peaks, respectively. These peaks are produced by the energy loss of transmitted electrons to the plasmon oscillation of the valence electrons. The $\pi+\sigma$ peak in each coating is positioned at a lower energy region than that of graphite. The π peaks in the W series and PECVD samples are also seen at a lower energy region than that of the graphite. However, the π peaks in the H-series are comparable to or higher than that of graphite (see Table 26.2). The plasmon oscillation frequency is proportional to the square root of the corresponding electron density to a first approximation. Therefore, the samples in the H-series most likely have a higher density of π electrons than the other samples.

Amorphous carbon coatings contain (mainly) a mixture of sp^2- and sp^3-bonds, even though there is some evidence for the presence of sp-bonds as well [26.82]. The PECVD coatings and the H-series coatings in this study have nearly the same mass density, as seen in Table 26.4, to be presented later, but the former have a lower concentration of hydrogen (18.1%) than the H-series (35–39%), as seen in Table 26.3, to be presented later. The relatively low energy position of π peaks of PECVD coatings, compared to those of the H-series, indicates that the PECVD coatings contain a higher fraction of sp^3-bonds than the sputtered hydrogenated carbon coatings (H-series).

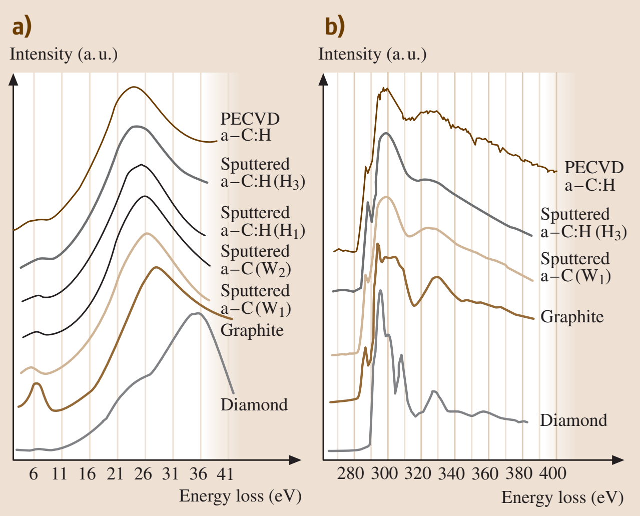

Fig. 26.5 **(a)** Low energy, and **(b)** high energy EELS of DLC coatings produced by DC magnetron sputtering and RF-PECVD techniques. Data for bulk diamond and polycrystalline graphite are included for comparison [26.35]

Figure 26.5b shows the EELS spectra associated with the inner-shell (K-shell) ionization. Again, the spectra for diamond and polycrystalline graphite are included for comparison. Sharp peaks are observed at

Table 26.2 Experimental results from EELS and Raman spectroscopy [26.35]

Sample	EELS peak position		Raman peak position		Raman FWHM^a		I_D/I_G^d
	π (eV)	$\pi+\sigma$ (eV)	G-band^b (cm^{-1})	D-band^c (cm^{-1})	G-band (cm^{-1})	D-band (cm^{-1})	
Sputtered *a*-C coating (W1)	5.0	24.6	1541	1368	105	254	2.0
Sputtered *a*-C coating (W2)	6.1	24.7	1560	1379	147	394	5.3
Sputtered *a*-C:H coating (H1)	6.3	23.3	1542	1334	95	187	1.6
Sputtered *a*-C:H coating (H3)	6.7	22.4	e	e	e	e	e
PECVD *a*-C:H coating	5.8	24.0	1533	1341	157	427	1.5
Diamond coating	...	...	1525^f	1333^g	...	8^g	...
Graphite (for reference)	6.4	27.0	1580	1358	37	47	0.7
Diamond (for reference)	...	37.0	...	1332^g	...	2^g	...

[a] Full width at half maximum width
[b] Peak associated with sp^2 "graphite" carbon
[c] Peak associated with sp^2 "disordered" carbon (not sp^3-bonded carbon)
[d] Intensity ratio of the D-band to the G-band
[e] Fluorescence
[f] Includes D and G band, signal too weak to analyze
[g] Peak position and width for diamond phonon

Table 26.3 Experimental results of FRS analysis [26.35]

Sample	Ar/H ratio	C (at% ± 0.5)	H (at% ± 0.5)	Ar (at% ± 0.5)	O (at% ± 0.5)
Sputtered *a*-C coating (W2)	100/0	90.5	9.3	0.2	...
Sputtered *a*-C:H coating (H2)	99/1	63.9	35.5	0.6	...
Sputtered *a*-C:H coating (H3)	97/3	56.1	36.5	...	7.4
Sputtered *a*-C:H coating (H4)	95/5	53.4	39.4	...	7.2
Sputtered *a*-C:H coating (H5)	93/7	58.2	35.4	0.2	6.2
Sputtered *a*-C:H coating (H6)	90/10	57.3	35.5	...	7.2
PECVD *a*-C:H coating	99.5% CH_4	81.9	18.1	...	...
Diamond coating	H_2-1 mole % CH_4	94.0	6.0	...	...

Table 26.4 Experimental results of physical properties [26.35]

Sample	Mass density (g/cm^3)	Nano-hardness (GPa)	Elastic modulus (GPa)	Electrical resistivity (Ohm-cm)	Compressive residual stress (GPa)
Sputtered *a*-C coating (W1)	2.1	15	141	1300	0.55
Sputtered *a*-C:H coating (W2)	1.8	14	136	0.61	0.57
Sputtered *a*-C:H coating (H1)	...	14	96	...	> 2
Sputtered *a*-C:H coating (H3)	1.7	7	35	$> 10^6$	0.3
PECVD *a*-C:H coating	1.6–1.8	33–35	∼ 200	$> 10^6$	1.5–3.0
Diamond coating	...	40–75	370–430	...	...
Graphite (for reference)	2.267	Soft	9–15	5×10^{-5a}, 4×10^{-3b}	0
Diamond (for reference)	3.515	70–102	900–1050	10^7–10^{20}	0

[a] Parallel to layer planes

[b] Perpendicular to layer planes

285.5 eV and 292.5 eV in graphite, while no peak is seen at 285.5 eV in diamond. The general features of the K-shell EELS spectra for the sputtered and PECVD carbon samples resemble those of graphite, but with the higher energy features smeared. The observation of the peak at 285.5 eV in the sputtered and PECVD coatings also indicates the presence of sp^2-bonded atomic sites in the coatings. All these spectra peak at 292.5 eV, similar to the spectra of graphite, but the peak in graphite is sharper.

Raman spectra from samples W1, W2, H1, and PECVD are shown in Fig. 26.6. Raman spectra could not be observed in specimens H2 and H3 due to high flourescence signals. The Raman spectra of single crystal diamond and polycrstalline graphite are also shown for comparison in Fig. 26.6. The results of the spectral fits are summarized in Table 26.2. We will focus on the G-band position, which has been shown to be related to the fraction of sp^3-bonded sites. Increasing the power density in the amorphous carbon coatings (W1 and W2) results in a higher G-band frequency, implying a smaller fraction of sp^3-bonding in W2 than in W1. This is consistent with higher density of W1. H1 and PEVCD have still lower G-band positions than W1, implying an even higher fraction of sp^3-bonding, which is presumably caused by the incorporation of H atoms into the lattice. The high hardness of H3 might be attributed to efficient sp^3 cross-linking of small, sp^2-ordered domains.

The Raman spectrum of a MPECVD diamond coating is shown in Fig. 26.6. The diamond Raman peak is at 1,333 cm^{-1} with a line width of 7.9 cm^{-1}. There is a small broad peak around 1,525 cm^{-1}, which is attributed to a small amount of a-C:H. This impurity peak is not intense enough to fit to separate G- and D-bands. The diamond peak frequency is very close to that of natural diamond (1,332.5 cm^{-1}, e.g. Fig. 26.6), indicating that the coating is not under stress [26.83]. The large line width compared to that of natural diamond (2 cm^{-1})

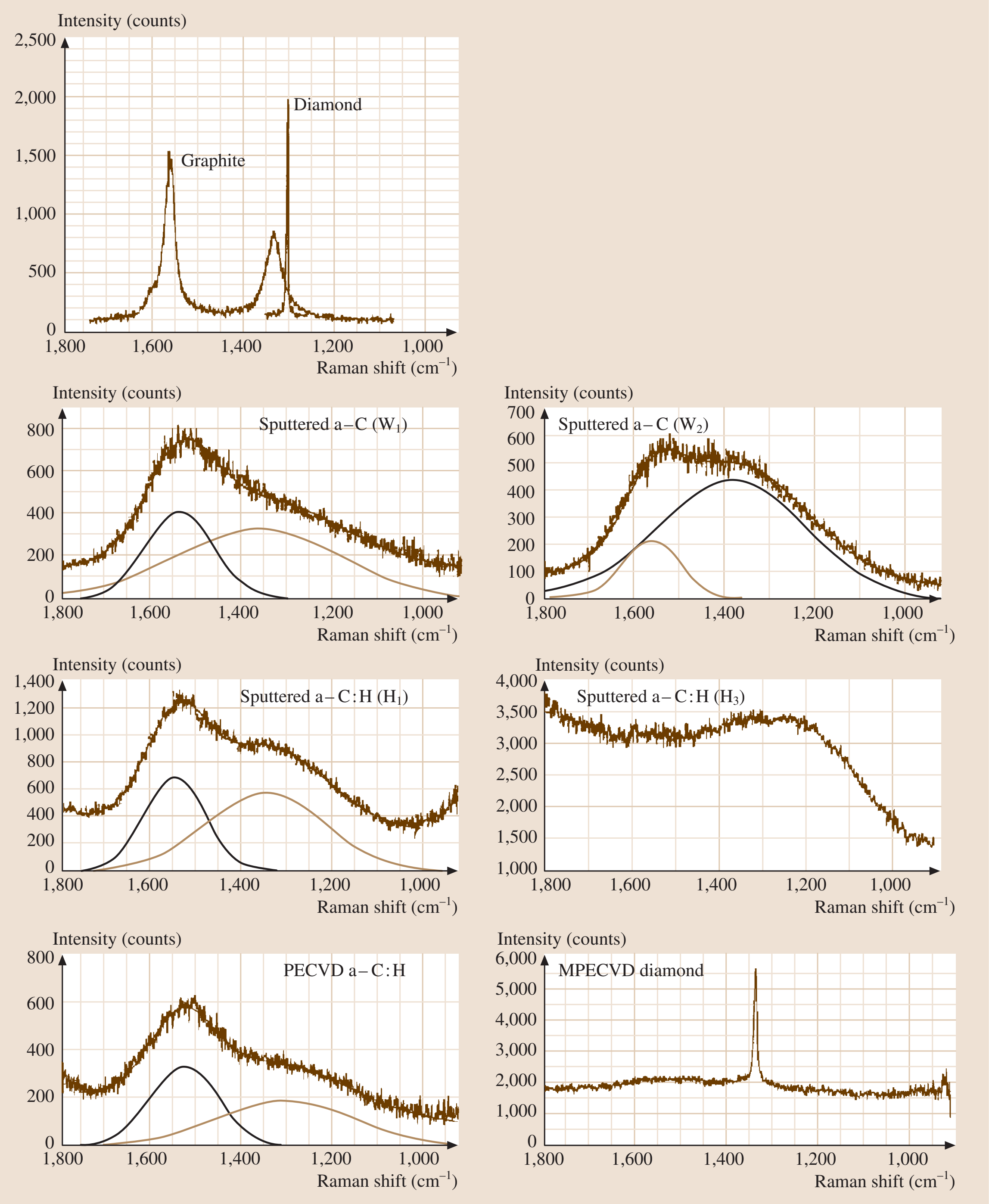

Fig. 26.6 Raman spectra of DLC coatings produced by DC magnetron sputtering and RF-PECVD techniques and a diamond film produced by MPECVD technique. Data for bulk diamond and microcrystalline graphite are included for comparison [26.35]

indicates that the microcrystallites likely have a high concentration of defects [26.84].

26.2.2 Hydrogen Concentrations

FRS analysis of six sputtered (W2, H2, H3, H4, H5, and H6) coatings, one PECVD coating, and one diamond coating was performed. Figure 26.7 shows an overlay of the spectra from the six sputtered samples. Similar spectra were obtained from the PECVD and the diamond films. Table 26.3 shows the H and C fractions, as well as the amount of impurities (Ar and O) in the films in atomic %. Most apparent is the large fraction of H in the sputtered films. Regardless of how much H_2 is in the Ar sputtering gas, the H content of the coatings is about the same, ~ 35 at%. Interestingly, there is still ~ 10% H present in the coating sputtered in pure Ar (W2). It is interesting to note that Ar is present only in coatings grown under low (< 1%) H content in the Ar carrier gas. The presence of O in the coatings, combined with the fact that the coatings were prepared approximately nine months before the FRS analysis, caused suspicion that they had absorbed water vapor, and that this may be the cause for the H peak in specimen W2.

All samples were annealed for 24 h at 250 °C in a flowing He furnace and then reanalyzed. Surprisingly, the H content of all coatings measured increased slightly, even though the O content decreased, and W2 still had a substantial amount of H_2. This slight increase in H concentration is not understood. However, since H concentration did not decrease with the oxygen as a result of annealing, it suggests that high H concentration is not due to adsorbed water vapor. The PECVD film has more H (~ 18%) than the sputtered films initially, but after annealing it has the same fraction as specimen W2, the film sputtered in pure Ar. The diamond film has the smallest amounts of hydrogen, as seen in Table 26.3.

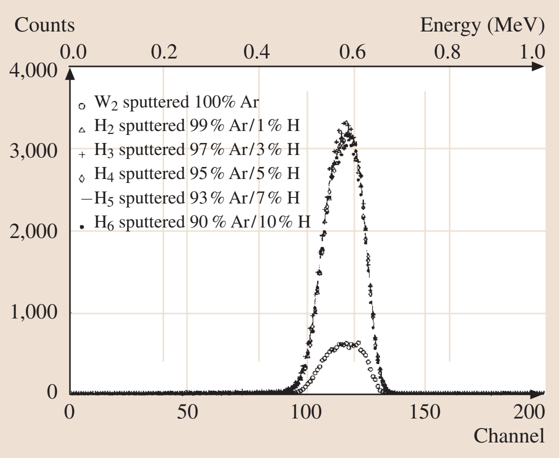

Fig. 26.7 FRS spectra of six DLC coatings produced by DC magnetron sputtering [26.35]

26.2.3 Physical Properties

Physical properties of the four sputtered (W1, W2, H1, and H3) coatings, one PECVD coating, one diamond coating, and bulk diamond and graphite are presented in Table 26.4. The hydrogenated carbon and the diamond coatings have very high resistivity compared to unhydrogenated carbon coatings. It appears that unhydrogenated carbon coatings have a higher density than the hydrogenated carbon coatings, although both groups are less dense than graphite. The density depends on the deposition technique and the deposition parameters. It appears that unhydrogenated sputtered coatings deposited at low power exhibit the highest density. Nanohardness of hydrogenated carbon is somewhat lower than that of the unhydrogenated carbon coatings. PECVD coatings are significantly harder than sputtered coatings. The nanohardness and modulus of elasticity of the diamond coatings are very high compared to that of DLC coatings, even though the hydrogen content is similar. The compressive residual stresses of the PECVD coatings are substantially higher than those of sputtered coatings, which is consistent with the hardness results.

Figure 26.8a shows the effect of hydrogen in the plasma on the residual stresses and the nanohardness for sputtered coatings W2 and H1 to H6. The coatings made with H_2 flow between 0.5 and 1.0% delaminate very quickly, even when only a few tens of nm thick. In pure Ar and at H_2 flows greater than 1%, the coatings appear to be more adhesive. The tendency of some coatings to delaminate can be caused by intrinsic stress in the coating, which is measured by substrate bending. All of the coatings in the figure are in compressive stress. The maximum stress occurs between 0 and 1% H_2 flow, but the stress cannot be quantified in this range because the coatings instantly delaminate upon exposure to air. At higher hydrogen concentrations the stress gradually diminishes. A generally decreasing trend is observed in the hardness of the coatings as hydrogen content increased. The hardness decreases slightly, going from 0% H_2 to 0.5% H_2, and then decreases sharply. These results are probably lower than the true values because of local delamination around the indentation point. This is especially likely for the 0.5% and 1.0% coatings where delamination is visually apparent, but may also be true to a lesser extent for the other coatings. Such an adjustment would bring the hardness profile into closer correlation with the stress profile. *Weissmantel* et al. [26.68] and *Scharff* et al. [26.85] observed a downturn in hardness for high bias and low pressure of hydrocarbon gas in ion plated carbon coating, and, therefore, presumably low hydrogen content in support of the above contention.

Figure 26.8b shows the effect of sputtering power (with no hydrogen added in the plasma) on the residual stresses and nanohardness for various sputtered coatings. As the power decreases, compressive stress does not seem to change while nanohardness slowly increases. The rate of change becomes more rapid at very low power levels.

The addition of H_2 during sputtering of carbon coatings increases H concentration in the coating. Hydrogen causes the character of the C-C bonds to shift from sp^2 to sp^3, and the rising number of C-H bonds, which ultimately relieves stress and produces a softer "polymer-like" material. Low power deposition, like the presence of hydrogen, appears to stabilize the formation of sp^3 C-C bonds, increasing hardness. These coatings have relieved stress and led to better adhesion. An increase in temperature during deposition at high power density results in graphitization of the coating material, responsible for a decrease in hardness with an increase in power density. Unfortunately, low power also means impractically low deposition rates.

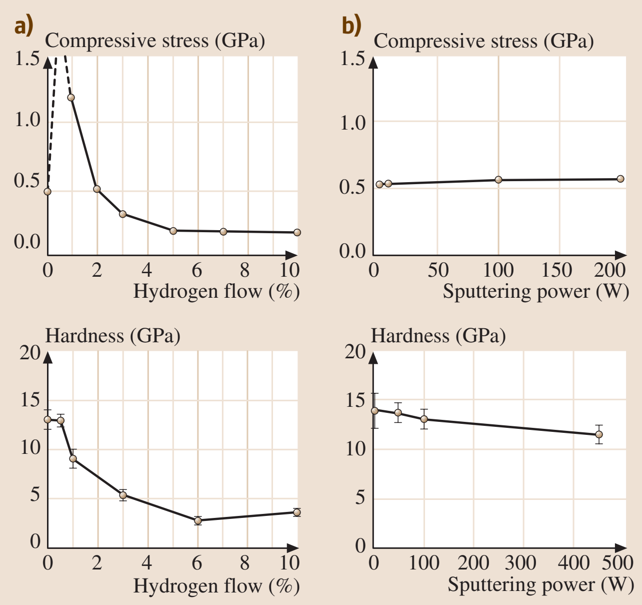

Fig. 26.8a,b Residual compressive stresses and nanohardness **(a)** as a function of hydrogen flow rate; sputtering power is 100 W and target diameter is 75 mm (power density = 2.1 W/cm^2), and **(b)** as a function of sputtering power over a 75-mm-diameter target with no hydrogen added in the plasma [26.40]

26.2.4 Summary

Based on the EELS and Raman data, all DLC coatings have both sp^2 and sp^3 bondings. The sp^2/sp^3 bonding ratio depends on the deposition techniques and parameters. The DLC coatings deposited by sputtering and PECVD contain significant concentrations of hydrogen, while the diamond coating contains only small amounts of hydrogen impurity. Sputtered coatings with no deliberate addition of hydrogen in the plasma contain a significant amount of hydrogen. Regardless of how much hydrogen is in the Ar sputtering gas, the hydrogen content of the coatings increases initially with no further increase.

Hydrogen flow and sputtering power density affect the mechanical properties of these coatings. Maximum compressive residual stress and hardness occur between 0 and 1% hydrogen flow, resulting in rapid delamination. Low sputtering power moderately increases hardness while relieving residual stress.

26.3 Micromechanical and Tribological Characterizations of Coatings Deposited by Various Techniques

26.3.1 Micromechanical Characterization

Common mechanical characterizations include measurement of hardness and elastic modulus, fracture toughness, fatigue life, and scratch and wear testing. Nanoindentation and atomic force microscopy (AFM) are used for mechanical characterization of ultrathin films.

Hardness and elastic modulus are calculated from the load displacement data obtained by nanoindentation at loads ranging typically from 0.2 to 10 mN using a commercially available nanoindenter [26.23,86].

This instrument monitors and records the dynamic load and displacement of the three-sided pyramidal diamond (Berkovich) indenter during indentation. For the fracture toughness measurement of ultrathin films ranging from 100 nm to a few µm, a nanoindentation-based technique is used in which through-thickness cracking in the coating is detected from a discontinuity observed in the load-displacement curve and energy released during the cracking is obtained from the curve [26.87–89]. Based on the energy released, fracture mechanics analysis is then used to calculate fracture toughness. An indenter with a cube-corner tip geometry is preferred because the through-thickness cracking of hard films can be accomplished at lower loads. In fatigue measurement, a conical diamond indenter having a tip radius of about one micron is used and load cycles of a sinusoidal shape are applied [26.90, 91]. The fatigue behavior of coatings is studied by monitoring the change in contact stiffness, which is sensitive to damage formation.

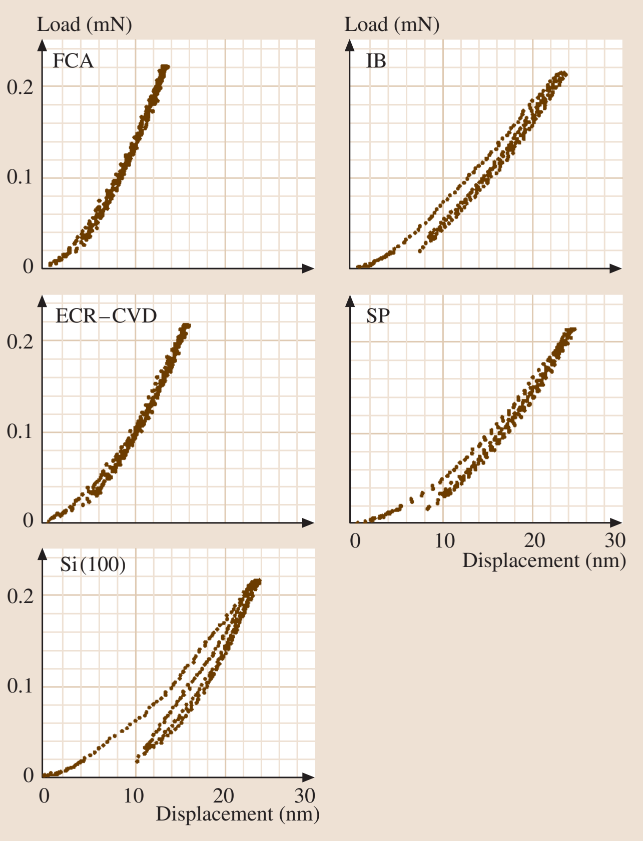

Fig. 26.9 Load versus displacement plots for various 100-nm-thick amorphous carbon coatings on single-crystal silicon substrate and bare substrate

Hardness and Elastic Modulus

For materials that undergo plastic deformation, high hardness and elastic modulus are generally needed for low friction and wear, whereas for brittle materials, high fracture toughness is needed [26.2, 3, 21]. DLC coatings used for many applications are hard and brittle, and values of hardness and fracture toughness need to be optimized.

Representative load-displacement plots of indentations made at 0.2 mN peak indentation load on 100-nm-thick DLC coatings deposited by the four deposition techniques on single-crystal silicon substrate are compared in Fig. 26.9. The indentation depths at the peak load range from about 18 to 26 nm, smaller than that of the coating thickness. Many of the coatings exhibit a discontinuity or pop-in marks in the loading curve, which indicate a sudden penetration of the tip into the sample. A nonuniform penetration of the tip into a thin coating possibly results from formation of cracks in the coating, formation of cracks at the coating-substrate interface, or debonding or delamination of the coating from the substrate.

The hardness and elastic modulus values at a peak load of 0.2 mN on the various coatings and single-crystal silicon substrate are summarized in Table 26.5 and Fig. 26.10 [26.47, 49, 89, 90]. Typical values for the peak and residual indentation depths ranged from 18 to 26 nm and 6 to 12 nm, respectively. The FCA coatings exhibit the highest hardness of 24 GPa and elastic modulus of 280 GPa of various coatings, followed by the ECR-CVD, IB, and SP coatings. Hardness and elastic modulus have been known to vary over a wide range with sp^3-to-sp^2 bonding ratio, which depends on the kinetic energy of the carbon species and amount of hydrogen [26.6, 30, 47, 92, 93]. The high hardness and elastic modulus of the FCA coatings are attributed to the high kinetic energy of carbon species involved in the FCA deposition [26.12, 47]. *Anders* et al. [26.57] also reported a high hardness, measured by nanoindentation, of about 45 GPa for cathodic arc carbon coatings. They observed a change in hardness from 25 to 45 GPa with pulsed bias voltage and bias duty cycle. The high hardness of cathodic arc carbon was attributed to the high percentage (more than 50%) of sp^3 bonding. *Savvides* and *Bell* [26.94] reported an increase in hardness from 12 to 30 GPa and an elastic modulus from 62 to 213 GPa with an increase of sp^3-to-sp^2 bonding ratio, from 3 to 6, for a C:H coating deposited by low-energy ion-assisted unbalanced magnetron sputtering of a graphite target in an Ar-H_2 mixture.

Bhushan et al. [26.35] reported hardnesses of about 15 and 35 GPa and elastic moduli of about 140 and 200 GPa, measured by nanoindentation, for a-C:H coatings deposited by DC magnetron sputtering and RF-plasma-enhanced chemical vapor deposition techniques, respectively. The high hardness of RF-PECVD a-C:H coatings is attributed to a higher concentration of sp^3 bonding than in a sputtered hydrogenated a-C:H coating. Hydrogen is believed to play a crucial role in the bonding configuration of carbon atoms by helping to stabilize tetrahedra-coordination (sp^3 bonding) of carbon species. *Jansen* et al. [26.78] suggested that the incorporation of hydrogen efficiently passivates the dangling bonds and saturates the graphitic bonding to some extent. However, a large concentration of hydrogen in the plasma in sputter deposition is undesirable. *Cho* et al. [26.33] and *Rubin* et al. [26.40] observed that hardness decreased from 15 to 3 GPa with increased hydrogen content. *Bhushan* and *Doerner* [26.95] reported a hardness of about 10–20 GPa and an elastic modulus of about 170 GPa, measured by nanoindentation, for 100-nm-thick DC-magnetron sputtered a-C:H on the silicon substrate.

Residual stresses measured using a well-known curvature measurement technique are also presented in Table 26.5. The DLC coatings are under significant compressive internal stresses. Very high compressive stresses in FCA coatings are believed to be partly responsible for their high hardness. However, high stresses result in coating delamination and buckling. For this reason, the coatings that are thicker than about 1 µm have a tendency to delaminate from the substrates.

Fracture Toughness

Representative load-displacement curves of indentations on the 400-nm-thick cathodic arc carbon coating on silicon at various peak loads are shown in Fig. 26.11. Steps are found in all curves, as shown by arrows in Fig. 26.11a. In the 30 mN SEM micrograph, in addition to several radial cracks, ring-like through-thickness cracking is observed with small lips of material overhanging the edge of indentation. The step at about 23 mN in the loading curves of indentations made at 30 and 100 mN peak indentation loads results from the ring-like through-thickness cracking. The step at 175 mN in the loading curve of the indentation made at 200 mN peak indentation load is caused by spalling and second ring-like through-thickness cracking.

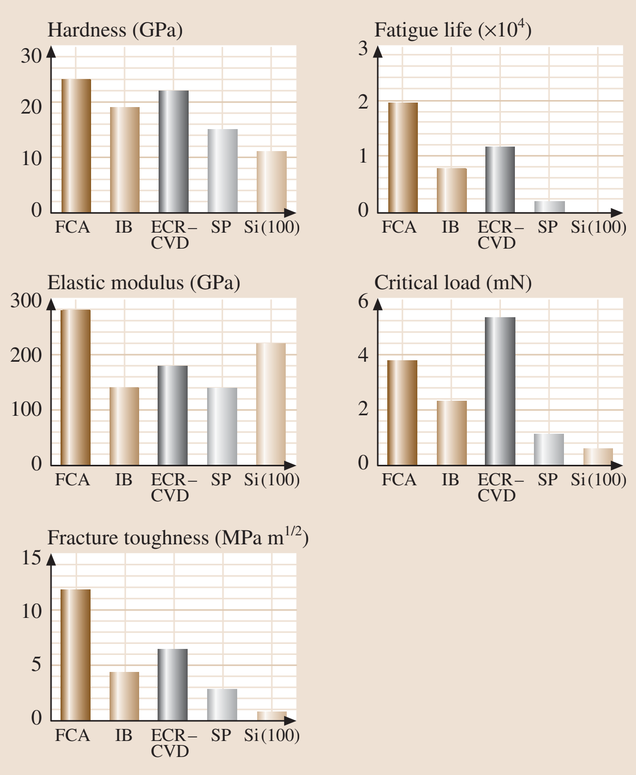

Fig. 26.10 Bar charts summarizing the data of various coatings and single-crystal silicon substrate. Hardness, elastic modulus, and fracture toughness were measured on 100-nm-thick coatings, and fatigue life and critical load during scratch were measured on 20-nm-thick coatings

Based on *Li* et al. [26.87], the fracture process progresses in three stages: (1) ring-like through-thickness cracks form around the indenter by high stresses in the contact area, (2) delamination and buckling occur around contact area at the coating/substrate interface by high lateral pressure, (3) second ring-like through-thickness cracks and spalling are generated by high bending stesses at the edges of the buckled coating, see Fig. 26.12a. In the first stage, if the coating under the indenter is separated from the bulk coating via the first ring-like through-thickness cracking, a corresponding step will be present in the loading curve. If discontinuous cracks form and the coating under the indenter is not separated from the remaining coating, no step appears in the loading curve, because the coating still supports the indenter and the indenter cannot suddenly advance into the material. In the second stage,

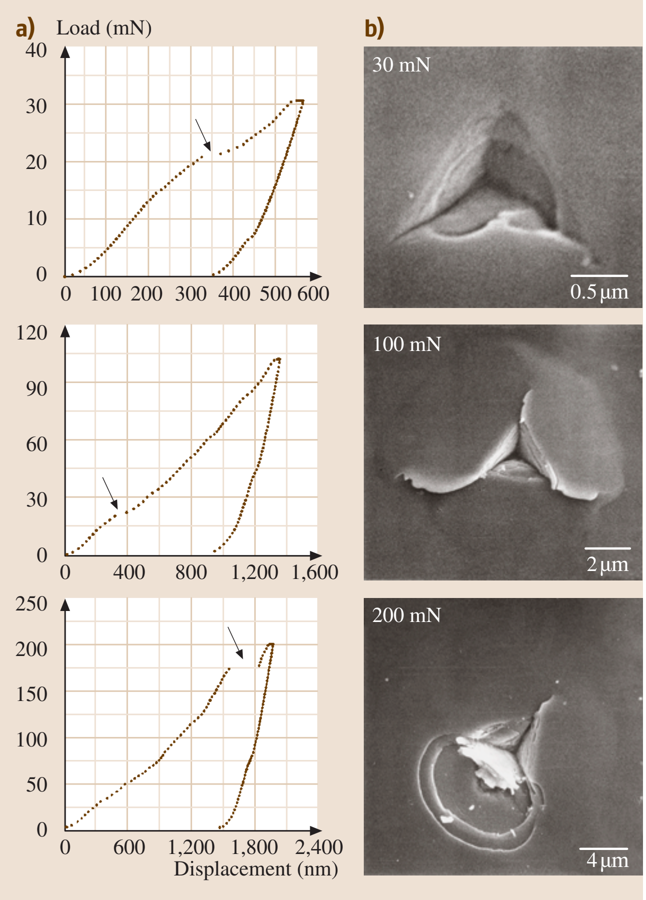

Fig. 26.11 **(a)** Load-displacement curves of indentations made at 30, 100, and 200 mN peak indentation loads using the cube corner indenter, and **(b)** the SEM micrographs of indentations on the 400-nm-thick cathodic arc carbon coating on silicon. *Arrows* indicate steps during loading portion of the load-displacement curve [26.87]

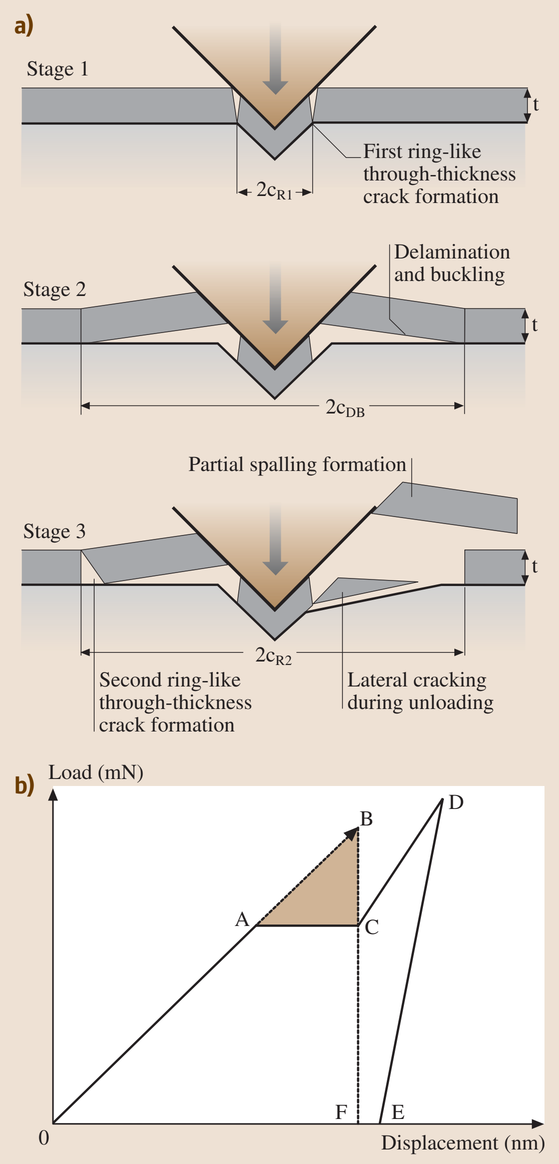
a)
Stage 1
t
First ring-like through-thickness crack formation
2c_R1
Delamination and buckling
Stage 2
t
2c_DB
Partial spalling formation
Stage 3
t
2c_R2
Second ring-like through-thickness crack formation
Lateral cracking during unloading
b)
Load (mN)
D
B
A
C
F
E
0
Displacement (nm)

Fig. 26.12 **(a)** Schematic of various stages in nanoindentation fracture for the coatings/substrate system, and **(b)** schematic of a load-displacement curve showing a step during the loading cycle and associated energy release

for the coating used in the present study, the advances of the indenter during the radial cracking, delamination, and buckling are not big enough to form steps in the loading curve, because the coating around the indenter still supports the indenter, but generate discontinuities that change the slope of the loading curve with increasing indentation loads. In the third stage, the stress concentration at the end of the interfacial crack cannot be relaxed by the propagation of the interfacial crack. With an increase in indentation depth, the height of the bulged coating increases. When the height reaches a critical value, the bending stresses caused by the bulged coating around the indenter will result in the second ring-like through-thickness crack formation and spalling at the edge of the buckled coating, as shown in Fig. 26.12a, which leads to a step in the loading curve. This is a single event and results in the separation of the part of the coating around the indenter from the bulk

Table 26.5 Hardness, elastic modulus, fracture toughness, fatigue life, critical load during scratch, coefficient of friction during accelerated wear testing and residual stresses of various DLC coatings on single-crystal silicon substrate

Coating	Hardness[a] [26.48] (GPa)	Elastic modulus[a] [26.48] (GPa)	Fracture toughness[a] [26.89] (MPa $m^{1/2}$)	Fatigue life[b], N_f[d] [26.90] $\times 10^4$	Critical load during scratch[b] [26.48] (mN)	Coefficient of friction during accelerated wear testing[b] [26.48]	Compressive residual stress[c] [26.47] (GPa)
Cathodic arc carbon coating (a-C)	24	280	11.8	2.0	3.8	0.18	12.5
Ion beam carbon coating (a-C:H)	19	140	4.3	0.8	2.3	0.18	1.5
ECR-CVD carbon coating (a-C:H)	22	180	6.4	1.2	5.3	0.22	0.6
DC sputtered carbon coating (a-C:H)	15	140	2.8	0.2	1.1	0.32	2.0
Bulk graphite (for comparison)	Very soft	9–15	–	–	–	–	–
Diamond (for comparison)	80–104	900–1050	–	–	–	–	–
Si(100) substrate	11	220	0.75	–	0.6	0.55	0.02

[a] Measured on 100-nm-thick coatings
[b] Measured on 20-nm-thick coatings
[c] Measured on 400-nm-thick coatings
[d] N_f was obtained at a mean load of 10 μN and a load amplitude of 8 μN

coating via cracking through coatings. The step in the loading curve results totally from the coating cracking and not from the interfacial cracking or the substrate cracking.

The area under the load-displacement curve is the work performed by the indenter during elastic-plastic deformation of the coating/substrate system. The strain energy release in the first/second ring-like cracking and spalling can be calculated from the corresponding steps in the loading curve. Fig. 26.12b shows a modeled load-displacement curve. OACD is the loading curve and DE is the unloading curve. The first ring-like through-thickness crack should be considered. It should be emphasized that the edge of the buckled coating is far from the indenter and, therefore, it does not matter if the indentation depth exceeds the coating thickness, or if deformation of the substrate occurs around the indenter when we measure fracture toughness of the coating from the released energy during the second ring-like through-thickness cracking (spalling). Suppose that the second ring-like through-thickness cracking occurs at AC. Now, let us consider the loading curve OAC. If the second ring-like through-thickness crack does not occur, it can be understood that OA will be extended to OB to reach the same displacement as OC. This means that the crack formation changes the loading curve OAB into OAC. For point B, the elastic-plastic energy stored in the coating/substrate system should be OBF. For point C, the elastic-plastic energy stored in the coating/substrate system should be OACF. Therefore, the energy difference before and after the crack generation is the area of ABC, i. e., this energy stored in ABC will be released as strain energy to create the ring-like through-thickness crack. According to the theoretical analysis by *Li* et al. [26.87], the fracture toughness of thin films can be written as

$$K_{\mathrm{Ic}} = \left[\left(\frac{E}{(1-\nu^2)\, 2\pi C_{\mathrm{R}}} \right) \left(\frac{U}{t} \right) \right]^{1/2} , \tag{26.2}$$

where E is the elastic modulus, ν is the Poisson's ratio, $2\pi C_{\mathrm{R}}$ is the crack length in the coating plane, t is the coating thickness, and U is the strain energy difference before and after cracking.

Using (26.2), the fracture toughness of the coatings is calculated. The loading curve is extrapolated along the tangential direction of the loading curve from the starting point of the step up to reach the same displacement as the step. The area between the extrapolated line and the step is the estimated strain energy difference before and after cracking. C_{R} is measured from SEM micrographs or AFM images of indentations. The second ring-like crack is where the spalling occurs. For example, for the 400-nm-thick cathodic arc carbon coating data presented in Fig. 26.11, U of 7.1 nNm is assessed from the steps in Fig. 26.11a at the peak indentation loads of 200 mN. For C_{R} of 7.0 µm from Fig. 26.11b, $E = 300$ GPa measured using a nanoindenter and an assumed value of 0.25 for ν, fracture toughness values are calculated as 10.9 MPa$\sqrt{m}$ [26.87, 88]. The fracture toughness and related data for various 100-nm-thick DLC coatings are presented in Fig. 26.10 and Table 26.5.

Nanofatigue

Delayed fracture resulting from extended service is called fatigue [26.96]. Fatigue fracturing progresses through a material via changes within the material at the tip of a crack, where there is a high stress intensity. There are several situations: cyclic fatigue, stress corrosion, and static fatigue. Cyclic fatigue results from cyclic loading of machine components. In a low-flying slider in a magnetic head-disk interface, isolated asperity contacts occur during use and the fatigue failure occurs in the multilayered thin-film structure of the magnetic disk [26.13]. In many MEMS components, impact occurs and the failure mode is cyclic fatigue. Asperity contacts can be simulated using a sharp diamond tip in an oscillating contact with the component.

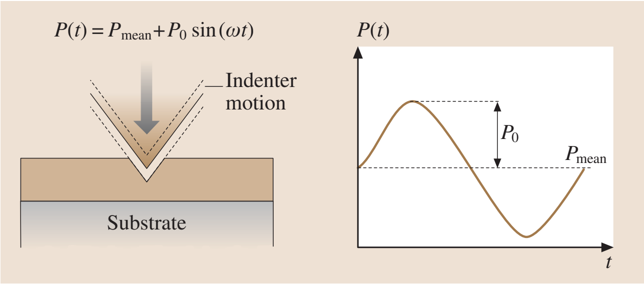

Fig. 26.13 Schematic of a fatigue test on a coating/substrate system using the continuous stiffness measurement technique

Figure 26.13 shows the schematic of a fatigue test on a coating/substrate system using a continuous stiffness measurement (CSM) technique. Load cycles are applied to the coating, resulting in a cyclic stress. P is the cyclic load, P_{mean} is the mean load, P_{o} is the oscillation load amplitude, and ω is the oscillation frequency. The following results can be obtained: (1) endurance limit, i. e., the maximum load below which there is no coating failure for a preset number of cycles; (2) number of cycles at which the coating failure occurs; and (3) changes in contact stiffness measured by using the unloading slope of each cycle, which can be used to monitor the propagation of the interfacial cracks during cyclic fatigue process.

Figure 26.14a shows the contact stiffness as a function of the number of cycles for 20-nm-thick FCA coatings cyclically deformed by various oscillation load amplitudes with a mean load of 10 µN at a frequency of 45 Hz. At 4 µN load amplitude, no change in contact stiffness was found for all coatings. This indicates that 4 µN load amplitude is not high enough to damage the coatings. At 6 µN load amplitude, an abrupt decrease in contact stiffness was found at a certain number of cycles for each coating, indicating that fatigue damage had occurred. With increasing load amplitude, the number of cycles to failure, $\mathrm{N_f}$, decreases for all coatings. Load amplitude versus $\mathrm{N_f}$, a so-called S-N curve, is plotted in Fig. 26.14b. The critical load amplitude, below which no fatigue damage occurs (an endurance limit), was identified for each coating. This critical load amplitude, together with mean load, are of critical importance to the design of head-disk interfaces or MEMS/NEMS device interfaces.

To compare the fatigue lives of different coatings studied, the contact stiffness as a function of the number of cycles for 20-nm-thick FCA, IB, ECR-CVD and SP coatings cyclically deformed by an oscillation load amplitude of 8 µN with a mean load of 10 µN at a frequency of 45 Hz is shown in Fig. 26.14c. FCA coating has the longest $\mathrm{N_f}$, followed by ECR-CVD, IB, and SP coatings. In addition, after the $\mathrm{N_f}$, the contact stiffness of the FCA coating shows a slower decrease than the other coatings. This indicates that after the $\mathrm{N_f}$, the FCA coating had less damage than the others. The fatigue behavior of FCA and ECR-CVD coatings of differ-

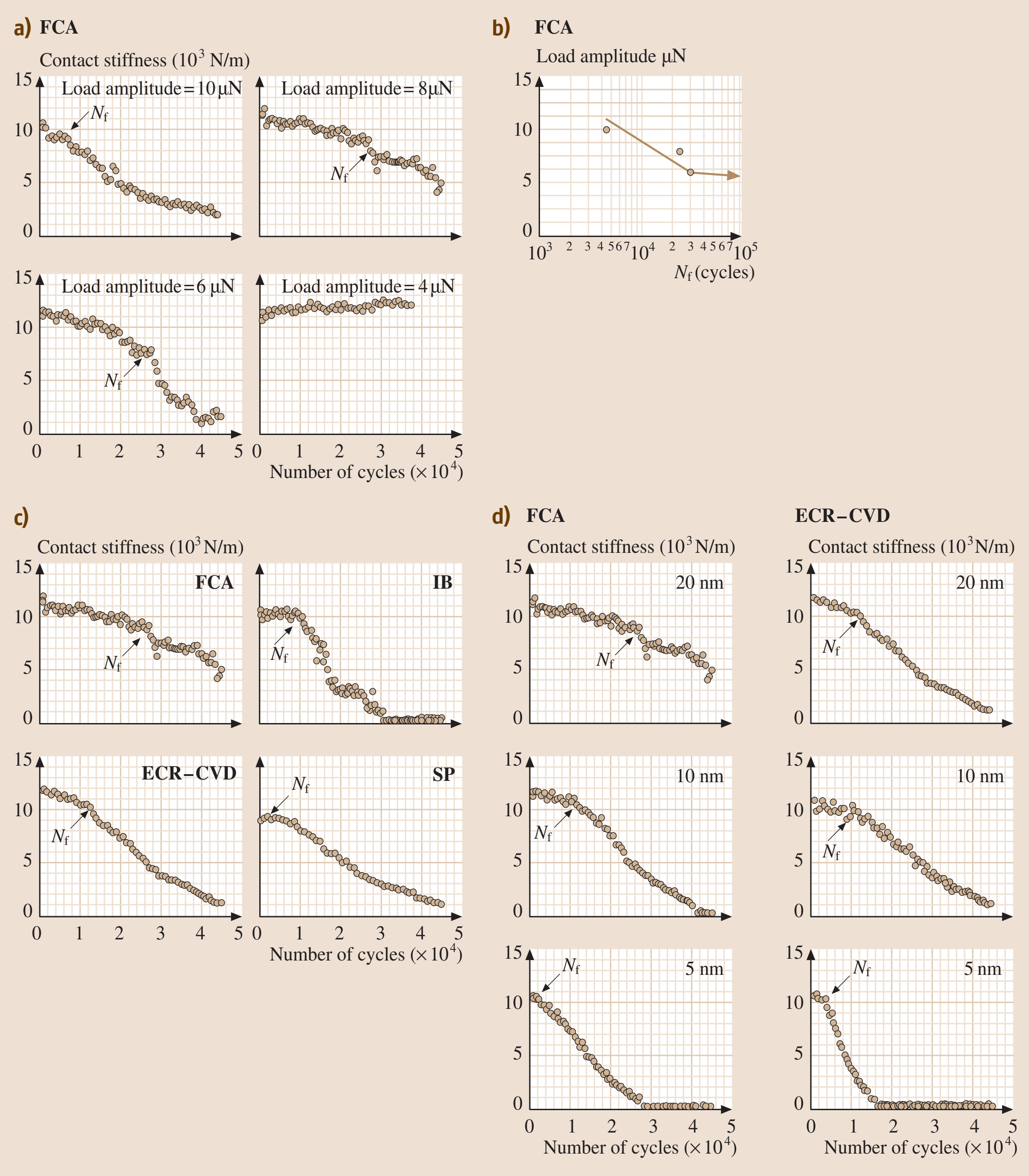

Fig. 26.14 **(a)** Contact stiffness as a function of the number of cycles for 20-nm-thick FCA coatings cyclically deformed by various oscillation load amplitudes with a mean load of 10 µN at a frequency of 45 Hz; **(b)** plot of load amplitude versus N_f; **(c)** contact stiffness as a function of the number of cycles for four different 20-nm-thick coatings with a mean load of 10 µN and load amplitude of 8 µN at a frequency of 45 Hz; and **(d)** contact stiffness as a function of the number of cycles for two coatings of different thicknesses at a mean load of 10 µN and load amplitude of 8 µN at a frequency of 45 Hz

ent thicknesses is compared in Fig. 26.14d. For both coatings, N_f decreases with decreasing coating thickness. At 10 nm, FCA and ECR-CVD have almost the same fatigue life. At 5 nm, ECR-CVD coating shows a slightly longer fatigue life than FCA coating. This indicates that even for nanometer-thick DLC coatings their microstructure and residual stresses are not uniform across the thickness direction. Thinner coatings are more influenced by interfacial stresses than thicker coating.

Figure 26.15a shows the high magnification SEM images of 20-nm-thick FCA coatings before, at, and after N_f. In the SEM images, the net-like structure is the gold film coated on the DLC coating, which should be ignored in analyzing the indentation fatigue damage. Before the N_f, no delamination or buckling was found except the residual indentation mark at magnifications up to 1,200,000× using SEM. This suggests that only plastic deformation occurred before the N_f. At the N_f, the coating around the indenter bulged upwards, indicating delamination and buckling. Therefore, it is believed that the decrease in contact stiffness at the N_f results from the delamination and buckling of the coating from the substrate. After the N_f, the buckled coating was broken down around the edge of the buckled area, forming a ring-like crack. The remaining coating overhung at the edge of the buckled area. It is noted that the indentation size increases with the increasing number of cycles. This indicates that deformation, delamination and buckling, and ring-like crack formation occurred over a period.

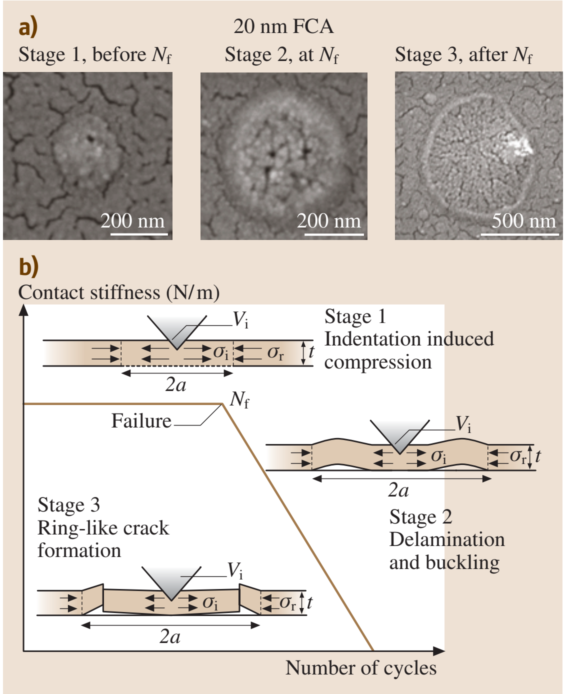

Fig. 26.15 **(a)** High magnification SEM images of a coating before, at, and after N_f, and **(b)** schematic of various stages in the indentation fatigue damage for a coating/substrate system [26.90]

The schematic in Fig. 26.15b shows various stages in the indentation fatigue damage for a coating/substrate system. Based on this study, three stages in the indentation fatigue damage appear to exist: (1) indentation-induced compression; (2) delamination and buckling; (3) ring-like crack formation at the edge of the buckled coating. The deposition process often induces residual stresses in coatings. The model shown in Fig. 26.15b considers a coating with the uniform biaxial residual compression σ_r. In the first stage, indentation induces elastic/plastic deformation, exerting an outward acting pressure on the coating around the indenter. Interfacial defects like voids and impurities act as original cracks. These cracks propagate and link up as the indentation compressive stress increases. At this stage, the coating, which is under the indenter and above the interfacial crack (with a crack length of $2a$), still maintains a solid contact with the substrate; the substrate still fully supports the coating. Therefore, this interfacial crack does not lead to an abrupt decrease in contact stiffness, but gives rise to a slight decrease in contact stiffness, as shown in Fig. 26.14. The coating above the interfacial crack is treated as a rigidly clamped disk. We assume that the crack radius, a, is large compared with the coating thickness t. Since the coating thickness ranges from 5 to 20 nm, this assumption is easily satisfied in this study (radius of the delaminated and buckled area, shown in Fig. 26.15a, is on the order of 100 nm). The compressive stress caused by indentation is given as [26.97]:

$$\sigma_i = \frac{E}{(1-\nu)}\varepsilon_i = \frac{E V_i}{2\pi t a^2 (1-\nu)} , \qquad (26.3)$$

where ν and E are the Poisson's ratio and elastic modulus of the coating, V_i is the indentation volume, t is the coating thickness, and a is the crack radius. With an increasing number of cycles, the indentation volume V_i increases. Therefore, the indentation compressive stress σ_i increases accordingly. In the second stage, buckling occurs during the unloading segment of the fatigue testing cycle when the sum of indentation compressive stress σ_i and the residual stress σ_r

exceed the critical buckling stress σ_b for the delaminated circular section as given by [26.98]

$$\sigma_b = \frac{\mu^2 E}{12\left(1-\nu^2\right)}\left(\frac{t}{a}\right)^2 , \tag{26.4}$$

where the constant μ equals 42.67 for a circular clamped plate with a constrained center point and 14.68 when the center is unconstrained. The buckled coating acts as a cantilever. In this case, the indenter indents a cantilever rather than a coating/substrate system. This ultrathin coating cantilever has much less contact stiffness than the coating/substrate system. Therefore, the contact stiffness shows an abrupt decrease at the N_f. In the third stage, with an increasing number of cycles, the delaminated and buckled size increases, resulting in a further decrease in contact stiffness since the cantilever beam length increases. On the other hand, a high bending stress acts at the edge of the buckled coating. The larger the buckled size, the higher the bending stress. The cyclically bending stress causes fatigue damage at the end of the buckled coating, forming a ring-like crack. The coating under the indenter is separated from the bulk coating (caused by the ring-like crack at the edge of the buckled coating) and the substrate (caused by the delamination and buckling in the second stage). Therefore, the coating under the indenter is not constrained, but is free to move with the indenter during fatigue testing. At this point, the sharp nature of the indenter is lost, because the coating under the indenter gets stuck on the indenter. The indentation fatigue experiment results in the contact of a relative huge blunt tip with the substrate. This results in a low contact stiffness value.

Compressive residual stresses result in delamination and buckling. A coating with a higher adhesion strength and a less compressive residual stress is required for a higher fatigue life. Interfacial defects should be avoided in the coating deposition process. We know that the ring-like crack formation occurs in the coating. Formation of fatigue cracks in the coating depends upon the hardness and fracture toughness. Cracks are more difficult to form and propagate in the coating with higher strength and fracture toughness.

It is now accepted that long fatigue life in a coating/substrate almost always involves "living with crack", that the threshold or limit condition is associated with the non-propagation of exiting cracks or defects, even though these cracks may be undetectable [26.96]. For all coatings studied, at 4 μN, contact stiffness does not change much. This indicates that delamination and buckling did not occur within the number of cycles tested in this study. This is probably because the indentation-induced compressive stress was not high enough to allow the cracks to propagate and link up under the indenter, or the sum of indentation compressive stress σ_i and the residual stress σ_r did not exceed the critical buckling stress σ_b.

Figure 26.10 and Table 26.5 summarize the hardness, elastic modulus, fracture toughness, and fatigue life of all coatings studied. A good correlation exists between fatigue life and other mechanical properties. Higher mechanical properties result in a longer fatigue life. The mechanical properties of DLC coatings are controlled by the sp^3-to-sp^2 ratio. The sp^3-bonded carbon exhibits the outstanding properties of diamond [26.51]. A higher deposition kinetic energy will result in a larger fraction of sp^3-bonded carbon in an amorphous network. Thus, the higher kinetic energy for the FCA could be responsible for its better carbon structure and higher mechanical properties [26.48–50, 99]. Higher adhesion strength between the FCA coating and substrate makes the FCA coating more difficult to delaminate from the substrate.

26.3.2 Microscratch and Microwear Studies

For microscratch studies, a conical diamond indenter (e.g., having a tip radius of about one micron and included angle of 60°) is drawn over the sample surface, and the load is ramped up (typically from 2 mN to 25 mN) until substantial damage occurs. The coefficient of friction is monitored during scratching. Scratch-induced damage of coating, specifically fracture or delamination, can be monitored by in situ friction force measurements and by optical and SEM imaging of the scratches after tests. A gradual increase in friction is associated with plowing, and an abrupt increase in friction is associated with fracture or catastrophic failure [26.100]. The load, corresponding to the abrupt increase in friction or an increase in friction above a certain value (typically 2× the initial value), provides a measure of scratch resistance or adhesive strength of a coating and is called "critical load". The depth of scratches with increasing scratch length or normal load is measured using an AFM, typically with an area of 10×10 μm [26.48, 49, 101].

The microscratch and microwear studies are also conducted using an AFM [26.23, 50, 99, 102, 103]. A square pyramidal diamond tip (tip radius ~ 100 nm) or a three-sided pyramidal diamond (Berkovich) tip with an apex angle of 60° and a tip radius of about 100 nm mounted on a platinum-coated, rectangular stainless steel cantilever of stiffness of about 40 N/m is scanned

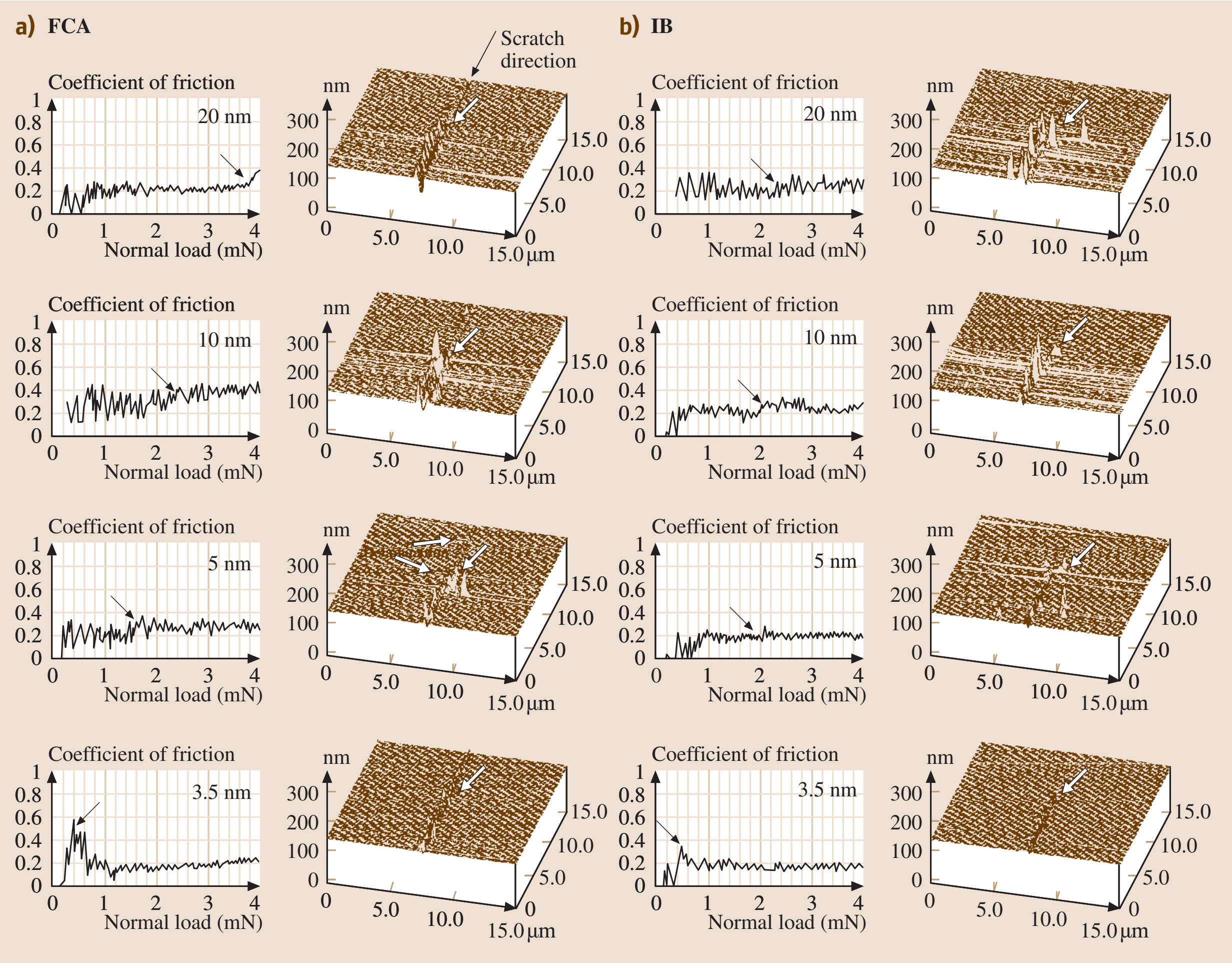

Fig. 26.16a,b

orthogonal to the long axis of the cantilever to generate scratch and wear marks. During the scratch test, the normal load is either kept constant or increased (typically from 0 to 100 µN) until damage occurs. Topography images of the scratch are obtained in situ with the AFM at a low load. By scanning the sample during scratching, wear experiments can be conducted. Wear at a constant load is monitored as a function of the number of cycles. Normal loads ranging from 10–80 µN are typically used.

Microscratch

Scratch tests conducted with a sharp diamond tip simulate a sharp asperity contact. In a scratch test, the cracking or delamination of a hard coating is signaled by a sudden increase in the coefficient of friction [26.23]. The load associated with this event is called the "critical load".

Wu [26.104], *Bhushan* et al. [26.70], *Gupta* and *Bhushan* [26.12,47], and *Li* and *Bhushan* [26.48,49,101] have used a nanoindenter to perform microscratch studies (mechanical durability) of various carbon coatings. The coefficient of friction profiles as a function of increasing normal load and AFM surface height maps of regions over scratches at the respective critical loads (indicated by the arrows in the friction profiles and AFM images) made on the various coatings of different thicknesses and single-crystal silicon substrate using a conical tip are compared in Figs. 26.16 and 26.17. *Bhushan* and *Koinkar* [26.102], *Koinkar* and

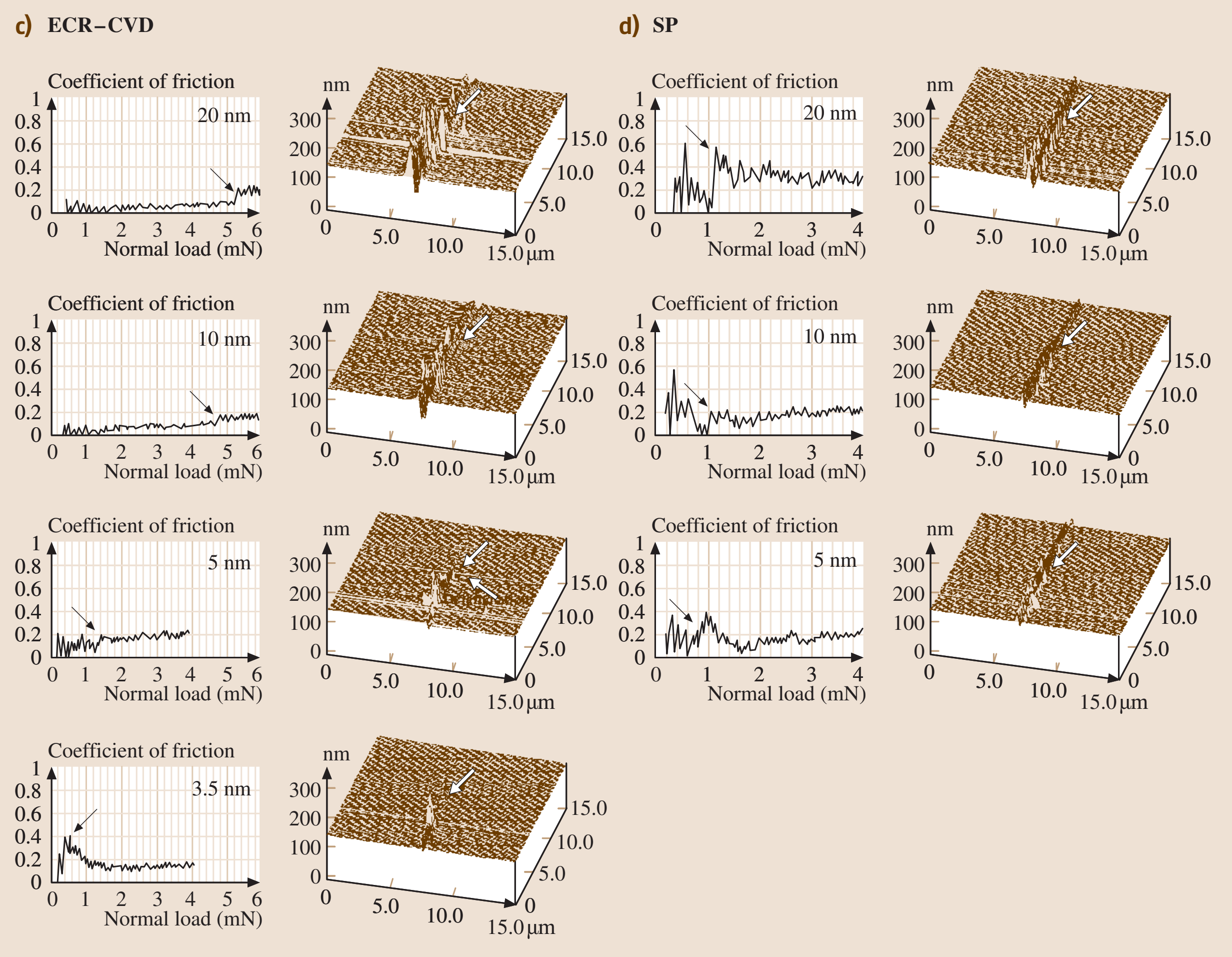

Fig. 26.16a–d Coefficient of friction profiles as a function of normal load and corresponding AFM surface height maps of regions over scratches at the respective critical loads (indicated by the arrows in the friction profiles and AFM images) for coatings of different thicknesses deposited by various deposition techniques: **(a)** FCA, **(b)** IB, **(c)** ECR-CVD, **(d)** SP

Bhushan [26.103], *Bhushan* [26.23], and *Sundararajan* and *Bhushan* [26.50, 99] have used an AFM to perform microscratch studies. Data for various coatings of different thicknesses and silicon substrate using a Berkovich tip are compared in Figs. 26.18 and 26.19. Critical loads for various coatings tested using a nanoindenter and AFM are summarized in Fig. 26.20. The selected data for 20-nm-thick coatings obtained using nanoindenter are also presented in Fig. 26.10 and Table 26.5.

It can be seen that a well-defined critical load exists for each coating. The AFM images clearly show that below the critical loads the coatings were plowed by the scratch tip, associated with the plastic flow of materials.

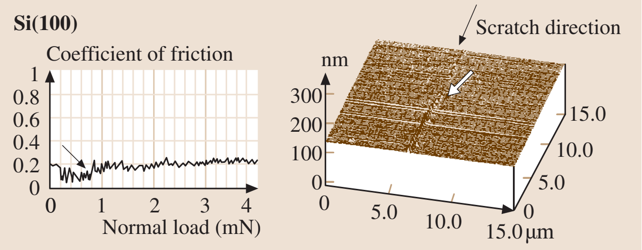

Fig. 26.17 Coefficient of friction profiles as a function of normal load and corresponding AFM surface height maps of regions over scratches at the respective critical loads (indicated by the arrows in the friction profiles and AFM images) for Si(100)

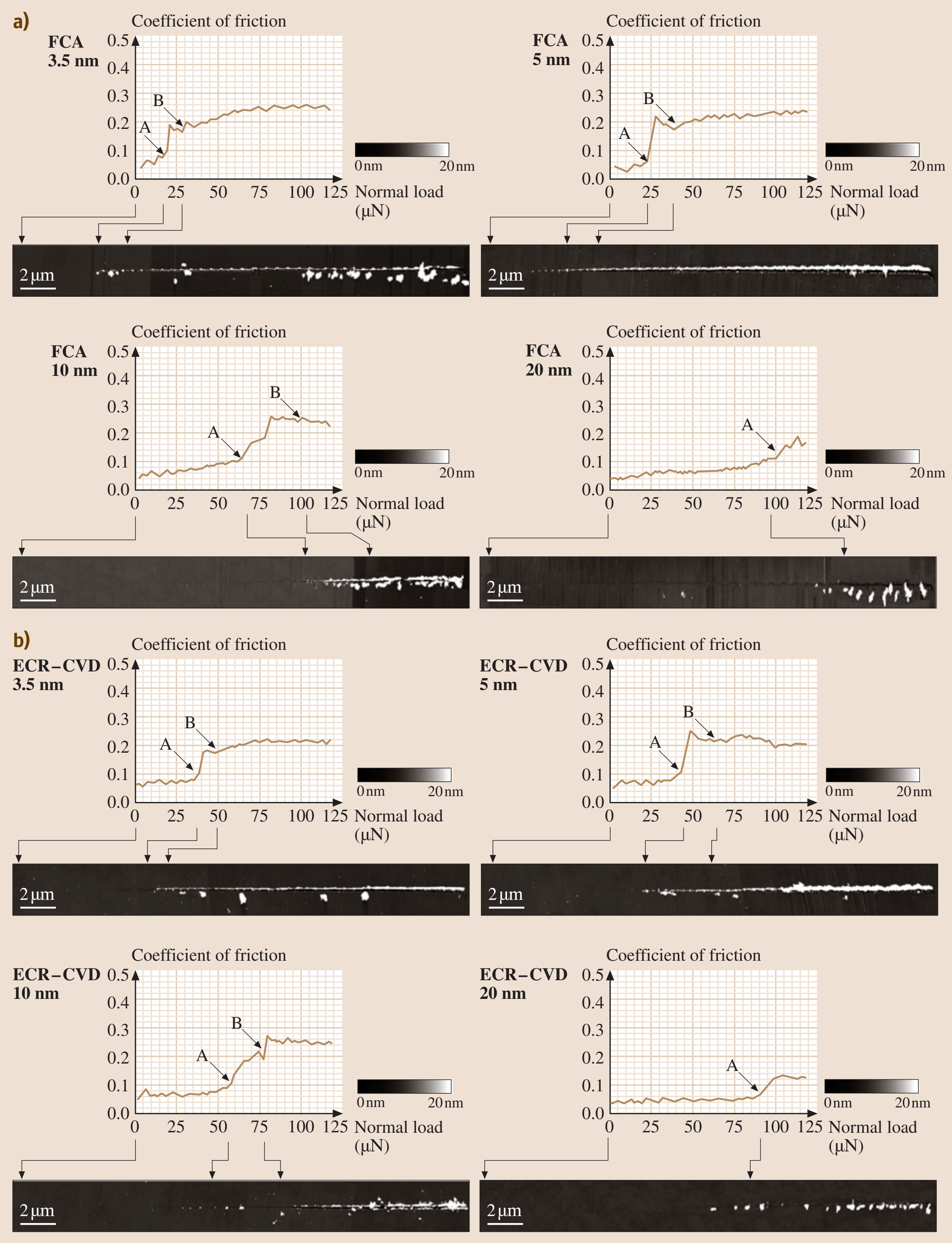
a)
FCA 3.5 nm
FCA 5 nm
FCA 10 nm
FCA 20 nm
b)
ECR–CVD 3.5 nm
ECR–CVD 5 nm
ECR–CVD 10 nm
ECR–CVD 20 nm
Coefficient of friction
0.0
0.1
0.2
0.3
0.4
0.5
A
B
0
25
50
75
100
125
Normal load (μN)
0nm
20nm
2 μm

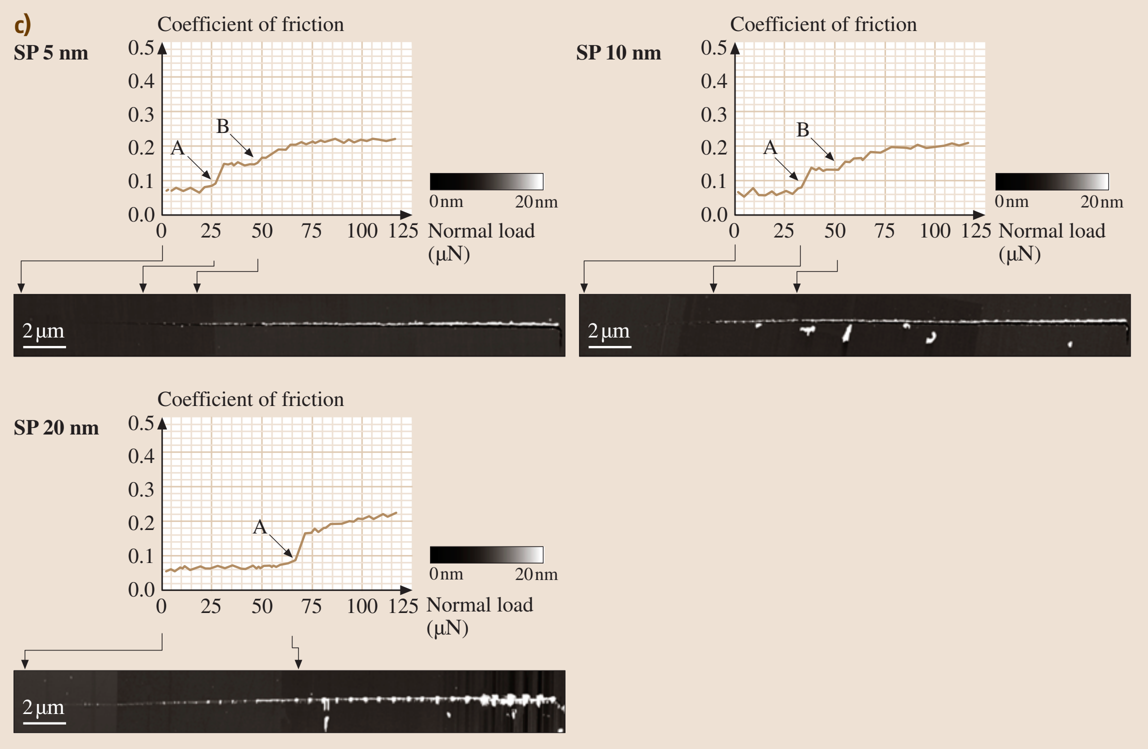

Fig. 26.18a–c Coefficient of friction profiles during scratch as a function of normal load and corresponding AFM surface height maps for **(a)** FCA, **(b)** ECR-CVD, and **(c)** SP coatings [26.99]

At and after the critical loads, debris (chips) or buckling was observed on the sides of scratches. Delamination or buckling can be observed around or after the critical loads, which suggests that the damage starts from delamination and buckling. For the 3.5- and 5-nm-thick FCA coatings, before the critical loads small debris is observed on the sides of scratches. This suggests that the thinner FCA coatings are not so durable. It is obvious that for a given deposition method, the critical loads increase with increasing coating thickness. This indicates that the critical load is determined not only by the adhesion strength to the substrate, but also by the coating thickness. We note that the debris generated on the thicker coatings is larger than that generated on the thinner coatings. For a thicker coating, it is more difficult to be broken; the broken coating chips (debris) for a thicker coating are larger than those for the thinner coatings. The difference in the residual stresses of the coatings of different thicknesses could also affect the size of debris. The AFM image shows that the silicon substrate was damaged by plowing, associated with the plastic flow of materials. At and after the critical load, small and uniform debris is observed and the amount of debris increases with increasing normal load.

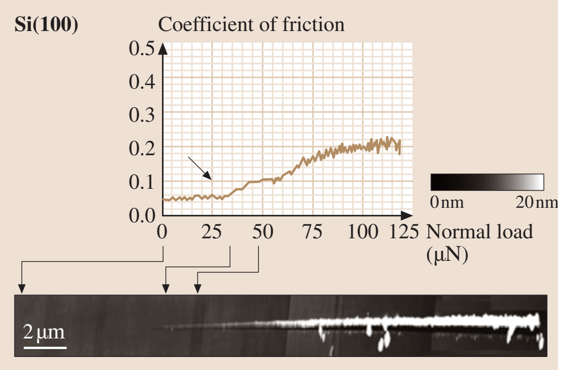

Fig. 26.19 Coefficient of friction profiles during scratch as a function of normal load and corresponding AFM surface height maps for Si(100) [26.99]

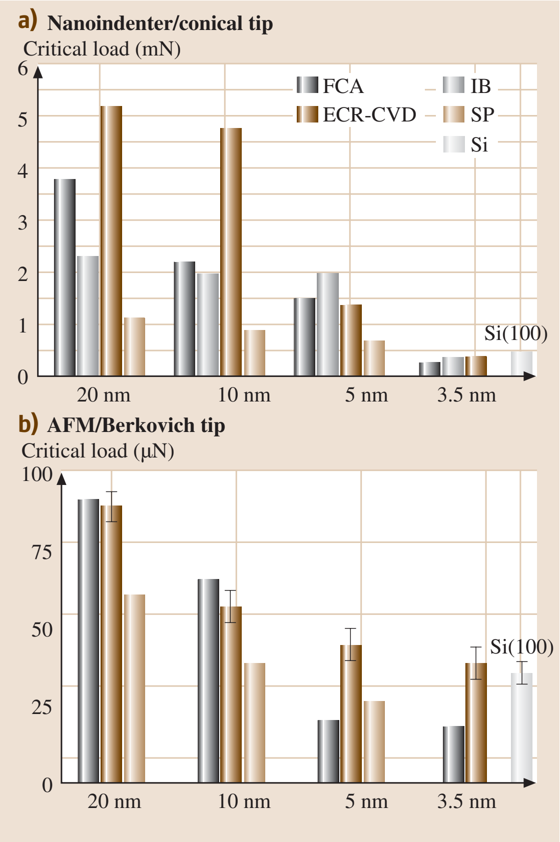

Fig. 26.20a,b Critical loads estimated from the coefficient of friction profiles from **(a)** nanoindenter and **(b)** AFM tests for various coatings of different thicknesses and Si(100) substrate

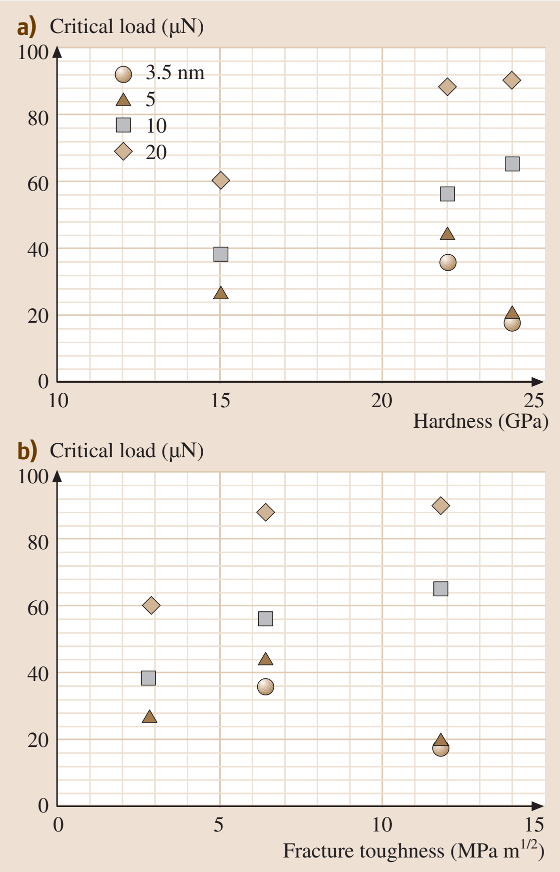

Fig. 26.21a,b Measured critical loads estimated from the coefficient of friction profiles from AFM tests as a function of **(a)** coating hardness and **(b)** fracture toughness. Coating hardness and fracture toughness values were obtained using a nanoindenter on 100-nm-thick coatings (Table 26.5)

Since at the critical load, the damage mechanism appears to be the onset of plowing, higher hardness and fracture toughness of a coating will therefore result in a higher load required for deformation and hence a higher critical load. Figure 26.21 shows critical loads of the various coatings, obtained with AFM tests, as a function of the coating hardness and fracture toughness (from Table 26.5). It can be seen that, in general, higher coating hardness and fracture toughness result in a higher critical load. The only exceptions are the FCA coatings at 5- and 3.5-nm coating thickness, which show the lowest critical loads despite their high hardness and fracture toughness. The brittleness of the thinner FCA coatings may be one reason for their low critical loads. The mechanical properties of coatings that are less than 10 nm thick are unknown. The FCA process may result in coatings with low hardness at such low thickness due to differences in coating stoichiometry and structure compared to the coatings of higher thickness. Also, at these thicknesses stresses at the coating-substrate interface may affect adhesion and load-carrying capacity of the coatings.

Based on the experimental results, a schematic of scratch damage mechanisms of the DLC coatings used in this study is shown in Fig. 26.22. Below the critical load, if a coating has a good combination of strength and fracture toughness, plowing associated with the plastic flow of materials is responsible for the damage of coating (Fig. 26.22a). Whereas if a coating has a lower fracture toughness, cracking could occur during plowing, asso-

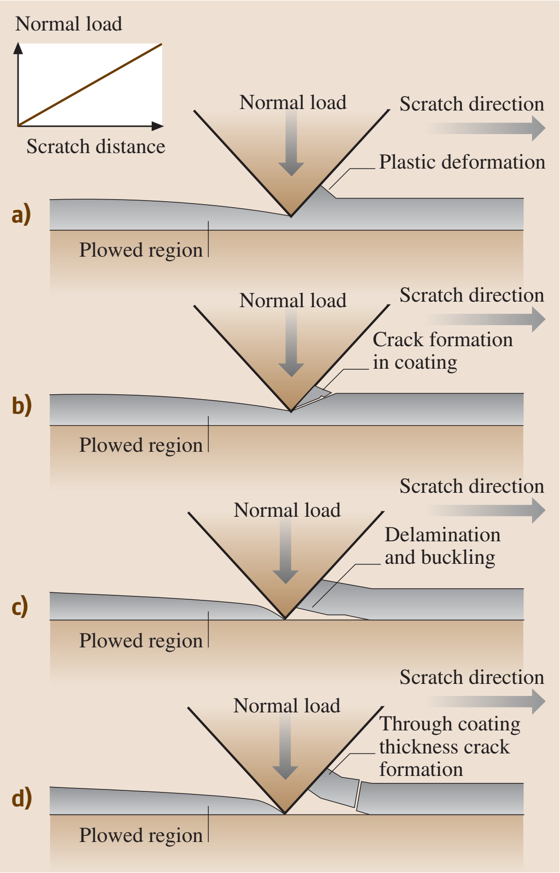

Fig. 26.22a–d Schematic of scratch damage mechanisms of the DLC coatings: **(a)** plowing associated with the plastic flow of materials, **(b)** plowing associated with the formation of small debris, **(c)** delamination and buckling at the critical load, and **(d)** breakdown via through coating thickness cracking at and after the critical load [26.48]

ciated with the formation of small debris (Fig. 26.22b). When normal load increases up to the critical load, delamination or buckling will occur at the coating/substrate interface (Fig. 26.22c). A further increase in normal load will result in the breakdown of coating via through coating thickness cracking, as shown in Fig. 26.22d. Therefore, adhesion strength plays a crucial role in the determination of critical load. If a coating has stronger adhesive strength with the substrate, the coating is more difficult to delaminate, which will result in a higher critical load. The interfacial and residual stresses of a coating could greatly affect the delamination and buckling [26.1]. The coating with higher interfacial and residual stresses is more easily delaminated and buckled, which will result in a low critical load. It has been reported earlier that the FCA coatings have higher residual stresses than the other coatings [26.47]. Interfacial stresses play a more important role when a coating gets thinner. A large mismatch in elastic modulus between the FCA coatings and the silicon substrate may cause large interfacial stresses. This may be why the thinner FCA coatings show relatively low critical loads compared with the thicker FCA coatings, even though the FCA coatings have a higher hardness and elastic modulus. The brittleness of the thinner FCA coatings may be another reason for the lower critical loads. The strength and fracture toughness of a coating also affect the critical load. Higher strength and fracture toughness will make the coating more difficult to be broken after delamination and buckling. The high scratch resistance/adhesion of FCA coatings is attributed to an atomic intermixing at the coating-substrate interface because of high kinetic energy (2 keV) plasma formed during the cathodic arc deposition process [26.57]. The atomic intermixing at the interface provides a graded compositional transition between the coating and the substrate materials. In all other coatings used in this study, the kinetic energy of the plasma was insufficient for atomic intermixing.

Gupta and *Bhushan* [26.12, 47] and *Li* and *Bhushan* [26.48,49] measured scratch resistance of DLC coatings deposited on Al_2O_3-TiC, Ni-Zn ferrite, and single-crystal silicon substrates. For good adhesion of DLC coating to other substrates, in most cases, an interlayer of silicon is required except for cathodic arc deposited coatings. The best adhesion with cathodic arc carbon coating is obtained on electrically conducting substrates such as Al_2O_3-TiC and silicon, as compared to Ni-Zn ferrite.

Microwear

Microwear studies can be conducted using an AFM [26.23]. For microwear studies, a three-sided pyramidal single-crystal natural diamond tip with an apex angle of about 80° and a tip radius of about 100 nm is used at relatively high loads of 1–150 μN. The diamond tip is mounted on a stainless steel cantilever beam with a normal stiffness of about 30 N/m. The sample is generally scanned in a direction orthogonal to the long axis of the cantilever beam (typically at a rate of 0.5 Hz). The tip is mounted on the beam such that one of its edges is orthogonal to the beam axis. In wear studies, typically an area of 2 μm × 2 μm is scanned for a selected number of cycles.

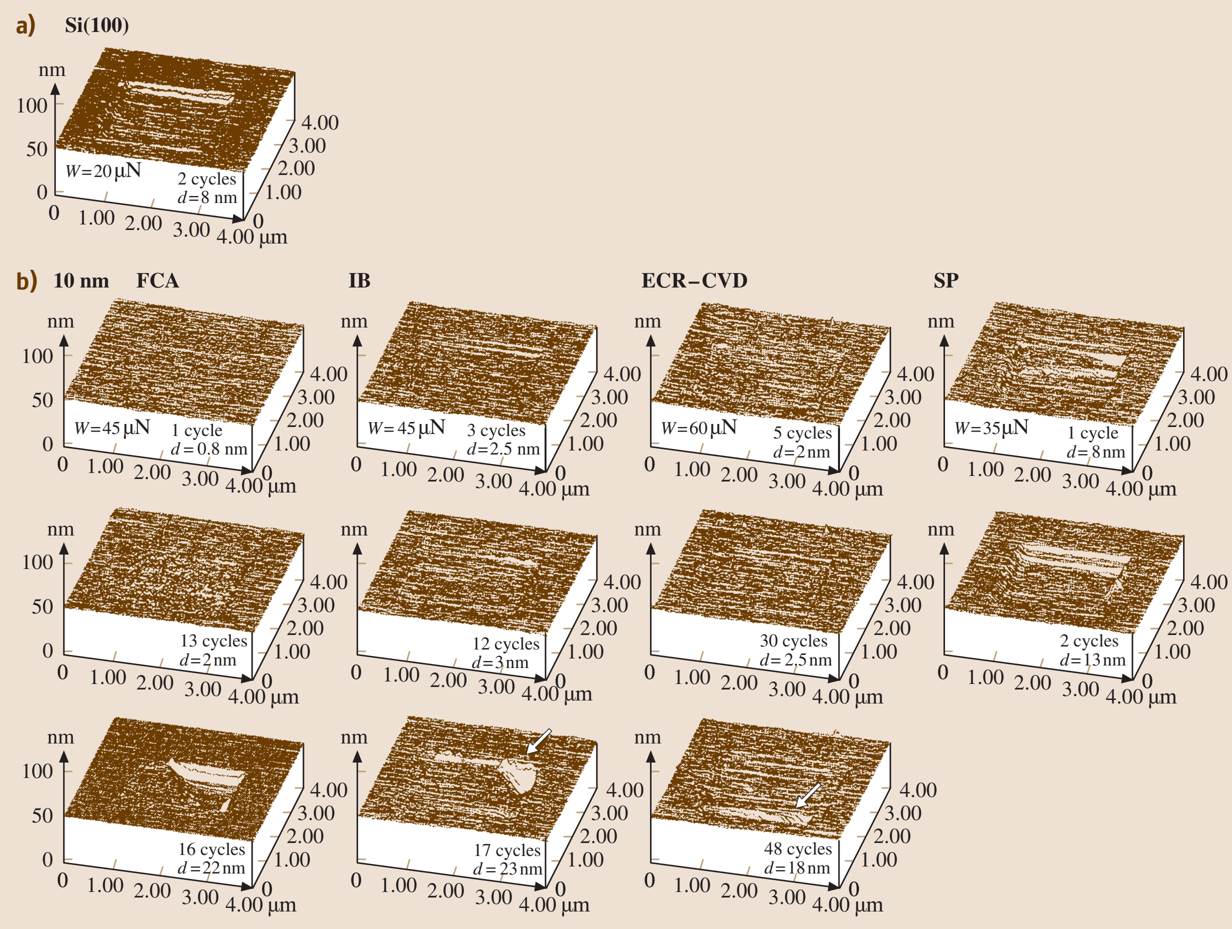
a)
Si(100)
nm
100
50
0
W=20 µN
2 cycles
d=8 nm
0
1.00
2.00
3.00
4.00 µm
b)
10 nm
FCA
IB
ECR–CVD
SP
W=45 µN
1 cycle
d=0.8 nm
W=45 µN
3 cycles
d=2.5 nm
W=60 µN
5 cycles
d=2 nm
W=35µN
1 cycle
d=8 nm
13 cycles
d=2 nm
12 cycles
d=3 nm
30 cycles
d=2.5 nm
2 cycles
d=13 nm
16 cycles
d=22 nm
17 cycles
d=23 nm
48 cycles
d=18 nm

Fig. 26.23a,b AFM images of wear marks on **(a)** bare Si(100), and **(b)** all 10-nm-thick DLC coatings [26.50]

Microwear studies of various types of DLC coatings have been conducted [26.50, 102, 103]. Fig. 26.23a shows a wear mark on uncoated Si(100). Wear occurs uniformly and material is removed layer by layer via plowing from the first cycle, resulting in the constant friction force seen during the wear (Fig. 26.24a). Figure 26.23b shows AFM images of the wear marks on all 10 nm coatings. It is seen that coatings wear nonuniformly. Coating failure is sudden and accompanied by a sudden rise in the friction force (Fig. 26.24b). Figure 26.24 shows the wear depth of Si(100) substrate and various DLC coatings at two different loads. FCA and ECR-CVD, 20-nm-thick coatings show excellent wear resistance up to 80 µN, the load that is required for the IB 20 nm coating to fail. In these tests, "failure" of a coating results when the wear depth exceeds the quoted coating thickness. The SP 20 nm coating fails at the much lower load of 35 µN. At 60 µN, the coating hardly provides any protection. Moving on to the 10 nm coatings, ECR-CVD coating requires about 45 cycles at 60 µN to fail as compared to IB and FCA, which fail at 45 µN. The FCA coating exhibits slight roughening in the wear track after the first few cycles, which leads to an increase in the friction force. The SP coating continues to exhibit poor resistance, failing at 20 µN. For the 5 nm coatings, the load required to fail the coatings continues to decrease. But IB and ECR-CVD still provide adequate protection as compared to bare Si(100) in that order, failing at 35 µN compared to FCA at 25 µN and SP at 20 µN. Almost all the 20, 10, and 5 nm coatings provide better wear resistance than bare silicon. At 3.5 nm, FCA coating

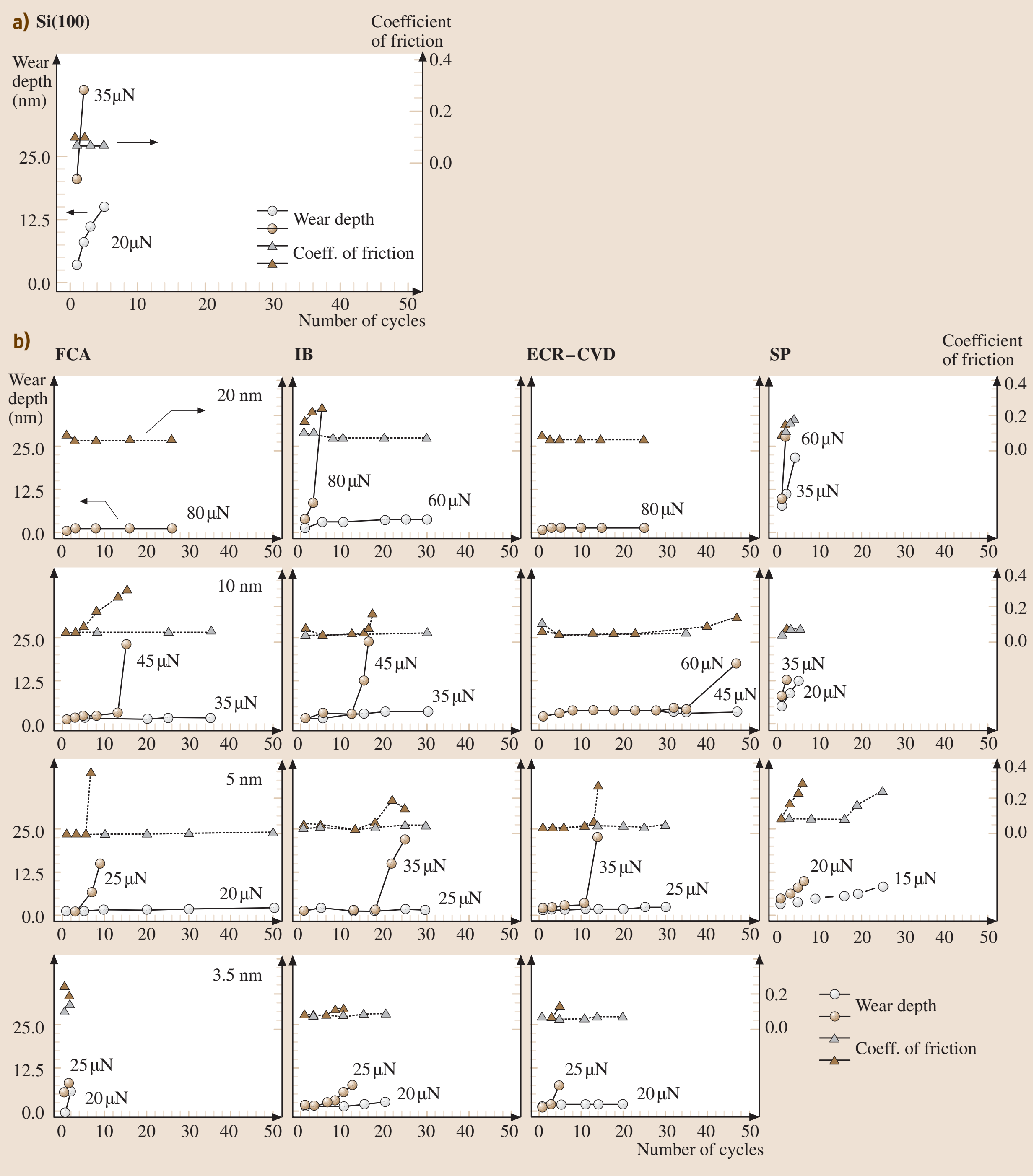

Fig. 26.24a,b Wear data of **(a)** bare Si(100) and **(b)** all DLC coatings. Coating thickness is constant along each row in **(b)**. Both wear depth and coefficient of friction during wear for a given cycle are plotted [26.50]

provides no wear resistance, failing almost instantly at 20 μN. The IB and ECR-CVD coating show good wear resistance at 20 μN compared to bare Si(100). But IB lasts only about 10 cycles and ECR-CVD about 3 cycles at 25 μN.

The wear tests highlight the differences in the coatings more vividly than the scratch tests. At higher thicknesses (10 and 20 nm), the ECR-CVD and FCA coatings appear to show the best wear resistance. This is probably due to higher hardness of the coatings (see Table 26.5). At 5 nm, IB coating appears to be the best. FCA coatings show poorer wear resistance with decreasing coating thickness. This suggests that the trends in hardness seen in Table 26.5 no longer hold at low thicknesses. SP coatings showed consistently poor wear resistance at all thicknesses. The IB 3.5 nm coating does provide reasonable wear protection at low loads.

26.3.3 Macroscale Tribological Characterization

So far, data on mechanical characterization and microscratch and microwear studies using a nanoindenter and an AFM have been presented. Mechanical properties affect tribological performance of the coatings, and microwear studies simulate a single asperity contact, which helps in developing a fundamental understanding of the wear process. These studies are useful in screening various candidates, as well as understanding the relationships between deposition conditions and properties of various samples. As a next step, macroscale friction and wear tests need to be conducted to measure tribological performance of the coatings.

Macroscale accelerated friction and wear tests to screen a large number of candidates and functional tests on selected candidates have been conducted. An accelerated test is designed to accelerate the wear process such that it does not change the failure mechanism. The accelerated friction and wear tests are generally conducted using a ball-on-flat tribometer under reciprocating motion [26.70]. Typically, a diamond tip with a 20 μm tip radius or a sapphire ball with a 3 mm diameter and surface finish of about 2 nm RMS is slid against the coated substrates at selected loads. Coefficient of friction is monitored during the tests.

Functional tests are conducted using an actual machine under close-to-actual operating conditions for which coatings are developed. Generally, the tests are accelerated somewhat to fail the interface in a short time.

Accelerated Friction and Wear Tests

Li and *Bhushan* [26.48] conducted accelerated friction and wear tests on DLC coatings deposited by various deposition techniques using a ball-on-flat tribometer.

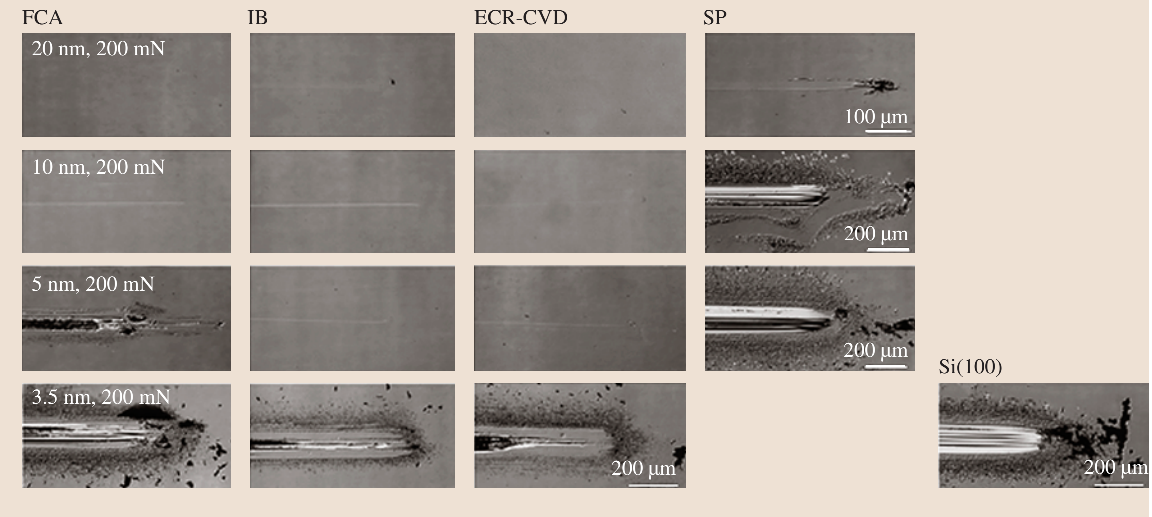

Fig. 26.25 Optical micrographs of wear tracks and debris formed on the various coatings of different thicknesses and silicon substrate when slid against a sapphire ball after sliding distance of 5 m. The end of the wear track is on the right-hand side of the image

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

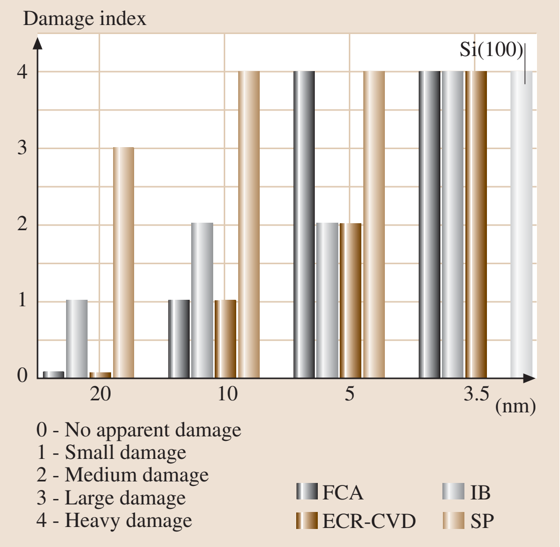

Fig. 26.26 Wear damage index bar chart of the various coatings of different thicknesses and Si(100) substrate based on optical examination of the wear tracks and debris

Average values of coefficient of friction are presented in Table 26.5. The optical micrographs of wear tracks and debris formed on all samples when slid against a sapphire ball after a sliding distance of 5 m are presented in Fig. 26.25. The normal load used for the 20- and 10-nm-thick coatings is 200 mN, and the normal load used for the 5- and 3.5-nm-thick coatings and silicon substrate is 150 mN.

Among the 20-nm-thick coatings, the SP coating exhibits a higher coefficient of friction of about 0.3 than the other coatings, which have comparable values of coefficient of friction of about 0.2. The optical micrographs show that the SP coating has a larger wear track and more debris than the IB coatings. No wear track and debris were found on the 20-nm-thick FCA and ECR-CVD coatings. The optical micrographs of 10-nm-thick coatings show that the SP coating was severely damaged, showing a large wear track with scratches and lots of debris. The FCA and ECR-CVD coatings show smaller wear tracks and less debris than the IB coatings.

For the 5-nm-thick coatings, the wear tracks and debris of the IB and ECR-CVD coatings are comparable. The bad wear resistance of the 5-nm-thick FCA coating is in good agreement with the low scratch critical load, which may be due to the higher interfacial and residual stresses as well as brittleness of the coating.

At 3.5 nm, all coatings exhibit wear. The FCA coating provides no wear resistance, failing instantly like the silicon substrate. Large block-like debris is observed on the sides of the wear track of the FCA coating. This indicates that large region delamination and buckling occurred during sliding, resulting in large block-like debris. This block-like debris, in turn, scratched the coating, making the coating damage even more severe. The IB and ECR-CVD coatings are able to provide some protection against wear at 3.5 nm.

In order to better evaluate the wear resistance of various coatings, based on the optical examination of the wear tracks and debris after tests, the wear damage index bar chart of the various coatings of different thicknesses and an uncoated silicon substrate is presented in Fig. 26.26. Among the 20- and 10-nm-thick coatings, the SP coatings show the worst damage, followed by FCA/ECR-CVD. At 5 nm, the FCA and SP coatings show the worst damage, followed by IB and ECR-CVD coatings. All 3.5-nm-thick coatings show the same heavy damage as the uncoated silicon substrate.

The wear damage mechanisms of thick and thin DLC coatings studied are believed to be as illustrated in Fig. 26.27. At the early stages of sliding, deformation zone and Hertzian and wear fatigue cracks formed beneath the surface extend within the coating on subsequent sliding [26.1]. Formation of fatigue cracks depend on the hardness and subsequent cycles. These are controlled by the sp^3-to-sp^2 ratio. For thicker coating, the cracks generally do not penetrate the coating. For a thinner coating, the cracks easily propagate down to the interface with the aid of the interfacial stresses and get diverted along the interface just enough to cause local delamination of the coating. When this happens, the coating experiences excessive plowing. At this point, the coating fails catastrophically, resulting in a sudden rise

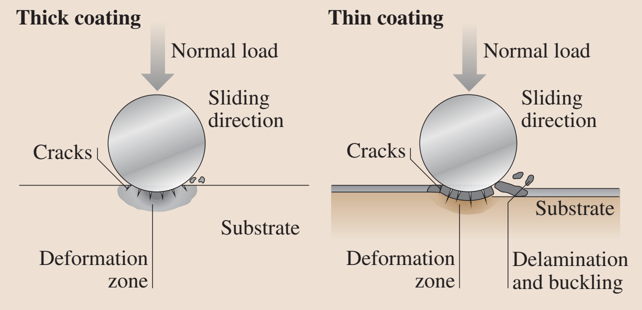

Fig. 26.27 Schematic of wear damage mechanisms of thick and thin DLC coatings [26.48]

in the coefficient of friction. All 3.5-nm-thick coatings failed much quicker than the thicker coatings. It appears that these thin coatings have very low load-carrying capacity and, therefore, the substrate undergoes deformation almost immediately. This generates stresses at the interface that weaken the coating adhesion and lead to delamination of the coating. Another reason may be that the thickness is insufficient to produce a coating comprised of the DLC structure. Instead, the bulk may be made up of a matrix characteristic of the interface region where atomic mixing occurs with the substrate and/or any interlayer used. This would also result in poor wear resistance and silicon-like behavior of the coating, especially in the case of FCA coatings, which show the worst performance at 3.5 nm. In contrast to the other coatings, all SP coatings show the worst wear performance at any thickness (Fig. 26.25). This may be due to their poor mechanical properties, such as lower hardness and scratch resistance, compared to the other coatings.

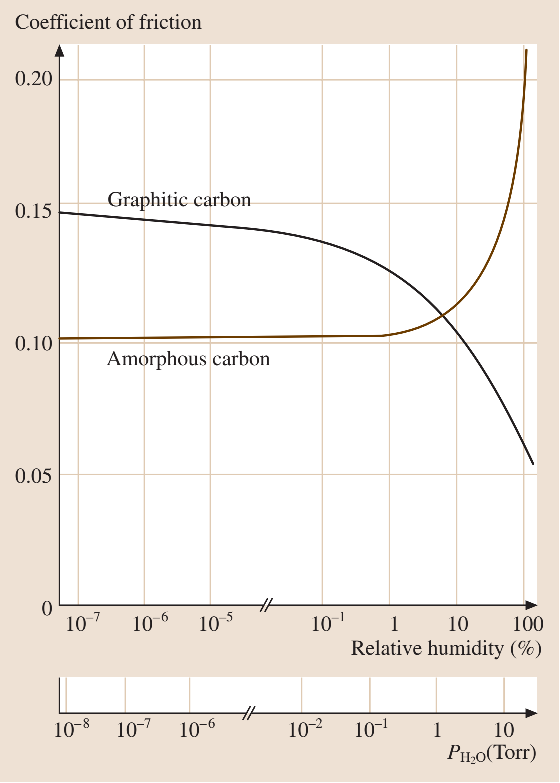

Fig. 26.28 Coefficient of friction as a function of relative humidity and water vapor partial pressure of RF-plasma deposited amorphous carbon coating and bulk graphitic carbon coating sliding against a steel ball

Comparison of Figs. 26.20 and 26.26 shows a very good correlation between the wear damage and scratch critical loads. Less wear damage corresponds to higher scratch critical load. Based on the data, thicker coatings do show better scratch and wear resistance than thinner coatings. This is probably due to better load-carrying capacity of the thick coatings compared to the thinner ones. For a given coating thickness, higher hardness and fracture toughness and better adhesion strength are believed to be responsible for the superior wear performance.

Effect of Environment

Friction and wear performance of amorphous carbon coatings are found to be strongly dependent on the water vapor content and partial gas pressure in the test environment. The friction data for an amorphous carbon film on the silicon substrate sliding against steel are presented as a function of the partial pressure of water vapor in Fig. 26.28 [26.1, 13, 69, 105, 106]. Friction increases dramatically above a relative humidity of about 40%. At high relative humidity, condensed water vapor forms meniscus bridges at the contacting asperities, and the meniscii result in an intrinsic attractive force that is responsible for an increase in friction. For completeness, the data for the coefficient of friction of bulk graphitic carbon are also presented in Fig. 26.28. Note that friction decreases with an increase in the relative humidity [26.107]. Graphitic carbon has a layered lattice crystal structure. Graphite absorbs po-

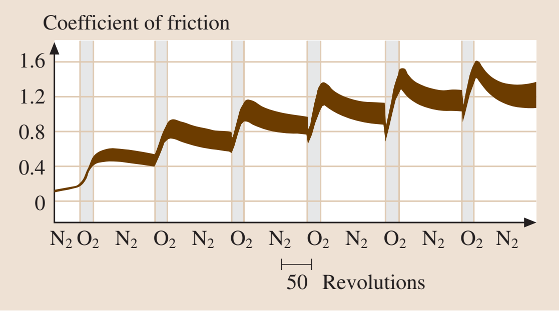

Fig. 26.29 Coefficient of friction as a function of sliding distance for a ceramic slider against a magnetic disk coated with 20-nm-thick DC magnetron sputtered DLC coating, measured at a speed of 0.06 m/s and 10 g load. The environment is alternated between oxygen and nitrogen gases [26.108]

lar gases (H_2O, O_2, CO_2, NH_3, etc.) at the edges of the crystallites, which weakens the interlayer bonding forces facilitating interlayer slip and results in lower friction [26.1].

To better study the effect of environment for carbon coated magnetic disks, a number of tests have been conducted in controlled environments. *Marchon* et al. [26.108] conducted tests in alternating environments of oxygen and nitrogen gases, Fig. 26.29. The coefficient of friction increases as soon as oxygen is added to the test environment, whereas in a nitrogen environment the coefficient of friction reduces slightly. Tribochemical oxidation of the DLC coating in the oxidizing environment is responsible for an increase in the coefficient of friction, implying wear. *Dugger* et al. [26.109], *Strom* et al. [26.110], *Bhushan* and *Ruan* [26.111], and *Bhushan* et al. [26.71] conducted tests with DLC coated magnetic disks (with about 2-nm-thick perfluoropolyether lubricant film) in contact with Al_2O_3-TiC sliders, in different gaseous environments including a high vacuum of 2×10^{-7} Torr, Fig. 26.30. The wear lives are shortest in high vacuum and longest in most atmospheres of nitrogen and argon with the following order (from best to worst): argon or nitrogen, $Ar+H_2O$, ambient, $Ar+O_2$, and $Ar+H_2O$, vacuum. From this sequence of wear performance, we can see that having oxygen and water in an operating environment worsens wear performance of the coatings, but having nothing in it (vacuum) is the worst of all. Indeed, failure mechanisms differ in various environments. In high vacuum, intimate contact between disk and slider surfaces results in significant wear. In ambient air, $Ar+O_2$, and $Ar+H_2O$, tribochemical oxidation of the carbon overcoat is responsible for interface failure. For experiments performed in pure argon and nitrogen, mechanical shearing of the asperities cause the formation of debris, which is responsible for the formation of scratch marks on the carbon surface as could be observed with an optical microscope [26.71].

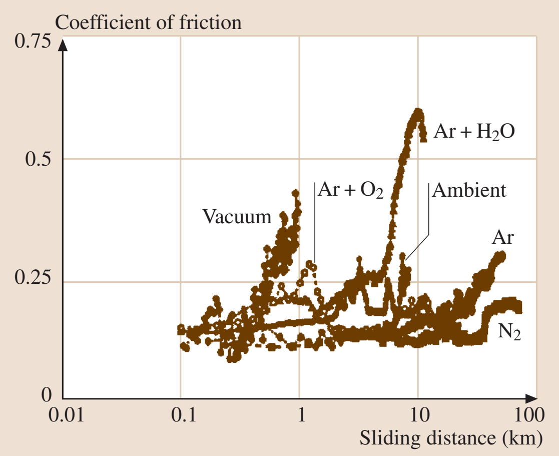

Fig. 26.30 Durability measured by sliding a Al_2O_3-TiC magnetic slider against a magnetic disk coated with 20-nm-thick DC sputtered amorphous carbon coating and 2-nm-thick perfluoropolyether film, measured at a speed of 0.75 m/s and 10 g load. Vacuum refers to 2×10^{-7} Torr [26.71]

Functional Tests

Magnetic thin-film heads made with Al_2O_3-TiC substrate are used in magnetic storage applications [26.13]. A multilayered thin-film pole-tip structure present on the head surface wears more rapidly than the much harder Al_2O_3-TiC substrate. Pole-tip recession (PTR) is a serious concern in magnetic storage [26.15–19, 112]. Two of the diamond-like carbon coatings superior in mechanical properties – ion beam and cathodic arc carbon – were deposited on the air bearing surfaces of Al_2O_3-TiC head sliders [26.15]. The functional tests were conducted by running a metal-particle (MP) tape in a computer tape drive. Average PTR as a function of sliding distance data are presented in Fig. 26.31. We note that PTR increases for the uncoated head, whereas for the coated heads there

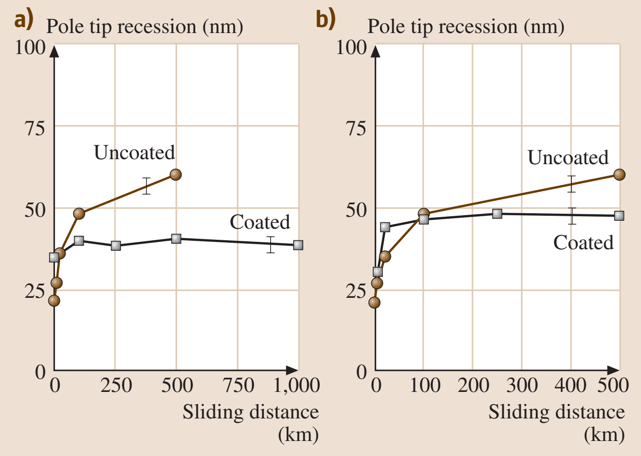

Fig. 26.31a,b Pole tip recession as a function of sliding distance measured with an AFM for **(a)** uncoated and 20-nm-thick ion beam carbon coated, and **(b)** uncoated and 20-nm-thick cathodic arc carbon coated Al_2O_3-TiC heads run against MP tapes [26.15]

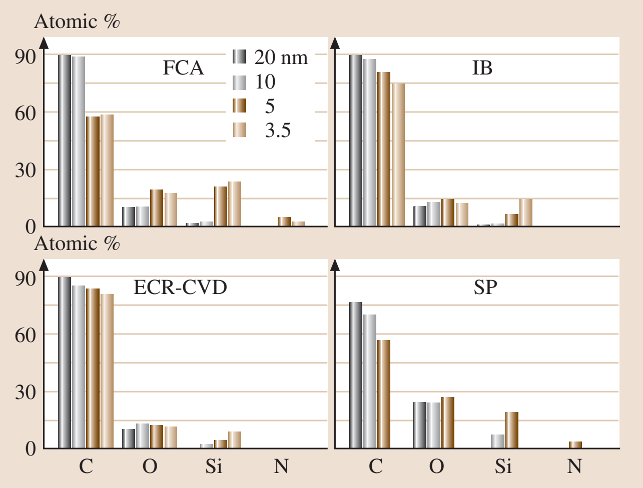

Fig. 26.32 Quantified XPS data for various DLC coatings on Si(100) substrate [26.50]. Atomic concentrations are shown

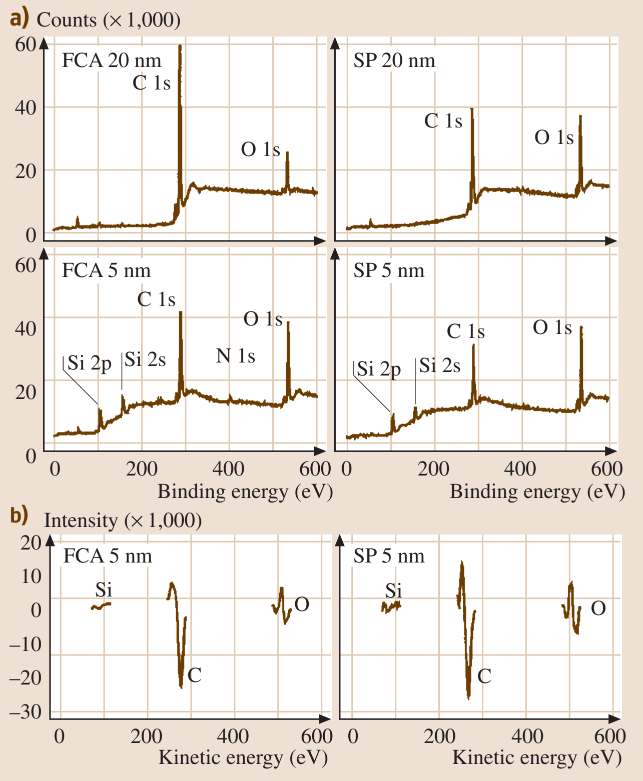

Fig. 26.33 (a) XPS spectra for FCA and SP coatings of 5 nm and 20 nm coating thicknesses on Si(100) substrate, and **(b)** AES spectra for FCA and SP coatings at 5 nm thickness on Si(100) substrate [26.50]

is a slight increase in PTR in early sliding followed by little change. Thus, coatings provide protection.

Micromechanical and accelerated and functional tribological data presented here clearly suggest that there is a good correlation between the scratch resistance and wear resistance measured using accelerated tests and functional tests. Thus, scratch tests can be successfully used to screen coatings for wear applications.

26.3.4 Coating Continuity Analysis

Ultrathin coatings less than 10 nm may not uniformly coat the sample surface. In other words, coating may be discontinuous and deposited in the form of islands on the microscale. A possible reason for poor wear protection and the nonuniform failure of the coatings may be due to poor coverage of the thin coatings on the substrate. Coating continuity can be studied by using surface analytical techniques such as Auger and/or XPS analyses. Any discontinuity with coating thickness less than the sampling depth of the instrument will detect locally the substrate species [26.49, 50, 102].

The results of XPS analysis on various coatings deposited on Si(100) substrates, over a 1.3 mm^2 region (single point measurement with spot diameter of 1,300 μm) are shown in Fig. 26.32. The sampling depth is about 2–3 nm. The poor SP coatings show much less carbon content ($< 75\%$ atomic concentration) as does the poor 5 nm and 3.5 nm FCA coatings ($< 60\%$) compared to the IB and ECR-CVD coatings. Silicon is detected in all 5 nm coatings. From the data it is hard to say if the Si is from the substrate or from exposed regions due to coating discontinuity. Based on the sampling depth any Si detected in 3.5 nm coatings would likely be from the substrate. The other interesting observation is that all poor coatings (all SP and FCA 5 and 3.5 nm) show almost twice the oxygen content than that of the other coatings. Any oxygen present may be due to leaks in the deposition chamber and is present in the form of silicon oxides.

AES measurements averaged over a scan area of 900 μm^2 were conducted on FCA and SP 5 nm coatings at six different regions on each sample. Very little silicon was detected on this scale, and the detected peaks were characteristic of oxides. The oxygen levels were comparable to that seen for the good coatings by XPS. These contrast the XPS measurements at a larger scale, suggesting that the coating possesses discontinuities only at isolated areas and that the 5 nm coatings are generally continuous on the microscale. Figure 26.33 shows representative XPS and AES spectra of selected samples.

References

26.1 B. Bhushan, B.K. Gupta: *Handbook of Tribology: Materials, Coatings, and Surface Treatments*, reprint edn. (Krieger, Malabar 1997)

26.2 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

26.3 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

26.4 B. Bhushan, B.K. Gupta, G.W. VanCleef, C. Capp, J.V. Coe: Fullerene (C_{60}) films for solid lubrication, Tribol. Trans. **36** (1993) 573–580

26.5 B.K. Gupta, B. Bhushan, C. Capp, J.V. Coe: Material characterization and effect of purity and ion implantation on the friction and wear of sublimed fullerene films, J. Mater. Res. **9** (1994) 2823–2838

26.6 B.K. Gupta, B. Bhushan: Fullerene particles as an additive to liquid lubricants and greases for low friction and wear, Lubr. Eng. **50** (1994) 524–528

26.7 B. Bhushan, V.V. Subramaniam, A. Malshe, B.K. Gupta, J. Ruan: Tribological properties of polished diamond films, J. Appl. Phys. **74** (1993) 4174–4180

26.8 B. Bhushan, B.K. Gupta, V.V. Subramaniam: Polishing of diamond films, Diam. Films Technol. **4** (1994) 71–97

26.9 P. Sander, U. Kaiser, M. Altebockwinkel, L. Wiedmann, A. Benninghoven, R.E. Sah, P. Koidl: Depth profile analysis of hydrogenated carbon layers on silicon by X-ray photoelectron spectroscopy, auger electron spectroscopy, electron energy-loss spectroscopy, and secondary ion mass spectrometry, J. Vac. Sci. Technol. A **5** (1987) 1470–1473

26.10 A. Matthews, S.S. Eskildsen: Engineering applications for diamond-like carbon, Diam. Relat. Mater. **3** (1994) 902–911

26.11 A.H. Lettington: Applications of diamond-like carbon thin films, Carbon **36** (1998) 555–560

26.12 B.K. Gupta, B. Bhushan: Mechanical and tribological properties of hard carbon coatings for magnetic recording heads, Wear **190** (1995) 110–122

26.13 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)

26.14 B. Bhushan: *Mechanics and Reliability of Flexible Magnetic Media*, 2nd edn. (Springer, New York 2000)

26.15 B. Bhushan, S.T. Patton, R. Sundaram, S. Dey: Pole tip recession studies of hard carbon-coated thin-film tape heads, J. Appl. Phys. **79** (1996) 5916–5918

26.16 J. Xu, B. Bhushan: Pole tip recession studies of thin-film rigid disk head sliders II: Effects of air bearing surface and pole tip region designs and carbon coating, Wear **219** (1998) 30–41

26.17 W.W. Scott, B. Bhushan: Corrosion and wear studies of uncoated and ultra-thin DLC coated magnetic tape-write heads and magnetic tapes, Wear **243** (2000) 31–42

26.18 W.W. Scott, B. Bhushan: Loose debris and head stain generation and pole tip recession in modern tape drives, J. Info. Storage Proc. Syst. **2** (2000) 221–254

26.19 W.W. Scott, B. Bhushan, A.V. Lakshmikumaran: Ultrathin diamond-like carbon coatings used for reduction of pole tip recession in magnetic tape heads, J. Appl Phys **87** (2000) 6182–6184

26.20 B. Bhushan: Macro- and microtribology of magnetic storage devices. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 1413–1513

26.21 B. Bhushan: Nanotribology and nanomechanics of MEMS devices, Proc. Ninth Annual Workshop on Micro Electro Mechanical Systems 1996 (IEEE, New York 91–98

26.22 B. Bhushan (Ed.): *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)

26.23 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

26.24 B. Bhushan: Macro- and microtribology of MEMS materials. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 1515–1548

26.25 S. Aisenberg, R. Chabot: Ion beam deposition of thin films of diamond like carbon, J. Appl. Phys. **49** (1971) 2953–2958

26.26 E.G. Spencer, P.H. Schmidt, D.C. Joy, F.J. Sansalone: Ion beam deposited polycrystalline diamond-like films, Appl. Phys. Lett. **29** (1976) 118–120

26.27 A. Grill, B.S. Meyerson: Development and status of diamondlike carbon. In: *Synthetic Diamond: Emerging CVD Science and Technology*, ed. by K.E. Spear, J.P. Dismukes (Wiley, New York 1994) pp. 91–141

26.28 Y. Catherine: Preparation techniques for diamond-like carbon. In: *Diamond and Diamond-like Films and Coatings*, ed. by R.E. Clausing, L.L. Horton, J.C. Angus, P. Koidl (Plenum, New York 1991) pp. 193–227

26.29 J.J. Cuomo, D.L. Pappas, J. Bruley, J.P. Doyle, K.L. Seagner: Vapor deposition processes for amorphous carbon films with sp^3 fractions approaching diamond, J. Appl. Phys. **70** (1991) 1706–1711

26.30 J.C. Angus, C.C. Hayman: Low pressure metastable growth of diamond and diamondlike phase, Science **241** (1988) 913–921

26.31 J.C. Angus, F. Jensen: Dense diamondlike hydrocarbons as random covalent networks, J. Vac. Sci. Technol. A **6** (1988) 1778–1782

26.32 D.C. Green, D.R. McKenzie, P.B. Lukins: The microstructure of carbon thin films, Mater. Sci. Forum **52–53** (1989) 103–124

26.33 N.H. Cho, K.M. Krishnan, D.K. Veirs, M.D. Rubin, C.B. Hopper, B. Bhushan, D.B. Bogy: Chemical structure and physical properties of diamond-like amorphous carbon films prepared by magnetron sputtering, J. Mater. Res. **5** (1990) 2543–2554

26.34 J. C. Angus: Diamond and diamondlike films, Thin Solid Films **216** (1992) 126–133

26.35 B. Bhushan, A. J. Kellock, N. H. Cho, J. W. Ager III: Characterization of chemical bonding and physical characteristic of diamond-like amorphous carbon and diamond films, J. Mater. Res. **7** (1992) 404–410

26.36 J. Robertson: Properties of diamond-like carbon, Surf. Coat. Technol. **50** (1992) 185–203

26.37 N. Savvides, B. Window: Diamondlike amorphous carbon films prepared by magnetron sputtering of graphite, J. Vac. Sci. Technol. A **3** (1985) 2386–2390

26.38 J. C. Angus, P. Koidl, S. Domitz: Carbon thin films. In: *Plasma Deposited Thin Films*, ed. by J. Mort, F. Jensen (CRC, Boca Raton 1986) pp. 89–127

26.39 J. Robertson: Amorphous carbon, Adv. Phys. **35** (1986) 317–374

26.40 M. Rubin, C. B. Hooper, N. H. Cho, B. Bhushan: Optical and mechanical properties of DC sputtered carbon films, J. Mater. Res. **5** (1990) 2538–2542

26.41 G. J. Vandentop, M. Kawasaki, R. M. Nix, I. G. Brown, M. Salmeron, G. A. Somorjai: Formation of hydrogenated amorphous carbon films of controlled hardness from a methane plasma, Phys. Rev. B **41** (1990) 3200–3210

26.42 J. J. Cuomo, D. L. Pappas, R. Lossy, J. P. Doyle, J. Bruley, G. W. Di Bello, W. Krakow: Energetic carbon deposition at oblique angles, J. Vac. Sci. Technol. A **10** (1992) 3414–3418

26.43 D. L. Pappas, K. L. Saegner, J. Bruley, W. Krakow, J. J. Cuomo: Pulsed laser deposition of diamondlike carbon films, J. Appl. Phys. **71** (1992) 5675–5684

26.44 H. J. Scheibe, B. Schultrich: DLC film deposition by laser-arc and study of properties, Thin Solid Films **246** (1994) 92–102

26.45 C. Donnet, A. Grill: Friction control of diamond-like carbon coatings, Surf. Coat. Technol. **94-95** (1997) 456

26.46 A. Grill: Tribological properties of diamondlike carbon and related materials, Surf. Coat. Technol. **94–95** (1997) 507

26.47 B. K. Gupta, B. Bhushan: Micromechanical properties of amorphous carbon coatings deposited by different deposition techniques, Thin Solid Films **270** (1995) 391–398

26.48 X. Li, B. Bhushan: Micro/nanomechanical and tribological characterization of ultra-thin amorphous carbon coatings, J. Mater. Res. **14** (1999) 2328–2337

26.49 X. Li, B. Bhushan: Mechanical and tribological studies of ultra-thin hard carbon overcoats for magnetic recording heads, Z. Metallkd. **90** (1999) 820–830

26.50 S. Sundararajan, B. Bhushan: Micro/nanotribology of ultra-thin hard amorphous carbon coatings using atomic force/friction force microscopy, Wear **225-229** (1999) 678–689

26.51 B. Bhushan: Chemical, mechanical, and tribological characterization of ultra-thin and hard amorphous carbon coatings as thin as 3.5 nm: Recent developments, Diam. Relat. Mater. **8** (1999) 1985–2015

26.52 I. I. Aksenov, V. E. Strel'Nitskii: Wear resistance of diamond-like carbon coatings, Surf. Coat. Technol. **47** (1991) 252–256

26.53 D. R. McKenzie, D. Muller, B. A. Pailthorpe, Z. H. Wang, E. Kravtchinskaia, D. Segal, P. B. Lukins, P. J. Martin, G. Amaratunga, P. H. Gaskell, A. Saeed: Properties of tetrahedral amorphous carbon prepared by vacuum arc deposition, Diam. Relat. Mater. **1** (1991) 51–59

26.54 R. Lossy, D. L. Pappas, R. A. Roy, J. J. Cuomo: Filtered arc deposition of amorphous diamond, Appl. Phys. Lett. **61** (1992) 171–173

26.55 I. G. Brown, A. Anders, S. Anders, M. R. Dickinson, I. C. Ivanov, R. A. MacGill, X. Y. Yao, K. M. Yu: Plasma synthesis of metallic and composite thin films with atomically mixed substrate bonding, Nucl. Instrum. Meth. Phys. Res. B **80–81** (1993) 1281–1287

26.56 P. J. Fallon, V. S. Veerasamy, C. A. Davis, J. Robertson, G. A. J. Amaratunga, W. I. Milne, J. Koskinen: Properties of filtered-ion-beam-deposited diamond-like carbon as a function of ion energy, Phys. Rev. B **48** (1993) 4777–4782

26.57 S. Anders, A. Anders, I. G. Brown, B. Wei, K. Komvopoulos, J. W. Ager III, K. M. Yu: Effect of vacuum arc deposition parameters on the properties of amorphous carbon thin films, Surf. Coat. Technol. **68-69** (1994) 388–393

26.58 S. Anders, A. Anders, I. G. Brown, M. R. Dickinson, R. A. MacGill: Metal plasma immersion ion implantation and deposition using arc plasma sources, J. Vac. Sci. Technol. B **12** (1994) 815–820

26.59 S. Anders, A. Anders, I. G. Brown: Transport of vacuum arc plasma through magnetic macroparticle filters, Plasma Sources Sci. **4** (1995) 1–12

26.60 D. M. Swec, M. J. Mirtich, B. A. Banks: Ion beam and plasma methods of producing diamondlike carbon films, Report No. NASATM102301, Cleveland 1988

26.61 A. Erdemir, M. Switala, R. Wei, P. Wilbur: A tribological investigation of the graphite-to-diamond-like behavior of amorphous carbon films ion beam deposited on ceramic substrates, Surf. Coat. Technol. **50** (1991) 17–23

26.62 A. Erdemir, F. A. Nicols, X. Z. Pan, R. Wei, P. J. Wilbur: Friction and wear performance of ion-beam deposited diamond-like carbon films on steel substrates, Diam. Relat. Mater. **3** (1993) 119–125

26.63 R. Wei, P. J. Wilbur, M. J. Liston: Effects of diamond-like hydrocarbon films on rolling contact fatigue of bearing steels, Diam. Relat. Mater. **2** (1993) 898–903

26.64 A. Erdemir, C. Donnet: Tribology of diamond, diamond-like carbon, and related films. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 871–908

26.65 J. Asmussen: Electron cyclotron resonance microwave discharges for etching and thin-film deposition, J. Vac. Sci. Technol. A **7** (1989) 883–893

26.66 J. Suzuki, S. Okada: Deposition of diamondlike carbon films using electron cyclotron resonance plasma chemical vapor deposition from ethylene gas, Jpn. J. Appl. Phys. **34** (1995) L1218–L1220

26.67 B.A. Banks, S.K. Rutledge: Ion beam sputter deposited diamond like films, J. Vac. Sci. Technol. **21** (1982) 807–814

26.68 C. Weissmantel, K. Bewilogua, K. Breuer, D. Dietrich, U. Ebersbach, H.J. Erler, B. Rau, G. Reisse: Preparation and properties of hard i-C and i-BN coatings, Thin Solid Films **96** (1982) 31–44

26.69 H. Dimigen, H. Hubsch: Applying low-friction wear-resistant thin solid films by physical vapor deposition, Philips Tech. Rev. **41** (1983/84) 186–197

26.70 B. Bhushan, B.K. Gupta, M.H. Azarian: Nanoindentation, microscratch, friction and wear studies for contact recording applications, Wear **181–183** (1995) 743–758

26.71 B. Bhushan, L. Yang, C. Gao, S. Suri, R.A. Miller, B. Marchon: Friction and wear studies of magnetic thin-film rigid disks with glass-ceramic, glass and aluminum-magnesium substrates, Wear **190** (1995) 44–59

26.72 L. Holland, S.M. Ojha: Deposition of hard and insulating carbonaceous films of an RF target in butane plasma, Thin Solid Films **38** (1976) L17–L19

26.73 L.P. Andersson: A review of recent work on hard i-C films, Thin Solid Films **86** (1981) 193–200

26.74 A. Bubenzer, B. Dischler, B. Brandt, P. Koidl: R.F. plasma deposited amorphous hydrogenated hard carbon thin films, preparation, properties and applications, J. Appl. Phys. **54** (1983) 4590–4594

26.75 A. Grill, B.S. Meyerson, V.V. Patel: Diamond-like carbon films by RF plasma-assisted chemical vapor deposition from acetylene, IBM J. Res. Develop. **34** (1990) 849–857

26.76 A. Grill, B.S. Meyerson, V.V. Patel: Interface modification for improving the adhesion of a-C:H to metals, J. Mater. Res. **3** (1988) 214

26.77 A. Grill, V.V. Patel, B.S. Meyerson: Optical and tribological properties of heat-treated diamond-like carbon, J. Mater. Res. **5** (1990) 2531–2537

26.78 F. Jansen, M. Machonkin, S. Kaplan, S. Hark: The effect of hydrogenation on the properties of ion beam sputter deposited amorphous carbon, J. Vac. Sci. Technol. A **3** (1985) 605–609

26.79 S. Kaplan, F. Jansen, M. Machonkin: Characterization of amorphous carbon-hydrogen films by solid-state nuclear magnetic resonance, Appl. Phys. Lett. **47** (1985) 750–753

26.80 H.C. Tsai, D.B. Bogy, M.K. Kundmann, D.K. Veirs, M.R. Hilton, S.T. Mayer: Structure and properties of sputtered carbon overcoats on rigid magnetic media disks, J. Vac. Sci. Technol. A **6** (1988) 2307–2315

26.81 B. Marchon, M. Salmeron, W. Siekhaus: Observation of graphitic and amorphous structures on the surface of hard carbon films by scanning tunneling microscopy, Phys. Rev. B **39** (1989) 12907–12910

26.82 B. Dischler, A. Bubenzer, P. Koidl: Hard carbon coatings with low optical-absorption, Appl. Phys. Lett. **42** (1983) 636–638

26.83 D.S. Knight, W.B. White: Characterization of diamond films by Raman spectroscopy, J. Mater. Res. **4** (1989) 385–393

26.84 J.W. Ager, D.K. Veirs, C.M. Rosenblatt: Spatially resolved Raman studies of diamond films grown by chemical vapor deposition, Phys. Rev. B **43** (1991) 6491–6499

26.85 W. Scharff, K. Hammer, O. Stenzel, J. Ullman, M. Vogel, T. Frauenheim, B. Eibisch, S. Roth, S. Schulze, I. Muhling: Preparation of amorphous i-C films by ion-assisted methods, Thin Solid Films **171** (1989) 157–169

26.86 B. Bhushan, X. Li: Nanomechanical characterization of solid surfaces and thin films, Int. Mater. Rev. **48** (2003) 125–164

26.87 X. Li, D. Diao, B. Bhushan: Fracture mechanisms of thin amorphous carbon films in nanoindentation, Acta Mater. **45** (1997) 4453–4461

26.88 X. Li, B. Bhushan: Measurement of fracture toughness of ultra-thin amorphous carbon films, Thin Solid Films **315** (1998) 214–221

26.89 X. Li, B. Bhushan: Evaluation of fracture toughness of ultra-thin amorphous carbon coatings deposited by different deposition techniques, Thin Solid Films **355–356** (1999) 330–336

26.90 X. Li, B. Bhushan: Development of a nanoscale fatigue measurement technique and its application to ultrathin amorphous carbon coatings, Scripta Mater. **47** (2002) 473–479

26.91 X. Li, B. Bhushan: Nanofatigue studies of ultrathin hard carbon overcoats used in magnetic storage devices, J. Appl. Phys. **91** (2002) 8334–8336

26.92 J. Robertson: Deposition of diamond-like carbon, Philos. Trans. R. Soc. London A **342** (1993) 277–286

26.93 S.J. Bull: Tribology of carbon coatings: DLC, diamond and beyond, Diam. Relat. Mater. **4** (1995) 827–836

26.94 N. Savvides, T.J. Bell: Microhardness and Young's modulus of diamond and diamondlike carbon films, J. Appl. Phys. **72** (1992) 2791–2796

26.95 B. Bhushan, M.F. Doerner: Role of mechanical properties and surface texture in the real area of contact of magnetic rigid disks, ASME J. Tribol. **111** (1989) 452–458

26.96 S. Suresh: *Fatigue of Materials* (Cambridge Univ. Press, Cambridge 1991)

26.97 D. B. Marshall, A. G. Evans: Measurement of adherence of residual stresses in thin films by indentation. I. Mechanics of interface delamination, J. Appl. Phys. **15** (1984) 2632–2638

26.98 A. G. Evans, J. W. Hutchinson: On the mechanics of delamination and spalling in compressed films, Int. J. Solids Struct. **20** (1984) 455–466

26.99 S. Sundararajan, B. Bhushan: Development of a continuous microscratch technique in an atomic force microscope and its application to study scratch resistance of ultrathin hard amorphous carbon coatings, J. Mater. Res. **16** (2001) 437–445

26.100 B. Bhushan, B. K. Gupta: Micromechanical characterization of Ni-P coated aluminum-magnesium, glass and glass-ceramic substrates and finished magnetic thin-film rigid disks, Adv. Info. Storage Syst. **6** (1995) 193–208

26.101 X. Li, B. Bhushan: Micromechanical and tribological characterization of hard amorphous carbon coatings as thin as 5 nm for magnetic recording heads, Wear **220** (1998) 51–58

26.102 B. Bhushan, V. N. Koinkar: Microscale mechanical and tribological characterization of hard amorphous coatings as thin as 5 nm for magnetic disks, Surf. Coat. Technol. **76–77** (1995) 655–669

26.103 V. N. Koinkar, B. Bhushan: Microtribological properties of hard amorphous carbon protective coatings for thin-film magnetic disks and heads, Proc. Inst. Mech. Eng., Part J **211** (1997) 365–372

26.104 T. W. Wu: Microscratch and load relaxation tests for ultra-thin films, J. Mater. Res. **6** (1991) 407–426

26.105 R. Memming, H. J. Tolle, P. E. Wierenga: Properties of polymeric layers of hydrogenated amorphous carbon produced by plasma-activated chemical vapor deposition: tribological and mechanical properties, Thin Solid Films **143** (1986) 31–41

26.106 C. Donnet, T. Le Mogne, L. Ponsonnet, M. Belin, A. Grill, V. Patel: The respective role of oxygen and water vapor on the tribology of hydrogenated diamond-like carbon coatings, Tribol. Lett. **4** (1998) 259

26.107 F. P. Bowden, J. E. Young: Friction of diamond, graphite and carbon and the influence of surface films, Proc. R. Soc. Lond. **208** (1951) 444–455

26.108 B. Marchon, N. Heiman, M. R. Khan: Evidence for tribochemical wear on amorphous carbon thin films, IEEE Trans. Magn. **26** (1990) 168–170

26.109 M. T. Dugger, Y. W. Chung, B. Bhushan, W. Rothschild: Friction, wear, and interfacial chemistry in thin film magnetic rigid disk files, ASME J. Tribol. **112** (1990) 238–245

26.110 B. D. Strom, D. B. Bogy, C. S. Bhatia, B. Bhushan: Tribochemical effects of various gases and water vapor on thin film magnetic disks with carbon overcoats, ASME J. Tribol. **113** (1991) 689–693

26.111 B. Bhushan, J. Ruan: Tribological performance of thin film amorphous carbon overcoats for magnetic recording rigid disks in various environments, Surf. Coat. Technol. **68/69** (1994) 644–650

26.112 B. Bhushan, G. S. A. M. Theunissen, X. Li: Tribological studies of chromium oxide films for magnetic recording applications, Thin Solid Films **311** (1997) 67–80

27. Self-Assembled Monolayers for Controlling Adhesion, Friction and Wear

Reliability of micro- and nanodevices, as well as magnetic storage devices require the use of lubricant films for the protection of sliding surfaces. To minimize high adhesion, friction, and because of small clearances in the devices, these films should be molecularly thick. Liquid films of low surface tension or certain hydrophobic solid films can be used. Ordered molecular assemblies with high hydrophobicity can be engineered using chemical grafting of various polymer molecules with suitable functional head groups, spacer chains and nonpolar surface terminal groups.

The classical approach to lubrication uses multi-molecular layers of liquid lubricants. Boundary lubricant films are formed by either physisorption, chemisorption, or chemical reaction. The physiosorbed films can be either monomolecularly or polymolecularly thick. The chemisorbed films are monomolecular, but stoichiometric films formed by chemical reaction can be multilayered. A good boundary lubricant should have a high degree of interaction between its molecules and the sliding surface. As a general rule, liquids are good lubricants when they are polar and thus able to grip on solid surfaces (or be adsorbed).

In this chapter, we focus on self-assembled monolayers (SAMs) for high hydrophobicity and/or low adhesion, friction, and wear. SAMs are produced by various organic precursors. We first present a primer to organic chemistry followed by an overview on suitable substrates, head groups, spacer chains, and end groups in the molecular chains and an overview of tribological properties of SAMs. The adhesion, friction, and wear properties of SAMs, having alkyl and biphenyl spacer chains with different surface terminal and head groups, are surveyed. The friction data are explained using a molecular spring model in which the local stiffness and intermolecular force govern its frictional performance. Based on the nanotribological studies of SAM films by AFM, they exhibit attractive hydrophobic and tribological properties.

27.1 **A Primer to Organic Chemistry** 834
27.1.1 Electronegativity/Polarity 834
27.1.2 Classification and Structure of Organic Compounds 835
27.1.3 Polar and Nonpolar Groups 838

27.2 **Self-Assembled Monolayers: Substrates, Head Groups, Spacer Chains, and End Groups** 839

27.3 **Tribological Properties of SAMs** 841
27.3.1 Surface Roughness and Friction Images of SAMs Films 844
27.3.2 Adhesion, Friction, and Work of Adhesion 844
27.3.3 Stiffness, Molecular Spring Model, and Micropatterned SAMs............. 848
27.3.4 Influence of Humidity, Temperature, and Velocity on Adhesion and Friction 850
27.3.5 Wear and Scratch Resistance of SAMs 853

27.4 **Closure** 856

References 857

Reliability of micro/nanodevices, also commonly referred to as micro/nanoelectromechanical systems (MEMS/NEMS), as well as magnetic storage devices (which include magnetic rigid disk drives, flexible disk drives, and tape drives) require the use of molecularly thick films for protection of sliding surfaces [27.1–8]. A solid or liquid film is generally necessary for acceptable tribological properties of sliding interfaces. However, a small quantity of high surface tension liquid present between smooth surfaces can substantially increase the adhesion, friction, and wear as a result of formation of menisci or adhesive bridges [27.9, 10]. It becomes a major concern in micro/nanoscale devices operating at ultralow loads, as the liquid mediated adhesive force may be on the same order as the external load.

The source of the liquid film can be either preexisting film of liquid and/or capillary condensates of water vapor from the environment. If the liquid wets the surface ($0 \leq \theta < 90°$, where θ is the contact angle between the liquid-vapor interface and the liquid-solid interface for a liquid droplet sitting on a solid surface, Fig. 27.1a), the liquid surface is thereby constrained to lie parallel to the surface [27.11], and the complete liquid surface must therefore be concave in shape, Fig. 27.1b. Surface tension results in a pressure difference across any meniscus surface, referred to as capillary pressure or Laplace pressure, and is negative for a concave meniscus [27.9, 10]. The negative Laplace pressure results in an intrinsic attractive (adhesive) force, which depends on the interface roughness (local geometry of interacting asperities and number of asperities), the surface tension, and the contact angle. During sliding, frictional effects need to be overcome, not only because of external load, but also because of intrinsic adhesive force.

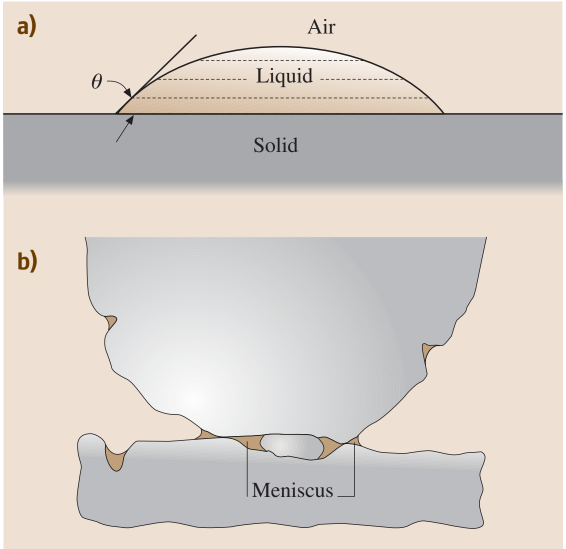

Fig. 27.1 **(a)** Schematic of a sessile-drop on a solid surface and the definition of contact angle, and **(b)** formation of meniscus bridges as a result of liquid present at an interface. The direct measurement of contact angle is most widely made from sessile-drops. The angle is generally measured by aligning a tangent with the drop profile at the point of contact with the solid surface using a telescope equipped with a goniometer eyepiece

Measured value of high static friction force contributed largely by liquid mediated adhesion (meniscus contribution) is generally referred to as "stiction". Basically, there are three ways to minimize the effect of liquid-mediated adhesion: increase in surface roughness, use of hydrophobic (water-fearing) rather than hydrophilic (water-loving) surfaces, and the use of a liquid with low surface tension [27.9, 10, 12, 13].

As an example, bulk silicon and polysilicon films used in the construction of micro/nanodevices can be dipped in hydrofluoric (HF) to make them hydrophobic (e.g., *Maboudian* [27.14]; *Scherge* and *Schaefer* [27.15]). In HF etching of silicon, hydrogen passivates the silicon surface by saturating the dangling bonds, and it results in a hydrogen-terminated silicon surface that is responsible for less adsorption of water. However, after exposure of the treated surface to the environment, it reoxidizes, which can adsorb water, and thus the surface again becomes hydrophilic.

The surfaces can also be treated or coated with a liquid with relatively low surface tension or a certain solid film to make them hydrophobic and/or to control adhesion, friction, and wear. The liquid lubricant film should be thin (about half of the composite roughness of the interface) to minimize liquid-mediated adhesion contribution [27.9, 10, 12, 13]. Thus, for ultra-smooth surfaces with RMS roughness on the order of a few nm, molecularly thick liquid films are required for liquid lubrication. The classical approach to lubrication uses freely supported multi-molecular layers of liquid lubricants [27.2, 4, 9, 10, 16–19]. Boundary lubricant films are formed by either physisorption, chemisorption, or chemical reaction. The physiosorbed films can be either monomolecularly or polymolecularly thick. The chemisorbed films are monomolecular, but stoichiometric films formed by chemical reaction can be multilayered. In general, stability and durability of surface films decrease in the following order: chemically reacted films, chemisorbed films, and physisorbed films. A good boundary lubricant should have a high degree of interaction between its molecules and the sliding surface. As a general rule, liquids are good lubricants when they are polar and thus able to grip on solid surfaces (or be adsorbed). Polar lubricants contain reactive functional end groups. Boundary lubrication properties are also dependent on the molecular conformation and lubricant spreading. It should be noted that the liquid films with a thickness on the order of a few nm, may be discontinuous and may deposit in an island form of nonuniform thickness with a lateral resolution on the nanometerscale.

Solid films are also commonly used for controlling hydrophobicity and/or adhesion, friction, and wear. Hydrophobic films have nonpolar surface terminal groups (to be described later) that repel water. These films have low surface energy (15–30 dyn/cm) and high contact angle ($\theta \geq 90°$), which minimize wetting (e.g., *Zisman* [27.20]; *Schrader* and *Loeb* [27.21]; *Neumann* and *Spelt* [27.22]). It should be noted that these films do not totally eliminate wetting. Multi-molecularly thick (few tenths of a nm) films of conventional solid-lubricants have been studied. *Hansma* et al. [27.23] reported the deposition of multi-molecularly thick, highly-oriented PTFE films from the melt or vapor phase or from solution by a mechanical deposition technique by dragging the polymer at controlled temperature, pressure, and speed against a smooth glass substrate. *Scandella* et al. [27.24] reported that the coefficient of nanoscale friction of MoS_2 platelets on mica, obtained by exfoliation of lithium intercalated MoS_2 in water, was a factor of 1.4 less than that of mica itself. However, MoS_2 is reactive to water and its friction and wear properties degrade with an increase in humidity [27.9, 10]. Amorphous diamond-like carbon (DLC) coatings can be produced with extremely high hardness and are commercially used as wear-resistant coatings [27.25, 26]. Their largest application is in magnetic storage devices [27.2]. Doping of the DLC matrix with elements like hydrogen, nitrogen, oxygen, silicon, and fluorine influences their hydrophobicity and tribological properties [27.25, 27, 28]. Nitrogen and oxygen reduce the contact angle (or increase the surface energy), due to the strong polarity that is formed when these elements are bonded to carbon. On the other hand, silicon and fluorine increase the contact angle ranging from 70–100 ° (or reduce the surface energy to 20–40 dyn/cm), making them hydrophobic [27.29, 30]. Nanocomposite coatings with a diamond-like carbon (a-C:H) network and a glass-like a-Si:O network are generally deposited using a plasma-enhanced chemical vapor deposition (PECVD) technique in which plasma is formed from a siloxane precursor using a hot filament. For a fluorinated DLC, CF_4 is added as the fluorocarbon source to an acetylene plasma. In addition, fluorination of DLC can be achieved by the post-deposition treatment of DLC coatings in a CF_4 plasma. Silicon- and fluorine-containing DLC coatings mainly reduce their polarity due to the loss of sp^2-bonded carbon (polarization potential of the involved π electrons) and dangling bonds of the DLC network. As silicon and fluorine are unable to form double bonds, they force carbon into a sp^3 bonding state [27.30]. Friction and wear properties of both silicon-containing and fluorinated DLC coatings have been reported to be superior to that of conventional DLC coatings [27.31,32]. However, DLC coatings require line of sight deposition process, which prevents deposition on complex geometries. Furthermore, it has been reported that some self-assembled monolayers (SAMs) are superior to DLC coatings in hydrophobicity and tribological performance [27.33].

Organized and dense molecular-scale layers of, preferably, long-chain organic molecules are known to be superior lubricants on both macro- and micro/nanoscales as compared with freely supported multi-molecular layers [27.4, 34–38]. Common techniques to produce molecular-scale organized layers are Langmuir–Blodgett (LB) deposition and chemical grafting of organic molecules to realize SAMs [27.39,40]. In the LB technique, organic molecules from suitable amphiphilic molecules are first organized at the air-water interface and then physisorbed on a solid surface to form mono- or multi-molecular layers [27.41]. In the case of SAMs, the functional groups of molecules chemisorb on a solid surface, which results in the spontaneous formation of robust, highly ordered, oriented, and dense monolayers [27.40]. In both cases, the organic molecules used have well distinguished amphiphilic properties (hydrophilic functional head and a hydrophobic aliphatic tail) so that an adsorption of such molecules on an active inorganic substrate leads to their firm attachment to the surface. Direct organization of SAMs on the solid surfaces allows coating in the tight areas such as the bearing and journal surfaces in an assembled bearing. The weak adhesion of classical LB films to the substrate surface restricts their lifetimes during sliding, whereas certain SAMs can be very durable [27.34]. As a result, SAMs are of great interest in tribological applications.

Much of the research in the application of SAMs has been carried out for the so-called soft lithographic technique [27.42, 43]. This is a non-photolithographic technique. Photolithography is based on a projection-printing system used for projection of an image from a mask to a thin-film photoresist and its resolution is limited by optical diffraction limits. In the soft lithography, an elastomeric stamp or mold is used to generate micropatterns of SAMs either by contact printing (known as microcontact printing or μCP [27.44]), by embossing (imprinting) [27.45], or by replica molding [27.46], which circumvents the diffraction limits of photolithography. The stamps are generally cast from photolithographically generated patterned masters, and the stamp material is generally polydimethylsiloxane (PDMS). In µCP, the ink is a SAM precursor to produce nm-thick resists with lines thinner than 100 nm. Although soft lithography requires little capital investment, it is doubtful that it would ever replace well established photolithography. However, µCP and embossing techniques may be used to produce microdevices that are a substantially cheaper and more flexible choice of material for construction than conventional photolithography (e.g., SAMs and non-SAM entities for µCP and elastomers for embossing).

Other applications for SAMs are in the areas of bio/chemical and optical sensors, for use as drug-delivery vehicles and in the construction of electronic components [27.39, 47]. Bio/chemical sensors require the use of highly sensitive organic layers with tailored biological properties that can be incorporated into electronic, optical, or electrochemical devices. Self-assembled microscopic vesicles are being developed to ferry potentially lifesaving drugs to cancer patients. By getting organic, metal, and phosphonate molecules (complexes of phosphorous and oxygen atoms) to assemble themselves into conductive materials, these can be produced as self-made sandwiches for use as electronic components. The application of interest here is an application requiring hydrophobicity and/or low friction and wear.

An overview of molecularly thick layers of liquid lubricants and conventional solid lubricant can be found in various references [27.2, 4, 9, 10, 18, 19, 25]. In this chapter, we focus on SAMs for high hydrophobicity and/or, low adhesion, friction, and wear. SAMs are produced by various organic precursors. We first present a primer to organic chemistry followed by an overview on suitable substrates, spacer chains, and end groups in the molecular chains, an overview of tribological properties of SAMs, and some concluding remarks.

27.1 A Primer to Organic Chemistry

All organic compounds contain the carbon (C) atom. Carbon, in combination with hydrogen, oxygen, nitrogen, and sulfur results in a large number of organic compounds. The atomic number of carbon is six, and its electron structure is $1s^2\,2s^2\,2p^2$. Two stable isotopes of carbon, ^{12}C and ^{13}C, exist. With four electrons in its outer shell, carbon forms four covalent bonds with each bond resulting from two atoms sharing a pair of electrons. The number of electron pairs that two atoms share determines whether or not the bond is single or multiple. In a single bond, only one pair of electrons is shared by the atoms. Carbon can also form multiple bonds by sharing two or three pairs of electrons between the atoms. For example, the double bond formed by sharing two electron pairs is stronger than a single bond, and it is shorter than a single bond. An organic compound is classified as saturated if it contains only the single bond and as unsaturated if the molecules possess one or more multiple carbon-carbon bonds.

27.1.1 Electronegativity/Polarity

When two different kinds of atoms share a pair of electrons a bond is formed in which electrons are shared unequally, one atom assumes a partial positive charge and the other a negative charge with respect to each other. This difference in charge occurs because the two atoms exert unequal attraction for the pair of shared electrons. The attractive force that an atom of an element has for shared electrons in a molecule or polyatomic ion is known as its electronegativity. Elements differ in their electronegativities. A scale of relative electronegatives, in which the most electronegative element, fluorine, is assigned a value of 4.0, was developed by Linus Pauling. The relative electronegativities of the elements in the periodic table can be found in most undergraduate chemistry textbooks (e.g., *Hein* et al. [27.48]). Relative electronegativity of the nonmetals is high compared to that of metals. Relative electronegativity of selected elements of interest with high values is presented in Table 27.1.

The polarity of a bond is determined by the difference in electronegativity values of the atoms forming the bond. If the electronegativities are the same, the bond is nonpolar and the electrons are shared equally. In this type of bond, there is no separation of positive and negative charge between atoms. If the atoms have greatly different electronegativities, the bond is very polar. A dipole is a molecule that is electrically asymmetrical, causing it to be oppositely charged at two points. As an example, both hydrogen and chlorine need one electron to form stable electron configurations. They share a pair of

Table 27.1 Relative electronegativity of selected elements

Element	Relative electronegativity
F	4.0
O	3.5
N	3.0
Cl	3.0
C	2.5
S	2.5
P	2.1
H	2.1

electrons in hydrogen chloride, HCl. Chlorine is more electronegative and, therefore, has a greater attraction for the shared electrons than hydrogen. As a result, the pair of electrons is displaced toward the chlorine atom, giving it a partial negative charge and leaving the hydrogen atom with a partial positive charge, Fig. 27.2. However, the entire molecule, HCl, is electrically neutral. The hydrogen atom with a partial positive charge (exposed proton on one end) can be easily attracted to the negative charge of other molecules, and this is responsible for the polarity of the molecule. A partial charge is usually indicated by δ, and the electronic structure of HClis given as:

$$\overset{\delta+}{\mathrm{H}}:\overset{\delta-}{\underset{\cdot\cdot}{\ddot{\mathrm{Cl}}}}: \; .$$

Similar to the HCl molecule, HF is polar, and both behave as a small dipole. On the other hand, methane (CH_4), carbon tetrachloride (CCl_4), and carbon dioxide (CO_2) are nonpolar. In CH_4 and CCl_4, the four C−H and C−Cl polar bonds are identical, and because these bonds emanate from the center to the corners of a tetrahedron in the molecule, the effect of their polarities cancels out one another. CO_2 (O=C=O) is nonpolar, because carbon-oxygen dipoles cancel each other by acting in opposite directions. Symmetric molecules are generally nonpolar. Water (H−O−H) is a polar molecule. If the atoms in water were linear as in CO_2, two O−H dipoles would cancel each other, and the molecule would be nonpolar. However, water has a bent structure with an angle of 105° between the two bonds, which is responsible for the water being a polar molecule.

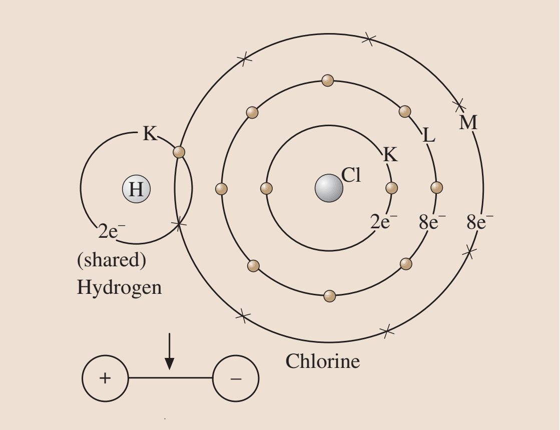

Fig. 27.2 Schematic representation of the formation of a polar HCl molecule

27.1.2 Classification and Structure of Organic Compounds

Tables 27.2–27.6 present selected organic compounds grouped into classes.

Hydrocarbons

Hydrocarbons are compounds that are composed entirely of carbon and hydrogen atoms bonded to each other by covalent bonds. Saturated hydrocarbons (alkanes) contain single bonds. Unsaturated hydrocarbons that contain a carbon-carbon double bond are called alkenes, and ones with a triple bond are called alkynes. Unsaturated hydrocarbons that contain an aromatic ring, e.g., benzene ring, are called aromatic hydrocarbons.

Saturated Hydrocarbons: Alkanes. The alkanes, also known as paraffins, are saturated hydrocarbons, straight- or branched-chain hydrocarbons with only single covalent bonds between the carbon atoms. The general molecular formula for alkanes is C_nH_{2n+2}, where n is the number of carbon atoms in the molecule. Each carbon atom is connected to four other atoms by four single covalent bonds. These bonds are separated by angles of 109.5° (the angle by lines drawn from the center of a regular tetrahedron to its corners). Alkane molecules contain only carbon-carbon and carbon-hydrogen bonds, which are symmetrically directed toward the corners of a tetrahedron. Therefore, alkane molecules are essentially nonpolar.

Common alkyl groups have the general formula C_nH_{2n+1} (one hydrogen atom less than the corresponding alkane). The missing H atom may be detached from any carbon in the alkane. The

Table 27.2 Names and formulas of selected hydrocarbons

Name	Formula
Saturated hydrocarbons	
Straight-chain alkanes	C_nH_{2n+2}
e.g., methane,	CH_4
ethane	C_2H_6 or CH_3CH_3
Alkyl groups	C_nH_{2n+1}
e.g., methyl,	$-CH_3$
ethyl	$-CH_2CH_3$
Unsaturated hydrocarbons	
Alkenes	$(CH_2)_n$
e.g., ethene,	C_2H_4 or $CH_2{=}CH_2$
propene	C_3H_6 or $CH_3CH{=}CH_2$
Alkynes	
e.g., acetylene	$HC{\equiv}CH$
Aromatic hydrocarbons	
e.g., benzene	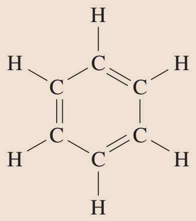

Table 27.3 Names and formulas of selected alcohols, ethers, phenols, and thiols

Name	Formula
Alcohols	$R{-}OH$
e.g., methanol,	CH_3OH
ethanol	CH_3CH_2OH
Ethers	$R{-}O{-}R'$
e.g., dimethyl ether,	$CH_3{-}O{-}CH_3$
diethyl ether	$CH_3CH_2{-}O{-}CH_2CH_3$
Phenols	C_6H_5OH or 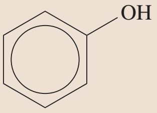
Thiols	$-SH$
e.g., methanethiol	CH_3SH

The letters R- and R′- represent an alkyl group. The R- groups in ethers can be the same or different and can be alkyl or aromatic (Ar) groups.

name of the group is formed from the name of the corresponding alkane by replacing -ane with -yl ending. Some examples are shown in Table 27.2.

Unsaturated Hydrocarbons. Unsaturated hydrocarbons consist of three families of compounds that contain fewer hydrogen atoms than the alkane with the corresponding number of carbon atoms and contain multiple bonds between carbon atoms. These include alkenes (with carbon-carbon double bonds), alkynes (with carbon-carbon triple bonds), and aromatic compounds (with benzene rings that are arranged in a six-member ring with one hydrogen atom bonded to each carbon atom and three carbon-carbon double bonds). Some examples are shown in Table 27.2.

Alcohols, Ethers, Phenols, and Thiols

Organic molecules with certain functional groups are synthesized for desirable properties. Alcohols, ethers, and phenols are derived from the structure of water by replacing the hydrogen atoms of water with alkyl groups (R) or aromatic (Ar) rings. For example, phenol is a class of compounds that have a hydroxy group attached to an aromatic ring (benzene ring). Organic compounds that contain the −SH group are analogs of alcohols and are known as thiols. Some examples are shown in Table 27.3.

Aldehydes and Ketones

Both aldehydes and ketones contain the carbonyl group, >C=O, a carbon-oxygen double bond. Aldehydes have at least one hydrogen atom bonded to the carbonyl group, whereas ketones have only an alkyl or aromatic group bonded to the carbonyl group. The general formula for the saturated homologous series of aldehydes and ketones is $C_nH_{2n}O$. Some examples are shown in Table 27.4.

Carboxyl Acids and Esters

The functional group of the carboxylic acids is known as a carboxyl group, represented as −COOH. Carboxylic acids can be either aliphatic (RCOOH) or aromatic (ArCOOH). The carboxylic acids with even numbers of carbon atoms, n, ranging from 4 to about 20 are called fatty acids (e.g., $n = 10$, 12, 14, 16 and 18 are called capric acid, lauric acid, myristic acid, palmitic acid, and stearic acid, respectively).

Esters are alcohol derivates of carboxylic acids. Their general formula is RCOOR′, where R may be a hydrogen, alkyl group, or aromatic group, and R′ may be an alkyl group or aromatic group but not a hydrogen. Esters are found in fats and oils. Some examples are shown in Table 27.5.

Table 27.4 Names and formulas of selected aldehydes and ketones

Name	Formula
Aldehydes	RCHO or 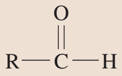 $R-\overset{\overset{O}{\|}}{C}-H$
	ArCHO or 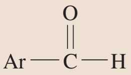 $Ar-\overset{\overset{O}{\|}}{C}-H$
e.g., methanal or	HCHO
formaldehyde ethanal or acetaldehyde	CH_3CHO
Ketones	RCOR′ or 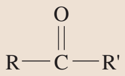 $R-\overset{\overset{O}{\|}}{C}-R'$
	RCOAr or 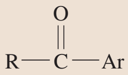 $R-\overset{\overset{O}{\|}}{C}-Ar$
	ArCOAr or 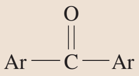 $Ar-\overset{\overset{O}{\|}}{C}-Ar$
e.g., butanone or methyl-ethyl ketone	$CH_3COCH_2CH_3$

The letters R and R′ represent alkyl groups, and Ar represents an aromatic group.

Table 27.5 Names and formulas of selected carboxylic acids and esters

Name	Formula
Carboxylic acid*	RCOOH or 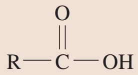 $\mathrm{R{-}\overset{O}{\overset{\parallel}{C}}{-}OH}$
	ArCOOH or 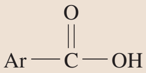 $\mathrm{Ar{-}\overset{O}{\overset{\parallel}{C}}{-}OH}$
e.g., methanoic acid (formic acid),	$HCOOH$
ethanoic acid (acetic acid),	CH_3COOH
octadecanoic acid (stearic acid)	$CH_3(CH_2)_{16}COOH$
Esters**	RCOOR′ or 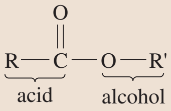 $\underbrace{\mathrm{R{-}\overset{O}{\overset{\parallel}{C}}}}_{\text{acid}}{-}\underbrace{\mathrm{O{-}R'}}_{\text{alcohol}}$
e.g., methyl propanoate	$CH_3CH_2COOCH_3$

* The letter R represents an alkyl group and Ar represents an aromatic group.

** The letter R represents hydrogen, an alkyl group, or an aromatic group and R′ represents an alkyl group or aromatic group

Table 27.6 Names and formulas of selected organic nitrogen compounds (amides and amines)

Name	Formula
Amides	$RCONH_2$ or 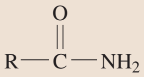 $\mathrm{R{-}\overset{O}{\overset{\parallel}{C}}{-}NH_2}$
e.g., methanamide (formamide),	$HCONH_2$
ethanamide (acetamide)	CH_3CONH_2
Amines	RNH_2 or 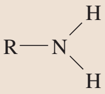 $\mathrm{R{-}N}\begin{array}{l}\diagup \mathrm{H}\\ \diagdown \mathrm{H}\end{array}$
	R_2NH
	R_3N
e.g., methylamine,	CH_3NH_2
ethylamine	$CH_3CH_2NH_2$

The letter R represents an alkyl group or aromatic group.

Amides and Amines

Amides and amines are organic compounds containing nitrogen. Amides are nitrogen derivates of carboxylic acids. The carbon atom of a carbonyl group is bonded directly to a nitrogen atom of a $-NH_2$, $-NHR$, or $-NR_2$ group. The characteristic structure of amide is $RCONH_2$.

An amine is a substituted ammonia molecule that has a general structure of RNH_2, R_2NH, or R_3N, where R is an alkyl or aromatic group. Some examples are shown in Table 27.6.

27.1.3 Polar and Nonpolar Groups

Table 27.7 summarizes polar and nonpolar groups commonly used in the construction of hydrophobic and hydrophilic molecules. Table 27.8 lists the relative polarity of selected polar groups [27.49]. Thiol, silane, carboxylic acid, and alcohol (hydroxyl) groups are the most commonly used polar anchor groups for their attachment to surfaces. Methyl and trifluoromethyl are commonly used end groups for hydrophobic film surfaces.

Springer Handbook of Nanotechnology
B. Bhushan • © Springer 2004

27.2 Self-Assembled Monolayers: Substrates, Head Groups, Spacer Chains, and End Groups

SAMs are formed as a result of spontaneous, self-organization of functionalized organic molecules onto the surfaces of appropriate substrates into stable, well-defined structures, Fig. 27.3. The final structure is close to or at thermodynamic equilibrium, and as a result, it tends to form spontaneously and rejects defects. SAMs consist of three building groups: a head group that binds strongly to a substrate, a surface terminal (tail) group that constitutes the outer surface of the film, and a spacer chain (backbone chain) that connects the head and surface terminal groups. The SAMs are named based on the surface terminal group followed by spacer chain and the head group (or type of compound formed at the surface). For a SAM to control hydrophobicity, adhesion, friction, and wear, it should be strongly adherent to the substrate, and the surface terminal group of the organic molecu-

Table 27.7 Some examples of polar (hydrophilic) and nonpolar (hydrophobic) groups

Name	Formula
Polar	
Alcohol (hydroxyl)	$-OH$
Carboxyl acid	$-COOH$
Aldehyde	$-COH$
Ketone	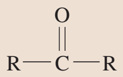 $R-\overset{O}{\overset{\|}{C}}-R$
Ester	$-COO-$
Carbonyl	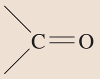
Ether	$R-O-R$
Amine	$-NH_2$
Amide	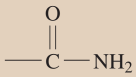 $-\overset{O}{\overset{\|}{C}}-NH_2$
Phenol	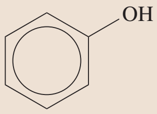
Thiol	$-SH$
Trichlorosilane	$SiCl_3$
Nonpolar	
Methyl	$-CH_3$
Trifluoromethyl	$-CF_3$
Aryl (benzene ring)	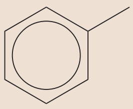

The letter R represents an alkyl group.

Table 27.8 Organic groups listed in the increasing order of polarity

Alkanes
Alkenes
Aromatic hydrocarbons
Ethers
Trichlorosilanes
Aldehydes, ketones, esters, carbonyls
Thiols
Amines
Alcohols, phenols
Amides
Carboxylic acids

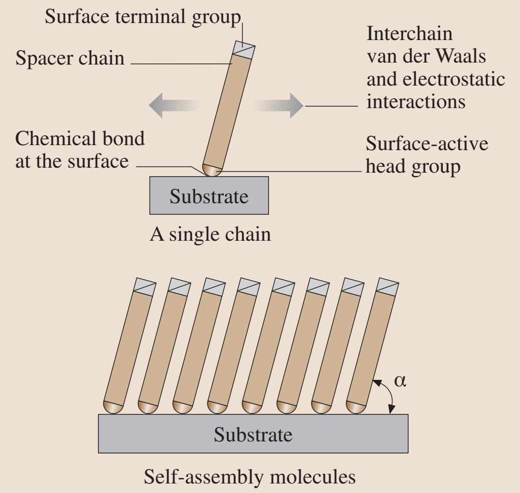
Surface terminal group
Spacer chain
Interchain
van der Waals
and electrostatic
interactions
Chemical bond
at the surface
Surface-active
head group
Substrate
A single chain
α
Substrate
Self-assembly molecules

Fig. 27.3 Schematic of a self-assembled monolayer on a surface and the associated forces

lar chain should be nonpolar. For strong attachment of the organic molecules to the substrate, the head group of the molecular chain should contain a polar end group resulting in the exothermic process (energies on the order of tens of kcal/mol), i. e., it results in an apparent pinning of the head group to a specific site on the surface through a chemical bond. Furthermore, molecular structure and any crosslinking would have a significant effect on their friction and wear performance. The substrate surface should have a high surface energy (hydropolic), so that there will be a strong tendency for molecules to adsorb on the surface. The surface should be highly functional with polar groups and dangling bonds (generally unpaired electrons), so that they can react with organic molecules and provide a strong bond. Because of the exothermic head group–substrate interactions, molecules try to occupy every available binding site on the surface, and during this process they generally push together molecules that have already adsorbed. The process results in the formation of crystalline molecular assemblies. The interactions between molecular chains are van der Waals or electrostatic-type with energies on the order of a few (< 10) kcal/mol, exothermic. The molecular chains in SAMs are not perpendicular to the surface; the tilt angle depends on the anchor group, as well as on the substrate and the spacer group. For example, the tilt angle for alkanethiolate on Au is typically about 30–35° angle with respect to substrate normal.

SAMs are usually produced by immersing a substrate in a solution containing precursor (ligand) that is reactive to the substrate surface, or by exposing the substrate to a vapor of the reactive chemical species [27.39]. Table 27.9 lists selected systems that have been used for formation of SAMs [27.43]. The spacer chain of SAM is mostly an alkyl chain ($-C_nH_{2n+1}$), or made of a derivatized alkyl group. By attaching different terminal groups at the surface, the film surface can be made to attract or repel water. The commonly used surface terminal group of a hydrophobic film with low surface energy, in the case of a single alkyl chain, is a nonpolar methyl ($-CH_3$) or trifluoromethyl ($-CF_3$) group. For a hydrophilic film, the commonly used surface terminal groups are alcohol ($-OH$) or carboxylic acid ($-COOH$) groups. Surface active head groups most commonly used are thiol ($-SH$), silane (e.g., trichlorosilane or $-SiCl_3$), and carboxyl ($-COOH$) groups. The substrates most commonly used are gold, silver, platinum, copper, hydroxylated (activated) surfaces of SiO_2 on Si, Al_2O_3 on Al, glass, and hydrogen-terminated single-crystal silicon ($H-Si$). Hydroxylation of oxide surfaces is important to make them hydrophilic. Thermally grown silica can be activated through a sulfochromic treatment. The sample is dipped in a solution consisting of 100 mL of concentrated H_2SO_4, 5 mL of water, and 2 g of potassium

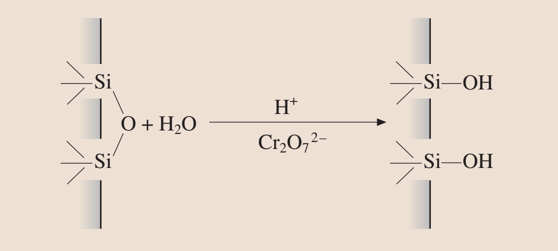

Fig. 27.4 Schematic showing the hydroxylation process occurring on a silica surface through a sulfochromic treatment

Table 27.9 Selected substrates and precursors that have commonly been used for formation of SAMs

Substrate	Precursor	Binding with substrate
Au	RSH (thiol)	RS−Au
Au	ArSH (thiol)	ArS−Au
Au	RSSR′ (disulfide)	RS−Au
Au	RSR′ (sulfide)	
Si/SiO_2, glass	$RSiCl_3$ (trichlorosilane)	Si−O−Si (Siloxane)
Si/Si−H	RCOOH (carboxyl)	R−Si
Metal oxides (e.g., Al_2O_3, SnO_2, TiO_2)	RCOOH (carboxyl)	RCOO− . . . MO_n

R represents alkane (C_nH_{2n+2}) and Ar represents aromatic hydrocarbon. It consists of various surface active head groups and mostly with methyl terminal group bichromate for 3–15 min [27.34]. The sample is then rinsed with flowing pure water. This results in a surface with silanol groups (−Si−OH) that is hydrophilic, Fig. 27.4. Bulk silicon, polysilicon film, or SiO_2 film surfaces can also be treated to produce activated hydrophilic silica surface by immersion in about three parts H_2SO_4 and one part H_2O_2 at temperatures ranging from ambient to 80 °C. For organic molecules to pack together and provide a better ordering, a substrate for given molecules should be selected such that the cross-sectional diameter of the spacer chains of the molecule is equal to or smaller than the distance between the anchor groups attached to the substrate. As an example, epitaxial Au film on glass, mica, or single-crystal silicon, produced by e-beam evaporation, is commonly used because it can be deposited on smooth surfaces as a film that is atomically flat and defect free. For the case of alkanethiolate film, the advantage of Au substrate over SiO_2 substrate is that it results in better ordering, because the cross-sectional diameter of the alkane molecule is slightly smaller than the distance between sulfur atoms attached to the Au substrate (~ 0.53 nm). The thickness of the film can be controlled by varying the length of the hydrocarbon chain, and the surface properties of the film can be modified by the terminal group.

Some of the SAMs have been widely reported. SAMs of long-chain fatty acids, $C_nH_{2n+1}COOH$ or $(CH_3)(CH_2)_nCOOH$ ($n = 10, 12, 14$ or 16), on glass or alumina have been used in films studies since the 1950s [27.16, 20]. Probably the most studied SAMs to date are n-alkanethiolate monolayers $CH_3(CH_2)_nS-$, prepared from adsorption of alkanethiol $-(CH_2)_nSH$ solution onto an Au film (n-alkyl and n-alkane are used interchangeably) [27.35–37, 43] and n-alkylsiloxane monolayers produced by adsorption of n-alkyltrichlorosilane $-(CH_2)_nSiCl_3$ solution onto hydroxylated Si/SiO_2 substrate [27.50] with siloxane (Si−O−Si) binding, Fig. 27.5. *Jung* et al. [27.51] have produced organosulfur monolayers – decanethiol $(CH_3)(CH_2)_9SH$, and didecyl $CH_3(CH_2)_9{-}S{-}(CH_2)_9CH_3$ on Au films. *Geyer* et al. [27.52], *Bhushan* and *Liu* [27.35], *Liu* et al. [27.36], and *Liu* and Bhushan [27.37] have produced monolayers of 1,1′-biphenyl-4-thiol (BPT) on Au surface, where the spacer chain of the film consists of two phenyl rings with a hydrogen end group. *Bhushan* and *Liu* [27.35] and *Liu* and *Bhushan* [27.37] have also reported monolayers of 4,4′-dihydroxybiphenyl on Si surface.

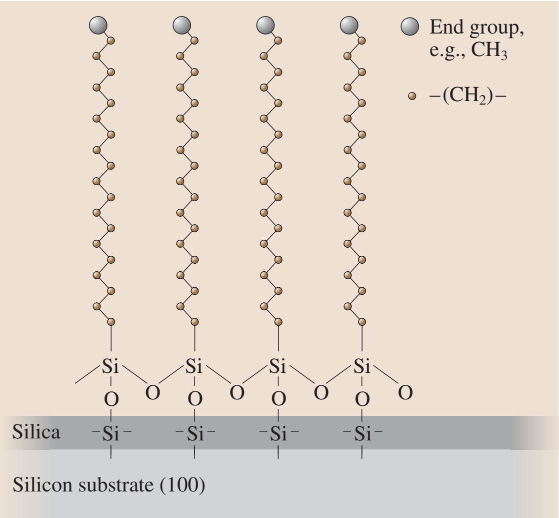

Fig. 27.5 Schematics of a methyl-terminated, n-alkylsiloxane monolayer on Si/SiO_2

27.3 Tribological Properties of SAMs

The basis for molecular design and tailoring of SAMs must start from a complete knowledge of the interrelationships between the molecular structure and tribological properties of SAMs, as well as a deep understanding of the adhesion, friction, and wear mechanisms of SAMs at the molecular level. Friction and wear properties of SAMs have been studied on macro- and nanoscales. Macroscale tests are conducted using a so-called pin-on-disk tribotester apparatus in which a ball specimen slides against a lubricated flat specimen [27.9, 10]. Nanoscale tests are conducted using an atomic force/friction force microscope (AFM/FFM) [27.4, 9, 10]. In the AFM/FFM experiments, a sharp tip of radius ranging from about 5–50 nm slides against a SAM specimen. A Si_3N_4 tip is commonly used for friction studies and a natural diamond tip is commonly used for scratch, wear, and indentation studies.

In early studies, the effect of chain length of the carbon atoms of fatty acid monolayers on the coefficient of friction and wear on the macroscale was studied by *Bowden* and *Tabor* [27.16] and *Zisman* [27.20]. *Zisman* reported that for the monolayers deposited on a glass surface sliding against a stainless steel surface, there is a steady decrease in friction with increasing chain length. At a significantly long chain length, the coefficient of friction reaches a lower limit, Fig. 27.6. He further reported that monolayers having a chain length below 12 carbon atoms behave as liquids (poor durability), those with chain length of 12–15 carbon atoms behave like a plastic solid (medium durability), whereas those with chain lengths above 15 carbon atoms behave like a crystalline solid (high durability). Investigations by *Ruhe* et al. [27.53] indicated that the lifetime of the alkylsilane monolayer coating on a silicon surface increases greatly with an increase in the chain length of the alkyl substituent. *DePalma* and *Tillman* [27.54] showed that a monolayer of n-octadecyltrichlorosilane (n-$C_{18}H_{37}SiCl_3$, OTS) is an effective lubricant on silicon, followed by n-undecyltrichlorosilane (n-$C_{11}H_{23}SiCl_3$, UTS) and (tridecafluoro-1,1,2,2-tetrahydrooct-1-yl) trichlorosilane (n-$C_6F_{13}CH_2CH_2SiCl_3$, FTS). *Ando* et al. [27.55] suggested that γ-(N,N-dioctadecylsuccinylamino) propyltriethoxysilane monolayer is a candidate lubricant. The film exhibited a low coefficient of kinetic friction of 0.1, without stick-slip. However, tests in all these investigations were carried out using a pin-on-disk tribotester under relatively large normal loads up to 0.15 N. Thus, the relevance of these tests is questionable for micro/nanodevice application.

With the development of AFM techniques, researchers have successfully characterized the nanotribological properties of self-assembled monolayers [27.1, 4, 34–38]. Studies by *Bhushan* et al. [27.34] showed that C_{18} alkylsiloxane films exhibit the lowest coefficient of friction and can withstand a much higher normal load during sliding compared to LB films, soft Au films, and hard SiO_2 coatings.

McDermott et al. [27.56] studied the effect of length of alkyl chains on the frictional properties of methyl-terminated n-alkylthiolate $CH_3(CH_2)_nS-$ films chemisorbed on Au(111) using an AFM. They reported that the longer chain monolayers exhibit a markedly lower friction and a reduced propensity to wear than shorter chain monolayers, Fig. 27.7. These results are in good agreement with the macroscale results by *Zisman* [27.20]. They also conducted infrared reflection spectroscopy to measure the bandwidth of the methylene stretching mode (ν_a (CH_2)), which exhibits a qualitative correlation with the packing density of the chains. They found that the chain structures of monolayers prepared with longer chain lengths are more ordered and more densely packed in comparison to those of monolayers prepared with shorter chain lengths. They further reported that the ability of the longer chain monolayers to retain molecular-scale order during shear leads to a lower observed friction. Monolayers with a chain length of more than 12 carbon atoms, preferably 18 or more, are desirable for tribological applications. (Incidentally, the monolayer with 18 carbon atoms, octadecanethiol films, are commonly studied.)

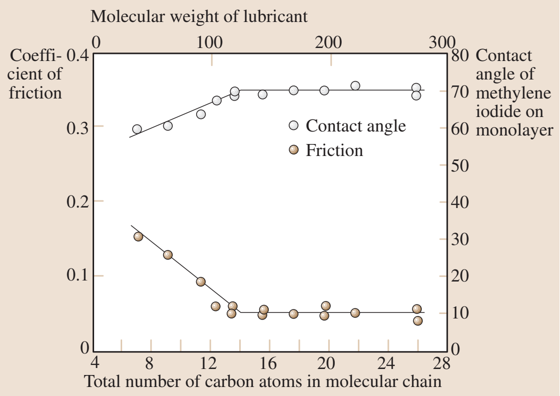

Fig. 27.6 Effect of chain length (or molecular weight) on coefficient of macroscale friction of stainless steel sliding on glass lubricated with a monolayer of fatty acid and contact angle of methyl iodide on condensed monolayers of fatty acid on glass [27.20]

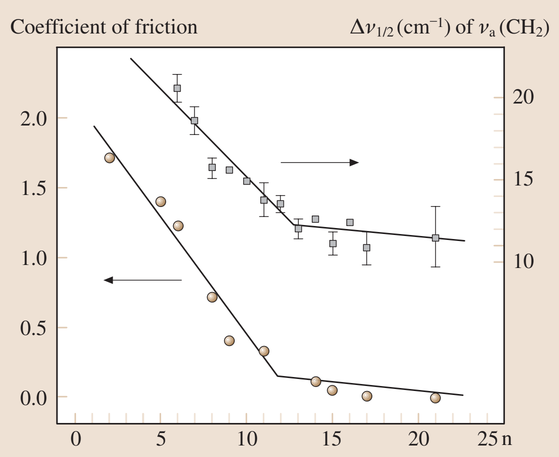

Fig. 27.7 Effect of chain length of methyl-terminated, n-alkanethiolate over Au film $AuS(CH_2)_nCH_3$ on the coefficient of microscale friction and peak bandwidth at half maximum ($\Delta\nu_{1/2}$) for the bandwidth of the methylene stretching mode [ν_a (CH_2)] [27.56]

Xiao et al. [27.57] and *Lio* et al. [27.58] also studied the effect of the length of the alkyl chains on the frictional properties of n-alkanethiolate films on gold and n-alkylsilane films on mica. Friction was found to be particularly high with short chains of less than eight carbon atoms. Thiols and silanes exhibit a similar friction force for the same n when $n > 11$; while for $n < 11$, silanes exhibit higher friction, larger than that for thiols by a factor of about 3 for $n = 6$. The increase in friction was attributed to the large number of dissipative modes in the less ordered chains that occurs when going from a thiol to a silane anchor or when decreasing n. Longer chains ($n > 11$), stabilized by van der Waals attraction, form more compact and rigid layers and act as better lubricants. *Schonherr* and *Vancso* [27.59] also correlated the magnitude of friction with the order of the alkane chains. The disorder of the short-chain hydrocarbon disulfides SAMs was found to result in a significant increase in the magnitude of friction.

Fluorinated carbon (fluorocarbon) molecules are commonly used for lubrication [27.2, 9, 10]. *Kim* et al. [27.60, 61] studied the influence of methyl-, isopropyl-, and trifluoromethyl-terminated alkanethiols on the friction properties of SAMs. They reported a factor of three increases in the friction in going from the methyl-terminated SAM to the trifluoromethyl-terminated SAM. They suggested that fluorinated monolayers exhibit higher frictional properties due to tighter packing at the interface, which arises from the larger van der Waals radii of the fluorine atoms. Subsequent steric and rotational factors between adjacent terminal groups give rise to long-range multi-molecular interactions in the plane of the $-CF_3$ groups. When these energetic barriers are overcome in the film structure, more energy is imparted to the film during sliding and results in higher friction for the fluorinated films.

Tsukruk et al. [27.62] and *Lee* et al. [27.63] studied the nanotribological properties of C_{60}-terminated alkyltrichlorosilanes and alkanedisulfides self-assembled monolayers. Both of their studies observed that the frictional forces of the C_{60}-terminated films are higher than those on normal methyl-terminated SAMs. However, *Tsukruk* et al. [27.62] reported that the C_{60}-terminated alkyltrichlorosilanes SAMs exhibit high wear resistance. *Tsukruk* and *Bliznyuk* [27.64] also studied the adhesion and friction between a Si sample and Si_3N_4 tip, in which both surfaces were modified by $-CH_3$, $-NH_2$, and $-SO_3H$ terminated silane-based SAMs. Various polymer molecules for the backbone were used. They reported that a very broad maximum adhesive force in the pH range of from 4 to 8 with a minimum adhesion at $pH > 9$ and $pH < 3$ were observed for all of the studied mating surfaces. This observation can be understood by considering a balance of electrostatic and van der Waals interaction between composite surfaces with multiple isoelectric points. The friction coefficient of NH_2/NH_2 and SO_3H/SO_3H mating SAMs is very high in aqueous solutions. Cappings of NH_2 modified surfaces (3-aminopropyltriethoxysilane) with rigid and soft polymer layers resulted in a significant reduction in adhesion to a level lower than that of untreated surfaces [27.65]. *Fujihira* et al. [27.66] studied the influence of surface terminal groups of SAMs and functional tip on adhesive force. It was found that the adhesive forces measured in air increase in the order of CH_3/CH_3, $CH_3/COOH$, and COOH/COOH.

Bhushan and *Liu* [27.35], *Liu* et al. [27.36], and *Liu* and *Bhushan* [27.37, 38] have studied the influence of spacer chains, surface terminal groups, and head groups on adhesion, friction, and wear properties of SAMs. They have explained the friction mechanisms using a molecular spring model in which local stiffness and the intermolecular forces govern the friction properties. They studied the influence of relative humidity, temperature, and velocity on adhesion and friction. They also investigated the wear mechanisms of SAMs by a continuous microscratch AFM technique.

To date, the nanotribological properties of alkyltrichlorosilanes, alkanethiols, and biphenyl thiol SAMs have been widely studied. In this chapter, we review, in some detail, the nanotribological properties of five kinds of SAMs, having alkyl and biphenyl spacer chains with different surface terminal groups ($-CH_3$, $-COOH$, and $-OH$), and head groups ($-SH$ and $-OH$), which have been investigated by AFM at various operating conditions, Fig. 27.8 [27.35–38]. Hexadecane thiol (HDT), 1,1′-biphenyl-4-thiol (BPT) and crosslinked BPT (BPTC), and 16-mercaptohexadecanoic acid thiol (MHA) were deposited on Au(111) substrates. A 4,4′-dihydroxybiphenyl (DHBp) film was deposited on a hydrogenated Si(111) substrate. Figure 27.8 shows that HDT and MHA have the same head groups and spacers, but different surface terminal groups; their surface terminals are $-CH_3$ and $-COOH$, respectively. BPT and DHBp have the same spacers, but the head and surface terminal groups of DHBp are $-OH$ instead of $-SH$ and $-CH$ in BPT. Crosslinked BPTC was produced by irradiation of BPT monolayers with low energy electrons.

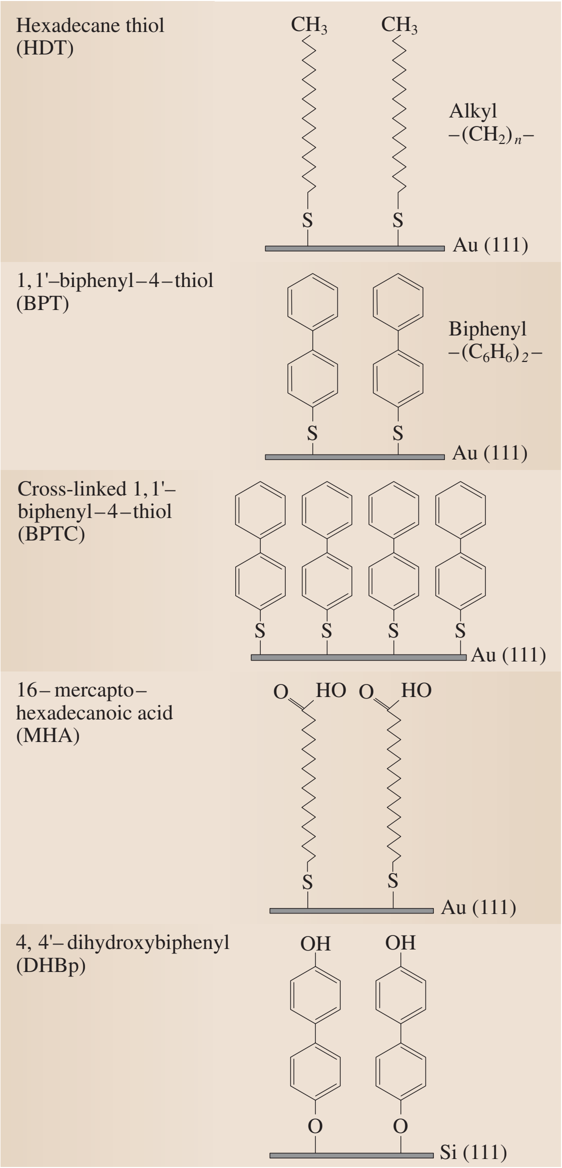

Fig. 27.8 Schematics of the structures of HDT, BPT, MHA, DHBp, and BPTC SAMs [27.35]

27.3.1 Surface Roughness and Friction Images of SAMs Films

Surface height and friction force images of SAMs were recorded simultaneously on an area of 1 µm × 1 µm by an AFM, Fig. 27.9 [27.35]. The topography of Au(111) film and SAMs deposited on Au(111) substrates appear to be granular. A good correlation between the surface height and the corresponding friction force images was observed. It was noticed that the friction force changed with surface height in the same direction (upwards and downwards) in both trace and retrace friction profiles of the friction loop. Thus, the change in friction force corresponds to transitions in surface slope [27.4, 35, 36, 67]. Figure 27.9 also indicates that the topography and friction images of Si(111) and DHBp/Si(111) are very similar; they exhibit featureless surfaces and their friction force images do not show any abrupt changes.

For further analysis presented later in this chapter, the roughness, thickness, tilt angles, and spacer chain lengths of Si(111), Au(111), and various SAMs are listed in Table 27.10 [27.35]. The roughness values of BPT and MHA are very close to that of Au(111). But the roughness of BPTC is lower than that of Au(111) and BPT; this is caused by electron irradiation. Table 27.10 indicates that the roughness values of HDT and DHBp are much higher than their substrate roughness of Au(111) and Si(111), respectively. This is caused by local aggregation of organic compounds on the substrates during SAMs deposition. Table 27.10 also indicates that the thickness of biphenyl thiol SAMs is generally thinner than the alkylthiol, which is responsible for shorter spacer chain in biphenyl thiol.

27.3.2 Adhesion, Friction, and Work of Adhesion

The average values and standard deviation of the adhesive force and coefficients of friction measured by contact mode AFM are presented in Fig. 27.10 [27.35]. Based on the data, the adhesive force and coefficient of friction of SAMs are less than their corresponding substrates. Among various films, HDT exhibits the lowest values. The ranking of adhesive forces F_a is in the following order: $F_{\text{a-Au}} > F_{\text{a-Si}} > F_{\text{a-DHBp}} \approx F_{\text{a-MHA}} > F_{\text{a-BPT}} > F_{\text{a-BPTC}} > F_{\text{a-HDT}}$. And the ranking of the coefficients of friction μ is in the following order: $\mu_{\text{Si}} > \mu_{\text{Au}} > \mu_{\text{DHBp}} > \mu_{\text{BPTC}} > \mu_{\text{BPT}} \approx \mu_{\text{MHA}} > \mu_{\text{HDT}}$. The ranking of various SAMs for adhesive force and coefficient of friction is similar. It suggests that alkylthiol and biphenyl SAMs can be used as effective molecular lubricants for micro/nanodevices fabricated from silicon.

In micro/nano scale contact, liquid capillary condensation is one of the factors that influence adhesion and friction. In the case of a sphere in contact with a flat surface, the attractive Laplace force caused by water

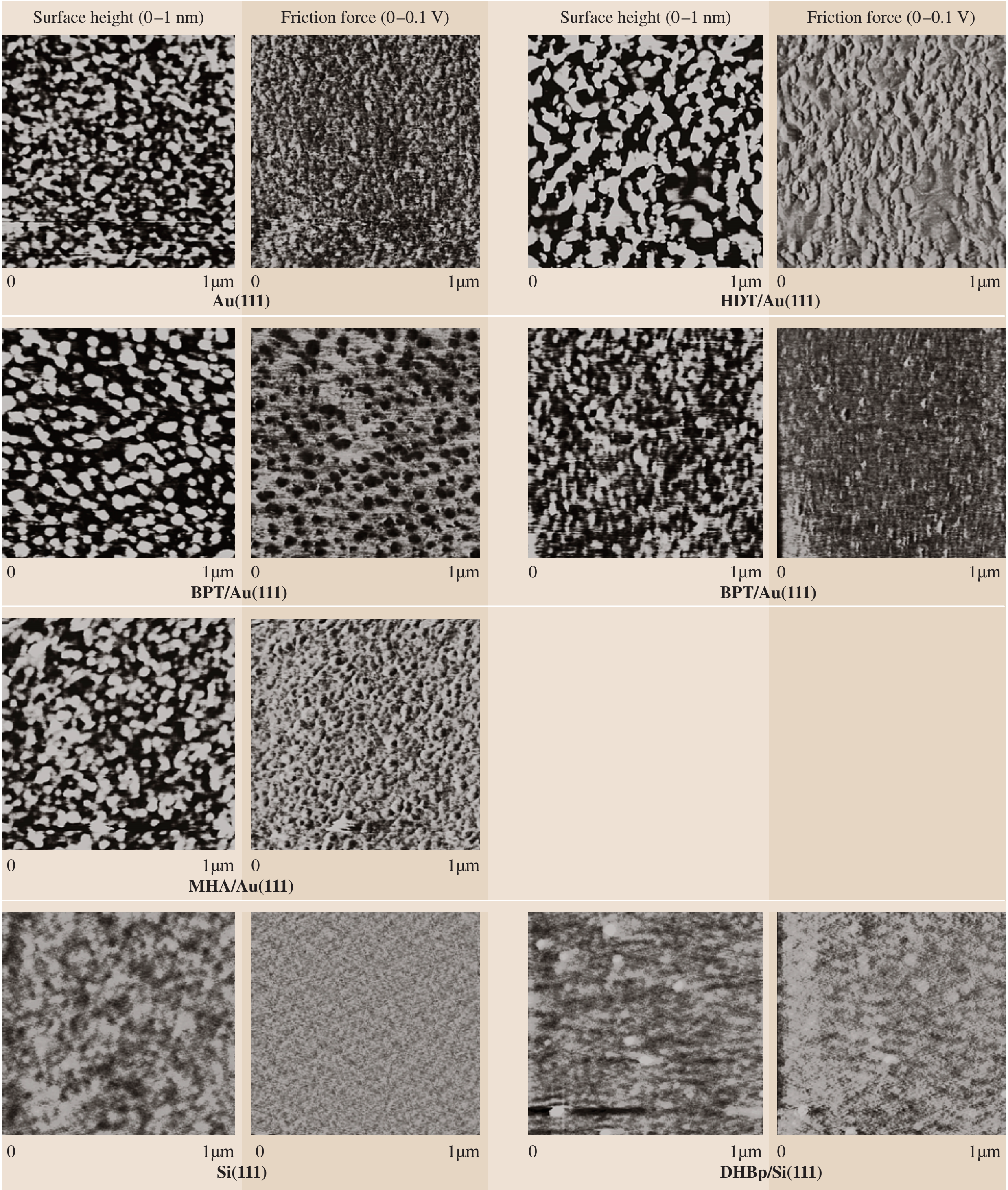
Surface height (0–1 nm)
Friction force (0–0.1 V)
Surface height (0–1 nm)
Friction force (0–0.1 V)
0
1 μm
Au(111)
HDT/Au(111)
BPT/Au(111)
BPT/Au(111)
MHA/Au(111)
Si(111)
DHBp/Si(111)

Fig. 27.9 AFM gray-scale surface height and friction force images of Au(111), Si(111), and various SAMs [27.35]

Table 27.10 The R_a roughness, thickness, tilt angles, and spacer chain lengths of SAMs

Samples	R_a roughness[1] (nm)	Thickness[2] (nm)	Tilt angle[2] (deg.)	Spacer length[3] (nm)
Si(111)	0.07			
Au(111)	0.37			
HDT	0.92	1.89	30	1.91
BPT	0.36	1.25	15	0.89
BPTC	0.14	1.14	25	0.89
MHA	0.37	2.01	30	1.91
DHBp	0.25	1.13	–	0.89

[1] Measured by an AFM with 1 µm × 1 µm scan size using a Si_3N_4 tip under 3.3 nN normal load.

[2] The thickness and tilt angles of BPT, BPTC, and DHBp are reported by *Geyer* et al. [27.52]. The thickness and tilt angles of HDT and MHA are reported by *Ulman* [27.40].

[3] The spacer chain lengths of alkylthiols were calculated by the method reported by *Miura* et al. [27.68]. The spacer chain lengths of biphenyl thiols were calculated by the data reported by *Ratajczak-Sitarz* et al. [27.69].

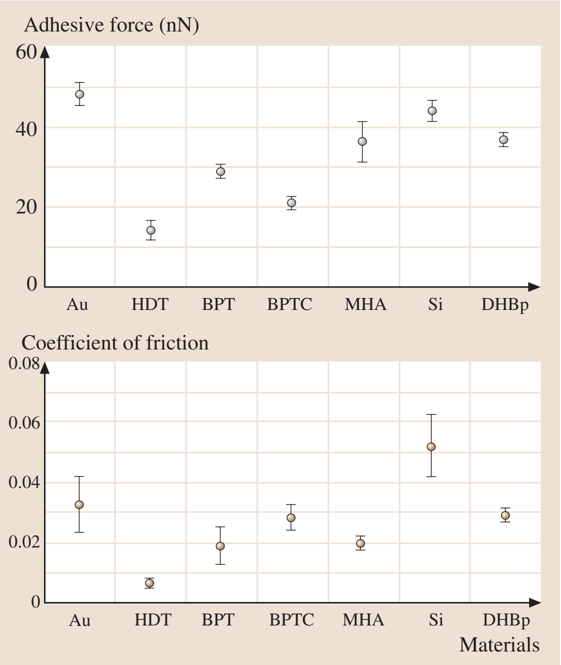

Fig. 27.10 Adhesive forces and coefficients of friction of Au(111), Si(111), and various SAMs [27.35]

capillary is:

$$F_L = 2\pi R \gamma_{la}(\cos\theta_1 + \cos\theta_2) , \qquad (27.1)$$

where R is the radius of the sphere, γ_{la} is the surface tension of the liquid against air, θ_1 and θ_2 are the contact angles between liquid and flat and spherical surfaces, respectively [27.9, 10]. In an AFM adhesive study, the tip-flat sample contact is just like a sphere in contact with a flat surface and the liquid is water. Since a single tip was used in the adhesion measurements, $\cos\theta_2$ can be treated as a constant. Therefore,

$$\begin{aligned} F_L &= 2\pi R \gamma_{la}(1 + \cos\theta_1) - 2\pi R \gamma_{la}(1 - \cos\theta_2) \\ &= 2\pi R \gamma_{la}(1 + \cos\theta_1) - C , \qquad (27.2) \end{aligned}$$

where C is a constant.

Based on the following Young-Dupre equation, work of adhesion W_a (the work required to pull apart the unit area of the solid-liquid interface) can be written as [27.70]

$$W_a = \gamma_{la}(1 + \cos\theta_1) . \qquad (27.3)$$

It indicates that W_a is determined by the contact angle of SAMs, i.e., is influenced by the surface chemistry properties (polarization and hydrophobicity) of SAMs. By substituting (27.3) into (27.2), F_L can be expressed as

$$F_L = 2\pi R W_a - C . \qquad (27.4)$$

When the influence on the adhesive force of other factors, such as van der Waals force, is very small, then adhesive force $F_a \approx F_L$. Thus the adhesive force F_a should be proportional to work of adhesion W_a.

Contact angle is a measure of the wettability of a solid by a liquid and determines the W_a value. The value of the contact angle changes with time. If the droplet expands, the contact angle is referred to as the advancing contact angle. The contact angles of distilled water on Si(111), Au(111), and SAMs were measured using a contact angle goniometer [27.35]. After dropping water on the sample surfaces, the dynamic advancing contact angle (DACA) was measured with a frequency of 60 s for spontaneous spreading for at least 420 s. The DACA as a function of time data is given in Fig. 27.11a. It indicates that there is a linear relationship between DACA and rest time. As stated by many authors, the static advancing contact angle (SACA) (at $t = 0$) should be used for surface characterization [27.71–73]. This SACA can be obtained by simple extrapolation of the data in Fig. 27.11a to $t = 0$. The SACA of Si(111), Au(111), and SAMs is summarized in Fig. 27.11b. For water, $\gamma_{\text{la}} = 72.6\,\text{mJ/m}^2$ at 22 °C, and by using (27.3), the W_a data are obtained and presented in Fig. 27.11c. The W_a can be ranked in the following order: $W_{\text{a-Si}}(117.0) > W_{\text{a-DHBp}}(98.8) \approx W_{\text{a-MHA}}(101.8) \approx W_{\text{a-Au}}(97.1) > W_{\text{a-BPT}}(86.8) > W_{\text{a-BPTC}}(82.1) > W_{\text{a-HDT}}(61.4)$. Excluding $W_{\text{a-Au}}$, this order exactly matches the order of adhesion force in Fig. 27.10. The relationship between F_a and W_a is summarized in Fig. 27.12 [27.35]. It indicates that the adhesive force F_a (nN) increases with work of adhesion W_a (mJ/m^2) by the following linear relationship:

$$F_\text{a} = 0.57 W_\text{a} - 22\,. \tag{27.5}$$

These experimental results agree well with the modeling prediction presented earlier in (27.4). It proves that on the nanoscale at ambient condition the adhesive force of SAMs is mainly influenced by the water capillary force. Comparing the results of Figure 27.12 with Figure 27.8, it is found that MHA and DHBp, which have polar surface terminal groups (−COOH and −OH), lead to larger W_a and eventually larger adhesive forces. Though both HDT and BPT do not have polar surface groups, the surface terminal of HDT has a symmetrical structure, which causes a smaller electrostatic attractive

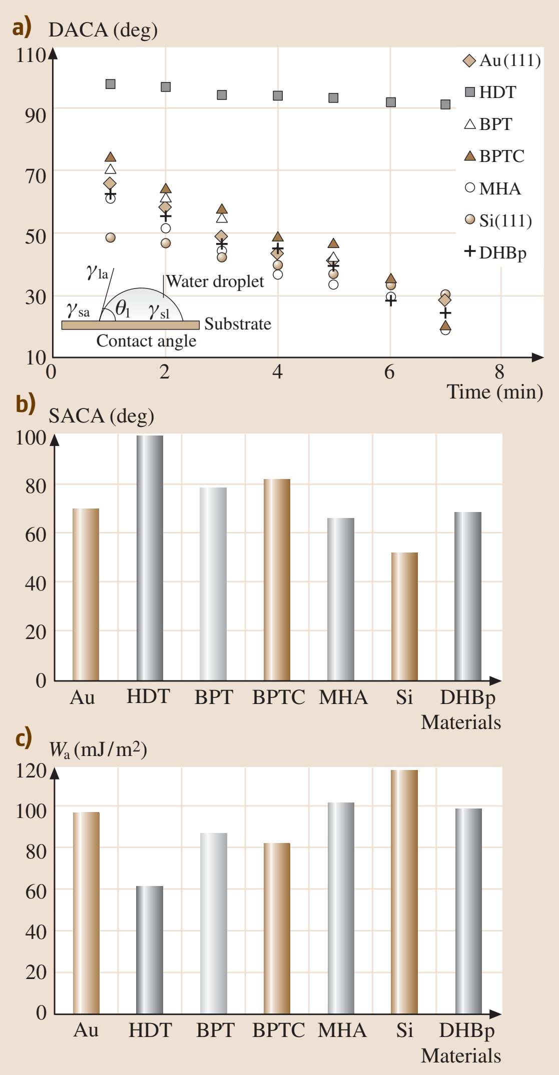

Fig. 27.11 **(a)** Variation of dynamic advancing contact angle (DACA) with time, **(b)** the static advancing contact angle (SACA), and **(c)** work of adhesion of Au(111), Si(111), and various SAMs. All of the points in this figure represent the mean value of six measurements. The uncertainty associated with the average contact angle is within ±2°. The insert schematic in **(a)** shows the contact angle of the water droplet on the sample surface [27.35]

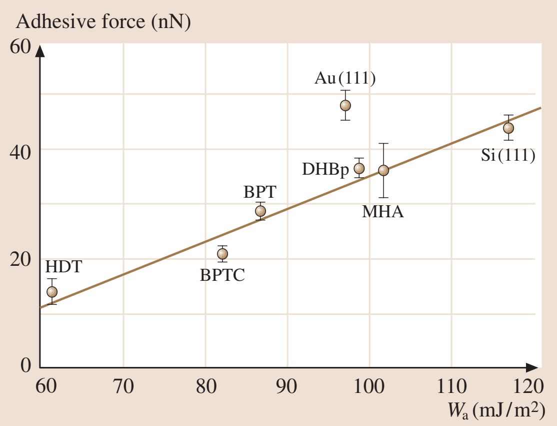

Fig. 27.12 Relationship between the adhesive forces and work of adhesion of different specimen [27.35]

force and yields smaller adhesive force than BPT. It is believed that the easy attachment of Au on the tip should be one of the reasons that cause the large adhesive force, which do not fit the linear relationship described by (27.5).

27.3.3 Stiffness, Molecular Spring Model, and Micropatterned SAMs

Next, the friction mechanisms of SAMs are examined. Monte Carlo simulation of the mechanical relaxation of $CH_3(CH_2)_{15}SH$ self-assembled monolayer performed by *Siepman* and *McDonald* [27.74] indicated that SAMs compress and respond nearly elastically to microindentation of an AFM tip under a critical normal load.

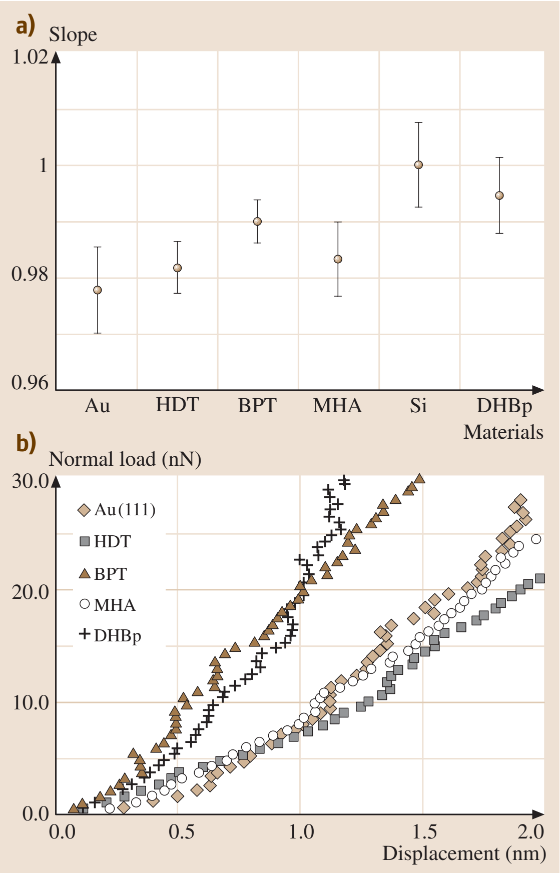

Fig. 27.13 **(a)** Slopes of cantilever deflection versus piezo displacement measured by force distance calibration mode AFM, and **(b)** normal load versus displacement curves of Au(111), Si(111), and various SAMs

Compression can lead to major changes in the mean molecular tilt (i. e., orientation), but the original structure is recovered as the normal load is removed.

To study the difference in the stiffness of various films, the stiffness properties were measured by an AFM in force-distance calibration mode, as well as in force modulation mode [27.4, 37]. Figure 27.13a shows the slope of cantilever deflection versus piezo movement obtained in the force-distance calibration mode. For an infinite hard material, since the surface can not be compressed, the cantilever deflection should equal the piezo movement distance, which means the slope of the cantilever deflection versus piezo movement equals 1. For a soft material, the surface can be compressed, causing a reduced cantilever deflection, and the slopes are smaller than 1. In this study the cantilever has been initially calibrated to slope $= 1$ against a clean, rigid Si(111) surface. Figure 27.13a indicates that the slopes of all of the SAMs are less than 1, which suggests that all of the SAMs can be compressed by a Si_3N_4 tip during the loading process. The slope value of SAMs can be ranked in the following order: $S_{\mathrm{DHBp}} > S_{\mathrm{BPT}} > S_{\mathrm{MHA}} \approx S_{\mathrm{HDT}}$. This order reflects the influence of molecular structure on the compression properties of SAMs. Since BPT and DHBp have rigid benzene structure, they are more difficult to compress than HDT and MHA. Figure 27.13b shows the variation of the displacement with normal load during indentation mode. It also clearly indicates that SAMs can be compressed. At a given normal load, long carbon chain structure SAMs such as HDT and MHA are easy to compress compared to rigid benzene-ring structure SAMs such as BPT and DHBp. *Garcia-Parajo* et al. [27.75] have also reported the compression and relaxation of octadecyltrichlorosilane (OTS) film in their loading and unloading tests.

In order to explain the frictional difference of SAMs, based on the friction and stiffness measurements by AFM and the Monte Carlo simulation, a molecular spring model is presented in Fig. 27.14. It is believed that the self-assembled molecules on a substrate are just like assembled molecular springs anchored to the substrates [27.35]. A Si_3N_4 tip sliding on the surface of SAMs is like a tip sliding on the top of "molecular springs or brush". The molecular spring assembly has compliant features and can experience compression and orientation under normal load. The orientation of the "molecular springs or brush" reduces the shearing force at the interface, which, in turn, reduces the friction force. The possibility of orientation is determined by the spring constant of a single molecule (local stiffness), as well as the interaction between the neighboring

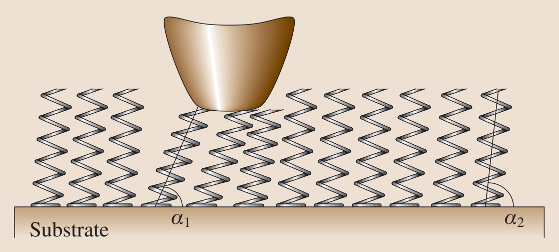

Fig. 27.14 Molecular spring model of SAMs. In this figure, $\alpha_1 < \alpha_2$, which is caused by the orientation under the normal load applied by an AFM tip. The orientation of the molecular springs reduces the shearing force at the interface, which, in turn, reduces the friction force. The molecular spring constant, as well as the inter-molecular forces can determine the magnitude of the coefficients of friction of SAMs. In this figure, the size of the tip and molecular springs are not drawn exactly to scale [27.35]

molecules, which can be reflected by packing density or packing energy. It should be noted that the orientation can lead to conformational defects along the molecular chains, which lead to energy dissipation. In the study of BPT by AFM, it was found that after the first several scans, the friction force is significantly reduced, but the surface height does not show any apparent change. This suggests that the molecular orientation can be facilitated by initial sliding and is reversible [27.38].

Based on the stiffness measurement results presented in Fig. 27.13 and the view of molecular structures in Fig. 27.14, biphenyl is a more rigid structure due to the contribution of two rigid benzene rings. Therefore, the spring constant of BPT is larger than that of HDT. The hydrogen (H^+) in a biphenyl chain has an electrostatic attractive force with the π electrons in the neighboring benzene ring. Thus, the intermolecular force between biphenyl chains is stronger than that for alkyl chains. The larger spring constant of BPT and stronger intermolecular force require a larger external force to allow it to orient, thus causing higher coefficient of friction. For MHA and DHBp, their basic chain structures are very close to HDT and BPT, respectively. But their surface terminals are different. The polar $-$COOH and $-$OH external functional groups in MHA and DHBp increase the adhesive force, thus leading to higher friction force than HDT and BPT, respectively. The crosslinking of BPT leads to a larger packing energy for BPTC. Therefore, it requires a larger external force to allow BPTC orientation, i. e., the coefficient of BPTC is higher than BPT.

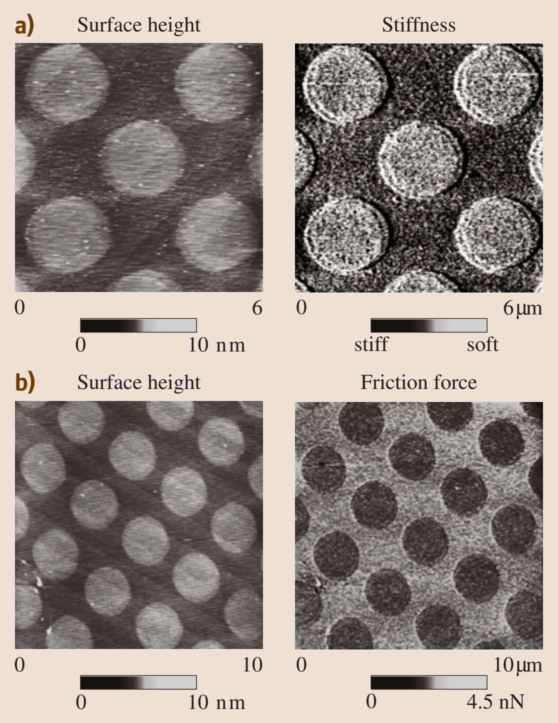

Fig. 27.15 **(a)** AFM gray-scale surface height and stiffness images, and **(b)** AFM gray-scale surface height and friction force images of micropatterned BDCS [27.37]

An elegant way to demonstrate the influence of molecular stiffness on friction is to investigate SAMs with different structures on the same wafer. For this purpose, a micropatterned SAM was prepared. First, the biphenyldimethylchlorosilane (BDCS) was deposited on silicon by the typical self-assembly method [27.37]. Then the film was partially crosslinked using mask technique by low energy electron irradiation. Finally, the micropatterned BDCS films were realized that had the as-deposited and crosslinked coatings on the same wafer. The local stiffness properties of these micropatterned samples were investigated by force modulation AFM technique [27.76]. The variation in the deflection amplitude provides a measure of the relative local stiffness of the surface. Surface height, stiffness, and friction images of the micropatterned biphenyldimethylchlorosilane (BDCS) specimen are obtained and presented in Fig. 27.15 [27.37]. The circular areas correspond to the as-deposited film, and the remaining area to the crosslinked film. Figure 27.15a indicates that crosslinking caused by the low energy electron irradiation leads to about a 0.5 nm decrease of the surface height of BDCS films. The corresponding stiffness images indicate that the crosslinked area has a higher stiffness than the as-deposited area. Figure 27.15b indicates that the as-deposited area (higher surface height area) has lower friction force. Obviously, the data of the micropatterned sample prove that the local stiffness of SAMs has an influence on their friction performance. Higher stiffness leads to larger friction force. These results provide a strong proof of the suggested molecular spring model.

In summary, it has been found that SAMs exhibit compliance and can experience compression and orientation under normal load. The orientation of SAMs reduces the shear stress at the interface; therefore SAMs can serve as good lubricants. The molecular spring constant (local stiffness), as well as the intermolecular forces can influence the magnitude of the coefficients of friction of SAMs.

27.3.4 Influence of Humidity, Temperature, and Velocity on Adhesion and Friction

The influence of relative humidity on adhesion and friction was studied in an environmentally controlled chamber at 22 °C [27.35]. The results are presented in Fig. 27.16. It shows that for Si(111), Au(111), DHBp and MHA, the adhesive and friction forces increase with relative humidity. For BPT and BPTC, the adhesive force only slightly increases with relative humidity when the relative humidity is higher than 40%, but it is very interesting that their friction decreases in the same range. For HDT, in the testing range, the adhesive and friction force are not sensitive to changes in relative humidity.

It has already been shown that the adhesive force is linearly related to the W_a value of the surfaces in Fig. 27.12. DHBp and MHA have polar surface terminals and larger W_a value. This must therefore lead to the larger adhesive forces in higher relative humidity, which, in turn, increase the friction force. In comparison, HDT has a nonpolar $-CH_3$ surface terminal and a very small W_a value. Thus, the adhesive force and friction force of HDT are not sensitive to the change of humidity. The W_a values of BPT and BPTC are between Si(111) and HDT (see Fig. 27.11c); therefore their adhesive forces show a slight increase when RH > 40%. The fact that BPT and BPTC show lower friction force when RH > 40% suggests that a thin adsorbed water layer can act as a lubricant [27.35, 37]. The reason why the adsorbed water layer acts as a lubricant may be related to the thickness of the absorbed layer.

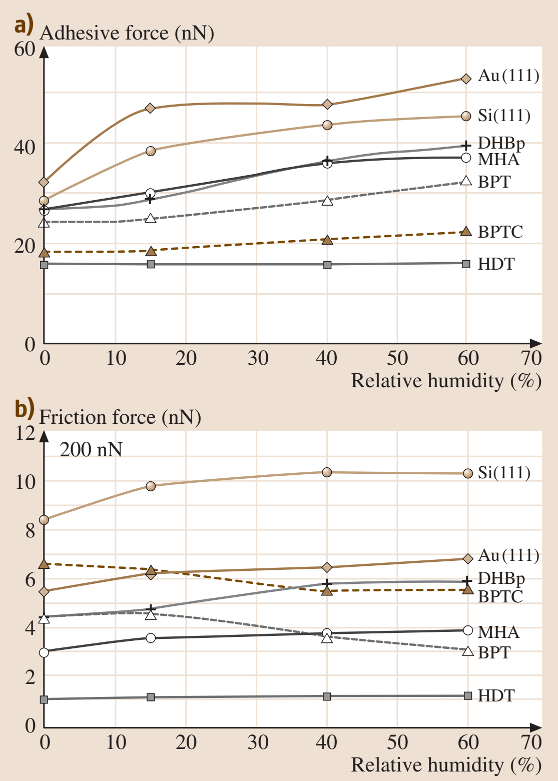

Fig. 27.16a,b The influence of relative humidity on the **(a)** adhesive force and **(b)** friction forces at 200 nN of Si(111), Au(111), and various SAMs [27.35]

Results of the influence of temperature on adhesion and friction of Si(111) and two selected typical SAMs – HDT and BPT – are presented in Fig. 27.17 [27.37]. The friction data are presented at various normal loads. In all cases, the normal load does not change the general trends. The data show that for Si(111), once the temperature is higher than 75 °C, the adhesive force decreases with temperature. The friction force of silicon decreases slightly with temperature from room temperature to 100 °C; once the temperature is higher than 100 °C, the friction force shows an apparent decrease. It is believed that desorption of the adsorbed water layer on Si(111) and the reduction of surface tension of water at high temperature are responsible for the decrease in adhesion and friction [27.77]. For HDT, in the testing range (22 °C–125 °C) its adhesion and friction show very slight change with temperature. For BPT, its adhesive force initially increases with temperature from 22 °C–75 °C, but when the temperature is higher than

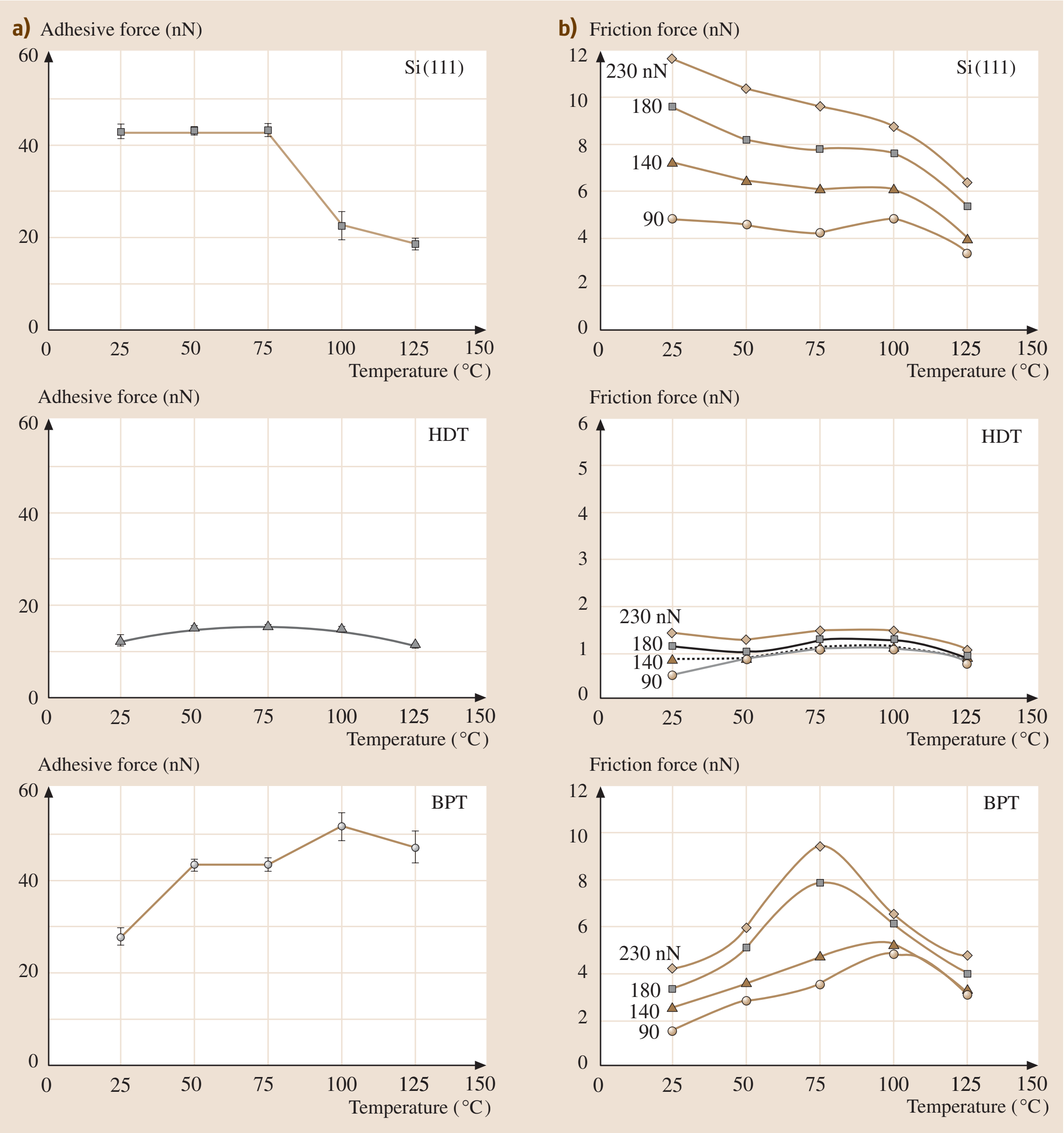

Fig. 27.17a,b The influence of temperature on **(a)** adhesive forces and **(b)** friction forces of Si(111), HDT, and BPT [27.37]

Table 27.11 Melting point of typical organic compounds similar to HDT and BPT SAMs

Compound type	Molecular formula	Melting points (°C) [27.77]
Linear carbon chain	$CH_3(CH_2)_{14}CH_2OH$	50
Benzene ring	$3\text{-}HO\text{-}C_6H_4C_6H_5$	78
	$2\text{-}HO\text{-}C_6H_4C_6H_5$	58–60

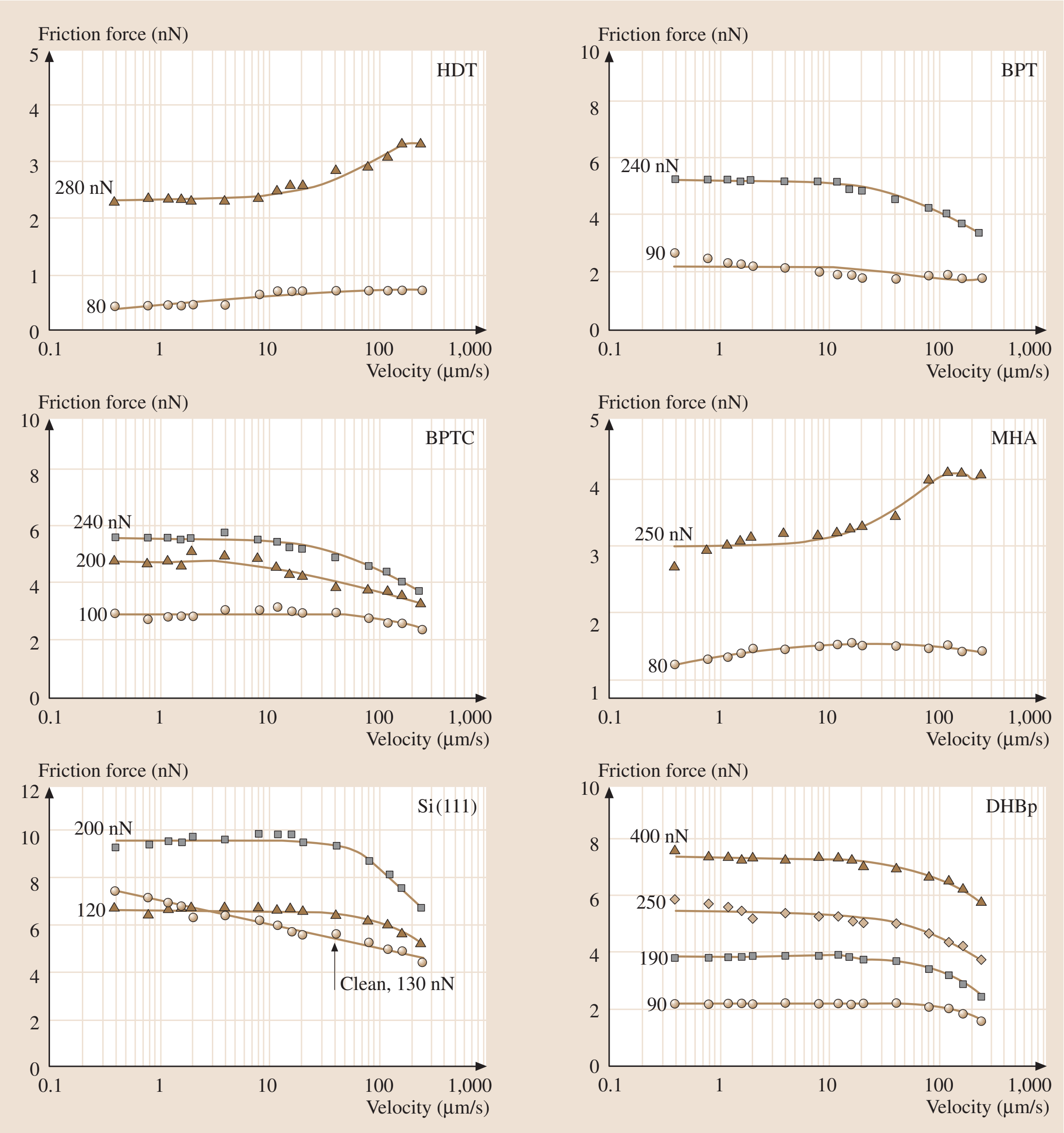

Fig. 27.18 The influence of velocity on friction forces of Si(111) and SAMs [27.37]

75 °C, the adhesive force remains steady within the experimental error. The variations in the friction force with temperature for BPT exhibit the maximum value, and the corresponding critical temperature is related to the normal load. The reason why BPT shows maximum friction force is believed to be caused by the melting of film at high temperature. When the temperature is below the melting point, an increase of temperature leads to the softening of SAM, which increases the real contact area and, consequently, the friction force. Once the temperature is higher than the melting point, the lubrication regime is changed from boundary lubrication in a solid

SAM to liquid lubrication in the melted SAM, and therefore the friction force is decreased. To date, the melting points of HDT and BPT SAMs are unknown. But from the literature [27.77], the melting points of typical linear carbon chain and benzene ring compounds are presented in Table 27.11. It is found that the reported melting of the benzene ring compounds is somewhat close to the frictional transition temperature of BPT measured by AFM. For HDT, their very compliant carbon chain can serve as an excellent lubricant even in solid state (i. e., at or near room temperature), so clear temperature effect is not observed as in the case of HDT.

Figure 27.18 shows the influence of velocity on the friction force of SAMs at different normal loads [27.37]. In all cases, normal load does not change the general trends. It is shown that for a silicon wafer, cleaned by Piranha solution just before the friction test, the friction force decreases linearly with increasing velocity. But for Si(111) without cleaning just before the AFM test, the friction force decreases only at high velocity. It is believed that this difference is caused by the adsorbed water and/or contamination on the surface during storage. Figure 27.18 also indicates that the velocity effects of SAMs depend on the molecular structures of SAMs. For SAMs that have compliant long carbon spacer chains, such as HDT and MHA, the friction force increases at high velocity, while for SAMs that have rigid biphenyl chains, such as BPT, BPTC, and DHBp, the friction force changes in the opposite way. The mechanisms responsible for the variation of the friction forces of SAMs with velocity are believed to be related to the viscoelastic properties of SAMs.

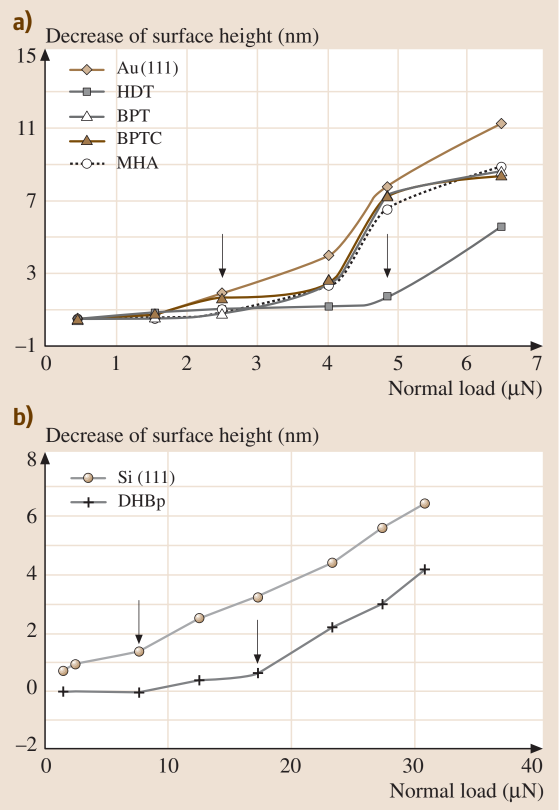

Fig. 27.19a,b Wear depth as a function of normal load after one scan cycle [27.35]. **(a)** on Au(111) and SAMs deposited on Au(111)/Si(111) **(b)** on Si(111) and DHPp deposited on Si(111)

27.3.5 Wear and Scratch Resistance of SAMs

Wear resistance was studied on an area of 1 µm × 1 µm using a diamond tip in an AFM. The variation of wear depth with normal loads is presented in Fig. 27.19 [27.35]. It clearly shows that in the whole testing range, DHBp on Si(111) exhibits much better wear resistance compared to Si(111), Au(111), and other SAMs that were deposited on the Au(111) substrate. For the SAMs deposited on Au(111), HDT exhibits the best wear resistance. For all of the tested SAMs, in the wear depth as a function of normal load curves, there appears a critical normal load, which is marked by arrows in Fig. 27.19. When the normal load is smaller than the critical normal load, the monolayer only shows slight height change in the scan areas. When the normal load is higher than the critical value, the height change of SAMs increase dramatically. Relocation and accumulation of BPT molecules have been observed during the initial several scans, which lead to the formation of a larger terrace. Wear studies of a single BPT terrace indicate that the wear life of BPT increases exponentially with terrace size [27.36, 37].

Scratch resistances of Si(111), Au(111), and SAMs were studied by a continuous AFM microscratch technique. Figure 27.20a shows coefficient of friction profiles as a function of increasing normal load and corresponding tapping mode AFM surface height images of the scratches captured on Si(111), Au(111), and SAMs [27.37]. Figure 27.20a indicates that there is an abrupt increase in the coefficients of friction for all of the tested samples. The normal load associated with this event is termed the critical load (indicated by the arrows labeled "A"). At the initial stages of the scratch, all the samples exhibit a low coefficient of friction, indicating that the friction force is dominated by the shear com-

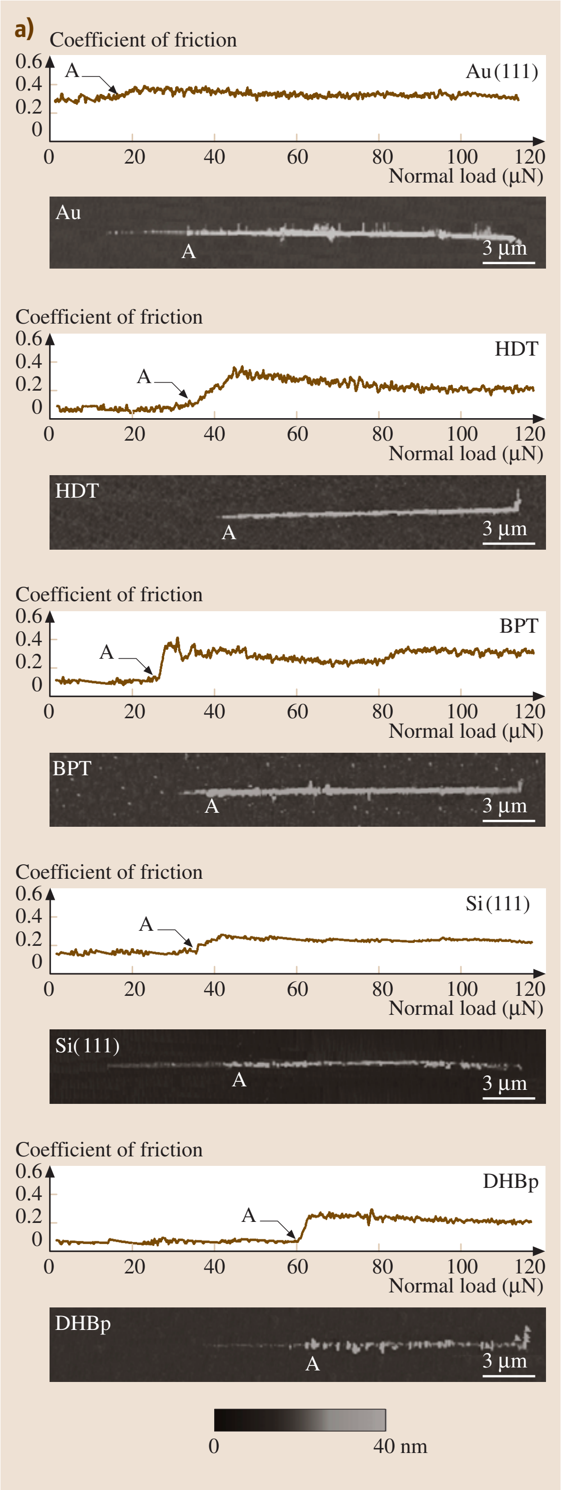

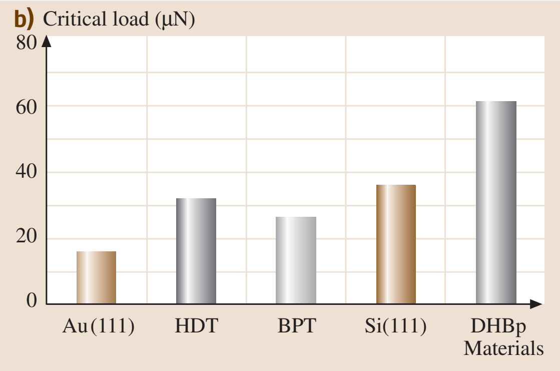

Fig. 27.20 **(a)** Coefficient of friction profiles during scratch as a function of normal load and corresponding AFM surface height images, and **(b)** critical loads estimated from the coefficient of friction profile and AFM images for Au(111), HDT/Au(111), BPT/Au(111), Si(111), and DHBp/Si(111) [27.37]

ponent. This is in agreement with analysis of the AFM images, which shows negligible damage on the surfaces prior to the critical load. At the critical load, a clear groove is formed, accompanied by the formation of material pileup at the sides of the scratch. This suggests that the initial damage that occurs at the critical load is due to ploughing associated with plastic deformation, which causes a sharp rise in the coefficient of friction. Beyond the critical load, debris can be seen in addition to material pile-up at the sides of the scratch. Figure 27.20b summarizes the critical loads for the various samples obtained in this study. It clearly indicates that all SAMs can increase the critical load of the corresponding substrate. DHBp, which is deposited on Si(111), shows superior scratch resistance in all of the tested samples.

Mechanisms responsible for a sudden drop in the decrease in surface height with an increase in load during wear and scratch tests need to be understood. *Barrena* et al. [27.78] observed that the height of self-assembled alkylsilanes decrease in discrete amounts with normal load. This step-like behavior is due to the discrete molecular tilts, which are dictated by the geometrical requirements of the close packing of molecules. Only certain angles are allowed due to the zigzag arrangement of the carbon atoms. The relative height of the monolayer under pressure can be calculated by the following equation:

$$\left(\frac{h}{L}\right) = \left[1+\left(\frac{na}{d}\right)^2\right]^{-\frac{1}{2}} , \qquad (27.6)$$

Table 27.12 Calculated $\left[1+\left(\frac{na}{d}\right)^2\right]^{-1/2}$ and measured $\left(\frac{h}{L}\right)$ relative heights of HDT and MHA self-assembled monolayers [27.35]

Steps (n)	Calculated[1] $\left[1+\left(\frac{na}{d}\right)^2\right]^{-1/2}$	Measured[2] $\left(\frac{h}{L}\right)$	
		HDT	MHA
1	0.883		
2	0.685	0.674[3]	0.643[3]
3	0.531	0.532[3]	0.552[3]
4	0.425	0.416[3]	
5	0.352	0.354[3]	
6	0.299		

[1] Calculations are based on the assumption that the molecules tilt in discrete steps (n) upon compression with a diamond AFM tip [27.78].

[2] Measured values are the mean values of three tests.

[3] These measured values correspond to the normal loads of 0.50 µN, 1.57 µN, 2.53 µN, and 4.03 µN, respectively.

where L is the total length of the molecule, h is the height of the SAMs in the tilt configuration (monolayer thickness), a is the distance between alternate carbon atoms in the molecule, d is the separation of the molecules, and n is the step number. For HDT and MHA, only the head groups are different in their work and this study. The spacer carbon chains in alkanethiol and alkylsilane are exactly the same. So the same a (0.25 nm) and d (0.47 nm) values are used in the calculation for HDT and MHA. The calculated and measured relative heights of HDT and MHA are listed in Table 27.12. When the normal loads are smaller than the critical values in Fig. 27.19, the measured relative height values of HDT and MHA are very close to the calculated values. This means that HDT and MHA underwent step tilting below critical normal loads.

Table 27.13 Calculated L_0 and measured residual film thickness for SAMs under critical load

	L_0^1 (nm)	Residual thickness[2] (nm)
HDT	0.24	0.25
BPT	0.39	0.42
BPTC	0.33	0.38
MHA	0.36	0.29
DHBp	–	0.52

[1] Calculated by the equation of $h = b\cos(\alpha)n + L_0$ [27.68]

[2] Measured by AFM using diamond tip under critical normal load. All of the data are the mean values of three tests

The residual SAMs thickness after wear under critical normal load was measured by profiling the worn film using AFM. The results are listed in Table 27.13. For an alkanethiol monolayer, the relationship between the monolayer thickness h and intercept length L_0 can be expressed as (see Fig. 27.21):

$$h = b\cos(\alpha)n + L_0\ , \tag{27.7}$$

where b is the length of the projection of the C−C bond onto the main chain axis ($b = 0.127$ nm for alkanethiol), n is the chain length defined by $CH_3(CH_2)_nSH$, and α is the tilt angle [27.68]. For BPT and BPTC, based on the same principle and using the bond lengths reported in reference [27.69], the L_0 values are also calculated, Table 27.13. It indicates that the measured residual thickness values of SAMs under critical load are very close to the calculated intercept length L_0 values. It means that under the critical normal load, the Si_3N_4 tip approaches the interface and SAMs wear severely away from the substrate. This is due to the interface chemical adsorption bond strength (HS−Au and Si−O) being generally smaller than the other chemical bond strengths in SAMs

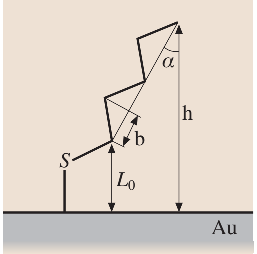

Fig. 27.21 Illustration of the relationship between the components of the equation $h = b\cos(\alpha)n + L_0$ [27.35]

Table 27.14 Bond strengths of the chemical bonds in SAMs [27.35]

SAMs	Bond	Bond strength[1] (kJ/mol)
Thiol on Au	S−Au	184 [27.58]
Hydroxyl on Si	O−Si	242.7 [27.79]
HDT	$H-CH_2$	464.8
	H−CH	421.7
	CH_3-t-C_4H_9	425.9±8
	t-C_4H_9−SH	286.2 ± 6.3
MHA	O=CO	532.2 ± 0.4
	$H-OCOC_2H_5$	445.2±8
	$HO-OCH_2C(CH_3)$	193.7 ± 7.9
	$C_6H_5CH_2-COOH$	280
BPT	CH_3-CH_3	376.0 ± 2.1
	$H_2C=CH_2$	733±8
	C_6H_5-SH	361.9±8
DHBp	C_6H_5-OH	361.9±8

[1] Most of the data are cited from [27.77], except where indicated. For MHA and DHBp, the bonds that are common as in HDT and DHBp are not repeated

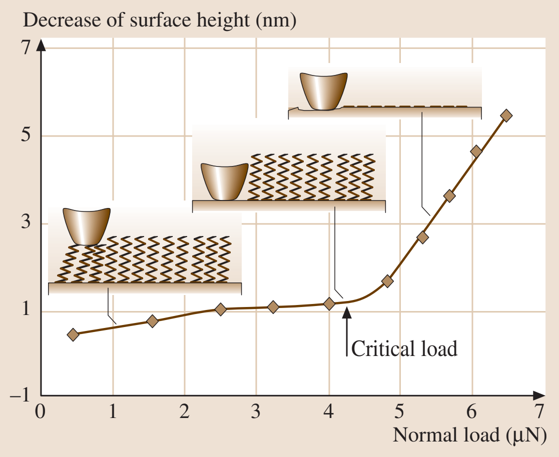

Fig. 27.22 Illustration of the wear mechanisms of SAMs with increasing normal load [27.37]

spacer chains (see Table 27.14). Based on this discussion, it is believed that the reason why DHBp has the best wear resistance is due to the rigid biphenyl ring structure (compared to linear carbon chain in alkylthiol), the hard Si(111) substrate (compared to Au(111) substrate), and the strong interface Si−O bond strength (compared to the weak S−Au bond strength in the other SAMs, see Table 27.14).

According to the wear and scratch results reported here and the above discussion, the transition of the wear mechanisms of SAMs with increasing normal load is illustrated in Fig. 27.22. Below the critical normal load, SAMs undergo step orientation, at the critical load SAMs wear away from the substrate due to the weak interface bond strengths, while above the critical normal load severe wear takes place on the substrate. In order to improve wear resistance, the interface bond must be enhanced; a rigid spacer chain and a hard substrate are also preferred.

27.4 Closure

Exposure of devices to humid environments results in condensates of water vapor from the environment. Condensed water, or a preexisting film of liquid, forms concave meniscus bridges between the mating surfaces. The negative Laplace pressure present in the meniscus results in an adhesive force that depends on the interface roughness, surface tension, and contact angle. The adhesive force can be significant in an interface with ultra-smooth surfaces, and it can be on the same order as the external load if the latter is small, such as in micro/nanodevices. Surfaces with high hydrophobicity can be produced by surface treatment. In many applications, hydrophobic films are expected to provide low adhesion, friction, and wear. To minimize high adhesion, friction, and/or because of small clearances in micro/nanodevices, these films should be molecularly thick. Liquid films of low surface tension or certain hydrophobic solid films can be used. Ordered molecular assemblies with high hydrophobicity can be engineered using chemical grafting of various polymer molecules with suitable functional head groups and nonpolar surface terminal groups.

The adhesion, friction, and wear properties of SAMs, having alkyl and biphenyl spacer chains with different surface terminal groups ($-CH_3$, $-COOH$, and $-OH$) and head groups ($-SH$ and $-OH$), studied using an AFM, are reported in this chapter. It is found that the adhesive force varies linearly with the W_a value of SAMs, which indicates that capillary condensation of water plays an important role to the adhesion of SAMs on nanoscale at ambient conditions. It has been found that HDT exhibits the smallest adhesive force and friction force, because of low W_a of $-CH_3$ surface terminal groups, and high-compliance long carbon spacer chain. The friction data are explained using a molecular spring model, in which the local stiffness and intermolecular force governs its frictional performance. The results of the stiffness and friction characterization of the micropatterned sample with different structures support this model. The influence of relative humidity on adhesion and friction of SAMs is dominated by the thickness of the adsorbed water layer. At higher humidity, water can either increase friction through increased adhesion by meniscus effect in the contact zone or reduce friction through an enhanced water-lubricating effect. In the case of Si(111), the desorption of the adsorbed water layer and reduction of surface tension of water with increasing temperature reduce the adhesive force and friction force. The increase of temperature does not show an apparent influence on HDT, but strongly influences the adhesion and friction properties of BPT, which are believed to be related to its melting. The effect of velocity on friction properties of SAMs depends on their molecular structures. For alkylthiol SAMs, their friction is increased by increasing velocity, while for biphenyl SAMs it changes in the opposite way. The mechanism responsible for the variation of the friction forces of SAMs with velocity is believed to be related to their viscoelastic properties. Wear and continuous microscratch tests show that among the SAMs on Au(111), HDT exhibits the best wear resistance. DHBp on Si(111), due to its rigid biphenyl spacer chains, strong interface bonds, and a hard substrate, has the best wear resistance among all of the tested samples. For all of the SAMs, the wear depth as a function of normal load curves shows critical normal loads. Below the critical normal load SAMs undergo step orientation, at the critical load SAMs wear away from the substrate due to the weak interface bond strengths, while above the critical normal load severe wear take place on the substrate.

Based on the nanotribological studies of SAM films by AFM, they exhibit attractive hydrophobic and tribological properties. SAM films should find many tribological applications, including in micro/nanodevices.

References

27.1 B. Bhushan, J. N. Israelachvili, U. Landman: Nanotribology: Friction, wear and lubrication at the atomic scale, Nature **374** (1995) 607–616

27.2 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)

27.3 B. Bhushan: *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)

27.4 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

27.5 K. F. Man, B. H. Stark, R. Ramesham: *A Resource Handbook for MEMS Reliability* (California Institute of Technology, Pasadena 1998)

27.6 S. Kayali, R. Lawton, B. H. Stark: MEMS reliability assurance activities at JPL, EEE Links **5** (1999) 10–13

27.7 D. M. Tanner, N. F. Smith, L. W. Irwin et al.: *MEMS Reliability: Infrastructure, Test Structure, Experiments, and Failure Modes* (Sandia National Laboratories, Albuquerque 2000)

27.8 S. Arney: Designing for MEMS reliability, MRS Bull. **26** (2001) 296–299

27.9 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

27.10 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

27.11 A. Ulman (Ed.): *Characterization of Organic Thin Films* (Butterworth-Heineman, Boston 1995)

27.12 B. Bhushan: Contact mechanics of rough surfaces in tribology: Multiple asperity contact, Tribol. Lett. **4** (1998) 1–35

27.13 B. Bhushan, W. Peng: Contact mechanics of multilayered rough surfaces, Appl. Mech. Rev. **55** (2002) 435–480

27.14 R. Maboudian: Surface processes in MEMS technology, Surf. Sci. Rep. **30** (1998) 209–269

27.15 M. Scherge, J. A. Schaefer: Surface modification and mechanical properties of bulk silicon. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 529–537

27.16 F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids, Part I* (Clarendon, Oxford 1950)

27.17 V. N. Koinkar, B. Bhushan: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy, J. Vac. Sci. Technol. A **14** (1996) 2378–2391

27.18 B. Bhushan, Z. Zhao: Macro- and microscale tribological studies of molecularly-thick boundary layers of perfluoropolyether lubricants for magnetic thin-film rigid disks, J. Info. Storage Proc. Syst. **1** (1999) 1–21

27.19 H. Liu, B. Bhushan: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM, Ultramicroscopy **97** (2003) 321–340

27.20 W. A. Zisman: Friction, durability and wettability properties of monomolecular films on solids. In: *Friction and Wear*, ed. by R. Davies (Elsevier, Amsterdam 1959) pp. 110–148
27.21 M. E. Schrader, G. I. Loeb (Eds.): *Modern Approaches to Wettability* (Plenum, New York 1992)
27.22 A. W. Neumann, J. K. Spelt (Eds.): *Applied Surface Thermodynamics* (Dekker, New York 1996)
27.23 H. Hansma, F. Motamedi, P. Smith, P. Hansma, J. C. Wittman: Molecular resolution of thin, highly oriented poly(tetrafluoroethylene) films with the atomic force microscope, Polymer Commun. **33** (1992) 647–649
27.24 L. Scandella, A. Schumacher, N. Kruse, R. Prins, E. Meyer, R. Luethi, L. Howald, H. J. Guentherodt: Tribology of ultra-thin MoS_2 platelets on mica: Studies by scanning force microscopy, Thin Solid Films **240** (1994) 101–104
27.25 B. Bhushan: Chemical, mechanical and tribological characterization of ultra-thin and hard amorphous carbon coatings as thin as 3.5 nm: Recent developments, Diam. Relat. Mater. **8** (1999) 1985–2015
27.26 A. Erdemir, C. Donnet: Tribology of diamond, diamond-like carbon, and related films. In: *Modern Tribology Handbook Vol. 2: Materials, Coatings, and Industrial Applications*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 871–908
27.27 V. F. Dorfman: Diamond-like nanocomposites (DLN), Thin Solid Films **212** (1992) 267–273
27.28 M. Grischke, K. Bewilogua, K. Trojan, H. Dimigan: Application-oriented modification of deposition process for diamond-like carbon based coatings, Surf. Coat. Technol. **74-75** (1995) 739–745
27.29 R. S. Butter, D. R. Waterman, A. H. Lettington, R. T. Ramos, E. J. Fordham: Production and wetting properties of fluorinated diamond-like carbon coatings, Thin Solid Films **311** (1997) 107–113
27.30 M. Grischke, A. Hieke, F. Morgenweck, H. Dimigan: Variation of the wettability of DLC coatings by network modification using silicon and oxygen, Diam. Relat. Mater. **7** (1998) 454–458
27.31 C. Donnet, J. Fontaine, A. Grill, V. Patel, C. Jahnes, M. Belin: Wear-resistant fluorinated diamondlike carbon films, Surf. Coat. Technol. **94-95** (1997) 531–536
27.32 D. J. Kester, C. L. Brodbeck, I. L. Singer, A. Kyriakopoulos: Sliding wear behavior of diamond-like nanocomposite coatings, Surf. Coat. Technol. **113** (1999) 268–273
27.33 H. Liu, B. Bhushan: Adhesion and friction studies of microelectromechanical systems/nanoelectromechanical systems materials using a novel microtriboapparatus, J. Vac. Sci. Technol. A **21** (2003) 1528–1538
27.34 B. Bhushan, A. V. Kulkarni, V. N. Koinkar, M. Boehm, L. Odoni, C. Martelet, M. Belin: Microtribological characterization of self-assembled and Langmuir–Blodgett monolayers by atomic and friction force microscopy, Langmuir **11** (1995) 3189–3198
27.35 B. Bhushan, H. Liu: Nanotribological properties and mechanisms of alkylthiol and biphenyl thiol self-assembled monolayers studied by atomic force microscopy, Phys. Rev. B **63** (2001) 245412-1–245412-11
27.36 H. Liu, B. Bhushan, W. Eck, V. Stadler: Investigation of the adhesion, friction, and wear properties of biphenyl thiol self-assembled monolayers by atomic force microscopy, J. Vac. Sci. Technol. A **19** (2001) 1234–1240
27.37 H. Liu, B. Bhushan: Investigation of nanotribological properties of alkylthiol and biphenyl thiol self-assembled monolayers, Ultramicroscopy **91** (2002) 185–202
27.38 H. Liu, B. Bhushan: Orientation and relocation of biphenyl thiol self-assembled monolayers, Ultramicroscopy **91** (2002) 177–183
27.39 A. Ulman: *An Introduction to Ultrathin Organic Films: From Langmuir–Blodgett to Self-Assembly* (Academic, San Diego 1991)
27.40 A. Ulman: Formation and structure of self-assembled monolayers, Chem. Rev. **96** (1996) 1533–1554
27.41 J. A. Zasadzinski, R. Viswanathan, L. Madsen, J. Garnaes, D. K. Schwartz: Langmuir–Blodgett films, Science **263** (1994) 1726–1733
27.42 J. Tian, Y. Xia, G. M. Whitesides: Microcontact printing of SAMs. In: *Thin Films – Self-Assembled Monolayers of Thiols*, Vol. 24, ed. by A. Ulman (Academic, San Diego 1998) pp. 227–254
27.43 Y. Xia, G. M. Whitesides: Soft lithography, Angew. Chem. Int. Edn. **37** (1998) 550–575
27.44 A. Kumar, G. M. Whitesides: Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol ink followed by chemical etching, Appl. Phys. Lett. **63** (1993) 2002–2004
27.45 S. Y. Chou, P. R. Krauss, P. J. Renstrom: Imprint lithography with 25-nanometer resolution, Science **272** (1996) 85–87
27.46 Y. Xia, E. Kim, X. M. Zhao, J. A. Rogers, M. Prentiss, G. M. Whitesides: Complex optical surfaces formed by replica molding against elastomeric masters, Science **273** (1996) 347–349
27.47 R. F. Service: Self-assembly comes together, Science **265** (1994) 316–318
27.48 M. Hein, L. R. Best, S. Pattison, S. Arena: *Introduction to General, Organic, and Biochemistry*, 6th edn. (Brooks/Cole, Pacific Grove 1997)
27.49 J. R. Mohrig, C. N. Hammond, T. C. Morrill, D. C. Neckers: *Experimental Organic Chemistry* (Freeman, New York 1998)
27.50 S. R. Wasserman, Y. T. Tao, G. M. Whitesides: Structure and reactivity of alkylsiloxane monolayers

formed by reaction of alkylchlorosilanes on silicon substrates, Langmuir **5** (1989) 1074–1089

27.51 C. Jung, O. Dannenberger, Y. Xu, M. Buck, M. Grunze: Self-assembled monolayers from organosulfur compounds: A comparison between sulfides, disulfides, and thiols, Langmuir **14** (1998) 1103–1107

27.52 W. Geyer, V. Stadler, W. Eck, M. Zharnikov, A. Golzhauser, M. Grunze: Electron-induced crosslinking of aromatic self-assembled monolayers: Negative resists for nanolithography, Appl. Phys. Lett. **75** (1999) 2401–2403

27.53 J. Ruhe, V. J. Novotny, K. K. Kanazawa, T. Clarke, G. B. Street: Structure and tribological properties of ultrathin alkylsilane films chemisorbed to solid surfaces, Langmuir **9** (1993) 2383–2388

27.54 V. DePalma, N. Tillman: Friction and wear of self-assembled tricholosilane monolayer films on silicon, Langmuir **5** (1989) 868–872

27.55 E. Ando, Y. Goto, K. Morimoto, K. Ariga, Y. Okahata: Frictional properties of monolayers of silane compounds, Thin Solid Films **180** (1989) 287–291

27.56 M. T. McDermott, J. B. D. Green, M. D. Porter: Scanning force microscopic exploration of the lubrication capabilities of n-alkanethiolate monolayers chemisorbed at gold: Structural basis of microscopic friction and wear, Langmuir **13** (1997) 2504–2510

27.57 X. Xiao, J. Hu, D. H. Charych, M. Salmeron: Chain length dependence of the frictional properties of alkylsilane molecules self-assembled on mica studied by atomic force microscopy, Langmuir **12** (1996) 235–237

27.58 A. Lio, D. H. Charych, M. Salmeron: Comparative atomic force microscopy study of the chain length dependence of frictional properties of alkanethiol on gold and alkylsilanes on mica, J. Phys. Chem. B **101** (1997) 3800–3805

27.59 H. Schonherr, G. J. Vancso: Tribological properties of self-assembled monolayers of fluorocarbon and hydrocarbon thiols and disulfides on Au(111) studied by scanning force microscopy, Mater. Sci. Eng. C **8–9** (1999) 243–249

27.60 H. I. Kim, T. Koini, T. R. Lee, S. S. Perry: Systematic studies of the frictional properties of fluorinated monolayers with atomic force microscopy: Comparison of CF_3- and CH_3-terminated groups, Langmuir **13** (1997) 7192–7196

27.61 H. I. Kim, M. Graupe, O. Oloba, T. Koini, S. Imaduddin, T. R. Lee, S. S. Perry: Molecularly specific studies of the frictional properties of monolayer films: A systematic comparison of CF_3-, $(CH_3)_2CH$-, and CH_3-terminated films, Langmuir **15** (1999) 3179–3185

27.62 V. V. Tsukruk, M. P. Everson, L. M. Lander, W. J. Brittain: Nanotribological properties of composite molecular films: C_{60} anchored to a self-assembled monolayer, Langmuir **12** (1996) 3905–3911

27.63 S. Lee, Y. S. Shon, T. R. Lee, S. S. Perry: Structural characterization and frictional properties of C_{60}-terminated self-assembled monolayers on Au(111), Thin Solid Films **358** (2000) 152–158

27.64 V. V. Tsukruk, V. N. Bliznyuk: Adhesive and friction forces between chemically modified silicon and silicon nitride surfaces, Langmuir **14** (1998) 446–455

27.65 V. V. Tsukruk, T. Nguyen, M. Lemieux, J. Hazel, W. H. Weber, V. V. Shevchenko, N. Klimenko, E. Sheludko: Tribological properties of modified MEMS surfaces. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 607–614

27.66 M. Fujihira, Y. Tani, M. Furugori, U. Akiba, Y. Okabe: Chemical force microscopy of self-assembled monolayers on sputtered gold films patterned by phase separation, Ultramicroscopy **86** (2001) 63–73

27.67 S. Sundararajan, B. Bhushan: Topography-induced contribution to friction forces measured using an atomic force/friction force microscopy, J. Appl. Phys. **88** (2000) 4825–4831

27.68 Y. F. Miura, M. Takenga, T. Koini, M. Graupe, N. Garg, R. L. Graham, T. R. Lee: Wettability of self-assembled monolayers generated from CF_3-terminated alkanethiols on gold, Langmuir **14** (1998) 5821–5825

27.69 M. Ratajczak-Sitarz, A. Katrusiak, Z. Kaluski, J. Garbarczyk: 4,4′-biphenyldithiol, Acta Crystallogr. C **43** (1987) 2389–2391

27.70 J. N. Israelachvili: *Intermolecular and Surface Forces*, 2nd edn. (Academic, London 1992)

27.71 R. J. Good, C. J. V. Oss: *Modern Approaches to Wettability – Theory and Applications* (Plenum, New York 1992)

27.72 Y. Unno, H. Kawamura, H. Kita, S. Sekiyama: Friction characteristics of sintered Si_3N_4 in oil-lubricated medium, Proc. Int. Symp. Canadian Institute of Mining, Metallurgy and Petroleum, Montreal 1995 (Canadian Institute of Mining, Metallurgy and Petroleum, Montreal 1995) 275–284

27.73 M. H. V. C. Adao, B. J. V. Saramago, A. C. Fernandes: Estimation of the surface properties of styrene-acrylonitrile random copolymers from contact angle measurements, J. Coll. Interface Sci. **217** (1999) 94–106

27.74 J. I. Siepman, I. R. McDonald: Monte Carlo simulation of the mechanical relaxation of a self-assembled monolayer, Phys. Rev. Lett. **70** (1993) 453–456

27.75 M. Garcia-Parajo, C. Longo, J. Servat, P. Gorostiza, F. Sanz: Nanotribological properties of octadecyltrichlorosilane self-assembled ultrathin films studied by atomic force microscopy: Contact and tapping modes, Langmuir **13** (1997) 2333–2339

27.76 D. DeVecchio, B. Bhushan: Localized surface elasticity measurements using an atomic force microscope, Rev. Sci. Instrum. **68** (1997) 4498–4505

27.77 D. R. Lide: *CRC Handbook of Chemistry and Physics*, 75th edn. (CRC, Boca Raton 1994)

27.78 E. Barrena, S. Kopta, D. F. Ogletree, D. H. Charych, M. Salmeron: Relationship between friction and molecular structure: Alkysilane lubricant films under pressure, Phys. Rev. Lett. **82** (1999) 2880–2883

27.79 T. Hoshino, C. Yayoi, K. Inage: Adsorption of atomic and molecular oxygen and desorption of silicon monoxide on Si(111) surface, Phys. Rev. B **59** (1999) 2332–2340

28. Nanoscale Boundary Lubrication Studies

Boundary films are formed by physisorption, chemisorption, and chemical reaction. With physisorption, no exchange of electrons takes place between the molecules of the adsorbate and those of the adsorbant. The physisorption process typically involves van der Waals forces, which are relatively weak. In chemisorption, there is an actual sharing of electrons or electron interchange between the chemisorbed species and the solid surface. The solid surfaces bond very strongly to the adsorption species through covalent bonds. Chemically reacted films are formed by the chemical reaction of a solid surface with the environment. The physisorbed film can be either monomolecularly or polymolecularly thick. The chemisorbed films are monomolecular, but stoichiometric films formed by chemical reaction can have a large film thickness. In general, the stability and durability of surface films decrease in the following order: chemically reacted films, chemisorbed films, and physisorbed films. A good boundary lubricant should have a high degree of interaction between its molecules and the sliding surface. As a general rule, liquids are good lubricants when they are polar and, thus, able to grip solid surfaces (or be adsorbed). In this chapter, we focus on PFPEs. We first introduce details of the commonly used PFPE lubricants; then present a summary of nanodeformation, molecular conformation, and lubricant spreading studies; followed by an overview of nanotribological properties of polar and nonpolar PFPEs studied by atomic force microscopy (AFM) and some concluding remarks.

28.1 **Lubricants Details** 862
28.2 **Nanodeformation, Molecular Conformation, and Lubricant Spreading** . 864
28.3 **Boundary Lubrication Studies** 866
    28.3.1 Friction and Adhesion................. 866
    28.3.2 Rest Time Effect.......................... 869
    28.3.3 Velocity Effect 871
    28.3.4 Relative Humidity and Temperature Effect................ 873
    28.3.5 Tip Radius Effect 876
    28.3.6 Wear Study................................ 879
28.4 **Closure** 880
References 881

Boundary films are formed by physisorption, chemisorption, and chemical reaction. With physisorption, no exchange of electrons takes place between the molecules of the adsorbate and those of the adsorbant. The physisorption process typically involves van der Waals forces, which are relatively weak. In chemisorption, there is an actual sharing of electrons or electron interchange between the chemisorbed species and the solid surface. The solid surfaces bond very strongly to the adsorption species through covalent bonds. Chemically reacted films are formed by the chemical reaction of a solid surface with the environment. The physisorbed film can be either monomolecularly or polymolecularly thick. The chemisorbed films are monomolecular, but stoichiometric films formed by chemical reaction can have a large film thickness. In general, the stability and durability of surface films decrease in the following order: chemically reacted films, chemisorbed films, and physisorbed films. A good boundary lubricant should have a high degree of interaction between its molecules and the sliding surface. As a general rule, liquids are good lubricants when they are polar and, thus, able to grip solid surfaces (or be adsorbed). Polar lubricants contain reactive functional groups with low ionization potential, or groups having high polarizability [28.1–3]. Boundary lubrication properties of lubricants are also dependent upon the molecular conformation and lubricant spreading [28.4–7].

Self-assembled monolayers (SAMs), Langmuir–Blodgett (LB) films, and perfluoropolyether (PFPE) films can be used as boundary lubricants [28.2,3,8–10].

PFPE films are commonly used for lubrication of magnetic rigid disks and metal evaporated magnetic tapes to reduce friction and wear of a head-medium interface [28.10]. PFPEs are well suited for this application because of the following properties: low surface tension and low contact angle, which allow easy spreading on surfaces and provide a hydrophobic property; chemical and thermal stability, which minimizes degradation under use; low vapor pressure,which provides low outgassing; high adhesion to substrate via organofunctional bonds; and good lubricity, which reduces the interfacial friction and wear [28.10–12]. While the structure of the lubricants employed at the head-medium interface has not changed substantially over the past decade, the thickness of the PFPE film used to lubricate the disk has steadily decreased from multilayer thicknesses to the sub-monolayer thickness regime [28.11, 13]. Molecularly thick PFPE films are also being considered for lubrication purposes of the evolving microelectromechanical systems (MEMS) industry [28.14]. It is well-known that the properties of molecularly thick liquid films confined to solid surfaces can be dramatically different from those of the corresponding bulk liquid. In order to efficiently develop lubrication systems that meet the requirements of the advanced rigid disk drive and MEMS industries, the impact of thinning the PFPE lubricants on the resulting of nanotribology should be fully understood [28.15, 16]. It is also important to understand lubricant-substrate interfacial interactions and the influence of the operating environment on the nanotribological performance of molecularly thick PFPEs.

An overview of nanotribological properties of SAMs and LB films can be found in many references such as [28.17]. In this chapter, we focus on PFPEs. We first introduce details of the commonly used PFPE lubricants; then present a summary of nanodeformation, molecular conformation, and lubricant spreading studies; followed by an overview of nanotribological properties of polar and nonpolar PFPEs studied by atomic force microscopy (AFM) and some concluding remarks.

28.1 Lubricants Details

Properties of two commonly used PFPE lubricants (Z-15 and Z-DOL) are reviewed here. Their molecular structures are shown schematically in Fig. 28.1. Z-15 has nonpolar $-CF_3$ end groups, whereas Z-DOL is a polar lubricant with hydroxyl ($-OH$) end groups. Their typical properties are summarized in Table 28.1. It shows that Z-15 and Z-DOL almost have the same density and surface tension. But Z-15 has larger molecular weight and higher viscosity. Both have low surface tension, low vapor pressure, low evaporation weight loss, and good oxidative stability [28.10, 12]. Generally, single-crystal Si(100) wafer with a native oxide layer was used as a substrate for deposition of molecularly thick lubricant films

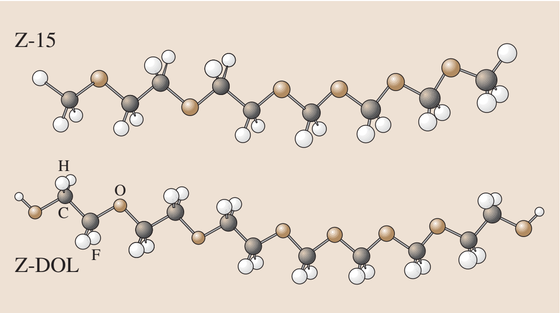

Fig. 28.1 Schematics of the molecular structures of Z-15 and Z-DOL. In this figure the m/n value, shown in Table 28.1, equals 2/3

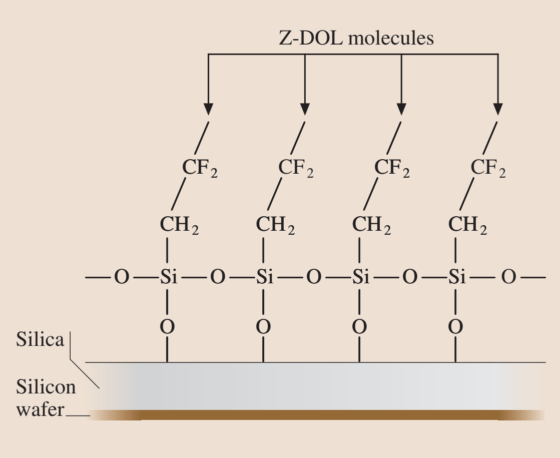

Fig. 28.2 Schematic of Z-DOL molecules that are chemically bonded on Si(100) substrate surface (which has native oxide) after thermal treatment at 150 °C for 30 min

Table 28.1 Typical properties of Z-15 and Z-DOL (data from Montefluous S.p.A., Milan, Italy)

	Z-15	**Z-DOL (2000)**
Formula	$CF_3{-}O{-}(CF_2{-}CF_2{-}O)_m{-}(CF_2{-}O)_n{-}CF_3$*	$HO{-}CH_2{-}CF_2{-}O{-}(CF_2{-}CF_2{-}O)_m{-}(CF_2{-}O)_n{-}CF_2{-}CH_2{-}OH$*
Molecular weight (Daltons)	9,100	2,000
Density *(ASTM D891)* 20 °C (g/cm^3)	1.84	1.81
Kinematic viscosity *(ASTM D445)* (cSt)		
20 °C	148	85
38 °C	90	34
99 °C	25	–
Viscosity index *(ASTM D2270)*	320	–
Surface tension *(ASTM D1331)* (dyn/cm) 20 °C	24	24
Vapor pressure (torr)		
20 °C	1.6×10^{-6}	2×10^{-5}
100 °C	1.7×10^{-5}	6×10^{-4}
Pour point *(ASTM D972)*		
°C	−80	–
Evaporation weight loss *(ASTM D972)*		
149 °C, 22 h (%)	0.7	–
Oxidative stability (°C)	–	320
Specific heat (cal/g °C)		
38 °C	0.21	–
* $m/n \sim 2/3$		

for nanotribological characterization. Z-15 and Z-DOL films can be deposited directly on the Si(100) wafer by dip coating technique. The clean silicon wafer is vertically submerged into a dilute solution of lubricant in hydrocarbon solvent (HT-70) for a certain time. The silicon wafers are vertically pulled up from the solution with a motorized stage at a constant speed for deposition of desired thicknesses of Z-15 and Z-DOL lubricants. The lubricant film thickness obtained in dip coating is a function of the concentration and pulling-up speed, among other factors. The Z-DOL film is bonded to the silicon substrate by heating the as-deposited Z-DOL samples in an oven at 150 °C for about 30 s. The native oxide layer of Si(100) wafer reacts with the −OH groups of the lubricants during thermal treatment [28.18–21]. Subsequently, fluorocarbon solvent (FC-72) washing of the thermal treated specimen removes loosely absorbed species, leaving chemically bonded phase on the substrate. The chemical bonding between Z-DOL molecules and silicon substrate is illustrated in Fig. 28.2. The bonded and washed Z-DOL film is referred to as Z-DOL(BW) in this chapter. The as-deposited Z-15 and Z-DOL films are mobile phase lubricants (i. e., liquid-like lubricant), whereas the Z-DOL(BW) films are fully bonded soft solid phase (i. e., solid-like) lubricants. This will be further discussed in the next section.

28.2 Nanodeformation, Molecular Conformation, and Lubricant Spreading

Nanodeformation behavior of Z-DOL lubricants was studied using an AFM by *Blackman* et al. [28.22,23]. Before bringing a tungsten tip into contact with a molecular overlayer, it was brought into contact with a bare clean-silicon surface, Fig. 28.3. As the sample approaches the tip, the force initially is zero, but at point A the force suddenly becomes attractive (top curve), which increases until at point B, where the sample and tip come into intimate contact and the force becomes repulsive. As the sample is retracted, a pull-off force of 5×10^{-8} N (point D) is required to overcome adhesion between the tungsten tip and the silicon surface. When an AFM tip is brought into contact with an unbonded Z-DOL film, a sudden jump into adhesive contact is also observed. A much larger pull-off force is required to overcome the adhesion. The adhesion is initiated by the formation of a lubricant meniscus surrounding the tip. This suggests that the unbonded Z-DOL lubricant shows liquid-like behavior. However, when the tip was brought into contact with a lubricant film, which was firmly bonded to the surface, the liquid-like behavior disappears. The initial attractive force (point A) is no longer sudden, as with the liquid film, but, rather, gradually increases as the tip penetrates the film.

According to *Blackman* et al. [28.22, 23], if the substrate and tip were infinitely hard with no compliance and/or deformation in the tip and sample supports, the line for B to C would be vertical with an infinite slope. The tangent to the force-distance curve at a given point is referred to as the stiffness at that point and was determined by fitting a least-squares line through the nearby data points. For silicon, the deformation is reversible (elastic), since the retracting (outgoing) portion of the curve (C to D) follows the extending (ingoing) portion (B to C). For bonded lubricant film, at the point where slope of the force changes gradually from attractive to repulsive, the stiffness changes gradually, indicating compression of the molecular film. As the load is increased, the slope of the repulsive force eventually approaches that of the bare surface. The bonded film was found to respond elastically up to the highest loads of 5 µN that could be applied. Thus, bonded lubricant behaves as a soft polymer solid.

Figure 28.4 illustrates two extremes for the conformation on a surface of a linear liquid polymer without any reactive end groups and at submonolayer coverages [28.4, 6]. At one extreme, the molecules lie flat on the surface, reaching no more than their chain diameter δ above the surface. This would be the case if a strong attractive interaction exists between the molecules and the solid. On the other extreme, when a weak attraction exists between polymer segments and the solid, the molecules adopt conformation close to that of the molecules in the bulk, with the microscopic thickness equal to about the radius of gyration R_g. *Mate* and *Novotny* [28.6] used AFM to study conformation of 0.5–1.3 nm-thick Z-15 molecules on clean Si(100) surfaces. They found that the thickness measured by AFM is thicker than that measured by ellipsometry with the offset ranging from 3–5 nm. They found that the offset was the same for very thin submonolayer coverages. If the coverage is submonolayer and inadequate to make a liquid film, the relevant thickness is then the height (h_e) of the molecules extended above the solid surface. The offset should be equal to $2h_e$, assum-

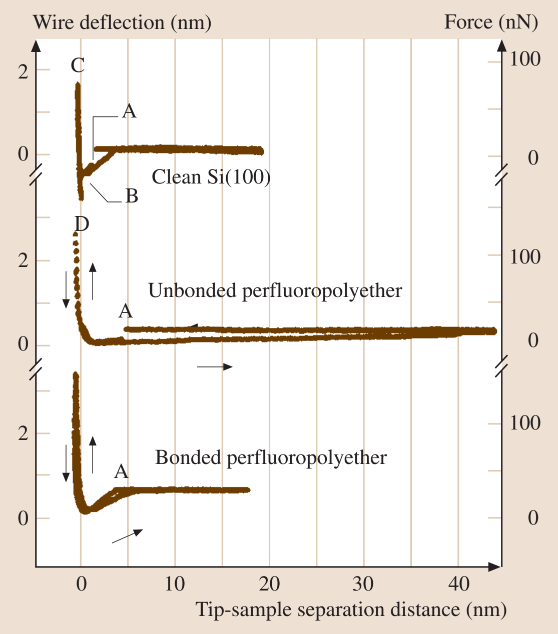

Fig. 28.3 Wire deflection (normal load) as a function of tip-sample separation distance curves comparing the behavior of clean Si(100) surface to a surface lubricated with free and unbonded PFPE lubricant, and a surface where the PFPE lubricant film was thermally bonded to the surface [28.22]

ing that the molecules extend the same height above both the tip and silicon surfaces. They therefore concluded that the molecules do not extend more than 1.5–2.5 nm above a solid or liquid surface, much smaller than the radius of gyration of the lubricants ranging between 3.2 and 7.3 nm, and to the approximate cross-sectional diameter of 0.6–0.7 nm for the linear polymer chain. Consequently, the height that the molecules extend above the surface is considerably less than the diameter of gyration of the molecules and only a few molecular diameters in height, implying that the physisorbed molecules on a solid surface have an extended, flat conformation. They also determined the disjoining pressure of these liquid films from AFM measurements of the distance needed to break the liquid meniscus that forms between the solid surface and the AFM tip. (Also see [28.7].) For a monolayer thickness of about 0.7 nm, the disjoining pressure is about 5 MPa, indicating strong attractive interaction between the liquid molecules and the solid surface. The disjoining pressure decreases with increasing film thickness in a manner consistent with a strong attractive van der Waals interaction between the liquid molecules and the solid surface.

Rheological characterization shows that the flow activation energy of PFPE lubricants is weakly dependent on chain length and is strongly dependent on the functional end groups [28.25]. PFPE lubricant films that contain polar end groups have lower mobility than those with nonpolar end groups of similar chain length [28.26]. The mobility of PFPE also depends on the surface chemical properties of the substrate. The spreading of Z-DOL on amorphous carbon surface has been studied as a function of hydrogen or nitrogen content in the carbon film, using scanning microellipsometry [28.24]. The diffusion coefficient data presented in Fig. 28.5 is thickness-dependent. It shows that the surface mobility of Z-DOL increased as the hydrogen content increased, but decreased as nitrogen content increased. The enhancement of Z-DOL surface mobility by hydrogenation may be understood from the fact that the interactions between Z-DOL molecules and the carbon surface can be significantly weakened, due to a reduction of the number of high-energy binding sites on the carbon surface. The stronger interactions between the Z-DOL molecules and carbon surface, as nitrogen content in the carbon coating increases, leads to the lowering of Z-DOL surface mobility.

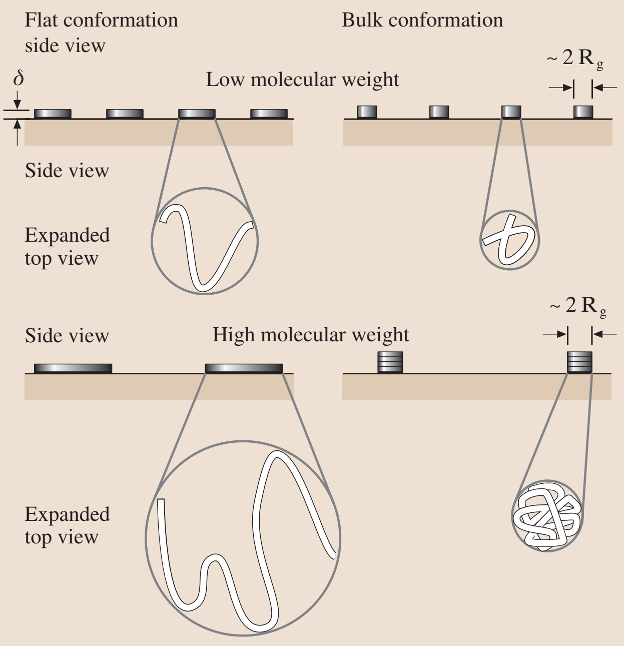

Fig. 28.4 Schematic representation of two extreme liquid conformations at the surface of the solid for low and high molecular weights at low surface coverage. δ is the cross-sectional diameter of the liquid chain, and R_g is the radius of gyration of the molecules in the bulk [28.6]

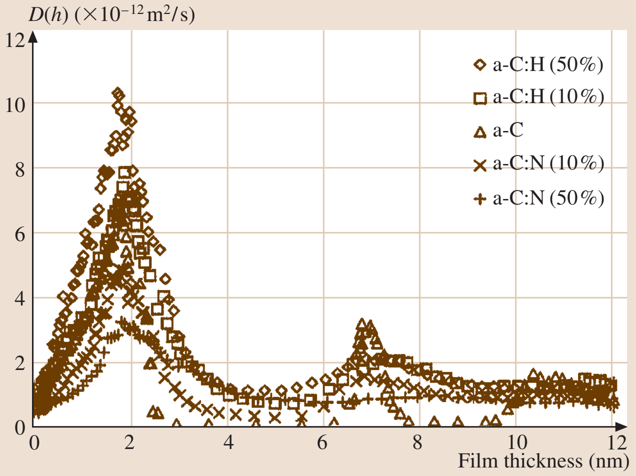

Fig. 28.5 Diffusion coefficient $D(h)$ as a function of lubricant film thickness for Z-DOL on different carbon films [28.24]

Molecularly thick films may be sheared at very high shear rates, on the order of $10^8–10^9\ \mathrm{s}^{-1}$ during sliding, such as during magnetic disk drive operation. During such shear, lubricant stability is critical to the protection of the interface. For proper lubricant selection, viscosity at high shear rates and associated shear thinning need to be understood. Viscosity measurements of eight different types of PFPE films show that all eight lubricants display Newtonian behavior and their viscosity remains constant at a shear rate up to $10^7\ \mathrm{s}^{-1}$ [28.27, 28].

28.3 Boundary Lubrication Studies

With the development of AFM techniques, studies have been carried out to investigate the nanotribological performance of PFPEs. *Mate* [28.29, 30], *O'Shea* et al. [28.31, 32], *Bhushan* et al. [28.15, 33], *Koinkar* and *Bhushan* [28.20, 34], *Bhushan* and *Sundararajan* [28.35], *Bhushan* and *Dandavate* [28.36], and *Liu* and *Bhushan* [28.21] used an AFM to provide insight into how PFPE lubricants function at the molecular level. *Mate* [28.29, 30] conducted friction experiments on bonded and unbonded Z-DOL and found that the coefficient of friction of the unbonded Z-DOL is about two times larger than the bonded Z-DOL (also see [28.31, 32]). *Koinkar* and *Bhushan* [28.20, 34] and *Liu* and *Bhushan* [28.21] studied the friction and wear performance of a Si(100) sample lubricated with Z-15, Z-DOL, and Z-DOL(BW) lubricants. They found that using Z-DOL(BW) could significantly improve the adhesion, friction, and wear performance of Si(100). They also discussed the lubrication mechanisms on the molecular level. *Bhushan* and *Sundararajan* [28.35] and *Bhushan* and *Dandavate* [28.36] studied the effect of tip radius and relative humidity on the adhesion and friction properties of Si(100) coated with Z-DOL(BW).

In this section, we review, in some detail, the adhesion, friction, and wear properties of two kinds of typical PFPE lubricants of Z-15 and Z-DOL at various operating conditions (rest time, velocity, relative humidity, temperature, and tip radius). The experiments were carried out using a commercial AFM system with pyramidal Si_3N_4 and diamond tips. An environmentally controlled chamber and a thermal stage were used to perform relative humidity and temperature effect studies.

28.3.1 Friction and Adhesion

To investigate the friction properties of Z-15 and Z-DOL(BW) films on Si(100), the friction force versus normal load curves were measured by making friction measurements at increasing normal loads [28.21]. The representative results of Si(100), Z-15, and Z-DOL(BW) are shown in Fig. 28.6. An approximately linear response of all three samples is observed in the load range of 5–130 nN. The friction force of solid-like Z-DOL(BW) is consistently smaller than that for Si(100), but the friction force of liquid-like Z-15 lubricant is higher than that of Si(100). *Sundararajan* and *Bhushan* [28.37] have studied the static friction force of silicon micromotors lubricated with Z-DOL by AFM. They also found that liquid-like lubricants of Z-DOL significantly increase the static friction force, whereas solid-like Z-DOL(BW) coating can dramatically reduce the static friction force. This is in good agreement with the results of *Liu* and *Bhushan* [28.21]. In Fig. 28.6, the nonzero value of the friction force signal at zero external load is due to the adhesive forces. It is well-known that the following relationship exists between the friction force F and external

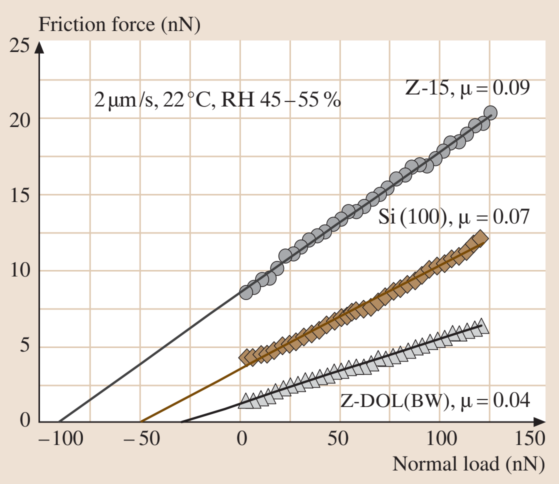

Fig. 28.6 Friction force versus normal load curves for Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at 2 µm/s, and in ambient air sliding against a Si_3N_4 tip. Based on these curves, coefficient of friction μ and adhesion force of W_a can be calculated [28.21]

normal load W [28.2, 3]

$$F = \mu(W + W_a)\,, \tag{28.1}$$

where μ is the coefficient of friction and W_a is the adhesive force. Based on this equation and the data in Fig. 28.6, we can calculate the μ and W_a values. The coefficients of friction of Si(100), Z-15, and Z-DOL are 0.07, 0.09, and 0.04, respectively. Based on (28.1), the adhesive force values are obtained from the horizontal intercepts of the friction force versus normal load curves at a zero value of friction force. Adhesive force values of Si(100), Z-15, and Z-DOL are 52 nN, 91 nN, and 34 nN, respectively.

The adhesive forces of these samples were also measured using a force calibration plot (FCP) technique. In this technique, the tip is brought into contact with the sample and the maximum force, needed to pull the tip and sample apart, is measured as the adhesive force. Figure 28.7 shows the typical FCP curves of Si(100), Z-15, and Z-DOL(BW) [28.21]. As the tip approaches the sample within a few nanometers (A), an attractive force exists between the tip and the sample surfaces. The tip is pulled toward the sample, and contact occurs at point B on the graph. The adsorption of water molecules and/or presence of liquid lubricant molecules on the sample surface can also accelerate this so-called "snap-in", due to the formation of meniscus of the water and/or liquid lubricant around the tip. From this point on, the tip is in contact with surface, and as the piezo extends further, the cantilever is further deflected. This is represented by the slope portion of the curve. As the piezo retracts, at point C the tip goes beyond the zero deflection (flat) line, because of the attractive forces, into the adhesive force regime. At point D, the tip snaps free of the adhesive forces and is again in free air. The adhesive force (pull-off force) is determined by multiplying the cantilever spring constant (0.58 N/m) by the horizontal distance between points C and D, which corresponds to the maximum cantilever deflection toward the samples before the tip is disengaged. Incidentally, the horizontal shift between the loading and unloading curves results from the hysteresis of the PZT tube.

The adhesive forces of Si(100), Z-15, and Z-DOL(BW) measured by FCP and friction force versus normal load plot are summarized in Fig. 28.8 [28.21]. The results measured by these two methods are in good agreement. Figure 28.8 shows that the presence of mobile Z-15 lubricant film increases the adhesive force as compared to that of Si(100). In contrast, the presence of solid phase Z-DOL(BW) film reduces the adhesive force as compared to that of Si(100). This result is in good agreement with the results of *Blackman* et al. [28.22] and *Bhushan* and *Ruan* [28.38]. Sources of adhesive forces between the tip and the sample surfaces are van der Waals attraction and long-range meniscus force [28.2, 3, 16]. Relative magnitudes of the forces from the two sources are dependent on various factors, including the distance between the tip and the sample surface, their surface roughness, their hydrophobicity, and relative humidity [28.39]. For most surfaces with some roughness, meniscus contribution dominates at moderate to high humidities.

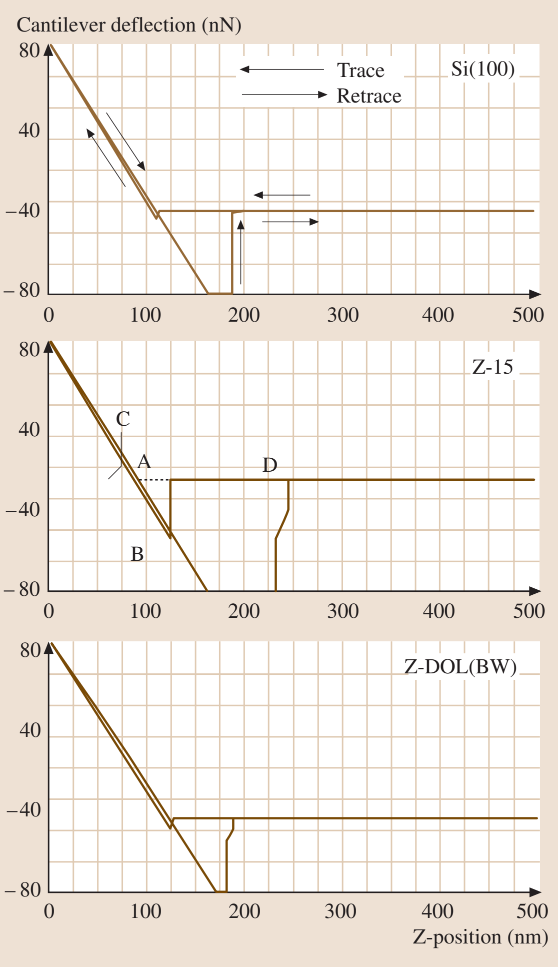

Fig. 28.7 Typical force calibration plots of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film in ambient air. The adhesive forces can be calculated from the horizontal distance between points C and D, and the cantilever spring constant of 0.58 N/m [28.21]

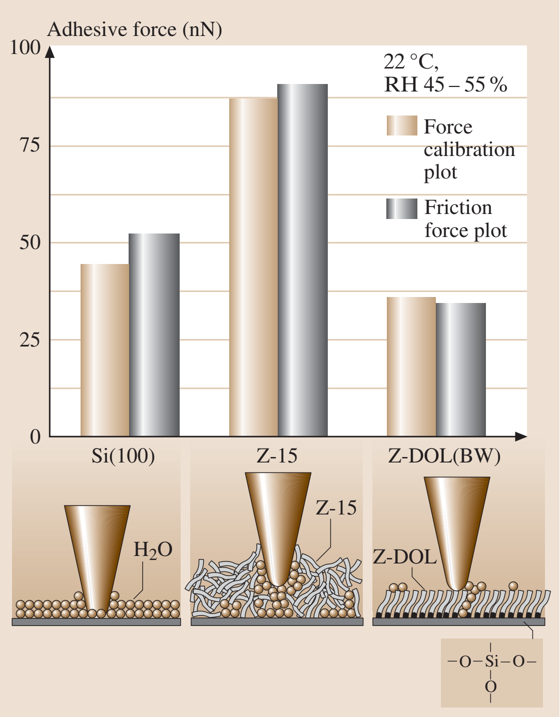

Fig. 28.8 Summary of the adhesive forces of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film measured by force calibration plots and friction force versus normal load plots in ambient air. The schematic (*bottom*) shows the effect of meniscus formed between the AFM tip and the sample surface on the adhesive and friction forces [28.21]

The schematic (bottom) in Fig. 28.8 shows relative size and sources of meniscus. The native oxide layer (SiO_2) on the top of Si(100) wafer exhibits hydrophilic properties, and some water molecules can be adsorbed on this surface. The condensed water will form meniscus as the tip approaches the sample surface. In the case of a sphere (such as a single-asperity AFM tip) in contact with a flat surface, the attractive Laplace force (F_L) caused by capillary is:

$$F_L = 2\pi R \gamma_{la} (\cos\theta_1 + \cos\theta_2) , \tag{28.2}$$

where R is the radius of the sphere, γ_{la} is the surface tension of the liquid against air, θ_1 and θ_2 are the contact angles between liquid and flat and spherical surfaces, respectively [28.2, 3, 40]. As the surface tension value of Z-15 (24 dyn/cm) is smaller than that of water (72 dyn/cm), the larger adhesive force in Z-15 cannot only be caused by the Z-15 meniscus. The nonpolarized Z-15 liquid does not have complete coverage and strong bonding with Si(100). In the ambient environment, the condensed water molecules will permeate through the liquid Z-15 lubricant film and compete with the lubricant molecules present on the substrate. The interaction of the liquid lubricant with the substrate is weakened, and a boundary layer of the liquid lubricant forms puddles [28.20, 34]. This dewetting allows water molecules to be adsorbed on the Si(100) surface as aggregates along with Z-15 molecules. And both of them can form meniscus while the tip approaches the surface. In addition, as the Z-15 film is pretty soft compared to the solid Si(100) surface, penetration of the tip in the film occurs while pushing the tip down. This leads to a large area of the tip involved to form the meniscus at the tip-liquid (water aggregates along with Z-15) interface. These two factors of the liquid-like Z-15 film result in higher adhesive force. It should also be noted that Z-15 has a higher viscosity compared to that of water. Therefore, Z-15 film provides higher resistance to sliding motion and results in a larger coefficient of friction. In the case of Z-DOL(BW) film, both of the active groups of Z-DOL molecules are strongly bonded on Si(100) substrate through the thermal and washing treatment. Thus, the Z-DOL(BW) film has relatively low free surface energy and cannot be displaced readily by water molecules or readily adsorb water molecules. Thus, the use of Z-DOL(BW) can reduce the adhesive force. We further note that the bonded Z-DOL molecules can be orientated under stress (behaves as a soft polymer solid), which facilitates sliding and reduces coefficient of friction.

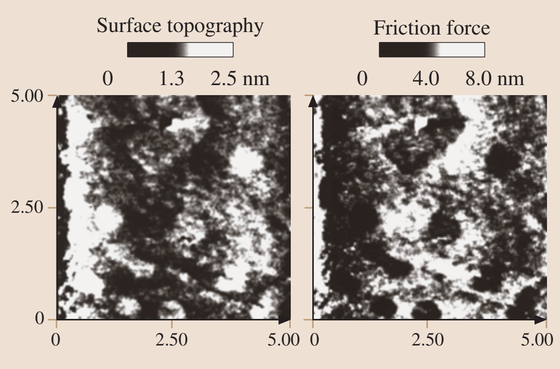

Fig. 28.9 Gray-scale plots of the surface topography and friction force obtained simultaneously for unbonded 2.3-nm-thick Demnum-type PFPE lubricant film on silicon [28.20]

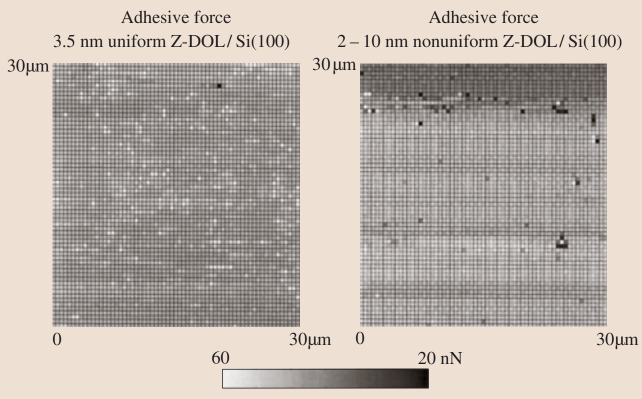

Fig. 28.10 Gray-scale plots of the adhesive force distribution of a uniformly coated, 3.5-nm-thick unbonded Z-DOL film on silicon and 3- to 10-nm-thick unbonded Z-DOL film on silicon that was deliberately coated nonuniformly by vibrating the sample during the coating process [28.36]

These studies suggest that if the lubricant films exist as liquid-like, such as Z-15 films, they easily form meniscus (by themselves and the adsorbed water molecules), and thus have higher adhesive force and higher friction force. Whereas if the lubricant film exists in solid-like phase, such as Z-DOL(BW) films, they are hydrophobic with low adhesion and friction.

In order to study the uniformity of lubricant film and its influence on friction and adhesion, friction force mapping and adhesive force mapping of PFPE have been carried out by *Koinkar* and *Bhushan* [28.34] and *Bhushan* and *Dandavate* [28.36], respectively. Figure 28.9 shows gray-scale plots of surface topography and friction force images obtained simultaneously for unbonded Demnum-type PFPE lubricant film on silicon [28.34]. The friction force plot shows well distinguished low and high friction regions corresponding roughly to high and low surface height regions in the topography image (thick and thin lubricant regions). A uniformly lubricated sample does not show such a variation in friction. Figure 28.10 shows the gray-scale plots of the adhesive force distribution for silicon samples coated uniformly and nonuniformly with Z-DOL lubricant. It can be clearly seen that there exists a region that has an adhesive force distinctly different from the other region for the nonuniformly coated sample. This implies that the liquid film thickness is nonuniform, giving rise to a difference in the meniscus forces.

28.3.2 Rest Time Effect

It is well-known that in the computer rigid disk drive, the stiction force increases rapidly with an increase in rest time between the head and magnetic medium disk [28.10, 11]. Considering that the stiction and friction are

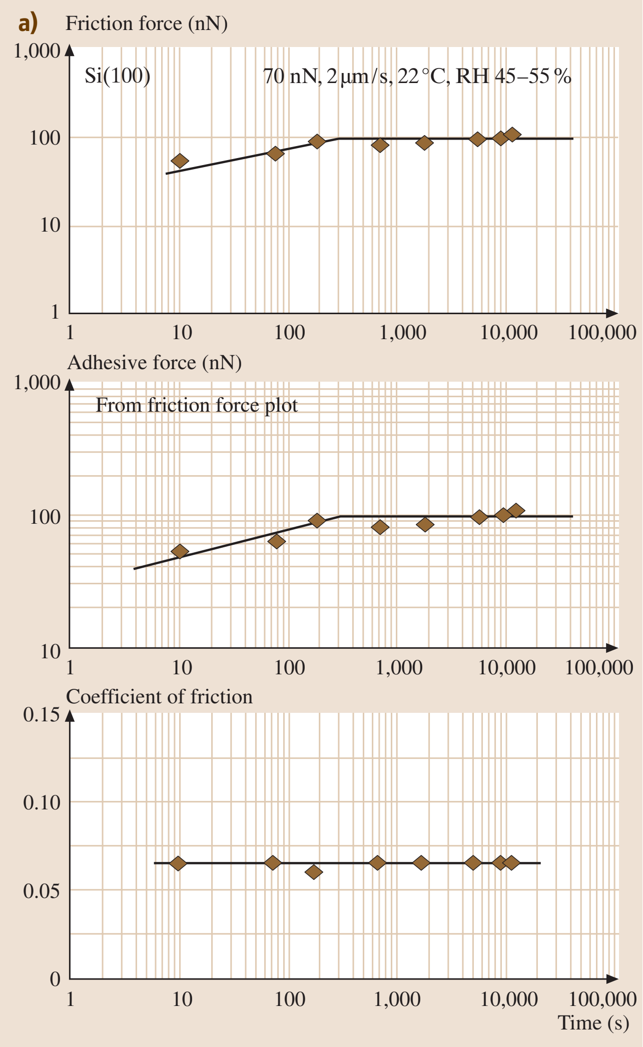

Fig. 28.11 (a) Rest time effect on friction force, adhesive force, and coefficient of friction of Si(100). **(b)** (*see next page*) Summary of the rest time effect on friction force, adhesive force, and coefficient of friction of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film. All of the measurements were carried out at 70 nN, 2 µm/s, and in ambient air [28.21]

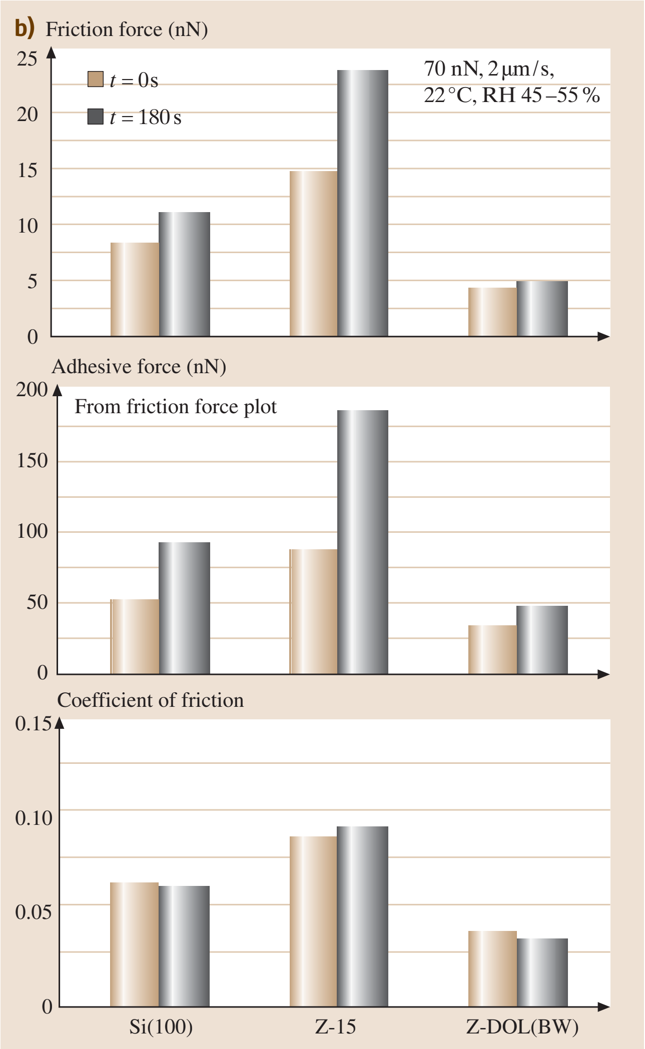

Fig. 28.11b

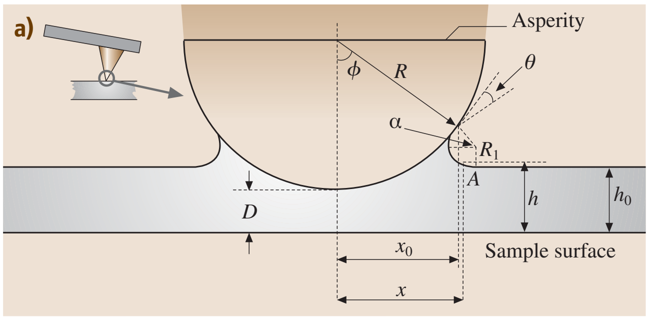

Fig. 28.12a

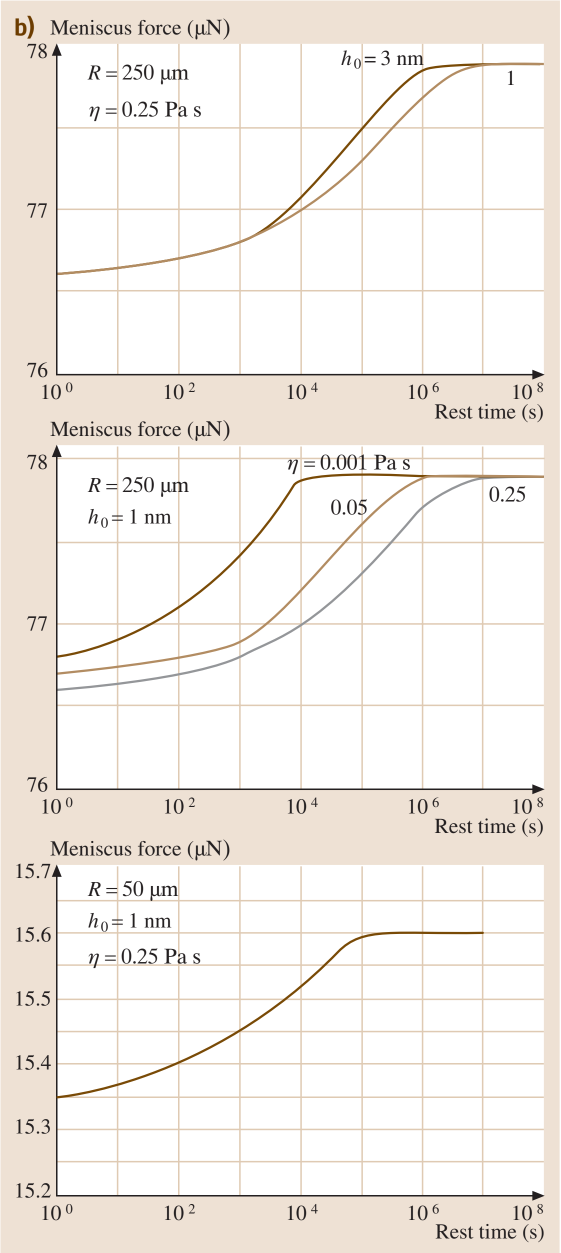

Fig. 28.12 **(a)** Schematic of a single asperity in contact with a smooth flat surface in the presence of a continuous liquid film when ϕ is large. **(b)** Results of the single asperity model. Effect of viscosity of the liquid, radius of the asperity, and film thickness is studied with respect to the time-dependent meniscus force [28.41]

two of the major issues that lead to the failure of computer rigid disk drives and MEMS, it is very important to find out if the rest time effect also exists on the nanoscale. First, the rest time effect on the friction force, adhesive force, and coefficient of Si(100) sliding against Si_3N_4 tip was studied, Fig. 28.11a [28.21]. It was found that the friction and adhesive forces logarithmically increase up to a certain equilibrium time after which they remain constant. Figure 28.11a also shows that the rest time does not affect the coefficient of friction. These results suggest that the rest time can result in growth of the meniscus, which causes a higher adhesive force, and in turn, a higher friction force. But in the whole testing range the friction mechanisms do not change with the rest time. Similar studies were also performed on Z-15 and Z-DOL(BW) films. The results are summarized in Fig. 28.11b [28.21]. It is seen that similar time effect has been observed on Z-15 film, but not on Z-DOL(BW) film.

AFM tip in contact with a flat sample surface is generally believed to represent a single asperity contact. Therefore, a Si_3N_4 tip in contact with Si(100) or Z-15/Si(100) can be modeled as a sphere in contact with a flat surface covered by a layer of liquid (adsorbed water and/or liquid lubricant), Fig. 28.12a. Meniscus forms around the contacting asperity and grows with time until equilibrium occurs [28.41]. The meniscus force, which is the product of meniscus pressure and meniscus area, depends on the flow of liquid phase toward the contact zone. The flow of the liquid toward the contact zone is governed by the capillary pressure P_c, which draws liquid into the meniscus, and the disjoining pressure Π, which tends to draw the liquid away from the meniscus. Based on the Young and Laplace equation, the capillary pressure, P_c, is:

$$P_c = 2K\gamma\ , \tag{28.3}$$

where $2K$ is the mean meniscus curvature ($= K_1 + K_2$, where K_1 and K_2 are the curvatures of the meniscus in the contact plane and perpendicular to the contact plane) and γ is the surface tension of the liquid. *Mate* and *Novotny* [28.6] have shown that the disjoining pressure decreases rapidly with increasing liquid film thickness in a manner consistent with a strong van der Waals attraction. The disjoining pressure, Π, for these liquid films can be expressed as:

$$\Pi = \frac{A}{6\pi h^3}\ , \tag{28.4}$$

where A is the Hamaker constant and h is the liquid film thickness. The driving forces that cause the lubricant flow that results in an increase in the meniscus force are the disjoining pressure gradient, due to a gradient in film thickness, and capillary pressure gradient, due to curved liquid-air interface. The driving pressure, P, can then be written as:

$$P = -2K\gamma - \Pi\ . \tag{28.5}$$

Based on these three basic relationships, the following differential equation has been derived by *Chilamakuri* and *Bhushan* [28.41], which can describe the meniscus at time t:

$$2\pi x_0\left(D + \frac{x_0^2}{2R} - h_0\right)\frac{\mathrm{d}x_0}{\mathrm{d}t} = \frac{2\pi h_0^3\gamma}{3\eta}\frac{(1+\cos\theta)}{D+a-h_0} - \frac{Ax_0}{3\eta h}\cot\alpha\ , \tag{28.6}$$

where η is the viscosity of the liquid and a is given as

$$a = R(1-\cos\phi) \sim \frac{R\phi^2}{2} \sim \frac{x_0^2}{2R}\ . \tag{28.7}$$

The differential equation (28.4) was solved numerically using Newton's iteration method. The meniscus force at any time t less than the equilibrium time is proportional to the meniscus area and meniscus pressure ($2K\gamma$), and it is given by

$$f_\mathrm{m}(t) = 2\pi R\gamma(1+\cos\theta)\left(\frac{x_0}{(x_0)_\mathrm{eq}}\right)^2\left(\frac{K}{K_\mathrm{eq}}\right)\ , \tag{28.8}$$

where $(x_0)_\mathrm{eq}$ is the value of x_0 at the equilibrium time

$$\left[(x_0)_\mathrm{eq}\right]^2 = 2R\left[\frac{-6\pi h_0^3\gamma(1+\cos\theta)}{A} + (h_0 - D)\right]. \tag{28.9}$$

This modeling work (at the microscale) showed that the meniscus force initially increases logarithmically with the rest time up to a certain equilibrium time after which it remains constant. Equilibrium time decreases with an increase in liquid film thickness, a decrease in viscosity, and a decrease in the tip radius, Fig. 28.12b. This early numerical modeling work and the data at the nanoscale in Fig. 28.11a are in good agreement.

28.3.3 Velocity Effect

To investigate the velocity effect on friction and adhesion, the friction force versus normal load relationships of Si(100), Z-15, and Z-DOL(BW) at different velocities were measured, Fig. 28.13 [28.21]. Based on these data,

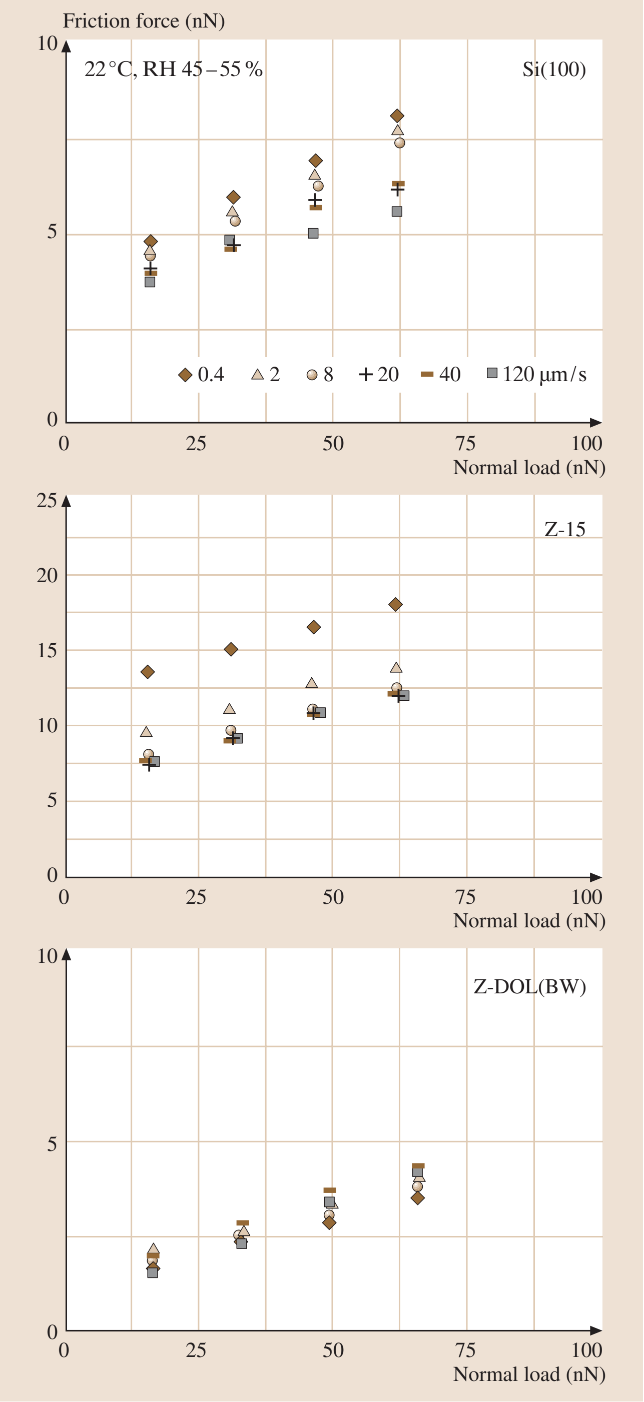

Fig. 28.13 Friction forces versus normal load data of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at various velocities in ambient air [28.21]

the adhesive force and coefficient of friction values can be calculated by (28.1). The variation of friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL(BW) as a function of velocity are summarized in Fig. 28.14. It indicates that for silicon wafer, the friction force decreases logarithmically with increasing velocity. For Z-15, the friction force decreases with increasing velocity up to 10 µm/s, after which it remains almost constant. The velocity has a much smaller effect on the friction force of Z-DOL(BW); it reduced slightly only at very high velocity. Figure 28.14 also indicates that the adhesive force of Si(100) is increased when the velocity is higher than 10 µm/s. The adhesive force of Z-15 is reduced dramatically with a velocity increase up to 20 µm/s, after which it is reduced slightly. And the adhesive force of Z-DOL(BW) also decreases at high velocity. In the testing range of velocity, only the coefficient of friction of Si(100) decreases with velocity, but the coefficients of friction of Z-15 and Z-DOL(BW) almost remain constant. This implies that the friction mechanisms of Z-15 and Z-DOL(BW) do not change with the variation of velocity.

The mechanisms of the effect of velocity on the adhesion and friction are explained based on the schematics shown in Fig. 28.14 (right). For Si(100), tribochemical reaction plays a major role. Although at high velocity the meniscus is broken and does not have enough time to rebuild, the contact stresses and high velocity lead to tribochemical reactions of Si(100) wafer and Si_3N_4 tip, which have native oxide (SiO_2) layers with water molecules. The following reactions occur:

$$SiO_2 + 2H_2O = Si(OH)_4 \tag{28.10}$$

$$Si_3N_4 + 16H_2O = 3Si(OH)_4 + 4NH_4OH\,. \tag{28.11}$$

The $Si(OH)_4$ is removed and continuously replenished during sliding. The $Si(OH)_4$ layer between the tip and Si(100) surface is known to be of low shear strength and causes a decrease in friction force and coefficient of friction in the lateral direction [28.42–46]. The chemical bonds of Si–OH between the tip and Si(100) surface induce large adhesive force in the normal direction. For Z-15 film, at high velocity the meniscus formed by condensed water and Z-15 molecules is broken and does not have enough time to rebuild. Therefore, the adhesive force and, consequently, friction force is reduced. For Z-DOL(BW) film, the surface can adsorb few water molecules in ambient condition, and at high velocity these molecules are displaced, which is responsible for a slight decrease in friction force and adhesive force. Even at high velocity range, the friction mechanisms for Z-15 and Z-DOL(BW) films still are shearing of the vis-

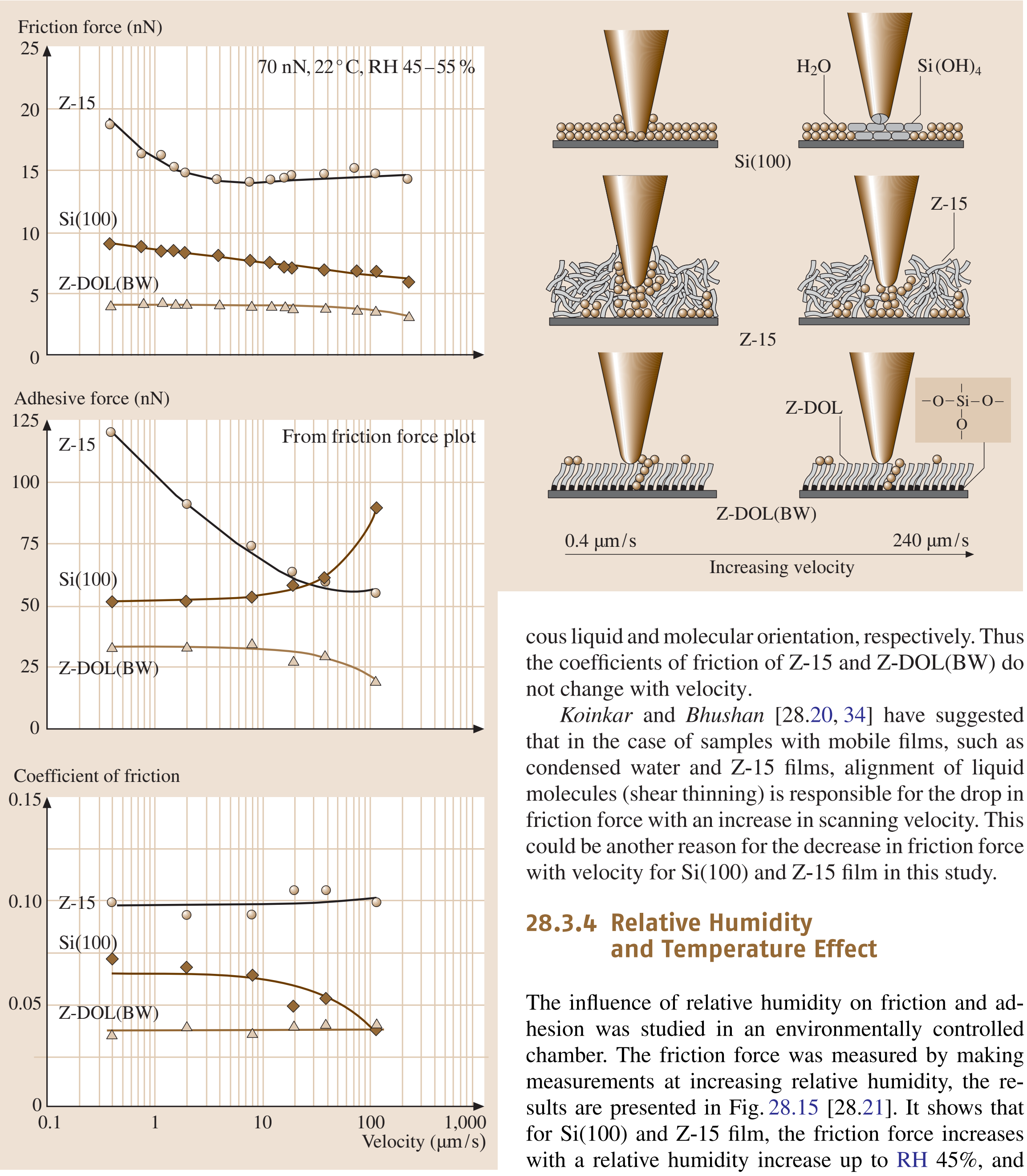

Fig. 28.14 The influence of velocity on the friction force, adhesive force, and coefficient of friction of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at 70 nN, in ambient air. The schematic (*right*) shows the change of surface composition (by tribochemical reaction) and change of meniscus with increasing velocity [28.21]

cous liquid and molecular orientation, respectively. Thus the coefficients of friction of Z-15 and Z-DOL(BW) do not change with velocity.

Koinkar and *Bhushan* [28.20, 34] have suggested that in the case of samples with mobile films, such as condensed water and Z-15 films, alignment of liquid molecules (shear thinning) is responsible for the drop in friction force with an increase in scanning velocity. This could be another reason for the decrease in friction force with velocity for Si(100) and Z-15 film in this study.

28.3.4 Relative Humidity and Temperature Effect

The influence of relative humidity on friction and adhesion was studied in an environmentally controlled chamber. The friction force was measured by making measurements at increasing relative humidity, the results are presented in Fig. 28.15 [28.21]. It shows that for Si(100) and Z-15 film, the friction force increases with a relative humidity increase up to RH 45%, and then it shows a slight decrease with a further increase in relative humidity. Z-DOL(BW) has a smaller friction force than Si(100) and Z-15 in the whole testing range. And its friction force shows a relatively apparent increase when the relative humidity is higher than RH 45%. For Si(100), Z-15, and Z-DOL(BW), adhesive

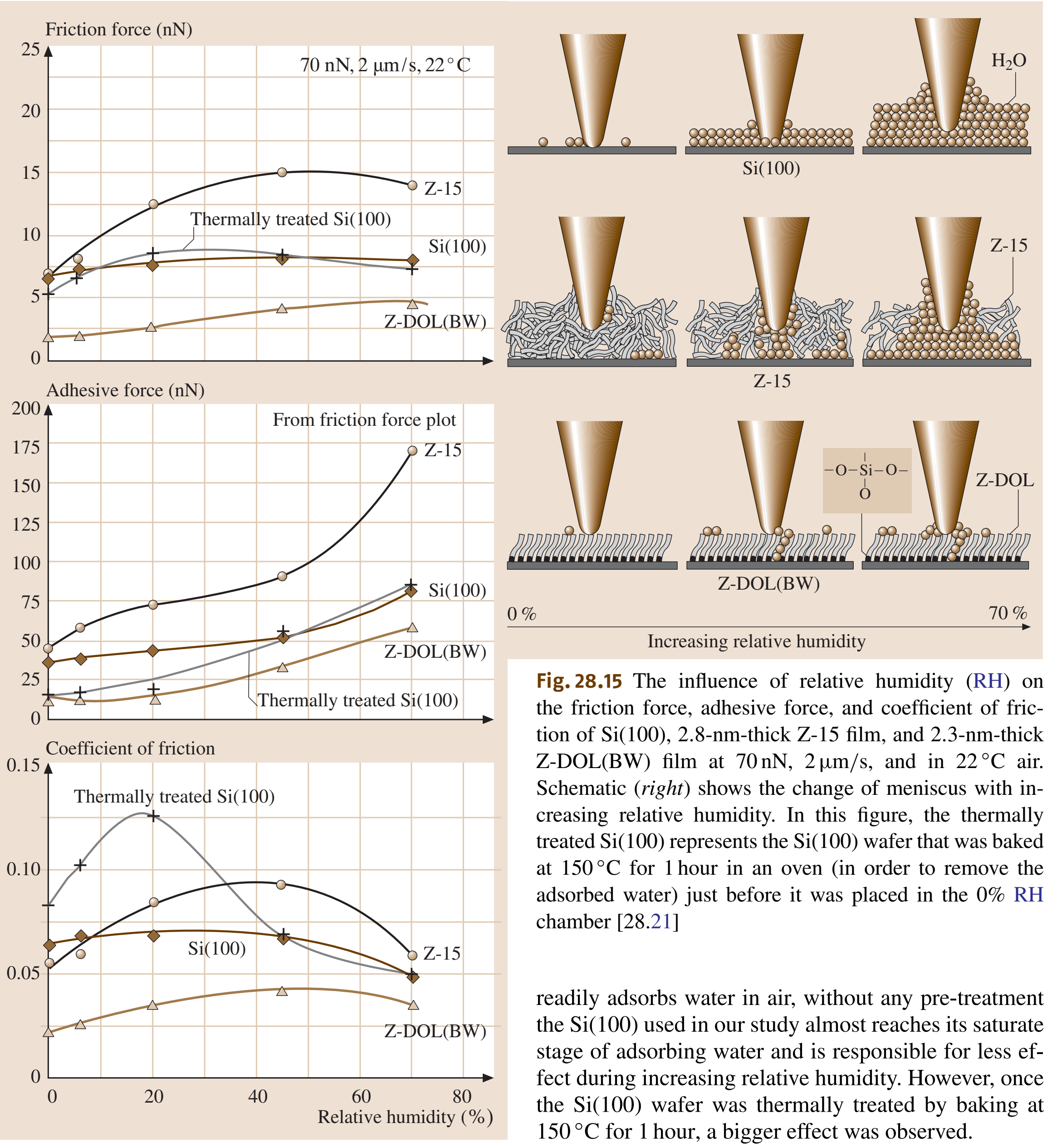

Fig. 28.15 The influence of relative humidity (RH) on the friction force, adhesive force, and coefficient of friction of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at 70 nN, 2 μm/s, and in 22 °C air. Schematic (*right*) shows the change of meniscus with increasing relative humidity. In this figure, the thermally treated Si(100) represents the Si(100) wafer that was baked at 150 °C for 1 hour in an oven (in order to remove the adsorbed water) just before it was placed in the 0% RH chamber [28.21]

forces increase with relative humidity. And their coefficients of friction increase with a relative humidity up to RH 45%, after which they decrease with a further increase of the relative humidity. It is also observed that the humidity effect on Si(100) really depends on the history of the Si(100) sample. As the surface of Si(100) wafer readily adsorbs water in air, without any pre-treatment the Si(100) used in our study almost reaches its saturate stage of adsorbing water and is responsible for less effect during increasing relative humidity. However, once the Si(100) wafer was thermally treated by baking at 150 °C for 1 hour, a bigger effect was observed.

The schematic (right) in Fig. 28.15 shows that because its high free surface energy Si(100) can adsorb more water molecules with increasing relative humidity. As discussed earlier, for Z-15 film in a humid environment, condensed water competes with the lubricant film present on the sample surface. Obviously, more water molecules can also be adsorbed on a Z-15 surface with increasing relative humidity. The more adsorbed water

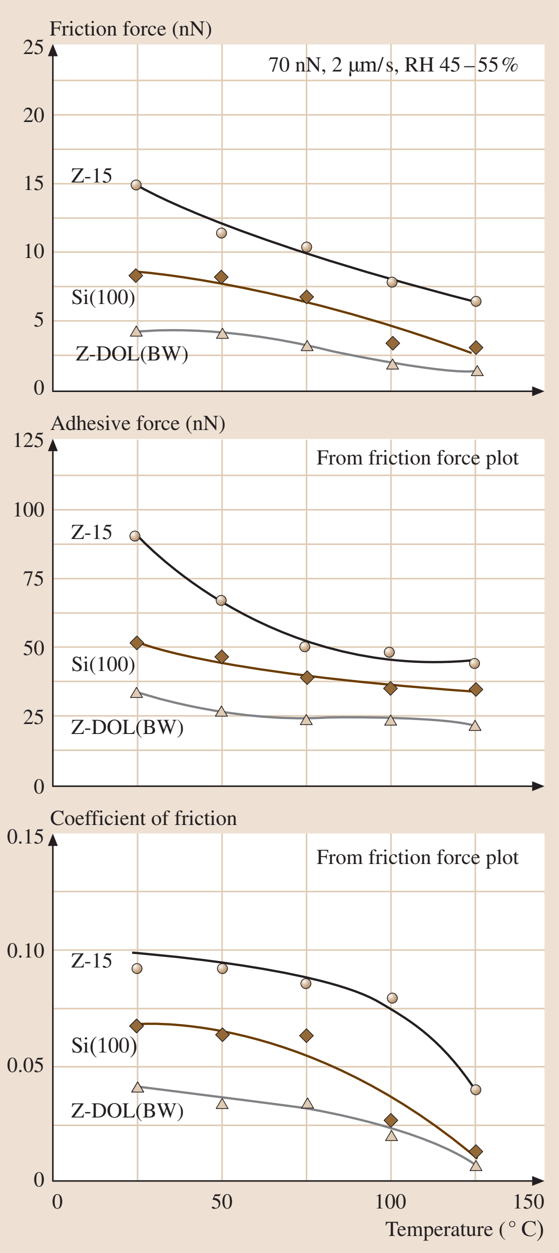

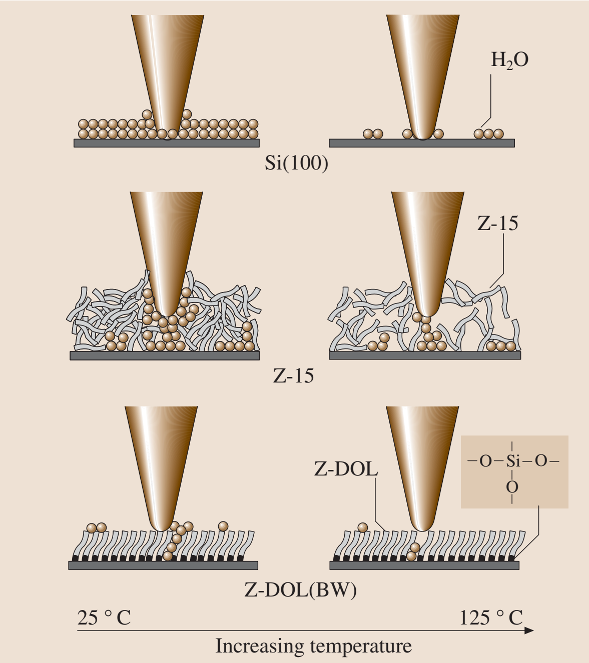

Fig. 28.16 The influence of temperature on the friction force, adhesive force, and coefficient of friction of Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at 70 nN, at 2 µm/s, and in RH 40–50 % air. The schematic (*right*) shows that at high temperature, desorption of water decreases the adhesive forces. And the reduced viscosity of Z-15 leads to the decrease of coefficient of friction. High temperature facilitates orientation of molecules in Z-DOL(BW) film, which results in lower coefficient of friction [28.21]

molecules in the case of Si(100), along with lubricant molecules in Z-15 film, form a bigger water meniscus, which leads to an increase of friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 with humidity. But at very high humidity of RH 70%, large quantities of adsorbed water can form a continuous water layer that separates the tip and sample surface, and acts as a kind of lubricant, which causes a decrease in the friction force and coefficient of friction. For Z-DOL(BW) film, because of its hydrophobic surface properties, water molecules can only be adsorbed at high humidity ($\geq$ RH 45%), which causes an increase in the adhesive force and friction force.

The effect of temperature on friction and adhesion was studied using a thermal stage attached to the AFM. The friction force was measured at increasing temperature from 22–125 °C. The results are presented in Fig. 28.16 [28.21]. It shows that the increasing temperature causes a decrease of friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL(BW). The schematic (right) in Fig. 28.16 indicates that at high temperature, desorption of water leads to the decrease of friction force, adhesive force, and coefficient of friction for all of the samples. Besides that, the reduction of surface tension of water also contributes to the decrease of friction and adhesion. For Z-15 film, the reduction of viscosity at high temperature has an additional contribution to the decrease of friction. In the case of Z-DOL(BW) film, molecules are more easily oriented at high temperature, which may also be responsible for the low friction.

Using a surface force apparatus, *Yoshizawa* and *Israelachvili* [28.47, 48] have shown that a change in the velocity or temperature induces phase transformation (from crystalline solid-like to amorphous, then to liquid-like) in surfactant monolayers, which are responsible for the observed changes in the friction force. Stick-slip is observed in a low velocity regime of a few μm/s, and adhesion and friction first increase followed by a decrease in the temperature range of 0–50 °C. Stick-slip at low velocity and adhesion and friction curves peaking at some particular temperature (observed in their study), have not been observed in the AFM study. It suggests that the phase transformation may not happen in this study, because PFPEs generally have very good thermal stability [28.10, 12].

As a brief summary, the influence of velocity, relative humidity, and temperature on the friction force of Z-15 film is presented in Fig. 28.17. The changing trends are also addressed in this figure.

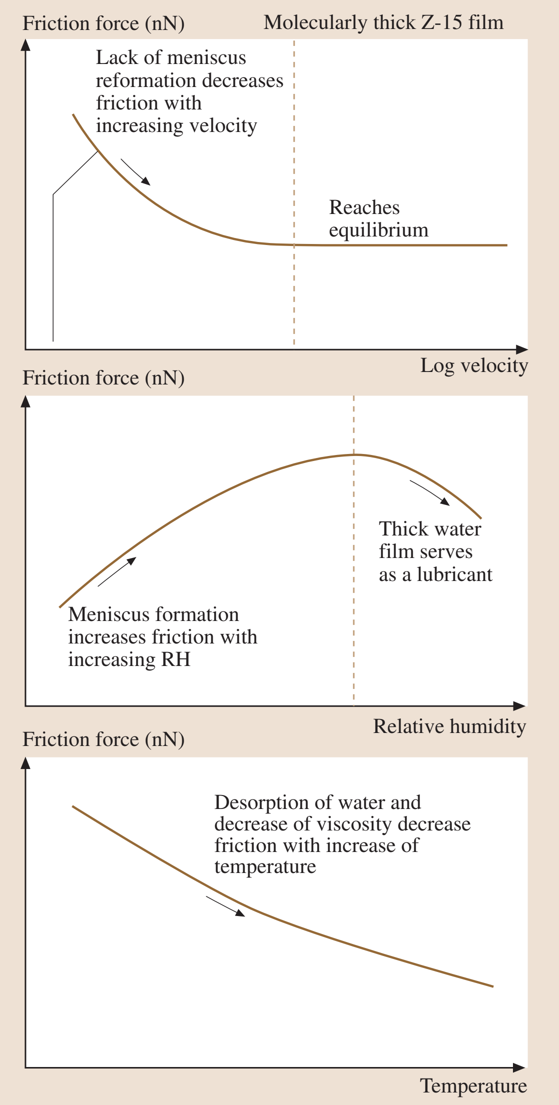
Friction force (nN)
Molecularly thick Z-15 film
Lack of meniscus reformation decreases friction with increasing velocity
Reaches equilibrium
Log velocity
Friction force (nN)
Meniscus formation increases friction with increasing RH
Thick water film serves as a lubricant
Relative humidity
Friction force (nN)
Desorption of water and decrease of viscosity decrease friction with increase of temperature
Temperature

Fig. 28.17 Schematic shows the change of friction force of molecularly thick Z-15 films with log velocity, relative humidity, and temperature [28.21]

28.3.5 Tip Radius Effect

The tip radius and relative humidity affect adhesion and friction for dry and lubricated surfaces [28.35, 36]. Figure 28.18a shows the variation of single point adhesive force measurements as a function of tip radius on a Si(100) sample for several humidities. The adhesive force data are also plotted as a function of relative humidity for various tip radii. Figure 28.18a indicates that the tip radius has little effect on the adhesive forces at low humidities, but the adhesive force increases with tip radius at high humidity. Adhesive force also increases with an increase in humidity for all tips. The trend in adhesive forces as a function of tip radii and relative humidity, in Fig. 28.18a, can be explained by the presence of meniscus forces, which arise from capillary condensation of water vapor from the environment. If enough liquid is present to form a meniscus bridge, the meniscus force should increase with an increase in tip radius based on (28.2). This observation suggests that thickness of the liquid film at low humidities is insufficient to form continuous meniscus bridges and to affect adhesive forces in the case of all tips.

Figure 28.18a also shows the variation in coefficient of friction as a function of tip radius at a given humidity and as a function of relative humidity for a given tip radius on the Si(100) sample. It can be observed that for RH 0%, the coefficient of friction is about the same for the tip radii except for the largest tip, which shows

a) **Si (100)**

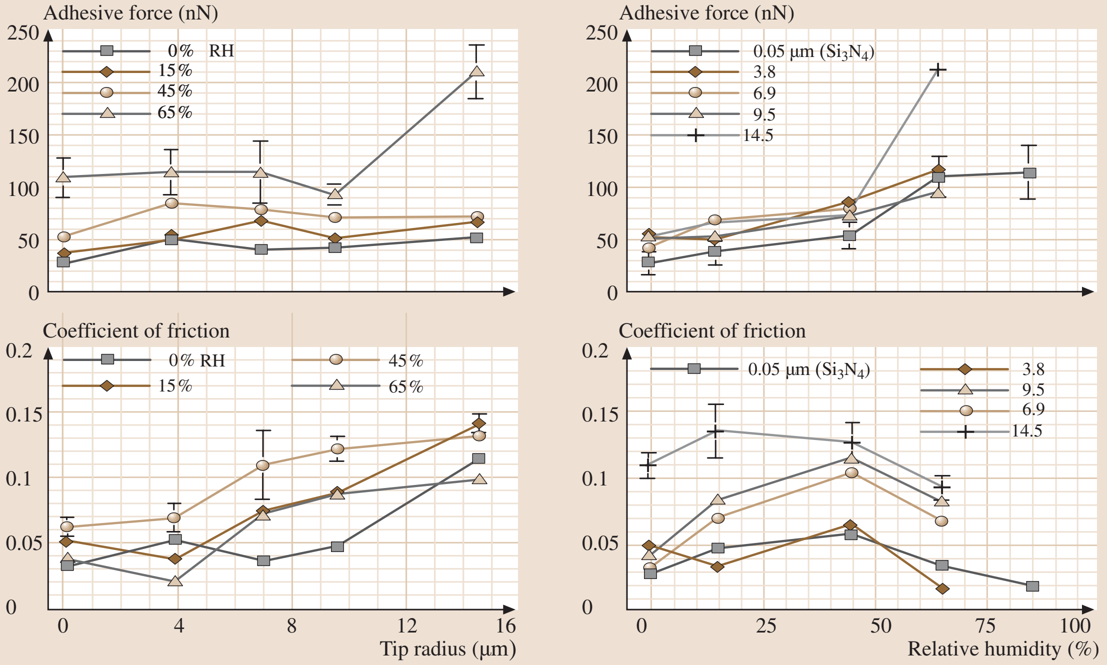

b) **Z-DOL(BW)/Si(100)**

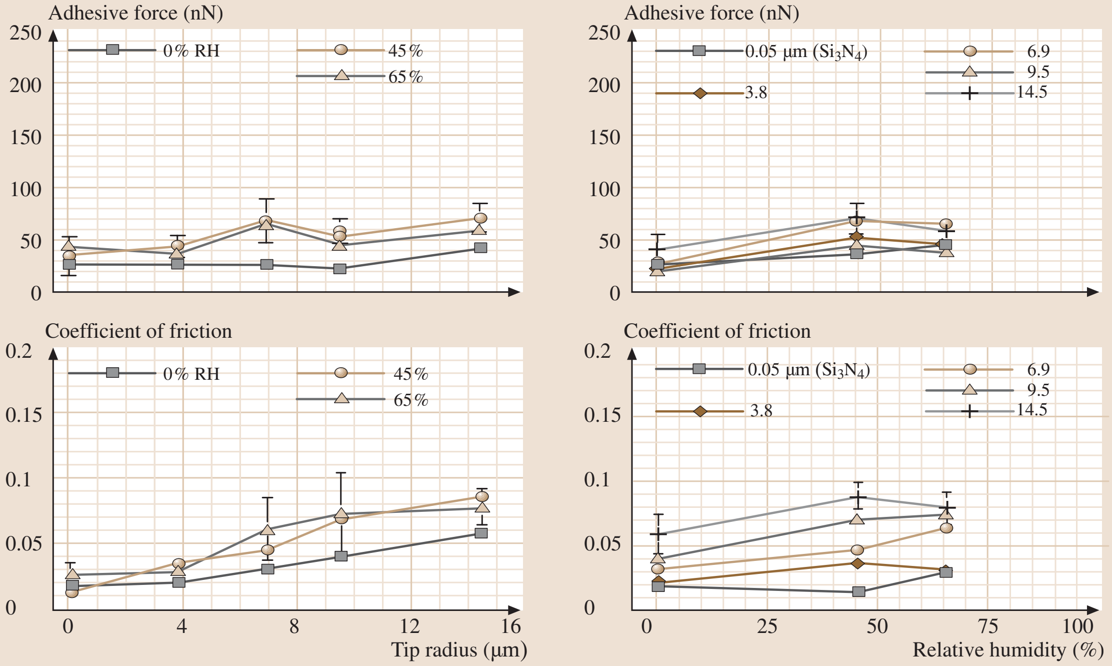

Fig. 28.18a,b Adhesive force and coefficient of friction as a function of tip radius at several humidities and as a function of relative humidity at several tip radii on **(a)** Si(100) and **(b)** 0.5-nm Z-DOL(BW) films [28.35]

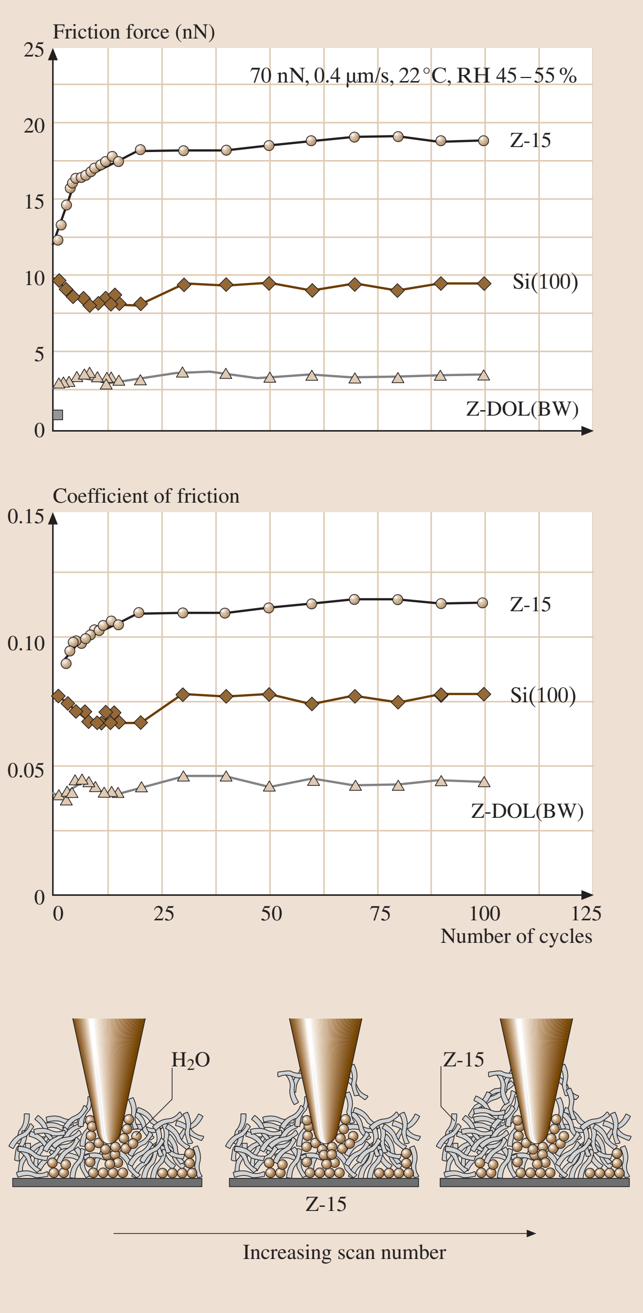

Fig. 28.19 Friction force and coefficient of friction versus number of sliding cycles for Si(100), 2.8-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film at 70 nN, 0.8 µm/s, and in ambient air. Schematic (*bottom*) shows that some liquid Z-15 molecules can be attached on the tip. The molecular interaction between the attached molecules on the tip with the Z-15 molecules in the film results in an increase of the friction force with multiple scanning [28.21]

a higher value. At all other humidities, the trend consistently shows that the coefficient of friction increases with tip radius. An increase in friction with tip radius at low to moderate humidities arises from increased contact area (i. e., higher van der Waals forces) and higher values of shear forces required for a larger contact area. At high humidities, similar to adhesive force data, an increase with tip radius occurs due to of both contact area and meniscus effects. It can be seen that for all tips, the coefficient of friction increases with humidity to about RH 45%, beyond which it starts to decrease. This is attributed to the fact that at higher humidities, the adsorbed water film on the surface acts as a lubricant between the two surfaces [28.21]. Thus the interface is changed at higher humidities, resulting in lower shear strength and, hence, lower friction force and coefficient of friction.

Figure 28.18b shows adhesive forces as a function of tip radius and relative humidity on Si(100) coated with 0.5-nm-thick Z-DOL(BW) film. Adhesive forces for all the tips with the Z-DOL(BW) lubricated sample are much lower than those experienced on unlubricated Si(100), shown in Fig. 28.18a. The data also show that even at a monolayer thickness of the lubricant there is very little variation in adhesive forces with tip radius at a given humidity. For a given tip radius, the variation in adhesive forces with relative humidity indicates that these forces slightly increase from RH 0% to RH 45%, but remain more or less the same with a further increase in humidity. This is seen even with the largest tip, which indicates that the lubricant is indeed hydrophobic; there is some meniscus formation at humidities higher than RH 0%, but the formation is very minimal and does not increase appreciably even up to RH 65%. Figure 28.18b also shows coefficient of friction for various tips at different humidities for the Z-DOL(BW) lubricated sample. Again, all the values obtained with the lubricated sample are much lower than the values obtained on unlubricated Si(100), shown in Fig. 28.18a. The coefficient of friction increases with tip radius for all humidities, as was seen on unlubricated Si(100), due to an increase in the contact area. Similar to the adhesive forces, there is an increase in friction from RH 0% to RH 45%, due to a contribution from an increased number of menisci bridges. But thereafter there is very little additional water film forming, due to the hydrophobicity of the Z-DOL(BW) layer, and, consequentially, the coefficient of friction does not change appreciably, even with the largest tip. These findings show that even a monolayer of Z-DOL(BW) offers good hydrophobic performance of the surface.

28.3.6 Wear Study

To study the durability of lubricant films at the nanoscale, the friction of Si(100), Z-15, and Z-DOL(BW) as a function of the number of scanning cycles was measured, Fig. 28.19 [28.21]. As observed earlier, friction force and coefficient of friction of Z-15 is higher than that of Si(100), and Z-DOL(BW) has the lowest values. During cycling, friction force and coefficient of friction of Si(100) show a slight variation during the initial few cycles then remain constant. This is related to the removal of the top adsorbed layer. In the case of Z-15 film, the friction force and coefficient of friction show an increase during the initial few cycles and then approach higher and stable values. This is believed to be caused by the attachment of the Z-15 molecules onto the tip. The molecular interaction between these attached molecules to the tip and molecules on the film surface is responsible for an increase in the friction. But after several scans, this molecular interaction reaches the equilibrium, and after that, friction force and coefficient of friction remain constant. In the case of Z-DOL(BW) film, the friction force and coefficient of friction start out low and remain low during the entire test for 100 cycles. It suggests that Z-DOL(BW) molecules do not get attached or displaced as readily as Z-15.

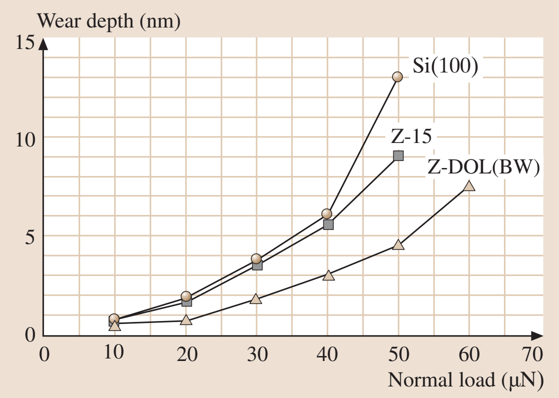

Fig. 28.20 Wear depth as a function of normal load using a diamond tip for Si(100), 2.9-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) after one cycle [28.20]

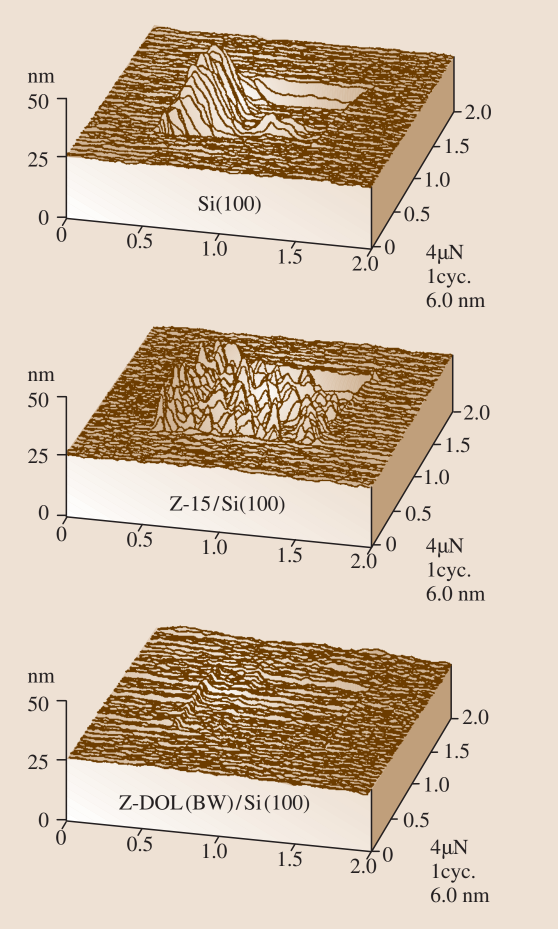

Fig. 28.21 Wear profiles for Si(100), 2.9-nm-thick Z-15 film, and 2.3-nm-thick Z-DOL(BW) film after wear studies using a diamond tip. Normal force used and wear depths are listed in the figure [28.20]

Koinkar and *Bhushan* [28.20, 34] conducted wear studies using a diamond tip at high loads. Figure 28.20 shows the plots of wear depth as a function of normal force, and Fig. 28.21 shows the wear profiles of the worn samples at 40 µN normal load. The 2.3-nm-thick Z-DOL(BW) lubricated sample exhibits better wear resistance than unlubricated and 2.9-nm-thick Z-15 lubricated silicon samples. Wear resistance of a Z-15 lubricated sample is little better than that of unlubricated sample. The Z-15 lubricated sample shows debris inside the wear track. Since the Z-15 is a liquid lubricant, debris generated is held by the lubricant and they become sticky. The debris moves inside the wear track

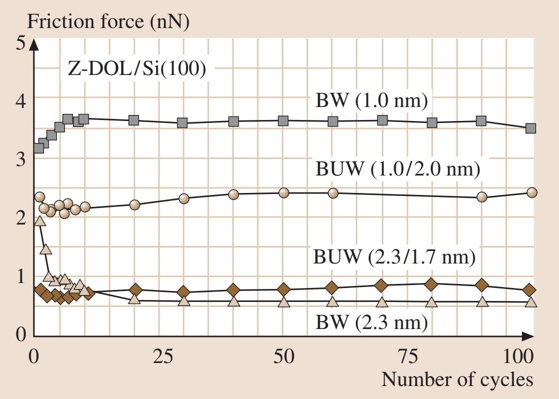

Fig. 28.22 Friction force as a function of number of cycles using a Si_3N_4 tip at a normal load of 300 nN for Z-DOL(BW) and Z-DOL(BUW) films with different film thicknesses [28.34]

and does damage, Fig. 28.20. These results suggest that using Z-DOL(BW) can improve the wear resistance of substrate.

To study the effect of the degree of chemical bonding, the durability tests were conducted on both fully bonded and partially bonded Z-DOL films. Durability results for Z-DOL(BW) and Z-DOL bonded and unwashed (Z-DOL(BUW), a partially bonded film that contains both bonded and mobile phase lubricants) with different film thicknesses are shown in Fig. 28.22 [28.34]. Thicker films, such as Z-DOL(BUW), with a thickness of 4.0 nm (bonded/mobile = 2.3 nm/1.7 nm) exhibit behavior similar to 2.3-nm-thick Z-DOL(BW) film. Figure 28.22 also indicates that Z-DOL(BW) and Z-DOL(BUW) films with a thinner film thickness exhibit a higher friction value. Comparing 1.0-nm-thick Z-DOL(BW) with 3.0-nm-thick Z-DOL(BUW) (bonded/mobile = 1.0 nm/2.0 nm), the Z-DOL(BUW) film exhibits a lower and stable friction value. This is because the mobile phase on a surface acts as a source of lubricant replenishment. A similar conclusion has also been reported by *Ruhe* et al. [28.19], *Bhushan* and *Zhao* [28.13], and *Eapen* et al. [28.49]. All of them indicate that using partially bonded Z-DOL films can dramatically reduce the friction and improve the wear life.

28.4 Closure

Nanodeformation study has shown that fully bonded Z-DOL lubricants behave as soft polymer solids, while the unbonded lubricants behave liquid-like. AFM studies have shown that the physisorbed nonpolar molecules on a solid surface have an extended, flat conformation. The spreading property of PFPE is strongly dependent on the molecular end groups and substrate chemistry.

Using solid-like Z-DOL(BW) film can reduce the friction and adhesion of Si(100), while using liquid-like lubricant of Z-15 shows a negative effect. Si(100) and Z-15 film show apparent time effect. The friction and adhesion forces increase as a result of growth of meniscus up to an equilibrium time, after which they remain constant. Using Z-DOL(BW) film can prevent time effects. High velocity leads to the rupture of meniscus and prevents its reformation, which leads to a decrease of friction and adhesive forces of Z-15 and Z-DOL(BW). The influence of relative humidity on the friction and adhesion is dominated by the amount of the adsorbed water molecules. Increasing humidity can either increase friction through increased adhesion by water meniscus, or reduce friction through an enhanced water-lubricating effect. Increasing temperature leads to desorption of water layer, decrease of water surface tension, decrease of viscosity, and easier orientation of the Z-DOL(BW) molecules. These changes cause a decrease of friction force and adhesion at high temperature. During cycling tests, the molecular interaction between the attached Z-15 molecules to the tip and the Z-15 molecules on the film surface causes the initial rise of friction. Wear tests show that Z-DOL(BW) can improve the wear resistance of silicon. Partially bonded PFPE film appears to be more durable than fully bonded films.

These results suggest that partially/fully bonded films are good lubricants for devices operating in different environments and under varying conditions.

References

28.1 B. Bhushan: Magnetic Recording Surfaces. In: *Characterization of Tribological Materials*, ed. by W. A. Glaeser (Butterworth-Heinemann, Boston 1993) pp. 116–133
28.2 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)
28.3 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)
28.4 V. J. Novotny, I. Hussla, J. M. Turlet, M. R. Philpott: Liquid polymer conformation on solid surfaces, J. Chem. Phys. **90** (1989) 5861–5868
28.5 V. J. Novotny: Migration of liquid polymers on solid surfaces, J. Chem. Phys. **92** (1990) 3189–3196
28.6 C. M. Mate, V. J. Novotny: Molecular conformation and disjoining pressures of polymeric liquid films, J. Chem. Phys. **94** (1991) 8420–8427
28.7 C. M. Mate: Application of disjoining and capillary pressure to liquid lubricant films in magnetic recording, J. Appl. Phys. **72** (1992) 3084–3090
28.8 G. G. Roberts: *Langmuir–Blodgett Films* (Plenum, New York 1990)
28.9 A. Ulman: *An Introduction to Ultrathin Organic Films* (Academic, Boston 1991)
28.10 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)
28.11 B. Bhushan: Macro- and microtribology of magnetic storage devices. In: *Modern Tribology Handbook Vol. 2: Materials, Coatings, and Industrial Applications*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 1413–1513
28.12 Anonymous: Fomblin Z Perfluoropolyethers, Data sheet Montedism Group, Milan (2002)
28.13 B. Bhushan, Z. Zhao: Macroscale and microscale tribological studies of molecularly thick boundary layers of perfluoropolyether lubricants for magnetic thin-film rigid disks, J. Info. Storage Proc. Syst. **1** (1999) 1–21
28.14 B. Bhushan: *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)
28.15 B. Bhushan, J. N. Israelachvili, U. Landman: Nanotribology: Friction, wear and lubrication at the atomic scale, Nature **374** (1995) 607–616
28.16 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)
28.17 B. Bhushan: Self-assembled monolayers for controlling hydrophobicity and/or friction and wear. In: *Modern Tribology Handbook Vol. 2: Materials, Coatings, and Industrial Applications*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 909–929
28.18 J. Ruhe, G. Blackman, V. J. Novotny, T. Clarke, G. B. Street, S. Kuan: Thermal attachment of perfluorinated polymers to solid surfaces, J. Appl. Polym. Sci. **53** (1994) 825–836
28.19 J. Ruhe, V. Novotny, T. Clarke, G. B. Street: Ultrathin perfluoropolyether films – influence of anchoring and mobility of polymers on the tribological properties, ASME J. Tribol. **118** (1996) 663–668
28.20 V. N. Koinkar, B. Bhushan: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy, J. Vac. Sci. Technol. A **14** (1996) 2378–2391
28.21 H. Liu, B. Bhushan: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM, Ultramicroscopy **97** (2003) 321–340
28.22 G. S. Blackman, C. M. Mate, M. R. Philpott: Interaction forces of a sharp tungsten tip with molecular films on silicon surface, Phys. Rev. Lett. **65** (1990) 2270–2273
28.23 G. S. Blackman, C. M. Mate, M. R. Philpott: Atomic force microscope studies of lubricant films on solid surfaces, Vacuum **41** (1990) 1283–1286
28.24 X. Ma, J. Gui, K. J. Grannen, L. A. Smoliar, B. Marchon, M. S. Jhon, C. L. Bauer: Spreading of PFPE lubricants on carbon surfaces: Effect of hydrogen and nitrogen content, Tribol. Lett. **6** (1999) 9–14
28.25 C. A. Kim, H. J. Choi, R. N. Kono, M. S. Jhon: Rheological characterization of perfluoropolyether lubricant, Polym. Prepr. **40** (1999) 647–649
28.26 M. Ruths, S. Granick: Rate-dependent adhesion between opposed perfluoropoly(alkylether) layers: Dependence on chain-end functionality and chain length, J. Phys. Chem. B **102** (1998) 6056–6063
28.27 U. Jonsson, B. Bhushan: Measurement of rheological properties of ultrathin lubricant films at very high shear rates and near-ambient pressure, J. Appl. Phys. **78** (1995) 3107–3109
28.28 C. Hahm, B. Bhushan: High shear rate viscosity measurement of perfluoropolyether lubricants for magnetic thin-film rigid disks, J. Appl. Phys. **81** (1997) 5384–5386
28.29 C. M. Mate: Atomic-force-microscope study of polymer lubricants on silicon surface, Phys. Rev. Lett. **68** (1992) 3323–3326
28.30 C. M. Mate: Nanotribology of lubricated and unlubricated carbon overcoats on magnetic disks studied by friction force microscopy, Surf. Coat. Technol. **62** (1993) 373–379
28.31 S. J. O'Shea, M. E. Welland, T. Rayment: Atomic force microscope study of boundary layer lubrication, Appl. Phys. Lett. **61** (1992) 2240–2242
28.32 S. J. O'Shea, M. E. Welland, J. B. Pethica: Atomic force microscopy of local compliance at solid–liquid interface, Chem. Phys. Lett. **223** (1994) 336–340

28.33 B. Bhushan, T. Miyamoto, V.N. Koinkar: Microscopic friction between a sharp diamond tip and thin-film magnetic rigid disks by friction force microscopy, Adv. Info. Storage Syst. **6** (1995) 151–161
28.34 V. N. Koinkar, B. Bhushan: Micro/nanoscale studies of boundary layers of liquid lubricants for magnetic disks, J. Appl. Phys. **79** (1996) 8071–8075
28.35 B. Bhushan, S. Sundararajan: Micro/nanoscale friction and wear mechanisms of thin films using atomic force and friction force microscopy, Acta Mater. **46** (1998) 3793–3804
28.36 B. Bhushan, C. Dandavate: Thin-film friction and adhesion studies using atomic force microscopy, J. Appl. Phys. **87** (2000) 1201–1210
28.37 S. Sundararajan, B. Bhushan: Static friction and surface roughness studies of surface micromachined electrostatic micromotors using an atomic force/friction force microscope, J. Vac. Sci. Technol. A **19** (2001) 1777–1785
28.38 B. Bhushan, J. Ruan: Atomic-scale friction measurements using friction force microscopy: Part II – application to magnetic media, ASME J. Tribol. **116** (1994) 389–396
28.39 T. Stifter, O. Marti, B. Bhushan: Theoretical investigation of the distance dependence of capillary and van der Waals forces in scanning probe microscopy, Phys. Rev. B **62** (2000) 13667–13673
28.40 J. N. Israelachvili: *Intermolecular and Surface Forces*, 2nd edn. (Academic, London 1992)
28.41 S. K. Chilamakuri, B. Bhushan: A comprehensive kinetic meniscus model for prediction of long-term static friction, J. Appl. Phys. **15** (1999) 4649–4656
28.42 H. Ishigaki, I. Kawaguchi, M. Iwasa, Y. Toibana: Friction and wear of hot pressed silicon nitride and other ceramics, ASME J. Tribol. **108** (1986) 514–521
28.43 T. E. Fischer: Tribochemistry, Annu. Rev. Mater. Sci. **18** (1988) 303–323
28.44 K. Mizuhara, S. M. Hsu: Tribochemical reaction of oxygen and water on silicon surfaces. In: *Wear Particles*, ed. by D. Dowson (Elsevier, New York 1992) pp. 323–328
28.45 S. Danyluk, M. McNallan, D. S. Park: Friction and wear of silicon nitride exposed to moisture at high temperatures. In: *Friction and Wear of Ceramics*, ed. by S. Jahanmir (Dekker, New York 1994) pp. 61–79
28.46 V. A. Muratov, T. E. Fischer: Tribochemical polishing, Annu. Rev. Mater. Sci. **30** (2000) 27–51
28.47 H. Yoshizawa, Y. L. Chen, J. N. Israelachvili: Fundamental mechanisms of interfacial friction I: Relationship between adhesion and friction, J. Phys. Chem. **97** (1993) 4128–4140
28.48 H. Yoshizawa, J. N. Israelachvili: Fundamental mechanisms of interfacial friction II: Stick slip friction of spherical and chain molecules, J. Phys. Chem. **97** (1993) 11300–11313
28.49 K. C. Eapen, S. T. Patton, J. S. Zabinski: Lubrication of microelectromechanical systems (MEMS) using bound and mobile phase of Fomblin Z-DOL, Tibol. Lett. **12** (2002) 35–41

29. Kinetics and Energetics in Nanolubrication

Lubrication, one of human kind's oldest engineering disciplines, in the 19th century gained from Reynolds' classical hydrodynamic description a theoretical base unmatched by most of the theories developed in tribology to date. In the 20th century, however, increasing demands on lubricants shifted the attention from bulk film to ultrathin film lubrication. Finite size limitations imposed constraints on the lubrication process that were not considered in bulk phenomenological treatments introduced by Reynolds. At this point, as is common in many engineering applications, empiricism took over. Functional relationships derived from the classical theories were tweaked to accommodate the new situation of reduced scales by introducing "effective" or "apparent" properties.

With the inception of nanorheological tools of complementary nature in the later decades of the 20th century (e.g., the surface forces apparatus and scanning force microscopy), tribology entered the realm of nanoscience. Through an increasing confidence in experimental findings on the nanoscale, kinetic and energetic theories incorporated interfacial and molecular constraints.

The very fundamentals have been challenged in recent years. Researchers have realized that bulk perceptions, such as "solid" and "liquid" are defied on the nanoscale. The reduction in dimensionality of the nanoscale imposes constraints that bring into question the use of classical statistical mechanics of decoupled events. The diffusive description of lubrication is failing in a system that is thermodynamically not well-equilibrated.

The challenge any nanotechnological endeavor encounters is the development of a theoretical framework based on an appropriate statistics. In tribology this is met with spectral descriptions of the dynamic sliding process. Statistical kernels are being developed for probability density functions to explain anomalous transport processes that involve long-range spatial or temporal correlations. With such theoretical developments founded in nanorheological experiments, a more realistic foundation will be laid to describe the behavior of lubricants in the confined geometries of the nanometer length scale.

29.1 **Background: From Bulk to Molecular Lubrication** 885
  29.1.1 Hydrodynamic Lubrication and Relaxation 885
  29.1.2 Boundary Lubrication 885
  29.1.3 Stick Slip and Collective Phenomena 885

29.2 **Thermal Activation Model of Lubricated Friction** 887

29.3 **Functional Behavior of Lubricated Friction** 888

29.4 **Thermodynamical Models Based on Small and Nonconforming Contacts** ... 890

29.5 **Limitation of the Gaussian Statistics – The Fractal Space** 891

29.6 **Fractal Mobility in Reactive Lubrication** .. 892

29.7 **Metastable Lubricant Systems in Large Conforming Contacts** 894

29.8 **Conclusion** 895

References 895

What is inaccessible today may become accessible tomorrow as has happened by the invention of the microscope. ... Coherent assumptions on what is still invisible may increase our understanding of the visible. ... Strong reasons have come to support a growing probability, and it can finally be said the certainty, in favor of the hypothesis of the atomists. (Jean Baptiste Perrin – Nobel Lecture, December 11, 1926)

Since technology is driving lubricant films to molecular thickness, kinetic friction and its dependence on the sliding parameters – especially the sliding velocity – have become of great interest. The complexity of the frictional resistance in lubricated sliding is illustrated in Fig. 29.1 with a *Stribeck Curve*. Various regimes of lubrication can be identified in the Stribeck curve. They express to what degree the hydrodynamic pressure is involved in the lubrication process. In the ultra-low speed regime, called the *boundary lubrication regime*, no hydrodynamic pressure is built up in the lubricant. Consequently the load is carried by contact asperities coated with adsorbed lubricant molecules. If the speed is raised, a hydrodynamic pressure builds that leads to a *mixed lubrication*, in which the load is carried by both asperities and hydrodynamic pressure. At even higher speeds, elastic contributions of the solid surfaces have to be considered paired with hydrodynamic pressure effects (elastohydrodynamic lubrication), until only *hydrodynamic lubrication* matters. Hence, the Stribeck curve combines various aspects of lubrication. The curve cannot be discussed without considering the lubricant thickness and the different models of asperity contact sliding.

In one of the first comprehensive physical models of "dry" friction, Bowden and Tabor introduced a plastic asperity model, in which the material's yield stress and adhesive properties play an important role [29.1]. Considering this model, which depends on surface energies and mechanical yield properties paired with all the properties that come along with a surface adsorbent lubricant, one can hardly grasp the difficulty level involved in describing the frictional kinetics in lubrication.

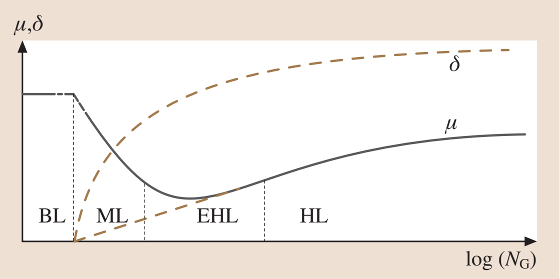

Fig. 29.1 Stribeck Curve (schematic) relates the fluid lubricant thickness, δ, and the friction coefficient μ to the Gumbel Number $N_G = \eta\omega P^{-1}$; i. e., the product of the liquid bulk viscosity, η, the sliding speed (or more precisely the shaft frequency), ω, and the inverse of the normal pressure, P. BL: Boundary Lubrication, ML: Mixed Lubrication, EHL: Elastohydrodynamic Lubrication, HL: Hydrodynamic Lubrication

Past and current engineering challenges in lubrication have been met with great and complex empiricism. The theoretical modeling of lubrication junctions generally involved only bulk property considerations with inadequately known adsorption mechanisms. The complexity of today's lubricants, most of them, such as motor oil, a product of empirical design over many years, increased exponentially, making it very difficult to meet future challenges. The problem of empiricism is that conventional laws and perceptions are unchallenged. *Effective* quantities are invented (e.g., effective viscosity), exponential fitting parameters are introduced (e.g., Kohlrausch relaxation parameter), and terminologies such as *solid* and *liquid* are taken as granted. Progress based on empiricism is only incremental and rarely revolutionary.

One of the reasons for empiricism is a lack of access to a system with fewer and better controlled parameters. In lubrication sliding that challenge has been addressed over the last two decades with the inception of the surface forces apparatus (SFA) by *Tabor* et al. [29.2] and scanning force microscopy (SFM) by *Binnig* et al. [29.3]. These two instrumental methods allow lubrication studies where roughness effects can be neglected, surface energies controlled, and wear from wearless friction distinguished. Lubricant properties can be studied at nearly mathematically described boundaries, atomistic friction events can be recorded, and fundamental models that have been considered to be mere Gedanken Experiments, such as the *Tomlinson model of friction*, can be verified. In the wake of these nanoscopic tools, exciting new theoretical lubrication and friction models have appeared.

This chapter considers these recent experimental and theoretical developments with a particular focus on sliding speed and real or apparent changes in the lubricant material properties. We will discuss kinetics and energetics in the "simplified" world of nanolubrication, in which our conventional perception is challenged. After a brief review of some of the classical lubrication concepts (hydrodynamic lubrication and boundary lubrication), we will turn our attention to a thermal activation model of friction, functional behavior of lubricated friction with velocity, and models based on small non-conforming contacts. We will critically discuss the limitation of the underlying Gaussian statistics, introduce fractal dynamics in lubrication, and will end our discussion with metastable lubricant systems.

29.1 Background: From Bulk to Molecular Lubrication

29.1.1 Hydrodynamic Lubrication and Relaxation

In the classical theories of tribology by da Vinci, Amonton, and Coulomb, not much attention was given to the dependence of kinetic friction on the sliding velocity. This clearly changed in the 19th century during the first industrial revolution, at which time lubricants became increasingly important, for instance, in ball and journal bearings. It was *Petrov* [29.4], *Tower* [29.5], and *Reynolds* [29.6] who established that the liquid viscous shear properties determine the frictional kinetics. *Reynolds* [29.6] combined the pressure-gradient determined *Poisseuille flow* with the bearing surface induced *Couette flow* assuming, based on Petrov's law [29.4], a no-slip condition at the interface between lubricant and solid. This led to the widely used linear relationship between friction and velocity. Reynolds' hydrodynamic theory of lubrication can be applied to steady state sliding at constant relative velocity and to transient decay sliding (sliding is stopped from an initial velocity v and a corresponding shear stress τ_0), which leads to the classical Debye exponential relaxation behavior, i. e.:

$$\tau = \tau_0 \exp\left(\frac{-D}{A\eta}t\right) \; ; \qquad \tau_0 \propto \frac{v\eta}{D} \, . \tag{29.1}$$

D is the lubricant thickness, A the area of the slider, and η the viscosity of the fluid. We will later see that this classical exponential relaxation behavior, obtained in a thermodynamically well-mixed three dimensional medium, is distorted when the liquid film thickness is reduced to molecular dimensions.

29.1.2 Boundary Lubrication

Reynolds hydrodynamic description of lubrication was found to work well for thick lubricant films but to break down for thinner films. One manifestation is that for films on the order of ten molecular diameters, the stress in the film does not allow the tension to return to zero. It was also found that the motion in the steady state sliding regime was disrupted, exhibiting a stick–slip-like slider motion [29.7]). Consequently, this non-Newtonian behavior was treated with a modified viscosity parameter (effective viscosity), which was composed of the pressure, temperature, and rate of shear.

The term *Boundary Lubrication* is used to describe a lubricant that is reduced in thickness to molecular dimension and effectively reduces friction between two opposing solid surfaces. *Hardy* et al. [29.8] recognized that molecular properties such as molecular weight and molecular arrangement, are governing the frictional force. This confined concept of lubrication, often visualized by two highly ordered opposing films with shear taking place somewhere in between the two layers, contains many of the rate dependent manifestations of frictional sliding; e.g., stick–slip, ultra-low friction, transitions from high to low friction, phase transitions, dissipation due to dislocations (e.g., gauche and cis-transformations), and memory effects.

Boundary lubrication was found to be in many respects unique [29.9]. In macroscopic experiments, which involved rough surfaces, friction–velocity plots resembled logarithmic functions at moderate speeds. No static stiction force peaks were observed in boundary lubricants close to zero speed. On the contrary, retractive slips could be observed upon halting, constituting a static friction coefficient exceeded by the dynamic friction coefficient [29.9]. These unique manifestations of boundary lubrication were discussed in terms of a lubricated asperity–junction mechanism, which associated "an increase in the coefficient of friction with a decrease in the adsorptive coverage of the rubbing surfaces by the lubricant substance" [29.9]. It was argued that in the course of the sliding process of a macroscopic slider, more adsorbed lubricant is expected to exist within the interfacial area than outside the contact zone. This would lead upon halting to a relaxation process of the elastic restraints on the slider, causing the slider to a retractively slip.

29.1.3 Stick Slip and Collective Phenomena

Based on numerous friction experiments at the initiation of sliding with rough macroscopic contact, it was argued that the distinction between static and kinetic friction is not categorical but rather a manifestation of the apparatus [29.9]. This was a widely held opinion prior to *Briscoe*'s et al. [29.10] molecularly smooth monolayer SFA experiments of aliphatic carboxylic acids and their soaps. Briscoe found that the character of sliding motion (continuous vs. discontinuous), depends not only on the apparatus but also on the properties (chemistry) of the monolayer. As in the rough boundary layer experiments discussed above, Briscoe's molecularly smooth monolayer experiments exhibited logarithmic-like friction–velocity behaviors.

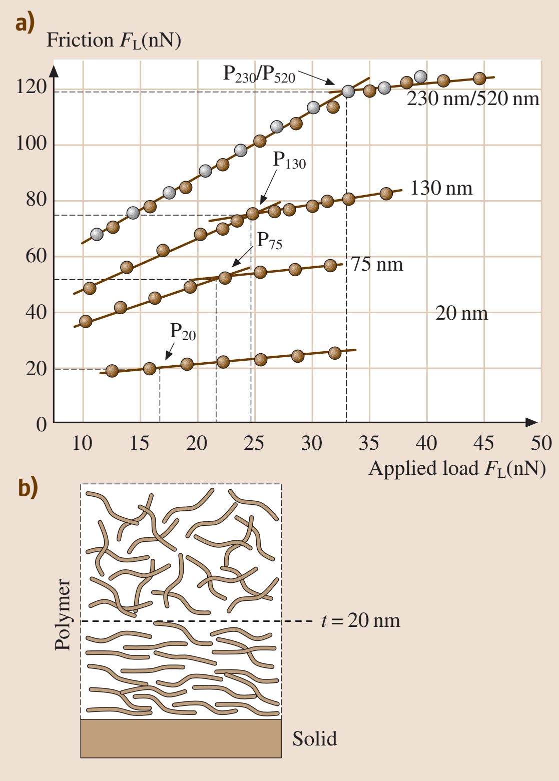

Fig. 29.2 (a) SFM friction measurements at a speed 1 µm/s: Cooperative molecular response of polyethylene co-propylene to frictional shear forces as a function of the applied load. P_t (t corresponds to the thickness of the polymer film) represents the critical activation load at which collective sliding is energetically more favorable than local plastic yielding. Adapted from [29.11]. **(b)** Sketch of the degree of disentanglement in the vicinity to the solid substrate surface

It was *Israelachvili* et al. [29.7] who, based on SFA experiments and computer simulations, provided a molecular picture of the stick–slip behavior caused by the lubricant material. The major achievement of this work was to draw our attention to the molecular structure of the lubricant, which is often different from the bulk and unstable during the sliding motion. It was recognized that bulk rheology failed to describe the lubrication process. Finally addressed in the Israelachvili study were in-plane structuring of simple liquids caused by compression forces and "freezing–melting" transitions due to shear.

The simple concept of a freezing–melting transition is based on a common perception of the two distinctive parts of a stick–slip occurrence: the solid (Hookian)-like sticking part and the liquid (Newtonian)-like slipping part. But a deformation of a solid can be both, coordinated or uncoordinated, and thus can exhibit both solid-like and liquid-like behavior. For instance, most of the plastic yielding processes are uncoordinated. On the other hand, slipping within a solid, along a crystal plane in a thermally activated strain-release process for instance, is a highly coordinated molecular process [29.12].

Similar arguments can be made for a liquid. For example, stick–slip behaviors were observed in more complex fluidic systems by *Reiter* et al. [29.13], who compared a molecularly "wet" lubricant film with a "dry" self-assembled monolayer lubricant. They concluded that sliding in liquid films is the result of slippage along an interface. In other words, the degree of molecular cooperation determined the frictional resistance.

The concept of local-*versus*-cooperative yield to shear is briefly illustrated here with a frictional-load study of a molecularly entangled polymer melt obtained in a SFM study of *Buenviaje* et al. [29.11]. Each of the curves presented in Fig. 29.2a represents a polymer film of polyethylene co-propylene of distinctly different degree of entanglement. Films of thickness above 230 nm exhibit the strongest entanglement strength. Films of 20 nm thickness or thinner are fully disentangled. The reason for the film thickness-dependent entanglement strength is given by the substrate distance-dependent shear strength during the spin coating process of the thin films. For entangled films SFM friction studies exhibit a critical applied load (identified by P_t, and the thickness t) that separates two friction regimes: One identified by a high friction coefficient and the other by a low friction coefficient. At loads below P_t the friction coefficients are high, indicating plastic yielding during sliding. In these plastic regimes of sliding, molecular cooperation is low, leading to high local shear stresses compensated by local yielding of the material. Above the critical load, the friction coefficient drops, independent of the film thickness, to a low value of 3.0, corresponding to the value obtained from the fully disentangled film. Note that the polymer molecules in the 20-nm-thick film experience high substrate tangential stresses during the spin coating process. Hence the disentangled polymer molecules can be considered to be aligned preferentially along the substrate surface as sketched in Fig. 29.2b. This leads to a decrease of the structural entropy the closer the material is to the solid substrate surface. Considering the matching friction coefficient of 0.3 above P_t for thicker films, we can assume that any entangled film above a critical load exhibits a similar molecular collective response toward shear as the 20 nm film during spin coating. The critical load and its related pressure represent a barrier that has to be overcome before a collective phenomenon is activated.

29.2 Thermal Activation Model of Lubricated Friction

With the discussion of shear in entangled polymer systems we have introduced structural entropy as one of the key players that affect frictional resistance in lubricants. We found that the structural entropy was affected by the load of the slider, which introduces an activation barrier in the form of a critical pressure. The terminology used here resembles the one of the Eyring theory of molecular liquid transport [29.14].

Eyring discussed a pure liquid at rest in terms of a thermal activation model. The individual liquid molecules experience a "cage-like" barrier that hinders molecular free motion, because of the close packing in liquids. To escape from the cage an activation barrier needs to be surmounted. In Eyring's model, two processes are considered in order to overcome the potential barrier: (i) shear stresses and (ii) thermal fluctuations. The potential barrier in the thermal activation model is depicted in Fig. 29.3 indicating the barrier modification by the applied pressure force P, and shear stress τ. *Briscoe* et al. [29.10] picked up on this idea to interpret the frictional behavior observed on molecularly smooth monolayer systems. Starting from the overall barrier height $E = Q + P\Omega - \tau\phi$ that is repeatedly overcome during a discontinuous sliding motion, using a Boltzmann distribution to determine the average time for single molecular barrier-hopping, and assuming a regular series of barriers and a high stress limit ($\tau\Phi/kT > 1$), the following shear strength versus velocity v relationship was derived [29.10]:

$$\tau = \frac{k_B T}{\phi} \ln\left(\frac{v}{v_0}\right) + \frac{1}{\phi}(Q + P\Omega) \ . \qquad (29.2)$$

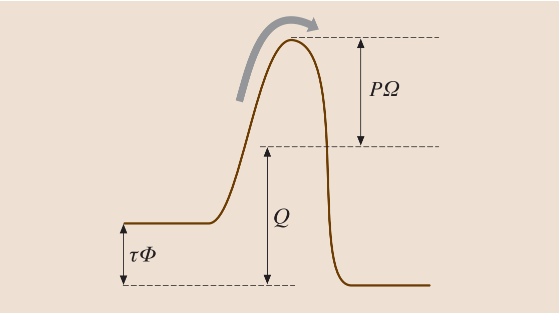

Fig. 29.3 Potential barrier in a lubricant based on Eyring's thermodynamic "cage-model." The normal pressure P and the shear stress τ are modifying the barrier height Q. Modified from [29.10]

The barrier height, E, is composed of the process activation energy Q, the compression energy $P\Omega$, where P is the pressure acting on the volume of the junction Ω, and the shear energy $\tau\phi$, where τ is the shear strength acting on the stress activation volume ϕ. T represents the absolute temperature. The stress activation volume ϕ can be conceived as a process coherence volume and interpreted as the size of the moving segment in the unit shear process, whether it is a part of a molecule or a dislocation line. The most critical parameter in (29.2), v_0, is a characteristic velocity related to the frequency of the process and to a jump distance (discussed further below).

From (29.2) the following iso-relationships can be directly deduced [29.10]:

$$\tau = \tau_0 + \alpha P \ ; \quad \tau_0 = \frac{1}{\phi}\left(k_B T \ln\left(\frac{v}{v_0}\right) + Q\right) \ ;$$

$$\alpha = \frac{\Omega}{\phi} \ ; \text{ at constant } v, T \ , \qquad (29.3a)$$

$$\tau = \tau_1 - \beta T \ ; \quad \tau_1 = \frac{1}{\phi}(Q + P\Omega) \ ;$$

$$\beta = -\frac{k}{\phi} \ln\left(\frac{v}{v_0}\right) \ , \quad \text{at constant } P, v \ , \qquad (29.3b)$$

$$\tau = \tau_2 + \theta \ln v \ ; \quad \tau_2 = \frac{1}{\phi}(Q + P\Omega - kT \ln v_0) \ ;$$

$$\theta = \frac{kt}{\phi} \ , \quad \text{at constant } P, T \ . \qquad (29.3c)$$

Thus Eyring's model predicts a linear relationship of friction (the product of the shear strength and the active process area) in pressure and temperature and a logarithmic relationship in velocity.

Eyring's model has been verified in lubrication experiments of solid (soap-like) lubricants by Briscoe and liquid lubricants by *He* et al. [29.15] within

Fig. 29.4 **(a)** Logarithmic $F_F(v)$-plots. $F_F(v) = F_0 + \alpha \ln(v[\mu m/s])$: ● "dry" contact (18% relative humidity) with $F_0 = 16.4$ nN and $\alpha = 0.91$ nN, △ OMCTS lubricated with $F_0 = 11.3$ nN and $\alpha = 3.4$ nN, and ■ n-hexadecane (n-$C_{16}H_{34}$) lubricated with $F_0 = 7.1$ nN and $\alpha = 2.5$ nN. The measurements were obtained with rectangular SFM cantilevers (0.4–0.8 N/m) at 100 nN load and 21 °C, both feedback-controlled. **(b)** Stress activation length, ϕ/A. (A area of contact) for OMCTS, n-hexadecane and dry contact. The *inset* provides a linear relationship between friction and temperature at a velocity of 1 µm/s and a normal load of 100 nN. Adapted from [29.15]

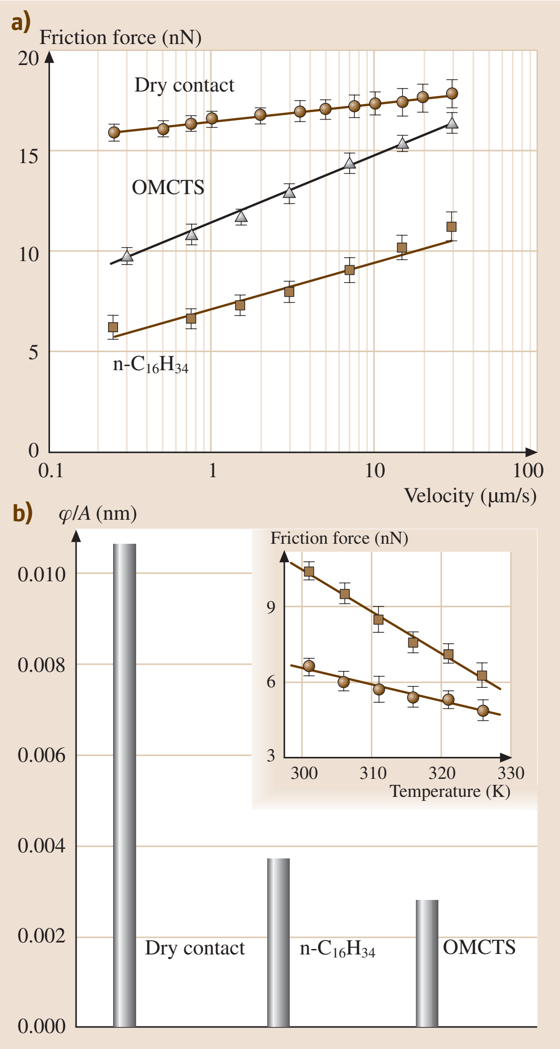

three logarithmic decades of velocities. While *Briscoe* et al. [29.10] employed a SFA that confines and pressurizes the film over several square microns, He et al. used a SFM system in which the contact is on the order of the lubricant molecular dimension.

He et al. determined the degree of interfacial structuring and its effect on lubrication of n-hexadecane and octamethylcyclotetrasiloxane (OMCTS). For spherically shaped OMCTS molecules, only an interfacial "monolayer" was found; in contrast, a 2-nm-thick entropically cooled layer was detected for n-hexadecane in the boundary regime to an ultra-smooth silicon wafer. SFM measurements of the two lubricants (with similar chemical affinity to silicon) identified the molecular shape of n-hexadecane responsible for augmented interfacial structuring. Consequently, interfacial liquid structuring was found to reduce lubricated friction, Fig. 29.4. Again as reasoned above, these results can be discussed in terms of a collective phenomenon, i. e., in terms of increased molecular coordination in n-hexadecane versus OMCTS.

29.3 Functional Behavior of Lubricated Friction

Friction-rate experiments are well suited to evaluate the rheological nature of interfacial liquids. In classical theories, such as the Reynolds' hydrodynamic theory discussed above, drag forces in lubricated sliding over thick liquid films were found to depend linearly on the rate of sliding and on the viscosity of the bulk fluid. In high-pressure lubrication, described by the elastohydrodynamic lubrication theory [29.16, 17], it was found that the linear relationship between friction and velocity can be retained by adjusting the (apparent) viscosity by introducing an *apparent viscosity term*. Qualitatively, the same linear relationship has been observed for highly confined simple liquids between ultrasmooth mica surfaces such as alkanes [29.18]. Note that the lubricated contact area in SFA experiments is on the micron-scale. It significantly exceeds the size of the confined molecules. For small and unbranched molecules, such as simple alkanes, it is possible that the confined material undergoes a pressure-induced phase reconstruction, which leads to material properties that deviate significantly from the bulk. Larger and more complex (branched) molecules are less likely to exhibit pressure-induced phase reconstruction due to internal constraints and poor mixing within the contact area. This was shown by *Drummond* et al. [29.19] in SFA shear experiments. They found that the linear friction–velocity dependence does not apply for branched hydrocarbon lubricants. Also Drummond discussed "molecular lubrication" in terms of a logarithmic friction–velocity relationship, which is in accordance with the above-discussed thermal activation model, the solid lubricant SFA study by *Briscoe* et al. [29.10], and the liquid lubricant SFM study by *He* et al. [29.15].

Common to the three studies by *Briscoe*, *He*, and *Drummond* is that they operate on a single material phase that is disrupted or relaxed over a very specific lateral length scale. In the Eyring model, the length scale is deduced by assuming a regular series of barriers, separated by a *virtual jump distance*. The distance is embedded in v_0, the characteristic velocity, which is the product of the jump distance and the frequency of the process. *Briscoe* et al. [29.10] used the lattice constant of the highly oriented monolayers as the virtual jump distance. It was assumed that the process frequency was related to the vibrational frequency of the molecules ($10^{11}\ \mathrm{s}^{-1}$), neglecting sliding velocity, temperature, and pressure effects. *He* et al. [29.15] assumed a jump distance of 0.2 nm and considered frequencies between a perfectly structured alkane layer (10^{11} Hz) and the bulk fluid (10^{13}–10^{15} Hz, estimated from infrared absorption data for typical covalent bonds). With these assumptions *He* determined total "jump-energies" of $4–8\times10^{-20}$ J. *Briscoe* and *He* pointed out that a friction–velocity study alone provides only a qualitative measure of the microscopic origin of friction. Additional measurements have to be conducted that quantitatively address jump distances and frequencies.

The issue of the jump distance has been addressed by *Overney* et al. [29.20] in a SFM study on a highly ordered lubricant model system. This study avoided two levels of difficulties *Briscoe* et al. [29.10] and *He* et al. [29.15] encountered: (a) large contact areas of SFA studies, and (b) complex rheology with unknown structure parameters as in liquid lubricant studies. It involved contact dimensions on the order of 1 nm^2, and the crystalline form a bilayer model-lipid-lubricant with in-plane lattice spacings of 0.6 and 1.1 nm. The study mainly focused on the effect of the depth of the corrugation potential (barrier height) on the static and dynamic friction force. This is illustrated in Fig. 29.5 in the form of stick–slip amplitude plotted as a function of the drag direction (i. e., sliding with respect to the anisotropic row-like film structure). Relevant to our discussion about jump distance in lubrication events is Overney's discussion about the sliding speed and its effect on the slip distance. They demonstrated that within sliding speeds of 36 nm/s to 100 nm/s, the jump distance corresponded to the lattice spacing. At higher velocities, however, they could observe jumps over multiples of lattice distances and found the jump length distribution to become increasingly stochastic. They proposed molecular (or atomistic) friction as a white-noise driven system, which obeys a Gaussian fluctuation–dissipation relation. Hence, based on this finding one should consider discussing kinetic friction in terms of a statistical fluctuation model and understand the jump distance as a statistical quantity.

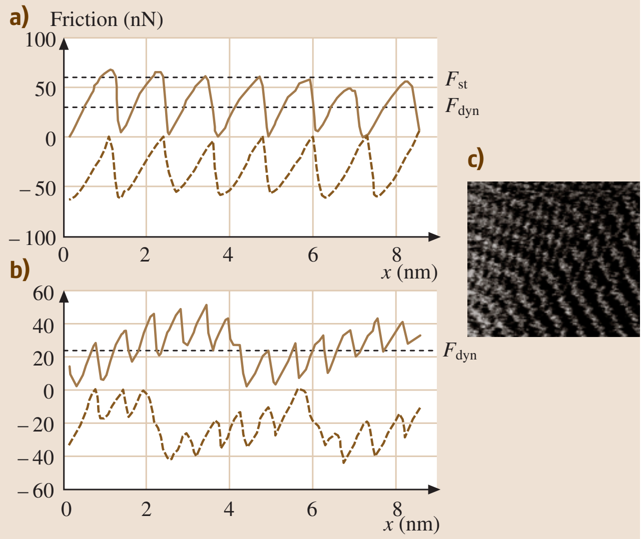

Fig. 29.5 SFM molecular stick-slip measurements of a bilayer lipid system (5-(4'-N,N-dihexadecylamino)benzylidene barbituric acid). **(a)** High amplitude frictional stick-slip behavior is observed for scans perpendicular to molecular rows as imaged in **(c)**. F_{st}, static friction, is assigned to the maximum force occurrence. The average value corresponds to the dynamic friction value, F_{dyn}, determined on large-scale micrometer scans. **(b)** A 30° out of row direction scan leads to decreased frictional stick-slip behavior due to smaller molecular corrugations. **(c)** $12\times12\ \mathrm{nm}^2$ SFM lateral force image of a highly structured lipid bilayer. Two crystalline domains with a boundary are imaged. The anisotropic row-like structure is responsible for directional dependent friction forces. The molecular corrugation between the rows is larger than the molecular corrugation in between a single row. **(a)** and **(b)** are adapted from [29.20] and **(c)** from [29.21].

29.4 Thermodynamical Models Based on Small and Nonconforming Contacts

The SFM approach simulates a single asperity contact with a very high compliance, provided by a microfabricated and etched ultra-sharp tip and a typically soft cantilever spring. From a realistic, tribological perspective, the SFM approach is targeted toward the study of the intrinsic lubricant properties of a thin film in close vicinity to the solid substrate. The small contact area on the order of the lubricant's molecular dimension allows discussing SFM results in terms of a thermodynamic equilibrium. The area is insufficient in reorganizing the lubricant molecules coherently, to cause an apparent material phase-transition, or to generate a metastable situation as observed in SFA experiments (see below). SFM is therefore not appropriate to reflect on tribological issues involving large area confinement effects.

In our prior molecular discussion of friction above, we introduced for solid and liquid lubrication a thermal activation model, the *Eyring model*, which employed a regular series of potential barriers. Note that the concept applies for a solid lubricant of an inherent, highly ordered structure (e.g., [29.10]), but also for a liquid system in which the series of potentials is built up and overcome in the course of the shear process (e.g., [29.15]).

Gnecco et al. [29.22] showed in a ultrahigh vacuum study on sodium chloride that the concept of the Eyring model also applies for dry SFM friction studies. Thus a molecular theory of lubricated friction involving a molecular contact could be derived from a very simplistic model of an apparent sinusoidal-corrugated surface over which a cantilever tip is pulled. In a first attempt one could assume that the corresponding wave length of the shear process corresponds to the apparent lattice spacing of the corrugated surface. With such a simple attempt it is, however, assumed that there is no noise, such as thermal noise, existing in the system, and thus the driven tip leaves the total potential well when the barrier vanishes at the instability point. In the presence of noise, the transition to sliding can be expected to occur before the top of the barrier is reached. Such barrier-hopping fluctuations have been theoretically discussed by *Sang* et al. [29.23] and *Dudko* et al. [29.24].

The relationship between thermal fluctuations and velocity must be handled thoughtfully. *Sang* et al. [29.23] pointed out that in previous considerations of thermal fluctuations by *Heslot* et al. [29.25], the fluctuations were proportionally related to the velocity, which led to a friction force that is logarithmically dependent on the velocity. In Heslot's *linear creep model*, the barrier height is proportional to the frictional force. *Sang* argued that if one considered an absorbing boundary condition (i. e., an elastic deformation of the overall potential which is accomplished by shifting the x-axis) the barrier height becomes proportional to a 3/2-power law in the friction force. *Sang*'s extended linear creep model resembles a *ramped creep model* and leads analytically to a logarithmic distorted dynamic friction-*versus*-velocity relationship; i. e.,

$$F = F_c - \Delta F \left|\ln v^*\right|^{2/3} . \qquad (29.4)$$

In (29.4) v^* represents a dimensionless velocity, $\Delta F \propto T^{2/3}$, and F_c is an experimentally determined constant (by plotting F versus $T^{2/3}$ for a fixed ratio $T/v = 1\,\mathrm{K/(nm/sec)}$) that contains the critical position of the cantilever support. The same relationship of friction with velocity was also derived for the maximum spring force by *Dudko* et al. [29.24]. ΔF and v^* in (29.4) were derived as follows by *Sang*:

$$v^* = 2\left(\frac{v\beta\omega_0^2 U_0}{k_B T\,\lambda}\right)\frac{\Omega_k^2}{\left(1-\Omega_k^4\right)^{1/2}} ;$$

$$\Omega_k = \frac{\omega_0}{2\pi\omega_k} ;$$

$$\omega_0 = \sqrt{\frac{M\lambda^2}{U_0}} ; \quad \omega_k = \sqrt{\frac{M}{k}} , \qquad (29.5a)$$

$$\Delta F = \frac{\pi U_0}{\lambda}\left(\frac{3}{2}\frac{k_B T}{U_0}\right)^{\frac{2}{3}}\left(\frac{\left(1-\Omega_k^4\right)^{\frac{1}{6}}}{1+\Omega_k^2}\right) . \qquad (29.5b)$$

In (29.5a,b) v is the velocity of the cantilever stage, β is the microscopic friction coefficient or dissipation (damping) factor, ω_0 is the frequency of the small oscillations of the tip in the minima of the periodic potential, λ is the lattice constant, U_0 is the surface barrier potential height, M and k represent the mass and the spring constant of the cantilever, respectively, and $2\pi\Omega_k$ represents the ratio of ω_0 with the intrinsic cantilever resonance frequency ω_k.

Sang's and *Dudko*'s model was experimentally confirmed by *Sills* and *Overney* [29.26] on an unstructured amorphous surface of atactic polystyrene, Fig. 29.6. *Dudko* determined that the typically used weak spring constants in SFM measurements are responsible for

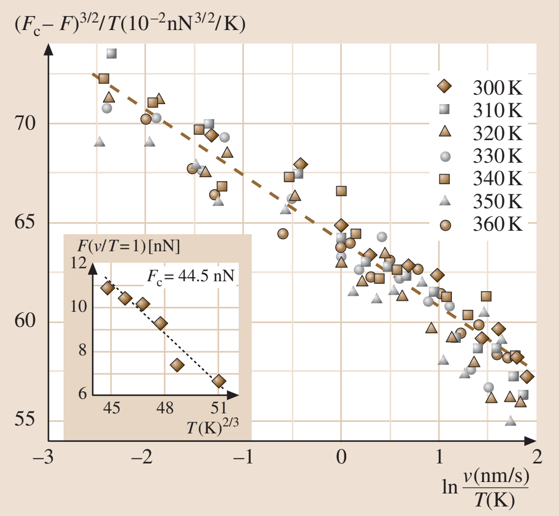

Fig. 29.6 Collapse of SFM friction data obtained on atactic polystyrene using the ramped creep scaling model. The regression parameters from the linear fit (*dashed line*) are $-2.158 \times 10^{-2}\,\mathrm{N}^{3/2}\,\mathrm{s/m}$ and $40.186 \times 10^{-2}\,\mathrm{nN}^{-3/2}/\mathrm{K}$ ($R^2 = 0.9124$). Lower *inset*: the constant F_c is determined from the intercept of the friction force F versus $T^{2/3}$ for a fixed ratio $T/v = 1\,\mathrm{K/(nm/s)}$: $F_c = 44.5\,\mathrm{nN}$. (Adapted from *Sills* and *Overney* [29.26])

the more pronounced logarithmic behavior of friction in velocity as found by *Gnecco* et al. [29.22] and *He* et al. [29.15]. The ramped creep model is also supported by numerical solutions of the Langevin equation.

The Langevin equation combines the equation of motion (including the sinusoidal potential and perfect cantilever oscillator in the total potential energy E) with the thermal noise in the form of the random force, $\xi(t)$, i. e.,

$$M\ddot{x} + M\beta\dot{x} + \frac{\partial E(x,t)}{\partial x} = \xi(t)\,, \qquad (29.6)$$

$$\text{where } E(x,t) = \frac{k}{2}\left(R(t) - x\right)^2 - U_0\left(\frac{2\pi x}{\lambda}\right)\,.$$

Equation (29.6) was solved numerically by both *Sang* et al. [29.23] and *Dudko* et al. [29.24] independently, assuming a Gaussian fluctuation–dissipation relation, $\langle \xi(t)\xi(t')\rangle = 2M\beta k_b T\delta(t-t')$ to express the random force. *Sang* confirmed the *ramped creep model*, and *Dudko* showed that a force reconstruction approach from the density of states (accumulated from the corresponding Fokker–Planck equation) is equivalent to the Langevin equation. From the dynamic spectral analysis, it could be concluded that the locked states (states within the potential wells) contribute mostly to the potential component of the friction force that dominates at low driving velocities, and sliding states contribute to viscous friction dominating at high driving velocities [29.24].

It should be noted that all of the above results considered an overdamped SFM system with respect to the driven spring (i. e., $\beta^2 > 4\,\mathrm{kM}$), and an underdamped system with respect to the periodic potential (i. e., $\beta^2 < 4(M\omega)^2$). This aspect will be further addressed below in our discussion of metastable lubricant systems in large conforming contacts.

29.5 Limitation of the Gaussian Statistics – The Fractal Space

The spectral description of dynamic processes involving probability density functions has recently been the focus of numerous theoretical papers that treat various statistical kernels [29.27–30]. *Dudko*'s et al. [29.24] Fokker–Planck discussion of kinetic friction and *Luedtke*'s et al. [29.27] Lévy flight model of slip diffusion of adsorbed nanoclusters are two examples in which statistical methods are applied to describe diffusive properties relevant to the kinetics in tribology.

Currently most models used to describe tribological processes assume Gaussian statistics (e.g., [29.23], and [29.24]). One of the limitations of a Gaussian statistics is that there are no correlations between statistical incidences. In other words, the Gaussian dynamic system is without memory. This is important to remember as the Eyring model discussed above, with its equally spaced potential barriers, used a Gaussian statistics. Simple "inert" lubricants, such as short chain alkanes embedded between silicon wavers, are described satisfactorily with such a statistics; however, confined complex liquids are not, including branched molecules, polymers, and generally chemically interactive and entropically confined systems (e.g., perfluropolyether lubricants as discussed below).

Confined complex liquids, for instance, easily exhibit strongly interacting glass-like behavior. The dynamic and stress relaxation behaviors in glasses, frequently discussed only as a low interacting system with

Arrhenius laws (Gaussian statistics), are often distorted from processes described by independently occurring microscopic processes. For instance, deviations from the Debye exponential relaxation as introduced in (29.1) are expressed in the form of an *extended exponential* Kohlrausch relaxation function over time t; i. e.,

$$F(t) \equiv \left[\frac{X(t)-X(\infty)}{X(0)-X(\infty)}\right] = \mathrm{e}^{-\left(\frac{t}{\tau}\right)^{b}} ; \quad 0 < b < 1 , \tag{29.7}$$

where X is the property that is relaxed. The exponent b, the Kohlrausch exponent, can theoretically be determined if one assumes that the process occurs in series, representing a well-determined microscopic origin that correlates the various degrees of freedom [29.31]. This approach is borrowed from magnetic spin models such as the *Ising spin model* [29.32]. The idea is that a given molecular motion is dependent on the availability of other degrees of freedom of mobile neighboring structural units. Finite relaxation times, $t_{\max}$, are gradually obtained with increasing spin levels (ergodic limit).

As mentioned above, the models by *Dudko* et al. [29.24] and *Sang* et al. [29.23] assumed Gaussian statistics. To illustrate how a diffusion process can deviate from Gaussian statistics, we introduce a simplified version of the Langevin equation, i. e.,

$$\ddot{x} = -\eta\dot{x} + \varsigma(t) \tag{29.8}$$

with the coordinate x, the dissipation (or dampening) parameter η, and the random acceleration $\zeta(t)$. Assuming Gaussian statistics, the mean squared displacement is

$$\left\langle x^2(t)\right\rangle = 2k_{\mathrm{B}}T\eta t = 2Dt , \tag{29.9}$$

where $D = \eta kT$ defines the diffusion constant [29.33]). It was already realized at the time of Smoluchovski at the beginning of the 20th century that a diffusive description of a dynamic process demands a thermodynamically well-equilibrated or mixed system. Especially in a confined tribological system that involves a third medium (e.g., a lubricant), it can be expected that the Markovian nature of the underlying stochastic process could be disturbed. Consequently, for a monolayer lubricant that is chemically interacting, a nonlinear relationship of the mean squared displacement in time can be expected. Manifestations of anomalous transport are long-range spatial or temporal correlations. Two extreme limits can be distinguished: (a) processes with strong temporal relations ("fractal time") [29.29], and (b) systems that exhibit long jumps ("Lévy flights") [29.30].

29.6 Fractal Mobility in Reactive Lubrication

The importance of the underlying kinetics is illustrated by ultra-thin wetting lubricants. The spreading of "completely wetting" polymer liquids on solid surfaces has revealed unexpected spatial and temporal features when examined at the molecular level. The spreading profile is typically characterized by the appearance of a precursor film of monomolecular thickness extending over macroscopic distances and, in many cases, a terracing (also on the order of molecular dimensions) of the fluid remaining in the reservoir [29.34]. These spatial features have been shown to be consistent with a Poiseuille-like flow in which the disjoining pressure gradients with film thickness drive the spreading process [29.35]. The temporal evolution of the spreading profile in this film thickness regime is, however, found to universally scale as $t^{1/2}$ even at short times [29.34]. That the spreading dynamics are reflective of a diffusive transport mechanism and not of a pressure driven "liquid" flow suggests that interfacial confinement substantially alters the mobility of molecularly thin polymer fluids [29.36].

The molecular mobility is of fundamental importance for monolayer lubrication purposes, such as in magnetic storage devices. It has, for instance, been shown that for low surface energy hydroxyl-terminated perfluoropolyether (PFPE-OH) films, the lubricant exhibits spatially terraced flow profiles indicative of film layering [29.35] and spreading dynamics that are diffusive in nature [29.37, 38]. In magnetic storage devices the hydroxylated chain ends of molecularly thin PFPE-OH films interact with the solid surface, an amorphous carbon surface, via the formation of hydrogen-bonds with the polar, carbon-oxygen functionalities located on the carbon surface. The bonding of the PFPE-OH polymer to carbon is predicated on the ability of the PFPE backbone to deliver spatially the hydroxyl end-group to within a sufficiently close distance to the surface active sites. Kinetic measurements probing the bonding of the PFPE-OH polymer to the carbon reveal two distinctive kinetic behaviors, as illustrated in Fig. 29.7 at two representative temperatures: 50 °C and 90 °C for the

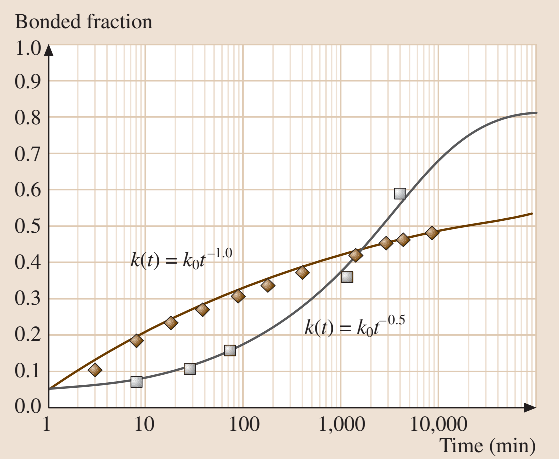

Fig. 29.7 Representative kinetic data for the bonding of PFPE-OH (tradename: Fomblin ZDOL) to amorphous carbon at $T = 50\,°\mathrm{C}$ and $T = 90\,°\mathrm{C}$. *Solid lines* represent fits using a rate coefficient of the form, $k(t) = k_0 t^{-\alpha}$ with $\alpha = 0.5$ for $T = 50\,°\mathrm{C}$ and $\alpha = 1.0$ for $T = 90\,°\mathrm{C}$

two temperature regimes below 56 °C and above 85 °C. Below 56 °C the kinetics are described with a time-dependent (fractal time dependent) rate coefficient of the form

$$k(t) = k_0 t^{-1/2} \tag{29.10}$$

and at temperatures above 85 °C with the form

$$k(t) = k_0 t^{-1.0}\,. \tag{29.11}$$

The initial bonding rate constants, k_0, increased abruptly as the temperature rose above 50 °C.

The bonding kinetics in the low-temperature regime is characteristic of a diffusion-limited reaction occurring from a glass-like state of the molecularly thin PFPE-OH film [29.39]. The mobility of the PFPE chain in the glass-like state is limited by the propagation of holes or packets of free volume, which facilitate configurational rearrangements of the chain. The onset of changes in the bonding kinetics at nominally $T > 56°$ signifies a fundamental change in the mobility of the molecularly thin PFPE-OH film. Specifically, the transition in the fractal time dependence suggests that delivery of the hydroxyl moiety to the surface is no longer limited by hole diffusion, and the increase in the initial rate constant indicates an enhancement in the backbone flexibility. These results are consistent with a transition in the film from a glass-like to a liquid-like state in which the enhanced PFPE-OH segmental mobility results from rotations about the ether oxygen linkages in the chain that become increasingly facile. The time dependence observed in the high-temperature rate coefficients, $k(t) = k_0 t^{-1.0}$ is characteristic of a process occurring from a confined liquid-like state in which the activation energy increases as the extent of the reaction increases.

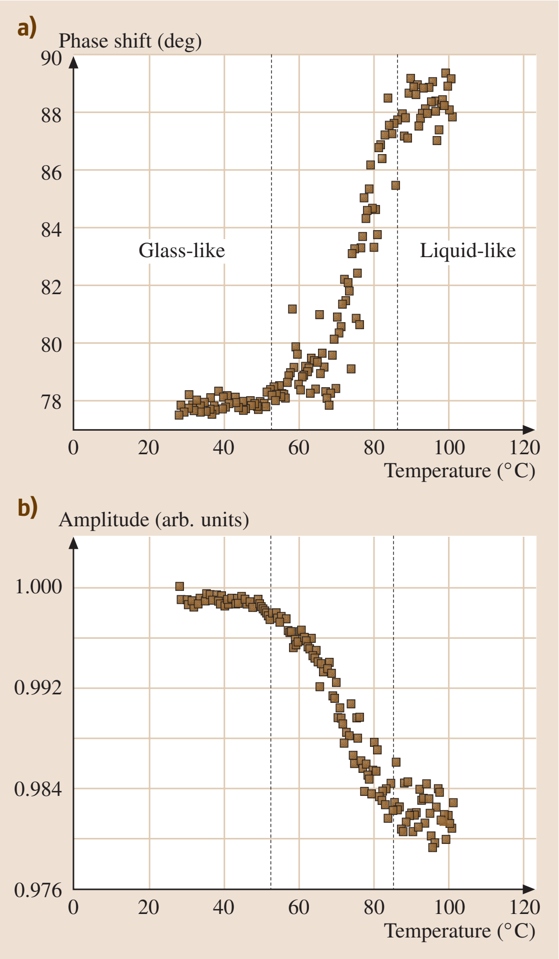

Fig. 29.8a,b Shear-modulated SFM experiments performed on a 10.7 ± 0.5 Å Fomblin Zdol film: **(a)** phase shift response between disturbance and response, and **(b)** contact stiffness response vs. temperature

The impact of this transition in the molecular mobility on tribology can be illustrated with sinusoidally modulated shear force experiments [29.40, 41]. In brief, a molecularly thin (10.7 ± 0.5 Å) PFPE-OH film is subjected to a local shear stress by means of a sinusoidal force applied *laterally* to a SFM probe (at constant load) where the modulation amplitude is initially set below that required to initiate sliding between the tip and the sample. The amplitude and phase-shift responses are measures of the contact stiffness and the effective viscous dampening, respectively. From Fig. 29.8 it can be inferred that the SFM-measured nanorheological properties of the PFPE-OH film exhibit the changes discussed above in the molecular mobility. Thus kinetic and rheological data suggest that the thermal transition observed is due to the formation of a two-dimensional (2-D) glass. The "glass transition" results from the preferential "freezing out" of the out-of-plane torsional motions of the energetically confined PFPE backbone. The confinement-induced solidification in the molecularly thin precursor film will significantly impact the lubrication properties and challenge thermodynamic, well-equilibrated models of lubrication as introduced above.

29.7 Metastable Lubricant Systems in Large Conforming Contacts

It is important to note that in an experiment of thermal activation, the critical time of the experiment t_{exp} decides the system response with its finite relaxation time $t_{\max}$. If $t_{\mathrm{exp}} > t_{\max}$, the system behaves in an ergodic manner, and thermodynamic laws apply for interpreting lubrication results. On the contrary, if $t_{\mathrm{exp}} < t_{\max}$ the thermal evolution cannot be described by classical statistical thermodynamics. A metastable configuration is generated. The experimental time depends strongly on the contact area, the parameter that most differs between SFA and SFM measurements. SFA experiments involve large micron-scale contacts while SFM measurements are conducted with contacts on the nanoscale

Because of the large contact area, SFA experiments are very susceptible to generating unequilibrated metastable lubricant configurations. The SFA study by *Yoshizawa* et al. [29.42], in which distinctively different *dynamic states of friction* were introduced, could be interpreted as such. To date there are three velocity regimes used to describe the dynamic state of friction for a system that is "underdamped" [29.43]. In an underdamped system, realized by a stiff spring compared to the friction constant, the characteristic slip time is comparable or smaller than the response time of the mechanical system. One distinguishes three velocity regimes, as depicted in Fig. 29.9. The three regimes distinguish themselves by a single discriminator, v_{c}, a material, pressure and temperature dependent critical velocity. The regimes are described as

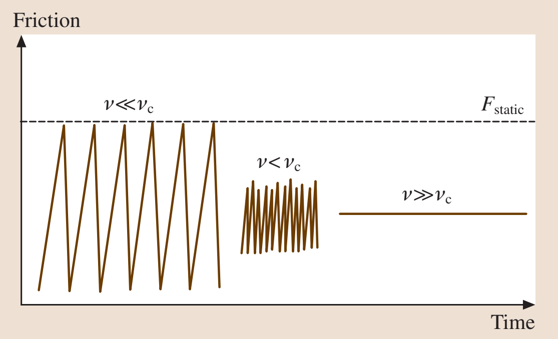

Fig. 29.9 Dynamic states of friction for an underdamped spring system

1. highly regular with high amplitude stick–slip for $v \ll v_{\mathrm{c}}$,
2. intermittent stochastic stick–slip for $v < v_{\mathrm{c}}$, and
3. smooth low friction sliding for $v > v_{\mathrm{c}}$.

Various models have been suggested to describe the different dynamic states, including melting freezing transition [29.44], chain adsorption on substrates [29.45], and embedded particle [29.43]. Until the embedded particle model by Rozman was introduced, the SFA approach seemed to be the tool of choice to investigate metastable lubricant configurations. Rozman's single particle model alerts us, however, to drawing unambiguous conclusions on the dynamical structure of a molecular system embedded between two plates and driven by an underdamped system. In Rozman's simple theoretical model, sketched in Fig. 29.10, a single particle is embedded between two corrugated

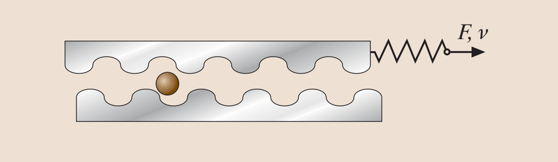

Fig. 29.10 Model of single particle embedded in a two-wave potential driven at constant force, F, and velocity v

surfaces (*two-wave potential*). The top plate is pulled at constant velocity by a linear spring, and the plate motion is monitored. Interestingly, the plate exhibits exactly the same dynamic state of frictional motion as introduced above with the three regimes, which were attributed to a rate-dependent configurational change of the lubricant.

Finally it shall be noted that the slip relaxation pattern depends condition under which the stick–slip motion is being studied. In an *overdamped* system, i. e., a system in which the spring constant is weak compared to the friction constant, *Rozman* et al. [29.46] found that one can control the experimentally observed relaxation pattern of the slip by controlling the spring constant.

29.8 Conclusion

Starting from a classical tribological Master Curve, the Stribeck curve, with its complex description of lubrication for thin lubricants, we launched a discussion of molecular lubrication pulling together disparate approaches to studying friction. We found that phenomenological descriptions of lubrication, such as the Reynolds' theory, were kept alive for ultra-thin lubricants by "adjusting" material properties, such as the viscosity. This is a common engineering approach to introduce "effective" or "apparent" properties, if tested models fail to describe new situations.

In the case of molecular lubrication, significant progress has been made in the last ten years. Two instrumental techniques have in particular contributed to this progress: the SFA and the SFM. The contributions of these two techniques have been complementary. While the SFA has tested lubricants under pressure constraints with large contacts in respect to the size of the trapped molecules, the SFM has probed the degree of collective mobility with disturbances in the size of the molecules themselves.

One feature common to interpreting lubrication results, and which was discussed here in detail, is the problem of finding the appropriate underlying statistics to describe the lubrication process. Most of the current molecular models that have been used to describe molecular friction and lubrication assumed a Gaussian statistics. Only recently has it been recognized important to consider statistics that embrace long-range spatial or temporal correlations. In the near future, it can be expected that the next leap in an improved fundamental understanding of kinetics and energetics in nanolubrication will come from interpretations that are challenging the Markovian nature of the underlying stochastic process.

References

29.1 F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids* (Clarendon, Oxford 1951)
29.2 D. Tabor, R. H. S. Winterton: The direct measurement of normal and retarded van der Waals forces, Proc. R. Soc. Lond. A **312** (1969) 435–450
29.3 G. Binnig, C. F. Quate, C. Gerber: Atomic Force Microscope, Phys. Rev. Lett. **56** (1986) 930–933
29.4 N. P. Petrov: *Friction in Machines and the Effect of the Lubricant* (St. Petersburg 1883)
29.5 B. Tower: First report on friction experiments (friction of lubricated bearings), Proc. Inst. Mech. Eng. (November 1883) 632–659
29.6 O. Reynolds: On the theory of lubrication and its application to Mr. Beauchamp tower's experiments, including an experimental determination of the viscosity of olive oil, Philos. Trans. Roy. Soc. Lond. **177** (1886) 157–234
29.7 J. P. Israelachvili, G. M. McGuiggan, M. Gee, A. Homola, M. Robbins, P. Thompson: Liquid dynamics in molecularly thin films, J. Phys. **2** (1990) 89–98
29.8 W. B. Hardy, I. Doubleday: Boundary lubrication – The paraffin series, Proc. Roy. Soc. Lond. A **100** (1922) 550–574
29.9 A. Dorinson, K. C. Ludema: *Mechanics and Chemistry in Lubrication* (Elsevier, Amsterdam 1985)
29.10 B. J. Briscoe, D. C. B. Evans: The shear properties of Langmuir–Blodgett layers, Proc. Roy. Soc. Lond. A **380** (1982) 389–407
29.11 C. Buenviaje, S. Ge, M. Rafailovich, J. Sokolov, J. M. Drake, R. M. Overney: Confined flow in polymer films at interfaces, Langmuir **19** (1999) 6446–6450
29.12 S. Blunier, H. Zogg, A. N. Tiwari, R. M. Overney, H. Haefke, P. Buffat, G. Kostorz: Lattice and thermal misfit dislocations in epitaxial $CaF_2/Si(111)$ and

BaF$_2$/CaF$_2$/Si(111) structures, Phys. Rev. Lett. **68** (1992) 3599–3602
29.13 G. Reiter, A. L. Demirel, J. Peanasky, L. L. Cai, S. Granick: Stick to slip transition and adhesion of lubricated surfaces in moving contact, J. Chem. Phys. **101** (1994) 2606–2615
29.14 S. Glasstone, K. J. Laidler, H. Eyring: *Theory of Rate Processes* (McGraw-Hill, New York 1941)
29.15 M. He, A. Szuchmacher Blum, G. Overney, R. M. Overney: Effect of interfacial liquid structuring on the coherence length in nanolubrication, Phys. Rev. Lett. **88**(15) (2002) 154302/1–4
29.16 K. L. Johnson: *Contact Mechanics* (Cambridge Univ. Press, Cambridge 1987)
29.17 E. Meyer, R. M. Overney, K. Dransfeld, T. Gyalog: *Nanoscience: Friction and Rheology on the Nanometer Scale* (World Scientific, Singapore 1998)
29.18 H. K. Christenson, D. W. R. Gruen, R. G. Horn, J. N. Israelachvili: Structuring in liquid alkanes between solid-surfaces – Force measurements and mean-field theory, J. Chem. Phys. **87**(3) (1987) 1834–1841
29.19 C. Drummond, J. Israelachvili: Dynamic behavior of confined branched hydrocarbon lubricant fluids under shear, Macromolecules **33**(13) (2000) 4910–4920
29.20 R. M. Overney, H. Takano, M. Fujihira, W. Paulus, H. Ringsdorf: Anisotropy in friction and molecular stick–slip motion, Phys. Rev. Lett. **72** (1994) 3546–3549
29.21 R. M. Overney, H. Takano, M. Fujihira: Elastic compliances measured by atomic force microscopy, Europhys. Lett. **26**(6) (1994) 443–447
29.22 E. Gnecco, R. Bennewitz, T. Gyalog, C. Loppacher, M. Bammerlin, E. Meyer, H.-J. Güntherodt: Velocity dependence of atomic friction, Phys. Rev. Lett. **84**(6) (2000) 1172–1175
29.23 Y. Sang, M. Dube, M. Grant: Thermal effects on atomic friction, Phys. Rev. Lett. **87**(17) (2001) 174301/1–4
29.24 O. K. Dudko, A. E. Filippov, J. Klafter, M. Urbakh: Dynamic force spectroscopy: A Fokker–Planck approach, Chem. Phys. Lett. **352** (2002) 499–504
29.25 F. Heslot, T. Baumberger, B. Perrin, B. Caroli, C. Caroli: Creep, stick–slip, and dry-friction dynamics: Experiments and a heuristic model, Phys. Rev. E **49** (1994) 4973–4988
29.26 S. Sills, R. M. Overney: Creeping friction dynamics and molecular dissipation mechanisms in glassy polymers, Phys. Rev. Lett. **91** (2003) 095501(1–4)
29.27 W. D. Luedtke, U. Landman: Slip diffusion and Levy flights of an adsorbed gold nanocluster, Phys. Rev. Lett. **82** (1999) 3835–3838
29.28 I. M. Sokolov: Levy flights from a continuous-time process, Phys. Rev. E **63** (2000) 011104/1–10
29.29 R. Metzler, J. Klafter: The random walks guide to anomalous diffusion: A fractional dynamics approach, Phys. Rep. **339** (2000) 1–77
29.30 R. Metzler, J. Klafter: Levy meets Boltzmann: Strange initial conditions for brownian and fractional Fokker–Planck equations, Physica A **302** (2001) 290–296
29.31 R. G. Palmer, D. L. Stein, E. Abrahams: Models of hierarchically constrained dynamics for glass relaxation, Phys. Rev. Lett. **53** (1984) 958–961
29.32 N. W. Ashcroft, N. D. Mermin: *Solid State Physics* (CBS Asia, Philadelphia 1976)
29.33 R. Becker: *Theorie der Wärme* (Springer, Berlin 1985)
29.34 F. Heslot, N. Fraysse, A. M. Cazabat: Molecular layering in the spreading of wetting liquid drops, Nature **338** (1989) 640–642
29.35 T. E. Karis, G. W. Tyndall: Calculation of spreading profiles for molecularly-thin films from surface energy gradients, J. Non-Newtonian Fluid Mech. **82** (1999) 287–302
29.36 S. F. Burlatsky, G. Oshanin, A. M. Cazabat, M. Moreau: Microscopic model of upward creep of an ultrathin wetting film, Phys. Rev. Lett. **76** (1996) 86–89
29.37 T. M. O'Connor, Y. R. Back, M. S. Jhon, B. G. Min, D. Y. Yoon, T. E. Karis: Surface diffusion of thin perfluoropolyalkylether films, J. Appl. Phys. **79** (1996) 5788–5790
29.38 X. Ma, J. Gui, L. Smoliar, K. Grannen, B. Marchon, C. L. Bauer, M. S. Jhon: Complex terraced spreading of perfluoropolyalkylether films on carbon surfaces, Phys. Rev. E **59** (1999) 722–727
29.39 A. Plonka, J. Bednarek, K. Pietrucha: Reaction dynamics in glass transition region: propagating radicals in ultraviolet-irradiated poly(methyl methacrylate), J. Chem. Phys. **104** (1996) 5279–5283
29.40 R. M. Overney, C. Buenviaje, R. Luginbuehl, F. Dinelli: Glass and structural transitions measured at polymer surfaces on the nanoscale, J. Therm. Anal. Calorimetry **59** (2000) 205–225
29.41 S. Ge, Y. Pu, W. Zhang, M. Rafailovich, J. Sokolov, C. Buenviaje, R. Buckmaster, R. M. Overney: Shear modulation force microscopy study of near surface glass transition temperature, Phys. Rev. Lett. **85**(11) (2000) 2340–2343
29.42 H. Yoshizawa, P. McGuiggan, J. N. Israelachvili: Identification of a second dynamic state during stick–slip motion, Science **259** (1993) 1305–1308
29.43 M. G. Rozman, M. Urbakh, J. Klafter: Stick-slip motion and force fluctuations in a driven two-wave potential, Phys. Rev. Lett. **77** (1996) 683–686
29.44 J. M. Carlson, A. A. Batista: Constitutive relation for the friction between lubricated surfaces, Phys. Rev. E **53** (1996) 4153–4164
29.45 Y. Braiman, F. Family, H. G. E. Hentschel: Array-enhanced friction in the periodic stick–slip motion of nonlinear oscillators, Phys. Rev. E **53** (1996) R3005–R3008
29.46 M. G. Rozman, M. Urbakh, J. Klafter: Controlling chaotic frictional forces, Phys. Rev. E **57** (1998) 7340–7343

Part E
Industrial

Part E Industrial Applications and Microdevice Reliability

30 Nanotechnology for Data Storage Applications
Dror Sarid, Tucson, USA
Brendan McCarthy, Tucson, USA
Ghassan E. Jabbour, Tucson, USA

31 The “Millipede” – A Nanotechnology-Based AFM Data-Storage System
Gerd K. Binnig, Rüschlikon, Switzerland
G. Cherubini, Rüschlikon, Switzerland
M. Despont, Rüschlikon, Switzerland
Urs T. Duerig, Rüschlikon, Switzerland
Evangelos Eleftheriou, Rüschlikon, Switzerland
H. Pozidis, Rüschlikon, Switzerland
Peter Vettiger, Rüschlikon, Switzerland

32 Microactuators for Dual-Stage Servo Systems in Magnetic Disk Files
Roberto Horowitz, Berkeley, USA
Tsung-Lin Chen, Shin Chu, Taiwan
Kenn Oldham, Berkeley, USA
Yunfeng Li, Berkeley, USA

33 Micro/Nanotribology of MEMS/NEMS Materials and Devices
Bharat Bhushan, Columbus, USA

34 Mechanical Properties of Micromachined Structures
Harold Kahn, Cleveland, USA

35 Thermo- and Electromechanics of Thin-Film Microstructures
Martin L. Dunn, Boulder, USA
Shawn J. Cunningham, Colorado Springs, USA

36 High Volume Manufacturing and Field Stability of MEMS Products
Jack Martin, Cambridge, USA

37 MEMS Packaging and Thermal Issues in Reliability
Yu-Ting Cheng, HsinChu, Taiwan
Liwei Lin, Berkeley, USA

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

30. Nanotechnology for Data Storage Applications

This chapter considers atomic force microscopy (AFM) as an enabling technology for data storage applications, considering already existing technologies such as hard disk drives (HDD), optical disk drives (ODD) and Flash Memories that currently dominate the nonvolatile data storage market, together with future devices based on magnetoresistive and phase change effects.
The issue at hand is the question of whether the novel AFM-based storage, dubbed "Probe Storage", can offer a competing approach to the currently available technologies by playing the role of a disruptive technology. Probe Storage will be contrasted to HDD and ODD who are purely mechanical, as they are based on a rotating disk that uses just a single probe to address billions of bits of data, and nonvolatile RAM that has no moving parts yet requires billions of interconnects. In particular, capacity, areal density, transfer rate, form factor and cost of various data storage devices will be discussed and the unique opportunity offered by Probe Storage in employing massive parallelism will be outlined. It will be shown that Probe Storage bridges the gap between HDD, ODD and other nonvolatile RAM, drawing from the strength of each one of these and adding a significant attribute neither of these has; namely, the possibility of addressing a very large number of nanoscale bits of data in parallel. This chapter differs from the other chapters in this book in that it addresses the important issue of whether a given scientific effort, namely, Probe Storage, is mature enough to evolve into a commercially viable technology. The answer seems to indicate that there indeed is a huge niche in the data storage arena that such a technology is uniquely qualified to fill, which is large enough to justify a major investment in research and development. Indeed, as other chapters indicate, such an effort is developing at a rapid pace, with hopes of having a viable product within a few years.

30.1 **Current Status of Commercial Data Storage Devices** 901
30.1.1 Non-Volatile Random Access Memory 904

30.2 **Opportunities Offered by Nanotechnology for Data Storage** 907
30.2.1 Motors 907
30.2.2 Sensors 909
30.2.3 Media and Experimental Results 913

30.3 **Conclusion** 918

References 919

This chapter will differ from the other chapters in this handbook in that it addresses the important issue of whether a given scientific effort is mature enough to evolve into a commercially viable technology. Specifically, we consider scanning probe microscopy (SPM), consisting of scanning tunneling microscopy (STM) [30.1–3] and atomic force microscopy (AFM) with all their variants, as a means for storing and retrieving nanoscale bits of data to and from a substrate. Indeed, many studies have already been published in which researchers demonstrated the feasibility of using SPM techniques that hold promise for realistic applications [30.4]. However, in approaching a topic from a practical point of view, that is, whether it has the ingredients that will spawn a commercial product, it is imperative to obtain a clear view of the status of competing technologies. The commercially available technologies that will concern us here entail two classes of devices, one consisting of hard disk drives (HDD) and optical disk drives (ODD), and the other consisting of Non-Volatile Random Access Memories (NVRAM) based on charge trapping (Flash Memory) that currently dominate the data storage market. The latter may soon be supplemented by magnetoresistive RAM (MRAM) and phase change RAM (PC-RAM). For brevity, we will refer to the first class as HDD and the one as NVRAM.

It is commonly thought that nanotechnology is a breakthrough technology that is a quantum leap beyond existing technologies in its capabilities. In reality, however, nanotechnology is in many cases just a limiting case to the already commercially available technologies. The latter may progress at such a high rate that the distinction between what is considered "conventional" and "nano" becomes blurred. A case in point is data storage in which distances and times are already specified by nanometers and nanoseconds. For example, the head of an HDD in a PC's hard disk flies across the storage medium (platter) at a speed of 10 m/s ($\sim$ 22 miles per h) at a height of $\sim$ 10 nm, reading and writing bits of data on a nanosecond time scale. Also, both giant magnetoresistance (GMR) devices that read magnetic data in an HDD and transistors belonging to NVRAM devices use nanoscale structures, one employing spintronic effects and the other tunneling junctions between pairs of transistors, where quantum laws of physics have to be used to describe their operation.

The issue at hand is, therefore, whether novel nanotechnologies can offer a competing approach to currently available technologies by playing the role of a "disruptive technology". This term has been coined in the high-tech industry to describe a situation where, figuratively speaking, a discovery developed in a basement can lead to a development of a new device that has the potential of becoming a commercial alternative to a commonly used one. Such a disruption of the success of a commercially established technology by a product developed by a new start-up, for example, has been witnessed more and more frequently in recent years. One would even venture to assume that the support companies give to research groups at universities can be attributed to their concern about a possible disruption of their own technology.

The important issue here is the question of what nanotechnology can offer to an already mature field such as data storage. As will be discussed later, HDD and NVRAM are worlds apart in technology. HDD is purely mechanical, as it is based on a rotating disk that uses just a single probe (head) to address billions of bits of data. NVRAM, on the other hand, has no moving parts yet requires billions of probes, i. e., interconnects, positioned at the intersection of horizontal and vertical interconnects with one probe for every bit of data. Another distinction between HDD and NVRAM is the time it takes to access a random bit of data. For an HDD the random access time is 10 ms (75 ms for ODD), while for NVRAM it is less than one µs. Clearly, NVRAM beats HDD (and ODD) in the random access arena, if that is the criteria for performance. On the other hand, the rate at which an HDD can read data sequentially is rather impressive. The current performance of an HDD stands at 436 Mb/s and 22 Mb/s for ODD, comparable to what NVRAM can offer, being of the order of 10 MByte/s. Another playing field in the battle between HDD and NVRAM performance concerns capacity and areal density of bits of data. Here, capacity measures the total number of bytes of data, while areal density measures the number of bits per square inch. The first addresses the ever growing need for storing larger and larger amounts of data, say, for high-definition TV applications, while the latter is related to what is commonly called the form factor, namely, the physical size of a storage device. The form factor is intimately connected to portability, as more and more hand-held devices have restrictions on the physical size of each of its components. Note also that a major consideration in comparing HDD and NVRAM is cost, a topic that will be addressed later.

As discussed before, both HDD and ODD use just a single probe that addresses all the bits of data on a platter where data is stored. NVRAM, on the other hand, has to resort to a grid of interconnects that contacts each bit of information individually. Such a grid, in which each bit is interfaced with one "word line" and one "bit line", occupies expensive real estate across a storage device, and, even more importantly, it requires that the interconnects be on the order of the bits of data they address. As long as these bits have dimensions on the order of, say, 1 µm, one is able to use conventional lithography for the fabrication of the interconnects. However, when the bits of data shrink to, say, 25 nm in size, interconnects are no longer a viable option due to lithographic limitations. HDD and ODD, in contrast, require no interconnects, so this limitation on the size of bits is not a concern. However, there are many other problems that arise as one tries to push the size of the stored bits to nanoscale dimensions, as will be discussed later.

By now we have touched on the heart of the challenge that nanotechnology-based storage faces if it is to compete with HDD and NVRAM in both performance and cost. Practically, one should restate this issue and ask a more modest question, namely, what niche nanotechnology will fill in consumer data-storage applications. To address this question, one notes that nanotechnology is concerned with the ability to miniaturize and characterize nanometer-sized structures. For data storage, the question is how to address nanoscale data bits rapidly without having to resort to the use of too many interconnects.

A possible solution to this challenge emerged from a brilliant idea developed at IBM and Stanford laboratories, as part of their scanning probe microscopy research [30.5–9].

A conventional AFM employs a single cantilever with a sharp tip at its end to raster scan a sample and its minute deflections, due to protrusion on the surface, generate a map of the surface topography. Under ideal conditions one can even obtain atomically resolved images. The IBM-Stanford group demonstrated that it is possible to employ a large number of such cantilevers and operate them in parallel. Such an operation provides a faster means of obtaining images of different parts of a sample simultaneously. These cantilever arrays are often referred to as types of microelectromechanical systems (MEMS), or if they are small enough, nanoelectromechanical systems (NEMS).

One can view this breakthrough concept as an enabling technology for data storage applications in that it bridges the gap between HDD and NVRAM technologies, drawing from the strength of each. To appreciate this concept, consider a medium whose area is sectioned into a thousand squares, each containing one million bits of data. Each square is addressed by its own interconnect whose width can now be as wide as a thousand bits of data. Consider now a square structure consisting of one thousand cantilevers whose tips are aligned such that each one addresses its own square. Such an arrangement, known as probe storage, and dubbed "millipede" by IBM, has some attributes belonging to an HDD, some attributes belonging to NVRAM, and potentially one significant attribute neither of these has: the possibility of addressing nanoscale bits of data. While random access is faster than that of a HDD, it is slower than that of NVRAM.

In this chapter, we will address the issues mentioned in the introduction with an emphasis on a comparison between existing and future development of conventional technologies and nanotechnologies. We will not address more futuristic approaches such as DNA and molecular electronics for data storage applications – fascinating topics on their own merit, yet deemed too far in the future for our more "technological" approach.

30.1 Current Status of Commercial Data Storage Devices

Before describing the current status of commercial storage devices, it is worthwhile to reiterate definitions and cover some new key features associated with HDD [30.10]. A schematic of a hard disk drive, shown in Fig. 30.1, is configured into sectors, each having radial tracks; there are typically 58,000 tracks per in. for HDD and 34,000 tracks per in. for ODD, 562 kbits per in. for HDD, and 100 kbits per in. for ODD. In 1993, a HDD head was flying above the storage media at a typical height of 100 nm, which dropped to 10 nm in 2002, and is expected to drop to 4 nm in 2004.

Hard Disk Drives and Optical Disk Drives

The time it takes the read-write head to get to a random bit along a track is called latency, while the time it takes the head to get to a random sector is called seek time. The time to get to any random bit is called access time. From 1990 to 2002 this time decreased linearly from 20 ms to 8 ms, and is expected to decrease to 4 ms by 2004.

Transfer time is the time it takes to transfer a given number of bits from the hard drive to a given destination in a computer. The total number of bits per unit area, usually given in bpsi (bits per square inch), is called areal density. The total number of bytes in a storage device is called capacity.

The physical size of a storage device and its price are paramount factors in its commercial viability. Disk drives are usually divided into magnetic and optical, although a combination of both is currently advanced by several companies and research groups around the world. A magnetic hard drive (HDD) has a higher capacity, lower cost, and higher performance than an optical hard drive (ODD), but its disk is not removable. The ODD, on the other hand, has a removable disk that can be mass produced by parallel replication (stamping), making it cheaper for large-scale distribution. Also, an ODD is configured to operate in such a way that the head reads data sequentially, while HDD is optimized for random access operation. Note that units of bits are usually denoted by a lower case "b" and used for areal density and transfer rate, while units of bytes are usually denoted by an upper case "B" and used for capacity. A note in passing: Areas are, oddly enough, expressed in square inches, and the holy grail for areal density is denoted by 1 tbpsi (one terabit per square inch).

Now, a crucial question when assessing the viability of a promised technology concerns the demand for the commercial product in question. Figure 30.2 estimates that a grand total of 2.12 million terabytes of information were produced in 1999 [30.11]. The breakdown is 83 terabytes for optical, 240 terabytes for paper, 427,000 terabytes for film, and 1.6 million terabytes for magnetic data storage. PC disk drives take the first place in demand: 700,000 terabytes of data!

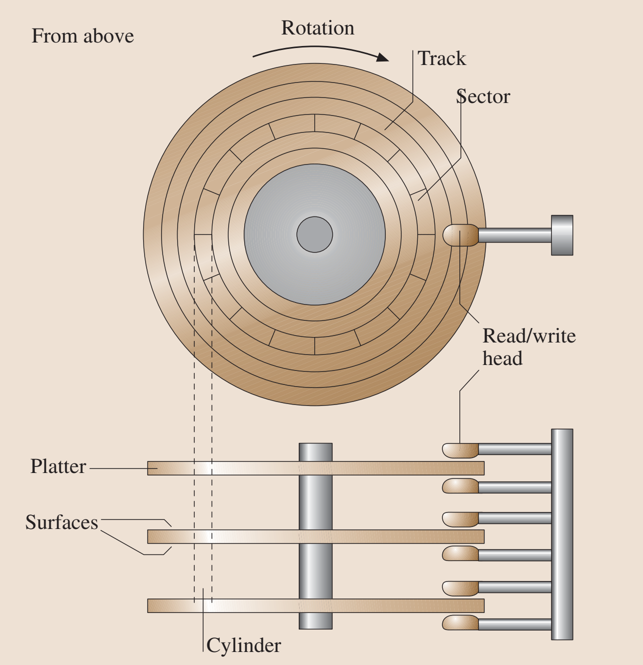

Fig. 30.1 A schematic diagram of a hard disk drive illustrating the concepts of cylinder, track, and sector

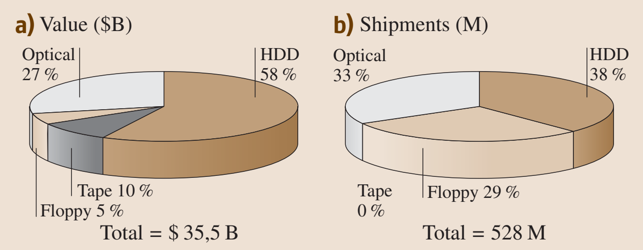

Fig. 30.3a,b Storage drive industry shipments in [30.10]. **(a)** Categorized by value and **(b)** units shipped

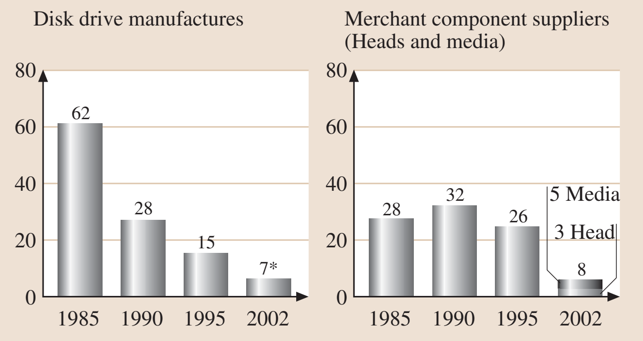

Fig. 30.4 Disk drive manufacturers and merchant component suppliers [30.12]. The number of disk drive manufacturers decreased from 62 in 1985 to a mere 7 by 2002

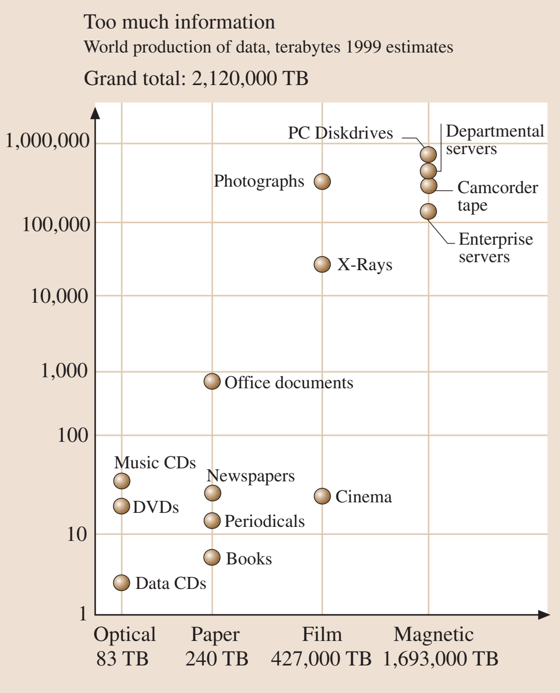

Fig. 30.2 World production of data in 1999 estimates

To appreciate the economic impact of the HDD storage drive industry (NVRAM will be discussed later), consider Fig. 30.3, which is current to 2002 [30.10]. Here, one observes that the industry shipped 528 million units with a value of $ 35.5 billion, 58% of the value consisting of HDD and 27% of ODD, while the numbers shipped are almost evenly divided between HDD, ODD, and floppies. Based on Fig. 30.3, one would think that such a huge market would involve a large number of companies competing with each other in price and performance.

However, as shown in Fig. 30.4, the cut-throat competition among various companies gave rise to the disappearance of many of them and mergers between the others [30.12]. Note that the number of disk drive manufacturers dropped from 62 in 1985 to only seven in 2002, while merchant components suppliers dropped from 28 to eight during that period! For example, Maxtor and Quantum merged, IBM sold its disk drive division to Hitachi, and Fujitsu exited desktop HDD altogether. In 2002, Seagate shipped 1,959,000 terabytes of storage, 10.2 million enterprise drives, 1.2 million 15,000 rpm drives, and 44.8 million personal storage devices. And they expect that there will be migration to smaller form factors and increased areal density in all segments of the industry and that mobility will be become increasingly important.

A projection of HDD areal density on a logarithmic scale is shown in Fig. 30.5 [30.13]. Whereas in 1990 the areal density was only 1 gigabit per sq. in., it reached 100 gigabits per square inch in 2002 and is projected to reach 1 Tbpsi in 2006. One wonders whether it will take nanotechnology to get to 50 Tbpsi in 2017, as the figure suggests. As quoted from a Seagate presentation, "in traditional applications, areal density has grown faster than consumption, moving to a single platter society in PC's that enables smaller form factors. In non-traditional markets, new applications drive capacity growth, especially with video content." [30.12].

Table 30.1 [30.10] presents several key parameters characterizing HDD and ODD that should serve as a reality check for the potential viability of any nanotechnology approach that aims to act as a disruptive technology. Thus any new concept should be judged against this current status of HDD and ODD and, as described later, NVRAM. Note that in contrast to these impressive parameters associated with HDD and ODD, seek time is limited to 10 ms and 75 ms, respectively.

A comparison of the perceived limits to HDD and ODD performance are presented in Table 30.2 [30.10]. For magnetic disk drives, the limit to areal density is the superparamagnetic effect in which too small a magnetic domain can be thermally excited into an opposite direction, losing the reliability of the stored data. Ways and means for approaching this limit and maintaining reliability will be discussed later. For optical data storage, the conventional limit is clearly the spot size of the focused laser beam. However, near-field methods are currently being developed to circumvent this limit. As far as speed goes, switching times for magnetic domains is around 1 ns, while for laser-induced thermal effects it is ten times slower.

The current and projected relationship between areal density and data rate for several technologies are

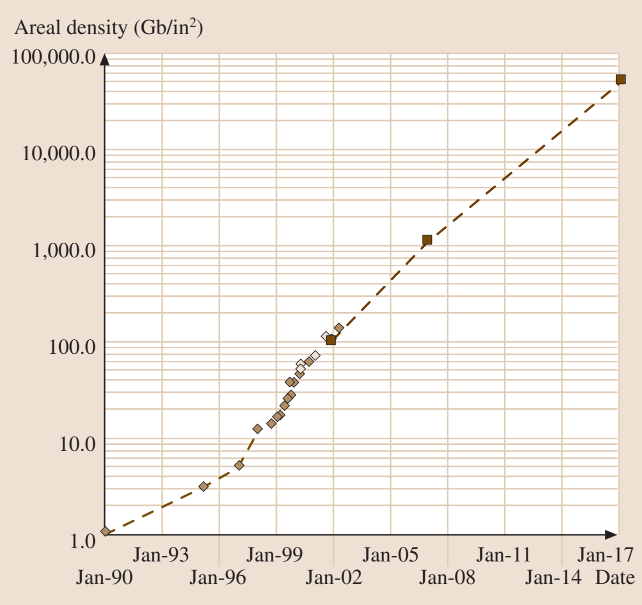

Fig. 30.5 Projected areal density for hard disk drives [30.13]

Table 30.1 A comparison of the parameters of HDD and ODD

Product comparison		
Parameter	HDD (80 Gb desktop)	ODD (4.7 Gb DVD-RAM)
Areal density	32.6 Gb/in.^{2}	3.3 Gb/in.^{2}
	Best = 50.0 Gb/in.^{2} (2.5″)	Best = 4.6 Gb/in.^{2} (50 mm)
TPI/BPI	58 K/562 K	34 K/100 K
Capacity/platter	40 Gb (2 side, 95 mm)	9.4 Gb (2 side, 120 mm)
Unit price (drive)	\$80–\$120	\$365
Price/Gb (drive)	\$1–\$1.5	\$39
Price/Gb (media)	–	∼ \$1.5–\$2
Seek time	10 ms	75 ms
Transfer rate (writing)	436 Mb/s	22 Mb/s
Research laboratory comparison		
Parameter	HDD (longitudinal)	ODD (blue near-field)
Areal density	130 Gb/in.^{2}	45 Gb/in.^{2}
TPI/BPI	213 K/610 K	141 K/319 K

Table 30.2 Perceived technical limits for hard disk drive technology [30.10]

	Magnetic	Optical
Areal density (Gb/in.^{2})	**Superparamagnetism** 1994: 35 (scaling) 1997: 100 (lower bit aspect ratio) 2002: 500 (perpendicular) 1,000 (patterned) > 1,000 (HAMR)	**Focused spot size** 1994: 0.7 ($\lambda = 780$ nm) 1997: 3.3 ($\lambda = 650$ nm) 2002: 19 (Blu-Ray) 2× (ML) 2–3× (NF) ? × (Volume)
Switching time (ns)	**Gyromagnetic ratio** ~ 1	**Media crystallization** ~ 10

shown in Fig. 30.6 [30.5]. These technologies consist of CD-ROM, magneto-optical, magnetic disks, magnetic tape, SIL, holography, and thermo-mechanical AFM. According to this projection, magnetic disks are in the forefront of technology and are expected to have the same areal density as that of the probe storage concept, yet have a much faster data rate. The latter technology, as predicted in this figure, will grow from a data rate of only 1 Mb/s to more than 10 Mb/s, accompanied by an increase in areal density from 10 gigabits per sq. in. to hundreds of gigabits per sq. in. More about the probe storage concept will be discussed later.

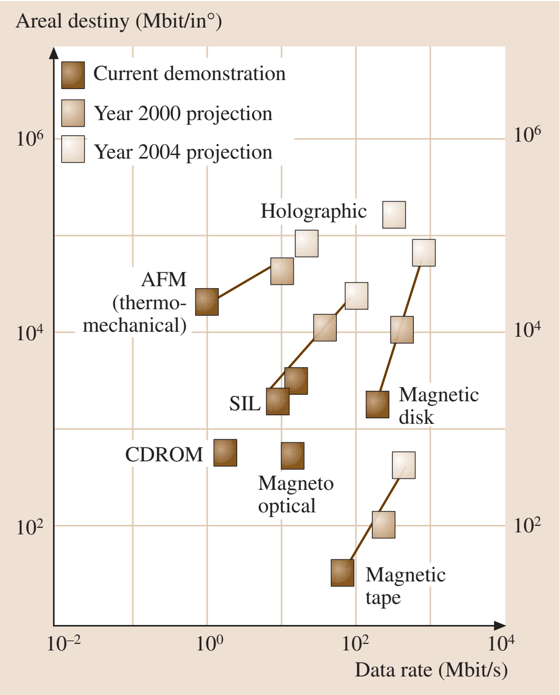

Fig. 30.6 Projected areal density vs. data rate for several technologies

30.1.1 Non-Volatile Random Access Memory

Demand for Flash Memory has been growing rapidly over the last few years, driven by the rapid expansion of several markets, including mobile communications and portable media devices (notably cameras and MP3 players). Flash Memory is by now the fastest growing segment of the memory market. Figure 30.7 [30.14] shows the total value of the Flash Memory market year by year. It can be seen that the market is increasing by nearly a factor of ten every five years and up to 100% every year. The market is projected to reach $ 23 billion by 2003, up from $ 2.5 billion in 1998. The reasons for this dramatic growth are depicted in Fig. 30.8 [30.14], which shows the demand by industry sector and the growth in these areas. By far the biggest driver of demand for Flash Memory is consumer electronics and, in particular, cellular phones, which account for nearly half the market and are growing at a compounded annual growth rate of 120%.

Flash Memory

Flash Memory, the most common non-volatile, programmable memory device, uses rows and columns of interconnects that address each data cell. At the intersection of these are two transistors, which comprise the Flash Memory cell, as shown in Fig. 30.9. Each cell is addressed by a network of interconnects, where the common connection to the transistor gates is called the word

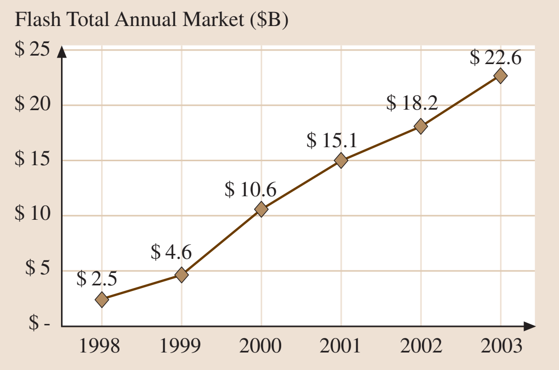

Fig. 30.7 Total value of the flash memory market, 1998–2003. The market has been growing rapidly over the last few years, and is now the fastest growing sector of the memory market

line and the common connection to the drain is called the bit line. Every cell in the memory array can be accessed by activating a unique combination of these lines. Here, the floating gate is linked to a word line via a second gate that serves as a control gate. The floating gate potential is altered by Fowler–Nordheim tunneling of electrons, an effect that takes place when the applied bias needed for switching is larger than the work function of the relevant electrode. The electrons arrive from the column, or bit line, and enter the floating gate from which they drain to the ground. A cell sensor monitors the flow of electrons through the gate. A value of 1 is assigned when the current is greater than a given threshold value, and 0 if it is below that value. Approximately 30,000 electrons are stored in the gate to make a 1 and 5,000 for a 0. The memory of the device derives from the very long storage time (tens of years) of the tunneling electrons in the gate capacitor. For the data to have a lifetime of ten years, the electrons can leak at a rate of no more than five a day. Flash memory can only be rewritten a limited number of times (10^{5-6}), as electrons get permanently trapped in the gate over time, impairing device efficiency. Memory devices using this technology are SmartMedia, CompactFlash cards, and Memory Sticks, for example, that read and write memory in 256- or 512-byte increments, enhancing their speed of operation. A Flash Memory device, in contrast to a hard disk drive, is noiseless, has no moving parts, is lighter and smaller, and, above all, has access times orders of magnitude faster than HDD. SmartMedia cards, which serve as solid-state floppy disk cards, were originally developed by Toshiba. They are currently available in a capacity of 128 Mbytes and higher and measure roughly 45 mm × 37 mm × 1 mm. In comparison to a HDD, however, SmartMedia cards are much more expensive. Note that a single storage cell occupies roughly 0.1 μm^2, a size limited by the complexity of its physical structure and width of the addressing interconnects.

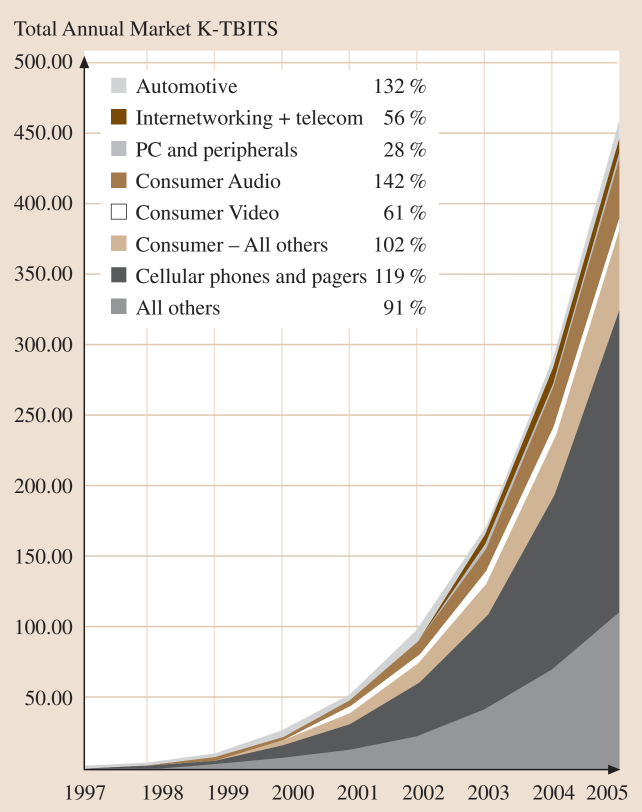

Fig. 30.8 Market segments that are driving the demand for flash memory where cellular phones are leading the demand

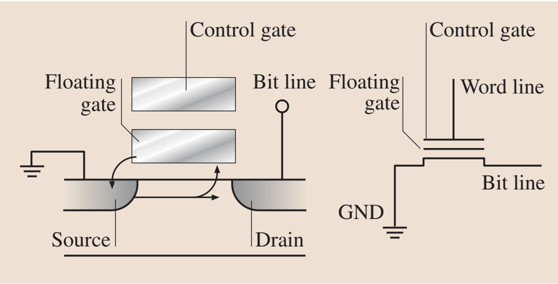

Fig. 30.9 Schematics diagram of a typical Flash Memory cell

Magnetoresistive Random-Access Memory

Figure 30.10 [30.15] is a schematic diagram of a magnetoresistive random-access memory (MRAM) that acts as a non-volatile storage device. As with Flash Memory, there are word lines and bit lines at the intersection of which is a cell with a switchable property. Whereas with Flash Memory the electric potential of a storage cell is modified by electron tunneling, here it is the magnetic domain orientation that is switched by the current flowing through the electrodes. Passage of electrons through the magnetic domain of the cell depends on the state of polarization of the spins of the injected electrodes, much the same as the operation of giant magnetoresistance elements used to read data in a hard disk drive. In contrast to HDD and Flash Memory, MRAM is still in a stage of development and, as such, is not yet commercially available. One of its main attributes is its high speed of programming, which is faster than that of Flash Memory. The main challenges confronting MRAM technology are: (i) the size of the cell is still too large, (ii) the power consumption required during the write mode is still too high, and (iii) there are manufacturing difficulties associated with high temperature steps that may damage the magnetic layers.

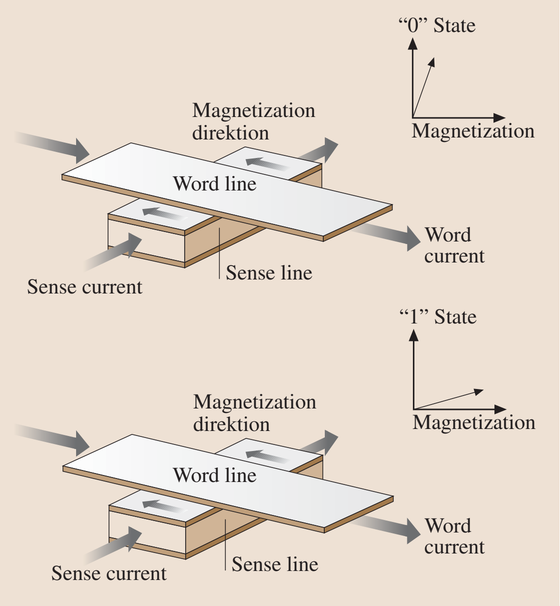

Fig. 30.10 A diagram of a MRAM cell

To exemplify the status of this technology, note that Motorola demonstrated a 1-Mbit MRAM test chip using a 0.6-µm process, yielding a cell size of 7.2 µm^2. Projections are that using a 0.18-µm process will shrink the cell size to about 0.7 µm^2, and thus be competitive with Flash Memory. Also, Sony demonstrated an MRAM test chip using a 0.35-µm process that yielded a cell size of 5.8 µm^2. Other companies like IBM, which is partnering with Infineon, are also active in MRAM research.

Phase-Change Random-Access Memory

Phase-change random-access memory is a technology that uses a chalcogenide as the data storage material in the memory cell. The medium is similar to the switchable material used in rewritable CDs. It is being pioneered by Ovonyx, in association with Intel. A schematic diagram of an Ovonyx Unified Memory (OUM) is shown in Fig. 30.11 [30.16]. The OUM is an example of a phase-change random-access memory that operates much as the Flash Memory and MRAM, as far as the electrodes are concerned, except that the memory bit cell consists of a chalcogenide material that undergoes a phase change. Upon passing current from the word line to the bit line, the chalcogenide, a GeTeSb alloy, changes its phase reversibly between amorphous and crystalline states, depending on the temperature it is heated to and the rate of cooling. This change in phase is accompanied by a several orders of magnitude change in conductivity, which makes it possible to obtain a high reading signal-to-noise ratio. Similar to Flash Memory and MRAM, the OUM requires a protective shield from the environment for reliable operation over a long period. We shall come back to this issue later.

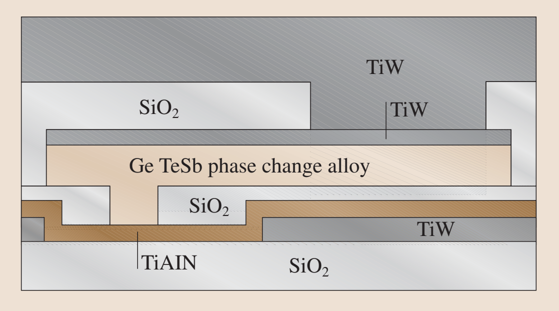

Fig. 30.11 A cross section of an OUM cell

30.2 Opportunities Offered by Nanotechnology for Data Storage

We have by now summarized the two mainstream commercial data storage devices, HDD and NVRAM, and the two near-term candidates for future commercial success, the MRAM and PC-RAM devices. As stated before, the technology utilized by these devices should be used as a measure against which novel nanotechnology approaches are judged. Note that even if a promising novel nanotechnology-based device comes into being, it will take time to bring it into maturity, and during that period, the state-of-the-art of HDD, Flash Memory, MRAM, and PC-RAM will steadily improve. One could state that we have here a competition on the "run" that will be decided based on the following four criteria: (i) capacity, (i) cell size, which translates to areal density or form factor, (iii) access time, and (iv) price, where each technology is looking for its own business niche.

Nanotechnology offers two unique opportunities to solve the problem of how to miniaturize storage devices. The first opportunity entails the use of a variety of chemical and physical processes to fabricate memory media capable of supporting nanoscale features that have reversible properties, be it magnetic, electric, or phase change. Such media should satisfy the first and second criteria, namely, a large data storage capacity and a small form factor. The second opportunity derives from the ability of scanning probe microscopy, the champion of nanotechnology, to characterize and modify structures down to atomic dimensions. Here, the use of a large number of probes operating in parallel should satisfy the third criterion, namely, fast access times. Also, because the fabrication of multiple cantilevers lends itself rather readily to mass production using currently available photolithographic processes, it should satisfy the fourth criterion: low cost.

In the next section, we describe how probe storage can be realized as a data storage technology that may find a commercially viable niche. It will be useful to divide this section into three parts, each one describing different aspects of what is known as probe storage. The first part, titled "Motors", will deal with the question of how one can access a large number of nanoscale memory cells using parallel and sequential processes. The second part, denoted by "Heads", will describe methods for reading and writing each individual memory cell using different SPM techniques. The third part, called "Media and Experimental Results", will describe the properties of different media applicable to data storage that have or can be used for probe storage.

30.2.1 Motors

This section considers two possible geometric configurations of data storage devices that operate in parallel. One consists of a rotating cylinder and an array of cantilevers positioned in an axial direction, as shown in Fig. 30.12, and the other is a 2-D, all-MEMS configuration, as shown in Fig. 30.13 [30.17]. Very little work has been done in the first, "mixed" configuration, and we will, therefore, present only a short survey of this work. Most of the section will be dedicated to the all-MEMS configuration, on which massive work has been proposed, tested, and published.

Cylinder

A cylindrical ROM media for optical data storage is currently being explored as an alternative to DVD devices, using an adhesive tape film as the recordable medium [30.18]. The purpose of the study is to find out whether this geometry will result in a smaller form factor

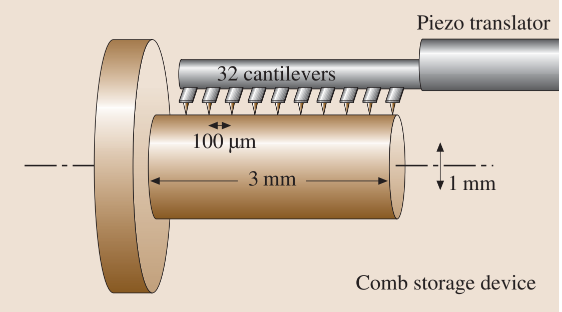

Fig. 30.12 A schematic diagram of a comb storage device

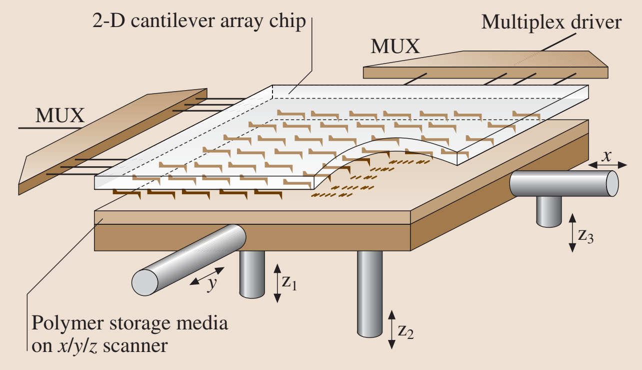

Fig. 30.13 A schematic diagram of a MEMS storage device

and reduced vibrations relative to a commercial DVD. These authors find that this geometry offers a larger writing area than that of a disk. For example, consider a disk with inner and outer radii of r and R, respectively, and a cylinder with the same radius and a height of h. For equal areas, the height of the cylinder is given by $h = \left(R^2 - r^2\right)/2R$. Therefore, for $r = 2.5$ mm and $R = 15$ mm, one obtains $h = 7.3$ mm. As a result, a cylinder with $R = h = 15$ mm offers more than twice the area of a disk. Such a device presents an interesting opportunity that the authors are currently pursuing. Figure 30.12 is an artist's concept of a possible geometry where a linear array of 32 cantilevers with their tips is mounted along the axis of a rotating cylinder. This array, which resembles a comb, is mounted on a piezoelectric actuator that can move the minute array very fast across a distance that equals the gap between adjacent cantilever tips. Let the cylinder circumference be divided into 32 sectors, each containing data bits arranged as circular tracks within each sector. As the cylinder rotates, each of the 32 tips reads independently a single track. Using the piezoelectric actuator, one can now move all the tips such that they read the next 32 tracks. Consider the particular geometry presented in Fig. 30.12, where the diameter of the cylinder is 1 mm, its length 3 mm, the distance between tips 100 μm, and a cylinder rotation rate of 10,020 rpm.

A crude assessment of the possible performance of such a device indicates promising values for the capacity, latency, access time, and transfer rate [30.19].

Note that while the rotational rates of these devices are expected to be similar to those of a HDD, one can conceive of MEMS-derived motors increasing the rotation rate. For example, a micro-engine with a diameter of 64 μm driven electrostatically at 300,000 rpm has recently been investigated [30.20]. The purpose here was to study the out-of-plane displacement and tilts about the rotating hub. Optical methods have been developed to characterize the operation of this device, together with a model that takes into account the various parameters relating to its operation. One wonders whether a combination of this MEMS technology with a disk or cylindrical recordable medium is a viable option for reducing latency and decreasing access time for data storage applications.

Probe Storage

The next section provides a detailed description of a probe storage device, so here we will only outline briefly the main principles of its operation [30.8, 9, 21–26]. Consider a typical schematic of a MEMS-based storage presented in Fig. 30.13, [30.17] where the cantilevers are laid out on a "sled" in both the x and y directions. The sled is translated, typically by electrostatic actuators, over the medium surface. This means that, unlike in conventional hard disks, the sectors are laid out adjacent to each other.

There are two possible architectures to this arrangement, one in which the media is moved relative to the tips and one in which the media is fixed and the tips are moved. The moving-tips model is very space inefficient, allowing only about 1% of the media area to be used for data storage, and, therefore, making it unviable in comparison to other storage technologies. In this configuration, the media is suspended by springs over the tips with actuators that control their positioning. In contrast, however, a typical design for a moving media model could have 10,000 tips in a 1 cm^2 area with a bit size of 50 nm^2. This geometry allows 30–50% of the area of the media to be utilized. Allowing for overheads such as error correction, this gives a storage capacity of 4 Gb/cm^2.

The disadvantage of this scheme is the greater mass that needs to be moved, resulting in large seek times. There is a settling time associated with each movement of the system, as the springs cause the sled to oscillate around its final position before coming to rest.

As shown in Fig. 30.14, the media is divided into $M \times N$ rectangular regions, each of which is accessible only by a single tip. The major advantage of this is that the data can be "striped" across multiple regions, dramatically increasing read times, as one seek operation is sufficient to read many bits.

The notation to address each bit is of the form $\langle x, y, tip\rangle$, where x and y are the horizontal and vertical coordinates of the bit within a rectangular region accessible by *tip*. A tip track is the data for which $\langle x, tip\rangle$ is the same, i. e., a vertical column within a rectangular region. By analogy with hard disk architecture, all data with identical x form a cylinder, accessible without moving the sled in the x direction. Cylinders are divided into tracks, which are accessible by the currently active tips, as not all tips can be active simultaneously for power considerations.

As mentioned, a benefit of this architecture is that data tracks are striped across tip tracks. This is depicted in Fig. 30.14 [30.23], where each data sector is mapped across several tip tracks. Sectors 1 and 2 can be read in parallel by activating tips 1–4 and moving them vertically downwards.

Simulated device performance for a prototype sled with parameters given in Table 30.3 is shown in Ta-

Table 30.3 A table of probe storage operating parameters

Sled mobility in x and y	100 µm
Bit cell width (area)	50 nm (0.0025 μm^2)
Number of tips	6,400
Simultaneously active tips	1,280
Tip sector length	80 bits (8 bytes)
Servo overhead	10 bits per tip sector
Device capacity (per sled)	2.1 GByte
Sled acceleration	114.8 m/s^2
Per-tip data rate	400 kbit/s
Settling time constants	1
Sled resonant frequency	220 Hz
Spring factor	75%

ble 30.4. Note that an important breakthrough in MEMS technology has recently been reported in which over one million cantilevers per sq. cm have been fabricated by anisotropic etching of silicon [30.27]. The cantilevers, which had tips at their end, measured a few micrometers in length, had spring constants of a few N/m, resonance frequencies around 10 MHz, and Q factors of 5 in air. Significantly, the resonance frequencies within the same row deviated only by 0.01%. Cantilevers as short as 100 nm and as thin as 20 nm were also fabricated, demonstrating the latitude that this technology offers. Also reported was the fabrication of a large number of cantilevers operating at frequencies up to 1 GHz with Q factors up to 8,000 [30.28]. These results will no doubt play a pivotal role in MEMS-related data storage technology for which mass parallelism combined with superfast sensors offer an important advantage.

Table 30.4 Probe storage projected performance

Average service time	1.49 ms (0.25)
Maximum service time	4.51 ms
Average seek time	1.27 ms (0.25)
Maximum seek time	1.66 ms
Average x seek time	1.24 ms (0.21)
Maximum x seek time	1.66 ms
Average y seek time	0.90 ms (0.31)
Maximum y seek time	1.62 ms
Settling time	0.72 ms
Average per request turnaround time	0.20 ms (0.20)
Maximum per request turnaround time	1.34 ms

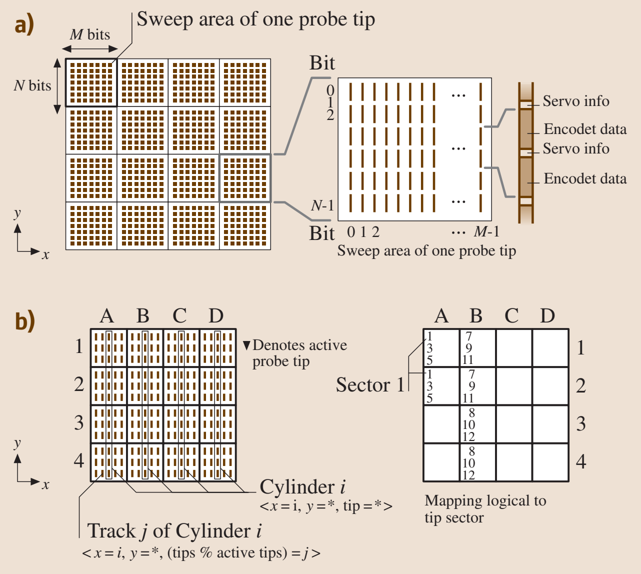

Fig. 30.14 A schematic diagram of a probe storage data organization

Table 30.4 [30.22] takes into account the physical performance of the device, including the acceleration of the actuators, settling time, etc. One observes that the results of the model compare favorably to current hard disk technology, having access times that are approximately six times faster. Note that recent results by the IBM group imply a potential bit areal densities of 400 Gbpsi [30.29].

30.2.2 Sensors

As stated before, the key problem associated with densely packed nanoscale data bits is how to address them, realizing that no interconnects of this small size exist. Clearly, the only known method is to utilize MEMS technology where sharp tips are used to read, write, and rewrite the data bits. Although such a technology is attractive, one has to realize that the number of physical phenomena available for effecting such an interaction is quite limited.

There are, in principle, three modes for interfacing, or sensing nanoscale data bits. In the first mode, denoted as contact mode, a sharp tip is constantly in contact with the medium on which the bits are written. Here, the apex of the tip has to be comparable in size to that of the bit. The second mode involves such a tip in close proximity to the bit without actually touching it. The third mode, dubbed the noncontact mode, is a combination of the other two modes. Here, the tip vibrates in close proximity to the media in such a way that on each cycle it touches or taps its surface for a short period. This mode is referred to as the tapping mode [30.30], which turns out to have several advantages relative to the other two modes. In either mode one can use an electric current within the tip to heat up a data bit, inject a current into it, or impose an electric field across it.

For the noncontact mode, one can use field emission from a sharp tip that is positioned several nanometers away from a sample, or use a tunneling current when the tip is within a fraction of a nanometer away from the sample. The current can then be used to heat a data bit and probe its conductivity as a means for data storage and retrieval. Yet both of these are accompanied with electronic noise when operating in ambient, thus requiring vacuum for a more steady operation. Note that a tunneling current is so sensitive to a tip-sample gap that a change of 0.1 nm in gap size will change the tunneling current by one order of magnitude. One can, therefore, rule out such a technique, as its restrictions on media flatness are too strenuous. In spite of some inherent difficulties, field emission maintained its position as a viable technology for data storage applications. A schematic diagram of the operation of a device based on this technology is shown in Fig. 30.15 [30.31].

Here, a multiplicity of stationary tips is positioned in close proximity to a conducting medium, such as a chalcogenide, and a controlled current emission for each tip is used to transform its phase reversibly from crystalline to amorphous. Either the tips or the media are raster scanned across each other such that, as with the probe storage device, one can address a large number of data bits. Both tips and media are placed under a mild vacuum that has to be maintained at all times.

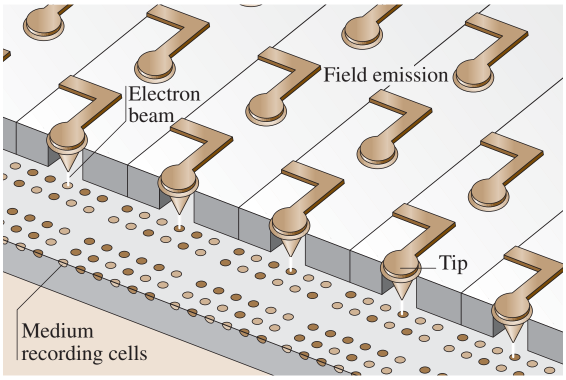

Fig. 30.15 A schematic diagram of field-emission tips

Although writing bits may be utilized efficiently, reading them back, namely, measuring their conductivity is a nontrivial task that requires further investigation. Along these lines, one may employ a MEMS-based approach and use conducting tips that are in direct contact with a conducting medium.

A different approach to locally heating a sample that does not require direct tip-sample contact, vacuum conditions, or conducting media, is to have a heating element embedded inside each tip. In spite of the fact that each tip has a complex structure, as shown in Fig. 30.16b, it was found possible to fabricate a 64 × 64 2-D array that raster scans a polymer medium. Heating a data bit and pressing a sharp tip into it generates a nanoscale pit that is estimated to persist for years. Sensing the pit for readout is accomplished by monitoring the heat

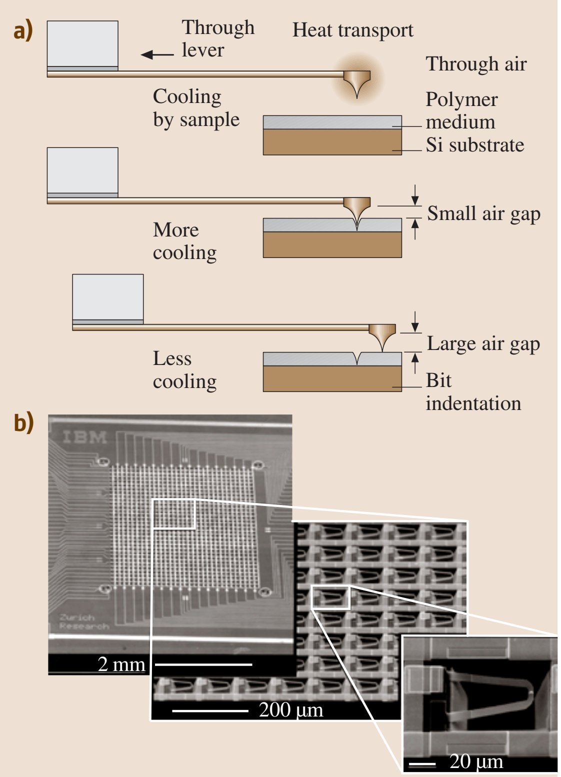

Fig. 30.16 (a) A schematic diagram of a cantilever with a heating and heat-sensing tip. Indentations are inscribed into a polymeric medium by means of a heated tip. The bits are read by sensing the increased heat loss in an indentation due to the greater surface area in contact. **(b)** Details of the tip structure

dissipation of the heated tip as a function of the media topography. When the tip is across a smooth surface, it is exposed mostly to air. When the tip is sunk into a pit, it is surrounded partially by the polymer material. Thus, the difference in heat conductivity between these two cases serves as a measure of the state of the polymer, whether it is smooth or dimpled. A key difficulty with this technology has been the erase mechanism, which is currently being addressed [30.29].

In the noncontact mode, the tip is vibrated at a distance such that the tip never touches the surface directly. However, the tip senses the surface of the sample indirectly through force gradients that modify the resonance frequency of the cantilever. This mode can be useful in magnetic recording if one attaches a small magnet to the tip and uses it to both modify and sense magnetic domains. This mode, however, is slow, as the tip has to vibrate many times during the sensing of a single data bit and, therefore, can be ruled out as a candidate for data storage.

We shall now consider the tapping mode that has three main advantages to the contact mode. First, the tip-sample contact time is reduced by almost two orders of magnitude, thus minimizing tip-media wear. The second advantage is that it touches the media in a much gentler way and, therefore, does not twist or bend. The third advantage is that the tip jumps between points of contacts, which constitute the data bits, so it is less prone to crash due to media roughness. If one uses the tapping mode with a conducting tip and taps on a conducting substrate, it is possible to follow the motion of the vibrating cantilever by monitoring the time the tip makes contact with the sample. This method makes it easy to control a feedback system that keeps the motion of the cantilever under control, provided the contact is good enough. The problem one encounters using this mode, however, is that such an operation is usually chaotic and, therefore, difficult to control. If one were to overcome this problem, this mode would be fast enough relative to the noncontact mode to be considered a viable method for data storage.

We will now consider the theory and experimental results of the tapping mode by first treating it as a grazing impact oscillator (GIO) for which one can obtain closed-form expressions connecting all of its parameters [30.32]. The GIO consists of a cantilever driven at a constant frequency and an amplitude by an external force. The vibrating cantilever is then brought into close proximity to the surface of a sample until its tip starts impacting the surface. At this point the amplitude of vibration of the GIO changes, together with the relative phase between the driver and the tip. Let the amplitude of vibration of the driver be denoted by a and the spring constant and quality factor of the cantilever be denoted by k_0 and Q, respectively. On impact, the GIO acquires amplitudes of vibration A and effective spring constant k_{eff}, establishing a new value for the input power and phase:

$$A = \frac{aQ}{\sqrt{Q^2\left(1-\frac{k}{k_0}\right)^2+\frac{k}{k_0}}}\,, \tag{30.1a}$$

$$\theta = \tan^{-1}\left(\frac{\sqrt{k/k_0}}{Q(1-k/k_0)}\right)\,, \tag{30.1b}$$

$$k = k_0\left[\sqrt{\left(\frac{a}{AA}\right)^2+\frac{1}{4}\left(\frac{1}{Q^4}-\frac{4}{Q^2}\right)} - \frac{1}{2Q^2}+1\right]\,. \tag{30.1c}$$

Equation (30.1) shows the relationship among the various parameters governing the operation of the GIO in a closed form, making it possible to model the effect that impacting has on the tapping cantilever. Figure 30.17a shows that on bringing the vibrating cantilever closer to the surface, its amplitude of vibration decreases and its effective spring constant increases. Figure 30.17b shows the phase of the vibrating cantilever as a function of the amplitude of vibration. Figure 30.17c and d depict the input power as a function of the amplitude of vibration of the cantilever and of its phase, respectively. The double-

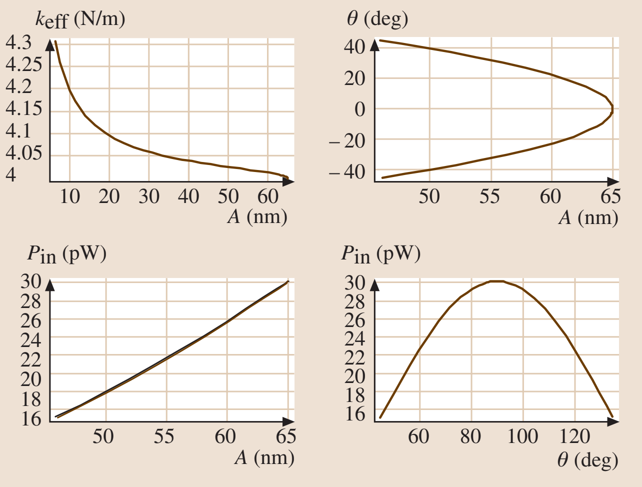

Fig. 30.17 Calculated tapping behavior. Modeling of the relationships among the amplitude A, effective spring constant k_{eff}, input power P_{in}, and phase ϕ

valued functions express the fact that one can operate either below or above the original resonance frequency of the cantilever.

This simplified model of a tapping mode neglects many effects such as indentation of the surface on impact and adhesion and other interface tip-sample forces. For our purpose, however, we will neglect these effects and explore the feasibility of using the tapping mode for data storage applications. The idea here is to have the tip jump from one pixel, or bit of stored data on the surface of a sample to another pixel without touching the surface in between. And while briefly touching the surface, the tip, which is conducting, will inject a pulse of current into the pixel and modify its conductivity [30.33]. Upon returning to the same pixel, the tip could monitor the conductivity of that particular pixel.

Considering the fact that the tapping mode involves a driven, damped oscillator, it is apparent that the motion of the cantilever traces a chaotic trajectory, necessitating an averaging process. For most imaging or nanofabrication purposes, such an averaging, combined with an appropriate image processing procedure, can indeed yield images with nanometer resolution. For data storage applications, however, it is desirable to have non-chaotic, reproducible interactions for every single tap because one could not afford an averaging procedure.

There are two key issues regarding the requirement that each tap produce a controlled interaction. The first concerns the identification of conditions under which (i) chaotic motion of the cantilever is minimized and (ii) good tip-sample contact is established. The second issue concerns the means by which one can observe these individual interactions. The difficulty in measuring tip-sample interactions during an individual tap stems from the fact that most of the time the tip is far away from the sample, and only during a small fraction of a cycle does it actually sense the presence of the sample. Also, it is the nature of the tapping mode of operation that the tip approaches the sample at a relatively small velocity, which is the reason why it acts as a grazing impact oscillator in the first place. One can choose the electrical contact between tip and sample as the measuring tool for the reproducibility of the interaction for individual taps. This choice is clearly compatible with storage modalities in which current injection into storage media gives rise to writing and reading of phase-change data.

The use of all-metal cantilevers tapping on a gold substrate was recently demonstrated, whereby carefully chosen parameters enable the generation and observation of individual current pulses [30.34]. These pulses were generated by applying a short bias pulse during the tip-sample contact time. Each current pulse, lasting up to 20 µs, could produce a current of ∼ 10 µA in a nm^2 region yielding a current density of 10^{13} A/m^2. This density is large enough to be applicable for phase-change. The cantilever was fabricated from a rectangular PtIr wire whose length, width, and thickness were 1,200 µm, 100 µm, and 50 µm, respectively. One side of the cantilever was attached to a driving bimorph, and the other side etched to produce a tip whose apex radius was estimated to be 70 nm. The spring constant, resonance frequency, amplitude of vibration, and quality factor of the cantilever were 300 N/m, 13 kHz, 100 nm, and 100, respectively. The all-metal cantilever ensured good conductivity and could withstand a large number of

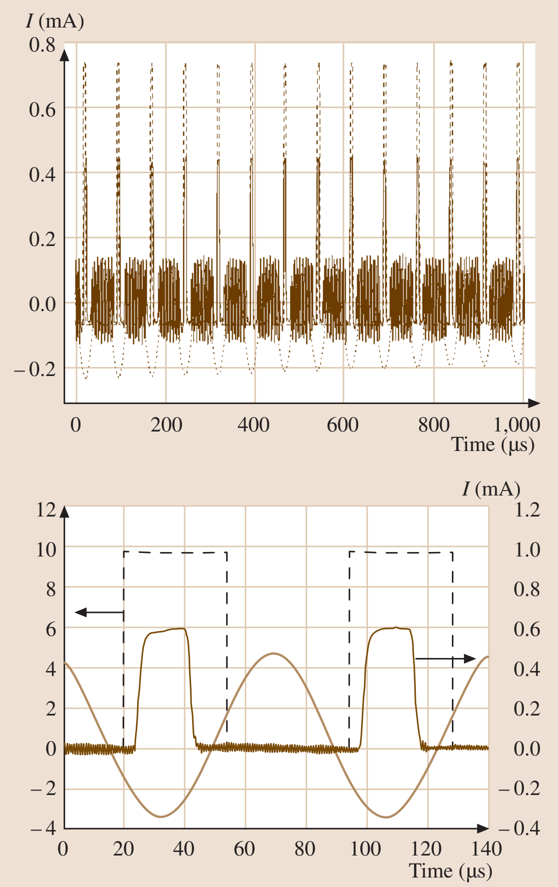

Fig. 30.18 **(a)** A typical current pulse train (*solid line*) obtained with pulsed voltage bias in phase with the cantilever's lowest position and the associated cantilever motion (*dotted line*), **(b)** two current pulses (*solid line*), bias voltage pulses (*dashed line*), and cantilever motion (*dotted line*)

tapping impacts without degrading and had a shiny surface that acted as a good mirror for the optical deflection technique.

The experimental results yielded long trains of individually injected current pulses. Figure 30.18 (left-hand side) shows 14 non-averaged experimental current pulses (solid line) and their associated cantilever motion (dashed line) with a background of instrumental noise [30.34]. The current pulses demonstrate the reproducibility of the conductance established during each tap. Figure 30.18 [30.34] (right-hand side) shows two averaged current pulses (solid line) with their input bias pulses (dashed line) and the associated cantilever motion (dotted line). Note that the tip-sample contact time is revealed by the duration of the current pulse, which is shorter than the duration of the applied bias pulse. Here, the amplitude of the bias pulse was 10 mV, its duration was 40 μs, and the timing of its application relative to the motion of the cantilever was provided by the output of the atomic force microscope photodiode. The current pulses, with a magnitude of 0.5 mA, are much larger than the tunneling current occurring while the tip is within a fraction of a nanometer from the surface. Also, the displacement current, due to the motion of the cantilever and the time varying nature of the applied bias, is expected to be much smaller than the observed current pulses. The contact conductance is shown in Fig. 30.19, where a linear relation between voltage and current implies a good ohmic contact. These results demonstrate that it is possible to minimize the chaotic behavior of a tapping cantilever and obtain reproducible conductance on each individual tap. The next step is to use the tapping mode in conjunction with current injection to write and read data bits on conducting media.

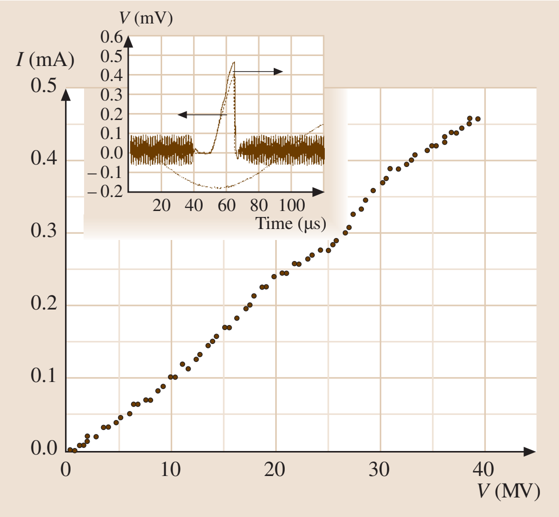

Fig. 30.19 An I–V curve taken with an individual ramped input voltage pulse demonstrating an ohmic contact. *Insert:* The current pulse (*solid line*), ramped input bias voltage (*dashed line*), and cantilever motion (*dotted line*)

30.2.3 Media and Experimental Results

This section summarizes several typical experimental results using different data storage media, electric charging, electric domain orientation, magnetic domain orientation, chemical modification, and phase change in chalcogenides and organic films.

Electric Charging

Quate's group at Stanford has studied charge trapping in a silicon nitride/silicon dioxide/silicon system using a scanning capacitance microscope [30.6]. The charge trapping in this system has been used for many years for non-volatilesemiconductor memory. Using chemical vapor deposition, 500 Å of silicon nitride was deposited onto a silicon wafer with a 50 Å oxide coating. By selectively storing charge into the substrate, they were able to write bits as small as 0.02 μm^2. Using this method, they were able to write the founding document of the Swiss Confederation onto a 120 μm^2 area, corresponding to 256 kilobits of information that yields a data density of 10 Gbpsi.

Electric Domains

Hidaka et al. have demonstrated the possibility of ultrahigh density charge storage in a ferroelectric film, $PbZr_{1-x}Ti_xO_3$ [30.35]. They show that a change in the topography of the film can be introduced by domain reorientation induced by the application of a 10 V pulse with a duration of 100 μs. This is due to the piezoelectric effect caused by domain structures. Data bits as small as 40 nm were written this way. This result implies a data density of approximately 300 Gbpsi, suggesting that this medium is a good candidate for high-density data storage.

Thin Oxide Films

IBM Research in Switzerland has demonstrated reproducible switching in thin oxide perovskite films [30.36]. By constructing metal-insulator-metal capacitor-like devices, with $SrZr0_3$ as the insulator, they were able to observe an order of magnitude increase in the leakage current of the device at a threshold voltage of −0.5 V. This process is perfectly reversible, and the leakage current can be restored to its initial state upon application of a positive voltage. The switching occurs very rapidly, taking less than 100 ns to change from one conductivity state to another. Interestingly, the magnitude of the transition depends on the voltage applied, suggesting the possibility of storing multiple bits in one pixel.

Magnetic Domains

Sun et al. have demonstrated a novel way to overcome the limitations of media currently used for magnetic data storage [30.37]. By reducing iron and platinum-containing compounds, they demonstrated the formation of large-scale self-ordered arrays of magnetic particles. Initial studies suggest that an assembly of magnetic particles as small as 4 nm can support stable magnetization transitions at room temperature. By using the particles in these arrays as bits, the possibility of terabit-per-square-inch magnetic data storage could be realized.

Figure 30.20 shows transmission electron microscopy of these arrays that indicate a remarkable uniformity of particle size and large-scale periodicity. The figure also shows assemblies of these particles deposited onto a silicon oxide substrate under varying experimental conditions.

Chemical Modification

We have discussed the fact that the tapping mode is considerably less damaging to the tip and sample, as the tip spends less time contacting the sample and does not drag across it while raster scanning. For this purpose, diamond-coated cantilevers have been found useful for experiments in which wear is a problem. For sample oxidation purposes, the tip is first oscillated away from the sample with peak-to-peak amplitude of 10–15 nm, and then a feedback system reduces the amplitude to 7–10 nm by pressing the tip further into the sample. A 50 ms-long pulse with a 10–12 V magnitude is then applied to oxidize a thin Ti film. It was found that this technique produces much finer structures. Figure 30.21 [30.38] is a tapping mode AFM image of four dots created in such a way. The dots in this image have a diameter of less than 10 nm and a height of 3.5 nm above the surrounding Ti surface with a center separation of 20 nm. Around the dots there is what appears to be a lower ring of material with a channel in between. This experiment demonstrates that one can use the tapping mode to write fine pixels, albeit not to erase them.

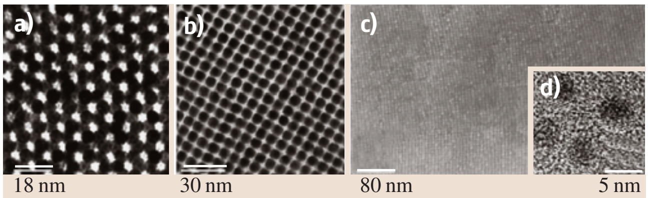

Fig. 30.20 **(a)** TEM micrograph of a 3-D assembly of 6-nm as-synthesized $Fe_{50}Pt_{50}$ particles deposited from a hexane/octane (v/v 1/1) dispersion onto a SiO-coated copper grid. **(b)** TEM micrograph of a 3-D assembly of a 6-nm $Fe_{50}Pt_{50}$ sample after replacing oleic acid/oleyl amine with hexanoic acid/hexylamine. **(c)** HRSEM image of a ∼ 180-nm-thick, 4-nm $F_{52}Pt_{48}$ nanocrystal assembly annealed at 560 °C for 30 min under 1 atm of N_2 gas. **(d)** High-resolution TEM image of 4-nm $Fe_{52}Pt_{48}$ nanocrystals annealed at 560 °C for 30 min on a SiO-coated copper grid reference

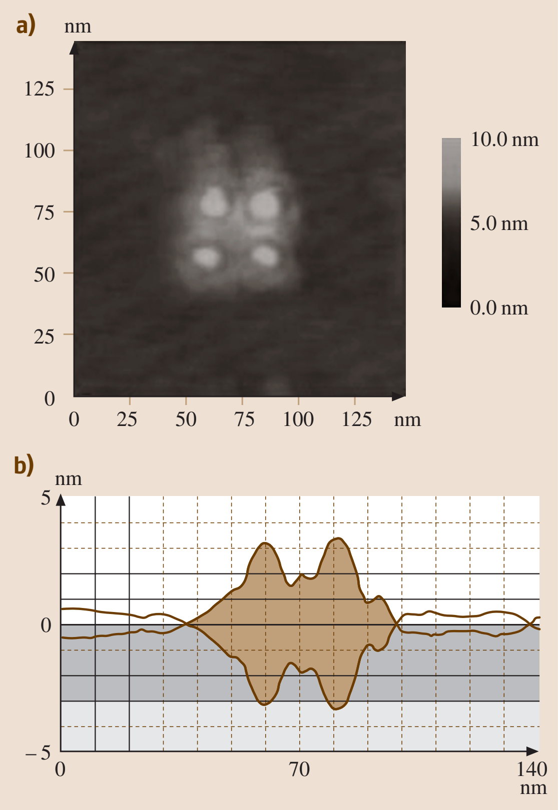

Fig. 30.21a,b Experimental results of local chemical modification using AFM tip: **(a)** 20-nm dots written on Ti film, **(b)** cross section of written bit

Phase Change in Chalcogenides

Chalcogenides, the active material in rewritable CDs, have the advantage of being a well-characterized and understood bistable phase change material that has already had successful commercial data storage applications, i. e., CDRWs. Currently, NVRAM using chalcogenides as the active material is being pursued by an alliance of Ovonyx and Intel. Chalcogenides are materials that have an easily reversible stable change between amorphous and crystalline phases. To switch from the crystalline to the amorphous phase, the material is heated sufficiently, so that the material subsequently cools rapidly into the amorphous phase. The reverse change is achieved by heating the material above the glass transition temperature, so that it recrystallizes upon cooling. For commercial CDRWs, writing is done by means of laser light heating and reading is done by measuring the difference in reflectivity of the laser light from each of these two phases.

Scanning probe microscopy can be used to switch the phase of a chalcogenide material by the application of a pulse of current that heats the material as it passes through it [30.39]. The state of the material can be read typically by measuring the resistance of the sample. Figure 30.22 shows the two different states, amorphous and crystalline of the chalcogenides. The reversibility of this transition is remarkable, and has been shown to be stable over 10^{12} repeated read-write cycles. *Gotoh* et al. demonstrated this effect using conducting AFM tips, where a transition from the high resistance amorphous state to the lower resistance crystalline state was observed both the in topography and the conductivity of the sample. The change from amorphous to crystalline phases was confirmed by X-ray diffraction. Figure 30.23a and a′ show a conductance AFM image and a profile of the surface after the application of varying voltage pulses. The topography of the substrate is shown in Fig. 30.23b. Figure 30.23c and c′ show the equivalent conductance AFM images after the substrate has been scanned at a negative voltage, erasing the lower power marks on the substrate.

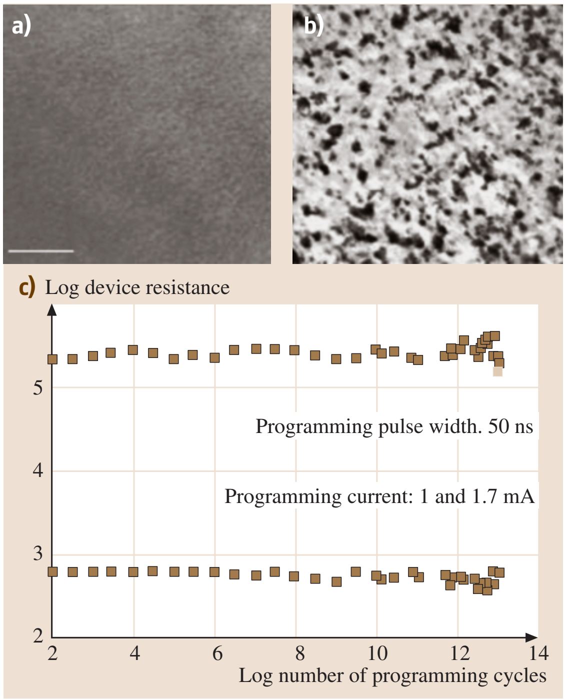

Fig. 30.22a–c Experimental results of phase change using interconnects. **(a)** Crystalline area, **(b)** amorphous area, **(c)** stability of bits over 1,012 rewrites (Ovonyx)

These materials have been further characterized by *Gidon* et al. [30.26], who showed that the resistivity of the film in its amorphous phase was 100 Ω cm^{-1}, while that of the crystalline state was 0.1 Ω cm^{-1}. They also show that a data bit requires a 200 ns pulse with a power of 50 µW, which translates to a total energy of 10 pJ.

Phase Change in Organics

Organics have proved important in a variety of electronic and optoelectronic applications such as light-emitting diodes, thin-film displays, transistors, photovoltaics, and

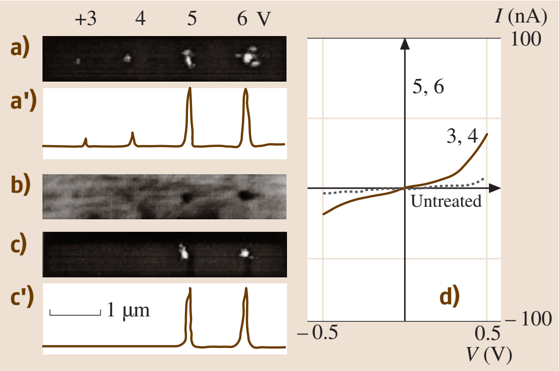

Fig. 30.23 Experimental results of phase change using an AFM tip **(a)**, **(a′)**, conducting AFM images after applying 100 ns pulses to the substrate, **(b)** topography AFM image of the substrate, **(c)**, **(c′)** conducting AFM images showing erasure after scanning the substrate with an applied bias of −1 V, **(d)** *I*–*V* curves of untreated (*dotted*) and treated (*solid*) regions

lasers. There are several advantages to organics for economic and manufacturing reasons, and a large body of research already exists in this field. As such, there is significant interest and motivation in finding data storage applications based on organic technology.

There has been much research recently into organic materials and systems capable of reversible switching and thus capable of being exploited for data storage applications. There are two broad categories of organics data storage; one where the active material (which can be homogenous or heterogeneous) has inherent bistability, and one where the bistability is a property of the device structure.

There are many different materials that have been investigated for data storage. The fundamental property is that it should have two states (electronic, morphological, or some other easily measured physical characteristic) that the material can exist in, and it should be stable indefinitely in either state. Furthermore, the material should be able to be easily and reversibly switched between these states, and this should be reproducible over many millions of cycles for it to be useful as a data storage medium.

Several groups have been working on reversible transitions in organic complexes suitable for switching by scanning probe techniques. Many complexes have been demonstrated to exhibit reversible switching, e.g., 3-nitrobenzal malonitrile (NBMN) and 1,4-phenylenediamine (pDA) [30.40], tetrathiofulvane (TTF) and m-nitrobenzylidene propanedinitrile (m-NBP) [30.41], and many other variants.

Li et al. [30.41] have studied organic films consisting of tetrathiofulvane (TTF), an electron donor, and m-nitrobenzylidene propanedinitrile (m-NBP), an electron acceptor. This complex demonstrates electrical bistability, which can be written and read by scanning tunneling microscopy.

The nature of the measurement means that it is difficult to ascertain whether this is due to solely conductive changes in the sample, or whether topographic modifications might also contribute. By applying pulses of 1 ms duration, bits of 1.2 nm diameter could be written. This implies an areal density of $\sim 10^{18}$ bits/in^2. This change was characterized electrically, and the resistance was measured as changing from $10^8\ \Omega$ to $10^3\ \Omega$ with a threshold switching voltage of 3.1 V. The proposed switching mechanism is charge transfer between the two film constituents. This change was stable over time and is reversible. This change is reversible, but only by heating for prolonged periods, making it unsuitable for applications as of yet.

Gao et al. have demonstrated stable, local, reversible nanoscale switching [30.40] using a charge transfer complex of 3-nitrobenzal malonitrile (NBMN) and 1,4-phenylenediamine (pDA). A 200-nm-thick film of this complex was deposited onto highly ordered pyrolitic graphite. Macroscopic electrical measurements demonstrate the bistability and switching of the material. Figure 30.24a shows the current-voltage characteristics of the sample in these two conductivity states. Initially, the material is in a low conductivity state, but upon reaching a certain threshold voltage, in this case 3.2 V, the material abruptly changes to a conductive state. The transition time is very abrupt, 80 ns, as shown in Fig. 30.24b.

Of further interest for data storage applications is that, having demonstrated reversible switching in macroscopic films, Gao et al. also showed that it was possible to switch these materials on the nanoscale using an STM

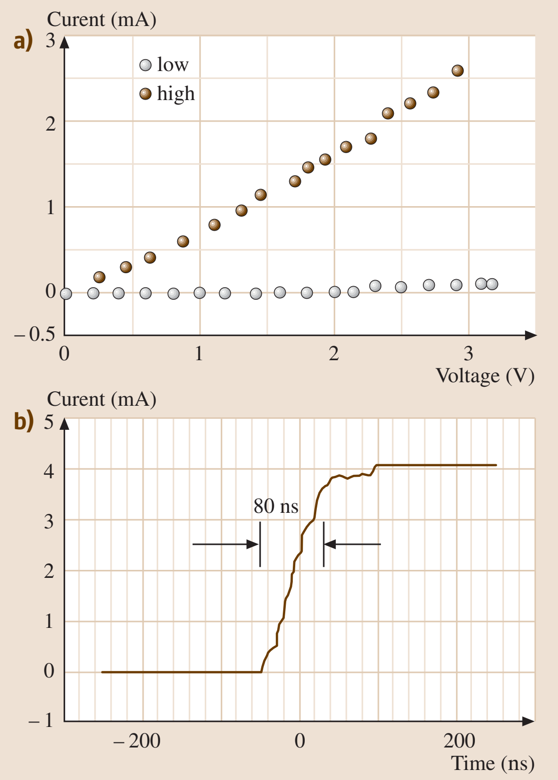

Fig. 30.24a,b Experimental results of bulk conductivity changes of organic films. **(a)** I–V curves of a 200-nm-thick film showing the two conductivity states. Switching voltage is 3.2 V. **(b)** Change in conductivity against time, showing a transition time of 80 ns

to apply pulses, i. e., to write bits, and also to measure conductivity, i. e., to read bits.

Figure 30.25a shows the local nanoscale film morphology of the sample. Figure 30.25b and c show patterns written on a sample by application of voltage pulses from the STM tip. Figure 30.25d demonstrates the erasure of a single bit from the pattern by application of a reverse-biased voltage pulse. Similar to the work of *Li* et al., the entire substrate could be erased by prolonged heating. Of much greater interest is the demonstration of individual bit erasure, achieved by applying reverse voltage pulses. The erasure is incomplete unless the pulse is reverse-biased, indicating that it is the combined effect of heating and electric field that erases the bits.

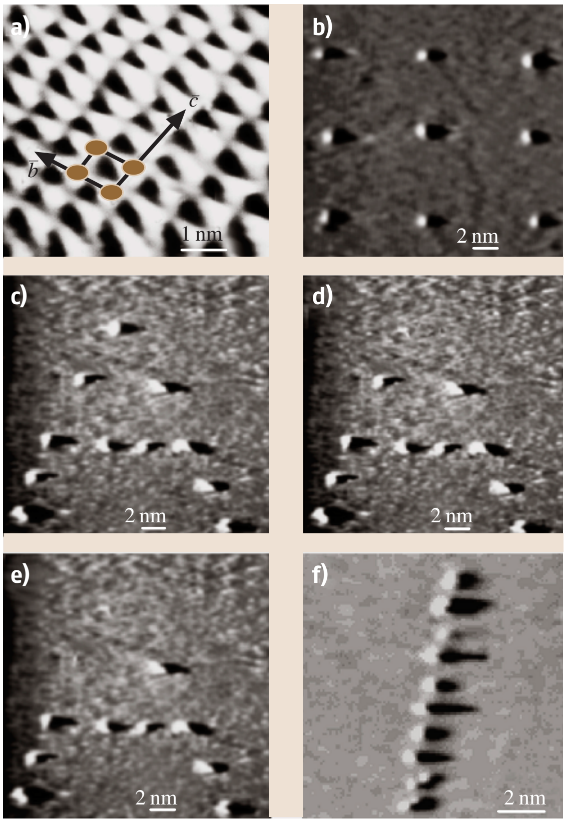

Fig. 30.25a–f STM of the NBMN-pDA film on HOPG. **(a)** 6 nm^2 image showing the crystalline order of the film; **(b)** a 3×3 array of bits (4 V, 1 µs); **(c)** an "A" pattern (3.5 V, 2 µs); **(d)**, **(e)**, **(f)** STM images after erasing marks with reverse-biased pulses (−4.5 V, 50 µs). Distance between neighboring marks is 1.7 nm

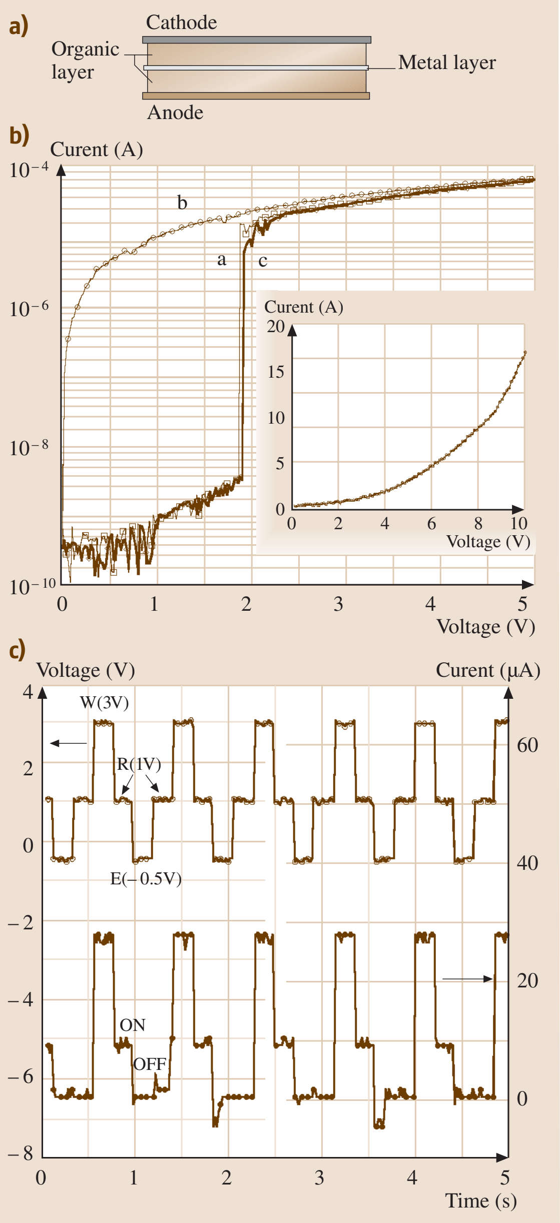

Fig. 30.26a–c Experimental results of conductivity changes of organic device structures: **(a)** the structure of the bistable device, **(b)** I–V characteristics. Voltage was scanned in steps of 0.1 V from 0 to 5 V. Curves a and b represent the I–V characteristics of the first and second bias scan, respectively. Curve c is the I–V curve of the third bias scan after the application of a reverse voltage pulse (−3 V). **(c)** Typical current responses of the device being written, read, erased, read over time

Several proposed mechanisms were investigated as the possible origin of this switching, including conformational switching and charge transfer. These could be discounted for reasons, including the stability and time scale of the switching. A more plausible hypothesis is a local phase transition from crystalline insulating regions to amorphous conducting regions. TEM and diffraction studies of the samples agree with this hypothesis. It should be pointed out that an alternative interpretation has been suggested by *Zhao* et al. [30.42]. The conductivity change could be due mainly to a change in the topography of the sample, reducing the tip-sample distance and, therefore, the tunneling distance, manifested as an apparent change in conductivity.

Ma et al. have developed a simple device structure that exhibits electrical bistability [30.43]. This device has a fairly simple structure consisting of alternating layers of aluminum and an organic compound, 2-amino-4,5-imidazoledicarbonitrile (AIDCN), as depicted in Fig. 30.26a. Several alternative organics (Alq3 and MEH-PPV) and metals (copper, silver, and gold) have been tried with varying degrees of success.

Figure 30.26b shows the current-voltage bistability of a device with structure Al/AIDCN (50 nm)/Al (20 nm)/AIDCN (50 nm)/Al. A first run shows an initially low current, followed by an abrupt and dramatic increase in the current of four orders of magnitude at 2 V. Subsequent repeated voltage runs (b) show that this transition is stable, being orders of magnitude higher than the initial measurement. The devices proved to be stable in this state over periods ranging from days to weeks.

Of further interest is that the device proved to be erasable by purely electrical means, the application of a −0.5 V pulse bringing the device back to its original state. This is denoted by c in Fig. 30.26b. Furthermore, the devices showed excellent stability over time, with over one million read-write-erase cycles being performed on the devices with excellent repeatability, as depicted in Fig. 30.26c. The origin of this bistability is unclear. It is clear that the central electrode is critical to the observed effect, as the bistability is only observed when the central electrode thickness is above a threshold value, 10 nm. Organics with high dielectric constants and low conductivity are believed to be important for device performance. The exact mechanism, however, is as yet unclear. Preliminary results on novel materials [30.44] have suggested the possibility of an all-organic variation of this concept, with an AFM tip switching between bistable states of the system.

30.3 Conclusion

Nanotechnology is perceived as a branch of science that deals with structures whose dimensions are measured in nanometers. Clearly, new phenomena become apparent when the number of atoms at the interface of such a structure becomes comparable to the number of atoms in its volume. As a result, nanotechnology evolved into a new kind of science that deals with the physics of small things. However, the suffix -technology here implies a far more reaching aspect than just a scientific curiosity; this branch of science should also be of some use. This poses a dilemma for researchers whose main interest is in understanding new phenomena, rather than in finding a use for them. Consider scanning probe microcopy, which is celebrating its twentieth birthday now with uncountable contributions to the science of measurement of surfaces and near-surface phenomena. Despite all the contributions scanning probe microscopy has made, it has always served as a tool for investigations, rather than as a practical device that can become a commercial commodity. And yet this is exactly what this chapter deals with. The challenge of converting scanning probe microscopy from a research tool into a probe storage device is daunting indeed and will require new ideas to make it happen. The danger associated with this challenge is that novel ideas developed by research groups are sometimes too immature to be measured against any existing technology, in general, and against data storage, in particular. One has to keep pursuing research opportunities whose outcomes are unknown and whose applications are unexpected with the hope that they will bring about good science and hopefully useful applications.

However, based on the technical breakthroughs delineated in this and the following chapter and judging from the history of data storage evolution, it seems plausible that probe storage will serve as a disruptive technology that replaces data storage devices in a particular market niche. These new devices should be non-volatile, have terabyte capacity with microsecond access times with gigabyte transfer rates, all for a cost that is highly competitive with current commercially available data storage devices.

References

30.1 D. Sarid: *Scanning Force Microscopy, with Applications to Electric, Magnetic, and Atomic Forces* (Oxford Univ. Press, Oxford 1991)

30.2 D. Sarid: *Scanning Force Microscopy, with Applications to Electric, Magnetic, and Atomic Forces*, revised edn. (Oxford Univ. Press, Oxford 1994)

30.3 D. Sarid: *Exploring Scanning Probe Microscopy with Mathematica* (Wiley, New York 1997)

30.4 B. G. Levi: Do atomic force microscope arrays have the write stuff?, Phys. Today **55**(10) (2002) 14–17

30.5 H. J. Mamin, B. D. Terris, L. S. Fan, S. Hoen, R. C. Barrett, D. Rugar: High-density data storage using proximal probe techniques, IBM J. Res. Dev. **39**(6) (1995) 681–699

30.6 R. C. Barrett, C. F. Quate: Large-scale charge storage by scanning capacitance microscopy, Ultramicroscopy **42–44** (1992) 262–267

30.7 H. J. Mamin, R. P. Ried, B. D. Terris, D. Rugar: High-density data storage based on the atomic force microscope, Proc. IEEE **87**(6) (1999) 1014–1027

30.8 P. Vettiger, J. Brugger, M. Despont, U. Drechsler, U. Durig, W. Haberle, H. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig: Ultrahigh density, high-data-rate NEMS-based AFM data storage system, Microelectron. Eng. **46** (1999) 11–17

30.9 P. Vettiger, M. Despont, U. Drechsler, U. Durig, W. Haberle, M. I. Lutwyche, H. E. Rothuizen, R. Stutz, R. Widmer, G. K. Binnig: The millipede-more than one thousand tips for future AFM data storage, IBM J. Res. Dev. **44**(3) (2000) 323–340

30.10 B. Schechtman: The future of optical disk storage, NSIC Annual Meeting Symposium, Monterey 2002, 1–14

30.11 P. Lyman, H. L. Varian: How much information (2000) retrieved from http://:www.sims.berkeley.edu/how-much-info

30.12 W. Perdue: Private Communication (September, 2002)

30.13 P. Frank: private communication, (2002)

30.14 H. De J. Ruiz, S. D. Morgan: , Witter Semiconductor and Systems Conference, Laguna Michael 2001

30.15 J. Daughton, NVE Inc., private communication

30.16 M. Gill, T. Lowrey, J. Park: Ovonics unified memory – A high performance nonvolatile memory technology for stand alone memory and embedded applications, 2002 IEEE Int. Solid State Circuits Conf., San Francisco 2002 (IEEE, Piscataway 2002)

30.17 M. Despont, J. Brugger, U. Drechsler, U. Dürig, W. Häberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig, H. Rohrer, P. Vettiger: VLSI–NEMS chip for AFM data storage, Technical Digest 12th IEEE Int. Conf. on Micro Electro Mechanical Systems (MEMS'99) (IEEE, Piscataway 1999) 564–569

30.18 K. Schulte-Wieking, S. Noehte, C. Dietrich, M. Mayer: A new cylindrically shaped optical data storage medium based on adhesive tape-concept and first results, IEEE Proc. Technical Digest of ISOM/ODS 2002, Joint Int. Symp. on Optical Memory and Optical Data Storage 2002 (IEEE, Piscataway 2002) 117–120

30.19 D. Sarid: unpublished results, (2002)

30.20 L. A. Romero, F. M. Dickey: A method for achieving constant rotation rates in a microorthogonal linkage system, J. Microelectromech. Syst. **9**(2) (2000) 236–244

30.21 L. R. Carley, G. R. Ganger, D. F. Nagle: MEMS-based integrated-circuit mass-storage systems, Commun. ACM **43**(11) (2000) 71–80

30.22 J. L. Griffin, S. W. Schlosser, G. R. Ganger, D. F. Nagle: Modeling and performance of MEMS-based storage devices, Int. Conf. on Measurement and Modeling of Computer Systems, Proc. of ACM Sigmetrics, Santa Clara 2000, **28**(1) (2000) 56–65, published as Performance Evaluation Review

30.23 S. W. Schlosser, J. L. Griffin, D. F. Nagle, G. R. Ganger: Designing computer systems with MEMS-based storage, Proc. of the 9th International Conf. on Architectural Support for Programming Languages and Operating Systems, Boston 2000, 1–12

30.24 L. R. Carley, G. Ganger, D. F. Guillou, D. Nagle: System design considerations for MEMS-actuated magnetic-probe-based mass storage, IEEE Trans. Magn. **37**(2) (2001) 657–662

30.25 R. T. El-Sayed, L. R. Carley: Performance analysis of beyond 100 Gb/in^2 MFM-based MEMS-actuated mass storage devices, IEEE Trans. Magn. **38**(5) (2002) 1–3

30.26 O. Bichet, S. Gidon, Y. Samson: Scanning probe based storage on phase change media, OSA Trends in Optics and Photonics Vol. 88 Optical Data Storage (OSA, Washington D.C. 2003)

30.27 H. Kawakatsu, D. Saya, A. Kato, K. Fukushima, H. Toshiyoshi, H. Fujita: Millions of cantilevers for atomic force microscopy, Rev. Sci. Instrum. **73**(3) (2002) 1188–1192

30.28 H. Kawakatsu, S. Kawai, D. Saya, M. Nagashio, D. Kobayashi, H. Toshiyoshi, H. Fujita: Towards atomic force microscopy up to 100 MHz, Rev. Sci. Instrum. **73**(6) (2002) 2317–2320

30.29 P. Vettiger, G. Cross, M. Despont, U. Drechsler, U. Dürig, B. Gotsman, W. Häberle, M. A. Lantz, H. E. Rothuizen, R. Stutz, G. K. Binnig: The 'millipede' – Nanotechnology entering data storage, IEEE Trans. Nanotechnol. **1** (2002) 39–55

30.30 Q. Zhong, D. Inniss, K. Kjoller, V. B. Elings: Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy, Surf. Sci. **290**(1–2) (1993) L688–L692

30.31 A. M. Hayashi: Punch cards of the future, Sci. Am. (May 2000)
30.32 J. P. Hunt, D. Sarid: Kinetics of lossy grazing impact oscillators, Appl. Phys. Lett. **72**(23) (1998) 2969–2971
30.33 D. Sarid: Tapping-mode scanning force microscopy: Metallic tips and samples, Comp. Mater. Sci. **5** (1996) 291–297
30.34 A. Fein, Y. Zhao, C. A. Peterson, G. E. Jabbour, D. Sarid: Individually injected current pulses with conducting-tip, tapping-mode atomic force microscopy, Appl. Phys. Lett. **79**(24) (2001) 3935–3937
30.35 T. Hidaka, T. Maruyama, M. Saitoh, N. Mikoshiba, M. Shimizu, T. Shiosaki, L. A. Willis, R. Hiskes, S. A. Dicarolis, J. Amano: Formation and observation of 50 nm polarized domains in $PbZr_{1-x}Ti_xO_3$ thin film using scanning probe microscope, Appl. Phys. Lett. **68**(17) (1996) 2358–2359
30.36 A. Beck, J. G. Bednorz, Ch. Gerber, C. Rossel, D. Widmer: Reproducible switching effect in thin oxide films for memory applications, Appl. Phys. Lett. **77**(1) (2000) 139–141
30.37 S. Sun, C. B. Murray, D. Weller, L. Folks, A. Moser: Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices, Science **287** (2000) 1989–1992
30.38 C. Peterson: Characterization of semiconductor devices through scanned probe microscopy. Ph.D. Thesis (University of Arizona, Tucson 2002)
30.39 T. Gotoh, K. Sugawara, K. Tanaka: Nanoscale electrical phase-change in $GeSb_2Te_4$ films with scanning probe microscopes, J. Non-Cryst. Solids **299–302** (2002) 968–972
30.40 H. J. Gao, K. Sohlberg, Z. Q. Xue, H. Y. Chen, S. M. Hou, L. P. Ma, X. W. Fang, S. J. Pang, S. J. Pennycook: Reversible, nanometer-scale conductance transitions in an organic complex, Phys. Rev. Lett. **84**(8) (2000) 1780–1783
30.41 J. C. Li, Z. Q. Xue, W. M. Liu, S. M. Hou, X. L. Li, X. Y. Zhao: Study on a new organic-complex thin film with electrical bistable properties using a scanning tunneling microscope, Phys. Lett. A **266** (2000) 441–445
30.42 Y. Zhao, A. Fein, C. A. Peterson, D. Sarid: Comment on reversible, nanometer-scale conductance transitions in an organic complex, Phys. Rev. Lett. **87**(17) (2001) 179706–1
30.43 L. P. Ma, J. Liu, Y. Yang: Organic electrical bistable devices and rewritable memory cells, Appl. Phys. Lett. **80**(16) (2002) 2997–2999
30.44 B. McCarthy, K. Yamnitskiy, G. E. Jabbour, D. Sarid: unpublished results, (2002)

31. The "Millipede" – A Nanotechnology-Based AFM Data-Storage System

The "millipede" concept presented here is a new approach for storing data at high speed and ultrahigh density. The interesting part is that millipede stores digital information in a completely different way from magnetic hard disks, optical disks, and transistor-based memory chips. The ultimate locality is provided by a tip, and high data rates are a result of massive parallel operation of such tips. As storage medium, polymer films are being considered, although the use of other media, in particular, magnetic materials, has not been ruled out. The current effort is focused on demonstrating the millipede concept with areal densities of up to $0.5-1\,\mathrm{Tb/in^2}$ and parallel operation of very large 2-D (up to 64×64) AFM cantilever arrays with integrated tips and write/read/erase functionality. The fabrication and integration of such a large number of mechanical devices (cantilever beams) will lead to what we envision as the VLSI age of micro- and nanomechanics.

In this chapter, the millipede concept for a MEMS-based storage device is described in detail. In particular, various aspects pertaining to AFM thermomechanical read/write/erase functions, 2-D array fabrication and characteristics, *x*/*y*/*z* microscanner design, polymer media properties, read channel modeling, servo control and synchronization, as well as modulation coding techniques suitable for probe-based data-storage devices are discussed.

31.1 **The Millipede Concept** 923

31.2 **Thermomechanical AFM Data Storage** 924

31.3 **Array Design, Technology, and Fabrication** 926

31.4 **Array Characterization** 927

31.5 ***x*/*y*/*z* Medium Microscanner** 929

31.6 **First Write/Read Results with the 32×32 Array Chip** 931

31.7 **Polymer Medium** 932
  31.7.1 Writing Mechanism 932
  31.7.2 Erasing Mechanism 935
  31.7.3 Overwriting Mechanism 937

31.8 **Read Channel Model** 939

31.9 **System Aspects** 943
  31.9.1 PES Generation for the Servo Loop .. 943
  31.9.2 Timing Recovery 945
  31.9.3 Considerations on Capacity and Data Rate 946

31.10 **Conclusions** 948

References 948

In the twenty-first century, the nanometer will probably play a role similar to the one played by the micrometer in the twentieth century. The nanometer scale will presumably also pervade the field of data storage, although there is so far no obvious way in conventional magnetic, optical, or transistor-based storage to achieve the nanometer scale in all three dimensions. After decades of spectacular progress, those mature technologies have entered the homestretch; imposing physical limitations loom before them.

One promising method involves the use of patterned magnetic media for which the ideal write/read concept still needs to be demonstrated. The biggest challenge, however, is to pattern the magnetic disk in a cost-effective way. If such approaches are successful, the basis for large-capacity storage in the twenty-first century might still be magnetism.

Other proposals call for totally different media and techniques such as local probes, near-field optics, magnetic super-resolution or holographic methods. In general, when an existing technology is about to reach its limits in the course of its evolution and alternatives are emerging in parallel, two things usually happen: First, the existing and well-established technology is explored further and every effort is made to push its limits to the utmost in order to get the maximum return on the consid-

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

erable investments made. Then, when all possibilities for improvement have been exhausted, the technology may still survive for certain niche applications, but the emerging technology will take over, opening new perspectives and new directions.

In many fields today we are witnessing the transition of structures from the micrometer to the nanometer scale, a dimension that nature has long been using to build the finest devices with a high degree of local functionality. Many of the techniques we use today are not suitable for the nanometer age; some will require minor or major modifications, others will be partially or entirely replaced. It is certainly difficult to predict which techniques will fall into which category. In key information technology areas, such as nano-electronics and data storage, it is not yet obvious which technologies and materials will be used in the future.

In any case, for an emerging technology to be seriously considered as a candidate to replace an existing technology that is approaching its inherent limits, it must provide a long-term perspective. For instance, the silicon microelectronics and storage industries are huge and have exacted correspondingly enormous investments, which makes them long-term oriented by nature. The consequence for storage is that any novel technology with higher areal storage density than today's magnetic recording [31.1, 2] should have long-term potential for further scaling, preferably down to the nanometer or even atomic scale.

The only available tool known today that is simple and yet provides this very long-term perspective is the nanometer-sharp tip. The simple tip is a very reliable tool that concentrates on one functionality: the ultimate local confinement of interaction. Techniques that use nanometer-sharp tips for imaging and investigating the structure of materials down to the atomic scale, such as the atomic force microscope (AFM) and the scanning tunneling microscope (STM) [31.3, 4], are suitable for the development of ultrahigh-density storage devices.

In the early 1990s, Mamin and Rugar at the IBM Almaden Research Center pioneered the capability of using an AFM tip for writing and readback of topographic features for data storage. In one of their schemes [31.5], writing and reading were demonstrated with a single AFM tip in contact with a rotating polycarbonate substrate. The writing was performed thermomechanically by heating the tip. In this way, storage densities of up to 30 Gb/in^2 were achieved, constituting a significant advance over the densities of that time. Later refinements included increasing readback speeds to attain a data rate of 10 Mb/s [31.6] and the implementation of track servoing [31.7].

When using single tips in AFM or STM operation for storage, one has to deal with their fundamental data-rate limitations. At present, the mechanical resonant frequencies of the AFM cantilevers limit the data rates of a single cantilever to a few Mb/s for AFM data storage [31.8, 9]. The feedback speed and low tunneling currents limit STM-based storage approaches to even lower data rates.

Currently a single AFM operates at best on the microsecond time scale. Conventional magnetic storage, however, operates at best on the nanosecond time scale, making it clear that AFM data rates have to be improved by at least three orders of magnitude to be competitive with current and future magnetic recording. One solution for substantially increasing the data rates achievable by tip-based storage devices is to employ micro-electro-mechanical system (MEMS) arrays of cantilevers operating in parallel, with each cantilever performing write/read/erase operations in an individual storage field. It is our conviction that very large-scale integrated (VLSI) micro/nanomechanics will greatly complement future micro- and nanoelectronics (integrated or hybrid) and may generate hitherto inconceivable applications of VLSI-MEMS.

Various efforts are under way to develop MEMS-based storage devices. For example, a MEMS-actuated magnetic-probe-based storage device capable of storing 2 GB of data on 2 cm^2 of die area and whose fabrication is compatible with a standard integrated circuit manufacturing process is described by *Carley* et al. [31.10]. In their device, a magnetic storage medium is positioned in the x/y plane, and writing is achieved magnetically by using an array of probe tips, each tip being actuated in the z-direction. Another concept is the atomic resolution storage described by *Gibson* et al. [31.11], who employ electron field emitters to change the state of a phase-change medium in a bit-wise fashion from polycrystalline to amorphous or vice versa. Reading is done with lower currents by detecting either back-scattered electrons or changes in the semiconductor properties in the media.

The "millipede" concept presented here is a new approach for storing data at high speed and ultrahigh density. The interesting part is that millipede stores digital information in a completely different way from magnetic hard disks, optical disks, and transistor-based memory chips. The ultimate locality is provided by a tip, and high data rates are a result of massive parallel operation of such tips. As storage medium, polymer films are being considered, although the use of other media, in particular, magnetic materials, has not been ruled out. Our current effort is focused on demonstrating the millipede concept with areal densities of up to 0.5–1 Tb/in^2 and parallel operation of very large 2-D (up to 64×64) AFM cantilever arrays with integrated tips and write/read/erase functionality. The fabrication and integration of such a large number of mechanical devices (cantilever beams) will lead to what we envision as the VLSI age of micro- and nanomechanics.

In this chapter, the millipede concept for a MEMS-based storage device is described in detail. In particular, various aspects pertaining to AFM thermomechanical read/write/erase functions, 2-D array fabrication and characteristics, $x/y/z$ microscanner design, polymer media properties, read channel modeling, servo control and synchronization, as well as modulation coding techniques suitable for probe-based data-storage devices are discussed.

31.1 The Millipede Concept

A 2-D AFM cantilever array storage technique [31.13–15], internally called the millipede, is illustrated in Fig. 31.1. Information is stored as sequences of indentations and no indentations written on nanometer-thick polymer films using the array of AFM cantilevers. The presence and absence of indentations will also be referred to as logical marks. Each cantilever performs write/read/erase operations over an individual storage field with an area on the order of 100×100 µm^2. Write/read operations depend on a mechanical parallel x/y scanning of either the entire cantilever array chip or the storage medium. The tip-medium contact is maintained and controlled globally, i. e., not on an individual cantilever basis, by using a feedback control in the z-direction for the entire chip, which greatly simplifies the system. This basic concept of the entire chip approach/leveling has been tested and demonstrated for the first time by parallel imaging with a 5×5 array chip [31.16, 17]. These parallel imaging results have shown that all 25 cantilever tips have approached the substrate within less than 1 µm of z-actuation, which indicates that overall chip tip-apex height control to within 500 nm is feasible. The stringent requirement for tip-apex uniformity over the entire chip is determined by the uniform force required to reduce tip and medium wear due to large force variations resulting from large tip-height nonuniformities [31.7]. Moreover, as the entire array is tracked without individual lateral cantilever positioning, thermal expansion of the array chip has to be small or well controlled. Thermal expansion considerations are a strong argument for a 2-D instead of a 1-D array arrangement.

Efficient parallel operations of large 2-D arrays can be achieved by a row/column time-multiplexed addressing scheme similar to that implemented in DRAMs. In the case of the millipede, the multiplexing scheme is used to address the array column by column with full parallel write/read operation within one column [31.18]. In particular, readback signal samples are obtained by applying an electrical read pulse to the cantilevers in a column of the array, low-pass filtering the cantilever response signals, and finally sampling the filter output signals. This process is repeated sequentially until all columns of the array have been addressed, and then restarted from the first column. The time between two pulses applied to the cantilevers of the same column corresponds to the time it takes for a cantilever to move from one logical mark position to the next. An alternative approach is to access all or a subset of the cantilevers simultaneously without resorting to the row/column multiplexing scheme. Clearly, the latter scheme yields higher data rates, whereas the former leads to lower implementation complexity of the channel electronics.

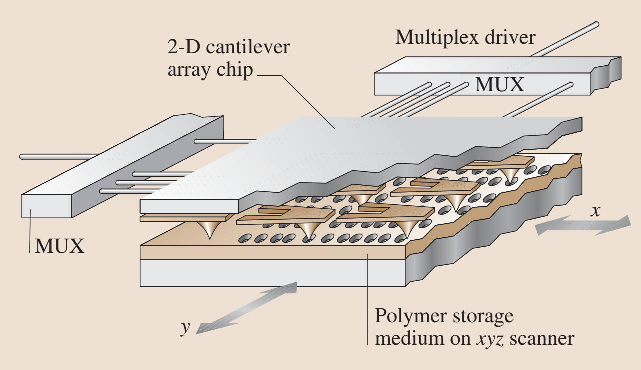
2-D cantilever array chip
Multiplex driver
MUX
MUX
x
y
Polymer storage medium on xyz scanner

Fig. 31.1 The millipede concept (after [31.12])

31.2 Thermomechanical AFM Data Storage

In recent years, AFM thermomechanical recording in polymer storage media has undergone extensive modifications mainly with respect to the integration of sensors and heaters designed to enhance simplicity and to increase data rate and storage density. Thermomechanical writing in polycarbonate films and optical readback was first investigated and demonstrated by Mamin and Rugar (1992). Using heater cantilevers, thermomechanical recording at 30 Gb/in^2 storage density and data rates of a few Mb/s for reading and 100 kb/s for writing have been demonstrated [31.5, 6, 19]. Although the storage density of 30 Gb/in^2 obtained originally was not overwhelming, the results were encouraging enough to consider using polymer films to achieve density improvements. The current millipede storage approach is based on a new thermomechanical write/read process in nanometer-thick polymer films.

Thermomechanical writing is achieved by applying a local force through the cantilever/tip to the polymer layer and simultaneously softening the polymer layer by local heating. Initially, the heat transfer from the tip to the polymer through the small contact area is very poor, but it improves as the contact area increases. This means that the tip must be heated to a relatively high temperature of about 400 °C to initiate softening. Once softening has been initiated, the tip is pressed into the polymer, and hence the indentation size is increased. Rough estimates [31.20, 21] indicate that at the beginning of the writing process only about 0.2% of the heating power is used in the very small contact zone (10–40 nm^2) to soften the polymer locally, whereas about 80% is lost through the cantilever legs to the chip body and about 20% is radiated from the heater platform through the air gap to the medium/substrate. After softening has started and the contact area has increased, the heating power available for generating the indentations increases at least ten times to reach 2% or more of the total heating power.

With this highly nonlinear heat-transfer mechanism it is very difficult to achieve small tip penetration and hence small bit sizes, as well as to control and reproduce the thermomechanical writing process. This situation can be improved if the thermal conductivity of the substrate is increased and if the depth of tip penetration is limited. We have explored the use of very thin polymer layers deposited on Si substrates to improve these characteristics [31.22, 23], as illustrated in Fig. 31.2. The hard Si substrate prevents the tip from penetrating farther than the film thickness, and it enables more rapid transport of heat away from the heated region, as Si is a much better conductor of heat than the polymer. Using coated Si substrates with a 50-nm film of polymethylmethacrylate (PMMA), we have achieved indentation sizes of between 10 and 50 nm. However, increased tip wear has occurred, probably caused by contact between the Si tip and the Si substrate during writing. Therefore, a 70-nm layer of crosslinked photoresist (SU-8) was introduced between the Si substrate and the PMMA film to act as a softer penetration stop that avoids tip wear, but remains thermally stable.

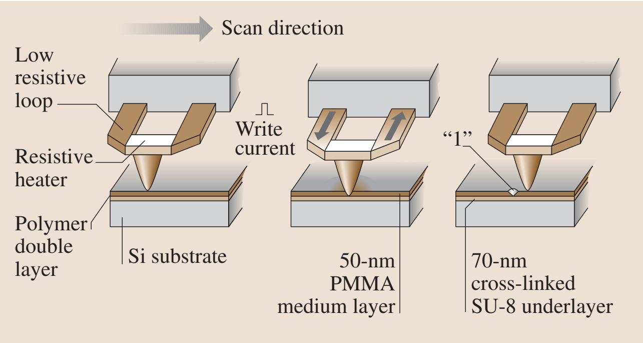

Fig. 31.2 New storage medium used for writing small bits. A thin, writable PMMA layer is deposited on a Si substrate separated by a cross-linked film of epoxy photoresist (after [31.15])

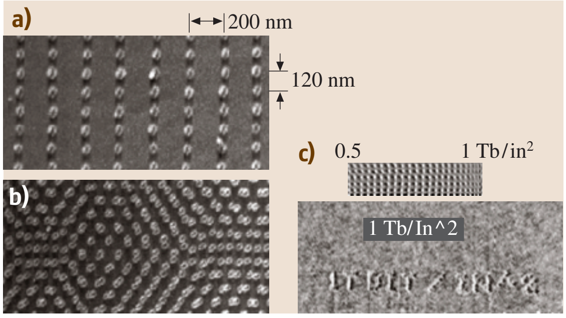

Fig. 31.3a–c Series of 40-nm indentations formed in a uniform array with **(a)** 120-nm pitch and **(b)** variable pitch ($\geq$ 40 nm), resulting in areal densities of up to 400 Gb/in^2. Images obtained with a thermal readback technique. **(c)** Ultra-high-density bit writing with areal densities approaching 1 Tb/in^2. The scale is the same for all three images (from [31.15] © 2002 IEEE)

Using this layered storage medium, indentations 40 nm in diameter have been written as shown in Fig. 31.3. These experiments were performed using a 1-µm-thick, 70-µm-long, two-legged Si cantilever [31.19]. The cantilever legs are made highly conducting by high-dose ion implantation, whereas the heater region remains low-doped. Electrical pulses 2 µs in duration were applied to the cantilever with a period of 50 µs. Figure 31.3a demonstrates that 40-nm bits can be written with 120-nm pitch or, as shown in Fig. 31.3b, very close to each other without merging, implying a potential areal density of 400 Gb/in^2. Figure 31.3c shows recent results from a single-lever experiment, where indentations are spaced as closely as 25 nm apart, resulting in areal densities of up to 1 Tb/in^2, although with a somewhat degraded write/read quality.

Imaging and reading are performed by a new thermomechanical sensing concept [31.24]. To read the written information, the heater cantilever originally used for writing is given the additional function of a thermal readback sensor by exploiting its temperature-dependent resistance. In general, the resistance increases nonlinearly with heating power/temperature from room temperature to a peak value at 500–700 °C. The peak temperature is determined by the doping concentration of the heater platform, which ranges from 1×10^{17} to 2×10^{18} at/cm^3. Above the peak temperature, the resistance drops as the number of intrinsic carriers increases due to thermal excitation [31.25]. For sensing, the resistor is operated at about 350 °C, a temperature that is not high enough to soften the polymer as in the case of writing. The principle of thermal sensing is based on the fact that the thermal conductance between the heater platform and the storage substrate changes according to the distance between them. The medium between the heater platform and the storage substrate, in our case air, transports heat from the cantilever to the substrate. When the distance between cantilever and storage substrate is reduced as the tip moves into an indentation, the heat transport through the air becomes more efficient. As a result, the evolution of the heater temperature differs in response to a pulse applied to the cantilever. In particular, the maximum value achieved by the heater temperature is higher in the absence of an indentation. As the value of the variable resistance depends on the temperature of the cantilever, the maximum value achieved by the resistance will be lower as the tip moves into an indentation. Therefore, during the read process, the cantilever resistance reaches different values depending on whether the tip moves into an indentation (logical bit "1") or over a region without an indentation (logical bit "0"). Figure 31.4 illustrates this concept.

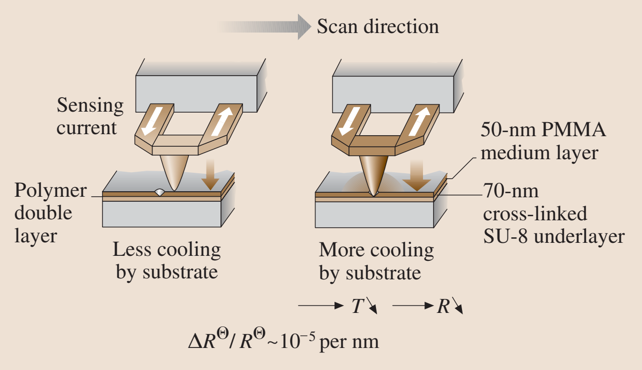
Scan direction
Sensing current
Polymer double layer
Less cooling by substrate
More cooling by substrate
50-nm PMMA medium layer
70-nm cross-linked SU-8 underlayer
→ T↘ → R↘
$\Delta R^{\Theta}/R^{\Theta} \sim 10^{-5}$ per nm

Fig. 31.4 Principle of AFM thermal sensing. The tip of the heater cantilever is continuously heated by a dc power supply while the cantilever is being scanned and the heater resistivity measured (after [31.15])

Under typical operating conditions, the sensitivity of thermomechanical sensing is even greater than that of piezoresistive-strain sensing, which is not surprising because thermal effects in semiconductors are stronger than strain effects. The good sensitivity is demonstrated by the images of the 40-nm-sized indentations in Fig. 31.3, which were obtained using the described thermal-sensing technique.

The thermomechanical cantilever sensor, which transforms temperature into an electrical signal that carries information, is the electrical equivalent, to a first degree of approximation, of a variable resistance. A detection circuit must, therefore, sense a voltage that depends on the value of the cantilever resistance to decide whether a "1" or a "0" is written. The relative variation of thermal resistance is on the order of 10^{-5}/nm. Hence a written "1" typically produces a relative change of the cantilever thermal resistance $\Delta R^{\Theta}/R^{\Theta}$ of about 10^{-4} to 5×10^{-4}. Note that the relative change of the cantilever electrical resistance is of the same order of magnitude. Thus, one of the most critical issues in detecting the presence or absence of an indentation is the high resolution required to extract the signal that contains the information about the logical bit being "1" or "0". The signal carrying the information can be regarded as a small signal superimposed on a very large offset signal. The large offset problem can be mitigated by subtracting a suitable reference signal [31.12, 26, 27].

31.3 Array Design, Technology, and Fabrication

Encouraged by the results of the 5×5 cantilever array [31.16, 17], a 32×32 array chip was designed and fabricated [31.18]. With the findings from the fabrication and operation of the 5×5 array and the very dense thermomechanical writing/reading in thin polymers with single cantilevers, some important changes of the chip functionality and fabrication processes were made. The major differences are (1) surface micromachining to form cantilevers at the wafer surface, (2) all-silicon cantilevers, (3) thermal instead of piezoresistive sensing, and (4) first- and second-level wiring with an insulating layer for a multiplexed row/column-addressing scheme.

As the heater platform functions as a read/write element and no individual cantilever actuation is required, the basic array cantilever cell becomes a simple two-terminal device addressed by a multiplexed *x*/*y* wiring, as shown in Fig. 31.5. The cell area and *x*/*y* cantilever pitch are 92×92 μm^2, which results in a total array size of less than 3×3 mm^2 for the 1,024 cantilevers. The cantilevers are fabricated entirely of silicon for good thermal and mechanical stability. They consist of a heater platform with the tip on top, legs acting as soft mechanical springs, and electrical connections to the heater. They are highly doped to minimize interconnect resistance and to replace the metal wiring on the cantilever in order to eliminate electromigration and parasitic *z*-actuation of the cantilever due to a bimorph effect. The resistive ratio between the heater and the silicon interconnect sections should be as high as possible; currently the resistance of the highly doped interconnections is ∼ 400 Ω and that of the heater platform is 5 kΩ (at 3 V reading bias).

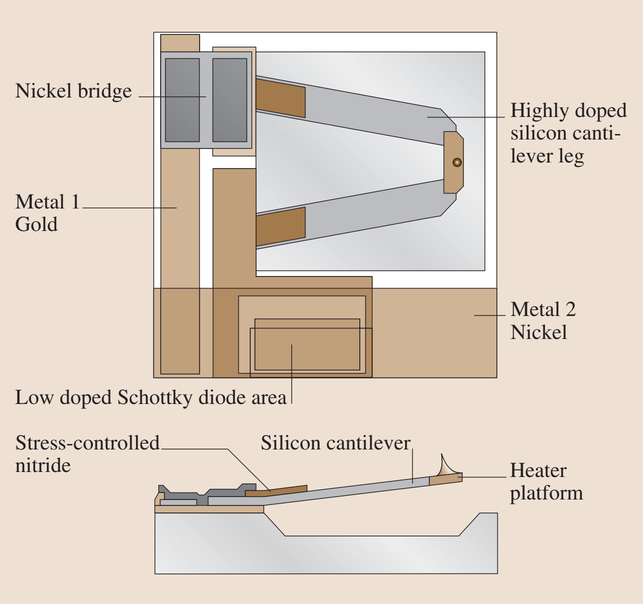
Nickel bridge
Metal 1
Gold
Highly doped silicon cantilever leg
Metal 2
Nickel
Low doped Schottky diode area
Stress-controlled nitride
Silicon cantilever
Heater platform

Fig. 31.5 Layout and cross section of one cantilever cell (after [31.28])

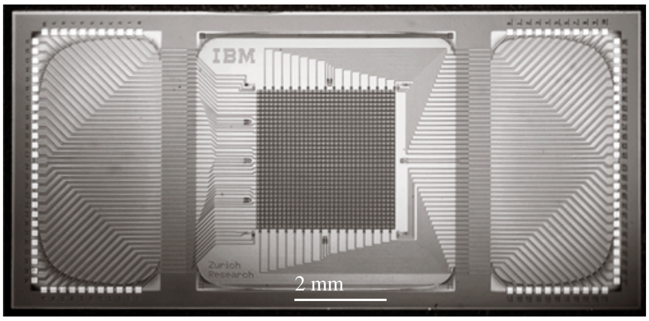

Fig. 31.6 Photograph of fabricated chip (14×7 mm^2). The 32×32 cantilever array is located at the center with bond pads distributed on either side (from [31.28] © 1999 IEEE)

The cantilever mass has to be minimized to obtain soft, high-resonant-frequency cantilevers. Soft cantilevers are required for a low loading force in order to eliminate or reduce tip and medium wear, whereas a high resonant frequency allows high-speed scanning. In addition, sufficiently wide cantilever legs are required for a small thermal time constant, which is partly determined by cooling via the cantilever legs [31.19]. These design considerations led to an array cantilever with 50-μm-long, 10-μm-wide and 0.5-μm-thick legs, and a 5-μm-wide, 10-μm-long and 0.5-μm-thick heater platform. Such a cantilever has a stiffness of ∼ 1 N/m and a resonant frequency of ∼ 200 kHz. The heater time constant is a few microseconds, which should allow a multiplexing rate of up to 100 kHz.

The tip height should be as small as possible because the heater platform sensitivity depends strongly on the platform-to-medium distance. This contradicts the requirement of a large gap between the chip surface and the storage medium to ensure that only the tips, and not the chip surface, make contact with the medium. Instead of making the tips longer, we purposely bent the cantilevers a few micrometers out of the chip plane by depositing a stress-controlled plasma-enhanced chemical vapor deposition (PECVD) silicon-nitride layer at the base of the cantilever (see Fig. 31.5). This bending as well as the tip height must be well controlled in order to maintain an equal loading force for all cantilevers of an array.

Cantilevers are released from the crystalline Si substrate by surface micromachining using either plasma or wet chemical etching to form a cavity underneath the cantilever. Compared to a bulk-micromachined through-wafer cantilever-release process, as was done for our 5 × 5 array [31.16, 17], the surface micromachining technique allows an even higher array density and yields better mechanical chip stability and heat sinking. As mentioned above, the entire array is tracked without individual lateral cantilever positioning, therefore thermal expansion of the array chip has to be small or well controlled. For a 3 × 3-mm^2 silicon array area and 10-nm tip-position accuracy, the temperature difference between array chip and medium substrate has to be controlled to about 1 °C. This is ensured by four temperature sensors in the corners of the array and heater elements on each side of the array.

The photograph in Fig. 31.6 shows a fabricated chip with the 32 × 32 array located in the center (3 × 3 mm^2) and the electrical wiring interconnecting the array with the bonding pads at the chip periphery. Figure 31.7 shows the 32 × 32 array section of the chip with the independent approach/heat sensors in the four corners and the heaters on each side of the array, as well as zoomed scanning electron micrographs (SEMs) of an array section, a single cantilever, and a tip apex. The tip height is 1.7 µm and the apex radius is smaller than 20 nm, which is achieved by oxidation sharpening [31.29]. The cantilevers are connected to the column and row address lines using integrated Schottky diodes in series with the cantilevers. The diode is operated in reverse bias (high resistance) if the cantilever is not addressed, thereby greatly reducing cross talk between cantilevers. More details about the array fabrication are given in [31.28, 30].

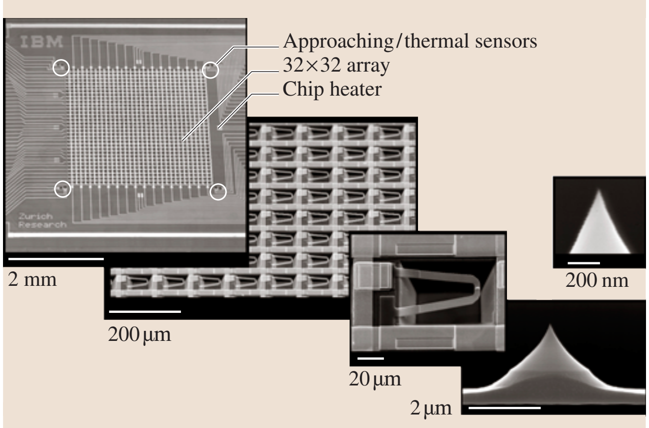

Fig. 31.7 SEM images of the cantilever array section with approaching and thermal sensors in the corners, array and single cantilever details, and tip apex (© 2000 International Business Machines Corporation; after [31.18])

31.4 Array Characterization

The array’s independent cantilevers, which are located in the four corners of the array and used for approaching and leveling the chip and storage medium, serve to initially characterize the interconnected array cantilevers. Additional cantilever test structures are distributed over the wafer; they are equivalent to but independent of the array cantilevers. Figure 31.8 shows an I/V curve of such a cantilever; note the nonlinearity of the resistance. In the low-power part of the curve, the resistance increases as a function of heating power, whereas in the high-power regime, it decreases.

In the low-power, low-temperature regime, silicon mobility is affected by phonon scattering, which depends on temperature, whereas at higher power, the intrinsic temperature of the semiconductor is reached, which results in a resistivity drop owing to the increasing number of carriers [31.25]. Depending on the heater-platform doping concentration of 1×10^{17} to 2×10^{18} at/cm^3, our calculations estimate a resistance maximum at a temperature of 500 to 700 °C, respectively.

The cantilevers within the array are electrically isolated from one another by integrated Schottky diodes. As every parasitic path in the array to the cantilever addressed contains a reverse-biased diode, the cross talk current is drastically reduced, as shown in Fig. 31.9. Thus, the current response of an addressed cantilever in an array is nearly independent of the size of the array, as demonstrated by the I/V curves in Fig. 31.9. Hence, the power applied to address a cantilever is not shunted by other cantilevers, and the reading sensitivity is not degraded – not even for very large arrays (32 × 32). The introduction of the electrical isolation using integrated

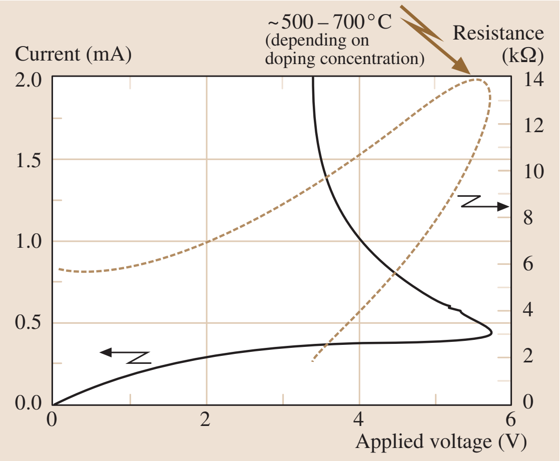

Fig. 31.8 I/V curve of one cantilever. The curve is nonlinear owing to the heating of the platform as the power and temperature are increased. For doping concentrations between 1×10^{17} and 2×10^{18} at/cm^3, the maximum temperature varies between 500 and 700 °C (after [31.30])

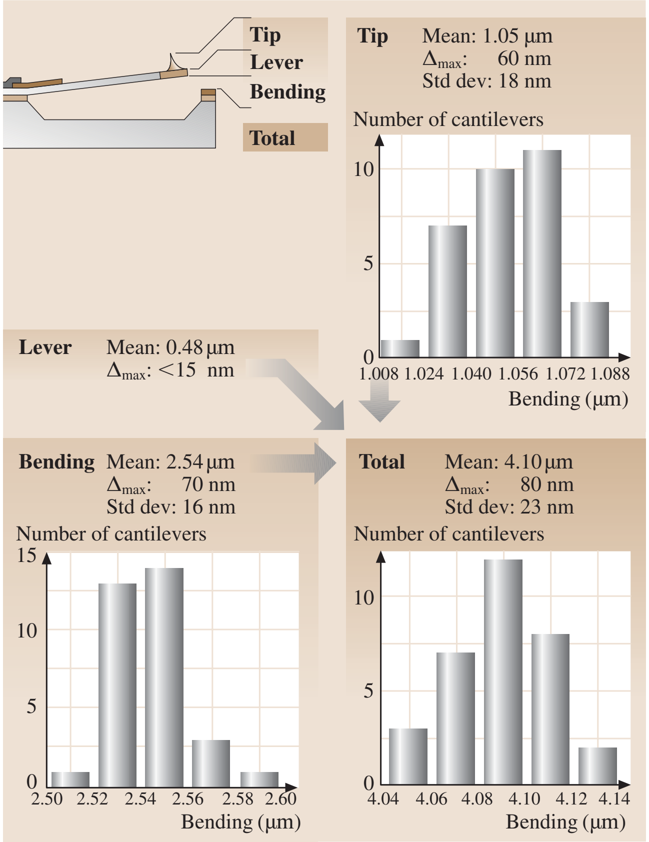

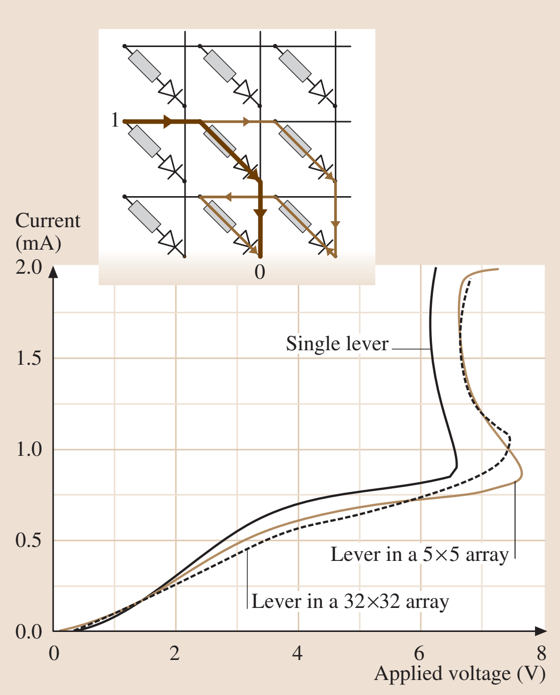

Fig. 31.9 Comparison of the I/V curve of an independent cantilever (*solid black line*) with the current response when addressing a cantilever in a 5×5 (*solid brown line*) or a 32×32 (*dashed line*) array with a Schottky diode serially connected to the cantilever. Little change is observed in the I/V curve between the different cases. Inset: sketch of the direct path (*bold line*) and a parasitic path (*thin line*) in a cantilever-diode array. In the parasitic path there is always one diode in reverse bias that reduces the parasitic current (after [31.30])

Schottky diodes turned out to be crucial for the successful operation of interconnected cantilever arrays with a simple time-multiplexed addressing scheme.

The tip-apex height uniformity within an array is very important because it determines the force of each cantilever while in contact with the medium and hence influences write/read performance, as well as medium and tip wear. Wear investigations suggest that a tip-apex height uniformity across the chip of less than 500 nm

Fig. 31.10 Tip-apex height uniformity across one cantilever row of the array with individual contributions from the tip height and cantilever bending. (© 2000 International Business Machines Corporation; after [31.18])

is required [31.7], with the exact number depending on the spring constant of the cantilever. In the case of the millipede, the tip-apex height is determined by the tip height and the cantilever bending. Figure 31.10 shows the tip-apex height uniformity of one row of the array (32 tips) due to tip height and cantilever bending. It demonstrates that our uniformity is on the order of 100 nm, thus meeting the requirements.

31.5 *x*/*y*/*z* Medium Microscanner

A key issue for the millipede concept is the need for a low-cost, miniaturized scanner with $x/y/z$ motion capabilities and a lateral scanning range on the order of 100 μm. Multiple-probe systems arranged as 1-D or 2-D arrays [31.18] must also be able to control, by means of tilt capabilities, the parallelism between the probe array and the sample [31.18, 32].

We have developed a microscanner with these properties based on electromagnetic actuation. It consists of a mobile platform supported by springs and containing integrated planar coils positioned over a set of miniature permanent magnets [31.33]. A suitable arrangement of the coils and magnets allows us, by electrically addressing the various coils, to apply magnetically induced forces to the platform and drive it in the x, y, z, and tilt directions. Our first silicon/copper-based version of this device has proved the validity of the concept [31.34], and variations of it have since been used elsewhere [31.35]. However, the undamped copper spring system gave rise to excessive cross talk and ringing when driven in an open loop, and its layout limited the compactness of the overall device.

We investigate a modified microscanner that uses flexible rubber posts as a spring system and a copper-epoxy-based mobile platform, Fig. 31.11. The platform is made of a thick, epoxy-based SU-8 resist [31.36], in which the copper coils are embedded. The posts are made of polydimethylsiloxane (PDMS, Sylgard 184 silicon elastomer, Dow Corning, Midland, MI) and are fastened at the corners of the platform and at the ground plate, providing an optimally compact device by sharing the space below the platform with the magnets. The shape of the posts allows their lateral and longitudinal stiffness to be adjusted, and the dissipative rubber-like properties of PDMS provide damping to avoid platform ringing and to suppress nonlinearities.

Figure 31.12 shows the layout of the platform, which is scaled laterally, so that the long segments of the "racetrack" coils used for in-plane actuation coincide with commercially available 24 mm^2 SmCo magnets. The thickness of the device is determined by that of the magnets (1 mm), the clearance between magnet and platform (500 μm), and the thickness of the platform itself, which is 250 μm and determined mainly by the aspect ratio achievable in SU-8 resist during the exposure of the coil plating mold. The resulting device volume is approximately $15 \times 15 \times 1.6$ mm^3.

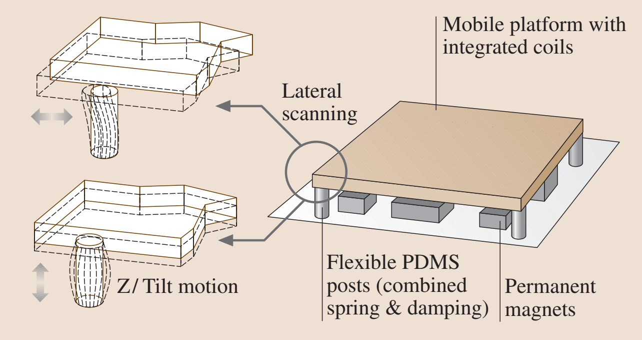
Mobile platform with integrated coils
Lateral scanning
Z/Tilt motion
Flexible PDMS posts (combined spring & damping)
Permanent magnets

Fig. 31.11 Microscanner concept using a mobile platform and flexible posts (after [31.31])

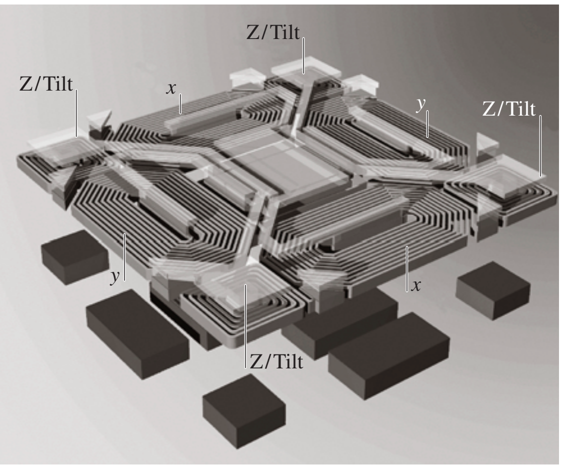

Fig. 31.12 Arrangement of the coils, the interconnects, and the permanent magnets, as well as the various motions addressed by the corresponding coils (from [31.15] © 2002 IEEE)

The SmCo magnets produce a measured magnetic field intensity of ∼ 0.14 T at the mid-thickness of the coils. The effective coil length is 320 mm, yielding an expected force $F_{x,y}$ of 45 µN per mA of drive current.

The principal design issue of the spring system is the ratio of its stiffnesses for in-plane and out-of-plane motion. Whereas for many scanning probe applications the required z-axis range need not be much larger than a few microns, it is necessary to ensure that the z-axis retraction of the platform due to the shortening of the posts as they take on an S-shape at large in-plane deflections can be compensated for at acceptable z-coil current levels. Various PDMS post shapes have been investigated to optimize and trade off the various requirements. Satisfactory performance was found for simple O-shapes [31.31].

The fabrication of the scanner, Fig. 31.13, starts on a silicon wafer with a seed layer and a lithographically patterned 200-µm-thick SU-8 layer, in which copper is electroplated to form the coils (Fig. 31.13a). The coils typically have 20 turns, with a pitch of 100 µm and a spacing of 20 µm. Special care was taken in the resist processing and platform design to achieve the necessary aspect ratio and to overcome adhesion and stress problems of SU-8. A second SU-8 layer, which serves as an insulator, is patterned with via holes, and another seed layer is then deposited (Fig. 31.13b). Next, an interconnect level is formed using a Novolac-type resist mask and a second copper-electroplating step (Fig. 31.13c). After stripping the resist, the silicon wafer is dissolved by a sequence of wet and dry etching, and the exposed seed layers are sputtered away to prevent shorts (Fig. 31.13d).

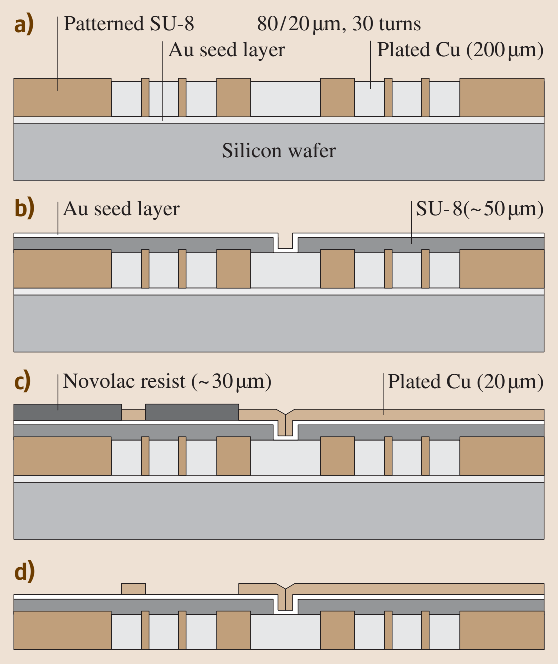

Fig. 31.13a–d Cross section of the platform fabrication process. **(a)** Coils are electroplated through an SU-8 resist mask, which is retained as the body of the platform; **(b)** an insulator layer is deposited; **(c)** interconnects are electroplated; **(d)** the platform is released from the silicon substrate (after [31.31])

The motion of the scanner was characterized using a microvision strobe technique [31.37]. The results presented below are based on O-type PDMS ports. Frequency response curves for in-plane motion (Fig. 31.14) show broad peaks (characteristic of a large degree of damping) at frequency values that are consistent with expectations based on the measured mass of the platform (0.253 mg). The amplitude response (Fig. 31.15) displays the excellent linearity of the spring system for displacement amplitudes up to 80 µm (160-µm displacement range). Based on these near-DC (10 Hz) responses (∼ 1.4 µm/mA) and a measured circuit resistance of 1.9 Ω, the power necessary for a 50-µm displacement amplitude is approximately equal to 2.5 mW.

Owing to limitations of the measurement technique, it was not possible to measure out-of-plane displacements greater than 0.5 µm. However, the small-

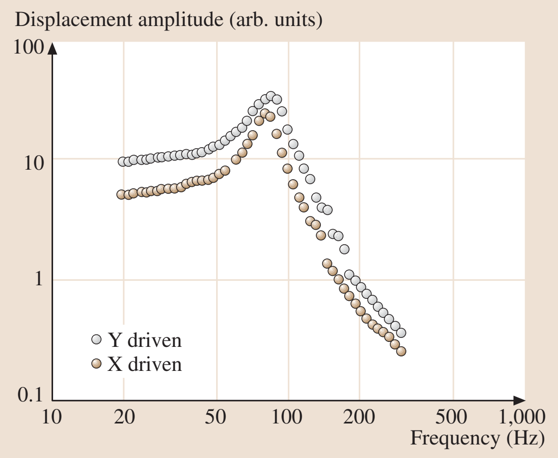

Fig. 31.14 Frequency response for in-plane x- and y-axis motion. The mechanical quality factors measured are between 3.3 and 4.6 (after [31.15])

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

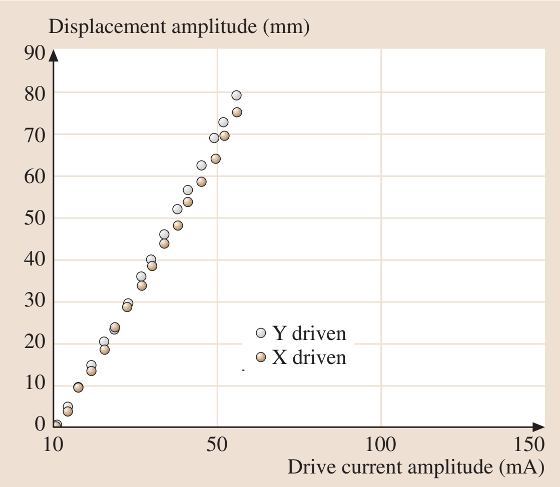

Fig. 31.15 In-plane displacement amplitude response for an ac drive current at 10 Hz (off resonance) (after [31.15])

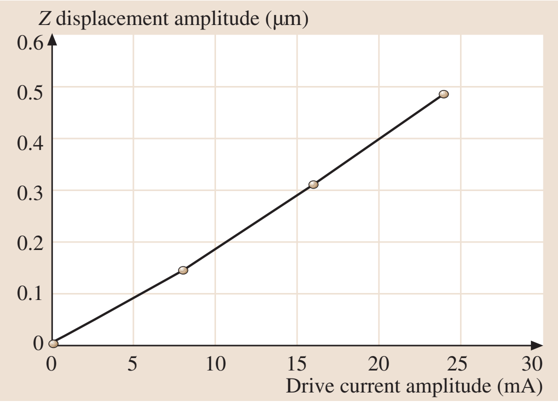

Fig. 31.16 Out-of-plane amplitude response for an ac drive current at 3 Hz. The drive current is the total for all four corner coils, which are driven in phase (after [31.31])

amplitude response for z-motion when all four corner coils are driven in-phase also displays good linearity over the range that can be measured (Fig. 31.16).

The electromagnetic scanner performs reliably and as predicted in terms of the scan range, device volume, and power requirements, achieving overall displacement ranges of 100 µm with approximately 3 mW of power. The potential access time in the 100-µm storage field is on the order of a few milliseconds. By being potentially cheap to manufacture, the integrated scanner presents a good alternative actuation system for many scanning probe applications.

31.6 First Write/Read Results with the 32×32 Array Chip

We have built a prototype that includes all the basic building blocks of the millipede concept (see Fig. 31.1) [31.38]. A 3 × 3-mm^2 silicon substrate is spin-coated with the SU-8/PMMA polymer medium, as described in Sect. 31.3. This storage medium is attached to the $x/y/z$ microscanner and approaching device. The magnetic z-approaching actuators bring the medium into contact with the tips of the array chip. The z-distance between the medium and the millipede chip is controlled by the approaching sensors in the corners of the array. The signals from these cantilevers are used to determine the forces on the z-actuators and, hence, the forces of the cantilever while it is in contact with the medium. This sensing/actuation feedback loop continues to operate during x/y scanning of the medium. The PC-controlled write/read scheme addresses the 32 cantilevers of one row in parallel. Writing is performed by connecting the addressed row for 20 µs to a high negative voltage and simultaneously applying data inputs (“0” or “1”) to the 32 column lines. The data input is a high positive voltage for a “1” and ground for a “0”. This row-enabling and column-addressing scheme supplies a heater current to all cantilevers, but only those cantilevers with high positive voltage generate an indentation (“1”). Those grounded are not hot enough to form an indentation and thus write a “0”. When the scan stage has moved to the next logical mark position, the process is repeated, and this is continued until the line scan is finished. In the read process, the selected row line is connected to a moderate negative voltage, and the column lines are grounded via a protection resistor of about 10 kΩ, which keeps the cantilevers warm. During scanning, the voltages across the resistors are measured. Depending on the topography of the recording surface, the degree of cooling of each cantilever varies, thus changing the resistance and voltage across the series resistor and allowing written data to be read back.

The results of writing and reading in this fashion can be seen in Fig. 31.17, which shows 1,024 images corresponding to the 1,024 storage fields and associ-

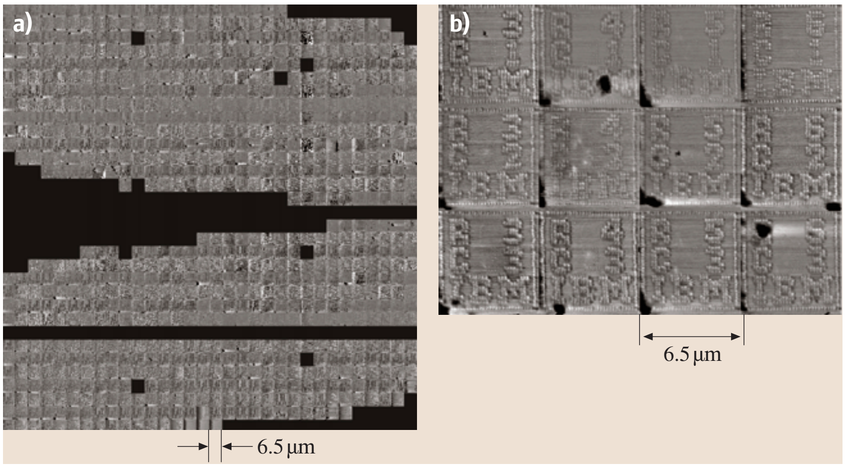

Fig. 31.17 **(a)** 1,024 images, each one obtained from a cantilever of the 2-D array. **(b)** Enlarged view of typical images from **(a)**. The numbers in the images indicate the row and column of each lever (from [31.38])

ated cantilevers. Of the 1,024 levers, 834 were able to write and read data back, i. e., a success rate of more than 80%. The sequence is as follows. First, a bit pattern is written simultaneously by each of the levers in row 1, then read back simultaneously, followed by row 2, etc. through row 32. The data sent to the levers is different, each lever writing its own row and column number in the array. The bit pattern consists of 64 × 64 bits, where odd-numbered bits are always set to 0. In this case, the area used is 6.5 × 6.5 μm^2. The readback image is a grey-scale bit map of 128 × 128 pixels. The distance between levers is 92 µm, so the images in Fig. 31.17 are also 92 µm apart. The areal density of the written information shown in Fig. 31.17 corresponds to 15–30 Gb/in^2, depending on the coding scheme adopted. More recently, an areal density of 150–200 Gb/in^2, at an array yield of about 60%, has been demonstrated.

Those levers that did not read back failed for one of four reasons: (i) a defective chip connector rendered an entire column unusable, (ii) a point defect occurred, meaning that a single lever or tip was broken, (iii) there was a nonuniformity of the tip contact due to tip/lever variability or storage substrate bowing due to mounting, (iv) there were thermal drifts. The latter two reasons were the most likely and major failure sources. At present, there is clearly a tradeoff between the number of working levers and the density, which will most likely be resolved by a better substrate/chip mounting technique and lower thermal drifts.

The writing and readback rates achieved with this system are 1 kb/s per lever, thus, the total data rate is about 32 kb/s. This rate is limited by the rate at which data can be transferred over the PC ISA bus, not by a fundamental time limitation of the read/write process.

31.7 Polymer Medium

The polymer storage medium plays a crucial role in millipede-like thermomechanical storage systems. The thin-film multilayer structure with PMMA as active layer (see Fig. 31.2) is not the only possible choice, considering the almost unlimited range of polymer materials available. The ideal medium should be easily deformable for writing, yet indentations should be stable against tip wear and thermal degradation. Finally, one would also like to be able to erase and rewrite data repeatedly. In order to be able to address all important aspects properly, some understanding of the basic physical mechanism of thermomechanical writing and erasing is required.

31.7.1 Writing Mechanism

In a *gedanken* experiment we visualize writing of an indentation as the motion of a rigid body (the tip) in a viscous medium (the polymer melt). Let us initially assume that the polymer, i. e., PMMA, behaves like a simple liquid after it has been heated above the glass-transition temperature in a small volume around the tip. As viscous drag forces must not exceed the loading force applied to the tip during indentation, we can estimate an upper bound for the viscosity ζ of the polymer melt using Stokes's equation:

$$F = 6\pi\zeta\varrho v\,. \tag{31.1}$$

In actual indentation formation, the tip loading force is on the order of $F = 50\,\text{nN}$ and the radius of curvature at the apex of the tip is typically $\varrho = 20\,\text{nm}$. Assuming a depth of the indentation of, say, $h = 50\,\text{nm}$ and a heat pulse of $\tau_h = 10\,\mu\text{s}$ duration, the mean velocity during indentation formation is on the order of $v = h/\tau_h = 5\,\text{mm/s}$. Note that thermal relaxation times are on the order of microseconds [31.20, 21] and, hence, the heating time can be equated to the time it takes to form an indentation. With these parameters we obtain $\zeta < 25\,\text{Pa s}$, whereas typical values for the shear viscosity of PMMA are at least seven orders of magnitude larger even at temperatures well above the glass-transition point [31.39].

This apparent contradiction can be resolved by considering that polymer properties are strongly dependent on the time scale of observation. At time scales on the order of 1 ms and below, entanglement motion is in effect frozen in and the PMMA molecules form a relatively static network. Deformation of the PMMA now proceeds by means of uncorrelated deformations of short molecular segments, rather than by a flow mechanism involving the coordinated motion of entire molecular chains. The price one has to pay is that elastic stress builds up in the molecular network as a result of the deformation (the polymer is in a so-called rubbery state). On the other hand, corresponding relaxation times are orders of magnitude smaller, giving rise to an effective viscosity at millipede time scales on the order of 10 Pa s [31.39], as required by our simple argument (see (31.1)). Note that, unlike normal viscosity, this high-frequency viscosity is basically independent of the detailed molecular structure of the PMMA, i. e., chain length, tacticity, polydispersity, etc. In fact, we can even expect that similar high-frequency viscous properties can be found in a large class of other polymer materials, which makes thermomechanical writing a rather robust process in terms of material selection.

We have argued above that elastic stress builds up in the polymer film during the formation of an indentation, creating a corresponding reaction force on the tip on the order of $F_r \sim 2\pi G\varrho^2$, where G denotes the elastic shear modulus of the polymer [31.40]. An important property for millipede operation is that the shear modulus drops by several orders of magnitude in the glass-transition regime, i. e., for PMMA from $\sim 1\,\text{GPa}$ below Θ_g to $\sim 0.5{-}1\,\text{MPa}$ above Θ_g, where Θ_g denotes the glass-transition temperature [31.39]. The bulk modulus, on the other hand, retains its low-temperature value of several GPa. Hence, in this elastic regime, formation of an indentation above Θ_g constitutes a volume-preserving deformation. For proper indentation formation, the tip load must be balanced between the extremes of the elastic reaction force F_r for temperatures below and above Θ_g, i. e., $F \ll 2.5\,\mu\text{N}$ for PMMA to prevent indentation of the polymer in the cold state and $F \gg 2.5\,\text{nN}$ to overcome the elastic reaction force in the hot state. Unlike the deformation of a simple liquid, the indentation represents a metastable state of the entire deformed volume, which is under elastic tension. Recovery of the unstressed initial state is prevented by rapid quenching of the indentation below the glass temperature with the tip in place. As a result, the deformation is frozen in, because below Θ_g motion of molecular-chain segments is, in effect, inhibited (see Fig. 31.18).

This mechanism also allows indentations to be erased locally – it suffices to heat the deformed volume locally above Θ_g, whereupon the indented volume reverts to its unstressed flat state driven by internal elastic stress. In addition, erasing is promoted by surface tension forces, which give rise to a restoring surface pressure on the order of $\gamma(\pi/\varrho)^2 h \approx 25\,\text{MPa}$, where $\gamma \sim 0.02\,\text{N/m}$ denotes the polymer-air surface tension.

One question immediately arises from these speculations: If the polymer behavior can be determined from the macroscopic characteristics of the shear modulus as a function of time, temperature, and pressure, can the time-temperature superposition principle also be applied in this case? The time-temperature superposition principle is a powerful concept of polymer physics [31.41]. It basically states that the time scale and the temperature are interdependent variables that determine the polymer behavior such as the shear modulus. A simple transformation can be used to translate time-dependent into temperature-dependent data and vice versa. It is not clear, however, whether this principle can be applied in our case, i. e., under such extreme conditions (high pressures, short time scales, and nanometer-sized volumes, which are clearly below the radius of gyration of individual polymer molecules).

To test this, we varied the heating time, the heating temperature, and the loading force in indentation-writing experiments on a standard PMMA sample. The results

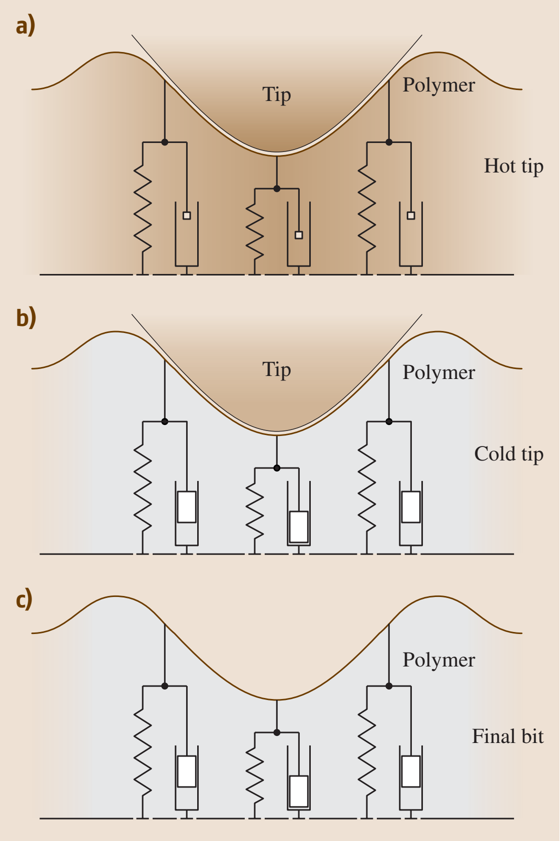

Fig. 31.18a–c Viscoelastic model of indentation writing. **(a)** The hot tip heats a small volume of polymer material to more than Θ_g. The shear modulus of the polymer drops drastically from GPa to MPa, which in turn allows the tip to indent the polymer. In response, elastic stress (represented as compression springs) builds up in the polymer. In addition, viscous forces (represented as pistons) associated with the relaxation time for the local deformation of molecular segments limit the indentation speed. **(b)** At the end of the writing process, the temperature is quenched on a microsecond time scale to room temperature: The stressed configuration of the polymer is frozen-in (represented by the locked pistons). **(c)** The final indentation corresponds to a metastable configuration. The original unstressed flat state of the polymer can be recovered by heating the indentation volume to more than Θ_g, which unlocks the compressed springs (after [31.15])

are summarized in Fig. 31.19. The minimum heater temperature at which the formation of an indentation starts for a given heating-pulse length and loading force was determined. This so-called threshold temperature is plotted against the heating-pulse length. A careful calibration of the heater temperature has to be done to allow a comparison of the data. The heater temperature was determined by assuming proportionality between temperature and electrical power dissipated in the heater resistor at the end of the heating pulse when the tip has reached its maximum temperature. An absolute temperature scale is established using two well-defined reference points. One is room temperature, corresponding to zero electrical power. The other is provided by the point of turnover from positive to negative differential resistance (see Fig. 31.8), which corresponds to a heater temperature of approximately 550 °C. The general shape of the measured threshold temperature versus heating time curves indeed shows the characteristics of time-temperature superposition. In particular, the curves

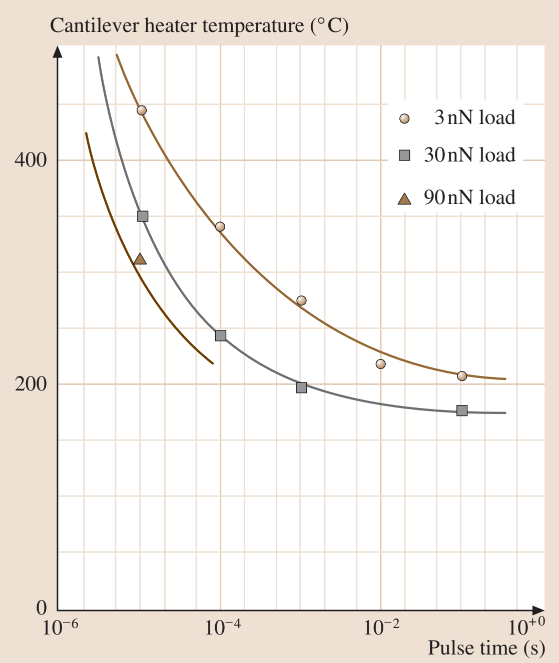

Fig. 31.19 Indentation-writing threshold measurements. The load was controlled by pushing the cantilever/tip into the sample with a controlled displacement and a known spring constant of the cantilever. When a certain threshold is reached, the indentations become visible in subsequent imaging scans (see also Fig. 31.21). The *solid lines* are guides to the eye. Curves of similar shape would be expected from the time-temperature superposition principle (after [31.15])

are identical up to a load-dependent shift with respect to the time axis. Moreover, we observe that the time it takes to form an indentation at constant heater temperature is inversely proportional to the tip load. This property is exactly what one would expect if internal friction (owing to the high-frequency viscosity) is the rate-limiting step in forming an indentation (see (31.1)).

The time it takes to heat the indentation volume of polymer material higher than the glass-transition temperature is another potentially rate-limiting step. Here, the spreading resistance of the heat flow in the polymer and the thermal contact resistance are the most critical parameters. Simulations suggest [31.20, 21] that equilibration of temperature in the polymer occurs within less than 1 µs. Very little is known, however, about the thermal coupling efficiency across the tip-polymer interface. We have several indications that the heat transfer between tip and sample plays a crucial role, one of them being the asymptotic heater temperature for long writing times, which according to the graph in Fig. 31.19 is approximately 200 °C. The exact temperature of the polymer is unknown. However, the polymer temperature should approach the glass-transition temperature (around 120 °C for PMMA) asymptotically. Hence, the temperature drop between heater and polymer medium is substantial. Part of the temperature difference is due to a temperature drop along the tip, which according to heat-flow simulations [31.20, 21] is expected to be on the order of 30 °C at most. Therefore, a significant temperature gradient must exist in the tip-polymer contact zone. Further experiments on the heat transfer from tip to surface are needed to clarify this point.

We also find that the heat transfer for a nonspherical tip is anisotropic. As shown in Fig. 31.20, in the case of a pyramid-shaped tip, the indentation not only exhibits sharp edges, but also the region around the indentation, where polymer material is piled up, is anisotropic. The pile-up characteristics will be discussed in detail below. At this point we take it as an indication of the relevance of the heat transfer to the measurements.

One of the most striking conclusions of our model of the indentation-writing process is that it should in principle work for most polymer materials. The general behavior of the mechanical properties as a function of temperature and frequency is similar for all polymers [31.41]. The glass-transition temperature Θ_g would then be one of the main parameters that determines the threshold writing temperature.

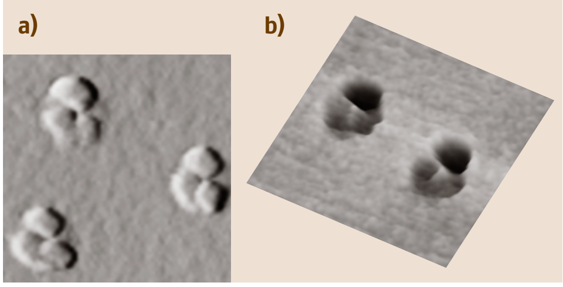

Fig. 31.20a,b Topographic image of individual indentations. **(a)** The region around the actual indentations clearly shows the threefold symmetry of the tip, here a three-sided pyramid. **(b)** The indentations themselves exhibit sharp edges, as can be seen from the inverted 3-D image. Image size is $2 \times 2\,\mu\mathrm{m}^2$ (from [31.15] © 2002 IEEE)

A verification of this was found experimentally by comparing various polymer films. The samples were prepared in the same way as the PMMA samples discussed above [31.18]: by spin-casting thin films (10–30 nm) onto a silicon wafer with a photo-resist buffer. Then, threshold measurements were done by applying heat pulses with increasing current (or temperature) to the tip while the load and the heating time were held constant (load about 10 nN and heating time 10 µs). Examples of such measurements are shown in Fig. 31.21, where the increasing size and depth of indentations can be seen for different heater temperatures. A threshold can be defined based on such data and compared with the glass-transition temperature of these materials. The results show a clear correlation between the threshold heater temperature and the glass-transition temperature (see Fig. 31.22).

With our simple viscoelastic model of writing we are able to formulate a set of requirements that potential candidate materials for millipede data storage have to fulfill. First, the material should ideally exhibit a well-defined glass-transition point with a large drop of the shear modulus at Θ_g. Second, a rather high value of Θ_g (on the order of 150 °C) is preferred to facilitate thermal read back of the data without destroying the information. We have investigated a number of materials to explore the Θ_g parameter space. The fact that all polymer types tested are suitable for forming small indentations leaves us free to choose which polymer type to optimize in terms of the technical requirements for a device such as lifetime of indentations, polymer endurance of the read and write process, power consumption, etc. These are subjects of ongoing research.

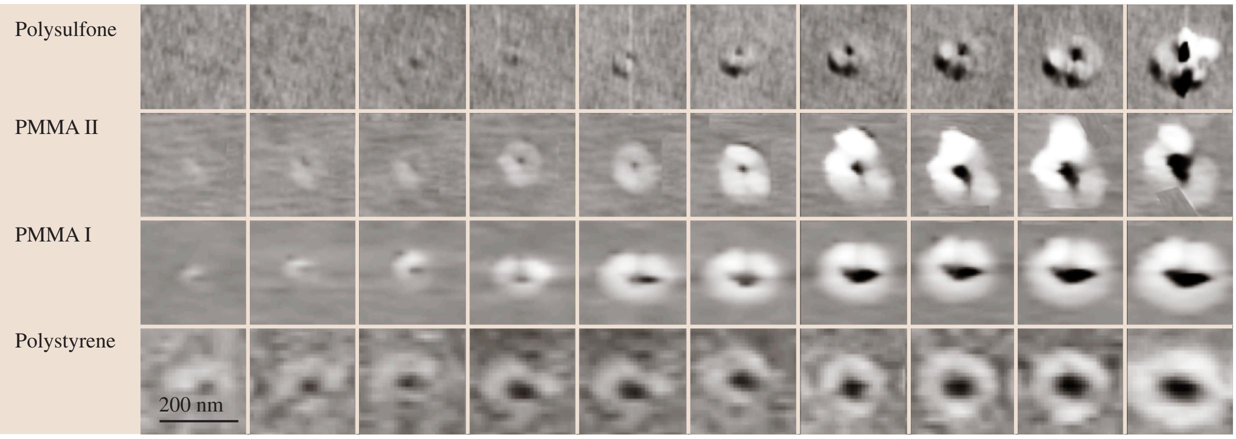

Fig. 31.21 Written indentations for different polymer materials. The heating pulse length was 10 µs, the load about 10 nN. The grey scale is the same for all images. The heater temperatures for the indentation on the left-hand side are 445, 400, 365, and 275 °C for the polymers Polysulfone, PMMA II (anionically polymerized PMMA, $M \approx 26\,\mathrm{k}$), PMMA I (Polymer Standard Service, Germany, $M \approx 500\,\mathrm{k}$), and Polystyrene, respectively. The temperature increase between events on the horizontal axis is 14, 22, 20, and 9 °C, respectively (from [31.15] © 2002 IEEE)

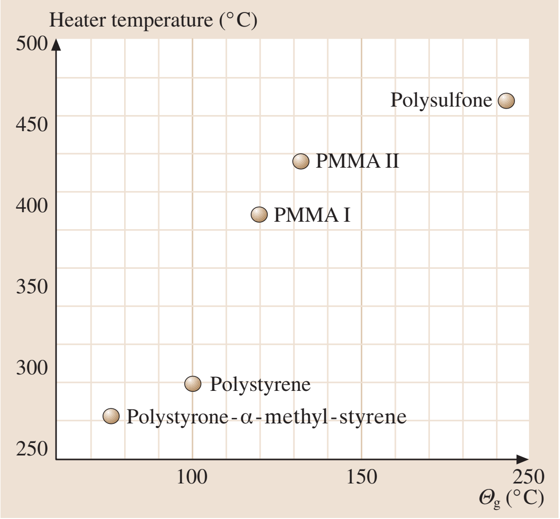

Fig. 31.22 The heater temperature threshold for writing indentations with the same parameters as in Fig. 31.21 is plotted against the glass-transition temperature for these polymers, including poly-α-methyl-styrene (after [31.15])

31.7.2 Erasing Mechanism

It is worthwhile to look at the detailed shapes of the written indentations. The polymer material around an indentation appears piled up as can be seen, for example, in Fig. 31.20. This is not only material that was pushed aside during indentation formation as a result of volume conservation. Rather, the flash heating by the tip and subsequent rapid cooling result in an increase of the specific volume of the polymer. This phenomenon, that the specific volume of a polymer can be increased by rapidly cooling a sample through the glass transition [31.41], is well-known. Our system allows a cooling time on the order of microseconds, which is much faster than the highest rates that can be achieved with standard polymer-analysis tools. However, a quantitative measurement of the specific volume change cannot be easily performed with our type of experiments. On the other hand, the pile-up effect serves as a convenient threshold thermometer. The outer perimeter of the pile-ups surrounding the indentations corresponds to the Θ_g isotherm, and the temperature in the enclosed area has certainly reached values greater than Θ_g during the indentation process. Based on our viscoelastic model, one would thus conclude that previously written indentations that overlap with the piled up region of a subsequently written indentation should be erased.

That this pile-up effect actually works against the formation of an indentation can clearly be seen in the line scans of a series of indentations written in Polysulfone (Fig. 31.23). Here, the heating of the tip was accompanied by a rather high normal force. The force was high enough to create a small indentation, even if the tip was too cold to modify the polymer (Fig. 31.23a).

Then, with increasing tip heating, the indentations initially fill up in the piled up region (Fig. 31.23b) before they finally become deeper (Fig. 31.23c).

The pile-up phenomenon turns out to be particularly beneficial for data-storage applications. The following example demonstrates the effect. If we look at the sequence of images in Fig. 31.24, taken on a standard PMMA sample, we find that the piled up regions can overlap each other without disturbing the indentations. However, if the piled up region of an individual writing event extends over the indented area of a previously written "1", the depth of the corresponding indentation decreases markedly (Fig. 31.24d). This can be used for erasing written data. On the other hand, if the pitch between two successive indentations is decreased even further, this erasing process will no longer work. Instead, a broader indentation is formed, as shown in Fig. 31.24e. Hence, to exclude mutual interference, the minimum pitch between successive indentations, which we denote by minimum-indentation pitch (BP_{min}), must be larger than the radius of the piled up area around an indentation.

In the example shown in Fig. 31.24, the temperature chosen was so high that the ring around the indentations was very large, whereas the depth of the indentation was limited by the stop layer underneath the PMMA material. Clearly, the temperature was too high here to form small indentations, the minimum pitch of which is around 250 nm. However, by carefully optimizing all parameters it is possible to achieve areal densities of up to 1 Tb/in^2, as demonstrated in Fig. 31.3c.

The new erasing scheme based on this volume effect switches from writing to erasing merely by decreasing the pitch of writing indentations. This can be done in a very controlled fashion, as shown in Fig. 31.25, where individual lines or predefined subareas are erased. Hence, this new erasing scheme can be made to work in a way that is controlled on the scale of individual indentations. Compared with earlier global erasing schemes [31.23], this simplifies erasing significantly.

31.7.3 Overwriting Mechanism

Overwriting data on some part of the storage medium can be achieved by first erasing the entire area and then writing the desired data on the erased surface. Although this process works well, it is time-consuming and dissipates a significant amount of power. In a millipede-based storage device, where data rate and power consumption are at a premium, such a two-step overwriting mech-

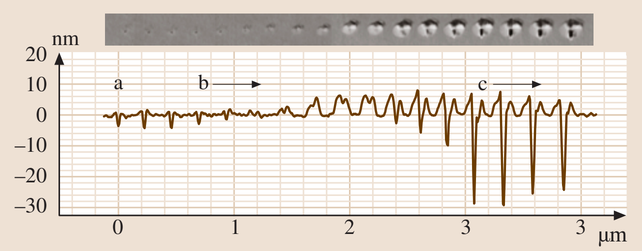

Fig. 31.23 Section through a series of indentations similar to Fig. 31.21. Here, a load of about 200 nN was applied before a heating pulse of 10-µs length was fired. The temperature of the heater at the end of the pulse has been increased from 430 to 610 °C in steps of about 10.6 °C. (a) The load was sufficient to form a plastic indentation even if the polymer was not heated enough to come near the glass transition. (b) With increasing heater temperature, the polymer swells. This eliminates the indentation, thus erasing previously written "cold" marks. (c) As this process continues, the thermomechanical formation of indentations begins to dominate until, finally, normal thermomechanical indentation writing occurs (after [31.15])

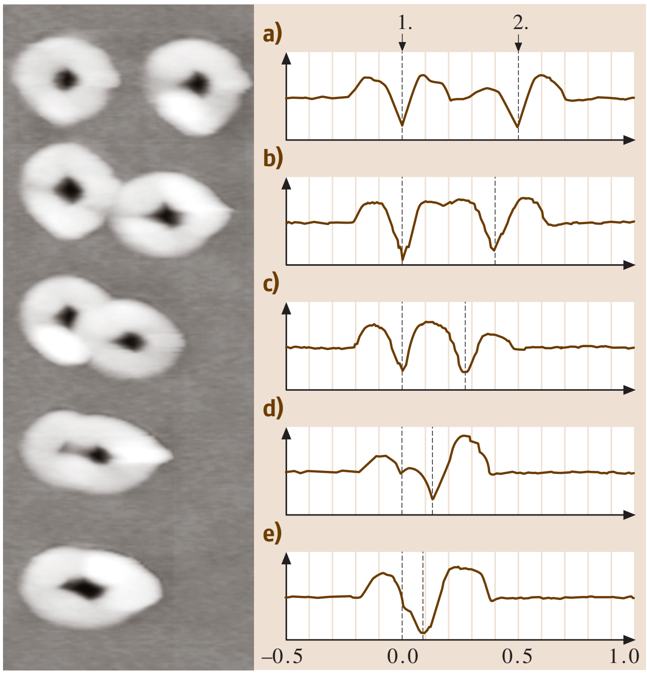

Fig. 31.24a–e Indentations in a PMMA film at several distances. The depth of the indentations is ∼ 15 nm, roughly the same as the thickness of the PMMA layer. The indentations on the left-hand side were written first, then a second series of indentations was made with decreasing distance from the first series going from **(a)** to **(e)** (after [31.15])

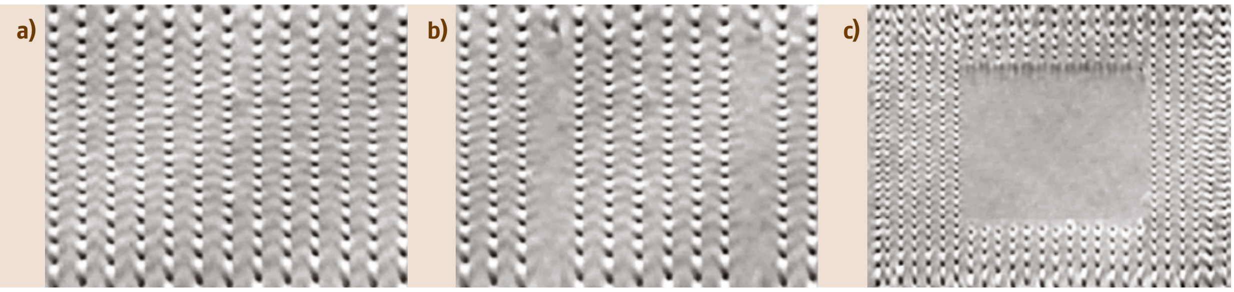

Fig. 31.25a–c Demonstration of the new erasing scheme: **(a)** A bit pattern recorded with variable pitch in the vertical axis (fast scan axis) and constant pitch in the horizontal direction (slow scan axis) was prepared. **(b)** Then two of the lines were erased by decreasing the pitch in the vertical direction by a factor of three, showing that the erasing scheme works for individual lines. One can also erase entire fields of indentations without destroying indentations at the edges of the fields. This is demonstrated in **(c)**, where a field has been erased from an indentation field similar to the one shown in **(a)**. The distance between the lines is 70 nm (from [31.15] © 2002 IEEE)

anism may be impractical. Instead, a one-step, direct overwriting process similar to those applied in magnetic hard-disk drives and rewritable optical drives is desired.

As discussed above, switching from writing to erasing may be achieved by decreasing the pitch of writing. It has been found experimentally that erasing can be performed effectively by halving the pitch of writing successive indentations, which is denoted as BP, provided the condition $\text{BP} \approx \text{BP}_{\text{min}}$ is satisfied. This suggests that the basic distance unit for combined write-erase operations should be BP/2. Written indentations are spaced n units apart, where $n \geq 2$. Let us recall that the presence of an indentation corresponds to a logical bit "1" and the absence of an indentation to a logical bit "0". Logical bits are then stored in the medium at the points of a regular lattice with minimum distance between points equal to BP/2 in the on-track direction, with successive "1"s separated by at least one "0". This condition is necessary in order to avoid mutual interference between successive "1"s. It is also the basis for an important category of codes known as (d, k) codes, which are described in Sect. 31.9.3. Coding can thus be used to enable direct data overwriting in an elegant way. Direct overwriting requires the simultaneous realization of two conditions: If previously written "1"s exist where "0"s are to be written, then these "1"s have to be *erased*. On the other hand, if "0"s exist where "1"s are to be written, then these "1"s have to be *written*. Writing an indentation is performed thermomechanically as described above. Erasing an existing indentation is done by writing another indentation next to it, at a distance of BP/2 units. However, as this operation creates a new indentation shifted by BP/2 with respect to the one erased, the erasing process must be performed repeatedly until the newly created indentation lies at a position corresponding to a "1" in the new data pattern. The basic principle of erasing is illustrated in Fig. 31.26a and b. The figures show how the four bit strings 001, 010, 100, and 101 are modified into the string 010. Figure 31.27 depicts the results of a rewriting experiment; the top track shows a prestored sequence, which is to be overwritten by another sequence, shown on the bottom track for comparison. The result of direct

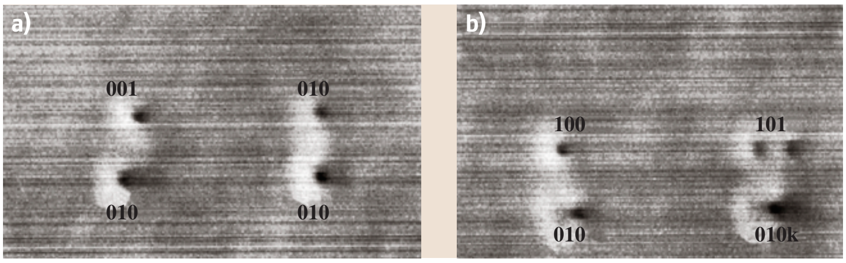

Fig. 31.26a,b Bit strings **(a)** 001 and 010, and **(b)** 100 and 101, overwritten to 010

overwriting of the prestored sequence is shown on the middle track.

Comparison with the sequence on the bottom track, which is written on a clean surface, illustrates the effectiveness of the proposed procedure. Although the write/read quality of overwritten data is somewhat inferior to that of data written on a clean storage surface, detection of the newly written sequence is not affected. However, repeated overwriting may further degrade the quality of stored data. As the extent and rate of degradation are important characteristics of a storage system, this remains an area of ongoing investigation.

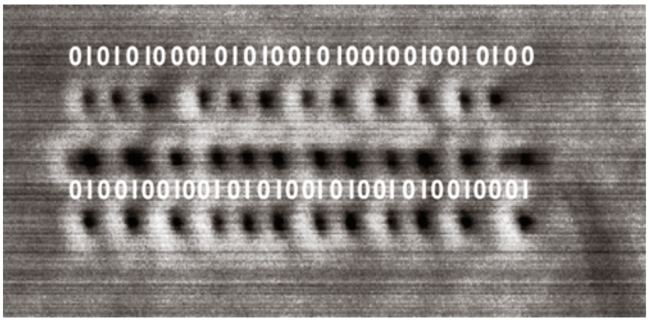

Fig. 31.27 Experimental result of overwriting a bit sequence

31.8 Read Channel Model

Let us now consider the readback channel for a single cantilever that is scanning a storage field where bits are written as indentations or absence of indentations in the storage medium. As discussed above, a cantilever is modeled as a variable resistor that depends on the temperature at the cantilever tip. The model of the read channel, used for the design and analysis of the detection system, is illustrated in Fig. 31.28 [31.12, 26, 27].

To evaluate the evolution of the temperature of a heated cantilever during the read process, we resort to a simple RC-equivalent thermal circuit, illustrated in Fig. 31.29, where $(1+\eta_x)R^{\Theta}$ and C^{Θ} denote the thermal resistance and capacitance, respectively. The parameter $\eta_x = \Delta R^{\Theta}(x)/R^{\Theta}$ indicates the relative variation of thermal resistance that results from the small change in air-gap width between the cantilever and the storage medium. The subscript x indicates the distance in the direction of scanning from the initial point. Therefore, the parameter η_x will assume the largest absolute value when the tip of the cantilever is located at the center of an indentation. The heating power that is dissipated in the cantilever heater region is expressed as

$$P^e[t, \Theta(t, x)] = \frac{V_C^2(t)}{R^e[\Theta(t, x)]} , \qquad (31.2)$$

where $V_C(t)$ is the voltage across the cantilever, $\Theta(t, x)$ is the cantilever temperature, and $R^e[\Theta(t, x)]$ is the temperature-dependent cantilever resistance.

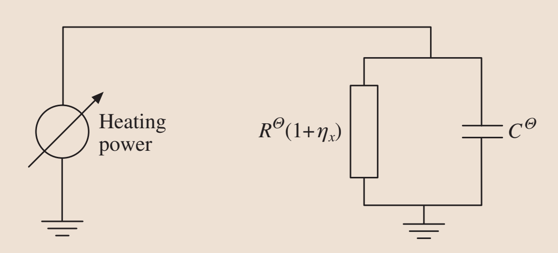

Fig. 31.29 RC-equivalent thermal model of the heat transfer process (after [31.12])

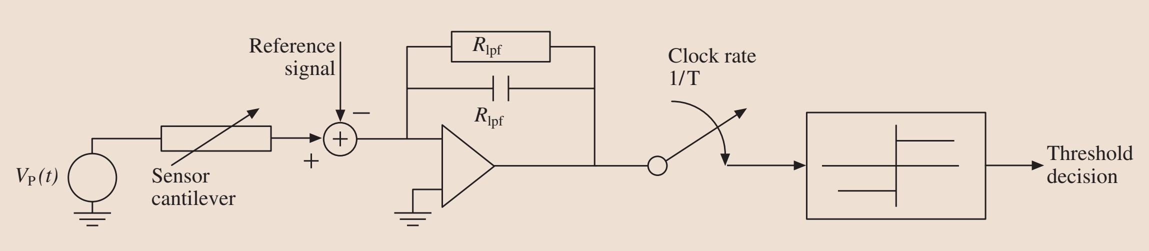

Fig. 31.28 Block diagram of the detection circuit

As the heat-transfer process depends on the value of thermal resistance and on the read pulse waveform, the cantilever temperature $\Theta(t,x)$ depends on time t and distance x. However, because the time it takes for the cantilever to move from the center of one logical mark to the next is greater than the duration of a read pulse, we assume that $\Theta(t,x)$ does not vary significantly as a function of x while a read pulse is being applied, and that it decays to the ambient temperature Θ_0 before the next pulse is applied. Therefore, the evolution of the cantilever temperature in response to a pulse applied at time $t_0 = x_0/v$ at a certain distance x_0 from the initial point of scanning and for a certain constant velocity v of the scanner obeys a differential equation expressed as

$$\begin{aligned}\Theta'(t,x_0) &+ \frac{1}{(1+\eta_{x_0})R^{\Theta}C^{\Theta}}\left[\Theta(t,x_0)-\Theta_0\right]\\ &= \frac{1}{C^{\Theta}}\,\frac{V_{\mathrm{C}}^2(t)}{R^e[\Theta(t,x_0)]}\,,\end{aligned} \tag{31.3}$$

where $\Theta'(t,x_0)$ denotes the derivative of $\Theta(t,x_0)$ with respect to time.

With reference to the block diagram of the read channel illustrated in Fig. 31.28, the source generates the read pulse $V_{\mathrm{P}}(t)$ applied to the cantilever. Clearly, because of the virtual ground at the operational amplifier input, the voltage $V_{\mathrm{C}}(t)$ across the cantilever variable resistance is equal to $\mathrm{V_P}(t)$. Furthermore, the active low-pass RC detector filter, where R_{lpf} and C_{lpf} denote the resistance and capacitance of the low-pass filter, respectively, is realized using an ideal operational amplifier that exhibits infinite input impedance, zero output impedance, and infinite frequency-independent gain. Therefore, the readback signal $V_o(t,x_0)$ obtained at the low-pass filter output in response to the applied voltage $V_{\mathrm{P}}(t) = A\,\mathrm{rect}\,[(t-T_0)/\tau]$, where

$$\mathrm{rect}\left(\frac{t}{\tau}\right) = \begin{cases}1 & \text{if } 0 \le t \le \tau\\ 0 & \text{otherwise}\end{cases}\,, \tag{31.4}$$

and A denotes the pulse amplitude, obeys the differential equation

$$\begin{aligned}V_o'(t,x_0) = \frac{1}{R_{\mathrm{lpf}}C_{\mathrm{lpf}}}\Bigg[&-V_o(t,x_0)\\ &+\frac{R_{\mathrm{lpf}}}{R^e\left[\Theta(t,x_0)\right]}V_P(t)\Bigg]\,.\end{aligned} \tag{31.5}$$

As the voltage at the output of the low-pass filter depends on the value of the variable resistance $R^e[\Theta(t,x_0)]$, the readback signal is determined by solving jointly the differential equations (31.3) and (31.5), with initial conditions $\Theta(t_0,x_0) = \Theta_0$ and $V_o(t_0,x_0) = 0$. For example, a comparison between experimental and synthetic readback signals is shown in Figs. 31.30 and 31.31 for a time constant of the low-pass filter $\tau_{\mathrm{lpf}} = 1.18\,\mu\mathrm{s}$ and two values of the duration of the applied rectangular pulse. For a given cantilever design the function $R^e(\Theta)$ is determined experimentally. Finally, the parameters R^{Θ} and C^{Θ} used in the simple readback channel model are obtained via simulated annealing, where the cost function is given by the mean-square error between experimental and synthetic signals at the low-pass filter output.

Assuming that ideal control of the scanner is performed such that the time of application of a read pulse corresponds either to the cantilever located at the center of an indentation for detecting a "1" bit, or away from an indentation for detecting a "0" bit, two possible responses are obtained at the output of the low-pass filter as solutions of (31.3) and (31.5), which we denote as $V_{o,1}(t,x_0)$ and $V_{o,0}(t,x_0)$, respectively. By sampling the readback signal at the instant $t_s = t_0 + \tau$, simple threshold detection may in principle be applied to detect a written bit, where the value of the threshold is given by

$$V_{\mathrm{Th}} = \frac{1}{2}\left[V_{o,1}(t_s,x_0) + V_{o,0}(t_s,x_0)\right]\,. \tag{31.6}$$

As mentioned in Sect. 31.2, one of the most critical issues in detecting the presence or absence of an indentation is the high resolution required to extract the small signal $V_{o,1}(t_s,x_0) - V_{o,0}(t_s,x_0)$ that contains the information about the bit being "1" or "0" superimposed on the offset signal $V_{o,0}(t,x_0)$. As illustrated in Fig. 31.28, this problem can be solved by subtracting a suitable reference signal $V_{o,\mathrm{ref}}(t,x_0)$ from the readback signal. The readback signal is thus given by

$$\tilde{V}_o(t,x_0) = V_o(t,x_0) - V_{o,\mathrm{ref}}(t,x_0)\,, \tag{31.7}$$

and the threshold is set at $\tilde{V}_{\mathrm{Th}} = \frac{1}{2}\left[V_{o,1}(t_s,x_0) - V_{o,0}(t_s,x_0)\right]$. A VLSI implementation of the detection scheme analyzed here is presented in [31.42].

Now consider read pulses of duration τ that are periodically applied at instants $t_n = nT$, where $1/T$ denotes the symbol rate. Assuming that the response of the previous pulse has vanished as a new pulse is applied and that the temperature of the cantilever has approached the ambient temperature, i. e., $V_o(t_n,x_n) = 0$ and $\Theta(t_n,x_n) = \Theta_0$, then the analysis presented above still holds. In particular, the readback signal samples ob-

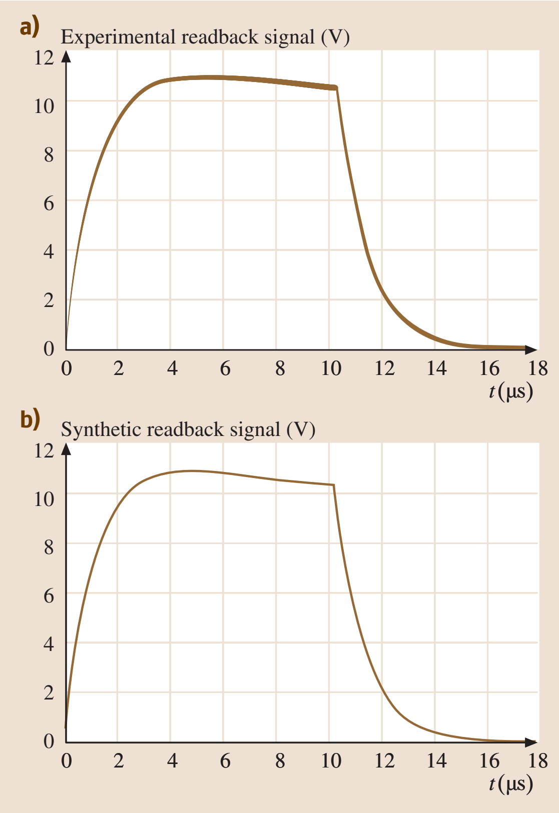

Fig. 31.30 **(a)** Experimental and **(b)** synthetic readback signal for $\tau = 10.25\,\mu\text{s}$ (after [31.12])

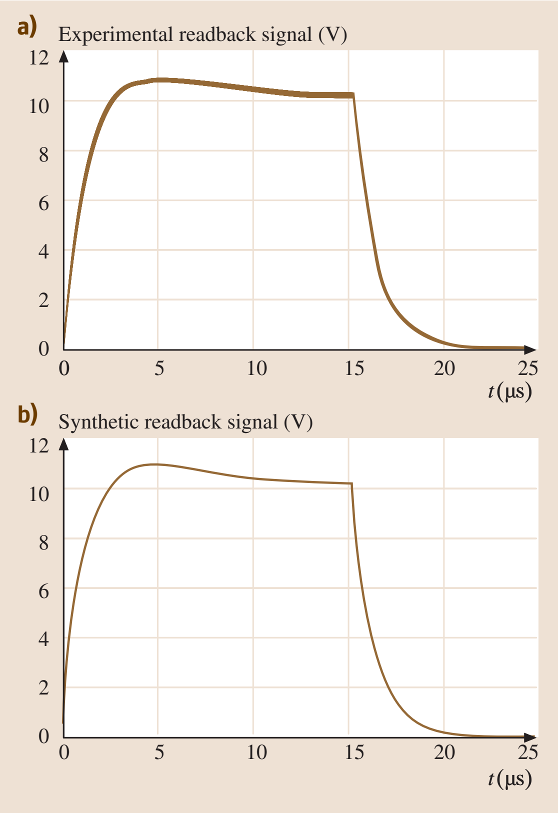

Fig. 31.31 **(a)** Experimental and **(b)** synthetic readback signal for $\tau = 15.25\,\mu\text{s}$ (after [31.12])

tained in response to N pulses applied to the cantilever for detecting a sequence of N binary written symbols are expressed as

$$s(t_{s,n}) = \tilde{V}_o(t_{s,n}, x_n), \quad t_{s,n} = nT + \tau, \quad n = 0, \ldots, N-1, \tag{31.8}$$

where $\tilde{V}_o(t, x_n)$ is given by (31.7) for pulses applied at time t_n and at distance $x_n = nT\upsilon$, $n = 0, \ldots, N-1$ from the initial point of scanning. Note that the functions $V_o(t, x_n)$ and $V_{o,\text{ref}}(t, x_n)$ in (31.7) are ideally given by the solution of the differential equations (31.3) and (31.5) for $\eta_x = \Delta R^{\Theta}(x_n)/R^{\Theta}$ and $\eta_x = 0$, respectively.

The readback signal (31.7) at the output of the low-pass filter is observed in the presence of additive noise. Therefore, the readback signal for the detection of the n-th binary symbol is given by

$$r(t_{s,n}) = s(t_{s,n}) + w(t_{s,n}), \tag{31.9}$$

where $w(t)$ denotes the noise signal. The components of the noise signal that must be taken into account are thermal noise (Johnson’s noise) from the sensor, which reaches a temperature of about $\Theta_1 = 350\,°\text{C}$ during the read process, as well as from front-end analog circuitry. However, note that besides thermal noise, medium-related noise also affects the overall system performance.

Based on the above analysis, the response to a pulse applied to the cantilever at a distance x from some initial point can be calculated given the parameter $\eta_x = \Delta R^{\Theta}(x)/R^{\Theta}$. Recall that the value of η_x is proportional to the distance of the cantilever from the storage medium at the current location of the tip. Therefore, during tip displacement due to scanner motion, η_x is modulated from the topographical features of the storage surface such as written indentations, rings, and dust particles. This indicates that the modeling of the readback signal is a two-step process. First,

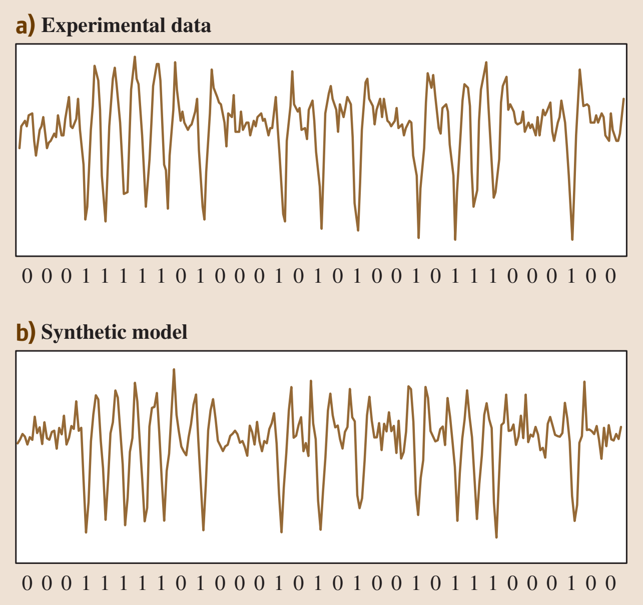

Fig. 31.32a,b Comparison between **(a)** the readback signal obtained experimentally along a data track and **(b)** the readback signal obtained by the synthetic model (after [31.12])

a model for the storage surface topography is developed, which directlyand then the above procedure is used to calculate the readback signal samples in response to pulses applied at selected points in the particular storage area.

In the absence of any imperfections during the manufacturing and the writing process, the storage surface would consist of completely flat regions interrupted by uniformly shaped indentations, possibly surrounded by polymer rings. A 1-D cross section of an indentation along the scanning direction is modeled by a function with one main lobe and two side lobes, one on each side of the main lobe, the magnitude and the extent of which can be varied independently. The side lobes are of opposite sign than the main lobe and simulate the polymer rings around written indentations. By varying their magnitude and extent while keeping the total extent of the pulse fixed, one can simulate indentations/rings of varying width and asymmetric ring formation, phenomena that are caused by different recording conditions. In practice, however, no polymer surface is entirely flat and indentation shapes are far from uniform. The deviation of indentations from uniformity is simulated by scaling the amplitude of each pulse shape by a random number drawn from a Gaussian distribution with unit mean and adjustable variance. The surface roughness is in turn simulated by adding white Gaussian noise to the height of every point in the area of interest. Note that surface roughness is a medium-related effect and manifests itself in the readback signal as a noise process, which is, however, of a very different nature than thermal noise. The advantage of the adopted two-step model is that it naturally decouples these unrelated noise sources, as well as the write and the read processes.

Figure 31.32 illustrates the experimental and synthetic readback signals obtained along a data track. The waveforms shown in Fig. 31.32 have been obtained by applying pulses at the oversampling rate of q/T, where q denotes the oversampling factor. For a more detailed comparison between model and actual signals, Figs. 31.33 and 31.34 illustrate 3-D views of isolated indentations from experimental and synthetic readback

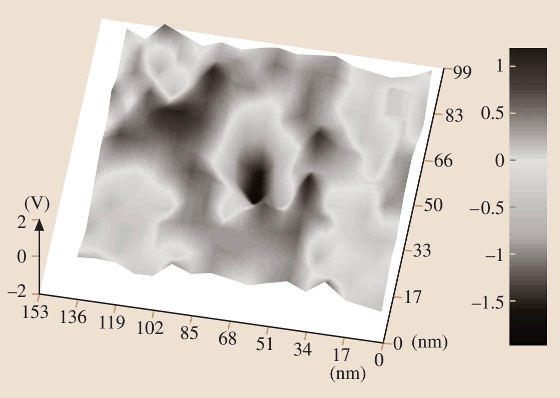

Fig. 31.33 Three-dimensional view of an isolated indentation obtained experimentally

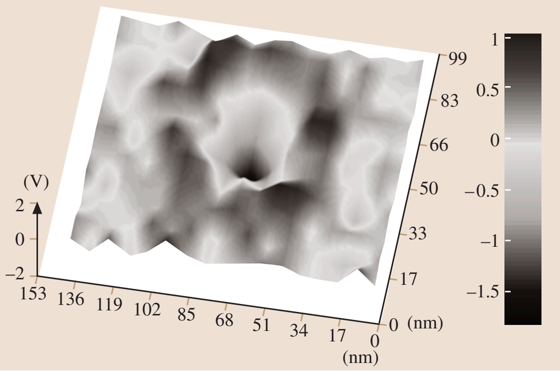

Fig. 31.34 Three-dimensional view of an isolated indentation obtained by the synthetic model

signals, respectively. The dark regions in the center of both figures correspond to the indentation centers, whereas dark regions around them are due to rings. Note also the irregular height of the surrounding surface, which is attributed to the roughness of the storage medium.

31.9 System Aspects

In this section, we describe various aspects of a storage system based on our millipede concept. Each cantilever can write data to and read data from a dedicated area of the polymer surface, called a *storage field*. As mentioned above, in each storage field the presence (absence) of an indentation corresponds to a logical "1" ("0"). All indentations are nominally of equal depth and size. The logical marks are placed at a fixed horizontal distance from each other along a data track. We refer to this distance, measured from one logical mark center to the next, as the *bit pitch* (BP). The vertical (cross-track) distance between logical mark centers, the *track pitch* (TP), is also fixed. To read and write data the polymer medium is moved under the (stationary) cantilever array at a constant velocity.

A robust way to achieve synchronization and servo control in an x/y-actuated large 2-D array is by reserving a small number of storage fields exclusively for timing recovery and servo-control purposes, as illustrated in Fig. 31.35. Because of the large number of levers in the millipede, this solution is advantageous in terms of overhead compared with the alternative of timing and servo information being embedded in all data fields.

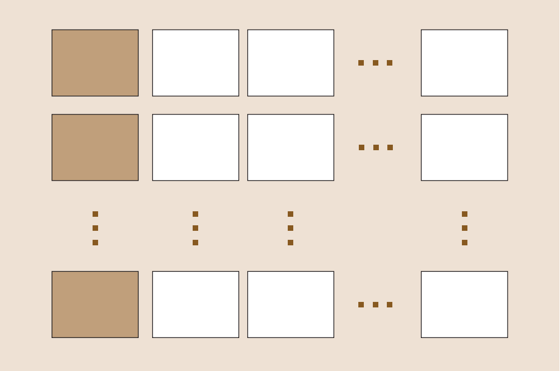

Fig. 31.35 Layout of data and servo/timing fields. *Dark boxes* represent dedicated servo/timing fields, *white boxes* represent data fields

31.9.1 PES Generation for the Servo Loop

With logical marks as densely spaced as in the millipede, accurate track following becomes a critical issue. Track following means controlling the position of each tip such that the tip is always positioned over the center of a desired track during reading. During writing, the tip position should be such that the written marks are aligned in a predefined way. In electromechanical systems, track following is performed in a servo loop, which is driven by an appropriate error signal called position-error signal (PES). Ideally, its magnitude is a direct estimate of the vertical (cross-track) distance of the tip from the closest track centerline, and its polarity indicates the direction of this offset.

Several approaches exist to generate a PES for AFM-based storage devices [31.9]. However, based on the results reported, none of these methods can achieve the track-following accuracy required for the millipede system. The quality of the PES directly affects the stability and robustness of the associated tracking servo loop [31.43].

We describe a method for generating a uniquely decodable PES for the millipede system [31.12, 27]. The method is based on the concept of *bursts* that are vertically displaced with respect to each other, arranged in such a way as to produce two signals in quadrature, which can be combined to provide a robust PES. This concept is borrowed from magnetic recording [31.43]. However, servo marks, as opposed to magnetic transitions, are placed in bursts labeled A and B for the in-phase signal and C and D for the quadrature signal. The centers of servo marks in burst B are vertically offset from mark centers in burst A by d' units of length. This amount of vertical spacing is related to the diameter of the written marks. The same principle applies to marks in the quadrature bursts C and D, with the additional condition that mark centers in burst C are offset by $d'/2$ units from mark centers in A in the cross-track direction. The latter condition is required in order to generate a quadrature signal. The configuration of servo bursts is illustrated in Fig. 31.36 for a case where TP $= 3d'/2$. Al-

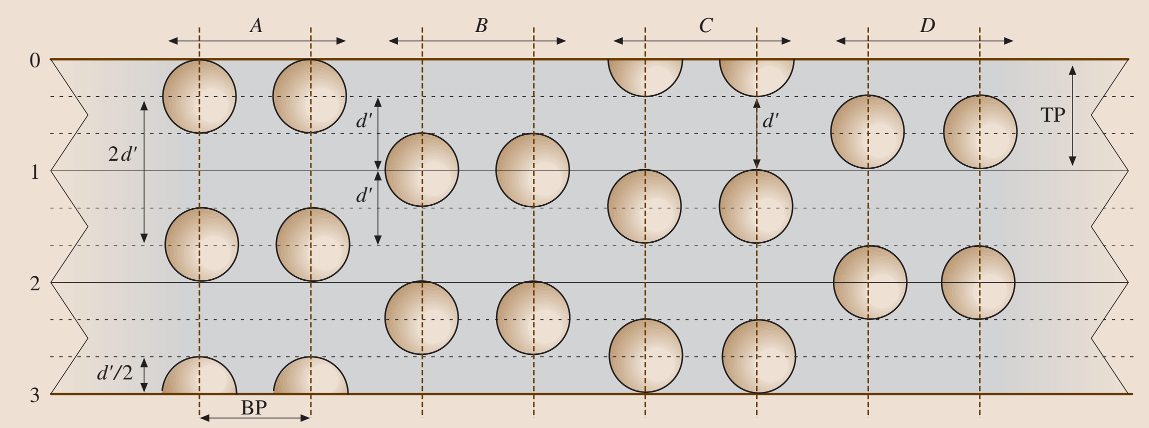

Fig. 31.36 Servo burst configuration (after [31.12])

though each burst typically consists of many marks to enable averaging of the corresponding readout signals, only two marks per burst are shown here to simplify the presentation. The solid horizontal lines depict track centerlines, and circles represent written marks, which are modeled here as perfect conical indentations on the polymer storage surface.

To illustrate the principle of PES generation, let us assume that marks in all bursts are spaced BP units apart in the longitudinal direction, and that sampling occurs exactly at mark centers, so that timing is perfect. The assumption of perfect timing is made only for the purpose of illustration. In actual operation, sampling is performed with the aid of a timing recovery loop, as described in Sect. 31.9.2. Referring to Fig. 31.36, let us further assume that the cantilever/tip is located on the line labeled "0" and moves vertically toward line "3" in a line crossing the centers of the left-most marks in burst A (shown as a brown dashed line). The tip moves from the edge of the top mark toward its center, then toward its bottom edge, then to a blank space, again to a mark, and so on. The readout signal magnitude decreases linearly with the distance from the mark center and reaches a constant, background level value at a distance greater than the mark radius from the mark center, according to the adopted (conical mark) model. To synthesize the in-phase signal, the readout signal is also captured as the tip (conceptually) moves in a vertical line crossing the mark centers of burst B (brown dashed line in Fig. 31.36). The in-phase signal is then formed as the difference $\bar{A}-\bar{B}$, where $\bar{A}$ and $\bar{B}$ stand for the measured signal amplitudes in bursts A and B, respectively. This signal is represented by the line labeled "I" in Fig. 31.37. It has zero-crossings at integer multiples of d', which do not generally correspond to track centers because we set $\mathrm{TP}=3d'/2$ in this example. Therefore, the I-signal is not a valid PES in itself. This is why the quadrature (Q) signal becomes necessary in this case. The Q-signal is generated from the servo readback signals of bursts C and D as $\bar{C}-\bar{D}$ and is also shown in Fig. 31.37 (Q-curve). Note that it exhibits zero-crossings at points where the I-signal has local extrema.

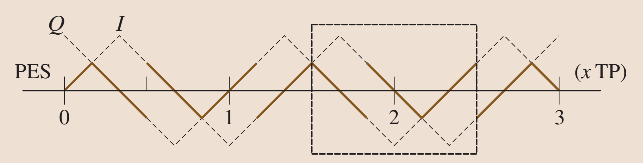

Fig. 31.37 Ideal position-error signal (after [31.12])

A certain combination of the two signals (I and Q), shown as solid lines in Fig. 31.37, has zero-crossings at all track center locations and constant (absolute) slope, which qualifies it as a valid PES. However, this PES exhibits zero-crossings at all integer multiples of $d'/2$. For our example of $\mathrm{TP}=3d'/2$, three such zero-crossings exist in an area of width equal to TP around any track centerline. This fact, however, does not hamper unique position decoding. At even-numbered tracks, it is the zero of the *in-phase* signal that indicates the track center. The zeros of the quadrature signal, in turn, can be uniquely mapped into a position estimate by examining the polarity of the in-phase signal at the corresponding positions. This holds for any value of the combined PES within an area of width equal to TP around each current track centerline. The signals exchange roles for odd-numbered tracks. The current track number, which is known a priori from the seek operation, is used to determine the mode of operation for the position demodulation procedure.

The principle of PES generation based on servo marks has been verified experimentally. For this purpose, A, B, C, and D bursts were written by an AFM cantilever/tip on an appropriate polymer medium consisting of a polymer coating on a silicon substrate. The bit pitch was set to 42 nm, and the track pitch was taken to be approximately equal to d', the cross-track distance between A (C) and B (D) bursts. An image created by reading the written pattern with the same cantilever is shown in Fig. 31.38. Shaded areas indicate indentations. The readout signal from the cantilever was also used for servo demodulation, as described above. The resulting in-phase and quadrature signals are shown in Fig. 31.39. The track centerlines are indicated by vertical dotted lines in the graph.

It can be observed that the zero-crossings of the in-phase signal are closely aligned with the track centerlines, as well as with the minima and maxima of the quadrature signal, as required for unique position decoding across all possible cross-track positions, at least in cases where $\mathrm{TP} \neq d'$. Moreover, the PES slope is nearly linear along a cross-track width of one track pitch around each track center, as $\mathrm{TP} \approx d'$ in this case, although deviations from the ideal signal shape exist. These deviations occur mainly because written indentations do not have perfect conical shapes, and also because of medium noise due to the roughness of the recording medium. Nevertheless, the experimentally generated error signals indicate that the proposed concept is valid and promising. Specifically, the results indicate that servo self-writing is feasible, that servo demodulation is almost identical to data readout and can be performed by any cantilever without special provisions, and that the PES generated closely approximates the desirable features described above.

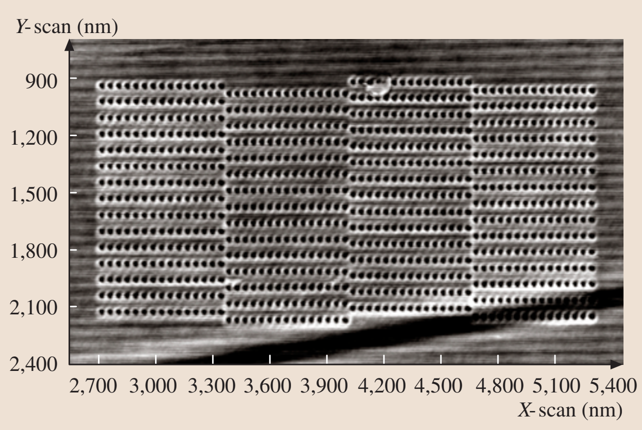

Fig. 31.38 Experimental A, B, C, and D servo bursts (BP = 42 nm) (after [31.12] © 2003 IEEE)

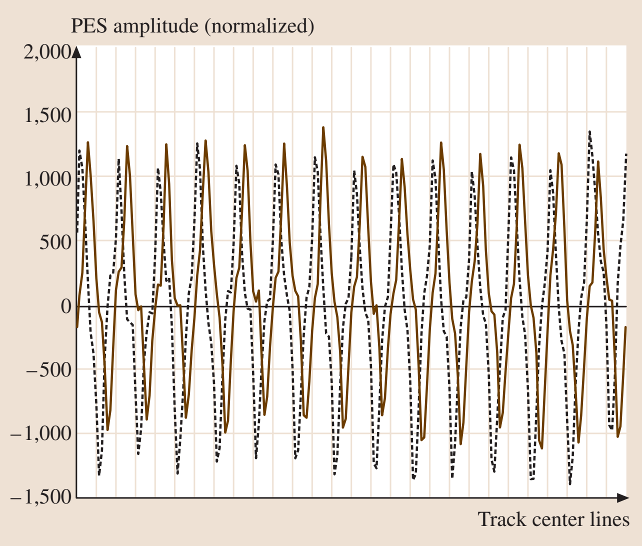

Fig. 31.39 Demodulated in-phase (*solid line*) and quadrature (*dashed line*) PES based on the servo burst of Fig. 31.38 (after [31.12])

31.9.2 Timing Recovery

Similar to obtaining servo information based on using dedicated servo fields, we employ separate dedicated clock fields for recovery of timing information [31.12, 26, 27]. The concept is to have continuous access to a pilot signal for synchronization after initial phase acquisition and gain estimation. The recovered clock is then distributed to all remaining storage fields to allow reliable detection of random data. Initial phase acquisition is obtained by a robust correlation algorithm, gain estimation is based on averaging of the readback signal obtained from a predefined stored pattern, and tracking of the optimum sampling phase is achieved by a second-order digital loop.

At the beginning of the read process, several signal parameters have to be estimated prior to data detection. Besides the clock phase and frequency, it is necessary to estimate the gain of the overall read channel. To solve the problem of initial estimation of signal parameters prior to data detection, the sequence written in the clock field consists of a preamble, followed by a pattern of all "1"s for tracking the optimum sampling phase during the detection of random data. The transition between the preamble and the pattern of all "1"s must be detected reliably, as it indicates the start of data records to the remaining storage fields. Assuming that the initial frequency offset is within a small, predetermined range, usually 1,000 parts-per-million (PPM), we distinguish the tasks needed for timing recovery as follows: (i) acquisition of the optimum sampling phase; (ii) estimation of the overall channel gain needed for threshold detection; (iii) detection of the transition between the preamble and the pattern of all "1"s; and (iv) tracking of the optimum sampling phase.

At the beginning of the acquisition process, an estimate of the optimum sampling phase is obtained by resorting to a correlation method. We rely on the knowledge of the preamble and of an ideal reference-channel impulse response, which closely resembles the actual impulse response (see Sect. 31.9). The channel output samples obtained at the oversampling rate q/T are first processed by removing the dc-offset, then averaging is performed to reduce the noise level, and finally the resulting sequence is correlated with the reference impulse response to determine the phase estimate.

After determining the estimate of the optimum sampling phase, an estimate of the overall channel gain is obtained by averaging the amplitude of the channel output samples at the optimum sampling instants. The gain estimate is obtained from an initial segment of the preamble corresponding to an "all 1" binary pattern. As mentioned above, it is necessary that the end of the preamble is indicated by a "sync" pattern, which marks the transition between acquisition mode and tracking mode. Detection of the sync pattern is also based on a robust correlation method. After the sync pattern, an "all 1" pattern, as in the case of robust phase acquisition and gain estimation, is employed for tracking. The "all 1" pattern corresponds to regularly spaced indentations, which convey reliable timing information.

Tracking of the optimum sampling phase is achieved by the second-order loop configuration shown in Fig. 31.40. Assuming data detection is performed at instants that correspond to integer multiples of the oversampling factor q, the deviation of the sampling phase from the optimum sampling phase is estimated as

$$\Delta\tau_n = r(t_{s,nq+1}) - r(t_{s,nq-1}) \,. \tag{31.10}$$

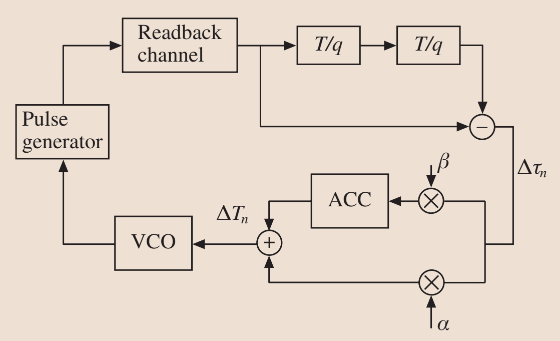

Fig. 31.40 Second-order loop for tracking the optimum sampling phase (after [31.12])

This estimate of the phase deviation is input to a second-order loop filter, which provides an output given by

$$\Delta T_n = u_n + \alpha\,\Delta\tau_n \,, \tag{31.11}$$

where the discrete-time integrator is recursively updated as

$$u_{n+1} = u_n + \beta\,\Delta\tau_n \,. \tag{31.12}$$

The loop-filter output then determines the control signal for a voltage-controlled oscillator (VCO).

Note that a similar concept for timing recovery can also be applied if no separate clock field is available. In this case, the timing information is extracted from the random user data on each storage field.

31.9.3 Considerations on Capacity and Data Rate

The ultimate locality provided by nanometer-sharp tips represents the pathway to the high areal densities that will be needed in the foreseeable future. The intrinsic nonlinear interactions between closely spaced indentations, however, may limit the minimum distance between successive indentations and, hence, the areal density. The storage capacity of a millipede-based storage device can be further increased by applying modulation or constrained codes [31.12].

With modulation coding, a desired constraint is imposed on the data-input sequence, so that the encoded data stream satisfies certain properties in the time or frequency domain. These codes are very important in digital recording devices and have become ubiquitous in all data-storage applications. The particular class

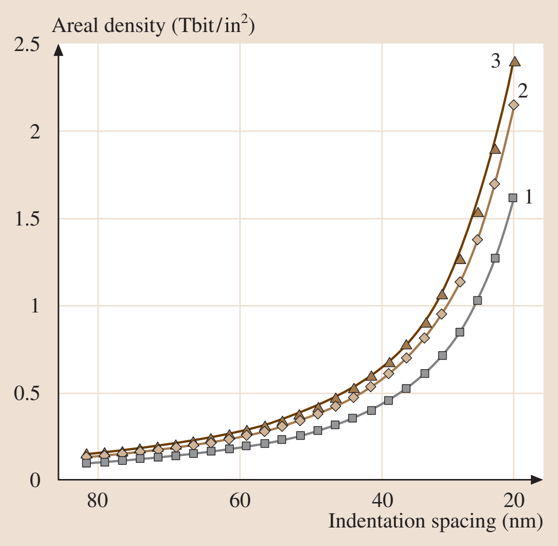

Fig. 31.41 Areal density versus indentation spacing. Curve 1: $d = 0$; curve 2: $d = 1$; and curve 3: $d = 2$

of codes that imposes restrictions on the number of consecutive "1"s and "0"s in the encoded data sequence, generally known as run-length-limited (RLL) (d, k) codes [31.44], can be used to facilitate overwriting and also increase the effective areal density of a millipede-based storage device. The code parameters d and k are nonnegative integers with $k > d$, where d and k indicate the minimum and maximum number of "0"s between two successive "1"s, respectively. In the past, the precoded (RLL) (d, k) codes were mainly used for spreading the magnetic transitions further apart via the d-constraint, thereby minimizing intersymbol interference and nonlinear distortion, and for preventing loss of clock synchronization via the k-constraint. In optical recording, precoded RLL codes are primarily used for increasing the shortest pit length in order to improve the reliability of bit detection, as well as for limiting the number of identical symbols, so that useful timing information can be extracted from the readback signal.

For the millipede application, where dedicated clock fields are used, the k-constraint does not really play an important role and, therefore, can, in principle, be set to infinity, thereby facilitating the code design process. In a precoderless RLL code design, where the presence or absence of an indentation represents a "1" or "0", respectively, the d-constraint is instrumental in limiting the interference between successive indentations, as well as in increasing the effective areal density of the storage device. In particular, the quantity $(d+1)R$, where R denotes the rate of the (d, k) code, is a direct measure of the increase in linear recording density. Clearly, the packing density can be made larger by increasing d. On the other hand, large values of d lead to codes with very low rate, which implies high recording symbol rates, thus rendering these codes impractical for storage systems that are limited by the clock speed. The choice of $d = 1$ and $k \geq 6$ guarantees the existence of a code with rate $R = 2/3$. Use of $(d = 1, k \geq 6)$ modulation coding reduces the bit distance by half while maintaining the pitch between "1"s constant, thereby

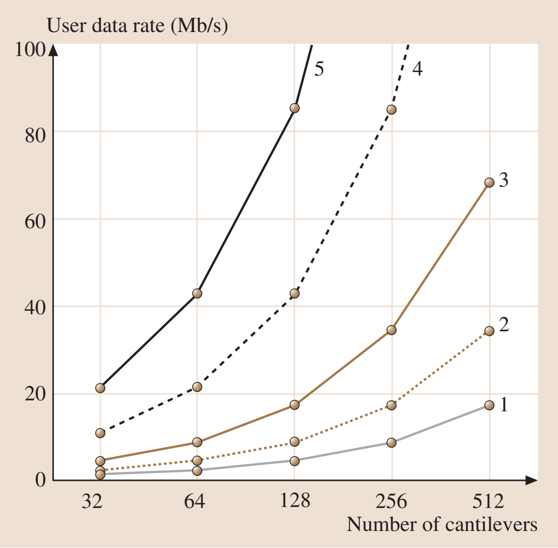

Fig. 31.42 User data rate versus number of active cantilevers for the $(d = 1, k \geq 6)$ coding scheme. Curve 1: $T = 20\,\mu\text{s}$; curve 2: $T = 10\,\mu\text{s}$; curve 3: $T = 5\,\mu\text{s}$; curve 4: $T = 2\,\mu\text{s}$; and curve 5: $T = 1\,\mu\text{s}$ (after [31.12])

Table 31.1 Areal density and storage capacity (after [31.12] with permission © 2003 IEEE)

Coding	Linear density (kb/in)	Track density (kt/in)	Areal density (Gb/in^2)	Capacity (Gb)
Uncoded	847	847	717	1.21
$(d = 1, k \geq 6)$	1,129	847	956	1.61
$(d = 2, k > 6)$	1,269	847	1,075	1.81

increasing the linear density by a factor of 4/3. Similarly, the choice of $d = 2$ and $k > 6$ guarantees the existence of a code with rate $R = 1/2$. Use of ($d = 2$, $k > 6$) modulation coding reduces the bit distance to a third while maintaining the pitch between "1"s constant, thereby increasing the linear density by a factor of 3/2. Figure 31.41 shows the areal density as a function of the indentation spacing for an uncoded system, as well as for systems coded with ($d = 1$, $k \geq 6$) and ($d = 2$, $k > 6$), where coding is applied only in the on-track direction.

Table 31.1 shows the achievable areal densities and storage capacities for a 32×32 cantilever array with 1,024 storage fields, each having an area of $100 \times 100\,\mu\text{m}^2$, resulting in a total storage area of $3.2 \times 3.2\,\text{mm}^2$. The indentation pitch and the track pitch are set equal to 30 nm. Finally, for the computation of the storage capacity an overall efficiency of 85% has been assumed, taking into account the redundancy of the outer error-correction coding, as well as the presence of dedicated servo and clock fields.

Figure 31.42 shows the user data rate as a function of the total number of cantilevers accessed simultaneously, for various symbol rates and a ($d = 1$, $k \geq 6$) modulation coding scheme. For example, for a 32×32 cantilever array, a system designed to access a maximum of 256 cantilevers every $T = 5\,\mu\text{s}$ provides a user data rate of 34.1 Mb/s. Alternatively, by resorting to the row/column multiplexing scheme with $T = 80\,\mu\text{s}$ a data rate of 8.5 Mb/s is achieved.

31.10 Conclusions

A very large 2-D array of AFM probes has been operated for the first time in a multiplexed/parallel fashion, and write/read/erase operations in a thin polymer medium have been successfully demonstrated at densities of or significantly higher than those achieved with current magnetic storage systems.

The millipede has the potential to achieve ultrahigh storage areal densities on the order of 1 Tb/in^2 or higher. The high areal storage density, small form factor, and low power consumption render the millipede concept a very attractive candidate as a future storage technology for mobile applications, as it offers several Gigabytes of capacity at data rates of several Megabytes per second. Dedicated servo and timing fields allow reliable system operation with a very small overhead. The read channel model provides the methodology for analyzing system performance and assessing various aspects of the detection and servo/timing algorithms that are key to achieving the system reliability required by the applications envisaged.

Although the first high-density storage operations with the largest 2-D AFM array chip ever built have been demonstrated, there are a number of issues that need further investigation such as overall system reliability, including long-term stability of written indentations, tip and medium wear, limits of data rates, array and cantilever size, as well as trade-offs between data rate and power consumption.

References

31.1 E. Grochowski, R. F. Hoyt: Future trends in hard disk drives, IEEE Trans. Magn. **32** (1996) 1850–1854

31.2 D. A. Thompson, J. S. Best: The future of magnetic data storage technology, IBM J. Res. Dev. **44** (2000) 311–322

31.3 G. Binnig, H. Rohrer, C. Gerber, E. Weibel: 7×7 reconstruction on Si(111) resolved in real space, Phys. Rev. Lett. **50** (1983) 120–123

31.4 G. Binnig, C. F. Quate, C. Gerber: Atomic force microscope, Phys. Rev. Lett. **56** (1986) 930–933

31.5 H. J. Mamin, D. Rugar: Thermomechanical writing with an atomic force microscope tip, Appl. Phys. Lett. **61** (1992) 1003–1005

31.6 R. P. Ried, H. J. Mamin, B. D. Terris, L. S. Fan, D. Rugar: 6-MHz 2-N/m piezoresistive atomic-force-microscope cantilevers with incisive tips, J. Microelectromech. Syst. **6** (1997) 294–302

31.7 B. D. Terris, S. A. Rishton, H. J. Mamin, R. P. Ried, D. Rugar: Atomic force microscope-based data storage: Track servo and wear study, Appl. Phys. A **66** (1998) S809–S813

31.8 H. J. Mamin, B. D. Terris, L. S. Fan, S. Hoen, R. C. Barrett, D. Rugar: High-density data storage using proximal probe techniques, IBM J. Res. Dev. **39** (1995) 681–699

31.9 H. J. Mamin, R. P. Ried, B. D. Terris, D. Rugar: High-density data storage based on the atomic force microscope, Proc. IEEE **87** (1999) 1014–1027

31.10 L. R. Carley, J. A. Bain, G. K. Fedder, D. W. Greve, D. F. Guillou, M. S. C. Lu, T. Mukherjee, S. Santhanam, L. Abelmann, S. Min: Single-chip computers with microelectromechanical systems-based magnetic memory, J. Appl. Phys. **87** (2000) 6680–6685

31.11 G. Gibson, T. I. Kamins, M. S. Keshner, S. L. Neberhuis, C. M. Perlov, C. C. Yang: Ultra-high density storage device, (1996) US Patent 5,557,596

31.12 E. Eleftheriou, T. Antonakopoulos, G. K. Binnig, G. Cherubini, M. Despont, A. Dholakia, U. Dürig, M. A. Lantz, H. Pozidis, H. E. Rothuizen, P. Vettiger: millipede – A MEMS-based scanning-probe data-storage system, IEEE Trans. Magn. **39** (2003) 938–945

31.13 G. K. Binnig, H. Rohrer, P. Vettiger: Mass-storage applications of local probe arrays, (1998) US Patent 5,835,477

31.14 P. Vettiger, J. Brugger, M. Despont, U. Drechsler, U. Dürig, W. Häberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig: Ultrahigh density, high-data-rate NEMS-based AFM data storage system, J. Microelectron. Eng. **46** (1999) 11–17

31.15 P. Vettiger, G. Cross, M. Despont, U. Drechsler, U. Dürig, B. Gotsmann, W. Häberle, M. A. Lantz, H. E. Rothuizen, R. Stutz, G. K. Binnig: The "millipede" – Nanotechnology entering data storage, IEEE Trans. Nanotechnol. **1** (2002) 39–55

31.16 M. Lutwyche, C. Andreoli, G. Binnig, J. Brugger, U. Drechsler, W. Häberle, H. Rohrer, H. Rothuizen, P. Vettiger: Microfabrication and parallel operation of 5 × 5 2D AFM cantilever array for data storage and imaging, Proc. IEEE 11th Int. Workshop MEMS, Heidelberg 1998 (IEEE, Piscataway 1998) 8–11

31.17 M. Lutwyche, C. Andreoli, G. Binnig, J. Brugger, U. Drechsler, W. Häberle, H. Rohrer, H. Rothuizen, P. Vettiger, G. Yaralioglu, C. Quate: 5 × 5 2D AFM cantilever arrays: A first step towards a terabit storage device, Sens. Actuators A **73** (1999) 89–94

31.18 P. Vettiger, M. Despont, U. Drechsler, U. Dürig, W. Häberle, M. I. Lutwyche, H. E. Rothuizen, R. Stutz, R. Widmer, G. K. Binnig: The "millipede" – More than one thousand tips for future AFM data storage, IBM J. Res. Dev. **44** (2000) 323–340

31.19 B. W. Chui, H. J. Mamin, B. D. Terris, D. Rugar, K. E. Goodson, T. W. Kenny: Micromachined heaters with 1-μs thermal time constants for AFM thermomechanical data storage, Proc. IEEE Transducers, Chicago 1997 (IEEE, Piscataway 1997) 1085–1088

31.20 W. P. King, J. G. Santiago, T. W. Kenny, K. E. Goodson: Modelling and prediction of sub-micrometer heat transfer during thermomechanical data storage, 1999 Microelectromechanical Systems (MEMS). Proc. ASME Intl. Mechanical Engineering Congress and Exposition, ed. by A. P. Lee, L. Lin, F. K. Forster, Y. C. Young, K. Goodson, R. S. Keynton (ASME, New York 1999) 583–588

31.21 W. P. King, T. W. Kenny, K. E. Goodson, G. L. W. Cross, M. Despont, U. Dürig, H. Rothuizen, G. Binnig, P. Vettiger: Design of atomic force microscope cantilevers for combined thermomechanical writing and thermal reading in array operation, J. Microelectromech. Syst. **11** (2002) 765–774

31.22 G. K. Binnig, M. Despont, W. Häberle, P. Vettiger: Method of forming ultrasmall structures and apparatus therefore, (March 1999) US Patent Office, Application No. 147865

31.23 G. Binnig, M. Despont, U. Drechsler, W. Häberle, M. Lutwyche, P. Vettiger, H. J. Mamin, B. W. Chui, T. W. Kenny: Ultra high-density AFM data storage with erase capability, Appl. Phys. Lett. **74** (1999) 1329–1331

31.24 G. K. Binnig, J. Brugger, W. Häberle, P. Vettiger: Investigation and/or manipulation device, (March 1999) US Patent Office, Application No. 147867

31.25 S. M. Sze: *Physics of Semiconductors Devices* (Wiley, New York 1981)

31.26 G. Cherubini, T. Antonakopoulos, P. Bächtold, G. K. Binnig, M. Despont, U. Drechsler, A. Dholakia, U. Dürig, E. Eleftheriou, B. Gotsmann, W. Häberle, M. A. Lantz, T. Loeliger, H. Pozidis, H. E. Rothuizen, R. Stutz, P. Vettiger: The millipede, a very dense, highly parallel scanning-probe data-storage system, ESSCIRC – Proceedings 28th European Solid-State Circuits Conference, ed. by A. Baschirotto, P. Malcovati (Univ. Bologna, Bologna 2002) 121–125

31.27 E. Eleftheriou, T. Antonakopoulos, G. K. Binnig, G. Cherubini, M. Despont, A. Dholakia, U. Dürig, M. A. Lantz, H. Pozidis, H. E. Rothuizen, P. Vettiger: "millipede": A MEMS-based scanning-probe data-storage system, Digest of the Asia-Pacific Magnetic Recording Conference 2002, APMRC '02 (IEEE, Piscataway 2002) CE-2-1–CE2-2

31.28 M. Despont, J. Brugger, U. Drechsler, U. Dürig, W. Häberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig, H. Rohrer, P. Vettiger: VLSI-NEMS chip for AFM data storage, Technical Digest 12th IEEE Int. Micro Electro Mechanical Systems Conf. "MEMS '99" (IEEE, Piscataway 1999) 564–569

31.29 T. S. Ravi, R. B. Marcus: Oxidation sharpening of silicon tips, J. Vac. Sci. Technol. B **9** (1991) 2733–2737

31.30 M. Despont, J. Brugger, U. Drechsler, U. Dürig, W. Häberle, M. Lutwyche, H. Rothuizen, R. Stutz, R. Widmer, G. Binnig, H. Rohrer, P. Vettiger: VLSI-NEMS chip for parallel AFM data storage, Sens. Actuators A **80** (2000) 100–107

31.31 H. Rothuizen, M. Despont, U. Drechsler, G. Genolet, W. Häberle, M. Lutwyche, R. Stutz, P. Vettiger: Compact copper/epoxy-based micromachined electromagnetic scanner for scanning probe applications, Technical Digest, 15th IEEE Int. Conf. on Micro Electro Mechanical Systems "MEMS 2002" (IEEE, Piscataway 2002) 582–585

31.32 S. C. Minne, G. Yaralioglu, S. R. Manalis, J. D. Adams, A. Atalar, C. F. Quate: Automated parallel high-speed atomic force microscopy, Appl. Phys. Lett. **72** (1998) 2340–2342

31.33 M. Lutwyche, U. Drechsler, W. Häberle, R. Widmer, H. Rothuizen, P. Vettiger, J. Thaysen: Planar micromagnetic x/y/z scanner with five degrees of freedom. In: *Magnetic Materials, Processes, and Devices: Applications to Storage and Micromechanical Systems (MEMS)*, Vol. 98-20, ed. by L. Romankiw, S. Krongelb, C. H. Ahn (Electrochemical Society, Pennington 1999) pp. 423–433

31.34 H. Rothuizen, U. Drechsler, G. Genolet, W. Häberle, M. Lutwyche, R. Stutz, R. Widmer, P. Vettiger: Fabrication of a micromachined magnetic x/y/z scanner for parallel scanning probe applications, Microelectron. Eng. **53** (2000) 509–512

31.35 J.-J. Choi, H. Park, K. Y. Kim, J. U. Jeon: Electromagnetic micro x-y stage for probe-based data storage, J. Semicond. Technol. Sci. **1** (2001) 84–93

31.36 H. Lorenz, M. Despont, N. Fahrni, J. Brugger, P. Vettiger, P. Renaud: High-aspect-ratio, ultrathick, negative-tone near-UV photoresist and its applications for MEMS, Sens. Actuators A **64** (1998) 33–39

31.37 C. Q. Davis, D. Freeman: Using a light microscope to measure motions with nanometer accuracy, Opt. Eng. **37** (1998) 1299–1304

31.38 M. I. Lutwyche, M. Despont, U. Drechsler, U. Dürig, W. Häberle, H. Rothuizen, R. Stutz, R. Widmer, G. K. Binnig, P. Vettiger: Highly parallel data storage system based on scanning probe arrays, Appl. Phys. Lett. **77** (2000) 3299–3301

31.39 K. Fuchs, C. Friedrich, J. Weese: Viscoelastic properties of narrow-distribution poly(methyl metacrylates), Macromolecules **29** (1996) 5893–5901

31.40 U. Dürig, B. Gotsman: This estimate is based on a fluid dynamic deformation model of a thin film, private communication

31.41 J. D. Ferry: *Viscoelastic Properties of Polymers, 3rd edition* (Wiley, New York 1980)

31.42 T. Loeliger, P. Bächtold, G. K. Binnig, G. Cherubini, U. Dürig, E. Eleftheriou, P. Vettiger, M. Uster, H. Jäckel: CMOS sensor array with cell-level analog-to-digital conversion for local probe data storage, ESSCIRC – Proceedings 28th European Solid-State Circuits Conference, ed. by A. Baschirotto, P. Malcovati (Univ. Bologna, Bologna 2002) 623–626

31.43 A. H. Sacks: Position signal generation in magnetic disk drives. Ph.D. Thesis (Carnegie Mellon University, Pittsburgh 1995)

31.44 K. A. S. Immink: *Coding Techniques for Digital Recorders* (Prentice Hall, Hemel 1991)

32. Microactuators for Dual-Stage Servo Systems in Magnetic Disk Files

This chapter discusses the design and fabrication of electrostatic MEMS microactuators and the design of dual-stage servo systems in disk drives. It introduces fundamental requirements of disk drive servo systems, along with challenges posed by storage densities increases. It describes three potential dual-stage configurations and focuses on actuated slider assemblies using electrostatic MEMS microactuators. The authors discuss major electrostatic actuator design issues, such as linear versus rotary motion, electrostatic array configuration, and differential operation. Capacitive and piezoresistive elements may be used to sense relative slider position, while integrated gimbals and structural isolation may prove useful in improving performance. A detailed design example based on a translational, micromolded actuator illustrates several of these concepts and is accompanied by theoretical and experimental results.

The chapter continues to discuss MEMS microactuator fabrication. It describes several processes for obtaining appropriate electrostatic devices including micromolding, deep-reactive ion etching, and electroplating. The primary goal of these processes is to obtain very high-aspect ratio structures, which improve both actuation force and structural robustness. Other fabrication issues, such as electrical interconnect formation, material selection, and processing cost, are also considered. Actuated slider fabrication is compared to that of actuated suspension and actuated head assemblies; this includes instrumentation of a suspension with strain sensors to aid in vibration detection.

The section on controller design reviews dual-stage servo control design architectures and methodologies. The considerations for controller design of MEMS microactuator dual-stage servo systems are discussed. The details of control designs using a decoupled SISO design method and the robust multivariable design method of μ-synthesis are also presented. In the decoupled design, a self-tuning control algorithm has been developed to compensate for the variations in the microactuator's resonance mode. In the μ-synthesis design, robust controllers can be synthesized by using additive and parametric uncertainties to characterize the un-modeled dynamics of the VCM and the variations in the microactuator's resonance mode. Finally, the chapter also introduces a dual-stage short-span seek control scheme based on decoupled feed forward reference trajectories generations.

32.1 **Design of the Electrostatic Microactuator** ... 952
  32.1.1 Disk Drive Structural Requirements ... 952
  32.1.2 Dual-Stage Servo Configurations ... 953
  32.1.3 Electrostatic Microactuators: Comb-Drives vs. Parallel-Plates ... 954
  32.1.4 Position Sensing ... 956
  32.1.5 Electrostatic Microactuator Designs for Disk Drives ... 958

32.2 **Fabrication** ... 962
  32.2.1 Basic Requirements ... 962
  32.2.2 Electrostatic Microactuator Fabrication Example ... 962
  32.2.3 Electrostatic Microactuator Example Two ... 963
  32.2.4 Other Fabrication Processes ... 966
  32.2.5 Suspension-Level Fabrication Processes ... 967
  32.2.6 Actuated Head Fabrication ... 968

32.3 **Servo Control Design of MEMS Microactuator Dual-Stage Servo Systems** ... 968
  32.3.1 Introduction to Disk Drive Servo Control ... 969
  32.3.2 Overview of Dual-Stage Servo Control Design Methodologies ... 969
  32.3.3 Track-Following Controller Design for a MEMS Microactuator Dual-Stage Servo System ... 971
  32.3.4 Dual-Stage Seek Control Design ... 976

32.4 **Conclusions and Outlook** ... 978

References ... 979

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

This chapter discusses the design and fabrication of electrostatic MEMS microactuators, and the design of dual-stage servo systems, in disk drives. The focus of the chapter is an actuated slider assembly using an electrostatic MEMS microactuator. We discuss major design issues, including linear versus rotary actuation, electrostatic array configuration, and integrated sensing capability. We describe several fabrication processes for obtaining the necessary devices, such as micromolding, deep-reactive ion etching, and electroplating. Dual-stage servo control design architectures and methodologies are then reviewed . We present in detail track-following controller designs based on a sensitivity function decoupling single-input-single-output design methodology and the robust μ-synthesis design methodology. Finally, we introduce a 2-DOF short span seek control design using a dual-stage actuator.

Since the first hard disk drive (HDD) was invented in the 1950s by IBM, disk drives' storage density has been following Moore's law, doubling roughly every 18 months. The current storage density is 10 million times larger than that of the first HDD [32.1]. Historically, increases in storage density have been achieved by almost equal increases in track density, the number of tracks encircling the disk, and bit density, the number of bits in each track. However, because of superparamagnetism limitations, future areal storage density increases in HDDs are predicted to be achieved mainly through an increase in track density [32.2].

Research in the HDD industry is now targeting an areal density of one terabit per square inch. For a predicted bit aspect ratio of 4 : 1, this translates to a linear bit density of 2 M bits per inch (BPI) and a radial track density of 500 K tracks per inch (TPI), which in turn implies a track width of 50 nm. A simple rule of thumb for servo design in HDDs is that three times the statistical standard deviation of the position error between the head and the center of the data track should be less than 1/10 of the track width. Thus, to achieve such a storage density, nanometer-level precision of the servo system will be required.

A disk drive stores data as magnetic patterns, forming bits, on one or more disks. The polarity of each bit is detected (read) or set (written) by an electromagnetic device known as the read/write head. The job of a disk drive's servo system is to position the read/write head over the bits to be read or written as they spin by on the disk. In a conventional disk drive, this is done by sweeping over the disk or disks a long arm consisting of a voice coil motor (VCM), an E-block, suspensions, and sliders, as shown in Figs. 32.1 and 32.2. A read/write head is fabricated on the edge of each slider (one for each disk surface). Each slider is supported by a suspension and flys over the surface of a disk on an air-bearing. The VCM actuates the suspensions and sliders about a pivot in the center of the E-block. We describe this operation in more detail in the following section.

The key to increasing HDD servo precision is to increase servo control bandwidth. However, the bandwidth of a traditional single-stage servo system as shown in Fig. 32.2 is limited by the multiple mechanical resonance modes of the pivot, the E-block, and the suspension between the VCM and the head. Nonlinear friction of the pivot bearing also limits achievable servo precision. Dual-stage actuation, with a second stage actuator placed between the VCM and the head, has been proposed as a solution that would increase servo bandwidth and precision.

Several different secondary actuation forces and configurations have been proposed, each having strengths and weaknesses given the requirements of HDDs. The dual-stage configurations can be categorized into three groups: actuated suspension, actuated slider, and actuated head. Within these, actuation forces include piezoelectric, electrostatic, and electromagnetic. In this chapter, we discuss design, fabrication and control of an electrostatic MEMS microactuator (MA) for actuated slider dual-stage positioning.

32.1 Design of the Electrostatic Microactuator

The servo system of a hard disk drive is the mechatronic device that locates and reads data on the disk. In essence, it is a large arm that sweeps across the surface of the disk. At the end of the arm is the read-write head, containing the magnetic reading and writing elements that transfer information to and from the disk.

32.1.1 Disk Drive Structural Requirements

This read-write head is contained in a box-like structure known as a slider. The slider has a contoured lower surface that acts as an air bearing between the head and the disk. The high-velocity airflow generated by the spinning disk pushes up on the air-bearing surface, maintaining the slider and read-write head at a constant distance from the disk, despite unevenness of the disk, permitting reliable data reading and writing.

The arm over an HDD's disk has three primary stages: the voice-coil motor (VCM), the E-block, and the suspension. In a conventional disk drive, the VCM performs all positioning of the head, swinging it back and forth across the disk. The E-block lies between the VCM and the suspension and contains the pivot point. The suspension projects from the E-block over the disk as a thin flexible structure, generally narrowing to a point at the location of the slider.

For a disk drive servo to operate effectively, it must maintain the read-write head at a precise height above the disk surface and within a narrow range between the disk tracks arranged in concentric circles around the disk. It must also be able to seek from one track on the disk to another. Information about the track that the head is following is encoded in sectors radiating from center of the disk, allowing the head to identify its position and distance from the center of the track. To maintain a correct flying height, the suspension must be designed with an appropriate stiffness in the vertical direction to balance an air-bearing force corresponding to the slider design in use in the drive. Meanwhile, the suspension must be flexible to roll and pitch at the slider location to permit adaptation to unevenness of the disk surface. This is accomplished using a gimbal structure. The suspension as a whole, however, should not bend or twist during the operation, as this would misdirect the head away from the track it is following.

As data densities in HDDs increase and track widths diminish, single-stage, conventional servo systems become less able to position the head precisely. Because the VCM/E-block/suspension assembly is large and massive as a unit, the speed at which the head can be controlled is limited. Furthermore, the assembly tends to have a low natural frequency, which can accentuate vibration in the disk drive and cause off-track errors. At track densities in the future approaching one Terabit per square inch, the vibration induced in a disk drive by airflow alone is enough to force the head off-track.

A solution to these problems is to complement the VCM with a smaller, secondary actuator to form a dual-stage servo system. The VCM continues to provide rough positioning, while the second stage actuator does fine positioning and damps out vibration and other disturbances. The smaller second stage actuator can typically be designed to have a much higher natural frequency and less susceptibility to vibration than the VCM. Any actuator used in a dual-stage system should be inexpensive to build, require little power to operate, and preserve the stiffness properties described above necessary to preserve the flying height.

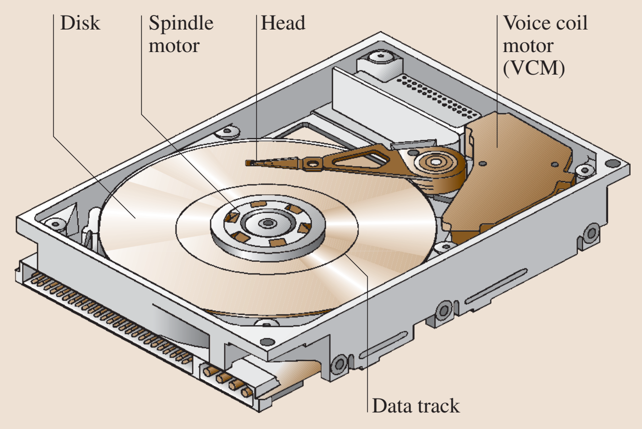

Fig. 32.1 A schematic diagram of an HDD

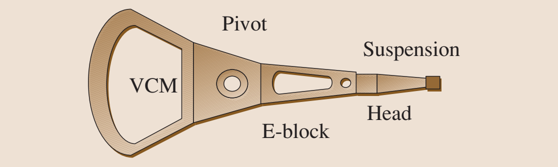

Fig. 32.2 VCM actuator in an HDD

32.1.2 Dual-Stage Servo Configurations

In the past six years, much research work has been dedicated to the exploration of suitable secondary actuators for constructing dual-stage servo systems for HDDs. These dual-stage configurations can be categorized into three groups: "actuated suspension," "actuated slider," and "actuated head."

Actuated Suspension

In this approach, the suspension is redesigned to accommodate an active component, typically a piezoelectric material. This piezoelectric material stretches or flexes the suspension to position the slider and magnetic head. Piezoelectric material is an active actuation element that produces a large actuation force but small actuation stroke. In the "actuated-suspension" configuration, therefore, the piezoelectric actuators are usually implemented in a leverage mechanism that can convert small actuation displacements into large head displacements. Typically this is done by placing the piezoelectric actuators away from the magnetic head (between the E-block and suspension) so that they can have a long leverage arm to gain mechanical amplification and produce a sufficient magnetic-head motion. The advantage of this approach is that the suspension can be fabricated by a conventional suspension-making process, and its dual-stage servo configuration is effective in attaining low frequency runout attenuation in the positioning servo loop. The major drawback of this approach is that the system is still susceptible to instabilities due to the excitation of suspension resonance modes. Thus, track-per-inch (TPI) servo performance can be increased but remains limited when compared to the alternative approaches. The actuated suspension approach, nonetheless is expected to be the first deployed in commercial HDDs [32.3, 4].

Actuated Slider

In this approach, a microactuator is placed between the slider and gimbal to position the slider/magnetic heads. The resulting servo bandwidth can be higher than the previous approach because the secondary actuation bypasses the mechanical resonances of the suspension. This approach uses existing sliders and microactuators that can be batch fabricated, and thus could be cost effective. However, the size and mass of the microactuator are significant relative to those of current sliders and may interfere with the slider flying stability. Current suspensions, therefore, need to be redesigned to adopt this secondary actuator. Suitable driving forces in this approach include electrostatic, electromagnetic, and piezoelectric [32.5–8]. To further reduce the assembly task of placing the microactuator in between gimbal structure and slider, some researchers have proposed microactuators that are either integrated with the gimbal structure [32.6] or the slider [32.9].

Actuated Head

In this approach, the slider is redesigned so that the microactuators can be placed inside the slider block and actuate the magnetic heads with respect to the rest of the slider body. As these microactuators are very small, they only slightly increase the slider weight and are thus capable of working with the current suspension assembly. Researchers have successfully demonstrated the integrated fabrication process for fabricating the electrostatic microactuators and magnetic heads within one piece of ceramic block (slider). The embedded electrostatic microactuator has its resonance close to 30 kHz and was able to position the magnetic heads relative to the rest of the slider body by 0.5 μm [32.10, 11]. Full-fledged integration of slider, actuator, and read/write head remains a challenge.

In this chapter, we focus on actuated slider configurations, as they involve a great deal of interesting microscale engineering. In particular, we will discuss electrostatic actuation, probably the most common method of implementing microactuation in microelectromechanical devices. Any such microactuator will exhibit certain features:

- a fixed base, which attaches to the suspension,
- a movable platform, upon which the slider rests,
- springs between the base and platform, flexible in the direction of desired motion, and stiff in all other directions, and
- an electrostatic actuation array that generates the force used to move the platform and slider.

Microactuators must also include a wiring scheme for transferring signals to and from the slider and often incorporate a structure for sensing the motion of the slider relative to the suspension. Electrostatic microactuators to be discussed in this chapter include HexSil and DRIE fabricated actuators from the University of California, Berkeley, and electroplating-formed actuators by IBM and the University of Tokyo.

Electromagnetic or piezoelectric force are alternatives to electrostatic actuation in the actuated-slider configuration. Electromagnetic microactuators use ferromagnetic films to produce force perpendicular to an applied electric field. This type of actuation has potentially low voltage requirements but requires special fabrication techniques to integrate the magnetic components into the assembly. A microactuator of this type for hard disk drives is under development at Seagate, with results as yet unpublished. Piezoelectric microactuators use a piezoelectric material, which expands or contracts in response to applied voltage, to move the slider. These actuators have simple fabrication, the patterning a piece of piezoelectric material to sit between the suspension and slider, but it is difficult to obtain an adequate range of motion. A short stroke from the piezoelectric piece must be leveraged into a much larger motion at the read/write head. A piezoelectric microactuator has been produced by TDK corporation with a 0.5 μm stroke length at 10 V, with a 10 V bias [32.12].

32.1.3 Electrostatic Microactuators: Comb-Drives vs. Parallel-Plates

Electrostatic microactuators have been studied as the secondary actuators in HDDs for their relative ease of fabrication, particularly in the configurations of "actuated slider" and "actuated head," since the structural material needs only to be conductive rather than ferromagnetic or piezoelectric. Electrostatic force is generated by applying a voltage difference between the moving shuttle and a fixed stator element. Depending on the designated motion for the shuttle, electrostatic actuators are often categorized into two groups: comb-drives and parallel-plates, as illustrated in Fig. 32.3.

The magnitude of the electrostatic force generated equals the rate of change of energy that is retained within the finger-like structure and varied by shuttle motion. Therefore, the electrostatic force for comb-drive actuators, in which the designated shuttle motion moves along the x-direction, as shown in Fig. 32.3, equals

$$F_{\text{comb}} = \frac{\partial E}{\partial x} = \frac{\epsilon h}{2d} V^2 , \tag{32.1}$$

where ϵ is the permittivity of air, x is the overlap between two adjacent plates, h is plate thickness, and d is the gap between two parallel plates. Similarly, the electrostatic force for parallel-plates actuators is

$$F_{\text{parallel}} = \frac{\partial E}{\partial y} = \frac{\epsilon x h}{2d^2} V^2 , \tag{32.2}$$

where x is the finger overlap.

As indicated in (32.1) and (32.2), the electrostatic force for comb-drives actuators does not depend on the displacement of the moving shuttle and thus allows a long stroke while maintaining a constant electrostatic force. The electrostatic force for parallel-plates actuators, in contrast, is a nonlinear function of its shuttle motion ($\propto 1/d^2$), and the maximum stroke is limited by the nominal gap between shuttle and stator. A longer stroke is achieved with a larger gap, at the expense of lower electrostatic force. For applications that require small stroke but large force output, parallel-plates actuators are preferred since the output force from parallel-plates can be x/d times larger than the force from comb-drives. The following equation (32.3) is easily derived from (32.1) and (32.2).

$$\frac{F_{\text{parallel}}}{F_{\text{comb}}} = \frac{x}{d} . \tag{32.3}$$

A simplified second-order differential equation is often utilized to describe the dynamic response of an electrostatic microactuator

$$m\ddot{x}(t) + b\dot{x}(t) + K_{\text{m}} x(t) = F\,[V, x(t)] , \tag{32.4}$$

where m is the mass of the moving shuttle, b is the damping coefficient of the microactuator, K_{m} is the spring constant of the mechanical spring that connects the moving shuttle to an anchor point, and F is the electrostatic force that can be obtained from (32.1) and (32.2).

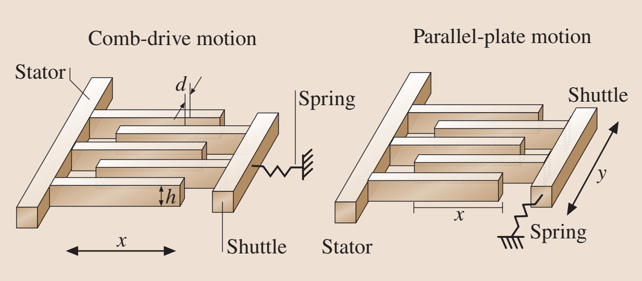

Fig. 32.3 Electrostatic microactuators. Comb-drives versus parallel-plates

Differential Drives

Because electrostatic force is always attractive, electrostatic microactuators need other features to actively control the direction of shuttle motion in servo applications, as opposed to relying on the restoring force from a mechanical spring. For this reason, the differential-drives approach, as shown in Fig. 32.4, is frequently adopted in electrostatic microactuator designs. Based on the differential-drive configuration, the simplified second-order differential equation (32.4) is rewritten as

$$\begin{aligned} m\ddot{x}(t)+b\dot{x}(t)+K_{\text{m}}x(t) = F\,[&V_{\text{bias}}+V_{\text{dr}}, x_o-x(t)]- \\ &F\,[V_{\text{bias}}-V_{\text{dr}}, x_o+x(t)] , \end{aligned} \tag{32.5}$$

where x_0 is the nominal position of the moving shuttle. If the differential drive is operated at the bias voltage (V_{bias}) with a small perturbation voltage (V_{dr}), the nonlinear force input in (32.5) can be linearized with a first-order

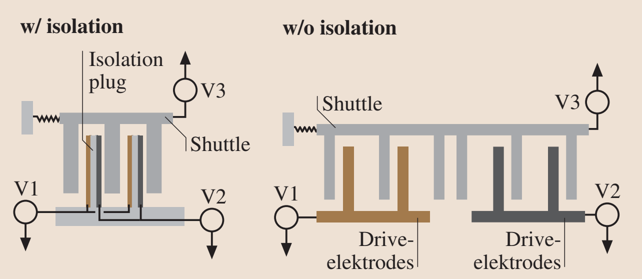

Fig. 32.4 Differential parallel-plates actuators and electrical-isolation features

approximation.

$$\begin{aligned} & m\ddot{x}(t)+b\dot{x}(t)+(K_{\mathrm{m}}-K_{\mathrm{e}})x(t)=K_v V_{\mathrm{dr}}\ , \\ & K_{\mathrm{e}}=2\frac{\partial F}{\partial x}\,|_{V_{\mathrm{bias}}}\ , \\ & K_{\mathrm{v}}=2\frac{\partial F}{\partial V_{\mathrm{dr}}}\,|_{V_{\mathrm{bias}}}\ . \end{aligned} \tag{32.6}$$

Here K_{e} represents a softening electrostatic spring constant, and K_{v} represents the voltage-to-force gain. The electrostatic spring constant acts as a negative spring during the electrostatic microactuator operation, and its value varies with the bias voltage (V_{bias}). When the electrostatic spring constant K_{e} exceeds the spring constant K_{m} from the mechanical spring, the microactuator becomes unstable; this is often described as pull-in instability. As shown in (32.5) and (32.6), the differential configuration cancels the even order harmonics in voltage and thus linearizes the voltage–force relation to some extent. Furthermore, in parallel-plates actuators, the differential configuration reduces the nonlinearity in actuation voltage as well as in shuttle displacement.

Electrical Isolation

Because electrostatic actuation requires multiple voltage levels for actuation force and position sensing (as discussed in the following section), electrical isolation is another challenge for designing an electrostatic microactuator. Generally speaking, when multiple voltage levels are needed in MEMS devices, electrical isolation is achieved by breaking up the parts that need to be on different voltage level and anchoring them separately to a nonconductive substrate. This approach has many drawbacks, not only because it requires a substrate in a device but also because structures have to be mechanically separated to be electrically isolated. The electrical isolation problem is far more severe in parallel-plates microactuators than comb-drive microactuators since parallel-plates actuation generally requires different voltage levels for stator fingers pulling in opposite directions.

Figure 32.4 shows an example of how an electrical-isolation feature can be utilized to increase the actuation force output in a differential parallel-plates microactuator design. As shown in the figure, without the proper electrical isolation, drive-electrodes with different voltage potentials have to be placed in separate groups and result in the same voltage difference between drive-electrodes and shuttle on both sides of each shuttle finger [32.5, 6]. Since electrostatic forces are always attractive, gaps on two sides of the interlaced structure cannot be made equal, otherwise the forces on two sides of a shuttle finger will be equal and the shuttle's movement direction will be uncontrollable. With such electrical-isolation features as the "isolation plug," shown on the left in Fig. 32.4, gaps of the interlaced structure can be the same width, since different voltages can be applied on the two sides of the shuttle fingers. As shown in Fig. 32.4, the design with integrated electrical-isolation features is more compact than without isolation features. Consequently, more finger structures can fit in the same amount of space, and the actuation voltage can be reduced.

32.1.4 Position Sensing

Most proposed HDD dual-stage servo controllers utilize only the position of the magnetic head relative to the center of data track, known in the industry as the position error signal, or PES, for closed-loop track following control. These systems have a single-input-multi-output (SIMO) control architecture. In some instances, however, it is also possible to measure the relative position error signal (RPES) of the magnetic head relative to the VCM. In this case, the control architecture is multi-input-multi-output (MIMO). As shown in [32.13], RPES can be used in a MIMO controller to damp out the second stage actuator's resonance mode and enhance the overall robustness of the servo system.

Capacitive position sensing and piezoresistive position sensing are two popular sensing mechanisms among electrostatic microactuator designs. Each of these sensing mechanisms is discussed in more detail in the following sections.

Capacitive Position Sensing

Capacitive position sensing is based on shuttle movement causing a capacitance change between the moving shuttle and fixed stators. By measuring the change in capacitance, it is possible to determine the shuttle location relative to the fixed stator. The output voltage (V_{o}) for both differential drives in Fig. 32.5 equals $2\delta C/C_{\mathrm{i}}V_{\mathrm{s}}$. The capacitance change due to shuttle movement ($\mathrm{d}C/\mathrm{d}x$) can be derived and the output voltage (V_{o}) for the comb-drives and parallel-plates can be formulated as a function of shuttle displacement.

$$\begin{aligned} V_{\mathrm{comb}} &= 2\frac{C_{\mathrm{s}}}{C_{\mathrm{i}}}\frac{\delta x}{x_0}V_{\mathrm{s}}\ , \\ V_{\mathrm{parallel}} &= 2\frac{C_{\mathrm{s}}}{C_{\mathrm{i}}}\frac{y_0\delta y}{y_0^2-\delta y^2}V_{\mathrm{s}}\ , \\ &\approx 2\frac{C_{\mathrm{s}}}{C_{\mathrm{i}}}\frac{\delta y}{y_0}V_{\mathrm{s}}\ , \end{aligned} \tag{32.7}$$

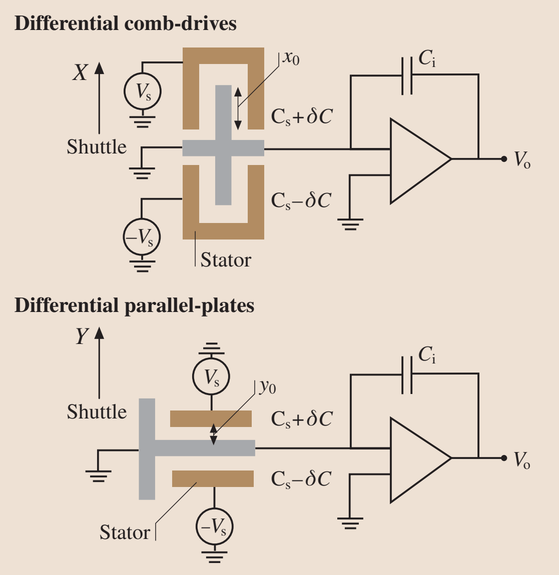
Differential comb-drives
X
Shuttle
V_s
x_0
C_i
$C_s+\delta C$
$C_s-\delta C$
V_o
$-V_s$
Stator
Differential parallel-plates
Y
Shuttle
V_s
y_0
C_i
$C_s+\delta C$
$C_s-\delta C$
V_o
Stator
$-V_s$

Fig. 32.5 Capacitive position sensing. Comb-drives versus parallel-plates motion

where x_0 is the nominal overlap for interlaced fingers and y_0 is the nominal gap between overlapped fingers, as shown in Fig. 32.5.

As indicated by (32.7), the voltage output for the comb-drives sensing structure, V_{comb}, is linear with shuttle displacement. On the other hand, the sensing configuration that makes use of parallel-plates motion has better sensitivity for detecting shuttle motion since y_0 is usually smaller than x_0. Although the nonlinearity in parallel-plates sensing can be linearized by the differential drive configuration to some extent, in a design example of 4 μm gap with 1 μm stroke, the linear model shown in (32.7) still produces 6% deviation from the nonlinear model.

Among electrostatic microactuators that use capacitive position sensing, the capacitance variation due to shuttle motion ($\mathrm{d}C/\mathrm{d}x$) is typically at the level of 100 fF/μm. In order to obtain 10 nm position sensing resolution, the capacitance sensing circuit must be able to detect capacitance variation of 1 fF in the presence of parasitic capacitance and offset/mismatches from op-amps, which can easily result in an output voltage orders of magnitude larger than the output voltage from the designated capacitance variation. In most capacitive position sensing, the limiting factor for the sensing resolution is not the thermal noise but the sensing circuit's design.

Here we introduce two basic capacitance sensing circuits suitable for high-resolution position sensing [32.14]. The concept of the "synchronous scheme," as shown in Fig. 32.6 I, is to reduce the impedance of sense capacitors as well as the offset and 1/f noises from op-amps by applying modulation techniques on the sense voltage (V_s). The R_{dc} resistor on the feedback loop sets the DC voltage level at the input nodes of the "charge integrator." The effect of the presence of parasitic capacitance (C_p) is nullified by the virtual ground condition from the op-amps. The major drawback of the synchronous scheme is that the DC-setting resistor (R_{dc}) has to be large to ensure the proper gain for the capacitance sensing [32.14], which introduces excessive thermal noise into the sensing circuit. In addition, a large resistor usually consumes a large die space in implementation.

A switched-capacitance scheme, as shown in Fig. 32.6 II, is one alternative that avoids the use of DC-setting resistor. The capacitance sensing period is broken into two phases: reset phase and sense phase. During the reset phase, input/output nodes of capacitors

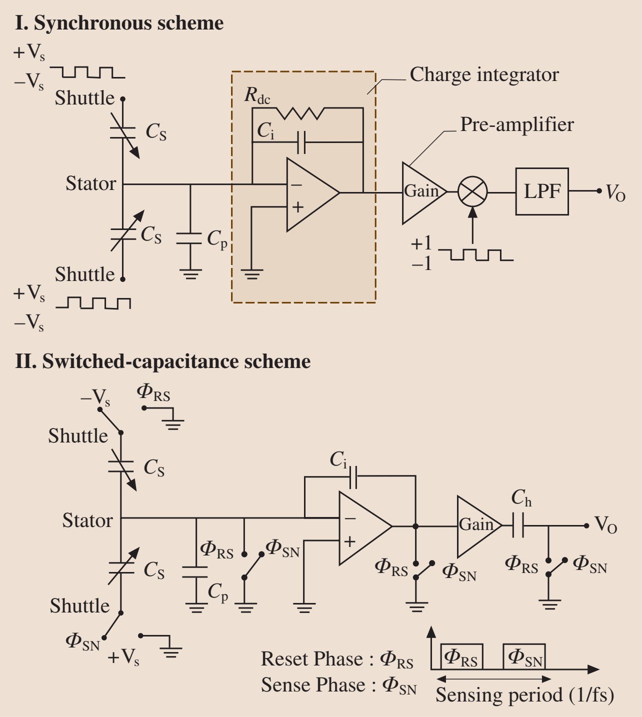
I. Synchronous scheme
$+V_s$
$-V_s$
Shuttle
C_S
Stator
C_S
C_p
Shuttle
$+V_s$
$-V_s$
R_{dc}
C_i
Charge integrator
Pre-amplifier
Gain
LPF
V_O
+1
−1
II. Switched-capacitance scheme
$-V_s$
Φ_{RS}
Shuttle
C_S
Stator
C_S
Shuttle
Φ_{SN}
$+V_s$
Φ_{RS}
Φ_{SN}
C_p
C_i
Gain
C_h
V_O
Φ_{RS}
Φ_{SN}
Φ_{RS}
Φ_{SN}
Reset Phase : Φ_{RS}
Sense Phase : Φ_{SN}
Φ_{RS}
Φ_{SN}
Sensing period (1/fs)

Fig. 32.6 Capacitive position sensing circuits. Synchronous scheme versus switched-capacitance scheme

and input voltage to op-amps are set to ground or reference level to ensure proper DC voltage for the charge integrator. During the sense phase, sense voltage $\pm V_s$ is applied to the sense capacitors, and the amount of charge proportional to the mismatch in the sense capacitors is integrated on the capacitor C_i, thus producing an output voltage proportional to the capacitance mismatch from the sense capacitors. This approach replaces the large DC-setting resistor by capacitors and switches and results in a much smaller die compared to the synchronous scheme. Furthermore, the switching technique allows more design flexibility for system integration and performance improvement because of the ability to allocate separate phases for various operations. The major drawback of this design is that it draws noise into sensing circuits from switches and sampling capacitor C_h. However, these noises can be compensated by dividing the sense phase into $2 \sim 3$ sub-sense phases [32.14], at the expense of a complicated circuit design.

Piezoresistive Sensing

Piezoresistive films have been widely used as strain-sensitive components in a variety of MEMS devices, including pressure sensors and vibration sensors. Generally speaking, piezoresistive sensing techniques require less complicated sensing circuits and perform better in a severe environment than other sensing techniques. When a piezoresistive film is subjected to stress, the film resistivity and dimensions change. The fractional change of resistance is proportional to the deformation of the piezoresistive film. For a small change of resistance, this relation can be expressed as,

$$\frac{\Delta R}{R} = K \cdot \epsilon \,, \tag{32.8}$$

where R is resistance of the piezoresistive film, K is its gage factor and ϵ is strain. In microactuator designs, the piezoresistive film is usually applied to the spring structure that connects the moving shuttle to an anchor point. When the shuttle moves, it stretches the spring as well as the piezoresistive film, consequently, the piezoresistive film produces a deformation signal proportional to the shuttle displacements. As a result, piezoresistive sensing is easier to implement than capacitive position sensing, but the sensing resolution is usually less accurate due to higher thermal noises introduced by the resistance of the piezoresistive film.

Another application of piezoresistive sensing, aside from measuring relative slider position, is to detect vibration in the suspension itself. The idea is to sense airflow-induced vibration of the suspension and feed that information forward to an actuated slider to damp out motion at the head. Piezoresistors used for this purpose can be made from metal or semiconductor materials, arranged as a strip or series of strips oriented along the direction of vibration strain. It is important that these sensors observe all vibration modes that contribute to off-track error, so a number of optimization schemes for locating the sensors have been developed [32.15]. One method is to maximize the minimum eigenvalue of the observability matrix of the sensor or sensors. This ensures that all relevant modes are observed. Another method is to minimize a linear quadratic gaussian control problem over potential sensor locations [32.16]. This serves to determine an optimal placement from the perspective of a linear controller.

32.1.5 Electrostatic Microactuator Designs for Disk Drives

Various electrostatic microactuators have been designed for secondary actuation in HDDs. To incorporate an electrostatic microactuator into a HDD without altering much of current suspension configuration, many design constrains are imposed. In this section, we will first discuss some design issues and then present one specific design example.

Translational Microactuators versus Rotary Microactuators

Depending on the motion of the magnetic head actuated, microactuator designs are categorized into two groups: translational actuators and rotary actuators. Either type can be implemented by comb-drives [32.8, 17] or parallel-plates [32.5, 6, 18] actuation.

When employing a translational microactuator in a dual-stage HDD servo, previous research [32.13] has shown that a force coupling between the suspension and the translational microactuator exists, consisting of transmitted actuation force from the VCM and suspension vibration induced by windage. The force coupling from the VCM not only complicates the dual-stage servo controller but also imposes a design constraint on a translational microactuator design, in that the translational microactuator has to provide a large force output to counterbalance the coupling force. When the VCM makes a large movement, as in seeking a new data track, the microactuator may be overpowered. One solution is to pull the actuator to one side and lock it momentarily in place. Even then, the use of the two actuation-stages must be carefully coordinated to moderate the influence of the VCM on the microactuator. On the other hand, the linear springs in the translational microactuator can also aid in damping out motion of the suspension. The portion of suspension vibration induced by windage mostly consists of high frequency excitation, so the resulting magnetic head's position error can be passively attenuated by low resonant frequency translational microactuators.

Generally speaking, rotary actuators are more difficult to design/analyze than translational actuators because of their nonuniform gap between shuttle and stator. Still, their different operating properties have both strengths and weaknesses. Unlike a translational actuator, no obvious force coupling is transmitted from the suspension to a rotary actuator, as the microactuator is nearly always attached to the end of suspension at the microactuator's center of rotation, which acts as a pivot point. With no mechanical coupling, the dual-stage servo system using a rotary microactuator does not suffer from the force coupling between VCM and microactuator seen in translational designs. However, the rotary microactuator has to compensate for the magnetic head's position error induced by suspension vibration without any passive attenuation of the vibration. Overall, a rotary acuator is likely to behave better than a translational microactuator during track seeking and worse during track following.

Gimballed Microactuator Design

Proper flying height and orientation of a slider and read-write head over a hard disk is maintained by the interaction of the suspension, the air bearing of the slider, and a gimbal structure. The gimbal structure is located at the tip of the suspension and holds the slider/microactuator in its center coupon. The dynamic characteristic requirements of gimbal structures are that they be flexible in pitch and roll motion but stiff in in-plane and out-of-plane bending motion. To meet all these requirements using only one piece of metal is a highly challenging task. For this reason, most commercially available suspension/gimbal designs consist of two to four pieces of steel, each with different thickness.

The goal of a gimballed microactuator design is to seamlessly integrate an actuator and gimbal into a one-piece structure, as shown in Fig. 32.7, to simplify both suspension design and HDD assembly. A full integrated suspension that includes suspension, gimbal, and microactuators in one part has also been proposed [32.19]. The dimple structure, existing in most current gimbal structures, is excluded from the gimballed microactuator design and electrical interconnects are in situ fabricated on the gimballed microactuator, replacing the flexible cable in current HDD suspension assembly.

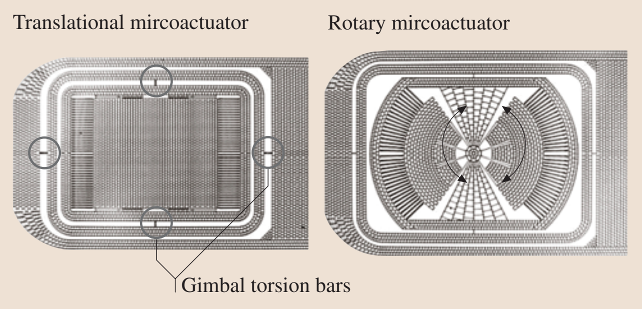

Fig. 32.7 Schematics of translational microactuator versus rotary microactuator. Courtesy from Lilac Muller

The dimple structure in current suspension assemblies provides out-of-plane stiffness while preserving the necessary torsional compliance in the gimbal structure. Without a dimple structure in the suspension assembly, the gimbal itself must provide high out-of-plane bending stiffness. Otherwise, it would unbalance the suspension pre-load, which is an overbend of the suspension that balances the upward air-bearing force on the suspension during operation. The electrical interconnects are implemented to transmit data between magnetic heads, located at the center coupon of the gimbal structure, and IC circuits located at the end of the suspension. The in situ fabricated electrical interconnects are inevitably passed through torsion bars of the gimbal and thus set a design constraint for the minimum width of torsional bars. Furthermore, both the gimbal structure and microactuator should be the same thickness to simplify the MEMS fabrication process.

To summarize the design constrains discussed above, the integrated gimbal structure has to meet performance requirements with a single, uniform piece of material that would previously have been achieved by two to four metal pieces with different thicknesses, while the minimum width of any torsion bars that may be used pre-determined. To solve this problem, *Muller* [32.6] proposed a T-shaped structure (a beam structure with overhang surface sheet) for the torsion bars, and *Chen* [32.19] proposed "double-flexured" torsion bars. Additionally, many suspension manufacturers have developed new gimbal structures for their suspensions designed specifically for use with MEMS microactuators, moving the gimbal location back to the suspension from the microactuator.

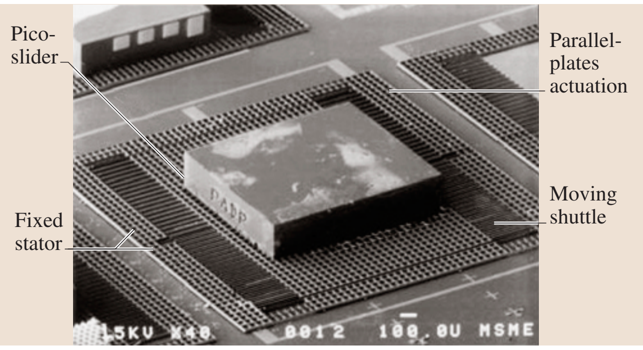

Fig. 32.8 Pico-slider mounted on a translational microactuator. Courtesy from Horsely 1998

An Electrostatic Microactuator Design Example

Figure 32.8 shows a translational microactuator design suitable for the HDD dual-stage actuation by *Horsley* in 1998 [32.5]. The translational electrostatic microactuator dimensions are 2.2 mm × 2.0 mm × 0.045 mm and weight 67 µg. The dimensions of the pico-slider on the top are 1.2 mm × 1.0 mm × 0.3 mm and weight 1.6 mg. This microactuator design does not include "electrical-isolation" features, and thus the electrical isolation and electrical-interconnects were fabricated on a separate substrate and subsequently bonded to the microactuator. This microactuator uses parallel-plates for actuation force but does not have dedicated position sensing structures due to fabrication process limitations. Table 32.1 summarizes key parameters of this microactuator design.

Based on these parameters, the characteristics of this electrostatic microactuator can be estimated by the linear differential equation shown in (32.6).

Figure 32.9 shows the schematics of a circuit design by *Wongkomet* in 1998 [32.14], in which actuation driving voltage and capacitive position sensing were implemented for the electrostatic microactuator designed by Horsley. As mentioned before, the electrostatic microactuator design doesn't have dedicated structure for a position sensing; as a consequence, the input nodes for actuation and output nodes for capacitive position sensing have to share the same electrodes. Capacitors C_c and C_{c0}, shown in Fig. 32.9, are carefully designed to shield the high voltage presented in actuation circuit from sensing circuit, which is mostly low voltage, and thus enable driving/sensing circuit integration.

The driving circuit, shown in the left in Fig. 32.9, demonstrates how to generate the bias voltage (V_{bias}, ±40 V) and drive voltage (V_{dr}, -40–$+40$ V) from 0–5 V CMOS compatible circuits. The switches at the output of the charge pumps were synchronized with the switching period ϕ_{RS}. During the sense phases ϕ_{SN1} and ϕ_{SN2}, therefore, the switches are left open and thus no voltage fluctuation is seen by the sensing circuits. This arrangement was utilized to reduce feedthrough from the driving circuit to position sensing circuit.

The design target for the capacitive position sensing circuit was to achieve position sensing resolution of 10 nm, and this goal was approached by two main techniques implemented in the circuit: differential sensing and Correlated-Double-Sampling (CDS). The main

Table 32.1 Parameters of the electrostatic microactuator design by *Horsley* 1998 [32.5]

Parameters	Source*	Value
Nominal gap	D	10 µm
Structure thickness	D	45 µm
Rotor mass, m	I	44 µg
$\mathrm{d}C/\mathrm{d}x$	C	68 fF/µm
Actuation voltage, bias voltage	D	40 V
Actuation voltage, maximum driving voltage	D	±40 V
Voltage-to-force gain K_v	I	50 nN/V
Mechanical spring constant K_m	I	29 N/m
Electrostatic spring constant K_e	I	9.6 N/m
Damping coefficient, b	I	1.03×10^{-4} N/(m/s)
Voltage-to-position DC gain	M	0.05 µm/V
Resonance frequency, w_r	M	550 Hz

*: D = Design value, C = Calculation, M = Measurement, I = Inferred from measurements

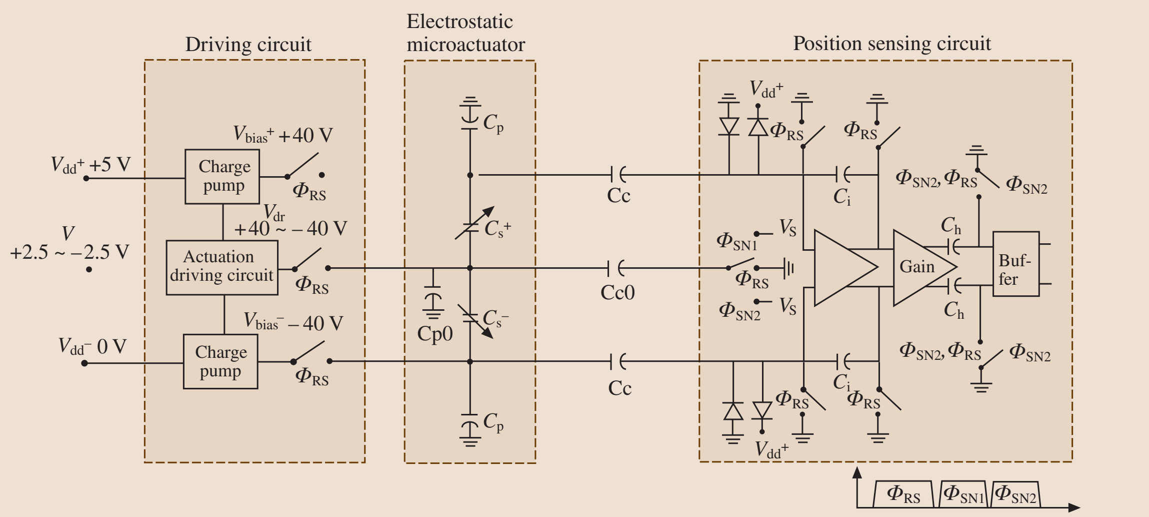
Driving circuit
Electrostatic microactuator
Position sensing circuit
V_{dd}^+ +5 V
V
+2.5 ~ −2.5 V
V_{dd}^- 0 V
V_{bias}^+ +40 V
Charge pump
Φ_{RS}
V_{dr}
+40 ~ −40 V
Actuation driving circuit
Φ_{RS}
V_{bias}^- −40 V
Charge pump
Φ_{RS}
C_p
C_s^+
Cp0
C_s^-
C_p
Cc
Cc0
Cc
V_{dd}^+
Φ_{RS}
Φ_{RS}
C_i
Φ_{SN2}, Φ_{RS}
Φ_{SN2}
Φ_{SN1}
V_S
Φ_{RS}
Φ_{SN2}
V_S
C_h
Gain
Buf-fer
C_h
Φ_{SN2}, Φ_{RS}
Φ_{SN2}
C_i
Φ_{RS}
Φ_{RS}
V_{dd}^+
Φ_{RS}
Φ_{SN1}
Φ_{SN2}

Fig. 32.9 A simplified schematic including driving and sensing circuits for the electrostatic microactuator. Courtesy from Wongkomet 1998

benefits of the differential sensing scheme are reduced noise coupling and feedthrough, elimination of even-order harmonics, and improvement of dynamic range by doubling the output swing. To adapt this differential sensing scheme in a differential parallel-plates electrostatic microactuator, a bias voltage (V_{bias}) was applied to the stator and the drive voltage (V_{dr}) was applied to the shuttle. The CDS technique, a modified capacitance sensing technique based on the switched-capacitance scheme, was implemented along with the differential sensing scheme to compensate for sensing noises including 1/f, KT/C, switch charge injection and offset from op-amps [32.14]. The concepts of CDS can be briefly described as follows: the sense period is broken into three phases: one reset phase, ϕ_{RS}, and two sense phases, ϕ_{SN1} and ϕ_{SN2}. During the ϕ_{RS} sense phase, the voltages for capacitors and input nodes to op-amps are set to the reference level of the DC-voltage setting for the charge integrator, same as for the switched-capacitance scheme discussed in Sect. 32.1.4. During the ϕ_{SN1} sense phase, a sensing voltage $-V_{\text{s}}$ is applied to the shuttle and results in a voltage difference, $\alpha(-V_{\text{s}}) + V_{\text{error}}$, across the sampling capacitor C_{h}, where α is the transfer function from sense voltage to the output voltage at the pre-amplifier, shown as *Gain* in the plot, and V_{error} is the voltage at the output node of pre-amplifier resulting from noise, leakage charge, and offset of op-amps. Lastly, during the ϕ_{SN2} sense phase, the sensing voltage is switched from $-V_{\text{s}}$ to V_{s} and the switch next to C_{h} is switched open. This results in a voltage output $\alpha(V_{\text{s}}) + V_{\text{error}}$ at the output node of the pre-amplifier and a voltage, $2\alpha(V_{\text{s}})$, at the input node of the buffer. As a consequence, the voltage resulting from sensing error (V_{error}), which appears at the output node of pre-amplifier, disappears from the voltage-input node to the buffer. Switches utilized in

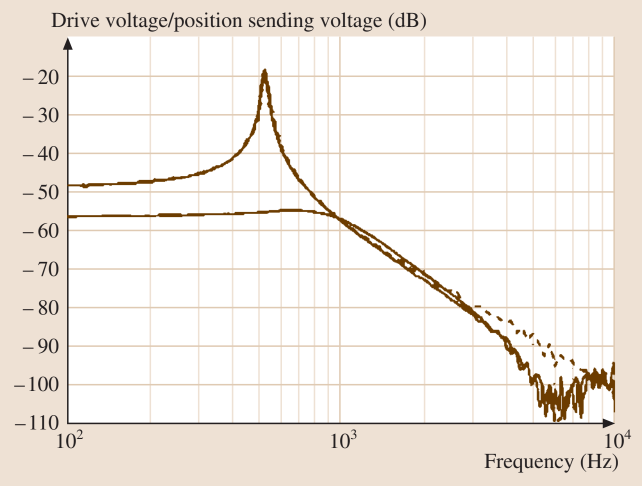

Fig. 32.10 Open-loop and closed-loop frequency response of the prototype microactuator with a pico-slider. The capacitive position measurement (*solid line*) is compared to the measurements from LDV (*dashed line*). Courtesy from Wongkomet

the circuit and their correspondent timing are shown in Fig. 32.9. Be aware that for presentation simplicity components in the pre-amplifier and buffer are not shown in detail in Fig. 32.9.

The frequency response, both open-loop and closed-loop with a PD controller, of the electrostatic microactuator measured by a capacitive position sensing circuit is shown in Fig. 32.10. The position measurements from an LDV are also shown in the same plot for comparison. The deviation between measurements from the different position sensing devices appears in the high frequency region of the plot and has been identified as the feedthrough from the capacitance sensing circuit. The effect from feedthrough was negligible at low frequency region but becomes significant and results in a deviation in magnitude for the transfer function of $-80\,\text{dB} \sim -90\,\text{dB}$ after 2 kHz. The feedthrough presented in the capacitance sensing circuits limited the position sensing resolution to the level of 10 nm.

32.2 Fabrication

While there are several approaches to building electrostatic microactuators suitable for hard disk drives, they all exhibit certain common features from a fabrication standpoint.

32.2.1 Basic Requirements

As we discussed earlier, nearly all electrostatic microactuators rely on a system of interlaced fingers or plates to provide actuation force. As a result, a method for producing arrays of these fingers or plates with narrow gaps between them is usually the central concern in developing a fabrication process. The resulting structure must then be strong enough to support both the slider on the microactuator and the microactuator on the suspension, particularly when loaded by the air-bearing that supports the slider above the hard drive's spinning disk. In addition, the design and fabrication process must include a way to perform electrical interconnection on the microactuator. This involves transferring signals to and from the actuator and slider and isolating the parts of the microactuator requiring different voltage levels.

Meanwhile, the microfabrication process is subject to certain basic constraints. The materials used in fabrication must either be thermally and chemically compatible with any processing steps that take place after their deposition or must somehow be protected during steps that would damage them. This often constrains the choice of materials, deposition techniques, and processing order for microdevices. Another concern is that the surface of the structure be planar within photoresist spinning capabilities and lithography depth-of-focus limits if patterning is to be performed. This can be a major challenge for disk drive microactuators, which are large in size and feature high-aspect ratio trenches compared to other MEMS devices.

32.2.2 Electrostatic Microactuator Fabrication Example

Section 32.1.5 described the design and operation of a translational electrostatic microactuator. This section examines the fabrication process by which that microactuator was built [32.5]. The process is a variation on a micro-molding process known as HexSil [32.20]. In a micro-molding process, a mold wafer defines the structure of the microactuator and may be reused many times like dies in macro-scale molding; in the HexSil version of the procedure, the mechanical parts of the microactuator are formed in the mold by polysilicon. A second wafer is used to create metallized, patterned target dies. Upon extraction from the mold, the HexSil structure is bonded to this target substrate, which is patterned to determine which sections of the HexSil structure are electrically connected.

Naturally, the mold wafer is the first item to be processed and is quite simple. It is a negative image of the desired structure, etched down into the wafer's surface, as shown in Fig. 32.11a. A deep but very straight etch is critical for successful fabrication and subsequent operation of the devices. If the trench is too badly bowed or is undercut, the finished devices will be stuck in the mold and difficult to release. On the other hand, a tall device will have a larger electrostatic array area, will generate more force, and be better able to support a slider while in a disk drive.

Fabrication of the HexSil structure forms the majority of the processing sequence Fig. 32.11b–g. HexSil fabrication begins by coating the surface of the mold wafer with a sacrificial silicon oxide. This layer, deposited by low-pressure chemical vapor deposition (LPCVD), must coat all surfaces of the mold, so that removal of the oxide at the end of the process will

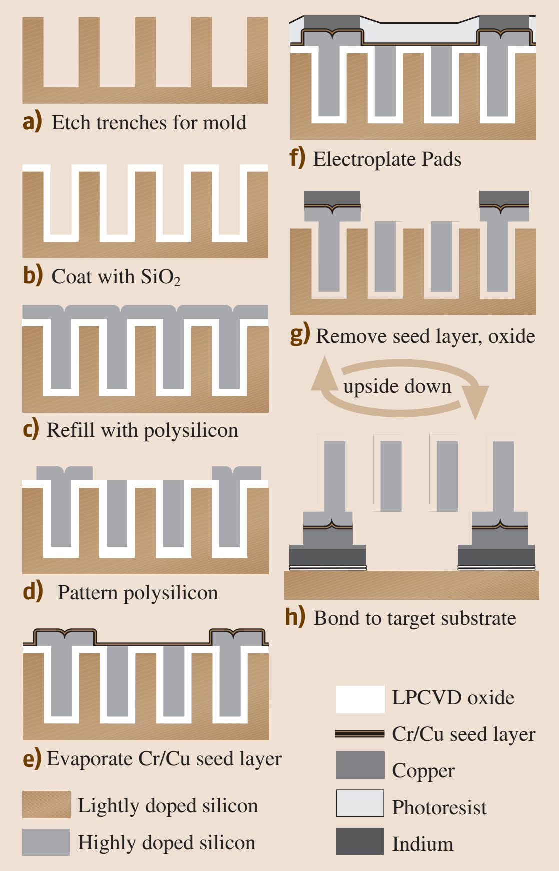

Fig. 32.11 Fabrication example 1: HexSil

leave the device completely free of the mold. This microactuator uses a 3 to 4 µm thick oxide layer to ensure clearance between the polysilicon structure and the trench walls during release. The mold is then refilled completely with LPCVD polysilicon, which will form the desired mechanical structure. Polysilicon is chosen for its conductivity and good conformality in refilling trenches. This also leaves a planar surface ready for photolithography to cover bonding locations. After the lithography step, the polysilicon deposited on top of wafer is removed, except where the bonding points were defined.

The structure is then prepared for bonding by forming a soldering surface. First a chrome/copper seed layer is evaporated on the surface of the wafer. The chrome promotes copper adhesion to the silicon, while the copper forms the starting point for electroplating. The seed layer is covered with photoresist, which is cleared only in the locations where contacts are desired. There, electroplating will produce a metal film, forming soldering points for the device to the target substrate that will form its base. In this case, copper is used as the plating material, thanks to its high conductivity and solderability. The device is released from the mold by ion milling away the thin seed layer, then dissolving the oxide lining with a hydroflouric (HF) acid wet etch. The free standing portion of the microactuator is thus ready for bonding to the target. The mold, meanwhile, may be used again after a simple cleaning step.

The target substrate may be prepared in several ways. The target shown above consists simply of indium solder bumps and interconnects on a patterned seed layer; this is known as a “plating bus” arrangement. The seed layer is used for the same reasons as on the HexSil structure, but in this case the seed layer is patterned immediately, with the unwanted portions removed by sputter etching. This provides isolation where desired without having to etch the seed layer after the indium is in place. Photoresist is spun again and patterned to uncover the remaining seed layer. A layer of indium is then electroplated everywhere the seed layer is visible. Indium acts as the solder when the two parts of the microactuator are pressed together, forming a cold weld to the copper at 200–300 MPa as shown in Fig. 32.11h, that was found suitable for hard disk requirements.

32.2.3 Electrostatic Microactuator Example Two

The high-aspect ratio microactuator described in this section is in development at the University of California, Berkeley, and demonstrates another way to create the features required of an actuated slider. The microactuator is translational, with parallel-plates actuation, and includes two of the design options described previously: integrated isolation plugs, as described in Sect. 32.1.3, and a capacative sensing array, as described in Sect. 32.1.4. A picture of the basic design is shown in Fig. 32.12; a central shuttle holding the slider is supported by four folded-flexure springs and driven by parallel-plates arrays.

Basic Process: Silicon-on-Insulator

Successful processing of the design is centered around the ability to etch very straight, narrow trenches using deep reactive ion etching (DRIE). DRIE is a special plasma etching sequence that uses polymerization of the sidewalls of a trench to keep the walls straight even at extreme aspect ratios. Deep, narrow trenches

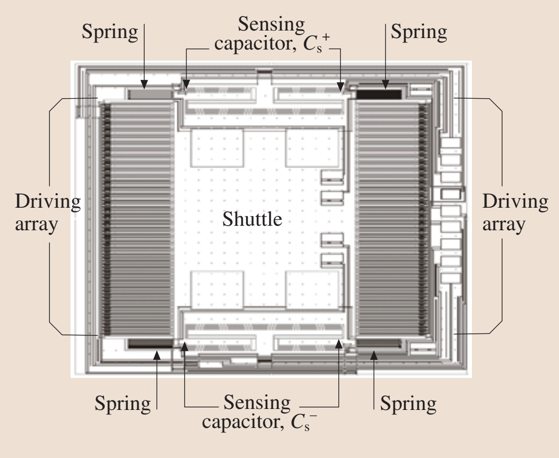

Fig. 32.12 Translational high-aspect ratio microactuator

make possible a microactuator with both closely packed electrostatic arrays and good mechanical strength. Conventional surface micromaching is used to produce the majority of the electrical interconnects.

The basic fabrication process uses a silicon-on-insulator (SOI) wafer to control DRIE trench depth. SOI wafers are very useful in microfabrication as they provide a layer of single-crystal silicon (the device layer) separated from the bulk of the wafer (the handle layer) by a thin layer of buried silicon oxide. This gives a very well controlled thickness to finished device but is very expensive. The microactuators described here are fabricated from an SOI wafer with a 100 µm device layer. A variation on the processing sequence and layout eliminates the need for SOI wafers and will be discussed in the following section.

The first stage of fabrication is creation of deep isolation trenches (see Fig. 32.13a–c). This procedure has been adapted to MEMS from integrated circuit processing for isolating thick MEMS structures [32.21]. The isolation pattern is formed by photolithography and etched by DRIE down to the buried oxide. With trenches only 2 µm wide, this corresponds to an aspect ratio of 50 : 1. The wafer surface and trenches are then coated with LPCVD silicon nitride, which acts as the electrical insulator. The trench is then refilled in its entirety with LPCVD polysilicon, which is a more conformal material better fills the trench than silicon nitride would alone; polysilicon also has much lower residual stresses than silicon nitride. The refill leaves a layer of polysilicon on the surface of the substrate, which must be etched or polished back to the silicon nitride. After etch-back, a second layer of LPCVD nitride completes isolation between regions.

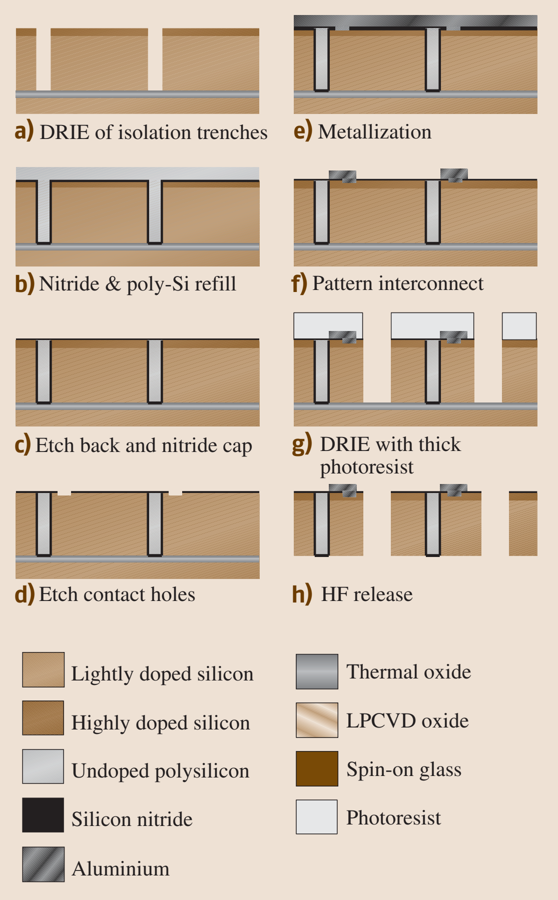
a) DRIE of isolation trenches
b) Nitride & poly-Si refill
c) Etch back and nitride cap
d) Etch contact holes
e) Metallization
f) Pattern interconnect
g) DRIE with thick photoresist
h) HF release

Fig. 32.13 Basic silicon-on-insulator fabrication process

Next, electrical interconnects are formed, as shown in Fig. 32.13d–f. Contact holes to the substrate are defined by photolithography and etched by reactive ion etching of the silicon nitride. Metal lines may then be patterned directly by photolithography or formed by lift-off. In a lift-off process, photoresist is deposited and patterned before the metal. The metal is then deposited vertically, so that the sidewalls of the resist are uncovered. Dissolving the resist allows the metal on top to float away, leaving behind the interconnect pattern. In the figures shown here, the interconnects are formed by evaporated aluminum, eliminating the need for electroplating and a seed layer. The ability to isolate portions of the substrate is useful for simplifying the interconnect layout, as one interconnect can cross another by passing underneath it through an isolated portion of the substrate.

Finally, structural lithography is performed and a second etch by DRIE is done, as shown in Fig. 32.13g. These trenches will define the shape of the microactuator, and are larger, 4 μm wide, to allow sufficient rotor travel. After this etch step, the device layer on the SOI buried oxide consists of fully defined microactuators. When the buried oxide is removed by wet etching, the microactuator is released from the substrate Fig. 32.13h and, after removal of the photoresist that protected the device surface from HF during release, ready for operation. Thanks to the high-aspect ratios of DRIE and the efficient integrated isolation scheme, the microactuator can theoretically be operated with a DC bias voltage of 15 V on the rotor, and a dynamic stator voltages below 5 V. In practice, higher voltages are required due to the spreading of trench walls over the course of DRIE. A picture of the electrostatic structure, with the rotor fingers on the left and the stator fingers on the right, divided down the center by isolation trenches, is shown in Fig. 32.14.

Variation: Anisotropic Backside Release

An alternative fabrication approach is to build the microactuator from a regular silicon wafer ($\sim 550\,\mu$m thick), but then, since the microactuator shouldn't be more than 100 μm thick to fit in a HDD, etch away the backside of the wafer to obtain devices of that size. The main reason for using a silicon-only wafer is to reduce the need for expensive SOI wafers, which cost approximately ten times as much as single-material wafers. Another benefit is the elimination of the need for etch holes to reach the buried oxide layer, which increase vulnerability of microactuators to particle or moisture contamination. However, the thickness of the devices cannot be controlled as accurately as with the SOI process, and the processing sequence is more complicated.

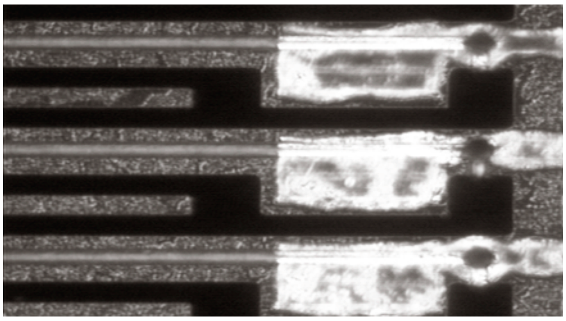

Fig. 32.14 Electrostatic driving array of microactuator from SOI

Control of backside etching is difficult, making it hard to obtain devices with uniform thickness across a wafer. A proposed solution is to use an anisotropic wet etchant during the backside etch step. Anisotropic etchants work very slowly on certain crystal planes of a single-crystal silicon wafer; a coating on trench sidewalls can protect fast-etching planes when the etchant reaches the trenches from the back, causing slow-etching crystal planes to begin coming together slowing the etch. The result is a nearly self-stopping etch with a V-shaped profile that prevents overetching even in the presence of nonuniformities in etch rate and trench depth across the wafer. The most common anisotropic etchants are potassium hydroxide (KOH) and tetramethyl ammonium hydroxide (TMAH). This release process is referred to here as an anisotropic backside release (ABR).

The key to a good release by this method is to achieve excellent protection of the top of the wafer and the deep trenches from the etchant, and this neccessitates changes in the process flow from that for an SOI wafer. Without a good protective film, the interconnects on the surface of the wafer could be destroyed, or the fast-etching planes might be attacked along the trench sidewalls. Of the etchants mentioned above, TMAH is preferred, due to high selectivity to both silicon oxide and nitride and, especially, thermally grown oxide, which can be used as protective films. However, growing thermal oxide and/or depositing high-quality nitride are high temperature processes not compatible with most metals, so the structural trenches must be etched and oxidized before forming the electrical interconnects. As shown in Fig. 32.15c–d, the isolation trenches are formed as before, but the structural trenches are patterned and etched in the very next step. A thermal oxide is then grown in the trenches to produce the protective oxide layer Fig. 32.15e.

To do lithography for the interconnects at this point, then, the surface of the wafer must be planarized. This is done by refilling the trenches with spin-on-glass. The spin-on-glass plugs the trenches nearly to the top and can be sacrificed along with the thermal oxide at the end of the process. Extra glass on the wafer surface can be removed by a quick HF dip or chemical mechanical polishing. The resulting surface is uneven, as shown in Fig. 32.15f, but a thick photoresist layer can be applied over the trenches evenly enough to accomplish contact and metal patterning lithographies. After metallization much the same as for the SOI wafer Fig. 32.15g, a final coating of etch resistant material is required to protect the metal during backside release, as shown in

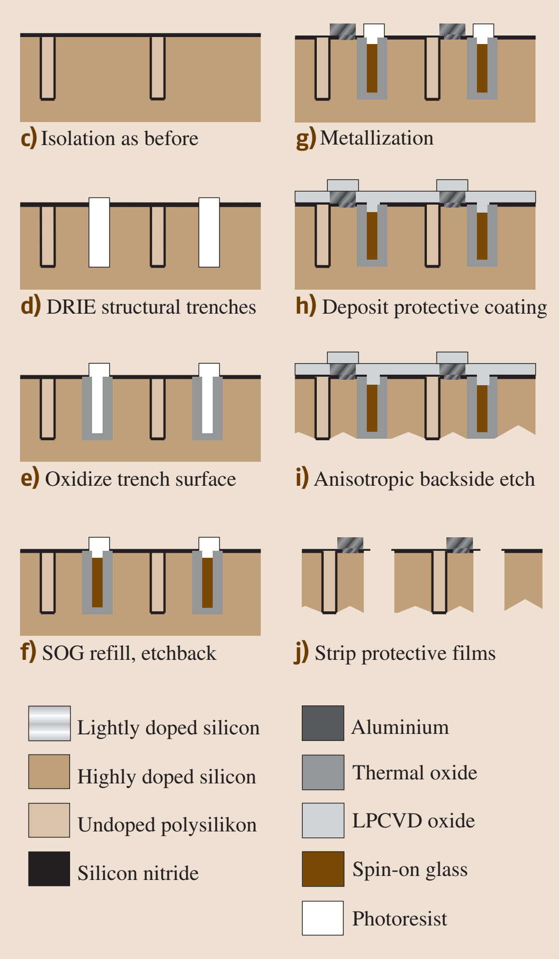
c) Isolation as before
g) Metallization
d) DRIE structural trenches
h) Deposit protective coating
e) Oxidize trench surface
i) Anisotropic backside etch
f) SOG refill, etchback
j) Strip protective films
Lightly doped silicon
Highly doped silicon
Undoped polysilikon
Silicon nitride
Aluminium
Thermal oxide
LPCVD oxide
Spin-on glass
Photoresist

Fig. 32.15 Deep trench fabrication process adapted for backside release

Fig. 32.15h–j. After ABR, the backside of the device will be shaped by crystal planes, but the thickness of the device at the edge of DRIE features will be fixed. While a combination of silicon oxide and silicon nitride has been effective over a "metallization" of undoped polysilicon, an effective combination of true metal interconnects and protective films has yet to be found and is an ongoing area of research for improving the ABR process sequence.

A prototype microactuator with integrated gimbal fabricated using ABR is shown in Fig. 32.16. This prototype was operated with a 30 V DC bias and ±8 V AC driving voltage for ±1 μm displacement, while a larger version (shown in Fig. 32.12) was operated at 15 V bias and ±3 V driving. This layout and fabrication concept is currently being adjusted to fit a MEMS-ready suspension for more advanced testing.

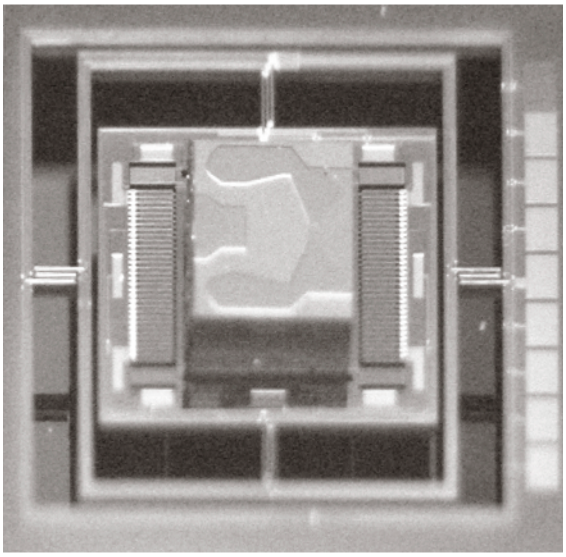

Fig. 32.16 Microactuator with integrated gimbal fabricated by ABR

32.2.4 Other Fabrication Processes

Where the DRIE-based process described above "digs" the microactuator out of a wafer, a procedure developed by IBM "grows" a microactuator with high-aspect ratio trenches on top of a wafer [32.22].

IBM Electroplated Microactuator

The microactuator is fabricated by a clever sequence of electroplating steps and sacrificial depositions. An oxide sacrificial film is first patterned then covered with a metal seed layer. When later removed, the sacrificial will have left a gap between the microactuator and substrate portions of the microactuator needed to move freely. A polymer, 40 μm thick, is spun onto the wafer and patterned with a reverse image of the main structural layer. Trenches are etched down to the sacrificial layer by plasma etching then refilled by electroplating, much like micro-molding. Another sacrificial polymer, photoresist this time, separates the top of the fixed portions of the structure from the platform upon which the slider will sit, while conveniently planarizing the surface for further lithography. The platform is created by two more electroplating steps on top of the parts of the structure that will move freely. Removal of the sacrificial layers (photoresist, polymer, and oxide) releases the microactuator for use.

This process has been used to produce both rotary and linear microactuators and is comparatively advanced from a commercial standpoint. The resulting devices are thinner than those described in the previous example, but a similar aspect ratio (20 : 1) can be achieved. This procedure also has the benefits of being a low temperature process, which helps decrease processing cost and is compatible with the thin film magnetic head manufacturing process. Moreover, it includes an upper structure that covers the fingers, which improves device reliability by shielding the fingers from particles.

Additional fabrication details and dynamic testing results for the IBM design may be found in references [32.22] and [32.23]. Certain control results are described in Sect. 32.3.2.

Electroplated Microactuators with Fine Gaps

Finally, fabrication processes developed at the University of Tokyo demonstrate additional techniques for obtaining very small gaps between electrostatic fingers [32.24]. Similar to the IBM process, electrostatic fingers are electroplated, in this case by nickel, within a polymer pattern. The gap between closely spaced fingers, however, is then formed by a sacrificial metal. Photoresist is used to cover the fingers except for the surfaces where a small gap is desired; these are electroplated with sacrificial copper, making narrow, well-defined gaps possible, in perhaps the most reliable of the processes described here. Continuing with a second nickel electroplating step creates the interlaced fingers, which will also support the slider; the wafer is polished down to level the structure. Finally the copper, photoresist, and seed layer are etched away, leaving the free standing structure.

In another version of process, photoresist alone separates the stator and rotor fingers. In this case, a thin layer of photoresist is left where small gaps are desired, instead of growing a layer of copper. This method is dependent on excellent alignment but can eliminate the need for polishing after the second electroplating step.

32.2.5 Suspension-Level Fabrication Processes

The long, highly integrated fabrication processes required by actuated slider microactuators contrast greatly with the fabrication of actuated suspension microactuators.

PZT Actuated Suspensions

In most cases, actuated suspensions use piezoelectric (PZT) drivers cut or etched as single, homogeneous pieces and attached to the steel suspension. Nevertheless, microfabrication techniques can still be useful in suspension processing. For example, some recent suspensions have used thin film PZT deposited on a substrate by sputtering or spinning-on of a sol-gel for incorporation in a suspension. This is intended to produce a higher quality PZT films and begins to introduce MEMS-style processing techniques even into suspension-scale manufacturing [32.25, 26].

Integrated Silicon Suspension

Another area where silicon processing has been suggested for use on a suspension is the gimbal region, or even the entire suspension, as described in [32.19]. An integrated silicon suspension would incorporate aspects of both actuated suspension and actuated slider fabrication. Force is provided by four piezes of bulk PZT cut down to size, as in other actuated suspensions, but the rest of the suspension is fabricated from single-crystal silicon by bulk micromachining.

The fabrication process for the integrated suspension is the precursor to the ABR process used to fabricate the microactuator in Sect. 32.2.3. Layers of LPCVD silicon nitride and polysilicon are deposited on a bare silicon wafers. The polysilicon is doped by ion implantation and covered with a second layer of nitride. The outline of the suspension and any spaces within it are then lithographically patterned. After a reactive-ion nitride etch to reach the substrate, trenches are etched by DRIE. The trenches are then plugged with spin-on-polymer for further lithography. First, the top layer of polysilicon is patterned into piezoresisistive strips for sensing vibration. Second, a layer of photoresist is patterned for copper lift-off, to create metal interconnects. Last, the planarizing polymer is removed, and the entire surface is coated with silicon nitride. This final coating protects the top surface during an anisotropic backside wet etch, in this case by potassium hydroxide.

Instrumented Suspensions

Finally, microfabrication techniques may be useful for installing sensors on conventional steel suspensions. The concept is to deposit a dielectric and a piezoresistive material on the stainless steel sheet that will be formed into the central piece of the suspension and then to pattern piezoresistive layer directly by photolithography and etching to form strain gages. A second lithography may be used to add metal lines back to the E-block. Vibration sensors on the suspension are very desirable for controlling slider position, and the use of thin-film deposition and photolithography permits the sensors to be located at the points most effective for detecting vibration. An example of a sensor formed by this method is shown in Fig. 32.17.

32.2.6 Actuated Head Fabrication

Several interesting fabrication processes have been proposed for another approach to creating a dual-stage disk drive servo: the actuated head. In an actuated head, the actuator is built into the slider and moves just the read/write head. This could potentially eliminate many of the mechanical limits of actuated suspensions or actuated sliders. Actuation techniques are typically similar to those of actuated sliders, with the key complication being the need to integrate an actuator fabrication process with a slider fabrication process. We discuss ways in which this might be done below.

One fabrication approach to actuating a read/write head is to enclose an electrostatic driving array inside a slider. The slider is built on a glass or silicon substrate, beginning with the air-bearing surface (ABS), followed by a microactuator, and then surrounded by the remainder of the slider. The ABS is formed by silicon oxide deposited over aluminum and tapered photoresist layers that give it a contoured shape. The majority of the ABS is then covered with photoresist, while a block in the center is plated with nickel and etched to form the electrostatic actuator. The electrostatic microactuator, in turn, is covered while the remainder of the slider body is electroplated. The slider may then be bonded to a suspension and the ABS surface released from the substrate by etching away the original aluminum and photoresist surface [32.9].

The sacrificial metal technique described in Sect. 32.2.4 is another possibility for building an electrostatic array onto a slider. It can be used to form small array with very narrow gaps, and the process is compatible with slider and head materials [32.27].

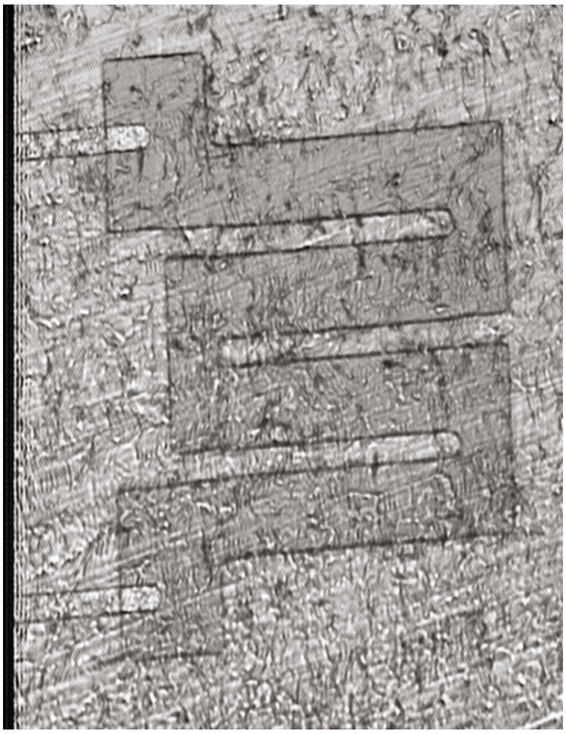

Fig. 32.17 Piezoresistive strain sensor on steel substrate

The third actuated-head process makes use of an SOI wafer and silicon based materials. For this structure, a 20 μm device layer is coated with silicon nitride. The nitride is patterned with contact holes for connections to the substrate and used as insulation over the rest of the device. A layer of molybdenum is sputter deposited to perform interconnections to both the head and the electrostatic array. This array is formed by DRIE to the buried oxide, using a lithographically patterned hard mask of tetraethylorthosilicate. The result is a simple, tiny, actuator that must be bonded to the edge of a slider after release from the SOI wafer [32.28].

32.3 Servo Control Design of MEMS Microactuator Dual-Stage Servo Systems

The objective of disk drive servo control is to move the read/write head to the desired track as quickly as possible, referred to as track seek control, and, once on-track, position the head on the center of the track as precisely as possible, referred to as track-following control, so that data can be read/written quickly and reliably. The implementation of the servo controller relies on the position error signal (PES), which is obtained by reading the position information encoded on the disk's data tracks.

32.3.1 Introduction to Disk Drive Servo Control

The position error is also called track mis-registration (TMR) in the disk drive industry. Major TMR sources in the track-following mode include spindle runout, disk fluttering, bias force, external vibration/shock disturbance, arm and suspension vibrations due to air turbulence, PES noise, written-in repeatable runout, and residual vibration due to seek/settling [32.29]. These sources can be categorized as runout, r, input disturbance, d, and measurement noise, n, by the locations they are injected into the control system, as shown in Fig. 32.18. In Fig. 32.18, $G_P(s)$ and $G_C(s)$ represent the disk drive actuator and the controller, respectively; r and x_p represent track runout and head position, respectively.

From Fig. 32.18, the PES can be written as

$$\mathrm{PES} = S(s)r + S(s)G_P(s)d - S(s)n\,, \tag{32.9}$$

where $S(s)$ is the closed loop sensitivity function defined by

$$S(s) = \frac{1}{1+G_C(s)G_P(s)}\,. \tag{32.10}$$

The higher the bandwidth of the control system, the higher the attenuation of the sensitivity function $S(s)$ below the bandwidth. Thus, one of the most effective methods to reduce PES and increase servo precision is to increase the control system bandwidth.

Traditional disk drive servo systems utilize a single voice coil motor (VCM) to move the head. Multiple structural resonance modes of the E-block arm and the suspension located between the pivot and the head impose a major limitation on the achievable control bandwidth. A dual-stage servo system using a MEMS actuated-slider microactuator can achieve high bandwidth because the microactuator is located between the suspension and the slider, thus by-passing the pivot, E-block arm, and suspension resonance modes.

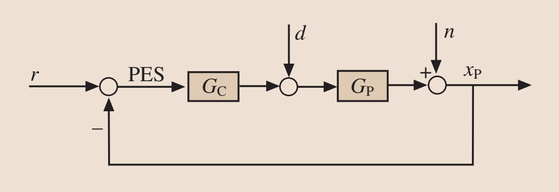

Fig. 32.18 Disk drive servo control

One basic task of dual-stage servo control design is to increase control bandwidth using the second stage actuator. Many design methodologies have been developed to accomplish this objective. In this section, we first review major dual-stage servo control design methodologies, and then we discuss design considerations for controller design of a MEMS microactuator dual-stage servo system. As design examples, the details of track-following control designs using a sensitivity function decoupling design method and the μ-synthesis design method are presented in Sect. 32.3.3. In Sect. 32.3.4, we introduce a short-span seek control scheme using a dual-stage actuator based on decoupled feed forward reference trajectories generations.

32.3.2 Overview of Dual-Stage Servo Control Design Methodologies

Various control design architectures and methodologies have been developed for dual-stage servo control design. They can be largely classified into two categories: those based on decoupled or sequential single-input-single-output (SISO) designs, and those based on modern optimal design methodologies, such as LQG, LQG/LTR, H_∞, and μ-synthesis, in which the dual-stage controllers are obtained simultaneously.

Two constraints must be considered in dual-stage servo control design. First, the contribution from each actuator must be properly allocated. Usually the first stage actuator, or the coarse actuator, has a large moving range but a low bandwidth while the second stage actuator, or the fine actuator, has a high bandwidth but small moving range. Second, the destructive effect, in which the two actuators fight each other by moving in opposite directions, must be avoided.

Classical SISO Design Methodologies

Several architectures and design methodologies have been proposed to transform the dual-stage control design problem into decoupled or sequential multiple SISO compensators design problems, for example, master–slave design, decoupled design [32.30], PQ method [32.31], and direct parallel design [32.32]. Figures 32.19 to 32.22 shows the block diagrams of dual-stage controller designs using these methods. In these figures, G_1 and G_2 represent the coarse actuator (VCM) and fine actuator (microactuator) respectively; x_1 is the position of the coarse actuator; x_r is the position of the fine actuator relative to the coarse actuator; x_p is the total position output; r is the reference input (runout), and PES is the position error, $e = r - x_\mathrm{p}$.

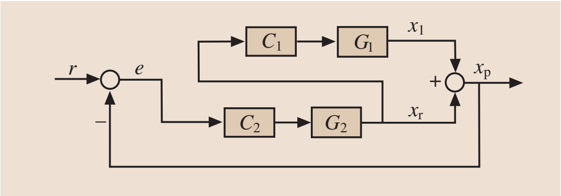

Fig. 32.19 Master–slave design structure

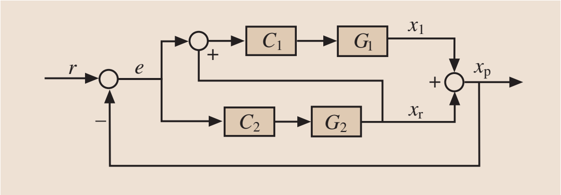

Fig. 32.20 Decoupled control design structure

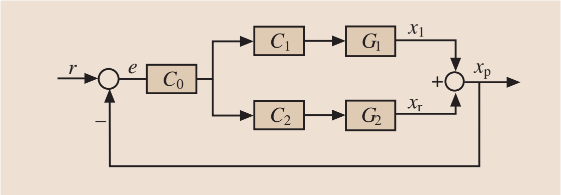

Fig. 32.21 PQ control design structure

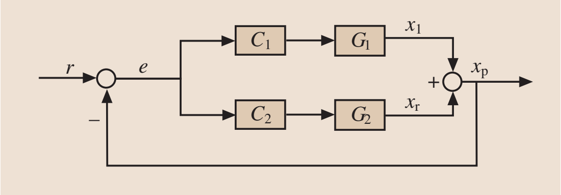

Fig. 32.22 Parallel control design structure

Master–slave Design. In a traditional master–slave structure, the absolute position error is fed to the fine actuator, and output of the fine actuator is fed to the coarse actuator, as shown in Fig. 32.19. The position error will be compensated by the high bandwidth fine actuator. The coarse actuator will follow the fine actuator to prevent its saturation.

Decoupled Design. Figure 32.20 shows the decoupled design structure [32.30], which is similar to the master–slave structure. A summation of the fine actuator output and the position of the coarse actuator, is fed the coarse actuator. A very nice feature of this structure is that the control system is decoupled into two independent control loops, and the total sensitivity function is the product of the sensitivity functions of each of the control loops [32.13]. Thus the two compensators C_1 and C_2 can be designed independently. Decoupled design is also referred to as decoupled master–slave design, or sensitivity function decoupling design. In Sect. 32.3.3, we will discuss details of a track-following controller for MEMS microactuator dual-stage servo system designed using this method.

Both the master–slave and the decoupled designs use the relative position of the fine actuator, x_r. If a relative position sensor is unavailable, x_r can be estimated using a model of the fine actuator.

PQ Design Method. The PQ method is another innovative design technique for control design of dual-input, single-output system [32.31]. A block diagram of a dual-stage control design using this method is shown in Fig. 32.21.

In PQ design, P is defined by

$$P = \frac{G_1}{G_2} , \tag{32.11}$$

and a dual-stage controller can be designed in two steps. The first involves the design of an auxiliary compensator Q for plant P, which is defined by

$$Q = \frac{C_1}{C_2} . \tag{32.12}$$

Q is designed to parameterize the relative contribution of the coarse and fine actuators. The 0 dB crossover frequency and phase margin of the open loop transfer function PQ are the design parameters in the design of Q. At frequencies below the 0 dB crossover frequency of PQ, the output is dominated by the coarse actuator, while at frequencies above the 0 dB crossover frequency, the output is dominated by the fine actuator. At the 0 dB crossover frequency, the contributions from the two actuators are equal. A large phase margin of PQ will ensure that the two actuators will not fight each other when their outputs are close in magnitude, thus avoiding any destructive effects.

The second step in the PQ design methodology is to design a compensator C_0 for SISO plant PQ such that the bandwidth (crossover frequency), gain margin, phase margin, and error rejection requirements of the overall control system are satisfied.

Direct Parallel Design. It is also possible to design the dual-stage controller directly using a parallel structure, as shown in Fig. 32.22, by imposing some design constraints and by sequential loop closing [32.32].

The two constraints for parallel design in terms of the PES open loop transfer functions are [32.32]:

$$C_1(s)G_1(s)+C_2(s)G_2(s) \to C_2(s)G_2(s)\ , \quad (32.13)$$

at high frequencies and

$$|C_1(s)G_1(s)+C_2(s)G_2(s)| \gg |C_2(s)G_2(s)|\ , \quad (32.14)$$

at low frequencies. The first constraint implies that open loop frequency response of the dual-stage control system at high frequencies approximately equals that of the fine actuator control loop. Thus the compensator $C_2(s)$ can be first designed independently as a SISO design problem to satisfy the bandwidth, gain margin, and phase margin requirements of the dual-stage control system. The second constraint ensures that the fine actuator will not be saturated. Compensator $C_1(s)$ can then be designed for the SISO plant model with the fine actuator control loop closed, such that the low frequency constraint and overall stability requirement are satisfied. This model is defined by:

$$G(s)=\frac{G_1(s)}{1+C_2(s)G_2(s)}\ . \quad (32.15)$$

A dual-stage controller for a MEMS dual-stage actuator servo system has been designed using this method and implemented at IBM. An open loop gain crossover frequency of 2.39 kHz with gain margin 5.6 dB and phase margin 33° was obtained. In experimental testing, a 1-σ TMR of 0.024 µm has been achieved [32.32].

Modern MIMO Design Methodologies

Since the dual-stage actuator servo system is a MIMO system, it is natural to utilize modern MIMO optimal design methodologies, such as LQG, LQG/LTR, H_∞, and µ-synthesis, to design the dual-stage controller. Usually MIMO optimal designs are based on the parallel structure shown in Fig. 32.22, augmented with noise/disturbances models and other weighting functions to specify the control design performance objectives.

Linear quadratic Gaussian (LQG) control combines a Kalman filter and optimal state feedback control based on the separation principle. However, the Kalman filter weakens the desirable robustness properties of the optimal state feedback control. Linear quadratic Gaussian/loop-transfer recovery (LQG/LTR) control recovers robustness by a Kalman filter redesign process. Examples of dual-stage control designs using LQG and LQG/LTR have been reported in [32.33–35].

The LQG design methodology minimizes the H_2 norm of the control system, while the H_∞ design methodology minimizes the H_∞ norm of the control system. µ-synthesis design methodology is based on the H_∞ design and accounts for plant model uncertainties during the controller synthesis process with guaranteed robustness. Examples of dual-stage control designs based on the H_∞ and µ-synthesis design methodologies have been reported in [32.33, 36, 37]. A control design using µ-synthesis for the MEMS microactuator dual-stage servo system will be presented in Sect. 32.3.3

Other advanced control theories also have been applied to dual-stage servo control designs, such as sliding mode control [32.38] and neural networks [32.39].

32.3.3 Track-Following Controller Design for a MEMS Microactuator Dual-Stage Servo System

MEMS microactuators for dual-stage servo application are usually designed to have a single flexure resonance mode between 1–2 kHz [32.8,40]. This resonance mode is usually very lightly damped and can have a ±15% variations due to the differences in the fabrication processes. Since the uncertain resonance frequency is relatively low and close to the open loop gain crossover frequency of the control system, robustness to these variations must be considered in the controller design.

Controller Design Considerations for MEMS Microactuator Dual-stage Servo System

Capacitive sensing can be utilized in MEMS microactuators to measure the position of the microactuator relative to the VCM actuator [32.14, 40]. However, this requires additional sensing electronics and wires to and from the head gimbal assembly (HGA), which result in additional fabrication and assembly costs. Thus, whether or not a relative position sensor will be used in MEMS dual-stage servo systems is still an open question. If the microactuator's resonance mode is lightly damped, it is susceptible to airflow turbulence and external disturbances. The relative position signal can be utilized to damp the microactuator's resonance mode. Dual-stage control can be classified as MIMO or SIMO designs according to the availability of the relative position signal.

Given the fact that the inertia of the MEMS microactuator is very small compared to that of the VCM, the effect of the motion of the microactuator on the VCM can be neglected. By feeding the control input of the VCM to the input of the microactuator with a proper gain, we can cancel the coupling effect of the VCM on the microactuator [32.13]. Thus the dual-stage control model can be decoupled as the sum of the outputs of the VCM and the microactuator, as shown in Fig. 32.23, in which G_V, G_M are the VCM and microactuator models respectively, and x_p is the absolute position of the read/write head, which equals the absolute position of the VCM, x_v, and the position of the microactuator relative to the VCM, RPES.

Decoupled Track-Following Control Design and Self-Tuning Control

Figure 32.24 shows a block diagram used for decoupled control design of a MEMS microactuator dual-stage servo systems [32.13].

Sensitivity Function Decoupling Design Methodology. The part enclosed in the dashed box on the upper-right corner is the dual-stage plant model shown in Fig. 32.23. In the figure, G_V and G_M are the VCM and microactuator models, respectively; x_P is the absolute position of the read/write head; x_v is the absolute position of the VCM actuator; and RPES is the position of the microactuator relative to the VCM; r represents the track runout, and PES is the position error signal of the head relative to the data track.

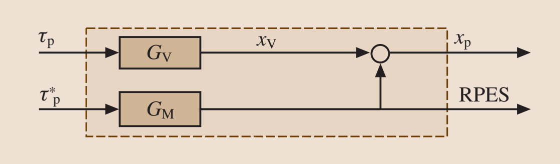

Fig. 32.23 Dual-stage control plant

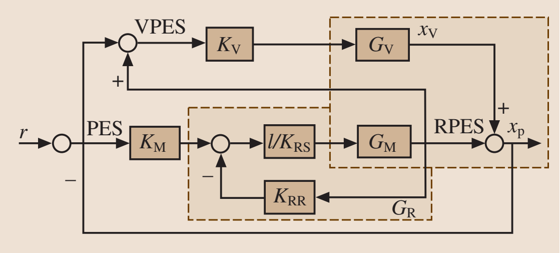

Fig. 32.24 Decoupled dual-stage control design block diagram

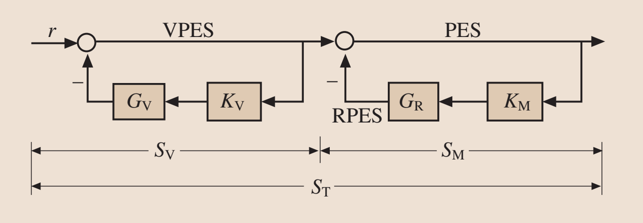

Fig. 32.25 The sensitivity block diagram

The decoupling control approach, originally introduced in [32.30], utilizes the PES and RPES signals to generate the position error of the VCM relative to the data track center, labeled as $VPES$,

$$VPES = \text{PES} + \text{RPES} = r - x_v , \qquad (32.16)$$

and this signal is fed to the VCM loop compensator.

As shown in the block diagram, three compensators need to be designed: the VCM loop compensator K_V; the microactuator PES loop compensator K_M; and the microactuator RPES minor loop compensator $K_{MM} = K_{RR}/K_{RS}$ is used to damp the microactuator's flexure resonance mode and place the closed loop poles of the microactuator RPES loop at appropriate locations. The damped microactuator closed loop transfer function G_R, shown in the lower-middle dashed box, is defined by

$$G_R = \frac{G_M}{K_{RS} + G_M K_{RR}} . \qquad (32.17)$$

The total dual-stage open loop transfer function from r to x_P, G_T, is

$$G_T = K_V G_V + K_M G_R + K_M G_R K_V G_V . \qquad (32.18)$$

It can be shown that the block diagram in Fig. 32.24 is equivalent to the sensitivity block diagram shown in Fig. 32.25, and the total closed loop sensitivity function from r to PES equals the product of the VCM and microactuator loop sensitivities, S_V and S_M, respectively:

$$S_T = \frac{1}{1 + G_T} = S_V S_M , \qquad (32.19)$$

where

$$S_V = \frac{1}{1 + K_V G_V} , \; S_M = \frac{1}{1 + K_M G_R} . \qquad (32.20)$$

Thus the dual-stage servo control design can be decoupled into two independent designs: the VCM loop and the microactuator loop designs. The VCM loop sensitivity can be designed using traditional single-stage servo design methodologies. The microactuator loop sensitivity is designed to expand the bandwidth and increase the attenuation of low frequency runout and disturbances.

Closed Loop Sensitivity Design by Pole Placement. For a typical sampled second order system with the discrete time transfer function

$$G\left(q^{-1}\right)=\frac{q^{-1}B_o\left(q^{-1}\right)}{A_o\left(q^{-1}\right)}\,, \tag{32.21}$$

where q^{-1} is one step delay operator; $B_o\left(q^{-1}\right)$ is the plant zero polynomial, and $A_o\left(q^{-1}\right)=1+a_1q^{-1}+a_2q^{-2}$ is the plant pole polynomial. The closed loop sensitivity function can be designed by pole placement, solving the following Diophantine equation [32.41]:

$$\begin{aligned}A_{\mathrm{c}}\left(q^{-1}\right)&=A_o\left(q^{-1}\right)S\left(q^{-1}\right)\\&\quad+q^{-1}B_o\left(q^{-1}\right)R\left(q^{-1}\right)\,,\end{aligned} \tag{32.22}$$

where $A_{\mathrm{c}}\left(q^{-1}\right)$ is the desired closed loop characteristic polynomial. The closed loop sensitivity function, G_{S}, is

$$G_{\mathrm{S}}\left(q^{-1}\right)=\frac{A_o\left(q^{-1}\right)\,S\left(q^{-1}\right)}{A_{\mathrm{c}}\left(q^{-1}\right)}\,, \tag{32.23}$$

and the discrete time controller $C\left(q^{-1}\right)$ is

$$C\left(q^{-1}\right)=\frac{R\left(q^{-1}\right)}{S\left(q^{-1}\right)}\,. \tag{32.24}$$

$A_{\mathrm{c}}\left(q^{-1}\right)$ can be chosen such that the required bandwidth and system response are satisfied. Usually it is more intuitive to describe it with its continuous time equivalent parameters: the damping ratio, ζ, and the natural frequency, ω_n. The damping ratio is directly related to the phase margin of the open loop transfer function. For a typical design, it can be chosen to be equal to or greater than one, in order to ensure adequate phase margin. ω_n is related to the control system's bandwidth, which is limited by sampling frequency and time delay, or the high frequency structural dynamics of the system.

Dual-Stage Sensitivity Function Design. The compensators in the dual-stage servo system depicted in Fig. 32.24 can be designed by a two-step design process, as illustrated in Fig. 32.26.

First, the VCM loop compensator K_{V} is designed to attain a desired VCM closed loop sensitivity S_{V}, as shown in the top part of Fig. 32.26. Its bandwidth, ω_{VC} in Fig. 32.26, is generally limited by the E-block and suspension resonance modes. The design of this compensator can be accomplished using conventional SISO frequency shaping techniques.

The second step of the design process involves the design of the microactuator loop compensators to attain additional attenuation S_{M}, as shown in the middle part of Fig. 32.26. This step is itself accomplished in two steps. First, the minor RPES loop compensator, $K_{\mathrm{RR}}/K_{\mathrm{RS}}$, is determined in order to damp the microactuator resonance mode and place the poles of G_{R}, or equivalently ω_{MR} in Fig. 32.26, at desired locations. The compensator can be obtained by solving a Diophantine equation by pole placement as discussed in the previous subsection. The poles of G_{R} will become the zeros of the microactuator loop sensitivity function, S_{M}.

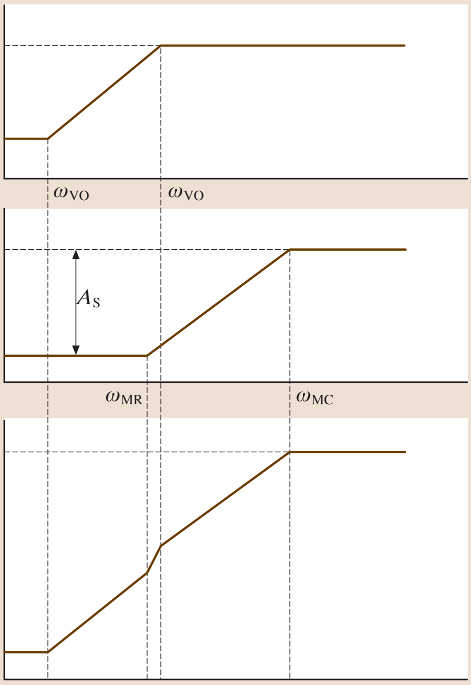

Fig. 32.26 Illustration of dual-stage sensitivity S_T design

Finally, the PES loop compensator K_{M} is designed to place the poles, or equivalently. ω_{MC} in Fig. 32.26, of the microactuator loop closed loop sensitivity S_{M}. ω_{MC} is limited by the PES sampling frequency and computational time delay [32.42].

The total dual-stage sensitivity is shown in the bottom part of Fig. 32.26. For a given ω_{MC}, the additional attenuation A_S, provided by the microactuator loop, will be determined by ω_{MR}. In our proposed procedure, the initial value of ω_{MR} can be chosen to be the same as ω_{VC}. It is then adjusted so the desired attenuation and phase margin requirements of the overall dual-stage system are satisfied. Decreasing ω_{MR} increases the low frequency attenuation of the closed loop sensitivity function S_{T} but also generally reduces the phase margin of open loop transfer function G_{T}.

For SIMO design when the RPES is not available, the microactuator model can be used as an open loop observer to estimate the RPES. The combination of the open loop observer and the minor loop compensator is equivalent to a notch filter.

Self-Tuning Control to Compensate Variations in the Microactuator's Resonance Mode. An adaptive control scheme can be combined with the decoupled discrete time pole placement design methodology described above to compensate variations in the microactuator's resonance mode by tuning the microactuator RPES inner loop compensator.

A block diagram for the microactuator inner loop self-tuning control is shown in Fig. 32.27. The parameter adaptation algorithm (PAA) is a direct self-tuning algorithm, based on the microactuator inner loop pole placement design.

Consider the microactuator open loop transfer function in (32.21). Since the microactuator's resonance mode is lightly damped, the zero of the microactuator's discrete time transfer function is very close to 1 ($z_o \approx 1$). Thus it is possible to factor out the "known" term $(1+z_o q^{-1})$ from the Diophantine equation (32.22). The resulting RPES minor-loop closed loop dynamics is given by

$$A_{\mathrm{c}}\left(q^{-1}\right) y(k) = q^{-1} b_0 \left(1+z_o q^{-1}\right) \times\left[K_{\mathrm{RS}}\left(q^{-1}\right) u_{\mathrm{M}}(k) + K_{\mathrm{RR}}\left(q^{-1}\right) y(k)\right],$$

where u_{M} is the control input to the microactuator; y denotes the position of the microactuator relative to the VCM, i.e. $y = \mathrm{RPES}$. Defining:

$$S\left(q^{-1}\right) = b_0 K_{\mathrm{RS}}\left(q^{-1}\right) = s_0 + s_1 q^{-1}, \qquad (32.25)$$

$$R\left(q^{-1}\right) = b_0 K_{\mathrm{RR}}\left(q^{-1}\right) = r_0 + r_1 q^{-1}, \qquad (32.26)$$

the regressor vector, $\phi(k)$ and filtered regressor vector, $\phi_{\mathrm{f}}(k)$,

$$\phi(k) = \left[u_{\mathrm{M}}(k)\ u_{\mathrm{M}}(k-1)\ y(k)\ y(k-1)\right], \qquad (32.27)$$

$$A_{\mathrm{c}}\left(q^{-1}\right)\phi_{\mathrm{f}}(k) = \left(1+z_o q^{-1}\right)\phi(k), \qquad (32.28)$$

and the controller parameter vector $\theta = [s_0\ s_1\ r_0\ r_1]^T$, the closed loop RPES dynamics (32.25) can be rewritten as

$$y(k) = \theta^T \phi_{\mathrm{f}}(k-1). \qquad (32.29)$$

From (32.29), the controller parameter vector estimate $\hat{\theta}(k) = [\hat{s}_0(k)\ \hat{s}_1(k)\ \hat{r}_0(k)\ \hat{r}_1(k)]^T$ can be updated using a standard recursive least square algorithm (RLS) [32.41]:

$$\hat{\theta}(k) = \hat{\theta}(k-1) + P(k)\phi_{\mathrm{f}}(k)e^o(k), \qquad (32.30)$$

$$e^o(k) = y(k) - \hat{\theta}^T(k-1)\phi_{\mathrm{f}}(k-1), \qquad (32.31)$$

$$P(k) = \left[P(k-1) - \frac{P(k-1)\phi_{\mathrm{f}}(k-1)\phi_{\mathrm{f}}^T(k-1)P(k-1)}{1+\phi_{\mathrm{f}}^T(k-1)P(k-1)\phi_{\mathrm{f}}(k-1)}\right], \qquad (32.32)$$

The control law is

$$\hat{S}\left(k,q^{-1}\right)u_{\mathrm{M}}(k) = \hat{S}_o(k)u_{\mathrm{M}}(k) - \hat{R}\left(k,q^{-1}\right)y(k), \qquad (32.33)$$

with

$$\hat{S}\left(k,q^{-1}\right) = \hat{s}_0(k) + \hat{s}_1(k)q^{-1}, \qquad (32.34)$$

$$\hat{R}\left(k,q^{-1}\right) = \hat{r}_0(k) + \hat{r}_1(k)q^{-1}, \qquad (32.35)$$

and $u_{\mathrm{M}}(k)$ being the output of the microactuator fixed outer loop compensator K_{M}, $u_{\mathrm{M}}(k)$ being the control input to the microactuator.

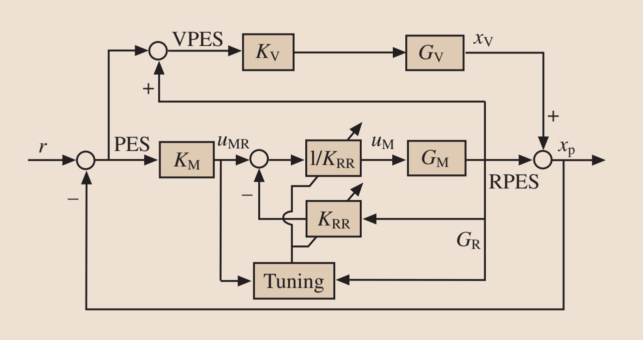

Fig. 32.27 Self-tuning control of the microactuator

Dual-Stage Track-Following Control Design Using μ-Synthesis

The structured singular value, μ, is a measure of how big a perturbation to a system needs to be to make the system unstable. By incorporating a fictitious uncertainty block, robust performance in terms of the H_∞ norm of the closed loop transfer function can be related to the value of μ. μ-synthesis is a robust optimal controller design technique that attempts to minimize μ through an iterative process [32.43].

Design Methodology. For controller design using μ-synthesis, the model uncertainties are represented using linear fractional transformations (LFT). Disturbances and outputs are weighted to characterize the real plant environment and the performance requirements. A block diagram used for dual-stage μ-synthesis controller design is shown in Fig. 32.28.

In the block diagram, two model uncertainties, δ_1 and δ_2, are considered. δ_1 is an additive uncertainty used to describe the VCM unmodeled dynamics. δ_2 represents the parametric uncertainty in the stiffness of the microactuator.

Disturbance signals accounted for in the model include the track runout r, input disturbances to the VCM d_{VCM}, input disturbances to the microactuator d_{MA}, PES sensor noise n_{PES}, and the $RPES$ sensor noise n_{RPES}. These disturbance signals are generated by passing normalized signals $\bar{r}$, $\overline{d_{\mathrm{VCM}}}$, $\overline{d_{\mathrm{MA}}}$, $\overline{n_{PES}}$, and $\overline{n_{RPES}}$ through weighting functions W_r, $W_{d\mathrm{VCM}}$, $W_{d\mathrm{MA}}$, W_{nPES}, and W_{nRPES}, respectively, which can be either constants or frequency shaping filters. These weights are selected by the designer so that disturbances are modeled with sufficient fidelity.

The output signals in the synthesis model are head position error signal, PES, the microactuator relative position signal, $RPES$, the VCM control input u_{VCM}, and the microactuator control input u_{MA}. These signals are multiplied by scaling factors, W_{PES}, W_{RPES}, $W_{u\mathrm{VCM}}$, and $W_{u\mathrm{MA}}$, respectively, to produce the weighted performance signals $\overline{PES}$, $\overline{RPES}$, $\overline{u_{\mathrm{VCM}}}$, $\overline{u_{\mathrm{MA}}}$. The scaling factors are selected to characterize the performance requirements.

Given a set of input and output weights and plant uncertainties, if the synthesized controller achieves

$$\mu \le \beta\,, \tag{32.36}$$

the closed loop transfer function, $\overline{T}$, from the normalized disturbances

$$\bar{d} = \left[\bar{r}\ \overline{d_{\mathrm{VCM}}}\ \overline{d_{\mathrm{MA}}}\ \overline{n_{\mathrm{PES}}}\ \overline{n_{\mathrm{RPES}}}\right], \tag{32.37}$$

to the weighted performance signals

$$\bar{e} = \left[\overline{\mathrm{PES}}\ \overline{\mathrm{RPES}}\ \overline{u_{\mathrm{VCM}}}\ \overline{u_{\mathrm{MA}}}\right], \tag{32.38}$$

will have infinity norm

$$\|\overline{T}\|_\infty \le \beta\,, \tag{32.39}$$

for perturbations

$$\|\Delta\|_\infty = \left\| \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix} \right\|_\infty \le \frac{1}{\beta}\,. \tag{32.40}$$

The interpretation of H_∞ as the RMS gain of the sinusoidal signals to help understand the design is as follows. Assume that each element $\overline{d_i}$ of the disturbance

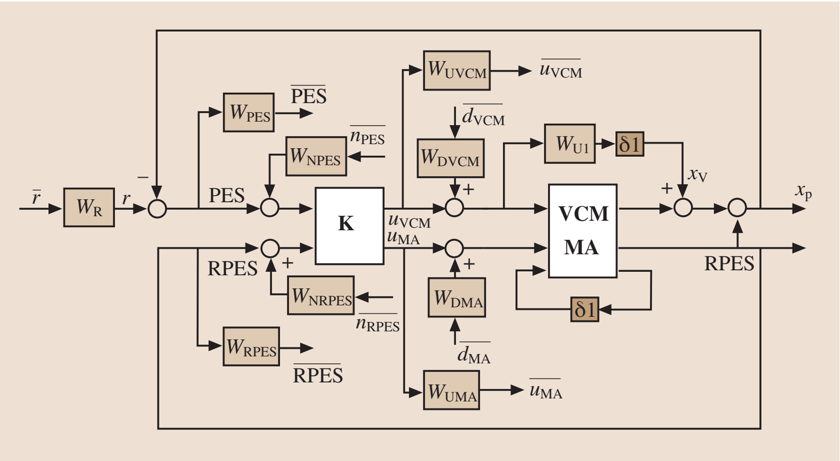
W_{UVCM}
$\overline{u_{\mathrm{VCM}}}$
$\overline{\mathrm{PES}}$
W_{PES}
$\overline{d_{\mathrm{VCM}}}$
$\overline{n_{\mathrm{PES}}}$
W_{U1}
δ1
W_{NPES}
W_{DVCM}
x_{V}
−
+
+
$\bar{r}$
W_{R}
r
PES
K
VCM
x_{p}
u_{VCM}
u_{MA}
MA
RPES
+
+
RPES
W_{NRPES}
W_{DMA}
δ1
$\overline{n_{\mathrm{RPES}}}$
W_{RPES}
$\overline{d_{\mathrm{MA}}}$
$\overline{\mathrm{RPES}}$
W_{UMA}
$\overline{u_{\mathrm{MA}}}$

Fig. 32.28 μ-synthesis design block diagram input vector in (32.37) is a sinusoid of the form

$$\overline{d_i}(t) = D_i \sin(\omega_i t + \psi_i) , \tag{32.41}$$

such that

$$\sum_{i=1}^{5} D_i^2 \le 1 . \tag{32.42}$$

Then the steady state response of output in (32.38) will also be a sinusoid of the form

$$\overline{e_i}(t) = E_i \sin(\omega_i t + \phi_i) , \tag{32.43}$$

and

$$\sum_{i=1}^{4} E_i^2 \le \beta , \tag{32.44}$$

under the perturbations defined in (32.40).

Weighting Functions for µ-Synthesis Design. Performance weights are chosen based on limits that the error signals should not exceed. The limit on the PES is based on a rule of thumb commonly used in the disk drive industry which states that the position error of the head should not exceed 1/10 of a track width during track following. RPES is limited by the stroke limit of the microactuator. Thus,

$$W_{\text{PES}} = \frac{10}{\text{track width}} ; \quad W_{\text{RPES}} = \frac{1}{\text{stroke limit}} . \tag{32.45}$$

Weights on control inputs are based on the control input saturations.

$$W_{U\text{VCM}} = \frac{1}{(i_{\text{V}})_{\max}} ; \quad W_{U\text{MA}} = \frac{1}{(v_{\text{M}})_{\max}} , \tag{32.46}$$

where $(i_{\text{V}})_{\max}$ and $(v_{\text{M}})_{\max}$ are the maximum drive current of the VCM and the maximum drive voltage of the microactuator, respectively.

The weights for PES noise, n_{PES}, and the disturbance to the VCM, $w_{d\text{VCM}}$, were determined by extrapolating the PES decomposition results presented in [32.29]. The weight of the RPES sensing noise, n_{RPES}, is assumed to be 8 nm [32.14]. The disturbance weight, $w_{d\text{MA}}$, is assumed to be 0.1 V in the design.

The runout weighting function is a low pass filter that captures the combined effects of track runout, mechanical vibrations, and bias force disturbance to PES,

$$W_{\text{r}} = \frac{2\times 10^{-8}(s+2\pi\times 10)^2}{(s+2\pi\times 1000)^2} . \tag{32.47}$$

The combination of W_{PES} and W_{r} determines the closed loop sensitivity function from r to PES, $S(s)$, of the dual-stage system. If the designed control system achieves a peak μ value of β, we have

$$\left\| W_{\text{PES}} S(s) W_{\text{r}} \right\|_\infty \le \beta . \tag{32.48}$$

If $\beta = 1$, the magnitude of bode plot of $S(s)$ lies below that of $1/W_{\text{PES}} W_{\text{r}}$.

For SIMO design, the controller to be synthesized has only one input. In the synthesis model, RPES is not fed to the controller K, and RPES noise n_{rpes} and its weights W_{nRPES} are not used.

Design and Simulation Results

Figure 32.29 shows the Bode plots of the open loop transfer functions from r to $x_{\text{p}}p$ and the closed loop sensitivity functions of the two designs, respectively. Sampling frequency was chosen to be 20 kHz in both designs.

For the decoupled MIMO design, the gain crossover frequency, gain margin, phase margin of the open loop system are 2,201 Hz, 9.1 dB, and 48.8°, respectively. For the decoupled SIMO design, they are 2,381 Hz, 8.7 dB, and 30.6°, respectively.

For µ-synthesis MIMO design, the gain crossover frequency, gain margin, phase margin of the open loop system are 2,050 Hz, 9.5 dB, 66.0°, respectively. For µ-synthesis SIMO design, they are 2,223 Hz, 2.9 dB, 32.0°, respectively.

The above controller designs have not been implemented on a MEMS microactuator dual-stage servo system. Controllers designed using both methods have been tested on a PZT actuated suspension dual-stage servo system, however, validating their effectiveness [32.44].

Figure 32.30 shows the simulation of control parameters adaptation for the self-tuning controller, when the real microactuator resonance frequency is 1.2 times its nominal value. The controller parameters converge to their desired values. Similar responses were obtained when the real resonance frequency is 0.8 times the nominal value.

32.3.4 Dual-Stage Seek Control Design

Because the inertia of a microactuator is much smaller than that of the VCM, it can produce a larger acceleration and move faster than the VCM. The motion range of a microactuator is usually limited to a few micrometers, however. Thus the performance improvement in seek control made possible by using

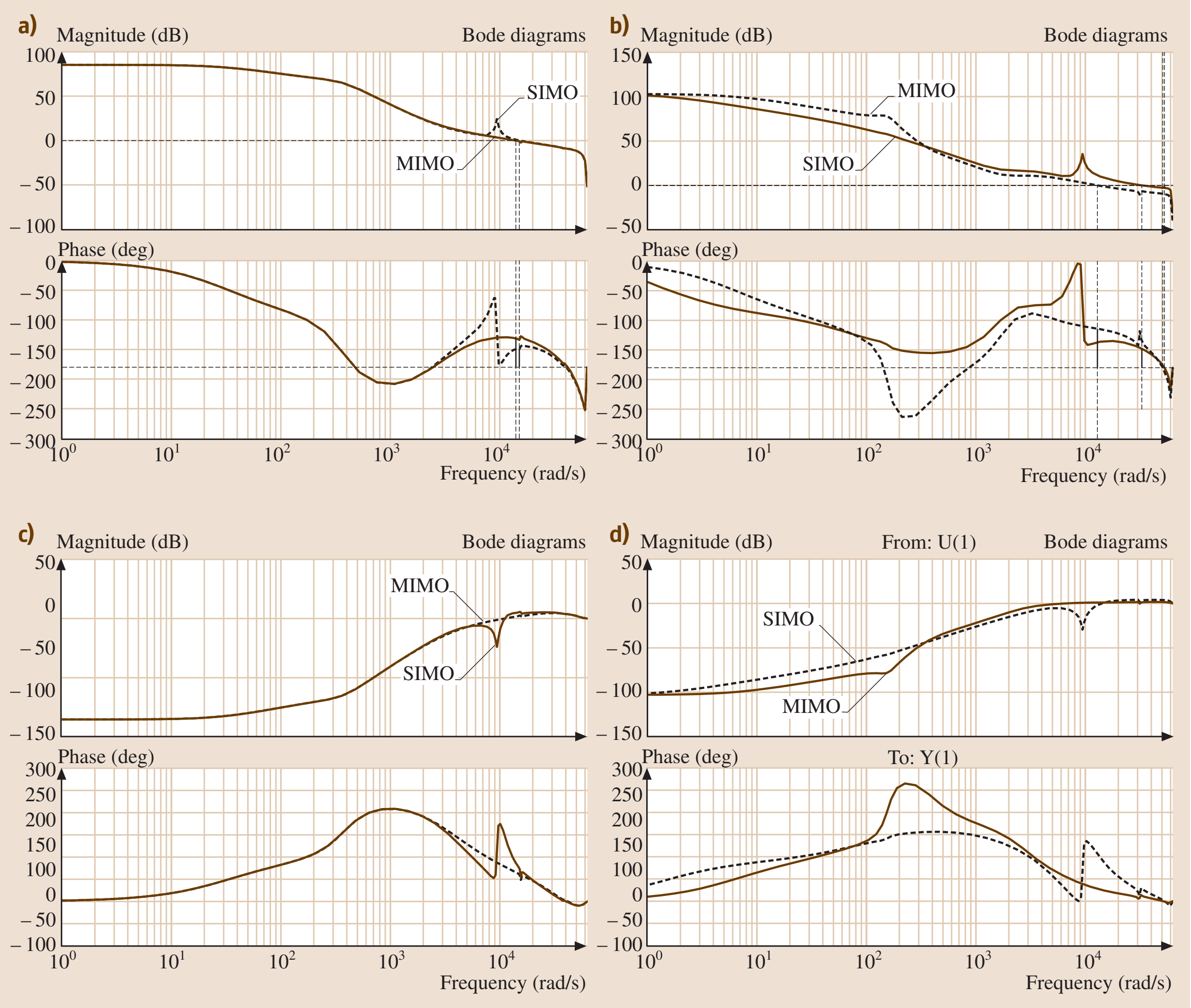

Fig. 32.29 **(a)** Open loop Bode plot of the decoupled design; **(b)** open loop Bode plot of the μ-synthesis design; **(c)** closed loop sensitivity Bode plot of the decoupled design; **(d)** closed loop sensitivity Bode plot of the μ-synthesis design

dual-stage actuation will mainly be in short distance seeks.

Two degree-of-freedom (DOF) position control has been a very popular control technique in short distance seeks. A 2-DOF control technique utilizing decoupled feed forward reference trajectories has been developed for short span seek control using a PZT actuated suspension dual-stage servo system [32.45]. The same technique can also be applied to seek control of MEMS microactuator dual-stage servo system.

Figure 32.31 shows a block diagram for dual-stage short span seek control design using this method. In the figure, x_{V}^{d} and x_{R}^{d} are the desired seek trajectories of the VCM and the microactuator respectively. K_{VF} and K_{RF} are Zero Phase Error Tracking Feed Forward Controllers (ZPETFFC) generated using with the VCM model G_{V} and the damped microactuator model G_{R} [32.46].

To minimize the residual vibration after seek operation, minimum jerk seek trajectories were applied to both the VCM and the microactuator. These can be generated

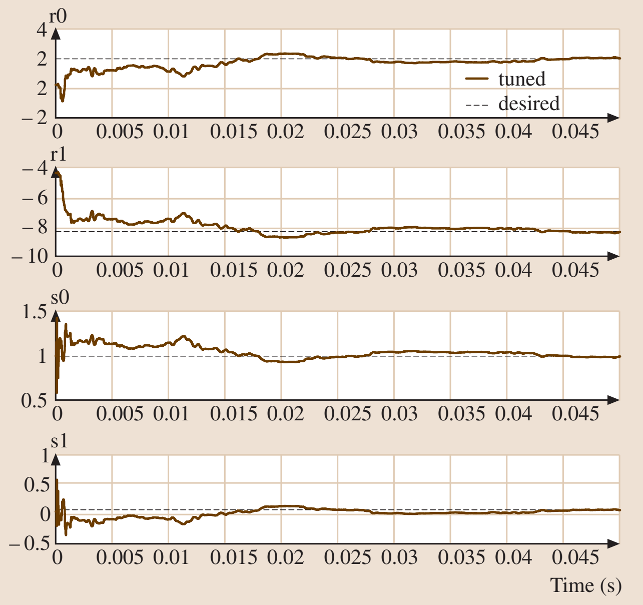

Fig. 32.30 Control parameters adaptation response

by [32.47]

$$x_{\mathrm{V}}^{d}(t)=60\,d_s\left[\frac{1}{10}\left(\frac{t}{T_{\mathrm{V}}}\right)^5-\frac{1}{4}\left(\frac{t}{T_{\mathrm{V}}}\right)^4+\frac{1}{6}\left(\frac{t}{T_{\mathrm{V}}}\right)^3\right], \tag{32.49}$$

$$x_{\mathrm{R}}^{d}=\begin{cases}60\,\left(d_s-x_{\mathrm{V}}^{d}(T)\right)\\ \times\left[\frac{1}{10}\left(\frac{t}{T}\right)^5-\frac{1}{4}\left(\frac{t}{T}\right)^4+\frac{1}{6}\left(\frac{t}{T}\right)^3\right] & t\le T\,,\\ d_s-x_{\mathrm{V}}^{d}(t) & t>T\end{cases} \tag{32.50}$$

where d_s is the distance of the head from the target track; T is the time when the head reaches the target track if dual-stage actuator is used, while T_{V} is the time when the head reaches the target track if only VCM is used. T_{V} and T can be chosen based on the control force saturation and seek performance requirements.

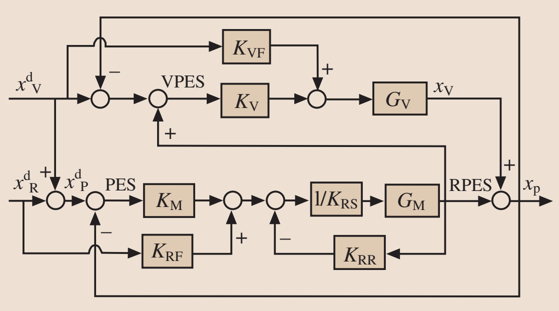

Fig. 32.31 Dual-stage seek control design

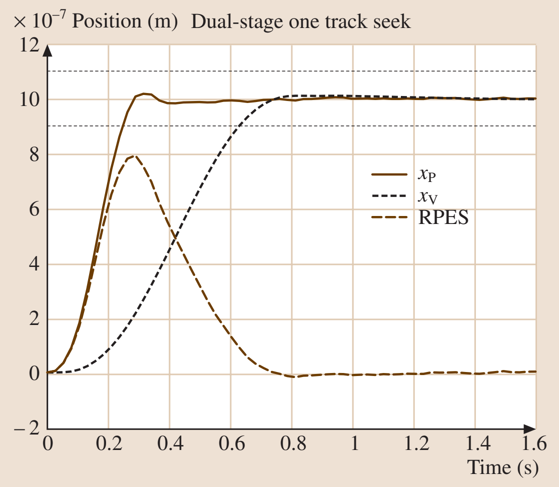

Fig. 32.32 Dual-stage short span seek response

Figure 32.32 shows the 1 µm seek responses of the dual-stage actuator for the MIMO design. Overshoot was eliminated, and no obvious residual vibrations occured. The seek time using the dual-stage actuator is about 0.25 ms, compared to a seek time of about 0.7 ms if only the VCM is used.

32.4 Conclusions and Outlook

In order to increase hard disk drive storage density, high bandwidth dual-stage servo systems are necessary to suppress disturbances and increase servo precision. Various prototype MEMS microactuators have been designed and fabricated to provide a dual-stage actuation system. The most common approach for these prototypes is to use electrostatic driving arrays produced by MEMS fabrication methods.

MEMS microactuators provide a potentially high performance and low cost solution for achieving the servo requirements for extremely high HDD storage density. Simulations and experiments show that many

MEMS microactuator features, such as integrated silicon gimbals and capacitive relative position sensing arrays, can meet important requirements for operation of HDDs. The use of structural electrical-isolation in the microactuator design could considerably reduce microactuator driving voltages, while processes that avoid the use of silicon-on-insulator wafers or high-temperature processing steps could reduce manufacturing costs. High bandwidth dual-stage track-following controllers have been designed using both a decoupled SISO design method and robust μ-synthesis design method, among others. In addition, short-span seeking control using a 2-DOF control structure with decoupled feed forward trajectory generation can greatly reduce the short span seek time.

Research remains to be done in the areas of system integration, reliability, and performance enhancement for this technology to be utilized in commercial products. Cost and reliability are probably the most important obstacles to commercial implementation of actuated slider dual-stage servo systems at this time. Streamlining of microactuator fabrication processes or development of a process compatible with that of the slider and head fabrication could reduce manufacturing costs, which are not quite yet economical. Meanwhile, dynamic behavior and reliability of the microactuator under disturbances from airflow, head-disk interaction, and particle presence are being experimentally studied by industry.

Further in the future, to achieve nanometer servo precision, third-generation dual-stage servo systems will have to employ an actuated head approach. Research in this area has just started, with the key problem being how to combine microactuator fabrication with read/write head fabrication. Another area of research is the use of MEMS technology to incorporate additional sensors, such as accelerometers and strain gauge vibration sensors, to suppress the TMR due to airflow and external disturbances excited structural vibrations. New robust, adaptive MIMO control architectures and algorithms must be developed for such a multisensing and multiactuation servo system.

Besides the application of MEMS microactuators in dual-stage servo control of traditional rotating media magnetic HDDs, recent advances in MEMS and nanotechnology have made possible the development of a new class of miniaturized ultra-high density storage devices that use probe arrays for recording and reading data. This new architecture abandons the traditional rotating media and flying slider paradigm in favor of parallel x-y scanning of entire probe arrays over a storage medium. The probe arrays and/or the media are moved with electrostatic or electromagnetic microactuators that generate x-y in-plane relative motion, as well as z-axis out-of-plane relative motion. One example of this new generation of storage devices is IBM's "millipede," which utilizes atomic force microscopy (AFM) recording technology [32.48]. Data is recorded on a polymer media by tiny depressions melted by a thermomechanical process by the AFM tips. 400 Gb/in.^{2} bit patterns have been demonstrated using thermal probes.

By any of the paths above, MEMS and nanotechnology will become a basic enabling technology for the development of future data storage devices, offering the capabilities of small size, low power consumption, ultra-high densities, low cost, and high performance.

References

32.1 T. Yamaguchi: Modelling and control of a disk file head-positioning system, Proc. Inst. Mech. Eng. I (J. Syst. Cont. Eng.) **215** (2001) 549–567

32.2 T. Howell, R. Ehrlich, M. Lippman: TPI growth is key to delaying superparamagnetism's arrival, Data Stor. (1999) 21–30

32.3 R. B. Evans, J. S. Griesbach, W. C. Messner: Piezoelectric microactuator for dual-stage control, IEEE Trans. Magn. **35** (1999) 977–981

32.4 I. Naniwa, S. Nakamura, S. Saegusa, K. Sato: Low voltage driven piggy-back actuator of hard disk drives, IEEE International MEMS 99 Conference (IEEE, Piscataway 1999) 49–52

32.5 D. Horsley: Microfabricated electrostatic actuators for magnetic disk drives. Ph.D. Thesis (University of California, Berkeley 1998)

32.6 L. Muller: Gimballed electrostatic microactuators with embedded interconnects. Ph.D. Thesis (University of California, Berkeley 2000)

32.7 S. Koganezawa, K. Takaishi, Y. Mizoshita, Y. Uematsu, T. Yamada: Development of integrated piggyback milli-actuator for high density magnetic recording, International Conference on Micromechatronics for Information and Precision Equipment 1997 (IEEE, New York 1997) 20–23

32.8 L.-S. Fan, T. Hirano, J. Hong, P. R. Webb, W. H. Juan, W. Y. Lee, S. Chan, T. Semba, W. Imaino,

T. S. Pan, S. Pattanaik, F. C. Lee, I. McFadyen, S. Arya, R. Wood: Electrostatic microactuator and design considerations for hdd application, IEEE Trans. Magn. **35** (1999) 1000–1005

32.9 T. Imamura, M. Katayama, Y. Ikegawa, T. Ohwe, R. Koishi, T. Koshikawa: MEMS-based integrated head/actuator/slider for hard disk drives, IEEE/ASME Trans. Mechatron. **3** (1998) 166–174

32.10 T. Imamura, T. Koshikawa, M. Katayama: Transverse mode electrostatic microactuator for MEM-based HDD slider, Proc. IEEE MEMS Workshop, San Diego 1996 (IEEE, New York 1996) 216–221

32.11 S. Nakamura, K. Suzuki, M. Ataka, H. Fujita: An electrostatic micro actuator for a magnetic head tracking system of hard disk drives, Transducers '97 (IEEE, Piscataway 1997) 1081–1084

32.12 Y. Soeno, S. Ichikawa, T. Tsuna, Y. Sato, I. Sato: Piezoelectric piggy-back microactuator for hard disk drive, IEEE Trans. Magn. **35** (1999) 983–987

32.13 Y. Li, R. Horowitz: Mechatronics of electrostatic microactuator for computer disk drive dual-stage servo systems, IEEE/ASME Trans. Mechatron. **6** (2001) 111–121

32.14 N. Wongkomet: Position sensing for electrostatic micropositioners. Ph.D. Thesis (University of California, Berkeley 1998)

32.15 A. Hac, L. Liu: Sensor and actuator location in motion control of flexible structures, J. Sound Vib. **167** (1993) 239–261

32.16 K. Hiramoto, H. Doki, G. Obinata: Optimal sensor/actuator placement for active vibration control using explicit solution of algebraic Riccati equation, J. Sound Vib. **229** (2000) 1057–1075

32.17 P. Cheung, R. Horowitz, R. Howe: Design, fabrication and control of an electrostatically driven polysilicon microactuator, IEEE Trans. Magn. **32** (1996) 122–128

32.18 T. Chen, Y. Li, K. Oldham, R. Horowitz: MEMS application in computer disk drive drive dual-stage servo systems, J. Soc. Instrum. Control Eng. **41** (2002) 412–420

32.19 T.-L. Chen: Design and fabrication of PZT-actuated silicon suspensions for hard disk drives. Ph.D. Thesis (University of California, Berkeley 2001)

32.20 C. G. Keller, R. T. Howe: HexSil tweezers for teleoperated micro-assembly, 10th Int'l Workshop on Micro Electro Mechaninical Systems (MEMS'97), Nagoya 1997 (IEEE, New York 1997) 72–77

32.21 T. J. Brosnihan, J. M. Bustillo, A. P. Pisano, R. T. Howe: Embedded interconnect and electrical isolation for high-aspect-ratio, SOI inertial instruments, International Conference on Solid-State Sensors and Actuators, New York 1997 (IEEE, New York 1997) 637–640

32.22 T. Hirano, L.-S. Fan, T. Semba, W. Y. Lee, J. Hong, S. Pattanaik, P. Webb, W.-H. Juan, S. Chan: Microactuator for tera-storage, IEEE Int'l MEMS 1999 Conference, Orlando 1999 (IEEE, Piscataway 1999) 6441–6446

32.23 T. Hirano, L.-S. Fan, T. Semba, W. Lee, J. Hong, S. Pattanaik, P. Webb, W.-H. Juan, S. Chan: High-bandwidth HDD tracking servo by a moving-slider micro-actuator, IEEE Trans. Magn. **35** (1999) 3670–3672

32.24 T. Iizuka, T. Oba, H. Fugita: Electrostatic micro actuators with high-aspect-ratio driving gap for hard disk drive applications, Int'l Symposium on micromechatronics and human science (IEEE, Piscataway 2000) 229–236

32.25 H. Kuwajima, K. Matsuoka: Thin film piezoelectric dual-stage actuator for HDD, InterMag Europe, Session BS04 (IEEE, Piscaraway 2000) BS4

32.26 Y. Lou, P. Gao, B. Qin, G. Guo, E.-H. Ong, A. Takada, K. Okada: Dual-stage servo with on-slider PZT microactuator for hard disk drives, InterMag Europe, Session BS03, Amsterdam 2002 (IEEE, Piscataway 2002)

32.27 H. Fujita, K. Suzuki, M. Ataka, S. Nakamura: A microactuator for head positioning system of hard disk drives, IEEE Trans. Magn. **35** (1999) 1006–1010

32.28 B.-H. Kim, K. Chun: Fabrication of an electrostatic track-following micro actuator for hard disk drives using SOI wafer, J. Micromech. Microeng. **11** (2001) 1–6

32.29 R. Ehrlich, D. Curran: Major HDD TMR sources, and projected scaling with TPI, IEEE Trans. Magn. **35** (1999) 885–891

32.30 K. Mori, T. Munemoto, H. Otsuki, Y. Yamaguchi, K. Akagi: A dual-stage magnetic disk drive actuator using a piezoelectric device for a high track density, IEEE Trans. Magn. **27** (1991) 5298–5300

32.31 S. J. Schroeck, W. C. Messner, R. J. McNab: On compensator design for linear time-invariant dual-input single-output systems, IEEE/ASME Trans. Mechatron. **6** (2001) 50–57

32.32 T. Semba, T. Hirano, L.-S. Fan: Dual-stage servo controller for HDD using MEMS actuator, IEEE Trans. Magn. **35** (1999) 2271–2273

32.33 T. Suzuki, T. Usui, M. Sasaki, F. Fujisawa, T. Yoshida, H. Hirai: Comparison of robust track-following control systems for a dual stage hard disk drive, Proc. of International Conference on Micromechatronics for Information and Precision Equipment, ed. by B. Bhushan, K. Ohno (Word Scientific, Singapore 1997) 101–118

32.34 X. Hu, W. Guo, T. Huang, B. M. Chen: Discrete time LQG/LTR dual-stage controller design and implementation for high track density HDDs, Proc. of American Automatic Control Conference (IEEE, Piscataway 1999) 4111–4115

32.35 S.-M. Suh, C. C. Chung, S.-H. Lee: Design and analysis of dual-stage servo system for high track density HDDs, Microsyst. Technol. **8** (2002) 161–168

32.36 D. Hernandez, S.-S. Park, R. Horowitz, A. K. Packard: Dual-stage track-following servo design for hard disk drives, Proc. of American Automatic Control Conference (IEEE, Piscataway 1999) 4116–4121

32.37 M. Rotunno, R. A. de Callafon: Fixed order H_∞ control design for dual-stage hard disk drives, Proc. of the 39th IEEE Conference on Decision and Control (IEEE, Piscataway 2000) 3118–3119

32.38 S.-H. Lee, S.-E. Baek, Y.-H. Kim: Design of a dual-stage actuator control system with discrete-time sliding mode for hard disk drives, Proc. of the 39th IEEE Conference on Decision and Control (IEEE, Piscataway 2000) 3120–3125

32.39 M. Sasaki, T. Suzuki, E. Ida, F. Fujisawa, M. Kobayashi, H. Hirai: Track-following control of a dual-stage hard disk drive using a neuro-control system, Eng. Appl. Artif. Intell. **11** (1998) 707–716

32.40 D. Horsley, N. Wongkomet, R. Horowitz, A. Pisano: Precision positioning using a microfabricated electrostatic actuator, IEEE Trans. Magn. **35** (1999) 993–999

32.41 K. J. Åström, B. Wittenmark: *Adaptive Control*, 2nd edn. (Addison-Wesley, Reading 1995)

32.42 M. T. White, W.-M. Lu: Hard disk drive bandwidth limitations due to sampling frequency and computational delay, Proc. of the 1999 IEEE/ASME International Conference on Intelligent Mechatronics (IEEE, Piscataway 1999) 120–125

32.43 G. J. Balas, J. C. Doyle, K. Glover, A. Packard, R. Smith: *μ-Analysis and Synthesis ToolBox* (MUSYN Inc. and The MathWorks, Natick 1995)

32.44 Y. Li, R. Horowitz: Design and testing of track-following controllers for dual-stage servo systems with PZT actuated suspensions [in HDD], Microsyst. Technol. **8** (2002) 194–205

32.45 M. Kobayashi, R. Horowitz: Track seek control for hard disk dual-stage servo systems, IEEE Trans. Magn. **37** (2001) 949–954

32.46 M. Tomizuka: Zero phase error tracking algorithm for digital control, Trans. ASME J. Dynam. Syst. Meas. Control **109** (1987) 65–68

32.47 Y. Mizoshita, S. Hasegawa, K. Takaishi: Vibration minimized access control for disk drive, IEEE Trans. Magn. **32** (1996) 1793–1798

32.48 P. Vettiger, G. Cross, M. Despont, U. Drechsler, U. Durig, B. Gotsmann, W. Haberle, M. A. Lantz, H. W. Rothuizen, R. Stutz, G. K. Binnig: The “millipede” – nanotechnology entering data storage, IEEE Trans. Nanotechnol. **1** (2002) 39–55

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

33. Micro/Nanotribology of MEMS/NEMS Materials and Devices

The field of MEMS/NEMS has expanded considerably over the last decade. The length scale and large surface-to-volume ratio of the devices result in very high retarding forces such as adhesion and friction that seriously undermine the performance and reliability of the devices. These tribological phenomena need to be studied and understood at the micro- to nanoscales. In addition, materials for MEMS/NEMS must exhibit good microscale tribological properties. There is a need to develop lubricants and identify lubrication methods that are suitable for MEMS/NEMS. Using AFM-based techniques, researchers have conducted micro/nanotribological studies of materials and lubricants for use in MEMS/NEMS. In addition, component level testing has also been carried out to aid in better understanding the observed tribological phenomena in MEMS/NEMS.

Macroscale and microscale tribological studies of silicon and polysilicon films have been performed. The effects of doping and oxide films and environment on the tribological properties of these popular MEMS/NEMS materials have also been studied. SiC film is found to be a good tribological material for use in high-temperature MEMS/NEMS devices. Hexadecane thiol self-assembled monolayers and bonded perfluoropolyether lubricants appear to be well suited for lubrication of microdevices under a range of environmental conditions. DLC coatings can also be used for low friction and wear. Surface roughness measurements of micromachined polysilicon surfaces have been made using an AFM. The roughness distribution on surfaces is strongly dependent on the fabrication process. Roughness should be optimized for low adhesion, friction, and wear. Adhesion and friction of microstructures can be measured using novel apparatuses. Adhesion and friction measurements on silicon-on-silicon confirm AFM measurements that hexadecane thiol and bonded perfluoropolyether films exhibit superior adhesion and friction properties. Static friction force measurements of micromotors have been performed using an AFM. The forces are found to vary considerably with humidity. A bonded layer of perfluoropolyether lubricant is found to satisfactorily reduce the friction forces in the micromotor.

AFM/FFM-based techniques can be satisfactorily used to study and evaluate micro/nanoscale tribological phenomena related to MEMS/NEMS devices.

This chapter presents a review of macro- and micro/nanoscale tribological studies of materials and lubrication studies for MEMS/NEMS and component-level studies of stiction phenomena in MEMS/NEMS devices.

33.1 **Introduction to MEMS** 985

33.2 **Introduction to NEMS** 988

33.3 **Tribological Issues in MEMS/NEMS** 989
  33.3.1 MEMS 989
  33.3.2 NEMS 994
  33.3.3 Tribological Needs 995

33.4 **Tribological Studies of Silicon and Related Materials** 995
  33.4.1 Tribological Properties of Silicon and the Effect of Ion Implantation 996
  33.4.2 Effect of Oxide Films on Tribological Properties of Silicon 998
  33.4.3 Tribological Properties of Polysilicon Films and SiC Film 1000

33.5 **Lubrication Studies for MEMS/NEMS** 1003
  33.5.1 Perfluoropolyether Lubricants 1003
  33.5.2 Self-Assembled Monolayers (SAMs) 1004
  33.5.3 Hard Diamond-like Carbon (DLC) Coatings 1008

33.6 **Component-Level Studies** 1009
  33.6.1 Surface Roughness Studies of Micromotor Components 1009
  33.6.2 Adhesion Measurements 1011
  33.6.3 Static Friction Force (Stiction) Measurements in MEMS 1014
  33.6.4 Mechanisms Associated with Observed Stiction Phenomena in Micromotors 1016

References 1017

Microelectromechanical systems (MEMS) refer to microscopic devices that have a characteristic length of less than 1 mm but more than 1 µm and combine electrical and mechanical components. Nanoelectromechanical systems (NEMS) refer to nanoscopic devices that have a characteristic length of less than 1 µm and combine electrical and mechanical components. In mesoscale devices, if the functional components are on the micro- or nanoscale, they may be referred to as MEMS or NEMS, respectively. To put the dimensions and masses in perspective, see Fig. 33.1 and Table 33.1. The acronym MEMS originated in the United States. The term commonly used in Europe and Japan is micro/nanodevices, which is used in a much broader sense. MEMS/NEMS terms are also now used in a broad sense. A micro/nanosystem, a term commonly used in Europe, is referred to as an intelligent miniaturized system comprising sensing, processing, and/or actuating functions.

Fabrication techniques include top-down methods, in which one builds down from the large to the small, and the bottom-up methods, in which one builds up from the small to the large. Top-down methods include micro/nanomachining methods and methods based on lithography, as well as nonlithographic miniaturization for MEMS and NEMS fabrication. In the bottom-up methods, also referred to as nanochemistry, the devices and systems are assembled from their elemental constituents for NEMS fabrication, much like the way nature uses proteins and other macromolecules to

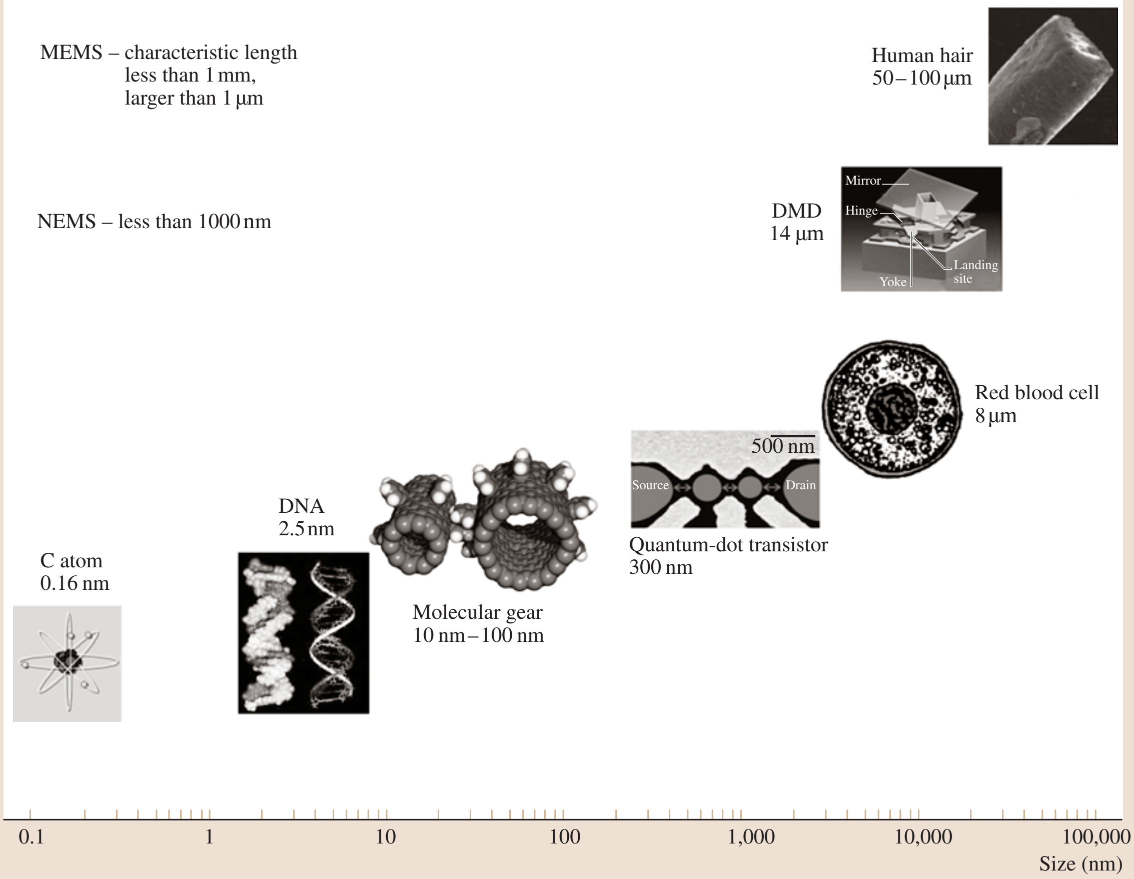
MEMS – characteristic length
less than 1 mm,
larger than 1 µm
NEMS – less than 1000 nm
Human hair
50–100 µm
DMD
14 µm
Mirror
Hinge
Landing site
Yoke
Red blood cell
8 µm
500 nm
Source
Drain
Quantum-dot transistor
300 nm
DNA
2.5 nm
Molecular gear
10 nm–100 nm
C atom
0.16 nm
0.1
1
10
100
1,000
10,000
100,000
Size (nm)

Fig. 33.1 Dimensions of MEMS and NEMS in perspective. An example of molecular dynamic simulations of carbon-nanotube-based gear is obtained from [33.1], quantum-dot transistor is obtained from [33.2], and DMD is obtained from [33.3]

Table 33.1 Dimensions and masses in perspective

A Dimensions in Perspective	
NEMS characteristic length	$< 1{,}000$ nm
MEMS characteristic length	< 1 mm and > 1 µm
Individual atoms	typically fraction of a nm in diameter
DNA molecules	~ 2.5 nm wide
Molecular gear	~ 50 nm
Biological cells	in the range of thousands of nm in diameter
Human hair	$\sim 75{,}000$ nm in diameter
B Masses in Perspective	
NEMS built with cross sections of about 10 nm	as low as 10^{-20} N
Micromachine silicon structure	as low as 1 nN
Water droplet	~ 10 µN
Eyelash	~ 100 nN

construct complex biological systems. The bottom-up approach has the potential to go far beyond the limits of top-down technology by producing nanoscale features through synthesis and subsequent assembly. Furthermore, the bottom-up approach offers the potential to produce structures with enhanced and/or completely new functions. It allows the combination of materials with distinct chemical composition, structure, and morphology.

MEMS and emerging NEMS are expected to have a major impact on our lives, comparable to that of semiconductor technology, information technology, or cellular and molecular biology [33.4, 5]. MEMS/NEMS are used in mechanical, information/communication, chemical, and biological applications. The MEMS industry in 2000 was worth about \$15 billion and, with a projected 10–20% annual growth rate, is expected to be more than \$100 billion by the end of this decade [33.6]. Growth of Si MEMS may slow down and non-silicon MEMS may pick up during this decade. The NEMS industry is expected to expand in this decade, mostly in biomedical applications, as well as in nanoelectronics or molecular electronics. The NEMS industry was worth about \$100 million in 2002 and integrated NEMS are expected to be more than \$25 billion by the end of this decade. Due to the enabling nature of these systems, and because of the significant impact they can have on both commercial and defense applications, industry, as well as the federal government have taken a special interest in seeing growth nurtured in this field. Micro- and nanosystems are the next logical step in the "silicon revolution."

33.1 Introduction to MEMS

The advances in silicon photolithographic process technology beginning in the 1960s led to the development of MEMS in the early 1980s. More recently, lithographic processes have been developed to process nonsilicon materials. The lithographic processes are being complemented with non-lithographic processes for fabrication of components or devices made from plastics or ceramics. Using these fabrication processes, researchers have fabricated a wide variety of miniaturized devices, including Si-based devices, chemical and biological sensors and actuators, and miniature non-silicon structures (e.g., devices made from plastics or ceramics) with dimensions in the range of a couple to a few thousand microns (see e.g., [33.7–15]). MEMS for mechanical applications include acceleration, pressure, flow, gas sensors, linear and rotary actuators, and other microstructures or microcomponents such as electric motors, gear trains, gas turbine engines, nozzles, fluid pumps, fluid valves, switches, grippers, and tweezers. MEMS for chemical applications include chemical sensors and various analytical instruments. Microoptoelectromechanical systems, or MOEMS, include micromirror arrays and fiber optic connectors. BIOMEMS include biochips. Radio frequency MEMS or RF-MEMS include inductors, capacitors, and antennas. High-aspect-ratio MEMS (HARMEMS) have also been introduced.

The fabrication techniques for MEMS devices include lithographic and non-lithographic techniques. The lithographic techniques fall into three basic categories: bulk micromachining, surface micromachining, and LIGA (a German acronym for Lithographie Galvanoformung Abformung), a German term for lithography, electroforming, and plastic molding. The first two approaches, bulk and surface micromachining, use planar photolithographic fabrication processes developed for semiconductor devices in producing two-dimensional (2-D) structures [33.10, 15–17]. The various steps involved in these two fabrication processes are shown schematically in Fig. 33.2. Bulk micromachining employs anisotropic etching to remove sections through the thickness of a single-crystal silicon wafer, typically 250 to 500 μm thick. Bulk micromachining is a proven high-volume production process and is routinely used to fabricate microstructures such as accelerometers, pressure sensors, and flow sensors. In surface micromachining, structural and sacrificial films are alternatively deposited, patterned and etched to produce a freestanding structure. These films are typically made of low-pressure chemical vapor deposition (LPCVD) polysilicon film with 2 to 20 μm thickness. Surface micromachining is used to produce sensors, actuators, and complex microdevices such as micromirror arrays, motors, gears, and grippers.

The LIGA process is based on the combined use of X-ray lithography, electroforming, and molding processes. The steps involved in the LIGA process are shown schematically in Fig. 33.3. LIGA is used to produce high-aspect-ratio MEMS (HARMEMS) devices that are up to 1 mm in height and only a few microns in width or length [33.18]. The LIGA process yields very sturdy 3-D structures due to their increased thickness. One of the limitations of silicon microfabrication processes originally used for fabrication of MEMS devices is the lack of suitable materials that can be processed. With LIGA, a variety of non-silicon materials such as metals, ceramics, and polymers can be processed. Nonlithographic micromachining processes, primarily in Europe and Japan, are also being used for fabrication of millimeter-scale devices using direct material microcutting or micromechanical machining (such as microturning, micromilling, and microdrilling), or removal by energy beams (such as microspark erosion, focused ion beam, laser ablation, and laser polymerization) [33.19, 20]. Hybrid technologies, including LIGA and high-precision micromachining techniques, have been used to produce miniaturized motors, gears, actuators, and connectors [33.21–24]. These millimeter-scale devices may find more immediate applications.

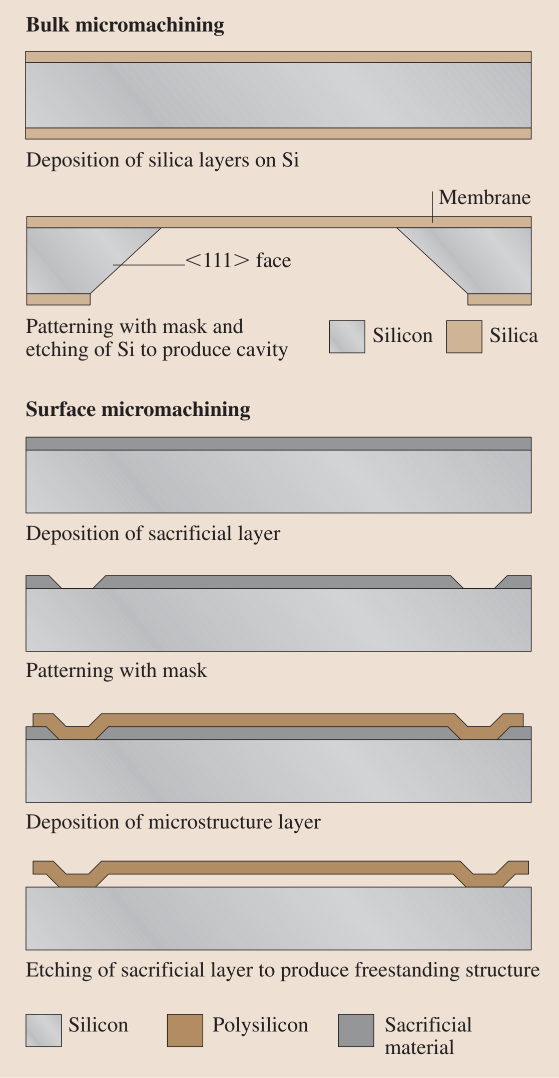
Bulk micromachining
Deposition of silica layers on Si
Membrane
<111> face
Patterning with mask and etching of Si to produce cavity
Silicon
Silica
Surface micromachining
Deposition of sacrificial layer
Patterning with mask
Deposition of microstructure layer
Etching of sacrificial layer to produce freestanding structure
Silicon
Polysilicon
Sacrificial material

Fig. 33.2 Schematics of process steps involved in bulk micromachining and surface micromachining fabrication of MEMS

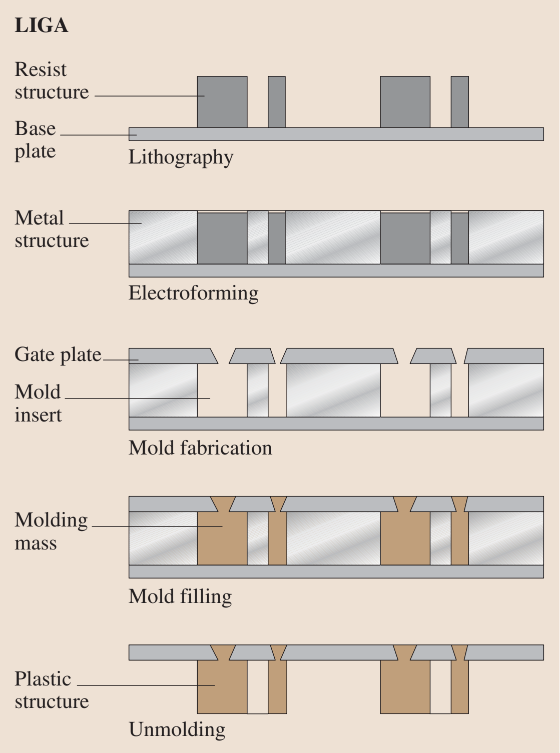

Fig. 33.3 Schematic of process steps involved in LIGA fabrication of MEMS

A microfabrication technique, so-called "soft lithography," is a nonlithographic technique [33.25, 26] in which an elastomeric stamp or mold is used to generate micropatterns by replica molding [33.27], embossing (imprinting) [33.28], or by contact printing (known as microcontact printing or µCP) [33.29]. Replica molding is commonly used for mass-produced, disposable plastic microcomponents, for example, microfluidic chips, generally made of poly(dimethylsiloxane) (PDMS) and polymethyl methacrylate (PMMA). The elastomeric stamps are generally cast from photolithographically generated micropatterned masters. This technique is substantially cheaper and more flexible in choice of materials for construction than conventional photolithography.

To assemble microsystems, microrobots are used. Microrobotics include building blocks such as steering links, microgrippers, conveyor system, and locomotive robots [33.13].

A variety of MEMS devices have been produced and some are commercially used [33.8, 10–12, 14, 15]. A variety of sensors are used in industrial, consumer, defense, and biomedical applications. Various microstructures or microcomponents are used in micro-instruments and other industrial applications such as micromirror arrays. Two of the largest "killer" industrial applications are accelerometers (about 85 million units in 2002) and digital micromirror devices (about $400 million in sales in 2001). Integrated capacitive-type, silicon accelerometers have been used in airbag deployment in automobiles since 1991 [33.30, 31]. Accelerometer technology was about a billion-dollar-a-year industry in 2001 dominated by Analog Devices, followed by Motorola and Bosch. Commercial digital light processing (DLP) equipments, using digital micromirror devices (DMD), were launched in 1996 by Texas Instruments for digital projection displays in portable and home theater projectors, as well as table top and projection TVs [33.32–34]. More than 1.5 million projectors were sold before 2002. Other major industrial applications include pressure sensors, inkjet printer heads and optical switches. Silicon-based piezoresistive pressure sensors for manifold absolute pressure sensing for engines were launched in 1991 by NovaSensor, and their annual sales were about 25 million units in 2002. Capacitive pressure sensors for tire pressure measurements were launched by Motorola. Annual sales of inkjet printer heads with microscale functional components were about 400 million units in 2002. Other applications of MEMS devices include chemical sensors, gas sensors, infrared detectors and focal plane arrays for earth observation, space science, and missile defense applications, pico-satellites for space applications, and many hydraulic, pneumatic, and other consumer products. MEMS devices are also being pursued in magnetic storage systems [33.35], where they are being developed for super-compact and ultrahigh-recording-density magnetic disk drives. Several integrated head/suspension microdevices have been fabricated for contact recording applications [33.36, 37]. High-bandwidth servo-controlled microactuators have been fabricated for ultrahigh-track-density applications that serve as the fine-position control element of a two-stage, coarse/fine servo system, coupled with a conventional actuator [33.38–40].

BIOMEMS are used increasingly in commercial and defense applications (e.g., [33.8, 41–44]). Applications of BIOMEMS include biofluidic chips (microfluidic chips or bioflips or simply biochips) for chemical and biochemical analyses (biosensors) in medical diagnostics (e.g., DNA, RNA, proteins, cells, blood pressure and assays, and toxin identification) and implantable pharmaceutical drug delivery. The biosensors, also referred to as lab-on-a-chip, integrate sample handling, separation, detection, and data analysis onto one platform. Biosensors are designed to either detect a single or class of (bio)chemicals or system-level analytical capabilities for a broad range of (bio)chemical species known as micro total analysis systems (μTAS). The chips rely on microfluidics and involve manipulation of tiny amounts of fluids in microchannels using microvalves. The chips consist of several basic components, including microfluidic channels and reservoirs, microvalves, micropumps, flow sensors, and biosensors. The test fluid is injected into the chip generally using an external pump or a syringe for analyses. Some chips have been designed with an integrated electrostatically actuated diaphragm-type micropump. Micropumps both with and without valves are used. The sample, which can have volume measured in nanoliters, flows through microfluidic channels via an electric potential and capillary action using microvalves (having various designs, including membrane type) for various analyses. The fluid is preprocessed and then analyzed using a biosensor. For a review on micropumps and microvalves, see [33.11, 45, 46]. Silicon-based disposable blood-pressure sensor chips were introduced in the early 1990s by NovaSensor for blood pressure monitoring (about 20 million units in 2002). A blood sugar monitor, referred to as GlucoWatch, was introduced in 2002. It automatically checks blood sugar every 10 min by detecting glucose through the skin, without having to draw blood. If glucose is out of the acceptable range, it sounds an alarm so the diabetic can address the problem quickly. A variety of biosensors, many using plastic substrates, are manufactured by various companies, including ACLARA, Agilent Technologies, Calipertech, and I-STAT.

After the tragedy of Sept. 11, 2001, concern over biological and chemical warfare has led to the development of handheld units with bio- and chemical sensors for detection of biological germs, chemical or nerve agents, and mustard agents, and chemical precursors to protect subways, airports, water supply, and population at large [33.47].

Other BIOMEMS applications include minimal invasive surgery, including endoscopic surgery, laser angioplasty, and microscopic surgery. Implantable artificial organs can also be produced.

Micro-instruments and micromanipulators are used to move, position, probe, pattern, and characterize nanoscale objects and nanoscale features. Miniaturized analytical equipment include gas chromatography and mass spectrometry. Other instruments include micro-STM, where STM stands for scanning tunneling microscope.

In some cases, MEMS devices are used primarily for their miniature size. While in others, as in the case of air bags, they are used because of their low-cost manufacturing techniques, since semiconductor-processing costs have reduced drastically over the last decade, allowing the use of MEMS in many fields.

33.2 Introduction to NEMS

NEMS are produced by nanomachining in a typical top-down approach (from large to small) and bottom-up approach (from small to large) largely relying on nanochemistry [33.48–53]. The top-down approach relies on fabrication methods, including advanced integrated-circuit (IC) lithographic methods – electron-beam lithography and STM writing by removing material atom by atom. The bottom-up approach includes chemical synthesis, the spontaneous “self-assembly” of molecular clusters (molecular self-assembly) from simple reagents in solution, or biological molecules (e.g., DNA) as building blocks to produce 3-D nanostructures, quantum dots (nanocrystals) of arbitrary diameter (about $10–10^5$ atoms), molecular beam epitaxy (MBE) and organometallic vapor phase epitaxy (OMVPE) to create specialized crystals one atomic or molecular layer at a time, and manipulation of individual atoms by an atomic force microscope or atom optics. The self-assembly must be encoded. That is, one must be able to precisely assemble one object next to another to form a designed pattern. A variety of nonequilibrium plasma chemistry techniques are also used to produce layered nanocomposites, nanotubes, and nanoparticles. The NEMS field, in addition to fabrication of nanosystems, has provided the impetus to develop experimental and computation tools.

Examples of NEMS include nanocomponents, nanodevices, nanosystems, and nanomaterials such as microcantilevers with integrated sharp nanotips for STM and atomic force microscopy (AFM), AFM array (millipede) for data storage, AFM tips for nanolithography, dip-pen nanolithography for printing molecules, biological (DNA) motors, molecular gears, molecularly thick films (e.g., in giant magnetoresistive or GMR heads, and magnetic media), nanoparticles (e.g., nanomagnetic particles in magnetic media), nanowires, carbon nanotubes, quantum wires (QWRs), quantum boxes (QBs), and quantum transistors. BIONEMS include nanobiosensors – a microarray of silicon nanowires, roughly a few nm in size, to selectively bind and detect even a single biological molecule such as DNA or protein by using nanoelectronics to detect the slight electrical charge caused by such binding, or a microarray of carbon nanotubes to electrically detect glucose, implantable drug-delivery devices – e.g., micro/nanoparticles with drug molecules encapsulated in functionized shells for a site-specific targeting applications and a silicon capsule with a nanoporous membrane filled with drugs for long-term delivery – , nanodevices for sequencing single molecules of DNA in the Human Genome Project, cellular growth using carbon nanotubes for spinal cord repair, nanotubes for nanostructured materials for various applications such as spinal fusion devices, organ growth, and growth of artificial tissues using nanofibers.

Nanoelectronics can be used to build computer memory using individual molecules or nanotubes to store bits of information, molecular switches, molecular or nanotube transistors, nanotube flat-panel displays, nanotube integrated circuits, fast logic gates, switches, nanoscopic lasers, and nanotubes as electrodes in fuel cells.

33.3 Tribological Issues in MEMS/NEMS

Tribological issues are important in MEMS/NEMS requiring intended and/or unintended relative motion. In MEMS/NEMS, various forces associated with the device scale down with size. When the length of the machine decreases from 1 mm to 1 μm, the area decreases by a factor of a million and the volume decreases by a factor of a billion. As a result, surface forces such as adhesion, friction, meniscus forces, viscous drag forces, and surface tension that are proportional to area become a thousand times larger than the forces proportional to the volume such as inertial and electromagnetic forces. In addition to the consequence of a large surface-to-volume ratio, since MEMS/NEMS are designed for small tolerances, physical contact becomes more likely, which makes them particularly vulnerable to adhesion between adjacent components. Since the start-up forces and torques involved in MEMS/NEMS operation available to overcome retarding forces are small, the increase in resistive forces such as adhesion and friction become a serious tribological concern that limits the life and reliability of MEMS/NEMS [33.10]. A large lateral force required to initiate relative motion between two surfaces, large static friction, is referred to as "stiction," which has been studied extensively in tribology of magnetic storage systems [33.35, 54, 55]. A large normal force required to separate two surfaces is also referred to as stiction. Adhesion, friction/stiction (static friction), wear, and surface contamination affect MEMS/NEMS performance and in some cases, can even prevent devices from working. Some examples of devices that experience tribological problems follow.

33.3.1 MEMS

Figure 33.4a shows examples of several microcomponents that can encounter the aforementioned tribological problems. The polysilicon electrostatic micromotor has 12 stators and a four-pole rotor and is produced by surface micromachining. The rotor diameter is 120 μm and the air gap between the rotor and stator is 2 μm [33.56]. It is capable of continuous rotation up to speeds of 100,000 RPM. The intermittent contact at the rotor-stator interface and physical contact at the rotor-hub flange interface result in wear issues, and high stiction between the contacting surfaces limits the repeatability of operation, or may even prevent the operation altogether. Next, a bulk micromachined silicon stator/rotor pair is shown with bladed rotor and nozzle guide vanes on the stator with dimensions less than a mm [33.57]. These are being developed for high-temperature micro-gas turbine generators with an operating speed up to 1 million RPM. Erosion of blades and vanes and wear of bearings used in the turbine generators are some of the concerns. Next, is an SEM micrograph of a surface micromachined polysilicon six-gear chain from Sandia National Lab. (For more examples of previous versions, see [33.58, 59].) As an example of non-silicon components, a milligear system produced using the LIGA process for a DC brushless permanent magnet millimotor (diameter = 1.9 mm, length = 5.5 mm) with an integrated milligear box [33.21–23] is also shown. The gears are made of metal (electroplated Ni-Fe), but can also be made from injected polymer materials (e.g.,

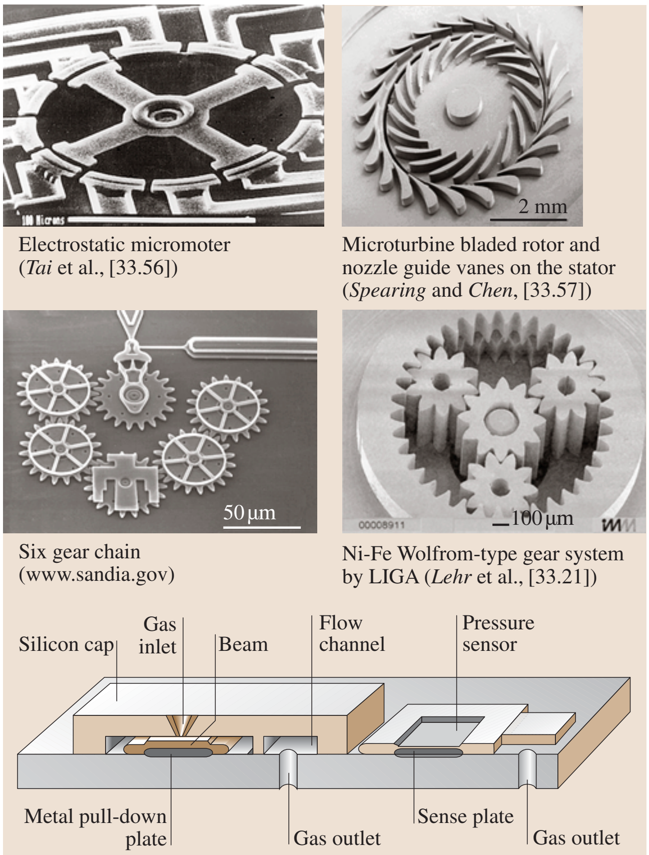

Electrostatic micromoter (*Tai* et al., [33.56])

Microturbine bladed rotor and nozzle guide vanes on the stator (*Spearing* and *Chen*, [33.57])

Six gear chain (www.sandia.gov)

Ni-Fe Wolfrom-type gear system by LIGA (*Lehr* et al., [33.21])

a) Low-pressure flow modulator with electrostatically actuated microvalves (*Robertson* and *Wise*, [33.60])

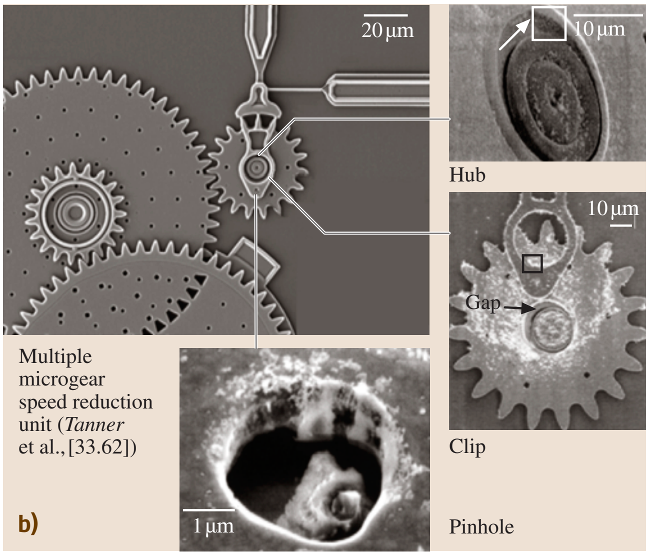

Multiple microgear speed reduction unit (*Tanner* et al., [33.62])

Fig. 33.4a,b Examples of MEMS devices and components that experience tribological problems: **(a)** several micro-component and **(b)** a polysilicon, multiple microgear speed reduction unit

Polyoxy-methylene, or POM) using the LIGA process. Even though the torque transmitted at the gear teeth is small, on the order of a fraction of a nNm, because of small dimensions of gear teeth, the bending stresses are large where the teeth mesh. Tooth breakage and wear at the contact of gear teeth are concerns. Next, in a micromachined flow modulator, several micromachined flow channels are integrated in series with electrostatically actuated microvalves [33.60]. The flow channels lead to a central gas outlet hole drilled in the glass substrate. Gas enters the device through a bulk micromachined gas inlet hole in the silicon cap. After passing through an open microvalve, the gas flows parallel to the glass substrate through flow channels and exits the device through an outlet. The normally open valve structure consists of a freestanding double-end-clamped beam, which is positioned beneath the gas inlet orifice. When deflected electrostatically upwards, the beam seals against the inlet orifice, and the valve is closed. In these microvalves used for flow control, the mating valve surfaces should be smooth enough to seal while maintaining a minimum roughness to ensure low adhesion [33.54, 55, 61]. High adhesion (stiction) is a major issue.

Figure 33.4b shows a polysilicon, multiple microgear speed reduction unit and its components after laboratory wear tests conducted in air [33.62]. These units have been developed for an electrostatically driven microactuator (microengine) developed at Sandia National Lab for operation in the kHz frequency range. Wear of various components is clearly observed in the figure. Humidity was shown to be a strong factor in the wear of rubbing surfaces. In order to improve the wear characteristics of rubbing surfaces, 20-nm-thick tungsten (W) coating using chemical vapor deposition (CVD) technique was used [33.63]. Tungsten-coated microengines tested for reliability showed improved wear characteristics with longer lifetimes than polysilicon microengines.

Commercially available MEMS devices also exhibit tribological problems. Figure 33.5 shows an integrated capacitive-type silicon accelerometer fabricated using surface micromachining by Analog Devices, a couple of mm in dimension, which is used for the deployment of airbags in automotives and more recently for

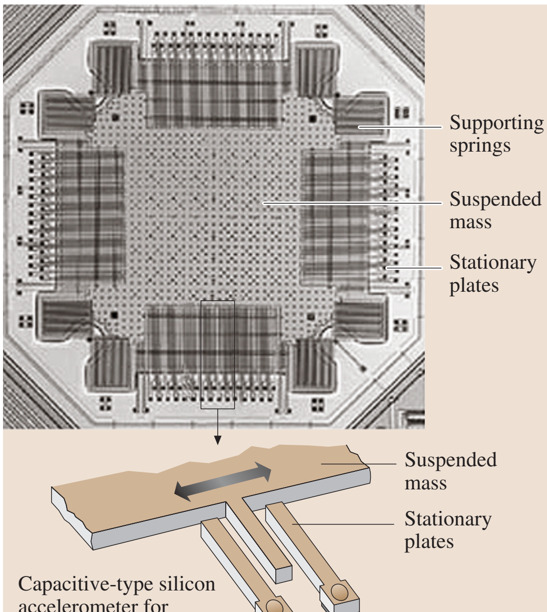
Supporting springs
Suspended mass
Stationary plates
Suspended mass
Stationary plates

Capacitive-type silicon accelerometer for automotive sensory applications (*Sulouff*, [33.64])

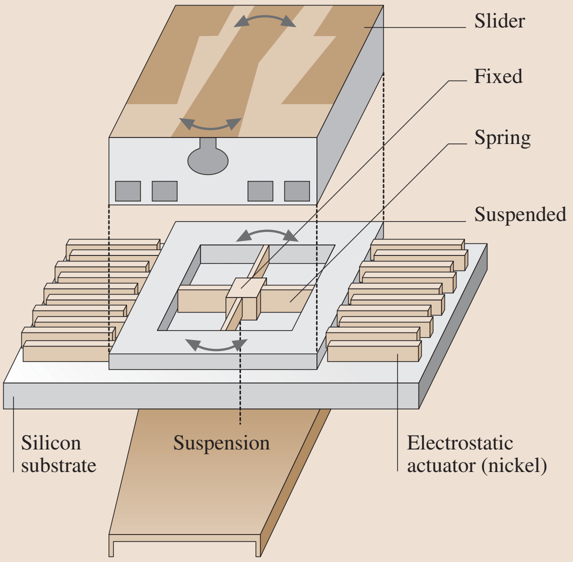
Slider
Fixed
Spring
Suspended
Silicon substrate
Suspension
Electrostatic actuator (nickel)

Electroplated-nickel rotary microactuator for magnetic disk drives (*Hirano* et al., [33.40])

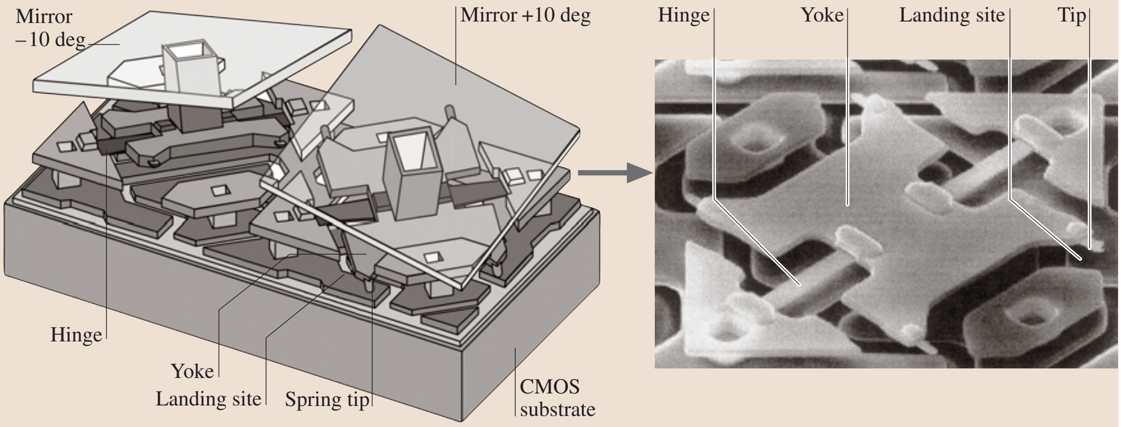
Mirror
–10 deg
Mirror +10 deg
Hinge
Yoke
Landing site
Spring tip
CMOS substrate
Hinge
Yoke
Landing site
Tip

Digital micromirror device for displays (*Hornbeck*, [33.33])

Fig. 33.5 Examples of commercial use MEMS devices that experience tribological problems

the consumer electronics market [33.30, 64, 65]. These accelerometers are now being developed for many other applications such as vehicle stability, rollover control, and gyro sensor applications. The central suspended beam mass (about 0.7 µg) is supported on the four corners by spring structures. The central beam has interdigitated, cantilevered electrode fingers (about 125 µm long and 3 µm thick) on all four sides that alternate with those of the stationary electrode fingers, as shown, with about a 1.3 µm gap. Lateral motion of the central beam causes a change in the capacitance between these electrodes, which is used to measure the acceleration. Here stiction between the adjacent electrodes, as well as stiction of the beam structure with the underlying substrate are detrimental to the operation of the sensor [33.30, 64]. Wear during unintended contacts of these polysilicon fingers is also a problem. A molecularly thick diphenyl siloxane lubricant film with high resistance to temperature and oxidation, applied by a vapor deposition process is used on the electrodes to reduce stiction and wear.

Figure 33.5 also shows two digital micromirror device (DMD) pixels used in digital light processing (DLP) technology for digital projection displays in portable and home theater projectors, as well as table-top and projection TVs [33.32–34]. The entire array (chip set) consists of a large number of rotatable aluminum micromirrors (digital light switches) that are fabricated on top of a CMOS static random access memory integrated circuit. The surface micromachined array consists of half of a million to more than two million of these independently controlled, reflective micromirrors (mirror size on the order of 14 µm square and 15 µm pitch) that flip backward and forward at a frequency on the order of 5,000 times per s. For the binary operation, a micromirror/yoke structure mounted on torsional hinges is rotated ±10° (with respect to the horizontal plane) as a result of electrostatic attraction between the micromirror structure and the underlying memory cell and is limited by a mechanical stop. Contact between cantilevered spring tips at the end of the yoke (four present on each yoke) with the underlying stationary landing sites is required for true digital (binary) operation. Stiction and wear during a contact between aluminum alloy spring tips and landing sites, hinge memory (metal creep at high operating temperatures), hinge fatigue, shock and vibration failure, and sensitivity to particles in the chip package and operating environment are some of the important issues affecting the reliable operation of a micromirror device [33.66–68]. Perfluorodecanoic acid (PFDA) self-assembled monolayers are used on the tip and landing sites to reduce stiction and wear. The spring tip is employed in order to use the spring stored energy to pop up the tip during pull-off. A lifetime estimate of over 100,000 operating hours with no degradation in image quality is the norm.

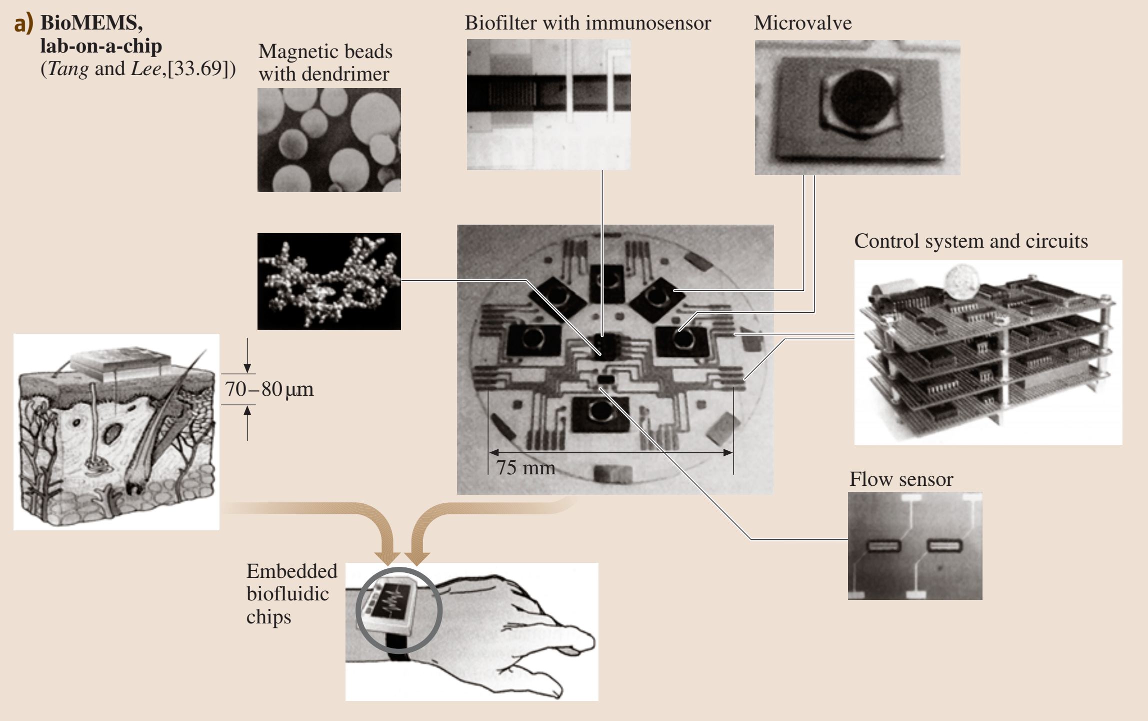

Fig. 33.6a,b Schematics of **(a)** a MEMS-based biofluidic chip, commonly known as disposable lab-on-a-chip, and **(b)** (*see next page*) a NEMS-based submicroscopic drug-delivery device, and an intravascular micro/nanoparticles for search-and-destroy diseased blood cells

A third MEMS device, shown in Fig. 33.5, is an electrostatically driven rotary microactuator for a magnetic disk drive surface-micromachined by a multilayer electroplating method [33.40]. This high-bandwidth servo-controlled microactuator, located between a slider and a suspension, is being developed for ultrahigh-track-density applications, which serves as the fine-position and high-bandwidth control element of a two-stage, coarse/fine servo system when coupled with a conventional actuator [33.38–40]. A slider is placed on top of the central block of a microactuator, which gives rotational motion to the slider. The bottom of the silicon substrate is attached to the suspension. The radial flexure beams in the central block give the rotational freedom of motion to the suspended mass (slider), and the electrostatic actuator drives the suspended mass. Actuation is accomplished via interdigitated, cantilevered electrode fingers, which are alternatingly attached to the central body of the moving part and to the stationary substrate to form pairs. A voltage applied across these electrodes results in an electrostatic force, which rotates the central block. The inter-electrode gap width is about $2\ \mu\mathrm{m}$. Any unintended contacts between the moving and stationary electroplated-nickel electrodes may result in wear and stiction.

An example of disposable plastic lab-on-a-chip using microfluidics and BIOMEMS technologies is shown in Fig. 33.6a [33.69]. This biofluidic chip integrates multiple fluidic components on a microfluidic motherboard that can be worn like a wristwatch. It enables the unobstructive assessment of a human's medical condition through continuous blood sampling. It can be used for the analysis of a number of different compounds, including lactate, glucose, carbon dioxide, and oxygen levels in the blood and for the detection of infectious diseases. The test fluid is injected into the microchannels using a micropump, and the flow is regulated using microvalves. A magnetic bead approach is used for identification of target biomolecules based on sandwich immunoassay and electrochemical detection. Adhesion in micropumps and microvalves involving moving parts and adhesion of fluid in microchannels are some of the tribological issues. In both active and passive microvalves, the main wear mechanism appears to be erosion, corrosion, and fatigue. As the magnetic bead particulates are pumped through the microfluidic system, they could erode the valve material. Also, several different reagents are used to perform the necessary analyses. The interaction of these reagents with the valve material could cause corrosive wear. Finally, because the valves are continuously being activated, fatigue failure becomes a concern.

b) **BioNEMS, submicroscopic drug-delivery device**
(*Ferrari* and *Liu*, [33.73])

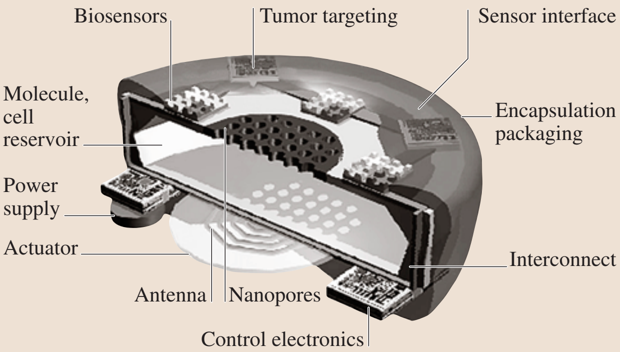

BioNEMS, intravascular micro/nanoparticles
(*Martin* and *Grove*, [33.74])

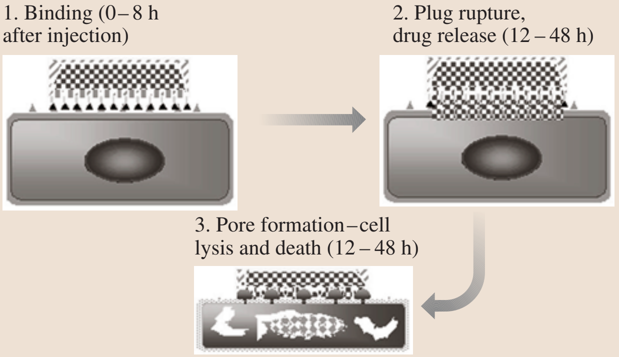
1. Binding (0–8 h after injection)
2. Plug rupture, drug release (12–48 h)
3. Pore formation–cell lysis and death (12–48 h)

Fig. 33.6b

Stiction/friction and wear clearly limit the lifetimes and compromise the performance and reliability of microdevices. Figure 33.7 summarizes tribological problems encountered in some of the MEMS devices and components just discussed. In addition, tribological issues are present in the processes used for fabrication of MEMS/NEMS. For example, in surface micromachining, the suspended structures can sometimes collapse and permanently adhere to the underlying substrate due to meniscus effects during the final rinse and dry process, as shown in Fig. 33.7 [33.70]. Adhesion is caused by water molecules adsorbed on the adhering surfaces and/or because of formation of adhesive bonds by silica residues that remain on the surfaces after the water has evaporated. This so called release stiction is overcome

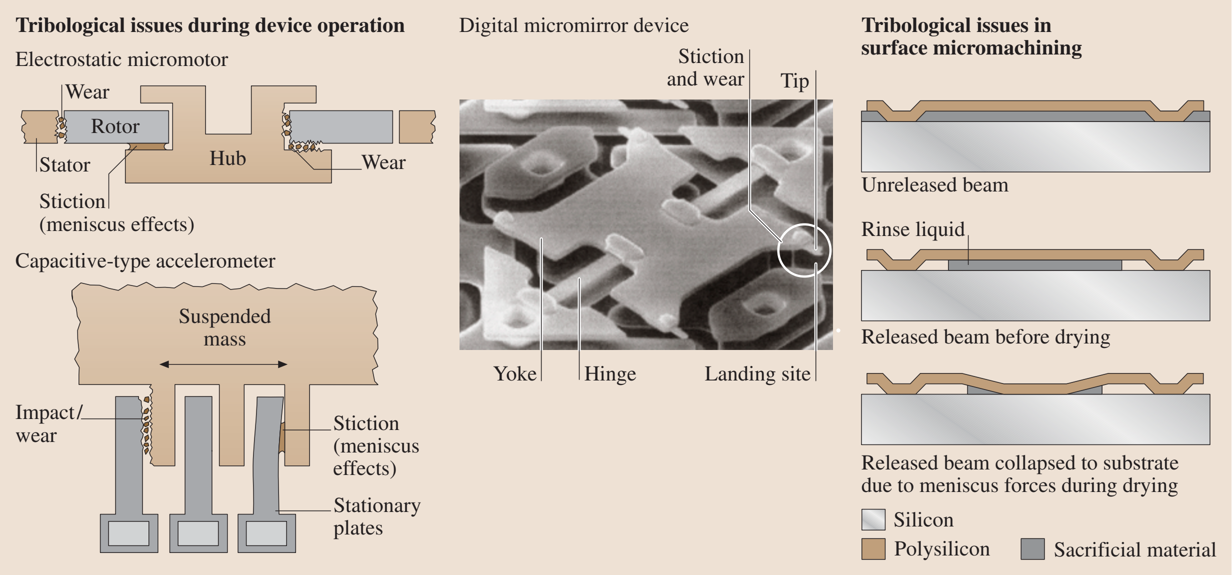
Tribological issues during device operation
Electrostatic micromotor
Wear
Rotor
Hub
Wear
Stator
Stiction
(meniscus effects)
Capacitive-type accelerometer
Suspended
mass
Impact/
wear
Stiction
(meniscus
effects)
Stationary
plates
Digital micromirror device
Stiction
and wear
Tip
Yoke
Hinge
Landing site
Tribological issues in
surface micromachining
Unreleased beam
Rinse liquid
Released beam before drying
Released beam collapsed to substrate
due to meniscus forces during drying
Silicon
Polysilicon
Sacrificial material

Fig. 33.7 Summary of tribological issues in MEMS device operation and fabrication via surface micromachining

by using dry release methods (e.g., CO_2 critical point drying or sublimation methods [33.71]).

33.3.2 NEMS

Figure 33.8 shows an AFM-based nanoscale data storage system for ultrahigh density magnetic recording that experiences tribological problems [33.72]. The system uses arrays of several thousand silicon microcantilevers ("millipede") for thermomechanical recording and playback on an about 40-nm-thick polymer (PMMA) medium with a harder Si substrate. The cantilevers are integrated with integrated tip heaters that have tips of nanoscale dimensions. Thermomechanical recording is a combination of applying a local force to the polymer layer, and softening it by local heating. The tip heated to about 400 °C is brought in contact with the polymer for recording. Imaging and reading are done using the heater cantilever, originally used for recording, as a thermal read-back sensor by exploiting its temperature-dependent resistance. The principle of thermal sensing is based on the fact that the thermal conductivity between the heater and the storage substrate changes according to the spacing between them. When the spacing between the heater and sample is reduced as the tip moves into a bit, the heater's temperature and hence its resistance will decrease. Thus, changes in temperature of the continuously heated resistor are monitored while the cantilever is scanned over data bits, providing a means of detecting the bits. Erasing for subsequent rewriting is carried out by thermal reflow of the storage field by heating the medium to 150 °C for a few seconds. The smoothness of the reflown medium allows multiple rewriting of the same storage field. Bit sizes ranging between 10 and 50 nm have been achieved by using a 32 × 32 (1024) array write/read chip (3 mm × 3 mm). It has been reported that tip wear occurs through the contact between tip and Si substrate during writing. Tip wear is considered a major factor in device reliability.

Figure 33.6b shows two examples of BIONEMS [33.73, 74]. The figure shows a conceptual model of a submicroscopic drug delivery device with the ability to localize in the areas of need. It is a silicon capsule with a nanoporous membrane filled with drugs for long-term delivery. It uses a nanomembrane with pores as small as 6 nm that are used as flux regulators for the long-term release of drugs. The nanomembrane also protects therapeutic substances from attack by the body's immune system. The figure also shows a conceptual model of an intravascular drug delivery device – micro/nanoparticle used for search-and-destroy disease cells. With the lateral dimensions of 1 µm or less, the particle is smaller than any blood cells. These particles can be injected into the blood stream and travel freely through the circulatory system. In order to direct these drug-delivery micro/nanoparticles to cancer sites, their external surfaces are chemically modified to carry molecules that have lock-and-key binding specificity with molecules that support a growing cancer mass. As soon as the particles dock on the cells, a compound is released that forms a pore on the membrane of the cells, which leads to cell death and ultimately to that of the cancer mass that was nourished by the blood vessel. Adhesion between micro/nanodevices and disease cells is required.

33.3.3 Tribological Needs

The MEMS/NEMS need to be designed to perform expected functions typically in the ms to ps range. Expected life of the devices for high speed contacts can vary from a few hundred thousand to many billions of cycles, e.g., over a hundred billion cycles for DMDs, which puts serious requirements on materials [33.10, 62, 75–78]. Most mechanical properties are known to be scale-dependent [33.79]. The properties of nanoscale structures need to be measured [33.80]. Tribology is an important factor affecting the performance and reliability of MEMS/NEMS [33.10, 49, 81]. There is a need for developing a fundamental understanding of adhesion, friction/stiction, wear, and the roles of surface contamination and environment [33.10]. MEMS/NEMS materials need to exhibit good mechanical and tribological properties on the micro/nanoscale. There is a need to develop lubricants and identify lubrication methods that are suitable for MEMS/NEMS. Component-level studies are required to provide a better understanding of the tribological phenomena occurring in MEMS/NEMS. The emergence of micro/nanotribology and atomic force microscopy-based techniques has provided researchers with a viable approach to address these problems [33.49, 82]. This chapter presents a review of macro- and micro/nanoscale tribological studies of materials and lubrication studies for MEMS/NEMS and component-level studies of stiction phenomena in MEMS/NEMS devices.

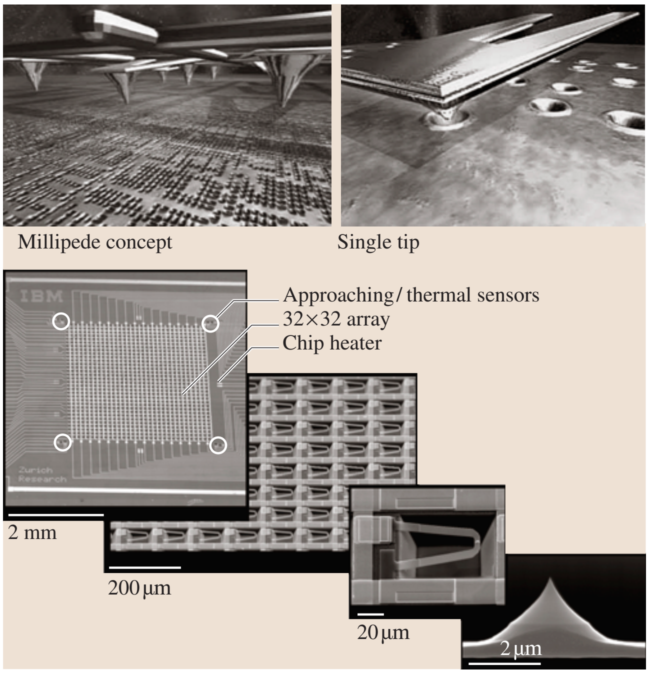

Fig. 33.8 Example of a NEMS device – AFM-based nanoscale data storage system with 32×32 tip array – that experiences tribological problems [33.72]

33.4 Tribological Studies of Silicon and Related Materials

Materials of most interest for planar fabrication processes using silicon as the structural material are undoped and boron-doped (p^+-type) single-crystal silicon for bulk micromachining and phosphorus (n^+-type) doped and undoped LPCVD polysilicon films for surface micromachining. Silicon-based devices lack high-temperature capabilities with respect to both mechanical and electrical properties. Researchers have been pursuing SiC as a material for high-temperature microsensor and microactuator applications, for some time [33.83, 84]. SiC is a likely candidate for such applications, since it has long been used in high temperature electronics, high frequency and high power devices. SiC can also be desirable for high-frequency micromechanical resonators, in the GHz range, because of its high modulus of elasticity and, consequently, high resonant frequency. Table 33.2 compares selected bulk properties of SiC and Si(100). Researchers have found low cost techniques of producing single-crystal 3C-SiC (cubic or β-SiC) films via epitaxial growth on large-area silicon substrates for bulk micromachining [33.85] and polycrystalline 3C-SiC films on polysilicon and silicon dioxide layers for surface micromachining of SiC [33.86].

Table 33.2 Selected bulk properties[a] of 3C (β- or cubic) SiC and Si(100)

Sample	Density (kg/m^3)	Hardness (GPa)	Elastic modulus (GPa)	Fracture toughness ($MPa\,m^{1/2}$)	Thermal conductivity[b] (W/m K)	Coeff. of thermal expansion[b] ($\times 10^{-6}$/°C)	Melting point (°C)	Band gap (eV)
β-SiC	3,210	23.5–26.5	440	4.6	85–260	4.5–6	2,830	2.3
Si(100)	2,330	9 –10	130	0.95	155	2 –4.5	1,410	1.1

[a] Unless stated otherwise, data shown were obtained from [33.87]

[b] Obtained from [33.88]

As will be shown, bare silicon exhibits inadequate tribological performance and needs to be coated with a solid and/or liquid overcoat or surface treated (e.g., oxidation and ion implantation, commonly used in semiconductor manufacturing), which exhibit lower friction and wear. SiC films exhibit good tribological performance. Both macroscale and microscale tribological properties of virgin and treated/coated silicon, polysilicon films, and SiC are presented next.

33.4.1 Tribological Properties of Silicon and the Effect of Ion Implantation

Friction and wear of single-crystalline and polycrystalline silicon samples, as well as the effect of ion implantation with various doses of C^+, B^+, N_2^+, and Ar^+ ion species at 200 keV energy to improve their friction and wear properties have been studied [33.89–91]. The coefficient of macroscale friction and wear factor of virgin single-crystal silicon and C^+-implanted silicon samples as a function of ion dose are presented in Fig. 33.9 [33.89]. The macroscale friction and wear tests were conducted using a ball-on-flat tribometer. Each data bar represents the average value of four to six measurements. The coefficient of friction and wear factor for bare silicon are very high and decrease drastically with ion dose. Silicon samples bombarded above the ion dose of 10^{17} C^+ cm^{-2} exhibit extremely low values of coefficients of friction (typically 0.03 to 0.06 in air) and wear factor (reduced by as much as four orders of magnitude). *Gupta* et al. [33.89] reported that a decrease in coefficient of friction and wear factor of silicon as a result of C^+ ion bombardment occurred because of formation of silicon carbide, rather than amorphization of silicon. *Gupta* et al. [33.90] also reported an improvement in friction and wear with B^+ ion implantation.

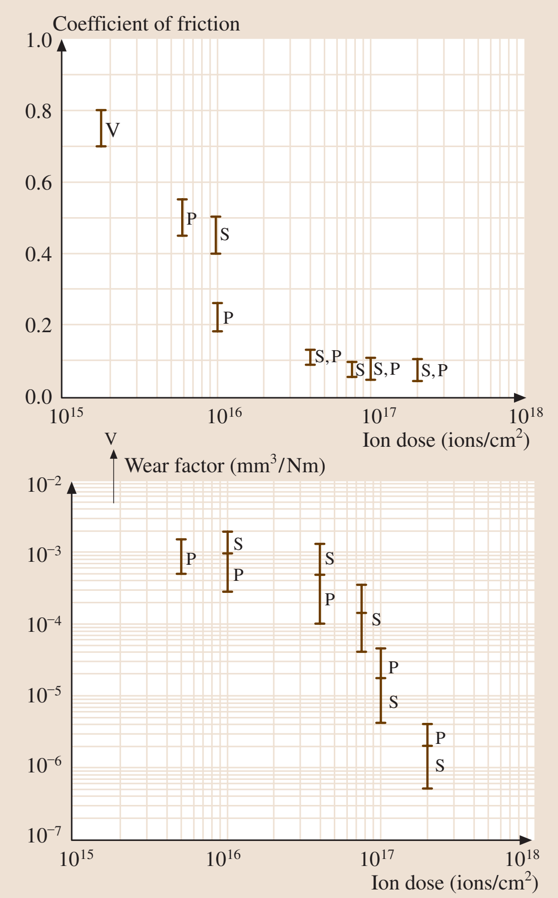

Fig. 33.9 Influence of ion doses on the coefficient of friction and wear factor on C^+ ion bombarded single-crystal and polycrystalline silicon slid against an alumina ball. V corresponds to virgin single-crystal silicon, while S and P denote tests that correspond to doped single- and polycrystalline silicon, respectively [33.89]

Table 33.3 Surface roughness and micro- and macroscale coefficients of friction of selected samples

Material	RMS roughness (nm)	Coefficient of microscale friction[a]	Coefficient of macroscale friction[b]
Si(111)	0.11	0.03	0.33
C^+-implanted Si(111)	0.33	0.02	0.18

[a] Versus Si_3N_4 tip, tip radius of 50 nm in the load range of 10–150 nN (2.5–6.1 GPa) at a scanning speed of 5 μm/s over a scan area of 1 μm × 1 μm in an AFM

[b] Versus Si_3N_4 ball, ball radius of 3 mm at a normal load of 0.1 N (0.3 GPa) at an average sliding speed of 0.8 mm/s using a tribometer

Microscale friction measurements were performed using an atomic force/friction force microscope (AFM/FFM) [33.49]. Table 33.3 shows values of surface roughness and coefficients of macroscale and microscale friction for virgin and doped silicon. There is a decrease in coefficients of microscale and macroscale friction values as a result of ion implantation. When measured for small contact areas and very low loads used in microscale studies, the indentation hardness and elastic modulus are higher than at the macroscale. This, added to the effect of the small apparent area of contact reducing the number of trapped particles on the interface, results in less plowing contribution and lower friction in the case of microscale friction measurements. Results of microscale wear resistance studies of ion-implanted silicon samples studied using a diamond tip in an AFM [33.92] are shown in Fig. 33.10a, b. For tests conducted at various loads on Si(111) and C^+-implanted Si(111), it is noted that wear resistance of an implanted sample is slightly poorer than that of virgin silicon up to about 80 μN. Above 80 μN, the wear resistance of implanted Si improves. As one continues to run tests at 40 μN for a larger number of cycles, the implanted sample exhibits higher wear resistance than the unimplanted sample. Damage from the implantation in the top layer results in poorer wear resistance. However, the implanted zone at the subsurface is more wear resistant than the virgin silicon.

Hardness values of virgin and C^+-implanted Si(111) at various indentation depths (normal loads) are presented in Fig. 33.10c [33.92]. The hardness at a small indentation depth of 2.5 nm is 16.6 GPa and drops

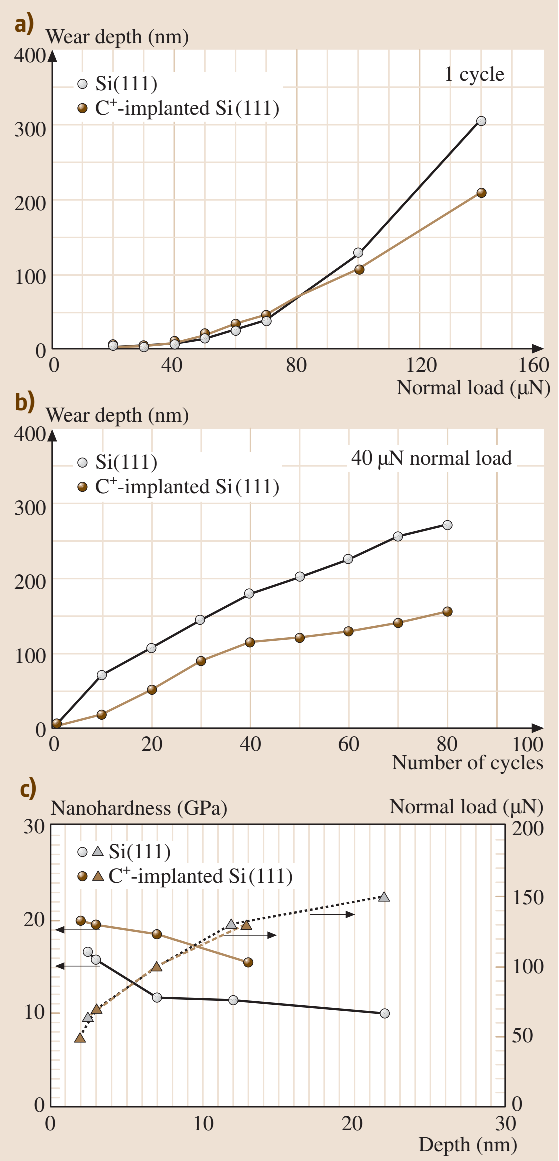

Fig. 33.10a–c Wear depth as a function of **(a)** load (after one cycle), and **(b)** cycles (normal load = 40 mN) for Si(111) and C^+-implanted Si(111). **(c)** Nanohardness and normal load as a function of indentation depth for virgin and C^+-implanted Si(111) [33.92]

to a value of 11.7 GPa at a depth of 7 nm and a normal load of 100 μN. Higher hardness values obtained in low-load indentation may arise from the observed pressure-induced phase transformation during the nanoindentation [33.93, 94]. An additional increase in the hardness at an even lower indentation depth of 2.5 nm reported here may arise from the contribution of complex chemical films (not from native oxide films) present on the silicon surface. At small volumes there is a lower probability of encountering material defects (dislocations, etc.). Furthermore, according to the strain gradient plasticity theory advanced by *Fleck* et al. [33.95], large strain gradients inherent in small indentations lead to the accumulation of geometrically necessary dislocations that cause enhanced hardening. These are some of the plausible explanations for an increase in hardness at smaller volumes. If the silicon material were to be used at very light loads such as in microsystems, the high hardness of surface films would protect the surface until it is worn.

From Fig. 33.10c, hardness values of C^+-implanted Si(111) at a normal load of 50 μN is 20.0 GPa with an indentation depth of about 2 nm, which is comparable to the hardness value of 19.5 GPa at 70 μN, whereas measured hardness value for virgin silicon at an indentation depth of about 7 nm (normal load of 100 μN) is only about 11.7 GPa. Thus, ion implantation with C^+ results in an increase in hardness in silicon. Note that the surface layer of the implanted zone is much harder compared to the subsurface and may be brittle, leading to higher wear on the surface. The subsurface of the implanted zone is harder than the virgin silicon, resulting in higher wear resistance, which is also observed in the results of the macroscale tests conducted at high loads.

33.4.2 Effect of Oxide Films on Tribological Properties of Silicon

Macroscale friction and wear experiments have been performed using a magnetic disk drive with bare, oxidized, and implanted pins sliding against amorphous-carbon coated magnetic disks lubricated with a thin layer of perfluoropolyether lubricant [33.96–99]. Representative profiles of the variation of the coefficient of friction with a number of sliding cycles for Al_2O_3-TiC slider and bare and dry-oxidized silicon pins are shown in Fig. 33.11. For bare Si(111), after an initial increase, the coefficient of friction drops to a steady state value of 0.1 following the increase, as seen in Fig. 33.11. The rise in the coefficient of friction for the Si(111) pin is associated with the transfer of amorphous carbon from the disk to the pin and the oxidation-enhanced fracture of pin material followed by tribochemical oxidation of the transfer film, while the drop is associated with the formation of a transfer coating on the pin, as shown in Fig. 33.12. Dry-oxidized Si(111) exhibits excellent characteristics, and no significant increase was observed over 50,000 cycles (Fig. 33.11). This behavior has been attributed to the chemical passivity of the oxide and lack of transfer of DLC from the disk to the pin. The behavior of PECVD-oxide (data are not presented here) was comparable to that of dry oxide, but for the wet oxide there was some variation in the coefficient of friction (0.26 to 0.4). The difference between dry and wet oxide was attributed to increased

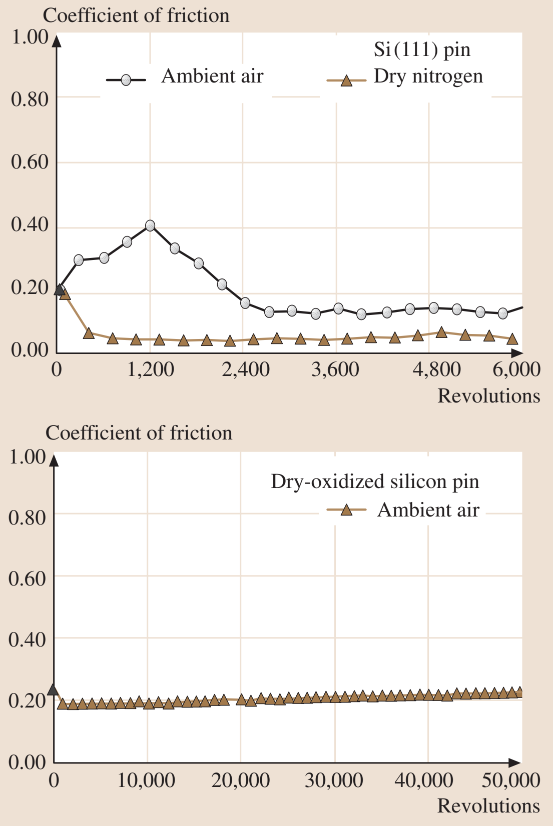

Fig. 33.11 Coefficient of friction as a function of number of sliding revolutions in ambient air for Si(111) pin in ambient air and dry nitrogen and dry-oxidized silicon pin in ambient air [33.96]

Ambient air, 6,000 cycles

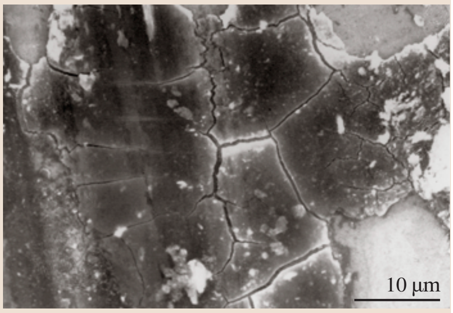

Dry nitrogen, 15,000 cycles

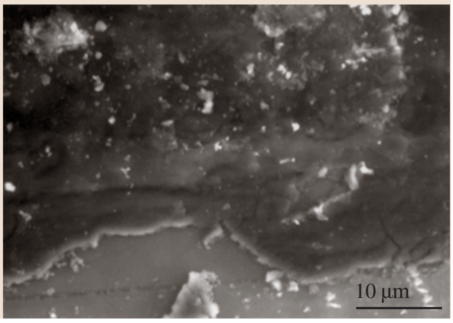

Fig. 33.12 Scanning electron micrographs of Si(111) after sliding against a magnetic disk in a rigid disk drive in ambient air for 6,000 cycles and in dry nitrogen after 15,000 cycles

porosity of the wet oxide [33.96]. Since tribochemical oxidation was determined to be a significant factor, experiments were conducted in dry nitrogen [33.97, 98]. The variation of the coefficient of friction for a silicon pin sliding against a thin-film disk in dry nitrogen is shown in Fig. 33.11. It is seen that in a dry nitrogen environment, the coefficient of friction of Si(111) sliding against a disk decreased from an initial value of about 0.2 to 0.05 with continued sliding. Based on SEM and chemical analysis, this behavior has been attributed to the formation of a smooth amorphous-carbon/lubricant transfer patch and suppression of oxidation in a dry nitrogen environment. Based on macroscale tests using disk drives, it was found that the friction and wear performance of bare silicon is not adequate. With dry-oxidized or PECVD SiO_2-coated silicon, no significant friction increase or interfacial degradation was observed in ambient air.

Table 33.4 and Fig. 33.13 show surface roughness, microscale friction and scratch data, and nanoindentation hardness for various silicon samples [33.92]. Scratch experiments were performed using a diamond tip in an AFM. Results on polysilicon samples are also shown for comparison. Coefficients of microscale friction values for all the samples are about the same. These samples could be scratched at 10 µN load. Scratch depth increased with normal load. Crystalline orientation of silicon has little influence on

Table 33.4 RMS, microfriction, microscratching/microwear, and nanoindentation hardness data for various virgin, coated, and treated silicon samples

Material	RMS roughness[a] (nm)	Coefficient of microscale friction[b]	Scratch depth[c] at 40 µN (nm)	Wear depth[c] at 40 µN (nm)	Nanohardness[c] at 100 µN (GPa)
Si(111)	0.11	0.03	20	27	11.7
Si(110)	0.09	0.04	20		
Si(100)	0.12	0.03	25		
Polysilicon	1.07	0.04	18		
Polysilicon (lapped)	0.16	0.05	18	25	12.5
PECVD-oxide coated Si(111)	1.50	0.01	8	5	18.0
Dry-oxidized Si(111)	0.11	0.04	16	14	17.0
Wet-oxidized Si(111)	0.25	0.04	17	18	14.4
C^+-implanted Si(111)	0.33	0.02	20	23	18.6

[a] Scan size of 500 nm × 500 nm using AFM.

[b] Versus Si_3N_4 tip in AFM/FFM, radius 50 nm; at 1 µm × 1 µm scan size.

[c] Measured using an AFM with a diamond tip of radius 100 nm.

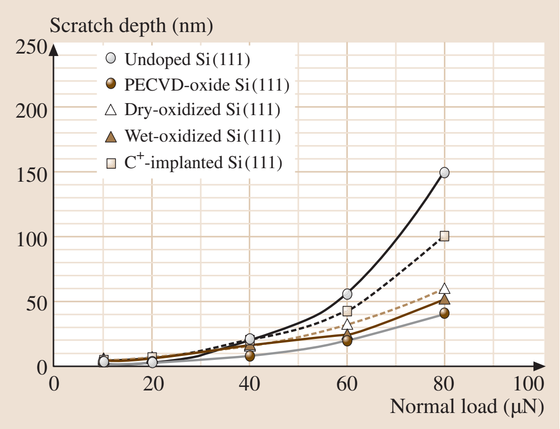

Fig. 33.13 Scratch depth as a function of normal load after 10 cycles for various silicon samples: virgin, treated, and coated [33.92]

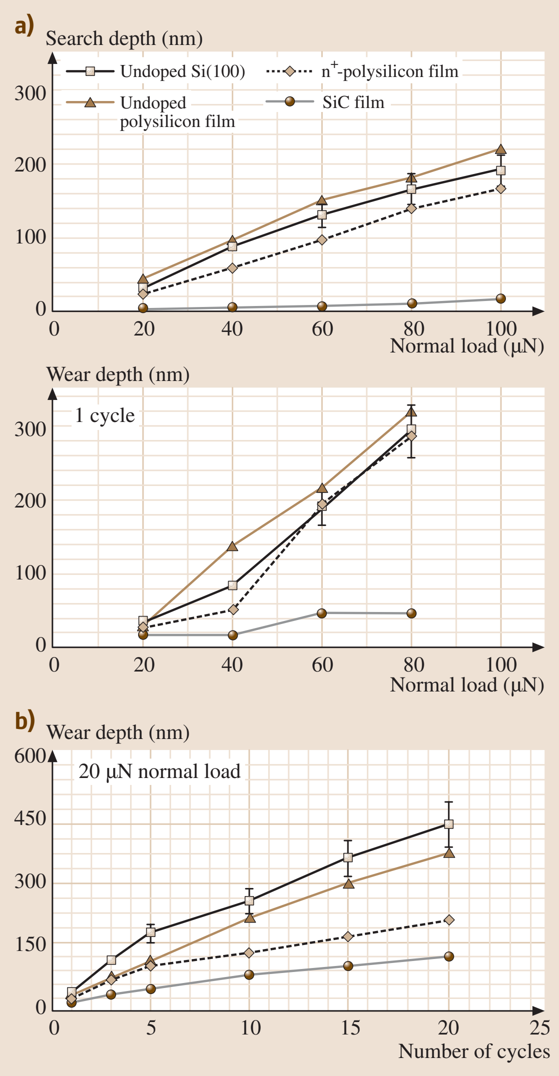

Fig. 33.14 **(a)** Scratch depths for 10 cycles as a function of normal load and **(b)** wear depths as a function of normal load and as a function of number of cycles for various samples [33.10]

scratch resistance, because natural oxidation of silicon in ambient masks the expected effect of crystallographic orientation. PECVD-oxide samples showed the best scratch resistance, followed by dry-oxidized, wet-oxidized, and ion-implanted samples. Ion implantation with C^+ does not appear to improve scratch resistance.

Wear data on the silicon samples are also presented in Table 33.1 [33.92]. PECVD-oxide samples showed superior wear resistance followed by dry-oxidized, wet-oxidized, and ion-implanted samples. This agrees with the trends seen in scratch resistance. In PECVD, ion bombardment during deposition improves the coating properties such as suppression of columnar growth, freedom from pinhole, decrease in crystalline size, and increase in density, hardness, and substrate-coating adhesion. These effects may help in improving the mechanical integrity of the sample surface. Coatings and treatments improved nanohardness of silicon. Note that dry-oxidized and PECVD films are harder than wet-oxidized films, as these films may be porous. High hardness of oxidized films may be responsible for measured high scratch/wear resistance.

33.4.3 Tribological Properties of Polysilicon Films and SiC Film

Studies have also been conducted on undoped polysilicon film, heavily doped (n^+-type) polysilicon film, heavily doped (p^+-type) single-crystal Si(100), and 3C-SiC (cubic or β-SiC) film [33.10, 100, 101]. The polysilicon films studied here are different from the ones discussed previously.

Table 33.5 presents a summary of the tribological studies conducted on polysilicon and SiC films. Values for single-crystal silicon are shown for comparison. Polishing of the as-deposited polysilicon and SiC films drastically affects the roughness, as the values reduce by two orders of magnitude. Si(100) appears to be

Table 33.5 Summary of micro/nanotribological properties of the sample materials

Sample	RMS roughness[a] (nm)	P–V distance[a] (nm)	Coefficient of friction		Scratch depth[d] (nm)	Wear depth[e] (nm)	Nano-hardness[f] (GPa)	Young's modulus[f] (GPa)	Fracture toughness[g], K_{IC} MPa m$^{1/2}$
			Micro[b]	Macro[c]					
Undoped Si(100)	0.09	0.9	0.06	0.33	89	84	12	168	0.75
Undoped polysilicon film (as deposited)	46	340	0.05						
Undoped polysilicon film (polished)	0.86	6	0.04	0.46	99	140	12	175	1.11
n^+-Type polysilicon film (as deposited)	12	91	0.07						
n^+-Type polysilicon film (polished)	1.0	7	0.02	0.23	61	51	9	95	0.89
SiC film (as deposited)	25	150	0.03						
SiC film (polished)	0.89	6	0.02	0.20	6	16	25	395	0.78

[a] Measured using AFM over a scan size of 10 µm × 10 µm

[b] Measured using AFM/FFM over a scan size of 10 µm × 10 µm

[c] Obtained using a 3-mm-diameter sapphire ball in a reciprocating mode at a normal load of 10 mN and average sliding speed of 1 mm/s after 4 m sliding distance

[d] Measured using AFM at a normal load of 40 µN for 10 cycles, scan length of 5 µm

[e] Measured using AFM at normal load of 40 µN for 1 cycle, wear area of 2 µm × 2 µm

[f] Measured using nanoindenter at a peak indentation depth of 20 nm

[g] Measured using microindenter with Vickers indenter at a normal load of 0.5 N

the smoothest followed by polished undoped polysilicon and SiC films, which have comparable roughness. The doped polysilicon film shows higher roughness than the undoped sample, which is attributed to the doping process. Polished SiC film shows the lowest friction followed by polished and undoped polysilicon film, which strongly supports the candidacy of SiC films for use in MEMS/NEMS devices. Macroscale friction measurements indicate that SiC film exhibits one of the lowest friction values as compared to the other samples. The doped polysilicon sample shows low friction on the macroscale as compared to the undoped polysilicon sample, possibly due to the doping effect.

Figure 33.14a shows a plot of scratch depth vs. normal load for various samples [33.10, 100]. Scratch depth increases with increasing normal load. Figure 33.15 shows AFM 3-D maps and averaged 2-D profiles of the scratch marks on the various samples. It is observed that scratch depth increases almost linearly with the normal load. Si(100) and the doped and undoped polysilicon films show similar scratch resistance. It is clear from the data that the SiC film is much more scratch resistant than the other samples. Figure 33.14b shows results from microscale wear tests on the various films. For all the materials, the wear depth increases almost linearly with increasing number of cycles. This suggests that the material is removed layer by layer in all the materials. Here also SiC film exhibits lower wear depths than the other samples. Doped polysilicon film wears less than the undoped film. Higher fracture toughness and hardness of SiC compared to Si(100) are responsible for its lower wear. Also the higher thermal conductivity of SiC (see Table 33.2) compared to the other materials leads to lower interface temperatures, which generally results in less degradation of the surface [33.35, 54, 55]. Doping of the polysilicon does not affect the scratch/wear resistance and hardness much. The measurements made on the doped sample are affected by the presence of grain boundaries. These studies indicate that SiC film exhibits desirable tribological properties for use in MEMS devices. Recently, researchers have fabricated SiC micromotors and have reported satisfactory operation at high temperatures [33.102].

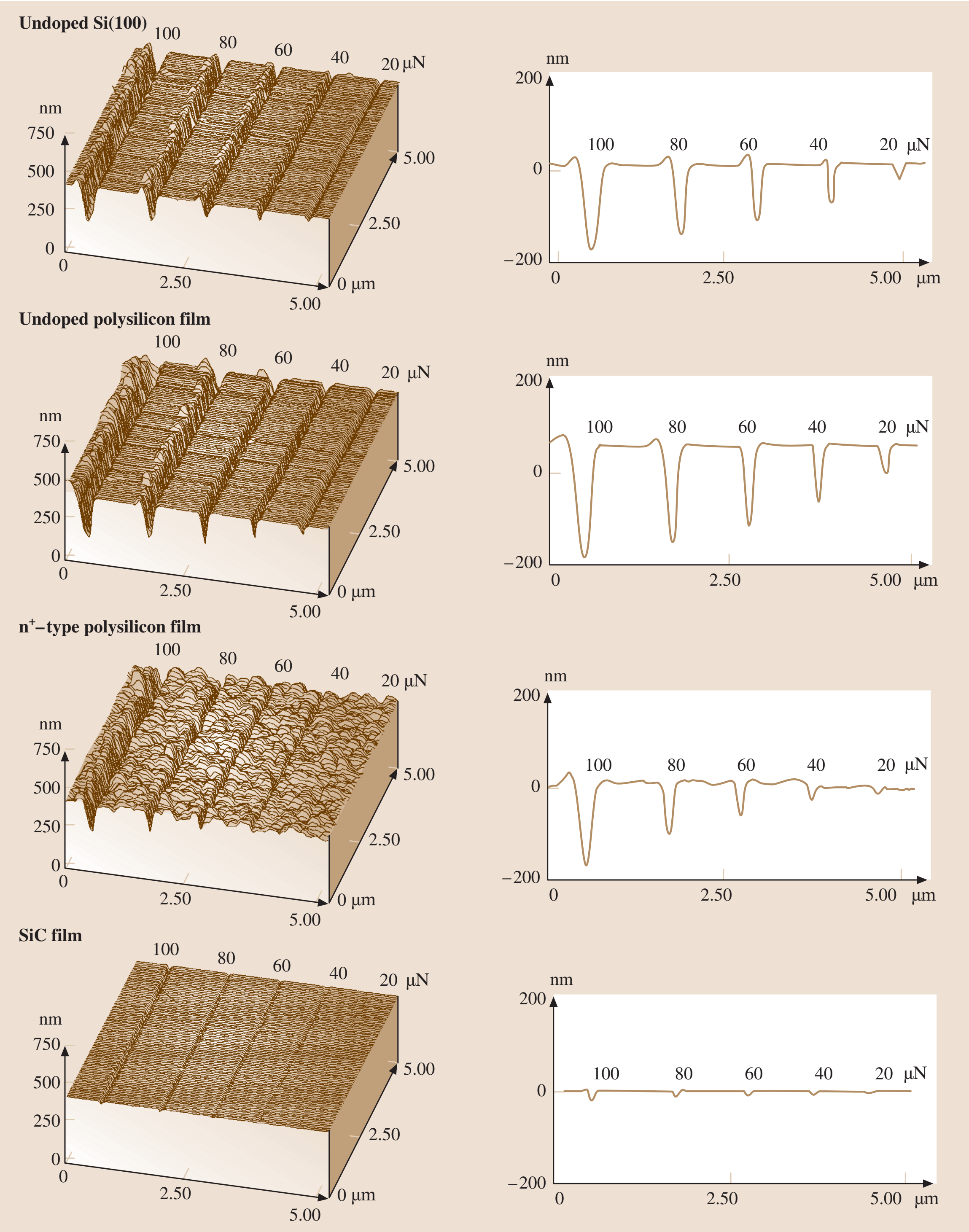

Fig. 33.15 AFM 3-D maps and averaged 2-D profiles of scratch marks on various samples [33.10]

33.5 Lubrication Studies for MEMS/NEMS

Several studies of liquid perfluoropolyether (PFPE) lubricant films, self-assembled monolayers (SAMs), and hard diamond-like carbon (DLC) coatings have been carried out for the purpose of minimizing adhesion, friction, and wear [33.49, 66, 99, 103–108]. Many variations of these films are hydrophobic (low surface tension and high contact angle) and have low shear strength, which provide low adhesion, friction, and wear. Relevant details are presented here.

33.5.1 Perfluoropolyether Lubricants

The classical approach to lubrication uses freely supported multimolecular layers of liquid lubricants [33.49, 54, 55]. The liquid lubricants are sometimes chemically bonded to improve their wear resistance. Partially chemically bonded, molecularly thick perfluoropolyether (PFPE) lubricants are widely used for lubrication of magnetic storage media [33.35] and are suitable for MEMS/NEMS devices.

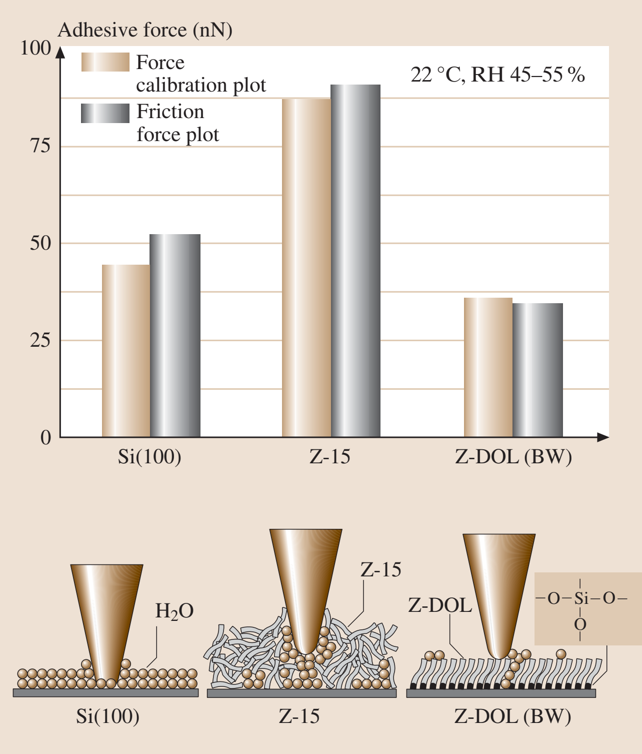

Fig. 33.16 Summary of the adhesive forces of Si(100) and Z-15 and Z-DOL (BW) films measured by force calibration plots and friction force versus normal load plots in ambient air. The schematic (*bottom*) showing the effect of meniscus, formed between the AFM tip and the surface sample, on the adhesive and friction forces [33.103]

Adhesion, friction, and durability experiments have been performed on virgin Si(100) surfaces and silicon surfaces lubricated with two commonly used PFPE lubricants – Z-15 (with -CF_3 nonpolar end groups) and Z-DOL (with -OH polar end groups) [33.49, 103, 105, 106]. Z-DOL film was thermally bonded at 150 °C for 30 min and an unbonded fraction was removed by a solvent (BW) [33.35]. The thicknesses of Z-15 and Z-DOL (BW) films were 2.8 nm and 2.3 nm, respectively. Nanoscale measurements were made using an AFM. The adhesive forces of Si(100), Z-15 and Z-DOL (BW) measured by force calibration plot and friction force versus normal load plot are summarized in Fig. 33.16. The results measured by these two methods are in good agreement. Figure 33.16 shows that the presence of mobile Z-15 lubricant film increases the adhesive force compared to that of Si(100) by meniscus formation [33.54, 55, 109]. The presence of solid phase Z-DOL (BW) film reduces the adhesive force compared to that of Si(100) due to the absence of mobile liquid. The schematic (bottom) in Fig. 33.16 shows the relative size and sources of meniscus. It is well-known that the native oxide layer (SiO_2) on the top of Si(100) wafer exhibits hydrophilic properties, and some water molecules can be adsorbed on this surface. The condensed water will form meniscus as the tip approaches the sample surface. The larger adhesive force in Z-15 is not only caused by the Z-15 meniscus. The nonpolarized Z-15 liquid does not have good wettability and strong bonding with Si(100). In the ambient environment, the condensed water molecules from the environment permeate through the liquid Z-15 lubricant film and compete with the lubricant molecules presented on the substrate. The interaction of the liquid lubricant with the substrate is weakened, and a boundary layer of the liquid lubricant forms puddles [33.105, 106]. This dewetting allows water molecules to be adsorbed on the Si(100) surface as aggregates, along with Z-15 molecules. And both of them can form meniscus while the tip approaches the surface. Thus, the dewetting of liquid Z-15 film results in higher adhesive force and poorer lubrication performance. In addition, as the Z-15 film is pretty soft compared to the solid Si(100) surface, penetration of the tip in the film occurs while pushing the tip downwards. This leads to the large area of the tip involved to form the meniscus at the tip-liquid (mixture of Z-15 and water) interface. It should also be noted that Z-15 has a higher viscosity compared to water, and therefore, Z-15 film provides higher resistance to motion and coefficient of friction. In the case of Z-DOL (BW) film, the active groups of Z-DOL molecules are mostly bonded on Si(100) substrate, thus the Z-DOL (BW) film has low free surface energy and cannot be displaced readily by water molecules, or readily adsorb water molecules. Thus, the use of Z-DOL (BW) can reduce the adhesive force.

To study the relative humidity effect on friction and adhesion, the variations of friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL (BW) as a function of relative humidity are shown in Fig. 33.17. It shows that for Si(100) and Z-15 film, the friction force increases with a relative humidity increase of up to 45% and then slightly decreases with a further increase in relative humidity. Z-DOL (BW) has a smaller friction force than Si(100) and Z-15 in the whole testing range. And its friction force shows a relative apparent increase when the relative humidity is higher than 45%. For Si(100), Z-15, and Z-DOL (BW), the adhesive forces increase with relative humidity. And their coefficients of friction increase with a relative humidity up to 45%, after which they decrease with a further increase in relative humidity. It is also observed that the humidity effect on Si(100) really depends on the history of the Si(100) sample. As the surface of Si(100) wafer readily adsorbs water in air, without any pre-treatment the Si(100) used in our study almost reaches its saturated stage of adsorbed water and is responsible for less effect during increasing relative humidity. However, once the Si(100) wafer was thermally treated by baking at 150 °C for 1 h, a bigger effect was observed.

The schematic (right) in Fig. 33.17 shows that Si(100), because its high free-surface energy, can adsorb more water molecules with increasing relative humidity. As discussed earlier, for the Z-15 film in the humid environment, the condensed water from the humid environment competes with the lubricant film present on the sample surface, and interaction of the liquid lubricant film with the silicon substrate is weakened, and a boundary layer of the liquid lubricant forms puddles. This dewetting allows water molecules to be adsorbed on the Si(100) substrate mixed with Z-15 molecules [33.105, 106]. Obviously, more water molecules can be adsorbed on Z-15 surface while increasing relative humidity. The more adsorbed water molecules in the case of Si(100), along with lubricant molecules in Z-15 film case, form bigger water meniscus, which leads to an increase in friction force, adhesive force, and coefficient of friction of Si(100) and Z-15 with humidity. But at a very high humidity of 70%, large quantities of adsorbed water can form a continuous water layer that separates the tip and sample surface and acts as a kind of lubricant, which causes a decrease in the friction force and coefficient of friction. For Z-DOL (BW) film, because of its hydrophobic surface properties, water molecules can be adsorbed at a humidity higher than 45% and cause an increase in the adhesive and friction forces.

To study the durability of lubricant films at the nanoscale, the friction of Si(100), Z-15, and Z-DOL (BW) as a function of the number of scanning cycles are shown in Fig. 33.18. As observed earlier, friction force and coefficient of friction of Z-15 is higher than that of Si(100) with the lowest values for Z-DOL (BW). During cycling, friction force and coefficient of friction of Si(100) show a slight decrease during the initial few cycles, then remain constant. This is related to the removal of the top adsorbed layer. In the case of Z-15 film, the friction force and coefficient of friction show an increase during the initial few cycles and then approach higher and stable values. This is believed to be caused by the attachment of the Z-15 molecules onto the tip. The molecular interaction between these attached molecules and molecules on the film surface is responsible for an increase in friction. But after several scans, this molecular interaction reaches the equilibrium, and after that, friction force and coefficient of friction remain constant. In the case of Z-DOL (BW) film, the friction force and coefficient of friction start out low and remain low during the entire test for 100 cycles, suggesting that Z-DOL (BW) molecules do not get attached or displaced as readily as Z-15.

33.5.2 Self-Assembled Monolayers (SAMs)

A preferred method of lubrication of MEMS/NEMS is by the deposition of organized and dense molecular-scale layers of long-chain molecules, as they have been shown to be superior lubricants [33.54, 55, 104, 107, 108, 110, 111]. Two common methods of producing monolayers are the Langmuir–Blodgett (L-B) deposition and self-assembled monolayers (SAMs) by chemical grafting of molecules. L-B films are physically bonded to the substrate by weak van der Waals forces, while SAMs are bonded covalently to the substrate and provide high durability.

SAMs can be formed spontaneously by the immersion of an appropriate substrate into a solution of an

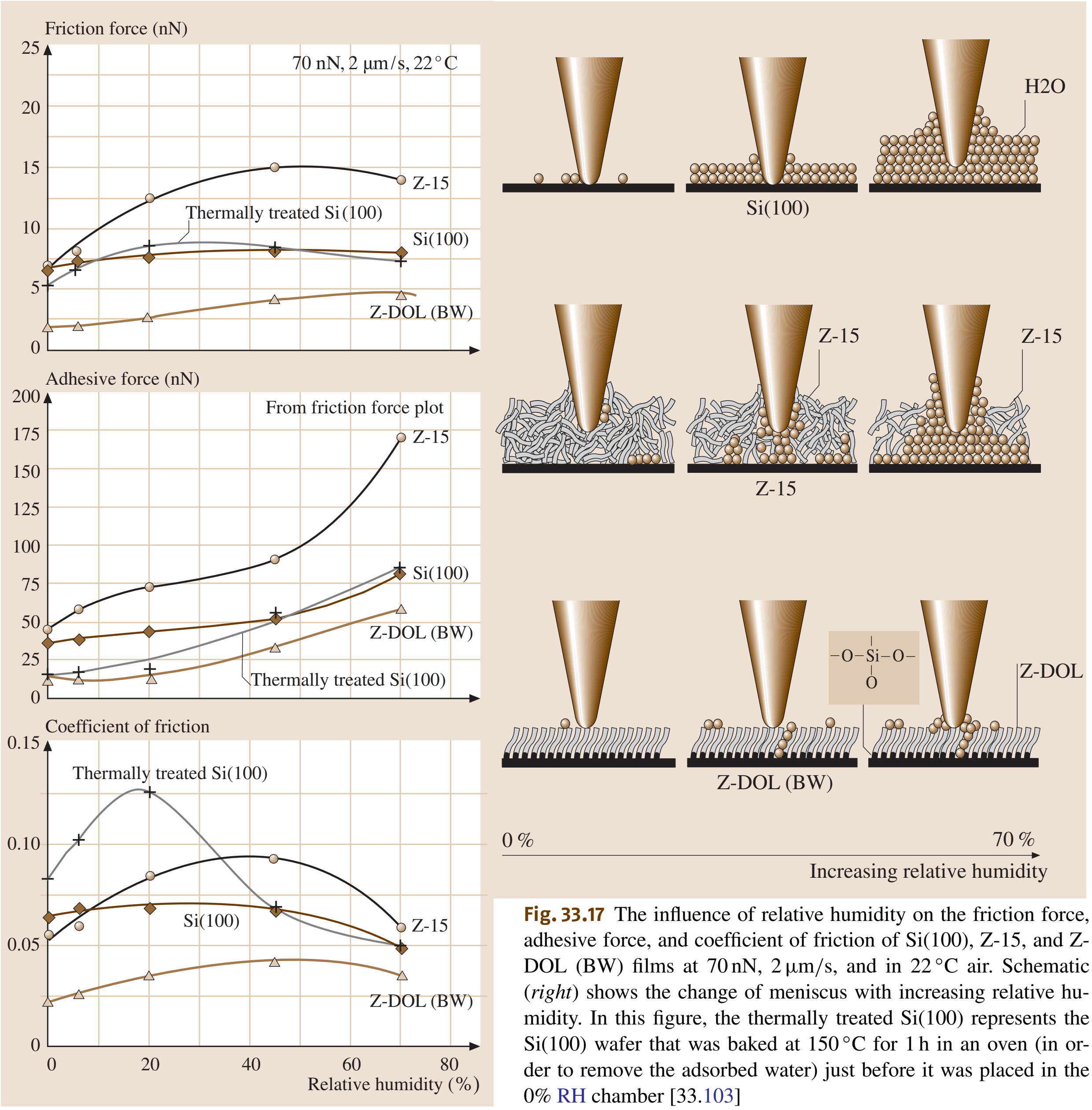

Fig. 33.17 The influence of relative humidity on the friction force, adhesive force, and coefficient of friction of Si(100), Z-15, and Z-DOL (BW) films at 70 nN, 2 µm/s, and in 22 °C air. Schematic (*right*) shows the change of meniscus with increasing relative humidity. In this figure, the thermally treated Si(100) represents the Si(100) wafer that was baked at 150 °C for 1 h in an oven (in order to remove the adsorbed water) just before it was placed in the 0% RH chamber [33.103]

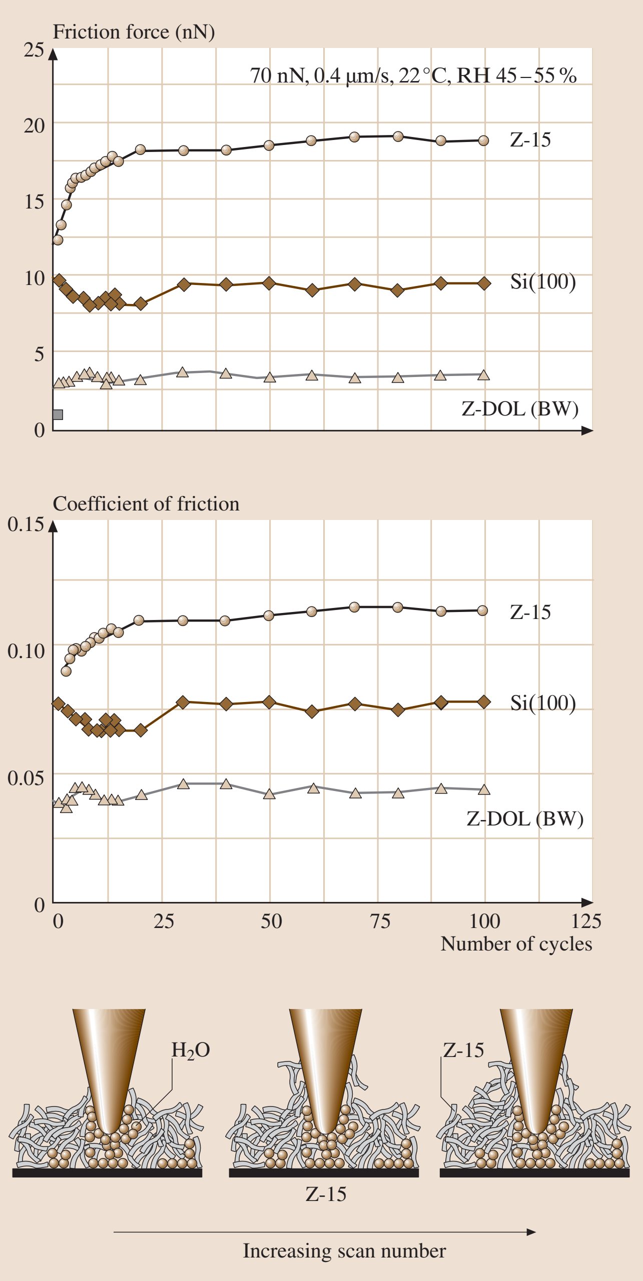

Fig. 33.18 Friction force and coefficient of friction versus number of sliding cycles for Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, 0.4 μm/s, and in ambient air. Schematic (*bottom*) shows that some liquid Z-15 molecules can be attached onto the tip. The molecular interaction between the attached molecules with the Z-15 molecules in the film results in an increase of the friction force with multi-scanning [33.103]

active surfactant in an organic solvent. SAMs offer the flexibility and advantage of molecular tailoring to obtain a variety of different tribological and mechanical properties. For example, researchers have shown that by changing the head groups, tail groups, chain lengths, or types of bonds within a chain, varying degrees of friction, adhesion and/or compliance can be obtained (e.g., [33.107, 108]). These studies indicate that the basis for molecular design and tailoring of SAMs must include a complete understanding of the interrelationships between the molecular structure and tribological properties of SAMs, as well as a deep understanding of the friction and wear mechanisms of SAMs at the molecular level.

Bhushan and *Liu* [33.107] and *Liu* and *Bhushan* [33.108] studied nanotribological properties of four different kinds of alkylthiol and biphenyl thiol monolayers with different surface terminals, spacer chains, and head groups using AFM/FFM techniques. These monolayers, along with a schematic of their structures and substrates, are listed in Fig. 33.19. Surface roughness, adhesion, and microscale friction studies were carried out using an AFM/FFM. The average values and standard deviation of the adhesive force and coefficient of microscale friction for the various SAMs are summarized in Fig. 33.20. It shows that SAMs can reduce the adhesive and friction forces. Alkylthiol and biphenyl thiol SAMs can be used as effective molecular lubricants for MEMS/NEMS fabricated from silicon. In order to explain the frictional differences in SAMs, a molecular spring model is presented in Fig. 33.21. It is suggested that the chemically adsorbed self-assembled molecules on a substrate are just like assembled molecular springs anchored to the substrate. An AFM tip sliding on the surface of a SAM is like a tip sliding on the top of "molecular springs or brush." The molecular assembly has compliant features and can experience orientation and compression under load. The orientation of the molecular springs under normal load reduces the shearing force at the interface, which, in turn, reduces the friction force. The possibility of orientation is determined by the spring constant of a single molecule (local stiffness), as well as the interaction between the neighboring molecules, which is reflected by the packing density or packing energy. It is noted that HDT exhibits the lowest adhesive and friction forces. Based on local stiffness measurements [33.108] and the view of molecular structures, biphenyl is a more rigid structure due to the contribution by two benzene rings. Therefore, the spring constant of BPT is larger than that of HDT. A more compliant film (HDT) exhibits lower friction force than a rigid film (BPT/BPTC).

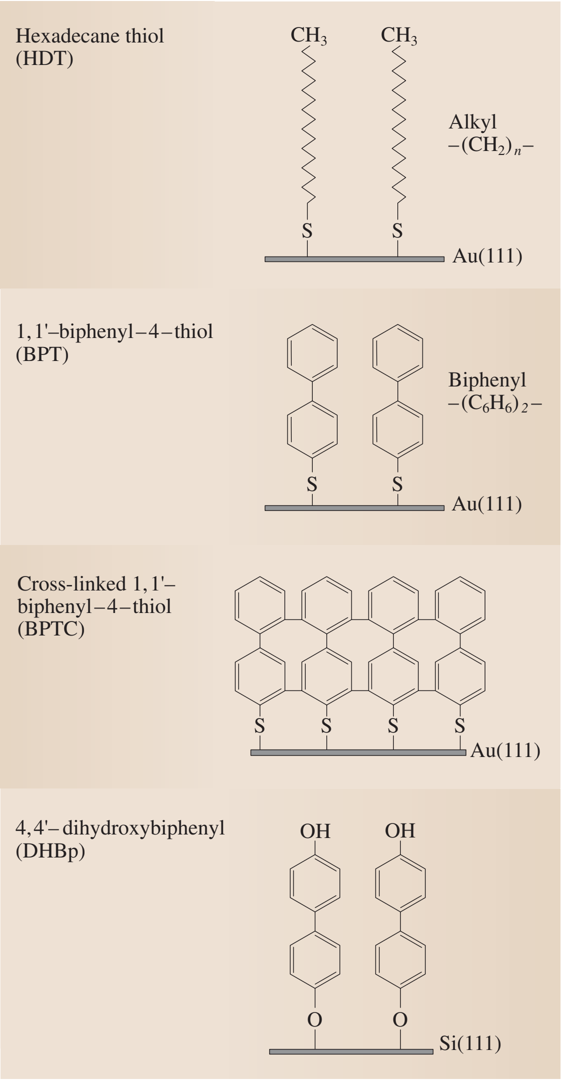
Hexadecane thiol (HDT)
CH_3
CH_3
Alkyl $-(CH_2)_n-$
S
S
Au(111)
1,1'–biphenyl–4–thiol (BPT)
Biphenyl $-(C_6H_6)_2-$
S
S
Au(111)
Cross-linked 1,1'–biphenyl–4–thiol (BPTC)
S
S
S
S
Au(111)
4,4'–dihydroxybiphenyl (DHBp)
OH
OH
O
O
Si(111)

Fig. 33.19 Structure of SAMs studied using AFM/FFM techniques

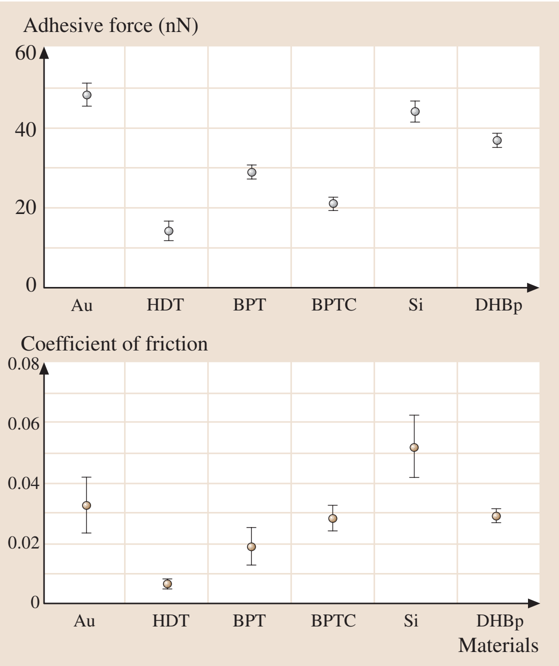

Fig. 33.20 Adhesive force and coefficient of friction values for the various SAMs studied

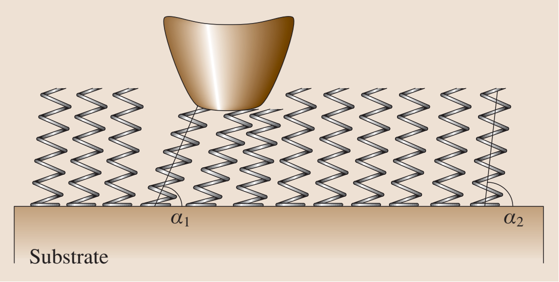
α_1
α_2
Substrate

Fig. 33.21 Molecular spring model of SAMs. In this figure, $\alpha_1 < \alpha_2$, which is caused by the further orientation under the normal load applied by an asperity tip [33.107]

Crosslinking of chains in BPTC resulted in higher coefficient of friction. It should be noted that the orientation can lead to conformational defects along the molecular chains that lead to energy dissipation.

The influence of relative humidity on adhesion and friction was studied in an environmentally controlled chamber. The results are given in Fig. 33.22, which show that for Si(111), Au(111), and DHBp, the adhesive and frictional forces increase with relative humidity. For BPT and BPTC (cross-linked BPT), the adhesive forces slightly increase with relative humidity when the relative humidity is higher than 40%, but it is very interesting that their coefficients of friction decrease slightly in the same range. For HDT, over the testing range, the adhesive and friction force do not seem to be sensitive to the change in relative humidity. The influence of relative humidity on adhesive and frictional forces can be mainly understood by comparing their surface terminal polarization properties and work of adhesion. Meniscus forces are related linearly to the works of adhesion of a surface, which in

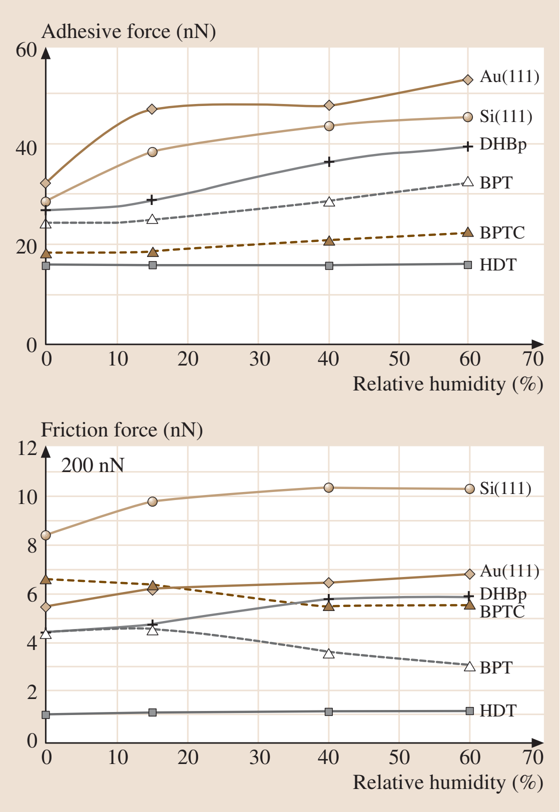

Fig. 33.22 The influence of relative humidity on the adhesive and friction forces at 200 nN of various SAMs studied [33.107]

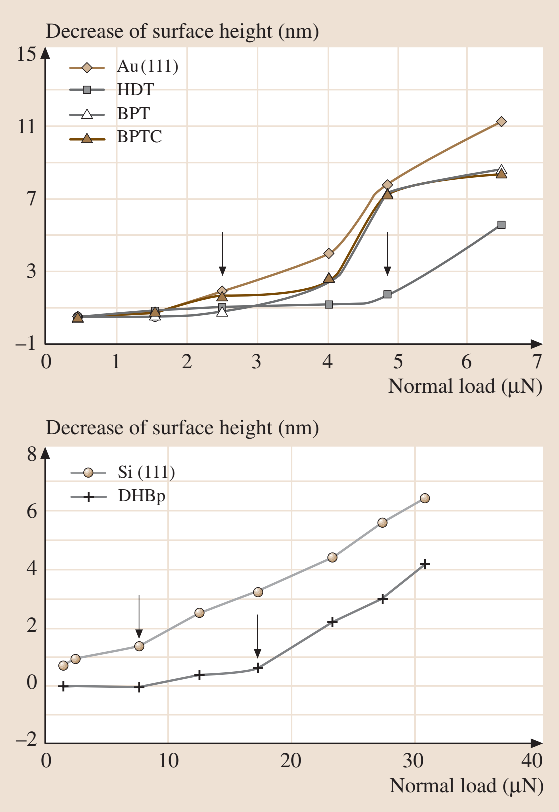

Fig. 33.23 Wear depth as a function of normal load for various SAMs [33.107]

the case of SAMs, is dependent on the surface terminals. Polar surface terminals such as DHBp on Si(111) result in higher work of adhesion and hence higher adhesive forces. Larger adhesive forces result in higher friction forces. Nonpolar surface terminals (HDT) have very small work of adhesion and hence low adhesive and friction forces. In higher humidity, water capillary condensation can either increase friction through increased adhesion in the contact zone, or reduce friction through an enhanced water-lubricating effect.

Figure 33.23 shows results of microscale wear tests on SAMs using a diamond tip in an AFM. DHBp on Si(111) shows the best wear resistance. For the SAMs deposited on gold substrate, HDT exhibits the best wear resistance. Scratch tests corroborated the wear results. The wear resistance of SAMs is influenced by the interfacial bond strength, the molecular structure of the spacers, and the substrate hardness. Based on the wear data, the SAMs with high-compliance long carbon chains, in addition to low friction, exhibit the best wear resistance [33.107, 108]. In wear experiments, the wear depth as a function of normal load curves shows critical normal load (Fig. 33.23). Below the critical normal load, SAMs undergo orientation. At the critical load, SAMs wear away from the substrate due to weak interfacial bond strength, while above the critical normal load, severe wear takes place on the substrate, Fig. 33.24.

According to these results, it is suggested that a dual SAM with a compliant layer on a stiff layer, deposited on hydrogenated silicon, may have optimized tribological performance for MEMS/NEMS applications.

33.5.3 Hard Diamond-like Carbon (DLC) Coatings

Hard amorphous carbon (a-C), commonly known as DLC (implying high hardness), coatings are deposited by a variety of deposition techniques including

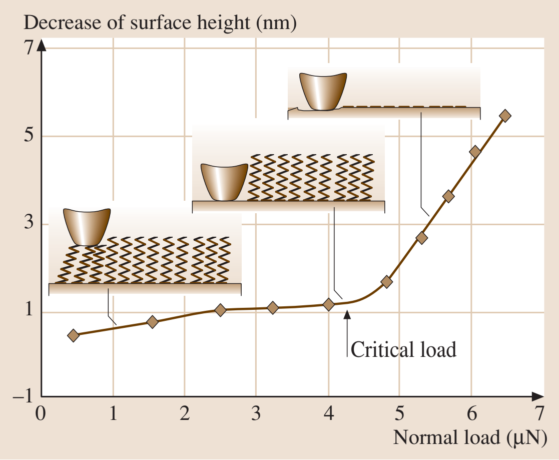

Fig. 33.24 Illustration of the wear mechanism of SAMs with increasing normal load [33.108]

filtered cathodic arc, ion beam, electron cyclotron resonance chemical vapor deposition (ECR-CVD), plasma-enhanced chemical vapor deposition (PECVD), and sputtering [33.83, 99]. These coatings are used in a wide range of applications, including tribological, optical, electronic, and biomedical applications. Ultrathin coatings (3.5 to 10 nm thick) are employed to protect against wear and corrosion in magnetic storage applications (thin-film rigid disks, metal evaporated tapes) and thin-film read/write head (Gillette Mach 3 razor blades, glass windows, and sunglasses). The coatings exhibit low friction, high hardness and wear resistance, chemical inertness to both acids and alkalis, lack of magnetic response, and optical band gap ranging from zero to a few eV, depending on the deposition technique and its conditions. Selected data on DLC coatings relevant for MEMS/NEMS applications are presented in a following section on adhesion measurements.

33.6 Component-Level Studies

33.6.1 Surface Roughness Studies of Micromotor Components

Most of the friction forces resisting motion in the micromotor are concentrated near the rotor-hub interface, where continuous physical contact occurs. Surface roughness of the surfaces usually has a strong influence on the friction characteristics on the micro/nanoscale. A catalog of roughness measurements on various components of a MEMS device does not exist in the literature. Using an AFM, measurements on various component surfaces was made for the first time by *Sundararajan* and *Bhushan* [33.112].

Table 33.6 shows various surface roughness parameters obtained from 5×5 µm scans of the various component surfaces of several unlubricated micromotors using the AFM in tapping mode. A surface with a Gaussian height distribution should have a skewness of zero and kurtosis of three. Although the rotor and stator top surfaces exhibit comparable roughness pa-

Table 33.6 Surface roughness parameters and microscale coefficient of friction for various micromotor component surfaces measured using an AFM. Mean and $\pm 1\sigma$ values are given

	RMS roughness[a] (nm)	Peak-to-valley distance[a] (nm)	Skewness[a], Sk	Kurtosis[a], K	Coefficient of microscale friction[b] µ
Rotor topside	21 ± 0.6	225 ± 23	1.4 ± 0.30	6.1 ± 1.7	0.07 ± 0.02
Rotor underside	14 ± 2.4	80 ± 11	−1.0 ± 0.22	3.5 ± 0.50	0.11 ± 0.03
Stator topside	19 ± 1	246 ± 21	1.4 ± 0.50	6.6 ± 1.5	0.08 ± 0.01

[a] Measured from a tapping mode AFM scan of size 5 µm × 5 µm using a standard Si tip scanning at 5 µm/s in a direction orthogonal to the long axis of the cantilever

[b] Measured using an AFM in contact mode at 5 µm × 5 µm scan size using a standard Si_3N_4 tip scanning at 10 µm/s in a direction parallel to the long axis of the cantilever

rameters, the underside of the rotors exhibits lower RMS roughness and peak-to-valley distance values. More importantly, the rotor underside shows negative skewness and lower kurtosis than the topsides both of which are conducive to high real area of contact and hence high friction [33.54, 55]. The rotor underside also exhibits higher coefficient of microscale friction than the rotor topside and stator, as shown in Table 33.6. Figure 33.25 shows representative surface height maps of the various surfaces of a micromotor measured using the AFM in tapping mode. The rotor underside exhibits varying topography from the outer edge to the middle and inner edge. At the outer edges, the topography shows smaller circular asperities, similar to the topside. The middle and inner regions show deep pits with fine edges that may have been created by the etchants used for etching the sacrificial layer. It is known that etching can affect the surface roughness of surfaces in surface micromachining. The residence time of the etchant near the inner region is high, responsible for larger pits. Figure 33.26 shows roughness of the surface directly beneath the rotors (the base polysilicon layer). There appears to be a difference in the roughness between the portion of this surface that was initially underneath the rotor (region B) during fabrication and the portion that was away from the rotor and hence always exposed (region A). The former region shows lower roughness than the latter region. This suggests that the surfaces at the rotor-hub interface that come into contact at the end of the fabrication process exhibit large real areas of contact that result in high friction.

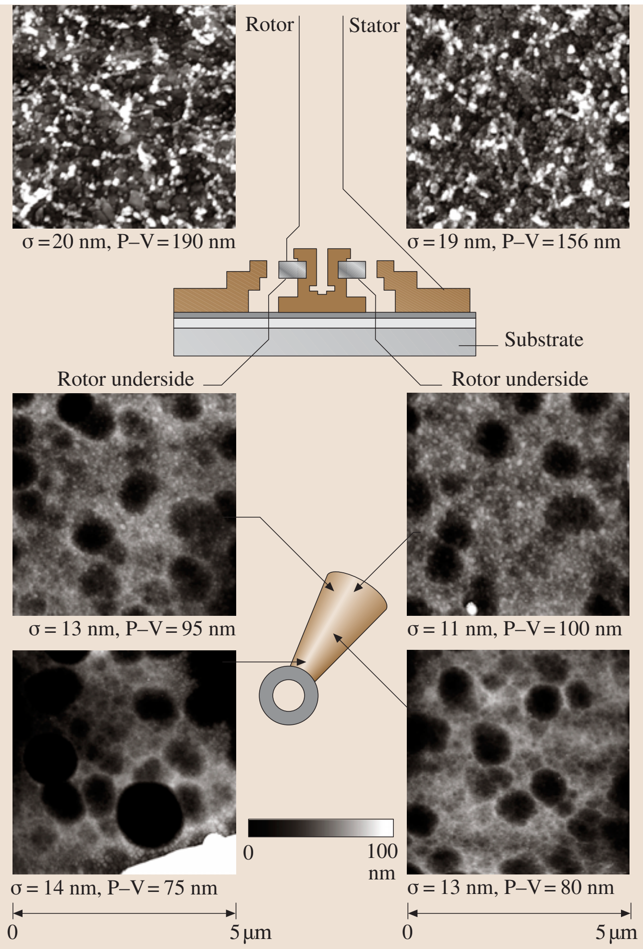

Fig. 33.25 Representative AFM surface height images obtained in tapping mode (5 µm × 5 µm scan size) of various component surfaces of a micromotor. RMS roughness and peak-to-valley (P-V) values of the surfaces are given. The underside of the rotor exhibits drastically different topography from the topside [33.112]

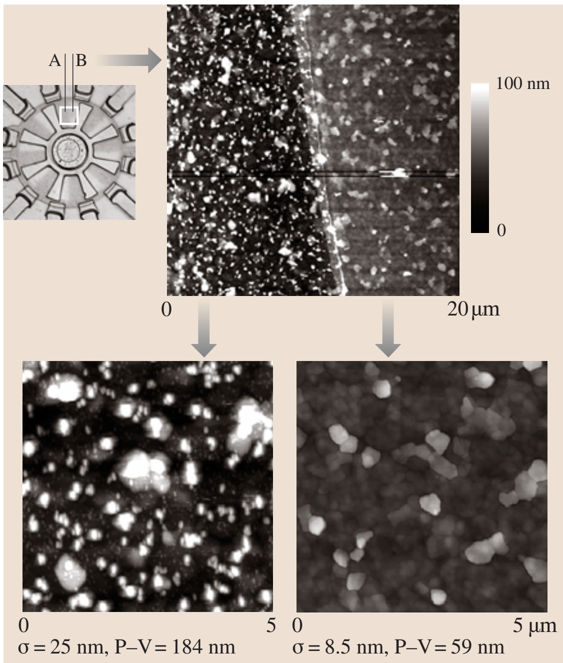

Fig. 33.26 Surface height images of polysilicon regions directly below the rotor. Region A is away from the rotor, while region B was initially covered by the rotor prior to the release etch of the rotor. During this step, slight movement of the rotor caused region B to be exposed [33.112]

33.6.2 Adhesion Measurements

Surface force apparatus (SFA) and AFMs are used to measure adhesion between two surfaces on micro- to nanoscales. In the SFA, the adhesion of liquid films sandwiched between two curved and smooth surfaces is measured. In an AFM, as discussed earlier, adhesion between a sharp tip and the surface of interest is measured. To measure adhesion between two beams in the mesoscopic length scale, a cantilever beam array (CBA) technique is used [33.113, 115–117]. The technique utilizes an array of micromachined polysilicon beams (for Si MEMS applications) anchored to the substrate at one end and with different lengths parallel to the surface. It relies on the peeling and detachment of cantilever beams. Change in free energy, or reversible work done to separate unit areas of two surfaces from contact is called work of adhesion. To measure the work of adhesion, electrostatic actuation is used to bring all beams in contact with the substrate, Fig. 33.27 [33.113, 117]. Once the actuation force is removed, the beams begin to peel themselves off the substrate, which can be observed with an optical interference microscope. For beams shorter than a characteristic length, the so-called detachment length, their stiffness is sufficient to completely free them from the substrate underneath. Beams larger than the detachment length remain adhered. The beams at the transition region start to detach and remain attached to the substrate just at the tips. For this case, by equating the elastic energy stored within the beam and the beam-substrate interfacial energy, the work of adhesion, W_{ad}, can be calculated by the following equation [33.117]:

$$W_{\mathrm{ad}} = \frac{3Ed^2t^3}{8l_{\mathrm{d}}^4}\,, \qquad (33.1)$$

where E is the Young's modulus of the beam, d is the spacing between the undeflected beam and the substrate, t is the beam thickness, and l_{d} is the detachment length. The technique has been used to screen methods for adhesion reduction in polysilicon microstructures.

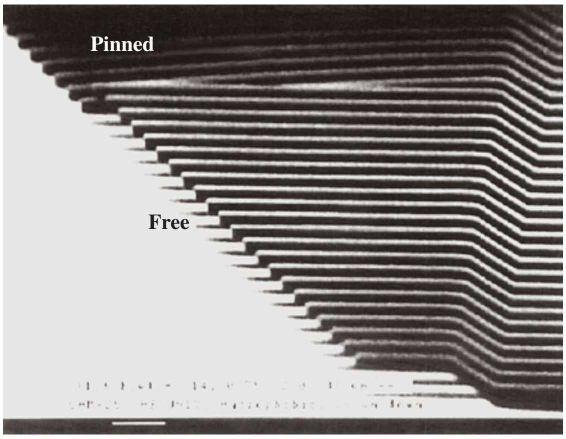

Fig. 33.27 SEM micrograph of a micromachined array of polysilicon cantilever beams of increasing length. The micrograph shows the onset of pinning for beams longer than 34 µm [33.113]

To measure adhesion, friction, and wear between two microcomponents, a microtriboapparatus has been used. Figure 33.28 shows a schematic of a microtriboapparatus capable of adopting MEMS components [33.114]. In this apparatus, an upper specimen mounted on a soft cantilever beam comes in contact with a lower specimen mounted on a lower specimen holder. The apparatus consists of two piezos (x- and z-piezos), and four fiber optic sensors (x- and z-displacement sensors, and x- and z-force sensors). For adhesion and friction studies, z- and x-piezos are used to bring the upper and lower specimens in contact and to apply a relative motion in the lateral direction, respectively. The x- and z-displacement sensors are used to measure the lateral position of the

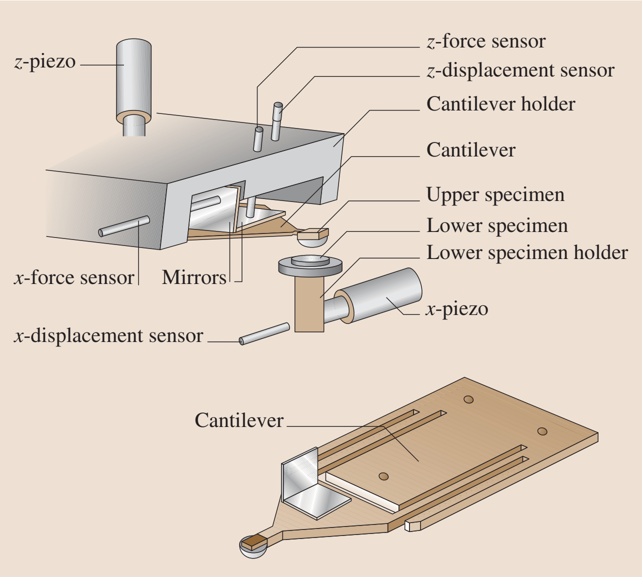

Fig. 33.28 Schematic of the microtriboapparatus, including specially designed cantilever (with two perpendicular mirrors attached on the end), lower specimen holder, two piezos (x- and z-piezos), and four fiber optic sensors (x- and z-displacement sensors and x- and z-force sensors) [33.114]

lower specimen and vertical position of the upper specimen, respectively. The x- and z-force sensors are used to measure friction force and normal load/adhesive force between these two specimens, respectively, by monitoring the deflection of the cantilever.

As most of the MEMS/NEMS devices are fabricated from silicon, the study of silicon-on-silicon contacts is important. This contact was simulated by a flat single-crystal Si(100) wafer (phosphorus doped) specimen sliding against a single crystal Si(100) ball (1 mm in diameter, 5×10^{17} atoms/cm^3 boron doped) mounted on a stainless steel cantilever [33.114, 118]. Both of them have a native oxide layer on their surfaces. The other materials studied were 10-nm-thick DLC deposited by filtered cathodic arc deposition on Si(100), 2.3-nm-thick chemically bonded PFPE (Z-DOL, BW) on Si(100), and hexadecane thiol (HDT) monolayer on evaporated Au(111) film to investigate their anti-adhesion performance.

It is well-known that in the computer rigid disk drives, the adhesive force increases rapidly with an increase in rest time between a magnetic head and a magnetic disk [33.35]. Considering that adhesion and friction are the major issues that lead to the failure of MEMS/NEMS devices, the rest time effect on microscale on Si(100), DLC, PFPE, and HDT was studied, and the results are summarized in Fig. 33.29a. It is found that the adhesive force of Si(100) increases logarithmically with the rest time to a certain equilibrium time ($t = 1{,}000$ s), after which it remains constant. Figure 33.29a also shows that the adhesive force of DLC, PFPE, and HDT does not change with rest time. Single-asperity contact modeling of the dependence of meniscus force on the rest time has been carried out by *Chilamakuri* and *Bhushan* [33.119], and the modeling results (Fig. 33.29b) verify experimental observations. Due to the presence of thin film adsorbed water on Si(100), meniscus forms around the contacting asperities and grows with time until equilibrium occurs, which causes the rest time effect on its adhesive force. The adhesive forces of DLC, PFPE, and HDT do not change with rest time, which suggests that either water meniscus is not present on their surfaces, or it does not increase with time.

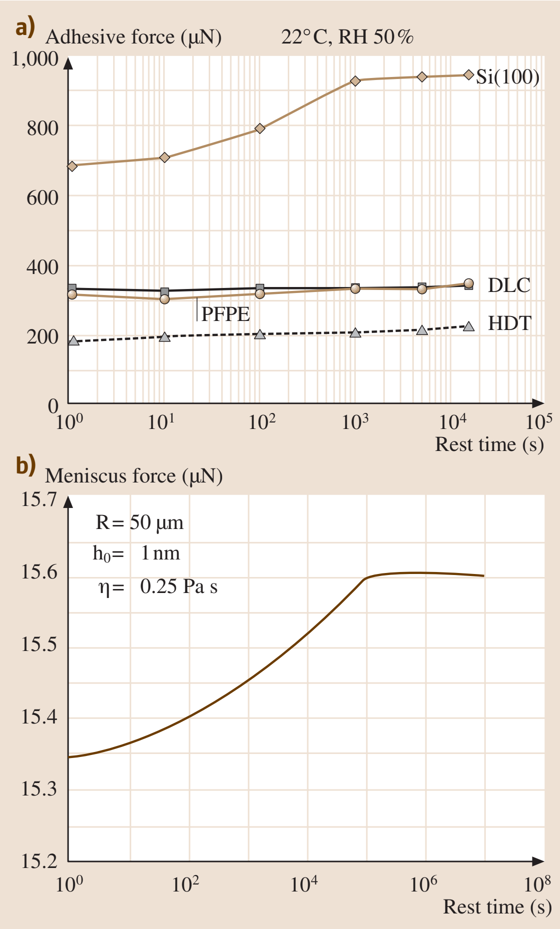

Fig. 33.29 **(a)** The influence of rest time on the adhesive force of Si(100), DLC, chemically bonded PFPE, and HDT, and **(b)** single asperity contact modeling results of the rest time effect on the meniscus force for an asperity of R in contact with a flat surface with a water-film thickness of h_0 and absolute viscosity of η_0 [33.119]

The measured adhesive forces of Si(100), DLC, PFPE, and HDT at rest time of 1 s are summarized in Fig. 33.30. It shows that the presence of solid films of DLC, PFPE, and HDT greatly reduces the adhesive force of Si(100), whereas HDT film has the lowest adhesive force. It is well-known that the native oxide layer (SiO_2) on the top of Si(100) wafer exhibits hydrophilic properties, and water molecules produced by capillary condensation of water vapor from the environment can be adsorbed easily on this surface. The condensed water will form meniscus as the upper specimen approaches the lower specimen surface. The meniscus force is a major contributor to the adhesive force. In the case of DLC, PFPE, and HDT, the films are found to be hydrophobic based on contact angle measurements, and the amount of condensed water vapor is low compared to that on Si(100). It should be noted that the measured adhesive force is generally higher than that measured in AFM,

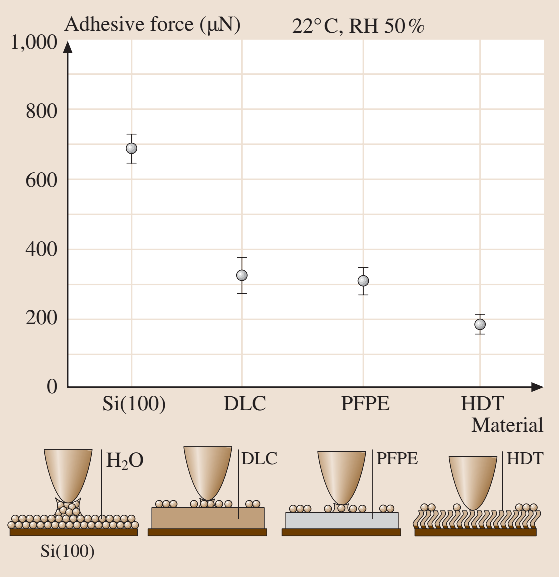

Fig. 33.30 Adhesive forces of Si(100), DLC, chemically bonded PFPE, and HDT at ambient conditions and a schematic showing the relative size of water meniscus on different specimens

because the larger radius of the Si(100) ball compared to that of an AFM tip induces larger meniscus and van der Waals forces.

To investigate the velocity effect on friction, the friction force as a function of velocity was measured and is summarized in Fig. 33.31a. It indicates that for Si(100) the friction force initially decreases with increasing velocity until equilibrium occurs. Figure 33.31a also indicates that the velocity almost has no effect on the friction properties of DLC, PFPE, and HDT. This implies that the friction mechanisms of DLC, PFPE, and HDT do not change with the variation in velocity. For Si(100), at high velocity, the meniscus is broken and does not have enough time to rebuild. In addition, it is believed that tribochemical reaction plays an important role. The high velocity leads to tribochemical reactions of Si(100), which has native oxide SiO_2, with water molecules to form a $Si(OH)_4$ film. This film is removed and continuously replenished during sliding. The $Si(OH)_4$ layer at the sliding surface is known to be of low shear strength. The breaking of the water meniscus and the formation of a $Si(OH)_4$ layer result in a decrease in friction force of Si(100). For DLC, PFPE, and HDT, surfaces exhibit hydrophobic properties and can only adsorb a few water molecules in ambient conditions. The above mentioned meniscus breaking and tribochemical reaction mechanisms do not exist for these films. Therefore, their friction force does not change with velocity.

The influence of relative humidity was studied in an environmentally controlled chamber. The adhesive and friction forces were measured at increasing relative humidity, and the results are summarized in Fig. 33.31b. The figure shows that for Si(100), the adhesive force increases with relative humidity, but the adhesive forces of DLC and PFPE only show slight increases when humidity is higher than 45%, while the adhesive force of HDT does not change with humidity. Figure 33.31b also shows that for Si(100), the friction force increases with a relative humidity increase up to 45%, and then it shows a slight decrease with a further increase in the relative humidity. For PFPE, there is an increase in the friction force when humidity is higher than 45%. In the whole testing range, relative humidity does not have any apparent influence on the friction properties of DLC and HDT. In the case of Si(100), the initial increase of relative humidity up to 45% causes more adsorbed water molecules and forms bigger water meniscus, which leads to an increase of friction force. But at a very high humidity of 65%, large quantities of adsorbed water can form a continuous water layer that separates the tip and sample surfaces and acts as a kind of lubricant, which causes a decrease in the friction force. For PFPE, dewetting of lubricant film at humidity larger than 45% results in an increase in the adhesive and friction forces. DLC and HDT surfaces show hydrophobic properties, and increasing relative humidity does not play much of a role in their friction forces.

The influence of temperature was studied using a heated stage. The adhesive force and friction force were measured at increasing temperatures from 22 °C to 125 °C. The results are presented in Fig. 33.31c. It shows that once the temperature is higher than 50 °C, increasing temperature causes a significant decrease in adhesive and friction forces of Si(100) and a slight decrease in the cases of DLC and PFPE. But the adhesion and friction forces of HDT do not show any apparent change with test temperature. At high temperature, desorption of water, and reduction of surface tension of water lead to the decrease of adhesive and friction forces of Si(100), DLC, and PFPE. However, in the case of HDT film, as only a few water molecules are adsorbed on the surface, the above mentioned mechanisms do not play a big role. Therefore, the adhesive and friction forces of HDT do not show any apparent change with temperature. Figure 33.31 shows that in the whole ve-

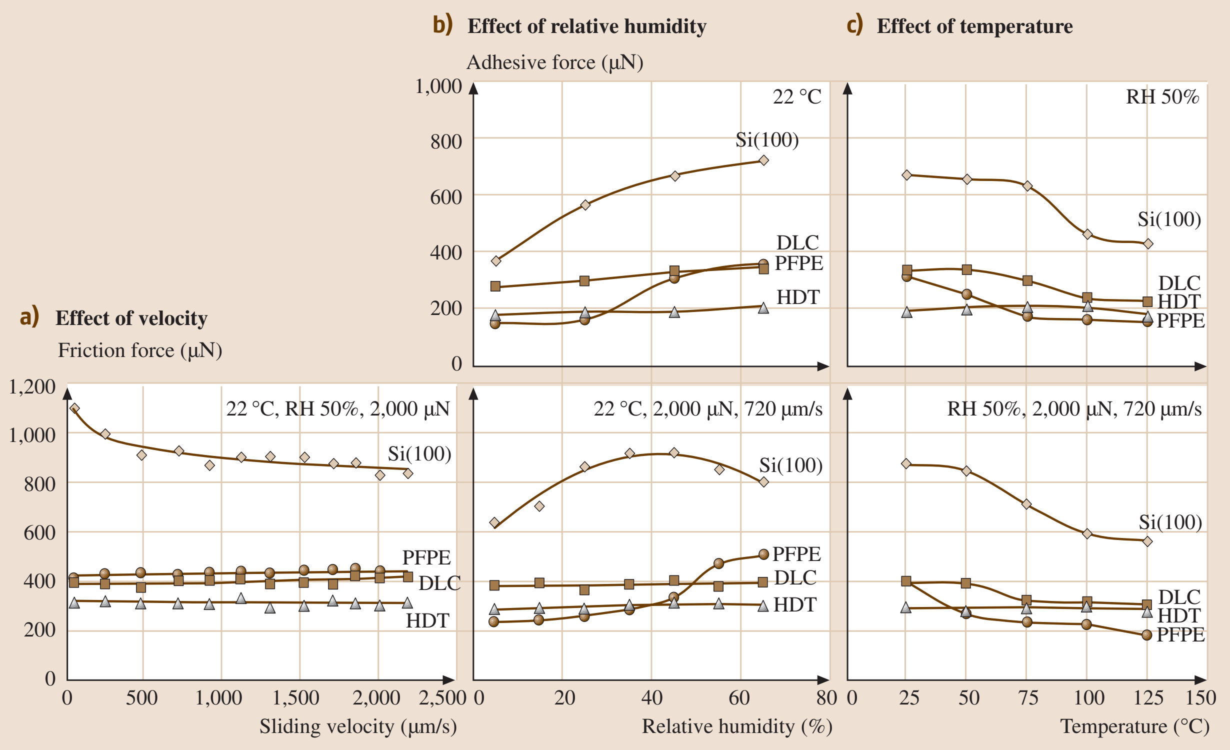

Fig. 33.31a–c The influence of **(a)** sliding velocity on the friction forces, **(b)** relative humidity on the adhesive and friction forces, and **(c)** temperature on the adhesive and friction forces of Si(100), DLC, chemically bonded PFPE, and HDT

locity, relative humidity, and temperature test range, the adhesive forces and friction forces of DLC, PFPE, and HDT are always smaller than that of Si(100), whereas HDT has the smallest value.

To summarize, several methods can be used to reduce adhesion in microstructures. MEMS surfaces can be coated with hydrophobic coatings such as PFPEs, SAMs, and passivated DLC coatings. It should be noted that other methods to reduce adhesion include the formation of dimples on the contact surfaces to reduce contact area [33.10, 54, 55, 116, 117]. Furthermore, an increase in hydrophobicity of the solid surfaces (high contact angle approaching 180°) can be achieved by using surfaces with suitable roughness, in addition to lowering their surface energy [33.120, 121]. The hydrophobicity of surfaces is dependent on a subtle interplay between surface chemistry and mesoscopic topography. The self-cleaning mechanism, or so-called "Lotus-effect", is closely related to the ultra-hydrophobic properties of these biological surfaces, which usually show microsculptures in specific dimensions.

33.6.3 Static Friction Force (Stiction) Measurements in MEMS

In MEMS devices involving parts in relative motion to each other, such as micromotors, large friction forces become the limiting factor to their successful operation and reliability. It is generally known that most micromotors cannot be rotated as manufactured and require some form of lubrication. It is therefore critical to determine the friction forces present in such MEMS devices. To measure in situ the static friction of a rotor-bearing interface in a micromotor, *Tai* and *Muller* [33.122] measured the starting torque (voltage) and pausing position for different starting positions under a constant-bias voltage. A friction-torque model was used to obtain the coefficient of static friction. To measure the in situ kinetic friction of the turbine and gear structures, *Gabriel* et al. [33.123] used a laser-based measurement system to monitor the steady-state spins and decelerations. *Lim* et al. [33.124] designed and fabricated a polysilicon microstructure to in situ measure the static friction of various films. The microstructure consisted of a shuttle suspended above the underlying electrode by a folded beam suspension. A known normal force was applied, and lateral force was measured to obtain the coefficient of static friction. *Beerschwinger* et al. developed a cantilever-deflection rig to measure friction of LIGA-processed micromotors [33.125, 126]. Table 33.7 presents static friction coefficients of various MEMS devices evaluated by various researchers. Most of these techniques employ indirect methods to determine the friction forces, or involve fabrication of complex structures.

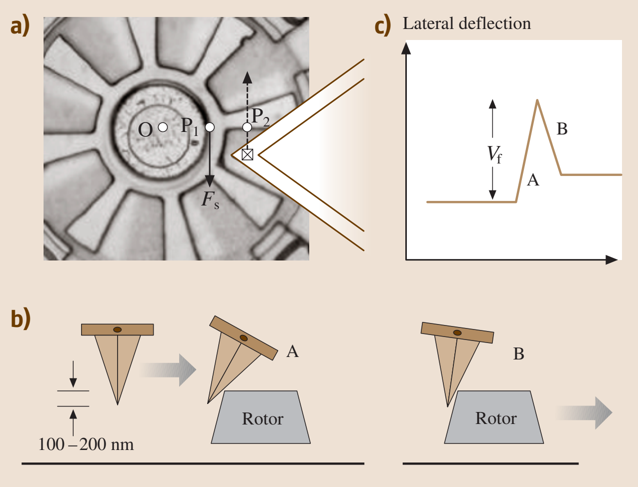

Fig. 33.32 **(a)** Schematic of the technique used to measure the force, F_s, required to initiate rotor movement using an AFM/FFM. **(b)** As the tip is pushed against the rotor, the lateral deflection experienced by the rotor due to the twisting of the tip prior to rotor movement is a measure of static friction force, F_s, of the rotors. **(c)** Schematic of lateral deflection expected from the above mentioned experiment. The peak V_f is related to the state of the rotor [33.112]

A novel technique to measure the static friction force (stiction) encountered in surface micromachined polysilicon micromotors using an AFM has been developed by *Sundararajan* and *Bhushan* [33.112]. Continuous physical contact occurs during rotor movement (rotation) in the micromotors between the rotor and lower hub flange. In addition, contact occurs at other locations between the rotor and the hub surfaces and between the rotor and the stator. Friction forces will be present at these contact regions during motor operation. Although the actual distribution of these forces is not known, they can be expected to be concentrated near the hub, where there is continuous contact. If we, therefore, represent the static friction force of the micromotor as a single force F_s acting at point P_1 (as shown in Fig. 33.32a), then the magnitude of the frictional torque about the center of the motor (O) that must be overcome before rotor movement can be initiated is:

$$T_s = F_s l_1 \, , \tag{33.2}$$

where l_1 is the distance OP_1, which is assumed to be the average distance from the center at which the friction force F_s occurs. Now consider an AFM tip moving against a rotor arm in a direction perpendicular to the long axis of the cantilever beam (the rotor arm edge closest to the tip is parallel to the long axis of the cantilever beam), as shown in Fig. 33.32a. When the tip encounters the rotor at point P_2, the tip will twist, generating a lateral force between the tip and the rotor (event A in Fig. 33.32b). This reaction force will generate a torque about the center of the motor. Since the tip is trying to move farther in the direction shown, the tip will continue to twist to a maximum value at which the lateral force between the tip and the rotor becomes high enough such that the resultant torque T_f about the center of the motor equals the static friction torque T_s. At this point, the ro-

Table 33.7 Published data on coefficient of static friction measurements of MEMS devices and structures

Reference	Test method	Device/structure	Material pairs	Environment	Coefficient of static friction
[33.122]	Starting voltage	IC-processed micromotor	PolySi/Si_3N_4	Air	0.20–0.40
[33.124]	Electrostatic loading	Comb-drive microstructure	PolySi/PolySi PolySi/Si_3N_4	Air	4.9 ± 1.0 2.5 ± 0.5
[33.127]	Pull-off force	Silicon microbeams	SiO_2/SiO_2	Air	2.1 ± 0.8
[33.126]	Cantilever/fiber deflection rig	LIGA micromotors	Ni/Alumina	Air	0.6–1.2

Fig. 33.33 Static friction force values of unlubricated motors and motors lubricated using PFPE lubricants, normalized over the rotor weight, as a function of rest time and relative humidity. Rest time is defined as the time elapsed between a given experiment and the first experiment in which motor movement was recorded (time 0). The motors were allowed to sit at a particular humidity for 12 h prior to measurements [33.112]

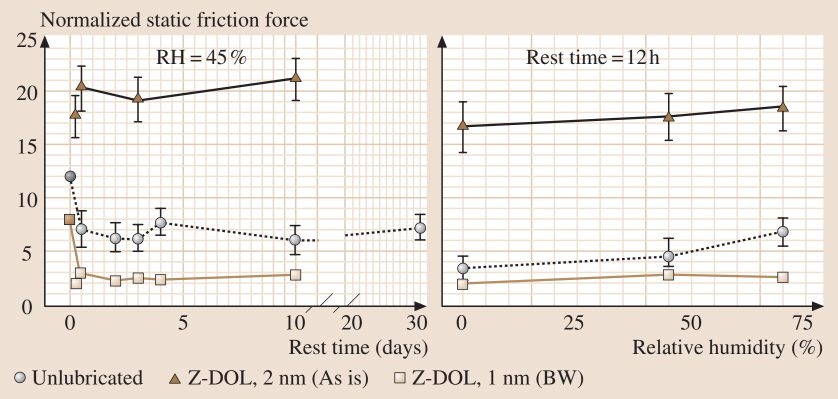

tor will begin to rotate, and the twist of the cantilever decreases sharply (event B in Fig. 33.32b). The twist of the cantilever is measured in the AFM as a change in the lateral deflection signal (in volts), which is the underlying concept of friction force microscopy (FFM). The change in the lateral deflection signal corresponding to the above mentioned events as the tip approaches the rotor is shown schematically in Fig. 33.32b. The value of the peak V_f is a measure of the force exerted on the rotor by the tip just before the static friction torque is matched and the rotor begins to rotate.

Using this technique, the viability of PFPE lubricants for micromotors has been investigated and the effect of humidity on the friction forces of unlubricated and lubricated devices was studied as well. Figure 33.33 shows static friction forces normalized over the weight of the rotor of unlubricated and lubricated micromotors as a function of rest time and relative humidity. Rest time here is defined as the time elapsed between the first experiment conducted on a given motor (solid symbol at time zero) and subsequent experiments (open symbols). Each open symbol data point is an average of six measurements. It can be seen that for the unlubricated motor and the motor lubricated with a bonded layer of Z-DOL (BW), the static friction force is highest for the first experiment and then drops to an almost constant level. In the case of the motor with an as-is mobile layer of Z-DOL, the values remain very high up to 10 days after lubrication. In all cases, there is negligible difference in the static friction force at 0% and 45% RH. At 70% RH, the unlubricated motor exhibits a substantial increase in the static friction force, while the motor with bonded Z-DOL shows no increase in static friction force due to the hydrophobicity of the lubricant layer. The motor with an as-is mobile layer of the lubricant shows consistently high values of static friction force that vary little with humidity.

33.6.4 Mechanisms Associated with Observed Stiction Phenomena in Micromotors

Figure 33.34 summarizes static friction force data for two motors, M1 and M2, along with schematics of the meniscus effects for the unlubricated and lubricated surfaces. Capillary condensation of water vapor from the environment results in the formation of meniscus bridges between contacting and near-contacting asperities of two surfaces in close proximity to each other, as shown in Fig. 33.34. For unlubricated surfaces, more menisci are formed at higher humidity, resulting in a higher friction force between the surfaces. The formation of meniscus bridges is supported by the fact that the static friction force for unlubricated motors increases at high humidity (Fig. 33.34). Solid bridging may occur near the rotor-hub interface due to silica residues after the first etching process. In addition, the drying process after the final etch can result in liquid bridging formed by the drying liquid due to meniscus force at these areas [33.54,55,113,116]. The initial static friction force therefore will be quite high, as evidenced by the solid data points in Fig. 33.34. Once the first movement of the rotor permanently breaks these solid and liquid bridges, the static friction force of the motors will drop (as seen in Fig. 33.34) to a value dictated predominantly by the adhesive energies of rotor and hub surfaces, the real area of contact between these surfaces and meniscus forces due to water vapor in the air, at which point, the effect of lubricant films can be observed. Lubrication with a mobile layer, even a thin one, results in very high static friction forces

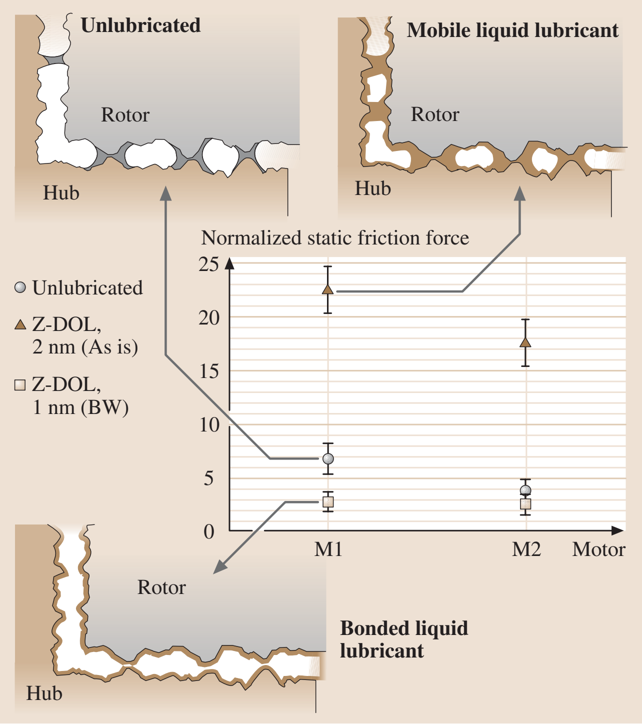

Fig. 33.34 Summary of effect of liquid and solid lubricants on static friction force of micromotors. Despite the hydrophobicity of the lubricant used (Z-DOL), a mobile liquid lubricant (Z-DOL as is) leads to very high static friction force due to increased meniscus forces, whereas a solid-like lubricant (bonded Z-DOL, BW) appears to provide some amount of reduction in static friction force

due to meniscus effects of the lubricant liquid itself at and near the contact regions. It should be noted that a motor submerged in a liquid lubricant would result in a fully flooded lubrication regime. In this case, there is no meniscus contribution and only the viscous contribution to the friction forces would be relevant. However, submerging the device in a lubricant may not be a practical method. A solid-like hydrophobic lubricant layer (such as bonded Z-DOL) results in favorable friction characteristics of the motor. The hydrophobic nature of the lubricant inhibits meniscus formation between the contact surfaces and maintains low friction even at high humidity (Fig. 33.34). This suggests that solid-like hydrophobic lubricants are ideal for lubrication of MEMS, while mobile lubricants result in increased values of static friction force.

References

33.1 NASA Ames Research Center, Mottett Field, CA, USA, http://www.ipt.arc.nasa.gov/gallery.html

33.2 W. G. van der Wiel, S. De Franceschi, J. M. Elzerman, T. Fujisawa, S. Tarucha, L. P. Kouwenhoven: Electron transport through double quantum dots, Rev. Mod. Phys. **75** (2003) 1–22

33.3 Texas Instruments DLP Products, Plano, TX, USA, http://www.dlp.com

33.4 Anonymous: *Microelectromechanical Systems: Advanced Materials and Fabrication Methods* (National Academy Press, Washington, D.C. 1997)

33.5 M. Roukes: Nanoelectromechanical systems face the future, Phys. World (February 2001) 25–31

33.6 M. A. Huff: A distributed MEMS processing environment, http://www.mems-exchange.org/ (2002)

33.7 R. S. Muller, R. T. Howe, S. D. Senturia, R. L. Smith, R. M. White: *Microsensors* (IEEE, New York 1990)

33.8 I. Fujimasa: *Micromachines: A New Era in Mechanical Engineering* (Oxford Univ. Press, Oxford 1996)

33.9 W. S. Trimmer (Ed.): *Micromachines and MEMS: Classic and Seminal Papers to 1990* (IEEE, New York 1997)

33.10 B. Bhushan (Ed.): *Tribology Issues and Opportunities in MEMS* (Kluwer, Dordrecht 1998)

33.11 G. T. A. Kovacs: *Micromachined Transducers Sourcebook* (WCB McGraw-Hill, Boston 1998)

33.12 S. D. Senturia: *Microsystem Design* (Kluwer, Boston 2001)

33.13 M. Gad-el-Hak: *The MEMS Handbook* (CRC, Boca Raton 2002)

33.14 T. R. Hsu: *MEMS and Microsystems* (McGraw-Hill, Boston 2002)

33.15 M. Madou: *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd edn. (CRC, Boca Raton 2002)

33.16 R. C. Jaeger: *Introduction to Microelectronic Fabrication*, 5th edn. (Addison-Wesley, Reading 1988)

33.17 J. W. Judy: Microelectromechanical systems (MEMS): Fabrication, design, and applications, Smart Mater. Struct. **10** (2001) 1115–1134

33.18 E. W. Becker, W. Ehrfeld, P. Hagmann, A. Maner, D. Munchmeyer: Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography, galvanoforming, and plastic moulding (LIGA process), Microelectron. Eng. **4** (1986) 35–56

33.19 C. R. Friedrich, R. O. Warrington: Surface characterization of non-lithographic micromachining. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 73–84

33.20 M. Madou: Facilitating choices of machining tools and materials for miniaturization science: A review. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 31–51

33.21 H. Lehr, S. Abel, J. Doppler, W. Ehrfeld, B. Hagemann, K. P. Kamper, F. Michel, Ch. Schulz, Ch. Thurigen: Microactuators as driving units for microrobotic systems, SPIE Proc. Microrobotics: Components and Applications **2906** (1996) 202–210

33.22 H. Lehr, W. Ehrfeld, B. Hagemann, K. P. Kamper, F. Michel, Ch. Schulz, Ch. Thurigen: Development of micro-millimotors, Min. Invas. Ther. Allied Technol. **6** (1997) 191–194

33.23 F. Michel, W. Ehrfeld: Microfabrication technologies for high performance microactuators. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 53–72

33.24 M. Tanaka: Development of desktop machining microfactory, Riken Rev. **34** (2001) 46–49

33.25 Y. Xia, G. M. Whitesides: Soft lithography, Angew. Chem. Int. Edn. **37** (1998) 550–575

33.26 H. Becker, C. Gaertner: Polymer microfabrication methods for microfluidic analytical applications, Electrophoresis **21** (2000) 12–26

33.27 Y. Xia, E. Kim, X. M. Zhao, J. A. Rogers, M. Prentiss, G. M. Whitesides: Complex optical surfaces formed by replica molding against elastomeric masters, Science **273** (1996) 347–349

33.28 S. Y. Chou, P. R. Krauss, P. J. Renstrom: Imprint lithography with 25-nanometer resolution, Science **272** (1996) 85–87

33.29 A. Kumar, G. M. Whitesides: Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol ink followed by chemical etching, Appl. Phys. Lett. **63** (1993) 2002–2004

33.30 T. A. Core, W. K. Tsang, S. J. Sherman: Fabrication technology for an integrated surface-micromachined sensor, Solid State Technol. **36** (1993) 39–47

33.31 J. Bryzek, K. Peterson, W. McCulley: Micromachines on the march, IEEE Spectrum (May 1994) 20–31

33.32 L. J. Hornbeck, W. E. Nelson: Bistable deformable mirror device. In: *OSA Technical Digest Series, Vol. 8, Spatial Light Modulators and Applications* (OSA, Washington 1988) pp. 107–110

33.33 L. J. Hornbeck: A digital light processing(tm) update – Status and future applications, Proc. SPIE **3634** (1999) 158–170

33.34 L. J. Hornbeck: The DMD^{TM} projection display chip: A MEMS-based technology, MRS Bull. **26** (2001) 325–328

33.35 B. Bhushan: *Tribology and Mechanics of Magnetic Storage Devices*, 2nd edn. (Springer, New York 1996)

33.36 H. Hamilton: Contact recording on perpendicular rigid media, J. Mag. Soc. Jpn. **15** (1991) (Suppl. S2) 483–481

33.37 T. Ohwe, Y. Mizoshita, S. Yonoeka: Development of integrated suspension system for a nanoslider with an MR head transducer, IEEE Trans. Magn. **29** (1993) 3924–3926

33.38 L. S. Fan, S. Woodman: Batch fabrication of mechanical platforms for high-density data storage, 8th Int. Conf. Solid State Sensors and Actuators (Transducers '95)/Eurosensors IX, Stockholm 1995, 434–437

33.39 D. A. Horsley, M. B. Cohn, A. Singh, R. Horowitz, A. P. Pisano: Design and fabrication of an angular microactuator for magnetic disk drives, J. Microelectromech. Syst. **7** (1998) 141–148

33.40 T. Hirano, L. S. Fan, D. Kercher, S. Pattanaik, T. S. Pan: HDD tracking microactuator and its integration issues, Proc. ASME **2** 2000, ed. by A. P. Lee, J. Simon, F. K. Foster, R. S. Keynton (ASME, New York 2000) 449–452

33.41 P. Gravesen, J. Branebjerg, O. S. Jensen: Microfluidics – A review, J. Micromech. Microeng. **3** (1993) 168–182

33.42 C. Lai Poh San, E. P. H. Yap (Eds.): *Frontiers in Human Genetics* (World Scientific, Singapore 2001)

33.43 C. H. Mastrangelo, H. Becker (Eds.): Microfluidics and BioMEMS, Proc. SPIE **4560** (2001)

33.44 H. Becker, L. E. Locascio: Polymer microfluidic devices, Talanta **56** (2002) 267–287

33.45 S. Shoji, M. Esashi: Microflow devices and systems, J. Micromech. Microeng. **4** (1994) 157–171

33.46 P. Woias: Micropumps – Summarizing the first two decades, *Proc. SPIE – Microfluidics and BioMEMS*, ed. by C. H. Mastrangelo, H. Becker, Proc. SPIE **4560** (2001) 39–52

33.47 M. Scott: MEMS and MOEMS for national security applications, *Reliability, Testing, and Characterization of MEMS/MOEMS II*, Proc. SPIE **4980** (2003) 37–44

33.48 K. E. Drexler: *Nanosystems: Molecular Machinery, Manufacturing and Computation* (Wiley, New York 1992)

33.49 B. Bhushan: *Handbook of Micro/Nanotribology*, 2nd edn. (CRC, Boca Raton 1999)

33.50 G. Timp (Ed.): *Nanotechnology* (Springer, New York 1999)

33.51 E. A. Rietman: *Molecular Engineering of Nanosystems* (Springer, New York 2001)

33.52 H. S. Nalwa (Ed.): *Nanostructured Materials and Nanotechnology* (Academic, San Diego 2002)

33.53 W.A. Goddard, D.W. Brenner, S.E. Lyshevski, G.J. Iafrate: *Handbook of Nanoscience, Engineering, and Technology* (CRC, Boca Raton 2003)

33.54 B. Bhushan: *Principles and Applications of Tribology* (Wiley, New York 1999)

33.55 B. Bhushan: *Introduction to Tribology* (Wiley, New York 2002)

33.56 Y.C. Tai, L.S. Fan, R.S. Muller: IC-processed micromotors: Design, technology and testing, Proc. IEEE Micro Electro Mechanical Systems 1989, 1–6

33.57 S.M. Spearing, K.S. Chen: Micro-gas turbine engine materials and structures, Ceramic Eng. Sci. Proc. **18** (2001) 11–18

33.58 M. Mehregany, K.J. Gabriel, W.S.N. Trimmer: Integrated fabrication of polysilicon mechanisms, IEEE Trans. Electron. Dev. **35** (1988) 719–723

33.59 E.J. Garcia, J.J. Sniegowski: Surface micromachined microengine, Sens. Actuators A **48** (1995) 203–214

33.60 J.K. Robertson, K.D. Wise: An electrostatically actuated integrated microflow controller, Sens. Actuators A **71** (1998) 98–106

33.61 B. Bhushan: Nanotribology and nanomechanics of MEMS devices, Proc. Ninth Annual Workshop on Micro Electro Mechanical Systems 1996 (IEEE, New York 1996) 91–98

33.62 D.M. Tanner, N.F. Smith, L.W. Irwin et al.: *MEMS Reliability: Infrastructure, Test Structures, Experiments, and Failure Modes* (Sandia National Laboratories, Albuquerque 2000) 2000–2091

33.63 S.S. Mani, J.G. Fleming, J.A. Walraven, J.J. Sniegowski et al.: Effect of W coating on microengine performance, Proc. 38th Annual Inter. Reliability Phys. Symp. 2000 (IEEE, New York 2000) 146–151

33.64 R.E. Sulouff: MEMS opportunities in accelerometers and gyros and the microtribology problems limiting commercialization. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 109–120

33.65 Analog Devices Inc., Berkeley, CA, USA, http://www.analog.com

33.66 S.A. Henck: Lubrication of digital micromirror devices, Tribol. Lett. **3** (1997) 239–247

33.67 M.R. Douglass: Lifetime estimates and unique failure mechanisms of the digital micromirror devices (DMD), Proc. 36th Annual Inter. Reliability Phys. Symp. 1998 (IEEE, New York 1998) 9–16

33.68 M.R. Douglass: DMD reliability: A MEMS success story, Reliability, Testing, and Characterization of MEMS/MOEMS II, Proc. SPIE **4980** (2003) 1–11

33.69 W.C. Tang, A.P. Lee: Defense applications of MEMS, MRS Bull. **26** (2001) 318–319. Also see www.darpa.mil/mto/mems

33.70 H. Guckel, D.W. Burns: Fabrication of micromechanical devices from polysilicon films with smooth surfaces, Sens. Actuators **20** (1989) 117–122

33.71 G.T. Mulhern, D.S. Soane, R.T. Howe: Supercritical carbon dioxide drying of microstructures, Proc. Int. Conf. on Solid-State Sensors and Actuators 1993 (IEEE, New York 1993) 269–299

33.72 P. Vettiger, J. Brugger, M. Despont, U. Drechsler, U. Duerig, W. Haeberle: Ultrahigh density, high data-rate NEMS based AFM data storage system, Microelectron. Eng. **46** (1999) 11–27, also see http://www.ibm.com

33.73 M. Ferrari, J. Liu: The engineered course of treatment, Mech. Eng. (December 2001) 44–47

33.74 F.J. Martin, C. Grove: Microfabricated drug delivery systems: Concepts to improve clinical benefits, Biomed. Microdev. **3** (2001) 97–108

33.75 K.F. Man, B.H. Stark, R. Ramesham: *A Resource Handbook for MEMS Reliability*, Rev. A (California Institute of Technology, Pasadena 1998)

33.76 S. Kayali, R. Lawton, B.H. Stark: MEMS reliability assurance activities at JPL, EEE Links **5** (1999) 10–13

33.77 S. Arney: Designing for MEMS reliability, MRS Bull. **26** (2001) 296–299

33.78 K.F. Man: MEMS reliability for space applications by elimination of potential failure modes through testing and analysis (2001), http://www.rel.jpl.nasa.gov/Org/5053/atop/products/Prod-map.html

33.79 B. Bhushan, A.V. Kulkarni, W. Bonin, J.T. Wyrobek: Nano/picoindentation measurement using a capacitance transducer system in atomic force microscopy, Philos. Mag. **74** (1996) 1117–1128

33.80 S. Sundararajan, B. Bhushan: Development of AFM-based techniques to measure mechanical properties of nanoscale structures, Sens. Actuators A **101** (2002) 338–351

33.81 B. Bhushan (Ed.): *Modern Tribology Handbook* (CRC, Boca Raton 2001)

33.82 B. Bhushan, J.N. Israelachvili, U. Landman: Nanotribology: Friction, wear and lubrication at the atomic scale, Nature **374** (1995) 607–616

33.83 J.S. Shor, D. Goldstein, A.D. Kurtz: Characterization of n-type β-SiC as a piezoresistor, IEEE Trans. Electron. Dev. **40** (1993) 1093–1099

33.84 M. Mehregany, C.A. Zorman, N. Rajan, C.H. Wu: Silicon carbide MEMS for harsh environments, Proc. IEEE **86** (1998) 1594–1610

33.85 C.A. Zorman, A.J. Fleischmann, A.S. Dewa, M. Mehregany, C. Jacob, S. Nishino, P. Pirouz: Epitaxial growth of 3C-SiC films on 4 in. diam Si(100) silicon wafers by atmospheric pressure chemical vapor deposition, J. Appl. Phys. **78** (1995) 5136–5138

33.86 C.A. Zorman, S. Roy, C.H. Wu, A.J. Fleischman, M. Mehregany: Characterization of polycrystalline silicon carbide films grown by atmospheric pressure chemical vapor deposition on polycrystalline silicon, J. Mater. Res. **13** (1998) 406–412

33.87 B. Bhushan, B.K. Gupta: *Handbook of Tribology: Materials, Coatings and Surface Treatments* (Krieger, Malabar 1997) Reprint

33.88 J. F. Shackelford, W. Alexander, J. S. Park (Eds.): *CRC Material Science and Engineering Handbook*, 2nd edn. (CRC, Boca Raton 1994)
33.89 B. K. Gupta, J. Chevallier, B. Bhushan: Tribology of ion bombarded silicon for micromechanical applications, ASME J. Tribol. **115** (1993) 392–399
33.90 B. K. Gupta, B. Bhushan, J. Chevallier: Modification of tribological properties of silicon by boron ion implantation, Tribol. Trans. **37** (1994) 601–607
33.91 B. K. Gupta, B. Bhushan: Nanoindentation studies of ion implanted silicon, Surf. Coat. Technol. **68-69** (1994) 564–570
33.92 B. Bhushan, V. N. Koinkar: Tribological studies of silicon for magnetic recording applications, J. Appl. Phys. **75** (1994) 5741–5746
33.93 G. M. Pharr: The anomalous behavior of silicon during nanoindentation. In: *Thin Films: Stresses and Mechanical Properties III*, Vol. 239, ed. by W. D. Nix, J. C. Bravman, E. Arzt, L. B. Freund (Materials Research Soc., Pittsburgh 1991) pp. 301–312
33.94 D. L. Callahan, J. C. Morris: The extent of phase transformation in silicon hardness indentation, J. Mater. Res. **7** (1992) 1612–1617
33.95 N. A. Fleck, G. M. Muller, M. F. Ashby, J. W. Hutchinson: Strain gradient plasticity: Theory and experiment, Acta Metall. Mater. **42** (1994) 475–487
33.96 B. Bhushan, S. Venkatesan: Friction and wear studies of silicon in sliding contact with thin-film magnetic rigid disks, J. Mater. Res. **8** (1993) 1611–1628
33.97 S. Venkatesan, B. Bhushan: The role of environment in the friction and wear of single-crystal silicon in sliding contact with thin-film magnetic rigid disks, Adv. Info Storage Syst. **5** (1993) 241–257
33.98 S. Venkatesan, B. Bhushan: The sliding friction and wear behavior of single-crystal, polycrystalline and oxidized silicon, Wear **171** (1994) 25–32
33.99 B. Bhushan: Chemical, mechanical, and tribological characterization of ultra-thin and hard amorphous carbon coatings as thin as 3.5 nm: Recent developments, Diam. Relat. Mater. **8** (1999) 1985–2015
33.100 S. Sundararajan, B. Bhushan: Micro/nanotribological studies of polysilicon and SiC films for MEMS applications, Wear **217** (1998) 251–261
33.101 X. Li, B. Bhushan: Micro/nanomechanical characterization of ceramic films for microdevices, Thin Solid Films **340** (1999) 210–217
33.102 A. A. Yasseen, C. H. Wu, C. A. Zorman, M. Mehregany: Fabrication and testing of surface micromachined polycrystalline SiC micromotors, IEEE Electron. Dev. Lett. **21** (2000) 164–166
33.103 H. Liu, B. Bhushan: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM, Ultramicroscopy **97** (2003) 321–340
33.104 B. Bhushan, A. V. Kulkarni, V. N. Koinkar, M. Boehm, L. Odoni, C. Martelet, M. Belin: Microtribological characterization of self-assembled and Langmuir–Blodgett monolayers by atomic force and friction force microscopy, Langmuir **11** (1995) 3189–3198
33.105 V. N. Koinkar, B. Bhushan: Micro/nanoscale studies of boundary layers of liquid lubricants for magnetic disks, J. Appl. Phys. **79** (1996) 8071–8075
33.106 V. N. Koinkar, B. Bhushan: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy, J. Vac. Sci. Technol. A **14** (1996) 2378–2391
33.107 B. Bhushan, H. Liu: Nanotribological properties and mechanisms of alkylthiol and biphenyl thiol self-assembled monolayers studied by AFM, Phys. Rev. B **63** (2001) 245412:1–11
33.108 H. Liu, B. Bhushan: Investigation of nanotribological properties of self-assembled monolayers with alkyl and biphenyl spacer chains, Ultramicroscopy **91** (2002) 185–202
33.109 T. Stifter, O. Marti, B. Bhushan: Theoretical investigation of the distance dependence of capillary and van der Waals forces in scanning force microscopy, Phys. Rev. B **62** (2000) 13667–13673
33.110 B. Bhushan: Self-assembled monolayers for controlling hydrophobicity and/or friction and wear. In: *Modern Tribology Handbook*, ed. by B. Bhushan (CRC, Boca Raton 2001) pp. 909–929
33.111 H. Liu, B. Bhushan, W. Eck, V. Stadler: Investigation of the adhesion, friction, and wear properties of biphenyl thiol self-assembled monolayers by atomic force microscopy, J. Vac. Sci. Technol. A **19** (2001) 1234–1240
33.112 S. Sundararajan, B. Bhushan: Static friction and surface roughness studies of surface micromachined electrostatic micromotors using an atomic force/friction force microscope, J. Vac. Sci. Technol. A **19** (2001) 1777–1785
33.113 C. H. Mastrangelo, C. H. Hsu: Mechanical stability and adhesion of microstructures under capillary forces – Part II: Experiments, J. Microelectromech. Syst. **2** (1993) 44–55
33.114 H. Liu, B. Bhushan: Adhesion and friction studies of microelectromechanical systems/nanoelectromechanical systems materials using a novel microtriboapparatus, J. Vac. Sci. Technol. A **21** (2003) 1528–1538
33.115 M. P. De Boer, T. A. Michalske: Accurate method for determining adhesion of cantilever beams, J. Appl. Phys. **86** (1999) 817
33.116 R. Maboudian, R. T. Howe: Critical review: Adhesion in surface micromechanical structures, J. Vac. Sci. Technol. B **15** (1997) 1–20
33.117 C. H. Mastrangelo: Surface force induced failures in microelectromechanical systems. In: *Tribology Issues and Opportunities in MEMS*, ed. by B. Bhushan (Kluwer, Dordrecht 1998) pp. 367–395
33.118 B. Bhushan, H. Liu, S. M. Hsu: Adhesion and friction studies of silicon and hydrophobic and low fric-

tion films and investigation of scale effects, ASME J. Tribol. in press

33.119 S. K. Chilamakuri, B. Bhushan: A comprehensive kinetic meniscus model for prediction of long-term static friction, J. Appl. Phys. **15** (1999) 4649–4656

33.120 J. Kijlstra, K. Reihs, A. Klamt: Roughness and topology of ultra-hydrophobic surfaces, Colloids Surf. A **206** (2002) 521–529

33.121 D. Quere, P. Aussillous: Non-stick droplets, Chem. Eng. Technol. **25** (2002) 925–928

33.122 Y. C. Tai, R. S. Muller: Frictional study of IC processed micromotors, Sens. Actuators A **21–23** (1990) 180–183

33.123 K. J. Gabriel, F. Behi, R. Mahadevan, M. Mehregany: In situ friction and wear measurement in integrated polysilicon mechanisms, Sens. Actuators A **21–23** (1990) 184–188

33.124 M. G. Lim, J. C. Chang, D. P. Schultz, R. T. Howe, R. M. White: Polysilicon microstructures to characterize static friction, Proc. IEEE Micro Electro Mechanical Systems (IEEE, New York 1990) 82–88

33.125 U. Beerschwinger, S. J. Yang, R. L. Reuben, M. R. Taghizadeh, U. Wallrabe: Friction measurements on LIGA-processed microstructures, J. Micromech. Microeng. **4** (1994) 14–24

33.126 D. Matheison, U. Beerschwinger, S. J. Young, R. L. Rueben, M. Taghizadeh, S. Eckert, U. Wallrabe: Effect of progressive wear on the friction characteristics of nickel LIGA processed rotors, Wear **192** (1996) 199–207

33.127 R. Maboudian: Adhesion and friction issues associated with reliable operation of MEMS, MRS Bull. (June 1998) 47–51

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

34. Mechanical Properties of Micromachined Structures

To be able to accurately design structures and make reliability predictions, in any field it is necessary first to know the mechanical properties of the materials that make up the structural components. In the fields of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), the devices are necessarily very small. The processing techniques and microstructures of the materials in these devices may differ significantly from bulk structures. Also, the surface-area-to-volume ratio in these structures is much higher than in bulk samples, and so the surface properties become much more important. In short, it cannot be assumed that mechanical properties measured using bulk specimens will apply to the same materials when used in MEMS and NEMS. This chapter will review the techniques that have been used to determine the mechanical properties of micromachined structures, especially residual stress, strength, and Young's modulus. The experimental measurements that have been performed will then be summarized, in particular the values obtained for polycrystalline silicon (polysilicon).

34.1 **Measuring Mechanical Properties of Films on Substrates** 1023
34.1.1 Residual Stress Measurements 1023
34.1.2 Mechanical Measurements Using Nanoindentation 1024

34.2 **Micromachined Structures for Measuring Mechanical Properties** 1024
34.2.1 Passive Structures 1025
34.2.2 Active Structures 1028

34.3 **Measurements of Mechanical Properties** . 1034
34.3.1 Mechanical Properties of Polysilicon 1034
34.3.2 Mechanical Properties of Other Materials 1036

References 1037

34.1 Measuring Mechanical Properties of Films on Substrates

In order to determine accurately the mechanical properties of very small structures, it is necessary to test specimens made from the same materials, processed in the same way, and of the same approximate size. Not surprisingly it is often difficult to handle specimens this small. One solution is to test the properties of films remaining on substrates. Micro- and nanomachined structures are typically fabricated from films that are initially deposited onto a substrate, are subsequently patterned and etched into the appropriate shapes, and then finally released from the substrate. If the testing is performed on the continuous film, before patterning and release, the substrate can be used as an effective "handle" for the specimen (in this case, the film). Of course since the films are adhered to the substrate, the types of tests possible are severely limited.

34.1.1 Residual Stress Measurements

One common measurement easily performed on films attached to substrates is residual film stress. The curvature of the substrate is measured before and after film deposition. Curvature can be measured in a number of ways. The most common technique is to scan a laser across the surface (or scan the substrate beneath the laser) and detect the angle of the reflected signal. Alternatively, profilometry, optical interferometry, or even atomic force microscopy can be used. As expected, tools that map a surface or perform multiple linear scans can give more accurate readings than tools that measure only a single scan.

Assuming that the film is thin compared to the substrate, the average residual stress in the film, σ_f, is given

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

by the Stoney equation,

$$\sigma_{\mathrm{f}} = \frac{1}{6}\frac{E_{\mathrm{s}}}{(1-\nu_{\mathrm{s}})}\frac{t_{\mathrm{s}}^2}{t_{\mathrm{f}}}\left(\frac{1}{R_1}-\frac{1}{R_2}\right) , \tag{34.1}$$

where the subscripts f and s refer to the film and substrate, respectively; t is thickness, E is Young's modulus, ν is Poisson's ratio, and R is radius of curvature before (R_1) and after (R_2) film deposition [34.2]. For the typical (100)-oriented silicon substrate, $E/(1-\nu)$ (also known as the biaxial modulus) is equal to 180.5 GPa, independent of in-plane rotation [34.3]. This investigation can be performed on the as-deposited film or after any subsequent annealing step, provided no changes occur to the substrate.

This measurement will reveal the average residual stress of the film. Typically, however, the residual stresses of deposited films will vary through the film thickness. One way to detect this, using substrate curvature techniques, is to etch away a fraction of the film and repeat the curvature measurement. This can be iterated any number of times to obtain a residual stress profile for the film [34.4]. Alternatively, tools have been designed to measure the substrate curvature during the deposition process itself, to obtain information on how the stresses evolve [34.5].

An additional feature of some of these tools is the ability to heat the substrates while performing the stress measurement. An example of the results obtained in such an experiment is shown in Fig. 34.1 [34.1], for an aluminum film on a silicon substrate. The slope of the heating curve gives the difference in thermal expansion between the film and the substrate. When the heating curve changes slope and becomes nearly horizontal, the yield strength of the film has been reached.

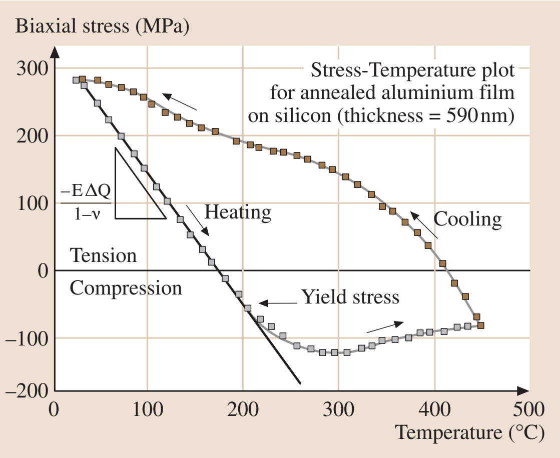

Fig. 34.1 Typical results for residual stress as a function of temperature for an aluminum film on a silicon substrate [34.1]. The stresses were determined by measuring the curvature of the substrate before and after film deposition, using the reflected signal of a laser scanned across the substrate surface

34.1.2 Mechanical Measurements Using Nanoindentation

Aside from residual stress, it is difficult to measure mechanical properties of films adhered to substrates without the measurement being affected by the presence of the substrate. Recent developments in nanoindentation equipment have allowed this technique to be used in some cases. With specially designed tools, indentation can be performed using very low loads. If the films being investigated are thick and rigid enough, measurements can be made that are not influenced by the presence of the substrate. Of course, this can be verified by depositing the same film onto different substrates. By continuously monitoring the displacement as well as the load during indentation, a variety of properties can be measured, including hardness and Young's modulus [34.6]. This area is covered in more detail in a separate chapter.

For brittle materials, cracks can be generated by indentation, and strength information can be gathered. But the exact stress fields created during the indentation process are not exactly known, and, therefore, quantitative values for strength are difficult to determine. Anisotropic etching of single crystal silicon has been done to create 30-μm-tall structures that were then indented to examine fracture toughness [34.7], however this is not possible with most materials.

34.2 Micromachined Structures for Measuring Mechanical Properties

Certainly the most direct way to measure the mechanical properties of small structures is to fabricate structures that would be conducive to such tests. Fabrication techniques are sufficiently advanced that virtually any design can be realized, at least in two dimensions. Two basic types of devices are used for mechanical property testing: "passive" structures and "active" structures.

34.2.1 Passive Structures

As mentioned previously, the main difficulty in testing very small specimens is in handling. One way to circumvent this problem is by using passive structures. These structures are designed to act as soon as they are released from the substrate and to provide whatever information they are designed to supply without further manipulation. For all of these passive structures, the forces acting on them come from the structural materials' own internal residual stresses. For devices on the micron scale or smaller, gravitational forces can be neglected, and therefore internal stresses are the only source of actuation force.

Stress Measurements

Since the internal residual stresses act upon the passive devices when they are released, it is natural to design a device that can be used to measure residual stresses. One such device, a rotating microstrain gauge, is shown in Fig. 34.2. There are many different microstrain gauge designs, but all operate under the same principle. In Fig. 34.2a, the large pads, labeled A, will remain anchored to the substrate when the rest of the device is released. Upon release, the device will expand or contract in order to relieve its internal residual stresses. A structure under tension will contract, and a structure under compression will expand. For the structure in Fig. 34.2, a compressive stress will cause the legs to lengthen. Since the two opposing legs are not attached to the central beam at the same point but are offset, they will cause the central beam to rotate when they expand. The device in Fig. 34.2 contains two independent gauges that point to one another. At the ends of the two central beams are two parts of a Vernier scale. By observing this scale, one can measure the rotation of the beams.

If the connections between the legs and the central beams were simple pin connections, the strain, ε, of the legs (the fraction of expansion or contraction) could be determined simply by the measured rotation and the geometry of the device, namely

$$\varepsilon = \frac{d_{\text{beam}} d_{\text{offset}}}{2 L_{\text{central}} L_{\text{leg}}} , \qquad (34.2)$$

where d_{beam} is the lateral deflection of the end of one central beam, d_{offset} is the distance between the connections of the opposing legs, L_{central} is the length of the central beam (measured to the center point between the leg connections), and L_{leg} is the length of the leg. But since the entire device was fabricated from a single polysilicon film, this cannot be the case; there must be some bending occurring at the connections. As a result, to get an accurate determination of the strain relieved upon release, finite element analysis (FEA) of the struc-

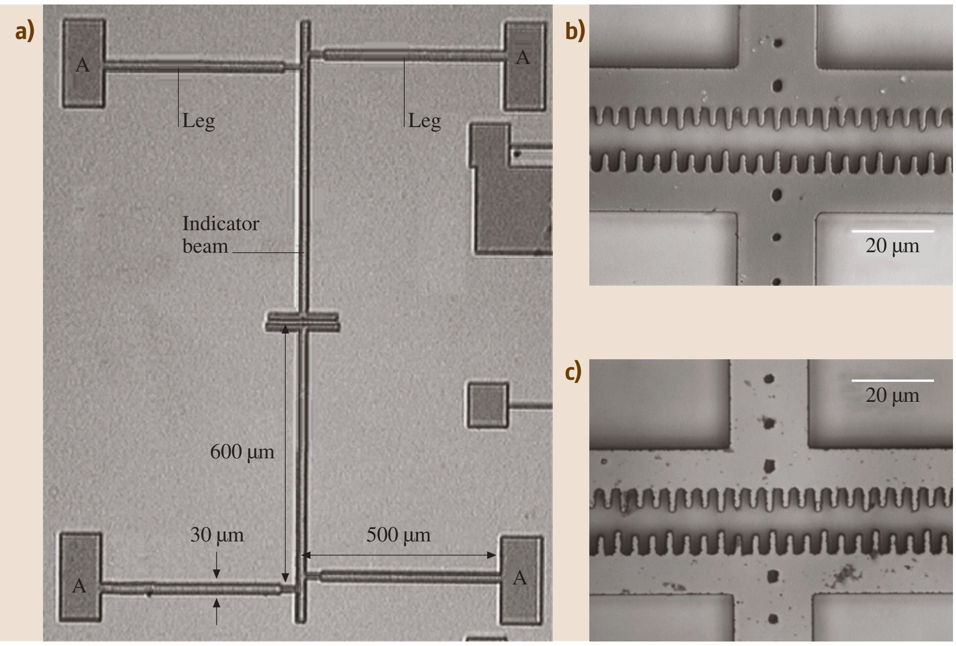

Fig. 34.2 **(a)** Microstrain gauge fabricated from polysilicon. **(b)** shows a close-up of the Vernier scale before release, and **(c)** shows the same area after release

ture must be performed. This is a common situation for microdevices. FEA is a powerful tool in determining the displacements and stresses of nonideal geometries. One drawback is that Young's modulus of the material must be known in order to do the FEA as well as to convert the measured strain into a stress value. But Young's moduli for many micromachined materials are known or can be measured using other techniques.

Other devices, besides rotating strain gauges, are designed to measure residual stresses. One of the most simple is a doubly clamped beam, a long, narrow beam of constant width and thickness that is anchored to the substrate at both ends. If the beam contains a tensile stress, it will remain straight. But if the beam contains a compressive stress, it will buckle if its length exceeds a critical value, l_{cr}, according to the Euler buckling criterion [34.8],

$$\varepsilon_{\mathrm{r}} = -\frac{\pi^2}{3}\left(\frac{h}{l_{\mathrm{cr}}}\right)^2 , \qquad (34.3)$$

where ε_{r} is the residual strain in the beam, and h is the width or thickness of the beam, whichever is less. To determine the residual strain, a series of doubly clamped beams are fabricated of varying lengths. In this way, the critical length, l_{cr}, for buckling can be deduced after release. One problem with this technique is that during the release process, any turbulence in the solution will lead to enhanced buckling of beams, and a low value for l_{cr} will be obtained.

For films with tensile stresses, a similar analysis can be done using ring-and-beam structures, also called Guckel rings after their inventor, Henry Guckel. A schematic of this design is shown in Fig. 34.3 [34.8]. A tensile stress in the outer ring will cause it to contract. This will lead to a compressive stress in the central beam, even though the material was originally tensile before release. The amount of compression in the central beam can be determined analytically from the geometry of the device and the residual strain of the material. Again, by changing the length of the central beam, the l_{cr} will be determined, and then the residual strain can be deduced.

Stress Gradient Measurements

For structures fabricated from thin deposited films, the stress gradient can be just as important as the stress itself. Figure 34.4 shows a portion of a silicon microactuator. The device is designed to be completely planar; however, stress gradients in the film cause the structures to bend. This figure illustrates the importance of characterizing and controlling stress gradients, and it also demonstrates that the easiest structure for measuring stress gradients is a simple cantilever beam. By measuring the end deflection δ, of a cantilever beam of length l, and thickness t, the stress gradient, $\mathrm{d}\sigma/\mathrm{d}t$ is determined by [34.9]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}t} = \frac{2\delta}{l^2}\frac{E}{1-\nu} . \qquad (34.4)$$

The magnitude of the end deflection can be measured by microscopy, optical interferometry, or any other technique.

Another useful structure for measuring stress gradients is a spiral. For this structure, the end of the spiral not anchored to the substrate will move out-of-plane. Also

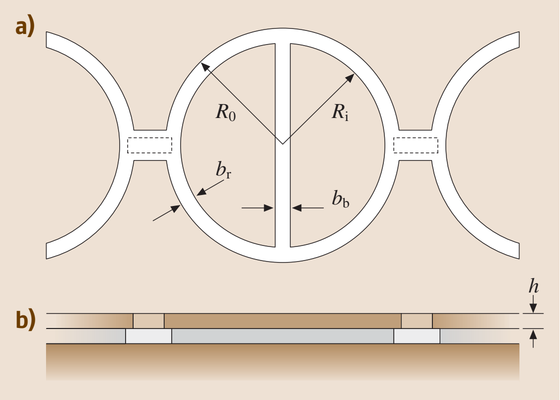

Fig. 34.3a,b Schematic **(a)** top view and **(b)** side view of Guckel ring structures [34.8]. The *dashed lines* in **(a)** indicate the anchors

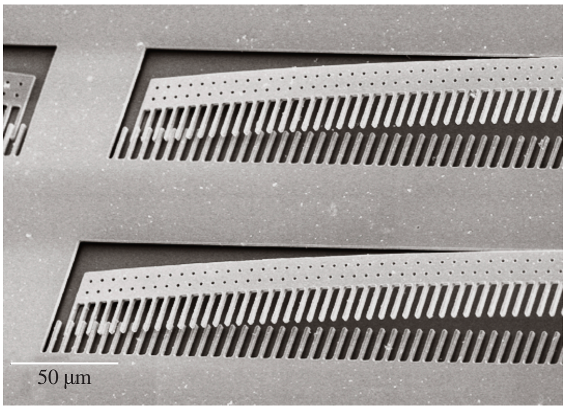

Fig. 34.4 Scanning electron micrograph (SEM) of a portion of a silicon microactuator. Residual stress gradients in the silicon cause the structure to bend

the diameter of the spiral will contract, and the free end of the spiral will rotate when released [34.10].

Strength and Fracture Toughness Measurements

As mentioned above, if a doubly clamped beam contains a tensile stress, it will remain taut when released because it cannot relieve any of its stress by contracting. This tensile stress can be thought of as a tensile load being applied at the ends of the beam. If this tensile load exceeds the tensile strength of the material, the beam will break. Since the tensile stress can be measured, as discussed in the Sect. 34.2.1, this technique can be used to gather information on the strength of materials. Figure 34.5 shows two different beam designs that have been used to measure strength. The device shown in Fig. 34.5a was fabricated from a tensile polysilicon film [34.11]. Different beams were designed with varying lengths of the wider regions (marked l_1 in the figure). In this manner, the load applied to the narrow center beam was varied, even though the entire film contained a uniform residual tensile stress. For l_1 greater than a critical value, the narrow center beam fractured, giving a measurement for the tensile strength of polysilicon.

The design shown in Fig. 34.5b was fabricated from a tensile $\mathrm{Si}_x\mathrm{N}_y$ film [34.12]. As seen in the figure, a stress concentration was designed into the beam, to ensure the fracture strength would be exceeded. In this case, a notch was etched into one side of the beam. Since the stress concentration is not symmetric with regard to the beam axis, this results in a large bending moment at that position, and the test measures the bend strength of the material. Again, like the beams shown in Fig. 34.5a, the geometry of various beams fabricated from the same film were varied, to vary the maximum stress seen at the notch. By seeing which beams fracture at the stress concentration after release, the strength can be determined.

Through a similar technique but by using an atomically sharp pre-crack instead of a stress concentration, the fracture toughness of a material can be determined. Sharp pre-cracks can be introduced into micromachined structures before release by placing a Vickers indent on the substrate, near the device; the radial crack formed by the indent will propagate into the overlying structure [34.14]. Accordingly, the beam with a sharp pre-crack, shown in Fig. 34.6, was fabricated using polysilicon [34.13]. Due to the stochastic nature of indentation, the initial pre-crack length will vary from beam to beam. Because of this, even though the geometry of the beams remains identical, the stress intensity, K, at the pre-crack tip will vary. Upon release, only those pre-cracks whose K exceeds the fracture toughness of the material, K_{Ic}, will propagate, and in this way upper and lower bounds for K_{Ic} for the material can be determined.

For all of the beams discussed in this section, finite element analysis is required to determine the stress concentrations and stress intensities. Even though approximate analytical solutions may exist for these designs, the actual fabricated structure will not have idealized geometries. For example, corners will never be perfectly sharp, and cracks will never be perfectly straight. This

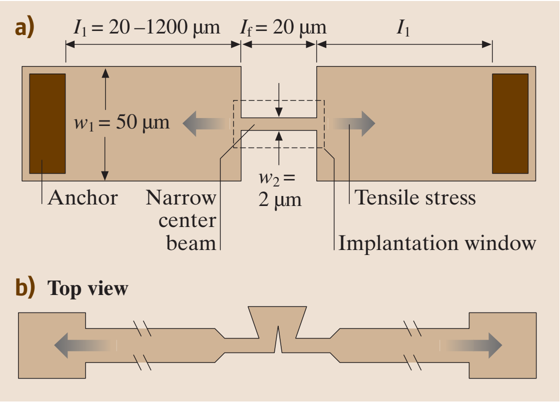

Fig. 34.5a,b Schematic designs of doubly clamped beams with stress concentrations for measuring strength. **(a)** was fabricated from polysilicon [34.11], and **(b)** was fabricated from $\mathrm{Si}_x\mathrm{N}_y$ [34.12]

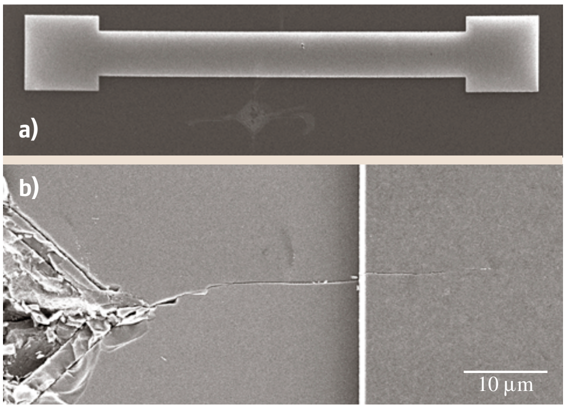

Fig. 34.6 **(a)** SEM of a 500 μm long polysilicon beam with a Vickers indent placed near its center; **(b)** higher magnification SEM of the area near the indent showing the pre-crack traveling from the substrate into the beam [34.13]

reinforces the idea that FEA is a powerful tool in determining mechanical properties of very small structures.

34.2.2 Active Structures

As discussed above, it is very convenient to design structures that act upon release to provide information on the mechanical properties of the structural materials. This is not always possible, however. For example, those passive devices just discussed rely on residual stresses to create the changes (rotation or fracture) that occur upon release, but many materials do not contain high residual stresses as deposited, or the processing scheme of the device precludes the generation of residual stresses. Also, some mechanical properties, such as fatigue resistance, require motion in order to be studied. Active devices are therefore used. These are acted upon by a force (the source of this force can be integrated into the device itself or can be external to the device) in order to create a change, and by the response to the force, the mechanical properties are studied.

Young's Modulus Measurements

Young's modulus, E, is a material property critical to any structural device design. It describes the elastic response of a material and relates stress, σ, and strain, ε, by

$$\sigma = E\varepsilon \,. \tag{34.5}$$

In bulk samples, E is often measured by loading a specimen in tension and measuring displacement as a function of stress for a given length. While this is considerably more difficult for small structures, such as those fabricated from thin deposited films, it can be achieved with careful experimental techniques. Figure 34.7 shows the schematic of one such measurement system [34.15]. The fringe detectors in the figure detect the reflected laser signal from two gold lines deposited onto the polysilicon specimen, which act as gage markers. In this manner, the strain in the specimen during loading can be monitored. Besides gold lines, Vickers indents placed in a nickel specimen can also serve as gage markers [34.16], or a speckle interferometry technique [34.17] can be used to determine strain in the specimen. Once the stress-versus-strain behavior is measured, the slope of the curve is equal to E.

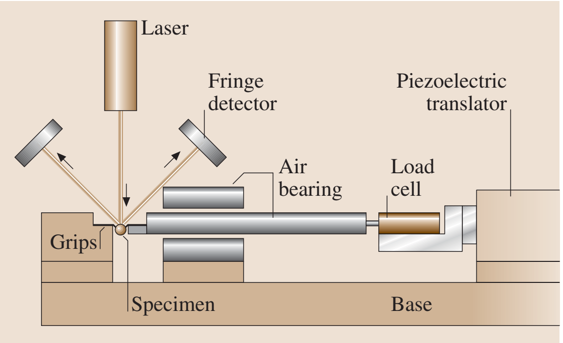

Fig. 34.7 Schematic of a measurement system for tensile loading of micromachined specimens [34.15]

In addition to the tensile test, Young's modulus can be determined by other measures of stress-strain behavior. As seen in Fig. 34.8, a cantilever beam can be bent by pushing on the free end with a nanoindenter [34.18]. The nanoindenter can monitor the force applied and the displacement, and simple beam theory can convert the displacement into strain to obtain E. A similar technique, shown schematically in Fig. 34.9 [34.19], involves pulling downward on a cantilever beam by means of an electrostatic force. An electrode is fabricated into the substrate beneath the cantilever beam, and a voltage is applied between the beam and the bottom electrode. The force acting on the beam is equal to the electrostatic force corrected to include the effects of fringing fields acting on the sides of the beam, namely

$$F(x) = \frac{\varepsilon_0}{2}\left(\frac{V}{g+z(x)}\right)^2\left(1+\frac{0.65[g+z(x)]}{w}\right) , \tag{34.6}$$

where $F(x)$ is the electrostatic force at x, ε_0 is the dielectric constant of air, g is the gap between the beam and the bottom electrode, $z(x)$ is the out-of-plane deflection of the beam, w is the beam width, and V is the applied voltage [34.19]. In this work, the deflection of the beam as a function of position is measured using optical interferometry. These measurements combine to give stress-strain behavior for the cantilever beam. An extension of this technique uses doubly clamped beams instead of cantilever beams. In this case, the deflection of the beam at a given electrostatic force depends on the residual stress in the material as well as Young's modulus. This method can also be used, therefore, to measure residual stresses in doubly clamped beams.

Another device possible to fabricate from a thin film and also possible to use for investigating stress-strain behavior is a suspended membrane, seen in Fig. 34.10 [34.20]. As shown in the schematic figure, the membrane is exposed to an elevated pressure on one side, causing it to bulge in the opposite direction. The deflection of the membrane is measured by optical or other techniques and related to the strain in the mem-

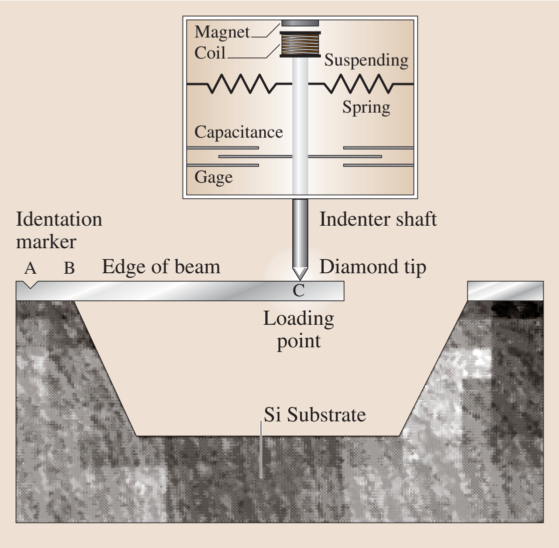

Fig. 34.8 Schematic of a nanoindenter loading mechanism pushing on the end of a cantilever beam [34.18]

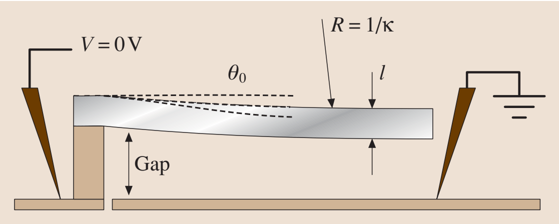

Fig. 34.9 Schematic of a cantilever beam bending test using an electrostatic voltage to pull the beam toward the substrate [34.19]

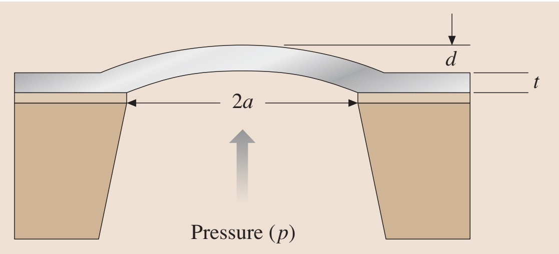

Fig. 34.10 Schematic cross section of a microfabricated membrane [34.20]

brane. These membranes can be fabricated in any shape, typically square or circular. Both analytical solutions and finite element analyses have been done to relate the deflection to the strain. Like the doubly clamped beams, both Young's modulus and residual stress play a role in the deflected shape. Both of these mechanical properties can therefore be determined by the pressure-versus-deflection performance of the membrane.

Another measurement besides stress-strain behavior that can reveal Young's modulus of a material is the determination of the natural resonance frequency. For a cantilever, the resonance frequency, f_r, for free undamped vibration is given by

$$f_\mathrm{r} = \frac{\lambda_i^2 t}{4\pi l^2}\left(\frac{E}{3\rho}\right)^{1/2} , \qquad (34.7)$$

where ρ, l, and t are density, length, and thickness of the cantilever; λ_i is the eigenvalue, where i is an integer that describes the resonance mode number; for the first mode $\lambda_1 = 1.875$ [34.22]. By measuring f_r and knowing the geometry and density, therefore, E can be determined. The cantilever can be vibrated by a number of techniques, including a laser, loudspeaker, or piezoelectric shaker. The frequency that produces the highest amplitude of vibration is the resonance frequency.

A micromachined device that uses an electrostatic comb drive and an AC signal to generate the vibration of the structure is known as a lateral resonator [34.23]. One example is shown in Fig. 34.11 [34.21]. When a voltage is applied across either set of interdigitated comb fingers shown in Fig. 34.11, an electrostatic attraction will be generated due to the increase in capacitance as the overlap between the comb fingers increases. The force,

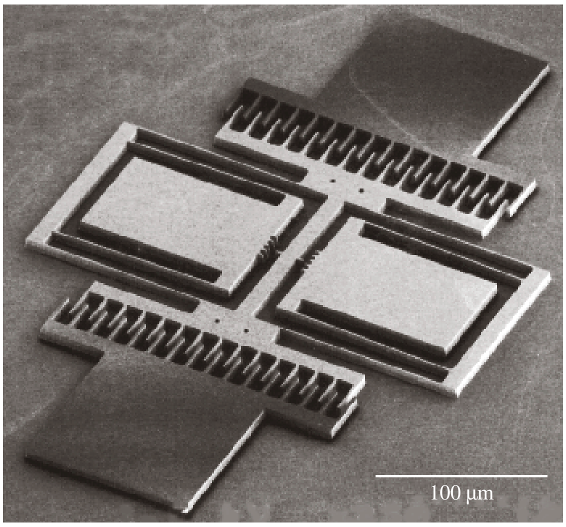

Fig. 34.11 SEM of a polysilicon lateral resonator [34.21]

F, generated by the comb drive is given by

$$F = \frac{1}{2}\frac{\partial C}{\partial x}V^2 = n\varepsilon\frac{h}{g}V^2 , \quad (34.8)$$

where C is capacitance, x is distance traveled by one comb-drive toward the other, n is the number of pairs of comb fingers in one drive, ε is the permittivity of the fluid between the fingers, h is the height of the fingers, g is the gap spacing between the fingers, and V is the applied voltage [34.23]. When an AC voltage at the resonance frequency is applied across either of the two comb drives, the central portion of the device will vibrate. In fact, since force depends on the square of the voltage for electrostatic actuation, for a time, t, dependent drive voltage, $v_{\mathrm{D}}(t)$, of

$$v_{\mathrm{D}}(t) = V_{\mathrm{P}} + v_{\mathrm{d}}\sin(\omega t) , \quad (34.9)$$

where V_{P} is the DC bias and v_{d} is the AC drive amplitude, the time dependent portion of the force will scale with

$$2\omega V_{\mathrm{P}} v_{\mathrm{d}}\cos(\omega t) + \omega v_{\mathrm{d}}^2\sin(2\omega t) \quad (34.10)$$

[34.23]. Therefore, if an AC drive signal is used with no DC bias, at resonance, the frequency of the AC drive signal will be one half the resonance frequency. For this device, the resonance frequency, f_{r}, will be

$$f_{\mathrm{r}} = \frac{1}{2\pi}\left(\frac{k_{\mathrm{sys}}}{M}\right)^{1/2} , \quad (34.11)$$

where k_{sys} is the spring constant of the support beams and M is the mass of the portion of the device that vibrates. The spring constant is given by

$$k_{\mathrm{sys}} = 24EI/L^3 , \quad (34.12)$$

$$I = \frac{hw^3}{12} , \quad (34.13)$$

where I is the moment of inertia of the beams, and L, h, and w are the length, thickness, and width of the beams. By combining these equations and measuring f_{r}, therefore, E can be determined.

One distinct advantage of the lateral resonator technique and the electrostatically pulled cantilever technique for measuring Young's modulus is that they require no external loading sources. Portions of the devices are electrically contacted, and a voltage is applied. For the pure tension tests, such as shown in Fig. 34.7, the specimen must be attached to a loading system, which for the very small specimens discussed here, can be extremely difficult, and any misalignment or eccentricity in the test could lead to unreliable results. But the advantage of the externally loaded technique is that there are no limitations on the type of materials that can be tested. Conductivity is not a requirement, nor is any compatibility with electrical actuation.

Strength and Fracture Toughness Measurements

As one might expect, any of the techniques discussed in the previous section that strain specimens in order to measure Young's modulus can also be used to measure fracture strength. Simply, the load is increased until the specimen breaks. As long as either the load or the strain is measured at fracture, and the geometry of the specimen is known, the maximum stress required for fracture, σ_{crit}, can be determined, either through analytical analysis or FEA. Depending on the geometry of the test, σ_{crit} will represent the tensile or bend strength of the material.

If the available force is limited, or if it is desired to have a localized fracture site, stress concentrations can be added to the specimens. These are typically notches micromachined into the edges of specimens.

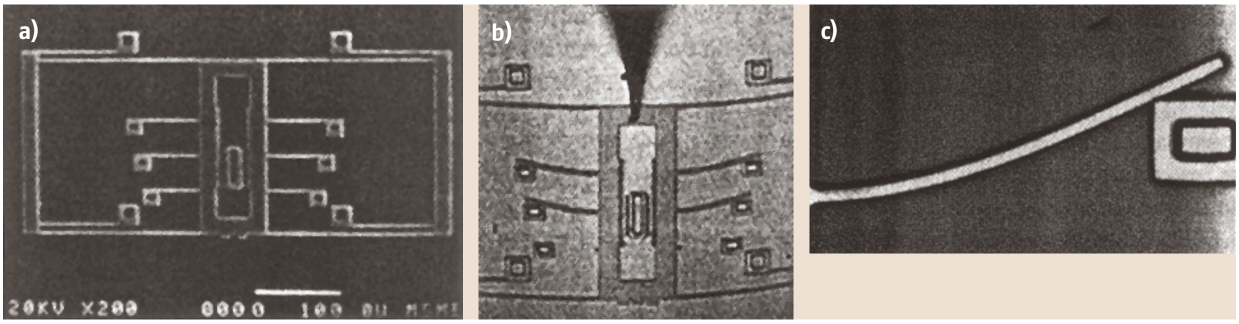

Fig. 34.12 **(a)** SEM of a device for measuring bend strength of polysilicon beams; **(b)** image of a test in process; **(c)** higher magnification view of one beam shortly before breaking [34.24]

Focused ion beams have also been used to carve stress concentrations into fracture specimens.

All of the external loading schemes, such as those shown in Figs. 34.7 and 34.8, have been used to measure fracture strength. Also, the electrostatically loaded doubly clamped beams can be pulled until they fracture. In this case, there is one complication. The electrostatic force is inversely proportional to the distance between the electrodes, and at a certain voltage, called the "pull-in voltage," the attraction between the beam and the substrate will become so great that the beam will immediately be pulled into contact with the bottom electrode. As long as the fracture takes place before the pull-in voltage is reached, the experiment will give valid results.

Other loading techniques have been used to generate fracture of microspecimens. Figure 34.12 [34.24] shows one device designed to be pushed by the end of a micromanipulated needle. The long beams that extend from the sides of the central shuttle come into contact with anchored posts, and, at a critical degree of bending, the beams will break off. Since the applied force cannot be measured in this technique, the experiment is optically monitored continuously during the test, and the image of the beams just before fracture is analyzed to determine σ_{crit}.

Another loading scheme that has been demonstrated for micromachined specimens utilizes scratch drive actuators to load the specimens [34.25]. These types of actuators work like inchworms, traveling across a substrate in discrete advances as an electrostatic force is repeatedly applied between the actuator and the substrate. The stepping motion can be made in the nm scale, depending on the frequency of the applied voltage, and so it can be an acceptable approximation to continuous loading. One advantage of this scheme is that very large forces can be generated by relatively small devices. The exact forces generated cannot be measured, so like the technique that used micromanipulated pushing, the test is continuously observed to determine the strain at fracture. Another advantage of this technique is that, like the lateral resonator and the electrically pulled cantilever, the loading takes place on-chip, and, therefore, the difficulties associated with attaching and aligning an external loading source are eliminated.

Another on-chip actuator used to load microspecimens is shown in Fig. 34.13, along with three different microspecimens [34.14]. Devices have been fabricated with each of the three microspecimens integrated with the same electrostatic comb-drive actuator. In all three cases, when a DC voltage is applied to the actuator, it moves downward, as oriented in Fig. 34.13. This pulls down on the left end of each of the three microspecimens, which are anchored on the right. The actuator contains 1,486 pairs of comb fingers. The maximum voltage that can be applied is limited by the breakdown

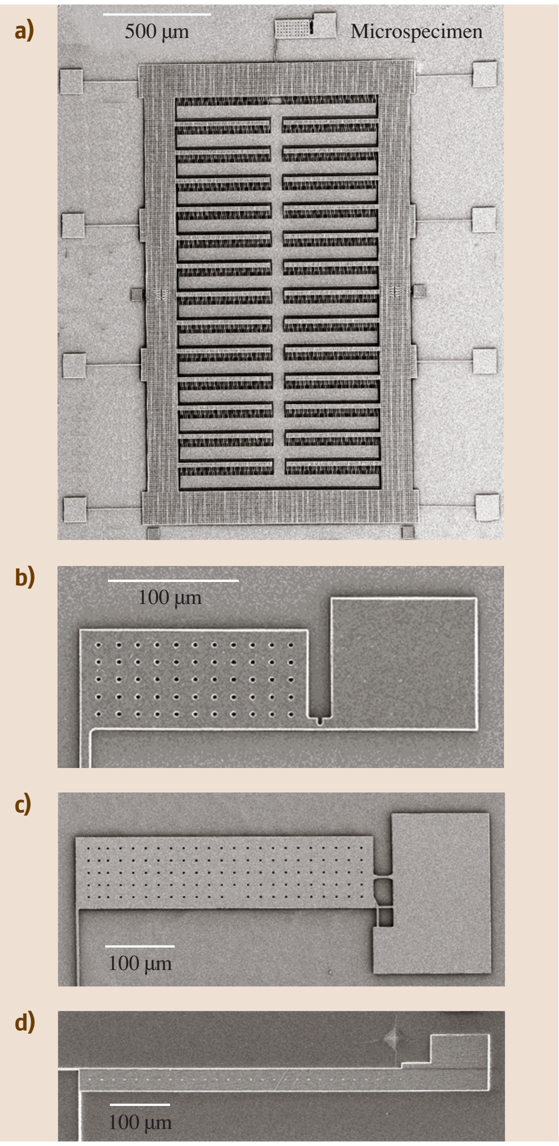

Fig. 34.13 **(a)** SEM of a micromachined device for conducting strength tests; the device consists of a large comb drive electrostatic actuator integrated with a microspecimen; **(b)–(d)** SEMs of various microspecimens for testing bend strength, tensile strength, and fracture toughness, respectively [34.14]

voltage of the medium in which the test takes place. In air, that limits the voltage to less than 200 V. As a result, given a finger height of 4 µm and a gap of 2 µm, and using (34.8), the maximum force generated by this actuator is limited to about 1 mN. Standard optical photolithography has a minimum feature size of about 2 µm. As a result, the electrostatic actuator cannot generate sufficient force to perform a standard tensile test on MEMS structural materials such as polysilicon. The microspecimens shown in Fig. 34.13 are therefore designed such that the stress is amplified.

The specimen shown in Fig. 34.13b is designed to measure bend strength. It contains a micromachined notch with a root radius of 1 µm. When the actuator pulls downward on the left end of this specimen, the notch serve as a stress concentration, and when the stress at the notch root exceeds σ_{crit}, the specimen fractures. The specimen in Fig. 34.13c is designed to test tensile strength. When the left end of this specimen is pulled downward, a tensile stress is generated in the upper thin horizontal beam near the right end of the specimen. As the actuator continues to move downward, the tensile stress in this beam will exceed the tensile strength, causing fracture. Finally, the specimen in Fig. 34.13d is similar to that in Fig. 34.13b, except that the notch is replaced by a sharp pre-crack that was produced by the Vickers indent placed on the substrate near the specimen. When this specimen is loaded, a stress intensity, K, is generated at the crack tip. When the stress intensity exceeds a critical value, K_{Ic}, the crack propagates. K_{Ic} is also referred to as the fracture toughness.

The force generated by the electrostatic actuator can be calculated using (34.8). But (34.8) assumes a perfectly planar, two-dimensional device. In fact, when actuated, the electric fields extend out of the plane of the device, and so (34.8) is just an approximation. Instead, like many of the techniques discussed in this section, the test is continuously monitored, and the actuator displacement at the time of fracture is recorded. Then FEA is used to determine the magnitude of the stress or stress intensity seen by the specimen at the point of fracture.

Fatigue Measurements

A benefit of the electrostatic actuator shown in Fig. 34.13 is that, besides monotonic loading, it can generate cyclic loading. This allows the fatigue resistance of materials to be studied. Simply by using an AC signal instead of a DC voltage, the device can be driven at its resonant frequency. The amplitude of the resonance depends on the magnitude of the AC signal. This amplitude can be increased until the specimen breaks; this will investigate the low-cycle fatigue resistance. Otherwise the amplitude of resonance can be left constant at a level below that is required for fast fracture, and the device will resonate indefinitely until the specimen breaks; this will investigate high-cycle fatigue. It should be noted that the resonance frequency of such a device is about 10 kHz. Therefore, it is possible to stress a specimen over 10^9 cycles in less than a day. In addition to simple cyclic loading, if a DC bias is added to the AC signal, a mean stress can be superimposed on the cyclic load. In this way, nonsymmetric cyclic loading (either with a large tensile stress alternating with a small compressive stress, or vice versa) can be studied.

Another device that can be used to investigate fatigue resistance in MEMS materials is shown in Fig. 34.14 [34.26]. In this case, a large mass is attached to the end of a notched cantilever beam. The mass contains two comb drives on opposite ends. When an AC signal is applied to one comb drive, the device will res-

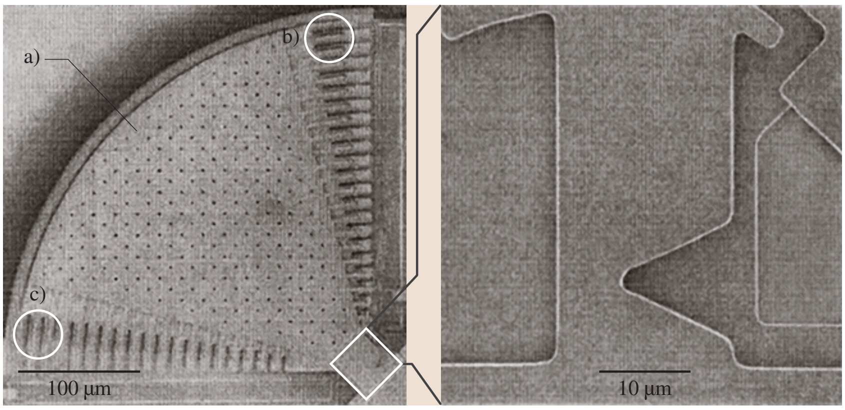

Fig. 34.14 SEM of a device for investigating fatigue; the image on the right is a higher magnification view of the notch near the base of the moving part of the structure [34.26] (a) mass, b) comb drive actuator, c) capacitive displacement sensor)

onate, cyclically loading the notch. The comb drive on the opposite side is used as a capacitive displacement sensor. This device contains many fewer comb fingers than the device shown in Fig. 34.13. As a result, it can apply cyclic loads by exploiting the resonance frequency of the device but cannot supply sufficient force to achieve monotonic loading.

Fatigue loading has also been studied using the same external loading techniques shown in Fig. 34.7. In this case, the frequency of the cyclic load is considerably lower, since the resonance frequency of the device is not being utilized. This leads to longer high-cycle testing times. Since the force is essentially unlimited, however, this technique allows a variety of frequencies to be studied to determine their effect on the fatigue behavior.

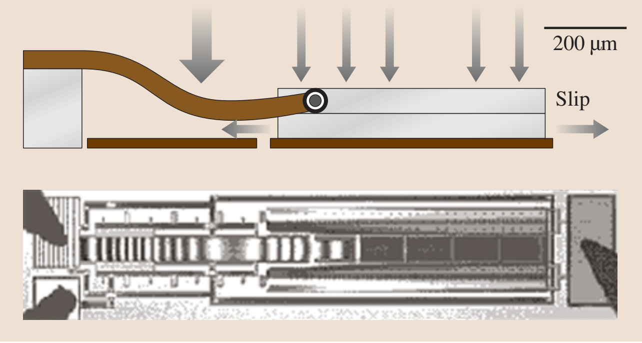

Fig. 34.16 Schematic cross section and top-view optical micrograph of a hinged-cantilever test structure for measuring friction in micromachined devices [34.28]

Friction Measurements

Friction is another property that has been studied in micromachined structures. To study friction, of course, two surfaces must be brought into contact with each other. This is usually avoided at all costs for these devices because of the risk of stiction. (Stiction is the term used when two surfaces that come into contact adhere so strongly that they cannot be separated.) Even so, a few devices have been designed to investigate friction. One of these is shown schematically in Fig. 34.15 [34.27]. It consists of a movable structure with a comb drive on one end and a cantilever beam on the other. Beneath the cantilever, on the substrate, is planar electrode. The device is moved to one side using the comb drive. Then a voltage is applied between the cantilever beam and the substrate electrode. The voltage on the comb drive is then released. The device would normally return to its original position, to relax the deflection in the truss suspensions, but the friction between the cantilever and the substrate electrode holds it in place. The voltage to the substrate electrode is slowly decreased until the device starts to slide. Knowing the electrostatic force generated by the

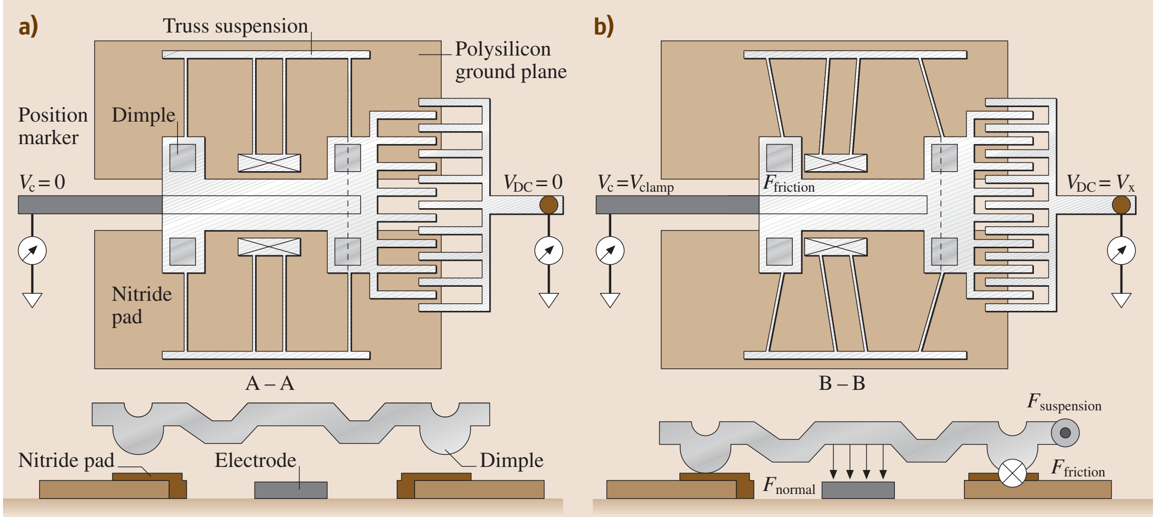

Fig. 34.15a,b Schematic drawings of a device used to study friction. **(a)** shows top and side views of the device in its original position, and **(b)** shows views of the device after it has been displaced using the comb drive and clamped using the substrate electrode [34.27]

substrate electrode and the stiffness of the truss suspensions, the static friction can be determined. For this device, bumps were fabricated on the bottoms of the cantilever beams. This limited the surface area that came into contact with the substrate and so lowered the risk of stiction.

Another device designed to study friction is shown in Fig. 34.16 [34.28]. This technique uses a hinged cantilever. The portion near the free end acts as the friction test structure, and the portion near the anchored end acts as the driver. The friction test structure is attracted to the substrate by means of electrostatic actuation, and when a second electrostatic actuator pulls down the driver, the friction test structure slips forward by a length proportional to the forces involved, including the frictional force. This distance, however, has a maximum of 30 nm, so all measurements must be exceedingly accurate in order to investigate a range of forces. This test structure can be used to determine the friction coefficients for surfaces with and without lubricating coatings.

34.3 Measurements of Mechanical Properties

All of the techniques discussed in Sects. 34.1 and 34.2 have been used to measure the mechanical properties of MEMS and NEMS materials. As a general rule, the results from the various techniques have agreed well with each other, and the argument becomes which of the measurement techniques is easiest and most reliable to perform. It is crucial to keep in mind, however, that certain properties, such as strength, are process dependent, and so the results taken at one laboratory will not necessarily match those taken from another. This will be discussed in more detail in Sect. 34.3.1.

34.3.1 Mechanical Properties of Polysilicon

In current MEMS technology, the most widely used structural material is polysilicon deposited by low-pressure chemical vapor deposition (LPCVD). One reason for the prevalence of polysilicon is the large body of processing knowledge for this material that has been developed by the integrated circuit community. Another reason, of course, is that polysilicon possesses a number of qualities beneficial for MEMS devices, particularly its high strength and Young's modulus. As a natural result, most of the mechanical properties investigations on MEMS materials have focused on polysilicon.

Residual Stresses in Polysilicon

The residual stresses of LPCVD polysilicon have been thoroughly characterized using both the wafer curvature technique, discussed in Sect. 34.1.1, and the microstrain gauges, discussed in Sect. 34.2.1. The results from both techniques give consistent values. Figure 34.17 summarizes the residual stress measurements as a function of deposition temperature taken from five different investigations at five different laboratories [34.29]. All five sets of data show the same trend. The stresses change from compressive at the lowest deposition temperatures to tensile at intermediate temperatures and back to compressive at the highest temperatures. The exact transition temperatures vary somewhat among the different investigations, probably due to differences in the deposition conditions: the silane or dichlorosilane pressure, the gas flow rate, the geometry of the deposition system, and the temperature uniformity. But in each data set the transitions are easily discernible. The origin of these residual stress changes lies with the microstructure of the LPCVD polysilicon films.

As with all deposited films, the microstructure of LPCVD polysilicon films is dependent on the deposition conditions. In general, the films are amorphous at

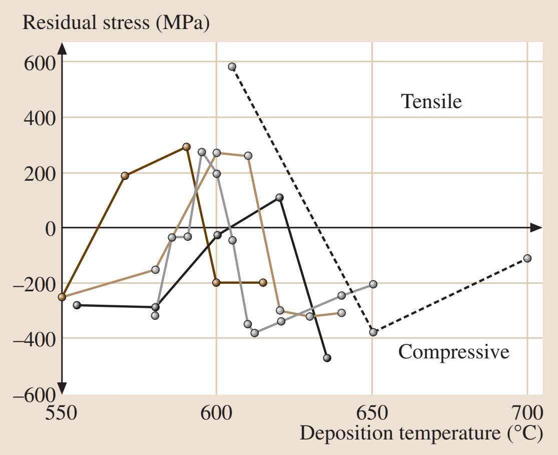

Fig. 34.17 Results for residual stress of LPCVD polysilicon films taken from five different investigations [34.29]. The data from each investigation are connected by a line

the lowest growth temperatures (lower than ∼ 570 °C), display fine (∼ 0.1 μm diameter) grains at intermediate temperatures (∼ 570 °C to ∼ 610 °C), and contain columnar (110)-textured grains with a thin fine-grained nucleation layer at the substrate interface at higher temperatures (∼ 610 °C to ∼ 700 °C) [34.29]. The fine-grained microstructure results from the homogeneous nucleation and growth of silicon grains within an as-deposited amorphous silicon film. In this regime, the deposition rate is just slightly faster than the crystallization rate. The as-deposited films will be crystalline near the substrate interface and amorphous at the free surface. (The amorphous fraction can be quickly crystallized by annealing above 610 °C.) The columnar microstructure seen at the higher growth temperatures results from the formation of crystalline silicon films as deposited, with growth being fastest in the ⟨110⟩ directions.

The origin of the tensile stress in the fine-grained polysilicon arises from the volume decrease that accompanies the crystallization of the as-deposited amorphous material. The origins of the compressive stresses in the amorphous and columnar films are less well understood. One proposed explanation for compressive stress generation during thin film growth postulates that an increase in the surface chemical potential is caused by the deposition of atoms from the vapor; the increase in surface chemical potential induces atoms to flow into newly formed grain boundaries, creating a compressive stress in the film [34.30].

Stress gradients are also typical of LPCVD polysilicon films. The partially amorphous films contain large stress gradients since they are essentially bilayers of compressive amorphous silicon on top of tensile fine-grained polysilicon. The fully crystalline films also exhibit stress gradients. The columnar compressive films are most highly stressed at the film-substrate interface, with the compressive stresses decreasing as the film thickness increases; the fine-grained films are less tensile at the film-substrate interface, with the tensile stresses increasing as the film thickness increases [34.29]. Both stress gradients are associated with microstructural variations. For the columnar films, the initial nucleation layer corresponds to a very high compressive stress, which decreases as the columnar morphology develops. For the fine-grained films, the region near the film-substrate interface has a slightly smaller average grain size, due to heterogeneous nucleation at the interface. This region displays a slightly lower tensile stress than the rest of the film, since the increased grain boundary area reduces the local density.

Young's Modulus of Polysilicon

The Young's modulus of polysilicon films has been measured using all the techniques discussed in Sect. 34.2.2. A good review of the experimental results taken from bulge testing, tensile testing, beam bending, and lateral resonators are contained in [34.31]. All of the reported results are in reasonable agreement, varying from 130 to 175 GPa, though many values are reported with a relatively high experimental scatter. The main origin of the error in these results is the uncertainties involving the geometries of the small specimens used to make the measurements. For example, from (34.13), the Young's modulus determined by the lateral resonators depends on the cube of the tether beam width, typically about 2 μm. In general, the beam width and other dimensions can be measured with scanning electron microscopy to within about 0.1 μm; however, the width of the beams is not perfectly constant along the entire length or even through the thickness. These uncertainties in geometry lead to uncertainties in modulus.

In addition, the various experimental measurements lie close to the Voigt and Reuss bounds for Young's modulus calculated using the elastic stiffnesses and compliances for single crystal silicon [34.31]. This strongly implies that Young's modulus of micro- and nanomachined polysilicon structures will be the same as for bulk samples made from polysilicon. This is not unexpected, since Young's modulus is a material property. It is related to the interatomic interactions and should have no dependence on the geometry of the sample. It should be noted that polysilicon can display a preferred crystallographic orientation depending on the deposition conditions, and that this could affect the Young's modulus of the material, since the Young's modulus of silicon is not isotropic. But the anisotropy is fairly small for the cubic silicon.

A more recent investigation that utilized electrostatically actuated cantilevers and interferometric deflection detection yielded a Young's modulus of 164 GPa [34.19]. They found the grains in their polysilicon films to be randomly oriented and calculated the Voigt and Reuss bounds to be 163.4–164.4 GPa. This appears to be a very reliable value for randomly oriented polysilicon.

Fracture Toughness and Strength of Polysilicon

Using the device shown in Fig. 34.13a and the specimen shown in Fig. 34.13d, the fracture toughness, K_{Ic}, of polysilicon has been shown to be $1.0 \pm 0.1\,\mathrm{MPa\,m^{1/2}}$ [34.32]. Several different polysilicon microstructures were tested, including fine-grained, columnar, and multilayered. Amorphous silicon was also investigated. All of the microstructures displayed the same K_{Ic}. This indicates that, like Young's modulus, fracture toughness is a material property, independent of the material microstructure or the geometry of the sample.

A tensile test, such as shown in Fig. 34.7 but using a sample with indentation-induced sharp pre-cracks, reveals a K_{Ic} of 0.86 MPa m$^{1/2}$ [34.33]. The passive, residual stress loaded beams with sharp pre-cracks, shown in Fig. 34.6, gave a K_{Ic} of 0.81 MPa m$^{1/2}$ [34.13].

Given that K_{Ic} is a material property for polysilicon, measured fracture strength, σ_{crit}, is related to K_{Ic} by

$$K_{\mathrm{Ic}} = c\sigma_{\mathrm{crit}}(\pi a)^{1/2} , \qquad (34.14)$$

where a is the crack-initiating flaw size, and c is a constant of order unity. The value for c will depend on the exact size, shape, and orientation of the flaw; for a semicircular flaw, c is equal to 0.71 [34.34]. Therefore, any differences in the reported fracture strength of polysilicon will be the result of changes in a.

A good review of the experimental results available in the literature for polysilicon strength is contained in [34.35]. The tensile strength data vary from about 0.5 to 5 GPa. Like many brittle materials, the measured strength of polysilicon is found to obey Weibull statistics. This implies that the polysilicon samples contain a random distribution of flaws of various sizes, and that the failure of any particular specimen will occur at the largest flaw that experiences the highest stress. One consequence of this behavior is that, since larger specimens have a greater probability of containing larger flaws, they will exhibit decreased strengths. More specifically, it was found that the most important geometrical parameter is the surface area of the sidewalls of a polysilicon specimen [34.35]. The sidewalls, as opposed to the top and bottom surfaces, are those surfaces created by etching the polysilicon film. This is not surprising since LPCVD polysilicon films contain essentially no flaws within the bulk, and the top and bottom surfaces are typically very smooth.

As a result, the etching techniques used to create the structures will have a strong impact on the fracture strength of the material. For single-crystal silicon specimens it was found that the choice of etchant could change the observed tensile strength by a factor of two [34.36]. In addition, the bend strength of amorphous silicon was measured to be twice that of polysilicon for specimens processed identically [34.32]. It was found that the reactive ion etching used to fabricate the specimens produced much rougher sidewalls on the polysilicon than on the amorphous silicon.

Fatigue of Polysilicon

Fatigue failure involves fracture after a number of load cycles, when each individual load is not sufficient by itself to generate catastrophic cracking in the material. For ductile materials, such as metals, fatigue occurs due to accumulated damage at the site of maximum stress and involves local plasticity. As a brittle material, polysilicon would not be expected to be susceptible to cyclic fatigue. But fatigue has been observed for polysilicon tensile samples [34.33], polysilicon bend specimens with notches [34.26, 37], and polysilicon bend specimens with sharp cracks [34.38]. The exact origins of the fatigue behavior are still subject to debate. But some aspects of the experimental data are that the fatigue lifetime does not depend on the loading frequency [34.33], the fatigue behavior is affected by the ambient [34.13, 38], and the fatigue depends on the ratio of compressive to tensile stresses seen in the load cycle [34.13].

Friction of Polysilicon

The friction of polysilicon structures has been measured using the techniques described in Sect. 34.2.2. The measured coefficient of friction was found to vary from 4.9 [34.27] to 7.8 [34.28].

34.3.2 Mechanical Properties of Other Materials

As discussed above, of all the materials used for MEMS and NEMS, polysilicon has generated the most interest as well as the most research in mechanical properties characterization. But measurements have been taken on other materials, and these are summarized in this section.

As discussed in Sect. 34.2.2, one advantage of the externally loaded tension test, as shown in Fig. 34.7, is that essentially any material can be tested using this technique. As such, tensile strengths have been measured to be 0.6 to 1.9 GPa for $\mathrm{SiO_2}$ [34.39] and 0.7 to 1.1 GPa for titanium [34.40]. The yield strength for electrodeposited nickel was found to vary from 370 to 900 MPa, depending on annealing temperature [34.16]. In addition, the yield strength was strongly affected by the current density during the electrodeposition process. Both the annealing and current density effects were correlated to changes in the microstructure of the material. Young's moduli were determined to be 100 GPa for titanium [34.40] and 215 GPa for electrodeposited nickel [34.16].

The technique of bending cantilever beams, shown in Fig. 34.8, can also be performed on a variety of materials. The yield strength and Young's modulus of gold were found to be 260 MPa and 57 GPa, respectively, using this method [34.18]. Another technique that can be used with a number of materials is the membrane deflection method, shown in Fig. 34.10. A polyimide membrane gave a residual stress of 32 MPa, a Young's modulus of 3.0 GPa, and an ultimate strain of about four percent [34.20]. Membranes were also fabricated from polycrystalline SiC films with two different grain structures [34.41]. The film with (110)-texture columnar grains had a residual stress of 434 MPa and a Young's modulus of 349 GPa. The film with equiaxed (110)- and (111)-textured grains had a residual stress of 446 MPa and a Young's modulus of 456 GPa.

Other devices used for measuring mechanical properties require more complicated micromachining, namely patterning, etching, and release, in order to operate. These devices are more difficult to fabricate with materials not commonly used as MEMS structural materials. The following examples, however, demonstrate work in this area. The structure shown in Fig. 34.5b was fabricated from Si_xN_y and revealed a nominal fracture toughness of 1.8 MPa $m^{1/2}$ [34.12]. Lateral resonators of the type shown in Fig. 34.11 were processed using polycrystalline SiC, and Young's modulus was determined to be 426 GPa [34.42]. The device shown in Fig. 34.12 was fabricated from polycrystalline germanium and used to measure a bend strength of 1.5 GPa for unannealed Ge and 2.2 GPa for annealed Ge [34.43].

References

34.1 W. Nix: Mechanical properties of thin films, Metall. Trans. A **20** (1989) 2217–2245

34.2 G. G. Stoney: The tension of metallic films deposited by electrolysis, Proc. R. Soc. Lond. A **82** (1909) 172–175

34.3 W. Brantley: Calculated elastic constants for stress problems associated with semiconductor devices, J. Appl. Phys. **44** (1973) 534–535

34.4 A. Ni, D. Sherman, R. Ballarini, H. Kahn, B. Mi, S. M. Phillips, A. H. Heuer: Optimal design of multilayered polysilicon films for prescribed curvature, J. Mater. Sci., in press

34.5 J. A. Floro, E. Chason, S. R. Lee, R. D. Twesten, R. Q. Hwang, L. B. Freund: Real-time stress evolution during $Si_{1-x}Ge_x$ heteroepitaxy: Dislocations, islanding, and segregation, J. Electron. Mater. **26** (1997) 969–979

34.6 X. Li, B. Bhusan: Micro/nanomechanical characterization of ceramic films for microdevices, Thin Solid Films **340** (1999) 210–217

34.7 M. P. de Boer, H. Huang, J. C. Nelson, Z. P. Jiang, W. W. Gerberich: Fracture toughness of silicon and thin film micro-structures by wedge indentation, Mater. Res. Soc. Symp. Proc. **308** (1993) 647–652

34.8 H. Guckel, D. Burns, C. Rutigliano, E. Lovell, B. Choi: Diagnostic microstructures for the measurement of intrinsic strain in thin films, J. Micromech. Microeng. **2** (1992) 86–95

34.9 F. Ericson, S. Greek, J. Soderkvist, J.-A. Schweitz: High sensitivity surface micromachined structures for internal stress and stress gradient evaluation, J. Micromech. Microeng. **7** (1997) 30–36

34.10 L. S. Fan, R. S. Muller, W. Yun, R. T. Howe, J. Huang: Spiral microstructures for the measurement of average strain gradients in thin films, Proc. IEEE Micro Electro Mechanical Systems Workshop, Napa Valley 1990 (IEEE, New York 1990) 177–182

34.11 M. Biebl, H. von Philipsborn: Fracture strength of doped and undoped polysilicon, Proc. Intl. Conf. Solid-State Sensors and Actuators, Stockholm 1995, ed. by S. Middelhoek, K. Cammann (Royal Swedish Academy of Engineering Sciences, Stockholm 1995) 72–75

34.12 L. S. Fan, R. T. Howe, R. S. Muller: Fracture toughness characterization of brittle films, Sens. Actuators A **21-23** (1990) 872–874

34.13 H. Kahn, R. Ballarini, J. J. Bellante, A. H. Heuer: Fatigue failure in polysilicon is not due to simple stress corrosion cracking, Science **298** (2002) 1215–1218

34.14 H. Kahn, N. Tayebi, R. Ballarini, R. L. Mullen, A. H. Heuer: Wafer-level strength and fracture toughness testing of surface-micromachined MEMS devices, Mater. Res. Soc. Symp. Proc. **605** (2000) 25–30

34.15 W. N. Sharpe Jr., B. Yuan, R. L. Edwards: A new technique for measuring the mechanical properties of thin films, J. Microelectromech. Syst. **6** (1997) 193–199

34.16 H. S. Cho, W. G. Babcock, H. Last, K. J. Hemker: Annealing effects on the microstructure and mechanical properties of LIGA nickel for MEMS, Mater. Res. Soc. Symp. Proc. **657** (2001) EE5.23.1–EE5.23.6

34.17 W. Suwito, M. L. Dunn, S. J. Cunningham, D. T. Read: Elastic moduli, strength, and fracture initiation at sharp notches in etched single crystal silicon microstructures, J. Appl. Phys. **85** (1999) 3519–3534

34.18 T. P. Weihs, S. Hong, J. C. Bravman, W. D. Nix: Mechanical deflection of cantilever microbeams: A new technique for testing the mechanical properties of thin films, J. Mater. Res. **3** (1988) 931–942

34.19 B. D. Jensen, M. P. de Boer, N. D. Masters, F. Bitsie, D. A. La Van: Interferometry of actuated microcantilevers to determine material properties and test structure nonidealities in MEMS, J. Microelectromech. Syst. **10** (2001) 336–346

34.20 M. G. Allen, M. Mehregany, R. T. Howe, S. D. Senturia: Microfabricated structures for the in situ measurement of residual stress, Young's modulus, and ultimate strain of thin films, Appl. Phys. Lett. **51** (1987) 241–243

34.21 H. Kahn, S. Stemmer, K. Nandakumar, A. H. Heuer, R. L. Mullen, R. Ballarini, M. A. Huff: Mechanical properties of thick, surface micromachined polysilicon films, Proc. IEEE Micro Electro Mechanical Systems Workshop, San Diego 1996, ed. by M. G. Allen, M. L. Redd (IEEE, New York 1996) 343–348

34.22 L. Kiesewetter, J.-M. Zhang, D. Houdeau, A. Steckenborn: Determination of Young's moduli of micromechanical thin films using the resonance method, Sens. Actuators A **35** (1992) 153–159

34.23 W. C. Tang, T.-C. H. Nguyen, R. T. Howe: Laterally driven polysilicon resonant microstructures, Sens. Actuators A **20** (1989) 25–32

34.24 P. T. Jones, G. C. Johnson, R. T. Howe: Fracture strength of polycrystalline silicon, Mater. Res. Soc. Symp. Proc. **518** (1998) 197–202

34.25 P. Minotti, R. Le Moal, E. Joseph, G. Bourbon: Toward standard method for microelectromechanical systems material measurement through on-chip electrostatic probing of micrometer size polysilicon tensile specimens, Jpn. J. Appl. Phys. **40** (2001) L120–L122

34.26 C. L. Muhlstein, E. A. Stach, R. O. Ritchie: A reaction-layer mechanism for the delayed failure of micron-scale polycrystalline silicon structural films subjected to high-cycle fatigue loading, Acta Mater. **50** (2002) 3579–3595

34.27 M. G. Lim, J. C. Chang, D. P. Schultz, R. T. Howe, R. M. White: Polysilicon microstructures to characterize static friction, Proc. IEEE Micro Electro Mechanical Systems Workshop, Napa Valley 1990 (IEEE, New York 1990) 82–88

34.28 B. T. Crozier, M. P. de Boer, J. M. Redmond, D. F. Bahr, T. A. Michalske: Friction measurement in MEMS using a new test structure, Mater. Res. Soc. Symp. Proc. **605** (2000) 129–134

34.29 J. Yang, H. Kahn, A. Q. He, S. M. Phillips, A. H. Heuer: A new technique for producing large-area as-deposited zero-stress LPCVD polysilicon films: The MultiPoly process, J. Microelectromech. Syst. **9** (2000) 485–494

34.30 E. Chason, B. W. Sheldon, L. B. Freund, J. A. Floro, S. J. Hearne: Origin of compressive residual stress in polycrystalline thin films, Phys. Rev. Lett. **88** (2002) 156103-1–156103-4

34.31 S. Jayaraman, R. L. Edwards, K. J. Hemker: Relating mechanical testing and microstructural features of polysilicon thin films, J. Mater. Res. **14** (1999) 688–697

34.32 R. Ballarini, H. Kahn, N. Tayebi, A. H. Heuer: Effects of microstructure on the strength and fracture toughness of polysilicon: A wafer level testing approach, ASTM STP **1413** (2001) 37–51

34.33 J. Bagdahn, J. Schischka, M. Petzold, W. N. Sharpe Jr.: Fracture toughness and fatigue investigations of polycrystalline silicon, Proc. SPIE **4558** (2001) 159–168

34.34 I. S. Raju, J. C. Newman Jr.: Stress intensity factors for a wide range of semi-elliptical surface cracks in finite-thickness plates, Eng. Fract. Mech. **11** (1979) 817–829

34.35 J. Bagdahn, W. N. Sharpe Jr., O. Jadaan: Fracture strength of polysilicon at stress concentrations, J. Microelectromech. Syst. (2003) 302–312

34.36 T. Yi, L. Li, C.-J. Kim: Microscale material testing of single crystalline silicon: Process effects on surface morphology an dtensile strength, Sens. Actuators A **83** (2000) 172–178

34.37 H. Kahn, R. Ballarini, R. L. Mullen, A. H. Heuer: Electrostatically actuated failure of microfabricated polysilicon fracture mechanics specimens, Proc. R. Soc. Lond. A **455** (1999) 3807–3923

34.38 W. W. Van Arsdell, S. B. Brown: Subcritical crack growth in silicon MEMS, J. Microelectromech. Syst. **8** (1999) 319–327

34.39 T. Tsuchiya, A. Inoue, J. Sakata: Tensile testing of insulating thin films; humidity effect on tensile strength of SiO_2 films, Sens. Actuators A **82** (2000) 286–290

34.40 H. Ogawa, K. Suzuki, S. Kaneko, Y. Nakano, Y. Ishikawa, T. Kitahara: Measurements of mechanical properties of microfabricated thin films, Proc. IEEE Micro Electro Mechanical Systems Workshop (IEEE, New York 1997) 430–435

34.41 S. Roy, C. A. Zorman, M. Mehregany: The mechanical properties of polycrystalline silicon carbide films determined using bulk micromachined diaphragms, Mater. Res. Soc. Symp. Proc. **657** (2001) EE9.5.1–EE9.5.6

34.42 A. J. Fleischman, X. Wei, C. A. Zorman, M. Mehregany: Surface micromachining of polycrystalline SiC deposited on SiO_2 by APCVD, Mater. Sci. Forum **264-268** (1998) 885–888

34.43 A. E. Franke, E. Bilic, D. T. Chang, P. T. Jones, T.-J. King, R. T. Howe, G. C. Johnson: Post-CMOS integration of germanium microstructures, Proc. Intl. Conf. Solid-State Sensors and Actuators (IEE Jpn., Tokyo 1999) 630–637

35. Thermo- and Electromechanics of Thin-Film Microstructures

Applications using thin-film micromechanical structures for actuation and sensing require the coupling of energy between various physical domains. This chapter focuses on two important couplings: thermomechanics and electromechanics. Thermomechanical phenomena is considered in Sect. 35.1 where we describe broad aspects of the deformation characteristics and stress states that arise when dealing with a large class of thin-film microstructures. These include the origin of stresses in multiplayer films and their qualitative evolution through processing and release from the substrate. A basic framework is described for the analysis of the thermo mechanics of multiplayer films, emphasizing linear response. Issues of geometric and material no linearity are then taken up, and equal emphasis is put on the generality of the analysis approach and specific applications. As much as possible, we show comparisons between theoretical predictions and companion experimental results.

A common use of electro mechanics in microsystems involves the application of an electric potential between two electrodes where one is fixed and the other is connected to a deformable elastic structure. The electric potential produces an electric field and an associated electrostatic force that deforms the structure, and in turn alters the electrostatic force, resulting in fully-coupled nonlinear behavior. At some point an instability can occur where the deformable structure snaps into contact with the fixed electrode. This phenomena, called pull-in is often used for switching applications. In Sect. 35.2 we describe the basic electromechanical phenomena using a parallel-plate electrostatic actuator as a reference. We discuss many important phenomena including pull-in, external forcing, stabilization, time response, the effects of dielectric charging, and breakdown of gases in small gaps. We address these phenomena for a wide range of micro mechanical structures including cantilevered beams and plates, torsionally suspended plates, and zipper actuators with curved electrodes.

35.1 **Thermomechanics of Multilayer Thin-Film Microstructures** .. 1041
  35.1.1 Basic Phenomena 1041
  35.1.2 A General Framework for the Thermomechanics of Multilayer Films 1046
  35.1.3 Nonlinear Geometry 1054
  35.1.4 Nonlinear Material Behavior 1058
  35.1.5 Other Issues 1061

35.2 **Electromechanics of Thin-Film Microstructures** 1061
  35.2.1 Applications of Electromechanics . 1061
  35.2.2 Electromechanics Analysis 1063
  35.2.3 Electromechanics – Parallel-Plate Capacitor 1064
  35.2.4 Electromechanics of Beams and Plates 1066
  35.2.5 Electromechanics of Torsional Plates 1068
  35.2.6 Leveraged Bending 1069
  35.2.7 Electromechanics of Zipper Actuators 1070
  35.2.8 Electromechanics for Test Structures 1072
  35.2.9 Electromechanical Dynamics: Switching Time 1073
  35.2.10 Electromechanics Issues: Dielectric Charging 1074
  35.2.11 Electromechanics Issues: Gas Discharge 1075

35.3 **Summary and Mention of Topics not Covered** 1078

References 1078

Microsystems rely heavily on thin-film technology: the deposition, patterning, etching, and so on, of multiple film layers to yield micromechanical structures. We do not discuss fabrication techniques here but instead refer the interested reader to many excellent references that describe the fabrication of micromechanical structures by means of thin-film technology [35.1, 2] and, in particular, surface micromachining [35.3]. Regardless of the fabrication process flow, the end result is a thin-film microstructure that is connected to a substrate through one or more *sacrificial* film layers, which must be released to render it freestanding and thus useful. Figure 35.1a illustrates a simple, yet typical in terms of the microstructure and the materials involved, case: a cantilevered beam made of polycrystalline silicon (polysilicon) anchored to a silicon substrate and encased in an SiO_2 sacrificial layer. Also shown is a metal electrode used for electrostatic actuation of the cantilever after release. This is the state of affairs after the thin-film fabrication steps; we refer to it as the as-processed but unreleased condition. To render the beam freestanding, although anchored to the substrate at one end, the sacrificial material must be removed (Fig. 35.1b). For SiO_2, this is usually accomplished by etching in an HF solution, followed by drying. The latter step is important because during wet etching, strong (at the microscale) capillary forces can pull the beam into contact with the substrate. If the microstructure is too compliant, adhesive forces between the microstructure and the substrate can pin the beam to the substrate, rendering it useless. This phenomena is often referred to as stiction, and many recent references describe the phenomena, its impact regarding reliability, and practical ways to overcome it [35.4, 5]. For our purposes, we assume that the microstructure can be successfully released so that it is free from the substrate where desired and anchored where desired.

High stresses can develop in the film layers during processing, and these can vary significantly from layer to layer and even within a single layer. When the sacrificial layers are removed, the stresses redistribute, but at the expense of deformation. This can be detrimental in applications in which planarity is essential, or it can be used advantageously for actuation purposes. Indeed many applications are based on the use of thermally induced bending of bilayer beams and plates that arises due to the difference in thermal expansion coefficients between the film layers. An example is shown in Fig. 35.2 where a micromirror is suspended above the substrate by 2.0-μm-thick gold/polysilicon bilayer beams that position the mirror (or any desired device) and can then be electrostatically actuated to control the position of the micromirror. Upon release, microstructures are typically used for actuation and/or sensing functions. This requires the coupling of the mechanical behavior of the microstructure with various energy domains. These couplings include, among others, thermomechanics, electromechanics, electrothermomechanics, and magnetomechanics. The most common, and perhaps the most easily realizable, is electromechanics, specifically the coupling of electrostatics and mechanics. This coupling has been successfully demonstrated by many applications as both sensors and actuators.

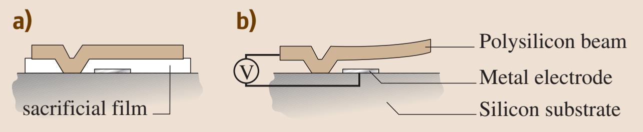

Fig. 35.1a,b Schematic of thin-film microstructures **(a)** attached to a substrate layer, and **(b)** freestanding after release from the substrate by etching of the sacrificial film

This chapter focuses on the behavior of thin-film microstructures under the action of thermomechanical and electromechanical loadings. Section 35.1 is devoted to the development of stresses and deformation in multilayer thin-film microstructures when subjected to thermomechanical loading. It focuses heavily on linear thermoelastic behavior but describes important issues regarding geometric and material nonlinearity. Section 35.2 is devoted to the electromechanics of microstructures, emphasizing general approaches to modeling and then focusing on the behavior of typ-

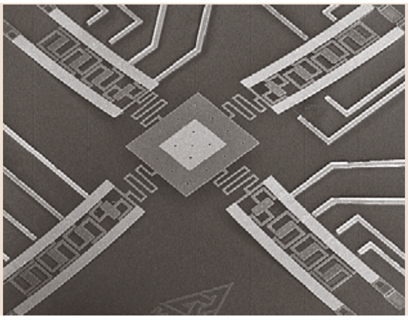

Fig. 35.2 Scanning electron microscope image of a micromirror (300 μm square) supported above the substrate by gold/polysilicon bilayer beams that can be individually actuated electrostatically to control the position of the micromirror

ical classes of electromechanical microstructures. The treatment is certainly not exhaustive, but we hope it is detailed enough to provide the interested reader with an entry-level understanding of the basic phenomena and some guidance regarding where to turn to obtain more in-depth information.

35.1 Thermomechanics of Multilayer Thin-Film Microstructures

In this section we describe aspects of the deformation characteristics and stress states in multilayer thin-film microstructures. Deformation and stress states will be considered for microstructures that are both attached to a substrate and freestanding after release from the substrate. In the modeling formalism that we will describe, the substrate simply serves as another layer. We will refer to microstructures as beams and/or plates, depending on the relevant dimensions. At times we will use the terms loosely, but we will describe the difference of their deformation characteristics in detail. The remainder of this section will be organized as follows: Section 35.1.1 provides an overview of the basic phenomena including the origin of stresses in multilayer films and their qualitative evolution through processing and then release from the substrate. Section 35.1.2 outlines a general framework for the analysis of the thermomechanical behavior of multilayer thin films, emphasizing *linear* response. Issues of *geometric* and *material* nonlinearity are then briefly taken up in Sects. 35.1.3 and 35.1.4, respectively. Section 35.1.5 touches on a number of related and important issues that were not covered in the other sections.

The objectives of this section are to describe the thermomechanical phenomena that arise when dealing with a large class of thin-film microstructures. A basic framework is described for analysis, and equal emphasis is put on the generality of the approach and specific applications. When possible, we show comparisons between theoretical predictions and companion experimental results. To this end, a specific material and geometric system must be chosen, and we use a gold/polysilicon multilayer for most of these cases. We note, though, that the phenomena discussed are not restricted to the gold/polysilicon system but arise in many multilayer film systems.

35.1.1 Basic Phenomena

During the multiple thin-film deposition and etching processes, many complicated mechanisms occur that result in straining of the film layers, and these generally differ from layer to layer. The requirement that the film layers maintain coherency at the interfaces between them means that these strains cannot occur freely, and thus they are constrained. This constraint leads to stresses in the layers and curvature of the multilayer film system; both can have significant practical implications. The mechanisms that lead to internal stresses include chemical reactions, lattice mismatch, and grain growth, among others, and are often called *intrinsic* stresses (see, for example [35.6]). The details of intrinsic stress development in thin films have been the subject of extensive study for both scientific and technological reasons, but they are beyond the scope of our study. In addition, stresses may arise due to thermal expansion mismatch between layers when subjected to a temperature change. Such stresses are often called *extrinsic* stresses. Now, in the unreleased configuration of Fig. 35.1a, a complicated stress distribution exists as a result of the intrinsic stresses. The release process to yield the configuration of Fig. 35.1b obviously alters the stress distribution as well as the state of deformation. In order to gain physical insight, the remainder of this section discusses rather qualitatively the nature of the stress states in the unreleased and released configurations. In subsequent sections details are given regarding analysis of these thermomechanical phenomena.

Transformation Strains, Misfit Strains, and Thin-Film Stresses

The important concepts here are most easily described in terms of a bilayer film system in which both layers are isotropic. Furthermore, they are easier to grasp in the context of a specific example, of which many would suffice. Once this understanding is in place, the generalization of the concepts to multilayer film systems follows naturally. To begin, we define a couple of terms. A *transformation strain* ε^* is a strain that occurs in a solid, but with no accompanying stress. It is thus an inelastic strain. In the literature a number of other terms are used to describe this concept including *stress-free strain* and *eigenstrain*. It can result, for example, from thermal expansion, a crystallographic phase transformation, or a variety of sources. When thermal expansion is the source, ε^* can be specified as $\varepsilon^* = \alpha T$ where

α is the thermal expansion coefficient of the material and T is the temperature change from a reference (usually stress-free) configuration. In general ε^* and α are second-order tensors. Stresses and deformation develop in bilayer film/substrate systems when, due to external or internal sources, each layer wants to undergo a *different* transformation strain, but the requirement of bonding of the films at the interface constrains each film to some degree. The source of the stress and deformation is then the difference in the transformation strains between the layers, i. e., $\Delta\varepsilon = \varepsilon_1^* - \varepsilon_2^*$ where ε_1^* and ε_2^* are the transformation strains of the two layers. For example, if a bilayer is subjected to a uniform temperature change T, then $\Delta\varepsilon = (\alpha_1 - \alpha_2)\,T = \Delta\alpha T$ where α_1 and α_2 are the thermal expansion coefficients of the two layers. $\Delta\varepsilon = \varepsilon_1^* - \varepsilon_2^*$ is termed the *misfit strain*. Its role as the direct source of deformation is obvious and clear for two-layer systems. The concept of misfit strain is not as direct for multilayer systems, and so we simply characterize the behavior in terms of the individual transformation strains of each layer.

It is important to understand the character of the stress distribution and deformation. To make the ideas concrete, consider the bilayer film system with planar dimensions (normal to the film thickness direction) that far exceed the total thickness. The two layers, denoted by 1 and 2, are perfectly bonded along the interface and have different isotropic thermal expansion coefficients and elastic moduli. The bilayer is initially flat, and each layer is stress free. Now assume the bilayer is subjected to a uniform temperature change T, which would lead to equibiaxial ($\varepsilon^* = \varepsilon_{xx}^* = \varepsilon_{yy}^*$, $\gamma_{xy}^* = 0$) transformation strains $\varepsilon_1^* = \alpha_1 T$ and $\varepsilon_2^* = \alpha_2 T$ in the two layers if they were not bonded. Because they are bonded, though, this results in the development of equibiaxial stresses ($\sigma = \sigma_{xx} = \sigma_{yy}$, $\sigma_{xy} = 0$) in the layers, a change of length in the planar directions (in-plane straining) and bending of the bilayer (curvature). This can be understood by the following thought experiment, which is often used in the calculation of film stresses. We will only use it, though, to explain the basics of the resulting deformation and will leave the actual calculations to the next section.

Consider the bilayer system shown schematically in Fig. 35.3a. The film layers are stress free and perfectly bonded. Our discussion will be in the context of biaxial transformation strains and stresses, but the ideas can be immediately applied to situations of uniaxial stress/strain behavior or plane stress or strain situations. We are interested in obtaining the deformation and stress state when the bilayer is subject to a uniform temperature change T. Since the films are stress free in Fig. 35.3a, one can cut the bonds that connect the two layers and separate the films (Fig. 35.3b) without generating any stresses or deforming either layer. Now we (Fig. 35.3c) subject the layers to a uniform temperature change T; each layer will undergo its stress-free biaxial strain, $\varepsilon_1^* = \alpha_1 T$ and $\varepsilon_2^* = \alpha_2 T$ (for simplicity, Fig. 35.3c illustrates the deformation for the case where $\varepsilon_2^* = 0$, $T < 0$, and $\alpha_1 > 0$, but the ideas are the same for more general situations). Now the interfaces, which were originally in perfect registry, are no longer. To bring them into registry, a uniform stress must be applied to the film (Fig. 35.3d) by means of the application of forces along the film edge. At this stage the film is subjected to the uniform biaxial stress $\sigma_1 = M_1\varepsilon_1^*$, where M_1 is the biaxial modulus of the film (see Sect. 35.1.2), the substrate is still stress free, $\sigma_2 = 0$, and the layers are in perfect registry. They can then be reconnected (imagine reattaching the original bonds that were cut in Fig. 35.3b), without generating additional stresses or deforming the layers (Fig. 35.3e). The stress

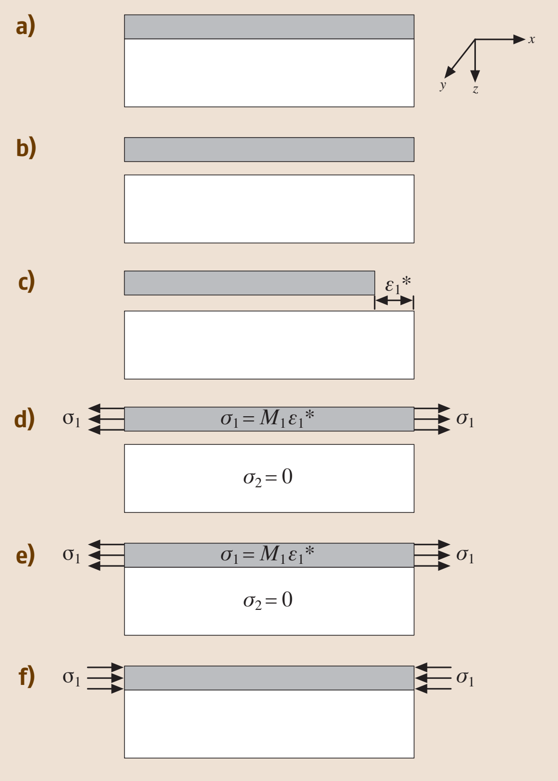

Fig. 35.3 Illustration showing the development of stresses and deformation in a bilayer film/substrate system

state in Fig. 35.3e is the same as that in Fig. 35.3d, the only difference is that the films are now bonded. Now, to recover the solution for the bilayer subjected to a uniform temperature change T, we only have to remove the edge forces. This can be accomplished by applying forces of equal magnitude but opposite sign to the edges (Fig. 35.3f). The solution is then obtained as the superposition of the problems in Fig. 35.3e, which is trivial, and Fig. 35.3f, which is difficult in general. Instead of now trying to solve the problem in Fig. 35.3f, we will simply describe the character of the resulting deformation and stress state. Two aspects are important: the behavior near the free edge where the stress in Fig. 35.3f is applied, and the behavior far away from the free edge. One can appeal to St. Venant's principle to simplify the latter, but more involved analysis is required for the former (see for example [35.7]).

At a distance of ten or so film thicknesses away from the edges, the solution to the problem in Fig. 35.3f can be decomposed, as shown in Fig. 35.4. The uniform force distribution on the edges of the film (Fig. 35.4a) can be replaced by a concentrated force acting at the middle z-coordinate of the film thickness (Fig. 35.4b). The solution to this problem can then be obtained by the superposition of the two problems in Fig. 35.4c and 35.4d, the loadings of which are statically equivalent to that in Fig. 35.4b. That in Fig. 35.4c is a concentrated force applied at the as-yet-unknown neutral surface of the bilayer; it will lead to a uniform in-plane (x-y plane) straining of the bilayer. That in Fig. 35.4d is a moment of magnitude Pd, where d is the distance between the z-coordinates of the middle of the film and the neutral surface of the bilayer; it will lead to bending of the bilayer. If there were also a film on the bottom of the substrate so that the layered film was symmetric in material properties and geometry (film thicknesses) in the z-direction, then no moment would result, and the layered system would not bend but would only strain in-plane.

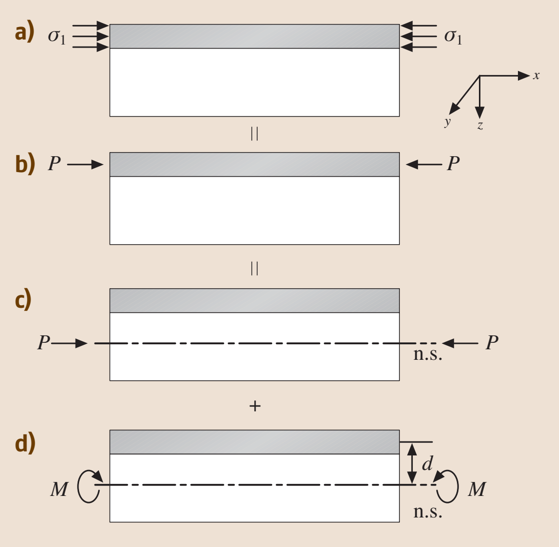

Fig. 35.4 Decomposition of the edge-loading problem into the superposition of a force and moment applied to the multilayer. Away from the edge, the force gives rise to uniform straining, and the moment gives rise to bending

It is probably obvious that very near the edge, the stress state is quite complicated. In fact, in the context of linear elasticity, a stress singularity generally exists at the intersection of the free edge and the interface between the two layers. This singular stress state is similar to that which exists at a crack tip, but the strength of the singularity differs depending on the mismatch in elastic constants of the materials [35.8, 9]. The region within which this stress state exists depends on the film thickness, and for very thin films it is often so small that it is insignificant. It has recently been shown, however, that in some cases it can be appreciable and in fact can be the primary source of failure [35.10]. As described in detail by *Hui* et al. [35.7], aside from the singular stresses, the stress state near the edges contains highly localized shear and normal stress (perpendicular to the interface) components. By St. Venant's principle these stress components decay over a rather short distance from the edge, but they can play a significant role in the durability of the film/substrate system as they can lead to delamination of the interface or cracking into one of the layers. The character of these stress components at the interface is shown qualitatively in Fig. 35.5. Note that in order to satisfy equilibrium, the normal stress (often called the peel stress) changes sign from tensile to compressive. Furthermore, since these stress components are highly localized at the edge, in the interior of the system the interface is free of normal or shear stresses. In this chapter we will not further consider these localized stresses but instead will concern ourselves with the stresses away from the edges and the companion deformation.

Our discussion has been based on the assumption that transformation strains are constant throughout the cross section of each film. In microsystems applications, though, films will often have transformation strains that vary through the thickness. This will lead to bending of the film after it is released from the substrate. Transformation strains that vary through the thickness, i. e.,

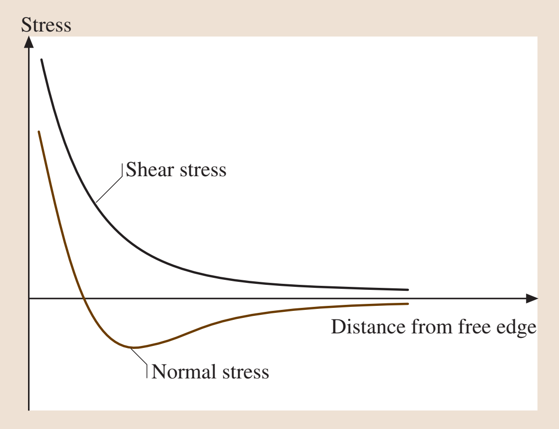

Fig. 35.5 Character of the normal and shear stress distribution at the film/substrate interface near the free edge

$\varepsilon^*(z)$, can be easily accommodated in the description of Fig. 35.3. It is altered in that now in Fig. 35.3c an in-plane strain results, the magnitude of which is given by the through-thickness average of $\varepsilon^*(z)$, but in addition, the film bends with constant curvature. Then, both a force (uniform stress) and moment (linearly varying stress) must be applied in Fig. 35.3d to flatten the film and lengthen it so that it is again in perfect registry with the substrate and can then be rebonded [35.11].

This description provides the starting point for understanding the effects of transformation strains on stresses and deformation in a multilayer thin-film system. It is an accurate picture of the state of affairs before release of a film from the substrate to yield a freestanding microstructure. In the next sections we will discuss the changes in the deformation that occur upon release as the constraint of the substrate is removed.

Single-Layer Beams

To keep things simple, in this and the following section we focus on *beams*, which have one nonzero normal stress component, that along the beam axis. In subsequent sections we will discuss the limitations of approximating such structures as beams and proceed with a more appropriate development in terms of *plates*. Consider the beam shown in Fig. 35.1. For simplicity we assume that the sacrificial material is much thinner than the substrate, has the same elastic moduli as the substrate, and so before release is mechanically equivalent to the substrate. In this way we can describe the stress distribution before release as that of a two-layer beam. To make the discussion more tangible, consider a specific but realistic case of a polysilicon beam on a silicon substrate. The top layer of thickness 1.5 µm is the polysilicon film, and the bottom layer of thickness 500 µm is the silicon substrate (including the very thin sacrificial layer that will be removed). Assume that the polysilicon film has an intrinsic mean stress S and a linear stress gradient β (we can easily consider a nonlinear stress gradient, but the simple linear gradient suffices to demonstrate the concepts involved). This intrinsic stress is interpreted as the stress state in the film if it were free, i. e., a transformation stress. It could be expressed in terms of a transformation strain $\varepsilon^*(z)$, but since reference is often made to thin-film stresses we will carry out this and the following example using this transformation stress rather than a transformation strain. We consider three cases to isolate the effects of the mean stress and the stress gradient: (a) $S = -10\,\text{MPa}$, $\beta = 0$, (b) $S = 0$, $\beta = 0.8\,\text{MPa}/\mu\text{m}$, and (c) $S = -10\,\text{MPa}$, $\beta = 0.8\,\text{MPa}/\mu\text{m}$. In all cases, the intrinsic stress in the substrate is zero.

In Table 35.1 we give calculated results using a simple beam theory (to be discussed in the following sections) for the three scenarios. Case (a) $S = -10\,\text{MPa}$, $\beta = 0$ results in a slight contraction and a negative curvature (bent downward) of the film/substrate system prior to release. The compressive stress in the polysilicon beam is uniform but slightly lower than the value of S. The stress induced in the substrate is small and varies linearly through its thickness. It changes sign through the substrate, a necessary requirement to satisfy equilibrium. After release the stress is zero throughout the beam. To relax the stress, the beam extends (the mid-plane strain ε^0 is positive), but because $\beta = 0$ it does not bend. Case (b) $S = 0\,\text{MPa}$, $\beta = 0.8\,\text{MPa}/\mu\text{m}$ results in a slight positive curvature (bent upward) but no extension or contraction of the film/substrate system prior to release. The mean compressive stress in the polysilicon beam is zero, and the gradient of the stress distribution through the film is decreased from the value of β. The stress induced in the substrate is again small and varies linearly through its thickness, changing sign. After release the stress is again zero throughout the beam. The beam does not extend or contract since $S = 0$, but it bends, assuming a curvature $\kappa = \beta/E$ where E is Young's modulus of polysilicon (see Sect. 35.1.2 for details regarding this result). Careful examination of Table 35.1 shows that the strain, curvature, and stress for case (c) $S = -10\,\text{MPa}$, $\beta = 0.8\,\text{MPa}/\mu\text{m}$ can be obtained as the superposition of the results of cases (a) and (b). Specifically, after release the beam extends and bends.

Table 35.1 Stresses, curvature, and strains in a film/substrate system consisting of a 1.5-µm-thick polysilicon film and a 500-µm-thick silicon film. The elastic moduli of the film and the substrate are assumed equal, and the film has an intrinsic mean stress and stress gradient. The sign convention for curvature is that (−) indicates bending downward and (+) indicates bending upward

		(a) $S = -10$ **MPa** $\beta = 0$ **MPa/µm**		**(b)** $S = 0$ **MPa** $\beta = 0.8$ **MPa/µm**		**(c)** $S = -10$ **MPa** $\beta = 0.8$ **MPa/µm**	
		Unreleased	**Released**	**Unreleased**	**Released**	**Unreleased**	**Released**
	ε^0 (µε)	0.0624	62.5	0	0	0.0624	62.5
	κ (m^{-1})	-7.5×10^{-4}	0	5.0×10^{-9}	5.0	-7.5×10^{-4}	5.0
σ (MPa)	Top	−9.96	0	0.2	0	−9.76	0
Polysilicon Film	Bottom	−9.96	0	−0.2	0	−10.16	0
σ (MPa)	Top	0.04	0	-2.0×10^{-7}	0	0.04	0
Silicon Substrate	Bottom	−0.02	0	2.0×10^{-7}	0	−0.02	0

These results illustrate the behavior of beams before and after release when they contain a mean stress and/or a linear stress gradient. The results also suggest natural means to attempt to measure the mean stress and the gradient. Test structures that are sensitive to length changes of beams after release can be used to characterize mean stress. For example, fracture of arrays of beams of varying width upon release can be used to estimate tensile stress, and fixed-fixed beams that buckle can be used to estimate compressive stress. The stress gradient in a beam can be determined by measuring its curvature after release. A survey of test structures suitable for these purposes can be found in [35.12].

Bilayer Beams

While stresses in single-layer microstructures are important for many applications, multilayer film microstructures abound in microsystems technology, and the development of stresses and deformation in them is also important. In order to convey the basic ideas, we consider an example of a two-layer beam fixed to a substrate in the same manner as that in the previous section. Specifically we consider the layered structure shown in the inset of Fig. 35.6: a polysilicon beam of thickness 1.5 µm attached to a silicon substrate of thickness 500 µm, covered with a gold film of thickness 0.5 µm. Photos of actual beams of this form are shown in Fig. 35.7 where the length of the longest beam shown is 600 µm, and the beams are 50 µm wide. Again we assume that the sacrificial material is very thin and has the same elastic moduli as the substrate, and so before release it is mechanically equivalent to the substrate. We assume that the polysilicon film has an intrinsic mean stress S and a linear stress gradient β as in the previous example, and that the gold film has an intrinsic mean stress but no stress gradient. We consider the reasonably typical stress states: $S = -10$ MPa and $\beta = 0.8$ MPa/µm in the polysilicon beam and $S = 50$ MPa in the gold film. The resulting curvature and stress distribution before and after release are shown in Fig. 35.6. They show that before release the stresses in the polysilicon and gold layer are nearly uniform, compared to the range of stresses throughout the layers. The multilayer experiences a positive curvature; recall that the system consisting of just the polysilicon

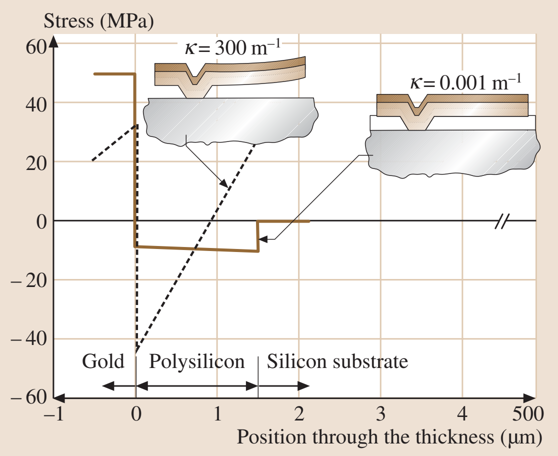

Fig. 35.6 Stress distribution through the thickness for a bilayer film consisting of a 0.5-µm-thick gold film on a 1.5-µm-thick polysilicon film subjected to an internal stresses and stress gradients. Stress distributions are shown before and after release from a 500-µm-thick silicon

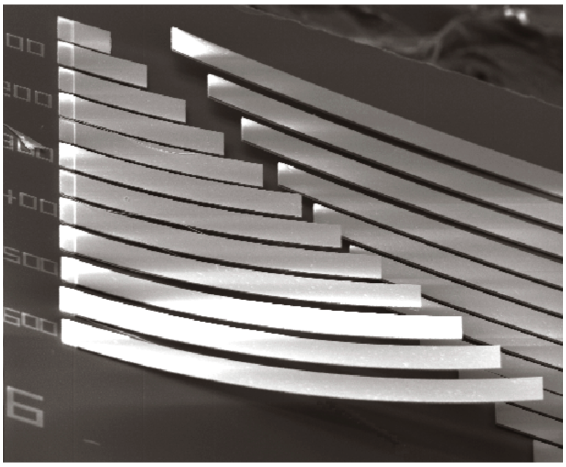

Fig. 35.7 Scanning electron microscope image of gold/polysilicon bilayer cantilever beams as released from the substrate. All beams have a 0.5-µm-thick gold film, but the beams on the *left* have a 1.5-µm-thick polysilicon film, and those on the *right* a 3.5-µm-thick polysilicon layer

film and the substrate experienced a negative curvature of about the same magnitude. Thus the tensile stress in the gold is significant enough to drive the curvature from negative (bending down) to positive (bending up).

After release from the substrate, the gold/polysilicon beam contracts and bends substantially to a curvature of about 300 m^{-1}. This should be compared with 5 m^{-1} for the polysilicon beam with a stress gradient. For a beam 300 µm long cantilevered from one end as shown in Fig. 35.6, this amounts to a tip deflection of about 13 µm. The stress states in both the gold and polysilicon are significantly altered upon release. Perhaps most significantly, large stress gradients exist through the thickness of each layer. Starting at the top of the gold film, the tensile stress increases through the gold thickness. Due to the discontinuity in elastic modulus and the residual stress and stress gradients, there is a jump in the stress of about −77 MPa at the gold/polysilicon interface. The stress is compressive at the top of the polysilicon film layer and increases linearly through the thickness of the polysilicon film, taking on its maximum tensile value at the bottom of the two-layer beam. In summary, the release process results in a reduction of the stress in the gold, an increase of the stress in the polysilicon, stress gradients in both layers, a small contraction, and a large curvature change.

These examples are meant to convey the basic phenomena and give a feel for the magnitudes of the quantities involved. The actual numerical results are based on calculations performed using a relatively simple multilayer beam theory and are reasonable estimates. Details regarding the analysis will be given in the following section, including a discussion of the suitability of the simple beam theory.

35.1.2 A General Framework for the Thermomechanics of Multilayer Films

Numerous studies have elucidated the basic thermomechanical response of multilayer material systems when subjected to temperature changes or other sources of transformation strains between the layers. These have come in the context of many technological applications, the most common being structural composite materials [35.13, 14] and thin film/substrate systems for microelectronics [35.15–22], which are directly applicable to the analysis of multilayer films.

The General Multilayer Film

When a layered film is subjected to transformation strains, two aspects of deformation play a primary role: straining of the midplane and bending. When the transverse deflections due to bending are of prime importance, as is often the case, one way to broadly characterize the deformation response, especially for plates with relatively large in-plane dimensions as compared to their thickness, is in terms of the average curvature developed as a function of the transformation strains. Formally the curvature is a second-rank tensor, and for the type of layered film problems considered here it can be wholly described by the two principal curvature components, e.g., in the x- and y- directions, κ_x and κ_y. The curvature is a pointwise quantity, meaning it varies from point to point over the in-plane dimensions of the plate. In terms of the average curvature variation as a function of misfit strain, three deformation regimes have been identified [35.16–19, 22, 23]. The first regime is a linear relation between the average curvature and temperature change where $\kappa_x = \kappa_y$, i. e., the average curvature is spherically symmetric (when the materials are isotropic). This symmetric deformation would not exist if the material properties were anisotropic. This deformation regime is characterized by both small transverse displacements and rotations, and so conventional thin-plate theory adequately describes the deformation. It is the subject of the analysis in this section. The second and third regimes, which arise due to geometric nonlinearity, will be taken up in Section 35.1.3.

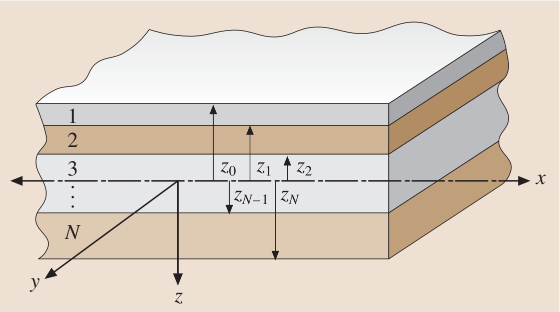

Fig. 35.8 Multilayer film system with a definition of the parameters involved in the analysis

Figure 35.8 shows a schematic of the geometry considered in this study; a multilayer film microstructure with lateral dimensions that far exceed the total thickness. As such we ignore the effects of the free edges and focus on a representative point in the layered film that lies at least a distance of a few film thicknesses from the edge. The layers are numbered $1, 2, \dots, N$ where 1 is the top layer and N is the bottom layer. Each layer has a thickness t_i $(i = 1 \rightarrow N)$, and the total thickness of the microstructure is t. Each of the film layers consists of an isotropic material with a Young's modulus E_i, Poisson's ratio, ν_i, and may undergo a transformation strain $\varepsilon_i^*(z)$, which may vary with position z through the film thickness. The source of $\varepsilon_i^*(z)$ may be arbitrary, and the effect of multiple sources can be considered as the sum of the individual ones. A common source, and one that we will focus on, is thermal expansion upon a uniform temperature change. In this case $\varepsilon_i^*(z) = \varepsilon_i^* = \alpha_i T$ where α_i is the thermal expansion coefficient and T is the temperature change from a reference state. In addition, each layer may contain an internal stress $\hat{\sigma}(z)$, which also may vary with position z through the thickness. For simplicity we focus our results on isotropic layers, but the framework is applicable to anisotropic behavior of each layer. As a result of the difference in stress-free strains and/or internal stresses between each layer, stresses are developed in the layers and the structure deforms; our intent is to compute these.

Analysis

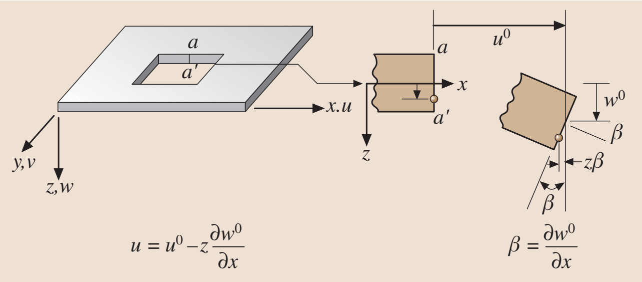
a
a'
x.u
y,v
z,w
a
u^0
x
a'
z
w^0
β
zβ
β
$u = u^0 - z\frac{\partial w^0}{\partial x}$
$\beta = \frac{\partial w^0}{\partial x}$

Fig. 35.9 Kinematics of the deformation of the thin-film multilayer

The analysis presented here is somewhat standard in the treatment of laminated composite plates but has not been often used in the analysis of multilayer thin film problems for microelectronics and/or microelectromechanical systems. Detailed accounts of parts of the theory can be found in texts on composite materials (see for example [35.24,25]), and so only a brief account is presented here.

The displacements u and v at any point z through the thickness of the laminate are assumed to be described by the classical Kirchoff hypothesis. As illustrated in Fig. 35.9, the Kirchoff hypothesis states that after loads are applied to the layered film, the line a–a' which is initially straight and normal to the geometric midplane remains straight, normal to the deformed geometric midplane, and its length is unchanged. In other words, the normal does not deform but only translates and rotates. As a consequence of this assumption, the displacements of any point in the layered film are given by:

$$\begin{aligned} u &= u^0 - z\frac{\partial w^0}{\partial x} \\ v &= v^0 - z\frac{\partial w^0}{\partial y} \\ w &= w^0 \, , \end{aligned} \tag{35.1}$$

where u^0, v^0, and w^0 are the displacements of the geometric midplane at $z = 0$. Note that in addition to this explicit dependence of the displacements on z, they are generally functions of x and y.

The nonzero normal and shear strains are then:

$$\begin{aligned} \varepsilon_x &= \frac{\partial u}{\partial x} = \frac{\partial u^0}{\partial x} - z\frac{\partial^2 w^0}{\partial x^2} = \varepsilon_x^0 + z\kappa_x \\ \varepsilon_y &= \frac{\partial v}{\partial y} = \frac{\partial v^0}{\partial x} - z\frac{\partial^2 w^0}{\partial x^2} = \varepsilon_y^0 + z\kappa_y \\ \gamma_{xy} &= \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = \frac{\partial u^0}{\partial y} + \frac{\partial v^0}{\partial x} - 2z\frac{\partial^2 w^0}{\partial x \partial y} \\ &= \gamma_{xy}^0 + z\kappa_{xy} \end{aligned} \tag{35.2a}$$

or in matrix form:

$$\{\varepsilon\} = \left\{\varepsilon^0\right\} + z\,\{\kappa\} \, , \tag{35.2b}$$

where:

$$\{\varepsilon\} = \begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{Bmatrix} \quad \left\{\varepsilon^0\right\} = \begin{Bmatrix} \varepsilon_x^0 \\ \varepsilon_y^0 \\ \gamma_{xy}^0 \end{Bmatrix} = \begin{Bmatrix} \dfrac{\partial u^0}{\partial x} \\ \dfrac{\partial v^0}{\partial y} \\ \dfrac{\partial u^0}{\partial y} + \dfrac{\partial v^0}{\partial x} \end{Bmatrix}$$

$$\{\kappa\} = \begin{Bmatrix} \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{Bmatrix} = \begin{Bmatrix} -z\dfrac{\partial^2 w^0}{\partial x^2} \\ -z\dfrac{\partial^2 w^0}{\partial y^2} \\ -2z\dfrac{\partial^2 w^0}{\partial x \partial y} \end{Bmatrix} . \tag{35.3}$$

In (35.3), $\left\{\varepsilon^0\right\}$ is the midplane strain and $\{\kappa\}$ is the midplane curvature. As with the displacements, $\left\{\varepsilon^0\right\}$ and $\{\kappa\}$ are generally functions of x and y.

The stress-strain relations for each layer are:

$$\begin{Bmatrix} \sigma_x(z) \\ \sigma_y(z) \\ \sigma_{xy}(z) \end{Bmatrix}_k = \begin{bmatrix} Q_{11} & Q_{12} & Q_{16} \\ Q_{12} & Q_{22} & Q_{26} \\ Q_{16} & Q_{26} & Q_{66} \end{bmatrix}_k \times \left[\begin{Bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{Bmatrix} - \begin{Bmatrix} \varepsilon_x^*(z) \\ \varepsilon_y^*(z) \\ \gamma_{xy}^*(z) \end{Bmatrix}_k \right] + \begin{Bmatrix} \hat{\sigma}_x(z) \\ \hat{\sigma}_y(z) \\ \hat{\sigma}_{xy}(z) \end{Bmatrix}_k \tag{35.4a}$$

or in compact form:

$$\{\sigma(z)\}_k = \{\overline{Q}\}_k \left[\{\varepsilon\} - \{\varepsilon(z)\}_k^*\right] + \{\hat{\sigma}(z)\}_k \ , \tag{35.4b}$$

where $k = 1 \rightarrow N$ denotes the layer number. Equation (35.4) is written in a global coordinate system for the layered structure, (x, y, z). As such, $\left[\bar{Q}\right]_k$ are the (generally anisotropic) elastic stiffness coefficients of the kth layer in the global coordinate system. They are obtained by transformation (see [35.24, 25], for example) of the stiffnesses, $[Q]_k$, from a natural coordinate system (x_1, x_2, x_3). For orthotropic materials the nonzero components of $[Q]$ for each layer are,

$$[Q] = \begin{bmatrix} \dfrac{E_1}{1-\nu_{12}\nu_{21}} & \dfrac{\nu_{12}E_2}{1-\nu_{12}\nu_{21}} & 0 \\ \dfrac{\nu_{21}E_1}{1-\nu_{12}\nu_{21}} & \dfrac{E_2}{1-\nu_{12}\nu_{21}} & 0 \\ 0 & 0 & G_{12} \end{bmatrix} , \tag{35.5a}$$

where E_i, ν_{ij}, and G_{ij} are the orthotropic Young's moduli, Poisson's ratios, and the shear modulus, respectively, of the layer, in the local coordinate system. Note that $Q_{12} = Q_{21}$, and so the equality of these expressions in (35.5a) shows the reciprocal relationship between E_i and ν_{ij}; only three of the four are independent. For isotropic materials $\left[\bar{Q}\right] = [Q]$, and the nonzero components for each layer are:

$$\left[\bar{Q}\right] = [Q] = \begin{bmatrix} \dfrac{E}{1-\nu^2} & \dfrac{\nu E}{1-\nu^2} & 0 \\ \dfrac{\nu E}{1-\nu^2} & \dfrac{E}{1-\nu^2} & 0 \\ 0 & 0 & G \end{bmatrix} \tag{35.5b}$$

In (35.4) $\left\{\hat{\sigma}(z)\right\}_k$ is the residual internal stress in the layer and may be an arbitrary function of z, i. e., a stress gradient may exist within each layer; $\{\varepsilon\}$ is the total strain, $\{\varepsilon^*\}_k$ is the inelastic transformation strain, and the term in rectangular brackets is the elastic strain. The stresses and strains are all defined in the (x, y, z) coordinate system. Substituting (35.2) into (35.4) yields the stress distribution through the thickness:

$$\{\sigma(z)\}_k = \{\overline{Q}\}_k \left[\left\{\varepsilon^0\right\} + z\{\kappa\} - \left\{\varepsilon^*(z)\right\}_k\right] + \left\{\hat{\sigma}(z)\right\}_k \ . \tag{35.6}$$

For the layered film system we define a resultant force $\{N\}$ and moment $\{M\}$ per unit length as:

$$\begin{Bmatrix} N_x \\ N_y \\ N_{xy} \end{Bmatrix} = \{N\} = \int_{-t/2}^{t/2} \{\sigma(z)\}_k \, \mathrm{d}z$$

$$\begin{Bmatrix} M_x \\ M_y \\ M_{xy} \end{Bmatrix} = \{M\} = \int_{-t/2}^{t/2} \{\sigma(z)\}_k \, z \, \mathrm{d}z \ . \tag{35.7}$$

For example, N_x is the force in the x-direction per unit length in the y-direction. $\{N\}$ and $\{M\}$ have units of $[F/L]$, $[F \cdot L/L]$, respectively.

Substituting (35.6) into (35.7) and breaking the integrals through the thickness into sums of integrals over each layer yields:

$$\begin{aligned} \{N\} &= [A]\left\{\varepsilon^0\right\} + [B]\{\kappa\} - \left\{N^\varepsilon\right\} + \left\{N^\sigma\right\} \\ \{M\} &= [B]\left\{\varepsilon^0\right\} + [D]\{\kappa\} - \left\{M^\varepsilon\right\} + \left\{M^\sigma\right\} \ , \end{aligned} \tag{35.8}$$

where:

$$
\begin{aligned}
[A] &= \int_{-t/2}^{t/2} \{\overline{Q}\}_k \, \mathrm{d}z = \sum_{k=1}^{N} \{\overline{Q}\}_k \, (z_k - z_{k-1}) \\
[B] &= \int_{-t/2}^{t/2} \{\overline{Q}\}_k \, z \, \mathrm{d}z = \frac{1}{2} \sum_{k=1}^{N} \{\overline{Q}\}_k \left(z_k^2 - z_{k-1}^2\right) \\
[D] &= \int_{-t/2}^{t/2} \{\overline{Q}\}_k \, z^2 \, \mathrm{d}z = \frac{1}{3} \sum_{k=1}^{N} \{\overline{Q}\}_k \left(z_k^3 - z_{k-1}^3\right) \\
\left\{N^{\varepsilon}\right\} &= \int_{-t/2}^{t/2} \{\overline{Q}\}_k \left\{\varepsilon^*(z)\right\}_k \, \mathrm{d}z \\
\left\{N^{\sigma}\right\} &= \int_{-t/2}^{t/2} \left\{\hat{\sigma}(z)\right\}_k \, \mathrm{d}z \\
\left\{M^{\varepsilon}\right\} &= \int_{-t/2}^{t/2} \{\overline{Q}\}_k \left\{\varepsilon^*(z)\right\}_k z \, \mathrm{d}z \\
\left\{M^{\sigma}\right\} &= \int_{-t/2}^{t/2} \left\{\hat{\sigma}(z)\right\}_k z \, \mathrm{d}z \, .
\end{aligned}
\tag{35.9}
$$

In the terminology of laminated composite materials, $[A]$, $[B]$, and $[D]$ are the extensional, coupling, and bending constants and are functions of the elastic moduli of each layer and the arrangement of the layers through the thickness. $[A]$, $[B]$, and $[D]$ have units of $[F/L]$, $[F]$, and $[F \cdot L]$, respectively. Physically, $[A]$ describes the connection between in-plane forces and straining of the midplane, $[D]$ describes the connection between moments and curvature, and $[B]$ connects moments to midplane strain and forces to curvature. If both the geometry and material properties, including $\varepsilon^*(z)$ and $\hat{\sigma}(z)$, of the layered film system are symmetric about $z = 0$, then $[B] = [0]$ and there is no coupling between in-plane straining and bending. Note that $[A]$, $[B]$, and $[D]$ are 3×3 matrices, as are $\left[\bar{Q}\right]_k$, while all other terms a 3×1 column vectors of the form of (35.3) and (35.4).

Equation (35.8) can be written in compact form as:

$$
\left\{ \begin{matrix} \{N\} \\ \{M\} \end{matrix} \right\} = \begin{bmatrix} [A] & [B] \\ [B] & [D] \end{bmatrix} \left\{ \begin{matrix} \left\{\varepsilon^0\right\} \\ \{\kappa\} \end{matrix} \right\} - \left\{ \begin{matrix} \{N^{\varepsilon}\} \\ \{M^{\varepsilon}\} \end{matrix} \right\} + \left\{ \begin{matrix} \{N^{\sigma}\} \\ \{M^{\sigma}\} \end{matrix} \right\} . \tag{35.10}
$$

Equation (35.10) is most easily implemented in matrix form as a 6×6 system of equations that hold at each (x, y) coordinate. These are the constitutive equations for the layered film, incorporating the Kirchoff hypothesis and the strain-displacement relations. In order to completely describe the response of a layered film due to external or internal loads, (35.10) must be supplemented with the stress equilibrium equations and appropriate boundary conditions. In general the solution of these equations for a specified multilayer film geometry and loading is a complex undertaking and recourse is often taken to numerical methods such as the finite element method (FEM). Alternatively, the equations can be formulated using an energy approach that will be advantageous when we consider geometric nonlinearity in Section 35.1.3.

Fortunately, using this formulation we can solve a number of important problems simply using (35.10). For example, if there are no externally applied loads then $\{N\} = \{M\} = \{0\}$ and (35.10) can be inverted to yield:

$$
\left\{ \begin{matrix} \left\{\varepsilon^0\right\} \\ \{\kappa\} \end{matrix} \right\} = \begin{bmatrix} [A] & [B] \\ [B] & [D] \end{bmatrix}^{-1} \times \left\{ \left\{ \begin{matrix} \{N^{\varepsilon}\} \\ \{M^{\varepsilon}\} \end{matrix} \right\} - \left\{ \begin{matrix} \{N^{\sigma}\} \\ \{M^{\sigma}\} \end{matrix} \right\} \right\} . \tag{35.11}
$$

Recall that all terms on the right-hand side are functions of the elastic moduli, the known transformation strains and stresses, and the geometrical arrangement of the layers. Once the midplane strain and curvature are computed using (35.11), the stress distribution in each layer can be computed using (35.6).

Special Cases

Here we apply (35.11) to yield some simple, yet important, results. Before doing so, it is useful to explicitly express the form of the transformation stress distribution $\left\{\hat{\sigma}(z)\right\}_k$. To this end we restrict $\left\{\hat{\sigma}(z)\right\}_k$ to be a linearly varying function through each layer. This can be written as:

$$
\left\{\hat{\sigma}(z)\right\}_k = \{S\}_k - \{\beta\}_k \left(z - \frac{z_{k-1} + z_k}{2}\right) , \tag{35.12}
$$

where $\{S\}_k$ is the average stress in layer k and $\{\beta\}_k$ is the stress gradient as shown in Fig. 35.10. For this form of the stress, $\{N^{\sigma}\}$ and $\{M^{\sigma}\}$ can be expressed

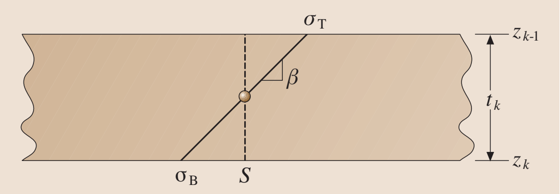

Fig. 35.10 Schematic showing a linear stress distribution characterized by a mean stress S and linear through-thickness stress gradient β

as:

$$\left\{N^{\sigma}\right\} = \sum_{k=1}^{N} \{S\}_k \left(z_k - z_{k-1}\right)$$
$$\left\{M^{\sigma}\right\} = \frac{1}{2}\sum_{k=1}^{N} \{S\}_k \left(z_k^2 - z_{k-1}^2\right) - \frac{1}{12}\sum_{k=1}^{N} \{\beta\}_k \left(z_k^3 - z_{k-1}^3\right) . \qquad (35.13)$$

Furthermore, it is useful to specify the form of $\{\varepsilon^*(z)\}$ when its source is thermal expansion during a uniform temperature change T from a reference temperature. In this case:

$$\left\{\varepsilon^*(z)\right\}_k = \left\{\varepsilon^*\right\}_k = \{\alpha\}_k T . \qquad (35.14)$$

For this form of the stress-free strain, $\{N^{\varepsilon}\}$ and $\{M^{\varepsilon}\}$ can be expressed as:

$$N^{\varepsilon} = \int_{-t/2}^{t/2} \left\{\overline{Q}\right\}_k \{\alpha\}_k T \,\mathrm{d}z = \Delta T \sum_{k=1}^{N} \left\{\overline{Q}\right\}_k \{\alpha\}_k \left(z_k - z_{k-1}\right)$$
$$M^{\varepsilon} = \int_{-t/2}^{t/2} \left\{\overline{Q}\right\}_k \{\alpha\}_k Tz \,\mathrm{d}z = \frac{\Delta T}{2} \sum_{k=1}^{N} \left\{\overline{Q}\right\}_k \{\alpha\}_k \left(z_k^2 - z_{k-1}^2\right) . \qquad (35.15)$$

In all of the examples described in this section we will consider only cases in which all layers have elastic constants, thermal expansion properties, and misfit strains that are isotropic. As a result, $A_{16} = A_{26} = B_{16} = B_{26} = D_{16} = D_{26} = 0$. The ramifications of this are that the midplane shear strains $\gamma_{xy}^0 = 0$ and the twist curvatures $\kappa_{xy} = 0$. In addition, the midplane strains and curvatures are equibiaxial, i. e., $\varepsilon_x^0 = \varepsilon_y^0 = \varepsilon^0$ and $\kappa_x = \kappa_y = \kappa$. So in the remainder of this section this equibiaxial deformation state is understood, and we refer to it simply by ε^0 and κ. While this is the case in the examples presented in this section, it is not the case when material anisotropy exists, the film is patterned in lines instead of as a blanket, or large deflections occur which result in geometric nonlinearity. These issues will be discussed briefly in later sections.

Single-Layer Plate. Consider a single-layer film with an internal biaxial mean stress $S = S_{11} = S_{22}$ and stress gradient $\beta = \beta_{11} = \beta_{22}$ that is subjected to a uniform temperature change T. This scenario models the deformation of a film after release from a substrate when subjected to a temperature change where before release the film had an internal stress S, and stress gradient β. In this case, after significant manipulation (35.11), with (35.12) and (35.13), yields:

$$\kappa = \frac{\beta}{M} \qquad (35.16)$$
$$\varepsilon^0 = -\frac{S}{M} + \alpha T , \qquad (35.17)$$

where $M = \frac{E}{1-\upsilon}$ is the biaxial modulus. As previously discussed, after release there is straining due to S and $\varepsilon^* = \alpha T$, and the plate bends with a biaxial curvature κ. The result of (35.16), and especially the simplification that results for beams (simply obtained by setting $\nu = 0$ and interpreting κ and ε^0 as the uniaxial curvature and midplane strain in the direction of the beam length) is commonly used in conjunction with measurements of the curvature after release to determine the stress gradient in a film.

Bilayer Plate. Another simple yet practically important special case is that of two isotropic layers of thicknesses t_1 and t_2, subject only to a uniform temperature change T. In this case, (35.11), with (35.12) and (35.13), yields:

$$\kappa = \frac{6\Delta\alpha T}{t_2} hm \left(\frac{1+h}{1+2hm\left(2+3h+2h^2\right)+h^4m^2}\right) \qquad (35.18)$$

$$\varepsilon^0 = T\left(\frac{\alpha_1 mh\left(1+3h+3h^2+mh^3\right)}{1+2hm\left(2+3h+2h^2\right)+h^4m^2} + \frac{\alpha_2\left[1+mh\left(3+3h+h^2\right)\right]}{1+2hm\left(2+3h+2h^2\right)+h^4m^2}\right) , \qquad (35.19)$$

where $h = \frac{t_1}{t_2}$, $m = \frac{M_1}{M_2}$, and $M_k = \frac{E_k}{1-\upsilon_k}$ (here $k = 1, 2$ corresponds to layers 1 and 2) and $\Delta\alpha = \alpha_2 - \alpha_1$.

It is often convenient to define a nondimensional curvature $\hat{\kappa}$:

$$\hat{\kappa} = \frac{\kappa\, t_2}{6\Delta\alpha T} \tag{35.20}$$

Equation (35.18) can then be written as

$$\hat{\kappa} = \frac{hm\,(1+h)}{1+2hm\left(2+3h+2h^2\right)+h^4m^2}\,. \tag{35.21}$$

Two limiting values of (35.21) are of interest. The first is the case where there is no elastic mismatch between the layers, i. e., $m = 1$. In this case (35.21) simplifies to:

$$\hat{\kappa} = \frac{h}{(1+h)^3}\,. \tag{35.22}$$

The other is the *thin-film limit* where $t_1 \ll t_2$. Expanding (35.18) in powers of h and retaining only the lowest order term recovers *Stoney's* [35.26] well-known result:

$$\begin{aligned}\kappa &= \frac{6\Delta\alpha T}{t_2}hm\\ \hat{\kappa} &= hm\,.\end{aligned} \tag{35.23}$$

In the thin film limit, (35.19) reduces to $\varepsilon^0 = \alpha_2 T$.

Figure 35.11 shows the normalized curvature $\hat{\kappa}$ as a function of the layer thickness ratio h for various ratios of the biaxial moduli, m. The curves from top to bottom are for $m = 10$, 0.66, and 0.1. $m = 10$ and 0.1 represent a stiff film on a compliant substrate and a compliant film on a stiff substrate, respectively. $m = 0.65$ represents a gold film on a polysilicon substrate. The solid curves are the exact results from (35.21) and the dashed curves are the Stoney's result for the thin-film limit, (35.23). In the latter case, $\hat{\kappa}$ varies linearly with h, with slope m. The results in Fig. 35.10 show how the range of applicability of Stoney's result depends on both h and m; the range decreases for increasing h and decreasing m. For most microstructure applications where the layer thicknesses are comparable ($h > 0.1$), Stoney's result is not accurate unless the film is much more compliant than the substrate. Thus physically, Stoney's result is accurate when the substrate stiffness controls the deformation, either because it is much thicker or has much higher elastic moduli than the film. Over the years many papers have discussed *corrections* to Stoney's result to increase the region of validity in terms of layer thicknesses and modulus mismatch. We emphasize, though, that (35.18) is valid for arbitrary layer thicknesses and modulus mismatch and is quite simple to use itself. We refer the reader interested in more details to [35.7, 18, 22].

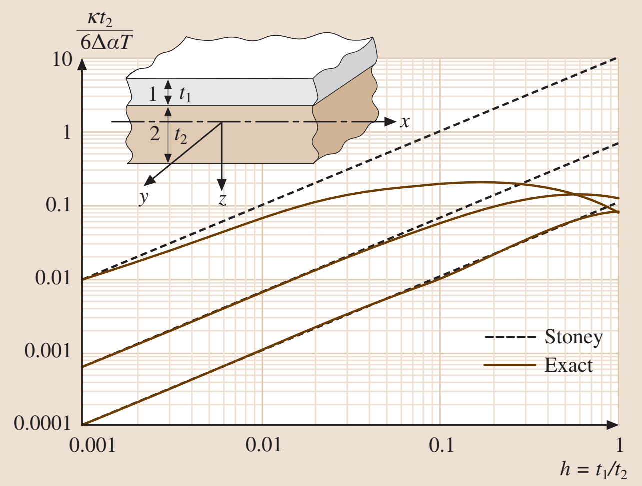

Fig. 35.11 Normalized curvature $\hat{\kappa}$ as a function of the film thickness ratio $h = t_1/t_2$ for three values of the biaxial modulus ratio, $m = M_1/M_2$. From *top* to *bottom*, the curves are for $m = 10$, 0.65 (gold/polysilicon), and 0.1. The *solid lines* are the exact solution, (35.21) and the *dashed lines* are Stoney's result in the thin-film limit, (35.23)

Equally important from a technological viewpoint is the equibiaxial stress distribution through the thickness $\sigma_x(z) = \sigma_y(z) = \sigma(z)$ obtained from (35.6):

$$\sigma(z)_k = M_k\left(\varepsilon^0 + z\kappa - \alpha_k T\right)\,, \tag{35.24}$$

where $k = 1, 2$ denotes the layer. In (35.24), κ and ε^0 are obtained from (35.18) and (35.19), respectively, and M_k and α_k are the values for the kth layer. In the thin-film limit the equibiaxial stress varies so little through the film thickness that it can be taken as constant, i. e., $\sigma_x = \sigma_y = \sigma$. This is the reason one often refers to the *stress* in a thin-film by a single number. In the thin film limit, the substrate (layer 2) thermal expansion dominates the deformation process, and so the film strains by an amount $\Delta\alpha T$, and the biaxial stress in the film is $\sigma_1 = M_1\Delta\alpha T$, or in terms of the curvature:

$$\sigma_1 = \frac{M_2}{6}\frac{t_2^2}{t_1}\kappa\,. \tag{35.25}$$

Note that, as written, the thermoelastic properties of the film do not enter (35.25). We will discuss the practical ramifications of this shortly.

Figure 35.12 shows the stress distribution through the thickness of both layers for a negative unit temperature change, (35.24), as a function of $h = t_1/t_2$ for

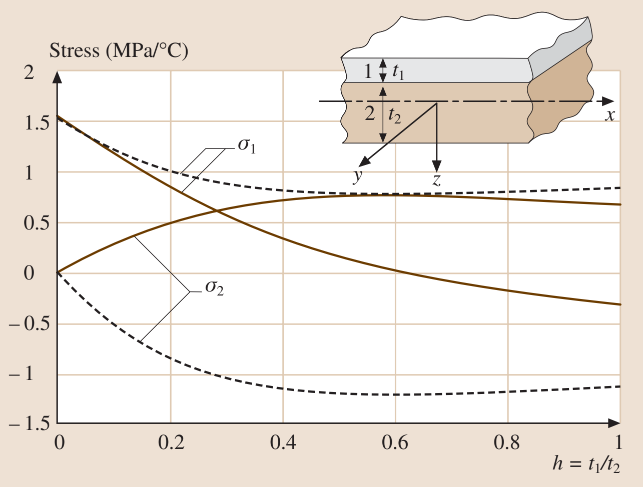

Fig. 35.12 Biaxial stress per unit temperature change, (35.24), for a gold/polysilicon bilayer ($m = 0.67$) for a 0.5-µm-thick gold film as a function of the polysilicon thickness $h = t_1/t_2$. Biaxial stresses are shown at four locations: the interface on the gold and polysilicon sides (*dashed lines*), and the surfaces of the gold and polysilicon (*solid lines*)

$m = 0.67$, and $t_1 = 0.5\,\mu\text{m}$, representative of a gold film on a silicon substrate. Equation (35.24) shows, and we have previously discussed, that the biaxial stress varies linearly through each material. Thus we can completely specify the stress distribution by plotting it at four points: the top of material 1, the interface in material 1, the interface in material 2, and the bottom of material 2. Referring to Fig. 35.8, these positions are $z = z_0$, z_1, and z_2. Because of the perfect bonding between the two films, the total strain must be continuous across the interface. The elastic moduli are different for the two films, though, and so the stresses will be discontinuous at the interface. Hence we interpret $z = z_1$ as a position just above, or just below, the interface, and use M_k and α_k in (35.24) as appropriate. As $h \to 0$ the stresses at the top and at the interface in the film (layer 1) converge and take on a maximum value given by $\sigma = M_1 \Delta\alpha T$. The stress in the substrate vanishes in this limit. As the substrate thickness increases (increasing h) the stress in the film decreases with the maximum value occurring at the interface. The stress in the substrate increases at the interface but decreases on the bottom surface. In general this results in a stress distribution in the substrate that is compressive on the bottom surface and tensile at the interface. Depending on h, the stresses in the film are either tensile throughout or tensile on the interface and compressive on the surface. The difference in stress between the interface and the surface is a measure of the stress gradient in each layer; it increases with increasing h. For different values of m, the behavior with varying h is similar to that in Fig. 35.12, with the stress in the film increasing with increasing m.

Equation (35.25) is the theoretical basis of what is perhaps the most common technique to measure thin-film stresses – the wafer curvature method [35.6]. It involves measuring the curvature of a thick (compared to the film thickness) wafer with and without a film, and then using the curvature difference between the two in (35.25) to determine the film stress. It is particularly convenient because the film stress is obtained from the measured curvature regardless of the source of the stress and the thermoelastic moduli of the film. It is necessary, though, to know accurately the thickness of the film and substrate, as well as the elastic properties (biaxial modulus, $E/(1-\nu)$) of the substrate. An appealing feature of the method is that the elastic properties of the film, typically difficult to measure themselves, do not need to be known. This is because the much thicker substrate controls the deformation response.

Three-Layer Symmetric Plate. Considerable simplification also results for the technologically important case of a thin-film multilayer subject to a uniform temperature change T that consists of three layers, but the top and bottom layer are identical in terms of thickness ($t_3 = t_1$) and material properties ($M_3 = M_1, \alpha_3 = \alpha_1$). In this case, from (35.9) $[B] = [0]$, and thus the plate does not bend, i. e., $\kappa = 0$. Equation (35.11), with (35.12) and (35.13), then yields:

$$\varepsilon^0 = \frac{(2\alpha_1 m h + \alpha_2)\,T}{2mh+1}\,. \tag{35.26}$$

In the thin-film limit $h \to 0$, $\varepsilon^0 = \alpha_2 T$.

The equibiaxial stress distribution is such that it does not vary through the thickness within each layer, i. e., $\sigma(z)_3 = \sigma(z)_1 = \sigma_1$, and $\sigma(z)_2 = \sigma_2$. From (35.6) with $\kappa = 0$ and ε^0 from (35.26):

$$\begin{aligned}\sigma_1 &= \frac{M_1 \Delta\alpha T}{2mh+1}\\ \sigma_2 &= \frac{-2mhM_2\Delta\alpha T}{2mh+1}\,.\end{aligned} \tag{35.27}$$

In the thin-film limit, the stresses of (35.27) reduce to $\sigma_1 = M_1 \Delta\alpha T$ and $\sigma_2 = -2M_1 h \Delta\alpha T$.

Figure 35.13 shows the uniform biaxial stress in each material for a negative unit temperature change as a function of $h = t_1/t_2$ for $m = 0.65$, and $t_1 = 0.5\,\mu\text{m}$, again representative of a gold film on a silicon substrate. As $h \to 0$, the stresses in the films (layers 1 and 3) are given by $\sigma_1 = M_1 \Delta\alpha T$, and that in the substrate (layer 2) approaches zero. The stress in the films is tensile and decreases with increasing h while that in the substrate is compressive, increasing in magnitude with increasing h.

Beams. Our analysis thus far has focused on multilayer plates where both in-plane dimensions far exceed the thickness and where the transformation strains, and thus the stresses, are equibiaxial. When one in-plane dimension, say x, is much larger than the other, the structure is often referred to as a beam. In the classical treatment of the flexure of beams, it is assumed that only one nonzero normal stress component exists, σ_x. This greatly simplifies the analysis for homogeneous beams, but the situation is not as straightforward for multilayers with biaxial transformation strains. From the results in the previous section, one can obtain results for a beam theory by simply setting $\nu_1 = \nu_2 = 0$ in the formulae developed for biaxial stresses, but these results should be used with caution as the transverse component of the moment that results from the equibiaxial transformation strains cannot be neglected as shown by three-dimensional finite element calculations [35.27, 28]. *Swanson* [35.29] describes how to apply the results from a multilayer plate theory to beams and specifically focuses on the effect of the beam width, giving easily usable results for narrow and wide beams. The main results are an effective structural stiffness $(EI)_{\text{eff}}$ that can be used to replace EI (E is Young's modulus and I is the cross-sectional moment of inertia) in existing beam theory results, such as those that will be presented in Sect. 35.2 for electromechanics analysis. Insight into the behavior can be seen in the results in Table 35.2, which show measurements (using interferometry as described by *Dunn* et al. [35.30]) and predictions based on (35.18) for plates and the simplified version for beams obtained by setting $\nu_1 = \nu_2 = 0$. In all calculations both the gold and polysilicon are modeled as linear thermoelastic with isotropic material properties. Input parameters to the finite element calculations are $E_2 = 163\,\text{GPa}$, $\nu_2 = 0.22$, $E_1 = 78\,\text{GPa}$, $\nu_1 = 0.42$ [35.31]. The thermal expansion coefficients of the materials were assumed to vary linearly with temperature, and values at 100(24) °C used are $\alpha_2 = 3.1(2.6) \times 10^{-6}\,/°\text{C}$, and $\alpha_1 = 14.6(14.2) \times 10^{-6}\,/°\text{C}$ [35.31]. Although some uncertainty exists in the values of these material properties for the gold and polysilicon films, we think they are sufficiently accurate for the purpose of modeling the observed phenomena. Young's modulus and Poisson's ratio of the polysilicon are in line with many measurements over many MUMPS runs [35.32] and agree adequately with bulk polycrystal averages of single-crystal elastic constants. Good agreement exists between the predictions using plate theory and the measurements, while predictions based on beam theory are significantly less accurate, demonstrating the need to use plate, rather than beam, theory.

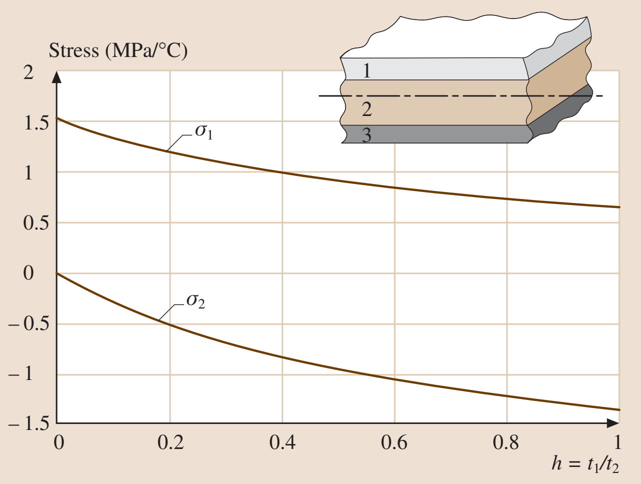

Fig. 35.13 Biaxial stress per unit temperature change for a gold/polysilicon/gold ($m = M_1/M_2 = 0.65$) three-layer film system. The calculations, (35.27), are for a 0.5-µm-thick gold film as a function of the polysilicon thickness $h = t_1/t_2$. Biaxial stresses are shown in the gold (*top curve*) and polysilicon (*bottom curve*)

Patterned Films. In microsystems applications, additional design freedom can be obtained by using multilayer films with patterned, rather than blanketed,

Table 35.2 Predicted and measured curvature per unit negative temperature change for gold/polysilicon microstructures as a function of the polysilicon thickness. In all cases the gold thickness is 0.5 µm. The microstructures are 300 µm long and 50 µm wide and cantilevered from one end

Polysilicon	$d\kappa/dT$ (m^{-1}/°C)		
thickness (µm)	Measured	Predicted (plate)	Predicted (beam)
1.5	−5.36	−5.56	−4.86
3.5	−1.42	−1.44	−1.16

films. In general one can use the finite element method with multilayer plate elements to model the deformation and stresses in a multilayer with arbitrary patterned layers. A useful practical example is a beam of width w_2 that is covered with a film in the pattern of a stripe of width w_1 as shown in Fig. 35.14. A simple modification of (35.18) yields the curvature along the length of the patterned beam as a function of the nondimensional linewidth, $c = w_1/w_2$:

$$\kappa_{\text{patterned}} = \frac{6\Delta\alpha T}{t_2} hcm \times \left(\frac{1+h}{1+2hcm\left(2+3h+2h^2\right)+h^4c^2m^2} \right) . \qquad (35.28)$$

Predictions from (35.28), normalized as $\eta = \kappa_{\text{patterned}}/T$, are shown in Fig. 35.14 for $m = 0.65$ (the gold/polysilicon bilayer) along with predictions from detailed finite element calculations and measurements over the entire range of $c = w_1/w_2$ [35.33]. Equation (35.28) accurately describes the effect of linewidth on curvature development and can be used to design bilayer beams.

Deflections. Thus far our discussion of deformation has been cast in terms of curvature. An important quantity for many applications where control of the deformation is important is the deflection of the multilayer. Formally this can be obtained by integrating the curvature-displacement relationship and applying the relevant boundary conditions. For example, assuming constant curvature, the tip deflection $w(x = L)$ for a cantilever beam of length L, is:

$$w(x = L) = \frac{\kappa L^2}{2} . \qquad (35.29)$$

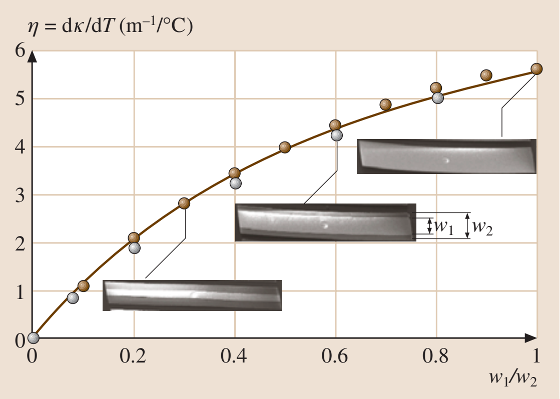

Fig. 35.14 Curvature per unit temperature change as a function of the ratio of gold to polysilicon film width w_1/w_2 for gold (0.5 µm thick)/polysilicon (1.5 µm thick) strip-patterned beams. *Grey circles* are finite element simulation results, the *solid line* is the analytical result of (35.28), and the *brown circles* are measurements

35.1.3 Nonlinear Geometry

In the previous section we noted that in general the deformation of a layered film microstructure subjected to transformation strains consists of three regimes but focused on the first, linear regime. Here we focus on the second and third regimes. To illustrate, consider the seemingly simple case of a plate with total thickness much less than the in-plane dimensions, composed of two isotropic layers with different material properties (elastic moduli and thermal expansion coefficients) subjected to a temperature change. Although we use a temperature change to demonstrate, the results hold for any type of transformation strains. In terms of the average curvature variation as a function of temperature change, three deformation regimes have been identified as illustrated in Fig. 35.15 [35.16–19, 22, 23]. The first regime, I, consists of a linear relation between the average curvature and temperature change where $\kappa_x = \kappa_y$, i. e., the average curvature is spherically symmetric. This deformation regime is characterized by both small transverse displacements and rotations, and so conventional thin-plate theory adequately describes the deformation. In Sect. 35.1.2 we outlined the analysis in this regime. The second regime, II, consists of a nonlinear relation between the average curvature and temperature, but again $\kappa_x = \kappa_y$. The behavior is due to *geometric nonlinearity* that results when the deflections become excessively large relative to the plate thickness, and they contribute significantly to the in-plane strains. It has been shown [35.15–19, 22, 23] that in these two regimes the symmetric deformation modes are stable. The second regime ends at the point when the deformation response bifurcates from a spherical to ellipsoidal deformation, i. e., $\kappa_x \neq \kappa_y$. At this point, the beginning of regime III, it becomes energetically favorable for the plate to assume the ellipsoidal shape because to retain the spherical deformation under an increasing temperature change requires increased mid-plane straining. After the bifurcation, the curvature in one direction increases while that perpendicular to it decreases; the plate tends toward a state of cylindrical curvature. This observation helps to explain the energetic argument, as unlimited cylindrical curvature can be obtained with no midplane straining, while spherical curvature cannot. This discussion has been cast in the context of linear material behavior. Additional deformation regimes result if material nonlinearity is present, for example, yielding [35.18].

Analysis

The understanding described above derives from a number of studies with different technological motivations, primarily structural laminated composites and thin films for microelectronics. Most of these studies are analytical [35.15–19, 22, 23, 34] and build upon the original work of Hyer [35.13, 14]. To illustrate the computational approach in a reasonably simple setting, here we consider a two-layer plate with layer thicknesses t_1 and t_2. Each layer is isotropic and characterized by the Young's modulus E_i, Poisson's ratio ν_i, and thermal expansion coefficient α_i $(i = 1, 2)$. We consider a square plate with side length L; a similar analysis can be carried out for circular [35.22] or other plate shapes. The plate is subject to a uniform temperature change, T. In other words, the stress-free strains $\{\varepsilon^*\}_k$ are due to thermal expansion as given by (35.14).

In order to compute the deformed shape when the plate is subject to a temperature change we use the approach of *Hyer* [35.14] as applied by *Masters* and *Salamon* [35.16]. The basic idea is to assume an admissible displacement field $w(x, y)$ in terms of unknown parameters (d_i) that are suitably chosen to be consistent with observed deformation modes. Values of the parameters d_i are then determined via a Ritz procedure so as to minimize the total potential energy of the system. Different choices of the assumed displacement field have been considered, and details of the procedures are given in the above references. Such analyses are sufficient to qualitatively, and in many cases quantitatively, explain the three regimes of deformation shown in Fig. 35.15. In fact, quite simple closed-form expressions result for special cases that provide illuminating descriptions of observed phenomena (see for example [35.22]). A disadvantage of the analytical approaches is that for simplicity a displacement field consistent with a spatially constant curvature deformation mode is usually chosen. Additionally, these formulations are useful for only simple plate shapes. While this may be adequate for the structures demonstrated here, it is not for more complex in-plane shapes, of either or both layers, that arise in microsystems applications. In this more general case the best approach is probably to use the finite element method for the approximate analysis.

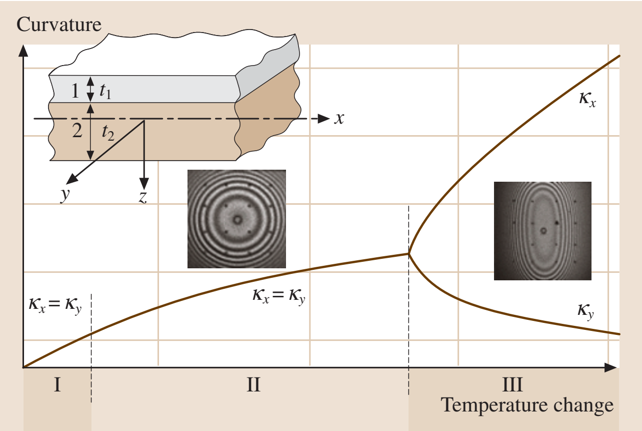

Fig. 35.15 General characteristics of the average curvature-versus-temperature change of a two-layer plate microstructure showing the three deformation regimes. The interferograms in the *inset* show measured displacement contours for gold/polysilicon plates in Regimes II and III

To proceed, we assume the transverse midplane displacement is of the form:

$$w^0(x, y) = d_1 x^2 + d_2 y^2 \,. \tag{35.30}$$

d_1 and d_2 are immediately recognized as one-half the curvature in the x- and y-directions, and the deformation is seen to be a constant curvature mode. The midplane displacements u^0 and v^0 are assumed to be described by third order polynomials in x and y, also with unknown constant coefficients d_3–d_8 to be determined:

$$\begin{aligned} u^0(x, y) &= d_3 x + \frac{d_5}{3} x^3 + d_7 x y^2 - \frac{d_1^2}{6} x^3 \\ v^0(x, y) &= d_4 y + \frac{d_6}{3} y^3 + d_8 x^2 y - \frac{d_2^2}{6} y^3 \,. \end{aligned} \tag{35.31}$$

The midplane strains are then computed from the nonlinear strain-displacement relations of the von Karman plate theory:

$$\begin{aligned} \varepsilon_x^0 &= \frac{\partial u^0}{\partial x} + \frac{1}{2}\left(\frac{\partial w^0}{\partial x}\right)^2 \\ \varepsilon_y^0 &= \frac{\partial v^0}{\partial y} + \frac{1}{2}\left(\frac{\partial w^0}{\partial y}\right)^2 \\ \gamma_{xy}^0 &= \frac{\partial u^0}{\partial y} + \frac{\partial v^0}{\partial x} + \left(\frac{\partial w^0}{\partial x}\right)\left(\frac{\partial w^0}{\partial y}\right) \,. \end{aligned} \tag{35.32}$$

The strains at any point through the thickness are then computed using the standard kinematic relations for thin plates, (35.2), and the stresses are computed from the strains using the conventional linear thermo-elastic constitutive relations for each layer, (35.6) with $\{\varepsilon^*\}_k = \{\alpha\}_k T$. The first term in (35.32) is recognized as the conventional small deformation, and the second term arises from the large transverse deflections w^0, which result in straining of the midplane. This is shown schematically in Fig. 35.16 for the strain component ε_x^0; similar results are obtained for ε_y^0 and γ_{xy}^0.

The potential energy density of each layer is computed from the stress and strain in each layer, and the total potential energy of the plate is then computed by integrating it over the volume of the plate. This yields an expression for the potential energy of the plate in terms of the unknown coefficients d_i:

$$U(d_i) = \int_{-L/2}^{L/2} \int_{-L/2}^{L/2} \int_{-t/2}^{t/2} \{\varepsilon\}^T \{\sigma\}_k \, \mathrm{d}z \, \mathrm{d}x \, \mathrm{d}y \,, \tag{35.33}$$

where the superscript T denotes the transpose of the 3×1 column vector. The coefficients d_i are determined by application of the principle of minimum potential energy:

$$\frac{\partial U(d_i)}{\partial d_i} = 0 \,. \tag{35.34}$$

This yields eight equations for the eight unknown d_i and thus solutions for the deformation response of Fig. 35.15. Complete details regarding this analysis can be found elsewhere [35.14, 16]; here we present only the pertinent results in what we hope is an accessible form. When the displacements are small, the solution of (35.34) and (35.35) recovers (35.18) and (35.19).

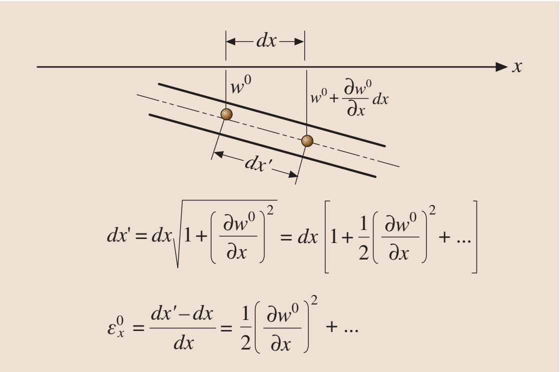
dx
x
w^0
$w^0 + \frac{\partial w^0}{\partial x} dx$
dx'
$dx' = dx\sqrt{1+\left(\frac{\partial w^0}{\partial x}\right)^2} = dx\left[1+\frac{1}{2}\left(\frac{\partial w^0}{\partial x}\right)^2 + \ldots\right]$
$\varepsilon_x^0 = \frac{dx'-dx}{dx} = \frac{1}{2}\left(\frac{\partial w^0}{\partial x}\right)^2 + \ldots$

Fig. 35.16 Illustration of the development of midplane strain due to bending when nonlinear geometry effects arise

In the nonlinear but symmetric deformation regime II, the relationship between both the curvature and midplane strain, and the thermal expansion mismatch is too complex to present here. This is also the case in the nonlinear regime III after bifurcation. An important difference between the linear and nonlinear response, though, is that the nonlinear response depends on the plate size. The critical curvature at which bifurcation occurs, though, can be obtained explicitly and is given by:

$$\kappa_{\mathrm{cr}} = \frac{12\sqrt{2}}{L^2}\sqrt{\frac{6+A_{12}/A_{11}}{1+A_{12}/A_{11}}}\,\frac{\sqrt{A_{66}D_{66}-B_{66}^2}}{A_{66}^2} \tag{35.35}$$

for a square plate of side length t, where A_{ij}, B_{ij}, and D_{ij} are the composite moduli of (35.8). In the simplified case where there is no elastic mismatch between the layers and the elastic, response can be expressed in terms of the Young's modulus E and Poisson's ratio ν, (35.35) reduces to:

$$\kappa_{\mathrm{cr}} = 2\sqrt{6}\sqrt{\frac{6+\nu}{1+\nu}}\,\frac{t}{L^2} \,, \tag{35.36}$$

where $t = t_1 + t_2$. These results have also been obtained by *Finot* and *Suresh* [35.18]; similar results exist for a circular plate [35.22, 30].

The analysis just discussed is made tractable by assuming a displacement field consistent with a constant curvature. In the nonlinear case, this assumption is questionable. While the assumed displacement field used in the Ritz procedure could be modified to incorporate the dependence of curvature on position, perhaps the simplest approach to tackle these more general problems is to use the finite element method to solve the geometrically nonlinear equations over an arbitrary spatial domain. This is also the most viable approach for more complicated in-plane geometries, including patterned films. Complete details regarding the use of the finite element method to carry out such calculations can be found in *Dunn* et al. [35.30]. Here we will simply show the result of some of these calculations. The input parameters used are the same as those given in Sect. 35.1.2.

Deformation Behavior

In order to illustrate the nature of the deformation phenomena in the nonlinear regimes, Fig. 35.17 shows contour plots of measured (using an interferometric microscope as described by *Dunn* et al. [35.30]) and predicted displacement fields $w(x, y)$ for square plates of four plate sizes that have been cooled from about 100 °C where they are flat to room temperature. Due to the thermal expansion mismatch between the polysilicon and gold, the $L = 150\,\mu\text{m}$ samples deform in a spherically symmetric manner; contours of constant transverse displacement $w(x, y)$ are nearly circles. This is also the case as the size increases to $L = 200$ and $250\,\mu\text{m}$, although the displacements increase as the plate size increases. At $L = 300\,\mu\text{m}$, though, the transverse displacement contours are not circular but elliptical, indicating that the deformation is no longer spherically symmetric. Thus when subjected to the same temperature change, both the magnitude and deformation mode depend on the plate size. As the in-plane dimension of the plate increases with the thickness held constant, the deformation mode changes from one of spherical symmetry to one more like cylindrical symmetry.

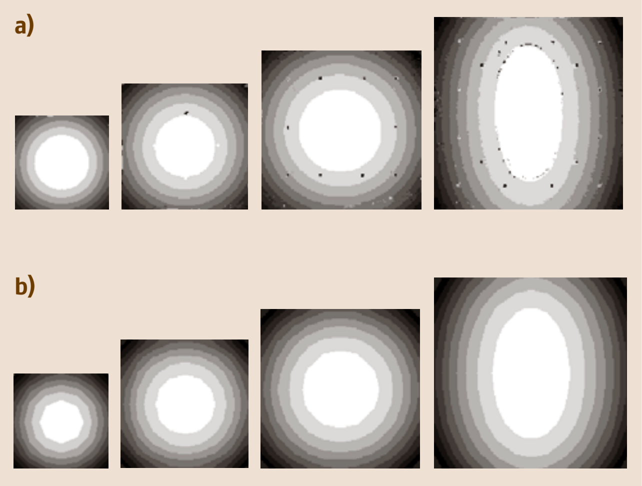

Fig. 35.17a,b Contour plots of the **(a)** measured, and **(b)** predicted transverse displacements $w(x, y)$ at room temperature following cooling from 100 °C for the four gold/polysilicon square plates: $L = 150$, 200, 250, and $300\,\mu\text{m}$ from *left* to *right*. Each contour band represents a displacement of 0.23, 0.35, 0.45, and $0.6\,\mu\text{m}$ for the $L = 150$, 200, 250, and $300\,\mu\text{m}$ plates, respectively

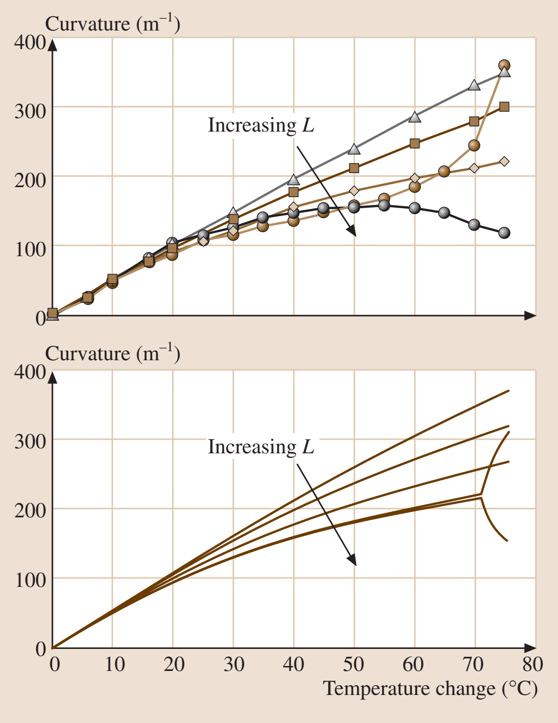

Fig. 35.18 Average measured (*top*) and predicted (*bottom*) curvature as a function of temperature change upon cooling from 100 °C to room temperature. The curves from *top* to *bottom* are for the $L = 150$, 200, 250, and $300\,\mu\text{m}$ structures, respectively

A complete picture of the deformation as a function of plate size and temperature change could be obtained by acquiring full-field displacements like those in Fig. 35.17 as a function of temperature. But a reasonable picture of the deformation behavior can be obtained in much simpler form by considering the *average* curvature in the x- and y-directions as a function of temperature. This is shown in Fig. 35.18, in which the average curvature in the x- and y-directions is plotted as a function of the magnitude of the temperature change during cooling. The temperature change is actually negative according to our convention, but its magnitude is plotted for convenience. The average curvatures are determined from the measured and computed $w(x, y)$ by averaging $\kappa_x = -\partial^2 w(x, 0)/\partial x^2$ and $\kappa_y = -\partial^2 w(0, y)/\partial y^2$ along the paths $y = 0$ and $x = 0$, respectively [35.30]. The x- and y-directions are taken to be aligned with the principal curvatures after bifurcation. The use of the average curvature as a measure of the plate deformation seems appropriate if the curvature is, or is close to, spatially uniform. This aspect will not be taken up in detail, but we refer the interested reader to [35.22, 30] for details. It is apparent from both the measurements and predictions that in regime I, the curvature-temperature response is independent of plate size and shape. In regime II, though, there is a strong dependence on plate size, and this is also the case in regime III. The major discrepancy between the measurements and predictions is the sharpness of the bifurcation for the $L = 300\,\mu m$ plate; it is quite sharp in the predictions but much more gradual in the measurements. To understand this we note that the source of the bifurcation is an *imperfection* of some sort that breaks the ideal symmetry, and in general it is difficult to accurately model the detailed imperfection.

Figure 35.19 demonstrates the connection between the thermomechanical loading (the temperature change), the geometry (plate size, L), and the boundaries between the three deformation regimes for gold/polysilicon plate microstructures. Specifically, it shows the temperature change necessary to initiate nonlinear effects (the transition between regions I and II), and bifurcation (the transition between regions II and III) as a function of polysilicon thickness when the gold film thickness is kept constant at $0.5\,\mu m$. Despite the fact that the constant curvature approximation becomes questionable for larger plate sizes, (35.35) is a good approximation as seen by the agreement with the finite element calculations, at least for the elastic mismatch and plate sizes considered here. In fact, although not shown, the simplified result of (35.36) for no elastic mismatch is in reasonable agreement with the finite element and the complete analytical results. The measurement for $L = 300\,\mu m$ is accurately described by both the analytical and finite element results.

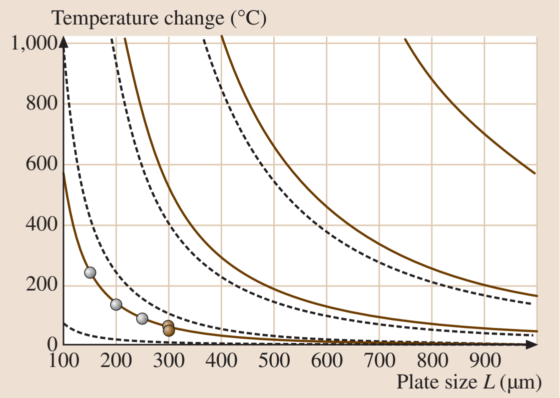

Fig. 35.19 Temperature change required for the initiation of nonlinear geometry effects (*dotted lines*) and critical temperature change for bifurcation (*solid lines*) as a function of square plate size and the thickness of polysilicon for the gold/polysilicon microstructures. Lines in each set from *left* to *right* represent $1.5\,\mu m$, $3.5\,\mu m$, $5.5\,\mu m$, and $8.5\,\mu m$ thick polysilicon, respectively. *Grey circles* are finite element calculations, and the *brown circle* is a measurement

In the analytical treatments discussed previously, it is assumed that the curvature is spatially uniform. The power of finite element calculations is that this requirement is relaxed and the spatial variation of the curvature can be studied theoretically. Full-field interferometry measurements allow one to study this experimentally as well. Briefly we discuss the resulting behavior for gold/polysilicon plate microstructures but refer the interested reader to [35.22, 30] for more details. In the linear regime, measurements and predictions both show that curvature is essentially uniform across the plate. In the nonlinear regime, though, the curvature varies appreciably with position, increasing by about a factor of two from the center to the periphery of the plate. The spatial nonuniformity of the curvature increases as the plate size increases. The spatial variation of the curvature raises concern regarding the suitableness of an analysis based on constant curvature. As mentioned previously, such an analysis may be adequate to describe the general deformation behavior but not for finer details.

35.1.4 Nonlinear Material Behavior

So far we have discussed in some detail the linear thermoelastic response of multilayer films and to a lesser degree, the effects of geometric nonlinearity. An additional complication is material nonlinearity. It can arise in numerous forms including plasticity, creep, stress relaxation, and evolution of the material microstructure (densification, grain growth, defect annihilation, etc.) during thermomechanical loading. In this section we briefly discuss nonlinear material behavior of multilayer films, focusing on the phenomena most relevant for the realization of reliable devices. The discussion in this section is concerned primarily with multilayer films where one of the films is a metal.

Nonlinear Material Behavior During Initial Thermal Cycles

As deposited, the material microstructure of a metal film is generally not stable. As the film is heated, microstructural evolution leads to the development of internal stress, which then can lead to changes in the curvature of the multilayer film. The general behavior is common to films made of many different metals, such as aluminum, copper, nickel, and gold. It has been studied extensively in the context of microelectronics applications (see for example [35.6, 35, 36]) where the substrate

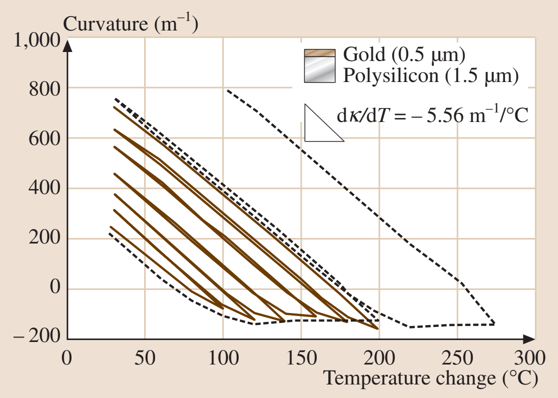

Fig. 35.20 Curvature versus temperature for gold (0.5-µm-thick)/polysilicon (1.5-µm-thick) 300 µm × 50 µm beam microstructures. The *solid lines* are tests with six cycles from room temperature to 200 °C with the maximum temperature in each successive cycle increasing by 20 °C. The *dashed lines* are tests with two cycles: the first from room temperature to 200 °C and then back to 30 °C, and the second to 275 °C and then back to 30 °C

is much thicker than the film and the thin-film limit results, which simplifies the interpretation of the results.

Figure 35.20 shows the general behavior of the curvature vs. temperature for a bilayer film consisting of a 0.5 µm gold film that was evaporated onto a 1.5 µm polysilicon film. During the initial heating the deformation is thermoelastic, characterized by a decrease in curvature with a constant $d\kappa/dT$, until the temperature reaches about 70 °C (the theoretical thermoelastic slope from (35.18) is shown on the figure). Between 70 °C and 100 °C, the curvature decreases at a much lower rate; in fact the curve flattens, i. e., $d\kappa/dT \approx 0$. Interestingly, between room temperature and 100 °C, the curvature changes from positive to negative, our convention being that a positive curvature means a beam with the gold on top is curved upward. The strong departure of the curvature from the thermoelastic behavior results because the microstructure of the evaporated film is not stable in the as-deposited condition. The temperature increase during the first cycle promotes microstructural changes in the film. These result in tensile straining of the film, which competes with the thermoelastic deformation (see for example [35.6, 36]). Upon cooling from 100 °C, the response is again thermoelastic, this time throughout the entire cooling process. Although it does not occur here, if the temperature change is large enough, yielding of the metal film can occur upon cooling. The curvature upon return to room temperature is greater than that initially at room temperature. It is important to understand this behavior because many post-fabrication and packaging processes expose a microstructure to elevated temperature excursions. During these, the film is susceptible to this nonlinear deformation. In the tests in Fig. 35.20, the material has not been taken to a temperature high enough to stabilize fully the gold microstructure in the first cycle. As a result, subsequent cycles going to successively higher temperatures appear to have the effect of continuing to stabilize the microstructure. The response during the second cycle is similar to the first with the exception that the temperature change required to initiate nonlinear behavior is increased. Indeed, over the range of temperature cycles studied, each subsequent cycle is similar, exhibiting thermoelastic response upon heating until a point where nonlinear κ-T behavior commences. The cooling process is again thermoelastic with the subsequent room temperature curvature increasing with each cycle. Also shown in Fig. 35.20 is a result for a second test on a nominally identical microstructure with only two cycles: the first from room temperature to 200 °C and back to 30 °C, and the second to 275 °C and back to 30 °C. The results of the first cycle to 200 °C neatly envelope the results of the set of tests with cycles up to 200 °C. This shows that although there is a path dependence of the deformation behavior, it is controlled by the maximum temperature reached during cycling, independent of how many increments are carried out to reach that temperature. The second cycle to 275 °C shows the same behavior observed in all cycles. The results in Fig. 35.20 have been obtained at a constant heating/cooling rate. In terms of the stabilization of the material microstructure that results in nonlinear material behavior, the rate, along with the time held at the elevated temperature, plays a significant role.

Figure 35.20 suggests that a simple description of the evolution of curvature with temperature can be obtained with parameters as shown in Fig. 35.21: the as-released curvature κ_r (and associated temperature, T_r), a critical temperature at which inelastic mechanisms are activated, T_{cr} (and the associated curvature, κ_{cr}), the slope of the inelastic κ-T behavior (assuming simple linear behavior) for $T > T_{cr}$, ξ, and the thermoelastic slope $\eta = d\kappa/dT$. The as-released curvature is determined primarily by the intrinsic stress development upon film deposition, cooling to room temperature, and release. These also influence T_{cr}. The parameter ξ is physically related to the curvature per unit temperature change due to inelastic mechanism(s) activated beyond T_{cr}. The situation in Fig. 35.20 is much simplified in that $\xi = 0$.

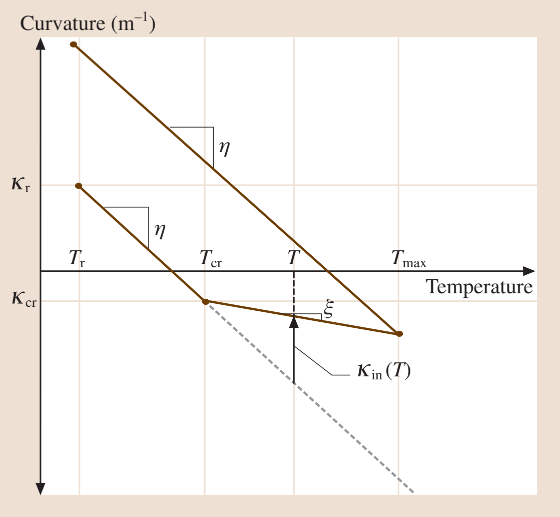

Fig. 35.21 Schematic diagram showing the parameters in the simple model that describes the curvature-versus-temperature behavior of the bilayer beam during thermal cycling after release

With knowledge of these parameters, one can compute the curvature following a defined thermal loading history. The qualitative behavior in Fig. 35.21 also results for a substrate with patterned lines (Sect. 35.1.2), and the same formalism can be used with κ_r, η, and ξ depending on linewidth [35.33]. For example, (35.28) gives the dependence of η on linewidth.

Plasticity

Plasticity in thin-film multilayers has been studied in depth for thin metal films on thick substrates (see, for example [35.6, 37, 38]), but little work has been directed toward freestanding multilayers. This is probably because, at least when subject to thermal loading, the stresses are typically smaller in freestanding multilayer films than in systems with thin films on thick substrates (see Fig. 35.12). Many interesting open issues exist with regard to the general issue of plasticity in microstructures including the understanding of size effects and strain gradient effects (see for example [35.39, 40]).

Creep and Stress Relaxation

For many microsystems applications, it is important to control accurately the deformation of thin-film microstructures over a significant period of time in order to meet performance and reliability requirements. This is especially important for microstructures subjected to thermal loading and/or operated at elevated temperatures. When one or more of the layers consists of a metal or polymer film, creep and stress relaxation in the film can significantly influence deformation and compromise device performance, and so their effects must be carefully considered. For example, *Miller* et al. [35.41] designed and fabricated microrelay switch arrays for RF communications applications using prestressed gold/polysilicon bimaterial beams as electrostatically actuatable switches. They observed a change in the switch shape and position over time and attributed it to stress relaxation in the gold. *Vickers-Kirby* et al. [35.42] report that creep in gold and nickel cantilever beams leads to voltage drops in micromachined tunneling accelerometers over time.

Creep and stress relaxation phenomena have been investigated in some detail for thin film on thick substrate systems, motivated primarily by microelectronics applications (see for example [35.11, 35, 43–46]). These studies have focused on measuring, understanding, and modeling the stress-versus-temperature behavior that occurs when a metal film-substrate system is heated from room temperature, then cooled back to room temperature, over one or a few cycles. The stress-temperature curves in the experimental works are typically obtained using the wafer curvature method that is now widely used to measure stress in thin films. During a typical test thermoelastic and inelastic mechanisms contribute to produce a complex nonlinear stress-temperature curve like that in Fig. 35.20. In most of these studies, the stress-temperature response is studied using wafer curvature measurements at a fixed heating or cooling rate so that rate dependent and independent phenomena are coupled in the response. Stress relaxation during an isothermal hold has been studied by [35.35, 43, 45, 47], among others. These studies all show that significant stress relaxation can occur over periods of only a few hours at modest temperatures of only about 100 °C for many metal thin-film systems. Due to the thin-film/thick-substrate system, the stress in the metal films is quite high, on the order of hundreds of MPa. Models incorporating power law creep of the metal film were successfully able to describe the observed response qualitatively and, to a large degree, quantitatively. The latter, however, sometimes required modification of the power-law exponent. *Shen* and *Suresh* [35.35] showed experimentally that a thin passivation layer on a metal film on thick silicon substrates can significantly reduce creep and stress relaxation. *Thouless* et al. [35.45] also studied the effect of a thin passivation layer on the stress relaxation of thin films with a focus on the deformation mechanisms. Their experimental results suggest that the presence of a passivation layer on the surface of a film can have a substantial effect on relaxation rates likely by suppressing mechanisms associated with diffusion and dislocation motion.

Thermomechanical Fatigue

It is also important to understand the development of damage and its effect on deformation during cyclic thermomechanical loading. To date, only very limited attention has been directed toward this issue. *Zhang* and *Dunn* [35.33] showed that if gold/polysilicon multilayers were cycled (Fig. 35.20) to an elevated temperature and then cooled to room temperature, they followed the thermoelastic path on subsequent cycling to a temperature below the maximum reached during the initial thermal cycle. Presumably over this range of temperature and time, the gold microstructure has been stabilized by the first cycle, and the polysilicon microstructure is not changing. This was confirmed for only a few, perhaps ten, cycles. *Gall* et al. [35.48] have shown that for thousands of cycles, though, a gradual shift of the thermoelastic curve downward is observed, possibly due to creep and stress relaxation in the gold, although the thermoelastic slope is maintained.

35.1.5 Other Issues

There are a number of additional issues regarding the thermomechanical behavior of thin-film microstructures that warrant discussion. Due to space limitations, however, we will only mention them and refer the reader to appropriate references.

In addition to the deformation itself, the stiffness characteristics of a microstructure, such as for example a beam or plate, can depend significantly on the stress in the beam. *Senturia* [35.49] describes the phenomena and discusses implications for microstructure design. This includes tensile stresses that tend to stiffen beams and plates as well as compressive stresses that can lead to buckling.

An important strategy has emerged to characterize the effects of processing on the residual stress state and thermomechanical behavior of films: the use of on-chip test structures to extract material properties including film stresses. Many test structures, both passive and active, have been developed to measure residual stresses (tensile and compressive), stress gradients, elastic moduli, strength, toughness, thermal expansion, and many others. Nice surveys of many test structures are given by Masters et al. [35.12] and *de Boer* et al. [35.50] where further references to more extensive details can be found.

35.2 Electromechanics of Thin-Film Microstructures

Electromechanics, the coupling of electrostatics and mechanics, is perhaps the most common of the energy couplings used in microsystems technology for sensing and actuation. In this section we summarize the contemporary application of electromechanics, the analysis of fundamental electromechanics problems, and the practical issues that arise in the design of an electromechanical microsystem.

35.2.1 Applications of Electromechanics

Our reference to electromechanics will be limited to the coupling of electrostatics and the mechanics of constrained elastic media such as a cantilever beam. In this case, the electrical energy is transformed into strain energy of the elastic media but with typically negligible energy dissipated due to material damping. An alternative application of electromechanics is the coupling of electrostatics with rigid body mechanics. In this case the mechanical structure is not constrained by a compliant support to the substrate but by a kinematic constraint, which is a pin joint between multiple structural layers or anchor between the substrate and a structural element. An example of these rigid body constraints is the anchor in the electrostatic micromotors. In this case, the electrostatic energy is dissipated by friction (Coulomb type friction between structural elements and pin joints, other structural elements, and/or the substrate) and is stored in the kinetic energy of the rigid body. In this case, the energy stored as strain energy in the structural element is by design negligible. In both electromechanics cases, the energy is dissipated by viscous forces under dynamic conditions. Finally the coupling of electrostatics and mechanics can more broadly; consider the coupling of electrostatics and fluids or electro-thermo-mechanics.

Microsystems have enabled many different commercial applications of electromechanics. These applications have included pressure sensors, accelerometers, gyros, resonators, micropumps, optical mirrors, optical shutters/VOAs, and DC and RF switches. In these examples, electrostatic forces are produced between a stationary electrode and a movable electrode, which is often the proof mass for an accelerometer or the diaphragm for a pump. The electrostatic forces are imparted by a voltage difference between the movable and stationary electrode. The electromechanics is typically implemented in two ways, with the first characterized by a stable, analog behavior, and the second characterized by an unstable, digital behavior. The stable behavior is implemented as a self-test capability for accelerometers such that a specific, stable displacement is produced with the application of the voltage. The unstable behavior is implemented with a switch so that when the instability point (snap-in or pull-in) is reached, the switch will instantaneously close.

Electrostatic actuation has been used for switches and microrelays [35.51–53], accelerometers [35.54,55], micropumps [35.56, 57], and micromirror [35.58–65]. The specific case of switches and relays involves multiple regions of electromechanical behavior and represents one of the richer electromechanics design problems. The general behavior is shown in Fig. 35.22. The first region is the stable actuation region. In this region, the switch is actuated to an instability point in the voltage displacement curve called the pull-in or snap-in voltage. Up to the instability, the displacement increases monotonically until the pull-in voltage is reached. At the pull-in voltage, the switch instantaneously closes the gap because of the instability and establishes contact with the substrate. In the case of the switch, this means the closure of a contact. The electromechanics problem continues with a new set of boundary conditions in the next region because the switch is now "simply" constrained at the contact and fixed at its anchors to the substrate. In Fig. 35.22, the switches tip displacement is shown, which is the reason the displacement is unchanged as the voltage increases. Region 2 is then characterized by a constant displacement for increasing voltage. The shape of the beam continues to change as the voltage is increased to the point that the electrodes could short if necessary precautions are not taken.

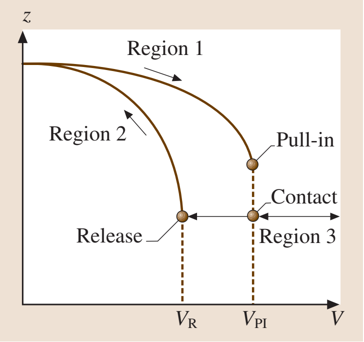

Fig. 35.22 Characteristic electromechanical hysteresis curve showing pull-in, contact, and release voltages

If the first consideration of the switch design is pull-in voltage, the second consideration of the switch design is the development of contact force, which is directly related to the contact resistance. The electromechanical analysis is continued beyond the pull-in voltage once contact is established to develop an increasing contact force. The electrostatic force is distributed between the anchor and the contacts, assuming other portions of the switch are not in contact. The electrostatic force increases as the square of the voltage difference, so the force acting on the contact will increase in proportion to this. Of course other physical limitations must be considered such as gas breakdown or a secondary pull-in that cause shorting between electrodes. The contact areas of the switch that are not the contacts should be minimized, so this does not detract from the contact force.

A third design consideration is the overdrive capability, defined as the maximum voltage the switch control electrodes can support before they touch and short. In some cases, the electrodes may be isolated from each other by a thin dielectric film, whose implications will be discussed in a later section. As suggested in the later section, the insulation material can be patterned to reduce the contacting area between the electrodes and the insulating material, in order to reduce charge accumulation. With the patterned isolation, the switch could be driven by such a large voltage to short the uninsulated regions. This failure mode could be evaluated by the same electromechanics analysis.

A fourth design issue would be the switch's release voltage, which is the voltage when the contact force reaches zero and the switch opens. This is shown as the beginning of region 3, where the deformation is reduced as the voltage is decreased. The release voltage is lower than the pull-in voltage because the electrostatic forces are much greater for the reduced gap than compared to the larger gap just before the pull-in instability. A fifth design consideration is the switch's self-actuation voltage, which is that which occurs across open contact and unintentionally causes the switch to close. By design, the switch's self-actuation voltage should be maximized in the electromechanical analysis. A final design issue is the determination of breakdown limits across the switch's control electrodes. This is not limited to the voltage supported across the equilibrium gap of the control electrodes and needs to be considered for different cases. The breakdown will be described by the Paschen curve or by vacuum breakdown proportional to the gap.

35.2.2 Electromechanics Analysis

An electromechanics analysis requires modeling of the structural domain, the electrostatic domain, and the coupling between the electrostatic fields and the structural elements. It has been approached at three levels of complexity. The simplest approach is the development of lumped parameter models, which have provided the basic understanding of instability, the impact of nonlinear springs, and the pull-in voltage-gap relationship. The lumped parameter model is comprised of the parallel plate capacitor and a spring (linear or nonlinear). One plate is constrained in space, by a substrate for example. The second plate is elastically constrained to one degree-of-freedom, typically, of motion, which is usually closing the "air" gap. The spring element represents the effective spring stiffness K of the structural element that could be a beam, a plate, a membrane. The parallel plate capacitor represents the electrostatic coupling between the fixed and moving elements.

Increasing in complexity, the second level of analysis considers continuum structural elements with a distributed electrostatic load. For example, the solution for an isotropic homogeneous beam would satisfy the Euler–Bernoulli beam equation:

$$EI\frac{\mathrm{d}^4 w}{\mathrm{d}x^4} = q(x)\ , \qquad (35.37)$$

where E is Young's modulus, I is the structure's moment of inertia, $w(x)$ is the transverse deflection of the beam, x is the coordinate along the length of the beam, and $q(x)$ represents the distributed electrostatic load. For a multilayer film, an effective EI can be used, as mentioned in Sect. 35.1.2. The third level of analysis is to model the fully coupled problem as continuum solids. The approach has used different implementations of finite difference (FD), finite element (FE), and boundary element (BE). A typical example would be to implement a finite element code to solve the continuum mechanics problem and a boundary element code to solve the electrostatics problem. Another implementation would use finite element methods to solve both the mechanics and the electrostatics problems. In each case, a similar algorithm is followed for developing a self-consistent solution to the coupled electromechanics analysis.

The coupled electromechanics problem can be described simply as a pair of conductors, where one conductor has a fixed constraint and the second conductor is attached to an elastic structure, such as a beam, a plate, a membrane, or an elastic substrate. The conductors support a voltage difference between them that causes charge to be induced on the surface of the conductors. This charge produces an electrostatic force that attracts the two conductors together. As the elastically supported conductor deforms toward the fixed conductor, the charge redistributes and modifies the field, thereby the force distribution. The process continues until a redistribution of charge is no longer necessary to maintain equilibrium. The typical algorithm begins by calculating the displacements of the undeformed geometry, which is updated before proceeding to the electrostatic analysis. In the electrostatic analysis, the surface charge density distribution is calculated and then used to calculate the electrostatic forces. The electrostatic force distribution is the important parameter because it is passed to the mechanical analysis as the loading condition on the conductors in the undeformed state to deform the structure until a state of equilibrium is reached.

Electromechanical Systems Energy Balance

A general analytical methodology can be applied to many different structures including parallel plate actuators with linear or nonlinear springs and plates supported by torsional suspensions [35.66]. They consider a single input system with charge control or voltage control. To begin, a generalized actuator is considered that consists of two conductors. One conductor is fixed. The second conductor is constrained by an elastic support.

In the charge controlled actuator, the total energy for the generalized actuator is written in terms of the mechanical energy as a function of the generalized coordinate χ and of the electrical energy stored in the capacitance of the actuator. The total energy is:

$$U_\mathrm{T}(\chi, Q) = U_\mathrm{M}(\chi) + \frac{Q^2}{2C(\chi)}\ , \qquad (35.38)$$

where U_T is the total energy, U_M is the mechanical energy, and the electrical energy is determined by the charge Q, and the capacitance $C(\chi)$. The first derivative of the total energy with respect to the generalized coordinate set equal to zero determines the equilibrium. The second derivative of the total energy with respect to the generalized coordinate set equal to zero determines the stability. The first and second derivatives are determined and then combined to determine what is referred to as the charge-controlled pull-in equation:

$$\frac{\partial U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi}\frac{\partial^2}{\partial \chi^2}\left(\frac{1}{C(\chi_{\mathrm{PI}})}\right) - \frac{\partial^2 U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi^2}\frac{\partial}{\partial \chi}\left(\frac{1}{C(\chi_{\mathrm{PI}})}\right) = 0 \tag{35.39}$$

and the pull-in charge Q_{PI}:

$$Q_{\mathrm{PI}} = \left(\frac{-2\dfrac{\partial U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi}}{\dfrac{\partial C(\chi_{\mathrm{PI}})}{\partial \chi}}\right)^{1/2} . \tag{35.40}$$

In the voltage controlled case, the total co-energy for the generalized actuator is used because the capacitor is nonlinear, and the actuator is voltage controlled. The co-energy is written in terms of the capacitance between the generalized conductors, the voltage across the conductors, and the mechanical energy. The total co-energy is:

$$U_{\mathrm{T}}^*(\chi, Q) = \frac{1}{2}C(\chi)V^2 - U_{\mathrm{M}}(\chi) . \tag{35.41}$$

The relationship between energy and co-energy is shown in Fig. 35.23.

In a similar process to the charge-controlled system, the voltage-controlled pull-in equation is determined by the first derivative of the total co-energy with respect to the generalized coordinate and by the second derivative of the total co-energy with respect to the generalized coordinate. The equations for the first and second derivative are combined to determine the voltage-controlled pull-in equation

$$\frac{\partial U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi}\frac{\partial^2 C(\chi_{\mathrm{PI}})}{\partial \chi^2} - \frac{\partial^2 U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi^2}\frac{\partial C(\chi_{\mathrm{PI}})}{\partial \chi} = 0 \tag{35.42}$$

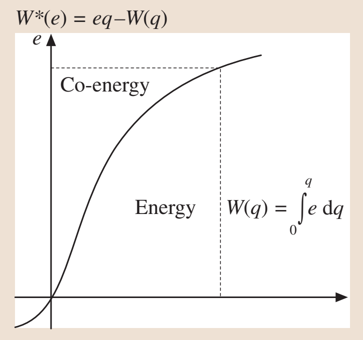

Fig. 35.23 Effort-Displacement curve showing the relationship between energy and co-energy

The pull-in voltage V_{PI} is:

$$V_{\mathrm{PI}} = \left(\frac{2\dfrac{\partial U_{\mathrm{M}}(\chi_{\mathrm{PI}})}{\partial \chi}}{\dfrac{\partial C(\chi_{\mathrm{PI}})}{\partial \chi}}\right)^{1/2} . \tag{35.43}$$

The application of this approach, to either the voltage-controlled or the charge-controlled case involves determining the mechanical energy and the capacitance. From the mechanical energy and the capacitance, the pull-in parameters and voltage are determined.

Lagrangian Approach to Electromechanics

Li and *Aluru* [35.67] and *Aluru* and *White* [35.68] have proposed a Lagrangian approach for both the electrostatic analysis and the mechanical analysis. Their algorithm begins by calculating the structural displacements for the undeformed geometry. They proceed to the electrostatic analysis to calculate the charge distribution on the conductors in the undeformed state. The charge distribution is used to calculate the electrostatic forces on the undeformed geometry until the system reaches equilibrium. The first advantage of their approach is that the algorithm does not require the structural geometry to be updated. A second advantage is the elimination of the integration error, wich occurs because flat panels are used in other approaches to approximate curved surfaces. The overall algorithm is shortened because the interpolation functions do not have to be updated as the structure shape changes.

35.2.3 Electromechanics – Parallel-Plate Capacitor

The classic element for understanding electromechanics is the simple parallel-plate capacitor model with one plate fixed and the second, moveable, plate, suspended by a spring. This basic electromechanics element is shown in Fig. 35.24. The capacitor is described by an area, A; an initial gap, g_0; and a dielectric constant of ε in the air gap (where ε is the product of the relative dielectric of the media filling the gap, ε_{r}, and the permittivity of free space, $\varepsilon_0 = 8.854 \times 10^{-12}\,\mathrm{F/m}$). The moveable plate is shown suspended by the spring, with spring constant K, and a single degree-of-freedom represented by the coordinate z. In this case, the equilibrium position is represented by having no charge on the plates (i. e. no electrostatic force) and no displacement of the spring (i. e. no spring force). It should be noted that as the spring length grows by increasing z, the gap decreases. The behavior of this simple model is described in detail in many sources including [35.49, 66].

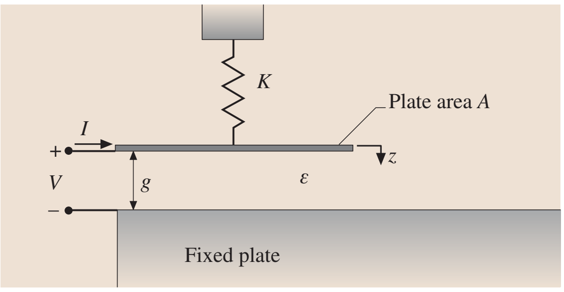

Fig. 35.24 Lumped parameter parallel plate capacitor electromechanical model

The parallel plate capacitor can be analyzed with either a voltage source or a current source. With a voltage source, the charge on the capacitor plates is

$$Q = \frac{\varepsilon A}{g} V \,, \tag{35.44}$$

where Q is the charge on the capacitor plates, ε is the dielectric constant in the gap, A is the capacitor plate area, g is the instantaneous gap between plates, and V is the voltage across the plates.

The electrostatic force on the capacitor plates is determined by the dielectric constant, the capacitor plate area, the instantaneous gap, and the voltage across the capacitor:

$$F = \frac{\varepsilon A V^2}{2g^2} \,. \tag{35.45}$$

The electrostatic force scales as the voltage squared and inversely diminishes as the instantaneous gap squared. This is the force that acts on the movable plate; hence the spring stretches (or is relaxed if the voltage is lowered). The amount of stretch in the spring is:

$$z = \frac{F}{K} \,. \tag{35.46}$$

The instantaneous gap is

$$g = g_0 - z = g_0 - \frac{\varepsilon A V^2}{2Kg^2} \,. \tag{35.47}$$

Electromechanics – Parallel Plate Capacitor with External Forcing

Nemirovsky et al. [35.66] examined the case of constant applied force acting on the suspended capacitor plate of a parallel-plate electrostatic actuator. This situation represents a force-balanced system, in which the electrostatic force is used to balance a pressure load acting on a membrane or an inertial load (e.g. acceleration induced) on a suspended proof mass. With an applied mechanical force, the mechanical energy is:

$$U_M(z) = \frac{1}{2} K z^2 - F_{ext} z \,, \tag{35.48}$$

where F_{ext} is the externally applied force. The solution to the pull-in equation is

$$\varsigma_{PI} = \frac{z}{z_{max}} = \frac{1}{3}\left(1 + 2\frac{F_{ext}/K}{z_{max}}\right) \,, \tag{35.49}$$

where ζ_{PI} is the ratio z/z_{max}. It can be seen when the applied force is zero. Equation (35.49) reduces to the familiar result of $z = z_{max}/3$. The pull-in voltage is

$$V_{PI} = \left(\frac{8K\left(z_{max} - \frac{F_{ext}}{K}\right)^3}{27\varepsilon_0 A}\right)^{1/2} \,. \tag{35.50}$$

Electromechanics – Parallel Plate Stabilization

An important practical consideration is the stabilization of the parallel plate electrostatic actuator. *Seeger* et al. [35.69] and Nemirovsky et al. [35.66] showed that the addition of a series capacitance (feedback capacitance) stabilizes the actuator such that it does not pull-in to the substrate. The series capacitance and the variable actuator capacitance combine to form a voltage divider that provides passive control of the voltage across the actuator. In this case, Nemirovsky et al. [35.66] identified three regimes, defined in terms of δ, which is the ratio of the initial actuator capacitance to the series capacitance (Nemirovsky et al. [35.66] had defined the series capacitance in terms of a dielectric layer on one capacitor plate). The first regime is defined for $0 < \delta < 0.5$, when the series capacitance is less than half the nominal actuator capacitance. In this case, the actuator is stable so the pull-in effect is eliminated. The second regime is defined by $\delta > 0.5$ and the pull-in occurs between $1/3$ of the gap and the full distance of the gap. In the third regime $\delta \gg 1$, which means the series capacitance is much greater than the nominal capacitance. The pull-in parameter is given by:

$$\varsigma_{PI} = \frac{1}{3}\left(1 + \frac{1}{\delta}\right) \,, \tag{35.51}$$

which reduces to $\zeta_{\mathrm{PI}} = \zeta_{\max}/3$ when the series capacitance is eliminated. The pull-in voltage is:

$$V_{\mathrm{PI}} = \left[\frac{8Kz_{\max}^3}{27\varepsilon_0 A} \left(1 + \frac{1}{\delta} \right)^3 \right]^{1/2} . \tag{35.52}$$

Electromechanics – Extending the Range of Motion

Electromechanical systems have two obvious characteristics: nonlinearity and instability. If one desires digital operation or fast switching times, the inherent electromechanical instability is a significant benefit. On the other hand if analog control is desired, the control is faced with both the nonlinearity and the instability. In both cases, the range of motion of the electromechanical device is limited. The range of motion to the instability was extended by the addition of a series capacitor, but this did not address the linearity. *Toshiyoshi* et al. [35.59] addressed the linearity of electrostatically actuated mirrors in an optical scanner. Their's was a two degree-of-freedom scanner, so the increased linearity was necessary to improve the crosstalk between the two different scanner modes. The scanner in question is a square plate suspended by torsion beams within a frame that is itself suspended by another pair of torsion beams. The plate and frame are suspended above the substrate, which has four stationary electrodes defined in the four quadrants of the plate as shown in Fig. 35.25. To actuate the mirror with increased linearity, a bias voltage V_{bias} was applied between each electrode and the plate. Independent differential control voltages are superimposed on the bias voltage supplied between each electrode and the plate. The driving voltages are described by two independent differential voltages such that $V_{\mathrm{diff1}} = (V_x + V_y)/2$, $V_{\mathrm{diff2}} = (-V_x + V_y)/2$, $V_{\mathrm{diff3}} = (-V_x + -V_y)/2$ and $V_{\mathrm{diff4}} = (V_x + -V_y)/2$. A second approach taken to improve the linearity was to consider the shape of the mirror, because of the geometric relationship that determines the net electrostatic torque. They considered such simple geometries as square, diamond (square rotated by 45), and a circular plate, as shown in Fig. 35.26, but more complex shapes could be considered. The square plate is presumed to have worse linearity because as the plate rotates the corners deflect toward the control electrodes, which increases the applied torque and further distorts the shape of the mirror. This suggested the removal of the corners could contribute to increased linearity and reduced distortion of the mirror surface. They found the circular plates reduced the distortion by 50% over the rectangular and diamond shaped plates, which is significant for the optical performance. The authors defined distortion as the normalized angle deviation from the ideal value. It is normalized relative to the maximum scan angle. *Nadal-Guardia* et al. [35.70] have described extending the range of travel by using current drive methods instead of voltage drive.

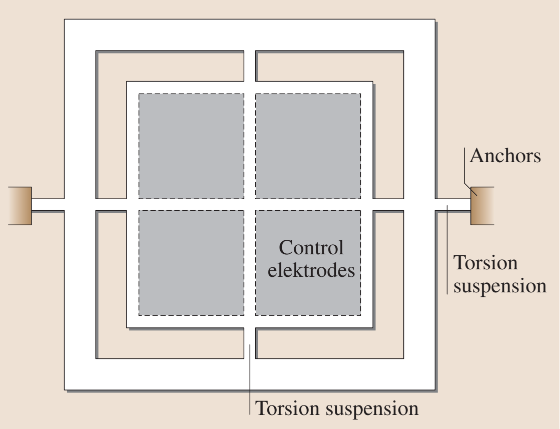

Fig. 35.25 Two degree-of-freedom torsional electromechanical actuator with a rectangular plate and a stationary electrode divided into four sections

35.2.4 Electromechanics of Beams and Plates

The basic understanding of electromechanics and pull-in phenomena has been developed around the simple parallel plate capacitor model developed above. Many MEMS designers innately understand that the electrostatic force is proportional to the square of the voltage, is inversely proportional to the square of the gap, and

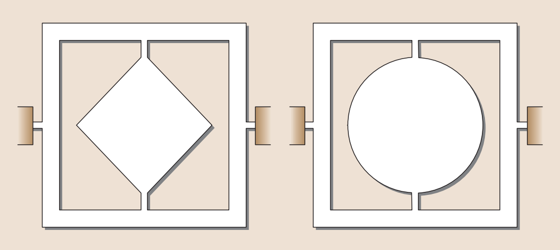

Fig. 35.26 Alternate two degree-of-freedom torsional, electromechanical actuators with diamond and circular plates

that the plate will pull in after displacing 1/3 of the gap. While this is good to first order, it is not adequate for the electromechanical design of other structures such as cantilevered beams/plates, beams/plates with multiple supports, plates with torsional supports, and other structures.

O'Brien et al. [35.71] and *Choi* and *Lovell* [35.72] have studied the electromechanics of cantilever beams by presenting two improved models for the cantilever beam electrostatic actuators. In both cases, the cantilever beam is a homogeneous material anchored to the substrate at one end and otherwise freely suspended along its length, which is supported above a stationary electrode. The new models are compared to the parallel plate capacitance model and a simple extension of the parallel plate model to a cantilever beam. The cantilever beam model uses a simple cantilever of width, W, length, L, thickness, T, and initial air gap of Z_0. The parallel plate capacitance model is based on a capacitance determined by the length and width of the beam. The beam deflection was represented by the displacement function of a beam with a force applied at the beam tip. The beam spring constant was determined by the ratio of the force at the beam tip to the displacement at the beam tip (maximum displacement). The capacitance was determined by integrating the parallel plate capacitance for a differential length element along the length of the beam, where the beam gap is defined by the initial gap in the mechanical displacement function. This showed the beam tip displaced 45% of the initial gap before pull-in occurred. In this case the pull-in voltage can be expressed as a function of the beam's Young's modulus E, cross-sectional moment of inertia I, length L, and width W, the initial gap Z_0, and the dielectric constant in the air gap as

$$V_1 = \sqrt{\frac{18EIZ_0^3}{5\varepsilon_0 L^4 W}} = 0.5477\sqrt{\frac{ET^3Z_0^3}{\varepsilon_0 L^4}} \quad \text{(Volts)} . \tag{35.53}$$

In the first improved model, the cantilever beam is defined by the same characteristics, but the electrostatic force acting along the length of the beam is replaced by an equivalent bending moment M applied at the beam tip. The beam displacement was determined for a uniform, homogeneous beam with a bending moment applied at its end. The beam's spring constant was determined by the ratio of the force at the beam tip and the displacement at the beam tip. The capacitance was determined by integrating the parallel plate capacitance for a differential length element along the length of the beam, where the beam gap is defined by the initial gap in the mechanical displacement function. This model showed the beam tip displacement was approximately 46% of the initial gap before pull-in occurred. In this case the pull-in voltage is

$$V_2 = \frac{43}{50}\sqrt{\frac{18EIZ_0^3}{5\varepsilon_0 L^4 W}} = 0.471\sqrt{\frac{ET^3Z_0^3}{\varepsilon_0 L^4}} \quad \text{(Volts)} . \tag{35.54}$$

In the second improved model, the cantilever beam was defined by a rigid beam, whose anchor was replaced by a simple hinge and continuum elasticity replaced by an equivalent spring at the beam's free end. Three new variables were introduced to describe the behavior in terms of radial coordinates centered at the hinge. The two radial coordinates describe the beginning of the electrode at r_0 and the end of the beam/electrode at r_1. The angle θ was introduced to describe the angular displacement of the beam. In this model, the tip of the beam displaced 44% of the initial gap before pulling into the substrate. The pull-in voltage is estimated by (35.55) or (35.56) where r_0 and r_1 have been replaced by 0 and L.

$$V_3 = \frac{11}{25r_1^2}\sqrt{\frac{77ET^3Z_0^3}{25\varepsilon_0\left[\frac{11(r_1-r_0)}{r_1-\frac{11}{25}r_0} - 14\ln\left(\frac{25r_1-11r_0}{14r_1}\right)\right]}} , \tag{35.55}$$

$$V_3 = \frac{11}{25L^2}\sqrt{\frac{77ET^3Z_0^3}{25\varepsilon_0}} = 0.4548\sqrt{\frac{ET^3Z_0^3}{\varepsilon_0 L^4}} \quad \text{(Volts)} . \tag{35.56}$$

As expected, the dependence on parameters, such as Young's modulus, beam thickness, beam length, initial air gap, and the air gap dielectric constant is the same for all cases presented. The difference in the three models is the coefficient of the pull-in voltage function and the tip deflection at pull-in. The coefficients of the pull-in voltage predictions are 0.5479, 0.471 and 0.4548, respectively. The beam tip deflected 45%, 44%, and 46% of the initial gap before pull-in occurred. This is in contrast to the parallel plate model that predicts the plate deflection will be 33% of the initial gap before pull-in. By comparison to experimental test results, the models presented in (35.54) and (35.56) were more accurate at predicting the pull-in voltage for cantilever beams. The parallel plate electromechanics model had greater error and under-predicted the pull-in voltage. The model represented by (35.53) over-predicted the pull-in voltage with greater error than either of (35.54) or (35.55). The beam models predicted more accurately for wider beams, which means they are better applied to beams with dimensions representative of a parallel plate model used as a basis of the capacitance estimate. In all three cases, the beam models presented in (35.53)–(35.55) predicted poorly the pull-in voltage for narrow beams, which are dominated by the effects of nonuniform fringing fields and not the uniform parallel plate fields. The typical application of the parallel plate capacitor model is to geometries with a width W and length L that are much greater than the capacitor gap Z_0 ($W \gg Z_0$, $L \gg Z_0$; where greater than a factor of 10 is a common rule of thumb). In Fig. 35.27, two capacitors are shown such that in one case the plate dimensions are much greater than the gap, and in the second case the plate dimensions are much smaller than the gap. When the plate dimensions are much larger than the gap, the uniform fields in the gap dominate the capacitive coupling. A small contribution to the capacitive coupling comes from the *fringing* fields at the edge of the plate (or the perimeter of the plate in three dimensions). As shown in Fig. 35.6, the fringing fields could depend on the thickness of the plate. When the plate dimensions are of the same order as the gap, the fringing fields may dominate or be of the same order as the contribution of the uniform field in the gap. This is the nature of the narrow cantilever beams that were not predicted well by a parallel plate capacitor model. A correction can be introduced to the parallel plate model to approximate the contribution of the fringing fields. This approximation is often referred to as the Love approximation.

Segmented Control of Electromechanical Membrane Deformation

Wang and *Hadaegh* [35.62] describe the electrostatic actuation of a deformable membrane mirror with segmented electrodes on the substrate. The membrane mirror, in this case silicon nitride with a conductive film, was suspended above the substrate containing the segmented electrode. The segmented electrodes are defined along angular and radial pattern. A specified voltage can be applied between the membrane and each electrode segment to develop a specific deformation for the mirror. This increases the complexity of the control system but also the flexibility of controlling the shape of the mirror surface. The optical properties could be optimized by individual characterization of membranes or by feedback from the optical output. This basic concept was also applied by *Zhang* and *Dunn* [35.73] who were able to linearize the voltage–deflection relationship by tailoring the shape of the electrodes and the beam structure used to support them.

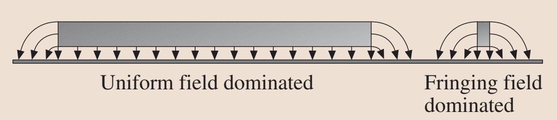

Fig. 35.27 Cross-sections of wide and narrow electrodes showing uniform electric fields in the gap and fringing fields at the edges (perimeter)

35.2.5 Electromechanics of Torsional Plates

The plate suspended by a torsional suspension instead of bending is the final configuration to be considered in terms of electromechanics. The electrostatic torsional plate has many applications, such as optical mirrors [35.58,63] and accelerometers [35.55]. The plate will define the optical surface of the mirror, the moveable plate of a varactor, or the proof mass of an accelerometer. It will be conductive or have a conductive film so that it can couple electrostatically to a stationary plate on the substrate. The moveable plate will be suspended by a torsional suspension such that the bending modes are suppressed and the primary kinematic mode is rotation of the plate. The variable capacitive coupling is used in two ways. First, it is used for actuation by applying a voltage between the moveable plate and a stationary plate. Second, it is used for sensing or use of the variable capacitance. Both actuation function and the variable capacitance function can be present in a device. The key elements of interest are the voltage-displacement function and the pull-in voltage.

Small Angular Deflection Approximation

Degani et al. [35.74], *Xiao* et al. [35.75], and Nemirovsky et al. [35.66] have considered the torsional electrostatic actuator from an analytical perspective with comparison to experimental results and other modeling results. The effective torsional spring constant is determined from the torque-twist angle relationship, where the torque is the electrostatic torque. The electrostatic torque is developed for both small and large angle cases. The pull-in angle and pull-in voltage are determined from the pull-in equations for voltage control (35.42, 43) and charge control (35.39, 40). Their method is an improved approach for modeling torsional electrostatic actuators over approximations involving parallel plate models and effective spring constants as they have shown by comparison to experimental measurements. The torsional plate considered is suspended by a torsional suspension, whose axis of rotation defines a gap of d. It is assumed the plates extend into the page (z-axis) by a depth b and have a total length a. The torsional plate electrostatic actuator is shown in Fig. 35.28. The solution assumes the tilted plates are semi-infinite, ignoring the effects of fringing fields hence their corrections, and are tilted by angles that satisfy the small angle approximation ($\tan\alpha \approx \sin\alpha \approx \alpha$ and $\cos\alpha \approx 1$). The capacitance is:

$$C(\Theta) = \frac{\varepsilon_0 b}{\alpha_{\max}} \frac{1}{\Theta} \ln\left(\frac{1}{1-\Theta}\right) , \tag{35.57}$$

where $\alpha_{\max} = d/a$ and $\Theta = \alpha/\alpha_{\max}$.

The capacitance function and the linear relationship for the elastic restoring force ($F_M = K\,a$) are used to develop the electrostatic and mechanical energy function to substitute in the pull-in equations. For voltage controlled actuation, the pull-in parameter is developed from (35.42) to yield:

$$4 + \frac{5\Theta_{PI} - 4}{(1-\Theta_{PI})^2} - 3\ln(1-\Theta_{PI}) = 0 . \tag{35.58}$$

Equation (35.58) is solved numerically to show the voltage-controlled pull-in angle is $\Theta_{PI} \approx 0.44$. The pull-in angle is substituted in (35.43) to determine the voltage controlled pull-in voltage:

$$V_{PI} = \left(\frac{0.827 K \alpha_{\max}^3}{\varepsilon_0 b}\right)^{1/2} . \tag{35.59}$$

The charge-controlled pull-in parameter relationship is determined by substituting the electrostatic energy and mechanical energy functions in the charge-controlled pull-in parameter (35.39). This provides the nonlinear relationship for the charge controlled pull-in angle:

$$\ln(1-\Theta_{PI}) - \frac{\Theta_{PI}}{(1-\Theta_{PI})^2} - \frac{2\Theta_{PI}^2}{(1-\Theta_{PI})^2 \ln(1-\Theta_{PI})} = 0 . \tag{35.60}$$

The charge-controlled pull-in angle is given by $\Theta_{PI} \approx 0.71$. The pull-in angle is used in the charge-controlled pull-in equation (35.40) to determine the pull-in charge relationship:

$$Q_{PI} = (1.798 K \varepsilon_0 b \alpha_{\max})^{1/2} . \tag{35.61}$$

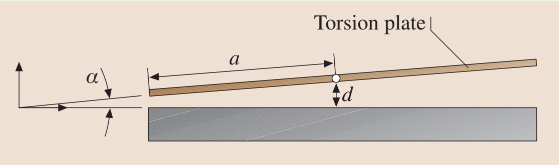

Fig. 35.28 Cross-section of a torsional actuator with a uniform cross section into the page showing the angular degree-of-freedom, α; plate length, a; and initial gap, d

Large Angular Deflection and Plate Geometry

Nemirovsky and *Bochobza-Degani* [35.66] have extended the work to the case of large angle deflections, which included using the trigonometric relationship in the capacitance function and a nonlinear relationship for the mechanical spring constant. These nonlinear relationships were used in the voltage-controlled pull-in equation. In addition, their methodology was extended to planar plates with different shapes (Fig. 35.29). They specifically considered triangular plates (the apex of the triangle is near the axis of rotation), a reversed triangular plate (the triangle apex is at the distal end of the plate), a square plate (with the rotational axis on the edge of the plate), and a circular plate (with the axis of rotation passing through a point on the perimeter and the opposite point on the perimeter defining the most distal point of the plate). Only simple geometries were considered, and their methodology can be applied to other more general geometry. The rectangular plate torsional, electrostatic actuator will provide the lowest actuation voltage relative to a fixed plate length. The reversed triangular plate has proved beneficial with regard to pixel area in mirror application and has been able to provide a higher torque for a specific plate length.

35.2.6 Leveraged Bending

As discussed, when the actuation electrode spans the entire length of the actuator, the stable controllable travel distance of the actuator plate is severely limited due to the pull-in instability. One can get around the problem of

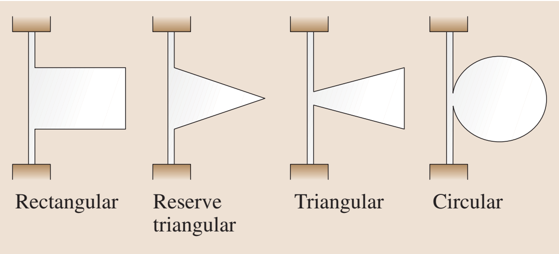

Fig. 35.29 Rectangular, triangular, reverse triangular, and circular plates are shown suspended by torsion beams for application to electromechanical mirrors or other actuators

pull-in by applying an electrostatic force to only part of the structure and using the rest of the structure as a lever to move the structure through a larger range of motion. This concept, referred to as leveraged bending [35.76], is shown schematically in Fig. 35.30 for a cantilever beam. The key idea is that the portions of the beam above the electrodes do not deflect far enough to violate the pull-in limit. The concept can be applied to other microstructure shapes and support conditions, but for fixed-fixed beams, the effect can be diminished by tensile residual stresses [35.76].

35.2.7 Electromechanics of Zipper Actuators

The design requirements for electromechanical actuators usually require minimizing the actuation voltage but maximizing the largest displacement. With the parallel plate actuator, be it a plate, beam, or torsional plate, the pull-in voltage is reduced by minimizing the air gap, but this limits the maximum displacement. Actuators with a zippering behavior have been implemented to achieve greater deflections for a lower actuation voltage. In one example (see [35.77, 78]), a plate is suspended by a torsional suspension. A long slender electrode extends from the plate and terminates with plate-like area. This electrode is supported above the fixed substrate electrode by a largely uniform gap. A voltage is applied between the substrate electrode and the extended electrode, such that the plate-like area is first pulled into the substrate. As the actuation voltage is increased, the long-slender electrode continues to collapse against the substrate electrode until a desired maximum deflection is achieved or until the drive voltage becomes impractical. This is the so-called zippering effect, which has been described for a motion perpendicular to the substrate. This uniform gap scenario is not the best suited for minimizing the pull-in voltage, so other configurations such as the curved electrode actuator have been proposed and applied to such devices as a switch and optical shutters as described by *Jin* et al. [35.79].

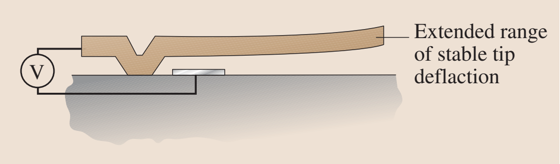

Fig. 35.30 Illustration of the concept of leveraged bending to increase the range of stable motion

Legtenberg et al. [35.80, 81] have described the basic operating features of the curved-electrode actuator. These actuators will provide large deflections in the plane of the substrate. The curved electrode actuator is composed of a stationary electrode and a compliant, moving electrode. The stationary electrode has a shape described by the function $s(x)$ shown in (35.62) and Fig. 35.31, which is written in terms of the maximum gap between electrodes $\delta_{\max}$, the length of the electrode L, the distance along the length of the electrode x, and the polynomial order of the curve n.

$$s(x) = \delta_{\max}\left(\frac{x}{L}\right)^n . \tag{35.62}$$

A necessary feature of the curved electrode actuators and zipper actuators, in general, is the contact between the moving and stationary electrodes. Without due consideration, these electrodes will short upon contact and damage the actuator. Two isolation solutions are implemented to prevent the shorting scenario. First, a continuous insulation film can be added between the electrodes to prevent them from shorting. In this case, the simplest solution is to add the insulation layer to the fixed, curved electrode. The alternative solutions include adding the insulating film to only the moving electrode or to both the moving and stationary electrodes. If the insulation is added to the moving electrode, it will become part of the stiffness of the beam. A second solution would be to add discrete, isolated contact bumpers that are made from the same conducting material as the moving electrode. As it closes the gap, the moving electrode makes contact with the isolated bumpers but not the actuation electrode. These have their operational differences as described by *Legtenberg* et al. [35.81]. Three other issues that must be considered during the design of a curved electrode actuator or zipper actuators are dielectric charging, stiction, and dielectric breakdown.

Legtenberg et al. [35.81] found that as the polynomial order increased for the electrode curve, the pull-in voltage will decrease for about the same maximum displacement. This means larger displacements can be achieved for lower voltages when compared to the parallel plate actuator. Once the moving electrode is pulled in to the curved electrode, the moving electrode will collapse to the maximum displacement. It was shown for polynomial curves of order above $n = 2$ that the maximum tip deflection becomes stable because of the geometric constraints provided by the curved electrode. This means to achieve the maximum displacement with the lowest pull-in voltage, a polynomial curve of order $n = 2$ should be specified. For polynomial curves with $n > 2$, stable continuous deflection along the curved electrode can be implemented with the maximum deflection occurring at lower voltages for polynomial orders just above two. If an approximate linear range of operation is desired (i. e. the tip deflection versus voltage is approximately linear), the higher order curves ($n > 2$) are desired. *Legtenberg* et al. [35.81] analyzed the curved electrode problem by application of analytical energy methods, specifically the Rayleigh–Ritz method and of three-dimensional self-consistent electromechanical numerical simulations using CoSolve-EM in MEMCAD (now integrated into CoventorWare). From an energy perspective, the total potential energy U_T is determined in terms of the mechanical energy associated with the strain energy of bending and the electrostatic energy stored in the electric field. For the curved electrode of Fig. 35.31, the total potential energy is:

$$\begin{aligned} U_{\mathrm{T}} &= U_{\mathrm{M}} + U_{\mathrm{E}} \\ &= \frac{1}{2}\int_0^L EI\left(\frac{\mathrm{d}^2 w\,(x)}{\mathrm{d}x^2}\right)^2 \mathrm{d}x \\ &\quad - \frac{1}{2}\int_0^L \frac{\varepsilon_0 h V^2}{\frac{d}{\varepsilon_r} + s(x) - w(x)}\,\mathrm{d}x\,, \end{aligned} \tag{35.63}$$

where L is the beam length, E is the beam's Young's modulus, I is the beam's cross-sectional moment of inertia, h is the thickness of the beam, V is the voltage applied between the beam and electrode, d is the initial gap at the start of the electrode, ε_0 is the permittivity of free space, and $w(x)$ is the beam's deflection curve. The next step in the solution of this problem is to determine an admissible trial function that satisfies the beam boundary conditions and can adequately approximate the solution. In this case the deflection curve for a uniformly loaded cantilever beam is chosen as the trial function:

$$\omega(x) = cx^2\left(6L^2 - 4Lx + x^2\right) = c\,f(x)\,, \tag{35.64}$$

where c is a constant to be determined for specific cases.

The approach of Nemirovsky et al. [35.66] can be taken by substituting the potential energy functions in the voltage-controlled pull-in equations. The simultaneous solution of the pull-in equations will result in the pull-in voltage and an implicit equation in terms of the constant c at pull-in. The pull-in voltage relationship is

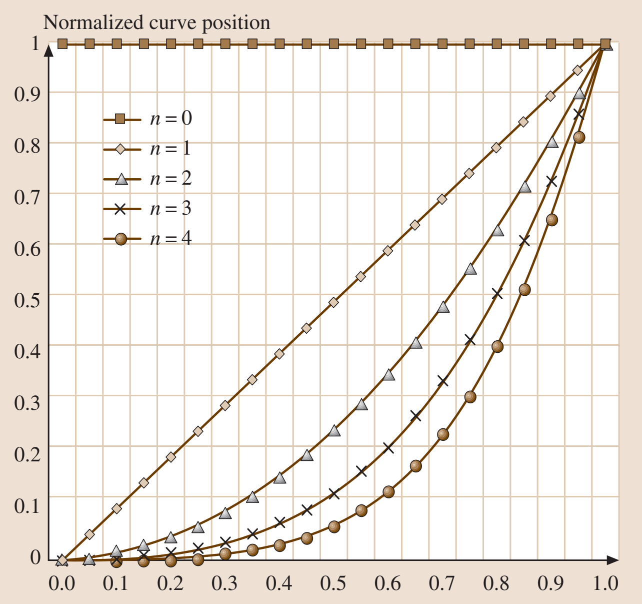

Fig. 35.31 Curved electrode shape curve shown as a function of order of the curve n

$$V_{\mathrm{PI}}^2 = \frac{EI}{\varepsilon_0 h}\,\frac{\int_0^L \left(\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}\right)^2 \mathrm{d}x}{\int_0^L \frac{f^2(x)}{[d/\varepsilon_r + s(x) - c_{\mathrm{PI}} f(x)]^3}\,\mathrm{d}x}\,. \tag{35.65}$$

The implicit equation to determine the coefficient of the beam deformation function is

$$\begin{aligned} &\int_0^L \frac{c_{\mathrm{PI}} f^2(x)\,\mathrm{d}x}{[d/\varepsilon_{\mathrm{r}} + s(x) - c_{\mathrm{PI}} f(x)]^3} \\ &\quad = \int_0^L \frac{f(x)\,\mathrm{d}x}{[d/\varepsilon_{\mathrm{r}} + s(x) - c_{\mathrm{PI}} f(x)]^2}\,. \end{aligned} \tag{35.66}$$

DeReus [35.82] designed the curved electrode actuator to convert the curvilinear motion into a translational motion at the tip of the moving electrode. The transformation is accomplished by a folded spring that connects the electrode to the body that will be translating. A solid model of this actuator design is shown in Fig. 35.32. A pull-in analysis was performed with the results before pull-in and after pull-in presented in Fig. 35.33. This shows that a 60 µm translation was accomplished

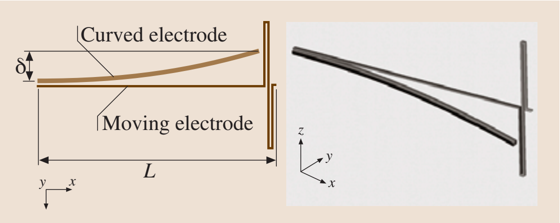

Fig. 35.32 Solid model of a curved electrode actuator with a folded suspension showing the curved electrode δ and the actuator beam length L

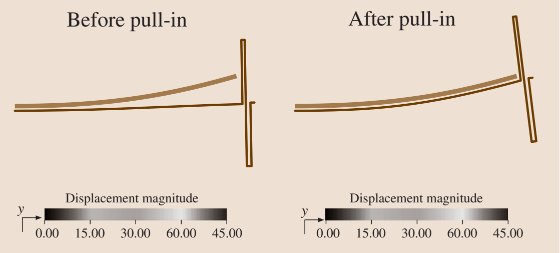

Fig. 35.33 Results of electromechanical pull-in analysis showing curved electrode actuator before and after pull-in has occurred

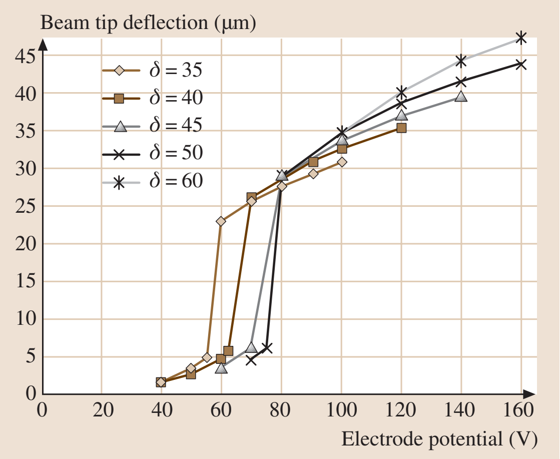

Fig. 35.34 Beam tip deformation as a function of electrode potential and δ for a suspension with a single fold

with this particular design, which has a twofold flexure and could have application as a optical shutter or VOA. In Fig. 35.34, electrode tip displacement is shown as a function of electrode voltage and electrode curve δ for a onefold flexure design. The displacement as a function of electrode potential and curve δ is shown for a twofold flexure in Fig. 35.35.

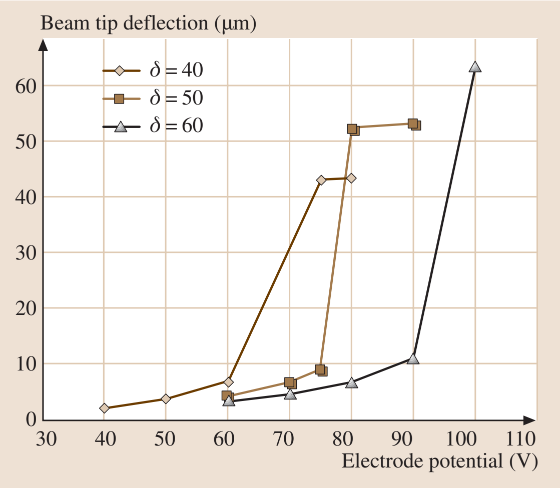

Fig. 35.35 Beam tip deformation as a function of electrode potential and δ for a suspension with a doubly folded suspension

35.2.8 Electromechanics for Test Structures

Electromechanics test structures are an important contributor to the development of microsystems products as they are used to extract information about material properties of thin films for specific process parameters. They are also used to monitor the material property variation during the fabrication process. In this case, they are distributed around the wafer on specific test die or they may be included in the dicing streets of more mature processes. These distributed test structures enable the manufacturer to track the variation in material properties with its associated within-wafer variation, wafer-to-wafer variation, and lot-to-lot variation. Another application of test structures is to characterize the impact of the first level packaging on the performance of the product. With all the critical information being extracted from the test structures, it is important for the information to be easy to interpret. Test structures have been developed that involve electrical, optical, and mechanical interrogation of the structure, but these are not easily instituted at all levels of characterization. The purely electrical test structures offer simple evaluation at the wafer, die, and package level so manufacturers often prefer them. The most common electrical test structures used for extracting information on mechanical properties involve electromechanics, of course. This section has provided a basis for interpreting and modeling the behavior of specific test structures, such as a homogeneous cantilever beam, and can be extended to other structures.

Osterberg and *Senturia* [35.83] describe in detail the application of electromechanics to the extraction of mechanical properties for microsystems design. They labeled their test structures as M-TEST for mechanical test structure. Their test structures comprised arrays of cantilever beams, fixed-fixed beams, and clamped, circular plates of various lengths, widths, or diameters. With these test structures, they demonstrated the extraction of important mechanical properties such as Young's modulus, plate modulus, and the residual film stress. Why choose an electromechanics test structure unless for the breadth and ease of application? An additional reason is the inherent instability that has been described as the pull-in of an electromechanical structure. The pull-in voltage is very distinct and can be measured in terms of a single parameter, the pull-in voltage. Given a control range of pull-in voltages for different structural layers in the manufacturing process, the pull-in parameter can be used as a simple control parameter.

Osterberg and *Senturia* [35.83] describe the methodology for extracting the mechanical properties. Their methodology includes the measurement of plan-view dimensions (i. e., width and length), out-of-plane curvature, pull-in voltage, film thickness, and air gap. With these measured parameters, the mechanical properties can be extracted. Their extraction methodology includes general closed form expressions for pull-in voltage that depend on the material properties and geometry of a homogeneous, isotropic material.

35.2.9 Electromechanical Dynamics: Switching Time

Switching time is an important factor in the design of such electromechanical devices as electrical or optical switches. The switching time analysis requires consideration of the dynamic equations representing the device and the electrical source. *Castaner* et al. [35.84–86] have considered the optimization of the speed-energy product. In this case, a parallel plate electrostatic actuator is described by an equivalent mechanical circuit that includes the mass m, the damping b, and the spring k, and a variable gap g. The variable gap depends on the initial gap g_0 and the displacement of the moving electrode x, such that $g = g_0 - x$. In the parallel plate, variable gap model the damping is described by squeeze film damping. (Other electrostatic actuators that move parallel to the substrate with variable gaps defined by groups of moving and stationary fingers.) In these cases, the damping model will need to include squeeze film damping and shear flow damping (Couette or Stokes Flow models). The electrical source is modeled as a voltage source and a source series resistance. A schematic of the model is shown in Fig. 35.36.

The equivalent electrical circuit of the parallel plate electrostatic actuator is modeled by (35.67), where V_s is the source voltage, V is the voltage across the variable capacitance, R is the series source resistance, and Q is the charge on the capacitor plates

$$\frac{V_\mathrm{s}}{R} = \frac{V}{R} + \frac{\mathrm{d}Q}{\mathrm{d}T} . \tag{35.67}$$

The equivalent mechanical circuit is described by:

$$m\frac{\mathrm{d}^2X}{\mathrm{d}T^2} + b\frac{\mathrm{d}X}{\mathrm{d}T} + kX = \frac{Q^2}{2C_0G_0} . \tag{35.68}$$

Equation (35.68) includes the sum of the inertial force, the damping force, the spring force, and the electrostatic force. Equation (35.68) is written in terms of the mass of the moving capacitor plate, m; the damping coefficient due to squeeze film damping, b; the elastic constant of the spring by which the capacitor plate is suspended, k; the charge on the capacitor plate, Q; the initial gap, G_0; and the initial capacitance, C_0. The initial capacitance is written in terms of the area of the capacitor A, the dielectric constant of the capacitor air gap ε, and the initial gap:

$$C_0 = \frac{\varepsilon A}{G_0} . \tag{35.69}$$

The dynamics of the parallel plate electrostatic actuation were divided into four events. The first occurrence is a short charging time for the capacitor. The second phase is the acceleration phase, which defines the rapidly increasing velocity of the mass, but the

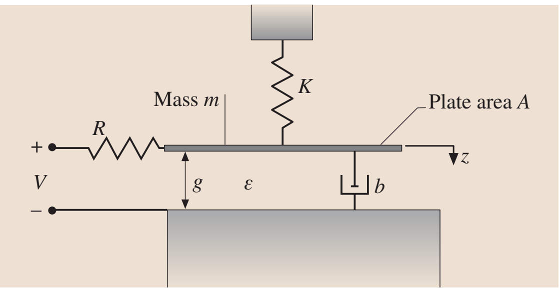

Fig. 35.36 Lumped parameter model of the parallel plate electromechanical actuator including damping and series resistance for dynamic analysis

position is still close to the initial equilibrium position. The third phase is the damping phase, defined by the velocity reaching a maximum, then decreasing due to the increased damping, and then finally increasing due to the increasing electrostatic forces as the gap closes. The fourth and final phase is the pull-in phase, characterized by a rapidly increasing velocity. *Hung* and *Senturia* [35.76] have described generating efficient dynamical models from a minimal set of finite element simulations. *Pons-Nin* et al. [35.87] have described the behavior of pull-in voltage and pull-in time for electromechanical actuators with current drives.

35.2.10 Electromechanics Issues: Dielectric Charging

Electrostatic actuators are operated in three different regimes depending on the application. In the first regime, the electrostatic actuator is operated over the voltage range that does not include the pull-in voltage. In this case, the actuated element does not contact the opposing substrate. In the second regime, the electrostatic actuator is operated through a voltage range that does include the pull-in voltage ($V_{act} \geq V_{PI}$). In this case, the desired response is to establish contact with or without some overdrive to achieve greater contact force or other response parameter. In the third regime, the electrostatic actuator is operated through a voltage range beyond the pull-in voltage. This is characteristic of electrostatic zipper actuators. In this case, the operating voltage is programmed to quickly establish pull-in, thereby contact, then the actuator operates in a range above the pull-in voltage ($V_{PI} < V_{act}$). In all three regimes, dielectric isolation layers are considered to prevent shorting between actuation electrodes. For the first and second case, the use of the insulating material is a preventative measure to protect against electrode shorting during an unintentional overdrive of the actuator. It is preventative because the actuation electrodes are not intended to touch. In the third case, the gap separating the moving and stationary actuation electrodes is reduced to zero so that an insulating layer is required. The insulating layer may be patterned or unpatterned.

The third regime defines the touch mode electrostatic actuators. *Cabuz* et al. [35.88] have described many of the failure modes that can be observed with the touch mode electrostatic actuators and that should be addressed during the design of the actuator and process. The observations and failure modes included: 1) a significant difference between the experimental actuation voltage (DC) and the theoretical actuation voltage, 2) a closed state electrostatic pressure lower than predicted, 3) an actuation voltage that increases with each subsequent cycle until electrical breakdown occurs, 4) actuator vibration observed between an open and a closed state, 5) actual release voltage is higher than theoretical, but the electrostatic forces should be much higher in the closed state such that the release voltage is always lower than the pull-in voltage, 6) the actuator remains closed temporarily after shorting, 7) permanent stiction will occur when in the closed state for a long period of time, 8) permanent stiction occurs after many actuation cycles.

Wibbeler et al. [35.89] have analyzed this problem in terms of one of the contributing factors, the accumulation and storage of localized charge in dielectrics. The common dielectrics in microsystems devices are SiO_2 and Si_3N_4, which show low mobility for surface charges and provide trap sites for positive or negative charge within their volume and at interfaces. *Kubo* et al. [35.90] showed this charge can be produced by the fabrication process by contact electrification, and by gas discharge in the air gap (the case focused on that by *Wibbeler* et al. [35.89]), and static electrification during handling (ESD). The contact electrification created electrical charge through the mechanical contact of two materials with different work functions and the separation of the two materials.

To understand the impact of the accumulated charge, *Wibbler* et al. [35.89] considered the impact of stored surface charge on the electrostatic force acting on a simple parallel plate capacitor. In Fig. 35.37, a simple capacitor is considered with a dielectric (ε_d) filling part of the air gap (ε). The dielectric has a surface charge of $\sigma_s(x, y)$. The capacitor plates carry charge shown as σ_1 and σ_2, which are provided by the battery, of voltage V, connected across the capacitor plates. The thicknesses of the dielectric and air gap are d_d and d, respectively. The force was estimated from the applied voltage and from the offset voltage due to the accumulated surface charge. The force acting on the capacitor plate as a function of the applied voltage, V; the offset

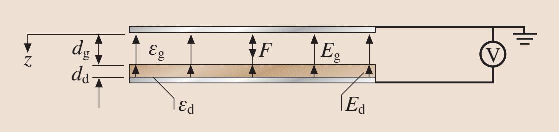

Fig. 35.37 Parallel plate capacitor with dielectric and parasitic charges

voltage, V_s; and the capacitance, C, is:

$$F = \frac{\mathrm{d}C}{\mathrm{d}z} \int_A \frac{[V - V_s(x, y)]^2}{2} \mathrm{d}A \,. \tag{35.70}$$

The offset voltage is shown in (35.71) as a function of the surface charge; the thickness of the dielectric, d_d; the dielectric constant, ε_d; or the dielectric's contribution to the total capacitance, C_d. The capacitance of the dielectric layer, per unit area, being defined as $C_d = \varepsilon_d / d_d$.

$$V_s = -\sigma_s(x, y) \frac{d_d}{\varepsilon_d} = -\frac{\sigma_s(x, y)}{C_d} \,. \tag{35.71}$$

Electromechanical Pull-In with Dielectric Charging

Nemirovsky et al. [35.66] and *Chan* et al. [35.91] studied the effect of residual charge at the interface of a dielectric and the air in the capacitor gap. With the dielectric in the capacitor gap and the residual charge, the total energy for the actuator is:

$$\begin{aligned} U_T = \frac{1}{2} K z^2 + \frac{Q_{res}^2}{2(C(z) + C_d)} \\ - \frac{C(z) C_d}{2(C(z) + C_d)} V^2 \,. \end{aligned} \tag{35.72}$$

In (35.72), Q_{res} is the residual charge, $C(z)$ is the air gap capacitance, and C_d is the dielectric capacitance. From the total energy, the pull-in parameter and pull-in equations (35.42) and (35.43) can be solved for the pull-in parameter ζ_{PI} and pull-in voltage V_{PI}:

$$\varsigma_{PI} = \frac{1}{3}\left(1 + \frac{1}{\delta}\right) \,, \tag{35.73}$$

$$V_{PI} = \left[\frac{8 K z_{max}^3}{27 \varepsilon_0 A} \left(1 + \frac{1}{\delta}\right)^3 - \frac{Q_{res}^2}{C_d^2} \right]^{1/2} \,. \tag{35.74}$$

In (35.73), ζ_{PI} is the ratio z/z_{max} and δ is the ratio $\varepsilon_d z_{max}/z_d$. Equation (35.74) shows that the dielectric layer will increase the pull-in voltage and the residual charge will lower the pull-in voltage. It can be seen that the residual charge does not impact the normalized pull-in coordinate shown in (35.73).

Test structures were fabricated to verify the shift predicted by (35.70) that would result with a dielectric in the air gap. Test structures with a dielectric did show a shift in the voltage–deflection curve, whereas a shift was not observed when a dielectric was not present. A shift of −25 V was observed. Following electrostatic discharge events, the offset voltages of −210 V and 145 V were observed during different phases of a charge-discharge cycle. This demonstrates the significant bias the accumulated surface charge can create. As a result, *Wibbler* et al. [35.89] did not recommend trying the fix shorting by adding dielectric films over the electrodes. They did recommend estimating the influence of accumulated charge on the response of the device being designed. Instead of complete dielectric film coverage of an electrode, they recommended patterning the dielectric such that it was limited to the edges of the electrodes or patterned at locations that will touch first. By patterning small areas of insulation, the surface area available to store charge is smaller and the distance to the electrode is shorter for more timely dissipation of stored charge.

Cabuz et al. [35.88] identified the effects of humidity, DC or AC driving voltage, and SAM coatings on the performance of electromechanical actuators. With a DC operating voltage, the devices operated as expected below 35% RH except in dry air or dry N_2. In the dry environment the actuator would temporarily remain closed, would remain closed after high fields are applied, or would remain closed after a high number of cycles. When operated above 35% RH, the DC voltage was higher than expected, the closed state electrostatic pressure was lower, the actuator did not release when expected, the actuator vibrated, and the voltage increased with subsequent cycles. When an AC actuation voltage is used the actuators operated up to 55% RH. At higher RH, the actuation voltage increased. The dry condition operation was improved with AC driving compared to DC operation. AC voltage control is recommended for eliminating RC effects but cannot overcome the effects of humidity. To avoid the effects of humidity, two SAM coatings (octadecyltrichlorosilane (OTS) and perfluorodecyltrichlorosilane (PFTS)) were tried to influence the hydration of the surfaces. The OTS eliminated the effects observed for uncoated samples up to a 95% RH level for both DC and AC voltages. The PTFS improved operation up to 95% RH for AC voltages and 70% RH for DC control voltages. The residual in-use stiction was determined to be associated with accumulated and trapped charge on the surfaces or at interfaces.

35.2.11 Electromechanics Issues: Gas Discharge

As described in previous sections, the electromechanical actuation of a microstructure is performed by applying a voltage across two plates. The two plates can be defined by two conductors (e.g. Al to Al), by semiconductors (e.g. polysilicon), or by a combination of a conductor and a semiconductor (e.g. Al and polysilicon). When these two plates (or other similarly described structures) are fabricated in proximity with each other, they are usually fabricated with an "air" gap less than 10 µm. With the electrostatic actuators, a small gap is desirable to minimize the voltage required for actuation.

The presence of small gaps in microsystems devices has prompted the further investigation of electrical breakdown in air and other ambient conditions at small gaps (i. e. less than 50 µm). *Slade* and *Taylor* [35.92] have demonstrated that breakdown in small gaps is defined by three regimes: less than 4 µm, 4 to 6 µm, and greater than 6 µm. In the first case, the breakdown event follows the behavior for a vacuum breakdown process at a larger gap. In the last case, the breakdown follows the behavior of Paschen's law for breakdown in air. The middle case of gaps between 4 and 6 µm, the breakdown process is in a transition from following the vacuum breakdown process to following the Paschen curve. In this case, the breakdown voltage is lower than predicted in either case, which indicates significant contributions from factors in both cases. The deviation from Paschen's curve was observed by *Torres* and *Dhariwal* [35.93] for gaps less than 4 µm.

The goal of this section is to describe some basic breakdown principles that can be used to determine whether or not a particular design will break down during an actuation event. Different processes will be considered for different gap regimes so that breakdown estimates can be made.

The vacuum breakdown region is characterized by a breakdown voltage that is a function of the gap separating the two electrodes. This contrasts with the breakdown behavior that follows the Paschen curve, which shows electrode gap-pressure product dependence. The breakdown voltage in vacuum is

$$V_B = K_V t_g \,. \tag{35.75}$$

Equation (35.75) shows the breakdown voltage V_B is proportional to the gap t_g where the proportionality constant is $K_v = 97\,V/\mu m$ for data from many sources as indicated by *Slade* and *Taylor* [35.92]. The breakdown data for vacuum was determined for electrode gaps between 35 and 200 µm. For electrode gaps between 0.2 and 40 µm in air, the proportionality constant was between 65 and 110 V/µm.

The electric field at the surface of the electrode is a critical parameter for determining breakdown in vacuum. This is estimated by the ratio of the voltage between the electrodes V and the gap between the electrodes t_g with a modification factors β_g. β_g is a geometrical enhancement factor that captures the field enhancement due to the macroscopic geometry:

$$E_g = \beta_g V/t_g \,. \tag{35.76}$$

The geometrical enhancement factor can be determined by performing finite element analysis to determine the macroscopic electric field as a function of the electrode gap for the electrode geometry under consideration. The geometric enhancement factor is determined from the ratio of the macroscopic electric field E_g, from finite element simulations, to the electric field estimated by V/t_g [35.92, 94]. The estimated electric field, V/t_g, can be used as a reasonable approximation when the electrode gap is much smaller than that radius of curvature of the cathode electrode. If this is not the case, the geometric enhancement factor needs to be estimated from numerical analysis such as finite element or boundary element methods. The geometric enhancement factor is shown, by *Slade* and *Taylor* [35.92], to be between 1.01 and 1.57 for gaps ranging from 0.2 to 40 µm, respectively, for the geometry of a needle cathode and a plate anode.

A second enhancement factor is attributed to the microscopic field enhancement due to the surface roughness of the electrodes. The electric field containing the effects of the geometric enhancement and the microscopic enhancement is:

$$E = \beta_m \beta_g V/t_g \,. \tag{35.77}$$

In (35.77) β_m is the microscopic enhancement factor. The total enhancement factor is $\beta_g\beta_m$, which can have a range of 100 to 250 for polished electrodes [35.95].

Following the estimate of the electric field at the surface of the electrodes, the current density for the field emission of electrons is calculated. This current density j_{FE} is described by the Fowler–Nordheim equation, which depends on the electric field E, the work function of the electrode material ϕ, and the dimensionless parameters of $t(y)$ and $\nu(y)$, where y is a parameter defined as $3.79\times10^{-5}\sqrt{E}/\phi$. Typically, $t(y)$ is one. $\nu(y)$ is given by $0.956-1.06y^2$. The Fowler–Nordheim equation is presented in (35.78) with j_{FE} in A/m², E in V/m, and ϕ in eV.

$$j_{FE} = \frac{1.54\times10^{-6}E^2}{\phi t(y)^2} \exp\left(\frac{-6.83\times10^9\phi^{3/2}\nu(y)}{E}\right) \,. \tag{35.78}$$

Equation (35.78) can be rewritten with E substituted from (35.77) and the current density replaced by an emission current I_e and the area A_e of a microprojection where the emission occurs. Equation (35.78) is shown rewritten in (35.79), such that the slope from the Fowler–Nordheim plot will provide the field enhancement factor as shown in (35.80) where m is the slope. The Fowler–Nordheim plot is the plot of I_e/V^2 versus $1/V$, which is constructed by measuring the voltage across the electrode gap and measuring I_e.

$$\begin{aligned}\mathrm{Log}_{10}\left(\frac{I_e}{V^2}\right) = {} & \mathrm{Log}_{10}\left(\frac{1.54\times10^{-6}A_e\beta^2}{\phi\, t(y)^2 t_g^2}\right)\\ & - \frac{1}{2.303}\left(\frac{6.83\times10^9\phi^{3/2}t_g v(y)}{\beta}\right)\frac{1}{V}\,,\end{aligned} \tag{35.79}$$

$$\beta = -\frac{2.303m}{6.83\times10^9\phi^{3/2}t_g v(y)}\,. \tag{35.80}$$

For gaps larger than 6 µm, the electrical breakdown as a function of gap does follow the Paschen curve for air and does not follow the vacuum breakdown curve for these larger gaps. The electrode gap is now larger than the mean free path of the air molecules so it is expected that emitted electrons will collide with the air molecules in the gap. For these larger electrode gaps, it is assumed the electrical breakdown will follow the Townsend electron avalanche theory. A similar process is followed to estimate the breakdown. First, the electric field at the surface of the cathode electrode is estimated including the enhancement factors for geometry and surface microstructure. An estimate of the total field enhancement factor is required for this. The electric field is substituted into the Fowler–Nordheim equation (35.78) to estimate the field emission current. In this case, the emitted electrons will interact with the background gas molecules causing an avalanche in the current. The current avalanche process is described by (35.79) and (35.80), where t_g is the electrode gap, α is the first Townsend coefficient, and γ is the second Townsend coefficient. The first Townsend coefficient represents the number of electrons produced per unit distance along the direction of the electric field. The second Townsend coefficient is the number of electrons generated by secondary processes per each primary avalanche. Breakdown will occur when the denominator of the expression for current approaches zero: $\gamma\exp(\alpha t_g)$ approaches unity.

In the transition region, the breakdown voltage can be lower than predicted by the Paschen curve or vacuum breakdown. It is suggested that in this regime, the partial electron avalanche process enhances the vacuum breakdown process to lead to a lower breakdown voltage [35.92].

The choice of materials is another consideration for the design of an electrostatic microactuator. As mentioned, the electrode pairs could comprise the following combinations: conductor-conductor, conductor-semiconductor, or semiconductor-semiconductor. In *Torres* and *Dhariwal* [35.93], the breakdown voltage was measured for electrode pairs comprised of nickel, brass, or aluminum. A significant difference was not observed between the breakdown voltages of the different conductors or for the size of the electrode. At larger gaps, the breakdown voltage was different for different electrode shapes (cylindrical versus spherical) but this can be explained by the relative difference in the geometric enhancement factor for the electric field for the different electrodes. *Ono* et al. [35.96] made a comparison between electrode combinations where both a semiconductor (silicon) and a silicon–metal combination were used. In the case of a silicon electrode with a metal electrode, the breakdown voltage deviated from the Paschen curve at approximately 6 µm. It was lower than the breakdown voltage predicted by the Paschen curve and trailed off with an approximate linear dependence on the electrode gap. With the silicon as one electrode, Au/Cr or Pt/Cr was used as the second electrode but little difference was found for the different metals. With the silicon-silicon electrode combination, the breakdown voltage followed the Paschen curve for air. For similar gaps, the breakdown voltage was significantly higher for the silicon-silicon electrodes than for the silicon-metal electrode. This was true for a thin silicon film over a Cr/Pt/Cr electrode.

Another option for reducing the electric breakdown in the air gap is to introduce an insulating material over the electrode. But *Wibbeler* et al. [35.89] have demonstrated this does not solve the problem of breakdown by itself and can lead to parasitic charging of the insulator. An electric breakdown will produce a large current, as described previously, that is destructive to the electrodes without insulation. The dielectric limits the current so that the electrical breakdown event is not so destructive. *Wibbeler* et al. [35.89] demonstrated through highly chaotic behavior in the voltage–displacement response of an electrostatic actuator. It is reasoned that the local field strength at the electrode reached a magnitude to cause an avalanche of free electrons and ions, as described here. These free charges travel the length of the air gap, but they are stopped and neutralized at the dielectric surface. This reduces the local electric field at the location of the surface charge so that the force on the cantilever is reduced and the process repeats itself. Throughout this process, a large amount of charge can be introduced onto the dielectric surface, which will typically have a very long dissipation time constant.

35.3 Summary and Mention of Topics not Covered

This chapter has focused on the thermomechanics and electromechanics of thin-film microstructures. Emphasis has been placed on the basic physical phenomena, approaches to analyze the phenomena, and the presentation of accessible results for special cases of practical importance in a variety of applications. We have not discussed many other energy domain couplings that play major roles in microsystems technology such as electrothermomechanics and magnetomechanics. We have also not discussed the role and behavior of materials such as piezoelectrics or shape memory alloys, which play important roles in electromechanics and thermomechanics, respectively. We refer the interested reader to *Senturia* [35.49] for an introduction and further references.

References

35.1 M. Madou: *Fundamentals of Microfabrication* (CRC, Boca Raton 1997)
35.2 M. Gad-el-Hak: *The MEMS Handbook* (CRC, Boca Raton 2001)
35.3 J.J. Sniegowski, M.P. de Boer: IC-compatible polysilicon surface micromachining, Annu. Rev. Mater. Sci. **30** (2000) 299–333
35.4 C. H. Mastrangelo: Adhesion-related failure mechanisms in microelectromechanical devices, Tribol. Lett. **3** (1997) 223–238
35.5 M. P. de Boer, T. M. Mayer: Tribology of MEMS, MRS Bull. **26** (2001) 302–304
35.6 W. D. Nix: Mechanical properties of thin films, Metall. Trans. A **20A** (1989) 2217–2245
35.7 C. Y. Hui, H. D. Conway, Y. Y. Lin: A reexamination of residual stresses in thin films and of the validity of Stoney's estimate, J. Electron. Packaging **122** (2000) 267–273
35.8 D. B. Bogy, K. C. Wang: Stress singularities at interface corners in bonded dissimilar materials, Int. J. Solids Struct. **7** (1971) 993–1005
35.9 V. L. Hein, F. Erdogan: Stress singularities in a two material wedge, Int. J. Fract. Mech. **7** (1971) 317–330
35.10 E. D. Reedy, T. R. Guess: Nucleation and propagation of an edge crack in a uniformly cooled epoxy/glass bimaterial, Int. J. Solids Struct. **39** (2002) 325–340
35.11 M. D. Thouless: Modeling the development and relaxation of stresses in films, Annu. Rev. Mater. Sci. **25** (1995) 69–96
35.12 N. D. Masters, M. P. de Boer, B. D. Jensen, M. S. Baker, D. Koester: Side-by-side comparison of passive MEMS residual strain test structures under residual compression. In: *Mechanical Properties of Structural Films*, ASTM STP **1413**, ed. by C. L. Muhlstein, S. B. Brown (2001) 168–200
35.13 M. W. Hyer: Calculation of the room-temperature shapes of unsymmetric laminates, J. Compos. Mater. **15** (1981) 296–310
35.14 M. W. Hyer: The room-temperature shape of four-layer unsymmetric cross-ply laminates, J. Compos. Mater. **16** (1982) 318–340
35.15 D. E. Fahnline, C. B. Masters, N. J. Salamon: Thin film stress from nonspherical substrate bending measurements, J. Vac. Sci. Technol. A **9** (1991) 2483–2487
35.16 C. B. Masters, N. J. Salamon: Geometrically nonlinear stress-deflection relations for thin film/substrate systems, Int. J. Eng. Sci. **31** (1993) 915–925
35.17 C. B. Masters, N. J. Salamon: Geometrically nonlinear stress-deflection relations for thin film/substrate systems with a finite element comparison, J. Appl. Mech. **61** (1994) 872–878
35.18 M. Finot, S. Suresh: Small and large deformation of thick and thin film multilayers: effects of layer geometry, plasticity and compositional gradients, J. Mech. Phys. Solids **44** (1996) 683–721
35.19 M. Finot, I. A. Blech, S. Suresh, H. Fujimoto: Large deformation and geometric instability of substrates with thin film deposits, J. Appl. Phys. **81** (1997) 3457–3464
35.20 L. B. Freund: The stress distribution and curvature of a general compositionally graded semiconductor layer, J. Cryst. Growth **132** (1993) 341–344
35.21 L. B. Freund: Some elementary connections between curvature and mismatch strain in compositionally graded thin films, J. Mech. Phys. Solids **44** (1996) 723–736

35.22 L. B. Freund: Substrate curvature due to thin film mismatch strain in the nonlinear deformation range, J. Mech. Phys. Solids **48** (2000) 1159–1174

35.23 N. J. Salamon, C. B. Masters: Bifurcation in isotropic thin film/substrate plates, Int. J. Solids Struct. **32** (1995) 473–481

35.24 R. M. Jones: *Mechanics of Composite Materials*, 2nd edn. (Taylor & Francis, London 1999)

35.25 M. W. Hyer: *Stress Analysis of Fiber-reinforced Composite Materials* (McGraw-Hill, Boston 1998)

35.26 G. G. Stoney: The tension of metallic films deposited by electrolysis, Proc. R. Soc. Lond. A **82** (1909) 172–175

35.27 C. D. Pionke, G. Wempner: The various approximations of the bimetallic thermostatic strip, J. Appl. Mech. **58** (1991) 1015–1020

35.28 P. Krulevitch, G. C. Johnson: Curvature of a cantilever beam subjected to an equibiaxial bending moment, Mater. Res. Symp. Proc. **518** (1998) 67–72

35.29 S. R. Swanson: *Introduction to Design and Analysis with Advanced Composite Materials* (Prentice-Hall, Upper Siddel River 1997)

35.30 M. L. Dunn, Y. Zhang, V. Bright: Deformation and structural stability of layered plate microstructures subjected to thermal loading, J. Microelectromech. Syst. **11** (2002) 372–383

35.31 J. A. King: *Materials Handbook for Hybrid Microelectronics* (Teledyne Microelectronics, Los Angeles 1988)

35.32 W. N. Sharpe: Mechanical properties of MEMS materials. In: *The CRC Handbook of MEMS*, ed. by M. Gad el Hak (CRC, Boca Raton 2001) Chap. 3

35.33 Y. Zhang, M. L. Dunn: Deformation of blanketed and patterned bilayer thin film microstructures during post-release thermal and cyclic thermal loading, J. Microelectromech. Syst. (2003) in press

35.34 B. D. Harper, C.-P. Wu: A geometrically nonlinear model for predicting the intrinsic film stress by the bending plate method, Int. J. Solids Struct. **26** (1990) 511–525

35.35 Y. L. Shen, S. Suresh: Thermal cycling and stress relaxation response of Si-Al and Si-Al-SiO$_2$ layered thin films, Acta. Metall. Mater. **43** (1995) 3915–3926

35.36 S. P. Baker, A. Kretschmann, E. Arzt: Thermomechanical behavior of different texture components in Cu thin films, Acta Mater. **49** (2001) 2145–2160

35.37 W. D. Nix: Elastic and plastic properties of thin films on substrates, Mater. Sci. Eng. A **234–236** (1997) 37–44

35.38 C. Thompson: The yield stress of polycrystalline thin films, J. Mater. Res. **8** (1993) 237–238

35.39 H. Gao, Y. Huang, W. D. Nix, J. W. Hutchinson: Mechanism-based strain gradient plasticity – I. Theory, J. Mech. Phys. Solids **47** (1999) 1239–1263

35.40 H. D. Espinosa, B. C. Prorok, M. Fischer: A novel method for measuring elasticity, plasticity, and fracture of thin films and MEMS materials, J. Mech. Phys. Solids **51** (2003) 47–67

35.41 D. C. Miller, M. L. Dunn, V. M. Bright: Thermally induced change in deformation of multimorph MEMS structures, Proc. SPIE **4558** (2001) 32–44

35.42 D. J. Vickers-Kirby, R. L. Kubena, F. P. Stratton, R. J. Joyce, D. T. Chang, J. Kim: Anelastic creep phenomena in thin film metal plated cantilevers for MEMS, Mater. Res. Soc. Symp. **657** (2001) EE2.5.1–EE2.5.6

35.43 R. M. Keller, S. P. Baker, E. Arzt: Stress-temperature behavior of unpassivated thin copper films, Acta Mater. **47** (1999) 415–426

35.44 M. D. Thouless, J. Cupta, J. M. E. Harper: Stress development and relaxation in copper films during thermal cycling, J. Mater. Res. **8** (1993) 1845–1852

35.45 M. D. Thouless, K. P. Rodbell, C. Cabral Jr.: Effect of a surface layer on the stress relaxation of thin films, J. Vac. Sci. Technol. **14** (1996) 2454–2461

35.46 R. P. Vinci, E. M. Zielinski, J. C. Bravman: Thermal stress and strain in copper thin films, Thin Solid Films **262** (1995) 142–153

35.47 J. Koike, S. Utsunomiya, Y. Shimoyama, K. Maruyama, H. Oikawa: Thermal cycling fatigue and deformation mechanism in aluminum alloy thin films on silicon, J. Mater. Res. **13** (1998) 3256–3264

35.48 K. Gall, M. L. Dunn, Y. Zhang, B. Corff: Thermal cycling response of layered gold/polysilicon MEMS structures, Mech. Mater. (2003) in press

35.49 S. D. Senturia: *Microsystems Design* (Kluwer, Dordrecht 2001)

35.50 M. P. de Boer, N. F. Smith, N. D. Masters, M. B. Sinclair, E. J. Pryputniewicz: Integrated platform for testing MEMS mechanical properties at the wafer scale by the IMaP methodology. In: *Mechanical Properties of Structural Films*, ASTM STP **1413** (2001) 85–95

35.51 Z. J. Yao, S. Chen, S. Eshelman, D. Denniston, C. Goldsmith: Micromachined low-loss microwave switches, J. Microelectromech. Syst. **8** (1999) 129–134

35.52 B. McCarthy, G. G. Adams, N. E. McGruer, D. Potter: A dynamic model, including contact bounce of an electrostatically actuated microswitch, J. Microelectromech. Syst. **11** (2002) 276–283

35.53 H.-S. Lee, C. H. Leung, J. Shih, S.-C. Chang, S. Lorincz, I. Nedelescu: Integrated microrelays: Concept and initial results, J. Microelectromech. Syst. **11** (2002) 147–153

35.54 S. J. Cunningham, S. Tatic-Lucic, J. Carper, J. Lindsey, L. Spangler: A high aspect ratio accelerometer fabricated using anodic bonding, dissolved wafer, and deep RIE processes, Proc. Transducers '99, the 10th Intl. Conf. Solid-State Sensors and Actuators, Sendai 1999, ed. by M. Esashi (IEE Jpn. 1999) 1522–1525

35.55 L. Spangler, C. Kemp: ISAAC: integrated silicon automotive acceleromter, Sens. Actuators A **54** (1996) 523–529

35.56 M. T. A. Saif, B. E. Alaca, H. Sehitoglu: Analytical modeling of electrostatic membrane actuator for micropumps, J. Microelectromech. Syst. **8** (1999) 335–345

35.57 C. Huang, C. Christophorou, K. Najafi, A. Naguib, H. M. Naguib: An electrostatic microactuator system for application in high-speed jets, J. Microelectromech. Syst. **11** (2002) 222–235

35.58 H. Toshiyoshi, H. Fujita: Electrostatic micro torsion mirrors for optical switch matrix, J. Microelectromech. Syst. **5** (1996) 231–237

35.59 H. Toshiyoshi, W. Piyawattanametha, C.-T. Chan, M. C. Wu: Linearization of electrostatically actuated surface micromachined 2-D optical scanner, J. Microelectromech. Syst. **10** (2001) 205–214

35.60 M. Fischer, M. Giousouf, J. Schaepperle, D. Eichner, M. Weinmann, W. von Münch, F. Assmus: Electrostatically deflectable polysilicon micromirrors-dynamic behaviour and comparison with the results from FEM modelling with Ansys, Sens. Actuators A **67** (1998) 89–95

35.61 P. K. C. Wang, R. C. Gutierrez, R. K. Bartman: A method for designing electrostatic-actuator electrode pattern in micromachined deformable mirrors, Sens. Actuators A **55** (1996) 211–217

35.62 P. K. C. Wang, F. Y. Hadaegh: Computation of static shapes and voltages for micromachined deformable mirrors with nonlinear electrostatic actuators, J. Microelectromech. Syst. **5** (1996) 205–220

35.63 J. Bühler, J. Funk, J. G. Korvink, F.-P. Steiner, P. M. Sarro, H. Baltes: Electrostatic aluminum micromirrors using double-pass metallization, J. Microelectromech. Syst. **6** (1997) 126–135

35.64 M. Fischer, H. Graef, W. von Münch: Electrostatically deflectable polysilicon torsional mirrors, Sens. Actuators A **44** (1994) 83–89

35.65 H. Schenk, P. Dürr, D. Kunze, H. Lakner, H. Kück: An electrostatically excited 2D-micro-scanning-mirror with an in-plane configuration of the driving electrodes, Proc. IEEE MEMS 2000, The 13th Ann. Intl. Conf. Micro Electro Mechanical Systems, Miyazaki 2000, ed. by I. Shimoyama, H. Kuwano (IEEE, Piscataway 2000)

35.66 Y. Nemirovsky, O. Bochobza-Degani: Methodology and model for the pull-in parameters of electrostatic actuators, J. Microelectromech. Syst. **10** (2001) 601–615

35.67 G. Li, N. R. Aluru: A Lagrangian approach for electrostatics analysis of deformable conductors, J. Microelectromech. Syst. **11** (2002) 245–254

35.68 N. R. Aluru, J. White: A multilevel Newton method for mixed-energy domain simulation of MEMS, J. Microelectromech. Syst. **8** (1999) 299–308

35.69 J. I. Seeger, S. B. Crary: Stabilization of electrostatically actuated mechanical devices, Proc. Transducers '97, 1997 Intl. Conf. Solid-State Sensors and Actuators, Chicago 1997, ed. by K. Wise (IEEE, Piscataway 1997) 1133–1136

35.70 R. Nadal-Guardia, A. Dehe, R. Aigner, L. M. Castaner: Current drive methods to extend the range of travel of electrostatic microactuators beyond the voltage pull-in point, J. Microelectromech. Syst. **11** (2002) 255–263

35.71 G. J. O'Brien, D. J. Monk, L. Lin: Electrostatic latch and release; a theoretical and empirical study, Proc. Micro-Electro-Mechanical Syst. (MEMS) 2000, MEMS-Vol. 2, The 2000 ASME Intl. Mech. Eng. Cong. and Expo., Orlando 2000, ed. by A. J. Malshe, Q. Tan, A.¶. Lee, F. R. Forster, R. S. Kenten (ASME, New York 2000) 19–26

35.72 B. Choi, E. G. Lovell: Improved analysis of microbeams under mechanical and electrostatic loads, J. Micromech. Microeng. **7** (1997) 24–29

35.73 Y. Zhang, M. L. Dunn: A vertical electrostatic actuator with extended digital range via tailored topology, Proc. SPIE **4700** (2002) 147–156

35.74 O. Degani, E. Socher, A. Lipson, T. Leitner, D. J. Setter, S. Kaldor, Y. Nemirovsky: Pull-in study of an electrostatic torsion microactuator, J. Microelectromech. Syst. **7** (1998) 373–379

35.75 Z. Xiao, X. Wu, W. Peng, K. R. Farmer: An angled-based design approach for rectangular electrostatic actuators, J. Microelectromech. Syst. **10** (2001) 561–568

35.76 E. S. Hung, S. D. Senturia: Generating efficient dynamical models for microelectromechanical systems from a few finite-element simulation runs, J. Microelectromech. Syst. **8** (1999) 280–289

35.77 J. R. Gilbert, R. Legtenberg, S. D. Senturia: 3D coupled electro-mechanics for MEMS: Application of CoSolve-EM, Proc. IEEE MEMS Conference, Amsterdam 1995, ed. by M. Elwenspoek, N. de Rooij (IEEE, Piscataway 1995) 122–127

35.78 J. R. Gilbert, G. K. Ananthasuresh: 3D modeling of contact problems and hysteresis in coupled electro-mechanics, Proc. IEEE MEMS 1996, The ninth Ann. Intl. Workshop on Micro Electro Mechanical Syst., San Diego 1996, ed. by M. G. Allen, M. L. Reed (IEEE, Piscataway 1996) 127–132

35.79 Y.-H. Jin, K.-S. Seo, Y.-H. Cho, S.-S. Lee, K.-C. Song, J.-U. Bu: An integrated SOI optical microswitch using electrostatic micromirror actuators with insulated touch-down beams and curved electrodes, Proc. Micro-Electro-Mechanical Syst. (MEMS) 2000, MEMS-Vol. 2. The 2000 ASME Intl. Mech. Eng. Cong. and Expo., Orlando 2000, ed. by A. J. Malshe, Q. Tan, A. P. Lee, F. R. Forster, R. S. Kenten (ASME, New York 2000) 177–181

35.80 R. Legtenberg, E. Berenschot, M. Elwenspoeke, J. Fluitman: Electrostatic curved electrode actuators, Proc. IEEE MEMS Conference, Amsterdam 1995 (IEEE, Piscataway 1995) 37–42

35.81 R. Legtenberg, J. Gilbert, S. D. Senturia, M. Elwenspoek: Electrostatic curved electrode actuators, J. Microelectromech. Syst. **6** (1997) 257–265

35.82 D. DeReus: Personal communications and internal reports (2002)

35.83 P. M. Osterberg, S. D. Senturia: M-TEST: A test chip for MEMS material property measurement using electrostatically actuated test structures, J. Microelectromech. Syst. **6** (1997) 107–118

35.84 L. M. Castaner, S. D. Senturia: Speed-energy optimization of electrostatic actuators based on pull-in, IEEE J. Microelectromech. Syst. **8** (1999) 290–298

35.85 L. Castaner, A. Rodriguez, J. Pons, S. D. Senturia: Measurement of power-speed product of electrostatic actuators, Proc. Transducers '99, the 10th Intl. Conf. Solid-State Sensors and Actuators, Sendai 1999, ed. by M. Esashi (IEE Jpn., Tokyo 1999) 1772–1775

35.86 L. Castaner, A. Rodriguez, J. Pons, S. D. Senturia: Pull-in time-energy product of electrostatic actuators: Comparison of experiments and simulation, Sens. Actuators **83** (2000) 263–269

35.87 J. Pons-Nin, A. Rodriguez, L. M. Castaner: Voltage and pull-in time in current drive of electrostatic actuators, J. Microelectromech. Syst. **11** (2002) 196–205

35.88 C. Cabuz, E. I. Cabuz, T. R. Ohnstein, J. Neus, R. Maboudian: Factors enhancing the reliability of touch-mode electrostatic actuators, Sens. Actuators **79** (2000) 245–250

35.89 J. Wibbeler, G. Pfeifer, M. Hietschold: Parasitic charging of dielectric surfaces in capacitive microelectromechanical systems (MEMS), Sens. Actuators A **71** (1998) 74–80

35.90 H. Kubo, T. Namura, K. Yoneda, H. Ohishi, Y. Todokoro: Evaluation of charge build-up in wafer processing by using MOS capacitors with charge collecting electrodes, ICMTS 1995 Proc. IEEE Intl. Conf. Microelectronic Test Structures, Vol. 8, Nara 1995, ed. by T. Sugano, K. Asada (IEEE, Piscataway 1995) 5–9

35.91 E. K. Chan, K. Garikipati, R. W. Dutton: Characterization of contact electromechanics through capacitance-voltage measurement and simulations, IEEE J. Microelectromech. Syst. **8** (1999) 208–217

35.92 P. G. Slade, E. D. Taylor: Electrical breakdown in atmospheric air between closely spaced (0.2 μm–40 μm) electrical contacts, Proc. 47th IEEE Holm Conf. Electrical Contacts, Montreal 2001, ed. by K. Leung (IEEE, Piscataway 2001) 245–250

35.93 J-M. Torres, R. S. Dhariwal: Electric field breakdown at micrometer separations, Nanotechnology **10** (1999) 102–107

35.94 R. Longwitz, H. van Lintel, R. Carr, C. Hollenstein, P. Renaud: Study of gas ionization schemes for microdevices, Proc. Transducers '01, Eurosensors XV, The 11th Intl. Conf. Solid-State Sensors and Actuators, Munich 2001, ed. by O. Obermeier (Springer, Berlin, Heidelberg 2001) 1258–1261

35.95 D. K. Davies, M. F. Biondi: Vacuum breakdown between plane-parallel copper plates, J. Appl. Phys. **37** (1966) 2969–2977

35.96 T. Ono, D. Y. Sim, M. Esashi: Micro-discharge and electric breakdown in a micro-gap, J. Micromech. Microeng. **10** (2000) 445–451

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

36. High Volume Manufacturing and Field Stability of MEMS Products

Low volume MEMS/NEMS production is practical when an attractive concept is implemented with business, manufacturing, packaging, and test support. Moving beyond this to high volume production adds requirements on design, process control, quality, product stability, market size, market maturity, capital investment, and business systems. In a broad sense, this chapter uses a case study approach: It describes and compares the silicon-based MEMS accelerometers, pressure sensors, image projection systems, and gyroscopes that are in high volume production. Although they serve several markets, these businesses have common characteristics. For example, the manufacturing lines use automated semiconductor equipment and standard material sets to make consistent products in large quantities. Standard, well controlled processes are sometimes modified for a MEMS product. However, novel processes that cannot run with standard equipment and material sets are avoided when possible. This reliance on semiconductor tools, as well as the organizational practices required to manufacture clean, particle-free products partially explains why the MEMS market leaders are integrated circuit manufacturers. There are other factors. MEMS and NEMS are enabling technologies, so it can take several years for high volume applications to develop. Indeed, market size is usually a strong function of price. This becomes a vicious circle, because low price requires low cost – a result that is normally achieved only after a product is in high volume production. During the early years, IC companies reduced cost and financial risk by using existing facilities for low volume MEMS production. As a result, product architectures are partially determined by capabilities developed for previous products. This chapter includes a discussion of MEMS product architecture with particular attention to the impact of electronic integration, packaging, and surfaces. Packaging and testing are critical, because they are significant factors in MEMS product cost. These devices have extremely high surface/volume ratios, so performance and stability may depend on the control of surface characteristics after packaging. Looking into the future, the competitive advantage of IC suppliers will decrease as small companies learn to integrate MEMS/NEMS devices on CMOS foundry wafers. Packaging challenges still remain, because most MEMS/NEMS products must interact with the environment without degrading stability or reliability. Generic packaging solutions are unlikely. However, packaging subcontractors recognize that MEMS/NEMS is a growth opportunity. They will spread the overhead burden of high-capital-cost-facilities by developing flexible processes in order to package several types of moderate volume integrated MEMS/NEMS products on the same equipment.

36.1 **Manufacturing Strategy** 1086
  36.1.1 Volume 1086
  36.1.2 Standardization 1086
  36.1.3 Production Facilities 1086
  36.1.4 Quality 1087
  36.1.5 Environmental Shield 1087

36.2 **Robust Manufacturing** 1087
  36.2.1 Design for Manufacturability 1087
  36.2.2 Process Flow and Its Interaction with Product Architecture 1088
  36.2.3 Microstructure Release 1095
  36.2.4 Wafer Bonding 1095
  36.2.5 Wafer Singulation 1097
  36.2.6 Particles 1098
  36.2.7 Electrostatic Discharge and Static Charges 1098
  36.2.8 Package and Test 1099
  36.2.9 Quality Systems 1101

36.3 **Stable Field Performance** 1102
  36.3.1 Surface Passivation 1102
  36.3.2 System Interface 1105

References 1106

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Solid-state pressure sensors were first reported in the 1960s and commercialized by a number of start-up companies in the 1970s. By 1983, a *Scientific American* cover story [36.1] described pressure sensors, accelerometers, and ink-jet print heads and featured a gas chromatograph that combined an injection valve, a detector, and a chromatographic column on a silicon wafer. The future of the budding micromechanical device industry looked bright, but commercial reality has seriously lagged the rosy market projections. The reason was simple: Market forecasters failed to recognize the difference between lab prototypes and high volume production of stable products. Many companies have demonstrated that small quantities can be produced for niche markets. However, routine high volume production of reliable packaged devices is much more challenging. It requires the production prototype skills plus the manufacturing disciplines that are critical to long-term stable production.

Successful companies understand and apply these lessons. Indeed, over a hundred million MEMS pressure sensor, accelerometer, gyro, gas flow, and optical projection devices are shipped to customers annually. There is no universally accepted definition of MEMS. This chapter focuses on silicon-based products with movable micromechanical elements. It excludes disk drive heads, ink-jet print heads, hearing aids, microscale plastic, ceramic, quartz and metal components, and test strips for in vitro diagnostics that are sometimes included within the definition of MEMS. Many MEMS products do not integrate support electronics with the MEMS element on the chip even though they are produced with IC equipment, materials, and processes. However, integrated MEMS products are becoming ubiquitous. By 2002 Analog Devices had shipped over a hundred million surface micromachined MEMS accelerometers with integrated electronics to customers.

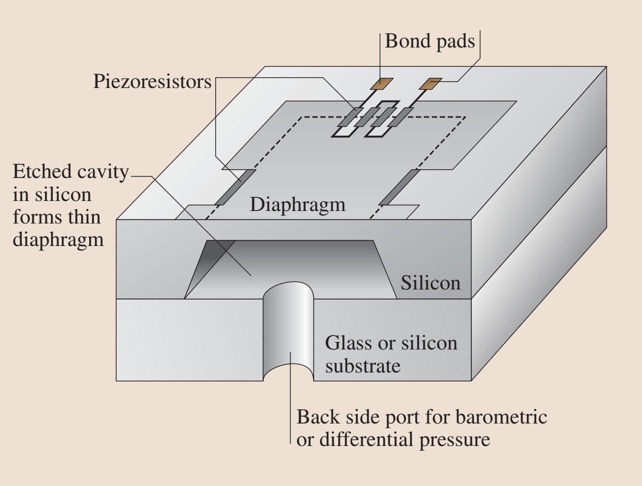
Bond pads
Piezoresistors
Etched cavity in silicon forms thin diaphragm
Diaphragm
Silicon
Glass or silicon substrate
Back side port for barometric or differential pressure

Fig. 36.1 Cross section of a piezoresistive silicon sensor after it is bonded to a silicon or glass substrate. When pressure is applied, the thin silicon diaphragm deforms. This causes changes in four implanted resistors that form a Wheatstone bridge. The resistors are located and oriented to cause resistance to increase in two resistors and decrease in the other two

The literature related to high volume production of MEMS products is quite uneven. Some suppliers have published detailed process and design descriptions, while others limit disclosures to market-focused publications. The two major suppliers of integrated MEMS products have discussed their designs and processes in detail. *Core* et al. [36.2], *Kuehnel* et al. [36.3], *Chau* et al. [36.4], and *Sulouff* [36.5] describe the Analog Devices *i*MEMS accelerometers and gyros. *Mignardi* et al. [36.6] give an in-depth review of the Texas Instruments DLP (Digital Light Processing) products.

MEMS devices are components in larger systems, so every MEMS product must define its system interface in a way that adds value from the perspective of the system designer. Some MEMS pressure sensor products are simply Wheatstone bridge elements in a circuit (Fig. 36.1). The image projection systems that use the Texas Instruments Digital Light Processing Technology$^{\text{TM}}$ could never be designed in this manner, because up to 1.3 million mirrors are driven on each chip (Fig. 36.2). Such a system would be impractical unless electronics are integrated with the MEMS mirrors on the chip. Air bag sensors also illustrate how integration adds value. Non-MEMS ball-and-spring air bag sensors were often used in the early 1990s. Upon impact, a ball would be released and detected as it passed through the sensor tube. Unfortunately, these binary sensors could not always distinguish a frontal collision from the jolt of a pothole or a side collision. Therefore, multiple ball-and-spring sensors were placed at different locations in a vehicle and wired together in order to reliably identify accidents that should result in air bag deployment. High cost limited the use of these distributed systems. Analog Devices changed the air bag market when it introduced the ADXL50 in 1993 (Fig. 36.3). The ADXL50 integrated electronics and an accelerometer on one chip to provide an analog output of deceleration versus time when an impact event occurred. This allowed automobile manufacturers to design air bag modules that compared the ADXL50 output to the known signature of a frontal collision. The result was greater reliability

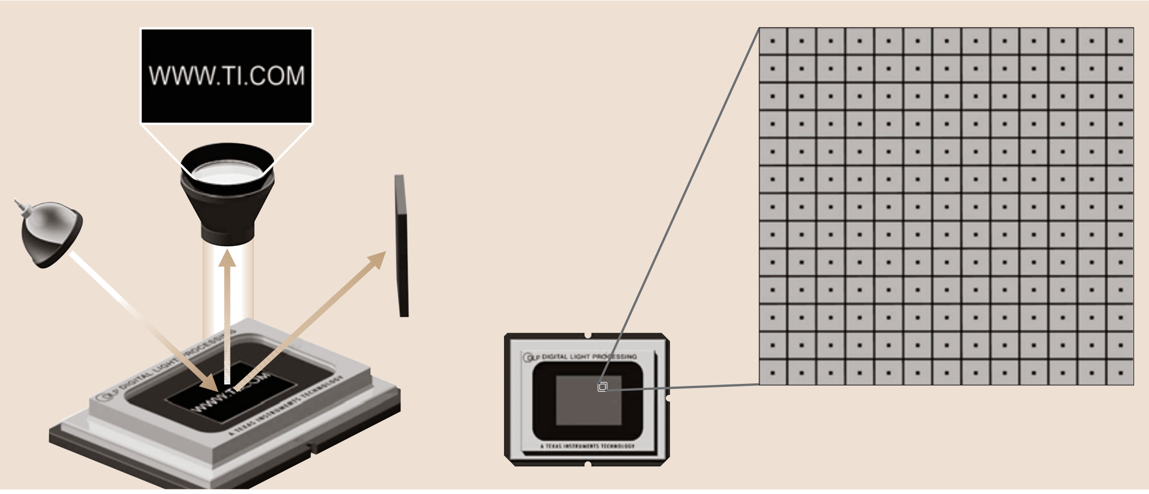

Fig. 36.2 Digital Light Processing technology^{TM} is based on Digital Micromirror Device (DMD) chips that address and drive a "sea of mirrors." The DMD is the key element in small computer-based projection systems. Each of the 14 or 17 µm mirrors represents one pixel. The diagram on the *left* represents light reflecting from a mirror to the system lens. The illustration on the *right* is an exploded view of a small region of mirrors in a packaged DMD. (Courtesy of Digital Light Processing^{TM}, Texas Instruments, Inc.)

and lower cost, which led to the installation of air bags in almost every new automobile.

This chapter will not recite detailed descriptions of particular commercial products. Instead, brief descriptions will be used to illustrate basic principles and concepts. Part of the chapter is organized according to MEMS-specific issues like wafer singulation. This is useful because some designs and unit processes are well suited for lab scale or low volume production, but impractical in a high volume manufacturing environment. Complete reliance on such an organization would be artificial, because partitioning omits the greatest challenge – combining these individual topics into a manufacturing flow that routinely produces products that meet cost, performance, and reliability expectations. The importance of process integration cannot be

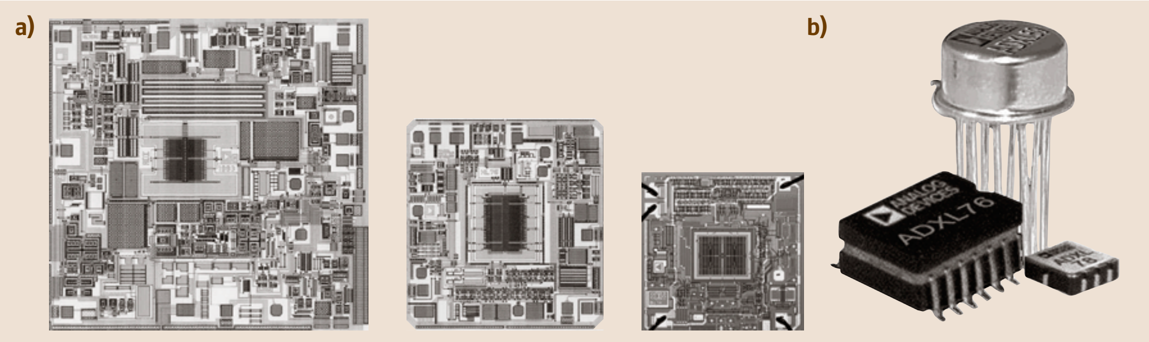

Fig. 36.3a,b Three generations of integrated accelerometers designed for single-axis air bag applications. **(a)** Die size is reduced by a factor of two about every four years. **(b)** Package technology also evolved to meet size, cost, and automated handling requirements. The ADXL50 was packaged in TO-100 seam-sealed metal headers. ADXL76 cerpacs (*bottom left*) are being replaced by the ADXL78 in 5 mm × 5 mm LCC packages (*bottom right*). (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

overemphasized, because changes in one step almost always affect other steps. Solutions may involve design modifications, so design-for-manufacturability is an important topic. Manufacturing and business skills are included to the extent that they apply to stable production of reliable MEMS products.

36.1 Manufacturing Strategy

MEMS manufacturing is based on existing IC technology. However, there are only a few examples of routine MEMS and IC production intermixed on the same equipment (or of MEMS devices with electronics integrated on the chip). Even when a MEMS process step is nominally the same as a standard IC process, the specific conditions or end point or some quality such as film stress will differ. Technologists from Analog Devices and Motorola have discussed these thin film issues in depth [36.7–10].

Manufacturers prefer to use dry etch processes when possible in order to achieve better control of features, minimize waste disposal, and utilize automated equipment developed for the IC industry. However, anisotropic wet etching provides unique value in the production of some MEMS products like pressure, mass flow, and yaw rate sensors [36.11].

Although MEMS production economics is closely tied to the semiconductor model, useful lessons can be drawn from many industries. Some of the key characteristics that affect success are as follows:

36.1.1 Volume

Semiconductor economics is rooted in the principle that expensive processes are cost-effective when many devices are made simultaneously. Obviously, a MEMS product that is sold in low quantities loses this critical advantage. MEMS components are “disruptive technologies” – they are the key element in higher-level products that were previously impractical. It can take several years to develop a higher-level product and build its customer base. Therefore, it is prudent to include an incubation phase in the business plan for a novel MEMS device. Low sales volume during this phase causes high unit costs – batch process cost benefits are achieved only after the market develops.

Start-up companies often do not understand the implications of design-for-volume. It is a major challenge to transfer products made on lab tools to manufacturing equipment. The challenge is sufficient to justify the use of manufacturing people and equipment for product development.

Realistic price-volume estimates are also essential to a good business plan. The highest volume MEMS products (accelerometers and pressure sensors) all represent a small part of the total system cost, even though the functions they provide are absolutely critical to system performance.

Market-pull business models that respond to the needs of a specific large application in a particular industry are usually more successful than technology-driven models that seek applications for new devices. Given this reality, it is prudent to thoroughly understand the target industry and its quirks. Biotech, wireless, image projection, optical communications, and consumer markets offer tremendous growth potential. However, the past and present growth in high volume MEMS products have been driven primarily by the automotive industry, with pressure sensors also benefiting from medical applications. *Sulouff* [36.12], *Weinberg* [36.13], *Marek* et al. [36.14], *Eddy* et al. [36.15], and *Verma* et al. [36.16] discuss automotive applications and the challenges a successful supplier must master.

36.1.2 Standardization

Mature industries develop standards in response to equipment vendors, material suppliers, and customer interface requirements. The MEMS industry is fragmented, so even after 30 years it is not mature. Deviations from semiconductor industry standards are inevitable because MEMS products impose unique requirements on wafer processing, packaging, handling, and testing. For example, KOH wet etching raises mobile ion concerns. Gold used in some optical products is incompatible with CMOS. Any deviation from standard process and material sets is serious, because the price-volume relationship is a steep curve in high volume applications.

36.1.3 Production Facilities

The semiconductor industry is divided into front-end (wafer fabrication) and back-end (product assembly and test) operations. IC design rules are based on pre-qualified fab, assembly, and test capabilities at specific sites. Consider how each of these functions matches the needs of MEMS production:

1. An IC product might be produced at an internal wafer fab, or designed to run on a standard foundry process. In contrast, production challenges and "know-how" have caused MEMS suppliers to retain wafer fab production within their internal facilities. High volume production in MEMS foundries was limited to pressure sensors. However, Motorola is now outsourcing accelerometer production. This suggests that the MEMS industry will follow the IC model, with foundry services becoming more substantial in the future.
2. IC suppliers have shifted assembly operations to subcontractor facilities located in Southeast Asia. Some MEMS suppliers have followed this model. It is common to assemble low volume products in North America, Europe, or Japan. However, semiconductor assembly and package technology are evolving rapidly and are driven by subcontractors based in Southeast Asia. MEMS companies that do not utilize these resources pay a significant cost penalty. Obviously, the break point at which offshore assembly is attractive differs from one product to the next. Products that have unique assembly requirements and support high market prices may never reach this point.
3. IC testing is often carried out at the assembly subcontractor site. MEMS package and testing comprise a substantial portion of the total product cost, so a high volume manufacturing strategy should consider Southeast Asia packaging and testing. This is complicated by the fact that MEMS testing requires unique stimuli that are not in the standard IC portfolio.

36.1.4 Quality

The goal in lab and low volume production environments is to make functioning devices. As quantities increase, other factors become more important: lot-to-lot repeatability, yield, device performance that meets well-defined specifications, process stability to ensure predictable on-time delivery, documentation and procedures that enable traceability and corrective actions when problems arise, etc. Methodologies that promote stable, high-quality manufacturing practices have been established in several industries. These standard procedures such as ISO9000, QS9000, and cGMP overlap to a large extent. QS9000 was developed to help automotive suppliers implement effective, well controlled processes to produce reliable products. It has several unique features, so it is used as the model in this chapter.

36.1.5 Environmental Shield

IC wafers are completely passivated before packaging in order to meet product stability and reliability requirements. Aside from ink-jet print heads, the active regions of all MEMS devices currently in high volume production replicate this hermetic barrier in some manner. This passivation requirement is noteworthy because it has been a primary limitation on the growth of the MEMS industry. For example, it limits electrochemical sensors to benign or "throw-away" applications.

36.2 Robust Manufacturing

Stable production requires a manufacturing flow that is well controlled when measured against the product performance specifications. This has several implications.

36.2.1 Design for Manufacturability

Low maintenance products and processes must be designed to run on standard equipment. The semiconductor industry has invested billions of dollars to develop equipment that is automated, maintainable, reliable, and capable of supporting well controlled processes. Processes that are implemented on custom-designed equipment put this experience base aside and invariably have a long learning curve.

Equipment

Microstructure release illustrates this issue and how it changes over time. Virtually all semiconductor processes operate in vacuum or at atmospheric pressure. In a brief, unpublished 1992 study related to ADXL50 development, the author demonstrated the use of supercritical CO_2 mixtures for microstructure release and particle cleaning. Unfortunately, supercritical processes were incompatible with the industry infrastructure, because they require pressures in excess of 7.38 MPa (72.8 atm; 1,070 psi). Implementation would have required process and equipment development, so a release process that ran on existing equipment was selected.

Supercritical equipment designed for MEMS release was introduced a few years later. However, maintenance and cycle time concerns, as well as single wafer capabilities limited its adoption to university and low volume manufacturing.

The intrinsic limitations of non-standard equipment have caused high volume MEMS suppliers to avoid supercritical processes. However supercritical equipment will continue to evolve and become competitive, particularly in applications where the design requirements make alternative processes unusable.

Materials

Process control can be no better than the materials used in the process. SAM (*s*elf-*a*ssembled *m*onolayer) coatings illustrate this limitation. These materials suppress stiction following aqueous release of microstructures by treating the wafers in a SAM solution before drying. Although straightforward in principle, high volume manufacturers have not (yet) adopted SAM processes (Sect. 36.2.3 reviews the processes that are used commercially). There are several reasons for this reticence. Classical SAM materials are chlorosilanes that have at least one organic substituent. Organosilanes have been used to treat the surface of inorganic materials for many years. Applications include coupling agents on the surface of fillers and reinforcing agents in polymers, as well as agglomeration control of particles. The chlorine sites hydrolyze when dissolved in solutions that contain a small amount of water. The resulting hydroxyl groups react with the microstructure surface oxide to produce a chemically bonded layer that reduces stiction due to the low surface energy and hydrophobic nature of the organic-rich surface. The concerns that cause manufacturers to use alternative stiction solutions include:

1. High variability due to the SAM chemical reactivity
2. Particles are common by-products of the reactivity
3. Many SAM process flows require organic baths. This raises health, safety, and waste disposal concerns
4. Possibility of chloride corrosion on aluminum interconnects

Maboudian et al. [36.17], *Srinivasan* et al. [36.18], *Kim* et al. [36.19], and *Pamidighantam* et al. [36.20] describe the use of SAM solutions to suppress stiction. There is also ongoing research focused on solving the manufacturing concerns [36.21–25].

Note that stiction also arises in packaged MEMS products. This yield and reliability problem is discussed in Sect. 36.3.1.

36.2.2 Process Flow and Its Interaction with Product Architecture

MEMS products have structures and functions that do not exist in standard IC devices, so it is unrealistic to expect that every fab, assembly, and test step will reapply a previously qualified IC process. The challenge is to maximize utilization of available IC technology within the constraints of the product function and cost requirements. This leads to fundamental choices in product architecture, product design, and process flow, as discussed below.

Integration of MEMS and Circuits

The integration of MEMS and electronics onto one chip has proven difficult. It adds little value in some MEMS products. However, as noted earlier, integration is essential to products based on the Texas Instruments (TI) DLP (Digital Light Processing$^{\mathrm{TM}}$). The third (and largest) category consists of applications where integrated and non-integrated products compete for market share. Pressure sensors and accelerometers are examples of this group.

Image Projection. The first projectors based on the DLP technology were shipped in 1996. DLP chips are produced on CMOS wafers in a mature TI process. The circuitry addresses and drives aluminum mirror arrays that function as on-off pixels – 1.3 million mirrors in the latest SXGA products (Fig. 36.2). *Mignardi* et al. [36.6] describe the DMD (Digital Mirror Device) fabrication process and illustrate some of the factors that must be considered when manufacturing flows are changed. The manufacturing flow patterns three aluminum depositions over SRAM cells that are positioned with 14 or 17 μm center-to-center spacing (Figs. 36.4 and 36.5). The first of these metal films form electrostatic drive electrodes and landing pads. The yoke and hinge layer has spring tips that land on the first metal layer when mirrors are rotated (Fig. 36.6). Mirrors are mechanically connected to the yokes through center pedestals. Sacrificial organic films separate the metal layers. These organics, along with a protective organic cover film, remain on the device until after the wafer is sawn and the chips are mounted in a ceramic package. They are removed by dry etching.

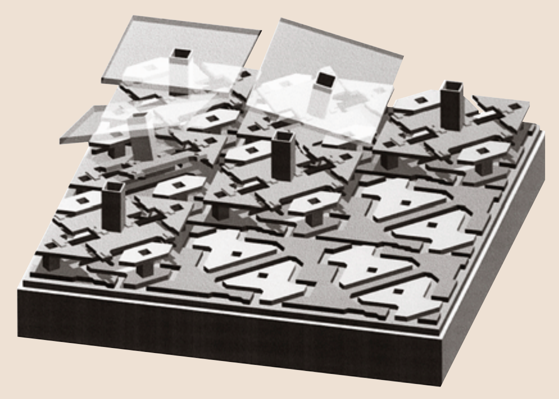

Fig. 36.4 DMD pixel array with tilted and non-tilted mirrors. (Courtesy of Digital Light Processing$^{\mathrm{TM}}$, Texas Instruments, Inc.)

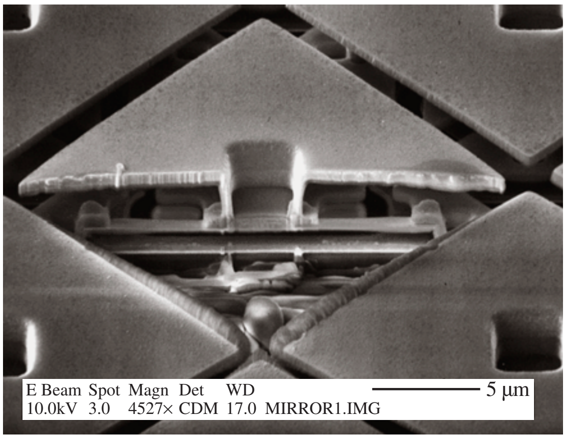

Fig. 36.5 Ion milled section of DMD pixel showing one mirror on its center support and its relation to the underlying layers. Note the close mirror spacing. (Courtesy of Digital Light Processing$^{\mathrm{TM}}$, Texas Instruments, Inc.)

Following surface passivation, inspection, and testing, an optical glass sub-assembly is hermetically sealed to the ceramic package (Fig. 36.7). At this point, the device is ready for burn-in and final test – a significant task because each mirror is actuated on chips with as many as 1.3 million mirrors.

In operation, a (nominal) 26 V electrode bias generates an electrostatic force that causes the mirror/yoke assembly to rotate either plus or minus 10 degrees around the hinge. Over one million DMD devices had been shipped by 2002 [36.26].

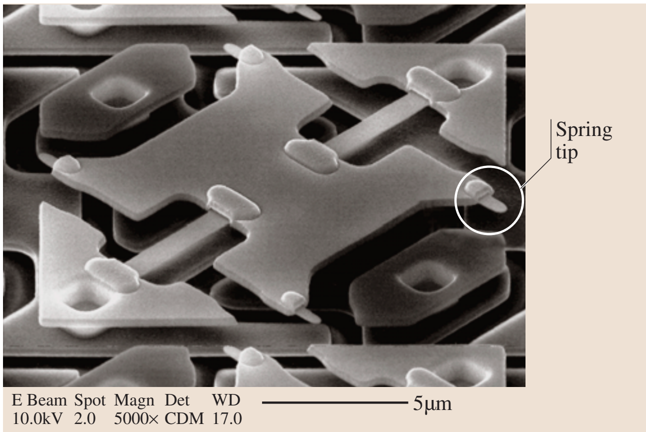

Fig. 36.6 DMD yoke and hinge layer. The spring tip touches landing pads when the drive electrode tilts the mirror. Elastic energy stored in the deformed spring is part of the restoring force that overcomes stiction when the drive electrode voltage is removed. (Courtesy of Digital Light Processing$^{\mathrm{TM}}$, Texas Instruments, Inc.)

Unlike the pressure sensors and accelerometers discussed below, TI had no choice but to integrate. It would be difficult, if not impossible, to devise a cost-effective multi-chip solution to control and drive a matrix of over a million mirrors. Successful integration of circuits and mirrors was only one step in the commercialization process. To remain competitive against lower cost LCD products, TI continues to develop their package and test technologies, because package and test comprise a major part of the total product cost. This is discussed further in Sect. 36.2.8.

Pressure Sensors. Early (1970s-era) bulk micromachined piezoresistive pressure sensors from Honeywell, ICT (later Foxboro-ICT), and Kulite were not integrated. However, in the 1980s, Motorola commercialized the

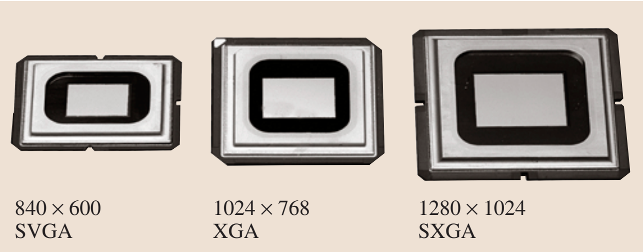

Fig. 36.7 DMD products in hermetic packages. (Courtesy of Digital Light Processing$^{\mathrm{TM}}$, Texas Instruments, Inc.)

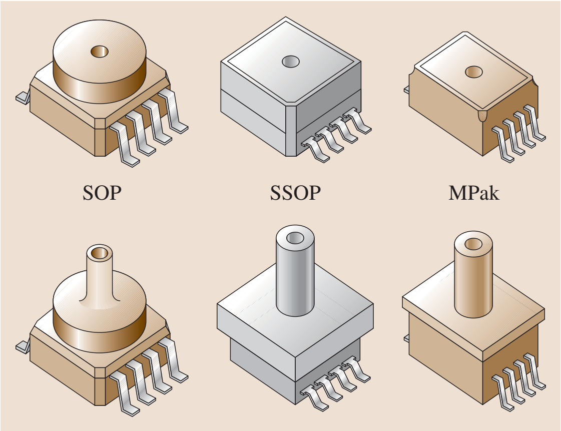

Fig. 36.8 Pressure sensor surface mount packages. (Courtesy of Semiconductor Products Sector, Motorola, Inc.)

first of its MPX5100 series piezoresistive pressure sensor products. These products had bipolar signal conditioning electronics and temperature compensation on the sensor chip. Current MPX products are offered in various plastic packages (Figs. 36.8 and 36.9) with maximum pressure ranges between 10 and 300 kPa (1.45–44 psi). Medical and tire pressure products are also available.

Integrated Motorola piezoresistive pressure sensor products are cited here for illustrative purposes, but capacitive and piezoresistive silicon pressure sensors are available from many companies. Capacitive designs are less temperature sensitive and require less power than piezoresistive designs. However, die size tends to be larger. The fabrication process and interface electronics are also more involved. Piezoresistive products (sensor plus signal processing) are less expensive, smaller, easier to manufacture and suitable for most applications. They rely on the fact that the resistance of silicon changes in response to strain. This effect is very temperature sensitive. However, configuring four piezoresistors as a Wheatstone bridge, judicious selection of the doping level and junction depth and the use of a temperature compensating resistor network minimizes temperature error.

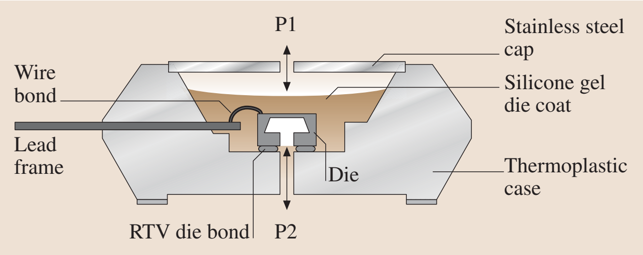

Fig. 36.9 Cross section of a typical gauge or differential pressure sensor. The silicon pressure sensor chip is bonded to a silicon substrate. This bonded unit is mounted in a pre-molded plastic package and protected from the ambient environment by silicone gel. (Courtesy of Semiconductor Products Sector, Motorola, Inc.)

Piezoresistive pressure sensors (Fig. 36.1) require double-sided polished wafers and front-to-back alignment. The piezoresistors are implanted into the front side of (100) wafers and precisely located with respect to the edges of thin diaphragms. The diaphragms are produced by anisotropically wet etching cavities into the back surface of the wafers. KOH etch solutions are usually used to avoid the safety, toxicity, and waste disposal concerns that arise when large quantities of organic etchants like EDP (Ethylenediamine-pyrocatechol-water) are used. Front-to-back side wafer alignment is critical in order to achieve precise resistor location with respect to the cavity edges. Alignment to the wafer crystal axis is also important, because the slow-etching (111) planes determine the cavity wall, and thus the location of the diaphragm edges. The piezoresistors are commonly placed near these edges to maximize the signal (the effect of pressure on diaphragm stress is greatest at the edges). An etch stop is sometimes used at additional cost because diaphragm thickness variations also have a large effect on sensitivity.

Control of mount stress is critical to achieving predictable performance. After wafer fab, some suppliers hermetically bond pressure-sensor wafers to a backup silicon wafer using glass frit. Others anodically bond them to a borosilicate glass wafer that has a thermal expansion coefficient close to silicon (Pyrex 7740 and Schott 8330 are two glasses used for this purpose). Absolute pressure sensors are produced when this bonding step is done in vacuum. As shown in Fig. 36.1 and Fig. 36.9, differential pressure and gauge pressure products incorporate through holes in the backup wafers. These ports allow fluid pressure to be applied through the cavities to the back side of the diaphragms.

After wafer singulation, the chips are mounted and sealed in cavity packages. Soft die attach materials, often silicone-based, are used to decouple the sensor from package and substrate stresses. In some products, the sensor chip is directly mounted to a substrate with a soft die attach without the intermediate backup wafer. Many package variations of MEMS pressure sensors are available, as summarized in Sect. 36.2.8.

The MEMS pressure sensor market is believed to exceed a hundred million pressure sensors annually [36.27]. Automotive applications like manifold absolute pressure (MAP) sensors and a wide range of gas and liquid pressure sensors form the largest market segment. Health care and medical uses such as disposable blood pressure transducers and sensors to measure pressure in angioplasty catheters, infusion pumps, and intrauterine products are the second largest market. Industrial products such as process control pressure and differential pressure transmitters, household appliance, and aeronautical products are also significant. *Maudie* et al. [36.28] review the performance and reliability issues related to appliance applications like household washing machines.

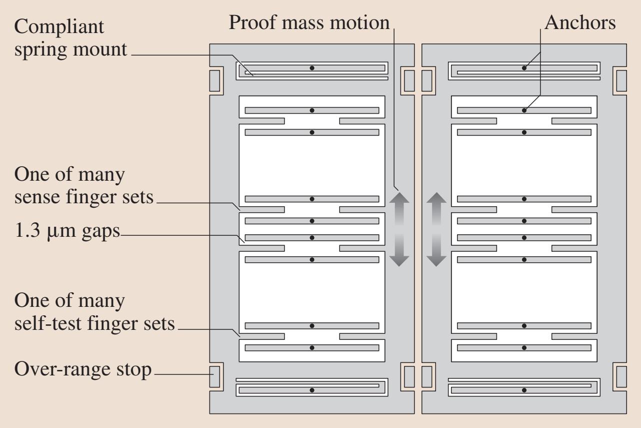

Fig. 36.10 Block diagram of the MEMS element in an ADXL78 single-axis accelerometer. The detailed design applies design-for-manufacturability principles to ensure close control of critical dimensions and residue-free removal of the sacrificial layer. (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

Accelerometers. Most MEMS accelerometer designs apply some variant of Newton's Law, $F = ma$, to sense the response of a proof mass. Many sensing principles, including piezoresistive, capacitive, piezoelectric, and resonant, have been examined [36.29]. Figure 36.10 has a conceptual view of the capacitive design used in Analog Devices ADXL accelerometers.

SensorNor was an early pioneer in silicon accelerometer manufacturing. Their product had a mass bonded to the end of a silicon piezoresistive element in an oil-filled package. However, the first MEMS accelerometer product to achieve large-scale market acceptance was the integrated ADXL50 air bag sensor from Analog Devices. When it was fully qualified in 1993, the ADXL50 was the first integrated surface-micromachined MEMS device of any type in production. The MEMS element in the ADXL50 was part of a closed loop differential capacitance circuit, but was designed to allow routine self-testing after the air bag module was installed in vehicles. This self-test capability eased concerns related to the adoption of a new technology in a critical safety application. The analog output feature also allowed the industry to design single point sensors – a significant system cost savings. *Core* et al. [36.2] and *Sulouff* [36.5] describe the ADXL wafer fabrication process. In essence, circuits are fabricated with a well-established BiCMOS process. The sensor polysilicon is deposited after these high temperature steps, followed by the lower temperature metal and passivation processes. The in situ doped polysilicon is connected to the circuit through N+ doped runners. Following wafer fabrication, thin film resistors are laser trimmed in an automated trim system to meet the specific end user's requirements. After wafer singulation (Sect. 36.2.5), the chips are assembled in hermetic cavity packages and screened in an automated test system to ensure compliance to the performance specification (Sect. 36.2.8).

Front air bag sensors have an output of 2 V at a full scale acceleration that is specified by the user (typically 35–50 g). Satellite air bag sensors are located in door pillars and near the front of the automobile and typically have higher ranges (250 g). Regardless of range, all air bag sensors must meet stringent cross-axis specifications, i. e., a front air bag sensor must not react to a side collision or to potholes in the road. This cross-axis requirement also applies to products that are designed to sense in two axes (Fig. 36.11).

Progressive reductions in chip and package sizes (Fig. 36.3) led to cost reductions and penetration into new markets. One example is the ADXL202 (Fig. 36.11). This two-axis, low-g accelerometer has a full scale of 2 g in each axis. It is widely used in industrial, consumer, and automotive applications. Production of integrated MEMS accelerometers at Analog Devices has grown rapidly in recent years (several million devices per month by 2001). It is worth noting that these products are developed using an organizational philosophy that emphasizes design-for-manufacturability. For example, development is conducted on production equipment and directly involves production personnel. This may appear to be expensive and unwieldy, however, it ensures that product designs remain within the bounds of practical man-

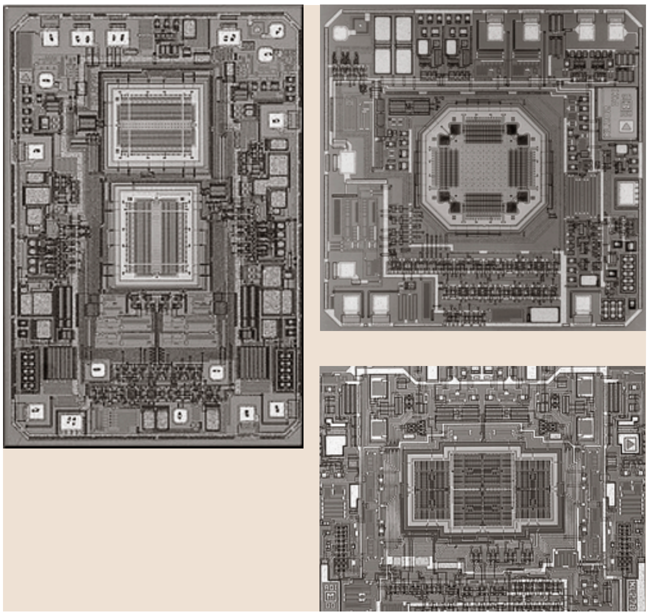

Fig. 36.11 Three two-axis accelerometers. *Top left:* ADXL276 was the first two-axis air bag accelerometer. The ADXL278 family (*bottom right*) is replacing it. *Top right:* 2-g ADXL202. The ADXL202 and ADXL278 are packaged in the LCC shown in Fig. 36.3. (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

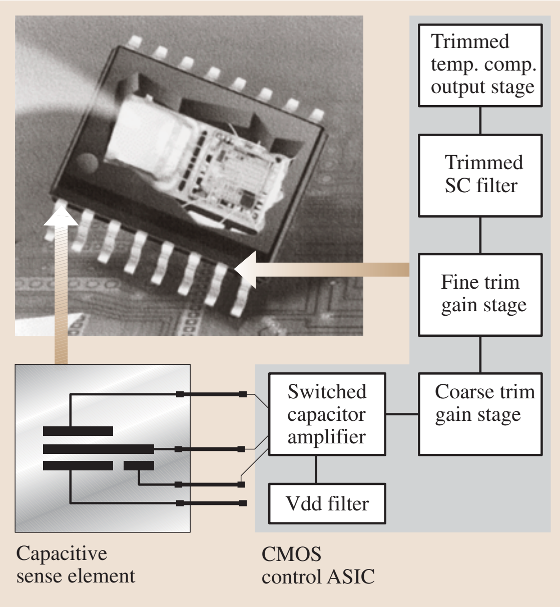

Fig. 36.12 Two-chip accelerometer in 16 lead plastic SOIC package. (Courtesy of Semiconductor Products Sector, Motorola, Inc.)

ufacturing. It also minimizes the difficulties that are normally encountered when new products are transferred to production.

Prior to the introduction of the ADXL50, most of the MEMS air bag candidates were nonintegrated, bulk-micromachined piezoresistive sensors [36.3]. However, Motorola and Bosch soon followed with an integrate-in-the-package strategy. Denso, Delphi, and VTI Hamlin are also active in some market sectors. The Motorola and Bosch products are nonintegrated capacitive MEMS products that are hermetically sealed at the wafer level (Sect. 36.2.4). By interconnecting a separate circuit chip and a sealed accelerometer chip in one package, they are able to use low cost CMOS for signal processing and near-standard plastic packaging (Sect. 36.2.8).

In contrast to the ADXL sensors, Motorola's MMA air bag products measure acceleration normal to the plane of the chip (z-axis). These differential capacitance sensors are formed from three layers of polysilicon. The middle layer is a "proof mass" that is suspended with ligaments and is free to move in response to an accelerating force. Ranges and outputs are similar to the ADXL products. Recent Motorola expansions add x-axis air bag products and a new line of low-g z-axis products (Fig. 36.12).

The Bosch x-axis sensor has interdigitated fingers that are patterned in a thick (11 μm) polysilicon film that is deposited in an epi reactor. Tight control of stress and stress gradients in these films is essential in order to meet sensor performance requirements [36.30, 31]. After patterning, the polysilicon is dry etched using fluorine-based chemistry [36.32] to form the differential capacitance sensor. An HF vapor process is used to remove the sacrificial oxide and release the microstructures. These thick structures have working capacitances near 1 pF [36.33]. *Laermer* et al. [36.34] explain how Bosch applies deep silicon etch processing to several production and experimental inertial sensors. Better known as the "Bosch etch," this deep reactive ion etching technique has pioneered the field of high-aspect-ratio silicon-based MEMS.

The Future of Integration. The high growth rate of integrated MEMS products demonstrates the value of integration in MEMS applications. To date, integration has been limited to large semiconductor companies that have in-house signal processing expertise and the willingness to invest substantial resources over a number of years. Each company identified a large volume market that could support a profitable integrated product. This "large semiconductor company" barrier-to-entry for integrated MEMS products will diminish in the future due to new SOI wafer technology (Sect. 36.2.4), new flip-chip equipment, and the development of low temperature MEMS processes. With these capabilities, MEMS-only companies will be able to integrate microstructures on circuit wafers that are preprocessed in CMOS foundries. To be successful, they will also have to develop cost-effective manufacturing and quality expertise (Sect. 36.2.9).

Risk

High risk is normal in research programs. However, the tremendous financial investments required to manufacture new products make the risk of failure a key consideration. This section discusses how the perception of manufacturing risk changes with time and is influenced by available knowledge and expertise.

Every development program makes fundamental product architecture decisions. These decisions can lead competing companies down different paths. For example, Analog Devices pursued chip-level integration, while Bosch and Motorola chose package-level integration. Why did these companies choose different paths? The answer is related to available resources, cost, time-to-market, and technical maturity. Analog Devices chose to integrate because of its expertise in process innovation and signal processing. This path did have risks. When the ADXL50 was designed, the fatigue life of polysilicon MEMS devices was unknown. Therefore, a closed loop circuit architecture was chosen in order to keep the proof mass motion under 100 Å. Polysilicon actually has excellent fatigue life. Once this was established, Analog changed their ADXL products to an open loop architecture that can detect capacitance changes as low as 20 zF (10^{-21} F). Most companies do not have the expertise to work in this regime, but the decision to integrate on one chip allowed Analog to make full use of its signal processing knowledge.

The two-chip Motorola and Bosch designs allow the circuit chip to be fabricated on any IC process. This product concept also removed the concern that integrating MEMS and circuits onto one chip might reduce overall product yield. However, it does require higher output signals from the MEMS chip in order to overcome parasitics that arise in the bond wires between the circuit and sensor chips. This requirement led Motorola toward large-area z-axis designs, while Bosch utilized its deep-etch process capability. To minimize cost and ensure reliability, both companies seal the sensors with glass at the wafer level. Sensor and circuit chips are then molded together in one plastic package. Each company built on expertise that it already had in order to minimize risk and cost. Motorola had in-house manufacturing expertise in wafer-level glass frit sealing of pressure sensors that are subsequently molded in plastic packages. Bosch had deep-etch process capability and the world's largest hybrid manufacturing plant, so screen printing and glass sealing were well-established. These considerations plus the concern that MEMS yield loss would undermine product yield contributed to their two-chip decision.

Semiconductor companies continually drive down cost and serve evolving markets by shrinking product size. Analog Devices believes that this long-term roadmap requires an integrated product. They chose to deal with one-chip risks, rather than package uncertainties. ADXL50 chips had been capped and molded in plastic long before the first qualified products were shipped in 1993 [36.35]. However, production-worthy capping equipment did not exist at that time. The effects of plastic package stress on yield and long-term device parametrics were also uncertain. By introducing the initial product in TO-100 hermetic metal packages, the risks associated with custom equipment and package stress were removed.

Reusable Engineering and Facilities to Achieve Economies of Scale

The IC industry produces many products on each process flow. Foundries and packaging subcontractors routinely apply this "economy-of-scale" principle. Only basic items like mask sets and bond pad diagrams are changed. The MEMS industry has been less successful in this area, although techniques like wafer-level laser trim and blowing poly fuses allow accelerometer suppliers to offer "specialty" products without losing the advantages of high volume production.

If only a few devices are produced in a capital-intensive facility, unit costs are high. This issue is particularly critical during the first years of product life when sales volume is normally low. One solution is to build the market by manufacturing the new product in an existing facility. The two-axis low-g ADXL202 (Fig. 36.11) was built using a new mask set with an existing air bag sensor process. Many ADXL202 consumer, industrial, and automotive applications are too small to justify the capital investment risk required to develop and produce a two-axis accelerometer, but use of an existing design and manufacturing base removed this constraint.

Market Dynamics Drive New Product Opportunities

In the 1980s, low cost and high reliability allowed MEMS pressure sensors to penetrate automotive applications like the measurement of manifold vacuum (MAP sensors). Once a technology is accepted, new opportunities usually develop. For example, tire pressure monitors with RF communication capability are now available in some car models. In the future, this type of monitor will be a safety feature expected by all consumers. Indeed, wireless communication capabilities will also be a common MEMS product requirement.

This example illustrates the fact that MEMS products are merely system components. They must support seamless integration into the larger system. MEMS opportunities in automobile safety systems will be affected by both communication protocols and interconnect technology (copper, fiber optic, and wireless). The growth spurt of the 1990s started with front-air-bag crash sensors located in the center module of automobiles. It has evolved to include satellite crash sensors behind bumpers, side-impact sensors in the B pillar (the vertical post between doors), impact sensors inside doors, rollover sensors, and vehicle dynamic control sensors to improve ride and handling. Each of these applications has unique characteristics that can favor one technology over another. For example, side-impact sensing systems must activate the air bags faster than front-impact systems. This response time characteristic can favor the use of pressure sensors that monitor air pressure inside door cavities – a good solution as long as the cavity remains sealed. An alternative is to place an accelerometer in the B pillar. These thin structural members have very little space, so z-axis accelerometers are sometimes easier to implement. Delphi and Motorola products have this characteristic. Analog's small x-axis accelerometers also meet this space requirement.

Gyros – New Products Produced with Preexisting Manufacturing Capabilities

Gyroscopes are transducers for rotational motion. They have a wide range of uses in platform stabilization and robotics, as well as automotive applications (dead reckoning backup for GPS navigation systems when satellite communication links are lost, skid control systems, and rollover detection). Unfortunately, the non-MEMS and resonant quartz gyros used in military and aerospace applications are rather expensive for many high volume applications. BEI Systron Donner tailored their high performance quartz gyro to meet automotive needs and has shipped several million gyros. Other lower cost gyro and angular rate products, like the Murata vibrating ceramic bimorph products, are used in cameras. However, as a group, these low cost products do not meet the performance required for many applications.

Analog Devices and Bosch reapplied their accelerometer capabilities to develop and manufacture low cost gyroscopes. *Geen* et al. [36.36] describe the Analog Devices product. *Funk* et al. [36.37] and *Lutz* et al. [36.38] describe the Bosch product. These three publications reference previous work and discuss the challenges of producing gyros that meet performance, reliability, and cost targets.

The Analog Devices ADXRS150 gyro (Fig. 36.13) illustrates the technology. It has a resonant microstructure that is integrated with two electronic systems on one chip. One of the electronic systems drives the MEMS resonator. When the ADXRS150 is rotated, Coriolis ac-

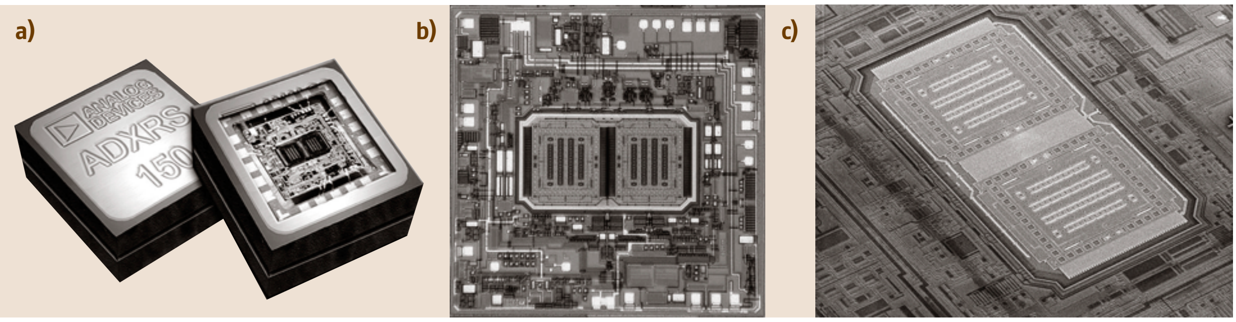

Fig. 36.13 **(a)** ADXRS150 gyro with integrated electronics in 7 mm × 7 mm BGA package. **(b)** ADXRS150 chip. **(c)** MEMS portion of the ADXRS150 chip. (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

celeration is generated in a direction that is perpendicular to the vibration axis of the resonator. This motion is detected by the second accelerometer system. Full scale is rated at 150 deg/s, although it can sustain overloads up to 10,000 deg/s. Integration of electronics on the chip, microstructure modeling, and predictable manufacturing were essential to success, because the full-scale Coriolis motion is only about 1 Å. The combination of differential capacitance detection, area averaging, and correlation techniques allow this motion to be resolved from thermal noise down to about 0.00016 Å [36.36].

Resonant gyros are not new. However, as noted above, their high cost restricts them to specialty applications. Many are non-MEMS, or utilize the piezoelectric properties of quartz, so chip-level integration is not an option. Most require expensive vacuum packaging to achieve high mechanical Q at resonance. Unfortunately, vacuum packaging also makes them susceptible to mechanical damage. Viscous damping in gas-filled devices like the ADXRS150 suppresses the destructive impact events that occur in vacuum-packaged MEMS products during shipping and handling and in-use conditions. Equally important are ADXRS150 design features that allow the measurement to reject mechanical shock, vibration, and other environmental noise sources. As a result, under-the-hood location is practical in automotive applications, an unmatched characteristic that significantly reduces system cost.

Analog Devices and Bosch both needed several years to develop their gyro products. Use of existing accelerometer infrastructure during the prototype and early production phases significantly reduced development costs and manufacturing risks. Early ADXRS150 gyros were sold as engineering grade products in socketable side braze packages. This allowed the development team and customers to understand the design and process parametrics that affect performance, reliability, and yield. Side braze packages are unacceptable in high volume applications, so the ADXRS150 is sold in a hermetic, ceramic, surface mount package.

36.2.3 Microstructure Release

Release stiction occurs when microstructures stick together after the sacrificial material is removed in a wet etch. It can cause considerable yield loss and must be considered in devices like capacitive pressure sensors and accelerometers that have closely spaced, mechanically compliant microstructures. Piezoresistive pressure sensors are not susceptible to release stiction, because these sensing diaphragms are not near other surfaces.

Release stiction is most frequently observed in silicon MEMS devices that use a sacrificial oxide. The oxide is removed by etching in wet HF, rinsed in deionized water, and then dried. Water promotes growth of a hydrophilic surface oxide. It also has a high surface tension. To minimize energy, the high surface tension in the shrinking water droplets causes the microstructures to be pulled together. When they touch, clean oxide surfaces stick, thus destroying device functionality. *Maboudian* et al. [36.17] review release stiction mechanisms, as well as the solutions that have been reported in the literature.

Some manufacturers do not publicly discuss their fab process. Those that do avoid release stiction by using gas phase processes to remove the sacrificial material. Texas Instruments uses photoresist, rather than silicon dioxide as the sacrificial material in their DMD manufacturing process. Two photoresist layers are under the aluminum mirrors. The wafer is also covered with a third organic layer to provide mechanical damage protection and allow for particle cleaning after the wafer is singulated into individual devices. By removing these organics in a dry etch process, the stiction problems that characterize wet etch processes are avoided.

Analog Devices avoids release stiction by dividing its accelerometer process flow into several steps, each of which is well controlled with standard manufacturing equipment. After the microstructure is formed, a few small channels are etched in the sacrificial oxide and filled with photoresist. These photoresist pedestals hold the microstructures in place when the remaining sacrificial oxide is etched, rinsed, and dried. The pedestals are then removed in a dry etch process. *Core* et al. [36.39] and *Sulouff* [36.5] describe this technique in more detail.

Bosch avoids release stiction by holding accelerometer wafers above an HF solution. HF vapors from the solution react with the sacrificial oxide, converting it to gaseous SiF_4 and water. The wafers are heated to prevent water from condensing. This process is noted in *Offenberg* et al. [36.40]. *Anguita* et al. [36.41] describe a similar process.

36.2.4 Wafer Bonding

To date, the growth of MEMS wafer bonding has been driven by three factors:

- Sealing microstructures by bonding wafers together addresses two sources of yield loss in MEMS manufacturing: wafer singulation and particles.

- Wafer-level mounting can be a cost-effective way to control stress in products that are sensitive to variations in mount stress.
- Fabrication of 3-D microstructures from elements that are formed on multiple wafers.

Piezoresistive pressure sensor wafers have been anodically bonded to borosilicate glass wafers for about 25 years. The earliest products were manufactured using little more than a hot plate and a high-voltage power supply. The era of custom built manual equipment has passed, because both Electronic Visions and Karl Suss now offer automated production-worthy tools for anodic, glass frit, organic, and silicon-silicon wafer bonding. *Schmidt* [36.42] reviews wafer-wafer bonding processes, while *Mirza* [36.43, 44] addresses MEMS wafer-level bonding applications and equipment.

Anodic Bonding

Anodic bonding is commonly used to seal glass wafers to the back side of bulk micromachined pressure sensor wafers. The process requires that a silicon wafer be placed in intimate contact with a glass wafer that contains a mobile ion at elevated temperature. An applied electric field causes mobile ions (usually sodium) in the glass to move away from the silicon interface toward the cathode at the other side of the glass. Bound negative charges remain in the glass near the silicon interface. These charges produce an electric field that pulls the wafers together and drives oxygen across the interface to anodically oxidize the silicon surface. Hermetic seals are routinely produced between flat wafers when particles are rigorously excluded. Process conditions depend on the glass composition and thickness, but 500–1,000 V at 400 °C with 10 min cycle times is typical. Borosilicate glasses with thermal expansion coefficients close to silicon are used to minimize stress. Anodic bonding promotes surface conformation so hermetic glass-silicon seals can be formed with surface grooves as deep as 50 nm [36.45]. This process has been demonstrated to work with many material combinations [36.46].

Glass Frit Bonding

Glass frit bonding reapplies techniques and materials developed for hybrid processes and cerdip sealing. It is often used to seal pressure sensor wafers to silicon substrate wafers. A second common application is to glass-frit seal bulk-micromachined silicon cap wafers over the microstructures in accelerometer wafers. In this process, fine glass powder is dispersed in an organic binder to form a paste. This paste is screen printed or stenciled in the desired pattern on a wafer. The wafer is then passed through a furnace (typically 430–500 °C). This causes the organics to burn off and the individual glass particles to coalesce into the desired pattern.

The wafer bonding process starts with a wafer prepared as described in the preceding paragraph. This wafer is aligned to a second wafer, where they are held in intimate contact and heated until the glass softens. Liquid glass readily wets the surface of the adjacent wafer when at least trace amounts of oxygen are present. The liquid-wetting mechanism produces hermetic seals even when the wafer surfaces are relatively rough. However, seal widths less than 100 μm are difficult to achieve. Fixtures that maintain wafer alignment through the bonding process are also required. Low-melting lead oxide and boron oxide glasses are commonly used in processes that typically run at 450 °C. Glass wet-out occurs quickly, but process cycle times are limited by the cooling rate (rapid cooling can cause stress gradients and cracking in the glass if proper design practices are not followed). Glass frit bonding to semiconductor wafers requires a good understanding of glass composition effects and process conditions [36.47].

Silicon–Silicon Wafer Bonding

Silicon-silicon bonding has been demonstrated in many IC and MEMS applications over the last 30 years. These processes bring highly polished wafers into intimate contact and are often promoted by a thin hydrophilic oxide that is left on the surfaces by the aqueous cleaning process. Clean, well polished wafers will bond at room temperature. However, heating (800–1,200 °C) increases bond strength by an order of magnitude. A fundamental difference between MEMS silicon–silicon bonding and either anodic or glass frit bonding is that it is used to produce complex microstructures that cannot be made in one wafer. In contrast, most anodic and glass frit bonding applications address assembly and packaging issues.

Recent silicon-silicon advances have solved technical factors like stress control that had limited MEMS applications. For example, Analog Devices balances the stresses at oxide interfaces and the handle wafer to produce flat, three- and four-layer SOI wafers [36.48, 49]. The thickness of each layer is precisely controlled using grinding and CMP (chemical-mechanical polishing). Intermediate layers can be patterned, electrically isolated, electrically interconnected and locally etched to produce single-crystal silicon microstructures that are hundreds of microns thick on CMOS compatible wafers. With this new capability, several organizations are fab-

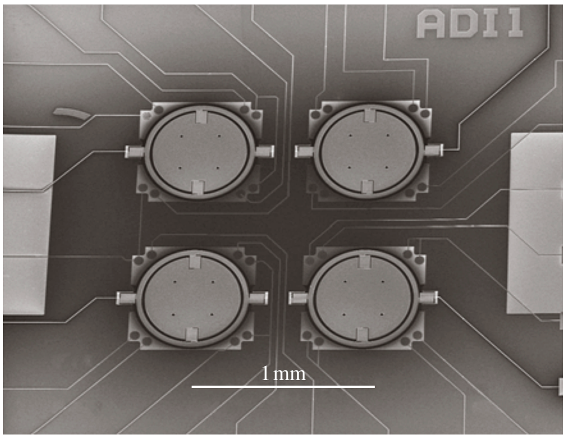

Fig. 36.14 Array of steerable mirrors designed for fiber optic network switching applications. The gimbal-mounted mirrors in this experimental product are electrostatically driven. Closed loop control based on capacitive sense electrodes ensures that the proper angle is maintained. High and low voltage circuits are integrated with mirrors on one chip. (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

ricating MEMS devices for biotech, automotive, and communications applications. One example is the array of gimbal-mounted mirrors shown in Figs. 36.14 and 36.15. These mirrors can be tilted in any axis to steer light between fibers in optical fiber bundles. The mirrors are driven by electrostatic force applied from electrodes that are patterned under the mirrors. Capacitance measurements based on a second set of electrodes are also part of a closed loop control system that maintains position control. Integration of high voltage drive electronics, low voltage control electronics, and MEMS mirrors would be impractical without this SOI capability and related MEMS processes.

36.2.5 Wafer Singulation

The cooling water and the particles generated in a standard IC diamond saw will destroy MEMS wafers unless some form of protection is used. Manufacturers solve this problem in several ways. Bosch and Motorola bulk micromachine cap wafers that are glass frit bonded to surface micromachined wafers. The caps protect the microstructures from water, particles, and mechanical damage so the wafers can be sawn with standard equipment.

The optical function of DMD chips required Texas Instruments to find a different solution for DMD wafers. Originally, they used a partial saw process on wafers that were protected with an organic film. This allowed wafer-level testing, but caused particle contamination and die loss when wafers were broken along the partial saw cuts. They now use a standard saw process, but do it before the mirrors are released. The wafers have the protective organic film over the mirrors, so normal cleaning processes are used. The singulated chips are

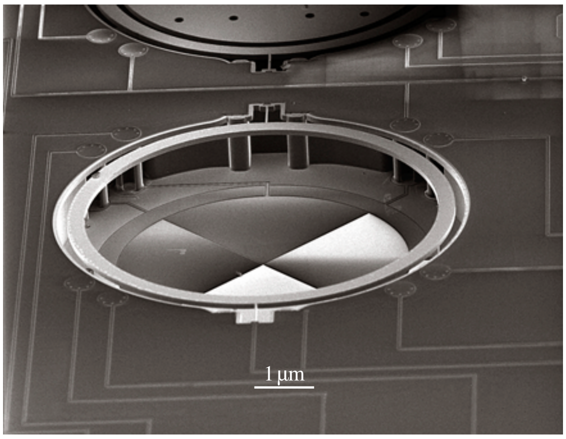

Fig. 36.15 View of the deep etched trenches, interlayer connections, and electrodes under a mirror of the SOI MEMS device illustrated in Fig. 36.14. (Courtesy of Micromachined Products Division, Analog Devices, Inc.)

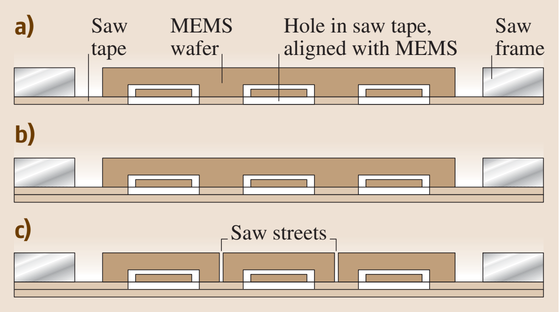

Fig. 36.16a–c Analog Devices "upside-down" saw process flow. **(a)** Saw tape is mounted on a saw frame. Holes are punched in the tape to match the microstructures on the wafer. The wafer is mounted-upside down on the saw tape and aligned so that the microstructures are in the holes. **(b)** A second layer of tape is placed over the first layer to form watertight pockets. **(c)** Wafer is aligned and sawn using standard equipment (adapted from [36.50])

mounted in ceramic packages before the organic layers are dry etched. This flow eliminates a primary source of particle contamination and problems associated with handling partially sawn wafers [36.6].

Analog Devices singulates uncapped MEMS wafers using an upside-down saw process and standard equipment and fixtures that are slightly modified (Fig. 36.16) (*Roberts* et al. [36.50]). A major attraction is that it allows wafer-level test and trim, so only chips that meet the product specification are assembled. It also avoids the added cost and yield loss associated with wafer capping. Although originally developed for accelerometers, this technology was extended to gyro wafers and to wafers that contain optical mirror arrays (see Figs. 36.13 and 36.14).

36.2.6 Particles

Particles cause yield loss in every product that has closely spaced movable microstructures. They also cause large area defects in anodic and silicon-silicon wafer bonded processes. Commercial MEMS suppliers use several solutions:

1. Cap the MEMS wafers before they leave the wafer fab clean area. Bosch and Motorola use this approach to produce accelerometer products (Fig. 36.12). It solves the wafer saw dilemma faced by most MEMS suppliers and allows the use of plastic packaging. Disadvantages include yield loss due to cap misalignment and stress, as well as the cost of the added capping steps. Hermeticity testing is also difficult because capped devices require sensitivity levels beyond the capabilities of standard analytical test equipment.
   Capping protects MEMS from particle contamination during assembly, but does not prevent wafer fab contamination. A greater concern is that glass-frit seal processes and equipment do not usually meet fab cleanliness standards. Once capped, inspection is essentially impossible.
2. Analog Devices and Texas Instruments assemble their MEMS products in clean rooms. This option increases capital costs. TI invested in environmental and process control, rather than develop optical quality capping for the DMD. Analog Devices had solved the wafer saw dilemma, so environmental and process control was the low risk path for particle control.

Particle control is central to high yield and reliability. The first control level is to minimize particle contamination through proper handling procedures, equipment maintenance, environment controls, and process design. The DMD protective organic film illustrates how process design can be used to control particles. This film allows the wafer to be cleaned after wafer saw and die attach.

Even with the best of controls, some particle contamination occurs in every clean room. Electrical tests in the automated wafer- and package-level test programs are designed to detect particles in Analog Devices products. Texas Instruments also finds particles in DMD products by driving each mirror in an array to ensure proper operation. Such screens are good but imperfect. Attempts have been made to replace visual inspection with automated particle inspection based on pattern recognition software. To date, however, machine vision systems have not been able to match the human eye and mind aided by a high quality microscope.

36.2.7 Electrostatic Discharge and Static Charges

Susceptibility to electrostatic discharges (ESD) led semiconductor manufacturers to make ESD avoidance a central criterion in equipment and fixture design, as well as handling procedures. Control of ESD events on the manufacturing floor is only the first step. Analog Devices and Texas Instruments incorporate ESD protective circuitry into their MEMS products as a standard practice to avoid failures in their customer assembly lines, as well as in the final system.

Most MEMS products use electrostatic force to actuate or control suspended microstructures. This design characteristic introduces performance parasitics and failure mechanisms that are not encountered in normal IC devices. For example, charges that build up on a dielectric surface can deflect unshielded microstructures, even when they are several millimeters apart. Such effects are insidious, because closed loop control systems do not always correct for electrostatic-induced errors.

Section 36.3.1 discusses surface treatments designed to suppress stiction caused by surface forces. These treatments often produce a dielectric surface. If the dielectric properties of the treatment and the device design cause the surface to hold a charge, that anti-stiction coating may actually promote electrostatic stiction! Thus, any “solution” must be critically evaluated in order to identify and remove undesirable side effects.

36.2.8 Package and Test

IC packages are environmental (mechanical, chemical, electromagnetic, optical) barriers that protect the chips from surrounding media. Only power and electrical signals and inertial forces pass unimpeded, so most MEMS sensors and actuators require that this barrier be selectively penetrated. Package and test functions are unique to each company's design and MEMS product type. Standard semiconductor packaging equipment and processes are modified when possible. However, even the line of demarcation that separates wafer fab from packaging is unique to each product.

High volume MEMS testing is even more challenging than MEMS packaging. In essence, the task is to measure and trim the response of products mounted in non-standard packages to calibrated stimuli over a range of temperatures. Automated IC test equipment does not have the calibrated pressure, acceleration, or optical test functionality required for this task. *Maudie* et al. [36.51] detail the test system elements that are required to support volume production of MEMS products.

Pressure Sensors

Pressure sensors are extremely susceptible to stress, so die-attach stress variations broaden performance distributions. The uniformity produced by wafer-level bonding is a primary reason why these processes are used to make the first-level package in many pressure sensor products. Soft silicone and fluorosilicone rubber products are often used to attach the sensor chip to the package to maximize isolation from mechanical stresses transmitted through the mount.

Most semiconductor products are assembled in plastic packages. The automated plastic presses transfer viscous epoxy compounds into multi-cavity molds at 175 °C and pressures near 6,900 kPa (1,000 psi). The hot liquid plastic creates high shear stress as it flows around devices in the mold cavities. Further stresses develop as the plastic hardens and cools. Even if this technology were adapted to create pressure ports in packages, the mold stresses would substantially affect pressure sensor performance.

Pre-molded plastic packages avoid the stress problem and incorporate pressure ports (Figs. 36.8 and 36.9). Polyphenylene sulfide (a high temperature thermoplastic) and medical-grade polysulfone are commonly used to make pre-molded pressure sensor packages.

The liquid or gas that is being measured can cause corrosion of interconnects and chip metallurgy, as well as parasitic leakage paths in the sensor. For that reason, a barrier such as silicone or fluorosilicone gel is often applied over the sensor. The soft gel transmits pressure with high fidelity over a wide temperature range and passivates the chip against many types of chemical attack. Parylene is also used for this purpose. This vapor-deposited organic coating is much stiffer than the gels, so thickness is typically controlled near 1 μm.

Petrovic et al. [36.52] tested the ability of fluorosilicone gels and several Parylene C thicknesses to protect powered pressure sensors against automotive and white goods benchmark liquids. One purpose of their study was to propose and demonstrate a formal media compatibility test protocol for pressure sensors similar to the IC industry standard tests. They found that the lifetime of sensors coated with both gel and parylene was considerably longer than sensors coated with either gel or parylene alone. *Petrovic* [36.53] later summarized the advantages and limitations of gels, parylene, and other techniques.

Figure 36.8 has examples of commercial plastic packaged pressure sensors. Pressure calibration equipment and software must be custom designed and built for these products. However, most pressure products are dimensioned to match IC packages in order to maximize compatibility with standard IC test handlers.

The harsh environment of some under-the-hood applications has been used to justify the high cost of metal cavity packages. For example, Bosch introduced a piezoresistive MAP sensor with signal conditioning that was hermetically sealed in an evacuated T08 header [36.54]. In-package trim was based on planar thyristors that were zapped as needed to bring the sensor within spec. The design avoided chip metallurgy corrosion and surface electrical parasitics by using the back surface of the silicon diaphragm as the fluid interface.

Much more expensive packaging is used in the pressure and differential pressure transmitters that are designed for industrial process control. These products seal the sensor in silicone oil behind thin metal diaphragms. Such products can measure pressure differences of 20 kPa superimposed on a pressure of 20,000 kPa (3,000 psi) over wide temperature ranges with high accuracy and stability. Soft die attach materials decouple the sensor from thermal stresses that arise when they are mounted in metal housings. Note, however, that silicone die attach materials cannot be used if the transmitter is filled with silicone oil. Fluorosilicones are one alternative. Piezoresistive sensors have been used in pressure and differential pressure transmitters since the 1980s. *Fung* et al. [36.55] describe a relatively new version. This product uses a piezoresistive polysilicon sensor to measure both absolute and differential pressure on the same chip. See *Chau* et al. [36.56] for early work on a multirange version of this technology.

Image Projection

The Texas Instruments DMD poses unusual challenges because the package lid must be an optical quality glass with anti-reflection coatings to improve optical performance and reduce heat load. The glass lid has opaque borders around the image area to minimize stray light effects and create a sharp edge on the projected image. It is fused to Kovar frames to produce lid sub-assemblies. The package base is a multilayer alumina substrate with co-fired tungsten to provide the electrical interconnects. The package sidewalls are formed by brazing a Kovar seal ring to the substrate. Heat dissipation through this substrate is a significant consideration because the service life of DMD mirror hinges (Fig. 36.6) is largely determined by operating temperature [36.57]. Excessive high temperature creep results in "hinge memory" and would cause a gradual drift in mirror orientation.

The DMD chips are die attached in the package cavity before the mirrors are released. A dry etch process is used to remove the organic sacrificial layers. The die are then passivated and tested. Getter strips are attached on the inside surface of the glass at the sides to control vapor composition in the package cavity before the lid is aligned and seam sealed to the package seal ring. Image quality requirements place stringent requirements on handling, alignment, and spacing tolerances. Thermal stresses also arise when dissimilar materials are joined in high temperature processes to produce the final hermetic package. The DMD packages (Fig. 36.7) and the related processes are discussed further in *O'Connor* [36.58], *Mignardi* et al. [36.6], *Bang* et al. [36.59], and *Poradish* et al. [36.60].

DMD mirror release does not occur until after die attach. This has advantages with respect to wafer saw and particle suppression, but it sacrifices the economic attraction of wafer-level testing. There is no commercial test system capable of combining CMOS electrical testing with 100% testing of optical mirrors. Therefore, a custom electro-optic test system was built around an x-y-θ translation stage with a CCD camera, light source, control hardware and software, and test programs [36.6]. A characteristic of the papers published by the DMD group is their effective use of this test system to examine problems and statistically validate the solutions.

Assembly of the DMD on the projector electronics board requires alignment with the system optics, in addition to electrical connections. Initially, the DMD was held with a plastic clamp and electrically connected through elastomer pads that had alternating layers of conductive and nonconductive material. This allowed easy replacement if a DMD was damaged during assembly. However, the impedance of this connector system was too high for new, higher speed products. Electrical intermittents were also observed. Therefore, it was replaced [36.61] by a grid of c-shaped springs (cLGA$^{\mathrm{TM}}$, Intercon Systems, Inc., Harrisburg, PA).

The DMD has about 40% of the business and entertainment image projection market [36.62]. Products based on liquid crystal technology serve the balance of the market. Market share is largely driven by price, so considerable effort has been placed on reducing the cost of DMD package and board-level assembly. *Migl* [36.61] reports on the assembly benefits achieved by replacing the epoxied heat sink with a mechanically attached heat sink. *Jacobs* et al. [36.62] describes the effort to replace the seam sealed window mount with a lower cost epoxy-bonded design. In theory, such a bond is not hermetic. However, the team realized that proper material selection, design, analysis, and use of moisture getters would produce a low humidity package through the product life. This change was not released to production. If it is implemented, success will require that adhesive bond integrity also be maintained, because the adhesive joint sustains thermal expansion and mechanical clamp stress cycling each time the projector is used.

Accelerometers and Gyros

The Bosch and Motorola capping processes are a first-level package that is applied at the end of wafer fab. After capping, the wafers can be sawn using standard equipment and do not require clean room assembly conditions. In principle, capped devices are compatible with standard plastic packaging. However, molding stresses can be a serious problem. To minimize stress, capped sensor chips are mounted on lead frames with a soft elastomer and often coated with a silicone gel to isolate the sensors from the package stress. Motorola products are calibrated by burning EPROMS after molding. Their early air bag accelerometers were molded in DIP and SIP packages for assembly on through-hole circuit boards. The shift to smaller surface mount SOIC packages reduces the quantity of plastic, so package stresses are reduced. To maintain compatibility with standard tooling in small thin plastic packages, either the cap or the MEMS wafers must be backlapped.

Analog Devices uses a near-standard saw process (Sect. 36.2.5 and Fig. 36.16), so capping is not required if hermetic cavity packages are used. Such packages are standard, but more expensive than molded plastic. A major attraction of cavity packages is that they eliminate plastic package stress (see page 1099), because the only mechanical connection between the chip and the package is through the die attach and bond wires. As a result, ADXL accelerometers and ADXRS gyro products are fully tested and trimmed on automated systems before the wafers are sawn. A lot-tracking system transfers this data to automated die attach systems. These systems are programmed to pick only chips that meet performance specifications from the wafer saw film frames. Thus, reject die are not assembled into packages. After packaging, devices are tested to ensure conformance to specification. However, package-level trim is not required, because the assembly process and cavity packages do not appreciably shift device parametrics.

The evolution of ADXL cavity packages is illustrated in Fig. 36.3. Initial products were packaged in TO-100 metal packages. This seam-sealed package is useful for development purposes. However, it is expensive and incompatible with the automated equipment used to assemble electronic circuit boards. Therefore, the early air bag sensors were soon switched to cerdips. Cerdips are made from low cost, molded, ceramic bases that have glass seal surfaces. The high process temperature requires use of a silver-glass die attach product. The full assembly flow includes two or three furnace passes near 450 °C to produce hermetic cavity packages.

IC products have been packaged in cerdips for decades, but through-hole circuit boards are seldom used today. A simple change in lead forming allows cerdips to be fully compatible with standard surface mount boards ("cerpacs"), so cerdips and cerpacs are assembled on the same equipment. Most of the ADXL products shipped in the late 1990s were packaged in cerpacs. However, small LCCs have become the package of choice for one- and two-axis accelerometers. The LCC (Leadless Chip Carrier) uses solder-sealed ceramic bases and metal lids. Organic die attach materials are practical, because the furnace gas is nitrogen, rather than air, and temperatures are about a hundred degrees lower than cerpac furnace temperatures. ADXRS gyros are assembled in a solder-sealed 7 mm × 7 mm ceramic ball grid array package (Fig. 36.13).

The fundamental message of this section is that IC package technology – both plastic and hermetic – is evolving very rapidly. Cost-competitive suppliers must remain cognizant of these trends in order to use them to best advantage. MEMS products that are packaged in a way that is not compatible with standard IC equipment and interfacing standards bear a significant cost premium. High package cost seriously limits market size because the price-volume curve is steep for most products. The package size trend is also critical to new market penetration. In summary, customer interface and new application requirements will continue to drive down both the cost and the size of most MEMS products.

The ADXL and ADXRS automated testers and handlers measure multiaxial linear and rotational acceleration. Analog Devices has a division that designs and manufactures automated test systems (many of the corporation's IC products have unique test requirements), so it avoided some of the problems associated with the procurement of custom-automated test equipment. MEMS suppliers have relied on custom-designed test equipment, but this situation is gradually changing. For example, Multitest GmbH now makes handlers with integrated shakers that are specifically designed for accelerometer testing.

36.2.9 Quality Systems

Quality systems are an intrinsic part of stable MEMS production. Management must drive a systemic approach to quality and set continuous improvement as a high priority goal.

Several quality systems define continuous improvement methodologies. In general, they formalize the process used to minimize defects and variations in products and in the processes used to manufacture them. Automotive supply companies must implement quality systems that meet QS-9000 [36.63] before product volume ramps up and continue them through the product life cycle. Four elements that are worthy of note:

1. *Failure Mode Effect Analysis* (*FMEA*). Early in the development phase, a team with representatives from several disciplines reviews the design (or process) in order to identify possible causes of failure. Each potential failure mode is given three numerical rankings. One ranking represents the likelihood of an occurrence, while the others rank the severity of that result and detectivity. These numerical scores are combined for each potential failure mode in order to identify the issues that merit the most attention before they become problems. FMEA spreadsheets are periodically updated as the product or process moves into production. More targeted versions are used in products or processes that are in stable production.

2. *Process Control*. Every significant process must have a "short loop" monitor to ensure stability. This measurement is tracked and statistically analyzed with respect to the control limits to identify changes before the process strays beyond the control limits.
3. *Review Boards*. Every change, unusual occurrence, and proposed solution is assessed by boards that meet on a regular basis to ensure that it does not put product quality at risk. Affected production material is put aside until the appropriate board approves its release. This may appear to be expensive and bureaucratic. However, the cost of a scraped wafer lot is insignificant when compared to the cost of a field replacement program.
4. *Procedures and Specifications*. Each step must be fully specified and identified in the process flow, along with the appropriate metrics for each lot. The compilation of this data in a retrievable form is an essential part of every continuous improvement program. By combining these product and process databases with lot tracking software, Pareto charts linking yield loss and test failures to process variations can be identified and eliminated. Experience has shown that unexpected second-order effects are present in every production line – but data-driven decisions on indirect effects cannot be made on small test populations.

Implementation of a continuous improvement program requires well-informed failure analysis teams, methodologies, and programs. Such analyses start with gathering facts and making relatively simple tests. Often this is sufficient. However, MEMS products are susceptible to uncommon failure modes, so *Walraven* et al. [36.64] gathered examples of more powerful analytical techniques and showed how they are applied to MEMS devices.

Each supplier uses the information and insights gathered from product performance, as well as control and yield data at different points in the process flow to refine their operations. *Douglass* [36.65], for example, outlined the yield loss and failure mechanisms observed in early Texas Instruments DMD products. Concerns like hinge memory, hinge fatigue, particles, stiction, and environmental robustness were each addressed and mitigated by focused teams. For example, *Mignardi* et al. [36.6] describe the partial wafer saw process used in the original DMD manufacturing flow process. Breaking these delicate wafers into individual product chips generated particles that caused yield loss and were potential sources of field failures. A new flow based on a standard full saw process eliminates this particle source. The result was an increase in both yield and reliability. Hinge fatigue characterization showed that bulk metal fatigue models do not properly describe thin film behavior. The hinge memory effort required an understanding of how thin film metal creep is affected by alloy composition and the environment [36.57]. Stiction control involves surface passivation, spring design, mirror dynamics, and moisture level [36.26, 65]. These publications give an insight into the quantities of data and the time required to bring robust products to market. They also illustrate how data-based evaluations can uncover unexpected effects like the acceleration of hinge creep by adsorbed moisture [36.26].

The second continuous improvement example applies to MEMS integration. High yield loss is perceived to be a serious risk when electronics and MEMS are integrated on one chip. Indeed, the initial yields on the Analog Devices integrated MEMS accelerometers were not impressive. However, it increased each year due to the work of many mission-oriented teams. Yield has reached defect density limited levels because the teams eliminated all significant failure mechanisms.

Coincident with the annual yield increases were reductions in customer failure rates. Failure rates in Analog's accelerometer products are in the low single digit ppm range. Quality does pay.

36.3 Stable Field Performance

Some topics discussed in Sect. 36.2 like particles and ESD are equally relevant to long-term stability. Most microstructure products have elements that are in close proximity, so stability is also affected by surface characteristics and mechanical shock.

36.3.1 Surface Passivation

Early, 1970s-era piezoresistive pressure sensors were not passivated. Like early ICs, these wafers only had oxide over the piezoresistors. Performance was inconsistent and drifted over time, because surface interactions with moisture and other atmospheric gases created a variety of shunts, parasitics, and charging issues.

Electrical Surface Passivation

The introduction of silicon nitride passivation and conductive field plates led to stable products that were hitherto unobtainable. The ability to achieve measurement stability was absolutely critical to growth of the MEMS pressure sensor industry.

Aside from pressure sensors, few MEMS products are passivated. Indeed, standard plasma nitride processes are usually impractical, because they produce dielectric coatings that support static charges. Such charges can cause electrical drift, or be the source of electrostatic forces that cause stiction.

There have been published reports attributing instability to the lack of passivation on microstructures. For example, Analog Devices implemented a special process in ADXL50 accelerometers before the product was released in order to suppress high temperature electrical drift [36.66]. This was not a moisture effect, because it occurred in dry nitrogen packages. Later ADXL designs eliminated the root cause.

Scientists at Lucent Technologies [36.67] observed anodic oxidation in polysilicon electrodes used to electrostatically drive mirrors in optical cross-connect products. If allowed to occur, this corrosion mechanism would cause the mirror position to drift and the product to eventually fail. The study concluded that, within the limits of the test, maintaining low moisture in these optical packages eliminates anodic oxidation.

The preceding paragraphs suggest that lack of MEMS passivation can affect long-term product stability. This issue is very design-related and can be driven by factors other than humidity. Even when low package humidity is determined to be adequate, design reviews should consider how normal manufacturing variations, outgassing and diffusion over the product life affect the moisture level in the gas adjacent to the microstructures.

Mechanical – Stiction and Wear

In-use stiction is difficult to predict and may not become an issue until manufacturing volumes increase, or when endusers handle the part in ways that are not anticipated [36.68]. The resulting product liability and field replacement programs carry substantial financial costs and have caused MEMS suppliers to withdraw from the market. End users have little tolerance for failure. Since in-use stiction is a failure mechanism in MEMS devices, a company that solves this problem creates an effective market barrier against competitors.

The scale of this problem varies between product areas. For example, the Texas Instruments DMD has moving mirrors that are designed to touch down on a substrate and later release. This product must address both stiction and wear. In contrast, the accelerometer products from Analog Devices, Bosch, and Motorola have proof masses that are suspended on compliant springs. By design, an automobile collision only displaces the proof mass by a few percent of the gap that separates it from adjacent surfaces. It is difficult to generate a shock wave in an air bag module that is sufficient to cause these proof masses to contact adjacent surfaces (acceleration levels of several thousand g are required). However, handling of the discrete packaged parts during test, shipment, and module assembly frequently cause shock events that bring MEMS surfaces into contact. For example, *Li* et al. [36.69] used both math modeling and tests to show that dropping a packaged MEMS accelerometer from the height of a table top generates several tens of thousands of g in the device when it lands on the floor. Thus, noncontacting MEMS products must be designed to withstand at least transient stiction.

Design. When shock, vibration, or functional operation causes MEMS elements to touch, the mounting springs are designed to pull them apart (Fig. 36.10). High stiffness springs are desirable to suppress stiction, but low stiffness springs increase measurement sensitivity and signal-to-noise ratio. Thus stiction is a fundamental MEMS design constraint that often limits product performance. Each manufacturer has proprietary design practices to address this performance versus reliability trade-off. In general, designers suppress stiction by minimizing the sources and applying supplementary techniques to ensure recovery when contact occurs:

1. Thorough analysis and testing to move harmful resonances beyond the range where they might be excited. Note that MEMS component-level analysis is insufficient – this is a system issue based on the packaged part as it is handled through manufacturing, system assembly, and end use.
2. Surface modification (discussed in the next section).
3. Minimizing contact area by integrating bumps and "stoppers" into the design [36.70].
4. Elimination of dielectric surfaces that may accumulate surface charges and result in electrostatic attraction.

5. Release of DMD mirrors from the touchdown position is assisted by pulsing the reset voltage to excite a mirror resonant frequency [36.65, 71].
6. Use of gas-filled packages to reduce contact velocity, as discussed later in this chapter.
7. ESD-protected designs that prevent voltage transients that may cause electrostatic attraction.

The only product in significant production that contains MEMS elements that are designed to touch is the Texas Instruments DMD. The proprietary aluminum alloy springs used in the mirror contact points touch and release millions of times during the product life. Considerable development was required to overcome metal creep that occurred in early designs.

Surface Modification. In addition to the touchdown springs, early DMD die were treated with perfluorodecanoic acid vapor before the glass lid was sealed onto the package. This treatment [36.72] created a low energy monolayer coating that suppressed initial stiction. However, improvements were required to achieve longer wear life, higher production rates, and lower contamination levels. A new process [36.73] places capsules inside DMD packages. After cleaning, the packages are immediately sealed. An oven bake releases anti-stiction vapor from the capsules to create a monolayer organic coating on surfaces inside these sealed packages. Excess anti-stiction material in the capsules maintains a low vapor pressure in the package cavities throughout the product life. Therefore, wear damage at the mirror contact springs is continuously repaired to prevent stiction failures before they occur.

Maintaining a low level of organic vapor inside the DMD reapplies a concept used with other organics to prevent corrosion of electronic metallurgy and contact fretting of separable connectors (see, for example, [36.74, 75]). These treatments have been shown to add years to the useful life of electronic systems that are in enclosures but also exposed to aggressive environments.

Stiction can be minimized by design and process spin-offs – stiff springs, control of contact area, etch residues, etc. Some suppliers design within these limits; others seek to modify the MEMS surfaces. The requirements of any surface modification technology are quite substantial. In addition to normal manufacturing requirements (stable, scalable process and materials), the process must be conformal and must uniformly treat all areas, including microstructure bottom and side surfaces. If a conductive material is used, it must somehow be patterned in order to avoid electrical shorting. If a dielectric surface is formed, it must be extremely thin because thick dielectrics support surface charges. Such charges cause performance shifts and induce electrostatic stiction.

Since devices with movable microstructures are particularly susceptible to stiction, most suppliers who develop production-worthy surface modification processes consider this information a trade secret. Analog Devices chose to patent portions of its technology. The following paragraphs give an overview of the published Analog Devices surface treatment technologies. *Martin* [36.66] describes the evolution of these technologies from 1992 to 1999. Further information is contained in *Martin* et al. [36.76] and *Martin* [36.77].

The earliest of the published Analog surface treatment technologies was implemented after stiction was observed during automated testing and module assembly of air bag sensors that were packaged in cerdips. Cerdip assembly typically includes at least two furnace processes in air at 440–500 °C, so they seldom contain organic materials. The high temperature furnace air thermo-oxidizes adsorbed materials and removes them from the MEMS surface. The resulting surfaces are extremely clean. Unfortunately, clean inorganics have high surface energies. Therefore, adjacent microstructures readily stick together if shock and vibration during handling cause them to touch. Cerdips (and the surface mount cerpacs) have very low moisture levels (Mil specs allow up to 5,000 ppm moisture, but < 100 ppm moisture is common in a well controlled manufacturing line). Deliberately raising the moisture level was found to be a potential solution because, within limits, moisture adsorption reduces surface energy. However, the solution adopted by Analog Devices in the mid-1990s was the anti-stiction agent (ASA) process. This process involved dispensing a controlled amount of a pure siloxane liquid (ASA) into each package immediately before sealing. ASA selection was based on thermo-oxidative stability, low volatility, purity, and the requirement that it be a liquid at the dispensing temperature. As packages heat in the furnace, adsorbed material on the microstructures is removed, leaving chemically active, high energy surfaces. Further heating volatilizes the ASA. Before escaping into the furnace, the ASA vapor surrounds the reactive MEMS surfaces and chemically bonds to them. The result is infusion of organic groups into the native oxide surface. Organic-rich surfaces are unreactive and have low surface energy. Thus the treatment is self-limiting after the first vapor molecules react. Obviously, ASA volatilization must be complete before the glass softens and seals the package cavity. Some

ASA degrades in this process. However, silicone thermo-oxidation simply adds a few angstroms of silicon oxide to the native oxide, so the ASA process, by design, is non-contaminating.

Although variations and refinements of the ASA process were developed, this technology required that each individual package be treated. Automated dispense equipment made ASA acceptable when product volumes were moderate. However, as production approached several million per month, it became evident that a wafer-level anti-stiction process was desirable.

Assessment of candidate processes and materials led to a vapor treatment based on a custom-synthesized solid polymeric siloxane. This polymer is volatilized in a standard CVD furnace that treats up to a hundred wafers at a time. Repeatability is extraordinary because thickness variation within a furnace run, or from run to run, is only about one angstrom.

Package Environment Effects. The viscosity and density of the gas that contacts MEMS resonators limit the device Q, so resonant devices are often designed to operate in vacuum. Unfortunately, this makes them susceptible to handling damage. Viscous fill gases reduce destructive excursions of high-Q microstructures when they are mechanically shocked. They also provide squeeze film damping to cushion the contact event as surfaces approach each other. None of the high volume MEMS products on the market are evacuated. Presumably, this is due to the high cost of vacuum packaging and the susceptibility to handling damage.

Many investigators believe that suppression of in-use stiction requires low humidity in the package. This view is not surprising because most MEMS research groups have observed stiction caused by aqueous surface tension during microstructure release. However, in-use stiction and wear test results published by several investigators lead to a different conclusion [36.76, 78–83]. In general, they find high stiction in both dry and moist environments. However, stiction is substantially suppressed at intermediate humidity levels (approximately 15–40 % relative humidity at room temperature). This is equivalent to a moisture level of 3,000–10,000 ppm. Note that the limits cited by different authors vary considerably due to the use of different materials and test conditions. Stiction at low humidity is probably caused by high surface energy on inorganic oxides that do not have adsorbed surface films. The low stiction and wear rates observed at intermediate humidity are attributed to the passivating and lubricating effects of adsorbed water. Capillary forces caused by the high surface tension of water become dominant at high humidity.

36.3.2 System Interface

MEMS products are mounted in systems that have much greater mass than the device itself. For this reason, it is difficult to transmit a mechanical shock wave to a MEMS sensor that is sufficient to cause failure after it is installed in a module. However, failures are possible when the devices are handled during module assembly. Improper functional response is also possible if the module mount system has a mechanical resonance that amplifies or reduces mechanical transmission under different conditions. This is a particular concern in automotive safety applications because every automobile platform has a different mechanical signature. Proper module mount design is required to ensure that the impact signal from an automobile crash is properly transmitted to air bag modules.

Suppliers of large IC die closely examine board-level stresses in order to minimize solder joint fatigue. Such stresses are more serious in MEMS products because the device, as well as the package are susceptible to mounting conditions. Furthermore, customer processes impose these stresses, so a supplier may have little information with respect to their origin or existence. One solution is to define and qualify a product-specific mount system, as Texas Instruments has done with the DMD. The ideal, however, is to devise products that are compatible with standard board footprint and assembly processes. This challenge became more difficult with the shift from through-hole to surface mount technology, because the mechanical isolation provided by the package pins is no longer present. Small package size, in-package, and within-chip isolation techniques, as well as thorough package-board stress analysis are used to address this concern.

References

36.1 J. B. Angell, S. C. Terry, P. W. Barth: Silicon micromechanical devices, Sci. Am. **248** (1983) 44–50

36.2 T. A. Core, W. K. Tsang, S. J. Sherman: Fabrication technology for an integrated surface-micromachined sensor, Solid State Technol. **36**(10) (1993) 39–47

36.3 W. Kuehnel, S. Sherman: A surface micromachined silicon accelerometer with on-chip detection circuitry, Sens. Actuators A **45** (1994) 7–16

36.4 K. H.-L. Chau, R. E. Sulouff Jr.: Technology for the high-volume manufacturing of integrated surface-micromachined accelerometer products, Microelectron. J. **29** (1998) 579–586

36.5 B. Sulouff: Integrated surface micromachined technology. In: *Sensors for Automotive Technology*, Sensors Applications, Vol. 4, ed. by J. Marek, H.-P. Trah, Y. Suzuki, I. Yokomori (Wiley-VCH, Weinheim 2003) Chap. 5.2

36.6 M. A. Mignardi, R. O. Gale, D. J. Dawson, J. C. Smith: The digital micromirror device – A micro-optical electromechanical device of display applications. In: *MEMS and MOEMS Technology and Applications*, ed. by P. Rai-Choudhury (SPIE, Bellingham 2000) Chap. 4

36.7 K. Nunan, G. Ready, J. Sledziewski: LPCVD & PECVD operations designed for iMEMS sensor devices, Vac. Technol. Coat. **2**(1) (2001) 26–37

36.8 M. Williams, J. Smith, J. Mark, G. Matamis, B. Gogoi: Development of low stress, silicon-rich nitride film for micromachined sensor applications, *Micromachining and Microfabrication Process Technology VI*, Proc. SPIE **4174**, ed. by J. Karam, J. Yasaitis, Proc. SPIE **4174** (2000) 436–442

36.9 Z. Zhang, K. Eskes: Elimination of wafer edge die yield loss for accelerometers, *Micromachining and Microfabrication Process Technology VI*, ed. by J. Karam, J. Yasaitis, Proc. SPIE **4174** (2000) 477–484

36.10 G. Bitko, A. C. McNeil, D. J. Monk: Effect of inorganic thin film material processing and properties on stress in silicon piezoresistive pressure sensors, Proc. Mat. Res. Soc. Symp. **444** (1997) 221–226

36.11 A. Hein, S. Finkbeiner, J. Marek, E. Obermeier: Material related effects on wet chemical micromachining of smart MEMS devices, *Micromachined Devices and Components V*, ed. by P. French, E. Peeters, Proc. SPIE **3876** (1999) 29–36

36.12 B. Sulouff: Commercialization of MEMS automotive accelerometers, 7th International Conference on the Commercialization of Micro and Nano Systems (COMS), Ypsilanti 2002 (MANCEF, Albuquerque 2002) 267–270

36.13 H. Weinberg: MEMS sensors are driving the automotive industry, Sensors **19**(2) (2002) 36–41

36.14 J. Marek, M. Illing: Microsystems for the automotive industry, Electron Devices Meeting, IEDM Technical Digest International, San Francisco 2000 (IEEE, New York 2000) 3–8

36.15 D. S. Eddy, D. R. Sparks: Application of MEMS technology in automotive sensors and actuators, Proc. IEEE **86**(8) (1998) 1747–1755

36.16 R. Verma, I. Baskett, B. Loggins: Micromachined electromechanical sensors for automotive applications, SAE Special Pub. **1312** (1998) 55–59

36.17 R. Maboudian, R. T. Howe: Critical review: Adhesion in surface micromechanical structures, J. Vac. Sci. Technol. B **15**(1) (1997) 1–20

36.18 U. Srinivasan, M. R. Houston, R. T. Howe, R. Maboudian: Alkyltrichlorosilane-based self-assembled monolayer films for stiction reduction in silicon micromachines, J. Microelectromech. Syst. **7**(2) (1998) 252–260

36.19 B. H. Kim, T. D. Chung, C. H. Oh, K. Chun: A new organic modifier for anti-stiction, J. Microelectromech. Syst. **10**(1) (2001) 33–40

36.20 S. Pamidighantam, W. Laureyn, A. Salah, A. Verbist, H. Tilmans: A novel process for fabricating slender and compliant suspended poly-Si micromechanical structures with sub-micron gap spacing, 15th IEEE 2002 Micro Electro Mechanical Systems (MEMS) Conf. (IEEE, New York 2002) 661–664

36.21 B. C. Bunker, R. W. Carpick, R. A. Assink, M. L. Thomas, M. G. Hankins, J. A. Voigt, D. Sipola, M. P. de Boer, G. L. Gulley: Impact of solution agglomeration on the deposition of self-assembled monolayers, Langmuir **16** (2000) 7742–7751

36.22 Y. Jun, V. Boiadjiev, R. Major, X.-Y. Zhu: Novel chemistry for surface engineering in MEMS, *Materials and Devices Characterization in Micromachining III*, ed. by Y. Vladimirsky, P. Coane, Proc. SPIE **4175** (2000) 113–120

36.23 R. Maboudian, W. R. Ashurst, C. Carraro: Self-assembled monolayers as anti-stiction coatings for MEMS: Characteristics and recent developments, Sens. Actuators A **82**(1) (2000) 219–223

36.24 W. R. Ashurst, C. Yau, C. Carraro, R. Maboudian, M. T. Dugger: Dichlorodimethylsilane as an anti-stiction monolayer for MEMS: A comparison to the octadecyltrichlorosilane self-assembled monolayer, J. Microelectromech. Syst. **10** (2001) 41–49

36.25 W. R. Ashurst, C. Yau, C. Carraro, C. Lee, G. J. Kluth, R. T. Howe, R. Maboudian: Alkene based monolayer films as anti-stiction coatings for polysilicon MEMS, Sens. Actuators A **91**(3) (2001) 239–248

36.26 S. J. Jacobs, S. A. Miller, J. J. Malone, W. C. McDonald, V. C. Lopes, L. K. Magel: Hermeticity and stiction in MEMS packaging, 40th Annual IEEE In-

ternational Reliability Physics Symp., Dallas 2002 (IEEE, New York 2002) 136–139
36.27 Nexus Task Force Report: *Market Analysis for Microsystems 1996-2002* (NEXUS, Grenoble 1998) www.nexus-mems.com
36.28 T. Maudie, J. Wertz: Pressure sensor performance and reliability, IEEE Industry Appl. Mag. **3**(3) (1997) 37–43
36.29 Lj. Ristic, R. Gutteridge, B. Dunn, D. Mietus, P. Bennett: Surface micromachined polysilicon accelerometer, IEEE 1992 Solid State Sensor and Actuator Workshop, IEEE 5th Technical Digest, Hilton Head 1992 (IEEE, New York 1992) 118–121
36.30 M. Furtsch, M. Offenberg, H. Munzel, J. R. Morante: Influence of anneals in oxygen ambient on stress of thick polysilicon layers, Sens. Actuators **76** (1999) 335–342
36.31 P. Lange, M. Kirsten, W. Riethmuller, B. Wenk, G. Zwicker, J. R. Morante, F. Ericson, J. A. Schweitz: Thick polycristalline silicon for surface micromechanical applications: Deposition, structuring and mechanical characterization, Proc. 8th International Conf. on Solid State Sensors and Actuators and Eurosensors IX, Transducers '95, Vol. 1 (IEEE, New York 1995) 202–205
36.32 F. Laermer, A. Schilp: Method of anisotropically etching silicon, (1996) U.S. Patent 5,501,893
36.33 M. Offenberg, H. Munzel, D. Schubert, O. Schatz, F. Laermer, E. Muller, B. Maihofer, J. Marek: Acceleration sensor in surface micromachining for airbag applications with high signal/noise ratio, SAE Special Pub. **1133** (1996) 35–41
36.34 F. Laermer, A. Schilp, K. Funk, M. Offenberg: Bosch deep silicon etching: Improving uniformity and etch rate for advanced MEMS applications, Proc 12th International Conf. on Micro Electro Mechanical Systems MEMS (IEEE, New York 1999) 211–216
36.35 J. R. Martin, C. M. Roberts Jr.: Package for sealing an integrated circuit die, (2001) U.S. Patent 6,323,550
36.36 J. A. Geen, S. J. Sherman, J. F. Chang, S. R. Lewis: Single chip surface micromachined integrated gyroscope with 50 deg/hour allan deviation, IEEE J. Solid-State Circuits **37**(12) (2002) 1860–1866
36.37 K. Funk, H. Emmerich, A. Schilp, M. Offenberg, R. Neul, F. Larmer: A surface micromachined silicon gyroscope using a thick polysilicon layer, Proc. 12th International Conf. on Micro Electro Mechanical Systems MEMS (IEEE, New York 1999) 57–60
36.38 M. Lutz, W. Golderer, J. Gerstenmeier, J. Marek, B. Maihofer, S. Mahler, H. Munzel, U. Bischof: A precision yaw rate sensor in silicon micromachining, 11th International Conf. on Solid State Sensors and Actuators, Transducers '97 (IEEE, New York 1997) 847–850
36.39 T. A. Core, R. T. Howe: Method for fabricating microstructures, (1994) U.S. Patent 5,314,572
36.40 M. Offenberg, F. Laermer, B. Elsner, H. Munzel, W. Reithmuller: Novel process for a monolithic integrated accelerometer, Proc. 8th International Conf. on Solid State Sensors and Actuators and Eurosensors IX, Transducers '95 (IEEE, New York 1995) 589–592
36.41 J. Anguita, F. Briones: HF/H_2O vapor etching of SiO_2 sacrificial layer for large-area surface-micromachined membranes, Sens. Actuators A **64** (1998) 247–251
36.42 M. A. Schmidt: Wafer-to-wafer bonding for microstructure formation, Proc. IEEE **86**(8) (1998) 1575–1585
36.43 A. R. Mirza: Wafer-level bonding technology for MEMS, Proc. 7th Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems, ITHERM 2000, Vol. 1 (IEEE, New York 2000) 113–119
36.44 A. R. Mirza: One micron precision, wafer-level aligned bonding for interconnect, MEMS and packaging applications, Proc. 50th Electronic Components & Technology Conf. (IEEE, New York 2000) 676–680
36.45 S. Mack, H. Baumann, U. Goesele: Gas tightness of cavities sealed by silicon wafer bonding, Proc. 10th Annual International Workshop on Micro Electro Mechanical Systems, MEMS; IEEE Micro Electro Mechanical Systems (MEMS) (IEEE, New York 1997) 488–493
36.46 G. Wallis: Field assisted glass sealing, SAE Automotive Engineering Congress, Detroit 1971 (Society of Automotive Engineers, New York 1971) Paper 71023
36.47 S. A. Audet, K. M. Edenfeld: Integrated sensor wafer-level packaging, Proc. International Conf. on Solid State Sensors and Actuators, Transducers '97, Vol. 1 (IEEE, New York 1997) 287–289
36.48 C. Gormley, A. Boyle, V. Srigengan, S. Blackstone: HARM processing techniques for MEMS and MOEMS devices using bonded SOI substrates and DRIE, *Micromachining and Microfabrication Process Technology VI*, ed. by J. Karam, J. Yasaitis, Proc. SPIE **4174** (2000) 98–110
36.49 K. Somasundram, D. Cole, C. McNamara, A. Boyle, P. McCann, C. Devine, A. Nevin: Fusion-bonded multilayered SOI for MEMS applications, Smart Sensors, Actuators and MEMS, Proc. SPIE **5116**, ed. by J.-C. Chiao, V. Varadan, C. Cané (2003) 12–19
36.50 C. M. Roberts Jr., L. H. Long, P. A. Ruggerio: Method for separating circuit dies from a wafer, (1994) U.S. Patent 5,362,681
36.51 T. Maudie, T. Miller, R. Nielsen, D. Wallace, T. Ruehs, D. Zehrbach: Challenges of MEMS device characterization in engineering development and final manufacturing, Proc. 1998 IEEE AUTOTESTCON (IEEE, New York 1998) 164–170

36.52 S. Petrovic, A. Ramirez, T. Maudie, D. Stanerson, J. Wertz, G. Bitko, J. Matkin, D. J. Monk: Reliability test methods for media-compatible pressure sensors, IEEE Trans. Industrial Electron. **45**(6) (1998) 877–885

36.53 S. Petrovic: Progress in media compatible pressure sensors, Proc. of InterPACK'01, the Pacific Rim/International Intersociety Electronic Packaging Technical/Business Conf. & Exhibition (ASME, New York 2001) IPACK2001-15517

36.54 H.-J. Kress, J. Marek, M. Mast, O. Schatz, J. Muchow: Integrated pressure sensors with electronic trimming, Automotive Eng. **103**(4) (1995) 65–68

36.55 C. Fung, R. Harris, T. Zhu: Multifunction polysilicon pressure sensors for process control, Sensors **16**(10) (1999) 75–79

36.56 K. H.-L. Chau, C. D. Fung, P. R. Harris, J. G. Panagou: High-stress and overrange behavior of sealed cavity polysilicon pressure sensors, IEEE 4th Technical Digest, Solid State Sensor and Actuator Workshop (IEEE, New York 1990) 181–183

36.57 A. B. Sontheimer: Digital micromirror device (DMD) hinge memory lifetime reliability modeling, Proc. 40th Annual IEEE International Reliability Physics Symp. (IEEE, New York 2002) 118–121

36.58 J. P. O'Connor: Packaging design considerations and guidelines for the digital micromirror device$^{\text{TM}}$, Proc. of InterPACK'01, the Pacific Rim/International Intersociety Electronic Packaging Technical/Business Conf. & Exhibition (ASME, New York 2001) IPACK2001-15526

36.59 C. Bang, V. Bright, M. A. Mignardi, D. J. Monk: Assembly and test for MEMS and optical MEMS. In: *MEMS and MOEMS Technology and Applications*, ed. by P. Rai-Choudhury (SPIE, Bellingham 2000) Chap. 7

36.60 F. Poradish, J. T. McKinley: Package for a semiconductor device, (1994) U.S. Patent 5,293,511

36.61 T. W. Migl: Interfacing to the digital micromirror device for home entertainment applications, Proc. of InterPACK'01, the Pacific Rim/International Intersociety Electronic Packaging Technical/Business Conf. & Exhibition (ASME, New York 2001) IPACK2001-15712

36.62 S. J. Jacobs, J. J. Malone, S. A. Miller, A. Gonzalez, R. Robbins, V. C. Lopes, D. Doane: Challenges in DMD$^{\text{TM}}$ assembly and test, Proc. Mater. Res. Soc. **657** (2001) EE6.1.1–EE6.1.12

36.63 Automotive Industry Action Group: *Quality Systems Requirements QS-9000*, 3rd edn. (Automotive Industry Action Group of the American Society for Quality, Milwaukee 1998)

36.64 J. A. Walraven, B. A. Waterson, I. De Wolf: Failure analysis of micromechanical systems (MEMS). In: *Microelectronic Failure Analysis*, 4th edn. (ASM, Materials Park 2002) 2002 Suppl.

36.65 M. R. Douglass: Lifetime estimates and unique failure mechanisms of the digital micromirror device (DMD), Proc. 1998 36th Annual IEEE International Reliability Physics Symp. (IEEE, New York 1998) 9–16

36.66 J. R. Martin: Surface characteristics of integrated MEMS in high volume production. In: *Nanotribology: Critical Assessment and Research Needs*, ed. by S. M. Hsu, Z. C. Ying (Kluwer, Dordrecht 2002) Chap. 14

36.67 H. R. Shea, A. Gasparyan, C. D. White, R. B. Comizzoli, D. Abusch-Magder, S. Arney: Anodic oxidation and reliability of MEMS poly-silicon electrodes at high relative humidity and high voltages, *MEMS Reliability for Critical Applications*; Proc. SPIE **4180** (2000) 117–122

36.68 J. Martin: Stiction suppression in high volume MEMS products, Proc. 2003 STLE/ASME Joint International Tribology Conf. (ASME, New York 2003) 2003TRIB-266

36.69 G. X. Li, F. A. Shemansky Jr.: Drop Test and analysis on micro-machined structures, Sens. Actuators **85** (2000) 280–286

36.70 R. T. Howe, H. J. Barber, M. Judy: Apparatus to minimize stiction in micromachined structures, (1996) U.S. patent 5,542,295

36.71 L. J. Hornbeck, W. E. Nelson: Spatial light modulator and method, (1992) U.S. Patent 5,096,279

36.72 L. J. Hornbeck: Low reset voltage process for DMD, (1994) U.S. Patent 5,331,454

36.73 E. C. Fisher, R. Jascott, R. O. Gale: Method of passivating a micromechanical device within a hermetic package, (1999) U.S. Patent 5,936,758

36.74 B. A. Miksic: Use of vapor phase inhibitors for corrosion protection of metal products, Proc. 1983 NACE Annual Conf., Corrosion 83 (National Assoc. Corrosion Engineers, Houston 1983) Paper 308

36.75 D. Vanderpool, S. Akin, P. Hassett: Corrosion inhibitors in the electronics industry: Organic copper corrosion inhibitors, Proc. 1986 NACE Annual Conf., Corrosion 86 (National Assoc. Corrosion Engineers, Houston 1986) Paper 1

36.76 J. R. Martin, Y. Zhao: Micromachined device packaged to reduce stiction, (1997) U.S. Patent 5,694,740

36.77 J. R. Martin: Process for wafer level treatment to reduce stiction and passivate micromachined surfaces and compounds used therefor, (2001) International Patent Application WO 01/57920 A1

36.78 S. T. Patton, J. S. Zabinski: Failure mechanisms of a MEMS actuator in very high vacuum, Tribol. Int. **35**(6) (2002) 373–379

36.79 S. T. Patton, K. C. Eapen, J. S. Zabinski: Effects of adsorbed water and sample aging in air on the µN level adhesion force between Si(100) and silicon nitride, Tribol. Int. **34**(7) (2001) 481–491

36.80 S. T. Patton, W. D. Cowan, K. C. Eapen, J. S. Zabinski: Effect of surface chemistry on the tribological performance of a MEMS electrostatic lateral output motor, Tribol. Lett. **9** (2000) 199–209

36.81 S. T. Patton, W. D. Cowan, J. S. Zabinski: Performance and reliability of a new MEMS electrostatic lateral output motor, Proc. 1999 37th Annual IEEE International Reliability Physics Symp. (IEEE, New York 1999) 179–188
36.82 D. M. Tanner, J. A. Walraven, L. W. Irwin, M. T. Dugger, N. F. Smith, W. P. Eaton, W. M. Miller, S. L. Miller: The effect of humidity on the reliability of a surface micromachined microengine, Proc. 1999 37th Annual IEEE International Reliability Physics Symp. (IEEE, New York 1999) 189–197
36.83 M. P. de Boer, P. J. Clews, B. K. Smith, T. A. Michalske: Adhesion of polysilicon microbeams in controlled humidity ambients, Proc. Mater. Res. Soc. Symp. **518** (1998) 131–136

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

37. MEMS Packaging and Thermal Issues in Reliability

The potential of MEMS/NEMS technologies has been viewed as a comparable or even bigger revolution than that of microelectronics. These scientific and engineering advancements in MEMS/NEMS could bring applications to reality previously unthinkable, from space systems, environmental instruments, to daily life appliances. As presented in previous chapters, the development of core MEMS/NEMS processes has already demonstrated a lot of commercial applications as well as future potentials with elaborated functionalities. However, a low cost and reliable package for the protection of these MEMS/NEMS products is still a very difficult task. Without addressing the packaging and reliability issues, no commercial products can be sold on the market. Packaging design and modeling, packaging material selection, packaging process integration, and packaging cost are main issues to be considered when developing a new MEMS packaging process. In this chapter, we will present the fundamentals of MEMS/NEMS packaging technology, including packaging processes, hermetic and vacuum encapsulations, thermal issues, packaging reliability, and future packaging trends. The future development of MEMS packaging will rely on the success of the implementation of several unique echniques, such as packaging design kits for system and circuit designer, low cost and high yield wafer level, chip-scale packaging techniques, effective testing techniques at the wafer-level to reduce overall testing costs; and reliable fabrication of an interposer [37.1] with vertical through-interconnects for device integrations.

37.1 **MEMS Packaging** 1111
  37.1.1 MEMS Packaging Fundamentals...... 1112
  37.1.2 Contemporary MEMS Packaging Approaches 1113

37.2 **Hermetic and Vacuum Packaging and Applications** 1116
  37.2.1 Integrated Micromachining Processes 1117
  37.2.2 Post-Packaging Processes............. 1118
  37.2.3 Localized Heating and Bonding...... 1119

37.3 **Thermal Issues and Packaging Reliability** 1122
  37.3.1 Thermal Issues in Packaging 1122
  37.3.2 Packaging Reliability 1124
  37.3.3 Long-Term and Accelerated MEMS Packaging Tests........................... 1125

37.4 **Future Trends and Summary** 1128

References 1129

37.1 MEMS Packaging

MEMS are miniaturized systems in a size of micrometer to millimeter that may have mechanical, chemical, or biomedical features integrated with IC circuitry for sensor or actuator applications [37.2]. For example, pressure [37.3], temperature, flow [37.4], accelerometers [37.5], gyroscopes [37.6], and chemical sensors [37.7] can be fabricated by MEMS technologies for sensing applications. Fluidic valves [37.8], pumps [37.9], and inkjet printer heads are examples of actuation devices for medical, environmental, office, and industrial applications. Silicon is typically used as the primary substrate material for MEMS fabrication because it can provide unique electrical, thermal, and mechanical properties however also can be easily micromachined in a form of batch-processing and be incorporated with a microelectronic circuits by using most of the conventional semiconductor manufacturing processes and tools. As a result, smaller size, lighter weight, lesser power consumption and cheaper fabrication cost become the advantages of MEMS devices as compared with the existing macroscale systems with similar functionalities. With the advances of MEMS fabrication technology in the past decades, the MEMS market at the component-level is currently in excess of

$ 5 billion and is driving the end-product markets of large than $ 100 billion [37.10].

Nevertheless, the road to the commercialization of MEMS doesn't look as promising as expected. Many industrial companies took the advantages of MEMS technology of high production volumes and high added value created by product integrations. Therefore, cost-efficiency becomes the major factor driving MEMS toward commercialization. Several MEMS devices have been developed for and applied in automotive industry and information technology, and they dominate the MEMS market due to high production volumes. Most of custom-designed MEMS products are still very diverse, aiming for different applications, and their initial costs in small-to-medium-scale production are still much higher than market-acceptable levels. Consequently, high packaging and testing costs have slowed the commercialization of MEMS. Furthermore, based on the past experience of IC industry, the cost of packaging processes is about 30%, and sometimes can be more than 70%, of the total production expenses. MEMS packaging process is expected to be even more costly because of the challenging and stringent packaging issues with regard to the additional MEMS components, in addition to the microelectronic circuitry, in a typical MEMS product [37.11].

37.1.1 MEMS Packaging Fundamentals

The functions of conventional IC packaging are to protect, power, and cool the microelectronic chips or components and provide electrical and mechanical connection between the microelectronic part and the outside world. With the increase of the needs of high performance and multifunctional consumer electronic products, IC packaging processes have incorporated more complex designs and advanced fabrication technologies, such as Cu interconnects [37.12], flip-chip bonding [37.13], ball grid array [37.14], wafer-level chip scale packaging [37.15], and 3-D packaging [37.16], etc., in order to satisfy the need for high I/O density, large die area, and high clock frequency. In addition to the requirements of electrical interconnects, MEMS components may need to interface outside environment (for example, fluidic interconnectors [37.17]), some other components may need to be hermetically sealed in vacuum (for example, accelerometers [37.5]). Therefore, MEMS packaging processes have to provide more functionalities including better mechanical protection, thermal management, hermetic sealing, and complex electricity and signal distribution.

It has been suggested that MEMS packaging should be incorporated in the device fabrication stage as part of the micromachining process. Although this approach solves the need for some specific devices, it does not solve the packaging need for general microsystems. Many MEMS devices are now fabricated by various foundry services [37.18, 19], and there is a tremendous need for a uniform packaging process. Figure 37.1 shows a typical MEMS device being encapsulated by a packaging cap. The most fragile part of this device is the suspended mechanical sensor, which is a freestanding mechanical, mass-spring microstructure. This mechanical part must be protected during the packaging and handling process. Moreover, vacuum encapsulation may be required for these microstructures in applications such as resonant accelerometers [37.5] or gyroscopes [37.6, 20]. A "packaging cap" with properly designed micro cavity is to be fabricated to encapsulate and protect the fragile MEMS structure as the first-level MEMS post-packaging process. The wafer can be diced afterward, and the well-established packaging technology in IC industry can follow and finish the final packaging step. Unlike the packaging requirements for ICs, however, the common MEMS packaging requirement is hermetic seal and sometimes vacuum encapsulations. Hermetic seal is important to ensure that no moisture or contaminant can enter the package

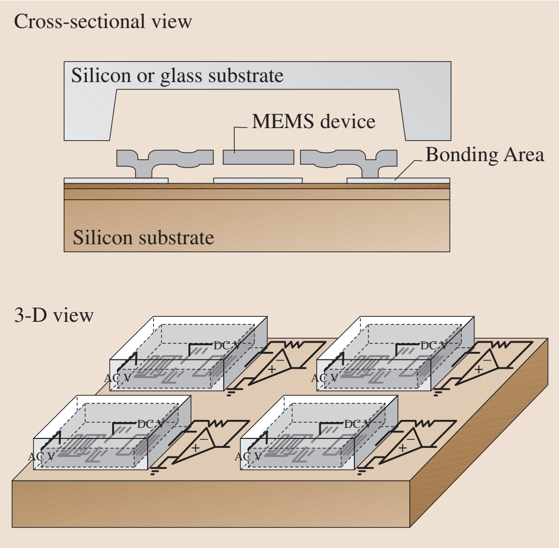

Fig. 37.1 A typical MEMS packaging illustration in which a MEMS structure is encapsulated and protected by the packaging cap

and affect the functionality of microstructures. This increases the difficulty of common IC packaging processes tremendously. Although most single function MEMS chips can employ typical IC packaging techniques, such as die-attached processes, wiring interconnects, molded plastic, ceramic, and metal for packaging [37.21], with the increase of complexity MEMS devices, more advanced packaging techniques, especially wafer-level packaging, are required for the integration of multi-chips for multifunctional applications.

Before the state-of-the-art MEMS packaging processes are discussed, several primary micro-fabrication processes for packaging applications are described. These processes may include Flip-Chip (FC) technique, Ball Grid Arrays (BGA), through-wafer etching, and plating. Other silicon-based processes, such as thin-film deposition, wet and dry chemical etching, lithography, lift-off, and wiring bonding processes can be found in many textbooks [37.22].

Flip-Chip Technique (FC)

This technique is commonly used in the assembly process between a chip with microelectronics and a package substrate [37.13]. The microelectronic chip is “flipped joined” with the packaging substrate, and metal solder bumps are used as both the bonding agents and electrical paths between bond pads on the microelectronic chip and metal pads on the package substrate. Because the vertical bonding space can be very small as controlled by the heights of the solder bumps, and the lateral distributions of bond pads can be on the whole chip instead of being only on the edge, this technique can provide high density Input/Output (I/O) connections. In the FC technique, solder bumps are generally fabricated by means of electroplating. Before the bumping process, multiple metal layers, such as TiW-Cu, Cr-Cu, Cr-Ni, TaN/Ta/Ni, have to be deposited as a seed layer for electroplating and as a barrier to prevent the diffusion of solder into its underneath electrical interconnect.

Ball Grid Arrays (BGA)

This technology is very similar to the FC technique. An area array of solder balls on a single chip module or multi-chip module are used in the packaging process as electrical, thermal, and mechanical connects to join the module with the next level package, usually a printed circuit board [37.14]. The major difference between a typical BGA and FC is the size of solder bump. In BGA chip, the bumps are in the order of 750 μm in diameter which is ten times larger than those commonly used in FC.

Through-Wafer Etching

This is a chemical etching process to make through-wafer channels on a silicon substrate for the fabrication of vertical through-wafer interconnects. The chemical etching process can be either wet or dry. Anisotropic or isotropic etching solutions can be used in the wet etching process. The dry etch process is based on plasma and ion-assisted chemical reactions that can be either isotropic or anisotropic. In order to create high density and high aspect ratio through-wafer vias, Deep Reactive Ion Etching (DRIE) is typically used. Two popular DRIE approaches, Bosch and Cyro, are well described in literature [37.23].

Electroplating

Electroplating is another common micro-fabrication process. It can be conducted for the deposition of an adherent metallic layer onto a conductive or nonconductive substrate. The process on a conductive substrate is called electrolytic plating, which utilizes a seed layer as the anode to transfer metal ions onto the cathode surface when a DC current is passed through the plating solution. The plating process without applying electrical current is called electroless plating, which can happen in both conducive and nonconductive surfaces. Electroless plating process requires a layer of noble metal such as Pd, Pt, or Ru on the substrate as the catalysis to trigger the self-decomposition reaction in the plating solution. The electroplating processes are very important for electrical interconnect and solder bump fabrications for packaging applications because of low process temperature and cost.

These processes are generally developed for providing electrical and thermal paths for various IC/MEMS packaging approaches.

37.1.2 Contemporary MEMS Packaging Approaches

Several MEMS packaging issues and approaches before 1985 were discussed in [37.24], and researchers have been working on MEMS packaging approaches continuously. For example, *Senturia* and *Smith* [37.25] discussed the packaging and partitioning issues for microsystems. *Smith* and *Collins* [37.26] used epoxy to bond glass and silicon for chemical sensors. Several MCM methods have been proposed. *Butler* et al. [37.27] proposed adapting multi-chip module foundries using chip-on-flex (COF) process. *Schuenemann* et al. [37.28] introduced a 3-D stackable packaging concept for the Top-Bottom Ball Grid Array (TB-BGA) that includes electric, fluidic, optic, and communication interfaces. *Lee* et al. [37.29] and *Ok* et al. [37.10] presented a direct-chip-attach MEMS packaging using through-wafer electrical interconnects. *Laskar* and *Blythe* [37.30] developed a multi-chip modules (MCM) type packaging process by using epoxy. *Reichl* [37.31] discussed different materials for bonding and interconnection. *Grisel* et al. [37.32] designed a special process to package micro-chemical sensors. Special processes have also been developed for MEMS packaging, such as packaging for microelectrodes [37.33], packaging for biomedical systems [37.34], and packaging for space systems [37.35]. These specially designed, device-oriented packaging methods are meant for individual systems. There is no reliable method yet that would qualify as a versatile post-packaging process for MEMS with the rigorous process requirements of low temperature, hermetic sealing and long-term stability.

An integrated process by using surface-micromachined microshells has previously been developed [37.36]. This process applies the concepts of sacrificial layer and LPCVD sealing to achieve wafer-level post-packaging. Similar processes have been demonstrated. For example, *Guckel* et al. [37.37] and *Sniegowski* et al. [37.38] developed a reactive sealing method to seal vibratory micromachined beams. *Ikeda* et al. [37.39] adopted epitaxial silicon to seal microstructures. *Mastrangelo* et al. [37.40] used silicon nitride to seal mechanical beams as light sources. *Smith* et al. [37.41] accomplished a new fabrication technology by embedding microstructures and CMOS circuitry. All of these methods have integrated the MEMS process together with the post-packaging such that no extra bonding is required. These schemes are highly process dependent, however, and not suitable for prefabricated circuitry.

Recently several new efforts for MEMS post-packaging processes have been reported. *Butler* et al. [37.27] demonstrated an advanced MCM packaging scheme. It adopts the high density interconnect (HDI) process consisting of embedding bare die into premilled substrates. Because the MEMS structures have to be released after the packaging process, it is undesirable for general microsystems. *Van der Groen* et al. [37.42] reported a transfer technique for CMOS circuits based on epoxy bonding. This process overcomes the surface roughness problem, but epoxy is not a good material for hermetic sealing. In 1996, *Cohn* et al. [37.43] demonstrated a wafer-to-wafer to-wafer vacuum packaging process by using Silicon–Gold eutectic bonding with a 2-μm-thick polysilicon microcap. But experimental results show substantial leakage after a period of 50 days. *Cheng* et al. [37.44] developed a vacuum packaging technology using localized aluminum/silicon-to-glass bonding. In 2002, *Chiao* and *Lin* [37.45] demonstrated vacuum packaging of microresonators by rapid thermal processing. These recent and ongoing research efforts indicate the strong need for a versatile MEMS post-packaging process.

Bonding Processes for MEMS Packaging Applications

Previously, silicon-bonding technologies have been used in many MEMS fabrication and packaging applications, including epoxy bonding, eutectic bonding, anodic bonding, fusion bonding, and solder bonding. For example, such devices as pressure sensors, micro-pumps, biomedical sensors or chemical sensors require mechanical interconnectors to be bonded on the substrate (see for example [37.7, 17, 46]). Glass has been commonly used as the bonding material by anodic bonding at a temperature of about $300 \sim 450\,^{\circ}\mathrm{C}$ (see for example [37.47, 48]). *Klaassen* et al. [37.49] and *Hsu* et al. [37.50] have demonstrated different types of silicon fusion bonding and Si–$\mathrm{SiO_2}$ bonding processes at very high temperatures of over $1{,}000\,^{\circ}\mathrm{C}$. *Ko* et al. [37.24], *Tiensuu* et al. [37.51], *Lee* et al. [37.52] and *Cohn* et al. [37.43] have used eutectic bonding for different applications. All of these bonding techniques have different mechanisms that determine the individual bonding characteristics and process parameters. This section discusses the details of these processes.

Fusion Bonding for MEMS Packaging

Silicon fusion bonding is an important fabrication technique of the SOI (Silicon On Insulator). The bonding is based on Si−O, Si−N, or Si−Si strong covalent bond. But very high bonding temperature (higher than $1{,}000\,^{\circ}\mathrm{C}$) and flat bonding surfaces (less than 6 nm) are the two basic requirements for strong, uniform, and hermetic bonding. Although hydrophilic surface treatment can lower the bonding temperature, an annealing step higher than $800\,^{\circ}\mathrm{C}$ is still needed to remove possible bubble formation at the bonding interface. *Bower* et al. [37.53] proposed that low temperature $\mathrm{Si_3N_4}$ fusion bonding could be achieved at the temperature under $300\,^{\circ}\mathrm{C}$. *Takagi* et al. [37.54] proposed that silicon fusion bonding could be made at room temperature by using Ar^+ beam treatment on the wafer surface, and the bond strength is comparable to the conventional fusion bonding. In summary, fusion bonding has been a popular fabrication technique in MEMS fabrication and packaging.

Anodic Bonding for MEMS Packaging Applications

The invention of anodic bonding dates back to 1969 when *Wallis* and *Pomerantz* [37.55] found that glass and metal can be bonded together at about $200 \sim 400\,^{\circ}\mathrm{C}$ below the melting point of glass with the aid of a high electrical field. This technology has been widely used for protecting on-board electronics in biosensors (see for example [37.56–58]) and sealing cavities in pressure sensors (see for example [37.59]). Many reports have also discussed the possibility of lowering the bonding temperature by different mechanisms [37.60, 61]. Anodic bonding forms Si−O or Si−Si covalent bond and is one of the strongest chemical bonds available for silicon-based systems. The bonding process can be accomplished on a hot plate with a temperature between 180 °C and 500 °C in atmosphere or vacuum environment. When a static electrical field is built within the Pyrex glass (7740 from Dow Corning) and silicon, the electrostatic force can pull two surfaces close for a strong bond. In order to create high electrical field, a flat bonding surface with less than 50 nm roughness is required. In addition, the electrical field required for bonding is larger than $3\times10^{6}\,\mathrm{V/cm}$ [37.24]. Such a high electrical field is generated by a power supply of 200 to 1,000 V, which may cause damages on integrated circuits. *Hanneborg* et al. [37.62] have successfully bonded silicon with other thin solid films, such as silicon dioxide, nitride, and polysilicon, together with an intermediate glass layer using anodic bonding technique. In practice, electrostatic bonding has become widely accepted in MEMS fabrications and packaging applications. Unfortunately, the possible contamination due to excessive alkali metal in the glass; possible damage to microelectronics due to the high electrical field; and the requirement of flat surface for bonding limit the application of anodic bonding to MEMS post-packaging applications [37.63].

Epoxy Bonding (Adhesive Bonding)

Four major components comprise epoxy : epoxy resin, filler – like silver slake, solvent or reactive epoxy diluent, and additives like hardener and catalyst [37.64, 65]. The bonding mechanism of epoxy is very complicated and depends on the type of epoxy. In general, the main source of bonding strength is the van der Waals force. Because epoxy is a soft polymer material and its curing temperature for bonding is only around 150 °C, low residual stress and process temperature are the major advantages of epoxy bonding. But the properties of epoxy can be easily changed with environmental humidity and temperature so the bonding strength decays over time. In addition, epoxy bonding has low moisture resistance and, due to its additives, is a dirty process. These disadvantages have made epoxy unfavorable for the special MEMS packaging need for hermetic or vacuum sealing.

Eutectic Bonding

In many binary systems, a eutectic point corresponds to the alloy composition with the lowest melting temperature. If the environmental temperature is kept higher than the eutectic point, two contacted surfaces containing two elements with the eutectic composition can form liquid phase alloy. The solidification of the eutectic alloy forms “eutectic bonding” at a temperature lower than the melting temperature of either element in the alloy. Eutectic bonding can be a strong bond. For example, in the case of Au/Si alloy system, eutectic temperature is only 363 °C when the composition is at the atomic ratio of 81.4% Au to 18.6% Si and bonding strength is higher than 5.5 GPa [37.66]. Because other alloy systems may have lower eutectic temperatures than Al/Si system, they present great potential for MEMS packaging applications. In addition to Au/Si, Al/Ge/Si, Au/SnSi, and Au/Ge/Si systems have been applied for MEMS packaging.

Solder Bonding

Solder bonding has been widely applied in microelectronic packaging [37.67]. Both low bonding temperature and high bonding strength are good characteristics for packaging. Furthermore, there are a variety of choices of solder material for specific applications. *Singh* et al. [37.68] have successfully applied solder bump bonding in the integration of electronic components and mechanical devices for MEMS fabrication [37.69]. In this case, indium metal was used for bonding two separated silicon surfaces together by applying 350 MPa pressure and the bonding strength was as strong as 10 MPa. Glass frits can also be treated as a solder material and have been extensively used for vacuum encapsulation in MEMS industry. Glass frits are ceramic materials that can provide strong bonding strength with silicon with good hermeticity. Its bonding temperature is lower than 400 °C and is suitable for electronic components. But a bonding area more than 200 µm wide is required to achieve good results, which may become a drawback because area is the measure of manufacturing cost in IC industry. Nevertheless, glass frit is the

Table 37.1 Summary of bonding mechanisms

Bonding methods	Temperature	Roughness	Hermeticity	Post-packaging	Reliability
Fusion bonding	very high	highly sensitive	yes	yes by LH	good
Anodic bonding	medium	highly sensitive	yes	difficult	good
Epoxy bonding	low	low	no	yes	???
Integrated process	high	medium	yes	no	good
Low Temp. bonding	low	highly sensitive	???	no	???
Eutectic bonding	medium	low	yes	yes by LH	???
Brazing	very high	low	yes	yes by LH	good
???: no conlusive data		LH: Localized Heating			

most popular bonding process used in the current MEMS products.

Localized Heating and Bonding

Low bonding temperature and short process time are desirable process parameters in MEMS packaging fabrication to provide less thermal budget and high throughput. Most chemical bonding reactions, however, require a minimum and sufficient thermal energy to overcome the reaction energy barrier, or activation energy, to start the reaction and to form a strong bond. As a result, high bonding temperature generally results in shorter processing time to reach the same bonding quality at a lower bonding temperature [37.70]. The common limitations for the above techniques are their individual bonding characteristics and temperature requirements. In general, MEMS packaging requires a good bonding for hermetic sealing while the processing temperature must be kept low at the wafer-level to have less thermal effects on the existing devices. For example, a MEMS device may have prefabricated circuitry, biomaterial, or other temperature-sensitive materials such as organic polymer, magnetic metal alloy, or pizeo-ceramic. Since the packaging step comes after the MEMS device fabrication processes, bonding temperature should be kept low to avoid high temperature effects on the system. Possible temperature effects include residual stress due to the mismatch of thermal expansion coefficient of bonding materials and substrates, electrical contact failure due to atom inter-diffusion at the interface, and contamination due to the outgas or evaporation of materials. In addition to the control of bonding temperature, Other factors to consider are the magnitude of applied force to create intimate contact for bonding and atmospherics environment control. Based on the heat transfer simulation study [37.71], it is possible to confine high temperature area in a small region by localized heating without heating the whole substrate. Assembly steps can therefore always be processed after device fabrication without having detrimental effects. As such, localized heating and bonding technique is introduced and implemented for the fabrication of MEMS packaging for post-processing approaches [37.70, 72].

Table 37.1 summarizes these MEMS packaging technologies and their limitations, including the localized heating and bonding approach. The localized heating approach introduces several new opportunities. First, better and faster temperature control can be achieved. Second, higher temperature can be applied to improve the bonding quality. Third, new bonding mechanisms that require high temperature such as brazing [37.73] may now be explored in MEMS applications. As such, it has potential applications for a widerange of MEMS devices and is expected to advance the field of MEMS packaging.

37.2 Hermetic and Vacuum Packaging and Applications

Hermetic packaging is desirable for MEMS packaging requirements in most cases because it provides a moisture free environment to avoid charge separation in capacitive devices, corrosion in metallization, or electrolytic conduction in order to prolong the lifetime of the electronic circuitry. In several device applications, vacuum encapsulation is necessary but can be costly. Many surface micromachined resonant devices need vacuum to improve the performance including comb-shape µ-resonators and ring-type µ-gyroscopes that have very large surface-to-volume ratios and vibrate in a very tight space [37.20,39]. Two major approaches of MEMS hermetic and vacuum packaging have been demonstrated: (1) the integrated encapsulation approach and (2) the post-process packaging approach. Both are discussed in this section. Vacuum encapsulation by means of localized heating and bonding is discussed separately as another example for issues related to hermetic and vacuum packaging.

37.2.1 Integrated Micromachining Processes

Several MEMS hermetic and vacuum packaging processes have been demonstrated before based on the integrated micromachining processes. An integrated vacuum sealing process by LPCVD is presented here as the illustration example. This integrated process can encapsulate comb-shape microresonators [37.74] in vacuum at the wafer level. Figure 37.2 illustrates the cross-sectional view of the manufacturing process. First, standard surface-micromachining process [37.75] is conducted by using four masks to define the first polysilicon layer, anchors to the substrate, dimples and, the second polysilicon layer as shown in Fig. 37.2a. The process so far is similar to the MCNC MUMPs process [37.19], and comb-shape microstructures are fabricated at the end of these steps. In the standard surface micromachining process, the sacrificial layer (oxide) is etched away to release the microstructures. In the MEMS post-packaging process, a thick PSG (phosphorus-doped glass) of 7 µm is deposited to cover the microstructure and patterned by using 5:1 BHF (buffered HF) to define the microshell area as shown in Fig. 37.2b. A thin PSG layer of 1 µm is then deposited and defined to form etch channels as illustrated in Fig. 37.2c. The microshell material, low-stress silicon nitride, is now deposited with a thickness of 1 µm. Etch holes are defined and opened on the silicon nitride layer by using a plasma etcher. Silicon dioxide inside the packaging shell is now etched away by concentrated HF, and the wafer is dried using the supercritical CO_2 drying process [37.76]. After these steps, Fig. 37.2d applies. A 2-µm-thick LPCVD low-stress nitride is now deposited at a deposition pressure of 300 mtorr to seal the shell in the vacuum condition. Finally, the contact pads are opened as shown in Fig. 37.2e. Figure 37.3 is the SEM (Scanning Electron Microscope) microphoto of a finished device with protected microshell on top. The total packaging area (microshell) is about 400×400 µm^2. A contact pad is shown with the covering nitride layer removed. The shape of the microresonator, with beams of 150-µm-long and 2-µm-wide is reflected on the surface of the microshell due to the integrated packaging process. The total height of the nitride shell is 12 µm, as seen standing above the substrate. Spectrum measurement of the comb resonator inside the packaging reveals that a vacuum level of about 200 mtorr has been accomplished [37.77].

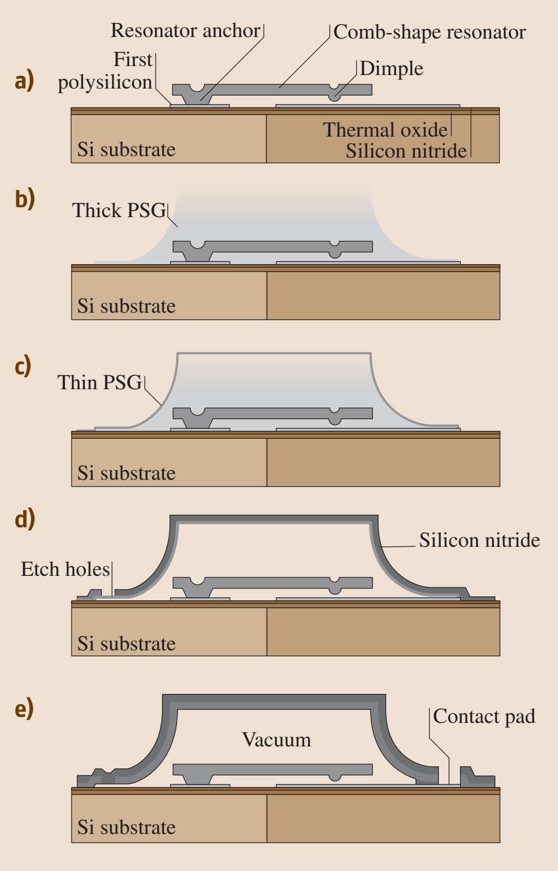

Fig. 37.2 An integrated vacuum encapsulation process using LPCVD nitride sealing to package micro-mechanical resonators [37.71]

Although the above vacuum sealing process successfully achieves MEMS hermetic and vacuum packaging, it has several drawbacks. First, several high temperature steps were used after the standard surface-micromachining process. As such, no circuitry (such as those shown in Fig. 37.1) or temperature-sensitive materials

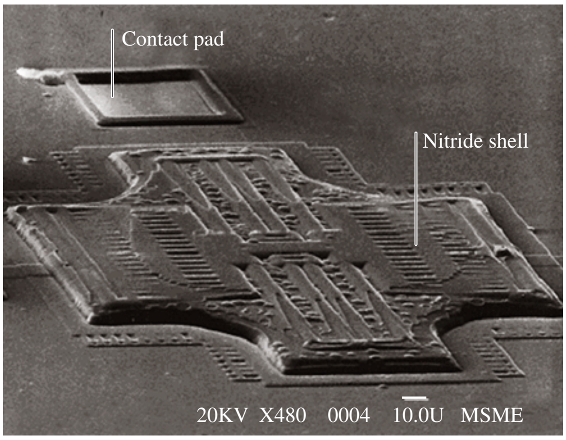

Fig. 37.3 SEM microphoto showing a vacuum packaged MEMS mechanical comb-shape resonator by an the integrated LPCVD sealing process as depicted in Fig. 37.2 [37.71]

will survive due to the global heating effect. Second, this post-packaging process is very specific and process dependent. MEMS companies or researchers have to adopt this post-packaging process with their own device manufacturing process. It can't be conducted in the multiuser MEMS process such as the MCNC MUMPs. Third, the thickness of the microshell is limited by the thin film deposition step that generally is in the range of a few micrometers. Wether the thin microshell can survive the high-pressure plastic molding process afterward in some device packaging design is a big concern. Finally, although integrated encapsulation can achieve low pressure by wafer-level fabrication and provide lower manufacturing cost, it does not provide the controllability of the cavity pressure.

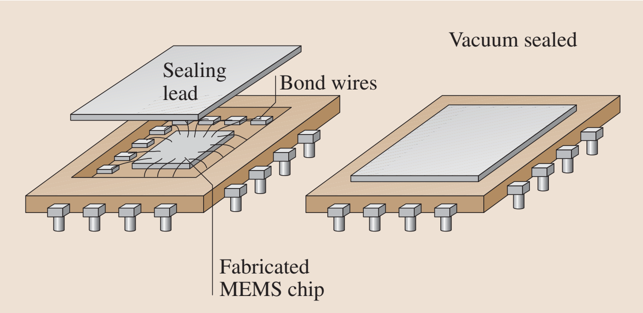

Fig. 37.4 The schematic diagram of industrial post-packaging (DIPS) using a ceramic holder to be covered by a sealing lid

37.2.2 Post-Packaging Processes

The second approach is defined as post-packaging process. The packaging process starts when the device fabrication processes are completed, so this approach has high flexibility for various microsystems. For example, Fig. 37.4 shows a common industrial post-hermetic packaging called Dual-in-line packaging (DIP) [37.78, 79]. A die is placed inside a ceramic holder covered by a sealing lid. Solder or ceramic joining is generally used for assembling lid and holder under a pressure-controlled environment. High cost is the major drawback of this method because of expensive ceramic holder and low fabrication throughput. Another example of post-packaging method is based on wafer bonding techniques combined with a microshell encapsulation. Devices are sealed by stacking another micromachined silicon or glass substrate as illustrated in Fig. 37.1. Integrated microsystems and protection shells are fabricated on different wafers, either silicon or glass, at the same time. After the two substrates are assembled together using silicon fusion, anodic, or low temperature solder bonding to achieve the final encapsulation, these micro-shells will provide mechanical support, thermal path, or electrical contact for the MEMS devices. Low packaging cost can be expected due to wafer-level processing.

The related packaging issues are discussed by using a specific example of hermetic and vacuum post-packaging by rapid thermal processing (RTP) that provides the advantages of wafer-level processing and low thermal budget. Figure 37.5 shows the

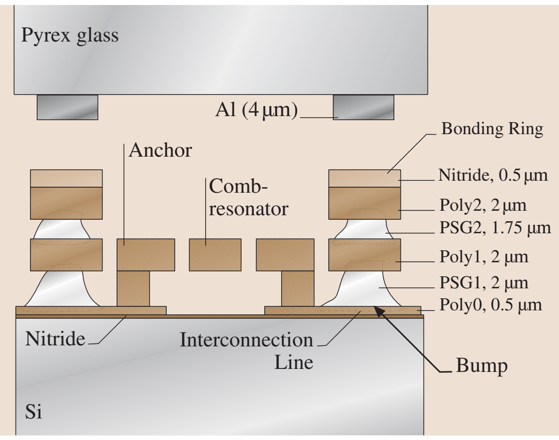

Fig. 37.5 The schematic diagram of the RTP (Rapid Thermal Processing) aluminum-to-nitride bonding setup [37.45]

schematic illustration of the RTP MEMS vacuum packaging process. Comb-shape microresonators are chosen as the vacuum packaging examples in this chapter, and a standard surface-micromachining process is used to fabricate these microresonators [37.19]. One major addition to a regular microstructure is that an integrated sealing ring using silicon nitride as the topmost layer is incorporated in the manufacturing process for the purpose of aluminum-to-nitride bonding [37.45]. A glass cap wafer is deposited and patterned with 4-µm-thick aluminum sealing-rings with width ranging from 100 to 250 µm and bonding area ranging from 450 × 450 to 1,000 × 1,000 μm^2. Before the vacuum packaging process, both the device and cap wafers are baked in vacuum at 300 °C for over four hours to dry water and gas species adhere at the surface. Afterward, the device and cap wafers are flip-chip assembled immediately and loaded into a quartz chamber and put into a RTP chamber. The base vacuum estimated at 10 mtorr inside the quartz chamber is achieved by using a mechanical pump. After heating for 10 seconds at 750 °C, the aluminum-to-nitride bond is formed. Figure 37.6 shows the measured spectrum of a vacuum-packaged, double-folded beam comb-drive resonator by using a micro-stroboscope [37.80]. The central resonant frequency is about 18,625 Hz and the quality factor is extracted as 1,800 ± 200 corresponding to a pressure level about 200 mtorr inside the package [37.81].

This type of post-packaging process at wafer level has become the favorite approach to fabricate a hermetic encapsulation because it can provide lower cost and more process flexibility. But the packaging process relies on "good" bonding techniques. A strong and reliable bonding between two substrates should be provided, and this procedure should be compatible with the other microsystem fabrication processes.

37.2.3 Localized Heating and Bonding

The approach of MEMS post-packaging by localized heating and bonding is proposed to address the problems of global heating effects. In this chapter, resistive micro-heaters are used as the examples to provide localized heating, although several other means of localized heating have been demonstrated recently, including laser welding [37.82], inductive heating [37.83], and ultrasonic bonding [37.84]. The principle of localized heating is to achieve high temperature for bonding while maintaining low temperature globally at the wafer level. Resistive heating by using micro-heaters on top of device substrate is applied to form strong bond with silicon or glass cap. According to the results of a 2-D heat conduction finite element analysis as shown in Fig. 37.7, the steady-state heating region of a 5-µm-wide polysilicon micro-heater capped with a Pyrex glass substrate

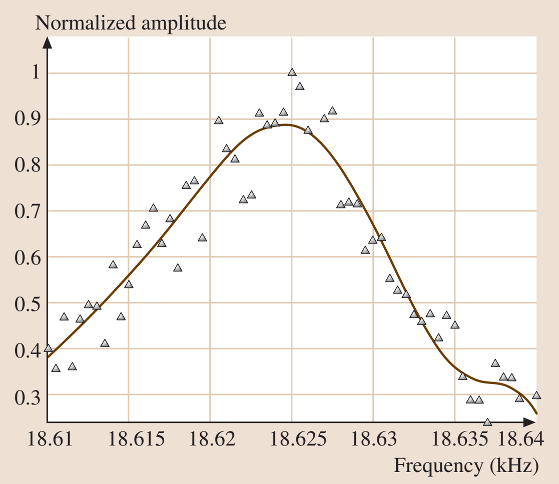

Fig. 37.6 Spectrum measurement results of a vacuum encapsulated comb-shape resonator by using the RTP aluminum-to-nitride bonding method [37.45]

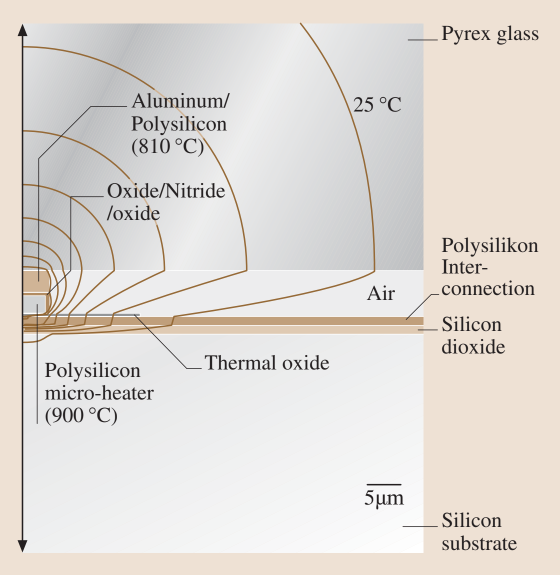

Fig. 37.7 The schematic diagram of 2-D heat transfer model, geometry and boundary conditions

can be confined locally as long as the bottom of the silicon substrate is constrained to the ambient temperature. The physics of localized heating behind this design can be understood by solving the governing heat conduction equations of the device structure without a cap [37.81]. As long as the width of micro-heater and the thickness of silicon substrate are much smaller than the die size and a good heat sink is placed underneath the silicon substrate, the heating can be confined locally. The temperature of silicon substrate can be kept low or close to room temperature. Several localized resistive heating and bonding techniques have been successfully developed for packaging applications including localized silicon-to-glass fusion bonding, gold-to-silicon eutectic bonding, and localized solder bonding. Several solder materials have been successfully tested, including PSG, indium, and aluminum alloy [37.85].

The vacuum packaging example presented in this chapter is based on localized aluminum/silicon-to-glass solder bonding technique. Built-in folded-beam comb drive μ-resonators are used to monitor the pressure of the package. Figure 37.8 shows the fabrication process of the package and resonators. Thermal oxide (2 μm) and LPCVD Si_3N_4 (3,000 Å) are first deposited on a silicon substrate for electrical insulation followed by the deposition of 3,000 Å LPCVD polysilicon. This polysilicon is used as both the ground plane and the electrical interconnect to the μ-resonators as shown in Fig. 37.8a. Figure 37.8b shows a 2 μm LPCVD SiO_2 layer that is deposited and patterned as a sacrificial layer for the fabrication of polysilicon μ-resonators using a standard surface micromachining process. A 2-μm-thick phosphorus doped polysilicon is used for both the structural layer of microresonators and the on-chip micro-heaters. This

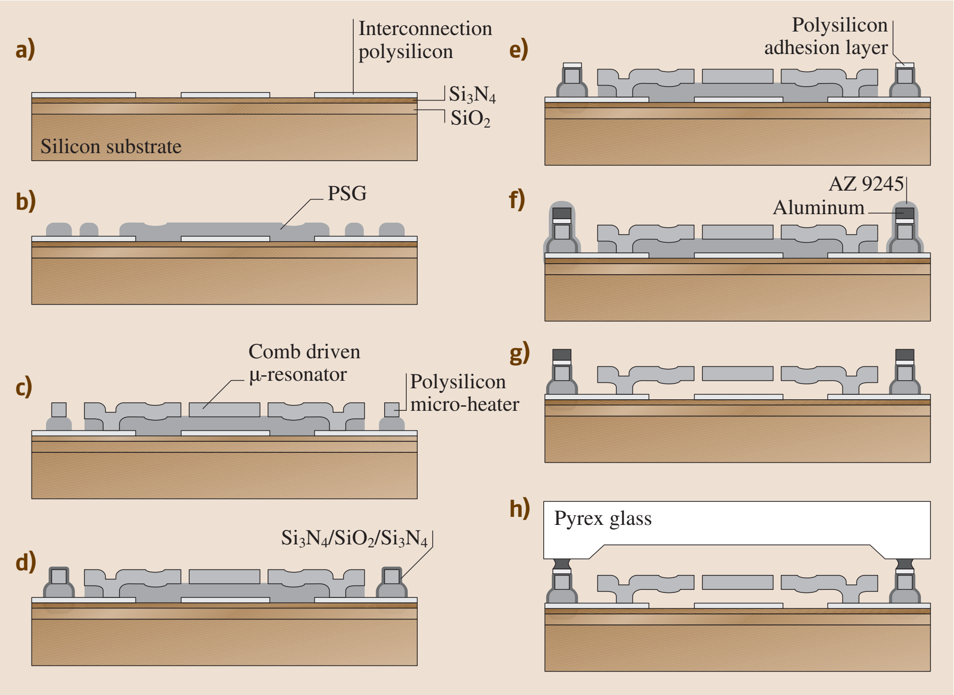

Fig. 37.8a–h The schematic process flow of vacuum encapsulation using localized aluminum/silicon-to-glass bonding, **(a)** electrical ground plane and device interconnects fabrication **(b)** a 2 μm LPCVD SiO_2 sacrificial layer deposition for the fabrication of polysilicon μ-resonator **(c)** a 2 μm-thick phosphorus doped polysilicon deposition and patterning for both the structural layer of micro resonators and the on-chip microheaters **(d)** a LPCVD Si_3N_4 (750 Å)/SiO_2 (1,000 Å)/Si_3N_4 (750 Å) sandwich layer deposition and patterning on top of the microheater for the electrical isolation **(e)** an aluminum (2.5 μm) and polysilicon (5,000 Å) bonding materials are deposition and patterning. **(f)** a thick AZ 9245 photoresist deposition and patterning for the protection of aluminum/silicon-to-glass bonding layers **(g)** 20 minutes sacrificial release in concentrated HF **(h)** vacuum encapsulation with 0.2 MPa contact pressure and 3.4 W input power under a 25 mtorr vacuum environment

layer is formed over the sacrificial oxide in two steps to achieve a uniform doping profile. Lower input power and better process compatibility are two major advantages in using the on-chip micro-heater in the glass package. The resonators are separated from the heater by a short distance, 30 µm, to effectively prevent their exposure to the high heater temperature as shown in Fig. 37.8c. This concludes the fabrication of µ-resonators.

In order to prevent the current supplied to the micro-heater from leaking into the aluminum solder during bonding, a LPCVD Si_3N_4 (750 Å)/SiO_2 (1,000 Å)/Si_3N_4 (750 Å) sandwich layer is grown and patterned on top of the micro-heater as shown in Fig. 37.8d. Figure 37.8e and f show that aluminum (2.5 µm) and polysilicon (5,000 Å) bonding materials are deposited and patterned. The sacrificial release is the final step to form freestanding µ-resonators. Figure 37.8f shows a thick AZ 9245 photoresist is applied to cover aluminum/silicon-to-glass bonding system to ensure that the system withstands the attack from concentrated hydrofluoric acid. After 20 minutes sacrificial release in concentrated HF, the system as shown in Fig. 37.8g is ready for vacuum packaging. A Pyrex glass cap with a 10-µm deep recess is then placed on top with an applied pressure of ∼ 0.2 MPa under a 25 mtorr vacuum, and the heater is warmed using 3.4 watt input power (exact amount depends on the design of the micro-heaters) for 10 minutes to complete the vacuum packaging process as shown in Fig. 37.8h.

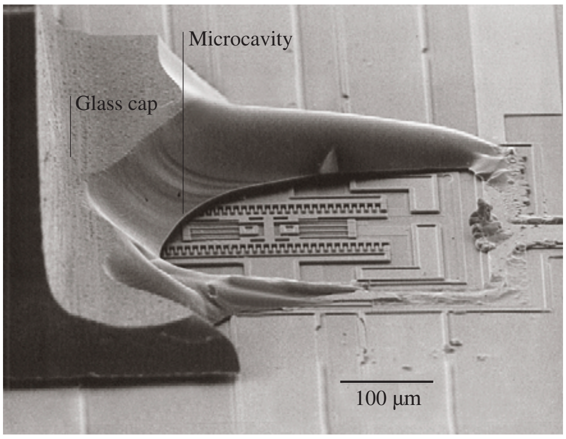

Fig. 37.9 The SEM microphoto of encapsulated micro-resonators after the glass cap is forcefully broken

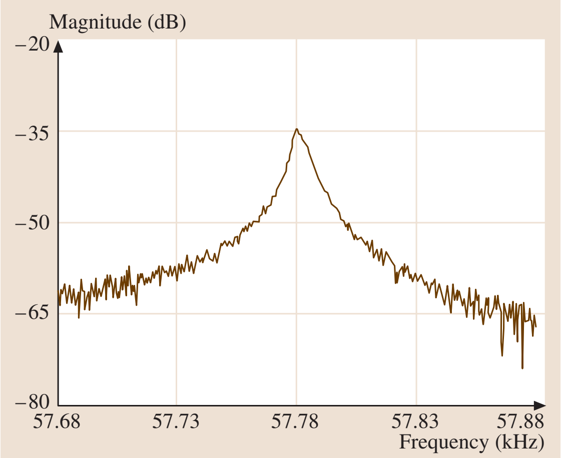

Fig. 37.10 The transmission spectrum of a glass-encapsulated µ-resonator after 120 minute pump down time in vacuum environment ($Q = 9{,}600$)

To evaluate the integrity of the resonators packaged using localized aluminum/silicon-to-glass solder bonding, the glass cap is forcefully broken and removed from the substrate. No damage is found on the µ-resonator and a part of the micro-heater is stripped away as shown in Fig. 37.9, demonstrating that a strong and uniform bond can be achieved without detrimental effects on the encapsulated device. Figure 37.10 shows a vacuum encapsulated unannealed µ-resonator (∼ 57 kHz) after

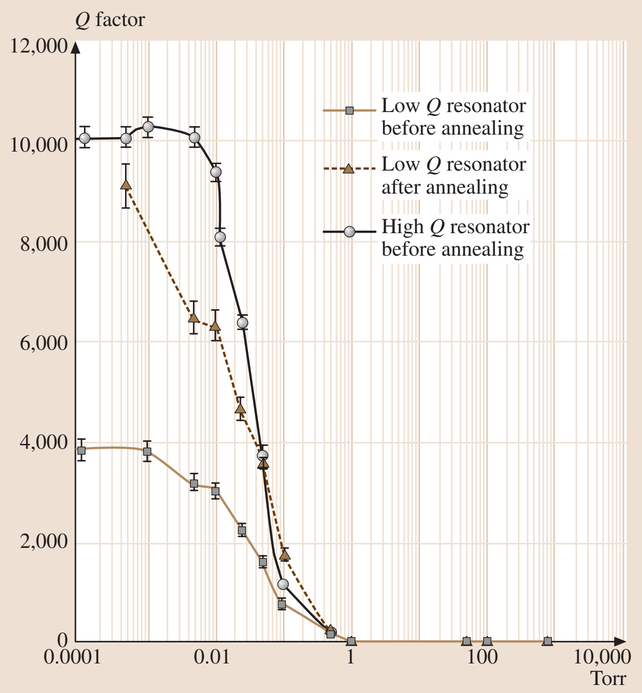

Fig. 37.11 Measured Q factor vs. pressure of unpackaged µ-resonators

120 minutes of wait time. The measured Q-factor after packaging is 9,600. Based on the measurement of Q vs. pressure of a high Q unpackaged μ-resonator as shown in Fig. 37.11, it is demonstrated that the pressure inside the packaging is comparable to the vacuum level of packaging chamber.

The post-process packaging using localized heating and bonding technique includes four basic components: (1) an electrical and thermal insulation layer such as silicon dioxide or silicon nitride should be used for localized heating, (2) resistive micro-heaters will be fabricated to provide the heating source for localized bonding, (3) materials, including metal and polysilicon that can provide good bonding and hermeticity with silicon or glass substrates are considered as the bonding materials, and (4) a good heat sink under the device substrate for localized heating is provided during the bonding experiments. MEMS devices will be fabricated on the device chip and hermetically sealed in the cavity formed by the device chip, resistive micro-heaters, and protection cap. The process can be either die level or wafer level. The schematic design of wafer-level packaging process is shown in Fig. 37.12. The resistive micro-heaters are parallel to each other and connected in order to ensure that identical current density is applied for individual packages at the same time. These heaters can be fabricated either on the chip or protection cap and can be built in a larger wafer for current inputs. The interconnections for these packaging cavities can be built in the dicing area, so no extra space is required for the packaging process.

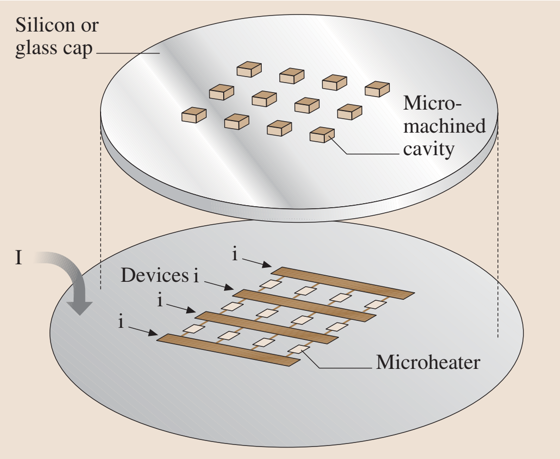

Fig. 37.12 Illustration of wafer-level vacuum packaging in the wafer level using localized heating and bonding technique

37.3 Thermal Issues and Packaging Reliability

37.3.1 Thermal Issues in Packaging

The two key thermal issues related to MEMS packaging are: 1) heat dissipation from actuators and integrated circuitry components and 2) thermal stress generated during the packaging process. These two topics are discussed separately.

Heat Dissipation Issues

In IC, heat dissipation is a serious problem when the size of transistor continues to shrink and the density of transistor on a chip continue to increase with the advance of IC fabrication technology. The trend of power packing into smaller packages has created increasing thermal management challenges [37.86]. Since the electrical characteristics of transistors change with working temperature, inefficient power dissipation could raise the working temperature and affect device performance. Present MEMS devices don't need high power, high performance microprocessors so power dissipation is not a problem. Nevertheless, some functional components in packaged MEMS are very sensitive to temperature variation, such as biomaterials or laser diodes. Several MEMS chemical sensors and other applications such as micro PCR chamber for DNA replication actually require elevated temperature for operation and microthermal platforms are built for these devices. Thermal management to maintain working temperature on these chips for stable operation is still an essential issue for packaging considerations. The geometrical complexity of MEMS resulting from packing various functional components in the tight space increases the difficulty of the thermal management. As the MEMS packaging integration process becomes more complex, the fabrication constraint in the packaging process will have great impact on the heterogeneous integration process in the front-end MEMS and IC processes. For example, low temperature requirement at the packaging process generally limits the possible choices of materials in the back-end process. In general, conventional IC packaging employs a heat sink attached to the chip to remove heat. The heat sink is generally made of a copper or stainless steel bar with an array of fin structures on one side for better natural or forced heat convection to dissipate heat into the environment. In addition to heat sinks, thermal vias, heat pipe cooling, immersion cooling, and thermoelectric cooling can also be used for effective heat removal. Because most MEMS packages still follow the typical IC packaging architecture, one promising thermal management method, heat pipe, is discussed for possible MEMS packaging applications.

A heat pipe is a sealed slender tube containing a wick structure and a working fluid, typically water in electronics cooling. It is composed of three sections, the evaporator section at one end, condenser section at the other end, and adiabatic section in the middle. In the evaporator section, heat is absorbed by the working fluid via phase transformation from liquid to vapor. In the section of condenser, heat dissipates into the outside environment. Thus the fluid goes back to liquid phase. Vapor phase is at a high pressure and temperature state that forces the vapor to flow into the condenser section at a lower temperature. Once the vapor condenses and gives up its latent heat, the condensed fluid is pumped back to the evaporator section by the capillary force developed in the wick structure. Therefore the middle adiabatic section contains two phases, the vapor phase in the core region and the liquid phase in the wick, flowing opposite to each other with no significant heat transfer between the fluid and the surrounding medium. Silicon has good thermal conductivity (1.41 W/cm/°C) and is easily micro-machined to fabricate the heat pipe. The implementation of the silicon micro heat pipe in IC and MEM packaging has therefore great potential and several approaches have been proposed on this topic [37.87–89].

Packaging Induced Thermal Stresses

Thermal based bonding processes have been used in MEMS packaging applications for many years, as introduced previously in this chapter. Thermal management is extremely important during the bonding process in order to avoid fracture in the substrate or MEMS device itself. Extremely high temperatures or rapid cooling conditions can cause damage and should be carefully evaluated both analytically and experimentally. There are many ways to provide heating energy including electrical resistive heating, oven heating or induction heating [37.83]. These bonding processes may be put into two categories. The first is localized bonding in which the heat is directly applied only to the adhesive material used to bond the package to the MEMS device. The second category is global heating where the entire system (MEMS device, adhesive, and packaging material) is heated to produce bonding of the materials. The latter is the common approach for all MEMS packaging processes. This section focuses therefore on the thermal stress effects during the heating and cooling procedures in MEMS packaging. The RTP aluminum-to-glass bonding process is used as the specific example for the discussion of thermal stresses [37.90]. The bonding process heats the packaging system to 750 °C for 10 seconds, and then cools it back to room temperature. To simulate this process, an ANSYS program [37.91] is established to examine the shear stress due to CTE (Coefficient of Thermal Expansion) variations in the bonding system as the result of temperature changes. The shear stress was recorded from the ANSYS analysis on the aluminum/Pyrex glass interface and the aluminum/silicon interface.

Two different models were analyzed. The first was the quartz-aluminum-silicon bonding system, and the second was the Pyrex glass-aluminum-silicon bonding system [37.92]. The results of the ANSYS analysis were then analyzed and compared with experimental observations. Figure 37.13 shows the ANSYS results for a Pyrex glass bonding system: the width of the aluminum solder is 100 μm and the maximum residual stress is be 60 MPa in glass, slightly lower than the fracture strength of Pyrex glass at 70 MPa. It was discovered that increasing the aluminum width leads to lower residual stresses. This likely occurred because the length of the Pyrex glass, quartz, and silicon remains constant. As a result, it will always want to contract the same amount with a constant temperature change independent of the aluminum width. When increasing the aluminum width, however, the stress is not in such a concentrated area and therefore decreases. For example, the maximum residual stress an-

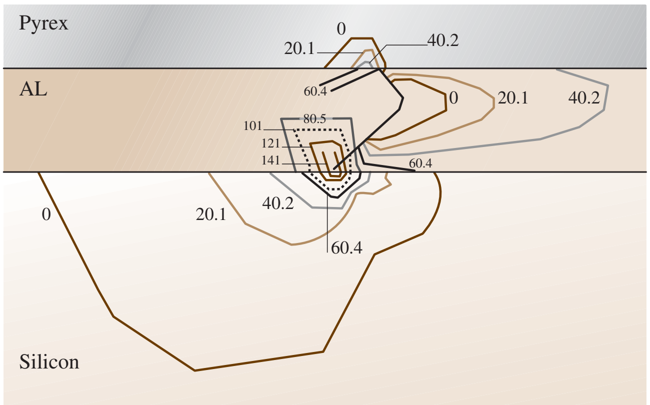

Fig. 37.13 Residual Stress (GPa) for an aluminum solder width of 100 μm under the RTP silicon-aluminum-glass bonding system

alyzed from ANSYS in the Pyrex glass bonding systems is 74.5 Gpa, 58 Gpa, and 60 GPa for aluminum widths of 30, 50, and 100 μm, respectively [37.92]. Pyrex glass has a documented strength of around 69 GPa [37.93]. Fracture should always occur with an aluminum width of 30 μm or less, according to the ANSYS analysis. Fracture may occur sporadically at widths of 50 or 100 μm, depending on the amount and magnitude of the flaws in the Pyrex glass. Experiments were done on the Pyrex glass bonding systems with a width of 100 μm. The samples were heated to 750 °C and then cooled by taking them out of the oven. In all four experimental cases, small cracks were observed in Pyrex glass as shown in Fig. 37.14. These cracks may have occurred consistently for several reasons. First, it may be a result of handling of the Pyrex glass before bonding. The Pyrex glass samples were kept in containers with each other, which may have resulted in abrasive contact and possibly caused flaws in the materials. These flaws can cause a reduced strength from the 69 GPa and therefore the predictions from the ANSYS analysis could be correct. Second, the observation is made that the cracks are small, only occurring tens of microns away from the aluminum and not propagating completely through the Pyrex glass. These cracks could be caused by the high stress applied, but the crack has not reached a critical size and therefore has not propagated completely through the Pyrex glass. Therefore, the strength remains at the theoretical strength of 69 GPa, and the Pyrex glass is only partially cracked. Experimental analysis done by *Chaio* and *Lin* [37.90] shows that fracture was not observed when using aluminum widths greater than 150 μm. This is consistent with the results of the ANSYS analysis which show that as the width of the aluminum is increased, the residual stress decreases.

The quartz bonding system ANSYS stress predictions are conducted, and the maximum stress predicted in the analysis is 207 GPa, 117 GPa, and 100 GPa, for aluminum widths of 30 μm, 50 μm, and 100 μm, respectively. All three of these stresses are much larger than the theoretical strength of Quartz at 48 GPa, and therefore fracture should always occur. Quartz has a much larger coefficient of thermal expansion (CTE) than silicon, which is why this was predicted. Experimentally, a quartz substrate was tested for the silicon-aluminum-quartz bonding test and the result is shown in Fig. 37.15.Cracks happen all over the place and cause serious damages on the quartz wafer. These cracks could be the failure mechanism of the hermetic package. Pyrex glass is therefore identified as a better bonding substrate than quartz.

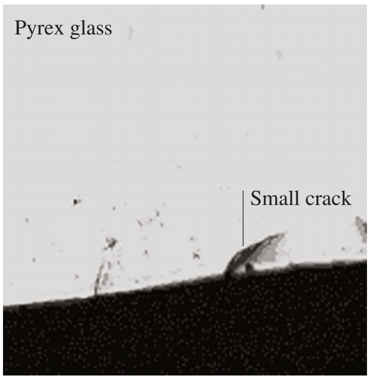

Fig. 37.14 Microphoto of the experimental result on Pyrex Glass-Aluminum-Silicon system. Small cracks can be observed

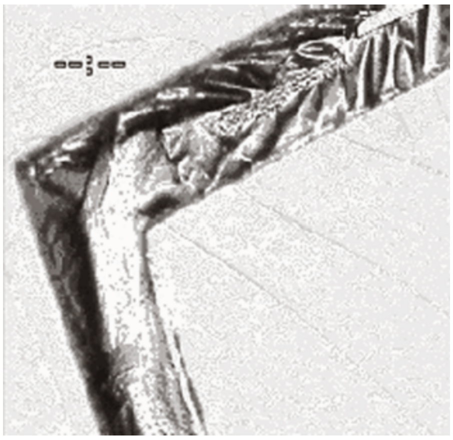

Fig. 37.15 Bonding result of Quartz-Aluminum-Silicon system, fracture can be observed

The thermal stresses generated in the packaging process in quartz are much larger than the stresses in Pyrex glass because of the difference between the CTE in the two systems. Quartz has a low CTE ($0.54\times10^{-6}\,K^{-1}$) compared to the aluminum CTE ($23\times10^{-6}\,K^{-1}$) and silicon CTE ($3.5\times10^{-6}\,K^{-1}$). On the other hand Pyrex glass has a much closer CTE ($3.2\times10^{-6}\,K^{-1}$) to silicon and aluminum, resulting in smaller stresses. The practical implications of the ANSYS results and the preceding information are that materials must be chosen carefully when carrying out bonding. Material should not be used with a much higher or lower CTE than silicon to ensure fracture will not occur. This finding is a valuable component of this particular packaging system and should benefit the other packaging processes involving bonding MEMS packaging because it supports the prediction that Pyrex glass is an excellent material to use when bonding with silicon in MEMS packaging, as long as sufficiently wide adhesive material is used.

37.3.2 Packaging Reliability

The thermal stress induced by the CTE mismatch is one of the main factors that affect the packaging reliability. As a matter of fact, the formation of the stress can happen not only during packaging process but also during the operation of devices. In particular, during device operation, the package will go through various temperature cycles because of environment change. Such temperature variation causes the expansion of packaging materials when they are constrained by the packaged assembly. As a result of such thermal mismatch, significant stresses are induced in the package and finally cause the device to fail. In addition to thermal mismatch, corrosion creep, fracture fatigue, crack initiation and propagation, and delamination of thin films are all possible factors to cause failure of packaged devices [37.94]. These failure mechanisms could be prevented or deferred under proper packaging designs. For instance, the thermally induced strain inside the packaging material is generally below the tolerance of the material and can't cause immediate catastrophic damage. But cyclic loading can generate and accumulate stresses, eventually causing failure. Several common designs in IC packaging have been used to prolong the lifetime of devices. For example, the strain in solder interconnects of BGA or flip-chip packaging can be effectively reduced by introducing a polymer underfill material between the chip and the substrate to distribute thermal stress induced by CTE mismatch [37.95]. The strain can further be reduced if excellent thermal paths are built around interconnects to diminish the thermal stress originated from the temperature gradient between ambient and operation temperatures. Delamination phenomenon is another source to cause reliability problems, and it happens in the interface of adjacent material layers. In MEMS, components made of dissimilar materials are commonly bonded together to provide specific functions. Delamination can result in electrical or mechanical failures of devices in the package such as mechanically cracking through the electrical via wall to make an electrical open because of the propagation of the delamination of metal line from the dielectric layer or over heating of the die because of the delamination of die from its underneath to make an opening on the heat dissipation path. Because the stress and thermal loading, the geometry, and the material properties are complex in MEMS, the development of the packaging designs to increase the reliability is very important and requires more extensive investigations.

Reliability testing is required before a new device can be delivered to the market. The test results can provide the information for the following improvement of packaging design and fabrication. How to analyze the failure data, called the "reliability metrology", is thus very important in the packaging industry. The analysis method is to use the mathematical tools of probability and statistical distributions to evaluate data to understand the patterns and identify the sources of failure. For example, a failure density function is defined as the time derivative of the cumulative failure function

$$f(t)=\frac{\mathrm{d}F(t)}{\mathrm{d}t}\,, \tag{37.1}$$

$$F(t)=\int_0^t f(s)\,\mathrm{d}s\,. \tag{37.2}$$

The cumulative failure function, $F(t)$, is the fraction of a group of original devices that has failed at time t. The Weibull distribution function is one of the analytic mathematic models commonly used in the packaging reliability evaluation to represent the failure density function [37.21].

$$f(t)=\frac{\beta}{\lambda}\left(\frac{t}{\lambda}\right)^{\beta-1}\exp\left[-\left(\frac{t}{\lambda}\right)^{\beta}\right]\,, \tag{37.3}$$

where β and λ are the Weibull parameters. The parameters β, called a shape factor, measures how the failure frequency is distributed around the average lifetime. The parameter λ is called the lifetime parameter and indicates the time at which 63.2% of the devices faile. By integrating both sides of the equation (37.3). $F(t)$ becomes:

$$F(t)=1-\exp\left[-\left(\frac{t}{\lambda}\right)^{\beta}\right]\,. \tag{37.4}$$

Using the Weibull distribution function with the two parameters extrapolated by experimental data, one can estimate the number of failures at any time during the test. Moreover, by knowing the meaning and values of the parameters, one can compare two sets of test data. For example, the λ with greater number means this set of samples has longer lifetime. Because all of the mathematical models are statistical approximations based on the real experimental data, more testing samples can provide more accurate estimation.

37.3.3 Long-Term and Accelerated MEMS Packaging Tests

The reliability of MEMS packages is best characterized by means of long-term tests with statistical data analyses. But it is very difficult and time-consuming to obtain these experimental data so accelerated tests that put samples in extreme environment for accelerated failure are commonly used to predict the lifetime of the devices. Unfortunately, few research publications deal with these two issues. In this section, two MEMS packaging examples that aim to address these two tests we discuss.

Figure 37.16 shows long-term measurements of the Q factor of a vacuum packaged µ-resonators using localized aluminum/silicon-to-glass bonding [37.81] vacuum encapsulation process is described in details as shown in Fig. 37.8. It is found that the vacuum package by means of localized heating and bonding provides stable vacuum environments for the µ-resonator and a quality factor of 9,600 has been achieved with no degradation for at least one year. Since the performance of high Q µ-resonator is very sensitive to environmental pressure as shown in Fig. 37.11, any leakage can be easily detected. The fact that this high Q value can hold for one year indicates the packaging process is well performed and both aluminum and Pyrex glass are suitable materials for vacuum packaging applications. According to a previous study of hermeticity in different materials, metal has lower permeability to moisture than other materials such as glass, epoxy, and silicon. With a width of 1 µm, metal can effectively block moisture for more than ten years [37.21]. In this vacuum package system, the bonding width is 30 µm so it can sufficiently block the diffusion process of moisture. On the other hand, the diffusion effects of air molecules into these tiny cavities have not been studied extensively, and the design guidelines for vacuum encapsulations are not clearly defined. Further investigations will be needed in this area, and the example presented here serves as a good starting point.

On the other hand, accelerated testing puts a large amount of samples in harsh environments, such as elevated temperature, elevated pressure, and 100% humidity, to accelerate the corrosion process. The statistical failure data are gathered and analyzed to predict the lifetime of packages under normal usage environment. As a result, the long-term reliability of the package can be predicted without going through the true long-term tests. Unfortunately, accelerated tests have not been addressed in MEMS research papers. Although the MEMS industry must have done some extensive reliability tests, they do not publish their results, probably due to liability concerns. Among the very limited publications, this section uses as the illustration example a specific MEMS packaging system that has gone through accelerated tests [37.45].

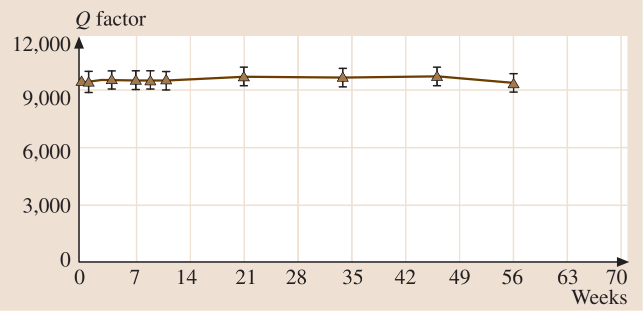

Fig. 37.16 Long-term measurement of encapsulated µ-resonators. No degradation of Q factors is found after 56 weeks

The MEMS package is accomplished by means of RTP (Rapid Thermal Processing) bonding as described previously in this chapter. The goal of the accelerated test is to examine the failure rate at the bonding interface. The accelerated tests start by putting the packaged samples into the autoclave chamber filled with high temperature and pressurized steam at 130 °C, 2.7 atm, and 100% relative humidity for accelerated testing. The pressurized steam can penetrate small cracks if any defect exists at the bonding interface [37.96]. Elevated temperature and humid environment speed up the corrosion process. A package is considered a failure if water is condensed or diffused into the package. Statistical failure data are gathered every 24 h under optical examination for a period of 864 h, during which time new failure is seldom observed. Because of the robustness of the samples, it is difficult to conduct the tests to the point at which all the packages may fail. The cumulative failure function is recorded, and it is found that most failures occur in the first 96 h and such high early failure rate reflects the yielding issue of the bonding process. Weibull and Lognormal models are compared to predict the lifetime of the packages, and it is found that Lognormal model better describes the statistical data. Figure 37.17 shows the inverse standard normal distribution function vs. ln(time) and the Maximum Likelihood Estimator (MLE) is then used to predict the mean, standard deviation, and the MTTF (mean time to failure). Table 37.2 shows the MLE calculation results of MTTF. The wide interval of confidence level comes from the fact that only a small number of samples failed at the end of the test. It is also observed that packages with larger bonding width and smaller bonding areas have larger MTTF values. The lower bound of the MTTF provides the worst-case scenario. For example, only 4 our of 31 samples failed when tests stopped in the case of ring width of 200 µm and sealing area of 450 × 450 µm^2. The MTTF predicts, in the worst-case scenario, that there

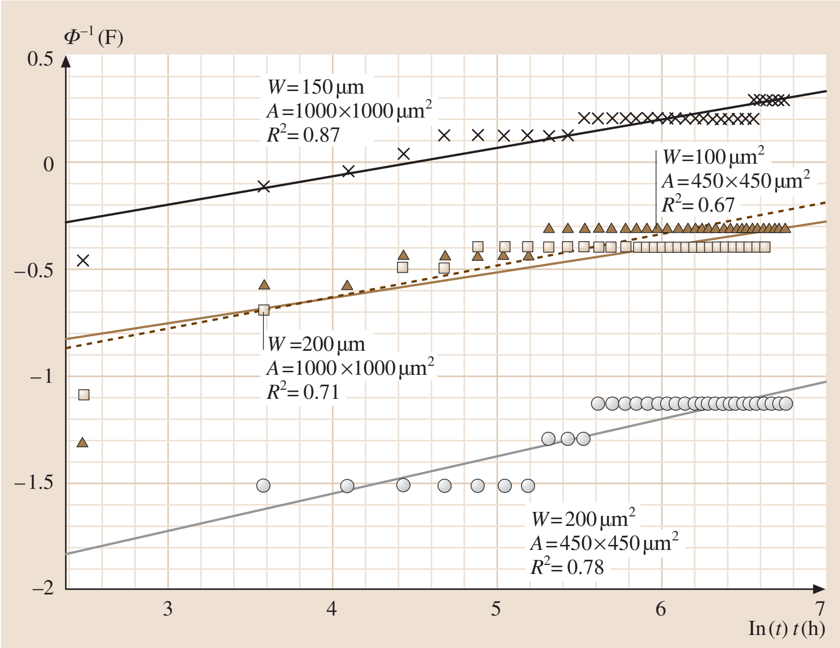

Fig. 37.17 Life data fitted by Lognormal distribution. R^2 is the coefficient of determination

Table 37.2 The Maximum Likelihood Estimation for Mean Time To Failure (MTTF)

MLE calculation resuls of MMTF[a]				
Bonding width, W (µm)	**Area, A (µm²)**	**MTTF**		**Worst cases in jungle condition (years)**
		UB (years)	**LB (years)**	
200	450× 450	1.8×10^7	0.57	1,700
100	450× 450	5.3	0.10	300
200	1,000×1,000	6.5×10^3	0.09	270
150	1,000×1,000	0.50	0.017	50

[a] UB is the upper bound, and LB is the lower bound of the 90% confidence interval, respectively. The MTTF LB times AF is the worst case MTTF used in jungle condition.

is 90% chance a package will fail in 0.57 years in the autoclave environment.

It is widely accepted that the acceleration factor (AF) for autoclave tests follows the Arrhenius equation [37.21] and can be modeled as:

$$\mathrm{AF} = \frac{(\mathrm{RH}^{-n}\,\mathrm{e}^{\Delta E_\mathrm{a}/kT})_\mathrm{normal}}{(\mathrm{RH}^{-n}\,\mathrm{e}^{\Delta E_\mathrm{a}/kT})_\mathrm{accelerated}}\,, \tag{37.5}$$

where RH is relative humidity (85%, RH = 85), k the Boltzmann constant and T the absolute temperature. The recommended value for n, an empirical constant, is 3.0 [37.97], and ΔE_a, the activation energy, is 0.9 eV for plastic dip package and 0.997 eV for anodically bonded glass-to-silicon package [37.58]. If $\Delta E_\mathrm{a} = 0.9$ is used to estimate the AF for the accelerated testing condition as compared with the jungle condition (35 °C, 1 atm and 95% RH), AF is about 3,000, and the worst-case lifetime values in jungle condition are also listed in Table 37.2. The high values of estimated MTTF in jungle condition could be a result of overestimation of AF because plastic dip package may have smaller AFs than those of glass packages. Nevertheless, these data and analyses provide important guidelines in the area of accelerated tests for MEMS packages.

37.4 Future Trends and Summary

In the past, the development of MEMS packaging is mainly originated from IC packaging advancement because existing packaging techniques can significantly reduce the development cost of MEMS. It is expected that the situation will be changed very soon, hoewever, in a way the MEMS packaging approaches will assist the IC packaging development. Recent advances in IC packaging aim to provide high I/O density and more chip integration capability for the needs of high speed and high data communication rate. In order to satisfy those requirements, several packaging concepts and techniques are developed, including 3-D packaging, wafer-level packaging, BGA, and flip-chip technique. Although all of the concepts and methods can provide a package with more I/O density, flexibility in chip integration, and lower manufacturing cost for IC fabrication, they are still insufficient for providing solutions for the future applications because of the increase of complexity and requirements of MEMS packaging. On the contrary, with the progress of MEMS fabrication technologies, several key processes such as DRIE (Deep Reactive Ion Etching), wafer bonding, and thick photoresist processes [37.98] have been utilized for IC packaging fabrication. Technologies developed in MEMS fabrication can therefore also assist the development of new IC packaging approaches.

In order to address the future needs of process integration, adaptive multi-chip module (MCM) [37.28] or 3-D packaging combined with vertical through substrate interconnects [37.10, 99] are promising approaches for the development of future MEMS packaging processes. Based on low temperature flip-chip solder bonding technique, these packaging methods can provide more flexibility in device fabrication and packaging. Devices can be fabricated before they are integrated to form microsystems to dramatically reduce the packaging cost. Vertical through substrate interconnects can have higher I/O density, smaller resistance, parasitic capacitance, and mutual inductance. Although this approach provides many possible advantages, technical challenges exist. For instance, metal is commonly used as the fill material inside the vertical vias as electrical interconnects, and it may introduce large thermal mismatch with respect to silicon substrate and generate huge thermal stress to cause packaging reliability problems. Moreover, it will be an interesting engineering challenge to fill the materials into those high aspect ratio vias.

The future development of MEMS packaging depends on the successful implementation of unique techniques as described in the following:

1. Development of mechanical, thermal, and electrical models for packaging designs and fabrication processes;
2. Wafer-level, chip-scale packaging with low packaging cost and high yield;
3. Effective testing techniques at the wafer level to reduce the testing costs;
4. Device integrations by vertical through-interconnects as an interposer [37.1] to avoid thermal mismatch problems.

In addition to these approaches and challenges, many other possibilities have not been listed but also require dedicated investigations. For example, several key nanotechnologies have been introduced in the previous chapters, but the packaging solutions for the NEMS devices have not been addressed. Because it is feasible to use MEMS as the platform for NEMS fabrication, all of the packaging issues discussed in the chapter can be directly applied to NEMS devices. On the other hand, nanotechnology may introduce new opportunities for MEMS/NEMS packaging applications by providing superior electrical, mechanical, and thermal properties [37.100–103]. For example, carbon nanotubes have very high thermal conductivity [37.102] and may be suitable to increase the thermal cooling effects for better IC/MEMS/NEMS packaging applications.

In summary, MEMS packaging issues have been introduced in the areas of fabrication, application, reliability, and future development. Packaging design and modeling, packaging material selection, packaging process integration, and packaging cost are the main issues to be considered when developing a new MEMS packaging process.

References

37.1 M. Matsuo, N. Hayasaka, K. Okumura, E. Hosomi, C. Takubo: Silicon interposer technology for high-density package, IEEE ECTC (2000) 1455–1459

37.2 K. E. Peterson: Silicon as a mechanical material, Proc. IEEE **70** (1982) 420–457

37.3 L. Lin, H.-C. Chu, Y.-W. Lu: A simulation program for the sensitivity and linearity of piezoresistive pressure sensors, IEEE/ASME J. Microelectromech. Syst. **8** (1999) 514–522

37.4 Y. C. Tai, R. S. Muller: Lightly-doped polysilicon bridge as a flow meter, Sens. Actuators **15** (1988) 63–75

37.5 D. Hicks, S.-C. Chang, M. W. Putty, D. S. Eddy: Piezoelectrically activated resonant bridge micro-accelerometer, 1994 Solid-State Sensors and Actuators Workshop (1994) 225–228

37.6 J. Berstein, S. Cho, A. King, A. Kourepenis, P. Maciel, M. Weinberg: A micromachined comb-drive tuning fork rate gyroscope, 6th IEEE Int. Conf. on MEMS (1993) 143–148

37.7 M. Madou: Compatibility and incompatibility of chemical sensors and analytical equipment with micromachining, Solid-State Sensor and Actuator Workshop, Hilton Head (1994) 164–171

37.8 M. J. Zdeblick, J. B. Angell: A microminiature electric-to-fluidic valve, Proceedings of Transducers'87, the 4th International Conference on Solid-State Transducers and Actuators (1987) 827–829

37.9 J. H. Tsai, L. Lin: A thermal bubble actuated micro nozzle-diffuser pump, IEEE/ASME J. Microelectromech. Syst. **11** (2002)

37.10 S. J. Ok, D. Baldwin: High density, aspect ratio through-wafer electrical interconnect vias for low cost, generic modular MEMS packaging, 8th IEEE Int. Symp. On Adv. Pack. Mater. (2002) 8–11

37.11 H. Reichl, V. Grosser: Overview and development trends in the field of MEMS packaging, 14th IEEE Int. Conf. on MEMS (2001) 1–5

37.12 P. Kapur, P. M. McVittie, K. Saraswat: Technology and reliability constrained future copper interconnects=Part II: Performance implication, Trans. Electron. Dev. **49** (2002) 598–604

37.13 J. Lau: *Flip Chip Technologies* (McGraw-Hill, New York 1996)

37.14 R. Prasad: *Surface Mount Technology: Principles and Practice*, 2nd edn. (Chapman, New York 1989)

37.15 M. Töpper, J. Auersperg, V. Glaw, K. Kaskoun, E. Prack, B. Keser, P. Coskina, D. Jäger, D. Petter, O. Ehrmann, K. Samulewicz, C. Meinherz, S. Fehlberg, C. Karduck, H. Reichl: Fab integrated packaging (FIP) a new concept for high reliability wafer-level chip size packaging, IEEE, ECTC (2000) 74–81

37.16 S. Savastiouk, O. Siniaguine, Ed. Korczynski: 3D wafer level packaging, IEEE, ECTC. (2000) 26–31

37.17 J. H. Tsai, L. Lin: Micro-to-macro fluidic interconnectors with an integrated polymer sealant, J. Micromech. Microeng. **11** (2001) 577–581

37.18 T. A. Core, W. K. Tsang, S. Sherman: Fabrication technology for an integrated surface-micromachined sensor, Solid State Technol. (October 1993) 39–47

37.19 K. Koester, R. Majedevan, A. Shishkoff, K. Marcus: *Multi-Sser MEMS Processes (MUMPS) Introduction and Design Rules* (MCNC MEMS Technology Applications Center, Research Triangle Park, NC 27709 1996)

37.20 R. Lengtenberg, H. A. C. Tilmans: Electrically driven vacuum-encapsulated polysilicon resonantor, Part I: Design and fabrication, Sens. Actuators A **45** (1994) 57–66

37.21 R. R. Tummala, E. J. Rymaszewski, A. G. Klopfenstein: *Microelectronics Packaging Handbook, Semiconductor Packaging* (Chapman, New York 1997)

37.22 S. Wolf: *Silicon Processing for the VLSI era, Vol. I: Process Technology* (Lattice Press, Sunset Beach 1995)

37.23 J. K. Bhardway, H. Ashraf: Advanced silicon etching using high density plasmas, SPIE Micromach. Fabr. Technol. **2639** (1995) 224–233

37.24 W. H. Ko, J. T. Suminto, G. J. Yeh: *Bonding Techniques for Microsensors, Micromachining and Micropackaging for Transducers* (Elsevier, Amsterdam 1985)

37.25 S. D. Senturia, R. L. Smith: Microsensor packaging and system partitioning, Sens. Actuators **15** (1988) 221–234

37.26 R. L. Smith, S. D. Collins: Micromachined packaging for chemical microsensors, IEEE Trans. Electron Dev. **35** (1988) 787–792

37.27 J. T. Butler, V. M. Bright, P. B. Chu, R. J. Saia: Adapting multichip module foundries for MEMS packaging, Int. Conf. on Multi. Mod. High Den. Pack. (1998) 106–111

37.28 M. Schuenemann, A. J. Kourosh, V. Grosser, R. Leutenbauer, G. Bauer, W. Schaefer, H. Reichl: MEM modular packaging and interfaces, IEEE ECTC (2000) 681–688

37.29 D. W. Lee, T. Ono, T. Abe, M. Esashi: Fabrication of microprobe array with sub-100nm nano-heater for nanometric thermal imaging and data storage, Proc. IEEE Micro Electro Mechanical Systems Conf., IEEE Int. Conf. on MEMS (January 2001) 204–207

37.30 A. S. Laskar, B. Blythe: Epoxy multichip modules, a solution to the problem of packaging and interconnection of sensors and signal-processing chips, Sens. Actuators A **36** (1993) 1–27

37.31 R. Reichl: Packaging and interconnection of sensors, Sens. Actuators A **25–27** (1991) 63–71

37.32 A. Grisel, C. Francis, E. Verney, G. Mondin: Packaging technologies for integrated electrochemical sensors, Sens. Actuators **17** (1989) 285–295

37.33 J. L. Lund, K. D. Wise: Chip-level encapsulation of implantable CMOS microelectrode arrays, 1994 Solid-State Sensor and Actuator Workshop, Hilton Head (1994) 29–32

37.34 T. Akin, B. Siaie, K. Najafi: Modular micromachined high-density connector for implantable biomedical systems, IEEE Micro Electro Mechanical Systems Workshop (1996) 497–502

37.35 L. Muller, M. H. Hecht, L. M. Miller et al.: Packaging qualification for MEMS-based space systems, IEEE Micro Electro Mechanical Systems Workshop 503–508

37.36 L. Lin, K. McNair, R. T. Howe, A. P. Pisano: Vacuum encapsulated lateral microresoantors, 7th Int. Conference on Solid State Sensors and Actuators (June 1993) 270–273

37.37 H. Guckel: Surface micromachined pressure transducers, Sens. Actuators A **28** (1991) 133–146

37.38 J. J. Sniegowski, H. Guckel, R. T. Christenson: Performance characteristics of second generation polysilicon resonating beam force transducers, IEEE Solid-State Sensor and Actuator Workshop, Hilton Head Island (1990) 9–12

37.39 K. Ikeda, H. Kuwayama, T. Kobayashi, T. Watanabe, T. Nishikawa, T. Oshida, K. Harada: Three dimensional micromachining of silicon pressure sensor integrating resonant strain gauge on diaphragm, Sens. Actuators A **21–23** (1990) 1001–1010

37.40 C. H. Mastrangelo, R. S. Muller: Vacuum-sealed silicon micromachined incandescent light source, IEEE IEDM (1989) 503–506

37.41 J. Smith, S. Montague, J. Sniegowski, R. Manginell, P. McWhorter, R. Huber: Characterization of the embedded micromechanical device approach to the monolithic integration of MEMS with CMOS, SPIE 2879

37.42 S. Van der Groen, M. Rosmeulen, P. Jansen, K. Baert, L. Deferm: CMOS compatible wafer scale adhesive bonding for circuit transfer, 1997 International Conference on Solid-State Sensors and Actuators, Transducers'97 (1997) 629–632

37.43 M. B. Cohn, Y. Liang, R. Howe, A. P. Pisano: Wafer to wafer transfer of microstructures for vacuum package, Solid State Sensor and Actuator Workshop, Hilton Head (1996) 32–35

37.44 Y. T. Cheng, Y. T. Hsu, L. Lin, C. T. Nguyen, K. Najafi: Vacuum packaging using localized aluminum/silicon-to-glass bonding, Int. Conf. on MEMS (2001) 18–21

37.45 M. Chiao, L. Lin: Vacuum packaging of microresonators by rapid thermal processing, Proc. of SPIE on Smart Electronics, MEMS, and Nanotechnology (2002) 17–21

37.46 M. Esashi, S. Shoji, A. Nakano: Normally closed microvalve and micropump fabricated on a silicon wafer, Sens. Actuators **20** (1989) 163–169

37.47 M. E. Poplawski, R. W. Hower, R. B. Brown: A simple packaging process for chemical sensors, Solid-State Sensor and Actuator Workshop, Hilton Head (1994) 25–28

37.48 S. F. Trautweiler, O. Paul, J. Stahl, H. Baltes: Anodically bonded silicon membranes for sealed and flush mounted microsensors, Micro Electro Mechanical Systems Workshop (1996) 61–66

37.49 E. H. Klaassen, K. Petersen, J. M. Noworolski, J. Logan, N. I. Malfu, J. Brown, C. Storment, W. McCulley, G. T. A. Kovac: Silicon fusion bonding and deep reactive ion etching: A new technology for microstructures, Sens. Actuators A **52** (1996) 132–139

37.50 C. H. Hsu, M. A. Schmidt: Micromachined structures fabricated using a wafer-bonded sealed cavity process, Solid State Sensor and Actuator Workshop, Hilton Head (1994) 151–155

37.51 A. L. Tiensuu, J. A. Schweitz, S. Johansson: In situ investigation of precise high strength micro assembly using Au-Si eutectic bonding, Int. Conf. Solid-State Sens. Actuators, Eurosens. IX (1995) 236–239

37.52 A. P. Lee, D. R. Ciarlo, P. A. Krulevitch, S. Lehew, J. Trevino, M. A. Northrup: Practical microgripper by fine alignment, eutectic bonding and SMA actuation, Int. Conf. Solid-State Sens. Actuators, Eurosens. IX (1995) 368–371

37.53 R. W. Bower, M. S. Ismail, B. E. Roberds: Low temperature Si_3N_4 direct bonding, Appl. Phys. Lett. **62** (1993) 3485–3487

37.54 H. Takagi, R. Maeda, T. R. Chung, T. Suga: Low temperature direct bonding of silicon and silicon dioxide by the surface activation method, Int. Conf. on Solid-State Sensors and Actuators, Transducer 97 (1997) 657–660

37.55 G. Wallis, D. Pomerantz: Filed assisted glass-metal sealing, J. Appl. Phys. **40** (1969) 3946–3949

37.56 L. Bowman, J. Meindl: The packaging of implantable integrated sensors, IEEE Trans. Biomed. Engin. **33** (1986) 248–255

37.57 M. Esashi: Encapsulated micro mechanical sensors, Microsyst. Technol. **1** (1994) 2–9

37.58 B. Ziaie, J. Von Arx, M. Dokmeci, K. Najafi: A hermetic glass-silicon micropackages with high-density on-Chip feedthroughs for sensors and actuators, J. Microelectromech. Syst. **5** (1996) 166–179

37.59 Y. Lee, K. Wise: A batch-fabricated silicon capacitive pressure transducer with low temperature sensitivity, IEEE Trans. Electron Dev. **29** (1982) 42–48

37.60 S. Shoji, H. Kicuchi, H. Torigoe: Anodic bonding below 180 degree C for packaging and assembling of MEMS using lithium aluminosilicate-beta-quartz glass-ceramic, Proc. 10th Annual Int. Workshop Micro Electro Mechanical Systems (1997) 482–487

37.61 M. Esashi, N. Akira, S. Shoji, H. Hebiguchi: Low-temperature silicon-to-silicon anodic bonding with intermediate low melting point glass, Sens. Actuators A **23** (1990) 931–934

37.62 A. Hanneborg, M. Nese, H. Jakobsen, R. Holm: Silicon-to-thin film anodic bonding, J. Micromech. Microeng. **2** (1992) 117–121

37.63 S. A. Audet, K. M. Edenfeld: Integrated sensor wafer-level packaging, International Conference on Solid-State Sensors and Actuators, Transducers'97 (1997) 287–289

37.64 R. C. Benson, N. deHaas, P. Goodwin, T. E. Phillips: Epoxy adhesives in microelectronic hybrid applications, Johns Hopkins APL Tech. Dig. **13** (1992) 400–406

37.65 M. Shimbo, J. Yoshikawa: New silicon bonding method, J. Electrochem. Soc. **143** (1996) 2371–2377

37.66 P. M. Zavracky, B. Vu: Patterned eutectic bonding with Al/Ge thin film for MEMS, SPIE **2639** (1995) 46–52

37.67 G. Humpston, D. M. Jacobson: Principles of soldering and brazing, ASM Int. (1993) 241–244

37.68 A. Singh, D. Horsely, M. B. Cohn, R. Howe: Batch transfer of microstructures using flip-chip solder bump bonding, International Conference on Solid State Sensors and Actuators, Transducer 97 (1997) 265–268

37.69 M. M. Maharbiz, M. B. Cohn, R. T. Howe, R. Horowitz, A. P. Pisano: Batch micropackaging by compression-bonded wafer-wafer transfer, 12th Int. Conf. MEMS (1999) 482–489

37.70 Y. T. Cheng, L. Lin, K. Najafi: Localized silicon fusion and eutectic bonding for MEMS fabrication and packaging, IEEE/ASME J. Microelectromech. Syst. **9** (2000) 3–8

37.71 L. Lin: Selective encapsulations of MEMS: micro channels, needles, resonators, and electromechanical filters. Ph.D. Thesis (University of California, Berkeley 1993)

37.72 Y. C. Su, L. Lin: Localized plastic bonding for micro assembly, packaging and liquid encapsulation, Proceedings of IEEE Micro Electro Mechanical Systems Conference, Interlaken (2001) 50–53

37.73 M. Schwartz: *Brazing* (Chapman, London 1995)

37.74 L. Lin, R. T. Howe, A. P. Pisano: Microelectromechanical Filters for Signal Processing, IEEE/ASME J. Microelectromech. Syst. **7** (1998) 286–294

37.75 W. C. Tang, C. T.-C. Nguyen, R. T. Howe: Laterally driven polysilicon resonant microstructures, Sens. Actuators A **20** (1989) 25–32

37.76 G. T. Mulhern, D. S. Soane, R. T. Howe: Supercritical carbon dioxide drying of microstructures, 7th International Conference on Solid State Sensors and Actuators, Yokohama (1993) 296–299

37.77 M. Judy: Micromechanisms using sidewall beams. Ph.D. Thesis (University of California, Berkeley 1994)

37.78 A. M. Leung, J. Jones, E. Czyzewska, J. Chen, B. Woods: Micromachined accelerometer based on convection heat transfer, International Conference on MEMS (1998) 627–630

37.79 D. R. Spark, L. Jordan, J. H. Frazee: Flexible vacuum-packaging method for resonating micromachines, Sens. Actuators A **55** (1996) 179–183

37.80 SensArray Corporation, 47451 Fremont Blvd. Fremont, CA 94538

37.81 Y. T. Cheng, W. T. Hsu, K. Najafi, C. T. Nguyen, L. Lin: Vacuum packaging technology using localized aluminum/silicon-to-glass bonding, IEEE/ASME J. Microelectromech. Syst. **11** (2002) 556–565

37.82 C. Luo, L. Lin: The application of nanosecond-pulsed laser welding technology in MEMS packaging with a shadow mask, Sens. Actuators A **97–98** (2002) 398–404

37.83 A. Cao, M. Chiao, Liwei. Lin: Selective and localized wafer bonding using induction heating, Technical Digest of Solid-State Sensors and Actuators Workshop (2002) 153–156

37.84 J. B. Kim, M. Chiao, L. Lin: Ultrasonic Bonding of In/Au and Al/Al for Hermetic Sealing of MEMS Packaging, Proceedings of IEEE Micro Electro Mechanical Systems Conference (2002) 415–418

37.85 Y. T. Cheng, L. Lin, K. Najafi: Localized bonding with PSG or indium solder as intermediate layer, 12th International Conference on MEMS (1999) 285–289

37.86 G. Thyrum, E. Cruse: Heat pipe simulation, a simplified technique for modeling heat pipe assisted heat sinks, Adv. Packaging (2001) 23–27

37.87 G. P. Peterson, A. B. Duncan, M. H. Weichold: Experimental investigation of micro heat pipes fabricated in silicon wafers, J. Heat Trans. (1993) 751–756

37.88 L. Jiang, M. Wong, Y. Zohar: Forced convection boiling in a microchannel heat sink, J. Microelectromech. Syst. **10** (2001) 80–87

37.89 F. Arias, S. R. J. Oliver, B. Xu, E. Holmlin, G. M. Whitesides: Fabrication of metallic exchangers using sacrificial polymer mandrils, IEEE/ASME J. Microelectromech. Syst. **10** (2001) 107–112

37.90 M. Chiao, L. Lin: Hermetic wafer bonding based on rapid thermal processing, Sens. Actuators A **91** (2001) 404–408

37.91 *ANSYS Modeling and Meshing Guide*, 3rd edn. (SAS IP, Inc., Canousburg 2002)

37.92 D. Bystrom, L. Lin: Residual stress analysis of silicon-aluminum-glass bonding processes, ASME International Mechanical Engineering Congress and Exposition, MEMS Symposium (November 2002)

37.93 H. Scholze: *Glass: Nature, Structure, and Properties*, 1st edn. (Springer, Berlin, Heidelberg 1991)

37.94 R. R. Tummala: *Fundamentals of Microsystems packaging* (McGraw-Hill, New York 2001)

37.95 S. J. Adamson: BGA, CSP, and flip chip, Adv. Packaging (2002) 21–24

37.96 D. Steoehle: On the penetration of water vapor into packages with cavities and on maximum allowable leak rate, 15th Annual Proceeding Reliability Physics Symposium (1977) 101–106

37.97 W. D. Brown: *Advanced Electronic Packaging* (IEEE Press, 1999)

37.98 F. Niklaus, P. Znoksson, E. Käluesten, G. Stemme: Void free full wafer adhesion bonding, Proceedings of IEEE Micro Electro Mechanical Systems Conf. 2000) 241–252

37.99 C. H. Cheng, A. S. Ergun, B. T. Khuri-Yakub: Electrical through-wafer interconnects with sub-picofarad parasitic capacitance, IEEE ECTC (2002) 18–21

37.100 D. Routkevitch, A. A. Tager, J. Haruyama, D. Almawlawi, M. Moskovits, J. M. Xu: Nonlithographic nano-array arrays: fabrication, physics, and device applications, IEEE Trans. Electron. Dev. **43** (1996) 1646–1658

37.101 J. Gou, M. Lundstrom, S. Datta: Performance projections for ballistic carbon nanotube fieldeffect transistors, Appl. Phys. Lett. **80** (2002) 3192–3194

37.102 S. U. S. Choi, Z. G. Zhang, W. Yu, F. E. Lockwood, E. A. Grulke: Anomalous thermal conductivity enhancement in nanotube suspensions, Appl. Phys. Lett. **79** (2001) 2252–2254

37.103 K. Velikov, A. Moroz, A. Blaaderen: Photonic crystals of core-shell colloidal particles, Appl. Phys. Lett. **80** (2002) 49–51

Part F Social and

Part F Social and Ethical Implication

38 Social and Ethical Implications of Nanotechnology
William S. Bainbridge, Arlington, USA

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

38. Social and Ethical Implications of Nanotechnology

Nanotechnology will have very broad applications across all fields of engineering, so it will be an amplifier of the social effects of other technologies. There is an especially great potential for it to combine with three other powerful trends – biotechnology, information technology, and cognitive science – based on the material unity of nature at the nanoscale and on technology integration from that scale. Technological convergence highlights such existing issues as the treatment of the disabled, communication breakdowns, economic stagnation, and threats to national security. Nanotechnology itself may possibly raise distinctive ethical and social issues in the future, but much of the public discussion to this point has been misdirected and misinformed, lacking a firm social scientific basis. Thus it will be important to integrate social and ethical studies into nanotechnology developments from their very beginning.

Social scientific and economic research can help manufacturers and governments make the right decisions when deploying a new technology, maximizing its benefit for human beings. In addition, technically competent research on the societal implications of nanotechnology will help give policymakers and the general public a realistic picture free of unreasonable hopes or fears. The costs of premature or excessive regulation would be extremely high, harming the very people it was intended to protect, and failure to develop beneficial nanotechnology applications would be unethical.

The significance of nanotechnology depends largely on how its development relates to wider trends going on in the world such as the impending population declines of most advanced industrial nations, the apparent diminishing returns to increased medical research and health care investment, and the threatened deceleration of progress in microelectronics. Well established social-scientific explanations for unethical behavior – such as learning, strain, control, and subculture theories – could help us understand possible future cases in nanotechnology industries. Ethics and social implications are largely matters of social perception, and the public conception of nanotechnology is still in the early stages of developing. Social science can now begin to examine its unfolding impacts in all sectors of the economy, in most spheres of life, and both short-term and long-term time scales.

38.1 **Applications and Societal Impacts** 1136
38.2 **Technological Convergence** 1139
38.3 **Major Socio-technical Trends** 1141
38.4 **Sources of Ethical Behavior** 1143
38.5 **Public Opinion** 1145
38.6 **A Research Agenda** 1148
References 1149

The views expressed in this essay do not necessarily represent the views of the National Science Foundation or the United States.

Science and engineering have only begun to explore the potential for discovery and creativity at the nanoscale, but already some intemperate voices call for government regulation or outright banning of nanotechnology [38.1–4]. Such action would be extremely premature, because we have just started on a very long road of research that we must traverse before we will know what is technically practical, what real-world applications nanotechnology might actually have, and what the appropriate societal responses would be to any hazards that might be directly or indirectly related to these applications. Some writers in the popular press seem to treat nanotechnology as a new kind of Frankenstein's Monster, but it is worth remembering that biological science is still not able to create such a monster, nearly two centuries after Mary Shelly imagined one.

A good starting point for consideration of social and ethical issues is a formal definition of the topic, but unfortunately a somewhat different definition has lodged in the public mind from that employed by professionals working in the field. For example, *Merriam-Webster's Collegiate Dictionary* defines *nanotechnology* as: "the art of manipulating materials on an atomic or molecular scale especially to build microscopic devices (as robots)." Similarly, the online reference About.com defines it thus: "The development and use of devices that have a size of only a few nanometres." Later, we will examine public conceptions more closely, but it is worth noting here that these definitions focus on robot-like devices constructed at the nanoscale and ignore the topics of most chapters in this volume.

A more authoritative if long-winded definition of *nanotechnology* has been provided by the U.S. Government's National Nanotechnology Initiative: "Research and technology development at the atomic, molecular or macromolecular levels, in the length scale of approximately 1–100 nm range, to provide a fundamental understanding of phenomena and materials at the nanoscale and to create and use structures, devices and systems that have novel properties and functions because of their small and/or intermediate size. The novel and differentiating properties and functions are developed at a critical length scale of matter typically under 100 nm. Nanotechnology research and development includes manipulation under control of the nanoscale structures and their integration into larger material components, systems and architectures."

While this definition includes the word "devices," it also covers materials, structures, and systems. Significantly, it envisions that many nanotechnology advances will be integrated into "larger material components, systems and architectures," which means that we would need to understand the social and ethical implications of the larger-scale systems of which nanoscale structures are only components.

There is good reason to believe that engineering at the nanoscale will have very broad applications across all fields of technology and most fields of science [38.5–7]. Therefore, it will be an enabler or amplifier of the effects of other technologies. This means that it is foolish to ask, "Is nanotechnology harmful?" Nanotechnology is not one thing, either ethical or unethical. Rather, it is a myriad of different things – many nanotechnologies – each of which has a range of potential uses and misuses in conjunction with other technologies and applied to different goals.

This line of reasoning could lead some people to conclude that we did not need to worry at all about the social and ethical implications of nanotechnology, because they will be subsumed by the implications of larger systems. Others might conclude that it is far too early in the history of nanoscale engineering to consider societal implications, because we cannot study the impact of a technology that does not yet exist.

However, the very fact that some writers are urging regulation or banning of nanotechnology suggests that the debate about its value to humanity has begun, and we need to bring the best possible thinking and information to that debate [38.8–10]. In addition, a few relatively simple but significant applications have begun to appear, so in those cases at least an impact can be observed. Social scientists, natural scientists, philosophers, and engineers have indeed begun considering the future of nanotechnology, so we can draw upon the very first serious intellectual efforts in this important field.

38.1 Applications and Societal Impacts

In the year 2000, the National Science and Technology Council sponsored a major workshop at the National Science Foundation, which led to a published report, *Societal Implications of Nanoscience and Nanotechnology* [38.11]. Sixty-four representatives of academic science, government laboratories, and corporate research listed a large number of potential nanotechnology applications, a few of which had already entered production, but many of which could be expected only after decades of research. Following are diverse examples of possible applications suggested by workshop participants concerning a variety of sectors of the economy and spheres of life:

- More efficient components for the semiconductor industry such as integrated circuits containing transistors constructed from carbon nanotubes.

- Nanostructured catalysts for the chemical industry and for more effective converters to handle pollution from automobile exhausts and other combustion.
- Lighter and stronger nanomaterials in bulk quantities to enable safer and more efficient transportation vehicles, including automobiles, aircraft, and train systems.
- Improved pharmaceuticals with features such as programmed delivery to desired targets like tumors and the ability to employ substances that are not soluble in water as medications.
- Cost-effective and reliable filters for water decontamination, desalination, containment of industrial pollutants, and air purification.
- More efficient solar energy conversion, thereby reducing current reliance on oil and offering an alternative to nuclear power for future electricity needs.
- Efficient fuel cells and hydrogen storage systems, leading to nonpolluting automobiles, trucks, and buses.
- More durable composite materials, such as nanoparticle-reinforced polymers, designed for optimal performance in specific uses with reduced waste and greatly increased lifetime.
- Molecularly engineered biodegradable fertilizers and insecticides designed for precision agriculture with efficient delivery to where they are needed and prevention of unwanted side effects.
- Revolutionary launch vehicles with high mass-ratios, nonpolluting but powerful engines, and low electric power needs in order to finally realize the promise of space exploration.
- Nanoscale components in sensor systems that can quickly detect and identify pollutants and disease organisms, as well as chemical or biological warfare agents to allow quick and appropriate medical treatment or security responses.
- Coatings that would give ordinary materials extraordinary properties like self-cleaning window glass and heat-reflecting exterior architectural surfaces.

Knowledgeable scientists and engineers reported that these and many other similar applications were plausible benefits of nanotechnology research, but they recognized that progress has costs. For example, the creation of a new industry often results in the decline or complete destruction of an older industry, as in the proverbial case of buggy-whip manufacturers who were rendered obsolete by the automobile. The potential benefit from a new technology must be calculated in a sophisticated manner, taking into account the full life-cycle costs [38.12]. For example, an electric car may reduce some forms of pollution while increasing others such as the release of lead or other dangerous substances into the environment during the manufacture of the electric battery and the disposal of it when the car is eventually junked.

Many ethical issues of fairness will arise within nanotechnology industries themselves. For example, it will be important to invest in appropriate safeguards for workers engaged in hazardous production processes. In general, scientific discoveries cannot be patented, whereas new products and production methods can, so there will be serious issues of intellectual property protection in nanoscience, a nontraditional field where the legal line between discovery and invention has not yet been clearly drawn. Often, it will be necessary to bring together information and expertise from several different sources in order to create an economically profitable application, drawing upon government-funded university research projects, as well as several specialized industrial corporations, which will require improved models of collaboration across organizations.

As the *Societal Implications* report observed, the general public has a significant stake in the National Nanotechnology Initiative (NNI) that must be managed both directly through public participation and indirectly through the involvement of social scientists:

> It is important to include a wide range of interests, values, and perspectives in the overall decision process that charts the future development of nanotechnology. Involvement of members of the public or their representatives has the added benefit of respecting their interests and enlisting their support. ... The inclusion of social scientists and humanistic scholars, such as philosophers of ethics, in the social process of setting visions for nanotechnology is an important step for the NNI. As scientists or dedicated scholars in their own right, they can respect the professional integrity of nanoscientists and nanotechnologists, while contributing a fresh perspective. Given appropriate support, they could inform themselves deeply enough about a particular nanotechnology to have a well-grounded evaluation. At the same time, they are professionally trained representatives of the public interest and capable of functioning as communicators between nanotechnologists and the public or government officials. Their input may help maximize the societal benefits of the technology while reducing the possibility of debilitating public controversies [38.11].

Social scientific and economic research can help manufacturers and the government make the right decisions when deploying a new technology, maximizing its benefit for human beings. In addition, technically competent research on the societal implications of nanotechnology will help give policymakers and the general public a realistic picture free from unreasonable hopes or fears. As the report notes, the National Nanotechnology Initiative is just now commencing, so “there is a rare opportunity to integrate the societal studies and dialogues from the very beginning and to include societal studies as a core part of the NNI investment strategy” [38.11].

The workshop developed a number of recommendations for government, academia, and industry suggesting how the social and ethical implications of nanotechnology could best be addressed [38.11]. Here are some of the most important principles in education, social science, and ethics:

- Nanoscale concepts should be introduced into science and engineering education at all levels, thereby giving the widest possible range of students a fundamental understanding of the field while intellectually linking nanotechnology to many other fields.
- The training of nanotechnologists should include societal implications and ethical sensitivity, so their future work will be guided by principles that will maximize human benefit.
- A sufficient number and variety of social and economic scientists should receive effective multidisciplinary training, so they will be well prepared to work in the nanotechnology area.
- Formal measurement methods, such as social and economic indicators, should be developed and consistently employed to chart the actual widespread changes in industry, education, and public welfare as they occur.
- Government agencies, private foundations, and industry should support a wide range of high-quality, theory-based social and economic research studies on nanotechnology, examining both the decision processes that shape the emerging technology and itsspecific societal impacts once it is deployed.
- A knowledge base and institutional infrastructure must be created to evaluate the probable future intellectual and societal impacts of nanoscience and nanotechnology over the short, medium, and long term.
- Government and the private sector should establish effective channels for informing the public promptly about new concepts, projects, potential applications, and ethical issues as they emerge.
- There should be formal organizations and mechanisms to ensure participation by diverse societal institutions and the general public in setting priorities for research and development and in providing timely input to decision makers.
- Government and industry should develop management plans and policies that can effectively incorporate all relevant information and legitimate stakeholder interests to ensure that we can respond flexibly to social, ethical, legal, and economic implications as they appear.

It will take time and effort to implement these recommendations, but some progress has already been made. The National Science Foundation has recently funded Rosalyn Berne at the University of Virginia to study the developing ethics of nanotechnology through a narrative analysis of discourse from engineers and scientists in the field, and a team headed by Michael Gorman at the same institution was funded to explore the potential for multidisciplinary research on social and ethical dimensions of nanotechnology. The Environmental Protection Agency, which has supported several research projects on how nanotechnology might remedy conventional forms of pollution, has funded Darrell Velegol and Kristen Fichthorn at Pennsylvania State University to study “Green Engineering of Dispersed Nanoparticles,” exploring how valuable nanoparticles might be produced without the use of polluting additives. In its review of the National Nanotechnology Initiative, the National Research Council has recommended the development of “a new funding strategy to ensure that the societal implications of nanoscale science and technology become an integral and vital component of the NNI” [38.13].

38.2 Technological Convergence

In the year following the *Societal Implications* workshop, NSF and the U.S. Department of Commerce staged an even larger meeting to examine the potential for *nanotechnology* to combine with three other powerful scientific and engineering trends: *biotechnology*, *information technology*, and *cognitive science* – four "NBIC" fields. The resulting report, *Converging Technologies for Improving Human Performance* [38.14], examined benefits of technological convergence that would specifically enhance the ability of humans to satisfy their needs and achieve their legitimate goals.

Convergence of diverse NBIC technologies will be based on the material unity of nature at the nanoscale and on technology integration from that scale. This is the scale at which much future biotechnology and the hardware for information technology will operate, and it is the scale from which fundamental research on the human nervous system will take place. Key transforming tools, including research instruments and production methods, will arise in the conjunction of previously separated fields of science. Drawing upon advances in mathematics and computation, it will be possible to model phenomena from the nanoscale to the cosmic scale as complex, coupled, hierarchical systems.

The *Converging Technologies* scientists, engineers, and policymakers sketched six general areas in which the human benefits and social issues are likely to result:

- Overall potential of converging technologies
- Expanding human cognition and communication
- Improving human health and physical capabilities
- Enhancing group and societal outcomes
- National security
- Unifying science and education

Numerous valuable applications were sketched in greater or lesser detail that could reasonably be expected to result from a decade or two of vigorous research and development, including the following ideas that depend on partnerships between nanotechnology and one or more of the other NBIC fields:

- Comfortable, wearable sensors and computers will enhance every person's awareness of his or her health condition, the environment, chemical pollutants, potential hazards, and information of interest about local businesses, natural resources, and the like.
- The human body will be more durable, healthier, more energetic, easier to repair, and more resistant to many kinds of stress, biological threats, and aging processes.
- Machines and structures of all kinds, from homes to aircraft, will be constructed of materials that have exactly the desired properties, including the ability to adapt to changing situations, high energy efficiency, and environmental friendliness.
- A combination of technologies and treatments will compensate for many physical and mental disabilities and will eradicate altogether some handicaps that have plagued the lives of millions of people.
- National security will be greatly strengthened by lightweight, information-rich war fighting systems, capable uninhabited combat vehicles, adaptable smart materials, invulnerable data networks, superior intelligence-gathering systems, and effective measures against biological, chemical, radiological, and nuclear attacks.
- The ability to control the genetics of humans, animals, and agricultural plants will greatly benefit human welfare; widespread consensus about ethical, legal, and moral issues will be built in the process.
- Agriculture and the food industry will greatly increase yields and reduce spoilage through networks of cheap, smart sensors that constantly monitor the condition and needs of plants, animals, and farm products.
- Transportation will be safe, cheap, and fast due to ubiquitous real-time information systems, extremely high-efficiency vehicle designs, and the use of synthetic materials and machines fabricated from the nanoscale for optimum performance.
- The work of scientists will be revolutionized by importing approaches pioneered in other sciences, for example, genetic research employing principles from natural language processing and cultural research employing principles from genetics.
- Formal education will be transformed by a unified but diverse curriculum based on a comprehensive, hierarchical intellectual paradigm for understanding the architecture of the physical world from the nanoscale through the cosmic scale.

Several of these application areas depend upon the development of nanoscale components for sensor, computation, and communication networks. Others involve a synergy between inorganic nanoscale engineering and organic chemistry or biology. Others extrapolate that it will be possible to build large-scale structures that are engineered at the nanoscale to enhance the qualities of architecture, transportation vehicles, or even the human body. Fundamental to all of the application areas is a conceptual revolution that unifies science and provides humanity with a comprehensive view of nature and technology.

Contributors to the report showed how nanoscale science and engineering were essential for progress in biology, information science, and cognitive science. The result of NBIC convergence will be the strengthening of science and the empowerment of humanity, as we will find solutions to many of humanity's problems in health, natural resources, environment, security, and even communication with fellow human beings. Often, the technologies based in the sister sciences will raise ethical issues. In such cases, nanotechnology may become implicated in controversies without being central to them. The correct policy response would then be to ignore the superficial nanotechnology aspects and focus directly on the core problem that lies in a different field. *Converging Technologies* contributors examined how technological convergence highlighted such existing issues as the treatment of the disabled, communication breakdowns, economic stagnation, and threats to national security.

For greater clarity, we can consider here the very specific example of the proliferation of nuclear weapons. Nuclear physics and atomic technology were specifically not included in the constellation of NBIC converging technologies considered by the workshop, so the following discussion supplements its findings. As scientists already understood in the early 1940s, there are fundamentally two ways that a rogue nation can build up its own supply of weapons-grade nuclear materials. One is breeding plutonium in a nuclear reactor fueled by uranium, transmuting the common uranium isotope of atomic weight 238. But this is hard to conceal from other nations, especially if reactors are open to international inspection. The other way is separation of the relatively rare uranium isotope ^{235}U, which, like plutonium, is suitable for use in weapons. The original separation method was gas diffusion, but then the centrifuge method was developed, and other approaches are possible such as laser or electromagnetic separation.

In the gas diffusion method, a chemical containing uranium (uranium hexaflouride) is passed through a porous barrier such as a ceramic wall that has tiny holes in it. Molecules containing ^{235}U travel through the barrier slightly faster than molecules containing ^{238}U. This is a very inefficient process, because the isotopes are very similar to each other, so many passes through a barrier are required to enrich the proportion of ^{235}U sufficiently for use in a nuclear weapon. The other approach uses centrifuges, rapidly spinning rotors containing the uranium hexaflouride to separate the two isotopes on the basis of their slight difference in mass. Both approaches are very difficult and costly, however. Conceivably, nanotechnology could slightly reduce the difficulty, by producing porous barriers with exactly the optimum sized holes and by providing lighter and stronger centrifuge rotors that could operate at higher speeds.

Therefore, without a more careful analysis someone might leap to the conclusion that nanotechnology is a danger because it facilitates proliferation of nuclear weapons. This ignores several crucial facts. First, the world already faces a very severe crisis of nuclear proliferation without nanotechnology. Second, the fundamental requirement for production of nuclear weapons is the availability of uranium, so the most logical preventive would be international monitoring and control over its mining and distribution. Third, there are many other technical requirements for production and delivery of nuclear weapons, and it is the whole collection of them that presents a threat, not any one alone. Fourth, nanotechnology is inherently nonnuclear, because nuclear reactions occur at a smaller scale than the nanoscale, so its connections to proliferation of nuclear weapons will always be highly indirect if they exist at all. Finally, the fundamental problem is not nuclear proliferation, but weapons of mass destruction in general, and nanotechnology has a major role to play in the defense against chemical and biological warfare agents in sensor systems that can detect them before they have caused harm.

Perhaps a very few nano products will need to be "controlled substances," analogous to narcotics, insecticides, asbestos, and antibiotics [38.15]. Conceivably, porous barriers with exactly the right pore size for uranium isotope separation would be controlled in some way. Some products could be produced and used under strict regulations, whereas a few others might be banned altogether. Most would not require restrictions of any kind.

Such decisions would need to be based on the careful analysis of facts related to the specific issue at hand. For example, it may be that no improvements in gas diffusion separation are possible through nanotechnology, or they are so slight as to be negligible. Perhaps technological developments in laser or electromagnetic separation, completely unrelated to nanotechnology, may render both the gas diffusion and centrifuge methods obsolete. Therefore, the mere fact that we can imagine a way in which nanotechnology might possibly encourage the proliferation of nuclear weapons is far too weak a justification for international regulation. We would need to go beyond vague fears to precise analysis based on rigorous scientific research to distinguish real dangers from imaginary ones and design appropriate safeguards in those rare cases when they are necessary.

The costs of excessive regulation would be extremely high, harming the very human beings it was intended to protect. Porous barriers precisely designed at the nanoscale are among the high benefit applications we can anticipate early in the history of nanotechnology. They would have a wide range of applications, from water purification and desalination to separation of valuable substances in biotechnology production processes. Similarly, new high-performance materials with great strength but low weight will have a myriad of applications throughout the civilian sectors of the economy. Among the products that would benefit millions of ordinary people are light but indestructible eyeglass frames, safety helmets, and hand tools.

Nanotechnology has the potential to increase the effectiveness of biotechnology, information technology, and technologies based on cognitive science through NBIC convergence. Therefore, it could become entangled in ethical issues that already involve those other technologies such as the conflict over genetically modified foods, the dispute over information privacy, and the debate over treatment of mental illness. These are not fundamentally nanotechnology issues, however, and it will be important to keep a proper perspective on their real nature.

38.3 Major Socio-technical Trends

The significance of nanotechnology depends not only on its own accomplishments and on convergence with other technologies, but also on how these purely technical developments relate to wider trends going on in the world [38.16, 17]. At present, many of these trends are confused and are the subject of great controversy in the social science and public policy arenas. It is always hazardous to extrapolate trends from any particular moment in history, but the present time appears so chaotic that the events of a single day might send humanity down a very different road.

A few years ago, there was a broad consensus that the rapid growth in the number of people on the planet, the so-called *population explosion*, was the dominant quantitative trend that needed to be factored into any projections [38.18]. But by the mid-1980s, it became obvious that population growth had been replaced by the threat of population decline in most advanced industrial nations, with the possible exception of the United States [38.19, 20]. Even before collapsing fertility begins to reduce the number of people in the affected nations, their average age and pension costs increase, social dynamism and creativity decrease, and governments become gridlocked trying to deal with ultimately insoluble problems of funding demanded services. This is one of the explicit reasons why the United States has begun to discount the influence of its European allies and Japan [38.21, 22].

Although nanotechnology has little obvious direct connection with fertility, there are two potentially counterbalancing, indirect connections. Rapid progress in nanotechnology could revitalize economic expansion and technical creativity, thereby to some extent offsetting the negative consequences of fertility collapse. However, the aversion to change that marks societies having a high proportion of elderly citizens may dampen the intellectual inquisitiveness and investment risk-taking required to develop nanotechnology and technological convergence.

A related trend is the constant improvement in health resulting from progress in medical science and the economic growth needed to fund increasingly costly health care. Continued progress in health is by no means assured, however. During the nineteenth century, American health may actually have declined significantly for a number of decades, despite economic growth [38.23]. We tend to attribute the undeniable improvement in health over the twentieth century to medical progress, but health education and public sanitation may have been more important [38.24]. The U.S. Centers for Disease Control argue that substantial improvements in health and longevity could be achieved by lifestyle changes, notably, exercising more, reducing the fats in our diet, and avoiding smoking [38.25, 26]. The introduction of antibiotics helped increase longevity, and modern cardiology saves the lives of many people who might otherwise die prematurely of heart attacks, but the progress against cancer, AIDS, and most aging-related illnesses has been agonizingly slow. On balance, economists find that the increasing investments in heath care are paying off, but not necessarily in all areas, or with very great benefit [38.27–29].

There are several reasons to be pessimistic about the future benefits of medicine, unless there are very major scientific breakthroughs. We have yet to find a cure for any chronic viral infections, of which AIDS is merely the most publicized example. Bacteria are rapidly evolving resistance to antibiotics, yet the rate of development of new prescription medicines is declining.

The average white American male born in 1900 could expect to live 48 years, but in 2000 this life expectancy had increased to 74 years. For white females, the average life expectancy increased from 51 to 80. Projecting these figures forward at the same rate of increase suggests that life expectancy in 2100 might be 114 for males and 125 for females. However, using more realistic assumptions, the U.S. Census Bureau has projected that life expectancy for Americans born in the year 2100 might be only around 88 for males and 92 for females [38.30]. This does represent progress, but at a steeply declining rate, and it depends upon continued economic growth.

Perhaps the most economically significant technical trend of recent years has been *Moore's Law*, Gordon Moore's observation that the density of transistors on the most advanced microchip doubles about every 18 months. The rapid improvement of computer and communications hardware has been fundamental to the implementation of new software applications, networking, and information technology in general. These, in turn, have been responsible for much economic growth. However, the semiconductor industry is approaching physical limits in the traditional methods of making integrated circuits, including computer chips, so Moore's Law will stall without breakthroughs in nanoscale technologies such as molecular logic gates and carbon nanotube transistors [38.31–33].

Arguably, several of the great thrusts of twentieth-century technology have already stalled. Except for the improved capacity of scientific instruments to collect and process astronomical data, the space program has hardly advanced since the mid-1970s. Civilian aviation and ground transportation have improved only in relatively minor design details since the 1950s. Despite great efforts, little progress has been made in controlling nuclear fusion for power generation. Nuclear fission power generation has largely been halted by public controversies, and no technological breakthroughs in safe operation of reactors or disposal of radioactive wastes have assuaged the public's concerns. The social sciences seem no better able to define good public policies than they were decades ago. Outside of nanotechnology itself, the two main areas where rapid progress obviously continues are information technology and genetic engineering. Convergence with nanotechnology is essential for further progress in both of these areas.

What would happen to the world economy if technological progress slowed or even halted? First, the advanced industrial nations would lose the advantage they currently hold over developing countries. For example, the American semiconductor industry would probably collapse if it lost its technological superiority, because nations with lower wage rates could produce the same things more cheaply. Technological products, from computer chips to pharmaceuticals, would become commodities, manufactured wherever it was cheapest to do so. With their heavy pension burdens and costly social services, governments of advanced nations would face fiscal crises of unprecedented magnitude. Indeed, the economies of these nations could fall rapidly down to the average of all nations, unleashing social discontent that could lead to unpredictable violence.

At the same time, some developing nations would advance economically, which would mean increased net world industrial production with concomitant increase in environmental pollution and resource depletion. Without technological progress to solve these problems, the average human welfare could decline, rather than remaining stable. In order to improve human welfare, technological progress must continue. Perhaps the best way to achieve that is through nanoscience and nanotechnology in convergence with the other NBIC fields, gaining understanding and control at the physical scale that is the basis for most science and technology.

It is impossible to predict how rapidly the conditions in the richest nations would deteriorate, or whether humanity could navigate through these crises to achieve a stable world. But clearly a halt in technological progress would be a shock to the world economy, global political institutions, and human welfare – which it would be wise to avoid.

38.4 Sources of Ethical Behavior

Ethical questions related to nanotechnology are not limited to the ways people might use it to harm others intentionally, but also include obligations to avoid potentially harmful unintended consequences [38.34]. An example of the former might be a weapon and of the latter, a kind of environmental pollution. In either case, the harm might be morally justified, as when a nation employs a weapon to defend against attack, or when limited pollution of the environment is offset by substantial benefits to humanity. Other ethical questions concern harm to the owners of nanotechnology, for example, if a competing company violates a nanotech patent, or if a government bans a nano-related product without careful examination of scientific evidence about its value. News reporters or popular writers who spread false information about nanotechnology may also be acting unethically, whether they are willfully lying or merely failing to be professionally diligent in checking their facts.

It is important to recognize that there are sins of omission as well as commission. To fail to develop a beneficial nanotechnology application could also be unethical. Imagine, for instance, that a laboratory had discovered how nanotechnology could enable a much more effective form of chemotherapy to cure cancer reliably and cheaply, but the company owning the patents prevented it from being developed, because its less effective traditional chemotherapy products were more profitable. This could be considered just as unethical as doing aggressive harm.

For over a century, social scientists have been studying the origins of ethical behavior – or its opposite, unethical behavior – in fields like the social psychology of groups, the sociology of deviance, and criminology. Although it is impossible to predict when a particular person will commit a specific ethical or unethical act, we can identify with good confidence the general factors that steer human behavior in one direction or another. An individual's behavior is determined by a complex of factors, few of which can be measured accurately, and chance seems to play a significant role in human affairs. Thus, the findings of social science tend to be statistical or conditional in nature. This does not render them useless, because one can develop rational laws, government policies, and investment strategies based on probabilities rather than certainties.

The theoretical and empirical scientific literature on the topic is vast, including literally thousands of publications dating back to the 1840s, expressing the views of countless schools of thought, often employing idiosyncratic terminology. For present purposes, we must distill that huge and incoherent intellectual heritage down to its most meaningful essentials. A good start is the following set of four themes, each of which can be expressed as a theory about why people might violate an ethical rule:

1. Learning – A person violates a rule because the person has learned it is rewarding to do so.
2. Strain – A person violates a rule because the person is frustrated in the attempt to fulfill one of the expectations of society.
3. Control – A person violates a rule because the person lacks stable social bonds that might prevent it.
4. Subculture – A person violates a rule because the person belongs to a group or network of other people who encourage violation.

When each of these ideas first appeared in the social-scientific literature, the theory's proponents tended to present it as the complete explanation for deviant behavior, but now we understand that each of these four has merit and the correct explanation is usually a mixture of two or more. We will illustrate each of the theories with hypothetical illustrations, because actual examples may not exist yet.

Learning Theory is sometimes interpreted in economic or "rational choice" terms, as meaning that a person will do whatever he or she perceives it is profitable to do. But human life is not spent entirely in a marketplace where one can compare prices and shop for a good deal. Rather, there are many quite different contexts in which people make decisions. Regardless of whether options can be quantified in dollars, people tend to rely upon their own past experiences in selecting between familiar courses of action [38.35–37].

If the managers of an industrial corporation have learned they can achieve higher profits by ignoring regulations against polluting the environment, and they have not suffered heavy fines from the government for doing so, they will be more likely to pollute in the future. Therefore, a corporation guilty of polluting with conventional technologies is apt to continue with this history of pollution when it adopts new nanotechnologies. This suggests that one way to reduce the likelihood that industrial companies will pollute when they adopt new nanotechnologies is to enact and enforce strong regulations against pollution in general, to make them learn to avoid polluting.

In Strain Theory, a person violates a norm because he or she is frustrated in the attempt to achieve one of the values endorsed by society [38.38, 39]. This idea has been used to explain why many members of poor communities or groups that have been subject to discrimination turn to crime to gain wealth and status, but it can also explain why a major corporation might begin to engage in dubious accounting practices when its growth in real business stalls. The central idea is that an individual or group becomes committed to a particular kind of success that society encourages – whether it is wealth, political power, widespread fame, or honor within a particular sector of society – then experiences frustration in attaining success, which motivates ethical violations.

Conceivably, a nanotechnology corporation could be affected by strain if it has committed itself to a new and untested line of products that turns out to be less effective, more harmful, or more costly than originally estimated. The corporation has committed not only its money and effort, but also its public prestige to success, but success may elude it. It may then unethically advertise its products as more effective than they really are, conceal knowledge of their danger to human health or the environment, and cook its financial records to deceive investors. Of course, many ordinary companies have done this in the past, and, ironically, they sometimes solve their problems before the public becomes aware of them, thereby becoming ethical and successful again (e.g., [38.40]).

When Merton proposed the most influential version of Strain Theory in 1938, he identified three modes of adaptation to three increasing levels of strain. If a person was unable to achieve society's values by following conventional norms, frustration might drive him or her to violate the norms in pursuit of the value, what Merton called *innovation*. If this did not work, he or she might substitute both new values and new norms for those of society, in what he called *rebellion*. But if this failed as well, the person might abandon values, as well as norms and fall into a demoralized state of inaction called *retreatism*. Retreatism is inconsequential, and successful rebellion tends to radicalize society, but innovation (in Merton's sense of the term) can promote scientific and technical progress. In other work, Merton [38.41, 42] was one of the founders of the sociology of science and technology, but he never followed through on the implications of Strain Theory for science, and his distinctive concept of innovation was chiefly used to explain criminal behavior by poor people who violated the law in order to the get the values (i. e., money) endorsed by capitalist society.

There are many examples in the history of science in which Strain Theory can be used to explain scientific and technological innovations that turned out to be beneficial, rather than unethical, and others where the result was a mixture of benefit and harm that presumably could have been managed to emphasize the benefit. In the seventeenth century, England led the world in scientific progress, perhaps in a nationalistic reaction to its inferiority in population and other resources relative to France and Spain [38.43]. In the late nineteenth and early twentieth centuries, Germany lacked access to as abundant a supply of natural resources as had the great colonial powers, Britain and France. This inferiority may have stimulated many advances in German chemistry such as Adolf Bayer's synthesis of indigo dye and the nitrogen fixation method developed by Fritz Haber and Karl Bosch, which allowed Germany to continue producing explosives for World War I after it had lost access to foreign sources of nitrates. When the Treaty of Versailles concluding World War I limited German long-range artillery, the Germans developed the V-2 rocket largely to get around this restriction, thereby ushering in the age of space flight [38.44, 45].

In Control Theory, a person violates norms because he or she lacks stable social bonds that might prevent it [38.46–48]. This is a very successful scientific explanation of many common crimes, such as ordinary theft, and many kinds of criminals tend to lack stable friendships or family relationships that might restrain their deviant behavior. Control Theory might seem irrelevant to nanotechnology ethics because no isolated individual is in a position to make major decisions about new applications. The theory perfectly fits the Hollywood image of Dr. Frankenstein living like a hermit in a remote castle, where he is free to conduct any ghastly experiment he wishes, but no real scientist lives like this.

However, a variant of Control Theory explains that a community or organization can suffer from a partial but pervasive form of isolation, called *social disorganization*, in which communication channels are poor and many people deviate to a moderate degree because they are somewhat detached from each other. In a socially disorganized research laboratory studying the properties of a new nanostructured material, the quality of the science could suffer, pieces of information could fail to get to scientists who need them, and management would get incorrect impressions about the material's properties. This could lead to a chain reaction of unethical behavior in which no one individual was primarily to blame. For instance, the result could be a corporate decision to put into production a new kind of aircraft component that was subject to unpredictable catastrophic failure, causing crashes and loss of life. In a case such as this, the unethical behavior is scattered in many small pieces, committed by many individuals, any time one of them presents a dubious observation as a scientific fact. The best way to reassert the truth-oriented professional norms of science would be to rebuild good channels of communication and cooperation, reattaching the researchers to each other and to the scientific community.

Subculture Theory is similar to Control Theory, but operates on the level of groups rather than individuals. A person violates a rule because he or she belongs to a subculture that rejects the rule and even encourages violation of it [38.49, 50]. A subculture is a group or network having a set of norms, roles, and values that differ from those of the surrounding culture. At the extreme, it can have a radical ideology or dogma, but practically every social group is a subculture to at least a very mild extent. Social psychologist *Janis* [38.51] has shown that even very powerful groups can become deviant subcultures, if they isolate themselves from the wider community, as the Nixon White House did during the Watergate episode.

Even if a research laboratory is functioning properly, the organization of which it is a part can make harmful decisions if the top management is a subculture practicing what Janis calls "groupthink," the refusal to accept information that conflicts with the group's beliefs and commitments. Sometimes this may result in harm to others, for instance, if a new nanotechnology-enabled treatment for cancer were put into production before safe methods for administering it to patients in hospital settings were worked out. Or the harm could fall on the organization itself, for example, if a semiconductor company decided to invest its future in molecular logic gates without fully assessing the potential of carbon nanotube transistors instead. In that case, we might say the company deserves what it gets, but there is the ethical issue of its responsibility to stockholders and employees, including to the scientists in the research laboratory who were ready to do the studies needed for a correct decision.

As the example of Strain Theory illustrates, sometimes strain has positive consequences, rather than negative ones. Similarly, Learning Theory explains why an individual or organization might develop habits of ethical innovation, and both Control Theory and Subculture Theory could show that a person or group may innovate best when somewhat detached from conventional thinking. Thus, the issue is not one of avoiding these four factors, but of finding the right levels of them to encourage healthy innovation without unleashing serious unethical behavior. This is the fundamental trade-off in any free society: How to liberate human creativity while restraining greed, immorality, and inhumanity.

Strain Theory has been used to explain the emergence of collective behavior and social movements [38.52], including reform movements such as the environmental movement or crusades to establish professional codes of ethics. Concern about the social and ethical implications of nanotechnology can, therefore, energize the development of appropriate institutions to manage it wisely. Learning Theory suggests that a rewarding market for beneficial nanotechnology products could teach scientists, engineers, and managers to invest their energy in projects that would serve the needs of humanity. Both Control Theory and Subculture Theory stress the importance of strong networks of communication in restraining unethical behavior. Convergence of nanotechnology with the other NBIC fields would support healthy communication by unifying science, providing a universal technical language for interaction across fields, and creating a network of collaborative relationships across sciences, institutions, and sectors of society.

38.5 Public Opinion

Ethics and social implications are largely matters of social perception, and the public conception of nanotechnology is still in the early stages of developing. Nanoscale science and engineering are evolving from conventional work in materials science, physics, chemistry, and related fields, as the gradual emergence of new methodologies permits many kinds of observation and experimentation with inorganic structures having at least one dimension less than about 100 nm. This progress in *real nanotechnology* is solidly based in existing scientific knowledge, even as it achieves new discoveries and inventions.

At the same time, something quite different has been emerging that also calls itself nanotechnology: a social movement based on metaphors, approximations, and hopes. *Tolles* [38.53] calls it an "irrational vision." In social science, popular enthusiasm for a loosely defined set of unreasonable hopes is often called a *mania* [38.54], so I will call this movement *nanomania*.

The key moment in the emergence of nanomania, but not of nanoscience and nanotechnology themselves, was the 1986 publication of *Engines of Creation* by Drexler. This is a popular book, inspirational rather than technical, filled with verbal metaphors about what might possibly be accomplished if nanoscale assemblers could be created, rather like industrial robot arms and assembly lines. The plausibility of Drexler's argument, such as it is, comes from his constant reminder that biological systems include "machines" that assemble themselves and manufacture or manipulate other nanoscale structures: proteins, RNA, and DNA.

In a rambling discourse that reminds us that Leonardo da Vinci could not have predicted the details of late twentieth century technology, Drexler asserts that some time in the future it will be possible to build nanoscale assemblers that are able to make *anything that can be designed*. At this point he has outstripped his biological analogy, because the "nanomachines" of biology cannot do this; they can make only those specific things that evolution has programmed them to make [38.55]. Drexler says that suitably programmed assemblers will be able to make copies of themselves, as living organisms do.

The specific image he offers is of tiny robot arms grabbing atoms from the surrounding raw material and putting them together in the correct configurations. These imaginary assemblers can make anything, he suggests, so they can quickly replace all the factories and construction companies on Earth, giving humankind perfect abundance of anything we might want. The book then scans quickly through other dreams of science fiction, including artificial intelligence, cheap space flight, near immortality, and the danger to human life of uncontrolled self-replicating intelligent machines.

In some ways the book is laudable, inspiring non-technical readers with many conceivable applications of nanotechnology, some of which might actually turn out to be practical in the long run, discussing their social implications, and offering potential solutions to problems that might arise. The problem is that many people have come to the unfounded conclusion that Drexler's self-replicating universal assemblers are not only theoretically possible, but likely to be created in the near future. This is the fundamental error of nanomania, and it has called forth a reaction in the form of a *nanophobia* – the fear that all kinds of horrible evils will soon emerge from nanotechnology.

At the 2000 *Societal Implications* workshop, Drexler's assemblers were debunked by several participants, including Nobel prize-winner Smalley, discoverer of fullerenes and a pioneer of carbon nanotubes. *Smalley* observes that technical details of chemistry and a correct understanding of the properties of atoms and molecules are largely absent from nanomania [38.56]. In particular, the universal nanoscale assemblers envisioned by Drexler are physically impossible: "Because the fingers of a manipulator arm must themselves be made out of atoms, they have a certain irreducible size. There just isn't enough room in the nanometer-size reaction region to accommodate all the fingers of all the manipulators necessary to have complete control of the chemistry. . . . Manipulator fingers on the hypothetical self-replicating nanobot are not only too fat; they are also too sticky: the atoms of the manipulator hands will adhere to the atom that is being moved. So it will often be impossible to release this miniscule building block in precisely the right spot" [38.57].

The difficulty of developing reasonable public policies about nanotechnology when popular thinking is dominated by fanciful notions is illustrated by Chen's paper, "The Ethics of Nanotechnology", distributed online both by the Markkula Center for Applied Ethics at Santa Clara University and by BioScience Productions, Inc., which describes itself as "a non-partisan organization whose goal is the promotion of public literacy in the biosciences". Following Drexler's approach, *Chen* takes assumes that nanotechnology means the creation of molecular machines, "which could be used as molecular assemblers and disassemblers to build, repair, or tear down any physical or biological objects" [38.58]. Recognizing that such "nanites" may not be practical in the near future, Chen nonetheless suggests that we consider taking three possible actions soon:

1. Nanotechnology R&D should be banned
2. A nongovernmental regulatory or advisory commission should be set up
3. Adopt design guidelines
   - Nanomachines should only be specialized, not general purpose
   - Nanomachines should not be self-replicating

- Nanomachines should not be made to use an abundant natural compound as fuel
- Nanomachines should be tagged so they can be tracked

Chen notes that the first choice would undesirably prevent beneficial applications of molecular machines, but might not prevent "rogue researchers" from developing nanites anyway. However, readers of his article who are unfamiliar with real nanotechnology may not realize that most research in the field has nothing to do with these hypothetical nanites. It would take considerable wisdom to figure out how a regulatory commission might be set up, empowered, and managed effectively in such a way that it could ensure nanotechnology will be used for human benefit without stifling worthwhile innovations. The design guidelines flow from Chen's assumptions about what nanotechnology is, but any formal regulations embodying them could have serious negative unintended consequences, given that they start with a flawed definition of nanomachines.

Chen's rule that nanomachines be "specialized, not general purpose" may mean little, because a fully "general purpose" machine is as mythical as a "universal solvent" or "perpetual motion machine." However, a regulatory commission could do great harm by adopting poorly conceived definitions of *nanomachine* and *general purpose*. Suppose a very useful new abrasive were developed consisting of nanoparticles that changed shape, perhaps in response to temperature or force fluctuations, in some complex way that helped them do their intended job. *Webster's II New Riverside University Dictionary* defines *machine* thus: "a system, usually of rigid bodies, constructed and connected to change, transmit, and direct applied forces in a predetermined way to accomplish a particular objective, as performance." By this definition, the nanoscale abrasive particles would be machines. If they could be used to shape a wide range of industrial materials, they could be called general purpose. Yet this abrasive could be no more dangerous to humanity than ordinary sandpaper.

Strictly speaking, no entities on Earth are "self-replicating." Living organisms cannot replicate themselves without the help afforded them by the particular environment in which they dwell. All life is part of an ecology. Any regulation prohibiting self-replication would have to define very clearly what *self-replication* meant in the context of particular environments. Today, factory robots build parts of robots, but nobody worries about the ethical implications, perhaps because anybody can see that the manufacture process is entirely under the control of human beings. Because nanites are imaginary, and only vaguely defined, there are no facts to give nontechnical people comparable confidence.

Chen's rule against nanomachines that use "an abundant natural compound as fuel" presumably would prevent a nanite from rampaging uncontrolled in the natural environment. But this could outlaw the development of non-replicating nanoscale devices that can function only in very restricted laboratory or industrial systems, because most chemicals used in motors or batteries are abundant natural compounds.

Requiring that all nanoscale machines be "tagged so that they can be tracked" is reminiscent of the British Red Flag Act of 1865, which required that each horseless carriage be preceded by someone on foot waving a red flag. In the context of Chen's recommendations, *tracked* would seem to mean located, not merely identified once found. This could add an extremely heavy technical burden such as radioactive tagging, addition of a distinctive adulterating chemical, or even complex design features above the nanoscale to allow the nanomachine to be detected by microwave radio locators. If the machines are incapable of escaping into the natural environment, this burden would be not only ill-advised but ludicrous.

Has nanomania distorted the reception real nanotechnology is receiving? A survey done in 1999 found that the American business community was rather unaware of nanotechnology, only 2% saying they knew what it was and a further 2% reporting they had heard of it, but did not know what it was [38.59]. Two years later, a Web-based survey sponsored by the National Geographic Society found that science-attentive members of the general public are very enthusiastic about nanotechnology, and a rather large number of ideas about its benefits have already entered popular culture [38.60]. Statistical analysis of responses from 3,909 people revealed that they associate nanotechnology with other science-based technology, but do not connect it with pseudoscience. Significantly, when 598 people wrote paragraphs about nanotechnology, not a single respondent expressed concern about hypothetical dangers from self-reproducing nanites. Future surveys can examine the deepening popular awareness of nanotechnology, as the field itself progresses, and assess any public concerns that develop.

38.6 A Research Agenda

Sociologist *Etzkowitz* [38.61] has noted that the social sciences can play three different but mutually supportive roles in the development of nanotechnology:

1. Analyzing and contributing to the improvement of the processes of scientific discoveries that increasingly involve organizational issues where the social sciences have a long-term research and knowledge base.
2. Analyzing the effects of nanotechnology, whether positive or negative, expected or unintended, hypothetically and proactively and as they occur in real-time.
3. Evaluation of public and private programs to promote nanoscience and nanotechnology.

There are several ways social scientists can study the nanotechnology innovation process as it occurs [38.11, 62]. Cultural anthropologists and participant observation sociologists can enter nanoscience research teams, ethnographically documenting the behavior of the researchers over time as they frame their scientific problems, seek solutions, and labor collectively to understand the meaning of research results [38.63]. Fruitful sites for research on the innovation process include academe, industry, government laboratories, federal agencies, and professional societies.

Historical methods for collecting and analyzing written documents are also useful, even in cases where the documents are only days or weeks old, rather than the decades-old materials employed by most professional historians. Formal interviews, content analysis of communications such as e-mail messages, and questionnaires can provide more systematic kinds of data.

Application of a scientific idea to a technical problem, a technology transfer, and the introduction of products into the marketplace can be tracked through economic statistics on research and development investments, patent applications, and advertisements for new products and services. Some geographic areas, strata of society, and kinds of individuals and institutions experience technological change earlier than others, notably, the first adopters who try out an innovation before influencing other people to try it as well [38.64, 65]. Start-up companies are in a sense early adopters, even though they are also innovators, and thus they may be an especially fruitful if logistically difficult field for research.

There is a long tradition of systematic research on scientific publication, tracking the introduction of new concepts, charting the changing patterns of citation in scientific literature, and measuring the bulk of publications in different areas [38.66]. The progress of nanotechnology in modern culture could also be studied by tracking the introduction of new topics and courses in school and university curricula, mentions of nanoscale phenomena in the popular press, and the proliferation of both commercial and noncommercial Websites dedicated to nanotechnology. Professional societies and entrepreneurs are already creating series of new forums, symposia, journals, and job fairs oriented toward nanoscience and nanotechnology.

Nanotechnology may have significant impacts in all sectors of the economy, in most spheres of life, and on both short-term and long-term time scales. The intended consequences of any particular innovation can be determined simply by interviewing its investors or promoters and by inspecting their publications about it. Unintended consequences are much harder to study, because they require not only extensive research on the impacts on diverse areas of society, but also good scientific intuition about where to look for them. Consequences have consequences, in a continuing cascade, so the second-order consequences will also have to be studied [38.67]. Ultimately, this would require an extensive, vigorous interdisciplinary scientific community that is dedicated to research on social and technological change in general.

Technology is an elaborate system, embedded in the even more elaborate system that is global society. Through feedback among complex social subsystems, major phenomena can produce chaotic and unpredictable effects. Nanotechnology will have substantial impacts on many aspects of society in often different ways, so predictions of its influence will be difficult to make and evaluate scientifically. This means that researchers will have to invest considerable sophistication and effort in their work.

The social and ethical implications of nanotechnology are the results not merely of what the technology itself does, but also of how people react to it. Therefore, it will be important to understand the social acceptance, resistance, or rejection of nanotechnology at different times and in different places and human contexts. Interviews, focus groups, and questionnaire surveys can measure emotional, cognitive, and psychosocial parameters [38.68]. The combined methodologies of political science, sociology, and socio-legal studies will be required to chart the process of regulatory review and approval, court decisions that actively permit or prohibit the use of the technology, mobilization of political support and opposition, and the activities of relevant social movements.

Ultimately, the test of the various nanotechnologies will be their benefit for human beings, as measured by economic growth, improved health and longevity, environmental protection, strengthened security, social vitality, and enhanced human capabilities. Convergence with other NBIC technologies will complicate the challenge for researchers and policymakers who wish to understand the social and ethical implications of nanotechnology, but it will also magnify the potential benefits to humanity many times over. Success will be most complete and most probable if scientific research on these implications is an integral part of the effort from the very beginning.

References

38.1 B. Joy: Why the future doesn't need us, Wired **8**(4) (April 2000) 238–262

38.2 B.J. Feder: Opposition to nanotechnology, New York Times (August 19 2002) (http://www.nytimes.com/)

38.3 G. H. Reynolds: *Forward to the Future: Nanotechnology and Regulatory Policy* (Pacific Research Institute, San Francisco 2002)

38.4 J. S. Brown, P. Duguid: Don't count society out: A response to Bill Joy. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 37–45

38.5 R. W. Siegel, E. Hu, M. C. Roco (Eds.): *Nanostructure Science and Technology* (Kluwer, Dordrecht 1999)

38.6 M. C. Roco, R. S. Williams, P. Alivisatos (Eds.): *Nanotechnology Research Directions* (Kluwer, Dordrecht 2000)

38.7 M. Roco, R. Tomellini: *Nanotechnology: Revolutionary Opportunities and Societal Implications* (Office for Official Publications of the European Communities, Luxembourg 2002)

38.8 M. S. Meyer: Socio-economic research on nanoscale science and technology: A european overview and illustration. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 278–311

38.9 R. H. Smith: Social, ethical, and legal implications of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 257–271

38.10 V. Weil: Ethical issues in nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 244–251

38.11 M. C. Roco, W. S. Bainbridge (Eds.): *Societal Implications of Nanoscience and Nanotechnology* (Kluwer, Dordrecht 2001)

38.12 L. B. Lave: Lifecycle/sustainability implications of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 205–212

38.13 National Research Council: *Small Wonders, Endless Frontiers: A Review of the National Nanotechnology Initiative* (National Academy Press, Washington D.C. 2002)

38.14 M. C. Roco, W. S. Bainbridge (Eds.): *Converging Technologies to Improve Human Performance* (Kluwer, Dordrecht 2003)

38.15 M. C. Suchman: Envisioning life on the nanofrontier. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 271–276

38.16 N. Gingrich: The age of transition. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 29–35

38.17 G. Yonas, S. T. Picraux: National needs drivers for nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 46–55

38.18 P. R. Ehrlich: *The Population Bomb* (Ballantine, New York 1968)

38.19 K. Davis, M. S. Bernstam, R. Ricardo-Campbell (Eds.): *Below-Replacement Fertility in Industrial Societies: Causes, Consequences, Policies (A supplement to Population and Development Review)* 1986)

38.20 B. J. Wattenberg: *The Birth Dearth* (Ballantine, New York 1987)

38.21 Central Intelligence Agency: *Long-Term Global Demographic Trends: Reshaping the Geopolitical Landscape* (Central Intelligence Agency, Washington D.C. 2001)

38.22 Demography and the west, Economist **364**(8287) (August 2002) 20–22

38.23 D. L. Costa, R. H. Steckel: Long-term trends in health, welfare, and economic growth in the United States. In: *Health and Welfare During Industrialization*, ed. by R. H. Steckel, R. Floud (Univ. Chicago Press, Chicago 1997) pp. 47–89

38.24 S. H. Preston: *American Longevity: Past, Present, and Future* (Center for Policy Research, Syracuse University, Syracuse 1996)
38.25 HHS (U.S. Department of Health and Human Services): *Chronic Diseases and Their Risk Factors: The Nation's Leading Causes of Death* (Centers for Disease Control, Atlanta 1999)
38.26 HHS (U.S. Department of Health and Human Services): *Reducing Tobacco Use: A Report of the Surgeon General* (Centers for Disease Control, Atlanta 2000)
38.27 D. M. Cutler, M. McClellan, J. P. Newhouse, D. Remler: Are medical prices declining, Quarterly J. Economics **113** (1998) 991–1024
38.28 D. M. Cutler, E. Richardson: Your money and your life: The value of health care and what affects it. In: *Frontiers in Health Policy Research*, ed. by A. M. Garber (MIT, Cambridge 1999) pp. 99–132
38.29 V. R. Fuchs: The future of health economics, J. Health Economics **19** (2000) 141–157
38.30 F. W. Hollmann, T. J. Mulder, J. E. Kallan: *Methodology and Assumptions for the Population Projections of the United States: 1999 to 2100* (United States Census Bureau, Washington D.C. 2000)
38.31 R. Doering: Social implications of scaling to nanoelectronics. In: *Social Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 84–93
38.32 R. S. Williams, P. J. Kuekes: W've only just began. In: *Social Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 103–108
38.33 Central Intelligence Agency: *Experts See Risks to US IT Superiority Beyond Moore's Law*, report OTI IA 2002-151 (Central Intelligence Agency, Directorate of Intelligence, Washington D.C. 2002) pp. 2002–151
38.34 E. Tenner: Nanotechnology and unintended consequences. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 311–318
38.35 G. C. Homans: *Social Behavior: Its Elementary Forms* (Harcourt Brace Jovanovich, New York 1974)
38.36 R. L. Akers: *Deviant Behavior: A Social Learning Approach* (Wadsworth, Belmont 1985)
38.37 R. Stark, W. S. Bainbridge: *A Theory of Religion* (Rutgers University Press, New Brunswick 1987)
38.38 R. K. Merton: *Social structure and anomie* (Free Press, New York 1968) pp. 185–214, first published 1938
38.39 R. A. Cloward, L. E. Ohlin: *Delinquency and Opportunity* (Free Press, New York 1960)
38.40 A. P. Sloan: *My Years with General Motors* (Doubleday, Garden City 1964)
38.41 R. K. Merton: *Science, Technology and Society in Seventeenth Century England* (H. Fertig, New York 1970)
38.42 R. K. Merton: *The Sociology of Science: Theoretical and Empirical Investigations* (University of Chicago Press, Chicago 1973)
38.43 L. Greenfeld: *Nationalism: Five Roads to Modernity* (Harvard Univ. Press, Cambridge 1992)
38.44 W. S. Bainbridge: *The Spaceflight Revolution* (Wiley, New York 1976)
38.45 W. S. Bainbridge: *Goals in Space: American Values and the Future of Technology* (State Univ. New York Press, Albany 1991)
38.46 T. Hirschi: *Causes of Delinquency* (University of California Press, Berkeley 1969)
38.47 C. Shaw, H. D. McKay: *Delinquency Areas* (University of Chicago Press, Chicago 1929)
38.48 M. R. Gottfredson, T. Hirschi: *A General Theory of Crime* (Stanford Univ. Press, Stanford 1990)
38.49 E. H. Sutherland: *Principles of Criminology* (Lippincott, Philadelphia 1947)
38.50 G. Sykes, D. Matza: Techniques of neutralization: A theory of delinquency, Am. Sociol. Rev. **22** (1957) 664–670
38.51 I. L. Janis: *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (Houghton Mifflin, Boston 1982)
38.52 N. J. Smelser: *Theory of Collective Behavior* (Free Press, New York 1962)
38.53 W. M. Tolles: National security aspects of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 218–237
38.54 W. S. Bainbridge: Collective behavior and social movements. In: *Sociology*, ed. by R. Stark (Wadsworth, Belmont 1985) pp. 493–523
38.55 V. Vogel: Societal impacts of nanotechnology in education and medicine. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 180–186
38.56 R. F. Service: Is nanotechnology dangerous?, Science **290** (2000) 1526–1527
38.57 R. E. Smalley: Of chemistry, love and nanobots, Sci. Am. **285** (2001) 76–77
38.58 A. Chen: *The Ethics of Nanotechnology* (BioScience Productions and Markkula Center for Applied Ethics, Santa Clara University, 2002) www.actionbioscience.org/newfrontiers/chen.html and www.scu.edu/ethics/publications/submitted/chen/nanotechnology.html
38.59 J. Canton: The strategic impact of nanotechnology on the future of business and economics. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 114–121
38.60 W. S. Bainbridge: Public attitudes toward nanotechnology, J. Nanopart. Res. **4** (2002) 561–570

38.61 H. Etzkowitz: Nano-science and society: Finding a social basis for science policy. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 121–128
38.62 J. S. Carroll: Social science research methods for assessing societal implications of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 238–244
38.63 B. A. Nardi: A cultural ecology of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 318–324
38.64 E. Katz, P. F. Lazarsfeld: *Personal Influence* (Free Press, Glencoe 1955)
38.65 E. M. Rogers: *Social Change in Rural Society* (Appleton-Century-Crofts, New York 1960)
38.66 D. Crane: *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (University of Chicago Press, Chicago 1972)
38.67 R. A. Bauer: *Second-Order Consequences: A Methodological Essay on the Impact of Technology* (MIT, Cambridge 1969)
38.68 P. B. Thompson: Social acceptance of nanotechnology. In: *Societal Implications of Nanoscience and Nanotechnology*, ed. by M. C. Roco, W. S. Bainbridge (Kluwer, Dordrecht 2001) pp. 251–256

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Acknowledgements

4 Nanowires

by Mildred S. Dresselhaus, Yu-Ming Lin, Oded Rabin, Marcie R. Black, Gene Dresselhaus

The authors gratefully acknowledge the stimulating discussions with Professors Charles Lieber, Gang Chen, S. T. Lee, Arun Majumdar, Peidong Yang, and Jean-Paul Issi, Dr. Joseph Heremans, and Ted Harman. The authors are grateful for support for this work by the ONR Grant #000140-21-0865, the MURI program subcontract PO #0205-G-7A114-01 through UCLA, and DARPA contract #N66001-00-1-8603.

6 Stamping Techniques for Micro and Nanofabrication: Methods and Applications

by John A. Rogers

The author extends his deepest thanks to all of the collaborators who contributed the work described here.

8 MEMS/NEMS Devices and Applications

by Darrin J. Young, Christian A. Zorman, Mehran Mehregany

The authors wish to thank Wen H. Ko for the helpful discussions and suggestions and Michael Suster and Joseph Seeger for preparing the figures.

10 Therapeutic Nanodevices

by Stephen C. Lee, Mark Ruegsegger, Philip D. Barnes, Bryan R. Smith, Mauro Ferrari

The authors gratefully acknowledge many friends and colleagues whose input helped shape this article. Particularly, we acknowledge Beth S. Lee for many, many helpful discussions and for unflagging support. We also acknowledge Phil Streeter for many of the same services, performed in the office of friend rather than spouse. We acknowledge the support services of Anita Bratcher in manuscript preparation and the artistic stylings of Vladimir Marukhlenko for the figures incorporated in the manuscript. We also acknowledge Carol Bozarth for her kind and spontaneous support of the physical process of manuscript editing. This work is dedicated to the memories of Mildred A. Lee, Antonio Ferrari and Marialuisa Ferrari and the multitude of others whose lives have been tragically shortened by cancer, with the determination that therapeutic nanotechnology be used to help abolish the terrible power of the disease over human life.

13 Noncontact Atomic Force Microscopy and Its Related Topics

by Seizo Morita, Franz J. Giessibl, Yasuhiro Sugawara, Hirotaka Hosoi, Koichi Mukasa, Akira Sasahara, Hiroshi Onishi

Thanks to Tom Albrecht, Alexis Baratoff, Hartmut Bielefeldt, Gerd Binnig, Dominik Brändlin, Peter van Dongen, Urs Dürig, Christoph Gerber, Stefan Hembacher, Markus Herz, Lukas Howald, Christian Laschinger, Ulrich Mair, Jochen Mannhart, Thomas Ottenthal, Calvin Quate, Marco Tortonese, and the BMBF for funding under project no. 13N6918.

18 Surface Forces and Nanorheology of Molecularly Thin Films

by Marina Ruths, Alan D. Berman, Jacob N. Israelachvili

This work was supported by ONR grant N00014-00-1–0214. M. R. Luths thanks the Academy of Finland for financial support.

19 Scanning Probe Studies of Nano-scale Adhesion Between Solids in the Presence of Liquids and Monolayer Films

by Robert W. Carpick, James D. Batteas

We gratefully acknowledge the help of Ms. Erin Flater, who provided valuable assistance and insights into the literature on capillary formation. RWC acknowledges support of a career award from the National Science Foundation, grant #CMS-0134571.

24 Mechanics of Biological Nanotechnology

by Rob Phillips, Prashant K. Purohit, Jané Kondev

We happily acknowledge useful discussions with Kai Zinn, Jon Widom, Bill Gelbart, Andy Spakowitz, Zhen-Gang Wang, Ken Dill, Carlos Bustamante, Tom Powers, Larry Friedman, Jack Johnson, Pamela Bjorkman, Paul Wiggins, Steve Williams, Wayne Falk, Adrian Parsegian, Alasdair Steven and Steve Quake. RP and PP acknowledge support of the NSF through grant number 9971922, the NSF supported CIMMS center, and the support of the Keck Foundation. JK is supported

by the NSF under grant number DMR-9984471 and is a Cottrell Scholar of Research Corporation.

30 Nanotechnology for Data Storage Applications

by Dror Sarid, Brendan McCarthy, Ghassan E. Jabbour

The authors would like to thank Digital Instruments (Veeco) for contributing their Multi-Mode Nanoscope III, the Department of Energy (DE-FG-3-02ER46013/A001) for a generous grant, EMC for their generous gift, and the Vice President for Research, University of Arizona, for equipment support.

31 The "Millipede" – A Nanotechnology-Based AFM Data-Storage System

by Gerd K. Binnig, G. Cherubini, M. Despont, Urs T. Duerig, Evangelos Eleftheriou, H. Pozidis, Peter Vettiger

It is our pleasure to acknowledge our colleagues T. Albrecht, T. Antonakopoulos, P. Bächtold, A. Dholakia, U. Drechsler, B. Gotsmann, W. Häberle, D. Jubin, M. A. Lantz, T. Loeliger, H. E. Rothuizen, R. Stutz, and D. Wiesmann for their invaluable contributions to the millipede project.

In addition, thanks and appreciation go to H. Rohrer for his contribution to the initial millipede vision and concept and to our former collaborators, J. Brugger, now at the Swiss Federal Institute of Technology, Lausanne (Switzerland), M. I. Lutwyche, now at Seagate, Pittsburg, IL, and W. P. King, now at Georgia Tech, Atlanta, GA, as well as to K. Goodson, T. W. Kenny, and C. F. Quate of Stanford University, CA.

We are also pleased to acknowledge stimulating discussions with and encouraging support from our colleagues W. Bux and P. F. Seidler of the IBM Zurich Research Laboratory, J. Mamin, D. Rugar, and B. D. Terris of the IBM Almaden Research Center, San Jose, CA, and G. Hefferon of IBM, East Fishkill, NY.

Special thanks go to J. Frommer, C. Hawker, J. Mamin, and R. Miller of the IBM Almaden Research Center for their enthusiastic support in identifying and synthesizing alternative polymer media materials, and to H. Dang, A. Sharma, and S. Sri-Jayantha of the IBM T. J. Watson Research Center, Yorktown Heights, NY, for their contributions to the work on servo control.

35 Thermo- and Electromechanics of Thin-Film Microstructures

by Martin L. Dunn, Shawn J. Cunningham

We are most grateful for the assistance of Ms. Yanhang Zhang, a Ph.D. student at the University of Colorado, for her help with the preparation of figures. Many of the calculations and measurements also draw from her work. MLD acknowledges support from DARPA, Sandia National Laboratories, and the AFOSR for support of aspects of his research that appear in this work.

About the Authors

Chong H. Ahn

University of Cincinnati
Department of Electrical
and Computer Engineering
and Computer Science
Cincinnati, OH, USA
chong.ahn@uc.edu
http://www.BioMEMS.uc.edu/

Dr. Chong Ahn is a Professor of Electrical and Computer Engineering at the University of Cincinnati. He obtained his Ph.D. degree in Electrical Engineering from the Georgia Institute of Technology in 1993 and then worked as a postdoctoral fellow at IBM T.J. Watson Research Center. His research interests include all aspects of design, fabrication, and characterization of magnetic MEMS devices, microfluidic devices, protein chips, lab-on-a-chips, nano biosensors, point-of-care testing and BioMEMS systems. He is an associate editor of the IEEE Sensors Journal.

Boris Anczykowski

nanoAnalytics GmbH
Münster, Germany
anczykowski@nanoanalytics.com
http://www.nanoanalytics.com

Dr. Boris Anczykowski is a physicist with an extensive research background in the field of dynamic Scanning Force Microscopy. He co-invented the Q-Control technique and received the Innovation Award Münsterland for Science and Economy in 2001 for this achievement. He is a managing director and co-founder of nanoAnalytics GmbH, a company specialized in the characterization of surfaces and interfaces on the micro- and nanometer scale.

Massood Z. Atashbar

Western Michigan University
Department of Electrical
and Computer Engineering
Kalamazoo, MI, USA
massood.atashbar@wmich.edu
http://homepages.wmich.edu/~zandim

Professor Atashbar's main area of research is nanotechnology and wireless sensors. He has been working in the field of micro/nanosensors for chemical and physical sensing specifically integrated smart wireless surface acoustic wave sensors and systems. He is a senior member of IEEE and an associate editor of the International Journal of Modelling and Simulation

Wolfgang Bacsa

Université Paul Sabatier
Laboratoire de Physique des Solides (LPST)
Toulouse, France
bacsa@lpst.ups-tlse.fr
http://www.lpst.ups-tlse.fr/users/
wolfgang/

Wolfgang Bacsa is an expert in the emerging field of Nano-Optics and Carbon Nanotubes. He has a Ph.D. from the Swiss Federal Institute of Technology (ETH) Zurich in Physics and has extensive experience in condensed matter physics, optics, microscopy, synthesis of ultra-thin films and nanostructured carbon. Professor Bacsa worked at the ETH Zürich, PennState Universiy and EPFL Lausanne.

William Sims Bainbridge

National Science Foundation
Division of Information
and Intelligent Systems
Arlington, VA, USA
wbainbri@nsf.gov
http://mysite.verizon.net/william.
bainbridge

William Sims Bainbridge earned his Ph.D. in Sociology from Harvard University in 1975 and is the author of 10 books, 4 textbook-software packages, and about 150 shorter publications in information science, social science of technology, and the sociology of culture. He has represented the social and behavioral sciences on five US government advanced technology initiatives: High Performance Computing and Communications, Knowledge and Distributed Intelligence, Digital Libraries, Information Technology Research, and Nanotechnology.

Antonio Baldi

Institut de Microelectronica
de Barcelona (IMB)
Centro National Microelectrónica
(CNM-CSIC)
Barcelona, Spain
Antoni.baldi@cnm.es

Antonio Baldi received his PhD from the Universitat Autonoma de Barcelona, Spain, in 2001. He is a postdoctoral fellow at the University of Minnesota. His current research is focused on the fabrication and testing of bioMEMS, on the development of new sensors and microfluidic devices incorporating stimuli-sensitive hydrogels and on the development of inductively coupled wireless MEMS.

Philip D. Barnes

Ohio State University
Biomedical Engineering Center
Columbus, OH, USA
d_skill@yahoo.com

Phillip D. Barnes is a graduate student at the Biomedical Engineering Center at the Ohio State University. His background is in electrical engineering and his current interests are in sensors and imaging devices.

James D. Batteas

National Institute
of Standards and Technology
Surface and Microanalysis Science
Division
Gaithersburg, MD, USA
james.batteas@nist.gov
http://www.cstl.nist.gov/div837/837.03/

Dr. James D. Batteas is a research chemist in the Surface and Interface Research Group of the National Institute of Standards and Technology. The group's current research involves developing and charcterizing the organization and properties of nanoscale materials and devices on surfaces, this includes nanoparticles for use in optoelectric and optomagnetic gating, molecular electronics, and sensors. His group is also investigating friction, adhesion and wear mechanisms on the nanoscale.

Roland Bennewitz

McGill University
Physics Department
Montreal, QC, Canada
roland@physics.mcgill.ca

Roland Bennewitz studied physics in Freiburg and Berlin where he received his PhD for work on defects at surfaces of insulators. He is now assistant at the University of Basel where his research activities focus on high-resolution force microscopy as a tool in nanotribology and surface science.

Alan D. Berman

Monitor Venture Enterprises
Los Angeles, CA, USA
alan.berman.2001@anderson.ucla.edu

Alan Berman received his Ph.D. from the Chemical Engineering Department at the University of California, Santa Barbara. He is involved in technology commercialisation, assisting large technology companies to maximize return on R&D investments by finding and developing applications for their intellectual property outside of their core business. Value is realized through technology licensing, joint ventures and venture capital funded new entities. Technical interests are in fields of tribology, MEMS, and nanotechnology.

Bharat Bhushan

The Ohio State University
Nanotribology Laboratory
for Information Storage
and MEMS/NEMS
Columbus, OH, USA
bhushan.2@osu.edu

Dr. Bharat Bhushan is an Ohio Eminent Scholar and The Howard D. Winbigler Professor in the Department of Mechanical Engineering, a Graduate Research Faculty Advisor in the Department of Materials Science & Engineering, and the Director of the Nanotribology Laboratory for Information Storage & MEMS/NEMS (NLIM) at the Ohio State University, Columbus, Ohio. He holds two M.S., a Ph.D. in mechanical engineering/mechanics, an MBA, and three semi-honorary and honorary doctorates. His research interests are in micro/nanotribology and its applications to magnetic storage devices and MEMS/NEMS (Nanotechnology). He has authored 5 technical books, 45 handbook chapters, more than 450 technical papers in referred journals, and more than 60 technical reports, edited more than 25 books, and holds 14 U.S. patents.

Gerd K. Binnig

IBM Zurich Research Laboratory
Micro-/Nanomechanics
Rüschlikon, Switzerland
gbi@zurich.ibm.com

Gerd Binnig obtained his Ph.D. from the Johann Wolfgang Goethe University, Frankfurt, Germany, and joined IBM Research in 1978. He was corecipient of the 1986 Nobel Prize in Physics for the invention of the scanning tunnelling microscope, and he also invented the atomic force microscope. His current research interests are micro- and nanosystem techniques and "Fractal Darwinism", a theory he developed to describe complex systems.

Marcie R. Black

Massachusetts Institute
of Technology
Department of Electrical Engineering and
Computer Science
Cambridge, MA, USA
marcie@alum.mit.edu
http://web.mit.edu/mrb/www/prof.html

Marcie Black recently received her Ph.D. from Prof. Dresselhaus's research group at MIT studying the optical properties of nanowires. In particular, she identified the dominant optical absorption mechanism in the IR of bismuth nanowires as an indirect interband transition that is enhanced over bulk bismuth. Currently she is studying organic opto-electronics with an emphasis on photovoltaics.

Jean-Marc Broto

University Toulouse III
Laboratoire National des Champs
Magnétiques Pulsés (LNCMP)
Toulouse, France
broto@insa-tlse.fr
http://www.insa-tlse.fr/

Jean-Marc Broto is Professor at the Universite Toulouse III, France. He is a specialist in electronic transport and magnetization properties under high magnetic fields and contributed to the discovery of the Giant Magnetoresistance in 1988.

Robert W. Carpick

University of Wisconsin-Madison
Department of Engineering Physics
Madison, WI, USA
carpick@engr.wisc.edu
http://www.engr.wisc.edu/ep/faculty/
carpick_robert.html

Robert Carpick has a Ph.D. in Physics (1997) from the University of California at Berkeley. He has been an Assistant Professor at the University of Wisconsin-Madison since 2000. He carries out research and publishes in the areas of nanotribology, nanomechanics, nanostructured materials, and scanning probe development. He serves on the editorial board of Review of Scientific Instruments. In 2002 he received a Faculty Early Career Development award from the U.S. National Science Foundation.

Tsung-Lin Chen

National Chiao Tung University
Department
of Mechanical Engineering
Shin Chu, Taiwan
tsunglin@mail.nctu.edu.tw

Tsung-Lin Chen received his B.S. and M.A. degrees in mechanical engineering from the National Tsing-Hua University, Taiwan in 1990 and 1992, respectively. He received Ph.D. degree in mechanical engineering from University of California at Berkeley, USA in 2001. He is currently an assistant professor in National Chiao Tung University in Taiwan. His research interests include MEMS devices design and MEMS fabrication process development.

Yu-Ting Cheng

National Chiao Tung University
Department
of Electronics Engineering
& Institute of Electronics
HsinChu, Taiwan
ytcheng@faculty.nctu.edu.tw

Yu-Ting Cheng received his Ph.D. degree in Electrical Engineering at the University of Michigan, Ann Arbor, in 2000. After his graduation, he worked for IBM Watson Research Center, Yorktown Heights, as a research staff member. He is a member of IEEE, IOP, and Phi Tau Phi. Currently he is Assistant Professor at the National Chiao Tung University, Taiwan. His research interests include the development of novel materials and fabrication technologies for MEMS/NEMS applications and microsystems integration.

Giovanni Cherubini

IBM Zurich Research Laboratory
Storage Technologies
Rüschlikon, Switzerland
cbi@zurich.ibm.com

Dr. Giovanni Cherubini received a Ph.D. degree in electrical engineering from the University of California, San Diego, in 1986, and joined IBM Research in 1987. His interests include high-speed data transmission and data storage systems. He is Editor for CDMA systems, for IEEE Transactions on Communications, and served as Guest Editor for the IEEE Journal on Selected Areas in Communications issues on access technologies (1995) and on multiuser detection techniques (2001–2002). He is co-author of the book Algorithms for Communications Systems and their Applications.

Jin-Woo Choi

Louisiana State University
Department of Electrical
and Computer Engineering
Baton Rouge, LA, USA
choi@ece.lsu.edu

Jin-Woo Choi received his B.S. and M.S. degree in Electrical Engineering from Seoul National University in Korea in 1994 and 1996, respectively. He received his Ph.D. degree in Electrical Engineering from the University of Cincinnati in 2000. Now he is an Assistant Professor at Louisiana State University, Baton Rouge, Louisiana. His current research activities include magnetic particle separators, microfluidic systems for biochemical detection, micro total analysis systems (μ-TAS), bioelectronics, and BioMEMS components and systems.

Shawn J. Cunningham

WiSpry, Inc.
Colorado Springs Design Center
Colorado Springs, CO, USA
shawn.cunningham@wispry.com

Shawn Cunningham is working on the development of RF MEMS switch and associated processes with wiSpry, Inc. His interests include materials characterization, reliability, and Design for MEMS Manufacturability. Prior to joining wiSpry, Shawn pursued MEMS research and product development at Coventor, Ford Microelectronics, and the University of Utah's Center for Engineering Design and in collaboration with the Univerisity of Colorado.

Michel Despont

IBM Zurich Research Laboratory
Micro-/Nanomechanics
Rüschlikon, Switzerland
dpt@zurich.ibm.com

Michel Despont received his Ph.D. in physics from the University of Neuchâtel, Switzerland, in 1996. After a postdoctoral fellowship at IBM's Zurich Research Laboratory, he was visiting scientist at the Seiko Instrument Research Laboratory in Japan in 1997. His current research at IBM focuses on the development of micro- and nanomechanical devices and of processes to fabricate so-called system-on-chip.

Gene Dresselhaus

Massachusetts Institute
of Technology
Francis Bitter Magnet Laboratory
Cambridge, MA, USA
gene@mgm.mit.edu
http://web.mit.edu/fbml/cmr/

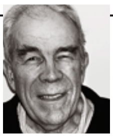

Gene Dresselhaus received his Ph.D. in physics from the University of California in 1955. He was a faculty member at the University of Chicago, and assistant professor at Cornell before joining MIT Lincoln Laboratory in 1960 as a staff member. In 1976 he assumed his current position at the MIT Francis Bitter Magnet Laboratory. His area of interest is the electronic structure of nanomaterials and he has co-authored with M.S. Dresselhaus several books on fullerenes, nanowires, and nanotubes.

Mildred S. Dresselhaus

Massachusetts Institute
of Technology
Department
of Electrical Engineering
and Computer Science
and Department of Physics
Cambridge, MA, USA
millie@mgm.mit.edu
http://www.eecs.mit.edu/faculty/index.html

Mildred Dresselhaus received her Ph.D. in physics from the University of Chicago in 1958. She joined the MIT faculty in 1967. She has been active in research across broad areas of solid state physics, especially in carbon science. Her present research activities focus on carbon nanotubes, bismuth nanowires, low dimensional thermoelectricity, and novel forms of carbon. She is the recipient of the National Medal of Science and 17 honorary degrees.

Martin L. Dunn

University of Colorado at Boulder
Department of Mechanical Engineering
Boulder, CO, USA
martin.dunn@colorado.edu

Martin L. Dunn received the Ph.D. in Mechanical Engineering from the University of Washington and was a postdoctoral appointee at Sandia National Laboratories. His research focuses on the micromechanical behavior of materials and structures. He has published over 75 articles in archival journals, and his research has been sponsored by NSF, DOE, NIST, DARPA, AFOSR, and Sandia National Laboratories.

Urs T. Dürig

IBM Zurich Research Laboratory
Micro-/Nanomechanics
Rüschlikon, Switzerland
drg@zurich.ibm.com

Urs Dürig received a Ph.D. degree from the Swiss Federal Institute of Technology, Zurich, in 1984. He joined IBM as a post-doc working on near-field optical microscopy. He is Research Staff Member since 1986: He worked in the field of scanning tunnelling and dynamic force microscopy. In 1997, he joined the Micro/Nanomechanics group focusing on polymer material issues and thermal modelling.

Evangelos Eleftheriou

IBM Zurich Research Laboratory
Storage Technologies
Rüschlikon, Switzerland
ele@zurich.ibm.com

He received a Ph.D. in Electrical Engineering from Carleton University, Ottawa, Canada, in 1985, and joined IBM Research in 1986. His research focuses on signal processing and coding for recording and transmission systems. His current research interests include nanotechnology, in particular probe-storage techniques.
Dr. Eleftheriou was elected IEEE Fellow in 2002.

Mauro Ferrari

Ohio State University
Biomedical Engineering Center
Columbus, OH, USA
Ferrari.5@osu.edu
http://bmew.bme.ohio- state.edu/
bmeweb3/bme_faculty.htm

Mauro Ferrari is the Edgar Hendrickson Professor of Biomedical Engineering of the Ohio State University. A mechanical engineer by training, he also holds appointments in Internal Medicine, Mechanical Engineering and Materials Science. Professor Ferrari is a globally recognized figure in biomedical nanotechnology and micro and nanofabrication. His current interests are application of novel fabrication technologies to innovative drug delivery devices for treatment of cardiovascular disease, palliative care and oncology.

Emmanuel Flahaut

Université Paul Sabatier
CIRIMAT (Centre Interuniversitaire
de Recherche et d'Ingénierie
des Matériaux)
Toulouse, France
flahaut@chimie.ups-tlse.fr
http://eflahaut.nano.free.fr

Emmanuel Flahaut obtained his Ph.D in Materials Science in Toulouse working on CCVD synthesis of carbon nanotubes (CNTs) and dense ceramic-based composites including CNTs. He spent then more than one year as a post-doctoral researcher in Malcolm Green's Group in Oxford to work mainly on the filling of CNTs. He is now a permanent CNRS researcher at the University of Toulouse.

Lásló Forró

Swiss Federal Institute
of Technology (EPFL)
Institute of Physics
of Complex Matter
Lausanne, Switzerland
laszlo.forro@epfl.ch
http://nanotubes.epfl.ch

Professor László Forró is working on the synthesis, physical properties and manipulation of carbon nanostuctures and nanostuctured arrays, as well as, mechanical properties of carbon nanotubes, carbon onions, biological tubular systems. Transport and electron spin resonance studies of molecular materials, quasi-one-dimensional organic metals, organic superconductors, cuprates, manganates and fullerenes – up to high pressures. Tunneling spectroscopy in cuprate and fullerene superconductors. Optical properties of strongly correlated systems and biomaterials.

Jane Frommer

IBM Almaden Research Center
Department of Science
and Technology
San Jose, CA, USA
frommer@Almaden.ibm.com

Following a Ph.D. in Organometallic Chemistry from Caltech, Dr. Jane Frommer has been involved in a diverse set of research areas, including electronically conducting polymers and scanning probe microscopy. Her present AFM lab at IBM collaborates with a wide variety of laboratories involved in materials research, including lithography, chromatography, and storage. Common to all these studies is Frommer's interest in the properties and structure of molecules in confined geometries.

Harald Fuchs

Universität Münster
Physikalisches Institut
Münster, Germany
fuchsh@uni-muenster.de
http://www.uni-muenster.de/Physik/PI/Fuchs

1984 Ph.D. Universität des Saarlandes with Prof. H. Prof Gleiter (nano crystalline Systems), 1984–1985, Post doc with IBM Research Lab. Zurich in the group of G. Binnig and H. Rohrer, 1985–1993 Project manager 'Ultrathin Organic Films' with BASF AG, Ludwigshafen, Germany. Since 1993 he is full Professor and Director at the Physical Institute of the University of Münster, 2000: Cofounder and Scientific Director of the Center for Nanotechnology (CeNTech).

Franz J. Giessibl

Universität Augsburg
Lehrstuhl für Experimentalphysik VI
Augsburg, Germany
franz.giessibl@physik.uni-augsburg.de
http://www.physik.uni-augsburg.de/exp6/

Dr. Giessibl is working in atomic force microscopy and scanning tunnelling microscopy in ultrahigh vacuum at room temperature and low temperatures. He is a Steering Committee Member of the International Conference on Noncontact Atomic Force Microscopy. He received the R&D 100 Award Chicago 1994, the German Nanoscience Award 2000, and the Rudolf Kaiser Price in 2001.

Enrico Gnecco

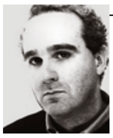

University of Basel
Department of Physics
Basel, Switzerland
Enrico.Gnecco@unibas.ch
http://monet.physik.unibas.ch/~gnecco/

Enrico Gnecco studied physics in Genoa, where he received his PhD for work on the mechanism of growth of nanostructured carbon films. He is now assistant at the University of Basel, where his research activities focus on friction force microscopy and molecular machinery.

Gérard Gremaud

Swiss Federal Institute
of Technology (EPFL)
Institute of Physics
of Complex Matter
Lausanne, Switzerland
gremaud@epfl.ch

Dr. Gérard Gremaud is a physicist and senior lecturer at the Ecole Polytechnique Fédérale de Lausanne (EPFL). He is active in the research fields of dislocation dynamic, acoustic and atomic force microscopy and granular physics. He is also responsible for the teaching of metrology and practical works to the physics students.

Jason H. Hafner

Rice University
Department of Physics
& Astronomy
Houston, TX, USA
hafner@rice.edu
http://www.ruf.rice.edu/~hafner

Jason Hafner earned his Ph.D. in physics from Rice University in 1998. He then held an NIH postdoctoral fellowship at Harvard University working on nanotube probes for high resolution biological atomic force microscopy. In 2001 he returned to Rice as an assistant professor where his group is pursuing various biophysical applications of scanned probe microscopy and nanomaterials. Dr. Hafner received a Beckman Young Investigator award in 2002.

Stefan Hengsberger

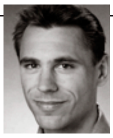

University of Applied Science
of Fribourg
Fribourg, Switzerland
stefan.hengsberger@eif.ch
http://www.anwalt-homburg.de/Stefan-index.htm

Stefan Hengsberger received his diploma in physics from the University of Saarbrücken in 1997. He started as a research scientist at Fraunhofer/Miami (Florida) where he worked until mai 1998. In July 1998 he joined the swiss federal institute of Technology (EPFL) where he earned his PhD in biomechanics in spring 2002. He stayed for another year at EPFL as a postdoc in biomechanics and physics until summer 2003. Since October 2003 he is Professor of Physics at the University of Applied Science in Fribourg, Switzerland.

Peter Hinterdorfer

Johannes Kepler University of Linz
Institute for Biophysics
Linz, Austria
peter.hinterdorfer@jku.at
http://at22.bphys.uni-linz.ac.at/bioph/staf/hipe/hipe.htm

Peter Hinterdorfer earned a Dr. tech. from the University of Linz, Austria, Institute for Biophysics in 1992. He was a postdoctoral fellow at the University of Virginia, Department of Molecular Physiology and Biological Physics (1992/1993). Since then he is at the University of Linz, Institute for Biophysics, where he holds a position as Associate Professor. His current and ongoing research includes single molecule force spectroscopy and high resolution topography and recognition imaging of biological samples.

Roberto Horowitz

University of California at Berkeley
Department
of Mechanical Engineering
Berkeley, CA, USA
horowitz@me.berkeley.edu
http://www.me.berkeley.edu/~horowitz

Roberto Horowitz joined the Department of Mechanical Engineering at the University of California at Berkeley in 1982, where he is currently a Professor. Dr. Horowitz teaches and conducts research in the areas of adaptive, learning, nonlinear and optimal control, with applications to Micro-Electromechanical Systems (MEMS), computer disk file systems, robotics, mechatronics and intelligent vehicle and highway systems (IVHS).

Hirotaka Hosoi

Japan Science and Technology
Corporation
Sapporo, Japan
hosoi@sapporo.jst-plaza.jp

Hirotaka Hosoi received the D.E. degree in electronic enginnering from Hokkaido University in 1999. Since 2002 he is at Innovation Plaza Hokkaido, Japan Science and Technology Corporation (JST). His main research focus is in high-resolution magnetic imaging of magnetic materials surfaces using a scanning force microscope. His current research interests includes magnetism on metal-oxide surfaces.

Jacob N. Israelachvili

University of California
Department of Chemical Engineering and
Materials Department
Santa Barbara, CA, USA
Jacob@engineering.ucsb.edu
http://www.chemengr.ucsb.edu/people/faculty/israelachvili

Jacob Israelachvili earned his Ph.D. 1971 at the Cavendish Laboratory, University of Cambridge, UK. He held various positions at the Department of Applied Mathematics at the Australian National University (1974–1986), including those of Professional Fellow and Head of Department. In 1986 he joined the faculty at University of California, Santa Barbara as Professor in the Department of Chemical Engineering and Materials Department. In 1988 he was elected a Fellow of the Royal Society of London, and in 1991 he was awarded the Alpha Chi Sigma Award for Chemical Engineering Research by the AIChE. He was elected as a foreign associate of the US National Academy of Engineering in 1996.

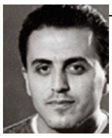

Ghassan E. Jabbour

University of Arizona
Optical Sciences Center
Tucson, AZ, USA
gej@optics.arizona.edu

Ghassan E. Jabbour is the head of the research group Organic Optoelectronic Materials and Devices at the Optical Sciences Center working on organic and hybrid materials and their applications to light emitting dvices, solar cells, memory storage, solid state lighting, and other areas. Professor Jabbour is a SPIE Fellow, Track Chair of the Nanotechnology Program for SPIE, and Associate Editor of the Journal of the Society for Information Display (JSID). He has over 200 publications, invited talks, and conference proceedings.

Harold Kahn

Case Western Reserve University
Department of Materials Science
and Engineering
Cleveland, OH, USA
kahn@cwru.edu

Harold Kahn is Researcher Associate Professor of Materials Science and Engineering at Case Western Reserve University, Cleveland, Ohio. His research is focused on MEMS device processing and testing, particularly wafer-level mechanical testing and shape-memory actuated microfluidics. He received a B.S. in metallurgical engineering from Lafayette College and a Ph.D. in electronic materials from the Massachusetts Institute of Technology.

András Kis

Swiss Federal Institute
of Technology (EPFL)
Institute of Physics
of Complex Matter
Lausanne, Switzerland
andras@igahpse.epfl.ch
http://www.andras.kis.name

András Kis is a Ph.D. student of Prof. László Forró at the Swiss Federal Institute of Technology (EPFL) where he works together with Andrzej Kulik. His main subject is the measurement of the mechanical properties of nanoscale objects, including carbon nanotubes and microtubules with the aim of understanding the interplay between their nanoscale structure and mechanical properties.

Jané Kondev

Brandeis University
Physics Department
Waltham, MA, USA
kondev@brandeis.edu
http://mattter.cc.brandeis.edu

Professor Kondev graduated with a PhD from Cornell University and did postdoctoral work at Brown University and Princeton University before joining the faculty at Brandeis University. He is a recipient of the CAREER award from the National Science Foundation and is a Cotrell Scholar of the Research Corporation.

Andrzej J. Kulik

Swiss Federal Institute
of Technology (EPFL)
Institute of Physics
of Complex Matter
Lausanne, Switzerland
andrzej.kulik@epfl.ch

Andrzej Kulik is the Head of the Biostructures and Nanomechanics Laboratory of the Swiss Federal Institute of Technology (EPFL) in Lausanne, Switzerland. His research concentrates on quantitative nanoscale materials properties, nanotribology. scanning probe microscopy, contact mechanics, nanoindentation, near-field ultrasonics, optical tweezers, nanolithography, and nanomanipulation.

Christophe Laurent

Université Paul Sabatier
CIRIMAT (Centre Interuniversitaire
de Recherche et d'Ingénierie
des Matériaux)
Toulouse, France
laurent@chimie.ups-tlse.fr
http://ncn.f2g.net/

Dr. Ch. Laurent is Professor of Materials Chemistry at University Paul Sabatier and is the head of the Nanocomposites and Carbon Nanotubes group of CIRIMAT. His research include in the synthesis, characterization and mechanical properties of ceramic-matrix nanocomposites, and since 1994, carbon nanotubes (synthesis of single- and double-walled CNTs, formation mechanisms, characterization, localized growth, hydrogen storage, ceramic-matrix composites).

Stephen C. Lee

Ohio State University
Biomedical Engineering Center
Columbus, OH, USA
Lee@bme.ohio-state.edu
http://medicine.osu.edu/mcbiochem/

Stephen C. Lee is a pioneer in the field of semi-biological nanodevices (nanobiological devices), having published the first monograph devoted to the topic in 1998. His interests are in enabling technologies for the incorporation of functional proteins and nucleic acids into nanodevices, particularly for application in oncology and cardiovascular disease. He is currently Associate Professor of Cellular and Molecular Biochemistry, Chemical Engineering and Biomedical Engineering at the Ohio State University.

Yunfeng Li

University of California at Berkeley
Department
of Mechanical Engineering
Berkeley, CA, USA
yunfeng@me.Berkeley.edu

Yunfeng Li received the B.S. and M.S. degrees from Beijing University of Aeronautics and Astronautics, Beijing, China, in 1992 and 1995, respectively. He is currently working towards the Ph.D. degree in the Department of Mechanical Engineering, University of California, Berkeley, CA, USA. His research interests include motion control, vibration control, control of MEMS devices, and mechatronics.

Liwei Lin

University of California at Berkeley
Mechanical Engineering Department
Berkeley, CA, USA
lwlin@me.berkeley.edu
http:// me.berkeley.edu/faculty/lin/index.html

Liwei Lin received his Ph.D. from the University of California, Berkeley, in 1993. He was an Associate Professor in the Institute of Applied Mechanics, National Taiwan University, Taiwan (1994–1996) and an Assistant Professor in Mechanical Engineering Department, University of Michigan (1996–1999). He joined UC-Berkeley in 1999 and is now an Associate Professor at Mechanical Engineering Department and Co-Director at Berkeley Sensor and Actuator Center.

Yu-Ming Lin

Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
Cambridge, MA, USA
yming@mgm.mit.edu

Lin Yu-Ming performed experimental and theoretical studies on Bi-based nanowires for next-generation thermoelectric materials. The electronic transport properties of these systems are studied to investigate quantum size effects. He received the Masterworks Award of the Department of EECS, MIT (2000) and the Gold Medal, MRS Graduate Student Award (2002)

Huiwen Liu

Ohio State University
Nanotribology Laboratory for Information Storage and MEMS/NEMS
Columbus, OH, USA
liu.403@osu.edu

Dr. Huiwen Liu is the associate director of the Nanotribology Laboratory for Information Storage and MEMS/NEMS at The Ohio State University. His research interests are study of mechanical, tribological, and physical properties of advanced materials and MEMS/NEMS devices on micro/nano scale. He obtained an Alexander von Humboldt fellowship and the Japanese Science and Technology Agency Fellowship in 1997 and 1998, respectively

Adrian B. Mann

Rutgers University
Department of Ceramics and Materials Engineering
Piscataway, NJ, USA
abmann@rci.rutgers.edu

Dr. Mann's research focuses on the nanomechanics of materials and the fabrication of nanostructured materials. His research is predominantly on biomedical materials, but also includes ceramics, polymers and metals. He is currently an Assistant Professor at Rutgers University, New Jersey. Prior to this he was a lecturer at the University of Manchester, England and a Fulbright scholar at The Johns Hopkins University, Maryland.

Othmar Marti

University of Ulm
Department of Experimental Physics
Ulm, Germany
Othmar.Marti@physik.uni-ulm.de
http://servex.physik.uni-ulm.de/marti

Profesor Othmar Marti is Head of the Department of Experimental Physics at the University of Ulm. His main research topics are polymers, scanning force microscopy, friction, near-field optics, and the optics of nanoparticles. Studies: diploma from ETH Zürich, Dr. sc. nat. ETH Zürich, Habilitation University Konstanz. He worked at IBM Research Zurich, Switzerland, University of California, USA, ETH Zurich, Switzerland and University of Konstanz, Germany.

Jack Martin

Analog Devices, Inc.
Micromachined Products Division
Cambridge, MA, USA
jack.martin@analog.com

Jack Martin has been a technologist and manager in the design, development and manufacture of industrial MEMS products for 25 years. His accomplishments include development of wafer fab and packaging processes that are used by Analog Devices to produce iMEMS integrated accelerometers and gyroscopes. He has a Ph.D. in Materials Science, a BS and MS in Chemical Engineering, and is a Licensed Professional Engineer.

Brendan McCarthy

University of Arizona
Optical Sciences Center
Tucson, AZ, USA
bmccarthy@optics.arizona.edu

Brendan Mc Carthy is a postdoctoral researcher in the Optical Sciences Center, University of Arizona. He received his PhD in 2001, from Trinity College, University of Dublin, Ireland. His thesis topic was the characterisation of composite materials based on carbon nanotubes and conjugated polymers, using optical and vibrational spectroscopy, scanning tunnelling microscopy, and transmission electron microscopy.

Mehran Mehregany

Case Western Reserve University
Department of Electrical Engineering and Computer Science
Cleveland, OH, USA
mxm31@cwru.edu
http://mems.cwru.edu/

Mehran Mehregany is currently the Chairman of the Electrical Engineering and Computer Science Department at Case and the Goodrich Professor of Engineering Innovation. His research interests include: silicon and silicon carbide micro/nano systems technology (including MEMS/NEMS); micromachining and microfabrication technologies; materials and modeling issues related to MEMS/NEMS and (in some cases) integrated circuits technologies; and MEMS packaging.

Ernst Meyer

University of Basel
Institute of Physics
Basel, Switzerland
Ernst.Meyer@unibas.ch
http://monet.unibas.ch/cgi-bin/people/meyer/

Ernst Meyer is professor of physics at the University of Basel. He is interested in friction force and dynamic force microscopy with true atomic resolution. He is also active in the field of sensors based upon micromechanics and magnetic spin resonance detection with force microscopy. Awarded from the Swiss Physical Society, he is member of the Swiss and American Physical Society, of the Editorial Board of Tribology Letters, and co-editor of books on atomic force microscopy.

Marc Monthioux

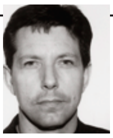

UPR A-8011 CNRS
Centre d'Elaboration des Matériaux et d'Etudes Structurales (CEMES)
Toulouse, France
monthiou@cemes.fr

Marc Monthioux has been working on carbon materials for more than 20 years. He is involved in research on carbon nanotubes since 1998, discovered the ability of single-wall nanotubes in being filled by foreign molecules the same year, associated with B. Smith and Prof. D.E. Luzzi from University of Pennsylvania. He is currently Director of Research at the French National Center for Scientific Research and European Associate Editor of CARBON Journal.

Markus Morgenstern

University of Hamburg
Institute of Applied Physics
Hamburg, Germany
mmorgens@physnet.uni-hamburg.de
http://www.nanoscience.de/group_r/stm-sts/

Markus Morgenstern earned his Ph.D. from the Institute of Interface Research and Vacuum Physics of the Forschungszentrum Jülich, Germany in 1996. After one year of research at the University of Paris VII he joined the Group of Prof. Dr. R. Wiesendanger in 1997 as a Senior Scientist. In 2002 he completed his Habilitation at the University of Hamburg with the subject Scanning tunnelling spectroscopy on semiconductor systems and nanostructures

Seizo Morita

Osaka University
Department of Electronic Engineering
Suita-Citiy, Osaka, Japan
smorita@ele.eng.osaka-u.ac.jp
http://www-e2.ele.eng.osaka-u.ac.jp/

Prof. Seizo Morita works in the atomic force microscopy (AFM). He has discovered two-dimensional friction with a lattice periodicity, and two-dimensional solid phase of densely contact-electrified electrons on SiO_2 thin films under ambient conditions. He has achieved mapping, discrimination and control of atomic force and atom with atomic resolution, and also atom manipulation based on a mechanical method using the noncontact AFM apparatus.

Koichi Mukasa

Hokkaido University
Nanoelectronics Laboratory
Sapporo, Japan
mukasa@nano.eng.hokudai.ac.jp
http://www.nano.eng.hokudai.ac.jp/nano/

K. Mukasa is a Professor of electronics at Hokkaido University, Sapporo, Japan. In 1980 he joined Alps Electric Co.Ltd, where he worked on the magnetic thin film heads and materials. In 1987 he moved to the university. His research interests include spin-polarized STM, exchange force microscopy, magnetic force microscopy, Mott spin detectors, nanostructure concerning electron spin and molecular/biological materials and devices.

Martin H. Müser

University of Western Ontario
Department of Applied Mathematics
London, Ontario, Canada
mmuser@uwo.ca
http://publish.uwo.ca/~mmuser/

Prof. Martin Müser received his Ph.D. (1995) and habilitation (2002) from the University of Mainz, Germany. He was a post doctoral researcher at Columbia University (1996/97) as Feodor Lynen fellow and at Johns Hopkins University (1998). He joined the faculty of the University of Western Ontario (U.W.O.) in 2002. His research interests are statistical physics with emphasis on computational materials science, tribology, and quantum solids.

Kenn Oldham

University of California at Berkeley
Department
of Mechanical Engineering
Berkeley, CA, USA
oldham@newton.berkeley.edu

Kenn Oldham is a graduate student in Mechanical Engineering at the University of California at Berkeley, pursuing M.S. and Ph.D. through National Science Foundation Graduate Fellowship. He received his B.S. from Carnegie Mellon University, Pittsburgh, PA. He is currently researching microdevices for hard disk drives. Research interests include MEMS design for control and reliability, optimal and robust control, and materials for microdevices.

Hiroshi Onishi

Kanagawa Academy
of Science and Technology
Surface Chemistry Laboratory
Kanagawa, Japan
oni@net.ksp.or.jp
http://home.ksp.or.jp/onishipro/home.eng/

Dr. Onishi Hiroshi is an experimental chemist at the Kanagawa Academy of Science and Technology interested in molecule-scale reaction kinetics at interfaces. He likes to observe molecules moving and reacting over metal oxide surfaces by time-lapse imaging with scanning probe microscopes. Domestic societies encouraged him with awards to further develop his research towards nano-scale chemistry.

René M. Overney

University of Washington
Department of Chemical Engineering
Seattle, WA, USA
roverney@u.Washington.edu
http://depts.washington.edu/nanolab/

René Overney received his Ph.D. in Physics at the University of Basel in 1992. After his postdoctoral years in Japan and at Exxon, CR, Annandale (NJ) he joined the University of Washington in 1996. His research interests are in mesoscale sciences involving nanorheological interpretations of processes that are as diverse as lubrication, membrane transport, or quantum yield in optoelectronic devices.

Alain Peigney

Université Paul Sabatier
CIRIMAT (Centre Inter-universitaire de
Recherches et d'Ingénierie
des Matériaux) – UMR CNRS 5085
Toulouse, France
peigney@chimie.ups-tlse.fr
http://ncn.f2g.net/

Ceramic Engineer and Doctor in Physical-Chemistry – Associate Professor of Materials Chemistry at the Paul Sabatier University – Researches in the synthesis, sintering and microstructural characterization of ceramics and ceramic matrix nanocomposites, and from 1994 on the synthesis on single- and double-walled carbon nanotubes and preparation of nanocomposites containing carbon nanotubes.

Oliver Pfeiffer

University of Basel
Institute of Physics
Basel, Switzerland
Oliver.Pfeiffer@stud.unibas.ch

Oliver Pfeiffer is PhD student at the University of Basel. The main topic of his work is the energy dissipation of oscillating cantilevers in non-contact AFM. Related to this research field is the examination of damping of torsional oscillations of cantilevers when approaching the sample.

Rob Phillips

California Institute of Technology
Mechanical Engineering
and Applied Physics
Pasadena, CA, USA
phillips@aero.caltech.edu
http://www.rpgroup.caltech.edu/people/phillips.html

Rob Phillips is Professor of Mechanical Engineering and Applied Physics at CALTECH, Pasadena. Phillips' work aims to examine nanomechanics of both crystalline solids and biological molecules and their assemblies. Recent efforts have been aimed at investigating mechanics of DNA packing in viruses, the tension-induced gating of ion channels and the mechanical response of bio-functionalized cantilevers.

Haralampos Pozidis

IBM Zurich Research Laboratory
Storage Technologies
Rüschlikon, Switzerland
hap@zurich.ibm.com

He received the Ph.D. in electrical engineering from Drexel University, Philadelphia, PA, in 1998. After working with Philips Research, Eindhoven, the Netherlands, on signal processing and coding for optical storage technologies, with focus on DVD and Blue-Ray-Disc, Dr. Pozidis joined IBM Research in 2001. His research focuses on receiver design for alternative storage technologies, particularly scanning probe microscopy-based techniques.

Prashant K. Purohit

California Institute of Technology
Mechanical Engineering
Pasadena, CA, USA
prashant@caltech.edu
http://www.rpgroup.caltech.edu/~prashant/

Purohit Prashant is currently a Postdoctoral Scholar in Applied Mechanics in the Mechanical Engineering department at Caltech. The overall theme of his research is to develop systematic methods to understand the physics operative at the nanometer scales. His research falls broadly into two categories. The first concerns the subject of coarse-graining methods for crystalline solids and the second mechanics problems in biology. His group hopes to extend the coarse-graining methods in solids to the arena of biological macromolecules.

Oded Rabin

Massachusetts Institute
of Technology
Department of Chemistry
Cambridge, MA, USA
oded@mgm.mit.edu

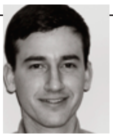

Oded Rabin is a Ph.D. candidate in Prof. Mildred S. Dresselhaus' group working on the thermoelectric properties of bismuth-antimony nanowire systems, and on electrochemistry-based nanowire synthesis methods. He earned a B.A. degree in Chemistry from the Technion – Israel Institute of Technology, Haifa, Israel and an M.A. degree in Chemistry from the Weizmann Institute of Science, Rehovot, Israel.

Françisco M. Raymo

University of Miami
Department of Chemistry
Coral Gables, FL, USA
fraymo@miami.edu
http://www.as.miami.edu/chemistry/FMRaymo.html

Françisco M. Raymo is Assistant Professor in the Department of Chemistry at the University of Miami. His research interests lie at the interface of chemistry and materials science. In particular, he is exploring innovative strategies to process optical signals with molecular switches, design fluorescent probes for chemical sensing and assemble nanostructured films from electroactive building blocks.

Manitra Razafinimanana

Université Paul Sabatier
Centre de Physique des Plasmas
et leurs Applications (CPPAT)
Toulouse, France
razafinimanana@cpat.ups-tlse.fr
http://cpat.ups-tlse.fr

Manitra Razafinimanana was born in Analalava, Madagascar, on June 28, 1951. He received the 3rd cycle degree, and Doctorat d'Etat ès Sciences Physiques degree from the Université Paul Sabatier, Toulouse, France, in 1982 and 1986 respectively. Since 2000, he was held the position of Professor. He has worked on plasma diagnostics, arc-electrode interaction, transport coefficients and thermodynamical properties calculation.

Mark O. Robbins

Johns Hopkins University
Department of Physics
and Astronomy
Baltimore, MD, USA
mr@jhu.edu
http://www.pha.jhu.edu/~mr

Mark Robbins is a Professor in the Department of Physics and Astronomy of the Johns Hopkins University. His research focuses on non-equilibrium processes like friction, adhesion, spreading, fluid invasion, and shear-induced phase transitions. A common goal is to understand the atomic origins of macroscopic behaviour. He is a fellow of the American Physical Society and was a Sloan Fellow and Presidential Young Investigator.

John A. Rogers

University of Illinois
Department of Materials Science and
Engineering
Urbana, IL, USA
jrogers@uiuc.edu
http://www.mse.uiuc.edu/faculty/
Rogers.html

John A. Rogers earned the Ph.D. degree in physical chemistry from M.I. T. in 1995. Until 2002, he was at Bell Labs serving first as a Member of Technical Staff and subsequently as research Director. John is currently Founder Professor of Engineering at University of Illinois at Urbana/ Champaign. His research interests include methods for micro/ nanofabrication, plastic and molecular electronics, photonics

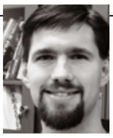

Mark Ruegsegger

Ohio State University
Biomedical Engineering Center
Columbus, OH, USA
mark@bme.ohio-state.edu

Mark Ruegsegger received his PhD in biomedical engineering at Case Western Reserve Unviersity and is currently an Assistant Professor of Biomedical Engineering at The Ohio State University. His research focus is the development of superparamagnetic particles that can target specific cell or tissue types, in particular atherosclerotic plaque. Other projects include biomimetic surface coatings on biomaterials and characterization of biomolecular flow in nanochannels.

Marina Ruths

Åbo Akademi University
Department of Physical Chemistry
Åbo, Finland
mruths@abo.fi

Marina Ruths received her Ph.D. from the University of California, Santa Barbara in 1996 followed by postdoctoral research at the University of Illinois at Urbana-Champaign, at the Max-Planck-Institute for Polymer Research, an at Åbo Akademi University. Her current research includes adhesion, friction and nanorheology of surfactant, polymer, and liquid crystal systems. She received an ASLA-Fulbright grant in 1991 and an Alexander von Humboldt fellowship in 1998.

Dror Sarid

University of Arizona
Optical Sciences Center
Tucson, AZ, USA
sarid@optics.arizona.edu
http://www.optics.arizona.edu/spm

Dror Sarid is Professor of Optical Sciences and Director of the Optical Data Storage Center. He is conducting research in scanning tunneling microcopy, atomic force microscopy and related systems and in particular probe storage, nano-optics and nano-technology. He published three books and more than 150 papers, mostly on optics and nanotechnology related topics.

Akira Sasahara

Kanagawa Academy
of Science and Technology
Surface Chemistry Laboratory
Kanagawa, Japan
ryo@net.ksp.or.jp
http://home.ksp.or.jp/onishipro/home.
eng/

Dr. Sasahara Akira is interested in local structures formed on solid surfaces. His current research focuses on elucidation of chemical and physical properties of nano-scale structures on metal oxide surfaces and their effect on chemical reaction.

André Schirmeisen

University of Münster
Institute of Physics
Münster, Germany
schira@uni-muenster.de
http://www.andre-schirmeisen.de

Dr. André Schirmeisen is currently working with his research group on nanoscale mechanical phenomena at Münster University. First to combine field ion microscopy with force microscopy to investigate atomically defined nanocontacts. He spent several years in Canada at McGill University to earn his PhD degree in physics. Before he worked as a strategic business consultant and is always interested in connecting nanoscience research with business applications.

Alexander Schwarz

University of Hamburg
Institute of Applied Physics
Hamburg, Germany
aschwarz@physnet.uni-hamburg.de
http://www.nanoscience.de/group_r/

Dr. Alexander Schwarz belongs to the scientific staff of the Center of Microstructure Research at the Institute of Applied Physics at the University of Hamburg, Germany. The group has 10 years (since 1993) experience in scientific research in the field of force microscopy and spectroscopy at cryogenic temperatures in ultrahigh vacuum and high magnetic fields. He is a Senior Scientist and works on Magnetic Force Microscopy (MFM) at low temperatures

Udo D. Schwarz

Yale University
Department
of Mechanical Engineering
New Haven, CT, USA
udo.schwarz@yale.edu
http://www.eng.yale.edu/
nanomechanics/

Udo D. Schwarz received his Ph.D. from the University of Basel in 1993 already using scanning force microscopy. Subsequently he moved to the University of Hamburg where he specialised on low-temperature scanning force microscopy and nanotribology. After spending a year at the Lawrence Berkeley National Laboratory he accepted a position as Associate Professor of Mechanical Engineering at Yale University in 2002.

Philippe Serp

Ecole Nationale Supèrieure d'Ingénieurs
en Arts Chimiques
et Technologiques
Laboratoire de Catalyse,
Chimie Fine et Polymères
Toulouse, France
Philippe.Serp@ensiacet.fr
http://www.ensiacet.fr/

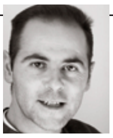

Philippe Serp is associate professor in the LCCFP at ENSIACET. After receiving is PhD from Paul Sabatier University, Toulouse, in 1994 where he worked on the preparation of supported catalyst, he moved to Universidade do Porto to carry out post-doctoral research on catalytic CVD to prepare carbon fibers. His current research interests include CVD preparation of nanostructured materials and catalysis.

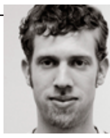

Bryan R. Smith

Ohio State University
Biomedical Engineering Center
Columbus, OH, USA
bryan@bme.ohio-state.edu

Bryan Smith is a Ph.D. candidate in biomedical engineering at The Ohio State University, specializing in the microfabrication and bioconjugation of micro- and nanoparticles for proteomics as well as imaging and drug delivery in breast cancer and atherosclerosis.

Anisoara Socoliuc

University of Basel
Institute of Physics
Basel, Switzerland
A.Socoliuc@unibas.ch
http://monet.physik.unibas.ch/gue/
uhvafm/

Anisoara Socoliuc studied physics in Iasi, Romania. She is now PhD student at the University of Basel. The main topic of her work is the study of wear processes on the nanometer scale by friction force microscopy.

Yasuhiro Sugawara

Osaka University
Department of Applied Physics
Suita, Japan
sugawara@ap.eng.osaka-u.ac.jp

Yasuhiro Sugawara received his Ph.D. in 1988 from Tohoku University and is Professor in the Department of Applied Physics of the Graduate School of Engineering at Osaka University since 2002. His research focuses on the further development of scanning probe microscopes and their applications, especially the noncontact atomic force microscope for the observation of solid surfaces at the atomic and molecular level. His aim is also to develop new nanomaterials and nanodevices by manipulation of single atoms and molecules using the atomic force microscope.

George W. Tyndall

IBM Almaden Research Center
Science and Technology
San Jose, CA, USA
tyndallgw@netscape.net

Dr. George Tyndall's current research focuses on the friction and adhesion in boundary lubricants especially as it pertains to the tribology of the magnetic recording head-disk interface in hard-disk drive applications.

Peter Vettiger

IBM Zurich Research Laboratory
Manager Micro-/Nanomechanics
Rüschlikon, Switzerland
pv@zurich.ibm.com

In 1963 Peter Vettiger joined IBM Research (Rüschlikon). He established and headed micro/nanoscale fabrication activities for superconducting, electronic and opto-electronic devices. Together with G.K. Binnig, he initiated Millipede probe-storage activities in 1995. His research interests are micro/nanomechanical devices and systems for probe-storage and biological applications. He is an IEEE Fellow and received a Doctor honoris causa from the University of Basel, Switzerland.

Darrin J. Young

Case Western Reserve University
Electrical Engineering
and Computer Science
Cleveland, OH, USA
djy@po.cwru.edu
http://home.cwru.edu/~djy

Darrin J. Young received his BS, MS, and PhD degrees from the EECS Department at University of California at Berkeley in 1991, 1993, and 1999, respectively. He pioneered RF MEMS high-Q tunable passive devices for wireless communications. He joined the EECS Department at Case Western Reserve University as an assistant professor in 1999. His main research interests include MEMS device design and fabrication.

Babak Ziaie

University of Minnesota
Department of Electrical
and Computer Engineering
Minneapolis, MN, USA
ziaie@ece.umn.edu

Dr. Babak Ziaie is an assistant professor in the Electrical and Computer Engineering Department of the University of Minnesota. His research interests are located at the boundaries between engineering and biology. He collaborates with physicians and biologists of all kinds and his lab is actively involved in the design and fabrication of biomedical micro and nanoelectromechanical systems.

Christian A. Zorman

Case Western Reserve University
Department of Electrical Engineering
and Computer Science
Cleveland, OH, USA
caz@po.cwru.edu
http://mems.cwru.edu/

Christian A. Zorman received a B.S. cum laude in physics and a B.A. cum laude in economics from The Ohio State University in 1988. He received M.S. and Ph.D. degrees in physics from Case Western Reserve University (CWRU) in 1991 and 1994, respectively. Dr. Zorman served in several research capacities at CWRU between 1994 and 2002, before being appointed an Associate Professor in 2002. His research interests involve materials and processes for MEMS.

Philippe K. Zysset

Technische Universität Wien
Institut für Leichtbau
und Flugzeugbau (ILFB)
Wien, Austria
philippe.zysset@epfl.ch
http://ilfb.tuwien.ac.at/

Philippe Zysset graduated in physics in 1987 and obtained a Ph.D. in biomechanics in 1994 at Ecole Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, Switzerland. He completed his postdoctoral training at the Orthopaedic Research Laboratories of the University of Michigan and returned to EPFL as an assistant professor in 1997. His current research interests are in bone biomechanics.

Detailed Contents

List of Tables XXIX
List of Abbreviations XXXIII

1 **Introduction to Nanotechnology** 1
  1.1 Background and Definition of Nanotechnology 1
  1.2 Why Nano? 2
  1.3 Lessons from Nature 2
  1.4 Applications in Different Fields 3
  1.5 Reliability Issues of MEMS/NEMS 4
  1.6 Organization of the Handbook 5
  References 5

Part A Nanostructures, Micro/Nanofabrication, and Micro/Nanodevices

2 **Nanomaterials Synthesis and Applications: Molecule-Based Devices** 9
  2.1 Chemical Approaches to Nanostructured Materials 10
    2.1.1 From Molecular Building Blocks to Nanostructures 10
    2.1.2 Nanoscaled Biomolecules: Nucleic Acids and Proteins 10
    2.1.3 Chemical Synthesis of Artificial Nanostructures 12
    2.1.4 From Structural Control to Designed Properties and Functions ... 12
  2.2 Molecular Switches and Logic Gates 14
    2.2.1 From Macroscopic to Molecular Switches 15
    2.2.2 Digital Processing and Molecular Logic Gates 15
    2.2.3 Molecular AND, NOT, and OR Gates 16
    2.2.4 Combinational Logic at the Molecular Level 17
    2.2.5 Intermolecular Communication 18
  2.3 Solid State Devices 22
    2.3.1 From Functional Solutions to Electroactive and Photoactive Solids ... 22
    2.3.2 Langmuir–Blodgett Films 23
    2.3.3 Self-Assembled Monolayers 27
    2.3.4 Nanogaps and Nanowires 31
  2.4 Conclusions and Outlook 35
  References 36

3 **Introduction to Carbon Nanotubes** 39
  3.1 Structure of Carbon Nanotubes 40
    3.1.1 Single-Wall Nanotubes 40
    3.1.2 Multiwall Nanotubes 43

3.2 Synthesis of Carbon Nanotubes 45
  3.2.1 Solid Carbon Source-Based Production Techniques for Carbon Nanotubes 45
  3.2.2 Gaseous Carbon Source-Based Production Techniques for Carbon Nanotubes 52
  3.2.3 Miscellaneous Techniques 57
  3.2.4 Synthesis of Aligned Carbon Nanotubes 58
3.3 Growth Mechanisms of Carbon Nanotubes 59
  3.3.1 Catalyst-Free Growth 59
  3.3.2 Catalytically Activated Growth 60
3.4 Properties of Carbon Nanotubes 63
  3.4.1 Variability of Carbon Nanotube Properties 63
  3.4.2 General Properties 63
  3.4.3 SWNT Adsorption Properties 63
  3.4.4 Transport Properties 65
  3.4.5 Mechanical Properties 67
  3.4.6 Reactivity 67
3.5 Carbon Nanotube-Based Nano-Objects 68
  3.5.1 Hetero-Nanotubes 68
  3.5.2 Hybrid Carbon Nanotubes 68
  3.5.3 Functionalized Nanotubes 71
3.6 Applications of Carbon Nanotubes 73
  3.6.1 Current Applications 73
  3.6.2 Expected Applications Related to Adsorption 76
References 86

4 **Nanowires** 99
4.1 Synthesis 100
  4.1.1 Template-Assisted Synthesis 100
  4.1.2 VLS Method for Nanowire Synthesis 105
  4.1.3 Other Synthesis Methods 107
  4.1.4 Hierarchical Arrangement and Superstructures of Nanowires 108
4.2 Characterization and Physical Properties of Nanowires 110
  4.2.1 Structural Characterization 110
  4.2.2 Transport Properties 115
  4.2.3 Optical Properties 126
4.3 Applications 131
  4.3.1 Electrical Applications 131
  4.3.2 Thermoelectric Applications 133
  4.3.3 Optical Applications 134
  4.3.4 Chemical and Biochemical Sensing Devices 137
  4.3.5 Magnetic Applications 137
4.4 Concluding Remarks 138
References 138

5 **Introduction to Micro/Nanofabrication** 147
5.1 Basic Microfabrication Techniques 148
5.1.1 Lithography 148
5.1.2 Thin Film Deposition and Doping 149
5.1.3 Etching and Substrate Removal 153
5.1.4 Substrate Bonding 157
5.2 MEMS Fabrication Techniques 159
5.2.1 Bulk Micromachining 159
5.2.2 Surface Micromachining 163
5.2.3 High-Aspect-Ratio Micromachining 166
5.3 Nanofabrication Techniques 170
5.3.1 E-Beam and Nano-Imprint Fabrication 171
5.3.2 Epitaxy and Strain Engineering 172
5.3.3 Scanned Probe Techniques 173
5.3.4 Self-Assembly and Template Manufacturing 176
References 180

6 **Stamping Techniques for Micro and Nanofabrication: Methods and Applications** 185
6.1 High Resolution Stamps 186
6.2 Microcontact Printing 187
6.3 Nanotransfer Printing 190
6.4 Applications 193
6.4.1 Unconventional Electronic Systems 193
6.4.2 Lasers and Waveguide Structures 198
6.5 Conclusions 200
References 200

7 **Materials Aspects of Micro- and Nanoelectromechanical Systems** . 203
7.1 Silicon 203
7.1.1 Single Crystal Silicon 204
7.1.2 Polysilicon 205
7.1.3 Porous Silicon 208
7.1.4 Silicon Dioxide 208
7.1.5 Silicon Nitride 209
7.2 Germanium-Based Materials 210
7.2.1 Polycrystalline Ge 210
7.2.2 Polycrystalline SiGe 210
7.3 Metals 211
7.4 Harsh Environment Semiconductors 212
7.4.1 Silicon Carbide 212
7.4.2 Diamond 215
7.5 GaAs, InP, and Related III-V Materials 217
7.6 Ferroelectric Materials 218
7.7 Polymer Materials 219
7.7.1 Polyimide 219
7.7.2 SU-8 220

7.7.3 Parylene 220
7.8 Future Trends 220
References 221

8 **MEMS/NEMS Devices and Applications** 225
8.1 MEMS Devices and Applications 227
8.1.1 Pressure Sensor 227
8.1.2 Inertial Sensor 229
8.1.3 Optical MEMS 233
8.1.4 RF MEMS 239
8.2 NEMS Devices and Applications 246
8.3 Current Challenges and Future Trends 249
References 250

9 **Microfluidics and Their Applications to Lab-on-a-Chip** 253
9.1 Materials for Microfluidic Devices and Micro/Nano Fabrication Techniques 254
9.1.1 Silicon 254
9.1.2 Glass 254
9.1.3 Polymer 255
9.2 Active Microfluidic Devices 257
9.2.1 Microvalves 258
9.2.2 Micropumps 260
9.3 Smart Passive Microfluidic Devices 262
9.3.1 Passive Microvalves 262
9.3.2 Passive Micromixers 265
9.3.3 Passive Microdispensers 266
9.3.4 Microfluidic Multiplexer Integrated with Passive Microdispenser 267
9.3.5 Passive Micropumps 269
9.3.6 Advantages and Disadvantages of the Passive Microfluidic Approach 269
9.4 Lab-on-a-Chip for Biochemical Analysis 270
9.4.1 Magnetic Micro/Nano Bead-Based Biochemical Detection System 270
9.4.2 Disposable Smart Lab-on-a-Chip for Blood Analysis 273
References 276

10 **Therapeutic Nanodevices** 279
10.1 Definitions and Scope of Discussion 280
10.1.1 Design Issues 281
10.1.2 Utility and Scope of Therapeutic Nanodevices 285
10.2 Synthetic Approaches: “top-down” versus “bottom-up” Approaches for Nanotherapeutic Device Components 285
10.2.1 Production of Nanoporous Membranes by Microfabrication Methods: A top-down Approach 285

10.2.2 Synthesis of Poly(amido) Amine (PAMAM) Dendrimers: A bottom-up Approach 286
10.2.3 The Limits of top-down and bottom-up Distinctions with Respect to Nanomaterials and Nanodevices 287
10.3 Technological and Biological Opportunities 288
10.3.1 Assembly Approaches 288
10.3.2 Targeting: Delimiting Nanotherapeutic Action in Three-Dimensional Space 296
10.3.3 Triggering: Delimiting Nanotherapeutic Action in Space and Time 298
10.3.4 Sensing Modalities 302
10.3.5 Imaging Using Nanotherapeutic Contrast Agents 304
10.4 Applications for Nanotherapeutic Devices 307
10.4.1 Nanotherapeutic Devices in Oncology 307
10.4.2 Cardiovascular Applications of Nanotherapeutics 310
10.4.3 Nanotherapeutics and Specific Host Immune Responses 311
10.5 Concluding Remarks: Barriers to Practice and Prospects 315
10.5.1 Complexity in Biology 315
10.5.2 Dissemination of Biological Information 315
10.5.3 Cultural Differences Between Technologists and Biologists 316
References 317

Part B Scanning Probe Microscopy

11 **Scanning Probe Microscopy – Principle of Operation, Instrumentation, and Probes** 325
11.1 Scanning Tunneling Microscope 327
11.1.1 Binnig et al.'s Design 327
11.1.2 Commercial STMs 328
11.1.3 STM Probe Construction 330
11.2 Atomic Force Microscope 331
11.2.1 Binnig et al.'s Design 333
11.2.2 Commercial AFM 333
11.2.3 AFM Probe Construction 338
11.2.4 Friction Measurement Methods 342
11.2.5 Normal Force and Friction Force Calibrations of Cantilever Beams 346
11.3 AFM Instrumentation and Analyses 347
11.3.1 The Mechanics of Cantilevers 347
11.3.2 Instrumentation and Analyses of Detection Systems for Cantilever Deflections 350
11.3.3 Combinations for 3-D-Force Measurements 358
11.3.4 Scanning and Control Systems 359
References 364

12 **Probes in Scanning Microscopies** 371
12.1 Atomic Force Microscopy 372
12.1.1 Principles of Operation 372
12.1.2 Standard Probe Tips 373
12.1.3 Probe Tip Performance 374
12.1.4 Oxide-Sharpened Tips 375
12.1.5 FIB tips 376
12.1.6 EBD tips 376
12.1.7 Carbon Nanotube Tips 376
12.2 Scanning Tunneling Microscopy 382
12.2.1 Mechanically Cut STM Tips 382
12.2.2 Electrochemically Etched STM Tips 383
References 383

13 **Noncontact Atomic Force Microscopy and Its Related Topics** 385
13.1 Principles of Noncontact Atomic Force Microscope (NC-AFM) 386
13.1.1 Imaging Signal in AFM 386
13.1.2 Experimental Measurement and Noise 387
13.1.3 Static AFM Operating Mode 387
13.1.4 Dynamic AFM Operating Mode 388
13.1.5 The Four Additional Challenges Faced by AFM 388
13.1.6 Frequency-Modulation AFM (FM-AFM) 389
13.1.7 Relation Between Frequency Shift and Forces 390
13.1.8 Noise in Frequency-Modulation AFM – Generic Calculation . 391
13.1.9 Conclusion 391
13.2 Applications to Semiconductors 391
13.2.1 Si(111)7×7 Surface 392
13.2.2 Si(100)2×1 and Si(100)2×1:H Monohydride Surfaces 393
13.2.3 Metal-Deposited Si Surface 395
13.3 Applications to Insulators 397
13.3.1 Alkali Halides, Fluorides, and Metal Oxides 397
13.3.2 Atomically Resolved Imaging of a NiO(001) Surface 402
13.3.3 Atomically Resolved Imaging Using Noncoated and Fe-Coated Si Tips 402
13.4 Applications to Molecules 404
13.4.1 Why Molecules and What Molecules? 404
13.4.2 Mechanism of Molecular Imaging 404
13.4.3 Perspectives 407
References 407

14 **Low Temperature Scanning Probe Microscopy** 413
14.1 Microscope Operation at Low Temperatures 414
14.1.1 Drift 414
14.1.2 Noise 415
14.1.3 Stability 415
14.1.4 Piezo Relaxation and Hysteresis 415
14.2 Instrumentation 415

14.2.1 A Simple Design for a Variable Temperature STM 416
14.2.2 A Low Temperature SFM Based on a Bath Cryostat 417
14.3 Scanning Tunneling Microscopy and Spectroscopy 419
14.3.1 Atomic Manipulation 419
14.3.2 Imaging Atomic Motion 420
14.3.3 Detecting Light from Single Atoms and Molecules 421
14.3.4 High Resolution Spectroscopy 422
14.3.5 Imaging Electronic Wave Functions 427
14.3.6 Imaging Spin Polarization: Nanomagnetism 431
14.4 Scanning Force Microscopy and Spectroscopy 433
14.4.1 Atomic-Scale Imaging 434
14.4.2 Force Spectroscopy 436
14.4.3 Electrostatic Force Microscopy 438
14.4.4 Magnetic Force Microscopy 439
References 442

15 **Dynamic Force Microscopy** 449
15.1 Motivation: Measurement of a Single Atomic Bond 450
15.2 Harmonic Oscillator: A Model System for Dynamic AFM 454
15.3 Dynamic AFM Operational Modes 455
15.3.1 Amplitude-Modulation/ Tapping-Mode AFMs 456
15.3.2 Self-Excitation Modes 461
15.4 *Q*-Control 464
15.5 Dissipation Processes Measured with Dynamic AFM 468
15.6 Conclusion 471
References 471

16 **Molecular Recognition Force Microscopy** 475
16.1 Ligand Tip Chemistry 476
16.2 Fixation of Receptors to Probe Surfaces 478
16.3 Single-Molecule Recognition Force Detection 479
16.4 Principles of Molecular Recognition Force Spectroscopy 482
16.5 Recognition Force Spectroscopy: From Isolated Molecules to Biological Membranes 484
16.5.1 Forces, Energies, and Kinetic Rates 484
16.5.2 Complex Bonds and Energy Landscapes 486
16.5.3 Live Cells and Membranes 489
16.6 Recognition Imaging 489
16.7 Concluding Remarks 491
References 492

Part C Nanotribology and Nanomechanics

17 **Micro/Nanotribology and Materials Characterization Studies Using Scanning Probe Microscopy** 497
17.1 Description of AFM/FFM and Various Measurement Techniques 499

17.1.1 Surface Roughness and Friction Force Measurements 500
17.1.2 Adhesion Measurements 502
17.1.3 Scratching, Wear and Fabrication/Machining 503
17.1.4 Surface Potential Measurements 503
17.1.5 In Situ Characterization of Local Deformation Studies 504
17.1.6 Nanoindentation Measurements 504
17.1.7 Localized Surface Elasticity and Viscoelasticity Mapping 505
17.1.8 Boundary Lubrication Measurements 507
17.2 Friction and Adhesion 507
17.2.1 Atomic-Scale Friction 507
17.2.2 Microscale Friction 507
17.2.3 Directionality Effect on Microfriction 511
17.2.4 Velocity Dependence on Microfriction 513
17.2.5 Effect of Tip Radii and Humidity on Adhesion and Friction 515
17.2.6 Scale Dependence on Friction 518
17.3 Scratching, Wear, Local Deformation, and Fabrication/Machining 518
17.3.1 Nanoscale Wear 518
17.3.2 Microscale Scratching 519
17.3.3 Microscale Wear 520
17.3.4 In Situ Characterization of Local Deformation 524
17.3.5 Nanofabrication/Nanomachining 526
17.4 Indentation 526
17.4.1 Picoindentation 526
17.4.2 Nanoscale Indentation 527
17.4.3 Localized Surface Elasticity and Viscoelasticity Mapping 528
17.5 Boundary Lubrication 530
17.5.1 Perfluoropolyether Lubricants 530
17.5.2 Self-Assembled Monolayers 536
17.5.3 Liquid Film Thickness Measurements 537
17.6 Closure 538
References 539

18 **Surface Forces and Nanorheology of Molecularly Thin Films** 543
18.1 Introduction: Types of Surface Forces 544
18.2 Methods Used to Study Surface Forces 546
18.2.1 Force Laws 546
18.2.2 Adhesion Forces 547
18.2.3 The SFA and AFM 547
18.2.4 Some Other Force-Measuring Techniques 549
18.3 Normal Forces Between Dry (Unlubricated) Surfaces 550
18.3.1 Van der Waals Forces in Vacuum and Inert Vapors 550
18.3.2 Charge Exchange Interactions 552
18.3.3 Sintering and Cold Welding 553
18.4 Normal Forces Between Surfaces in Liquids 554
18.4.1 Van der Waals Forces in Liquids 554
18.4.2 Electrostatic and Ion Correlation Forces 554
18.4.3 Solvation and Structural Forces 557

18.4.4 Hydration and Hydrophobic Forces 559
18.4.5 Polymer-Mediated Forces 561
18.4.6 Thermal Fluctuation Forces 563
18.5 Adhesion and Capillary Forces 564
18.5.1 Capillary Forces 564
18.5.2 Adhesion Mechanics 566
18.5.3 Effects of Surface Structure, Roughness, and Lattice Mismatch 566
18.5.4 Nonequilibrium and Rate-Dependent Interactions: Adhesion Hysteresis 567
18.6 Introduction: Different Modes of Friction and the Limits of Continuum Models 569
18.7 Relationship Between Adhesion and Friction Between Dry (Unlubricated and Solid Boundary Lubricated) Surfaces 571
18.7.1 Amontons' Law and Deviations from It Due to Adhesion: The Cobblestone Model 571
18.7.2 Adhesion Force and Load Contribution to Interfacial Friction 572
18.7.3 Examples of Experimentally Observed Friction of Dry Surfaces 576
18.7.4 Transition from Interfacial to Normal Friction with Wear 579
18.8 Liquid Lubricated Surfaces 580
18.8.1 Viscous Forces and Friction of Thick Films: Continuum Regime 580
18.8.2 Friction of Intermediate Thickness Films 582
18.8.3 Boundary Lubrication of Molecularly Thin Films: Nanorheology 584
18.9 Role of Molecular Shape and Surface Structure in Friction 591
References 594

19 **Scanning Probe Studies of Nanoscale Adhesion Between Solids in the Presence of Liquids and Monolayer Films** 605
19.1 The Importance of Adhesion at the Nanoscale 605
19.2 Techniques for Measuring Adhesion 606
19.3 Calibration of Forces, Displacements, and Tips 610
19.3.1 Force Calibration 610
19.3.2 Probe Tip Characterization 611
19.3.3 Displacement Calibration 612
19.4 The Effect of Liquid Capillaries on Adhesion 612
19.4.1 Theoretical Background 612
19.4.2 Experimental and Theoretical Studies of Capillary Formation with Scanning Probes 614
19.4.3 Future Directions 618
19.5 Self-Assembled Monolayers 618
19.5.1 Adhesion at SAM Interfaces 618
19.5.2 Chemical Force Microscopy: General Methodology 619
19.5.3 Adhesion at SAM-Modified Surfaces in Liquids 620
19.5.4 Impact of Intra- and Inter-Chain Interactions on Adhesion 621

19.5.5 Adhesion at the Single-Bond Level 622
19.5.6 Future Directions 623
19.6 Concluding Remarks 624
References 624

20 **Friction and Wear on the Atomic Scale** 631
20.1 Friction Force Microscopy in Ultra-High Vacuum 632
20.1.1 Friction Force Microscopy 632
20.1.2 Force Calibration 632
20.1.3 The Ultra-high Vacuum Environment 635
20.1.4 A Typical Microscope in UHV 635
20.2 The Tomlinson Model 636
20.2.1 One-dimensional Tomlinson Model 636
20.2.2 Two-dimensional Tomlinson Model 637
20.2.3 Friction Between Atomically Flat Surfaces 637
20.3 Friction Experiments on Atomic Scale 638
20.3.1 Anisotropy of Friction 642
20.4 Thermal Effects on Atomic Friction 642
20.4.1 The Tomlinson Model at Finite Temperature 642
20.4.2 Velocity Dependence of Friction 644
20.4.3 Temperature Dependence of Friction 645
20.5 Geometry Effects in Nanocontacts 646
20.5.1 Continuum Mechanics of Single Asperities 646
20.5.2 Load Dependence of Friction 647
20.5.3 Estimation of the Contact Area 647
20.6 Wear on the Atomic Scale 649
20.6.1 Abrasive Wear on the Atomic Scale 649
20.6.2 Wear Contribution to Friction 650
20.7 Molecular Dynamics Simulations of Atomic Friction and Wear 651
20.7.1 Molecular Dynamics Simulation of Friction Processes 651
20.7.2 Molecular Dynamics Simulations of Abrasive Wear 652
20.8 Energy Dissipation in Noncontact Atomic Force Microscopy 654
20.9 Conclusion 656
References 657

21 **Nanoscale Mechanical Properties – Measuring Techniques and Applications** 661
21.1 Local Mechanical Spectroscopy by Contact AFM 662
21.1.1 The Variable-Temperature SLAM (T-SLAM) 663
21.1.2 Example One: Local Mechanical Spectroscopy of Polymers ... 664
21.1.3 Example Two: Local Mechanical Spectroscopy of NiTi 665
21.2 Static Methods – Mesoscopic Samples 667
21.2.1 Carbon Nanotubes – Introduction to Basic Morphologies and Production Methods 667
21.2.2 Measurements of the Mechanical Properties of Carbon Nanotubes by SPM 668
21.2.3 Microtubules and Their Elastic Properties 673

21.3 Scanning Nanoindentation: An Application to Bone Tissue 674
21.3.1 Scanning Nanoindentation 674
21.3.2 Application of Scanning Nanoindentation 674
21.3.3 Example: Study of Mechanical Properties of Bone Lamellae Using SN 675
21.3.4 Conclusion 681
21.4 Conclusions and Perspectives 682
References 682

22 **Nanomechanical Properties of Solid Surfaces and Thin Films** 687
22.1 Instrumentation 688
22.1.1 AFM and Scanning Probe Microscopy 688
22.1.2 Nanoindentation 689
22.1.3 Adaptations of Nanoindentation 690
22.1.4 Complimentary Techniques 691
22.1.5 Bulge Tests 691
22.1.6 Acoustic Methods 692
22.1.7 Imaging Methods 693
22.2 Data Analysis 694
22.2.1 Elastic Contacts 694
22.2.2 Indentation of Ideal Plastic Materials 694
22.2.3 Adhesive Contacts 695
22.2.4 Indenter Geometry 696
22.2.5 Analyzing Load/Displacement Curves 696
22.2.6 Modifications to the Analysis 699
22.2.7 Alternative Methods of Analysis 700
22.2.8 Measuring Contact Stiffness 701
22.2.9 Measuring Viscoelasticity 702
22.3 Modes of Deformation 702
22.3.1 Defect Nucleation 702
22.3.2 Variations with Depth 704
22.3.3 Anisotropic Materials 704
22.3.4 Fracture and Delamination 704
22.3.5 Phase Transformations 705
22.4 Thin Films and Multilayers 707
22.4.1 Thin Films 707
22.4.2 Multilayers 709
22.5 Developing Areas 711
References 712

23 **Atomistic Computer Simulations of Nanotribology** 717
23.1 Molecular Dynamics 718
23.1.1 Model Potentials 719
23.1.2 Maintaining Constant Temperature 720
23.1.3 Imposing Load and Shear 721
23.1.4 The Time-Scale and Length-Scale Gaps 721
23.1.5 A Summary of Possible Traps 722

23.2 Friction Mechanisms at the Atomic Scale 723
  23.2.1 Geometric Interlocking 723
  23.2.2 Elastic Instabilities 724
  23.2.3 Role of Dimensionality and Disorder 727
  23.2.4 Elastic Instabilities vs. Wear in Atomistic Models 727
  23.2.5 Hydrodynamic Lubrication and Its Confinement-Induced Breakdown 729
  23.2.6 Submonolayer Films 731
23.3 Stick-Slip Dynamics 732
23.4 Conclusions 734
References 735

24 **Mechanics of Biological Nanotechnology** 739
24.1 Science at the Biology–Nanotechnology Interface 740
  24.1.1 Biological Nanotechnology 740
  24.1.2 Self-Assembly as Biological Nanotechnology 740
  24.1.3 Molecular Motors as Biological Nanotechnology 740
  24.1.4 Molecular Channels and Pumps as Biological Nanotechnology 741
  24.1.5 Biologically Inspired Nanotechnology 742
  24.1.6 Nanotechnology and Single Molecule Assays in Biology 743
  24.1.7 The Challenge of Modeling the Bio-Nano Interface 744
24.2 Scales at the Bio-Nano Interface 746
  24.2.1 Spatial Scales and Structures 747
  24.2.2 Temporal Scales and Processes 749
  24.2.3 Force and Energy Scales: The Interplay of Deterministic and Thermal Forces 750
24.3 Modeling at the Nano-Bio Interface 752
  24.3.1 Tension Between Universality and Specificity 752
  24.3.2 Atomic-Level Analysis of Biological Systems 753
  24.3.3 Continuum Analysis of Biological Systems 753
24.4 Nature's Nanotechnology Revealed: Viruses as a Case Study 755
24.5 Concluding Remarks 760
References 761

25 **Mechanical Properties of Nanostructures** 763
25.1 Experimental Techniques for Measurement of Mechanical Properties of Nanostructures 765
  25.1.1 Indentation and Scratch Tests Using Micro/Nanoindenters 765
  25.1.2 Bending Tests of Nanostructures Using an AFM 765
  25.1.3 Bending Tests Using a Nanoindenter 769
25.2 Experimental Results and Discussion 770
  25.2.1 Indentation and Scratch Tests of Various Materials Using Micro/Nanoindenters 770
  25.2.2 Bending Tests of Nanobeams Using an AFM 773
  25.2.3 Bending Tests of Microbeams Using a Nanoindenter 777

25.3 Finite Element Analysis of Nanostructures with Roughness and Scratches ... 778
25.3.1 Stress Distribution in a Smooth Nanobeam ... 779
25.3.2 Effect of Roughness in the Longitudinal Direction ... 781
25.3.3 Effect of Roughness in the Transverse Direction and Scratches ... 781
25.3.4 Effect on Stresses and Displacements for Materials That Are Elastic, Elastic-Plastic, or Elastic-Perfectly Plastic ... 784
25.4 Closure ... 785
References ... 786

Part D Molecularly Thick Films for Lubrication

26 **Nanotribology of Ultrathin and Hard Amorphous Carbon Films** ... 791
26.1 Description of Commonly Used Deposition Techniques ... 795
26.1.1 Filtered Cathodic Arc Deposition Technique ... 798
26.1.2 Ion Beam Deposition Technique ... 798
26.1.3 Electron Cyclotron Resonance Chemical Vapor Deposition Technique ... 799
26.1.4 Sputtering Deposition Technique ... 799
26.1.5 Plasma-Enhanced Chemical Vapor Deposition Technique ... 799
26.2 Chemical Characterization and Effect of Deposition Conditions on Chemical Characteristics and Physical Properties ... 800
26.2.1 EELS and Raman Spectroscopy ... 800
26.2.2 Hydrogen Concentrations ... 804
26.2.3 Physical Properties ... 804
26.2.4 Summary ... 805
26.3 Micromechanical and Tribological Characterizations of Coatings Deposited by Various Techniques ... 805
26.3.1 Micromechanical Characterization ... 805
26.3.2 Microscratch and Microwear Studies ... 813
26.3.3 Macroscale Tribological Characterization ... 822
26.3.4 Coating Continuity Analysis ... 826
References ... 827

27 **Self-Assembled Monolayers for Controlling Adhesion, Friction and Wear** ... 831
27.1 A Primer to Organic Chemistry ... 834
27.1.1 Electronegativity/Polarity ... 834
27.1.2 Classification and Structure of Organic Compounds ... 835
27.1.3 Polar and Nonpolar Groups ... 838
27.2 Self-Assembled Monolayers: Substrates, Head Groups, Spacer Chains, and End Groups ... 839
27.3 Tribological Properties of SAMs ... 841
27.3.1 Surface Roughness and Friction Images of SAMs Films ... 844
27.3.2 Adhesion, Friction, and Work of Adhesion ... 844

27.3.3 Stiffness, Molecular Spring Model, and Micropatterned SAMs 848
27.3.4 Influence of Humidity, Temperature, and Velocity on Adhesion and Friction 850
27.3.5 Wear and Scratch Resistance of SAMs 853
27.4 Closure 856
References 857

28 **Nanoscale Boundary Lubrication Studies** 861
28.1 Lubricants Details 862
28.2 Nanodeformation, Molecular Conformation, and Lubricant Spreading 864
28.3 Boundary Lubrication Studies 866
28.3.1 Friction and Adhesion 866
28.3.2 Rest Time Effect 869
28.3.3 Velocity Effect 871
28.3.4 Relative Humidity and Temperature Effect 873
28.3.5 Tip Radius Effect 876
28.3.6 Wear Study 879
28.4 Closure 880
References 881

29 **Kinetics and Energetics in Nanolubrication** 883
29.1 Background: From Bulk to Molecular Lubrication 885
29.1.1 Hydrodynamic Lubrication and Relaxation 885
29.1.2 Boundary Lubrication 885
29.1.3 Stick Slip and Collective Phenomena 885
29.2 Thermal Activation Model of Lubricated Friction 887
29.3 Functional Behavior of Lubricated Friction 888
29.4 Thermodynamical Models Based on Small and Nonconforming Contacts 890
29.5 Limitation of the Gaussian Statistics – The Fractal Space 891
29.6 Fractal Mobility in Reactive Lubrication 892
29.7 Metastable Lubricant Systems in Large Conforming Contacts 894
29.8 Conclusion 895
References 895

Part E Industrial Applications and Microdevice Reliability

30 **Nanotechnology for Data Storage Applications** 899
30.1 Current Status of Commercial Data Storage Devices 901
30.1.1 Non-Volatile Random Access Memory 904
30.2 Opportunities Offered by Nanotechnology for Data Storage 907
30.2.1 Motors 907
30.2.2 Sensors 909
30.2.3 Media and Experimental Results 913
30.3 Conclusion 918

References 919

31 **The "Millipede" –**
A Nanotechnology-Based AFM Data-Storage System 921
31.1 The Millipede Concept 923
31.2 Thermomechanical AFM Data Storage 924
31.3 Array Design, Technology, and Fabrication 926
31.4 Array Characterization 927
31.5 *x*/*y*/*z* Medium Microscanner 929
31.6 First Write/Read Results with the 32×32 Array Chip 931
31.7 Polymer Medium 932
31.7.1 Writing Mechanism 932
31.7.2 Erasing Mechanism 935
31.7.3 Overwriting Mechanism 937
31.8 Read Channel Model 939
31.9 System Aspects 943
31.9.1 [peserror]PES Generation for the Servo Loop 943
31.9.2 Timing Recovery 945
31.9.3 Considerations on Capacity and Data Rate 946
31.10 Conclusions 948
References 948

32 **Microactuators for Dual-Stage Servo Systems**
in Magnetic Disk Files 951
32.1 Design of the Electrostatic Microactuator 952
32.1.1 Disk Drive Structural Requirements 952
32.1.2 Dual-Stage Servo Configurations 953
32.1.3 Electrostatic Microactuators:
Comb-Drives vs. Parallel-Plates 954
32.1.4 Position Sensing 956
32.1.5 Electrostatic Microactuator Designs for Disk Drives 958
32.2 Fabrication 962
32.2.1 Basic Requirements 962
32.2.2 Electrostatic Microactuator Fabrication Example 962
32.2.3 Electrostatic Microactuator Example Two 963
32.2.4 Other Fabrication Processes 966
32.2.5 Suspension-Level Fabrication Processes 967
32.2.6 Actuated Head Fabrication 968
32.3 Servo Control Design
of MEMS Microactuator Dual-Stage Servo Systems 968
32.3.1 Introduction to Disk Drive Servo Control 969
32.3.2 Overview of Dual-Stage Servo Control Design Methodologies 969
32.3.3 Track-Following Controller Design
for a MEMS Microactuator Dual-Stage Servo System 971
32.3.4 Dual-Stage Seek Control Design 976
32.4 Conclusions and Outlook 978
References 979

33 **Micro/Nanotribology of MEMS/NEMS Materials and Devices** 983
- 33.1 Introduction to MEMS 985
- 33.2 Introduction to NEMS 988
- 33.3 Tribological Issues in MEMS/NEMS 989
  - 33.3.1 MEMS 989
  - 33.3.2 NEMS 994
  - 33.3.3 Tribological Needs 995
- 33.4 Tribological Studies of Silicon and Related Materials 995
  - 33.4.1 Tribological Properties of Silicon and the Effect of Ion Implantation 996
  - 33.4.2 Effect of Oxide Films on Tribological Properties of Silicon 998
  - 33.4.3 Tribological Properties of Polysilicon Films and SiC Film 1000
- 33.5 Lubrication Studies for MEMS/NEMS 1003
  - 33.5.1 Perfluoropolyether Lubricants 1003
  - 33.5.2 Self-Assembled Monolayers (SAMs) 1004
  - 33.5.3 Hard Diamond-like Carbon (DLC) Coatings 1008
- 33.6 Component-Level Studies 1009
  - 33.6.1 Surface Roughness Studies of Micromotor Components 1009
  - 33.6.2 Adhesion Measurements 1011
  - 33.6.3 Static Friction Force (Stiction) Measurements in MEMS 1014
  - 33.6.4 Mechanisms Associated with Observed Stiction Phenomena in Micromotors 1016
- **References** 1017

34 **Mechanical Properties of Micromachined Structures** 1023
- 34.1 Measuring Mechanical Properties of Films on Substrates 1023
  - 34.1.1 Residual Stress Measurements 1023
  - 34.1.2 Mechanical Measurements Using Nanoindentation 1024
- 34.2 Micromachined Structures for Measuring Mechanical Properties 1024
  - 34.2.1 Passive Structures 1025
  - 34.2.2 Active Structures 1028
- 34.3 Measurements of Mechanical Properties 1034
  - 34.3.1 Mechanical Properties of Polysilicon 1034
  - 34.3.2 Mechanical Properties of Other Materials 1036
- **References** 1037

35 **Thermo- and Electromechanics of Thin-Film Microstructures** 1039
- 35.1 Thermomechanics of Multilayer Thin-Film Microstructures 1041
  - 35.1.1 Basic Phenomena 1041
  - 35.1.2 A General Framework for the Thermomechanics of Multilayer Films 1046
  - 35.1.3 Nonlinear Geometry 1054
  - 35.1.4 Nonlinear Material Behavior 1058
  - 35.1.5 Other Issues 1061
- 35.2 Electromechanics of Thin-Film Microstructures 1061
  - 35.2.1 Applications of Electromechanics 1061
  - 35.2.2 Electromechanics Analysis 1063

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

35.2.3 Electromechanics – Parallel-Plate Capacitor 1064
35.2.4 Electromechanics of Beams and Plates 1066
35.2.5 Electromechanics of Torsional Plates 1068
35.2.6 Leveraged Bending 1069
35.2.7 Electromechanics of Zipper Actuators 1070
35.2.8 Electromechanics for Test Structures 1072
35.2.9 Electromechanical Dynamics: Switching Time 1073
35.2.10 Electromechanics Issues: Dielectric Charging 1074
35.2.11 Electromechanics Issues: Gas Discharge 1075
35.3 Summary and Mention of Topics not Covered 1078
References 1078

36 **High Volume Manufacturing and Field Stability of MEMS Products** 1083
36.1 Manufacturing Strategy 1086
36.1.1 Volume 1086
36.1.2 Standardization 1086
36.1.3 Production Facilities 1086
36.1.4 Quality 1087
36.1.5 Environmental Shield 1087
36.2 Robust Manufacturing 1087
36.2.1 Design for Manufacturability 1087
36.2.2 Process Flow and Its Interaction with Product Architecture . 1088
36.2.3 Microstructure Release 1095
36.2.4 Wafer Bonding 1095
36.2.5 Wafer Singulation 1097
36.2.6 Particles 1098
36.2.7 Electrostatic Discharge and Static Charges 1098
36.2.8 Package and Test 1099
36.2.9 Quality Systems 1101
36.3 Stable Field Performance 1102
36.3.1 Surface Passivation 1102
36.3.2 System Interface 1105
References 1106

37 **MEMS Packaging and Thermal Issues in Reliability** 1111
37.1 MEMS Packaging 1111
37.1.1 MEMS Packaging Fundamentals 1112
37.1.2 Contemporary MEMS Packaging Approaches 1113
37.2 Hermetic and Vacuum Packaging and Applications 1116
37.2.1 Integrated Micromachining Processes 1117
37.2.2 Post-Packaging Processes 1118
37.2.3 Localized Heating and Bonding 1119
37.3 Thermal Issues and Packaging Reliability 1122
37.3.1 Thermal Issues in Packaging 1122
37.3.2 Packaging Reliability 1124
37.3.3 Long-Term and Accelerated MEMS Packaging Tests 1125

37.4 Future Trends and Summary 1128
References 1129

Part F Social and Ethical Implication

38 **Social and Ethical Implications of Nanotechnology** 1135
38.1 Applications and Societal Impacts 1136
38.2 Technological Convergence 1139
38.3 Major Socio-technical Trends 1141
38.4 Sources of Ethical Behavior 1143
38.5 Public Opinion 1145
38.6 A Research Agenda 1148
References 1149

Acknowledgements 1153
About the Authors 1155
Detailed Contents 1171
Subject Index 1189

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

Subject Index

1, 1′-biphenyl-4-thiol (BPT) 843
4, 4′-dihydroxybiphenyl (DHBp) 843
μCP (microcontact printing) 187, 197, 834
μTAS (micro-total analysis systems) 257
γ-modified geometry 700
1,4-phenylenediamine (pDA) 916
1/f noise 387, 957
16-mercaptohexadecanoic acid thiol (MHA) 843
1-D localization
– effects 120
– theory 120
1-D to 3-D transition in magnetic field 122
1-ethyl-3-(3-diamethylaminopropyl) carbodiimide (EDC) 80, 296
2-D FKT model 638
2-DEG
– two-dimensional electron gas 438
2-D-histogram technique 647
2-amino-4,5-imidazoledicarbonitrile (AIDCN) 918
2-mercaptoethylamine HCl 478
2-pyridyldithiopropionyl (PDP) 478
3,3′-dimethyl bipyridinium 14
3-D bulk state 428
3-D coil inductor 242
3-D switching architecture 238
3-D-force measurements 358
3-nitrobenzal malonitrile (NBMN) 916

A

abrasive wear 649, 652
Abrikosov lattice 440
ABS (air-bearing surface) 968
ac electrochemical deposition 104
ac electrodeposition 103
accelerated friction 822
acceleration energy 798
accelerometer 230
– lateral 231
– packaging 1100
– test 1101
– three-axis 233
– vertical 232
– wafer fabrication 1091
acetonitrile 14, 19
acetylene (C_2H_2) 799
acid-treated SWNT 64
acoustic emission 690
actin-myosin motor 751
activation energy barrier 482
active linearization 361
active matrix electronic paper 195
active microfluidic devices 257
active microvalves 258
active structures 1028
actuated head 954
– fabrication 968
actuated slider 954, 968
– configuration 954
actuated suspension 953
actuation
– pneumatic 258
actuator
– miniaturized 986
actuator failure modes 1074
adatoms 392
addressing nanoscale bits of data 901
adenine 10
adenosine triphosphate (ATP) 741
adhesion 325, 498, 507, 515, 605, 799, 844, 930, 983
– at SAM interfaces 618
– control 832
– energy 619
– force 452, 546, 547, 552, 564, 566, 567, 619, 621
– force, quantized 558, 584
– hysteresis 567, 568, 573
– hysteresis, relation to friction 573, 576–579, 585
– influence of humidity on 850
– measurement 1011
– measurement techniques 606
– mechanics 564–566
– performance 866
– primary minimum 544, 555, 559
– promoter 188
– quantization 623
– rate-dependent 567, 568
adhesion force
– total 616
adhesion-controlled friction 571, 572, 574
adhesive
– coating 479
– interaction 732
– material 610
– surfaces 731
– wear 728
adhesive force 335, 646, 662, 867, 1004
– increase 876
– intrinsic 832
– mapping 537, 869
– measurement 503
adhesive tape film 907
adiabatic limit 482
adsorbate 404, 695
adsorbed
– insulator substrate 400
– water 875
– water film 878
adsorption 64, 79, 476
– capacity 80
– sites in MWNT 65
AES (Auger electron spectroscopy) measurement 826
AFAM (atomic force acoustic microscopy) 326
affinity maturation 316
affinity selection 297
affinity-based targeting 301
AFM (atomic force microscope) 173–175, 743, 765
– adhesion measurement 620
– *Binnig* design 333
– calculated sensitivity 355
– cantilever 332, 350, 610
– cantilever array 923
– commercial 333
– construction 288
– contact mode 331
– control electronics 362
– design optimization 348
– designs 376
– feedback loop 362
– for UHV application 635
– image 392
– imaging signal 386
– instrumentation 347
– manufacturers 333
– microscratch technique 853
– probe construction 338
– probes 378
– pull-off force 619

– resolution 375
– set-up 450, 453
– spectroscopy 458
– surface height map 817
– test 818
– thermomechanical recording 924
– tip 288, 379, 386, 450, 452, 477, 480, 1013
– tip containing antibodies 489
– tip radius 518
– tip sensor design 476
– tip size 622
AFM image
– tapping mode 914
Ag 395, 397
– on Nb(110) 429
– trimer 395
air bag sensor 1084, 1091
air damping 470
air gap capacitance 1075
air induced oscillations 348
air/water interface 479
air-bearing surface (ABS) 968
Al_2O_3-TiC head 825
Al_2O_3 54, 462, 633
– grains 508
– ultrafiltration membrane 669
Al_2O_3-TiC composite 508
Al_2O_3-TiC slider 998
$Al_xGa_{1-x}As$ 248
AlAs 248
alcohol 837
aldehydes 837
aligned carbon nanotubes 58
alignment accuracy 195
alignment of nanowires 109
alkali halides 397
alkaline phosphatase (AP) 271
alkane 835, 888, 889
alkanethiolate 187
– film 841
– on gold 187
– on palladium 187
– on silver 187
alkanethiols 404
alkylsilane film 619
all-fiber interferometer 418
all-MEMS configuration 907
all-metal cantilever 912
all-optical logic gates 21
all-optical switching network 237
alloy 662
all-silicon cantilever 926
alumina template 127
aluminum
– gates 34
– mirror array 1088
aluminum oxide 54, 462, 633
aluminum-based surface micromachining 241
AM (amplitude modulation)
– AFM 388
– mode 455
amide 838
amine 838
Amontons law 454, 517, 571, 574, 577, 580, 646, 718
amorphous
– oxide layer 106
– surfaces 544
amorphous carbon 49, 793
– chemical structure 800
– coatings 800, 806
– phase 48
amperometric time-based detection method 272
amphiphilic molecular building blocks 23
amplitude feedback 465
amplitude modulation (AM) 388, 456
– mode 331
– SFM (scanning force microscope) 433
amplitude of vibration 911
anchor group 840
AND operator 15
angle of helicity 41
angle of twist formula 355
anion-terminated tip 399
anisotropy 704
– of friction 642
annealing effect 1036
anodic alumina (Al_2O_3) 100
– oxide films 102
– templates 103
anodic bonding 1096, 1115
anodization 100
ANSYS
– analysis 1123
– stress prediction 1124
anthracene 20
– channel 21
antibiotic-resistant pathogens 293
antibody 296
– directed enzyme-prodrug therapy (ADEPTS) 301
– production 312
antibody–antigen 475
– binding 283
– complex 484
– recognition 491
antibody-coated beads 271
antiferromagnetic spin ordering 403
antigen presenting cell (APC) 312
antigen recognition 282
antimicrobial chemotherapy 293
anti-stiction
– agent (ASA) 1104
– material 1104
AP (alkaline phosphatase) 271
APC (antigen presenting cell) 312
APCVD (atmospheric pressure chemical vapor deposition) 770
apparent viscosity 888
applications
– electrical 131
arc discharge 798
areal density 900, 901, 907, 925, 932, 947, 952
Arg-Gly-Asp (RGD) 489
armchair-type SWNT 42
array
– cantilever 927
– characterization 927
– chip 923
– of aligned nanowires 115
– of nanotubes 53
Arrhenius equation 1127
artifacts 415
artificial double helix 12
as-deposited film 1035
aspect ratio 127
asperity-asperity interaction 618
as-released curvature 1059
assembled nanotube probes 378
assembly approaches 288
assembly of nanostructures 294
association process 476
atmospheric pressure chemical vapor deposition (APCVD) 770
atomic
– interaction force 452
– manipulation 419
– motion imaging 420
atomic force acoustic microscopy (AFAM) 326
atomic force microscope (AFM) 114, 288, 289, 303, 325, 331, 371, 385, 404, 476, 498, 546, 547, 549, 585, 588, 688, 717, 743, 841, 861, 862, 899, 922
atomic resolution 330, 382, 388, 389, 427, 476, 480
– imaging 391, 394–397, 462
atomic-scale
– dissipation 655
– force measurement 347

– friction 507
– hysteresis 468
– image 327, 342, 434
– roughness 721
atomistic computer simulation 398, 724
ATP hydrolysis 742
ATP(adenosintriphosphate) synthase 741
attraction
– long-range 550, 562
attractive force–distance profile 480
attractive interaction 864
Au film 771
Au microbeam 777, 778, 785
Au tip 619
Au(111) 423, 848, 850
– surface 619
Aubry transition 728
Au-coated AFM tip 619
austenite 665
average
– distance 749
– lifetime 482
– stress 1049
axial strength 672
azopyridine 19

B

backbone chain 839
backflow 261
bacmid 283
bacteriophage 744, 755
bacteriorhodopsin 742
BaF_2 398
ball and spring model 724
ball grid arrays (BGA) 1113
ballistic phonon transport 126
ballistic transport 115, 121
balloon angioplasty 189
bamboo texture 52, 61
band gap tunability 128
band structure
– electronic 128
barcode reading 233
barcode tags 136
barrier-hopping 887
– fluctuations 890
basic switching operations 15
batch fabrication 249
– techniques 373
batch nanotube tip fabrication 380
B-cell epitopes 311
bc-MWNT (bamboo-concentric) 44
BDCS (biphenyldimethylchlorosilane) 849
beam
– two-layer 1045
beam bending energy 745
beam deflection 455
beam failure 775
beam-deflection FFM 633
beam-steering mirror 239
behavior
– unethical 1143
Bell’s formula 483
bending
– moment 767, 783
– stiffness 349
– strength 767, 774, 775, 785
– stress 767, 781
– test 765, 766, 769
Berkovich
– indenter 765
– indenter tip 677
– pyramid 696
– tip 815
BGA (ball grid arrays) 1113
bh-MWNT (bamboo-herringbone) 44
$Bi_2Sr_2CaCu_2O_{8+\delta}$ (BSCCO) 431
$Bi_{1-x}Sb_x$ nanowires 123
bias voltage 328
biaxial strain 1042
biaxial stress 1053
BiCMOS 229
bidirectional micropump 261
bifurcation 1056
bilayer
– beam 1054
– cantilever 1046
– film 1059
– model-lipid-lubricant 889
– plate 1050
– system 1042
bimetallic catalyst 51, 62
binary compound 397
bio force probe (BFP) 480
biochemical fluids 254
biochemical reaction 262, 270
biocompatibility 263, 311
bioconjugate chemistry 295, 296
bioconjugation method 288
biofilter 271
– surface-mounted 272
biofluidic chip 273, 987
biofunctionalized cantilever 742, 743, 745
biogenic amorphous silicon shells 314
biological
– activity 316
– affinity 297
– affinity reagents 303
– device 742
– device components 304
– evolution 751
– macromolecules 299
– nanotechnology 739, 760
– self-assembly 293
biomaterials 711
biomechanics 675
biomedical device 189
BioMEMS (Biological or Biomedical Microelectromechanical Systems) 253, 262, 764, 987
– actuation 269
biomimetic valves 262
biomolecules bound to carbon nanotubes 303
biomolecules in therapeutic nanodevices 293
bio–nano interface 744
BioNEMS (Biological or Biomedical Nanoelectromechanical Systems) 764, 989
bio-sampling
– magnetic bead-based 271
biosensors 80, 302
bio-surfaces 489
biotechnology 1139
– limits of 281
biotin 80, 476
biotin-avidin 622
– spectrum 487
biotin-directed IgG 476
biotinylated AFM tip 489
biphenyl 419
bipyridine
– building block 12, 31
– centered LUMOs 26
birefringent crystal 352
bismuth nanowire 118, 134
bistability 459, 916
bit pitch (BP) 943
bit strings 938
bits per inch (BPI) 952
blister test 691
Bloch states 428
Bloch wave 427
block co-polymer 291
block-like debris 823
blood cell analysis 310
blood flow measurement 304
blood-pool agents 306
BN nanotubes 68

Bode plot 976
Bohr radius of excitons 135
Boltzmann ansatz 482
Boltzmann distribution 887
bond
– angle 67
– breakage 483
– lifetime 482
– rupture 487
– scission 623
– strength 550
bonded lubricant film 864
bonded PFPE 1012
bonding
– energy 386
– kinetics 893
– techniques 1114
bone
– lamellae 675
– lamellation 679
– tissue 675
boron
– diffusion 227
– doped Si 205
– ion implantation 330
bottom-up
– approach 99, 107, 285
– chemical strategies 10
– materials 281
– scheme 133
boundary
– element (BE) 1063
– film formation 861
– lubricant 729, 861
– lubricant film 832
– lubrication 530, 569–571, 580, 583, 585, 852, 866, 885
– lubrication measurement 507
– lubrication regime 884
– scattering effects 125
bovine serum albumin (BSA) 476
bpsi (bits per square inch) 901
BPT (1, 1′-biphenyl-4-thiol) 854, 1007
BPTC (cross-linked BPT) 1007
branched hydrocarbon lubricants 889
breakdown
– electrical 1076
bridging of polymer chains 562, 563
broken beams 777
broken coating chip 817
brush *see* polymer brush
brush-like structures 109
buckling 808, 812, 1026
– force 377
– stress 813
bulge test 691
bulk
– addressing 22
– atoms 387
– conduction band 424
– diamond 793, 801, 803
– etched silicon wafer 263
– fluid transport 261
– graphitic carbon 824
– micromachining 229
– Si wafers 205
– state 428
– viscosity 731
– xenon 436

C

C_{70} 792
C_2 moieties 51
C_{60} 32, 41, 49, 404, 647, 792
– film 330
– fullerenes 289
– island 641, 642
– multilayered film 404
– SWNT (single wall nanotube) 70
C_{60}-terminated film 843
Ca ion 399
cadherin-mediated adhesion 488
CaF_2 398
CaF_2 tip 652
$CaF_2(111)$ 399
$CaF_2(111)$ surface 399, 652
calculated tapping behavior 911
calculus of variations 745
calibration 610, 698, 699
calorimetric experiments 482
cantilever 387, 390, 457, 481
– all-metal 912
– all-silicon 926
– array 926, 927
– axis 418
– base 469
– beam 1015, 1061
– beam array (CBA) 1011
– beam model 1067
– biofunctionalized 742, 743, 745
– biosensors 303
– cell 926
– deflection 332, 372, 387, 389, 418, 465, 480, 632, 745, 848
– deflection calculation 344
– diode array 928
– driven 455
– effective mass 633
– eigenfrequency 415
– elasticity 636
– fabrication 909
– flexible 332
– foil 373
– heater 925
– material 339, 387
– motion 331, 913
– mount 336
– oscillation 457, 469
– Q-factor 433
– resonance 451
– resonance behavior 350
– resonance frequency 481, 611, 890, 1029
– sensor 925
– spring 454
– spring constant 485
– stainless steel 813
– stiffness 338, 505, 610, 926
– temperature 940
– thickness 632, 633
– tip 303, 454
– triangular 339, 349, 354
– untwisted 335
cantilever-based probes 433
capacitance detection 356
capacitive
– accelerometers 230
– detection 361
– detector 331
– displacement sensor 1033
– forces 465
– position sensing 956
– sensor 227, 230
– transducer 245
capacitor-like device 913
capacity growth 903
capillarity-driven stop valve 265
capillary
– effect on adhesion 607
– electrophoresis (CE) 255
– force 544, 559, 564, 619, 662
– force curve 617
– formation 614
– pressure 265
– stiffness 614
capsid wall 760
carbon 792
– crystalline 792
– film 865
– magnetron sputtered 799
– shells 52
– source 57
– spacer chain 853
– superactivated 77

carbon coating
– unhydrogenated 804
carbon nanofiber
– vapor grown 56
carbon nanotube (CNT) 55, 105, 287, 425, 667
– adsorption properties 76
– application 73
– catalyst-free growth 59
– catalyst-supporting materials 75
– chemical reactivity 67
– diameters 63
– FETS 303
– field emission 74
– formation 46
– growth mechanisms 59
– heterogeneity 63
– in situ filling 69
– maximum current density 74
– mechanical properties 376
– molten state filling 69
– oxidation 72
– production 46
– properties 63
– quantum wires 66
– sublimation filling 70
– synthesis conditions 61
– tip 342, 375, 376, 501
– wet chemistry filling 69
carbon–carbon distance 436
carboxyl acid 837
carboxylates ($RCOO^-$) 404
cardiovascular
– sensors 310
– system 310
– tissue engineering 310
carrier
– density 118, 122
– gas 804
– mean free path 116, 118, 121
– mobility 118
Casimir force 545
casting 186
catalysis 397
catalysis-enhanced disproportionation 54
catalysis-enhanced thermal cracking 54
catalyst 49, 106, 668
– nanoparticles 58
– preparation 54
catalyst-based SWNT 61
catalyst-free 57
catalytic chemical vapor deposition (CCVD) 52, 55, 58, 60, 61
catalytic decomposition 54
catalytically grown MWNT 673
cathode deposit 49, 51
cathodic arc carbon 806
CCVD *see* catalytic chemical vapor deposition
CCVD method 61
CDS (correlated double sampling) technique 961
CdSe nanorods 136
cell adhesion 489
cellular immune responses 311
cellular phone 904
$CeO_2(111)$ 400
ceramic
– matrix composites 81
– slider 824
– tip 74
cerpacs 1101
CFM (chemical force microscopy) 73, 489, 619
chain adsorption 894
chalcogenide 906, 910, 915
change of meniscus 1005
changing of surface conditions 117
channel etching process 255
channel-liquid pair 261
chaotic mixer 266
characteristic distance 588
characteristic slip time 894
characteristic velocity 889
charge density wave (CDW) 428
charge exchange interactions 545, 552, 553
charge fluctuation forces *see* ion correlation forces
charge separation 1116
charge transfer interactions *see* charge exchange interactions
charge trapping 913
charge-controlled pull-in equation 1064
check valve design 260
chelate metal cations 12
chemical
– bond 386
– bonding 388, 863
– bonding force 331
– characterization 800
– composition 115
– detection devices 137
– force 386
– force microscopy (CFM) 73, 489, 619
– heterogeneity 568
– input 16
– interaction force 452
– sensors 74
– signal 19, 20
– synthesis 12
– vapor deposition (CVD) 104, 793
– vapor infiltration 57
chemisorption 861
chemistry route 57
chemoselective conjugation 295, 296
chemotherapy 307
chip level integration 1093
chip-on-flex (COF) 1113
clamped-clamped mechanical beams 247
classical finite size 116
classical size effect 120
clock field 946
clogging 56
closed loop sensitivity 973
clustered acetate 406
CMOS electronics 230
CMOS static random access memory (SRAM) 235
CMOS sustaining electronics 245
c-MWNT (concentric multiwall nanotube) 43, 48, 60, 67
Co clusters 425
CO disproportionation 54
CO on Cu(110) 419
coated particles 306
coating
– continuity 826
– damage 823
– failure 820
– hardness 818
– hydrophobic 1014
– mass density 801
– microstructure 800
– thickness 810, 812
coating–substrate interface 819
coefficient
– effective 647
coefficient of friction 501, 510, 516, 535, 579, 634, 771, 773, 809, 813, 823, 824, 844, 868, 872, 996, 1005
– Si(100) 531
– Z-15 531
– Z-DOL (BW) 531
coefficient of friction *see* friction coefficient
coefficient of friction relationship 343
coefficient of thermal expansion (CTE) 1124
co-energy 1064
coercivity 137

COF (chip-on-flex) 1113
cognate receptors 298
cognitive ligands 476
cognitive science 1139
coherence length 423
cohesive
– potential 759
– surface model 759
coil inductor 242
cold welding 553, 579
collagen fibers 679
collective phenomenon 885, 888
colloidal
– forces 545
– probe 548
colossal magneto resistive effect 439
comb storage device 907
comb-drive 955, 1030
– actuator 955
– resonator 245
combined AFM nanoindenter device 676
commensurate Cu(111) 728
commensurate system 723
commercial data-storage device 907
commercial MEMS devices 991
communicating
– between compatible molecular components 19
– force 748
– molecular switches 19
compact bone
– lamellae 678
– mean hardness 680
complex
– bonds 486
– dielectric function 127
– fluidic systems 886
– logic functions with molecular switches 19
component
– failure 764
– supplier 902
– surface 1009
composite 662
composite material 81
compositionally modulated nanowires 106
compression energy 887
compressive stress 805, 812, 1035, 1044
computer simulation 718
– forces 553, 557, 561, 564, 565, 585
– friction 553, 573, 574, 579, 580, 583, 585, 586, 590, 591, 594
concentration
– critical 650
concentric
– graphenes 43
– texture 61
– type (c-MWNT) 43
conditional release of therapeutics 288
conductance 117
conductance quantization in metallic nanowires 115
conducting polymers 103, 304
conductive polymeric materials 304
conductivity 136
– electrical 133
confined
– complex liquids 891
– nanostructure 68
– simple liquids 888
confinement 544, 545, 570, 584, 593
conformation 864
conformational defect 849
conforming contacts 894
connector
– miniaturized 986
constant
– amplitude (CA) 464
– amplitude FM mode 468
– current mode 328
– excitation mode (CE) 464
– force mode 362
– height mode 328, 399
constrained elastic media 1061
consumer electronics 904
consumer nanotherapeutic device 288
contact
– AFM (atomic force microscope) 455
– AFM dynamic method 662
– angle 616, 833
– angle goniometer 847
– angle of SAM 846
– conductance 649, 913
– elastic 703
– mechanics *see* adhesion mechanics
– meniscus 619
– mode 373, 909
– mode photolithography 195
– printing 186, 200, 833
– radius 678
– resistance 690
– stiffness 677, 689, 698, 701, 810, 894
– stress 872
– value theorem 575
contact area 548, 650, 651, 677, 699, 709
– apparent macroscopic 571, 576, 580
– true molecular 547, 549, 566, 571, 574–576, 594
contamination 332
continuous micropump 269
continuous stiffness measurement (CSM) 810
continuum
– analysis 753
– model 638
– theory 554, 556, 558, 559, 569, 580
contour plot 1057
contour-mode disk resonator 246
contraction 1044
contrast 435
– agent design 306
– agents 304
– enhancing agents for medical imaging 304
– enhancing nanoparticle 306
– formation 397
control
– of nanotherapeutic action 302
– of position 109
– over nanowire diameter distribution 106
– system 359, 362
– theory 1144
controlled desorption 419
controlled evaporation 269
controlled geometry (CG) 330
controlled nanoscale architecture 294
controlled substances 1140
controlled triggering of therapeutic action 298
controller design 969
CoO 402
Co-O film on PET 524
Cooper pairs 429
coordinated molecular process 886
copper adhesion 963
copper hexadecafluorophthalocyanine 196
copper tip 652, 727
corannulene 61
cores of carbon nanotubes 105
core-sheath nanowire 114

core-shell structure 132
Coriolis acceleration 232, 1095
correlated-double-sampling (CDS) 960
corrosion 1116, 1125
cost of goods (COGs) 285, 307
co-transporter 489
Couette flow 569, 593, 885
Coulomb
– force 396
– interaction 66
– law of friction 644
coupled electromechanics 1063
covalent
– bonds 10
– scaffolds 12
Cr binding layer 619
Cr coating 432
Cr(001) 427
crack spacing 524
cracks 807
creativity at the nanoscale 1135
creep 415, 691, 1125
– effect 528
– relaxation 1060
creep model
– linear 890
critical
– concentration 650
– curvature 1056
– degree of bending 1031
– load 809
– magnetic field strength 122
– normal load 853
– position 636
– shear stress 570–572, 577, 585, 586
– temperature 423, 434, 1059
– time 894
– velocity 585, 587, 590, 591
crosshair alignment 195
crosslinked BPT (BPTC) 843
crosstalk 363, 927, 1066
cross-track distance 945
cryostat 415
crystal
– growth direction 103
– orientation 206
– structure 333
– surfaces 375
crystal growth orientation, preferred 104
crystalline
– carbon 792
– silicon 254
– surfaces 544
CSM (continuous stiffness measurement) 810
CTE (coefficient of thermal expansion) 1124
CTL (cytotoxic T-cell) responses 311
Cu(001) 397
Cu(100) 640
Cu(100) substrate 728
Cu(100) tip 728
Cu(111) 423, 639, 640, 655
Cu(111) surface states 427
Cu(111) tip 653
cube corner 696, 704
culture of engineering 315
cuprates 431
current density effect 1036
current rectification 116
current-voltage characteristics 116, 916
curvature 1044
– critical 1056
– effects 723
– measured 1053
– variation 1046
curved electrode
– actuator 1070
– shape 1071
custom-designed MEMS product 1112
cut-off distance 552, 573
CVD (chemical vapor deposition) 104, 793
cyclic
– fatigue 810, 1036
– olefin copolymers (COC) 257
– thermomechanical load 1061
– voltammograms 29
cytoplasmic surface 478
cytosine 10
cytoskeletal filament 747
cytotoxic therapeutics 307
cytotoxic T-lymphocytes (CTLs) 311
cytotoxin deposition 309

D

(*d*, *k*) codes 947
D_2O 425
damage 553, 569–588
damage mechanism 818
damped harmonic oscillator 454
damping 664
– effective 470
– effects 451
– pneumatic 418
data
– lifetime 905
– rate limitation 922
– rates 948
– storage 900
– storage application 912
– storage device 922
– storage system 994
dc electrochemical deposition 103, 104
de Broglie wavelength 116
Deborah number 583
debris 823
Debye exponential relaxation 885, 892
Debye length 555, 557
decoupled design 970
decoupled track-following control 972
deep reactive ion etching (DRIE) 156, 162, 254, 963
defect motion 415
defect nucleation 702
defect production 651
defects in channel 255
deflection 457, 1054
– curve 1071
– maximum 1071
– measurement 387
– noise 433
deformation 547, 549, 1041, 1047
– behavior 1058
– elastic 450, 566, 581, 647
– of microtubule 674
– plastic 553, 580
– regime 1046
deformed beam 778
degradability 311
degrees of freedom 753
delamination 704, 804, 808, 812, 1125
delivery of therapeutics 296
Demnum-type PFPE 537
– lubricant film 869
dendrimer-based therapeutic 284
dendritic
– cells (DCs) 312
– polymers 310
density
– functional theory (DFT) 78
– modulation 730
– of defects 193, 195
density of states (DOS) 66
– electronic 133
dental enamel 711

depletion
– attraction 545, 562
– interaction 562
– stabilization 545, 562
deployment of functional biomolecules 288
deposition
– conditions 206
– rate 798
– techniques 795, 796
Derjaguin approximation 546, 557, 559
Derjaguin–Landau–Verwey–Overbeek (DLVO) theory 555
Derjaguin–Muller–Toporov (DMT)
– model 647
– theory 566
design rule 348
designed molecules 12
design-for-manufacturability 1086
design-for-volume 1086
detection 939
– circuit 939
– systems 332, 350
device
– architectures 288
– components communication 304
– design 280
– electronic 133
– fluidic 253
– molecule-based 24, 28
– scaling 989
dewetting 868, 1003, 1013
DFM (dynamic force microscopy) 433, 471, 490, 501
DFS (dynamic force spectroscopy) 484
DHBp (4, 4′-dihydroxybiphenyl) 843, 854, 1007
diameter-dependent 116
diamond 67, 373, 379, 673, 698, 792
– coated cantilever 914
– coating 804, 833
– film 793
– like amorphous carbon coating 800
– like carbon (DLC) 524, 1003
– powder 51
– tip 504, 651, 662, 676, 814, 819, 841, 866, 999
diatoms 314
diblock copolymers 102, 103, 138
dicationic BIPY 17
dielectric
– breakdown 691
– capacitance 1075
– charging 1074
– function 127, 129
differential scanning calorimetry (DSC) 664
differential-drive configuration 955
diffuser micropumps 261
diffusion 1061
– based extractor 266
– coefficient 265, 865
– flame synthesis 58
– limited reaction 893
– parameters 420
– thermally activated 415
diffusive
– communication 750
– relaxation 482
– transport 115, 116
digital
– feedback 362
– light processing (DLP) 987
– micromirror device (DMD) 235, 987, 992
– signal processor (DSP) 328
– transmission between molecules 20
dilation 573, 584, 586, 593
dimension 387
dimensions of spatial structures 747
dimer structure 394
dimer-adatom-stacking (DAS) 392
Diophantine equation 973
DIP (dual in-line packaging) 1118
dip coating technique 863
diphtheria toxin (DT) 300
dipole molecule 834
dip-pen nanolithography (DPN) 175, 288
diprotic acid (11-thioundecyl-1-phosphonic acid) 621
direct overwriting 938
direct parallel design 970
direct write electron beam/focused ion beam lithography 186
directly growing nanotubes 379
discontinuous sliding motion 887
disjoining pressure 871
disk drive
– manufacturer 902
– servo control 968
dislocation
– line tension 704
– motion 570
– nucleation 696
dispersion
– force 406
– mixing 266
displacement 1047, 1057
– amplitude 930
– calibration 612
– controlled scanning probe 610
– field 1055
– maximum 1070
– resolution 681
– vertical 779
display
– paper-like 193
disposable biosensors 274
disposable plastic lab-on-a-chip 993
dissemination of biological information 315
dissipated power 939
dissipation 468
– force 461
– measurement 631
dissipative tip-sample interaction 468
dissociation 476
distortion 360
distributed
– Bragg reflector (DBR) 199
– feedback (DFB) 199
– laser resonators 199
dithio-bis(succinimidylundecanoate) 476
dithio-phospholipids 479
D-L amino acids (aas) 291
DLC 1012
– coating 524, 794, 796, 809, 819, 826, 1008
– coating microstructure 812
DLP technology 1088
DLVO interactions 549, 555–558
DMD (digital mirror device)
– fabrication 1088
– manufacturing 1095
– packaging 1100
DMT (Derjaguin–Muller–Toporov)
– model 695
– theory 724
DNA 404, 464, 467, 476, 747
– analysis 272
– analysis system 265
– nanostructure 294
– nanowire 33
– packing 756
domain
– pattern 439
– re-orientation 913
donor impurity concentration 127

doped
– anode 49
– polysilicon film 1001
– silicon 996
– silicon wafer 31
DOS structure 423
double barrier resonant tunneling 124
double-layer
– force 621
– interaction 545, 555–557
double-wall nanotubes (DWNTs) 51
doubly clamped beam 1028
doxorubicin 308
DRAM 923
Drexler 1146
DRIE (deep reactive ion etching) 156, 162, 166, 168–170, 254, 954, 963
dried tissue properties 676
drive amplitude 470
drive current
– maximum 976
driven cantilever 455
driving frequency 351
driving/sensing circuit integration 960
drug
– delivery 285
– resistant organism 293
– targeting 297
dry
– etching 205
– nitrogen environment 999
– (plasma) etching 254
– thermal oxidation 286
dry surfaces 571, 587
– forces 550, 552
– friction 571, 576, 588, 589
dual-axis gyroscope 233
dual-in-line packaging (DIP) 1118
dual-stage
– configuration 952
– sensitivity 973
– servo system 953
– short span seek 977
Dupré equation 569
DWNT (double-wall nanotubes) 51
dynamic
– interactions 544–546, 570, 581, 584
– mode 331
– operation mode 388
– oscillation force 456
– spectral analysis 891
– states of friction 894
– viscosity 264
dynamic AFM 387, 453–455, 461, 468, 469
dynamic force microscopy (DFM) 457, 471, 490, 501
dynamic force spectroscopy (DFS) 484
dynamic friction *see* kinetic friction

E

E. coli
– bacterium 748
– cells 756
early failure 1126
early tumor detection 310
EBD (electron beam deposited) tips 376
ECR (electron cyclotron resonance) CVD 795, 815, 1009
– coating 822
EDP (ethylene diamine pyrocatechol) 254
EELS (electron resonance loss spectroscopy) 800
EFC (electrostatic force constant) 678
effect of surface roughness 779, 781
effective
– coefficient of friction 647
– damping constant 470
– force gradient 463
– ligand concentration 485
– mass 433
– medium 127
– medium theory 127
– shear stress 648
– spring constant 637, 645, 721
– structural stiffness 1053
– tether length 485
– viscosity 583, 593, 885
effects of doping and annealing 120
EHD
– induction type 262
– injection type 262
eigenstrain 1041
elastic
– contact 703
– energy 732
– force 727, 757
– Frenkel–Kontorova chain 727
– Hamiltonian 755
– instabilities 727
– mismatch 1051, 1056
– modulus 664, 698, 767, 770, 774, 805, 806, 809
– properties 673
– properties of ssDNA 746
– stiffness coefficient 1048
elastic deformation 450, 647
– long-range 722
elasticity 433, 663
– of DNA 757
– two-dimensional 754
elastic-plastic deformation 785
elastohydrodynamic lubrication 569, 570, 581, 583
elastomer precursor 256
elastomeric stamp 190
electric charging 913
electric field gradient microscopy (EFM) 114
electric force gradient 338
electrical
– applications 131
– breakdown 1076
– conductivity 133
– isolation 956
– surface passivation 1103
electric-arc
– method 61
– reactor 48
electroactive
– fragments 14
– layer 24
– solids 22
electrochemical
– AFM 338
– cell 24
– degradation 304
– deposition 103, 675
– detection 270
– etch 383
– etching 373
– immunosensor 272
– STM 330
electrochemically etched tips 383
electrode
– arrays 26
– geometry 1076
– pair 1077
electrodeposition 103, 153, 180
electrodes with molecular layers 27
electrohydrodynamic (EHD) pumping 261
electroless deposition 188
electro-luminescence (EL) 132, 135
electrolytic
– conduction 1116

– method 57
electromagnetic
– actuators 259
– forces 550
– microvalves 259
electromechanical
– coupling 245
– load 1040
– mirror 1069
electromechanics
– analysis 1063
– application 1061
– test structure 1072
electromigration 419
electron
– beam deposition (EBD) 376
– cyclotron resonance chemical vapor deposition (ECR-CVD) 795, 799
– energy loss spectrometer (EELS) 113
– energy loss spectroscopy (EELS) 798
– Fermi wavelength 116
– field emission 133
– interactions 428
– mean free path 115
– scattering 120
– tunneling 327
electron-beam lithography 674
electronegativity 834
electron-electron interaction 423
electronic
– band structure 128
– density of states 133
– device 133
– ink 195
– newspaper 196
– noise 333
electron-phonon interaction 428
electro-optical modulation 136
electro-osmosis (EO) 261
electro-osmotic
– flow (EOF) 255
– pumping 260
electroplated microactuator 966
electroplating 153, 1113
electro-rheological fluid 260
electrostatic
– actuation 236, 244, 248
– binding 478
– coupling 1063
– discharge (ESD) 1098
– interaction 393, 402, 464
– microactuator 952
– micromotor 1061
– potential 398, 438
– short-range interaction 398
– spring constant 956
electrostatic force 386, 437, 544–558, 1062, 1103
– constant (EFC) 677
– interaction 393
– microscopy 438
electrostatically actuated devices 258
electrostatic-induced errors 1098
embedded atom method (EAM) 720
embedded system 721
embossing 833
– technique 191
end caps 47
end deflection 1026
end group 864
endofullerene 71
endothelial cell surfaces 488
energy
– barrier 482
– conservation 469
– dissipation 468, 567–569, 572–575, 578, 587, 654, 849
– elastic 732
– electrical 1063
– mechanical 1063
– resolution 422
– scales 750
– separation between subband 115
– total 1063
energy dispersive X-ray spectrometer (EDS) 113
engineered nanostructures 13
engineering materials 758
enhanced permeability and retention effect (EPR) 307
entangled states 570
entanglement 622, 886
– strength 886
entropic force 564
entropically
– confined systems 891
– cooled layer 888
entropy
– structural 886, 887
environmental shield 1087
enzymatic cleavage 287
enzymatic prodrug activator 301
enzymes 482
epitaxial silicon seal 1114
epitaxy 152, 170, 172
epitope mapping 491
epoxy bonding 1115
EPR tumor targeting 310
equibiaxial
– deformation state 1050
– stress distribution 1051
equilibrium
– equipartition 720
– interactions *see* static interactions
– true (full) or restricted 562
equivalent mechanical circuit 1073
erasing mechanism 936
ESD protective circuitry 1098
ester 837
etching 186, 1041
ethanolamine 477
ether 837
ethical rule violation 1143
ethics of nanotechnology 1138
ethylene diamine 287
ethylene diamine pyrocatechol (EDP) 254
Euler
– buckling criterion 1026
– equation 349
Euler–Bernoulli beam equation 1063
eutectic bonding 1115
evolution of curvature 1059
examples of NEMS 988
exchange
– force 402
– force interaction 402
– interaction energy 402
exchangeable carrier plate 636
excitation
– external 456
excitation amplitude 461
exciton 127, 136
– binding energy 135
– confinement 135
– laser action 135
extended exponential Kohlrausch relaxation function 892
external
– excitation 456
– excitation frequency 466
– noise 415
– normal load 867
– triggering 284
– vibrations 329
extraction of mechanical properties 1073
extravasation 297, 307
extrinsic stress 1041
Eyring
– model 644, 887, 890
– theory 887

F

Fab molecule 476
fabrication
– of membrane 314
– of microfluidic systems 254
– of mold masters 255
– techniques 254
– technology 225
face-centered cubic (fcc) 792
faceted 109, 111, 135
failure
– density function 1125
– mechanism 825, 1125
– mode 1101
– mode effect analysis (FMEA) 1101
– of MEMS/NEMS device 1012
– rate 1126
Fano resonance 425
fatigue 810
– behavior 769
– crack 776, 823, 1125
– damage 810
– failure 1036
– life 809
– measurement 1032
– properties 768, 776
– resistance 1028
– strength 776
– test 776, 813
fatty acid monolayer 841
FC (flip chip) technique 1113
FCA (filtered cathodic arc) 815
– coating 806, 811
$Fe(NO_3)_3$ 381
feasible simulation
– maximum 719
Fe-coated tip 403
feedback
– architecture 228
– circuit 333, 464, 465
– loop 363, 373, 450, 456, 461, 465, 485, 634, 957
– network 327
– signal 388
Fe-N/Ti-N multilayer 675
FeO 402
Feridex 306
Fermi
– energy 122
– level 33, 422
– points 428
ferrocene 59
– hydrophobic 24
ferroelectric film 913
ferromagnetic
– particles 305
– probe 439
– tip 441
FE-SEM image 136
FFM (friction force microscopy) 632
– dynamic mode 652
– on atomic scale 644
– tip 507, 636, 639
FIB (focused ion beam) tips 376
fiber
– optic telecommunication 233
– optical interferometer 352
– optics 299
FIB-milled probe 342
FID
– free induction decay 305
field
– emission 910
– emission tip 378, 910
– emission-based display 58
– emission-based screen 74
– ion microscope (FIM) 452
field-effect transistor (FET) 85, 116, 132, 137, 289, 290, 302
figure of merit 133
filled nanotube 69
filling efficiency 71
film/substrate system 1045
films on substrate 1023
film–substrate interface 1035
film–substrate system 1060
filtered cathodic arc (FCA) deposition 796–811
filtering 246
FIM (field ion microscope) 452
– technique 464
finite
– difference (FD) 1063
– element (FE) 1063
– element analysis 778
– element method (FEM) 764, 1049
– element modeling (FEM) 699
– size effect 131
first generation NEMS devices 247
first principle calculation 396
first principle simulation 434
FKT (Frenkel–Kontorova–Tomlinson)
– model 637
flagellar motor rotation rate 750
flap type valves 261
flash memory 899, 904–907
flat punch 697
flavin adenine dinucleotide (FAD) 303
flavoenyzme glucose oxidase (Gox) 303
flexible cantilever 332
flexible electronic devices 193
flip chip technique (FC) 1113
floating catalyst method 55
flocculation 555
Flory temperature 561
flow
– activation energy 865
– laminar 264
– rate 269
– resistance 266
fluid mosaic membranes 293
fluidic
– devices 253
– motion 270
– sampling 253
fluorescence
– quantum yield 17
– spectroscopy 127
fluoride 397, 398
fluoride (111) surface 400
fluorinated DLC 833
fluorinated silane monolayer 190
FM (frequency modulation)
– AFM 464
– AFM images 435
– mode 461
FMEA (failure mode effect analysis) 1101
focused ion beam (FIB) 330, 342, 376, 693, 709
foil cantilever 373
Fokker–Planck 891
folded protein structures 293
force
– advection 266
– and displacement calibration 618
– between macroscopic bodies 551, 556
– between surfaces in liquids 554
– calibration 610, 632
– calibration mode 526
– calibration plot (FCP) 867, 1003
– cantilever-based 433
– constant 618
– curve 382
– detection 348
– distribution 1043
– effective gradient 463
– elastic 727, 757
– extension curve 744
– induced unbinding 487

– long-range 389, 398, 437
– mapping 490
– measurement 387
– measuring techniques 546, 547, 549
– modulation mode (FMM) 663
– modulation technique 505
– packing curves 758
– repulsive 480
– resolution 433, 481
– scales 750
– sensing tip 358
– sensitivity 481
– sensor 332, 389, 635
– spectroscopy (FS) 362, 434, 436, 476
– titration 621
– undulation 545, 564
– velocity curve 752
force field spectroscopy
– three-dimensional 437
force-displacement
– curve 608, 677, 681
– plot 607
force-distance
– calibration 848
– curve 338, 451, 456, 621, 662, 668
– cycle 480
– diagram 469
form factor 900, 907
formate
– covered surface 405
– (HCOO$^-$) 404
formation of SAM 840
four helical bundle proteins 316
four-quadrants photodetectors 632
Fowler–Nordheim
– equation 1076
– plot 1077
– tunneling 905
fractal
– mobility 892
– space 891
– time 892
– time dependence 893
fracture 704, 1124, 1125
– strength 1036
– stress 767
– surfaces 777
– toughness 765, 767, 771, 775, 806, 807, 809, 810, 818
– toughness measurement 1027, 1030
Frank-Read source 703
free carrier density 127
free induction decay (FID) 305
free surface energy 874
freezing-melting transitions 588, 590, 886
Frenkel–Kontorova–Tomlinson (FKT) 637
– elastic chain 727
– model 725
frequency
– measurement precision 358
– modulation (FM) 388
– modulation (FM) mode 331
– modulation AFM (FM-AFM) 389
– modulation SFM (FM-SFM) 415
– shift calculation 390
– shift curve 463, 464
friction 325, 498, 507, 515, 606, 618, 631, 717, 793, 844, 983
– anisotropy 642
– characteristics 1009
– coefficient 569–572, 576, 577, 579, 580, 583, 590, 592, 593, 886
  load dependence 518
– control 832
– directionality effect 512
– effect of humidity on 515
– effect of tip on 515
– experiments on atomic scale 638
– global internal 667
– image 640
– kinetic 571–573, 583, 585–587, 589, 590, 593, 718, 728, 729, 884, 885
– lateral 453
– loop 637, 638, 649
– macroscale 842, 998
– macroscopic 732
– map 583, 640
– measurement 1033
– measurement methods 342
– measuring techniques 547, 548, 576, 583, 585
– mechanism 510, 723, 843, 848, 871, 1013
– molecular dynamics simulation 651
– performance 866
– rate experiments 888
– scale dependent 518
– torque 1015
– torque model 1014
– total 726
friction force 331, 335, 501, 516, 642, 813, 843, 850, 866, 1004, 1005
– calibration 335, 346
– curve 363
– decrease of 875
– image 845
– kinetic 722
– magnitude 345
– map 511, 648
– mapping 869
– microscope (FFM) 325, 498, 499, 841
– microscopy (FFM) 537, 631, 632
– Z-15 531
– Z-DOL (BW) 531
friction models
– cobblestone model 572, 574
– Coulomb model 572
– creep model 589
– distance-dependent model 589
– interlocking asperity model 572
– phase transitions model 587, 589, 590
– rate-and-state dependent 590
– rough surfaces model 588
– surface topology model 588
– velocity-dependent model 587, 589
fringe detector 1028
fringing field 1068
frustules 314
fullerene 41, 792
fullerene-like structure 49
fullerite 42
functionalized nanotube 71, 72
fundamental resonant frequency 350
fused silica 699, 700, 704
– hardness 678
fusion bonding 1114

G

GaAs 248
– stamp 191
– wafers 192
GaAs/AlGaAs
– heterostructures 438
– quantum well heterostructures 248
gage marker 1028
gain 465
gain control circuit 389
GaN nanowires 135
gap stability 327
GaP(110) 435
gas separation 79
gate potential
– floating 905

gauche defect 622
Gaussian
– fluctuation–dissipation relation 889, 891
– statistics 884, 891, 892
Gd on Nb(110) 429
Ge membranes 210
gear
– miniaturized 986
generalized actuator 1064
genetic engineering 1142
genomic libraries 756
genuine graphite 40
geometric
– interlocking 724
– nonlinearity 1054
– parameters 228
– restrictions 249
geometry effects in nanocontacts 646
germanium 705
g-factor 424
giant magnetoresistance (GMR) 900
gimballed microactuator design 959
gimbal-mounted mirror 1097
GIO (grazing impact oscillator) 911
glass 619
– bulk etching 255
– frit bonding 1096
– frit sealing 1093
– transition 894
– transition temperature 256, 570, 579, 665
glassiness 593
glass-like behavior 891
glass-to-glass direct bonding 271
glass-to-silicon package 1127
global internal friction 667
glucose oxidase (GOD) 274
GMR
– giant magnetoresistance 900
gold 330
– coated tip 476
– electrodes 27
– film on polysilicon substrate 1051
– nanowires 136
– to-silicon eutectic bonding 1120
Grahame equation 555
grain
– boundary 553, 1035
– boundary scattering 120, 121
– size 189, 793
graphene 40, 63, 667
– defect 70
graphite 415, 417, 507, 638, 651, 792
– cathode 798
– pellet 47
– surface 463
graphite(0001) 427
gratings 200
grazing impact oscillator (GIO) 911
Griffith fracture theory 778
grinding 382
growth
– by pressure injection 102
– of nanotubes 668
guanine 10
gyro
– packaging 1100
– test 1101
gyroscope 230, 232, 1094

H

H_2O 425
Hagen–Poiseuille equation 264
half-generation precursor 287
Hall resistance 438
Hall–Petch behavior 710
Hamaker constant 386, 464, 551–577, 621, 871
Hamilton–Jacobi method 390
hand-held biochip analyzer 274
handling damage 1105
haptenization 313
haptenized nanostructure 313
hard amorphous carbon coatings 795
hard disk drive (HDD) 899–907, 952
– head 901
– performance 903
hard disk drive technology
– limits for 904
hard-core or steric repulsion 545
hardness 504, 770, 800, 802, 805, 809, 1024
harmonic oscillator 454
harpooning interaction 545, 553
HARPSS 166, 169, 170
HDT (hexadecanethiol) 188, 843, 854, 1006, 1012
head
– displacement 953
– gimbal assembly (HGA) 971
– group 839
– position 969
health care 1141
heat
– curable prepolymer 186
– dissipation 1122
– pipe cooling 1123
– transfer model 1119
heater
– cantilever 925
– temperature 934
heavy ion bombardment 441
heavy metal toxicity 306
helicity 41
helper T-cell epitopes 311
hermetic
– packaging 1116
– seal 1112
herringbone
– MWNT 44, 60, 76
– texture 61
Hershey–Chase experiment 755
Hertz model 694
Hertzian contact model 695
Hertz-plus-offset relation 647
heterodyne interferometer 351
heterogeneous CCVD 53
hetero-nanotube 68
hexadecanethiol (HDT) 188, 843, 854, 1006, 1012
HexSil 166, 168, 169, 954
– micro-molding process 962
hierarchical arrangement 108
hierarchy
– of structures 747
– of temporal processes 749
high modulus elastomer 188
high quality manufacturing 1087
high reflectivity 136
high speed scanning 467
high temperature conditions 61
high temperature superconductivity (HTCS) 419, 431, 440
high volume production 1093
high-aspect-ratio
– MEMS (HARMEMS) 986
– tips 376, 377
high-definition television 235
higher orbital tip states 422
highest resolution images 382
highly ordered monolayers 187
highly oriented pyrolytic graphite (HOPG) 337, 507
high-performance material 1141
high-resolution
– electrodes 196
– FM-AFM 436
– image 305
– imaging 407

– MR images 305
– patterning 193
– printing 186
– spectroscopy 422
– stamps 186
– tips 379
high-temperature operation STM 417
Hill coefficient 488
hinged cantilever 1034
HiPCo process 56
HiPCo technique 48
h-MWNT (herringbone) 44, 60, 76
homodyne interferometer 351
honeycomb
– chained trimer (HCT) 395
– lattice 68
– pattern 399
Hooke's law 433, 450, 454, 469, 480
horizontal coupling 329
host immune responses 311
hot embossing 257
HtBDC (hexa-tert-butyl-decacyclene) on Cu(110) 420
HTCS (high temperature superconductivity) 419, 431, 440
Huber-Mises 695
human bone tissue 675
humanity 1136
humidity 645
– effect 1105
hybrid
– carbon nanotube 68
– manufacturing 1093
– nanotube tip fabrication 380
hybridization 41, 397
hydration forces 545–575
hydration regulation 560
hydrocarbon 53, 57, 835
– chains 731
– precusors 799
– unsaturated 836
hydrodynamic
– forces 545, 549
– lubrication 729, 884
– radius 561
hydrofluoric (HF) 832
hydrogen 394
– bonding 10, 561
– bonding force 622
– concentrations 804
– content 804
– end group 841
– flow rate 805
– in carbon nanotube 77
– sensor 137
– storage 77
– storage capacity 78
– termination 394
hydrogenated carbon 800
hydrogenated coating 800
hydrophilic
– control 262
– surfaces 544, 560
– tip 616
hydrophobic
– coating 1014
– control 262
– ferrocene 24
– force 559, 560
– surfaces 544, 560
– tip 616
hydroxyl polymethacrylamide (HPMA) 308
hydroxylated surfaces 29
hydroxylation 840
hydroxyl-terminated perfluoropolyether 892
Hysitron 676
hysteresis 360, 382, 388, 415, 439, 459, 469, 681
– loop 333, 363

I

IB (ion beam)
– coating 822
IBD (ion beam deposition) 796, 798, 815
IC (integrated circuit) 226, 988
– industry 1112
– package technology 1101
– packaging 1112
ideal nanobiological devices 282
IFM (interfacial force microscope) 610
IgG
– repetitive 484
image
– effects 710
– processing software 364
– topography 331, 406
image projection 1100
– system 1084
imaging
– bandwidth 339
– electronic wave functions 427
– signal noise 389, 391
– tools 111
immunoassay
– magnetic bead-based 271
immunogenic synthetic nanomaterials 313
immunoisolation 286, 313
immunosensor 271
immunosuppression 313
immunosurveillance 313
impregnation method 54
imprinted polymers 200
in situ
– environmental TEM chamber 112
– screening 137
– sharpening of the tips 328
in vivo properties 676
InAs 417
InAs(110) 423, 435
incipient wetness impregnation 76
incommensurate surface 725, 731
indentation 498, 526, 923, 932, 935, 936
– creep process 528
– depth 527, 678, 806, 809, 998
– fatigue damage 812
– hardness 325, 527
– induced compression 812
– modulus 679
– size 527, 924
– size effect (ISE) 704
– technique 765
independent molecular operators 19
indirect sensor-device coupling 304
indirect transition in bismuth 129
indium tin oxide (ITO) 193
induction type EHD 262
inductive heating 1119
inelastic tunneling 419
inertial sensor 229
infected cell 756
influence of humidity
– on adhesion 850
information technology 1139, 1142
initial contact 703
injection type EHD 262
ink
– electronic 195
ink jet printing 193
inorganic-organic solar cells 136
InP(110) 392
in-phase signal 944
in-plane
– actuation 929
– mechanical strain 188
instability position 461
instability, jump-to-position 388, 610

instrumented suspension 967
insulator 397
insuline pump 269
integrated
– circuit (IC) 226, 988
– MEMS 259, 1084, 1092
– MEMS accelerometer 1091
– micromachining process 1117
– silicon suspension 967
– suspension 967
– tip 341, 373
– vacuum sealing 1117
Intel 915
intellectual property protection 1137
interaction
– energy 757
– force 434, 757
– potential 718
interatomic
– attractive force 344
– bonding 393
– force 332
– force constants 387
– interaction 1035
– spring constant 332
interbulk interactions 728
intercalated MoS_2 833
intercellular adhesion molecule-1 (ICAM-1) 489
intercellular concentration 748
intercept length 855
interchain hydrogen bonding 621
interchain interactions 484
interconnect 900
interconnect technology 1094
interconnected operators 18
interdiffusion 568
interdigitation 568
interface sensor components 302
interfacial
– chemistry 607
– defects 813
– energy (tension) *see* surface energy (tension)
– force microscope (IFM) 610
– friction *see* also boundary lubrication, 571–574, 579, 580, 585, 862
– liquid structuring 888
– potential 726
– stress 819
– structuring 888
– wear 606
interferometeric detection sensitivity 353
interlocking macrocycles 26
intermediate or mixed lubrication 569, 570, 582, 583
intermittent contact mode 373, 460
intermolecular communication 18
intermolecular force 289, 481, 484, 606
internal friction 935
internal stress 1025, 1041
intersymbol interference 947
intertube distance 43
intraband transitions 423
intramolecular
– forces 484
– self-assembly 293
intrinsic
– adhesive force 832
– damping 469
– mean stress 1045
– stress 804
iodine 419
iodobenzene 419
ion
– correlation forces 554, 557
– displacement 398
– implantation 205, 528, 925, 996
– implanted silicon 997
– plating techniques 795
– source 798
ion beam
– deposition (IBD) 796, 798
– etching 248
– sputtered carbon 798
ion correlation forces 545
ionic bond 545
ionic strength 619
Ising spin model 892
isolated nanotubes 381
isotropic etching 189
isotropic layer 1050
itinerant nanotube levels 425
ITO (indium-tin-oxide) 195
I–*V* characterization 116, 124

J

JKR (Johnson–Kendall–Roberts)
– model 619, 621, 695
– relation 646
– theory 566
Joule
– dissipation 655
– heating 439
jump length distribution 889
jump-to-contact 388, 436, 452
junction area 132, 135

K

KBr 397
KBr(100) 640, 650
$KCl_{0.6}Br_{0.4}$(001) surface 398
Kelvin
– equation 564, 612
– radius 613, 614
Kelvin probe force microscopy (KPFM) 407, 503
ketones 837
key scale at bio-nano interface 746
kidney exclusion limit 296
kinesin 741, 745
kinetic
– friction 571–573, 583, 585–587, 589, 590, 593, 718, 728, 729, 884, 885
– friction force 722
– friction ultralow 591
– measurements 892
– off-rate 487, 488
– processes 703
– rate constant 482, 486
kinetics 883
– of capillary formation 614
Kirchoff hypothesis 1047
knife-edge blocking 361
Koanda effect 266
KOH 330
Kohlrausch
– exponent 892
– relaxation 884
Kondo
– effect 414, 425
– temperature 425
K-shell EELS spectra 802

L

L amino acid 291
lab-on-a-chip 270, 988
– application 270
– concept 253
– disposable plastic 993
– systems 265
– technology 255
lac repressor 754
Lagrangian approach 1064
Lamb waves 693
laminar
– flow 264
– flow characteristics 266
laminated n-channel transistor 196
Landau
– levels 122, 415, 438

– quantization 424
Langevin equation 891, 892
Langmuir–Blodgett (LB) 861
– deposition 536, 833
– film (ethyl-2,3-dihydroxyoctadecanoate) 467
– methodology 35
Laplace
– force 844, 868
– pressure 544, 564, 613–618, 832
– pressure force 616
large angular deflection 1069
large array of nanowires 109
larger-scale systems 1136
large-scale stick-slip motion 733
Larmor frequency 305
laser
– ablation method 61
– beam deflection method 453
– deflection sensing 453
– deflection technique 333
– devices 46
– interference lithography 186
– vaporization method 61
– welding 1119
laser-to-fiber coupler 236
lateral
– accelerometer 231
– contact stiffness 631
– cracks 693
– deflection 671
– deflection signal 1016
– displacement 669
– force 632, 643, 669, 718, 1015
– force calculation 354
– force microscope (LFM) 325, 499, 632
– force microscopy (LFM) 187, 404, 548
– friction 453
– resolution 328, 331, 477, 500, 606, 663
– resonator 1037
– scanning range 929
– spring constant 338, 636
– stiffness 362, 376, 649
lattice
– constant 725
– imaging 332
– vibration 726
layer of resist 186
layered structure 1045
LCC (Leadless Chip Carrier) 1101
leakage 1126
– rate 263
learning theory 1143
length scale 722, 747
Lennard-Jones potential 436, 457, 719, 730
Lennard-Jonesium 727
leukocyte function-associated antigen-1 (LFA-1) 489
leveraged bending 1070
levers 932
lever-sample displacement 608
levitated copper inductor 242
Lévy flight model 891
LiF 397
LiF(100) surface 652
life-cycle costs 1137
lifetime broadening 419, 423
lifetime-force relation 485
Lifshitz theory 551, 552, 554
lift mode 338, 503
LIGA 166–168, 986
– fabrication of MEMS 987
– technique 765
ligand concentration
– effective 485
ligand-receptor affinity 288, 316
ligand-receptor interaction 476
light beam deflection galvanometer 353
light emission 135
light-emitting diode (LED) 132, 135
limits for hard disk drive technology 904
limits of biotechnology 281
linear
– creep model 890
– fractional transformations (LFT) 975
– friction–velocity dependence 889
– quadratic Gaussian (LQG) 971
– recording density 947
– stress gradient 1045
– variable differential transformers (LVDT) 361
lipid 747
lipid film 644
liposome 742
liquid
– capillaries 612
– capillary condensation 844
– film meniscus 612
– film thickness 537, 885
– helium 415
– helium operation STM 417
– lubricant 530, 832
– lubricated surfaces 572, 576, 578, 587
– mediated adhesion 832
– nitrogen (LN) 416
– perfluoropolyether (PFPE) 1003
– solid interface 832
– vapor interface 832
load
– critical 809
load contribution to friction 572–574
load dependence of friction 647
load-carrying capacity 824
load-controlled 610
load-controlled friction 571, 579
load-displacement 528
load-displacement curve 774, 778, 808, 809
loading curve 701
loading rate 483
local deformation 504, 518
local deformation of material 524
local density of states (LDOS) 422
local mechanical spectroscopy 664
local pinning 727
local shear stress 731
local stiffness 362, 1006
localization
– effect 122–125, 134
– theory 120
localized heat bonding 1116–1120
logarithmic-like friction–velocity behavior 885
logic
– devices 132
– gates 14, 15, 132
– protocols 15
Lognormal distribution 1127
London
– dispersion interaction 545, 550, 552
– penetration depth 431, 440
longitudinal
– magnetoresistance 120, 121
– piezo-resistive effect 356
– relaxation time (T_1) 305
long-range
– attraction 550, 562
– elastic deformation 722
– force 389, 398, 437
– tip-molecule force 406
long-term
– measurements 414
– memory 734
– reliability 1126
– stability 414, 948, 1102

– test 1125
loop-transfer recovery (LTR) 971
loss modulus 702
lot-to-lot variation 1072
low cost gyroscope 1094
low pressure chemical vapor deposition (LPCVD) 770
low temperature
– AFM/STM 332
– condition 60
– NC-AFM 402
– SFM (LTSFM) 417
low-cycle fatigue resistance 1032
lower lasing threshold 135
low-noise measurement 636
low-temperature microscope operation 414
low-temperature scanning tunneling spectroscopy (LTSTM) 419
low-temperature SPM (LTSPM) 414
LPCVD (low pressure chemical vapor deposition) 770, 962, 1034
– nitride sealing 1117
LTR
– loop-transfer recovery 971
LTSFM 417
LTSTM 419
lubricant 723
– atoms 732
– film 618
– meniscus 864
– spreading 864
– thickness 885
lubricated sliding 884
lubrication 717
– elastohydrodynamic 569, 581, 583
– intermediate or mixed 569, 570, 582, 583
– method 995
lumped
– parameter 1065
– parameter model 1063, 1073
Luttinger liquid 66
LVDT
– linear variable differential transformers 361

M

macrocyclic polyether 13, 25
macromolecular
– building block 747
– interaction 475
macromolecules as elastic rods 753
macroscale friction 842, 998
macroscopic
– building blocks 10
– friction 732
magnetic
– actuators 259
– applications 137
– dipole interaction 403
– disk 824
– disk drive 794
– information 138
– Ni 431
– ordering 129
– particles
self-ordered arrays 914
– quantum flux 440
– recording 911, 922
– storage device 795, 831
– tape 466, 513, 524, 862
– thin-film head 825
– tip 465
magnetic field
– critical strength 122
– length 122
– microscopy (MFM) 114
– triggering 299
magnetic force 386
– gradient 338
– microscope (MFM) 338
– microscopy (MFM) 326, 439
magnetic resonance
– force microscopy (MRFM) 441
– imaging (MRI) 304
magnetically coated tip 490
magnetohydrodynamic (MHD) pumping 261
magnetomotive transduction 248
magneto-optics 128
magnetoresistance 120
– longitudinal 120, 121
– transverse 121
magnetoresistive RAM (MRAM) 899, 906
magnetostatic interaction 439
magnetron sputtered carbon 799
major histocompatibility complex (MHC) 311
manganites 431
manifold absolute pressure (MAP) 1091
manipulation of individual atoms 325
manually assembled MWNT tips 378
manufacturability 1087
manufacturing “risk” change 1093
MAP sensor 1094
marginal dimension 727
martensite 665
mask, high resolution stamp 186
mass
– effective 433
mass of cantilever 331
master–slave design 970
material
– for microfluidic devices 254
– hardness 765
– molecule-based 35
– property variation 1072
– structural 1028
Matthiessen rule 120
maximum likelihood estimator (MLE) 1126
MBE (molecular beam epitaxy) 152, 172
MBI (multiple beam interferometry) 547
mean time to failure (MTTF) 1127
mean-field theory 556
measurement of hardness 505
mechanical
– coupling 588
– dissipation in nanoscopic device 631
– instability 452, 609
– micropumps 260
– protection 1112
– relaxation 457, 848
– resonance 591
– spectroscopy 664
– stability 927
– surface relaxations 468
– wear 263
mechanical properties 764, 1023
– characterization 1036
– of bone 675
– of carbon nanotubes 668, 672
– of DLC coating 813
mechanically cut tips 382
mechanics of cantilevers 347
mechatronic device 952
media flatness 910
mediator molecules 304
medium
– effective 127
medium theory
– effective 127
melting freezing transition 894
melting point of SAM 853
membrane
– deflection method 1037
– fabrication of 314

– proteins 375
membrane-embedded machine 741
membranes as elastic media 754
memory
– cells and switches 116
– distance 588
– long-term 734
MEMS 147–171, 225, 270, 376, 606, 617, 764, 795, 831, 862, 901, 983, 1023, 1084
– accelerometer manufacturing 1091
– accelerometers 230
– applications 985
– capacitive switch 244
– components 1011
– device 247, 987, 990
– device operation 994
– fabrication 1111
– fiber optic switching 238
– industry 1093
– integration 1102
– manufacturing 1086
– microactuator (MA) 952, 971
– mounting condition 1105
– packaging 1111, 1116, 1128
– packaging requirement 1112
– passivation 1103
– post-packaging 1114, 1117
– pressure sensor production 1090
– production economics 1086
– resonators 244
– sensor 229
– standardization 1086
– storage device 907
– structure 1112
– surface 1104
– switch, metal-to-metal 244
– technology 233, 237, 249, 1034
– testing 1099
– tribological problems 989
– tunable capacitors 241
– vacuum packaging 1117
– wafer bonding 1095
– wafer protection 1097
MEMS/NEMS 530
MEMS-based
– microfluidic system 270
– storage concept 904
– storage device 922
meniscus
– bridge 831, 876
– force 516, 871
– of liquid 564
mercaptopropyltrimethoxysilane (MPTMS) 191
merocyanine 18
mesoporous molecular sieves 102
mesoscale osmotic actuator 269
messenger RNA 750
metabolic facility 282
metal 103
– catalyst 379
– cavity packaging 1099
– cluster agents 306
– electroplating 255
– evaporated (ME) tape 793
– matrix composites 81
– nanowires 69
– oxide 397, 400
– particle catalyst 54
– porphyrin (Cu-TBPP) 404
– seed layer 966
metal/insulator/metal (MIM) 197
metal-catalyzed
– chemical vapor deposition (CVD) 379
– polymerization method 287
metal-deposited Si surface 395
metallic nanotube 34
metallic SWNT 66
metal-organic compound 56
metal-particle (MP) tape 825
metal-to-metal MEMS switch 244
metastable lubricant systems 894
methylene stretching mode 842
methyl-terminated SAM 843
Meyer's law 708
MFM sensitivity 338
MgO tip 398
MgO(001) surface 403
Mg-terminated tip 398
MHD (magnetohydrodynamic) pumping 261
mica 102, 337, 615, 619, 640, 888
– films 100
– muscovite 615, 650
– surface 641
micelles 291, 292
micro/nanofabrication 147–150, 154
microactuator 990
– design 960
– fabrication requirements 962
– loop 972
– shape 965
microbially-derived antibiotics 293
microcantilever 994
microchannel 260
microcomponent 989
microcontact printing (μCP) 187, 834
microcrystalline graphite 803
microdevice 1026
microdispenser 266
microelectromechanical
– motor 226
– resonators 244
– switch 244
microelectronics 397, 749
micro-electro-rheological valve 260
micro-engine 908
microfabricated
– cantilever 372, 453
– silicon cantilevers 387
microfabrication 187, 226, 249
– methods 285
– techniques 285, 372
microfluidic 253
– applications 255
– channel 254, 269
– channels on silicon substrates 254
– control 260
– line 258
– mixers 265
– motherboards 270
– multiplexer 268
– network 268
– passive 262
– pumping 260
– structures 256
microfluidic device 269
– disadvantages 270
– materials for 254
– passive 257, 269
microfluidic dispenser
– passive 273
microfriction 511
micro-hinge technology 236
microindentation 689
micro-injection molding 273
micromachined
– high-Q resonators 245
– inductors 242
– notch 1032
– silicon 310
micromachining 147, 150, 151, 154, 155, 159–163, 166–168, 170
micromanipulator 988
micromechanical
– cantilevers 303
– device 1084
– structures 204
micromirror 992, 1040
– array 237
micromixer category 265

micromotor 1016
micron-sized channels 265
micropatterned SAM 848
micropipette aspiration 549
micropump 260
– passive 269
micro-Raman spectroscopy 693
microrelay 1062
micro-resonators 244
microscale
– diffusion 265
– friction 1006, 1010
– material removal 521
– scratching 503
– silicon gate 34
– wear 503, 520
– wear test 1008
microscanner 929
microscope eigenfrequency 348
microscopic friction 723
microscopic origin of friction 889
microscratch 813
– test 675
microscratching measurement 519
microsensors 229
microstrain gauge 1025, 1034
microstructure 1040
microsystem 249
micro-total analysis systems (µTAS) 257
microtransformer 197
microtriboapparatus 1011
microtubule 740
– bending 673
– buckling 673
microvalve 258
– design 260
microwear 813, 819
midplane 1046
– curvature 1048
– strain 1048, 1049
milligear system 989
millipede 901, 994
– chip 931
– system concept 921, 923, 943
MIM capacitors 197
MIMO design 971
miniaturized
– actuator 986
– connector 986
– gear 986
– motor 986
minimal detectable depression 374
minimally invasive screening 285
minimum-indentation pitch 937
mirror/yoke assembly 1089
misfit
– angle 638
– strain 1042
mismatch
– elastic 1051, 1056
mismatch of crystalline surfaces 566, 567, 593
mitochondria 747
mixed C-N nanotube 56
mixed lubrication *see* intermediate or mixed lubrication, 884
mixing performance 265
MMA air bag sensor 1092
Mn on Nb(110) 429
Mn on W(110) 432
m-nitrobenzylidene propanedinitrile (*m*-NBP) 916
MOCVD (metallorganic chemical vapor deposition) 152, 172
mode coupling gratings 199
model potentials 719
modeling
– at the nanoscale 752
– mechanics of systems 752
– tribological process 753
modular multifunctionality 280
modulation codes 946
modulus
– elastic 664, 698, 767, 770, 774, 805, 806, 809
modulus of elasticity 325, 516
molarity 749
molding 186
molecular
– AND gate 17
– assembly 26
– beam epitaxy (MBE) 988
– building blocks 12, 35, 747
– chain 839
– conduction band 33
– conformation 864
– cooperation 886
– dynamics (MD) 476, 718, 749
– dynamics (MD) calculation 639
– dynamics simulation (MDS) 488, 622, 631, 651, 753
– engineering 281
– interaction 879
– interconverting state 15
– logic gates 15
– lubrication 885
– machines 13
– mobility 894
– motor 13, 673, 740
– NOT gate 16
– placement by force microscopy 285
– precursor 58
– pump 741, 748
– recognition force microscopy (MRFM) 476
– recognition force spectroscopy principles 482
– resolution 421
– reversible transformation 15
– scale layer 833
– shape 558, 559, 586, 591, 592
– shuttle 14
– spring 536
– spring model 843, 848
– stiffness 849
– switch 14, 35
– theory of lubricated friction 890
– wires 289
molecule-based
– device 24, 28
– materials 35
– switches 15
moment of inertia 1030
monodispersity 286
monolayer 885
– switches 28
– thickness 855
monolithic accelerometers 231
monolithic high-Q MEMS 245
Monte Carlo simulation 848
MoO_3 film 614
Moore's law 1142
Morse potential 464
MoS_2 friction 640, 641, 668
motion
– unperturbed 390
motor
– actin-myosin 751
– flagellar 750
– micro 1016
– microelectromechanical 226
– miniaturized 986
– molecular 13, 752
– myosin 741, 745
– oil 884
– voice coil 952, 969
movement-constrained macromolecules 305
moving media model 908
moving-tips model 908
MRAM (magnetoresistive RAM) 899–907
– cell 906
MRFM

– magnetic resonance force microscopy 441
MRI contrast agents 305
MTTF (mean time to failure) 1126
multicomponent nanodevice 288
multi-input-multi-output (MIMO) 956
multilayer 709
– devices 197
– Fe-N/Ti-N 675
– structures 104
– thin-film 825
multimode AFM 338
multiple beam interferometry (MBI) 547
multiple cantilever 907
multiple therapeutic nanodevice 294
multiple-asperity contact 517
multiplexed sensor arrays 304
multiplication of dislocations 703
multivalent targeting strategy 298
multiwall carbon nanotubes (MWCNT) 43–49, 289, 342, 376, 667, 727
muscle contraction 740
muscovite mica 615, 650
μ-synthesis 974
μ-synthesis controller design 975
MWCNT
– multiwall carbon nanotubes 43–49, 289, 342, 376, 667, 727
MWNT (multiwall nanotube) 64
– based catalyst-support 76
– bc (bamboo-concentric) 44
– bh (bamboo-herringbone) 44
– bunches 58
– catalytically grown 673
– composites
  MWNT-Al composite 81
– flexural modulus 67
– purified 378
– surface area 64
– tips
  manually assembled 378
– Young's modulus 669
myosin motor 751

N

N acceptors 10
Na on Cu (111) 421
NaCl 397
– island on Cu(111) 655
– islands 468
– thin film on Cu(111) 398
NaCl(001) 397
NaCl(100) 641–643, 648
NaF 397, 640
NAND gate 15
nanites 1146
nanoasperity 515, 695
nanobeam 669
– array 766, 774
nanobiological
– applications 313
– design 312
– devices 281
– vaccine design 312
nanobiotechnological device 300
nanobiotechnological therapeutic device 313
nanochannel glass 100, 102
nanochemistry 984
nanocluster 400
nanocomposite
– coating 833
– materials 28
nanocrystallites 800
nanodeformation 864, 880
nanodevice 85, 281
– multiple therapeutic 294
nanoelectrodes 32
nanoelectromechanical systems (NEMS) 226, 246, 606, 764, 901, 984
nanoelectronic device 32, 34
nano-electronics 84
nanofabrication 147, 170, 175, 187, 326, 526
– techniques 253
nanofatigue 810
nanofiber 44, 53
nanofilament 44, 60
nanogaps 31
nanohardness 695, 804, 805, 1000
nanohelix 12
nanoimprint 171
nanoindentation 674, 688, 697, 805, 1024
– curve 677
– measurement 504
nanoindenter 765, 769, 818, 1028
nanolubrication 883
nanomachine 1147
nanomachining 325, 326
– techniques 205
nanomagnetism 431
nanomania 1146
nanomaterial 662
nanomechanical Si beams 247
nanometer resolution 186
nanometer servo precision 979
nano-objects 68
nanoparticle 49, 374
– contrast agents 305, 306
nanoparticulate carriers 310
nanophobia 1146
nanoporous membranes 285, 314
– microfabrication 286
nano-positioning 359
nanorheology 584
nanoscale
– antimicrobials 293
– architectural property 283
– assemblers 1146
– biomolecules 10, 12
– control of molecular events 310
– data bit 909
– drug delivery 291
– electrodes 23, 27
– FET 117
– film morphology 917
– interventions 285
– mechanical properties 661, 662
– molecular therapeutic device 308
– molecules 35
– Schottky barrier 34
– supramolecular assembly 407
– therapeutic device 281, 315
– therapeutic platforms 284
– wear 518
Nanoscope I 328
nanoscratch 522
– studies 765
nanostructures 147, 170, 171, 179
– containing proteins 307
– mechanical properties 764, 1023
nanoswitches 15
nanotechnological
– actions 755
– therapy 307
nanotechnology 280, 918, 1139
– definition 1136
– ethical issues 1136
– in the living world 740
– possible applications 1136
– recommendations 1138
– role of social sciences 1148
– social issues 1136
– unintended consequences 1143
nanotherapeutic 280, 288, 303
– contrast agents 304, 310
– delivery device 298
– design paradigms 282
– device application 307
– device components 285
– devices in oncology 307

– drug delivery 311
– triggering external stimuli 299
– vaccines 312
nanotransfer printing (nTP) 190
nanotransistor 32
nanotribology 717, 862
nanotube 12, 105, 133
– AFM tips 379
– buckling 377
– chemistry 67
– defects 47
– functionalization of wall 72
– functionalized 71, 72
– growth 45
– length 382
– morphology 49
– nucleation 45
– oxidation 72
– resistance 34
– strain 671
– surface energy 380
– tip fabrication 380
– tips, surface growth 380
– transistor 34
– volume fraction 83
– yield 49, 59
nanotube production 56
– efficiency 49
– electric-arc method 48
– laser ablation 46
– solar furnace 52
– techniques 57
nanotube-based
– emitter 74
– hybrid material 71
– sensor 75
– SPM tip 73
nanotube-Co-MgO 81
nanotube-containing
– composite 82
– electrical device 289
nanotube-Fe/Co-MgAl$_2$O$_4$ 81
nanotube-Fe-Al$_2$O$_3$ 81
nanotube-polymer composite 83
nano-tweezer 85
nanowire 31
– applications 131, 138
– metal 69
– nucleation 110
– photodetectors 135
– semimetallic 118
– stress-induced crystalline 107
– superlattice 132
– Zn 120, 123
– ZnO 135
nanowire device
– two-terminal 132
naphthalene channel 21
National Nanotechnology Initiative (NNI) 1137
native oxide layer 868
Nb superconductor 423
NBIC convergence 1140
NbSe$_2$ 431, 649
NC-AFM (noncontact atomic force microscopy) 397, 433
near-field
– method 903
– scanning optical microscopy (NSOM) 128, 372
– technique 434
near-surface mechanical properties 665
necrotic domains 307
negative logic convention 26
negative magnetoresistance 121
neighbouring indents 678
NEMS 831, 983, 988, 1023
– devices 246
– nanoelectromechanical systems 226, 246, 606, 764, 901, 984
neocarzinostatin (NCS) 309
nested-arc structure 680
net electrostatic torque 1066
net fluid transport 261
newspaper
– electronic 196
Newton's first law 454
Newtonian flow 570, 580, 581, 583
Newton–Raphson method 753
n-hexadecane 888
N-hydroxysuccinimidyl (NHS) 478
Ni(001) tip on Cu(100) 653
Ni(111) tip on Cu(110) 653
Ni, Co catalyst 47
Ni/Y catalyst 49
Ni^{2+}-chelating 485
nickel-iron permalloy 259
Ni-Fe valve membrane 259
NIH3T3 fibroblast cell 489
NiO 402, 417
NiO(001) 397, 400, 402
NiO(001) surface 403
Ni-P
– beam 785
– cantilever 778
– film 770, 771
– microbeam 777, 785
NiTi 665
– transformation behavior 666
nitrogenated carbon 800
NMP (no moving part) valves 265
noble metal surfaces 424
n-octadecyltrichlorosilane (n-C$_{18}$H$_{37}$SiCl$_3$, OTS) 616, 842
n-octadecyltrimethoxysilane (OTE) 615
noise 387, 391, 453
– 1/f 387, 957
– electronic 333
– external 415
– performance 391
– signal 941
– source 355
– vertical 391
Nomarski-interferometer 352
nonconducting film 798
nonconductive
– materials 404
– sample 329
– surface 462
nonconforming contacts 890
noncontact
– dynamic force microscopy 654
– friction 438, 655
– imaging 331
– mode 373, 388, 909
noncontact atomic force microscopy (NC-AFM) 374, 385, 391, 393, 397, 404, 654
nondestructive contact-mode measurement 652
nondimensional curvature 1051
nonequilibrium interactions *see* dynamic interactions
nonequilibrium simulations 720
noninvasive device 283
nonlinear
– force 460
– geometry effect 1058
– *I*–*V* behavior 124
– material behavior 1058, 1059
– optical property 129
– spring 491
– strain-displacement relation 1055
nonliquidlike behavior 584
nonmagnetic Zn 431
non-Newtonian behavior 885
non-Newtonian flow 570, 581, 583
nonpolar group 838
nonspherical tip 647
nonsymmetrical passive valves 268
nontherapeutic nanobiological device 283
nonvolatile random access memories (NVRAM) 899–913
nonwetting 560
NOR gate 15

normal friction 569, 577, 579, 580
normal load 721, 866
– external 867
normalized frequency shift 463
no-slip boundary condition 729
NOT gate 15
N-succinnimidyl-3-(S-acethylthio)propionate (SATP) 478
NTA (nitrilotriacetate) - His$_6$ 478
nTP (nanotransfer printing) 197
nuclear weapon 1140
nucleation of nanowires 110
nucleic
– acid (NA) 294, 299
– acid hybridization 294, 300
nucleotides 747
n-undecyltrichlorosilane (n-$C_{11}H_{23}SiCl_3$, UTS) 842

O

O acceptors 10
$(OCN)^-$ 398
octamethylcyclotetrasiloxane 888
ODD performance 903
odd-even effect 621
off-track error 958
OH-terminated tip 398
on-chip
– actuator 1031
– electronic detection 231
– integration 210
– microvalves 260
– spiral inductors 242
oncological nanotherapeutics 291
one-dimensional quantum effect 115
one-dimensional system 100
open loop architecture 1093
operation 387
opsonized particle 296
optical
– absorption 129
– applications 134
– beam deflection 607
– cavity 135
– data switching 237
– deflection systems 372
– detector 331
– disk drives (ODD) 899, 900
– gap behavior 127
– head 336
– head mount 337
– integration 236
– MEMS 233
– output 16
– property 126, 127
– reflection 127
– switch 135
– transmission 127
– trap 480
– tweezers (OT) 480, 549, 744
optical lever 353
– angular sensitivity 355
– deflection method 362
– optimal sensitivity 355
optimal beam waist 355
optimum sampling phase 946
optoelectronic
– components 198
– module 236
– switch 136
OR
– operation 23
– operator 15
order
– in-plane 567, 584, 585, 593
– long-range 570
– out-of-plane 557–559, 567, 570, 584, 585, 591
– parameter 570, 586
ordering nanowires 108
organelles 281
organic
– compounds 835
– device structure 917
– film 916
– inverter circuit 196
– technology 916
– transistor 193, 196
organics 915
– data storage 916
organized molecular arrays 23
organometallic vapor phase epitaxy (OMVPE) 988
Orowan strengthening 710
orthogonal conjugation 295, 296
orthotropic Young’s modulus 1048
oscillating tip 334, 469
oscillation
– amplitude 389, 453, 454, 457, 461
– loop 462
oscillatory force 544, 545, 557–561, 567
osmosis response time 269
osmotic
– interactions 545, 555, 562, 564
– pressure 555, 562, 575
– stress technique 549
OTE (octadecyltrimethoxysilane) 615
OTS (octadecyltrichlorsilane) 848, 1075
OUM cell 906
out-of-plane stiffness 959
overall barrier height 887
overdamped SFM system 891
overdrive capability 1062
overwriting mechanism 937
Ovonyx 915
Ovonyx Unified Memory (OUM) 906
oxide layer 703
– amorphous 106
oxide-enhanced nanowire growth mechanism 107
oxide-sharpened tips 375
oxidized nanotube 72
oxygen
– content 826
– sensor 274

P

package level integration 1093
packaging 248, 249
– cap 1112
– equipment 1099
– induced thermal stress 1123
– reliability 1125
– uniform 1112
packing density
– chains 842
palladium nanowires 137
PAMAM (polyamido amine) dendrimer 286, 287
p-aminophenyl phosphate (PAPP) 271
paper-like displays 193
paraboloid load displacement 697
paraffins 835
parallel-plate
– actuator 955
– capacitor model 1064
– sensing 957
paramagnetic particles 305
parasitic charging 1077
particle
– contamination 1098
– control 1098
– opsonized 296
– track-etched mica films 103
particle model
– embedded 894
Paschen curve 1076
passive
– check valves 262

– linearization 361
– microfluidic device 257, 269
– microfluidic dispenser 273
– microfluidics 262
– micromixers 265
– micropump 269
– microvalves 257, 262
– structure 1025
– valve geometry 264
patient-monitoring systems 274
pattern of ink 186
pattern transfer 191
patterned
– conductive substrate 110
– film 1053
– self-assembled monolayer (SAM) 188
– silicon chip 25
patterning 1075
Pb on Ge(111) 428
PbI_2 filled SWNT 70
PDP (2-pyridyldithiopropionyl) 478
peak indentation load 810
peapods 71, 289
PECVD (plasma enhanced chemical vapor deposition) 833, 926, 1009
– carbon sample 800
PEG (polyethylene glycol)
– crosslinkers 479
pegylated molecules 313
Peierls instability 428
per molecule adhesion force 620
perfluorodecanoic acid (PFDA) 992
perfluoropolyether (PFPE) 530, 861
perfluoropolyether Z-DOL 614
performance of electromechanical actuator 1075
performance parasitics 1098
periodic
– force 725
– potential 637, 891
permanent dipole moment 407
perpendicular scan 344
persistence length 488
perturbation approach 390
perylene 404
PES (photoemission spectroscopy) 956
– generation 944
– loop compensator 973
PFPE (perfluoropolyether) 864, 1012
– lubricant 862, 868, 1003, 1016
– OH 892
pH of water environment 619, 621
pH sensor 137
phage display 297, 298
pharmacoeconomics 285
phase
– angle image 530
– breaking length 120
– curve 458
– images 470
– lag spectrum 666
– measurement 471
– signal change 458
– transformation 690, 705
phase change
– material 915
– medium 922
– RAM (PC-RAM) 899, 906
phenol 837
phenoxy-centered HOMOs 26
phonon
– confinement effects 125, 127, 131
– excitation 468
– scattering 120
phosphorus-doped polysilicon 206
photoacid 19
photoactive
– fragments 14
– solids 22
– stopper 14
photo-conductance 128
photo-crosslinker 478
photodetector sensitivity 766
photoemission spectroscopy (PES) 423
photogenerated hole 30
photoinduced electron transfer 16, 17
photoinduced proton transfer 19
photolithography 285
– near-field conformal 187
– proximity mode 186
photolithography patterns 186
photo-luminescence (PL) 127, 135
– optical imaging 128
photoresist 188, 190, 1095, 1121
photothermal effect 465
physical properties of nanowires 110
physical vapor deposition 104
physisorbed protein layer 476
physisorption 861
pick-up
– SWNT tips 382
– tip method 381
pico-slider 960
piezo
– ceramic material 360
– effect 359
– element 461
– excitation 454
– hysteresis 415
– relaxation 415
– stacks 361
– tube 359
piezoelectric
– actuation 259
– actuator 908
– drive 327
– leg 636
– positioning elements 414
– scanner 372, 415, 676
piezoresistive
– cantilever 355
– coefficients 356
– detection 355, 387
– pressure sensor 1089
– sensing 958
– sensor 227
piezoscanner 363, 382
piezotranslator 363
piezotube calibration 329
pigment particles 195
pile-up, nanoindent 693, 699, 700, 709, 711
– characteristic 935
piling-up behavior 674
pilot signal 945
pinning 441
– forces 727
pin-on-disk tribotester 841
PL (photoluminescence) NSOM imaging 135
plane-to-plane separation 12
plasma
– etch 286
– frequency 127
– resonance 127
plasma enhanced chemical vapor deposition (PECVD) 197, 339, 770, 795, 799
plasminogen activator (PA) 302
plasmon mode 421
plastic
– asperity model 884
– circuit 193
– contact regime 517
– deformation 522, 728, 812, 854
– electronics 197
– fluidic chips 273
– large area circuit 194
– packaging 1099
plasticity 1060
platelet nanofiber 60

platform 929
– fabrication process 930
platinum-iridium 330
– tip 29
PMMA (poly(methyl methacrylate)) 247, 257, 671, 924, 933, 936, 994
p-n
– diodes 116
– junction diode 132
– junctions 132, 135
pneumatic
– actuation 258
– damping 418
p-nitrophenolate 19
point probes 688
Poisseuille flow 885
Poisson statistics 620
Poisson's ratio 698, 1047
polar
– group 838
– lubricant 832, 861
polarization dependence 129
polishing 793, 1000
poly(amido) amine (PAMAM) dendrimer 286
poly(ethylene glycol) (PEG) 477
poly(ethyleneterephthalate) (PET) 192
poly(methyl methacrylate) (PMMA) 247, 257, 671, 924, 933, 936, 994
polyaromatic shells 63
polycarbonate (PC) 257
polycrystalline
– films 210
– graphite 801
– Si 226
– SiGe 210
polydimethylsiloxane (PDMS) 187, 834
polyethylene (PE) 257
polyethylene co-propylene 886
polyethylene terephthalate (PET) 524
poly-Ge
– mechanical properties 210
polymer 255, 610, 664, 910
– blend 471, 662
– brush 563, 575, 581
– chemistry 286, 291
– conjugation (pegylation) 296
– fluids 892
– intermediates 287
– matrix composites 82
– medium 932
– melt 886
– membranes 103
– microfabrication techniques 255
– microstructures 257
– mushroom 562
– nanojunction 32
– properties 263
– therapeutics 307
– transparent microcapsules 195
polymeric
– liquids (melts) 561, 581, 583
– magnetic tape 526
polypeptides 10
polypropylene 404, 471
polysilicon 963, 967, 1010, 1040
– beam 1044
– fatigue 1036
– film 229, 995, 1027
– fracture strength 1036
– fracture toughness 1035
– friction 1036
– layer 263
– mechanical properties 1034
– MEMS 206
– microstructure 1035
– residual stress 1034
– resonators 245
– strip 245
– surface micromachining 210
– Young's modulus 1035
polystyrene (PS) 257, 936
polysulfone 936
polytetrafluoroethylene (PTFE) 639
polyurethane 471
poor mixing 889
pop-ins 690
pop-outs 690
population explosion 1141
porous
– silica or alumina 120
– vycor glass 102
position
– accuracy 363
– error 972
– error signal (PES) 943, 968
– sensing resolution 957
positioning 109
positive logic convention 18, 26
post-packaging process 1118
potential barrier 887
potential energy
– total 720
power dissipation 237, 470, 654
power to weight ratio 752
PQ design method 970
Prandtl–Tomlinson model 725–727
pre-crack length 1027
precursor 56, 69, 70
– gases 210
preferential flow direction 261
preparation improvements 54
pressure 871
– injection 104
pressure sensor 227
– fabrication process 1090
– packaging 1099
pressure-induced phase reconstruction 888
primary minimum adhesion 544, 555, 559
printed
– coil 197
– DFB resonator 200
probability density functions 891
probe
– FIB-milled 342
probe storage 901, 907, 908
– data organization 909
– operating parameters 909
probe tip 372
– characterization 611
– performance 374
probe-sample distance 457, 462
probe-surface distance 455
process
– control 1088
– flow 1088
– gas 799
process coherence volume 887
processing conditions 44
prodrugs 298
product architecture 1088
production efficiency 47
programmable microfluidic systems (sPROMs) 266
programmed assembler 1146
projection mode photolithography 186
propagation of cracks 524
properties of a coating 795
protection of sliding surface 831
protein 477
– bioactivity 294
– coupled to nanomaterials 294
– folding 293
– primary structure 293
– secondary structure 293
– synthetic chemistry 294
– tertiary structure 293
protofilament 673
prototyping of biological nanodevices 282
protrusion force 545, 564

P-selectin glycoprotein ligand-1 (PSGL-1) 488
PSG (phosphorus-doped glass) 1117
Pt alloy tip 330
PTCDA 404
PTFE (polytetrafluoroethylene)
– coated Si-tip 640
– film 833
Pt-Ir
– tip 330
– wire cantilever 912
pull-hold-release cycle 485
pull-in 1072
– angle 1068
– charge 1064
– instability 956
– voltage 1062, 1064, 1067, 1068, 1071, 1075
pull-off force *see* adhesion force, 608, 614, 615, 617, 635, 864
pulse-etching 379
pumping chamber 261
purified MWNT 378
pyramidal
– AFM tip 615
– etch 373
pyrazoline 16
pyrolysis of hydrocarbons 59
pyrolytic method 58
PZT (led zirconate titanate)
– actuated suspensions 967
– scanner 500
– tube scanner 328, 334

Q

Q-Control 464, 465
Q-factor 388, 436, 455, 1126
quad photodetector 334
quadrant detector 354
quadrature signal 943–945
quality factor Q 339, 348, 389, 454, 911
quality system 1101
quantization energy 117
quantum
– corrals 427
– dot 134, 172, 173
– effects 126, 134
– Hall regime 438
– limit for the thermal conductance 125
– subbands 116
– well 421
– wire superlattices 123
– yield 16
quantum confinement 99, 126, 127, 129
– effect 116, 135, 138
quantum size
– effect 118, 134
– regimes 116
quantum-confined structures 133
quartz-crystal microbalance (QCM) 717
quasi-optical experiment 421
quasi-static
– bending test 775, 777
– mode 388
– strain rate 677

R

radial cracking 808
radio-frequency (RF) MEMS 239
radiotherapy 307
radius of gyration 561
Raleigh's method 348
Raman
– spectra 131, 802
– spectroscopy 800
ramped creep model 890
random access memory (RAM) 28
random access time 900
random force 891
randomly oriented polysilicon 1035
rapid thermal processing (RTP) 1118
raster scan 901
Rayleigh wave 692
Rayleigh–Ritz method 1071
RbBr 397
RC-oscillators 358
reaction kinetics 483
reactive
– lubrication 892
– sealing 1114
– spreading 189
reactive ion etching (RIE) 205, 247, 254
read
– channel 939
– time 908
readback
– rates 932
– signal 940, 941
readout
– electronics 354
– signal 944
receptor binding 295
receptor-ligand
– bond 482
– interaction 484, 545
– unbinding 483
– unbinding force 476
receptor-mediated uptake of synthetic nanomaterial 296
recognition force microscopy (RFM) 476
recognition force spectroscopy (RFS) 484
recognition imaging 489
recombinant protein therapeutics 285
rectangular cantilever 339, 610, 656
rectifying properties 132
recursive least square algorithm (RLS) 974
reduced modulus 690
reflection interference contrast microscopy (RICM) 550
reflectivity 129
relative humidity (RH) 613
– effects of 533, 873, 1004
relative position error signal (RPES) 956
relaxation
– mechanical 457, 848
relaxation time 570–574, 583, 587, 588, 593, 894
relaxation time (T_1)
– longitudinal 305
relaxation time (T_2)
– transverse 305
release
– long-term of drugs 994
– stiction 1095
– voltage 1062
reliability 831
– long-term 1126
– of MEMS 995
– testing 1125
relief on a surface 186
remanence 137
remote detection system 354
removable disk 901
repetitive IgG 484
replica molding 833
repulsive force 480
residual
– film stress 1023
– stress 692, 699, 709, 804, 809, 812
– thickness 855
resistive micro-heater 1119
resolution 233, 374
– vertical 331

resonance
– curve detection 339
– frequency 454
– mechanical 591
resonant
– enhancement effects 131
– frequency 373
Resovist 306
response time 467
rest time 869
retardation effect 550, 552
reticuloendothelial system (RES) 296
retractive slips 885
Reynolds hydrodynamic theory of lubrication 885
RF magnetron sputtering 795
rheological
– characterization 865
– model 469, 664
rhombhedral R-8 706
rigid
– biphenyl chain 853
– disk drive 862, 869
– surface asperities 723
Ritz procedure 1055
RLL (run-length-limited) code 947
RMS roughness 1010
RNA polymerase 747
robust manufacturing 1087
rolling friction 552, 568
root mean square 433
ropes (SWNT) 66
rotary microactuator 958, 993
rotating cylinder 907
rotating strain gauge 1026
rotaxane 13
rough macroscopic contact 885
roughness 558, 559, 566–568, 594
RTP (rapid thermal processing) MEMS vacuum packaging 1119
Ru(II)-trisbipyridine 13
rule of mixtures 708, 709
run-length-limited (RLL) 947
rupture forces 488

S

sacrificial film layers 1040
sales volume 1093
SAM (self-assembled monolayer) 165, 176, 178, 622, 1088
– coating 1075
– melting point of 853
– modified surface 620
sample holder 416
sandwich immunoassays 270
sandwiched molecules 26
saturated calomel electrode 24
saturated hydrocarbons 835
scale-dependent surface tension 614
scaling factor 975
scan
– area 335
– direction 345
– frequency 335
– head 417
– range 359, 360
– rate 329
– size 337
– speed 360
scanner, piezo 436
scanning
– acoustic microscopy (SAM) 663, 692
– capacitance microscopy (SCM) 326
– chemical potential microscopy (SCPM) 326
– electrochemical microscopy (SEcM) 326
– electron microscope (SEM) 110, 373, 521
– electrostatic force microscopy (SEFM) 326
– head 328
– ion conductance microscopy (SICM) 326
– Kelvin probe microscopy (SKPM) 326
– lateral range 929
– local-acceleration microscopy (SLAM) 663
– magnetic microscopy (SMM) 326
– nanoindentation (SN) 674
– nanoindenter 662
– near field optical microscopy (SNOM) 326
– polarization force microscopy (SPFM) 615
– speed 335
– system 359
– thermal microscopy (SThM) 114, 326
– tunneling microscope (STM) 385, 396, 450, 498, 922
– tunneling microscopy (STM) 114, 325, 371, 382, 397, 899
– tunneling probe 114
– velocity 519
scanning force
– acoustic microscopy (SFAM) 326
– microscopy (SFM) 326, 433, 663, 884
– spectroscopy (SFS) 433
scanning probe
– lithography 186
– methods 606
– microscope (SPM) 326
– microscopy (SPM) 371, 385, 612, 663, 899
schematic diagram of T-SLAM 663
Schottky diode 132
scission 622
scope of therapeutic nanodevices 285
scratch
– critical load 824
– damage mechanism 818, 819
– depth 771, 773, 1000
– depth profile 773
– drive actuator 1031
– profile 519
– resistance 765, 771, 853, 1001
– test 520, 814
scratch-induced damage 813
scratching measurement 325
SCREAM 161
screen printing 193
sealing 1112
second anodization 100
secondary signaling drug delivery 301
Seebeck coefficient 123, 133
seek
– control 976
– time 901
selected area electron diffraction (SAED) 111
selective electrodeposition 108
self-actuation voltage 1062
self-aligning optical system 236
self-assembled
– growth 420
– microscopic vesicles 834
– monolayers (SAMs) 27, 165, 176, 187, 476, 536, 606, 618, 703, 833, 839, 861, 1003, 1004
– nanoparticles 312
– nanotube bunches 58
– structures 291, 310
self-assembled monolayer (SAM)
– patterned 188
self-assembly 131, 147, 170, 173, 176, 178, 179, 288
– in biological systems 740
– of nanostructures 289
– of organic layers 23

self-enhancing instability 458
self-excitation
– modes 461
– scheme 464
self-lubrication 652, 653
self-regulation of a therapeutic device 302
self-replicating nanobot 1146
self-tuning control 974, 976
semibiological nanodevices 281
semiconductor 103, 391
– economics 1086
– quantum dots 428
– surface 392
– SWNT 74, 85
semimetallic nanowire 118
semimetal–semiconductor transition 118, 123
sensing
– channel 303
– element 227
– modalities of therapeutic nanotechnology 302
sensing devices
– biological 137
– chemical 137
sensitivity 356, 357, 390, 467, 479, 633, 972
sensitizer 30
sensor 233
– and effector communication 283
– applications 117
– noise 975
– systems 302
– transducer interface 304
sequential logic circuits 20
sequential microfluidic manipulation 268
servo
– demodulation 945
– field 943
– loop 943
– self-writing 945
servo control design 968
sessile-drop 832
SFA (surface forces apparatus) 884, 894, 1011
– measurement 613
SFM (scanning force microscope) 884, 888, 890, 894
– shear force experiments 893
shallow indentation measurement 679
shape memory alloys (SMA) 260
sharpened metal wire tip 382
shear
– energy 887
– force *see* kinetic friction, static friction
– melting 570, 586
– modulus 633, 933
– rate 866
– strength 646
– stress 850, 887
– thinning 581, 583, 584, 589
shear flow
– damping 1073
– detachment (SFD) 479
shear stress
– critical 571, 572, 577, 585, 586
– effective 648
shearing force 848
shearing interface 721
shear-melting transitions 734
SH-group 476
short-cut carbon nanotubes 428
short-range
– chemical force 437, 464
– chemical interaction 402
– contribution 386
– electrostatic attraction 399
– electrostatic interaction 398
– energy dissipation 468
– interatomic force 455
– magnetic interaction 397, 402, 403
shot noise 355
Shubnikov–de Haas quantum oscillatory effect 122
shuttle
– displacement 957
– finger 956
Si 373, 375, 619
– adatom 393
– cantilever 373, 458, 465
– MEMS 247, 985
– nanobeam 773, 781, 785
– stamp 190
– tip 374, 392
– trimer 395
– wafer 373, 459, 619
Si(001)2 × 1 394
Si(001)2 × 1-H surface 394
Si(100) 515, 520, 614, 818, 820, 862, 867–869, 875, 877, 878, 995, 1012
– wafer 863
Si(100)2 × 1:H monohydride 393
Si(111) 519, 527, 848, 850, 999
Si(111) surface 404
Si(111)$\sqrt{3}\times\sqrt{3}$-Ag 395, 396
Si(111)(7 × 7) 434, 468, 640
Si(111)(7 × 7) surface 389, 397, 462, 639
$Si(OH)_4$ film 1013
Si_3N_4 375, 501, 619
– based tip 619
– cantilever 373
– layer 373
– tip 341, 374, 531, 841, 848, 866, 871, 880
Si-Ag covalent bond 397
Si-based devices 764
SiC 995
– bulk properties 995
– film 770
signal
– distribution 1112
– transduction 20
silane (SiH_4) 206
silica
– thermally grown 840
silicon 332, 356, 388, 610, 693, 705, 1111
– AFM tip 478
– cantilever 387, 633
– chemistry 478
– crystalline 254
– dioxide (SiO_2) 254
– dioxide layer 31, 33
– FET 303
– fractures 777
– fusion 1118
– grain 1035
– membrane 263
– MEMS 1095
– micromachining 227
– micromotor 866
– oxide 106, 477
– surface 477
– tip 392, 434, 463, 635
– wafer 25, 27, 186, 189, 228, 286, 453, 614, 930
silicon nitride (Si_3N_4) 254, 332, 477, 610, 615, 913, 964, 968
– cantilever 372
– layer 1117
– tip 375, 635
silicon-based MEMS sensor 229
silicon-based mold masters 255
silicon-on-insulator (SOI) 205, 241, 964
– substrates 247
silicon-only wafer 965
silicon-on-silicon contact 1012
silicon-silicon bonding 1096
silicon-terminated tip 435

silicon-to-glass fusion bonding 1120
siloxane 833
simultaneous imaging 397
– of a $TiO_2(110)$ surface 401
single asperity 646
– AFM tip 868
– contact 517, 1012
– size of 722
single bond 622
single crystal
– aluminum (100) 522
– nanowires 104
– Si(100) 770, 862
– silicon 504, 964, 995, 1024
– silicon cantilever 341, 611
– wafers 204
single crystalline iron 675
single domain magnetic 138
single magnetic flux quantum 122
single molecular motor 480
single molecule
– assays 743
– biology 743
– detection 477
– experiment 760
– studies 485
single particle model 894
single point sensor 1091
single protein therapeutics 281
single receptor-ligand pair 480
single stranded DNA (ssDNA) 746
single-input-multi-output (SIMO) 956
single-input-single-output (SISO) 969
single-layer
– beam 1044
– film 1050
single-particle wave function 422
single-stranded nucleic acid molecule (SSNA) 302
single-wall carbon nanotube (SWCNT) 34, 40, 289, 315, 342, 376, 668
single-wall nanocapsule 49
singularity in joint density of states 134
sink-in 693, 699, 700, 709, 711
– behavior 674
sintering 553
SiO_2 54, 373, 375, 531, 615
– film 770
– nanobeam 773, 774, 785
– thin films 205
SiOH groups 478
$Si-SiO_2$ stress formation 375
SISO 973
– design 969
site-specific drug delivery 298
Si-wafer 460
SLAM 663
slider 952
sliding
– contact 651
– direction 585
– distance 587
– friction 734
– of tip 632
– speed 889
– velocity 574, 585, 644, 884
SLIGA 168
s-like tip states 422
slip distance 732, 889
slipping and sliding regimes 586
SMA (shape memory alloys) 260
– actuation schemes 260
– driven micropumps 261
small angular deflection 1068
small specimens handling 1025
SMANCS (S-Methacryl-neocarzinostatin) 309
– albumin complexes 309
smart media card 905
smart plastic biochip 273
smooth sliding 585
snap-in process 609
$SnO_2(110)$ 400
soap-like lubricants 887
social
– perception 1145
– services 1142
societal impacts 1136
socio-technical trends 1141
soft
– cantilever 453
– coatings 708
– lithography 833
– polymer stamps 190
SOI wafer 1096
solar furnace
– devices 52
– method 61
solder bonding 1115, 1118
solid boundary lubricated surfaces 571, 572, 576–578, 587, 588
solid solution catalyst 55
solid xenon 436
SOLID95 779
solidification 578, 582
solid-like
– behavior 593
– Z-DOL(BW) 866
solution-phase synthesis 107
solvation
– effect 621, 623
– forces 545, 557, 558, 561, 567
sonolubrication 513
soot 46
SP 815
spacer chain 839–841
– length 844
spark plasma sintering (SPS) 81
spatial variation curvature 1058
specific accumulation of nanoparticles 311
specificity 752
spectroscopic resolution in STS 419
speed-energy product 1073
spin
– casting 192
– coating 887
– density waves 431
– quantization 423
spin-on-glass 965
spiropyran 18
SPM (scanning probe microscopy) 73, 372
– methods 672
– tip 452
spreading
– dynamics 892
– profile 892
spring
– sheet cantilever 341
– system 327, 929
spring constant 377, 387, 390, 393, 433, 450, 451, 469, 480, 483, 611, 633, 671, 911, 956, 1030, 1067
– calculation 347
– changes 415
– effective 637, 645, 721
– lateral 338, 636
– measurement 346
– torsional 1068
– vertical 338
sputtered coatings
– physical properties 804
sputtering
– deposition 799
– power 805
sp^3 bonding 806
sp^3-bonded carbon 813
squeeze film damping 248, 1073
SRAM
– cells 1088

– CMOS static random access memory 235
SrF_2 398, 399
$SrTiO_3$ 402
$SrTiO_3(100)$ 400, 401
stability
– mechanical 927
stabilization 1065
stable field performance 1102
stainless steel cantilever 813
stamp
– Au coated PDMS 192
– Au/Ti coated 191
– composite 188
– depth 190
– fabrication 186
– mechanical properties 187
– positioning 195
standardization 249
static
– advancing contact angle (SACA) 847
– AFM 388, 469
– charge 1103
– deflection AFM 450
– friction 571, 576, 579, 583, 585–587, 589, 590, 593, 718, 731, 1034
– friction force 722, 866, 1014
– interactions 544, 584
– mode 387, 433, 453
– mode AFM 354
statistical fluctuation model 889
statistical kernels 891
steady-state sliding *see* kinetic friction
steady-state velocity 733
step tilting 855
stepping motor 636
stereoelectronic properties 12
stick slip 592
sticking regime 586
stick–slip 571, 576, 579, 583–585, 587, 590, 876, 885, 886, 889, 894
– mechanism 636
– motion 732
stiction 579, 832, 885, 991, 1088, 1103
– phenomena in micromotors 1016
– suppression 1103
stiffness 390, 663, 819
– torsional 349
stiffness coefficient
– elastic 1048
stiffness measurement
– continuous (CSM) 810
STM (scanning tunneling microscope) 173–176, 325, 397
– cantilever 330
– cantilever material 339
– principle 327
– probe construction 330
– tip 325, 417
stoichiometry of compound nanowires 115
Stokes's equation 933
Stone–Wales defect 62, 68
Stoney equation 1024
storage
– capacity 946
– drive industry 902
– field 943
– modulus 702
– surface topography 942
strain
– energy 609
– energy difference 810
– engineering 147, 170
– theory 1144
– thermally induced 1125
– transformation 1041
strain–displacement relation 1049
stray capacitance 357
strength 1023
streptavidin 478
– mutants 484
stress 751, 930, 1023
– activation volume 887
– component 1043
– distribution 779, 1049
– field 1024
– gradient 1026
– maximum 1027
– measurement 1025
– relaxation 1060
– singularity 1043
– tensor 759
– transformation 1044
– uniaxial 1042
stresses within the capsid walls 758
stress–strain
– behavior 1028
– curve 671, 784
– relation 1048
stress–temperature response 1060
stretch modulus 751
Stribeck curve 569, 584, 884
strip domain 439
structural
– characterization 110
– entropy 886, 887
– forces 545, 557, 559, 566
– integrity 764
– material 1028
structurally programmable microfluidic system (sPROMs) 264
structure factor 730
structuring of simple liquids 886
SU-8 resist 929
sub-Angstrom deflections 372
subcontractor 1087
subculture theory 1145
sublimation 52
submicron
– electron-beam lithography 246
– ICs 247
– lithographic techniques 246
submonolayer
– coverage 864
– lubrication 731
subnanometer precision 10
sub-nanonewton precision 452
substrate 839, 840
substrate curvature technique 1024
substrate–slider interactions 727
subsurface water irrigation 676
super modulus 691
supercapacitor 85
superconducting
– gap 431
– magnetic levitation 327
– matrix 440
superconductivity 419
superconductors 103, 429
– type-I 431
– type-II 431
superficial nanotechnology aspects 1140
superlattice
– nanowire 103, 114
– structure 108
superlubricity 642
superparamagnetic materials 305
superparamagnetism 138
superstructures of nanowires 108
supramolecular
– assembly 17
– forces 35
surface
– adsorption 263
– amorphous 544
– asperity effect 513
– atom layer 388
– band 423
– barrier potential height 890
– charge 557, 621, 1074
– charge density 555–557

– chemistry 190, 296
– crystalline 544
– diffusion 190
– effects 129
– elasticity 529
– energy 619, 792, 1088
– energy (tension) 547, 552, 566, 567, 572, 574, 578
– forces 544
– forces apparatus (SFA) 479, 498, 546–548, 552, 576, 585, 588, 606, 717, 731, 884
– free energy 745, 746
– hydrophobic 544, 560
– micromachined accelerometers 231
– micromachining 163, 226, 229
– micromechanical properties 498
– microtribological properties 498
– mobility 865
– nanomechanical properties 498
– nanotribological properties 498
– passivation 1102
– potential 419, 555–557, 798
– potential measurement 503
– slope 844
– state lifetime 423
– stiffness 506
– structure 111, 400, 544, 558, 559, 566, 567, 570, 586, 591
– tension 265, 516, 618, 832, 846, 862
– terminal group 839
– topography 632, 778, 901
– transfer chemistry 190
– treatment 1104
surface roughness 383, 674, 678, 724, 773, 778, 844, 942, 997, 1009
– map 510
– measurement 500
surface-mounted biofilter 272
surfactant monolayers 547, 564, 568–570, 577, 580, 584, 585, 588
surfactants 103, 107
surgical resection 307
suspended membrane 1028
suspension 967
– vibration 959
suspension-level fabrication 967
sustainable internal pressure
– maximum 759
sustained drug delivery 269
switch
– two-state 19
switch design 1062
switches
– molecule-based 15
– monolayer 28
switching
– architecture 238
– devices 15
– energy 135
– speed 135
– time 1073
SWNT (single wall nanotube) 40, 52
– adsorption properties 63
– based FET 85
– based materials 64
– catalyst-based 61
– conductance 75
– electronic structure 66
– epoxy composite 82
– flexural modulus 67
– magneto-resistance 67
– matrix interaction 82
– production techniques 45
– rope 668, 670
– structure 41
– surface area 64
– tensile strength 67
– thermal conductivity 66
SWNT-PMMA composite 83
symmetric deformation mode 1054
sync pattern 946
synchronization 945
synthesis 100
– reactor 47
– yield 47
synthetic
– devices 281
– nanoporous membrane 313
– readback signal 941
system
– underdamped 894

T

tailoring of SAMs 841, 1006
tapping amplitude 529
tapping-mode 332, 373, 456, 501, 910
– AFM 456, 459
– etched silicon probe (TESP) 341
– resistance measurement 914
– tip assembly 507
targeting of nanomaterials 296
Taylor dispersion pattern 266
technical PVC 664
technological convergence 1139
Teflon layer 652
temperature
– annealing 227
– critical 423, 434, 1059
– dependence of friction 645
– domain 664
– effect 873
– regulation 720
– sensitive microcantilever 303
temperature-dependent resistance 118
template
– manufacturing 170, 176, 179
– synthesis 100
template-assisted synthesis 100
templates 103
templating technique 57
temporal scales 749
tensile
– load 1027
– maximum stress 783
– straining 1059
– stress 767, 1026, 1035, 1046
– test 1036
terabit 903
terabyte 901
test environment 824
tethering
– flexible 491
tetraarylmethane 13
tetracene 20
– channel 21
tetramethyl-aluminum hydroxide (TMAH) 204, 254
tetrapods 108
tetrathiofulvane (TTF) 916
T-filter 266
therapeutic
– bioactivity 312
– nanodevices 280
– nanomaterials 281
– nanotechnology 285
– photo-oxidation 299
thermal
– activation 894
– activation model of friction 884
– activation model of lubricated friction 887
– conductivity 125, 133
– drift 414, 417
– effect 642
– energy 481
– expansion coefficient 388, 1042
– expansion mismatch 1056
– fluctuation 633
– fluctuation forces 564
– fluctuations and velocity 890
– force 751

– frequency noise 415
– management 1112, 1122
– mismatch 1125
– noise 481, 891
– sensing 925
– vibration amplitude 377
thermocouples 123
thermodynamic forces 293
thermoelastic behavior 1040
thermoelectric
– applications 125, 133
– figure of merit 125, 133
– properties 122, 123
thermo-gravimetric analysis (TGA) 72
thermomechanical
– load 1040, 1058
– noise 433
– write/read 924
thermomechanics of multilayer film 1046
thermoplastics 257
thermopneumatic actuation 258
theta condition 561
thick film lubrication *see* elastohydrodynamic lubrication
thickness 1047
thickness of nanotubes 668
thin film 701, 707
thin oxide film 913
thin-film
– limit 1051
– lubrication 498
– microstructure 1040, 1041
thin-plate theory 1046, 1054
thiol 464, 837
three-axis accelerometer 233
three-digit input/output strings 20
three-dimensional force field spectroscopy 437
three-dimensional micromixers 265
three-input NOR gate 22
three-layer symmetric plate 1052
three-state molecular switch 18
three-terminal devices 31
three-way microvalve system 258
threshold potential 726
through wafer etching 1113
through-thickness
– average 1044
– cracking 807
thymine 10
Ti atom 405
TiAl alloy 675
TiC grains 508
tight-binding approximation 402
tilt angle of SAMs 855
time scales 721
timing field 943
timing recovery 945
TiO_2 400, 402, 425
– substrate 405
$TiO_2(100)$ 400
$TiO_2(110)$ 405
TiO_2-terminated layer 401
tip
– anion-terminated 399
– apex 435, 477, 655, 927
– artifact 392
– atom(s) 415
– atomic structure 383
– characterization 618
– deflection 503
– displacement 941
– Fe-coated 403
– ferromagnetic 441
– geometry 377
– hydrophilic 616
– load 933
– material 383
– mount 329
– oscillation 373
– oscillation amplitude 654
– performance 382
– preparation in UHV 635
– preparation method 330
– properties 382
– radius effect 339, 515, 813, 876, 878
– shape 611
– surface 477
– vibration 460
– vibration amplitude 664
tip-bound
– antigens 484
– biomolecule 303
tip-broadened image 374
tip–cantilever assembly 376
tip-induced atomic relaxation 435
tip-induced quantum dot 424
tip–liquid interface 531, 868
tip–molecule
– distance 419
– gap 406
– long-range force 406
tip–particle distance 29
tip–polymer interface 935
tip–sample 451
– contact 467
– dissipative interaction 468
– distance 388, 396, 399, 401, 435, 451, 459, 507
– electric field 656
– energy dissipation 389
– force 373, 388, 433, 436, 450, 453, 455, 460, 461
– force gradient 391
– gap 910
– interaction 391, 433, 441, 455, 456, 462, 654
– interaction force 608
– interaction potential 464
– interface 372
– junction 468
– potential 390
– separation 467, 608, 609
– system 664
tip–surface
– distance 406, 468
– interaction 490, 651
– potential 636
– separation 480
tip-terminated atom 400
tire pressure monitor 1094
tissue dendritic cells (DCs) 296
tissue engineering scaffolds 310
titin 485
TMR sources 969
Tomlinson model 631, 636
– finite temperature 642
– of friction 884
– one-dimensional 636
– two-dimensional 637
tooth enamel 675
top-down
– approach 10, 133
– materials 281
topographic
– AFM image 611
– images 331
topographical asymmetry 403
topography measurement 334
torsion mirror
– vertical 238
torsional plate 1068
torus model calculation 360
total internal reflection microscopy (TIRM) 550
touch mode architecture 228
Townsend electron avalanche theory 1077
toxicity monitoring 274
trabecular bone 679
track
– centerline 943, 945
– following 943
– mis-registration (TMR) 969
– pitch (TP) 943

track-etched
– polycarbonate membranes 102
– polymers 100
track-following controller 971
tracking servo loop 943
tracks per inch (TPI) 952
transducer component 302
transfer time 901
transformation
– strain 1041
– stress 1044
transistors 132
transition metal complexes 31
transition metal oxides 402
transitions between smooth and stick–slip sliding 585
translational microactuator 958
transmembrane channel 741
transmission 129
transmission electron microscopy (TEM) 111, 374, 521, 693
transplant rejection 311
transport properties of nanowires 115, 118
transversal piezo-resistive effect 356
transverse
– magnetoresistance 121
– midplane displacement 1055
– relaxation time (T_2) 305
traveling direction of the sample 346
Tresca criterion 695
triangular cantilever 339, 349, 354
triaryl 13
tribochemical
– oxidation 999
– reaction 531, 872
tribological
– characterization of coating 805
– computer simulations 721
tribological issues
– in MEMS 989
– in NEMS 989
tribological performance 996
– of coatings 822
tribological properties
– of SAM 841
– of silicon 996
tribometer 996
(tridecafluoro-1,1,2,2-tetrahydro-oct-1-yl) trichlorosilane (n-$C_6F_{13}CH_2CH_2SiCl_3$, FTS) 842
triethanolamine 14
triethoxysilane 644
trifluoromethyl-terminated SAM 843
triggering 288
– external 284
– strategy 298
trimer tip apex 452
true atomic resolution 389, 392, 434
true tip-sample interaction 608
truth table 17, 21
T-SLAM 663
tubulin monomer 740
tumor
– architecture 307
– associated antigens (TAAs) 307
– deposition 309
– properties 307
– targeting 307
– vasculature 307
tungsten 330
– sphere 633
– tip 330, 468, 614
tunneling
– current 327, 388, 389, 424
– detector 331
– junctions 25
– tip 431
turn-on voltage 117
two-body potentials 719
two-digit input strings 20
two-dimensional
– elasticity 754
– electron gas (2-DEG) 438
– electron system (2-DES) 428
two-layer beam 1045
two-state switch 19
two-terminal nanowire device 132

U

UHV environment 638
UHV-AVM 635
ultrahigh vacuum (UHV) 391, 418, 461, 607, 635
ultrasonic
– bonding 1119
– lubrication 513
– transducer 663
ultrasound triggering 299
ultrathin DLC coatings 795
umklapp processes 125
unbinding
– force 476, 481
– force distribution 485
– pathway 488
uncapped MEMS wafer 1097, 1098
uncoated monolithic cantilever 611
uncommon failure mode 1102
unconstrained binding 477
unidirectional electron transfer 28
unimolecular level 22
universality 752
unlubricated sample 879
unlubricated surfaces *see* dry surfaces
upward cantilever deflection 746
UV-LIGA lithography 267

V

vacuum encapsulation 1116, 1117, 1126
vacuum packaged MEMS 1118
vacuum packaging 231, 1116, 1121
van der Waals 607, 617, 621
– adhesion 616
– attraction 515
– contact 759
– force 372, 386, 398, 404, 406, 450, 544, 545, 550–554, 557, 558, 618
– interaction 58, 289, 402, 457, 843, 865
– surfaces 435
vanadium carbide 691
vapor deposition 104, 120
vapor grown
– carbon fiber 53
– carbon nanofiber 56
vapor-liquid-solid (VLS) mechanism 105
varactor diodes 241
variable capacitor 240
variable force mode 362
variable temperature STM
– setup 416
variable-temperature SLAM (T-SLAM) 663
vascular
– address system mapping 297, 298
– prosthesis 310
VCM loop compensator 972
vehicle stability 232
velocity
– critical 585, 587, 590, 591
velocity dependence of friction 644
velocity effect 871
Verlet algorithm 457
vertical coupling 329
vertical rms-noise 434
vertically aligned MWNT 59
Verwey transition temperature 439

very large-scale integration (VLSI) 922, 940
vibration 749
– external 329
vibratory rate gyroscope 232
vibromotor 236
Vickers
– hardness 674, 678, 695
– indentations 771
– indenter 765
vinylidene fluoride 404
viral life cycle 755
virgin single-crystal silicon 996
virtual jump distance 889
viruses 747
viscoelastic
– effect 468
– model 934
– properties 681
– relaxation time 731
– response 731
viscoelasticity 663, 702, 711
– mapping 505, 528
viscosity 730, 885
viscous
– damping 469
– force 569, 580, 584, 1061
VLS (Vapor-Liquid-Solid) 60
– growth method 116
– mechanism 60
– method 105, 108, 137
VLSI (very large scale integration) 922, 940
voice coil motor (VCM) 952
voltage bias 383
voltage-controlled oscillator (VCO) 239, 946
voltage-displacement function 1068
voltage-to-force gain 956
von Karman plate theory 1055
vortex in superconductor 440
V-shaped cantilever 341, 610, 634

W

W tip 330
wafer
– curvature technique 1034
– fabrication 1087
– saw dilemma 1097, 1098
wafer-level vacuum packaging 1122
wafer-to-wafer
– vacuum packaging 1114
– variation 1072
waste disposal concern 1090
water 613
– capillary force 847
– meniscus 461
water vapor 691, 695
– content 824
Watson-Crick base pairing 484
wear 498, *see* damage, 618, 793, 809
– contribution to friction 650
– control 832
– damage 824
– damage mechanism 823
– debris 521
– depths 1000
– measurement 325
– mechanical 263
– mechanism 522
– of tip 728
– performance 866
– process 822
– region 523
– resistance 822, 823, 853, 856, 879, 997, 1008
– study 879
– test 822
wearless
– friction 884
– static friction 726
Weibull
– distribution 774
– distribution function 1125
– statistics 1036
weight function 390
weighting functions for μ-synthesis 976
well-aligned nanowires 111
wet (chemical) etching 254
wet etching in HF 247
wettability 846
wetting 569
– lubricants 892
white-noise driven system 889
wire boundary scattering 121
wire cantilever 339
wireless communication 1094
work
– hardening 704, 709
– of adhesion 844, 847, 1011
– of indentation 701
write/read scheme 931
writing mechanism 932

X

xenon 417
XNOR 17
X-ray
– analysis 115
– diffraction (XRD) patterns 103
– lithographic techniques 286

Y

yield point 570, 571, 585
– load 703
yield stress 703
yielding 1055
Young's modulus 81, 346, 362, 377, 415, 611, 646, 668, 677, 779, 1023, 1024, 1044, 1071
– measurement 1028
Young–Dupre equation 846
Young–Laplace equation 612

Z

Z-15 530, 862, 874–876, 878, 1006
– film 531, 879
– lubricant 1003
– lubricant film 867
– properties 863
z-axis accelerometer 230
z-axis gyroscope 233
Z-DOL 530, 614, 862, 869
– partially bonded film 880
– properties 863
Z-DOL(BW) 868, 874, 877–879, 1006
– film 530, 867, 1003
zeolites 54
zeta potential 261
zig-zag axis 42
zig-zag-type SWNT 42
zinc 704
zipper actuator 1070
Zn nanowire 120, 123
ZnO nanowire 135

Springer Handbook *of* Nanotechnology
B. Bhushan • © Springer 2004

ENCYCLOPEDIA OF MODERN OPTICS

EDITED BY
B.D. GUENTHER
A. MILLER
L. BAYVEL
J.E. MIDWINTER

VOLUME ONE

EDITOR-IN-CHIEF

Robert D. Guenther
Duke University
Durham, NC, USA

EDITORS

Duncan G. Steel
University of Michigan
Ann Arbor, MI, USA

Leopold Bayvel
London, UK

CONSULTING EDITORS

John E. Midwinter
University College London, London, UK

Alan Miller
University of St Andrews, Fife, UK

EDITORIAL ADVISORY BOARD

R Alferness
Lucent Technologies Bell Laboratories
Holmdel, NJ, USA

D J Brady
Duke University
Durham, NC, USA

J Caulfield
Diversified Research Corporation
Cornersville, TN, USA

R L Donofrio
Display Device Consultants LLC
Ann Arbor, MI, USA

J Dudley
Université de Franche-Comté
Besançon, France

J G Eden
University of Illinois
Urbana, IL, USA

H O Everitt
U.S. Army Research Office
Research Triangle Park, NC, USA

E L Hecht
Adelphi University
Garden City, NY, USA

H S Hinton
Utah State University
Logan, UT, USA

T N Hurst
Hewlett-Packard Laboratories
Tucson, AZ, USA

D Killinger
University of South Florida
Tampa, FL, USA

T A King
University of Manchester
UK

E Leith
University of Michigan
Ann Arbor, MI, USA

A Macleod
Thin Film Center, Inc.
Tucson, AZ, USA

P F McManamon
Air Force Research Laboratory
Wright-Paterson AFB, OH, USA

J N Mait
U.S. Army Research Laboratory
Adelphi, MD, USA

D Phillips
Imperial College
London, UK

C R Pidgeon
Heriot-Watt University
Edinburgh, UK

R Pike
King's College London
UK

E Van Stryland
University of Central Florida
Orlando, FL, USA

Preface

We live in a world powered by light, but much of the understanding of light was developed around the time of the French Revolution. Before the 1950s, optics technology was viewed as well established and there was little expectation of growth either in scientific understanding or in technological exploitation. The end of the Second World War brought about huge growth in scientific exploration, and the field of optics benefited from that growth. The key event was the discovery of methods for producing a source of coherent radiation, with the key milestone being the demonstration of the first laser by Ted Maiman in 1960. Other lasers, nonlinear optical phenomena, and technologies such as holography and optical signal processing, followed in the early 1960s. In the 1970s the foundations of fiber optical communications were laid, with the development of low-loss glass fibers and sources that operated in the wavelength region of low loss. The 1980s saw the most significant technological accomplishment: the development of efficient optical systems and resulting useful devices. Now, some forty years after the demonstration of the first coherent light source, we find that optics has become the enabling technology in areas such as:

- information technology and telecommunications;
- health care and the life sciences;
- sensing, lighting, and energy;
- manufacturing;
- national defense;
- manufacturing of precision optical components and systems; and
- optics research and education.

We find ourselves depending on CDs for data storage, on digital cameras and printers to produce our family photographs, on high speed internet connections based on optical fibers, on optical based DNA sequencing systems; our physicians are making use of new therapies and diagnostic techniques founded on optics.

To contribute to such a wide range of applications requires a truly multidisciplinary effort drawing together knowledge spanning many of the traditional academic boundaries. To exploit the accomplishments of the past forty years and to enable a revolution in world fiber-optic communications, new modalities in the practice of medicine, a more effective national defense, exploration of the frontiers of science, and much more, a resource to provide access to the foundations of this field is needed. The purpose of this Encyclopedia is to provide a resource for introducing optical fundamentals and technologies to the general technical audience for whom optics is a key capability in exploring their field of interest.

Some 25 internationally recognized scientists and engineers served as editors. They helped in selecting the topical coverage and choosing the over 260 authors who prepared the individual articles. The authors form an international group who are expert in their discipline and come from every part of the technological community spanning academia, government and industry. The editors and authors of this Encyclopedia hope that the reader finds in these pages the information needed to provide guidance in exploring and utilizing optics.

As Editor-in-Chief I would like to thank all of the topical editors, authors and the staff of Elsevier for each of their contributions. Special thanks should go Dr Martin Ruck of Elsevier who provided not only organizational skills but also technological knowledge which allowed all of the numerous loose ends to be tied.

B D Guenther
Editor-in-Chief

A

ALL-OPTICAL SIGNAL REGENERATION

O Leclerc, Alcatel Research & Innovation,
Marcoussis, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The breakthrough of optical amplification, combined with the techniques of wavelength division multiplexing (WDM) and dispersion management, have made it possible to exploit a sizeable fraction of the optical fiber bandwidth (several terahertz). Systems based on 10 Gbit/s per channel bit-rate and showing capacities of several terabit/s, with transmission capabilities of hundreds or even thousands of kilometers, have reached the commercial area.

While greater capacities and spectral efficiencies are likely to be reached with current technologies, there is potential economic interest in reducing the number of wavelength channels by increasing the channel rate (e.g., 40 Gbit/s). However, such fourfold increase in the channel bit-rate clearly results in a significant increase in propagation impairments, stemming from the combined effects of noise accumulation, fiber dispersion, fiber nonlinearities, and inter-channel interactions and contributing to two main forms of signal degradation. The first one is related to the amplitude domain; power levels of marks and spaces can suffer from random deviations arising from interaction between signal and amplified spontaneous emission (ASE) noise or with signals from other channels through cross-phase modulation (XPM) from distortions induced by chromatic dispersion. The second type of signal degradations occurs in the time domain; time position of pulses can also suffer from random deviations arising from interactions between signal and ASE noise through fiber dispersion. Preservation of high power contrast between '1' and '0', and of both amplitude fluctuations and timing jitter below some acceptable levels, are mandatory for high transmission quality, evaluated through bit-error-rate (BER) measurements or estimated by Q-factors. Moreover, in future optical networks, it appears mandatory to ensure similar but high optical signal quality at the output of whatever nodes in the networks, as to enable successful transmission of the data over arbitrary distance.

Among possible solutions to overcome such systems limitations is the implementation of Optical Signal Regeneration, either in-line for long-haul transmission applications or at the output of network nodes. Such Signal Regeneration performs, or should be able to perform, three basic signal-processing functions that are Re-amplifying, Re-shaping, and Re-timing, hence the generic acronym '3R' (Figure 1). When Re-timing is absent, one usually refers to the regenerator as '2R' device, which has only re-amplifying and re-shaping capabilities. Thus, full 3R regeneration with retiming capability requires clock extraction.

Given system impairments after some transmission distance, two solutions remain for extending the actual reach of an optical transmission system or the scalability of an optical network. The first consists in segmenting the system into independent trunks, with full electronic repeater/transceivers at interfaces (we shall refer to this as 'opto-electronic regeneration' or O/E Regeneration forthwith). The second solution, i.e., all-optical Regeneration, is not the optical version of the first which would have higher bandwidth capability but still performs the same signal-restoring functions with far reduced complexity. At this point, it should be noted that Optical 3R techniques are not necessarily void of any electronic functions (e.g., when using electronic clock recovery and O/E modulation), but the main feature is that these electronic functions are narrowband (as opposed to broadband in the case of electronic regeneration).

Some key issues have to be considered when comparing such Signal Regeneration approaches. The first is that today's and future optical transmission systems or/and networks are WDM networks.

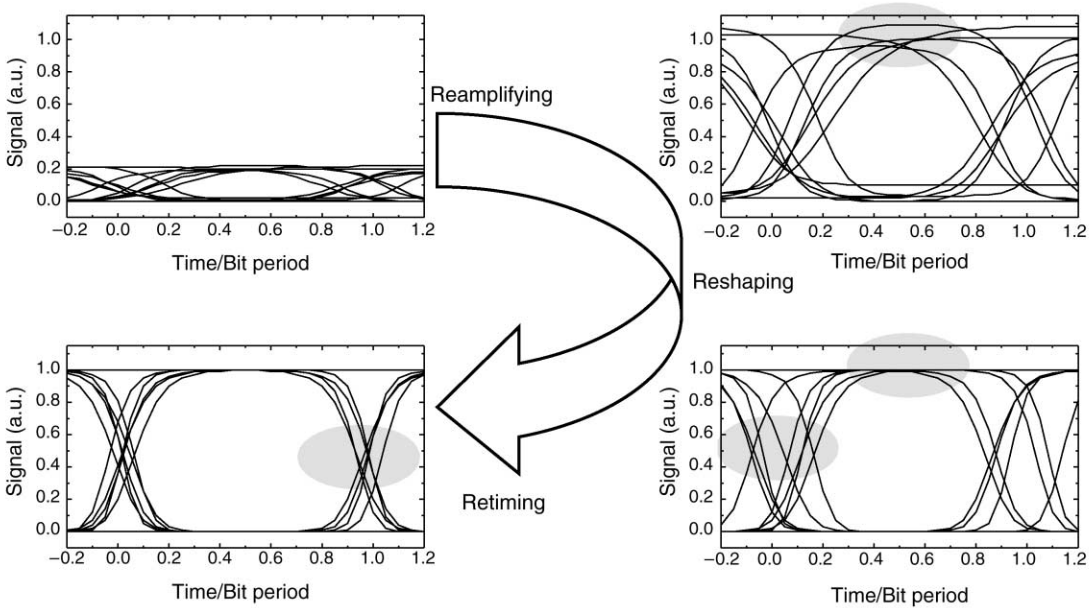

Figure 1 Principle of 3R regeneration, as applied to NRZ signals. (a) Re-Amplifying; (b) Re-Shaping; (c) Re-Timing. NB: Such eye diagrams can be either optical or electrical eye diagrams.

Under this condition, the WDM compatibility – or the fact that any Regeneration solution can simultaneously process several WDM channels – represents a key advantage. The maturity of the technology – either purely optical or opto-electronic – also plays an important role in the potential (pre-)development of such solutions. But the main parameter that will decide the actual technology (and also technique) relies on the tradeoff between actual performance of the regeneration solutions and their costs (device and implementation), depending on the targeted applications (long-haul system, medium haul transport, wide area optical network, etc.).

In this article, we review the current alternatives for all-optical Signal Regeneration, considering both theoretical and experimental performance and practical implementation issues. Key advantages and possible drawbacks of each solutions are discussed, to sketch the picture in this field. However, first we must focus on some generalities about Signal Regeneration and the way to define (and qualify) such regenerator performance. In a second part, we will detail the currently-investigated optical solutions for Signal Regeneration with a specific highlight for semiconductor-based solutions using either semiconductor optical amplifiers (SOA) technology or newly-developed saturable absorbers. Optical Regeneration techniques based on synchronous modulation will also be discussed in a third section. The conclusion will summarize the key features of each solution, so as to underline the demanding challenge optical components are facing in this application.

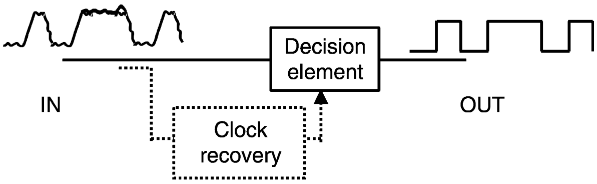

Figure 2 Generic structure of Signal 2R/3R Regenerator based on Decision Element (2R) and Decision Element and Clock Recovery (3R).

Generalities on Signal Regeneration

Principles

In the general case, Signal Regeneration is performed using a decision element exhibiting a nonlinear transfer function. Provided with a threshold level and when associated with an amplifier, such an element then performs the actual Re-shaping of the incoming data (either in the electrical or optical domain) and completes a 2R Signal Regenerator. Figure 2 shows the generic structure of such a Signal Regenerator in the general case as applied to non return to zero (NRZ) data. A clock recovery block can be added (dotted lines) to provide the decision element with time reference and hence perform the third R (Re-timing) of full Signal 3R Regeneration. At this point, it should be mentioned that the decision element can operate either on electrical signals (standard electrical DFF) provided that optical → electrical and electrical → optical signal conversion stages are added or directed onto optical signals using the different techniques described below. The clock signal can be of an electrical nature, as for electrical decision element in O/E regenerator – or either an electrical or a purely optical signal in all-optical regenerators.

Prior to reviewing and describing the various current technology alternatives for such Optical Signal Regeneration, the issue of the actual characterization of regenerator performance needs to be explained and clarified. As previously mentioned, the core element of any Signal Regenerator is the decision element showing a nonlinear transfer function that can be of varying steepness. As will be seen in Figure 3, the actual regenerative performance of the regenerator will indeed depend upon the degree of nonlinearity of the decision element transfer function.

Figure 3 shows the principle of operation of a regenerator incorporating a decision element with two steepnesses of the nonlinear transfer function. In any case, the ‘1’ and ‘0’ symbols amplitude probability densities (PD) are squeezed after passing through the decision element. However, depending upon the addition of a clock reference for triggering the decision element, the symbol arrival time PD will be also squeezed (clocked decision = 3R regeneration) or enlarged (no clock reference = 2R regeneration) resulting in conversion of amplitude fluctuations to time position fluctuations.

As for system performance – expressed through BER – regenerative capabilities of any regenerator simultaneously depend upon both the output amplitude and arrival time PD of the ‘1’ and ‘0’ symbols. In the unusual case of 2R regeneration (no clocked decision), a tradeoff has then to be derived, considering both the reduction of amplitude PD and the enlarged arrival time PD induced by the regenerator, to ensure sufficient signal improvement. In Figure 3a, we consider a step function for the transfer function of the decision element. In this case, amplitude PD are squeezed to Dirac PD after the decision element, and depending upon addition or not of a clock reference, arrival time PD is reduced (3R) or dramatically enlarged (2R). In Figure 3b, the decision element exhibits a moderately nonlinear transfer function. This results in an asymmetric and less-pronounced squeezing of the amplitude PD compared to the previous case, but in turn results in a significantly less enlarged arrival time PD when no clock reference is added (2R regeneration). Comparison of these two decision element of different nonlinear transfer function indicates that for 3R regeneration applications, the more nonlinear the transfer function of the decision element the better performance, the ideal case being the step function. In the case of 2R regeneration applications, a tradeoff between the actual reduction of the amplitude PD and enlargement of timing PD is to be derived and clearly depends upon the actual shape of the nonlinear transfer function of the decision element.

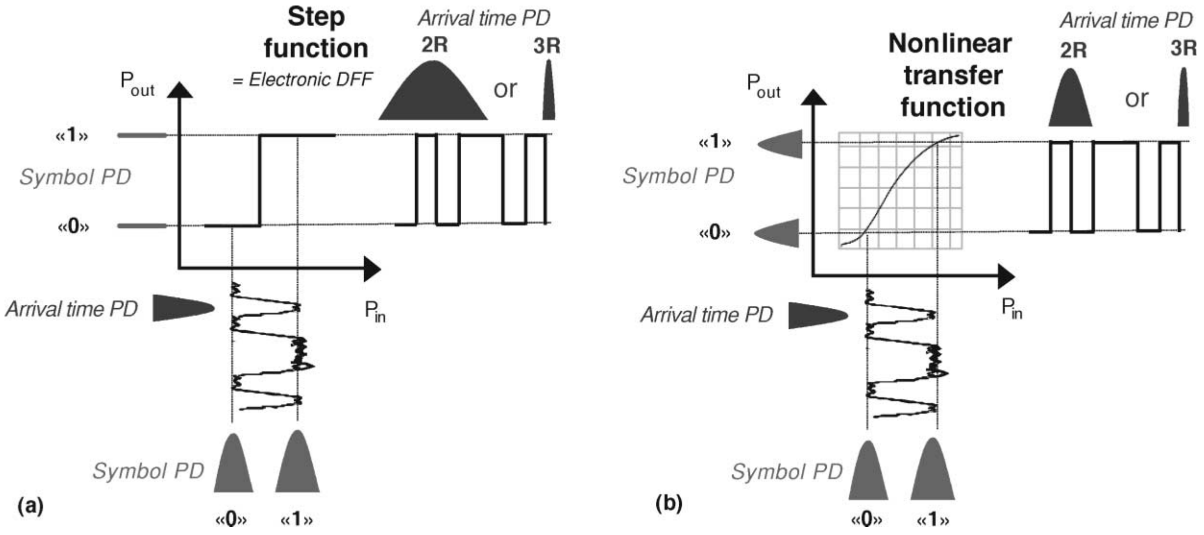

Figure 3 Signal Regeneration process using Nonlinear Gates. (a) Step transfer function (= Electronic DFF); (b) ‘moderately’ nonlinear transfer function. As an illustration of the Regenerator operation ‘1’ and ‘0’ symbols amplitude probability density (PD) and arrival time probability density (PD) are shown in light gray and dark gray, respectively.

Qualification of Signal Regenerator Performance

To further illustrate the impact of the actual shape of the nonlinear transfer function of the decision element in 3R application, the theoretical evolution of BER with number of concatenated regenerators have been plotted for regenerators having different nonlinear responses. Figure 4 shows the numerically calculated evolution of the BER of a 10 Gbit/s NRZ signal with fixed optical signal-to-noise ratio (OSNR) at the input of the 3R regenerator, as a function of the cascaded regenerator incorporating nonlinear gates with nonlinear transfer function of different depths. From Figure 4, can be seen different behaviors depending on the nonlinear function shape. As previously stated, the best regeneration performance is obtained with an ideal step function (case a), which is actually the case for O/E regenerator using electronic decision flip-flop (DFF). In that case, BER linearly increase (i.e., more errors) in the cascade. Conversely, when nonlinearity is reduced (cases (b) and (c)), both BER and noise accumulate, until the concatenation of nonlinear functions reach some steady-state pattern, from which BER linearly increases. Concatenation of nonlinear devices thus magnifies shape differences in their nonlinear response, and hence their regenerative capabilities.

Moreover, as can be seen in Figure 4, all curves standing for different regeneration efficiencies pass through a common point defined after the first device. This clearly indicates that it is not possible to qualify the regenerative capability of any regenerator when considering the output signal after only one regenerator. Indeed, the BER is the same for either a 3R regenerator or a mere amplifier if only measured after a single element. This originates from the initial overlap between noise distributions associated with marks and spaces, that cannot be suppressed but only minimized by a single decision element through threshold adjustment.

As a general result, the actual characterization of the regenerative performance of any regenerator should *in fine* be conducted considering a cascade of regenerators. In practice this can easily be done with the experimental implementation of the regenerator under study in a recirculating loop. Moreover, such an investigation tool will also enable access to the regenerator performance with respect to the transmission capabilities of the regenerated signal, which should not be overlooked.

Let us now consider the physical implementation of such all-optical Signal Regenerators, along with the key features offered by the different alternatives.

All-Optical 2R/3R Regeneration Using Optical Nonlinear Gates

Prior to describing the different solutions for all-optical in-line Optical Signal Regeneration, it should be mentioned that since the polarization states of the optical data signals cannot be preserved during propagation, it is required that the regenerator exhibits an extremely low polarization sensitivity. This clearly translates to a careful optimization of all the different optical components making up the 2R/3R regenerator. It should be noted that this also applies to the O/E solutions but is of limited impact, since only the photodiode has to be polarization insensitive.

Figure 5 illustrates the generic principle of operation of an all-optical 3R Regenerator using optical nonlinear gates. Contrary to what occurs in O/E, regenerator where the extracted clock signal drives the electrical decision element, the incoming and distorted optical data signal triggers the nonlinear gate, hence generating a switching window which is

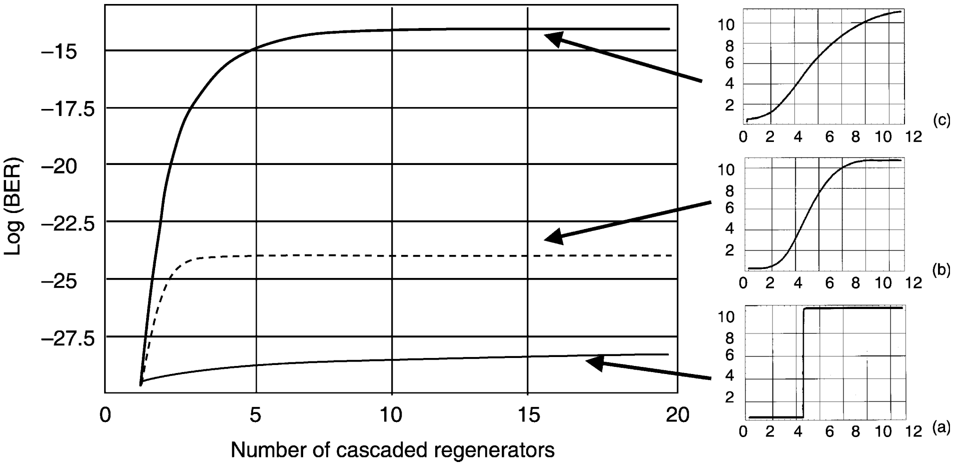

Figure 4 Evolution of the BER with concatenated regenerators for nonlinear gates with nonlinear transfer function of decreasing depths from case (a)–(c) (step function).

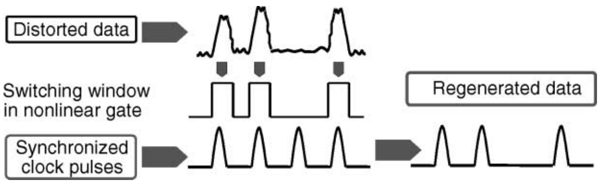

Figure 5 Principle of operation of an all-Optical Signal 3R Regenerator using nonlinear gates.

applied to a newly generated optical clock signal so as to reproduce the initial data stream on the new optical carrier.

In the case of 2R Optical Signal Regeneration, a continuous wave (CW) signal is substituted for the synchronized optical clock pulses. As previously mentioned, the actual regeneration performance of the 2R/3R devices will mainly depend upon the nonlinearity of the transfer function of the decision element but in 3R applications the quality of the optical clock pulses has also to be considered. In the following, we describe current solutions to explain the two main building blocks of all-optical Signal Regenerators: the decision element (i.e., nonlinear gate) and the clock recovery (CR) elements.

Optical Decision Element

In the physical domain, optical decision elements with ideal step response – as for electrical DFF – do not exist. Different nonlinear optical transfer functions, approaching more or less the ideal case, can be realized in various media such as fiber, SOA, electroabsorption modulators (EAM), and lasers. Generally, as described below, the actual response (hence the regenerative properties of the device) of such optical gates directly depends upon the incoming signal instantaneous power. Under these conditions, it appears essential to add an adaptation stage so as to reduce intensity fluctuations (as caused by propagation or crossing routing/switching node) and provide the decision element with fixed power conditions. In practice, this results in the addition of a control circuitry (either optical or electrical) in the Re-amplification block, whose complexity directly depends on actual system environment (ultra-fast power equalization for packet-switching applications and compensation of slow power fluctuations in transmission applications).

As previously described the decision gate performs Re-shaping (and Re-timing when clock pulses are added) of the incoming distorted optical signal, and represent the regenerator's core element. Ideally, it should also act as a transmitter with characteristics ensuring the actual propagation of the regenerated data stream. In that respect, the chirp possibly induced by the optical decision gate onto the regenerated signal – and the initial quality of the optical clock pulses in 3R applications – should be carefully considered (ideally by means of loop transmission) as to adequately match line transmission requirements.

Different solutions for the actual realization of the optical decision element have been proposed and extensively investigated using, for example, cross gain modulation in semiconductor optical amplifier (SOA) devices but the most promising and flexible devices probably are interferometers, for which, descriptions of the generic principle of operation follows. Consider a CW signal (probe) at λ_2 wavelength injected into an optical interferometer, in which one arm incorporates a nonlinear medium in which an input signal carried by λ_1 wavelength (command) is, in turn, injected. Such a signal, at λ_1 wavelength, induces a phase shift through cross-phase modulation (XPM) in this arm of the interferometer, the amount depending upon power level $P_{\mathrm{in},\lambda 1}$. In turn, such phase modulation (PM) induces amplitude modulation (AM) on the signal at λ_2 wavelength when recombined at the output of the interferometer and translates the information carried by wavelength λ_1 onto λ_2. Under these conditions, such optical gates clearly act as wavelength converters (it should be mentioned that Wavelength Conversion is not necessarily equivalent to Regeneration; i.e., a linear transfer function performs suitable Wavelength Conversion but by no means Signal Regeneration).

Optical interferometers can be classified according to the nature of the nonlinearity exploited to achieve a π phase shift. In the case of fiber-based devices, such as the nonlinear optical loop mirror (NOLM), the phase shift is induced through the Kerr effect in an optical fiber. The key advantage of fiber-based devices such as NOLM lies in the near-instantaneous (fs) response of the Kerr nonlinearity, making them very attractive for ultra-high bit-rate operation ($\geq$160 Gbit/s). Polarization-insensitive NOLMs have been realized, although with the same drawbacks concerning integrability. With recent developments in highly nonlinear (HNL) fibers, however, the required NOLM fiber length could be significantly reduced, hence dramatically reducing environmental instability.

A second type of device is the integrated SOA-based Mach-Zehnder interferometers (MZI). In MZIs, the phase shift is due to the effect of photo-induced carrier depletion in the gain saturation regime of one of the SOAs. The control and probe can be launched in counter- or co-directional ways. In the first case, no optical filter is required at the output of the device for rejecting the signal at λ_1 wavelength but operation of the MZI is limited by its speed. At this point, one should mention that the photo-induced modulation effects in SOAs are intrinsically limited in speed by the gain recovery time, which is a function of the carrier lifetime and the injection current. An approach referred to as differential operation mode (DOM) and illustrated on Figure 6, which takes advantage of the MZI's interferometric properties, makes it possible to artificially increase the operation speed of such 'slow' devices up to 40 Gbit/s.

As discussed earlier, the nonlinear response is a key parameter for regeneration efficiency. Combining two interferometers is a straightforward means to improve the nonlinearity of the decision element transfer function, and hence regeneration efficiency. This approach was validated at 40 Gbit/s using a cascade of two SOA-MZI, (see Figure 7 (**left**)). Such a scheme offers the advantage of restoring data polarity and wavelength, hence making the regenerator inherently transparent. Finally, the second conversion stage can be used as an adaptation interface to the transmission link achieved through chirp tuning in this second device.

Such an Optical 3R Regenerator was upgraded to 40 Gbit/s, using DOM in both SOA-MZIs with validation in a 40 Gbit/s loop RZ transmission. The 40 Gbit/s eye diagram monitored at the regenerator output after 1, 10, and 100 circulations are shown in Figure 7 (**right**) and remain unaltered by distance. With this all-optical regenerator structure, the minimum OSNR tolerated by the regenerator (1 dB sensitivity penalty at 10^{-10} BER) was found to be as low as 25 dB/0.1 nm. Such results clearly illustrate the high performance of this SOA-based regenerator structure for 40 Gbit/s optical data signals.

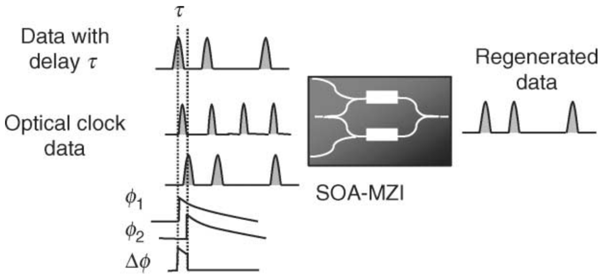

Figure 6 Schematic and principle of operation of SOA-MZI in differential mode.

Such a complex mode of operation for addressing 40 Git/s bit-rates will probably be discarded when we consider the recent demonstration of standard-mode wavelength conversion at 40 Gbit/s, which uses a newly-designed active-passive SOA-MZI incorporating evanescent-coupling SOAs. The device architecture is flexible in the number of SOAs, thus enabling easier operation optimization and reduced power consumption, leading to simplified architectures and operation for 40 Gbit/s optical 3R regeneration.

Based on the same concept of wavelength conversion for Optical 3R Regeneration, it should be noted many devices have been proposed and experimentally validated as wavelength converters at rates up to 84 Gbit/s, but with cascadability issues still to be demonstrated to assess their actual regenerative properties.

Optical Clock Recovery (CR)

Next to the decision, the CR is a second key function in 3R regenerators. One possible approach for CR uses electronics while another only uses optics. The former goes with OE conversion by means of a photodiode and subsequent EO conversion through a modulator. This conversion becomes more complex and power-hungry as the data-rate increases. It is clear that the maturity of electronics gives a current advantage to this approach. But considering the pros and cons of electronic CR for cost-effective implementation, the all-optical approach seems more promising, since full regenerator integration is potentially possible with reduced power consumption. In this view, we shall focus here on the optical approach and more specifically on the self-pulsating effect in three-sections distributed feedback (DFB) lasers or more recently in distributed Bragg reflector

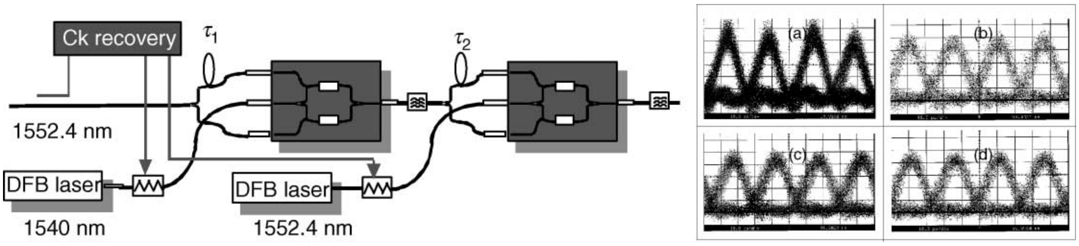

Figure 7 Optimized structure of a 40 Gbit/s SOA-based 3R regenerator. 40 Gbit/s eye diagram evolution: (a) B-to-B; (b) 1 lap; (c) 10 laps; (d) 100 laps.

(DBR) lasers. Recent experimental results illustrate the potentials of such devices for high bit rates (up to 160 Gbit/s), broad dynamic range, broad frequency tuning, polarization insensitivity, and relatively short locking times (1 ns). This last feature makes these devices good candidates for operation in asynchronous-packet regimes.

Optical Regeneration by Saturable Absorbers

We next consider saturable absorbers (SA) as nonlinear elements for optical regeneration. Figure 8 (**left**) shows a typical SA transfer function and illustrates the principle of operation. When illuminated with an optical signal with peak power below some threshold (P_{sat}), the photonic absorption of the SA is high and the device is opaque to the signal (low transmittance). Above P_{sat}, the SA transmittance rapidly increases and asymptotically saturates to transparency (passive loss being overlooked). Such a nonlinear transfer function only applies to 2R optical regeneration.

Different technologies for implementing SAs are available, but the most promising approach uses semiconductors. In this case, SA relies upon the control of carrier dynamics through the material's recombination centers. Parameters such as on–off contrast (ratio of transmittance at high and low incident powers), recovery time (1/e) and saturation energy, are key to device optimization. In the following, we consider a newly-developed ion-irradiated MQW-based device incorporated in a micro-cavity and shown on Figure 8 (**right**). The device operates as a reflection-mode vertical cavity, providing both a high on/off extinction ratio by canceling reflection at low intensity and a low saturation energy of 2 pJ. It is also intrinsically polarization-insensitive. Heavy ion-irradiation of the SA ensures recovery times (at $1/e$) shorter than 5 ps (hence compatible with bit-rate above 40 Gbit/s), while maintaining a dynamic contrast in excess of 2.5 dB at 40 GHz repetition rate.

The regenerative properties of SA make it possible to reduce cumulated amplified spontaneous emission (ASE) in the '0' bits, resulting in a higher contrast between mark and space, hence increasing system performance. Yet SAs do not suppress intensity noise in the marks, which makes the regenerator incomplete. A solution for this noise suppression is optical filtering with nonlinear (soliton) pulses. The principle is as follows. In absence of chirp, the soliton temporal width scales in the same way as the reciprocal of its spectral width (Fourier-transform limit) times its intensity (fundamental soliton relation). Thus, an increase in pulse intensity corresponds to both time narrowing and spectral broadening. Conversely, a decrease in pulse intensity corresponds to time broadening and spectral narrowing. Thus, the filter causes higher loss when intensity increases, and lower loss when intensity decreases. The filter thus acts as an automatic power control (APC) in feed-forward mode, which causes power stabilization. The resulting 2R regenerator (composed by the SA and the optical filter) is fully passive, which is of high interest for submarine systems where the power consumption must be minimal, but it does not include any control in the time domain (no Re-timing).

System demonstrations of such passive SA-based Optical Regeneration have been reported with a 20 Gbit/s single-channel loop experiment. Implementation of the SA-based 2R Regenerator with 160 km-loop periodicity made it possible to double the error-free distance ($Q = 15.6$ dB or 10^{-9} BER) of a 20 Gbit/s RZ signal. So as to extend the capability of passive 2R regeneration to 40 Gbit/s systems, an improved configuration was derived from numerical optimization and experimentally demonstrated in a 40 Gbit/s WDM-like, dispersion-managed loop transmission, showing more than a fourfold increase in the WDM transmission distance at 10^{-4} BER (1650 km without the SA-based regenerator and 7600 km when implementing the 2R regenerator with 240 km periodicity).

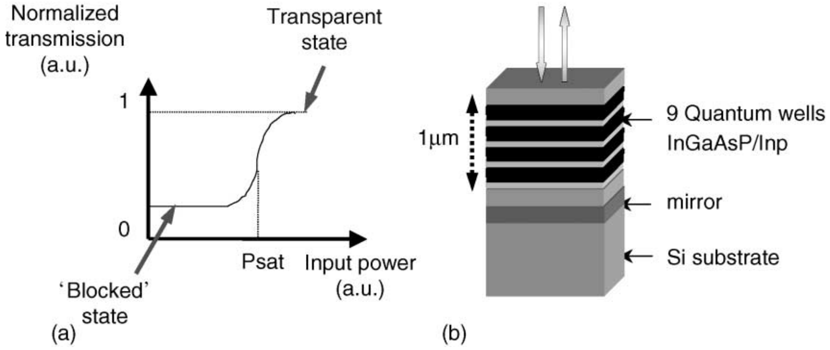

Figure 8 (a) Saturable Absorber (SA) ideal transfer function. (b) Structure of Multi-Quantum Well SA.

Such a result illustrates the potential high interest of such passive optical 2R regeneration in long-haul transmission (typically in noise-limited systems) since the implementation of SA increases the system's robustness to OSNR degradation without any extra power consumption. Reducing both saturation energy and insertion loss along with increasing dynamic contrast represent key future device improvements. Regeneration of WDM signals from the same device, such as one SA chip with multiple fibers implemented between Mux/DMux stages, should also be thoroughly investigated. In this respect, SA wavelength selectivity in quantum dots could possibly be advantageously exploited.

Synchronous Modulation Technique

All-optical 3R regeneration can be also achieved through in-line synchronous modulation (SM) associated with narrowband filtering (NF). Figure 9 shows the basic layout of such an Optical 3R Regenerator. It is composed of an optical filter followed by an Intensity and Phase Modulator (IM/PM) driven by a recovered clock. Periodic insertion of SM-based modulators along the transmission link provides efficient jitter reduction and asymptotically controls ASE noise level, resulting in virtually unlimited transmission distances. Re-shaping and Re-timing provided by IM/PM intrinsically requires nonlinear (soliton) propagation in the trunk fiber following the SM block. Therefore, one can refer to the approach as distributed optical regeneration. This contrasts with lumped regeneration, where 3R is completed within the regenerator (see above with Optical Regeneration using nonlinear gates), and is independent of line transmission characteristics.

However, when using a new approach referred to 'black box' optical regeneration (BBOR), it is possible to make the SM regeneration function and transmission work independently in such a way that any type of RZ signals (soliton or non-soliton) can be transmitted through the system. The BBOR technique includes an adaptation stage for incoming RZ pulses in the SM-based regenerator, which ensures high regeneration efficiency regardless of RZ signal format (linear RZ, DM-soliton, C-RZ, etc.). This is achieved using a local and periodic soliton conversion of RZ pulses by means of launching an adequate power into some length of fiber with anomalous dispersion. The actual experimental demonstration of the BBOR approach and its superiority over the 'classical' SM-based scheme for DM transmission was experimentally investigated in 40 Gbit/s DM loop transmission. Under these conditions, one can then independently exploit dispersion management (DM) techniques for increasing spectral efficiency in long-haul transmission, while ensuring high transmission quality through BBOR.

One of the key properties of the SM-based all-optical Regeneration technique relies on its WDM compatibility. The first (Figure 10, **left**) and straightforward solution to apply Signal Regeneration to WDM channels amounts to allocating a regenerator to each WDM channel. The second consists in sharing a single modulator, thus processing the WDM channels at once in serial fashion. This approach requires WDM synchronicity, meaning that all bits be synchronous with the modulation, that can be achieved either by use of appropriate time-delay lines located within a DMux/Mux apparatus (Figure 10, **upper right**), or by making the WDM channels inherently time-coincident at specific regenerator locations (Figure 10, **bottom right**). Clearly, the serial regeneration scheme is far simpler and cost-effective than the parallel version; however, optimized re-synchronization schemes still remain to be

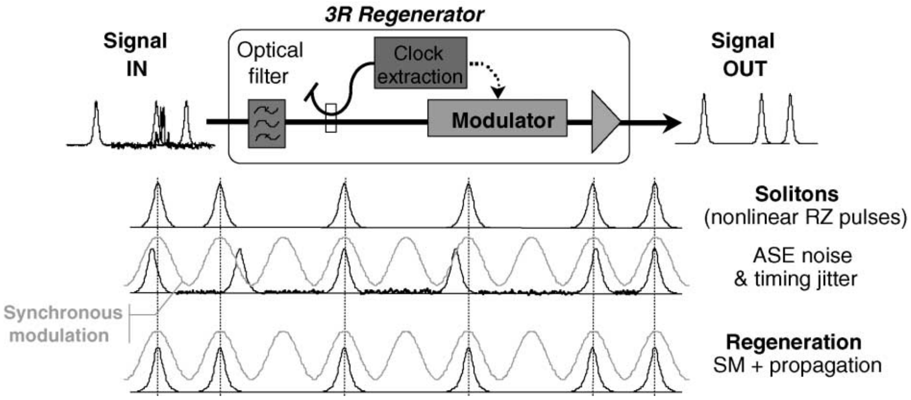

Figure 9 Basic layout of the all-optical Regenerator by Synchronous Modulation and Narrowband Filtering and illustration of the principle of operation.

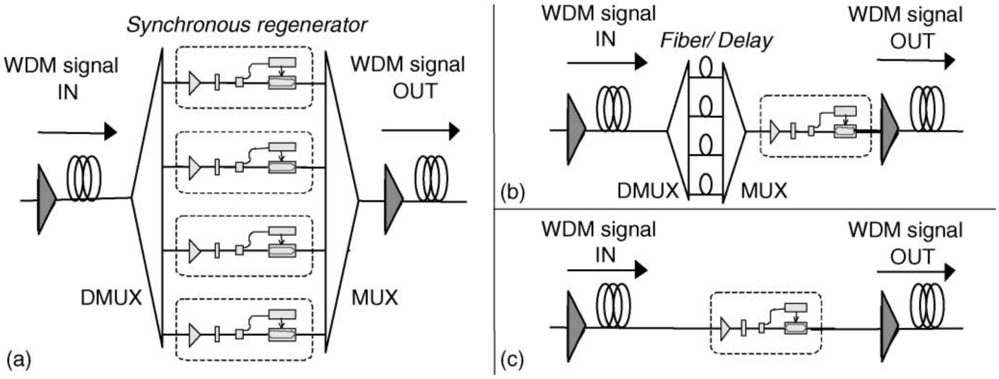

Figure 10 Basic implementation schemes for WDM all-Optical Regeneration. (a) parallel asynchronous; (b) serial re-synchronized; (c) serial self-synchronized.

developed for realistic applications. Experimental demonstration of this concept was assessed by means of a 4 × 40 Gbit/s dispersion-managed transmission over 10 000 km (BER $< 5.10^{-8}$) in which a single modulator was used for the simultaneous regeneration of the 4 WDM channels.

Considering next all-optical regeneration schemes with ultra-high speed potential, a compact and loss-free 40 Gbit/s Synchronous Modulator, based on optically-controlled SOA-MZI, was proposed and loop-demonstrated at 40 Gbit/s with an error-free transmission distance in excess of 10 000 km. Moreover, potential ultra-high operation of this improved BBOR scheme was recently experimentally demonstrated by means of a 80 GHz clock conversion with appropriate characteristics through the SOA-MZI. One should finally mention all fiber-based devices such as NOLM and NALM for addressing ultra-high speed SM-based Optical Regeneration, although no successful experimental demonstrations have been reported so far in this field.

Conclusion

In summary, optical solutions for Signal Regeneration present many key advantages. These are the only advantages to date to possibly ensure WDM compatibility of the regeneration function (mostly 2R related). Such optical devices clearly exhibits the best 2R regeneration performance (wrt to O/E solutions) as a result of the moderately nonlinear transfer function (which in turn can be considered as a drawback in 3R applications), but the optimum configuration is still to be clearly derived and identified depending upon the system application. Optics also allow to foresee and possibly target ultrafast applications above 40G, for signal regeneration if needed. Among the current drawbacks, one should mention the relative lack of wavelength/formats flexibility of these solutions (compared to O/E solutions). It is complex or difficult to restore the input wavelength or address any C-band wavelength at the output of the device or to successfully regenerate modulation formats other than RZ. In that respect, investigations should be conducted to derive new optical solutions capable of processing more advanced modulation formats at 40G. Finally, the fact that the nonlinear transfer function of the optical is in general triggered by the input signal instantaneous power also turns out to be a drawback since it requires control circuitry. The issue of cost (footprint, power consumption, etc.) of these solutions, compared to O/E ones, is still open. In this respect, purely optical solutions incorporating all-optical clock recovery, the performance of which is still to be technically assessed, are of high interest for reducing costs. Complete integration of an all-optical 2R/3R regenerator or such parallel regenerators onto a single semiconductor chip should also contribute to make all-optical solutions cost-attractive even though acceptable performance of such fully integrated devices is still to be demonstrated.

From today's status concerning the two alternative approaches for in-line regeneration (O/E or all-optical), it is safe to say that the choice between either solution will be primarily dictated by both engineering and economical considerations. It will result from a tradeoff between overall system performance, system complexity and reliability, availability, time-to-market, and rapid returns from the technology investment.

See also

Interferometry: Overview. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Bigo S, Leclerc O and Desurvire E (1997) All-optical fiber signal processing for solitons communications. *IEEE Journal of Selected Topics on Quantum Electronics* 3(5): 1208–1223.

Dagens B, Labrousse A, Fabre S, *et al*. (2002) New modular SOA-based active–passive integrated Mach–Zehnder interferometer and first standard mode 40 Gb/s all-optical wavelength conversion on the C-band. paper PD 3.1, Proceedings of ECOCOL 02, Copenhague.

Durhuus T, Mikkelsen B, Joergensen C, Danielsen SL and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14(6): 942–954.

Lavigne B, Guerber P, Brindel P, Balmefrezol E and Dagens B (2001) Cascade of 100 optical 3R regenerators at 40 Gbit/s based on all-active Mach–Zehnder interferometers. paper We.F.2.6, Proceedings of ECOC 2001, Amsterdam.

Leclerc O, Lavigne B, Chiaroni D and Desurvire E (2002) All-optical regeneration: Principles and WDM implementation. In: Kaminow IP and Li T (eds) *Optical Fiber Technologies IVA*, chap. 15, pp. 732–784. San Diego: Academic Press.

Leuthold J, Raybon G, Su Y, *et al*. (2002) 40 Gbit/s transmission and cascaded all-optical wavelength conversion over 1,000,000 km. *Electronic Letters* 38(16): 890–892.

Öhlén P and Berglind E (1997) Noise accumulation and BER estimates in concatenated nonlinear optoelectronic repeaters. *IEEE Photon Technol. Letters* 9(7): 1011–1013.

Otani T, Suzuki M and Yamamoto S (2001) 40 Gbit/s optical 3R regenerator for all-optical networks. We.F.2.1, Proceedings of ECOC 2001, Amsterdam.

Oudar J-L, Aubin G, Mangeney J, *et al*. (2004) Ultra-fast quantum-well saturable absorber devices and their application to all-optical regeneration of telecommunication optical signals. *Annales des Télécommunications* 58(11–12): 1667–1707.

Sartorius B, Mohrle M, Reichenbacher S, *et al*. (1997) Dispersive self-Q-switching in self-pulsating DFB lasers. *IEEE Journal of Quantum Electronics* 33(2): 211–218.

Ueno Y, Nakamura S, Sasaki J, *et al*. (2001) Ultrahigh-speed all-optical data regeneration and wavelength conversion for OTDM systems. Th.F.2.1, Proceedings of ECOC 2001, Amsterdam.

B

BABINET'S PRINCIPLE

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In the development of the theory of diffraction, the diffraction field is due to a surface integral, the Fresnel integral, but no restrictions are imposed on the choice of the surface over which the integration must be performed. This fact leads to a very useful property called Babinet's principle, first stated by Jacques Babinet (1794–1872) in 1837 for scalar waves. We will discuss only the scalar Babinet's principle; discussion of the rigorous vector formulation can be found in a number of books on electromagnetic theory.

To introduce Babinet's principle, we label a plane, separating the source, S, and observation point, P, as Σ. If no obstructions are present, a surface integration over Σ yields the light distribution at P. If we place an opaque obstruction in this plane with a clear aperture, Σ_1, then the field at P is given by integrating over only Σ_1; contributions from Σ, outside of Σ_1, are zero since the obstruction is opaque.

We may define an aperture, Σ_2, as complementary to Σ_1 if the obstruction is constructed by replacing the transparent regions of Σ, i.e., Σ_1, by opaque surfaces and the opaque regions of Σ by clear apertures. Figure 1 shows two complementary obstructions, where the shaded region indicates an opaque surface.

The surface integral over Σ generates the field, E, in the absence of any obstruction. If obstruction Σ_1 is present then the diffraction field is E_1, obtained by integrating over Σ_1. According to Babinet's principle, the diffraction field due to obstruction Σ_2, must be

$$\tilde{E}_2 = \tilde{E} - \tilde{E}_1$$

We will look at examples of the application of the principle for both Fraunhofer and Fresnel diffraction.

Fraunhofer Diffraction

The electric field due to Fraunhofer diffraction is given by

$$E_P(\omega_x, \omega_y) = \frac{i\alpha e^{-ikR_0}}{\lambda R_0} \iint_{\Sigma} f(x, y) e^{-i(\omega_x x + \omega_y y)}\, dx\, dy$$

where α is the amplitude of the incident plane wave and R_0 is the distance from the obstruction to the observation point. We have defined the spatial frequencies, ω_x and ω_y by the equations

$$\omega_x = kL = -\frac{2\pi\xi}{\lambda R_0} \qquad \omega_y = kM = -\frac{2\pi\eta}{\lambda R_0}$$

This surface integral is an approximation of the Hygens–Fresnel integral and can be identified as a Fourier transform of the aperture transmission function, $f(x, y)$ in two dimensions. For our discussion we will consider only one dimension and ignore the parameters in front of the integral.

Assume that $a(x)$ is the amplitude transmission of an aperture in an obstruction, $b(x)$ is the amplitude transmission function of the complementary obstruction and g is the amplitude of the wave when no

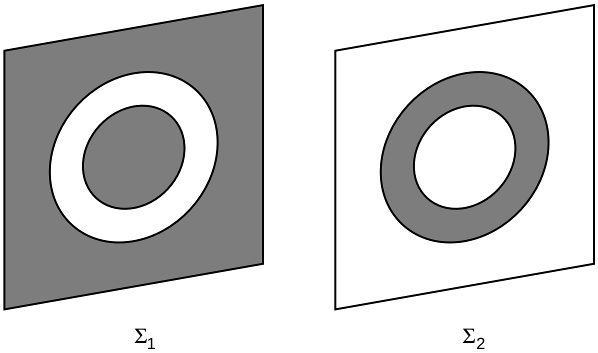

Figure 1 An aperture Σ_1 shown by the unshaded region and its complement Σ_2. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

obstruction is present. The Fourier transforms of the one-dimensional amplitude transmission functions of the two apertures are equal to the Fraunhofer diffraction patterns that will be generated by the two apertures

$$A(k) = \int_{-\infty}^{\infty} a(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

$$B(k) = \int_{-\infty}^{\infty} b(x)\mathrm{e}^{-ikx}\,\mathrm{d}x$$

With no aperture present, the far field amplitude is

$$G(k) = g\int_{-\infty}^{\infty} \mathrm{e}^{-ikx}\,\mathrm{d}x$$

Babinet's principle states that

$$B(x) = G - A(x)$$

must be the Fraunhofer diffraction field of the complementary aperture. We may rewrite this equation for the diffraction field as

$$B(k) = G\delta(k) + A(k)\mathrm{e}^{\mathrm{i}\pi} \qquad [1]$$

The first term on the right of the Fraunhofer diffraction field for the complementary obstruction [1] is located at the origin of the observation plane and is proportional to the amplitude of the unobstructed wave. The second term in the equation for the Fraunhofer diffraction pattern [1] is identical to the Fraunhofer diffraction pattern of the original aperture, except for a constant phase. Thus the Fraunhofer diffraction from the two complementary apertures are equal, except for a constant phase and the bias term at the origin. Physically this means that the diffraction intensity distributions of complementary apertures will be identical but their brightness will differ!

Fresnel Diffraction

We can calculate the Fresnel diffraction from an opaque disk by applying Babinet's principle to the diffraction pattern calculated for a circular aperture.

We assume that a circular aperture of radius a is illuminated by a point source a distance Z' from the aperture. We observe the transmitted wave at the point P, located a distance Z from the aperture (see Figure 2).

The Fresnel diffraction integral is

$$\tilde{E}_\mathrm{P} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD} \times \int\int f(x,y)\mathrm{e}^{-\frac{ik}{2\rho}\left[(x-x_0)^2+(y-y_0)^2\right]}\,\mathrm{d}x\,\mathrm{d}y \qquad [2]$$

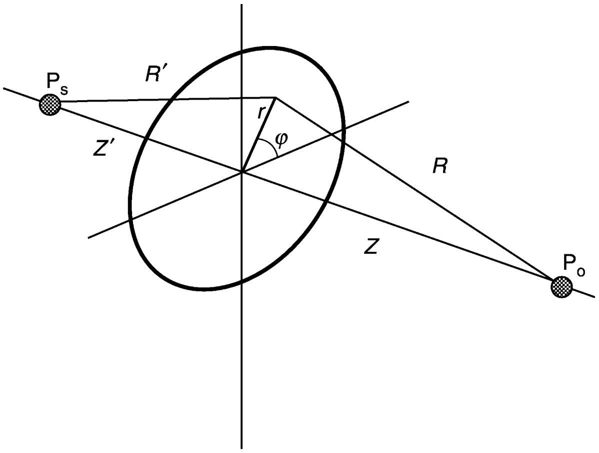

Figure 2 Geometry for the analysis of Fresnel diffraction of a circular aperture. Reproduced with permission from Guenther R (1990) *Modern Optics*. New York: Wiley. Copyright © 1990 John Wiley & Sons.

where the distance between the source and observation point is

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z+Z')} \qquad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'}$$

and the coordinates of the stationary point are

$$x_0 = \frac{Z'\xi + Zx_s}{Z+Z'} \qquad y_0 = \frac{Z'\eta + Zy_s}{Z+Z'}$$

The parameters of the integral are simplified by the geometry selected for Figure 2. Both the observation point and source are located on the axis of symmetry of the aperture which lies along the z-axis of a cylindrical coordinate system thus

$$x_0 = y_0 = \xi = \eta = 0 \qquad \text{and} \qquad D = Z' + Z$$

The approximation for the distance from the source and observation point to the aperture yields

$$R + R' \approx Z + Z' + \frac{r^2}{2\rho} = D + \frac{r^2}{2\rho}$$

where $r^2 = x^2 + y^2\cdot$ The cylindrical coordinate version of [2] is then

$$\tilde{E}_\mathrm{P} = \frac{\mathrm{i}\alpha}{\lambda\rho D}\mathrm{e}^{-\mathrm{i}kD}\int_0^a\int_0^{2\pi} f(r,\varphi)\mathrm{e}^{-\mathrm{i}\frac{kr^2}{2\rho}}\,r\,\mathrm{d}r\,\mathrm{d}\phi \qquad [3]$$

We assume that the transmission function of the aperture is a constant, $f(r,\phi) = 1$, to make the integral as simple as possible. After performing the integration over the angle ϕ, [3] can be rewritten as

$$\tilde{E}_{\mathrm{Pa}} = \frac{\mathrm{i}\alpha\pi}{D}\mathrm{e}^{-\mathrm{i}kD} \int_0^{\frac{a^2}{\rho\lambda}} \mathrm{e}^{-\mathrm{i}\pi r^2/\rho\lambda}\,\mathrm{d}\left(\frac{r^2}{\rho\lambda}\right) \qquad [4]$$

Performing the integration of [4], results in the Fresnel diffraction amplitude

$$\tilde{E}_{\mathrm{Pa}} = \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\left[1 - \mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda}\right] \qquad [5]$$

The intensity of the diffraction field is

$$\begin{aligned} I_{\mathrm{Pa}} &= 2I_0\left(1 - \cos\frac{\pi a^2}{\rho\lambda}\right) \\ &= 4I_0 \sin^2\frac{\pi a^2}{2\rho\lambda} \qquad [6] \end{aligned}$$

[6] predicts a sinusoidal variation in intensity as the observation point moves toward the aperture.

Babinet's principle can now be used to evaluate Fresnel diffraction by an opaque disk, the same size as the circular aperture, i.e., of radius *a*. The field for the disk is obtained by subtracting the field, diffracted by a circular aperture, from the field of an unobstructed spherical wave, E_{P}:

$$\begin{aligned} \tilde{E}_{\mathrm{Pd}} &= \tilde{E}_{\mathrm{P}} - \tilde{E}_{\mathrm{Pa}} \\ \tilde{E}_{\mathrm{Pd}} &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD} - \tilde{E}_{\mathrm{Pa}} \\ &= \frac{\alpha}{D}\mathrm{e}^{-\mathrm{i}kD}\mathrm{e}^{-\mathrm{i}\pi a^2/\rho\lambda} \qquad [7] \end{aligned}$$

The intensity is

$$I_{\mathrm{Pd}} = I_0$$

The result of the application of Babinet's principle is the conclusion that, at the center of the geometrical shadow cast by the disk, there is a bright spot with the same intensity as would exist if no disk were present. This is Poisson's spot. The intensity of this spot is independent of the choice of the observation point along the *z*-axis.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction.

Further Reading

Abramowitz M and Stegun IA (1964) *Handbook of Mathematical Functions*. NBS Applied Mathematics Series-55 Washington, DC: National Bureau of Standards.

Born M and Wolf E (1959) *Principles of Optics*. New York: Pergamon Press.

Guenther R (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

C

CHAOS IN NONLINEAR OPTICS

R G Harrison and W Lu, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Instabilities in laser emission, notably in the form of spontaneous coherent pulsations, have been observed almost since the first demonstration of laser action. Indeed, the first laser operated in 1960 by Maiman produced noisy spiked output even under conditions of quasi-steady excitation and provided the early impetus for studies of such effects. Subsequent theoretical efforts towards understanding these phenomena occurred up to the 1980s at a modest level, due in part to the wide scope of alternative areas of fertile investigation provided by lasers during this period. However, since the 1990s, there has been a major resurgence of interest in this area. This has been due to profound mathematical discoveries over this period which have revolutionized our understanding of classical dynamical systems. It is now clear that many systems containing some form of nonlinearity are dynamically unstable and even chaotic and that such behavior is deterministic. Further, the discovery that chaos evolves through particular routes with well-defined scenarios and that such routes are universal, has stimulated experimentalists to search for physical systems that exhibit these properties. These phenomena have now been observed in areas as diverse as fluid flow, chemical reactions, population ecology, and superconductor devices. The laser, perhaps the most recent newcomer to the field, is a paradigm for such investigation, owing to its simplicity both in construction and in the mathematics that describe it, and already a wide range of these phenomena have been observed. Along with a proliferation of new observations, many of the instabilities known to occur in these systems are now being reinterpreted in lieu of our new insight in this area. From this, fascinating new concepts of control and synchronization of chaos have emerged and spawned new fields of applications, not least in secure communication. In this article we introduce the general principles of dynamical instabilities, chaos, and control in lasers. For in-depth study, the reader is referred to texts and review articles cited in the Further Reading section at the end of this article.

Laser Physics

Our conventional understanding of laser physics is concerned with how cooperative order evolves from randomness. This transition is explained by first considering the ensemble of lasing atoms in thermal equilibrium. Every atom executes a small motion, giving rise to an oscillating electric dipole, described by linear dynamics. For larger motions, the atomic dipoles interfere with each other's motion and beyond a particular threshold, the motion becomes 'co-operative' or ordered over a long range. The role of the electromagnetic field in the laser cavity in ordering the induced atomic dipoles, is collectively described as giving rise to a macroscopic polarization, the magnitude of which depends on the number of dipoles (excited atoms). The dynamical interplay between the cavity field amplitude (E) as one variable and the atomic material variables of polarization (P) and population (D) of excited atoms, all of which are in general complex quantities, provide a full description of lasing action. For the simplest laser, a single mode two-level system with a homogeneously broadened gain lasing at resonance, these reduce to just three real variables, the Maxwell–Bloch equations:

$$E = -\kappa E + \kappa P \qquad [1]$$

$$P = \gamma_{\perp} ED - \gamma_{\perp} P \qquad [2]$$

$$D = \gamma_{II}(\lambda + 1) - \gamma_{II} D - \gamma_{II} \lambda EP \qquad [3]$$

where κ is the cavity decay rate, $\gamma_{\perp}$ is the decay rate of atomic polarization, γ_{II} is the decay rate of population inversion, and λ the pumping parameter.

Described in this way the laser acts as a single oscillator in much the same way as a mechanical nonlinear oscillator, where laser output represents damping of the oscillator and excitation of the atoms is the driving mechanism to sustain lasing (oscillation). For ordered or coherent emission, it is necessary for one or both of the material variables (which are responsible for further lasing emission) to respond sufficiently fast to ensure a phase correlation with the existing cavity field. This is readily obtained in many typical lasers with output mirrors of relatively high reflectivity, since the field amplitude within the laser cavity will then vary slowly compared to the fast material variables which may then be considered through their equilibrium values. This situation, commonly referred to as adiabatic elimination of fast variables, reduces the dynamics of lasing action to that of one (or two) variables, the field (and population), all others being forced to adapt constantly to the slowly varying field variable. Our familiar understanding of DC laser action presupposes such conditions to hold. However when, for example, the level of excitation of the atoms is increased to beyond a certain value, i.e., the second laser threshold, all three dynamical variables may have to be considered, which satisfies the minimum requirement of three degrees of freedom (variables) for a system to display chaos. So this simple laser is capable of chaotic motion, for which the emission is aperiodic in time and has a broadband spectrum. Prediction of such behavior was initially made by Haken in 1975 through establishing the mathematical equivalence of the Maxwell–Bloch equations to those derived earlier by Lorentz describing chaotic motion in fluids.

Nonlinear Dynamics and Chaos in Lasers

In general there is no general analytical approach for solving nonlinear systems such as the Maxwell–Bloch equations. Instead solutions are obtained by numerical means and analyzed through geometric methods originally developed by Poincaré as early as 1892. This involves the study of topological structures of the dynamical trajectories in phase space of the variables describing the system. If an initial condition of a dissipative nonlinear dynamical system, such as a laser, is allowed to evolve for a long time, the system, after all the transients have died out, will eventually approach a restricted region of the phase space called an attractor. A dynamical system can have more than one attractor, in which case different initial conditions lead to different types of long-time behavior. The simplest attractor in phase space is a fixed point which is a solution with just one set of values for its dynamical variables; the nonlinear system is attracted towards this point and stays there, giving a DC solution. For other control conditions the system may end up making a periodic motion; the attractor of this motion is called a limit cycle. However, when the operating conditions exceed a certain critical value, the periodic motion of the system breaks down into a more complex dynamical trajectory, which never repeats. This motion represents a third kind of attractor in phase space called a chaotic or strange attractor.

Figure 1 shows a sequence of trajectories of the Lorentz equations in the phase space of (x, y, z) on increasing the value of one of the control parameters; this corresponds to (E, P, D) in the Maxwell–Block equations on increasing the laser pump. For a control setting near zero, all trajectories approach stable equilibrium at the origin, the topological structure of the basin of attraction being hyperbolic about zero (Figure 1a). For the laser equations, this corresponds to operation below lasing threshold for which the magnitude of the control parameter (laser gain) is insufficient to produce lasing. As the control parameter is increased, the basin lifts at its zero point to create an unstable fixed point here but with now two additional fixed points located in the newly formed troughs on either side of the zero point, which is now a saddle point. This is illustrated in Figure 1b where the system eventually settles to one or other of the two new and symmetrical fixed points, depending on the initial conditions. This corresponds to DC or constant lasing as defined by the parameter values of one or other of these points which are indistinguishable. Pictorially, think of a conventional saddle shape comprising a hyperbolic and inverted hyperbolic form in mutually perpendicular planes and connected tangentially at the origin. With the curve of the inverted hyperbola turned up at its extremity and filling in the volume with similar profiles which allow the two hyperbolic curves to merge into a topological volume, one sees that a ball placed at the origin is constrained to move most readily down either side of the inverted hyperbolas into one or other of the troughs formed by this volume. For the Lorentz system, chaotic behavior occurs at larger values of the control parameter when all three fixed points become saddles. Since none of the equilibrium points are now attracting, the behavior of the system cannot be a steady motion. Although perhaps difficult to visualize topologically, it is then possible to find a region in this surface enclosing all three points and

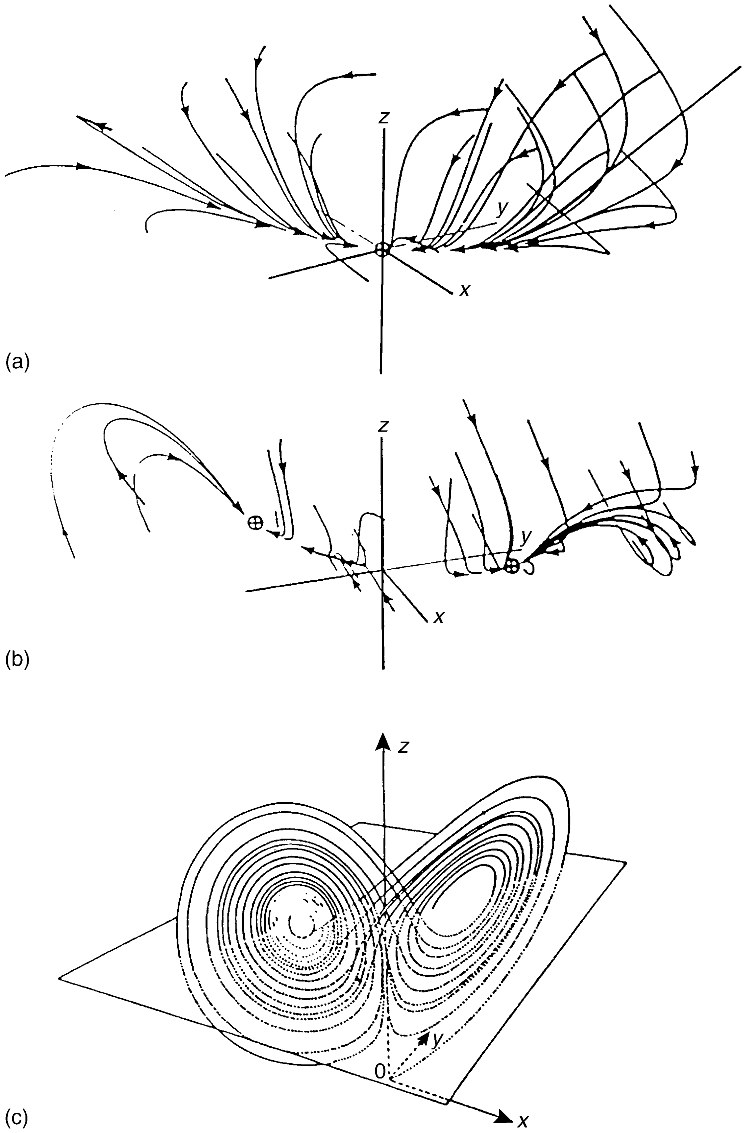

Figure 1 Trajectories in the three-dimensional phase space of the Lorentz attractor on increasing one of the control parameters. (Reproduced with permission from Thompson JMT and Stewart HB (1987) *Nonlinear Dynamics and Chaos*. New York: John Wiley.)

large enough so that no trajectories leave the region. Thus, all initial conditions outside the region evolve into the region and remain inside from then on. A corresponding chaotic trajectory is shown in Figure 1c. A point outwardly spirals from the proximity of one of the new saddle points until the motion brings it under the influence of the symmetrically placed saddle, the trajectory then being towards the center of this region from where outward spiraling again occurs. The spiraling out and switching over continues forever though the trajectory never intersects. In this case, arbitrarily close initial conditions lead to trajectories, which, after a sufficiently long time, diverge widely. Since a truly exact assignment of the initial condition is never possible, even numerically, a solution comprising several such trajectories therefore evolves and, as a consequence, long-term predictability is impossible. This is in marked contrast to that of the fixed point and limit cycle attractors, which settle down to the same solutions. A recording of one of these variables in time, say for a laser the output signal amplitude (in practice recorded as signal intensity, proportional to the square of the field amplitude), then gives oscillatory emission of increasing intensity with sudden discontinuities (resulting from flipping from one saddle to the other in Figure 1c) as expected.

While the Lorentz/Haken model is attractive for its relative simplicity, many practical lasers cannot be reduced to this description. Nevertheless, the

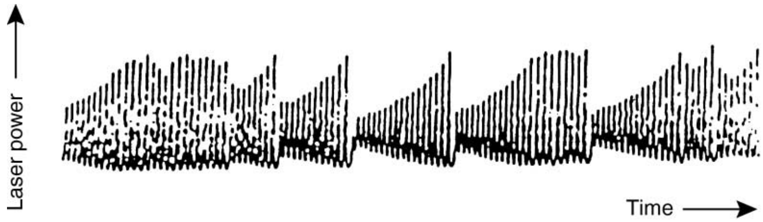

Figure 2 Lorentz-type chaos in the NH_3 laser (emitting at 81 μm optically pumped by a N_2O laser.

underlying topology of the trajectory of solutions in phase space for these systems is often found to be relatively simple and quite similar, as for three-level models descriptive of optically pumped lasers. An experimental example of such behavior for a single mode far infared molecular laser is shown in Figure 2. In other lasers, for which there is adiabatic elimination of one or other of the fast variables, chaotic behavior is precluded if, as a consequence, the number of variables of the system is reduced to less than three. Indeed this is the case for many practical laser systems. For these systems the addition of an independent external control parameter to the system, such as cavity length or loss modulation, have been extensively used as a means to provide the extra degrees of freedom. Examples include gas lasers with saturable absorbers, solid state lasers with loss modulation, and semiconductor lasers with external cavities, to name but a few. In contrast, for multimode rather than single mode lasers, intrinsic modulation of inversion (or photon flux) by multimode parametric interaction ensures the additional degrees of freedom. Furthermore, when the field is detuned from gain center, the dynamical variables E, P, D are complex, providing five equations for single mode systems, which is more than sufficient to yield deterministic chaos for suitable parameter values. Also, of significance is the remarkably low threshold found for the generation of instabilities and chaos in single-mode laser systems in which the gain is inhomogeneously broadened, an example being the familiar He–Ne laser.

Erratic and aperiodic temporal behavior of any of the system's variables implies a corresponding continuous spectrum for its Fourier transform, which is thus a further signature of chaotic motion. Time series, power spectra, and routes to chaos collectively provide evidence of deterministic behavior. Of the wide range of possible routes to chaos, three have emerged as particularly common and are frequently observed in lasers. These are period doubling, intermittency, and two-frequency scenarios. In the first, a solution, which is initially stable is found to oscillate, the period of which successively doubles at distinct values of the control parameter. This continues until the number of fixed points becomes infinite at a finite parameter value, where the variation in time of the solution becomes irregular. For the intermittency route, a signal that behaves regularly in time becomes interrupted by statistically distributed periods of irregular motion, the average number of which increases with the external control parameter until the condition becomes chaotic. The two-frequency route is more readily identified with early concepts of turbulence, considered to be the limit of an infinite sequence of instabilities (Hopf bifurcation) evolving from an initial stable solution each of which creates a new basic frequency. It is now known that only two or three instabilities (frequencies) are sufficient for the subsequent generation of chaotic motion.

Applications of Chaos in Lasers

It has been accepted as axiomatic since the discovery of chaos that chaotic motion is in general neither predictable nor controllable. It is unpredictable because a small disturbance will produce exponentially growing perturbation of the motion. It is uncontrollable because small disturbances lead to other forms of chaotic motion and not to any other stable and predictable alternative. It is, however, this very sensitivity to small perturbations that has been more recently used to stabilize and control chaos, essentially using chaos to control itself. Among the many methods proposed in the late 1980s, a feedback control approach was proposed by Ott, Grebogi, and Yorke (OGY) in which tiny feedback was used to stabilize unstable periodic orbits or fixed points of chaotic attractors. This control strategy can be best understood by a schematic of the OGY control algorithm for stabilizing a saddle point P^*, as shown in Figure 3. Curved trajectories follow a stable (unstable) manifold towards (away from) the saddle point. Without perturbation of a parameter, the starting state, s_1, would evolve to the state s_2.

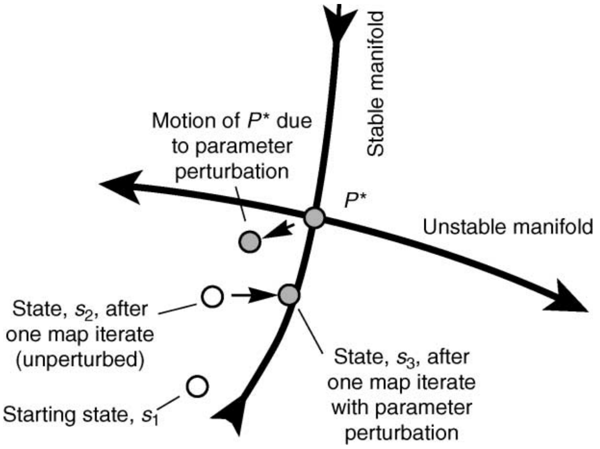
Stable manifold
Motion of P^* due to parameter perturbation
P^*
Unstable manifold
State, s_2, after one map iterate (unperturbed)
Starting state, s_1
State, s_3, after one map iterate with parameter perturbation

Figure 3 The OGY control strategy for stabilizing an orbit in a chaotic attractor.

The effect of changing a parameter of the system is depicted as shifting states near P^* along the solid black arrows, whereas the combination of the unperturbed trajectory and the effect of the perturbation is to bring the state to a point, s_3, on the stable manifold. Once on the stable manifold, the trajectory naturally tends towards the desired point. This algorithm has been successfully applied in numerous experimental systems without *a priori* modeling of these systems, examples being in cardiology, electronics, and lasers. Based on the concept of feedback control, various other approaches have been developed where emphasis has been given to algorithms which are more readily implemented in practical systems, in particular that utilizing occasional proportional feedback (OPF). These pioneering studies have since inspired prolific activities, both in theory and experiment, of control of chaos across many disciplines, opening up possibilities of utilizing chaos in many diverse systems. In optics, Roy *et al.* first demonstrated dynamical control of an autonomously chaotic and high-dimensional laser system on microsecond time-scales. The laser, a diode pumped solid-state Nd:YAIG system with an intracavity KTP doubling crystal, exhibited chaotic behavior from coupling of the longitudinal modes through nonlinear sum-frequency generation. To control the system, the total laser output was sampled within a window of selected offset and width. A signal proportional the deviation of the sampled intensity from the center of the window was generated and applied to perturb the driving current of the diode laser. The sampling frequency is related to the relaxation oscillation frequency of the system. This control signal repeatedly attempts to bring the system closer to a periodic unstable orbit that is embedded in the chaotic attractor, resulting in a realization of the periodic orbit. By adjusting the frequency of the feedback signal they observed the basic (relaxation oscillation) and many higher-order periodic waveforms of the laser intensity. In a subsequent experiment, Gills *et al.* showed that an unstable steady state of this laser could also be stabilized by the OPF control technique and tracked with excellent stability to higher output power with increase of the pump excitation.

Another area to receive considerable attention in the last few years is that of synchronization of chaos. It may be expected that chaotic systems will defy synchronization because two such identical systems started at nearly the same initial conditions have dynamical trajectories that quickly becomes uncorrelated. However, it has now been widely demonstrated that when these systems are linked, their chaotic trajectories converge to be the same and remain in step with each other. Further, such synchronization is found to be structurally stable and does not require the systems to be precisely identical. Not surprisingly, these findings have attracted interest from the telecommunication community; the natural masking of information by chaotic fluctuations offering a means to a certain degree of security. In this new approach, a chaotic carrier of information can be considered as a generalization of the more traditional sinusoidal carrier. In communication systems that use chaotic waveforms, information can be recovered from the carrier using a receiver synchronized, or tuned, to the dynamics of the transmitter. Optical systems are particularly attractive since they display fast dynamics, offering the possibility of communication at bandwidths of a hundred megahertz or higher. Van Wiggeren and Roy first demonstrated data transmission rates of 10 Mbites per second using erbium-doped fiber ring lasers. These lasers are particularly well-suited for communication purposes since their lasing wavelength is close to the minimum-loss wavelength in optical fiber. Figure 3 shows a schematic of this system. The tiny message, in order of one thousandth of the carrier intensity, is decoded in the transmitter. The receiver tends to synchronize its behavior to the chaotic part of the transmitted wave (but not the message). Subtracting the waveform created in the receiver from the transmitted signal recovers the tiny message. In principle, it is possible to communicate information at ultrahigh data rate with the use of this scheme because the chaotic dynamics in the ring laser has a very large spectral width. In a later experiment they showed that a receiver can recover information at 126 Mbits/sec from the chaotic carrier (Figure 4).

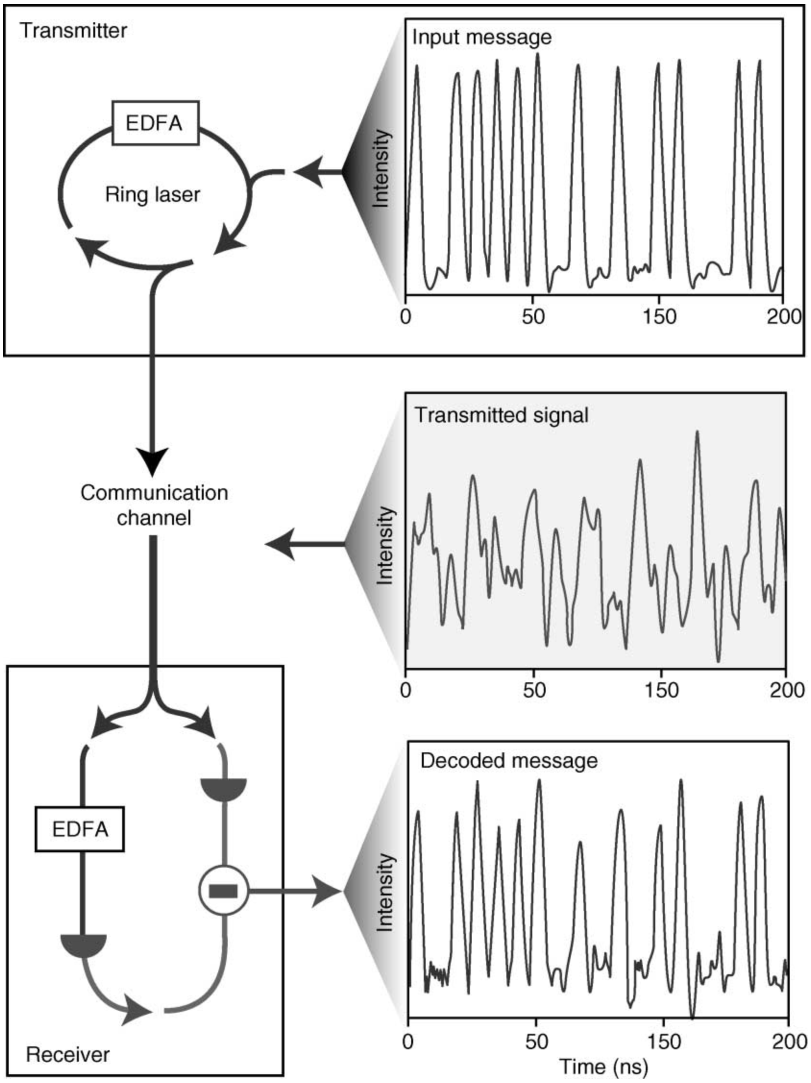
Transmitter
EDFA
Ring laser
Input message
Intensity
0
50
150
200
Communication channel
Transmitted signal
Intensity
0
50
150
200
EDFA
Receiver
Decoded message
Intensity
0
50
150
200
Time (ns)

Figure 4 Schematic of communication with chaotic erbium-doped fiber amplifiers (EDFAs). Injecting a message into the transmitter laser folds the data into the chaotic frequency fluctuations. The receiver reverses this process, thereby recovering a high-fidelity copy of the message. (Reproduced with permission from Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.)

See also

Fourier Optics. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Polarization:** Introduction.

Further Reading

Abraham NB, Lugiato LA and Narducci LM (eds) (1985) Instabilities in active optical media. *Journal of the Optical Society of America B* 2: 1–272.

Arecchi FT and Harrison RG (1987) *Instabilities and Chaos in Quantum Optics*. Synergetics Series 34, 1–253. Berlin: Springer Verlag.

Arecchi FT and Harrison RG (eds) (1994) *Selected Papers on Optical Chaos*. SPIE Milestone Series, vol. MS 75, SPIE Optical Engineering Press.

Gauthier DJ (1998) Chaos has come again. *Nature* 279: 1156–1157.

Gills Z, Iwata C and Roy R (1992) Tracking unstable steady states: extending the stability regime of a multimode laser system. *Physical Review Letters* 69: 3169–3172.

Haken H (1985) Light. In: *Laser Light Dynamics*, vol. 2. Amsterdam: North-Holland.

Harrison RG (1988) Dynamical instabilities and chaos in lasers. *Contemporary Physics* 29: 341–371.

Harrison RG and Biswas DJ (1985) Pulsating instabilities and chaos in lasers. *Progress in Quantum Electronics* 10: 147–228.

Hunt ER (1991) Stabilizing high-period orbits in a chaotic system – the diode resonator. *Physical Review Letters* 67: 1953–1955.

Ott E, Grebogi C and Yorke JA (1990) Controlling chaos. *Physical Review Letters* 64: 1196–1199.

Pecore LM and Carrol TL (1990) Synchronization of chaotic systems. *Physical Review Letters* 64: 821–824.

Roy R, *et al.* (1992) Dynamical control of a chaotic laser: experimental stabilization of a globally coupled system. *Physical Review Letters* 68: 1259–1262.

VanWiggeren GD and Roy R (1998) Communication with chaotic lasers. *Science* 279: 1198–1200.

VanWiggeren GD and Roy R (1998) Optical communication with chaotic waveforms. *Physical Review Letters* 279: 1198–1200.

CHEMICAL APPLICATIONS OF LASERS

Contents

Detection of Single Molecules in Liquids
Diffuse-Reflectance Laser Flash Photolysis
Laser Manipulation in Polymer Science
Nonlinear Spectroscopies
Photodynamic Therapy of Cancer
Pump and Probe Studies of Femtosecond Kinetics
Time-Correlated Single-Photon Counting
Transient Holographic Grating Techniques in Chemical Dynamics

Detection of Single Molecules in Liquids

A J de Mello, J B Edel and E K Hill, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A significant challenge facing experimentalists in the physical and biological sciences is the detection and identification of single molecules. The ability to perform such sensitive and selective measurements is extremely valuable in applications such as DNA analysis, immunoassays, environmental monitoring, and forensics, where small sample volumes and low analyte concentrations are the norm. More generally, most experimental observations of physical systems provide a measurement of ensemble averages, and yield information only on mean properties. In contrast, single molecule measurements permit observation of the interactions and behavior of a heterogeneous population in real time.

Over the past few years a number of techniques with sufficient sensitivity have been developed to detect single molecules. Scanning probe microscopies (most notably scanning tunneling microscopy and atomic force microscopy) have been used to great effect in the analysis of surface bound species, but for the detection of single molecules in liquids, optical methods incorporating the measurement of absorption or emission processes, have proved most successful.

The Absorption–Emission Cycle

The key concept underlying most emissive approaches to single molecule detection is that a single molecule can be cycled repeatedly between its ground state and an excited electronic state to yield multiple photons. The process can be understood by reference to Figure 1. Fluorescence emission in the condensed phase can be described using a four-step cycle. Excitation from a ground electronic state to an excited state is followed by rapid (internal) vibrational relaxation. Subsequently, radiative decay to the ground state is observed as fluorescence emission and is governed by the excited state lifetime. The final stage is internal relaxation back to the original ground state. Under saturating illumination, the rate-limiting step for this cycle is governed by the fluorescence lifetime (τ_f), which is typically of the order of a few nanoseconds. If a single molecule diffuses through an illuminated zone (e.g., the focus of a laser beam) it may reside in that

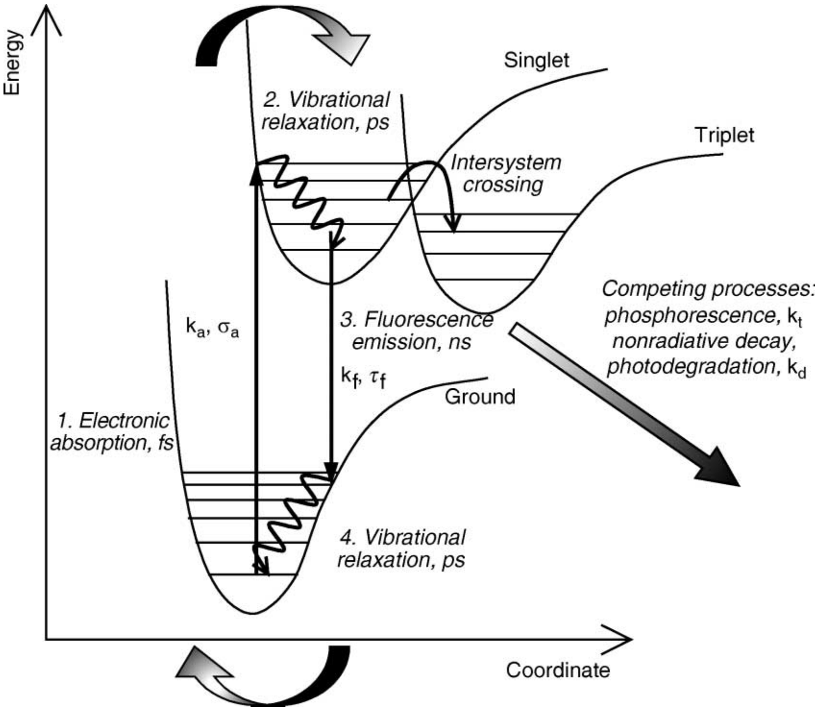
Energy
2. Vibrational relaxation, ps
Singlet
Triplet
Intersystem crossing
Competing processes: phosphorescence, k_t nonradiative decay, photodegradation, k_d
k_a, σ_a
3. Fluorescence emission, ns
k_f, τ_f
Ground
1. Electronic absorption, fs
4. Vibrational relaxation, ps
Coordinate

Figure 1 Schematic illustration of the molecular absorption–emission cycle and timescales for the component processes. Competing processes that may reduce the ultimate photon yield are also shown.

region for several milliseconds. The rapid photon absorption-emission cycle may therefore be repeated many times during the residence period, resulting in a burst of fluorescence photons as the molecule transits the beam. The burst size is limited theoretically by the ratio of the beam transit time (τ_r) and the fluorescence lifetime:

$$N_{\text{photons}} = \frac{\tau_r}{\tau_f} \qquad [1]$$

For typical values of τ_r (5 ms) and τ_f (5 ns) up to one million photons may be emitted by the single molecule. In practice, photobleaching and photodegradation processes limit this yield to about 10^5 photons. Furthermore, advances in optical collection and detection technologies enable registration of about 1–5% of all photons emitted. This results in a fluorescence burst signature of up to a few thousand photons or photoelectrons.

Successful single molecule detection (SMD) depends critically on the optimization of the fluorescence burst size and the reduction of background interference from the bulk solvent and impurities. Specifically a molecule is well suited for SMD if it is efficiently excited by an optical source (i.e., possesses a large molecular absorption cross-section at the wavelength of interest), has a high fluorescence quantum efficiency (favoring radiative deactivation of the excited state), has a short fluorescence lifetime, and is exceptionally photostable.

Ionic dyes are often well suited to SMD as fluorescence quantum efficiencies can be close to unity and fluorescence lifetimes below 10 nanoseconds. For example, xanthene dyes such as Rhodamine 6G and tetramethyl-rhodamine isothiocyanate are commonly used in SMD studies. However, other highly fluorescent dyes such as fluorescein are unsuitable for such applications due to unacceptably high photodegradation rate coefficients. Furthermore, some solvent systems may enhance nonradiative processes, such as intersystem crossing, and yield significant reduction in the photon output. Structures of three common dyes suitable for SMD are shown in Figure 2.

Signal vs Background

The primary challenge in SMD is to ensure sufficient reduction in background levels to enable discrimination between signal and noise. As an example, in a 1 nM aqueous dye solution each solute molecule occupies a volume of approximately 1 fL. However, this same volume also contains in excess of 10^{10} solvent molecules. Despite the relatively small scattering cross-section for an individual water molecule ($\sim 10^{-28}$ cm^2 at 488 nm), the cumulative scattering

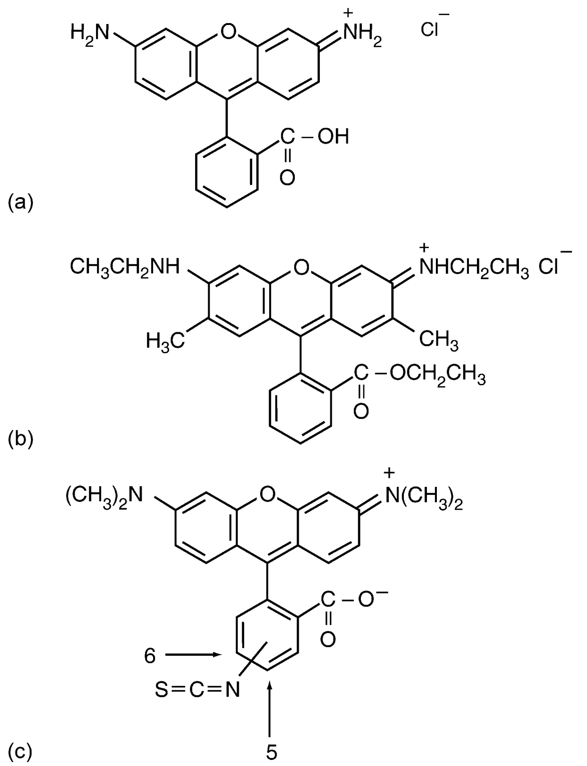

Figure 2 Structures of common dye molecules suitable for SMD in solution: (a) 3,6-diamino-9-(2-carboxyphenyl)-chloride (rhodamine 110); (b) 9-(2-(ethoxycarbonyl)phenyl)-3,6-bis (ethylamino)-2,7-dimethyl chloride (rhodamine 6G); (c) 9-(2-carboxy-isothiocyanatophenyl)-3,6-bis(dimethylamino)-inner salt (tetra methylrhodamine-5-(and-6)-isothiocyanate).

signal from the solvent may swamp the desired fluorescence signal. The principal method of reducing the solvent background is to minimize the optical detection volume: the signal from a single molecule is independent of probe volume dimensions, but the background scales proportionally with the size of the detection region. Although there are several experimental approaches to SMD in solution, several factors hold common:

1. Tiny detection volumes (10^{-12}–10^{-15} L) are used to reduce background signals.
2. A low analyte concentration combined with the small observation volume, ensures that less than one analyte molecule is present in the probe volume on average.
3. High-efficiency photon collection (optics) and detection maximize the proportion of the isotropic fluorescence burst that is registered.
4. Background reduction methods are employed to improve signal-to-noise ratios. These include: optical rejection of Raman and Rayleigh scatter, time-gated discrimination between prompt scatter and delayed emission, and photobleaching of the solvent immediately before detection.

The minute volumes within which single molecules are detected can be generated in a variety of ways. Picoliter volumes can be defined by mutually orthogonal excitation and detection optics focused in a flowing stream. Much smaller, femtoliter probe volumes are generated using confocal microscopes. At this level, background emission is significantly reduced and high signal-to-noise ratios can be achieved. Confocal detection techniques are versatile and have been widely adopted for SMD in freely diffusing systems. Consequently, confocal methods will be discussed in detail in this article. The other general approach to performing SMD in solution involves the physical restriction of single molecules within defined volumes. Of particular note are techniques where single molecules are confined within a stream of levitated microdroplets. Droplet volumes are typically less than 1 fL and imaging of the entire microdroplet enables single molecule fluorescence to be contrasted against droplet 'blanks' with good signal-to-noise. Furthermore, since molecules are confined within discrete volumes, the technique can be utilized for high-efficiency molecular counting applications. More recently, spatial confinement of molecules in capillaries and microfabricated channels (with submicron dimensions) has been used to create probe volumes between 1 fL and 1 pL, and immobilized molecules on surfaces have been individually probed using wide-field microscopy with epi-illumination or evanescent wave excitation.

Single Molecule Detection using Confocal Microscopy

As previously stated, the confocal fluorescence microscope is an adaptable and versatile tool for SMD. In its simplest form, a confocal microscope is one in which a point light source, a point focus in the object plane, and a pinhole detector are all confocal with each other. This optical superposition generates superior imaging properties and permits definition of ultra-small probe volumes. The concepts behind a basic confocal microscope and its use in SMD are schematically illustrated in Figure 3. Coherent light (typically from a laser and tuned to an optical transition of the molecule under investigation) behaves as a point light source and is focused into a sample chamber using a high-numerical aperture objective lens. As a single molecule traverses the laser beam it is continuously cycled between the ground and an excited electronic state, emitting a burst of fluorescence photons. Fluorescence emission is isotropic (spontaneous emission), so photons are emitted in all directions (4π steradians).

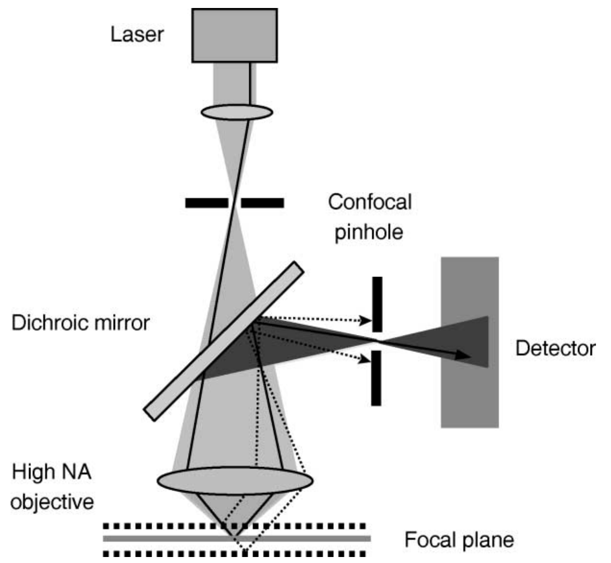

Figure 3 Principle of confocal detection. A confocal pinhole only selects light that emanates from the focal region. Dashed lines indicate paths of light sampled above and below the focal plane that are rejected by the pinhole. The solid ray derives from the focal point, and is transmitted through the pinhole to a detector.

Consequently, the high numerical aperture is used to collect as large a fraction of photons emitted from the focal plane as possible. Designing the objective to be used with an immersion medium, such as oil, glycerin, or water, can dramatically increase the objective numerical aperture, and thus the number of collected photons. Light collected by the objective is then transmitted towards a dichroic beam splitter. In the example shown, fluorescence photons (of lower energy) are reflected towards the confocal detector pinhole, whereas scattered radiation (of higher energy) is transmitted through the dichroic towards the light source. Creation of a precise optical probe volume is effected through the definition of the confocal pinhole. The detector is positioned such that only photons that pass through the pinhole are detected. Consequently, light emanating from the focal plane in the sample is transmitted through the pinhole and detected, whereas light not deriving from the focal plane is rejected by the aperture, and therefore not detected (Figure 3).

To ensure that the maximum number of photons are detected by the system, high efficiency detectors must be used. Photomultiplier tubes (the most common detectors for light-sensing applications) are robust and versatile but have poor detection efficiencies (approximately 5% of all photons that fall on the photocathode yield an electrical signal). Consequently, the most useful detectors for SMD (or low light level) applications are single photon-counting avalanche photodiodes (SPADs). A SPAD is essentially a p–n junction reverse biased above the breakdown voltage, that sustains an avalanche current when triggered by a photon-generated carrier. Detection efficiencies for typical SPADs are normally between 60–70% and are thus ideal for SMD in solution. An approximation of the overall detection efficiency of a confocal system for SMD can be made using eqn [2], which incorporates an estimation of photon losses at all stages of the collection/detection process. Typical transmission efficiencies for each step are also shown.

$$\begin{array}{ccccccccc}
\text{overall} & & \text{objective} & & \text{dichroic} & & \text{additional} & & \text{detector}\\
\text{detection} & \approx & \text{collection} & \times & \text{transmission} & \times & \text{optical} & \times & \text{efficiency}\\
\text{efficiency} & & \text{efficiency} & & \text{coefficient} & & \text{losses} & & \\
 & & & & & & & & \\
\mathbf{0.06} & & 0.24 & & 0.9 & & 0.5 & & 0.6
\end{array} \qquad [2]$$

Optical Probe Volumes

The precise nature of the probe volume is determined by the image of the pinhole in the sample and the spherical aberration of the microscope objective. Routinely, confocal probe volumes are approximated as resembling a cylinder with a radius defined by the diffraction-limited waist of a Gaussian beam. This approximation is useful when the incident beam is narrow or not tightly focused. However, when the radius of the incident beam is large, the corresponding diffraction limited focus is narrowed, and the probe volume more closely resembles a pair of truncated cones. Figure 4a illustrates the dependence of the curvature of the $1/e^2$ intensity contour on the collimated beam radius.

Consequently, it is clear that a simple cylindrical approximation for the probe volume breaks down for wide, tightly focused beams. If a broad incident beam (diameter >1.5 mm) is used, a large noncylindrical contribution to the probe volume is anticipated and a more appropriate model is required. An alternative and more accurate model for the confocal probe volume considers the Gaussian profile of the focused beam. The $1/e^2$ intensity contour radius of a Gaussian waveform with wavelength λ, at some distance z from the beam waist radius w_0, is given by eqn [3]:

$$w(z) = w_0\sqrt{1 + \left(\frac{\lambda z}{\pi w_0^2}\right)^2} \qquad [3]$$

In this case, the probe volume V is given by the volume of rotation of $w(z)$ around the z-axis between Z' and $-Z'$. The volume of rotation can therefore be

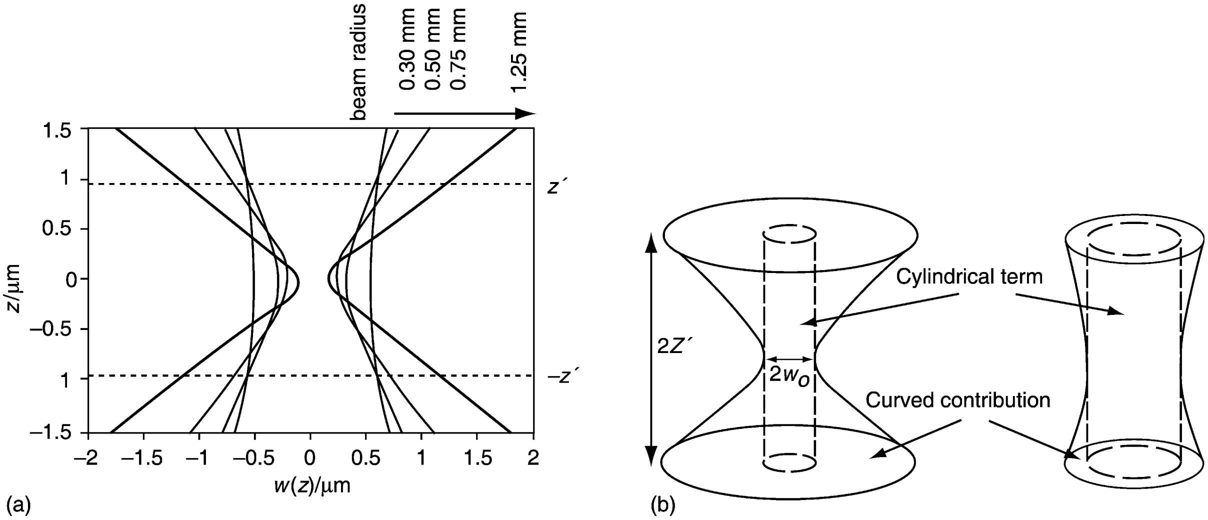

Figure 4 (a) $1/e^2$ Gaussian intensity contours plotted for a series of laser beam radii ($\lambda = 488$ nm, $f = 1.6$ mm, $n = 1.52$). (b) Cylindrical and curved components of the Gaussian probe volume. The curved contribution is more significant for larger beam radii and correspondingly tight beam waists.

simply defined according to,

$$V = \int_{-Z'}^{Z'} \pi w(z)^2 dz \qquad [4]$$

Solution of eqn [4] yields

$$V = 2\pi w_0^2 Z' + \frac{2\lambda^2}{3\pi w_0^2} Z'^3 \qquad [5]$$

The Gaussian volume expression contains two terms. The first term, $2\pi w_0^2 Z'$, corresponds to a central cylindrical volume; the second term has a more complex form that describes the extra curved volume (Figure 4b). The diffraction-limited beam waist radius w_0 can be defined in terms of the focusing objective focal length f, the refractive index n, and the collimated beam radius R according to

$$w_0 = \frac{\lambda f}{n\pi R} \qquad [6]$$

Substitution in eqn [5] yields,

$$\begin{aligned} V &= 2\pi\left(\frac{\lambda f}{n\pi R}\right)^2 Z' + \frac{2\lambda^2}{3\pi}\left(\frac{n\pi R}{\lambda f}\right)^2 Z'^3 \\ &= \frac{2\lambda^2 f^2}{\pi n^2 R^2} Z' + \frac{2\pi n^2 R^2}{3 f^2} Z'^3 \end{aligned} \qquad [7]$$

The volume is now expressed in terms of identifiable experimental variables and constants. Once again, the first term may be correlated with the cylindrical contribution to the volume, and the second term is the additional volume due to the curved contour. It is clear from Figure 4a that, for a given focal length, the noncylindrical contribution to the probe volume increases with incident beam diameter, when the diffraction limited focus is correspondingly sharp and narrow. Furthermore, it can also be seen that the second term in eqn [7] is inversely proportional to f^2, and thus the extent to which the probe volume is underestimated by the cylindrical approximation increases with decreasing focal length. This fact is significant when performing confocal measurements, since high numerical aperture objectives with short focal lengths are typical. Some realistic experimental parameters give an indication of typical dimensions for the probe volume in confocal SMD systems. For example, if $\lambda = 488$ nm, $f = 1.6$ mm, $Z' = 1.0$ μm and $n = 1.52$, a minimum optical probe volume of 1.1 fL is achievable with a collimated beam diameter of 1.1 mm.

Intensity Fluctuations: Photon Burst Statistics

When sampling a small volume within a system that may freely exchange particles with a large surrounding analyte bath, a Poisson distribution of particles is predicted. A Poisson distribution is a discrete series that is defined by a single parameter μ equating to the mean and variance of the distribution:

$$P(n = x) = \frac{\mu^x e^{-\mu}}{x!} \qquad [8]$$

Common Poisson processes include radioactive disintegration, random walks and Brownian motion. Although particle number fluctuations in the excitation volume are Poissonian in nature, the corresponding fluorescence intensity modulation induces a stronger correlation between photon counts. For a single molecular species the model is described by two parameters: an intrinsic molecular brightness and the average occupancy of the observation volume. A super-Poissonian distribution has a width or variance that is greater than its mean; in a Poisson distribution the mean value and the variance are equal. The fractional deviation Q is defined as the scaled difference between the variance and the expectation value of the photon counts, and gives a measure of the broadening of the photon counting histogram (PCH). Q is directly proportional to the molecular brightness factor ε and the shape factor γ of the optical point spread function. (γ is constant for a given experimental geometry.)

$$Q = \frac{\langle \Delta n^2 \rangle - \langle n \rangle}{\langle n \rangle} = \gamma \varepsilon \qquad [9]$$

A pure Poisson distribution has $Q = 0$; for super-Poissonian statistics $Q > 0$. Deviation from the Poisson function is maximized at low number density and high molecular brightness.

In a typical SMD experiment raw data are generally collected with a multichannel scalar and photons are registered in binned intervals. Figure 5 illustrates typical photon burst scans demonstrating the detection of single molecules (R-phycoerythrin) in solution. Fluorescence photon bursts, due to single molecule events, are clearly distinguished above a low background baseline (top panel) of less than 5 counts per channel in the raw data. It is noticeable that bursts vary in both height and size. This is in part due to the range of possible molecular trajectories through the probe volume, photobleaching kinetics, and the nonuniform illumination intensity in the probe region. In addition, it can be seen that the burst frequency decreases as bulk solution concentration is reduced. This effect is expected since the properties of any given single-molecule event are determined by molecular parameters alone (e.g., photophysical and diffusion constants) and concentration merely controls the frequency/number of events.

Although many fluorescence bursts are clearly distinguishable from the background, it is necessary to set a count threshold for peak discrimination in order to correctly identify fluorescence bursts above the background. A photocount distribution can be used as the starting point for determining an appropriate threshold for a given data set. The overlap between signal and background photocount distributions affects the efficiency of molecular detection. Figure 6 shows typical signal and background photocount probability distributions, with a threshold set at approximately 2 photocounts per channel. The probability of spurious (or false)

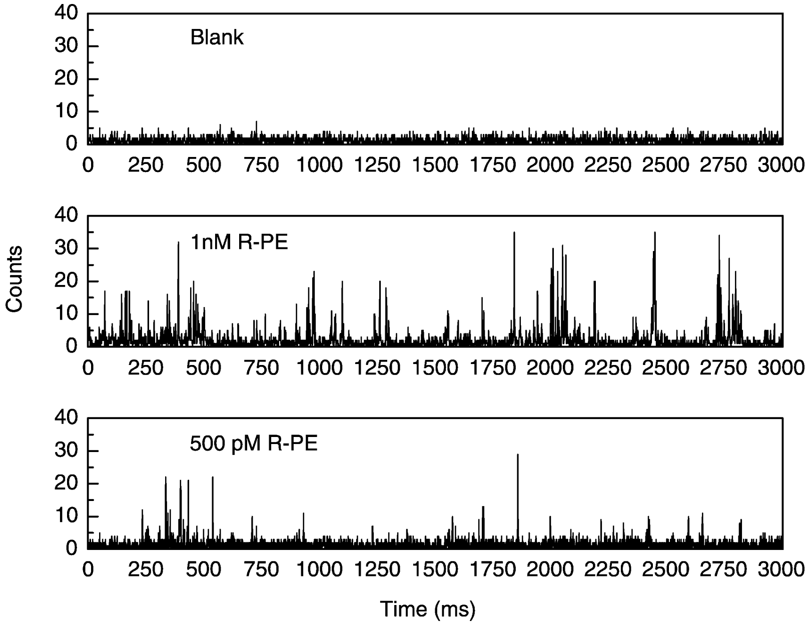

Figure 5 Photon burst scans originating from 1 nM and 500 pM R-phycoerythrin buffered solutions. Sample is contained within a 50 μm square fused silica capillary. Laser illumination = 5 μW, channel width = 1 ms. The top panel shows a similar burst scan originating from a deionized water sample measured under identical conditions.

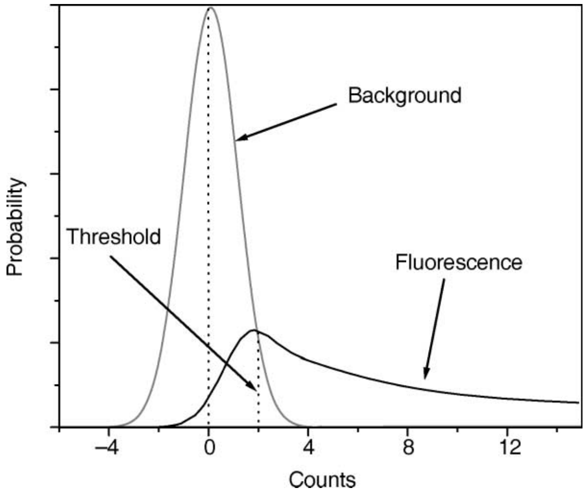

Figure 6 Simulated fluorescence and background photocount probability distributions. The vertical dashed line at 2 counts represents an arbitrarily defined threshold value for peak determination.

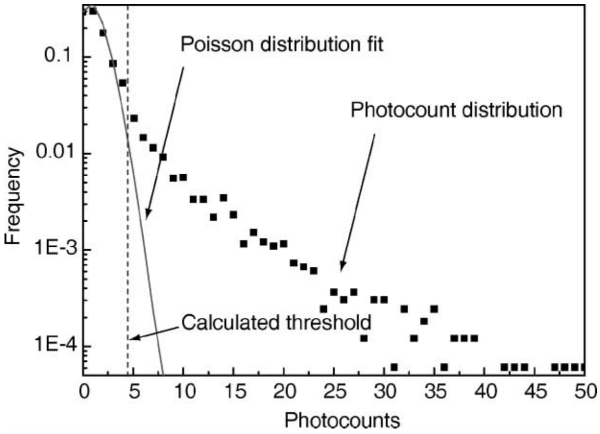

Figure 7 A photon counting histogram generated from a 16 second photon burst scan originating from a 10 μg/mL solution of 1000 nm fluorescent microbeads. The dotted curve shows a least-squares fit of early channels to a Poisson distribution, and the dashed vertical line marks the peak threshold (defined as $\mu + 3\sqrt{\mu} = 4.47$ counts).

detection resulting from statistical fluctuations in the background can be quantified by the area under the 'background' curve at photocount values above the threshold. Similarly, the probability that 'true' single molecule events are neglected can be estimated by the area under the 'fluorescence' curve at photocount values below the threshold. Choice of a high threshold value will ensure a negligible probability of calling a false positive, but will also exclude a number of true single molecule events that lie below the threshold value. Conversely, a low threshold value will generate an unacceptably high number of false positives. Consequently, choice of an appropriate threshold is key in efficient SMD.

Since the background shot noise is expected to exhibit Poisson statistics, the early part of the photocount distribution (i.e., the portion that is dominated by low, background counts) can be modeled with a Poisson distribution, to set a statistical limit for the threshold. Photon counting events above this threshold can be defined as photon bursts associated with the presence of single molecules. In an analogy with Gaussian systems the selected peak discrimination threshold can be defined as three standard deviations from the mean count rate:

$$n_{\text{threshold}} = \mu + 3\sqrt{\mu} \qquad [10]$$

Adoption of a threshold that lies 3 standard deviations above the mean yields a confidence limit that is typically greater than 99%. Figure 7 illustrates a sample photocount distribution, a least-squares fit to an appropriate Poisson distribution, and the calculated threshold that results. Once the threshold has been calculated, its value is subtracted from all channel counts and a peak search utility used to identify burst peaks in the resulting data set.

Data Filtering

As stated, the primary challenge in detecting single molecules in solution is not the maximization of the detected signal, but the maximization of the signal-to-noise ratio (or the reduction of background interferences). Improving the signal-to-noise ratio in such experiments is important, as background levels can often be extremely high.

Several approaches have been used to smooth SMD data with a view to improving signal-to-noise ratios. However, the efficacy of these methods is highly dependent on the quality of the raw data obtained in experiment. As examples, three common methods are briefly discussed. The first method involves the use of a weighted quadratic sum (WQS) smoothing filter. The WQS function creates a weighted sum of adjacent terms according to

$$\tilde{n}_{k,WQS} = \sum_{j=0}^{m-1} w_j (n_{k+j})^2 \qquad [11]$$

The range of summation m is the same order as the burst width, and the weighting factors w_j are chosen

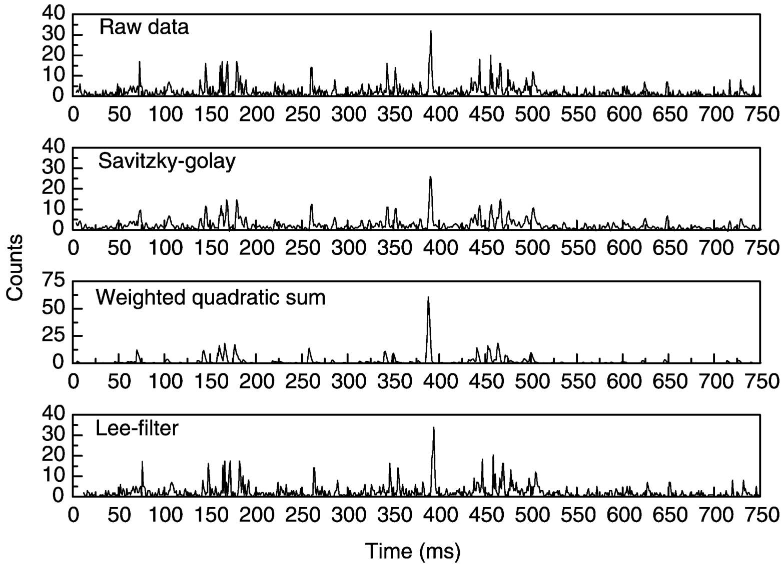

Figure 8 Effect of various smoothing filters on a 750 ms photon burst scan originating from a 1 nM R-phycoerythrin buffered solution. Raw data are shown in the top panel.

to best discriminate between single-molecule signals and random background fluctuations. This method proves most useful for noisy systems, in which the raw signal is weak. There is a practical drawback, in that peak positions are shifted by the smoothing function, and subsequent burst analysis is therefore hampered.

Another popular smoothing filter is the Lee-filtering algorithm. The Lee filter preferentially smoothes background photon shot noise and is defined according to

$$\tilde{n}_k = \bar{n}_k + (n_k - \bar{n}_k)\frac{\sigma_k^2}{\sigma_k^2 + \sigma_0^2} \qquad [12]$$

where the running mean (n_k) and running variance (σ_k^2) are defined by

$$\bar{n}_k = \frac{1}{(2m+1)}\sum_{j=-m}^{m} n_{k+j} \quad m < k \leq N - m \qquad [13]$$

$$\sigma_k^2 = \frac{1}{(2m+1)}\sum_{j=-m}^{m} (n_{k+j} - \bar{n}_{k+j})^2 \quad 2m < k \leq N - 2m \qquad [14]$$

for a filter $(2m+1)$ channels wide. Here, n_k is the number of detected photons stored in a channel k, σ_0 is a constant filter parameter, and N is the total number of data points.

A final smoothing technique worth mentioning is the Savitsky-Golay smoothing filter. This filter uses a least-squares method to fit an underlying polynomial function (typically a quadratic or quartic function) within a moving window. This approach works well for smooth line profiles of a similar width to the filter window and tends to preserve features such as peak height, width and position, which may be lost by simple adjacent averaging techniques. Figure 8 shows the effects of using each approach to improve signal-to-noise for raw burst data.

Photon Burst Statistics

A valuable quantitative analysis method for analysis of fluorescence bursts utilizes the analysis of Poisson statistics. Burst interval distributions are predicted to follow a Poissonian model, in which peak separation frequencies adopt an exponential form. The probability of a single molecule (or particle) event occurring after an interval Δt is given by eqn [15]:

$$N(\Delta t) = \lambda \exp(-\beta t) \qquad [15]$$

where λ is a proportionality constant and β is a characteristic frequency at which single molecule events occur. The recurrence time τ_R can then be simply defined as

$$\tau_R = \frac{1}{\beta} \qquad [16]$$

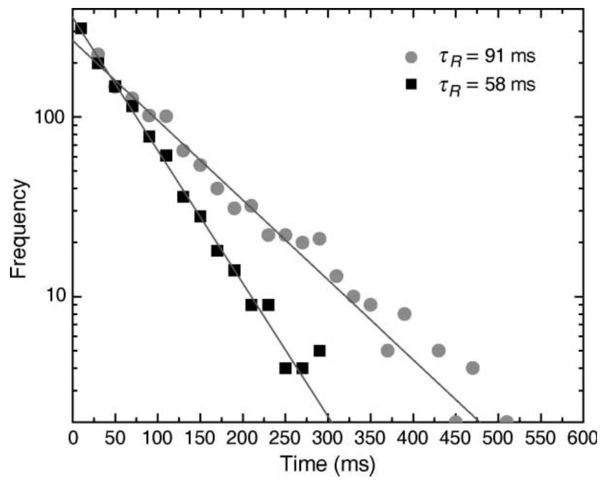

Figure 9 Burst interval distribution analysis of photon burst scans. Data originate from 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 200 nL min^{-1} (circles) and 1000 nL min^{-1} (squares). Least squares fits to a single exponential function are shown by the solid lines.

Equation [6] simply states that longer intervals between photon bursts are less probable than shorter intervals. Furthermore, the recurrence time reflects a combination of factors that control mobility, probe volume occupancy, or other parameters in the single molecule regime. Consequently, it is expected that τ_R should be inversely proportional to concentration, flow rate or solvent viscosity in a range of systems. Figure 9 shows an example of frequency $N(\Delta t)$ versus time plots for two identical particle systems moving at different velocities through the probe volume. A least-squares fit to a single exponential function yields values of $\tau_R = 91$ ms for a volumetric flow rate of 200 nL/min and $\tau_R = 58$ ms for a volumetric flow rate of 1000 nL/min.

Temporal Fluctuations: Autocorrelation Analysis

Autocorrelation analysis is an extremely sensitive method for detecting the presence of fluorescence bursts in single molecule experiments. This approach essentially measures the average of a fluctuating signal as opposed to the mean spectral intensity. As previously discussed, the number of molecules contained within a probe volume at any given time is governed by Poisson statistics. Consequently, the root mean square fluctuation can be defined according to eqn [17],

$$\frac{\sqrt{\langle(\delta N)^2\rangle}}{\langle N\rangle} = \frac{\sqrt{\langle(N-\langle N\rangle)^2\rangle}}{\langle N\rangle} = \frac{1}{\sqrt{\langle N\rangle}} \qquad [17]$$

where N is the number of molecules. It is observed that the relative fluctuation diminishes as the number of particles measured is increased. Hence, it is important to minimize the number of molecules present in the probe volume. It should be noted, however, that if there are too few molecules in the solution there will be long dark periods were no single molecule bursts are observed. If the probe volume is bathed in radiation of constant intensity, fluctuation of the resulting fluorescence signal can simply be defined as deviations from the temporal signal average:

$$\langle F(t)\rangle = \frac{1}{T}\int_0^T F(t)\mathrm{d}t \qquad [18]$$

Here, t is defined as the total measurement time, $F(t)$ is the fluorescence signal at time t, and $\langle F(t)\rangle$ is the temporal signal average. Fluctuations in the fluorescence intensity, $\delta F(t)(\delta F(t) = F(t) - \langle F(t)\rangle)$, with time t, about an equilibrium value $\langle F\rangle$, can be statistically investigated by calculating the normalized autocorrelation function, $G(\tau)$, where

$$G(\tau) = \frac{\langle F(t+\tau)F(t)\rangle}{\langle F\rangle^2} = \frac{\langle \delta F(t+\tau)\delta F(t)\rangle}{\langle F\rangle^2} + 1 \qquad [19]$$

In dedicated fluorescence correlation spectroscopy experiments, the autocorrelation curve is usually generated in real time in a high-speed digital correlator. Post data acquisition calculation is also possible using the following expression

$$G(\tau) = \sum_{t=0}^{N-1} g(t)g(t+\tau) \qquad [20]$$

Here $g(t)$ is the total number of counts during the time interval $(t, t+\Delta t)$, $g(t+\tau)$ is the number of counts detected in an interval of Δt at a later time $t+\tau$, and N is the total number of time intervals in the dataset. In a diffusion controlled system with a single fluorescent molecule that is irradiated with a three dimensional Gaussian intensity profile, the autocorrelation curve is governed by the mean probe volume occupancy N and the characteristic diffusion time (τ_D). The laser beam waist radius ω and the probe depth $2z$ describe the Gaussian profile:

$$G(\tau) = 1 + \frac{1}{N}\left(1 + \frac{\tau}{\tau_d}\right)^{-1}\left(1 + \left(\frac{\omega}{z}\right)^2 \frac{\tau}{\tau_d}\right)^{-1/2} \qquad [21]$$

The diffusion time is a characteristic molecular residence time in the probe volume and inversely

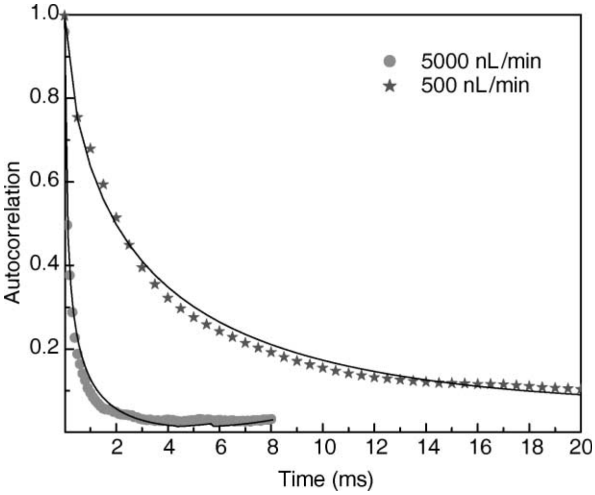

Figure 10 Autocorrelation analysis of photon burst scans of 1 μm fluorescent beads moving through 150 μm wide microchannels at flow rates of 500 nL min^{-1} (stars) and 1000 nL min^{-1} (circles). Solid lines represent fits to the data according to eqn [23].

related to the translational diffusion coefficient for the molecule:

$$D = \frac{w^2}{4\tau_D} \qquad [22]$$

In a flowing system, the autocorrelation function depends on the average flow time through the probe volume τ_{flow}. A theoretical fit to the function can be described according to

$$G(\tau) = 1 + \frac{1}{N} A \exp\left\{ \left(\frac{\tau}{\tau_{\text{flow}}} \right)^2 A \right\}$$

$$A = \left(1 + \frac{\tau}{\tau_d}\right)^{-1} \left(1 + \left(\frac{\omega}{z}\right)^2 \frac{\tau}{\tau_d}\right) \qquad [23]$$

where N is the mean probe volume occupancy; the flow velocity v can then be extracted from the characteristic flow time according to

$$v = \frac{w}{\tau_{\text{flow}}} \qquad [24]$$

It should be noted that in the case that directed flow is negligible or defined to be zero, the autocorrelation function simplifies to eqn [21].

Figure 10 illustrates experimentally determined autocorrelation curves for two identical particle systems moving at different velocities through the probe volume. As particle flow velocity is increased, the width of the autocorrelation curves is seen to narrow as a result of the reduced residence time in the probe volume. A plot of the reciprocal of the full width half maximum of the autocorrelation curve as a function of volumetric flow rate is linear, and provides a simple way of calculating particle/molecule velocities within flowing systems.

Applications

The basic tools and methods outlined in this chapter have been used to perform SMD in a variety of chemically and biologically relevant systems, and indeed there is a large body of work describing the motion, conformational dynamics and interactions of individual molecules (see Further Reading). A primary application area has been in the field of DNA analysis, where SMD methods have been used in DNA fragment sizing, single-molecule DNA sequencing, high-throughput DNA screening, single-molecule immunoassays, and DNA sequence analysis. SMD methods have also proved highly useful in studying protein structure, protein folding, protein-molecule interactions, and enzyme activity.

More generally, SMD methods may prove to be highly important as a diagnostic tool in systems where an abundance of similar molecules masks the presence of distinct molecular anomalies that are markers in the early stages of disease or cancer.

See also

Microscopy: Confocal Microscopy.

Further Reading

Ambrose WP, Goodwin PM, Jett JH, Van Orden A, Werner JH and Keller RA (1999) Single molecule fluorescence spectroscopy at ambient temperature. *Chem. Rev.* 99: 2929–2956.

Barnes MD, Whitten WB and Ramsey JM (1995) Detecting single molecules in liquids. *Anal. Chem.* 67: 418A–423A.

Basche T, Orrit M and Rigler R (2002) *Single Molecule Spectroscopy in Physics, Chemistry, and Biology.* New York: Springer-Verlag.

Zander CJ, Enderlein J and Keller RA (2002) *Single Molecule Detection in Solution: Methods and Applications*. New York: John Wiley & Sons.

Diffuse-Reflectance Laser Flash Photolysis

D R Worrall and S L Williams, Loughborough University, Loughborough, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Flash photolysis was developed as a technique to study short-lived intermediates in photoinduced reactions by George Porter (later Lord Porter of Luddenham) and Ronald Norrish in 1949, drawing on Porter's wartime experience with radar techniques. Such was the impact of this development that it earned the Nobel prize jointly for Porter, Norrish, and Eigen in 1967. In this article we will describe the application of flash photolysis to opaque, scattering samples, detailing how light propagation in such samples can be treated theoretically, and will discuss methods by which the data obtained from diffuse reflectance flash photolysis experiments may be analyzed.

The technique of flash photolysis was originally based on using an intense flash of light (the photolysis flash) from a xenon tube to excite the sample, followed a certain time delay later by a spectroscopic flash of lower intensity from a second flash tube, the light from the latter being detected using a photographic plate. The photolysis flash is of sufficient intensity to produce a large population of intermediates (radicals, ions, excited states, isomers) in the sample, which then absorb light from the spectroscopic flash depending on the concentration of intermediates according to the Beer–Lambert law:

$$A = \varepsilon c l \qquad [1]$$

where A is the sample absorbance, ε the molar absorption coefficient, c the concentration and l the pathlength. The absorbance is related to the incident and transmitted intensities as

$$A = \log_{10} I_0 / I \qquad [2]$$

With I_0 the incident and I the transmitted intensities, there is therefore an exponential fall-off of intensity with pathlength for a homogeneous absorber. Hence, by monitoring the evolution of the absorption spectra, the changes in concentration of the photoproduced intermediates and hence the kinetics of the processes in which they are involved are elucidated.

Flash photolysis has evolved subsequently to make use of laser sources and sophisticated electronic detection apparatus to push the limits of time resolution to the femtosecond regime. Indeed, recently the Nobel prize for chemistry was awarded to Ahmed Zewail for his work with ultrafast pump-probe techniques. However, in its conventional geometry, flash photolysis is limited to transparent samples, since it is necessary to be able to probe the excited species by monitoring the absorption spectra.

Many biological systems and industrially important samples are opaque or highly scattering, and hence the attenuation of light through the sample is no longer described by the Beer–Lambert law. In 1981, Frank Wilkinson and Rudolph Kessler had the idea of using diffusely reflected light to interrogate the changes in concentration within a scattering sample subjected to a high-intensity excitation pulse. When photons enter a sample, they may be absorbed or scattered. Those which are scattered may re-emerge from the irradiated surface as diffusely reflected light. The intensity of diffusely reflected light emerging from the surface at a particular wavelength is a unique function of the ratio of scattering to absorption. The more scattering events occurring before absorption, the more likely the photon is to escape from the sample as diffusely reflected light. Hence the probability of escape decreases as the absorption probability increases, and the diffusely reflected light is deficient in those wavelengths where the absorption is strongest, i.e., the ratio of the incident to absorbed light intensity at a given wavelength is related to the absorption of the sample at that wavelength.

Kubelka–Munk Theory of Reflectance

The theory which describes the relationship between incident and scattered light intensity, absorption and scatter, and concentration which is widely applied in this context, is the Kubelka–Munk theory of reflectance. The theory was originally developed to describe the reflectance characteristics of paint films but it works quite well for many samples containing a homogeneous distribution of scatterers and absorbers. The limiting assumption in this theory is that the scatterers from which the scattering layer is composed are very much smaller than the total layer thickness. Additionally, the layer should be optically thick such that all of the light entering the layer should be either absorbed or reflected, with a negligible fraction transmitted.

For a layer of thickness X diffusely irradiated with monochromatic light, the diagram shown in Figure 1

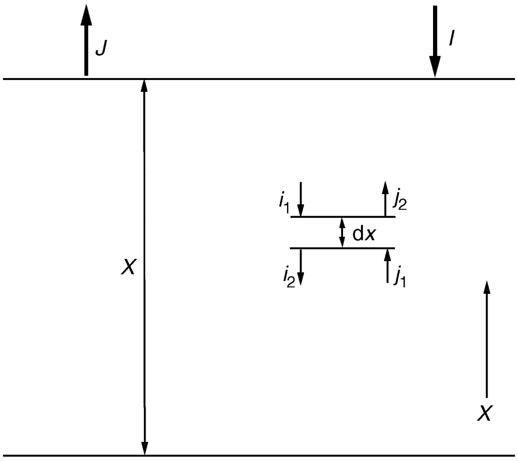

Figure 1 Schematic of counterpropagating fluxes in a diffusely irradiated scattering medium.

can be constructed, with I the incident flux and J the diffusely reflected flux, and i and j the fluxes traveling upwards and downwards through an infinitesimally small thickness element dx. Two further parameters may be defined which are characteristic of the medium.

K The absorption coefficient. Expresses the attenuation of light due to absorption per unit thickness.

S The scattering coefficient. Expresses the attenuation of light due to scattering per unit thickness.

Both of these parameters can be thought of as arising due to the particle (or chromophore) acting as a sphere of characteristic size, which casts a shadow either due to the prevention of on-axis transmission or due to absorption. The scattering or absorption coefficient then depends on the effective size of this sphere, and the number density of spheres in the medium. In each case, the probability P of a photon being transmitted through a particular thickness X of a medium is related exponentially to the absorption or scattering coefficient:

$$P = \exp(-KX) \qquad [3]$$

$$P = \exp(-SX) \qquad [4]$$

The scattering coefficient S depends on the refractive index difference between the particle and the dispersion medium. The scattering coefficient is also dependent upon the particle size, and shows an inverse correlation, i.e., the scattering coefficient increases as the particle size decreases. This effect is a function of the size of the particle relative to the wavelength of light impinging on it, with the scattering coefficient increasing as the wavelength decreases. However, this change with wavelength is small provided the particle size is large relative to the wavelength of the light.

The effect of the material in the element dx on the counterpropagating fluxes i and j depends on the absorption and scattering coefficients. Both i and j will be attenuated by both absorption and scattering:

$$i_2 = i_1 - i_1(S + K)\mathrm{d}x \qquad [5]$$

$$j_2 = j_1 - j_1(S + K)\mathrm{d}x \qquad [6]$$

Both i and j are reinforced by backscattering from the other flux:

$$i_2 = i_1 + j_1 S\mathrm{d}x \qquad [7]$$

$$j_2 = j_1 + i_1 S\mathrm{d}x \qquad [8]$$

The net effect of this in the attenuation of the flux propagating into the sample (i) and the flux backscattered from the sample (j) can be expressed as the following differential equations:

$$\mathrm{d}i = (S + K)i_1\mathrm{d}x - j_1 S\mathrm{d}x \qquad [9]$$

$$\mathrm{d}j = -(S + K)j_1\mathrm{d}x + i_1 S\mathrm{d}x \qquad [10]$$

For a sample of infinite thickness, these equations can be solved to give an analytical solution for the observed reflectance of the sample in terms of the absorption and scattering coefficients:

$$\frac{K}{S} = \frac{(1 - R_\infty)^2}{2R_\infty} \qquad [11]$$

where R_∞ is the reflectance from a sample of such thickness that an increase in sample thickness has no effect on the observed reflectance. The absorption coefficient K is dependent upon the concentration of absorbers in the sample through

$$K = 2\varepsilon c \qquad [12]$$

with ε the naperian absorption coefficient, c the concentration of absorbers and the factor 2 which is the geometrical factor for an isotropic scatterer. For multiple absorbers, K is simply the linear sum of absorption coefficients and concentrations. Hence for a diffusely scattering medium, an expression analogous to the Beer–Lambert law can be derived to relate concentration to a physically measurable parameter, in this case the sample reflectance. The ratio K/S is usually referred to as the Kubelka–Munk remission function, and is the parameter usually quoted in this context. It is important to note that the relationship with concentration is only valid for a homogeneous distribution of absorbers within a sample.

Transient Concentration Profiles

As has been discussed in the previous section, a light flux propagating through a scattering medium is attenuated by both scattering and absorption events, whilst in a nonturbid medium attenuation is by absorption only. Hence even in a sample with a very low concentration of absorbers, the light flux is rapidly attenuated as it penetrates into the sample by scattering events. This leads to a significant flux gradient through the sample, resulting in a reduction in the concentration of photo-induced species as penetration depth into the sample increases. There are three distinct concentration profiles which may be identified within a scattering sample which are dependent on the scattering and absorption coefficients. These concentration profiles are interrogated by a beam of analyzing light, and hence an understanding of the effect of these differing profiles on the diffusely reflected intensity is vital in interpreting transient absorption data. The transient depth profiles are illustrated in Figure 2.

Kubelka–Munk Plug

This occurs when the photolysis flash, i.e., laser fluence is high and the absorption coefficient is low at the laser wavelength. If we assume a simple photophysical model involving simply the ground state S_0, the first excited singlet state S_1 and first excited triplet state T_1, and we assume that either the quantum yield of triplet state production is high or the S_1 lifetime is very much shorter than the laser pulse allowing re-population of the ground state, then it is possible at high enough fluence to completely convert all of the S_0 states to T_1 states within the sample. Provided the fluence is high enough, this complete conversion will penetrate some way into the sample (Figure 2). Under circumstances where the T_1 state absorbs strongly at some wavelength other than the photolysis wavelength, probing at this wavelength will result in the probe beam being attenuated significantly within a short distance of the surface of the sample, and thus will only interrogate regions where there is a homogeneous excited state concentration. Under these circumstances the reflectance of the sample as a function of the concentration of excited states is described by the Kubelka–Munk equation, and the change in remission function can be used to probe the change in excited state concentration.

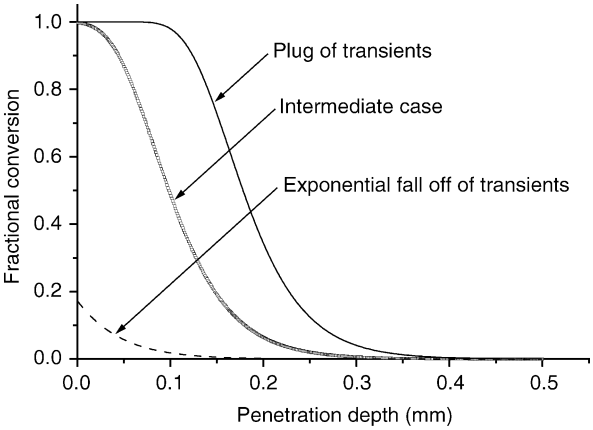

Figure 2 Transient concentration profiles following pulsed laser excitation of a scattering and absorbing sample.

Exponential Fall-Off of Concentration

This occurs when the laser fluence is low and the absorption coefficient at the laser wavelength high. Again considering the simple photophysical model, most of the laser flux will be absorbed by the ground state in the first few layers of sample, and little will penetrate deeply into the sample. In the limiting case this results in an exponential fall-off of transient concentration with penetration depth into the sample (Figure 2). Here the distribution of absorbers is not random and the limiting Kubelka–Munk equation is no longer applicable since the mean absorption coefficient varies with sample penetration depth. Lin and Kan solved eqs [9] and [10] with the absorption coefficient K varying exponentially with penetration depth and showed that the series solution converges for changes in reflectance of less than 10% such that the reflectance change is a linear function of the number of absorbing species.

Intermediate Case

Between the two extremes described above is a case where significant transient conversion takes place at the front surface, but with little penetration into the sample. This can occur, for example, with high laser fluences and large ground state absorption coefficients at the laser wavelength. Under these circumstances, illustrated in Figure 2, the analyzing light interrogates not only the transient concentration profile, but also a significant amount of the analyzing light may penetrate through the transient layer into the unconverted sample behind, if the transient absorption at the analyzing wavelength is low. This creates a more complex problem for analysis since effectively the sample is irradiated from both front and back faces, with consequent effects on the diffusely reflected intensity. It is possible to numerically model the reflectance change as a function of transient concentration under these circumstances but a precise knowledge of the absorption and scattering coefficients is required. Under most circumstances, this case is avoided and diffuse reflectance flash photolysis experiments are arranged such that one of the two limiting cases above, generally the exponential fall-off (usually achieved by attenuation of the laser beam), prevails in a particular experiment.

Sample Geometry

In the case of conventional nanosecond laser flash photolysis, it is generally the case that right-angled geometry is employed for the photolysis and probe beams (Figure 3a). This geometry has a number of advantages over alternatives. The photolysis beam and analyzing beam are spatially well separated, such that the analyzing beam intensity is largely unaffected by scattered photolysis light. Also, the fluorescence generated by the sample will be emitted in all directions, whilst the analyzing beam is usually collimated allowing spatial separation of fluorescence and analyzing light; sometimes an iris is used to aid this spatial discrimination. This geometry is appropriate for quite large beam diameters and fluences; where smaller beams are used collinear geometry may be more appropriate in order to achieve long interaction pathlengths.

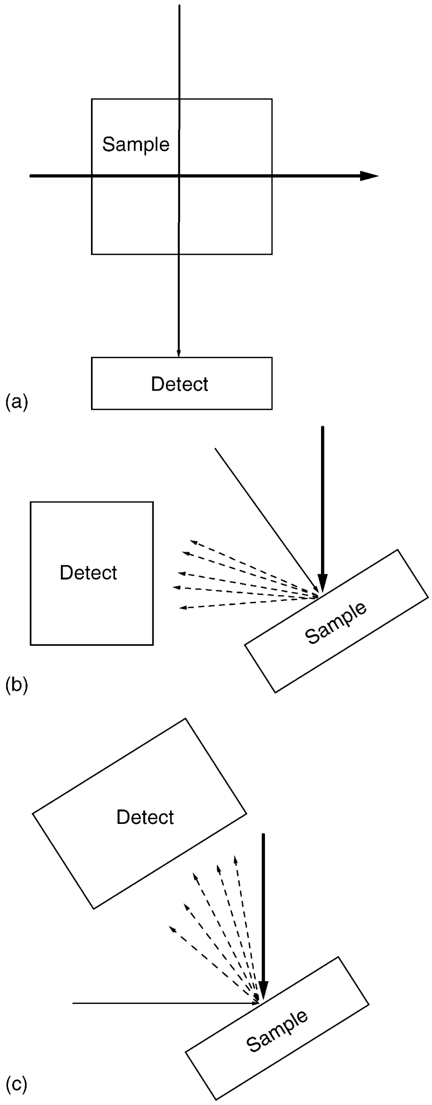

Figure 3 Sample geometries for (a) conventional and (b,c) diffuse reflectance flash photolysis. ⟶, photolysis beam; ⟶, analyzing beam.

In the case of nanosecond diffuse reflectance flash photolysis, the geometry required is quite different (Figure 3b,c). Here, the photolysis beam and analyzing beam must be incident on the same sample surface, and the diffusely reflected analyzing light is collected from the same surface. The geometry is often as shown in Figure 3b, where the analyzing light is incident almost perpendicularly on the sample surface, with the photolysis beam incident at an angle such that the specular reflection of the photolysis beam passes between detector and analyzing beam (not shown). Alternatively, the geometry shown in Figure 3c may be employed, where diffusely reflected light is detected emerging perpendicular to the sample surface. In both cases, the geometry is chosen such that specular (mirror) reflection of either exciting or analyzing light from the sample is not detected, since specular reflection is light which has not penetrated the sample and therefore contains no information regarding the concentrations of species present. A requirement, as in conventional flash photolysis, is that the analyzing beam probes only those areas which are excited by the photolysis beam, requiring the latter to be larger than the former. The nature of the scattering described previously means that the required geometry does not give spatial discrimination at the detector between photolysis and analyzing light, and fluorescence. This is since both analyzing and photolysis light undergo scattering and absorption processes (although with wavelength-dependent absorption and scattering coefficients) and emerge with the same spatial profiles. Fluorescence, principally that stimulated by the photolysis beam since this is of greatest intensity, originates from within the sample but again undergoes absorption and scattering and emerges with the same spatial distribution as the exciting light. In diffuse reflectance flash photolysis, separation of the analyzing light from the excitation or fluorescence must be achieved using spectral (filters and/or monochromators) or temporal, rather than spatial discrimination. Time-gated charge coupled devices

(CCD) or photodiode array detectors can be effectively used in nanosecond laser flash photolysis to exclude excitation light and fluorescence, since these occur on time-scales usually much shorter than the transient absorption of species of interest. The usual analyzing light source used in nanosecond diffuse reflectance flash photolysis is a xenon arc lamp, due to its good spectral coverage and high intensity. High probe intensity is particularly important in diffuse reflectance flash photolysis since the scattered intensities are often low, and the scattered light emerges in a large solid angle, only part of which can be effectively collected and detected. Also, high intensities allow for the light from the analyzing source to dominate over fluorescence if the latter is relatively weak.

In femtosecond diffuse reflectance laser flash photolysis, sample geometry considerations are also important. Such experiments are performed using the pump-probe technique, with the probe often being a femtosecond white-light continuum. The sample geometry employed is illustrated in Figure 4.

Here the pump and probe are almost colinear, and are incident on the same sample area; again, the requirement is for the pumped area to be larger than the probed area. Diffusely reflected light is then collected and analyzed, time resolution being achieved by varying the delay between pump and probe beams. It should be noted that the temporal resolution is worse than in conventional pump-probe techniques. In conventional femtosecond flash photolysis, the time resolution is generally limited by the widths of pump and probe pulses; in diffuse reflectance mode, the pulses undergo numerous refractions, reflections, and diffractions such that the pulse is temporally broadened during its transit through the material. The extent of this broadening is a sensitive function of the optical properties of the individual sample.

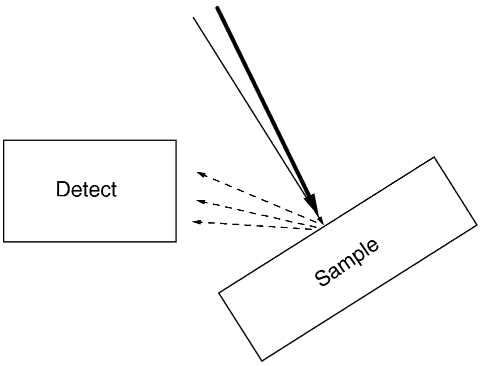

Figure 4 Sample geometry for femtosecond diffuse reflectance flash photolysis. ⟶, photolysis beam; ⟶, analyzing beam.

Kinetic Analysis

Kinetic analysis, and of course time-resolved spectroscopic analysis, require a quantitative treatment of the concentration changes within a sample following an excitation pulse as a function of time. When studying transient absorption phenomena in opaque samples, it is usual to define the reflectance change in terms of the sample reflectance before and after excitation, such that in spectrally resolved data a transient difference spectrum rather than an absolute reflectance spectrum of the transient species is obtained. The latter can, however, be reconstructed from a knowledge of the absorption coefficients and concentration of the species involved. It is possible to define the reflectance change as

$$\frac{\Delta J_t}{J_0} = 1 - \frac{R_t}{R_0} \qquad [13]$$

where R_0 and R_t represent the intensity of probe light diffusely reflected from the sample before excitation and at a time t after excitation, respectively. Frequently reflectance change is expressed as '% absorption', which is defined in eqn [14].

$$\%\ \text{absorption} = 100 \times \left(1 - \frac{R_t}{R_0}\right) \qquad [14]$$

These parameters are often used as being proportional to transient concentration, subject to satisfying the criteria for an exponential fall-off of transient concentration with penetration depth as discussed previously, and are used to replace transient concentration in kinetic equations used in data analysis.

It is generally the case that the samples studied using diffuse reflectance laser flash photolysis have either some degree of heterogeneity, for example paper, microcrystalline cellulose, silica gel or alumina, or have well-defined porous structures such as zeolites. Molecules adsorbed to these supports may be present on the surface, within micro- or mesopores, or intimately included within the structure. Hence each molecule may experience its own unique environment, and this will obviously influence its observed photophysics. It is therefore the case that in these systems, even very simple photo-induced reactions such as unimolecular photoisomerizations do not follow first-order kinetics; rather, a distribution of rates is observed which reflect the differing environments experienced by the molecules, and hence the molecules act as probes of this heterogeneity.

There are a number of approaches to kinetic analysis in opaque, often heterogeneous systems, as described below.

Rate Constant Distributions

Here the sample heterogeneity is treated as producing a series of micro-environments, and the reaction studied will have its own unique rate constant in each of these environments. The width of the observed rate constant distribution therefore describes the number of these possible environments, with the distribution amplitude at any particular rate constant, reflecting the probability of the molecule existing in the corresponding environment.

Exponential series lifetime distribution analysis is the simplest approach to distributed lifetime analysis. This analysis has the inherent advantage that there are no presuppositions regarding the kinetic model describing the data. Rather, a large number of exponentials with fixed rate constants and with amplitudes as adjustable parameters are applied to the data, and a least-squares procedure used to optimize the amplitude of each exponential. Generally there is a relationship between the rate constants, they being equally spaced on either a linear or logarithmic scale. Physically this can be interpreted as a large number of first-order or pseudo-first-order (in the case of bimolecular reactions) rates arising from the sample heterogeneity. Hence the rate constant distribution emerges naturally from the fitting procedure, with no preimposed constraints. Since there are a large number of adjustable parameters, these models are applied most successfully to data with good signal-to-noise.

Alternative approaches involve imposing a distribution shape onto a set of exponentials, and optimizing this distribution to the data. This approach has the advantage that by assuming a rate constant distribution, the relative amplitudes of the exponentials are fixed and hence the number of fitting parameters greatly reduced. One of the more successful models applied in this context is that developed by Albery *et al.* In the development of their model it is assumed that, for the reaction in question, the free energy of activation, $\Delta G^{\neq}$, is distributed normally around a mean value $\Delta\bar{G}^{\neq}$ according to eqn [15]:

$$\Delta G^{\neq} = \Delta\bar{G}^{\neq} - \gamma x RT \qquad [15]$$

with γ being the width of the distribution for $\infty \geq x \geq -\infty$. This assumed normal distribution of the free energy of activation leads to a log-normal distribution of the decay rate constants distributed around some average rate constant $\bar{k}$. The dispersion in the first-order rate constant is then

$$\ln(k) = \ln(\bar{k}) + \gamma x \qquad [16]$$

The equation used to describe the data is given as:

$$\frac{c}{c_0} = \frac{\int_{-\infty}^{\infty} \exp(-t^2)\exp\left[-\bar{k}t\exp(\gamma t)\right]\mathrm{d}t}{\int_{-\infty}^{\infty} \exp(-t^2)\mathrm{d}t} \qquad [17]$$

with c and c_0 being concentration at time $t = t$ and $t = 0$ relative to the excitation flash, respectively. Where reflectance change is proportional to transient concentration, reflectances can be used directly in place of concentrations. This equation can be solved by making the following substitutions:

$$\int_{-\infty}^{\infty} \exp(-t^2)\mathrm{d}t = \pi^{1/2} \qquad [18]$$

$$\int_{-\infty}^{\infty} \exp(-t^2)\exp\left[-\bar{k}t\exp(\gamma t)\right]\mathrm{d}t = \int_0^1 \lambda^{-1}\exp\left(-(\ln\lambda^2)\right)\left(\exp(-\bar{k}t\lambda^{\gamma})+\exp(-\bar{k}t\lambda^{-\gamma})\right)\mathrm{d}\lambda \qquad [19]$$

Hence here there are only two fitting parameters, namely the width of the distribution γ and the distribution center, $\bar{k}$. Note that for $\gamma = 0$, eqn [17] reduces to a single exponential function.

A further model which constrains the rate constant distribution, which has been successfully applied to describe the rates of photo-induced processes on silica gel surfaces, is a Lévy stable distribution of rate constants, described as:

$$\mathrm{P}(\bar{k}, \alpha, \gamma) = \frac{1}{\pi}\int_0^{\infty} \exp(-\gamma q^{\alpha})\cos(\bar{k}\mathrm{q})\mathrm{dq} \qquad [20]$$

where α is the characteristic power law exponent ($0 < \alpha \leq 2$) and γ is the distribution width ($\gamma > 0$). Special cases of the Lévy stable distribution occur for $\alpha = 1$ and $\alpha = 2$, where the distribution shape becomes Lorentzian and Gaussian, respectively. The Lévy distribution gives an increasing weighting to the tails of the distribution as the distribution width decreases, and can be described as a random walk consisting of long jumps followed by several short walks. It has been shown that this type of foraging behavior is more efficient at seeking out randomly distributed targets than a simple random walk. The Lévy stable distribution has three adjustable parameters which allows greater flexibility in the distribution of rate constants than does the Gaussian model, but still constrains the distribution to be symmetrical about some mean value.

Physical Models

A number of these have been applied to very specific data sets, where parameters of the systems are accurately known. These include random walk models in zeolites, and time-dependent fractal-dimensional rate constants to describe kinetics on silica gel surfaces. The advantage of these methods over rate constant distributions is that since they are based on the physical description of the system, they yield physically meaningful parameters such as diffusion coefficients from analysis of kinetic data. However, for many systems the parameters are known in insufficient detail for accurate models to be developed.

Examples

An example of bimolecular quenching data is shown in Figure 5. Here, anthracene at a concentration of 1 μmol g^{-1} is co-adsorbed to silica gel from acetonitrile solution together with azulene at a concentration of 0.8 μmol g^{-1}. Laser excitation at 355 nm from an Nd:YAG laser produces the excited triplet state of the anthracene, which undergoes energy transfer to the co-adsorbed azulene molecule as a result of the rapid diffusion of the latter.

The data shown in Figure 5 are recorded monitoring at 420 nm, and the laser energy (approximately 5 mJ per pulse) is such that an exponential fall-off of transient concentration with penetration depth is expected such that reflectance change is proportional to transient concentration (see section on Transient Concentration Profiles above). The data are shown on a logarithmic time axis for clarity. The fitted line is obtained by applying the model of Albery *et al.* (see section on Rate Constant Distributions above) with fitting parameters $\bar{k} = 1.01 \times 10^4\ s^{-1}$ and $\gamma = 0.73$.

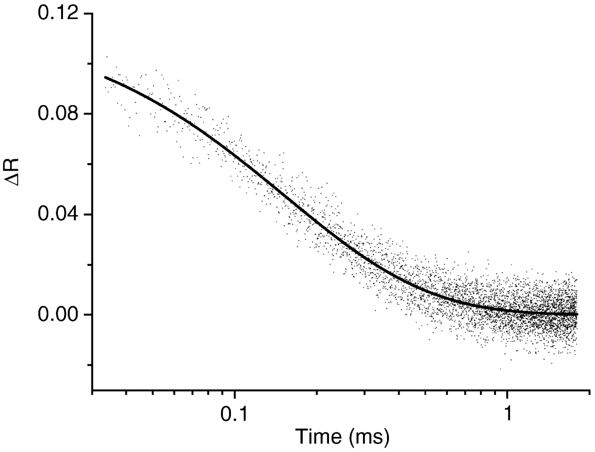

Figure 5 Transient absorption decay for anthracene (1 μmol g^{-1}) co-adsorbed with azulene (0.8 μmol g^{-1}) to silica gel monitored at 420 nm. Fitted using the model of Albery *et al.*

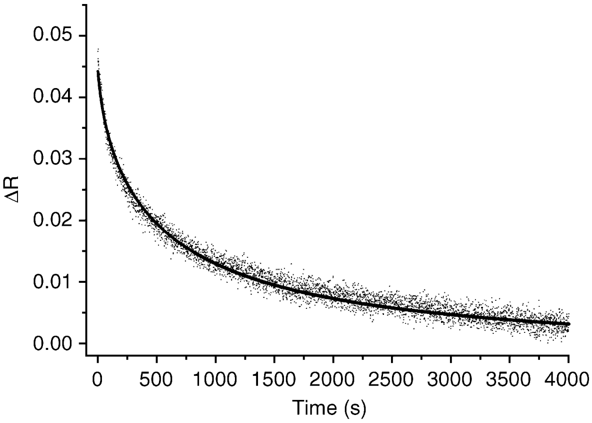

Figure 6 Transient absorption decay for the naphthalene (1 μmol g^{-1}) radical cation monitored at 680 nm on silica gel. Fitted using a Lévy stable distribution.

Where ion–electron recombination is concerned, the Albery model often fails to adequately describe the data obtained since it does not allow sufficient small rate constants relative to the value of $\bar{k}$, given the constraints of the Gaussian distribution. The Lévy stable distribution is ideal in this application since it allows greater flexibility in the shape of the distribution.

Figure 6 shows example data for naphthalene adsorbed to silica gel (1 μmol g^{-1}). The laser pulse energy at 266 nm (approximately 40 mJ per pulse) is such that photo-ionization of the naphthalene occurs producing the radical cation. The subsequent decay of the radical cation via ion–electron recombination can be monitored at 680 nm. Note that decay is observed on a time-scale of several thousand seconds. The fitted line is according to a Lévy stable distribution with parameters $\bar{k} = 8.2 \times 10^{-4}\ s^{-1}$, $\gamma = 0.5$ and $\alpha = 1.7$ (see section on Rate Constant Distributions above). This model allows the shape of the distribution to deviate from a Gaussian, and can be more successful than the model of Albery *et al.* in modeling the complex kinetics which arise on surfaces such as silica gel. Note that where the model of Albery *et al.* can successfully model the data, the data can also be described by a Lévy stable distribution with $\alpha = 2$.

List of Units and Nomenclature

A	Sample absorbance
c	Concentration
I	Incident flux
J	Diffusely reflected flux

k	Rate constant
$\bar{k}$	Mean rate constant
K	Absorption coefficient (absorption per unit thickness) [cm^{-1}]
l	Pathlength [cm]
P	Probability of transmission of a photon through a sample of defined thickness
R_∞	Reflectance of an infinitely thick sample
S	Scattering coefficient (scattering per unit thickness) [cm^{-1}]
t	Time
X	Thickness of sample layer [cm]
α	Characteristic Lévy power law exponent
γ	Width of rate constant distribution
$\Delta G^{\neq}$	Free energy of activation [kJ mol^{-1}]
$\Delta\bar{G}^{\neq}$	Mean free energy of activation [kJ mol^{-1}]
ε	Molar decadic or naperian absorption coefficient

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics. **Optical Materials**: Measurement of Optical Properties of Solids. **Scattering**: Scattering from Surfaces and Thin Films; Scattering Theory.

Further Reading

Albery WJ, Bartlett PN, Wilde CP and Darwent JR (1985) A general-model for dispersed kinetics in heterogeneous systems. *Journal of the American Chemical Society* 107: 1854–1858.

Anpo M (ed.) (1996) Surface photochemistry. *Wiley Series in Photoscience and Photoengineering*, vol. 1. Chichester, UK: John Wiley and Sons.

Anpo M and Matsuura T (eds) (1989) Photochemistry on solid surfaces. *Elsevier Series in Studies in Surface Science and Catalysis*, vol. 47. Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Asahi T, Furube A, Fukumura H, Ichikawa M and Masuhara H (1998) Development of a femtosecond diffuse reflectance spectroscopic system, evaluation of its temporal resolution, and applications to organic powder systems. *Review of Scientific Instruments* 69: 361–371.

Bertoin J (1998) *Lévy Processes*. Cambridge, UK: Cambridge University Press.

Hapke B (1993) *Introduction to the Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kamat PV (1993) Photochemistry on nonreactive and reactive (semiconductor) surfaces. *Chemical Reviews* 93: 287–300.

Kan HKA and Lin TP (1970) Calculation of reflectance of a light diffuser with non-uniform absorption. *Journal of the Optical Society of America* 60: 1252–1256.

Kessler RW and Wilkinson F (1981) Diffuse reflectance triplet–triplet absorption spectroscopy of aromatic hydrocarbons chemisorbed on γ-alumina. *Journal of the Chemical Society – Faraday Transactions I* 77: 309–320.

Kortum G (1969) *Reflectance Spectroscopy*. Berlin: Springer-Verlag.

Kubelka P and Munk F (1931) Ein beitrag zur optik der farbanstriche. *zeitschrift fuer Technische Physik* 12: 593–601.

Ramamurthy V (ed.) (1991) *Photochemistry in Organised & Constrained Media*. New York: VCH Publishers.

Thomas JK (1993) Physical aspects of photochemistry and radiation chemistry of molecules adsorbed on SiO_2, γ-Al_2O_3, zeolites and clays. *Chemical Reviews* 93: 301–320.

Thomas JK and Ellison EH (2001) Various aspects of the constraints imposed on the photochemistry of systems in porous silica. *Advances in Colloid and Interface Science* 89–90: 195–238.

Wendlandt WW and Hecht HG (1966) Reflectance Spectroscopy. *Chemical Analysis Series*, vol. 21. New York: Interscience Publishers.

Wilkinson F and Kelly G (1989) Diffuse reflectance flash photolysis. In: Scaiano JC (ed.) *Handbook of Organic Photochemistry*, vol. 1, pp. 293–314. Florida: CRC Press.

Laser Manipulation in Polymer Science

S Ito, Y Hosokawa and H Masuhara, Osaka University, Osaka, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Laser manipulation is a method for manipulating single particles with a size of less than a few tens of microns in which optical pressure of a focused laser beam is used to trap and control the particles without any mechanical contact. As infrared lasers, as a trapping light source, have become more user-friendly, even operators who are not familiar with lasers and microscopes can perform laser manipulation. Those laser manipulators already on sale have attracted significant attention as a new tool. It is especially interesting to combine this method with nanotechnology and biotechnology, which have progressed rapidly during the 1990s. In this article, the principle and method of laser manipulation is described and then its applications and possibilities in nanotechnology and biotechnology are summarized.

Principle and Method

When light is reflected or refracted at the interface between two media with different refractive indices, the momentum of photons is changed, which leads to the generation of photon force as a reaction to the momentum change. For example, when a laser beam is reflected by a mirror, the momentum of photons is changed by Δp (Figure 1a). As a result, a photon force, $\mathbf{F}_{\text{phot}}$, acts on the mirror to deflect it vertically away from the reflected beam. In refraction, the momentum of photons is also changed, so that the photon force acts on the interface, as shown in Figure 1b. Thus light does not only give its energy to materials via absorption but also applies a mechanical force to them. However, when we are exposed to light, such as from a halogen lamp, we are never aware of the photon force because its magnitude is less than an order of pN. However, the photon force of the pN order acquired from a focused laser beam is sufficient to manipulate nm–μm-sized particles in solution, under an optical microscope.

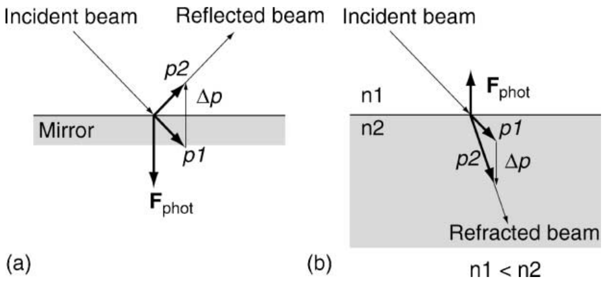

Figure 1 Optical pressure originating from the momentum change of photons.

Principle of Laser Trapping

If the size of a particle is larger than the wavelength of a trapping laser beam (μm-sized particles), the principle of single-beam gradient force optical trapping can be explained in terms of geometrical optics. When a tightly focused laser beam is irradiated onto a transparent dielectric particle, an incident laser beam is refracted at the interface between the particle and medium, as represented by Figure 2a. The propagation direction of the beam is changed, i.e., the momentum of the photon is changed, and consequently, a photon force is generated. As the laser beam leaves the particle and enters the surrounding medium, this refraction causes a photon force to be exerted again on that interface. Summing up the force contributions of all rays, if the refractive index of the particle is higher than that of the medium, a resultant force exerted on the particle is directed toward the focal point as an attractive force. However, reflection at the surface of the particle is negligible if that particle is transparent and the refractive index ratio of the particle to medium is close to unity. Therefore, the incident beam is reflected at two surfaces by a small amount, and as a result, the particle is directed slightly to the propagation direction of the incident light. Where the particle absorbs the incident beam, the photon force is also generated to push it in the propagation direction. The negligible effects of the

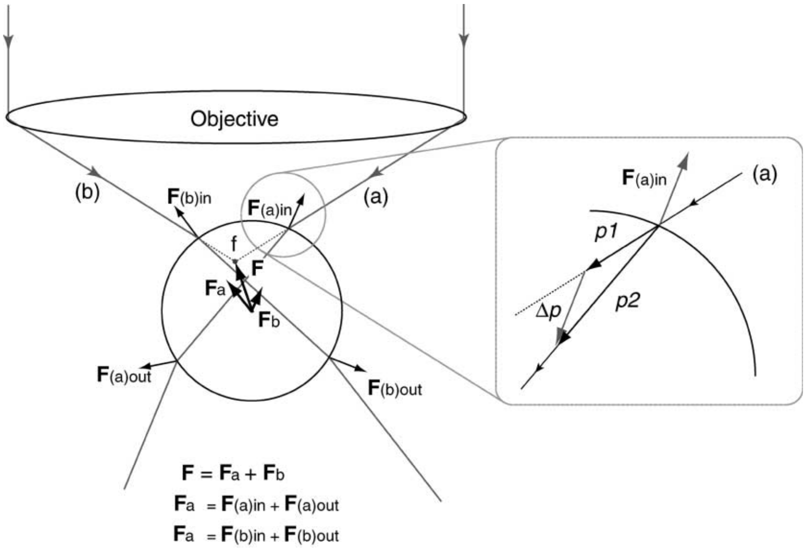

Figure 2 Principle of single-beam gradient force optical trapping explained by ray optics.

reflection and absorption occur in trapping experiment of transparent particles, such as polymer particles, silica microspheres, etc. However, particles with high reflectance and absorption coefficients such as metallic particles, exert a more dominant force at the surface where absorption and reflection occurs. This force is a repulsive force away from the focal point; consequently, metallic microparticles cannot be optically trapped.

If a dielectric particle is much smaller than the wavelength of a trapping light (nm-sized), it can be regarded as a point dipole (Rayleigh approximation) and the photon force ($\mathbf{F}_{\text{phot}}$) acted on it is given by

$$\mathbf{F}_{\text{phot}} = \mathbf{F}_{\text{grad}} + \mathbf{F}_{\text{scat}} \qquad [1]$$

Here, $\mathbf{F}_{\text{grad}}$ and $\mathbf{F}_{\text{scat}}$ are called the gradient force and scattering force, respectively. The scattering force is caused by the scattering of light, and it pushes the particle toward the direction of light propagation. On the other hand, the gradient force is generated when a particle is placed in a heterogeneous electric field of light. If the dielectric constant of the particle is higher than that of the surrounding medium, the gradient force acts on the particle to push it toward the higher intensity region of the beam. In the case of laser trapping of dielectric nanoparticles, the magnitude of gradient force is much larger than the scattering force. Consequently, the particle is trapped at the focal point of the trapping laser beam, where the beam intensity (electric field intensity) is maximum. The photon force can be expressed as follows:

$$\mathbf{F}_{\text{phot}} \cong \mathbf{F}_{\text{grad}} = \frac{1}{2}\varepsilon_{\text{m}}\alpha\nabla|\mathbf{E}|^2 \qquad [2]$$

$$\alpha = 3V\frac{\varepsilon_{\text{p}} - \varepsilon_{\text{m}}}{\varepsilon_{\text{p}} + 2\varepsilon_{\text{m}}} \qquad [3]$$

where, $\mathbf{E}$ is electric field of the light, V is volume of the particle, and ε_{p} and ε_{m} are dielectric constants of the particle and medium. In trapping nanoparticles that absorb the laser beam, such as gold nanoparticles, the complex dielectric constant and attenuation of the electric field in the particle need to be taken into consideration. Although the forces due to light absorption and scattering are both propelling the particle toward the direction of light propagation, the magnitude of gradient force is much larger than these forces. This is in contrast to the μm-sized metallic particle, which cannot be optically trapped.

Laser Manipulation System

An example of a laser micromanipulation system, with a pulsed laser to induce photoreactions, is shown in Figure 3. A linearly polarized laser beam from a CW Nd^{3+}:YAG laser is modulated to the circularly polarized light by a λ/4 plate and then split into horizontally and vertically polarized beams by a

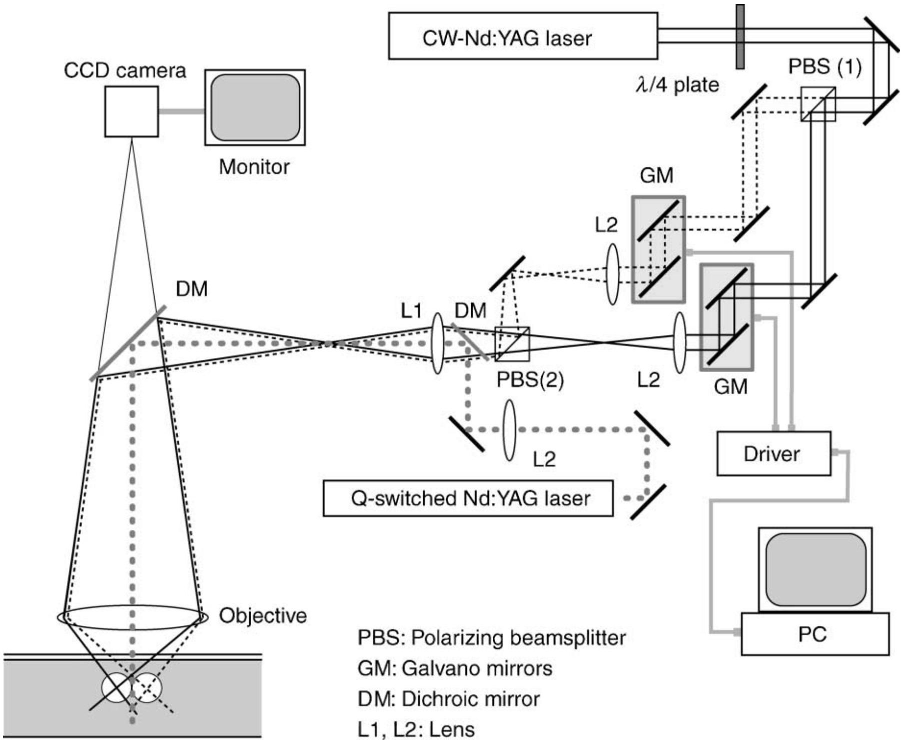

Figure 3 A block diagram of a dual beam laser manipulation-reaction system.

polarizing beamsplitter (PBS1). Both laser beams are combined by another (PBS2), introduced coaxially into an optical microscope, and then focused onto a sample through a microscope objective. The focal spots of both beams in the sample solution can be scanned independently by two sets of computer-controlled galvano mirrors. Even if the beams are spatially overlapping, interference does not take place because of their orthogonal polarization relation. Pulsed lasers, such as a Q-switched Nd^{3+}:YAG laser, are used to induce photoreactions, such as photopolymerization, photothermal reaction, laser ablation, etc.

Laser Manipulation and Patterning of Nanoparticles

Laser manipulation techniques enable us to capture and mobilize fine particles in solution. Most studies using this technique have been conducted on μm-sized objects such as polymer particles, microcrystals, living cells, etc. Because it is difficult to identify individual nanoparticles with an optical microscope, laser manipulation techniques have rarely been applied to nanotechnology and nanoscience. However, the laser trapping technique can be a powerful tool for manipulation of nanoparticles in solution where individual particles can be observed. This is achieved by detection of fluorescence emission from labeled dyes or scattered light. Now, even single molecule fluorescence spectroscopy has been achieved by the use of a highly sensitive photodetector, and a single metallic nanoparticle can be examined by detecting the light scattered from it. There have been several reports on the application of laser manipulation techniques for patterning of nm-sized particles. Here, fixation methods of nanoparticles, using the laser manipulation technique and local photoreactions, are introduced.

Patterning of Polymer Nanoparticles

Patterning of individual polymer nanoparticles onto a substrate can be achieved by using local photopolymerization. The following example shows the strength of this method. Polystyrene nanoparticles of 220 nm with fluorescent dye were dispersed in ethylene glycol solution containing polymerizable vinyl monomer (acrylamide, 31 wt%), crosslinker (*N*,*N*′-methylenebisacrylamide, 2.2 wt%), and radical photoinitiator (Irgacure2959, Ciba Specialty Chemicals, 1.1 wt%). When the sample was irradiated by blue light from a high-pressure mercury lamp, green fluorescence from dye molecules within each nanoparticle was observed. A nanoparticle that entered the region, irradiated by a near-infrared laser beam (1064 nm), was trapped at the focal point, and moved onto the surface of the glass substrate by handling the 3D stage of the microscope. An additional fixation laser beam (355 nm, 0.03 μJ, pulse duration ~6 ns, repetition rate ~5 Hz) was then focused to the same point for ~10 s, which led to the generation of acrylamide gel around the trapped nanoparticle. By repeating the procedure, patterning of single polymer nanoparticles on a glass substrate was achieved, and a fluorescence image of single nanoparticles as a letter ‘H’, is shown in Figure 4. A magnified atomic force microscope (AFM) image of one of the fixed nanoparticles is also shown, which confirms that only one polymer nanoparticle was contained in the polymerized gel.

Multiple polymer nanoparticles can also be gathered, patterned, and fixed on a glass substrate by scanning both trapping and fixation laser beams with the use of two pairs of galvano mirrors. The optical transmission and fluorescence images of the ‘H’ patterned nanoparticles on a glass substrate are shown in Figure 5a and b, respectively. The letter ‘H’ consists of three straight lines of patterned and fixed nanoparticles. The trapping laser beam (180 mW) was scanned at 30 Hz along each line with a length of ~10 μm on a glass substrate for 300 s. Nanoparticles were gathered and patterned along the locus of the focal point on the substrate. Then the fixation laser beam (0.097 μJ) was scanned for another 35 s. As a result, the straight lines of

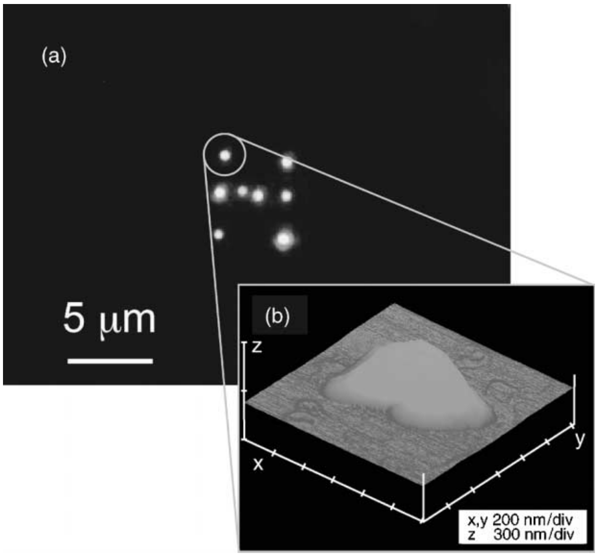

Figure 4 (a) A fluorescence image of spatially patterned individual polymer nanoparticles as the letter ‘H’ in distilled water. (b) A magnified AFM image of one of the produced acrylamide gels containing only one polymer nanoparticle on a glass substrate in the air.

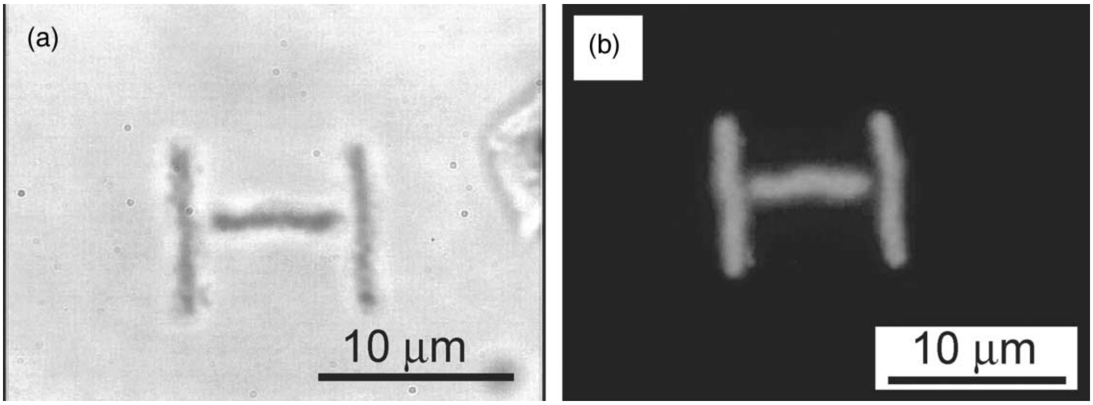

Figure 5 Optical transmission (a) and fluorescence (b) images of patterned and fixed polymer nanoparticles on the glass substrate, as the letter ‘H’.

patterned nanoparticles were fixed in generated acrylamide gel on the substrate. Combining several simple fixed patterns, a more complex arrangement can be created with use of the present manipulation and fixation techniques.

Fixation of Individual Gold Nanoparticles

Manipulation and fixation of single metallic nanoparticles in solution has been achieved by means of photo-induced transient melting of nanoparticles. As an example, the following method describes this technique. Gold nanoparticles (diameter ~80 nm) were dispersed in ethylene glycol. In order to identify a single gold nanoparticle, extinction of light from the halogen lamp of the optical microscope by the particle was utilized. Thus, the trapped particle was observed as a black spot in transmission image. The optically trapped single gold nanoparticle was then transferred to a precise position on the surface of glass substrate in the sample solution. The focused 355 nm pulse was additionally irradiated onto the pressed nanoparticle, which led to a transient temperature elevation of the gold nanoparticle to enable its fixation. It was confirmed by AFM observation that, at suitable laser fluence (32–64 mJ cm^{-2}), a gold nanoparticle was fixed on the glass substrate without fragmentation. Repeating the same manipulation and fixation procedure, single gold nanoparticles could be patterned on a glass substrate. Figure 6 shows an AFM image of successive spatial patterning of single 80 nm gold particles.

The significance of the laser manipulation-fixation technique is that we can trap, manipulate, and fix single and/or many nanoparticles in solution at room temperature. We have already demonstrated assembling entangled polymer chains of 10–20 nm mean radius by laser trapping, and formation of a single μm-sized particle. Thus, these various kinds of nanoscopic materials can be well manipulated using these techniques and it is believed that the present nanomanipulation-fixation method will be useful for future nanoscience and nanotechnology.

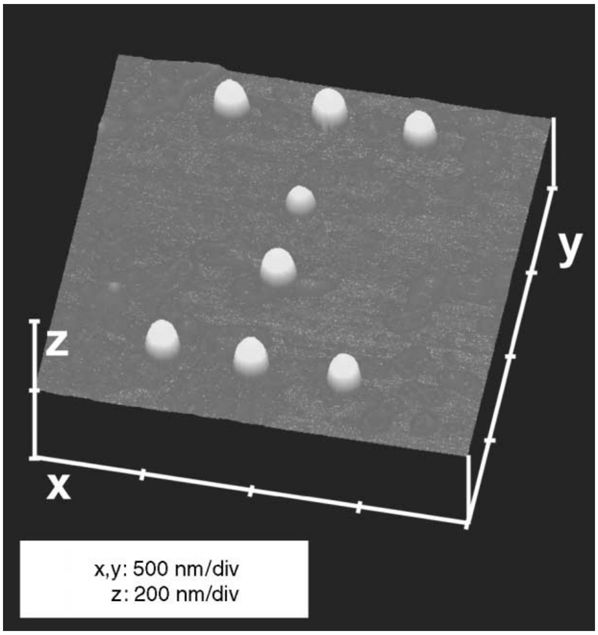

Figure 6 An AFM image of fixed and patterned gold nanoparticles, as the letter I.

Application to Biotechnology

Recent progress in biotechnology, using single cell manipulation by laser and microscope, has been attracting significant attention. In conventional methods, cell manipulation has been performed by mechanical manipulation using microneedles and micropipettes. However, laser manipulation can be applied as microtweezers. In comparison with the former, this form of laser manipulation has the

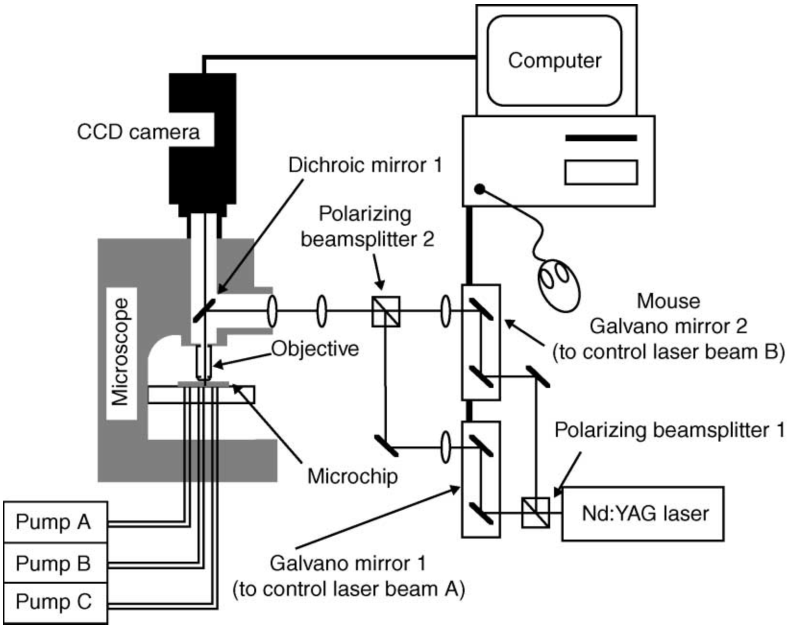
Computer
CCD camera
Dichroic mirror 1
Polarizing beamsplitter 2
Microscope
Objective
Mouse
Galvano mirror 2 (to control laser beam B)
Polarizing beamsplitter 1
Microchip
Nd:YAG laser
Pump A
Pump B
Pump C
Galvano mirror 1 (to control laser beam A)

Figure 7 An experimental setup of dual-beam laser manipulation and microfluidic devices.

advantage of no contact with cells and can perform some complex and characteristic movement, for example, rotation, separation, accumulation, etc. Furthermore, by combining it with a microcutter using laser ablation, cell fusion and extraction/injection of organs from/into the cells can be achieved. In this section, single-cell manipulation achieved by using the laser manipulation, is introduced.

Noncontact Rotation of Cells Using Dual Beam Laser Manipulation

As a special feature of laser-cell manipulation, it is noticeable that multiple laser tweezers can be operated independently without any contact. Here, noncontact rotation of the fission yeast, *Schizosaccharomyces pombe* (h^-), demonstrates this method. The yeast has the shape of an ellipse of length 8 and 3 μm for major and minor axes, and is shown set on a dual-beam laser manipulation system in Figure 7. One trapping beam A was focused at the end of the cell to anchor it, and another beam, B, was at the other end and scanned circularly around beam A by controlling the Galvano mirrors. The rotation of the yeast cell was realized at a frequency of less than 2 Hz, as shown in Figure 8. Such a cell manipulation is impossible by mechanical manipulation and indicates the superior performance of laser manipulation.

Transfer of Cells in Microchannel

A flow cytometry to physically separate and identify specific types of cells from heterogeneous populations by fluorescence, which is called fluorescence-activated cell sorter (FACS), has attracted significant attention as an important technique in biotechnology. In the separation process, charged single droplets containing single, fluorescence labeled cells are selected by passing them between two high-voltage deflection plates. Since it is a sequential process, which does not use microscopy, flexible and high-purity cell separation is limited. To overcome this problem, a selective cell separation system,

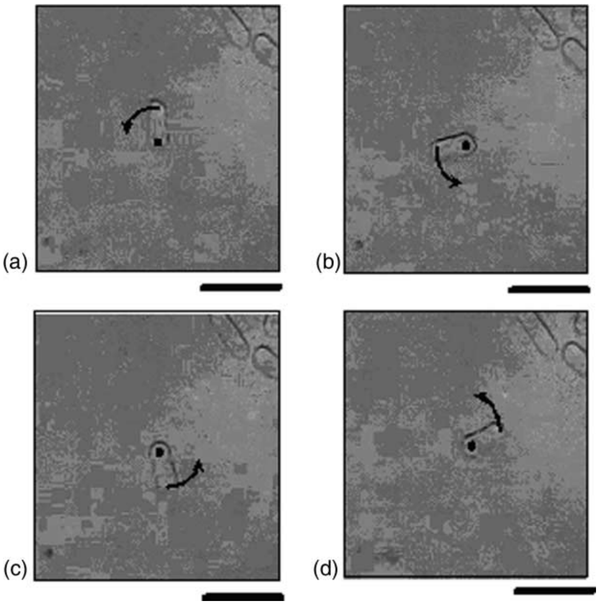

Figure 8 Rotation of fission yeast cells using dual-beam laser manipulation. The bars are 10 μm.

combining the laser trapping with the microchannel, is used. The method using microscopy is useful and flexible, because cells are not simply detected by fluorescence but also identified by their shape, size, surface morphology, absorption, etc. However, cell sorting under a microscope is not processed efficiently by laser trapping. A more effective method is described by the following.

The microchannel was prepared with laser microfabrication and photopolymerization, whose top view is shown in Figure 9a, and was then set on the microscope stage. Three syringes were equipped with homemade pumps A, B, and C, and connected with a microchip. There were two wing-like chambers in the microchip, that were connected with pumps A and C. Between chambers, there was a microchannel 100 μm wide, which joined these chambers and crossed a long microchannel connected with the pump B, thus forming a drain. Thickness of these chambers was 100 μm. By controlling these pumps, the left and right chambers were filled by the culture medium including the yeast cells and the neat culture medium, respectively, as shown in Figure 9b.

Individual cells were transferred from left to right chambers by using a single laser beam and handling an electrically movable microscope stage. The trapping laser irradiated a yeast cell in the left chamber and the trapped cell was transferred to the right chamber. A representative demonstration is given in Figure 9c. The position of the trapped yeast cell was controlled by the microscope stage with the velocity of 20 μm/s. The single yeast cells were successfully transferred from left to right chambers at the rate of 26 s. By combining laser trapping with the microchannel, single separation of cells was achieved under a microscope.

Collection and Alignment of Cells

When cell sorting using laser trapping is applied to biotechnology, the time to transfer cells should be short compared to the above-mentioned times. One method is to transfer multiple cells simultaneously by developing dual-beam laser manipulation, which is shown schematically in Figure 10. One trapping beam A was scanned on a line given as an arrow in Figure 10a at a rate faster than motions of cells.

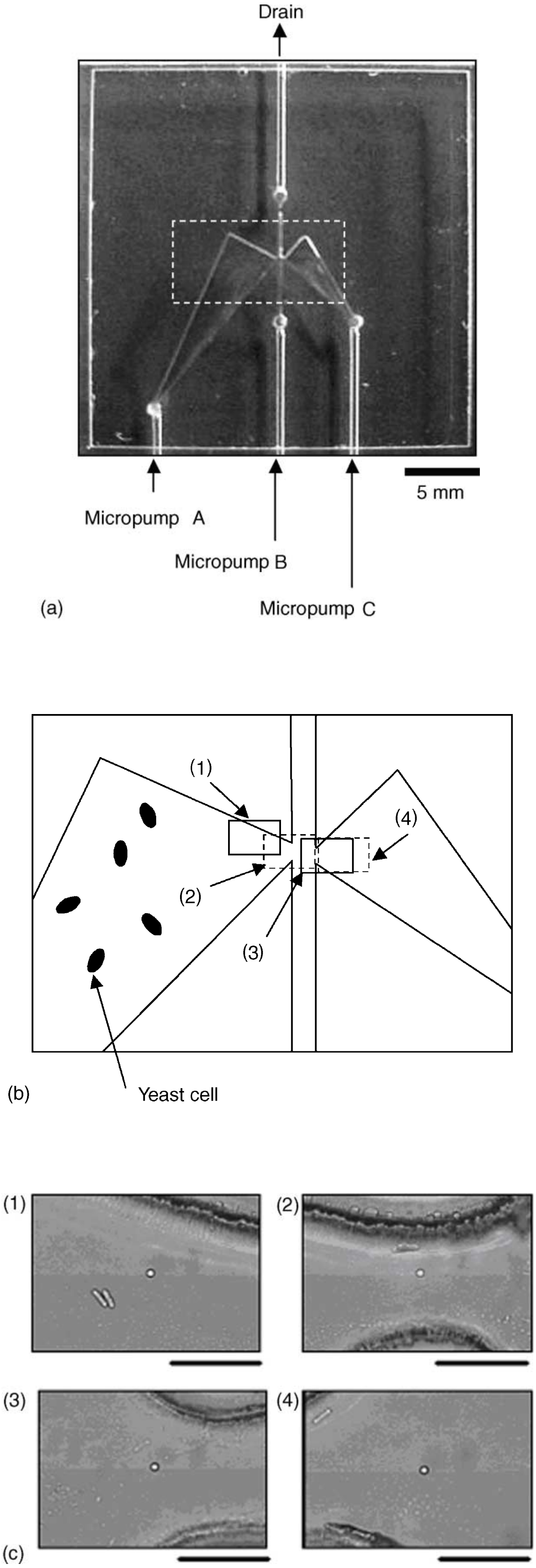

Figure 9 (a) A microchip fabricated by laser polymerization technique and (b) its schematic representation corresponding to dashed area in the microchip. (c) Cell transfer in microchip using laser trapping. The bars are 100 μm. A trapping laser beam is fixed at the center of each picture, by which a particle is trapped and transferred from left to right chambers. The position of each picture is shown in (b).

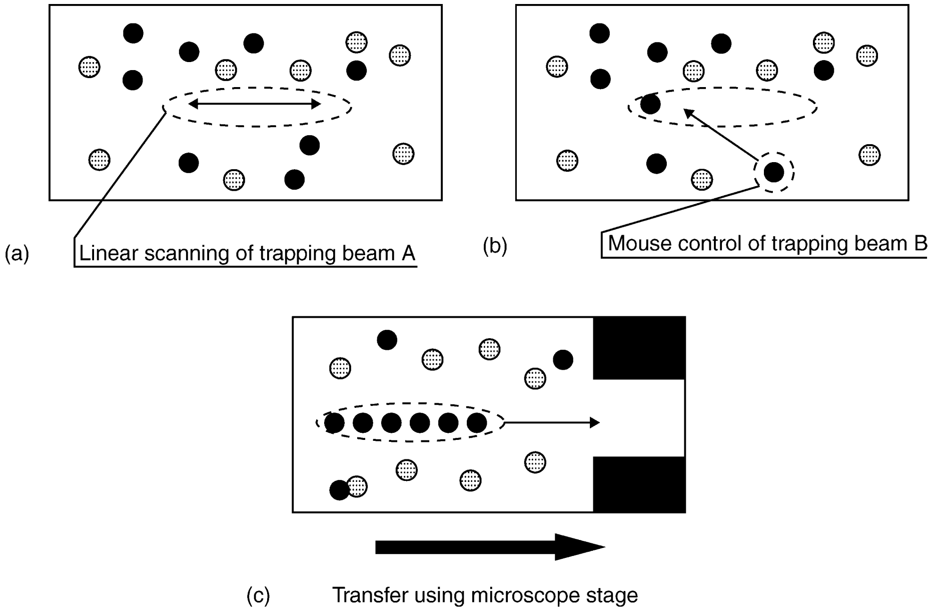
(a)
Linear scanning of trapping beam A
(b)
Mouse control of trapping beam B
(c)
Transfer using microscope stage

Figure 10 A schematic illustration of efficient collection and transfer of cells based by dual-beam laser manipulation.

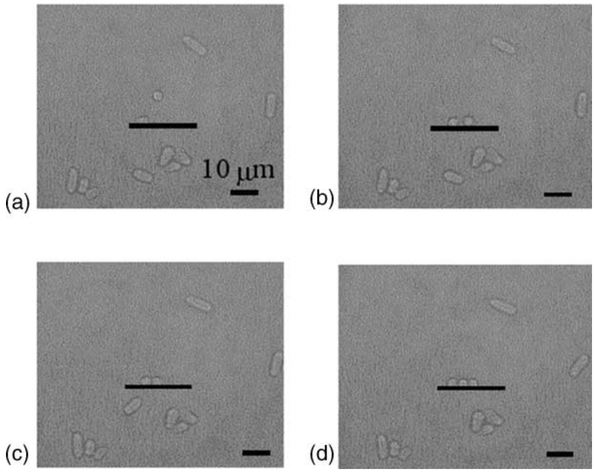

Figure 11 Collection and alignment of fission yeast cells using dual-beam laser manipulation.

The other trapping beam B was used to move individual cells to the line sequentially (Figure 10b). Finally, by moving the microscope stage, many particles trapped at the scanning line can be transferred to another chamber (Figure 10c).

An example is shown in Figure 11. The trapping beam A was scanned on a line with the rate of 15 Hz whose length was 30 μm. The trapping beam B was used to trap a yeast cell and move it to the line, which was controlled by a computer mouse pointer. The cells stored on the line were successfully isolated and with this method three cells can be collected within 15 s. Furthermore, the cells could be transferred with the velocity of 20 μm/s by driving the microscope stage. By improving the presented cell-sorting system, it is expected to realize flexible and high-purity cell separation.

List of Units and Nomenclature

Laser fluence (light energy density per pulse)	[mJ cm^{-2}]

See also

Time-Resolved Fluorescence: Measurements in Polymer Science.

Further Reading

Ashkin A, Dziedzic MJ, Bjorkholm JE and Chu S (1986) Observation of a single-beam gradient force optical trap for dielectric particles. *Optical Letters* 11: 288–290.

Ashkin A, Dziedzic MJ and Yamane T (1987) Optical trapping and manipulation of single cells using infrared laser beams. *Nature* 330: 769–771.

Berns WM, Aist J, Edwards J, *et al.* (1981) Laser microsurgery in cell and developmental biology. *Science* 213: 505–513.

Hoffmann F (1996) Laser microbeams for the manipulation of plant cells and subcellular structures. *Plant Science* 113: 1–11.

Hosokawa Y, Masuhara H, Matsumoto Y and Sato S (2002) Dual-beam laser micromanipulation for sorting biological cells and its device application. *Proceedings of SPIE* 4622: 138–142.

Ito S, Yoshikawa H and Masuhara H (2001) Optical patterning and photochemical fixation of polymer nanoparticles on glass substrates. *Applied Physics Letters* 78: 2566–2568.

Ito S, Yoshikawa H and Masuhara H (2002) Laser manipulation and fixation of single gold nanoparticles in solution at room temperature. *Applied Physics Letters* 80: 482–484.

Kamentsky AL, Melamed RM and Derman H (1965) Spectrophotometer: new instrument for ultrarapid cell analysis. *Science* 150: 630–631.

Kim BH, Kogi O and Kitamura N (1999) Single-microparticle measurements: laser trapping-absorption microscopy under solution-flow conditions. *Analytical Chemistry* 71: 4338–4343.

Masuhara H, De Schryver FC, Kitamura N and Tamai N (eds) (1994) *Microchemistry: Spectroscopy and Chemistry in Small Domains*. Amsterdam: Elsevier-Science.

Sasaki K, Koshioka M, Misawa H, Kitamura N and Masuhara H (1991) Laser-scanning micromanipulation and spatial patterning of fine particles. *Japanese Journal of Applied Physics* 30: L907–L909.

Schütze K, Clement-Segelwald A and Ashkin A (1994) Zona drilling and sperm insertion with combined laser microbeam and optical tweezers. *Fertility and Sterility* 61: 783–786.

Shikano S, Horio K, Ohtsuka Y and Eto Y (1999) Separation of a single cell by red-laser manipulation. *Applied Physics Letters* 75: 2671–2673.

Svoboda K and Block SM (1994) Optical trapping of metallic Rayleigh particles. *Optical Letters* 19: 930–932.

Nonlinear Spectroscopies

S R Meech, University of East Anglia, Norwich, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemistry is concerned with the induction and observation of changes in matter, where the changes are to be understood at the molecular level. Spectroscopy is the principal experimental tool for connecting the macroscopic world of matter with the microscopic world of the molecule, and is, therefore, of central importance in chemistry. Since its invention, the laser has greatly expanded the capabilities of the spectroscopist. In linear spectroscopy, the monochromaticity, coherence, high intensity, and high degree of polarization of laser radiation are ideally suited to high-resolution spectroscopic investigations of even the weakest transitions. The same properties allowed, for the first time, the investigation of the nonlinear optical response of a medium to intense radiation. Shortly after the foundations of nonlinear optics were laid, it became apparent that these nonlinear optical signals could be exploited in molecular spectroscopy, and since then a considerable number of nonlinear optical spectroscopies have been developed. This short article is not a comprehensive review of all these methods. Rather, it is a discussion of some of the key areas in the development of the subject, and indicates some current directions in this increasingly diverse area.

Nonlinear Optics for Spectroscopy

The foundations of nonlinear optics are described in detail elsewhere in the encyclopedia, and in some of the texts listed in the Further Reading section at the end of this article. The starting point is usually the nonlinearity in the polarization, P_i, induced in the sample when the applied electric field, E, is large:

$$P_i = \varepsilon_0\left[\chi_{ij}^{(1)}E_j + \frac{1}{2}\chi_{ijk}^{(2)}E_jE_k + \frac{1}{4}\chi_{ijkl}^{(3)}E_jE_kE_l + \cdots\right] \quad [1]$$

where ε_0 is the vacuum permittivity, $\chi^{(n)}$ is the nth order nonlinear susceptibility, the indices represent directions in space, and the implied summation over repeated indices convention is used. The signal field, resulting from the nonlinear polarization, is calculated by substituting it as the source polarization in Maxwell's equations and converting the resulting field to the observable, which is the optical intensity.

The nonlinear polarization itself arises from the anharmonic motion of electrons under the influence of the oscillating electric field of the radiation. Thus, there is a microscopic analog of eqn [1] for the induced molecular dipole moment, μ_i:

$$\mu_i = [\varepsilon_0\alpha_{ij}d_j + \varepsilon_0^{-2}\beta_{ijk}d_jd_k + \varepsilon_0^{-3}\gamma_{ijkl}d_jd_kd_l + \cdots] \quad [2]$$

in which α is the polarizability, β the first molecular hyperpolarizability, γ the second, etc. The power series is expressed in terms of the displacement field **d** rather than **E** to account for local field effects. This somewhat complicates the relationship between the molecular hyperpolarizabilities, for example, γ_{ijkl} and the corresponding macroscopic susceptibility, $\chi^{(3)}_{ijkl}$, but it is nevertheless generally true that a molecule exhibiting a high value of γ_{ijkl} will also yield a large third-order nonlinear polarization. This relationship between the macroscopic and microscopic parameters is the basis for an area of chemistry which has influenced nonlinear optics (rather than the inverse). The synthesis of molecules which have giant molecular hyperpolarizabilities has been an active area, because of their potential applications in lasers and electro-optics technology. Such molecules commonly exhibit a number of properties, including an extended linear π electron conjugation and a large dipole moment, which changes between the ground and excited electronic states. These properties are in accord with the predictions of theory and of quantum chemical calculations of molecular hyperpolarizabilities.

Nonlinear optical spectroscopy in the frequency domain is carried out by measuring the nonlinear signal intensity as a function of the frequency (or frequencies) of the incident radiation. Spectroscopic information is accessible because the molecular hyperpolarizability, and therefore the nonlinear susceptibility, exhibits resonances: the signal is enhanced when one or more of the incident or generated frequencies are resonant with the frequency of a molecular transition. The rich array of nonlinear optical spectroscopies arises in part from the fact that with more input fields there are more accessible resonances than is the case with linear spectroscopy. As an example we consider the third-order susceptibility, $\chi^{(3)}_{ijkl}E_jE_kE_l$, in the practically important case of two incident fields at the same frequency, ω_1, and a third at ω_2. The difference between the two frequencies is close to the frequency of a Raman active vibrational mode, Ω_{rg} (Figure 1). The resulting susceptibility can be calculated to have the form:

$$\chi^{(3)}_{ijkl}(-2\omega_1+\omega_2;\omega_1,\omega_1,-\omega_2)$$

$$=\frac{N\Delta\rho_{rg}\Omega_{rg}(\alpha^{R}_{ij}\alpha^{R}_{kl}+\alpha^{R}_{ik}\alpha^{R}_{jl})}{12\hbar[\Omega^2_{rg}-(\omega_1-\omega_2)^2+\Gamma^2-2i(\omega_1-\omega_2)\Gamma]} \quad [3]$$

in which the α^{R} are elements of the Raman susceptibility tensor, Γ is the homogeneous linewidth of the Raman transition, $\Delta\rho_{rg}$ a population difference, and N the number density. A diagram illustrating this process is shown in Figure 1, where the signal field is at the anti-Stokes Raman frequency $(2\omega_1-\omega_2)$. The spectroscopic method which employs this scheme is called coherent anti-Stokes Raman spectroscopy (CARS) and is one of the most widely applied nonlinear optical spectroscopies (see below). Clearly, from eqn [3] we can see that the signal will be enhanced when the difference frequency is resonant with the Raman transition frequency.

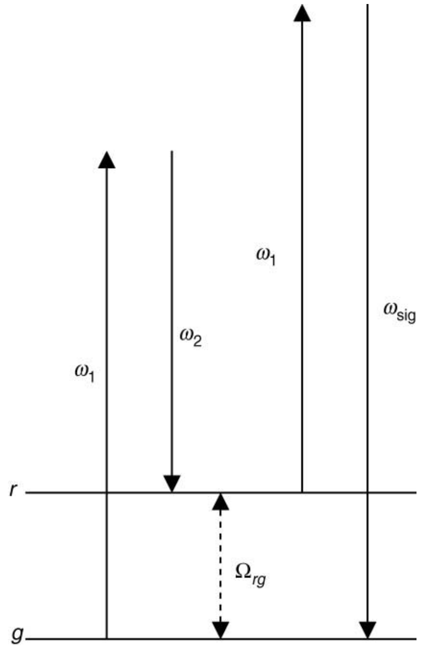

Figure 1 An illustration of the resonance enhancement of the CARS signal, $\omega_{sig}=2\omega_1-\omega_2$ when $\omega_1-\omega_2=\Omega_{rg}$.

With essentially the same experimental geometry there will also be nonlinear signals generated at both the Stokes frequency $(2\omega_2-\omega_1)$ and at the lower of the incident frequencies, ω_2. These signals have distinct phase matching conditions (see below), so they can easily be discriminated from one another, both spatially and energetically. Additional resonance enhancements are possible if either of the individual frequencies is resonant with an electronic transition of the molecule, in which case information on Raman active modes in the excited state is also accessible.

It is worthwhile noting here that there is an equivalent representation of the nth order nonlinear susceptibility tensor $\boldsymbol{\chi}^{(n)}(\omega_{sig}:\omega_1,\ldots,\omega_n)$ as a time domain response function, $\mathrm{R}^{(n)}(\tau_1,\ldots,\tau_n)$. While it is possible to freely transform between them, the frequency domain representation is the more commonly used. However, the response function approach is increasingly applied in time domain nonlinear optical spectroscopy when optical pulses shorter than the homogeneous relaxation time are used. In that case, the time ordering of the incident fields, as well as their frequencies, is of importance in defining an experiment.

An important property of many nonlinear optical spectroscopies is the directional nature of the signal, illustrated in Figure 2. The directional nonlinear

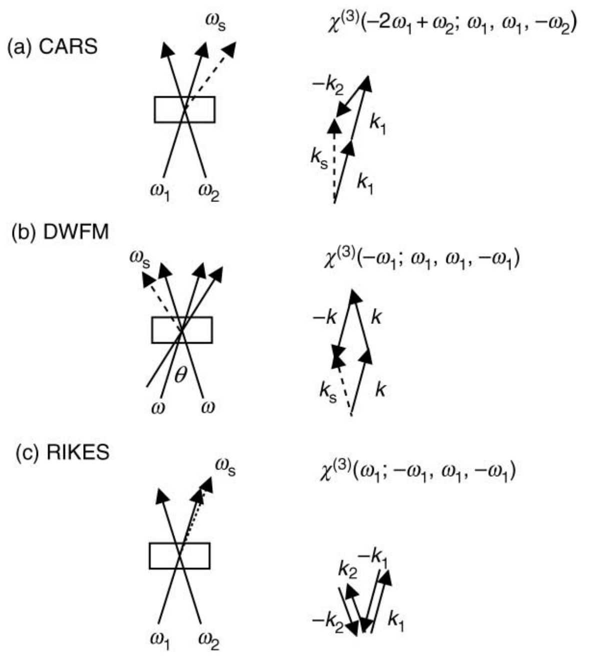
(a) CARS
ω_s
$\chi^{(3)}(-2\omega_1+\omega_2;\ \omega_1,\ \omega_1,\ -\omega_2)$
$-k_2$
k_1
k_s
k_1
ω_1
ω_2
(b) DWFM
ω_s
$\chi^{(3)}(-\omega_1;\ \omega_1,\ \omega_1,\ -\omega_1)$
$-k$
k
θ
k_s
k
ω
ω
(c) RIKES
ω_s
$\chi^{(3)}(\omega_1;\ -\omega_1,\ \omega_1,\ -\omega_1)$
k_2
$-k_1$
$-k_2$
k_1
ω_1
ω_2

Figure 2 Experimental geometries and corresponding phase matching diagrams for (a) CARS (b) DFWM (c) RIKES. Note that only the RIKES geometry is fully phase matched.

signal arises from the coherent oscillation of induced dipoles in the sample. Constructive interference can lead to a large enhancement of the signal strength. For this to occur the individual induced dipole moments must oscillate in phase – the signal must be phase matched. This requires the incident and the generated frequencies to travel in the sample with the same phase velocity, $k_\omega/\omega = c/n_\omega$ where n_ω is the index of refraction and k_ω is the wave propagation constant at frequency ω. For the simplest case of second-harmonic generation (SHG), in which the incident field oscillates at ω and the generated field at 2ω, phase matching requires $2k_\omega = k_{2\omega}$. The phase-matching condition for the most efficient generation of the second harmonic is when the phase mismatch, $\Delta k = 2k_\omega - k_{2\omega} = 0$, but this is not generally fulfilled due to the dispersion of the medium, $n_\omega < n_{2\omega}$. In the case of SHG, a coherence length L can be defined as the distance traveled before the two waves are 180° out of phase, $L = |\pi/\Delta k|$. For cases in which the input frequencies also have different directions, as is often the case when laser beams at two frequencies are combined (e.g., CARS), the phase-matching condition must be expressed as a vectorial relationship, hence, for CARS, $\Delta\mathbf{k} = \mathbf{k}_{AS} - 2\mathbf{k}_1 + \mathbf{k}_2 \approx 0$, where $\mathbf{k}_{AS}$ is the wavevector of the signal at the anti-Stokes frequency (Figure 2). Thus for known input wavevectors one can easily calculate the expected direction of the output signal, $\mathbf{k}_s$. This is illustrated for a number of important cases in Figure 2.

Nonlinear Optical Spectroscopy in Chemistry

As already noted, there is a vast array of nonlinear optical spectroscopies, so it is clear some means of classification will be required. For coarse graining the order of the nonlinear process is very helpful, and that is the scheme we will follow here.

Second-Order Spectroscopies

Inspection of eqn [1] shows that in gases, isotropic liquids, and solids where the symmetry point group contains a center of inversion, $\boldsymbol{\chi}^{(2)}$ is necessarily zero. This is required to satisfy the symmetry requirement that polarization must change sign when the direction of the field is inverted, yet for a quadratic, or any even order dependence on the field strength, it must remain positive. Thus second-order nonlinearities might not appear very promising for spectroscopy. However, there are two cases in which second-order nonlinear optical phenomena are of very great significance in molecular spectroscopy, harmonic generation and the spectroscopy of interfaces.

Harmonic conversion

Almost every laser spectroscopist will have made use of second-harmonic or sum frequency generation for frequency conversion of laser radiation. Insertion of an oscillating field of frequency ω into eqn [1] yields a second-order polarization oscillating of 2ω. If two different frequencies are input, the second-order polarization oscillates at their sum and difference frequencies. In either case, the second-order polarization acts as a source for the second-harmonic (or sum, or difference frequency) emission, provided $\boldsymbol{\chi}^{(2)}$ is nonzero. The latter can be arranged by selecting a noncentrosymmetric medium for the interaction, the growth of such media being an important area of materials science. Optically robust and transparent materials with large values of $\boldsymbol{\chi}^{(2)}$ are available for the generation of wavelengths shorter than 200 nm to longer than 5 μm. Since such media are birefringent by design, a judicious choice of angle and orientation of the crystal with respect to the input beams allows a degree of control over the refractive indices experienced by each beam. Under the correct phase matching conditions $n_\omega \approx n_{2\omega}$, and very long interaction lengths result, so the efficiency of signal generation is high.

Higher-order harmonic generation in gases is an area of growing importance for spectroscopy in the deep UV and X-ray region. The generation of such short wavelengths depends on the ability of amplified ultrafast solid state lasers to generate extremely high instantaneous intensities. The mechanism is somewhat different to the one outlined above. The intense pulse is injected into a capillary containing an appropriate gas. The high electric field of the laser causes ionization of atoms. The electrons generated begin to oscillate in the applied laser field. The driven recombination of the electrons with the atoms results in the generation of the high harmonic emission. Although the mechanism differs from the SHG case, questions of phase matching are still important. By containing the gas in a corrugated waveguide phase matching is achieved, considerably enhancing the intensity of the high harmonic. Photon energies of hundreds of electronvolts are possible using this technique. The generation of such high energies is not yet routine, but a number of potential applications are already apparent. A powerful coherent source of X-ray and vacuum UV pulses will certainly aid surface analysis techniques such as UV photoemission and X-ray photoelectron spectroscopy. Much excitement is currently being generated by the possibility of using intense ultrashort X-ray pulses to record structural dynamics on an ultrafast timescale.

Surface second-order spectroscopy

At an interface inversion symmetry is absent by definition, so the second-order nonlinear susceptibility is finite. If the two bulk phase media are themselves isotropic, then even a weak second-order signal necessarily arises from the interface. This surface-specific all-optical signal is unique, because it can be used to probe the interface between two condensed phases. This represents a great advantage over every other form of surface spectroscopy. In linear optical spectroscopy, the signal due to species at the interface are usually swamped by contributions from the bulk phase. Other surface-specific signals do exist, but they rely on the scattering of heavy particles (electrons, atoms) and so can only be applied to the solid vacuum interface. For this reason the techniques of surface SHG and SFG are widely applied in interface spectroscopy.

The most widely used method is sum frequency generation (SFG) between temporally overlapped tuneable infrared and fixed frequency visible lasers, to yield a sum frequency signal in the visible region of the spectrum. The principle of the method is shown in Figure 3. The surface nonlinear susceptibility exhibits resonances at vibrational frequencies, which are detected as enhancements in the visible SFG intensity. Although the signal is weak, it is directional and background free, so relatively easily measured by photon counting techniques. Thus, SFG is used to measure the vibrational spectra of essentially any optically accessible interface. There are, however, some limitations on the method. The surface must exhibit a degree of order – if the distribution of adsorbate orientation is isotropic the signal again becomes zero by symmetry. Second, a significant enhancement at the vibrational frequency requires the transition to be both IR and Raman allowed, as suggested by the energy level diagram (Figure 3).

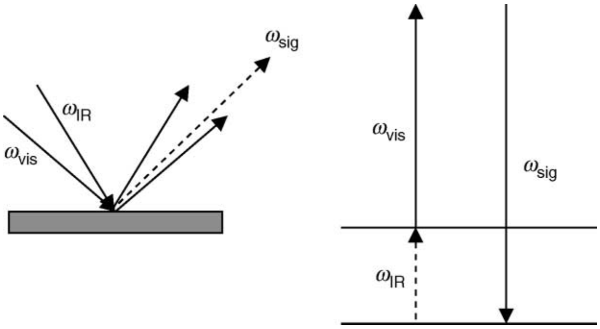

Figure 3 The experimental geometry for SFG, and an illustration of the resonance enhancement of $\omega_{sig} = \omega_{IR} + \omega_{vis}$ at a Raman and IR allowed vibrational transition.

The SHG signal can also be measured as a function of the frequency of the incident laser, to recover the electronic spectrum of the interface. This method has been used, particularly in the study of semiconductor surfaces, but generally the electronic spectra of adsorbates contain less information than their vibrational spectra. However, by measuring the SHG intensity as a function of time, information on adsorbate kinetics is obtained, provided some assumptions connecting the surface susceptibility to the molecular hyperpolarizability are made. Finally, using similar assumptions, it is possible to extract the orientational distribution of the adsorbate, by measuring the SHG intensity as a function of polarization of the input and output beams. For these reasons, SHG has been widely applied to analyze the structure and dynamics of interfaces.

Third-Order Spectroscopies

The third-order coherent Raman spectroscopies were introduced above. One great advantage of these methods over conventional Raman is that the signal is generated in a coherent beam, according to the appropriate phase matching relationship (Figure 2). Thus, the coherent Raman signal can easily be distinguished from background radiation by spatial filtering. This has led to CARS finding widespread application in measuring spectra in (experimentally) hostile environments. CARS spectroscopy has been widely applied to record the vibrational spectra of flames. Such measurements would obviously be very challenging for linear Raman or IR, due to the strong emission from the flame itself. The directional CARS signal in contrast can be spatially filtered, minimizing this problem. CARS has been used both to identify unstable species formed in flames, and to probe the temperature of the flame (e.g., from measured population differences eqn [3]). A second advantage of the technique is that the signal is only generated when the two input beams overlap in space. Thus, small volumes of the sample can be probed. By moving the overlap position around in the sample, spatially resolved information is recovered. Thus, it is possible to map the population of a particular transient species in a flame.

CARS is probably the most widely used of the coherent Raman methods, but it does have some disadvantages, particularly in solution phase studies. In that case the resonant $\boldsymbol{\chi}^{(3)}$ signal (eqn [3]) is accompanied by a nonresonant third-order background. The interference between these two components may result in unusual and difficult to interpret lineshapes. In this case, some other coherent Raman methods are more useful. The phase matching scheme for Raman-induced Kerr effect spectroscopy (RIKES) was shown in Figure 2. The RIKES signal is always phase matched, which leads to a long interaction length. However, the signal is at ω_2 and in the direction of ω_2, which would appear to be a severe disadvantage in terms of detection. Fortunately, if the input polarizations are correctly chosen the signal can be isolated by its polarization. In an important geometry, the signal (ω_2) is isolated by transmission through a polarizer oriented at 45° to a linearly polarized pump (ω_1). The pump is overlapped in the sample with the probe (ω_2) linearly polarized at $-45°$. Thus the probe is blocked by the polarizer, but the signal is transmitted. This geometry may be viewed as pump-induced polarization of the isotropic medium to render it birefringent, thus inducing ellipticity in the transmitted probe, such that the signal leaks through the analyzing polarizer (hence the alternative name optical Kerr effect).

In this geometry, it is possible to greatly enhance signal-to-noise ratios by exploiting interference between the probe beam and the signal. Placing a quarterwave plate in the probe beam with its fast axis aligned with the probe polarization, and reorienting it slightly ($<1°$) yields a slightly elliptically polarized probe before the sample. A fraction of the probe beam, the local oscillator (LO), then also leaks through the analyzing polarizer, temporally and spatially overlapped with the signal. Thus, the signal and LO fields are seen by the detector, which measures the intensity as:

$$\begin{aligned} I(t) &= \frac{nc}{8\pi}|\mathbf{E}_{\mathrm{LO}}(t) + \mathbf{E}_{\mathrm{S}}(t)|^2 \\ &= I_{\mathrm{LO}}(t) + I_{\mathrm{S}}(t) + \frac{nc}{4\pi}\mathrm{Re}[\mathbf{E}_{\mathrm{S}}^*(t)\cdot\mathbf{E}_{\mathrm{LO}}(t)] \end{aligned} \qquad [4]$$

where the final term may be very much larger than the original signal, and is linear in $\boldsymbol{\chi}^{(3)}$. This term is usually isolated from the strong I_{LO} by lock-in detection. This method is called optical heterodyne detection (OHD), and generally leads to excellent signal to noise. It can be employed with other coherent signals by artificially adding the LO to the signal, provided great care is taken to ensure a fixed phase relationship between LO and signal. In the RIKES experiment, however, the phase relationship is automatic. The arrangement described yields an out-of-phase LO, and measures the real part of $\boldsymbol{\chi}^{(3)}$, the birefringence. Alternatively, the quarterwave plate is omitted, and the analyzing polarizer is slightly reoriented, to introduce an in-phase LO, which measures the imaginary part of $\boldsymbol{\chi}^{(3)}$, the dichroism. This is particularly useful for absorbing media. The OHD-RIKES method has been applied to measure the spectroscopy of the condensed phase, and has found particularly widespread application in transient studies (below).

Degenerate four wave mixing (DFWM) spectroscopy is a simple and widely used third-order spectroscopic method. As the name implies, only a single frequency is required. The laser beam is split into three, and recombined in the sample, in the geometry shown in Figure 2. The technique is also known as laser-induced grating scattering. The first two beams can be thought of as interfering in the sample to write a spatial grating, with fringe spacing dependent on the angle between them. The signal is then scattered from the third beam in the direction expected for diffraction from that grating. The DFWM experiment has been used to measure electronic spectra in hostile environments, by exploiting resonances with electronic transitions. It has also been popular in the determination of the numerical value of $\boldsymbol{\chi}^{(3)}$, partly because it is an economical technique, requiring only a single laser, but also because different polarization combinations make it possible to access different elements of $\boldsymbol{\chi}^{(3)}$. The technique has also been used in time resolved experiments, where the decay of the grating is monitored by diffraction intensity from a time delayed third pulse.

Two-photon or, more generally, multiphoton excitation has applications in both fundamental spectroscopy and analytical chemistry. Two relevant level

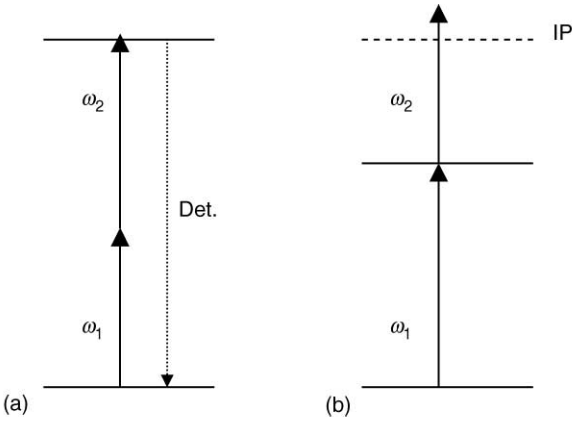

Figure 4 Two cases of resonant two photon absorption. In (a) the excited state is two-photon resonant, and the process is detected by the emission of a photon. In (b) the intermediate state is resonant, and the final energy is above the ionization potential (IP) so that photocurrent or mass detection can be used.

schemes are shown in Figure 4. Some property associated with the final state permits detection of the multiphoton absorption, for example, fluorescence in (a) and photocurrent in (b).

Excitation of two-photon transitions, as in Figure 4a, is useful in spectroscopy because the selection rules are different to those for the corresponding one-photon transition. For example the change in angular momentum quantum number, ΔL, in a two-photon transition is 0, ± 2, so, for example, an atomic S to D transition can be observed. High spectroscopic resolution may be attained using Doppler free two-photon absorption spectroscopy. In this method, the excitation beams are arranged to be counter-propagating, so that the Doppler broadening is cancelled out in transitions where the two excitation photons arise from beams with opposing wavevectors. In this case, the spectroscopic linewidth is governed only by the homogeneous dephasing time.

The level scheme in Figure 4b is also widely used in spectroscopy, but in this case the spectrum of the intermediate state is obtained by monitoring the photocurrent as a function of ω_1. The general technique is known as resonance enhanced multiphoton ionization (REMPI) and yields high-quality spectra of intermediate states which are not detectable by standard methods, such as fluorescence. The sensitivity of the method is high, and it is the basis of a number of analytical applications, often in combination with mass spectrometry.

Ultrafast Time Resolved Spectroscopy

The frequency and linewidth of a Raman transition may be extracted from the CARS measurement, typically by combining two narrow bandwidth pulsed lasers, and tuning one through the resonance while measuring the nonlinear signal intensity. The time resolved analogue requires two pulses, typically of a few picoseconds duration (and therefore a few wavenumbers bandwidth) at ω_1 and ω_2 to be incident on the sample. This pair coherently excites the Raman mode. A third pulse at ω_1 is incident on the sample a time t later, and stimulates the CARS signal at $2\omega_1 - \omega_2$ in the phase-matched direction. The decay rate of the vibrational coherence is measured from the CARS intensity as a function of the delay time. Thus, the frequency of the mode is measured in the frequency domain, but the linewidth is measured in the time domain. If very short pulses are used (such that the pulsewidth is shorter than the inverse frequency of the Raman active mode) the Raman transition is said to be impulsively excited, and the CARS signal scattered by the time delayed pulse reveals an oscillatory response at the frequency of the Raman active mode, superimposed on its decay. Thus, in this case, spectroscopic information is measured exclusively in the time domain. In the case of nonlinear Raman spectroscopy, similar information is available from the frequency and the time domain measurements, and the choice between them is essentially one of experimental convenience. For example, time domain CARS, RIKES, and DFWM spectroscopy have turned out to be particularly powerful routes to extracting low-frequency vibrational and orientational modes of liquids and solutions, thus providing detailed insights into molecular interactions and reaction dynamics in the condensed phase.

Other time domain experiments contain information that is not accessible in the frequency domain. This is particularly true of photon echo methods. The name suggests a close analogy with nuclear magnetic resonance (NMR) spectroscopy, and the (optical) Bloch vector approach may be used to describe both measurements, although the transition frequencies and time-scales involved differ by many orders of magnitude. In the photon echo experiment, two or three ultrafast pulses with carefully controlled interpulse delay times are resonant with an electronic transition of the solute. In the two-pulse echo, the echo signal is emitted in the phase match direction at twice the interpulse delay, and its intensity as a function of time yields the homogeneous dephasing time associated with the transition. In the three-pulse experiment the pulses are separated by two time delays. By measuring the intensity of the stimulated echo as a function of both delay times it is possible to separately determine the dephasing time and the population relaxation time associated with the resonant transition. Such information is not accessible from linear spectroscopy, and can be extracted only with difficulty in the frequency domain. The understanding of photon echo spectroscopy has expanded well beyond the simple description given here, and it now provides unprecedented insights into optical dynamics in solution, and thus informs greatly our understanding of chemistry in the condensed phase. The methods have recently been extended to the infra red, to study vibrational transitions.

Higher-Order and Multidimensional Spectroscopies

The characteristic feature of this family of spectroscopies is the excitation of multiple resonances, which may or may not require measurements at $\chi^{(n)}$ with $n > 3$. Such experiments require multiple frequencies, and may yield weak signals, so they only became experimentally viable upon the availability of stable and reliable solid state lasers and optical parametric generators. Measurements are made in either the time or the frequency domain, but in either case benefit from heterodyne detection.

One of the earliest examples was two-dimensional Raman spectroscopy, where multiple Raman active modes are successively excited by temporally delayed pulse pairs, to yield a fifth-order nonlinear signal. The signal intensity measured as a function of both delay times (corresponding to the two dimensions) allows separation of homogeneous and inhomogeneous contributions to the line shape. This prodigiously difficult $\chi^{(5)}$ experiment has been completed in a few cases, but is plagued by interference from third-order signals.

More widely applicable are multidimensional spectroscopies using infrared pulses or combinations of them with visible pulses. The level scheme for one such experiment is shown in Figure 5 (which is one of many possibilities). From the scheme, one can see that the nonlinear signal in the visible depends on two resonances, so both can be detected. This can be regarded as a multiply resonant nondegenerate four-wave mixing (FWM) experiment. In addition, if the two resonant transitions are coupled, optical excitation of one affects the other. Thus, by measuring the signal as a function of both frequencies, the couplings between transitions are observed. These appear as cross peaks when the intensity is plotted in the two frequency dimensions, very much as with 2D NMR. This technique is already providing novel information on molecular structure and structural dynamics in liquids, solutions, and proteins.

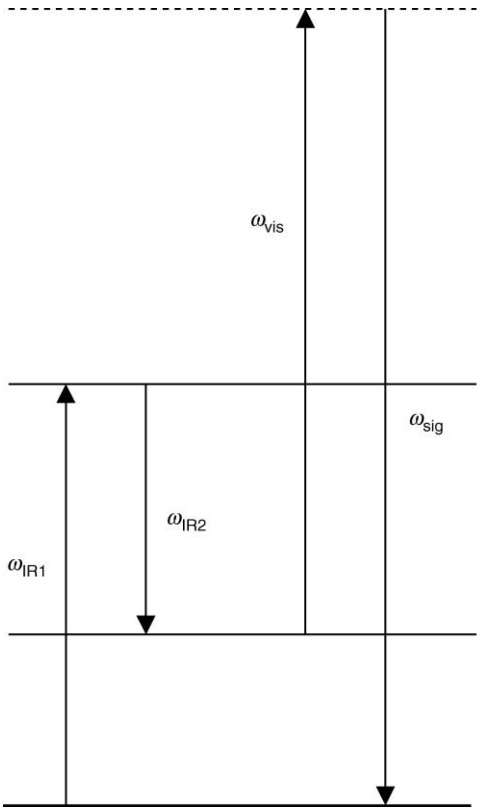

Figure 5 Illustration of multiple resonance enhancements in a FWM geometry, from which 2D spectra may be generated.

Spatially Resolved Spectroscopy

A recent innovation is nonlinear optical microscopy. The nonlinear dependence of signal strength on intensity means that nonlinear processes are localized at the focal point of a lens. When focusing is strong, such as in a microscope objective, spatial localization of the nonlinear signal can be dramatic. This is the basis of the two-photon fluorescence microscopy method, where a high repetition rate source of low-energy ultrafast pulses is focused by a microscope objective into a sample labeled with a fluorescent molecule, which has absorption at half the wavelength of the incident photons (Figure 4). The fluorescence is necessarily localized at the focal point because of its dependence on the square of the incident intensity. By measuring intensity while scanning the position of the focal point in space, a 3D image of the distribution of the fluorophore is constructed. This technique turns out to have a number of advantages over one photon fluorescence microscopy, most notably in terms of ease of implementation, minimization of sample damage, and depth resolution. The technique is widely employed in cell biology.

Stimulated by the success of two-photon microscopy, further nonlinear microscopies have been developed, all relying on the spatial localization of the signal. CARS microscopy has been demonstrated to yield 3D images of the distribution of vibrations in living cells. It would be difficult to recover such data by linear optical microscopy. SHG has been applied in microscopy. In this case, by virtue of the symmetry selection rule referred to above, a 3D image of orientational order is recovered. Both these and other nonlinear signals provide important new information on complex heterogeneous samples, most especially living cells.

See also

Spectroscopy: Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Andrews DL and Allcock P (2002) *Optical Harmonics in Molecular Systems*. Weinheim, Germany: Wiley-VCH.

Andrews DL and Demidov AA (eds) (2002) *An Introduction to Laser Spectroscopy*. New York: Kluwer Academic.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Cheng JX and Xie XS (2004) Coherent anti-Stokes Raman scattering microscopy: Instrumentation, theory, and applications. *Journal of Physical Chemistry B* 108: 827.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99.

Eisenthal KB (1992) Equilibrium and dynamic processes at interfaces by 2nd harmonic and sum frequency generation. *Annual Review of Physical Chemistry* 43: 627.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Oxford University Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109.

Fourkas JT (2002) Higher-order optical correlation spectroscopy in liquids. *Annual Review of Physical Chemistry* 53: 17.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of Chemistry Society Faraday Transactions* 90: 1.

Heinz TF (1991) Second order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353.

Hesselink WH and Wiersma DA (1983) Theory and experimental aspects of photon echoes in molecular solids. In: Hochstrasser RM and Agranovich VM (eds) *Spectroscopy and Excitation Dynamics of Condensed Molecular Systems*, chap. 6. Amsterdam: North Holland.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. San Diego, CA: Academic Press.

Meech SR (1993) Kinetic application of surface nonlinear optical signals. In: Lin SH, Fujimura Y, and Villaeys A (eds) *Advances in Multiphoton Processes and Spectroscopy*, vol. 8, pp. 281. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rector KD and Fayer MD (1998) Vibrational echoes: a new approach to condensed-matter vibrational spectroscopy. *International Review of Physical Chemistry* 17: 261.

Richmond GL (2001) Structure and bonding of molecules at aqueous surfaces. *Annual Review of Physical Chemistry* 52: 357.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Smith NA and Meech SR (2002) Optically heterodyne detected optical Kerr effect – Applications in condensed phase dynamics. *International Review of Physical Chemistry* 21: 75.

Tolles WM, Nibler JW, McDonald JR and Harvey AB (1977) Review of theory and application of coherent anti-Stokes Raman spectroscopy (CARS). *Applied Spectroscopy* 31: 253.

Wright JC (2002) Coherent multidimensional vibrational spectroscopy. *International Review of Physical Chemistry* 21: 185.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1368.

Zyss J (ed.) (1994) *Molecular Nonlinear Optics*. San Diego: Academic Press.

Photodynamic Therapy of Cancer

A J MacRobert and T Theodossiou, University College London, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

One of the most active areas of photomedical research, in recent years, has been the exploration of the use of light-activated drugs known as photosensitizers. These compounds may be activated using light, usually provided by a laser via an optical fiber which is placed at the site of the target lesion. This treatment is known as photodynamic therapy (PDT) and is being applied to the local destruction of malignant tumors and certain nonmalignancies. Activation of the photosensitizer results in the generation of reactive oxidizing intermediates which are toxic to cells, and this process ultimately leads to tumor destruction. PDT is a relatively low-power, nonthermal, photochemical technique that uses fluence rates not exceeding 200 mW/cm^2 and total light doses or fluences of typically 100 J/cm^2. Generally red or near-infrared light is used, since tissue is relatively transparent at these wavelengths. It is a promising alternative approach to the local destruction of tumors for several reasons. First, it is a minimally invasive treatment since laser light can be delivered with great accuracy to almost any site in the body via thin flexible optical fibers with minimal damage to overlying normal tissues. Second, the nature of PDT damage to tissues is such that healing is safer and more complete than after most other forms of local tissue destruction (e.g., radiotherapy). PDT is also capable of some degree of selectivity for tumors when the sensitizer levels, light doses, and irradiation geometry are carefully controlled. This selectivity is based on the higher sensitizer retention in tumors after administration relative to the adjacent normal tissues in which the tumor arose (generally 3:1 for extracranial tumors but up to 50:1 for brain tumors).

The photosensitizer is administered intravenously to the patient and time allowed (3–96 hours depending on the sensitizer) for it to equilibrate in the body before the light treatment (Figure 1). This time is called the drug–light interval. PDT may also be useful for treating certain nonmalignant conditions, in particular, psoriasis and dysfunctional menorrhagia (a disorder of the uterus), and local treatment of infections; such as genital papillomas, infections in the mouth and the upper gastrointestinal tract. In certain cases, the photosensitizer may be applied directly to the lesions, particularly for treatment of skin tumors, as discussed later. The main side-effect of PDT is skin photosensitivity, owing to retention of the drug in the skin, so patients must avoid exposure to sunlight in particular for a short period following treatment. Retreatment is then possible once the photosensitizer has cleared from the skin since these drugs have little intrinsic toxicity, unlike many conventional chemotherapy agents.

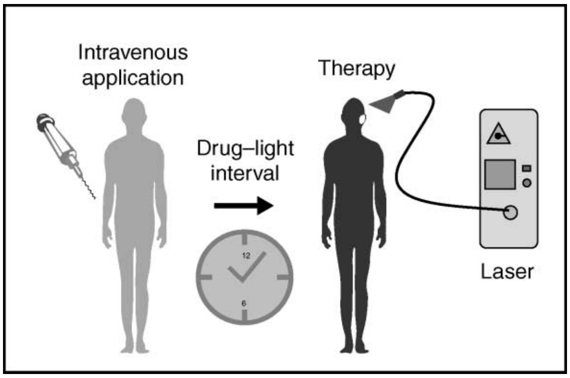

Figure 1 Photodynamic therapy: from sensitization to treatment.

Photoproperties of Photosensitizers

By definition, there are three fundamental requirements for obtaining a photodynamic effect: (a) light of the appropriate wavelength matched to the photosensitizer absorption; (b) a photosensitizer; and (c) molecular oxygen. The ideal photochemical and biological properties of a photosensitizer may be easily summarized, although the assessment of a sensitizer in these terms is not as straightforward as might be supposed because the heterogeneous nature of biological systems can sometimes profoundly affect these properties. Ideally though, a sensitizer should possess the following attributes: (a) red or near infrared light absorption; (b) nontoxic, and with low skin photosensitizing potency; (c) selective retention in tumors relative to normal adjacent tissue; (d) an efficient generator of cytotoxic species, usually singlet oxygen; (e) fluorescence, for visualization; (f) a defined chemical composition, and (g) preferably water soluble. A list of several photosensitizers possessing the majority of the above-mentioned attributes is given in Table 1.

The reasons for these requirements are partly self-evident, but worth amplifying. Strong absorption is desirable in the red and near-infrared spectral region, where tissue transmittance is optimum enabling penetration of the therapeutic light within the tumor (Figure 2). To minimize skin photosensitization by solar radiation, the sensitizer absorption spectrum should ideally consist of a narrow red wavelength band, with little absorption at other wavelengths down to 400 nm, below which solar irradiation falls off steeply. Another advantage of red wavelength irradiation is that the potential mutagenic effects encountered with UV-excited sensitisers (e.g., psoralens) are avoided. Since the object of the treatment is the selective destruction of tumor tissue, leaving surrounding normal tissue undamaged, some degree of selective retention of the dye in tumor tissue is desirable. Unfortunately, the significance of this aspect has been exaggerated in the literature and in many cases treatment selectivity owes more to careful light irradiation geometry. Nevertheless, many normal tissues have the capacity to heal safely following PDT damage.

The key photochemical property of photosensitizers is to mediate production of some active

Table 1 Properties of several photosensitizers

Compound	*λ/nm (ε/M^{-1} cm^{-1})*	*Drug dose/mg kg^{-1}*	*Light dose/J cm^{-2}*	*Diseases treated*
Hematoporphyrin (HpD–Photofrin–Photosan)	628 (3.0×10^3)	1.5–5	75–250	Early stage esophagus, bladder, lung, cervix, stomach, and mouth cancers. Palliative in later stages
ALA (converted to protoporphyrin IX)	635 (5×10^3)	60	50–150	Skin, stomach, colon, bladder, mouth cancers. Esophageal dysplasia. Various nonmalignant conditions
Benzoporphyrin derivative (BpD)	690 (3.5×10^4)	4	150	Age-related macular degeneration (AMD)
Tin etiopurpurin (SnET2–Purlytin)	665 (3.0×10^4)	1.2	150–200	Breast and skin cancers, AMD
Monoaspartyl chlorin e_6 (MACE)	660 (4.0×10^4)	1.0	25–200	Skin cancers
Lutetium texaphyrin (Lu–Tex)	732 (4.2×10^4)	1.0	150	Metastatic brain tumors, breast cancers, atherosclerotic plaques
Aluminum disulfonated phthalocyanine	675 (2.0×10^5)	1.0	50–200	Brain, colon, bladder, and pancreatic cancers. Head and neck cancers in animal studies only
Metatetrahydroxychlorin (mTHPC–temoporphin–Foscan)	652 (3.5×10^4)	0.15	5–20	Head, neck, prostate, pancreas, lung, brain, biliary tract, and mouth cancers. Superior to HpD and ALA in mouth cancers
Palladium pheophorbide (Tookad)	763 (8.6×10^4)	2.0	–	Prostate cancer
Hypericin	590 (3×10^4)	–	–	Psoriasis

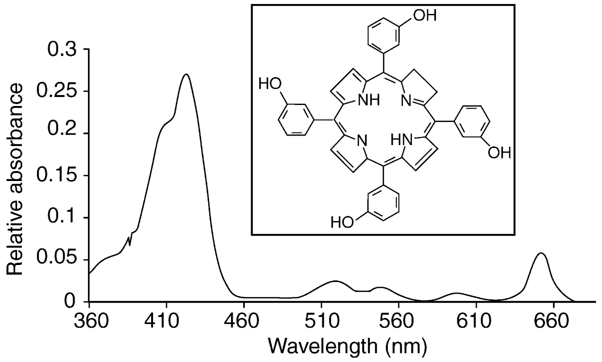

Figure 2 The absorption spectrum of m-THPC. The structure of m-THPC is shown in the inset.

molecule which is cytotoxic, that is, will destroy cells. The first electronically excited state of molecular oxygen, so-called singlet oxygen, fulfills this role very well and may be produced via the interaction of an excited electronic state of the sensitizer with oxygen present in the tissue. Thus, in summary, to achieve effective photosensitization, a sensitizer should exhibit appreciable absorption at red to near-infrared wavelengths and generate cytotoxic species via oxygen-dependent photochemical reactions. The first clinical photosensitizer was hematoporphyrin derivative (HpD), which is derived synthetically from hematoporphyrin by reaction with acetic and sulfuric acids to give a complex mixture of porphyrins. A purified fraction of these (Photofrin) is available commercially, and this has been used most widely in clinical applications to date. Second-generation (photosensitizers are now becoming available, including phthalocyanine and chlorin compounds as shown in Table 1.

A new approach to PDT has recently emerged involving the administration of a natural porphyrin precursor, 5-aminolaevulinic acid (ALA), which is metabolized within cells to produce protoporphyrin IX (see Table 1). This porphyrin is known to be a powerful photosensitizing agent but suffers from the drawback of being a poor tumor localizer when used directly. In contrast, administration of ALA induces protoporphyrin biosynthesis, particularly in rapidly proliferating cells, which may then be destroyed using irradiation at 630 nm. Therefore, this new therapy may offer enhanced treatment selectivity with little risk of skin photosensitivity owing to the rapid clearance after 24 h of the protoporphyrin. Investigation of this new approach has already proved successful in clinical treatment of skin tumors using topical application of ALA in a thick emulsion.

Considerable effort is also being expended on exploiting the retention of sensitizers in tumors for diagnostic purposes, although the results are rather mixed to date. The prospects with ALA-induced protoporphyrin IX are, however, more promising, as illustrated in Figure 3 where the fluorescence is selectively confined to the skin tumor (basal cell carcinoma).

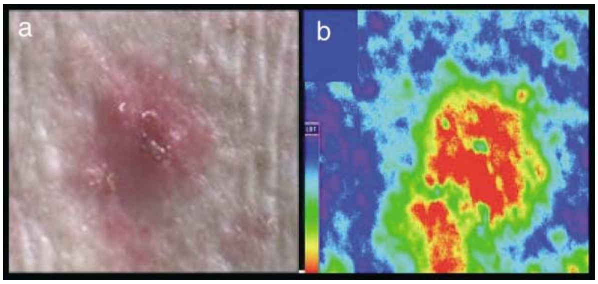

Figure 3 (a) Image of a basal cell carcinoma; (b) fluorescence imaging of the same lesion after ALA sensitization and 405 nm light excitation.

Mechanisms of Photodynamic Therapy

The main principles of the photosensitization mechanism are now well established with the initial step being excitation of the sensitizer from its electronic ground state to the short-lived fluorescent singlet state. The lifetime of the singlet state is generally only a few nanoseconds and the main role of this state in the photosensitization mechanism is to act as a precursor of the metastable triplet state through intersystem crossing. Efficient formation of this metastable state is required because it is the interaction of the triplet state with tissue components that generates cytotoxic species such as singlet oxygen. Thus the triplet state quantum yield (i.e., probability of triplet state formation per photon absorbed) of photosensitizers should ideally approach unity. Interaction of the metastable triplet state (which in de-aerated solutions has a lifetime extending to the millisecond range) with tissue components may proceed via either a type I or II mechanism or a combination (see Figure 4). A type I process can involve hydrogen abstraction from the sensitizer to

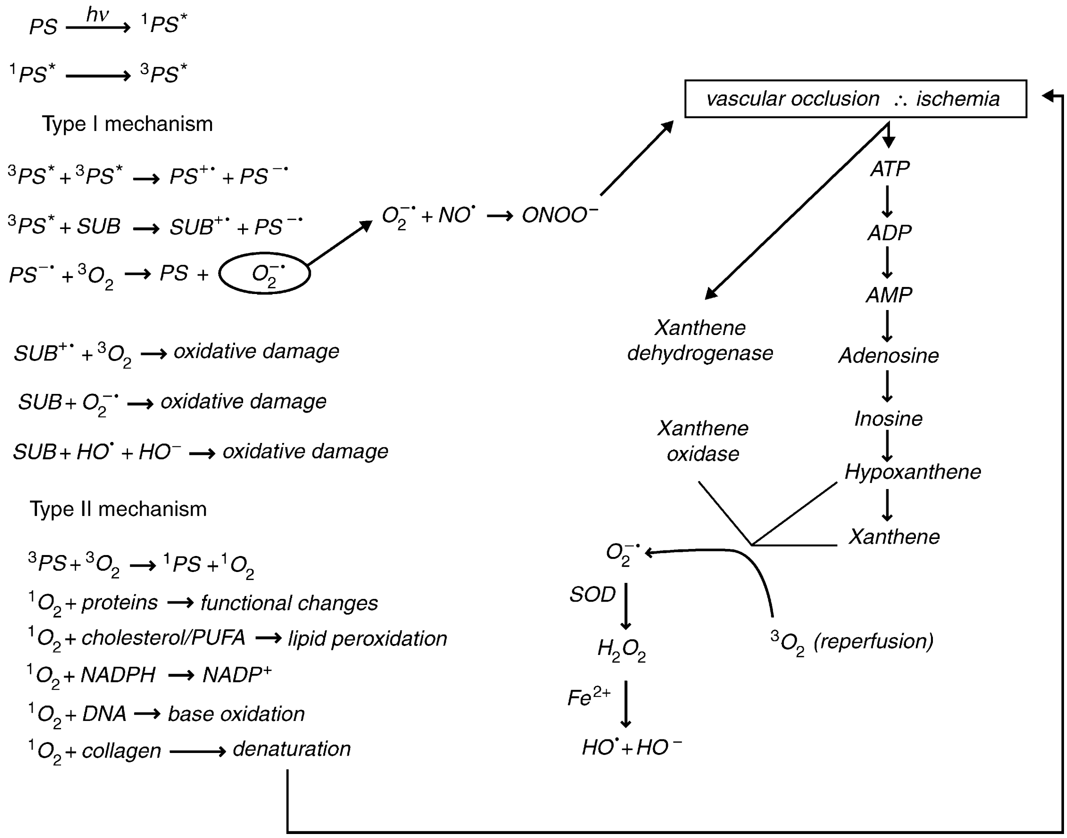
$PS \xrightarrow{h\nu} {}^1PS^*$
${}^1PS^* \longrightarrow {}^3PS^*$
Type I mechanism
${}^3PS^* + {}^3PS^* \longrightarrow PS^{+\bullet} + PS^{-\bullet}$
${}^3PS^* + SUB \longrightarrow SUB^{+\bullet} + PS^{-\bullet}$
$PS^{-\bullet} + {}^3O_2 \longrightarrow PS + O_2^{-\bullet}$
$O_2^{-\bullet} + NO^\bullet \longrightarrow ONOO^-$
vascular occlusion ∴ ischemia
$SUB^{+\bullet} + {}^3O_2 \longrightarrow$ oxidative damage
$SUB + O_2^{-\bullet} \longrightarrow$ oxidative damage
$SUB + HO^\bullet + HO^- \longrightarrow$ oxidative damage
Type II mechanism
${}^3PS + {}^3O_2 \longrightarrow {}^1PS + {}^1O_2$
${}^1O_2$ + proteins ⟶ functional changes
${}^1O_2$ + cholesterol/PUFA ⟶ lipid peroxidation
${}^1O_2 + NADPH \longrightarrow NADP^+$
${}^1O_2$ + DNA ⟶ base oxidation
${}^1O_2$ + collagen ⟶ denaturation
ATP
ADP
AMP
Adenosine
Inosine
Hypoxanthene
Xanthene
Xanthene dehydrogenase
Xanthene oxidase
$O_2^{-\bullet}$
SOD
H_2O_2
Fe^{2+}
$HO^\bullet + HO^-$
${}^3O_2$ (reperfusion)

Figure 4 The photophysical and photochemical mechanisms involved in PDT. PS denotes the photosensitizer, SUB denotes the substrate, either biological, a solvent or another photosensitizer; * denotes the excited state and • denotes a radical.

produce free radicals or electron transfer, resulting in ion formation. The type II mechanism, in contrast, exclusively involves interaction between molecular oxygen and the triplet state to form singlet oxygen which is highly reactive in biological systems. The near-resonant energy transfer from the triplet state to O_2 can be highly efficient and the singlet oxygen yield can approach the triplet state yield provided that the triplet state energy exceeds 94 kJ/mol, the singlet oxygen excitation energy. It is widely accepted that the type II mechanism underlies the oxygen-dependent phototoxicity of sensitizers used for photodynamic therapy. Both proteins and lipids (the main constituents of membranes) are susceptible to photooxidative damage induced via a type II mechanism which generally results in the formation of unstable peroxide species. For example, unsaturated lipids may be oxidized by the 'ene' reaction where the singlet oxygen reacts with a double bond. Other targets include such important biomolecules as cholesterol, certain amino acid residues, collagen, and the coenzyme, NADPH. Another synergistic mode of action involves the occlusion of the microcirculation to tumors and reperfusion injury. Reperfusion injury relates to the formation of xanthene oxidase from xanthene dehydrogenase in anoxic conditions which reacts with xanthene or hypoxanthene products of ATP dephosphorylation to convert the restored oxygen into superoxide anion which directly or via the Fenton reaction causes cellular damage.

An important feature of the type II mechanism, which is sometimes overlooked, is that when the sensitizer transfers electronic energy to O_2 it returns to its ground state. Thus the cytotoxic singlet oxygen species is generated without chemical transformation of the sensitizer which may then absorb another photon and repeat the cycle. Effectively, a singlet photosensitizer molecule is capable of generating many times its own concentration of singlet oxygen, which is clearly a very efficient means of photosensitization provided the oxygen supply is adequate.

Lasers in PDT

Lasers are the most popular light source in PDT since they have several key characteristics that differentiate them from conventional light sources, namely coherence, monochromaticity, and collimated output. The two main practical features that make them so useful in PDT are their monochromaticity and their combination with fiber-optic delivery. The monochromaticity is important since the laser can be tuned to a specific absorption peak of a photosensitizer, thus ensuring that all the energy delivered is utilized for the excitation and photodynamic activation of the photosensitizer. This is not true for a conventional light source (as for example a tungsten lamp) where the output power is divided over several hundred nanometers throughout the UV, visible, and IR regions and only a fraction of the power lies within the absorption band of the photosensitizer. To numerically illustrate this, let us simplify things by representing the lamp output as a square profile and also consider a photosensitizer absorption band about 30 nm full width half maximum. If a laser with a frequency span of about 2 nm and a lamp with the same power but with a frequency span of about 600 nm are used, then the useful portion of it is about 0.05 and if we consider a Gaussian profile for the sensitizer absorption band, the above fraction drops even lower, perhaps even to 0.01. That means that the lamp would achieve 100 times less the excitation rate of the laser, or in other words, we would require a lamp with an output power of 100 times that of the laser to achieve the same rate of excitation and consequently the same time of treatment, provided we use a low enough laser power to remain within the linear regime of the photosensitizer excitation. The other major disadvantage of a lamp source is the lack of collimation which results in low efficiency for fiber-optic delivery.

We now review the different lasers that have found application in PDT. Laser technology has significantly advanced in recent years and there are now a range of options in PDT laser sources for closely matching the absorption profiles of the various photosensitizers. Moreover, since PDT is becoming more widely used clinically, practical considerations, such as portability and turn-key operation are increasingly important. In Figure 5 the optical spectrum is shown from about 380 nm (violet–blue) to 800 nm (near-infrared, NIR). We have superimposed on this spectrum, the light penetration depth of tissue to roughly illustrate how deeply light can penetrate (and consequently activate photosensitizer molecules) into lower-lying tissue layers. The term 'penetration depth' is a measure of the light attenuation by the tissue so the figure of 2 mm corresponds to $1/e$ attenuation of the incident intensity at a tissue depth of 2 mm. It is obvious that light at the blue end of the spectrum has only a superficial effect whereas the more we shift further into the red and NIR regions the penetration depth increases. This is due to two main reasons: absorption and scattering of light by various tissue component molecules. For example, in the blue/green region for the absorption of melanin and hemoglobin is relatively high. But there is a noticeable increase in penetration depth beyond about 600 nm leading to an optimal wavelength region, or

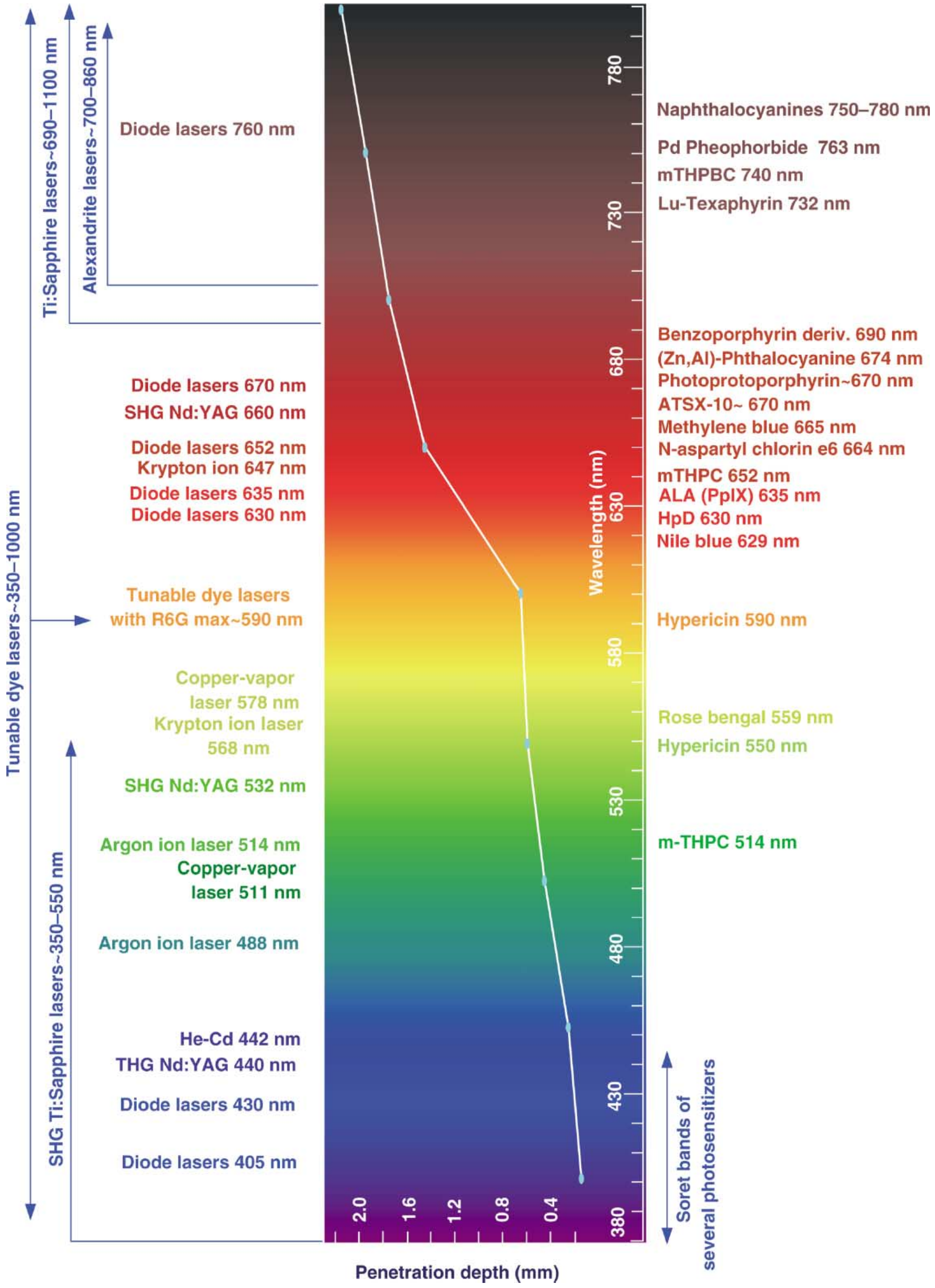

Figure 5 Photosensitizers and laser light sources available in the visible and near-infrared spectral regions.

‘therapeutic window’ for laser therapy of around 600–1100 nm (Figure 5).

The fundamental wavelength of the Nd:YAG laser lies within this therapeutic window at 1064 nm and this laser is now widely used in thermal laser therapy. The Nd:YAG lasers can be operated either in cw (output power ~ 200 W multimode), long-pulse (~ 500 W average power at 50 Hz), or Q-switched

(50 MW peak power at around 10 ns pulse duration) modes. Nd:YAG is a solid-state laser with a yttrium aluminum garnet crystal doped with about 1% of trivalent Nd ions as the active medium of the laser; and with another transition of the Nd ion, this laser (with the choice of different optics) can also operate at 1320 nm. Although dyes are available which absorb from 800–1100 nm, the generation of singlet oxygen via the type II mechanism is energetically unfavorable because the triplet state energies of these dyes are too low. However, the fundamental frequency of the laser can be doubled (second-harmonic generation, SHG) or tripled (third-harmonic generation, THG) with the use of nonlinear crystals, to upconvert the output radiation to the visible range, thus rendering this laser suitable for PDT: i.e., frequency doubling to 532 nm and tripling to 355 nm. Note that the 532 nm output is suitable for activation of the absorption band of hypericin with a maximum at 550 nm even though it is not tuned to the maximum of this absorption.

In the early days of PDT and other laser therapies, ion lasers were widely used. The argon ion laser uses ionized argon plasma as gain medium, and produces two main wavelengths at 488 nm and 514 nm. The second of the two lies exactly at the maximum of the 514 nm m-THPC absorption. Argon ion lasers are operated in cw mode and usually have output powers in the region of 5–10 W at 514 nm (the most powerful argon ion line). Krypton ion lasers, which emit at 568 or 647 nm, are similar to their argon ion counterparts; however, they utilize ionized krypton plasma as gain medium. The 568 nm output can be used to activate the Rose Bengal peak at 559 nm, whereas the 647 nm most powerful line of krypton ion lasers, has been used for activation of m-THPC (652 nm).

Ideally, for optimal excitation, it is best to exactly tune the laser wavelength to the maxima of photosensitizer absorption bands. For this reason tunable organic dye lasers have been widely used for PDT. In these lasers the gain medium is an organic dye with a high quantum yield of fluorescence. Due to the broad nature of their fluorescence gain profile, tuning elements within the laser cavity can be used for selection of the lasing frequency within the gain band. Tunable dye lasers with the currently available dyes can quite easily cover the range from about 350–1000 nm. However, their operation is not of a ‘turn key’ nature since they require frequent replacement of their active material either for tuning to a different spectral region or for a better performance when the dye degrades. Tunable dye lasers also need to be pumped by some other light source like an argon ion laser, an excimer laser, a solid state laser (e.g., Nd:YAG 532 or 355 nm), a copper vapor laser, or lamps. Dye lasers have been used for the excitation of HpD at 630 nm, protoporphyrin IX at 635 nm, photoprotoporphyrin at 670 nm, and phthalocyanines around 675 nm. A further example is hypericin which, apart from the band at 550 nm, has a second absorption band at 590 nm. This is the optimum operation wavelength of dye lasers with rhodamine 590, better known as rhodamine 6G.

Dye lasers can operate in either pulsed or cw mode. However, there is a potential disadvantage in using a pulsed laser for PDT which becomes apparent when using lower repetition rates and higher pulse energies (>1 mJ per pulse) laser excitation. High pulse energies can induce saturation or transient bleaching of the sensitizer during the laser pulse and consequently much of the energy supplied is not efficiently absorbed by the sensitizer. It has been suggested that this effect accounts for the lack of photosensitized damage in tissue sensitized with a phthalocyanine and irradiated by a 5 Hz, 25 mJ-per-pulse flash lamp pumped dye laser. However, using a low pulse energy, high repetition rate copper vapor pumped dye laser (see below), the results were indistinguishable from cw irradiation with an Argon ion dye laser.

Another laser that has found clinical application in PDT is the copper vapor laser. In this laser the active medium is copper vapor at high temperature maintained in the tube by a repetitively pulsed discharge current. The copper vapor lasers are pulsed with typical pulse duration of about 50 ns and repetition rates reaching 20 kHz. Copper vapor lasers have been operated at quite high average powers up to about 40 W. The output radiation is produced at two distinct wavelengths, namely 511 and 578 nm both of which have found clinical use. Copper vapor pumped dye lasers have also been widely used for activating red-absorbing photosensitizers, and the analogous gold vapor lasers operating at 628 nm have been used to activate HpD.

A relatively new class of tunable lasers is the solid state Ti:Sapphire laser. The gain medium in this laser is a $\sim 1\%$Ti doped sapphire (Al_2O_3) crystal and its output may be tuned throughout the 690–1100 nm region, covering many photosensitizer absorption peaks. For example, most of the bacteriochlorin type photosensitizers have their absorption bands in that spectral region, e.g., m-THPBC with an absorption maximum at 740 nm. Also, Texaphyrin sensitizers absorb highly in this spectral region: lutetium texaphyrin has an absorption maximum at 732 nm.

Although tunable dye or solid-state lasers are widely used in laboratory studies for PDT, fixed-wavelength semiconductor diode lasers are gradually supplanting tunable lasers, owing to their practical convenience. It is now generally the case that when a photosensitizer enters clinical trials or enters standard clinical use, laser manufacturers provide dedicated diode lasers with outputs matched to the chosen sensitizer. In this context there are diode lasers available at 630 nm for use with HpD, at 635 nm for use with ALA-induced protoporphyrin IX, 670 nm for ATSX-10 or phthalocyanines, or even in the region of 760 nm, for use with bacteriochlorins. The advantage of these diode lasers is that they are tailormade for use with a particular photosensitizer, they are highly portable, and they have a relatively easy turn-key operation.

So far we have concentrated on red wavelengths but most of the porphyrin and chlorin family of photosensitizers exhibit quite strong absorption bands in the blue region, known as 'Soret' bands. Despite the fact that tissue penetration in this spectral region is minimal, these bands are far more intense than the corresponding red absorption bands of the sensitizer. In this respect it is possible to activate these sensitizers at blue wavelengths, especially in the case of treatment of superficial (e.g., skin) malignant lesions. There are now blue diode lasers (and light-emitting diodes) for the activation of these bands, in particular at 405 nm where the Soret band of protoporphyrin IX lies or at 430 nm for selective excitation of the photoprotoporphyrin species. For very superficial lesions (e.g., in the retina) it may even be possible to use multiphoton excitation provided by femtosecond diode lasers operating at about 800 nm.

Clinical Applications of Lasers in PDT

In the previous section we reviewed the various types of lasers being used for PDT but their combination with fiber-optic delivery is also of key importance for their clinical application. The laser light may be delivered to the target lesion either externally using surface irradiation, or internally within the lesion which is denoted as interstitial irradiation, as depicted in Figure 6. For example, in case of superficial cutaneous lesions the laser irradiation is applied externally (Figure 6a). The area of the lesion is marked prior to treatment to determine the diameter of the laser spot. For a given fluence rate (100–200 W/cm^2) and the area of the spot, the required laser output power is then calculated. A multimode optical fiber, terminated with a microlens to ensure uniform irradiation, is then positioned at a distance from the lesion yielding the desired beam waste, and the surrounding normal tissue is shielded with dark material. However, if the lesion is a deeper-lying solid tumor, interstitial PDT is employed

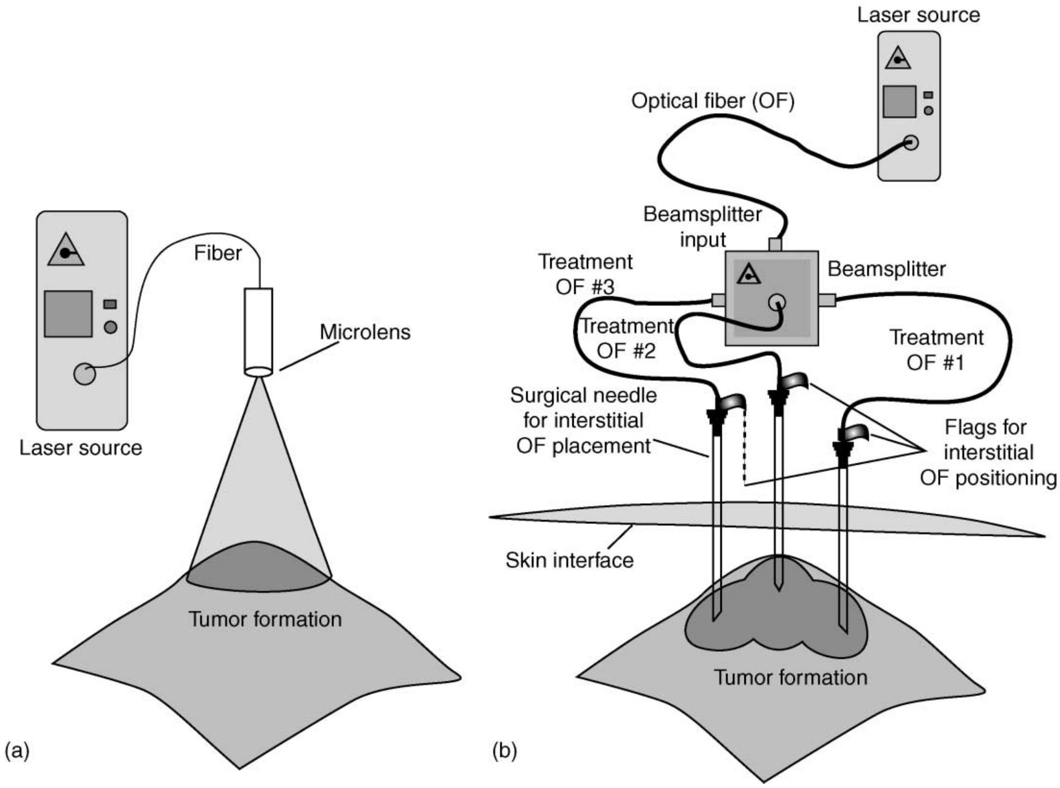

Figure 6 Clinical application of PDT. (a) Surface treatment. (b) Interstitial treatment.

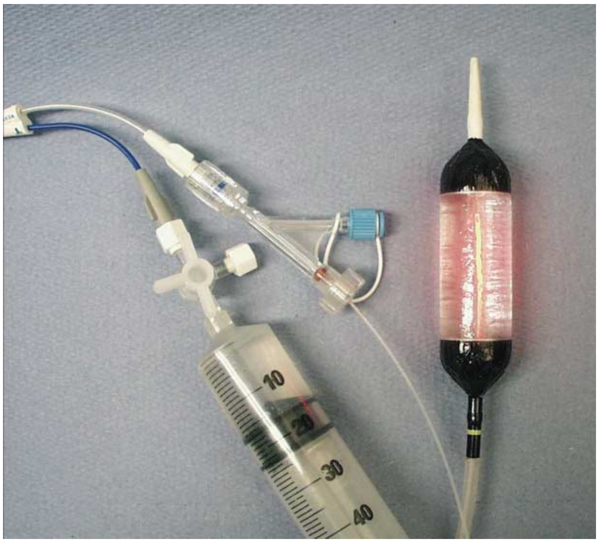

Figure 7 Balloon applicator and diffuser fiber used for PDT in Barrett's esophagus.

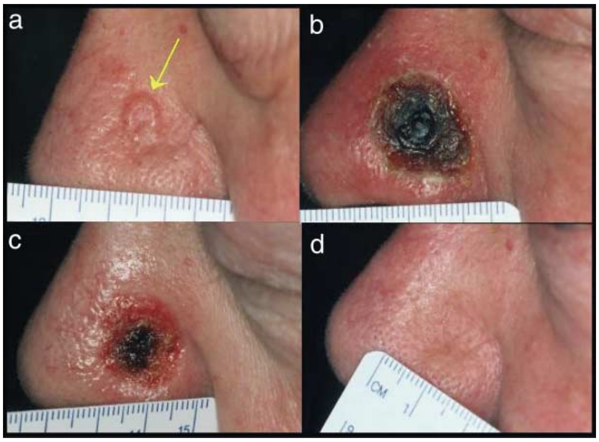

Figure 8 PDT treatment of a nasal basal cell carcinoma with m-THPC (Foscan®): (a) prior to PDT; (b) tissue necrosis a week after treatment; (c) four weeks after treatment; gradual recession of the necrosis and inflammation; (d) healing two months after treatment. Courtesy of Dr Alex Kuebler.

(Figure 6b). Surgical needles are inserted in an equidistant parallel pattern within the tumor, either freehand or under image (MRI/CT) guidance. Bare tip optical fibers are guided through the surgical needles to the lesion and flagged for position. A beamsplitter is used to divide the laser output into 2–4 components so that two to four optical fibers can be simultaneously employed. The laser power is adjusted so that the output fluence of each of the optical fibers is 100–200 mW. Even light distribution in the form of overlapping spheres approximately 1 cm diameter ensures the treatment of an equivalent volume of tissue around each fiber tip. Once each 'station' is likewise treated the fibers are repositioned accordingly, employing a pull-back technique in order for another station to be treated. Once the volume of the tumor is scanned in this manner the treatment is complete. Interstitial light delivery allows PDT to be used in the treatment of large, buried tumors and is particularly suitable for those in which surgery would involve extensive resection.

The treatment of tumors of internal hollow organs is also possible with PDT. The most representative example of this is the treatment of precancerous lesions in the esophagus, known medically as 'Barrett's esophagus'. In this case, a balloon applicator is used to house a special fiber with a light-diffusing tip or 'diffuser' (Figure 7). These diffusers are variable in length and transmit light uniformly along the whole of their length in order to facilitate treatment of circumferential lesions. The balloon applicator is endoscopically inserted into the patient. The balloon is inflated to stretch the esophageal walls and the flagged diffuser fiber is placed in position, central to the balloon optical window. In this way the whole of the lesion may be treated circumferentially at the same time.

Finally, the clinical response to PDT treatment is shown in Figure 8. The treated area becomes necrotic and inflamed within two to four days following PDT. This necrotic tissue usually sloughs off or is mechanically removed. The area eventually heals with minimal scarring (Figure 8).

Future Prospects

A major factor in the future development of PDT will be the application of relatively cheap and portable semiconductor diode lasers. High-power systems consisting of visible diode arrays coupled into multimode fibers with an output of several Watts, have recently become available which will greatly simplify the technical difficulties which have held back PDT. The use of new photosensitizers, in the treatment of certain tumors by PDT, is making steady progress, and although the toxicology testing and clinical evaluation are lengthy and costly processes, the expectation is that these compounds will become more widely available for patient use during the next five years. Regarding the current clinical status of PDT, Photofrin has recently been approved for treatment of lung and esophageal cancers in Europe, USA, and Japan. Clinical trials are in progress for several of the second-generation photosensitizers which offer significant improvements over Photofrin in terms of chemical purity, photoproperties, and skin clearance. With the increasing clinical application of compact visible diode lasers the prospects for photodynamic therapy are therefore encouraging.

See also

Lasers: Dye Lasers; Free Electron Lasers; Metal Vapor Lasers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bonnet R (2000) *Chemical Aspects of Photodynamic Therapy*. Amsterdam: Gordon and Breach Science Publishers.

Dolmans DEJGJ, Fukumura D and Jain RK (2003) Photodynamic therapy for cancer. *Nature Reviews (Cancer)* 3: 380–387.

Hopper C (2000) Photodynamic therapy: a clinical reality in the treatment of cancer. *Lancet Oncology* 1: 212–219.

Milgrom L and MacRobert AJ (1998) Light years ahead. *Chemistry in Britain* 34: 45–50.

Rosen GM, Britigan BE, Halpern HJ and Pou S (1999) *Free Radicals Biology and Detection by Spin Trapping*. New York: Oxford University Press.

Svelto O (1989) *Principles of Lasers*, 3rd edn. New York: Plenum Press.

Vo-Dinh T (2003) *Biomedical Photonics Handbook*. Boca Raton, FL: CRC Press.

Pump and Probe Studies of Femtosecond Kinetics

G D Scholes, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Pump-probe spectroscopy is a ubiquitous time-resolved optical spectroscopy that has found application to the study of all manner of ultrafast chemical processes that involve excited states. The principle of the measurement is that a 'pump pulse' – usually an intense, short laser pulse – impulsively excites a sample, thus defining the start time for the ensuing photophysical dynamics. A probe pulse interrogates the system at later times in order to capture 'snap shots' of the state of the system. Many kinds of pump-probe experiments have been devised. The present article will mainly focus on transient absorption measurements.

Femtosecond pump-probe experiments have been employed to reveal the kinetics of excited state chemical processes such as solvation, isomerization reactions, proton transfer, electron transfer, energy transfer, or photochromism. It is becoming more common to use pump-probe techniques to study systems of increasing complexity, such as photosynthetic proteins, photoreceptor proteins, molecular aggregates, nanostructured materials, conjugated polymers, semiconductors, or assemblies for artificial light harvesting. However, the underlying chemical kinetics are not always easily revealed by pump-probe signals.

The principle of the pump-probe measurement is related to the seminal flash-photolysis experiment devised by Norrish and Porter. However, in order to achieve ultrafast time resolution – beyond that attainable by electronic detection – the probe pulse is temporally short and is controlled to arrive at the sample at variable time delays after the pump pulse has excited the sample. The change in intensity of the probe pulse, as it is transmitted through the sample, is monitored at each pump-probe time delay using a 'slow' detector that integrates over the probe pulse duration. One possible xprocess that can diminish the probe transmission is transient absorption. After formation of an excited state S_1 by the pump pulse, the probe pulse can monitor a resonance with a higher excited state $S_1 \rightarrow S_n$ absorption. Figure 1 shows transient absorption spectra, corresponding to excited state absorption at various time delays, that has been probed by a white light continuum probe and monitored by a multichannel CCD detector. The transient spectrum is seen to blue-shift with

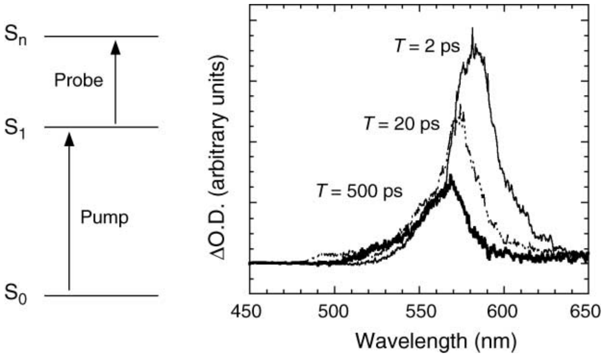

Figure 1 Excited state absorption spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent. The pump wavelength is 277 nm (100 nJ, 150 fs, 40 kHz) and the probe is a white light continuum (500 to 700 nm) generated in a sapphire crystal. The pump-probe delay *T* for each spectrum is indicated on the plot.

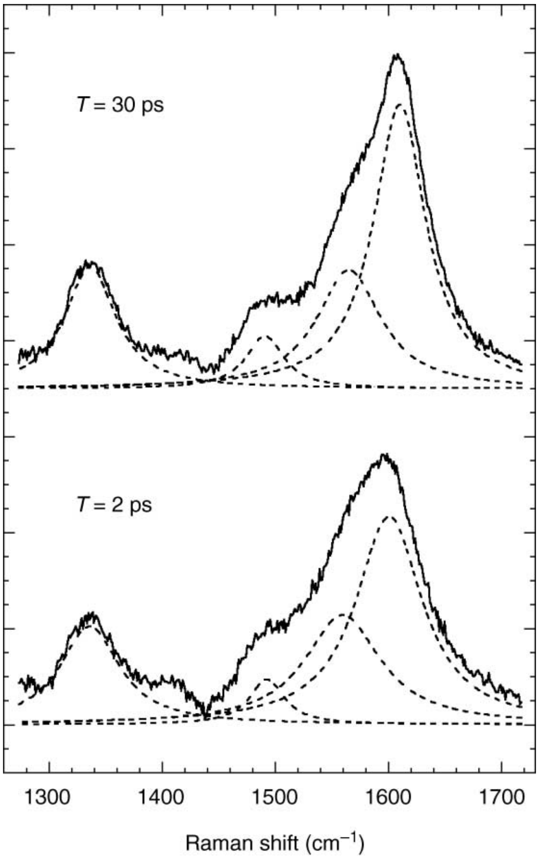

Figure 2 TR^3 spectra of 4′-*n*-pentyl-4-cyanoterphenyl in octanol solvent, recorded using the same set-up as summarized in Figure 1. The dashed lines are deconvolutions of the data to show the Raman bands.

pump-probe delay owing to solvation of the S_1 state of the molecule – the dynamic Stokes' shift. The dynamic Stokes' shift is the shift to lower energy of the emission spectrum as the solvent relaxes in response to the change in dipole moment of the excited electronic state compared to the ground state. The overall intensity of the induced absorption decays according to the lifetime of the excited state.

Although they will not be discussed in detail in the present article, various other pump-probe experiments are possible. For example, when in resonance with this transient absorption, the probe pulse can induce resonance Raman scattering, which reveals the evolution of the vibrational spectrum of the excited state. In this kind of pump-probe experiment we do not monitor the change in probe transmission through the sample. Instead, we monitor the difference between probe-induced resonance Raman scatter from the excited and ground states of the system. The resonance Raman (TR^3) scatter, time-resolved as a function of pump-probe delay, can be detected provided that it is not overwhelmed by fluorescence emission. Typical TR^3 data for different pump-probe delays are shown in Figure 2. The dashed lines show a deconvolution of the spectra to reveal the underlying vibrational bands. These bands are rather broad owing to the spectral width of the 1 ps laser pulses used in this experiment. It is clear that the time-resolution of a TR^3 experiment is limited to the picosecond regime, otherwise most of the vibrational band information is lost. TR^3 and complementary pump-probe infrared transient absorption experiments, utilizing an infrared probe beam, have been useful for observing the evolution of structural information, such as geometry changes, subsequent to photoexcitation.

In the present article we will describe the foundation of experiment and theory necessary to understand resonant ultrafast pump-probe spectroscopy applied to chemical processes in the condensed phase. By resonant, it is meant that the pump and probe pulses have frequencies (energies) resonant with electronic transitions of the molecules being studied. The implications of condensed phase are that the experiment interrogates an ensemble of molecules and the electronic transitions of these molecules are coupled to the random motions and environments of a bath, for example the solvent.

Experimental Measurement of Pump-Probe Data

The experimental setup for a one-color pump-probe experiment (i.e. pump and probe of the same color) is shown in Figure 3. The setup is easily adapted for a two-color experiment. Most of the laser intensity (70%) is transmitted through the first beam splitter so that the pump is more intense than the probe. A second beam splitter is used to split off a small amount of probe light for the reference beam. Note that the arrangement of these beam splitters results in both the pump and probe passing through the same amount of dispersive material on their way to the sample. The pump and probe beams each propagate through half-wave plates in order to control polarization. Usually the probe polarization is set to the magic angle (54.7°) relative to the pump in order to remove polarization bias. The beams are sent towards the sample by way of retroreflectors mounted on $x-y-z$ translation stages. The pump retroreflector is mounted on a precision computer-controlled translation stage such that it can be scanned in the x-direction to arrive at variable delays before and after the probe. The pump and probe beams are aligned using the $y-z$ controls on the retroreflector translation stages together with the pick-off mirrors

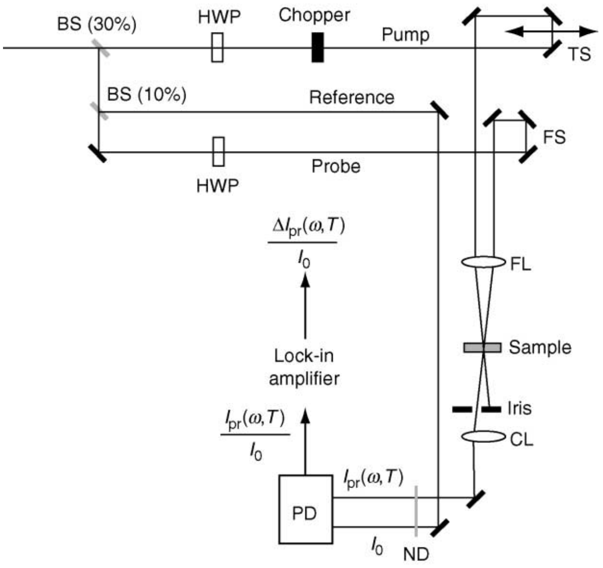

Figure 3 An experimental layout for femtosecond pump-probe spectroscopy. The reference beam ideally passes through the sample (not shown as such in aid of clarity). See text for a description. HWP, half-wave plate; BS, beamsplitter (% reflection); FL, focusing lens; CL, collimating lens; ND, neutral density filter; PD, photodetectors.

such that they travel in parallel towards the sample. The overall setup of the beams is in a box geometry so that pump and probe arms have the same path length.

The pump and probe beams are focused into the sample using a transmissive (i.e. lens) or reflective optic FL. The focal length is typically 20 to 30 cm, providing a small crossing angle to improve phase matching. The pump beam should have a larger spot size at its focal point in the sample than the probe beam, in order to avoid artifacts arising from the wandering of the pump and probe spatial overlap as a function of delay time – a potential problem, particularly for long delays. Good spatial overlap of the beams in the sample is important. It can be checked using a 50 μm pinhole. Fine adjustment of the pump-probe delay, to establish temporal overlap of the pump and probe pulses, can be dictated by autocorrelating the beams in a second harmonic generating crystal mounted at the sample position. At the same time the pulse compression is tweaked to ensure that the pulse dispersion in the experiment has been pre-compensated.

The sample is usually flowed or mounted in a spinning cell if it is liquid, in order to avoid thermal effects or photochemical bleaching. The path length of the sample is typically ≤1 mm. The transmitted probe beam is spatially isolated from the pump beam using an iris, is collimated, and then directed onto the photodetector.

The reference beam, which provides the relative reference intensity I_0 should preferably pass through the sample (away from the pump spot). This provides the probe-only transmission – that is, the reference beam intensity is attenuated according to sample ground state optical density. The photodetector outputs are used to ratio the pump-probe arm intensity transmitted through the sample with the reference intensity, $I_{pr}(\omega, T)/I_0$. Since the pump beam is being chopped at the lock-in reference frequency, the lock-in amplifier outputs the pump-induced fractional change in transmission intensity, $\Delta I_{pr}(\omega, T)/I_0$, usually simply written as $\Delta T/T$. When the pump-induced signal I_{pr} is small compared to the reference intensity I_0, then the detected $\Delta I_{pr}(\omega, T)/I_0$ signal is approximately equal to the change in optical density, ΔO.D.:

$$\Delta\text{O.D.} = \log\left(\frac{I_0 - \Delta I_{pr}}{I_0}\right) \approx \frac{-\Delta I_{pr}}{I_0} \equiv \frac{-\Delta T}{T} \qquad [1]$$

Here we have used a Taylor expansion $\log(1 + x) \approx x$ for small x.

The power dependence of the signal in the $\chi^{(3)}$ limit should be linear in both the pump and probe intensities. It is also possible to pump the sample using two-photon absorption, then probe an excited state absorption. This is a $\chi^{(5)}$ experiment, so that the signal depends quadratically on the pump intensity.

What is Measured in a Pump-Probe Experiment?

Origin of the Signal

The pump-probe measurement is a third-order nonlinear spectroscopy. A complete calculation of the signal requires computation of the third-order polarization induced by the interaction of the pump and probe with the material sample, together with an account for the kinetic evolution of populations (e.g., excited state reactant and product states, etc.).

The pump pulse, wavevector $\mathbf{k}_1$, interacts twice with the sample, thereby creating excited state population density $|e\rangle\langle e|$ and a hole-in-the-ground state population density $|g\rangle\langle g|$. These population densities propagate until the probe pulse, with wavevector $\mathbf{k}_2$, interacts with the system to induce a polarization that depends on the state of the system at time T after the pump. This induced polarization $P_s(t)$ is radiated in the $\mathbf{k}_s = \mathbf{k}_1 - \mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_2$ direction. Because the signal is radiated in the probe direction, the probe pulse acts as a local oscillator to heterodyne the signal. The change in probe-pulse spectral intensity after transmission through an

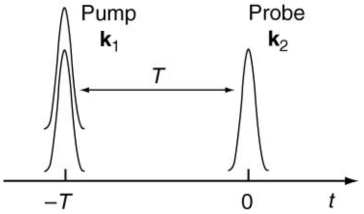

Figure 4 Labeling and time variable for the description of a pump-probe experiment. T defines the pump-probe delay.

optically thin sample of path length l, with concentration of absorbers c and refractive index n, is given by eqn [2]

$$\Delta I_{\mathrm{pr}}(\omega, T) \propto \frac{cl\omega_{\mathrm{pr}}}{n} \mathrm{Im}\left[\int dt E^*_{\mathrm{pr}}(t) P^{(3)}(0, T, t) e^{i\omega t}\right] \quad [2]$$

In eqn [2] $E_{\mathrm{pr}}(t)$ is the probe pulse electric field, ω is the center frequency of the probe, and * indicates complex conjugate. The time variables correspond to those indicated in Figure 4, and the signal is resolved as a function of frequency by Fourier transformation of $t \to \omega$. Experimentally this is achieved by dispersing the transmitted probe using a spectrograph. The first-time variable, the time interval between the two pump pulse interactions, is set to zero because we are interested here in the limit where the pulse duration is much shorter than T. In this regime – impulsive pump-probe – the measurement is sensitive to formation and decay of excited state species.

The induced polarization is given by eqn [3]

$$P^{(3)}(0, T, t) = \int_0^\infty \mathrm{d}t \int_0^\infty \mathrm{d}T [R^{\mathrm{SE}} - R^{\mathrm{ESA}} + R^{\mathrm{GSR}}] \times E_1^*(t - T) E_1(t - T) E_2(t) \quad [3]$$

where the response functions that contain information about the lineshape functions and kinetics are given by eqn [4]:

$$R^{\mathrm{SE}}(0, T, t) = |\mu_{eg}|^2 |\mu_{eg}|^2 \exp(-i\omega_{eg} t) \times \exp[-g^*(t) + 2ig''(T + t) - 2ig''(T)] \times K^{\mathrm{SE}}(T) \quad [4a]$$

$$R^{\mathrm{ESA}}(0, T, t) = |\mu_{eg}|^2 |\mu_{fe}|^2 \exp(-i\omega_{fe} t) \times \exp[-g^*(t) - 2ig''(T + t) + 2ig''(T)] \times K^{\mathrm{ESA}}(T) \quad [4b]$$

$$R^{\mathrm{GSR}}(0, T, t) = |\mu_{eg}|^2 |\mu_{eg}|^2 \exp(-i\omega_{eg} t) \exp[-g(t)] \times K^{\mathrm{GSR}}(T) \quad [4c]$$

The overall intensity of each contribution to the signal is scaled by the transition dipole moment that connects the ground and excited state μ_{eg} or excited state and a higher excited state μ_{fe}. The intensity is further influenced by resonance of the probe-pulse spectrum with the transition frequencies ω_{eg} and ω_{fe}. These transition energies have a time-dependence owing to the relaxation of the excited state density in the condensed phase environment – the dynamic Stokes' shift. The details of the dynamic Stokes' shift depend on the bath, and are therefore contained in the imaginary part of the lineshape function $g(t) = g'(t) + ig''(t)$. The evolution of the Stokes' shift with pump-probe delay T is clearly seen in the excited state absorption spectrum shown in Figure 1. The lineshape function contains all the details of the time-scales of fluctuations of the bath and the coupling of the electronic transition of the molecule to these fluctuations. The lineshape function can be cast in the simple form given by eqn [5] using the Brownian oscillator model in the high temperature, high friction limit:

$$g(t) \approx (2\lambda k_B T_B / \hbar\Lambda)[\Lambda t - 1 + \exp(-\Lambda t)] + i(\lambda/\Lambda)[1 - \exp(-\Lambda t)] \quad [5]$$

The Brownian oscillator model attributes fluctuations of the transition frequencies to coupling between electronic states and an ensemble of low-frequency bath motions. In eqn [5]; λ is the solvent reorganization energy (half the Stokes' shift) and Λ is the modulation frequency of the solvent oscillator (τ_{L}^{-1} for a Debye solvent, where τ_{L} is the longitudinal dielectric relaxation time); k_{B} is the Boltzmann constant, and T_{B} is the temperature.

Information about the kinetics of the evolution of the system initiated by the pump pulse is contained in the terms K^{SE}, K^{ESA}, and K^{GSR}. These terms will be discussed in the next section.

The evolution of the excited state density is mapped onto the response functions R^{SE} and R^{ESA}. R^{SE} is the contribution arising from the probe pulse stimulating emission from the excited state e in the probe direction, which increases the probe intensity. Thus the spectrum of the signal arising from R^{SE} resembles the fluorescence emission of the sample, as shown in Figure 5. The excited state absorption contribution to the signal R^{ESA} depletes the probe intensity, and corresponds to an e → f electronic resonance, such as that shown in Figure 1. Note that the ESA signal contributes to the overall $\Delta I_{\mathrm{pr}}(\omega, T)$ with an opposite sign to the SE and GSR contributions. The ground-state recovery term R^{GSR} arises from depletion of the ground-state population by the pump pulse, thereby increasing the transparency of the sample over the

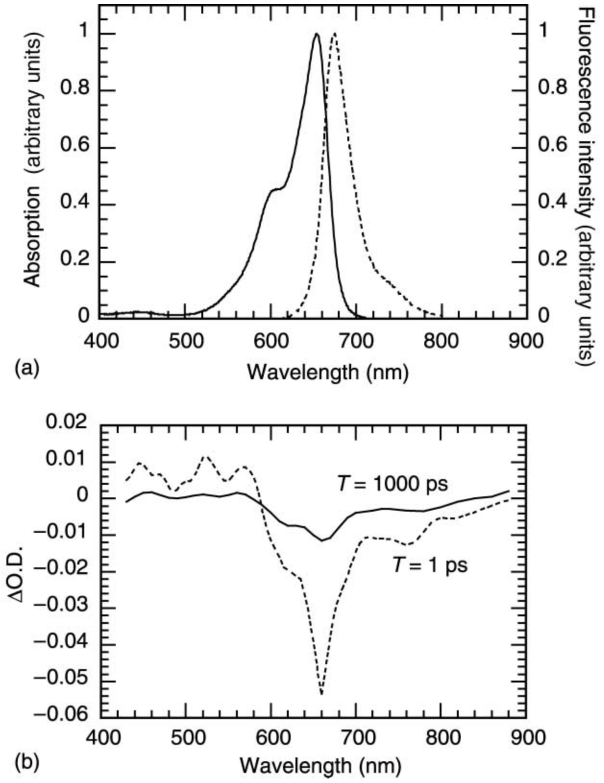

Figure 5 (a) shows the absorption and emission spectrum of a dilute solution of oxazine-1 in water. (b) shows the transient absorption spectrum of this system at various delays after pumping at 600 nm using picosecond pulses. The transient spectrum, ground state recovery and stimulated emission resemble the sum of the absorption and emission spectra.

spectrum of the ground-state absorption (see Figure 5). Thus the transmitted probe intensity is increased relative to that without the pump pulse. The GSR contribution to $\Delta I_{\mathrm{pr}}(\omega, T)$ decreases with T according to the excited state lifetime of the photoexcited species.

The Coherent Spike

To calculate the pump-probe signal over delay times that are of the order of the pulse duration, one must calculate all possible ways that a signal can be generated with respect to all possible time orderings of the two pump interactions and the probe interactions (Liouville space pathways). Such a calculation reveals coherent as well as sequential contributions to the response of the system owing to the entanglement of time orderings that arise from the overlap of pulses of finite duration. The nonclassical, coherent, contributions dominate the signal for the time period during which pump and probe pulses overlap, leading to the coherent spike (or coherent 'artifact') (Figure 6).

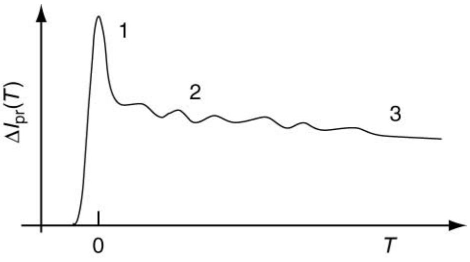

Figure 6 A schematic depiction of a pump-probe signal. The coherent spike is labeled 1. In region 2 we can see quantum beats as a sum of sinusoidal modulations of the signal. The label 3 denotes the long time decay of the signal, as determined by the population dynamics, K^{SE}, K^{GSR}, and K^{ESA}.

Nuclear Wavepackets

The electronic absorption spectrum represents a sum over all possible vibronic transitions, each weighted by a corresponding nuclear overlap factor according to the Franck–Condon principle. It was recognized by Heller and co-workers that in the time-domain representation of the electronic absorption spectrum, the sum over vibrational overlaps is replaced by the propagation of an initially excited nuclear wavepacket on the excited state according to $|i(t)\rangle = \exp(-Ht/\hbar)\,|i(0)\rangle$. Thus the absorption spectrum is written as

$$\sigma_i(\omega) = \frac{2\pi\omega}{3c}\int_{-\infty}^{\infty} \mathrm{d}t\,\exp[-i(\omega-\omega_{eg}^{i})t - g(t)] \times \langle i(0)\,|i(t)\rangle \qquad [6]$$

for mode i. The Hamiltonian H is determined by the displacement of mode i, Δ_i. A large displacement results in a strong coupling between the vibration and the electronic transition from state g to e. In the frequency domain this is seen as an intense vibronic progression. In the time domain this translates to an oscillation in the time-dependent overlap $\langle i(0)\,|i(t)\rangle$, with frequency ω_i, thereby modulating the intensity of the linear response as a function of t.

It follows that an impulsively excited coherent superposition of vibrational modes, created by an optical pulse that is much shorter than the vibrational period of each mode, propagates on the excited state as a wavepacket. The oscillations of this wavepacket then modulate the intensity of the transmitted probe pulse as a function of T to introduce features into the pump-probe signal known as quantum beats (see Figure 6). In fact, nuclear wavepackets propagate on both the ground and excited states in a pump-probe experiment. The short probe pulse interrogates the evolution of the nuclear wavepackets by projecting them onto the manifold of vibrational wavefunctions in an excited or ground state.

Femtosecond Kinetics

According to eqn [4] the kinetics of the evolution of the system initiated by the pump pulse are governed by the terms K^{SE}, K^{ESA}, and K^{GSR}. Each of these terms describes the kinetics of decay, production, or recovery of excited state and ground state populations. In the simplest conceivable case, each might be a single exponential function. The stimulated emission term K^{SE} contains information about the depopulation of the initially excited state, which for example, may occur through a chemical reaction involving the excited state species or energy transfer from that state. The excited state absorption term K^{ESA} contains information on the formation and decay of all excited state populations that give a transient absorption signal. The ground-state recovery term K^{GSR} contains information on the time-scales for return of all population to the ground state, which will occur by radiative and nonradiative processes. Usually the kinetics are assumed to follow a multi-exponential law such that

$$K^m(T, \lambda_{exc}, \omega) = \sum_{j=1}^{N} A_j(\lambda_{exc}, \omega) \exp(-t/\tau_j) \qquad [7]$$

where the amplitude A_j, but not the decay time coefficient τ_j, of each contribution in the coupled N-component system depends on the excitation wavelength λ_{exc} and the signal frequency ω. The index m denotes SE, ESA, or GSR.

In order to extract meaningful physical parameters from analysis of pump-probe data, a kinetic scheme in the form of eqn [7] needs to be constructed, based on a physical model to connect the decay of initially excited state population to formation of a product excited state and subsequent recovery of the ground-state population. An example of such a model is depicted in Figure 7. Here the stimulated emission contribution to the signal is attenuated according to the rate that state 1 goes to 2. The transient absorption appears at that same rate and decays according to the radiationless process 2 to 3. The rate of ground state recovery depends on 1 to 2 to 3 to 5 and possibly pathways involving the isomer 4.

The kinetic scheme is not easily extracted from pump-probe data because contributions from each of K^{SE}, K^{ESA}, and K^{GSR} contribute additively at each detection wavelength ω, as shown in eqns [2]–[4]. Moreover, additional, wavelength-dependent, decay or rise components may appear in the data on ultrafast time-scales owing to the time-dependent Stokes' shift (the $g''(t)$ terms indicated in eqns [4]). A spectral shift can resemble either a decay or rise component. In addition, the very fastest dynamics may be hidden beneath the coherent spike. For these reasons, simple fitting of pump-probe data usually cannot reveal the underlying kinetic model. Instead the data should be simulated according to eqns [3] and [4].

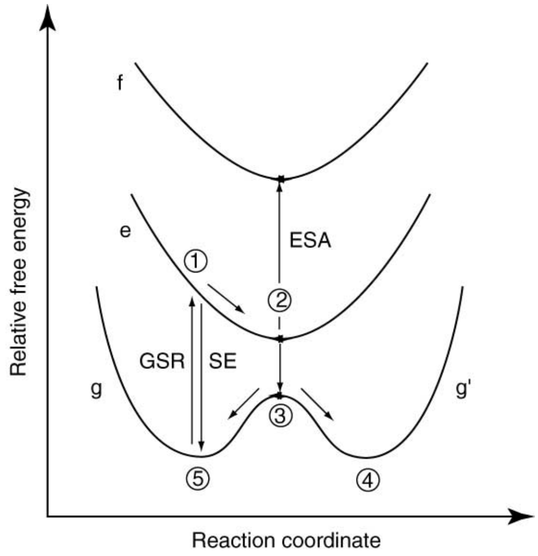

Figure 7 Free energy curves corresponding to a model for an excited state isomerization reaction, with various contributions to a pump-probe measurement indicated. See text for a description. (Courtesy of Jordanides XJ and Fleming GR.)

Alternative strategies involve global analysis of wavelength-resolved data to extract the kinetic evolution of spectral features, known as 'species associated decay spectra'. Global analysis of data effectively reduces the number of adjustable parameters in the fitting procedure and allows one to obtain physically meaningful information by associating species with their spectra. These methods are described in detail by Holzwarth.

List of Units and Nomenclature

ESA	excited state absorption
GSR	ground state recovery
SE	stimulated emission
S_n	nth singlet excited state
TR^3	time-resolved resonance Raman
$\chi^{(3)}$	nth-order nonlinear optical susceptibility

See also

Coherent Lightwave Systems. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. New York: Oxford University Press.

Heller EJ (1981) The semi-classical way to molecular spectroscopy. *Accounts of Chemical Research* 14: 368–375.

Holzwarth AR (1996) Data analysis of time-resolved measurements. In: Amesz J and Hoff AJ (eds) *Biophysical Techniques in Photosynthesis*, pp. 75–92. Dordrecht, Germany: Kluwer.

Ippen EP, Shank CV, Wiensenfeld JM and Migus A (1980) Subpicosecond pulse techniques. *Philosophical Transactions of the Royal Society of London A* 298: 225–232.

Jonas DM and Fleming GR (1995) Vibrationally abrupt pulses in pump-probe spectroscopy. In: El-Sayed MA, Tanaka I and Molin Y (eds) *Ultrafast Processes in Chemistry and Photobiology*, pp. 225–256. Oxford, UK: Blackwell Scientific.

Klimov VI and McBranch DW (1998) Femtosecond high-sensitivity, chirp-free transient absorption spectroscopy using kilohertz lasers. *Optics Letters* 23: 277–279.

Lee SY (1995) Wave-packet model of dynamic dispersed and integrated pump-probe signals in femtosecond transition-state spectroscopy. In: Manz J and Wöste L (eds) *Femtosecond Chemistry*, vol. 1, pp. 273–298. Weinheim, Germany: VCH.

Lin SH, Alden R, Islampour R, Ma H and Villaeys AA (1991) *Density Matrix Method and Femtosecond Processes*. Singapore: World Scientific.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Norrish RGW and Porter G (1949) Chemical reactions produced by very high light intensities. *Nature* 164: 658.

Towrie M, Parker AW, Shaikh W and Matousek P (1998) Tunable picosecond optical parametric generator-amplifier system for time resolved Raman spectroscopy. *Measurement Science and Technology* 9: 816–823.

Yan YJ and Mukamel S (1990) Femtosecond pump-probe spectroscopy of polyatomic molecules in condensed phases. *Physical Reviews A* 41: 6485–6504.

Time-Correlated Single-Photon Counting

A Beeby, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Time-correlated single-photon counting, TCSPC, is a well-established technique for the determination of fluorescence lifetimes and related time-resolved fluorescence properties. Since the 1980s, there have been dramatic changes in the laser sources used for this work, which in turn have opened up the technique to a wide user base.

TCSPC: The Basics

In the TCSPC experiment, the sample is repeatedly excited by a high-repetition rate, low-intensity light source. Fluorescence from the sample is collected and may be passed though a cut-off filter or monochromator to remove scattered excitation light and to select the emission wavelength. The fluorescence is then focused onto a detector, typically a photomultiplier (PMT) or single-photon avalanche diode (SPAD) which detects single emitted photons. The 'raw' signal from the detector fluctuates considerably from event to event. In order to remove any timing error that could be induced by this, the signal is processed first by a constant fraction discriminator (CFD) in order to provide a stable and consistent timing pulse. This signal is used as an input for the time-to-amplitude converter (TAC) which determines the time interval between the excitation of the sample and the emission of the photon. The output from the TAC is a relatively slow pulse whose intensity is proportional to the time interval between the start and stop signals. A pulse height analyzer (PHA) is used to process the output from the TAC and increments a channel in its memory corresponding to a time window during which the photon was detected. As the measurement is repeated many times over, a histogram is constructed in the memory of the PHA, showing the number of photons detected as a function of the time interval. This histogram corresponds to the fluorescence decay.

As outlined below, it is desirable to have a light source with a moderately high repetition rate, typically of the order of MHz. Because it is not practical to start the TAC each time the light source 'fires', it is more usual to operate the TAC in the so-called 'reversed mode', whereby the start signal is derived from the detected photon and the stop is derived from the light source. The electronic circuitry is illustrated in Figure 1. In practice, many researchers still use modular components that are wired together manually, although there is increasingly a tendency towards the use of single PC-based cards which contain all the necessary discriminators, timing circuits, and data acquisition electronics. There are a number of commercial PC cards that contain all the timing electronics, etc., required for TCSPC, for example, Becker and Hickl (Germany).

This approach to obtaining the fluorescence decay profile can only work if the detected photons are 'randomly selected' from the total emission from the sample. In practice, this means that the rate of detection of photons should be of the order of 1% of the excitation rate: that is, only one in a hundred excitation pulses gives rise to a detected photon. This histogram mirrors the fluorescence decay of the sample convolved with the instrument response function (IRF). The IRF is itself a convolution of the temporal response of the optical light pulse, the optical path in the collection optics, and the response of the detector and ancillary electronics. Experimentally, the IRF is usually recorded by placing a scattering material in the spectrometer.

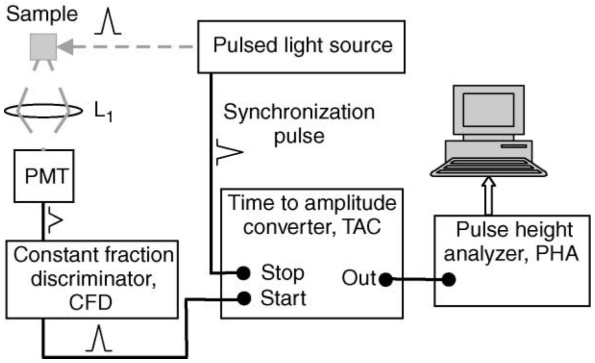
Sample
Pulsed light source
L_1
Synchronization pulse
PMT
Time to amplitude converter, TAC
Pulse height analyzer, PHA
Constant fraction discriminator, CFD
Stop
Start
Out

Figure 1 Schematic diagram of a TSCPC spectrometer. Fluorescence from the sample is collected by L_1, which may also include a monochromator/wavelength selection filter. A synchronization pulse is obtained either from the laser driver electronics directly, or via a photodiode and second constant fraction discriminator. The TAC is shown operating in a 'reversed mode' as is normal with a high repetition rate laser source.

In principle, the decay function of the sample can be extracted by deconvolution of the IRF from the decay function, although in practice it is more common to use the method of iterative re-convolution using a model decay function(s). Thus, the experimentally determined IRF is convolved with a model decay function and parameters within the function are systematically varied to obtain the best fit for a specific model. The most commonly used model assumes the decaying species to follow one or more exponential functions (sum of exponentials) or a distribution of exponentials. Further details regarding the mechanics of data analysis and fitting procedures can be found in the literature. A typical fit is illustrated in Figure 2.

The main aim of this review is to illustrate how recent advances in laser technology have changed the face of TCSPC and to discuss the types of laser light sources that are in use. The easiest way of approaching this topic is to define the criteria for an ideal light source for this experiment, and then to discuss the ways in which new technologies are meeting these criteria.

High Repetition Rate

A typical decay data set may contain something of the order of 0.1–5 million 'counts': it is desirable to

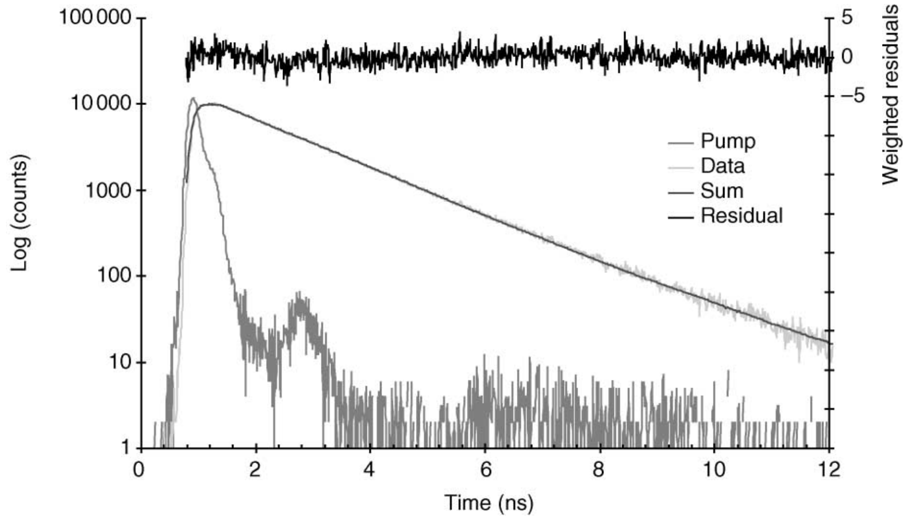

Figure 2 The fluorescence decay of Rhodamine B in water. The sample was excited using a 397 nm pulsed laser diode (1 MHz, < 200 ps FWHM), and the emission was observed at 575 nm. Deconvolution of the IRF (dark gray trace) and observed decay (light gray trace) with a single exponential function gave a lifetime of 1.50 ns, with a $\chi^2 R = 1.07$, Durbin–Watson parameter = 1.78. The fitted curve is shown overlaid in gray, and the weighted residuals are shown offset in black.

acquire such a number of events in order that we can reliably employ statistical parameters to judge the quality of the fitted decay function(s). As stated above, the rate of photon acquisition should be of the order of 1% of the excitation rate in order to obtain reliable, good-quality data. It is clear then that a high repetition rate source is desirable in order to reduce the overall data acquisition time. However, it is important that the repetition rate of the source is not too high, otherwise the data show 'overlap' effects, caused by the fluorescence intensity from one excitation pulse not completely decaying before the next laser pulse arrives at the sample. Under these conditions the observed fluorescence decay is perturbed and extraction of complex kinetic behavior is more difficult. As a general rule of thumb, it is recommended that the inter-pulse separation is greater than five times the lifetime of the longest component of the fluorescence decay. Fluorescence lifetimes can be as long as 100 ns, indicating a maximum repetition rate of 2 MHz, although when investigating fluorophores with a shorter lifetime correspondingly higher frequencies can be selected. Clearly the ability to vary the repetition rate according to the type of sample under investigation is an advantage.

Short Pulse Duration

Typically the minimum lifetime that can be determined by a TCSPC spectrometer is limited by the instrument response function: as a general rule it is possible to extract fluorescence lifetimes of about one tenth of the full width at half maximum (FWHM) of the IRF by reconvolution methods. As discussed above, the IRF is the product of a number of factors, including the optical pulse duration, differences in the optical path traversed by the fluorescence, the response time of the detector, and the temporal response of the associated timing electronics. Ideally the excitation source should be as short as possible, and with modern laser-based TCSPC systems, it is usually the case that the time-response of the detector is the limiting factor. The introduction of high-speed microchannel plate (MCP) photomultipliers have lead to very short IRFs, which can be <50 ps FWHM. When using these detectors is it essential to keep the optical transit-time spread as low as possible: that is, the path taken by light through a monochromator should be independent of wavelength and trajectory taken through the optics. Many research groups advocate the use of a subtractive dispersive double monochromator for this purpose.

Tunability

Early TCSPC experiments were carried out using hydrogen- or nitrogen-filled flashlamps which produced broadband ns-duration pulses with low repetition rates (<100 kHz). Variable wavelength excitation is possible using these sources, but their low intensity and low repetition rates mean that data acquisition periods can be very long. The early 1980s saw the introduction of cavity-dumped ion lasers and synchronously pumped mode-locked cavity-dumped dye lasers, which provided significantly higher repetition rates, typically up to 4 MHz, and greater intensities. However, the TCSPC experiments that could be carried out using these laser-based sources were somewhat limited by the wavelengths that could be generated by the argon ion and synchronously pumped dye lasers, and their second harmonics. The dye lasers were restricted to the use of Rhodamine 6G as the gain medium, which has a limited tuning range (~580–620 nm). It is often desirable to vary the excitation wavelength as part of a photophysical study, either as a systematic study of the effects of excitation energy upon the excited state behavior, or in order to optimize excitation of a particular chromophore in a system. To some extent in the past, the wavelengths available from the laser systems dictated the science that they could be used to address. The mid- to late-1980s saw the introduction of mode-locked continuous wave (CW) Nd:YAG lasers as pump lasers and increased number of dyes that could be synchronously pumped, providing a broader range of excitation wavelengths as well as increased stability. One drawback of these complex and often fickle laser systems was the level of expertise and attention they required to keep their performance at its peak.

Since the 1990s, the introduction of two important new types of laser have had a significant impact on the use of TCSPC as an investigative method. The first is based upon the mode-locked titanium sapphire laser which offers tuneable radiation from the deep UV to the NIR, whilst the second series of laser are based upon (relatively) low-cost, turn-key solid state diode lasers. The developments and merits of these new laser sources, which now dominate the market for TCSPC sources, are discussed below.

Ti-Sapphire

The discovery of the titanium doped sapphire laser has led to a revolution in ultrafast laser sources. Coupled with efficient diode-pumped solid state

Nd:YAG pump lasers, this medium currently dominates the entire ultrafast laser market, from high repetition, low pulse energy systems through to systems generating very high pulse energies at kHz repetition rates.

Ti-sapphire is, in many respects, an ideal laser gain medium: it has a high energy storage capacity, it has a very broad-gain bandwidth, and it has good thermal properties. Following the first demonstration of laser action in a Ti-sapphire crystal, the CW Ti-sapphire laser was quickly commercialized. It proved to be a remarkable new laser medium, providing greater efficiency, with outputs of the order of 1 W. The laser also provided an unprecedented broad tuning range, spanning from just below 700 nm to 1000 nm. The most important breakthrough for time-resolved users came in 1991, when self-mode-locking in the medium was demonstrated. This self-mode-locking, sometimes referred to as Kerr-lens mode-locking, arises from the nonlinear refractive index of the material and is implemented by simply placing an aperture within the cavity at an appropriate point. The broad-gain bandwidth of the medium results in very short pulses: pulse durations of the order of tens of femtoseconds can be routinely achieved with commercial laser systems.

The mode-locked Ti-sapphire laser was quickly seized upon and commercialized. The first generation of systems used multiwatt CW argon ion lasers to pump them, making the systems unwieldy and expensive to run, but in recent years these have been replaced by diode-pumped Nd:YAG lasers, providing an all solid-state, ultrafast laser source. Equally significantly, the current generation of lasers are very efficient and do not require significant electrical power or cooling capacity. There are several commercial suppliers of pulsed Ti-sapphire lasers, including Coherent and Spectra Physics.

A typical mode-locked Ti-sapphire laser provides an 80 MHz train of pulses with a FWHM of ~100 fs, with an average power of up to 1.5 W and are tuneable from 700 nm to almost 1000 nm. The relatively high peak powers associated with these pulses means that second- or even third-harmonic generation is relatively efficient, providing radiation in the ranges 350–480 nm and 240–320 nm. The output powers of the second and third harmonics far exceeds that required for TCSPC experiments, making them an attractive source for the study of fluorescence lifetime measurements. The simple mode-locked Ti-sapphire laser has some drawbacks: the very short pulse duration provides a broad-bandwidth excitation, the repetition rate is often too high with an inter-pulse separation of ~12.5 ns, and the laser does not provide continuous tuning from the deep-UV though to the visible. Incrementally, all of these points have been addressed by the laser manufacturers.

As well as operating as a source of femtosecond pulses, with their associated bandwidths of many tens of nm, some modern commercial Ti-sapphire lasers can, by adjustment of the optical path and components, be operated in a mode that provides picosecond pulses whilst retaining their broad tuning curve. Under these conditions, the bandwidth of the output is greatly reduced, making the laser resemble a synchronously pumped mode-locked dye laser.

As stated above, a very high repetition rate can be undesirable if the system under study has a long fluorescence lifetime due to the fluorescence decay from an earlier pulse not decaying completely before the next excitation pulse. Variation of the repetition of output frequency can be addressed by two means. The first is to use a pulse-picker. This is an opto-accoustic device which is placed outside the laser cavity. An electrical signal is applied to a device, synchronized to the frequency of the mode-locked laser to provide pulses at frequencies from ~4 MHz down to single-shot. The operation of a pulse picker is somewhat inefficient, with less than half of the original pulse energy being obtained in the output, although output power is not usually an issue when using these sources. More importantly, when using pulse-pickers, it is essential to ensure that there is no pre- or post-pulse which can lead to artifacts in the observed fluorescence decays. The second means of reducing the pulse duration is cavity dumping. This is achieved by extending the laser cavity and placing an opto-acoustic device, known as a Bragg cell, at a beam-waist within the cavity. When a signal is applied to the intracavity Bragg cell, synchronized with the passage of a mode-locked pulse though the device, a pulse of light is extracted from the cavity. The great advantage of cavity-dumping is that it builds up energy within the optical cavity and when pulses are extracted these are of a higher pulse energy than the mode-locked laser alone. This in turn leads to more efficient harmonic generation. This addition of the cavity dumper system extends the laser cavity and hence reduces the overall frequency of the mode-locked laser to ~50 MHz, and can provide pulses at 9 MHz down to single shot.

The extension of the tuning range of the Ti-sapphire laser has been facilitated by the introduction of the optical parametric oscillator (OPO). In the OPO, a short wavelength pump beam passes through a nonlinear optical crystal and generates two tuneable

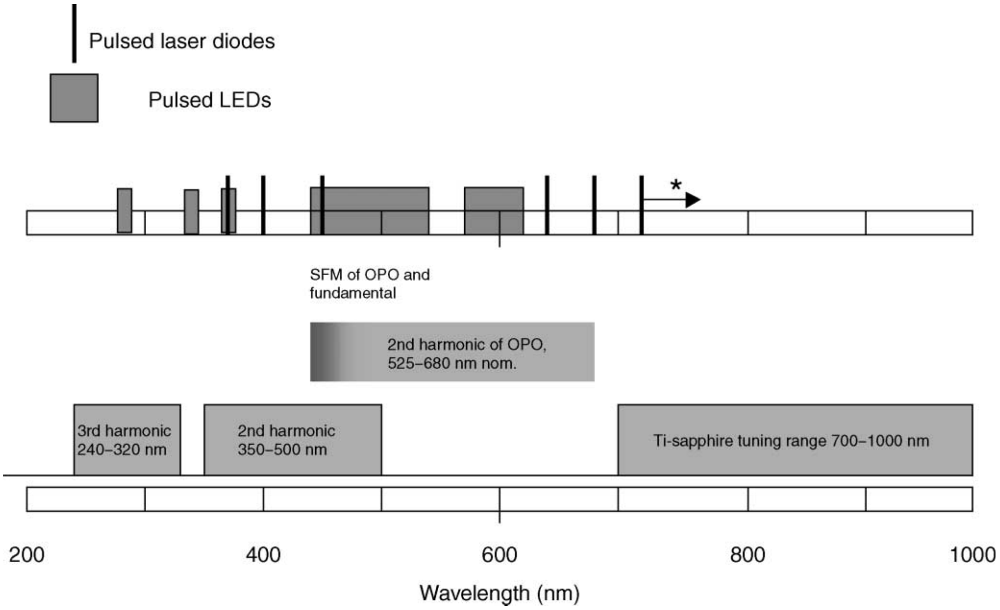

Figure 3 An illustration of the spectral coverage provided by currently available pulsed semiconductor lasers and LEDs, and a modern Ti-sapphire laser and OPO. Note that the OPO's output is frequency doubled, and the wavelengths in the range 525– ~ 440 nm may be obtained by sum frequency mixing of the Ti-sapphire's fundamental output with the output of the OPO. $*$ denotes that there are many long wavelength pulsed laser diodes falling in the range $\lambda > 720$ nm.

beams of longer wavelengths. Synchronous pumping of the OPOs with either a fs or ps Ti-sapphire laser ensures reasonably high efficiencies. The outputs from the OPO may be frequency doubled, with the option of intracavity doubling enhancing the efficiency still further. Using an 800 nm pump wavelength, an OPO provides typical signal and idler tuning ranges of 1050– > 1350 nm and 2.1–3.0 μm, respectively. The signal can be doubled to give 525–680 nm, effectively filling the visible spectrum. The average powers obtainable from Ti-sapphire pumped OPOs makes them eminently suitable for TCSPC experiments. A summary of the spectral coverage of the Ti-sapphire laser and OPO, and low-cost solid-state sources are illustrated in Figure 3.

An additional advantage of the Ti-sapphire laser is the possibility of multiphoton excitation. The combination of high average power and very short pulse duration of a mode-locked Ti-sapphire laser means that the peak-powers of the pulses can be very high. Focusing the near-IR output of the laser gives rise to exceedingly high photon densities within a sample which can lead to the simultaneous absorption of two or more photons. This is of particular significance when recording fluorescence lifetimes in conjunction with a microscope, particularly when the microscope is operated in a confocal configuration. Such systems are now being used for a derivative of TCSPC, known as fluorescence lifetime imaging (FLM).

In summary, the Ti-sapphire laser is a very versatile and flexible source that is well suited to the TCSPC experiment. Off-the-shelf systems provide broad spectral coverage and variable repetition rates. Undoubtedly this medium will be the mainstay of TCSPC for some time to come.

Diode Lasers

Low-power diode lasers, with outputs in the range 630–1000 nm, have been commercially available for some time. These devices, which are normally operated as CW sources, can be operated in a pulsed mode, providing a high frequency train of sub-ns pulses. However, the relatively long output wavelengths produced by this first generation of diode lasers severely restricted the range of chromophores that could be excited and this proved to be a limiting factor for their use as TCSPC sources. Attempts to generate shorter wavelengths from these NIR laser diodes by second harmonic generation (SHG), were never realized: the SHG process is extremely inefficient due to the very low peak output power of these lasers. Shorter wavelengths could be obtained by pulsing conventional light-emitting diodes (LEDs), which were commercially available with emission wavelengths covering the visible spectrum down to ~450 nm. However, these had the disadvantage of exhibiting longer pulse durations than those achievable from laser diodes and have much broader emission bandwidths than a laser source.

The introduction of violet and blue diode lasers and UV-LEDs by the Nichia Corporation (Japan) in the late 1990s resulted in a renewed interest in sources utilizing these small, solid-state devices. At least three manufacturers produce ranges of commercial TCSPC light sources based upon pulsed LEDs and laser diodes and can now boast a number of pulsed laser sources in their ranges including 375, 400, 450, 473, 635, 670, and 830 nm. Currently commercial pulsed laser diodes and LEDs are manufactured by Jobin-Yvon IBH (UK), Picoquant (Germany), and Hamamatso (Japan). These small, relatively low-cost devices offer pulse durations of <100 ps and excitation repetition rates of up to tens of MHz making them ideally suited to 'routine' lifetime measurements. Although these diode laser sources are not tunable and do not span the full spectrum, they are complemented by pulsed LEDs which fill the gap in the visible spectrum. Pulsed LEDs generally have a longer pulse duration and broader, lower intensity spectral output, but are useful for some applications. At the time of writing there are no diode laser sources providing wavelengths of <350 nm, although Jobin-Yvon IBH (UK) have recently announced pulsed LEDs operating at 280 nm and 340 nm. As LEDs, and possibly diode lasers, emitting at these shorter wavelengths become more widely available, the prospects of low running costs and ease of use makes these semiconductor devices an extremely attractive option for many TCSPC experiments. Commercial TCSPC spectrometers based around these systems offer a comparatively low-cost, turnkey approach to fluorescence lifetime based experiments, opening up what was a highly specialized field to a much broader range of users.

See also

Lasers: Noble Gas Ion Lasers. **Light Emitting Diodes**. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Lakowicz JR (ed.) (2001) *Principles of Fluorescence Spectroscopy*. Plenum Press.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*.

Spence DE, Kean PN and Sibbett W (1991) *Optics Letters* 16: 42–44.

Transient Holographic Grating Techniques in Chemical Dynamics

E Vauthey, University of Geneva, Geneva, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past two decades, holographic techniques have proved to be valuable tools for investigating the dynamics of chemical processes. The aim of this chapter is to give an overview of the main applications of these techniques for chemical dynamics. The basic principle underlying the formation and the detection of elementary transient holograms, also called transient gratings, is first presented. This is followed by a brief description of a typical experimental setup. The main applications of these techniques to solve chemical problems are then discussed.

Basic Principle

The basic principle of the transient holographic technique is illustrated in Figure 1. The sample material is excited by two laser pulses at the same wavelength and crossed at an angle $2\theta_{pu}$. If the two pump pulses have the same intensity, I_{pu}, the intensity distribution at the interference region, assuming plane waves, is

$$I(x) = 2I_{pu}\left[1 + \cos\left(\frac{2\pi x}{\Lambda}\right)\right] \qquad [1]$$

where $\Lambda = \lambda_{pu}/(2\sin\theta_{pu})$ is the fringe spacing, λ_{pu} is the pump wavelength, and I_{pu} is the intensity of one pump pulse.

As discussed below, there can be many types of light–matter interactions that lead to a change in the optical properties of the material. For a dielectric material, they result in a spatial modulation of the optical susceptibility and thus of the complex refractive index, $\tilde{n}$. The latter distribution can be described as a Fourier cosine series:

$$\tilde{n}(x) = \tilde{n}_0 + \sum_{m=1}^{\infty} \tilde{n}_m \cos\left(\frac{m2\pi x}{\Lambda}\right) \qquad [2]$$

where $\tilde{n}_0$ is the average value of $\tilde{n}$. In the absence of saturation effects, the spatial modulation of $\tilde{n}$ is harmonic and the Fourier coefficients with $m > 1$ vanish. In this case, the peak to null variation of the complex refractive index, $\Delta\tilde{n}$, is equal to the Fourier coefficient $\tilde{n}_1$. The complex refractive index can be split into its real and imaginary components:

$$\tilde{n} = n + iK \qquad [3]$$

where n is the refractive index and K is the attenuation constant.

The hologram created by the interaction of the crossed pump pulses consists in periodic, one-dimensional spatial modulations of n and K. Such distributions are nothing but phase and amplitude gratings, respectively. A third laser beam at the probe wavelength, λ_{pr}, striking these gratings at Bragg angle, $\theta_B = \arcsin(\lambda_{pr}/2\Lambda)$, will thus be partially diffracted (Figure 1b). The diffraction efficiency, η, depends on the modulation amplitude of the optical properties. In the limit of small diffraction efficiency ($\eta < 0.01$), this relationship is given by

$$\eta = \frac{I_{dif}}{I_{pr}} \cong \left[\left(\frac{\ln 10\Delta A}{4\cos\theta_B}\right)^2 + \left(\frac{\pi d \Delta n}{\lambda_{pr}\cos\theta_B}\right)^2\right] \times \exp\left(-\frac{\ln 10 A}{\cos\theta_B}\right) \qquad [4]$$

where I_{dif} and I_{pr} are the diffracted and the probe intensity respectively, d is the sample thickness, and $A = 4\pi dK/(\lambda \ln 10)$ is the average absorbance.

The first and the second terms in the square bracket describe the contribution of the amplitude and phase gratings respectively, and the exponential term accounts for the reabsorption of the diffracted beam by the sample.

The main processes responsible for the variation of the optical properties of an isotropic dielectric material are summarized in Figure 2.

The modulation of the absorbance, ΔA, is essentially due to the photoinduced concentration change,

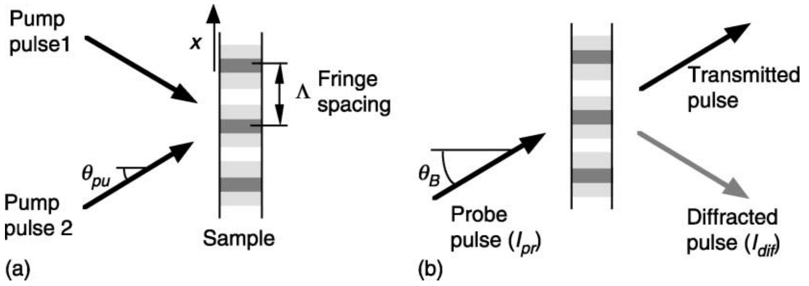
Pump
pulse1
x
Λ
Fringe
spacing
θ_{pu}
Pump
pulse 2
Sample
(a)
θ_B
Probe
pulse (I_{pr})
Transmitted
pulse
Diffracted
pulse (I_{dif})
(b)

Figure 1 Principle of the transient grating technique: (a) grating formation (pumping), (b) grating detection (probing).

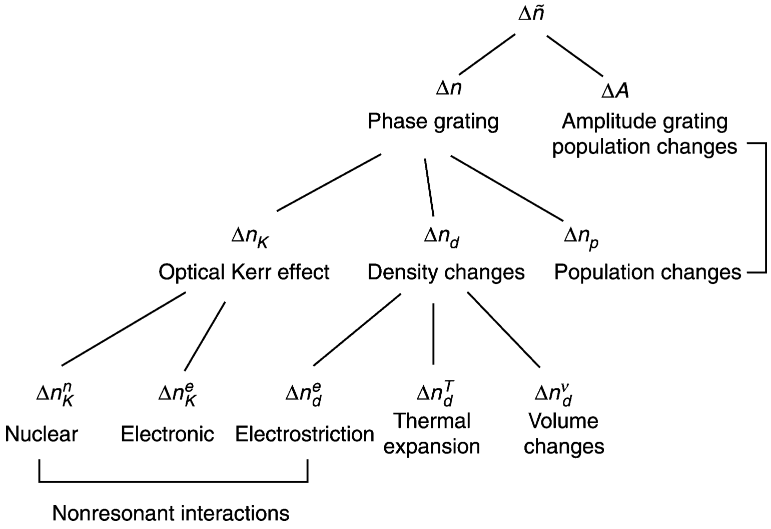
$\Delta\tilde{n}$
Δn
Phase grating
ΔA
Amplitude grating
population changes
Δn_K
Optical Kerr effect
Δn_d
Density changes
Δn_p
Population changes
Δn_K^n
Nuclear
Δn_K^e
Electronic
Δn_d^e
Electrostriction
Δn_d^T
Thermal
expansion
Δn_d^v
Volume
changes
Nonresonant interactions

Figure 2 Classification of the possible contributions to a transient grating signal.

ΔC, of the different chemical species i (excited state, photoproduct, ...):

$$\Delta A(\lambda_{pr}) = \sum_i \varepsilon_i(\lambda_{pr})\Delta C_i \qquad [5]$$

where ε_i is the absorption coefficient of the species i.

The variation of the refractive index, Δn, has several origins and can be expressed as

$$\Delta n = \Delta n_K + \Delta n_p + \Delta n_d \qquad [6]$$

Δn_K is the variation of refractive index due to the optical Kerr effect (OKE). This nonresonant interaction results in an electronic polarization (electronic OKE) and/or in a nuclear reorientation of the molecules (nuclear OKE) along the direction of the electric field associated with the pump pulses. As a consequence, a transient birefringence is created in the material. This effect is usually discussed within the framework of nonlinear optics in terms of intensity dependent refractive index or third order nonlinear susceptibility. Electronic OKE occurs in any dielectric material under sufficiently high light intensity. On the other hand, nuclear OKE is mostly observed in liquids and gases and depends strongly on the molecular shape.

Δn_p is the change of refractive index related to population changes. Its magnitude and wavelength dependence can be obtained by Kramers–Kronig transformation of $\Delta A(\lambda)$ or $\Delta K(\lambda)$:

$$\Delta n_p(\lambda) = \frac{1}{2\pi^2}\int_0^\infty \frac{\Delta K(\lambda')}{1-(\lambda'/\lambda)^2}\,\mathrm{d}\lambda' \qquad [7]$$

Δn_d is the change of refractive index associated with density changes. Density phase gratings can have essentially three origins:

$$\Delta n_d = \Delta n_d^t + \Delta n_d^v + \Delta n_d^e \qquad [8]$$

Δn_d^t is related to the temperature-induced change of density. If a fraction of the excitation energy is converted into heat, through a nonradiative transition or an exothermic process, the temperature becomes spatially modulated. This results in a variation of density, hence to a modulation of refractive index with amplitude Δn_d^t. Most of the temperature dependence of n originates from the density. The temperature-induced variation of n at constant density is much smaller than Δn_d^t.

Δn_d^v is related to the variation of volume upon population changes. This volume comprises not only the reactant and product molecules but also their environment. For example, in the case of a photodissociation, the volume of the product is larger than that of the reactant and a positive volume change can be expected. This will lead to a decrease of the density and to a negative Δn_d^v.

Finally, Δn_d^e is related to electrostriction in the sample by the electric field of the pump pulses. Like OKE, this is a nonresonant process that also contributes to the intensity dependent refractive index. Electrostriction leads to material compression in the regions of high electric field strength. The periodic compression is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$, where $\vec{i}$ is the unit vector along the modulation axis. The interference of these acoustic waves leads to a temporal modulation of Δn_d^e at the acoustic frequency v_{ac}, with $2\pi v_{ac} = k_{ac}v_s$, v_s being the speed of sound. As Δn_d^e oscillates between negative and positive values, the diffracted intensity, which is proportional to $(\Delta n_d^e)^2$, shows at temporal oscillation at twice the acoustic frequency. In most cases, Δn_d^e is weak and can be neglected if the pump pulses are within an absorption band of the sample.

The modulation amplitudes of absorbance and refractive index are not constant in time and their temporal behavior depends on various dynamic processes in the sample. The whole point of the transient grating techniques is precisely the measurement of the diffracted intensity as a function of time after excitation to deduce dynamic information on the system.

In the following, we will show that the various processes shown in Figure 2, that give rise to a diffracted signal, can in principle be separated by choosing appropriate experimental parameters, such as the timescale, the probe wavelength, the polarization of the four beams, or the crossing angle.

Experimental Setup

A typical experimental arrangement for pump–probe transient grating measurements is shown in Figure 3. The laser output pulses are split in three parts. Two parts of equal intensity are used as

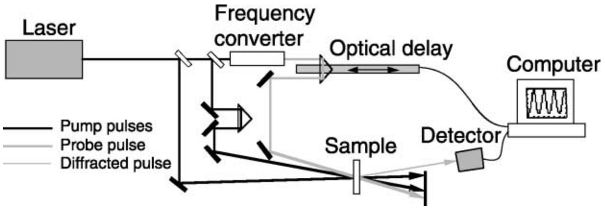

Figure 3 Schematic of a transient grating setup with pump–probe detection.

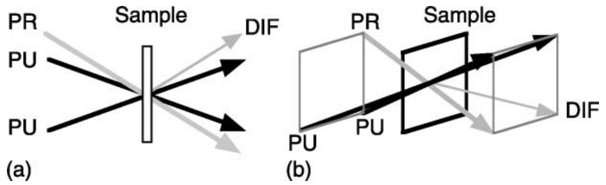

Figure 4 Beam geometry for transient grating: (a) in plane; (b) boxcars.

pump pulses and are crossed in the sample. In order to ensure time coincidence, one pump pulse travels along an adjustable optical delay line. The third part, which is used for probing, can be frequency converted using a dye laser, a nonlinear crystal, a Raman shifter, or white light continuum generation. The probe pulse is sent along a motorized optical delay line before striking the sample at the Bragg angle. There are several possible beam configurations for transient grating and the two most used are illustrated in Figure 4. When the probe and pump pulses are at different wavelengths, they can be in the same plane of incidence as shown in Figure 4a. However, if the pump and probe wavelengths are the same, the folded boxcars geometry shown in Figure 4b has to be used. The transient grating technique is background free and the diffracted signal propagates in a well-defined direction. In a pump–probe experiment, the diffracted signal intensity is measured as a function of the time delay between the pump and probe pulses. Such a setup can be used to probe dynamic processes occurring in timescales going from a few fs to a few ns, the time resolution depending essentially on the duration of the pump and probe pulses. For slower processes, the grating dynamics can be probed in real time with a *cw* laser beam and a fast photodetector.

Applications

The Transient Density Phase Grating Technique

If the grating is probed at a wavelength far from any absorption band, the variation of absorbance, $\Delta A(\lambda_{pr})$, is zero and the corresponding change of refractive index, $\Delta n_p(\lambda_{pr})$, is negligibly small. In this case, eqn [4] simplifies to

$$\eta \cong \left(\frac{\pi d}{\lambda_{pr}\cos\theta_B}\right)^2 \times \Delta n_d^2 \qquad [9]$$

In principle, the diffracted signal may also contain contributions from the optical Kerr effect, Δn_K, but we will assume here that this non-resonant and ultrafast response is negligibly small. The density change can originate from both heat releasing processes and volume differences between the products and reactants, the former contribution usually being much larger than the latter. Even if the heat releasing process is instantaneous, the risetime of the density grating is limited by thermal expansion. This expansion is accompanied by the generation of two counterpropagating acoustic waves with wave vectors, $\vec{k}_{ac} = \pm(2\pi/\Lambda)\vec{i}$. One can distinguish two density gratings: 1) a diffusive density grating, which reproduces the spatial distribution of temperature and which decays by thermal diffusion; and 2) an acoustic density grating originating from the standing acoustic wave and whose amplitude oscillates at the acoustic frequency v_{ac}.

The amplitudes of these two gratings are equal but of opposite sign. Consequently, the time dependence of the modulation amplitude of the density phase grating is given by

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V\right)\rho\left(\frac{\partial n}{\partial \rho}\right)R(t) \qquad [10a]$$

with

$$R(t) = 1 - \cos(2\pi v_{ac}t) \times \exp(-\alpha_{ac}v_s t) \qquad [10b]$$

where ρ, β, C_v, and α_{ac} are the density, the volume expansion coefficient, the heat capacity, and the acoustic attenuation constant of the medium, respectively, Q is the amount of heat deposited during the photoinduced process, and ΔV is the corresponding volume change. As the standing acoustic wave oscillates, its corresponding density grating interferes with the diffusive density grating. Therefore, the total modulation amplitude of the density and thus Δn_d exhibits the oscillation at v_{ac}. Figure 5 shows the time profile of the diffracted

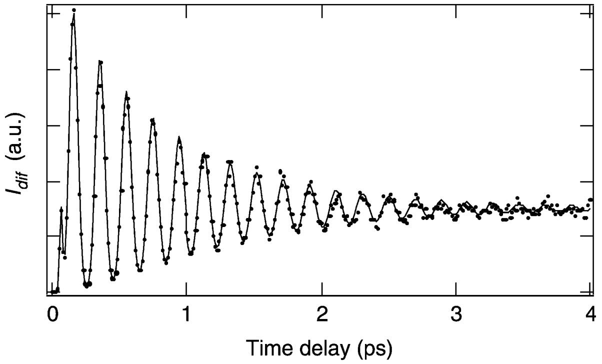

Figure 5 Time profile of the diffracted intensity measured with a solution of malachite green after excitation with two pulses with close to counterpropagating geometry.

intensity measured with a solution of malachite green. After excitation to the S_1 state, this dye relaxes nonradiatively to the ground state in a few ps. For this measurement, the sample solution was excited with two 30 ps laser pulses at 532 nm crossed with an angle close to 180°. The continuous line is the best fit of eqns [9] and [10]. The damping of the oscillation is due to acoustic attenuation. After complete damping, the remaining diffracted signal is due to the diffusive density grating only.

$R(t)$ can be considered as the response function of the sample to a prompt heat release and/or volume change. If these processes are not instantaneous compared to an acoustic period ($\tau_{ac} = \nu_{ac}^{-1}$), the acoustic waves are not created impulsively and the time dependence of Δn_d is

$$\Delta n_d(t) = \left(\frac{\beta Q}{\rho C_v} + \Delta V \right) \rho \left(\frac{\partial n}{\partial \rho} \right) F(t) \qquad [11a]$$

with

$$F(t) = \int_{-\infty}^{t} R(t - t') \cdot f(t') \mathrm{d}t' \qquad [11b]$$

where $f(t)$ is a normalized function describing the time evolution of the temperature and/or volume change. In many cases, $f(t) = \exp(-k_r t)$, k_r being the rate constant of the process responsible for the change. Figure 6 shows the time profile of the diffracted intensity calculated with eqns [9] and [11] for different values of k_r.

If several processes take place, the total change of refractive index is the sum of the changes due to the individual processes. In this case, Δn_d should be expressed as

$$\Delta n_d(t) = \sum_i \left(\frac{\beta Q_i}{\rho C_v} + \Delta V_i \right) \rho \left(\frac{\partial n}{\partial \rho} \right) F_i(t) \qquad [12]$$

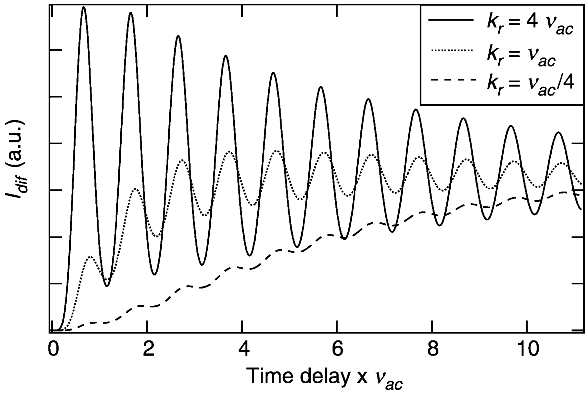

Figure 6 Time profiles of the diffracted intensity calculated using eqns [9] and [11] with various k_r values.

The separation of the thermal and volume contributions to the diffracted signal is problematic. Several approaches have been proposed, the most used being the measurement in a series of solvents with different expansion coefficient β. This method requires that the other solvent properties like refractive index, dielectric constant, or viscosity, are constant or that the energetics of the system investigated does not depend on them. This separation is easier when working with water because its β value vanishes at 4 °C. At this temperature, the density variations are due to the volume changes only.

The above equations describe the growth of the density phase grating. However, this grating is not permanent and decays through diffusive processes. The phase grating originating from thermal expansion decays via thermal diffusion with a rate constant k_{th} given by

$$k_{th} = D_{th} \left(\frac{2\pi}{\Lambda} \right)^2 \qquad [13]$$

where D_{th} is the thermal diffusivity. Table 1 shows k_{th} values in acetonitrile for different crossing angles.

The decay of the phase grating originating from volume changes depends on the dynamics of the population responsible for ΔV (vide infra).

Time resolved optical calorimetry

A major application of the density phase grating in chemistry is the investigation of the energetics of photo-induced processes. The great advantage of this technique over other optical calorimetric methods, like the thermal lens or the photoacoustic spectroscopy, is its superior time resolution. The time constant of the fastest heat releasing process that can be time resolved with this technique is of the order of the acoustic period. The shortest acoustic period is achieved when forming the grating with two counterpropagating pump pulses ($2\theta_{pu} = 180°$). In this case the fringe spacing is

Table 1 Fringe spacing, Λ, acoustic frequency, ν_{ac}, thermal diffusion rate constant, k_{tr}, for various crossing angles of the pump pulses, 2 θ_{pu}, at 355 nm and in acetonitrile

$2\ \theta_{pu}$	Λ (μm)	ν_{ac}(s^{-1})	k_{th} (s^{-1})
0.5°	40.7	3.2×10^7	4.7×10^3
50°	0.42	3.1×10^9	4.5×10^7
180°	0.13	9.7×10^9	4.7×10^8

$\Lambda = \lambda_{pu}/2n$ and, with UV pump pulses, an acoustic period of the order of 100 ps can be obtained in a typical organic solvent.

For example, Figure 7a shows the energy diagram for a photoinduced electron transfer (ET) reaction between benzophenone (BP) and an electron donor (D) in a polar solvent. Upon excitation at 355 nm, $^1BP^*$ undergoes intersystem-crossing (ISC) to $^3BP^*$ with a time constant of about 10 ps. After diffusional encounter with the electron donor, ET takes place and a pair of ions is generated. With this system, the whole energy of the 355 nm photon ($E = 3.49$ eV) is converted into heat. The different heat releasing processes can be differentiated according to their timescale. The vibrational relaxation to $^1BP^*$ and the ensuing ISC to $^3BP^*$ induces an ultrafast release of 0.49 eV as heat. With donor concentrations of the order of 0.1 M, the heat deposition process due to the electron transfer is typically in the ns range. Finally, the recombination of the ions produces a heat release in the microsecond timescale. Figure 7b shows the time profile of the diffracted intensity measured after excitation at 355 nm of BP with 0.05 M D in acetonitrile. The 30 ps pump pulses were crossed at 27° ($\tau_{ac} = 565$ ps) and the 10 ps probe pulses were at 590 nm. The oscillatory behavior is due to the ultrafast heat released upon formation of $^3BP^*$, while the slow dc rise is caused by the heat dissipated upon ET. As the amount of energy released in the ultrafast process is known, the energy released upon ET can be determined by comparing the amplitudes of the fast and slow components of the time profile. The energetics as well as the dynamics of the photoinduced ET process can thus be determined. Figure 7c shows the decay of the diffracted intensity due to the washing out of the grating by thermal diffusion. As both the acoustic frequency and the thermal diffusion depends on the grating wavevector, the experimental time window in which a heat releasing process can be measured, depends mainly on the crossing angle of the pump pulses as shown in Table 1.

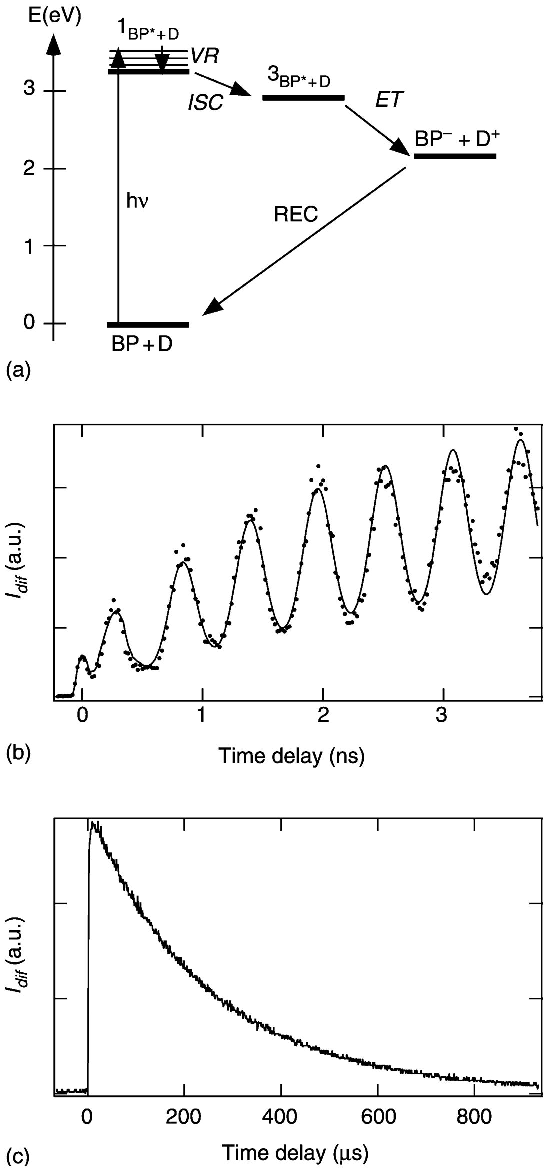

Figure 7 (a) Energy diagram of the states involved in the photoinduced electron transfer reaction between benzophenone (BP) and an electron donor (D) (VR: vibrational relaxation, ISC: intersystem crossing; ET: electron transfer; REC: charge recombination). (b) Time profile of the diffracted intensity measured at 590 nm after excitation at 355 nm of a solution of BP and 0.05 M D. (c) Same as (b) but measured at 1064 nm with a *cw* laser.

Determination of material properties

Another important application of the transient density phase grating technique is the investigation of material properties. Acoustics waves of various frequencies, depending on the pump wavelength and crossing angle, can be generated without physical contact with the sample. Therefore, acoustic properties, such as the speed of sound and the acoustic attenuation of the material, can be easily obtained from the frequency and the damping of the oscillation of the transient grating signal.

Similarly, the optoelastic constant of a material, $\rho\partial n/\partial\rho$, can be determined from the amplitude of the signal (see eqn [11]). This is done by comparing the signal amplitude of the material under investigation with that obtained with a known standard.

Finally, the thermal diffusivity, D_{th}, can be easily obtained from the decay of the thermal density phase grating, such as that shown in Figure 7c. This technique can be used with a large variety of bulk materials as well as films, surfaces and interfaces.

Investigation of Population Dynamics

The dynamic properties of a photogenerated species can be investigated by using a probe wavelength within its absorption or dispersion spectrum. As shown by eqn [4], either ΔA or Δn_p has to be different from zero. For this application, Δn_K and Δn_d should ideally be equal to zero. Unless working with ultrashort pulses (< 1 ps) and a weakly absorbing sample, Δn_K can be neglected. Practically, it is almost impossible to find a sample system for which some fraction of the energy absorbed as light is not released as heat. However, by using a sufficiently small crossing angle of the pump pulses, the formation of the density phase grating, which depends on the acoustic period, can take as much as 30 to 40 ns. In this case, Δn_d is negligible during the first few ns after excitation and the diffracted intensity is due to the population grating only:

$$I_{dif}(t) \propto \sum_i \Delta C_i^2(t) \qquad [14]$$

where ΔC_i is the modulation amplitude of the concentration of every species i whose absorption and/or dispersion spectrum overlaps with the probe wavelength.

The temporal variation of ΔC_i can be due either to the dynamics of the species i in the illuminated grating fringes or to processes taking place between the fringes, such as translational diffusion, excitation transport, and charge diffusion. In liquids, the decay of the population grating by translational diffusion is slow and occurs in the microsecond to millisecond timescale, depending on the fringe spacing. As thermal diffusion is typically hundred times faster, eqn [14] is again valid in this long timescale. Therefore if the population dynamics is very slow, the translational diffusion coefficient of a chemical species can be obtained by measuring the decay of the diffracted intensity as a function of the fringe spacing. This procedure has also been used to determine the temperature of flames. In this case however, the decay of the population grating by translational diffusion occurs typically in the sub-ns timescale.

In the condensed phase, these interfringe processes are of minor importance when measuring the diffracted intensity in the short timescale, i.e., before the formation of the density phase grating. In this case, the transient grating technique is similar to transient absorption, and it thus allows the measurement of population dynamics. However, because holographic detection is background free, it is at least a hundred times more sensitive than transient absorption.

The population gratings are usually probed with monochromatic laser pulses. This procedure is well suited for simple photoinduced processes, such as the decay of an excited state population to the ground state. For example, Figure 8 shows the decay of the diffracted intensity at 532 nm measured after excitation of a cyanine dye at the same wavelength. These dynamics correspond to the ground state recovery of the dye by non-radiative deactivation of the first singlet excited state. Because the diffracted intensity is proportional to the square of the concentration changes (see eqn [14]), its decay time is twice as short as the ground state recovery time. If the reaction is more complex, or if several intermediates are involved, the population grating has to be probed at several wavelengths. Instead of repeating many single wavelength experiments, it is preferable to perform multiplex transient grating. In this case, the grating is probed by white light pulses generated by focusing high intensity fs or ps laser pulses in a dispersive medium. If the crossing angle of the pump pulses is small enough ($< 1°$), the Bragg angle for probing is almost independent on the wavelength. A transient grating spectrum is obtained by dispersing the diffracted signal in a spectrograph. This spectrum consists in the sum of the square of the transient absorption and transient dispersion spectra. Practically, it is very similar to a transient absorption spectrum, but with a much superior signal to noise ratio. Figure 9 shows the transient grating spectra measured at various time delays after excitation of a solution of chloranil (CA) and

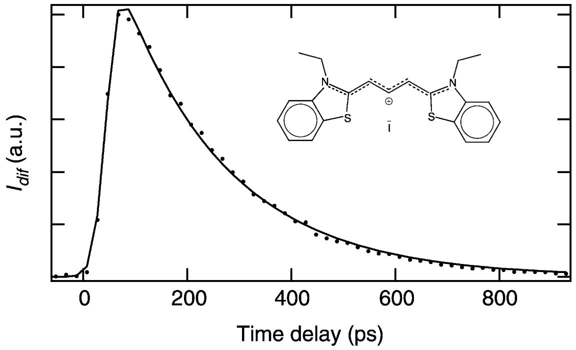

Figure 8 Time profile of the diffracted intensity at 532 nm measured after excitation at the same wavelength of a cyanine dye (inset) in solution and best fit assuming exponential ground state recovery.

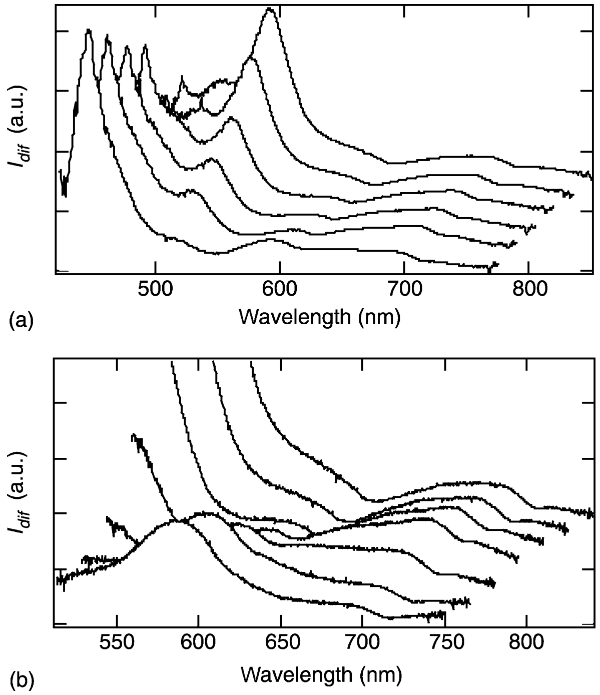

Figure 9 Transient grating spectra obtained at various time delays after excitation at 355 nm of a solution of chloranil and 0.25 M methylnaphthalene (a): from top to bottom: 60, 100, 180, 260, 330 and 600 ps; (b): from top to bottom: 100, 400, 750, 1100, 1500 ps).

methylnaphthalene (MNA) in acetonitrile. The reaction that takes place is

$$^3\mathrm{CA}^* \cdot \mathrm{MNA} \xrightarrow{\text{diffusion}} (^3\mathrm{CA}^* \cdot \mathrm{MNA}) \xrightarrow{\text{ET}}$$

$$^3(\mathrm{CA}^- \cdot \mathrm{MNA}^+) \xrightarrow{\text{MNA}} {}^3(\mathrm{CA}^- \cdot \mathrm{MNA}^+)$$

After its formation upon ultrafast ISC, $^3\mathrm{CA}^*$, which absorbs around 510 nm, decays upon ET with MNA to generate CA^-(450 nm) and MNA^+ (690 nm). The ion pair reacts further with a second MNA molecule to form the dimer cation (580 nm).

The time profile of the diffracted intensity reflects the population dynamics as long as these populations follow first- or pseudo-first order kinetics. Higher order kinetics leads to inharmonic gratings and in this case the time dependence of the diffracted intensity is no longer a direct measure of the population dynamics.

Polarization Selective Transient Grating

In the above applications, the selection between the different contributions to the diffracted signal was essentially made by choosing the probe wavelength and the crossing angle of the pump pulses. However, this approach is not always sufficient. Another important parameter is the polarization of the four waves involved in a transient grating experiment. For example, when measuring population dynamics, the polarization of the probe beam has to be at magic angle (54.7°) relatively to that of the pump pulses. This ensures that the observed time profiles are not distorted by the decay of the orientational anisotropy of the species created by the polarized pump pulses.

The magnitude of ΔA and Δn depends on the pump pulses intensity, and therefore the diffracted intensity can be expressed by using the formalism of nonlinear optics:

$$I_{dif}(t) = C\int_{-\infty}^{+\infty} \mathrm{d}t\, I_{pr}(t - t'') \times \left[\int_{-\infty}^{t} \mathrm{d}t' |R_{ijkl}(t'' - t')| I_{pu}(t')\right]^2 \qquad [15]$$

where C is a constant and R_{ijkl} is an element of the fourth rank tensor $\mathbf{R}$ describing the nonlinear response of the material to the applied optical fields. In isotropic media, this tensor has only 21 nonzero elements, which are related as follows:

$$R_{1111} = R_{1122} + R_{1212} + R_{1221} \qquad [16]$$

where the subscripts are the Cartesian coordinates. Going from right to left, they design the direction of polarization of the pump, probe, and signal pulses. The remaining elements can be obtained by permutation of these indices ($R_{1122} = R_{1133} = R_{2211} = \ldots$). In a conventional transient grating experiment, the two pump pulses are at the same frequency and are time coincident and therefore their indices can be interchanged. In this case, $R_{1212} = R_{1221}$ and the number of independent tensor elements is further reduced:

$$R_{1111} = R_{1122} + 2R_{1212} \qquad [17]$$

The tensor $\mathbf{R}$ can be decomposed into four tensors according to the origin of the sample response: population and density changes, electronic and nuclear optical Kerr effects:

$$\mathbf{R} = \mathbf{R}(p) + \mathbf{R}(d) + \mathbf{R}(K, e) + \mathbf{R}(K, n) \qquad [18]$$

As they describe different phenomena, these tensors do not have the same symmetry properties. Therefore, the various contributions to the diffracted intensity can, in some cases, be measured selectively by choosing the appropriate polarization of the four waves.

Table 2 shows the relative amplitude of the most important elements of these four tensors.

The tensor $\mathbf{R}(p)$ depends on the polarization anisotropy of the sample, r, created by excitation with polarized pump pulses:

$$r(t) = \frac{N_{\parallel}(t) - N_{\perp}(t)}{N_{\parallel}(t) + 2N_{\perp}(t)} = \frac{2}{5}\mathrm{P}_2[\cos(\gamma)]\exp(-k_{re}t) \qquad [19]$$

where $N_{\parallel}$ and $N_{\perp}$ are the number of molecules with the transition dipole oriented parallel and perpendicular to the polarization of the probe pulse, respectively, γ is the angle between the transition dipoles involved in the pump and probe processes, P_2 is the second Legendre polynomial, and k_{re} is the rate constant for the reorientation of the transition dipole, for example, by rotational diffusion of the molecule or by energy hopping.

The table shows that $R_{1212}(d) = 0$, i.e., the contribution of the density grating can be eliminated with the set of polarization $(0°, 90°, 0°, 90°)$, the so-called crossed grating geometry. In this geometry, the two pump pulses have orthogonal polarization, the polarization of the probe pulse is parallel to that of one pump pulse and the polarization component of the signal that is orthogonal to the polarization of the probe pulse is measured. In this geometry, $R_{1212}(p)$ is nonzero as long as there is some polarization anisotropy, ($r \neq 0$). In this case, the diffracted intensity is

$$I_{dif}(t) \propto |R_{1212}(p)|^2 \propto [\Delta C(t) \times r(t)]^2 \qquad [20]$$

The crossed grating technique can thus be used to investigate the reorientational dynamics of molecules, through $r(t)$, especially when the dynamics of r is faster than that of ΔC. For example, Figure 10 shows the time profile of the diffracted intensity measured in the crossed grating geometry with rhodamine 6G in ethanol. The decay is due to the reorientation of the molecule by rotational diffusion, the excited lifetime of rhodamine being about 4 ns.

Table 2 Relative value of the most important elements of the response tensors $\mathbf{R}(i)$ and polarization angle ζ of the signal beam, where the contribution of the corresponding process vanishes for the set of polarization $(\zeta, 45°, 0°, 0°)$.

Process	R_{1111}	R_{1122}	R_{1212}	ζ
Electronic OKE	1	1/3	1/3	−71.6°
Nuclear OKE	1	−1/2	3/4	63.4°
Density	1	1	0	−45°
Population:				
$\gamma = 0°$	1	1/3	1/3	−71.6°
$\gamma = 90°$	1	2	−1/2	−26.6°
No correlation: $r = 0$	1	1	0	−45°

On the other hand, if the decay of r is slower than that of ΔC, the crossed grating geometry can be used to measure the population dynamics without any interference from the density phase grating. For example, Figure 11 shows the time profile of the diffracted intensity after excitation of a suspension of TiO_2 particles in water. The upper profile was measured with the set of polarization $(0°, 0°, 0°, 0°)$ and thus reflects the time dependence of $R_{1111}(p)$ and $R_{1111}(d)$. $R_{1111}(p)$ is due to the trapped electron population, which decays by charge recombination, and $R_{1111}(d)$ is due to the heat dissipated upon both charge separation and recombination. The lower time profile was measured in the crossed grating geometry and thus reflects the time dependence of R_{1212} and in particular that of $R_{1212}(p)$.

Each contribution to the signal can be selectively eliminated by using the set of geometry $(\zeta, 45°, 0°, 0°)$ where the value of the 'magic angle' ζ for each

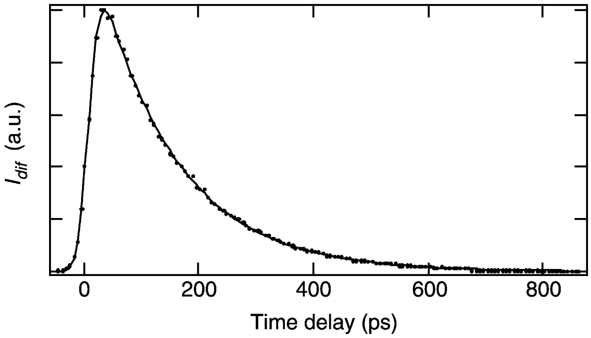

Figure 10 Time profile of the diffracted intensity measured with a solution of rhodamine 6G with crossed grating geometry.

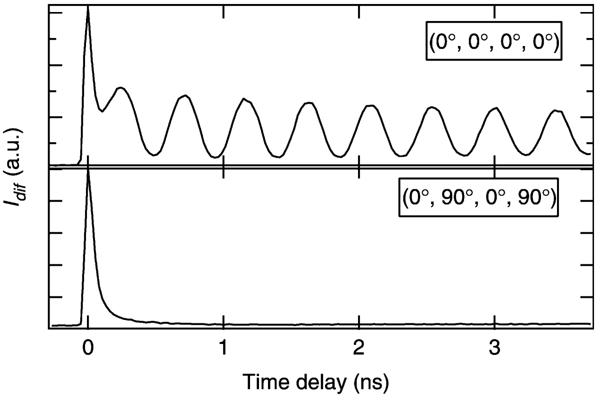

Figure 11 Time profiles of the diffracted intensity after excitation at 355 nm of a suspension of TiO_2 particles in water and using different set of polarization of the four beams.

contribution is listed in Table 2. This approach allows for example the nuclear and electronic contributions to the optical Kerr effect to be measured separately.

Concluding Remarks

The transient grating techniques offer a large variety of applications for investigating the dynamics of chemical processes. We have only discussed the cases where the pump pulses are time coincident and at the same wavelength. Excitation with pump pulses at different wavelengths results to a moving grating. The well-known CARS spectroscopy is such a moving grating technique. Finally, the three pulse photon echo can be considered as a special case of transient grating where the sample is excited by two pump pulses, which are at the same wavelength but are not time coincident.

List of Units and Nomenclature

C	concentration [mol L^{-1}]
C_v	heat capacity [J K^{-1} kg^{-1}]
d	sample thickness [m]
D_{th}	thermal diffusivity [m^2 s^{-1}]
I_{dif}	diffracted intensity [W m^{-2}]
I_{pr}	intensity of the probe pulse [W m^{-2}]
I_{pu}	intensity of a pump pulse [W m^{-2}]
k_{ac}	acoustic wavevector [m^{-1}]
k_r	rate constant of a heat releasing process [s^{-1}]
k_{re}	rate constant of reorientation [s^{-1}]
k_{th}	rate constant of thermal diffusion [s^{-1}]
K	attenuation constant
n	refractive index
$\tilde{n}$	complex refractive index
N	number of molecule per unit volume [m^{-3}]
r	polarization anisotropy
$\mathbf{R}$	fourth rank response tensor [m^2 V^{-2}]
v_s	speed of sound [m s^{-1}]
V	volume [m^3]
α_{ac}	acoustic attenuation constant [m^{-1}]
β	cubic volume expansion coefficient [K^{-1}]
γ	angle between transition dipoles
Λ	fringe spacing [m]
Δn_d	variation of refractive index due to density changes
Δn_d^e	Δn_d due to electrostriction
Δn_K	variation of refractive index due to optical Kerr effect
Δn_d^p	Δn_d due to volume changes
Δn_p	variation of refractive index due to population changes
Δn_d^t	Δn_d due to temperature changes
Δx	peak to null variation of the parameter x
ε	molar decadic absorption coefficient [cm L mol^{-1}]
ζ	angle of polarization of the diffracted signal
η	diffraction efficiency
θ_B	Bragg angle (angle of incidence of the probe pulse)
θ_{pu}	angle of incidence of a pump pulse
λ_{pr}	probe wavelength [m]
λ_{pu}	pump wavelength [m]
ρ	density [kg m^{-3}]
τ_{ac}	acoustic period [s]
υ_{ac}	acoustic frequency [s^{-1}]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Holography, Applications:** Holographic Recording Materials and their Processing. **Holography, Techniques:** Holographic Interferometry. **Materials Characterization Techniques:** $\chi^{(3)}$. **Nonlinear Optics Basics:** Four-Wave Mixing. **Ultrafast Technology:** Femtosecond Chemical Dynamics, Gas Phase; Femtosecond Condensed Phase Spectroscopy – Structural Dynamics.

Further Reading

Bräuchle C and Burland DM (1983) Holographic methods for the investigation of photochemical and photophysical properties of molecules. *Angewandte Chemie International Edition Engl* 22: 582–598.

Eichler HJ, Günter P and Pohl DW (1986) *Laser-Induced Dynamics Gratings*. Berlin: Springer.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford: Oxford University Press.

Fourkas JT and Fayer MD (1992) The transient grating: a holographic window to dynamic processes. *Accounts of Chemical Research* 25: 227–233.

Hall G and Whitaker BJ (1994) Laser-induced grating spectroscopy. *Journal of the Chemistry Society Faraday Trans* 90: 1–16.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*, revised edn. Boston: Academic Press.

Mukamel S (1995) *Nonlinear Optical Spectroscopy*. Oxford: Oxford University Press.

Rullière C (ed.) (1998) *Femtosecond Laser Pulses*. Berlin: Springer.

Terazima M (1998) Photothermal studies of photophysical and photochemical processes by the transient grating method. *Advances in Photochemistry* 24: 255–338.

cross-sectional area 10^4 times and dispersed the beam across it. The CPA technique makes it possible to use conventional laser amplifiers and to stay below the onset of nonlinear effects.

See also

Diffraction: Diffraction Gratings. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

COHERENCE

Contents

Overview
Coherence and Imaging
Speckle and Coherence

Overview

A Sharma and A K Ghatak, Indian Institute of Technology, New Delhi, India

H C Kandpal, National Physical Laboratory, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherence refers to the characteristic of a wave that indicates whether different parts of the wave oscillate in tandem, or in a definite phase relationship. In other words, it refers to the degree of confidence by which one can predict the amplitude and phase of a wave at a point, from the knowledge of these at another point at the same or a different instant of time. Emission from a thermal source, such as a light bulb, is a highly disordered process and the emitted light is incoherent. A well-stabilized laser source, on the other hand, generates light in a highly ordered manner and the emitted light is highly coherent. Incoherent and coherent light represent two extreme cases. While describing the phenomena of physical optics and diffraction theory, light is assumed to be perfectly coherent in both spatial as well as temporal senses, whereas in radiometry it is generally assumed to be incoherent. However, practical light sources and the fields generated by them are in between the two extremes and are termed as partially coherent sources and fields. The degree of order that exists in an optical field produced by a source of any kind may be described in terms of various correlation functions. These correlation functions are the basic theoretical tools for the analyses of statistical properties of partially coherent light fields.

Light fields generated from real physical sources fluctuate randomly to some extent. On a microscopic level quantum mechanical fluctuations produce randomness and on macroscopic level the randomness occurs as a consequence of these microscopic fluctuations, even in free space. In real physical sources, spontaneous emission causes random fluctuations and even in the case of lasers, spontaneous emission cannot be suppressed completely. In addition to spontaneous emission, there are many other processes that give rise to random fluctuations of light fields. Optical coherence theory was developed to describe the random nature of light and it deals with the statistical similarity between light fluctuations at two (or more) space–time points.

As mentioned earlier, in developing the theory of interference or diffraction, light is assumed to be perfectly coherent, or, in other words, it is taken to be monochromatic and sinusoidal for all times. This is, however, an idealization since the wave is obviously generated at some point of time by an atomic or molecular transition. Furthermore, a wavetrain generated by such a transition is of a finite duration, which is related to the finite lifetime of the atomic or molecular levels involved in the transition. Thus, any wave emanating from a source is an ensemble of a large number of such wavetrains of finite duration, say τ_c. A simplified visualization of such an ensemble is shown in Figure 1 where a wave is shown as a series of wavetrains of duration τ_c. It is evident from the figure that the fields at time t and $t + \Delta t$ will have a definite phase relationship if $\Delta t \ll \tau_c$ and will have no or negligible phase relationship when $\Delta t \gg \tau_c$. The time τ_c is known as the coherence time of the

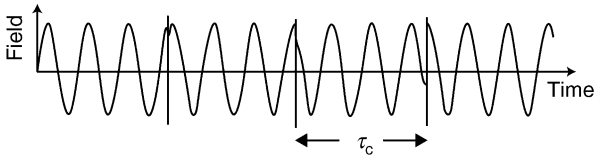

Figure 1 Typical variation of the radiation field with time. The coherence time $\sim \tau_c$.

radiation and the field is said to remain coherent for time $\sim \tau_c$. This property of waves is referred to as the time coherence or the temporal coherence and is related to the spectral purity of the radiation. If one obtains the spectrum of the wave shown in Figure 1 by taking the Fourier transform of the time variation, it would have a width of $\Delta\nu$ around ν_0 which is the frequency of the sinusoidal variation of individual wavetrains. The spectral width $\Delta\nu$ is related to the coherence time as

$$\Delta\nu \sim 1/\tau_c \qquad [1]$$

For thermal sources such as a sodium lamp, $\tau_c \sim 100$ ps, whereas for a laser beam it could be as large as a few milliseconds. A related quantity is the coherence length l_c, which is the distance covered by the wave in time τ_c,

$$l_c = c\tau_c \sim \frac{c}{\Delta\nu} = \frac{\lambda_0^2}{\Delta\lambda} \qquad [2]$$

where λ_0 is the central wavelength ($\lambda_0 = c/\nu_0$) and $\Delta\lambda$ is the wavelength spread corresponding to the spectral width $\Delta\nu$. In a two-beam interference experiment (e.g., Michelson interferometer, Figure 2), the interfering beam derived from the same source will produce good interference fringes if the path difference between the two interfering beams is less than the coherence length of the radiation given out by the source.

It must be added here that for the real fields, generated by innumerable atoms or molecules, the individual wavetrains have different lengths around an average value τ_c. Furthermore, several wavetrains in general are propagating simultaneously, overlapping in space and time, to produce an ensemble whose properties are best understood in a statistical sense as we shall see in later sections.

Another type of coherence associated with the fields is the space coherence or the spatial coherence, which is related to the size of the source of radiation. It is evident that when the source is an ideal point source, the field at any two points (within the coherence length) would have definite phase relationship. However, the field from a thermal source of finite area S can be thought as resultant of the fields from each point on the source. Since each point source would usually be independent of the other, the phase relationship between the fields at any two points would depend on their position and size of the source. It can be shown that two points will have a strong phase relationship if they lie within the solid angle $\Delta\Omega$ from the source (Figure 3) such that

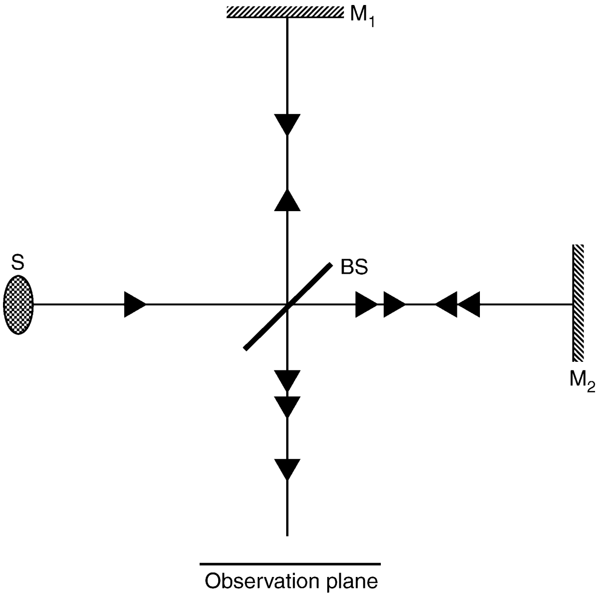

Figure 2 Michelson interferometer to study the temporal coherence of radiation from source S; M_1 and M_2 are mirrors and BS is a 50%–50% beamsplitter.

$$\Delta\Omega \sim \frac{\lambda_0^2}{S} \qquad [3]$$

Thus, on a plane R distance away from the source, one can define an area $A_c = R^2\Delta\Omega$ over which the field remains spatially coherent. This area A_c is called the coherence area of the radiation and its square root is sometimes referred to as the transverse coherence length. It is trivial to show that

$$A_c \sim \frac{\lambda_0^2}{\Delta\Omega_S} \qquad [4]$$

where $\Delta\Omega_S$ is the solid angle subtended by the source on the plane at which the coherence area is defined. In Young's double-hole experiment, if the two pinholes lie within the coherence area of the radiation from the primary source, good contrast in fringes would be observed.

Combining the concepts of coherence length and the coherence area, one can define a coherence volume as $V_c = A_c l_c$. For the wavefield from a thermal source, this volume represents that portion of the space in which the field is coherent and any

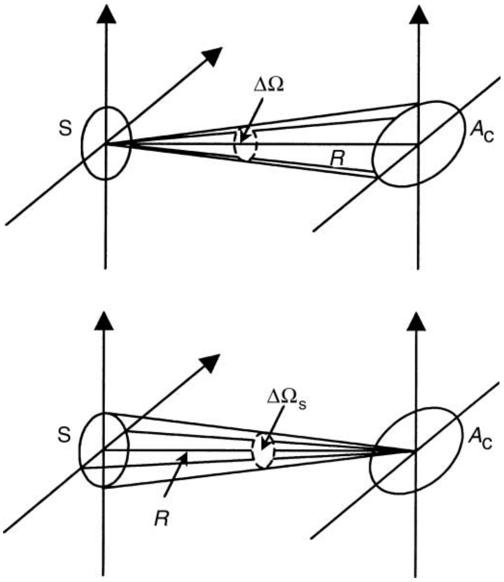

Figure 3 Spatial coherence and coherence area.

interference produced using the radiation from points within this volume will produce fringes of good contrast.

Mathematical Description of Coherence

Analytical Field Representation

Coherence properties associated with fields are best analyzed in terms of complex representation for optical fields. Let the real function $V^{(r)}(\mathbf{r}, t)$ represent the scalar field, which could be one of the transverse Cartesian components of the electric field associated with the electromagnetic wave. One can then define a complex analytical signal $V(\mathbf{r}, t)$ such that

$$\begin{aligned} V^{(r)}(\mathbf{r}, t) &= \mathrm{Re}[V(\mathbf{r}, t)], \\ V(\mathbf{r}, t) &= \int_0^\infty \tilde{V}^{(r)}(\mathbf{r}, \nu)\mathrm{e}^{-2\pi i\nu t}\,\mathrm{d}\nu \end{aligned} \qquad [5]$$

where the spectrum $\tilde{V}^{(r)}(\mathbf{r}, \nu)$ is the Fourier transform of the scalar field $V^{(r)}(\mathbf{r}, t)$ and the spectrum for negative frequencies has been suppressed as it does not give any new information since $\tilde{V}^{(r)^*}(\mathbf{r}, \nu) = \tilde{V}^{(r)}(\mathbf{r}, -\nu)$.

In general, the radiation from a quasi-monochromatic thermal source fluctuates randomly as it is made of a large number of mutually independent contributions from individual atoms or molecules in the source. The field from such a source can be regarded as an ensemble of a large number of randomly different analytical signals such as $V(\mathbf{r}, t)$. In other words, $V(\mathbf{r}, t)$ is a typical member of an ensemble, which is the result of a random process representing the radiation from a quasi-monochromatic source. This process is assumed to be stationary and ergodic so that the ensemble average is equal to the time average of a typical member of the ensemble and that the origin of time is unimportant. Thus, the quantities of interest in the theory of coherence are defined as time averages:

$$\langle f(t)\rangle = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f(t)\mathrm{d}t \qquad [6]$$

Mutual Coherence

In order to define the mutual coherence, the key concept in the theory of coherence, we consider Young's interference experiment, as shown in Figure 4, where the radiation from a broad source of size S is illuminating a screen with two pinholes P_1 and P_2. The light emerging from the two pinholes produces interference, which is observed on another screen at a distance R from the first screen. The field at point P, due to the pinholes, would be $K_1V(\mathbf{r}_1, t - t_1)$ and $K_2V(\mathbf{r}_2, t - t_2)$ respectively, where $\mathbf{r}_1$ and $\mathbf{r}_2$ define the positions of P_1 and P_2, t_1 and t_2 are the times taken by the light to travel to P from P_1 and P_2, and K_1 and K_2 are imaginary constants that depend on the geometry and size of the respective pinhole and its distance from the point P. Thus, the resultant field at point P would be given by

$$V(\mathbf{r}, t) = K_1V(\mathbf{r}_1, t - t_1) + K_2V(\mathbf{r}_2, t - t_2) \qquad [7]$$

Since the optical periods are extremely small as compared to the response time of a detector, the detector placed at point P will record only the time-averaged intensity:

$$I(P) = \langle V^*(\mathbf{r}, t)V(\mathbf{r}, t)\rangle \qquad [8]$$

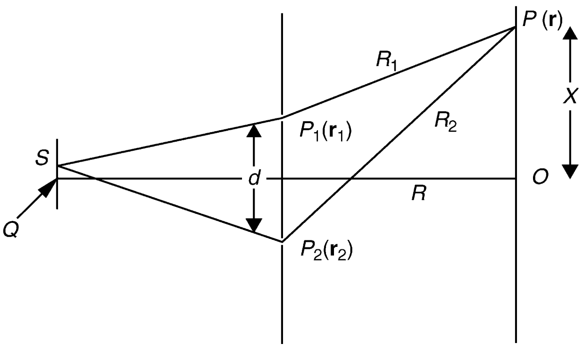

Figure 4 Schematic of Young's double-hole experiment.

where some constants have been ignored. With eqn [7], the intensity at point P would be given by

$$I(P) = I_1 + I_2 + 2\,\mathrm{Re}\left\{K_1 K_2^* \Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2)\right\} \quad [9]$$

where I_1 and I_2 are the intensities at point P, respectively, due to radiations from pinholes P_1 and P_2 independently (defined as $I_i = \langle |K_i V(\mathbf{r}_i, t)|^2 \rangle$) and

$$\begin{aligned}\Gamma(\mathbf{r}_1, \mathbf{r}_2, t - t_1, t - t_2) &\equiv \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\\ &= \left\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau)\right\rangle \quad [10]\end{aligned}$$

is the mutual coherence function of the fields at P_1 and P_2, and it depends on the time difference $\tau = t_2 - t_1$, since the random process is assumed to be stationary. This function is also sometimes denoted by $\Gamma_{12}(\tau)$. The function $\Gamma_{ii}(\tau) \equiv \Gamma(\mathbf{r}_i, \mathbf{r}_i, \tau)$ defines the self-coherence of the light from the pinhole at P_i, and $|K_i|^2 \Gamma_{ii}(0)$ defines the intensity I_i at point P, due to the light from pinhole P_i.

Degree of Coherence and the Visibility of Interference Fringes

One can define a normalized form of the mutual coherence function, namely:

$$\begin{aligned}\gamma_{12}(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) &= \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}}\\ &= \frac{\langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t + \tau)\rangle}{\left[\left\langle |V(\mathbf{r}_1, t)|^2\right\rangle \left\langle |V(\mathbf{r}_2, t)|^2\right\rangle\right]^{1/2}} \quad [11]\end{aligned}$$

which is called the complex degree of coherence. It can be shown that $0 \le |\gamma_{12}(\tau)| \le 1$. The intensity at point P, given by eqn [9], can now be written as

$$I(P) = I_1 + I_2 + 2\sqrt{I_1 I_2}\,\mathrm{Re}\left[\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\right] \quad [12]$$

Expressing $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ as

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \exp\left[\mathrm{i}\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi \mathrm{i} \nu_0 \tau\right] \quad [13]$$

where $\alpha(\vec{r}_1, \vec{r}_2, \tau) = \arg[\gamma(\vec{r}_1, \vec{r}_2, \tau)] + 2\pi\nu_0\tau$, and ν_0 is the mean frequency of the light, the intensity in eqn [12] can be written as

$$\begin{aligned}I(P) = {} & I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|\\ & \times \cos\left[\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau) - 2\pi\nu_0\tau\right] \quad [14]\end{aligned}$$

Now, if we assume that the source is quasi-monochromatic, i.e., its spectral width $\Delta\nu \ll \nu_0$, the quantities $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ and $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$ vary slowly on the observation screen, and the interference fringes are mainly obtained due to the cosine term. Thus, defining the visibility of fringes as $v = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$, we obtain

$$v = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| \quad [15]$$

which shows that for maximum visibility, the two interfering fields must be completely coherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$). On the other hand, if the fields are completely incoherent ($|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 0$), no fringes are observed ($I_{\max} = I_{\min}$). The fields are said to be partially coherent when $0 < |\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| < 1$. When $I_1 = I_2$, the visibility is the same as $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)|$. The relation in eqn [15] shows that in an interference experiment, one can obtain the modulus of the complex degree of coherence by measuring I_1, I_2, and the visibility. Similarly, eqn [14] shows that from the measurement the positions of maxima, one can obtain the phase of the complex degree of coherence, $\alpha(\mathbf{r}_1, \mathbf{r}_2, \tau)$.

Temporal and Spatial Coherence

If the source illuminating the pinholes is a point source of finite spectral width situated at point Q, the fields at point P_1 and P_2 (Figure 4) at any given instant are the same and the mutual coherence function becomes

$$\begin{aligned}\Gamma_{11}(\tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_1, \tau) &= \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_1, t + \tau)\rangle\\ &= \langle V^*(\mathbf{r}_2, t) V(\mathbf{r}_2, t + \tau)\rangle = \Gamma_{22}(\tau) \quad [16]\end{aligned}$$

The self-coherence, $\Gamma_{11}(\tau)$, of the light from pinhole P_1, is a direct measure of the temporal coherence of the source. On the other hand, if the source is of finite size and we observe the interference at point O which corresponds to $\tau = 0$, the mutual coherence function would be

$$\Gamma_{12}(0) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t)\rangle \equiv J_{12} \quad [17]$$

which is called the mutual intensity and is a direct measure of the spatial coherence of the source. In general, however, the function $\Gamma_{12}(\tau)$ measures, for a source of finite size and spectral width, a combination of the temporal and spatial coherence, and in only some limiting cases, the two types of coherence can be separated.

Spectral Representation of Mutual Coherence

One can also analyze the correlation between two fields in the spectral domain. In particular, one can define the cross-spectral density function $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ which defines the correlation between the amplitudes of the spectral components of frequency ν of the light at the points P_1 and P_2. Thus:

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu)\delta(\nu - \nu') = \langle V^*(\mathbf{r}_1, \nu)V(\mathbf{r}_2, \nu')\rangle \qquad [18]$$

Using the generalized Wiener–Khintchine theorem, the cross-spectral density function can be shown to be the Fourier transform of the mutual coherence function:

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu)\mathrm{e}^{-2\pi \mathrm{i}\nu t}\,\mathrm{d}\nu \qquad [19]$$

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^\infty \Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)\mathrm{e}^{2\pi \mathrm{i}\nu t}\,\mathrm{d}\tau \qquad [20]$$

If the two points P_1 and P_2 coincide (i.e., there is only one pinhole), the cross-spectral density function reduces to the spectral density function of the light, which we denote by $S(\mathbf{r}, \nu)[\equiv W(\mathbf{r}, \mathbf{r}, \nu)]$. Thus, it follows from eqn [20] that the spectral density of light is the inverse Fourier transform of the self-coherence function. This leads to the Fourier transform spectroscopy, as we shall see later. One can also define spectral degree of coherence at frequency ν as

$$\begin{aligned}\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{W(\mathbf{r}_1, \mathbf{r}_1, \nu)W(\mathbf{r}_2, \mathbf{r}_2, \nu)}} \\ &= \frac{W(\mathbf{r}_1, \mathbf{r}_2, \nu)}{\sqrt{S(\mathbf{r}_1, \nu)S(\mathbf{r}_2, \nu)}}\end{aligned} \qquad [21]$$

It is easy to see that $0 \leq |\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)| \leq 1$. It is also sometimes referred to as the complex degree of coherence at frequency ν. It may be noted here that in the literature the notation $G^{(1,1)}(\nu) \equiv G(\mathbf{r}_1, \mathbf{r}_2, \gamma)$ has also been used for $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$.

Propagation of Coherence

The van Cittert–Zernike Theorem

Perfectly coherent waves propagate through diffraction formulae, which have been discussed in this volume elsewhere. However, incoherent and partially coherent waves would evidently propagate somewhat differently. The earliest treatment of propagation noncoherent light is due to van Cittert, which was later generalized by Zernike to obtain what is now the van Cittert–Zernike theorem. The theorem deals with the correlations developed between fields at two points, which have been generated by a quasi-monochromatic and spatially incoherent planar source. Thus, as shown in Figure 5, we consider a quasi-monochromatic ($\Delta\nu \ll \nu_0$) planar source σ, which has an intensity distribution $I(\mathbf{r}')$ on its plane and is spatially incoherent, i.e., there is no correlation between the fields at any two points on the source.

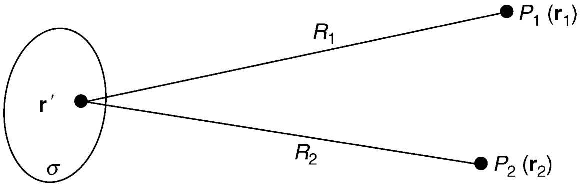

Figure 5 Geometry for the van Cittert–Zernike theorem.

The field, due to this source, would develop finite correlations after propagation, and the theorem states that

$$\begin{aligned}\gamma(\mathbf{r}_1, \mathbf{r}_2, 0) &= \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{12}(0)}} \\ &= \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}} \iint_\sigma I(\mathbf{r}') \frac{\mathrm{e}^{2\pi \mathrm{i}\nu_0(R_2 - R_1)/c}}{R_1 R_2}\mathrm{d}^2 r'\end{aligned} \qquad [22]$$

where $R_i = |\mathbf{r}_i - \mathbf{r}'|$ and $I(\mathbf{r}_i)$ is the intensity at point P_i. This relation is similar to the diffraction pattern produced at point P_1 due to a wave, with a spherical wavefront converging towards P_2 and with an amplitude distribution $I(\mathbf{r}')$ when it is diffracted by an aperture of the same shape, size, and the intensity distribution as those of the incoherent source σ. Thus, the theorem shows that a radiation, which was incoherent at the source, becomes partially coherent as it propagates.

Generalized Propagation

The expression $\mathrm{e}^{-2\pi \mathrm{i}\nu_0 R_1/c}/R_1$ can be interpreted as the field obtained at P_1 due to a point source located at the point $\mathbf{r}'$ on the planar source. Thus, this expression is simply the point spread function of the homogeneous space between the source and the observation plane containing the points P_1 and P_2. Hopkins generalized this to include any linear optical system characterized by a point spread function $h(\mathbf{r}_j - \mathbf{r}')$ and obtained the formula for the complex degree of coherence of a radiation emerging from an incoherent, quasi-monochromatic planar source after it has propagated through such a linear optical system:

$$\gamma(\mathbf{r}_1,\mathbf{r}_2,0) = \frac{1}{\sqrt{I(\mathbf{r}_1)I(\mathbf{r}_2)}} \iint_\sigma I(\mathbf{r}')h(\mathbf{r}_1 - \mathbf{r}')h^*(\mathbf{r}_2 - \mathbf{r}')\mathrm{d}^2 r' \qquad [23]$$

It would thus seem that the correlations propagate in much the same way, as does the field. Indeed, Wolf showed that the mutual correlation function

$\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$ satisfies the wave equations:

$$\nabla_1^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)$$
$$\nabla_2^2\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau)=\frac{1}{c^2}\frac{\partial^2}{\partial\tau^2}\Gamma(\mathbf{r}_1,\mathbf{r}_2,\tau) \qquad [24]$$

where ∇_j^2 is the Laplacian with respect to the point $\mathbf{r}_i$. Here the first of eqns [24], for instance, gives the variation of the mutual coherence function with respect to $\mathbf{r}_1$ and τ for fixed $\mathbf{r}_2$. Further, the variable τ is the time difference defined through the path difference, since $\tau=(R_2-R_1)/c$, and the actual time does not affect the mutual coherence function (as the fields are assumed to be stationary).

Using the relation in eqn [20], one can also obtain from eqn [24], the propagation equations for the cross-spectral density function $W(\mathbf{r}_1,\mathbf{r}_2,\nu)$:

$$\nabla_1^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0$$
$$\nabla_2^2 W(\mathbf{r}_1,\mathbf{r}_2,\nu)+k^2W(\mathbf{r}_1,\mathbf{r}_2,\nu)=0 \qquad [25]$$

where $k=2\pi\nu/c$.

Thompson and Wolf Experiment

One of the most elegant experiments for studying various aspects of coherence theory was carried out by Thompson and Wolf by making slight modifications in the Young's double-hole experiment. The experimental set-up shown in Figure 6 consists of a quasi-monochromatic broad incoherent source S of diameter $2a$. This was obtained by focusing filtered narrow band light from a mercury lamp (not shown in the figure) on to a hole of size $2a$ in an opaque screen A. A mask consisting of two pinholes, each of diameter $2b$ with their axes separated by a distance d, was placed symmetrically about the optical axis of the experimental setup at plane B between two lenses L_1 and L_2, each having focal length f. The source was at the front focal plane of the lens L_1 and the observations were made at the plane C at the back focus of the lens L_2. The separation d was varied to study the spatial coherence function on the mask plane by measuring the visibility and the phase of the fringes formed in plane C.

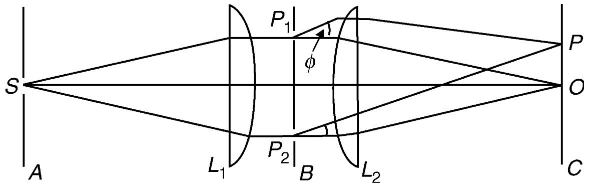

Figure 6 Schematic of the Thompson and Wolf experiment.

Using the van Cittert–Zernike theorem, the complex degree of coherence at the pinholes P_1 and P_2 on the mask after the lens L_1 is

$$\gamma_{12}=|\gamma_{12}|e^{i\beta_{12}}=\frac{2J_1(v)}{v}\text{ with }v=\left(\frac{2\pi\nu ad}{cf}\right) \qquad [26]$$

for symmetrically placed pinholes about the optical axis. Here β_{12} is the phase of the complex degree of coherence and in this special case where the two holes are equidistant from the axis, β_{12} is either zero or π, respectively, for positive or negative values of $2J_1(v)/v$. Let us assume that the intensities at two pinholes P_1 and P_2 are equal. The interference pattern observed at the back focal plane of lens L_2 is due to the superposition of the light diffracted from the pinholes. The beams are partially coherent with degree of coherence given by eqn [26]. Since the pinholes are symmetrically placed, the intensity due to either of the pinholes at a point P at the back focal plane of the lens L_2 is the same and is given by the Fraunhofer formula for diffraction from circular apertures, i.e.:

$$I_1(P)=I_2(P)=\left|\frac{2J_1(u)}{u}\right|^2,\text{ with }u=\frac{2\pi\nu}{c}b\sin\phi \qquad [27]$$

and ϕ is the angle that the diffracted beam makes from normal to the plane of the pinholes. The intensity of the interference pattern produced at the back focal plane of the lens L_2 is:

$$I(P)=2I_1(P)\left[1+\left|\frac{2J_1(v)}{v}\right|^2\cos(\delta+\beta_{12})\right] \qquad [28]$$

where $\delta=\mathrm{d}\sin\phi$ is the phase difference between the two beams reaching P from P_1 and P_2. For the on-axis point O, the quantity δ is zero.

The maximum and minimum values of $I(P)$ are given by

$$I_{\max}(P)=2I_1(P)\left[1+|2J_1(v)/v|^2\right] \qquad [29(a)]$$

$$I_{\max}(P)=2I_1(P)\left[1-|2J_1(v)/v|^2\right] \qquad [29(b)]$$

Figure 7 shows an example of the observed fringe patterns obtained by Thompson (1958) in a subsequent experiment.

Types of Fields

As mentioned above, $\gamma_{12}(\tau)$ is a measure of the correlation of the complex field at any two points P_1

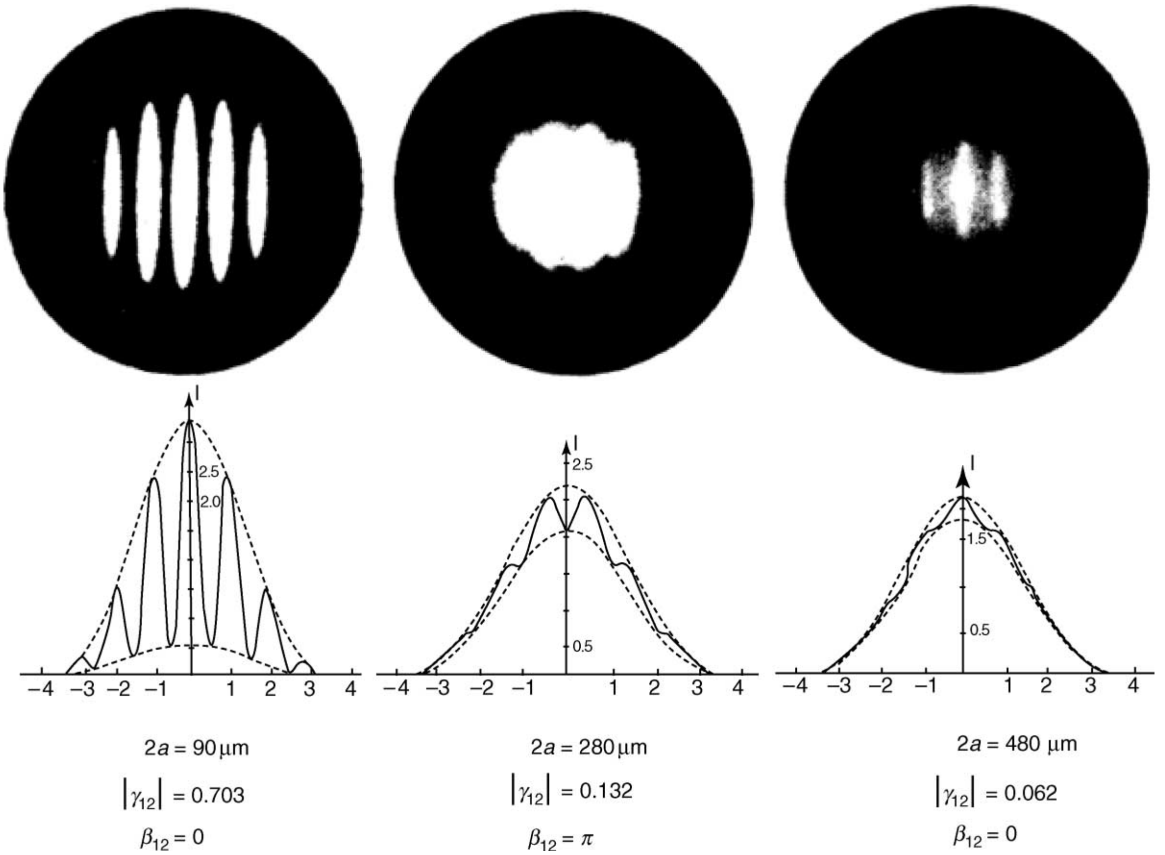

Figure 7 Two beam interference patterns obtained by using partially coherent light. The wavelength of light used was 0.579 µm and the separation *d* of the pinholes in the screen at B was 0.5 cm. The figure shows the observed fringe pattern and the calculated intensity variation for three sizes of the secondary source. The dashed lines show the maximum and minimum intensity. The values of the diameter, 2*a*, of the source and the corresponding values of the magnitude, $|\gamma_{12}|$, and the phase, β_{12}, are also shown in each case. Reproduced with permission from Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

and P_2 at specific time delay τ. Extending this definition, an optical field may be coherent or incoherent, if $|\gamma_{12}(\tau)| = 1$ or $|\gamma_{12}(\tau)| = 0$, respectively, for all pairs of points in the field and for all time delays. In the following, we consider some specific cases of fields and their properties.

Perfectly Coherent Fields

A field would be termed as perfectly coherent or self-coherent at a fixed point, if it has the property that $|\gamma(\mathbf{R}, \mathbf{R}, \tau)| = 1$ at some specific point $\mathbf{R}$ in the domain of the field for all values of time delay τ. It can also be shown that $\gamma(\mathbf{R}, \mathbf{R}, \tau)$ is periodic in time, i.e.:

$$\gamma(\mathbf{R}, \mathbf{R}, \tau) = \mathrm{e}^{-2\pi \mathrm{i}\nu_0\tau} \quad \text{for} \quad \nu_0 > 0 \qquad [30]$$

For any other point $\mathbf{r}$ in the domain of the field, $\gamma(\mathbf{R}, \mathbf{r}, \tau)$ and $\gamma(\mathbf{r}, \mathbf{R}, \tau)$ are also unimodular and also periodic in time.

A perfectly coherent optical field at two fixed points has the property that $|\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau)| = 1$ for any two fixed points $\mathbf{R}_1$ and $\mathbf{R}_2$ in the domain of the field for all values of time delay τ. For such a field, it can be shown that $\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \exp[\mathrm{i}(\beta - 2\pi\nu_0\tau)]$ with ν_0 (> 0) and β are real constants. Further, as a corollary, $|\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau)| = 1$ and $|\gamma(\mathbf{R}_2, \mathbf{R}_2, \tau)| = 1$ for all τ, i.e., the field is self-coherent at each of the field points $\mathbf{R}_1$ and $\mathbf{R}_2$, and

$$\gamma(\mathbf{R}_1, \mathbf{R}_1, \tau) = \gamma(\mathbf{R}_2, \mathbf{R}_2, \tau) = \exp(-2\pi \mathrm{i}\nu_0\tau) \text{ for } \nu_0 > 0$$

It can be shown that for any other point $\mathbf{r}'$ within the field

$$\begin{aligned} &\gamma(\mathbf{r}, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\gamma(\mathbf{r}, \mathbf{R}_1, \tau) = \gamma(\mathbf{r}, \mathbf{R}_1, \tau)\mathrm{e}^{\mathrm{i}\beta} \\ &\gamma(\mathbf{R}_1, \mathbf{R}_2, \tau) = \gamma(\mathbf{R}_1, \mathbf{R}_2, 0)\exp[-2\pi \mathrm{i}\nu_0\tau] \end{aligned} \qquad [31]$$

A perfectly coherent optical field for all points in the domain of the field has the property that $|\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)| = 1$ for all pairs of points $\mathbf{r}_1$ and $\mathbf{r}_2$ in the domain of the field, for all values of time delay τ.

It can then be shown that $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \exp\{i[\alpha(\mathbf{r}_1) - \alpha(\mathbf{r}_2)] - 2\pi i \nu_0 \tau\}$, where $\alpha(\mathbf{r}')$ is real function of a single position $\mathbf{r}'$ in the domain of the field. Further, the mutual coherence function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ for such a field has a factorized periodic form as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = U(\mathbf{r}_1)U^*(\mathbf{r}_2)\ \exp(-2\pi i \nu_0 \tau) \qquad [32]$$

which means the field is monochromatic with frequency ν_0 and $U(\mathbf{r})$ is any solution of the Helmholtz equation:

$$\nabla^2 U(\mathbf{r}) + k_0^2 U(\mathbf{r}) = 0, \quad k_0 = 2\pi\nu_0/c \qquad [33]$$

The spectral and cross-spectral densities of such fields are represented by Dirac δ-functions with their singularities at some positive frequency ν_0. Such fields, however, can never occur in nature.

Quasi-Monochromatic Fields

Optical fields are found in practice for which spectral spread $\Delta\nu$ is much smaller than the mean frequency $\bar{\nu}$. These are termed as quasi-monochromatic fields. The cross-spectral density $W(\mathbf{r}_1, \mathbf{r}_2, \nu)$ of the quasi-monochromatic fields attains an appreciable value only in the small region $\Delta\nu$, and falls to zero outside this region.

$$W(\mathbf{r}_1, \mathbf{r}_2, \nu) = 0, \quad |\nu - \bar{\nu}| > \Delta\nu \ \text{ and } \ \Delta\nu \ll \bar{\nu} \qquad [34]$$

The mutual coherence function $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ can be written as the Fourier transform of cross-spectral density function as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = e^{-2\pi i \bar{\nu}\tau} \int_0^\infty W(\mathbf{r}_1, \mathbf{r}_2, \nu) e^{-2\pi i(\nu - \bar{\nu})\tau} d\nu \qquad [35]$$

If we limit our attention to small τ such that $\Delta\nu|\tau| \ll 1$ holds, the exponential factor inside the integral is approximately equal to unity and eqn [35] reduces to

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \exp(-2\pi i \bar{\nu}\tau) \int_0^\infty [W(\mathbf{r}_1, \mathbf{r}_2, \nu)] d\nu \qquad [36]$$

which gives

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \Gamma(\mathbf{r}_1, \mathbf{r}_2, 0)\ \exp(-2\pi i \bar{\nu}\tau) \qquad [37]$$

Equation [37] describes the behavior of $\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ for a limited range of τ values for quasi-monochromatic fields and in this range it behaves as a monochromatic field of frequency $\bar{\nu}$. However, due to the factor $\Gamma(\mathbf{r}_1, \mathbf{r}_2, 0)$, the quasi-monochromatic field may be coherent, partially coherent, or even incoherent.

Cross-Spectrally Pure Fields

The complex degree of coherence, if it can be factored into a product of a component dependent on spatial coordinates and a component dependent on time delay, is called reducible. In the case of perfectly coherent light, the complex degree of coherence is reducible, as we have seen above, e.g., in eqn [32], and in the case of quasi-monochromatic fields, this is reducible approximately as shown in eqn [37]. There also exists a very special case of a cross-spectrally pure field for which the complex degree of coherence is reducible. A field is called a cross-spectrally pure field if the normalized spectrum of the superimposed light is equal to the normalized spectrum of the component beams, a concept introduced by Mandel. In the space-frequency domain, the intensity interference law is expressed as the so-called spectral interference law:

$$S(\mathbf{r}, \nu) = S^{(1)}(\mathbf{r}, \nu) + S^{(2)}(\mathbf{r}, \nu) + 2\left[\sqrt{S^{(1)}(\mathbf{r}, \nu)}\sqrt{S^{(2)}(\mathbf{r}, \nu)}\right] \mathrm{Re}[\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) e^{-2\pi i \nu\tau}] \qquad [38]$$

where $\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)$ is the spectral degree of coherence, defined in eqn [21] and τ is the relative time delay that is needed by the light from the pinholes to reach any point on the screen; $S^{(1)}(\mathbf{r}, \nu)$ and $S^{(2)}(\mathbf{r}, \nu)$ are the spectral densities of the light reaching P from the pinholes P_1 and P_2, respectively (see Figure 4) and it is assumed that the spectral densities of the field at pinholes P_1 and P_2 are the same $[S(\mathbf{r}_1, \nu) = CS(\mathbf{r}_2, \nu)]$. Now, if we consider a point for which the time delay is τ_0, then it can be seen that the last term in eqn [38] would be independent of frequency, provided that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)\ \exp(-2\pi i \nu\tau_0) = f(\mathbf{r}_1, \mathbf{r}_2, \tau_0) \qquad [39]$$

where $f(\mathbf{r}_1, \mathbf{r}_2, \tau_0)$ is a function of $\mathbf{r}_1$, $\mathbf{r}_2$ and τ_0 only and the light at this point would have the same spectral density as that at the pinholes. If a region exists around the specified point on the observation plane, such that the spectral distribution of the light in this region is of the same form as the spectral distribution of the light at the pinholes, the light at the pinholes is cross-spectrally pure light.

In terms of the spectral distribution of the light $S(\mathbf{r}_1, \nu)$ at pinhole P_1 and $S(\mathbf{r}_2, \nu)[= CS(\mathbf{r}_1, \nu)]$ at pinhole P_2, the mutual coherence function at the pinholes can be written as

$$\Gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \sqrt{C} \int S(\mathbf{r}_1, \nu)\mu(\mathbf{r}_1, \mathbf{r}_2, \nu)\ \exp(-2\pi i \nu\tau) d\nu \qquad [40]$$

and using eqn [39], we get the very important condition for the field to be cross-spectrally pure, i.e.:

$$\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\gamma(\mathbf{r}_1, \mathbf{r}_1, \tau - \tau_0) \qquad [41]$$

The complex degree of coherence $\gamma(\mathbf{r}_1, \mathbf{r}_2, \tau)$ is thus expressible as the product of two factors: one factor characterizes the spatial coherence at the two pinholes at time delay τ_0 and the other characterizes the temporal coherence at one of the pinholes. Equation [41] is known as the reduction formula for cross-spectrally pure light. It can further be shown that

$$\mu(\mathbf{r}_1, \mathbf{r}_2, \nu) = \gamma(\mathbf{r}_1, \mathbf{r}_2, \tau_0)\ \exp(2\pi i \nu \tau_0) \qquad [42]$$

Thus, the absolute value of the spectral degree of coherence is the same for all frequencies and is equal to the absolute value of the degree of coherence for the point specified by τ_0. It has been shown that cross-spectrally pure light can be generated, for example, by linear filtering of light that emerges from the two pinholes in Young's interference experiments.

Types of Sources

Primary Sources

A primary source is a set of radiating atoms or molecules. In a primary source the randomness comes from true source fluctuations, i.e., from the spatial distributions of fluctuating charges and currents. Such a source gives rise to a fluctuating field. Let $Q(\mathbf{r}, t)$ represent the fluctuating source variable at any point $\mathbf{r}$ at time t, then the field generated by the source is represented by fluctuating field variable $V(\mathbf{r}, t)$. The source is assumed to be localized in some finite domain such that $Q(\mathbf{r}, t) = 0$ at any time $t > 0$ outside the domain. Assuming that field variable $V(\mathbf{r}, t)$ and the source variable $Q(\mathbf{r}, t)$ are scalar quantities, they are related by an inhomogeneous equation as

$$\nabla^2 V(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} V(\mathbf{r}, t) = -4\pi Q(\mathbf{r}, t) \qquad [43]$$

The mutual coherence functions of the source $\Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle Q^*(\mathbf{r}_1, t) Q(\mathbf{r}_2, t+\tau)\rangle$ and of the field $\Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) = \langle V^*(\mathbf{r}_1, t) V(\mathbf{r}_2, t+\tau)\rangle$ characterize the statistical similarity of the fluctuating quantities at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. Following eqn [20], one can define, respectively, the cross-spectral density function of the source and the field as

$$W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_Q(\mathbf{r}_1, \mathbf{r}_2, \tau) \mathrm{e}^{-2\pi i \nu t}\, \mathrm{d}\tau \qquad [44]$$

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_{-\infty}^{\infty} \Gamma_V(\mathbf{r}_1, \mathbf{r}_2, \tau) \mathrm{e}^{-2\pi i \nu t}\, \mathrm{d}\tau$$

The cross-spectral density functions of the source and of the field are related as

$$(\nabla_2^2 + k^2)(\nabla_1^2 + k^2) W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = 4\pi^2 W_Q(\mathbf{r}_1, \mathbf{r}_2, \nu) \qquad [45]$$

The solution of eqn [45] is represented as

$$W_V(\mathbf{r}_1, \mathbf{r}_2, \nu) = \int_S \int_S W_Q(\mathbf{r'}_1, \mathbf{r'}_2, \nu) \frac{\mathrm{e}^{ik(R_2 - R_1)}}{R_1 R_2} \mathrm{d}^3 r'_1\, \mathrm{d}^3 r'_2 \qquad [46]$$

where $R_1 = |\mathbf{r}_1 - \mathbf{r'}_1|$ and $R_2 = |\mathbf{r}_2 - \mathbf{r'}_2|$ (see Figure 8). Using eqn [46], one can then obtain an expression for the spectrum at a point $(\mathbf{r}_1 = \mathbf{r}_2 = \mathbf{r} = r\mathbf{u})$ in the far field $(r \gg r'_1, r'_2)$ of a source as

$$S^{\infty}(r\mathbf{u}, \nu) = \frac{1}{r^2}\int_S \int_S W_Q(\mathbf{r'}_1, \mathbf{r'}_2, \nu) \mathrm{e}^{-ik\mathbf{u}\cdot(\mathbf{r'}_1 - \mathbf{r'}_2)} \mathrm{d}^3 r'_1\, \mathrm{d}^3 r'_2 \qquad [47]$$

where $\mathbf{u}$ is the unit vector along $\mathbf{r}$. The integral in eqn [47], i.e., the quantity defined by $r^2 S^{\infty}(r\mathbf{u}, \nu)$, is also defined as the radiant intensity, which represents the rate of energy radiated at the frequency ν from the source per unit solid angle in the direction of $\mathbf{u}$.

Secondary Sources

Sources used in a laboratory are usually secondary planar sources. A secondary source is a field, which arises from the primary source in the region outside the domain of the primary sources. This kind of source is an aperture on an opaque screen illuminated either directly or via an optical system by primary sources. Let $V(\boldsymbol{\rho}, t)$ represent the fluctuating field in a secondary source plane σ at $z = 0$ (Figure 9) and $W_0(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \nu)$ represent its cross-spectral density (the subscript 0 refers to $z = 0$). One can then solve eqn [25] to obtain the propagation of the cross-spectral density from this planar source. For two points P_1 and P_2 located at distances which are large compared to wavelength, the cross-spectral density is

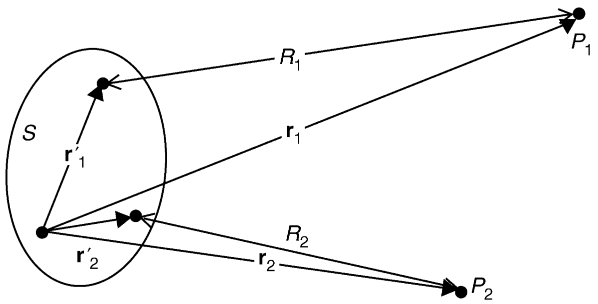

Figure 8 Geometry of a 3-dimensional primary source S and the radiation from it.

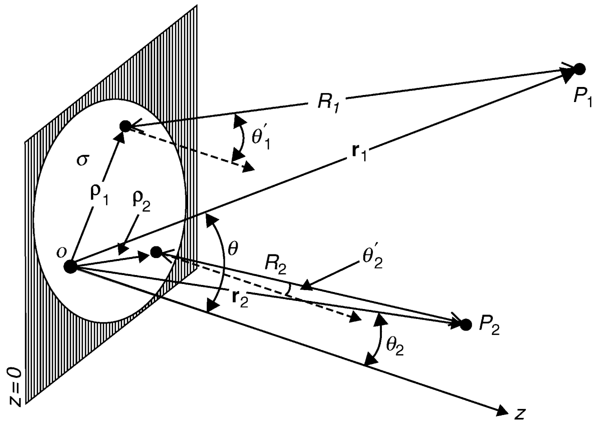

Figure 9 Geometry of a planar source σ and the radiation from it.

then given by

$$W(\mathbf{r}_1,\mathbf{r}_2,\nu)=\left(\frac{k}{2\pi}\right)^2\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu) \times\frac{e^{ik(R_2-R_1)}}{R_1R_2}\cos\theta'_1\cos\theta'_2\,d^2\rho_1\,d^2\rho_2 \qquad [48]$$

where $R_j=|\mathbf{r}_j-\boldsymbol{\rho}_j|$, $(j=1,2)$, θ'_1 and θ'_2 are the angles that R_1 and R_2 directions make with the z-axis (Figure 9). Using eqn [48], one can then obtain an expression for the spectrum at a point $(\mathbf{r}_1=\mathbf{r}_2=\mathbf{r}=r\mathbf{u})$ in the far field $(r\gg\rho_1,\rho_2)$ of a planar source as

$$S^\infty(r\mathbf{u},\nu)=\frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_\sigma\int_\sigma W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2,\nu)\times e^{-ik\mathbf{u}_\perp\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}d^2\rho_1\,d^2\rho_2 \qquad [49]$$

where $\mathbf{u}_\perp$ is the projection of the unit vector $\mathbf{u}$ on the plane σ of the source.

Schell-Model Sources

In the framework of coherence theory in space–time domain, two-dimensional planar model sources of this kind were first discussed by Schell. Later, the model was adopted for formulation of coherence theory in space–frequency domain. Schell-model sources are the sources whose degree of spectral coherence $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ (for either primary or secondary source) is stationary in space. It means that $\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)$ depends on $\mathbf{r}_1$ and $\mathbf{r}_2$ only through the difference $\mathbf{r}_2-\mathbf{r}_1$, i.e., of the form:

$$\mu_A(\mathbf{r}_1,\mathbf{r}_2,\nu)\equiv\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \qquad [50]$$

for each frequency ν present in the source spectrum. Here A stands for field variables V, in the case of a Schell-model secondary source and for source variables Q, in the case of a Schell-model primary source. The cross-spectral density function of a Schell-model source is of the form:

$$W_A(\mathbf{r}_1,\mathbf{r}_2,\nu)=[S_A(\mathbf{r}_1,\nu)]^{1/2}[S_A(\mathbf{r}_2,\nu)]^{1/2}\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu) \qquad [51]$$

where $S_A(\mathbf{r},\nu)$ is the spectral density of the light at a typical point in the primary source or on the plane of a secondary source. Schell model sources do not assume low coherence, and therefore, can be applied to spatially stationary light fields of any state of coherence. The Schell-model of the form given in eqn [51] has been used to represent both three-dimensional primary sources and two-dimensional secondary sources.

Quasi-Homogeneous Sources

Useful models of partially coherent sources that are frequently encountered in nature or in the laboratory are the so-called quasi-homogeneous sources. These are an important sub-class of Schell-model sources. A Schell-model source is called quasi-homogeneous if the intensity of a Schell model source is essentially constant over any coherence area. Under these approximations, the cross-spectral density function for a quasi-homogeneous source is given by

$$\begin{aligned}W_A(\mathbf{r}_1,\mathbf{r}_2,\nu)&=S_A\left[\tfrac{1}{2}(\mathbf{r}_1+\mathbf{r}_2),\nu\right]\mu_A(\mathbf{r}_2-\mathbf{r}_1,\nu)\\&=S_A[\mathbf{r},\nu]\mu_A(\mathbf{r}',\nu)\end{aligned} \qquad [52]$$

where $\mathbf{r}=(\mathbf{r}_1+\mathbf{r}_2)/2$, and $\mathbf{r}'=\mathbf{r}_2-\mathbf{r}_1$. The subscript A stands for either V or Q for the field variable or a source variable, respectively. It is clear that for a quasi-homogeneous source the spectral density $S_A(\mathbf{r},\nu)$ varies so slowly with position that it is approximately constant over distances across the sources that are of the order of the correlation length Λ, which is a measure of the effective width of $|\mu_A(\mathbf{r}',\nu)|$. Therefore, $S_A(\mathbf{r},\nu)$ is a slowly varying function of $\mathbf{r}$ (Figure 10b) and $|\mu_A(\mathbf{r}',\nu)|$ is a fast varying function of $\mathbf{r}'$ (Figure 10a). In addition, the linear dimensions of the source are large compared with the wavelength of light and with the correlation length Δ (Figure 10c).

Quasi-homogeneous sources are always spatially incoherent in the 'global' sense, because their linear dimensions are large compared with the correlation length. This model is very good for representing two-dimensional secondary sources with sufficiently low coherence such that the intensity does not vary over the coherence area on the input plane. It has also been applied to three-dimensional primary sources,

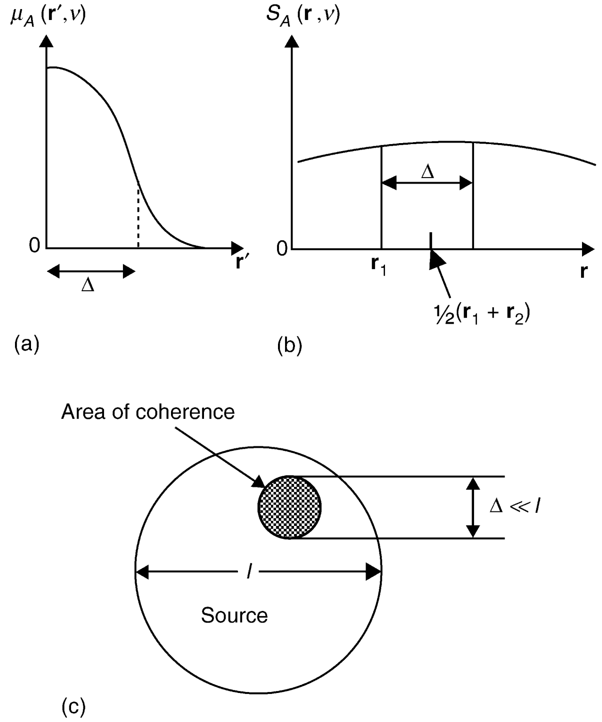

Figure 10 Concept of quasi-homogeneous sources.

three-dimensional scattering potentials, and two-dimensional primary and secondary sources.

Equivalence Theorems

The study of partially coherent sources led to the formulations of a number of equivalence theorems, which show that sources of any state of coherence can produce the same distribution of the radiant intensity as a fully spatially coherent laser source. These theorems provide conditions under which sources of different spatial distribution of spectral density and of different state of coherence will generate fields, which have the same radiant intensity. It has been shown, by taking examples of Gaussian–Schell model sources, that sources of completely different coherence properties and different spectral distributions across the source generate identical distribution of radiant intensity. Experimental verifications of the results of these theorems have also been carried out. For further details on this subject, the reader is referred to Mandel and Wolf (1995).

Correlation-Induced Spectral Changes

It was assumed that spectrum is an intrinsic property of light that does not change as the radiation propagates in free-space, until studies on partially coherent sources and radiations from them in 1980s, revealed that this was true only for specific type of sources. It was discovered on general grounds that the spectrum of light, which originates in an extended source, either a primary source or a secondary source, depends not only on the source spectrum but also on the spatial coherence properties of the source. It was also predicted theoretically by Wolf that the spectrum of light would, in general, be different from the spectrum of the source, and be different at different points in space on propagation in free space.

For a quasi-homogeneous planar secondary source defined by eqn [52], whose normalized spectrum is the same at each source point, one can write the spectral density as

$$S_0(\boldsymbol{\rho},\nu)=I_0(\boldsymbol{\rho})g_0(\nu) \quad \text{with} \quad \int_0^\infty g_0(\nu)\mathrm{d}\nu=1 \qquad [53]$$

where $I_0(\boldsymbol{\rho})$ is the intensity of light at point $\boldsymbol{\rho}$ on the plane of the source, $g_0(\nu)$ is the normalized spectrum of the source and the subscript 0 refers to the quantities of the source plane. Using eqn [49], one can obtain an expression for the far-field spectrum due to this source as

$$S^{\infty}(r\mathbf{u},\nu)=\frac{k^2}{2\pi^2}\frac{\cos^2\theta}{r^2}\int_\sigma\int_\sigma I_0(\boldsymbol{\rho})g_0(\nu)\mu_0(\boldsymbol{\rho}',\nu) \times \mathrm{e}^{-ik\mathbf{u}_\perp\cdot(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)}\,\mathrm{d}^2\rho_1\,\mathrm{d}^2\rho_2 \qquad [54]$$

Noting that $\boldsymbol{\rho}=(\boldsymbol{\rho}_1+\boldsymbol{\rho}_2)/2$ and $\boldsymbol{\rho}'=\boldsymbol{\rho}_2-\boldsymbol{\rho}_1$, one can transform the variables of the integration and obtain after some manipulation:

$$S^{(\infty)}(r\mathbf{u},\nu)=\frac{k^2\cos^2\theta}{(2\pi r)^2}\tilde{I}_0\tilde{\mu}_0(k\mathbf{u}_\perp,\nu)g_0(\nu) \qquad [55]$$

where

$$\tilde{\mu}_0(k\mathbf{u}_\perp,\nu)=\int_\sigma \mu_0(\boldsymbol{\rho}',\nu)\mathrm{e}^{-ik\mathbf{u}_\perp\cdot\boldsymbol{\rho}'}\,\mathrm{d}^2\rho' \qquad [56]$$

and

$$\tilde{I}_0=\int_\sigma I_0(\boldsymbol{\rho})\mathrm{d}^2\rho \qquad [57]$$

Equation [55] shows that the spectrum of the field in the far-zone depends on the coherence properties of the source through its spectral degree of coherence $\mu_0(\boldsymbol{\rho}',\nu)$ and on the normalized source spectrum $g_0(\nu)$.

Scaling Law

The reason why coherence-induced spectral changes were not observed until recently is that the usual thermal sources employed in laboratories or commonly encountered in nature have special coherence properties and the spectral degree of coherence has the function form:

$$\mu^{(0)}(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=f[k(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1)] \quad \text{with} \quad k=\frac{2\pi\nu}{c} \tag{58}$$

which shows that the spectral degree of coherence depends only on the product of the frequency and space coordinates. This formula expresses the so-called scaling law, which was enunciated by Wolf. Commonly used sources satisfy this property. For example, the spectral degree of coherence of Lambertian sources and black-body sources can be shown to be

$$\mu_0(\boldsymbol{\rho}_2-\boldsymbol{\rho}_1,\nu)=\frac{\sin(k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|)}{k|\boldsymbol{\rho}_2-\boldsymbol{\rho}_1|} \tag{59}$$

This expression evidently satisfies the scaling law. If the spectral degree of coherence does not satisfy the scaling law, the normalized far-field spectrum will, in general, vary in different directions in the far-zone and will differ from the source spectrum.

Spectral Changes in Young's Interference Experiment

Spectral changes in Young's interference experiment with broadband light are not as well understood as in experiments with quasi-monochromatic, probably because in such experiments no interference fringes are formed. However, if one were to analyze the spectrum of the light in the region of superposition, one would observe changes in the spectrum of light in the region of superposition in the form of a shift in the spectrum for narrowband spectrum and spectral modulation for broadband light. One can readily derive an expression for the spectrum of light in the region of superposition. Let $S^{(1)}(P,\nu)$ be the spectral density of the light at P which would be obtained if the small aperture at P_1 alone was open; $S^{(2)}(P,\nu)$ has a similar meaning if only the aperture at P_2 was open (Figure 4).

Let us assume, as is usually the case, that $S^{(2)}(P,\nu)\approx S^{(1)}(P,\nu)$ and let d be the distance between the two pinholes. Consider the spectral density at the point P, at distance x from the axis of symmetry in an observation plane located at distance of R from the plane containing the pinholes. Assuming that $x/R \ll 1$, one can make the approximation $R_2 - R_1 \approx xd/R$. The spectral interference law (eqn [38]) can then be written as

$$\begin{aligned} S(P,\nu) \approx\ & 2S^{(1)}(P,\nu)\{1+|\mu(P_1,P_2,\nu)| \\ & \times\cos[\beta(P_1,P_2,\nu)+2\pi\nu xd/cR]\} \end{aligned} \tag{60}$$

where $\beta(P_1,P_2,\nu)$ denotes the phase of the spectral degree of coherence. Equation [60] implies the two results:

(i) at any fixed frequency ν, the spectral density varies sinusoidally with the distance x of the point from the axis, with the amplitude and the phase of the variation depending on the (generally complex) spectral degree of coherence $\mu(P_1,P_2,\nu)$; and

(ii) at any fixed point P in the observation plane the spectrum $S(P,\nu)$ will, in general, differ from the spectrum $S^{(1)}(P,\nu)$, the change also depending on the spectral degree of coherence $\mu(P_1,P_2,\nu)$ of the light at the two pinholes.

Experimental Confirmations

Experimental tests of the theoretical prediction of spectral invariance and noninvariance due to correlation of fluctuations across the source were performed just after the theoretical predictions. Figure 11 shows results of one such experiment in which spectrum changes in the Young's experiment were studied. Several other experiments also reported confirmation of the source correlation-dependent spectral changes. One of the important applications of these observations has been to explain the discrepancies in the maintenance of the spectroradiometric scales by national laboratories in different countries. These studies also have potential application in determining experimentally the spectral degree of coherence of partially coherent fields. The knowledge of spectral degree of coherence is often important in remote sensing, e.g., for determining angular diameters of stars.

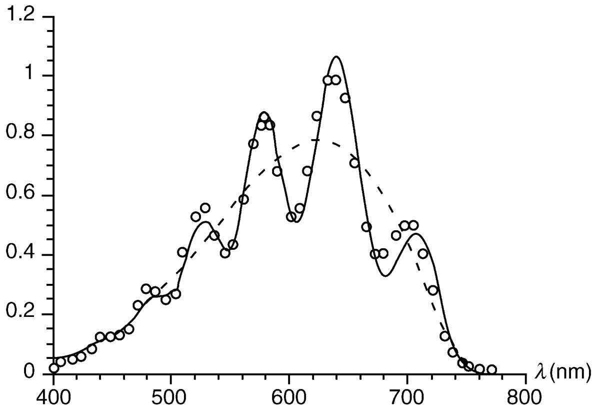

Figure 11 Correlation-induced changes in spectrum in Young's interference. Dashed line represents the spectrum when only one of the slits is open, the continuous curve shows the spectrum when both the slits are open and the circles are the measured values in the latter case. Reproduced with permission from Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

Applications of Optical Coherence

Stellar Interferometry

The Michelson stellar interferometer, named after Albert Michelson, was used to determine the angular diameters of stars and also the intensity distribution across the star. The method was devised by Michelson without using any concept of coherence, although subsequently the full theory of the method was developed on the basis of propagation of correlations. A schematic of the experiment is shown in Figure 12. The interferometer is mounted in front of a telescope, a reflecting telescope in this case. The light from a star is reflected from mirrors M_1 and M_2 and is directed towards the primary mirror (or the objective lens) of the telescope. The two beams thus collected superpose in the focal plane F of telescope where an image crossed with fringes is formed. The outer mirrors M_1 and M_2 can be moved along the axis defined as $M_1M_3M_4M_2$ while the inner mirrors M_3 and M_4 remain fixed. The fringe spacing depends on the position of mirrors M_3 and M_4 and hence is fixed, while the visibility of the fringes depends on the separation of the mirrors M_1 and M_2 and hence, can be varied. Michelson showed that from the measurement of the variation of the visibility with the separation of the two mirrors, one could obtain information about the intensity distribution of the stars, which are rotationally symmetric. He also showed that if the stellar disk is circular and uniform, the visibility curve as a function of the separation d of the mirrors M_1 and M_2 will have zeros for certain values of d, and that the smallest of these d values for which zero occurs is given by $d_0 = 0.61\lambda_a/\alpha$, where α is the semi-angular diameter of the star and λ_a is the mean wavelength of the filtered quasi-monochromatic light from the star. Angular diameters of several stars down to 0.02 second of an arc were determined.

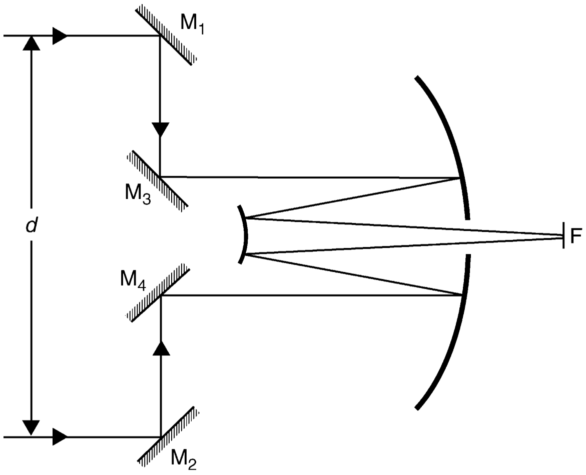

Figure 12 Schematic of the Michelson stellar interferometer.

From the standpoint of second-order coherence theory the principles of the method can be readily understood. The star is considered an incoherent source and according to the van Cittert–Zernike theorem, the light reaching the outer mirrors M_1 and M_2 of the interferometer will be partially coherent. This coherence would depend on the size of and the intensity distribution across the star. Let (x_1, y_1) and (x_2, y_2) be the coordinates of the positions of the mirrors M_1 and M_2, respectively, and (ξ, η) the coordinates of a point on the surface plane of the star which is assumed to be at a very large (astronomical) distance R from the mirrors. The complex degree of coherence at the mirrors would then be given by eqn [22] which can now be written as

$$\gamma(\Delta x, \Delta y, 0) = \frac{\int_\sigma I(u, v) e^{-ik_a(u\Delta x + v\Delta y)}\, du\, dv}{\int_\sigma I(u, v) du\, dv} \qquad [61]$$

where $I(u, v)$ is the intensity distribution across the star disk σ as a function of the angular coordinates $u = \xi/R$, $v = \eta/R$, $\Delta x = x_1 - x_2$, $\Delta y = y_1 - y_2$, and $k_a = 2\pi/\lambda_a$, λ_a being the mean wavelength of the light from the star. Equation [61] shows that the equal-time ($\tau = 0$) complex degree of coherence of the light incident at the outer mirrors of the interferometer is the normalized Fourier-transform of the intensity distribution across the stellar disk. Further, eqn [15] shows that the visibility of the interference fringes is the absolute value of γ, if the intensity of the two interfering beams is equal, as in the present case. The phase of γ can be determined by the position of the intensity maxima of the fringe pattern (eqn [14]). If one is interested in determining only the angular size of the star and the star is assumed to be a circularly symmetric disk of angular diameter 2α and of uniform intensity [$I(u, v)$ is constant across the disk], then eqn [61] reduces to

$$\gamma(\Delta x, \Delta y) = \frac{2J_1(v)}{v}, \quad v = \frac{2\pi\alpha}{\lambda_a} d, \quad d = \sqrt{(\Delta x)^2 + (\Delta y)^2} \qquad [62]$$

The smallest separation of the mirrors for which the visibility γ vanishes corresponds to $v = 3.832$, i.e., $d_0 = 0.61\lambda_a/\alpha$, which is in agreement with Michelson's result.

Interference Spectroscopy

Another contribution of Michelson, which was subsequently identified as an application of the coherence theory, was the use of his interferometer (Figure 2) to determine the energy distribution in the spectral lines. The method he developed is capable to resolving spectral lines that are too narrow to be analyzed by the spectrometer. The visibility of the interference fringes depends on the energy distribution in the spectrum of the light and its measurement can give information about the spectral lines. In particular, if the energy distribution in a spectral line is symmetric about some frequency ν_0, its profile is simply the Fourier transform of the visibility variation as a function of the path difference between the two interfering beams. This method is the basis of interference spectroscopy or the Fourier transform spectroscopy.

Within the framework of second-order coherence theory, if the mean intensity of the two beams in a Michelson's interferometer is the same, then the visibility of the interference fringes in the observation plane is related to the complex degree of coherence of light at a point at the beamsplitter where the two beams superimpose (eqn [15]). The two quantities are related as $v(\tau) = |\gamma(\tau)|$ where $\gamma(\tau) \equiv \gamma(\mathbf{r}_1, \mathbf{r}_1, \tau)$ is the complex degree of self-coherence at the point $\mathbf{r}_1$ on the beamsplitter. Following eqn [19], $\gamma(\tau)$ can be represented by a Fourier integral as

$$\gamma(\tau) = \int_0^\infty s(\nu)\exp(-i2\pi\nu\tau)d\nu \qquad [63]$$

where $s(\nu)$ is the normalized spectral density of the light defined as $s(\nu) = S(\nu)/\int_0^\infty S(\nu)d\nu$ and $S(\nu) \equiv S(\mathbf{r}_1, \nu) = W(\mathbf{r}_1, \mathbf{r}_1, \nu)$ is the spectral density of the beam at the point $\mathbf{r}_1$. The method is usually applied to very narrow spectral lines for which one can assume that the peak that occurs at ν_0, and $\gamma(\tau)$ can be represented as

$$\begin{aligned} &\gamma(\tau) = \bar{\gamma}(\tau)\exp(-2\pi i\nu_0\tau) \quad \text{with} \\ &\bar{\gamma}(\tau) = \int_{-\infty}^\infty \bar{s}(\mu)\exp(-i2\pi\mu\tau)d\mu \end{aligned} \qquad [64]$$

where $\bar{s}(\mu)$ is the shifted spectrum such that

$$\begin{aligned} \bar{s}(\mu) &= s(\nu_0 + \mu) & \mu &\geq -\nu_0 \\ &= 0 & \mu &< -\nu_0 \end{aligned}$$

From the above one readily gets

$$v(\tau) = |\gamma(\tau)| = \left| \int_{-\infty}^\infty \bar{s}(\mu)\, \exp(-i2\pi\mu\tau)d\mu \right| \qquad [65]$$

If the spectrum is symmetric about ν_0, then $v(\tau)$ would be an even function of τ and the Fourier inversion would give:

$$\bar{s}(\mu) = s(\nu_0 + \mu) = 2\int_0^\infty v(\tau)\cos(2\pi\mu\tau)d\tau \qquad [66]$$

which can be used to calculate the spectral energy distribution for a symmetric spectrum about ν_0 from the visibility curve. However, for an asymmetric spectral distribution, the visibility and the phase of the complex degree of coherence must be determined as the Fourier transform of the shifted spectrum is no longer real everywhere.

Higher-Order Coherence

So far we have considered correlations of the fluctuating field variables at two space–time $(\mathbf{r}, t)$ points, as defined in eqn [10]. These are termed as second-order correlations. One can extend the concept of correlations to more than two space–time points, which will involve higher-order correlations. For example, one can define the space–time cross correlation function of order (M, N) of the random field $V(\mathbf{r}, t)$, represented by $\Gamma^{(M,N)}$, as an ensemble average of the product of the field $V(\mathbf{r}, t)$ values at N space–time points and $V^*(\mathbf{r}, t)$ at other M points. In this notation, the mutual coherence function as defined in eqn [10] would now be $\Gamma^{(1,1)}$. Among higher-order correlations, the one with $M = N = 2$, is of practical significance and is called the fourth-order correlation function, $\Gamma^{(2,2)}(\mathbf{r}_1, t_1, \mathbf{r}_2, t_2; \mathbf{r}_3, t_3, \mathbf{r}_4, t_4)$. The theory of Gaussian random variables tells us that any higher correlation can be written in terms of second-order correlations over all permutations of pairs of points. In addition, if we assume that $(\mathbf{r}_3, t_3) = (\mathbf{r}_1, t_1)$ and $(\mathbf{r}_4, t_4) = (\mathbf{r}_2, t_2)$, and that the field is stationary, then $\Gamma^{(2,2)}$ is called the intensity–intensity correlation and is given as

$$\begin{aligned} &\Gamma^{(2,2)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) \\ &\quad = \langle V(\mathbf{r}_1, t_1)V(\mathbf{r}_2, t_2)V^*(\mathbf{r}_1, t_1)V^*(\mathbf{r}_2, t_2)\rangle \\ &\quad = \langle I(\mathbf{r}_1, t_1)I(\mathbf{r}_2, t_2)\rangle \\ &\quad = \langle I(\mathbf{r}_1, t_1)\rangle\langle I(\mathbf{r}_2, t_2)\rangle(1 + |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)|^2) \end{aligned} \qquad [67]$$

where

$$\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1) = \frac{\Gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, t_2 - t_1)}{[\langle I(\mathbf{r}_1, t_1)\rangle]^{1/2}[\langle I(\mathbf{r}_2, t_2)\rangle]^{1/2}} \qquad [68]$$

We now define fluctuations in intensity at $(\mathbf{r}_j, t_j)$ as

$$\Delta I_j = I(\mathbf{r}_j, t_j) - \langle I(\mathbf{r}_j, t_j)\rangle$$

and then the correlation of intensity fluctuations becomes

$$\begin{aligned}\langle \Delta I_1 \Delta I_2 \rangle &= \langle I(\mathbf{r}_1,t_1) I(\mathbf{r}_2,t_2) \rangle - \langle I(\mathbf{r}_1,t_1)\rangle \langle I(\mathbf{r}_2,t_2) \rangle \\ &= \langle I(\mathbf{r}_1,t_1)\rangle \langle I(\mathbf{r}_2,t_2) \rangle |\gamma^{(1,1)}(\mathbf{r}_1,\mathbf{r}_2,t_2-t_1)|^2\end{aligned} \quad [69]$$

where we have used eqn [67]. Equation [69] forms the basis for intensity interferometry.

Hanbury-Brown and Twiss Experiment

In this landmark experiment conducted both on the laboratory scale and astronomical scale, Hanbury-Brown and Twiss demonstrated the existence of intensity–intensity correlations in terms of the correlations in the photocurrents in the two detectors and thus measured the squared modulus of complex degree of coherence. In the laboratory experiment, the arc of a mercury lamp was focused onto a circular hole to produce a secondary source. The light from this source was then divided equally into two parts through a beamsplitter. Each part was, respectively, detected by two photomultiplier tubes, which had identical square apertures in front. One of the tubes could be translated normally to the direction of propagation of light and was so positioned that its image through the splitter could be made to coincide with the other tube. Thus, by suitable translation a measured separation d between the two square apertures could be introduced. The output currents from the photomultiplier tubes were taken by cables of equal length to a correlator. In the path of each cable a high-pass filter was inserted, so that only the current fluctuations could be transmitted to the correlator. Thus, the normalized correlations between the two current fluctuations:

$$C(d) = \frac{\langle \Delta J_1(t) \Delta J_2(t) \rangle}{\langle [\Delta J_1(t)]^2 \rangle^{1/2} \langle [\Delta J_2(t)]^2 \rangle^{1/2}} \quad [70]$$

as a function of detector separation d could be measured. Now when the detector response time is much larger than the time-scale of the fluctuations in intensity, then it can be shown that the correlations in the fluctuations of the photocurrent are proportional to the correlations of the fluctuations in intensity of the light being detected. Thus, we would have

$$C(d) \approx \delta |\gamma^{(1,1)}(\mathbf{r}_1, \mathbf{r}_2, 0)|^2 \quad [71]$$

where δ is the average number of photocounts of the light of one polarization during the time-scale of the fluctuations (for general thermal sources, this is much less than one). Equation [71] represents the Hanbury-Brown–Twiss effect.

Stellar Intensity Interferometry

Michelson stellar interferometry can resolve stars which have angular sizes of the order of 0.01″, since for smaller stars, the separation between the primary mirrors runs into several meters and maintaining stability of mirrors such that the optical paths do not change, even by fraction of a wavelength, is extremely difficult. The atmospheric turbulence further adds to this problem and obtaining stable fringe pattern becomes next to impossible for very small stars. Hanbury-Brown and Twiss applied the intensity interferometry based on their photoelectric correlation technique for determining the angular sizes of such stars. Two separate parabolic mirrors collected light from the star and the output of the photodetectors placed at the focus of each mirror was sent to a correlator. The cable lengths were made unequal so as to compensate for the time difference of the light arrival at the two mirrors. The normalized correlation of the fluctuations of the photocurrents was determined as described above. This would give the variation of the modulus-squared degree of coherence as a function of the mirror separation d from which the angular size of the stars can be estimated. The advantage of the stellar intensity interferometer over the stellar (amplitude) interferometer is that the light need not interfere as in the latter, since the photodetectors are mounted directly at the focus of the primary mirrors of the telescope. Thus, the constraint on the large path difference between the two beams is removed and large values of d can now be used. Moreover, the atmosphere turbulence and the mirror movements have very small effect. Stellar angular diameters as small as 0.0004″ of arc with resolution of 0.00003″ could be measured by such interferometers.

See also

Coherence: Coherence and Imaging. **Coherent Lightwave Systems**. **Coherent Transients:** Coherent Transient Spectroscopy in Atomic and Molecular Vapours. **Coherent Control:** Applications in Semiconductors; Experimental; Theory. **Interferometry:** Overview. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Beran MJ and Parrent GB (1964) *Theory of Partial Coherence*. Englewood Cliffs, NJ: Prentice-Hall.

Born M and Wolf E (1999) *Principles of Optics*. New York: Pergamon Press.

Carter WH (1996) Coherence theory. In: Bass M (ed.) *Hand Book of Optics*. New York: McGraw-Hill.

Davenport WB and Root WL (1960) *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill.

Goodman JW (1985) *Statistical Optics*. Chichester: Wiley.

Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: I basic theory: the correlation between photons in coherent beams of radiation. *Proceedings of the Royal Society* 242: 300–324.

Hanbury-Brown R and Twiss RQ (1957) Interferometry of intensity fluctuations in light: II an experimental test of the theory for partially coherent light. *Proceedings of the Royal Society* 243: 291–319.

Kandpal HC, Vaishya JS and Joshi KC (1994) Correlation-induced spectral shifts in optical measurements. *Optical Engineering* 33: 1996–2012.

Mandel L and Wolf E (1965) Coherence properties of optical fields. *Review of Modern Physics* 37: 231–287.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Marathay AS (1982) *Elements of Optical Coherence*. New York: John Wiley.

Perina J (1985) *Coherence of Light*. Boston, MA: M.A. Reidel.

Santarsiero M and Gori F (1992) Spectral changes in Young interference pattern. *Physics Letters* 167: 123–128.

Schell AC (1967) A technique for the determination of the radiation pattern of a partially coherent aperture. *IEEE Transactions on Antennas and Propagation* AP-15: 187.

Thompson BJ (1958) Illustration of phase change in two-beam interference with partially coherent light. *Journal of the Optical Society of America* 48: 95–97.

Thompson BJ and Wolf E (1957) Two beam interference with partially coherent light. *Journal of the Optical Society of America* 47: 895–902.

Wolf E and James DFV (1996) Correlation-induced spectral changes. *Report on Progress in Physics* 59: 771–818.

Coherence and Imaging

J van der Gracht, HoloSpex, Inc., Columbia, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical imaging systems are strongly affected by the coherence of the light that illuminates the object of interest. In many cases, the light is approximately coherent or incoherent. These approximations lead to simple mathematical models for the image formation process and allow straightforward analysis and design of such systems. When the light is partially coherent, the mathematical models are more complicated and system analysis and design is more difficult. Partially coherent illumination is often used in microscopy, machine vision, and optical lithography. The intent of this article is to provide the reader with a basic understanding of the effects of coherence on imaging. The information should enable the reader to recognize when coherence effects are present in an imaging system and give insight into when coherence can be modified to improve imaging performance. The material relies mostly on concepts drawn from the Fourier optics perspective of imaging and a rigorous coherence theory treatment is avoided. We encourage the reader to consult the Further Reading list at the end of this article, for more complete definitions of terms for coherence theory. A number of experimental results highlight the effects of spatial coherence of the illumination on image formation. Discussions of the role of coherence in such key applications are interspersed throughout the article.

Image Formation – Ideal and Optimal

An image is typically defined as the reproduction or likeness of the form of an object. An image that is indistinguishable from the original object is generally considered to be ideal. In a general context, the sound of a voice coming from a loudspeaker can be thought of as the image of the sound coming directly from the original speaker's mouth. In optical imaging, the ideal image replicates the light emanating from the object. Taken to the extreme, the ideal image replicates the light leaving the object in terms of intensity, wavelength, polarization, and even coherence. When the final image is viewed by the eye, the ideal image only needs to replicate the spatial distribution of the light leaving the object in terms of color and relative intensity at each point on the object. (In this article, intensity is defined as optical power per unit area (watts per meter squared).)

A general diagram of a direct view image formation system is shown in Figure 1. The condenser optics gathers light from a primary source and illuminates a transmissive object having a complex wave amplitude transmission of $O(x, y)$. The imaging optics produce an optical image that is viewed directly by the viewer. A growing number of image formation systems now include a solid-state image detector as shown in Figure 2. The raw intensity optical image is converted to an electronic signal that can be digitally processed and then displayed on a monitor. In this system, the spatial intensity distribution emanating from the monitor is the final image. The task of the optical system to the left of the detector is to gather spatial information about the light properties of the object. In fact, the information gathered by the detector includes information about the object, the illumination system, and the image formation optics. If the observer is only interested in the light transmission properties of the object, the effects of the illumination and image formation optics must be well understood. When the light illuminating the object is known to be coherent or incoherent, reasonably simple models for the overall image formation process can be used. More general, partially coherent illumination can produce optically formed images that differ greatly from the intensity pattern leaving the object. Seen from this perspective, partially coherent illumination is an undesirable property that creates nonideal images and complicates the image analysis.

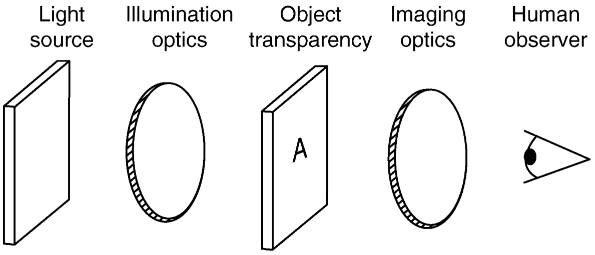

Figure 1 Direct view optical system. Imaging optics conveys the light from an illuminated object directly to the human viewer.

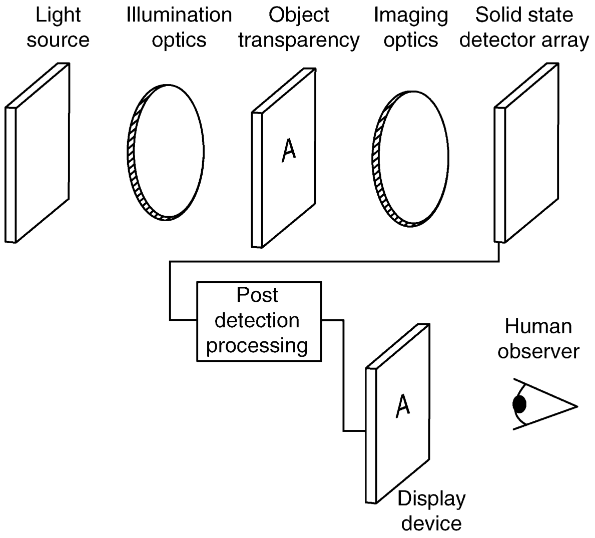

Figure 2 Indirect view optical system. The raw intensity optical image is detected electronically, processed and a final image is presented to the human observer on a display device.

The ideal image as described above is not necessarily the optimum image for a particular task. Consider the case of object recognition when the object has low optical contrast. Using the above definition, an ideal image would mimic the low contrast and render recognition difficult. An image formation system that alters the contrast to improve the recognition task would be better than the so-called ideal image. The image formation system that maximized the appropriate performance metric for a particular recognition task would be considered optimal. In optimizing the indirect imaging system of Figure 2, the designer can adjust the illumination, the imaging optics, and the post-detection processing. In fact, research microscopes often include an adjustment that modifies illumination coherence and often alters the image contrast. Darkfield and phase contrast imaging microscopes usually employ partially coherent illumination combined with pupil modification to view otherwise invisible objects. Seen from this perspective, partially coherent illumination is a desirable property that provides more degrees of freedom to the imaging system designer. Quite often, partially coherent imaging systems provide a compromise between the performance of coherent and incoherent systems.

Elementary Coherence Concepts

Most readers are somewhat familiar with the concept of temporal coherence. In Figure 3, a Michelson interferometer splits light from a point source into two paths and recombines the beams to form interference fringes. The presence of interference fringes indicates that the wave amplitude fluctuations of the two beams are highly correlated so the light adds in wave amplitude. If the optical path difference between the two paths can be made large without reducing the fringe contrast, the light is said to be

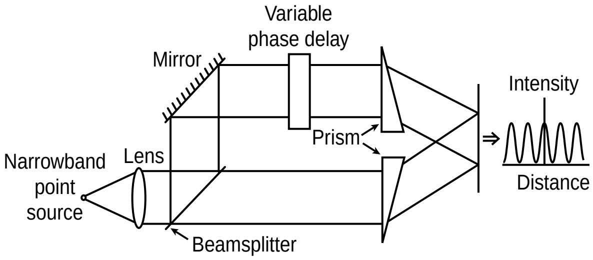

Figure 3 The presence of high contrast fringes in a Michelson interferometer indicates high temporal coherence.

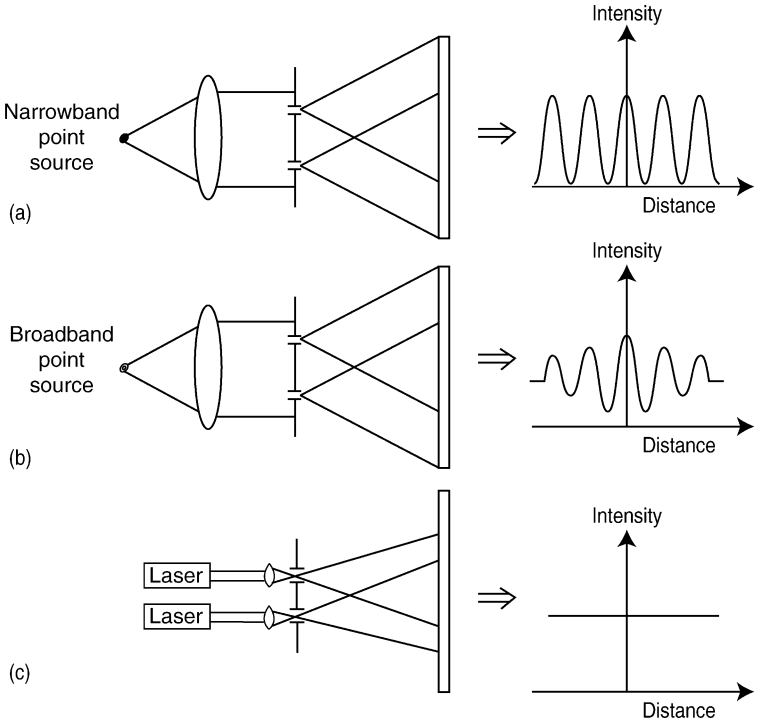

Figure 4 A Young's two pinhole interferometer produces: (a) a uniform high contrast fringe pattern for light that is highly spatially and temporally coherent; (b) high contrast fringes only near the axis for low temporal coherence and high spatial coherence light; and (c) no fringe pattern for high temporal coherence, low spatial coherence light.

highly temporally coherent. Light from very narrowband lasers can maintain coherence over very large optical path differences. Light from broadband light sources requires very small optical path differences to add in wave amplitude.

Spatial coherence is a measure of the ability of two separate points in a field to interfere. The Young's two pinhole experiment in Figure 4 measures the coherence between two points sampled by the pinhole mask. Only a one-dimensional pinhole mask is shown for simplicity. Figure 4a shows an expanded laser beam illuminating two spatially separated pinholes and recombining to form an intensity fringe pattern. The high contrast of the fringe indicates that the wave amplitude of the light from the two pinholes is highly correlated. Figure 4b shows a broadband point source expanded in a similar fashion. The fringe contrast is high near the axis because the optical path difference between the two beams is zero on the axis and relatively low near the axis. For points far from the axes, the fringe pattern disappears because the low temporal coherence from the broadband source results in a loss in correlation of the wave amplitudes.

A final Young's experiment example in Figure 4c shows that highly temporally coherent light can be spatially incoherent. In the figure, the light illuminating the two pinholes comes from two separate and highly temporally coherent lasers that are designed to have the same center frequency. Since the light from the lasers is not synchronized in phase, any fringes that might form for an instant will move rapidly and average to a uniform intensity pattern over the integration time of a typical detector. Since the fringe contrast is zero over a practical integration time, the light at the two pinholes is effectively spatially incoherent.

Two-point Imaging

In a typical partially coherent imaging experiment we need to know how light from two pinholes adds at the optical image plane, as shown in Figure 5. Diffraction and system aberrations cause the image of a single point to spread so that the images of two spatially separated object points overlap in the image plane. The wave amplitude image of a single pinhole is called the coherent point spread function (CPSF). Since the CPSF is compact, two CPSFs will only overlap when the corresponding object points are closely spaced. When the two points are sufficiently close, the relevant optical path differences will be small so full temporal coherence can be assumed. Spatial coherence will be the critical factor in determining how to add the responses to pairs of point images. We assume full temporal coherence in the subsequent analyses and the light is said to be *quasimonochromatic*.

Coherent Two-Point Imaging

Consider imaging a two-point object illuminated by a spatially coherent plane wave produced from a point source, as depicted in Figure 6a. In the figure, the imaging system is assumed to be space-invariant. This means that the spatial distribution of the CPSF is the same regardless of position of the input pinhole. The CPSF in the figure is broad enough such that the image plane point responses overlap for this particular pinhole spacing. Since the wave amplitudes from the two pinholes are correlated, the point responses add in wave amplitude, resulting in a two-point image intensity given by

$$I_{2\mathrm{coh}}(x) = I_0|h(x - x_1) + I_0 h(x - x_2)|^2 \qquad [1]$$

where $h(x)$ is the normalized CPSF and I_0 is a scaling factor that determines the absolute image intensity value. Since the CPSF has units of wave amplitude, the magnitude squaring operation accounts for the square law response of the image plane detector.

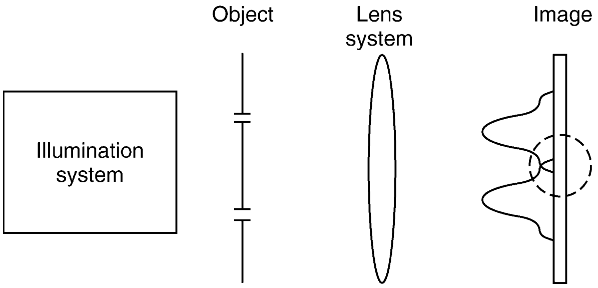

Figure 5 Imaging of two points. The object plane illumination coherence determines how the light adds in the region of overlap of the two image plane points.

More generally, the image intensity for an arbitrary object distribution can be found by breaking the object into a number of points and adding the CPSFs due to each point on the object. The resultant image plane intensity is the spatial convolution of the object amplitude transmittance with the CPSF and is given by

$$I_{\mathrm{coh}}(x) = I_0\left|\int O(\xi)h(\xi - x)\mathrm{d}\xi\right|^2 \qquad [2]$$

where $O(x)$ is the object amplitude transmittance.

Incoherent Two-Point Imaging

Two pinhole imaging with spatially incoherent light is shown in Figure 6b, where a spatially extended blackbody radiator is placed directly behind the pinholes. Once again the imaging optics are assumed to be space-invariant and have the same CPSF as the system in Figure 6a. Since the radiation is originating from two completely different physical points on the source, no correlation is expected between the wave amplitude of the light leaving the pinholes and the light is said to be spatially incoherent. The resultant image intensity corresponding to the two pinhole object with equal amplitude transmittance values is calculated by adding the individual intensity responses to give

$$I_{2\mathrm{inc}}(x) = I_0|h(x - x_1)|^2 + I_0|h(x - x_2)|^2 \qquad [3]$$

For more general object distributions, the image intensity is the convolution of the object intensity transmittance with the incoherent PSF, and is given by

$$I_{\mathrm{inc}}(x) = I_0\int |o(\xi)|^2|h(x - \xi)|^2\mathrm{d}\xi \qquad [4]$$

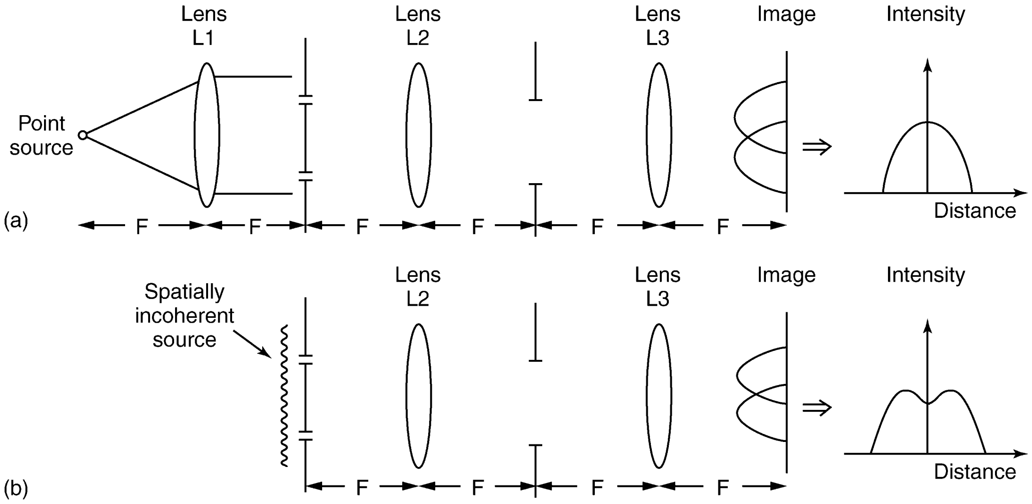

Figure 6 Imaging of two points for (a) spatially coherent illumination and (b) spatially incoherent imaging.

where the intensity transmittance is the squared magnitude of the amplitude transmittance and the incoherent PSF is the squared magnitude of the CPSF.

Partially Coherent Two-Point Imaging

Finally we consider two pinhole imaging illuminated by partially spatially coherent light. Now the two-point responses do not add simply in amplitude or intensity. Rather, the image intensity is given by the more general equation:

$$I_{2\mathrm{pc}}(x) = I_0\big[|h(x-x_1)^2 + |h(x-x_2)|^2 + 2\mathrm{Re}\{\mu(x_1;x_2)h(x-x_1)h^*(x-x_2)\}\big] \quad [5]$$

where the $*$ denotes complex conjugations and $\mu(x_1;x_2)$ is the normalized form of the mutual intensity of the object illumination evaluated at the object points in question. The mutual intensity function is often denoted as $J_0(x_1;x_2)$ and is a measure of the cross correlation of the wave amplitude distributions leaving the two pinholes. Rather than providing a rigorous definition, we note that the magnitude of $J_0(x_1;x_2)$ corresponds to the fringe contrast that would be produced if the two object pinholes were placed at the input of a Young's interference experiment. The phase is related to the relative spatial shift of the fringe pattern. When the light is uncorrelated, $J_0(x_1;x_2)=0$ and eqn [5] collapses to the incoherent limit of eqn [4]. When the light is coherent, $J_0(x_1;x_2)=1$ and eqn [5] reduces to the coherent form of eqn [2].

Image formation for a general object distribution is given by the bilinear equation:

$$I_{\mathrm{pc}}(x) = I_0\int o(\xi_1)o^*(\xi_2)h(x-\xi_1)h^*(x-\xi_2) \times J_o(\xi_1-\xi_2)\mathrm{d}\xi_1\,\mathrm{d}\xi_2 \quad [6]$$

Note that, in general, $J_0(x_1;x_2)$ must be evaluated for all pairs of object points. Close examination of this equation reveals that just as a linear system can be evaluated by considering all possible single point responses, a bilinear system requires consideration of all possible pairs of points. This behavior is much more complicated and does not allow the application of the well-developed linear system theory.

Source Distribution and Object Illumination Coherence

According to eqn [6], the mutual intensity of all pairs of points at the object plane must be known, to calculate a partially coherent image. Consider the telecentric Kohler illumination imaging system shown in Figure 7. The primary source is considered to be spatially incoherent and illuminates the object after passing through lens L1 located one focal distance away from the source and one focal distance away from the object. Even though the primary source is spatially incoherent, the illumination at the object plane is partially coherent. The explicit mutual intensity function corresponding to the object plane illumination is given by applying the van Cittert Zernike theorem:

$$J_0(\Delta x) = \int S(\xi)\exp[\mathrm{j}2\pi\Delta x\xi/\lambda F]\mathrm{d}\xi \quad [7]$$

where $S(x)$ is the intensity distribution of the spatially incoherent primary source, F is the focal length of the lens, λ is the illumination wavelength, and $\Delta x = x_1 - x_2$. The van Cittert Zernike theorem reflects a Fourier transform relationship between the source image intensity distribution and the mutual intensity at the object plane. When the source plane is effectively spatially incoherent, the object plane mutual intensity is only a function of separation distance. For a two-dimensional object, the mutual intensity needs to be characterized for all pairs of unique spacings in x and y. In Figure 7, the lens arrangement ensures that the object plane is located in the far field of the primary source plane. In fact, the van Cittert Zernike theorem applies more generally, even in the Fresnel propagation regime, as long as the standard paraxial approximation for optics is valid.

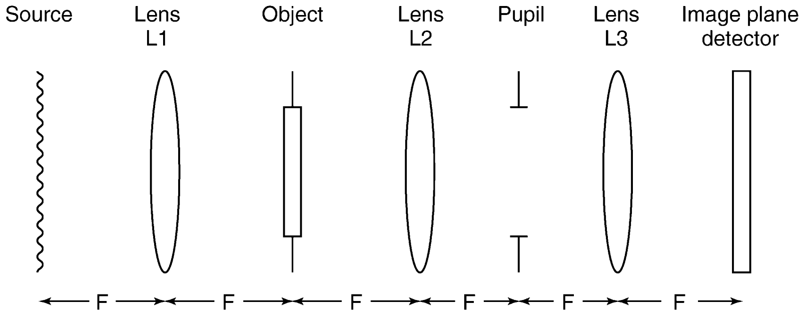

Figure 7 A telecentric Kohler illumination system with a spatially incoherent primary source imaged onto the pupil.

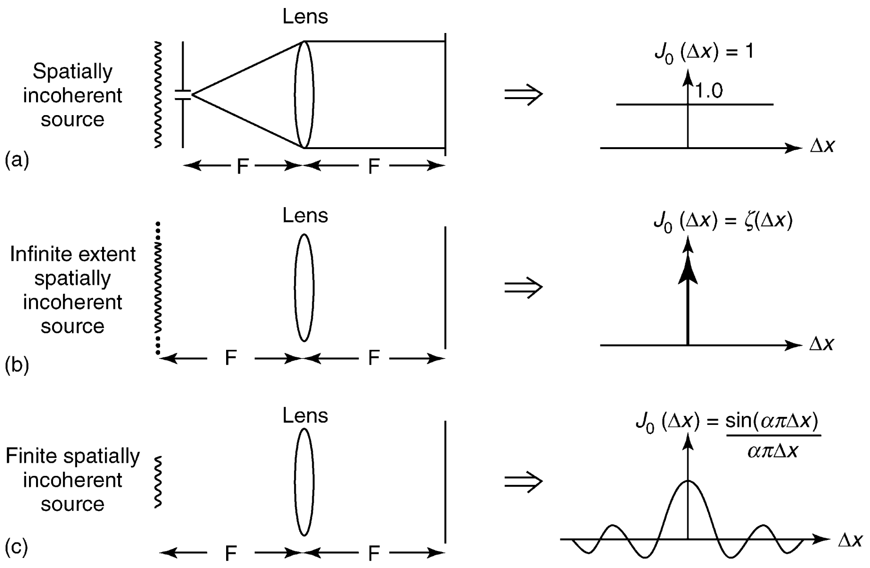
Lens
Spatially incoherent source
(a)
F
F
$J_0(\Delta x) = 1$
1.0
Δx
Lens
Infinite extent spatially incoherent source
(b)
F
F
$J_0(\Delta x) = \zeta(\Delta x)$
Δx
Lens
Finite spatially incoherent source
(c)
F
F
$J_0(\Delta x) = \frac{\sin(\alpha\pi\Delta x)}{\alpha\pi\Delta x}$
Δx

Figure 8 Light propages from the source plane to the object plane and produces (a) coherent illumination from a single source point; (b) incoherent illumination from an infinite extent source; and (c) partially coherent illumination from a finite extent primary source.

Figure 8 shows some simple examples of primary source distributions and the corresponding object plane mutual intensity function. Figure 8a assumes an infinitesimal point source and the corresponding mutual intensity is 1.0 for all possible pairs of object points. Figure 8b assumes an infinite extent primary source and results in a dirac delta function for the mutual intensity function. This means that there is no correlation between any two object points having a separation greater than zero so the object plane illumination is spatially incoherent. In Figure 5c, a finite extent uniform source gives a mutual intensity function of the form $\sin(\alpha\pi\Delta x)/(\alpha\Delta x)$. The finite-sized source corresponds to partially coherent imaging and shows that the response to pairs of points is affected by the spatial distribution of the source in a complicated way. Note that a large primary source corresponds to a large range of angular illumination at the object plane.

Varying the size of the source in the imaging system of Figure 8 will affect the spatial coherence of the illumination and hence the optical image intensity. Many textbook treatments discuss the imaging of two points separated by the Rayleigh resolution criterion which corresponds to the case where the first minimum of one point image coincides with the maximum of the adjacent point image. With a large source that provides effectively incoherent light, the two point image has a modest dip in intensity between the two points, as shown in Figure 9a. Fully coherent illumination of two points separated by the Rayleigh resolution produces a single large spot with no dip in intensity, as shown in Figure 9b. Varying the source size and hence the illumination spatial coherence produces a dip that is less than the incoherent intensity dip. This result is often used to suggest that coherent imaging gives poorer resolution than incoherent imaging. In fact, generalizations about two-point resolution can be misleading.

Recent developments in optical lithography have shown that coherence can be used to effectively increase two-point resolution beyond traditional diffraction limits. In Figure 9c, one of the pinholes in a coherent two-point imaging experiment has been modified with a phase shift corresponding to one half of a wavelength of the illuminating light. The two images add in wave amplitude and the phase shift creates a distinct null at the image plane and effectively enhances the two-point resolution. This approach is termed phase screen lithography and has been exploited to produce finer features in lithography by purposely introducing small phase shift masks at the object mask. In practice, the temporal and spatial coherence of the light is engineered to give sufficient coherence to take advantage of the two-point enhancement while maintaining sufficient incoherence to avoid speckle-like artifacts associated with coherent light.

Spatial Frequency Modeling of Imaging

Spatial frequency models of image formation are also useful in understanding how coherence affects image formation. A spatially coherent imaging system has a particularly simple spatial frequency model. In the

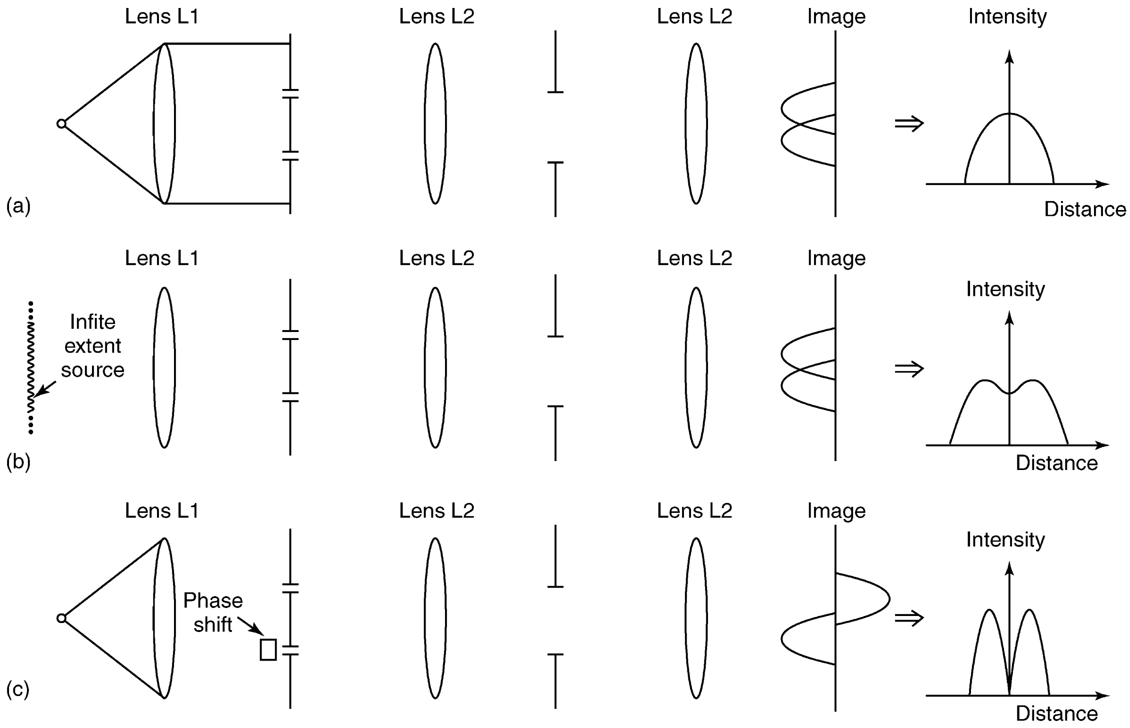
Lens L1
Lens L2
Lens L2
Image
Intensity
⇒
Distance
(a)
Lens L1
Lens L2
Lens L2
Image
Infite extent source
Intensity
⇒
Distance
(b)
Lens L1
Lens L2
Lens L2
Image
Phase shift
Intensity
⇒
Distance
(c)

Figure 9 Comparison of coherent and two point imaging of two pinholes separated by the Rayleigh distance. (a) Coherent image cannot resolve the two points. (b) Incoherent image barely resolves the two points. (c) Coherent illumination with a phase shifting plate at one pinhole produces a null between the two image points.

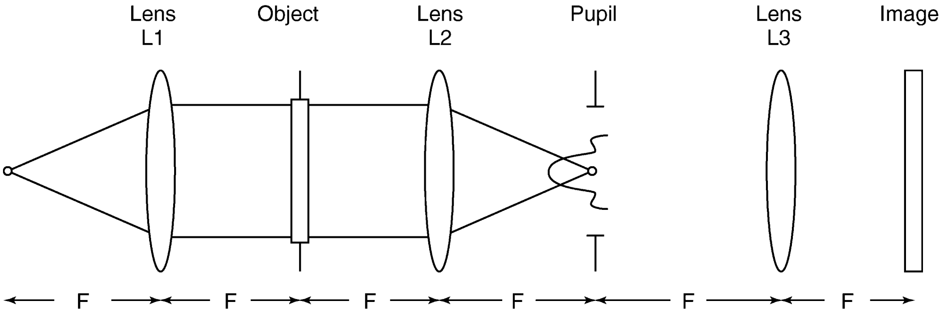
Lens L1
Object
Lens L2
Pupil
Lens L3
Image
F
F
F
F
F
F

Figure 10 The object Fourier transform is filtered by the pupil function in a coherent imaging system.

spatially coherent imaging system shown in Figure 10, the on-axis plane wave illumination projects the spatial Fourier transform of the object, or Fraunhofer pattern, onto the pupil plane of the imaging system. The pupil acts as a frequency domain filter that can be modified to perform spatial frequency filtering. The complex transmittance of the pupil is the wave amplitude spatial frequency transfer function of the imaging system. It follows that the image plane wave amplitude frequency spectrum, $\tilde{U}_c(f_x)$, is given by

$$\tilde{U}_c(f_x) = \tilde{O}(f_x)H(f_x) \qquad [8]$$

where $\tilde{O}(f_x)$ is the spatial Fourier transform of the object amplitude transmittance function and $H(f_x)$ is proportional to the pupil plane amplitude transmittance. Equation [8] is the frequency domain version of the convolution representation given by eqn [2], but does not account for the squared magnitude response of the detector.

In the previous section, we learned that an infinite extent source is necessary to achieve fully incoherent imaging for the imaging system of Figure 7. As a thought experiment, one can start with a single point source on the axis and keep adding mutually incoherent source points to build up from a coherent imaging system to an incoherent imaging system. Since the individual source points are assumed to be incoherent with each other, the images from each point source can be added in intensity. In Figure 11,

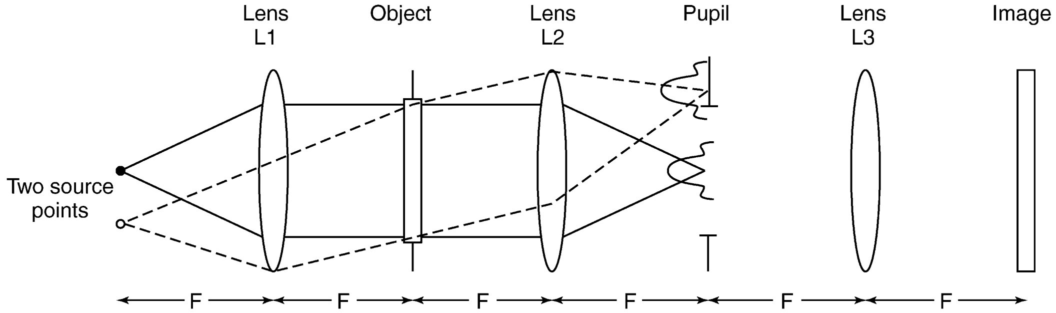

Figure 11 A second source point projects a second object Fourier transform onto the pupil plane. The images produced by the two source points add in intensity at the image plane.

only two point sources are shown. Each point source projects an object Fraunhofer pattern onto the pupil plane. The centered point source will result in the familiar coherent image. The off-axis point source projects a displaced object Fraunhofer pattern that is also filtered by the pupil plane before forming an image of the object. The image formed from this second point source can be modeled with the same coherent imaging model with a shifted pupil plane filter. Since the light from the two source points is uncorrelated, the final image is calculated by adding the intensities of the two coherently formed images.

The two source point model of Figure 11 can be generalized to an arbitrary number of source points. The final image is an intensity superposition of a number of coherent images. This suggests that partially coherent imaging systems behave as a number of redundant coherent imaging systems, each having a slightly different amplitude spatial frequency transfer function due to the relative shift of the pupil filter with respect to the object Fraunhofer pattern. As the number of point sources is increased to infinity, the primary source approaches an infinite extent spatially incoherent source. In practice, the source need not be infinite. When the source is large enough to effectively produce linear-in-intensity imaging, the imaging system is effectively spatially incoherent. The corresponding image intensity in the spatial frequency domain, $\tilde{I}_{\text{inc}}(f_x)$, is given by

$$\tilde{I}_{\text{inc}}(f_x) = \tilde{I}_{\text{obj}}(f_x)\text{OTF}(f_x) \qquad [9]$$

where $\tilde{I}_{\text{obj}}(f_x)$ is the spatial Fourier transform of the object intensity transmittance and $\text{OTF}(f_x)$ is the incoherent optical transfer function which is proportional to the spatial autocorrelation of the pupil function.

Many texts include detailed discussions comparing the coherent transfer function and the OTF and attempt to make comparisons about the relative performance of coherent and incoherent systems. These comparisons are often misleading since one transfer function describes the wave amplitude spatial frequency transfer function and the other describes the intensity spatial frequency transfer function. Here we note that incoherent imaging systems do indeed allow higher object amplitude spatial frequencies to participate in the image formation process. This argument is often used to support the claim that incoherent systems have higher resolution. However, both systems have the same image intensity spatial frequency cutoff. Furthermore, the nature of the coherent transfer function tends to produce high contrast images that are typically interpreted as higher-resolution images than their incoherent counterparts. Perhaps the real conclusion is that the term resolution is not well defined and direct comparisons between coherent and incoherent imaging must be treated carefully.

The frequency domain treatment for partially coherent imaging of two-dimensional objects involves a four-dimensional spatial frequency transfer function that is sometimes called the bilinear transfer function or the transmission cross-coefficient model. This model describes how constituent object wave amplitude spatial frequencies interact to form image plane intensity frequencies. The utility of this approach to analysis is limited for someone new to the field, but is often used in numerical simulations of partially coherent imaging systems used in optical lithography.

Experimental Examples of Important Coherence Imaging Phenomena

Perhaps the best way to gain an understanding of coherence phenomena in imaging is to examine experimental results. In the following section we use experimental data to see how coherence affects edge response, noise immunity, and depth of field. Several experimental configurations were used to collect the image data, but all of the systems can be represented generically by the Kohler illumination system shown in Figure 12. Kohler illumination is often employed in many projection illumination systems. In order to obtain Kohler illumination, the primary spatially incoherent source is imaged onto the pupil plane of the imaging portion of the system and the object is placed at the pupil plane of the condenser optics.

Primary Source Generation

Figure 13 shows a highly coherent illumination system produced by focusing a laser beam to a point and imaging the focused spot onto the pupil of the imaging system to produce spatially coherent illumination. The use of a laser produces highly temporally coherent light.

Two methods were used to create an extended spatially incoherent primary source with control over the spatial intensity distribution. Figure 14a shows a collimated laser beam with a 633 nm center wavelength illuminating a rotating diffuser. A photographically produced mask defines the spatial shape of the primary source. The laser provides highly temporally coherent light and the diffuser destroys the spatial coherence of the light. Consider the thought experiment of two pinholes placed immediately to the right of the moving diffuser. Without the diffuser, the two wave amplitudes would be highly correlated. Assuming that the diffuser can be modeled as a spatially random phase plate, a fixed diffuser would only introduce a fixed phase difference between the amplitudes leaving the two pinholes and would not destroy the coherence. When the diffuser is rotated, the light from each pinhole encounters a different phase modulation that is changing over time. This random modulation destroys the effective correlation between the wave amplitude of the light leaving the two pinholes provided that the rotation speed is sufficiently fast.

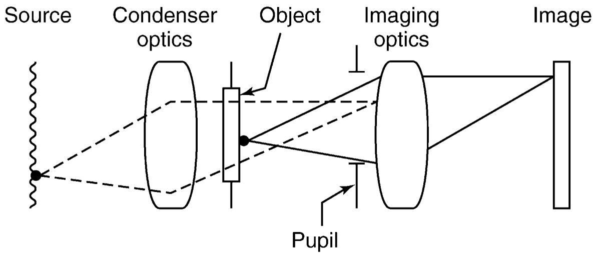

Figure 12 General representation of the Kohler illumination imaging systems used in the experimental result section. The primary incoherent source is imaged onto the imaging system pupil plane and the object resides in the pupil of the condenser optics.

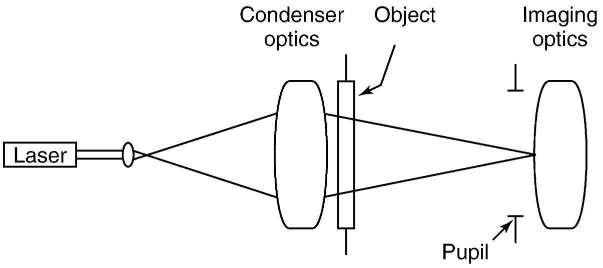

Figure 13 Highly spatially and temporally coherent illumination produced by imaging a focused laser beam into the imaging system pupil.

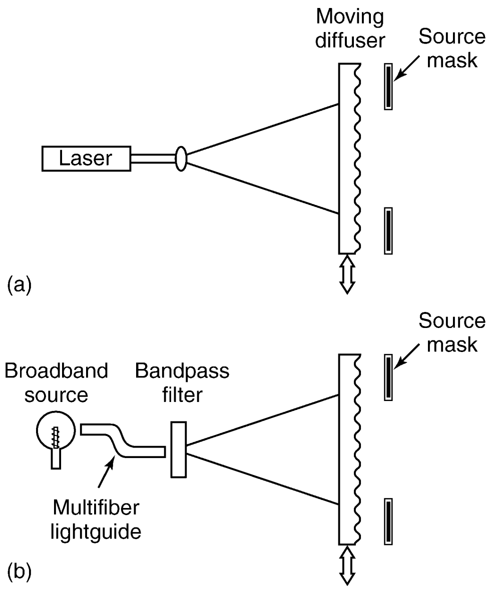

Figure 14 Two methods for generating a spatially incoherent primary source. (a) An expanded laser beam passes through a moving diffuser followed by a mask to define the extent and shape of the source. (b) A broadband source exits a multifiber lightguide and propagates to a moving diffuser followed by a source mask.

The moving diffuser method is light inefficient but is a practical way of exploring coherence in imaging in a laboratory environment. The choice of the diffuser is critical. The diffuser should spread light out uniformly over an angular subtense that overfills the object of interest. Many commercially available diffusers tend to pass too much light in the straight through direction. Engineered diffusers can be purchased to produce an optimally diffused beam. When quick and inexpensive results are required, thin plastic sheets used in day-to-day packaging often serve as excellent diffusers. When the diffusion angles are not high enough, a number of these plastic sheets can be layered on top of each other to achieve the appropriate angular spread.

The second method for producing a spatially incoherent source is shown in Figure 14b. Broadband light is delivered through a large multifiber lightguide and illuminates a moving diffuser. The main purpose of the diffuser is to ensure that the entire object is illuminated uniformly. The low temporal coherence of the source and the long propagation distances are usually enough to destroy any spatial coherence at the plane of the diffuser. The rotation further ensures that no residual spatial coherence exists at the primary source plane. A chromatic filter centered around 600 nm, with a spectral width of approximately 100 nm, is shown in the light path. The filter was used to minimize the effects of chromatic aberrations in the system. The wide spectral width certainly qualifies as a broadband source in relation to a laser. High-contrast photographically defined apertures determine the spatial intensity distribution at the plane of the primary source. In the following results, the primary source distributions were restricted to circular sources of varying sizes as well as annular sources. These shapes are representative of most of the sources employed in microscopy, machine vision, and optical lithography.

Noise Immunity

Coherent imaging systems are notorious for introducing speckle-like noise artifacts at the image. Dust and optically rough surfaces within an optical system result in a complicated textured image plane intensity distribution that is often called speckle noise. We refer to such effects as coherent artifact noise. Figure 15a shows the image of a standard binary target as imaged by a benchtop imaging system, with the highly coherent illumination method shown in Figure 13. Some care was taken to clean individual lenses and optical surfaces within the system and no dust was purposely introduced. The image is corrupted by a complicated texture that is due in

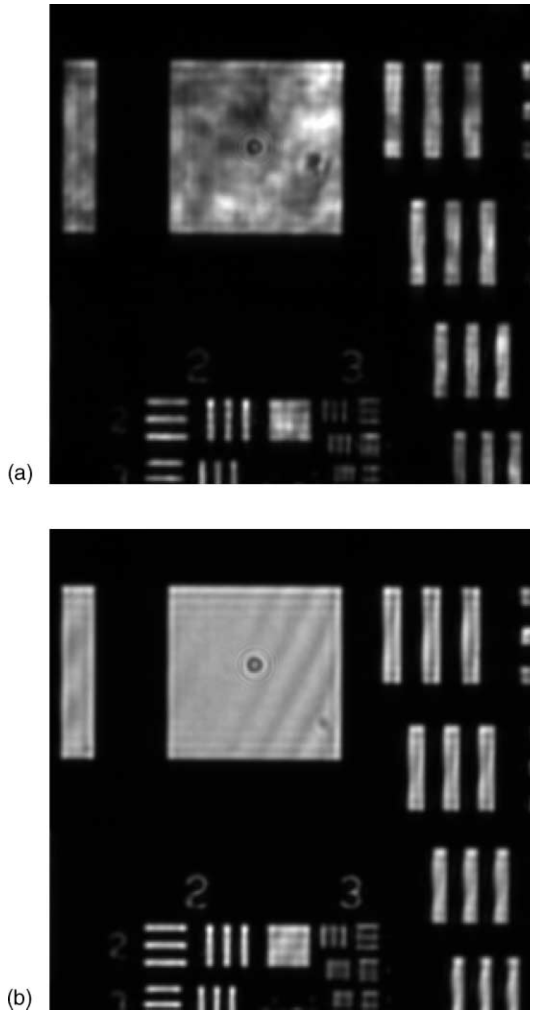

Figure 15 High temporal coherence imaging with disk sources corresponding to (a) extremely high spatial coherence ($K = 0.05$) and (b) slightly reduced but high spatial coherence ($K = 0.1$).

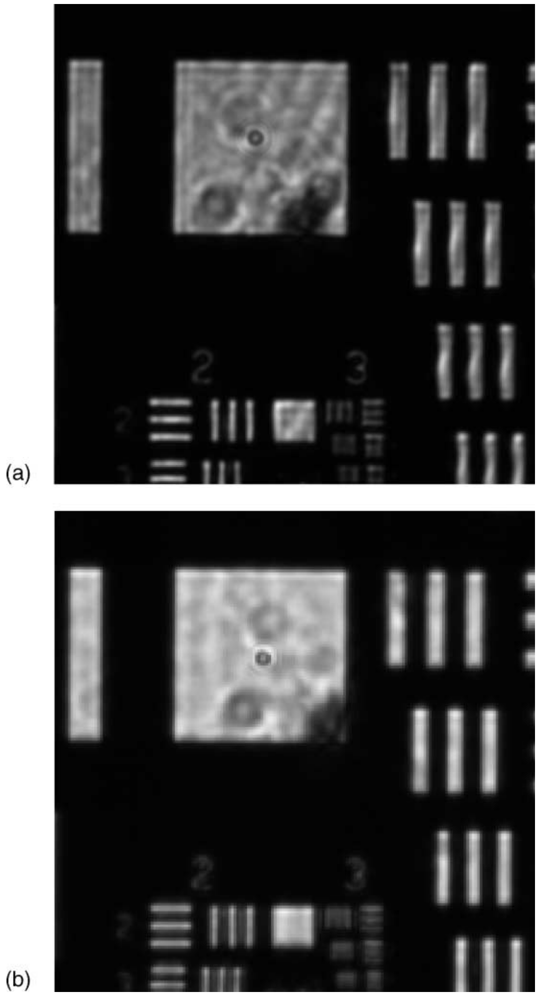

Figure 16 High spatial coherence disk illumination ($K = 0.1$) imaging through a dust covered surface with (a) narrowband laser illumination and (b) with broadband illumination.

part to imperfections in the laser beam itself and in part to unwanted dust and optically rough surfaces. Certainly, better imaging performance can be obtained with more attention to surface cleanliness and laser beam filtering, but the result shows that it can be difficult to produce high-quality coherent images in a laboratory setting.

Lower temporal coherence and lower spatial coherence will reduce the effect of speckle noise. The image of Figure 15b was obtained with laser illumination according to Figure 14a with a source mask corresponding to a source to pupil diameter ratio of $K = 0.1$. The object plane illumination is still highly temporally coherent and the spatial coherence has been reduced but is still very high. The artifact noise has been nearly eliminated by the modest reduction of spatial coherence. The presence of diagonal fringes in some regions is a result of multiple reflections produced by the cover glass in front of the CCD detector. The reflections produce a weak secondary image that is slightly displaced from the main image and the high temporal coherence allows the two images to interfere.

The noise performance was intentionally perturbed in the image of Figure 16a by inserting a slide with a modest sprinkling of dust at an optical surface in between the object and the pupil plane. The illumination conditions are the same as for the image of Figure 15b. The introduction of the dust has further degraded the image. The image of Figure 16b maintains the same spatial coherence ($K = 0.1$) but employs the broadband source. The lower temporal coherence eliminates the unwanted diagonal fringes but the speckle noise produced by the dust pattern does not improve significantly relative to the laser illumination $K = 0.1$ system. Figures 17a–c show that increasing the source size (and hence the range of angular illumination) reduces

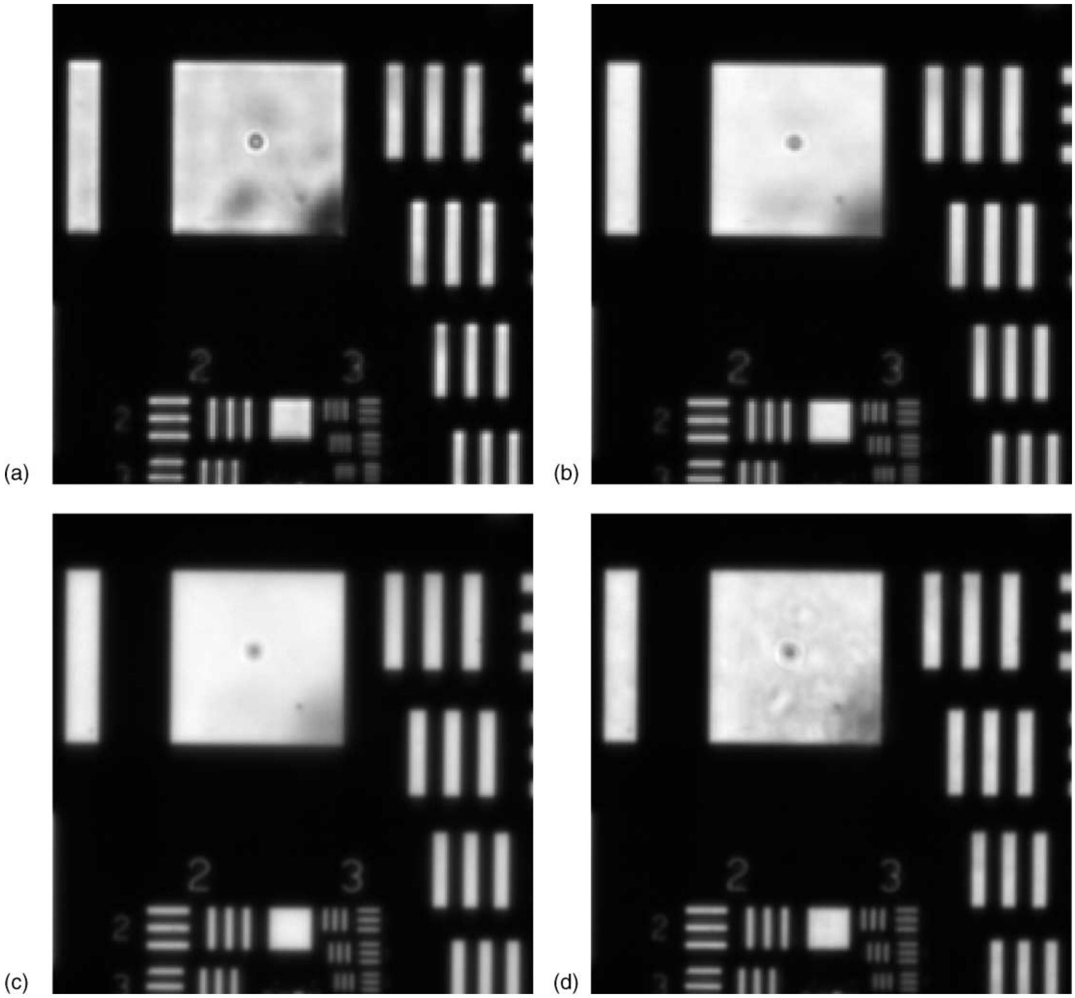

Figure 17 Low temporal coherence imaging through a dust-covered surface with various source sizes and shapes: (a) disk source with $K = 0.3$; (b) disk source with $K = 0.7$; (c) disk source with $K = 2.0$; and (d) a thin annular source with an outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

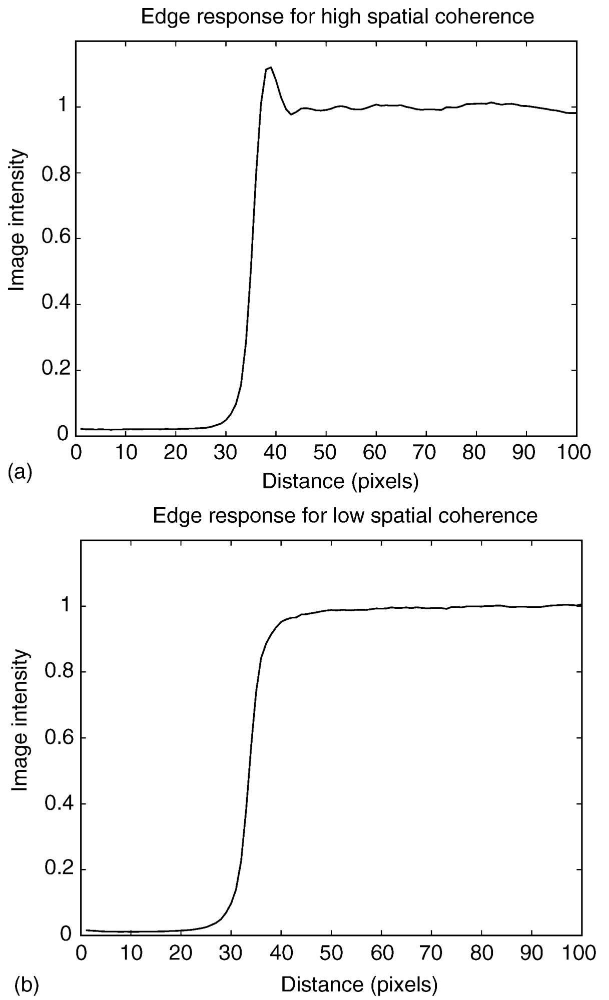

Figure 18 Slices of edge intensity images of high contrast edges taken by a laboratory system with low temporal coherence and (a) high spatial coherence (disk source with $K = 0.1$) and (b) effectively incoherent (disk source with $K = 2$).

the spatial coherence and virtually eliminates the unwanted speckle pattern. Finally, Figure 17d shows that an annular source can also provide some noise immunity. The amount of noise immunity is related to the total area of the source rather than the outer diameter of annular source. Seen from the spatial frequency model, each source point produces a noise-corrupted coherent image. Since the effect of the noise is different for each coherent image, the image plane noise averages out as the images add incoherently. This perspective suggests that an extended source provides redundancy in the transfer of object information to the image plane.

Edge Response

Spatial coherence has a strong impact on the images of edges. The sharp spatial frequency cutoff of spatially coherent imaging systems creates oscillations at the images of an edge. Figure 18 shows slices of experimentally gathered edge imagery as a function of the ratio of source diameter to pupil diameter. Higher coherence systems produce sharper edges, but tend to have overshoots. The sharper edges contribute to the sense that high-coherence systems produce higher-resolution images. One advantage to an incoherent imaging system is that the exact location of the edge corresponds to the 50% intensity location. The exact location of an edge in a partially coherent imaging system is not as easily determined. The presence or lack of edge ringing can be used to assess whether a given imaging system can be modeled as spatially incoherent or partially coherent.

Depth of Field

Coherent imaging systems exhibit an apparent increase in depth-of-field compared to spatially incoherent systems. Figure 19 shows spatially incoherent imagery for four different focus positions. Figure 20 shows imagery with the same amount of defocus produced with highly spatially coherent light. Finally, Figure 21 gives an example of how defocused imagery depends on the illumination coherence. The images of a spoke target were all gathered with a fixed amount of defocus and the source size was varied to control the illumination coherence. The images of Figure 21 differ greatly, even though the CPSF was the same for all the cases.

Digital Post-detection Processing and Partial Coherence

The model of Figure 2 suggests that post-detection image processing can be considered as part of the image formation system. Such a general view can result in imaging that might be otherwise unobtainable by classical means. In fact, microscopists routinely use complicated deblurring methods to reconstruct out-of-focus imagery and build up three-dimensional images from multiple image slices.

The coherence of the illumination should be considered when undertaking such image restoration. Rigorous restoration of partially coherent imagery is computationally intensive and requires precise knowledge of the coherence. In practice, a linear-in-intensity model for the image formation is almost always used in developing image restoration algorithms. Even nonlinear restoration algorithms have built-in linear assumptions about the image formation models which imply spatially incoherent illumination.

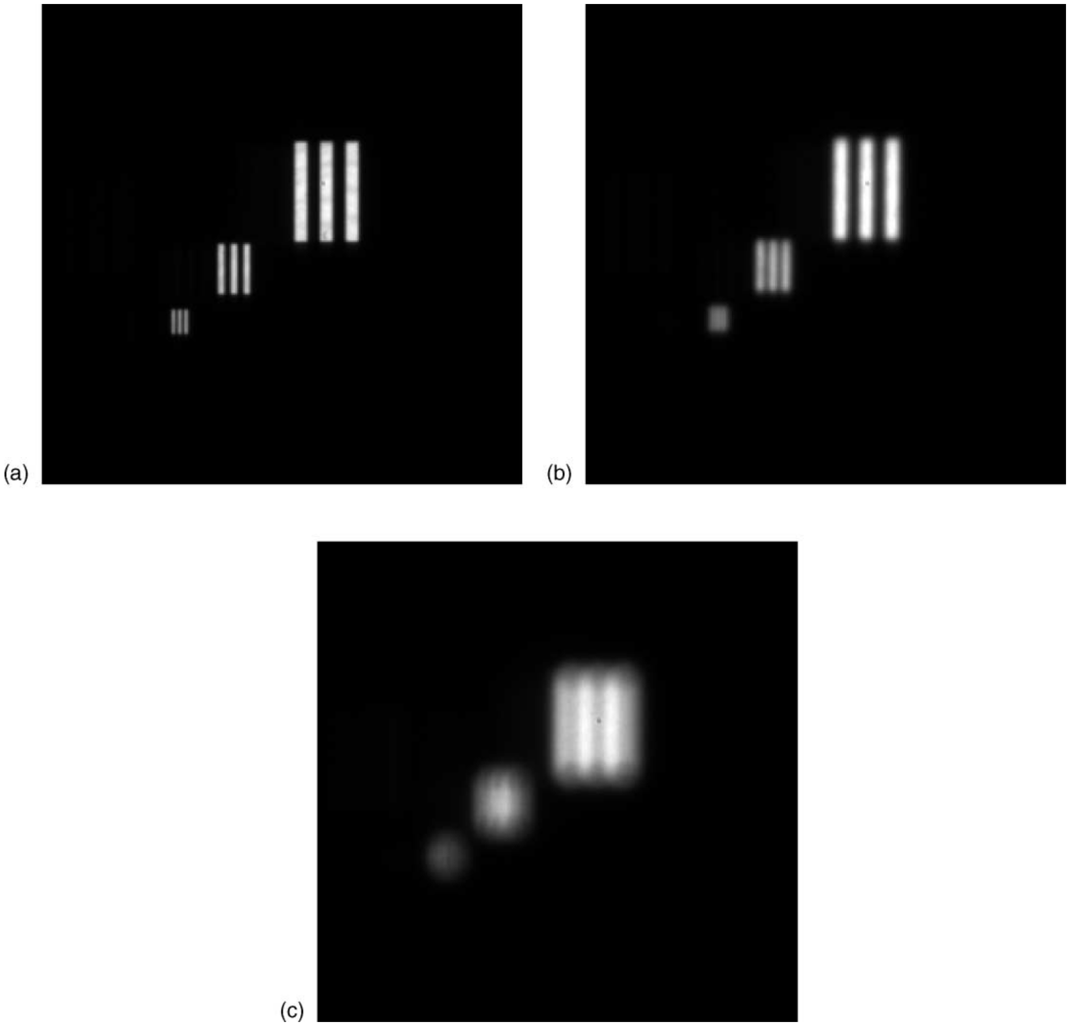

Figure 19 Spatially incoherent imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

The images of Figures 22a and b are digitally restored versions of the images of Figure 19b and Figure 20b. The point spread function was directly measured and used to create a linear restoration filter that presumed spatially incoherent illumination. The restored image in Figure 22a is faithful since the assumption of spatially incoherent light was reasonable. The restored image of Figure 22b suffers from a loss in fidelity since the actual illumination coherence was relatively high. The image is visually pleasing and correctly conveys the presence of three bar targets. However, the width of the bars is not faithful to the original target which had spacings equal to the widths of the bars.

As discussed earlier, the ideal image is not always the optimal image for a given task. In general, restoration of partially coherent imaging with an implicit spatially incoherent imaging model will produce visually pleasing images that are not necessarily faithful to the object plane intensity. While they are not faithful, they often preserve and even enhance edge information and the overall image may appear sharper than the incoherent image. When end task performance is improved, the image may be considered to be more optimal than the ideal image. It is important to keep in mind that some of the spatial information may be misleading and a more complete understanding of the coherence may be needed for precise image plane measurements. This warning is relevant in microscopy where the user is often encouraged to increase image contrast by reducing illumination spatial coherence. Annular sources can produce highly complicated spatial coherence functions that will strongly impact the restoration of such images.

Summary and Discussion

The coherence of the illumination at the object plane is important in understanding image formation.

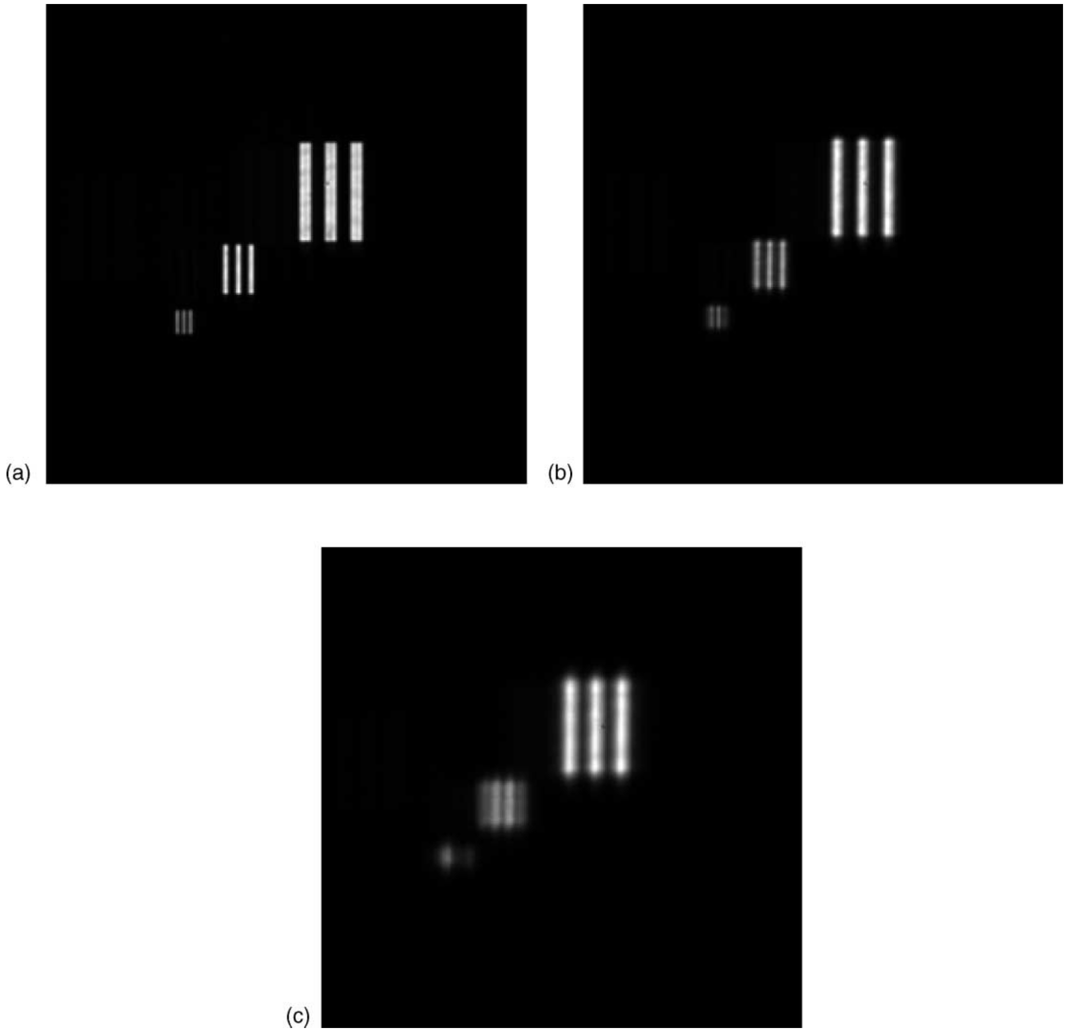

Figure 20 High spatial coherence disk source ($K = 0.2$) imaging for (a) best focus, (b) moderate misfocus, and (c) high misfocus.

Highly coherent imaging systems produce high-contrast images with high depth of field and provide the opportunity for sophisticated manipulation of the image with frequency plane filtering. Darkfield imaging and phase contrast imaging are examples of frequency plane filtering. Unfortunately, coherent systems are sensitive to optical noise and are generally avoided in practical system design. Relaxing the temporal coherence of the illumination can provide some improvement, but reducing the spatial coherence is more powerful in combating noise artifacts. Research microscopes, machine vision systems, and optical lithography systems are the most prominent examples of partially coherent imaging systems. These systems typically employ adjustable spatially incoherent extended sources in the shapes of disks or annuli. The exact shape and size of the primary source, shapes the angular extent of the illumination at the object plane and determines the spatial coherence at the object plane. Spatial coherence effects can be significant, even for broadband light. Control over the object plane spatial coherence allows the designer to find a tradeoff between the various strengths and weaknesses of coherent and incoherent imaging systems.

As more imaging systems employ post-detection processing, there is an opportunity to design fundamentally different systems that effectively split the image formation process into a physical portion and a post-detection portion. The simple example of image deblurring cited in this article shows that object plane coherence can affect the nature of the digitally restored image. The final image can be best understood when the illumination coherence effects are well understood. The spatially incoherent case results in the most straightforward model for image interpretation, but is not always the best choice since the coherence can often be manipulated to increase the contrast and hence the amount of useful information in the raw image. A completely general approach to

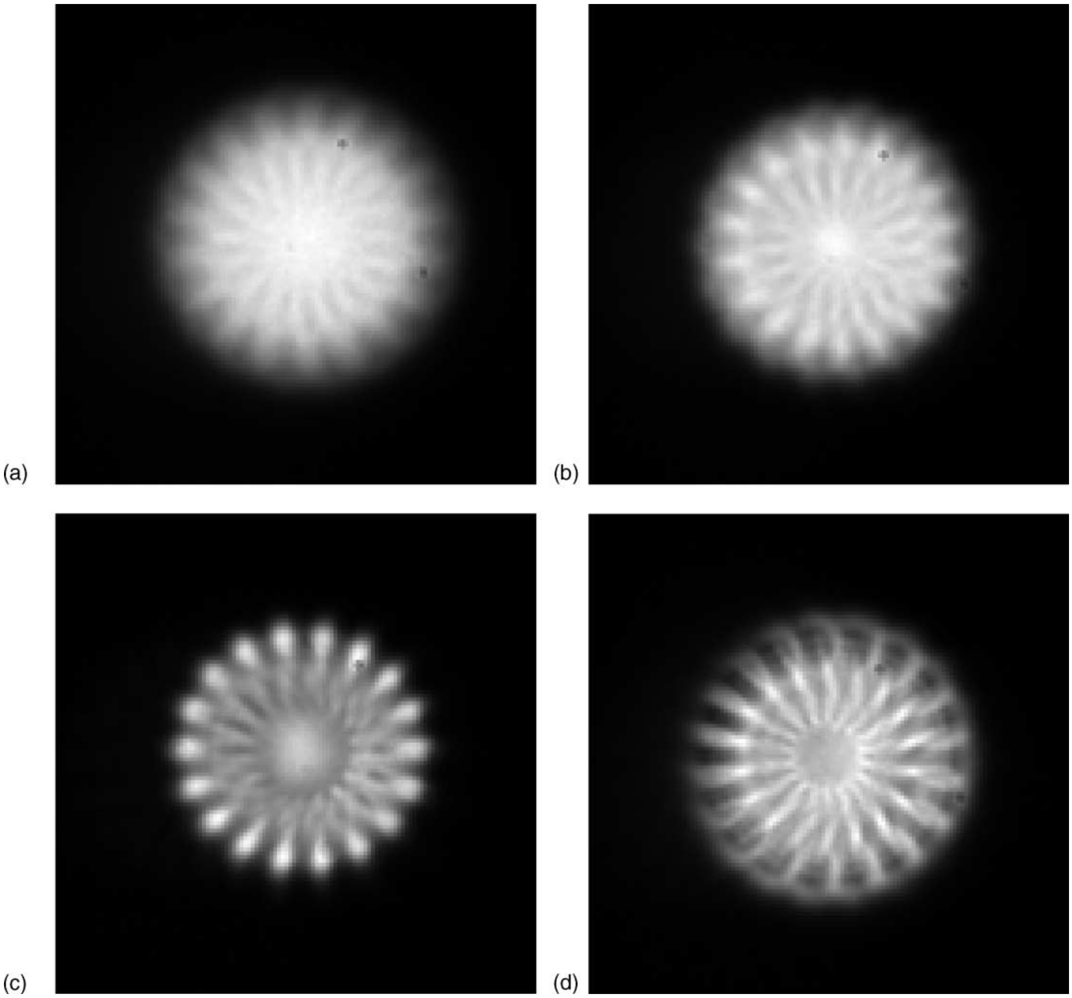

Figure 21 Imaging with a fixed amount of misfocus and varying object spatial coherence produced by: (a) highly incoherent disk source illumination ($K = 2$); (b) moderate coherence disk source ($K = 0.5$); (c) high coherence disk source ($K = 0.2$); and (d) annular source with outer diameter corresponding to $K = 0.5$ and inner diameter corresponding to $K = 0.45$.

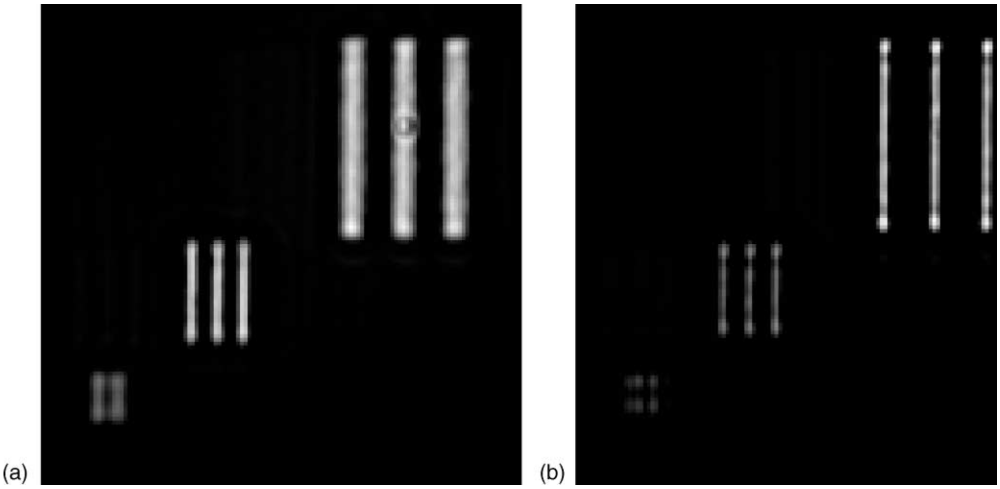

Figure 22 Digital restoration of blurred imagery with an inherent assumption of linear-in-intensity imaging for (a) highly spatially incoherent imaging ($K = 2$) and (b) high spatial coherence ($K = 0.2$).

imaging would treat the coherence of the source, the imaging optics, and the post-detection restoration all as free variables that can be manipulated to produce the optimal imaging system for a given task.

See also

Coherent Lightwave Systems. **Information Processing:** Coherent Analogue Optical Processors. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Bartelt H, Case SK and Hauck R (1982) Incoherent Optical Processing. In: Stark H (ed.) *Applications of Optical Fourier Transforms*, pp. 499–535. London: Academic Press.

Born M and Wolf E (1985) *Principles of Optics*, 6th edn. (corrected), pp. 491–554. Oxford, UK: Pergamon Press.

Bracewell RN (1978) *The Fourier Transform and its Applications*. New York: McGraw Hill.

Gonzales RC and Woods EW (1992) *Digital Image Processing*, 3rd edn., pp. 253–304. Reading, MA: Addison-Wesley.

Goodman JW (1985) *Statistical Optics*, 1st edn. New York: John Wiley & Sons.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Hecht E and Zajac A (1979) *Optics*, 1st edn., pp. 397–435, Reading, MA: Addison-Wesley.

Inoue S and Spring KR (1997) *Video Microscopy*, 2nd edn. New York: Plenum.

Reynolds GO, DeVelis JB, Parrent GB and Thompson JB (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*, 1st edn. New York: SPIE, Bellingham and AIP.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*, pp. 343–381. New York: John Wiley & Sons.

Speckle and Coherence

G Häusler, University of Erlangen-Nürnberg,
Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Opticists are aware that the amount of coherence plays a significant role in imaging systems: laser speckles are known to add significant noise to the image, as well as parasitic interferences from dusty lenses. Optical systems are often called coherent, if a laser is used (right), and incoherent if other sources come into play (wrong). Many users of optical systems are unaware that it is not the high temporal coherence but the spatial coherence that commonly afflicts the image quality, and that this parasitic spatial coherence is ubiquitous, even though not obvious. Coherent artifacts can occur without the use of lasers, although speckle noise is more prominent with lasers. Even opticists sometimes underestimate the damage that residual coherent noise can cause, and as laser oriented sensor funding programs are 'en vogue', nonexperts are disappointed if some metrology device does not include a laser.

This encyclopedia addresses the many uses of lasers. In this article, we will discuss the costs of coherence. The commonly pretended incoherent approach of everyday optics may lead to significant quantitative measuring errors of illumination or reflectivity, 3d shape, distance or size. Spatial coherence is the dominant source of noise. We will give some rules of thumb to estimate these errors and a few tricks to reduce coherent noise. These rules will help to minimize coherent noise, however, it turns out that as spatial coherence is ubiquitous, there are only limited options available to clear it. One of the options to build good sensors that measure shape, reflectivity, etc. is avoiding the use of lasers!

To become familiar with some basics of the theory of coherence we refer the reader the Further Reading section at the end of this article.

Coherence can be boon or disaster for the opticist, as is explained in other articles of this encyclopedia about interferometry, diffraction, and holography. A specific topic is information acquisition from coherently scattered light. An enlightening example is where speckles in white light interferometry at rough surfaces and in speckle interferometry are exploited. We will briefly discuss white light interferometry at rough surfaces in the section on speckles as carriers of information below.

Practical Coherence Theory

A major issue of this chapter will be corrupting properties of coherence in the daily life of an optical metrologist. We will demonstrate that 'speckle' noise is ever present, and essentially unavoidable, in the images of (diffusively reflecting) objects. Its influence on the quality of optical measurements leads to a lower limit of the physical measuring uncertainty.

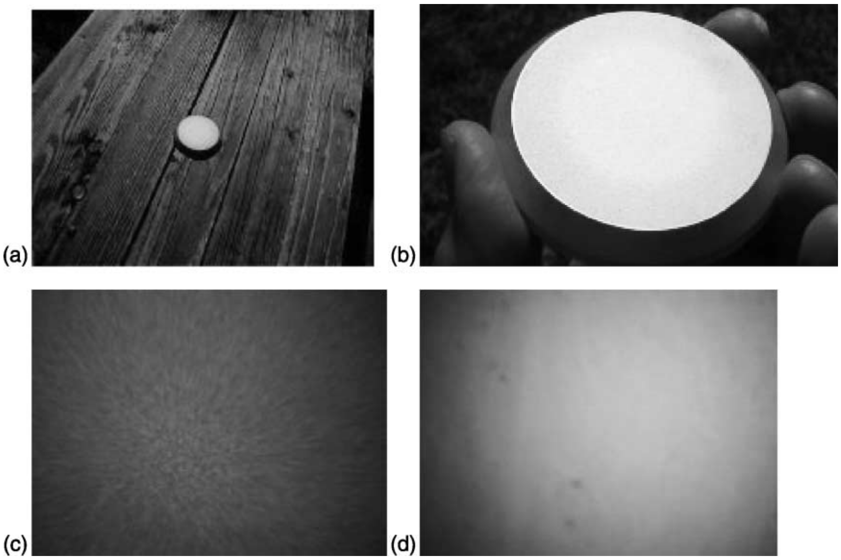

Figure 1 Ground glass in sunlight: (a) the image of a ground glass, illuminated by the sun, observed with small aperture (from large distance), displays no visible speckles; (b) a medium observation aperture (smaller distance) displays weak visible speckles; (c) observation with very high aperture displays strong contrast speckles; and (d) the image of the ground glass on a cloudy day does not display speckles, due to the very big aperture of illumination. Reproduced with permission from Häusler G (1999) Optical sensors and algorithms for reverse engineering. In: Donah S (ed.) Proceedings of the 3rd Topical Meeting on Optoelectronic Distance/Displacement Measurements and Applications. IEEE-LEOS.

This 'coherent noise' limit is, not surprisingly, identical to Heisenberg's limit. We will start by summarizing the important results, and will give some simple explanations later.

Some Basic Observations

(i) The major source of noise in optical measurements is spatial coherence, the temporal coherence status, generally, does not play a significant role.

(ii) An optically rough surface displays subjective speckles with contrast $C = 1$, if the observation aperture $\sin u_o$ is larger than the illumination aperture $\sin u_q$.

(iii) The speckle contrast is $C \sim \sin u_q / \sin u_o$, if the observation aperture is smaller than the illumination aperture.

Figure 1 illustrates the situation by a simple experiment.

The results of Figure 1 can be explained simply by summarizing some 80 pages of coherence theory in a nutshell, by useful simplifications and approximations (Figure 2).

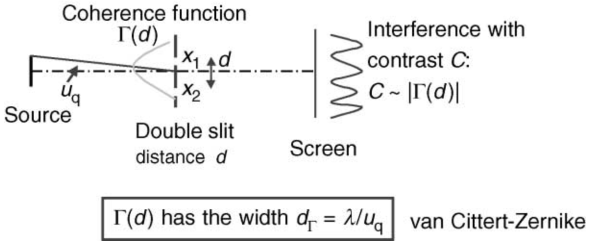

Figure 2 Basic experiment for spatial coherence. An extended source, such as the sun or some incandescent lamp illuminates a double slit, from a large distance. On a screen behind the double slit we can observe 'double slit interference', if the waves coming from x_1 and x_2 display coherence. The contrast of the interference fringes is given by the magnitude of the coherence function Γ (neglecting some normalization factor). Γ is a function of the two slit locations x_1 and x_2.

If the source is at an infinite distance, the coherence function no longer depends on the two variables but just on the slit distance $d = |x_1 - x_2|$. In this case, the coherence function $\Gamma(d)$ can be easily calculated as the Fourier transform of the spatial intensity distribution of the source (van Cittert-Zernike theorem). Let us assume a one dimensional source with an angular size of $2u_q$, where u_q is commonly called the aperture of the illumination. In this case, the coherence function will be:

$$\Gamma(d) \sim \sin c(2u_q d/\lambda) \qquad [1]$$

The width of the coherence function which gives the size of the coherently illuminated area (coherence area) can be approximated from eqn [1]:

$$d_\Gamma = \lambda / u_q \qquad [2]$$

Equation [1] includes some approximations: for circular sources such as the sun, we get an additional factor 1.22 for the width of the coherence function – the ‘sinc-function’ has to be replaced by the Bessel function $J_1(r)/(r)$. There is another approximation: we should replace the aperture angle u_q by $\sin u_q$, for larger angles. We choose the wavelength of the maximum sensitivity of the eye, at $\lambda = 0.55\ \mu\text{m}$, which is the wavelength of the maximum emission of the sun also.

In conclusion, if two points at the object are closer than d_Γ these points will be illuminated with a high degree of coherence. The light waves, scattered from these object-locations can display interference contrast. Specifically, they may cancel each other out, causing a low signal-to-noise ratio of only 1:1. This does not happen if the points have a distance larger than d_Γ.

From Figure 3 we learn that the width of the spatial coherence function from sunlight illumination at the earth’s surface is ~110 μm. (Stars more distant than the sun appear smaller, and hence have a wider coherence function, at the earth’s surface. Michelson was able to measure the angular size of some close stars, by measuring the width of this coherence function, which was about $d_\Gamma \sim 10$ m.)

Figure 4 again displays an easy to perform experiment. We can see speckles at sunlight illumination. We can observe speckles as well in shops because they often use small halogen spot illumination.

So far we have only discussed the first stage (from the source to the object). We still have to discuss the way from the object to the retina (or CCD-chip of a video camera). This is illustrated by Figure 5.

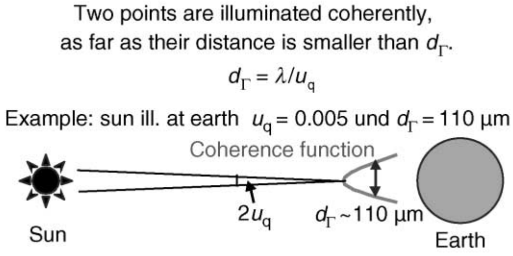
Two points are illuminated coherently,
as far as their distance is smaller than d_Γ.
$d_\Gamma = \lambda/u_q$
Example: sun ill. at earth $u_q = 0.005$ und $d_\Gamma = 110\ \mu\text{m}$
Coherence function
Sun
$2u_q$
$d_\Gamma \sim 110\ \mu\text{m}$
Earth

Figure 3 Coherence from sunlight illumination. With the van Cittert–Zernike theorem and an approximate illumination aperture of the sun $u_q^{\text{sun}} \sim 0.005$ (0.25°), we get a width of the coherence function at the Earth’s surface of about $d_\Gamma \sim 110\ \mu\text{m}$.

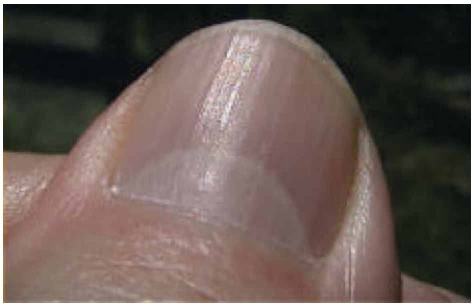

Figure 4 Finger nail in sunlight, with speckles.

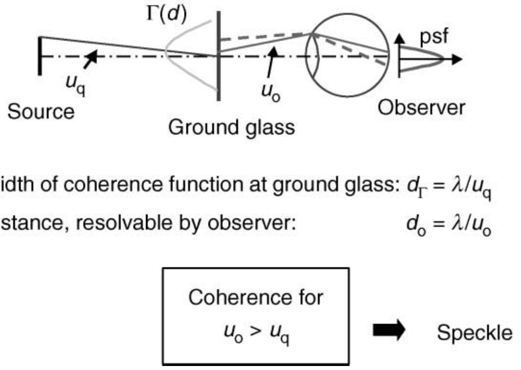
$\Gamma(d)$
psf
u_q
u_o
Source
Ground glass
Observer
• Width of coherence function at ground glass: $d_\Gamma = \lambda/u_q$
• Distance, resolvable by observer: $d_o = \lambda/u_o$
Coherence for $u_o > u_q$
Speckle

Figure 5 Coherence in a nutshell: spatial coherence and observer.

As an object, we choose some diffusely reflecting or transmitting surface (such as a ground glass). The coherence function at the ground glass has again a width of d_Γ. What happens if we image this ground glass onto our retina? Let us assume our eye to be diffraction limited (which, in fact, it is at sunlight conditions, where the eye pupil has only a diameter $\Phi_{\text{pupil}} = 2.5$ mm diameter or less). Then a point x_1 at the ground glass will cause a diffraction spot at the retina, with a size of:

$$d'_{\text{diffr}} = \lambda/u'_o = 8\ \mu\text{m} \qquad [3]$$

The image sided aperture of observation u'_0 of the eye is calculated from Φ_{pupil} and from its focal length $f_{\text{eye}} = 18$ mm as $u'_o = \Phi_{\text{pupil}}/2f_{\text{eye}} = 0.07$. If we project the diffraction spot at the retina back onto the object, we get its size d from:

$$d_{\text{diffr}} = \lambda/u_o = \lambda/(\Phi_{\text{pupil}}/2z_o) \qquad [4]$$

with z_o = observation distance of the object from the eye. We call $u_o = \Phi_{\text{pupil}}/2z_o$ the object sided observation aperture. Let us calculate the laterally resolvable distance d_{diffr} at the object, with a distance $z_o = 250$ mm, which is called the ‘comfortable viewing distance’:

$$d_{\text{diffr}}(z_o = 250\text{ mm}) = 110\ \mu\text{m} \qquad [5]$$

After these preparations we come to the crucial issue: how does the image at the retina appear if we cannot resolve distances at an object smaller than d_Γ, and how does it appear if the resolution of the eye is sufficient to resolve distances smaller than d_Γ?

Let us start with the first assumption: $d_{\text{diffr}} \gg d_\Gamma$. Now the images of points at the object, over an area of the diffraction spot are more or less incoherently averaged at the retina, so we will see little interference contrast or ‘speckles’. From eqn [4] we see that at this incoherent averaging starts for distances larger than 250 mm, which is the comfortable viewing distance. Adults cannot see objects from much closer distances. We generally do not see speckles, not just at the limit of maximum accommodation. For larger distances, if we take $z_o = 2500$ mm, the resolvable distance at the object will be 1.1 mm, which is 10 times larger than the diameter of the coherence area. Averaging over such a large area will drop the interference contrast by a factor of 10. Note that such a small interference contrast might not be visible, but it is not at zero!

Coming to the second assumption we can laterally resolve distances smaller than d_Γ. In order to understand this, we first have to learn what is an 'optically rough' object surface. Figure 6 illustrates the problem.

A surface commonly is called 'rough' if the local height variations are larger than λ. However, a surface appears only rough, if the height variation within the distance that can be resolved by the observer, is larger than λ (reflected light will travel the distance twice, hence a roughness of $\lambda/2$ will be sufficient). Then the scattered different phasors, or 'Huygens elementary waves':

$$u_s \sim \exp[i2kz(x, y)] \qquad [6]$$

scattered from the object at x, y may have big phase differences so that destructive interferences (and speckle) can occur in the area of the diffraction image. So the attribute 'rough' depends on the object as well as on the observation aperture. With a high observation aperture (microscope), the diffraction image is small and within that area the phase differences might be small as well. So a ground glass may then look locally like a mirror, while with a small observation aperture it appears 'rough'.

With the 'rough' observation mode assumption, we can summarize what we have assumed: with resolving distances smaller than the coherence width d_Γ we will see high interference contrast; in fact we see a speckle contrast of $C = 1$, if the roughness of the object is smaller than the coherence length of the source. This is the case for most metal surfaces, for ground glasses, or for worked plastic surfaces. The assumption is not true for 'translucent' surfaces, such as skin, paper, or wood. This will be discussed below.

The observation that coherent imaging is achieved, if we can resolve distances smaller than the coherence width d_Γ, is identical to the simple rule that fully coherent imaging occurs if the observation aperture u_o is larger than the illumination aperture u_q.

As mentioned, we will incoherently average in the image plane, over some object area, determined by the size of the diffraction spot. According to the rules of speckle-averaging, the speckle contrast C decreases with the inverse square root of the number N of incoherently averaged speckle patterns. This number N is equal to the ratio of the area A_{diff} of the diffraction spot divided by the coherence area $A_\Gamma = d_\Gamma^2$. So we obtain for the speckle contrast C:

$$C = 1 \quad \text{for } u_q < u_o \qquad [7a]$$

$$C = 1/\sqrt{(N)} = u_o/u_q, \quad \text{for } u_q >= u_o \qquad [7b]$$

We summarize the results in Figure 7.

Equation [7b] has an interesting consequence: we never get rid of speckle noise, even for large illumination apertures and small observation apertures. In many practical instruments, such as a slide projector, the illumination 'aperture stop' cannot be greater than the imaging lens. Fortunately, the observer's eye commonly has a pupil smaller than that of the projector, and/or looks from a distance at the projection screen. Laser projection devices however cause strong and disturbing speckle effects for the user and significant effort is invested to cope with this effect.

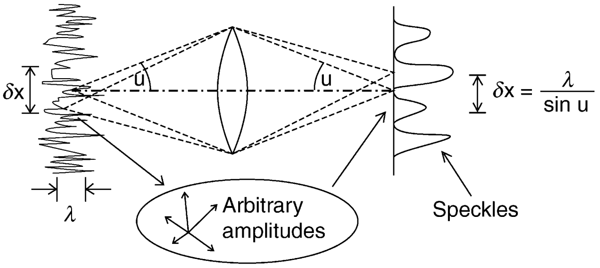

Elementary phasors have phase differences bigger than +/– 180°. Destructive interference may occur.

Figure 6 What is a rough surface?

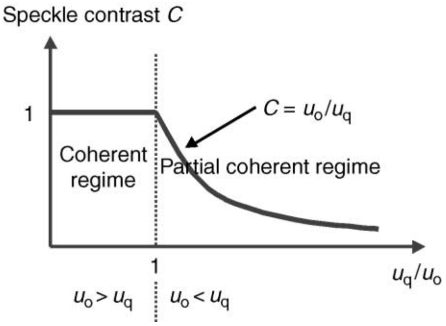

Figure 7 Coherence in a nutshell: what are the conditions for incoherent and coherent imaging?

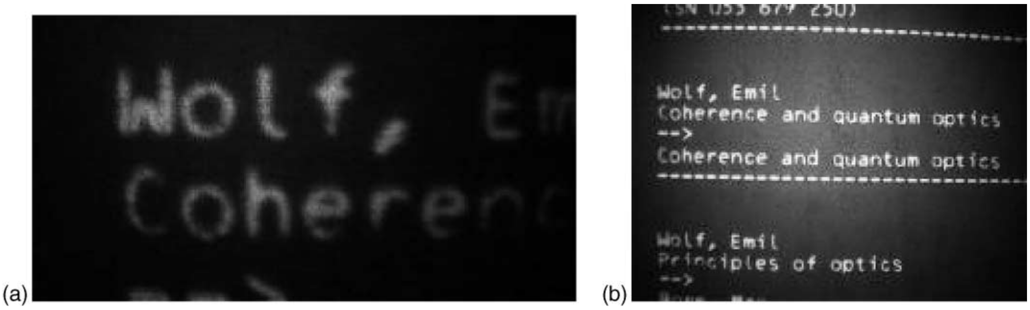

Figure 8 Microfiche projection. For high observation aperture (a) strong speckles occur, and for small observation aperture (b), the speckle contrast is low.

In Figure 8, an image is depicted from a microfiche reading projector. Figure 8a displays the close-up look (high observation aperture) and Figure 8b depicts a more distant view (small observation aperture). The amount of noise is much less in the second case, as it should be according to the rules of eqn [7].

Let us finish this section with some speculation. How would a fly see the sunny world, with a human type of eye? With high observation aperture and at short distance, the world is full of speckle noise. Fortunately, nature invented the facet-eye, for insects, as depicted in Figure 9.

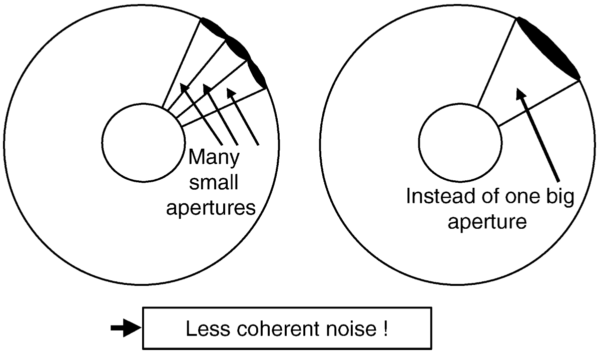

Figure 9 Facet eye and ‘human’ eye. The facet eye consists of many low aperture lenses, in contrast to the human eye.

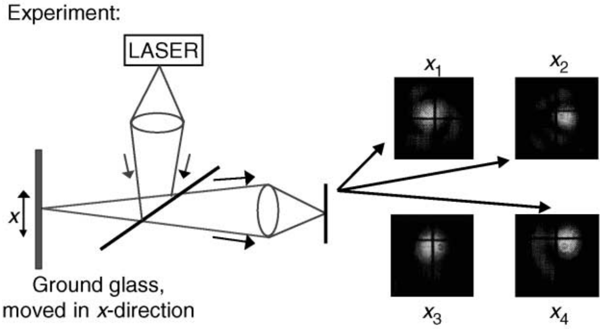

Figure 10 Coherence makes it impossible to localize objects with very high accuracy. Laser spots projected onto a ground glass cannot be imaged without some uncertainty of the lateral position. The four pictures of the spot images, taken with different lateral positions of the ground glass, display the cross bar (true position) and the apparent position of the spot images. The position uncertainty is equal to the uncertainty calculated from Heisenberg's uncertainty principle.

Speckle Limits of Metrology

The consequences of the effects discussed above are often underestimated. We should be suspicious, if some effect in nature that disturbs us – like our coherent noise – is ubiquitous, that there might be some deep underlying principle that does not allow us to know everything about the object under observation. Indeed, it turns out that Heisenberg's uncertainty principle is strongly connected with coherent noise (Figure 10).

We can see this from the following experiment: a laser spot is projected onto a ground glass, and imaged with high magnification by a video camera, with an aperture sin u_o. The ground glass is macroscopically planar. When the ground glass is laterally shifted, we find that the observed spot is ‘dancing’ at the video target. Its observed position is not constant, although its projected position is. It turns out that the standard deviation of the observed position is equal to the uncertainty calculated from the aperture by Heisenberg's principle.

We can calculate the limit for the distance measuring uncertainty, from speckle theory, as well from Heisenberg's principle (within some factor of order 1):

$$\delta z \cdot \delta p_z \gg h/4\pi \qquad [8]$$

where δp_z is the uncertainty of the photon impulse h/λ in the z-direction (along the optical axis of the measuring system) and h is Planck's constant. For a small measurement uncertainty of the distance, we should allow a big uncertainty δp_z. This can be achieved by a large aperture of the observation system, giving the photons a wide range of possible directions to the lens. We can also allow different wavelengths, and come to white light interferometry, see section on speckles as carriers of information below. The result is – not surprising – the same as Rayleigh scattering found for the depth of field:

$$\delta z = \lambda / \sin^2 u_{\mathrm{o}} \qquad [9]$$

Coherent noise is the source of the fundamental limit of the distance measuring uncertainty δz of triangulation based sensors.

$$\delta z = C \cdot \lambda / (\sin u_{\mathrm{o}} \sin \theta) \qquad [10]$$

where C is the speckle contrast, λ is the wavelength, $\sin u_{\mathrm{o}}$ is the aperture of observation, and θ is the angle of triangulation (between the direction of the projection and the direction of observation). For a commercially available laser triangulation sensor, (see Figure 11), with $\sin u_{\mathrm{o}} \sim 0.01, C = 1$ and $\theta = 30°$, the measuring uncertainty δz will be larger than 100 μm. We may add, that for $\sin u_{\mathrm{o}} = \sin \theta$, which is valid for auto-focus sensors (such as the confocal scanning microscope), eqn [10] degenerates to the well known Rayleigh depth of field (eqn [9]).

The above results are remarkable in that we cannot know the accurate position of the projected spot or about the position of an intrinsic reflectivity feature; we also cannot know the accurate local reflectivity of a coherently illuminated object. A further consequence is that we do not know the accurate shape of an object in 3D-space. We can calculate this 'physical measuring uncertainty' from the considerations above.

These consequences hold for optically rough surfaces, and for measuring principles that exploit the local intensity of some image, such as with all triangulation type of sensors. The consequences of these considerations are depressing: triangulation with a strong laser is not better than triangulation with only one single photon. The deep reason is, that all coherent photons stem from the same quantum mechanical phase cell and are indistinguishable. Hence, many photons do not supply more information than one single photon.

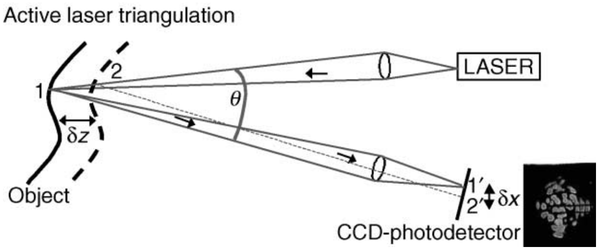

Figure 11 Principle of laser triangulation. The distance of the projected spot is calculated from the location of the speckled spot image and from the angle of triangulation. The speckle noise introduces an ultimate limit of the achievable measuring uncertainty. Reproduced with permission from *Physikalische Plätter* (May 1997): 419.

Why Are We Not Aware of Coherence

Generally, coherent noise is not well visible, even in sophisticated technical systems the visibility might be low. There are two main reasons; first, the observation aperture might be much smaller than the aperture of illumination. This is often true, for large distance observation, even with small apertures of illumination. The second reason holds for technical systems, if the observation is implemented via pixelized video targets. If the pixel size is much larger than a single speckle, which is commonly so, then, by averaging over many speckles, the noise is greatly reduced. However, we have to take into account that we pay for this by loss of lateral resolution $1/\delta x$. We can formulate another uncertainty relation:

$$\delta x > \lambda / (C \cdot \sin u_{\mathrm{o}}) \qquad [11]$$

which says that if we want to reduce the speckle contrast C (noise), by lateral averaging, we can do this but lose lateral resolution $1/\delta x$.

Can We Overcome Coherence Limits?

Since the daily interaction of light with matter is coherent scattering, we can overcome the limit of eqn [10] only by looking for measuring principles that are not based on the exploitation of local reflectivity, i.e., on 'conventional imaging'. Concerning optical 3D-measurements, the principle of triangulation uses the position of some image detail, for the calculation of the shape of objects, and has to cope with coherent noise.

Are there different mechanisms of photon–matter interaction, with noncoherent scattering? Fluorescence and thermal excitation are incoherent mechanisms. It can be shown (see Figure 12) that triangulation utilizing fluorescent light, displays much less noise than given by eqn [10]. This is exploited in fluorescent confocal microscopy.

Thermally excited matter emits perfectly incoherent radiation as well. We use this incoherence to measure the material wear in laser processing (see Figure 13), with, again, much better accuracy than given by eqns [9,10].

The sensor is based on triangulation. Nevertheless, by its virtually zero coherent noise, it allows a measuring uncertainty which is limited only by camera noise and other technical imperfections. The uncertainty of the depth measurement through the narrow aperture of the laser nozzle is only a few

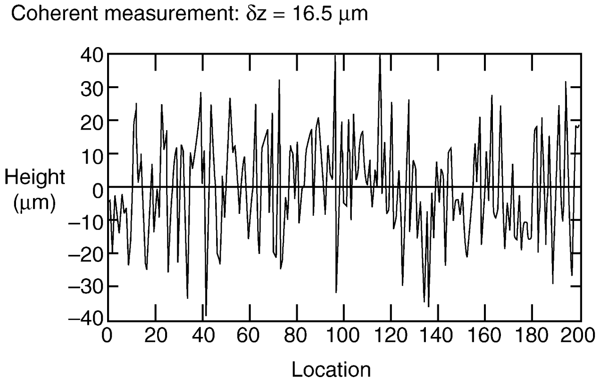

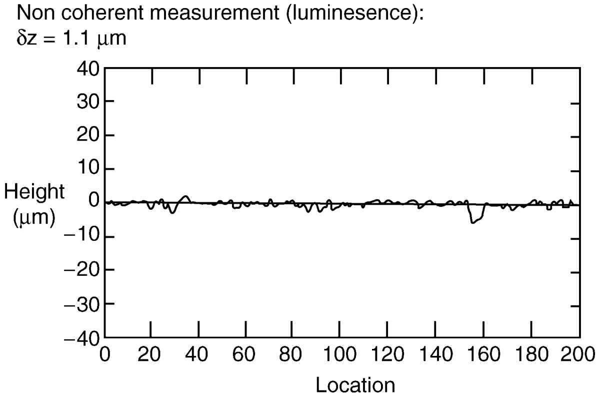

Figure 12 Reduction of speckle noise by triangulation with fluorescent light. A flat metal surface is measured by laser triangulation. Top: surface measured by laser triangulation with full speckle contrast. Bottom: the surface is covered with a very thin fluorescent layer, and illuminated by the same laser, however, the triangulation is done after suppressing the scattered laser light, and utilizing the (incoherent!) fluorescent light. This experiment proves that it is the coherent noise that causes the measuring uncertainty.

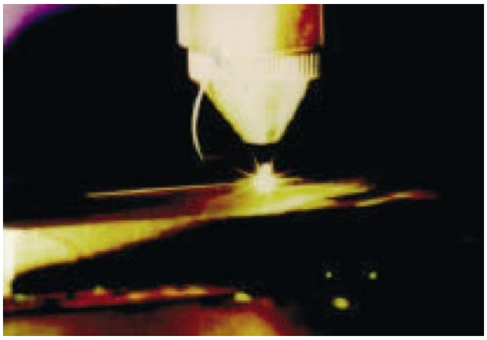

Figure 13 A 100 W laser generates a plasma that emits perfectly incoherent radiation at the object surface. The emitted light can be used to measure on-line the distance of the object surface. A triangulation sensor, exploiting the incoherent plasma emission, controls surface ablation with an accuracy of only a few microns. Reproduced with permission from F & M, *Feinwerktechnik Mikrotechnik Masstechnik* (1995) Issue 9.

microns, beating the limitation of eqn [10] by a factor of about 5, even in the presence of a turbulent plasma spot with a temperature of 3000 K.

Broadband Illumination

Speckle noise can hardly be avoided by broadband illumination. The speckle contrast C is related to the surface roughness σ_z and the coherence length l_c by eqn [12]:

$$C^2 \sim l_c/2\sigma_z \quad \text{for} \quad \sigma_z \gg l_c \qquad [12]$$

So, only for very rough surfaces, and white light illumination, C can be reduced. However, the majority of technical objects is smooth, with a roughness of a few micrometers, hence speckles can be observed even with white light illumination ($l_c \sim 3$ μm). This situation is different for ‘translucent’ objects such as skin, paper, or some plastic material.

Speckles as a Carrier of Information

So far we discussed the problems of speckles when measuring rough surfaces. We can take advantage of speckles, if we do not stick to triangulation as the basic mechanism of distance measurement. We may utilize the fact that although the phase within each speckle has a random value, this phase is constant over some area. This observation can be used to build an interferometer that works even at rough surfaces. So, with proper choice of the illumination aperture and observation aperture, and with a surface roughness less than the coherence length, we can measure distances by the localization of the temporal coherence function. This function can easily be measured by moving the object under test along the optical axis and measuring the interference intensity (correlogram) at each pixel of the camera. The signal generating mechanism of this ‘white light interferometer’ distance measurement is not triangulation but ‘time of flight’. Here we do not suffer from the limitations of eqn [10]. The limits of white light rough surface interferometry are discussed in the Further Reading section at the end of this article.

The ‘coherence radar’ and the correlograms of different speckles are depicted in Figures 14 and 15.

The physical mechanism of the signal formation in white light interferometry at rough surfaces is different from that of white light interferometry at smooth surfaces. That is why the method is sometimes referred as ‘coherence radar’. We will briefly summarize some advantageous, and – probably unexpected – features of the coherence radar.

We call the real shape of the object, or the surface profile, $z(x, y)$, where z is the local distance at the position (x, y). We give the measured data an index ‘m^2’:

(i) The physical measuring uncertainty of the measured data is independent from the observation aperture. This is quite a remarkable property, as it enables us to take accurate measurements at the bottom of deep boreholes.

(ii) The standard deviation σ_z of the object can be calculated from the measured data z_{m}: $\sigma_z = <|z_{\mathrm{m}}|>$. According to German standards, σ_z corresponds to the roughness measure R_{q}.

This measure R_{q} was calculated from measurements at different roughness gauges, with different observation apertures. Two measurements are depicted in Figure 16.

The experiments shown in Figure 16 display the correct roughness measure, even if the higher frequencies of the object spectrum $Z(\nu, \mu)$ (Figure 17), are not optically resolved. The solution of this paradox might be explained as follows.

From the coherently illuminated object each point scatters a spherical wave. The waves scattered from different object points have, in general, a different phase. At the image plane, we see the laterally

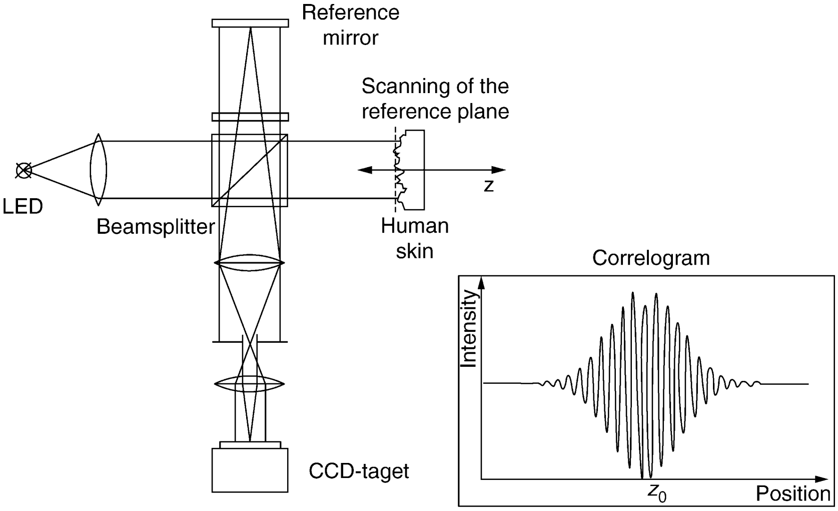

Figure 14 ‘Coherence radar’, white light interferometry at rough surfaces. The surface under test is illuminated with high spatial coherence and low temporal coherence. We acquire the temporal coherence function (correlogram) within each speckle, by scanning the object under test in depth.

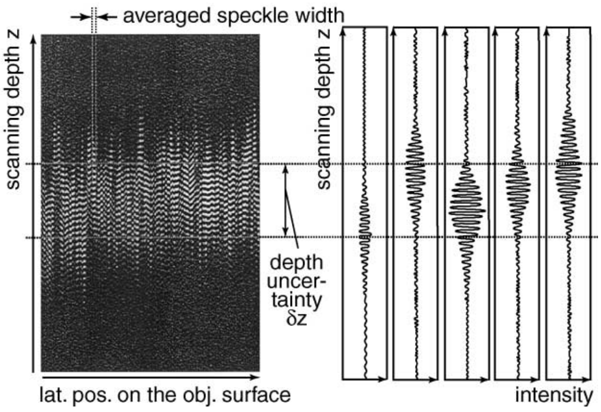

Figure 15 White light speckles in the $x-z$ plane. The left side displays the acquired correlograms, for a couple of speckles, the right side shows the graphs of some of the correlograms. We see that the correlograms are not located at a constant distance, but display some ‘distance uncertainty’. This uncertainty does not originate from the instrument, but from the roughness of the object.

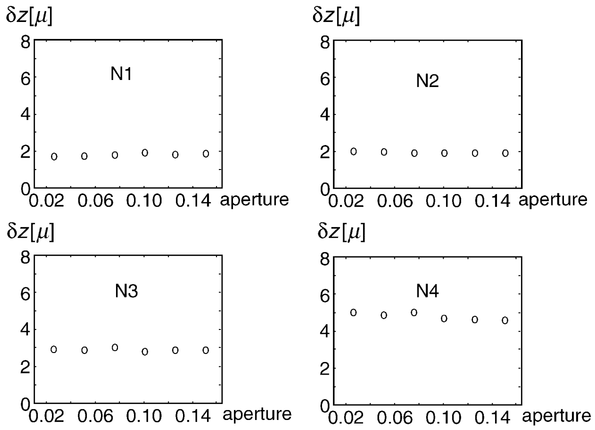

Figure 16 Optical measurement of the roughness beyond the Abbe resolution limit. The roughness is measured for 4 roughness standards N1–N4, with different observation apertures. The roughness is measured correctly, although the microtopology is not resolved by the observing optics.

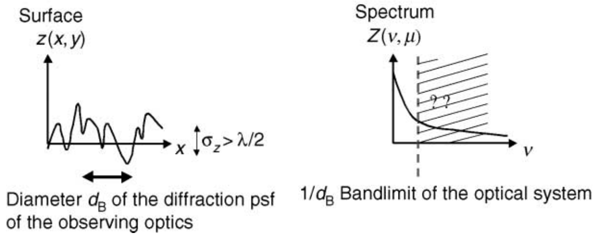
Surface
$z(x, y)$
x
$\sigma_z > \lambda/2$
Diameter d_B of the diffraction psf of the observing optics
Spectrum
$Z(\nu, \mu)$
??
ν
$1/d_B$ Bandlimit of the optical system

Figure 17 Signal generation beyond the Abbe limit. The microtopology of the object $z(x, y)$ is not resolved, due to the large diffraction image with diameter d_B. Yet we get information from beyond the bandlimit $1/d_B$ of the observing optics.

averaged complex amplitude $<u>$, according to eqn [6]. The averaging is due to the diffraction limited resolution.

Equation [6] reveals that the (spatially) averaged field amplitude $<u>$ is a nonmonotonic, nonlinear function of the surface profile $z(x, y)$, if the surface is rough (i.e., $z \gg \lambda$). We do not average over the profile z but over the complex amplitude. As a consequence of this nonlinearity, the limited observation aperture does not only collect the spatial frequencies ν, μ of the spatial Fourier spectrum $Z(\nu, \mu)$ within the bandlimit $1/d_B$, but also acquires information beyond this bandlimit. The reason is the 'down conversion' of higher spatial frequencies by nonlinear mixing. Simulations and experiments confirm the consequences.

Thus we can evaluate the surface roughness, even without laterally resolving the microtopology. However, there is a principal uncertainty on the measured data. We do not see the true surface but a surface with a 'noise' equivalent to the roughness of the real surface.

Nonlinear nonmonotonic nonlinearities in a sequence of operations may cause 'chaotic' behavior, in other words, small parameter changes such as vibrations, humidity on the surface, etc. may cause large differences of the outcome. In fact, we observe a significant variation of the measured data $z_m(x, y)$ within a sequence of repeated measurements. Our conjecture is that the complex signal formation may be involved in this irregular behavior. The hypothesis may be supported by the observation that much better repeatability can be achieved at specular surfaces, where eqn [6] degenerates to a linear averaging over the surface profile:

$$<u> \sim 1 + ik <z(x, y)>, \quad \text{for } z \ll \lambda \qquad [13]$$

Summary

Spatial coherence is ubiquitous and unavoidable. Spatial coherence disturbs most optical measurements, where optically rough surfaces exist. Then, coherence gives the ultimate limit of the achievable measuring uncertainty. There are simple rules to estimate the measuring uncertainty, by calculating the width of the coherence function and the resolution of the observation system. Spatial coherence and Heisenberg's uncertainty principle lead to the same results of measuring uncertainty.

On the other hand – the formation of signals from coherent light that is scattered at rough surfaces is quite complex and a strongly nonlinear process, which sometimes might encode information which is not otherwise available. One example appears to be the measurement of the roughness with white light interferometry – which is possible, even if the microtopology of the surface is not optically resolved.

See also

Coherence: Overview; Coherence and Imaging.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Brooker G (2003) *Modern Classical Optics*. Oxford: Oxford University Press.

Dainty JC (ed.) (1984) *Laser Speckle and Related Phenomena*. Berlin, Heidelberg, New York: Springer Verlag.

Dorsch R, Herrmann J and Häusler G (1994) Laser triangulation: fundamental uncertainty in distance measurement. *Applied Optics* 33: 1306–1314.

Dresel T, Venzke H and Häusler G (1992) 3D-sensing of rough surfaces by coherence radar. *Applied Optics* 31: 919–925.

Ettl P (2002) *Über die Signalenstehung bei Weisslichtinterferometrie*. PhD Dissertation, University of Erlangen-Nürnberg.

Häusler G, Ettl P, Schenk M, Bohn G and Laszlo I (1999) Limits of optical range sensors and how exploit them. In: Asakura T (ed.) *Trends in Optics and Phototonics*. Ico IV, Springer Series in Optical Sciences, vol. 74, pp. 328–342. Berlin, Heidelberg, New York: Springer Verlag.

Häusler G and Herrmann J (1992) Physical limits of 3D-sensing. Proceedings of the SPIE Conf. 1822. In: Svetkov D (ed.) *Optics, Illumination, and Image Sensing for Machine Vision VII*, pp. 150–158. Boston, MA: SPIE.

Klein M and Furtak TE (1986) *Optics*. New York: John Wiley & Sons.

Klinger P, Spellenberg B, Herrmann JM and Häusler G (2001) *In Process 3D-Sensing for Laser Material Processing*. Proceedings of Third Int. Conf. on 3-D Digital Imaging and Modeling, Quebec City, Canada. Los Alamitos: IEEE Computer Society, 38–41.

Pedrotti F and Pedrotti L (1993) *Introduction to Optics*, 2nd edn. New Jersey: Prentice Hall.
Pérez J-Ph (1996) *Optik, Spektrum Akademischer*. Heidelberg: Verlag.
Sharma A, Kandpal HC and Ghatak AK (2004) *Optical Coherence*. See this Encyclopedia.
van de Gracht (2004) *Coherence and Imaging*. See this Encyclopedia.
Young M (1993) *Optics and Lasers, Including Fibers and Optical Wave Guides*, 4th revised edn. Berlin, Heidelberg: Springer Verlag.

COHERENT CONTROL

Contents

Theory
Experimental
Applications in Semiconductors

Theory

H Rabitz, Princeton University, Princeton, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the inception of quantum mechanics almost a century ago, a prime activity has been the observation of quantum phenomena in virtually all areas of chemistry and physics. However, the natural evolution of science leads to the desire to go beyond passive observation to active manipulation of quantum mechanical processes. Achieving control over quantum phenomena could be viewed as engineering at the atomic scale guided by the principles of quantum mechanics, for the alteration of system properties or dynamic behavior. From this perspective, the construction of quantum mechanically operating solid-state devices through selective material growth would fall into this category. The focus of this article is principally on the manipulation of quantum phenomena through tailored laser pulses. The suggestion of using coherent radiation for the active alteration of microworld processes may be traced to the early 1960s, almost immediately after the discovery of lasers. Since then, the subject has grown enormously to encompass the manipulation of (1) chemical reactions, (2) quantum electron transport in semiconductors, (3) excitons in solids, (4) quantum information systems, (5) atom lasers, and (6) high harmonic generation, amongst other topics. Perhaps the most significant use of these techniques may be their provision of refined tools to ultimately better understand the basic physical interactions operative at the atomic scale.

Regardless of the particular application of laser control over quantum phenomena, there is one basic operating principle involved: active manipulation of constructive and destructive quantum wave interferences. This process is depicted in Figure 1, showing the evolution of a quantum system from the initial state $|\psi_i\rangle$ to the desired final state $|\psi_f\rangle$ along three of possibly many interfering pathways. In general, there may be many possible final accessible states $|\psi_f\rangle$, and often the goal is to achieve a high amplitude in one of these states and low amplitude in all the others. The target state might actually be a superposition of states, and an analogous picture to

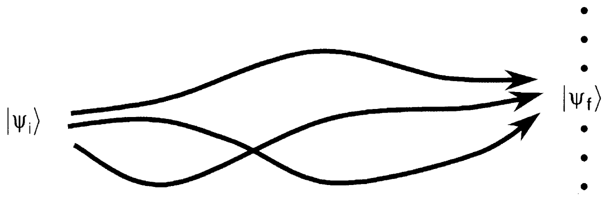

Figure 1 The evolution of a quantum system under laser control from the initial state $|\psi_i\rangle$ to the final state $|\psi_f\rangle$. The latter state is chosen to have desirable physical properties, and three of possibly many pathways between the states are depicted. Successful control of the process $|\psi_i\rangle \rightarrow |\psi_f\rangle$ by a tailored laser pulse generally requires creating constructive quantum wave interferences in the state $|\psi_f\rangle$ from many pathways, and destructive interferences in all other accessible final states $|\psi_{f'}\rangle \neq |\psi_f\rangle$.

that in Figure 1 would also apply to steering about the density matrix. The process depicted in Figure 1 may be thought of as a microscale analog of the classic double-slit experiment for light waves. As the goal is often high-finesse focusing into a particular target quantum state, success may call for the manipulation of many quantum pathways (i.e., the notion of a 'many-slit' experiment). Most applications are concerned with achieving a desirable outcome for the expectation value $\langle \psi_f | O | \psi_f \rangle$ associated with some observable Hermitian operator O.

The practical realization of the task above becomes a control problem when the system is expressed through its Hamiltonian $H = H_0 + V_c$, where H_0 is the free Hamiltonian describing the dynamics without explicit control; it is assumed that the free evolutionary dynamics under H_0 will not satisfactorily achieve the physical objective. Thus, a laboratory-accessible control term V_c is introduced in the Hamiltonian to achieve the desired manipulation. Even just considering radiative interactions, the form of V_c could be quite diverse depending on the nature of the system (e.g., nuclear spins, electrons, atoms, molecules, etc.) and the intensity of the radiation field. Many problems may be treated through an electric dipole interaction $V_c = -\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ where $\boldsymbol{\mu}$ is a system dipole moment and $\boldsymbol{\varepsilon}(t)$ is the laser electric field. Where appropriate, this article will consider the interaction in this form, but other suitable radiative control interactions may be equally well treated. Thus, the laser field $\boldsymbol{\varepsilon}(t)$ as a function of time (or frequency) is at our disposal for attempted manipulation of quantum systems.

Before considering any practical issues associated with the identification of shaped laser control fields, a fundamental question concerns whether it is, in principle, possible to steer about any particular quantum system from an arbitrary initial state $|\psi_i\rangle$ to an arbitrary final state $|\psi_f\rangle$. Questions of this sort are addressed by a controllability analysis of Schrödinger's equation

$$i\hbar \frac{\partial |\psi(t)\rangle}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_i\rangle \qquad [1]$$

Controllability concerns whether, in principle, some field $\boldsymbol{\varepsilon}(t)$ exists, such that the quantum system described by $H_0 = H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)$ permits arbitrary degrees of control. For finite-dimensional quantum systems (i.e., those described by evolution amongst a discrete set of quantum states), the formal tools for such a controllability analysis exist both for evaluating the controllability of the wave function, as well as the more general time evolution operator $U(t)$, which satisfies

$$i\hbar \frac{\partial U}{\partial t} = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]U, \quad U(0) = 1 \qquad [2]$$

Analyses of this type can be quite insightful, but they require detailed knowledge about H_0 and $\boldsymbol{\mu}$. Cases involving quantum information science applications are perhaps the most demanding with regard to achieving total control. Most other physical applications would likely accept much more modest levels of control and still be categorized as excellent achievements.

Theoretical tools and concepts have a number of roles in considering the control of quantum systems, including (1) an exploration of physical/chemical phenomena under active control, (2) the design of viable control fields, (3) the development of algorithms to actively guide laboratory control experiments towards achieving their dynamical objectives, and (4) the introduction of special algorithms to reveal the physical mechanisms operative in the control of quantum phenomena. Activities (1)–(3) have thus far been the primary focus of theoretical studies in this area, and it is anticipated that item (4) will grow in importance in the future.

A few comments on the history of laser control over quantum systems are relevant, as they speak to the special nature of the currently employed successful closed-loop quantum control experiments. Starting in the 1960s and spanning roughly 20 years, it was thought that the design of lasers to manipulate molecular motion could be achieved by the application of simple physical logic and intuition. In particular, the thinking at the time focused on using cw laser fields resonant with one or more local modes of the molecule, as state or energy localization was believed to be the key to successful control. Since quantum dynamics phenomena typically occur on ultrafast time-scales, possibly involving spatially dispersed wave packets, expecting to achieve high quality control with a light source operating at one or two resonant frequencies is generally wishful thinking. Quantum dynamics phenomena occur in a multifrequency domain, and controls with a few frequencies will not suffice. Over approximately the last decade, it became clear that successful control often calls for manipulating multiple interfering quantum pathways (cf., Figure 1). In turn, this recognition led to the need for broad-bandwidth laser sources (i.e., tailored laser pulses). Fortunately, the necessary laser pulse-shaping technologies have become available, and these sources continue to expand into new frequency ranges with enhanced bandwidth capabilities. Many chemical and physical applications of control over quantum phenomena involve performing a significant amount of work (i.e., quantum mechanical action), and sufficient laser field intensity is required. A commonly quoted adage is 'no field, no yield', and at the other extreme, there was much speculation that operating nonlinearly at high field intensities would lead to a loss of control because the quantum system would effectively act as an amplifier of even weak laser field noise. Fortunately, the latter outcome has not occurred, as is evident from a number of successful high field laser control experiments.

Quantum control may be expressed as an inverse problem with a prescribed chemical or physical objective, and the task being the discovery of a suitable control field $\boldsymbol{\varepsilon}(t)$ to meet the objective. Normally, Schrödinger's equation [1] is viewed as linear, and this perspective is valid if the Hamiltonian is known a priori, leaving the equation to be solved for the wave function. However, in the case of quantum control, neither the wave function nor the control field is known a priori, and the two enter bilinearly on the right-hand side of eqn [1]. Thus, quantum control is mathematically a nonlinear inverse problem. As such, one can anticipate possibly diverse behavior in the process of seeking successful controls, as well as in the ensuing quantum dynamical control behavior. The control must take into account the evolving system dynamics, thereby resulting in the control field being a function(al) of the current state of the system $\boldsymbol{\varepsilon}(\psi)$ in the case of quantum control design, or dependent on the system observations $\boldsymbol{\varepsilon}(\langle O\rangle)$ in the case of laboratory closed-loop field discovery efforts where $\langle O\rangle = \langle\psi|O|\psi\rangle$. Furthermore, the control field at the current time t will depend in some implicit fashion on the future state of the evolving system at the final target time T. It is evident that control field design and laboratory discovery present highly nontrivial tasks.

Although the emphasis in this review is on the theoretical aspects of quantum control theory, the subject exists for its laboratory realizations, and those realizations, in turn, intimately depend on the capabilities of theoretical and algorithmic techniques for their successful implementations. Accordingly, theoretical laser control field design techniques will be discussed, along with algorithmic aspects of current laboratory practice. This material respectively will focus on optimal control theory (OCT) and optimal control experiments (OCEs), as seeking optimality provides the best means to achieve any posed objective. Finally, the last part of the article will present some general conclusions on the state of the quantum control field.

Quantum Optimal Control Theory for Designing Laser Fields

When considering control in any domain of application, a reasonable approach is to computationally design the control for subsequent implementation in the laboratory. Quantum mechanical laser field design has taken on a number of realizations. At one limit is the application of simple intuition for design purposes, and in some special cases (e.g., the weak field perturbation theory regime), this approach may be applicable. Physical insights will always play a central role in laser field design, but to be especially useful, they need to be channeled into the proper mathematical framework. Many of the interesting applications operate in the strong-field nonperturbative regime. In this domain, serious questions arise regarding whether sufficient information is available about the Hamiltonian to execute reliable designs. Regardless of whether the Hamiltonian is known accurately or is an acknowledged model, the theoretical study of quantum control can provide physical insight into the phenomena involved, as well as possibly yielding trial laser pulses for further refinement in the laboratory.

Achieving control over quantum mechanical phenomena often involves a balance of competing dynamical processes. For example, in the case of aiming to create a particular excitation in a molecule or material, there will always be the concomitant need to minimize other unwanted excitations. Often, there are also limitations on the form, intensity, or other characteristics of the laser controls that must be adhered to. Another goal is for the control outcome to be as robust as possible to the laser field fluctuations and Hamiltonian uncertainties. Overriding all of these issues is merely the desire to achieve the best possible physical result for the posed quantum mechanical objective. In summary, all of these desired extrema conditions translate over to posing the design of control laser fields as an optimization problem. Thus, optimal control theory forms a basic foundation for quantum control field design.

Control field design starts by specification of the Hamiltonian components H_0 and $\boldsymbol{\mu}$ in eqn [1], with the goal of finding the best control electric field $\boldsymbol{\varepsilon}(t)$ to balance the competing objectives. Schrödinger's equation must be solved as part of this process, but this effort is not merely a forward propagation task, as the control field is not known a priori. As an optimal design is the goal, a cost functional $J = J(\text{objectives, penalties, } \boldsymbol{\varepsilon}(t))$ is prescribed, which contains the information on the physical objectives, competing penalties, and any costs or constraints associated with the structure of the laser field. The functional J could contain many terms if the competing physical goals are highly complex. As a specific simple illustration, consider the common goal of steering the system to achieve an expectation value $\langle\psi(T)|O|\psi(T)\rangle$ as close as possible to the target value O_{target} associated with the observable operator O at time T. An additional cost is imposed to minimize the laser fluence. These criteria may be embodied in a cost functional of the form

$$J = \left(\langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}}\right)^2 + \omega\int_0^T \varepsilon^2(t)\,\mathrm{d}t - 2\Im\int_0^T \mathrm{d}t\left\langle\lambda(t)|\mathrm{i}\hbar\frac{\partial}{\partial t} - H_0 + \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)|\psi(t)\right\rangle \quad [3]$$

Here, the parameter $\omega \geq 0$ is a weight to balance the significance of the fluence term relative to achieving the target expectation value, and $|\lambda(t)\rangle$ serves as a Lagrange multiplier, assuring that Schrödinger's equation is satisfied during the variational minimization of J with respect to the control field. Carrying out the latter minimization will lead to eqn [1], along with the additional relations

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\lambda(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(t)]|\lambda(t)\rangle, \quad |\lambda(T)\rangle = 2\sigma O|\lambda(T)\rangle \quad [4]$$

$$\boldsymbol{\varepsilon}(t) = \frac{1}{\omega}\Im\langle\lambda(t)|\boldsymbol{\mu}|\psi(t)\rangle \quad [5]$$

$$\sigma = \langle\psi(T)|O|\psi(T)\rangle - O_{\text{target}} \quad [6]$$

Equations [1] and [4]–[6] embody the OCT design equations that must be solved to yield the optimal field $\boldsymbol{\varepsilon}(t)$ given in eqn [5]. These design equations have an unusual mathematical structure within quantum mechanics. First, insertion of the field expression in eqn [5] into eqns [1] and [4] produces two cubically nonlinear coupled Schrödinger-like equations. These equations are in the same family as the standard nonlinear Schrödinger equation, and as such, one may expect that unusual dynamical behavior could arise. Although eqn [4] for $|\lambda(t)\rangle$ is identical in form to the Schrödinger equation [1], only $|\psi(t)\rangle$ is the true wavefunction, with $|\lambda(t)\rangle$ serving to guide the controlled quantum evolution towards the physical objective. Importantly, eqn [1] is an initial value problem, while eqn [4] is a final value problem. Thus, the two equations together form a two-point boundary value problem in time, which is an inherent feature of temporal engineering optimal control as well. The boundary value nature of these equations typically leads to the existence of multiple solutions, where σ in eqn [6] plays the role of a discrete eigenvalue, specifying the quality of the particular achieved control field design. The fact that there are generally multiple solutions to the laser design equations can be attractive from a physical perspective, as the designs may be sorted through to identify those being most attractive for laboratory implementation.

The overall mathematical structure of eqns [1] and [4]–[6] can be melded together into a single design equation

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\psi(t)\rangle = [H_0 - \boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}(\psi,\sigma,t)]|\psi(t)\rangle, \quad |\psi(0)\rangle = |\psi_{\mathrm{i}}\rangle \quad [7]$$

The structure of this equation embodies the comments in the introduction that control field design is inherently a nonlinear process. The main source of complexity arising in eqn [7] is through the control field $\boldsymbol{\varepsilon}(\psi,\sigma,t)$ depending not only on the wave function at the present time t, but also on the future value of the wavefunction at the target time T contained in σ. This structure again reflects the two-point boundary value nature of the control equations.

Given the typical multiplicity of solutions to eqns [1] and [4]–[6], or equivalently, eqn [7], it is attractive to consider approximations to these equations (except perhaps for the inviolate Schrödinger equation [1], and much work continues to be done along these lines. One case involves what is referred to as tracking control, whereby a path is specified for $\langle\psi(t)|O|\psi(t)\rangle$ evolving from $t = 0$ out to $t = T$. Tracking control eliminates σ to produce a field of the form $\boldsymbol{\varepsilon}(\psi,t)$, thereby permitting the design equations to be explicitly integrated as a forward-marching problem toward the target; the tracking equations are still nonlinear with respect to the evolving state. Many variations on these concepts can be envisioned, and other approximations may also be introduced to deal with special circumstances. One technique appropriate for at least few-level systems is stimulated Raman adiabatic passage (STIRAP), which seeks robust adiabatic passage from an initial state to a particular final state, typically with nearly total destructive interference occurring in the intermediate states. This design technique can have special attractive robustness characteristics with respect to field errors. STIRAP, tracking, and various perturbation theory-based control design techniques can be expressed as special cases of OCT, with suitable cost functionals and constraints. Further approximation methods will surely be developed, with the rigorous OCT concepts forming the general foundation for control field design.

As the OCT equations are inherently nonlinear, their solution typically requires numerical iteration, and a variety of procedures may be utilized for this purpose. Almost all of the methods employ local search techniques (e.g., gradient methods), and some also show monotonic convergence with respect to each new iteration step. Local techniques typically evolve to the nearest solution in the space of possible control fields. Such optimizations can be numerically efficient, although the quality of the attained result can depend on the initial trial for the control field and the details of the algorithm involved. In contrast, global search techniques, such as simulated annealing or genetic algorithms, can search more broadly for the best solution in the control field space. These more expansive searches are attractive, but at the price of typically requiring more intensive computations. Effort has also gone into seeking global or semiglobal input → output maps relating the electric field $\boldsymbol{\varepsilon}(t)$ structure to the observable $O[\boldsymbol{\varepsilon}(t)]$. Such maps may be learned, hopefully, from a modest number of computations with a selected set of fields, and then utilized as high-speed interpolators over the control field space to permit the efficient use of global search algorithms to attain the best control solution possible. At the foundation of all OCT design procedures is the need to solve the Schrödinger equation. Thus, computational technology to improve this basic task is of fundamental importance for designing laser fields.

Many computations have been carried out with OCT to attain control field designs for manipulating a host of phenomena, including rotational, vibrational, electronic, and reactive dynamics of molecules, as well as electron motion in semiconductors. Every system has its own rich details, which in turn, are compounded by the fact that multiple control designs will typically exist in many applications producing comparable physical outcomes. Collectively, these control design studies confirm the manipulation of constructive and destructive quantum wave interferences as the general mechanism for achieving successful control over quantum phenomena. This conclusion may be expressed as the following principle:

> **Control field–system cooperativity principle**: Successful quantum control requires that the field must have the proper structure to take full advantage of all of the dynamical opportunities offered by the system, to best satisfy the physical objective.

This simple statement of system–field cooperativity embodies the richness, as well as the complexity, of seeking control over quantum phenomena, and also speaks to why simple intuition alone has not proved to be a generally viable design technique. Quantum systems can often exhibit highly complex dynamical behavior, including broad dispersion of the wave packet over spatial domains or multitudes of quantum states. Handling such complexity can require fields with subtle structure to interact with the quantum system in a global fashion to manage all of the motions involved. Thus, we may expect that successful control pulses will often have broad bandwidth, including amplitude and phase modulation. Until relatively recently, laser sources with these characteristics were not available, but the technology is now in hand and rapidly evolving (see the discussion later).

The cooperativity principle above is of fundamental importance in the control of all quantum phenomena, and a simple illustration of this principle is shown in Figure 2 for the control of wave packet motion on an excited state of the NO molecule. The target for the control is a narrow wave packet located over the well of the excited B state. The optimal control field consists of two coordinated pulses, at early and late times, with both features having internal structure. The figure indicates that the ensuing excited state wave packet has two components, one of which actually passes through the target region during the initial evolution, only to return again and meet the second component at just the right place and time, to achieve the target objective as best as possible. Similar cooperativity interpretations can be found in virtually all implementations with OCT.

The main reason for performing quantum field design is to ultimately implement the designs in the laboratory. The design procedures must face laboratory realities, which include the fact that most Hamiltonians are not known to high accuracy (especially for polyatomic molecules and complex solid-state structures), and secondly, a variety of laboratory field imperfections may unwittingly be present. Notwithstanding these comments, OCT has been fundamental to the development of quantum control, including laying out the logic for how to perform the analogous optimal control experiments. Design implementations and further OCT development will continue to play a basic role in the quantum control field. At present, perhaps the most important contribution of OCT has been to (1) highlight the basic control cooperativity principle above, and (2) provide the basis for developing algorithms to successfully guide optimal control experiments, as discussed below.

Algorithms for Implementing Optimal Control Experiments

The ultimate purpose of considering control theory within quantum mechanics is to take the matter into

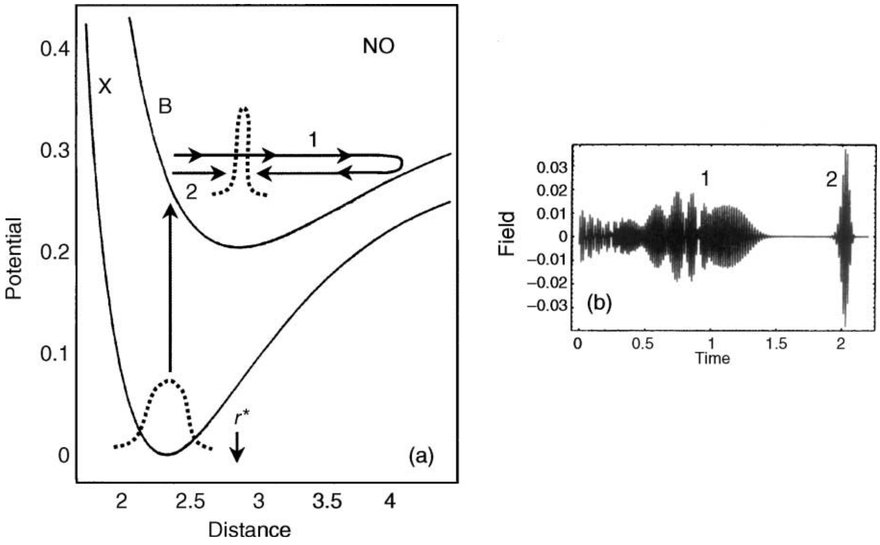

Figure 2 Control of wave packet evolution on the electronically excited B state of NO, with the goal of creating a narrow final packet over the excited state well. (a) indicates that the ground-state packet is brought up in two pieces using the coordinated dual pulse in (b). The piece of the packet from the first pulse actually passes through the target region, then bounces off the right side of the potential, to finally meet the second piece of the packet at just the right time over the target location r^* for successful control. This behavior is an illustration of the control field–system cooperativity principle stated in the text.

the laboratory for implementation and exploitation of its capabilities. Ideally, theoretical control field designs may be attained using the techniques discussed above, followed by the achievement of successful control upon execution of the designs in the laboratory. This appealing approach is burdened with three difficulties: (1) Hamiltonians are often imprecisely known, (2) accurately solving the design equations can be a significant task, and (3) realization of any given design will likely be imperfect due to laboratory noise or other unaccounted-for systematic errors. Perhaps the most serious of these difficulties is point (1), especially considering that the best quality control will be achieved by maximally drawing on subtle constructive and destructive quantum wave interference effects. Exploiting such subtleties will generally require high-quality control designs that, in turn, depend on having reliable Hamiltonians. Although various designs have been carried out seeking robustness with respect to Hamiltonian uncertainties, the issue in point (1) should remain of significance in the foreseeable future, especially for the most complex (and often, the most interesting!) chemical/physical applications. Mitigating this serious problem is the ability to create shaped laser pulse controls and apply them to a quantum system, followed by a probe of their effects at an unprecedented rate of thousands or more independent trials per minute. This unique capability led to the suggestion of partially, if not totally, sidestepping the design process by performing closed-loop experiments to let the quantum system teach the laser how to achieve its control in the laboratory. Figure 3 schematically shows this closed loop process, drawing on the following logic:

1. **The molecular view**. Although there may be theoretical uncertainty about the system Hamiltonian, the actual chemical/physical system under study 'knows' its own Hamiltonian precisely! This knowledge would also include any unusual exigencies, perhaps associated with structural or other defects in the particular sample. Furthermore, upon exposure to a control field, the system 'solves' its own Schrödinger equation impeccably accurately and as fast as possible in real time. Considering these points, the aim is to replace the offline arduous digital computer control field design effort with the actual quantum system under study acting as a precise analog computer, solving the true equations of motion.
2. **Control laser technology**. Pulse shaping under full computer control may be carried out using even hundreds of discrete elements in the frequency domain controlling the phase and amplitude structure of the pulse. This technology is readily available and expanding in terms of pulse center frequency flexibility and bandwidth capabilities.

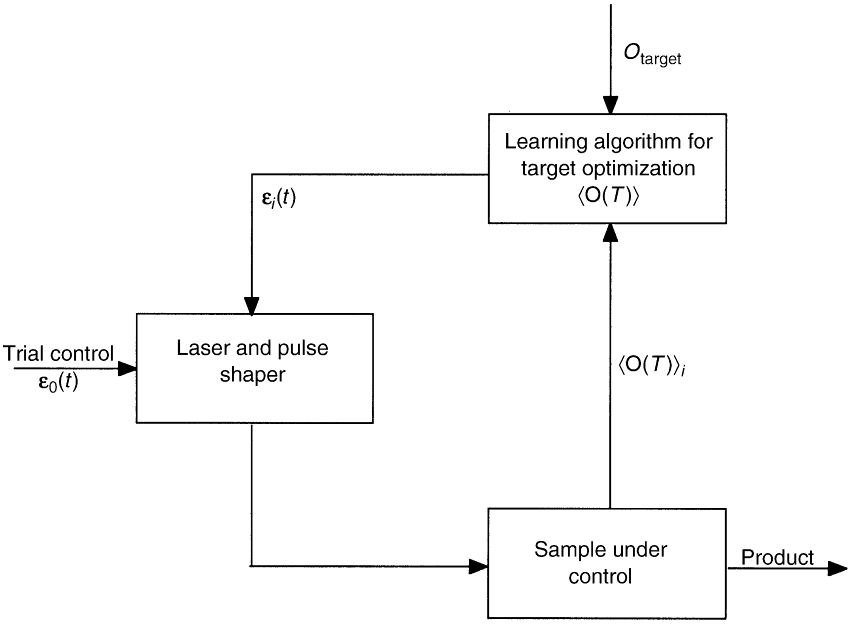

Figure 3 A schematic of the closed-loop concept for allowing the quantum system to teach the laser how to achieve its control. The actual quantum system under control is in a loop with a laser and pulse shaper all slaved together with a pattern recognition algorithm to guide the excursions around the loop. The success of this concept relies on the ability to perform very large numbers of control experiments in a short period of time. In principle, no knowledge of the system Hamiltonian is required to steer the system to the desired final objective, although a good trial design $\varepsilon_0(t)$ may accelerate the process. The ith cycle around the loop attempts to find a better control field $\varepsilon_i(t)$ such that the system response $\langle O(T)\rangle_i$ for the observable operator O comes closer to the desired value O_{target}.

3. **Quantum mechanical objectives.** Many chemical/physical objectives may be simply expressed as the desire to steer a quantum mechanical flux out one clearly defined channel versus another.
4. **Detection of the control action.** Detection of the control outcome can often be carried out by a second laser pulse or any other suitable high duty cycle detection means, such as laser-induced mass spectrometry. The typical circumstance in point (3) implies that little, if any, time-consuming offline data analysis will be necessary beyond that of simple signal averaging.
5. **Fast learning algorithms.** All of the four points above may be slaved together using pattern recognition learning algorithms to identify those control fields which are producing better results, and bias the next sequence of experiments in their favor for further improvement. Although many algorithms may be employed for this purpose, the global search capabilities of genetic algorithms are quite attractive, as they may take full advantage of the high throughput nature of the experiments.

In linking together all of these components into a closed-loop OCE, it is important that no single operational step significantly lags behind any other, for efficiency reasons. Fortunately, this economy is coincident with all the tasks and technologies involved. For achieving control alone, there is no requirement that the actual laser pulse structure be identified on each cycle of the loop. Rather, the laser control 'knob' settings are adequate information for the learning algorithm, as the only criterion is to suitably adjust the knobs to achieve an acceptable value for the chemical/physical objective. The learning algorithm in point (5) operates with a cost functional J, analogous to that used for computational design of fields, but now the cost functional can only depend on those quantities directly observable in the laboratory (i.e., minimally, the current achieved target expectation value). In cases where a physical understanding is sought about the control mechanism, the laser field structure must also be measured. This additional information is typically also not a burden to attain, as it often is only required for the single best control field at the end of the closed-loop learning experiments.

The OCE process in Figure 3 and steps (1)–(5) above constitute the laboratory process for achieving optimal quantum control, and exactly the same reasons put forth for OCT motivate the desire for attaining optimal performance: principally, seeking the best control result that can possibly be obtained. Seeking optimal performance also ensures some degree of inherent robustness to field noise, as fleeting observations would not likely survive the signal averaging carried out in the laboratory. Additionally, robustness as a specific criterion may also be included in the laboratory cost functional in item (5). When a detailed physical interpretation of the controlled dynamics is desired, it is essential to remove any extraneous control field features that have little impact on the system manipulations. Without the latter clean-up carried out during the experiments, the final field may be contaminated by structures that have little physical impact. Although quantum dynamics typically occurs on femtosecond or picosecond time-scales, the loop excursions in Figure 3 need not be carried out on the latter time-scales. Rather, the loop may be traversed as rapidly as convenient, consistent with the capabilities of the apparatus. A new system sample is introduced on each cycle of the loop. This process is referred learning control, to distinguish it from real-time feedback control.

The closed-loop quantum learning control algorithm can be self-starting, requiring no prior control field designs, under favorable circumstances. Virtually all of the current experiments were carried out in this fashion. It remains to be seen if this procedure of ‘going in blind’ will be generally applicable in highly complex situations where the initial state is far from the target state. Learning control can only proceed if at least a minimal signal is observed in the target state. Presently, this requirement has been met by drawing on the overwhelmingly large number of exploratory experiments that can be carried out, even in a brief few minutes, under full computer control. However, in some situations, the performance of intermediate observations along the way to the target goal may be necessary in order to at least partially guide the quantum dynamics towards the ultimate objective. Beneficial use could also be made from prior theoretical laser control designs $\boldsymbol{\varepsilon}_0(t)$ capable of at least yielding a minimal target signal.

The OCE closed-loop procedure was based on the growing number of successful theoretical OCT design calculations, even with all of their foibles, especially including less than perfect Hamiltonians. The OCT and OCE processes are analogous, with their primary distinction involving precisely what appears in their corresponding cost functionals. Theoretical guidance can aid in identifying the appropriate cost functionals and learning algorithms for OCE, especially for addressing the needs associated with attaining a physical understanding about the mechanisms governing controlled quantum dynamics phenomena.

A central issue is the rate that learning control occurs upon excursions around the loop in Figure 3. A number of factors control this rate of learning, including the nature of the physical system, the choice of objective, the presence of field uncertainties and measurement errors, the number of control variables, and the capabilities of the learning algorithm employed. Achieving control is a matter of discrimination, and some of these factors, whether inherent to the system or introduced by choice, may work against attaining good-quality discrimination. At this juncture, little is known quantitatively about the limitations associated with any of these factors.

The latter issues have not prevented the successful performance of a broad variety of quantum control experiments, and at present the ability to carry out massive numbers of closed-loop excursions has overcome any evident difficulties. The number of examples of successful closed-loop quantum control is rapidly growing, with illustrations involving the manipulation of laser dye fluorescence, chemical reactivity, high harmonic generation, semiconductor optical switching, fiber optic pulse transmission, and dynamical discrimination of similar chemical species, amongst others. Perhaps the most interesting cases are those carried out at high field intensities, where prior speculation suggested that such experiments would fail due to even modest laser field noise being amplified by the quantum dynamics. Fortunately, this outcome did not occur, and theoretical studies have indicated that surprising degrees of robustness to field noise may accompany the manipulation of quantum dynamics phenomena. Although the presence of field noise may diminish the beneficial influence of constructive and destructive interferences, evidence shows that field noise does not appear to kill the actual control process, at least in terms of obtaining reasonable values for the control objectives. One illustration of control in the high-field regime is shown in Figure 4, demonstrating the learning process for the dissociative rearrangement of acetophenone to form toluene. Interestingly, this case involves breaking two bonds, with the formation of a third, indicating that complex dynamics can be managed with suitably tailored laser pulses.

Operating in the strong-field regime, especially for chemical manipulations, also has the important benefit of producing a generic laser tool to control a broad variety of systems by overcoming the long-standing problem of having sufficient bandwidth. For example, the result in Figure 4 uses a Ti:sapphire laser in the near-infrared regime. Starting with a bandwidth-limited pulse of intensity near

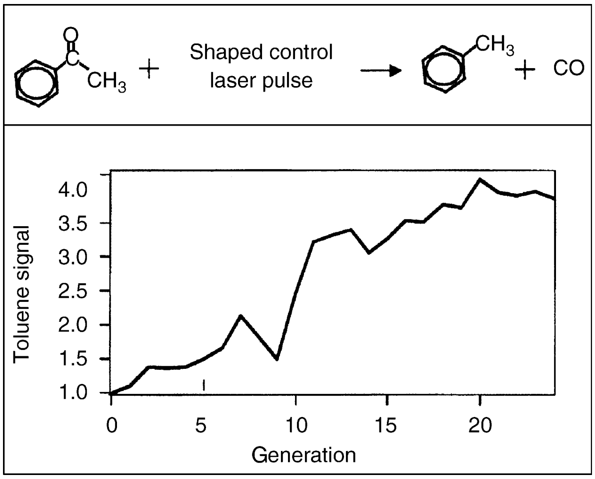

Figure 4 An illustration of dissociative rearrangement achieved by closed-loop learning control in Figure 3. The optimally deduced laser pulse broke two bonds in the parent acetophenone molecule and formed a new one to yield the toluene product. The laboratory learning curve is shown for the toluene product signal as a function of the generations in the genetic algorithm guiding the experiments. The fluctuations in the learning curve for optimizing the toluene yield corresponds to the algorithm searching for the optimal yield as the experiments proceed.

$\sim 10^{14}$ W cm^{-2}, in this regime, the dynamic power broadening can easily be on the order of ~ 1 eV or more, thereby effectively converting the otherwise discrete spectrum of the molecule into an effective continuum for ready multiphoton matching by the control laser pulse. Operating under these physical conditions is very attractive, as the apparatus is generic, permitting a single shaped-pulse laser source (e.g., a Ti:sapphire system with phase and amplitude modulation) to be utilized with virtually any chemical system where manipulations are desired. Although every desired physical/chemical goal may not be satisfactorily met (i.e., the goals must be physically attainable!), the means are now available to explore large numbers of systems. Operating under closed loop in the strong-field tailored-pulse regime eliminates the prior serious limitation of first finding a physical system to meet the laser capabilities; now, the structure of the laser pulse can be shaped to meet the molecule's characteristics. Strong field operations may be attractive for these reasons, but other broadband laser sources, possibly working at weaker field intensities, might be essential in some applications (e.g., controlling electron dynamics in semiconductors, where material damage is to be avoided). It is anticipated that additional broadband laser sources will become available for these purposes.

The coming years should be a period of rapid expansion, including a thorough exploration of closed-loop quantum control capabilities. As with the use of any laboratory tool, certain applications may be more amenable than others to attaining successful control. The particular physical/chemical questions need to be well posed, and controls need to have sufficient flexibility to meet the objectives. The experiments ahead should be able to reveal the degree to which drawing on optimality in OCE, combined with the performance of massive numbers of experiments can lead to broad-scale successful control of quantum phenomena. One issue of concern is the richness associated with the large numbers of phase and amplitude control knobs that may be adjusted in the laboratory. Some experiments have already operated with hundreds of knobs, while others have restricted their number in a variety of ways, to simplify the search process. Additional technological and algorithmic advances may be required to manage the high-dimensional control space searches. Fortunately, for typical applications, the search does not reduce to seeking a needle in a haystack, as generally there are multiple control solutions, possibly all of very good quality.

As a final comment on OCE, it is useful to appreciate the subtle distinction of the procedure from closed-loop feedback experiments. This distinction is illustrated in Figure 5, pointing to three types of closed-loop quantum control laboratory experiments. The OCE procedure in Figure 5a produces a generic laser tool capable of controlling a broad variety of systems with an emphasis in the figure placed on the point that each cycle around the loop starts with a new sample for control. The replacement of the sample on each cycle eliminates a number of difficulties, principally including concerns about sample damage, avoidance of operating the loop at the ultrafast speed of the quantum mechanical processes, and elimination of the effect that 'to observe is to disturb' in quantum mechanics. Thus, learning control provides a generic practical procedure, regardless of the nature of the quantum system. Figure 5b has the same structure as that of Figure 5a, except the loop is now closed around the same single quantum system, which is followed throughout its evolution. All of the issues mentioned above that are circumvented by operating through laboratory learning control in Figure 5a must now be directly faced in the setup of Figure 5b. The procedure in Figure 5b will likely only be applicable in special circumstances, at least for the reason that many quantum systems operate on time-scales far too fast for opto-electronics and computers to keep up with. However, there are certain quantum mechanical processes that are sufficiently slow to meet the criteria. Furthermore, a period of free evolutionary

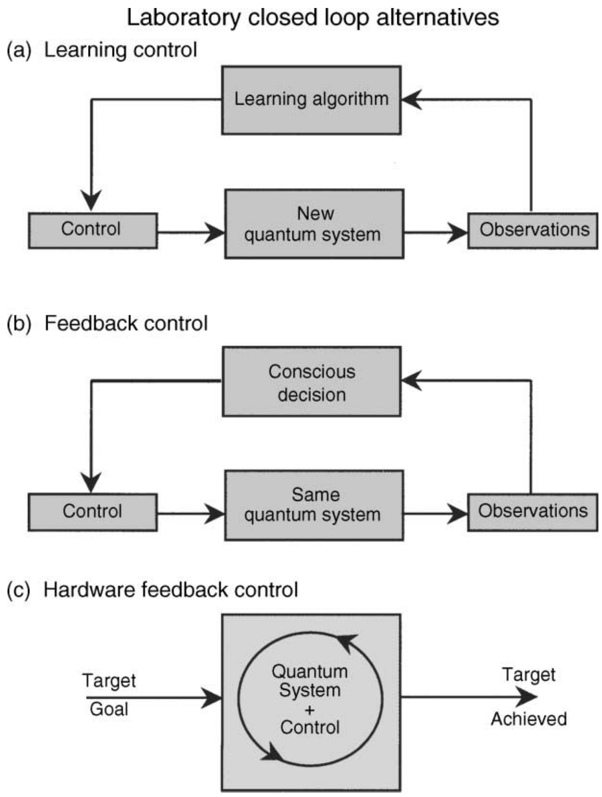

Figure 5 Three possible closed-loop formulations for attaining laser control over quantum phenomena. (a) is a learning control process where a new system is introduced on each excursion around the loop. (b) utilizes feedback control with the loop being traversed with the same sample, generally calling for an accurate model of the system to adjust the controls on each loop excursion. (c) is, at present, a dream machine where the laser and the sample under control are one unit operating without external algorithmic guidance. The closed-loop learning control procedure in (a) appears to form the most practical means to achieve laser control over quantum phenomena, especially for complex systems.

quantum dynamics may be permitted to occur between one control pulse and the next, to make the time issue manageable. While the learning control process in Figures 3 and 5a can be performed model-free, the feedback algorithm in Figure 5b generally must operate with a sufficiently reliable system Hamiltonian to carry out fast design corrections to the control field, based on the previous control outcome. These are very severe demands, but they may be met under certain circumstances. One reason for considering closed-loop feedback control in Figure 5b is to explore the basic physical issue of quantum mechanical limitations inherent in the statement 'to observe is to disturb'. It is suggestive that control over many types of quantum phenomena may fall into the semiclassical regime, lying somewhere between classical engineering behavior and the hard limitations of quantum mechanics. Experiments of the type in Figure 5b will be most interesting for exploring this matter. Finally, Figure 5c introduces a *gedanken* experiment in the sense that the closed-loop process is literally built into the hardware. That is, the laser control and quantum system act as a single functioning unit operating in a stable fashion, so as to automatically steer the system to the desired target. This process may involve engineering the quantum mechanical system prior to control in cases where that freedom exists, as well as engineering the laser components involved. The meaning of such a device in Figure 5c can be understood by considering an analogy with airplane flight, where the aircraft is constructed to have an inherent degree of aerodynamic stability and will essentially fly (glide) on its own accord when pushed forward. It is an open question whether closing the loop in the hardware can be attained for quantum control, and an exploration of this concept may require additional laser technologies.

Conclusions

This article presented an overview of theoretical concepts and algorithmic considerations associated with the control of quantum phenomena. Theory has played a central role in this area by revealing the fundamental principles underlying quantum control, as well as by providing algorithms for designing controls and guiding experiments to discover successful controls in the laboratory. The challenge of controlling quantum phenomena, in one sense, is an old subject, going back at least 40 years. However, roughly the first 30 of these years would best be described as a period of frustration, due to a lack of full understanding of the principles involved and the nature of the lasers needed to achieve success. Thus, from another perspective, the subject is quite young, with perhaps the most notable development being the introduction of closed-loop laboratory learning procedures. At the time of writing this article, these procedures are just beginning to be explored for their full capabilities. To appreciate the young nature of this subject, it is useful to note that the analogous engineering control disciplines presently occupy many thousands of engineers worldwide, both theoreticians and practitioners, and have done so for many years. Yet, engineering control is far from considered a mature subject. Armed now with the basic concepts and proper laboratory tools, one may anticipate a thorough exploration of control over quantum phenomena, including its many possible applications.

Acknowledgments

Support for this work has come from the National Science Foundation and the Department of Defense.

Further Reading

Brixner T, Damrauer NH and Gerber G (2001) Femtosecond quantum control. *Advances in Atomic, Molecular, and Optical Physics* 44: 1–54.

Rabitz H and Zhu W (2000) Optimal control of molecular motion: Design, implementation and inversion. *Accounts of Chemical Research* 33: 572–578.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824–828.

Rice SA and Zhao M (2000) Optical Control of Molecular Processes. NewYork: John Wiley.

Experimental

R J Levis, Temple University, Philadelphia, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The experimental control of quantum phenomena in atoms, molecules, and condensed phase materials has been a long-standing challenge in the physical sciences. For example, the idea that a laser beam could be used to selectively cleave a chemical bond has been pursued since the demonstration of the laser in the early 1960s. However, it was not until very recent times that one could realize selective bond cleavage. One key to the rapid acceleration of such experiments is the capability of manipulating the relative phases between two or more paths leading to the same final state. This is done in practice by controlling the phase of two (or two thousand!) laser frequencies coupling the initial state to some desired final state. The present interest in phase-related control stems in part from research focusing on quantum information sciences, controlling chemical reactivity, and developing new optical technologies, such as in biophotonics for imaging cellular materials. Ultimately, one would like to identify coherent control applications having impact similar to linear regime applications like compact disk reading (and writing), integrated circuit fabrication, and photodynamic therapy. Note that in each of these the phase of the electromagnetic field is not an important parameter and only the brightness of the laser at a particular frequency is used. While phase is of negligible importance in any one photon process, for excitations involving two or more photons, phase not only becomes a useful characteristic, it can modulate the yield of a desired process between zero and one hundred percent.

The use of optical phase to manipulate atomic and molecular systems rose to the forefront of laser science in the 1980s. Prior to this there was considerable effort directed toward using the intensity and high spectral resolution features of the laser for various applications. For example, the high spectral resolution character was employed in purifying nuclear isotopes. Highly intense beams were also used in the unsuccessful attempt to pump enough energy into a single vibrational resonance to selectively dissociate a molecule. The current revolution in coherent control has been driven in part by the realization that controlling energy deposition into a single degree of freedom (such as selectively breaking a bond in a polyatomic molecule) can not be accomplished by simply driving a resonance with an ever more intense laser source. In the case of laser-induced photo dissociation, such excitation ultimately results in statistical dissociation throughout the molecule due to vibrational mode coupling. In complex systems, controlling energy deposition into a desired quantum state requires a more sophisticated approach and one scheme involves coherent control.

The two distinct regimes for excitation in coherent control, the time and frequency domain experiments, are formally equivalent. While the experimental approach and level of success for each domain is vastly different, both rely on precisely specifying the phase relations for a series of excitation frequencies. The basic idea behind phase control is simple. One emulates the interference effects that can be easily visualized in a water tank experiment with a wave propagating through two slits. In the coherent control experiment, two or more indistinguishable pathways must be excited that connect one quantum state to another. In either water waves, or coherent control, the probability for observing constructive or destructive interference at a target state depends on the phase difference between the paths connecting the two locations in the water tank, or the initial and final states in the quantum system. In the case of coherent control, the final state may have some technological value, for instance, a desirable chemical reaction product or an intermediate in the implementation of quantum computation.

Frequency domain methods concentrate on interfering two excitation routes in a system by controlling the phase difference between two distinct frequencies coupling an initial state to a final state. Perhaps the first mention of the concept of phase control in the optical regime can be traced to investigations in Russia in the early 1960s, where the interference between a one- and three-photon path was theoretically proposed as a means to modulate a two-photon absorption. This two-path interference idea lay dormant until the reality of controlling the relative phase of two distinct frequencies was achieved in 1990 using gas phase pressure modulation for frequency domain experiments. The idea was theoretically extended to the control of nuclear motion in 1988. The first experimental example of two-color phase control over nuclear motion was demonstrated in 1991. It is interesting that the first demonstrations of coherent control did not, however, originate in the frequency domain approach, but came from time domain investigations in the mid-1980s. This was work that grew directly out of the chemical dynamics community and will be described below.

While frequency domain methods were developed exclusively for the observation of quantum interference, the experimental implementation of time-domain methods was first developed for observing nuclear motion in real time. In the time domain, a superposition of multiple states is prepared in a system using a short duration (~50 fs) laser pulse. The superposition state will subsequently evolve according to the phase relationships of the prepared states and the Hamiltonian of the system. At some later time the evolution must be probed using a second, short-time-duration laser pulse. The time-domain methods for coherent control were first proposed in 1985 and were based on wavepacket propagation methods developed in the 1970s. The first experiments were reported in the mid-1980s.

To determine whether the time or frequency domain implementation is more applicable for a particular control context, one should compare the relevant energy level spacing of the quantum system to the bandwidth of the excitation laser. For instance, control of atomic systems is more suited to frequency domain methods because the characteristic energy level spacing in an atom (several electron volts) is large compared to the bandwidth of femtosecond laser pulses (millielectron volts). In this case, preparation of a superposition of two or more states is impractical with a single laser pulse at the present time, dictating that the control scheme must employ the initial state and a single excited eigenstate. In practice, quasi-CW nanosecond duration laser pulses are employed for such experiments because current ultrafast (fs duration) laser sources cannot typically prepare a superposition of electronic states. One instructive exception involves the excitation of Rydberg states in atoms. Here the electronic energy level spacing can be small enough that the bandwidth of a femtosecond excitation laser may span several electronic states and thus can create a superposition state. One should also note that the next generation of laser sources in the attosecond regime may permit the preparation of a wavepacket from low-lying electronic states in an atom. In the case of molecules, avoiding a superposition state is nearly impossible. Such experiments employ pico- to femtosecond duration laser pulses, having a bandwidth sufficient to excite many vibrational levels, either in the ground or excited electronic state manifold of a molecule. To complete the time-dependent measurement, the propagation of the wavepacket to the final state of interest must be observed (probed) using a second ultrafast laser pulse that can produce fluorescence, ionization, or stimulated emission in a time-resolved manner. The multiple paths that link the initial and final state via the pump and the probe pulses are the key to the equivalence of the time and frequency domain methods.

There is another useful way to classify experiments that have been performed recently in coherent control, that is into either open-loop or closed-loop techniques. Open-loop signifies that a calculation may be performed to specify the required pulse shape for control before the experiment is attempted. All of the frequency-based, and most of the time-based experiments fall into this category. Closed-loop, on the other hand, signifies that the results from experimental measurements must be used to assist in determining the correct parameters for the next experiment, this pioneering approach was suggested in 1992. It is interesting that in the context of control experiments performed to date, closed-loop experiments represent the only route to determining the optimal laser pulse shape for controlling even moderately complex systems. In this approach, no input from theoretical models is required for coherent control of complex systems. This is unlike typical control engineering environments, where closed-loop signifies the use of experiments to refine parameters used in a theoretical model.

There is one additional distinction that is of value for understanding the state of coherent control at the present time. Experiments may be classified into systems having Hamiltonians that allow calculation of frequencies required for the desired interference patterns (thus enabling open-loop experiments) and those having ill-defined Hamiltonians (requiring closed-loop experiments). The open-loop experiments have demonstrated the utility of various control schemes, but are not amenable to systems having even moderate complexity. The closed-loop experiments are capable of controlling systems of moderate complexity but are not amenable to precalculation of the required control fields. Systems requiring closed-loop methods include propagation of short pulses in an optical fiber, control of high harmonic generation for soft X-ray generation, control of chemical reactions and control of biological systems. In the experimental regime, the value of learning control for dealing with complex systems has been well documented.

The remainder of this article represents an overview of the experimental implementation of coherent control. The earliest phase control experiments involved time-dependent probing of molecular wavepackets, and were followed by two-path interference in an atomic system. The most recent, and general, implementation of control involves tailoring a time-dependent electromagnetic field for a desired objective using closed-loop techniques. Since the use of tailored laser pulses appears to be rather general, a more complete description of the experiment is provided.

Time-Dependent Methods

The first experiments demonstrating the possibility of coherent control were performed to detect nuclear motion in molecules in real time. These investigations were the experimental realization of the pump-dump or pulse-timing control methods as originally described in 1985. The experimental applications have expanded to numerous systems including diatomic and triatomic molecules, small clusters, and even biological molecules. For such experiments, a vibrational coherence is prepared by exciting a superposition of states. These measurements implicitly used phase control both in the generation of the ultrashort pulses and in the coherent probing of the superposition state by varying the time delay between the pump and probe pulses, in effect modulating the outcome of a quantum transition. Since the earliest optical experiment in the mid-1980s, there have been many hundreds of such experiments reported in the literature. The motivation for the very first experiments was not explicitly phase control, but rather the observation of nuclear motion of molecules in real time. In such later measurements, oscillations in the superposition states were observable for many cycles, up to many tens of picoseconds in favorable cases, i.e., diatomic molecules in the gas phase. The loss of signal in the coherent oscillations is ultimately due to passage of the superposition state in to surrounding bath states. This may be due to collision with another molecule, dissociation of the system or redistribution of the excitation energy into other modes of the system not originally excited in the superposition. It is notable that coherence can be maintained in solution phase systems for tens of picoseconds and for the case of resonances having weak coupling to other modes, coherence can even be modulated in large biological systems on the picosecond time scale.

Decoherence represents loss of phase information about a system. In this context, phase information represents our detailed knowledge of the electronic and nuclear coordinates of the system. In the case of vibrational coherence, decoherence may be represented by intramolecular vibrational energy transfer to other modes of the molecule. This can be thought of as the propensity of vibrational modes of a molecule to couple together in a manner that randomizes deposited energy throughout the molecule. Decoherence in a condensed phase system includes transfer of energy to the solvent modes surrounding the system of interest. For an electronically excited molecule or atom, decoherence can involve spontaneous emission of radiation (fluorescence), or dissociation in the case of molecules. Certain excited states may have high probability for fluorescence and would represent a significant decoherence pathway and an obstacle for coherent control. These states may be called ‘lossy’ and are often used in optical pumping schemes as well as for stimulated emission pumping. An interesting question arises as to whether one can employ such states in a pathway leading to a desired final state, or must such lossy states be avoided at all costs in coherent control. This question has been answered to some degree by the method of rapid adiabatic passage. In this experiment, an initial state is coupled to a final state by way of a lossy state. The coupling is a coherent process and the preparation involves dressed states that are eigenstates of the system. In this coherent superposition the lossy state is employed, but no population is allowed to build up and thus decoherence is circumvented.

Two-Path Interference Methods

Interference methods form perhaps the most intuitive example of coherent control. The first experimental demonstration was reported in 1990 and involved the modulation of ionization probability in Hg by interfering a one-photon and three-photon excitation pathway to an excited electronic state (followed by two-photon ionization). The long delay between the introduction of the two-path interference concept in 1960 and the demonstration in 1990 was the difficulty in controlling the relative phase of the one- and three-photon paths. The solution involved generating the third harmonic of the fundamental in a nonlinear crystal, and copropagating the two beams through a low-density gas. Changing the pressure of the gas changes the optical phase retardance at different rates for different frequencies, thus allowing relative phase control between two colors. The ionization yield of Hg was clearly modulated as a function of pressure. Such schemes have been extended to diatomic molecular systems and the concept of determining molecular phase has been investigated. While changing the phase of more than two frequencies has not been demonstrated using pressure tuning (because of experimental difficulties), a more flexible method has been developed to alter relative phase of up to 1000 frequency bands, as will be described next.

Closed-Loop Control Methods

Perhaps the most exciting new aspect of control in recent years concerns the prospect for performing closed-loop experiments. In this approach a more complex level of control in matter is achieved in comparison to the two-path interference and pump-probe control methods. In the closed-loop method, an arbitrarily complex time-dependent electromagnetic field is prepared that may have multiple subpulses and up to thousands of interfering frequency bands. The realization of closed-loop control over complex processes relies on the confluence of three technologies: (i) production of large bandwidth ultrafast laser pulses; (ii) spatial modulation methods for manipulating the relative phases and amplitudes of component frequencies of the ultrafast laser pulse; and (iii) closed-loop learning control methods for sorting through the vast number of pulse shapes available as potential control fields. In the closed-loop method, the result of a laser molecule interaction is used to tailor an optimized pulse.

In almost all of the experimental systems used for closed-loop control today, Kerr lens mode-locking in the Ti:sapphire crystal is used to phase lock the wide bandwidth of available frequencies (750–950 nm) to create the ultrafast pulse. In this phenomenon, the intensity variation in the short laser pulse creates a transient lens in the Ti:sapphire lasing medium that discriminates between free lasing and mode-locked pulses. As a result, all of the population inversion available in the lasing medium may be coerced into enhancing the ultrashort traveling wave in the cavity in this way. The wide bandwidth available may then be amplified and tailored into a shaped electromagnetic field. Spatial light modulation is one such method for modulating the relative phases and amplitudes of the component frequencies in the ultrafast laser pulse to tailor the shaped laser pulse. Pulse shaping first involves dispersing the phase-locked frequencies on an optical grating, the dispersed radiation is then collimated using a cylindrical lens and the individual frequencies are focused to a line forming the Fourier plane. At the Fourier plane, the time-dependent laser pulse is transformed into a series of phase-locked continuous wave (CW) laser frequencies, and thus Fourier transformed from time to frequency space. The relative phases and amplitudes may be modulated in the Fourier plane using an array of liquid crystal pixels. After altering the spectral phase and amplitude profile, the frequencies are recombined on a second grating to form the shaped laser pulse. With one degree of phase control and 10 pixels there are an astronomic number (360^{10}) of pulse shapes that may be generated in this manner. (At the present time, pulse shapers have up to 2000 independent elements.) Evolutionary algorithms are employed to manage the available phase space. Realizing that the pulse shape is nothing more than the summation of a series of sine waves, each having an associated phase and amplitude, we find that a certain pulse shape can be represented by a genome consisting of an array of these frequency-dependent phases and amplitudes. This immediately suggests that the methods of evolutionary search strategies may be useful for determining the optimal pulse shape for a desired photoexcitation process.

The method of closed-loop optimal control for experiments was first proposed in 1992 and the first experiments were reported in 1997 regarding optimization of the efficiency of a laser dye molecule in solution. Since that time a number of experiments have been performed in both the weak and strong field regime. In the weak field regime, the intensity of the laser does not alter the field free structure of the excited states of the system under investigation. In the strong field, the field free states of the system are altered by the electric field of the laser. Whether one is in the weak or strong field regime depends on parameters including the characteristic level spacing of the states of the system to be controlled and the intensity of the laser coupling into those states. Examples of weak field closed-loop control include optimization of laser dye efficiency, compression of an ultrashort laser pulse, dissociation of alkali clusters, optimization of coherent anti-Stokes

Raman excitation, and optical pulse propagation in fibers. In the strong field, the laser field is used to manipulate the excited states of the molecule as well as induce population transfer among these states. In this sense, the laser is both creating and exciting resonances, in effect allowing universal excitation with a limited frequency bandwidth of 750–850 nm. For chemical applications, it is worth noting that most molecules do not absorb in this wavelength region in the weak field. Strong field control has been used for controlling bond dissociation, chemical rearrangement, photonic reagents, X-ray generation, molecular centrifuges and the manipulation of mass spectral fragmentation intensities. In the latter case, a new sensing technology has emerged wherein each individual pulse shape represents an independent sensor for a molecule. Given the fact that thousands of pulse shapes can be tested per minute, this represents a new paradigm for molecular analysis.

At the time of this writing, the closed-loop method for adaptively tailoring control fields, has far outstripped our theoretical understanding of the process (particularly in the strong field regime). Adaptive control is presently an active field of investigation, encompassing the fields of physics, engineering, optics, and chemistry. In the coming years, there will be as much work done on the mechanism of control as in the application of the methods. Due to small energy level spacings and complexity of the systems, we anticipate the majority of applications to be in the chemical sciences.

See also

Coherence: Coherence and Imaging; Overview. **Interferometry:** Overview. **Microscopy:** Interference Microscopy.

Further Reading

Brumer PW and Shapiro M (2003) *Principles of the Quantum Control of Molecular Processes*. New York: Wiley Interscience.

Chen C and Elliott DS (1990) Measurement of optical phase variations using interfering multiphoton ionization processes. *Physics Review Letters* 65: 1737.

Levis RJ and Rabitz H (2002) Closing the loop on bond selective chemistry using tailored strong field laser pulses. *Journal of Physical Chemistry A* 106: 6427–6444.

Levis RJ, Menkir GM and Rabitz H (2001) Selective bond dissociation and rearrangement with optimally tailored, strong-field laser pulses. *Science* 292(5517): 709–713.

Rice SA and Zhao M (2000) *Optical Control of Molecular Dynamics*. New York: Wiley Interscience.

Stolow A (2003) Femtosecond time-resolved photoelectron spectroscopy of polyatomic molecules. *Annual Review of Physical Chemistry* 54: 89–119.

Zewail AH, Casati G, Rice SA, *et al.* (1997) Femtochemistry: Chemical reaction dynamics and their control. *Chemical Reactions and Their Control on the Femtosecond Time Scale, Solvay Conference on Chemistry* 101: 3–45.

Applications in Semiconductors

H M van Driel and J E Sipe, University of Toronto, Toronto, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference phenomena are well-known in classical optics. In the early 19th century, Thomas Young showed conclusively that light has wave properties, with the first interference experiment ever performed. He passed quasi-monochromatic light from a single source through a pair of double slits using the configuration of Figure 1a and observed an interference pattern consisting of a series of bright and dark fringes on a distant screen. The intensity distribution can be explained only if it is assumed that light has wave or phase properties. In modern terminology, if E_1 and E_2 are the complex electric fields of the two light beams arriving at the screen, by the superposition principle of field addition the intensity at a particular point on the screen can be written as

$$\begin{aligned} I &\propto |E_1 + E_2|^2 \\ &= |E_1|^2 + |E_2|^2 + |E_1||E_2|\cos(\phi_1 - \phi_2) \end{aligned} \qquad [1]$$

where $\phi_1 - \phi_2$ is the phase difference of the beams arriving at the screen. While this simple experiment was used originally to demonstrate the wave properties of light, it has subsequently been used for many other purposes, such as measuring the wavelength of light. However, for our purposes, here the Young's double slit apparatus can also be viewed as a device that redistributes light, or

(a) Optical interferometer

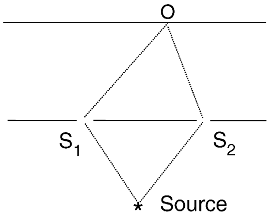

(b) Matter interferometer

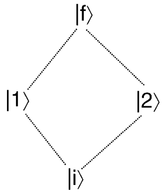

(c) Conduction-valence band transitions

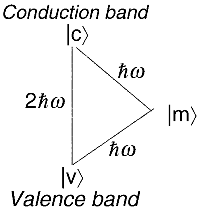

Figure 1 (a) Interference effects in the Young's double slit experiment; (b) illustration of the general concept of coherence control via multiple quantum mechanical pathways; (c) interference of single- and two-photon transitions connecting the same valence and conduction band states in a semiconductor.

controls its intensity at a particular location. For example, let's consider one of the slits to be a source, with the other slit taken to be a gate, with both capable of letting through the same amount of light. If the gate is closed, the distant screen is nearly uniformly illuminated. But if the gate is open, at a particular point on the distant screen there might be zero intensity or as much as four times the intensity emerging from one slit because of interference effects, with all the light merely being spatially redistributed. In this sense the double slit system can be viewed as a device to redistribute the incident light intensity. The key to all this is the superposition principle and the properties of optical phase.

The superposition principle and interference effects also lie at the heart of quantum mechanics. For example, let's now consider a system under the influence of a perturbation with an associated Hamiltonian that has phase properties. In general the system can evolve from one quantum state $|i\rangle$ to another $|f\rangle$ via multiple pathways involving intermediate states $|m\rangle$. Because of the phased perturbation, interference between those pathways can influence the system's final state. If a_m is the (complex) amplitude associated with a transition from the initial to the final state, via intermediate virtual state $|m\rangle$, then for all possible intermediate states the overall transition probability, W, can be written as

$$W = \left| \sum_m a_m \right|^2 \qquad [2]$$

In the case of only two pathways, as illustrated in Figure 1b, W becomes

$$\begin{aligned} W &= |a_1 + a_2|^2 \\ &= |a_1|^2 + |a_2|^2 + |a_1||a_2|\cos(\phi_1 - \phi_2) \end{aligned} \qquad [3]$$

where ϕ_1 and ϕ_2 are now the phases of the two transition amplitudes. While this expression strongly resembles that of eqn [1], here ϕ_1 and ϕ_2 are influenced by the phase properties of the perturbation Hamiltonian that governs the transition process; these phase factors arise, for example, if light fields constitute the perturbation. But overall it is clear that the change in the state of the system is affected by both the amplitude and phase properties of the perturbation. Analogous with the classical interferometer, equality of the transition amplitudes leads to maximum contrast in the transition rate. Although it has been known since the early days of quantum mechanics that the phase of a perturbation can influence the evolution of a system, generally phase has only been discussed in a passive role. Only in recent years has phase been used as a control parameter for a system, on the same level as the strength of the perturbation itself.

Coherence Control

Coherence control, or quantum control as it is sometimes called, refers to the active process through which one can use phase-dependent perturbations originating with, for example, coherent light waves, to control one or more properties of a quantum system, such as state population, momentum, or spin. This more general perspective of interference leads to a picture of interference of matter waves, rather than simply light waves, and one can now speak of an effective 'matter interferometer'. Laser beams, in particular, are in a unique position to play a role in such processes, since they offer a macroscopic 'phase handle' with which to create such effects. This was recognized by the community of atomic and molecular scientists who emphasized the active manifestations of quantum interference effects, and proposed that branching ratios in photochemical reactions might be controlled through laser-induced interference processes. The use of phase as a control parameter also represents a novel horizon of applications for the laser since most previous applications had involved only amplitude (intensity).

Of course, just as the ‘visibility’ of the screen pattern for a classical Young’s double slit interferometer is governed by the coherence properties of the two slit source, the evolution of a quantum system reflects the coherence properties of the perturbations, and the degree to which one can exercise phase control via external perturbations is also influenced by the interaction of the system of interest with a reservoir – essentially any other degrees of freedom in the system – which can cause decoherence and reduce the effectiveness of the ‘matter interferometer’. Since the influence of a reservoir will generally increase with the complexity of a system, one might think that coherence control can only be effectively achieved in simple atomic and molecular systems. Indeed, some of the earliest suggestions for coherence control of a system involved using interference between single and three photon absorption processes, connecting the same initial and final states with photons of frequency 3ω and ω, respectively, and controlling the population of excited states in atomic or diatomic systems. However, it has been difficult to extend this ‘two color’ paradigm to more complex molecular systems. Since the reservoir leads to decoherence of the system overall, shorter and shorter pulses must be used to overcome decoherence effects. However, short pulses possess a large bandwidth with the result that, for complex systems, selectivity of the final state can be lost. One must therefore consider using interference between multiple pathways in order to control the system, as dictated by the details of the unperturbed Hamiltonian of the system.

For a polyatomic molecule or a solid, the complete Hamiltonian is virtually impossible to determine exactly and it is therefore equally difficult to prescribe the optimal spectral (amplitude and phase) content of the pulses that should be used for control purposes. An alternative approach has therefore emerged, in which one foregoes knowledge of the eigenstates of the Hamiltonian and details of the different possible interfering transition paths in order to achieve control of the end state of a system. This branch of coherence control has come to be known as optimal control. In optimal control one employs an optical source for which one can (ideally) have complete control over the spectral and phase properties. One then uses a feedback process in which experiments are carried out, the effectiveness of achieving a certain result is determined, and the pulse characteristics are then altered to obtain a new result. A key component is the use of an algorithm to select the new pulse properties as part of the feedback system. In this approach the molecule teaches the external control system what it ‘requires’ for a certain result to be optimally achieved. While the ‘best’ pulse properties may not directly reveal details of the multiple interference process required to achieve the optimal result, this technique can nonetheless be used to gain some insight into the properties of the unperturbed Hamiltonian and, regardless of such understanding, achieve a desirable result. It has been used to control chemical reaction rates involving several polyatomic molecules, with considerable enhancement in achieving a certain product relative to what can be done using simple thermodynamics.

Coherence Control in Semiconductors

Since coherence control of chemical reactions involving large molecules has represented a significant challenge, it was generally felt that ultrafast decoherence processes would also make control in solids, in general, and semiconductors, in particular, difficult. Early efforts therefore focused on atomic-like situations in which the electrons are bound and associated with discrete states and long coherence times. In semiconductors, the obvious choice is the excitonic system, defect states, or discrete states offered by quantum wells. Population control of excitons and directional ionization of electrons from quantum wells has been clearly demonstrated, similar to related population control of and directional ionization from atoms. Other manifestations of coherence control of semiconductors include control of electron–phonon interactions and intersub-band transitions in quantum wells.

Among the different types of coherence control in semiconductors is the remarkable result that it is possible to coherently control the properties of free electrons associated with continuum states. Although optimal control might be used for some of these processes, we have achieved clear illustrations of control phenomena based on the use of harmonically related beams and interference of single- and two-photon absorption processes connecting the continuum valence and conduction band states as generically illustrated in Figure 1c. Details of the various processes observed can be found elsewhere. Of course, the momentum relaxation time of electrons or holes in continuum states is typically of the order of 100 fs at room temperature, but this time is sufficiently long that it can permit phase-controlled processes. Indeed, with respect to conventional carrier transport, this ‘long time’ lapse is responsible for the typically high electrical mobilities in crystalline semiconductors such as Si and GaAs. In essence, a crystalline semiconductor with translational symmetry has crystal momentum as a good quantum number, and selection rules for scattering prevent the momentum relaxation time from being prohibitively small. Since electrons or holes in any continuum state can participate in such control processes, one need not be concerned about pulsewidth or bandwidth, unless the pulses were to be so short that carriers of both positive and negative effective mass were generated within one band.

Coherent Control of Electrical Current Using Two Color Beams

We now illustrate the basic principles that describe how the interference process involving valence and conduction band states can be used to control properties of a bulk semiconductor. We do so in the case of using one or two beams to generate and control carrier population, electrical current, and spin current. In pointing out the underlying ideas behind coherence control of semiconductors, we will skip or suppress many of the mathematical details which are required for a full understanding but which might obscure the essential physics. In particular we consider how phase-related optical beams with frequencies ω and 2ω interact with a direct gap semiconductor such that $\hbar\omega < E_g < 2\hbar\omega$ where E_g is the electronic bandgap. For simplicity we consider exciting electrons from a single valence band via single-photon absorption at 2ω and two-photon absorption at ω. As is well known, within an independent particle approximation, the states of electrons and holes in semiconductors can be labeled by their vector crystal momentum **k**; the energy of states near the conduction or valence bandedges varies quadratically with $|\mathbf{k}|$. For one-photon absorption the transition amplitude can be derived using a perturbation Hamiltonian of the form $H = e/mc\, \mathbf{A}^{2\omega} \cdot \mathbf{p}$ where $\mathbf{A}^{2\omega}$ is the vector potential associated with the light field and **p** is the momentum operator. The transition amplitude is therefore of the form

$$\mathrm{a}^{2\omega} \propto E^{2\omega} e^{i\phi_{2\omega}} p_{cv} \qquad [4]$$

where p_{cv} is the interband matrix element of **p** along the field ($E^{2\omega}$) direction; for illustration purposes the field is taken to be linearly polarized. The overall transition rate between two particular states of the same **k** can be expressed as $W_1 \propto \mathrm{a}^{2\omega}(\mathrm{a}^{2\omega})^* \propto I^{2\omega}|p_{cv}|^2$ where $I^{2\omega}$ is the intensity of the beam. This rate is independent of the phase of the light beam as well as the sign of **k**. Hence the absorption of light via single-photon transitions generally populates states of equal and opposite momentum with equal probability or, equivalently, establishes a standing electron wave with zero crystal momentum. This is not surprising since photons possess very little momentum and, in the approximation of a uniform electric field, do not give any momentum to an excited electron. The particular states that are excited depends on the light polarization and crystal orientation. However, the main point is that while single-photon absorption can lead to anisotropic filling of electron states, the distribution in momentum space is not polar (dependent on sign of **k**). Similar considerations apply to the excited holes, but to avoid repetition we will focus on the electrons only.

For two-photon absorption involving the ω photons and connecting states similar to those connected with single-photon absorption, one must employ the 2nd order perturbation theory using the Hamiltonian $H = e/mc\mathbf{A}^{\omega} \cdot \mathbf{p}$. For two-photon absorption there is a transition from the valence band to an (energy nonallowed) intermediate state followed by a transition from the intermediate state to the final state. To determine the total transition rate one must sum over all possible intermediate states. For semiconductors, by far the dominant intermediate state is the final state itself, so that the associated transition amplitude has the form

$$\mathrm{a}^{\omega} \propto (E^{\omega} e^{i\phi_{\omega}} p_{cv})(E^{2\omega} e^{i\phi_{\omega}} p_{cc}) \propto (E^{\omega})^2 e^{2i\phi_{\omega}} p_{cv} \hbar k \qquad [5]$$

where the matrix element p_{cc} is simply the momentum of the conduction band state ($\hbar k$) along the direction of the field. Note that unlike an atomic system, p_{cc} is nonzero, since Bloch states in a crystalline solid do not possess inversion symmetry. If two-photon absorption acts alone, the overall transition rate between two particular **k** states would be $W_2 \propto (I^{\omega})^2 |p_{cv}|^2 k^2$. As with single-photon absorption, because this transition rate is independent of the sign of k, two-photon absorption leads to production of electrons with no net momentum.

When both single- and two-photon transitions are present simultaneously, the transition amplitude is the sum of the transition amplitudes expressed in eqn [3]. The overall transition rate is then found using eqn [1] and yields $W = W_1 + W_2 + Int$ where the interference term Int is given by

$$Int \propto E^{2\omega} E^{\omega} E^{\omega} \sin(\phi_{2\omega} - 2\phi_{\omega})k \qquad [6]$$

Note, however, that the interference effect depends on the sign of k and hence can be constructive for one part of the conduction band but destructive for other parts of that band, depending on the value of the relative phase, $\Delta\phi = \phi_{2\omega} - 2\phi_{\omega}$. In principle one can largely eliminate transitions with $+\mathbf{k}$ and enhance those with $-\mathbf{k}$. This effectively generates a net momentum for electrons or holes and hence, at least temporarily, leads to an electrical current in the absence of any external bias. The net momentum of the carriers, which is absent in the individual processes, must come from the lattice. Because the carriers are created with net momentum during the pulses one has a form of current injection that can be written as

$$dJ/dt \propto E^{2\omega}E^{\omega}E^{\omega}\sin(\Delta\phi) \qquad [7]$$

where J is the current density. This type of current injection is allowed in both centrosymmetric and noncentrosymmetric materials. The physics and concepts behind the quantum interference leading to this form of current injection are analogous with the Young's double slit experiment. Note as well that if this 'interferometer' is balanced (single- and two-photon transition rates are similar), then it is possible to control the overall transition rate with great contrast. Roughly speaking, one balances the two arms of the 'effective interferometer' involved in the interference process, one 'arm' corresponding to the one-photon process and the other to the two-photon process. As an example, let's consider the excitation of GaAs, which has a room-temperature bandgap of 1.42 eV (equivalent to 870 nm). For excitation of GaAs using 1550 and 775 nm light under balanced conditions, the electron cloud is injected with a speed close to 500 km s^{-1}. Under 'balanced' conditions, nearly all the electrons are moving in the same direction. Assuming that the irradiance of the 1550 nm beam is 100 MW cm^{-2}, while that of the second harmonic beam is only 15 kW cm^{-2} (satisfying the 'balance' condition), with Gaussian pulse widths of 100 fs, one obtains a surprisingly large peak current of about 1 kA cm^{-2} for a carrier density of only 10^{14} cm^{-3} if scattering effects are ignored. When scattering is taken into account, the value of the peak current is reduced and the transient current decays on a time-scale of the momentum relaxation time. Figure 2a shows an experimental setup which can be used to demonstrate coherence control of electrical current using the above parameters. Figure 2a shows the region between a pair of electrodes on GaAs being illuminated by a train of harmonically related pulses. Figure 2b illustrates how the steady-state experimental voltage across the capacitor changes as the phase parameter $\Delta\phi$ is varied. Transient electrical currents, generated though incident femtosecond optical pulses, have also been detected through the emission of the associated Terahertz radiation.

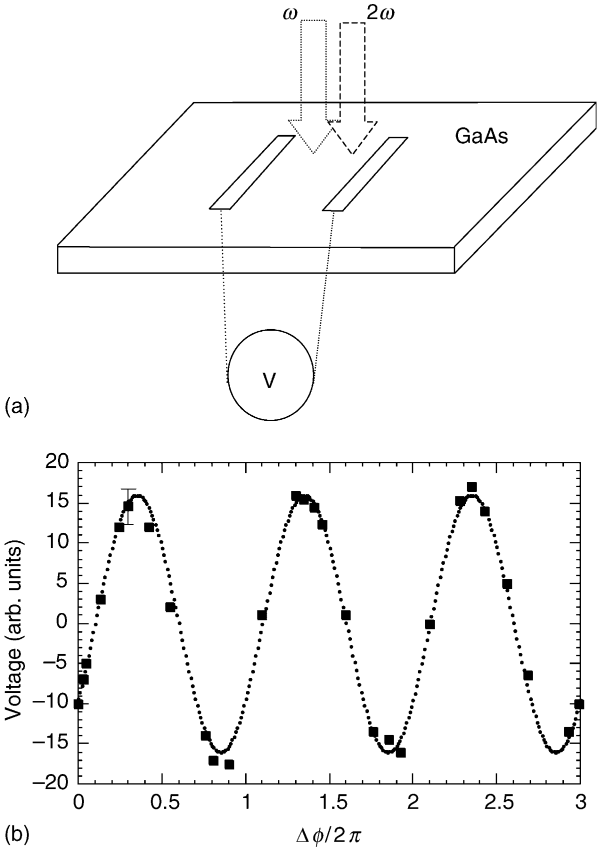

Figure 2 (a) Experimental setup to measure steady-state voltage (V) across a pair of electrodes on GaAs with the intervening region illuminated by phased radiation at ω and 2ω; (b) induced voltage across a pair of electrodes on a GaAs semiconductor as a function of the phase parameter associated with two harmonically related incident beams.

The current injection via coherence control and conventional cases (i.e., under a DC bias) differs with respect to their evolution. The current injected via coherent control has an onset determined by the rise time of the optical pulses. In the case of normal current production, existing carriers are accelerated by an electric field, and the momentum distribution is never far from isotropic. For a carrier density of 10^{14} cm^{-3} a DC field ~80 kV cm^{-1} is required to produce a current density of 1 kA cm^{-2}. In GaAs, with an electron mobility of 8000 cm^{2} V^{-1} s^{-1} this current would occur about 1/2 ps after the field is 'instantaneously' turned on. This illustrates that the coherently controlled phenomenon efficiently and quickly produces a larger current than can be achieved with the redirecting of statistically distributed electrons.

A more detailed analysis must take into account the actual light polarization, crystal symmetry, crystal face, and orientation relative to the optical polarization. For given optical intensities of the two beams, the maximum electrical current injection in the case of GaAs occurs for linearly polarized beams both oriented along the (111) or equivalent direction. However, large currents can also be observed with the light beams polarized along other high symmetry directions.

Coherent Control of Carrier Density, Spin Population, and Spin Current Using Two Color Beams

The processes described above do not exhaust the coherent control effects that can be observed in bulk semiconductors using harmonic beams. Indeed, for noncentrosymmetric materials, certain light polarizations and crystal orientations allow one to coherently control the total carrier generation rate with or without generating an electrical current. In the case of the cubic material GaAs, provided the ω beam has electric field components along two of the three principal crystal axes, with the 2ω beam having a component along the third direction, one can coherently control the total carrier density. However, the overall degree of control here is determined by how 'noncentrosymmetric' the material is.

The spin degrees of freedom of a semiconductor can also be controlled using two color beams. Due to the spin–orbit interaction, the upper valence bands of a typical semiconductor have certain spin characteristics. To date, optical manipulation of electron spin has been largely based on the fact that partially spin-polarized carriers can be injected in a semiconductor via one-photon absorption of circularly polarized light from these upper valence bands. In such carrier injection – where in fact two-photon absorption could be used as well – spins with no net velocity are injected, and then are typically dragged by a bias voltage to produce a spin-polarized current. However, given the protocols discussed above it should not come as a surprise that the two color coherence scheme, when used with certain light polarizations, can coherently control the spin polarization, making it dependent on $\Delta\phi$. Furthermore, for certain polarization combinations it is also possible to generate a spin current with or without an electrical current. Given the excitement surrounding the field of spintronics, where the goal is the use of the spin degree of freedom for data storage and processing, the control of quantities involving the intrinsic angular momentum of the electron is of particular interest.

Various polarization and crystal geometries can be examined for generating spin currents with or without electrical current. For example, with reference to Figure 3a, for both beams propagating in the **z** (*normal*) direction of a crystal and possessing the same circular polarization, an electrical current can be injected in the (*xy*) plane, at an angle from the crystallographic **x** direction dependent on the relative phase parameter, $\Delta\phi$; this current is spin-polarized in the **z** direction. As well, the injected carriers have a ±**z** component of their velocity, as many with one component as with the other; but those going in one direction are preferentially spin-polarized in one direction in the (*xy*) plane, while those in the other direction are preferentially spin-polarized in the opposite direction. This is an example of a spin current in the absence of an electrical current.

Such a pure spin current that is perhaps more striking is observable with the two beams cross linearly polarized, for example with the fundamental beam in the **x** direction and the second harmonic beam in the **y** direction as shown in Figure 3b. Then there is no net spin injection; the average spin in any

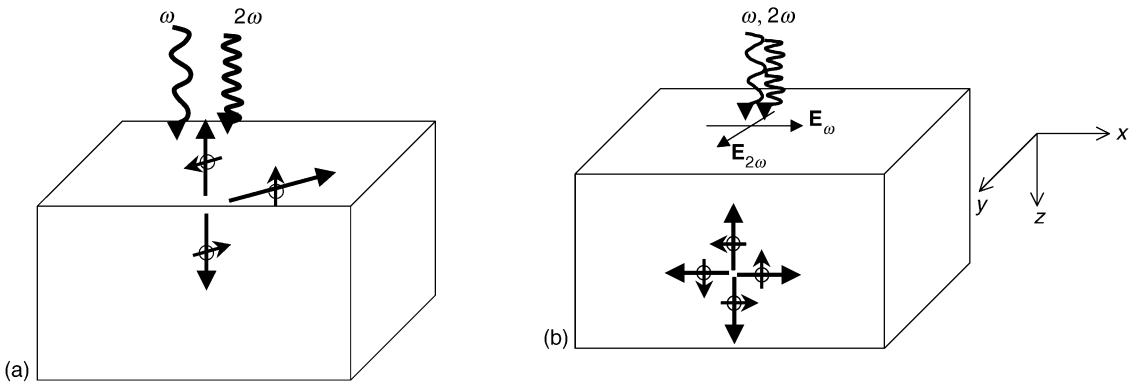

Figure 3 (a) Excitation of a semiconductor by co-circularly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the $x-y$ plane in a direction dependent on $\Delta\phi$ while a pure spin current is generated in the beam propagation (z) direction. (b) Excitation of a semiconductor by orthogonally, linearly polarized ω and 2ω pulses. Arrows indicate that a spin polarized electrical current is generated in the direction of the fundamental beam polarization as well as along the beam propagation (z) direction.

direction is zero. And for a vanishing phase parameter $\Delta\phi$ there is no net electrical current injection. Yet, for example, the electrons injected with a $+\mathbf{x}$ velocity component will have one spin polarization with respect to the **z** direction, while those injected with a $-\mathbf{x}$ component to their velocity will have the opposite spin polarization with respect to the **z** direction.

The examples given above are the spin current analog of the two-color electrical current injection discussed above. There should also be the possibility of injecting a spin current with a single beam into crystals lacking center-of-inversion symmetry.

The simple analysis presented above relies on simple quantum mechanical ideas and calculations using essentially nothing more than Fermi's Golden Rule. Yet the process involves a mixing of two frequencies, and can therefore be thought of as a nonlinear optical effect. Indeed, one of the key ideas to emerge from the theoretical study of such phenomena is that an interpretation of coherence control effects alternate to that provided by the simple quantum interference picture that is provided by the usual susceptibilities of nonlinear optics. These susceptibilities are, of course, based on quantum mechanics, but the macroscopic viewpoint allows for the identification and classification of the effects in terms of 2nd order nonlinear optical effects, 3rd order optical effects, etc. Indeed, one can generalize many of the processes we have discussed above to a hierarchy of frequency mixing effects or high-order nonlinear processes involving multiple beams with frequency $n\omega, m\omega, p\omega\ldots$ with n, m, p being integers. Many of theses schemes require high intensity of one or more of the beams, but can occur in a simple semiconductor.

Coherent Control of Electrical Current Using Single Color Beams

It is also possible to generate coherence control effects using beams at a single frequency (ω), if one focuses on the two orthogonal components (e.g., x and y) polarization states and uses a noncentrosymmetric semiconductor of reduced symmetry such as a strained cubic semiconductor or a wurtzite material such as CdS or CdSe. In this case, interference between absorption pathways associated with the orthogonal components can lead to electrical current injection given by

$$dJ/dt \propto E^{\omega}E^{\omega}\sin(\phi_{\omega}^{x} - \phi_{\omega}^{y}) \qquad [8]$$

Since in this process current injection is linear in the beam's intensity, the high intensities necessary for two-photon absorption are not necessary. Nonetheless, the efficacy of this process is limited by the fact that it relies on the breaking of center-of-inversion symmetry of the underlying crystal. It is also clear that the maximum current occurs for circularly polarized light and that right and left circularly polarized light lead to a difference in sign of the current injection. Finally, for this particular single beam scheme, when circularly polarized light is used the electrical current is partially spin polarized.

Conclusions

Through the phase of optical pulses it is possible to control electrical and spin currents, as well as carrier density, in bulk semiconductors on a time-scale that is limited only by the rise time of the optical pulse and the intrinsic response of the semiconductor. A few optically based processes that allow one to achieve these types of control have been illustrated here. Although at one level one can understand these processes in terms of quantum mechanical interference effects, at a macroscopic level one can understand these control phenomena as manifestations of nonlinear optical phenomena. For fundamental as well as applied reasons, our discussion has focused on coherence control effects using the continuum states in bulk semiconductors, although related effects can also occur in quantum dot, quantum well, and superlattice semiconductors. Applications of these control effects will undoubtedly exploit the all-optical nature of the process, including the speed at which the effects can be turned on or off. The turn-off effects, although not discussed extensively here, are related to transport phenomena as well as momentum scattering and related dephasing processes.

See also

Semiconductor Physics: Spin Transport and Relaxation in Semiconductors; Spintronics. **Interferometry**: Overview. **Coherent Control:** Experimental; Theory.

Further Reading

Brumer P (1995) Laser control of chemical reactions. *Scientific American* 272: 56.

Rabitz H, de Vivie-Riedle R, Motzkus M and Kompa K (2000) Whither the future of controlling quantum phenomena? *Science* 288: 824.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. New York: Springer.

van Driel HM and Sipe JE (2001) Coherent control of photocurrents in semiconductors. In: Tsen T (ed.) *Ultrafast Phenomena in Semiconductors*, p. 261. New York: Springer Verlag.

COHERENT LIGHTWAVE SYSTEMS

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical communication systems are based on the intensity modulation (IM) of an optical carrier by an electrical data signal and direct detection (DD) of the received light. The simplest scheme employs on/off keying (OOK), whereby turning on or off, an optical source transmits a binary 1 or 0. The signal light is transmitted down an optical fiber and at the receiver is detected by a photodetector, as shown in Figure 1. The resulting photocurrent is then processed to determine if a 1 or 0 was received. In the ideal case, where a monochromatic light source is used and where no receiver noise (detector dark current and thermal noise) is present, the probability of error is determined by the quantum nature (shot noise) of the received light. In this case, the number of photons/bit required (sensitivity) at the receiver, to achieve a bit-error-rate (BER) of 10^{-9}, is 10. This is called the quantum-limit. In practical receivers without an optical preamplifier, operating in the 1.3 μm and 1.55 μm optical fiber communication windows, the actual sensitivity is typically 10–30 dB less than the theoretical sensitivity. This is due to receiver noise, in particular thermal noise. This means that the actual sensitivity is between 100 to 10 000 photons/bit. The capacity of IM/DD fiber links can be increased using wavelength division multiplexing (WDM). At the receiver the desired channel is selected using a narrowband optical filter. Some advanced commercial IM/DD systems utilize dense WDM with 50 GHz spacing and single-channel bit rates as high as 40 Gb/s. Indeed, IM/DD systems with terabit (10^{12} bit/s) capacity are now possible.

Compared to IM/DD systems, coherent optical communication systems have greater sensitivity and selectivity. In the context of coherent lightwave systems, the term coherent refers to any technique employing nonlinear mixing between two optical waves on a photodetector, as shown in Figure 2. Typically, one of these is an information-bearing signal and the other is a locally generated wave (local oscillator). The result of this heterodyne process is a modulation of the photodetector photocurrent at a frequency equal to the difference between the signal and local oscillator frequencies. This intermediate frequency (IF) electronic signal contains the information in the form of amplitude, frequency, or phase that was present in the original optical signal. The IF signal is filtered and demodulated to retrieve the transmitted information. An automatic frequency control circuit is required to keep the local-oscillator frequency stable.

It is possible for the information signal to contain a number of subcarriers (typically at microwave frequencies), each of which can be modulated by a separate data channel. It is a simple matter to select the desired channel at the receiver by employing electronic heterodyning and low pass filtering, as shown in Figure 3. In IM/DD systems, channel selection can only be carried out using narrowband optical filters.

Coherent optical communications utilize techniques that were first investigated in radio communications. Most of the basic research work on coherent optical communications was carried out in the 1980s and early 1990s, and was primarily motivated by the need for longer fiber links using no repeaters. Improving receiver sensitivity using coherent techniques made it possible to increase fiber link spans. In the mid-1990s, reliable optical fiber amplifiers became available. This made it possible to construct fiber links using optical amplifiers spaced at appropriate intervals to compensate for fiber and other transmission losses.

The sensitivity of an IM/DD receiver can be greatly improved by the use of an optical preamplifier. Indeed the improvement is so marked that coherent techniques are not a viable alternative to IM/DD in the vast majority of commercial optical communication systems due to their complexity and higher cost. Moreover, because semiconductor lasers with very precise wavelengths and narrowband optical filters can be fabricated, the selectivity advantage of coherent optical communications has become less important. This has meant that the research work currently being carried out on coherent optical communications is very limited, and in particular, very few field trials have been carried out since the early 1990s.

We will review the basic principles underlying coherent optical communications, modulation schemes, detection schemes, and coherent receiver sensitivity. The effects of phase noise and polarization on coherent receiver performance will be outlined. A brief summary of pertinent experimental results will also be presented.

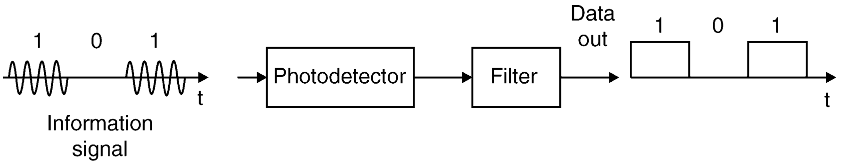
1 0 1
t
Information signal
Photodetector
Filter
Data out
1 0 1
t

Figure 1 IM/DD optical receiver.

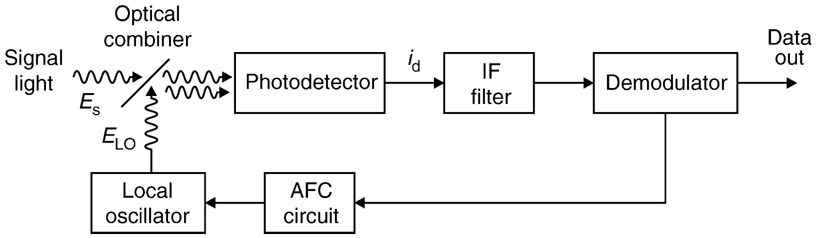
Optical combiner
Signal light
E_s
E_{LO}
Photodetector
i_d
IF filter
Demodulator
Data out
Local oscillator
AFC circuit

Figure 2 Basic coherent optical receiver. AFC: automatic frequency control.

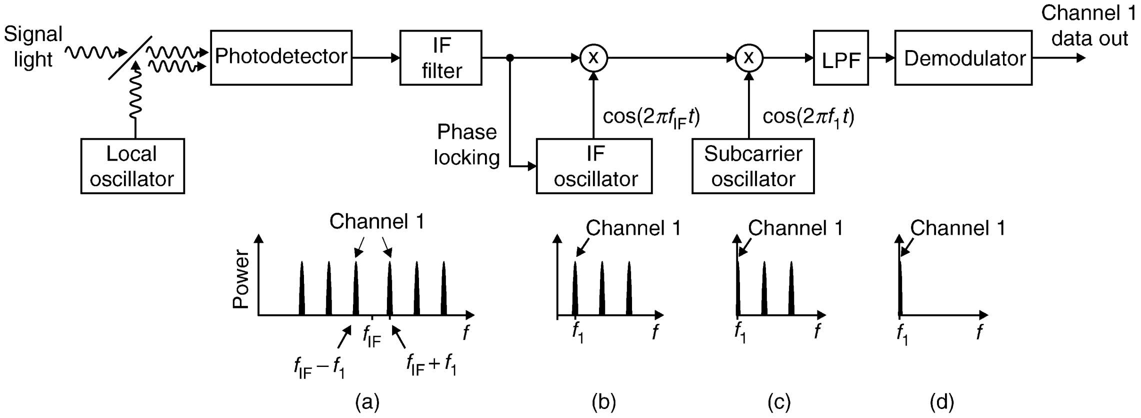
Signal light
Photodetector
IF filter
Phase locking
IF oscillator
$\cos(2\pi f_{IF}t)$
Subcarrier oscillator
$\cos(2\pi f_1 t)$
LPF
Demodulator
Channel 1 data out
Local oscillator
Power
Channel 1
f_{IF}
$f_{IF}-f_1$
$f_{IF}+f_1$
f
f_1
(a)
(b)
(c)
(d)

Figure 3 Coherent detection and processing of a subcarrier modulated signal lightwave. The power spectrum of the detected signal is shown: (a) after IF filtering; (b) after multiplication by the IF oscillator; (c) after multiplication with the subcarrier oscillator (in this case channel 1's carrier frequency); and (d) after low pass filtering prior to demodulation and data detection. LPF: lowpass filter.

Basic Principles

In coherent detection, a low-level optical signal field E_s is combined with a much larger power optical signal field E_{LO} from a local oscillator laser. E_s and E_{LO} can be written as

$$E_s(t) = \sqrt{2P_s}\cos[2\pi f_s t + \phi_s(t)] \qquad [1]$$

$$E_{LO}(t) = \sqrt{2P_{LO}}\cos[2\pi f_{LO} t + \phi_{LO}(t)] \qquad [2]$$

where P_s, f_s, and ϕ_s are the signal power, optical frequency, and phase (including phase noise), respectively. P_{LO}, f_{LO}, and ϕ_{LO} are the equivalent quantities for the local oscillator. For ideal coherent detection, the polarization states of the signal and local oscillator must be equal. If this is the case, the photocurrent is given by

$$i_d(t) = R[E_s(t) + E_{LO}(t)]^2 \qquad [3]$$

where R is the detector responsivity. Because the detector cannot respond to optical frequencies, we have

$$\begin{aligned} i_d(t) = R\big\{&P_s + P_{LO} + 2\sqrt{P_s P_{LO}} \\ &\times \cos[2\pi f_{IF} t + \phi_s(t) - \phi_{LO}(t)]\big\} \end{aligned} \qquad [4]$$

The IF signal is centered at frequency $f_{IF} = f_s - f_{LO}$. In homodyne detection $f_s = f_{LO}$ and the homodyne signal is centered at baseband. In heterodyne detection $f_s \neq f_{LO}$ and the signal generated by the photodetector is centered around f_{IF} (typically three to six times the bit rate). The IF photocurrent is proportional to $\sqrt{P_s}$, rather than P_s as is the case in direct detection, and is effectively amplified by a factor proportional to $\sqrt{P_{\mathrm{LO}}}$. If P_{LO} is large enough, the signal power can be raised above the receiver noise, thereby leading to a greater sensitivity than possible with conventional IM/DD receivers using p–i–n or avalanche photodiodes. Theoretically it is possible to reach shot noise limited receiver sensitivity although this is not possible in practice. Homodyne optical receivers usually require less bandwidth and are more sensitive than heterodyne receivers but require that the local oscillator is phase locked to the information signal. This requires an optical phase locked loop (OPLL), which is very difficult to design.

Modulation Schemes

The three principal modulation schemes used in coherent optical communications; amplitude shift keying (ASK), frequency shift keying (FSK) and phase shift keying (PSK), are shown in Figure 4. ASK modulation is essentially the same as the OOK scheme used in IM/DD systems. The optical signal is given by

$$E_{\mathrm{s}}(t) = 2a(t)\sqrt{2P_{\mathrm{s}}}\cos[2\pi f_{\mathrm{s}}t + \phi_{\mathrm{s}}(t)] \qquad [5]$$

where $a(t) = 0$ or 1 for transmission of a 0 or 1, respectively. P_{s} is the average signal power, assuming that it is equally likely that a 0 or 1 is transmitted. ASK modulation can be achieved by modulating the output of a semiconductor laser using an external modulator. The most common type is based on $\mathrm{LiNbO_3}$ waveguides in a Mach–Zehnder interferometer configuration.

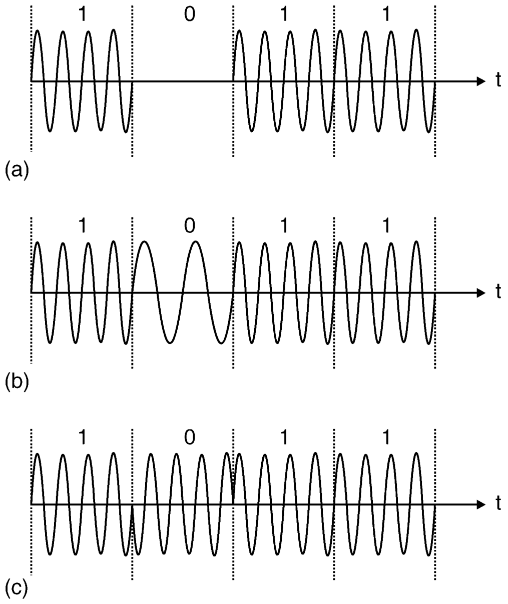

Figure 4 Modulation schemes for coherent optical communications: (a) ASK; (b) FSK; and (c) PSK.

In FSK modulation, the optical frequency of the signal lightwave is changed slightly by the digital data stream. In binary FSK, the optical signal is given by

$$E_{\mathrm{s}}(t) = \sqrt{2P_{\mathrm{s}}}\cos[2\pi(f_{\mathrm{s}} + a(t)\Delta f)t + \phi_{\mathrm{s}}(t)] \qquad [6]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The choice of the frequency deviation Δf depends on the available bandwidth, the bit rate B_{T} and the demodulation scheme used in the receiver. The modulation index m is given by

$$m = \frac{2\Delta f}{B_{\mathrm{T}}} \qquad [7]$$

The minimum value of m that produces orthogonality (for independent detection of the 0s and 1s) is 0.5. If $m = 0.5$, the signal is referred to as minimum FSK (MFSK). When $m \geq 2$, the signal is referred to as wide-deviation FSK, and when $m < 2$, as narrow-deviation FSK. Wide-deviation FSK receivers are more resilient to phase noise than narrow-deviation FSK receivers. FSK modulation can be achieved using $\mathrm{LiNbO_3}$ external modulators, acoustic optic modulators and distributed feedback semiconductor lasers. The phase of the lightwave does not change between bit transitions.

In PSK modulation, the digital data stream changes the phase of the signal lightwave. In binary PSK the optical signal is given by

$$E_{\mathrm{s}}(t) = \sqrt{2P_{\mathrm{s}}}\cos\left[2\pi f_{\mathrm{s}}t + a(t)\theta + \frac{\pi}{2} + \phi_{\mathrm{s}}(t)\right] \qquad [8]$$

where $a(t) = -1$ or 1 for transmission of a 0 or 1, respectively. The phase shift between a 0 and 1 is $\theta + \pi/2$. Most PSK schemes use $\theta = \pi/2$. PSK modulation can be achieved by using an external phase modulator or a multiple quantum well electro-absorption modulator.

Detection Schemes

There are two classes of demodulation schemes: synchronous and asynchronous. The former exploits the frequency and phase of the carrier signal to perform the detection. The latter only uses the envelope of the carrier to perform the detection. In the following we review the principal demodulation schemes.

PSK Homodyne Detection

In an ideal PSK homodyne receiver, shown in Figure 5, the signal light and local oscillator have identical

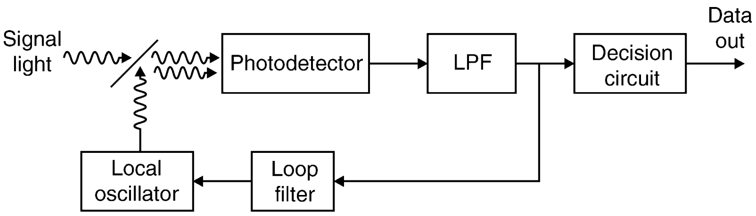
Signal light
Photodetector
LPF
Decision circuit
Data out
Local oscillator
Loop filter

Figure 5 PSK homodyne receiver.

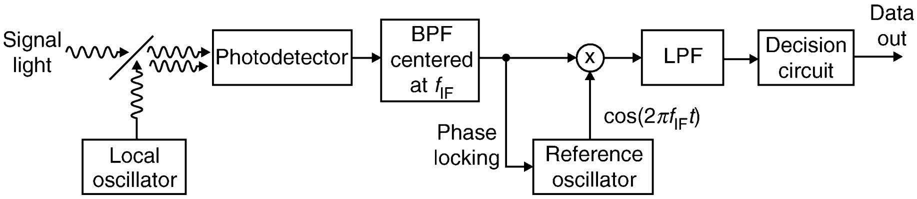
Signal light
Photodetector
BPF centered at f_{IF}
x
LPF
Decision circuit
Data out
$\cos(2\pi f_{IF}t)$
Phase locking
Reference oscillator
Local oscillator

Figure 6 PSK synchronous heterodyne receiver. BPF: bandpass filter.

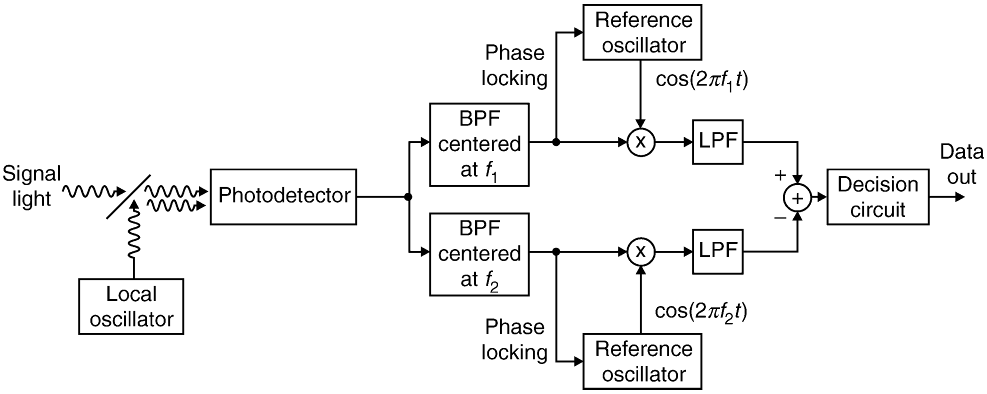
Reference oscillator
Phase locking
$\cos(2\pi f_1 t)$
BPF centered at f_1
x
LPF
Signal light
Photodetector
+
+
−
Decision circuit
Data out
BPF centered at f_2
x
LPF
$\cos(2\pi f_2 t)$
Phase locking
Reference oscillator
Local oscillator

Figure 7 FSK heterodyne synchronous receiver.

optical frequency and phase. If the resulting baseband information signal is positive, then a 1 has been received. If the baseband information signal is negative, then a 0 bit has been received. This scheme has the highest theoretical sensitivity but requires an OPLL and is also highly sensitive to phase noise.

ASK Homodyne Detection

An ideal ASK homodyne receiver is identical to an ideal PSK homodyne receiver. Because the baseband information signal is either zero (0 bit) or nonzero (1 bit), the receiver sensitivity is 3 dB less than the PSK homodyne receiver.

PSK Heterodyne Synchronous Detection

Figure 6 shows an ideal PSK synchronous heterodyne receiver. The IF signal is filtered by a bandpass filter, multiplied by the phase locked reference oscillator to move the signal to baseband, low pass filtered and sent to a decision circuit that decides if the received bit is a 0 or 1. The receiver requires synchronization of the reference oscillator with the IF signal. This is not easy to achieve in practice because of the large amount of semiconductor laser phase noise.

ASK Heterodyne Synchronous Detection

This scheme is basically the same as the PSK heterodyne detection and is similarly difficult to design.

FSK Heterodyne Synchronous Detection

This scheme is shown in Figure 7 for a binary FSK signal consisting of two possible frequencies $f_1 = f_s - \Delta f$ and $f_2 = f_s + \Delta f$. There are two separate branches in which correlation with the two possible IF signals at f_1 and f_2 is performed. The two resulting signals are subtracted from each other and a decision taken on the sign of the difference. Because of the requirement for two electrical phase locked loops, the scheme is not very practical.

ASK Heterodyne Envelope Detection

In this scheme, shown in Figure 8, the IF signal is envelope detected to produce a baseband signal proportional to the original data signal. The output of the envelope detector is relatively insensitive to laser phase noise.

FSK Heterodyne Dual-Filter Detection

In this scheme, shown in Figure 9, two envelope detectors are used to demodulate the two possible IF signals at f_1 and f_2 and a decision taken based on the outputs. The scheme is tolerant to phase noise and can achieve high sensitivity. However, it needs high-bandwidth receiver electronics because a large frequency deviation is required.

FSK Heterodyne Single-Filter Detection

This scheme omits one branch of the FSK heterodyne dual-filter receiver and performs a decision in the same way as for the ASK heterodyne envelope detection. Because half the power of the detected signal is not used, the sensitivity of the scheme is 3 dB worse than the dual-filter scheme.

CPFSK Heterodyne Differential Detection

This scheme, shown in Figure 10, uses the continuous phase (CP) characteristic of an FSK signal, i.e., there is no phase discontinuity at bit transition times. After bandpass filtering, the IF photocurrent is

$$i_{\mathrm{IF}}(t) = A\cos\{2\pi[(f_{\mathrm{IF}} + a(t)\Delta f)]t\} \qquad [9]$$

where $A = R\sqrt{P_s P_{\mathrm{LO}}}$. The output from the delay-line demodulator, after low pass filtering to remove the double IF frequency term, is

$$x(t) = \frac{A^2}{2}\cos\{2\pi[f_{\mathrm{IF}} + a(t)\Delta f]\tau\} \qquad [10]$$

The function of the delay-line demodulator is to act as a frequency discriminator. We require that $x(t)$ is a maximum when $a(t) = 1$ and a minimum when $a(t) = -1$. This is the case if the following relations are satisfied:

$$2\pi(f_{\mathrm{IF}} + \Delta f)\tau = 2\pi k \qquad [11]$$

$$2\pi(f_{\mathrm{IF}} - \Delta f)\tau = (2k - 1)\pi, \qquad k \text{ an integer}$$

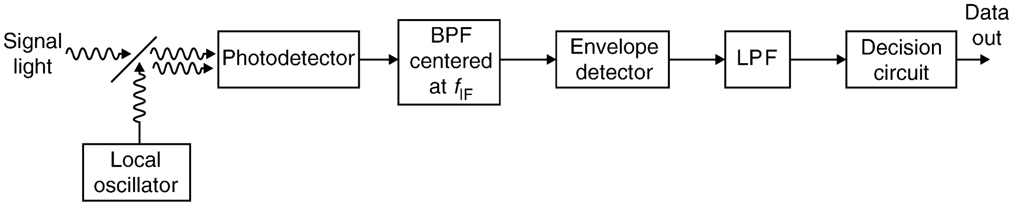
Signal light
Photodetector
BPF centered at f_{IF}
Envelope detector
LPF
Decision circuit
Data out
Local oscillator

Figure 8 ASK heterodyne envelope receiver.

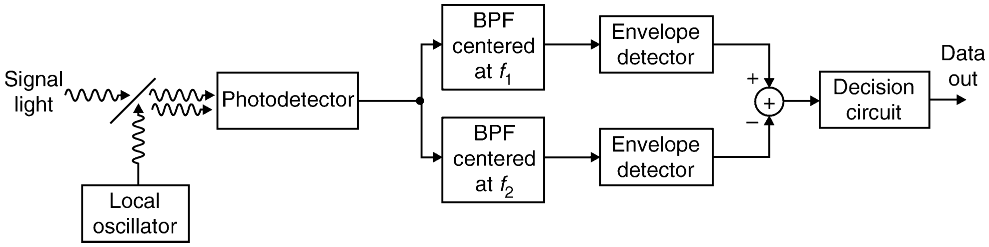
Signal light
Photodetector
BPF centered at f_1
BPF centered at f_2
Envelope detector
Envelope detector
+
−
+
Decision circuit
Data out
Local oscillator

Figure 9 FSK heterodyne dual-filter receiver.

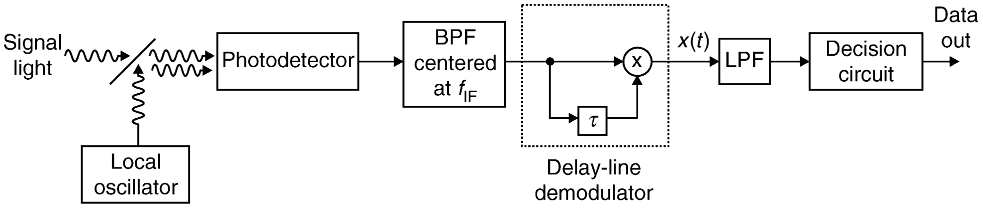
Signal light
Photodetector
BPF centered at f_{IF}
x
τ
Delay-line demodulator
x(t)
LPF
Decision circuit
Data out
Local oscillator

Figure 10 CPFSK heterodyne differential detection receiver.

Hence, we require that

$$\Delta f \tau = \frac{1}{4} \qquad [12]$$

$$f_{\mathrm{IF}}\tau = \left[k - \frac{1}{4}\right], \qquad k \text{ integer}$$

In terms of the modulation index we have

$$\tau = \frac{1}{2mB_{\mathrm{T}}} \qquad [13]$$

The above equation shows that it is possible to decrease the value of τ required by increasing m, thereby reducing the sensitivity of the receiver to phase noise. The minimum value of m possible is 0.5. This scheme can operate with a smaller modulation index compared to the dual-filter and single-filter schemes for a given bit rate, thereby relaxing the bandwidth requirements of the receiver electronics.

DPSK Heterodyne Differential Detection

In differential phase shift key (DPSK) modulation, a 0 is sent by changing the phase of the carrier by π with respect to the previous signal. A 1 is sent by not changing the phase. The receiver configuration is identical to the CPFSK heterodyne differential detection receiver, except that the time delay τ is equal to the inverse of the data rate. Compared to other PSK detection schemes, DPSK differential detection is relatively simple to implement and is relatively immune to phase noise.

Coherent Receiver Sensitivity

The signal-to-noise ratio γ_{c} of an ideal PSK homodyne receiver, in the shot-noise limit, is given by

$$\gamma_{\mathrm{c}} = \frac{2RP_{\mathrm{s}}}{eB_{\mathrm{c}}} \qquad [14]$$

where B_{c} is the receiver bandwidth. The detector responsivity is given by

$$R = \frac{\eta e}{hf_{\mathrm{s}}} \qquad [15]$$

where η is the detector quantum efficiency (≤ 1). The average signal power $P_{\mathrm{s}} = N_{\mathrm{s}}hf_{\mathrm{s}}B_{\mathrm{T}}$, where N_{s} is the average number of photons per bit. If it is assumed that $B_{\mathrm{c}} = B_{\mathrm{T}}$, then

$$\gamma_{\mathrm{c}} = 2\eta N_{\mathrm{s}} \qquad [16]$$

The actual detected number of photons per bit $N_{\mathrm{R}} = \eta N_{\mathrm{s}}$, so

$$\gamma_{\mathrm{c}} = 2N_{\mathrm{R}} \qquad [17]$$

If the noise is assumed to have a Gaussian distribution, then the bit error rate (BER) is given by

$$\mathrm{BER} = \frac{1}{2}\mathit{erfc}\left(\sqrt{2N_{\mathrm{R}}}\right) \qquad [18]$$

where *erfc* is the complementary error function; $\mathit{erfc}(x) \approx \mathrm{e}^{-x^2}/(x\sqrt{\pi})$ for $x > 5$. The receiver sensitivity for a $\mathrm{BER} = 10^{-9}$ is 9 photons/bit.

BER expressions and sensitivity for the demodulation schemes discussed above are listed in Table 1. The sensitivity versus N_{R} is plotted for some of the demodulation schemes in Figure 11. The received signal power required for a given BER is proportional to B_{T}. This dependency is shown in Figure 12 for some of the demodulation schemes.

Table 1 Shot-noise limited BER expressions and sensitivity for coherent demodulation schemes

Modulation scheme	*BER*	*Sensitivity (photons/bit)*
Homodyne		
PSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{2N_{\mathrm{R}}}\right)$	9
ASK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{N_{\mathrm{R}}}\right)$	18 (peak 36)
Synchronous heterodyne		
PSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{N_{\mathrm{R}}}\right)$	18
ASK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{\frac{N_{\mathrm{R}}}{2}}\right)$	36 (peak 72)
FSK	$\frac{1}{2}\mathit{erfc}\left(\sqrt{\frac{N_{\mathrm{R}}}{2}}\right)$	36
Asynchronous heterodyne		
ASK (envelope)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{2}\right)$	40 (peak 80)
FSK (dual-filter)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{2}\right)$	40
FSK (single-filter)	$\frac{1}{2}\exp\left(-\frac{N_{\mathrm{R}}}{4}\right)$	80
Differential detection		
CPFSK	$\frac{1}{2}\exp(-N_{\mathrm{R}})$	20
DPSK	$\frac{1}{2}\exp(-N_{\mathrm{R}})$	20

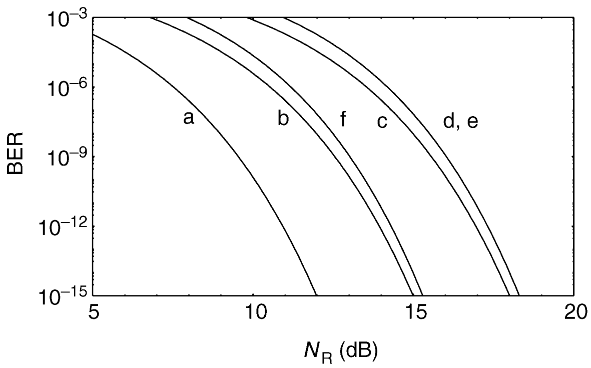

Figure 11 Shot-noise limited BER versus number of received photons/bit for various demodulation schemes: (a) PSK homodyne; (b) synchronous PSK; (c) synchronous FSK; (d) ASK envelope; (e) FSK dual-filter; and (f) CPFSK and DPSK.

Phase Noise

A major influence on the sensitivity possible with practical coherent receivers is laser phase noise, which is responsible for laser linewidth (3 dB bandwidth of the laser power spectrum). The effect of phase noise on system performance depends on the modulation and demodulation scheme used. In general, the effects of phase noise are more severe in homodyne and synchronous heterodyne schemes than in asynchronous schemes. If $\Delta\nu_{\mathrm{IF}}$ is the beat linewidth (without modulation present) between the signal and local oscillator lasers, the normalized beat linewidth $\Delta\nu$ is defined as

$$\Delta\nu = \frac{\Delta\nu_{\mathrm{IF}}}{B_{\mathrm{T}}} \qquad [19]$$

where $\Delta\nu_{\mathrm{IF}}$ is the IF phase noise bandwidth. Semiconductor lasers usually have a Lorentzian spectrum in which case $\Delta\nu_{\mathrm{IF}}$ is equal to the sum of the signal and local-oscillator laser linewidths. The requirements placed on $\Delta\nu$ become less severe as the bit rate increases. Table 2 compares the $\Delta\nu$ requirement for a 1 dB power penalty at a BER of 10^{-9} for some of the principal modulation/demodulation schemes.

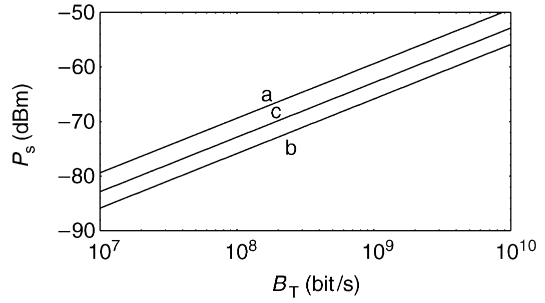

Figure 12 Required signal power versus bit rate for a BER = 10^{-9} for various shot-noise limited demodulation schemes: (a) PSK homodyne; (b) ASK envelope; and (c) CPFSK. The signal wavelength is 1550 nm and the detector is assumed to have unity quantum efficiency.

Table 2 Required normalized beat linewidth at 1 dB power penalty for a BER = 10^{-9} for various modulation/demodulation schemes

	Modulation scheme	*$\Delta\nu$ (%)*
Homodyne	PSK	6×10^{-4} (balanced loop) 0.01 (decision-driven loop)
Synchronous heterodyne	PSK	0.2–0.5
Asynchronous heterodyne	PSK	0.3–0.7
	ASK (envelope)	3
	FSK (dual-filter)	10–20
Differential detection	CPFSK	0.33 ($m = 0.5$) 0.66 ($m = 1.0$)
	DPSK	0.33

PSK Homodyne Detection – Phase Locked Loop Schemes

Practical homodyne receivers require an OPLL to lock the local-oscillator frequency and phase to that of the signal lightwave. There are a number of OPLL schemes possible for PSK homodyne detection, including the balanced PLL, Costas-type PLL and the decision-driven PLL.

The balanced PLL scheme, shown in Figure 13, requires that the PSK signal uses a phase shift of less than π between a 0 and 1, $E_{\mathrm{s}}(t)$ and $E_{\mathrm{LO}}(t)$ can be written as

$$E_{\mathrm{s}}(t) = \sqrt{2P_{\mathrm{s}}}\cos\left[2\pi f_0 t + a(t)\theta + \phi_{\mathrm{N,s}}(t) + \frac{\pi}{2}\right] \qquad [20]$$

$$E_{\mathrm{LO}}(t) = \sqrt{2P_{\mathrm{LO}}}\cos[2\pi f_0 t + \phi_{\mathrm{N,LO}}(t)] \qquad [21]$$

where f_0 is the common signal and local oscillator optical frequency and $\phi_{\mathrm{N,s}}(t)$ and $\phi_{\mathrm{N,LO}}(t)$ are the signal and local oscillator phase noise, respectively. When $\theta \neq \pi/2$, $E_{\mathrm{s}}(t)$ contains an unmodulated carrier component (residual carrier) in quadrature with the information-bearing signal. $E_{\mathrm{s}}(t)$ and $E_{\mathrm{LO}}(t)$ are combined by an optical 180° hybrid. The output optical signals from the π-hybrid are given by

$$E_1(t) = \frac{1}{\sqrt{2}}[E_{\mathrm{s}}(t) + E_{\mathrm{LO}}(t)] \qquad [22]$$

$$E_2(t) = \frac{1}{\sqrt{2}}[E_{\mathrm{s}}(t) - E_{\mathrm{LO}}(t)] \qquad [23]$$

The voltage at the output of the summing circuit is given by

$$\begin{aligned} v_{\mathrm{L}}(t) = {} & 2RR_{\mathrm{L}}\sqrt{P_{\mathrm{s}}P_{\mathrm{LO}}}\{a(t)\sin\theta\cos[\phi_{\mathrm{e}}(t)] \\ & + \cos\theta\sin[\phi_{\mathrm{e}}(t)]\} \end{aligned} \qquad [24]$$

where the photodetectors have the same responsivity and load resistance R_{L}. The phase error is given by

$$\phi_{\mathrm{e}}(t) = \phi_{\mathrm{N,s}}(t) - \phi_{\mathrm{N,LO}}(t) - \phi_{\mathrm{c}}(t) \qquad [25]$$

$\phi_{\mathrm{c}}(t)$ is the controlled phase determined by the control signal $v_{\mathrm{c}}(t)$ at the output of the loop filter. v_{L} contains two terms; an information-bearing signal proportional to $\sin\theta$, which is processed by the data detection circuit, and a phase error signal proportional to $\cos\theta$ used by the PLL for locking. The power penalty, due to the residual carrier transmission, is $10\log_{10}(1/\sin^2\theta)$ dB.

This scheme is not practical because the $\Delta\nu$ needed for a power penalty of <1 dB is typically less than 10^{-5}. This would require that $\Delta\nu_{\mathrm{IF}} < 10$ kHz for

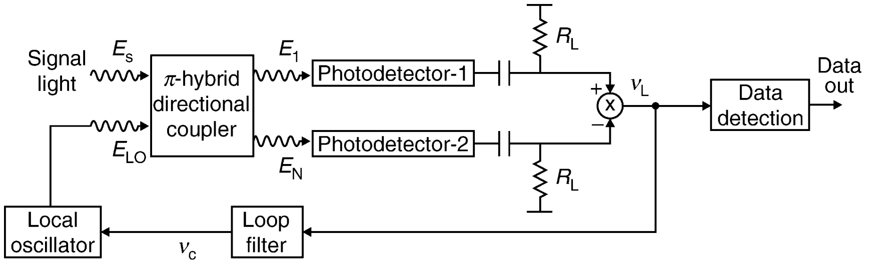
Signal light
E_s
E_1
π-hybrid directional coupler
Photodetector-1
Photodetector-2
E_{LO}
E_N
R_L
R_L
+
−
v_L
Data detection
Data out
Local oscillator
v_c
Loop filter

Figure 13 PSK homodyne balanced PLL receiver.

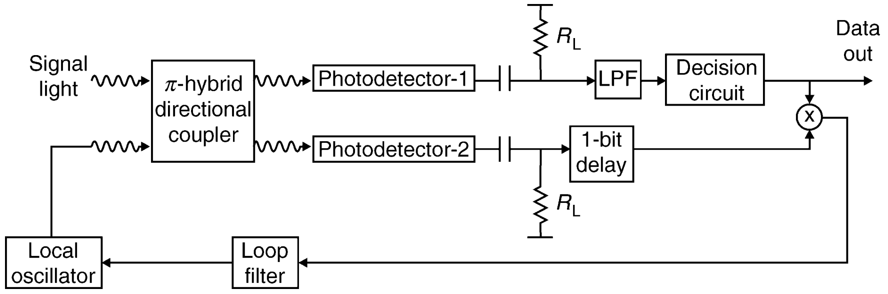
Signal light
π-hybrid directional coupler
Photodetector-1
Photodetector-2
R_L
R_L
LPF
Decision circuit
Data out
1-bit delay
Local oscillator
Loop filter

Figure 14 PSK decision-driven PLL.

a data rate of 1 Gb/s. The narrow signal and local-oscillator linewidths required are only achievable with external cavity semiconductor lasers.

The Costas-type and decision-driven PLLs also utilize π-hybrids and two photodetectors but as they have full suppression of the carrier ($\theta = \pi/2$), nonlinear processing of the detected signals is required to produce a phase error signal that can be used by the PLL for locking. In the decision-driven PLL, shown in Figure 14, the signal from one of the photodetector outputs is sent to a decision circuit and the output of the circuit is multiplied by the signal from the second photodetector. The mixer output is sent back to the local oscillator laser for phase locking. The $\Delta\nu$ required for a power penalty of <1 dB is typically less than 2×10^{-4}. This is superior to both the balanced PLL and Costas-type PLL receivers.

Heterodyne Phase Locking

In practical synchronous heterodyne receivers phase locking is performed in the electronic domain utilizing techniques from radio engineering. In the PSK case, electronic analogs of the OPLL schemes used in homodyne detection can be used. Synchronous heterodyne schemes have better immunity to phase noise than the homodyne schemes.

Asynchronous Systems

Asynchronous receivers do not require optical or electronic phase locking. This means that they are less sensitive to phase noise than synchronous receivers. They do have some sensitivity to phase noise because IF filtering leads to phase-to-amplitude noise conversion.

Differential Detection

Differential detection receivers have sensitivities to phase noise between synchronous and asynchronous receivers because, while they do not require phase locking, they use phase information in the received signal.

Phase-Diversity Receivers

Asynchronous heterodyne receivers are relatively insensitive to phase noise but require a much higher receiver bandwidth for a given bit rate. Homodyne receivers only require a receiver bandwidth equal to the detected signal bandwidth but require an OPLL, which is difficult to implement. The phase diversity receiver, shown in Figure 15, is an asynchronous homodyne scheme, which does not require the use of an OPLL and has a bandwidth approximately equal to synchronous homodyne detection. This is at the expense of increased receiver complexity and reduced sensitivity compared to synchronous homodyne detection. The phase-diversity scheme can be used with ASK, DPSK, and CPFSK modulation. ASK modulation uses an squarer circuit for the demodulator component, while DPSK and CPFSK modulation use a delay line and mixer.

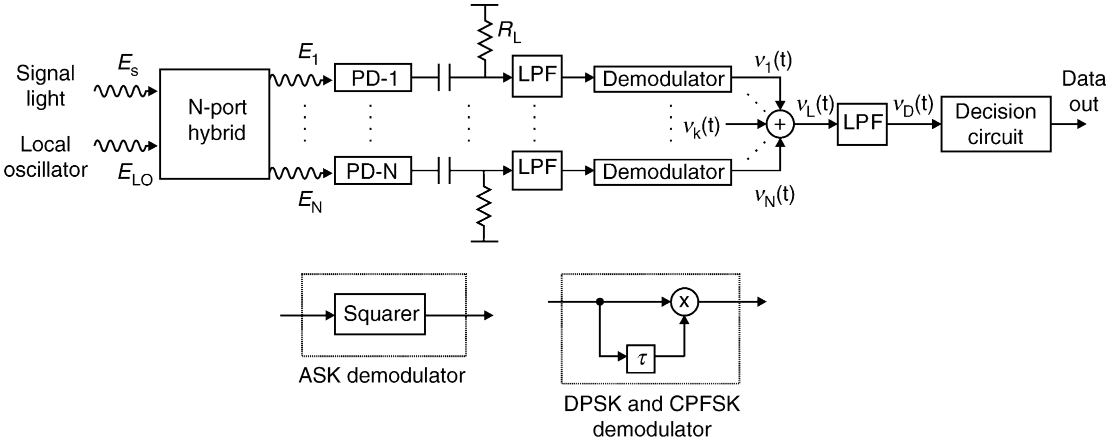

ASK demodulator

DPSK and CPFSK demodulator

Figure 15 Multiport phase-diversity receiver. PD: photodetector.

In the case of ASK modulation, the outputs of the N-port hybrid can be written as

$$E_k(t) = \frac{1}{\sqrt{N}}\left\{a(t)\sqrt{2P_s}\cos\left[2\pi f_s t + \phi_s(t) + \frac{2\pi k}{N}\right] + \sqrt{2P_{LO}}\cos[2\pi f_{LO}t + \phi_{LO}(t)]\right\}, \quad k = 1,2,\ldots,N,\ N \geq 3 \qquad [26]$$

For $N = 2$, the phase term is $\pi/2$ for $k = 2$. The voltage input to the k-th demodulator (squarer), after the dc terms have been removed by the blocking capacitor, is given by

$$v_k(t) = a(t)\frac{RR_L}{N}2\sqrt{P_s P_{LO}}\cos\left[2\pi(f_s - f_{LO}) + \phi_s(t) - \phi_{LO}(t) + \frac{2\pi k}{N}\right] \qquad [27]$$

These voltages are then squared and added to give an output voltage

$$v_L(t) = \alpha\left(\frac{RR_L}{N}\right)^2 4P_s P_{LO} a(t) \sum_{k=1}^{N}\left\{1 + \cos\left[4\pi \times \left(f_s - f_{LO}\right)t + 2\phi_s(t) - 2\phi_{LO}(t) + \frac{4\pi k}{N}\right]\right\} \qquad [28]$$

where α is the squarer parameter. The lowpass filter at the summing circuit output effectively integrates $v_L(t)$ over a bit period. If $2(f_s - f_{LO}) \ll B_T$ and the difference between the signal and local oscillator phase noise is small within a bit period, then the input voltage to the decision circuit is

$$v_D(t) = \alpha\frac{(RR_L)^2}{N}4P_s P_{LO} a(t) \qquad [29]$$

which is proportional to the original data signal. In shot-noise limited operation, the receiver sensitivity is 3 dB worse than the ASK homodyne case. The scheme is very tolerant of laser phase noise. The $\Delta\nu$ for a power penalty of <1 dB is of the same order as for ASK heterodyne envelope detection.

Polarization

The mixing efficiency between the signal and local-oscillator lightwaves is a maximum when their polarization states are identical. In practice the polarization state of the signal arriving at the receiver is unknown and changes slowly with time. This means that the IF photocurrent can change with time and in the worst case, when the polarization states of the signal and local-oscillator are orthogonal, the IF photocurrent will be zero. There are a number of possible solutions to this problem.

1. Polarization maintaining (PM) fiber can be used in the optical link to keep the signal polarization from changing. It is then a simple matter to adjust the local-oscillator polarization to achieve optimum mixing. However, the losses associated with PM fiber are greater than for conventional single-mode (SM) fiber. In addition most installed fiber is SM fiber.
2. An active polarization controller, such as a fiber coil using bending-induced birefringence, can be used to ensure that the local-oscillator polarization tracks the signal polarization.
3. A polarization scrambler can be used to scramble the polarization of the transmitted signal at a speed greater than the bit rate. The sensitivity degradation due to the polarization scrambling is 3 dB compared to a receiver with perfect polarization matching. However, this technique is only feasible at low bit rates.
4. The most general technique is to use a polarization-diversity receiver, as shown in Figure 16. In this scheme, the signal light and local-oscillator

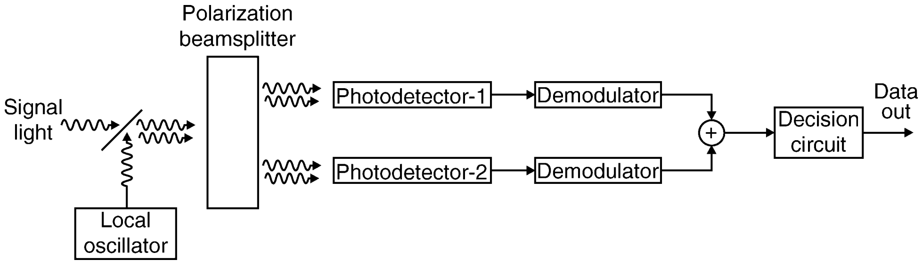

Figure 16 Polarization-diversity receiver.

Table 3 General comparisons between various modulation/demodulation schemes. The sensitivity penalty is relative to ideal homodyne PSK detection

Modulation/demodulation scheme	*Sensitivity penalty (dB)*	*Immunity to phase noise*	*IF bandwidth/bit rate*	*Complexity*
Homodyne PSK	0	Very poor	1	Requires an OPLL
Synchronous heterodyne PSK	3	Poor	3–6	Requires electronic phase locking
Asynchronous heterodyne FSK (dual-filter)	6.5	Excellent	3–6	High bandwidth IF circuits and two envelope detectors required
Differential detection CPFSK	3.5	Moderate	≈2	Relatively simple
DPSK	3.5	Moderate	≈2	Relatively simple

are split into orthogonally polarized lightwaves by a polarization beamsplitter. The orthogonal components of the signal and local-oscillator are separately demodulated and combined. The resulting decision signal is polarization independent. The demodulator used depends on the modulation format.

Comparisons Between the Principal Demodulation Schemes

General comparisons between the demodulation schemes discussed above are given in Table 3. FSK modulation is relatively simple to achieve. FSK demodulators have good receiver sensitivity and low immunity to phase noise. PSK modulation is also easy to achieve and gives good receiver sensitivity. ASK modulation has no real merit compared to FSK or PSK.

In general, homodyne schemes have the best theoretical sensitivity but require optical phase locking and are very sensitive to phase noise. Phase diversity receivers are a good alternative to homodyne receivers, in that they do not require an OPLL. Asynchronous heterodyne receivers have similar sensitivities to synchronous receivers but are much less complex. However, they require an IF bandwidth much greater than the data rate.

Differential detection schemes have good receiver sensitivity, do not require wide bandwidth IF circuits and have moderate immunity to phase noise. The advantage of CPFSK over DPSK is that the immunity to phase noise can be increased by increasing the FSK modulation index.

Table 4 Coherent systems experiments. The penalty is with respect to the theoretical sensitivity

Receiver	*Bit rate (Mb/s)*	*Sensitivity (dBm)*		*Penalty (dB)*
		dBm	*Photons/bit*	
ASK	400	−47.2	365	9.7
FSK (dual-filter)	680	−39.1	1600	16
CPFSK ($m = 0.5$)	1000	−37	1500	18.6
differential detection	4000	−31.3	1445	18.6
DPSK	400	−53.3	45.1	3.5
	2000	−39.0	480	13.8

Polarization diversity receivers can be used with any modulation/demodulation scheme and offer the most general approach to overcoming polarization effects.

System Experiments

Most coherent communication system experiments date from the 1980s and early 1990s and were based on the technology available at that time. This restricted the bit rates possible to typically <4 Gb/s. Some coherent systems experimental results obtained around this time are listed in Table 4. All of the listed experimental trials used lasers (usually external cavity lasers) with negligible phase noise in the 1550 nm region. It can be seen that the measured sensitivities deteriorate at higher bit rates. This is attributable to the nonuniform response and bandwidth limitations of the modulator, photodetector, and demodulator. There is no doubt that the performance could be improved by the use of current optoelectronic technologies.

Conclusion

Coherent optical communication systems offer the potential for increased sensitivity and selectivity compared to conventional IM/DD systems. However, coherent optical receivers are more complex and have more stringent design parameters compared to IM/DD receivers. Moreover, advances in optoelectronic component design and fabrication along with the advent of reliable optical amplifiers has meant that coherent systems are not a viable alternative to IM/DD systems in the vast majority of optical communication systems.

List of Units and Nomenclature

e	Electronic charge
h	Planck constant

See also

Modulators: Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Lightwave Transmitters; Wavelength Division Multiplexing. **Polarization:** Introduction.

Further Reading

Barry JR and Lee EA (1990) Performance of coherent optical receivers. *IEEE Proceedings* 78: 1369–1394.

Cvijetic M (1996) *Coherent and Nonlinear Lightwave Communications*. Boston, MA: Artech House.

Kazovsky LG (1989) Phase- and polarization-diversity coherent optical techniques. *Journal of Lightwave Technology* 7: 279–292.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.

Linke RA and Gnauck AH (1988) High-capacity coherent lightwave systems. *Journal of Lightwave Technology* 6: 1750–1769.

Ryu S (1995) *Coherent Lightwave Systems*. Boston, MA: Artech House.

COHERENT TRANSIENTS

Contents

Coherent Transient Spectroscopy in Atomic and Molecular Vapors
Foundations of Coherent Transients in Semiconductors
Ultrafast Studies of Semiconductors

Coherent Transient Spectroscopy in Atomic and Molecular Vapors

P R Berman, University of Michigan, Ann Arbor, MI, USA

R G Brewer, Stanford University, Stanford, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spectroscopy is a method that allows one to determine atomic and molecular transition frequencies, as well as relaxation rates in atomic and molecular vapors. The transition frequencies serve as ‘fingerprints’ that can be used to identify and classify atoms and molecules. Optical spectroscopic methods fall into two broad categories, continuous wave (cw) or stationary spectroscopy and time-dependent or transient spectroscopy. In cw spectroscopy, the absorption or emission line shapes are measured as a function of the frequency of a probe field. On the basis of theoretical line shape formulas, it is then possible to extract the relevant transition frequencies and relaxation rates from the line shapes. In transient optical spectroscopy, pulsed optical fields are used to create atomic state populations or coherences between atomic states. Following the excitation, the time-evolution of the atoms is monitored, from which transition frequencies and relaxation rates can be obtained.

The earliest coherent transient effects were observed with atomic nuclear spins in about 1950 by H C Torrey, who discovered transient spin nutation, and by E L Hahn who detected spin echoes and free induction decay (FID). The optical analogs of these and other nuclear magnetic resonance (NMR) effects have now been detected using either optical pulses or techniques involving Stark or laser frequency switching.

In this brief introduction to coherent optical transients, several illustrative examples of coherent transient phenomena are reviewed. The optical Bloch equations are derived and coupled to Maxwell's Equations. The Maxwell–Bloch formalism is used to analyze free precession decay, photon echoes, stimulated photon echoes, and optical Ramsey fringes. Experimental results are presented along with the relevant theoretical calculations. In addition, coherent transient phenomena involving quantized matter waves are discussed. Although the examples considered in this chapter are fairly simple, it should be appreciated that sophisticated coherent transient techniques are routinely applied as a probe of complex structures such as liquids and solids or as a probe of single atoms and molecules.

Optical Bloch Equations

Many of the important features of coherent optical transients can be illustrated by considering the interaction of a radiation field with a two-level atom. The lower state $|1\rangle$ has energy $-\hbar\omega_0/2$ and upper state $|2\rangle$ has energy $\hbar\omega_0/2$. For the moment, the atom is assumed to be fixed at $\mathbf{R} = 0$ and all relaxation processes are neglected. The incident electric field is

$$\mathbf{E}(\mathbf{R} = 0, t) = \frac{1}{2}\boldsymbol{\varepsilon}[E(t)e^{-i\omega t} + E(t)e^{i\omega t}] \qquad [1]$$

where $E(t)$ is the field amplitude, $\boldsymbol{\varepsilon}$ is the field polarization and ω is the carrier frequency. The time dependence of $E(t)$ allows one to consider pulses having arbitrary shape. A time-dependent phase for the field could have been included, allowing one to consider the effect of arbitrary frequency 'chirps,' but such effects are not included in the present discussion. It is convenient to expand the atomic state wave function in a field interaction representation as

$$|\psi(t)\rangle = c_1(t)e^{i\omega t/2}|1\rangle + c_2(t)e^{-i\omega t/2}|2\rangle \qquad [2]$$

The atom-field interaction potential is

$$V(\mathbf{R}, t) = -\boldsymbol{\mu}\cdot\mathbf{E}(\mathbf{R}, t) \qquad [3]$$

where $\boldsymbol{\mu}$ is a dipole moment operator. When eqn [2] is substituted into Schrödinger's equation and rapidly varying terms are neglected (rotating-wave approximation), one finds that the state amplitudes evolve according to

$$i\hbar\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}t} = \tilde{\mathbf{H}}\mathbf{c}; \qquad \tilde{\mathbf{H}} = (\hbar/2)\begin{pmatrix} -\delta_0 & \Omega_0(t) \\ \Omega_0(t) & \delta_0 \end{pmatrix} \qquad [4]$$

where $\mathbf{c}$ is a vector having components (c_1, c_2)

$$\delta_0 = \omega_0 - \omega \qquad [5]$$

is an atom-field detuning

$$\Omega_0(t) = -\frac{\mu E(t)}{\hbar} = -\frac{\mu}{\hbar}\sqrt{\frac{2S(t)}{\varepsilon_0 c}} \qquad [6]$$

is a Rabi frequency, $\mu = \langle 1|\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}|2\rangle = \langle 2|\boldsymbol{\mu}\cdot\boldsymbol{\varepsilon}|1\rangle$ is a dipole moment matrix element, ε_0 is the permittivity of free space, and $\mathbf{S}(t)$ is the time-averaged Poynting vector of the field. Equation [4] can be solved numerically for arbitrary pulse envelopes.

Expectation values of physical observables are conveniently expressed in terms of density matrix elements defined by

$$\rho_{ij} = c_i c_j^* \qquad [7]$$

which obey equations of motion

$$\begin{aligned} \dot{\rho}_{11} &= -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\ \dot{\rho}_{22} &= i\Omega_0(t)(\rho_{21} - \rho_{12})/2 \\ \dot{\rho}_{12} &= -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 + i\delta_0\rho_{12} \\ \dot{\rho}_{21} &= i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - i\delta_0\rho_{21} \end{aligned} \qquad [8]$$

An alternative set of equations in terms of real variables can be obtained if one defines new parameters

$$\begin{aligned} u &= \rho_{12} + \rho_{21}; & \rho_{12} &= (u + iv)/2 \\ v &= i(\rho_{21} - \rho_{12}); & \rho_{21} &= (u - iv)/2 \\ w &= \rho_{22} - \rho_{11}; & \rho_{22} &= (m + w)/2 \\ m &= \rho_{11} + \rho_{22}; & \rho_{11} &= (m - w)/2 \end{aligned} \qquad [9]$$

which obey

$$\begin{aligned} \dot{u} &= -\delta_0 v \\ \dot{v} &= \delta_0 u - \Omega_0(t)w \\ \dot{w} &= \Omega_0(t)v \\ \dot{m} &= 0 \end{aligned} \qquad [10]$$

The last of these equations reflects the fact that $\rho_{11} + \rho_{22} = 1$, while the first three can be rewritten as

$$\dot{\mathbf{U}} = \boldsymbol{\Omega}(t) \times \mathbf{U} \qquad [11]$$

where the *Bloch vector* **U** has components (u, v, w) and the *pseudofield vector* $\boldsymbol{\Omega}(t)$ has components $[\Omega_0(t), 0, \delta_0]$.

Equations [8] or [10] constitute the optical Bloch equations without decay. The vector **U** has unit magnitude and traces out a path on the Bloch sphere as it precesses about the pseudofield vector. The component w is the population difference of the two atomic states, while u and v are related to the quadrature components of the atomic polarization (see below).

Equations [8] and [10] can be generalized to include relaxation. A simplified relaxation scheme that has a wide range of applicability is one in which state 2 decays spontaneously to state 1 with rate γ_2, while the relaxation of the coherence ρ_{12} is characterized by a complex decay parameter Γ_{12}, resulting from phase-changing collisions with a background gas. When such processes are included in eqn [8], they are modified as

$$\dot{\rho}_{11} = -i\Omega_0(t)(\rho_{21} - \rho_{12})/2 + \gamma_2\rho_{22} \qquad [12a]$$

$$\dot{\rho}_{22} = i\Omega_0(t)(\rho_{21} - \rho_{12})/2 - \gamma_2\rho_{22} \qquad [12b]$$

$$\dot{\rho}_{12} = -i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{12} + i\delta\rho_{12} \qquad [12c]$$

$$\dot{\rho}_{21} = i\Omega_0(t)(\rho_{22} - \rho_{11})/2 - \gamma\rho_{21} - i\delta\rho_{21} \qquad [12d]$$

where $\gamma = \gamma_2/2 + \mathrm{Re}(\Gamma_{12})$ and the detuning $\delta = \delta_0 - \mathrm{Im}(\Gamma_{12})$ is modified to include the collisional shift. With the addition of decay, the length of the Bloch vector is no longer conserved. The quantity γ_2 is referred to as the longitudinal relaxation rate. Moreover, one usually refers to $T_1 = \gamma_2^{-1}$ as the longitudinal relaxation time and $T_2 = \gamma^{-1}$ as the transverse relaxation time. In the case of purely radiative broadening, $\gamma_2 = 2\gamma$ and $T_1 = T_2/2$.

Maxwell–Bloch Equations

As a result of atom-field interactions, it is assumed that a polarization is created in the medium of the form

$$\mathbf{P}(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[P(Z,t)e^{i[kZ-\omega t]} + P^*(Z,t)e^{-i[kZ-\omega t]}\right] \qquad [13]$$

which gives rise to a signal electric field of the form

$$\mathbf{E}_s(\mathbf{R}, t) = \frac{1}{2}\boldsymbol{\varepsilon}\left[E_s(Z,t)e^{i[kZ-\omega t]} + E_s^*(Z,t)e^{-i[kZ-\omega t]}\right] \qquad [14]$$

The z-axis has been chosen in the direction of **k**. It follows from Maxwell's equations that the quasi-steady-state field amplitude is related to the polarization by

$$\frac{\partial E_s(Z,t)}{\partial Z} = \frac{ik}{2\varepsilon_0}P(Z,t) \qquad [15]$$

To arrive at eqn [15], it is necessary to assume that transverse field variations can be neglected, that the *phase matching* condition $k = \omega/c$ is met, and that the complex field amplitudes $P(Z,t)$ and $E_s(Z,t)$ vary slowly in space compared with e^{ikZ} and slowly in time compared with $e^{i\omega t}$.

The polarization $\mathbf{P}(\mathbf{R}, t)$, defined as the average dipole moment per unit volume, is given by the expression

$$\mathbf{P}(\mathbf{R}, t) = \mathcal{N}\Big[\boldsymbol{\mu}_{21}\langle\rho_{12}(Z,t)\rangle e^{-i[kZ-\omega t]} + \boldsymbol{\mu}_{12}\langle\rho_{21}(Z,t)\rangle e^{i[kZ-\omega t]}\Big] \qquad [16]$$

where $\mathcal{N}$ is the atomic density and $\rho_{12}(Z,t)$ and $\rho_{21}(Z,t)$ are single-particle density matrix elements that depend on the atomic properties of the medium. As such, eqn [16] provides the link between Maxwell's equations and the optical Bloch equations. The $\langle\ \rangle$ brackets in eqn [16] indicate that there may be additional averages that must be carried out. For example, in a vapor, there is a distribution of atomic velocities that must be summed over, while, in a solid, there may be an inhomogeneous distribution of atomic frequencies owing to local strains in the media. By combining eqns [9], [13], [15] and [16], and using the fact that $\mu = \mu_{12}\cdot\varepsilon = \mu_{21}\cdot\varepsilon$, one arrives at

$$\begin{aligned}\frac{\partial E_s(Z,t)}{\partial Z} &= \frac{ik\mathcal{N}\mu}{\varepsilon_0}\langle\rho_{21}(Z,t)\rangle \\ &= \frac{ik\mathcal{N}\mu}{2\varepsilon_0}\langle u(Z,t) - iv(Z,t)\rangle \qquad [17]\end{aligned}$$

which, together with eqn [12], constitute the Maxwell–Bloch equations. From eqn [17], it is clearly seen that the coherence $\langle\rho_{21}(Z,t)\rangle$ drives the signal field. If $\langle\rho_{21}(Z,t)\rangle$ is independent of Z, the intensity of radiation exiting a sample of length L is proportional to

$$I(L,t) = \frac{k\mathcal{N}\mu L}{\varepsilon_0}^2|\langle\rho_{21}(t)\rangle|^2 \qquad [18]$$

Alternatively, the outgoing field can be heterodyned with a reference field to extract the field amplitude rather than the field intensity.

Coherent Transient Signals

The Maxwell–Bloch equations can be used to describe the evolution of coherent transient signals. In a typical experiment, one applies one or more ‘short’ radiation pulses to an atomic sample and monitors the radiation field emitted by the atoms following the pulses. A ‘short’ pulse, is one satisfying

$$|\delta|\tau, ku\tau, \gamma\tau, \gamma_2\tau \ll 1 \qquad [19]$$

where τ is the pulse duration. Conditions [19] allow one to neglect any effects of detuning (including Doppler shifts or other types of inhomogeneous broadening) or relaxation during the pulse’s action. The evolution of the atomic density matrix during each pulse is determined solely by the pulse area defined by

$$\theta = \int_{-\infty}^{\infty} \Omega_0(t)\mathrm{d}t \qquad [20]$$

In terms of density matrix elements in a standard interaction representation defined by

$$\begin{aligned} \rho_{12}^I &= \rho_{12}e^{-i[kZ+\delta t]}, & \rho_{21}^I &= \rho_{21}e^{i[kZ+\delta t]}, \\ \rho_{22}^I &= \rho_{22}, & \rho_{11}^I &= \rho_{11} \end{aligned}$$

the change in density matrix elements for an interaction occurring at time T_j is given by

$$\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^+ = \frac{1}{2}\begin{pmatrix} 1+\cos\theta_j & 1-\cos\theta_j & i\sin\theta_j e^{i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} \\ 1-\cos\theta_j & 1+\cos\theta_j & -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{-i\Phi(T_j)} \\ i\sin\theta_j e^{-i\Phi(T_j)} & -i\sin\theta_j e^{-i\Phi(T_j)} & 1+\cos\theta_j & (1-\cos\theta_j)e^{-2i\Phi(T_j)} \\ -i\sin\theta_j e^{i\Phi(T_j)} & i\sin\theta_j e^{i\Phi(T_j)} & (1-\cos\theta_j)e^{2i\Phi(T_j)} & 1+\cos\theta_j \end{pmatrix}\begin{pmatrix} \rho_{11}^I \\ \rho_{22}^I \\ \rho_{12}^I \\ \rho_{21}^I \end{pmatrix}^- \qquad [21]$$

where $\mp$ superscripts refer to times just before and after the pulse

$$\Phi(T_j) = \mathbf{k}\cdot\mathbf{R}(T_j) + \delta T_j$$

$$\mathbf{R}(T_j) = \mathbf{R} - \mathbf{v}(t - T_j)$$

θ_j is the pulse area, and $\mathbf{v}$ is the atomic velocity. Between pulses, or following the last pulse, the atoms evolve freely in the absence of any applied fields. The density matrix elements evolve as

$$\begin{aligned} \dot{\rho}_{11}^I &= \gamma_2\rho_{22}^I; & \dot{\rho}_{22}^I &= -\gamma_2\rho_{22}^I; \\ \dot{\rho}_{12}^I &= -\gamma\rho_{12}^I; & \dot{\rho}_{21}^I &= -\gamma\rho_{21}^I \end{aligned} \qquad [22]$$

A coherent transient signal is constructed by piecing together field interaction zones and free evolution periods. There are many different types of coherent transient signals. A few illustrative examples are given below.

Free Polarization Decay

Free polarization decay (FPD), also referred to as optical FID, is the optical analog of free induction decay (FID) in NMR. This transient follows preparation of the atomic sample by any of three methods: (1) an optical pulse resonant with the atomic transition; (2) a Stark pulse that switches the transition frequency of a molecule (atom) into resonance with an incident *cw* laser beam; or (3) an electronic pulse that switches the frequency of a *cw* laser into coincidence with the molecular (atomic) transition frequency. In methods (2) and (3), the FPD emission propagates coherently in the forward direction along with the *cw* laser beam, automatically generating a heterodyne beat signal at a detector that can be orders of magnitude larger than the FPD signal. The FPD field intensity itself exiting the sample is equal to

$$\begin{aligned} I(L,t) &= \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2 \\ &= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 \left|\int \mathrm{d}\mathbf{v}\, W_0(\mathbf{v}) e^{-[\gamma - i(\delta + \mathbf{k}\cdot\mathbf{v})]t}\right|^2 \end{aligned} \qquad [23]$$

For a Maxwellian velocity distribution

$$W_0(\mathbf{v}) = \frac{1}{(\pi u^2)^{3/2}} e^{-(v/u)^2} \qquad [24]$$

where u is the most probable atomic speed, one finds

$$I(L,t) = \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta\right)^2 e^{-2\gamma t} e^{-(kut)^2/2} \qquad [25]$$

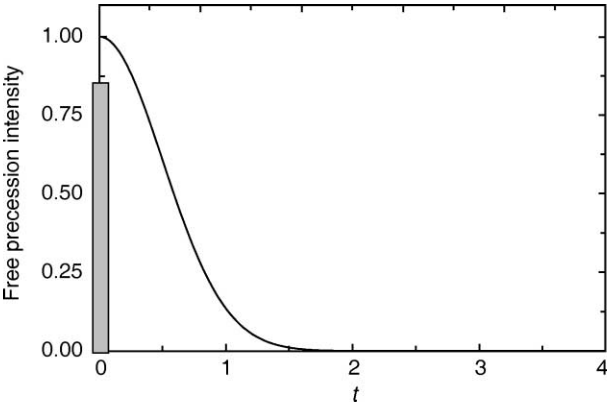

Figure 1 Pulse sequence and signal intensity for free precession decay. In this and subsequent figures, time is measured in units of $2/ku$.

If $ku \gg \gamma$, the signal decays mainly owing to inhomogeneous broadening, as shown in Figure 1. Owing to different Doppler shifts for atoms having different velocities, the optical dipoles created by the pulse lose their relative phase in a time of order $T_2^* = 2/ku$. The FPD signal can be used to measure T_2^*, which can be viewed as an inhomogeneous, transverse relaxation time. At room temperature ku/γ is typically of order of 100.

When FPD signals are observed using laser frequency switching, method 3 above, the *cw* field can modify the molecular velocity distribution, exciting only those atoms having $\mathbf{k} \cdot \mathbf{v} = -\delta \pm \gamma'$, where γ' is a power broadened homogeneous width. In the linear field regime, the FPD signal again decays on a time scale of order $(ku)^{-1}$, since the velocity distribution is unchanged to first order in the field amplitude. When one considers nonlinear interactions with the field, however, the *cw* field excites a narrow velocity subclass that is no longer subject to inhomogeneous broadening. These atoms give rise to a contribution to the FPD signal that decays with rate $T_2^{-1} = (\gamma + \gamma')$. Thus by using nonlinear atom-field interactions, one can extract homogeneous decay rates in situations where there is large inhomogeneous broadening. The price one pays is that only a small percentage of the atoms (those having velocities for which the applied field is resonant) contribute to the signal.

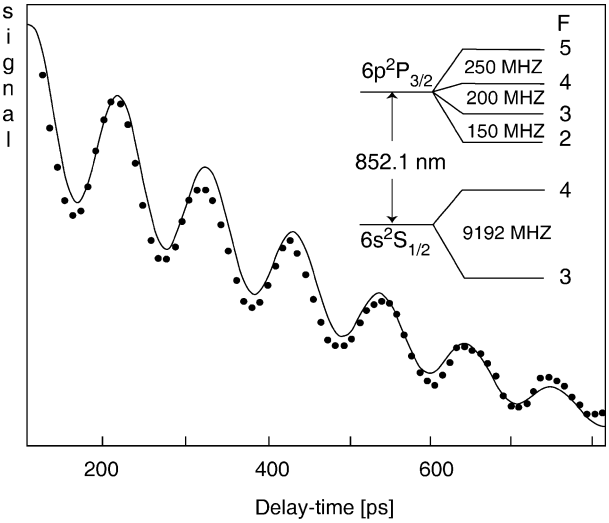

Figure 2 Optical free precession decay in Cs using an excitation pulse having 20 ns duration. Reproduced with permission from Lehmitz H and Harde H (1986) In: Prior Y, Ben-Reuven A and Rosenbluh M (eds) *Methods of Laser Spectroscopy*, pp. 109–112. New York: Plenum.

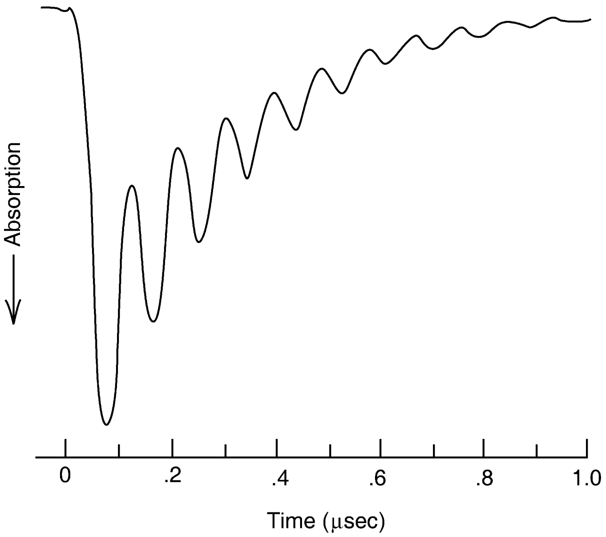

Figure 3 Optical free precession in NH_2D using the Stark switching technique. Reproduced with permission from Brewer RG and Shoemaker RL (1972) Optical free induction decay. *Physical Review A* 6: 2001. Copyright (1972) by the American Physical Society.

A FPD signal obtained on the D_2 transition in Cs is shown in Figure 2, where $T_2^* = 1.4$ ns. Oscillations in the signal originate from ground state hyperfine splitting. Figure 3 represents a FPD signal obtained in NH_2D at 10.6 μm using the method of Stark switching with *cw* state preparation. The oscillations are the heterodyne beat signal while the slowly varying increase in the signal is the result of optical nutation of molecules switched into resonance by the Stark pulse. The FPD signal manifests itself as a reduction of the amplitude of the oscillation with time. This amplitude decays with the (power-broadened) homogeneous decay rate γ'.

Photon Echo

Although the FPD signal produced by short excitation pulses decays in a time of order $(ku)^{-1}$, the coherence of individual atoms decays in a much longer time, T_2.

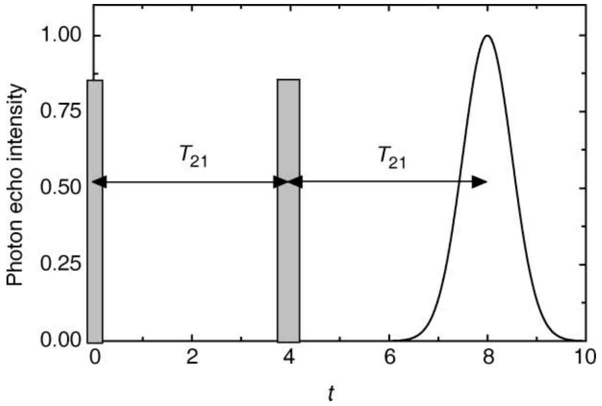

Figure 4 Pulse sequence and signal intensity for the photon echo.

The question arises as to whether it is possible to bring all the atomic dipoles back into phase so they can radiate a coherent signal. The photon echo accomplishes this goal, although it has very little to do with either photons or echoes. The pulse sequence and echo signal are shown in Figure 4. The first pulse creates a phased array of dipoles, which begin to dephase owing to inhomogeneous broadening. At time T_{21} the second pulse provides a nonlinear interaction that couples ρ_{12} and ρ_{21}. As a result of the second pulse, the dipoles begin a rephasing process that is completed at time $t = 2T_{21}$. The 'echo' appears at this time. The echo intensity exiting the sample is

$$\begin{aligned} I(L,t) &= \left(\frac{k\mathcal{N}\mu L}{\varepsilon_0}\right)^2 |\langle\rho_{21}(t)\rangle|^2 \\ &= \left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin^2(\theta_2/2)\right)^2 \\ &\quad \times e^{-4\gamma T_{21}} e^{-k^2u^2(t-2T_{21})^2/2} \end{aligned} \qquad [26]$$

If the echo intensity is measured as a function of T_{21}, one can extract the homogeneous decay rate $\gamma = T_{21}^{-1}$. It is interesting to note that the echo intensity near $t = 2T_{21}$ mirrors the FPD intensity immediately following the first pulse. For experimental reasons, it is often convenient to use a different propagation vector for the second pulse. If $\mathbf{k}_1$ and $\mathbf{k}_2$ are propagation vectors of the first and second pulses, the echo signal is in the direction $\mathbf{k} = 2\mathbf{k}_2 - \mathbf{k}_1$, provided $|\mathbf{k}_2 - \mathbf{k}_1|/k_1 \ll 1$.

The echo signal is sensitive only to $\rho_{21}(t)$ and $\rho_{12}(t)$ in the entire time interval of interest. Any interaction that results in a degradation of the dephasing–rephasing process for these density matrix elements leads to a decrease in the echo intensity. As such, echo signals can serve as a probe of transverse relaxation. Transverse relaxation generally falls into two broad categories. First, there are dephasing processes which produce an exponential damping of the coherences and contribute to γ. Second there is *spectral diffusion*; in a vapor, spectral diffusion is produced by velocity-changing collisions that change the effective field frequency seen by the atoms owing to the Doppler effect. Such terms enter the optical Bloch equations as *integral* terms, transforming the equations into differentio-integral equations. In general, the phase-changing and velocity-changing aspects of collisions are entangled; however, if the collisional interaction is *state-independent*, as it is for some molecular transitions, then collisions are purely velocity-changing in nature, leading to an echo that decays exponentially as T_{21}^3 for early times and T_{21} for later times. For electronic transitions, collisions are mainly phase-changing in nature, but there is a velocity-changing contribution that persists in the forward diffractive scattering cone. Photon echo techniques are sufficiently sensitive to measure the relaxation caused by such diffractive scattering.

The photon echo was first observed by NA Kurnit and co-workers, using a pulsed ruby laser (Figure 5). In Stark switching experiments, the photon echo follows two Stark pulses as shown in Figure 6. The effects of velocity-changing collisions on the photon echo signal in a CH_3F molecular vapor is represented in Figure 7. The signal exhibits exponential decay varying as T_{21}^3 for early times and T_{21} for later times. A theoretical interpretation of the observed echo decay permits a determination of the binary collision parameters for CH_3F, namely, the characteristic velocity jump $\Delta u_{\mathrm{rms}} = 200\ \mathrm{cm\,s^{-1}}$ and the cross section of $\sigma = 580\ \text{Å}^2$.

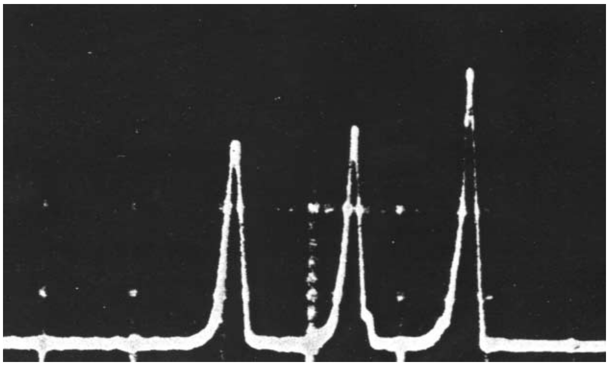

Figure 5 A photon echo signal from ruby. Time increases to the right with a scale of 100 ns/division. The pulse on the right is the echo signal, while the first two pulses are the (attenuated) input pulses. Reprinted with permission from Kurnit NA, Abella ID and Hartmann SR (1964) Observation of a photon echo. *Physical Review Letters* 13: 567. Copyright (1964) by the American Physical Society.

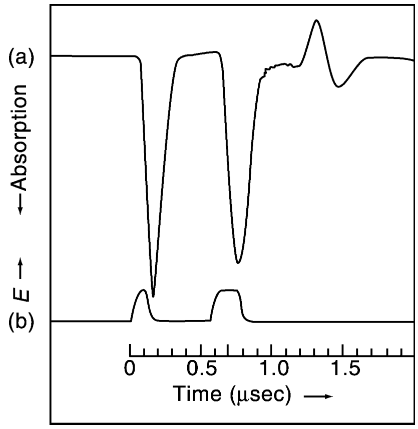

Figure 6 A photon echo signal in $^{13}CH_3F$ using the Stark switching method. The third pulse in trace (a) is the echo heterodyne beat signal. This signal is preceded by 'nutation' transient signals following each of the two Stark pulses shown in trace (b). Reprinted with permission from Brewer RG and Shoemaker RL (1971) Photo echo and optical nutation in molecules. *Physical Review Letters* 27: 631. Copyright (1971) by the American Physical Society.

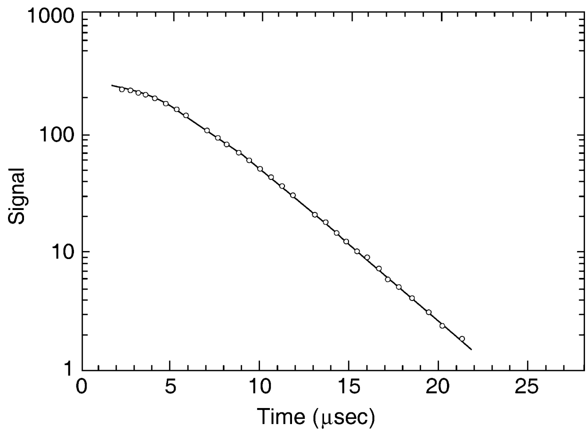

Figure 7 Decay of the photon echo signal in $^{13}CH_3F$ as a function of pulse separation. Reprinted with permission from Schmidt J, Berman PR and Brewer RG (1973) Coherent transient study of velocity-changing collisions. *Physical Review Letters* 31: 1103. Copyright (1973) by the American Physical Society.

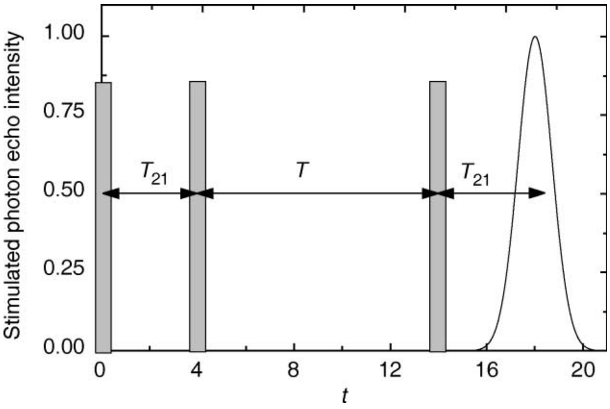

Figure 8 Pulse sequence and signal intensity for the stimulated photon echo.

Stimulated Photon Echo

In contrast to conventional photon echoes, *stimulated photon echoes* can be used to *simultaneously* measure both transverse and longitudinal relaxation times. Stimulated photon echoes have become an important diagnostic probe of relaxation in condensed matter systems as well as atomic vapors. The pulse sequence is shown in Figure 8 and consists of three pulses, having areas $\theta_1, \theta_2, \theta_3$, and equal propagation vectors $\mathbf{k}$. The time interval between the first two pulses is $T_{21} \gg T_2^*$ and pulse 3 occurs at time $t = T_{21} + T$. The maximum echo amplitude is given by

$$I(L, t_e) = \frac{1}{2}\left(\frac{k\mathcal{N}\mu L}{2\varepsilon_0}\sin\theta_1 \sin\theta_2 \sin\theta_3\right)^2 \times e^{-2\gamma_2 T} e^{-4\gamma T_{21}} \qquad [27]$$

and occurs when $t_e - (T_{21} + T) = T_{21}$. The optimal pulse sequence consists of three $\pi/2$ pulses.

If slightly different $\mathbf{k}$ vectors are chosen such that $\mathbf{k}_1 = \mathbf{k}_3 \approx \mathbf{k}_2$ the echo intensity varies as

$$e^{-2\gamma_2 T} e^{-4\gamma T_{21}} e^{-|\mathbf{k}_1 - \mathbf{k}_2|^2 u^2 (T + 2T_{21})^2/2}$$

By varying the angle between $\mathbf{k}_1$and $\mathbf{k}_2$, one can determine the Doppler width ku. By monitoring the echo signal as a function of $T_{21}(T)$, one obtains information on the transverse (longitudinal) relaxation.

The situation changes dramatically if there is an *additional* ground state level (level 0) to which the excited state can decay. In that case, the system is 'open' since population can leave the 'two-level' subsystem. The total population of the 1–2 state subsystem is no longer conserved, requiring an *additional* decay rate to account for relaxation. Let us suppose that all states decay with rate Γ_t as a result of their finite time in the laser beams. Moreover, let $\Gamma_{2,1}$ and $\Gamma_{2,0}$ be the decay rates of level 2 to levels 1 and 0, respectively. Assuming that $\gamma_2 T \gg 1$, one finds that the factor $e^{-2\gamma_2 T}$ in eqn [27] must be replaced by $(\Gamma_{2,0}/(\Gamma_{2,1} + \Gamma_{2,0}))^2 e^{-2\Gamma_t T}$. If $\Gamma_{2,0} \neq 0$, there is a long-lived component in the ground state population that contributes to the echo signal. One can exploit this feature of open systems to study spectral diffusion or velocity-changing collisions with very high sensitivity. Echoes can occur for time separations T much greater than the excited state lifetime. The creation of such long-lived stimulated photon echoes can be attributed to the fact that, in open systems, a portion of the velocity modulation created by the first two pulses in the ground state remains following spontaneous decay from the excited state.

The stimulated photon echo observed on the D_1 transition in sodium is shown in Figure 9. This is

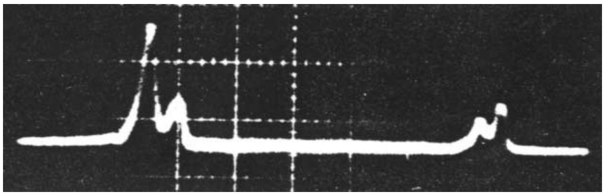

Figure 9 Stimulated photon echo observed on the D_1 transition in sodium. Reprinted with permission from Mossberg T, Flusberg A, Kachru R and Hartmann SR (1979) Total scattering cross section for Na on He measured by stimulated photon echoes. *Physical Review Letters* 42: 1665. Copyright (1979) by the American Physical Society.

an ‘open’ system, owing to ground state hyperfine structure. For the data shown, T is 17 times the 16 ns lifetime of the excited state. The first three pulses represent light scattered from the three input pulses and the fourth pulse is the echo. The echo appears at $t = T + 2T_{21}$.

Optical Ramsey Fringes

We have seen that coherent transients can be used to measure relaxation rates. With a slight modification of the stimulated photon echo geometry, it is also possible to use coherent transient signals to measure transition frequencies. If one chooses $\mathbf{k}_3$ in the $-\mathbf{k}_1$ direction instead of the $\mathbf{k}_1$ direction, it is possible to generate a phase matched signal in the

$$\mathbf{k} = \mathbf{k}_1 - \mathbf{k}_2 + \mathbf{k}_3 = -\mathbf{k}_2$$

direction. To simplify matters, it is assumed that $\mathbf{k}_1 \approx \mathbf{k}_2 = -\mathbf{k}$ and terms of order $|\mathbf{k}_1 - \mathbf{k}_2| u \times (T + 2T_{21})$ are neglected.

The averaged density matrix element in the vicinity of the echo is

$$\langle \rho_{21}(t) \rangle = (i/8) \sin \theta_1 \sin \theta_2 \sin \theta_3 e^{-\gamma_2 T} e^{-2\gamma T_{21}} \times e^{i\delta[(t-T_{21}-T)+T_{21}]} e^{-k^2 u^2 |(t-T_{21}-T)-T_{21}|^2/4} \qquad [28]$$

An echo is formed at time $t_e = T + 2T_{21}$, just as for the stimulated photon echo, but there is an additional phase factor, given by $e^{2i\delta T_{21}}$ when $t = t_e$, that is absent for the nearly collinear geometry. This phase factor is the optical analog of the phase factor that is responsible for the generation of Ramsey fringes in the microwave domain. The phase factor can be measured directly by heterodyning the signal field with a reference field or, indirectly, by converting the off-diagonal density matrix element into a population by the addition of a *fourth* pulse in the $\mathbf{k}_3$ direction at time $t = T + 2T_{21}$. In either case, the signal varies as $\cos(2\delta T_{21})$. The central fringe, corresponding to $\delta = 0$ can be isolated using one of two methods. If the experiment is carried out using an atomic beam rather than atoms in a cell, $T_{21} = L/u_0$, will be different for atoms having different u_0 (L = spatial separation of the first two pulses and u_0 is the longitudinal velocity of the atoms). On averaging over a distribution of u_0, the fringe having $\delta = 0$ will have the maximum amplitude. Experiments of this type allow one to measure optical frequencies with accuracy of order T_{21}^{-1}. For experiments using temporally-separated pulses acting on atoms in a cell, it is necessary to take data as a function of δ for several values of T_{21}, and then average the data over T_{21}. It might seem remarkable that narrow pulses having large bandwidths can be used to measure optical frequencies with arbitrary accuracy as T_{21} is increased. Of course, it is the *total* duration of the pulse sequence rather than the duration of a single pulse that is the relevant time parameter. For the optical coherence, T_{21} is the appropriate time parameter (during the interval T, it is population rather than coherence that contributes to the signal). The price one pays by using large T_{21} is a signal that decays as $e^{-4\gamma T_{21}}$.

Figure 10 illustrates an optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four-field zones were used. The most probable value of T_{21} was about 10^{-5} s for an effusive beam having a most probable longitudinal speed equal to $800\ \mathrm{ms}^{-1}$, giving a central fringe width of order 60 kHz.

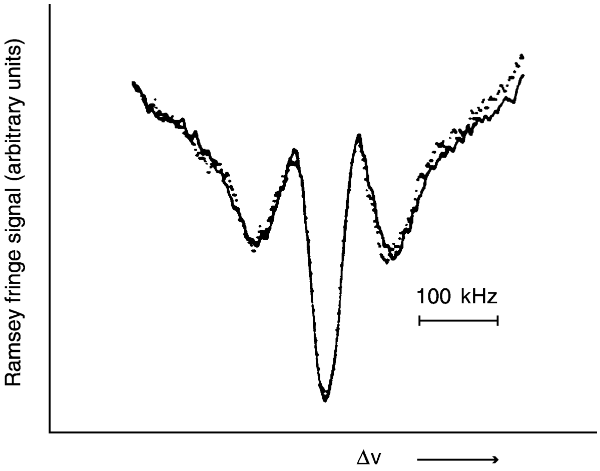

Figure 10 An optical Ramsey fringe signal on the 657 nm intercombination line in Ca. Four field zones were used. The solid and dotted lines represent runs with the directions of the laser field reversed, to investigate phase errors in the signals. Reprinted from Ito N, Ishikawa J and Moringa A (1994) Evaluation of the optical phase shift in a Ca Ramsey fringe stabilized optical frequency standard by means of laser-beam reversal. *Optics Communications* 109: 414, with permission from Elsevier.

Atom Interferometry

Matter wave atom interferometers rely on the wave nature of the center-of-mass motion for their operation. Such interferometers illustrate some interesting features of coherent optical transients not found in NMR. As an example, one can consider the response of a cold atomic ensemble to two, off-resonant standing-wave optical pulses separated in time by T. The electric field amplitude of pulse $j(j = 1, 2)$ is given by $\mathbf{E}_j(\mathbf{Z}, t) = \boldsymbol{\varepsilon} E_j(t)\cos(kZ)\cos(\omega t)$. The net effect of the field is to produce a spatially-modulated, ac Stark or light shift of the ground state energy. As a result, pulse j modifies the ground state amplitude by the phase factor $\exp[i\theta_j \cos(2kZ)]$. In other words, the standing wave field acts as a phase grating for the atoms. The phase factor can be interpreted in terms of momentum exchange between the two traveling wave components of the standing wave field. Owing to the fields' interaction with the atoms, all even integral multiples of $2\hbar k$ can be exchanged by the fields, imparting impulsive momenta of $2n\hbar k$ (n is a positive or negative integer) to the atoms. For an atom having momentum $\mathbf{P}$ and mass M, the frequency change associated with this momentum change is

$$\begin{aligned} E_{P,\mathbf{P}\pm 2n\hbar k\hat{Z}}/\hbar &= \left(\frac{P^2}{2M} - \frac{|\mathbf{P} \pm 2n\hbar k\hat{Z}|^2}{2M}\right)\Big/\hbar \\ &= \mp\frac{2nP_Z k}{M} - \omega_{2nk} \end{aligned} \qquad [29]$$

where

$$\omega_\kappa = \frac{\hbar\kappa^2}{2M}$$

There are two contributions to the frequency eqn [29]. The first part is independent of $\hbar$ and represents a classical Doppler shift, while the second part is proportional to $\hbar$ and represents a quantum, matter-wave effect. The quantum contribution will become important for times of order ω_{2nk}^{-1}. As with the echoes described previously, the Doppler effect results in a dephasing and rephasing of the optical response. However, in contrast to the previous echoes, the ground state density is no longer constant for times $t > \omega_{2nk}^{-1}$, when matter wave effects begin to play a role. At such times the phase grating created by the pulses can be transformed into amplitude gratings of the atomic density.

Long interaction times are possible, limited only by the time the atoms spend in the field. As such, atom interferometers offer exciting new possibilities for precision measurements of rotation rates, fundamental constants such as $\hbar$, gravitational acceleration, and gravitational gradients. Moreover, one can take advantage of the nonlinear atom-field interaction to create atomic density patterns having period $\lambda/2n$ from a standing wave optical field having wavelength λ. In this manner one has the possibility to construct atom interferometers and atom structures that operate on the nano-scale (spacings on the order of tens of nanometers).

Conclusion

Optical coherent transients are now used routinely as probes of atomic vapors, liquids, and condensed matter. Sophisticated techniques have been developed using fast and ultrafast pulses (having duration of 10 to 100 femtoseconds) to probe such systems and obtain the relevant relaxation rates. The manner in which optical coherent transients can be used to extract information on transition frequencies and relaxation rates in atomic and molecular vapors has been reviewed. In atomic vapors, optical coherent transients provide a means for obtaining both longitudinal and transverse relaxation rates. Moreover, echo experiments serve as an extremely sensitive probe of velocity changing collisions. Similar techniques are being rediscovered as an important tool in the analysis of cold atomic vapors and Bose–Einstein condensates. Only the most basic optical coherent transient phenomena have been described in this chapter. More elaborate, multi-pulse excitation schemes are joining established methods in attempts to understand the complex decay dynamics that one encounters in many branches of physics, chemistry and biology.

Acknowledgments

The research of PRB is supported by the National Science Foundation under Grants No. PHY-0098016 and the FOCUS Center Grant, and by the U. S. Army Research Office under Grant No. DAAD19-00-1-0412.

List of Units and Nomenclature

$\mathbf{E}(\mathbf{R}, t)$	Electric field vector
$E(t)$	Electric field amplitude
$E_s(Z, t)$	Complex signal electric field amplitude
$I(L, t)$	Signal intensity exiting the sample
$\mathbf{k}$	Field propagation vector
L	Sample length
$\mathcal{N}$	Atomic density
$\mathbf{P}$	Center-of-mass momentum
$\mathbf{P}(\mathbf{R}, t)$	Polarization vector
$P(Z, t)$	Complex polarization field amplitude
T_{21}, T	Time interval between pulses
u	Most probable atomic speed

(u, v, w)	Elements of Bloch vector
$\mathbf{U}(t)$	Bloch vector
$\mathbf{v}$	Atomic velocity
$W_0(\mathbf{v})$	Atomic velocity distribution
γ	Transverse relaxation rate
γ_2	Excited state decay rate or longitudinal relaxation rate
δ	Atom-field detuning
θ	Pulse area
μ	Dipole moment matrix element
ρ_{ij}^I	Density matrix element in 'normal' interaction representation
$\rho_{ij}(Z, t)$	Density matrix element in a field interaction representation
τ	Pulse durations
ω	Field frequency
ω_k	Recoil frequency
ω_0	Atomic transition frequency
$\Omega(t)$	Generalized Rabi frequency
$\boldsymbol{\Omega}(t)$	Pseudofield vector
$\Omega_0(t)$	Rabi frequency

See also

Interferometry: Gravity Wave Detection; Overview.

Further Reading

Allen L and Eberly J (1975) *Optical Resonance and Two-Level Atoms*, Chapters 2,4,9, New York: Wiley.

Berman PR and Steel DG (2001) *Coherent optical transients*. In: Bass M, Enoch JM, Van Stryland E and Wolf WL (eds) Handbook of Optics, vol. IV. chap. 24 New York: McGraw-Hill. (Material presented in this chapter is a condensed version of this reference.).

Brewer RG (1977) Coherent optical spectroscopy. In: Balian R, Haroche S and Liberman S (eds) *Frontiers in Laser Spectroscopy, Les Houches, Session XXVII*, vol. 1, pp. 341–396. Amsterdam: North Holland.

Brewer RG (1977) Coherent optical spectroscopy. In: Bloembergen N (ed.) *Nonlinear Spectroscopy*, Proc. Int. School Phys., Enrico Fermi, Course 64, Amsterdam: North-Holland.

Brewer RG (1977) *Coherent optical transients*. In *Physics Today*, vol. 30(5), 50–59.

Brewer RG and Hahn EL (1984) Atomic memory. In *Scientific American*, vol. 251(6), 50–57.

Journal of Optical Society America (1986) Special issue on coherent optical spectroscopy, Vol. A3(4).

Levenson MD and Kano SS (1988) *Introduction to Nonlinear Laser Spectroscopy*, Chapter 6. Boston: Academic Press.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*, 3rd edn. Chapter 12. Berlin: Springer-Verlag.

Shen YR (1984) *The Principles of Nonlinear Optics*, Chapter 17. New York: Wiley.

Shoemaker RL (1978) Coherent transient infrared spectroscopy. In: Steinfeld JI (ed.) *Laser and Coherence Spectroscopy*, pp. 197–371. New York: Plenum.

Zinth W and Kaiser W (1988) *Ultrafast Coherent Spectroscopy*. In: Kaiser W (ed.) *Ultrashort Laser Pulses and Applications*, vol. 60, pp. 235–277. Berlin: Springer-Verlag.

Foundations of Coherent Transients in Semiconductors

T Meier and S W Koch, Philipps-University, Marburg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Electromagnetic irradiation of matter generates a coherent superposition of the excited quantum states. The attribute 'coherent' relates to the fact that the material excitations originally have a well defined phase dependence which is imposed by the phase of the excitation source, often a laser in the optical regime. Macroscopically, the generated superposition state can be described as an optical polarization which is determined by the transition amplitudes between the participating quantum states.

The optical polarization is a typical non-equilibrium quantity that decays to zero when a system relaxes to its equilibrium state. Coherent effects are therefore only observable in a certain time window after pulsed photo-excitation, or in the presence of a continuous-wave (cw) beam. As coherent transients one usually refers to phenomena that can be observed during or shortly after pulsed laser excitation and critically depend on the presence of the induced optical polarization.

Many materials such as atoms, molecules, metals, insulators, semiconductors including bulk crystals, heterostructures, and surfaces, as well as organic and biological structures are studied using coherent optical spectroscopy. Depending on the particular system, the states participating in the optical transitions, the interactions among them, and the resulting time-scale for the decay of the induced polarization may be very different. As a result, the time window during which coherent effects are observable can be as long as seconds for certain atomic transitions, or it may be as short as a few femtoseconds (10^{-15}s), e.g., for metals, surfaces, or highly excited semiconductors.

Coherent spectroscopy and the analysis of coherent transients has provided valuable information on the nature and dynamics of the optical excitations. Often it is possible to learn about the interaction processes among the photoexcitations and to follow the temporal evolution of higher-order transitions which are only accessible if the system is in a non-equilibrium state.

The conceptually simplest experiment which one may use to observe coherent transients is to time resolve the transmission or reflection induced by a single laser pulse. However, much richer information can be obtained if one excites the system with several pulses. A pulse sequence with well-controlled delay times makes it possible to study the dynamical evolution of the photo-excitations, not only by time resolving the signal, but also by varying the delay. Prominent examples of such experiments are pump-probe measurements, which usually are performed with two incident pulses, or four-wave mixing, for which one may use two or three incident pulses.

Besides the microscopic interaction processes, the outcome of an experiment is determined by the quantum mechanical selection rules for the transitions and by the symmetries of the system under investigation. For example, if one wants to investigate coherent optical properties of surface states one often relies on phenomena, such as second-harmonic or sum-frequency generation, which give no signal in perfect systems with inversion symmetry. Due to the broken translational invariance, such experiments are therefore sensitive only to the dynamics of surface and/or interface excitations.

In this article the basic principles of coherent transients are presented and several examples are presented. The basic theoretical description and its generalization for the case of semiconductors are introduced.

Basic Principles

In the absence of free charges and currents, Maxwell's equations show that the electromagnetic field interacts with matter via the optical polarization. This polarization **P**, or more precisely its second derivative with respect to time $(\partial^2/\partial t^2)\mathbf{P}$, appears as a source term in the wave equation for the electric field **E**. Consequently, if the system is optically thin such that propagation effects within the sample can be ignored and if measurements are performed in the far field region, i.e., at distances exceeding the characteristic optical wavelength λ, the emitted electric field resulting from the polarization is proportional to its second time derivative, $\mathbf{E} \propto (\partial^2/\partial t^2)\mathbf{P}$. Thus the measurement of the emitted field dynamics yields information about the temporal evolution of the optical material polarization.

Microscopically the polarization is determined by the transition amplitudes between the different states of the system. These may be the discrete states of atoms or molecules, or the microscopic valence and conduction band states in a dielectric medium, such as a semiconductor. In any case, the macroscopic polarization **P** is computed by summing over all microscopic transitions p_{cv} via $\mathbf{P} = \sum_{c,v}(\boldsymbol{\mu}_{cv} p_{cv} + c.c.)$, where $\boldsymbol{\mu}_{cv}$ is the dipole matrix element which determines the strength of the transitions between the states v and c, and *c.c.* denotes the complex conjugate. If ε_c and ε_v are the energies of these states, their dynamic quantum mechanical evolution is described by the phase factors $e^{-i\varepsilon_c t/\hbar}$ and $e^{-i\varepsilon_v t/\hbar}$, respectively. Therefore, each p_{cv} is evolving in time according to $e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar}$. Assuming that we start at $t = 0$ with $p_{cv}(t = 0) = p_{cv,0}$, which may be induced by a short optical pulse, we have for the optical polarization $\mathbf{P}(t) = \sum_{c,v}(\boldsymbol{\mu}_{cv} p_{cv,0} e^{-i(\varepsilon_c - \varepsilon_v)t/\hbar} + c.c.)$. Thus $\mathbf{P}(t)$ is given by a summation over microscopic transitions which all oscillate with frequencies proportional to the energy differences between the involved states. Hence, the optical polarization is clearly a coherent quantity which is characterized by amplitude and phase. Furthermore, the microscopic contributions to $\mathbf{P}(t)$ add up coherently. Depending on the phase relationships, one may obtain either constructive superposition, interference phenomena like quantum beats, or destructive interference leading to a decay (dephasing) of the macroscopic polarization.

Optical Bloch Equations

The dynamics of photo-excited systems can be conveniently described by a set of equations, the optical Bloch equations, named after Felix Bloch (1905–1983) who first formulated such equations to analyze the spin dynamics in nuclear magnetic resonance. For the simple case of a two-level model, the Bloch equations can be written as

$$i\hbar \frac{\partial}{\partial t} p = \Delta\varepsilon p + \mathbf{E}\cdot\boldsymbol{\mu} I \qquad [1]$$

$$i\hbar \frac{\partial}{\partial t} I = 2\mathbf{E}\cdot\boldsymbol{\mu}(p - p^*) \qquad [2]$$

Here, $\Delta\varepsilon$ is the energy difference and I the inversion, i.e., the occupation difference between upper and lower state. The field **E** couples the polarization to the product of the Rabi energy $\mathbf{E}\cdot\boldsymbol{\mu}$

and the inversion I. In the absence of the driving field, i.e., $\mathbf{E} \equiv 0$, eqn [1] describes the free oscillation of p discussed above.

The inversion is determined by the combined action of the Rabi energy and the transition p. The total occupation N, i.e., the sum of the occupations of the lower and the upper states, remains unchanged during the optical excitation since light does not create or destroy electrons; it only transfers them between different states. If N is initially normalized to 1, then I can vary between -1 and 1. $I = -1$ corresponds to the ground state of the system, where only the lower state is occupied. In the opposite extreme, i.e., for $I = 1$, the occupation of the upper state is 1 and the lower state is completely depleted.

One can show that eqns [1]–[2] contain another conservation law. Introducing $\mathcal{P} = p + p^* = 2Re[p]$ and $\mathcal{J} = i(p - p^*) = -2Im[p]$, which are real quantities and often named polarization and polarization current, respectively, one finds that the relation

$$\mathcal{P}^2 + \mathcal{J}^2 + I^2 = 1 \qquad [3]$$

is fulfilled. Thus the coupled coherent dynamics of the transition and the inversion can be described on a unit sphere, the so-called Bloch sphere. The three terms $\mathcal{P}, \mathcal{J}$, and I can be used to define a three-dimensional Bloch vector $\mathbf{S} = (\mathcal{P}, \mathcal{J}, I)$. With these definitions, one can reformulate eqns [1]–[2] as

$$\frac{\partial}{\partial t}\mathbf{S} = \boldsymbol{\Omega} \times \mathbf{S} \qquad [4]$$

with $\boldsymbol{\Omega} = (-2\boldsymbol{\mu}\cdot\mathbf{E}, 0, \Delta\varepsilon)/\hbar$ and × denoting the vector product.

The structure of eqn [4] is mathematically identical to the equations describing either the angular momentum dynamics in the presence of a torque or the spin dynamics in a magnetic field. Therefore, the vector $\mathbf{S}$ is often called pseudospin. Moreover, many effects which can be observed in magnetic resonance experiments, e.g., free decay, quantum beats, echoes, etc. have their counterparts in photo-excited optical systems.

The vector product on the right-hand side of eqn [4] shows that $\mathbf{S}$ is changed by $\boldsymbol{\Omega}$ in a direction perpendicular to $\mathbf{S}$ and $\boldsymbol{\Omega}$. For vanishing field the torque $\boldsymbol{\Omega}$ has only a z-component. In this case the z-component of $\mathbf{S}$, i.e., the inversion, remains constant, whereas the x- and y-components, i.e., the polarization and the polarization current, oscillate with the frequency $\Delta\varepsilon/\hbar$ on the Bloch sphere around $\boldsymbol{\Omega}$.

Semiconductor Bloch Equations

If one wants to analyze optical excitations in crystallized solids, i.e., in systems which are characterized by a periodic arrangement of atoms, one has to include the continuous dispersion (bandstructure) into the description. In translationally invariant, i.e., ordered systems, this can be done by including the crystal momentum $\hbar\mathbf{k}$ as an index to the quantities which describe the optical excitation of the material, e.g., p and I in eqns [1]–[2]. Hence, one has to consider a separate two-level system for each crystal momentum $\hbar\mathbf{k}$. As long as all other interactions are disregarded, the bandstructure simply introduces a summation over the uncoupled contributions from the different $\mathbf{k}$ states, leading to inhomogeneous broadening due to the presence of a range of transition frequencies $\Delta\varepsilon(\mathbf{k})/\hbar$.

For any realistic description of optical processes in solids, it is essential to go beyond the simple picture of non-interacting states and to treat the interactions among the elementary material excitations, e.g., the Coulomb interaction between the electrons and the coupling to additional degrees of freedom, such as lattice vibrations (electron–phonon interaction) or other bath-like subsystems. If crystals are not ideally periodic, imperfections, which can often be described as a disorder potential, need to be considered as well.

All these effects can be treated by proper extensions of the optical Bloch equations introduced above. For semiconductors the resulting generalized equations are known as semiconductor Bloch equations, where the microscopic interactions are included at a certain level of approximation. For a two-band model of a semiconductor, these equations can be written schematically as

$$i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}} = \Delta\varepsilon_{\mathbf{k}}p_{\mathbf{k}} + \Omega_{\mathbf{k}}(n^c_{\mathbf{k}} - n^v_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}}|_{\text{corr}} \qquad [5]$$

$$i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}} = (\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^c_{\mathbf{k}}|_{\text{corr}} \qquad [6]$$

$$i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}} = -(\Omega^*_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^*_{\mathbf{k}}) + i\hbar\frac{\partial}{\partial t}n^v_{\mathbf{k}}|_{\text{corr}} \qquad [7]$$

Here $p_{\mathbf{k}}$ is the microscopic polarization and $n^c_{\mathbf{k}}$ and $n^v_{\mathbf{k}}$ are the electron populations in the conduction and valence bands (c and v), respectively. Due to the Coulomb interaction and possibly further processes, the transition energy $\Delta\varepsilon_{\mathbf{k}}$ and the Rabi energy $\Omega_{\mathbf{k}}$ both depend on the excitation state of the system, i.e., they are functions of the time-dependent polarizations and populations. This leads, in particular, to a coupling among the excitations for all different values of the crystals momentum $\hbar\mathbf{k}$. Consequently, in the presence of interactions, the optical excitations can no longer be described as independent two-level systems but have to be treated as a coupled many-body system. A prominent and important example in this context is the appearance of strong exciton resonances which, as a consequence of the Coulomb interaction, show up in the absorption spectra of semiconductors energetically below the fundamental bandgap.

The interaction effects lead to significant mathematical complications since they induce couplings between all the different quantum states in a system and introduce an infinite hierarchy of equations for the microscopic correlation functions. The terms given explicitly in eqns [5]–[7] arise in a treatment of the Coulomb interaction on the Hartree-Fock level. Whereas this level is sufficient to describe excitonic resonances, there are many additional effects, e.g., excitation-induced dephasing due to Coulomb scattering and significant contributions from higher-order correlations, like excitonic populations and biexcitonic resonances, which make it necessary to treat many-body correlation effects that are beyond the Hartree-Fock level. These contributions are formally included by the terms denoted as $|_{\mathrm{corr}}$ in eqns [5]–[7]. The systematic truncation of the many-body hierarchy and the analysis of controlled approximations is the basic problem in the microscopic theory of optical processes in condensed matter systems.

Examples

Radiative Decay

The emitted field originating from the non-equilibrium coherent polarization of a photo-excited system can be monitored by measuring the transmission and the reflection as function of time. If only a single isolated transition is excited, the dynamic evolution of the polarization and therefore of the transmission and reflection is governed by radiative decay. This decay is a consequence of the coupling of the optical transition to the light field described by the combination of Maxwell- and Bloch-equations. Radiative decay simply means that the optical polarization is converted into a light field on a characteristic time scale $2T_{\mathrm{rad}}$. Here, T_{rad} is the population decay time, often also denoted as T_1-time. This radiative decay is a fundamental process which limits the time on which coherent effects are observable for any photo-excited system. Due to other mechanisms, however, the non-equilibrium polarization often vanishes considerably faster.

The value of T_{rad} is determined by the dipole matrix element and the frequency of the transition, i.e., $T_{\mathrm{rad}}^{-1} \propto |\boldsymbol{\mu}|^2 \Delta\omega$, with $\Delta\omega = \Delta\varepsilon/\hbar$. The temporal evolution of the polarization and the emitted field are proportional to $e^{-i\Delta\omega t - t/(2T_{\mathrm{rad}})}$. Usually one measures the intensity of a field, i.e., its squared modulus, which evolves as $e^{-t/T_{\mathrm{rad}}}$ and thus simply shows an exponential decay, the so-called *free-induction decay* (see Figure 1). T_{rad} can be very long for transitions with small matrix elements. For semiconductor quantum wells, however, it is in the order of only 10 ps, as the result of the strong light-matter interaction.

The time constant on which the optical polarization decays is often called T_2. In the case where this decay is dominated by radiative processes, we thus have $T_2 = 2T_{\mathrm{rad}}$.

Superradiance

The phenomenon of superradiance can be discussed considering an ensemble of N two-level systems which are localized at certain positions $\mathbf{R}_i$. In this case

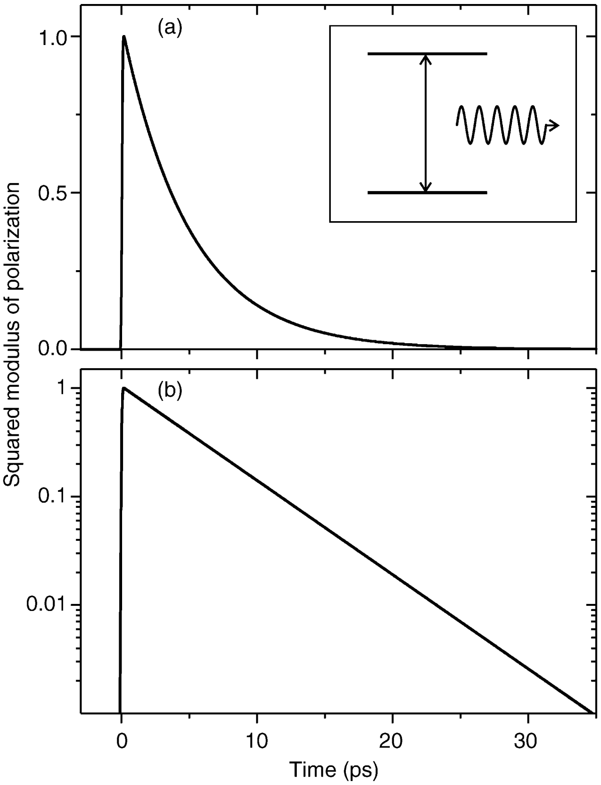

Figure 1 A two-level system is excited by a short optical pulse at $t = 0$. Due to radiative decay (inset) and possibly other dephasing mechanisms, the polarization decays exponentially as function of time with the time constant T_2, the dephasing time. The intensity of the optical field which is emitted as a result of this decay is proportional to the squared modulus of the polarization, which falls off with the time constant $T_2/2$. The decay of the squared modulus of the polarization is shown in (a) on a linear scale and in (b) on a logarithmic scale. The dephasing time was chosen to be $T_2 = 10$ ps which models the radiative decay of the exciton transition of a semiconductor quantum well.

Maxwell's equations introduce a coupling among all these resonances since the field emitted from any specific resonance interacts with all other resonances and interferes with their emitted fields. As a result, this system is characterized by N eigenmodes originating from the radiatively coupled optical resonances.

A very spectacular situation arises if one considers N identical two-level systems regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance, i.e., $\lambda = c/\Delta\omega$, where c is the speed of light in the considered material (see Figure 2a). In this case all emitted fields interfere constructively and the system behaves effectively like a single two-level system with a matrix element increased by $\sqrt{N}$. Consequently, the radiative decay rate is increased by N and the polarization of the coupled system decays N-times faster than that of an isolated system (see Figure 2b). This effect is called superradiance.

It is possible to observe superradiant coupling effects, e.g., in suitably designed semiconductor heterostructures. Figure 2c compares the measured time-resolved reflection from a single quantum well (dashed line) with that of a Bragg structure, i.e., a multiple quantum-well structure which consists of 10 individual wells that are separated by $\lambda/2$ (solid line), where λ is the wavelength of the exciton resonance. For times greater than about 2 ps the direct reflection of the exciting pulse has decayed sufficiently and one is left with the exponential decay of the remaining signal. Figure 2c shows that this decay is much more rapid for the Bragg structure than for the single quantum well, due to superradiance introduced by the radiative coupling among the quantum wells.

Destructive Interference

As the next example we now consider a distribution of two-level systems which have slightly different transition frequencies characterized by the distribution function $g(\Delta\omega - \bar{\omega})$ of the transition frequencies which is peaked at the mean value $\bar{\omega}$ and has a spectral width of $\delta\omega$ (see Figure 3a). Ignoring the radiative coupling among the resonances, the optical polarization of the total system evolves after excitation with a short laser pulse at $t = 0$ proportional to $\int d\omega\, g(\Delta\omega - \bar{\omega})e^{-i\Delta\omega t} \propto \tilde{g}(t)e^{-i\bar{\omega}t}$, where $\tilde{g}(t)$ denotes the Fourier transform of the frequency distribution function. $\tilde{g}(t)$ and thus the optical polarization decays on a time-scale which is inversely proportional to the spectral width of the distribution function $\delta\omega$. Thus the destructive interference of many different transition frequencies results in a rapid decay of the polarization (see Figure 3b).

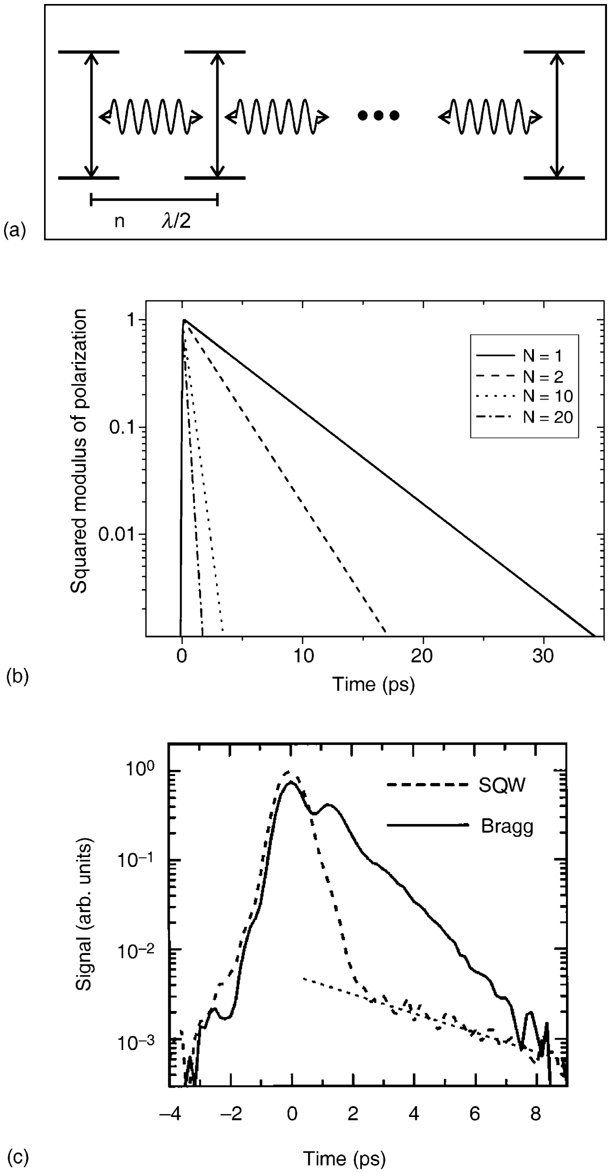

Figure 2 (a) N identical two-level systems are depicted which are regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance. Due to their coupling via Maxwell's equations all fields emitted from the two-level systems interfere constructively and the coupled system behaves effectively like a single two-level system with an optical matrix element increased by $\sqrt{N}$. (b) The temporal decay of the squared modulus of the polarization is shown on a logarithmic scale. The radiative decay rate is proportional to N and thus the polarization of the coupled system decays N-times faster than that of an isolated system. This effect is called superradiance. (c) Measured time-resolved reflection of a semiconductor single quantum well (dashed line) and a $N = 10$ Bragg structure, i.e., a multi quantum well where the individual wells are separated by $\lambda/2$ (solid line), on a logarithmic scale. [Part (c) is reproduced with permission from Haas S, Stroucken T, Hübner M, *et al.* (1998) Intensity dependence of superradiant emission from radiatively coupled excitons in multiple-quantum-well Bragg structures. *Physical Review B* 57: 14860. Copyright (1998) by the American Physical Society].

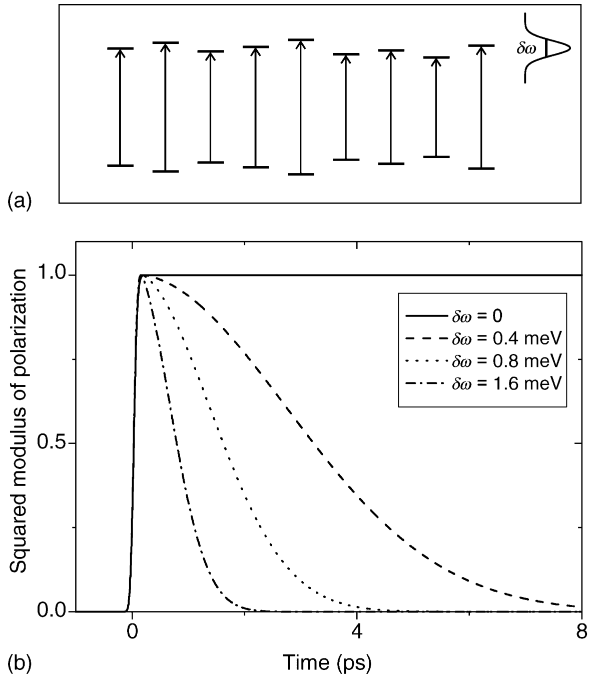

Figure 3 (a) An ensemble of two-level systems is shown where the resonance frequencies are randomly distributed according to a Gaussian distribution function of width $\delta\omega$ around an average value. (b) The temporal dynamics of the squared modulus of the polarization of ensembles of two-level systems after excitation with a short optical pulse at $t = 0$ is shown on a linear scale. Since the dephasing time is set to infinity, i.e., $T_2 \rightarrow \infty$, the polarization of an ensemble of identical two-level system ($\delta\omega = 0$) does not decay. However, for a finite width of the distribution, $\delta\omega > 0$, the individual polarizations of the ensemble oscillate with different frequencies and, therefore, due to destructive interference the polarization of the ensemble decays as function of time. Since the Fourier transform of a frequency-domain Gaussian is a Gaussian in the time domain, the dashed, dotted, and dashed dotted line have a Gaussian shape with a temporal width which is inversely proportional to $\delta\omega$.

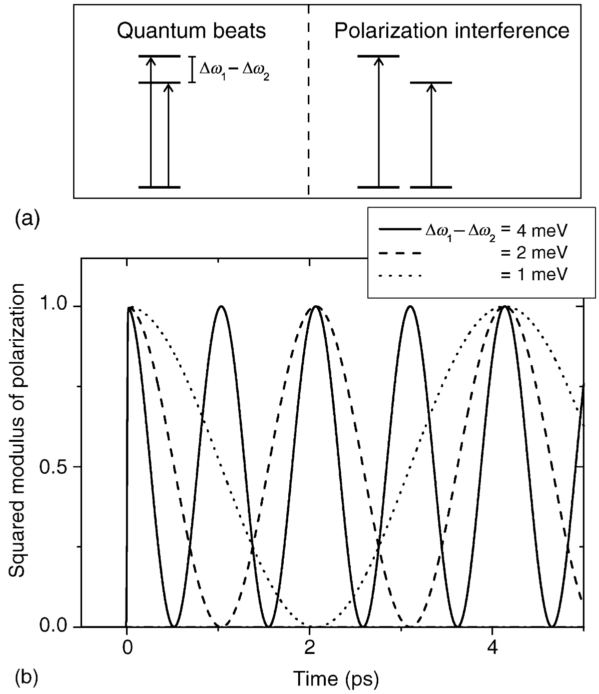

Figure 4 (a) Two optical resonances with frequency difference $\Delta\omega_1 - \Delta\omega_2$ may be realized by either a three-level system or by two uncoupled two-level systems. After impulsive excitation with an optical pulse the linear polarization of both types of systems show a modulation of the squared modulus of the polarization with the time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$. For the intrinsically coupled three-level system these modulations are called quantum beats, whereas for an interference of uncoupled systems they are named polarization interferences. Using nonlinear optical techniques, e.g., four-wave mixing and pump probe, it is possible to distinguish quantum beats and polarization interferences since coupled and uncoupled systems have different optical nonlinearities. (b) The temporal dynamics of the squared modulus of the polarization is shown for systems with two resonances and frequency difference $\Delta\omega_1 - \Delta\omega_2$, neglecting dephasing, i.e., setting $T_2 \rightarrow \infty$. After excitation with a short optical pulse at $t = 0$ the squared modulus of the polarization is periodically modulated with a time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$.

In the spectral domain this rapid decay shows up as inhomogeneous broadening. Depending on the system under investigation, there are many different sources for such an inhomogeneous broadening. One example is the Doppler broadening in atomic gases or disorder effects such as well-width fluctuations in semiconductor quantum wells or lattice imperfections in crystals.

In the nonlinear optical regime it is, under certain circumstances, possible to reverse the destructive interference of inhomogeneously broadened coherent polarizations. For example, in four-wave mixing, a second pulse may lead to a rephasing of the contributions with different frequencies, which results in the photon echo (see discussion below).

Quantum Beats

The occurrence of quantum beats can be understood most easily in a system where the total optical polarization can be attributed to a finite number of optical transitions. Let us assume for simplicity that all these transitions have the same matrix element. In this case, after excitation with a short laser pulse at $t = 0$ the optical polarization of the total system evolves proportional to $\sum_i e^{-i\Delta\omega_i t}$. The finite number of frequencies results in a temporal modulation with time periods $2\pi/(\Delta\omega_i - \Delta\omega_j)$ of the squared modulus of the polarization which is proportional to the emitted field intensity. For the case of two frequencies the squared modulus of the polarization is proportional to $[1 + \cos((\Delta\omega_1 - \Delta\omega_2)t)]$, i.e., due to the interference of two contributions the polarization varies between a maximum and zero (see Figure 4b).

In the linear optical regime it is impossible to distinguish whether the optical transitions are uncoupled or coupled. As shown in Figure 4, the two uncoupled two-level systems give the same linear polarization as a three-level systems where the two transitions share a common state. It is, however, possible to decide about the nature of the underlying transitions if one performs nonlinear optical spectroscopy. This is due to the fact that the so-called quantum beats, i.e., the temporal modulations of the polarization of an intrinsically coupled system, show a different temporal evolution as the so-called polarization interferences, i.e., the temporal modulations of the polarization of uncoupled systems. The former ones are also much more stable in the presence of inhomogeneous broadening than the latter ones.

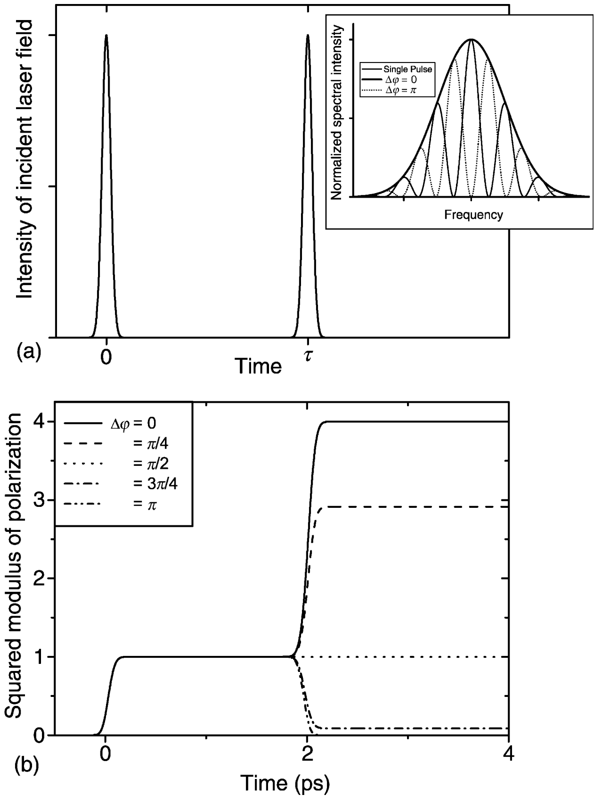

Figure 5 (a) Two Gaussian laser pulses separated by the time delay τ can lead to interferences. As shown by the thick solid line in the inset, the spectral intensity of a single pulse is a Gaussian with a maximum at the central frequency of the pulse. The width of this Gaussian is inversely proportional to the duration of the pulse. The spectral intensity of the field consisting of both pulses shows interference fringes, i.e., is modulated with a spectral period that is inversely proportional to τ. As shown by the thin solid and the dotted lines in the inset, the phase of the interference fringes depends on the phase difference $\Delta\varphi$ between the two pulses. Whereas for $\Delta\varphi = 0$ the spectral intensity has a maximum at the central frequency, it vanishes at this position for $\Delta\varphi = \pi$. (b) The temporal dynamics of the squared modulus of the polarization is shown for a two-level system excited by a pair of short laser pulses neglecting dephasing, i.e., setting $T_2 \to \infty$. The first pulse excites at $t = 0$ an optical polarization with a squared modulus normalized to 1. The second pulse, which has the same intensity as the first one, also excites an optical polarization at $t = 2$ ps. For $t > 2$ ps the total polarization is given by the sum of the polarizations induced by the two pulses. Due to interference, the squared modulus of the total polarization depends sensitively on the phase difference $\Delta\varphi$ between the two pulses. For constructive interference the squared modulus of the total polarization after the second pulse is four times bigger than that after the first pulse, whereas it vanishes for destructive interference. One may achieve all values in between these extremes by using phase differences $\Delta\varphi$ which are no multiples of π. It is thus shown that the second pulse can be used to coherently control the polarization induced by the first pulse.

In semiconductor heterostructures, quantum-beat spectroscopy has been widely used to investigate the temporal dynamics of excitonic resonances. Also the coupling among different optical resonances has been explored in pump-probe and four-wave-mixing measurements.

Coherent Control

In the coherent regime the polarization induced by a first laser pulse can be modified by a second, delayed pulse. For example, the first short laser pulse incident at $t = 0$ induces a polarization proportional to $e^{-i\Delta\omega t}$ and a second pulse arriving at $t = \tau > 0$ induces a polarization proportional to $e^{-i\varphi}e^{-i\Delta\omega(t-\tau)}$. Thus for $t \geq \tau$ the total polarization is given by $e^{-i\Delta\omega t}(1 + e^{-i\Delta\varphi})$ with $\Delta\varphi = \varphi - \Delta\omega\tau$. If $\Delta\varphi$ is an integer multiple of 2π, the polarizations of both pulses interfere constructively and the amplitude of the resulting polarization is doubled as compared to a single pulse. If, on the other hand, $\Delta\varphi$ is an odd multiple of π, the polarizations of both pulses interfere destructively and the resulting polarization vanishes after the action of both pulses. Thus by

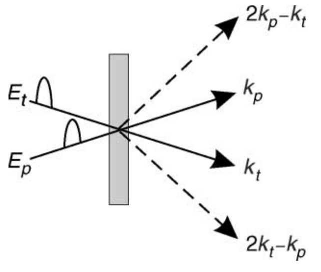

Figure 6 Transient optical nonlinearities can be investigated by exciting the system with two time delayed optical pulses. The pump pulse (E_p) puts the system in an excited state and the test (probe) pulse (E_t) is used to measure its dynamics. The dynamic nonlinear optical response may be measured in the transmitted or reflected directions of the pump and test pulses, i.e., $\pm k_p$ and $\pm k_t$, respectively. In a pump-probe experiment one often investigates the change of the test absorption induced by the pump pulse, by measuring the absorption in the direction k_t with and without E_p. One may also measure the dynamic nonlinear optical response in scattering directions like $2k_p - k_t$ and $2k_t - k_p$ as indicated by the dashed lines. This is what is done in a four-wave-mixing or photon-echo experiment.

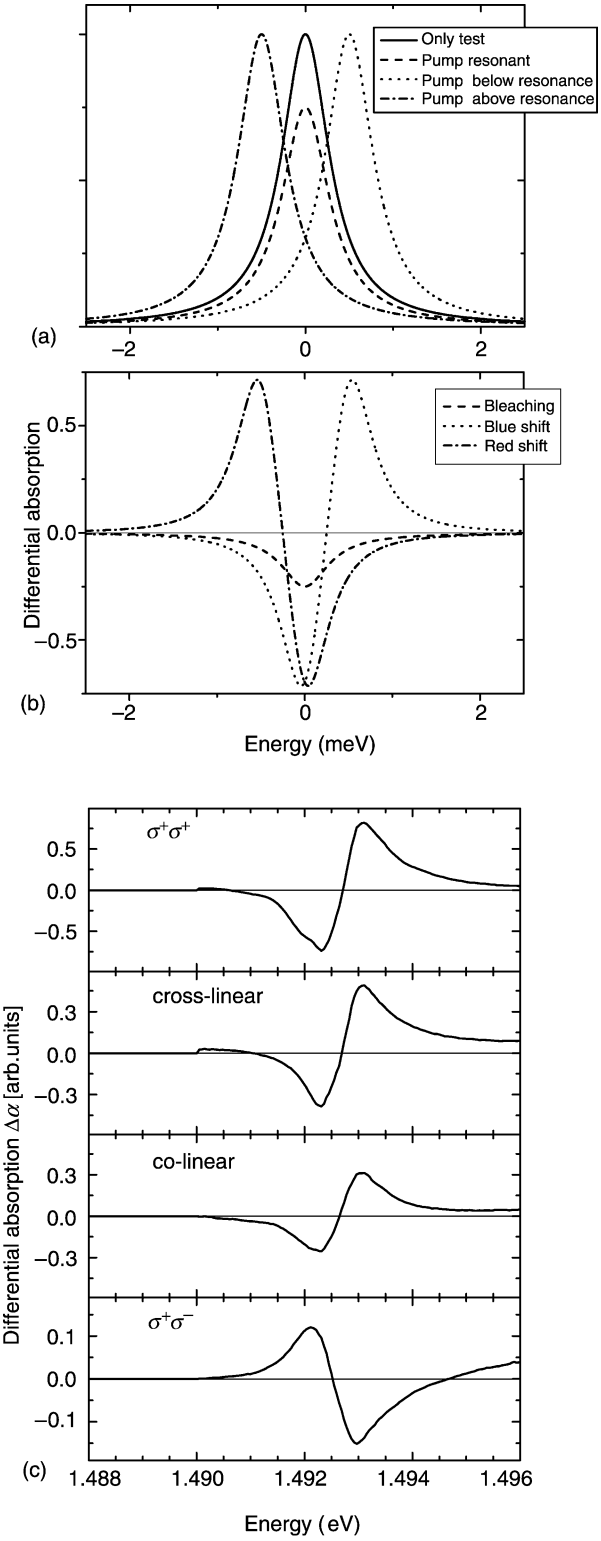

Figure 7 (a) Absorption spectra of a two-level system. The solid line shows the linear optical absorption spectrum as measured by a weak test pulse. It corresponds to a Lorentzian line that is centered at the transition frequency $\Delta\omega$, which has been set to zero. The width of the line is inversely proportional to the dephasing time T_2. In the presence of a pump pulse which resonantly excited the two-level system, the absorption monitored by the test pulse is reduced in amplitude, i.e., bleached, since the pump puts the system in an excited state (dashed line). In the optical Stark effect the frequency of the pump is nonresonant with the transition frequency. If the pump is tuned below (above) the transition frequency, the absorption is shifted to higher (lower) frequencies, i.e., shifted to the blue (red) part of the spectrum, see dotted (dashed-dotted) line. (b) The differential absorption obtained by taking the difference between the absorption in the presence of the pump and the absorption without pump. The dashed line shows the purely negative bleaching obtained using a resonant pump. The dotted and the dashed-dotted lines correspond to the dispersive shape of the blue and red shift obtained when pumping below and above the resonance, respectively. (c) Measured differential absorption spectra of a semiconductor quantum well which is pumped off resonantly 4.5 *meV* below the 1s heavy-hole exciton resonance. The four lines correspond to different polarization directions of the pump and probe pulses as indicated. [Part (c) is reproduced with permission from Sieh C, Meier T, Jahnke F, *et al.* (1999) Coulomb memory signatures in the excitonic optical Stark effect. *Physical Review Letters* 82: 3112. Copyright (1999) by the American Physical Society.]

varying the phase difference of the two pulses it is possible to coherently enhance or destroy the optical polarization (see Figure 5b).

One can easily understand the coherent control in the frequency domain by considering the overlap between the absorption of the system and the pulse spectrum. For excitation with two pulses that are temporally delayed by τ, the spectrum of the excitation shows interference fringes with a spectral oscillation period that is inversely proportional to τ. The positions of the maxima and the minima of the fringes depend on the phase difference between the two pulses. For constructive interference, the excitation spectrum is at a maximum at the resonance of the system, whereas for destructive interference the excitation spectrum vanishes at the resonance of the system (see Figure 5a).

Coherent control techniques have been applied to molecules and solids to control the dynamical evolution of electronic wavepackets and also the coupling to nuclear degrees of freedom. In this context, it is sometimes possible to steer certain chemical reactions into a preferred direction by using sequences of laser pulses which can be chirped, i.e., have a time-dependent frequency. Furthermore, in semiconductors and heterostructures, coherent control has been used to optically inject electronic currents on an ultrafast time-scale.

Transient Absorption Changes

In a typical pump-probe experiment one excites the system with a pump pulse (E_p) and probes its dynamics with a weak test pulse (E_t) (see Figure 6). With such experiments one often measures the differential absorption $\Delta\alpha(\omega)$ which is defined as the difference between the probe absorption in the presence of the pump $\alpha_{\text{pump on}}(\omega)$ and the probe absorption without the pump $\alpha_{\text{pump off}}(\omega)$.

For resonant pumping and for a situation where the pump precedes the test (positive time delays $\tau > 0$), the absorption change is usually negative in the vicinity of the resonance frequency $\Delta\omega$, indicating the effect of absorption bleaching (see Figure 7a,b). There may be positive contributions spectrally around the original absorption line due to resonance broadening and, at other spectral positions, due to excited state absorption, i.e., optical transitions to energetically higher states which are only possible if the system is in an excited state. The bleaching and the positive contributions are generally present in coherent, and also in incoherent situations, where the polarization vanishes but occupations in excited states are present.

For detuned pumping, the resonance may be shifted by the light field as for example, in the optical Stark effect. Depending on the excitation configuration and the system, this transient shift may be to higher (blue shift) or lower energies (red shift) (see Figure 7a,b). With increasing pump-probe time delay the system gradually returns to its unexcited state and the absorption changes disappear.

As an illustration we show in Figure 7c experimentally measured differential absorption spectra of a semiconductor quantum well which is pumped spectrally below the exciton resonance. For a two-level system one would expect a blue shift of the absorption, i.e., a dispersive shape of the differential absorption with positive contributions above and negative contributions below the resonance frequency. This is indeed observed for most polarization directions of the pump and probe pulses. However, if both pulses are oppositely circularly polarized, the experiment shows a resonance shift in the reverse direction, i.e., a red shift. For the explanation of this behavior one has to consider the optical selection rules of the quantum well system. One finds that the signal should actually vanish for oppositely circularly

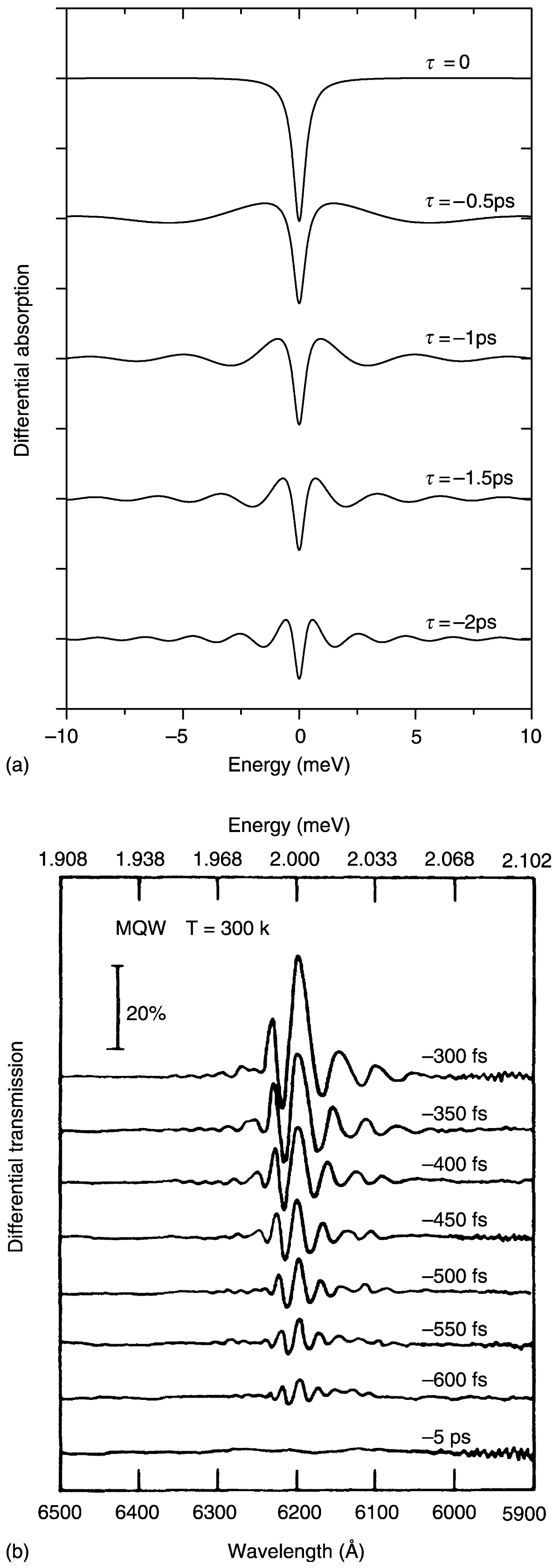

Figure 8 (a) Differential absorption spectra are shown for resonant pumping of a two-level system for various time delays τ between the pump and the test pulse. When the pump precedes the test, i.e., $\tau < 0$, the differential absorption exhibits spectral oscillations with a period which is inversely proportional to the time delay. When τ approaches zero, these spectral oscillations vanish and the differential absorption develops into purely negative bleaching. (b) Measured differential transmission spectra of a multi quantum well for different negative time delays as indicated. [Part (b) is reproduced with permission from Sokoloff JK, Joffre M, Fluegel B, *et al.* (1988) Transient oscillations in the vicinity of excitons and in the band of semiconductors. *Physics Review B* 38: 7615. Copyright (1988) by the American Physical Society.]

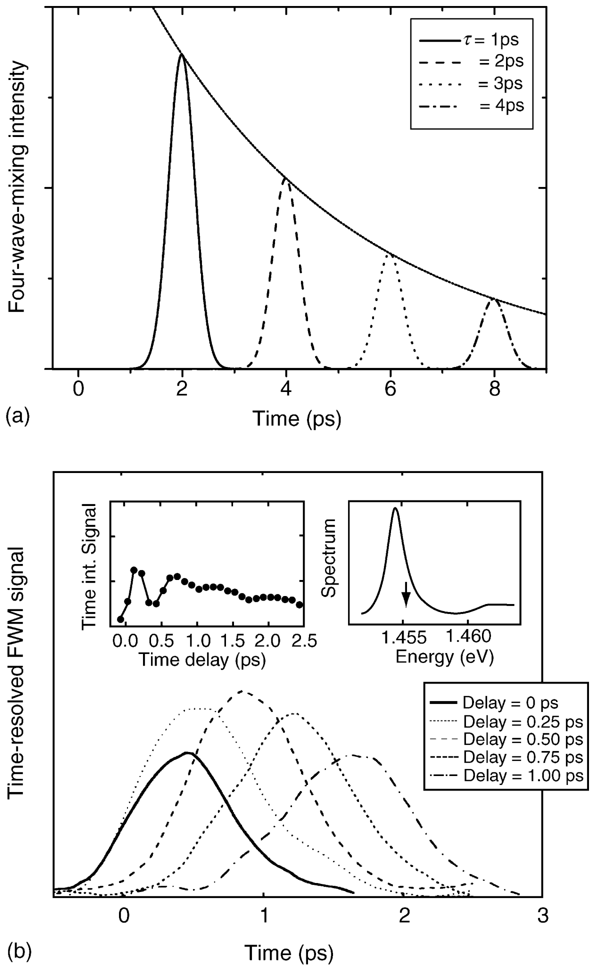

Figure 9 (a) The intensity dynamics as measured in a four-wave-mixing experiment on an ensemble of inhomogeneously broadened two-level systems for various time delays τ of the two incident pulses. The first pulse excites the system at $t = 0$. In the time period $0 < t < \tau$ the individual polarizations of the inhomogeneously broadened system oscillate with their respective frequencies and due to destructive interference the total polarization decays, cp. **Figure 3**. The second pulse arriving at $t = \tau$ leads to a partial rephasing of the individual polarizations. Due to phase conjugation all individual polarizations are in phase at $t = 2\tau$ which results in a macroscopic measurable signal, the photon echo. Due to dephasing processes the magnitude of the photon echo decays with increasing τ, see solid, dashed, dotted and dashed-dotted lines. Describing the dephasing via a T_2 time corresponds to an exponential decay as indicated by the thin dotted line. By measuring the decay of the echo with increasing τ, one thus gets experimental information on the dephasing of the optical polarization. (b) Time-resolved four-wave-mixing signal measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays as indicated. The origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. The insets show the corresponding time-integrated four-wave-mixing signal and the linear absorption. [Part (b) is reproduced with permission from Jahnke F, Koch M, Meier T, *et al.* (1994) Simultaneous influence of disorder and Coulomb interaction on photon echoes in semiconductors. *Physics Review B* 50: 8114. Copyright (1994) by the American Physical Society.]

polarized pulses, as long as many-body correlations are neglected. Thus in this case the entire signal is due to many-body correlations and it is their dynamics which gives rise to the appearance of the red shift.

Spectral Oscillations

For negative time delays $\tau < 0$, i.e., if the test pulse precedes the pump, the absorption change $\Delta\alpha(\omega)$ is characterized by spectral oscillations around $\Delta\omega$ which vanish as τ approaches zero (see Figure 8a). The spectral period of the oscillations decreases with increasing $|\tau|$. Figure 8b shows measured differential transmission spectra of a multiple quantum-well structure for different negative time delays. As with the differential absorption, also the differential transmission spectra are dominated by spectral oscillations around the exciton resonance whose period and amplitude decrease with increasing $|\tau|$.

The physical origin of the coherent oscillations is the pump-induced perturbation of the free-induction decay of the polarization generated by the probe pulse. This perturbation is delayed by the time τ at which the pump arrives. The temporal shift of the induced polarization changes in the time domain leads to oscillations in the spectral domain, since the Fourier transformation translates a delay in one domain into a phase factor in the conjugate domain.

Photon Echo

In the nonlinear optical regime one may (partially) reverse the destructive interference of a coherent, inhomogeneously broadened polarization. For example, in four-wave mixing, which is often performed with two incident pulses, one measures the emitted field in a background-free scattering direction (see Figure 5). The first short laser pulse excites all transitions at $t = 0$. As a result of the inhomogeneous broadening the polarization decays due to destructive interference (see Figure 3). The second pulse arriving at $t = \tau > 0$ is able to conjugate the phases ($e^{i\varphi} \rightarrow e^{-i\varphi}$) of the individual polarizations of the inhomogeneously broadened system. The subsequent unperturbed dynamical evolution of the polarizations leads to a measurable macroscopic signal at $t = 2\tau$. This photon echo occurs since at this point in time all individual polarizations are in phase and add up constructively (see Figure 9a). Since this rephasing process leading to the echo is only possible as long as the individual polarizations remain coherent, one can analyze the loss of coherence (dephasing) by measuring the decay of the photon echo with increasing time delay.

Echo experiments are frequently performed using more complicated geometries, e.g., more than two laser pulses. Further echo phenomena can also be observed as the result of the coupling of the photoexcitation to nuclear degrees of freedom and also due to spatial motion of the electrons, i.e., electronic currents.

Time-resolved four-wave-mixing signal, measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays, is presented in Figure 9b. Compared to Figure 9a, the origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t = \tau$. Due to the inhomogeneous broadening the maximum of the signal is shifting to longer times with increasing delay. Further details of the experimental results, in particular, the exact position of the maximum and the width of the signal, cannot be explained on the basis of non-interacting two-level systems, but require the analysis of many-body effects in the presence of inhomogeneous broadening.

See also

Coherent Control: Applications in Semiconductors; Experimental; Theory. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley.

Baumert T, Helbing J, Gerber G, *et al.* (1997) Coherent control with femtosecond laser pulses. *Advanced Chemical Physics* 101: 47–82.

Bergmann K, Theuer H and Shore BW (1998) Coherent population transfer among quantum states of atoms and molecules. *Review of Modern Physics* 70: 1003–1025.

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Brewer RG (1977) *Coherent Optical Spectroscopy*. In: Balian, *et al.* (ed.). Course 4 in Les Houches Session XXVII, *Frontiers in Laser Spectroscopy*, vol. 1. Amsterdam: North-Holland Publ.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) *Photons and Atoms*. New York: Wiley.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific.

Koch SW, Peyghambarian N and Lindberg M (1988) Transient and steady-state optical nonlinearities in semiconductors. *Journal of Physics* C 21: 5229–5250.

Macomber JD (1976) *The Dynamics of Spectroscopic Transitions*. New York: Wiley.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford.

Peyghambarian N, Koch SW and Mysyrowicz A (1993) *Introduction to Semiconductor Optics*. Englewood Cliffs, NJ: Prentice-Hall.

Pötz W and Schroeder WA (eds) (1999) *Coherent Control in Atoms, Molecules, and Semiconductors*. Dordrecht: Kluwer.

Schäfer W and Wegener M (2002) *Semiconductor Optics and Transport Phenomena*. Berlin: Springer.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*, 2nd edn. Berlin: Springer.

Yeazell JA and Uzer T (eds) (2000) *The Physics and Chemistry of Wave Packets*. New York: Wiley.

Ultrafast Studies of Semiconductors

J Shah, Arlington, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear optical spectroscopy, using the techniques of absorption, transmission, reflection and light scattering, has provided invaluable information about the electronic and vibrational properties of atoms, molecules, and solids. Optical techniques also possess some additional unique strengths: the ability to

1. generate nonequilibrium distributions functions of electrons, holes, excitons, phonons, etc. in solids;
2. determine the distribution functions by optical spectroscopy;
3. determine the nonequilibrium distribution functions on femtosecond timescales;
4. determine femtosecond dynamics of carrier and exciton transport and tunneling; and
5. investigate interactions between various elementary excitations as well as many-body processes in semiconductors.

Researchers have exploited these unique strengths to gain fundamental new insights into nonequilibrium, nonlinear, and transport physics of semiconductors and their nanostructures over the past four decades.

A major focus of these efforts has been devoted to understanding how a semiconductor in thermodynamic equilibrium, excited by an ultrashort optical pulse, returns to the state of thermodynamic equilibrium. Four distinct regimes can be identified:

(a) the coherent regime in which a well-defined phase relationship exists between the excitation created by the optical pulse and the electromagnetic (optical) field creating it;
(b) the nonthermal regime in which the coherence (well-defined phase relationships) has been destroyed by various collision and interference processes but the distribution function of excitations is nonthermal (i.e., cannot be described by a Maxwell–Boltzmann distribution or its quantum (degenerate) counterparts);
(c) the hot carrier regime in which the distributions for various excitations are thermal, but with different characteristic temperatures for different excitations and thermal bath; and
(d) the isothermal regime in which the excitations and the thermal bath are at the same temperature but there is an excess of excitation (e.g., electron–hole pairs) compared to thermodynamic equilibrium.

Various physical processes take the semiconductor from one regime to the next, and provide information not only about the fundamental physics of semiconductors but also about the physics and ultimate performance limits of electronic, optoelectronic, and photonic devices.

In addition to coherent and relaxation dynamics, ultrafast studies of semiconductors provide new insights into tunneling and transport of carriers, and demonstrate novel quantum mechanical phenomena and coherent control in semiconductors.

The techniques used for such studies have also advanced considerably over the last four decades. Ultrafast lasers with pulse widths corresponding to only a few optical cycles (1 optical cycle ≈ 2.7 fs at 800 nm) have been developed. The response of the semiconductor to an ultrafast pulse, or a multiple of phase-locked pulses, can be investigated by measuring the dynamics of light emitted by the semiconductor using a streak camera (sub-ps time resolution) or a nonlinear technique such as luminescence up-conversion (time resolution determined by the laser pulse width for an appropriate nonlinear crystal). The response of the semiconductor can also be investigated using a number of two or three beam pump-probe techniques such as transmission, reflection, light scattering or four-wave mixing (FWM). Both the amplitude and the phase of the emitted radiation or the probe can be measured by phase-sensitive techniques such as spectral interferometry. The lateral transport of excitation can be investigated by using time-resolved spatial imaging and the vertical transport and tunneling of carriers can be investigated using the technique of 'optical markers'. Electro-optic sampling can be used for transmission/reflection studies or for measuring THz response of semiconductors. More details on these techniques, and indeed many topics discussed in this brief article, can be found in the Further Reading.

Coherent Dynamics

The coherent response of atoms and molecules is generally analyzed for an ensemble of independent (noninteracting) two-level systems. The statistical properties of the ensemble are described in terms of the density matrix operator whose diagonal components relate to population of the eigenstates and off-diagonal components relate to coherence of the superposition state. The time evolution of the density matrix is governed by the Liouville variant of the Schrödinger equation $i\hbar\dot{\rho} = [H, \rho]$ where the system Hamiltonian $H = H_0 + H_{\text{int}} + H_R$ is the sum of the unperturbed, interaction (between the radiation field and the two-level system), and relaxation Hamiltonians, respectively. Using a number of different assumptions and approximations, this equation of motion is transformed into the optical Bloch equations (OBE), which are then used to predict the coherent response of the system, often using iterative procedures, and analyze experimental results.

Semiconductors are considerably more complex. Coulomb interaction profoundly modifies the response of semiconductors, not only in the linear regime (e.g., strong exciton resonance in the absorption spectrum) but also in the coherent and nonlinear regime. The influence of Coulomb interaction may be introduced by renormalizing the electron and hole energies (i.e., introducing excitons), and by renormalizing the field–matter interaction strength by introducing a renormalized Rabi frequency. These changes and the relaxation time approximation for the relaxation Hamiltonian lead to an analog of the optical Bloch equations, the semiconductor Bloch equations which have been very successful in analyzing the coherent response of semiconductors.

In spite of this success, it must be stressed that the relaxation time approximation, and the Boltzmann kinetic approach on which it is based, are not valid under all conditions. The Boltzmann kinetic approach is based on the assumption that the duration of the collision is much shorter than the interval between the collisions; i.e., the collisions are instantaneous. In the non-Markovian regime where these assumptions are not valid, each collision does not strictly conserve the energy and momentum and the quantum kinetic approach becomes more appropriate. This is true for photo-excited semiconductors as well as for the quantum transport regime in semiconductors. Also, semiconductor Bloch equations have been further extended by including four-particle correlations. Finally, the most general approach to the description of the nonequilibrium and coherent response of semiconductors following excitation by an ultrashort laser pulse is based on nonequilibrium Green's functions.

Experimental studies on the coherent response of semiconductors can be broadly divided into three categories:

1. investigation of novel aspects of coherent response not found in other simpler systems such as atoms;
2. investigation of how the initial coherence is lost, to gain insights into dephasing and decoherence processes; and
3. coherent control of processes in semiconductors using phase-locked pulses.

Numerous elegant studies have been reported. This section presents some observations on the trends in this field.

Historically, the initial experiments focused on the decay of the coherent response (time-integrated FWM as a function of delays between the pulses), analyzed them in terms of the independent two-level model, and obtained very useful information about various collision processes and rates (exciton–exciton, exciton–carrier, carrier–carrier, exciton–phonon, etc.). It was soon realized, however, that the noninteracting two-level model is not appropriate for semiconductors because of the strong influence of Coulomb interactions. Elegant techniques were then developed to explore the nature of coherent response of semiconductors. These included investigation of time- and spectrally resolved coherent response, and of both the amplitude and phase of the response to complement intensity measurements. These studies provided fundamental new insights into the nature of semiconductors and many-body processes and interactions in semiconductors. Many of these observations were well explained by the semiconductor Bloch equations. When lasers with <10 fs pulse widths became laboratory tools, the dynamics was explored on a time-scale much shorter than the characteristic phonon oscillation period (≈115 fs for GaAs longitudinal optical phonons), the plasma frequency (≈150 fs for a carrier density of $5 \times 10^{17}\ \mathrm{cm}^{-3}$), and the typical electron–phonon collision interval (≈200 fs in GaAs). Some remarkable features of quantum kinetics, such as reversal of the electron–phonon collision, memory effects, and energy nonconservation, were demonstrated. More recent experiments have demonstrated the influence of four-particle correlations on coherent nonlinear response of semiconductors such as GaAs in the presence of a strong magnetic field.

The ability to generate phase-locked pulses with controllably variable separation between them provides an exciting opportunity to manipulate a variety of excitations and processes within a semiconductor. If the separation between the two phase-locked pulses is adjusted to be less than the dephasing time of the system under investigation, then the amplitudes of the excitations produced by the two phase-locked pulses can interfere either constructively or destructively depending on the relative phase of the carrier waves in the two pulses. The nature of interference changes as the second pulse is delayed over an optical period. A number of elegant experiments have demonstrated coherent control of exciton population, spin orientation, resonant emission, and electrical current. Phase-sensitive detection of the linear or nonlinear emission provides additional insights into different aspects of semiconductor physics.

These experimental and theoretical studies of ultrafast coherent response of semiconductors have led to fundamental new insights into the physics of semiconductors and their coherence properties. Discussion of novel coherent phenomena is given in a subsequent section. Coherent spectroscopy of semiconductors continues to be a vibrant research field.

Incoherent Relaxation Dynamics

The qualitative picture that emerges from investigation of coherent dynamics can be summarized as follows. Consider a direct gap semiconductor excited by an ultrashort laser pulse of duration τ_L (with a spectral width $h\Delta\nu_L$) centered at photon energy $h\nu_L$ larger than the semiconductor bandgap E_g. At the beginning of the pulse the semiconductor does not know the pulse duration so that coherent polarization is created over a spectral region much larger than $h\Delta\nu_L$. If there are no phase-destroying events during the pulse (dephasing rate$1/\tau_D \ll 1/\tau_L$), then a destructive interference destroys the coherent polarization away from $h\nu_L$ with increasing time during the excitation pulse. Thus, the coherent polarization exists only over the spectral region $h\Delta\nu_L$ at the end of the pulse. This coherence will be eventually destroyed by collisions and the semiconductor will be left in an incoherent (off-diagonal elements of the density matrix are 0) nonequilibrium state with peaked electron and hole population distributions whose central energies and energy widths are determined by $h\nu_L$, $h\Delta\nu_L$ and E_g, if the energy and momentum relaxation rates $(1/\tau_E, 1/\tau_M)$ are much smaller than the dephasing rates.

The simplifying assumptions that neatly separate the excitation, dephasing and relaxation regimes are obviously not realistic. A combination of all these (and some other) physical parameters will determine the state of a realistic system at the end of the exciting pulse. However, after times $\sim \tau_D$ following the laser pulse, the semiconductor can be described in terms of electron and hole populations whose distribution functions are not Maxwell–Boltzmann or Fermi–Direct type, but nonthermal in a large majority of cases in which the energy and momentum relaxation rates are much smaller than the dephasing rates. This section discusses the dynamics of this incoherent state.

Extensive theoretical work has been devoted to quantitatively understand these initial athermal distributions and their further temporal evolution. These include Monte Carlo simulations based on the Boltzmann kinetic equation approach (typically appropriate for time-scales longer than ~100 fs, the carrier–phonon interaction time) as well as the quantum kinetic approach (for times typically less than ~100 fs) discussed above. We present below some observations on this vast field of research.

Nonthermal Regime

A large number of physical processes determines the evolution of the nonthermal electron and hole populations generated by the optical pulse. These include electron–electron, hole–hole, and electron–hole collisions, including plasma effects if the density is high enough, intervalley scattering in the conduction band and intervalence band scattering in valence bands, intersub-band scattering in quantum wells, electron–phonon, and hole–phonon scattering processes. The complexity of the problem is evident when one considers that many of these processes occur on the same time-scale. Of these myriad processes, only carrier–phonon interactions and electron–hole recombinations (generally much slower) can alter the total energy in electronic systems. Under typical experimental conditions, the redistribution of energy takes place before substantial transfer of energy to the phonon system. Thus, the nonthermal distributions first become thermal distributions with the same total energy. Investigation of the dynamics of how the nonthermal distribution evolves into a thermal distribution provides valuable information about the nature of various scattering processes (other than carrier–phonon scattering) described above.

This discussion of separating the processes that conserve energy in the electronic system and those that transfer energy to other systems is obviously too simplistic and depends strongly on the nature of the problem one is considering. The challenge for the experimentalist is to devise experiments that can isolate these phenomena so each can be studied separately. If the laser energy is such that $h\nu_L - E_g < \hbar\omega_{LO}$, the optical phonon energy, then majority of the photo-excited electrons and holes do not have sufficient energy to emit an optical phonon, the fastest of the carrier–phonon interaction processes. The initial nonthermal distribution is then modified primarily by processes other than phonon scattering and can be studied experimentally without the strong influence of the carrier–phonon interactions. Such experiments have indeed been performed both in bulk and quantum well semiconductors. These experiments have exhibited spectral hole burning in the pump-probe transmission spectra and thus demonstrated that the initial carrier distributions are indeed nonthermal and evolve to thermal distributions. Such experiments have provided quantitative information about various carrier–carrier scattering rates as a function of carrier density. In addition, experiments with $h\nu_L - E_g > \hbar\omega_{LO}$ and $h\nu_L - E_g >$ intervalley separation have provided valuable information about intravalley as well as intervalley electron–phonon scattering rates. Experiments have been performed in a variety of different semiconductors, including bulk semiconductors and undoped and modulation-doped quantum wells. The latter provide insights into intersub-band scattering processes and quantitative information about the rates.

These measurements have shown that under typical experimental conditions, the initial nonthermal distribution evolves into a thermal distribution in times $<$ or ~ 1 ps. However, the characteristic temperatures of the thermal distributions can be different for electrons and holes. Furthermore, different phonons may also have different characteristic temperatures, and may even be nonthermal even when the electronic distributions are thermal. This is the hot carrier regime that is discussed in the next subsection.

Hot Carrier Regime

The hot carriers, with characteristic temperatures T_c (T_e and T_h for electrons and holes, respectively), have high energy tails in the distribution functions that extend several kT_c higher than the respective Fermi energies. Some of these carriers have sufficient energy to emit an optical phonon. Since the carrier–optical phonon interaction rates and the phonon energies are rather large, this may be the most effective energy loss mechanism in many cases even though the number of such carriers is rather small. The emitted optical phonons have relatively small wavevectors and nonequilibrium populations larger than expected for the lattice temperature. These optical phonons are often referred to as hot phonons although nonequilibrium phonons may be a better term since they probably have nonthermal distributions. The dynamics of hot phonons will be discussed in the next subsection, but we mention here that in case of a large population of hot phonons, one must consider not only phonon emission but also phonon absorption. Acoustic phonons come into the picture at lower temperatures when the fraction of carriers that can emit optical phonons is extremely small or negligible, and also in the case of quantum wells with sub-band energy separation smaller than optical phonons.

The dynamics of hot carriers have been studied extensively to obtain a deeper understanding of carrier–phonon interactions and to obtain quantitative measures of the carrier–phonon interaction rates. Time resolved luminescence and pump-probe transmission spectroscopy are the primary tools used for such studies. For semiconductors like GaAs, the results indicate that polar optical phonon scattering (longitudinal optical phonons) dominates for electrons whereas polar and nonpolar optical phonon scattering contribute for holes. Electron–hole scattering is sufficiently strong in most cases to maintain a common electron and hole temperature for times $> \sim 1$ ps. Since many of these experiments are performed at relatively high densities, they provide important information about many-body effects such as screening. Comparison of bulk and quasi-two-dimensional semiconductors (quantum wells) has been a subject of considerable interest. The consensus appears to be that, in spite of significant differences in the nature of electronic states and phonons, similar processes with similar scattering rates dominate both types of semiconductors. These studies reveal that the energy loss rates are influenced by many factors such as the Pauli exclusion principle, degenerate (Fermi–Dirac) statistics, hot phonons, and screening and many-body aspects.

Hot carrier distribution can be generated not only by photo-excitation, but also by applying an electric field to a semiconductor. Although the process of creating the hot distributions is different, the processes by which such hot carriers cool to the lattice temperature are the same in two cases. Therefore, understanding obtained through one technique can be applied to the other case. In particular, the physical insights obtained from the optical excitation case can be extremely valuable for many electronic devices that operate at high electric fields, thus in the regime of hot carriers. Another important aspect is that electrons and holes can be investigated separately by using different doping. Furthermore, the energy loss rates can be determined quantitatively because the energy transferred from the electric field to the electrons or holes can be accurately determined by electrical measurements.

Isothermal Regime

Following the processes discussed above, the excitations in the semiconductor reach equilibrium with each other and the thermal bath. Recombination processes then return the semiconductor to its thermodynamic equilibrium. Radiative recombination typically occurs over a longer time-scale but there are some notable exceptions such as radiative recombination of excitons in quantum wells occurring on picosecond time-scales.

Hot Phonons

As discussed above, a large population of nonequilibrium optical phonons is created if a large density of hot carriers has sufficient energy to emit optical phonons. Understanding the dynamics of these nonequilibrium optical phonons is of intrinsic interest.

At low lattice temperatures, the equilibrium population of optical phonons is insignificant. The optical phonons created as a result of photoexcitation occupy a relatively narrow region of wavevector (k) space near $k = 0$ ($\sim 1\%$ of the Brillouin zone). This nonequilibrium phonon distribution can spread over a larger k-space by various processes, or anharmonically decay into multiple large-wavevector acoustic phonons. These acoustic phonons then scatter or decay into small-wavevector, low-energy acoustic phonons that are eventually thermalized.

Pump-probe Raman scattering provides the best means of investigating the dynamics of nonequilibrium optical phonons. Such studies, for bulk and quantum well semiconductors, have provided invaluable information about femtosecond dynamics of phonons. In particular, they have provided the rate at which the nonequilibrium phonons decay. However, this information is for one very small value of the phonon wavevector so the phonon distribution function within the optical phonon branch is not investigated. The large-wavevector acoustic phonons are even less accessible to experimental techniques. Measurements of thermal phonon propagation provide some information on this subject. Finally, the nature of phonons is considerably more complex in quantum wells, and some interesting results have been obtained on the dynamics of these phonons.

Tunneling and Transport Dynamics

Ultrafast studies have also made important contributions to tunneling and transport dynamics in semiconductors. Time-dependent imaging of the luminescing region of a sample excited by an ultrashort pulse provides information about spatial transport of carriers. Such a technique was used, for example, to demonstrate negative absolute mobility of electrons in p-modulation-doped quantum wells. The availability of near-field microscopes enhances the resolution of such techniques to measure lateral transport to the subwavelength case.

A different technique of ‘optical markers’ has been developed to investigate the dynamics of transport and tunneling in a direction perpendicular to the surface (vertical transport). This technique relies heavily on fabrication of semiconductor nanostructures with desired properties. The basic idea is to fabricate a sample in which different spatial regions have different spectral signatures. Thus if the carriers are created in one spatial region and are transported to another spatial region under the influence of an applied electric field, diffusion, or other processes, the transmission, reflection, and/or luminescence spectra of the sample change dynamically as the carriers are transported.

This technique has provided new insights into the physics of transport and tunneling. Investigation of perpendicular transport in graded-gap superlattices showed remarkable time-dependent changes in the spectra, and analysis of the results provided new insight into transport in a semiconductor of intermediate disorder. Dynamics of carrier capture in quantum wells from the barriers provided information not only about the fundamentals of capture dynamics, but also about how such dynamics affects the performance of semiconductor lasers with quantum well active regions.

The technique of optical markers has been applied successfully to investigate tunneling between the two quantum wells in an a-DQW (asymmetric double quantum well) structure in which two quantum wells of different widths (and hence different energy levels) are separated by a barrier. The separation between the energy levels of the system can be varied by applying an electric field perpendicular to the wells. In the absence of resonance between any energy levels, the wavefunction for a given energy level is localized in one well or the other. In this nonresonant case, it is possible to generate carriers in a selected quantum well by proper optical excitation. Dynamics of transfer of carriers to the other quantum well by tunneling can then be determined by measuring dynamic changes in the spectra. Such measurements have provided much insight into the nature and rates of tunneling, and demonstrated phonon resonances. Resonant tunneling, i.e., tunneling when two electronic levels are in resonance and are split due to strong coupling, has also been investigated extensively for both electrons and holes. One interesting insight obtained from such studies is that tunneling and relaxation must be considered in a unified framework, and not as sequential events, to explain the observations.

Novel Coherent Phenomena

Ultrafast spectroscopy of semiconductors has provided valuable insights into many other areas. We conclude this article by discussing some examples of how such techniques have demonstrated novel physical phenomena.

The first example once again considers an a-DQW biased in such a way that the lowest electron energy level in the wide quantum well is brought into resonance with the first excited electron level in the narrow quantum well. The corresponding hole energy levels are not in resonance. By choosing the optical excitation photon energy appropriately, the hole level is excited only in the wide well and the resonant electronic levels are excited in a linear superposition state such that the electrons at $t = 0$ also occupy only the wide quantum well. Since the two electron eigenstates have slightly different energies, their temporal evolutions are different with the result that the electron wavepacket oscillates back and forth between the two quantum wells in the absence of damping. The period of oscillation is determined by the splitting between the two resonant levels and can be controlled by an external electric field. Coherent oscillations of electronic wavepackets have indeed been experimentally observed by using the coherent nonlinear technique of four-wave mixing. Semiconductor quantum wells provide an excellent flexible system for such investigations.

Another example is the observation of Bloch oscillations in semiconductor superlattices. In 1928 Bloch demonstrated theoretically that an electron wavepacket composed of a superposition of states from a single energy band in a solid undergoes a periodic oscillation in energy and momentum space under certain conditions. An experimental demonstration of Bloch oscillations became possible only recently by applying ultrafast optical techniques to semiconductor superlattices. The large period of a superlattice (compared to the atomic period in a solid) makes it possible to satisfy the assumptions underlying Bloch's prediction. The experimental demonstration of Bloch oscillations was also performed using four-wave mixing techniques. This provides another prime example of the power of ultrafast optical studies.

Summary

Ultrafast spectroscopy of semiconductors and their nanostructures is an exciting field of research that has provided fundamental insights into important physical processes in semiconductors. This article has attempted to convey the breadth of this field and the diversity of physical processes addressed by this field. Many exciting developments in the field have been omitted out of necessity. The author hopes that this brief article inspires the readers to explore some of the Further Reading.

See also

Coherence: Coherence and Imaging. **Coherence Control:** Experimental; Theory. **Coherent Transients:** Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Scattering Theory.

Further Reading

Allen L and Eberly JH (1975) *Optical Resonance and Two-Level Atoms*. New York: Wiley Interscience.

Chemla DS (1999) Chapter 3. In: Garmire E and Kost AR (eds) *Nonlinear Optics in Semiconductors*, pp. 175–256. New York: Academic Press.

Chemla DS and Shah J (2001) Many-body and correlation effects in semiconductors. *Nature* 411: 549–557.

Haug H (2001) Quantum kinetics for femtosecond spectroscopy in semiconductors. In: Willardson RK and Weber ER (eds) *Semiconductors and Semimetals*, vol. 67, pp. 205–229. London: Academic Press.

Haug H and Jauho AP (1996) *Quantum Kinetics in Transport and Optics of Semiconductors*. Berlin: Springer-Verlag.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Kash JA and Tsang JC (1992) Chapter 3. In: Shank CV and Zakharchenya BP (eds) *Spectroscopy of Nonequilibrium Electrons and Phonons*, pp. 113–168. Amsterdam: Elsevier.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. New York: Oxford University Press.

Shah J (1992) *Hot Carriers in Semiconductor Nanostructures: Physics and Applications*. Boston, MA: Academic Press.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. 2nd edn. Heidelberg, Germany: Springer-Verlag.

Steel DG, Wang H, Jiang M, Ferrio KB and Cundiff ST (1994) In: Phillips RT (ed.) *Coherent Optical Interactions in Solids*, pp. 157–180. New York: Plenum Press.

Wegener M and Schaefer W (2002) *Semiconductor Optics and Transport*. Heidelberg, Germany: Springer-Verlag.

Zimmermann R (1988) *Many-Particle Theory of Highly Excited Semiconductors*. Leipzig, Germany: Teubner.

COLOR AND THE WORLD

G N Rao, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The human visual system is capable of detecting a small part of the electromagnetic spectrum which is called light. The visible region has a typical wavelength range of about 380–780 nm. Under optimal conditions the range is 360–830 nm. The most important point is that color is a perception. It is how human beings perceive the visual part of the electromagnetic spectrum. The color of an object perceived depends on the physical characteristics of the radiation illuminating the object, spectral scattering and reflectivity of the object, physiology of the eye, photochemistry of the sensors in the retina, the complex data processing that takes place in the visual system, and the psychology of color perception. In this review, we present an overview of color in nature and the world. Initially, color and wavelength is presented, then the human visual system and color perception. Then we go on to discuss measuring color and colorimetry and then color in nature, followed by a brief account of color in art and holography and the structural colors in nature.

Color and Wavelength (Spectral and Nonspectral Colors)

The ingenious yet simple experiments performed by Isaac Newton in the year 1666 and his ideas presented in his famous treatise 'Opticks', are the foundations of our understanding of color and color perception. Based on his experiments, Newton showed that different regions of the visible spectrum are perceived as different colors and the solar radiation (white light) consists of spectral colors violet, blue, cyan, green, yellow, orange, and red. Thus, natural day light is quite colorful though we perceive it as colorless. Table 1 gives approximate wavelength ranges of the spectral colors. Superposition of all the spectral colors results in a perception of white.

Colors can be produced in a number of ways:

1. *Gas discharges* emit characteristic radiation in the visible region. For example, neon discharge is red in color, argon discharge blue, sodium discharge yellow, helium-neon laser gives red radiation and argon ion laser blue-green radiation. A variety of fluorescent lamps are available and they provide continuous spectral power distribution with characteristic lines of mercury in the blue-green region. Gas-filled sodium and fluorescent lamps offer long life and high efficiency and are popular for indoor and outdoor lighting.
2. *Transition metal* (Cr, Mn, Fe, Co, Ni, Cu) compounds or transition metal impurities are responsible for the colors of the minerals, paints, gems, and pigments. They involve inorganic compounds of transition metal or rare earth ions with unpaired d or f orbitals. About 1% chromium sesquioxide (Cr_2O_3) in colorless aluminum oxide results in beautiful red ruby, a gemstone. The Cr^{3+} impurity in pure beryllium aluminum silicate gives a beautiful emerald green resulting in sapphire, another gemstone. Transition metal impurities are added to glass in the molten state to prepare colored glass. For example, adding Cr^{3+} gives green, Mn^{3+} purple, Fe^{3+} pale yellow, Co^{3+} reddish blue, Ni^{2+} brown, and Cu^{2+} greenish blue color.
3. *Color in Organic Compounds* – Most natural and synthetic dyes and biological (vegetable and animal origin) colorants are complicated organic molecules. They absorb light, exciting molecular orbitals. The dyes are used for food coloration, clothing, photographic, and sensitizer purposes.
4. *Colors of metals, semiconductor materials and color centers* – The colors of these materials are caused by the electronic transitions involving the energy bands.

   Diode lasers emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. They are inexpensive, compact, and can be operated using dry cell batteries. They are extensively used in compact disks, barcode readers, and optical communications.

   Light emitting diodes (LED) are also prepared using semiconductor materials and they emit radiation corresponding to the bandgap of the semiconductor materials used for their fabrication. LEDs are available in a variety of colors, red (GaAsP), orange (AlInGaP), yellow (InGaAs), green (CdSSe), blue (CdZnS) and are widely used in clocks, toys, tuners, displays, electronic billboards, and appliances.

   Color centers may be produced by irradiating some alkali halides and glass materials with electromagnetic radiation (gamma rays, X-rays, ultraviolet, and visible radiation) or with charged or uncharged particles (electrons, protons, neutrons). The irradiation produces an electron hole pair and the electron is trapped forming an electron center. The electron or hole, or both, form a color center absorbing part of the visible radiation. The color centers can be reversed by heating to high temperatures.
5. *Optical Phenomenon* lead to spectacular colors in nature and in biological materials. Scattering of light by the atmosphere is responsible for colors of the sky, beautiful sunsets and sunrise, twilight, blue moon, and urban glows. Dispersion and polarization cause rainbows and halos. Interference gives rise to the beautiful colors of thin films on water, foam, bubbles, and some biological colors (butterflies). Diffraction is responsible for the colors of liquid crystals, coronae, color of gems (opal), color of animals (sea mouse), the colors of butterfly (*Morpho rhetenor*), and diffraction gratings.

Table 1 Approximate wavelength ranges of different spectral colors

Color	*Wavelength range (nm)*
Violet	380–435
Blue	435–500
Cyan	500–520
Green	520–565
Yellow	565–597
Orange	597–625
Red	625–780

The Human Visual System and Color Perception

Photoreceptors in the Eye: Rods and Cones

As light enters the eye, the cornea and the eye lens focus the image on the retina. The retina has two kinds of photoreceptors: rods and cones so named because of their shapes. Typically, there are about 7 million cones and 110–125 million rods. Cones are activated during photopic (daylight) conditions and the rods during scotopic (night) conditions. Cones are responsible for color perception whereas rods can perceive only white and gray shades. Cones are predominantly located near the fovea, the central region of the retina. When light falls on the rods and the cones, the photosensitive material in them, namely rhodopsin, is activated, generating an electrical signal which is further processed in the visual system and transmitted to the visual cortex. Considerable data processing takes place within the retina.

Color Perception

Subjective nature of color

Color is a perception and is subjective. There is no way one can define color in an absolute or quantitative means, even though it can be precisely characterized by the spectral power distribution (SPD) of the radiation. For monochromatic light of different wavelengths we see the spectral colors detailed in Table 1. In general, the SPD of light from an object is complex and to predict the exact color of such a spectrum is quite difficult. However, in the following we discuss some simple color schemes.

Additive color mixing

Newton's experiments demonstrated that using the three primary color light sources, blue, green, and red and by varying their relative intensities, one can generate most of the colors by additive color mixing. Additive color mixing is employed in stage lighting, and large screen color television projection systems. If we are concerned only with the hue and ignore saturation and brightness, the superposition of colored light beams of equal brightness on a white screen gives the following simple additive color scheme:

$$\text{Green} + \text{Red} \Rightarrow \text{Yellow}$$

$$\text{Blue} + \text{Red} \Rightarrow \text{Magenta}$$

$$\text{Blue} + \text{Green} \Rightarrow \text{Cyan}$$

$$\text{Blue} + \text{Green} + \text{Red} \Rightarrow \text{White}$$

Subtractive color mixing

When white light passes through, say, a red filter, the radiation corresponding to the red radiation is transmitted and all other parts of the visible spectrum are absorbed. If we insert more red filters the brightness of the light transmitted becomes less though the hue is likely to be more saturated. If you insert red and blue filters together in a white beam, you will see no light transmitted because the red filter blocks the entire spectrum except the radiation corresponding to red color which in turn is blocked by the blue filter.

Subtractive color scheme is also applicable when paints of various colors are mixed together. The subtractive primary colors are magenta, yellow, and cyan. The artists often call them red, yellow, and blue. By appropriately mixing the primary colored paints one can, in principle, generate most of the colors.

How many colors can we differentiate?

If we do a simple experiment beginning from the lowest wavelength of the visible spectrum, say 380 nm, and gradually move to the long wavelength end of the visible spectrum, one can differentiate about 150 spectral colors. For each spectral color there will be a number of colors that vary in brightness and saturation. Additionally, we can have additive mixtures of these spectral colors, which in turn have variations in brightness and saturation. It is estimated that under optimal conditions we can differentiate as many as 7 to10 million different colors and shades.

Temporal Response of the Human Visual System

Positive afterimages

The visual system responds to changes in the stimulation in time. However, the response of the visual system is both delayed and persists for some time after the stimulation is off. Because of the persistence of vision, one would see positive afterimages. The afterimages last about 1/20 s under normal conditions but are shorter at high light conditions. A rapid succession of images show results in a perception of continuous motion which is the basis for movies and TV.

Close your eyes for about five minutes so that any after images of the objects you have seen before are cleared. Open your eyes to see a bright colored object and then close your eyes again; you will see positive afterimages of the object in its original colors. Even though the positive afterimage at the initial stages will be of the same colors as the original, they gradually change color showing that different cones recover at different rates.

Negative afterimages

If you stare at a color picture intensely for about a minute and look at a white background, such as a sheet of white paper, you will see the complementary colors of the picture. The image of the object falls on a particular part of the retina and if we look at a green colored object, the cones that are more sensitive to the mid-wavelength region are activated and are likely to be saturated. If you look at a bright sheet of white paper immediately, the cones that are sensitive to the middle wavelength region are inactivated since they need some time to replenish the active chemical whereas the cones that have maximum sensitivity for the low and high wavelengths are active, therefore you will see the object in complementary colors. Yellow will be seen after looking at blue image, cyan after looking at red image and vice versa.

Spatial Resolution of the Human Visual System

Lateral inhibition (color constancy)

By and large the perceived color of an object is independent of the nature of the illuminating white light source. The color of the object does not change markedly under different white light sources of illumination. For example, we see basically the same color of the object in the daytime in the light of the blue sky or under an incandescent light source even though the spectral power distributions of these sources are different and the intensity of light reaching the eye in both cases is also drastically different. The color constancy is a result of chromatic lateral inhibition. The human visual system has a remarkable way of comparing the signals from the object and manipulating them with the signals generated from the surroundings and picking up the correct color independent of the illuminating source. The color constancy and simultaneous lightness contrast are due to lateral inhibition. Lateral inhibition was exploited by a number of artists to provide certain visual effects that would not have been possible otherwise. For example, Georges Seurat in his famous painting 'La Poseuse en Profil' improved the light intensity contrast in different regions of his painting by employing edge enhancement effects resulting from lateral inhibition. Victor Vasarely's art work 'Arcturus' demonstrates the visual effects of lateral inhibition in different colors.

Spatial assimilation

The color of a region assimilates the color of the neighboring region, thus changing the perception of its color. Closely spaced colors, when viewed from a distance, tend to mix partitively and are not seen as distinct colors but as a uniform color due to additive color mixing. Georges Seurat in his painting 'Sunday Afternoon on the Island of La Grande Jatte', uses thousands of tiny dots of different colors to produce a variety of color sensations, a technique which is known as pointillism. The visual effects of spatial assimilation were clearly demonstrated by Richard Anuszkiewicz in his painting 'All Things Do Live in the Three'.

Color Perception Models

Thomas Young, in 1801, postulated that there are three types of sensors thus proposing trichromacy to explain the three attributes of color perception: hue, saturation, and brightness. Data on microspectrophotometry of excised retinas, reflection densitometry of normal eyes, and psychophysical studies of different observers confirm the existence of three types of cones. As shown in Figure 1, the three types of cones have different responsivities to light. The cones that have maximum responsivity in the short wavelength region are often refereed as S-cones, the cones that have maximum in the intermediate wavelength range as I-cones, and the cones that have maximum in the long wavelength region as L-cones.

The three response curves have considerable overlap. The overlap of the response curves and the complex data processing that takes place in the visual system enable us to differentiate as many as 150 spectral colors and about seven to ten million different colors and shades. An important consequence of the existence of only three types of color sensors is that different spectral power distributions can produce identical stimuli resulting in the perception of the same color. Such spectral power distributions are called metamers.

The neural impulses generated by the cones are processed through a complex cross-linking of bipolar cells, horizontal cells, amacrine cells, and communicated to ganglion cells leading to the optic chiasmas through the optic nerve. Information gathered by about 100 million photoreceptors, after considerable processing, is transmitted by about a million ganglion cells.

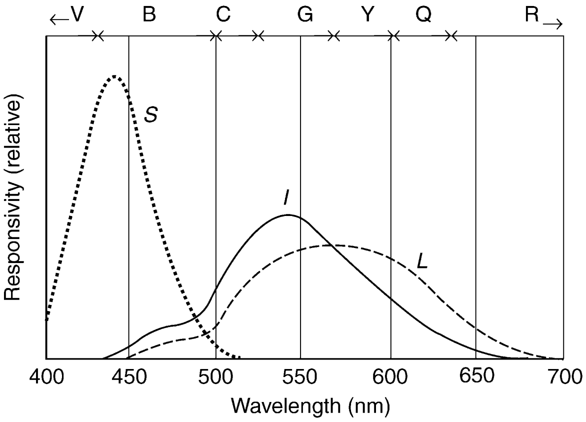

Figure 1 Relative spectral absorption of the three types of cones. *S*, *I*, and *L* respectively stand for the cones that have maximum responsivity for the short, intermediate, and long wavelength regions.

The zone theory of color vision assumes two separate zones where the signals are sequentially processed. In the first zone the signals are generated by the cones and in the second zone, the signals are processed to generate one achromatic and two opponent chromatic signals. Possibly other zones exist where the signals generated in the first and the second zones are further processed and communicated to the visual cortex where they are processed and interpreted based on the temporal and spatial response and comparisons with the information already available in the memory.

Measuring Color and Colorimetry

Since color is a human perception, how do we measure color? Colorimetry is based on the human visual system and how different observers perceive color. Though, the spectral power distribution of light is helpful for characterizing spectral hues, it is very difficult to infer the color from the spectral power distribution.

The three attributes of color perception are hue, saturation, and brightness. Hue is the color such as red, green, and blue. Saturation refers to the purity of the color. A color is said to be more saturated if the whiteness in the color is less and vice versa. Brightness is related to the intensity of light. For the sun at sunrise the hue is red, saturated, and low brightness. About half an hour after sunrise, the hue is yellow, less saturated, and higher brightness. Though a variety of color schemes are available in the literature, in the following we discuss the Munsell system and the C.I.E. diagram only.

Munsell Color System

Albert Munsell, a painter and art teacher, devised a color atlas in 1905 by arranging different colors in an ordered three-dimensional space. He used the three color attributes hue, chroma, and value corresponding to hue, saturation, and brightness. In this system there are 10 basic hues, each of which is divided into 10 equal gradations, resulting in 100 equally spaced hues. Each hue has a chart with a number of small chips arranged in rows and columns. Under daylight illumination, the chips in any column are supposed to have colors of equal chroma and the chips in any row are supposed to be of equal brightness. The brightness increases from the bottom of the chart to the top in steps that are perceptually equal. The saturation of the color increases from the inside edge to the outside edge of the chart in steps that are also perceptually equal. Each chip is identified by hue, value/chroma. In the Munsell scale, black is given a value of 0 and white 10, and the nine grays are uniformly placed in between. Though the Munsell system has gone through many revisions, typically it has a total of about 1450 chips. The Munsell color scheme, though subjective in nature, is widely used because of its simplicity and ease.

The Commission Internationale de l'Echlairage (C.I.E.) Diagram

Even though the visible light reaching the eye has a very complex SPD, color perception is mostly dependent on the signals generated by the three different types of cones. However, we do not know precisely the absorption and sensitivity of the cones and therefore are not in a position to calculate the signal strengths generated by the three different cones. Additionally, the response of the cones is specific to the individual. Therefore, the C.I.E. system is based on the average response of a large sample of normal observers. The C.I.E. system employs three color matching functions $\bar{x}(\lambda), \bar{y}(\lambda)$, and $\bar{z}(\lambda)$ (Figure 2), based on the psychological observations of a large number of standard observers to calculate the 'tristimulus values' X, Y, and Z defined as

$$X = c\int S(\lambda)\bar{x}(\lambda)\mathrm{d}\lambda \qquad [1]$$

$$Y = c\int S(\lambda)\bar{y}(\lambda)\mathrm{d}\lambda \qquad [2]$$

$$Z = c\int S(\lambda)\bar{z}(\lambda)\mathrm{d}\lambda \qquad [3]$$

Here, λ is the wavelength, $S(\lambda)$ the spectral power distribution, and c is the normalization constant. The integration is carried out over the visible region, normally 380 to 780 nm. The system is based on the assumption of additivity and linearity. The sum of the tristimulus values $(X + Y + Z)$ is normalized to 1. We choose $y(\lambda)$ so that the Y tristimulus value is proportional to the luminance which is a quantitative measure of the intensity of light leaving the surface. We further calculate x and y which are called chromaticity coordinates:

$$x = \frac{X}{(X + Y + Z)} \qquad [4]$$

$$y = \frac{Y}{(X + Y + Z)} \qquad [5]$$

The psychological color is then specified by the coordinates (Y, x, y). Relationship between colors is usually displayed by plotting the x and y values on

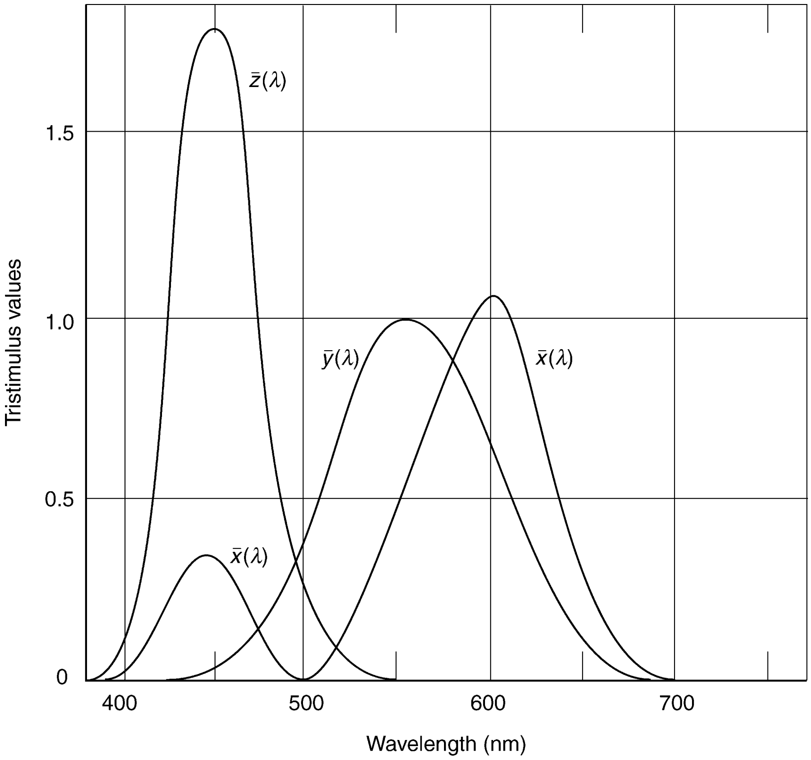

Figure 2 Color-matching functions $\bar{x}$ (λ),$\bar{y}$ (λ), and $\bar{z}$ (λ) of the C.I.E. 1931 standard observer.

a two-dimensional Cartesian Coordinate system with z as an implicit parameter ($z = 1 - x - y$) which is known as the C.I.E. chromaticity diagram (Figure 3). In this plot, the monochromatic hues are on the perimeter of a horseshoe shaped curve and is called spectrum locus. A straight line joining the ends of the spectrum is called the line of purples. Every chromaticity is represented by means of two coordinates (x, y) in the chromaticity diagram. The point corresponding to $x = 1/3$ and $y = 1/3$ (also $z = 1/3$) is the point of 'equal energy' and represents achromatic point. The complementary colors are on the opposite sides of the achromatic point. Complementary wavelength of a color is obtained by drawing a line from the color through the achromatic point to the perimeter on the other side.

Continuum (white) sources

A perfect white source is the one that has constant SPD over the entire visible region. Such a source will not distort the color of the objects they illuminate. A number of sources in general are considered as white sources; they include incandescent lamps, day skylight, a variety of fluorescent lamps, etc. It should be noted that the SPD of these sources is not constant and varies with the source. Typical SPDs of commonly used light sources are given in Figure 4.

Color temperature

The thermal radiation emitted by a black body has a characteristic SPD which depends solely on its temperature. The wavelength λ of the peak of this distribution is inversely proportional to its absolute temperature. As the absolute temperature of the blackbody increases, the peak shifts towards shorter wavelengths (blue region) and also the width of the distribution decreases. Since the peak of the black body radiation is dependent on the absolute temperature, the color of the blackbody can be approximately defined by its temperature.

The color temperature of a body is that temperature at which the SPD of the body best matches with that of the blackbody radiation at that temperature. The concept of color temperature is often used to characterize the sources. Low color temperature sources have a reddish tinge and high color temperature sources tend to be slightly bluish. For example, the color temperature of red looking star Antares is about 5000 K and that of the bluish white looking Sirius is about 11 000 K. One should note that the color temperature of a body is different from its actual temperature. The fluorescent lamps have a

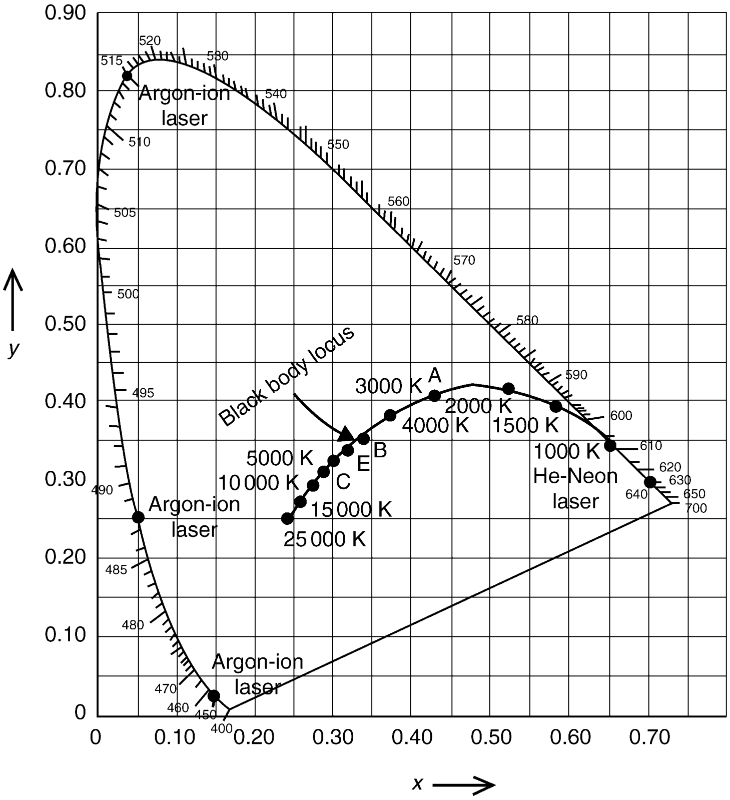

Figure 3 Chromaticity diagram of the C.I.E. 1931 standard observer. The chromaticities of the incandescent sources and their color temperatures are given inside the curve. A, B, and C are the C.I.E. standard illuminants. The chromaticities of the helium-neon laser (red, 633 nm) and the argon-ion laser (green 515 nm, blue 488 nm, and indigo, 458 nm) are also shown.

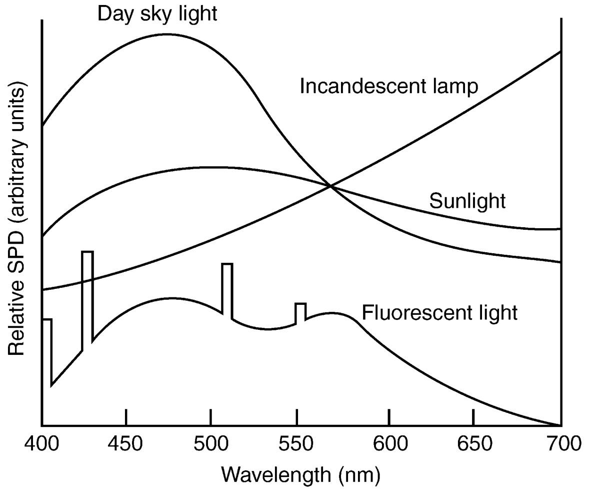

Figure 4 Relative SPDs of commonly used white light sources.

color temperature range of 5000–7500 K, incandescent lamps about 3000 K, direct sunlight ~4870 K and overcast skylight has a color temperature of ~7000 K.

The C.I.E. chromaticity diagram along with the color temperatures of the commonly employed sources, is shown in Figure 3.

Color in the Nature

As the solar radiation passes through the atmosphere, it is partly absorbed and partly scattered by the atmospheric constituents: air, water vapor, and dust particles. Absorption depends on the molecular properties of the air molecules, water vapor, and the dust particles, and is strongly wavelength dependent. Absorbed energy increases the internal energy of the molecule finally manifesting itself as thermal energy. When the sun is at the zenith, the solar radiation passes through one air mass loosing approximately 32% of its power. The SPD of the solar radiation above the atmosphere and at the Earth's surface at air mass 1 is shown in Figure 5. Scattering and absorption of the solar radiation by the atmosphere are responsible for a variety of interesting colorful views in the nature and they are briefly discussed in this section.

Atmospheric Colors

1. *Colors of the sky.* Scattering occurs over most of the visible part of the electromagnetic spectrum and is a function of the frequency of

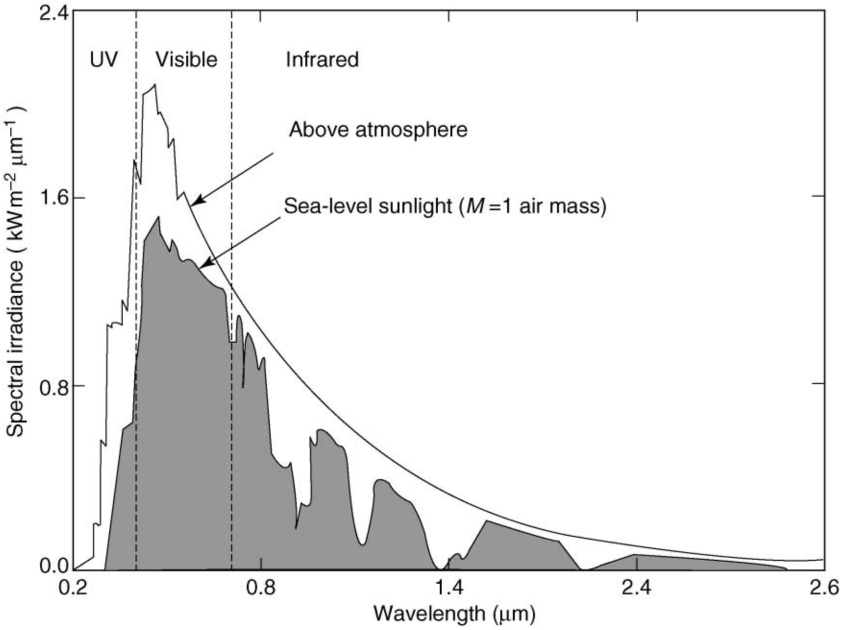

Figure 5 Spectral irradiance (kW m^{-2} µm^{-1}) of the solar radiation measured above the atmosphere and at the Earth's surface at sea level on a clear day and the sun at the zenith.

the radiation. According to Lord Raleigh's theory of scattering, the scattering probability varies inversely as a function of the fourth power of the wavelength of the radiation. If we consider the radiation corresponding to blue (400 nm) and the radiation corresponding to say red (700 nm), the photons corresponding to blue radiation are scattered by about 9.4 times that of photons corresponding to the red radiation. During noontime, when we look at the sky, not directly at the sun, the sky looks unsaturated blue because of the predominant scattering experienced by the high frequency radiation corresponding to blue color. However, it should be noted that the daytime skylight contains all the spectral colors of the solar radiation though the spectral power distribution has changed resulting in an increase in the intensity of the high frequency (blue) radiation. During noontime at the horizon, the sky looks close to white in color with a bluish tinge, because the solar radiation has passed through many air masses. As the thickness of the atmosphere increases, the intensity of the blue radiation decreases and ultimately as the radiation reaches the horizon, all the spectral colors are scattered about the same level resulting in white or bluish white sky.

2. *Polarization of the solar radiation by the atmosphere*. Light scattered by a molecule is polarized. The blue skylight is polarized with a maximum polarization of about 75% to 85% at 90° from the sun.
3. *Distant mountains seem bluish in color.* When an artist wishes to paint a distant mountain in perspective, he/she paints it blue; why? This is called airlight. The low wavelength part of the solar radiation is scattered more prominently, giving it a blue perspective when observed from a distance. However, it should be mentioned that the distant mountains may look any color depending on the scattering and absorption of the solar radiation sometimes resulting in 'Purple Mountain's majesty.'
4. *Sun looks red at the sunrise and sunset*. At sunset as the sun approaches the horizon, the sun changes its color from dazzling white to yellow, orange, bright red, and dull red and the order is reversed at sunrise. At the sunsets and sunrise, the solar radiation travels through many airmasses compared to approximately one airmass when the sun is at the zenith. Because of the large mass of air the radiation has to pass through, first the blue part of the spectrum is removed and the transmitted green to red spectral colors give the appearance of yellow. As this spectrum passes through more airmass, green is scattered most, leaving spectral yellow, orange, and red. As the radiation further passes through air, yellow and orange are also scattered, leaving only red.

5. *Twilight*. After sunset (dusk), or before the sunrise (dawn), the sky is illuminated by light which in turn illuminates the landscape. At sunsets and sunrise the solar radiation passes through a large mass of air. At a solar depression of about 0°, most of the blue is lost by scattering, resulting in a yellow arch and bright sky, and as the solar depression increases to about 10°, the arch will be reddish and the brightness of the sky reduced.
6. *Alpenglow*. Snow-covered mountains glow orange yellow from the low sun and after the sunset the same mountains look purple because of the reflection of the twilight.
7. *Twinkling of stars*. The stars twinkle. The local variations in the density of air are responsible for the twinkling of stars. As the density of air changes, so does the index of refraction and the starlight is continually refracted and dispersed. Twinkling is more prominent when the stars are on the horizon or on cold clear and windy nights.
8. *Mirages*. Mirages are refracted images produced because of the existence of refractive index gradients in the atmosphere. This results in the sight of shimmering water on a highway, inverted view of a car, etc.
9. *Aurora Borealis*. Aurorae are diffused slowly moving lights seen at high latitudes. They are often greenish yellow with different shapes. The lights in the northern hemisphere are called 'Aurora Borealis' and the lights in the southern hemisphere 'Aurora Australis.' They are caused primarily by the high-energy electrons deflected by the Earth's magnetic field towards the Earth's magnetic poles interacting with the atmospheric molecules resulting in molecular excitation and ionization. When the molecules de-excite, the characteristic electromagnetic radiation is commonly emitted. The commonly seen yellow-green light is due to the molecular emission of oxygen at 557 nm.
10. *Urban glows*. The scattered light from cities can be observed from space. The metropolitan Los Angeles can be observed from about 250 km in space. Extensively used sodium lamps with strong spectral yellow and blue-green mercury and fluorescent lights have unique color spectra of their own.

Colors due to Water and Water Droplets

1. *What is the color of pure water?* The transmission coefficient of pure water as a function of wavelength for different thicknesses of water column is shown in Figure 6. Our perception of the color of water changes with the thickness of water. For 1 mm thick water column, all the wavelengths are transmitted, resulting in the perception of white; 1 m thick water column looks unsaturated bluish green. As the thickness of water increases, the bluish green color will get more and more saturated. The color of water perceived by us depends on: (i) light reflected from the water surface; (ii) the light refracted at the water surface and diffused and scattered by the water molecules; and (iii) the refracted light reaches the bottom and gets reflected, which is further diffused and scattered and refracted at the water surface again. What we see is an additive sum of all these components.

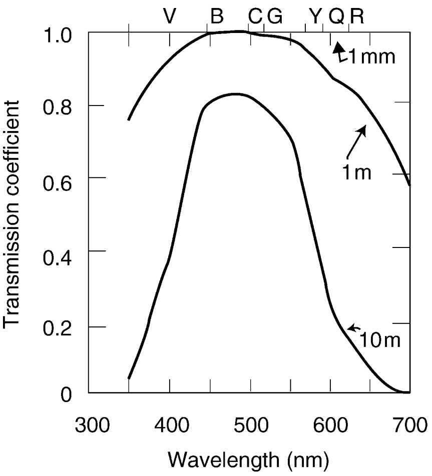

Figure 6 Transmission coefficient of water as a function of wavelength for different thicknesses of water (1 mm, 1 m, and 10 m).

2. *Refraction through water–air interface*. When light travels from an optically denser medium to an optically lighter medium, the light is deviated away from the normal. This is the reason why the apparent depth of a fish located in a pond is less than the true depth and a pencil struck in water seems to bend. If you are a diver when you look at the horizon from under water, the hemisphere seems to be compressed into a circle of diameter 97.2° instead of 180° which is known as 'Optical Manhole.'
3. *Glitter.* When you look at the reflection of the sun or the moon in wavy water, you notice a bright elongated reflected light known as glitter. It has a brilliant reddish orange color when the sun is above the horizon. Since the water is wavy, light reflected by a certain part of the wave at a certain time has the correct angle of reflection to reach your eye, which is known as glint. An ensemble of a large number of glints

coming from different waves and different parts of the waves results in glitter. Since you are looking at the reflected light, it will be polarized.

4. *Sky pools and landpools*. The light coming from a wavy water surface of a lake consists of oval shaped bright colored patterns that continually change in shape. This results from the reflection of skylight by the wavy water. If the waves are not turbulent the reflected light reaching your eye continually changes resulting in bright colored patterns. Landpools are caused by a nearby sloping landscape and the reflected light pattern from the landscape.
5. *Thin films*. Oil floats on water and spreads into thin films because of its low surface tension. When viewed in reflected natural white light, the light reflected from the top surface of the oil and the light reflected from the interfacial surface of oil and water interfere to produce colorful patterns. Brilliant colored patterns are commonly seen on the roads after rain because of gasoline on water. Foam ordinarily consists of bubbles ranging typically 10^{-4} mm to several mm. Even though each bubble produces colorful light due to the interference effects, the light coming out of the foam adds up to white light.
6. *Wet surfaces look darker*. A water spot on a concrete road looks darker. Since concrete is not porous a thin layer of water sticks to the surface. At the water surface part of the light incident is reflected, and a large part is transmitted. The transmitted light is diffusively scattered, part of which goes through total internal reflection and only a small part reaches the eye, making the surface look darker.

   When a fabric is wet it looks darker because of a slightly different mechanism. The surface of the fabric gets coated with a thin layer of water (refractive index $= 1.33$) which is smaller than the refractive index of the fabric. When the cloth is wet more light is transmitted, and less gets reflected than before. The transmitted light penetrates the fabric and gets scattered, resulting in a darker appearance.
7. *Rainbows*. Water with a refractive index of 1.33 is a dispersive medium. White light consisting of different wavelengths travel at different speeds in the medium and emerge at different directions. If the dispersion is adequate, different spectral colors can be seen by the naked eye. The dew covered lawn on a sunny day displays the spectral colors. If you stand close to the dew drop without blocking the sun's radiation falling on the drop, at some angle one should see the spectral colors. The primary rainbow is produced by the water droplets in the atmosphere after rain. The radius of the primary rainbow is about 42° with a width of about 2°. The color of the primary rainbow would be blue to red from the center of the rainbow. A rainbow is observed at the antisolar point. The brightness of the rainbow depends on the ambient sunlight, the size of the droplets, and whether the sun is low or high in the sky. The secondary rainbow, which is fainter, is often seen outside the primary rainbow at about 51°. The secondary rainbow will have red to blue from the center of the secondary rainbow. The secondary rainbow is approximately half as bright as the primary rainbow. The secondary rainbow is caused by two total internal reflections within the droplet whereas the primary rainbow is the result of one total internal reflection only. Because of the total internal reflection process, light is about 95% tangentially polarized in the primary rainbow.
8. *Heiligenschein*. Heiligenschein or the 'holy light' can be seen around the shadow of your head on a lawn on a sunny morning when dew drops are present. Usually, the dew drops are held by the tiny hairs of the grass at some distance from the blade of the grass. As light falls on the droplet, the spherical drop converges light at its focus and if the grass blade is located at that point, the light gets retroreflected approximately towards the sun resulting in holy light around your shadow.
9. *Coronae*. Colored concentric patchy rings seen around the moon and the sun are called coronae. They are primarily produced by the water droplets in the thin clouds. Even though the water droplets are randomly located, the diffracted light from the droplets has a defined angle and the superposition of the diffraction patterns from different water droplets results in colored ring structures. Coronae can be easily seen around the full moon viewed through thin clouds.
10. *The Glory*. On a foggy day with the sun at your back, one may see bright colored rings in the form of arches of about 5° to 10° at the solar point. The light in the glory is polarized and the ring structure is produced by reflection of the solar radiation by the fog.
11. *Clouds*. Clouds contain suspended water droplets or ice crystals. The size of the droplets ranges from less than a micrometer to about 100 micrometers (10^{-6} m). When the air is supersaturated with water vapor, the water condenses into water droplets, resulting in a definite shape to the cloud. When the sun's radiation falls on

the cloud, the light is scattered by the water droplets. Even though one would expect that the scattered light may have some color distribution, since we look at the scattered light from a very large number of droplets, the summation of all the colors results in white. Some clouds look dark or gray because, either they are in the shadow of other clouds or they are so thick that most of the light is absorbed. During sunset and sunrise, the clouds look bright orange because of the colored light illuminating them. Even though all the clouds are basically white, they may appear colored because of the selective absorption of certain colors in the atmosphere or due to the colored background radiation. Often, you see a silver lining at the edges of a cloud, because at the edges the density of water is low and when the sun illuminates it from the back, the radiation is scattered in the forward direction quite efficiently, resulting in a bright and shiny 'silver lining' to the cloud.

12. *Lightning*. Lightning is caused by an intense electrical discharge due to the high potential differences developed. On a cloudy and rainy day, as the water droplets grow in size, the falling drops get distorted to ellipsoidal shape and get polarized. The top smaller portion becoming negative and the lower larger part becoming positive. As the deformation increases, the drop ultimately breaks and the smaller drops being negatively charged and the larger ones positively charged. The wind currents easily lift the buoyant small drops to higher altitudes and the heavier positively charged droplets fall towards the Earth. The final result is that the upper region of the cloud is negatively charged whereas the lower region of the cloud is positively charged. The portion of the Earth under the cloud becomes negatively charged due to induction. The localized ions in the positively charged cloud produce a strong electric field which is further augmented by the induced negative charges on the Earth below. The strong electric field ionizes the air molecules setting up a current discharge to the earth. This discharge is called stepped leader. Following the stepped leader, the charges in the upper region of the cloud also discharge resulting in what is called return stroke. The sequence of leader and return strokes occur typically about ten times in a second. Since the clouds are charged and they are closer to each other, lightening is more frequent between the oppositely charged clouds than cloud and Earth. It is estimated that the lightning between the clouds is about five times more frequent compared to the lightening events between a cloud and the Earth. In this process, the nitrogen and oxygen molecules in air are excited and ionized. The flash is produced by the de-excitation of these molecular species. Since the temperatures involved are quite high (30 000 K), the flash has a strong unsaturated bluish hue. Some hydrogen lines are also observed in the flash. Hydrogen is produced as a result of the dissociation of water molecules during the discharge process.
13. *Ice and Halos*. Snow reflects most of the light falling with a reflectivity of about 95%. Snowflakes are crystalline and have hexagonal structure. In nature they occur in assorted sizes and shapes. As white light falls on them, light may be simply reflected from their surfaces making them shining white, which are often called glints. The light may also pass through the crystalline material get dispersed and emerge with different colors, which are known as sparkles.

    The halos appear as thin rings around the sun or the moon. They are produced because of the minimum deviation of light in snow crystals. The 22° halo is most common and is often seen a number of times in a year. The halos are circular and have a faint reddish outer edge and a bluish white diffused spot at the center. They often look white because of the superposition of different colors resulting in a more or less white color.

Color in Art and Holography

Sources of Color

There are two major color sources, dyes and paint pigments. Dyes are usually dissolved in a solvent. The dye molecules have selective absorption resulting in selective transmission. Paint pigments are powdered and suspended in oil or acrylic. The paint particles reflect part of the radiation falling on them, selectively absorb part of the radiation and transmit the balance. Lakes are made of translucent materials such as tiny grains of alumina soaked in a dye of appropriate color.

In the case of water colors and printer's inks, part of the light incident is reflected from the surface and the remaining transmitted which is selectively absorbed by the dye. The transmitted light reflected by the paper, passes through the dye again and emerges. Each transmission produces color by subtractive color scheme and the emerging light is mixed partitively by the observer.

In the case of paints, the color depends on the size and concentration of the pigment particles and the medium in which they are suspended. Part of the light incident on the artwork gets reflected at the surface and the remaining selectively absorbed and transmitted which in turn is further reflected and transmitted by other pigment particles each time following a subtractive color scheme. The light reaching the canvas gets reflected and passes through the medium and emerges. Again, all the emerging beams are partitively mixed by the observer.

Holograms may be broadly categorized as transmission and reflection holograms. Both transmission and reflection (volume) holograms are prepared by producing the interference pattern of the reference and object laser beams. Transmission holograms can be viewed in the presence of the reference laser beam that was employed for its preparation. Spectacular three-dimensional colored images are reconstructed by the reflection holograms depending on the angle of incidence of the white light and the direction of observation. Embossed holograms are popular for advertising, security applications (credit cards), display, and banknotes. Some studios prepare three-dimensional holographic portraits that can be viewed in natural light. In the future, three-dimensional videos and even holographic three-dimensional movies may be a reality.

Structural Colors in Nature

Nanoscale Photonic Lattices

Newton observed that iridescence (change in color with the direction of observation) is caused by optical interference. The colors are produced by the interference or the light diffracting characteristics of the microstructures and not by the selective reflection or absorption of light, as in the case of pigments. Photonic band structures consist of periodic arrangement of dielectric materials even though they are transparent, have a bandgap resulting in inhibition of certain wavelengths. Light with frequencies in the gaps between the bands get reflected resulting in the specific color.

The structural colors of light can be seen in the animal world and minerals. For example, opal, a sedimentary gem does not have crystalline structure, but consists of tiny spheres of silica packed together. The diffraction of light through these structures results in spectacular colors determined by the size of the spheres and their periodicity. The spines of the sea mouse Aphrodite are covered with natural photonic structures that produce iridescence resulting in brilliant colors depending on the direction of the incident light and the direction of observation.

In the case of the butterfly, *Morpho rhetenor*, the metallic blue color is produced by the periodic structure of its wings. Its wings have layers of scales, each of about 200 mm long and 80 mm wide and about 1300 thin parallel structures per mm which form a diffraction grating.

The bird feathers viewed through an optical microscope show the structural components which are responsible for their coloration. The structural details of the peacock's eye pattern employing the scanning electron microscopy show photonic structures on peacock barbules. Barbules are the strap like branches on the peacock feathers. The four principal colors on the peacock feathers blue, green, yellow, and brown were identified wth the specific structures. The interference between the light reflected from the front surface to that reflected from the back was shown to be responsible for the brown color in the peacock's eye. The photonic structure of the green barbule has a periodic structure of melanin cylinders with a spacing of about 150 nm whereas the blue barbule has a periodic structure of 140 nm, and yellow barbules have a periodicity of 165 nm.

See also

Dispersion Management. **Holography, Techniques:** Color Holography. **Photon Picture of Light**. **Polarization:** Introduction. **Scattering:** Scattering Theory.

Further Reading

Blau SK (2004) Light as a feather: structural elements give peacock plumes their color. *Physics Today* 57: 18–21.

Color Science, http://www.physics.sfasu.edu/astro/color.html.

Colourware, Colour FAQ, http://www.colourware.co.uk/cpfaq.htm.

CVRL Color & Vision database, http://www.cvrl.org/.

Hecht E (2002) *Optics*, 4th edn. San Francisco, CA: Addison Wesley.

Lynch DK and Livingston W (2001) *Color and Light in Nature*, 2nd edn. Cambrige, UK: Cambridge University Press.

Minnaert MGJ (1993) *Light and Color in the Outdoors*. Heidelbergh: Springer-Verlag.

Nassau K (2000) *The Physics and Chemistry of Color*. New York: John Wiley and Sons.

Poynton CA, Color technology, http://www.Poynton.com/ColorFAQ.html.

Tayeb G, Gralak B and Enoch S (2003) Structural colors in nature and butterfly-wing modeling. *Optics and Photonics News* 14(2): 38–43.

Wyszecki G and Stiles WS (2000) *Color Science*. New York: John Wiley.

D

DETECTION

Contents

Fiber Sensors
Heterodyning
Image Post-Processing and Electronic Distribution
Smart Pixel Arrays

Fiber Sensors

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The growth in the fiber optic communications and optoelectronics industries has led to the development and commercialization of many types of optical components. Fiber sensor devices and systems are a major user of such technologies. This has resulted in the development of commercial fiber sensors that can compete with conventional sensors. Fiber sensors are capable of measuring a wide variety of stimuli including: mechanical (displacement, velocity, acceleration, strain, pressure); temperature; electromagnetic (electric field, magnetic field, current); radiation (X-ray, nuclear); chemical composition; flow and turbulence in liquids; and biomedical. The main advantages of fiber sensors are that they are low cost, compact and lightweight, robust, passive, immune to electromagnetic interference, and highly sensitive.

Fiber sensors can be grouped into two basic classes: intrinsic and extrinsic. In an intrinsic fiber sensor the sensing is carried out by the optical fiber itself. In an extrinsic sensor the fiber is simply used to carry light to and from an external optical device where the sensing takes place. Point fiber sensors are localized to discrete regions; quasi-distributed sensors utilize point sensors at various locations along a fiber and distributed sensors are capable of sensing over the entire length of fiber.

A basic fiber sensor system, as shown in Figure 1, consists of a light source, fiber sensor, optical detector, and signal processing electronics. A measurand causes some change (intensity, phase, polarization, spectrum, etc.) in the light propagating through the sensor. This change is detected and processed to give an output signal proportional to the measurand. A large part of fiber sensor research concerns the development of fiber sensors sensitive to particular measurands and appropriate signal processing techniques. We describe some of the most important fiber sensors and also consider quasi-distributed and distributed sensing, including the use of multiplexing techniques.

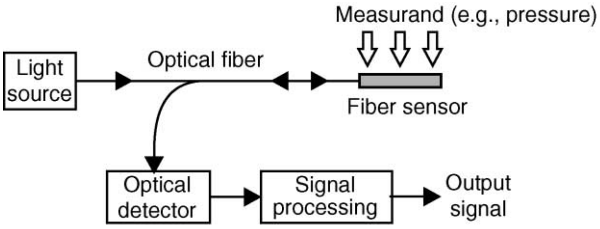

Figure 1 Basic fiber sensor system.

Intensity-Based Fiber Sensors

The intensity modulation (IM) of light is a simple method for optical sensing. There are several mechanisms that can produce a measurand-induced change in the optical intensity propagated by an optical fiber. Perhaps the simplest type of IM fiber sensor is the microbend sensor shown in Figure 2. The sensor consists of two grooved plates between which passes an optical fiber. The upper plate can move in response to pressure. When the fiber is bent sufficiently, light escapes into the fiber cladding and is lost. The greater the pressure on the plates the more loss occurs.

Coupling-based fiber sensors are useful for measurement of displacement or dynamic pressure. Transmission and reflective configurations are possible as shown in Figure 3. The transmission coupling-based sensor consists of two fibers with a small gap between them. The amount of light coupled to the second fiber depends on the fiber acceptance angles and the distance between the fibers. One of the fibers can move in response to vibration or pressure thereby changing the distance between the fibers and hence the coupling loss. The reflection-based sensor operates in a similar fashion, where light is reflected from a flexible diaphragm back into a collecting fiber. The reflected light intensity changes as the diaphragm is flexed. Once the coupling relationship between the input fiber, diaphragm and collecting fiber is known, intensity changes can be related to the applied displacement or pressure.

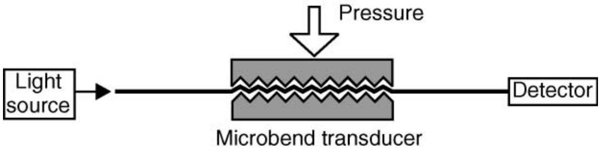

Figure 2 Microbend fiber sensor. The transducer moves in response to pressure and in doing so changes the bending radius of the fiber and thereby the fiber loss.

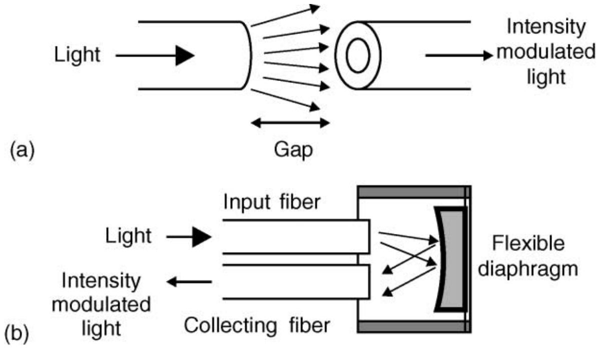

Figure 3 (a) Transmission and (b) reflective coupling-based fiber sensors.

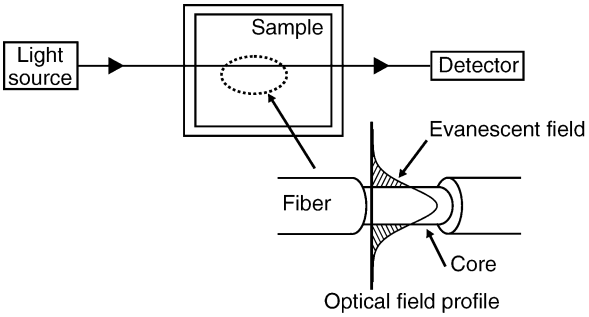

Figure 4 Evanescent wave fiber chemical sensor.

Evanescent wave fiber sensors exploit the fact that some of the energy in the guided mode of an optical fiber penetrates a short distance from the core into the cladding. The penetration of light energy into the cladding is called the evanescent wave. It is possible to design a sensor where energy is absorbed from the evanescent wave in the presence of certain chemicals as shown in Figure 4. This is achieved by stripping the cladding from a section of the fiber and using a light source having a wavelength that can be absorbed by the chemical that is to be detected. The resulting change in light intensity is a measure of the chemical concentration. Measurements can also be performed in a similar fashion by replacing the cladding with a material such as an organic dye whose optical properties can be changed by the chemical under investigation. Evanescent wave fiber sensors have found many applications in the biomedical field, such as blood component meters for detection of cholesterol and uric acid concentrations in blood.

A linear position sensor based on time division multiplexing is shown in Figure 5. It uses a square-wave modulated light source, optical delay loops and an encoded card. The delay loops separate the return signal from the encoded card by a time that is greater than the pulse duration. The encoded return signal can be decoded to determine the card position and velocity.

Interferometric Sensors

Interferometric fiber sensors operate on the principle of interference between two or more light beams to convert phase differences to intensity changes. Common configurations used include Michelson, Mach–Zehnder, and Sagnac interferometers, polarimetric systems, grating and etalon-based interferometers and ring resonators. Interferometric fiber sensors have extremely high sensitivity and are able to resolve path differences of the order of 10^{-6} of the light

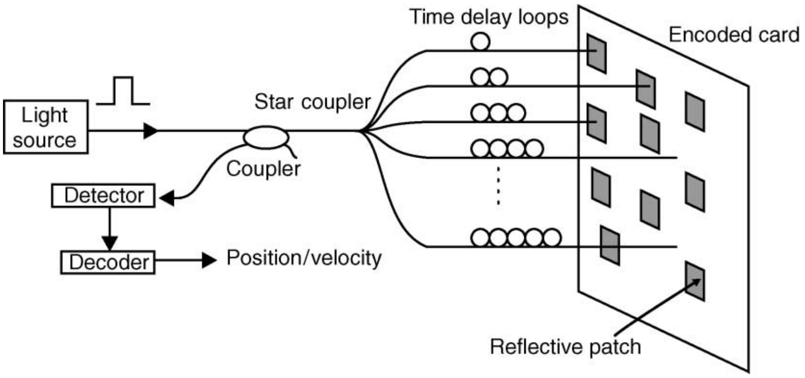
Time delay loops
Encoded card
Light source
Star coupler
Coupler
Detector
Decoder
Position/velocity
Reflective patch

Figure 5 Linear position sensor using time division multiplexing.

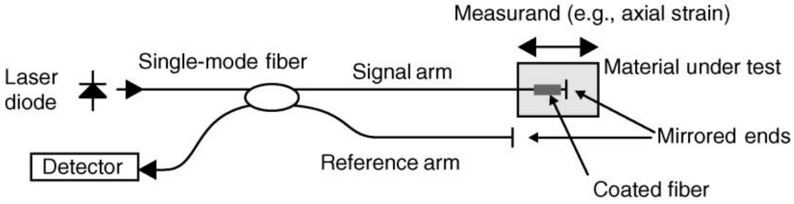
Measurand (e.g., axial strain)
Laser diode
Single-mode fiber
Signal arm
Material under test
Mirrored ends
Detector
Reference arm
Coated fiber

Figure 6 Michelson interferometric strain sensor. Part of the signal arm is embedded in a material, such as concrete (in civil engineering structures) or carbon composite (used in the aerospace industry).

source wavelength. To realize high sensitivity, spectrally pure light sources such as semiconductor lasers must be used. In addition, single mode fibers and components are used to maintain the spatial coherence of the light beams. Factors that affect the performance of interferometric sensors include the optical source phase noise and the polarization states of the interfering light beams.

In two-beam interferometric sensors, such as the Michelson interferometric strain sensor shown in Figure 6, one of the light beams travels through a fiber where its phase can be modulated by the measurand. The two reflected beams recombine on a detector, the output of which is of the form $1 + V \cos \phi$ as shown in Figure 7. ϕ is the relative phase shift between the return beams and V the interferometer visibility. Compensating techniques must be used to ensure that the interferometer sensitivity is maximized. Active techniques involve the insertion of an active device, such as a fiber stretcher, in one of the interferometer arms to control the mean phase difference between the beams. Passive schemes have the advantage in that no active components are required, but usually involve complex signal processing schemes such as active homodyne or synthetic heterodyne demodulation. The sensitivity of the fiber to the measurand can be further improved through the use of specialized coatings.

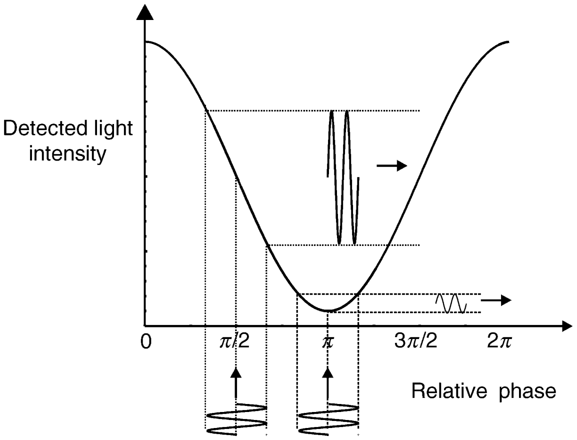

Figure 7 Two-beam interferometer transfer function. The sensitivity is maximized when the mean phase difference between the detected light beams is an odd integer multiple of $\pi/2$. When this is the case the interferometer is said to be operating in quadrature.

Compact interferometric sensors can be constructed using intrinsic or extrinsic Fabry–Perot type configurations, the most common of which are shown in Figure 8. If the Fabry–Perot cavity reflectivities are $\ll 1$, then it can be considered to be a two-beam interferometer.

Polarimetric fiber sensors relate changes induced in the polarization state of light to the measurand.

These sensors usually use high birefringence (Hi-Bi) fiber. The phase difference between the two orthogonally polarized modes of Hi-Bi fiber is given by

$$\phi = \Delta\beta l \qquad [1]$$

where $\Delta\beta$ is the fiber (linear) birefringence and l the fiber length. Both $\Delta\beta$ and l can be changed by the measurand. A polarimetric strain sensor using Hi-Bi fiber embedded in carbon fiber composite, is shown in Figure 9. Light from a linearly polarized He–Ne laser is launched into the Hi-Bi fiber, through a half-waveplate and lens. The waveplate is used to rotate the light plane of polarization so that it is at 45° to the principal axes of the fiber. This ensures that half of the input light power is coupled to each of the fast and slow modes of the Hi-Bi fiber. The output light from the Hi-Bi fiber is passed through a polarization beamsplitter. The beamsplitter separates the light into two orthogonally polarized beams, which are then detected. The output from each of the detectors is given by $V_1 = I_0 \sin^2\phi$ and $V_2 = I_0 \cos^2\phi$, respectively, where I_0 is the input light intensity. The state of polarization (SOP) of the output light from the Hi-Bi fiber is given by

$$\mathrm{SOP} = \frac{V_1 - V_2}{V_1 + V_2} = \cos^2\phi + \sin^2\phi \qquad [2]$$

It is a simple matter to calculate the SOP, from which ϕ can be determined. The form of eqn [2] is very similar to the two-beam interferometer transfer function, shown in Figure 7.

A polarimetric electric current sensor based on the Faraday effect is shown in Figure 10. The Faraday effect provides a rotation of the light's polarization state when a magnetic field is parallel to the optical path in glass. If an optical fiber is closed around a current-carrying conductor, the Faraday rotation is directly proportional to the current. By detecting the polarization rotation of light in the fiber, the current can be measured.

Linearly polarized light is equivalent to a combination of equal intensity right and left circularly polarized components. The linear state of polarization of the polarized laser light rotates in the presence of a magnetic field because the field produces

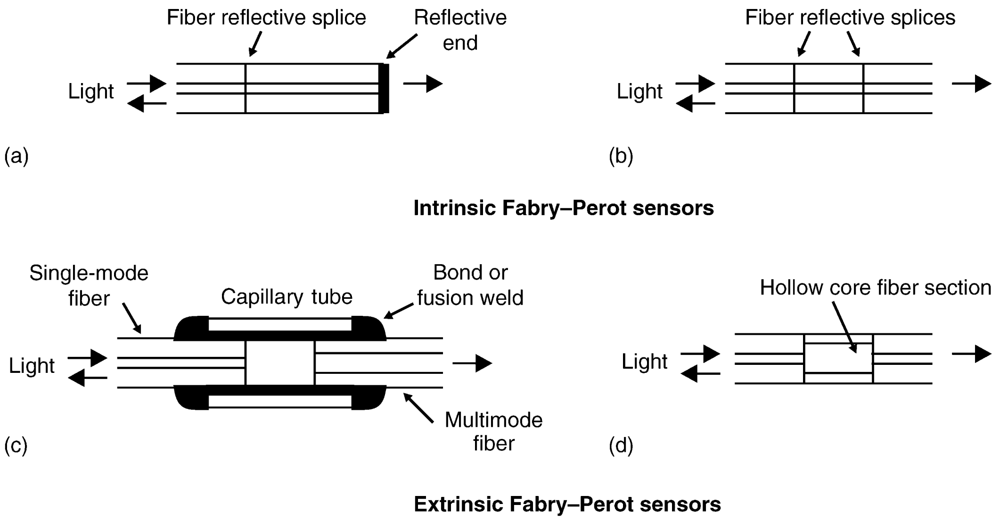
Fiber reflective splice
Reflective end
Light
(a)
Fiber reflective splices
Light
(b)
Intrinsic Fabry–Perot sensors
Single-mode fiber
Capillary tube
Bond or fusion weld
Light
Multimode fiber
(c)
Hollow core fiber section
Light
(d)
Extrinsic Fabry–Perot sensors

Figure 8 Interferometric fiber Fabry–Perot configurations.

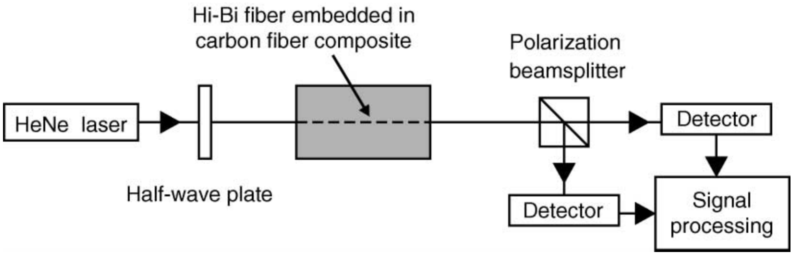
Hi-Bi fiber embedded in carbon fiber composite
Polarization beamsplitter
HeNe laser
Detector
Half-wave plate
Detector
Signal processing

Figure 9 Polarimetric strain sensor system.

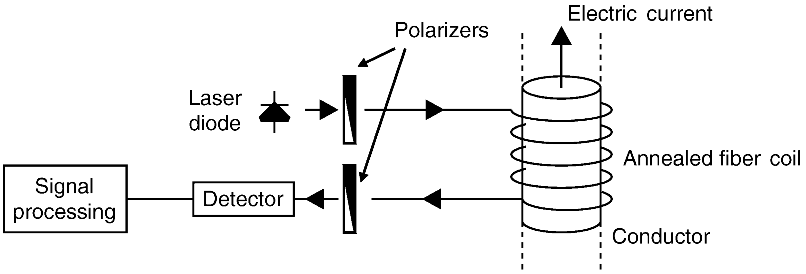

Figure 10 Polarimetric electric current sensor system.

circular birefringence in the fiber. This means that right and left circularly polarized light will travel at different speeds and accumulate a relative phase difference given by

$$\phi = VBl \qquad [3]$$

where V is the Verdet constant (rad Tesla^{-1} m^{-1}) of the glass, B the magnetic field flux density (Tesla), and l the length of the fiber exposed to the magnetic field. In the polarimetric current sensor, the polarization rotation is converted to an intensity change by the output polarizer. The output to the detector is proportional to $1 + \sin(2\phi)$, from which ϕ and the current can be determined.

The linear birefringence of conventional silica fiber is much greater than its circular birefringence. To make a practical current sensor, the linear birefringence must be removed from the fiber and this can be achieved by annealing the fiber. Annealing involves raising the temperature of the fiber, during manufacture, to a temperature above the strain point for a short period of time and slowly cooling back to room temperature. This reduces stresses in the glass, which are the principal cause of linear birefringence, and also cause physical and chemical changes to the glass. Waveguide-induced birefringence cannot be removed by annealing, but can be significantly reduced by twisting the fiber.

White light – or more accurately low-coherence – interferometry utilizes broadband sources such as LEDs and multimode lasers in interferometric measurements. Optical path differences (OPDs) are observed through changes in the interferometric fringe pattern. A processing interferometer is required in addition to the sensing interferometer to extract the fringe information.

The processing of white light interferometry signals relies on two principal techniques. The first technique involves scanning the OPD of the processing interferometer to determine regions of optical path balance. The second technique involves determination of the optical spectrum using an optical spectrum analyzer. The resulting fringe pattern fringe spacing is then related to the OPD of the sensing interferometer.

An example of the optical path scanning technique is shown in Figure 11a. The sensing interferometer is designed such that its OPD is much greater than the coherence length of the light source, so at its output no interference fringes are observed. The output of the sensing interferometer is fed to the processing interferometer. The OPD of the processing interferometer is scanned using a piezoelectric fiber stretcher driven by a suitable scanning voltage. As the processing interferometer is scanned, interference fringes are observed at two distinct points as shown in Figure 11b. The first set of fringes occurs when the OPD of the processing interferometer is within the coherence length of the optical source (close to zero). The second set of fringes occurs when the OPDs in the two interferometers are equal. The OPD in the sensing interferometer can be determined by measuring the OPD in the processing interferometer between the two positions of maximum visibility in the output signal from the detector. This, in turn, can be related to OPD changes due to the action of the measurand, in this case strain.

Sagnac interferometric sensors can be used to create highly sensitive gyroscopes that can be used to sense angular velocity (e.g., in aircraft navigation systems). It is based on the principle that the application of force (e.g., centrifugal force) will alter the wavelength of light as it travels around a coil of optical fiber. A basic open-loop fiber gyroscope is shown in Figure 12. The broadband source (e.g., superluminescent diode) is split into two counter propagating light beams traveling in the clockwise and anticlockwise directions. The polarizer is used to ensure the reciprocity of the counter propagating waves. The inherent nonlinear response of the gyroscope can be overcome by using a phase modulator and signal processing techniques. In the ideal case the detector output is proportional to $1 + \cos\phi_s$. The Sagnac phase shift ϕ_s between the two returning beams is given by

$$\phi_s = \frac{8\pi AN\Omega}{c\lambda_0} \qquad [4]$$

A is the area enclosed by a single loop of the fiber, N the number of turns, Ω the component of the angular velocity perpendicular to the plane of the loop, λ_0 the free-space wavelength of the optical source, and c the speed of light in a vacuum. Sensitivities greater than 10^{-8} rad/s and dynamic ranges in excess of 40 dB have been achieved with open-loop fiber gyroscopes. More advanced gyroscopes can greatly improve on this performance.

Fiber Grating Sensors

Many types of fiber gratings can be used in sensing applications including Bragg, long-period, and

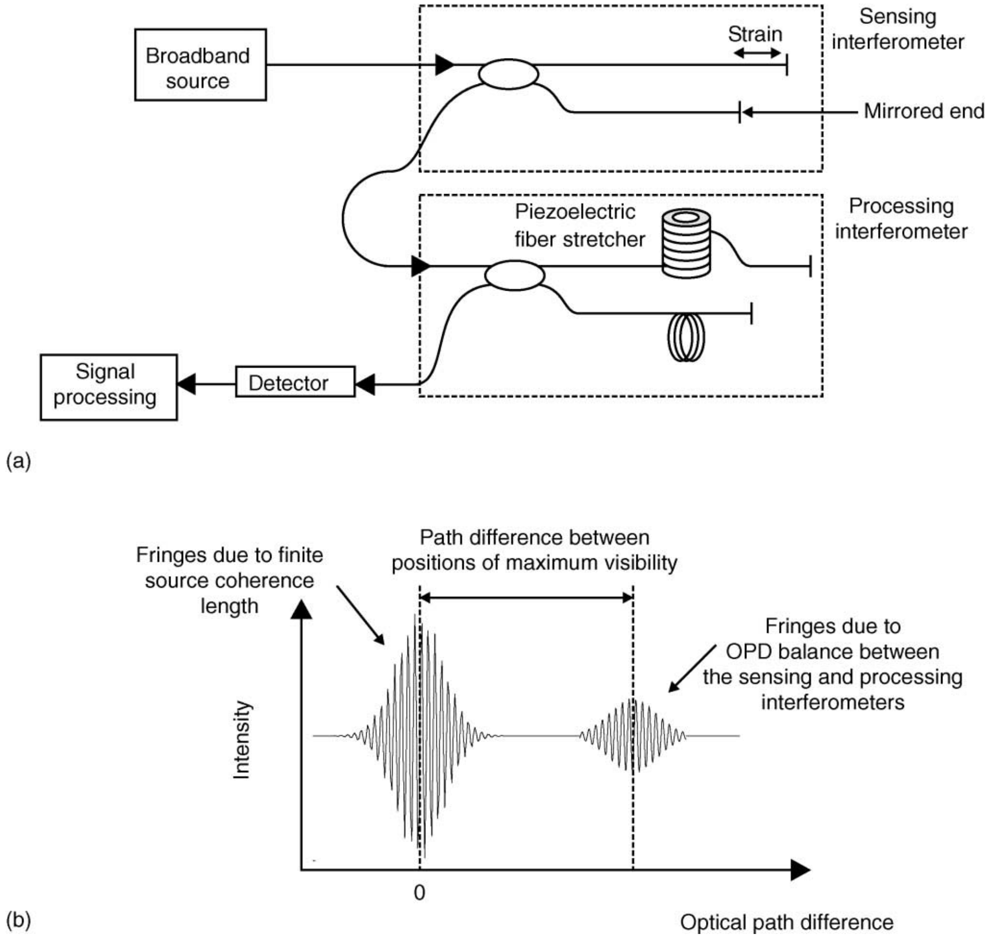
Broadband source
Strain
Sensing interferometer
Mirrored end
Piezoelectric fiber stretcher
Processing interferometer
Signal processing
Detector
(a)
Path difference between positions of maximum visibility
Fringes due to finite source coherence length
Fringes due to OPD balance between the sensing and processing interferometers
Intensity
0
(b)
Optical path difference

Figure 11 Low-coherence interferometric sensor. (a) Schematic diagram. (b) Output of the scanned processing interferometer.

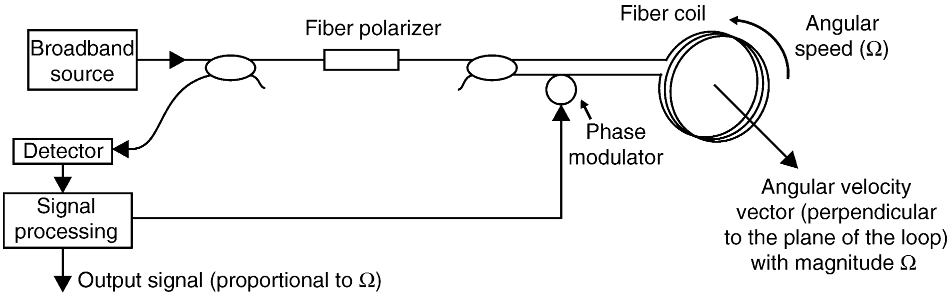
Broadband source
Fiber polarizer
Fiber coil
Angular speed (Ω)
Detector
Phase modulator
Signal processing
Angular velocity vector (perpendicular to the plane of the loop) with magnitude Ω
Output signal (proportional to Ω)

Figure 12 Basic open-loop fiber gyroscope.

chirped gratings. Because fiber gratings are small and have a narrow wavelength response, they can be used for both point and quasi-distributed sensing. They can be embedded in composite materials for smart structure monitoring and also in civil engineering structures such as bridges.

Fiber Bragg gratings (FBGs) are the most popular type of grating sensor. A typical FBG consists of a short section (typically a few mm) of single-mode fiber with a periodic modulation (typically 100s of nm) of the core refractive index. The index modulation causes light in the forward propagating core mode to be coupled into the backward core mode. This causes the FBG to act as a highly wavelength-selective rejection filter. The wavelength of peak reflectivity is the Bragg wavelength $\lambda_B = 2n_{\mathrm{eff}}\Lambda$, where n_{eff} is the effective refractive index of the guided mode in the fiber and Λ the index modulation period. Both n_{eff} and Λ can be changed by an external measurand, resulting in a shift in λ_B.

The basic principle of FBG sensors is the measurement of an induced shift in the wavelength of an optical source due to a measurand, such as strain or temperature. A basic reflective FBG sensor system is shown in Figure 13. A broadband light source is used to interrogate the grating, from which a narrowband slice is reflected. The peak wavelength of the reflected spectrum can be compared to λ_B, from which strain or temperature can be inferred. The shift in the Bragg wavelength $\Delta\lambda_B$ with applied microstrain $\Delta(\mu\varepsilon)$ and change in temperature ΔT, for silica fiber is given approximately by

$$\frac{\Delta\lambda_B}{\lambda_B} \approx 0.78\times10^{-6}\Delta(\mu\varepsilon) + 6.67\times10^{-6}\Delta T \qquad [5]$$

A wavelength resolution of ~1 pm is required (at 1.3 μm) to resolve a temperature change of 0.1 °C or a strain change of 1 $\mu\varepsilon$. The response of the grating to strain can be improved through the use of specialized coatings. Polyimide coatings are commonly used as an efficient strain transducer for gratings embedded in composite materials. Because thermal effects in materials are usually very slow, it is relatively easy to measure dynamic strain (>1 Hz). However, in structural monitoring it can be necessary to distinguish between wavelength shifts due to static strain and those due to temperature. One technique is to use two collocated gratings whose response to strain and temperature is significantly different. In addition to strain and temperature measurement, grating sensors have also been used to measure flow, vibration, electromagnetic fields and chemical effects.

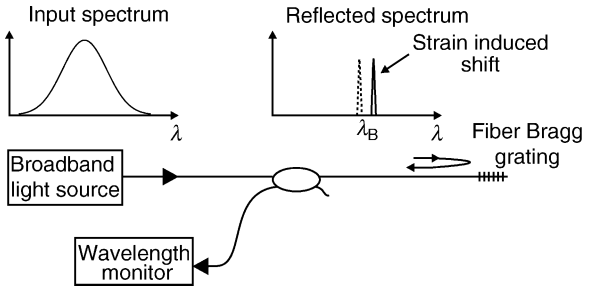

Figure 13 Basic reflective FBG sensor system.

An example of quasi-distributed strain sensing, using a wavelength division multiplexed array of FBGs, is shown in Figure 14. Each FBG in the array has a unique Bragg wavelength. The return light from the FBG array is passed through a tunable narrowband Fabry–Perot filter. As the filter is tuned, the wavelengths returned by the individual FBGs can be analyzed and the strain present at each grating determined.

FBGs can be used as narrowband reflectors for creating fiber laser sensors, capable of measuring temperature, static strain, and very high-resolution dynamic strain. The basic form of an FBG laser sensor system shown in Figure 15 consists of a doped fiber section between two identical FBGs. The doped fiber is optically pumped to provide gain and thereby enable lasing to occur. Single-mode or multi-mode lasing is possible depending on the cavity length. In single-mode operation the FBG laser linewidth can be much smaller than the linewidth of diode lasers. This means that FBG laser sensors have greater sensitivities compared to passive FBG sensors.

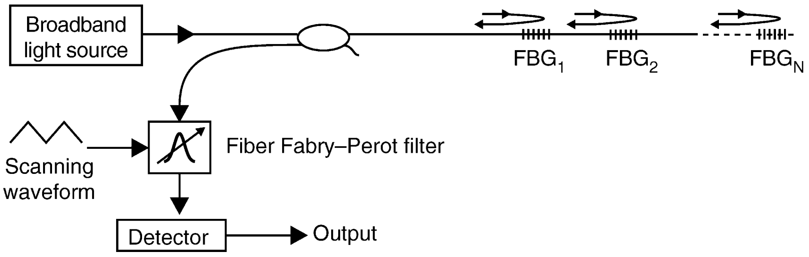

Figure 14 Quasi-distributed strain sensing using a wavelength division multiplexed array of FBGs.

When the cavity is subject to weak dynamic strain, the laser output is frequency modulated. This frequency modulation can be detected by using interferometric techniques. The main advantage of the FBG laser sensor over direct interferometry is that it is possible to obtain comparable strain sensitivities using a much shorter length of fiber.

Long-period fiber gratings (LPFGs) are attracting much interest for use in sensing applications. They are more sensitive to measurands than FBGs and easier to manufacture. A typical LPFG has a length of tens of mm with a grating period of 100s of μm. Its operation is different to an FBG in that coupling occurs between the forward propagating core mode and co-propagating cladding modes. The high attenuation of the cladding modes results in a series of minima occurring in the transmission spectrum of the fiber. This means that the spectral response is strongly influenced by the optical properties of the cladding and surrounding medium. This can be exploited for chemical sensing as shown in Figure 16, where a broadband source is used to interrogate an LPFG. The wavelength shifts of the output spectrum minima can be used to determine the concentration of particular chemicals in the substance surrounding the grating. The longest wavelength attenuation bands are the most sensitive to the refractive index of the substance surrounding the grating. This is because higher order cladding modes extend a greater distance into the external medium. LPFGs can also be used as strain, temperature, refractive index, bend, and load sensors.

Fiber Laser Doppler Velocimeter

Laser Doppler velocimetry (LDV) is a technique used for measuring velocity, especially of fluid flows. A fiber LDV system and associated scattering geometry is shown in Figure 17. In LDV two coherent laser beams intersect in a small measurement volume where they can interfere. The light reflected by a seed particle passing through the measurement volume is modulated at a frequency proportional to the spatial frequency (Doppler difference frequency Δf) of the interference fringes and the component of its velocity

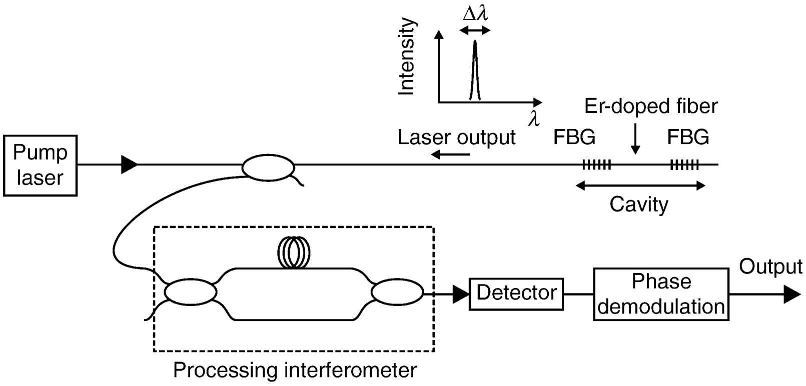

Figure 15 FBG laser sensor with interferometric detection.

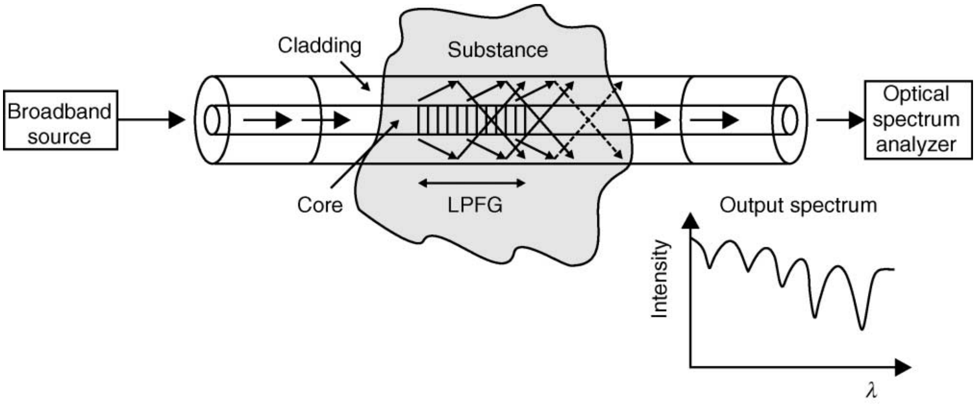

Figure 16 Detail of LPFG sensor.

normal to the interference fringes. Δf is given by

$$\Delta f = \frac{2nV\cos\beta}{\lambda_0}\sin(\theta/2) \qquad [6]$$

where n is the fluid refractive index, V the particle velocity and λ_0 the laser free-space wavelength. The output of the detector is processed to extract Δf and therefore $V\cos\beta$. Δf is independent of the scattered light direction so collection of the scattered light using a lens increases the system sensitivity. The form of eqn [6] indicates that the direction of flow cannot be ascertained. The simplest technique to resolve this ambiguity is to apply a frequency shift Δf_s to one of the input beams. This can be achieved by the use of a piezoelectric frequency shifter. The frequency shift causes a phase shift to appear between the two beams. The phase shift increases linearly with time. This results in a fringe pattern of spacing s, which moves with constant velocity $V_f = s\Delta f_s$. In this case the measured Δf will be less than or greater than Δf_s, depending on whether the particle is moving with or against the fringe motion. There will then be an unambiguous detectable range of velocities from zero to V_f.

Luminescence-Based Fiber Sensors

Luminescence-based fiber sensors are usually based on fluorescence or amplified spontaneous emission occurring in rare earth materials. They can be used in many applications such as chemical, humidity, and temperature sensing. It is possible to connect a fiber to luminescent material or to introduce luminescent dopants into the fiber. An example of the latter, used to detect chemical concentration is shown in Figure 18. A laser pulse causes the doped section of the fiber to luminesce at a longer wavelength than the

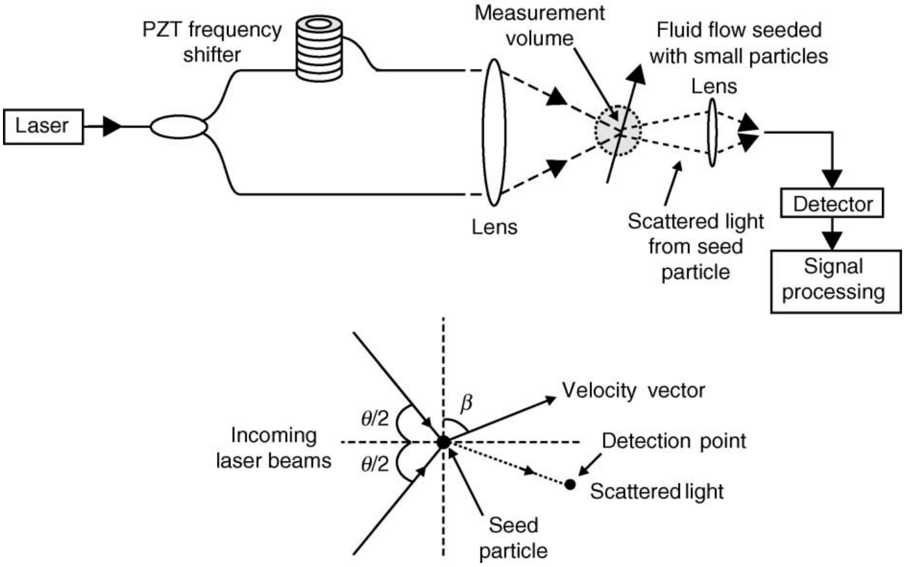
PZT frequency shifter
Measurement volume
Fluid flow seeded with small particles
Lens
Laser
Detector
Lens
Scattered light from seed particle
Signal processing
β
Velocity vector
θ/2
Incoming laser beams
θ/2
Detection point
Scattered light
Seed particle

Figure 17 Fiber Doppler velocimeter and scattering geometry.

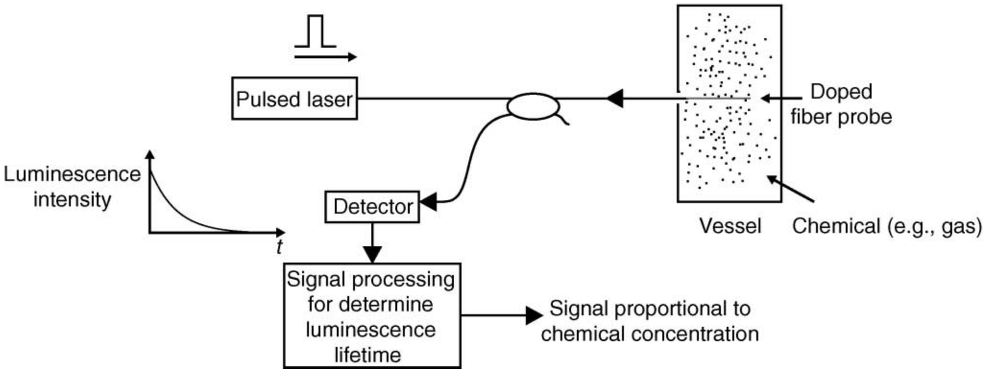
Pulsed laser
Doped fiber probe
Luminescence intensity
t
Detector
Vessel
Chemical (e.g., gas)
Signal processing for determine luminescence lifetime
Signal proportional to chemical concentration

Figure 18 Chemical sensor based on fluorescence in doped optical fiber.

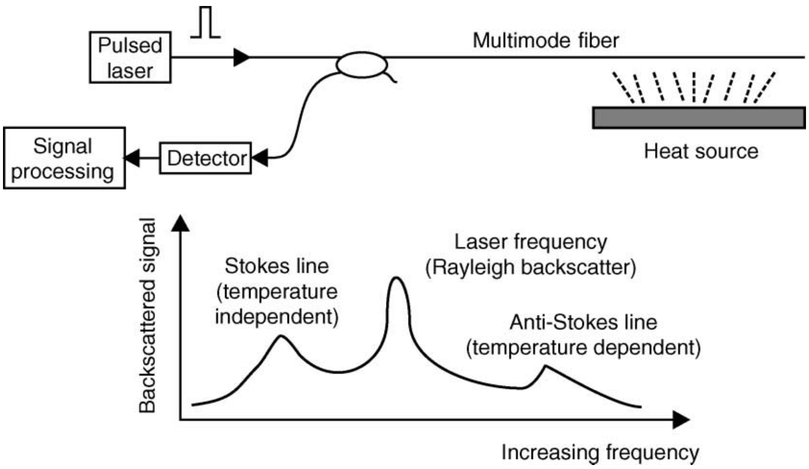
Pulsed laser
Multimode fiber
Heat source
Signal processing
Detector
Backscattered signal
Stokes line (temperature independent)
Laser frequency (Rayleigh backscatter)
Anti-Stokes line (temperature dependent)
Increasing frequency

Figure 19 Anti-Stokes Raman thermometry system.

laser. The luminescence intensity *I(t)* at the detector has an exponential decay profile given by

$$I(t) = I_0 \exp\big(-(k_1 + k_2 C)t\big) \qquad [7]$$

where I_0 is the initial luminescence intensity, t time, k_1, k_2 constants, and C the chemical concentration. The luminescence time constant $1/(k_1 + k_2 C)$ can be determined by comparing the luminescence intensity at various times after excitation by the laser pulse, from which C can be determined. The use of time division multiplexing allows quasi-distributed measurement of chemical concentration. The use of plastic optical fiber for luminescence-based sensors is attracting much interest.

Distributed Fiber Sensing

We have seen that both point and quasi-distributed sensing are possible using fiber sensors. Distributed sensing can be achieved through the use of linear or nonlinear backscattering or forward scattering techniques. In linear backscattering systems, light backscattered from a pulse propagating in an optical fiber is time resolved and analyzed to obtain the spatial distribution of the measurand field, e.g., polarization optical time domain reflectometry analyzes the polarization state of backscattered light to determine the spatial distribution of electromagnetic fields. Nonlinear backscattering schemes use effects such as Raman or Brillouin scattering. An important example of the former is the anti-Stokes Raman thermometry system shown in Figure 19. A light pulse is transmitted down the sensing fiber. Spontaneous Raman scattering causes Stokes and anti-Stokes photons to be generated along the fiber. Some of these photons travel back along the fiber to a fast detector. The intensity of the Stokes line is temperature independent. The anti-Stokes line intensity is a function of temperature. The ratio of the two intensities provides a very accurate measurement of temperature. The measurement location is determined by timing of the laser pulse.

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Fiber Gratings**. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Optical Materials:** Smart Optical Materials.

Further Reading

Culshaw B and Dakin JP (1988, 1989, 1997) *Optical Fiber Sensors*, vols. I, II, III. Norwood, MA: Artech House.

Grattan KTV and Meggitt BT (1994, 1997, 1999, 1999) *Optical Fiber Sensor Technology*, vols. I, II, III and IV. Boston, CA: Kluwer Academic.

Grattan KTV and Meggitt BT (2000) *Optical Fiber Sensor Technology*. Boston, MA: Kluwer Academic.

Kersey AD, Davis MA, Patrick HJ, *et al.* (1997) Fiber grating sensors. *Journal of Lightwave Technology* 15: 1442–1463.

Krohn DA (1999) *Fiber Optic Sensors: Fundamentals and Applications*, 3rd edn. Research Triangle Park, NC: Instrumentation Society of America.

Lopez-Higuera JM (2001) *Handbook of Optical Fiber Sensing Technology*. New York: Wiley.

Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.

Udd E (1991) *Fiber Optic Sensors: An Introduction for Engineers and Scientists*. New York: Wiley.

Udd E (1995) *Fiber Optic Smart Structures*. New York: Wiley.

Yu FTS and Yin S (2002) *Fiber Optic Sensors*. New York: Marcel Dekker.

Heterodyning

T-C Poon, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Heterodyning, also known as frequency mixing, is a frequency translation process. Heterodyning has its root in radio engineering. The principle of heterodyning was discovered in the late 1910s by radio engineers experimenting with radio vacuum tubes. Russian cellist and electronic engineer Leon Theremin invented the so-called Thereminvox, which was one of the earliest electronic musical instruments that generates an audio signal by combining two different radio signals. American electrical engineer Edwin Armstrong invented the so-called super-heterodyne receiver. The receiver shifts the spectrum of the modulated signal, that is, the frequency contents of the modulated signal, so as to demodulate the audio information from it. Commercial amplitude modulation (AM) broadcast receivers nowadays are super-heterodyne type. After illustrating the basic principle of heterodyning by some simple mathematics, we will discuss some examples on how the principle of heterodyning is used in radio and in optics.

For a simple case, when signals of two different frequencies are heterodyned or mixed, the resulting signal produces two new frequencies, the sum and difference of the two original frequencies. Figure 1 illustrates how signals of two frequencies are mixed or heterodyned to produce two new frequencies, $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$, by simply multiplying the two signals $\cos(\omega_1 t + \theta_1)$ and $\cos(\omega_2 t)$, where ω_1 and ω_2 are radian frequencies of the two signals and θ_1 is the phase angle between the two signals. Note that when the frequencies of the two signals to be heterodyned are the same, i.e., $\omega_1 = \omega_2$, the phase information of $\cos(\omega_1 t + \theta_1)$ can be extracted to get $\cos\theta_1$ if we use an electronic lowpass filter (LPF) to filter out the term $\cos(2\omega_1 t + \theta_1)$. This is shown in Figure 2. Heterodyning is often referred to as homodyning for the mixing of two signals of the same frequency.

For a general case of heterodyning, we can have a signal represented by $s(t)$ with its spectrum $S(\omega)$, which is given by the Fourier transform of $s(t)$, and when it is multiplied by $\cos(\omega_2 t)$, the resulting spectrum is frequency-shifted to new locations in the frequency domain as:

$$\mathcal{F}\{s(t)\cos(\omega_2 t)\} = \frac{1}{2}S(\omega - \omega_2) + \frac{1}{2}S(\omega + \omega_2) \quad [1]$$

where $\mathcal{F}\{s(t)\} = S(\omega)$ and $\mathcal{F}\{\cdot\}$ denotes the Fourier transform of the quantity being bracketed. The spectrum of $s(t)\cos(\omega_2 t)$, along with the spectrum of $s(t)$, is illustrated in Figure 3. It is clear that multiplying $s(t)$ with $\cos(\omega_2 t)$ is a process of heterodyning, as we have translated the spectrum of the signal $s(t)$. This process is known as modulation in communication systems.

In order to appreciate the process of heterodyning let us now consider, for example, heterodyning in radio. In particular, we consider AM. While the frequency content of an audio signal (from 0 to around 3.5 kHz) is suitable to be transmitted over a pair of wires or coaxial cables, it is, however, difficult to be transmitted in air. By impressing the audio information onto a higher frequency, say 550 kHz, as one of the broadcast bands in AM radio, i.e., by

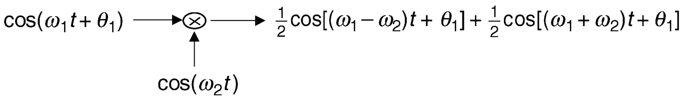
$\cos(\omega_1 t + \theta_1)$
$\frac{1}{2}\cos[(\omega_1 - \omega_2)t + \theta_1] + \frac{1}{2}\cos[(\omega_1 + \omega_2)t + \theta_1]$
$\cos(\omega_2 t)$

Figure 1 Heterodyning of two sinusoidal signals.

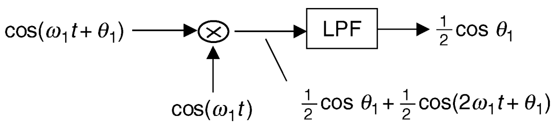
$\cos(\omega_1 t + \theta_1)$
LPF
$\frac{1}{2}\cos\theta_1$
$\cos(\omega_1 t)$
$\frac{1}{2}\cos\theta_1 + \frac{1}{2}\cos(2\omega_1 t + \theta_1)$

Figure 2 Heterodyning becomes homodyning when the two frequencies to be mixed are the same: homodyning allows the extraction of the phase information of the signal.

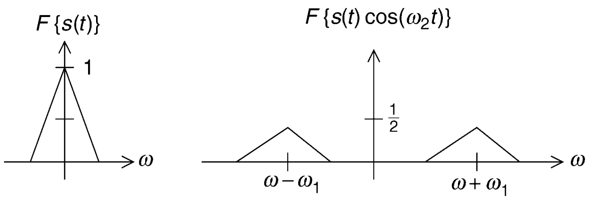

Figure 3 Spectrum of $s(t)$ and $s(t)\cos(\omega_2 t)$. It is clear that spectrum of $s(t)$ has been translated to new locations upon multiplying a sinusoidal signal.

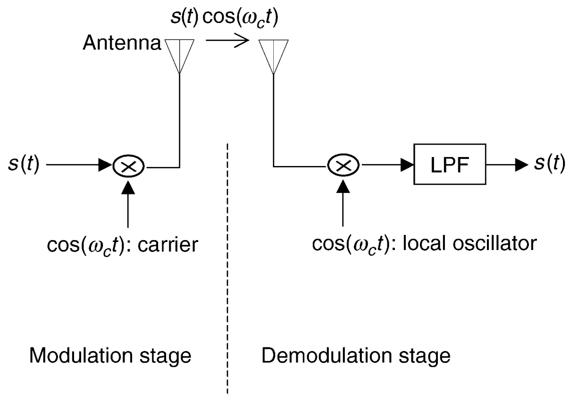

Figure 4 Radio link: heterodyning in the demodulation stage is often known as heterodyne detection.

modulating the audio signal, the resulting modulated signal can now be transmitted using antennas of reasonable dimensions. In order to recover (demodulate) the audio signal from the modulated signal, the modulated signal is first received by an antenna in a radio receiver, multiplied by the so-called local oscillator with the same frequency as the signal used to modulate the audio signal, then followed by a lowpass filter to eventually obtain the audio information back. The situation is illustrated in Figure 4. We shall now describe the process mathematically with reference to Figure 4.

We denote $s(t)$ as the audio signal. After multiplying by $\cos(\omega_c t)$, which is usually called a carrier in radio in the modulation stage, where ω_c is the radian frequency of the carrier. The modulated signal $s(t)\cos(\omega_c t)$. is transmitted via an antenna. When the modulated signal is received in the demodulation stage, the modulated signal is then multiplied by the local oscillator of signal waveform $\cos(\omega_c t)$. The output of the multiplier is given by

$$s(t)\cos(\omega_c t)\cos(\omega_c t) = s(t)\tfrac{1}{2}[1+\cos(2\omega_c t)] \qquad [2]$$

Note that the two signals in the demodulation stage, $s(t)\cos(\omega_c t)$ and $\cos(\omega_c t)$, are heterodyned to produce two new frequencies, the sum $\omega_c + \omega_c = 2\omega_c$ to give the term $\cos(2\omega_c t)$ and the difference $\omega_c - \omega_c = 0$ to give the term $\cos(0) = 1$. Since the two frequencies to be multiplied in the demodulation stage are the same, this is homodyning as explained above. Now by performing lowpass filtering (LPF), we can recover our original audio information $s(t)$. Note that if the frequency of the local oscillator in the demodulation stage within the receiver is higher than the frequency of the carrier used in the modulation stage, heterodyning in the receiver is referred to as superheterodyning, which most receivers nowadays use for amplitude modulation. In general, we have two heterodyning processes in the radio system just described, one in the modulation stage and the other in the demodulation stage. However, it is unusual to speak of heterodyning in the modulation stage, and so we just refer to the process in the demodulation stage as 'heterodyne detection.'

In summary, heterodyne detection can extract the information from a modulated signal and can also extract the phase information of a signal if homodyning is used. We shall see, in the next section, how optical heterodyne detection is employed.

Optical information is usually carried by coherent light such as a laser. Let $\psi_p(x, y)$ be a complex amplitude of the light field, which may physically represent a component of the electric field. We further assume that the light field is oscillating at temporal frequency ω_0. Therefore, we can write the light field as $\psi_p \exp(i\omega_0 t)$. By taking the real part of $\psi_p \exp(i\omega_0 t)$, i.e., $\mathrm{Re}[\psi_p \exp(i\omega_0 t)]$, we recover the usual physical real quantity. For a simple example, if we let $\psi_p = A\exp(-ik_0 z)$, where A is some constant and k_0 is the wavenumber of the light, $\mathrm{Re}[\psi_p \exp(i\omega_0 t)] = \mathrm{Re}[A\exp(-ik_0 z)\exp(i\omega_0 t)] = A\cos(\omega_0 t - k_0 z)$, which is a plane wave propagating along the positive z direction in free space. To detect light energy, we use a photodetector (PD), as shown in Figure 5. Assuming a plane wave for simplicity, as $\psi_p = A$, we have also taken $z = 0$ at the surface of the photodetector. As the photodetector responds to intensity, i.e., $|\psi_p|^2$, which gives the current, i, as output by spatially integrating the intensity:

$$i \propto \int_S |\psi_p \exp(i\omega_0 t)|^2 \mathrm{d}x\mathrm{d}y = A^2 S \qquad [3]$$

where S is the surface area of the photodetector. We can see that the photodetector current is proportional to the intensity, A^2, of the incident light. Hence the output current varies according to the intensity of the

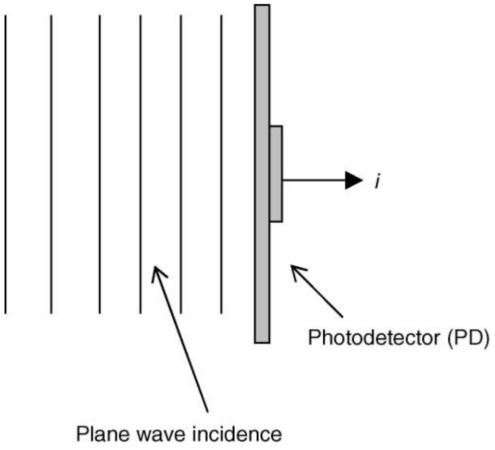

Figure 5 Optical direction detection or optical incoherent detection.

optical signal intensity. This mode of photodetection is called direct detection or incoherent detection in optics.

Let us now consider the heterodyning of two plane waves on the surface of the photodetector. We assume an information-carrying plane wave, also called the signal plane wave, $A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)] \times \exp(-ik_0 x \sin\phi)\}$, and a reference plane wave, $A_r \exp(i\omega_0 t)$, or called a local oscillator in radio. The situation is shown in Figure 6.

Note that the frequency of the signal plane wave is ω_s higher than that of the reference signal, and it is inclined at an angle ϕ with respect to the reference plane wave, which is normal incident to the photodetector. Also, the information content $s(t)$ is in the phase of the signal wave. In the situation shown in Figure 6, we see that the two plane waves are interfered on the surface of the photodetector, giving the total light field ψ_t, given by

$$\psi_t = A_r \exp(i\omega_0 t) + A_s \exp\{i[(\omega_0 + \omega_s)t + s(t)]\} \times \exp(-ik_0 x \sin\phi) \qquad [4]$$

Again, the photodetector responds to intensity, giving the current

$$\begin{aligned} i &\propto \int_S |\psi_t|^2 \mathrm{d}x\mathrm{d}y \\ &= \int_{-a}^{a}\int_{-a}^{a} [A_r^2 + A_s^2 + 2A_r A_s \cos(\omega_s t + s(t) \\ &\quad - k_0 x \sin\phi)]\mathrm{d}x\mathrm{d}y \end{aligned} \qquad [5]$$

where we have assumed that the photodetector has a $2a \times 2a$ square area. The current can be evaluated to be:

$$i(t) \propto 2a(A_r^2 + A_s^2) + 4A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \qquad [6]$$

The current output has two parts: the DC current and the AC current. The AC current at frequency ω_s is commonly known as the heterodyne current. Note that the information content $s(t)$ originally embedded in the phase of the signal plane wave has now been preserved and transferred to the phase of the heterodyne current. The above process is called optical heterodyning. In optical communications, it is often referred to as optical coherent detection. In contrast, if the reference plane wave has not been used for the detection, we have the incoherent detection. The information content carried by the signal plane wave would be lost, as it is evident that for $A_r = 0$, the above equation gives only a DC current at a value proportional to the intensity of the plane wave, A_s^2.

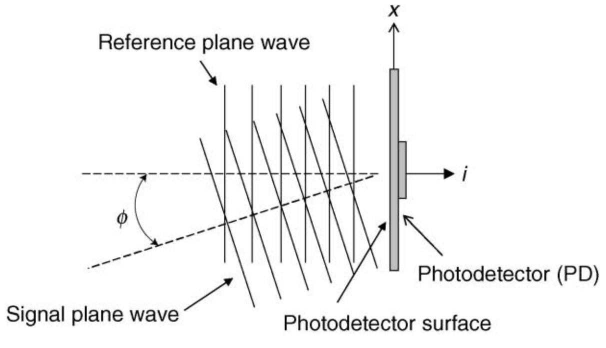

Figure 6 Optical heterodyne detection.

Let us now consider some of the practical issues encountered in coherent detection. Again, the AC part of the current given by the above equation is the heterodyne current $i_{het}(t)$, given by

$$i_{het}(t) \propto A_r A_s \frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} \cos(\omega_s t + s(t)) \qquad [7]$$

We see that since the two plane waves propagate in slightly different directions, the heterodyne current output is degraded by a factor of

$$\frac{\sin(k_0 a \sin\phi)}{k_0 \sin\phi} = a\mathrm{sinc}(k_0 a \sin\phi)$$

where $\mathrm{sinc}(x) = \sin(x)/x$. For small angles, i.e., $\sin\phi \approx \phi$, the current amplitude falls off as $\mathrm{sinc}(k_0 a\phi)$. Hence, the heterodyne current is at a maximum when the angular separation between the signal plane wave and the reference plane wave is zero, i.e., the two plane waves are propagating exactly in the same direction. The current will go to zero when $k_0 a\phi = \pi$, or $\phi = \lambda_0/2a$, where λ_0 is the wavelength of the laser light. To see how critical it is for the angle ϕ to be aligned in order to have any heterodyne output, we assume the size of the photodetector $2a = 1$ cm and the laser used is red, i.e., $\lambda_0 \approx 0.6$ μm; ϕ is calculated to be about 2.3×10^{-3} degrees. Hence to be able to work with coherent detection, we need to have precise optomechanical mounts for angular rotation.

Finally we describe a famous application of heterodyning in optics – holography. As we have seen, the mixing of two light fields with different temporal frequencies will produce heterodyne current at the output of the photodetector. But we can record the two spatial light fields of the same temporal frequency with photographic films instead of using electrical devices. Photographic films respond to light intensity as well. We discuss holographic recording of a point source object as a simple example. Figure 7 shows a collimated laser split into two plane waves and recombined by the use of two mirrors (M) and two beamsplitters (BS). One plane wave is used to illuminate the pinhole aperture (our point source object), and the other

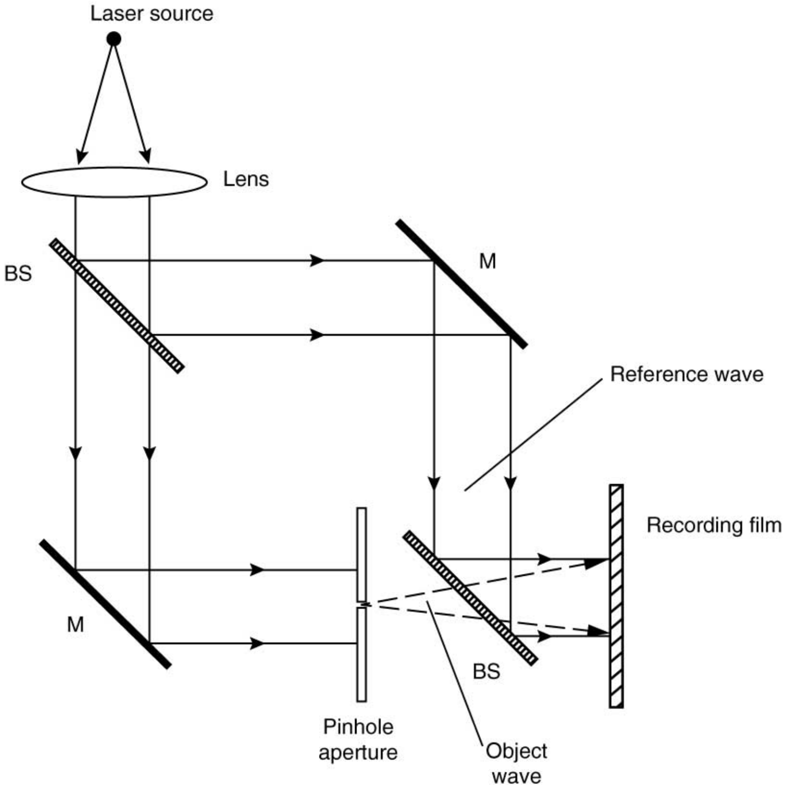

Figure 7 Holographic recording of a point source object.

illuminates the recording film directly. The plane wave that is scattered by the point source object, which is located z_0 away from the film, generates a diverging spherical wave. This diverging wave is known as an object wave in holography. The object wave arising from the point source object on the film is given by, according to Fresnel diffraction:

$$\psi_0 = A_0 \exp(-ik_0 z_0)\ \frac{ik_0}{2\pi z_0} \times \exp[-ik_0(x^2 + y)^2/2z_0]\exp(i\omega_0 t) \qquad [8]$$

This object wave is a spherical wave, where A_0 is the amplitude of the spherical wave.

The plane wave that directly illuminates the photographic plate is known as a reference wave, ψ_r. For the reference plane wave, we assume that the plane wave has the same phase with the point source object at a distance z_0 away from the film. Its field distribution on the film is, therefore, $\psi_\mathrm{r} = A_r \exp(-ik_0 z_0)\exp(i\omega_0 t)$, where A_r is the amplitude of the plane wave. The film now records the interference of the reference wave and the object wave, i.e., what is recorded on the film as a 2D pattern is given by $t(x, y) \propto |\psi_\mathrm{r} + \psi_0|^2$. The resulting recorded 2D pattern, $t(x, y)$, is called the hologram of the object. This kind of recording is known as holographic recording, distinct from a photographic recording in which the reference wave does not exist and hence only the object wave is recorded. Now:

$$\begin{aligned} t(x,y) &\propto |\psi_\mathrm{r} + \psi_0|^2 \\ &= \left| A_r \exp(-ik_0 z_0)\exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \right. \\ &\quad \left. \times \frac{ik_0}{2\pi z_0}\exp[-ik_0(x^2+y)^2/2z_0]\exp(i\omega_0 t)\right|^2 \\ &= A + B\sin\left\{\frac{k_0}{2z_0}[(x^2+y)^2/2z_0]\right\} \end{aligned} \qquad [9]$$

where A and B are some inessential constants. Note that the result of recording two spatial light fields of the same temporal frequency preserves the phase information (noticeably the depth parameter z_0) of the object wave. This is considered optical homodyning as it is clear from the result shown in Figure 2.

The intensity distribution being recorded on the film, upon being developed, will have transmittance given by the above equation. This expression is called the Fresnel zone plate, which is the hologram of a point source object and we shall call it the point-object hologram. Figure 8 shows the hologram for a particular value of z_0 and k_0. The importance of this phase-preserving recording is that when we illuminate the hologram with a plane wave, called the reconstruction wave, a point object is reconstructed at the same location of the original point source object if we look towards the hologram as shown in Figure 9, that is the point source is reconstructed at a distance z_0 from the hologram as if it were at the same distance away from the recording film. For an arbitrary 3D object, we can imagine the object as a collection of points, and therefore, we can see that we have a collection of Fresnel zone plates on the hologram. Upon reconstruction, such as the illumination of the hologram by a plane wave, the observer would see a 3D virtual object behind a hologram.

As a final example, we discuss a state-of-the-art holographic recording technique called optical scanning holography, which employs the use of optical heterodyning and electronic homodyning to achieve real-time holographic recording without the use of films. We will take the holographic recording of a point object as an example. Suppose we superimpose a plane wave and a spherical wave of different temporal frequencies and use the resulting intensity pattern as an optical scanning beam to raster scan a point object located z_0 away from the point source that generates the spherical wave. The situation is shown in Figure 10. Point C is the point source that generates the spherical wave (shown with dashed lines) on the pinhole object, our point object. This point source, for example, can be generated by a focusing lens. The solid parallel rays represent the plane wave. The interference of the two waves generates a Fresnel zone plate type pattern on the pinhole object:

$$
\begin{aligned}
I_s(x, y;\ z_0, t) \\
&= \left| A_r \exp(-ik_0 z_0) \exp(i\omega_0 t) + A_0 \exp(-ik_0 z_0) \frac{ik_0}{2\pi z_0} \right. \\
&\quad \left. \times \exp[-ik_0 (x^2 + y)^2 / 2z_0] \exp[i(\omega_0 + \Omega)t] \right|^2 \\
&= 1 + \left(\frac{1}{\lambda_0 z_0} \right)^2 + \frac{1}{\lambda_0 z_0} \sin\left[\frac{\pi}{\lambda_0 z_0} (x^2 + y^2) - \Omega t \right]
\end{aligned}
\qquad [10]
$$

assuming $A_r = A_0 = 1$ for simplicity. Note that the plane wave is at temporal frequency ω_0, and the spherical wave is at temporal frequency $\omega_0 + \Omega$. The expression $I_s(x, y;\ z_0, t)$ is a temporally modulated Fresnel zone plate and is known as the time-dependent Fresnel zone plate (TDFZP). If we freeze time, say at $t = t_0$, we have the Fresnel zone plate pattern on the pinhole object as shown in Figure 10. However, if we let the time run in the above expression, physically we will have running zones that would be moving away from the center of the pattern. These running zones are the result of optical heterodyning of the plane wave and the spherical wave of different temporal frequencies. It is this TDFZP that is used to raster scan a 3D object to obtain holographic information of the scanned object and such technique is called optical scanning holography.

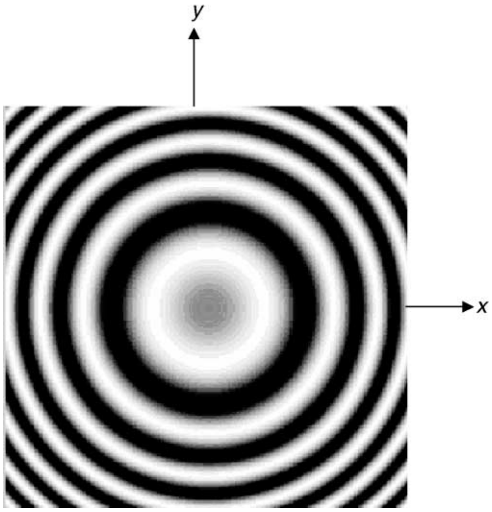

Figure 8 Point-object hologram.

Upon scanning by the TDFZP, the photodetector, captures all the transmitted light and delivers a current, which consists of a heterodyne current at

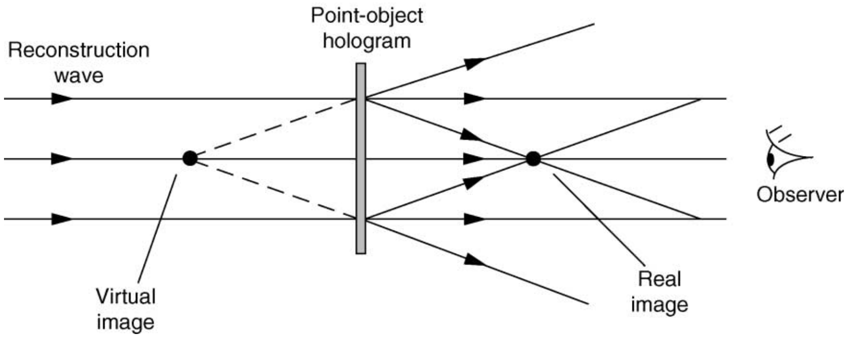

Figure 9 Reconstruction of a point-object hologram.

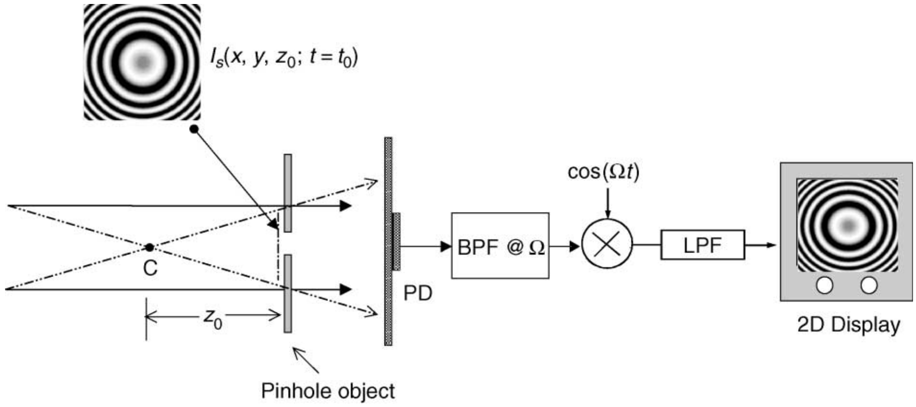

Figure 10 Optical scanning holography (use of optical heterodyning and electronic homodyning to record holographic information upon scanning the object in two dimensions).

frequency Ω. After electronic bandpass filtering (BPF) at Ω, the heterodyne current is homodyned electronically by cos(Ωt) and lowpass filtered (LPF) to extract the phase information of the current. When this final scanned and processed current is display in a 2D display, as shown in Figure 10, we have the Fresnel zone plate, which is the hologram of the pinhole object, on the display. We can photograph this 2D display to obtain a transparency and have it illuminated by a plane wave to have the reconstruction as shown in Figure 9, or, since the hologram is now in electronic form, we can store it in a PC and reconstruct it digitally. When holographic reconstruction is performed digitally, we have so-called digital holography.

See also

Diffraction: Fresnel Diffraction. **Holography, Techniques:** Digital Holography.

Further Reading

Banerjee PP and Poon TC (1991) *Principles of Applied Optics*. MA: Irwin.

Palais JC (1988) *Fiber Optic Communications*. New Jersey: Prentice-Hall.

Poon TC and Banerjee PP (2001) *Contemporary Optical Image Processing with MATLAB*. Oxford: Elsevier.

Poon TC and Qi Y (2003) Novel real-time joint-transform correlation using acousto-optic heterodyning. *Applied Optics* 42: 4663–4669.

Poon TC, Wu MH, Shinoda K and Suzuki Y (1996) Optical scanning holography. *Proc. IEEE* 84: 753–764.

Pratt WK (1969) *Laser Communication Systems*. New York: Wiley.

Stremler FG (1990) *Introduction to Communication Systems*. MA: Addison-Wesley.

VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Image Post-Processing and Electronic Distribution

K M Iftekharuddin, The University of Memphis, Memphis, TN, USA

F Ahmed, The Catholic University of America, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The spectacular development of imaging sensors, communication, and internet infrastructure has immensely facilitated the creation, processing, and distribution of electronic digital images. One nontrivial aspect of the digital image is its complete application dependence. The image may be created using a number of different acquisition methods and sensors. It is then processed in different ways, depending upon myriad of digital imaging applications. Consequently, the effectiveness of a useful electronic image processing engine may involve development of robust pre-processing techniques, different types of goal-directed processing such as detection, recognition, identification, classification, tracking, reconstruction and/or registration, as well as post-processing algorithms for different application areas such as multimedia, biomedical, astronomical, defense, consumer, industrial, etc. A notional schematic of this electronic image processing cycle is shown in Figure 1. Note that while the goal-directed techniques may be different, based on problems at hand, the pre- and post-processing techniques and tools are very similar. Thus, most of the common image pre-processing techniques are also applicable for image post-processing purposes.

Electronic image post-processing techniques involve three broad domains such as: (i) spatial; (ii) frequency; and (iii) time-frequency. Each of these domain techniques may be grouped into different types of processing such as: (i) spatial domain technique to include filtering, spatial processing such as histogram modification, morphological processing, texture processing, etc.; (ii) frequency domain technique to include filter; and (iii) time-frequency domain technique to include short-time Fourier transform, wavelet, and Wigner-ville transforms.

On the other hand, it is useful to identify the typical lifecycle of a digital image from the distribution perspective as shown in Figure 2. The image may be created using a number of different acquisition methods and sensors as mentioned above. The processing block next combines all different types of processing, as mentioned in Figure 1. The image may then be transmitted over the internet or it can leave the digital world to go to the print world. It may then go through wear/tear, subsequently being scanned back to digital domain again. After some processing, it may again be distributed over the internet.

Consequently, the image may be copied legally or fraudulently, processed, and used by a user. Further, the threats and vulnerabilities of the internet have also increased proportionately. The rapid development and availability of high-end computers, printers, and scanners has made the tools of counterfeiting easily accessible and more affordable. With these products, one may counterfeit digital images, which are easy to distribute and modify. As a result, copyright ownership poses a formidable challenge to image post-processing and distribution.

In the subsequent sections, we briefly review relevant background in image post-processing and distribution techniques. We also present a few representative application examples of image post-processing and copyright protection in image distribution.

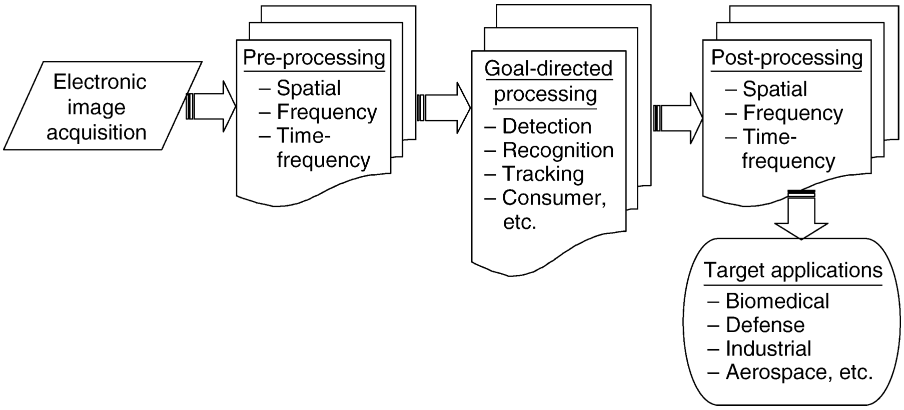
Electronic image acquisition
Pre-processing
– Spatial
– Frequency
– Time-frequency
Goal-directed processing
– Detection
– Recognition
– Tracking
– Consumer, etc.
Post-processing
– Spatial
– Frequency
– Time-frequency
Target applications
– Biomedical
– Defense
– Industrial
– Aerospace, etc.

Figure 1 Notional schematic of electronic image processing.

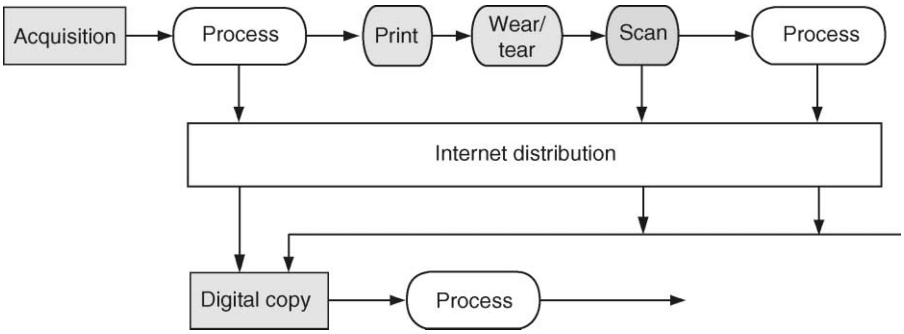
Acquisition
Process
Print
Wear/tear
Scan
Process
Internet distribution
Digital copy
Process

Figure 2 Typical lifecycle of a digital image distribution cycle.

Techniques in Image Post-Processing

Image Formation

We begin by describing the image formation technique briefly. Consider that a three-dimensional (3D) object or scene is imaged onto a 2D imaging plane by means of a recording system such as a camera. If the image formation process is linear, the recorded image may be modeled as the output of the system shown in Figure 3, which is given mathematically by

$$g(x, y) = s\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y; s, t) f(s, t) \mathrm{d}s \, \mathrm{d}t\right\} + n(x, y) \qquad [1]$$

where $g(x, y)$ denotes the recorded output image, and $f(x, y)$ is the ideal electronic image. The function $h(x, y; s, t)$ is the 2D impulse response or the point-spread function (PSF) of the image formation system. The PSF determines the radiant energy distribution in the image plane due to a point source located in the object plane. Usually the PSF is normalized such that

$$\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y; s, t) \mathrm{d}s \, \mathrm{d}t\right\} = 1 \qquad [2]$$

The noise contribution $n(x, y)$ is shown as an additive random process that is statistically uncorrelated with the image. This is a simplification because noises such as film-grain noise and the noise caused by photon statistics, which often corrupt images, are not uncorrelated with the input. This simplification nonetheless leads to reasonable and useful results. One of the primary goals of image pre-processing (and post-processing as well) is to devise techniques that reduce this additive random noise.

The PSF is a function of only the argument differences $x - s$ and $y - t$ for stationary image and object fields. Thus, it is possible to envision an image formation system that is space invariant; hence the PSF is independent of position. In this case, we may rewrite the superposition integral in eqn [1] into a more familiar convolution integral as follows:

$$g(x, y) = \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x - s, y - t) f(s, t) \mathrm{d}s \, \mathrm{d}t\right\} + n(x, y) \qquad [3]$$

Equivalently, eqn [3] is written as

$$g(x, y) = \{h(x, y)^{*} f(x, y)\} + n(x, y) \qquad [4]$$

where ($*$) is used to denote 2D convolution. Converting from a continuous image $f(x, y)$ to its discrete representation $f(i, j)$ requires the process of sampling. In the ideal sampling system, $f(x, y)$ is multiplied by an ideal 2D impulse train given as

$$f(i,j) = \sum_{m=-\infty}^{+\infty}\sum_{m=-\infty}^{+\infty} f(mX_0, nY_0)\,\delta(x - mX_0, y - Y_0) \qquad [5]$$

where X_0 and Y_0 are the sampling distances or intervals, $\delta(^*,^*)$ is the ideal impulse function. Note that the square sampling implies that $X_0 = Y_0$. Sampling with an impulse function corresponds to sampling with an infinitesimally small point. If one is interested in image processing, one should choose a sampling density based upon the classical Nyquist sampling theory. Thus, in discrete domain implementation, eqns [3] and [4] are expressed as

$$g(i,j) = \left\{\sum_{\forall(k,l)} h(i,j;k,l) f(k,l)\right\} + n(i,j), \; 0 \le i, j \le N - 1 \qquad [6]$$

and,

$$g(i,j) = s\left\{\sum_{\forall(k,l)} h(i - k, j - l) f(k,l)\right\} + n(i,j) \qquad [7]$$

Thus, we may rewrite eqn [6] in discrete domain as follows:

$$g(i,j) = h(i,j)^{*} f(i,j) + n(i,j) \qquad [8]$$

where ($*$) represents convolution operation. Equation [8] also describes the spatial domain filtering wherein $h(x,y;s,t)$ may be considered as the spatial domain filter mask. The discrete Fourier transform may be used to yield the frequency domain model as follows:

$$G(u,v) = H(u,v)F(u,v) + N(u,v) \qquad [9]$$

where G, H, F, and N are corresponding Fourier transforms and u and v are the dummy variables. In eqn [9], $H(u,v)$ is again the frequency domain filter.

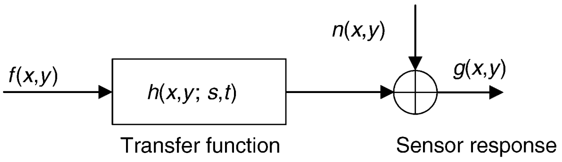

Figure 3 Model for image formation and recording.

We discuss the frequency domain filtering concept further, later in this article.

Spatial Domain Processing

The image pixel domain operations are frequently described as the spatial domain processing. These techniques are important for image pre- as well as post-processing operations. We describe two representative spatial domain processing techniques such as histogram processing and fractal processing for texture analysis below.

Histogram processing

One of the important issues in post-processing is image enhancement. Distortions in image may be introduced due to different noise factors, relative motion between a camera and the object, a camera that is out of focus, or atmospheric turbulence. Noise may be due to thermal measurement errors, recording medium, transmission medium, or digitization process. The histogram processing operation is one of the simplest techniques for enhancing the distorted image for certain type of noises. Histogram equalization is a contrast enhancement technique with the objective to obtain a new enhanced image with a uniform histogram. This can be achieved by using the normalized cumulative histogram as the gray scale mapping function. Histogram modeling is usually introduced using continuous domain functions. Consider that the images of interest contain continuous intensity levels in the interval [0,1] and that the transformation function τ which maps an input image $f(x, y)$ onto an output image $w(x, y)$ is continuous within this interval. We assume that the transfer function, $K_\mathrm{f} = \tau(K_\mathrm{w})$, is single-valued and monotonically increasing such that it is possible to define the inverse law $K_\mathrm{w} = \tau^{-1}(K_\mathrm{f})$. An example of such a transfer function is illustrated in Figure 4. All pixels in the input image, with densities in the interval D_f to $D_\mathrm{f} + \mathrm{d}D_\mathrm{f}$, are mapped using the transfer function $K_\mathrm{f} = \tau(K_\mathrm{w})$, such that they assume an output pixel density value in the interval from D_w to $D_\mathrm{w} + \mathrm{d}D_\mathrm{w}$. Consider that if the histogram h is regarded as a continuous probability density function p, describing the distribution of the intensity levels, then we obtain:

$$p_\mathrm{w}(w) = p_\mathrm{f}(D_\mathrm{f})/\mathrm{d}(D_\mathrm{f}) \qquad [10]$$

where

$$D_\mathrm{f} = \mathrm{d}u(f)/\mathrm{d}f \qquad [11]$$

In the case of histogram equalization, the output probability densities are all an equal fraction of the maximum number of intensity levels in the input image, thus generating a uniform intensity distribution. Sometimes it is desirable to control the shape of the output histogram in order to enhance certain intensity levels in an image. This can be accomplished by the histogram specialization operator which maps a given intensity distribution into a desired distribution using a histogram equalized image as an intermediate stage.

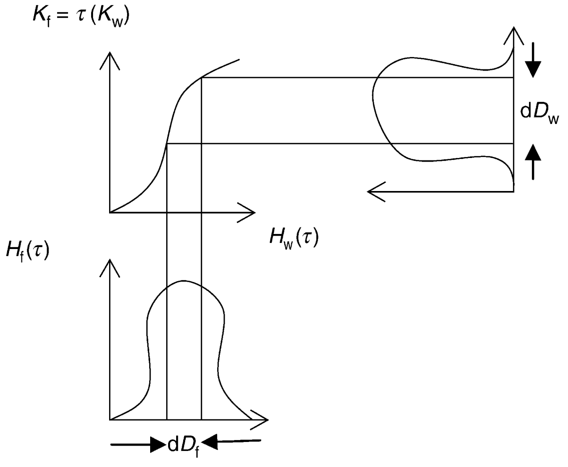

Figure 4 A histogram transformation function.

Fractals for Texture Analysis

The fractal aspect of an image has successfully been exploited for image texture analysis. The fractal concept developed by Mandelbrot, who coined the term fractal from the Latin 'fractus', provides a useful tool to explain a variety of naturally occurring phenomena. A fractal is an irregular geometric object with an infinite nesting of structure at all scales. Fractal objects can be found everywhere in nature such as coastlines, fern trees, snowflakes, clouds, mountains, and bacteria. Some of the most important properties of fractals are self-similarity, chaos, and noninteger fractal dimension (FD). The FD of an object can be correlated with the object properties. Object textures in images are candidates for characterization using fractal analysis because of their highly complex structure. The fractal theory developed by Mandelbrot is based, in part, on the work of mathematicians Hausdorff and Besicovitch. The Hausdorff–Besicovitch dimension, D_H, is defined as:

$$D_\mathrm{H} = \lim_{\varepsilon \to 0^+} \frac{\ln N}{\ln 1/r} \qquad [12]$$

where N is the number of elements of box size r required to form a cover of the object. The FD can be defined as the exponent of the number of self-similar pieces (N) with magnification factor ($1/r$) into which a figure may be broken. The FD is a noninteger value in contrast to objects that lie strictly in Euclidean space. The equation for FD is as follows:

$$\mathrm{FD} = \frac{\ln(\text{number of self-similar pieces})}{\ln(\text{magnification factor})} = \frac{\ln N}{\ln(1/r)} \qquad [13]$$

Frequency Domain Image Filtering

Ignoring noise contribution, we may obtain the basic filtering expression using eqns [7] and [8] as follows:

$$f(i,j) = \Im^{-1}[G(u,v)/H(u,v)] \qquad [14]$$

where $\Im^{-1}$ is the inverse Fourier transform and $[1/H(u,v)]$ is the inverse frequency domain filter. The key importance of this result is that, instead of performing a convolution as shown in eqn [7] to obtain the output of the system, we obtain the inverse of the Fourier transform of the impulse response $h(x,y;s,t)$ and the output image $g(x,y)$, multiply them and then take the inverse Fourier transform of the product to obtain $f(x,y)$. This is the basis for all the filtering as well as other types of image post-processing such as image reconstruction works. The filter function in eqn [9] may have different forms and shapes based on different types of functions such as uniform, Gaussian, lowpass, highpass, etc. While these are all linear filters, there are examples of nonlinear filters such as median, mean, etc. that are equally applicable for image post-processing.

Time-Frequency Domain Processing

One of the fundamental aspects of early image representation is the notion of scale. An inherent property of objects in the world and details in image is that they only exist as meaningful entities over certain ranges of scale. Consequently, a multiscale representation is of crucial importance if one aims at describing the structure of the world, or more specially the structure of projections of the three-dimensional world onto two-dimensional electronic images. The multiscale image post-processing may be facilitated by the wavelet (WT) computation.

A wavelet is, by definition, an amplitude-varying short waveform with a finite bandwidth. Naturally, wavelets are more effective than the sinusoids of Fourier analysis for matching and reconstructing signal features. In wavelet transformation and inversion, all transient or periodic data features can be detected and reconstructed by stretching or contracting a single wavelet to generate the matching building blocks. Consequently, wavelet analysis provides many flexible and effective ways to post-process images that surpass classical techniques – making it very attractive for data analysis, and modeling. The continuous WT is defined as

$$\Phi(a,b) = |a|^{-1/2} \int_{-W}^{W} x(t)\Psi\left[\frac{t-b}{a}\right]\mathrm{d}t \qquad [15]$$

where a and b are scaling and translation factors and $\Psi(t)$ is the mother wavelet. The WT may be viewed as a multiresolution analysis with shifted and scaled versions of the mother wavelet. Corresponding WT coefficients $c_{m,n}$ are

$$c_{m,n} = |a_0|^{-m/2} \int x(t)\Psi([a_0^{-m}t - nb_0])\mathrm{d}t \qquad [16]$$

The reconstruction of the signal is achieved using the admissibility condition, $\int \Psi(t)\mathrm{d}t = 0$, as follows:

$$x(t) = c \int \int_{a>0} \left[\Phi(a,b)\frac{\Psi a, b(t)}{a}\right]\mathrm{d}a\,\mathrm{d}b \qquad [17]$$

where c is a constant. For compact dyadic wavelets, binary scaling and dyadic translation of the mother wavelet form the basis functions. A dyadic translation is a shift by the amount $n/2^m$ that is an integer multiple of the binary scale factor and of the width of the wavelet as well. The discrete WT (DWT) is the natural choice for digital implementation of WT. The DWT basis and 'sub-band coding' process for discretization of a and b are identified. The iterative bandpass filter that finally leads to the wavelet is given as

$$\Psi m, n(x) = 2^{m/2}\Psi(2^{-m}x - n) \qquad [18]$$

The above eqn [18] forms the basis for 'sub-band' implementation for multiresolution analysis (MRA) of wavelet processing.

Image Post-Processing Examples

Some of the important application areas for electronic image post-processing involve the following:

(i) Biomedical image analysis and object detection;
(ii) Image compression for transmission and distribution;
(iii) Satellite image processing and noise removal;
(iv) IR/SAR/detection and clutter removal; and
(v) Image classification and machine vision.

In this article, we provide examples of some of these image-post-processing applications.

Example I: Spatial Domain Processing: Image Post-Processing in Biomedical Image

The mega voltage X-ray images are the X-ray images taken using high voltage X-radiations. These images are often characterized by low resolution, low contrast, and blur. The importance of mega voltage X-ray images arises in the context of radiation therapy for cancer treatment wherein the mega voltage X-radiation is used as a therapeutic agent. Figure 5 shows such mega voltage X-ray images, also known as an electronic portable digital (EPID) image with a very low level of detail. Figure 5a shows poor dynamic range. In order to improve the contrast of this image, without affecting the structure (i.e., geometry) of the information contained therein, we can apply the histogram equalization operation discussed in eqn [10]. Figure 5b presents the result of the histogram image enhancement post-processing process. The low level of detail in the bony anatomy is easily visible in Figure 5b.

Example II: Frequency Domain Processing: Image Post-Processing in Biomedical Tissue Detection

In medical imaging, texture is a fundamental characteristic that can be exploited to analyze tissues and pathologies in a biomedical image. The image texture is often characterized by the random nature of the objects such as tissues and body structures, noise, imaging modality, imaging limitations, and a variety of measurement parameters affecting the image acquisition process. Extensive research suggests that the fractal analysis may be combined with MRA techniques to post-process the texture content of an image in general and tumor detection in particular. In addition, the stochastic fractal Brownian motion (fBm) is well suited to describe texture and tumor characteristics since fBm offers an elegant integration of fractal and multiresolution analysis.

The fBm is a part of the set of $1/f$ processes, corresponding to a generalization of the ordinary Brownian motion $B_H(s)$. It is a nonstationary zero-mean Gaussian random function, defined as

$$B_H(t) - B_H(s) = \frac{1}{\Gamma(H+0.5)} \times \left\{ \int_{-\infty}^{0} [(t-s)^{H-0.5} - s^{H-0.5}]\,\mathrm{d}B(s) + \int_{0}^{\infty} (t-s)^{H-0.5}\,\mathrm{d}B(s) \right\}, \quad B_H(0) = 0 \qquad [19]$$

where the Hurst coefficient H, restricted to $0 < H < 1$, is the parameter that characterizes fBm; t and s correspond to different observation times of the process B_H, and Γ to the Euler's Gamma function. From the frequency domain perspective, it is known that some of the most frequently seen structures in fractal geometry, generally known as $1/f$ processes, show a power spectrum satisfying the power law relationship:

$$S(\omega) \propto \frac{k}{|\omega|^{\gamma}} \qquad [20]$$

where ω corresponds to the spatial frequency, and $\gamma = 2H + 1$. The observations of stationarity and self-similarity of fBm suggest that the use of time-frequency signal decomposition techniques, such as MRA, are well suited for the analysis of fractal signals. Thus, the fBm analysis framework, along with eqn [13], may be successfully exploited to estimate H, and hence, FD as follows:

$$\mathrm{FD} = D_E + 1 - H \qquad [21]$$

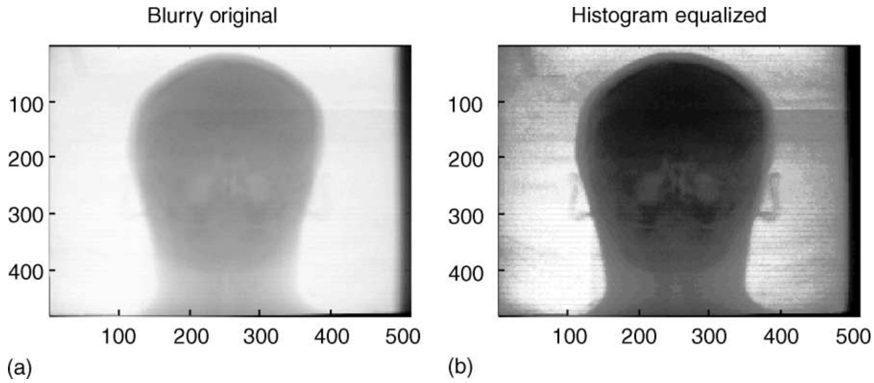

Figure 5 (a) Original EPID image; and (b) Histogram enhanced image.

where D_E is the Euclidean dimension that contains the fBm (i.e., the position of each point of the process is described with the vector $x = (x_1, \ldots, x_E)$).

We exploit the wavelet MRA combined with complementary signal processing techniques such as time-frequency descriptions and parameter estimation of stochastic signals in fBm framework for tumor detection in CT images. We demonstrate an example of image post-processing for tumor segmentation. We test our models to estimate FD in brain MR images as shown in Figures 6a and b, respectively. The images are scanned pixel-by-pixel with a sliding window that defines the subimage for analysis. Pre-processing, such as median filtering, is used to remove noise from the original images.

The results for the spectral analyses of CT image are shown in Figure 7. The results suggest that the FD values do not depend on the mother wavelet used. The window size that offers the best results is the 16 × 16 pixels, while better discrimination between low and high H values is obtained using biorthogonal wavelet. To improve tumor detection performance of the spectral technique, we apply post-processing filters on the results in Figures 8a and b, respectively. In Figure 8, for an example, we show the filtered fractal analysis results corresponding to a 16 × 16 pixels window with Daubechies 4. It is possible to obtain an even better description of the location and shape of the tumor after applying a two-dimensional post-processing filter to the matrices derived from the variance estimation algorithms with biorthogonal 1.5 wavelet and an 8 × 8 pixels window, respectively. The results are shown in Figure 9 and are similar for both averaging and Gaussian filters as well.

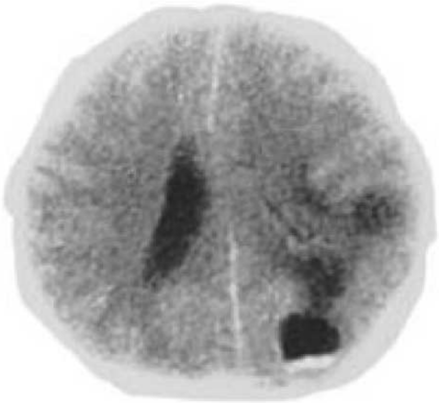

Figure 6 Biomedical images used to test the proposed algorithms.

Example III: Time-Frequency Domain Filtering: Image Post-Processing in Noise Reduction (De-Noising)

This section discusses an example of image post-processing in image noise removal (or de-noising) using a scale-frequency technique such as wavelets. The wavelet transform has been proved to be very successful in reducing noise (de-noising) in an image. We recently propose a new family of wavelets namely Joint Harmonic Wavelet Transform (JHWT) that offers better speckle noise reduction for image post-processing. Unlike wavelets generated by discrete dilation functions, as shown in eqn [18], whose shape cannot be expressed in functional form, JHWT has a simple structure, which can be written, in the

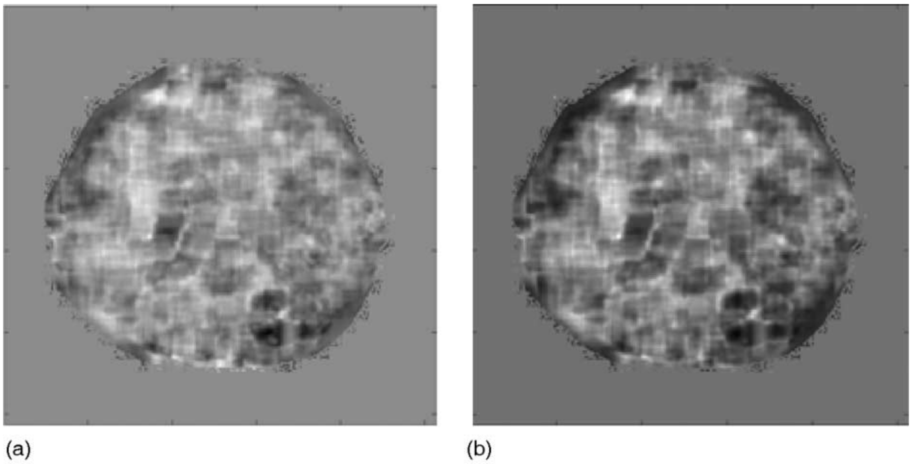

Figure 7 MR-Fractal analysis using spectral method (a) and (b) correspond to a 16 × 16 analyzing window for Daubechies 4 and biorthogonal 1.5 wavelets, respectively.

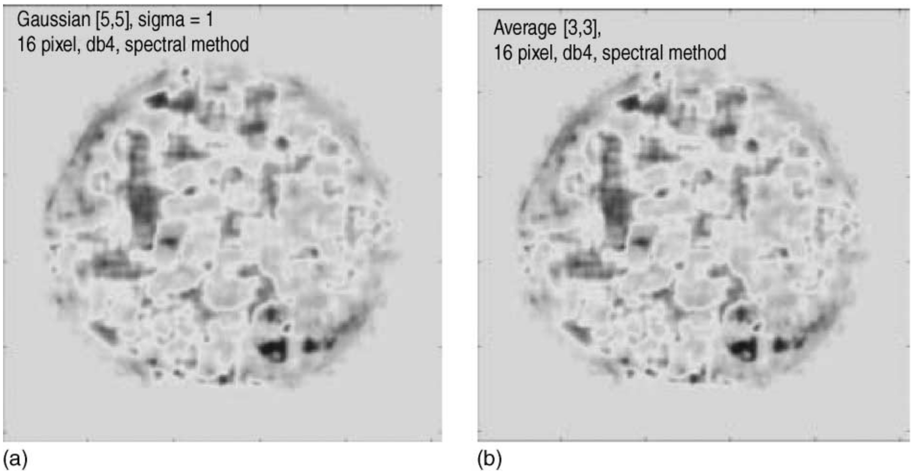
Gaussian [5,5], sigma = 1
16 pixel, db4, spectral method
Average [3,3],
16 pixel, db4, spectral method
(a)
(b)

Figure 8 Post-processing results using: (a) 5 × 5 Gaussian filtering with $\sigma = 1$ and (b) 3 × 3 averaging filter.

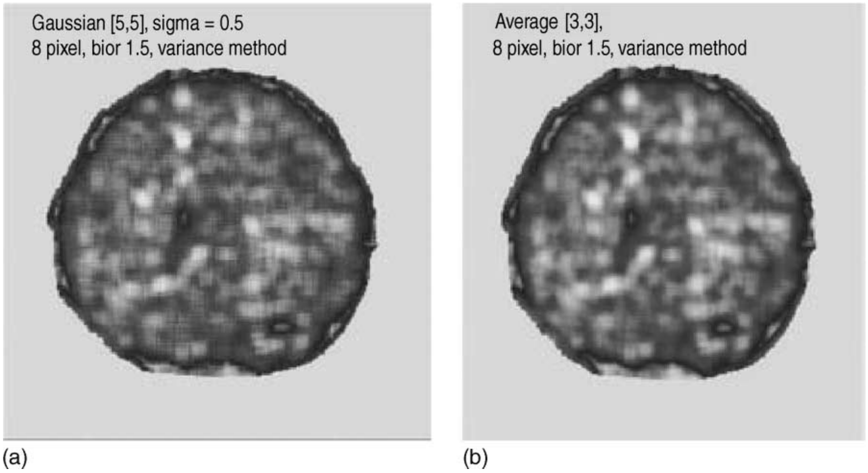
Gaussian [5,5], sigma = 0.5
8 pixel, bior 1.5, variance method
Average [3,3],
8 pixel, bior 1.5, variance method
(a)
(b)

Figure 9 Post-processing on the fractal analysis results of Figure 8 for (a) 5 × 5 Gaussian filtering with $\sigma = 1$, and (b) 3 × 3 averaging filtering.

frequency domain:

$$W(\omega) = \int_{-\infty}^{\infty} w(x)\exp(-i\omega x)\mathrm{d}x, \\ = \begin{cases} 1 & \text{for } 2\pi \leq \omega < 4\pi \\ 0 & \text{elsewhere} \end{cases} \qquad [22]$$

Now, at level j, we let $g(x) = w(2^j x)$ and shift it by $k/2^j$ to obtain the harmonic wavelet as:

$$g(x - k/2^j) = \{\exp\{i4\pi(2^j x - k)\} \\ - \exp\{i2\pi(2^j x - k)\}\}/i2\pi(2^j x - k) \qquad [23]$$

where integers j and k define the scale and translation of the wavelet $w(x)$, respectively on the x-axis. Now consider a real function $f(t)$ is represented by the sequence f_r of length $(n - 1)$, where $r = 0$ to $2^n - 1$. To obtain JHWT coefficients, we take FFT of f_r such that,

$$|F_{2^j+s}|^2 = (1/n^2)\sum_{k=0}^{n-1} |a_{2^j+k}|^2 \exp(-i4\pi sk/n), \\ \text{where } r = 0 \text{ to } n-1 \text{ and } s = 0 \text{ to } n-1 \qquad [24]$$

where j denotes the level of the wavelet, k is the translation, m and n are adjacent levels of wavelets with $n = 2^j$, and $s = m - n$. The JHWT in eqn [24] offers all the properties and advantages of standard discrete wavelet transform. Putting $n = 2^j$ and $m = s + n$, eqn [24] yields all the JHWT coefficients from

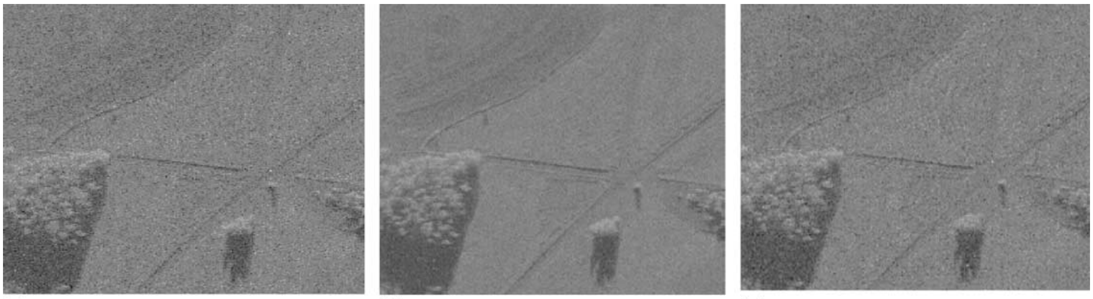

(a) (b) (c)

Figure 10 (a) Noisy image; (b) de-noised image with JHWT-filtering; and (c) de-noised image with WP-filtering.

the Fourier coefficients as

$$\sum_{j=0}^{N-1} |F_{2^j+4}|^2 = |a_0|^2 + \sum_{j=0}^{n-2} (1/2^j) \times \sum_{k=0}^{2^j-1} \left(|a_{2^j+k}|^2 + |a_{N-2^j-k}|^2\right) + (a_{N/2})^2 = (1/N) \sum_{r=0}^{N-1} f_r^2 \qquad [25]$$

Unlike conventional correlation techniques, the perfect octave band limited characteristics of JHWT eliminate the necessity of any multiplication. We show an example of image de-noising in Figure 10 using JHWT.

Image Post-processing, Transmission, and Distribution

One of the areas wherein image post-processing plays a vital role is in image transmission. A significant problem for any digital library is the network bandwidth required to transmit large digital images. The bandwidth consumption depends on the number of images used by a single user, and with the number of users requesting the services of the library. The popularity of database browsers such as Gopher and Mosaic has serious implications for image transmission and distribution. There has been active research in innovative new compression schemes and transmission protocols that reduce bandwidth requirements for many users along with software servers and visualization. Further, compression of digital image and video data plays an important role in reducing the transmission and distribution cost for visual communication.

One of the requirements for effective image transmission and distribution is reduced data size for faster speed. Image compression schemes are often used for image data reduction. Most of the successful image compression schemes exploit subdividing the image into blocks for increased compression. However, this subdivision of image renders a blocky effect on the compressed image. There has been intense research in removing the blocky effect on the compressed image. To achieve high bit rate reduction while maintaining the best possible perceptual quality, post-processing techniques provide one attractive solution. There has been a significant amount of work done in post-processing of compressed image/video. Some of them are developed from the image enhancement point of view while others are from image restoration. In the case of image restoration, the post-processing is considered as a reconstruction problem. Image enhancement methods are performed based on the local contents of the decoded image as discussed above.

The image post-processing algorithms are, generally, very effective in removing blocking artifacts. A straightforward approach to remove blocking artifacts is to apply lowpass filtering to the region, where artifacts occur. Consequently, the filter must be adapted to the block boundaries, i.e., the image partition. Further, if high frequency details such as edges and texture are to be preserved, then lowpass filter coefficients should be selected appropriately in regions where such details occur. Therefore, adaptive image post-processing such as spatial filtering methods may be applied to deal with this problem. In low-bit-rate image and video compression and distribution systems, high frequency transformed coefficients are often lost due to aggressive quantization or uneven error protection schemes. However, with the block-based encoding and transmission methods, the amount of high frequency texture loss is uneven between adjacent blocks. Thus, it is possible to exploit this texture-level correlation between adjacent image blocks to reconstruct the lost texture.

A relevant problem of robust transmission and distribution of compressed electronic images involves unreliable networks. The pre- and post-processing of the image data may play a vital role in addressing this problem. With the rapid growth of mobile wireless communication systems, wireless multimedia has recently undergone enormous development. However, to design an efficient multimedia communication system over wireless channels, there still exist many challenges due to severe wireless channel conditions, special characteristics of the compressed multimedia data, and resource limitation, such as power supply. Since multimedia data, such as an image, contain redundancy, to efficiently utilize limited resources, source compression may be necessary. Further, as mentioned above, the ease of image creation, processing, transmission, and distribution has also magnified the associated counterfeiting problems. We discuss the relevant issues of image distribution in the following sections.

Techniques in Image Distribution

The spectacular development of imaging sensors, communication, and internet infrastructure has also made the counterfeiting tools easily accessible and more affordable. In addition, the threats and vulnerabilities of the internet have also increased manyfold. All of these together pose a formidable challenge to the secure distribution and post-processing of images.

Requirements of Secure Digital Image Distribution

In light of the above-mentioned vulnerabilities of image distribution channels and processes, below are a few requirements of secure image distribution:

- *Data Integrity and Authentication*: Digital image is easy to tamper with. Powerful publicly available image processing software packages such as Adobe Photoshop or PaintShop Pro make digital forgeries a reality. Integrity of the image content and authentication of the source and destination of the image distribution are, therefore, of utmost importance.
- *Digital Forensic Analysis*: Consider a litigation involving a medical image that is used for diagnosis and treatment of a disease, as shown in the post-processing example in Figures 7 or 8 above. Integrity of the image again is vital for the proper forensic analysis.
- *Copyright Protection*: Digital copyright protection has become an essential issue to keep the healthy growth of the internet and to protect the rights of on-line content providers. This is something that a textual notice alone cannot do, because it can so easily be forged. An adversary can steal an image, use an image processing program to replace the existing copyright notice with a different notice and then claim to own the copyright himself. This requirement has resulted in a generalized digital rights management initiative that involves copyright protection technologies, policies, and legal issues.
- *Transaction Tracking/Fingerprinting*: In the continuous (re)distribution of digital images through internet and other digital media (CD, camera, etc.), it is important to track the distribution points to track possible illegal distribution of images.

Next, we will discuss how the digital watermarking technology can help enforce these security requirements.

Example of secure image distribution: digital watermarking

Digital watermarking is the process of embedding a digital code (watermark) into a digital content like image, audio, or video. The embedded information, sometimes called a watermark, is dependent on the security requirements mentioned above. For example, if it is copyright application, the embedded information could be the copyright notice. Figure 11 shows a block diagram of the digital watermarking process.

Consider that we want to embed a message M in an image, f_o, also known as the original image. For the sake of brevity, we drop the indices representing the pixel coordinates. The message is first encoded using source coding and optionally with the help of error correction and detection coding, represented in Figure 11 as $e(M)$ such that

$$W_m = e(M) \tag{26}$$

The encoded message W_m is then combined with the key-based reference pattern W_k, in addition to the scaling factor α, to result in W_a which is the signal that is actually added to the original image, given as

$$W_a = \alpha(f_o)[W_M \otimes W_K] \tag{27}$$

Note that the scaling coefficient α may be a constant or image dependent or derived from a model of image perceptual quality. When the scaling factor α is dependent on the original image, we can have informed embedding which may result in more perceptually adaptive watermarking. The operation involving W_m and W_k is the modulation part, which is dependent on the actual watermarking algorithm used. As an example, in a typical spread-spectrum based watermarking, this operation is usually an exclusive NOR (XNOR) operation. The final operation is an additive or multiplicative embedding of W_a with the original image. We show a simple additive watermarking process as

$$f_w = f_o + W_a \tag{28}$$

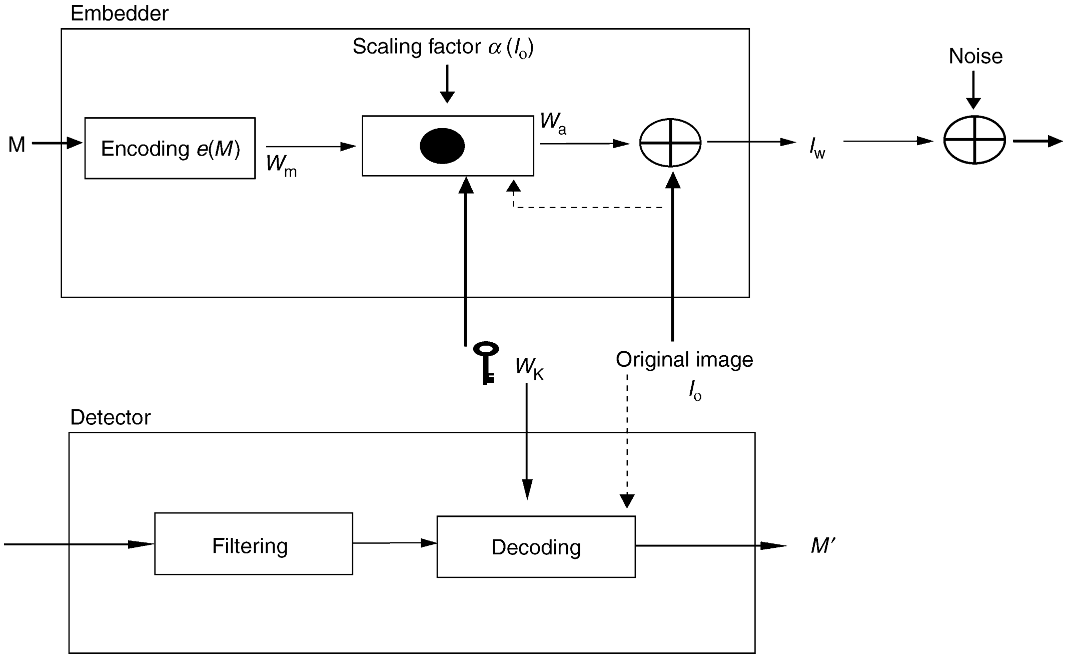

Figure 11 A simple watermarking process: embedder and detector.

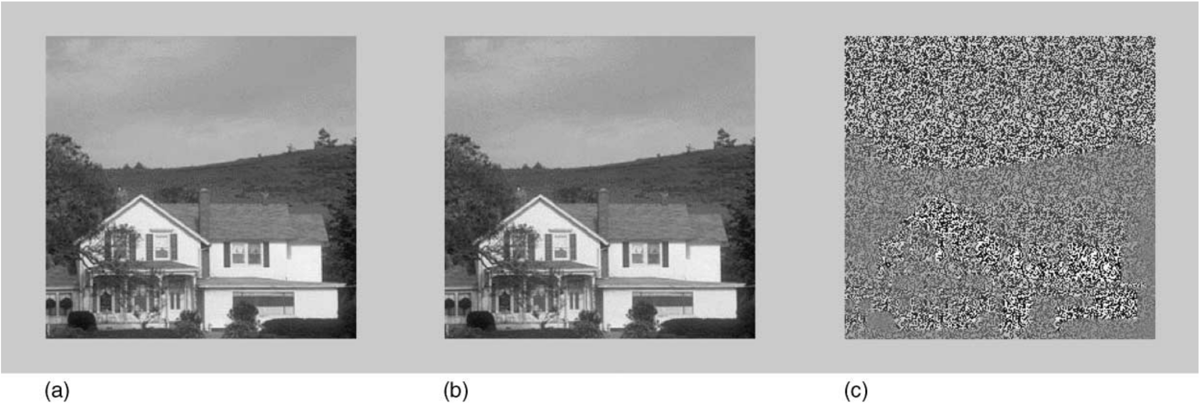

Figure 12 (a) Original image, (b) watermarked image, and (c) the difference image.

The result of this process is the watermarked image f_w which may actually need some clipping and clamping operations to have the image pixel values in the desired range.

The detector, as shown in Figure 11, does the inverse operations of the embedder. If the original image is known, as in the case of nonblind watermarking, it is first subtracted from the watermarked image. Blind watermarking detector, on the other hand, uses some filtering with a goal to estimate and subtract the original image component. The result of filtering is then passed to the decoder, which collects the message bits with the help of the shared key, and finally gets back the decoded message M'.

Figure 12 shows the original image, watermarked image, and the difference image using a spread-spectrum based watermarking algorithm, respectively. The enhanced difference image represents a noise-like image, wherein the scaling coefficient α is clearly dependent on the original image. As evident from Figure 12b, the hidden message is imperceptible as well as the change in the original image. The imperceptibility is one of the primary criteria of digital watermarking.

There are other criteria such as unobtrusiveness, capacity, robustness, security, and detectability. Thus, digital watermarking is a multidimensional problem, which attempts to make a trade-off among a number of different criteria. The inserted watermark should not be obtrusive to the intended use of the original image. Because the watermarking process does not increase the size of the original image, it may be desirable to add as much information as possible. However, generally speaking, the more information one adds, the more severe will be the impact on the perceptual quality of the image. From the detection point of view, watermark should be robust enough to tolerate a range of image post-processing and degradation. In addition, the embedded watermark should also be secure enough so that it may not be removed from the watermarked image. Finally, detectability of the embedded watermark is an important criterion that places some constraints on the embedding algorithm.

In order to fully appreciate the use of watermarking for image distribution, it is instructive to look at different classes of watermark, which is delineated in Figure 13.

For an example, depending on the robustness criterion of watermarking, they can be classified into three categories such as robust, fragile, and semi-fragile. Note that while the robustness of watermark is important, it is not always equally desirable in all applications. Let us take the authentication as an example. If you, as the owner of a digital image, embed some authentication information into the image using watermarking, you might want to see the authentication fail, when some one steals your image. This is an example of nonrobust watermark, where the fragility of the embedded watermark is required to detect any forging done on the digital image.

Now, consider how watermark provides the security requirements for secure image distribution. In particular, we look into the authentication of watermark. Figure 14 shows the use of watermark for image authentication, which has some similarities with the cryptographic authentication using digital signature. However, there are potential benefits to using watermarks in content authentication and verification. Unlike cryptography, watermarking offers both in-storage and in-transit authentication. Watermarking is also faster compared to encryption, which is very important in internet-based image distribution. By comparing a watermark against a known reference, it may be possible to infer not just that an alteration occurred but what, when and where changes have occurred. As shown in Figure 14, first we identify some features from the image and then we compute the digital signature, which is then embedded into the image. The detector takes a test image, computes the signature and extracts the embedded

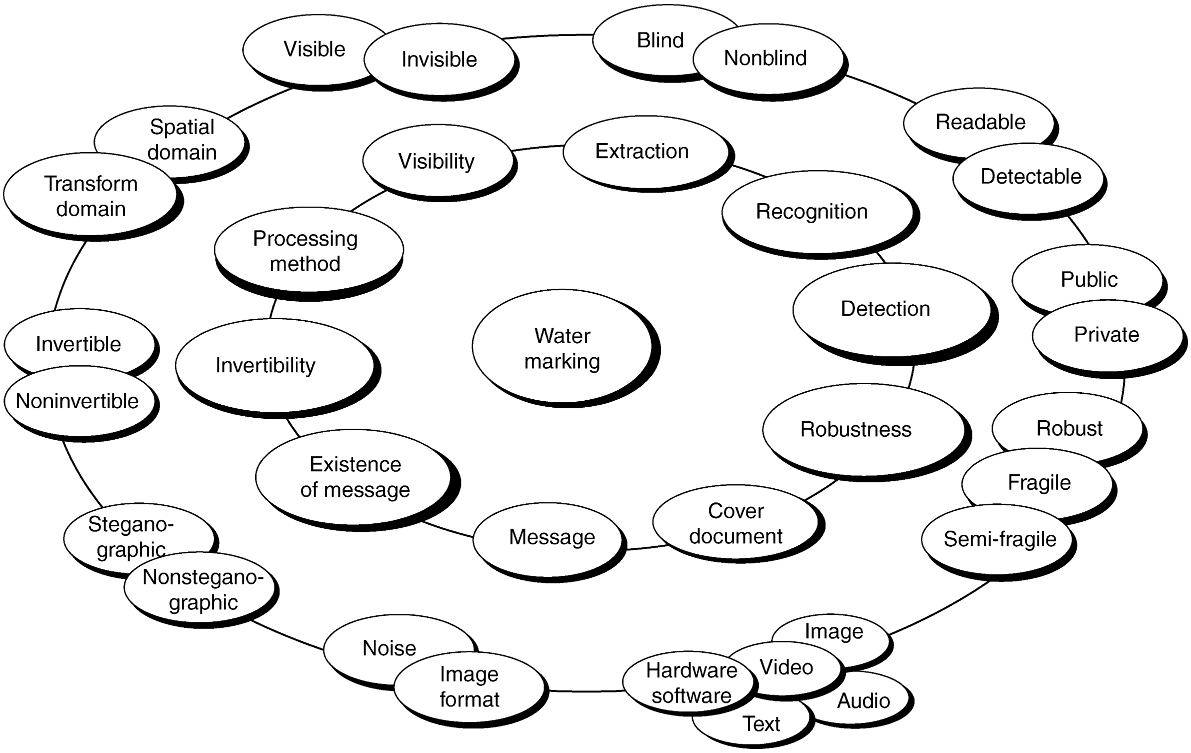
Visible
Invisible
Blind
Nonblind
Spatial domain
Transform domain
Visibility
Extraction
Readable
Detectable
Recognition
Processing method
Detection
Public
Private
Water marking
Invertible
Noninvertible
Invertibility
Robustness
Robust
Fragile
Existence of message
Cover document
Semi-fragile
Message
Stegano-graphic
Nonstegano-graphic
Noise
Image format
Hardware software
Image
Video
Audio
Text

Figure 13 Watermark classifications.

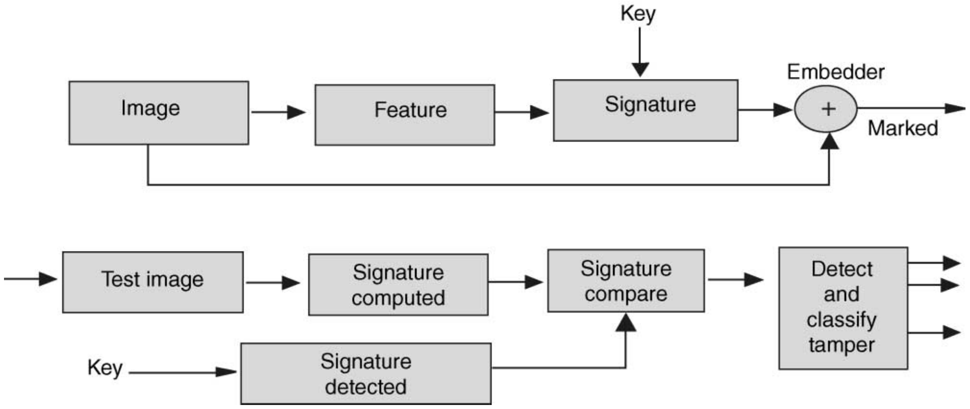

Figure 14 Image authentication using watermark.

signature, and then compares the two. If they match, the image is authenticated. If they do not match, then the image may have gone through some forging/tampering effects and the algorithm may optionally detect the tamper area as well.

One of the limitations of the watermarking approach in secure image distribution is that the technology is not yet widely deployed, and nor is the protocol satisfactorily standardized. Other major limitation is that the watermarking process cannot be made sufficiently robust to arbitrary types of different attacks. In reality, therefore, the hybrid combination of cryptography and watermarking is expected to improve the secure distribution of images across the unreliable internet.

Concluding Remarks

In this chapter, we discussed the principles of image post-processing and distribution. We showed a few representative examples for post-processing and distribution. We also explained the relevant issues in image distribution and present digital watermarking as one of the solutions for secure image distribution.

Acknowledgment

One of the authors (KMI) wishes to acknowledge the partial support of this work through a grant (RG-01-0125) from the Whitaker Foundation.

See also

Imaging: Information Theory in Imaging.

Further Reading

Ahmed F and Moskowitz IS (2004) A correlation-based watermarking method for image authentication applications, accepted for publication in *Optical Engineering*.

Barlaud M (1994) *Wavelets in Image Communication, Advances in Image Communication 5*. Amsterdam: Elsevier.

Chui A (1992) *An Introduction to Wavelets*. San Diego: Academic Press.

Cox NI, Bloom J and Miller M (2001) *Digital Watermarking: Principles & Practice*. Francis, CO: Morgan Kauffman Publishers.

Daubechies I (1990) The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* 36(5): 961–1005.

Gonzalez RC and Woods RE (1993) *Digital Image Processing*. New Jersey: Addison-Wesley Publishing Company.

Iftekharuddin KM and Parra C (2004) Multiresolution-fractal analysis of medical images. Under Review. *Neuroimaging*, vol. 1. New York: Wiley/IEEE Press, Chapter 4.

Jain AK (1989) *Fundamental of Digital Image Processing*. New Jersey: Prentice-Hall.

Lu N (1997) *Fractal Imaging*. San Diego: Academic Press.

Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(7): 674–693.

Mandelbrot BB (1983) *The Fractal Geometry of Nature*. San Francisco, CA: Freeman.

Schmucker AM and Wolthusen SD (2003) *Techniques and Applications of Digital Watermarking and Content Protection*. Norwood, MA: Artech House.

Serra J (1982) *Image Analysis and Mathematical Morphology*. San Diego: Academic Press.

Stallings W (2003) *Cryptography and Network Security – Principles and Practices*, 3rd edn. New Jersey: Prentice-Hall.

Wu M and Liu B (2002) *Multimedia Data Hiding*. New York: Springer.

Smart Pixel Arrays

P Seitz, Centre Suisse d'Electronique et de Microtechnique SA, Zurich, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

For more than 50 years silicon has been the semiconductor on which the enormous progress of microelectronics is based. Since silicon is also an excellent detector for electromagnetic radiation, in particular in the visible spectrum between 400 and 700 nm, the advances in semiconductor technology have directly influenced the progress of solid-state image sensing. Once the relentlessly shrinking dimensions of transistors have reached a critical minimum limit, it has become economically feasible to co-integrate analog and digital circuitry with each light-sensitive pixel. This capability is at the heart of the technological revolution called smart pixel arrays.

The first section relates how the shrinkage of minimum feature size on silicon chips paralleled the reduction of solid-state pixel geometry, as a function of time. This development ceased once the pixel size reached the optical diffraction limit, which marked the birth of the field of smart pixel arrays. In the second section, the basic functionality and the most important physical limitations of conventional pixels and photosensors are given. The third section deals with opto-electronic functionality that is preferred today, of smart pixel arrays; in particular, the capability of spatio-temporal demodulation is very much in demand. In the fourth section, some of the most useful modular building blocks for smart pixel arrays are presented, together with a few key applications. An outlook on photonic systems-on-chip (SoC) is given in the fifth section, with a discussion of single-chip machine vision systems (so-called seeing chips), organic semiconductors for even higher levels of integration of photonic microsystems, and smart pixel arrays that are sensitive to physical parameters other than electromagnetic radiation.

Silicon Technology for Image Sensing

Long before it was realized that silicon would be the material on which modern semiconductor industry and the information technology revolution is based, in 1940 Russell Ohl discovered that silicon is an excellent detector of electromagnetic radiation. In particular, silicon is an excellent detector for light in the visible and near infrared part of the spectrum, for wavelengths between 400 and 1,000 nm. This is very fortunate because it allows designers of image sensors to profit directly from the amazing progress of silicon technology. Since the 1970s, the minimum feature size on silicon chips showed a reliable exponential reduction of about 10% per year, as illustrated in Figure 1. This is a direct consequence of the observation by Intel's Gordon Moore, sometimes called 'Moore's law', that the number of transistors per integrated circuit doubles about every two years. The evolution of minimum feature size is paralleled by the evolution of minimum pixel size of solid-state image sensor, as illustrated in Figure 2. Once the pixel size reached a record low value of 2.4 μm in the year 1996, no further reduction in pixel size seemed desirable, because the optical diffraction limit was essentially reached.

The combination of technological availability of smaller and smaller electronic components and pixel size that demanded no further reduction, inspired researchers to provide each pixel with an increasing amount of electronic functionality, all of which is available simultaneously for massively parallel information processing. This represents the heart of the smart pixel array revolution; pixels that not only sense the incident electromagnetic radiation but that are capable of pre-processing the acquired information in a relevant fashion.

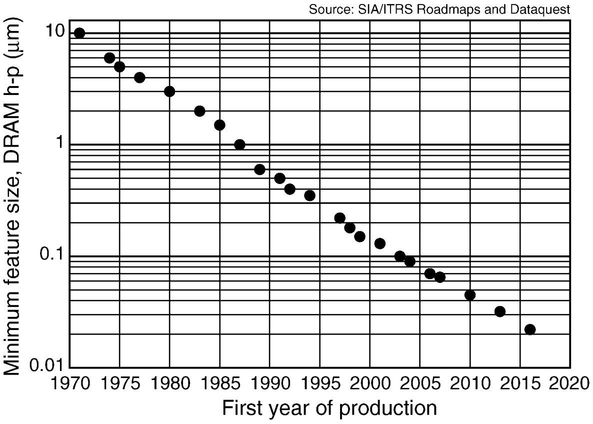

Figure 1 Evolution of minimum feature size (DRAM half-pitch) in state-of-the-art silicon technology, showing an exponential reduction for the past 30 years, which is expected to continue for at least another decade.

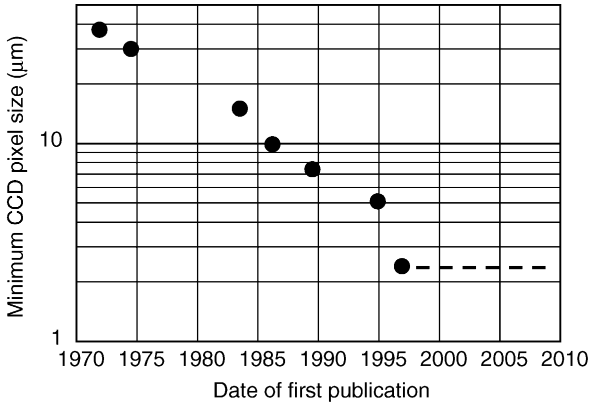

Figure 2 Evolution of minimum pixel pitch in CCD image sensors, tracking the exponential reduction of minimum feature size shown in Figure 1, until a lower limit was reached in 1997, imposed by optical diffraction. Since then, no smaller pixel pitches have been described in the literature; this is indicated graphically with the hatched 'diffraction limit' line.

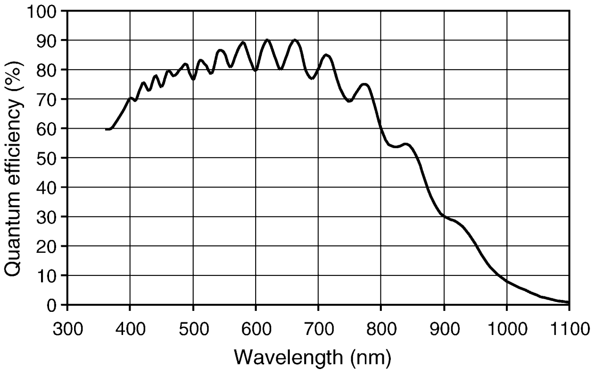

Figure 3 External quantum efficiency of a photosensor realized with an industry standard CMOS process, showing quantum efficiencies above 50% in the whole visible spectral region.

Basic Functionality and Physical Limitations of Conventional Solid-State Photosensors

The prime task of a pixel is the detection of incident electromagnetic radiation. The detection process consists of six basic steps:

1. The incident photon impinges on the surface of the image sensor, and it has to reach the bulk of the semiconductor.
2. In the bulk of the semiconductor the photon has to interact with the semiconductor, converting its energy into a number of electron–hole pairs.
3. The photogenerated charge carriers must be transported to the sensor surface.
4. At the surface, the photogenerated charge needs to be collected and accumulated in the storage part of the pixel.
5. The accumulated photocharge must be transferred to the input of an electronic circuit capable of detecting this charge.
6. The charge detection circuit converts photocharge into a corresponding voltage, exhibiting a monotonous (but not necessarily linear) transfer function.

Since the interaction length of electromagnetic radiation in silicon is in the tens of micrometers for incident photon energies between 1 eV and 10 keV (corresponding to a wavelength range of about 0.1 to 1,000 nm), silicon is an excellent detector for this large spectral range. This is illustrated for the visible and near infrared spectral range in Figure 3, showing the absolute external quantum efficiency of a pixel realized with an industry standard silicon process. In the visible spectral range, absolute external quantum efficiencies close to 100% can be reached, even with standard silicon processes for the microelectronics industry that were not optimized for photosensing applications.

Photocharge transport to the surface of the semiconductor is mainly provided by diffusion, i.e. random thermal motion of the charge carriers that exhibit a velocity of about 10^5 m/s in silicon at room temperature. The mean distance L traveled by a diffusing charge carrier during time t is given by

$$L = \sqrt{Dt} \tag{1}$$

with the diffusivity D, which has a value of around 40 cm^2/s for electrons in low-doped silicon at room temperature.

Once the photocharge carriers are close to the surface of the semiconductor, they are transported by drift in an electric field E, exhibiting a linear dependence of the transport distance L with time t:

$$L = \mu E t \tag{2}$$

with the mobility μ which has a value of around 1300 cm^2/Vs for electrons in low-doped silicon at room temperature.

There are two basic principles for creating the electric field with which the photocharge carriers are transported to their storage site within the pixel. A photodiode consists of a semiconductor junction between a p-doped and an n-doped region, as illustrated in Figure 4. Such a photodiode also represents a capacitance on which photocharge can be accumulated and stored. Photodiodes are used as photosensors in one

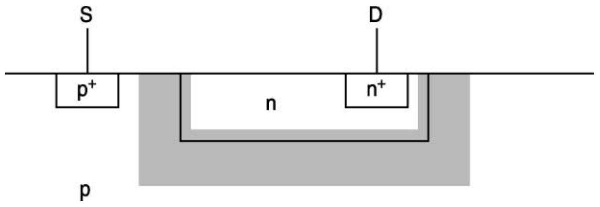

Figure 4 Cross-section of a photodiode, consisting of an n-doped region in a p-doped substrate. The substrate is contacted using a heavily p-doped diffusion S, and the n-region is contacted using a heavily n-doped diffusion D. The space–charge region is illustrated as hatched area at the pn-junction.

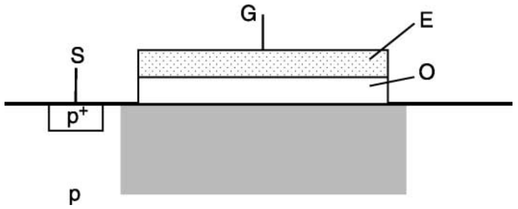

Figure 5 Cross-section of an MOS capacitance, the basic element of a CCD image sensor. On top of a p-substrate a thin transparent insulator O and a transparent electrode E are fabricated. The substrate is contacted with a heavily p-doped diffusion S, and the electrode E is contacted with the gate contact G. The hatched area illustrates the space–charge region under the electrode.

popular type of image sensor, so-called APS (active pixel sensor) or CMOS (complementary metal oxide semiconductor) imagers.

An alternative to a photodiode is an MOS (metal oxide semiconductor) capacitor, consisting of a transparent electrode that covers an insulating layer on top of semiconducting material, as illustrated in Figure 5. An MOS capacitor is the basic element of a charge coupled device (CCD), the other widespread type of image sensor.

The charge transfer between photodiodes and photodetector circuits is based on electrical conduction, occurring essentially with the speed of light. In CCDs, the charge transfer is effectuated with clocks, and excellent charge transfer efficiencies can be observed at clock frequencies of several tens of MHz.

The major limitation in semiconductor photosensors is the last element in the line, i.e. the electronic charge detection circuit. In Figure 6, a so-called source follower is illustrated, which is by far the most frequently used photocharge detection circuit. Its performance is limited by the thermal (Johnson) noise in the channel of the first field effect transistor (FET). This corresponds to a statistical uncertainty with which an amount of photocharge can be measured at the gate of the FET. The input referred charge noise

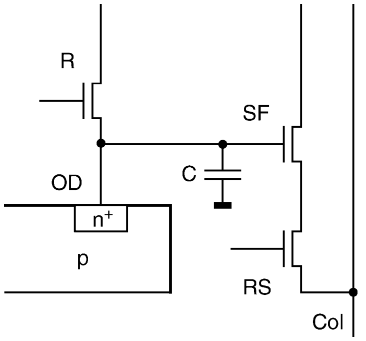

Figure 6 Schematic diagram of a photosensor's output diffusion OD connected to a reset transistor R and a source follower transistor SF, the most widely used charge detection circuit for image sensors and smart pixel arrays. The photocharge is stored on the effective capacitance C at the gate of transistor SF, which is read out using a row select transistor RS, connecting the source follower to the readout column Col.

variance s_Q is given approximately by

$$s_Q \cong C^2 \frac{4kTB}{g} \qquad [3]$$

where C is the effective capacitance seen by the gate of the FET, k is Boltzmann's constant, T is the absolute temperature, B is the effective bandwidth at the FET's channel, and g is the transconductance of the FET.

Desired Optoelectronic Functionality in Smart Pixel Arrays

From the previous section it becomes clear that even conventional pixel and photosensors can be improved in several desirable respects:

- The spectral response can be improved in the ultraviolet (50–400 nm), or it can be expanded significantly into the X-ray region (below 0.1 nm) and into the infrared region (above 1,000 nm).
- The speed of the photosensors can be increased by replacing diffusion based photocharge transfer mechanisms with the directed, more efficient drift in electric fields. This is true for the vertical as well as for the lateral charge transport direction.
- The sensitivity of the electronic charge detection process can be improved in two ways. It can be achieved by decreasing the effective detection capacitance C at the input of the charge detection circuit, as can easily be seen from eqn [3]. An alternative is to make use of a physical charge amplification process (i.e. avalanche multiplication in a high electric field), that has low intrinsic noise.

- A derivative improvement that is desired is the increase of the dynamic range D/R of the complete detection process, as defined, for example, by

  $$D/R = 10^{10} \log \frac{Q_{\max}^2}{s_Q} \qquad [4]$$

  where $Q_{\max}$ denotes the maximum (or saturation) charge and s_Q is the total charge noise variance of the electronic detection circuit.

In addition to the above described improvements of conventional pixel functionality, a huge demand exists for the capability to measure the parameters of modulated electromagnetic signals, i.e. to demodulate such signals. This demand is related to the large number of very powerful electronic and opto-electronic measurement principles, capable of improving the sensitivity of the measurement process significantly by employing modulated signals, either in the temporal or in the spatial domain. For this reason, smart pixel arrays are sought that offer this combination of photodetection and demodulation.

Modular Building Blocks for Smart Pixel Arrays

Towards the end of the last century, a large body of literature has been created, addressing various approaches for the realization of one of the above described desired opto-electronic functionalities. The result is a powerful ‘toolbox’ of modular building blocks for custom smart pixel arrays, with which optimized solutions for optical measurement problems and practical applications can be constructed.

High-Sensitivity Charge Detection

From the noise formula (eqn [3]) two possibilities for the improvement of the sensitivity of image sensors can be deduced: A first approach is the reduction of the capacitance value. An excellent example for this is the invention of the ‘double gate’ FET, in which the charge to be measured is not placed on a gate electrode at the surface but resides in the semiconductor bulk. In this way the effective capacitance of the pixel can be reduced by at least one order of magnitude compared to conventional pixels. The resulting capacitance of less than 1 fF leads to a readout noise of about one electron at room temperature and for a readout bandwidth of several MHz. Unfortunately the small capacitance allows the storage of only around ten thousand electrons. For such small charge packets, quantum noise is still relevant, leading to significantly degraded image quality, even when the images are taken under favorable lighting conditions.

A second approach to high-sensitivity photosensing exploits the dependence of the charge detection variance on the measurement bandwidth, as indicated in eqn [3]. One method for exploiting this consists of reading out the image sensor at a significantly reduced bandwidth of a few 10 kHz. At the same time, the image sensor is cooled to temperatures of around −100 °C. This becomes necessary because readout at such slow speeds increases the time, to read one complete image, to many minutes, requiring a substantial reduction of the dark current. Instead of working at low readout bandwidths, it is also possible to read out the same information repeatedly and non-destructively at high speed, and to average many measurements of the same amount of photocharge. Both methods have been shown to lead to readout noise levels of about one electron; on the other hand, both approaches require very long times for the acquisition of complete images.

This problem can be resolved by exploiting the massive parallelism that smart pixel arrays offer: low-bandwidth techniques can be employed without losing effective readout speed. Pixels and columns are read out with low-bandwidth amplifiers that are all active simultaneously, and only the final off-chip readout amplifier needs to operate at a large bandwidth. In practice, this approach has been used successfully in single-chip digital cameras with Megapixel resolution, exhibiting a readout noise of only a very few electrons.

A completely different approach is based on the use of a physical amplification mechanism with low intrinsic noise, before detecting the amplified charge packets with a conventional electronic circuit; the avalanche multiplication effect in high electric fields is very well suited to this approach. Each pixel is supplied with its own combination of stabilized avalanche multiplier and conventional charge detection circuit. This approach has been used for the realization of a fully integrated 2D array of single photon detectors that can be fabricated with industry standard CMOS processes. Similar performance is obtained with the Impactron™ series of high-sensitivity CCD image sensors offered by Texas Instruments Inc., making use of a single multistage avalanche multiplier at the output of the image sensor. For both high-sensitivity smart pixel arrays, single-photoelectron performance is reported at room temperature and at readout frequencies of several MHz, as required for video applications.

Photosensing with Improved Dynamic Range

Conventional solid-state image sensors and cameras usually have a very limited dynamic range, as defined in eqn [4]. Typical values are between 40 and 80 dB, rarely exceeding four orders of magnitude. In natural scenes, however, much higher *D*/*R* values can be encountered. To cope with this, the human eye is capable of spanning a dynamic range of more than 150 dB. This realization has spurred the invention of several methods with which the *D*/*R* of smart pixel arrays can be increased.

One approach consists of acquiring several images of the same scene with different exposure times, with the subsequent combination into one high-dynamic-range composite picture. *D*/*R* values exceeding 120 dB have been obtained with such multi-exposure methods. Fast-moving objects, however, will have relocated from one exposure to the next, resulting in false, blurred representations of these objects in the composite picture.

This can be overcome using nonlinear response characteristics similar to the ones found in natural vision. Impinging light produces a photocurrent that is highly linear with the incident light intensity. Because the voltage across the photodiode is logarithmically dependent on the photocurrent, a logarithmic voltage-illumination response is obtained, with a *D*/*R* exceeding 120 dB. Such logarithmic pixels are well-suited for applications providing high illumination levels but they suffer from excessive noise and slow response speed at low light levels.

This disadvantage led to the development of linear-logarithmic pixels, exhibiting a high-sensitivity linear response at low light levels, and a gradual changeover to logarithmic response once a certain, programmable light level is reached. The design principle is illustrated in Figure 7. An otherwise conventional CMOS/APS pixel is supplied with a device that exhibits logarithmic voltage–current response, such as a diode. As long as the cathode is at a higher voltage than the anode, no current flows across the diode, and the pixel response is linear. Once the photocurrent has lowered the cathode voltage below the anode voltage, the pixel response becomes logarithmic, as described above. The *D*/*R* of such a linear-logarithmic pixel reaches unprecedented high values; almost 200 dB have already been obtained in practice. Because of the simplicity and robustness of this principle, such linear-logarithmic smart pixel arrays have found their way into commercial sensor and camera products that are available commercially, as illustrated in Figure 8. A circuit board and a light bulb are imaged in linear mode (Figure 8a), showing highly saturated areas, and in linear-logarithmic mode (Figure 8b), in which even the light bulb's filament can clearly be seen.

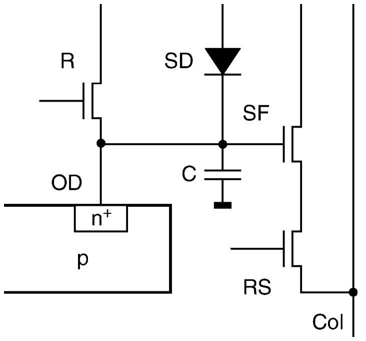

Figure 7 Schematic diagram of a high-dynamic-range pixel, consisting of a conventional APS/CMOS pixel, as illustrated in Figure 6, in combination with a shunt diode SD. The photocurrent lowers the voltage at the gate of transistor SF continuously. Below a certain value of this voltage, the diode starts to become forward biased, leading to a linear-logarithmic illumination response, with a dynamic range that can be as high as 200 dB.

Extension of the Spectral Sensitivity Range

The spectral sensitivity of silicon is essentially restricted to the spectral wavelength range between 0.1 and 1,100 nm. With a suitable choice of materials that cover the silicon sensor's surface, this sensitive wavelength range can be extended. For high-energy radiation, scintillator materials are employed, capable of converting incident high-energy photons into a proportional number of visible photons. Scintillators render smart image sensors particularly sensitive in the UV (300–400 nm), deep UV (below 300 nm), X-ray (0.002–1 nm), and gamma ray (below 0.002 nm) spectral region. Typically, the scintillator materials are fabricated as platelets that are glued on the surface of the produced CMOS/CCD image sensor.

The extension of the sensitive spectral range to the infrared can be achieved with semiconductor materials with smaller energy bandgaps than silicon. Examples include Ge, SiGe, SiC, PtSi, IrSi, InAs, InGaAs, or InGaAsP. Typically, these materials are deposited on the produced CMOS/CCD image sensor in an ultra-high vacuum environment. An alternative is the use of bump-bonding, with which the CMOS/CCD image sensor can be connected to a small-bandgap material chip at each pixel site. With these technologies, silicon-based smart pixel arrays become sensitive to wavelengths of several microns.

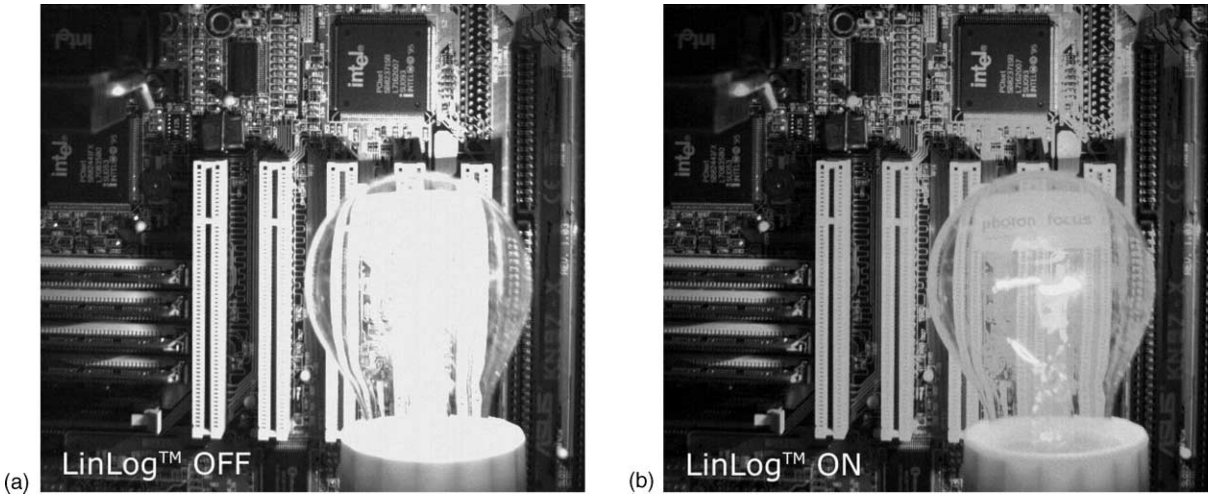

Figure 8 Images of a scene with high dynamic range, taken with an image sensor exhibiting linear illumination response (a) and linear-logarithmic illumination response (b). A similar pixel architecture as illustrated in Figure 7 is employed. (Courtesy of Photonfocus AG, Lachen, Switzerland.)

High-Speed Image Sensing

Increasing the speed of the photosensing process necessitates improvements in two key areas. Photocharges need to be collected and transported faster, and the acquired information must be read out faster.

A comparison of eqns [1] and [2] immediately reveals that the replacement of diffusion by drift transport provides the desired speed improvement. In the vertical direction, the electric field of photodiodes or MOS capacitances reaches more deeply into the semiconductor bulk if lower doping levels of the substrate are used or the bias voltage is increased. Improving the lateral transport speed is more difficult. A highly successful approach in so-called drift-field pixels is illustrated in Figure 9. On top of a thin (a few tens of nanometers) transparent insulator, a highly resistive layer – often of polysilicon – is placed. When a voltage difference is applied at the two ends, a small current flows across the resistive layer, inducing a linear potential variation at the semiconductor surface. A simple calculation shows that very small transit times for photocharges can be obtained in this way. Assuming a (quite large) pixel size of 25 μm and a voltage difference of 5 V, the maximum transit time of photoelectrons that have to move from one end of the pixel to the other, is less than 1 ns. This is more than two orders of magnitude faster than can be achieved with diffusion transport.

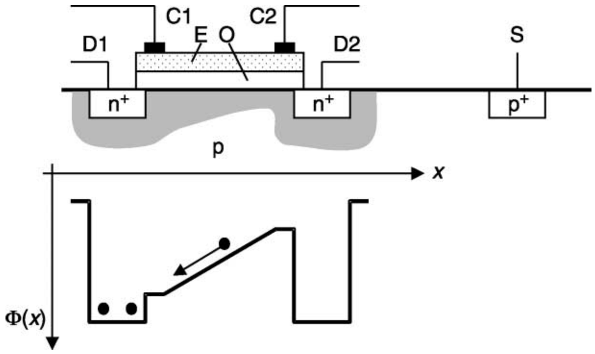

Figure 9 Cross-section of a drift field pixel, exploiting the lateral drift field at the semiconductor–oxide interface. The corresponding sloped potential distribution $\Phi(x)$ is created with a highly resistive transparent electrode E on top of a thin transparent insulator O; this conductor has two contacts C1 and C2 at its ends, to which a voltage difference is applied. Depending on the sign of the slope, the photoelectrons are collected either in diffusion D1 or D2. The hatched area indicates the space-charge region.

Smart pixel arrays provide a simple solution to the problem of speeding up the readout of the acquired picture information; a large degree of parallelism. This is illustrated with the architecture of a split-field high-speed image sensor, shown schematically in Figure 10. The photosensitive area is split into two parts, so that the column amplifiers of one half have time to settle to their proper voltages, while the columns of the other half are being read out. No time is lost, therefore, when another row of the image is selected for readout. A large number of output amplifiers are provided for highly parallel readout of the image data, each with a typical readout rate of 40–100 MHz. In this way, sustainable pixel rates of several GHz can be achieved in practice.

Static Spatial Demodulation

The spatially varying geometry of the photosensitive part of a pixel can be used for the static spatial

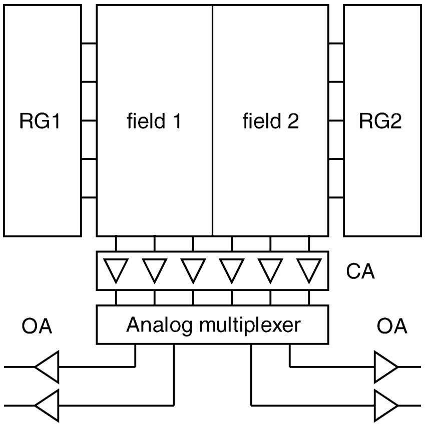

Figure 10 Block diagram of a high-speed image sensor, making use of the large amount of parallelism possible in smart pixel arrays. The light-sensitive area is split into two fields whose rows can be addressed independently with two separate row address generators RG. In this way, one row of one field can be selected and the corresponding column amplifiers CA can reach equilibrium, while the columns of the other field are being read out using several parallel output amplifiers OA.

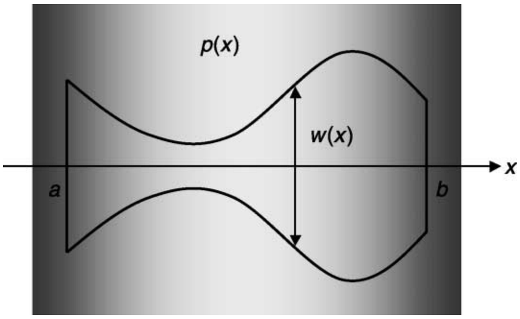

Figure 11 Illustration of a one-dimensional intensity distribution $p(x)$ of light which is incident on a shaped photodiode with position-dependent width $w(x)$, between the two boundaries a and b.

demodulation of an impinging light field. When a one-dimensional intensity distribution $p(x)$ is incident on a photodiode with spatially varying width $w(x)$ between the limits a and b, as illustrated in Figure 11, an output signal r is produced:

$$r = \int_a^b p(x)w(x)\mathrm{d}x \qquad [5]$$

By shaping the photodiode appropriately, any type of parametrized linear transform can be realized, as illustrated by the photomicrograph in Figure 12. In this practical example, four photodiodes are shaped sinusoidally and cosinusoidally, weighted with a slowly varying windowing function. Their signals are used for determining the Fourier coefficients (the sine and cosine transform) of an incident light distribution for a fixed spatial frequency. This is employed, for example, in cost-effective optical encoders, capable of measuring linear displacement with an accuracy below 50 nm.

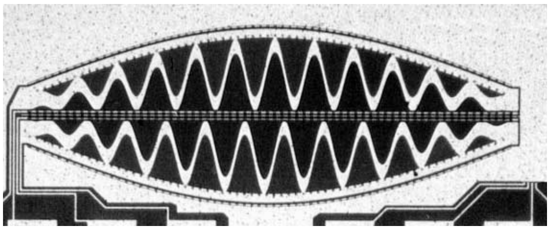

Figure 12 Photomicrograph of two photodiode pairs that are shaped as spatial sine and cosine functions, weighted with a slowly varying windowing function. The four photosignals can be employed to determine the phase and amplitude of a one-dimensional intensity signal that has the same spatial frequency as the sine and cosine. In practice, the accuracy of the phase determination is as high as 0.1% of the spatial wavelength. The width of the photomicrograph corresponds to about 1 mm in reality.

Dynamic Spatial Convolution

Many image-processing tasks require the convolution of the pixel data $P_{i,j}$ at the position (i, j) in an image with a two-dimensional convolution kernel $w_{k,l}$:

$$R_{k,l} = \sum_{i=-N}^{N} \sum_{j=-N}^{N} P_{i,j} w_{k-i,l-j} \qquad [6]$$

Instead of employing a purely geometrical solution, as described in the previous section, an alternate approach is found when interpreting the convolution eqn [6] in the following way. The resulting image R is obtained by displacing the original image P laterally by an amount (k,l), by multiplying the shifted pixel values with the corresponding weight $w_{k,l}$, and by adding up all these products.

The lateral displacement is realized with a CCD structure, implementing the 2D lateral movement of the acquired image charge packets, the multiplications with the convolution kernel values are realized by using various exposure times for each lateral shift position, proportional to the corresponding weight value, and the summation is just the natural charge accumulation of CCD structures. The problem of negative weights is solved by separately determining the positive and negative parts of the weight function, each stored separately in the pixel, and by calculating the difference of the two values during readout, either on-chip or off-chip. In this way, such convolution image sensors can be used for the determination of complete, programmable, real-time image convolutions, with arbitrary kernels that are fully under the dynamic control of digital logic or microprocessors. Practical experience with such convolution imagers shows that the convolution results exhibit inaccuracies of less than 1%.

Temporal Demodulation

In many relevant optical measurement techniques, temporally modulated light fields are employed. For this reason, smart pixel arrays are desired, providing synchronous demodulation of these signals. An electromagnetic signal $S(t)$ that is harmonic as a function of time with given frequency ω, having unknown modulation amplitude A, offset B and phase delay φ, is described by

$$S(t) = A \sin(\omega t + \varphi) + B \qquad [7]$$

The three unknowns A, B and φ can be measured if three or more sampling points are sampled synchronously per period in so-called 'lock-in' pixels. An example of a lock-in pixel employing the drift-field transport principle with four accumulation taps, is illustrated in Figure 13. Demodulation frequencies of several tens of MHz have already been demonstrated in practice.

The most important application of such synchronous demodulation pixels is in optical time-of-flight range cameras. Such a camera incorporates a light source, usually an array of infrared or visible LEDs, that is modulated at a few tens of MHz. The modulated light is reflected diffusely by the various objects, and the camera's lens images it onto a lock-in image sensor. Each of its pixels carries out synchronous sampling/accumulation of photocharges. At the end of the exposure time, the accumulated photocharge samples are read out, and the three modulation parameters of the reflected light are determined for each pixel. The phase shift is a direct measure of the local distance. It has been shown that for a wide range of operating conditions, the distance resolution of such optical range cameras is already limited just by the photo shot noise. A typical result of such a depth image is shown in Figure 14.

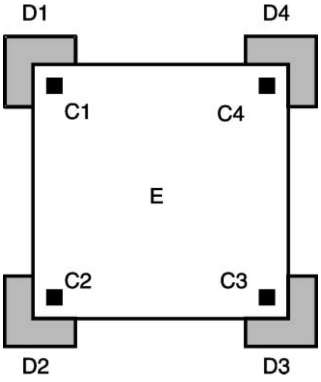

Figure 13 Schematic top view of a lock-in pixel based on the lateral drift-field approach illustrated in Figure 9. A square electrode is fabricated with a highly resistive transparent material on top of a thin insulator. The electrode contacts C1 to C4 are employed to create a 'reeling' potential distribution that rotates with the modulation frequency of the incident light that requires synchronous demodulation, so that the photocharges are collected at diffusions D1 to D4, respectively, at the appropriate time during each period.

Photonic Systems-on-chip

The microelectronics revolution is based on the key notion of integration. More functionality integrated monolithically on the same chip leads to faster, smaller, better performing, and cheaper products. The same is true for smart pixel arrays, for which progress through integration is attained in several respects.

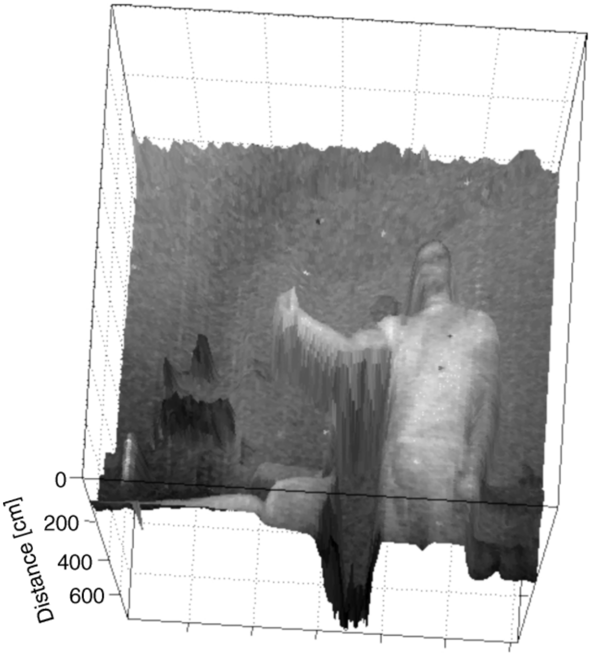

Figure 14 Example of a distance picture (a person) out of a 3D range image sequence, taken with an optical time-of-flight distance camera making use of lock-in pixels. Depending on the available light level, the background illumination and the modulation frequency, such solid-state range cameras can have a resolution of a few mm.

Seeing Chips

Already, in the mid-1980s, researchers were excited about the possibility of integrating local signal processing capabilities in each pixel, implementing some functions of biological vision systems. It was envisioned that an increasing amount of bio-inspired functionality would lead to smart image sensors, exhibiting a primitive sense of vision. The idea of a 'seeing chip', with the capability to interpret a scene, was born.

Unfortunately, even today, our understanding of the enormously successful biological vision systems has not yet reached a mature enough state allowing us to implement signal processing functions that approximate any 'image understanding'. Published algorithms are still incapable of interpreting a scene and 'perceiving' its contents. Obviously, no set of vision algorithms is yet known with which seeing chips could be realized. Although progress is slower than envisioned, single-chip machine vision systems of growing complexity and increasing functionality are being developed.

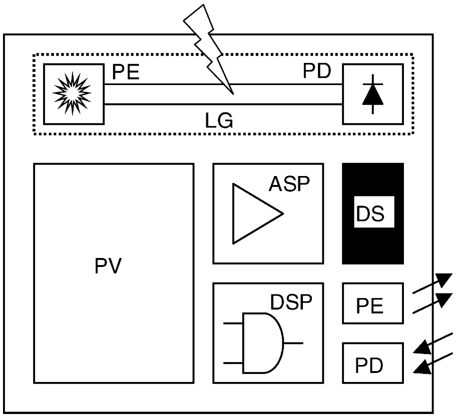

Figure 15 Schematic block diagram of a monolithically integrated photonic microsystem, that is expected to be producable in the near future with organic semiconductors. All optoelectronic devices are fabricated on the same planar surface: photovoltaic power supplies (PV); analog and digital signal processing (ASP and DSP); photodetectors (PD); photoemitters (PE) – used for bidirectional communication or as parts of a photonic detector subsystem – a light guiding structure (LG) that can be sensitized for specific sensing applications, and a display subsystem (DS).

Organic Semiconductors for Monolithic Photonic Microsystems

The realization that organic materials with conjugated double bonds, in particular noncrystalline polymers, are direct bandgap semiconductors, has recently led to a flurry of R&D activities in the domain of organic photonics. Organic light-emitting diodes (OLEDs), color displays, lighting panels, photodiodes, image sensors, photovoltaic cells, RF transponders, and tags, and several other organic photonic devices have not only been demonstrated but many are already commercially available. The special appeal of organic semiconductors stems from their potential for monolithic, low-cost integration of complete, large-area photonic microsystems, even on thin, flexible substrates. In Figure 15, such a monolithic photonic microsystem is illustrated schematically, combining a photovoltaic power supply, a photonic detector subsystem, analog and digital signal processing, as well as bidirectional optical communications with the outside world, possibly including also an on-chip display. The photonic detector subsystem typically consists of a light source, a light guiding structure that is sensitized – for example chemically – and a photodetector/image sensor subsection. In such a monolithic photonic microsystem, the term 'smart pixel array' obtains a much broader meaning, opening up a huge number of practical applications in essentially all domains of our daily and professional lives. This domain of organic semiconductors for integrated photonic microsystems is still largely unexplored, and it is offering exciting opportunities for research and product development.

Smart Pixel Arrays Beyond Electromagnetic Field Imaging

The sensing capabilities of smart pixel arrays are not restricted to electromagnetic radiation; this technology can also be used for the acquisition of other two- or three-dimensional physical fields. As an example, electrostatic imaging can be used for the temporally and spatially resolved acquisition of surface potential distributions in living cells. Electrostatic potential values of several hundred mV can be acquired with a precision of 1 mV and a temporal resolution below 1 millisecond. Other acquisition modalities include magnetostatic field imaging, for example by employing the Hall effect, or imaging of local electrostatic charge or current density distributions.

Summary

The relentless advances of semiconductor technology make it possible to integrate an increasing amount of photosensitive, analog, and digital functionality in the same smart pixel and on the same smart pixel array. Unusual photosensor devices can be fabricated with commercially available semiconductor processes, implementing innovative designs and layouts, as well as novel electronic circuits, with which the physical limits of photosensing and photocharge processing may be reached. Novel opto-electronic microsystems can be developed, based on application-specific smart pixel arrays, either by realizing known optical measurement techniques more efficiently or by implementing new techniques, whose realization requires the availability of nonconventional image sensing functionality.

Already, smart pixel arrays are opening up a large number of new opportunities in photonics research and product development. The continuously increasing levels of integration lead to new forms of system-on-chip solutions for a wide range of practical applications. Smart pixel arrays can perform at the basic limits imposed by physics, and they have become indispensable tools in almost any type of research laboratory and any field of science and engineering.

List of Units and Nomenclature

Bandwidth	B	s^{-1}
Boltzmann's constant	k	J/K
Capacitance	C	F
Charge	Q	C
Charge carrier mobility	μ	$m^2/(V\ s)$
Charge noise variance	s_Q	C^2
Diffusivity	D	m^2/s
Distance	L	m
Dynamic range	D/R	dB
Field effect transistor transconductance	g	A/V
Integrated power measured with shaped photodiode	r	W
Intensity of electromagnetic wave	S, A, B	W/m^2
One-dimensional light intensity distribution	$p(x)$	W/m^2
One-dimensional width distribution	$w(x)$	m
Phase delay	φ	–
Pixel value at position (i,j) in an image	$P_{i,j}; R_{i,j}$	–
Temperature	T	K
Temporal frequency	ω	s^{-1}
Time	t	s
Two-dimensional convolution kernel	$w_{k,l}$	–

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides. **Semiconductor Materials:** Group IV Semiconductors, Si/SiGe.

Further Reading

Forrest SR (2000) Active optoelectronics using thin-film organic semiconductors. *IEEE Journal on Selected Topics in Quantum Electronics* 6: 1072–1083.

Fossum ER (1997) CMOS image sensors: electronic camera-on-a-chip. *IEEE Transactions on Electron Devices* 44: 1689–1698.

Hosticka BJ, Seitz P and Simoni A (2004) Optical time-of-flight range imaging for solid-state 3D vision. In: Grimes G and Dickey EC (eds) *Encyclopedia of Sensors*. Stevenson Ranch, CA: American Scientific Publishers (in print).

Koch C (1989) Seeing chips: analog VLSI circuits for computer vision. *Neural Computation* 1: 184–200.

Lutz G (1999) *Semiconductor Radiation Detectors*. Berlin: Springer.

Mead C (1989) *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley and Sons.

Seitz P (1999) Solid-state image sensing. In: Jaehne B (ed.) *Handbook of Computer Vision and Applications*. London: Academic Press.

Seitz P and Blanc N (2003) CCD and APS/CMOS technology for smart pixels and image sensors. *Proceedings of SPIE* 5251 (in print).

Sze SM (1994) *Semiconductor Sensors*. New York: John Wiley and Sons.

Theuwissen AJP (1995) *Solid State Imaging with Charge Coupled Devices*. Dordrecht, Germany: Kluwer.

DIFFRACTION

Contents

Diffraction Gratings
Fraunhofer Diffraction
Fresnel Diffraction

Diffraction Gratings

J Turunen and T Vallius, University of Joensuu, Joensuu, Finland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The classic textbook example of a diffraction grating is a periodic arrangement of parallel opaque wires: if a plane wave is incident upon such a structure, it is divided into a multitude of plane waves propagating in well-defined directions. In particular, the dramatic wavelength-dependence of these directions has attracted great interest ever since the potential of diffraction gratings in spectroscopy was realized in the late 1700s. Today spectroscopy is only one, though important, application of diffraction gratings. Gratings of many different forms belong to the basic building blocks in modern wave-optics-based design of optical elements and systems.

Grating Equations

Some examples of the wide variety of possible grating structures are illustrated in Figure 1, which shows the profiles of some commonly encountered diffraction gratings that are invariant in the y direction. These are known as linear gratings. Each grating may be of reflection type (the refractive index n_{III} is complex) or of transmission type (n_{III} is real). In all cases n_1 and n_2 can both be real or one of them can be complex. Here we assume that the incident field is a plane wave with wave vector $\mathbf{k}$ in the $x-z$ plane; more general illumination waves can be represented as superpositions of plane waves.

Figure 1a illustrates a binary grating: in the modulated region $0 < z < H$ the refractive index alternates between n_1 and n_2, changing from one value to another at certain transition points within the period d. The structure in Figure 1b represents a sinusoidal surface-relief grating as an example of smooth grating profiles. Finally, Figure 1c shows a triangular grating profile.

The periodicity of a grating, illuminated by a plane wave, usually leads to the appearance of a set of reflected and transmitted plane waves (diffraction orders of the grating) that propagate in directions determined by the grating period d, the illumination wavelength λ, and the refractive indices n_{I} and n_{III} (see Figure 2a). These directions are obtained from the transmission grating equation:

$$n_{\mathrm{III}} \sin \theta_m = n_{\mathrm{I}} \sin \theta + m\lambda/d \qquad [1]$$

where m is the index of the diffraction order and θ_m is its propagation angle. For reflected orders (superscript r) one simply replaces n_{III} by n_{I} in eqn [1].

Obviously only a limited number of orders satisfy the condition $|\theta_m| < \pi/2$ and thereby represent conventional propagating plane waves. Higher, so-called evanescent orders are inhomogeneous waves that propagate along the grating surface and decay exponentially in the z direction. However, they cannot be ignored in grating analysis, as we shall see.

In the case of a biperiodic grating with a period $d_x \times d_y$ the propagation direction of the incident beam is defined by two angles θ and ϕ illustrated in Figure 2b. The grating equations now read as:

$$n_{\mathrm{III}} \sin \theta_{mn} \cos \phi_{mn} = n_{\mathrm{I}} \sin \theta \cos \phi + m\,\lambda/d_x \qquad [2]$$

$$n_{\mathrm{III}} \sin \theta_{mn} \sin \phi_{mn} = n_{\mathrm{I}} \sin \theta \sin \phi + n\lambda/d_y, \qquad [3]$$

where θ_{mn} and ϕ_{mn} define the propagation directions of the diffraction order with indices m and n. For reflected orders n_{III} is again replaced by n_{I}. These expressions apply also to the special case of linear gratings (limit $d_y \rightarrow \infty$) illuminated by a plane

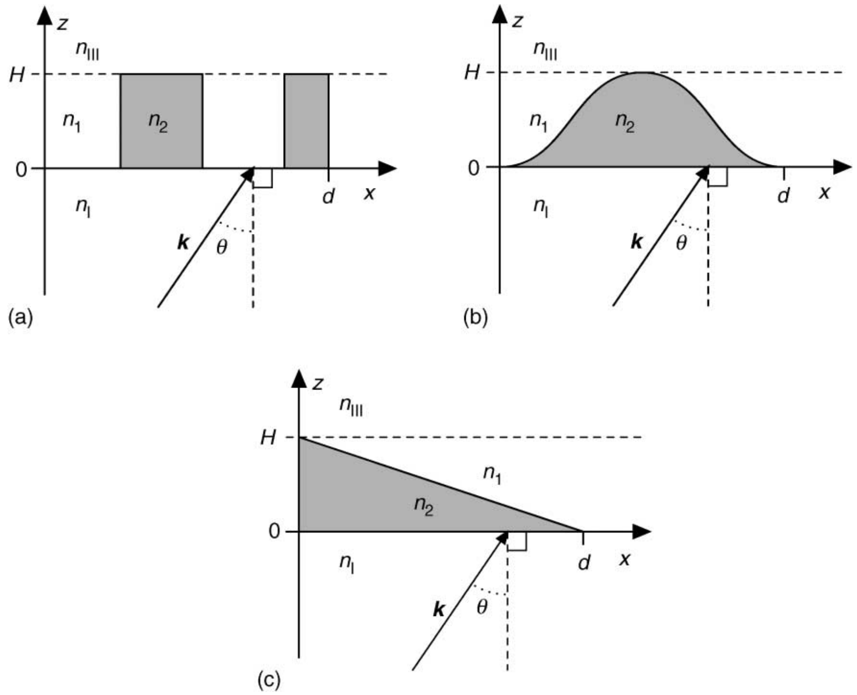

Figure 1 Examples of linear gratings. (a) Binary grating; (b) sinusoidal grating; (c) triangular grating. Here d is the grating period, H is the thickness of the modulated region, θ is the angle of incidence, and n denotes refractive index.

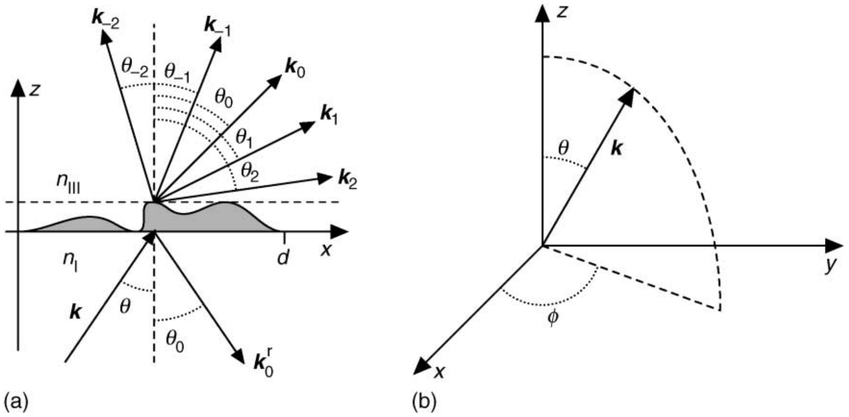

Figure 2 Diffraction by gratings. (a) Linear grating; (b) definitions of direction angles in the case of a biperiodic grating.

wave not propagating in the x–z plane (conical incidence).

Some two-dimensionally periodic gratings are illustrated in Figure 3. The structure illustrated in Figure 3a is an extension of a binary grating to the biperiodic geometry: here a single two-dimensional feature is placed within the grating period but, of course, more features with different outlines could be included as well. Moreover, each period could contain a continuous 'mountain-range' or consist of a 'New-York'-style pixel structure. Figure 3b illustrates a structure consisting of an array of rectangular-shaped holes pierced in a flat screen, while Figure 3c defines an array of isolated particles on a flat substrate. If the grid or the particles are metallic, surrounded by dielectrics from all sides, one speaks of inductive and capacitive structures, respectively: in the former case relatively free-flowing currents are induced in the metallic grid structure, but in the latter case the isolated metal particles become polarized (and therefore capacitive) when illuminated by an electromagnetic field.

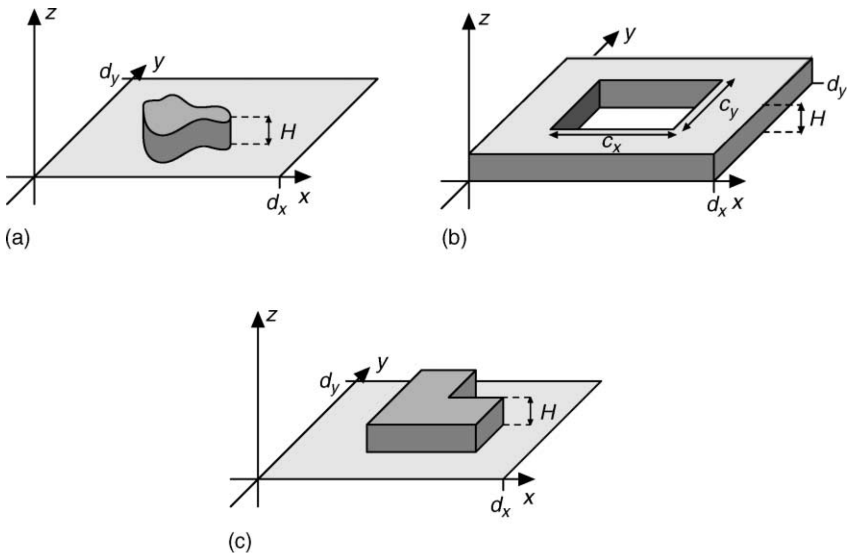

Figure 3 Examples of biperiodic gratings. (a) General binary grating; (b) inductive structure; (c) capacitive structure.

Grating Theory

Gratings equations provide exact knowledge of the propagation directions of the diffraction orders, but they give no information about the distribution of light among different orders. To obtain such information, one must in general solve Maxwell's equations exactly (rigorous diffraction theory), but within certain limits simpler methods are available. In all cases, however, the transmitted and reflected fields are represented as superpositions of plane waves that propagate in the directions allowed by the grating equations (these are known as Rayleigh expansions).

The most popular method for rigorous grating analysis is the Fourier Modal Method (FMM), though many alternative analysis methods exist. To understand the principle of FMM, consider first binary gratings. The field inside the modulated region $0 < z < H$ is represented as a superposition of waveguide modes, determined by imagining that the modulated region would extend from $x = -\infty$ to $x = \infty$. With given transition points the electric permittivity, its inverse, and the electric field inside the grating, can be represented in the form of Fourier series and one ends up with a matrix eigenvalue problem with finite dimensions when the matrices are truncated. A numerical solution of this problem provides a set of orthogonal modes that can propagate in the structure with certain propagation constants. The relative weights of the modes are then determined by applying the electromagnetic boundary conditions at the planes $z = 0$ and $z = H$ to match the modal expansions inside the grating with the plane-wave (Rayleigh) expansions of the fields in the homogeneous materials on both sides of the grating. This matching procedure, which is numerically implemented by solving a set of simultaneous linear equations, yields the complex amplitudes of the electric and magnetic field components associated with the reflected and transmitted diffraction orders of the grating.

Of main interest are the diffraction efficiencies, η_m (or η_{mn} for biperiodic gratings) of the orders, which tell the distribution of energy between different propagating orders. In electromagnetic theory the diffraction efficiency is defined as the z-component of the time-averaged Poynting vector associated with the particular order, divided by that of the incident plane wave.

Gratings with z-dependent surface profiles can be handled similarly if they are 'sliced' into a sufficient number of layers, in each of which the grating is modeled as a binary structure. The electromagnetic boundary conditions are applied at each slice boundary and the approach models the real continuous structure with arbitrary degree of accuracy when the number of the slices is increased; in practice it is usually sufficient to use approximately 20–30 slices for sufficiently accurate modeling of any continuous grating profile by FMM.

The main problem with rigorous techniques such as FMM is bad scaling. If, in the case of 1D modulated (or y-invariant) gratings, the period is increased by a factor of two, the computation time required to solve the eigenvalue problem is increased by a factor of around eight. Thus we have a d^3 type scaling law. The situation is much worse for biperiodic gratings, for which the exponent is around six. Thus no foreseeable developments in computer technology will solve the problem, and approximate analysis methods have to be sought. A multitude of such methods are under active development, but in this context we discuss the simplest example only.

If $d \gg \lambda$ and the grating profile does not contain wavelength-scale features, one can describe the effects of a grating on the incident field merely by considering it as a complex-amplitude-modulating object, which imposes a position-dependent phase shift and amplitude transmission (or reflection) coefficient on the incident field. This model is known as the thin-element approximation (TEA).

At normal incidence the plane-wave expansion of any scalar field component U immediately behind the grating (at $z = H$) takes the form:

$$U(x, y, H) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{mn} \times \exp[i2\pi(mx/d_x + ny/d_y)] \qquad [4]$$

where

$$T_{mn} = \frac{1}{d_x d_y} \int_0^{d_x} \int_0^{d_y} U(x, y, H) \times \exp[-i2\pi(mx/d_x + ny/d_y)]\mathrm{d}x\,\mathrm{d}y \qquad [5]$$

are the complex amplitudes of the transmitted diffraction orders. In TEA the field at $z = H$ may be connected to the field at $z = 0$ (a unit-amplitude plane wave is assumed here) by optical path integration along a straight line though the modulated region of the grating:

$$U(x, y, H) = \int_0^H \exp\left[i\frac{2\pi}{\lambda}\hat{n}(x, y, z) \right] \mathrm{d}z \qquad [6]$$

where $\hat{n}(x, y, z) = n(x, y, z) + i\kappa(x, y, z)$ is the complex refractive-index distribution in the region $0 < z < H$. With TEA the diffraction efficiencies can be calculated from a simple formula:

$$\eta_{mn} = |T_{mn}|^2 \qquad [7]$$

which, however, ignores Fresnel reflections.

Equations [5]–[7] allow one to obtain closed-form expressions for the diffraction efficiencies η_{mn} in many grating geometries. For example, a y-invariant triangular profile with $H = \lambda/|n_{\mathrm{I}} - n_{\mathrm{III}}|$ gives $\eta_{-1} = 100\%$ if $n_1 = n_{\mathrm{III}}$, $n_2 = n_{\mathrm{I}}$, and $\theta = 0$. For a Q-level stair-step-quantized version of this grating we obtain:

$$\eta_m = \sin c^2(1/Q) \qquad [8]$$

where $\sin cx = \sin(\pi x)/(\pi x)$. It should be noted, however, that these results are valid only for gratings with $d \gg \lambda$, even if they are corrected by Fresnel transmission losses at the interface between the two media.

Fabrication of Gratings

Gratings can be manufactured in a number of different ways, perhaps the simplest being to plot an equidistant array of straight black and white lines and to reduce the scale of the pattern photographically. This method can be readily adapted to the generation of more complex grating structures and diffractive elements, but it is not suitable for the production of high-quality gratings required in, for example, spectroscopy.

Ruling engines represent the traditional way of making gratings with periods of the order of the wavelength of light, especially for spectroscopic applications. These are machines equipped with a triangularly shaped diamond ruling edge and an interferometrically controlled two-dimensional mechanical translation stage that moves the substrate with respect to the ruling edge. Triangular-profile gratings of high quality can be produced with such machines, but the fabrication process is slow and the cost of the manufactured gratings is therefore high. In addition, even with real-time interferometric control of the ruling-edge position, it is impossible to avoid small local errors in the grating geometry. Such errors are particularly harmful if they are periodic since, according to the grating equation, these errors will generate spurious spectral lines known as ghosts. Such lines generated by the grating 'super-period' are weak but they may be confused with real spectral lines in the source spectrum. If the ruling errors are random, they result in 'pedestal-type' background noise. While such errors do not introduce ghosts, they nevertheless reduce the visibility of weak original spectral lines.

The introduction of lasers in the early 1960s made it possible to form stable and high-contrast interference patterns over a sufficient time-scale to allow the exposure of photographic films and other photosensitive materials such as polymeric photoresists. This interferometric technique offers a direct method for the fabrication of large diffraction gratings without appreciable periodic errors. With this technique it is possible to make highly regular gratings for spectroscopy, but it does not permit easy generation of arbitrarily shaped grating structures. The two intersecting plane waves can also be generated with a linear grating using diffraction orders $m = +1$ and $m = -1$, suppressing the zero order by appropriate design of the grating profile and choosing the period in such a way that all higher transmitted orders are evanescent. This so-called phase mask technique is suitable, for example, for the exposure of Bragg gratings in fibers.

To generate arbitrary gratings one usually employs lithographic techniques. Some of the possible processes are illustrated in Figure 4. The exposure can be performed with scanning photon, electron, or ion beams, etc. The process starts with substrate preparation, i.e., coating it with a beam-sensitive layer known as resist, and in the case of electron beam exposure with a thin conductive layer to prevent charging. Following the exposure the resist is developed chemically, and in this process either the exposed or nonexposed parts of the resist are dissolved (depending on whether the resist is of positive or negative type), leading to a surface-relief profile. Several alternative fabrication steps may follow the development, depending on the desired grating type. If a high-contrast resist is used, a slightly 'undercut' profile as illustrated in Figure 4c is obtained. This can be coated with a metal layer and the resist can be 'lifted off' chemically. Thus a binary metallic grating is formed. To produce a binary dielectric profile one bombards the substrate with ions either purely mechanically (ion beam milling) or with chemical assistance (reactive ion etching), and finally the residual metal is dissolved.

Alternatively, the particle beam dose can be varied with position to obtain an analog profile in the resist after development. This resist profile can be transformed into the substrate by proportional etching, which leads to an analog dielectric surface-relief structure (see Figure 4f). This profile can be made

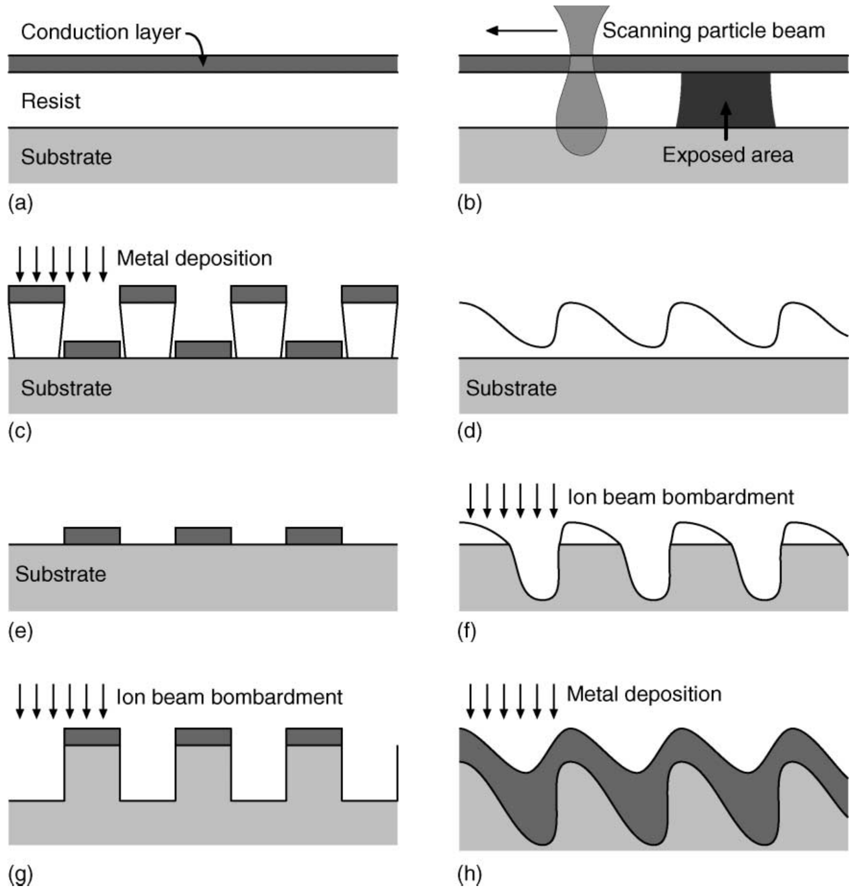

Figure 4 Lithographic processes for grating fabrication. (a) Substrate after preparation for electron beam exposure; (b) exposure by a scanning electron beam; (c) slightly undercut profile in hight-contrast resist after development, under metal film evaporation; (d) analog resist profile after variable-dose exposure and resist development; (e) finished binary metallic grating after resist lift-off; (f) transfer of the analog surface-relief profile into the substrate by proportional etching; (g) etching of a binary profile into the substrate with a metal mask; (h) growth of a metal film on an analog surface profile.

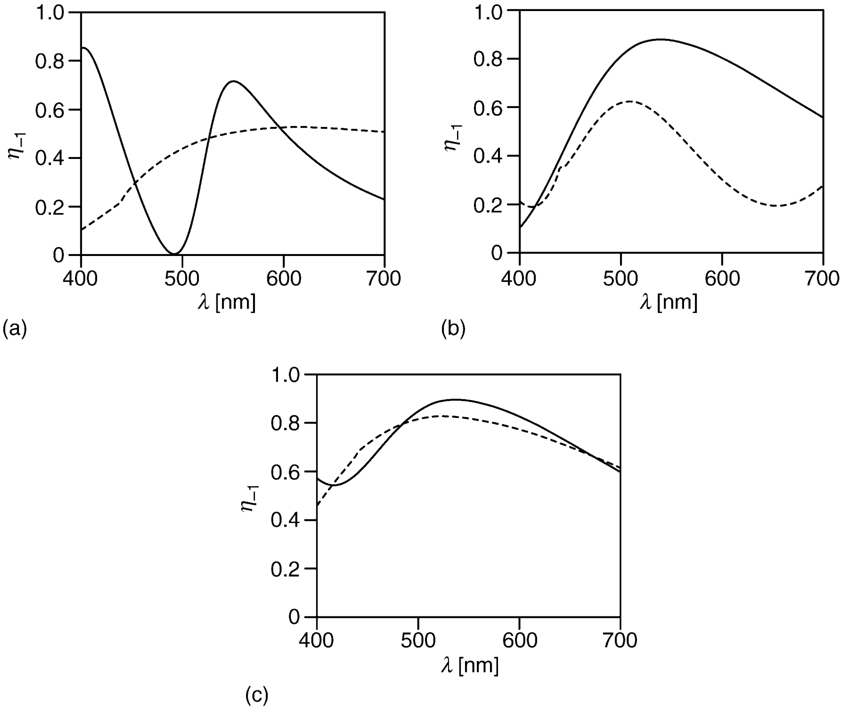

Figure 5 Spectral diffraction efficiency curves of spectroscopic gratings in TE polarization. (a) Binary; (b) sinusoidal; and (c) triangular reflection gratings illuminated at Bragg angle. Solid lines: $d = \lambda_0$. Dashed lines: $d = 2\lambda_0$.

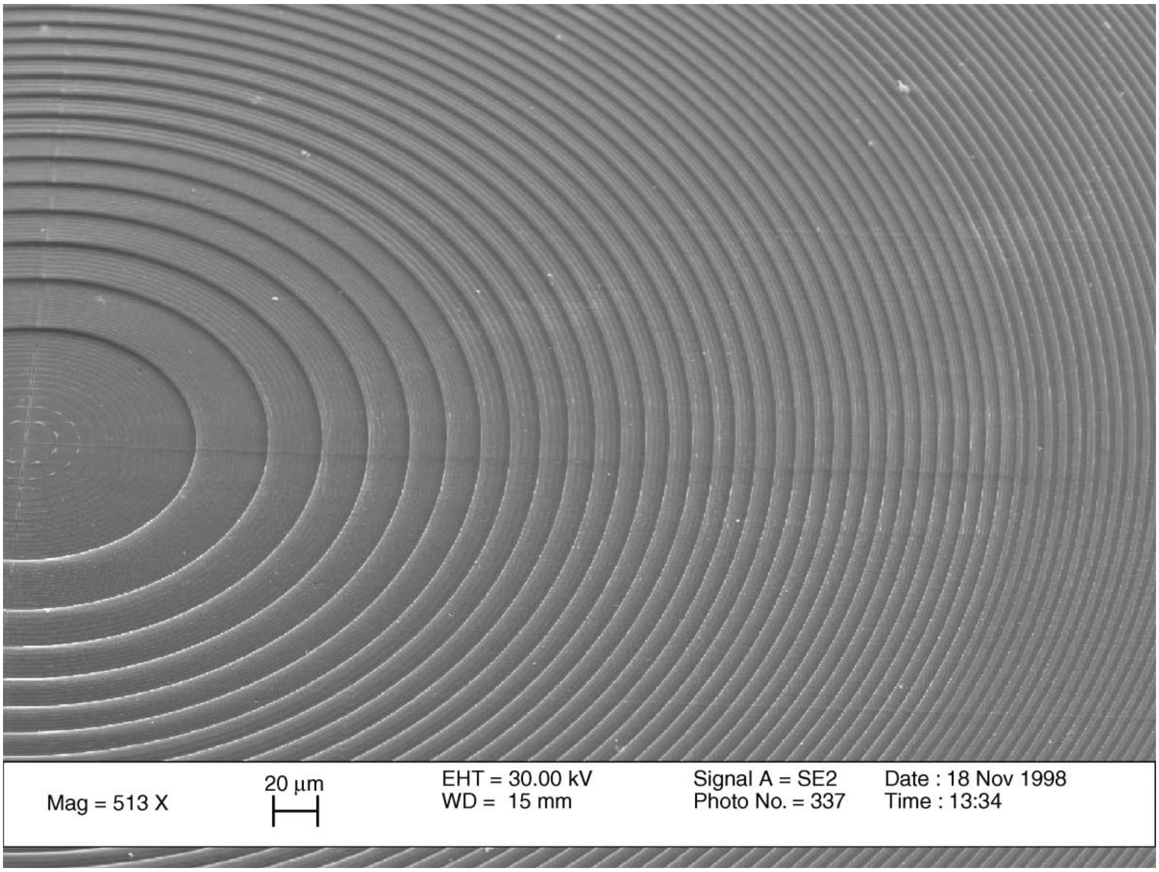
Mag = 513 X
20 μm
EHT = 30.00 kV
WD = 15 mm
Signal A = SE2
Photo No. = 337
Date : 18 Nov 1998
Time : 13:34

Figure 6 A scanning electron micrograph of a part of a diffractive lens.

reflective by depositing a thin metal film on top of it. If large-scale fabrication of identical gratings is the goal, a thick (~100 μm) metal layer (for example nickel) can be grown on the surface by electroplating. After separation from the substrate, the electroplated metal film, which is a negative of the original profile, can be used as a stamp in replicating the original profile in plastics using methods such as hot embossing, ultraviolet radiation curing, and injection molding. In particular, the latter method provides the key to high-volume and low-cost production of various types of diffraction gratings and other grating-based diffractive structures in the industrial scale.

Some Grating Applications

We now proceed to describe a collection of representative grating applications. More exhaustive coverage of grating applications can be found in the Further Reading section at the end of this article.

Spectroscopic Gratings

The purpose of a spectroscopic grating is to divide the incident light into a spectrum, i.e., to direct different frequency components of the incident beam into different directions given by the grating (eqn [1]). When the grating period is reduced, the angular dispersion $\partial\theta_m/\partial\lambda$, and therefore the resolution of the spectrograph, increases.

In the numerical examples presented in Figure 5 we assume a reflection grating in a Littrow mount, i.e., $\theta_m^r = -\theta$. This mount is common in spectroscopy; it obviously requires that the grating is rotated when the spectrum is scanned. We see from eqn [1] that in the Littrow mount:

$$\sin\theta = \frac{m\lambda}{2d} \qquad [9]$$

By inserting this result back into eqn [1] and differentiating we obtain a quantitative measure for dispersion in the Littrow mount:

$$\frac{\partial\theta_m}{\partial\lambda} = -\frac{m}{2d\cos\theta} \qquad [10]$$

We consider TE polarization (electric field is linearly polarized in the y-direction) and order $m = -1$, optimize H to give maximum efficiency at $\lambda_0 = 550$ nm, and plot the diffraction efficiency η_{-1} over the visible wavelength range for two values of the grating period, $d = \lambda_0$ and $d = 2\lambda_0$. Rigorous diffraction theory is used with $n_I = n_2 = 1$, and the metal is assumed to be aluminum (n_I and n_{III} are complex-valued).

In Figure 5a binary gratings with one groove (width $d/2$) per period is considered. Parametric optimization gives $H = 422$ nm if $d = 1\lambda$, and $H = 168$ nm if $d = 2\lambda$. Figure 5b illustrates the results for sinusoidal gratings with $H = 345$ nm if $d = 1\lambda$, and

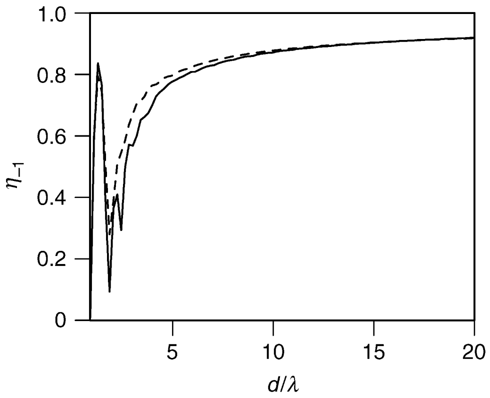

Figure 7 Diffraction efficiency versus wavelength for triangular transmission gratings illuminated at normal incidence. Solid lines: TE polarization. Dashed lines: TM polarization.

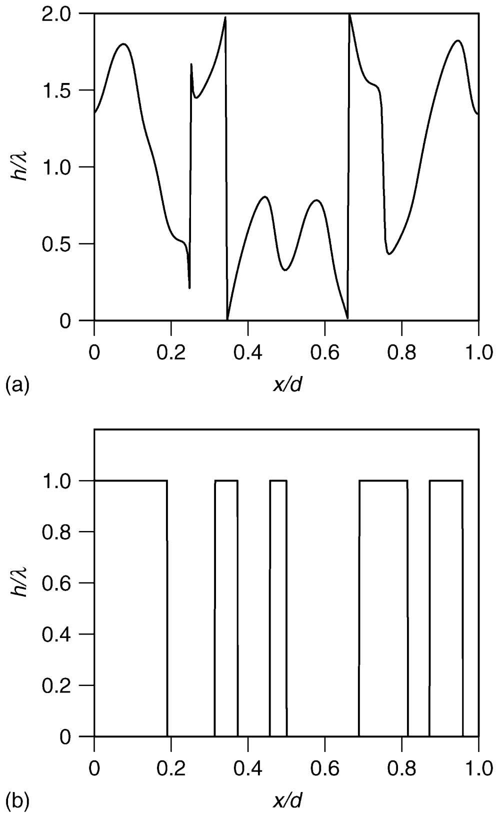

Figure 8 Surface-relief profiles of (a) continuous and (b) binary 1 → 8 beamsplitter gratings.

$H = 520$ nm if $d = 2\lambda$. Finally, triangular gratings with $H = 431$ nm if $d = 1\lambda$, and $H = 316$ nm if $d = 2\lambda$, are considered. The triangular grating profile is seen to be the best choice in general.

Diffractive Lenses

Diffractive lenses are gratings with locally varying period and groove orientation. Figure 6 illustrates the central part of a diffractive microlens fabricated by lithographic methods. It is seen to consist of concentric zones with radially decreasing width.

The phase transmission function of an ideal diffractive focusing lens can be obtained by a simple optical path calculation, and is given by

$$\phi(r) = \frac{2\pi}{\lambda_0}\left(f - \sqrt{f^2 + r^2}\right) \approx -\frac{\pi}{\lambda_0 f} r^2 \qquad [11]$$

where f is the focal length, r is the radial distance from the optical axis, and the approximate expression is valid in the paraxial region $r \ll f$. Here λ_0 is the design wavelength, which determines the modulation depth of the locally blazed grating profile:

$$H = \frac{\lambda_0}{n(\lambda_0) - 1} \qquad [12]$$

and $n(\lambda_0)$ is the refractive index of the dielectric material at $\lambda = \lambda_0$.

The zone boundaries r_n are defined by the condition $\phi(r_n) = -2\pi n$, and therefore we obtain from eqn [11]:

$$r_n = \sqrt{2nf\lambda_0 + (n\lambda_0)^2} \approx \sqrt{2nf\lambda_0} \qquad [13]$$

Obviously, when the numerical aperture of the lens is large, the zone width in the outer regions of the lens is reduced to values of the order of λ. Figure 7 illustrates the local diffraction efficiency in different part of a diffractive lens: we use the order $m = -1$ of a locally triangular diffraction grating of the

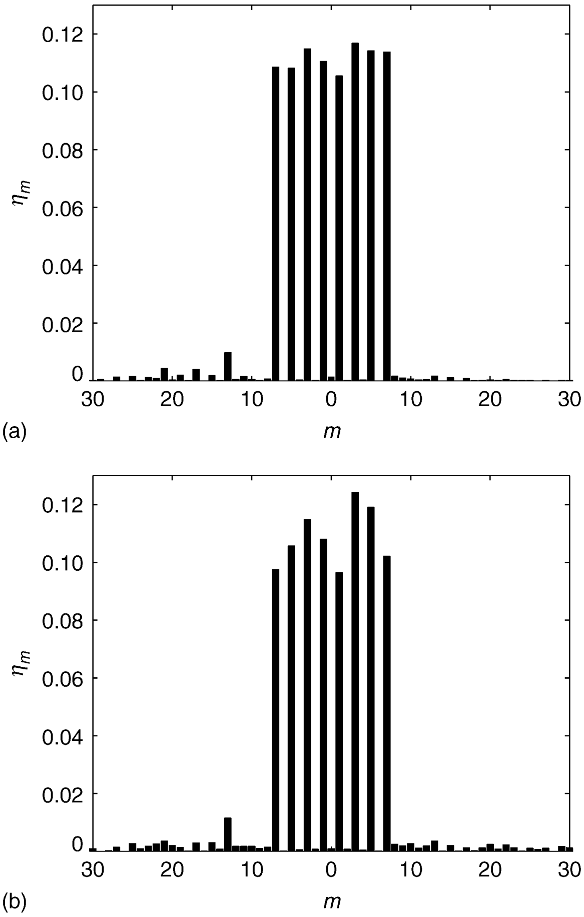

Figure 9 Diffraction pattern produced by a continuous-profile beamsplitter grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

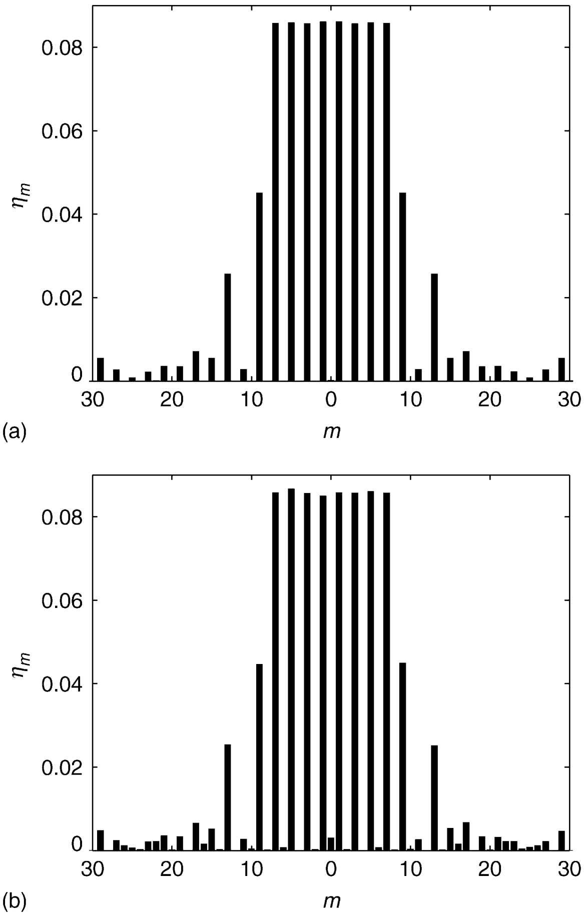

Figure 10 Diffraction pattern produced by a binary beamsplitter grating grating when (a) $d = 200\lambda$ and (b) $d = 50\lambda$.

type shown in Figure 1c with $\theta = 0$, assuming that $n_{\mathrm{I}} = n_2 = 1.5$ (fused silica) and $n_1 = n_{\mathrm{III}} = 1$. The efficiency η_{-1} is calculated as a function of λ using rigorous diffraction theory for both TE and TM polarization (the electric vector points in the groove direction in the TE case and is perpendicular to it in the TM case). It is seen that the local efficiency of the lens becomes poor in regions where the local grating period is small, and thus diffractive lenses operate best in the paraxial region (it is, however, possible to obtain improved nonparaxial performance by optimizing the local grating profile).

A factor that limits the use of diffractive lenses with polychromatic light is their large chromatic aberration: in the paraxial limit the focal length depends on the wavelength according to the formula:

$$f(\lambda) = \frac{\lambda_0}{\lambda} f \qquad [14]$$

However, since the dispersion of a grating has an opposite sign compared to the diffraction of a prism, a weak positive diffractive lens can be combined with a relatively strong positive refractive lens to form a single-element achromatic lens. Such a hybrid lens is substantially thinner and lighter than a conventional all-refractive achromatic lens made of crown and flint glasses.

Beam Splitter Gratings

Multiple beamsplitters, also known as array illuminators, are gratings with sophisticated periodic structure that are capable of transforming an incident plane wave into a set of diffraction orders with a specified distribution of diffraction efficiencies. Most often one wishes to obtain a one- or two-dimensional array of diffracted plane waves with equal efficiency. This goal can be achieved by appropriate design of the grating profile, which may be of binary, multilevel, or continuous form.

Figure 8 illustrates binary and continuous grating profiles capable of dividing one incident plane wave into eight diffracted plane waves (central odd-numbered orders of the grating) with equal efficiency in view of TEA. Figure 9 shows the distribution of energy among the different diffraction orders in and around the array produced by the continuous profile, evaluated by rigorous theory for two different grating periods, while Figure 10 gives corresponding results for the binary grating.

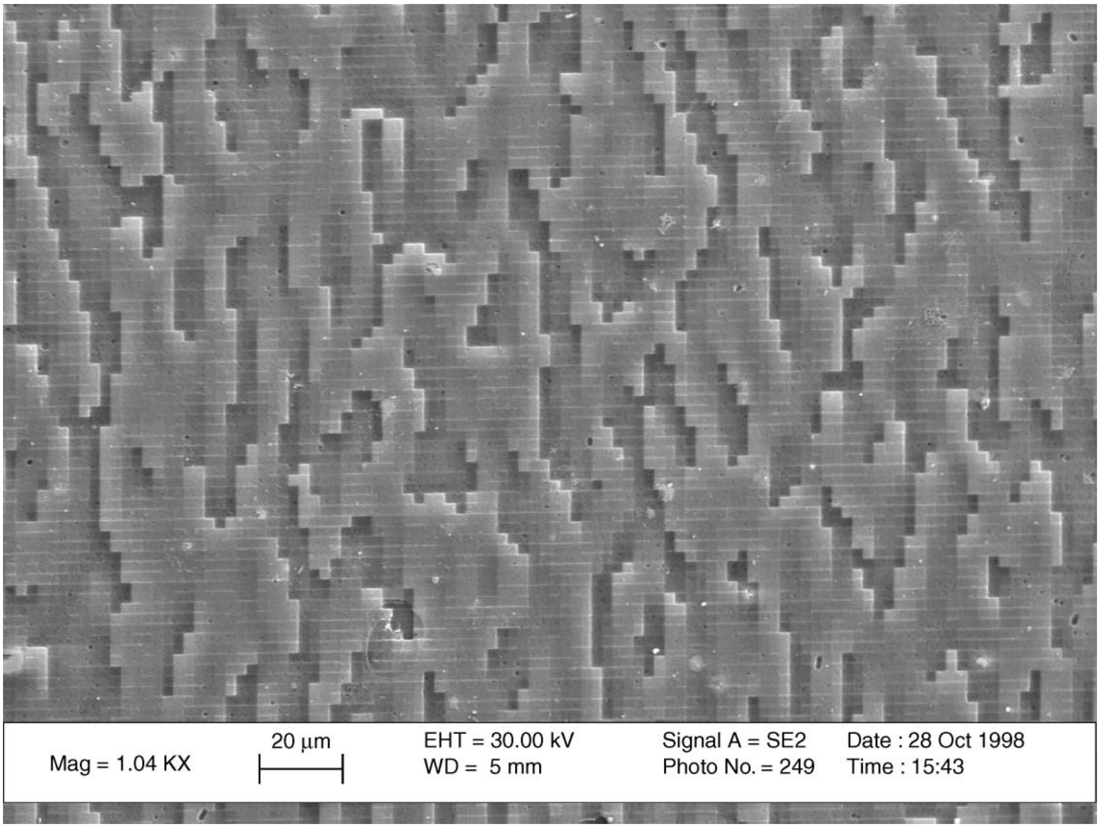

Figure 11 Scanning electron micrograph of a sophisticated beamsplitter grating.

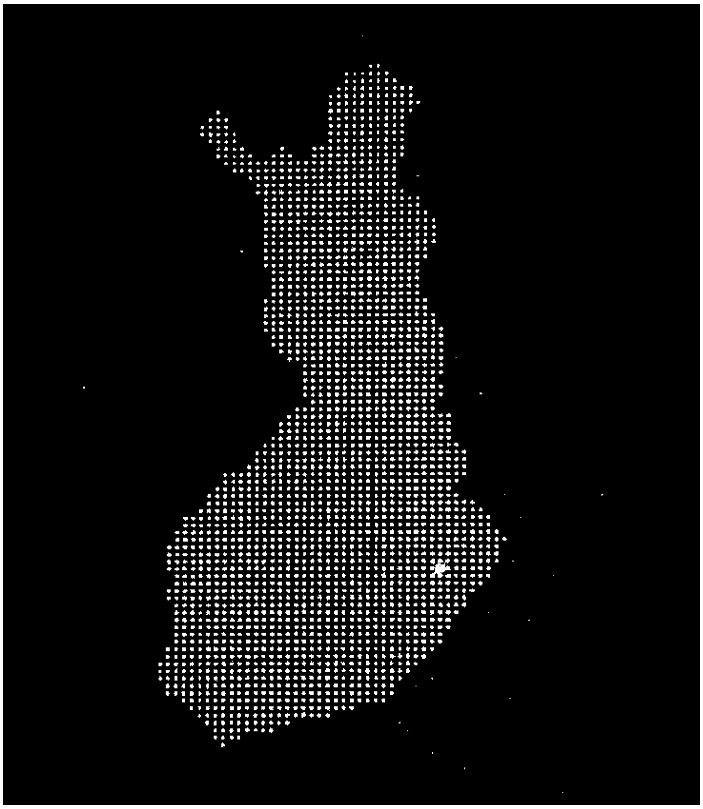

Figure 12 The far-field diffraction pattern produced by the grating in Figure 11.

The fraction of incident energy that ends up in the desired eight orders is 72% for the binary grating and 96% for the continuous grating according to TEA, making the latter more attractive from a theoretical viewpoint. However, the continuous profile is more difficult to fabricate accurately than the binary one, especially for small values of the ratio d/λ. Moreover, the array uniformity produced by the continuous profile degrades faster than that of the binary profile when the period is reduced.

An example of the structure of a biperiodic grating that is capable of forming a large, shaped array of diffraction orders with equal efficiency is presented in Figure 11. Figure 12 shows the far-field diffraction pattern produced by this element: the shape of Finland with the location of the town of Joensuu highlighted by placing the zero diffraction order on the appropriate position on the map (the apparent nonuniformity in the array is mostly due to the finite resolution of the CCD detector).

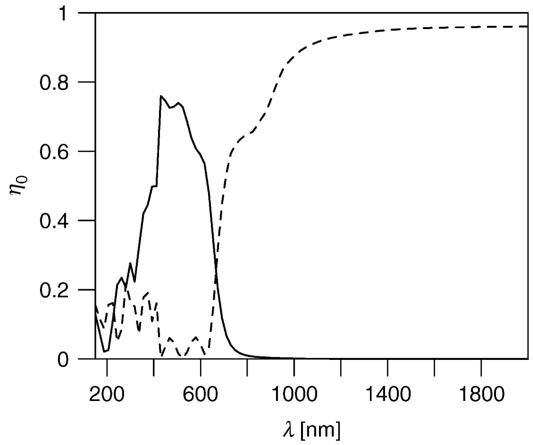

Figure 13 Zero order spectral transmission (solid line) and reflection (dashed line) of an aluminium inductive grid filter with 400 nm period.

Inductive Grid Filters

As a final example we consider a scaled-down version of a device well known from the door of a microwave oven: one can see in through the holes pierced in the metallic screen in the oven door, but the microwaves are trapped inside because their wavelengths are large in comparison with the dimensions of the holes. Another example of these inductive grid filters is found in large radio telescope mirrors for centimeter-scale wavelengths, which are wide-grid structures rather than smoothly curved metal surfaces.

Figure 13 illustrates the spectral transmission and reflection curves of an inductive grid filter of the type illustrated in Figure 3b. We assume that $d_x = d_y = d = 400$ nm, $c_x = c_y = c = 310$ nm, $H = 500$ nm, and that the grid is a self-supporting structure ($n_\mathrm{I} = n_\mathrm{III} = 1$) made of aluminum. Only the zero reflected and transmitted orders propagate when $\lambda > d$, but other orders (with rather low efficiency) appear for smaller wavelengths. We see that the filter reflects infrared radiation effectively, but transmits a substantial part of radiation at visible and shorter wavelength regions.

List of Units and Nomenclature

Complex field amplitude	U
Diffraction angles [rad]	θ, ϕ
Diffraction efficiency	η
Focal length [m]	f, f_0
Grating period [m]	d
Grating thickness [m]	H
Number of quantization levels	Q
Phase function [rad]	$\phi(r)$
Refractive index	n
Wavelength [m]	λ, λ_0
Wave vector [m^{-1}]	$\mathbf{k}$

See also

Fiber Gratings.

Further Reading

Chandezon J, Maystre D and Raoult G (1980) A new theoretical method for diffraction gratings and its numerical application. *Journal of Optics* 11: 235–241.

Gaylord TK and Moharam MG (1985) Analysis and applications of optical diffraction by gratings. *Proceedings of IEEE* 73: 894–937.

Herzig HP (ed.) (1997) *Micro-optics: Elements, Systems and Applications*. London: Taylor & Francis.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Li L (2001) Mathematical reflections on the fourier modal method in grating theory. In: Bao G, Cowsar L and Masters W (eds) *Mathematical Modelling in Optical Science*, pp. 111–139. Philadelphia: SIAM.

Li L, Chandezon J, Granet G and Plumey J-P (1999) Rigorous and efficient grating-analysis method made easy for optical engineers. *Applied Optics* 38: 304–313.

Petit R (ed.) (1980) *Electromagnetic Theory of Gratings*. Berlin: Springer-Verlag.

Rayleigh JWS (1907) On the dynamical theory of gratings. *Proceeding of the Royal Society A* 79: 399.

Solymar L and Cooke DJ (1981) *Volume Holography and Volume Gratings*. London: Academic Press.

Turunen J, Kuittinen M and Wyrowski F (2000) Diffractive optics: electromagnetic approach. In: Wolf E (ed.) *Progress in Optics XL*. Amsterdam: Elsevier.

Turunen J and Wyrowski F (eds) (1997) *Industrial and Commercial Applications of Diffractive Optics*. Berlin: Akademie Verlag.

van Renesse RL (1993) *Optical Document Security*. Boston, MA: Artech House.

Fraunhofer Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Geometrical optics does a reasonable job of describing the propagation of light in free space unless that light encounters an obstacle. Geometrical optics predicts that the obstacle would cast a sharp shadow. On one side of the shadow's edge would be a bright uniform light distribution, due to the incident light, and on the other side there would be darkness. Close examination of an actual shadow edge reveals, however, dark fringes in the bright region and bright fringes in the dark region. This departure from the predictions of geometrical optics was given the name diffraction by Grimaldi.

There is no substantive difference between diffraction and interference. The separation between the two subjects is historical in origin and is retained for pedagogical reasons. Since diffraction and interference are actually the same physical process, we expect the observation of diffraction to be a strong function of the coherence of the illumination. With incoherent sources, geometrical optics is usually all that is needed to predict the performance of an optical system and diffraction can be ignored except at dimensions on the scale of a wavelength. With light sources having a large degree of spatial coherence, it is impossible to neglect diffraction in any analysis.

Even though diffraction produced by most common light sources and objects is a small effect, it is possible to observe diffraction without special equipment. By viewing a light source through a slit formed by two fingers that are nearly touching, fringes can be observed. Diffraction is used by photographers and television cameramen to provide artistic highlights in an image containing small light sources. They accomplish this by placing a screen in front of the camera lens. Light from point sources in the field of view is diffracted by the screen and produces 'stars' in the image. A third place where diffraction effects are easily observed is when a mercury or sodium streetlight, a few hundred meters away, is viewed through a sheer curtain.

Fresnel originally developed an approximate scalar theory based on Huygens' principle which states that:

> Each point on a wavefront can be treated as a source of a spherical wavelet called a secondary wavelet or a Huygens' wavelet. The envelope of these wavelets, at some later time, is constructed by finding the tangent to the wavelets. The envelope is assumed to be the new position of the wavefront.

Rather than simply using the wavelets to construct an envelope, Fresnel's scalar theory assumes that the Huygens' wavelets interfere to produce a new wavefront (*see* **Diffraction:** Fresnel Diffraction).

A rigorous diffraction theory is based on Maxwell's equations and uses the boundary conditions associated with the obstacle to calculate a field scattered by the obstacle. The origins of this scattered field are currents induced in the obstacle by the incident field. The scattered field is allowed to interfere with the incident field to produce a resultant diffracted field. The application of the rigorous theory is very difficult and for most problems an approximate scalar theory developed by Kirchhoff and Sommerfeld is used.

Kirchhoff Theory of Diffraction

Kirchhoff's theory was based on the elastic theory of light but can be reformulated into a vector theory. Here we will limit our discussion to the scalar formulation. Given an incident wave, φ, we wish to calculate the optical wave at point P_0, in Figure 1, located at $\mathbf{r}_0$, in terms of the wave's value on a surface, S, that we construct about the observation point.

To obtain a solution to the wave equation at the point, P_0, we select as a Green's function the one proposed by Kirchhoff (Box 1), a unit amplitude spherical wave, expanding about the point at $\mathbf{r}_0$, denoted by Ψ, and then apply Green's theorem (see Box 2). We will run into boundary condition problems as we proceed with the derivation of the solution. Sommerfeld removed the problems by assuming a slightly more complex Green's function (see Box 1). We will limit our discussion to the Kirchhoff Green's function.

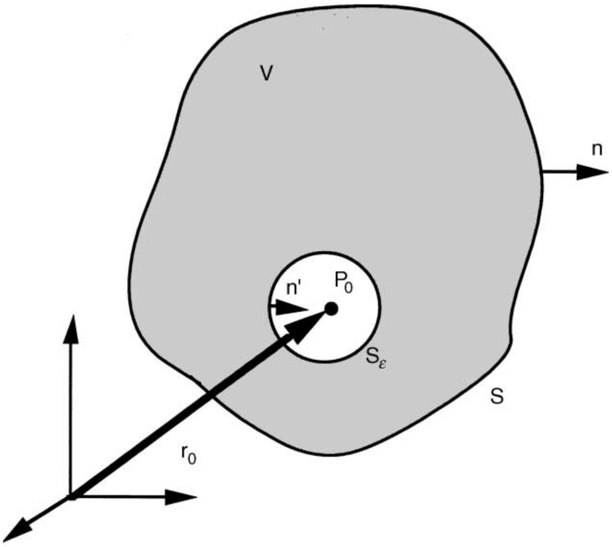

Figure 1 Region of integration for solution of Kirchhoff's diffraction integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Green's theorem requires that there be no sources (singularities) inside the surface S but our Green's function [1] is based on a source at r_0. We eliminate the source by constructing a small spherical surface S_ε, of radius ε, about $\mathbf{r}_0$, excluding the singularity at $\mathbf{r}_0$ from the volume of interest, shown in gray in Figure 1. The surface integration that must be performed in Green's theorem is over the surface $S' = S + S_\varepsilon$.

Within the volume enclosed by S', the Green's function, Ψ, and the incident wave, φ, satisfy the scalar Helmholtz equation so that the volume integral can be written as

$$\iiint_V (\Psi\nabla^2\varphi - \varphi\nabla^2\Psi)\,d\nu = -\iiint_V (\Psi\varphi k^2 - \varphi\Psi k^2)\,d\nu \qquad [7]$$

The right side of [7] is identically equal to zero. This fact allows us to use [6] to produce a simplified statement of Green's theorem:

$$\iint_{S'} \left(\Psi\frac{\partial\varphi}{\partial n} - \varphi\frac{\partial\Psi}{\partial n}\right) ds = 0 \qquad [8]$$

Box 1

Green's Function

1. **Kirchhoff Green's function**: Assume that the wave, Ψ, is due to a single point source at $\mathbf{r}_0$. At an observation point, $\mathbf{r}_1$, the Green's function is given by

$$\Psi(\mathbf{r}_1) = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}}, \qquad [1]$$

   where $r_{01} = |\mathbf{r}_{01}| = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from the source of the Green's function, at $\mathbf{r}_0$, to the observation position, at $\mathbf{r}_1$.
2. **Sommerfeld Green's function**: Assume that there are two point sources, one at P_0 and one at P_0' (see Figure 2 where the source, P_0', is positioned to be the mirror image of the source at P_0 on the opposite side of the screen, thus

$$r_{01} = r_{01}' \qquad \cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) = -\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}') \qquad [2]$$

   A Green's function that could be used is

$$\Psi' = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} - \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}'}}{r_{01}'} \qquad [3a]$$

   or

$$\Psi' = \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}}{r_{01}} + \frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}'}}{r_{01}'} \qquad [3b]$$

Box 2

Green's Theorem

To calculate the complex amplitude, $\tilde{E}(\mathbf{r})$, of a wave at an observation point defined by the vector $\mathbf{r}$, we need to use a mathematical relation known as Green's Theorem. Green's theorem states that, if φ and Ψ are two scalar functions that are well behaved, then

$$\int_S\!\!\int (\varphi\nabla\Psi\cdot\hat{\mathbf{n}} - \Psi\nabla\varphi\cdot\hat{\mathbf{n}})\mathrm{d}s = \int\!\!\int_V\!\!\int (\varphi\nabla^2\Psi - \Psi\nabla^2\varphi)\mathrm{d}v \qquad [4]$$

The vector identities

$$\nabla\Psi\cdot\hat{\mathbf{n}} = \frac{\partial\Psi}{\partial n} \qquad \nabla\varphi\cdot\hat{\mathbf{n}} = \frac{\partial\varphi}{\partial n} \qquad [5]$$

allow Green's theorem to be written as

$$\int_S\!\!\int \left[\varphi\frac{\partial\Psi}{\partial n} - \Psi\frac{\partial\varphi}{\partial n}\right]\mathrm{d}s = \int\!\!\int_V\!\!\int [\varphi\nabla^2\Psi - \Psi\nabla^2\varphi]\mathrm{d}v \qquad [6]$$

This equation is the prime foundation of scalar diffraction theory but only the proper choice of Ψ, φ and the surface, S, allows direct application to the diffraction problem.

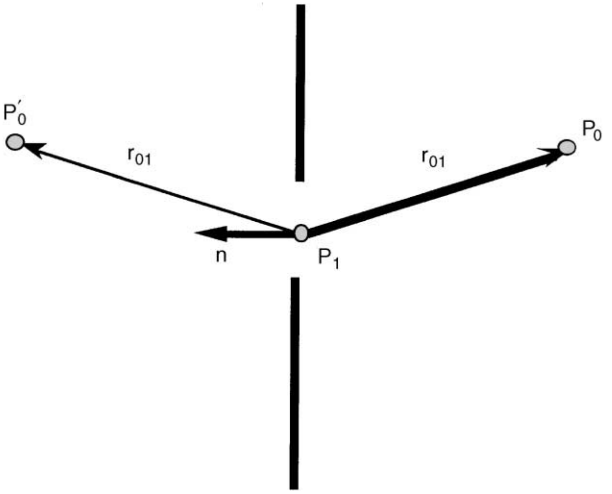

Figure 2 Geometry for the Sommerfeld Green's function. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Because the integral over the combined surfaces, S and S_ε, is equal to zero, the integral over the surface S must equal the negative of the integral over the surface, S_ε:

$$-\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s=\iint_{S}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s \qquad [9]$$

To perform the surface integrals, we must evaluate the Green's function on the surface, S'. We therefore select $\mathbf{r}_1$, the vector defining the observation point, to be on either of the two surfaces that make up S'. The derivative, $\partial\Psi/\partial n$, to be evaluated is

$$\begin{aligned}\frac{\partial\Psi}{\partial n}&=\frac{\partial\Psi}{\partial r}\frac{\partial r}{\partial n}\\&=\left(-ik\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}}-\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}^2}\right)\cos(n,\mathbf{r}_{01}) \qquad [10]\end{aligned}$$

where $\cos(\hat{\mathbf{n}},\mathbf{r}_{01})$ is the cosine of the angle between the outward normal $\hat{\mathbf{n}}$ and the vector $\mathbf{r}_{01}$, the vector between points P_0 and P_1. (Note from Figure 1 that the outward normal on S_ε is inward, toward P_0, while the normal on S is outward, away from P_0.)

For example if $\mathbf{r}_1$ were on S_ε, then

$$\cos(\mathbf{n},\mathbf{r}_{01})=-1, \qquad [11]$$

$$\frac{\partial\Psi(\mathbf{r}_1)}{\partial n}=\left[\frac{1}{\varepsilon}+ik\right]\frac{e^{-ik\varepsilon}}{\varepsilon}, \qquad [12]$$

where ε is the radius of the small sphere we constructed around the point at P_0.

We now use [10] to rewrite the two integrals in [9]. The integral over the surface S_ε is

$$\begin{aligned}&\iint_{S_\varepsilon}\left(\Psi\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial\Psi}{\partial n}\right)\mathrm{d}s\\&=\iint_{S_\varepsilon}\left[\frac{\partial\varphi}{\partial n}\frac{e^{-ik\varepsilon}}{\varepsilon}-\varphi\frac{e^{-ik\varepsilon}}{\varepsilon}\left(\frac{1}{\varepsilon}+ik\right)\right]\\&\quad\times\varepsilon^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \qquad [13]\end{aligned}$$

while the integral over the surface S is

$$\iint_S\left[\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \qquad [14]$$

The omitted volume contained within the surface S_ε is allowed to shrink to zero by taking the limit as $\varepsilon\to0$. Equation [14] will not be affected by taking the limit. The first and last terms of the right side of [13] go to zero as $\varepsilon\to0$ because they contain $\varepsilon e^{-ik\varepsilon}$. The second term in [13] contains $e^{-ik\varepsilon}$ which goes to 1 as $\varepsilon\to0$. Therefore, in the limit as $\varepsilon\to0$, [9] becomes

$$\varphi(\mathbf{r}_0)=\frac{1}{4\pi}\iint_S\left[\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}}\frac{\partial\varphi}{\partial n}-\varphi\frac{\partial}{\partial n}\left(\frac{e^{-ik\cdot\mathbf{r}_{01}}}{r_{01}}\right)\right]\mathrm{d}s \qquad [15]$$

which is sometimes called the integral theorem of Kirchhoff. By carefully selecting the surface of integration in [15], we can calculate the diffraction field observed at P_0 produced by an aperture in an infinite opaque screen.

Assume that a source at P_2 in Figure 3 produces a spherical wave that is incident on an infinite, opaque screen from the left. To find the field at P_0 in Figure 3, we apply [15] to the surface $S_1+S_2+\Sigma$, where S_1 is a plane surface adjacent to the screen, Σ is that portion of S_1 in the aperture and S_2 is a large spherical surface, of radius R, centered on P_0, see Figure 3. The first question to address is how to calculate the contribution from S_2 or better yet how do we show we can ignore contributions from S_2. As R increases, Ψ and φ will decrease as $1/R$. However, the area of integration increases as R^2 so the $1/R$ fall-off is not a sufficient reason for neglecting the contribution of the integration over S_2.

On the surface S_2, the Green's function and its derivative are

$$\Psi=\frac{e^{-ikR}}{R} \qquad [16]$$

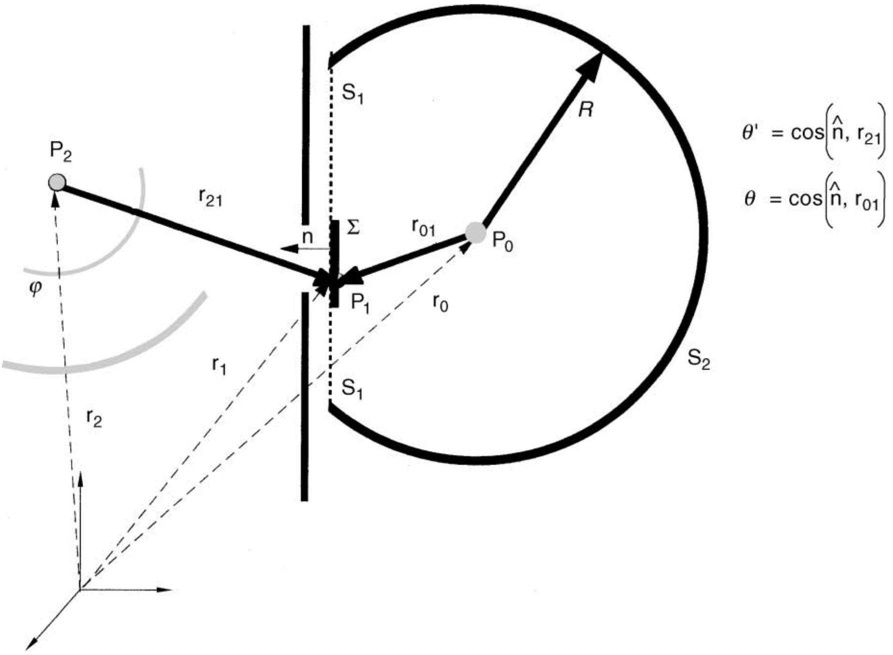

Figure 3 Integration surface for diffraction calculation. P_2 is the source and P_0 is the point where the field is to be calculated. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\frac{\partial \Psi}{\partial n} = \left(-\mathrm{i}k - \frac{1}{R} \right) \frac{\mathrm{e}^{-\mathrm{i}kR}}{R} \approx -\mathrm{i}k\Psi \qquad [17]$$

where the approximate equality is for large R. Thus, the integral over S_2 for large R is

$$\iint_{S_2} \left[\Psi \frac{\partial \varphi}{\partial n} + \varphi(\mathrm{i}k\Psi) \right] \mathrm{d}s = \iint_{S_2} \Psi \left(\frac{\partial \varphi}{\partial n} + \mathrm{i}k\varphi \right) R^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \qquad [18]$$

The quantity $R\Psi = \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}$ is finite as $R \to \infty$ so for the integral to vanish, we must have

$$\lim_{R\to\infty} R\left(\frac{\partial \varphi}{\partial n} + \mathrm{i}k\varphi \right) = 0 \qquad [19]$$

This requirement is called the Sommerfeld radiation condition and is satisfied if $\varphi \to 0$ as fast as $1/R$ (for example, a spherical wave). Since the illumination of the screen will be a spherical wave, or at least a linear combination of spherical waves, we should expect the integral over S_2 to make no contribution (it is exactly zero).

Another way to insure that surface S_2 makes no contribution is to assume that the light turns on at time t_0. At a later time, t, when we desire to know the field at P_0, the radius of S_2 is $R > c\ (t - t_0)$ which physically means that the light has not had time to reach S_2. In this situation, there can be no contribution from S_2. This is not a perfect solution because when the wave is of finite duration it can no longer be considered monochromatic.

We should expect the major contribution in the integral over S_1 to come from the portion of the surface in the aperture, Σ. We make the following assumptions about the incident wave, φ (these assumptions are known as St. Venaut's hypothesis or Kirchhoff's boundary conditions)

1. We assume that, in the aperture, φ and $\partial\varphi/\partial n$ have the values they would have if the screen were not in place.
2. On the portion of S_1 not in the aperture, φ and $\partial\varphi/\partial n$ are identically zero.

The Kirchhoff boundary conditions allow the screen to be neglected, reducing the problem to an integration only over the aperture. It is surprising that the assumptions about the contribution of the surface S_2 and the screen yield very accurate results (as long as polarization is not important, the aperture is large with respect to the wavelength, and we don't get too close to the aperture).

Mathematically the Kirchhoff boundary conditions are incorrect. The two Kirchhoff boundary conditions imply that the field is zero everywhere behind the screen, except in the aperture, which makes Ψ and $\partial\Psi/\partial n$ discontinuous on the boundary of the aperture. Another problem with the boundary conditions is that Ψ and $\partial\Psi/\partial n$ are known only if the problem has already been solved.

The Sommerfeld Green's functions will remove the inconsistencies of the Kirchhoff theory. If we use the Green's function [3a] then Ψ vanishes in the aperture while $\partial\Psi/\partial n$ can assume the value required by Kirchhoff's boundary conditions. If we use the Green's functions [3b] then the converse holds, i.e. $\partial\Psi/\partial n$ is zero in the aperture. This improved Green's function makes the problem harder with little gain in accuracy so we will retain the Kirchhoff formalism.

As a result of the above discussion, the surface integral [15] is only performed over the surface, Σ, in the aperture. The evaluation of the Green's function on this surface can be simplified by noting that normally $r_{01} \gg \lambda$, i.e., $k \gg 1/r_{01}$, thus on Σ,

$$\frac{\partial\Psi}{\partial n} = \cos(\mathbf{n}, \mathbf{r}_{01})\left(-ik - \frac{1}{r_{01}}\right)\frac{e^{-ik\cdot r_{01}}}{r_{01}} \approx -ik\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\frac{e^{-ik\cdot r_{01}}}{r_{01}} \qquad [20]$$

Substituting this approximate evaluation of the derivative into [15] yields

$$\varphi(\mathbf{r}_0) = \frac{1}{4\pi}\iint_{\Sigma}\frac{e^{-ik\cdot r_{01}}}{r_{01}}\left[\frac{\partial\varphi}{\partial n} + ik\varphi\cos(\hat{\mathbf{n}}, \mathbf{r}_{01})\right]ds \qquad [21]$$

The source of the incident wave is a point source located at P_2, with a position vector, $\mathbf{r}_2$, measured with respect to the coordinate system and a distance $|\mathbf{r}_{21}|$ away from P_1, a point in the aperture (see Figure 3). The incident wave is therefore a spherical wave of the form

$$\varphi(\mathbf{r}_{21}) = A\frac{e^{-ik\cdot r_{21}}}{r_{21}} \qquad [22]$$

which fills the aperture. Here also we will assume that $r_{21} \gg \lambda$ so that the derivative of the incident wave assumes the same form as [20]. Then [21] can be written

$$\varphi(\mathbf{r}_0) = \frac{iA}{\lambda}\iint_{\Sigma}\frac{e^{-ik\cdot(r_{21}+r_{01})}}{r_{21}r_{01}} \times\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right]ds \qquad [23]$$

This relationship is called the Fresnel–Kirchhoff diffraction formula. It is symmetric with respect to r_{01} and r_{21}, making the problem identical when the source and measurement point are interchanged.

A physical understanding of [23] can be obtained if it is rewritten as

$$\varphi(\mathbf{r}_0) = \iint_{\Sigma}\Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}}ds \qquad [24]$$

The field at P_0 is due to the sum of an infinite number of secondary Huygens sources in the aperture Σ. The secondary sources are point sources radiating spherical waves of the form

$$\Phi(\mathbf{r}_{21})\frac{e^{-ik\cdot r_{01}}}{r_{01}} \qquad [25]$$

with amplitude $\Phi(r_{21})$, defined by

$$\Phi(\mathbf{r}_{21}) = \frac{i}{\lambda}\left[A\frac{e^{-ik\cdot r_{21}}}{r_{21}}\right]\left[\frac{\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})}{2}\right] \qquad [26]$$

The imaginary constant, i, causes the wavelets from each of these secondary sources to be phase shifted with respect to the incident wave. The obliquity factor, in the amplitude [26]

$$\tfrac{1}{2}[\cos(\hat{\mathbf{n}}, \mathbf{r}_{01}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_{21})] \qquad [27]$$

causes the secondary sources to have a forward directed radiation pattern.

If we had used the Sommerfeld Green's function, the only modification to our result would be a change in the obliquity factor to $\cos(\mathbf{n},\mathbf{r}_{01})$. In our discussions below we will assume that the angles are all small, allowing the obliquity factor to be replaced by 1 so that in the remaining discussion the choice of Green's function has no impact.

We will assume the source of light is at infinity, $z_1 = \infty$ in Figure 3, so that the aperture is illuminated by a plane wave, traveling parallel to the z-axis. With this assumption

$$\theta' = \cos(n, \mathbf{r}_{21}) \approx -1 \qquad [28]$$

We will also make the paraxial approximation which assumes that the viewing position is close to the z-axis, leading to

$$\theta = \cos(\mathbf{n}, \mathbf{r}_{01}) \approx 1 \qquad [29]$$

With these assumptions eqn [24] becomes identical to the Huygens–Fresnel integral discussed in the section on Fresnel diffraction:

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda}\iint_{\Sigma}\frac{\tilde{E}_i(\mathbf{r})}{R}e^{-ik\cdot R}\,ds \qquad [30]$$

The modern interpretation of the Fresnel–Kirchhoff (or now in the small-angle approximation, the Fresnel–Huygens) integral is to view it as a convolution integral. By considering free space to be a linear system, the result of propagation can be calculated by convolving the incident (input) wave with the impulse response of free space (our Green's function),

$$\frac{i}{\lambda}\frac{e^{-ik\cdot R}}{R} \qquad [31]$$

The job of calculating diffraction has only begun with the derivation of the Fresnel–Kirchhoff integral. In general, an analytic expression for the integral cannot be found because of the difficulty of performing the integration over R. There are two approximations that will allow us to obtain analytic expressions of the Fresnel–Kirchhoff integral. In both approximations, all dimensions are assumed to be large with respect to the wavelength. In one approximation, the viewing position is assumed to be far from the obstruction; the resulting diffraction is called Fraunhofer diffraction and will be discussed here. The second approximation, which leads to Fresnel diffraction, assumes that the observation point is nearer the obstruction, to be quantified below. This approximation, which is the more difficult mathematically, is discussed in the article on **Diffraction:** Fresnel Diffraction.

Fraunhofer Approximation

In Fraunhofer diffraction, we require that the source of light and the observation point, P_0, be far from the aperture so that the incident and diffracted wave can be approximated by plane waves. A consequence of this requirement is that the entire waveform passing through the aperture contributes to the observed diffraction.

The geometry to be used in this derivation is shown in Figure 4. The wave incident normal to the aperture, Σ, is a plane wave and the objective of the calculation is to find the departure of the transmitted wave from its geometrical optical path. The calculation will provide the light distribution, transmitted by the aperture, as a function of the angle the light is deflected from the incident direction. We assume that diffraction makes only a small perturbation on the predictions of geometrical optics. The deflection angles encountered in this derivation are, therefore, small, as assumed above, and we will be able to use the paraxial approximation.

The distance from a point P, in the aperture, to the observation point P_0, of Figure 4, is

$$R^2 = (x-\xi)^2 + (y-\eta)^2 + Z^2 \qquad [32]$$

From Figure 4 we see R_0 is the distance from the center of the screen to the observation point, P_0,

$$R_0^2 = \xi^2 + \eta^2 + Z^2 \qquad [33]$$

The difference between these two vectors is

$$\begin{aligned} R_0^2 - R^2 &= (R_0 - R)(R_0 + R) \\ &= \xi^2 + \eta^2 + Z^2 - Z^2 - (x^2 - 2x\xi + \xi^2) \\ &\quad - (y^2 - 2y\eta + \eta^2) \\ &= 2(x\xi + y\eta) - (x^2 + y^2) \end{aligned} \qquad [34]$$

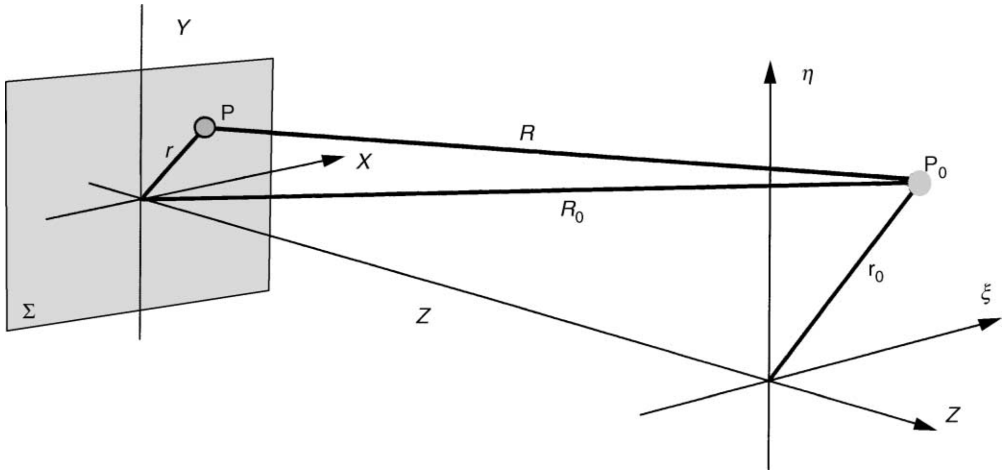

Figure 4 Geometry for Fraunhofer diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The equation for the position of point P in the aperture can be written in terms of [34]:

$$r = R_0 - R = \frac{R_0^2 - R^2}{R_0 + R},$$
$$= [2(x\xi + y\eta) - (x^2 + y^2)]\frac{1}{R_0 + R} \qquad [35]$$

The reciprocal of $(R_0 + R)$ can be written as

$$\frac{1}{R_0 + R} = \frac{1}{2R_0 + R - R_0} = \frac{1}{2R_0}\left(1 + \frac{R - R_0}{2R_0}\right)^{-1} \qquad [36]$$

Now using [36]

$$R_0 - R = \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{R_0 - R}{2R_0}\right]^{-1}$$
$$= \left[\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right]\left[1 - \frac{r}{2R_0}\right]^{-1} \qquad [37]$$

If the diffraction integral [30] is to have a finite (nonzero) value, then

$$k|R_0 - R| \ll kR_0 \qquad [38]$$

This requirement insures that all the Huygens's wavelets, produced over the aperture from the center out to the position r, will have similar phases and will interfere to produce a nonzero amplitude at P_0. This requirement results in

$$\frac{1}{1 - \frac{r}{2R_0}} \approx 1 \qquad [39]$$

From this equation we are tempted to state that the aperture dimension must be small relative to the observation distance. However, for the exponent to provide a finite contribution in the integral, it is kr that is small, not the aperture dimension, r.

Using the approximation for $(R_0 - R)$, we obtain the diffraction integral

$$\tilde{E}_P = \frac{\mathrm{i}\alpha\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\iint_{\Sigma} f(x,y)\mathrm{e}^{+\mathrm{i}k\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)}\,\mathrm{d}x\,\mathrm{d}y \qquad [40]$$

The change in the amplitude of the wave due to the change in R, as we move across the aperture, is neglected, allowing R in the denominator of the Huygens–Fresnel integral to be replaced by R_0 and moved outside of the integral. We have introduced the complex transmission function, $f(x, y)$, of the aperture to allow a very general aperture to be treated. If the aperture function described the variation in absorption of the aperture as a function of position, as would be produced by a photographic negative, then $f(x, y)$ would be a real function. If the aperture function described the variation in transmission of a biological sample, it might be entirely imaginary.

The argument of the exponent in [40] is

$$\mathrm{i}k\left(\frac{x\xi + y\eta}{R_0} - \frac{x^2 + y^2}{2R_0}\right)$$
$$= \mathrm{i}2\pi\left(\frac{x\xi + y\eta}{\lambda R_0} - \frac{x^2 + y^2}{2\lambda R_0}\right) \qquad [41]$$

If the observation point, P_0, is far from the screen, we can neglect the second term and treat the phase variation across the aperture as a linear function of position. This is equivalent to assuming that the diffracted wave is a collection of plane waves. Mathematically, the second term in [41] can be neglected if

$$\frac{x^2 + y^2}{2\lambda R_0} \ll 1 \qquad [42]$$

This is called the far-field approximation and the theory yields Fraunhofer diffraction. If the quadratic term of [41] is on the order of 2π then the fraction is

$$\frac{x^2 + y^2}{2\lambda R_0} \approx \Theta(1) \qquad [43]$$

and we must retain the quadratic term and the theory yields Fresnel diffraction.

We define spatial frequencies in the x and y direction by

$$\omega_x = \frac{2\pi}{\lambda}\sin\theta_x = -\frac{2\pi\xi}{\lambda R_0}$$
$$\omega_y = \frac{2\pi}{\lambda}\sin\theta_y = -\frac{2\pi\eta}{\lambda R_0} \qquad [44]$$

We define the spatial frequencies with a negative sign to allow equations involving spatial frequencies to have the same form as those involving temporal frequencies. The negative sign is required because ωt and $\mathbf{k}\cdot\mathbf{r}$ appear in the phase of the wave with opposite signs and we want the Fourier transform in space and in time to have the same form.

With the variables defined in [44], the integral becomes

$$\tilde{E}_P(\omega_x, \omega_y) = \frac{\mathrm{i}\alpha\, \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\iint_{\Sigma} f(x,y)\mathrm{e}^{-\mathrm{i}(\omega_x x + \omega_y y)}\,\mathrm{d}x\,\mathrm{d}y \qquad [45]$$

The result of our derivation is that the Fraunhofer diffraction field, $\tilde{E}_P$, can be calculated by performing the two-dimensional Fourier transform of the aperture's transmission function.

Definition of Fourier Transform

In this discussion the function $F(\omega)$ is defined as the *Fourier transform* of $f(t)$:

$$\mathcal{F}\{f(t)\} \equiv F(\omega) \equiv \int_{-\infty}^{\infty} f(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \qquad [46]$$

The transformation from a temporal to a frequency representation given by [46] does not destroy information; thus, the inverse transform can also be defined

$$\mathcal{F}^{-1}\{F(\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega \qquad [47]$$

By assuming that the illumination of an aperture is a plane wave and that the observation position is in the far field, the diffraction of an aperture is found to be given by the Fourier transform of the function describing the amplitude transmission of the aperture. The amplitude transmission of the aperture, $f(x, y)$, may thus be interpreted as the superposition of mutually coherent plane waves.

Diffraction by a Rectangular Aperture

We will now use Fourier transform theory to calculate the Fraunhofer diffraction pattern from a rectangular slit and point out the reciprocal relationship between the size of the diffraction pattern and the size of the aperture.

Consider a rectangular aperture with a transmission function given by

$$f(x, y) = \begin{cases} 1 & |x| \le x_0, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [48]$$

Because the aperture is two-dimensional, we need to apply a two-dimensional Fourier transform but it is not difficult because the amplitude transmission function is separable, in x and y. The diffraction amplitude distribution from the rectangular slit is simply one-dimensional transforms carried out individually for the x and y dimensions:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\int_{-x_0}^{x_0} f(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \int_{-y_0}^{y_0} f(y)\mathrm{e}^{-\mathrm{i}\omega_y y}\,\mathrm{d}y \qquad [49]$$

Since both $f(x)$ and $f(y)$ are defined as symmetric functions, we need only calculate the cosine transforms to obtain the diffracted field:

To calculate the Fourier transform defined by [46] we can rewrite the transform as

$$F(\omega) = \int_{-\infty}^{\infty} f(\tau)\cos\omega\tau\,\mathrm{d}\tau - \mathrm{i}\int_{-\infty}^{\infty} f(\tau)\sin\omega\tau\,\mathrm{d}\tau \qquad [50]$$

If $f(\tau)$ is a real, even function then the Fourier transform can be obtained by simply calculating the cosine transform:

$$\int_{-\infty}^{\infty} f(\tau)\cos\omega\tau\,\mathrm{d}\tau \qquad [51]$$

$$\tilde{E}_P = \mathrm{i}\frac{4x_0 y_0 \alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\omega_x x_0}{\omega_x x_0}\frac{\sin\omega_y y_0}{\omega_y y_0} \qquad [52]$$

The intensity distribution of the Fraunhofer diffraction produced by the rectangular aperture is

$$I_P = I_0\frac{\sin^2\omega_x x_0}{(\omega_x x_0)^2}\frac{\sin^2\omega_y y_0}{(\omega_y y_0)^2} \qquad [53]$$

The maximum intensities in the x and y directions occur at

$$\omega_x x_0 = \omega_y y_0 = 0 \qquad [54]$$

The area of this rectangular aperture is defined as $A = 4x_0 y_0$, resulting in an expression for the maximum intensity of

$$I_0 = \left[\frac{\mathrm{i}4x_0 y_0 \alpha\,\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}}{\lambda R_0}\right]^2 = \frac{A^2\alpha^2}{\lambda^2 R_0^2} \qquad [55]$$

The minima of [53] occur when $\omega_x x_0 = n\pi$ or when $\omega_y y_0 = m\pi$. The location of the zeros can be specified as a dimension in the observation plane or, using the paraxial approximation, in terms of the angle defined by [44]

$$\begin{aligned} \sin\theta_x &\approx \theta_x \approx \frac{\xi}{R_0} = \frac{n\lambda}{2x_0} \\ \sin\theta_y &\approx \theta_y \approx \frac{\eta}{R_0} = \frac{m\lambda}{2y_0} \end{aligned} \qquad [56]$$

The dimensions of the diffraction pattern are characterized by the location of the first zero, i.e., when $n = m = 1$, and are given by the observation plane coordinates, ξ and η. The dimensions of the diffraction pattern are inversely proportional to the dimensions of the aperture. As the aperture dimension expands, the width of the diffraction pattern decreases until, in the limit of an infinitely wide aperture, the diffraction pattern becomes a delta function.

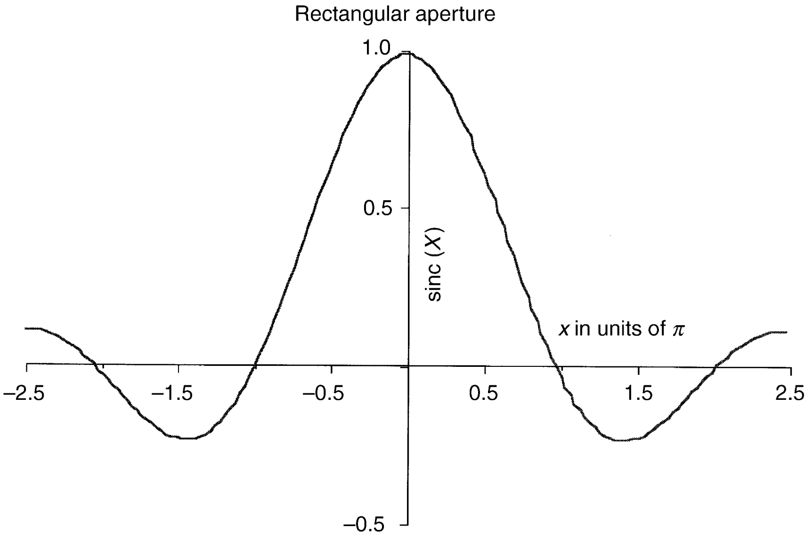

Figure 5 The amplitude of the wave diffracted by a rectangular slit. This sinc function describes the light wave's amplitude that would exist in both the *x* and *y* directions. The coordinate would have to be scaled by the dimension of the slit. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Figure 5 is a theoretical plot of the amplitude of the diffraction wave from a rectangular slit.

Diffraction from a Circular Aperture

To obtain the diffraction pattern from a circular aperture we first convert from the rectangular coordinate system we have used to a cylindrical coordinate system. The new cylindrical geometry is shown in Figure 6. At the aperture plane

$$\begin{aligned} x &= s\cdot\cos\varphi & y &= s\cdot\sin\varphi \\ f(x,y) &= f(s,\varphi) & \mathrm{d}x\,\mathrm{d}y &= s\,\mathrm{d}s\,\mathrm{d}\varphi \end{aligned} \qquad [57]$$

At the observation plane

$$\xi = \rho\cdot\cos\theta \qquad \eta = \rho\cdot\sin\theta \qquad [58]$$

In the new, cylindrical, coordinate system at the observation plane, the spatial frequencies are written as

$$\begin{aligned} \omega_x &= -\frac{k\xi}{R_0} = -\frac{k\rho}{R_0}\cos\theta \\ \omega_y &= -\frac{k\eta}{R_0} = -\frac{k\rho}{R_0}\sin\theta \end{aligned} \qquad [59]$$

From Figure 6 we see that the observation point, P, can be defined in terms of the angle ψ, where

$$\sin\psi = \frac{\rho}{R_0} \qquad [60]$$

This allows an angular representation of the size of the diffraction pattern if it is desired.

Using [57] and [59], we may write

$$\begin{aligned} \omega_x x + \omega_y y &= -\frac{ks\rho}{R_0}(\cos\theta\cos\varphi + \sin\theta\sin\varphi), \\ &= -\frac{ks\rho}{R_0}\cos(\theta-\varphi) \end{aligned} \qquad [61]$$

The Huygens–Fresnel integral for Fraunhofer diffraction can now be written in terms of cylindrical coordinates:

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}_0}\int_0^{\frac{a}{2}}\int_0^{2\pi} f(s,\varphi)\mathrm{e}^{-\mathrm{ik}\frac{s\rho}{R_0}\cos(\theta-\varphi)}\,s\,\mathrm{d}s\,\mathrm{d}\varphi \qquad [62]$$

We will use this equation to calculate the diffraction amplitude from a clear aperture of diameter a, defined by the equation

$$f(s,\varphi) = \begin{cases} 1 & s \le \frac{a}{2},\ \text{all } \varphi \\ 0 & s > \frac{a}{2} \end{cases} \qquad [63]$$

The symmetry of this problem is such that

$$f(s,\varphi) = f(s)g(\varphi) = f(s) \qquad [64]$$

$$\mathcal{F}\{f(s,\phi)\} = \mathcal{F}\{f(s)\} = F(\rho) \qquad [65]$$

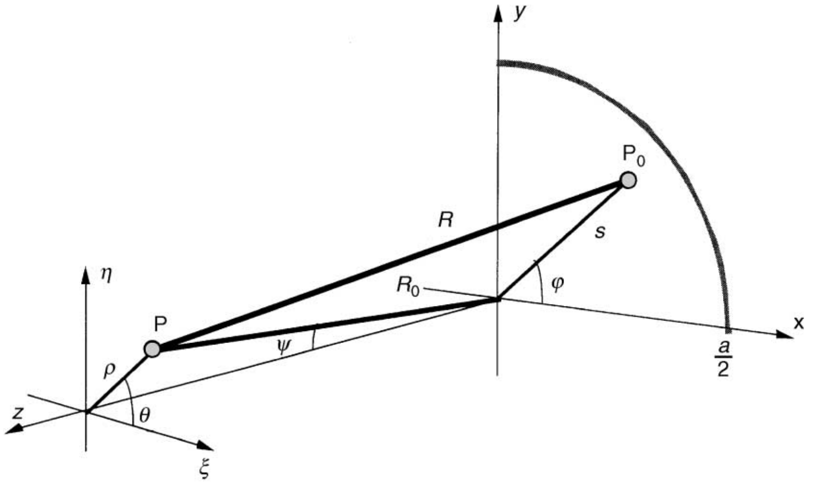

Figure 6 Geometry for calculation of Fraunhofer diffraction from a circular aperture. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$F(\rho, \theta) = \int_0^\infty f(s)s\, \mathrm{d}s \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}\rho \cdot s \cos(\varphi - \theta)}\, \mathrm{d}\varphi \qquad [66]$$

$$F(\rho) = \int_0^\infty f(s)\, s\, \mathrm{d}s \int_0^{2\pi} \mathrm{e}^{-\mathrm{i}\rho \cdot s \cos \varphi}\, \mathrm{d}\varphi \qquad [67]$$

The second integral in [67] belongs to a class of functions called the Bessel function defined by the integral

$$\mathrm{J}_n(s\rho) = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{e}^{\mathrm{i}[s\rho \sin \varphi - n\varphi]} \mathrm{d}\varphi \qquad [68]$$

In [67] the integral corresponds to the $n = 0$, zero-order, Bessel function. Using this definition, we can rewrite [67] as

$$F(\rho) = \int_0^\infty f(s) \mathrm{J}_0(s\rho)s\, \mathrm{d}s \qquad [69]$$

This transform is called the Fourier–Bessel transform or the Hankel zero-order transform.

Using these definitions, the transform of the aperture function, $f(s)$, is

$$F(\rho) = \int_0^{\frac{a}{2}} \mathrm{J}_0(s\rho)s\, \mathrm{d}s \qquad [70]$$

We use the identity

$$x\mathrm{J}_1(x) = \int_0^x \chi \mathrm{J}_0(\chi)\, \mathrm{d}\chi \qquad [71]$$

to obtain

$$F(\rho) = \frac{a}{2\rho} \mathrm{J}_1\left(\frac{\rho a}{2}\right) \qquad [72]$$

We can now write the amplitude of the diffracted wave as

$$\tilde{E}_P = \frac{\mathrm{i}\alpha}{\lambda} \mathrm{e}^{-\mathrm{i}\mathbf{k} \cdot \mathbf{R}_0} \left[\frac{\pi a}{k\rho} \mathrm{J}_1\left(\frac{ka\rho}{2R_0}\right) \right] \qquad [73]$$

A plot of the function contained within the bracket of [73] is given in Figure 7. If we define

$$u = \frac{ka\rho}{2R_0} \qquad [74]$$

then the spatial distribution of intensity in the diffraction pattern can be written in a form known as the Airy formula,

$$I = I_0 \left[\frac{2\mathrm{J}_1(u)}{u} \right]^2 \qquad [75]$$

where we have defined

$$I_0 = \left(\frac{\alpha A}{\lambda R_0} \right)^2 \qquad [76]$$

with A representing the area of the aperture, $A = \pi(\frac{a}{2})^2$.

The intensity pattern described by [75] is called the Airy pattern. The intensity at $u = 0$ is the same as was obtained for a rectangular aperture of the same area, because, in the limit,

$$\lim_{u \to 0} \left[\frac{2\mathrm{J}_1(u)}{u} \right] = 1 \qquad [77]$$

For the Airy pattern, 84% of the total area is contained in the disk between the first zeros of [75]. Those zeros occur at $u = \pm 1.22\pi$, which corresponds

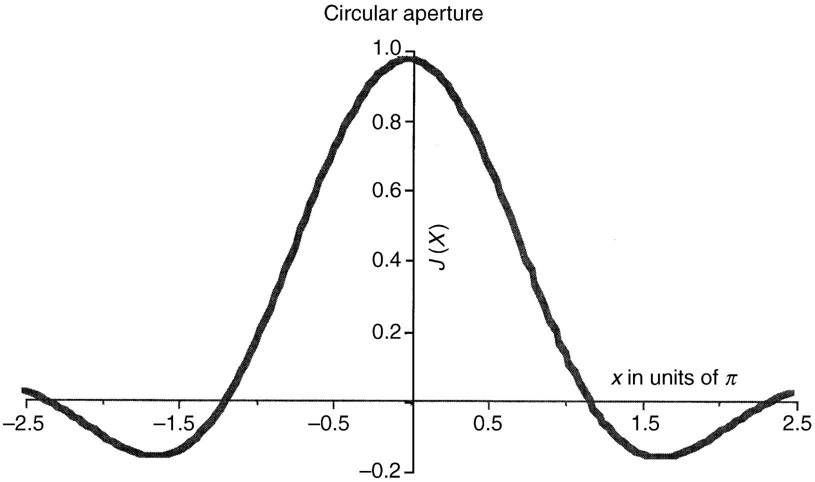

Figure 7 Amplitude of a wave diffracted by a circular aperture. The observed light distribution is constructed by rotating this Bessel function around the optical axis. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

to a radius in the observation plane of

$$\rho = \frac{1.22(\lambda R_0)}{a} \qquad [78]$$

91% of the light intensity is contained within the circle bounded by the second minimum at $u = 2.233\pi$. The intensities in the secondary maxima of the diffraction pattern of a rectangular aperture [53] are much larger than the intensities in the secondary maxima of the Airy pattern of a circular aperture of equal area. The peak intensities, relative to the central maximum, of the first three secondary maxima of a rectangular aperture are 4.7%, 1.6%, and 0.8%, respectively. For a circular aperture, the same quantities are 1.7%, 0.04%, and 0.02%, respectively.

Array Theorem

There is an elegant mathematical technique for handling diffraction from multiple apertures, called the array theorem. The theorem is based on the fact that Fraunhofer diffraction is given by the Fourier transform of the aperture function and utilizes the convolution integral.

The convolution of the functions $a(t)$ and $b(t)$ is defined as

$$g(t) = a(t) \otimes b(t) = \int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t \qquad [79]$$

The Fourier transform of a convolution of two functions is the product of the Fourier transforms of the individual functions:

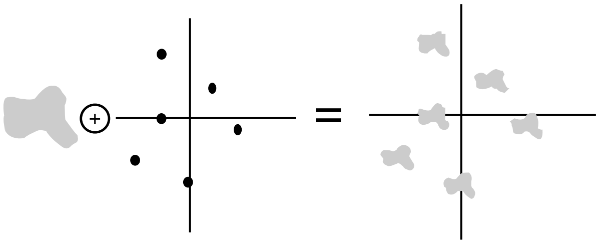

Figure 8 The convolution of an aperture with an array of delta functions will produce an array of identical apertures, each located at the position of one of the delta functions. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

$$\mathcal{F}\{a(t) \otimes b(t)\} = \mathcal{F}\left\{\int_{-\infty}^{\infty} a(t)b(\tau - t)\,\mathrm{d}t\right\} = A(\omega)B(\omega) \qquad [80]$$

We will demonstrate the array theorem for one dimension, where the functions represent slit apertures. The results can be extended to two dimensions in a straightforward way.

Assume that we have a collection of identical apertures, shown on the right of Figure 8. If one of the apertures is located at the origin of the aperture plane, its transmission function is $\psi(x)$. The transmission function of an aperture located at a point, x_n, can be written in terms of a generalized aperture function, $\psi(x - \alpha)$, by the use of the sifting property of the delta function

$$\psi(x - x_n) = \int \psi(x - \alpha)\delta(\alpha - x_n)\,\mathrm{d}\alpha \qquad [81]$$

The aperture transmission function representing an array of apertures will be the sum of the distributions of the individual apertures, represented graphically in Figure 8 and mathematically by the summation

$$\Psi(x) = \sum_{n=1}^{N} \psi(x - x_n) \qquad [82]$$

The Fraunhofer diffraction from this array is $\Phi(\omega_x)$, the Fourier transform of $\Psi(x)$,

$$\Phi(\omega_x) = \int_{-\infty}^{\infty} \Psi(x)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \qquad [83]$$

which can be rewritten as

$$\Phi(\omega_x) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \psi(x - x_n)\mathrm{e}^{-\mathrm{i}\omega_x x}\,\mathrm{d}x \qquad [84]$$

We now make use of the fact that $\psi(x - x_n)$ can be expressed in terms of a convolution integral. The Fourier transform of $\psi(x - x_n)$ is, from the convolution theorem [80], the product of the Fourier transforms of the individual functions that make up the convolution:

$$\begin{aligned}\Phi(\omega_x) &= \sum_{n=1}^{N} \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\{\delta(\alpha - x_n)\}\\ &= \mathcal{F}\{\psi(x-\alpha)\}\sum_{n=1}^{N}\mathcal{F}\{\delta(\alpha - x_n)\}\\ &= \mathcal{F}\{\psi(x-\alpha)\}\mathcal{F}\left\{\sum_{n=1}^{N}\delta(\alpha - x_n)\right\} \qquad [85]\end{aligned}$$

The first transform in [85] is the diffraction pattern of the generalized aperture function and the second transform is the diffraction pattern produced by a set of point sources with the same spatial distribution as the array of identical apertures. We will call this second transform the array function.

To summarize, the array theorem states that the diffraction pattern of an array of similar apertures is given by the product of the diffraction pattern from a single aperture and the diffraction (or interference) pattern of an identically distributed array of point sources.

N Rectangular Slits

An array of N identical apertures is called a diffraction grating in optics. The Fraunhofer diffraction patterns produced by such an array have two important properties; a number of very narrow beams are produced by the array and the beam positions are a function of the wavelength of illumination of the apertures and the relative phase of the waves radiated by each of the apertures.

Because of these properties, arrays of diffracting apertures have been used in a number of applications.

- At radio frequencies, arrays of dipole antennas are used to both radiate and receive signals in both radar and radio astronomy systems. One advantage offered by diffracting arrays at radio frequencies is that the beam produced by the array can be electrically steered by adjusting the relative phase of the individual dipoles.
- An optical realization of a two-element array of radiators is Young's two-slit experiment and a realization of a two-element array of receivers is Michelson's stellar interferometer. Two optical implementations of arrays, containing more diffracting elements, are diffraction gratings and holograms. In nature, periodic arrays of diffracting elements are the origin of the colors observed on some invertebrates.
- Many solids are naturally arranged in three-dimensional arrays of atoms or molecules that act as diffraction gratings when illuminated by X-ray wavelengths. The resulting Fraunhofer diffraction patterns are used to analyze the ordered structure of the solids.

The array theorem can be used to calculate the diffraction pattern from N rectangular slits, each of width a and separation d. The aperture function of a single slit is equal to

$$\psi(x, y) = \begin{cases} 1 & |x| \le \frac{a}{2}, |y| \le y_0 \\ 0 & \text{all other } x \text{ and } y \end{cases} \qquad [86]$$

The Fraunhofer diffraction pattern of this aperture has already been calculated and is given by [52]:

$$\mathcal{F}\{\psi(x, y)\} = K\frac{\sin\alpha}{\alpha} \qquad [87]$$

where the constant K and the variable α are

$$\begin{aligned}K &= \frac{\mathrm{i}2ay_0\alpha}{\lambda R_0}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}_0}\frac{\sin\omega_y y_0}{\omega_y y_0}\\ \alpha &= \frac{ka}{2}\sin\theta_x \qquad [88]\end{aligned}$$

The array function is

$$A(x) = \sum_{n=1}^{N}\delta(x - x_n) \quad \text{where } x_n = (n-1)d \qquad [89]$$

and its Fourier transform is given by

$$\mathcal{F}\{A(x)\} = \frac{\sin N\beta}{\sin \beta} \quad [90]$$

$$\beta = \frac{kd}{2}\sin\theta_x$$

The Fraunhofer diffraction pattern's intensity distribution in the x-direction is thus given by

$$I_\theta = I_0 \underbrace{\frac{\sin^2\alpha}{\alpha^2}}_{\text{shape factor}} \underbrace{\frac{\sin^2 N\beta}{\sin^2\beta}}_{\text{grating factor}} \quad [91]$$

We have combined the variation in intensity in the y-direction into the constant I_0 because we assume that the intensity variation in the x-direction will be measured at a constant value of y.

A physical interpretation of [91] views the first factor as arising from the diffraction associated with a single slit; it is called the shape factor. The second factor arises from the interference between light from different slits; it is called the grating factor. The fine detail in the spatial light distribution of the diffraction pattern is described by the grating factor and arises from the coarse detail in the diffracting object. The coarse, overall light distribution in the diffraction pattern is described by the shape factor and arises from the fine detail in the diffracting object.

Young's Double Slit

The array theorem makes the analysis of Young's two-slit experiment a trivial exercise. This application of the array theorem will demonstrate that the interference between two slits arises naturally from an application of diffraction theory. The result of this analysis will support a previous assertion that interference describes the same physical process as diffraction and the division of the two subjects is an arbitrary one.

The intensity of the diffraction pattern from two slits is obtained from [91] by setting $N = 2$:

$$I_\theta = I_0 \frac{\sin^2\alpha}{\alpha^2}\cos^2\beta \quad [92]$$

The sinc function describes the energy distribution of the overall diffraction pattern, while the cosine function describes the energy distribution created by interference between the light waves from the two slits. Physically, α is a measure of the phase difference between points in one slit and β is a measure of the phase difference between similar points in the two slits. Zeros in the diffraction intensity occur whenever $\alpha = n\pi$ or whenever $\beta = \frac{1}{2}(2n+1)\pi$. Figure 9 shows the interference maxima, from the grating factor, under the central diffraction peak, described by the shape factor. The number of interference maxima contained under the central maximum is given by

$$\frac{2d}{a} - 1 \quad [93]$$

In Figure 9 three different slit spacings are shown with the ratio d/a equal to 3, 6, and 9, respectively.

The Diffraction Grating

In this section we will use the array theorem to derive the diffraction intensity distribution of a large number of identical apertures. We will discover that the positions of the principal maxima are a function of the illuminating wavelength. This functional relationship has led to the application of a diffraction grating to wavelength measurements.

The diffraction grating normally used for wavelength measurements is not a large number of diffracting apertures but rather a large number of reflecting grooves cut in a surface such as gold or aluminum. The theory to be derived also applies to these reflection gratings but a modification must be made to the theory. The shape of the grooves in the reflection grating can be used to control the fraction of light diffracted into a principal maximum and we will examine how to take this into account. A grating whose groove shape has been controlled to enhance the energy contained in a particular principal maximum is called a blazed grating. The use of special groove shapes is equivalent to the modification of the phase of individual elements of an antenna array at radio frequencies.

The construction of a diffraction grating for use in an optical device for measuring wavelength was first suggested by David Rittenhouse, an American astronomer, in 1785, but the idea was ignored until Fraunhofer reinvented the concept in 1819. Fraunhofer's first gratings were fine wires spaced by wrapping the wires in the threads of two parallel screws. He later made gratings by cutting (*ruling*) grooves in gold films deposited on the surface of glass. H.A. Rowland made a number of well designed ruling machines, which made possible the production of large-area gratings. Following a suggestion by Lord Rayleigh, Robert Williams Wood (1868–1955) developed the capability to control the shape of the individual grooves.

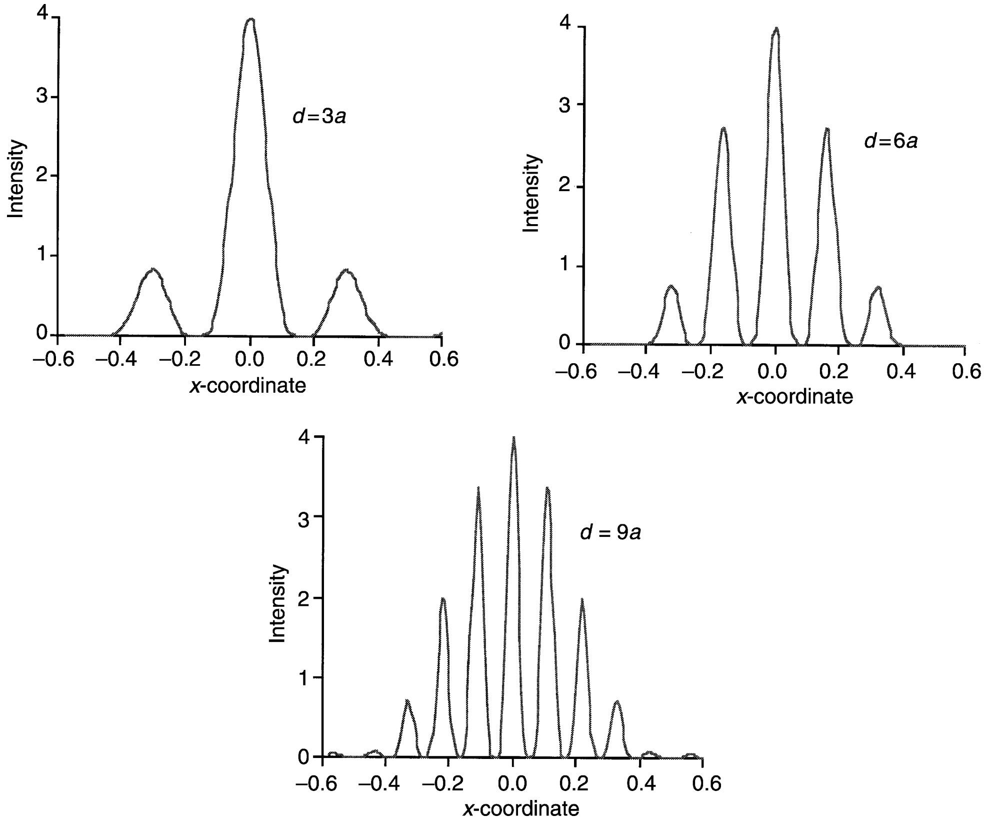

Figure 9 The number of interference fringes beneath the main diffraction peak of a Young's two-slit experiment with rectangular apertures. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If N is allowed to assume values much larger than 2, the appearance of the interference fringes, predicted by the grating factor, changes from a simple sinusoidal variation to a set of narrow maxima, called principal maxima, surrounded by much smaller, secondary maxima. To calculate the diffraction pattern we must evaluate [91] when N is a large number.

Whenever $N\beta = m\pi$, $m = 0, 1, 2, \ldots$, the numerator of the second factor in [91] will be zero, leading to an intensity that is zero, $I_\theta = 0$. The denominator of the second factor in [91] is zero when $\beta = l\pi$, $l = 0, 1, 2, \ldots$ If both conditions occur at the same time, the ratio of m and N is equal to an integer, $m/N = l$, and instead of zero intensity, we have an indeterminate value for the intensity, $I_\theta = 0/0$. To find the actual value for the intensity, we must apply L'Hospital's rule

$$\lim_{\beta \to l\pi} \frac{\sin N\beta}{\sin\beta} = \lim_{\beta \to l\pi} \frac{N\cos N\beta}{\cos\beta} = N \qquad [94]$$

L'Hospital's rule predicts that whenever

$$\beta = l\pi = \frac{m}{N}\pi \qquad [95]$$

where l is an integer, a principal maximum in the intensity will occur with a value given by

$$I_{\theta P} = N^2 I_0 \frac{\sin^2\alpha}{\alpha^2} \qquad [96]$$

Secondary maxima, much weaker than the principal maxima, occur when

$$\beta = \left(\frac{2m+1}{2N}\right)\pi \qquad m = 1, 2, \ldots \qquad [97]$$

(When $m = 0$, m/N is an integer, thus the first value that m can have in [97] is $m = 1$.) The intensity of each secondary maximum is given by

$$\begin{aligned} I_{\theta S} &= I_0 \frac{\sin^2\alpha}{\alpha^2} \frac{\sin^2 N\beta}{\sin^2\beta} \\ &= I_0 \frac{\sin^2\alpha}{\alpha^2} \left[\frac{\sin^2\left(\frac{2m+1}{2}\right)\pi}{\sin^2\left(\frac{2m+1}{2N}\right)\pi} \right] \\ &= I_0 \frac{\sin^2\alpha}{\alpha^2} \left[\frac{1}{\sin\left(\frac{2m+1}{2N}\right)\pi} \right]^2 \qquad [98] \end{aligned}$$

The quantity $(2m+1)/2N$ is a small number for large values of N, allowing the small angle approximation to be made:

$$I_{\theta S} \approx I_0 \frac{\sin^2\alpha}{\alpha^2}\left[\frac{2N}{\pi(2m+1)}\right]^2 \qquad [99]$$

The ratio of the intensity of a secondary maximum and a principal maximum is given by

$$\frac{I_{\theta S}}{I_{\theta P}} = \left[\frac{2}{\pi(2m+1)}\right]^2 \qquad [100]$$

The strongest secondary maximum occurs for $m = 1$ and, for large N, has an intensity that is about 4.5% of the intensity of the neighboring principal maximum.

The positions of principal maxima occur at angles specified by the grating formula.

$$\beta = \frac{kd\sin\theta}{2} = l\pi = \frac{m}{N}\pi \qquad [101]$$

The angular position of the principal maxima (the diffracted angle θ_d) is given by the Bragg equation

$$\sin\theta_d = \frac{l\lambda}{d} \qquad [102]$$

where l is called the grating order. This simple relationship between the angle and the wavelength can be used to construct a device to measure wavelength, the grating spectrometer.

The model we have used to obtain this result is based on a periodic array of identical apertures. The transmission function of this array is a periodic square wave. If, for the moment, we treat the grating as infinite in size, we discover that the principal maxima in the diffraction pattern correspond to the terms of the Fourier series describing a square wave.

The zero order, $m = 0$, corresponds to the temporal average of the periodic square wave and has an intensity proportional to the spatially averaged transmission of the grating. Because of its equivalence to the temporal average of a time-varying signal, the zero-order principal maximum is often called the dc term.

The first-order principal maximum corresponds to the fundamental spatial frequency of the grating and the higher orders correspond to the harmonics of this frequency.

The dc term provides no information about the wavelength of the illumination. Information about the wavelength of the illuminating light can only be obtained by measuring the angular position of the first or higher order principal maxima.

Grating Spectrometer

The curves shown in Figure 10 display the separation of the principal maxima as a function of $\sin\theta_d$. The angular separation of principal maxima can be converted to a linear dimension by assuming a distance, R, from the graing to the observation plane. In the lower right-hand curve of Figure 10, a distance of 2 meters was assumed. Grating spectrometers are classified by the size in meters of R used in their design. The larger R, the easier it is to resolve wavelength differences. For example, a 1 meter spectrometer is a higher-resolution instrument than a 1/4 meter spectrometer.

The fact that the grating is finite in size causes each of the orders to have an angular width that limits the resolution with which the illuminating wavelength can be measured. To calculate the resolving power of the grating, we first determine the angular width of a principal maximum. This is accomplished by measuring the angular change, of the principal maximum's position, when β changes from $\beta = l\pi = m\pi/N$ to $\beta = (m+1)\pi/N$, i.e., $\Delta\beta = \pi/N$. Using the definition of β

$$\beta = \frac{kd\sin\theta_d}{2} \qquad [103]$$

$$\Delta\beta = \frac{\pi d\cos\theta_d\Delta\theta}{\lambda}$$

and the angular width is

$$\Delta\theta = \frac{\lambda}{Nd\cos\theta_d} \qquad [104]$$

The derivative of the grating formula gives

$$\Delta\lambda = \frac{d}{l}\cos\theta_d\Delta\theta \qquad [105]$$

$$\Delta\theta = \frac{l\Delta\lambda}{d\cos\theta_d}$$

Equating this spread in angle to eqn [104] yields

$$\frac{l\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \qquad [106]$$

The resolving power of a grating is therefore

$$\frac{\lambda}{\Delta\lambda} = Nl \qquad [107]$$

The improvement of resolving power with N can be seen in Figure 10. A grating 2 inches wide and containing 15 000 grooves per inch would have a resolving power in second order ($l = 2$) of 6×10^4. At a wavelength of 600 nm this grating could resolve two waves, differing in wavelength by 0.01 nm.

The diffraction grating is limited by overlapping orders. If two wavelengths, λ and $\lambda + \Delta\lambda$, have successive orders that are coincident, then

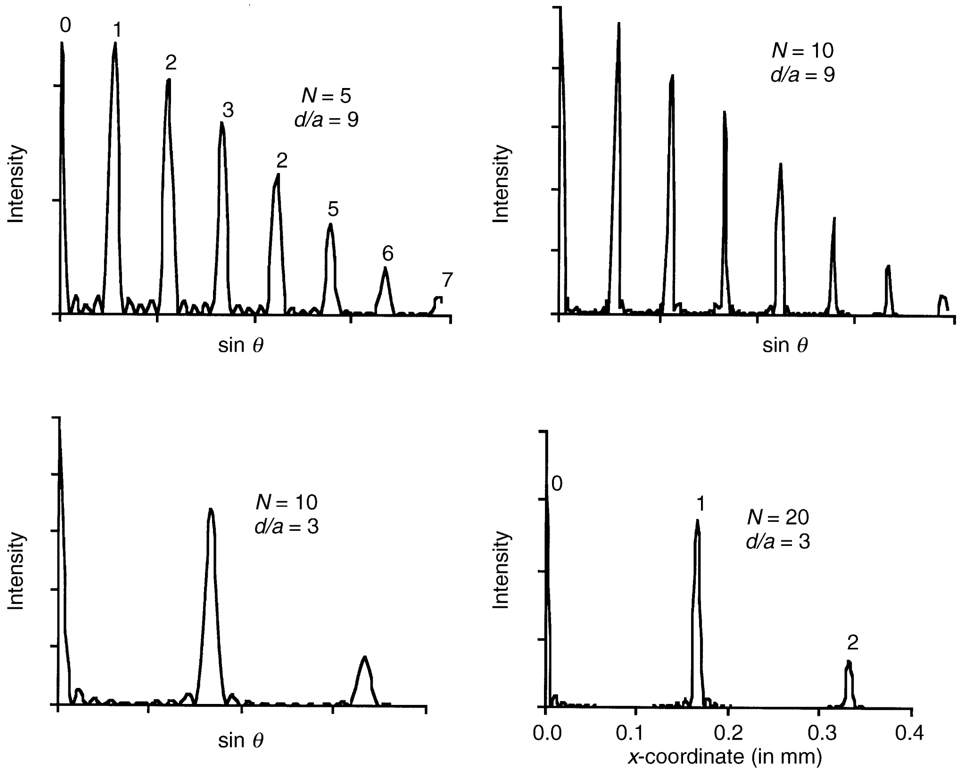

Figure 10 Decrease in the width of the principal maxima of a transmission grating with an increasing number of slits. The various principal maxima are called orders, numbering from zero, at the origin, out to as large as seven in this example. Also shown is the effect of different ratios, *d*/*a*, on the number of visible orders. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

from eqn [102]

$$(l + 1)\lambda = l(\lambda + \Delta\lambda) \qquad [108]$$

The minimum wavelength difference for which this occurs is defined as the free spectral range of the diffraction grating

$$(\Delta\lambda)_{\mathrm{SR}} = \frac{\lambda}{l} \qquad [109]$$

Blazed Gratings

We have been discussing amplitude transmission gratings. Amplitude transmission gratings have little practical use because they waste light. The light loss is from a number of sources.

1. Light is diffracted simultaneously into both positive and negative orders (the positive and negative frequencies of the Fourier transform). The negative diffraction orders contain redundant information and waste light.
2. In an amplitude transmission grating, light is thrown away because of the opaque portions of the slit array.
3. The width of an individual aperture leads to a shape factor for a rectangular aperture of $\mathrm{sinc}^2\alpha = (\sin^2\alpha)/\alpha^2$, which modulates the grating factor and causes the amplitude of the orders to rapidly decrease. This can be observed in **Figure 10**, where the second order is very weak. Because of the loss in intensity at higher orders, only the first few orders ($l = 1, 2$ or 3) are useful. The shape factor also causes a decrease in diffracted light intensity with increasing wavelength for higher-order principal maxima.
4. The location of the maximum in the diffracted light, i.e., the angular position for which the shape factor is a maximum, coincides with the location of the principal maximum due to the zero-order interference. This zero-order maximum is independent of wavelength and is not of much use.

One solution to the problems created by transmission gratings would be the use of a grating that modified only the phase of the transmitted wave. Such gratings would operate using the same physical processes as a microwave phased array antenna, where the location of the shape factor's maximum is controlled by adding a constant phase shift to each antenna element. The construction of an optical transmission phase grating with a uniform phase variation across the aperture of the grating is very difficult. For this reason, a second approach, based on the use of reflection gratings, is the practical solution to the problems listed above.

By tilting the reflecting surface of each groove of a reflection grating, Figure 11, the position of the shape factor's maximum can be controlled. Problems 1, 3, and 4 are eliminated because the shape factor maximum is moved from the optical axis out to some angle with respect to the axis. The use of reflection gratings also removes Problem 2 because all of the incident light is reflected by the grating.

Robert Wood, in 1910, developed the technique of producing grooves of a desired shape in a reflective grating by shaping the diamond tool used to cut the grooves. Gratings, with grooves shaped to enhance their performance at a particular wavelength, are said to be blazed for that wavelength. The physical properties on which blazed gratings are based can be understood by using Figure 11. The groove faces can be treated as an array of mirror surfaces. The normal to each of the groove faces makes an angle θ_B with the normal to the grating surface. We can measure the angle of incidence and the angle of diffraction with respect to the grating normal or with respect to the groove normal, as shown in Figure 11. From Figure 11, we can write a relationship between the angles

$$\theta_i = \varphi_i - \theta_B \qquad -\theta_d = -\varphi_d + \theta_B \tag{110}$$

(The sign convention used defines positive angles as those measured in a counterclockwise rotation from the normal to the surface. Therefore, θ_B is a negative angle.) The blaze angle provides an extra degree of freedom that will allow independent adjustment of the angular location of the principal maxima of the grating factor and the zero-order, single-aperture, diffraction maximum. To see how this is accomplished, we must determine, first, the effect of off-axis illumination of a diffraction grating.

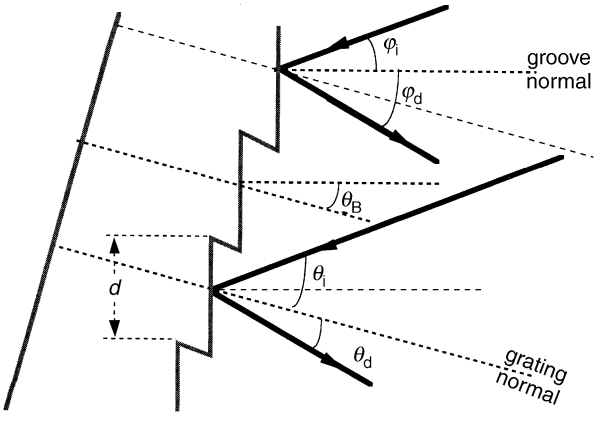

Figure 11 Geometry for a blazed reflection grating. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Off-axis illumination is easy to incorporate into the equation for the diffraction intensity from an array. To include the effect of an off-axis source, the phase of the illuminating wave is modified by changing the incident illumination from a plane wave of amplitude, E, traveling parallel to the optical axis, to a plane wave with the same amplitude, traveling at an angle θ_i to the optical axis: $\tilde{E}\,e^{-ikx\sin\theta_i}$. (Because we are interested only in the effects in a plane normal to the direction of propagation, we ignore the phase associated with propagation along the z-direction, $kz\cos\theta_i$.) The off-axis illumination results in a modification of the parameter for single-aperture diffraction from

$$\alpha = \frac{ka}{2}\sin\theta_d \tag{111}$$

to

$$\alpha = \frac{ka}{2}(\sin\theta_i + \sin\theta_d) \tag{112}$$

and for multiple aperture interference from

$$\beta = \frac{kd}{2}\sin\theta_d \tag{113}$$

to

$$\beta = \frac{kd}{2}(\sin\theta_i + \sin\theta_d) \tag{114}$$

The zero-order, single-aperture, diffraction peak occurs when $\alpha = 0$. If we measure the angles with respect to the groove face, this occurs when

$$\alpha = \frac{ka}{2}(\sin\varphi_i + \sin\varphi_d) = 0 \tag{115}$$

The angles are therefore related by

$$\sin\varphi_i = -\sin\varphi_d \tag{116}$$

$$\varphi_i = -\varphi_d$$

We see that the single-aperture, diffraction maximum (the shape factor's maximum) occurs at the same angle that reflection from the groove faces occurs. We can write this result in terms of the angles measured with respect to the grating normal

$$\theta_i = -(\theta_d + 2\theta_B) \tag{117}$$

The blaze condition requires the single aperture diffraction maximum to occur at the lth principal maximum, for wavelength λ_B. At that position

$$l\pi = \frac{2\pi d}{\lambda_B}(\sin\theta_i + \sin\theta_d) \qquad [118]$$

$$l\lambda_B = 2d\sin\tfrac{1}{2}(\theta_i + \theta_d)\cos\tfrac{1}{2}(\theta_i - \theta_d)$$

For the special geometrical configuration called the Littrow condition, where $\theta_i = \theta_d$, we find that [117] leads to the equation

$$l\lambda_B = 2d\sin\theta_B \qquad [119]$$

By adjusting the blaze angle, the single-aperture diffraction peak can be positioned on any order of the interference pattern. Typical blaze angles are between 15° and 30° but gratings are made with larger blaze angles.

See also

Diffraction: Fresnel Diffraction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*. 3rd edn. Reading, MA: Addison-Wesley.

Jackson JD (1962) *Classical Electrodynamics*. New York: Wiley.

Klein MV (1970) *Optics*. New York: Wiley.

O'Neill EL (1963) *Introduction to Statistical Optics*. Reading, MA: Addison-Wesley.

Rossi B (1957) *Optics*. Reading, MA: Addison-Wesley.

Rubinowicz A (1984) Tahe Miyamoto–Wolf diffraction wave. In: Wolf E (ed.) *Progress in Optics*, vol. IV, p. 201. Amsterdam: North-Holland.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

Fresnel Diffraction

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Fresnel was a civil engineer who pursued optics as a hobby, after a long day of road construction. Based on his own very high-quality observations of diffraction, Fresnel used the wave propagation concept developed by Christiaan Huygens (1629–1695) to develop a theoretical explanation of diffraction. We will use a descriptive approach to obtain the Huygens–Fresnel integral by assuming an aperture can be described by N pinholes which act as sources for Huygens' wavelets. The interference between these sources will lead to the Huygens–Fresnel integral for diffraction.

Huygens' principle views wavefronts as the product of wavelets from various luminous points acting together. To apply Huygens' principle to the propagation of light through an aperture of arbitrary shape, we need to develop a mathematical description of the field from an array of Huygens' sources filling the aperture. We will begin by obtaining the field from a pinhole that is illuminated by a plane wave,

$$\tilde{\mathrm{E}}_i(\mathbf{r}, t) = \tilde{\mathrm{E}}_i(\mathbf{r})e^{i\omega t}$$

Following the lead of Fresnel, we will use theory of interference to combine the fields from two pinholes and then generalize to N pinholes. Finally by letting the areas of each pinhole approach an infinitesimal value, we will construct an arbitrary aperture of infinitesimal pinholes. The result will be the Huygens–Fresnel integral.

We know that the wave obtained after propagation through an aperture must be a solution of the wave equation,

$$\nabla^2\tilde{\mathrm{E}} = \mu\varepsilon\frac{\partial^2\tilde{\mathrm{E}}}{\partial t^2}$$

We will be interested only in the spatial variation of the wave so we need only look for solutions of the Helmholtz equation

$$(\nabla^2 + k^2)\tilde{\mathrm{E}} = 0$$

The problem is further simplified by replacing this vector equation with a scalar equation,

$$(\nabla^2 + k^2)\tilde{E}(x, y, z) = 0$$

This replacement is proper for those cases where $\mathbf{n}E(x,y,z)$ [where $\mathbf{n}$ is a unit vector] is a solution of the vector Helmholtz equation. In general, we cannot substitute $\mathbf{n}E$ for the electric field $\mathbf{E}$ because of Maxwell's equation

$$\nabla\cdot\mathbf{E} = 0$$

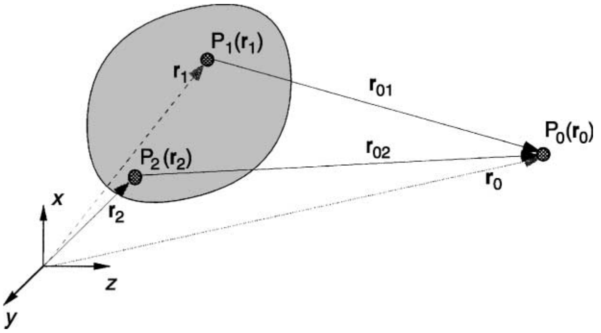

Figure 1 Geometry for application of Huygens' principle to two pinholes. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

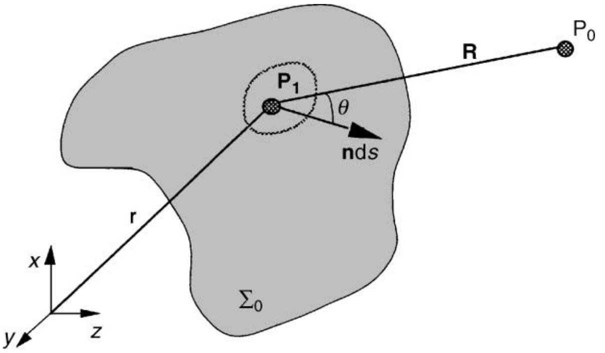

Figure 2 The geometry for calculating the field at P$_0$ using [9]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Rather than working with the magnitude of the electric field, we should use the scalar amplitude of the vector potential. We will neglect this complication and assume the scalar, E, is a single component of the vector field, $\mathbf{E}$. A complete solution would involve a scalar solution for each component of $\mathbf{E}$.

The pinhole is illuminated by a plane wave, and the wave that leaves the pinhole will be a spherical wave which is written in complex notation as

$$\tilde{E}(\mathbf{r})\mathrm{e}^{\mathrm{i}\omega t} = A\frac{\mathrm{e}^{-\mathrm{i}\delta}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}}{r}\mathrm{e}^{\mathrm{i}\omega t}$$

The complex amplitude

$$\tilde{E}(\mathbf{r}) = A\frac{\mathrm{e}^{-\mathrm{i}\delta}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}}{r} \qquad [1]$$

is a solution of the Helmholtz equation. The field at P$_0$, in Figure 1, from two pinholes: one at P$_1$, located a distance $\mathbf{r}_{01} = |\mathbf{r}_0 - \mathbf{r}_1|$ from P$_0$, and one at P$_2$, located a distance $\mathbf{r}_{02} = |\mathbf{r}_0 - \mathbf{r}_2|$ from P$_0$, is a generalization of Young's interference. The complex amplitude is

$$\tilde{E}(\mathbf{r}_0) = \frac{A_1}{r_{01}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}} + \frac{A_2}{r_{02}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{02}} \qquad [2]$$

We have incorporated the phase δ_1 and δ_2 into the constants A_1 and A_2 to simplify the equations.

The light emitted from the pinholes is due to a wave, E_i, incident onto the screen from the left. The units of E_i are per unit area so to obtain the amount of light passing through the pinholes, we must multiply E_i by the areas of the pinholes. If $\Delta\sigma_1$ and $\Delta\sigma_2$ are the areas of the two pinholes, respectively, then

$$A_1 \propto \tilde{E}_\mathrm{i}(\mathbf{r}_1)\Delta\sigma_1 \qquad A_2 \propto \tilde{E}_\mathrm{i}(\mathbf{r}_2)\Delta\sigma_2$$

$$\tilde{E}(\mathbf{r}_0) = C_1\frac{\tilde{E}_\mathrm{i}(\mathbf{r}_1)}{r_{01}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{01}}\Delta\sigma_1 + C_2\frac{\tilde{E}_\mathrm{i}(\mathbf{r}_2)}{r_{02}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{02}}\Delta\sigma_2 \qquad [3]$$

where C_i is the constant of proportionality. The constant will depend on the angle that $\mathbf{r}_{0i}$ makes with the normal to $\Delta\sigma_i$. This geometrical dependence arises from the fact that the apparent area of the pinhole decreases as the observation angle approaches 90°.

We can generalize [3] to N pinholes

$$\tilde{E}(\mathbf{r}_0) = \sum_{j=1}^{N} C_j\frac{\tilde{E}_\mathrm{i}(\mathbf{r}_j)}{r_{0j}}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_{0j}}\Delta\sigma_j \qquad [4]$$

The pinhole's diameter is assumed to be small compared to the distances to the viewing position but large compared to a wavelength. In the limit as $\Delta\sigma_j$ goes to zero, the pinhole becomes a Huygens source. By letting N become large, we can fill the aperture with these infinitesimal Huygens' sources and convert the summation to an integral.

It is in this way that we obtain the complex amplitude, at the point P$_0$, see Figure 2, from a wave exiting an aperture, Σ_0, by integrating over the area of the aperture. We replace $\mathbf{r}_{0j}$, in [4], by $\mathbf{R}$, the position of P$_1$, the infinitesimal Huygens' source of area, ds, measured with respect to the observation point, P$_0$. We also replace $\mathbf{r}_j$ by $\mathbf{r}$, the position of the infinitesimal area, ds, with respect to the origin of the coordinate system. The discrete summation [4] becomes the integral

$$\tilde{E}(\mathbf{r}_0) = \iint_A C(\mathbf{r})\frac{\tilde{E}_\mathrm{i}(\mathbf{r})}{R}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}\mathrm{d}s \qquad [5]$$

This is the Fresnel integral. The variable $C(\mathbf{r})$ depends upon θ, the angle between $\mathbf{n}$, the unit vector normal to the aperture, and $\mathbf{R}$, shown in Figure 2. We now need to determine how to treat $C(\mathbf{r})$.

The Obliquity Factor

When using Huygens' principle, a problem arises with the spherical wavelet produced by each

Huygens' source. Part of the spherical wavelet propagates backward and results in an envelope propagating toward the light source, but such a wave is not observed in nature. Huygens neglected this problem. Fresnel required the existence of an obliquity factor to cancel the backward wavelet but was unable to derive its functional form. When Kirchhoff placed the theory of diffraction on a firm mathematical foundation, the obliquity factor was generated quite naturally in the derivation. In this intuitive derivation of the Huygens–Fresnel integral, we will present an argument for treating the obliquity factor as a constant at optical wavelengths. We will then derive the constant value it must be assigned.

The obliquity factor, $C(\mathbf{r})$, in [5], causes the amplitude per unit area of the transmitted light to decrease as the viewing angle increases. This is a result of a decrease in the effective aperture area with viewing angle. When Kirchhoff applied Green's theorem to the scalar diffraction problem, he found the obliquity factor to have an angular dependence given by

$$\frac{\cos(\hat{\mathbf{n}}, \mathbf{R}) - \cos(\hat{\mathbf{n}}, \mathbf{r}_s)}{2} \qquad [6]$$

which includes the geometrical effect of the incident wave arriving at the aperture, at an angle with respect to the normal to the aperture from a source located at a position $\mathbf{r}_s$, with respect to the aperture. If the source is at infinity, then the incident wave can be treated as a plane wave, incident normal to the aperture, and we may simplify the angular dependence of the obliquity factor to

$$\frac{1 + \cos\theta}{2}$$

where $\theta \rightarrow (\hat{\mathbf{n}}, \mathbf{R})$ is the angle between the normal to the aperture and the vector $\mathbf{R}$. This is the configuration shown in Figure 2. The obliquity factor provides an explanation of why it is possible to ignore the backward-propagating wave that occurs in application of Huygens' principle. For the backward wave, $\theta = \pi$, and the obliquity factor is zero.

The obliquity factor increases the difficulty of working with the Fresnel integral and it is to our benefit to be able to treat it as a constant. We can neglect the angular contribution of the obliquity factor by making an assumption about the resolving power of an optical system operating at visible wavelengths.

Assume we are attempting to resolve two stars that produce plane waves at the aperture of a telescope with an aperture diameter, a. The wavefronts from the two stars make an angle ψ with respect to each other, Figure 3:

$$\tan\psi \approx \psi = \frac{\Delta x}{a}$$

The smallest angle, ψ, that can be measured is determined by the smallest length, Δx, that can be measured. We know we can measure a fraction of wavelength with an interferometer but, without an interferometer, we can measure a length no smaller than λ, leading to the assumption that $\Delta x \geq \lambda$. This reasoning leads to the assumption that the smallest angle we can measure is

$$\psi \geq \frac{\lambda}{a} \qquad [7]$$

The resolution limit established by the above reasoning places a limit on the minimum separation that can be produced at the back focal plane of the telescope. The minimum distance on the focal plane between the images of star 1 and 2 is given by

$$d = f\psi \qquad [8]$$

From [7]

$$d = \frac{\lambda f}{a}$$

The resolution limit of the telescope can also be expressed in terms of the cone angle produced when the incident plane wave is focused on the back focal plane of the lens. From the geometry of Figure 3, the cone angle is given by

$$\tan\theta = \frac{a}{2f}$$

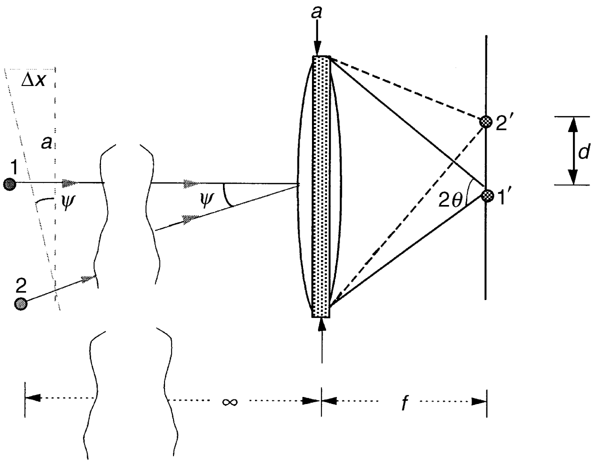

Figure 3 Telescope resolution. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The separation between the stars 1 and 2 at the back focal plane of the lens in Figure 3 is thus

$$d = \frac{\lambda}{2\tan\theta} \qquad [9]$$

If we assume that the minimum d is 3λ (this is four times the resolution of a typical photographic film at visible wavelengths), then the largest obliquity angle that should be encountered in a visible imaging system is

$$\tan\theta = \frac{\lambda}{2d} = \frac{\lambda}{6\lambda} = \tfrac{1}{6} = 0.167$$

$$\theta = 9.5^\circ \approx 10^\circ$$

yielding a value for $\cos\theta = 0.985$ and $(1 + \cos\theta)/2 = 0.992$. The obliquity factor has only changed by 0.8% over the angular variation of 0° to 10°. The obliquity factor as a function of angle is shown in Figure 4 for an incident plane wave. The obliquity factor undergoes a 10% change when θ varies from 0° to about 40°; therefore, the obliquity factor can safely be treated as a constant in any optical system that involves angles less than 40°.

While we have shown that it is safe to ignore the variability of C, we still have not assigned a value to the obliquity factor. To find the proper value for C, we will compare the result obtained using [5] with the result predicted by using geometric optics.

We illuminate the aperture Σ_0 in Figure 5 with a plane wave of amplitude α, traveling parallel to the z-axis. Geometrical optics (equivalently the propagation of an infinite plane wave) predicts a field at P_0, on the z-axis, a distance z_0 from the aperture, given by

$$\tilde{E}_{\text{geom}} = \alpha e^{-ikz_0} \qquad [10]$$

The area of the infinitesimal at P_1 (the Huygens' source) is

$$ds = r\,dr\,d\phi$$

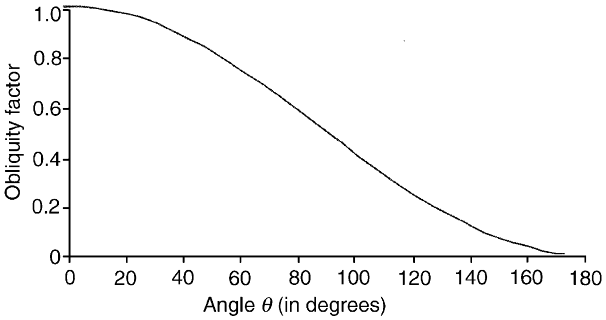

Figure 4 The obliquity factor as a function of the angle defined in eqn [6] for an incident plane wave. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

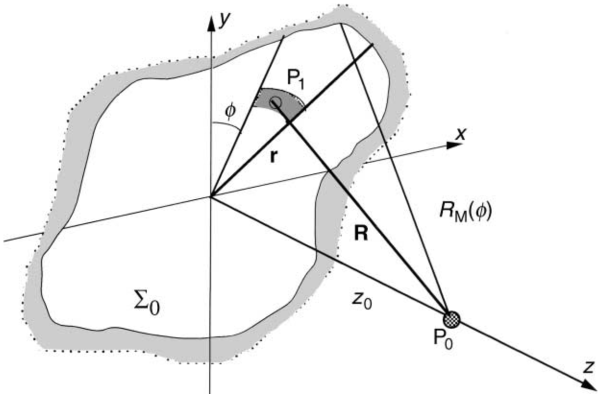

Figure 5 Geometry for evaluating the constant in the Fresnel integral. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Based on our previous argument, the obliquity factor can be treated as a constant, C, that can be removed from under the integral. The incident wave is a plane wave whose value at $z = 0$ is $E(r) = \alpha$. Using these parameters, the Fresnel integral [5] can be written as

$$\tilde{E}(z_0) = C\alpha \iint \frac{e^{-ik\cdot R}}{R} r\,dr\,d\phi \qquad [11]$$

From the geometry in Figure 5, the distance from the Huygens source to the observation point is

$$z_0^2 + r^2 = R^2$$

where z_0 is a constant equal to the distance from the observation plane to the aperture plane. The variable of integration can be written in terms of R:

$$r\,dr = R\,dR$$

The limits of integration over the aperture extend from $R = z_0$ to the maximum value of R, $R = R_M(\phi)$.

$$\tilde{E}(z_0) = C\alpha \int_0^{2\pi} \int_{z_0}^{R_M(\phi)} e^{-ikR} dR\,d\phi \qquad [12]$$

The integration over R can now be carried out to yield

$$\tilde{E}(z_0) = \underbrace{\frac{C\alpha}{ik} e^{-ikz_0} \int_0^{2\pi} d\phi}_{\text{Geometrical Optics}} - \underbrace{\frac{C\alpha}{ik} \int_0^{2\pi} e^{-ikR_M(\phi)} d\phi}_{\text{Diffraction}} \qquad [13]$$

The first integration in [13] is easy to perform and contains the amplitude at the observation point due to geometric optics. The second term may be interpreted as interference of the waves diffracted by the boundary of the aperture. An equivalent statement is that the second term is the interference of the waves scattered from the aperture's boundary, an interpretation of diffraction that was first suggested by Young.

The aperture is irregular in shape, at least on the scale of a wavelength, thus in general, kR_M will vary over many multiples of 2π as we integrate around the aperture. For this reason, we should be able to ignore the second integral in [13] if we confine our attention to the light distribution on the z-axis. After neglecting the second term, we are left with only the geometrical optics component of [13]:

$$\tilde{E}(z_0) = \frac{2\pi C}{ik} \alpha e^{-ikz_0} \qquad [14]$$

For [14] to agree with the prediction of geometric optics [10] the constant C must be equal to

$$C = \frac{ik}{2\pi} = \frac{i}{\lambda} \qquad [15]$$

The Huygens–Fresnel integral can be written, using the value for C just derived,

$$\tilde{E}(\mathbf{r}_0) = \frac{i}{\lambda} \int\int_{\Sigma} \frac{\tilde{E}_i(\mathbf{r})}{R} e^{-ik\cdot R}\, ds \qquad [16]$$

The job of calculating diffraction has only begun with the derivation of the Huygens–Fresnel integral [16]. Rigorous solutions exist for only a few idealized obstructions. To allow discussion of the general properties of diffraction, it is necessary to use approximate solutions. To determine what type approximations we can make let's look at the light propagation path shown in Figure 6.

For the second term in [13] to contribute to the field at point P_0, in Figure 6, the phase of the exponent that we neglected in [13] must not vary over 2π when the integration is performed over the aperture, i.e.,

$$\Delta = \mathbf{k}\cdot(\mathbf{R}_1 - \mathbf{R'}_1 + \mathbf{R}_2 - \mathbf{R'}_2) < 2\pi$$

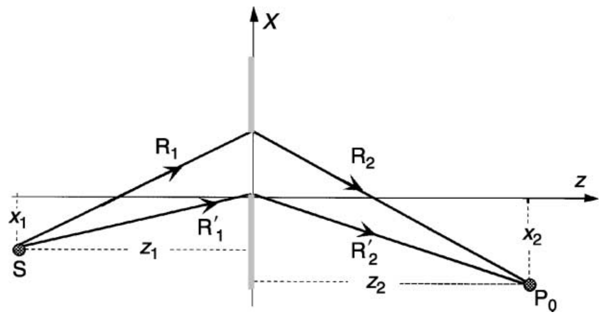

Figure 6 One-dimensional slit used to establish Fresnel approximations. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

From the geometry of Figure 6, the two paths from the source, S, to the observation point, P_0, are

$$R_1 + R_2 = \sqrt{z_1^2 + (x_1 + b)^2} + \sqrt{z_2^2 + (x_2 + b)^2}$$

$$R_1' + R_2' = \sqrt{z_1^2 + x_1^2} + \sqrt{z_2^2 + x_2^2}$$

By assuming that the aperture, of width b, is small compared to the distances z_1 and z_2, the difference between these distances, Δ, can be rewritten as a binomial expansion of a square root

$$\sqrt{1+b} = 1 + \frac{b}{2} - \frac{b^2}{8} + \ldots \qquad [17]$$

thus

$$\Delta \approx k\left(\frac{x_1}{z_1} + \frac{x_2}{z_2}\right)b + \frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 + \ldots \qquad [18]$$

If we assume that the second term of this expansion is small, i.e.,

$$\frac{k}{2}\left(\frac{1}{z_1} + \frac{1}{z_2}\right)b^2 \ll 2\pi$$

$$\frac{b^2}{2}\frac{1}{z_1} + \frac{1}{z_2} < \lambda$$

we see that the phase of the wavefront in the aperture is assumed to have a quadratic dependence upon aperture coordinates. Diffraction predicted by this assumption is called Fresnel diffraction.

To discover the details of the approximation consider Figure 7. In the geometry of Figure 7, the Fresnel integral [16] becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda} \int\int_{\Sigma} f(x, y) \frac{e^{-ik\cdot(R+R')}}{RR'}\, dx\, dy \qquad [19]$$

As we mentioned in our discussion of [13] the integral that adds diffractive effects to the geometrical optics field is nonzero only when the phase of the integrand is stationary. For Fresnel diffraction, we can insure that the phase is nearly constant if R does not differ appreciably from Z, or R' from Z'. This is equivalent to stating that only the wave in the aperture, around a point in Figure 7 labeled **S**, called the stationary point, will contribute to E_p. The stationary point, **S**, is the point in the aperture plane where the line, connecting the source and observation positions, intersects the plane. Physically, only light propagating over paths nearly equal to the path predicted by geometrical optics (obtained from Fermat's principle) will contribute to E_{P_0}.

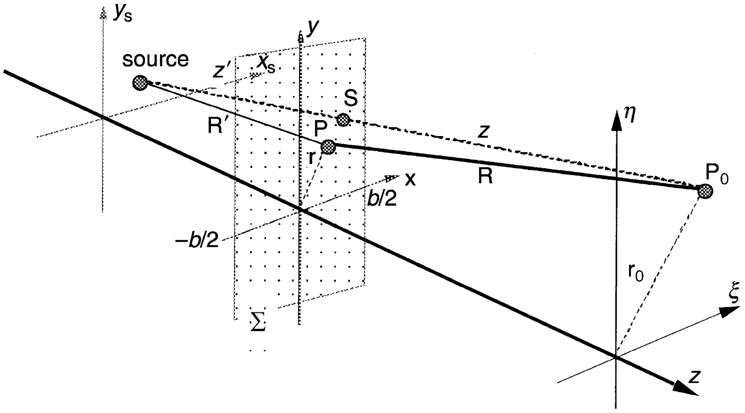

Figure 7 Geometry for Fresnel diffraction. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The geometry of Figure 7 allows the distances R and R' to be written as

$$R^2 = (x - \xi)^2 + (y - \eta)^2 + Z^2 \qquad [20]$$

$$R'2 = (x_s - x)^2 + (y_s - y)^2 + Z'2$$

To solve these expressions for R and R', we apply the binomial expansion [17] and retain the first two terms:

$$R + R' \approx Z + Z' + \left[\frac{(x-\xi)^2}{2Z} + \frac{(x_s - x)^2}{2Z'} \right] + \left[\frac{(y-\eta)^2}{2Z} + \frac{(y_s - y)^2}{2Z'} \right] \qquad [21]$$

In the denominator of [19], R is replaced by Z and R' by Z'. (By making this replacement, we are implicitly assuming that the amplitudes of the spherical waves are not modified by the differences in propagation distances over the area of integration. This is a reasonable approximation because the two propagation distances are within a few hundred wavelengths of each other). In the exponent, we must use [21], because the phase changes by a large fraction of a wavelength, as we move from point to point in the aperture.

If the wave, incident on the aperture, were a plane wave we can assume $R' = \infty$ and [19] becomes

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda} \frac{e^{-ikZ}}{Z} \int\int_{\Sigma} f(x, y) e^{-\frac{ik}{2Z}[(x-\xi)^2 + (y-\eta)^2]} dx\, dy \qquad [22]$$

The physical interpretation of [22] states that when a plane wave illuminates the obstruction, the field at point P_0 is a spherical wave, originating at the aperture, a distance Z away from P_0:

$$\frac{e^{-ikZ}}{Z}$$

The amplitude and phase of this spherical wave are modified by the integral containing a quadratic phase dependent on the obstruction's spatial coordinates.

By defining three new parameters,

$$\rho = \frac{ZZ'}{Z + Z'} \quad \text{or} \quad \frac{1}{\rho} = \frac{1}{Z} + \frac{1}{Z'} \qquad [23a]$$

$$x_0 = \frac{Z'\xi + Zx_s}{Z + Z'} \qquad [23b]$$

$$y_0 = \frac{Z'\eta + Zy_s}{Z + Z'} \qquad [23c]$$

the more general expression for Fresnel diffraction of a spherical wave can be placed in the same format as [22]. The newly defined parameters x_0 and y_0 are the coordinates, in the aperture plane, of the stationary point, S. The parameters in [23] can be used to express, after some manipulation, the spatial dependence of the phase, in the integrand of [19]

$$R + R' = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} + \left[\frac{(x - x_0)^2 + (y - y_0)^2}{2\rho} \right]$$

A further simplification to the Fresnel diffraction integral can be made by obtaining an expression for the distance, D, between the source and the observation point:

$$D = Z + Z' + \frac{(\xi - x_s)^2 + (\eta - y_s)^2}{2(Z + Z')} \qquad [24]$$

We use this definition of D to also write

$$\frac{1}{ZZ'} = \frac{Z+Z'}{ZZ'}\,\frac{1}{Z+Z'} \approx \frac{1}{\rho D}$$

Using the parameters we have just defined, we may rewrite [19] as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D}e^{-ikD}\iint f(x,y)e^{-\frac{ik}{2\rho}\left[(x-x_0)^2+(y-y_0)^2\right]}dx\,dy \qquad [25]$$

With the use of the variables defined in [23], the physical significance of the general expression of the Huygens–Fresnel integral can be understood in an equivalent fashion as was [22]. At point P_0, a spherical wave

$$\frac{e^{-ikD}}{D}$$

originating at the source, a distance D away, is observed, if no obstruction are present. Because of the obstruction, the amplitude and phase of the spherical wave are modified by the integral in [25]. This correction to the predictions of geometrical optics is called Fresnel diffraction.

Mathematically, [25] is an application of the method of stationary phase, a technique developed in 1887 by Lord Kelvin to calculate the form of a boat's wake. The integration is nonzero only in the region of the critical point we have labeled **S**. Physically, the light distribution, at the observation point, is due to wavelets from the region around **S**. The phase variations of light, coming from other regions in the aperture, are so rapid that the value of the integral over those spatial coordinates is zero. The calculation of the integral for Fresnel diffraction is complicated because, if the observation point is moved, a new integration around a different stationary point in the aperture must be performed.

Rectangular Apertures

If the aperture function, $f(x,y)$, is separable in the spatial coordinates of the aperture, then we can rewrite [25] as

$$\tilde{E}_{P_0} = \frac{i\alpha}{\lambda\rho D}e^{-ikD}\int_{-\infty}^{\infty} f(x)e^{-ig(x)}dx\int_{-\infty}^{\infty} f(y)e^{-ig(y)}dy$$

$$\tilde{E}_{P_0} = A[C(x) - iS(x)][C(y) - iS(y)]$$

A represents the spherical wave from the source, a distance D away,

$$A = \frac{i\alpha}{2D}e^{-ikD}$$

If we treat the aperture function as a simple constant, C and S are integrals of the form

$$C(x) = \int_{x_1}^{x_2}\cos[g(x)]dx \qquad [26a]$$

$$S(x) = \int_{x_1}^{x_2}\sin[g(x)]dx \qquad [26b]$$

The integrals, $C(x)$ and $S(x)$, have been evaluated numerically and are found in tables of Fresnel integrals. To use the tabulated values for the integrals, [26] must be written in a general form

$$C(w) = \int_0^w \cos\left(\frac{\pi u^2}{2}\right)du \qquad [27]$$

$$S(w) = \int_0^w \sin\left(\frac{\pi u^2}{2}\right)du \qquad [28]$$

The variable u is an aperture coordinate, measured relative to the stationary point, $\mathbf{S}(x_0,y_0)$, in units of

$$\sqrt{\frac{\lambda\rho}{2}}$$

$$u = \sqrt{\frac{2}{\lambda\rho}}(x - x_0) \text{ or } u = \sqrt{\frac{2}{\lambda\rho}}(y - y_0) \qquad [29]$$

The parameter w in [27] and [28] specifies the location of the aperture edge relative to the stationary point **S**. The parameter w is calculated through the use of [29] with x and y replaced by the coordinates of the aperture edge.

The Cornu Spiral

The plot of $S(w)$ versus $C(w)$, shown in Figure 8, is called the Cornu spiral in honor of M. Alfred Cornu (1841–1902) who was the first to use this plot for graphical evaluation of the Fresnel integrals.

To use the Cornu spiral, the limits w_1 and w_2 of the aperture are located along the arc of the spiral. The length of the straight line segment drawn from w_1 to w_2 gives the magnitude of the integral. For example, if there were no aperture present, then, for the x-dimension, $w_1 = -\infty$ and $w_2 = \infty$. The length of the line segment from the point $(-1/2,1/2)$ to $(1/2,1/2)$ would be the value of E_{P_0}, i.e., $\sqrt{2}$. An identical value is obtained for the y-dimension, so that

$$\tilde{E}_{P_0} = 2A = \frac{\alpha e^{-ikD}}{D}$$

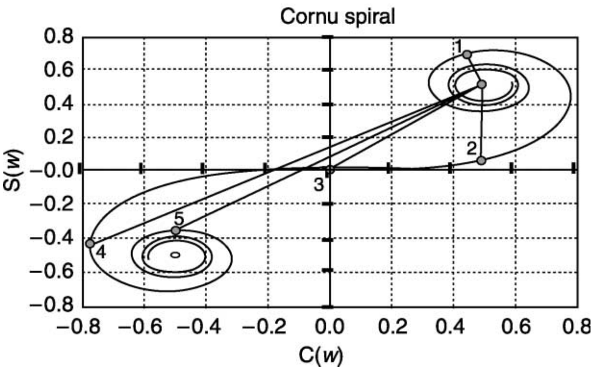

Figure 8 Cornu spiral obtained by plotting the values for *C*(*w*) versus *S*(*w*) from a table of Fresnel integrals. The numbers indicated along the arc of the spiral are values of *w*. The points 1–5 correspond to observation points shown in Figure 9 and are used to calculate the light intensity diffracted by a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

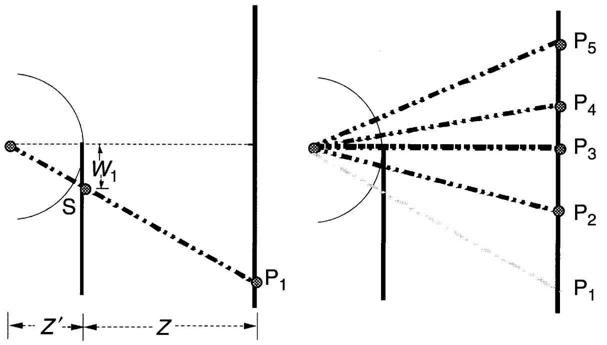

Figure 9 A geometrical construct to determine w_1 for the stationary point S associated with five different observation points. The values of *w* are then used in Figure 8 to calculate the intensity of light diffracted around a straight edge. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Table 1 Fresnel integrals for straight edge

w_1	$C(w_1)$	$S(w_1)$	I_P/I_0	P_j
∞	0.5	0.5	0	
2.0	0.4882	0.3434	0.01	
1.5	0.4453	0.6975	0.021	P_1
1.0	0.7799	0.4383	0.04	
0.5	0.4923	0.0647	0.09	P_2
0	0	0	0.25	P_3
−0.5	−0.4923	−0.0647	0.65	
−1.0	−0.7799	−0.4383	1.26	P_4
−1.2	−0.7154	−0.6234	1.37	
−1.5	−0.4453	−0.6975	1.16	
−2.0	−0.4882	−0.3434	0.84	P_5
−2.5	−0.4574	−0.6192	1.08	
−∞	−0.5	−0.5	1.0	

This is the spherical wave that is expected when no aperture is present.

As an example of the application of the Fresnel integrals, assume the diffracting aperture is an infinitely long straight edge reducing the problem to one dimension:

$$f(x, y) = \begin{Bmatrix} 1 & x \geq x_1 \\ 0 & x < x_1 \end{Bmatrix}$$

The value of E_{P_0} in the y-dimension is from above $\sqrt{2}$ since $w_1 = -\infty$ and $w_2 = \infty$. In the x-dimension, the value of w at the edge is

$$w_{x_1} = \sqrt{\frac{2}{\lambda\rho}}(x_1 - x_0)$$

We will assume that the straight edge blocks the negative half-plane so that $x_1 = 0$. The straight edge is treated as an infinitely wide slit with one edge located at infinity so that $w_2 = \infty$ in the upper half-plane. When the observation point is moved, the coordinate x_0 of S changes and the origin of the aperture's coordinate system moves. New values for the w's must therefore be calculated for each observation point.

Figure 9 shows the assumed geometry. A set of observation points, P_1–P_5, on the observation screen are selected. The origin of the coordinate system in the aperture plane is relocated to a new position (given by the position of S) when a new observation point is selected. The value of w_1, the position of the edge with respect to S, must be recalculated for each observation point. The distance from the origin to the straight edge, w_1, is positive for P_1 and P_2, zero for P_3, and negative for P_4 and P_5.

Figure 8 shows the geometrical method used to calculate the intensity values at each of the observation points in Figure 9 using the Cornu spiral. The numbers labeling the straight line segments in Figure 8 are associated with the labels of the observation points in Figure 9.

To obtain an accurate calculation of Fresnel diffraction from a straight edge, a table of Fresnel integrals provides the input to the following equation

$$I_P = I_0\left\{\left[\tfrac{1}{2} - C(w_1)\right]^2 + \left[\tfrac{1}{2} - S(w_1)\right]^2\right\}$$

where $I_0 = 2A^2$. Table 1, shows the values extracted from the table of Fresnel integrals to find the relative intensity at various observation points. The result obtained by using either method for calculating the light distribution in the observation plane, due to the straight edge in Figure 9, is plotted in Figure 10. The relative intensities at the observation points

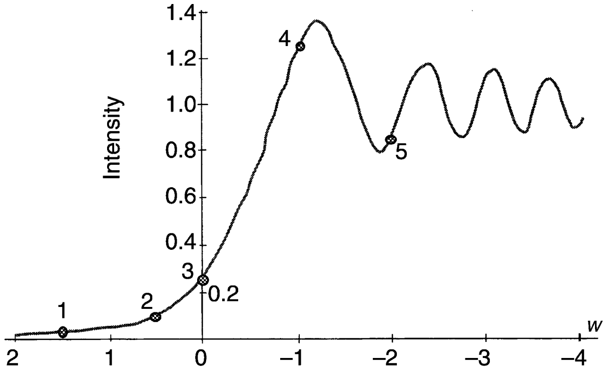

Figure 10 Light diffracted by a straight edge with its edge located at $w = 0$. The points labeled 1–5 were found using the construction in Figure 9 to obtain w. This was then used in Figure 8 to find a position on the Cornu spiral. The length of the lines shown in Figure 8 from the (1/2,1/2) point to the numbered positions led to the intensities shown. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

depicted in Figure 9 are labeled on the diffraction curve of Figure 10.

Fresnel Zones

Fresnel used a geometrical construction of zones to evaluate the Huygens–Fresnel integral. The Fresnel zone is a mathematical construct, serving the role of a Huygens source in the description of wave propagation. Assume that at time t, a spherical wavefront from a source at P_1 has a radius of R'. To determine the field at the observation point, P_0, due to this wavefront, a set of concentric spheres of radii, $Z, Z + (\lambda/2), Z + 2(\lambda/2), \ldots, Z + j(\lambda/2), \ldots$ are constructed, where Z is the distance from the wavefront to the observation point on the line connecting P_1 and P_0 (see Figure 11). These spheres divide the wavefront into a number of zones, $\zeta_1, \zeta_2, \ldots, \zeta_j \ldots$, called Fresnel zones, or half-period zones.

We treat each zone as a circular aperture illuminated from the left by a spherical wave of the form

$$\tilde{E}_j(\mathbf{R}') = \frac{A\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}'}}{R'} = \frac{A\mathrm{e}^{-ikR'}}{R'}$$

R' is the radius of the spherical wave. The field at P_0 due to the jth zone is obtained by using [5]

$$\tilde{E}_j(P_0) = \frac{A}{R'}\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}'} \iint_{\zeta_j} C(\varphi)\frac{\mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{R}}}{R}\mathrm{d}s \qquad [30]$$

For the integration over the jth zone, the surface element is

$$\mathrm{d}s = R'2\sin\theta\mathrm{d}\theta\,\mathrm{d}\phi \qquad [31]$$

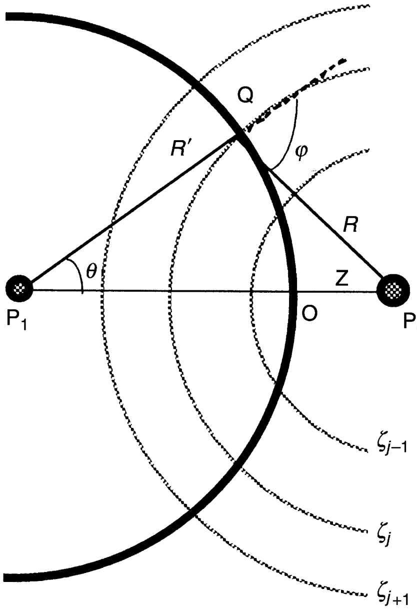

Figure 11 Construction of Fresnel zones observed from a position P_0 on a spherical wave originating from a source at point P_1. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

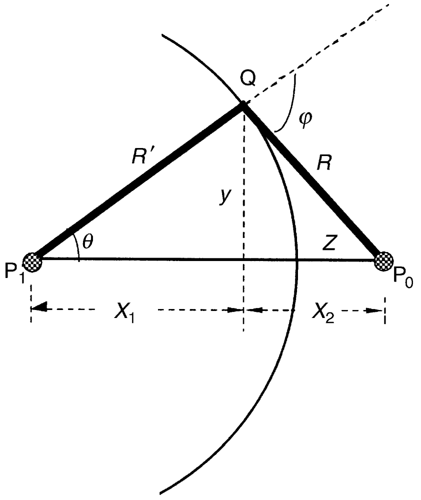

Figure 12 Geometry for finding the relationship between R and R' yielding eqn [32]. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The limits of integration extend over the range

$$Z + (j-1)\tfrac{\lambda}{2} \le R \le Z + j\tfrac{\lambda}{2}$$

The variable of integration is R; thus, a relationship between R' and R must be found. This is accomplished by using the geometrical construction shown in Figure 12.

A perpendicular is drawn from the point Q on the spherical wave to the line connecting P_1 and P_0, in Figure 12. The distance from the source to the observation point is $x_1 + x_2$ and the distance from the source to the plane of the zone is the radius of the incident spherical wave, R'. The distance from P_1 to P_0 can be written

$$x_1 + x_2 = R' + Z$$

The distance from the observation point to the zone is

$$R^2 = x_2^2 + y^2 = [(R' + Z) - x_1]^2 + (R' \sin\theta)^2$$
$$= R'2 + (R' + Z)^2 - 2R'(R' + Z)\cos\theta$$

The derivative of this expression yields

$$R\,\mathrm{d}R = R'(R' + Z)\sin\theta\,\mathrm{d}\theta \qquad [32]$$

Substituting [32] into [31] gives

$$\mathrm{d}s = \frac{R'}{R' + Z} R\mathrm{d}R\mathrm{d}\phi \qquad [33]$$

The integration over ϕ is accomplished by rotating the surface element about the P_1P_0 axis. After integrating over ϕ between the limits of 0 and 2π, we obtain

$$\tilde{E}_j(\mathrm{P}_0) = \frac{2\pi A}{R' + Z} \mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}'} \int_{Z+(j-1)\frac{\lambda}{2}}^{Z+\frac{j\lambda}{2}} \mathrm{e}^{-\mathrm{ik}\cdot\mathrm{R}} C(\varphi)\mathrm{d}R \quad [34]$$

We assume that $R',\ Z \gg \lambda$ so that the obliquity factor is a constant over a single zone, i.e., $C(\varphi) = C_j$. The obliquity factor is not really a constant as we earlier assumed but rather a very slowly varying parameter of j, changing by less than two parts in 10^4 when 'j' changes by 500 (for $Z = 1$ meter and $\lambda = 500$ nm). In fact it only changes by 2 parts in 10^{-7} across one zone. We are safe in applying the assumption that the obliquity factor is a constant over a zone and this allows the integral in [34] to be calculated:

$$\tilde{E}_j(\mathrm{P}_0) = \frac{2\pi \mathrm{i} C_j A}{k(R' + Z)} \mathrm{e}^{-\mathrm{i}k(R'+Z)} \mathrm{e}^{-\mathrm{i}kj\frac{\lambda}{2}} \left[1 - \mathrm{e}^{-\mathrm{i}k\frac{\lambda}{2}}\right] \quad [35]$$

Using the identity $k\lambda = 2\pi$ and the definition for the distance between the source and observation point [24] modified for this geometry, $D = R' + Z$ [35] can be simplified to

$$\tilde{E}_j(\mathrm{P}_0) = 2\mathrm{i}\lambda(-1)^j \frac{C_j A}{D} \mathrm{e}^{-\mathrm{i}kD} \qquad [36]$$

The physical reasons for the behavior predicted by [36] are quite easy to understand. The distance from P_0 to a zone changes by only $\lambda/2$ as we move from zone to zone and the area of a zone is almost a constant, independent of the zone number; thus, the amplitudes of the Huygens' wavelets from each zone should be approximately equal. The alternation in sign, from zone to zone, is due to the phase change of the light wave from adjacent zones because the propagation paths for adjacent zones differ by $\lambda/2$.

To find the total field strength at P_0 due to N zones, the collection of Huygens' wavelets is added:

$$E(\mathrm{P}_0) = \sum_{j=1}^{N} \tilde{E}_j(\mathrm{P}_0) = \frac{2\mathrm{i}\lambda A}{D} \mathrm{e}^{-\mathrm{i}kD} \sum_{j=1}^{N} (-1)^j C_j \qquad [37]$$

To evaluate the sum, the elements of the sum are regrouped and rewritten as

$$-\sum_{j=1}^{N} (-1)^j C_j = \frac{C_1}{2} + \left(\frac{C_1}{2} - C_2 + \frac{C_3}{2}\right) + \left(\frac{C_3}{2} - C_4 + \frac{C_5}{2}\right) + \ldots$$

Because the C's are very slowly varying functions of j, even out to 500 zones, we are justified in setting the quantities in parentheses equal to zero. With this approximation, the summation can be set equal to one of two values, depending upon whether there is an even or odd number of terms in the summation

$$-\sum_{j=1}^{N} (-1)^j C_j = \begin{Bmatrix} \frac{1}{2}(C_1 + C_N) & N \text{ odd} \\ \frac{1}{2}(C_1 - C_N) & N \text{ even} \end{Bmatrix}$$

For very large N, the obliquity factor approaches zero, $C_N \to 0$, as was demonstrated in Figure 4. Thus, the theory has led us to the conclusion that the total field produced by an unobstructed wave is equal to one half the contribution from the first Fresnel zone, i.e., $E = E_1/2$. Stating this result in a slightly different way, we obtain a surprising result – the contribution from the first Fresnel zone is twice the amplitude of the unobstructed wave!

The zone construction can be used to analyze the effect of the obstruction of all or part of a zone. For example, by constructing a circular aperture with a diameter equal to the diameter of the first Fresnel zone, we have just demonstrated that it is possible to produce an intensity at the point P_0 equal to four times the intensity that would be observed if no aperture were present. To analyze the effects of a smaller aperture, we subdivide a half-period zone into a number of subzones such that there is a constant phase difference between each subzone. The individual vectors form a curve know as the vibrational curve. Figure 13 shows a vibrational curve produced by the addition of waves from nine subzones of the first Fresnel zone. The vibrational curve in Figure 13 is an arc with the appearance of a half-circle.

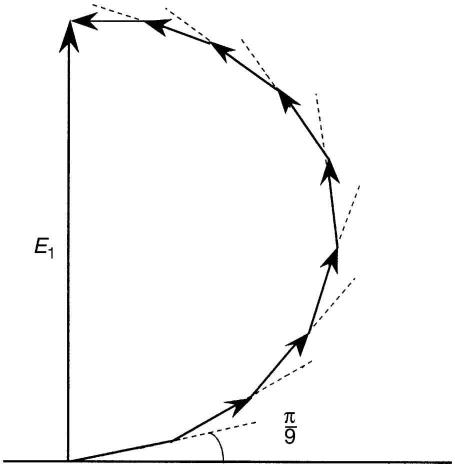

Figure 13 Vector addition of waves from nine subzones constructed in the first Fresnel zone. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the radius of curvature of the arc were calculated, we would discover that it is a constant, except for the contribution of the obliquity factor,

$$\rho = \frac{\lambda R'}{R' + Z} C(\varphi)$$

Because the obliquity factor for a single zone is a constant to 2 parts in 10^{-7}, the radius of curvature of the vibration curve can be considered a constant over a single zone. If we let the number of subzones approach infinity, the vibrational curve becomes a semicircle whose chord is equal to the wavelet produced by zone one, i.e., E_1. If we subdivide additional zones, and add the subzone contributions, we create other half-circles whose radii decrease at the same rate as the obliquity factor. The vibrational curve for the total wave is a spiral, constructed of semicircles which converges to a point halfway between the first half-circle, see Figure 14. The length of the vector from the start to the end of the spiral is $E = E_1/2$, as we derived above. When this same construction technique is applied to a rectangular aperture, the vibrational curve generated is the Cornu spiral.

Circular Aperture

The vector addition technique, described in Figure 13, can be used to evaluate Fresnel diffraction, at a point P_0, from a circular aperture and yields the intensity distribution along the axis of symmetry of the circular aperture. The zone concept will also allow a qualitative description of the light distribution normal to this axis.

To develop a quantitative estimate of the intensity at an observation point on the axis of symmetry of the circular aperture, we construct a spiral, Figure 14, to represent the Fresnel zones of a spherical wave incident on the aperture. The point B on the spiral shown in Figure 14 corresponds to the portion of the spherical wave unobstructed by the screen. The length of the chord, AB, represents the amplitude of the light wave at the observation point P_0. As the diameter of the aperture increases, B moves along the spiral, in a counterclockwise fashion, away from A. The first maximum occurs when B reaches the point labeled A_1 in Figure 14; the aperture has then uncovered the first Fresnel zone. At this point, the amplitude is twice what it would be with no obstruction. Four times the intensity!

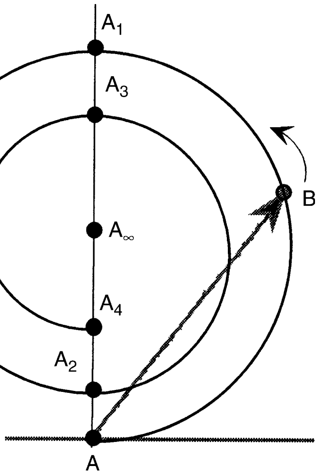

Figure 14 Vibration curve for determining the Fresnel diffraction from a circular aperture. The change of the diameter of the half-circles making up the spiral has been exaggerated for easy visualization. The actual changes are one part in 10^{-5}. The position of point B is determined by the aperture size. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

If the aperture's diameter continues to increase, B reaches the point labeled A_2 in Figure 14 and the amplitude is very nearly zero; two zones are now exposed in the aperture. Further maxima occur when an odd number of zones are in the aperture and further minima when an even number of zones are exposed. Figure 15 shows an aperture containing four exposed Fresnel zones. The amplitude at the observation point would correspond to the chord drawn from A to A_4 in Figure 14.

The aperture diameter can be fixed and the observation point P_0 can move along, or perpendicular to, the axis of symmetry of the circular aperture. As P_0 is moved away from the aperture, along the symmetry axis, i.e., as Z increases, the radius of the Fresnel zones increase without limit. For small values of n, the radius of the nth zone can be approximated by

$$r_n \approx \sqrt{nZ\lambda} \qquad [38]$$

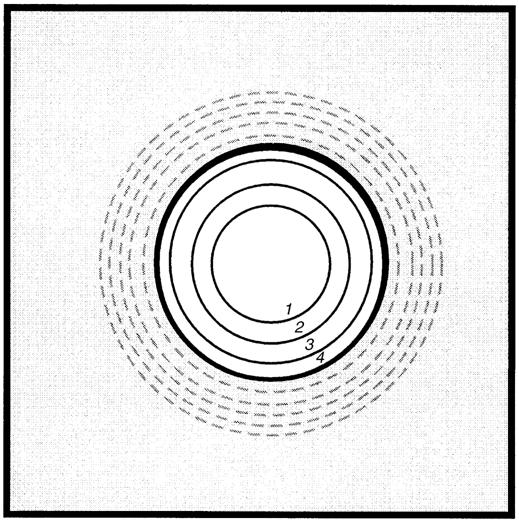

Figure 15 Aperture with four Fresnel zones exposed. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

At Z_{max}, the light intensity is a maximum, given by the chord length from A to A_1 in Figure 14. If $\lambda = 500$ nm and $a = 0.5$ mm then this maximum occurs when $Z = 0.5$ m

$$Z_{\mathrm{max}} = \frac{a^2}{\lambda} \qquad [39]$$

If we start at Z_{max} and move toward the aperture, along the axis, as Z decreases in value, a point will be reached when the intensity on the axis becomes a minimum. The value of Z where the first minimum in intensity is observed is equal to

$$Z_{\mathrm{min}} = \frac{a^2}{2\lambda}$$

In Figure 14 the chord would extend from A to A_2.

As the observation point, P_0, is moved along the axis toward the aperture, and Z assumes values less than Z_{min}, the point B in Figure 14 spirals inward, toward the center of the spiral and the intensity cycles through maximum and minimum values. The cycling of the intensity, as the observation point moves toward the aperture, will not continue indefinitely. At some point, the field on the axis will approach the field observed without the aperture because the distance between Z_{max} and Z_{min} shrinks to the wavelength of light making the intensity variation unobservable.

For values of Z that exceed Z_{max} of [39], the aperture radius, a, will be smaller than the radius of the first zone, and Fraunhofer diffraction will be observed because the aperture contains only one zone and the phase in the aperture is a constant.

Opaque Screen

If the screen containing a circular aperture, of radius a, is replaced by an opaque disk of radius a, the intensity distribution on the symmetry axis, behind the disk, is found to be equal to the value that would be observed with no disk present. This prediction was first derived by Poisson, to demonstrate that wave theory was incorrect; however, experimental observation supported the prediction and verified the theory.

We construct a spiral, shown in Figure 16, similar to the one for the circular aperture in Figure 14. The point B on the spiral represents the edge of the disk. The shaded portion of the spiral from A to B does not contribute because that portion of the wave is covered by the disk and the zones, associated with that portion of the wave, cannot be seen from the observation point. The amplitude at P_0 is the length of the chord from B to A_∞ shown in Figure 16. If the observation point moves toward the disk, then B moves along the spiral toward A_∞. There is always intensity on the axis for this configuration, though it slowly decreases until it reaches zero when the observation point reaches the disk; this corresponds to point B reaching point A_∞ on the spiral. Physically, zero intensity occurs when the disk blocks the entire light wave. There are no maxima or minima observed as the disk diameter, a, increases or as the observation distance changes. If the observation point is moved perpendicular to the symmetry axis, a set of concentric bright rings are observed. The origin of these bright rings can be explained using Fresnel zones, in a manner similar to the one used to explain the bright rings observed in a Fresnel diffraction pattern from a circular aperture.

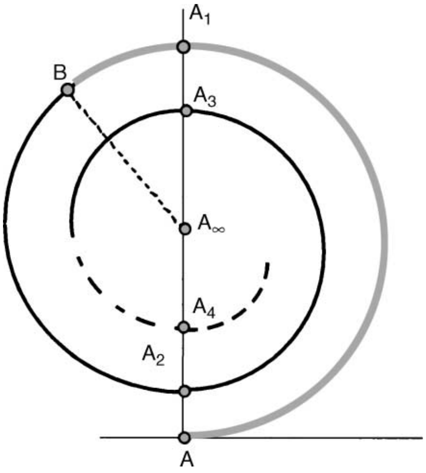

Figure 16 Vibration curve for opaque disk. The shaded region makes no contribution because it is associated with the portion of the wave obstructed by the opaque object. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

Zone Plate

In the construction of Fresnel zones, each zone was assumed to produce a Huygens wavelet, out of phase with the wavelets produced by its nearest neighbors. If every other zone were blocked, then there would be no negative contributions to [37]. The intensity on-axis would be equal to the square of the sum of the amplitudes produced by the in-phase zones – exceeding the intensity of the incident wave. An optical component made by the obstruction of either the odd or the even zones could therefore be used to concentrate the energy in a light wave.

The boundaries of the opaque zones, used to block out-of-phase wavelets, are seen from [38] to increase as the square root of the integers. An array of opaque rings, constructed according to this prescription is called a zone plate, see Figure 17.

A zone plate will perform like a lens with focal length

$$f = \pm \frac{r_1^2}{\lambda} \qquad [40]$$

The zone plate, shown in Figure 17, will act as both a positive and negative lens. What we originally called the source can now be labeled the object point, O, and what we called the observation point can now be labeled the image point, I. The light passing through the zone plate is diffracted into two paths, labeled C and D in Figure 17. The light waves, labeled C, converge to a real image point I. For these waves the zone plate performs the role of a positive lens with a focal length given by the positive value of [40]. The light waves, labeled D in Figure 17, appear to originate from the virtual image point labeled $\mathbb{I}$. For these waves the zone plate performs the role of a negative lens with a focal length given by the negative value of [40].

The zone plate will not have a single focus, as is the case for a refractive optical element, but rather will have multiple foci. As we move toward the zone plate, from the first focus, given by [40], the effective Fresnel zones will decrease in diameter. The zone plate will no longer obstruct out-of-phase Fresnel zones and the light intensity on the axis will decrease. However, additional maxima, of the on-axis intensity, will be observed at values of Z for which the first zone plate opening contains an odd number of zones. These positions can also be labeled as foci of the zone plate; however, the intensity at each of these foci will be less than the intensity at the primary focus.

Lord Rayleigh suggested that an improvement of the zone plate design would result if, instead of blocking every other zone, we shifted the phase of alternate zones by 180°. The resulting zone plate, called a phase-reversal zone plate, would, more efficiently, utilize the incident light. R.W. Wood was the first to make such a zone plate. Holography provides an optical method of constructing the phase-reversal zone plates in the visible region of the spectrum. Semiconductor lithography has been used to produce zone plates in the x-ray region of the spectrum.

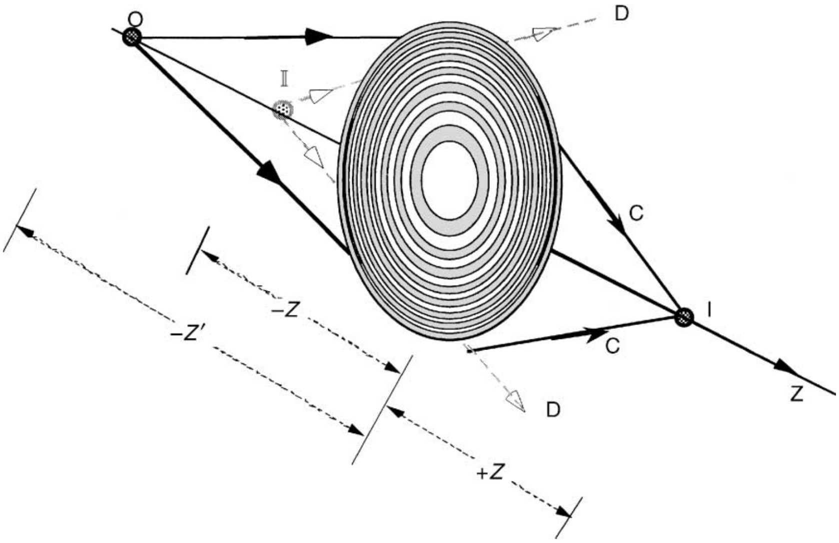

Figure 17 A zone plate acts as if it were both a positive and a negative lens. Light from the object, O, is diffracted into waves traveling in both the D and the C directions. The light traveling in the C direction produces a real image of O at I. The light traveling in the D direction appears to originate from a virtual image at I. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

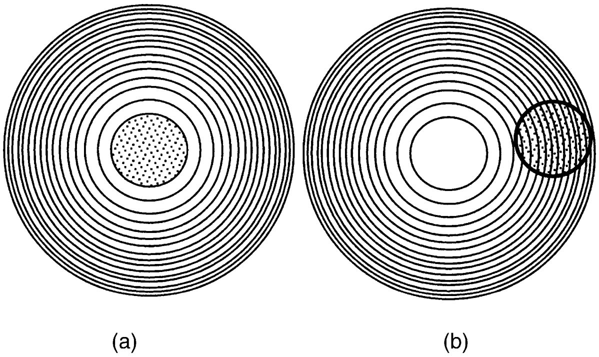
(a) (b)

Figure 18 (a) A set of Fresnel zones has been constructed about the optical path taken by some light ray. If the optical path of the ray is varied over the cross-hatched area shown in the figure, then the optical path length does not change. This cross-hatched area is equal to the first Fresnel zone and is described as the neighborhood of the light ray. (b) The neighborhood defined in (a) is moved so that it surrounds an incorrect optical path for a light ray. We see that this region of space would contribute no wave amplitude at the observation point because of the destructive interference between the large number of partial zones contained in the neighborhood. Reprinted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley & Sons.

The resolving power of a zone plate is a function of the number of zones contained in the plate. When the number of zones exceeds 200, the zone plate's resolution approaches that of a refractive lens.

Fermat's Principle

The Fresnel zone construction provides physical insight into the interpretation of Fermat's principle which states that, if the optical path length of a light ray is varied in a neighborhood about the true path, there is no change in path length. By constructing a set of Fresnel zones about the optical path of a light ray, we can discover the physical meaning of a neighborhood.

The rules for constructing a Fresnel zone require that all rays passing through a given Fresnel zone have the same optical path length. The true path will pass through the center of the first Fresnel zone constructed on the actual wavefront. A neighborhood must be the area of the first Fresnel zone, for it is over this area that the optical path length does not change. Figure 18a shows the neighborhood, about the true optical path, as a cross-hatched region equal to the first Fresnel zone.

Light waves do travel over wrong paths but we do not observe them because the phase differences for those waves that travel over the 'wrong' paths are such that they destructively interfere. By moving the neighborhood, defined in the previous paragraph as the first Fresnel zone, to a region displaced from the true optical path, we can use the zone construction to see that this statement is correct. In Figure 18b the neighborhood is constructed about a ray that is improper according to Fermat's principle. We see that this region of space would contribute no energy at the observation point because of destructive interference between the large number of partial zones contained in the neighborhood.

This leads us to another interpretation of diffraction. If an obstruction is placed in the wavefront so that we block some of the normally interfering wavelets, we see light not predicted by Fermat's principle, i.e., we see diffraction.

Further Reading

Backer BB and Copson EJ (1969) *The Mathematical Theory of Huygens' Principle*. London: Oxford University Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Haus HA (1984) *Waves and Fields in Optoelectronics*. Englewood Cliffs, NJ: Prentice-Hall.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Rubinowicz A (1984) *Tahe Miyamoto–Wolf diffraction wave*. In: Wolf E (ed.) Progress in Optics, vol. IV.

Sayanagi I (1967) Pinhole Imagery. *Journal of the Optical Society of America*. 57: 1091–1099.

Sears FW (1949) *Optics*. Reading, MA: Addison-Wesley.

Sommerfeld A (1964) *Optics*. New York: Academic Press.

Stegliani DJ, Mittra R and Semonin RG (1967) Resolving power of a zone plate. *Journal of the Optical Society of America* 57: 610–613.

Stone JM (1963) *Radiation and Optics*. New York: McGraw-Hill.

DIFFRACTIVE SYSTEMS

Contents

Aberration Correction with Diffractive Elements
Applications of Diffractive and Micro-Optics in Lithography
Design and Fabrication of Diffractive Optical Elements
Diffractive Laser Resonators
Diffractives in Animals
Microstructure Fibers
Omnidirectional Surfaces and Fibers
Wave Optical Modeling and Design

Aberration Correction with Diffractive Elements

N Lindlein, University of Erlangen-Nürnberg, Erlangen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Diffractive optical elements (DOEs) can be used for the correction of aberrations in optical systems because of two fundamental properties:

1. The phase function Φ of a DOE can be chosen quite arbitrarily so that for a given wavelength λ the DOE can act like an aspheric optical element.
2. The dispersion of a DOE due to the local grating equation is very high and its sign is opposite to that of the dispersion of a refractive optical element with normal dispersion of the refractive index.

The first property allows the correction of monochromatic aberrations in an optical system similar to the case of using an aspheric surface in a refractive optical system. The second property allows the construction of hybrid achromatic optical elements that consist of a refractive lens and a diffractive lens whose optical powers have the same sign (e.g., two positive lenses for a positive achromatic element or two negative lenses for a negative achromatic element).

This article will first describe two methods for the simulation of DOEs by ray tracing because ray tracing is the basic simulation tool for aberration correction. Afterwards, the properties of DOEs for the correction of monochromatic and chromatic aberrations and some applications will be discussed. The name DOE will always mean either binary, or multiple-step or blazed surface relief diffractive optical elements which change only the phase of the incident wave and not the amplitude. All examples will be given for transmission elements but in principle reflective elements can also be described with quite similar formulas.

Ray Tracing Simulation of DOEs

First of all it seems straightforward to use a diffractive analysis to simulate a diffractive optical element. However, the global diffractive analysis of a general DOE with perhaps 1000 or more grating structures and locally changing grating period and especially the diffractive analysis of a complete optical system containing DOEs and other elements is so complex that it can be done only for very simple DOEs and systems. Therefore, the diffractive analysis of a DOE is mostly restricted to a local analysis using some periods (e.g., 10–50 periods) to calculate, e.g., the different local diffraction orders, their diffraction efficiencies and possibly local phase shifts. For the global analysis another simulation method has to be used.

The most important simulation method for the correction of aberrations in traditional optical systems is ray tracing. Therefore, ray tracing methods are also used for DOEs. These methods can calculate the path of a ray for a given diffraction order, but they cannot calculate the diffraction efficiency for this ray. For the calculation of the diffraction efficiency local diffractive analysis methods (scalar or rigorous methods depending on the local grating frequency and the angle of the incident ray) have to be used in addition. In particular, for applications with a broad wavelength spectrum (e.g., an achromatic lens) or a broad spectrum of angles of the incident rays, the diffraction efficiency has to be analyzed carefully to obtain the energy distribution in the different diffraction orders. But this is beyond the scope of this article about aberration correction and we assume that we know the local diffraction efficiency for each ray.

Sweatt Model for the Ray Tracing Simulation of DOEs

An aspheric refractive surface with height h as a function of the lateral coordinates x and y and a refractive index n below the surface and air (refractive index 1) above the surface can be approximately replaced by a DOE with the phase function (see Figure 1)

$$\Phi(x, y) = 2\pi \frac{(n-1)h(x, y)}{\lambda} \qquad [1]$$

Here, λ is the wavelength of the light for which the DOE is designed and the phase function is a continuous representation of the structure of the DOE. To encode the phase function of the DOE the wrapped phase function $\Phi \bmod 2\pi$ has to be considered. The encoding itself can be done by building a blazed, a multiple-step or a binary structure (see Figure 1), whereby the first-order diffraction efficiency decreases from the blazed to the binary structure (at maximum 100% diffraction efficiency for the blazed structure and about 40% for the binary structure). The replacement of an aspheric refracting surface by a DOE using eqn [1] is only valid in the paraxial approximation since the phase function of the DOE is encoded in a plane whereas the refracting surface of the aspheric is typically curved. But, the agreement between the aspheric surface and the DOE with a phase function calculated with the help of eqn [1] increases with increasing refractive index n since then the curvature of the aspheric is quite small and the refraction also nearly takes place in a plane. Of course, the refractive index of real materials cannot exceed a value of more than about 3.5 (for silicon used in the near infrared). But, in a numerical simulation the refractive index can be arbitrarily high, e.g., several thousand. This is the reason why the so called Sweatt model for tracing rays through DOEs uses a weakly curved aspheric surface with height h_{sweatt} and a very high refractive index n_{sweatt} to simulate a diffractive phase function Φ:

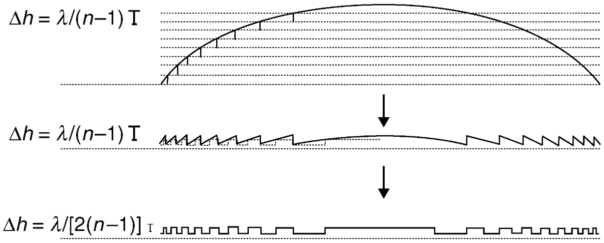

Figure 1 Transition from an aspheric refracting surface to a blazed diffractive optical element and finally to a binary diffractive optical element.

$$h_{\mathrm{sweatt}}(x, y) = \frac{\lambda}{n_{\mathrm{sweatt}} - 1} \frac{\Phi(x, y)}{2\pi} \qquad [2]$$

To take into account the wavelength dispersion of the DOE the refractive index n_{sweatt} in the Sweatt model has to be proportional to the wavelength λ

$$n_{\mathrm{sweatt}}(\lambda) = n_0 \frac{\lambda}{\lambda_0} \qquad [3]$$

where n_0 is the refractive index at the wavelength λ_0.

The validity of the Sweatt model increases with increasing refractive index and becomes nearly exact for $n_{\mathrm{sweatt}} \to \infty$ (in practice values of several thousand for the refractive index are normally sufficient and a rule of thumb is to take the numerical value of the utilized wavelength specified either in nanometers for infrared light or angstroms for visible light). The Sweatt model was introduced as a means to simulate DOEs in ray tracing programs unable to implement DOEs directly but capable of simulating aspheric refractive surfaces. However, a disadvantage of this model is the numerical error generated by the finite value of the refractive index.

Modern ray tracing programs therefore use a more appropriate model based on the local grating equation.

Local Grating Model for the Ray Tracing Simulation of DOEs

The basic equation for the simulation of DOEs by the local grating model is that the phase Φ_{out} of a diffracted ray in the plane of the DOE can be calculated from the phase Φ_{in} of the incident ray by adding the phase function Φ of the DOE multiplied by the diffraction order m:

$$\Phi_{\mathrm{out}}(x, y, z_0) = \Phi_{\mathrm{in}}(x, y, z_0) + m\Phi(x, y) \qquad [4]$$

Here, z_0 is the z-coordinate of the DOE plane which is assumed to be perpendicular to the z-axis (see Figure 2). The diffraction order m will be only one for an ideal blazed DOE but it can have an arbitrary integer value for binary DOEs. In practice, the optical designer will calculate the system for one diffraction order and analyze the system for other diffraction orders afterwards if necessary.

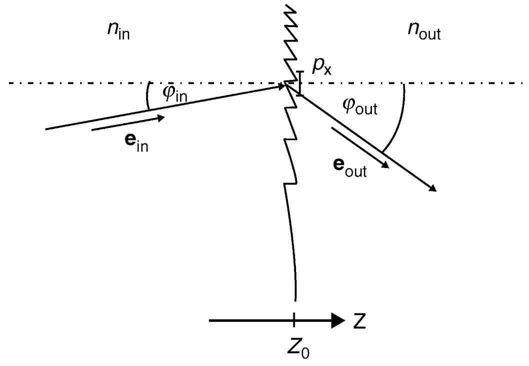

Figure 2 Schematic view in two-dimensions of the path of a ray which is diffracted at a DOE.

For the phases Φ_{in} of the incident and Φ_{out} of the diffracted ray the eikonal equation of geometrical optics is valid,

$$|\nabla\Phi_j|^2 = k_j^2 = \left(\frac{2\pi n_j}{\lambda}\right)^2 \qquad [5]$$

where the index j can have the two values $j \in \{\text{in}, \text{out}\}$. n_{in} or n_{out} are the refractive indices of the materials in which the incident or diffracted rays propagate. λ is the wavelength in vacuum. This form of the eikonal equation is obtained from the more common form $|\nabla L_j|^2 = n_j^2$, which is derived in any textbook of optics, by multiplying it with $(2\pi/\lambda)^2$ and using the definition of the optical phase $\Phi_j = 2\pi L_j/\lambda$. Here, L_j is the optical path length which is often called the eikonal. The curves of constant phase $\Phi_j(x, y, z) = \text{constant}$ are the wavefronts and the gradient $\nabla\Phi_j(x, y, z)$ points in the direction of the local ray which is perpendicular to the wavefront at the point (x, y, z) and describes its direction of propagation. Furthermore, the unit vectors $\mathbf{e}_{in}$ and $\mathbf{e}_{out}$ along the ray direction of the incident and diffracted ray, respectively, can be calculated by normalizing the gradient:

$$\mathbf{e}_j = \frac{\nabla\Phi_j}{|\nabla\Phi_j|} = \frac{\lambda}{2\pi n_j}\nabla\Phi_j\,; \qquad j \in (\text{in}, \text{out}) \qquad [6]$$

Since $\mathbf{e}_j$ is a unit vector, i.e., $e_{x,j}^2 + e_{y,j}^2 + e_{z,j}^2 = 1$, it can be seen from eqn [6] (and also from the original eqn [5]) that the z-component $e_{z,j}$ of the ray direction vector can be calculated (apart from its sign) if the two partial derivatives of the phase with respect to x and y are known. This means that it is sufficient to know the phase in a plane as stated in eqn [4].

By calculating the partial derivatives of eqn [4] with respect to x and y the equations for the ray tracing at a DOE are derived:

$$\begin{aligned}
\frac{\partial\Phi_{out}(x,y)}{\partial x} &= \frac{\partial\Phi_{in}(x,y)}{\partial x} + m\frac{\partial\Phi(x,y)}{\partial x}\\
&\Rightarrow n_{out}e_{x,out}(x,y)\\
&= n_{in}e_{x,in}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial x}\\
\frac{\partial\Phi_{out}(x,y)}{\partial y} &= \frac{\partial\Phi_{in}(x,y)}{\partial y} + m\frac{\partial\Phi(x,y)}{\partial y}\\
&\Rightarrow n_{out}e_{y,out}(x,y)\\
&= n_{in}e_{y,in}(x,y) + m\frac{\lambda}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [7]$$

The local grating frequencies ν_x and ν_y in x and y direction are calculated from the phase function Φ by:

$$\begin{aligned}
\nu_x(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial x}\\
\nu_y(x,y) &= \frac{1}{2\pi}\frac{\partial\Phi(x,y)}{\partial y}
\end{aligned} \qquad [8]$$

Therefore, given $\mathbf{e}_{in}$ and eqn [6], the ray direction vector $\mathbf{e}_{out}$ of the diffracted ray is

$$n_{out}\mathbf{e}_{out} = n_{out}\begin{pmatrix} e_{x,out}\\ e_{y,out}\\ e_{z,out}\end{pmatrix} = \begin{pmatrix} n_{in}e_{x,in} + m\lambda\nu_x\\ n_{in}e_{y,in} + m\lambda\nu_y\\ \sqrt{n_{out}^2 - (n_{in}e_{x,in} + m\lambda\nu_x)^2 - (n_{in}e_{y,in} + m\lambda\nu_y)^2}\end{pmatrix} \qquad [9]$$

The dependence of the lateral coordinates x and y is left out in eqn [9] to simplify notation. It can be easily checked that this equation fulfills the required condition for the unit vector: $e_{x,out}^2 + e_{y,out}^2 + e_{z,out}^2 = 1$.

As stated above, for a blazed DOE fabricated to the exact depth and illuminated by the design wavelength only the first diffraction order $m = 1$ is generated. But in practice other diffraction orders m are generated due to fabrication errors or illuminating with other wavelengths. In particular, by using binary DOEs there are inherently different diffraction orders. Therefore, it may be necessary to analyze the optical system for several values m one after the other. Each value m defines one wavefront generated by the DOE and the wavefront can be calculated by evaluating the optical path lengths along the rays at different points. In practice, the optical system has to be analyzed for those values m with sufficiently high diffraction efficiency. Then, it can be checked whether undesired diffraction orders of the DOE are blocked in the optical system by aperture stops or whether they will disturb the output of the optical system. The local diffraction efficiency of each ray for a value m can be calculated to a good approximation by using a diffractive analysis of the corresponding local grating, i.e., that of an infinitely extended grating which would have the same grating period and the same structure shape (e.g., depth) as the DOE at the specific point.

If the y component of the incident ray and the local grating frequency ν_y are both zero the ray components in the x direction can be expressed via the sine of the angle of incidence φ_{in} or the angle of the diffracted ray φ_{out} (see Figure 2). In this case the first row of eqn [9] can be written as

$$n_{\text{out}} \sin \varphi_{\text{out}} = n_{\text{in}} \sin \varphi_{\text{in}} + m \frac{\lambda}{p_x} \qquad [10]$$

where the local grating period $p_x = 1/\nu_x$ has been introduced. This is the well-known grating equation.

In this article eqns [4]–[10] assume that the DOE is on a planar substrate and that x and y are the lateral coordinates on the substrate. However, the equations can also be written in a vector formulation which is valid for DOEs on curved substrates (see Further Reading).

So, we have now a powerful simulation method to design and analyze in the following several optical systems containing DOEs to correct monochromatic or chromatic aberrations.

Correction of Monochromatic Aberrations

For practical applications it is important to have an efficient algorithm to calculate the phase function of a DOE to correct the aberrations in a given optical system. Several different cases have to be distinguished. An easy case is the design of a DOE to produce a desired output wavefront Φ_{out} from a known input Φ_{in}. In this case, eqn [4] can be used to calculate the phase function Φ of the DOE directly. Two examples are given in the following. Afterwards, some aspects of a more general numerical optimization procedure are described.

Aberration Correction for Interferometrically Recorded Holograms

Although the following example of the aberration correction of interferometrically recorded holograms by using DOEs seems to be quite special it shows exemplarily several principal steps which are necessary to design a DOE for correcting monochromatic aberrations.

In this article the name DOE is used for diffractive elements with a microstructured surface (binary, multistep or blazed). Such elements have the property that their diffraction efficiency depends only weakly on the wavelength or the angle of incidence of the illuminating light. Another interesting class of diffractive elements are holograms in volume media, which are recorded interferometrically using two coherent recording waves. The refractive index in a volume hologram is modulated periodically and the thickness has to amount to several periods of the interference pattern along the z-direction in order to have a so-called thick hologram with a high diffraction efficiency. If there are only a few periods along the z-direction the quite small modulation of the refractive index (typically about 0.02) will result in a low diffraction efficiency. Therefore, if the angle between the two recording waves is small the resulting interference pattern will have large periods and only a few periods are recorded in the photosensitive layer along the z-direction. But this means that the holographic optical element (HOE) can only have a small diffraction efficiency. From this fact, it is clear that also by using the recorded HOE the deflection angle between the incident and the diffracted wave must not be smaller than a certain value because otherwise the thickness has to be very big or the diffraction efficiency decreases. Therefore, such elements with high diffraction efficiency, which will be named in the following holographic optical elements (HOE), are always off-axis elements because for on-axis elements all rays in the neighborhood of the axis are only deflected by a small angle and then the diffraction efficiency would be low. For the correct wavelength and angle of incidence, so that the Bragg condition of holography is fulfilled, HOEs, like blazed DOEs, exhibit only one dominant diffraction order $m = m_{\text{dom}}$ (normally $m_{\text{dom}} = 1$). But, if the angle of incidence or the wavelength change, the diffraction efficiency in the order m_{dom} decreases rapidly and light passes through the HOE without being deflected, i.e., most of the light is then in the order $m = 0$. For this reason, a HOE can be used for example for a head-up display to fade in some information under the Bragg angle without hindering the normal view onto the environment.

Traditional materials for HOEs like dichromated gelatin (DCG) are sensitive only in the blue–green spectral region (e.g., for light of an argon ion laser with $\lambda = 488$ nm). However, once recorded it will be illuminated in many applications (e.g., optical interconnects) by a wavelength in the red–infrared spectral region. Unfortunately, the wavelength mismatch between recording and reconstruction illumination introduces aberrations.

If the two desired reconstruction waves at the wavelength λ_2 have the phases $\Phi_{r,\lambda 2}$ and $\Phi_{o,\lambda 2}$ in the surface plane of the HOE (the subscripts r and o refer to interferometry reference and object, respectively), the phase function Φ of the HOE which has to be recorded is

$$\Phi(x,y) = \Phi_{o,\lambda 2}(x,y) - \Phi_{r,\lambda 2}(x,y) \qquad [11]$$

If our discussion here is restricted to holographic lenses, the two reconstruction waves will be spherical (or plane) waves (see Figure 3b) with the center of curvature at the points $\mathbf{r}_{r,\lambda 2}$ and $\mathbf{r}_{o,\lambda 2}$ and eqn [11] is:

$$\begin{aligned}\Phi(x,\ y) = {} & \frac{2\pi\sigma_{o,\lambda 2}}{\lambda_2}\sqrt{(x-x_{o,\lambda 2})^2+(y-y_{o,\lambda 2})^2+z_{o,\lambda 2}^2} \\ & - \frac{2\pi\sigma_{r,\lambda 2}}{\lambda_2}\sqrt{(x-x_{r,\lambda 2})^2+(y-y_{r,\lambda 2})^2+z_{r,\lambda 2}^2}\end{aligned} \qquad [12]$$

The parameters $\sigma_{o,\lambda 2}$ and $\sigma_{r,\lambda 2}$ are either $+1$ for a divergent spherical wave or -1 for a convergent spherical wave. The surface of the HOE itself is in the plane at $z = 0$.

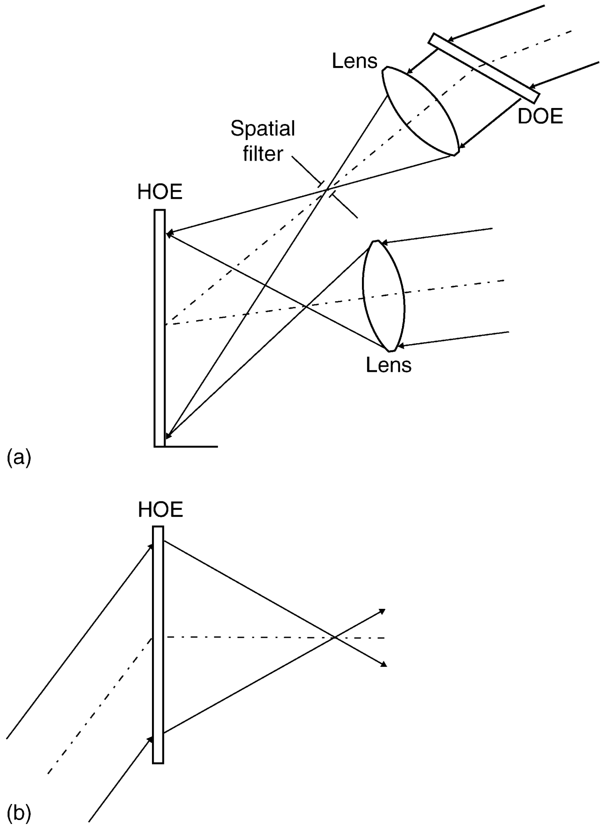

Figure 3 Schematic view of the (a) recording and (b) reconstruction setup for a holographic lens. The recording is made at wavelength λ_1 and the reconstruction at wavelength λ_2.

The task is now to find two recording waves which construct nearly the same three-dimensional interference pattern in the hologram at the recording wavelength λ_1. This means that eqn [12] has to be fulfilled exactly for the difference of the phases of the two recording waves in the plane $z = 0$ and, additionally, the Bragg condition has to be fulfilled at least approximately. So, the next step is to calculate one of the recording waves with phase $\Phi_{r,\lambda 1}$ so that the Bragg condition is fulfilled. This can be done approximately with a third-order theory resulting in a spherical wave or nearly exactly with an iterative numerical procedure resulting generally in an aspheric recording wave. If $\Phi_{r,\lambda 1}$ is known, the other recording wave with phase $\Phi_{o,\lambda 1}$ is calculated by

$$\Phi_{o,\lambda 1}(x,y) = \Phi(x,y) + \Phi_{r,\lambda 1}(x,y) \qquad [13]$$

At least this recording wave will be aspheric. To generate such an aspheric wave it is traced by ray tracing to a plane where the two recording waves are spatially separated. There, either a DOE or a combination of a DOE and a lens are used to generate the aspheric wave. The calculations of the phase function of the DOE are done again by using an equation like [11] whereby one wave is the propagated aspheric wave (possibly propagated also through a lens) and the other is normally a plane off-axis wave with which the DOE will be illuminated in the real recording setup (see Figure 3a). Here, the off-axis angle is necessary to allow spatial filtering of disturbing diffraction orders of the DOE which will be normally fabricated as a binary DOE.

Experiments of our own showed (see the literature) that nearly diffraction-limited holographic lenses with a numerical aperture of NA = 0.4 can be fabricated by correcting the aberrations of about 28 wavelengths peak-to-valley (for the reconstruction at 633 nm wavelength) due to the wavelength mismatch between recording at 488 nm wavelength and reconstruction at 633 nm wavelength by using a DOE in the recording setup.

Testing of Aspheric Surfaces by Using a DOE as Aberration Compensating Element

An important application for DOEs is testing aspheric surfaces and nowadays more and more companies use DOEs to test aspheric surfaces which are used for example in the objectives of lithographic wafer steppers to produce highly integrated microelectronic circuits.

In this testing setup the DOE is illuminated by a wave with phase Φ_{in} to generate a wave with phase Φ_{out} that is incident perpendicular to the aspheric surface (see Figure 4). After reflection at the asphere

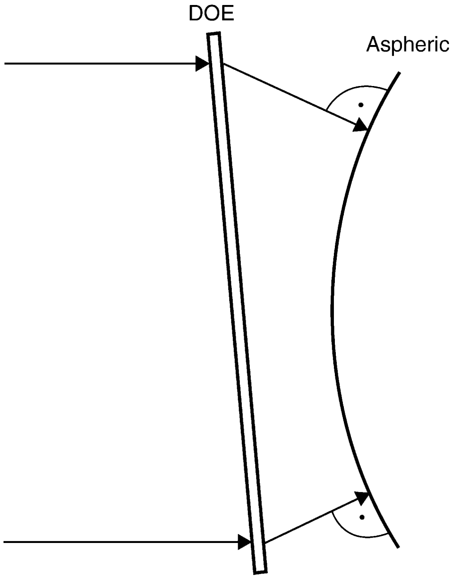

Figure 4 Scheme showing the basic elements for testing an aspheric surface.

the wave is retro-reflected and takes the same path if the asphere has no surface deviations and no misalignments. This means that the setup is completely symmetric for a reflection at the aspheric surface. Therefore, the optical path lengths have to be identical for all rays on the asphere. To calculate the phase function of the DOE the following steps are made:

- A grid of N rays starting perpendicularly on the aspheric surface and having all phase zero (due to the fact that the phase has to be identical it can be set to zero) is traced to the plane of the DOE. N depends on the shape and numerical aperture of the asphere and the grid will be in most cases two dimensional and regular or one dimensional (along one radial section of the asphere) and regular if the asphere and the system are completely rotationally symmetric. As a rule of thumb the number N of rays has to be more than twice the number of polynomial coefficients if the asphere is described by a polynomial.
- The phase of these rays is the function $-\Phi_{\text{out}}$ (the negative sign is necessary because the rays take the opposite direction) and is now known at some points (x_i, y_i) with $i \in \{1, 2, \ldots, N\}$.
- The phase Φ_{in} of the incident wave which is in practice either a plane wave or a spherical wave is calculated in the plane of the DOE. If the incident wave is generated by a lens having aberrations it is again calculated by ray tracing, otherwise it can be calculated analytically.
- Now, the phase function Φ of the DOE can in principle be calculated with the help of eqn [4].

However, the phase Φ_{out} (and possibly also Φ_{in}) is known only at some points (x_i, y_i), but it has to be known at all points of the DOE. Therefore in practice an analytical representation of the phase functions of the incident and output wave has to be calculated. This can be done by fitting a polynomial to the phase values using a least squares fit. In this case the coordinates, at which the phase values are known, can be quite arbitrary as long as the sampling is dense enough. For a rotationally symmetric system the polynomial should also be rotationally symmetric, i.e., a polynomial P_{rot} depending only on the radius coordinate r

$$r = \sqrt{x^2 + y^2} \tag{14}$$

$$P_{\text{rot}}(r) = \sum_{j=0}^{G} a_j r^j \quad \text{or} \quad P_{\text{rot}}(r) = \sum_{j=0}^{G/2} a_{2j} r^{2j} \tag{15}$$

where G is the degree of the polynomial, i.e., the highest power in r. In these two cases the number N of sampling points lying along one radial section of the rotationally symmetric optical system should be larger than $2G$ or G, respectively.

For general non-symmetric systems a Cartesian polynomial P_{cart} can be used, e.g.,

$$P_{\text{cart}}(x, y) = \sum_{j=0}^{G} \sum_{k=0}^{G-j} a_{jk} x^j y^k \tag{16}$$

Here, the number of coefficients of the polynomial is $(G+1)(G+2)/2$ and the number N of sampling points should again be larger than twice the number of coefficients, i.e., $N > (G+1)(G+2)$.

For both types of polynomials the coefficients a_j or a_{jk} have to be calculated so that the merit function M

$$M(\mathbf{a}) = \sum_{i=1}^{N} (P(x_i, y_i) - \Phi_j(x_i, y_i))^2; \quad j \in \{\text{in, out}\} \tag{17}$$

is minimized. The symbol P means either P_{rot} or P_{cart}. The vector $\mathbf{a}$ symbolizes the coefficients $\mathbf{a}_j$ or $\mathbf{a}_{jk}$. By differentiating M with respect to all coefficients and setting the partial derivatives to zero a system of linear equations is obtained which can be solved by standard procedures. As mentioned before, the number N of independent sampling points has to be at least double the size of the number of coefficients of the polynomial. Otherwise, the polynomial can 'swing' between the sampling points and the polynomial fits well to the desired phase function only at the sampling points but not at other points in between.

For very irregular aspherics with a fast change of the local curvature it might by necessary not to take a global polynomial fit but a spline fit, i.e., a set of adjacent local polynomials with several continuity conditions at the border area of each local polynomial. In this case the total number of coefficients used to describe the phase functions and Φ_{in} and Φ_{out} of the DOE will be increased compared to the description by a global polynomial. Then, the number N of rays for the calculation has also to be increased accordingly.

So, in this example the DOE acts as a null-compensator for the asphere and at first sight it is not a typical example for aberration correction. However, by regarding the complete system the DOE compensates the aberrations, i.e., deviations from a spherical wave, generated by the asphere which are primary and higher-order spherical aberrations if the asphere is rotationally symmetric and more general aberrations otherwise. This example shows how the phase function of a DOE can be designed straightforwardly for correcting aberrations if there is only one well-defined wavefront propagating through the optical system. This design method can also be used for other optical systems where the input and the output wavefront of the DOE are well defined.

Numerical Optimization of DOEs in Optical Systems

In a general optical system, where a DOE is not only used to compensate the aberrations for one specific incident wave but for different incident waves (optical system with a field), it is not possible to calculate the optimum phase function in such a straightforward way as given above. However, a DOE is in principle similar to an asphere in an optical system, so that the design rules for including aspherics in optical systems can be used in some cases as long as monochromatic aberrations are considered. Nevertheless, one should keep in mind that a DOE is, according to the Sweatt model, an asphere with high refractive index and small height, i.e., the deflection of the light rays occurs in a plane (or on a curved surface if the DOE is situated on a curved substrate). Therefore, the design rules for aspheric surfaces can be used only if the high refractive index of the Sweatt model is used in the design.

In the general case only a numerical optimization can be used. For this, it is very important to select an appropriate mathematical representation of the phase function of the DOE.

For rotationally symmetric optical systems also the phase function of the DOE should be rotationally symmetric. Therefore, in such a case it is useful to write the phase function Φ as a polynomial in the radius coordinate r as given in eqns [14] and [15]. Then, the $(G + 1)$ or $(G/2 + 1)$ coefficients a_j have to be optimized, whereby a_2 is proportional to the focal power of the DOE, whereas the other coefficients represent aspheric terms. Depending on the numerical aperture and on the application the degree G of the polynomial can be increased in order to achieve a better solution. Nevertheless, the degree in the optimization should be as small as necessary in order to avoid 'swinging effects' and numerical instabilities of the polynomial. To determine the smallest adequate degree an iterative method has to be used by starting with a quite small degree (e.g., 4). If the result of the optimization is not good enough the degree has to be increased until the result of the optimization is satisfying.

For non-rotationally symmetric optical systems a Cartesian polynomial like in eqn [16] can be taken and all the $(G + 2)(G + 1)/2$ coefficients a_{jk} have to be optimized. Since the number of free parameters is in this case larger than the number of free parameters in the rotationally symmetric case the convergence of the optimization may take a long time and the risk to break down in a local minimum of the multidimensional parameter space increases.

In principle, the calculation of a DOE used, e.g., for testing an aspheric surface can also be done by optimizing the coefficients of the polynomial representing the phase function of the DOE. Nevertheless, as a general rule it can be said that a direct calculation of the phase function is always more efficient than a numerical optimization, if the direct calculation is possible. In practice, it can be seen that the maximum degree of a polynomial, which is directly calculated by the method described in the last section, is always smaller than the degree which is necessary in a numerical optimization to achieve convergence of the optimization.

Of course, in such cases where several different types of aberrations (on-axis aberrations like spherical aberration, off-axis aberrations like astigmatism and coma or field aberrations like field curvature or distortion) have to be corrected in an optical system a direct calculation of the phase function is not possible so that only the numerical optimization can provide a solution.

Up to now only monochromatic aberrations have been discussed. But as stated at the beginning of this article the strong dispersion of a DOE having opposite sign to the dispersion of a refractive element offers additional applications of DOEs by correcting chromatic aberrations. The principle of doing this is described in the following.

Correction of Chromatic Aberrations

The focal power $1/f$ (f is the focal length of the lens) of a thin refractive lens consisting of two spherical surfaces with radii of curvature R_1 and R_2 and made of a material (glass) with refractive index n is

$$\frac{1}{f(\lambda)} = (n(\lambda) - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad [18]$$

where λ is the wavelength of the illuminating light.

This equation is valid for all wavelengths and so also for the commonly used wavelength $\lambda_d = 587.6$ nm (yellow d-line of helium) which is near the center of the visible spectrum. Then $1/R_1 - 1/R_2$ can be expressed as $1/((n_d - 1)f_d)$ and the focal power depends on the wavelength as

$$\frac{1}{f(\lambda)} = \frac{n(\lambda) - 1}{n_d - 1}\frac{1}{f_d} \qquad [19]$$

where n_d and f_d are the refractive index and the focal length at $\lambda_d = 587.6$ nm, respectively. So, by combining two thin lenses made of materials with refractive indices n_1 and n_2 and focal lengths f_1 and f_2 the focal power of the doublet assuming a zero distance between the two lenses is:

$$\begin{aligned}\frac{1}{f_{ges}(\lambda)} &= \frac{1}{f_1(\lambda)} + \frac{1}{f_2(\lambda)} \\ &= \frac{n_1(\lambda) - 1}{n_{1,d} - 1}\frac{1}{f_{1,d}} + \frac{n_2(\lambda) - 1}{n_{2,d} - 1}\frac{1}{f_{2,d}} \qquad [20]\end{aligned}$$

For an achromatic doublet the focal power should be equal for two different wavelengths which are commonly the wavelengths $\lambda_F = 486.1$ nm (blue F-line of hydrogen) and $\lambda_C = 656.3$ nm (red C-line of hydrogen) which mark the limits of the well visible spectrum (of course the visible spectrum itself is broader but the sensitivity of the human eye to wavelengths below λ_F or beyond λ_C is quite bad). Therefore, the following equations result:

$$\begin{aligned}\frac{1}{f_{ges}(\lambda_F)} &= \frac{1}{f_{ges}(\lambda_C)} \Rightarrow \frac{n_1(\lambda_F) - n_1(\lambda_C)}{n_{1,d} - 1}\frac{1}{f_{1,d}} \\ &= -\frac{n_2(\lambda_F) - n_2(\lambda_C)}{n_{2,d} - 1}\frac{1}{f_{2,d}} \qquad [21]\end{aligned}$$

By introducing the Abbe number V_d

$$\begin{aligned}V_d &= \frac{n_d - 1}{n_F - n_C} \\ &= \frac{n(\lambda_d = 587.6\ \text{nm}) - 1}{n(\lambda_F = 486.1\ \text{nm}) - n(\lambda_C = 656.3\ \text{nm})} \qquad [22]\end{aligned}$$

equation [21] can be simplified to

$$V_{1,d}f_{1,d} = -V_{2,d}f_{2,d} \qquad [23]$$

Now, the wavelength dispersion of all conventional materials (glasses, polymers, etc.) is such that the refractive index decreases with increasing wavelength, i.e. the material has a so-called normal dispersion. This means that the Abbe number V_d is always positive and ranges from about 20 for high-dispersive glasses (e.g., SF59) to more than about 80 for low-dispersive glasses (e.g. Ultran 20). So, to fabricate a refractive achromatic doublet with a positive focal power it is necessary to take a positive lens with focal length f_1 made of a low-dispersive glass with a large Abbe number $V_{1,d}$ and a negative lens with focal length f_2 made of a high-dispersive glass with a small Abbe number $V_{2,d}$ (see Figure 5a). Therefore, the resulting focal power of the achromatic doublet

$$\frac{1}{f_{ges,d}} = \frac{1}{f_{1,d}} + \frac{1}{f_{2,d}} = \frac{1}{f_{1,d}}\left[1 - \frac{V_{2,d}}{V_{1,d}}\right] \qquad [24]$$

is always smaller than the focal power of the positive lens with focal power $1/f_{1,d}$ alone.

The situation is completely different if a refractive and a diffractive lens are used. From the Sweatt model

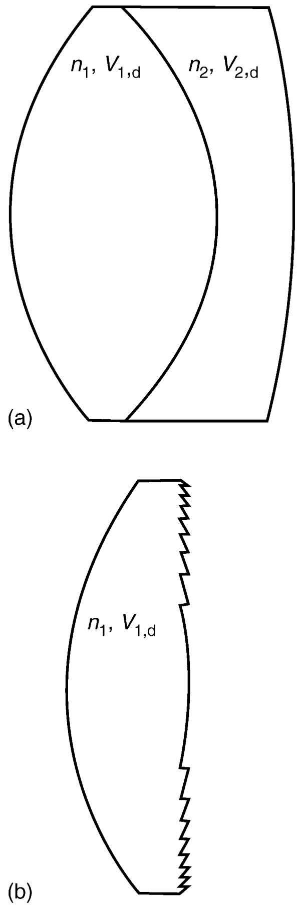

Figure 5 Structure of achromats: (a) refractive achromatic doublet consisting of a positive lens with low dispersion and a negative lens with high dispersion; (b) hybrid achromatic lens consisting of a positive refractive lens and a positive diffractive lens.

of a DOE (eqn [3]) it is known that the refractive index n_{sweatt} is proportional to the wavelength λ

$$n_{\mathrm{sweatt}}(\lambda) = \frac{n_0}{\lambda_0}\lambda = \frac{n_{\mathrm{d}}}{\lambda_{\mathrm{d}}}\lambda \qquad [25]$$

Therefore, the Abbe number of a DOE is

$$\begin{aligned} V_{\mathrm{d}} &= \frac{n_{\mathrm{d}} - 1}{n_{\mathrm{F}} - n_{\mathrm{C}}} \\ &= \frac{\lambda_{\mathrm{d}}}{\lambda_{\mathrm{F}} - \lambda_{\mathrm{C}}}\left(1 - \frac{1}{n_{\mathrm{d}}}\right) \overset{n_{\mathrm{d}}\to\infty}{=} \frac{\lambda_{\mathrm{d}}}{\lambda_{\mathrm{F}} - \lambda_{\mathrm{C}}} \\ &= -3.452 \end{aligned} \qquad [26]$$

Here, we used the fact that the accuracy of the Sweatt model increases with increasing refractive index so that the Abbe number of a DOE is achieved in the limit of an infinite refractive index. So, it can be seen that the Abbe number of a DOE is a small negative constant. This means that the dispersion of a DOE is high and opposite in sign to that of a refractive element. Therefore, according to eqns [23] and [26] a positive refractive lens with a large focal power (i.e., small focal length) and a positive diffractive lens with a small focal power (i.e., long focal length) can be combined to form an achromatic doublet with a focal power larger than the focal power of one of the two lenses alone (see Figure 5b). By using eqns [19] and [25] the focal power of a diffractive lens can be calculated with the Sweatt model (n very big $\rightarrow n - 1 \approx n$):

$$\frac{1}{f_{\mathrm{DOE}}(\lambda)} = \frac{\lambda}{\lambda_{\mathrm{d}} f_{\mathrm{d}}} \qquad [27]$$

In practice, the lenses are neither thin nor is the distance between them zero. In this case the equations for the design of a hybrid achromatic doublet are more complex (see Further Reading). Nevertheless, the given equations can be used for a first design and to understand the principal behavior.

Another advantage of a hybrid achromatic doublet is that the DOE can in addition behave like an asphere. Therefore, e.g., the spherical aberration can be corrected with the aspheric terms of the DOE provided that the spherical aberration is considerably smaller than the phase term encoded in the DOE to correct the chromatic aberrations. Otherwise, the correction of the monochromatic (spherical) aberration will disturb the correction of the chromatic aberrations. A comparison of a simple hybrid achromat, consisting of a planoconvex lens with the diffractive structure made directly onto the plane surface of the refractive lens and a conventional refractive achromatic doublet shows that the optical performance of the hybrid achromat is a little bit worse than that of the refractive achromatic doublet regarding, for example, off-axis aberrations. A simple explanation of this is that a refractive achromatic doublet has three spherical surfaces and two glasses, i.e., in total five parameters, whereas a hybrid achromat has one spherical surface, one glass and the DOE with a paraxial term and an aspheric term, i.e. in total about four parameters. Of course, the merit of these different types of parameters is not identical so that this assessment is only qualitative.

In the case of hybrid achromatic elements for a broad wavelength range (e.g., for the whole visible wavelength range) the dependence of the diffraction efficiency on the wavelength may be a limiting factor because the other diffraction orders behave like stray light and reduce the maximum contrast of the image. Therefore, special laminated blazed diffractive optical elements made of laminae of two or more different materials are described in the literature which show a nearly achromatic behavior of the diffraction efficiency but the 'normal' dispersion of DOEs. DOEs are in the meantime used in zoom lenses (see the cited patent in Further Reading) to correct chromatic aberrations.

Summary

In this article the possibilities of correcting monochromatic and chromatic aberrations in optical systems by using diffractive optical elements have been discussed. There are several advantages and disadvantages of using DOEs.

Advantages:

- DOEs offer many degrees of freedom in the design. By using modern lithographic fabrication techniques the fabrication of a DOE with a strongly aspheric phase function is just as expensive as fabricating a simple Fresnel zone lens with a quadratic phase function. So, DOEs can be used in some cases instead of aspheric surfaces which are more expensive. In addition, the lithographic fabrication technique guarantees a quite good accuracy concerning the phase function of the DOE compared to the more difficult fabrication of well-defined aspheric surfaces.
- The strong chromatic dispersion with the opposite sign to that of a refractive element allows the correction of chromatic aberrations. DOEs also give in this case an additional degree of freedom in the design because aspheric terms of the phase function can correct in addition monochromatic aberrations of the system as long as these aberrations are considerably smaller than the quadratic term of the phase function which corrects the chromatic aberrations.

Disadvantages:

- DOEs often exhibit not only one diffraction order but several diffraction orders due to fabrication errors, a broad spectrum of application wavelengths or because some DOEs naturally have several diffraction orders like, for example, binary DOEs. This can produce disturbing light in the optical system and limit the application.
- The strong dispersion can restrict the application in monochromatic systems because it requires in some cases (e.g., for testing aspheric surfaces) not only a relative constancy, i.e., constant but the absolute value does not have to be hit exactly, of the illuminating wavelength but an absolute constancy because otherwise systematic aberrations are introduced by the DOE.

So, there are many advantages of DOEs compared to other optical elements but on the other hand a DOE is not a magic bullet in correcting aberrations in optical systems. Depending on the system a careful assessment of the advantages and disadvantages has to be performed in order to find the best solution.

List of Units and Nomenclature

Abbe number	real number	V_d
Angle	[rad]	φ
Cartesian coordinates	[m]	x, y, z
Coefficients of a polynomial	different negative powers of m	a
Components of the ray vector	real numbers	e_x, e_y, e_z
Degree of a polynomial	integer number	G
Diffraction order	integer number	m
Focal length	[m]	f
Grating frequencies in x and y direction	[m^{-1}]	ν_x, ν_y
Height of a lens surface	[m]	h
Local grating periods in x and y direction	[m]	p_x, p_y
Merit function	real valued function	M
Modulus of the wave vector	[m^{-1}]	k
Nabla differential operator $\begin{pmatrix} \partial/\partial x \\ \partial/\partial y \\ \partial/\partial z \end{pmatrix}$	[m^{-1}]	∇
Number of sampling points/rays	integer number	N
Optical phase, phase function	[rad]	Φ
Polynomial	real valued function	P
Radial coordinate	[m]	r
Refractive index	real number	n
Sign variable	either +1 or −1	σ
Symbol for partial differentiation		∂
Unit vector in ray direction	vector with three real components	$\mathbf{e}$
Wavelength	[m]	λ

See also

Geometrical Optics: Aberrations.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Fairchild RC and Fienup JR (1982) Computer-originated aspheric holographic optical elements. *Optical Engineering* 21: 133–140.

Falkenstörfer O, Lindlein N, Keinonen T and Schwider J (1996) Diffraction-limited holographic lenses in dichromated gelatine. *Optical Engineering* 35: 2026–2033.

Herzig HP (1997) *Micro-Optics*. London: Taylor & Francis.

Kogelnik H (1969) Coupled wave theory for thick hologram gratings. *The Bell System Technology Journal* 48: 2909–2947.

Kress B and Meyrueis P (2000) *Digital Diffractive Optics*. Chicester, UK: Wiley.

Latta JN (1971) Computer-based analysis of holography using ray tracing. *Applied Optics* 10: 2698–2710.

Lindlein N, Schwider J, Schrader M, *et al.* (1993) Wavelength mismatch correcting elements for volume holograms written on a laser pattern generator. *Journal of Modern Optics* 40: 647–661.

MacGovern AJ and Wyant JC (1971) Computer generated holograms for testing optical elements. *Applied Optics* 10: 619–624.

Meier RW (1965) Magnification and third-order aberrations in holography. *Journal of the Optical Society of America* 55: 987–992.

Nishio A (1999) Zoom lens with diffractive optical elements. United States Patent 5978153.

Offner A (1966) Ray tracing through a holographic system. *Journal of the Optical Society of America* 56: 1509–1512.

Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim: Wiley-VCH.

Stone Th and George N (1988) Hybrid diffractive-refractive lenses and achromats. *Applied Optics* 27: 2960–2971.

Sweatt WC (1977) Describing holographic optical elements as lenses. *Journal of the Optical Society of America* 67: 803–808.

Upatnieks J, Vander Lugt A and Leith E (1966) Correction of lens aberrations by means of holograms. *Applied Optics* 5: 589–593.

Welford WT (1975) A vector raytracing equation for hologram lenses of arbitrary shape. *Optical Communications* 14: 322–323.

Winick K (1982) Designing efficient aberration-free holographic lenses in the presence of a construction-reconstruction wavelength shift. *Journal of the Optical Society of America* 72: 143–148.

Applications of Diffractive and Micro-Optics in Lithography

V Kettunen, Heptagon, Zurich, Switzerland

H P Herzig, University of Neuchâtel, Neuchâtel, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Most aspects of modern life are thoroughly influenced by the micro-electronics industry through microchips, such as microprocessors and memory chips. Optical lithography has been a key factor in the rapid pace of integration that drives the development of the micro-electronics industry. The photolithography process not only facilitates the fabrication of the chips, but also directly influences the limits of the obtainable chip performance in terms of speed or storage capacity, through the minimum feature size that can be created in the lithography steps. The increasing resolution demands of the IC industry, outlined in the International Technology Roadmap for Semiconductors as technology nodes, mean continuing demand for improvements in the photolithography systems.

The challenges to the improvements are encapsulated in the equation for the resolution of an optical lithography system:

$$R = k_1 \lambda / NA \qquad [1]$$

which expresses the resolution R in terms of the smallest resolvable half-pitch as a function of the wavelength λ, the numerical aperture NA of the exposure system and the unit less constant k_1, which is used as a measure of lithographic difficulty, with smaller values indicating a more challenging process. Traditionally, the increase in the resolution has been obtained by decreasing the used wavelength, with the current mark for volume production systems currently standing at 193 nm. Within each wavelength, further reduction has been traditionally obtained by increasing the numerical aperture towards the mathematical limit of 1. Finally, several so-called Resolution Enhancement Techniques (RET) have been introduced to reduce k_1, in order to reach subwavelength technology nodes within each lithography wavelength generation. It is in connection with these techniques that micro-optics has found a role in modern lithography systems.

In this chapter we discuss the ways micro-optics are being utilized in lithography systems. We will start with a brief overview of conventional lithography systems, of the essential parts of such systems as well as the special resolution enhancement techniques utilized in the systems. We will then discuss the utilization of micro-optics within the context of lithography systems, along with the unique challenges in terms of design and fabrication that are associated with such applications. Finally, we will briefly discuss nonconventional lithography applications, especially those based on utilization of micro-optics. It should be noted that the motivation of this article is to describe the issues related to the use of micro-optical elements in a modern lithography system and is not intended as a review of the modern lithography systems, nor is it a review of micro-optics.

Overview of Conventional Lithography Systems

Photolithography is a process whereby a specified pattern or mask is transferred onto a wafer using photons. In the process a photosensitive material, commonly known as a resist, is initially used to record the pattern that is generated when the mask is illuminated. After exposure, the resist is developed. The resist is referred to as positive or negative, depending on whether it is removed or remains, respectively, after the development of the irradiated regions. After its development, the pattern in the resist is then transferred onto the wafer using suitable chemical processes, such as wet chemical etching or dry plasma etching.

Several different lithographical techniques exist. The oldest and simplest is contact printing, where a photomask is brought into contact with a layer of photoresist on the wafer. Although this technique can reproduce high-resolution features, the contact between the wafer and the mask can lead to problems with cross contamination, and can damage either the wafer or the mask. Contact printing is mainly used nowadays in laboratory environments for small series of photolithography steps. A slightly more complicated technique is proximity printing, in which the photomask is held in close proximity to the surface of the wafer without actual contact. This avoids the problems of contamination and damage, but due to diffraction effects, reduces the maximum resolution of the process. Nevertheless, proximity printing remains a valuable technique for lithography applications, that do not require the highest possible resolution.

Modern volume production photolithography systems are based on the use of a projection system to image a mask onto a wafer through a complex system of lenses. There exist several variations of the basic idea. In some systems the entire mask is imaged at once onto the wafer. This approach works only when the feature size on the mask is reasonably large, in the range of a few microns. Reproducing smaller features with this approach is generally not possible, as the demands on the optical quality of the projection system, in the form of allowed aberrations, become extremely difficult to meet, especially when the physical size of the optical components increases with the reducing feature size. Therefore, so-called wafer steppers are used in modern microchip fabrication. The basic idea of a wafer stepper is to image only a small region of the mask onto the wafer, and pattern the entire surface area in consecutive steps.

The centerpiece of a wafer stepper is the exposure system. Its task is to image desired patterns from the photomask to their proper positions on the surface of the wafer. In many systems this process includes a size reduction between the patterns on the photomask and on the wafer. This reduces the dimensional accuracy requirements, such as feature size and positioning for the mask fabrication by whatever is the de-magnification of the lithography system. However, it also increases the complexity of the exposure system, as significance of aberrations is increased. The requirements of microchip production mean that the image of the photomask with submicron features must be reproduced on the wafer with a dimensional accuracy of only a few tens of nanometers and aligned to a specific position within a fraction of the linewidth. Furthermore, these tolerances must be guaranteed over the entire exposure field that is typically several square centimeters. It is not surprising that the resulting complex piece of optics costs several millions dollars.

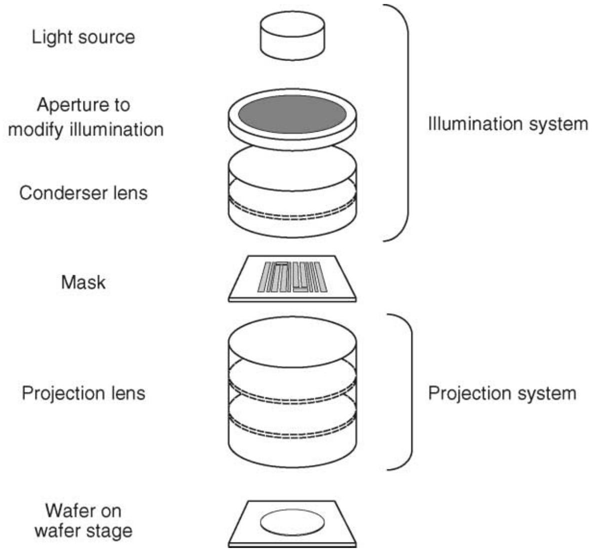

Figure 1 Schematic illustration of the exposure system.

The exposure system consists of a projection lens, illumination system, and wafer management system, as illustrated schematically in Figure 1.

A typical projection lens is a complicated lens system. As the resolution of a lens system is inversely proportional to the diameter of the lenses, the lenses in photolithographic systems can have diameters up to 250 mm, and the whole system is arranged inside a massive barrel-frame. The aim of the projection lens in the entire exposure system is to image the pattern from the photomask to the wafer. The projection lens is primarily responsible for the resolution and must therefore fulfill the highest optical requirements. To produce the smallest possible image, diffraction effects alone must essentially limit the resolution of the system. This means that the optical aberrations over the exposure field of several square centimeters must be virtually eliminated, by the wavefront deviations being less than a fraction of the wavelength.

The illumination system transfers the illumination source through the photomask to the entrance of the projection system. The main requirements for the illumination system are to collect most of the light into the system and to ensure that the irradiance of the photomask is uniform to within a few percent. In addition, resolution enhancement techniques, based on off-axis illumination, require that the illumination pattern on the photomask has a specific shape. We will discuss these requirements later in more detail. In general terms, the illumination system is less complicated than the projection system, especially with regard to the elimination of aberrations, due to the fact that precise imaging is not required in the illumination system.

Although the wafer management system is outside the scope of this article, we note in passing that it maintains the position, orientation, and movement of the wafers through the exposure system. Again, high precision is required to ensure the successful exposure of the desired pattern to the wafer.

Resolution Enhancement Techniques

In most modern lithography systems the achievable maximum resolution is increased via the use of one or more special resolution enhancement techniques (RET). These are essentially creative refinements of the lithography technology, based on theoretical optics papers published often decades before their first application in the lithography process in the late 1980s and early 1990s. There are essentially three RET approaches: Off-Axis Illumination (OAI), Optical Proximity Correction (OPC), and Phase Shifted Masks (PSM).

In optical proximity correction the basic idea is to predistort the mask pattern in order to overcome the proximity effects caused by the optics and the different processes. This is done either using a set of pre-existing rules describing the connection between a mask pattern and its realization on the wafer, using advanced models to predict the proximity effects during the lithography process, or with some combination of both. The key difficulties with OPC are related to the fabrication of the predistorted mask and to the accuracy of the used distortion rules or proximity effect models. Finally, it should be noted that OPC is strictly speaking not a true resolution enhancement technique, as it only enhances the effective resolution and not the theoretical resolution limit of the process defined by the optics.

In phase shifted masks the improvement in the resolution is obtained by altering the phase of light passing through different portions of the mask, creating regions of destructive interference in the image. This leads to higher contrast, which helps to improve the effective patterning resolution. However, as with OPC, there is no true resolution enhancement in terms of the theoretical resolution limit. There are two main difficulties with application of PSM. First, it is not straightforward to introduce required phase shifts into an arbitrary design layout, because some geometrical shapes, such as junctions are harder to realize than others. Another problem lies with the mask fabrication, which requires careful calibration of the material layers and can be very complicated and costly compared to fabrication of conventional binary masks.

For the purposes of this article, only off-axis illumination is discussed, as both OPC and PSM are essentially mask-based techniques and thus provide no opportunity for micro-optics. Off-axis illumination refers to any illumination scheme that significantly reduces or eliminates the light hitting the mask at near normal incidence, leading to the diffraction pattern of the mask being shifted on the projection lens. The principle behind the resolution enhancement, obtained by introducing such a shift, is schematically illustrated in Figure 2 for periodic pattern.

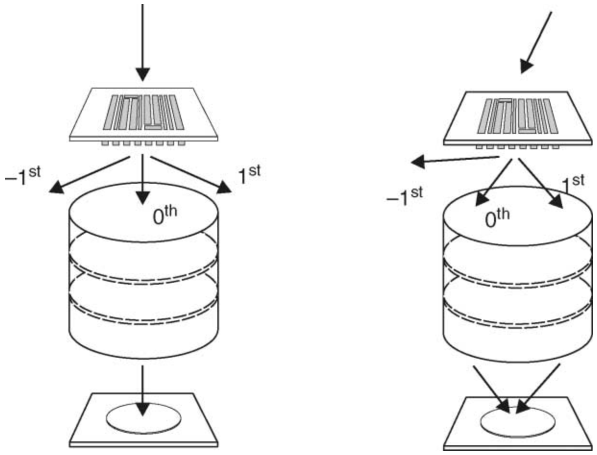

Figure 2 Schematic illustration of effect of off-axis illumination on image formation.

The mask pattern can be imaged if at least two diffraction orders generated by the mask pass through the pupil of the projection lens. Using on-axis illumination, the minimum feature size of the system is reached when the period of the mask pattern is so short that the ± 1st diffraction orders can no longer pass through the pupil. If the illumination is moved off-axis, the diffraction pattern is shifted and the second diffraction order (in this case +1st) again passes the pupil. Thus the resolution has been increased. A secondary improvement in contrast may be obtained by shaping the diffraction orders falling onto the pupil, such that the filling of the pupil is optimized, i.e., such that less unwanted light contributing to the background illumination passes the pupil. Finally, the depth of focus of the system can be optimized by ensuring that the contributing orders (in this case 0th and +1st) are symmetric with regards to the optical axis.

The conventional way to realize off-axis illumination schemes is to introduce a suitable aperture to the illumination optics between the source and the mask. In order to effectively use off-axis illumination, the shape and size of the aperture, i.e., the illumination pattern, must be optimized for the specific mask shape and pitch. This optimization leads to four basic shapes, as illustrated in Figure 3. In most modern lithography systems these shapes

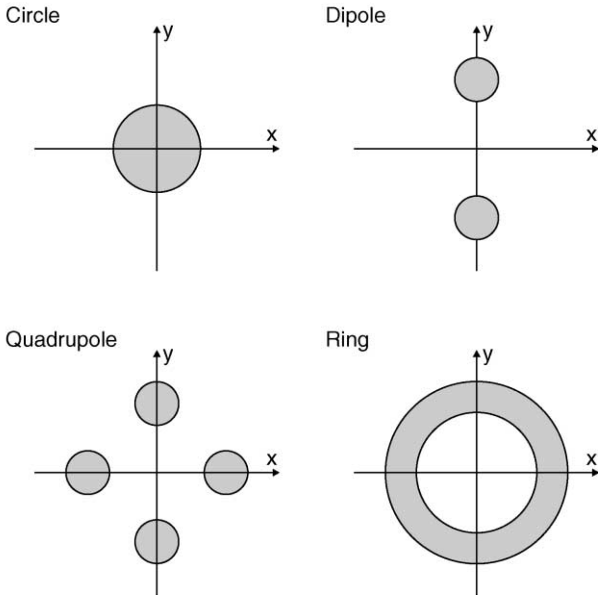

Figure 3 The four typical illumination patterns: disk, ring or annular, quadrupole, and dipole.

are available with different values of partial coherence factor, governed by the ratio of the numerical aperture of the illumination lens and projection lens, allowing control of proximity effects and edge sharpness.

Overview of Micro-Optics

There exists no generally agreed definition of micro-optics, but it typically means optical systems built to significantly smaller scales than conventional table-top systems, using components such as microlenses and microprisms with physical dimensions in the range of a couple of millimeters at most. Micro-optical systems also include so-called diffractive optical elements, which are usually optical components whose operation cannot be explained by Snell's law or the law of reflection. Such elements are usually macroscopically planar microstructures, consisting of features with dimensions from a couple of wavelengths to a few tens of microns, and are designed by advanced numerical algorithms based on diffraction theory.

In the past decade, micro-optics has emerged as a powerful new addition to the field of optics and optical systems, following improvements in micro-optics fabrication techniques, both via adaptation from established microelectronics fabrication processes and development of novel approaches, such as variable dose direct writing using either UV light or an electron beam. Additionally, emergence of replication techniques, such as injection molding or UV casting, allows the low-cost mass production of micro-optical elements, making micro-optics economically feasible for a wide range of consumer applications. Consequently, micro-optical elements, utilizing either refractive or diffractive surfaces, are now found in industrial applications such as material processing, in the optical telecom field, and in consumer electronics such as the digital video camera.

One of the major strengths of micro-optics, as compared to conventional optics, is their ability to realize complex optical functions into a compact form. This is especially true for micro-optics using diffractive optical elements, as diffractive elements may be designed to realize several optical functions such as focusing, filtering, or beam splitting at once allowing integration of several classical optical components into a single diffractive element.

Micro-Optics in Conventional Lithography Systems

The simplicity of contact or proximity printing means that there is little need or opportunity to incorporate additional functionality through inclusion of micro-optical elements. The only exception to this will be discussed in the final section of this article, where we will consider a nonconventional technique, which utilizes micro-optics and can be incorporated into proximity printing systems. However, in this section we will limit the discussion to projection-based lithography systems such as wafer steppers.

Let us consider the possible uses of micro-optics in the three main parts of the exposure system identified in the earlier section. It is obvious that the wafer management system, which is essentially mechanical, cannot be enhanced by inclusion of micro-optics. In principle, micro-optics could be incorporated into the projection system to reduce its weight and, to a lesser degree, complexity. However, it is important to note that using micro-optics to reduce the physical size, i.e., the diameter of the lenses in the projection system is not possible, as such reduction would increase the theoretical minimum feature size the system can resolve which is inversely proportional to the lens diameter. In practical terms, even the most advanced fabrication approaches used to realize micro-optical elements cannot yet produce elements with the precision and quality sufficient to meet the optical requirements associated with a lithographic projection lens. In general, the state of the art in micro-optics is such that it does not make sense to replace refractive lenses with diffractive ones unless the weight of the optics is a critical parameter in determining the merit of the system. This is not the case for wafer steppers.

Even though the optical requirements for the lenses in the illumination system are not as high as in the projections system, they are in practice still far too demanding for micro-optical elements. Therefore, the only area where micro-optics could be used in the illumination system is the generation of various illumination patterns needed for the off-axis illumination schemes used for resolution enhancement, as well as overall homogenization of light. An element meeting these goals is essentially a custom-designed diffuser, where both the generated angular range and the relative intensity of illumination in any given angular direction are controlled to a high degree.

Custom Diffusers for Illumination

As mentioned earlier, one of the strengths of micro-optics lies in its ability to combine different optical functions into a single element. When used in lithographic illumination systems, micro-optical elements can serve a dual purpose. The element can serve as an artificial aperture, effectively distributing the light over some predetermined angular range and at the same time minimizing light outside the range, allowing generation of various illumination patterns needed for the resolution enhancement via off-axis illumination. Unlike conventional solutions based on the use of actual apertures to generate the desired illumination patterns, distribution of light using micro-optical elements can be done with efficiencies up to 95%, with typical target efficiency being in the range of 90%, depending on the pattern complexity. At the same time, the element can be used to smooth the light distribution, improving the uniformity of the illumination. Typically the required uniformity error, i.e., allowed variations from the average illumination, is in the range of 5%, but can be as low as 1%.

In addition to the above-mentioned high-efficiency and -uniformity targets, optical quality requirements of modern lithography systems introduce unique challenges to the micro-optics used in such systems. First, micro-optical elements used in lithography systems must also exhibit low amounts of stray light, especially in formation of hot-spots. Typically, hot-spots are allowed to be only a few percent of the average illumination level. Because of the fact that the size, shape, and profile of the illumination beam can vary, partly due to variations in the source and partly because such variations improve the functionality of the lithography system, by, for example, controlling the partial coherence factor, the micro-optical elements must be virtually space-invariant. The elements must perform the same function regardless of where and how the illumination coming from the source hits the element. Furthermore, the elements themselves must have very low absorption, because the high power used could otherwise lead to destruction of the elements. This means that the materials must be carefully selected. Finally, the optical wavelengths used in advanced lithography systems are 248 nm or less. Because the feature size of a diffractive element scales down with the wavelength, the features in elements used are much smaller than in the case of elements used for applications in the visible region. Consequently the accurate fabrication of such elements becomes a formidable challenge, because most fabrication errors, such as mask misalignment, etch depth error or the way the sharp corners in diffractive elements get rounded during fabrication, increase in significance when the feature size is decreased.

There exist two distinct approaches for obtaining such custom-designed diffusers. We will now discuss the benefits and weaknesses of these two techniques in terms of efficiency, pattern uniformity, stray light, and ease of fabrication, in order to illustrate the trade-off a designer creating micro-optical elements for a lithography system has to consider.

Aperture Modulated Diffusers

One approach to acquire custom diffusers for illumination systems is to use an array of refractive or diffractive micro-lenses. More generally, the lens phase can be replaced by any phase obtained by geometrical considerations, i.e., by a geometrical map transform between the input and output planes of the optical system. The basic principle of an aperture modulated diffuser is to use the phase function to accurately control the angular divergence of the diffuser, while the shape of the desired intensity distribution is created by introducing a properly shaped aperture in which the phase function is contained. This basic principle is shown in Figure 4.

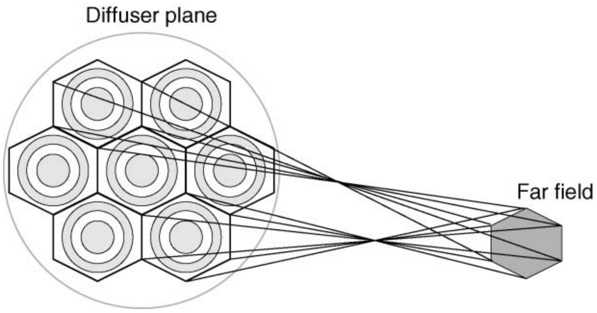

Figure 4 Basic principle of aperture modulated diffuser. The lens f-number determines the angular extend of the far field pattern, while the lens aperture shape determines the shape of the pattern.

The main benefit of using a lens phase lies in the ease of the design process, which essentially consists of selecting the numerical aperture of the lens to meet the angular size of the desired pattern, while more complicated phase profiles can be used to improve some aspects of the performance of the diffuser.

In terms of optical performance, aperture modulated diffusers typically achieve efficiency of up to 95% and low levels of stray light, easily fulfilling the target specifications outlined earlier, especially when refractive surfaces are used. However, since the lens aperture in the element plane controls the shape of the intensity distribution, the range of intensity distribution shapes is limited. This is necessary because it must be possible to tile the required aperture shape in order to realize an array of elements with high fill factor, i.e., to realize a space-invariant diffuser. Furthermore, introducing an aperture into the element plane means that intensity oscillations, due to boundary diffraction at the edge of the apertures, cannot be avoided. Consequently, the uniformity of the generated intensity distributions suffers. This is especially true in cases where the desired intensity pattern has sharp corners, where the intensity oscillations due to boundary diffraction, can easily be up to 50% of the average intensity of the desired intensity pattern. Figure 5 illustrates a typical far field intensity distribution obtainable with aperture modulated diffusers.

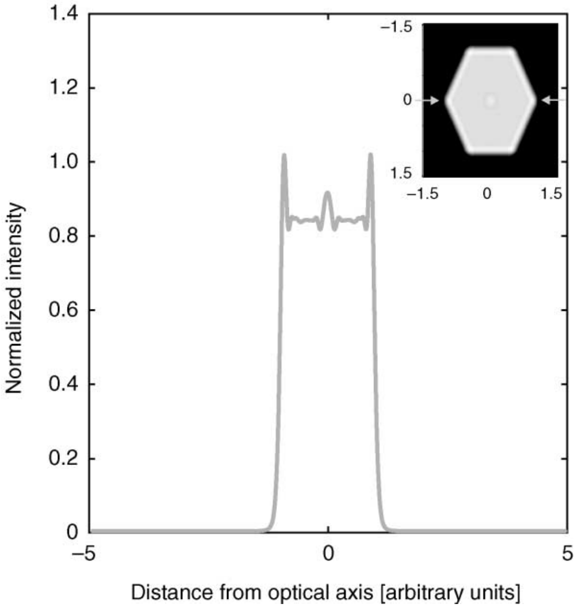

Figure 5 Typical intensity distribution obtainable with an aperture modulated diffuser. The overall distribution is smooth apart from the slow intensity oscillations at the edges resulting from diffraction at the aperture boundary.

One needs to consider several issues with regard to the fabrication of aperture modulated diffusers. For refractive aperture modulated diffusers, the main issue is the accurate realization of the required surface profile. The fabrication of large arrays of refractive elements having arbitrary surface profiles is still challenging. Resist reflow techniques, where the micro-lenses are realized by melting small islands of resist, lead to a spherical lens profile. Complicated and hard-to-control profile reshaping techniques are required if dramatic alterations from the spherical lens profiles are needed. Even though fabrication through direct writing techniques, such as laser writing offers more freedom, realization of arbitrary refractive surface remains a challenge. Therefore, use of refractive elements is mainly limited to cases where the lens profile can be close to the spherical shape. The most important of such cases are patterns with uniform circular or rectangular distributions. The former can be realized by proper selection of the numerical aperture of the lens, while the latter can be realized by combining two properly selected cylindrical lenses on opposite sides of the substrate.

For diffractive aperture modulated diffusers there are considerably less limitations in terms of what phase profiles can be attained. As long as the opening angles, i.e., the angular size of the pattern is below 10 degrees, the phase can be realized as a diffractive element. With diffractive elements, the main issue to consider is therefore the effect of typical fabrication errors such as profile depth and shape errors. First, the typical consequence of fabrication errors in the element is loss of efficiency via increase of stray light. Fortunately the errors do not generate sharp stray light peaks, with the obvious exception of the axis hot-spot. This results from the fact that both lens functions and more complicated phase functions designed by geometrical considerations, exhibit locally periodic structures, and the introduction of common fabrication errors then leads to redistribution of light away from the desired signal orders of these local structures to higher or lower orders. On the level of the whole intensity distribution, such redistribution appears as a constant rise in the background noise level, but will not lead to sharp intensity fluctuations.

Another issue related to the fabrication of aperture modulated diffusers appears when diffuser elements are fabricated using a series of photomasks, as it is typically the case with a relatively large size of the element. Although mask misalignment is essentially a random error, i.e., the direction and amount of the error are difficult to predict before hand, it easily leads to systematic and partly symmetric uniformity problems with aperture modulated diffusers.

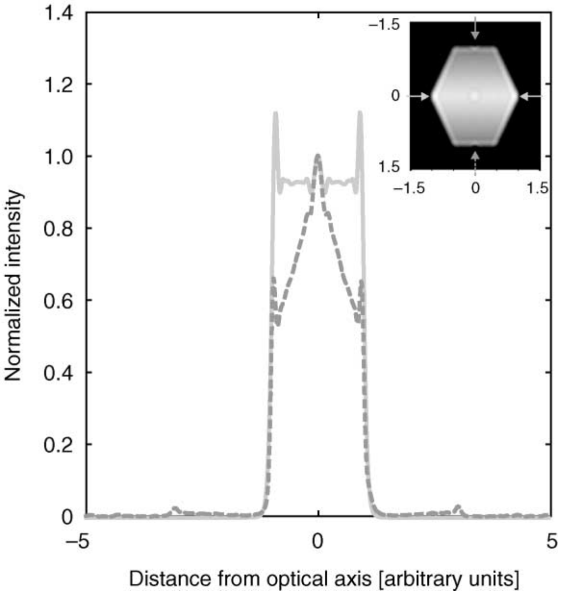

Figure 6 Effect of mask misalignment on the intensity distribution with aperture modulated diffusers. Essentially random error leads to symmetric and systematic uniformity error due to locally periodic nature of the design.

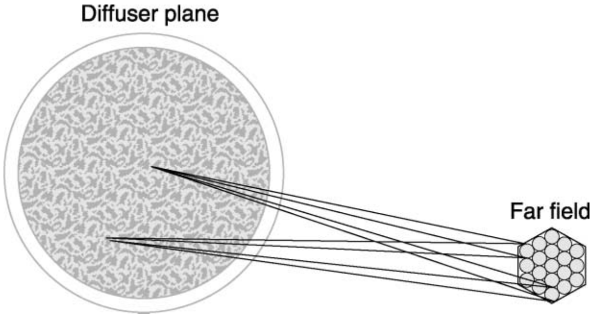

Figure 7 Basic principle of fan-out based diffusers. A grating is used to generate a high number of closely spaced diffraction orders arranged in the shape of the desired pattern.

Again the reason for this lies in the locally periodic nature of the elements. A single mask misalignment along some direction results in an erroneous profile shape in the structures, which are periodic in the direction of the alignment error. Consequently, the relative intensity distribution along that direction is changed. On the other hand, structures, where the period is perpendicular to the misalignment, are realized correctly. Furthermore, the local period of the phase function is often changing from the center of the element to the edge, meaning that the relative magnitude of the mask misalignment is different in different parts of the element. These facts lead to asymmetry in the overall intensity distribution, which is especially harmful with off-axis illumination patterns, as the asymmetry results in loss of the resolution gain obtained with off-axis illumination. Furthermore, as the source of the asymmetry is essentially a random error, compensation in the design is very difficult. Figure 6 illustrates the above-mentioned phenomena.

Fan-Out Based Diffusers

An alternative technique uses specially designed fan-out elements to produce a large number of closely spaced diffraction orders arranged in a desired pattern, as illustrated in Figure 7. Such an approach offers all the flexibility of diffractive elements, allowing generation of nearly arbitrary intensity distributions, especially if the elements can be realized as so-called multilevel elements, i.e., if the surface profile can have more than two distinct levels.

The flexibility in the design comes at a cost, where specially developed design tools, such as the iterative Fourier transform algorithm, are required. Fortunately such tools are starting to become commercially available as emphasis and interest in wave-optical engineering, an extension of the classical optical engineering approaches, increases.

In terms of optical performance, fan-out based diffusers offer high uniformity as there are no boundary diffraction effects as with aperture modulated diffusers. Furthermore, by choosing the beam separation of the fan-out to be clearly smaller than the divergence of the incidence beam, it is possible to average out any local intensity variations from order to order, leading to a smooth overall intensity distribution. The remaining variations in the intensity distribution generated by a fan-out based diffuser can be as small as 1% of the average intensity.

As a trade-off for the improved performance in terms of uniformity compared to aperture modulated diffusers, fan-out based diffusers typically exhibit lower efficiency and, consequently, higher levels of stray light, even when no fabrication errors are present. This is a result of additional design freedoms being used to remove unwanted intensity oscillations from the far field regions, belonging to the desired intensity distribution, by introducing some noise outside these regions. Typically the efficiency of a fan-out based diffuser is in the range of 85–90%, or 5–10 percentage points lower than the efficiency of a comparable aperture modulated diffuser (should one exist). Furthermore, unlike aperture modulated diffusers, fan-out based diffusers can have stray light that forms hot-spots. If care is not taken during the design, these hot-spots may be up to 10% of the average intensity of the illumination pattern, even in the absence of fabrication errors. Figure 8 illustrates

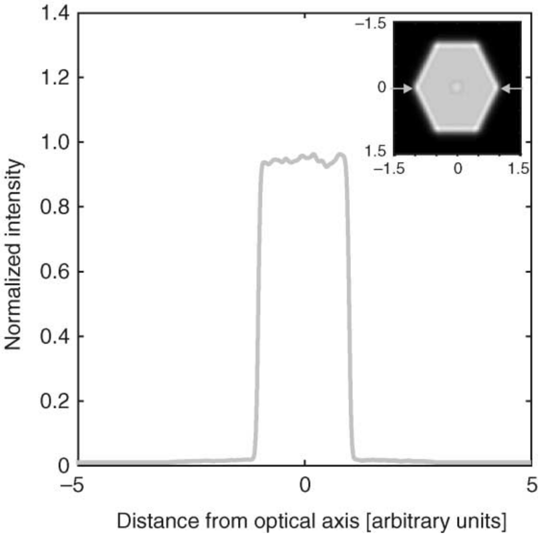

Figure 8 Typical intensity distribution obtainable with a fan-out based diffuser. Overall distribution is smooth, but stray light appears outside the main pattern.

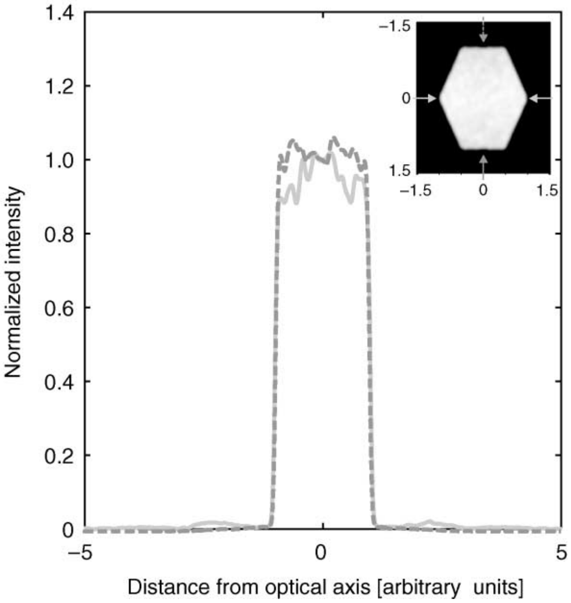

Figure 9 Effect of mask misalignment on the intensity distribution with fan-out based diffusers. The uniformity error appears randomly, allowing statistical compensation schemes.

a typical far field intensity distribution obtainable with fan-out based diffusers.

Fan-out based diffusers offer several advantages in terms of fabrication compared to mapping based diffusers. First, random errors, such as mask misalignment, do not lead to symmetric or systematic uniformity problems, but instead result in random local oscillations in the intensity pattern as illustrated in Figure 9. This allows special compensation schemes where the element is divided into segments, each of which has different fabrication tolerances but generates the same far field intensity distribution. If the fabrication tolerances of the individual designs are balanced correctly against the typical fabrication errors of the used fabrication methods, the loss of uniformity due to fabrication errors can be eliminated almost entirely. Furthermore, the flexibility of diffractive design allows for fabrication error-tolerant designs. This follows from the ability to combine several optical functions in a single optical element. The typical trade-off with designs with relaxed fabrication tolerances is that the optical performance of the element in terms of efficiency is worse than that of a conventional design if fabrication errors are not present.

Micro-Optics and Non-Conventional Lithography

In the previous section we have considered micro-optics in connection with conventional lithography approaches and shown that micro-optical elements can find a significant role in such systems. We conclude this article by considering briefly nonconventional lithography systems, focusing especially on a technique called micro-lens lithography, which is fundamentally based on the use of micro-optical elements.

The most straightforward way to improve existing lithography approaches is to further reduce the wavelength used in the exposure process. In X-ray lithography, the light source used in conventional lithography approaches is replaced by a cyclotron emitting light with wavelengths deep in the X-ray region. This leads to minimum resolution far lower than achievable by conventional techniques, but with the additional difficulty and cost of maintaining and running such a source.

The techniques that are currently considered as serious alternatives for conventional photolithography approaches are based on the use of particles other than photons in the fundamental exposure process. Techniques such as electron- or ion-beam lithography are routinely used to realize structures with feature size and positioning resolution of some tens of nanometers for use in both micro-electronics and micro-optics. The main drawback of these approaches is the cost involved in patterning large surface areas. Unlike optical lithography techniques, particle beam lithography cannot expose large surface areas at once, instead requiring painstaking scanning of the surface area with a relatively small writing beam. This leads to processing times that scale linearly with the required surface area, and consequently the throughput of such techniques is small. Therefore, particle beam lithography is not yet widely used in commercial fabrication of micro-electronics. However, it is worth noting that especially electron beam lithography is widely used in the origination of micro-optical components, as the relatively high cost of creating the master component can be offset in the final product by the inexpensive replication process. Such an approach allows the replicated end-product to meet the high optical requirements achievable only by high resolution surface control, yet remain affordable for use in, for example, consumer electronics. This has lead some in the micro-optics community to conclude that electrons are better suited for fabrication of components controlling photons, while photons are better suited for fabrication components controlling electrons.

In addition to particle beam lithography, other noteworthy techniques exist. For our purposes, the most interesting technique is the so-called micro-lens array lithography, where one or more micro-lens arrays are used to image the pattern of the photomask to the resist layer on the substrate. Although the resolution of this approach is limited to 3–5 microns, for applications such as with flat panel displays, micromechanics, or multichip module fabrication, this technique is of considerable interest.

It is well known that the geometrical aberrations scale down proportionally with the physical lens size. Thus, large areas can be imaged favorably by use of arrays of micro-lenses, where each lens images only a small section of the whole field. Such a setup is shown in Figure 10.

With this setup it is possible to expose a comparatively large area with a single exposure, thus eliminating the need for costly stepping operations and therefore increasing the throughput of the system. Analysis of the basic system shows that an array based on spherical micro-lenses of 1 mm focal length and 300 micron aperture can produce a nearly diffraction limited resolution on the order of 3–5 micrometers over a 300 micron field.

In terms of conventional lithography systems, the micro-lens array lithography has most in common with proximity printing. Both techniques achieve comparable resolutions, yet avoid damaging the mask or substrate by keeping the two separated by some finite working distance. However, micro-lens array lithography offers two distinct advantages over conventional proximity printing. First, the working distance between the imaging system and the substrate surface can be much larger, which further minimizes the risk of mask or substrate damage. Furthermore, added working distance allows lithography steps on nonflat surfaces, e.g., inside holes or grooves where conventional proximity printing would not be possible. Another advantage of the micro-lens array lithography is the larger depth of focus. With proximity printing, the resolution rapidly decreases with increasing distance. In micro-lens array lithography, the optimum resolution is obtained at the image plane, while a region of reasonable resolution extends forward and backward from this plane (Figure 11). This extension means that in the depth of field, resolution is less sensitive to changes in the working distance, which makes it better suited for practical applications.

In addition to the optical benefits of micro-lens array lithography, there also exist other practical benefits. Many micro-devices have a repetitive and

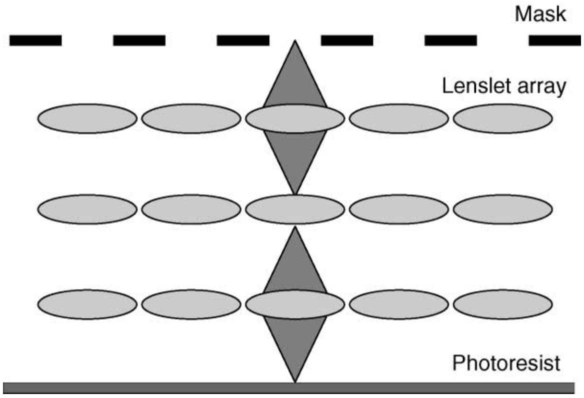

Figure 10 Basic geometry of micro-lens array lithography.

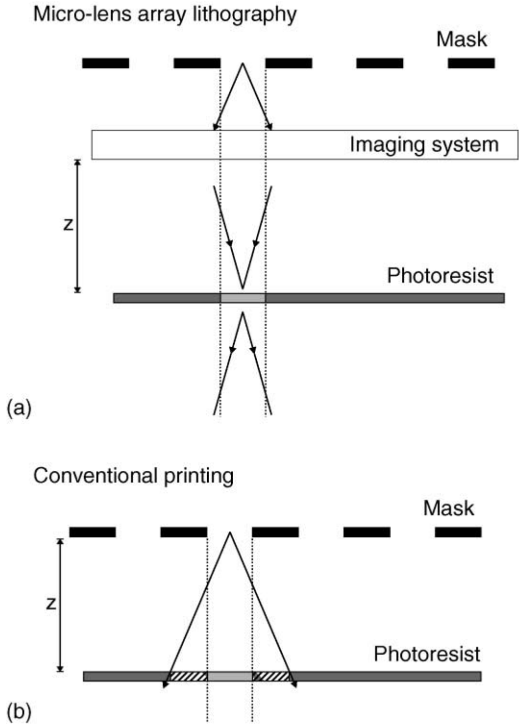

Figure 11 Depth of focus with micro-lens array lithography and conventional printing. (a) With micro-lens array lithography, the image of the mask is extended both forward and backward from the working plane at distance *z*. (b) In conventional printing, the resolution is rapidly reduced with increasing working distance *z*.

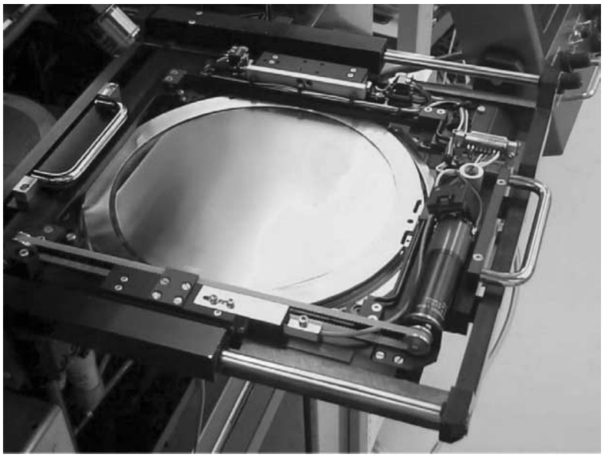

Figure 12 Ultra-flat micro-lens projection system (courtesy of R. Völkel, SUSS MicroOptics, Neuchâtel, Switzerland).

discrete nature with respect to some unit element. Such a unit element lends itself well to the imaging properties particular to lens arrays, as each lens of the array can be made to correspond to the unit cell of the particular micro-device making the imaging highly efficient. A photograph of a micro-lens projection system is shown in Figure 12.

One could also imagine that instead of an array of imaging systems, an array of pattern generators is used that generates a specific pattern, such as lines or points. In the ideal case such a system would replace the mask and the imaging system altogether. However, this approach is only possible for simple patterns and large production series.

Outlook

As long as optical lithography stays attractive, micro-optics will be used as beam-homogenizers and sophisticated diffusers to generate optimized illumination patterns. How far micro-optics will be used in the imaging system is difficult to predict. However, it can be assumed that the difference between micro-optics and macro-optics fabrication technology will disappear.

Recent research is also investigating near-field optics to further reduce the resolution limit. Micro-optics technology is well suited to realize such systems. A serious problem of near-field optics is the short distance (10 nm–100 nm) between the optical element and the interface to be patterned. In addition, near-field optics is based on scanning systems, which increases the exposure time considerable. However, there is still room for improvements or as Richard Feynman said 'there is plenty of room at the bottom'.

Further Reading

Dändliker R, Gray S, Clube F, Herzig HP and Völkel R (1995) Non-conventional techniques for optical lithography. *Microelectronic Engineering* 27: 205–211.

Herzig HP (1998) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Herzig HP and Kipfer P (1999) Aperture modulated diffusers (AMDs). In: Asakura T (ed.) *International Trends in Optics and Photonics*. Berlin: Springer.

Naumann H and Schröder G (1992) *Bauelemente der Optik*. München, Germany: Carl Hanser Verlag.

Sheats JR and Smith BW (eds) (1998) *Microlithography: Science and Technology*. New York: Marcel Decker.

Turunen J and Wyrowski F (eds) (1997) *Diffractive Optics for Industrial And Commercial Applications*. Berlin: Wiley-VCH.

Völkel R, Herzig HP, Nussbaum P, Dändliker R and Hugle WB (1996) Microlens array imaging system for photolithography. *Optical Engineering* 35: 3323–3330.

Design and Fabrication of Diffractive Optical Elements

D W Prather, T Dillon, A Sure and X Gao,
University of Delaware, Newark, DE, USA

J N Mait, US Army Research Laboratory, Adelphi, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Diffractive optical elements (DOEs) are passive devices that redirect and focus light through the division and mutual interference of a propagating electromagnetic wave. This is in contrast to refractive elements, which redirect and focus light through variations in indices of refraction. From a historical perspective, diffraction was traditionally considered a detriment in optics and caused images to blur. However, we now know that it is possible to use diffraction to enhance the performance of optical systems. For example, DOEs can be used to correct chromatic aberrations and to implement optical functions that conventional refractive and reflective optics cannot, such as multiple beamsplitting.

Further, recent advances in computing and fabrication technologies have increased the number of applications in which diffractive elements are used, which has established diffractive optics as yet another tool in an optical designer's kit.

The first modern diffractive element was the hologram. A hologram is a photographic record of the constructive and destructive interference between two coherent beams known as the object and write beams. The object beam contains information about the object recorded. The write beam is necessary to record the hologram but contains no information itself. When the hologram is played back with a third beam, the read beam, the recorded interference pattern diffracts it in such a way that it reconstructs the object beam.

Unfortunately, the functions that a natural hologram can produce are limited to those whose wavefields can be produced in a laboratory. Thus, although it is possible to record a hologram that can focus or deflect a single beam, it is difficult to record a hologram that splits a single beam into multiple foci. However, the advent of computing technology made it possible to design such holograms analytically. The first diffractive element designed with the aid of a computer was demonstrated in 1966. Since then the ability to create elements that respond arbitrarily has had widespread application in imaging systems, telecommunications, and sensing.

The two most basic diffractive elements are the grating and the Fresnel zone plate. The diffractive behavior of a grating is predominately transverse to the optic axis. It causes light to split into discrete beams, referred to as diffraction orders, each with a different angle of propagation relative to the optic axis. The diffractive behavior of a Fresnel zone plate also divides incoming light into discrete orders, however, the orders are produced along the optic axis. This is useful in producing the diffractive equivalent of a refractive lens. Arbitrary transverse and longitudinal behavior is realized by combining characteristics of these basic two diffractive elements. In the remainder of this article, we discuss the design and fabrication of general diffractive elements.

Design

Once an optical system designer determines the functional behavior required for a particular application, for example, focusing or multiple beamsplitting, he or she needs to design a diffractive element that realizes this function. However, unlike refractive and reflective optical elements, whose material properties and surface curvatures determine their functionality, the functionality of a diffractive element is determined by a pattern of wavelength scale features. Determining this pattern is the goal of diffractive design but to do so one must first understand the physical relationship between the DOE and the wavefield it produces.

Scalar diffraction theory is one of the simpler models that relates the effect of a surface relief, or diffractive profile, $d(x, y)$, on an incident wavefield $U_{\text{in}}(x, y)$:

$$U_{\text{out}}(x, y) = \exp[\mathrm{i}\theta(x, y)]U_{\text{in}}(x, y) \qquad [1]$$

where the phase transformation $\theta(x, y)$ is related to the diffractive profile $d(x, y)$ by

$$\theta(x, y) = 2\pi(n - 1)d(x, y)/\lambda \qquad [2]$$

The constant n is the refractive index of the optical substrate and λ is the wavelength of illumination in a vacuum.

The effects of the phase change are not apparent until the wavefield $U_{\text{out}}(x, y)$ propagates over some distance z. In scalar diffraction theory the relationship between $U_{\text{out}}(x, y)$ and the wavefield $U_z(x, y)$ at a plane a distance z away, is described by the Rayleigh–Sommerfeld diffraction equation:

$$U_z(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\text{out}}(\zeta,\eta) \times \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x-\zeta/z)^2+(y-\eta/z)^2}\right\}}{\mathrm{i}z\lambda\left[1+(x-\zeta/z)^2+(y-\eta/z)^2\right]}\mathrm{d}\zeta\,\mathrm{d}\eta \qquad [3]$$

Huygens' principle is reflected in this formulation by the convolution between the transmitted field $U_{\text{out}}(x,y)$ and a spherical wave $h(x,y)$, which represents free space propagation over the distance z:

$$U_z(x,y) = U_{\text{out}}(x,y)^{**}h(x,y) \qquad [4]$$

where

$$h(x,y) = \frac{\exp\left\{\mathrm{i}(2\pi z/\lambda)\sqrt{1+(x/z)^2+(y/z)^2}\right\}}{\mathrm{i}z\lambda\left[1+(x/z)^2+(y/z)^2\right]} \qquad [5]$$

The Rayleigh–Sommerfeld formulation is valid so long as the angles at which energy diffracts from the aperture are small. If the distance z is sufficiently large that the angles stay clustered about the optic axis, one can replace the spherical wave response of freespace with a parabolic wave:

$$h(x,y) \approx \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\exp[\mathrm{i}(\pi/\lambda z)(x^2+y^2)] \qquad [6]$$

This simplification is referred to as the paraxial, or Fresnel, approximation. The diffraction integral is now given by:

$$\begin{aligned} U_z(x,y) \approx\ & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\mathrm{out}}(\zeta,\eta) \\ & \times \exp\{\mathrm{i}(\pi/\lambda z)[(x-\zeta)^2+(y-\eta)^2]\}\mathrm{d}\zeta\,\mathrm{d}\eta \\ =\ & \frac{\exp(\mathrm{i}2\pi z/\lambda)}{\mathrm{i}z\lambda}\exp[\mathrm{i}(\pi/\lambda z)(x^2+y^2)] \\ & \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U_{\mathrm{out}}(\zeta,\eta)\exp[\mathrm{i}(\pi/\lambda z)(\zeta^2+\eta^2)] \\ & \times \exp\{-\mathrm{i}2\pi[\zeta(x/\lambda z)+\eta(y/\lambda z)]\}\mathrm{d}\zeta\,\mathrm{d}\eta \end{aligned} \qquad [7]$$

The second representation results from an expansion of the exponent and allows the diffraction integral to be written as a Fourier transform. The quadratic phase terms can be removed if a positive lens with focal length z is placed next to the diffractive element and a second positive lens with focal length z is placed in the image plane (see Figure 1). Under these conditions, $U_{\mathrm{out}}(x,y)$ and $U_z(x,y)$ form a Fourier transform pair:

$$\begin{aligned} U_z(x,y) &= \mathrm{FT}[U_{\mathrm{out}}(x,y)]|_{u=x/\lambda z, v=y/\lambda z} \\ &= \mathrm{FT}\{\exp[\mathrm{i}\theta(x,y)]\}^{**}\mathrm{FT}[U_{\mathrm{in}}(x,y)]|_{u=x/\lambda z, v=y/\lambda z} \end{aligned} \qquad [8]$$

The objective of diffractive design is to find a surface relief $d(x,y)$ that generates a desired response $q(x,y)$ somewhere in the response $U_z(x,y)$. However, eqns [3] and [7] neglect the fabrication process, which affects the fields that can actually be realized. A diffractive optic designer can incorporate fabrication constraints into their design either directly or indirectly. In a direct approach, the designer represents the diffractive structure parametrically, for example, a multilevel phase element can be represented by a discrete set of allowed phase values and a binary phase element can be represented by the contours that separate one phase value from the other. A designer defines the diffractive element by specifying and modifying these parameters.

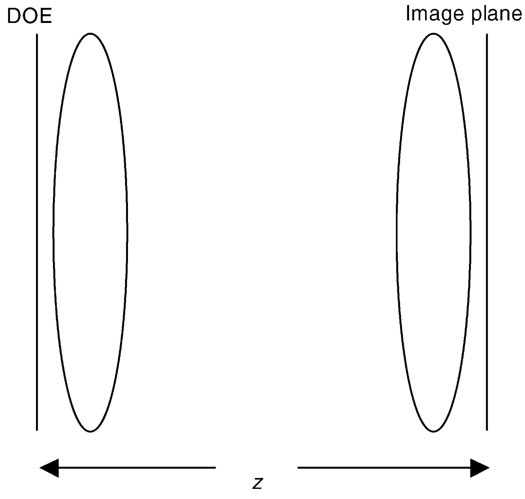

Figure 1 Optical Fourier transform module.

In an indirect approach, the designer first determines an intermediate function without consideration for fabrication and afterward applies the fabrication constraints to this function. For example, if $f(u,v)$ is the desired response one wishes to generate in the z-plane using $U_{\mathrm{in}}(x,y)$ as the input, from eqn [8], the optical function required to do is

$$Q(x,y) = F(x,y)/U_{\mathrm{in}}(x,y) \qquad [9]$$

where $F(x,y)$ is the Fourier transform of $f(u,v)$. The designer must then find a phase $\theta(x,y)$ that satisfies the fabrication constraints and insures that $\exp[\mathrm{i}\theta(x,y)] \approx Q(x,y)$.

For example, fabrication using multiple binary masks yields a phase structure quantized to 2^N levels. An effective design should, therefore, yield a quantized phase structure that minimizes reconstruction error. A naïve approach is to set the magnitude of $Q(x,y)$ in eqn [9] to unity and quantize the phase. However, this introduces considerable error. A direct approach to design searches only over the set of quantized phase structures for the one that, for example, generates the desired response with maximum energy and minimum error. An indirect design seeks first a phase structure whose response produces a minimum amount of error between it and $Q(x,y)$ and then quantizes this structure.

Performance of the element can be improved if one exploits certain freedoms available in design. For example, all systems have a finite-sized detector. It is, therefore, unnecessary to specify the performance of the diffractive element everywhere in the image plane. Its behavior within the detector window is all that is important. If only the intensity of the response is specified, then the phase of the response within the window is also a design freedom. Finally, one can also control the amount of energy contained in the signal window by scaling the amplitude of the response. This allows a designer to balance diffraction efficiency with error in the signal window.

Two general nonlinear optimization algorithms for the design of DOEs are represented in Figure 2. Indicative of the nature of data flow, Figure 2a is

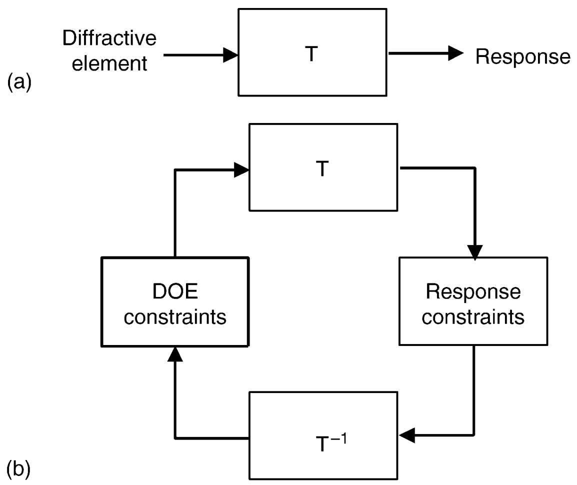

Figure 2 Algorithms used for diffractive design. (a) Unidirectional and (b) bidirectional algorithms.

referred to as a unidirectional algorithm and Figure 2b as bidirectional. In a unidirectional algorithm, changes are made only to the diffractive element. The changes are driven by how well the element performed in a previous iteration. If models exist for both forward and inverse propagation (T and T^{-1}, respectively), a designer can use a bidirectional algorithm, which allows modification of both the diffractive element and its response. This allows bidirectional algorithms to converge to a solution, in general, more rapidly than unidirectional algorithms. In scalar theory, T is a Fourier transform but more rigorous models exist. If the model of the optical system cannot be inverted, a unidirectional algorithm must be used.

Fabrication

Regardless of which design algorithm is used to determine $d(x, y)$, ultimately it needs to be realized physically. Although numerous techniques exist for doing so, by far the most common are based on microlithographic processes. The process requires exposing a layer of photoresist deposited on top of an optical substrate and etching the photoresist to transfer the pattern into the substrate. These methods are especially appealing because they leverage the existing manufacturing infrastructure of the microelectronics industry and, as a consequence, have enabled the efficient and economical mass manufacture of DOEs.

Two basic methods exist for patterning photoresist, direct writing using a modulated beam and indirect patterning using masks. Further, the mask process can be either a single or multistep process. The multistep approach uses N masks and N exposures to fabricate a multilevel diffractive element that has 2^N phase levels (see Figure 2). However, profiles that contain very small features are difficult to fabricate in this manner. As a consequence, alternative processes using single-step grayscale lithography have also been developed.

Although grayscale photolithography is more time-efficient and less prone to mask alignment errors than multistep lithography, the photoresist must respond linearly to illumination. Unfortunately, most photoresists used to fabricate microelectronic circuits are highly nonlinear. However, one can obtain a linear response by adjusting the exposure and development process. In the following, we review fabrication using both multiple binary masks and a single grayscale mask.

Multilevel Fabrication Using Binary Masks

In this fabrication method, one must first generate a sequence of N masks that are related, in a recursive manner, to the quantized phase $\theta(x, y)$:

$$m_k(x, y) = \mathrm{mod}(M_k(x, y), 2), \quad k = [1, N] \qquad [10]$$

where

$$M_{k+1}(x, y) = \frac{M_k(x, y) - m_k(x, y)}{2} \qquad [11]$$

and

$$M_1(x, y) = \theta(x, y)\frac{N}{2\pi} \qquad [12]$$

The masks can be produced using either optical or electron beam (e-beam) pattern generators from integrated circuit fabrication. The pattern generators expose a thin layer of photoresist on the surface of a chrome-covered quartz substrate. Development of the photoresist leaves the desired diffractive pattern in the unexposed areas. The diffractive pattern is transferred to the chrome layer by etching any chrome not covered by photoresist. After the chrome etch process is completed, any remaining photoresist is washed away to yield the finished lithographic mask. This process is repeated for each mask.

The completed mask set is used to fabricate the DOE using techniques similar to those used to generate the mask set. An optical substrate is coated with a thin layer of photoresist, placed in intimate contact with the first lithographic mask $m_1(x, y)$, and illuminated with an ultraviolet (UV) exposure lamp. During photoresist development, exposed resist is washed away, which leaves a binary pattern in the remaining developed photoresist.

When the pattern is transferred into the substrate material via etching, the photoresist acts as an etch stop. For the first mask, the etch depth into the substrate corresponds to a π-phase shift. The process is repeated for each subsequent mask, that is, the substrate is coated with a layer of photoresist, UV illumination exposes the photoresist through the mask, the photoresist is developed, and the substrate etched by the remaining pattern. The etch depth of each mask is half that of the previous mask. Figure 3 illustrates this process.

The substrate can be etched using either wet or dry techniques. In wet etching, one uses reactive chemicals, such as hydrofluoric acid, to etch the substrate. However, wet etching is isotropic, that is, substrate etching occurs equally in all directions. As a result, it is difficult to preserve vertical sidewalls and to achieve high fidelity in etching high-aspect ratio features. As a consequence, dry etching is more commonly used.

A dry etch uses a beam of ions to remove substrate material. The process for removal can be either chemical or kinetic. The process used is determined by the value of the plasma power that creates the source of ions relative to the bias power used to accelerate the ions toward the substrate. If the bias power is large and the plasma power small, a small stream of ions are accelerated rapidly toward the substrate. In this case, the operational mode is similar to a micromilling machine, wherein inert particles chip away at the surface by virtue of their kinetic energy. Alternatively, a small bias power and a large plasma power produce a large population of ions with no preferential directionality. This yields an isotropic process similar to wet etching.

Two methods for performing dry etching are reactive ion etching (RIE) and inductively coupled plasma (ICP) etching. However, in RIE systems the bias and plasma power are not independent due to capacitive coupling. Thus, adjusting one affects the other. ICP, on the other hand provides independent control over the bias and plasma powers. Independent control allows one to control precisely the anisotropic nature of an etch process. As mentioned above, this is achieved by varying the bias power relative to the plasma power, which controls the directionality of the acceleration.

Independent control is important because the ideal process for a given DOE depends on the substrate material and the characteristics of the diffractive profile. Independent control allows one to adjust the respective powers, as well as gas flow rates, and their partial pressures, to determine an optimal etch process. Dry etching allows one to incorporate endpoint detection, which provides precise control

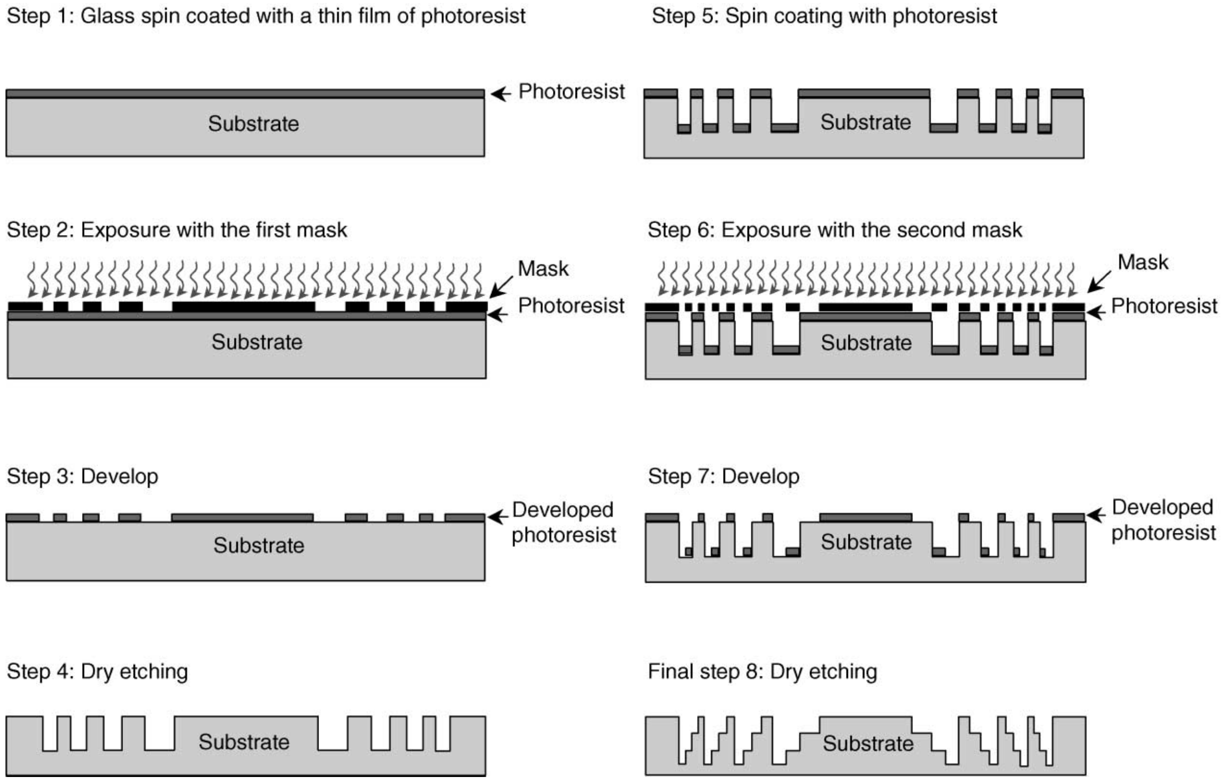
Step 1: Glass spin coated with a thin film of photoresist
Photoresist
Substrate
Step 2: Exposure with the first mask
Mask
Photoresist
Substrate
Step 3: Develop
Developed photoresist
Substrate
Step 4: Dry etching
Substrate
Step 5: Spin coating with photoresist
Substrate
Step 6: Exposure with the second mask
Mask
Photoresist
Substrate
Step 7: Develop
Developed photoresist
Substrate
Final step 8: Dry etching
Substrate

Figure 3 Fabrication of a multilevel quantized phase DOE using binary masks.

over the etch depth. As presented below, achieving the proper etch depth is important to generating responses with high diffraction efficiency.

Care must be used when determining the number of phase levels appropriate for a particular application. The number of phase levels has a direct impact on the portion of input energy one can hope to diffract into the desired response. This is referred to as the diffraction efficiency. According to scalar diffraction theory, the relationship between the maximum diffraction efficiency η and the number of masks N is:

$$\eta = \mathrm{sinc}^2\left(\frac{1}{N}\right) \qquad [13]$$

Thus, using two masks to create a 4-level phase element, as opposed to a single mask to create a binary-phase element, doubles the upper bound on the diffraction efficiency from 41% to 81%. The maximum diffraction efficiency for 8-level and 16-level phase elements is 95 and 99%, respectively.

However, this increase in diffraction efficiency incurs fabrication costs. After the first etch step, each subsequent exposure requires careful mask alignment. Alignment is accomplished using a standard tool from integrated circuit fabrication known as a mask aligner. The quality of a mask aligner is determined by its precision and the majority of commercially available aligners can register a mask and substrate to within less than a micrometer. This degree of alignment accuracy is sufficient to fabricate many useful multilevel diffractive elements.

To summarize, a multilevel surface relief profile uses a set of N masks to generate a surface profile that exhibits 2^N phase levels. The etch depth for the kth mask is $d_k = \lambda/2^k(n-1)$. In the next section, we describe a process for fabricating continuous DOE profiles in a single expose and etch cycle.

Continuous DOE Profiles Using Grayscale Lithography

Grayscale photolithography has become increasingly popular because of its flexibility and convenience in fabricating continuous diffractive profiles. Methods including directly writing partially exposed e-beam resist, resist reflow lithography, and using grayscale masks in a single step exposure. Here we discuss only the latter method because the processes are similar to those used in multiple mask fabrication. Thus, the same facilities can be used to produce both. In particular, we consider the application of high-energy-beam-sensitive (HEBS) glass to grayscale lithography and the fabrication of continuous profile DOEs.

The process for creating a diffractive element using HEBS glass is similar to fabricating a binary-phase element from a single mask. However, mask creation is markedly different. HEBS glass is a white crown glass within which a thin buried layer of silver ions is incorporated. The layer of silver ions is created via ion exchange in an aqueous acidic solution and essentially serves the same function as silver grains in photographic film. When the glass is exposed to a beam of electrons, the silver-containing crystal complexes are chemically reduced and the glass darkens. By controlling the electron dose, one controls the opacity of the glass to produce a grayscale pattern.

Although the attraction of single-step grayscale lithography is reduced costs in fabrication, the process incurs high calibration costs to insure quality pattern transfer through several steps. The parameter that has the greatest impact on the overall process is the desired size of the smallest diffractive feature. The minimum feature size affects the selection of the accelerating potential, beam current, addressing grid, and electron dose of the e-beam used to expose the HEBS glass.

The e-beam accelerating potential sets the resolution limit of the process and affects the sensitivity of the glass to the electron dose, the total charge per area incident upon the glass. For example, for a constant electron dose, increasing the accelerating potential increases the darkening response of the material. The large accelerating potential produces a large flux, which achieves the desired electron dose in a short period using energetic electrons. This, in turn, enhances the sensitivity of the glass. However, increasing the accelerating potential also increases scattering, which limits the minimum feature size achievable. Thus, low beam energies are generally used to write small features. But this, unfortunately, reduces the range of available optical densities because the material is less sensitive at low beam energies. To compensate for the reduced sensitivity, it is possible to increase the electron dose by increasing the exposure time. However, this undermines to some extent using a low energy beam to achieve small features due to broadening that results from long exposure times.

Once the accelerating potential for writing the mask is chosen, the beam current and addressing grid need to be specified. The beam current is related to the spot size and, therefore, affects directly the length of time to write the mask. High current levels produce large spots. Since write times can easily reach into hours for large structures, high current levels are desirable to reduce them.

The addressing grid represents the physical grid that the electron beam is translated over in order to complete the exposure. This grid should be chosen as finely as possible, however, too fine a grid can significantly increase the required dwell time on each grid point. In this respect, a compromise is required. High beam currents produce short dwell times, which may necessitate using a coarse address grid to insure the beam is capable of performing the exposure.

Once the process parameters are set, calibration between the HEBS transmission and the electron dose is necessary. Analyzing the optical density of written test structures by some experimental means and determining its functional dependence on the electron dose completes the calibration. This allows one to specify the electron doses required to achieve the features dictated by the device design.

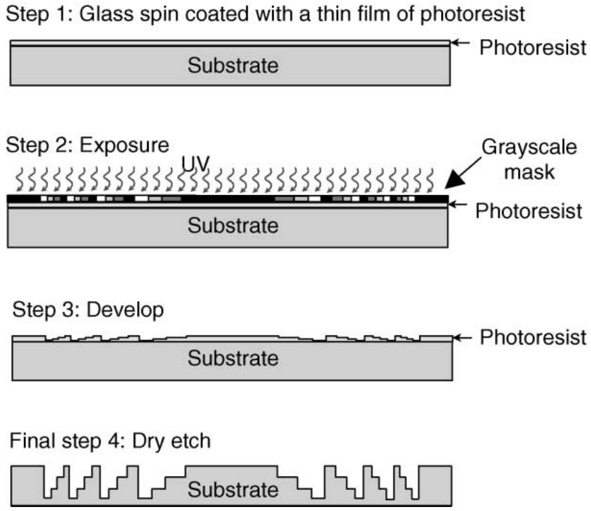

Figure 4 Fabrication of a continuous phase DOE using a grayscale mask.

If resolution requirements cannot be met through judicious selection of the accelerating potential, then some of the contrast in the mask can be sacrificed for an increase in available resolution. As we have mentioned, reducing the maximum optical density, or contrast, can alleviate scattering-induced broadening. It is possible, however, to compensate to some degree for the reduced contrast during the photolithography process, where UV exposure and develop times provide additional latitude in the pattern transfer process.

Figure 4 represents the fabrication of a continuous profile DOE using a grayscale mask. The substrate is spin-coated with a thin film of photoresist, placed in intimate contact with the grayscale mask, and UV illuminated to expose the resist film. The resist is developed and the pattern transferred into the substrate using either wet or dry etching techniques. Because the solubility of the resist is proportional to exposure, developing the photoresist produces a variable height profile.

Grayscale lithography is challenging because modern photoresists are also designed to be highly nonlinear yet linear response is required. Traditional photoresists are also designed to amplify mask contrast and produce the sharpest resist profile, thus they either dissolve away quickly and completely, or remain impervious to developer. The exposure threshold of traditional resists is, therefore, steep and considerable effort is required to

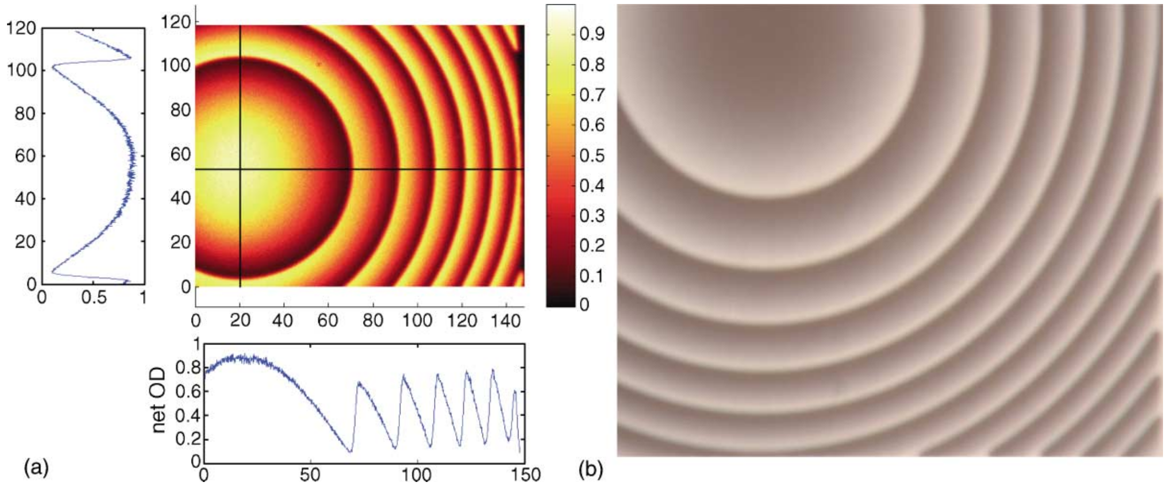

Figure 5 HEBS glass fabrication. (a) Data map of the optical density cross-section of an exposed HEBS grayscale mask and (b) its corresponding visible image.

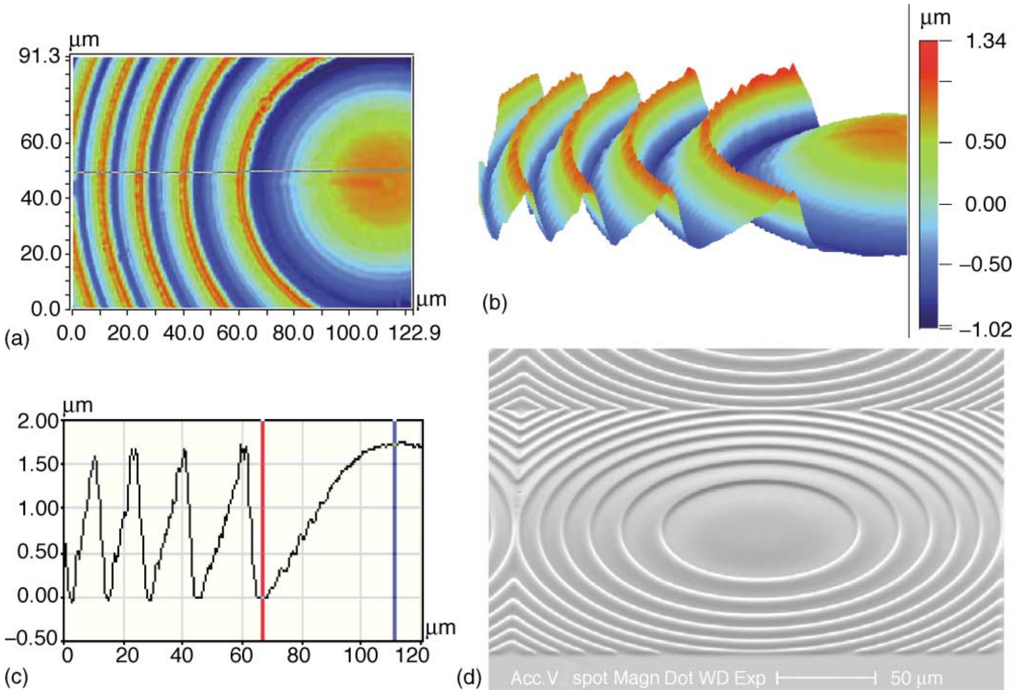

Figure 6 Characterization of a continuous phase diffractive lens. (a) Two-dimensional profile. (b) Three-dimensional profile. (c) Line scan of the two-dimensional plot. (d) Image of element from a scanning electron microscope.

insure that resist exposure is on the linear portion of the curve.

Although the resist exposure process is sensitive to all parameters, the behavior of a resist can be characterized using simulation models. The details of the photoresist models can be found elsewhere, but the simulations predict that the greatest utilization of the available resist occurs for short exposure times (seconds), and that these short exposures require long developing times (minutes). Further, the models predict that low mask contrast does not adversely affect the usable resist height. It is, therefore, not unreasonable to use a low contrast mask to achieve small features.

Through simulation it is possible to determine process parameters that maximize the linear response range of the photoresist. However, these parameters must be validated in the laboratory. The results presented in Figures 5 and 6 provide convincing evidence of the efficacy of these techniques.

Final Remarks

Advances in passive photonic devices have been made possible by leveraging fabrication processes from the electronics industry. Modifying standard photolithographic techniques and photoresist processes has produced an industrial capacity for fabricating diffractive elements. This capability will no doubt lead to future advances in telecommunications, signal processing, and imaging.

See also

Diffraction: Fresnel Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Holography, Techniques:** Overview.

Further Reading

Dascher W, Long P, Stein R, Wu C and Lee SH (1997) Cost-effective mass fabrication of multilevel diffractive optical elements by use of a single optical exposure with a gray-scale mask on high-energy beam-sensitive glass. *Applied Optics* 36: 4675–4680.

Dillon T, Balcha A, Murakowski JA and Prather DW (2004) Continuous-tone grayscale mask fabrication using high-energy-beam-sensitive (HEBS) glass. *Journal of Microlithography, Microfabrication, and Microsystems.*

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19: 297–306.

Goodman JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Kohler M (1998) Etching in microsystem technology. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*. Berlin: Wiley-VCH.

LeCompte M, Gao X and Prather DW (2001) Photoresist characterization and linearization procedure for the grayscale fabrication of diffractive optical elements. *Applied Optics* 40: 5921–5927.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America A* 52: 1123–1130.

Lohmann AW and Paris DP (1967) Binary Fraunhofer holograms generated by computer. *Applied Optical* 6: 1739–1748.

Mait JN (1995) Understanding diffractive optical design in the scalar domain. *Journal of the Optical Society of America A* 12: 2145–2158.

Stern MB and Rubico-Jay T (1994) Dry etching for coherent refractive microlens arrays. *Optical Engineering* 33: 3547–3551.

Swanson GJ (1989) Binary optics technology: the theory and design of multi-level diffractive optical elements, Technical Report 854, MIT.

Swanson GJ (1991) Binary optics technology: theoretical limits on the diffraction efficiency of multi-level diffractive optical elements, Technical Report 914, MIT.

Swanson GJ and Veldkamp WB, (1990) High-efficiency, multilevel diffractive optical elements. United States Patent 4,895,790.

Wang MR and Su H (1998) Laser direct-write gray-level mask and one-step etching for diffractive microlens fabrication. *Applied Optics* 37: 7568–7576.

Wu CK (1994) High Energy Beam Sensitive Glasses. U.S. Patent 5,285,517.

Wyrowski F (1991) Digital holography as part of diffractive optics. *Report on Progress in Physics* : 1481–1571.

Diffractive Laser Resonators

U D Zeitner, Fraunhofer Institute of Applied Optics and Precision Engineering, Jena, Germany

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

All lasers consist of three basic functional components: an active medium, an optical arrangement to pump the active medium, and a resonator, which we consider in detail here. The basic function of the resonator is to back-couple a portion of the emitted laser light into the active medium for further amplification. Simultaneously, in steady-state laser operation, the geometry of the resonator defines the spatial field distribution of the generated light field.

In the most simple, and most common, case the laser resonator consists of two planar or spherical mirrors. This configuration was proposed in 1958 and has been treated extensively in the literature. However, these configurations are only special cases of a more general resonator concept in which arbitrary mirror surface profiles are used. Historically, resonators consisting of surface profiled mirrors played only a small role in laser physics. This was due, not only to difficulties in their numerical design and analysis, but also to the lack of suitable fabrication methods for the mirrors.

However, since the 1990s, these difficulties have been overcome and in 1992 it was demonstrated for the first time experimentally a stable CO_2 laser resonator that has a user-defined mode shape generated by surface structured mirrors.

In the following we derive the basic design rules for stable and unstable diffractive laser resonators and discuss the advantages and disadvantages of the technology.

Modes in Laser Resonators

Before presenting the details of diffractive laser resonators we look first at some general properties of the electromagnetic field inside a laser resonator. As is typical in laser resonator theory, we use scalar and paraxial descriptions of the fields.

The complex amplitude $U(x, y, z)$ represents one scalar, monochromatic component of the electromagnetic field. To handle wave propagation inside the resonator, we use a general operator formalism in which the influence of each optical component on the wave $U(x, y, z)$ is represented by an operator. One of the most important operators is the so-called round trip operator $\mathbf{Z}$, which describes the effect of a round trip inside the resonator on the wave $U(x, y, z)$.

If $U(x, y, z)$ denotes the wave at the beginning of a resonator round trip, the resulting wave $\tilde{U}(x, y, z)$ at its end is

$$\tilde{U}(x, y, z) = \mathbf{Z}U(x, y, z) \qquad [1]$$

With reference to Figure 1, the round trip operator can be decomposed into a series of operators that represent the different elements in the resonator. For the common two-mirror resonator a round trip from mirror 1 to mirror 2, reflection at mirror 2, back propagation to mirror 1, and final reflection at

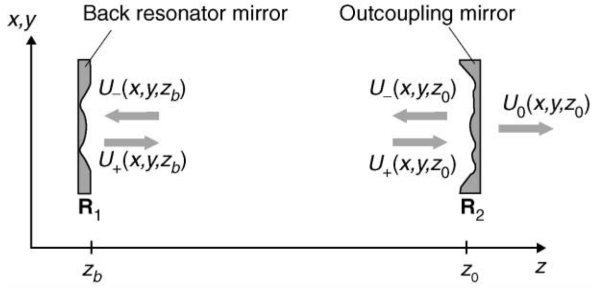

Figure 1 General diffractive laser resonator arrangement.

mirror 1, is represented by

$$\mathbf{Z} = \mathbf{R}_1 \mathbf{P}_{\Delta z} \mathbf{R}_2 \mathbf{P}_{\Delta z} \qquad [2]$$

The operators $\mathbf{R}_1$ and $\mathbf{R}_2$ represent the reflection at mirrors 1 and 2, respectively, and the propagation along the resonator length Δz is represented by the propagation operator $\mathbf{P}_{\Delta z.}$

In steady-state laser operation, the wave circulating inside the resonator reproduces itself, up to a constant factor γ, after a complete round trip:

$$\tilde{U}(x, y, z) = \gamma U(x, y, z) \qquad [3]$$

Together eqns [1] and [3] give the eigenvalue equation:

$$\gamma U(x, y, z) = \mathbf{Z} U(x, y, z) \qquad [4]$$

which, in general, has a set of solutions $U_p(x, y, z)$.

Thus, in steady-state, the waves $U_p(x, y, z)$ circulating inside the resonator are eigenfunctions of the round trip operator $\mathbf{Z}$ with the connected eigenvalue γ_p. The eigenfunctions of the resonator are called modes and the real-valued eigenvalues γ_p are connected with the round trip losses V, according to the equation:

$$V = 1 - |\gamma_p|^2 \qquad [5]$$

As we can see, the spatial properties of the modes are determined by the optical elements in the resonator. In the simple two-mirror resonator, the mirror profiles, which are related to the operators $\mathbf{R}_1$ and $\mathbf{R}_2$, can be used to specify the shape of the modes.

To derive the rules for resonator design we must look closer at the orthogonality of the resonator modes. The normalized eigenfunctions of the round trip operator $\mathbf{Z}$ do not obey the usual power-orthogonality relationship:

$$\iint U_m(x, y, z) U_n^*(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = \delta_{mn} \qquad [6]$$

but instead satisfy the biorthogonality relationship:

$$\iint U_{m;+}(x, y, z) U_{n;-}(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = \delta_{mn} \qquad [7]$$

The δ_{mn} is the so-called Kronecker symbol and $*$ denotes the complex conjugate. In eqn [7], $U_{m;+}(x, y, z)$ is the mode in plane z traveling in the $+z$ direction and $U_{m;-}(x, y, z)$ is the mode in the same plane z propagating in the opposite direction. The physical reason for the difference between eqns [6] and [7] are diffractive mode losses at limiting apertures. However, if a mode $U_{m;+}(x, y, z)$ travels without losses through the whole resonator, a power normalization:

$$\iint U_m(x, y, z) U_m^*(x, y, z)\, \mathrm{d}x\, \mathrm{d}y = 1 \qquad [8]$$

is valid in all planes z. Equations [7] and [8] will be used in the next section to derive the design strategy for resonators with predefined fundamental modes.

If the mirrors, either planar or spherical, are assumed to be infinite in extent, it is possible to derive analytic solutions for the modes. This yields either the Gauss–Hermite or Gauss–Laguerre modes, which are orthogonal in the sense of eqn [6]. For resonators whose mirror shape differs from either a plane or sphere, the analytical calculation of its eigenfunctions or resonator modes is difficult and in most cases even impossible.

However, in such cases the analysis can be performed numerically. One of the most common methods is the Fox-Li-algorithm which uses the physical propagation of the modes through the resonator. The algorithm begins with a random field distribution in some plane of the resonator. This field is then numerically propagated through the resonator until a round trip is completed. The effect of the resonator optical elements must be accounted for in this procedure. At the end of the resonator round trip, the starting field will have changed due to diffraction and serve as a new starting wave for a succeeding round trip calculation. This procedure is repeated until a static field distribution is obtained, which is then an eigenfunction of the resonator.

Design Principles

The condition in eqn [7] on the biorthogonality of resonator modes, together with the eqn [8] valid for lossless modes, can be used to develop principles to design resonators with an arbitrary fundamental mode shape. Consider a stable laser resonator characterized by negligible diffraction loss of the fundamental mode. Following the discussion above, its lossless modes must simultaneously satisfy eqns [7] and [8]. A comparison of both equations yields:

$$U_+(x, y, z) = U_-^*(x, y, z) \qquad [9]$$

For reasons of simplicity we neglect here, and in the following, the subscript m for the mode order if only one mode is considered. Because this equation must hold in all planes z, it is satisfied in the mirror resonator planes z_0 and z_b. Further, by definition of the operators $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$U_+(x, y, z_b) = \mathbf{R}_1 U_-(x, y, z_b) \qquad [10]$$

$$U_-(x, y, z_0) = \mathbf{R}_2 U_+(x, y, z_0) \qquad [11]$$

The combination of eqns [10] and [11] with eqn [9] yields the following expressions for the mirror operators:

$$\mathbf{R}_1 = \frac{U_+(x, y, z_b)}{U_+^*(x, y, z_b)} \qquad [12]$$

$$\mathbf{R}_2 = \frac{U_-(x, y, z_0)}{U_-^*(x, y, z_0)} \qquad [13]$$

Equations [12] and [13] imply that lossless modes are phase-conjugated on reflection at the resonator mirrors. Because we have not made any assumptions on the shape of the mode, these equations represent a general property of lossless resonator modes. Furthermore, they provide a general design rule for the construction of stable laser resonators with predefined fundamental mode shape.

If one considers limiting apertures inside the resonator, an equivalent statement for lossy modes can be made by making assumptions about phase conjugation in a lossy system and showing that they lead to a contradiction. Without loss of generality, we assume an aperture on the outcoupling resonator mirror that clips the mode. We now consider the wave $U_-(x, y, z_0)$ after reflection at this mirror and assume this wave is still phase-conjugated after propagation through the resonator during its reflection at the back resonator mirror according to:

$$U_+(x, y, z_b) = U_-^*(x, y, z_b) \qquad [14]$$

If now the wave $U_+(x, y, z_b)$ is back propagated to the outcoupling mirror, we end up with the wave $U_+(x, y, z_0)$ which must, because of the phase conjugation in z_b, have the same amplitude as the wave $U_-(x, y, z_0)$ we started with. As a result, because both waves are connected to each other via the reflection operator of the outcoupling mirror by eqn [11], $\mathbf{R}_2$ cannot influence the amplitude distribution which, however, is a contradiction to the assumption of a limiting aperture. Thus, we finally can state that lossy modes do not have the property to be phase conjugated during their reflection at both mirrors. As a consequence, the design of diffractive mirrors for unstable laser resonators, in general characterized by energy outcoupling via diffractive losses, is more complex than simply calculating a phase-conjugating mirror structure for stable laser resonators.

Up to now, design methods described in the literature for unstable laser resonators with user-defined output beam shapes have all been based on an iterative calculation of the mirror reflection function. Because a detailed description of these methods is rather complex, we restrict ourselves to a sketch of the principal ideas behind them.

Without loss of generality, we consider an unstable laser resonator from which the light is coupled out by a mirror with spatially varying transmission and reflection. The second mirror is fully reflective within its aperture. The particular application defines only few design constraints which are in most cases the light wavelength, the dimensions of the resonator, and the shape of the outcoupled beam $U_0(x, y, z_0)$. As a consequence of the outcoupling principle, the shape of the mode inside the resonator is not finally defined by the outcoupled beam. In regions where the outcoupling mirror is reflective and not transmissive, the mode amplitude can be chosen freely. Also the amplitude of the mode in the plane of the back resonator mirror is not fixed but influenced by the amplitude and phase of the mode in the plane of the outcoupling mirror.

The general starting situation for an unstable resonator design is sketched in Figure 2. The design procedure must yield the complex amplitudes of four waves: the mode directly in front of the outcoupling mirror $U_-(x, y, z_0)$, the mode after reflection at the outcoupling mirror $U_+(x, y, z_0)$, and the complex amplitudes of the forward and reverse modes in the plane of the back resonator mirror $U_+(x, y, z_b)$ and $U_-(x, y, z_b)$, respectively. Since we assume no losses at

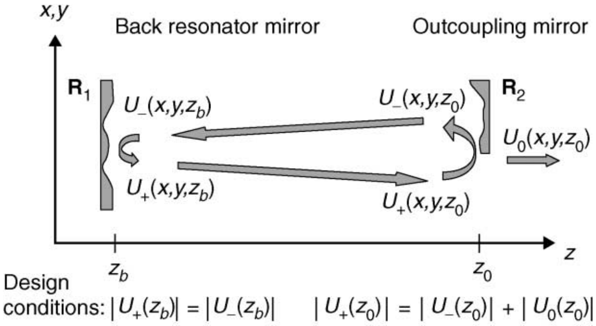

Figure 2 Starting situation for design of an unstable diffractive laser resonator.

this mirror, the amplitudes must satisfy the condition:

$$|U_+(x,y,z_b)| = |U_-(x,y,z_b)| \tag{15}$$

If we further assume no absorption at the outcoupling mirror, the amplitudes of the three waves in z_0 are related according to:

$$|U_-(x,y,z_0)| + |U_0(x,y,z_0)| = |U_+(x,y,z_0)| \tag{16}$$

To design the resonator one starts with the choice of initial intensity distributions for the wavefields $U_-(x,y,z_0)$ and $U_+(x,y,z_b)$. In order to obtain a good filling of the available mode volume, a homogeneous intensity across the mirror apertures is recommendable. To avoid strong diffraction effects at the mirror edges, the corresponding intensity distributions should be smoothed out. Together with the desired outcoupled beam amplitude $|U_0(x,y,z_0)|$ one can calculate the two remaining amplitudes for the wave fields $U_-(x,y,z_b)$ and $U_+(x,y,z_0)$ according to eqns [15] and [16].

An iterative algorithm, such as the well known Gerchberg–Saxton algorithm, can be used to find the phases of the beams $U_-(x,y,z_0)$ and $U_+(x,y,z_0)$ that insure the wavefields have almost equal amplitudes once they are propagated or back-propagated, respectively, to the plane z_b. In order to find wavefields that satisfy eqns [15] and [16], two successive iteration runs with different constraints on the complex amplitudes have to be performed. During the first iteration run no change in the initial amplitudes in plane z_0 are allowed. The iteration only look for phases for $U_-(x,y,z_0)$ and $U_+(x,y,z_0)$ which result in the amplitude $|U_+(x,y,z_b)|$ once they are propagated and back-propagated, respectively, to plane z_b. Usually the results of a Gerchberg–Saxton algorithm, with such constraints, are wavefields $U_+(x,y,z_b)$ and $U_-(x,y,z_b)$ with similar but not equal amplitude. Thus, in order to determine wavefields that satisfy eqn [15], a second iteration run with the Gerchberg–Saxton algorithm has to be added in which slight variations in wavefield amplitudes also in plane z_0 are allowed. Because the application of the constraints to ensure convergence of the iteration is rather complex, we do not go into details here but refer the reader to the Further Reading at the end of this article.

Upon completion of the iterations, the mirror reflectance operators $\mathbf{R}_1$ and $\mathbf{R}_2$ are calculated according to:

$$\mathbf{R}_1 = \frac{U_+(x,y,z_b)}{U_-(x,y,z_b)} \tag{17}$$

$$\mathbf{R}_2 = \frac{U_-(x,y,z_0)}{U_+(x,y,z_0)} \tag{18}$$

The design method described so far yields a mathematical expression for the resonator mirror reflection operator. Practical realization of the resonator requires this expression be converted into a description of a physical property of the mirror. In most practical cases this is the surface profile. Since the design method considers only phase-only functions, the physical description of the reflection operators are:

$$\begin{aligned}\mathbf{R}_1 &= \exp[-\mathrm{i}\phi_1(x,y)] \text{ and}\\ \mathbf{R}_2 &= \exp[-i\phi_2(x,y)]\end{aligned} \tag{19}$$

Using the so-called thin-element-approximation, the surface profiles $h_1(x,y)$ and $h_2(x,y)$ of the two mirrors follow directly from the phases $\phi_1(x,y)$ and $\phi_2(x,y)$:

$$\begin{aligned}h_1(x,y) &= -\frac{\lambda}{4\pi}\phi_1(x,y) \text{ and}\\ h_2(x,y) &= -\frac{\lambda}{4\pi}\phi_2(x,y)\end{aligned} \tag{20}$$

with λ being the laser wavelength. The expressions in eqn [20] contain a factor of 1/2 due to reflection. These surface profiles are commonly fabricated by microlithographic techniques. For stable laser resonators, where the mirrors perform a static phase conjugation, the mirror profile is a surface of equal phase of the fundamental mode.

Use of General Resonator Modes

Applications of diffractive laser resonators can be classified into beam shaping for a particular task external to the laser and resonator optimization via modifying the eigen-space of the round trip operator.

Internal beam shaping for external tasks is relevant to applications where it is advantageous to use a fixed beam profile other than a Gaussian, as is the case for instance in materials processing. Shaping the beam internally, avoids the energy losses usually associated with external beam shaping techniques. However, lasers that have a nonGaussian outcoupled beam profile are often less flexible in their general applicability.

However, there is considerable potential for diffractive laser resonators to enhance resonator characteristics and thereby improve the lasing operation itself. The conventional Gaussian-shaped fundamental mode in customary laser resonators varies considerably in intensity from its central maximum to its outer wings. The transfer of energy from the pumped active medium into the propagating mode is dependent upon the local transverse intensity and is optimal for homogeneous illumination. As a result, the efficiency with which energy is transferred into the fundamental mode can be increased by using diffractive resonator mirrors to match the shape of the lateral fundamental mode to that of the cross-section of the active medium.

An important resonator property, especially in high power laser systems, is the difference of round trip losses V between the fundamental mode and the higher, usually undesired, laser modes. This difference is referred to as modal discrimination. With no limiting apertures, the wavefronts of all Gauss–Hermite or Gauss–Laguerre modes of common laser resonators are phase conjugated upon reflection at the spherical mirrors. As we have shown above, there is a tight connection between the phase conjugation property of resonator modes and their property of having no round trip losses. Thus, the phase conjugation property of Gauss–Hermite and Gauss–Laguerre modes directly implies small round trip losses. In fact, in resonators without apertures they all have an eigenvalue $\gamma_p = 1$ and according to eqn [5] the round trip losses are zero. High-order modes, whose diameters are large, can be suppressed by decreasing the size of the mirror aperture, however, because the difference in mode diameter between modes of neighboring order m is small, the achievable discrimination is limited to a maximum ratio $V_{m+1}/V_m \approx 10$. The structured surface profile of the mirrors in a diffractive laser resonator whose fundamental mode is nonGaussian is tightly adapted to the phase of the design mode. The wavefront of the higher-order modes in these resonators often do not fit the mirror profile and it can be observed that their round trip losses are significantly higher. Thus, diffractive laser resonators usually show an improved modal discrimination compared to resonators with spherical mirrors. This effect can be further improved by inserting into the resonator additional phase plates with diffuser-like properties.

Shaping the surface profile of resonator mirrors or insertion of additional phase plates significantly improves the flexibility for the resonator design. As is demonstrated by the examples below, there are resonator configurations in which high modal discrimination or adaptation of the mode volume to the cross-section of the active medium are combined with a Gaussian shape of the outcoupled beam.

In the design procedure described so far, the resonators were treated as passive optical systems. The propagation of the mode through the resonator has not been influenced by the active medium. However, in real laser systems, this simplification is often not applicable. Due to the pumping process and the connected heat insertion, the refractive index distribution inside the active medium is not homogeneous, which affects the propagation of the modes. Furthermore, changes or fluctuations in the pump energy and cooling mechanisms introduce temporal changes in the optical function associated with the refractive index of the active medium. Further, although static or averaged, optical power of the active medium are included in the round trip operator and calculation of the mirror profiles, experiments with lasers that have diffractive resonator mirrors indicate that the temporal fluctuations can strongly influence the shape and round trip losses of the fundamental resonator mode. In general, diffractive laser resonators are more sensitive to deviations from design than conventional resonators.

Examples

We consider as an example the adaptation of fundamental mode shape to the cross-section of the active medium. In solid state lasers, such as Nd:YAG lasers, the active medium is often a crystal rod with a circular cross-section. Instead of using this sharp-edged cross-section shape as intensity distribution for the desired fundamental mode, a smoothed version, which is approximated by a super-Gaussian intensity distribution in the center plane of the rod, is chosen to be the fundamental mode. The super-Gaussian intensity distribution has a more desirable propagation characteristic than a sharp circular one. However, even these modes do not exhibit the advantageous property of propagation-invariant beam shape characteristic of Gaussian beams.

In our first example, we consider the design of a stable diffractive laser resonator with a super-Gaussian fundamental mode shape inside the active medium and a Gaussian beam coupled out from the resonator. The resonator configuration is based upon a beam shaping arrangement for a Gauss to super-Gauss beam transformation. Its optical setup is sketched in Figure 3. In the plane of the outcoupling

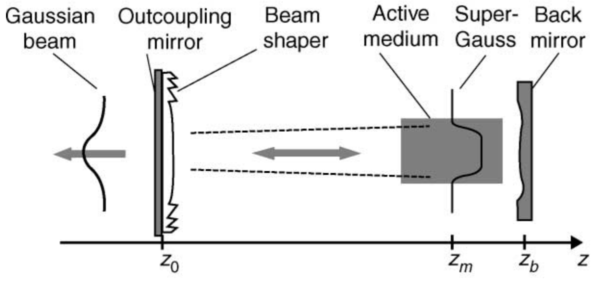

Figure 3 Setup for the Gauss-to-super-Gauss resonator.

mirror z_0, the fundamental mode has a Gaussian amplitude $U_g(x, y, z_0)$. A phase element, which is located directly in front of the outcoupling mirror, shapes the mode amplitude into a super-Gaussian one $U_{sg}(x, y, z_m)$ in the center plane z_m of the active medium. For this special beam shaping problem it is possible to calculate a phase plate surface profile that transforms the beams without energy losses. The corresponding calculation procedure is described in the Further Reading section at the end of this article.

The mirror profiles have to be realized to satisfy the previously derived phase conjugation property for the described beam. Since the Gaussian beam has its minimum waist at the outcoupling mirror, its phase is planar in that plane. Thus, to insure phase conjugation a plane mirror is all that is required in front of the beam shaping phase plate. In the plane of the back resonator mirror the beam has a modulated wavefront due to its transformation from a Gaussian to a super-Gaussian amplitude distribution. Thus, a surface modulated mirror profile is required to phase conjugate this beam during reflection. Figures 4a and b represent the phase profile of the beam shaping element and the back resonator mirror, respectively, for the following set of parameters: laser wavelength $\lambda = 1064\,\text{nm}$, radius of the Gaussian and super-Gaussian beams in the plane z_0 and z_m, respectively, $w_g = 1\,\text{mm}$ and $w_{sg} = 1\,\text{mm}$, order of the super-Gauss $n = 10$, distance $z_m - z_0 = 200\,\text{mm}$, and resonator length $z_b - z_0 = 250\,\text{mm}$. The amplitude distribution for the fundamental mode in this resonator simulated with the Fox–Li-algorithm is shown in Figure 4c for the plane of the outcoupling mirror and in Figure 4d for the center of the active medium.

In our second example, the flexibility of diffractive laser resonators is demonstrated by a resonator with a user-defined far-field intensity distribution. In our case the far-field intensity $|U_f(x, y, z_f)|^2$ is shaped like the Greek letter lambda, λ. Because we are concerned only with the intensity of the wavefield, the phase $\phi_f(x, y, z_f)$ of the far-field $U_f(x, y, z_f)$ can be chosen freely. As shown in the following, this phase freedom can be used to improve the interaction of the mode with the active medium, as was the case in our first design example. Again, the mode shape in the center of the active medium has a super-Gaussian intensity distribution, which insures high extraction efficiency. We used an iterative Fourier-transform algorithm to calculate a far-field phase that yields a nearly super-Gaussian intensity distribution in the outcoupling plane of the resonator. The resulting wave:

$$U_0(x, y, z_0) = a_0(x, y, z_0)\exp[\mathrm{i}\varphi_0(x, y, z_0)] \qquad [21]$$

is generated directly as the laser output. The amplitude $a_0(x, y, z_0)$ is generated as a mode inside the cavity using a surface-modulated back resonator mirror. This performs a static phase conjugation of the desired mode in the mirror plane. The phase $\varphi_0(x, y, z_0)$ is provided by an additional element on the outside of the outcoupling mirror. A sketch of the resonator configuration is shown in Figure 5.

For the experimental realization of this example the thermal index gradient of the active medium was included into the resonator design and the two diffractive elements were fabricated on a quartz substrate using electron beam lithography and ion-beam etching. The measured mode shape in the plane of the outcoupling mirror and the far-field

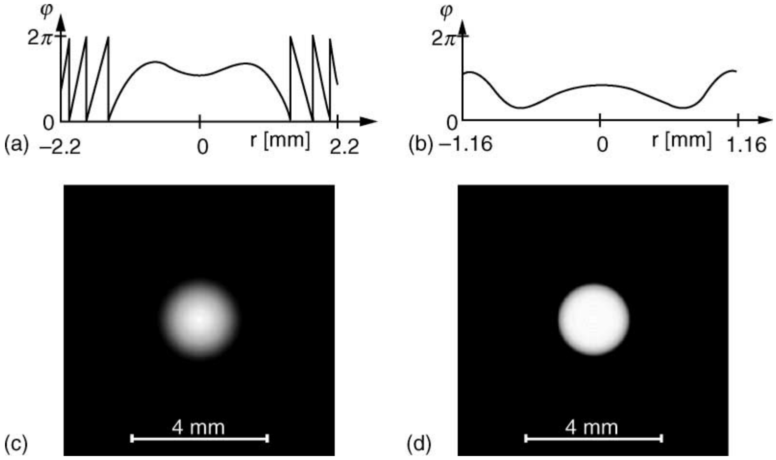

Figure 4 Gauss-to-super-Gauss resonator: (a) phase of beam shaping element; (b) phase of back resonator mirror, simulated amplitude distributions for the fundamental resonator mode; (c) in the plane of the outcoupling mirror; and (d) in the center of the active medium.

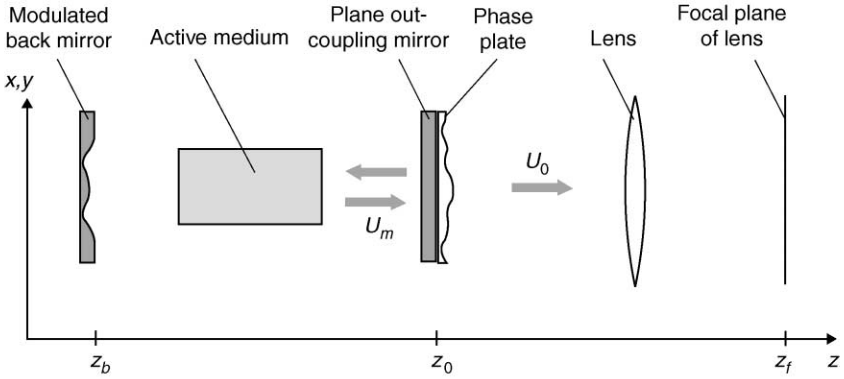

Figure 5 Configuration for resonator internal generation of a user defined far-field intensity distribution. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2001 IEEE.)

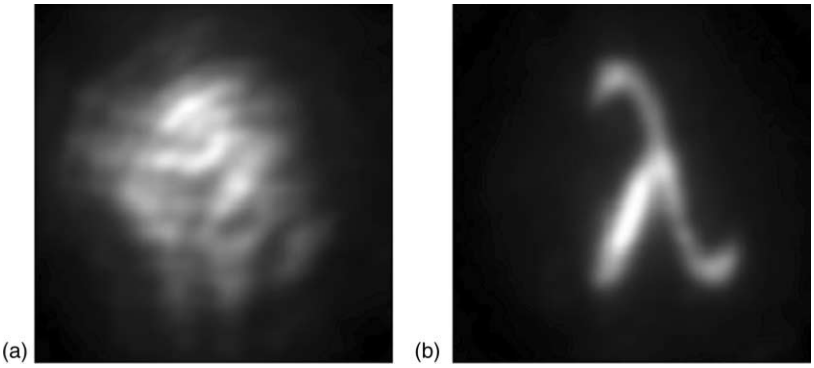

Figure 6 Measured mode intensities for the resonator of Figure 5. (a) In the plane of the outcoupling mirror; and (b) in the far-field of the resonator. (With permission from Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109. © 2000 IEEE.)

intensity distribution are shown in Figures 6a and b, respectively. The mode in the outcoupling plane of the resonator fills an area of about 2 mm diameter with nearly homogeneous intensity. The far-field shows the desired λ-shaped intensity distribution. The variations in intensity are due to the nonlinear response of the active medium. As mentioned above, the response can be improved if these effects are included in the design of the resonator mirrors.

Conclusion

Diffractive laser resonators are a generalization of common spherical mirror resonators. The use of arbitrary resonator mirror profiles and additional internal phase plates offer a huge flexibility for the generation of user defined mode shapes and optimization of resonator properties like extraction efficiency and modal discrimination. The drawback of the use of structured mirrors is an increased adjustment effort, because the diffractive mirrors have more degrees of freedom to be aligned. Furthermore, diffractive resonators tend to be more sensitive against changes of the optical properties of the active medium than spherical mirror resonators. Such changes may occur due to temporal fluctuations in pump power or cooling power.

Most promising applications for diffractive resonators are:

- lasers for materials processing which require a fixed intensity pattern for the sample treatment and a fixed laser power;
- high efficient homogeneous laser illumination, for example, for data projection or sensing applications; and
- suppression of higher order modes and maintaining single-mode operation for high brightness lasers.

Some example applications and solutions have been presented here which cannot be achieved by conventional resonator optics. For stable laser resonators, the phase-conjugation property of the modes is used for a straightforward calculation of the required diffractive mirror profiles. Unstable resonator geometries with diffractive outcoupling require more complex iterative design procedures. However, especially for high-power and high-gain lasers, the latter sometimes have more advantageous properties, such as better discrimination of undesired higher-order modes and more simple mirror cooling options.

See also

Lasers: Carbon Dioxide Lasers. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beam splitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Belanger PA, Lachance RL and Pare C (1992) Super-Gaussian output from a CO_2 laser by using a graded-phase mirror resonator. *Optics Letters* 17: 739–741.

Fox A and Li T (1961) Resonant modes in a maser interferometer. *Bell Systems Technical Journal* 40: 453–488.

Gerchberg R and Saxton W (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35: 237–246.

Herzig H (1997) *Micro Optics*. London: Taylor & Francis.

Leger JR (1997) Diffractive laser resonators. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, pp. 165–188. Berlin: Akademie Verlag.

Leger JR, Chen D and Dai K (1994) High modal discrimination in a Nd:YAG laser resonator with internal phase gratings. *Optics Letters* 19: 1976–1978.

Makki S and Leger JR (2001) Mode shaping of a graded-reflectivity-mirror unstable resonator with an intra-cavity phase element. *IEEE Journal of Quantum Electronics* 37: 80–86.

Siegman AE (1986) *Lasers*. Mill Valley, CA: Univ. Science Books.

Zeitner UD and Wyrowski F (2001) Design of unstable laser resonators with user-defined mode shape. *IEEE Journal of Quantum Electronics* 37: 1594–1599.

Zeitner UD, Wyrowski F and Zellmer H (2000) External design freedom for optimization of resonator originated beam shaping. *IEEE Journal of Quantum Electronics* 36: 1105–1109.

Diffractives in Animals

A R Parker, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Structural coloration involves the selective reflectance of incident light by the physical nature of a structure. Although the color effects often appear considerably brighter than those of pigments, structural colors often result from completely transparent materials. In addition, animal structures can be designed to provide the opposite effect to a light display: they can be antireflectors, causing 'all' of the incident light to penetrate a surface (like glass windows, smooth surfaces cause some degree of reflection).

The study of natural structural colors began in the seventeenth century, when Hooke and Newton correctly explained the structural colors of silverfish (Insecta) and peacock feathers (Figure 1), respectively. Nevertheless, accurate, detailed studies of the mechanisms of structural colors did not begin until the introduction of the electron microscope in 1942.

Structural colors generally may be formed by one of three mechanisms: thin-film reflectors, diffraction gratings, or structures causing scattering of light waves. This is a simplistic characterization resulting from the application of geometric optics to the reflective elements. Some structures, however, rather fall between the above categories, such as photonic crystals (Figure 2) – these become apparent when quantum optics is considered. In some cases, this has led to confusion over the reflector type. For example, the reflectors in some scarab beetles have been categorized by different authors as multilayer reflectors, three-dimensional diffraction gratings, photonic crystals, and liquid crystal displays. Perhaps all are correct! Since the above categories are academic, I place individual cases of structural colors in their most appropriate, not unequivocal, category. Also, in this article I will employ geometric optical treatments only.

The array of structural colors found in animals today results from millions of years of evolution. Structures that produce metallic colors have also been identified in extinct animals. Confirmation of this

Figure 1 Newton's subject.

fact, from ultrastructural examination of exceptionally well-preserved fossils, such as those from the Burgess Shale (Middle Cambrian, British Columbia), 515 million years old, permits the study of the role of light in ecosystems throughout geological time, and consequently its role in evolution. In some cases, such as for ostracod crustaceans, diffraction gratings have been found to increase in their optical sophistication throughout 350 million years of evolution.

Multilayer Reflectors

Light may be strongly reflected by constructive interference between reflections from the different interfaces of a stack of thin films (of actual thickness *d*) of alternately high and low refractive index (*n*). For this to occur, reflections from successive interfaces must emerge with the same phase and this is achieved when the so-called 'Bragg Condition' is fulfilled. The optical path difference between the light reflected from successive interfaces is an integral number of wavelengths and is expressed in the equation:

$$2nd \cos \Theta = (m + 1/2)\lambda \qquad [1]$$

from which it can be seen that the effect varies with angle of incidence (Θ, measured to the surface normal), wavelength (λ), and the optical thickness of the layers (*nd*). There is a phase change of half a wavelength in waves reflected from every low to high refractive index interface only. The optimal narrowband reflection condition is therefore achieved where the optical thickness (*nd*) of every layer in the stack is a quarter of a wavelength. In a multilayer consisting of a large number of layers with a small variation of index the process is more selective than one with a smaller number of layers with a large difference of index. The former therefore gives rise to more saturated colors corresponding to a narrow spectral bandwidth and these colors therefore vary more with a change of angle of incidence. Both conditions can be found in animals – different colored effects are appropriate for different functions under different conditions. For an oblique angle of incidence, the wavelength of light that interferes constructively will be shorter than that for light at normal incidence. Therefore, as the angle of the incident light changes, the observed color also changes (Figure 3).

Single layer reflectors are found in nature, where light is reflected, and interferes, from the upper and lower boundaries (Figure 4). A difference in the thickness of the layer provides a change in the color observed from unidirectional polychromatic light. The wings of some houseflies act as a single thin film and appear different colors as a result of this phenomenon. A single quarter-wavelength film of guanine in cytoplasm, for example, reflects about 8% of the incident light. However, in a multilayer reflector with ten or more high index layers, reflection efficiencies can reach 100%. Thus, animals possessing such reflectors may appear highly metallic.

The reflectance of the multilayer system increases very rapidly with increasing number of layers. If the dimensions of the system deviate from the quarter-wave condition (i.e., *nd* is not equal for all layers), then the reflector is known as 'nonideal' in a theoretical sense (may be 'ideal' for some natural situations). 'Nonideal' reflectors have a reduced proportional reflectance (not always a significant reduction) for a given number of layers and this

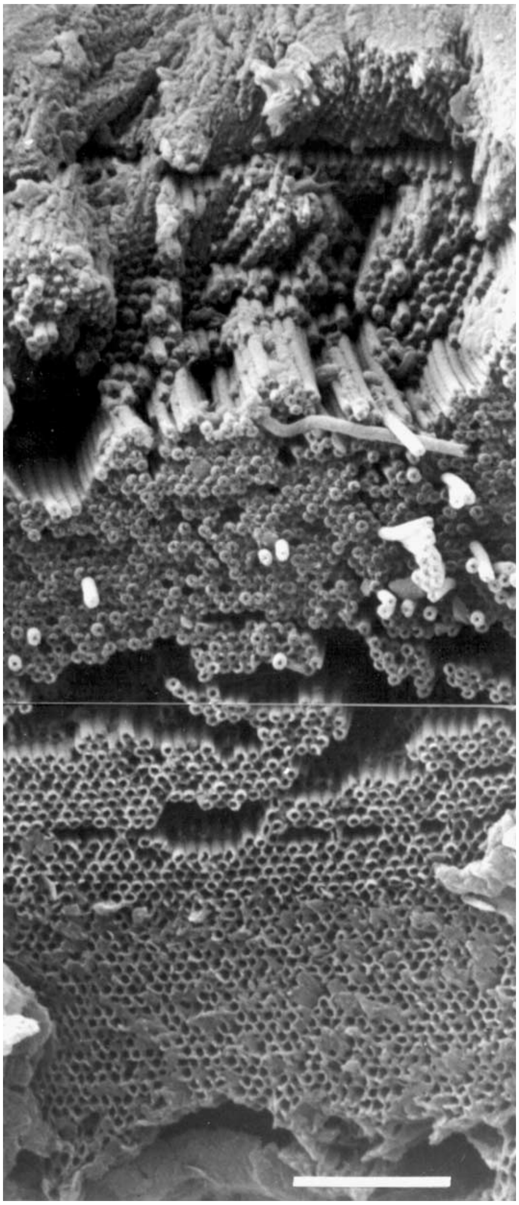

Figure 2 Scanning electron micrograph of a cross-section of the wall of a cylindrical spine of the sea mouse *Aphrodita* sp. (Polychaeta). The wall is composed of small cylinders with varying internal diameters (increasing with depth in the stack), arranged in a hexagonal array, that form a photonic crystal fiber. The opal-type photonic crystal – hexagonally packed nano-spheres (250 nm diameter) – are also found in animals (e.g. weevils). Scale bar represents 8 μm.

reflectance has a narrower bandwidth. A narrow bandwidth, less conspicuous reflection is sometimes selected for in animals, this will be discussed later in this article. Multilayer reflectors polarize light incident at Brewster's angles. This is about 54° for a quarter-wave stack of guanine and cytoplasm. At very oblique angles, all wavelengths are strongly reflected.

Multilayer reflectors are common in animals. They are usually extra-cellular, produced by periodic secretion and deposition, but sometimes occur within cells. Guanine ($n = 1.83$) is a common component in invertebrate reflectors because it is one of very few biological materials with a high refractive index and is readily available to most invertebrates as a nitrogenous metabolite. However, arthropods, including insects, crustaceans, and spiders, have largely ignored guanine in favor of pteridines. Also surprising is that the reflector material of closely related species, e.g., the molluscs *Pecten* (scallop) and *Cardium* (cockle), may differ.

Multilayers produce effects in beetle cuticle from highly metallic colors ('ideal' system) to rather dull greens ('nonideal' system in combination with scattering; Figure 5), and colors from the wings of many butterflies. Often in butterflies, layers of chitin ($n =$ about 1.56) are supported by vertical vanes of the scales. Air ($n = 1.0$) fills in the spaces and provides the alternate layers of the system. Under white light at normal incidence, the blue color of the butterfly *Arhopala micale* turns to green when the air is replaced by acetone. This is due to an increase in the refractive index of the low n layer (the actual thickness of this layer remains the same but the optical thickness increases) that effectively makes the system less 'ideal'. If the refractive indices of both layers were made equal, there would be no internal interference. A layer of melanin often underlies the reflector and intensifies the metallic colored effect by absorbing the transmitted portion of incident light. For example, in beetles, the elytra of *Anopognathus parvulus* appears metallic gold, green, or yellow in reflected light, and diffuse brown in transmitted light. Individual butterfly scales have been examined in detail to reveal a number of variations of quarter-wave stacks, sometimes in combination with other optical structures, to provide a range of colored effects.

The crustaceans *Limnadia* (Conchostraca), *Tanais tennicornis* (Tanaidacea), *Ovalipes molleri* (Decapoda), and the males of *Sapphirina* (Copepoda) all bear multilayer reflectors in their cuticles, in different forms (Figure 6). In contrast to the usual continuous thin layers, male sapphirinids have ten to fourteen layers of interconnecting hexagonal platelets within the epidermal cells of the dorsal integument. The reflector of *O. molleri* comprises layers that are corrugated and also slightly out of phase (Figure 7). At close to normal incidence this structure reflects red light, but at an angle of about 45° blue light is reflected. The corrugation, however, functions to broaden the reflectance band, at the expense of reducing the intensity of reflection.

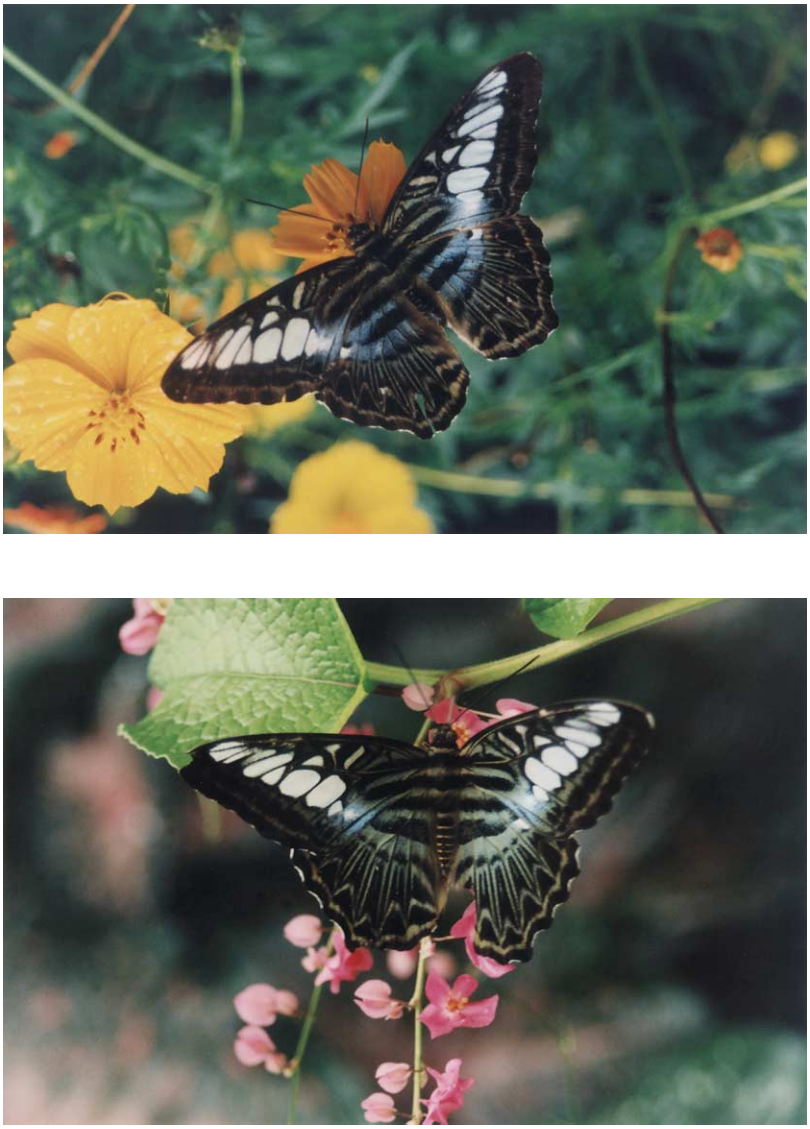

Figure 3 Butterfly (*Parthenos sylvia* (Nymphalidae) from Malaysia), showing the change in color with angle of wings. This animal may have evolved a temporal signal rather than a static one.

A broadband wavelength-independent reflectance, appearing silver or mirror-like to the human eye, can be achieved in a multilayer stack in at least three ways in invertebrates (Figure 8). These are (a) a composite of regular multilayer stacks each tuned to a specific wavelength, (b) a stack with systematically changing optical thicknesses with depth in the structure, termed a 'chirped' stack, and (c) a disordered arrangement of layer thicknesses about a mean value, termed a 'chaotic' stack (Figure 8). The nauplius eye of the copepod *Macrocyclops* (Crustacea) has regularly arranged platelets ~100 nm thick in stacks of 20–60, achieving the first condition. Silver beetles and the silver and gold chrysalis of butterflies in the genera *Euoplea* and *Amauris*, owe their reflection to the second condition. The mirror-like reflectors in the scallop *Pecten* eye comprise alternating layers of cytoplasm ($n = 1.34$) and guanine crystals ($n = 1.83$) and approximate an 'ideal' quarter-wave system in the same manner as within fish skin using the third mechanism. The ommatidia of the superposition compound eyes of *Astacus* (Crustacea) are lined with a multilayer of isoxanthopterin (a pteridine) crystals, which again fall into the third category. Multilayer reflectors can also be found in the eyes of certain spiders, butterflies, and possibly flies, where they assist vision, as discussed further below.

Squid and cuttlefish, for example, possess mirror-like reflectors in photophores (light organs) and

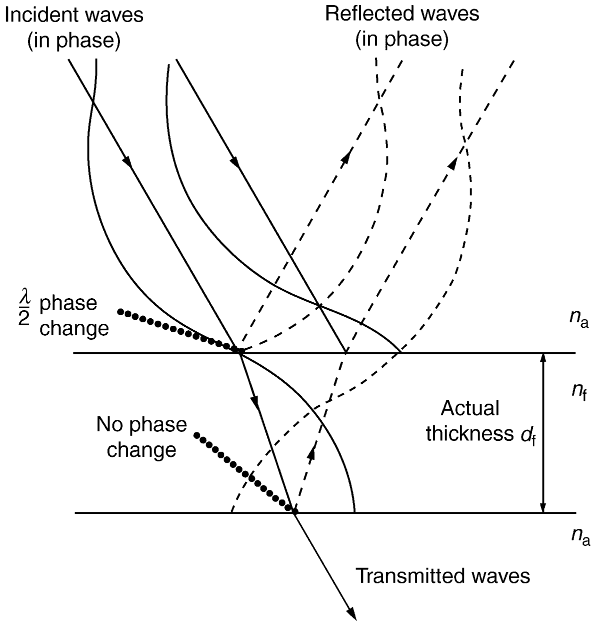

Figure 4 Schematic diagram of thin-film reflection. The direction of wave (straight line) and profile of electric (or magnetic) component are illustrated. Incident waves are indicated by a solid line, reflected waves by broken lines. Refraction occurs at each media interface. The refractive index of the film (n_f) is greater than the refractive index of the surrounding medium (n_a). Constructive interference of the reflected waves is occurring. As the angle of incidence changes, different wavelengths constructively interfere. At normal incidence constructive interference occurs where $n_f \times d_f = \lambda/4$.

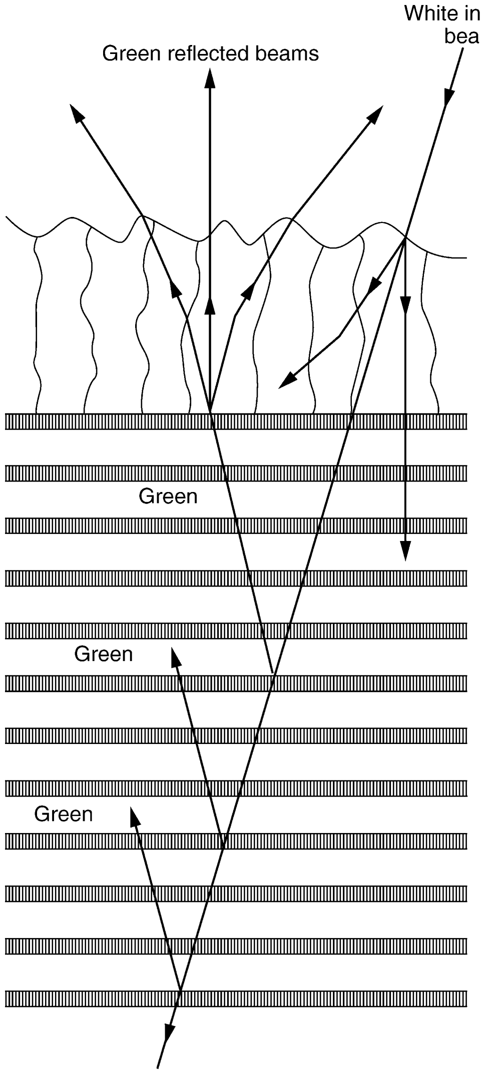

Figure 5 Generalized diagram of a multilayer reflector in the cuticle of the green beetle *Calloodes grayanus* (high refractive index material is shown shaded). Outer layer causes scattering.

iridophores. Iridophores are cells that, in this case, contain groups (iridosomes) of flexible layers of thin lamellae with cytoplasm between, forming a quarter-wave stack. The platelets of both squids and octopods develop from the rough endoplasmic reticulum and are separated by extracellular space. Euphausiid crustaceans possess photophores with very elaborate mirror-like reflectors. Up to 60 dense layers, about 70 nm thick and 75–125 nm apart, are formed from the aggregation of granules (probably a type of chitin) and surround the main photogenic mass. An intricate ring consisting of very flattened cells, forming the dense layers of a multilayer reflector (about 175 nm thick, separated by 90 nm) surrounds the lens of the photophore and reflects blue light at acute angles of incidence.

Dead invertebrates may not display their original colors. Following death, one (or both) of the layers in a multilayer reflector may become gradually reduced. For example, water may be lost from the system. This occurs in beetles of the genus *Coptocycla*; their brassy yellow color quickly changes through green, blue, and violet until the brown of melanin is finally observed. The color progression may subsequently be reversed by water uptake. This is an important consideration when examining fossils for multilayer reflectors. Multilayer reflectors have been identified in the shells of some ammonites, preserved as fossils.

Diffraction Gratings

When light interacts with a periodic surface consisting for example of a series of parallel grooves, it may be deviated from the direction of simple transmission or reflection. For this to happen, light which is scattered or diffracted from successive grooves should be out of phase by integral values of 2π. This occurs when, for a given direction of propagation, the optical path difference via successive grooves is $M\lambda$, where M is an integer known as the circle number.

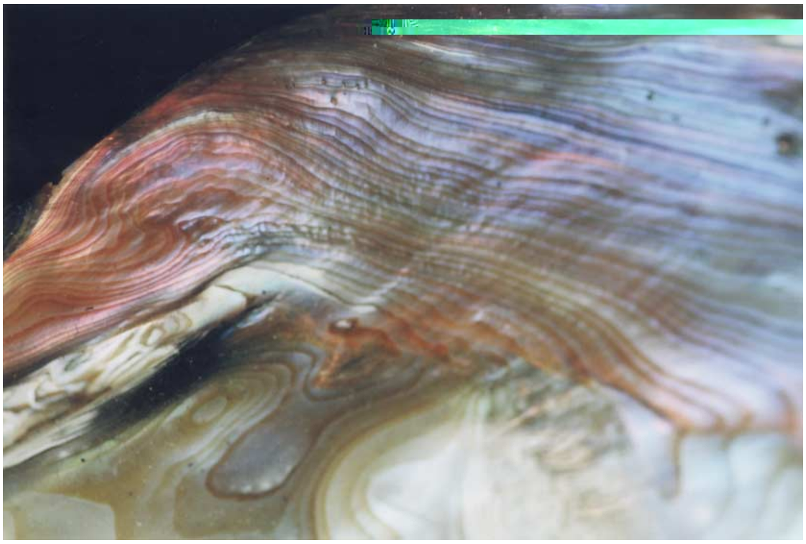

Figure 6 Bivalve shell with outer pigment removed to reveal an incidental multilayer reflector. Layers evolved to provide strength and prevent crack propagation.

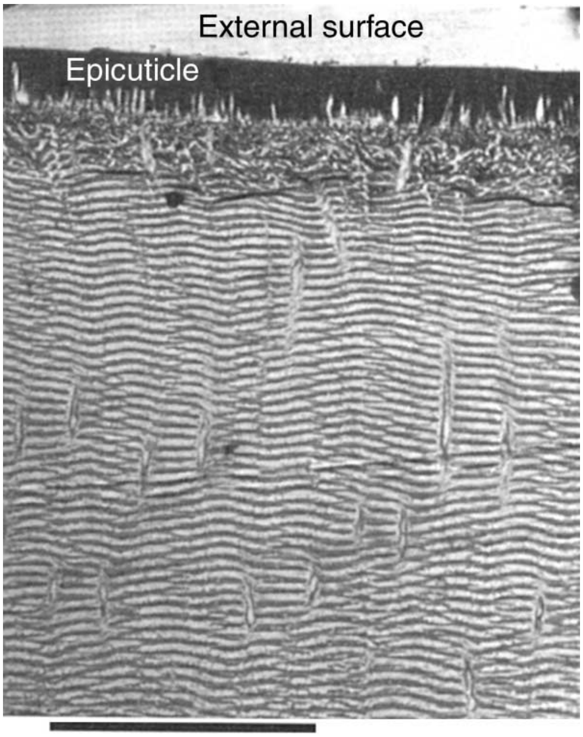

Figure 7 Transmission electron micrograph of a multilayer reflector in the cuticle of a swimming paddle of the crab *Ovalipes molleri* (Crustacea: Decapoda). Layers of reflector are slightly sinuous and out of phase; note the unusual side branches of each high refractive index (dark) layer, which provide support for these solid layers within a liquid matrix. Scale bar represents 5 μm.

This may be expressed by the grating equation:

$$2d(\sin\alpha - \sin\beta) = M\lambda \qquad [2]$$

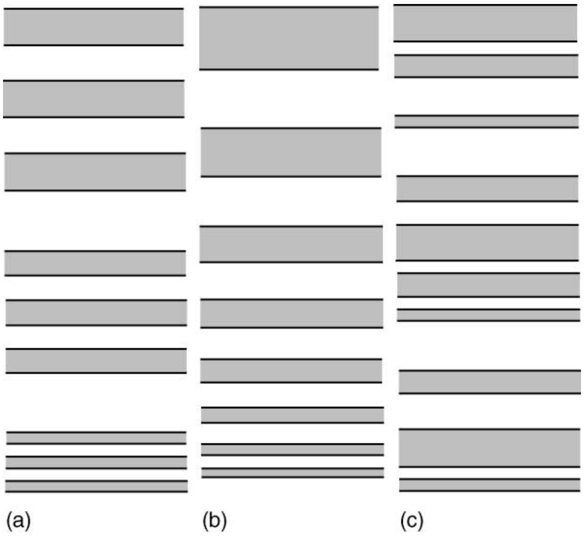

Figure 8 Three ways of achieving a broadband, wavelength-independent reflector in a multilayer stack (high refractive index material is shown shaded). (a) Three quarter-wave stacks, each tuned to a different wavelength. (b) A 'chirped' stack. (c) A 'chaotic' stack.

where α and β are angles of incidence and diffraction, and d is the period.

As with multilayers, the effect gives rise to coloration because different wavelengths are diffracted into different directions. Although the effect changes with angle of incidence, it is less critical than it is with thin films and the visual appearance is different. For a parallel beam of white light incident upon a multilayer, one wavelength will be reflected as determined by the Bragg Condition. The same beam

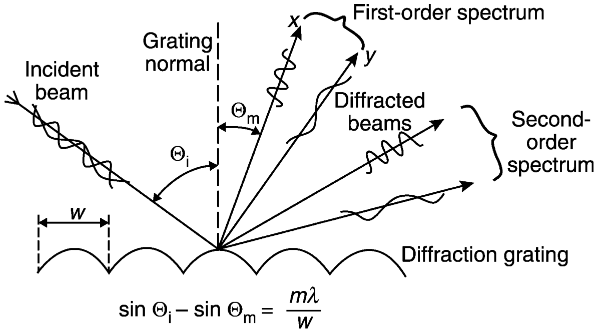

Figure 9 Reflection-type diffraction grating (w = periodicity, m = order of reflected beam). For white light of fixed angle of incidence (Θ_i), the color observed is dependent on the point of observation (e.g., violet light can be seen at point x, red at point y, within the first order spectrum).

incident upon a grating will be dispersed into spectra (e.g., Figure 9). The complete spectrum reflected nearest to the perpendicular (grating normal) is the first order. The first-order spectrum is reflected over a smaller angle than the second-order spectrum, and the colors are more saturated and appear brighter within the former. Diffraction gratings have polarizing properties, but this is strongly dependent on the grating profile.

Diffraction gratings were believed to be extremely rare in nature, but have recently been revealed to be common among invertebrates. They are particularly common on the seate or setules (hairs) of Crustacea, such as on certain first antennal setules of male Myodocopina ostracods or ‘seed shrimps’ (Crustacea) (Figure 10). Here, the grating is formed by the external surface of juxtaposed rings with walls circular in cross-section (Figure 11). The width of the rings, and consequently the periodicity of the grating, is about 700 nm in *Azygocypridina lowryi*. Different colors are viewed with varying angles of observation under a fixed light source. The ostracod *Euphilomedes carcharodonta*, for example, additionally houses a diffraction grating on the rostrum, a continuous flattened area of the carapace that is corrugated to form periodic ridges. The dark brown beetle *Serica sericea*, bears gratings on its elytra with 800 nm periodicity, which causes a brilliant iridescence in sunlight.

Many polychaetes possess gratings on their setae (hairs). For example, the opheliid *Lobochesis longiseta* bear gratings with periodicities in the order of 500 nm (Figure 12), appearing iridescent. The wings of the neurochaetid fly *Neurotexis primula* bear diffraction gratings only on their dorsal surfaces, and the iridescent effect remains after the insect is gold coated for electron microscopy. These gratings cause iridescence with a higher reflectance than the iridescence of the membranous wings of other insects, which reflect light by interference. Iridescence caused by interference disappears after gold coating because transmission of light through the outer surface is prevented.

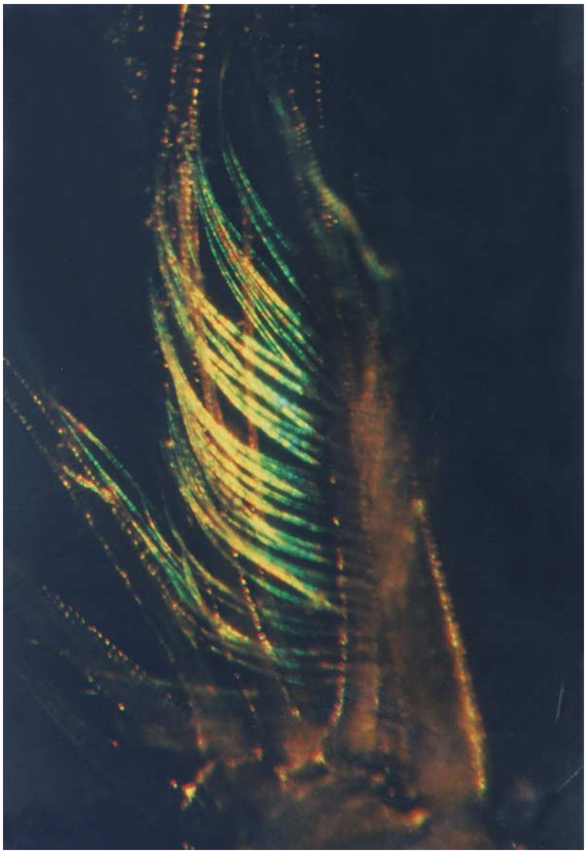

Figure 10 Diffraction gratings on the halophores (hairs) on the first antenna of the seed shrimp *Azygocypridina lowryi* (Ostracoda).

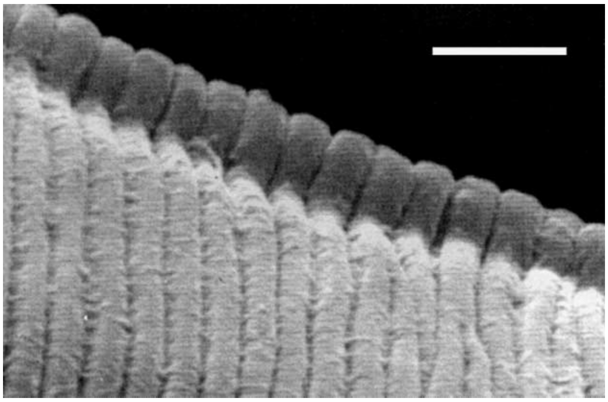

Figure 11 Scanning electron micrograph of diffraction grating on a halophore (setule) of *Azygocypridina lowryi* (Ostracoda). Scale bar represents 2 μm.

Very closely spaced, fine setules may also form the ridges of a diffraction grating. Cylindroleberidid

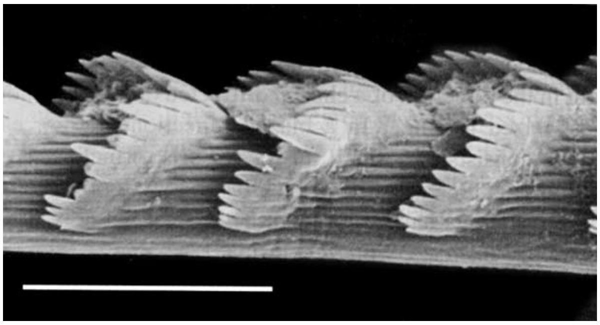

Figure 12 Scanning electron micrograph of diffraction gratings on a hair of *Lobochesis longiseta* (Polychaeta). Scale bar represents 10 μm.

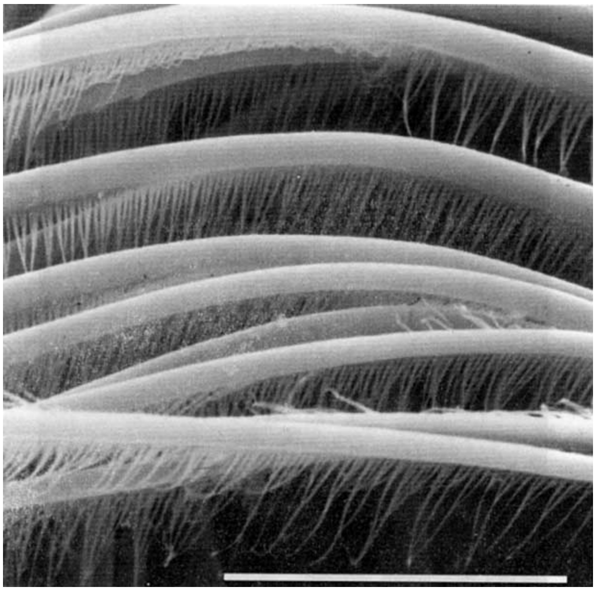

Figure 13 Scanning electron micrograph of *Tetraleberis brevis* (Ostracoda), hairs of comb of maxilla; setules (orientated vertically) form the ridges of a grating. Scale bar represents 2 μm.

ostracods (seed-shrimps) possess a comb on their maxilla bearing numerous setules on each seta (hair), collectively forming a grating with a periodicity of about 500 nm (Figure 13).

The 'helicoidal' arrangement of the microfibrils comprising the outer 5–20 μm of the cuticle (exocuticle) of certain scarabeid beetles, such as *Plusiotis resplendens*, also gives rise to metallic colors. Here, the fibrils are arranged in layers, with the fibril axis in each layer arranged at a small angle to the one above, so that after a number of layers the fibrillar axis comes to lie parallel to the first layer. Thus, going vertically down through the cuticle, two corresponding grating layers will be encountered with every 360° rotation of the fibrils – the 'pitch' of the system. Polarized light encounters an optically reinforcing plane every half turn of the helix. The system can be treated as a three-dimensional diffraction grating, with a peak reflectance at $\lambda = 2nd$, where d is the separation of analogous planes, or half the pitch of the helix. The diffracted light resembles that from a linear grating except for the polarization; the three-dimensional grating reflects light that is circularly or elliptically polarized. It should be noted that the diffracted color does not depend on the total film thickness as it does in interference, but on the layer repeat distance within the film as in the diffraction grating (analogous to 'liquid crystals').

When each groove of a grating is so formed that it independently, by means of geometrical optics, redirects the light in the direction of a chosen diffracted order, it is known as a blazed grating. In a blazed reflection grating, each groove consists of a small mirror (or prism) inclined at an appropriate angle (i.e., the grating has a 'saw-tooth' profile). Blazed gratings have been identified on the wing scales of the moth *Plusia argentifera* (Plusinae).

When the periodicity of a grating reduces much below the wavelength of light, it becomes a zero-order grating and its effect on light waves changes. This difference in optical effect occurs because when the periodicity of the grating is below the wavelength of light the freely propagating diffracted orders are suppressed and only the zero-order is reflected when the illumination is normal to the plane of the grating. To describe accurately the optical properties of a zero-order grating, rigorous electromagnetic theory is required. In contrast to gratings with freely propagating orders, zero-order structures can generate saturated colors even in diffuse illumination. Such structures occur on the setae (hairs) of the first antenna of some isopod crustaceans, such as the giant species of *Bathynomus*. Here, there are diffracted orders and the spectral content of the light within the grating is controlled by the groove profile. In an optical system that only accepts the zero-order, what is seen is white light minus that diffracted into the ±1 orders.

A zero-order grating can cause total transmission (i.e., there is no reflection). Such antireflective structures are found on the corneal surfaces of each ommatidium (visual unit) in the eye of *Zalea minor* (Diptera) (Figure 14). The periodicity of the corneal gratings of this fly is 242 nm. Another form of antireflection grating is formed on the transparent wings of the hawkmoth *Cephonodes hylas* on the corneal surface of each visual unit (ommatidium) of the eyes of moths and butterflies (Figure 15). Here, optical-impedance matching is achieved by means of a hexagonal array of tapered cylindrical protuberances, each of about 250 nm diameter, thus forming a 'tri-grating' with grooves transecting at 120°.

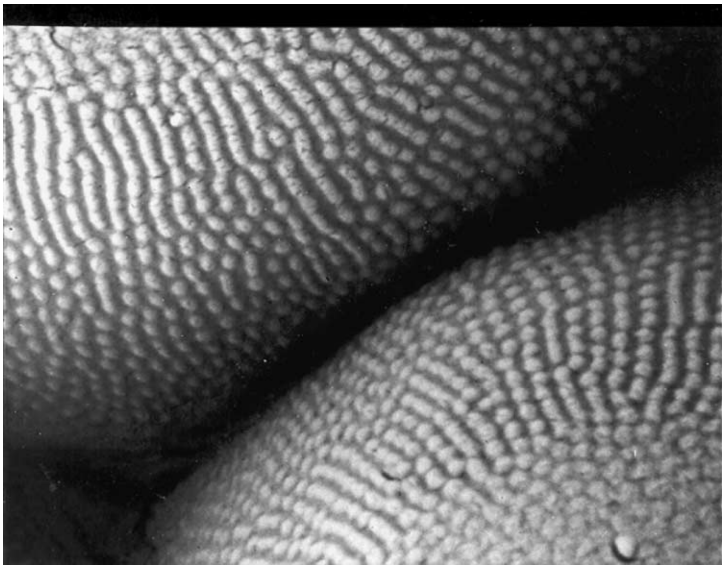

Figure 14 Scanning electron micrograph of the compound eye of *Zalea minor* (Diptera). The corneal surface at the junction of two ommatidia is shown with antireflection gratings (positioned on sloping regions, i.e., at high angles to the normal of the eye) phasing into moth-eye-type protuberances. Scale bar represents 2 μm.

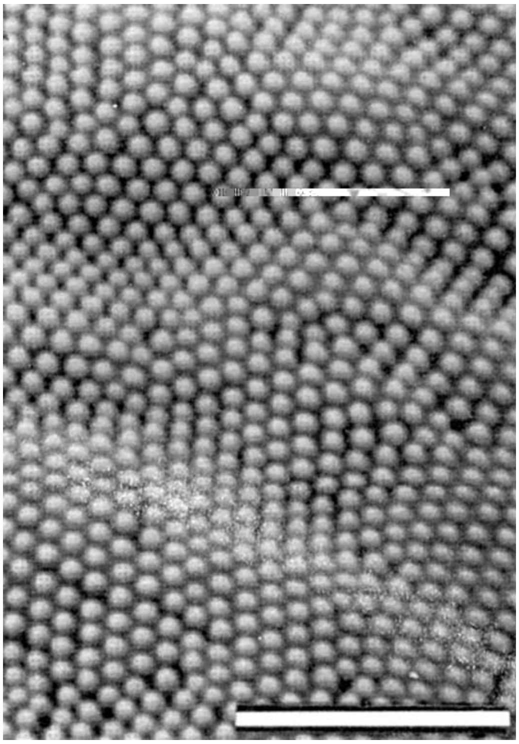

Figure 15 Tri-grating (antireflector) on the corneal surface of a butterfly eye. Scale bar represents 2 μm.

The protuberances provide a graded transition of refractive index between the air and the cornea/wing. Hence the refractive index at any depth is the average of that of air and the corneal/wing material.

The grooves of a grating may also create parallel rows in two directions, forming a bi-grating.

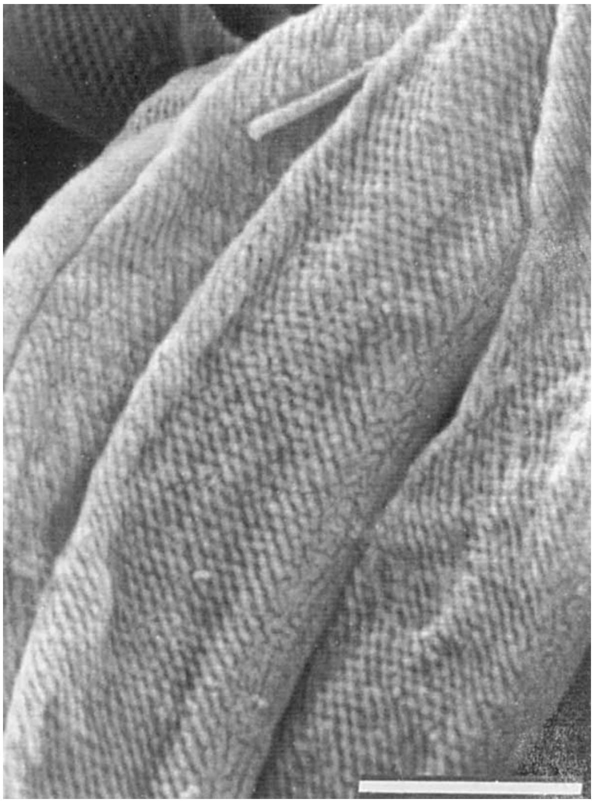

Figure 16 Bi-grating on the callynophores (hairs) of the amphipod crustacean *Waldeckia australiensis*. Scale bar represents 2 μm.

Bi-gratings can be found in some crustaceans and flies. In the amphipod crustacean *Waldeckia australiensis*, two effectively superimposed gratings subtend angles of about 60/120° (Figure 16).

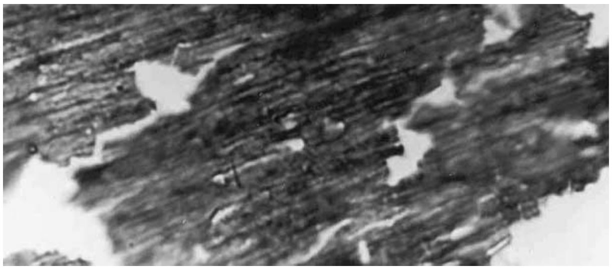

Figure 17 Light micrograph of the surface of a hair of *Canadia spinosa* (Middle Cambrian, 515 million years old, British Columbia), isolated by acid maceration of the rock matrix, showing gratings running longitudinally. Periodicity 0.9 µm.

In terms of fossil colors, diffraction gratings are responsible for the nacreous luster of brachiopod shells, such as those from the Devonian, around 360 million years old. Here, tabular aragonite platelets averaging 600 nm in thickness, each comprising a linear diffraction grating, form layers and consequently a three-dimensional diffraction grating. Antireflective, zero-order gratings have been identified on the eye of an Eocene fly, 45 million years old, preserved in Baltic amber. Linear diffraction gratings causing color have been discovered on the sclerites of *Wiwaxia corrugata* from the Burgess Shale (Middle Cambrian, 515 million years old) (Figure 17). This finding has led to the 'light switch theory', in that the evolution of the first eye caused the big bang in animal evolution.

Scattering (Incoherent Reflectors)

Simple, equal scattering of all spectral wavelengths results in the observation of a diffuse white effect. This commonly arises from the effects of a nonperiodic arrangement of colloidally dispersed matter where the different materials involved have different refractive indices (Figure 18), or from solid colorless materials in relatively concentrated, thick layers. In the colloidal system, the particles are larger than the wavelength of light and can be thought of as mirrors oriented in all directions. The reflection is polarized unless the incident light is at normal incidence on the system and, in the colloidal system, spherical or randomly arranged particles are involved.

The colloidal system involves either a gas-in-solid, gas-in-liquid, liquid-in-liquid (emulsions), or solid-in-liquid. For example, the gas-in-liquid system is partly responsible for the white body and/or tentacles of certain anemones. Light is reflected and refracted at the surfaces of the particles of matter or spaces (with dimensions >1 µm), regardless of the color of the materials involved (except opaque brown and black compounds, such as melanin). In insects, the materials involved typically have very low transparencies.

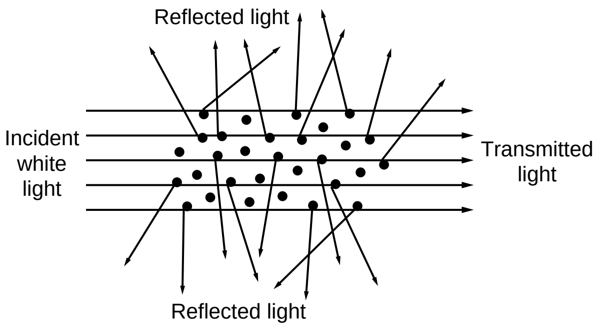

Figure 18 Scattering of white light by small particles (represented by black circles). The reflected light is randomly directed at different angles.

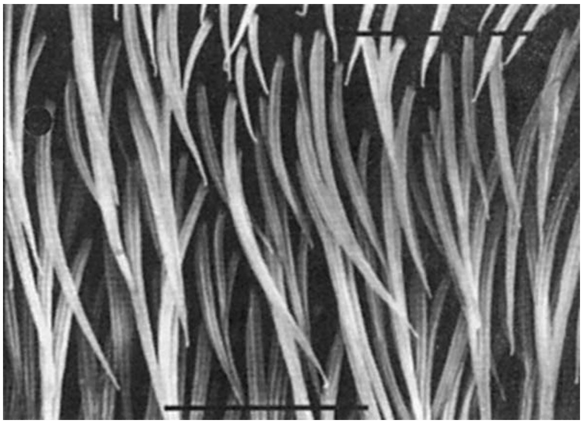

Figure 19 Scanning electron micrograph of a section of a white reflecting patch on the abdomen of the fly *Amenia* sp., showing closely packed, randomly arranged hairs that scatter incident light in all directions (Figure 11). Scale bar represents 20 µm.

An unordered (as opposed to periodic) group of closely spaced setae (hairs), such as those in patches on the fly *Amenia* sp., may form a white reflection via random scattering or reflection (Figure 19). However, if the arrangement becomes periodic to some degree, a diffraction grating may be formed, such as the grating of *Tetraleberis brevis* (Ostracoda) (Figure 13).

From some scales of butterfly wings, light is scattered uniformly and completely in all directions, due to the chaotic disposition of the surfaces. Matt or pearly whites may be observed depending on the complexity of the arrangement of the structures, which affects the relative degree of scattering. The structures may be so small that the molecular topography of the surface has an effect. The chromatic effects of the butterfly scales are greatly intensified if a dark, absorbing pigment screen lies beneath. This screen prevents reflection of white or colored light from the background that would dilute or alter the color. Additionally, if a dark pigment such as melanin is interspersed with the scattering elements, the reflection will appear a shade of gray or brown. The cells of the reflector in the photophore of a beetle ('fire-fly') are packed with sphaerocrystals of urate that cause a diffuse reflection (Figure 20).

Reflection and refraction that occur at the interfaces of strata with different refractive indices may result in the display of white light. The degree of whiteness depends upon the difference in refractive indices. This mechanism is evident in the shells of many lamellibranch molluscs. Between the outer, often pigmented layer and the mantle, is a thick middle layer of crystalline calcium carbonate. The inner surface of this (nacreous) layer is lined with multiple laminations of the same salt. In most species these laminations are sufficiently thick (>1 μm) to render the inner lining white, although in some species they become thin so as to form a multilayer reflector. Calcium carbonate similarly produces whiteness in Foraminifera and in calcareous sponges, corals, echinoderms, and crustacean cuticles. Also, in the class of white solids is silica in diatom tests and skeletons of hexactinellid sponges.

Other forms of scattering also exist and result in a blue colored effect (red when the system is viewed in transmission). Tyndall or Mie scattering occurs in a colloidal system where the particle size approximates the wavelength of light. Here, diffraction is important. Rayleigh scattering occurs in molecules in a two-photon process by which a photon is absorbed and raises the molecule to an excited electronic state from which it re-radiates a photon when it returns to the ground state. Diffraction is not involved here.

Tyndall scattered light is polarized under obliquely incident light. The intensity of the resultant blue is increased when it is viewed against a dark background, such as melanin. The relative sizes of particles determine the shade of blue. If the particles responsible for the scattering coalesce to form particles with a diameter greater than about 1 μm, then white light is observed (see above). A gradation from blue to white scattering ('small' to 'large' particles) occurs on the wings of the dragonfly *Libellula pulchella*.

Scattered blues can also be found in other dragonflies. In the aeschnids and agrionids, the epidermal cells contain minute colorless granules and a dark base. The males of libellulids and agrionids produce a waxy secretion that scatters light similarly over their dark cuticle. The green of the female *Aeschna cyanea* is the combined result of Tyndall scattering and a yellow pigment, both within the epidermal cells (degradation of the yellow pigment turns the dead dragonfly blue).

Scattered blues are also observed from the skin of the cephalopod *Octopus bimaculatus*, where a pair of ocelli are surrounded by a broad blue ring. Blue light is scattered from this region as a result of fine granules of purine material within cells positioned above melanophore cells. The color and conspicuousness of the ring are controlled by the regulation of the melanophores, by varying the distribution of melanin

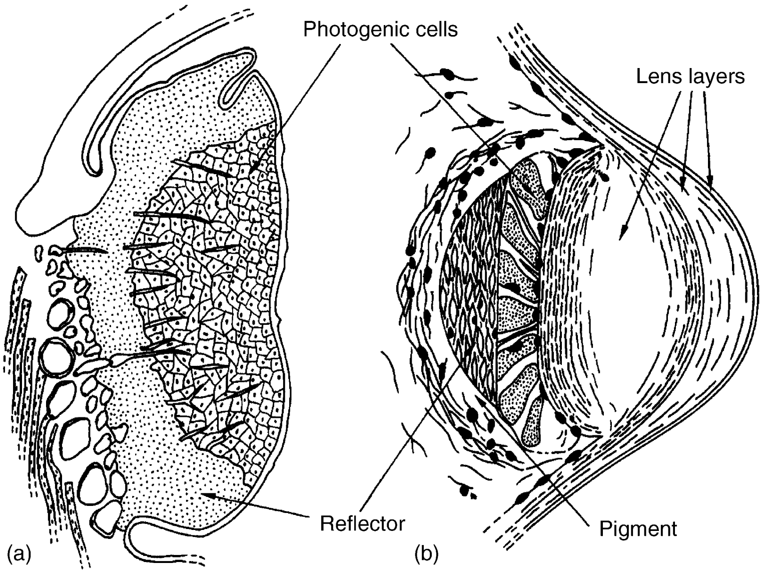

Figure 20 Sections of photophores (bioluminescent organs), external surface of animals on the right. (a) The 'fire-fly' beetle *Pyrophorus* sp., abdominal photophore of male – reflector is based on a scattering system. (b) The shrimp *Sergestes prehensilis*, showing lens layers and absorptive pigment – reflector is based on a multilayer type.

and consequently the density of the absorbing screen. The squid *Onychia caribbaea* can produce rapidly changing blue colors similarly. The bright blue patterns produced by some nudibranch molluscs result from reflecting cells containing small vesicular bodies, each composed of particles about 10 nm in diameter and therefore appropriate for Rayleigh scattering.

See also

Diffraction: Diffraction Gratings. **Scattering:** Scattering Theory.

Further Reading

Fox DL (1976) *Animal Biochromes and Structural Colours*. Berkeley, CA: University of California Press.

Fox HM and Vevers G (1960) *The Nature of Animal Colours*. London: Sidgwick and Jackson Ltd.

Gale M (1989) Diffraction, beauty and commerce. *Physics World* 2: 24–28.

Harvey EN (1952) *Bioluminescence*. New York: Academic Press Inc.

Herring PJ, Cambell AK, Whitfield M and Maddock L (1990) *Light and Life in the Sea*. Cambridge, UK: Cambridge University Press.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Land MF and Nilsson D-E (2002) *Animal Eyes*. Oxford, UK: Oxford University Press.

Pain S (1998) Seeing the light. *New Scientist* 2161: 42–46.

Parker AR (1999) Light-reflection strategies. *American Scientist* 87: 248–255.

Parker AR (2003) *In the Blink of an Eye*. London: Simon and Schuster/Cambridge (USA): Perseus Publishing.

Parker AR, Welch VL, Driver D and Martini N (2003) An opal analogue discovered in a weevil. *Nature* 426: 786–787.

Parker AR, McPhedran RC, McKenzie DR, Botten LC and Nicorovici N-AC (2001) Aphrodite's iridescence. *Nature* 409: 36–37.

Microstructure Fibers

R S Windeler, OFS Laboratories, Murray Hill, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Microstructure fibers have unique properties that cannot be obtained from traditional fibers (i.e. all glass, doped silica fibers) and can deliver functionalities superior to many of today's best transmission and specialty fibers. Their unique properties are obtained from an intricate cross-section of high and low index regions that traverse the length of the fiber. The vast majority of fibers consist of silica for the high-index region and air for the low-index region. These fibers are known by several different names including microstructure fiber, holey fiber, and photonic crystal fiber. The term microstructure fiber is used in this chapter because it is the most generic, encompassing all the fiber types.

Microstructure fiber properties vary greatly and are determined by the size and arrangement of air holes in the fiber. For example, fibers have been fabricated such that the numerical aperture of the core or cladding approaches one. They have been fabricated with unique dispersion profiles, such as a zero-group velocity dispersion in the near visible regime or an ultraflat dispersion over hundreds of nanometers wide. They have been fabricated to generate a super continuum over two octaves wide. Some guide light with an air core via bandgap guidance. In addition, the air holes in these microstructure fibers can be filled with highly tunable materials, giving one the capability of controlling the fiber properties for use in energy efficient devices.

Microstructure fibers can have doped regions like traditional fibers. These hybrid fibers combine the benefits of traditional and microstructure fibers. The doped core typically guides most of the light, but its guidance properties can be strongly influenced by the air regions. Applications for hybrid fibers include dispersion compensation, polarization maintaining, and bend insensitive fibers.

Loss is always an important factor when determining whether a microstructure fiber will compete with or replace a traditional optical fiber. The index profiles that make these fibers so special can also lead to fibers that have an inherently high loss at connections or along the length of the fiber. Loss at connections typically occurs due to mode mismatch between the two fibers, undesired hole collapse, or poor alignment. Loss along the fiber length can occur due to impurities, hole surface roughness, or poor confinement.

The dispersion profile (zero dispersion wavelength, dispersion slope, normal or anomalous) of microstructure fibers can be optimized more than a traditional fiber, because the hole size and placement can be arranged to make the cladding index strongly wavelength dependent. Dispersion causes a pulse to spread because the phase velocity is wavelength dependent. Normal dispersion is when longer wavelengths travel faster than shorter wavelengths. Anomalous dispersion is just the opposite – shorter wavelengths travel faster than longer wavelengths. Dispersion determines the amount of interaction between different wavelengths. For applications such as transmission fiber, one wants little interaction and for high nonlinear applications, one wants significant interaction.

This chapter examines the methods for fabricating microstructure fibers and reviews the different types of microstructure fibers and their properties. Fibers that guide by total internal reflection are reviewed separately from bandgap guided fibers because their properties are significantly different. Lastly, a description and example of an active device is presented in which the air regions of a microstructure fiber are filled with controllable material.

Microstructure Fiber Fabrication

At first glance, fabrication of microstructure fibers looks similar to traditional fibers in that the fibers are fabricated in a two-step process. First a preform is fabricated and then it is drawn (stretched) into a fiber. However, when the process is examined in more detail, both the preform fabrication and draw depart significantly from traditional methods. First, a novel preform fabrication method is used to incorporate air holes that run the length of the preform. Second, novel drawing procedures are used to keep the air holes open as the preform diameter is reduced by several orders of magnitude during draw.

Preform Fabrication

Microstructure preforms are cylinders of amorphous material usually less than a few centimeters in diameter that have an index profile running the length of the preform similar to the desired fiber. The most common preform consists of air and a single material such as silica, polymer, or high nonlinear glass. These air-containing preforms are relatively easy to fabricate but are difficult to draw because the structures can deform.

The basic parameters of the fiber are usually determined by the geometry of the holes in the preform. These include hole diameter (d), hole position, pitch (center-to center hole spacing, Λ), core diameter (a), and number of layers. However, changes in the size and shape of the holes in the fiber can be made purposely or accidentally during draw, which will cause deviations between the fiber and preform profiles.

Several methods are used to fabricate microstructure preforms including stacking, casting, extrusion, and drilling. Each process has its advantages and disadvantages. By far, the most common method is stacking in which tubes, rods, and core rods (rods containing doped cores fabricated by traditional optical fiber techniques) are stacked in a close packed pattern such as a triangular or hexagonal lattice (Figure 1). The bundle of tubes and rods is then bound together with a large tube, called an overclad tube. Fibers can be made with complicated or asymmetric index profiles by strategically placing tubes with the same outer diameter (for good stacking) but a different inner diameter to change the index in that location. For example, a fiber with smaller holes on opposite sides of the core produces high birefringence.

The main advantages of stacking are that no special equipment is needed for fabricating microstructure preforms and that doped cores are easily added. However, there are several disadvantages. First, the stacking method is limited to simple geometries of air holes because the tubes are stacked in a close pack arrangement. Second, unless hexagonal tubes are used, interstitial areas are created between the tubes that may not be desired in the final fiber. Third, the method is labor intensive and requires significant glass handling.

A variation of the stacking method is used for fabricating air-clad fibers. Air-clad preforms are created by placing a layer of small tubes around the perimeter of a large core rod. The assembly is then overcladed for strength (Figure 2).

Extrusion and casting of glass powder, polymer, or sol–gel slurry are also used for making single material

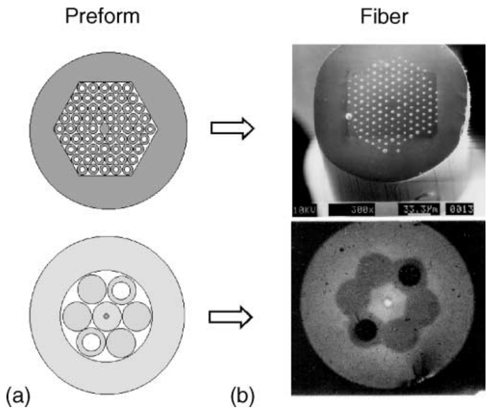

Figure 1 (a) Tubes, rods, and core rods are stacked together in a close packed arrangement and held together with an overclad tube. (b) During draw the desired air region stay open forming a microstructure fiber.

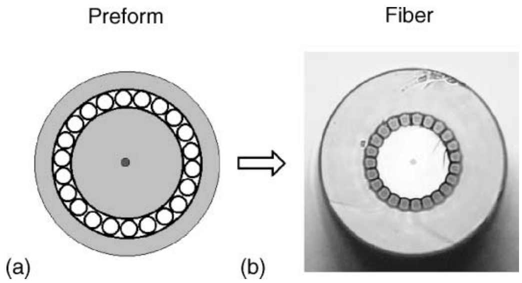

Figure 2 (a) An air-clad preform is fabricated by stacking a ring of small, thin walled tubes around a large core rod. (b) Photograph of the resulting fiber.

microstructure preforms. The main advantage of extrusion and casting is that complicated structures can be fabricated in which the position, size, and shape of the air regions are independent of one another. The disadvantage of this method is that preform fabrication is more difficult compared to the other fabricate methods. This method may become more common as complex air structures are needed to make advanced microstructure fibers.

The last method consists of drilling holes in a preform or rod. Drilling is well understood and used for other specialty fibers. The advantage of this method is that it is easy to drill holes of various sizes in any position in a preform, including doped regions. The disadvantages of drilling are that the holes cannot be drilled very deep compared to a traditional preform length; the distance between holes may be limited due to cracking; and the fiber may experience high loss due to surface roughness of the holes and impurities incorporated during drilling.

Fiber Draw

It is significantly more difficult to draw a microstructure preform with air holes than a traditional preform. This is because the air holes tend to close due to surface tension. This force can be overcome or minimized using two different techniques. The first is to draw the preform under very high tension (low temperature). Minimal hole collapse occurs when the draw tension is significantly larger than the surface tension. The problem with this method is that the break rate increases substantially at higher tension.

The second method is to apply an external pressure to the holes to counteract the surface tension. If a single pressure source is used, the holes can be made larger or smaller during draw by changing the pressure as needed. However, the process can become unstable because a large hole needs a lower pressure to maintain its size than a small hole. If holes of different sizes exist, a larger hole will grow at the expense of a smaller one regardless of the pressure used. To minimize the instability, this method is typically performed in conjunction with high-tension drawing.

To alleviate the draw problem, a second solid material can be used in the preform to obtain the desired index profile. With the air regions removed from the preform, the drawing process becomes significantly easier. In addition, when a preform consists of two solid materials, one of the materials does not have to be continuous, as in single material preforms. This allows for simpler designs such as concentric rings, which may be easier to fabricate and model (Figure 13c). The disadvantage of this approach is finding two materials that have compatible physical and chemical properties, have low loss, and have a large index of refraction difference.

Microstructure Fiber Types

Microstructure fibers are categorized by their method of guidance, properties, and function. Index guided microstructure fibers most closely resemble traditional fibers and are described first. These fibers guide light by total internal reflection and have a core index of refraction greater than the cladding. Bandgap fibers are examined next. The unique guidance of bandgap fibers allows the core index to be lower than the cladding index. Tunable microstructure fibers are described last. These microstructure fibers are filled with tunable materials so that the fiber's properties can be actively manipulated.

Index Guided Fibers

Most microstructure fibers guide light by total internal reflection. The size and spacing of air holes determine the guiding properties. For example, a solid silica core surrounded by a cladding consisting of small air holes that are spread apart creates a small core-to-cladding index difference similar to what can be achieved in traditional fibers. At the other extreme, a solid silica core surrounded by a cladding of large holes, which are closely spaced together, forms a large core-to-cladding index difference. This structure creates a large numerical aperture in the core and optically isolated regions within the fiber. The ability to create regions of significantly different indices in various parts of the fiber distinguishes the index guided microstructure fibers from traditional fiber. Different types of microstructure fiber are formed by varying the size, spacing and pattern of the air holes.

Different techniques are used to model these fibers, depending on spacing and size of the air holes relative to the wavelength of the guided light. If the hole diameter and spacing is much smaller than the wavelength of the guided light, the cladding index can be modeled as an average index weighted by the volume fraction of the two materials (typically air and silica). As the hole diameter or spacing approaches the wavelength of the guided light, complicated models are used such as finite difference method, finite element method, beam propagation method, or multipole method.

Endlessly single-mode fiber

The endlessly single-mode fiber consists of a solid core surrounded by a two dimensional array of small air holes in the cladding (Figure 3). The size and position of the air holes are arranged such that only the fundamental mode exists in the core regardless of the wavelength of the guided light.

The prevention of higher order modes can be understood by examining the V number. The V number is a dimensionless parameter that describes the guiding nature of the waveguide. In traditional fiber, when the V number is less than 2.405, the fiber will guide only the fundamental mode. The V number is given by

$$V = \frac{2\pi a}{\lambda}(n_{\text{core}}^2 - n_{\text{clad}}^2)^{1/2} \qquad [1]$$

where a is the core diameter, λ is the wavelength of the guided light, and n_{core} and n_{clad} are the index of refractions of the core and cladding, respectively. In traditional fiber, the core and cladding index (n_{core} and n_{clad}) are for the most part constant, so the V number increases as the wavelength decreases.

However, for the endlessly single mode microstructure fiber, the V number does not increase indefinitely with decreasing wavelength, but approaches a constant value. Obtaining a V number independent of wavelength is understood by examining the distance the light propagates into the cladding as a function of wavelength. As the wavelength becomes shorter, light penetrates a shorter distance into the cladding, interacting with less of the air regions. This causes the effective cladding index (n_{clad}) to increase with decreasing wavelength, such that V approaches a constant value.

Microstructure fibers are made endlessly single-moded over all wavelengths by properly designing the hole diameter-to-pitch ratio (d/Λ). When the d/Λ is low enough, the microstructure fiber guides light only in the single mode regardless of wavelength and core size (Figure 4). However, these large core diameter fibers are limited by long wavelength bend loss in the same way as large core single mode traditional fibers.

High delta core fiber

High delta core microstructure fibers (HDCMF) consist of a small (typically less than 2 microns) solid silica core surrounded by one or more layers of large air holes closely spaced together (Figure 5). This creates a very large index difference between the core and cladding. A core-to-cladding index difference of 0.4 is easily achieved in a HDCMF, which is an order of magnitude larger than can be achieved with traditional fiber.

Even though the core is very small, these fibers are almost always multimoded due to the large index difference between the core and cladding. However, because of its small core, it is difficult to launch anything other than the fundamental mode into the fiber. Once a mode is guided in the fiber it remains in that mode due to the large difference in effective indices between modes (Figure 6). When higher-order modes are purposely generated in the fiber, they also

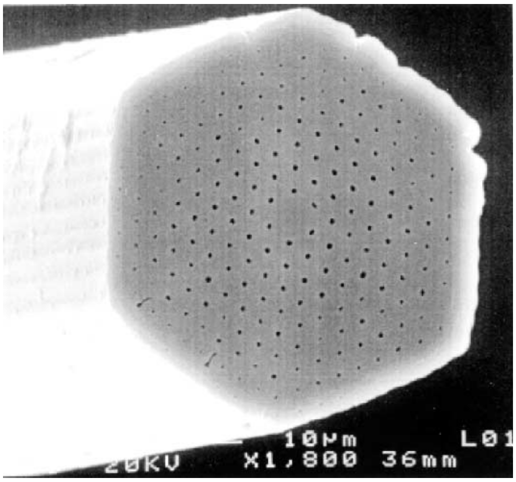

Figure 3 Photograph of an endlessly single mode fiber (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

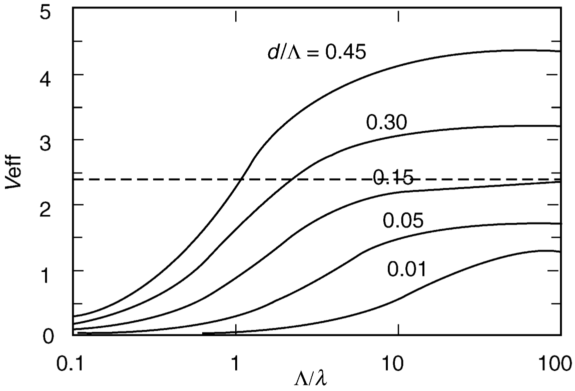

Figure 4 The effective V number as a function of Λ/λ for various d/Λ (Reprinted with permission from Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22: 961–963).

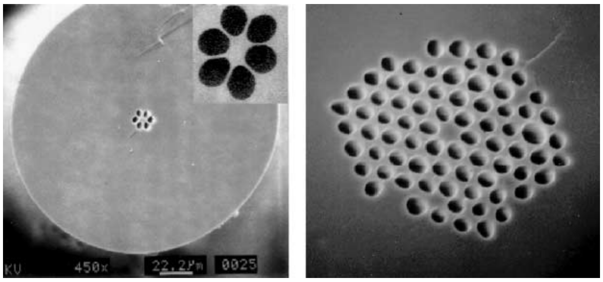

Figure 5 High delta core fibers consist of a small solid core surrounded by at least one layer of large air holes closely spaced together. Right-hand panel reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

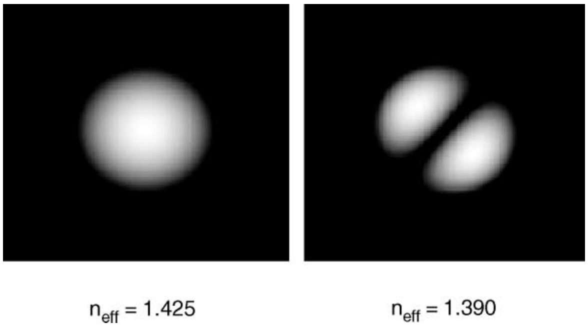

Figure 6 Individual modes do not couple well due to large differences in effective indices between modes. The figure shows the mode profile and effective index of the fundamental and the next higher mode for a fiber with a two-micron diameter core with a core-to-cladding index difference of 0.45 (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

guide over long lengths without coupling to other modes (Figure 7). This strong guidance makes the fiber extremely bend insensitive.

HDCMF can have a unique dispersion profile because it can guide the fundamental mode exclusively in the multimode regime. This allows greater flexibility in controlling the shape of the dispersion profile, especially the ability to move the zero dispersion to shorter wavelengths. The HDCMF is the first silica fiber with anomalous dispersion of single-mode light below 1270 nm (the wavelength at which the material dispersion of silica is zero). These fibers are often made to have a zero-group velocity dispersions around 770 nm for use with Ti-sapphire lasers.

HDCMF is ideal for performing nonlinear experiments in the near-visible regime because the fiber's small, high-index core creates high intensity light, and the fiber's low dispersion creates long interaction times. For example, a broadband continuum can be generated over two octaves using less than a meter of HDCMF fiber (Figures 8 and 9).

Tapered microstructure fiber (TMF)

Coupling light in and out of HDCMF is difficult and results in a large loss due to its small core. When free space optics are used to couple the light, frequent realignment of the beam is required to minimize loss due to drifts in the equipment. Splicing HDCMF fiber to traditional fiber also results in high loss due to a large modal mismatch. These coupling problems are solved with tapered microstructure fibers.

Tapered microstructure fibers (TMF) consist of a traditional germanium single-mode core surrounded by a small silica cladding (Figure 10). The inner

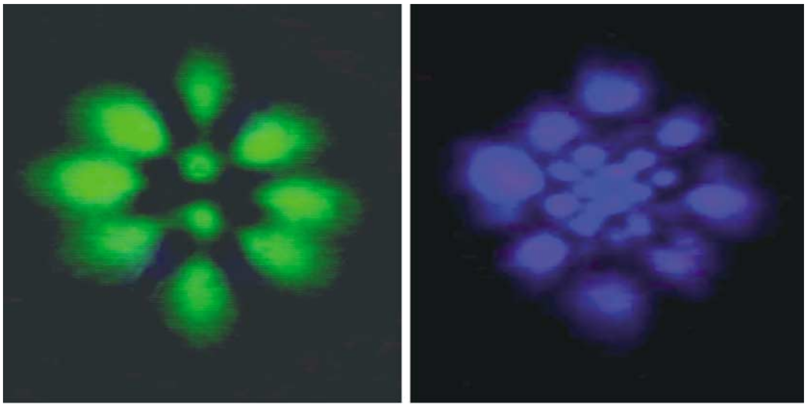

Far-field pattern, 532 nm light
5 meter fiber length

Far-field pattern, 355 nm light
2 meter fiber length

Figure 7 Far-field patterns show that higher order modes propagate long distances in the HDCMF without coupling to other modes (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798).

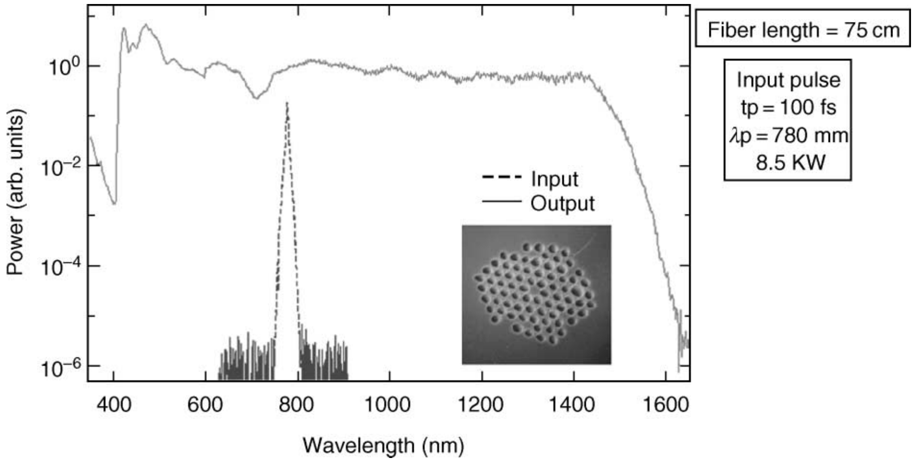

Figure 8 Plot of a supercontinuum spectrum over two octaves wide generated from a 75 cm piece of HDCMF (Reprinted with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25: 25–27).

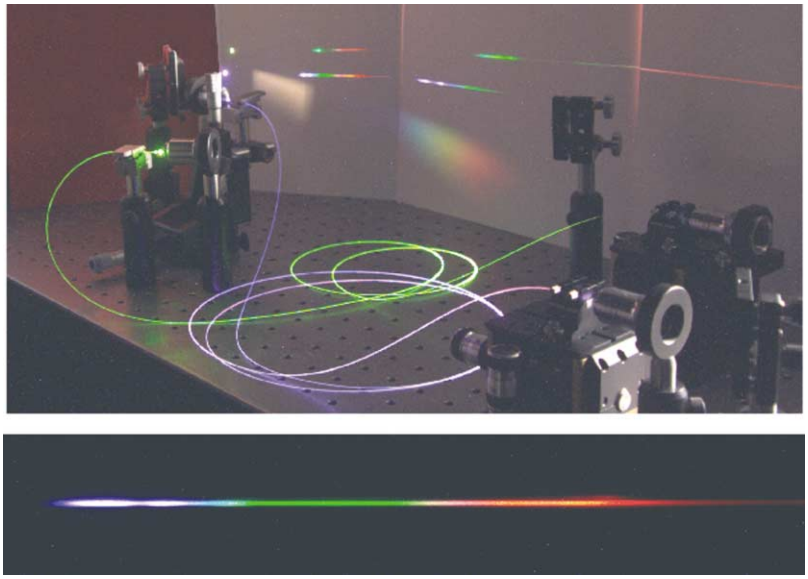

Figure 9 Photograph of fiber generating the supercontinuum. The bottom photo shows the continuum after passing through a prism.

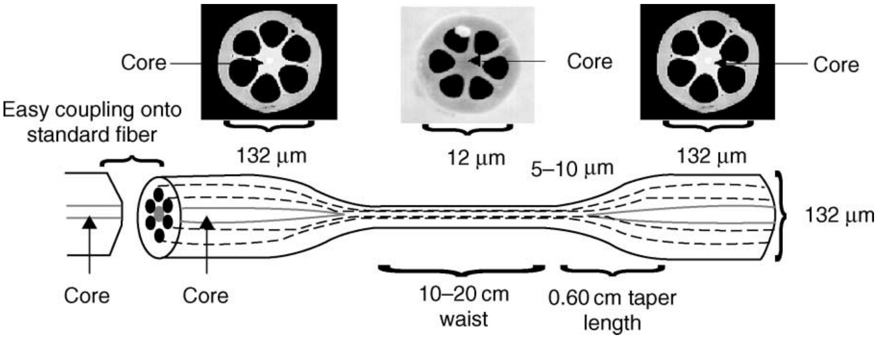
Core
Core
Core
Easy coupling onto standard fiber
132 μm
12 μm
5–10 μm
132 μm
132 μm
Core
Core
10–20 cm waist
0.60 cm taper length

Figure 10 Schematic drawing of the tapered microstructure fiber (Reprinted with permission from Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1) (©2001 IEEE)).

cladding is surrounded by a layer of air holes and then a protective outer silica cladding for strength. Since the core of the TMF is very similar to a traditional single-mode fiber core, the splice loss is low. Splice losses below 0.075 dB are typical.

Properties similar to the HDCMF are obtained by adiabatically tapering the TMF while maintaining the cross-sectional profile of the fiber. In the taper region, the fundamental mode is no longer guided by the germanium core and evolves into the fundamental mode of the silica region, where it is confined by the ring of air holes. The taper waist is typically ten to twenty centimeters long and has properties identical to the HDCMF. When the fiber is adiabatically expanded back to its original size, the light is guided back into the standard diameter germanium core, making it easy to splice to traditional fiber with low loss.

The same effect can be obtained by stretching a traditional fiber from 125 microns down to 2 microns. However, there are several disadvantages compared to the tapered microstructure fiber. The taper region is much longer in the traditional fiber; the optimal diameter of the silica core is harder to maintain; the taper region must remain clean and uncoated; and the taper region is much weaker.

Air-clad fiber

Air-clad fibers consist of a doped core surrounded by a silica inner cladding. The inner cladding is surrounded by a ring of air, which in turn is surrounded by an outer silica cladding for strength (Figure 2). Thin silica webs connect the inner and outer claddings to hold the inner fiber region in place. The region consisting of air and thin silica webs creates an optical barrier, preventing most of the light from escaping the inner cladding. Such fibers are referred to as double clad fiber.

In traditional double-clad fibers, a material of lower index (typically another layer of glass or a layer of polymer) immediately surrounds the inner silica cladding. A key optical parameter describing the light guiding properties of the inner cladding is the numerical aperture (NA). The NA is defined as the sine of the largest (acceptance) angle of light that will be guided in the inner cladding. The upper NA values of traditional double-clad fibers (~0.22 and ~0.45 for glass and polymer outer claddings, respectively) are limited by the refractive indices of the available cladding materials.

For the air-clad fiber, the NA is determined by the ability of light to escape from the inner cladding through the silica webs. The NA of the air-clad fiber can be calculated by modeling the web as an infinitely long slab waveguide. As seen in Figure 11, the NA increases as the webs become thinner or the wavelength of light becomes longer. To obtain NA similar to what is achieved by coating a traditional fiber with a low index polymer (NA ~0.45), the web thickness must be about equal to the wavelength of the light. The NA of the air-clad fiber increases dramatically as the web thickness becomes less than half the wavelength of the guided light. Numerical apertures above 0.90 have been experimentally demonstrated.

The large NA values achievable with air-clad fibers are an important advantage for advanced double-clad

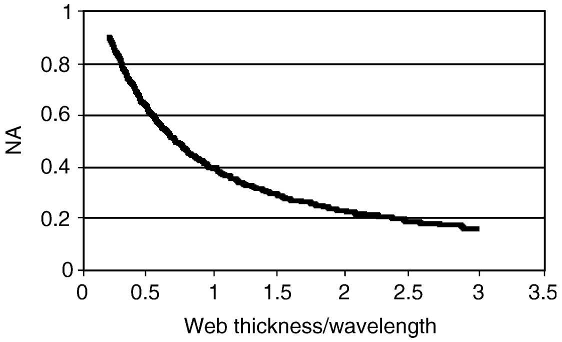

Figure 11 Predicted NA as a function of the web thickness divided by the wavelength using an infinite slab model.

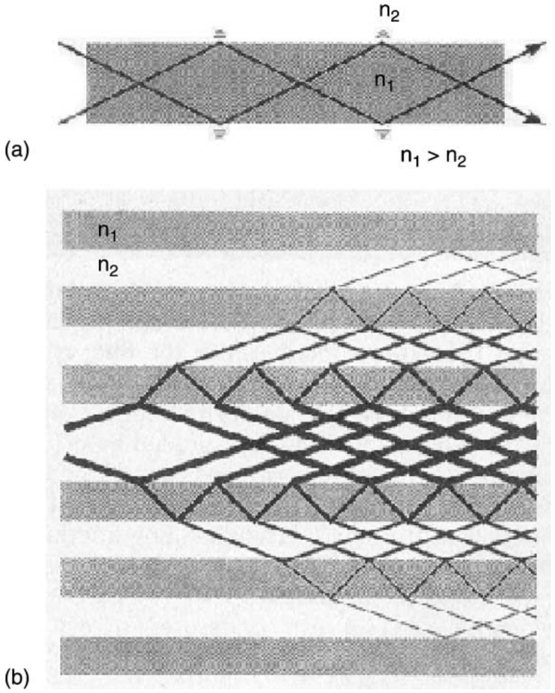

Figure 12 Schematic of the mechanism of (a) index guided light and (b) band gap guided light (Reprinted with permission from Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285. Copyright 1999 American Association for the Advancement of Science).

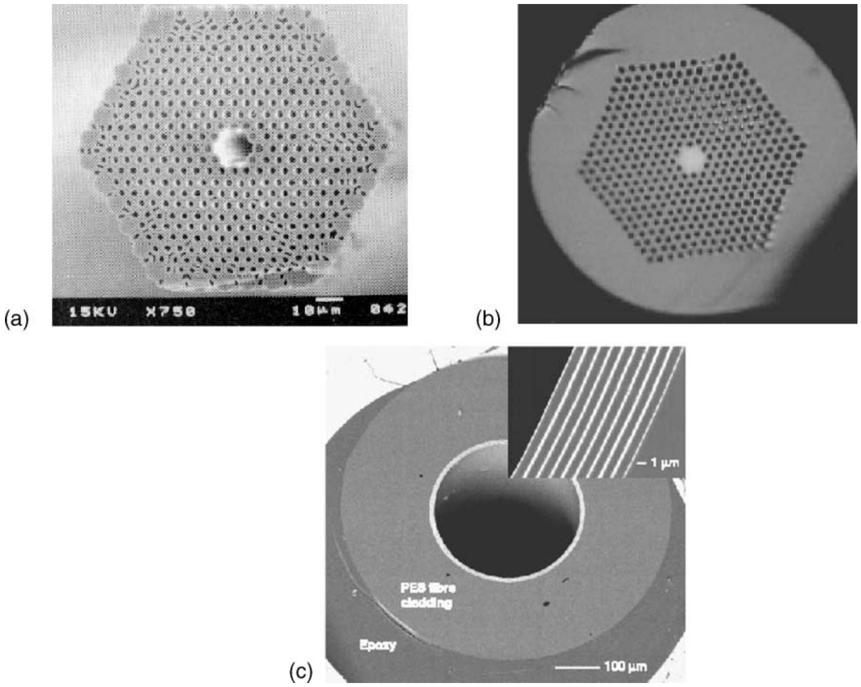

Figure 13 Photographs and drawing of three types of bandgap fibers: (a) triangular array cladding with an air core (Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* vol. 285), (b) tunable with a silica core, (c) concentric rings with an air core (Reprinted with permission from Temelkuran *et al.* *Nature* 420. Copyright 1999 American Association for the Advancement of Science).

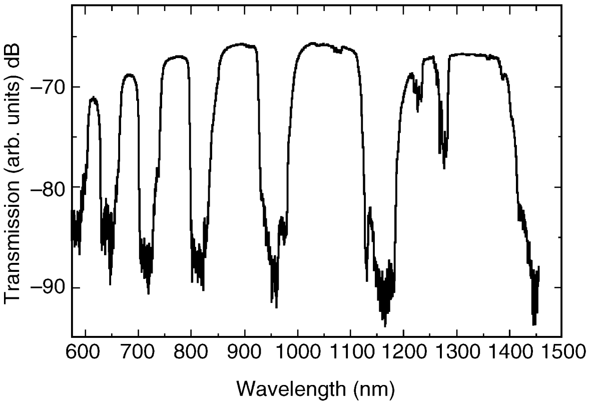

Figure 14 Spectrum of a typical bandgap fiber.

amplifiers and lasers. A double-clad amplifier or laser (also known as cladding pumped amplifier or laser) contains a double-clad fiber with a rare earth doped core. High-power pump light, launched into and guided within the inner cladding, excites the rare earth ions as the pump light crosses through the core. The excited ions, which emit light with the same wavelength and phase as the signal, amplify the signal. The fiber's efficiency and the rare earth inversion level increase with the amount of pump light absorbed by the core. The light absorbed by the core, in turn, increases with the core-to-inner cladding area ratio and the pump light intensity in the inner cladding.

Traditional fibers have a practical limit to their core-to-cladding ratio. The maximum core diameter is limited by bend loss. And, while ultimately the minimum fiber diameter is limited by draw capabilities, in practice it is limited by the ability to couple pump light into a fiber having a limited inner-cladding NA. For a pump source of given brightness, the product of beam diameter and NA cannot be increased. Such light can be successfully coupled into a larger inner cladding with smaller NA or a smaller inner cladding with larger NA.

Air-clad fibers have other advantages over traditional double-clad fibers. The outer glass cladding in air-clad fibers is optically inactive. This allows the inner cladding diameter to be decoupled from the physical fiber diameter, eliminating draw-related constraints on minimum inner cladding size. Since only the inner cladding is optically active, its size can be optimized (made smaller) while keeping the total fiber diameter at the standard 125 microns. In addition, the pump light does not interact with the polymer coating, which can be important if the polymer properties are affected by the surroundings.

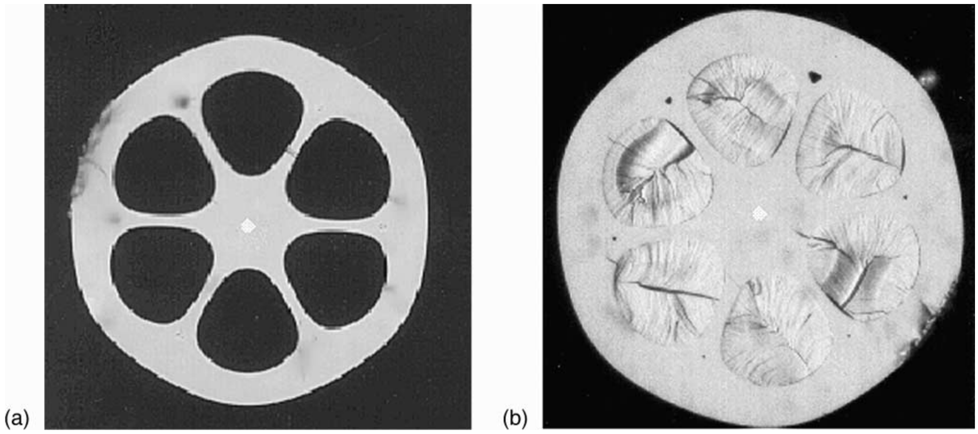

Figure 15 Holes in the microstructure fibers are filled with controllable materials to affect the fiber properties actively. Photos show (a) original fiber and (b) fiber after polymer is inserted and cured in the microstructure fiber.

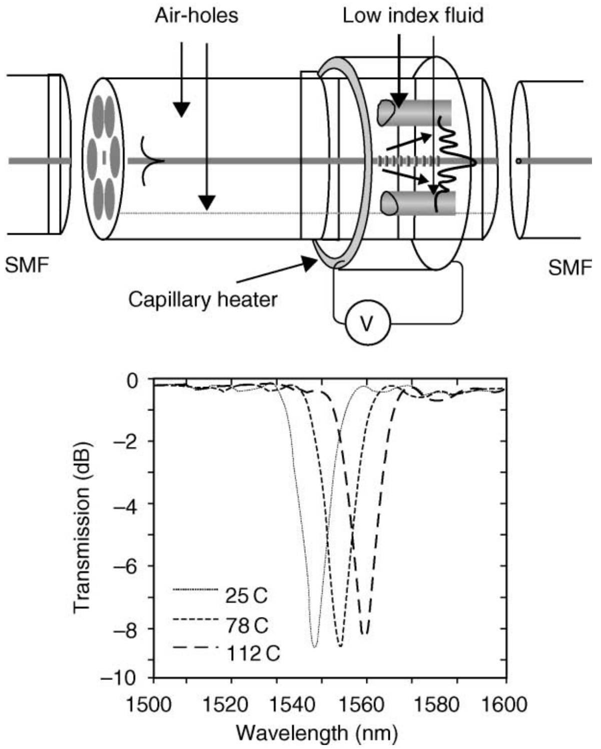

Figure 16 Schematic of a tunable device in which low index liquid is positioned over the grating. The plot shows the shift in the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184, with permission from Elsevier).

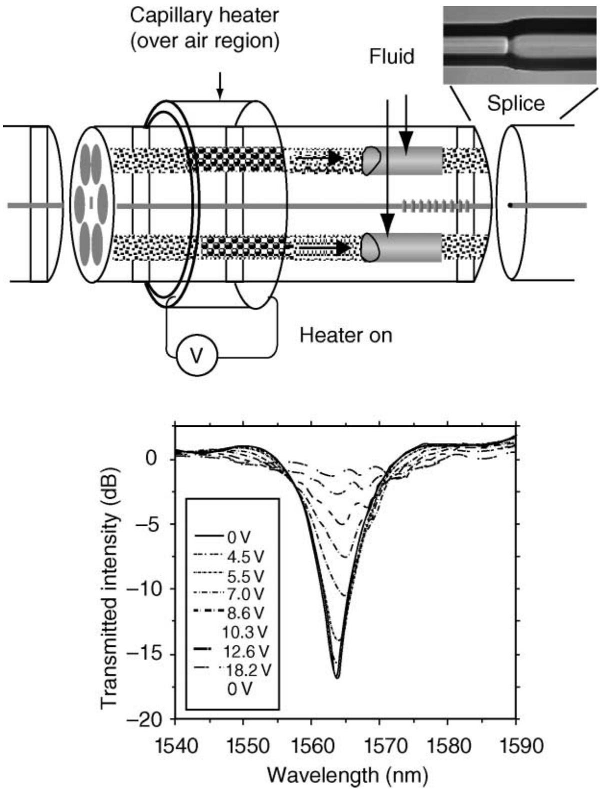

Figure 17 Schematic of a tunable device in which a high index liquid plug is moved over the grating. The plot shows the intensity of the spectrum as a function of temperature (Reproduced from Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber *Optics Communications* 204: 179–184, with permission from Elsevier).

Bandgap Guided Fibers

Bandgap fibers guide light in the core (also referred to as a defect) by confining the light through constructive interference due to Bragg scattering (Figure 12). Unlike traditional fibers, this mechanism allows light to propagate in a core that has an index lower than the cladding. Bandgap fibers with an air core could theoretically be very low loss, propagate high powers, have a large effective area core, and exhibit very low nonlinearities. In addition, the fibers can be used in novel devices by filling the core with special gasses or liquids for nonlinear processes.

Bandgap fibers can be generalized into three types (Figure 13). The first type of fiber has an air core surrounded by a cladding that consists of a periodic array of air holes in silica. Fibers of this type have been designed to guide a single mode and hold the greatest promise of guiding light over long distances with very low loss. The second type of bandgap fiber consists of a silica core and a silica cladding with a periodic array of holes which are filled with a high-index, tunable liquid. This design allows the bandgap properties to be tuned and is of interest for use in devices. The third fiber type consists of rings of alternating high- and low-index dielectric material with a center air core. They typically have a large, multimode core which is ideal for sending high powers. In addition, very little light is in the dielectric materials so the fiber can be designed to guide any wavelength with relatively little loss.

The spectrum of a bandgap fiber is quite different than an index guided fiber (Figure 14). Frequencies that lie within the bandgap cannot propagate in the cladding and are confined to the defect in the lattice (the core). Bandgap fibers with ideal geometry are predicted to have losses over an order of magnitude lower than can be obtained in an ideal traditional fiber. Although the loss of bandgap fibers has not been demonstrated to be less than traditional fiber at telecommunication wavelengths, bandgap fibers have been shown to have much lower loss than traditional fibers at other wavelengths.

Tunable Microstructure Fiber Devices

Unique tunable devices are made by filling the air holes of microstructure fibers with materials whose properties can be controlled actively (Figure 15). The materials can be positioned in the core or cladding to affect the corresponding modes. Most commonly, the index of refraction is changed using temperature sensitive liquids or polymers, although

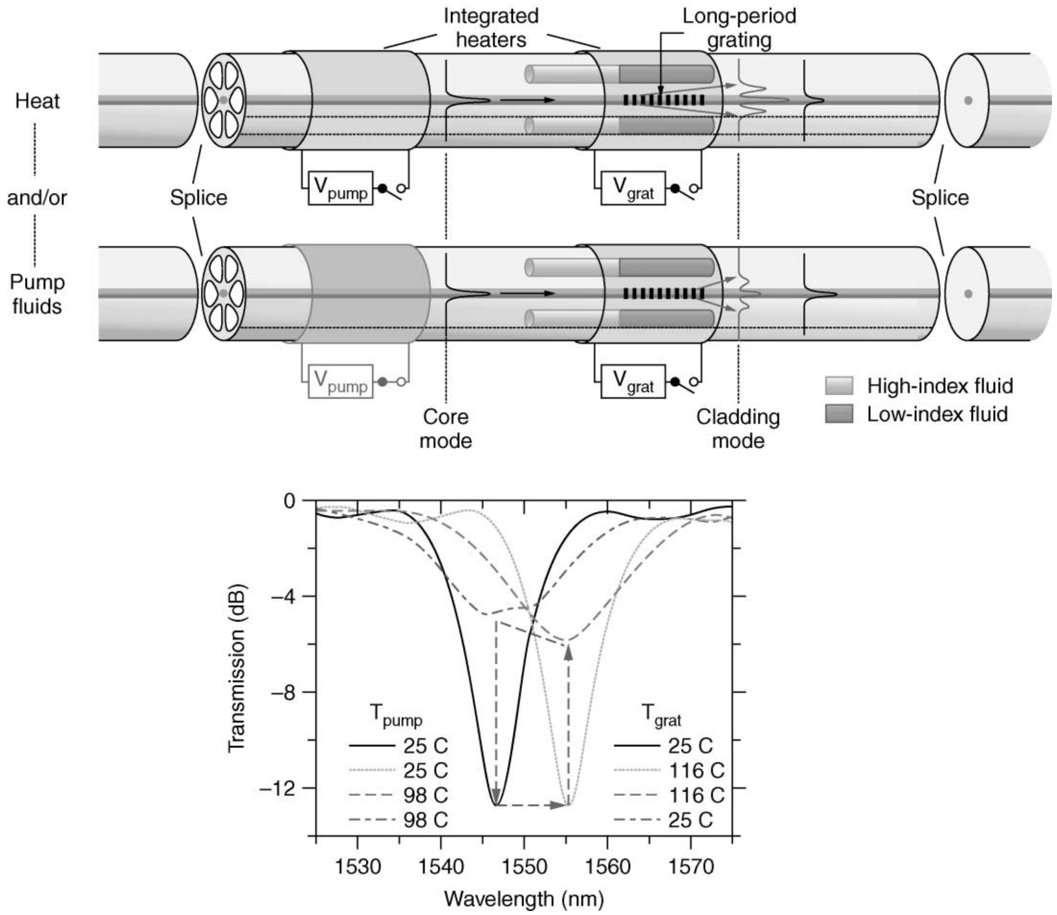

Figure 18 Schematic of a tunable device using both high and low index liquids with two heaters. The plot shows that the position and strength of the spectrum are adjusted independently (Reproduced with permission from Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* 80(23). Copyright (2002) by the American Physical Society).

materials that respond to electric or magnetic fields enable a quicker response time. These material-filled active fibers can be used to fabricate devices like tunable filters, switches, broadband attenuators, and fibers with tunable birefringence. Below, a few examples of tunable filters are presented in which the device properties depend on the type and position of the liquid inserted in the fiber holes.

A microstructure tunable filter works by controlling the spectral position, shape, or strength of a longitudinal long-period grating written in the fiber. A grating is written in the fiber before filling the holes with tunable liquid. The position of the peak absorption wavelength is controlled by placing a liquid, with an index close to silica and strongly temperature dependent, in air holes in the cladding region over the grating (Figure 16). The cladding index is tuned by varying the temperature of the liquid, using a small thin film heater located over the liquid. The position of the peak absorption wavelength moves as the cladding index is changed.

The strength or depth of the filter is controlled by inserting a moveable plug of high-index fluid in a sealed air section of a microstructure fiber (Figure 17). The strength of the grating is determined by the fraction of the grating surrounded by high index fluid. The position of the liquid plug relative to the grating can be finely adjusted with a thin film heater located over the air region adjacent to the material plug. As heat is applied to the air, the air expands and pushes the liquid plug over the grating, decreasing the coupling of the fundamental mode to higher order cladding modes. Figure 18 shows an example in which the position and strength of a filter are varied independently using two heaters and two adjacent materials.

See also

Photonic Crystals: Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Birks TA, Knight JC and Russell P (1997) Endlessly single mode photonic crystal fiber. *Optics Letters* 22(13).

Bjarklev A, Broeng J and Bjarklev AS, *Photonic Crystal Fibers*, Boston: Kluwer Academic Publishers.

Broeng J, Mogilevstev D, Bjarklev A and Barkou SE (1999) Photonic crystal fibers: A new class of optical waveguides. *Optical Fiber Technology*, vol. 5.

Chandalia JK, Eggleton BJ, Windeler RS, Kosinski SG, Liu X and Xu C (2001) Adiabatic coupling in tapered air-silica microstructures optical fiber. *IEEE Photonics Tech. Letters* 13(1).

Cregan RF, Mangan BJ, Knight JC, Birks TA, Russell PJ, Roberts PJ and Allan DC (1999) Single-mode photonic band gap guidance of light in air. *Science* 285.

Eggleton BJ, Westbrook PS, White CA, Kerbage C, Windeler RS and Burdge GL (2000) Cladding-mode resonance in air-silica microstructure optical fiber. *Journal Lightwave Technology* 18(8).

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. New Jersey: Princeton University Press.

Johnson SG, Ibanesc M, Skorobogatiy M, Weisberg M, Engeness O, Soljacic M, Jacobs SA, Joannopoubs JD and Fink Y (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9(13).

Kerbage C, Hale A, Yablon A, Windeler RS and Eggleton BJ (2001) Integrated all fiber variable attenuator based on hybrid microstructure fiber. *Applied Physics Letters* 79(19).

Kerbage C, Windeler RS, Eggleton BJ, Dolinskoi M, Mach P and Rogers JA (2002) Tunable devices based on dynamic positioning of micro-fluids in microstructure optical fiber. *Optics Communications* 204: 179–184.

Knight JC, Birks TA, Russell PJ and Atkin DM (1996) All-silica single-mode optical fiber with photonic crystal cladding. *Optics Letters* 21(19).

Knight JC, Birks TA and Russell PJ (2001) 'Holey' silica fibers. In: Markel VA and George TF (eds). *Optics of nanostructured materials*. New York: John Wiley & Sons.

Mach P, Dolinski M, Baldwin KW, Rogers JA, Kerbage C, Windeler RS and Eggleton BJ (2002) Tunable microfluidic optical fiber. *Applied Physics Letters* vol. 80.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optics Letters* 25(1).

Ranka JK, Windeler RS and Stentz AJ (2000) Optical properties of high-delta air-silica microstructure optical fibers. *Optics Letters* 25: 796–798.

Omnidirectional Surfaces and Fibers

S Hart, G Benoit and Y Fink, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Omnidirectional Reflecting Mirrors

Mirrors are probably the most prevalent of optical devices. Known to the ancients and used by them as objects of worship and beauty, mirrors are currently employed for imaging, solar energy collection, and also in laser cavities. Their intriguing optical properties have captured the imagination of scientists as well as artists and writers.

One can distinguish between two types of mirrors, the age-old metallic, and the more recent multilayer dielectric. Metallic mirrors reflect light over a broad range of frequencies incident from arbitrary angles (i.e., omnidirectional reflectance). However, at infrared and optical frequencies a few percent of the incident power is typically lost due to absorption.

Multilayer dielectric mirrors are used primarily to reflect a narrow range of frequencies incident from a particular angle or particular angular range. Unlike their metallic counterparts, dielectric reflectors can be extremely low loss. The ability to reflect light of arbitrary angle of incidence for all-dielectric structures has been associated with the existence of a complete photonic bandgap, which can exist only in a system with a dielectric function that is periodic along three orthogonal directions. In fact, a sufficient condition for the achievement of omnidirectional reflection in a periodic system with an interface, is the existence of an overlapping bandgap regime in phase space above the light cone of the ambient media.

Consider a system that is constructed of alternating dielectric layers coupled to a homogeneous medium – characterized by n_0 (such as air with $n_0 = 1$), at the interfaces. Electromagnetic waves are incident upon the multilayer film from the homogeneous medium. While such a system was analyzed extensively in the literature the possibility of omnidirectional reflectivity was not recognized until recently.

The generic system is described by the index of refraction profile in Figure 1, where h_1 and h_2 are the layer thickness, and n_1 and n_2 are the indices of refraction of the respective layers. The incident wave has a wavevector $\vec{k} = k_x\hat{e}_x + k_y\hat{e}_y$, and frequency of $\omega = c|k|$. The wavevector together with the normal to the periodic structure defines a mirror plane of symmetry which allows us to distinguish between two independent electromagnetic modes: transverse electric (TE) modes and transverse magnetic (TM) modes. For the TE mode the electric field is perpendicular to the plane, as is the magnetic field for the TM mode.

General features of the transport properties of the finite structure can be understood when the properties of the infinite structure are elucidated. In a structure with an infinite number of layers, translational symmetry along the direction perpendicular to the layers leads to Bloch wave solutions of the form:

$$u_K(x, y) = E_K(y)\, e^{iKy}\, e^{ik_x x} \qquad [1]$$

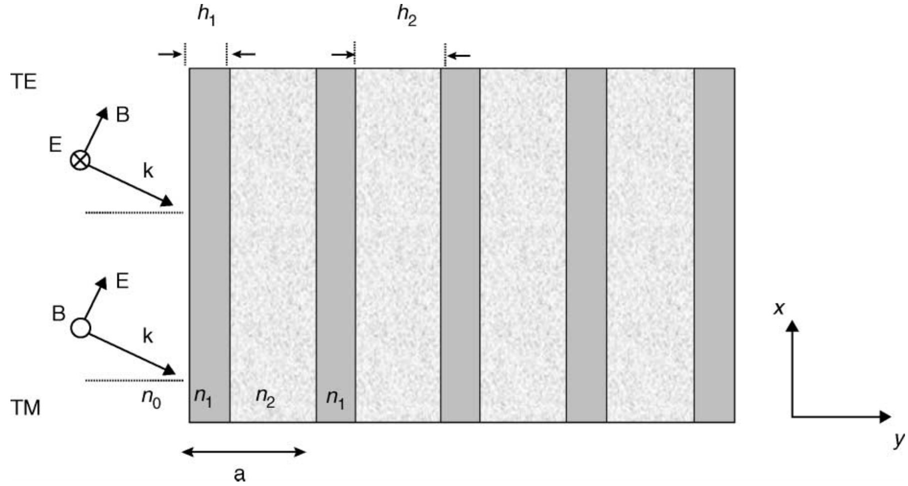

Figure 1 Schematic of the multilayer system showing the layer parameters (n_α, h_α – index of refraction and thickness of layer α), the incident wavevector $\vec{k}$ and the electromagnetic mode convention.

where $E_K(y)$ is periodic, with a period of length a, and K is the Bloch wave number. These waves represent solutions to an eigenvalue problem and are completely and uniquely defined by the specification of K, k_x and ω.

Solutions can be propagating or evanescent, corresponding to real or imaginary Bloch wave numbers, respectively. It is convenient to display the solutions of the infinite structure by projecting the $\omega(K, k_x)$ function onto the $\omega - k_x$ plane; Figure 2a,b are examples of such projected structures.

The gray background areas highlight phase space where K is strictly real, that is, regions of propagating states, while the white areas represent regions containing evanescent states. The shape of the projected bandstructures for the multilayer film can be understood intuitively. At $k_x = 0$, the bandgap for waves traveling normal to the layers is recovered. For $k_x > 0$, the bands curve upward in frequency. As $k_x \to \infty$, the modes become largely confined to the slabs with the high index of refraction and do not couple between layers (and are therefore independent of k_x).

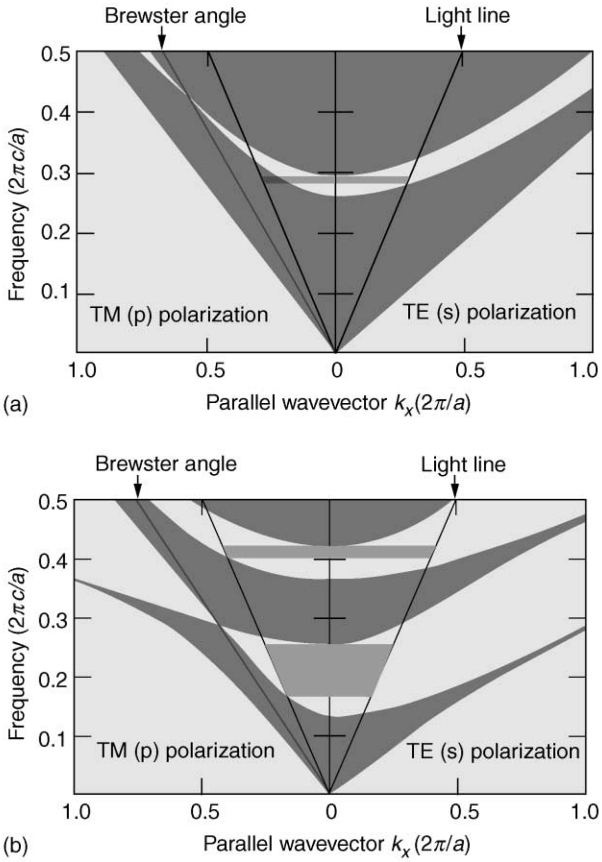

Figure 2 (a) Projected bandstructure of a multilayer film with the light line and Brewster line, exhibiting a reflectivity range of limited angular acceptance with ($n_1 = 2.2$ and $n_2 = 1.7$ and a thickness ratio of $h_2/h_1 = 2.2/1.7$). (b) Projected bandstructure of a multilayer film together with the light line and Brewster line, showing an omnidirectional reflectance range at the first and second harmonic (propagating states – dark gray, evanescent states – white, omnidirectional reflectance range – light grey). The film parameters are $n_1 = 4.6$ and $n_2 = 1.6$ with a thickness ratio of $h_2/h_1 = 1.6/0.8$. These parameters are similar to the actual polymer–tellurium film parameters measured in the experiment. Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

In a finite structure, the translational symmetry in the directions parallel to the layers is preserved, hence k_x remains a conserved quantity and can be used to label solutions even though these solutions will no longer be of the Bloch form. The relevance of the band diagram to finite structures is that it allows for the prediction of regions of phase space where waves are evanescently decaying in the multilayer structure.

Since we are primarily interested in waves originating from the homogeneous medium external to the periodic structure, we will focus only on the portion of phase space lying above the light line. Waves originating from the homogeneous medium satisfy the condition $\omega \geq ck_x/n_0$, where n_0 is the refractive index of the homogeneous medium, and therefore they must reside above the light line. States of the homogeneous medium with $k_x = 0$ are normal incident, and those lying on the $\omega = ck_x/n_0$ line with $k_y = 0$ are incident at an angle of 90°.

The states in Figure 2a, that are lying in the restricted phase space defined by the light line and that have a (ω, k_x) corresponding to the propagating solutions (gray areas) of the crystal can propagate in both the homogeneous medium and in the structure. These waves will partially or entirely transmit through the film. Those with (ω, k_x) in the evanescent regions (white areas) can propagate in the homogeneous medium but will decay in the crystal – waves corresponding to this portion of phase space will be reflected off the structure.

The multilayer system leading to Figure 2a represents a structure with a limited reflectivity cone, since for any frequency one can always find a k_x vector for which a wave at that frequency can propagate in the crystal – and hence transmit through the film. The necessary and sufficient criterion for omnidirectional reflectivity at a given frequency is that there exists no transmitting state of the structure inside the light cone – this criterion is satisfied by frequency ranges marked in light gray in Figure 2b. In fact, the system leading to Figure 2b exhibits two omnidirectional reflectivity ranges.

The omnidirectional range is defined from above by the normal incidence bandedge $\omega_h(k_y = \pi/a, k_x = 0)$ (Figure 2b) and below by the intersection of the top of the TM allowed bandedge with the light line $\omega_l(k_y = \pi/a, k_x = \omega_l/c)$ (Figure 2b).

A dimensionless parameter used to quantify the extent of the omnidirectional range is the range to mid-range ratio defined as $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$.

Figure 3 is a plot of this ratio as a function of n_2/n_1 and n_1/n_0, where ω_h and ω_l are determined by solutions of eqn [2] with quarter wave layer thickness. The contours in this figure represent various equi-omnidirectional ranges for different material index parameters and could be useful for design purposes.

At normal incidence, there is no distinction between TM and TE modes. At increasingly oblique angles, the gap of the TE mode increases, whereas the gap of the TM mode decreases. In addition, the center of the gap shifts to higher frequencies. Therefore, the criterion for the existence of omnidirectional reflectivity can be restated as the occurrence of a frequency overlap between the gap at normal incidence and the gap of the TM mode at 90°. Analytical expressions for the range to mid-range ratio can be obtained by setting:

$$\omega_h = \frac{2c}{h_2 n_2 + h_1 n_1} \cos^{-1}\left(- \left| \frac{n_1 - n_2}{n_1 + n_2} \right| \right)$$

$$\omega_l = \frac{2c}{h_2\sqrt{n_2^2 - 1} + h_1\sqrt{n_1^2 - 1}} \times \cos^{-1}\left(\left| \frac{n_1^2\sqrt{n_2^2 - 1} - n_2^2\sqrt{n_1^2 - 1}}{n_1^2\sqrt{n_2^2 - 1} + n_2^2\sqrt{n_1^2 - 1}} \right| \right) \quad [2]$$

Moreover, the maximum range width is attained for thickness values that are not equal to the quarter wave stack though the increase in bandwidth gained by deviating from the quarter wave stack is typically only a few percent.

In general, the TM mode defines the lower frequency edge of the omnidirectional range; an example can be seen in Figure 2b for a particular choice of the indices of refraction. This can be proven by showing that:

$$\left.\frac{\partial \omega}{\partial k_y}\right|_{\mathrm{TM}} \geq \left.\frac{\partial \omega}{\partial k_y}\right|_{\mathrm{TE}} \quad [3]$$

in the region that resides inside the light line. The physical reason for eqn [3] lies in the vectorial nature of the electric field. In the upper portion of the first band the electric field concentrates its energy in the high dielectric regions. Away from normal incidence, the electric field in the TM mode has a component in the direction of periodicity, this component forces a larger portion of the electric field into the low dielectric regions. The group velocity of the TM mode is therefore enhanced. In contrast, the electric field of the TE mode is always perpendicular to the direction of periodicity and can concentrate its energy primarily in the high dielectric region. While the omnidirectional reflection criteria can be used to confine light in various geometries, it is the cylindrical fiber configuration that appears to present significant application opportunities.

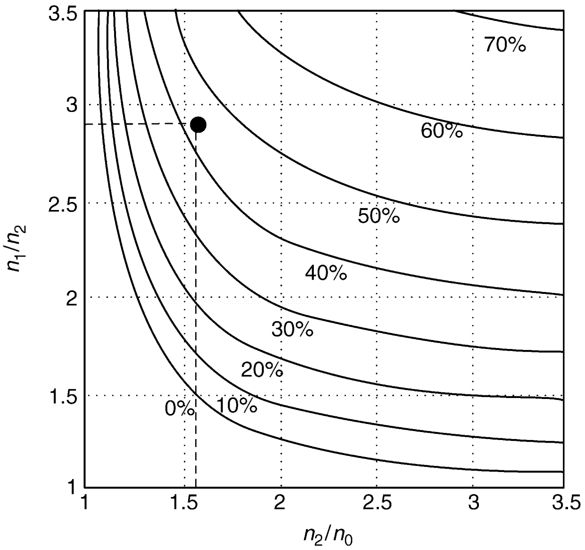

Figure 3 The range to mid-range ratio $(\omega_h - \omega_l)/\frac{1}{2}(\omega_h + \omega_l)$, for the fundamental frequency range of omnidirectional reflection, plotted as contours. Here, the layers were set to quarter wave thickness and $n_1 > n_2$. The ratio for our materials is approximately 45%. ($n_1/n_2 = 2.875$, $n_2/n_0 = 1.6$) is located at the intersection of the dashed lines (black dot). Reproduced with permission from Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682. Copyright 1998 American Association for the Advancement of Science.

Mesostructured Fibers for External Reflection Applications

Polymer fibers are ubiquitous in applications such as textile fabrics, due to their excellent mechanical properties and the availability of low-cost, high-volume processing techniques; however, the control over their optical properties has so far remained relatively limited. Conversely, dielectric mirrors are used to precisely control and manipulate light in high performance optical applications, but the fabrication of these typically fragile mirrors has been mostly restricted to planar geometries and typically involves multiplicity of deposition sequences in a high vacuum thin film deposition system.

Fabrication Approach

The fabrication of extended lengths of omnidirectionally reflecting fibers with large bandgaps and

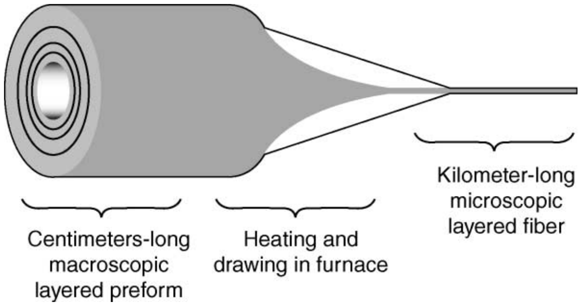
Centimeters-long macroscopic layered preform
Heating and drawing in furnace
Kilometer-long microscopic layered fiber

Figure 4 Conceptual preform based fabrication process for meso-structured fibers.

high layer counts poses considerable challenges. To illustrate the nature of this formidable task, one needs to consider the necessity of maintaining the uniformity of sub-100 nm layer thicknesses over kilometer-length scales; creating continuous layers with an aspect ratio of $\sim 10^{10}$ in a single process! To meet this and other challenges associated with the fabrication of mesostructured fibers, we have developed a preform-based fabrication approach (Figure 4). A scaled-up version of the final fiber, called a preform, is fabricated which shares the geometry and materials of the final fiber but exhibits macroscopic lateral features. The preform is heated up and drawn under tension into the fiber using a simple cylindrical furnace. The macroscopic layers are reduced in the process to microscopic dimensions while maintaining the overall geometry and symmetry of the original preform. Conservation of mass determines the final length of the fiber – typically this length is equal to the lateral reduction factor squared multiplied by the length of the preform. The nature of the process and requirements on the fiber's optical properties lead to the definition of materials selection rules. First, in order to achieve omnidirectional reflectivity, one needs to identify two solid materials exhibiting an index contrast given by the plot in Figure 3. Second, to enable the codrawing of the two dissimilar materials both will need to have viscosities that are lower than $\sim 10^8$ poise at the drawing temperature. In order to maintain high draw speeds, the majority component needs to be amorphous and the adhesion between these two materials needs to be sufficient to prevent delamination. Finally, their thermal expansion coefficients need to be close or alternatively at least one of the materials needs to be capable of relieving the stress due to CTE mismatch.

Materials Selection Criteria

Pairs of materials that are compatible with the process and fiber property requirements have been identified, including: high glass transition temperature (T_g), thermoplastic polymers such as poly(ether-sulfone) (PES), poly (ether-imide) (PEI) and members of the chalcogenide glass family, such as arsenic triselenide (As_2Se_3) or arsenic trisulfide. Pairs of these materials will have substantially different refractive indices, as shown in Figure 5, which is a measurement of the real and imaginary indices of refraction of PES and As_2Se_3, obtained using a broadband spectroscopic ellipsometer (SOPRA GES5). They nevertheless exhibit similar thermo-mechanical properties within a certain thermal processing window.

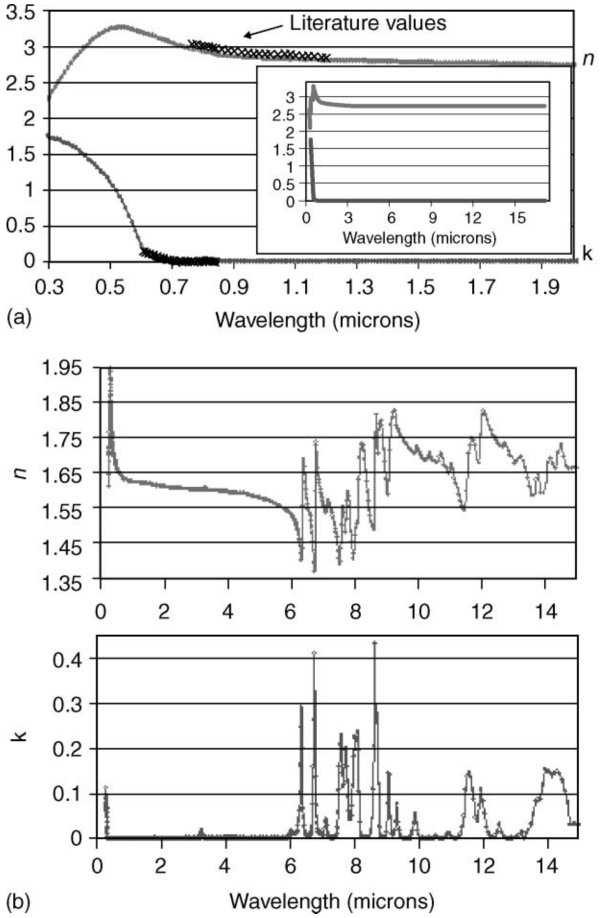

Figure 5 (a) Real part (*n* – light gray) and imaginary part (*k* – dark gray) of the refractive index of annealed As_2Se_3. The black crosses correspond to literature values. (b) Real part (*n* – upper) and imaginary part (*k* – lower) of the refractive index of PES.

Adhesion and extensional viscosity in the fluid state are difficult to measure in general, and the measurement of high-temperature surface tension is quite involved. Thus, limited data on these properties are available and it was necessary to empirically identify materials that could be used to draw out mirror-fibers. Various high-index chalcogenide (S, Se, and Te containing) glasses and low-index polymers were identified as potential candidates based on their optical properties and overlapping thermal softening regimes. Adhesion and viscosity matching were tested by thermal evaporation of a chalcogenide glass layer on top of a polymer film or rod and elongation of the coated substrate at elevated temperatures. The choice of a high-temperature polymer, PES, and a simple chalcogenide glass, As_2Se_3, resulted in excellent thermal co-deformation without film cracking or delamination. Approximate matching of extensional viscosity in this manner was also demonstrated using As_2Se_3 and PEI. The properties, processing, and applications of chalcogenide glasses have been explored extensively elsewhere. One advantage in choosing As_2Se_3 for this application is that not only is it a stable glass, but it is a stoichiometric compound that can be readily deposited in thin films through thermal evaporation or sputtering without dissociation. Additionally, As_2Se_3 is transparent to IR radiation from approximately 0.8 to 17 μm as shown in the ellipsometric data (Figure 5a), and has a refractive index of ~2.8 in the mid-IR. PES is a high-performance, dimensionally stable thermoplastic with a refractive index of ~1.6 and good transparency to EM waves in a range extending from the visible regime into the mid-IR (~5 μm), as shown in Figure 5b.

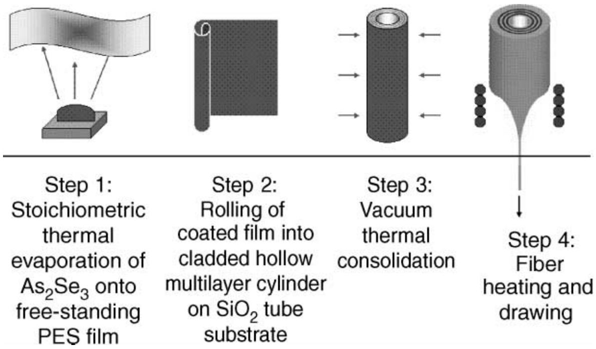
Step 1: Stoichiometric thermal evaporation of As_2Se_3 onto free-standing PES film
Step 2: Rolling of coated film into cladded hollow multilayer cylinder on SiO_2 tube substrate
Step 3: Vacuum thermal consolidation
Step 4: Fiber heating and drawing

Figure 6 Multilayer preform fabrication sequence.

Preform Construction Process and Fiber Draw

The selected materials were used to construct a multilayer preform rod, which essentially is a macroscale version of the final fiber. In order to fabricate the dielectric mirror fiber preform, an As_2Se_3 film was deposited through thermal evaporation on either side of a free-standing PES film which was then rolled on top of a PES tube substrate, forming a structure having 21 alternating layers of PES and As_2Se_3, as shown in Figure 6.

The resulting multilayer fiber preform was subsequently drawn down using an optical fiber draw tower into hundreds of meters of multilayer fiber with a precisely controlled submicron layer thickness, creating a photonic bandgap in the mid-IR. Fibers of outer diameters varying from 175–500 μm with a typical standard deviation of 10 μm from target, were drawn from the same preform to demonstrate adjustment of the reflectivity spectra through thermal deformation. The spectral position of the photonic bandgap was controlled by the optical monitoring of the outer diameter (OD) of the fiber during draw, which was later verified by reflectivity measurements on single and multiple fibers of different diameters. Scanning electron micrographs (SEMs) of the cross-section of these fibers are depicted in Figure 7.

Bandstructure for Multilayer Fibers for External Reflection Applications

In theoretically predicting the spectral response of these fibers, it is helpful to calculate the photonic

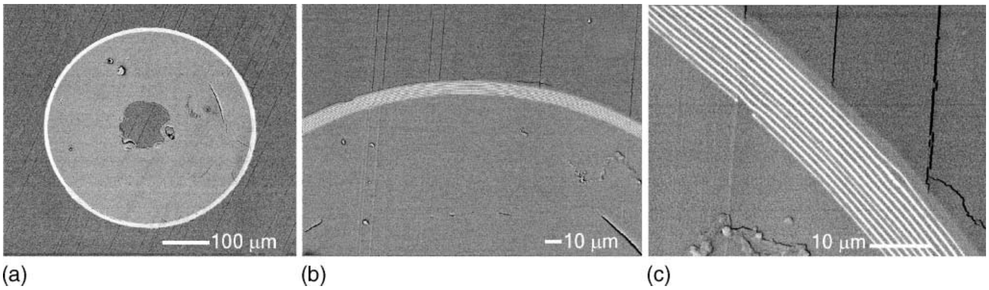

Figure 7 SEM micrographs of 400 μm OD fiber cross-section. The entire fiber is embedded in epoxy. (a) shows the entire fiber cross-section, with mirror structure surrounding the PES core; (b) demonstrates that the majority of the fiber exterior is free of significant defects and that the mirror structure adheres well to the fiber substrate; and (c) reveals the ordering and adhesion within the alternating layers of As_2Se_3 (bright layers) and PES. Stresses developed during sectioning caused some cracks in the mounting epoxy that are deflected at the fiber interface. Fibers from this batch were used in the reflectivity measurements recorded below in Figure 9a. Reproduced with permission from Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513. Copyright 2002 American Association for the Advancement of Science.

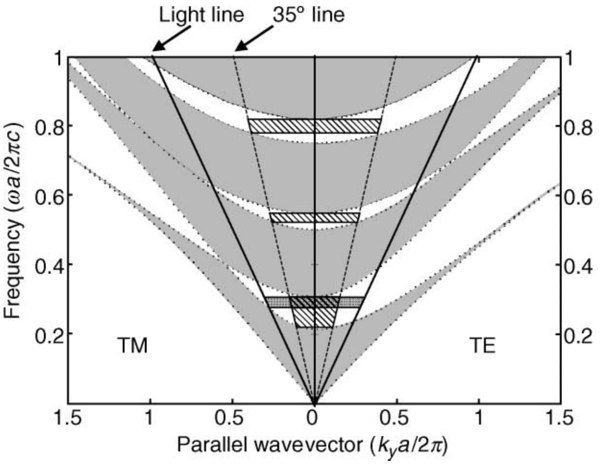

Figure 8 Photonic band diagram for a one-dimensional photonic crystal having a periodic refractive index alternating between 2.8 and 1.55. Gray regions represent propagating modes within the structure, while white regions represent evanescent modes. Hatched regions represent photonic bandgaps where high reflectivity can be expected for external EM waves over an angular range extending from normal to 35° incidence. The lower dark shaded trapezoid represents a region of external omnidirectional reflection.

bandstructure that corresponds to an infinite one-dimensional photonic crystal. This allows for the analysis of propagating and evanescent modes in the structure, corresponding to real or imaginary Bloch wave number solutions.

The electric or magnetic field vector is parallel to the mirror layer interfaces for the TE and TM polarized modes, respectively. The parallel wavevector (k_y) is the component of the incident electromagnetic (EM) wavevector that is parallel to the layer interfaces. The phase space accessible from an external ambient medium is contained between the light lines (defined by the glancing-angle condition $\omega = ck_y/n_0$), and the modes between the 35° lines correspond to those sampled experimentally. Axes are normalized to the thickness a of one mirror bilayer (a period consisting of one high- and one low-index layer). Figure 8 depicts the photonic band diagram for an infinite structure having similar periodicity and refractive indices to the mirror structures fabricated here. Three photonic bandgaps are present where high reflectivity is expected within the 0–35° angular range, and the fundamental gap contains a region of external omnidirectional reflectivity.

Optical Characterization of 'Mirror Fibers'

Mirror fiber reflectivity was measured from both single fibers and parallel fiber arrays using a Nicolet/SpectraTech NicPlan Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860). The microscope objective (SpectraTech 15 ×, Reflachromat) used to focus on the fibers had a numerical aperture (NA) of 0.58. This results in a detected cone where the angle of reflection with respect to the surface normal of the structure could vary from normal incidence to ~35°, which is determined by the NA of the microscope objective. As a background reference for the reflection measurements, we used gold-coated PES fibers of matching diameters. Dielectric mirror fibers, drawn to 400 μm OD, exhibited a very strong reflection band centered at 3.4 μm wavelength (Figure 9a). Measured reflectivity spectra agree well with planar-mirror transfer matrix method (TMM) simulations, where the reflectivity was averaged across the aforementioned angular range for both polarization modes. Fibers drawn down to 200 μm OD show a similar strong fundamental reflection band centered near 1.7 μm (Figure 9b). This shifting of the primary photonic bandgap clearly illustrates the precise tuning of the reflectivity spectra over wide frequency ranges through thermal deformation processing. Strong optical signatures are measurable from single fibers as small as 200 μm OD. Fiber array measurements,

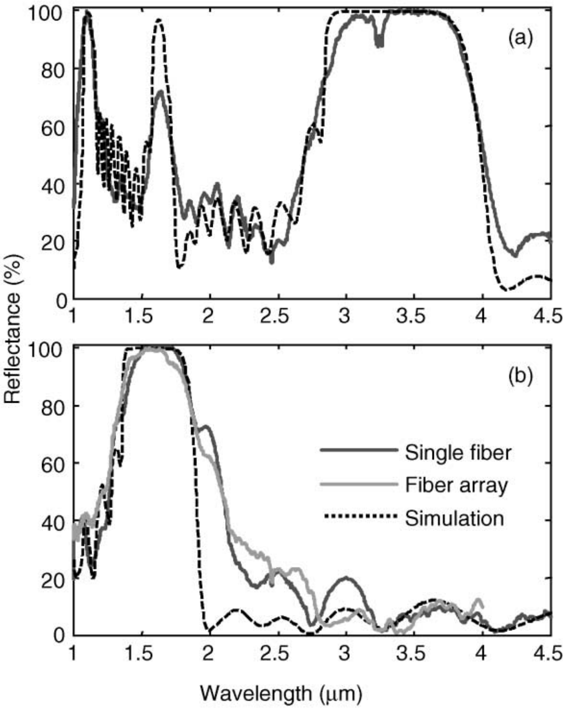

Figure 9 Measured reflectance spectra for 400 μm OD (a) and 200 μm OD (b) dielectric mirror fibers relative to gold-coated fibers of the same diameter. (a) shows a single-fiber reflectivity measurement, while (b) compares single-fiber reflectivity to that measured from a multifiber array. Simulations were performed using the transfer matrix method.

simultaneously sampling reflected light from multiple fibers, agree quite well with single-fiber data (Figure 9b).

These reflectivity results are strongly indicative of uniform layer thickness control, good interlayer adhesion, and low interdiffusion through multiple thermal treatments. This was confirmed by SEM inspection of fiber cross-sections (Figure 7). The layer thicknesses observed ($a = 0.90\ \mu\text{m}$ for the 400 μm fibers; $a = 0.45\ \mu\text{m}$ for the 200 μm fibers) correspond well to the measured reflectivity spectra. The fibers have a hole in the center, due to the choice of a hollow rod as the preform substrate, which experienced some nonuniform deformation during draw. The rolled-up mirror structure included a double outer layer of PES for mechanical protection, creating a noticeable absorption peak in the reflectivity spectrum at ~3.2 μm (Figure 9a).

A combination of spectral and direct imaging data demonstrates excellent agreement with the photonic band diagram. The measured gap width (range to mid-range ratio) of the fundamental gap for the 400 μm OD fiber is 27%, compared to 29% in the photonic band diagram.

Tunable ‘Fabry–Perot’ Fibers

The fabrication of fibers surrounded by or lined with alternating layers of materials with a large disparity in their refractive indices, presents interesting opportunities for passive and active optical devices. While a periodic multilayer structure, such as the one reported above, leads to the formation of photonic bandgaps and an associated range of high reflectivity, it is the incorporation of intentional deviations from periodicity, also called ‘defects’, which allows for the creation of localized electromagnetic modes in the vicinity of the defect. These structures, sometimes called optical cavities, in turn can provide the basis for a large number of interesting passive and active optical devices such as vertical cavity surface emitting lasers (VCSELs), bi-stable switches, tuneable dispersion compensators, tuneable drop filters, etc. Here we report on the fabrication of a fiber surrounded by a Fabry–Perot cavity structure and demonstrate that by application of axial mechanical stress, the spectral position of the resonant Fabry–Perot mode can be reversibly tuned.

Structure and Optical Properties of the Fabry–Perot Fibers

The fabrication technique described above allows for the accurate placement of optical cavities that can encompass the entire or partial fiber circumference. The fibers discussed above are made of As_2Se_3 and PES and have a low index Fabry–Perot cavity (Figure 10).

This structure was achieved by introducing an extra polymer layer in the middle of the periodic multilayer structure of the preform, thus generating a defect mode in the photonic bandgaps of the drawn fibers. The position of the bandgap center is linearly related to the optical thickness of the layers by the Bragg condition. Through accurate outer diameter control afforded by the laser micrometer mounted on the draw tower we have been able to place gaps (Figure 11) at wavelengths ranging from 11 microns to below 1 micron. Such cost-effective tuneable optical filters could lead to applications such as optical

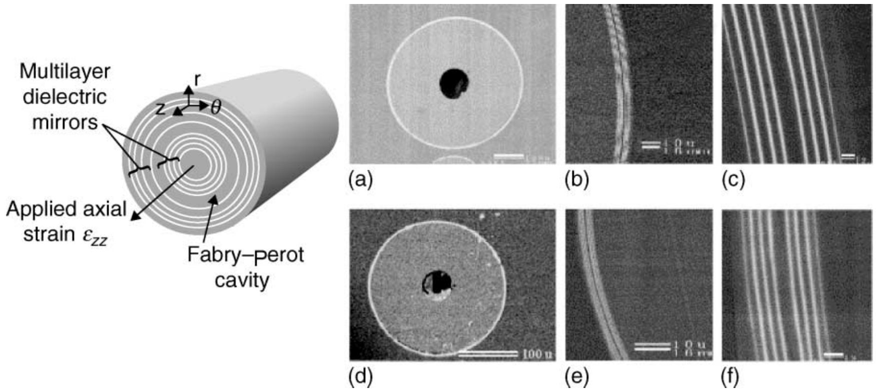
Multilayer dielectric mirrors
r
z
θ
Applied axial strain ε_{zz}
Fabry–perot cavity
(a)
(b)
(c)
(d)
(e)
(f)

Figure 10 Schematic of the structure of dielectric mirror fibers made of As_2Se_3 (light gray) and PES (dark gray) with low index Fabry–Perot cavity. The local cylindrical coordinate system is represented as well as the applied axial strain ε_{zz}. Typical radii are ~150 μm ± 100 μm. Backscattered SEM micrographs of the cross-sections of a 460 micron diameter fiber (a, b, c) and of a 240 micron diameter fiber (d, e, f) embedded in epoxy and microtomed. (a) and (d) show the entire cross-section of the fibers, (b) and (e) demonstrate long-range layer uniformity, and (c) and (f) reveal the ordering and adhesion of the Fabry–Perot cavity structure. Bar scales have been redrawn for clarity. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

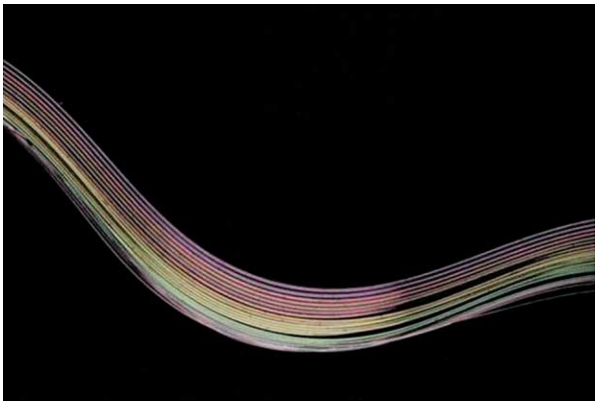

Figure 11 Array of parallel fibers with outer diameter ranging from ~420 μm (bottom) to ~100 μm (top). The colors are due to the narrow 4th order photonic bandgap in the visible.

switches for wavelength-division-multiplexing (WDM) systems and sensors.

The cross-sectional structures of a 240 micron and a 460 micron diameter fiber were observed by an SEM using a backscattered electron detector (Figure 10). The structure is composed of a hollow core polymer rod surrounded by six bilayers of As_2Se_3 and PES, separated in the middle by an extra polymer layer, which forms the Fabry–Perot cavity. An extra polymer layer protects the fiber surface. For the 240 micron (460 micron) diameter fiber, the glass layers are ~135 nm (250 nm) thick, except for the first and the last ones which are half as thick due to the fabrication technique; the polymer layers are ~270 nm (540 nm) thick and the defect layer is ~610 nm (1170 nm) thick.

Reflectivity spectra measurements were performed under a microscope using a (Nicolet SpectraTech NicPlan) Infrared Microscope and Fourier Transform Infrared Spectrometer (Magna 860) with a lens numerical aperture (NA) corresponding to 30 degrees of angular spread. They exhibited a single-mode Fabry–Perot resonant mode at 1.74 and 3.2 μm for the 240 and 460 micron diameter fiber, respectively (Figure 12).

Because of the range of incident angles, the measured quality factor (Q-factor, defined as its spectral position divided by the full width half maximum (FWHM)) was equal to 31. Using the different thicknesses and the real and imaginary part of the refractive index (n & k) of As_2Se_3 and PES carefully measured with a broadband (300 nm to 16 microns) spectroscopic ellipsometer (Sopra GES-5), the reflectivity spectra of these fibers were computed with the TMM by approximating them as a one-dimensional planar stacking. By averaging the calculated spectra over the accessible incident angles and polarizations, we obtained a very good agreement between the simulation and the measurements.

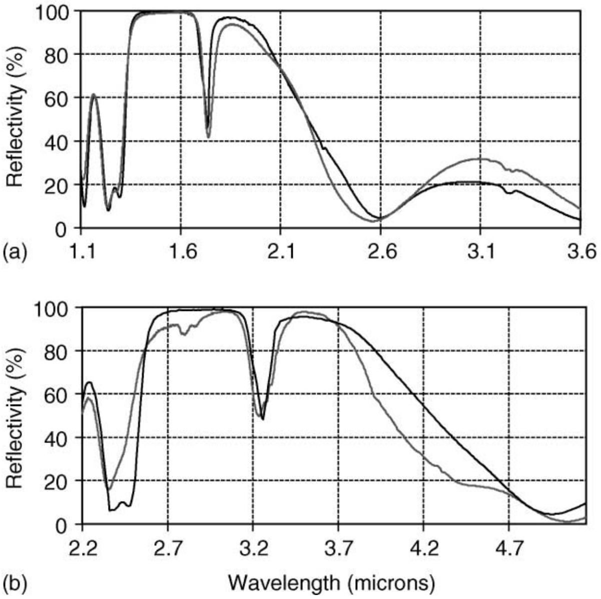

Figure 12 Computed (black lines) and measured (grey lines) reflectivity spectra for the 240 micron (a) and of the 460 micron diameter fibers (b) with Fabry–Perot resonant modes at 1.74 and 3.2 μm, respectively. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

Simulation of the Opto-Mechanical Behavior of the Fabry–Perot Fibers

We focused our analysis on the elastic regime, which ultimately limits the operational range of these fibers. The fiber's Young's modulus (E) can be approximated by modeling this multilayer structure as independent parallel springs under axial (along the z-axis of the fiber) strain assumed to be equal for all the layers, leading to a Young's modulus of 2.64 GPa (using $E_{\mathrm{PES}} = 2.4$ GPa for bulk PES and $E_{\mathrm{As_2Se_3}} = 15$ GPa reported for 1.5 μm thick films).

Neglecting the possible strain-induced refractive index variation of the materials, the normalized shift of the Fabry–Perot resonant mode can be related to the applied axial strain by calculating the radial stresses σ_{rr} (resulting from the difference between the Poisson ratios of the materials – $\nu_{\mathrm{PES}} = 0.45$ and $\nu_{\mathrm{As_2Se_3}} = 0.289$ – and the adhesion condition between the layers) and displacements u_r in each layer under axial strain. Starting from the equilibrium equations:

$$\frac{\partial \sigma_{zz}}{\partial z} = 0 \qquad [4]$$

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{\sigma_{rr} - \sigma_{\theta\theta}}{r} = 0 \Leftrightarrow \frac{\mathrm{d}^2 u_r}{\mathrm{d}^2 r} + \frac{1}{r}\frac{\mathrm{d}u_r}{\mathrm{d}r} - \frac{u_r}{r^2} = 0 \qquad [5]$$

and using Hooke's law and Lame's equations, we can derive a general expression for u_r and σ_{rr}, with two unknowns A and B per layer:

$$u_r = \frac{A}{r} + Br \quad [6]$$

$$\sigma_{rr} = -\frac{2GA}{r^2} + 2(\eta + G)B + \lambda\varepsilon_{zz} \quad [7]$$

where $G = E/[2(1+\nu)]$ is the shear modulus and $\eta = E\nu/[(1+\nu)(1-2\nu)]$ the Lame modulus. The continuity of the displacement and the radial stress at each interface, plus the boundary conditions (σ_{rr} vanishes at free surfaces) can be expressed as a linear system whose unique solution allows us to relate the radial strain in each layer to the applied axial strain as a linear relation $\varepsilon_{rr} = C\varepsilon_{zz}$, where C can be interpreted as an effective Poisson ratio. Finally, taking into account that the Fabry–Perot resonant mode itself is linearly shifted within the bandgap because of the different effective Poisson ratios of the glass and polymer layers, we obtain a linear relation between the normalized shift of the Fabry–Perot resonant mode and the applied axial strain:

$$\frac{\Delta\lambda}{\lambda} = -0.373\varepsilon_{zz} \quad [8]$$

All the layers (except the outer protective polymer layer) are under tensile radial stress whose maximum (0.22 MPa under 1% axial strain) is located at the interface between the layers and the polymer core where delamination is most likely to occur.

Mechanical Tuning Experiment and Discussion

Measurements were performed on fibers ~30 cm long, which were fixed at one end with epoxy to a load cell (Transducer Techniques MDB-2.5) while the other end was attached with strong tape to a pole mounted on a stepper rotational stage (Newport PR50) (Figure 13). This end of the fiber was also screwed to the pole to further secure it in place.

The diameter of the pole was equal to 2.1 cm, leading to a normalized shift precision below 0.005%. The uncertainty on the Young's modulus, due to the precision of the load cell, was lower than 30 MPa. All the reflectivity spectra were normalized to a background taken with a flat gold mirror. The measurements were realized as far as possible from the fixed ends of the fiber where edge effects are likely to occur. These edge effects result in a reduction of the length of the fiber that deforms uniformly and consequently increase the real strain far from the edges by a factor of 1.14 and 1.15 for the 240 and the 460 micron diameter fiber, respectively (determined experimentally by measuring the position of two reference points on the fiber) compared to the strain calculated from the rotation of the pole. Moreover, to avoid measuring slight variations in the spectral position of the bandgap resulting from outer diameter variations, typically of the order of 4–5 microns over meters of fibers, the measurements were realized at a fixed reference position on the fiber.

By focusing on the fundamental bandgap of the 240 and the 460 micron diameter fiber, we demonstrated the tuning of the Fabry–Perot resonant mode under increasing axial strain (Figure 14).

A drop in the reflectivity of 13% was observed at 1.71 μm (dash line) for the 240 micron diameter fiber when increasing the applied axial strain from 0.23% (light gray) to 1.07% (dark gray). The normalized-shift versus strain curves (Figure 15) appeared to be linear for both fibers up to approximately 0.9% (dash line) with a slope equal to −0.3859 and −0.3843 for the 240 and the 460 micron diameter fiber, respectively, close to the predicted value (−0.373, eqn [8]).

The normalized shift was equal to −0.347% for 0.9% applied axial strain, which seems to be the limit of the elastic regime and corresponds to small applied

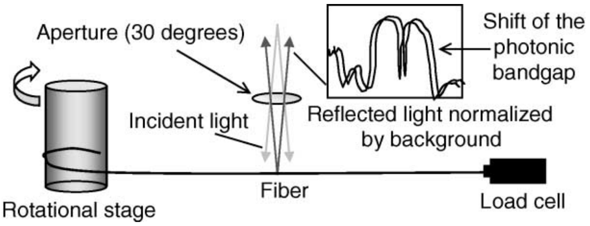
Aperture (30 degrees)
Shift of the photonic bandgap
Incident light
Reflected light normalized by background
Fiber
Load cell
Rotational stage

Figure 13 Experimental setup used for mechanical tuning demonstration. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

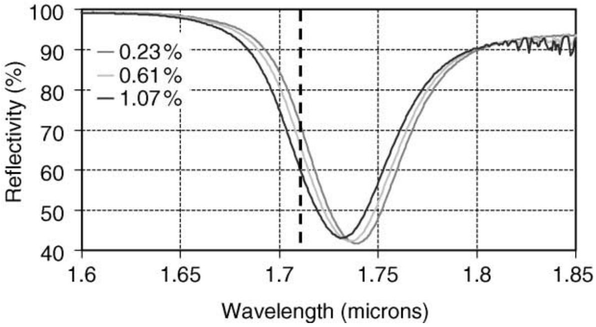

Figure 14 Reflectivity versus wavelength plot showing the shift of the Fabry–Perot resonant mode of the 240 micron diameter fiber for three increasing values of the applied axial strain. Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

loads: 76 and 300 g for the 240 and 460 micron diameter fiber, respectively. Under higher strains, the normalized-shift versus strain curves were no longer linear and the measured loads would start decreasing slightly with time, which could be a sign of delamination between the layers and the polymer core, of plastic deformation and/or of relaxation in the layers. The stress–strain curves also exhibit an elastic regime up to approximately 1% strain with a corresponding Young's modulus equal to 2.59 GPa for the 240 μm OD fiber and to 2.39 GPa for the 460 μm OD fiber. These values are slightly lower than the predicted value, possibly due to relaxation effects in the polymer and the glass layers (for As_2Se_3 and PES, $T_g = 175$ and 220 °C, respectively).

Wavelength-Scalable Hollow Optical Fibers with Large Photonic Bandgaps for CO_2 Laser Transmission

Hollow optical transmission fibers offer the potential to circumvent fundamental limitations associated with conventional index guided fibers and thus have been the subject of active research in recent years. Here we report on the materials selection, design, fabrication, and characterization of extended lengths of hollow optical fiber lined with an interior omnidirectional dielectric mirror. These fibers consist of a hollow air core surrounded by multiple alternating submicron-thick layers of a high-refractive-index glass and a low-index polymer, resulting in large infrared photonic bandgaps. These gaps provide strong confinement of optical energy in the hollow fiber core and lead to light guidance in the fundamental and up to fourth-order gaps. We show that the fiber transmission windows can be scaled over a large wavelength range covering at least 0.75 to 10.6 microns. The utility of our approach is further demonstrated by the design and fabrication of tens of meters of hollow photonic bandgap fibers for 10.6 micron radiation transmission. We demonstrate transmission of carbon dioxide (CO_2) laser light with high power-density through more than 4 meters of hollow fiber and measure the losses to be less than 1.0 dB/m at 10.6 microns. This establishes suppression of fiber waveguide losses by orders of magnitude compared to the intrinsic fiber material losses.

Silica optical fibers have been extremely successful in telecommunications applications, and other types of solid-core fibers have been explored at wavelengths where silica is not transparent. However, all fibers that rely on light propagation principally through a solid material have certain fundamental limitations stemming from nonlinear effects, light absorption by electrons or phonons, material dispersion, and Rayleigh scattering that limit maximum optical transmission power and increase attenuation losses. These limitations have, in turn, motivated the study of a fundamentally different light-guiding methodology: the use of hollow waveguides having highly reflecting walls. Light propagation through air in a hollow fiber eliminates or greatly reduces the problems of nonlinearities, thermal lensing, and end-reflections, facilitating high-power laser guidance and other applications which may be impossible using conventional fibers. Hollow metallic or metallo-dielectric waveguides have been studied fairly extensively and found useful practical application, but their performance has been bounded by the notable losses occurring in metallic reflections at visible and infrared (IR) wavelengths, as well as by the limited length and mechanical flexibility of the fabricated waveguides. Hollow all-dielectric fibers, relying on specular or attenuated total reflection, have also been explored, but high transmission losses have prevented their broad application. More recently, all-dielectric fibers,

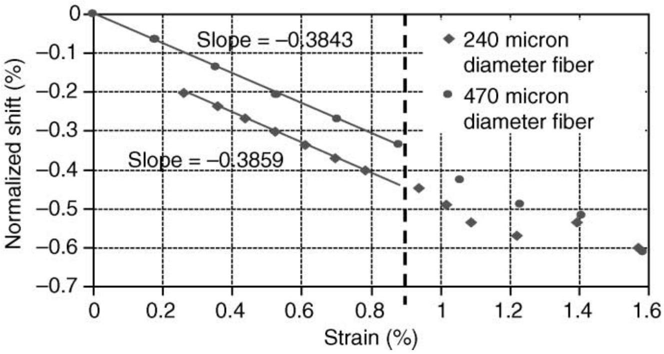

Figure 15 Normalized shift of the Fabry–Perot resonant mode (defined as $\Delta\lambda/\lambda$) versus applied axial strain for the 240 micron (diamonds) and the 460 micron diameter fiber (dots). The curve obtained for the 240 micron diameter fiber is lower because the experimental 0% strain is likely to correspond to a nonzero positive strain, necessary to keep this thinner fiber straight for the measurement (equivalent to a load less than 10 g). Reproduced with permission from Benoit G, *et al.* (2003) Static and dynamic properties of optical cavities in photonic bandgap gains. *Advanced Materials* 15: 2053–2056.

consisting of a periodic array of air holes in silica, have been used to guide light through air using narrow photonic bandgaps. Solid-core, index-guiding versions of these silica photonic crystal fibers have also been explored for interesting and important applications, such as very large core single-mode fibers, nonlinear enhancement and broadband supercontinuum generation, polarization maintenance, and dispersion management. However, the air-guiding capabilities of such waveguides thus far remain inferior to transmission through solid silica, due to various factors such as the difficulties in fabricating long, uniform fibers which must have a high volume fraction of air and many air–hole periods, as well as by the large electromagnetic (EM) penetration depths associated with the small photonic bandgaps achievable in these air–silica structures. In our fiber, the hollow core is surrounded by a solid high-refractive-index-contrast multilayer structure leading to large photonic bandgaps and omnidirectional reflectivity. The pertinent theoretical background and recent analyses indicate that such fibers may be able to achieve ultralow losses and other unique transmission properties. The large photonic bandgaps result in very short EM penetration depths within the layer structure, significantly reducing radiation and absorption losses while increasing robustness. Omnidirectional reflectivity is expected to reduce intermode coupling losses.

To achieve high index contrast in the layered portion of the fiber, we combined a chalcogenide glass with a refractive index of ~2.8 As_2Se_3, and a high-performance polymer with a refractive index of ~1.55 PES. We recently demonstrated that these materials could be thermally co-drawn into precisely layered structures without cracking or delamination, even under large temperature excursions. The same polymer was used as a cladding material, resulting in fibers composed of ~98% polymer by volume (not including the hollow core) and thus combine high optical performance with polymeric processability and mechanical flexibility. We fabricated a variety of fibers by depositing a 5–10 micron thick As_2Se_3 layer through thermal evaporation onto a 25–50 micron thick PES film and the subsequent 'rolling' of that coated film into a hollow multilayer tube called a fiber preform. This hollow macroscopic preform was consolidated by heating under vacuum and cladded with a thick outer layer of PES; the layered preform was then placed in an optical fiber draw tower and drawn down into tens or hundreds of meters of fiber having well-controlled submicron layer thicknesses. The nominal positions of the photonic bandgaps were determined by laser monitoring of the fiber OD during the draw process. Typical standard deviations in the fiber OD were ~1% of the OD. The resulting fibers were designed to have large hollow cores, useful in high-energy transmission.

SEM analysis (Figure 16) reveals that the drawn fibers maintain proportionate layer thickness ratios and that the PES and As_2Se_3 films adhere well during rigorous thermal cycling and elongation. Within the multilayer structure shown in Figure 1, the PES layers (gray) have a thickness of 900 nm, and the As_2Se_3 layers (bright) are 270 nm thick (except for the first and last As_2Se_3 layers, which are 135 nm). Broadband fiber transmission spectra were measured with a Fourier transform infrared (FTIR) spectrometer (Nicolet Magna 860), using a parabolic mirror to couple light into the fiber and an external detector. The results of these measurements are shown in the lower panel of Figure 2 for fibers having two different layer structures. For each spectrum, light is guided at the fundamental and high-order photonic bandgaps. Also shown in the upper panel of Figure 2, is the corresponding photonic band diagram for an infinite periodic multilayer structure calculated using the experimental parameters of our fiber (layer thicknesses and indices). Good agreement is found between the positions of the measured transmission peaks and the calculated bandgaps, corroborated with the SEM-measured layer thicknesses, verifying that transmission is dominated by the photonic bandgap mechanism. In order to demonstrate 'wavelength scalability' (i.e., the control of transmission through the fiber's structural parameters) another fiber was produced, having the same cross-section but with thinner layers. We compared the transmission

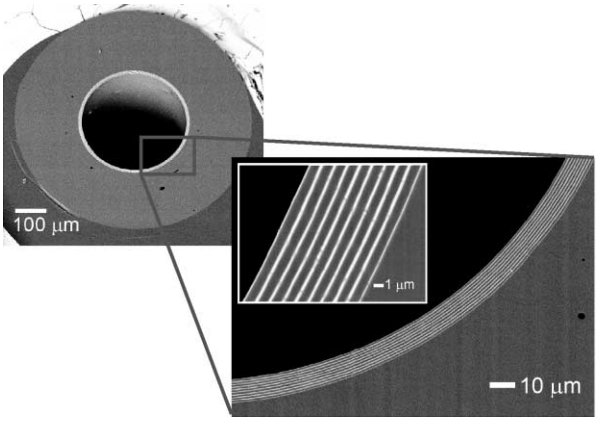

Figure 16 Cross-sectional SEM micrographs at various magnifications of hollow cylindrical multilayer fiber mounted in epoxy. The hollow core appears black, the PES layers and cladding gray, and the As_2Se_3 layers bright white. This fiber has a fundamental photonic bandgap at a wavelength of ~3.55 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

spectra for the original 3.55 micron bandgap fibers to the fiber with the scaled-down layer thicknesses.

Figure 17 shows the shifting of the transmission bands, corresponding to fundamental and high-order photonic bandgaps, from one fiber to the next. The two fibers analyzed in Figure 17 were fabricated from the same fiber preform using different draw-down ratios (fibers with a fundamental bandgap centered near 3.55 microns have an OD of 670 microns; those with a gap at 3.1 microns have an OD of 600 microns). The high-order bandgaps are periodically spaced in frequency, as expected for such a photonic crystal structure.

The wavelength scalability of our fibers was further demonstrated in the fabrication of hollow fibers designed for the transmission of 10.6 micron EM radiation. This not only shows that these structures can be made to guide light at extremely disparate wavelengths, but that specific useful bandgap wavelengths can be accurately targeted during fabrication and fiber drawing. Powerful and efficient CO_2 lasers are available that emit at 10.6 microns and are used in such applications as laser surgery and materials processing, but waveguides operating at this wavelength have remained limited in length or loss levels. Using the fabrication techniques outlined above, we produced fibers having hollow core diameters of 700–750 microns and ODs of 1300–1400 microns with a fundamental photonic bandgap spanning the 10–11 micron wavelength regime, centered near 10.6 microns. Figure 5 depicts a typical FTIR transmission spectrum for these fibers, measured using ~30 cm long straight fibers.

In order to quantify the transmission losses in these 10.6 micron bandgap hollow fibers, fiber cut-back measurements were performed. This involved the comparison of transmitted intensity through ~4 meters of straight fiber with the intensity of transmission through the same section of fiber cut to shorter lengths (Figure 18 inset). This test was performed on multiple sections of fiber, and the results found to be nearly identical for the different sections tested. The measurements were performed using a 25 watt CO_2 laser (GEM-25, Coherent-DEOS) and high power detectors (Newport 818T-10). The fiber was held straight, fixed at both ends as well as at multiple points in the middle to prevent variations in the input coupling and propagation conditions during fiber cutting. The laser beam was sent through focusing lenses as well as 500 micron diameter pinhole apertures and the input end face of the fiber was coated with a metal film to prevent accidental

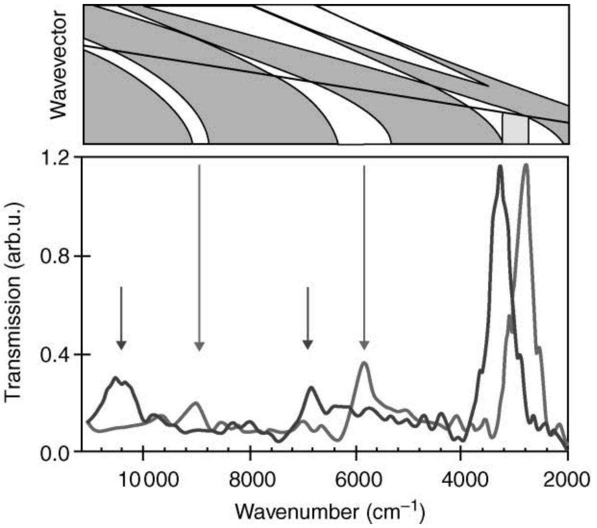

Figure 17 Upper panel: Calculated photonic bandstructure associated with the dielectric mirror lining of the hollow fiber. Modes propagating through air and reflected by the fiber walls lie in the bandgaps (white) and within the light cone defined by the glancing-angle condition (black line). The gray regions represent modes radiating through the mirror. The fundamental bandgap has the widest range-to-mid-range ratio, with a region of omnidirectional reflectivity highlighted in black. Lower panel: Comparison of transmission spectra for two different hollow fibers of ~30 cm length having similar structure but scaled photonic crystal (multilayer) period dimensions. The spectrum in gray is from a fiber with a fundamental photonic bandgap at 3.55 microns; the black spectrum is from a fiber where the corresponding bandgap is at 3.1 microns. High-order bandgaps (indicated by arrows) are periodically spaced in frequency. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

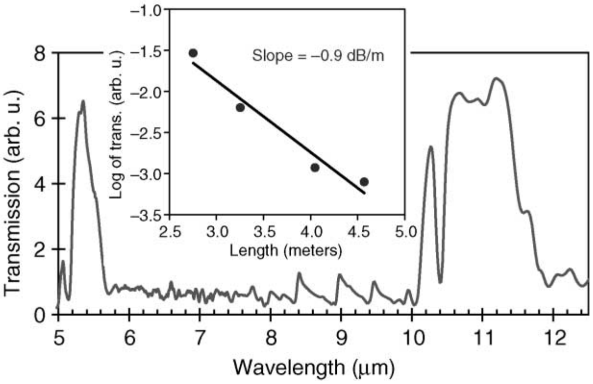

Figure 18 Typical transmission spectrum of hollow fibers designed to transmit CO_2 laser light. The fundamental photonic bandgap is centered near a wavelength of 10.6 microns and the second-order gap is at ~5 microns. Inset: Log of transmitted power (arbitrary units) versus length of fiber (meters). The slope of this graph is the loss in dB/m. The measured fiber has a hollow core diameter of 700 microns. Reproduced with permission from Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

laser damage from misalignment. The transmission losses in the fundamental bandgap at 10.6 microns were measured to be 0.95 dB/m, as shown in the inset of Figure 18, with an estimated measurement uncertainty of 0.15 dB/m. These loss measurements are comparable to some of the best reported loss values for other types of waveguides operating at 10.6 microns. A bending analysis for fibers with a bandgap centered at 10.6 microns revealed bending losses below 1.5 dB for 90 degree bends with bending radii from 4–10 cm. We expect that these loss levels could be lowered even further by increasing the number of layers, through optimization of the layer thickness ratios and by creating a cylindrically symmetric multilayer fiber with no inner seam (present here because of the 'rolling' fabrication method). In addition, using a polymer with lower intrinsic losses should greatly improve the transmission characteristics.

One reasonable figure of merit for optical transmission losses through hollow all-dielectric photonic bandgap fibers is to compare the hollow fiber losses to the intrinsic losses of the materials used to make the fiber. As_2Se_3 has been explored as an IR-transmitting material, yet the losses at 10.6 microns reported in the literature are ~10 dB/m for highly purified material, and more typically are greater than 10 dB/m for commercially available materials such as those used in our fabrication. Based on FTIR transmission and spectroscopic ellipsometer measurements that we have performed on PES, the optical losses associated with propagation through solid PES should be greater than 40 000 dB/m at 10.6 microns. This demonstrates that guiding light through air in our hollow bandgap fibers leads to waveguide losses that are orders of magnitude lower than the intrinsic fiber material losses, which has been one of the primary goals of hollow photonic bandgap fiber research. These comparatively low losses are made possible by the very short penetration depths of EM waves in the high refractive index contrast photonic crystal structure, allowing these materials to be used at wavelengths that may have been thought improbable. Another long-standing motivation of infrared fiber research has been the transmission of high-power laser light. As a qualitative demonstration of the potential of these fibers for such applications, both straight and smoothly bent fibers of lengths varying from 0.3–2.5 meters were used to transmit enough CO_2 laser energy to burn holes through paper. The maximum laser power density coupled into our fibers in these trials was approximately 300 W/cm^2, more than sufficient to burn a homogeneous polymer material, including PES. No damage to the fibers was observed when the laser beam was properly coupled into the hollow fiber core. These results indicate the feasibility of using hollow multilayer photonic bandgap fibers as a low-loss wavelength-scalable transmission medium for high-power laser light.

See also

Diffractive Systems: Design and Fabrication of Diffractive Optical Elements. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber. **Lasers:** Carbon Dioxide Lasers. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abeles F (1950) Investigations on the propagation of sinusoidal electromagnetic waves in stratified media. Application to thin films. *Annales de Physique* 5: 706.

Benoit G and Fink Y, Spectroscopic Ellipsometry Database, http://mit-pbg.mit.edu/Pages/Ellipsometry.html

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press, p. 67.

Engeness TD, Ibanescu M, Johnson SG, *et al.* (2003) Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. *Optics Express* 11: 1175–1198. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-11-10-1175

Fink Y, Winn JN, Fan S, *et al.* (1998) A dielectric omnidirectional reflector. *Science* 282: 1679–1682.

Fink Y, Ripin DJ, Fan S, Chen C, Joannopoulos JD and Thomas EL (1999) Guiding optical light in air using an all-dielectric structure. *Journal of Lightwave Technology* 17: 2039–2041.

Hart SD, Maskaly GR, Temelkuran B, Prideaux PH, Joannopoulos JD and Fink Y (2002) External reflection from omnidirectional dielectric mirror fibers. *Science* 296: 510–513.

Ibanescu M, Johnson SG, Soljacic M, *et al.* (2003) Analysis of mode structure in hollow dielectric waveguide fibers. *Physics Review E* 67: 046608-1–8.

Johnson SG, Ibanescu M, Skorobogatiy M, *et al.* (2001) Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Optics Express* 9: 748–779. http://www.opticsexpress.org/abstract.cfm?URI = OPEX-9-13-748

Kuriki K, Shapira O, Hart SD, *et al.* (2004) Hollow multilayer photonic bandgap fibers for NIR applications. *Optics Express* 12: 8.

Soljacic M, Ibanescu M, Johnson SG, Joannopoulos JD and Fink Y (2003) Optical bistability in axially modulated OmniGuide fibers. *Optics Letters* 28: 516–518.

Temelkuran B, Hart SD, Benoit G, Joannopoulos JD and Fink Y (2002) Wavelength-scalable hollow optical fibres with large photonic bandgaps for CO_2 laser transmission. *Nature* 420: 650–653.

Yeh P, Yariv A and Hong C-H (1977) Electromagnetic propagation in periodic stratified media. 1. General theory. *Journal of the Optical Society of America* 67(4): 423.

Yeh P, Yariv A and Marom E (1978) Theory of Bragg fiber. *Journal of the Optical Society* 68: 1196–1201.

Wave Optical Modeling and Design

F Wyrowski, Friedrich-Schiller University, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Why do ray optics dominate the literature on optical engineering? The major practical reason may be that historically optical engineering has dealt with systems for image forming. These systems transform divergent spherical waves from points in an object, into spherical waves converging to focal points in the image. Thus, the basic operation of an imaging system performed by a lens is changing the radius of curvature of spherical fields. The propagation of spherical waves is modeled well by geometrical optics, so long as truncation by apertures and other high frequency aberrations are not significant. Moreover, the basic merit function in imaging evaluates the appearance of the image point at its correct location by so-called dot diagrams. Evaluation of this merit function requires only a ray bundle representation of the output field. The combination of a ray representation with geometrical optics propagation is called raytracing, which explains the dominance of raytracing in conventional optical engineering. Because the design of illumination systems is also often based on principles of imaging, raytracing remains a suitable technique in that case too. However, systems that perform general light transformations must apply wave-optical engineering techniques, which constitutes a generalization of conventional optical engineering. In this sense it includes the modeling of image-forming systems as a special case.

Modeling Principles

Modeling in wave-optical engineering addresses the propagation of electromagnetic fields through systems as illustrated in Figure 1. The introduction of operators and operands as mathematical tools simplifies a systematic discussion of the modeling concepts. Formally, an electromagnetic field may be expressed by the operand $\mathbf{f} \in \mathsf{F}$ with the set of fields F. The actual form of $\mathbf{f} \in \mathsf{F}$ depends on the physical approach chosen to model electromagnetic fields. For instance, $\mathbf{f} \in \mathsf{F}$ may stand for a ray bundle, a complex amplitude, or a coherence function. The photon source generates a field which acts as the input field $\mathbf{f}$ into the optical system. The quality of the output field $\mathbf{f}_{\text{out}}$ is evaluated with the help of a merit function $\boldsymbol{\Omega}$. This function may consist of M components which yields:

$$\boldsymbol{\Omega}(\mathbf{f}_{\text{out}}) = \{\Omega_1(\mathbf{f}_{\text{out}}), \ldots, \Omega_\mu(\mathbf{f}_{\text{out}}), \ldots, \Omega_M(\mathbf{f}_{\text{out}})\} \quad [1]$$

with $\Omega_\mu : \mathsf{F} \to \mathsf{R}^+$ and $\mu = 1, \ldots, M$. Each component Ω_μ defines a mathematical way to obtain one positive valued quantity for a given $\mathbf{f}_{\text{out}}$. Examples are the rms-value to evaluate the quality of a wavefront, M^2-value to judge the profile of a laser beam, center of gravity of a light distribution, light transformation efficiency, and the value of an overlap integral to determine the coupling efficiency into a fiber. Any property of the output field can be evaluated by an appropriate definition of a function Ω_μ, as long as the demanded field parameters are accessible for the field model in use. However, it should be mentioned, that it is not clear *a priori*, that all quality criteria Ω_μ can be satisfied simultaneously. How the different Ω_μ are related is dependent on the application.

The propagation of the input field through the system is mathematically summarized by the operator equation:

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} \quad [2]$$

with the operator $S : \mathsf{F} \to \mathsf{F}$. p_α stands for all parameters p_1, p_α, p_A which specify the system, for instance position of elements, refractive index, surface profiles, and layer thickness. Obviously a full electromagnetic model of S does not exist for arbitrary systems. The propagation problem eqn [2] is much too complex for a wave-optical solution. However, for a sequential approach the situation is different. If the propagation through a system is performed sequentially through the elements and the homogenous regions, at least approximate wave-optical methods are available to model the propagation S. In a sequential approach, either the transmitted or reflected field is considered per region

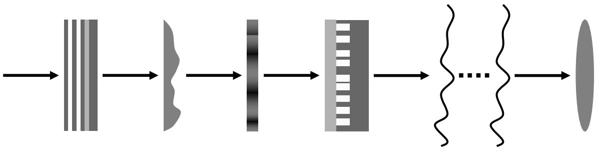

Figure 1 An optical system may include various types of elements to perform the desired optical function. Layered media, free-form surfaces, index-modulated regions, microstructured interfaces, lenses, and other structured media offer the flexibility needed to realize general light transformations. Typically structured media and homogenous regions follow each other.

of the system. Although this is an approximation, it turns out to be very useful in wave-optical engineering. With regard to Figure 2 the sequential approach is mathematically expressed by

$$\mathbf{f}_{\text{out}} = S(p_\alpha)\mathbf{f}_{\text{in}} = \mathcal{P}_J(n_J,\Delta z_J)S_{J-1}(p_{\alpha_{J-1}})\cdots S_2(p_{\alpha_2}) \times \mathcal{P}_1(n_1,\Delta z_1)\mathbf{f}_{\text{in}} \qquad [3]$$

$\mathcal{P}_j(n_j,\Delta z_j)$ expresses the propagation through a homogenous dielectric with refractive index n_j and extent Δz_j. This type of propagation is also called free propagation. Sometimes the term free propagation is restricted to propagation in vacuum; Here it is used more generally. $S_j(p_{\alpha_j})$ models the propagation through an optical component which is specified by the parameters p_{α_j}. Optical engineering needs physical models for the field operand $\mathbf{f}$, the free propagation $\mathcal{P}$, and the propagation S_j through elements. The most fundamental demand in wave-optical engineering is a wave-optical model to express the field $\mathbf{f}$ in homogenous regions. It is common to restrict to harmonic fields. The generalization to more general radiation is often possible by suitable superposition of harmonic fields. This approach assumes linear optical systems as typical in optical engineering. From Maxwell's equations it follows that harmonic fields in homogenous regions are completely specified if two independent field components are determined in one plane z_0 within the region. Therefore, the electromagnetic field operand takes the form:

$$\mathbf{f}(x,y,z_0) = [E_x(x,y,z_0), E_y(x,y,z_0)] \qquad [4]$$

with the complex amplitudes E_x and E_y of the x and y components of the electric field vector. Because the knowledge of $\mathbf{f}(x,y,z_0)$ allows the direct calculation of the missing complex amplitudes E_z, H_x, H_y, H_z via Maxwell's equations, the entire electromagnetic harmonic field in a homogenous region is specified by eqn [4].

With respect to the two components of the field operand $\mathbf{f}$ propagation operators S are 2×2-matrices of the form:

$$S = \begin{Bmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{Bmatrix} \qquad [5]$$

In the case of a diagonal matrix, that is $S_{xy} = S_{yx} = 0$, E_x and E_y propagate independently and are considered decoupled channels of the system. If in addition $S_{xx} = S_{yy}$, both channels behave identically, and it is sufficient to model only one channel. This is often referred to as the scalar model. Of course, also in the scalar model the electromagnetic field still consists of two independent components and maintains its vectorial nature.

Numerical implementation of eqn [3] for the field operand of eqn [4] constitutes the backbone of modeling in wave-optical engineering.

Free Propagation

The propagation $\mathcal{P}$ of harmonic fields $\mathbf{f}$ through homogenous regions of a system has a rigorous electromagnetic solution in form of the spectrum of plane waves integral which yields:

$$\begin{aligned} \mathbf{f}(x,y,z \geq z_0) &= \mathcal{P}_{\text{SPW}}\mathbf{f}(x,y,z_0) \\ &= \begin{pmatrix} \mathcal{P}_{\text{SPW}} & 0 \\ 0 & \mathcal{P}_{\text{SPW}} \end{pmatrix} \begin{pmatrix} E_x(x,y,z_0) \\ E_y(x,y,z_0) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{F}^{-1}\left([\mathcal{F}E_x(x,y,z_0)]\mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z \Delta z)}\right) \\ \mathcal{F}^{-1}\left([\mathcal{F}E_y(x,y,z_0)]\mathrm{e}^{\mathrm{i}(k_x x + k_y y + k_z \Delta z)}\right) \end{pmatrix} \end{aligned} \qquad [6]$$

with the Fourier transform $\mathcal{F}$ and $\Delta z = z - z_0$. The term $\mathcal{F}E(x,y,z_0)$ is called spectrum of planewaves (SPW). The operator matrix is diagonal and $\mathcal{P}_{\text{SPW}} = \mathcal{P}_{xx} = \mathcal{P}_{yy}$, that means the x- and y-components propagate independent of each other according to the same formula in homogenous dielectrics.

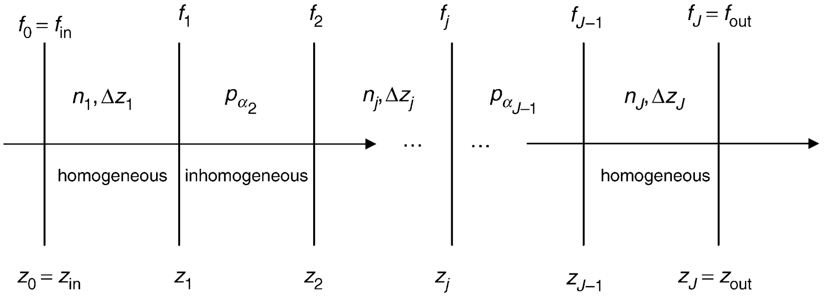

Figure 2 Formally a system may be subdivided into a sequence of homogenous dielectrics like air and inhomogeneous regions which include the elements. We indicate a region by the letter *j*.

In this sense, free propagation is a scalar problem which allows the convenient use of $U(x,y,z)$ instead of two field components for the sake of simpler mathematical expressions.

Equation [6] allows the derivation of other rigorous versions of the free propagation operator $\mathcal{P}_{\text{SPW}}$, for instance Rayleigh's integral formula, which is well-known from diffraction theory. Moreover, eqn [6] leads to important approximate operators for far-field propagation and the propagation of paraxial fields. Propagation of a field into its far field is given by the operator:

$$\begin{aligned} U(r\mathbf{s}) &= \mathcal{P}_{\text{far}} U(x,y,z_0) \\ &= -\mathrm{i}2\pi k s_z [\mathcal{F}U(x,y,z_0)]|_{(ks_x,ks_y)} \frac{\exp(\mathrm{i}kr)}{r} \end{aligned} \qquad [7]$$

with the wave number $k = 2\pi n/\lambda$, $r = |\mathbf{r}| = (x^2 + y^2 + (\Delta z)^2)^{1/2}$, and the direction vector $\mathbf{s} = (s_x, s_y, s_z) = \mathbf{r}/r$. Here, and in what follows, λ denotes the vacuum wavelength. $|_{(ks_x,ks_y)}$ indicates the variables to be inserted into the Fourier transformed field. Though the far-field formula is derived for $r \rightarrow \infty$, it is also of great practical value for the propagation of fields generated by nonparaxially emitting sources like laser diodes. Roughly speaking eqn [7] predicts a spherical wave in the far field which is modulated by the spectrum of planewaves of the field in z_0. $\mathcal{P}_{\text{far}}$ propagates from the plane z_0 to the spherical surface with radius r and not to a plane z.

The paraxial approximation of eqn [6] follows if the spectrum of planewaves $\mathcal{F}U(x,y,z_0)$ possesses significant values only for small k_x and k_y, that is if both electric field components propagate approximately along the z-axis. The resulting propagation operator is known as Fresnel's integral formula. In practice, its formulation with the Fourier transform $\mathcal{F}$ is typically more suitable:

$$\begin{aligned} U(x,y,z \geq z_0) &= \mathcal{P}_{\text{paraxial}} U(x,y,z_0) \\ &= \frac{k \exp[\mathrm{i}k\Delta z]}{\mathrm{i}2\pi\Delta z} \exp\left[\frac{\mathrm{i}k(x^2+y^2)}{2\Delta z}\right] \\ &\quad \times \left| \left\{ \mathcal{F}\left(\exp\left[\frac{\mathrm{i}k(x^2+y^2)}{\Delta z}\right] U(x,y,z_0) \right) \right\} \right|_{(x,y)n/\lambda\Delta z} \end{aligned} \qquad [8]$$

$|_{(x,y)n/\lambda\Delta z}$ indicates the variables to be inserted after Fourier transform. The far-field approximation of eqn [8] leads to the well-known Fraunhofer diffraction formula.

Equations [6]–[8] provide fundamental operators for the propagation of fields in wave-optical engineering. All integrals are based on Fourier transforms which allows the application of the fast Fourier transform (FFT) in practice. Dependent on the propagation technique, one or two FFTs must be performed. In practice, the sampling of fields is of great concern. Naturally the amount of data should be as small as possible. According to the rigorous propagation operator eqn [6] the (x,y)-coordinate system and therefore the sampling period in z_0 and z are identical, that is, scaling does not occur. As illustrated in Figure 3, that can lead to an enormous amount of data for large propagation distances. For demonstration purpose the field amplitude depicted at the top is chosen. The field window has the size 1 mm × 4 mm. The wavelength λ is 632.8 nm, $n = 1$, and the sampling distance 10 μm. That results in 100 × 400 sampling points. Propagation of this field with the spectrum of planewave operator maintains the size of the field window. The resulting field for a propagation distance $\Delta z = 10$ mm is shown in the middle. Because of diffraction, the field widens. As long as the initial window is large enough to encompass the propagated field, the propagation method works well. At the bottom, the resulting field for $\Delta z = 200$ mm is shown. The field exceeds the size of the window and that leads to numerical errors (aliasing). This can be avoided by embedding the initial field into a window which is large enough with respect to the expected size of the propagated field. As a result, one obtains the

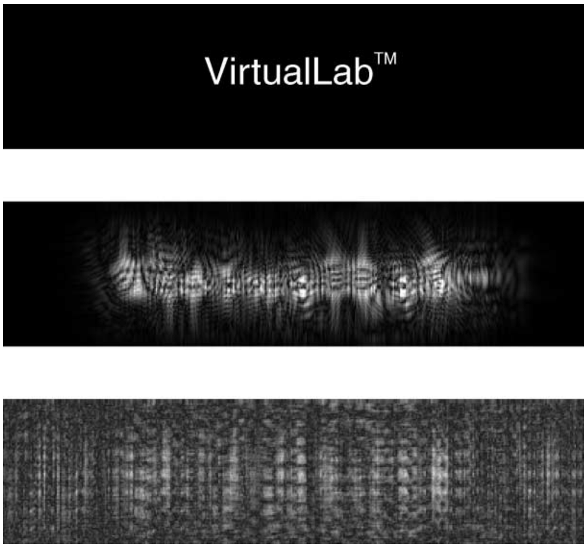

Figure 3 Illustration of numerical properties of the spectrum of plane wave propagation technique. The field amplitude depicted at the top is numerically propagated in free space and the distributions in the middle and the bottom are obtained. The simulations in this and all following figures were performed. The simulations in this and all following figures were performed with the wave-optical engineering software VirtualLab™ from LightTrans GmbH.

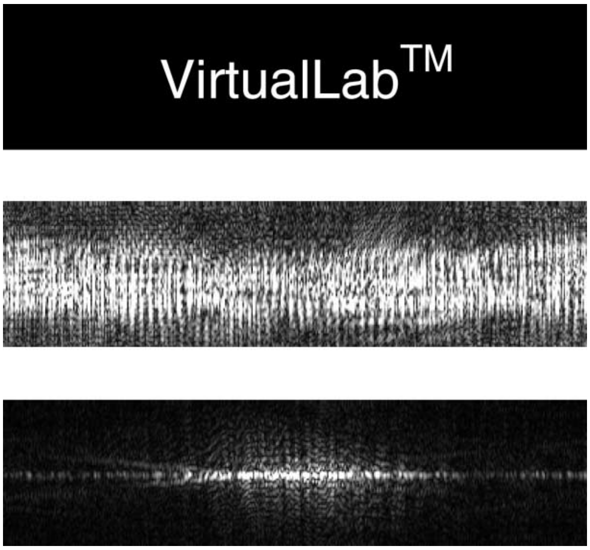

Figure 4 Starting from the same input field and the same distances as in Figure 3 the propagated fields depicted in the middle and the bottom are obtained using the paraxial propagation operator. Because of the inherent scaling the field window sizes are 632.8 μm × 632.8 μm and 12.656 mm × 12.656 mm, respectively. The number of sampling points remains unchanged. Obviously, the field in the middle is too small to encompass the propagated field and numerical errors occur. For the long distance the field size is perfectly matched to the field size obtained by propagation.

demanded accuracy at cost of increasing data memory requirements and computation time. Thus in practice the SPW propagation operator is particularly suitable for short distances. Fortunately the paraxial propagation operator eqn [6] possesses a complementary behavior. Because of $(x, y)n/\lambda\Delta z$, the (x, y)-coordinate system changes its scaling with the distance. As illustrated in Figure 4, that makes it well suited for large propagation distances but problems occur for small ones. Therefore, all paraxial propagation problems have an efficient numerical solution by a suitable combination of $\mathcal{P}_{\text{paraxial}}$ and $\mathcal{P}_{\text{SPW}}$. In the nonparaxial case the far-field operator eqn [7] often helps to realize a numerically efficient propagation for large distances. However, there are sometimes situations, in which free propagation of nonparaxial fields causes significant numerical problems.

Propagation Through Elements

The propagation of harmonic fields $\mathbf{f}$ through elements is a challenging task. For example, Figure 5 illustrates propagation through smooth or microstructured interfaces between two homogenous media. The field $\mathbf{f}_1(x, y, z_1)$ should be determined if $\mathbf{f}_0(x, y, z_0)$, the shape of the interface, and the refractive indices n_0 and n_1 are known. The reflection mode of operation can be formulated analogously. It is remarkable that this basic problem possesses a rigorous solution only for a few special interface geometries. Of particular importance are the rigorous solutions for plane and periodic interfaces. In the first case, the solution comprizes Snell's law of refraction, the law of reflection, and Fresnel's formulae. Grating theory provides the solution for periodic interfaces. It is possible to use these methods also for nonperiodic interfaces but at the cost of immense computational effort. Thus, they are only practicable for structures with detail no larger than a few wavelengths. For general interfaces, accurate approximate methods are required. One approach is to decompose a general propagation problem into a set of simpler problems that can be solved rigorously. The results are then combined to obtain an approximate solution of the original problem. Two basic concepts for subdividing the general problem are: (i) Decomposition of the incident field into laterally truncated fragments which illuminate elementary fractions of the interface. (ii) Decomposition of the response of the interaction of the incident field with the interface into approximately independent ones related to local interface features. The first technique is referred to as local elementary interface approximation (LEIA) and the second one as local independent response approximation (LIRA). Elementary interfaces of particular importance are plane, periodic, and spherical ones.

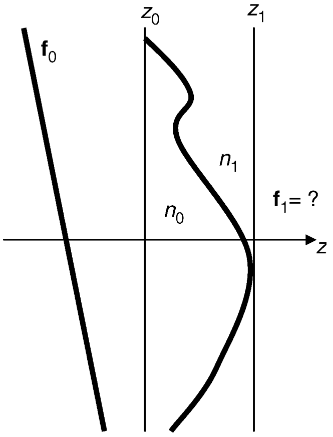

Figure 5 A planewave $\mathbf{f}_0$ propagates through an interface between two homogeneous dielectrics with refractive index n_0 and n_1. The interface is of general shape and therefore a rigorous solution of the propagation problem does not exist. The virtual planes in z_0 and z_1 encompass the interface in a way which minimizes the distance $z_1 - z_0$.

The local plane interface approximation (LPIA) is of particular importance in wave-optical engineering. Figure 6 illustrates the basic concept for the example

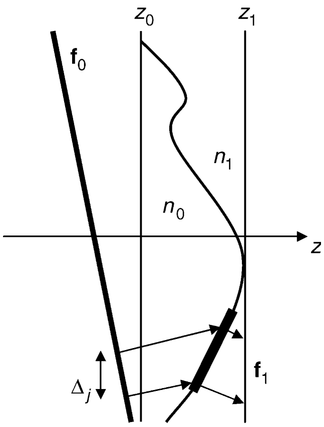

Figure 6 If a lateral fraction of the input field is propagated towards the interface, it is often possible to laterally restrict the major part of the propagated field to a local plane interface. Because of diffraction it is never possible to do that rigorously. Neglecting diffraction in the region between z_0 and z_1 provides a simple model to propagate a section of the incident plane field through the interface. Fresnel's formulae and Snell's law of refraction can then be applied directly to obtain the transmitted plane field, which is propagated to the plane z_1.

of a plane incident field. It has been proven that the approximation of geometrical optics propagation within the regions of the component, here that means between z_0 and z_1, provides results which are surprisingly accurate in numerous situations of practical interest. A full wave-optical version of LPIA is currently under research and development. In general, however, LPIA is applicable if the characteristic lateral extent of the local modulation of the interface is not close to the wavelength. It should be emphasized that LPIA is, as most other approximate propagation techniques, a 2 × 2 matrix, like eqn [5], with nonvanishing components. That means, for instance, that in general, an E_y-component occurs in z_1 even if $E_y = 0$ in z_0. That is, the two field channels experience crosstalk. A further approximation in geometrical optics LPIA, popular in Fourier optics holography, and diffractive optics is the thin element approximation (TEA). Geometrical optics LPIA leads to TEA if refraction and the local dependence of Fresnel losses at the interface are neglected. Obviously, that is allowed only if all angles involved, including the slope of the interface and the direction of the input field, are small. Therefore, TEA is a paraxial and geometrical optics approximation. Moreover, small local slopes of the interface lead to a restriction to thin elements, which explains the name TEA. It is also referred to as scalar approximation, because the operator matrix becomes diagonal with $S_{\mathrm{TEA}} = S_{xx} = S_{yy}$, that means the two field channels propagate independently according the same formula through the interface. In the thin element only approximation the dependence of the optical path through the element on position (x, y) is considered and mathematically takes the form:

$$\begin{aligned} U(x,y,z_1) &= S_{\mathrm{TEA}}\,U(x,y,z_0) \\ &= c_{\mathrm{Fresnel}}\,U(x,y,z_0)\exp\left[\frac{\mathrm{i}2\pi n_1(z_1-z_0)}{\lambda}\right] \\ &\quad \times \exp\left[\frac{\mathrm{i}2\pi h(x,y)(n_0-n_1)}{\lambda}\right] \end{aligned} \qquad [9]$$

with the vacuum wavelength λ. The profile height $h(x, y)$ is expressed as the distance from z_0 to the interface. Propagation through an interface between dielectrics modeled with TEA affects the phase of the incident field proportional to the profile height. The magnitude remains unchanged besides a constant factor c_{Fresnel}, due to Fresnel's equations. If complex valued refractive indices are taken into consideration, absorption results in an amplitude modulation completely analogous to the phase modulation described in eqn [9]. This way it is also possible to address the propagation through apertures and the combination of S_{TEA} with $\mathcal{P}$ leads to the well-known diffraction integrals. In fact, Kirchhoff's boundary conditions of field propagation through a hole in a completely opaque screen can be interpreted as TEA for complex valued refractive indices.

The simple form of the TEA operator allows a straightforward numerical implementation. The field operand and the profile are equidistantly sampled and the resulting field is sampled analogously. That makes the combination of S_{TEA} with numerical implementations of the free propagation operators $\mathcal{P}$ easy. As soon as refraction is considered in geometrical optics LPIA, the equidistant sampling of the output field is destroyed which enforces additional interpolation effort before a free propagation by fast Fourier transforms can be implemented, because fast numerical Fourier transform algorithms require an equidistant sampling.

In wave-optical engineering the combination of geometrical optics LPIA with a wave-optical field model is mandatory. In this context it is instructive to mention, that its combination with a ray-bundle representation of the incident electromagnetic field leads to raytracing through the interface, that is it reduces to the basic technique in conventional optical engineering.

Similar to LPIA it is possible to consider linear gratings as elementary local interface and the local linear grating approximation (LLGA) results. It is of special concern for all types of diffractive elements, for which the interface can locally considered periodically. An important example is given for diffractive lenses in the outer zones of high numerical aperture lenses. In the center and for low numerical aperture TEA is appropriate for propagating through diffractive lenses.

Design Principles

The design problem in wave-optical engineering can be stated as follows: specify system parameters p_α in a way that ensures for a given input field $\mathbf{f}_{\text{in}}$ an output field $\mathbf{f}_{\text{out}}$ which satisfies the quality criteria defined by the merit functions Ω_μ. In addition, one may demand a system which is most simple, particularly cheap, or various other system properties of concern in practice.

Design constitutes an optimization problem. Restrictions to the system parameters and the merit function sets design constraints as well as design freedoms.

System parameters can be restricted, for instance, by the materials available, the maximum size of a system, and types of element profiles. Within these limits the system parameters are chosen freely during optimization to achieve the optical function.

The merit functions determines which parameters of the output field are important and should be emphasized, for instance the intensity distribution. All output fields, whose parameters satisfy the constraints satisfy the design problem. In most applications, not all field parameters are constrained. For instance, in laser materials processing or illumination applications, the merit function is related to energy quantities and then the phase of the output field can be used as a free parameter distribution. Moreover, the output field is typically not constrained in the entire output plane, but only in so-called signal regions, which introduces freedom of amplitude outside the signal regions. Hence, dependent on the application, there exist field parameters that can be chosen partly or completely free during design.

Successful design strategies use the freedoms to improve the matrix function. It is often not clear a priori, if a solution does exist. Then the field constraints are satisfied as good as possible in terms of the merit function.

In imaging systems, typical system parameters include, for instance, distances, refractive indices, curvature of lenses, and parameters to specify aspherical surfaces. For high-end lens systems, the resulting number of parameters can be large, but they are still small enough to utilize parameter optimization algorithms if the designer starts with a reasonable initial guess. In wave-optical engineering, the number of system parameters can become huge due to the generalized nature of the structures, which typically prevents the use of system parameter optimization techniques. For example, the profile of diffractive elements is typically specified by 10^4–10^6 parameters. Thus a direct optimization of the system parameters, which requires the evaluation of $S(p_\alpha)$ (see eqn [3]) after any change of system parameters p_α, is not realistic.

Instead of optimizing the system parameters directly, which is referred to as design in the structural embodiment of the system, a technique wellknown from Fourier optics and diffractive optics can be generalized for wave-optical engineering. It does not start with structure related element parameters but replaces all or at least some of the operators $S_j(p_{\alpha_j})$ of eqn [3] by mathematical operators $\mathcal{T}_j(t_{\tau_j})$. t_τ denotes the parameters of $\mathcal{T}$. The propagation operator through a system expressed in this functional embodiment has the form:

$$\begin{aligned}\mathbf{f}_{\text{out}} &= S_{\text{functional}}\mathbf{f}_{\text{in}} \\ &= \mathcal{P}_J(n_J, \Delta z_J)\mathcal{T}_{J-1}(t_{\tau_{J-1}})\cdots\mathcal{T}_2(t_{\tau_2}) \\ &\quad\times \mathcal{P}_1(n_1, \Delta z_1)\mathbf{f}_{\text{in}}\end{aligned} \qquad [10]$$

if all propagations through elements are expressed in functional form. A very important example of a functional operator $\mathcal{T}$ is given by a simple multiplication of both field components with identical or different complex functions $t(x, y)$, that is:

$$\mathcal{T}_t(t_\tau)\mathbf{f}_{\text{in}} = \begin{pmatrix} t_{xx}(x, y) & 0 \\ 0 & t_{yy}(x, y) \end{pmatrix}\mathbf{f}_{\text{in}}(x, y) \qquad [11]$$

The values of both $t(x, y)$ form the parameters t_τ of $\mathcal{T}$. The functions $t(x, y)$ are also called ideal transmission functions of elements. Of special interest are phase-only transmission functions $t(x, y) = \exp[i\tau(x, y)]$ (see discussion below). If $t(x, y) = t_{xx}(x, y) = t_{yy}(x, y)$, the design is simplified. Fortunately, in most actual applications this assumption is valid.

Designing in the functional embodiment requires the evaluation of $S_{\text{functional}}$ instead of S of eqn [3], which reduces the numerical complexity enormously. That improves the likelihood for developing efficient design algorithms to optimize the distances Δz_j, the refractive indices n_j, and parameters t_{τ_j}.

The design in the functional embodiment does not deliver the structure parameters of elements but only parameters t_τ of the operator $\mathcal{T}$ which gives information about how a field should be changed at the position z where $\mathcal{T}$ takes effect. Thus, in a next design step, elements must be obtained which cause the required effect. This so-called structure design step is also generally challenging because it requires the solution of an inverse propagation problem through interfaces. For a further discussion the most important $\mathcal{T}$, that is the transmission type operator of eqn [11], is considered. For the sake of simplicity $t(x, y) = t_{xx}(x, y) = t_{yy}(x, y)$ is assumed. Then, a solution which uses geometrical optics LPIA in combination with phase-only transmissions is known in the form of a recursive structure design strategy. In its paraxial approximation, that is TEA, the solution becomes very simple. For a transmission $t(x, y) = \exp[i\tau(x, y)]$, obtained from a functional design, an element with a surface profile:

$$h(x, y) = \frac{\lambda\tau(x, y)}{2\pi(n_{\text{element}} - n_{\text{free}})} \qquad [12]$$

results from eqn [9]. The refractive index of the element is denoted by n_{element} and the homogenous region behind the element has the index n_{free}, eqn [12] constitutes a direct relationship between the phase values of the transmission function and the profile $h(x, y)$. This structure design technique is accurate for all paraxial situations. However, it is also a reasonable method for nonparaxial structure design since it provides an initial profile for further optimization. Obviously it is applicable for phase-only transmission functions only. That underlines the importance of this type of functional operator for the design in the functional embodiment. Often the fabrication of profiles is done with lithographic techniques which result in surface reliefs with a few discrete height levels, that means the profile $h(x, y)$ is quantized. According to eqn [12] a quantization of $h(x, y)$ directly leads to quantized phase values $\tau(x, y)$. This quantization must be taken into account when designing $\tau(x, y)$. An example should serve to discuss further aspects of structure design according to eqn [9]. A fundamental transmission in various applications is the phase of a spherical wave and its paraxial version, the quadratic phase. An example of a spherical phase $\tau(x, y)$ is shown in Figure 7 where the 2π-modularity of the exponential function is taken into account – as normal for phase data which has been obtained by numerical calculations. Inserting this phase into eqn [9] leads to the profile of a diffractive lens. Because of the use of inverse TEA for structure design it works as demanded for moderate numerical apertures and starts to suffer from aberrations for large numerical apertures because of electromagnetic effects in the very outer zones. It is also possible to unwrap the phase before inserting it into eqn [9]. Then a smooth surface, that is a refractive spherical lens, results. It is well-known that a spherical lens works accurately in the paraxial domain and suffers from aberrations for moderate and large numerical apertures, which are more severe than for the diffractive lens. The use of inverse geometrical optics LPIA as a structure design technique is not restricted to the paraxial domain and therefore it leads to a suitable aspherical surface in order to ensure a spherical phase transmission for larger numerical apertures too. This is illustrated in Figure 8. The lens example shows another advantage of the two-step design method, which consists of the design in the functional embodiment prior to the one in the structural form of the element. In the functional

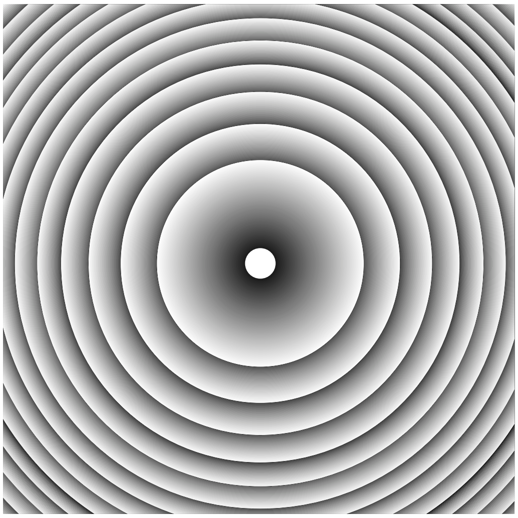

Figure 7 The phase of a spherical phase in the 2π-modulo representation. This phase can be unwrapped which leads to a smooth spherical phase which varies from 0 to various 2π. The noisy edges are due to the use of a cartesian coordinate system.

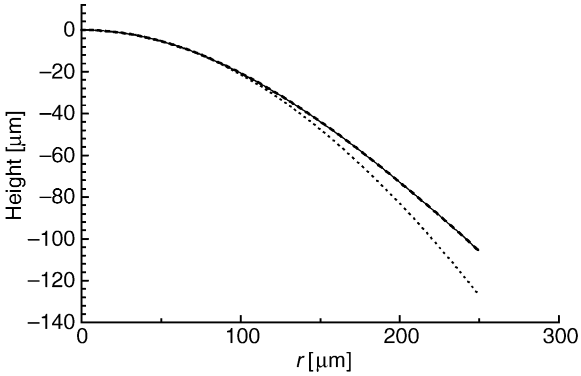

Figure 8 Application of the thin lens approximation for structure design leads to a spherical surface for a spherical phase-only transmission. The corresponding surface is depicted by the dotted line. If geometrical optics LPIA is used for structure design, an aspherical surface is obtained as required. The aspherical surface is depicted by the solid line. The height of the surfaces are shown as function of the radius *r*. Naturally the difference between TEA and geometrical optics LPIA occurs for larger *r* only.

form the transmission is obtained. It is not necessary to decide about the use of a diffractive or a refractive element in an early stage of the design. Moreover, it provides the designer with insight into selecting optical elements to achieve the function. If the phase-only transmission can be unwrapped, a refractive as well as a diffractive element can be used. An optical engineer has the freedom to decide about the suitable choice taking fabrication, size, costs, and other additional constraints into account. However, as soon as the phase-only transmission includes so-called vortices, unwrapping is not possible and diffractive optics is mandatory. An example of a vortex is shown in Figure 9. It turns out that vortices are more likely to appear in wave-optical engineering the less image-like the optical function is (see also last section). It is important to note, that functional design assumes a diagonal operator matrix, that is, it excludes crosstalk of the two field channels. Transferring the system into a structural form may violate this assumption because of vectorial effects in the nonparaxial domain. For instance, geometrical optics LPIA in general does not correspond to a diagonal operator. Therefore, it is necessary to analyze the final system for $\mathbf{f}_{\text{in}}$ – even in the case of linear polarized light – using more accurate propagation techniques than S_{TEA} if it is not ensured that the paraxial approximation is accurate enough.

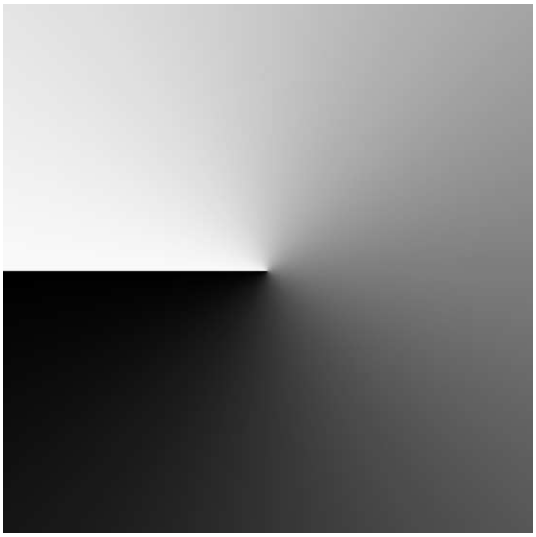

Figure 9 Illustration of a single vortex of charge one. The phase varies from 0 to 2π. Because of the vortex, this phase is not unwrappable, that means a smooth phase function cannot be obtained. The transition in the phase function requires a microstructure in the profile of the element, which generates this phase-only transmission.

The so-called thin lens, or ideal lens, in conventional optical engineering constitutes a special case of the two-step design strategy. In the first design step, distances and lens focal lengths are determined, without taking into account the actual shape of a lens. In the second step lenses, or even lens systems, are inserted according to the position and the focal length specified in the first step. Thus, wave-optical engineering extends our understanding by broadening established methods in optical engineering.

Fundamental Systems

A smart approach to design minimizes the number of elements required to achieve a desired light transformation. Thus, it is logical to try a functional design with a system of the form:

$$\begin{aligned} U_{\text{out}}(x, y) &= S_{\text{functional}} U_{\text{in}}(x, y) \\ &= \mathcal{P}(n_{\text{free}}, \Delta z_{\text{out}}) \exp[\mathrm{i}\tau(x, y)] \\ &\quad \times \mathcal{P}(n_{\text{free}}, \Delta z_{\text{in}}) U_{\text{in}}(x, y) \end{aligned} \quad [13]$$

which is illustrated in Figure 10. One phase-only transmission, properly placed between the input and the output plane, should realize the desired transformation of the function. Scalar terminology is used, because both components are decoupled in the functional embodiment.

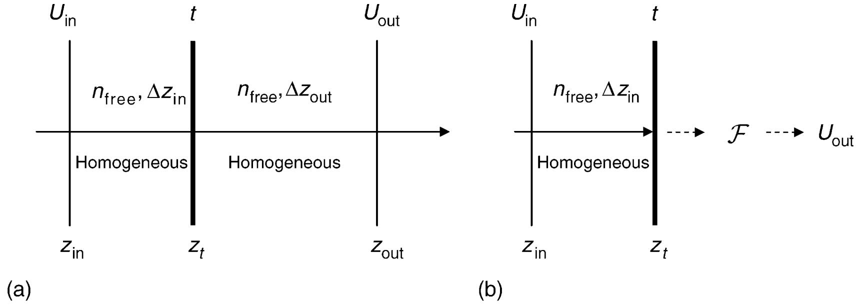

Figure 10 A basic system for light transformation can be described functionally by two propagations through homogeneous dielectrics with refractive index n_{free} and one phase-only transmission function. (b) If the second propagation is replaced by an operation in functional form, here the Fourier transform, the setup shown on the right side results.

If the output field is completely specified by the merit function, for instance: if the complex amplitude of a laser beam should be transformed into another one, the so-called amplitude matching strategy is a very powerful technique to solve the design problem stated by eqn [13]. In amplitude matching one determines the Distances Δz_{in} and Δz_{out} such that the inverse propagated U_{out} and the forward propagated U_{in} possess identical magnitudes. If such distances exist, the transformation can be realized using a phase-only transmission in the plane z_t. The phase simply follows from the ratio of the propagated fields. If no distances exist where amplitude matching occurs, the design problem has no solution and at least one more transmission function must be introduced.

As mentioned above, in numerous applications the output field is not specified completely. In particular, phase is often a free parameter. Then, an even simpler system approach than the one suggested by eqn [13], often leads to reasonable solution. Instead of $\mathcal{P}(n_{\text{free}}, \Delta z_{\text{out}})$, a Fourier transform is assumed, then if,

$$\begin{aligned} U_{\text{out}}(k_x, k_y) &= S_{\text{functional}} U_{\text{in}}(x, y) \\ &= \mathcal{F}\{\exp[\mathrm{i}\tau(x, y)]\mathcal{P}(n_{\text{free}}, \Delta z_{\text{in}})U_{\text{in}}(x, y)\} \end{aligned} \quad [14]$$

The output field is expressed in k-coordinates as long as the system is discussed in functional form. For a functional design the distance Δz_{in} is of no concern, however it must be chosen together with the structure design as discussed below. Thus, in functional design, one concentrates only on the design of the transmission function. The iterative Fourier transform algorithm (IFTA), an extremely powerful algorithm, has been developed to accomplish this design. Its principle is illustrated in Figure 11. The algorithm sequentially applies constraints on the transmission and in the Fourier plane. Properly controlled, the algorithm converges to a Fourier pair for which the transmission plane constraint is satisfied completely and the Fourier domain constraint is satisfied as accurately as required by the merit function. The initial distribution is of enormous importance for a successful transmission design. Typically the initial distribution is defined in the Fourier plane as shown in Figure 11. A field $U_{\text{sig}}(k_x, k_y)$ is chosen, which satisfies the merit function. It is called the signal field. Because of design freedoms such as phase, typically there exists a set of signal fields F_{sig}. If phase freedom is available, any field with the desired amplitude and an arbitrary phase is an element of F_{sig} and can be used as initial distribution of the IFTA. Dependent on the application, the choice of the phase is of fundamental concern. In the following we assume phase freedom is available, but IFTA is also useful if this is not the case.

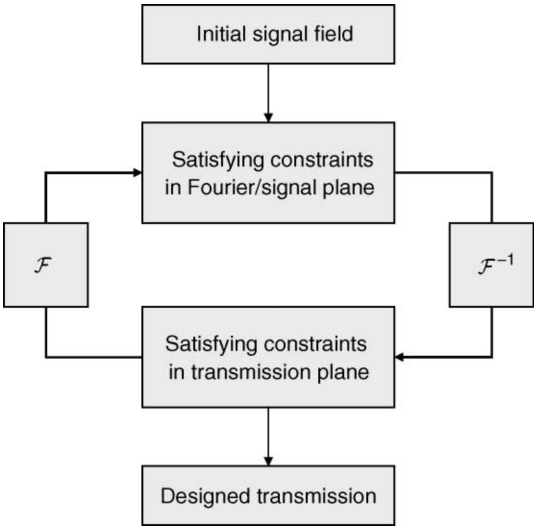
Initial signal field
Satisfying constraints in Fourier/signal plane
$\mathcal{F}$
$\mathcal{F}^{-1}$
Satisfying constraints in transmission plane
Designed transmission

Figure 11 Illustration of the basic principle of the iterative Fourier transform algorithm (IFTA). By a sequential satisfaction of constraints in the signal and the transmission plane, the design in the functional embodiment can be accomplished. If both constraints are satisfied accurate enough, the IFTA is terminated with a final satisfaction of the transmission plane constraints. The depicted IFTA version can be generalized by replacing $\mathcal{F}$ by $\mathcal{P}$ or any S.

It is helpful to distinguish between two types of signal fields. If amplitudes for only a finite number of discrete k-values, (i.e., angles or directions), need to be determined, the transmission acts as a diffractive beamsplitter. In all other cases more general light transformations are concerned.

The constraints in the transmission plane follow from the structure design. In most cases the structure design is done by TEA according to eqn [12]. Then, phase-only transmissions are required. Typically the fabrication of the corresponding element profile demands a phase quantization in addition. Both constraints have been intensively investigated for the IFTA.

As mentioned above, practically signal fields are always restricted to one or more signal regions of finite extent. For example, a detector has finite extent. Outside of these regions the signal may be unrestricted or only partially restricted. This freedom in amplitude outside the signal regions is fundamental for a successful design of phase-only transmissions. It has been shown, that the only phase-only function whose Fourier transform exists over a finite area is a linear phase. The transform of all other phase-only functions contribute energy outside of the signal regions. In other words, the use of amplitude freedom is mandatory. Energy losses due to stray light outside the signal region cannot be avoided. As a result, it is not possible to generate the desired response with 100% efficiency. For any given signal field U_{sig} an upper bound $\eta_{\mathrm{bound}}(U_{\mathrm{sig}})$ of the achievable efficiency $\eta(U_{\mathrm{sig}})$ of light transformation with a phase-only transmission has been derived, that is $\eta(U_{\mathrm{sig}}) \leq \eta_{\mathrm{bound}}(U_{\mathrm{sig}})$. The equal sign is only valid for the special case of a linear phase transmission, for which the signal field consists of one k-direction only. Then $\eta_{\mathrm{bound}}(U_{\mathrm{sig}}) = \eta(U_{\mathrm{sig}}) = 1$ is obtained. After structure design for a linear phase-only transmission with TEA, a so-called blazed grating results. If TEA is a sufficient approximation one concludes that a blazed grating has a theoretical efficiency of 100%. For all other signals the theoretical efficiency is smaller than 100%. Moreover, the upper bound is a function of U_{sig} and thus also of its phase. Therefore, in a first step of a transmission design it is recommendable to perform a phase synthesis which maximizes the upper bound. This first design step delivers information about achieveable efficiency and a suitable initial signal field for the IFTA.

If fabrication requires phase quantization, amplitude freedom can be used to separate quantization noise from the signal regions. Unfortunately this reduces the achievable efficiency even more. If the phase to be quantized is homogenously distributed, the upper bound for a phase-only transmission quantized to Q equidistant phase levels is given by:

$$\eta_{\mathrm{bound}}(U_{\mathrm{sig}}, Q) = \mathrm{sinc}^2\left[\frac{1}{Q}\right]\eta_{\mathrm{bound}}(U_{\mathrm{sig}}) \qquad [15]$$

For the special case of linear phase, the well-known formula $\eta(Q) = \mathrm{sinc}^2[1/Q]$ follows. The values of $\mathrm{sinc}^2[1/Q]$ and therefore of quantized blazed gratings in the paraxial approximation are 0.405, 0.812, 0.950, 0.987, and 0.997 for $Q = 2$, 4, 8, 16, and 32, respectively. This result does not mean that, for example, a diffractive beamsplitter with a splitting ratio $1 : M_{\mathrm{beams}}$ and 16 height levels will provide ~99% efficiency. According to eqn [15], the theoretical values are smaller and depend on M_{beams} because $\eta_{\mathrm{bound}}(U_{\mathrm{sig}})$ depends via U_{sig} on M_{beams}. If fabrication processes and the inherent imperfections are taken into consideration, the values decrease even more. As a rule of thumb, it is possible to claim that efficiencies larger than 90% are always challenging to obtain. With sophisticated design and fabrication techniques, efficiencies larger than 80% are reasonable.

Up to now, we have discussed only the importance of the signal field phase to efficiency. But the phase of the initial signal field has another extremely important effect on the result of the transmission design. From mapping-type considerations, a smooth initial signal field phase can be constructed in order to realize a light transformation. An example is illustrated in Figure 12. It shows the central part of the phase of the transmission function which transforms a Gaussian input beam into a flat-top line beam. Because of the smooth signal phase also the transmission phase is smooth and therefore unwrappable. This approach is typical for a so-called beam shaping design. It requires a well-defined input beam and high accuracy in adjustment of the beam shaping optics to obtain the output field with good quality. If, instead of a smooth initial phase, a random one is applied, the situation illustrated in Figure 13 (left) is achieved. Here the phase of the transmission looks random, which is typical for beamsplitter and diffuser. Because such phase functions possess numerous vortices, they do not allow an unwrapping. Thus, diffractive elements are the logical result of structure design. An example is shown in Figure 13 (right).

Figure 14 shows how light from a laser diode passes through the diffuser element of Figure 13 (right). The radiation is transformed into an intensity pattern that depicts the ‘skyline’ of Jena. Although globally the intensity possesses the requested shape, a closer look shows that the intensity distribution is speckled. However, in numerous applications this is not a problem so long as the size of the speckles is well adapted to the resolution of the detector.

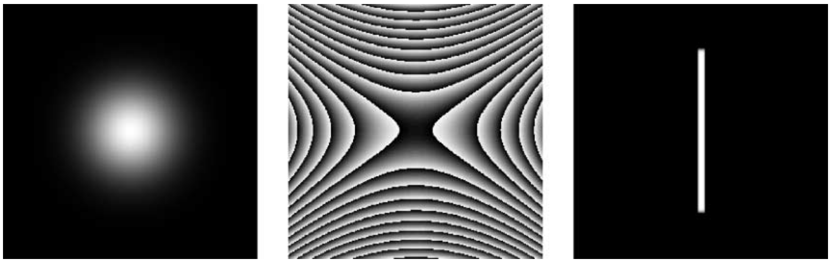

Figure 12 The input beam on the left propagates through the transmission function the central part of the phase of which is shown in the middle and in the signal plane the amplitude shown on the right results. This Gaussian-to-line transformation is a typical beam shaping design task.

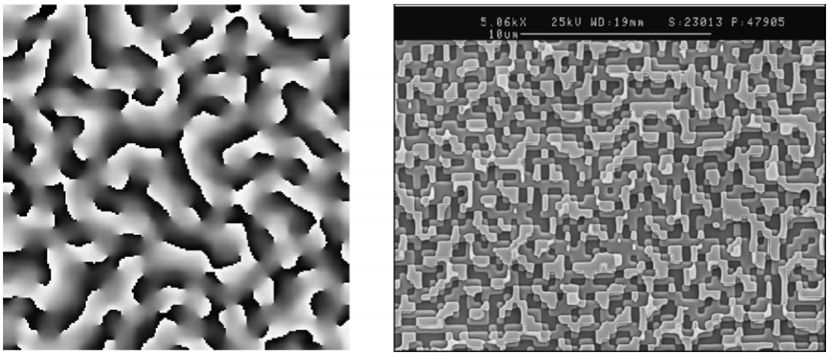

Figure 13 A typical phase distribution of a transmission (left) which results from design with IFTA for a random initial signal field phase. Numerous vortexes prevent its unwrapping. Therefore, structure design leads to diffractive elements. Then, a quantization of the phase is demanded in most cases. An example is shown on the right-handed side. It depicts a portion of a four level diffuser which results in the output field shown in Figure 14. The pixel size is 400 nm and the corresponding element was designed and fabricated in a cooperation of LightTrans GmbH and the Institute of Applied Physics at the University of Jena.

This technique is called diffuser-type light transformation. The appearance of the global intensity distribution is robust against variations in the input beam and thus is well-suited also for partial coherent illumination, for example radiation from an excimer laser and LED. In that case, even the appearance of speckles is reduced or vanishes completely. The mapping-type and the diffuser-type approach to choose the initial signal phase constitute two extreme cases of a great variety to control basic aspects of light transformation using different initial phases.

After choosing a suitable initial signal field $U_{\mathrm{sig}}(k_x, k_y)$ the IFTA typically performs well and a transmission function is obtained, which leads to $|U_{\mathrm{out}}(k_x, k_y)|^2 \propto |U_{\mathrm{sig}}(k_x, k_y)|^2$ in the signal region with minimum error. In the next step the structural design must be performed. The structure of the element is calculated via eqn [12].

Besides designing structure to produce the desired transmission function one needs to realize optically the Fourier transformation. Placing the observation plane in the far field behind the element is the most straightforward way to achieve this. However, field Fourier transformation can also be realized by lenses as shown in Figure 15. From simple geometric considerations using paraxial approximation one obtains:

$$\alpha = \frac{2\pi n_{\mathrm{free}}}{\lambda f_{\mathrm{e}}} \qquad [16]$$

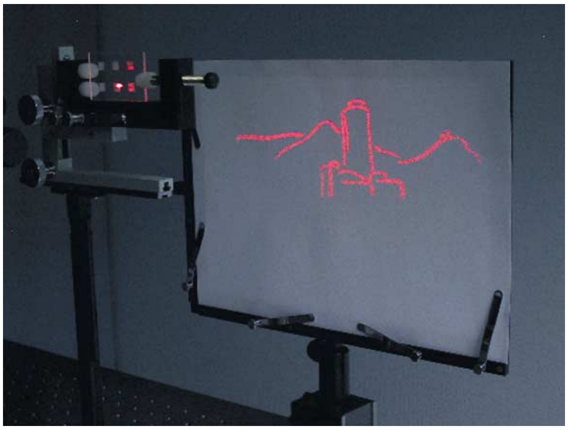

Figure 14 Optical demonstration of the use of a diffuser (see in Figure (right)) transforming a laser diode beam into an intensity displaying the ‘skyline’ of Jena. No additional optical element is needed! Stray light appears but is laterally separated from the desired signal and thus it is easy to be filtered.

Each system produces a Fourier transform that is scaled specially by α, where the output field is considered in a dielectric with refractive index n_{free}. The effective focal length f_{e} is equal to the focal length of the lens if the element is placed in front of the lens (see top of Figure 15), the distance of the element to the lens is of no concern for the resulting intensity, however it is recommended to minimize the distance between element and lens for elements with deflection angles close to the paraxial limit. That avoids light losses due to truncation by the lens aperture. If the focal length itself is chosen as the distance of the element to the lens, a so-called $2f$-setup results, which generate, in addition to the correct magnitude, the correct phase of the Fourier transformation in the focal plane. If an alternative distance to the lens is used, an additional quadratic phase occurs in the focal plane. If the element is placed behind the lens (see middle row of Figure 15), f_{e} is the distance between element and focal plane. This setup provides some freedom in adjusting the output scaling. On the other hand, it requires an accurate adjustment of position to obtain a specific scaling. This is not required for the first setup. In order to obtain a specific

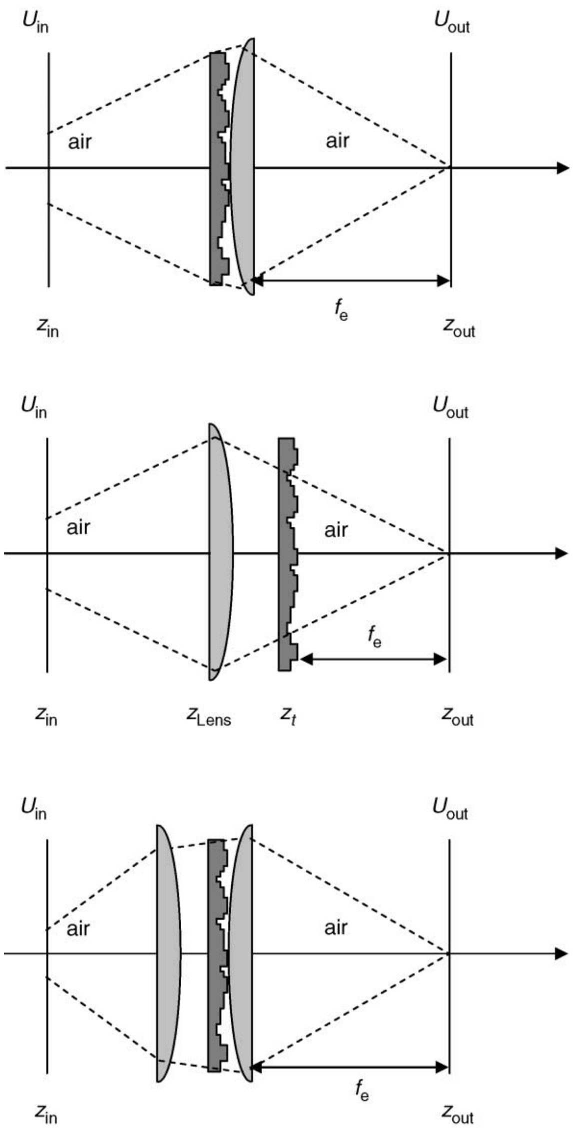

Figure 15 Three setups to realize a Fourier transformation for a beam shaping or diffuser/splitter element, which deflects the light only in a paraxial way. In all cases the output is obtained in the focal plane of the lens.

scaling of the output field $U_{\text{out}}(\alpha x, \alpha y)$, the effective focal length and the size of $U_{\text{sig}}(k_x, k_y)$ in the k-space must be chosen properly according to eqn [16]. In addition to the size of the output field, its resolution is also of fundamental importance. According to eqn [14] and the convolution theorem the output field is given by

$$U_{\text{out}}(k_x, k_y) = \mathcal{F} \exp[\mathrm{i}\tau(x, y)] * \mathcal{F}\mathcal{P}(\Delta z_{\text{in}}) U_{\text{in}}(x, y) \qquad [17]$$

where $*$ denotes the convolution. For a constant transmission function one would obtain the focus of the input beam. In functional embodiment that is $\mathcal{F}\mathcal{P}(\Delta z_{\text{in}}) U_{\text{in}}(x, y)$ only. In the structural form, the focus of the beam must be calculated by propagating the input field through the system without the transmission function. The Fourier transformed transmission function, that is $\mathcal{F} \exp[\mathrm{i}\tau(x, y)]$, is convolved with this beam focus in order to obtain the resulting output field. In beamsplitting that means a replication of the beam. In general the output field can be understood as being composed of numerous replicated beam foci which are overlapping. That makes plausible that the size of the focus of the input beam limits the resolution of the whole output field. The size of the focus is determined by the basic beam focusing lens system and not by the diffractive or refractive element. In conclusion, a highly resolved output field requires a focusing optic with high numerical aperture. For examples, it is not possible to obtain a flat-top beam profile from a Gaussian beam where focus has the same size as the flat-top should have. The steeper the sidewalls of the flat-top profile, the smaller must be the focal spot without the beam shaping element. The choice of a suitable numerical aperture of the lens system must be done before the transmission design is performed because it also determines the extent of the transmission function. The distance Δz_{in} must be chosen in such a way that ensures that the propagated input field illuminates the whole aperture of the system.

If a high numerical aperture on the output side is required to achieve a high resolution, the setup in the middle of Figure 15 should be avoided and the first setup is preferable. Otherwise the element would be illuminated by a highly nonparaxial field which violates the structure design with TEA. With the same argument, but now on the input side, an input field with high divergence should be at least partly collimated to achieve a paraxial illumination of the element. Then, the setup at the bottom of Figure 15 results.

For nonparaxial light transformation, that means the output field is large and hence the transformation requires nonparaxial deflection of the input field; the described transmission and structure design is still useful to obtain an initial design for a further optimization by wave-optical modeling techniques.

Though a Fourier approach to obtain general light transformations seems rather restricted, it covers a broad field of applications. The combination of suitable beam optics with the refractive free-form or diffractive element can be very simple (see optical setup of Figure 14) or very complex. That mainly depends on the resolution and size of the demanded output field as it is well known from imaging systems.

Each design should be finalized with a complete wave-optical analysis, typically including adjustment and fabrication error tolerancing, to achieve a full understanding of the optical properties of the system. After this final investigation it is ready for fabrication and application providing a further innovation through optics and photonics on the base of wave-optical engineering.

See also

Fourier Optics.

Further Reading

Aagedal H, Schmid M, Beth T, Teiwes S and Wyrowski F (1996) Theory of speckles in diffractive optics and its application to beam shaping. *Journal of Modern Optics* 43(7): 1409–1421.

Aagedal H, Schmid M, Egner S, Müller-Quade J, Beth T and Wyrowski F (1997) Analytical beam shaping with application to laser-diode arrays. *Journal of the Optical Society of America A* 14(7): 1549–1553.

Aagedal H, Wyrowski F and Schmid M (1997) Paraxial beamsplitting and shaping. In: Turunen J and Wyrowski F (eds) *Diffractive Optics for Industrial and Commercial Applications*, chapter 6, pp. 165–188. Berlin: Akademie Verlag.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. New York: Pergamon Press.

Bryngdahl O (1974) Optical map transformations. *Optics Communications* 10(2): 164–168.

Bühling S, Schimmel H and Wyrowski F (2003) Wave-optical engineering with virtuallab. *Proceedings SPIE*, 5182: 24–33.

Bühling S, Schimmel H and Wyrowski F (2003) Challenges in industrial applications of diffractive beam splitting. *Proceedings SPIE* 5182: 123–131.

Collier RJ, Burckhardt CB and Lin LH (1971) *Optical Holography*. London: Academic Press.

Fienup JR (1980) Iterative method applied to image reconstruction and to computer-generated holograms. *Optical Engineering* 19(3): 297–305.

Gerchberg RW and Saxton WO (1972) A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35(2): 237–246.

Goodmann JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Goodman JW and Silvestri AM (1970) Some effects of Fourier-domain phase quantization. *IBM Journal of Reserved Development* 14: 478–484.

Krackhardt U, Mait JN and Streibl N (1992) Upper bound on the diffraction efficiency of phase-only fanout elements. *Applied Optics* 31(1): 27–37.

Lajunen H, Tervo J, Turunen J, Vallius T and Wyrowski F (2003) Simulation of light propagation by local spherical interface approximation. *Applied Optics* 34: 6804–6810.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge: Cambridge University Press.

Pfeil AV and Wyrowski F (2000) Wave-optical structure design with the local plane-interface approximation. *Journal of Modern Optics* 47(13): 2335–2350.

Pfeil AV, Wyrowski F, Drauschke A and Aagedal H (2000) Analysis of optical elements with the local plane-interface approximation. *Applied Optics* 39(19): 3304–3313.

Schimmel H, Bühling S and Wyrowski F (2003) Use of ray tracing in wave-optical engineering. *Proceedings SPIE* 5182: 6–14.

Schimmel H and Wyrowski F (2000) Amplitude matching strategy for wave-optical design of monofunctional systems. *Journal of Modern Optics* 47(13): 2295–2321.

Sergienko N, Stamnes JJ, Kettunen V, Kuittinen M, Turunen J, Vahimaa P and Friberg AT (1999) Comparison of electromagnetic and scalar methods for evaluation of diffractive lenses. *Journal of Modern Optics* 46(1): 65–82.

Smith WJ (1997) *Practical Optical System Layout*. New York: McGraw-Hill.

Smith WJ (2000) *Modern Optical Engineering*, 3rd edn. Bellingham: SPIE Press.

Turunen J (1997) Diffraction theory of microrelief gratings. In: Herzig HP (ed.) *Micro-Optics Elements, Systems and Applications*, chapter 2, pp. 31–52. London: Taylor & Francis.

Turunen J and Wyrowski F (1997) *Diffractive Optics for Industrial and Commercial Applications*. Berlin: Akademie Verlag.

Vallius T, Jefimovs K, Kettunen V, Kuittinen M, Laakkonen P and Turunen J (2001) Design of non-paraxial array illuminators by step-transition perturbation approach. *Journal of Modern Optics* 48(12): 1869–1879.

Wyrowski F (1990) Diffractive optical elements: Iterative calculation of quantized, blazed phase structures. *Journal of the Optical Society of America A* 7: 961–969.

Wyrowski F (1991) Digital phase encoded inverse filter for optical pattern recognition. *Applied Optics* 30: 4650–4657.

Wyrowski F (1991) Considerations on convolution and phase factors. *Optics Communications* 81(6): 353–358.

Wyrowski F (1991) Upper bound of the diffraction efficiency of diffractive phase elements. *Optics Letters* 16(24): 1915–1917.

Wyrowski F (1993) Design theory of diffraction elements in the paraxial domain. *Journal of the Optical Society of America A* 10: 1553–1561.

Wyrowski F (2003) Comments on wave-optical engineering. *Proceedings SPIE* 5182: 1–5.

Wyrowski F and Turunen J (2002) Wave-optical engineering. In: Guenther AH (ed.) *International Trends in Applied Optics*, chapter 21, pp. 471–496. Bellingham: SPIE Press.

DISPERSION MANAGEMENT

A E Willner, Y-W Song, J Mcgeehan and Z Pan, University of Southern California, Los Angeles, CA, USA

B Hoanca, University of Alaska Anchorage, Anchorage, AK, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical communications systems have grown explosively in terms of the capacity that can be transmitted over a single optical fiber. This trend has been fueled by two complementary techniques, those being the increase in data-rate-per-channel coupled with the increase in the total number of parallel wavelength channels. However, there are many considerations as to the total number of wavelength-division-multiplexed (WDM) channels that can be accommodated in a system, including cost, information spectral efficiency, nonlinear effects, and component wavelength selectivity.

Dispersion is one of the critical roadblocks to increasing the transmission capacity of optical fiber. The dispersive effect in an optical fiber has several ingredients, including intermodal dispersion in a multimode fiber, waveguide dispersion, material dispersion, and chromatic dispersion. In particular, chromatic dispersion is one of the critical effects in a single mode fiber (SMF), resulting in a temporal spreading of an optical bit as it propagates along the fiber. At data rates ≤ 2.5 Gbit/s, the effects of chromatic dispersion are not particularly troublesome. For data rates ≥ 10 Gbit/s, however, transmission can be tricky and the chromatic dispersion-induced degrading effects must be dealt with in some way, perhaps by compensation. Furthermore, the effects of chromatic dispersion rise quite rapidly as the bit rate increases – when the bit rate increases by a factor of four, the effects of chromatic dispersion increase by a factor of 16! This article will only deal with the management of chromatic dispersion in single mode fiber.

One of the critical limitations of optical fiber communications comes from chromatic dispersion, which results in a pulse broadening as it propagates along the fiber. This occurs as photons of different frequencies (created by the spreading effect of data modulation) travel at different speeds, due to the frequency-dependent refractive index of the fiber core. Compounding the problems cause by chromatic dispersion is the fact that as bit rates rise, chromatic dispersion effects rise quadratically with respect to the increase in the bit rate. One can eliminate these effects using fiber with zero chromatic dispersion, known as dispersion shifted fiber (DSF). However, with zero dispersion, all channels in a wavelength-division-multiplexed (WDM) system travel at the same speed, in-phase, and a number of deleterious nonlinear effects such as cross-phase modulation (XPM) and four-wave mixing (FWM) result. Thus, in WDM systems, some amount of chromatic dispersion is necessary to keep channels out-of-phase, and as such chromatic dispersion compensation is required. Any real fiber link may also suffer from ‘dispersion slope’ effects, in which a slightly different dispersion value is produced in each WDM channel. This means that while one may be able to compensate one channel exactly, other channels may progressively accumulate increasing amounts of dispersion, which can severely limit the ultimate length of the optical link and the wavelength range that can be used in a WDM system. This article will address the concepts of chromatic dispersion and dispersion slope management followed by some examples highlighting the need for tunability to enable robust optical WDM systems in dynamic environments. Some dispersion monitoring techniques are then discussed and examples given.

Chromatic Dispersion in Optical Fiber Communication Systems

In any medium (other than vacuum) and in any waveguide structure (other than ideal infinite free space), different electromagnetic frequencies travel at different speeds. This is the essence of chromatic dispersion. As the real fiber-optic world is rather distant from the ideal concepts of both vacuum and infinite free space, dispersion will always be a concern when one is dealing with the propagation of electromagnetic radiation through fiber. The velocity in fiber of a single monochromatic wavelength is constant. However, data modulation causes a broadening of the spectrum of even the most monochromatic laser pulse. Thus, all modulated data have a nonzero spectral width which spans several wavelengths, and the different spectral components of modulated data travel at different speeds. In particular, for digital data intensity modulated on an optical carrier, chromatic dispersion leads to pulse broadening – which in turn leads to chromatic dispersion limiting the maximum data rate that can be transmitted through optical fiber (see Figure 1).

Considering that the chromatic dispersion in optical fibers is due to the frequency-dependent nature of the propagation characteristics, for both the material (the refractive index of glass) and the waveguide structure, the speed of light of a particular wavelength λ will be expressed as follows, using a Taylor series expansion of the value of the refractive index as a function of the wavelength:

$$v(\lambda) = \frac{c_0}{n(\lambda)} = \frac{c_0}{n_0(\lambda_0) + \dfrac{\partial n}{\partial \lambda}\delta\lambda + \dfrac{\partial^2 n}{\partial \lambda^2}(\delta\lambda)^2} \qquad [1]$$

Here, c_0 is the speed of light in vacuum, λ_0 is a reference wavelength, and the terms in $\partial n/\partial\lambda$ and $\partial^2 n/\partial\lambda^2$ are associated with the chromatic dispersion and the dispersion slope (i.e., the variation of the chromatic dispersion with wavelength), respectively. Transmission fiber has positive dispersion, i.e. longer wavelengths result in longer propagation delays. The units of chromatic dispersion are picoseconds per nanometer per kilometer, meaning that shorter time pulses, wider frequency spread due to data modulation, and longer fiber lengths will each contribute linearly to temporal dispersion. Higher data rates inherently have both shorter pulses and wider frequency spreads. Therefore, as network speed increases, the impact of chromatic dispersion rises precipitously as the square of the increase in data rate. The quadratic increase with the data rate is a result of two effects, each with a linear contribution. On one hand, a doubling of the data rate makes the spectrum twice as wide, doubling the effect of dispersion. On the other hand, the same doubling of the data rate makes the data pulses only half as long (hence twice as sensitive to dispersion). The combination of a wider signal spectrum and a shorter pulse width is what leads to the overall quadratic impact. Moreover, the data modulation format used can significantly affect the sensitivity of a system to chromatic dispersion. For example, the common nonreturn-to-zero (NRZ) data format, in which the optical power stays high throughout the entire time slot of a ‘1’ bit, is more robust to chromatic dispersion than is the return-to-zero (RZ) format, in which the optical power stays high in only part of the time slot of a ‘1’ bit. This difference is due to the fact that RZ data have a much wider channel frequency spectrum compared to NRZ data, thus incurring more chromatic dispersion. However, in a real WDM system, the RZ format increases the maximum allowable transmission distance by virtue of its reduced duty cycle (compared to the NRZ format), making it less susceptible to fiber nonlinearities as can be seen in Figure 2.

A rule for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by

$$L_\mathrm{D} = \frac{1}{D\cdot B\cdot\Delta\lambda} = \frac{1}{D\cdot B\cdot(cB)} \propto \frac{1}{B^2} \qquad [2]$$

(see Figure 3). For example, for single mode fiber, $D = 17$ ps/nm/km, so for 10 Gbit/s data the distance is $L_\mathrm{D} = 52$ km. In fact, a more exact calculation

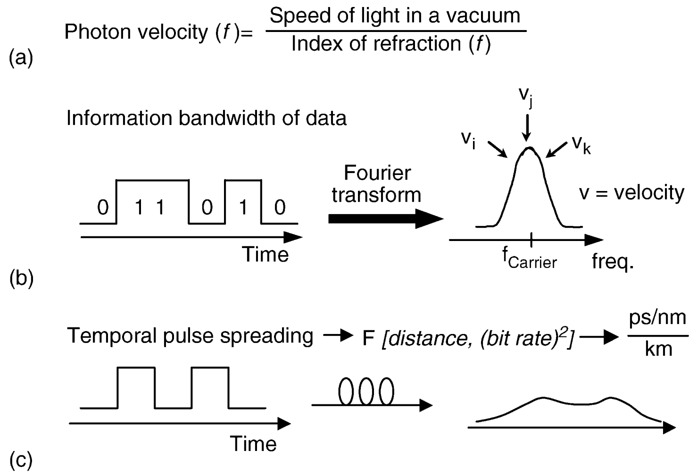
Photon velocity (f)= Speed of light in a vacuum / Index of refraction (f)
(a)
Information bandwidth of data
0
1 1
0
1
0
Time
Fourier transform
v_j
v_i
v_k
v = velocity
f_Carrier
freq.
(b)
Temporal pulse spreading
F [distance, (bit rate)²]
ps/nm / km
000
Time
(c)

Figure 1 The origin of chromatic dispersion in data transmission. (a) Chromatic dispersion is caused by the frequency-dependent refractive index in fiber. (b) The nonzero spectral width due to data modulation. (c) Dispersion leads to pulse broadening, proportional to the transmission distance and the data rate.

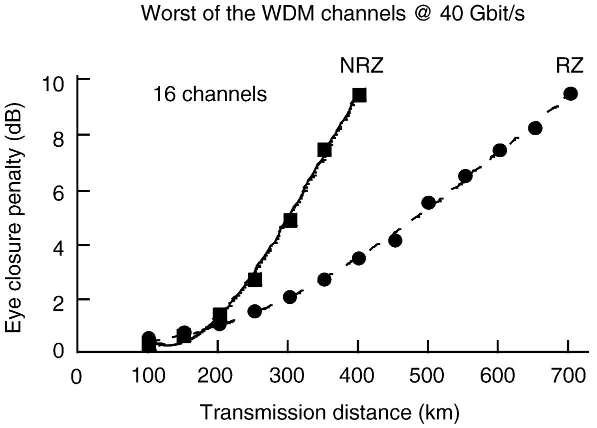

Figure 2 Performances of RZ and NRZ formats in a real fiber transmission link. (Reproduced with permission from Hayee I and Willner AE (1999) NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems. *IEEE Photon. Tech. Lett.* 11(8): 991–993.)

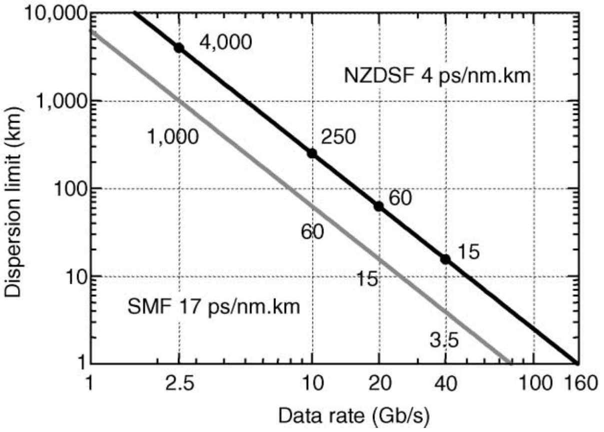

Figure 3 Transmission distance limitations due to uncompensated dispersion in SMF as a function of data rate for intensity modulated optical signals. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference.*)

shows that for 60 km, the dispersion induced power penalty is less than 1 dB (see Figure 4). The power penalty for uncompensated dispersion rises exponentially with transmission distance, and thus to maintain good signal quality, dispersion compensation is required.

Chromatic Dispersion Management

Optical Nonlinearities as Factors to be Considered in Dispersion Compensation

Even though it is possible to manufacture fiber with zero dispersion, it is not practical to use such fiber for WDM transmission, due to large penalties induced by fiber nonlinearities. Most nonlinear effects originate from the nonlinear refractive index of fiber, which is

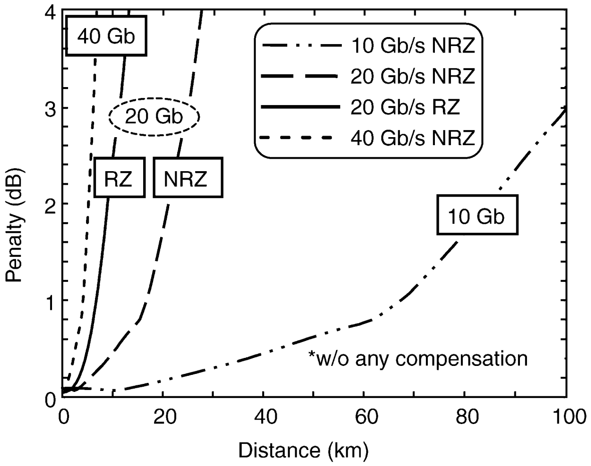

Figure 4 Power penalties due to uncompensated dispersion in single mode fiber (SMF) as a function of distance and data rate. (Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference.*)

not only dependent on the frequency of light but also on the intensity (optical power), and is related to the optical power as:

$$\bar{n}(f,P) = n(f) + n_2 \frac{P}{A_{\text{eff}}} \qquad [3]$$

where $n(f)$ is the linear part of the refractive index, P is the optical power inside the fiber, and n_2 is the nonlinear-index coefficient for silica fibers. The typical value of n_2 is 2.6×10^{-20} m^2/W. This number takes into account the averaging of the polarization states of the light as it travels in the fiber. The intensity dependence of the refractive index gives rise to three major nonlinear effects.

Self-phase modulation (SPM)

A million photons 'see' a different glass than does a single photon, and a photon traveling along with many other photons will slow down. SPM occurs because of the varying intensity profile of an optical pulse on a single WDM channel. This intensity profile causes a refractive index profile and, thus, a photon speed differential. The resulting phase change for light propagating in an optical fiber is expressed as:

$$\Phi_{\text{NL}} = \gamma P L_{\text{eff}} \qquad [4]$$

where the quantities γ and L_{eff} are defined as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}} \text{ and } L_{\text{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \qquad [5]$$

where A_{eff} is the effective mode area of the fiber and α is the fiber attenuation loss. L_{eff} is the effective nonlinear length of the fiber that accounts for fiber loss, and γ is the nonlinear coefficient measured in rad/km/W. A typical range of values for γ is between 10–30 rad/km/W. Although the nonlinear coefficient is small, the long transmission lengths and high optical powers, that have been made possible by the use of optical amplifiers, can cause a large enough nonlinear phase change to play a significant role in state-of-the-art lightwave systems.

Cross-phase modulation (XPM)

When considering many WDM channels co-propagating in a fiber, photons from channels 2 through N can distort the index profile that is experienced by channel 1. The photons from the other channels 'chirp' the signal frequencies on channel 1, which will interact with fiber chromatic dispersion and cause temporal distortion. This effect is called cross-phase modulation. In a two-channel system, the frequency chirp in channel 1, due to power fluctuation within both channels, is given by

$$\Delta B = \frac{\mathrm{d}\Phi_{\mathrm{NL}}}{\mathrm{d}t} = \gamma L_{\mathrm{eff}}\frac{\mathrm{d}P_1}{\mathrm{d}t} + 2\gamma L_{\mathrm{eff}}\frac{\mathrm{d}P_2}{\mathrm{d}t} \qquad [6]$$

where, $\mathrm{d}P_1/\mathrm{d}t$ and $\mathrm{d}P_2/\mathrm{d}t$ are the time derivatives of the pulse powers of channels 1 and 2, respectively. The first term on the right-hand side of the above equation is due to SPM, and the second term is due to XPM. Note that the XPM-induced chirp term is double that of the SPM-induced chirp term. As such, XPM can impose a much greater limitation on WDM systems than can SPM, especially in systems with many WDM channels.

Four-wave-mixing (FWM)

The optical intensity propagating through the fiber is related to the electric field intensity squared. In a WDM system, the total electric field is the sum of the electric fields of each individual channel. When squaring the sum of different fields, products emerge that are beat terms at various sum and difference frequencies to the original signals. Figure 5 depicts that if a WDM channel exists at one of the four-wave-mixing beat-term frequencies, then the beat term will interfere coherently with this other WDM channel and potentially destroy the data.

Dispersion Maps

While zero-dispersion fiber is not a good idea, a large value of the accumulated dispersion at the end of a fiber link is also undesirable. An ideal solution is to have a 'dispersion map,' alternating sections of positive and negative dispersion as can be seen in Figure 6. This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, eliminating FWM and XPM, but the total dispersion at the end of the fiber link is zero, so that no pulse broadening is induced (Table 1). The most advanced systems require periodic dispersion compensation, as well as pre- and post-compensation (before and after the transmission fiber).

The addition of negative dispersion to a standard fiber link has been traditionally known as 'dispersion compensation,' however, the term 'dispersion management' is more appropriate. SMF has positive dispersion, but some new varieties of nonzero dispersion-shifted fiber (NZDSF) come in both positive and negative dispersion varieties. Some examples are shown in Figure 7. Reverse dispersion fiber is also now available, with a large dispersion

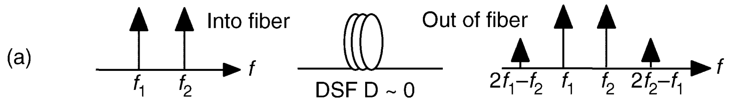

(b) Degradation on optical spectrum (Unequal channel spacing)

(c) Degradation on bit stream (equal channel spacing)

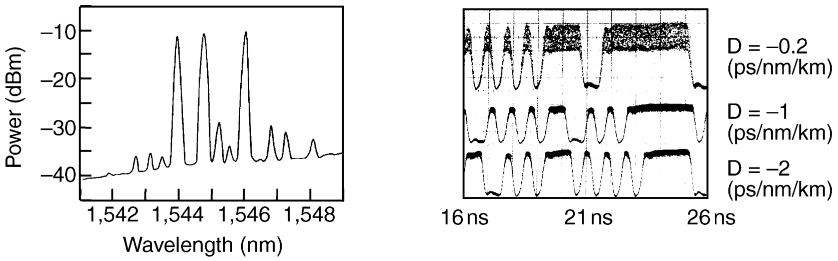

Figure 5 (a) and (b) FWM induces new spectral components via nonlinear mixing of two wavelength signals. (c) The signal degradation due to FWM products falling on a third data channel can be reduced by even small amounts of dispersion. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnauck AH and Derosier RM (1995) Four-photon mixing and high-speed WDM systems. *Journal of Photon Technology* 13(5): 841–849.)

comparable to that of SMF, but with the opposite sign. When such flexibility is available in choosing both the magnitude and sign of the dispersion of the fiber in a link, dispersion-managed systems can be fully optimized to the desired dispersion map using a combination of fiber and dispersion compensation devices (see Figure 8).

Dispersion is a linear process, so first-order dispersion maps can be understood as linear systems. However, the effects of nonlinearities cannot be ignored, especially in WDM systems, with many tens of channels, where the launch power may be very high. In particular, in systems deploying dispersion compensating fiber (DCF), the large nonlinear coefficient of the DCF can dramatically affect the dispersion map.

Corrections to Linear Dispersion Maps

Chromatic dispersion is a necessity in WDM systems, to minimize the effects of fiber nonlinearities. A chromatic dispersion value as small as a few ps/nm/km is usually sufficient to make XPM and FWM negligible. To mitigate the effects of nonlinearities but maintain small amounts of chromatic dispersion, NZDSF is commercially available. Due to these nonlinear effects, chromatic dispersion must be managed, rather than eliminated.

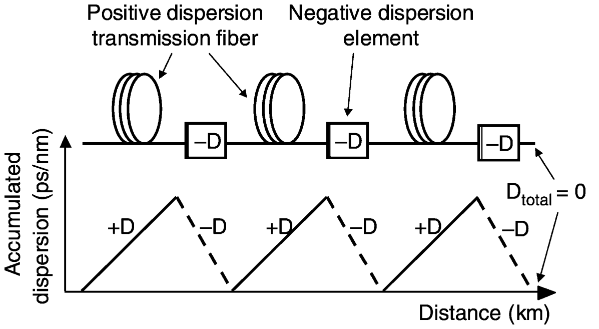
Positive dispersion transmission fiber
Negative dispersion element
–D
–D
–D
Accumulated dispersion (ps/nm)
$D_{total} = 0$
+D
–D
+D
–D
+D
–D
Distance (km)

Figure 6 Dispersion map of a basic dispersion managed system. Positive dispersion transmission fiber alternates with negative dispersion compensation elements such that the total dispersion is zero end-to-end.

If a dispersion-management system was perfectly linear, it would be irrelevant whether the dispersion along a path is small or large, as long as the overall dispersion is compensated to zero (end to end). Thus, in a linear system the performance should be similar, regardless of whether the transmission fiber is SMF, and dispersion compensation modules are deployed every 60 km, or the transmission fiber is NZDSF (with approximately a quarter of the dispersion value of SMF) and dispersion compensation modules are deployed every 240 km. In real life, optical nonlinearities are very important, and recent results seem to favor the use of large, SMF-like, dispersion values in the transmission path and correspondingly high dispersion compensation devices. A recent study of performance versus channel spacing showed that the capacity of SMF could be more than four times that of NZDSF. This is because the nonlinear coefficients are much higher in NZDSF than in SMF, and for dense WDM the channel interactions become a limiting factor. A critical conclusion is that not all dispersion compensation maps are created equal: a simple calculation of the dispersion compensation, to cancel the overall dispersion value, does not lead to optimal dispersion map designs.

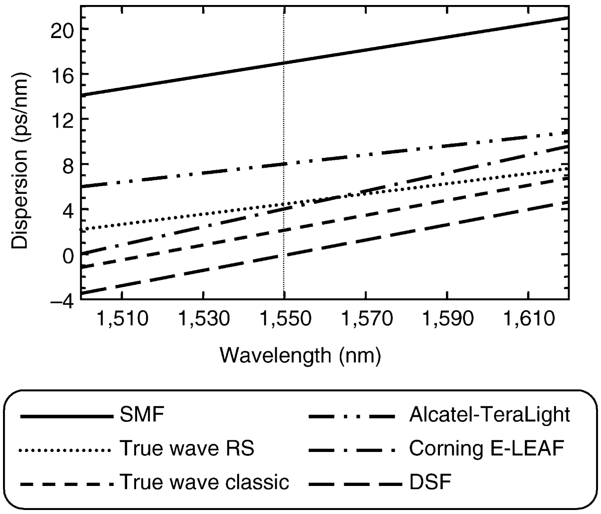

Figure 7 Chromatic dispersion characteristics of various commercially available types of transmission fiber.

Table 1 Commercially available fibers and their characteristics

	Dispersion @ 1,550 nm [ps/nm/km]	*Dispersion slope [ps/nm^2/km]*	*Attenuation [dB]*	*Mode field diameter [μm]*	*PMD [ps/km$^{0.5}$]*
SMF	18	0.08	≤0.25	11	≤0.5
Tera Light	8.0	0.058	≤0.2	9.2	≤0.04
TW-RS	4.4	0.043	≤0.25	8.4	≤0.03
LEAF	4.0	0.085	≤0.25	9.6	≤0.08
Standard DCF	−90	−0.22	≤0.5	5.2	≤0.08

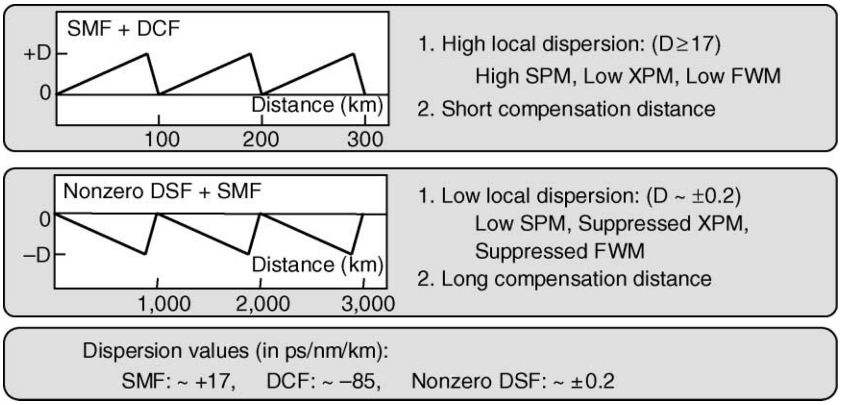
SMF + DCF
+D
0
Distance (km)
100
200
300
1. High local dispersion: (D≥17)
High SPM, Low XPM, Low FWM
2. Short compensation distance
Nonzero DSF + SMF
0
−D
Distance (km)
1,000
2,000
3,000
1. Low local dispersion: (D ~ ±0.2)
Low SPM, Suppressed XPM,
Suppressed FWM
2. Long compensation distance
Dispersion values (in ps/nm/km):
SMF: ~ +17, DCF: ~ −85, Nonzero DSF: ~ ±0.2

Figure 8 Various dispersion maps for SMF-DCF and NZDSF-SMF.

Additionally, several solutions have been shown to be either resistant to dispersion, or have been shown to rely on dispersion itself for transmission. Such solutions include chirped pulses (where prechirping emphasizes the spectrum of the pulses so that dispersion does not broaden them too much), dispersion assisted transmission (where an initial phase modulation tailored to the transmission distance leads to full-scale amplitude modulation at the receiver end due to the dispersion), and various modulation formats robust to chromatic dispersion and nonlinearities.

Dispersion Management Solutions

Fixed Dispersion Compensation

From a systems point of view, there are several requirements for a dispersion compensating module: low loss, low optical nonlinearity, broadband (or multichannel) operation, small footprint, low weight, low power consumption, and clearly low cost. It is unfortunate that the first dispersion compensation modules, based on DCF only, met two of these requirements: broadband operation and low power consumption. On the other hand, several solutions have emerged that can complement or even replace these first-generation compensators.

Dispersion compensating fiber (DCF)

One of the first dispersion compensation techniques was to deploy specially designed sections of fiber with negative chromatic dispersion. The technology for DCF emerged in the 1980s and has developed dramatically since the advent of optical amplifiers in 1990. DCF is the most widely deployed dispersion compensator, providing broadband operation and stable dispersion characteristics, and the lack of a dynamic, tunable DCF solution has not reduced its popularity.

As can be seen in Figure 9, the core of the average dispersion compensating fiber is much smaller than that of standard SMF, and beams with longer wavelengths experience relatively large changes in mode size (due to the waveguide structure) leading to greater propagation through the cladding of the fiber, where the speed of light is greater than that of the core. This leads to a large negative dispersion value. Additional cladding layers can lead to improved DCF designs that can include negative dispersion slope to counteract the positive dispersion slope of standard SMF.

In spite of its many advantages, DCF has a number of drawbacks. First, it is limited to a fixed compensation value. In addition, DCF has a weakly guiding structure and has a much smaller core cross-section, 19 μm^2, compared to the 85 μm^2 of SMF. This leads to higher nonlinearity, higher splice losses, as well as higher bending losses. Last, the length of DCF required to compensate for SMF dispersion is rather long, about one-fifth of the length of the transmission fiber for which it is compensating. Thus DCF modules induce loss, and are relatively bulky and heavy. The bulk is partly due to the mass of fiber, but also due to the resin used to hold the fiber securely in place. One other contribution to the size of the module is the higher bend loss associated with the refractive index profile of DCF; this limits the radius of the DCF loop to 6–8 inches, compared to the minimum bend radius of 2 inches for SMF.

Traditionally, DCF-based dispersion compensation modules are usually located at amplifier sites. This serves several purposes. First, amplifier sites offer relatively easy access to the fiber, without requiring any digging or unbraiding of the cable. Second, DCF has high loss (usually at least double that of standard SMF), so a gain stage is required

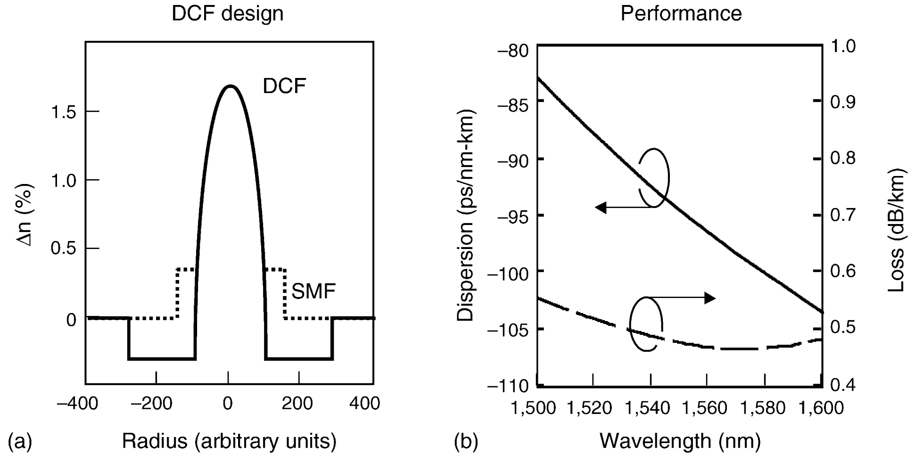

Figure 9 Typical DCF (a) refractive index profile and (b) dispersion and loss as a function of wavelength. Δ*n* is defined as refractive index variation relative to the cladding.

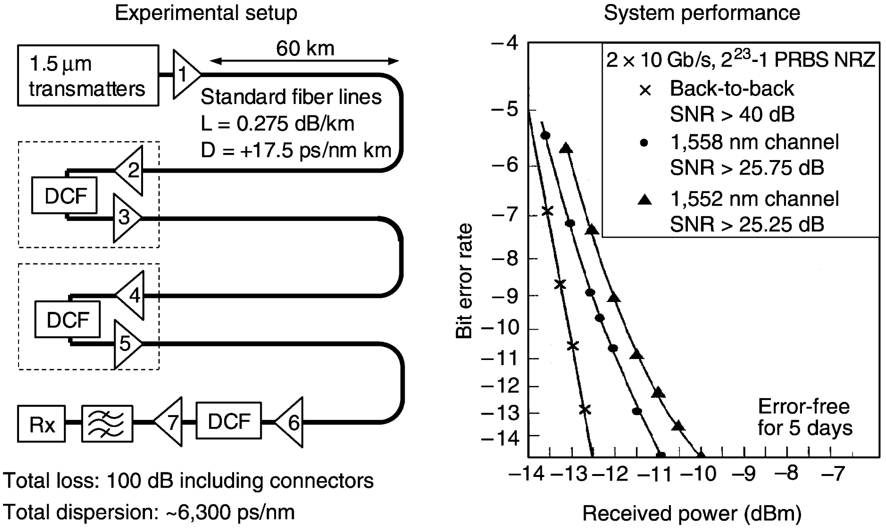

Figure 10 System demonstration of dispersion compensation using DCF. (Reproduced with permission from Park YK, Yeates PD, Delavaux J-MP, *et al.* (1995) A field demonstration of 20-Gb/s capacity transmission over 360 km of installed standard (non-DSF) fiber. *Photon. Technol. Lett.* 7(7): 816–818.)

before the DCF module to avoid excessively low signal levels. DCF has a cross-section four times smaller then SMF, hence a higher nonlinearity, which limits the maximum launch power into a DCF module. The compromise is to place the DCF in the mid-section of a two-section EDFA. This way, the first stage provides pre-DCF gain, but not to a power level that would generate excessive nonlinear effects in the DCF. The second stage amplifies the dispersion compensated signal to a power level suitable for transmission though the fiber link. This launch power level is typically much higher than could be transmitted through DCF without generating large nonlinear effects. Many newer dispersion compensation devices have better performance than DCF, in particular lower loss and lower nonlinearities. For this reason, they may not have to be deployed at the mid-section of an amplifier. Figure 10 shows the real demonstration results using the DCF.

Chirped fiber Bragg gratings

Fiber Bragg gratings have emerged as major components for dispersion compensation because of their low loss, small footprint, and low optical nonlinearity. Bragg gratings are sections of single-mode fiber in which the refractive index of the core is modulated in a periodic fashion, as a function of the spatial coordinate along the length of the fiber. When the spatial periodicity of the modulation matches what is known as a Bragg condition with respect to the wavelength of light propagating through the grating, the periodic structure acts like a mirror, reflecting the optical radiation that is traveling through the core of the fiber. An optical circulator is traditionally used to separate the reflected output beam from the input beam.

When the periodicity of the grating is varied along its length, the result is a chirped grating which can be used to compensate for chromatic dispersion. The chirp is understood as the rate of change of the spatial frequency as a function of position along the grating. In chirped gratings the Bragg matching condition for different wavelengths occurs at different positions along the grating length. Thus, the roundtrip delay of each wavelength can be tailored by designing the chirp profile appropriately. Figure 11 compares the chirped FBG with uniform FBG. In a data pulse that has been distorted by dispersion, different frequency components arrive with different amounts of relative delay. By tailoring the chirp profile such that the frequency components see a relative delay which is the inverse of the delay of the transmission fiber, the pulse can be compressed back. The dispersion of the grating is the slope of the time delay as a function of wavelength, which is related to the chirp.

The main drawback of Bragg gratings is that the amplitude profile and the phase profile as a function of wavelength have some amount of ripple. Ideally, the amplitude profile of the grating should have a flat (or rounded) top in the passband, and the phase profile should be linear (for linearly chirped gratings) or polynomial (for nonlinearly chirped gratings). The grating ripple is the deviation from the ideal profile shape. Considerable effort has been expended on reducing the ripple. While early gratings were plagued by more than 100 ps of ripple, published results have shown vast improvement to values close to ±3 ps.

Higher order mode dispersion compensation fiber

One of the challenges of designing standard DCF is that high negative dispersion is hard to achieve unless the cross-section of the fiber is small (which leads to high nonlinearity and high loss). One way to reduce both the loss and the nonlinearity is to use a higher-order mode (HOM) fiber (LP_{11} or LP_{02} near cutoff instead of the LP_{01} mode in the transmission fiber).

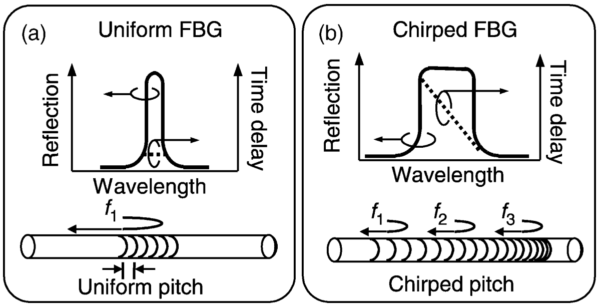

Figure 11 Uniform and chirped FBGs. (a) A grating with uniform pitch has a narrow reflection spectrum and a flat time delay as a function of wavelength. (b) A chirped FBG has a wider bandwidth, a varying time delay and a longer grating length.

Such a device requires a good-quality mode converter between LP_{01} and LP_{02} to interface between the SMF and HOM fiber. HOM fiber has a dispersion per unit length greater than six times that of DCF. Thus, to compensate for a given transmission length in SMF, the length of HOM fiber required is only one sixth the length of DCF. Thus, even though losses and nonlinearity per unit length are larger for HOM fiber than for DCF, they are smaller overall, because of the shorter HOM fiber length. As an added bonus, the dispersion can be tuned slightly by changing the cutoff wavelength of LP_{02} (via temperature tuning). A soon to be released HOM fiber-based commercial dispersion compensation module is not tunable, but can fully compensate for dispersion slope.

Tunable Dispersion Compensation

The need for tunability

In a perfect world, all fiber links would have a known, discrete, and unchanging value of chromatic dispersion. Network operators would then deploy fixed dispersion compensators periodically along every fiber link to exactly match the fiber dispersion. Unfortunately, several vexing issues may necessitate that dispersion compensators are tunability, that they have the ability to adjust the amount of dispersion to match system requirements.

First, there is the most basic business issue of inventory management. Network operators typically do not know the exact length of a deployed fiber link nor its chromatic dispersion value. Moreover, fiber plants periodically undergo upgrades and maintenance, leaving new and nonexact lengths of fiber behind. Therefore, operators would need to keep in stock a large number of different compensator models, and even then the compensation would only be approximate. Second, we must consider the sheer difficulty of 40 Gbit/s signals. The tolerable threshold for accumulated dispersion for a 40 Gbit/s data channel is 16 times smaller than at 10 Gbit/s. If the compensation value does not exactly match the fiber to within a few percent of the required dispersion value, then the communication link will not work. Tunability is considered a key enabler for this bit rate (see Figures 12 and 13). Third, the accumulated dispersion changes slightly with temperature, which begins to be an issue for 40 Gbit/s systems and 10 Gbit/s ultra long-haul systems. In fiber, the zero-dispersion wavelength changes with temperature at a

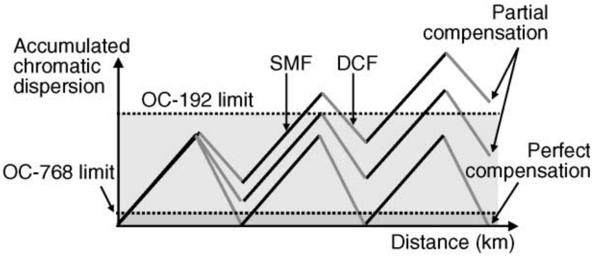
Accumulated chromatic dispersion
Partial compensation
SMF
DCF
OC-192 limit
OC-768 limit
Perfect compensation
Distance (km)

Figure 12 The need for tunability. The tolerance of OC-768 systems to chromatic dispersion is 16 times lower than that of OC-192 systems. Approximate compensation by fixed in-line dispersion compensators for a single channel may lead to rapid accumulation of unacceptable levels of residual chromatic dispersion.

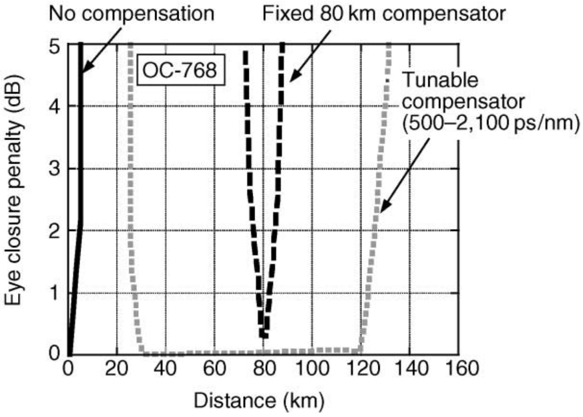

Figure 13 Tunable dispersion compensation at OC-768 (40 Gb/s) is essential for achieving a comfortable range of acceptable transmission distances (80 km for tunable, only ~4 km for fixed compensation).

typical rate of 0.03 nm/°C. It can been shown that a not-uncommon 50 °C variation along a 1,000 km 40 Gbit/s link can produce significant degradation (see Figure 14). Fourth, we are experiencing the dawn of reconfigurable optical networking. In such systems, the network path, and therefore the accumulated fiber dispersion, can change. It is important to note that even if the fiber spans are compensated span-by-span, the pervasive use of compensation at the transmitter and receiver suggests that optimization and tunability based on path will still be needed.

Other issues that increase the need for tunability include: (i) laser and (de)mux wavelength drifts for which a data channel no longer resides on the flat-top portion of a filter, thereby producing a chirp on the signal that interacts with the fiber's chromatic dispersion; (ii) changes in signal power that change both the link's nonlinearity and the optimal system dispersion map; and (iii) small differences that exist in transmitter-induced signal chirp.

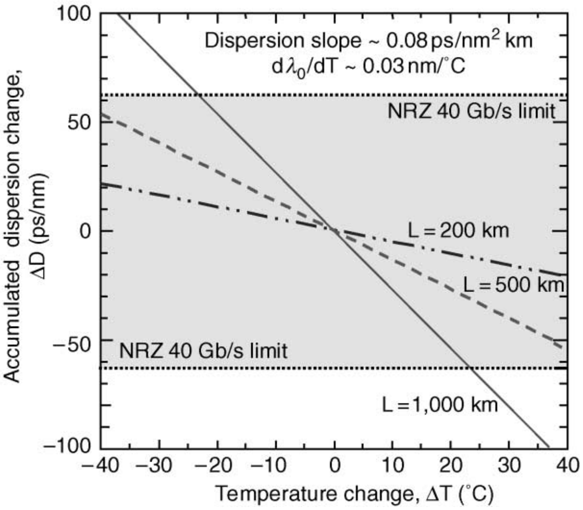

Figure 14 Accumulated dispersion changes as a function of the link length and temperature fluctuation along the fiber link.

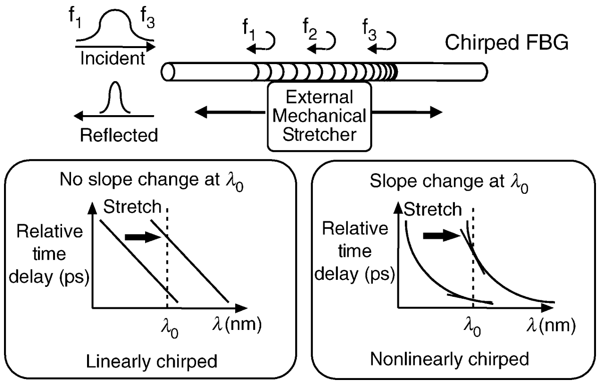
f1
f3
Incident
f1
f2
f3
Chirped FBG
Reflected
External Mechanical Stretcher
No slope change at λ0
Stretch
Relative time delay (ps)
λ0
λ(nm)
Linearly chirped
Slope change at λ0
Stretch
Relative time delay (ps)
λ0
λ(nm)
Nonlinearly chirped

Figure 15 Tuning results for both linearly and nonlinearly chirped FBGs using uniform stretching elements. The slope of the dispersion curve at a given wavelength λ_0 is constant when the linearly chirped grating is stretched, but changes as the nonlinearly chirped grating is stretched.

Approaches to tunable dispersion compensation

A host of techniques for tunable dispersion compensation have been proposed in recent years. Some of these ideas are just interesting research ideas, but several have strong potential to become viable technologies.

Fiber gratings offer the inherent advantages of fiber compatibility, low loss, and low cost. If a FBG has a refractive-index periodicity that varies nonlinearly along the length of the fiber, it will produce a time delay that also varies nonlinearly with wavelength (see Figure 15). Herein lies the key to tunability. When a linearly chirped grating is stretched uniformly by a single mechanical element, the time delay curve is shifted towards longer wavelengths, but the slope of the ps-vs.-nm curve remains constant at all wavelengths within the passband. When a nonlinearly-chirped grating is stretched, the time delay curve is shifted toward longer wavelengths, but the slope of the ps-vs.-nm curve at a specific channel wavelength changes continuously. Ultimately, tunable dispersion compensators should accommodate multichannel operation. Several WDM channels can be accommodated by a single chirped FBG in one of two ways: fabricating a much longer (i.e., meters-length) grating, or using a sampling function when writing the grating, thereby creating many replicas of transfer function of the FBG in the wavelength domain (see Figure 16).

One free-space-based tunable dispersion compensation device is the virtually imaged phased array (VIPA), based on the dispersion of a Fabry–Perot interferometer. The design requires several lenses, a movable mirror (for tunability), and a glass plate with a thin film layer of tapered reflectivity for good mode matching. Light incident on the glass plate undergoes several reflections inside the plate. As a result, the beam is imaged at several virtual locations, with a spatial distribution that is wavelength-dependent.

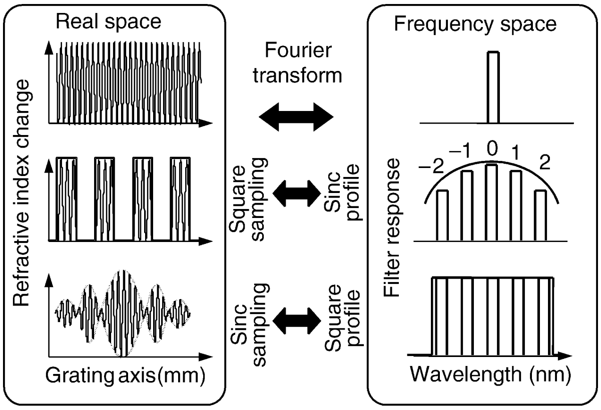

Figure 16 The concept of ‘sampled’ FBGs, where a superstructure is written on top of the grating that produces a Fourier transform in the frequency domain, leading to multiple grating passbands.

Several devices used for dispersion compensation can be integrated on a chip, using either an optical chip media (semiconductor-based laser or amplifier medium) or an electronic chip. One such technology is the micro-ring resonator, a device that, when used in a structure similar to that of an all-pass filter (see Figure 17), can be used for dispersion compensation on a chip-scale. Although these technologies are not yet ready for deployment as dispersion compensators, they have been used in other applications and have the potential to offer very high performance at low cost.

As the ultimate optical dispersion compensation devices, photonic bandgap fibers (holey fibers) are an interesting class in themselves (see Figure 18). These are fibers with a hollow structure, with holes engineered to achieve a particular functionality. Instead of being drawn from a solid preform, holey fibers are drawn from a group of capillary tubes fused together. This way, the dispersion, dispersion slope, the nonlinear coefficients could in principle all be precisely designed and controlled, up to very small or very large values, well outside the range of those of the solid fiber.

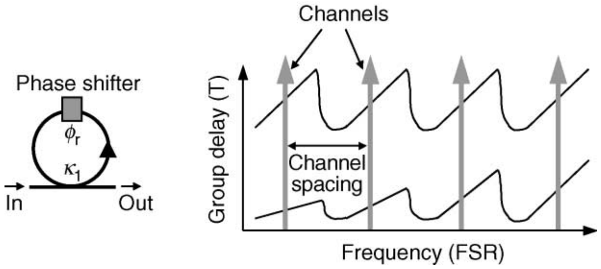

Figure 17 Architecture of an all-pass filter structure for chromatic dispersion and slope compensation. (Reproduced with permission from Madsen CK, Lenz G, Bruce AJ, *et al.* (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(12): 1623–1625.)

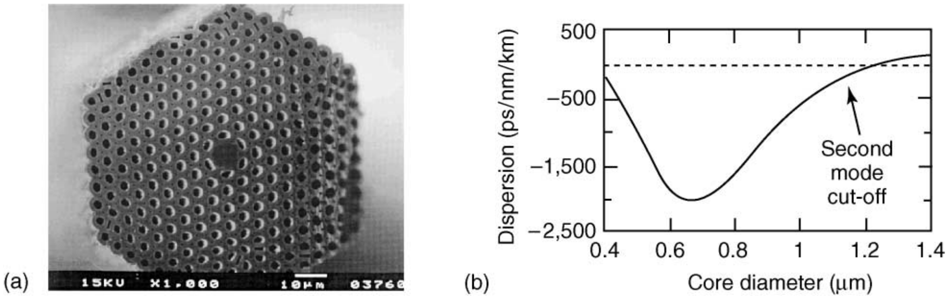

Figure 18 (a) SEM image of a photonic crystal fiber (holey fiber), and (b) net dispersion of the fiber at 1550 nm as a function of the core diameter. (Reproduced with permission from Birk TA, Mogihetsev D, Knight JC and Russell PSt.J (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(6): 674–676.)

Dispersion Slope Mismatch

Transmission fiber, especially fiber with dispersion compensation built in, may suffer from a dispersion slope in which a slightly different dispersion value is produced for each WDM channel (see Figure 19). Even though the compensator would be able to cancel the dispersion of the fiber at the design wavelength, there will be residual dispersion left at the other wavelength channels unless the compensator can match the slope of the dispersion curve of the fiber as well. Some solutions for dispersion slope compensation are described in this section.

First, DCF, with negative dispersion slope, is a prime candidate for deployment as a dispersion slope compensator even though it cannot easily be made tunable. By designing the DCF with the same ratio of dispersion to dispersion slope as that of a real fiber link, new types of DCF can be used to compensate for both dispersion and dispersion slope, much like DCF is used for dispersion compensation today. DCF's popularity, wideband functionality, and stable dispersion characteristics make this a particularly attractive solution. Designing the DCF to match the dispersion characteristics of the transmission fiber is the critical engineering challenge in this slope compensation scheme. Second, third-order nonlinearly chirped FBG can act as a tunable dispersion slope compensator. A simple modification of the nonlinearly chirped FBG allows tuning of the compensated dispersion slope value via stretching the grating. The grating is prepared such that the time delay as a function of wavelength has a cubic profile that covers several WDM channels over a continuous bandwidth of many nanometers. Since the resulting dispersion curve is quadratic over the grating bandwidth, the dispersion slope experienced by the WDM channels can be tuned by stretching the grating using a single mechanical element (see Figure 20). Third, combining the VIPA, with either a 3D mirror or diffraction grating, can also provide tunable free-space dispersion slope compensation. Slope tuning is achieved by dynamically controlling the MEMS-based 3D mirror or the diffraction grating. Fourthly, using an FBG with many spaced thin-film heater sections can also enable tunable dispersion slope. Each heater can be individually electrically controlled, allowing the time delay profile of the grating to be dynamically tuned via changing the temperature along the length of the grating. Advanced applications of this technique can allow alterations to the entire time delay profile of the grating, providing a truly flexible dispersion and dispersion slope compensation mechanism.

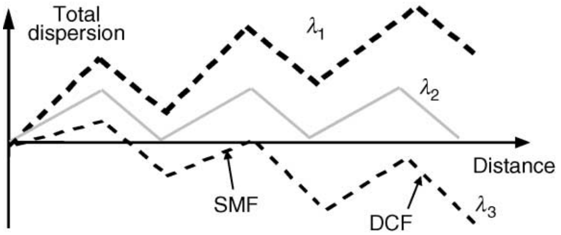

Figure 19 The chromatic dispersion slope mismatch caused by the different slopes of transmission fiber (SMF or NZDSF) and DCF.

Chromatic Dispersion Monitoring

Another important issue related to dispersion management is dispersion monitoring techniques. In a reconfigurable system, it is necessary to reconfigure

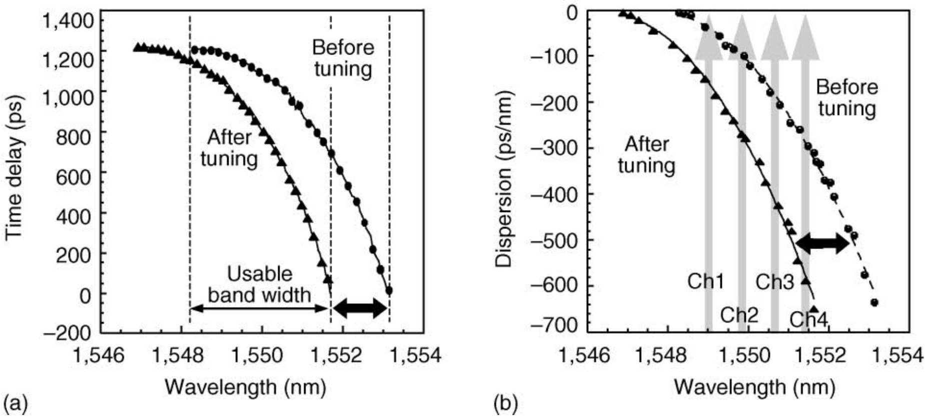

Figure 20 Tunable dispersion slope compensation using a third-order nonlinearly chirped FBG. (a) Cubic time delay curves of the grating, and (b) quadratic dispersion curves showing the change in dispersion in each channel before and after tuning. (Reproduced with permission from Song YW, Motoghian SMR, Starodubov D, *et al.* (2002) Tunable dispersion slope compensation for WDM systems using a non-channelized third-order-chirped FBG. *Optical Fiber Communication Conference* 2002. Paper ThAA4.)

any tunable chromatic dispersion compensation modules on the fly as the network changes. An in-line chromatic dispersion monitor can quickly measure the required dispersion compensation value while data are still being transmitted through the optical link. This is very different from the more traditional chromatic dispersion measurement techniques where dark fiber is used and the measurement is done off-line over many hours (or days).

Chromatic dispersion monitoring is most often done at the receiving end, where the Q-factor of the received data or some other means is employed to assess the accumulated dispersion. Existing techniques that can monitor dispersion in-line, fast, and with relatively low cost include: (i) general performance monitoring using the bit-error rate (BER) or eye opening, but this approach cannot differentiate among different degrading effects; (ii) detecting the intensity modulation induced by phase modulation at the transmitter; (iii) extracting the bit-rate frequency component (clock) from photo-detected data and monitoring its RF power (see Figure 21); (iv) inserting a subcarrier at the transmitter and subsequently monitoring the subcarrier power degradation; (v) extracting several frequency components from the optical data using very narrow-band optical filters and detecting the optical phase; (vi) dithering the optical-carrier frequency at the transmitter and measuring the resultant phase modulation of the clock extracted at receiver with an additional phase-locked loop; and (vii) using an optical filter to select the upper and lower vestigial sideband (VSB) signals in transmitted optical data and determine the relative group delay caused by dispersion (see Figure 22).

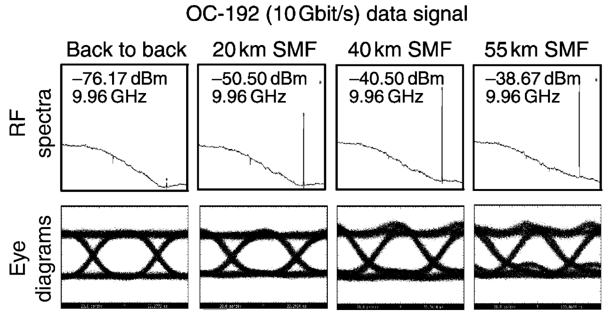

Figure 21 Clock regenerating effect due to chromatic dispersion for NRZ data. As the amount of residual dispersion increases, so does the amount of power at the clock frequency. This power can be used to monitor the amount of uncompensated chromatic dispersion. (Reproduced with permission from Pan Z, Yu Q, Xie Y, *et al.* (2001) Chromatic dispersion monitoring and automated compensation for NRZ and RZ data using clock regeneration and fading without adding signalling. *Optical Fiber Communication Conference 2001.* Paper WH5.)

Conclusion

Chromatic dispersion is a phenomenon with profound implications for optical fiber communications systems. It has negative effects, broadening data pulses, but it also helps reduce the effects of fiber nonlinearities. For this reason, managing dispersion, rather than trying to eliminate it altogether, is the key. Fixed dispersion components are suitable for point-to-point OC-192 systems. Tunable dispersion components are essential for dispersion management in reconfigurable and OC-768 systems. These dispersion compensation elements must have low loss, low nonlinearity, and must be cost effective.

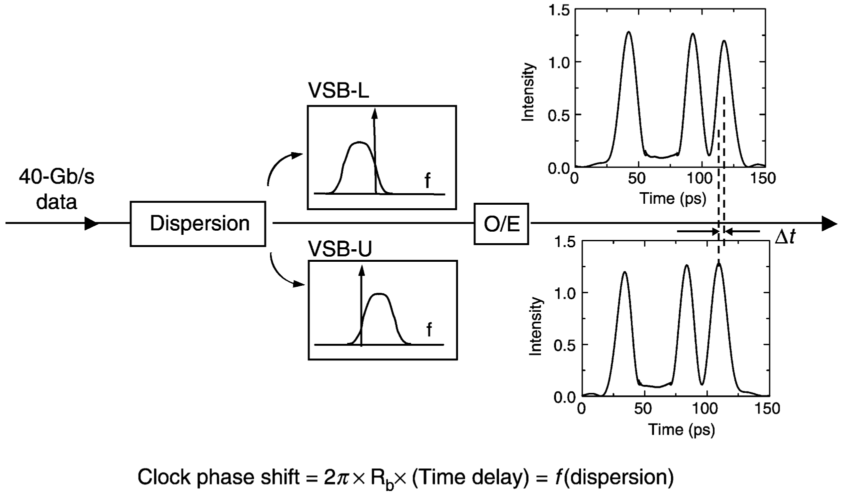

Figure 22 Chromatic dispersion monitoring using the time delay (Δt) between two VSB signals, which is a function of chromatic dispersion. (Reproduced with permission from Yu Q, Yan L-S, Pan Z and Willner AE (2002) Chromatic dispersion monitor for WDM systems using vestigial-sideband optical filtering. *Optical Fiber Communication Conference 2002.* Paper WE3.)

Although several technologies have emerged that meet some or all of the above requirements, no technology is a clear winner. The trend is towards tunable devices, or even actively self-tunable compensators, and such devices will allow system designers to cope with the shrinking system margins and with the emerging rapidly reconfigurable optical networks.

List of Units and Nomenclature

Chromatic dispersion	The time-domain pulse broadening in an optical fiber caused by the frequency dependence of the refractive index. This results in photons at different frequencies traveling at different speeds [ps/nm/km].
Chromatic dispersion slope	Different wavelengths have different amounts of dispersion values in an optical fiber [ps/nm^2/km].
Conventional single mode fiber (SMF)	Fiber that transmits only a single optical mode by virtue of a very small core diameter relative to the cladding. Provides lowest loss in the 1,550 nm region and has zero dispersion at 1,300 nm.
Cross-phase modulation (XPM)	A nonlinear Kerr effect in which a signal undergoes a nonlinear phase shift induced by a copropagating signal at a different wavelength in a WDM system.
Dispersion compensating fiber (DCF)	Optical fiber that has both a large negative dispersion and dispersion slope around 1,550 nm.
Fiber Bragg gratings (FBGs)	A small section of optical fiber in which there is a periodic change of refractive index along the core of the fiber. An FBG acts as a wavelength-selective mirror, reflecting only specific wavelengths (Bragg wavelengths) and passing others.
Four-wave mixing (FWM)	A nonlinear Kerr effect in which two or more signal wavelengths interact to generate a new wavelength.
Kerr effect	Nonlinear effect in which the refractive index of optical fiber varies as a function of the intensity of light within the fiber core.
Modulation	The process of encoding digital data onto an optical signal so it can be transmitted through an optical network.
Nonlinear effects	Describes the nonlinear responses of a dielectric to intense electromagnetic fields. One such effect is the intensity dependence of the refractive index of an optical fiber (Kerr effect).
Nonzero dispersion-shifted fiber (NZDSF)	Optical fiber that has a small amount of dispersion in the 1,550 nm region in order to reduce deleterious nonlinear effects.
Polarization mode dispersion (PMD)	Dispersion resulting from the fact that different light polarizations within the fiber core will travel at different speeds through the optical fiber [ps/km$^{0.5}$].
Self-phase modulation (SPM)	A nonlinear Kerr effect in which a signal undergoes a self-induced phase shift during propagation in optical fiber.
Wavelength division multiplexing (WDM)	Transmitting many different wavelengths down the same optical fiber at the same time in order to increase the amount of information that can be carried.

See also

Nonlinear Optics, Basics: Four-Wave Mixing. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. Rochester, NY: John Wiley & Sons.

Agrawal GP (1997) *Nonlinear Fiber Optics*, 2nd edn. Academic Press.

Gowar J (1993) *Optical Communication Systems*, 2nd edn. Prentice Hall.

Kaminow IP and Koch TL (1997) *Optical Fiber Telecommunications IIIA & IIIB*. Academic Press.

Kashyap R (1999) *Fiber Bragg Gratings*. Academic Press.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Artech House.

Willner AE and Hoanca B (2002) Fixed and tunable management of fiber chromatic dispersion. In: Kaminow IP and Li T (eds) *Optical Fiber Telecommunications IV*. Academic Press.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. John Wiley & Sons.

DISPLAYS

R L Donofrio, Display Device Consultants LLC, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article is a brief review of display technology being used today. It emphasizes some of the optical aspects of displays. The details of the basic construction of the display devices are also discussed to aid in understanding. The displays covered are the CRT, VFD, FED, PDP, LED, OLED, LCD-AMLCD, transmissive and transflective LCDs.

Definition of a Display

A display can be defined as an object which transfers information to the viewer and/or an output unit that gives a visual representation of data.

Optical Characteristics of a Display

A display has (in general) the optical characteristics of luminance, reflectivity, contrast ratio, color, surface gloss, resolution, and image sharpness (sometimes called small area contrast). Measurement methods can be obtained from the EIA (Electronics Industries Alliance), JT-31 (Optical Characteristics of Display Devices), SAE (Society of Automotive Engineers and ISO Standards) and VESA Standards, to name a few.

Luminance

The luminance of a display is its light output and is given in units of candela/square meter or cd/m^2 (in the English system fL). Luminance is measured using a photometer, photometer/colorimeter, or spectroradiometer.

Illuminance

The illuminance is the light incident on the surface being measured. Its units are Lumen per m^2 or Lux (and in the English system Ft.cd). This is measured using illuminance meters.

Reflectivity

The reflectivity for CRTs (cathode ray tubes) is measured with incident light at 45 degrees (in some cases for LCDs, a smaller 30 degree angle is used). The reflected light from the surface is measured with a detector normal to (perpendicular to) the surface. Devices such as 'tube face reflectivity' (TFR) can be used to measure this parameter and it is published in the EIA-JT-31 measurement methods. Other similar methods are used for flat panel displays.

Color

Although color perception is different for different individuals, standard responses to color stimuli have been developed and we can measure the color of different objects and display images. Color can be measured with such equipment as a colorimeter or spectroradiometer. The output data are given in various color systems. Color systems, such as the 1931 CIE (Commission International de I'Eclairage) Color System (x, y), the 1960 CIE (u, v) or 1976 CIE (u', v') color coordinates are all used. All these color systems can be converted to each other by simple mathematical relations and these conversions are usually available to the user in the data output. The choice of the color space/system is dependent on application but the system in most general use for displays is the 1976 CIE (u', v'). This color system has become more widely used because equally perceived color differences in this color system are at almost equal distances in this color space. In reality, the quest for a system with equal distances equating to equal color differences can be achieved through a complicated matrix approach in which the constants used in the matrix vary from point to point in the color space.

Gloss

Gloss is a measure of surface reflectivity under Snell's Law criteria (equal incidence and reflection angles). The measurement angles of 20, 45, 60, and 85 degrees are used to measure the gloss of common surfaces. Display glass surface measurements tend to use the 60 degree gloss angle. The surface of a display can be made rougher to scatter incident light away from the viewer or it can be optically coated to reduce the reflected light to achieve improved viewing. Some times both methods are used. Reduced gloss surfaces also are involved with the reduction in surface glare. The level of increased roughness of the surface affects the user's ability to see detail.

Contrast Ratio

When people talk about contrast they generally mean 'contrast ratio'. The contrast ratio (CR) of an area on a display is a unitless parameter and is defined as the excited luminance (L_e) divided by the unexcited luminance (L_u). The excited luminance (L_e) equals the glass transmission (T) times some excited internal luminance (L_{ei}) and for simplicity, L_u equals T times an internal zero ambient unexcited luminance light level of (L_{ui}). Factors, such as internal scattered light in LCDs and scattered electrons and scattered light in CRTs, give a nonzero un-excited luminance L_u.

When dealing with the CR, the effect of ambient illumination A contributes to the numerator and the denominator. Thus, when the ambient illumination is increased, it lowers the contrast ratio. In a simple case where the ambient illumination is reflected from the surface of the display whose reflectivity is ρ, the contrast ratio is given below:

$$\mathrm{CR} = (L_e + A\rho)/(L_u + A\rho) \qquad [1]$$

For the case of the reflectivity of a phosphor screen (reflectance R) coated panel with a glass transmission (T), the reflectivity to first approximation is

$$\rho = cRT^2 \qquad [2]$$

where c is a constant.

So we see that changes in contrast can come about through changing the display's glass panel transmission. For example, if the L_e of the white field of a CRT was 100 fL with a 90% panel glass (along with $R = 0.5$) and had a L_u of 2 fL, then the zero ambient contrast would be 50. However, with an ambient illumination of 50 Ftcd the contrast would be 4.76. Now, when we use a 50% transmission panel glass at the ambient illumination of 50 Ft.cd the CR is increased to approximately 8.39. Equations [1] and [2] show how to calculate the contrast. It should be noted that decreased panel glass transmission or the use of 'dark glass' increases contrast because the reflectivity is a 'T squared' effect and the luminance L is a function 'T'. There are second-order effects involved with the reduction of reflected light and unexcited luminance which make the equations more complicated. These second-order effects will not be covered as they are beyond the basic discussion of this article.

In comparing display CR, one must compare the displays at the same ambient illumination. Additionally, you will find that a number of flat panel display advertising reports will report the CR at zero ambient illumination. As we have shown, these CR values are the highest ones for the display. The CR for emissive displays tends to be constant over a wide range of viewing angles, whereas the contrast ratio for liquid crystal displays (LCDs) is more variable over large-angle viewing. Recent new electrode position developments, such as the 'in-plane switching' technology for active matrix liquid crystal displays (AMLCDs) has gone a long way to reduce this reduced angle-of-view effect in that type of LCD display.

Figure 1 shows how a viewer in a sport utility vehicle (SUV) would see the changes in contrast with viewing angle for two types of twisted nematic LCDs (TN-LCDs). One of the TN LCDs does not

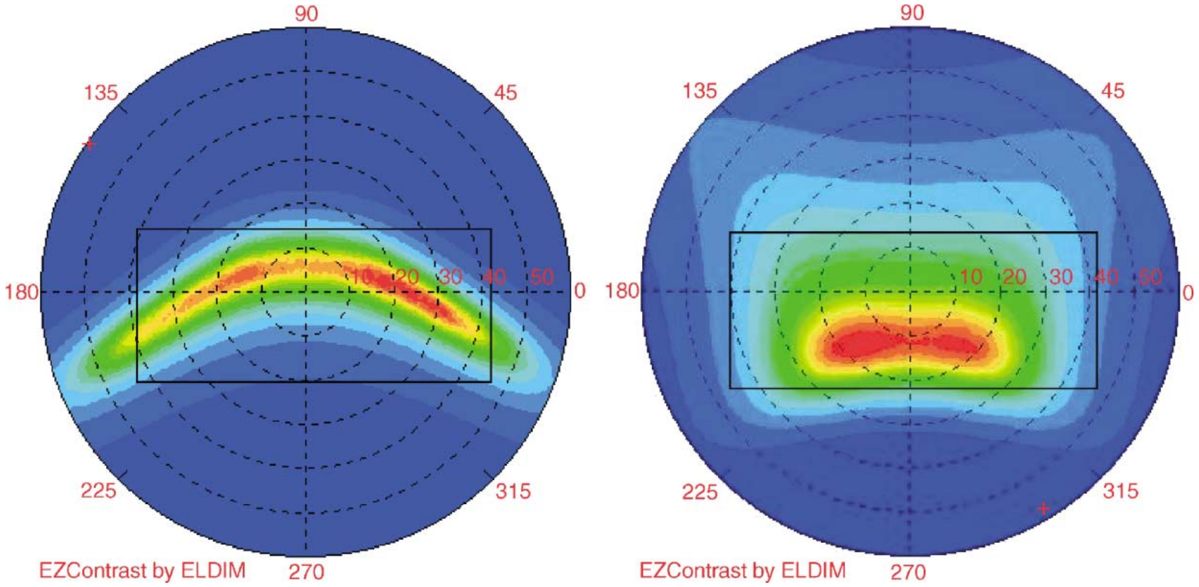

Figure 1 A typical rectangular viewing envelope for a mid-size SUV vehicle superimposed on LCD iso-contrast plots. Left is standard TN LCD. Right, is film-compensated TN LCD. Reproduced from Smith-Gillespie R, *et al.* (2001) Design requirements for automotive entertainment displays. *Proceedings of the 8th Annual Symposium on Vehicle Displays (2001).* Permission for reprint courtesy of the Society for Information Displays.

have the film compensation and the other does. For relative contrast appraisal, the red areas show the highest contrast regions and the blue the lowest. It is obvious that different technologies can give significant changes in viewed contrast.

Phosphor Aging

The fall off of luminance as the phosphor screen is bombarded with the electron beam is called phosphor aging. This effect is due to two main factors: (i) the thermal quenching of the phosphor luminescence (which is reversible); and (ii) a nonreversible aging due to damage to the screen components (such as phosphor, binder, and aluminum coating). It has been found that the aging (when we are using an electron beam) is a function of the number of electrons bombarding the screen. Aging models help us to understand and predict the amount of aging a tube will show after a fixed amount of time. The first aging model in 1951 was on ion bombardment and showed that:

$$I = I_0/(1 + AN) \qquad [3]$$

where I_0 is the initial luminance before aging, and I is the final luminance.

In 1961 a similar (one parameter) Model $I = I_0/(1 + CN)$ was used by Pfhanl, where C is the aging constant to half luminance and N is the number of coulombs. In 1981, a two-component model for low current aging was presented, where:

$$I = I_0/(1 + N(B/Q^D)) \qquad [4]$$

Here we have two constants B and D, and Q is the cumulated charge in number of coulombs. When $D = 0$ then the equation given above goes to the Pfhanl case.

Later, in 1996 the three-parameter model was described and in 1997 the Gaussian beam aging model was described.

Ambient Illumination and Reflection

To see the information shown on a display, the information must be spatially detected by the human eye over the reflected ambient illumination (such as room and sunlight). The ambient illumination is generally from localized or diffuse sources. As we have seen previously, the effect of diffuse illumination can be reduced through the use of neutral density filters. But, when we are dealing with very high ambient light such as that entering the canopy of a fighter aircraft then, the monochromatic information from the display can be viewed using selective spectral filters. However, for the usual computer monitor, the ‘specular reflection’ (Snell’s Law type of reflection) off the front of the display is due to a localized external light source. This reflection is reduced by methods such as anti-glare (AG) and anti-reflecting (AR) surfaces applied to the front surface of the display. A flat versus curved front surface of the display can also be a glare reduction technique. Surface roughness is used in AG surfaces to scatter a portion of the specular light away from the viewer. Surface roughness can be achieved through surface abrading, surface plastic coating, surface silicate coating, and surface acid etching to name a few methods. AR surfaces use optical interference techniques to reduce the reflected light to the viewer. In some cases, both methods are applied to a display surface. Surface gloss of the display is usually measured with a gloss meter in 60 degree gloss values but 20, 45, and 80 degree angles are also used. Resolution is measured in line pairs/mm or pixels/mm and image sharpness is associated with modulation transfer function (MTF) measurements. We should note here that resolution and image sharpness are not the same thing. High resolution is equated with the high frequency limit of the MTF measurement. Since the reduction in specular reflection through the use of surface roughness can affect the image sharpness, we see that a system of trade-offs can be developed to optimize a display for a given viewing environment. AR coatings improve glare reduction without scattering and thus give higher-resolution capability to the displays but this is usually obtained at a higher price then the AG coatings.

Types of Displays

Displays can be divided into two types; emissive and nonemissive. By emissive we mean that the display surface gives off light. In the nonemissive, the display or a small portion of the display acts as a shutter or spatial filter which also can be reflective in nature. In this nonemissive mode, the display can pass light from a source in back of or in front of the shutter (or both, as we shall see later).

Table of Emissive and Nonemissive Displays

Thus from Table 1 we see that under the title of emissive displays we have the cathode ray tube (CRT), projection CRT, vacuum fluorescent display (VFD), field emissive display (FED), electroluminescent (EL) display, light emitting diode (LED), organic light emitting diode (OLED), and plasma. These devices, which are in the realm of emissive displays, use certain types of emissive materials called

Table 1 Table of emissive and nonemissive displays

Emissive	*Nonemissive*
CRT	LCD and AMLCD
VFD	Bi-Stable LCD
FED	Digital Mirror D
EL – inorganic	LCOS (Liquid Crystal on Si)
LED	Electronic Inks
OLED	
PLASMA	

lumophors (or phosphors). The lumophors are called cathodoluminescent, photoluminescent, and electroluminescent, depending whether the excitation action was due to the interaction of the phosphor with cathode rays, photons, or low-voltage electrons. Cathodoluminescent lumophors are used in cathode ray tubes (CRTs – television tubes, monitor tubes, projection CRTs), (VFDs), and (FEDs).

EL displays can use inorganic lumophors (phosphors) or organic lumophors such as those used in OLEDs. Under photoluminescence displays, we have 'plasma displays' (plasma TV), IR-up conversion, and fluorescent lamps. In plasma displays, the excitation of the plasma in a localized region gives rise to UV photons which strike the phosphor screen in that region. In IR-up conversion, IR photons are absorbed in such a way that two IR photons can excite an electron with twice the energy of one IR photon. This then allows the use of IR lasers to create displays through the excitation of these up-conversion phosphors. In fluorescent lamps, the UV created by the excitation of Hg or Xe, strikes the special lamp phosphors (coated on the inside of the glass envelope), which become luminescent.

Under the heading of shutters we can consider LCDs which consist of AMLCDs, super twisted nematic liquid crystal displays (STN LCDs) and bistable LCDs. Another shutter type is the liquid crystal on silicon (LCOS) display and the movable digital mirror display (DMD), and the digital light processing (DLP) projecting engines. Another spatial display type is both normal and dynamic inks. A dynamic ink can be a 'polarizable' ink display in which a two shade (light/dark) ink sphere element can rotate under the presence of an electric field to show the light or dark side. Although they are very interesting, we will not cover ink displays in this article.

The shutter action is also used in some projection TV sets (or projectors) in which the LCD either transmits or reflects (such as LCOS) the light from an arc lamp. The DMD movable mirrors reflect the digital information to the lens system and the unwanted rays are reflected away from the lens to a black light capture area.

The Cathode Ray Tube (CRT)

The CRT is a display which has been in use for over 100 years (the Braun CRT tube was developed in 1898). The CRT uses a phosphor screen which is bombarded by energetic electrons and emits visible light radiation. The single electron gun CRT was used for years in early monochrome television sets and monochrome computer displays. Monochrome green, amber, and white phosphors have been used in monitor tubes. Some color selectivity can be obtained with the single electron gun using the following three methods; (i) voltage penetration phosphors, (ii) current sensitive phosphors and (iii) beam index methods. In a CRT, the penetration of the electron beam follows Hans Bethe's energy loss equation for electron penetration in solids and thus is a function of the electron energy or CRT voltage.

Hans Bethe's energy loss equation is:

$$dE/dx = -2\Pi e^4 (N_e/E)\ln(\alpha E/I) \qquad [5]$$

where: E = energy of the incident electron; x = depth of material; N_e = density of atomic electrons; I = mean excitation energy; and α = a constant.

In the voltage penetration tube, a cascaded screen is used to give a color change when the electron voltage is increased. For example, if a cascade screen is made from a green phosphor layer placed on top of the red phosphor layer then the electrons strike the green phosphor first. Then with increased voltage, the electrons pass through the top phosphor and excite the bottom red phosphor. In the current sensitive color tube, the phosphor screen is composed of (for example) red and green phosphor whose luminance versus current curves are different. In these cases, we are referring to the linearity of the change in luminescence from the phosphor as a result of a change in electron beam current. The red phosphor (generally a rare earth oxide) is a linear phosphor, whereas the green is nonlinear. Changing the current thus can change the color of the display. In the beam index tube, a UV (or secondary electron emitters have also been used) emitting index layer is put along side a red, green, or blue triad (for example, between the blue and the red phosphor stripes). The scanning electron beam, through the use of a UV sensor (or secondary electron detector) mounted in the rear portion of the tube and the scanning electronics, can determine the location of the red, green, and blue phosphors. The electron beam can then excite only the individual phosphor areas of interest. With this tube, generally, the electron beam is at a low current to give a small enough cross-section to only excite one color and thus has limited application.

Another application of the monochrome is the projection CRT system. Here, three tubes for red, green, and blue are used to project an image on a viewing screen. The terminology is such that, when the CRTs are in the back of the screen and the viewer in front of the screen, it is called a 'rear' projection set and if the CRTs and the viewer are in front of the screen it is a 'front' projection set. The principle here is that, either using one CRT tube at a time or by a superposition two or three of the CRTs, the color of the displayed image will be within the color gamut defined by the primary colors.

The Color CRT

The more ubiquitous method of obtaining a color display is the device called the shadow mask CRT (and the similar Trinitron CRT). This color CRT is designed in such a manner that through the use of an aperture mask, (seen in Figure 1) the electrons from the red electron gun strike the red phosphor and likewise electrons from the green and blue electron guns strike the green and blue phosphors. The combination allows the viewer to see a color image. This color separation action is seen in Figure 2. Although Figure 2 shows a delta tri-color design, there are also in-line systems in which the mask consists of an array of slots which are somewhat rectangular in nature. In this array, many slots can be on a line, or a mask made can be made of a series of wires as used in the Trinitron design.

The disadvantage of the shadow mask technique is that the mask can capture as much as 80% of the electrons from the electron guns. This yields lower screen luminance and lower contrast due to scattered electrons. As the screen size gets larger from 19 inches to 36 inches or more, additional temperature compensation methods must be used to reduce the effects of mask heating. A number of methods to control the shift of the electron beam on the phosphor as the mask heats up have been developed. Initially, axially positioned bimetal springs were used on masks to compensate for the heating effect. Then lighter masks were used with corner lock suspension systems. Another method is using mask materials with lower coefficients of expansion such as INVAR.

Contrast in color CRTs has been improved by first applying a carbon coating between the panel glass and the phosphor screen. This requires additional process steps which form 'windows' in the carbon coating (matrix) through which the excited phosphor is seen. The use of pigmented phosphors also helped to reduce the reflectivity of the phosphor/matrix screen.

The phosphors used in the shadow mask color CRT are the red Y_2O_2S:Eu or, in some cases, the Y_2O_3:Eu phosphor. The green phosphor is ZnS:Cu and the blue phosphor is ZnS:Ag.

The CRT seen in Figure 3 shows the new high-definition television-type (HDTV) tube with the 16 × 9 aspect ratio as contrasted with the older 4 × 3 aspect ratio (screen width × height).

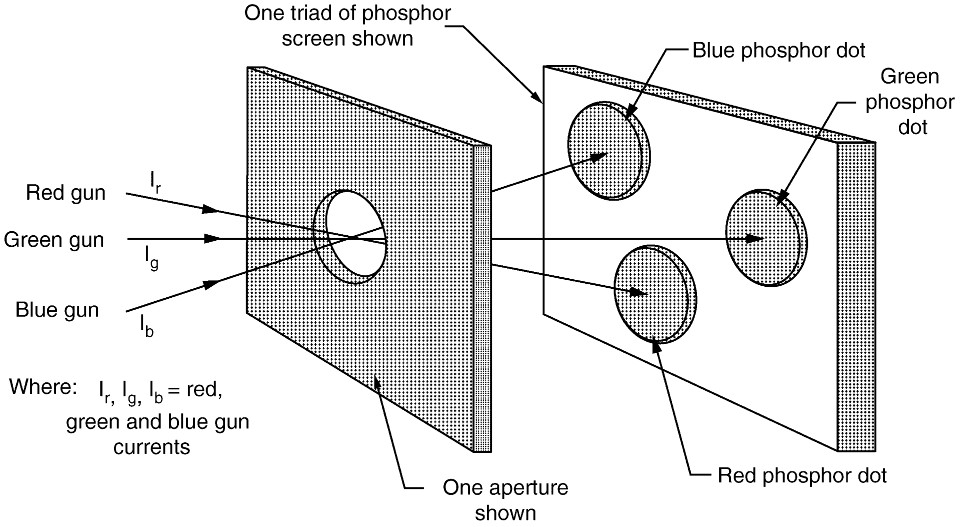

Figure 2 The action of a shadow mask.

Figure 3 A typical HDTV color CRT. Reproduced from Donofrio RL (2002) The future of direct-view CRTs. *SID Information Display Magazine* 6: 10. Permission for reprint courtesy of the Society for Information Display.

The Vacuum Fluorescent Display (VFD)

The VFD can be thought of as a low-voltage CRT. Just as in the CRT, electrons from a cathode are controlled by a grid which allows them to be accelerated and strike the phosphor screen anode. However, the electrons in a CRT can be accelerated with a voltage of 30 000 V, whereas the VFD accelerates the electrons with voltages of 12 to 40 V.

In the operation of the VFD, electrons are emitted by the filament. The filament is a very thin tungsten wire coated with a tri-carbonate oxide (barium, strontium, and calcium carbonates) which are the same constituents of the CRT oxide cathode. The filament is held between a filament support and an anchor which applies tension to the filament wire. A current is passed through the filament to achieve a temperature to ~600 °C. At that temperature, thermonic emission occurs. Between the filament and the phosphor elements is a stainless steel mesh grid. By the application of a positive or negative voltage on this grid, electrons can either be cut off or accelerated to the phosphor on the anode. The anode can be a conductor such as graphite, coated with a phosphor layer such as ZnO:Zn (zinc oxide). This is a low-voltage cathodoluminescent phosphor which emits a green color with an emission maxima of about 505 nm. Other phosphors are also used to obtain red, orange, and blue (Figure 4).

In addition to the filament, grid and anode, the VFD is a vacuum device (like the CRT) and contains a getter to assist achieving and maintaining a good vacuum and an internal transparent conductive coating to protect the display from external electrostatic fields. The VFD display is used frequently in the auto industry and aging of the phosphor is very small, achieving 100 000 hrs to half of the initial luminance.

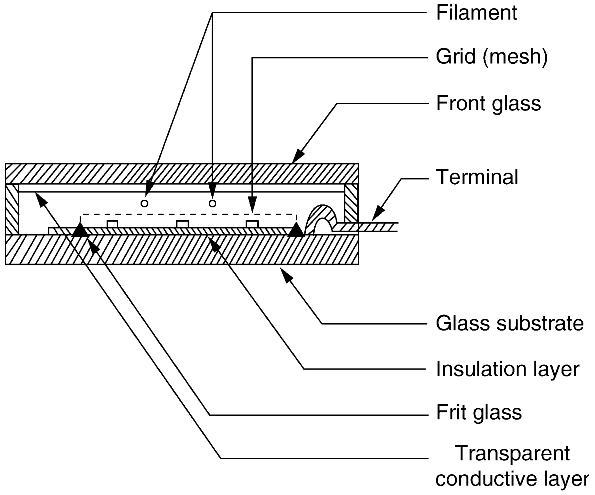

Figure 4 The cross-section of a vacuum fluorescent display assembly.

VFD units have been made with phosphor elements on either side of the filament. This gives rise to a 3D effect. Graphic and fixed segment VFDs have been made for multimedia products and 16 × 16 graphics for Kanji and other characters.

Experimentally, 2D VFD arrays have been made similar to the FEDs covered in the next section.

Field Emissive Displays (FEDs)

The FED is similar to the CRT and the VFD in that electrons from a source bombard a phosphor screen. In both the CRT and the VFD thermionic electrons are used. However, in the FED the electrons arise from field emission. Field emission devices create electrons though quantum mechanical tunneling from the emitter surface into the vacuum by the action of a large electrical field. This process is governed by the Fowler–Nordheim equation. This equation takes the form:

$$J = A(\beta V)^2/\Phi \exp(-B\Phi^{3/2}/\beta V) \qquad [6]$$

where, J is the electron current density; V is the applied voltage; Φ is the work function of the emitting tip; A & B are constants; and β is the field enhancement factor.

The emitters can be of the Molybdenum Spindt type and also the polydiamond or diamond-like carbon (DLC) coated emitters. We have the field enhancement with a sharp tip and reduced work function with the DLC. There has been renewed interest in FEDs, with the development of the 'carbon nanotube' as an electron field emitting source. The carbon nanotube is a fullerene material, which is a distinct form of carbon. This 'bucky ball'-like

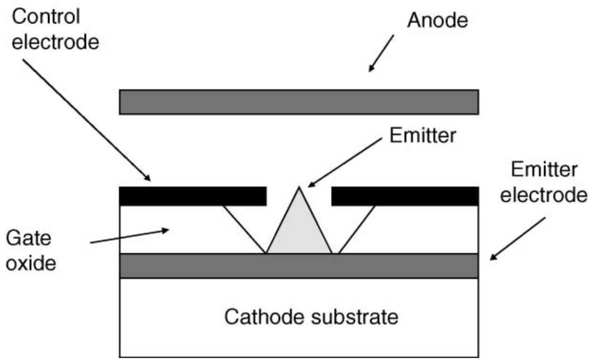

Figure 5 Field emission display cross-section. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

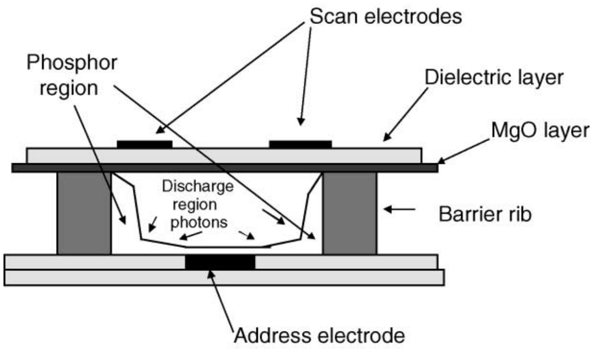

Figure 6 ACPDP partial cross-section view. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

material is cylindrical instead of spherical in nature, has great strength, and high conductivity.

There are two types of FEDs, a low-voltage (LVFED) and a high-voltage (HVFED). The HVFED uses an aluminum film over the phosphor to protect the phosphor screen from aging. The HVFED can use the high voltage (2–10 KV) electrons to bombard the aluminized screen and it can use the more efficient, somewhat standard CRT phosphors ZnS:Cu green, ZnS:Ag Blue, and Y_2O_3:Eu red. The LVFED (12–200 V) cannot use the standard CRT phosphors but uses the low-voltage phosphor such as the ZnO:Zn green phosphor. This does not use an aluminum layer, since about 2 KV is needed to penetrate the Al layer. The commonly used CRT red phosphor Y_2O_2S:Eu was found unsuitable for HVFED use due to release of the sulfur during electron bombardment.

Figure 5 shows that for this basic FED structure, the electrons from the Spindt emitter are controlled by the voltages of the anode, control electrode and emitter electrode.

Plasma Displays

Earlier plasma displays, made using excited Ne gas with the characteristic amber glow required no phosphor and were used in the printing industry. Present plasma displays used for television and information signs are photoluminescent displays. As shown in Figure 6, the excitation of the plasma gas made of elements such as He, Ne, Ar, and Xe give rise to a glow discharge and the creation of vacuum ultraviolet photons that strike the screen and cause it to luminesce. In construction, the display is similar to the FED in that it is a matrix addressed device. It has been described in literature that the similarity goes even further with the need for anode and cathode surfaces, a means of evacuating and a perimeter seal. There are two types of plasma displays, DC and AC. The AC design appears to be the more successful. Typical PDP phosphors for the triad are $(Y,Gd)BO_3$:Eu for the red, Zn_2SiO_4:Mn for the green, and $BaMgAl_{10}O_{17}$:Eu (called BAM for short) for the blue. Figure 6 below shows the basic structure of an AC plasma display. One phosphor element is shown with barrier ribs and address and scan electrodes. The discharge of the plasma takes place in the region of the phosphor and excites the photoluminescent phosphor, which is shaped (in this example) along the side of the barrier ribs and over the address electrode.

Electroluminescence

Electroluminescence was first seen in the 1930s by Destriau. EL is the generation of (nonthermal) light by use of an electric field. High electric field EL units emit light through the impact of high energy electrons in luminescent centers in a semi-conducting material. The electric field can be alternating current (AC) or direct current (DC). Powder phosphor AC EL displays appear to be one of the successful inorganic EL display methods. Because of their rugged nature, alternating current electroluminescence ACEL displays have been used in trucks, tractors and the M1 Abrams Tank. Thick-film EL panels are also used in some vehicle displays as back lights for LCDs. These light panels are free from mercury and are more environmentally friendly.

Inorganic LED Displays

Light-emitting diodes are being used as backlights for LCD displays, replacing some of the fluorescent light backlighting methods. They are also used in indicator lights. Large displays have been made of LEDs. Using a matrix array row and column connectors, we can digitally select a given color LED and pattern in an LED array and generate a color image. Similar methods were carried out many years ago using incandescent lights. However, now they are bright and fast reacting. LEDs operate by semiconductor P–N junction rules and the flow of electrons and holes under forward bias operation give rise to junction light emission. Materials such as GaP:Zn,O are used to create a red LED with an emission maxima wavelength of 699 nm.

Red, green, and blue-UV/blue LEDs are now available, with an increasing demand for white LEDs. Two methods are used to produce a white LED. One method is to use a blue LED with the front of the LED coated with a green photoluminescent phosphor. If the phosphor is thin enough, the blue light penetrates the screen and excites the green to give a blue/green white. Another method is to use a blue LED and coat it with a white photoluminescent phosphor.

Organic LEDs

The basic process of light generation in OLEDs is the formation of a P–N junction. Electrons injected from the N-type material and holes from the P-type material recombine at the junction and light is generated. There are presently two competing methods to fabricate an OLED. The first developed by Kodak in 1987 is the small molecule molecular weight material approach. In OLED operation, the electron hole pair recombine to form excitons (on the average one singlet and three triplet excitons are formed for every four electron hole pairs). The light for the Kodak OLED is from the recombination of singlet excitons and the energy transferred to the fluorescent dye. The terminology is such that we have small molecule-OLEDs (sm-OLEDs) and polymer-OLED (PLED). In later developments, the sm-OLED device structure was reported to have three layers. The first is the hole transport layer (hole transport) enhancing layer copper phthalocyanine (CuPc), a layer NPB (60 nm), the light emitting aluminum tris-8-hydroxyquinolate Alq layer (75 nm), and electron transport layer. The electron transport layer is in contact with the three layer structure. The top electrode is a low work function material, such as MgAg (75–200 nm thick) called an electron injection layer and topped with a transparent indium tin oxide (ITO) layer for transparent OLEDS. All of these coatings are applied in a vacuum chamber.

In the phosphorescent material method, the light emission is from the triplet excitons through the use of phosphorescent dyes and the polymer-based devices. The common base material is a polyphenylene-vinylene PPV and other polymers such as polyfluorene have been used. The construction of the phosphorescent OLED is similar in that the ITO is also used as a starting substrate but the hole transport layer is applied by spin casting from solution. Spin casting is also used for the electron transfer material. After drying (removal of solvents) a calcium cathode material is applied. A contrast of 3000/1 has been achieved (for a military type application) using a green OLED with an AR coated circular polarizer and at an ambient light level of 500 lux, and a 6/1 contrast ratio has been achieved with an ambient illumination of 50 000 lux and an output of 370 cd/m^2. Additionally, OLEDs can be made on flexible (F) substrates and are called FOLEDs. The use of a transparent (T) cathode gives rise to a top emission OLED or a TOLED.

The first OLED display used in auto products was developed by Pioneer and used in a radio display and the Ford Motor Company has shown a White OLED display in one of their ‘Concept Cars’. This display allows them to have a reconfigurable instrument

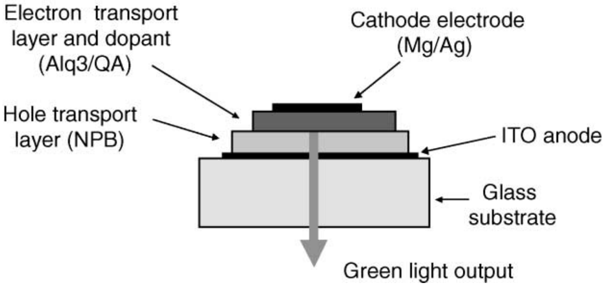
Electron transport layer and dopant (Alq3/QA)
Cathode electrode (Mg/Ag)
Hole transport layer (NPB)
ITO anode
Glass substrate
Green light output

Figure 7 OLED partial cross-section – Kodak type. Reproduced from Green P and Haven D (1998) Fundamentals of emissive displays. *SID Short Course S-3*. Permission for reprint courtesy of the Society for Information Display.

Figure 8 Philips poled display in an auto mirror. Reproduced with permission of Philips Mobile Display Systems.

panel (IP) and is made up of a number of 5 inch × 8 inch displays across the instrument panel (Figure 7).

The POLED as seen in Figure 8 is now being used by Philips Mobile Systems in a rear view car mirror, as seen below. It is the display from a video camera looking at the rear of a car to see any obstacles to backing up safely.

Thin Film EL (TFEL) Displays

A simple type of TFEL display is the color by white structure shown below. In this case, the color is generated using a white phosphor and red, green, and blue optical filters. To make this ACEL display function properly, insulators are used on either side of the phosphor. In Figure 9, the ITO conducting stripes (depicted) horizontally on the top and vertically on the bottom, give the means of addressing each color pixel. The white phosphor, for example, is a ZnS:Mn, SrS:Ce blend which gives a white electroluminescence.

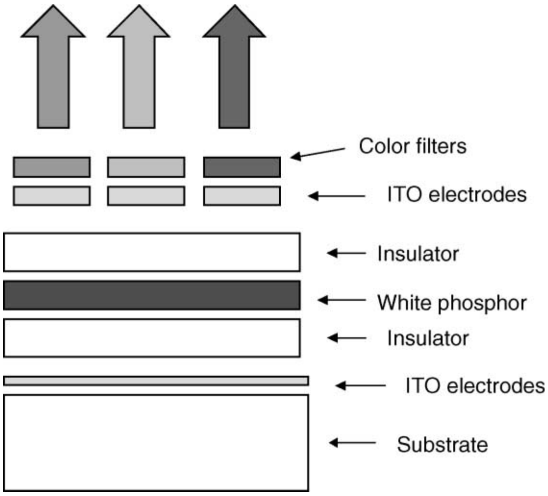

Figure 9 Color thin film electroluminescent display (TFEL) cross-section. Reproduced from Dakkala A (1998) New EL displays for vehicular applications. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98.* Permission for reprint courtesy of the Society for Information Display.

Other phosphors such as ZnS:Mn can be used but its yellow luminescence requires the use of a green and red filter and only two metal electrodes on the bottom to give a two-color display. Better red green color separation can be achieved through the use of a $\mathrm{Zn}_{1-x}\mathrm{Mg}_x\mathrm{S}$:Mn and ZnS:Mn phosphor mix. New EL blue phosphor developments have allowed larger color gamuts to be achieved.

One of the important characteristics of the alternating current thin-film electroluminescence (ACTFEL) is that it has an operating temperature range from −40 °C to 85 °C which is key for automotive use. It has a viewing angle of 160 degrees with a life of more than 50 000 hours. Active matrix (active addressed) AMEL structures allow fast

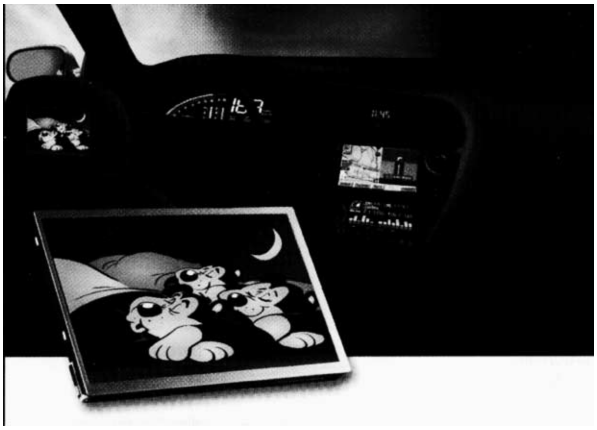

Figure 10 AMLCDs used in automobiles. Reproduced from Donofrio RL (2003) The road show. *SID Information Display Magazine* 7: 26. Permission for reprint courtesy of the Society for Information Display.

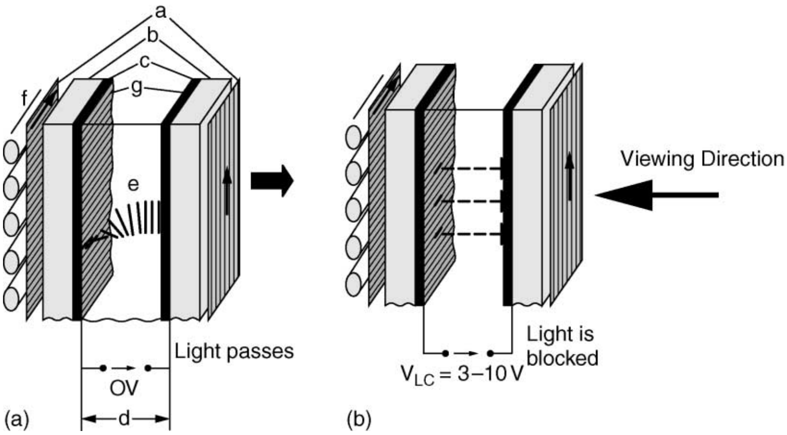

Figure 11 The operation of a simple liquid crystal display. Reproduced from Lueder E (1997) Fundamentals of passive and active-addressed liquid crystal displays. *SID Short Course S-1* (1997). Permission for reprint courtesy of the Society for Information Display.

response and moving graphic displays. These have found application in 'heads up displays' (HUD) for the military and medical industries and in consumer goods such as portable computers, pagers, cell phones, and 3D gaming.

The Active Matrix LCD

The active matrix liquid crystal display (AMLCD) device consists of a series of liquid crystal cells or windows whose characteristics can be controlled by an applied voltage to pass or restrict the passage of light through the windows via the external polarizers and the polarizing action of the LCD cells. These displays are used in laptop and tabletop computers, TVs, and for automotive displays (for driving information and entertainment). Figure 10 shows some present applications for AMLCDs in the automobile.

When the simple LCD is used without a transistor matrix, it is termed a passive display rather than an active one. Figure 11 below shows how a simple LCD cell operates.

The action shown in this figure is the twisted alignment nature of the liquid crystal material. When the polarizer on the left has its high transmission axis horizontal and the polarizer on the right of this cell has perpendicular alignment then the light from the backlight (f) is passed through the crossed polarizers. However, when voltage is applied to the cell, the liquid crystal is no longer twisted and the crossed polarizers prevent the light from going though the cell to the viewer. If the polarizers were not crossed but the polarization transmission axis was in the same direction in both polarizers, then the opposite effect would occur.

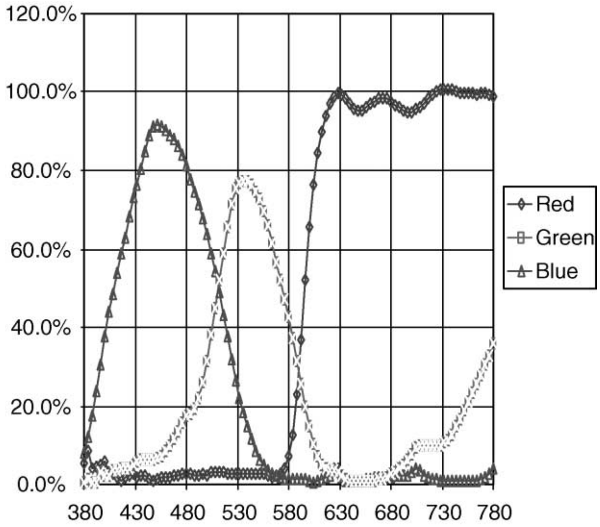

Figure 12 Typical color LCD filter transmission characteristics. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

In order to show a color image, the LCD is designed to have three or more cells clustered per color pixel. The following (Figure 12) is the transmission profiles of each of these red, green, and blue filters.

The backlight used is a fluorescent lamp with a white phosphor (made up of red, green, and blue photoluminescent phosphors). Figure 13 shows what would happen if only the blue phosphor was used and also shows the superposition of the mercury lines that penetrate the phosphor and contribute to the final spectral energy distribution.

Reflective LCD

For applications where there is suitable external light to view the display, no internal light source is used.

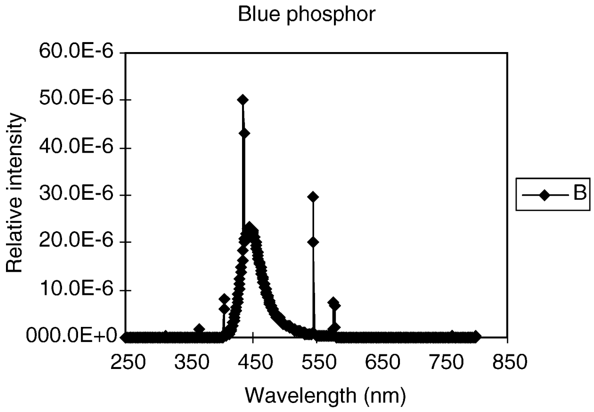

Figure 13 The relative luminance spectrum of the LCD. Backlight blue phosphor showing mercury lines. Reproduced from Donofrio RL and Abileah A (1998) LCD backlighting system – phosphors and colors. In: Donofrio RL and Musa S (eds) *Flat Panel and Vehicle Display '98, p. 123*. Permission for reprint courtesy of the Society for Information Display.

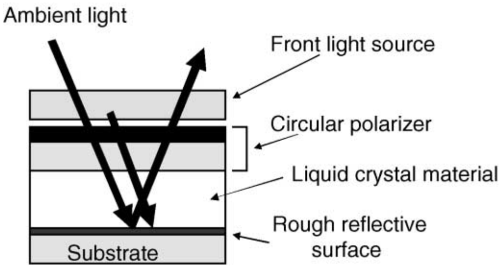

Figure 14 The action and construction of a reflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

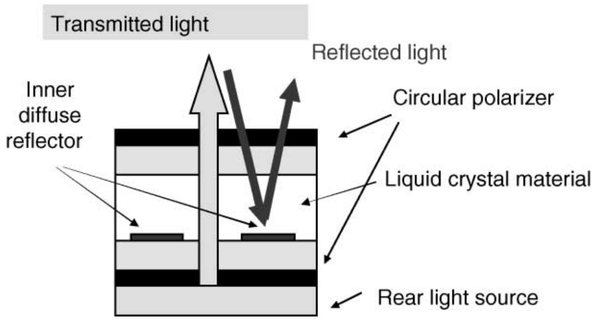

Figure 15 The action and cross-section of a transflective LCD. Reproduced with permission from Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35. Permission for reprint courtesy of the Society for Information Display.

This reflective LCD consists of a polarizer and retarder plates stacked on the front side of the cell. The cell with spacers contains the liquid crystal material and on the bottom of the cell is located a reflecting surface. It has been found that a dimpled reflecting surface makes a diffuse reflecting surface (Figure 14).

Transflective LCD

The importance of the transflective device is that it has better contrast under high ambient illumination than the standard transmission back lighted LCD and fair contrast for low ambient illumination.

The construction of this device consists of front and rear circular polarizers, back light, reflective elements, and novel filters. The back light is in the rear of the display. However, like the reflective LCD, internal reflectors are used on each pixel. The filters are designed spatially to be more transmissive for the reflective light than for the back light. Additionally the cell gap for the reflective light is smaller then the transmitted light region because the reflected light passes the cell twice whereas the transmitted light passes the cell once. This type of device seen in Figure 15, is finding uses in the auto industry and where the LCD is needed for both low and high ambient light levels.

Conclusions

In this brief review of displays we have tried to cover many different types of displays. The reader will in time find that there are many other variations on the displays discussed. The object is to give the reader an up-to-date review of the topics involved with displays and the various types of display in use today. By its nature it is a broad approach.

See also

Color and the World. **Instrumentation:** Photometry. **Light Emitting Diodes**. **Materials for Nonlinear Optics:** Liquid Crystals for NLO.

Further Reading

Donofrio RL (1981) Low Current Density Aging. *Proceedings of the Electro-Chemical Society*, Abstract #152, May 11, 1981.

Donofrio RL (2003) The road show. *Information Display* 7(3): 26–29.

Donofrio RL and Abileah A (1998) LCD backlighting. *Flat Panel and Vehicle Displays '98* : 119–124. Metropolitan Detroit Chapter of the SID.

Donofrio RL, Musa S and Wise K (eds) (1999) Vehicle Displays and Microsensors '99. *Proceedings of the 6th Strategic Forum and Technical Symposium*, Sept 22–23, 1999.

Flat Panel Display Measurements Standard, Version 1.0 VESA (Video Electronics Standards Association) May 15, 1998.

Forrest S (2002) *Fundamentals of Organic Light-Emitting Devices*. Short Course S-2 SID.

Green P and Haven D (1998) *Fundamentals of Emissive Displays*, Short Course S-3, SID.

Keller P (1991) *The Cathode-Ray Tube, Technology, History, and Applications*. New York: Palisades Press.

Leverence HW (1968) *An Introduction to Luminescence of Solids*. Dover Publications.

MacDonald W and Lowe A (1997) *Display Systems – Design and Applications*. SID Series in Display Technology, Wiley and Sons Ltd.

Sherr S (1993) *Electronic Displays*, 2nd edn. John Wiley.

Shionoya S and Yen WM (1998) *Phosphor Handbook*. CRC Press.

Stupp EH and Brennesholtz MS (1999) *Projection Displays*. SID Series in Display Technology, Wiley and Sons Ltd.

Watanabe R and Tomita O (2003) Active-Matrix LCDs for Mobile Telephones in Japan. *SID Information Display Magazine* 7(3): 30–35.

Web site atis.org/tg2k/. American National Standard TI.523-2001 'Def of Display Device'.

Wu ST and Yang DK (2001) *Reflective Liquid Crystal Displays*. SID Series in Displays Technology, Wiley and Sons Ltd.

ELECTROMAGNETICALLY INDUCED TRANSPARENCY

J P Marangos, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we examine how the process of electromagnetically induced transparency (EIT) can be used to increase the efficiency of non-linear optical frequency mixing schemes. We illustrate the basic ideas by concentrating upon the enhancement of resonant four-wave mixing schemes in atomic gases. To start we introduce the quantum interference phenomenon that gives rise to EIT. The calculation of EIT effects in a three-level atomic medium using a density matrix approach is presented. Next we examine the modified susceptibilities that result and that can be used to describe nonlinear mixing, and show how large enhancements in nonlinear conversion efficiency can result. Specific examples are briefly discussed along with a further discussion of the novel aspects of refractive index and pulse propagation modifications that result from EIT. The potential benefits to nonlinear optics of electromagnetically induced transparency are discussed. In an article of this nature it is not possible to credit all of the important contributions to this field, but we include a bibliography for further reading indicating some of the seminal contributions.

Electromagnetically induced transparency is the term applied to the process by which an otherwise opaque medium can be rendered transparent through laser-induced quantum mechanical interference processes. The linear susceptibility is substantially modified by EIT. Absorption is cancelled due to destructive interference between excitation pathways. The dispersion of the medium is likewise modified such that where there is no absorption the refractive index is unity and there can be a large value of normal dispersion leading to low values of group velocity. In contrast to the linear susceptibility the nonlinear susceptibility involving these same resonant laser fields will undergo constructive rather than destructive interference. This leads to a strong enhancement in nonlinear frequency mixing. For a normally transparent medium the nonlinear optical couplings will be small unless large-amplitude fields are applied. The most important consequence of EIT, in contrast to the usual case when a medium is transparent, is that the dispersion is not vanishing and the nonlinear couplings can be very large.

To understand how this comes about we must consider the system illustrated in Figure 1a where because of the resonant condition satisfied for the applied fields the atom can be simplified to just three-levels. The important parameters in this model system are that state |2> is metastable and has a dipole-allowed coupling to state |3>. The coupling between states |1> and |3> is also dipole-allowed

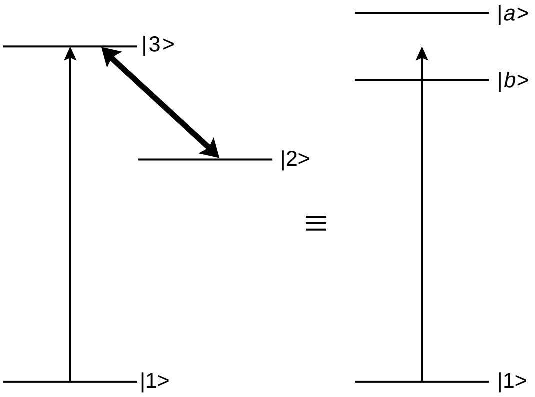

Figure 1 A three-level atomic system coupled to laser fields in the lambda configuration is shown on the left-hand side. In the limit of a strong coupling field this is equivalent to the dressed states |a> and |b> coupled to the ground state by the probe field alone as shown on the right-hand side.

and as a consequence state |3> can decay radiatively to either |1> or |2>. Let a coupling field be applied resonantly to the |2> − |3> transition, which may be cw or pulsed but should in the latter case be transform-limited. We define the Rabi coupling as $\Omega_{23} = \frac{\mu_{23}E}{\hbar}$ where E is the laser electric field amplitude and μ_{23} is the dipole moment of the transition. The Rabi frequency needs to be larger than the inhomogeneous broadening of the sample. A second field, the probe, typically of much lower intensity, is then applied in resonance with the |1> − |3> transition. If the condition $\Omega_{23} \gg \Omega_{13}$ is satisfied then it is convenient to replace the bare atomic states |2> and |3> with the dressed states (see Figure 1b):

$$|a> = \frac{1}{\sqrt{2}}[|2> + |3>] \qquad [1a]$$

$$|b> = \frac{1}{\sqrt{2}}[|2> - |3>] \qquad [1b]$$

Transitions from state |1> induced by the probe field to this pair of near-resonant dressed states are subject to exact cancellation at resonance if |2> is perfectly metastable. This is because the only contribution to the excitation amplitude comes from the matrix elements $<1|\mathbf{r}\cdot\mathbf{E}|3>$ as the |1> − |2> transition is dipole forbidden. The contributions from the equally detuned states |a> and |b> thus enter into the overall amplitude with opposite signs and equal magnitude as can be seen by inspection of eqns [1a] and [1b]. This leads to cancellation of the absorption amplitude. This type of absorption cancellation is well known and closely related to the so-called *dark states*.

Theoretical Treatment of EIT in a Three-Level Medium

It was realized by several workers in the 1970s that laser-induced interference effects could lead to a cancellation of absorption at certain frequencies. To gain a more quantitative understanding of the effects of the coupling field upon the optical properties of a dense ensemble of three-level atoms we require a treatment that computes the optical susceptibility of the medium. A treatment originally carried out by Harris *et al.* for a Λ scheme similar to that illustrated in Figure 1 was the first to derive the modified susceptibilities that will be discussed below. In that treatment the state amplitudes in the three-level atom were solved in the steady-state limit and from these the linear and nonlinear susceptibilities (see below) are then derived. In what follows we adopt a density matrix treatment as employed by various workers. This readily allows the inclusion of dephasing as well as population decay terms. The critical parameters in this system are the strengths of the fields (described in terms of the Rabi couplings), the detuning of the fields from resonance $\Delta\omega_{13} = \omega_{13} - \omega_P$ and $\Delta\omega_{23} = \omega_{23} - \omega_C$ (see Figure 2), the radiative decays from |3> to |1> and |2>, γ_c and γ_d and from |2>−|1> γ_a (although the latter is anticipated to be smaller). Extension to other configurations of the three states, such as a V or ladder scheme is implicit within this general treatment.

Figure 2 illustrates the system considered. This treatment will also address the nonlinear response in a four-wave mixing scheme completed by the two-photon resonant coupling applied between state |1> and |2>.

We write the interaction Hamiltonian as:

$$H = H_0 + V \qquad [2]$$

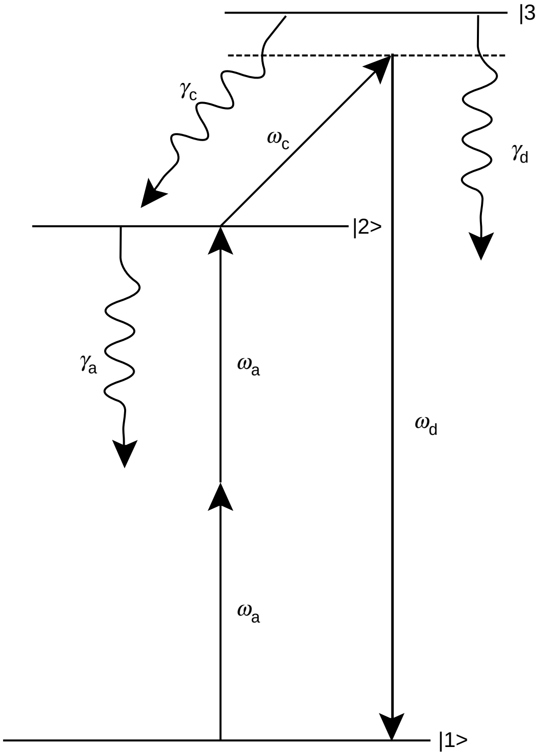

Figure 2 A four-wave mixing scheme incorporating the three-level lambda system in which electromagnetically induced transparency has been created for the generated field ω_d by the coupling field ω_c. The decay rates γ_i from the three atomic levels are also shown. For a full explanation of this system see text.

where H_0 is the unperturbed Hamiltonian of the system and is written as

$$H_0 = \hbar\omega_1|1\rangle\langle 1| + \hbar\omega_2|2\rangle\langle 2| + \hbar\omega_3|3\rangle\langle 3| \qquad [3]$$

and V is the interaction Hamiltonian and can be expressed

$$\begin{aligned} V = {} & \hbar\Omega_{\mathrm{a}}\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{a}}\mathrm{t}}|2\rangle\langle 1| + \hbar\Omega_{\mathrm{c}}\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{c}}\mathrm{t}}|2\rangle\langle 3| \\ & + \hbar\Omega_{\mathrm{d}}\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{d}}\mathrm{t}}|3\rangle\langle 1| + c.c \end{aligned} \qquad [4]$$

Note that the Rabi frequencies in the equation can be described as $\hbar\Omega_{ij} = \mu_{ij}|E(\omega_{ij})|$, where μ_{ij} is the dipole moment of the transition between states |i> and |j>, the Rabi frequency Ω_{a} is a two-photon Rabi frequency that characterizes the coupling between the laser field *a* and the atom for this two-photon transition. We have assumed for simplicity that $\omega_{\mathrm{a}} = \omega_{\mathrm{b}} = \omega_{12}/2$.

Assuming all the fields lie close to the resonance, the rotating wave approximation can be applied to the interaction picture and the interaction Hamiltonian V^I is given as

$$\begin{aligned} V^I = {} & \hbar\Omega_{\mathrm{a}}\mathrm{e}^{\mathrm{i}\Delta_{\mathrm{a}}\mathrm{t}}|2\rangle\langle 1| + \hbar\Omega_{\mathrm{c}}\mathrm{e}^{\mathrm{i}\Delta_{\mathrm{c}}\mathrm{t}}|2\rangle\langle 3| \\ & + \hbar\Omega_{\mathrm{d}}\mathrm{e}^{\mathrm{i}\Delta_{\mathrm{d}}\mathrm{t}}|3\rangle\langle 1| + c.c \end{aligned} \qquad [5]$$

where Δ_{a}, Δ_{c} and Δ_{d} refer the detunings of the fields and can be written as:

$$\begin{aligned} \Delta_{\mathrm{a}} &= \omega_{12} = 2\omega_{\mathrm{a}} \\ \Delta_{\mathrm{c}} &= \omega_{32} - \omega_{\mathrm{c}} \\ \Delta_{\mathrm{d}} &= \omega_{13} - \omega_{\mathrm{d}} \end{aligned} \qquad [6]$$

For the evaluation of the density matrix with this interaction V^I, the Schrödinger equation can be restated in terms of the density matrix components. This form is called the Liouville equation and can be written as:

$$\hbar\frac{\partial}{\partial t}\varrho_{ij}(t) = -\mathrm{i}\sum_k H_{ik}(t)\varrho_{ij}(t) + \mathrm{i}\sum_k \varrho_{ij}(t)H_{ik}(t) + \Gamma_{ij} \qquad [7]$$

where Γ_{ij} represents phenomenologically added decay terms (i.e. spontaneous decays, collisional broadening, etc.). This formalism leads to a set of nine differential equations for nine different density matrix elements that describe the three-level system.

To remove the optical frequency oscillations, a coordinate transform is needed and to incorporate the relevant oscillatory detuning terms into the off-diagonal elements we make the substitution:

$$\begin{aligned} \tilde{\varrho}_{12} &= \varrho_{12}\mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{a}}\mathrm{t}} \\ \tilde{\varrho}_{23} &= \varrho_{23}\mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{c}}\mathrm{t}} \\ \tilde{\varrho}_{31} &= \varrho_{31}\mathrm{e}^{-\mathrm{i}\Delta_{\mathrm{d}}\mathrm{t}} \end{aligned} \qquad [8]$$

This operation conveniently eliminates the time dependencies in the rate equations and the equations of motion for the density matrix are given by:

$$\begin{aligned} \frac{\partial}{\partial t}\varrho_{11} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{12} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{13} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}^*\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}^*\tilde{\varrho}_{31} + \Gamma_{11} \\ \frac{\partial}{\partial t}\varrho_{22} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}^*\tilde{\varrho}_{23} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}^*\tilde{\varrho}_{21} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{32} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{12} + \Gamma_{22} \\ \frac{\partial}{\partial t}\varrho_{33} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}^*\tilde{\varrho}_{23} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}^*\tilde{\varrho}_{31} + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{13} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{32} + \Gamma_{33} \\ \frac{\partial}{\partial t}\varrho_{23} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{22} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{33} - \mathrm{i}\Delta_{\mathrm{a}}\tilde{\varrho}_{23} \\ &\quad + \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{21} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{13} + \Gamma_{23} \\ \frac{\partial}{\partial t}\varrho_{21} &= \mathrm{i}\Omega_{\mathrm{a}}\tilde{\varrho}_{22} + \mathrm{i}\Omega_{\mathrm{a}}\tilde{\varrho}_{33} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{23} - \mathrm{i}\Delta_{\mathrm{c}}\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{31} - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{a}}\tilde{\varrho}_{31} + \Gamma_{21} \\ \frac{\partial}{\partial t}\varrho_{31} &= \mathrm{i}\frac{1}{2}\Omega_{\mathrm{d}}\tilde{\varrho}_{22} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{33} + \mathrm{i}\Omega_{\mathrm{d}}\tilde{\varrho}_{23} - \mathrm{i}\Delta_{\mathrm{c}}\tilde{\varrho}_{21} \\ &\quad - \mathrm{i}\frac{1}{2}\Omega_{\mathrm{c}}\tilde{\varrho}_{31} - \mathrm{i}\frac{1}{2}\Omega_{a}\tilde{\varrho}_{31} + \Gamma_{31} \end{aligned} \qquad [9]$$

Using the set of equations above and the relation $\tilde{\varrho}_{ij} = \tilde{\varrho}_{ij}^*$ we obtain equations for $\tilde{\varrho}_{12}$, $\tilde{\varrho}_{32}$, and $\tilde{\varrho}_{13}$. For the incoherent population relaxation the decays can be written:

$$\begin{aligned} \Gamma_{11} &= \gamma_{\mathrm{a}}\varrho_{22} + \gamma_{\mathrm{d}}\varrho_{33} \\ \Gamma_{22} &= -\gamma_{\mathrm{a}}\varrho_{22} + \gamma_{\mathrm{c}}\varrho_{33} \\ \Gamma_{33} &= -(\gamma_{\mathrm{c}} + \gamma_{\mathrm{d}})\varrho_{33} \end{aligned} \qquad [10]$$

and for the coherence damping:

$$\begin{aligned} \Gamma_{21} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{a}} + \gamma_{\mathrm{c}}] + \gamma_{21}^{\mathrm{col}}\right\}\varrho_{21} \\ \Gamma_{23} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{a}} + \gamma_{\mathrm{c}} + \gamma_{\mathrm{d}}] + \gamma_{23}^{\mathrm{col}}\right\}\varrho_{23} \\ \Gamma_{31} &= -\left\{\frac{1}{2}[\gamma_{\mathrm{d}} + \gamma_{\mathrm{c}}] + \gamma_{31}^{\mathrm{col}}\right\}\varrho_{31} \end{aligned} \qquad [11]$$

where $\gamma_{ij}^{\mathrm{col}}$ represents collisional dephasing terms which may be present.

This system of equations can be solved by various analytical or numerical methods to give the individual density matrix elements. Analytical solutions are possible if we can assume for instance that $\Omega_c \gg \Omega_a$, Ω_d and that there are continuous-wave fields so a steady-state treatment is valid. For pulsed fields or in the case where the generated field may be significant numerical solutions are in general required.

We are specifically interested in the optical response to a probe field at ω_d (close to resonance with the $|1\rangle - |3\rangle$ transition) that is governed by the magnitude of the coherence ϱ_{13}. We find ϱ_{13} making the assumption that only the coupling field is strong, i.e., that $\Omega_c \gg \Omega_a, \Omega_d$ holds. From this quantity the macroscopic polarization is obtained from which the susceptibility can be computed. The macroscopic polarization P at the transition frequency ω_{13} can be related to the microscopic coherence ϱ_{13} via the expression:

$$P_{13} = N\mu_{13}\varrho_{13} \qquad [12]$$

where N is the number of equivalent atoms in the ground state within the medium, and μ_{13} is the dipole matrix element associated with the transition. In this way real and imaginary parts of the linear susceptibility χ at frequency ω can be directly related to ϱ_{13} via the macroscopic polarization since the latter can be defined as:

$$P_{13}(\omega) = \varepsilon_0\chi(\omega)E \qquad [13]$$

where E is the electric field amplitude at frequency ω_d. The linear susceptibility (real and imaginary parts) is given by the following expressions:

$$\mathrm{Re}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0\hbar} \times \left[\frac{-4\Delta_{21}(|\Omega_2|^2 - 4\Delta_{21}\Delta_{32}) + 4\Delta_{31}\gamma_a^2}{4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \qquad [14]$$

$$\mathrm{Im}\chi_D^{(1)}(-\omega_d,\omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0\hbar} \times \left[\frac{8\Delta_{21}^2\gamma_d + 2\gamma_a(|\Omega_2|^2 + \gamma_d\gamma_a)}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \qquad [15]$$

We now consider the additional two-photon resonant field ω_a. This leads to a four-wave mixing process that generates a field at ω_d (i.e., the probe frequency). The nonlinear susceptibility that describes the coupling of the fields in a four-wave mixing process is given by the expression:

$$\chi_D^{(3)}(-\omega_d,\omega_a,\omega_a,\omega_c) = \frac{\mu_{23}\mu_{13}N}{6\varepsilon_0\hbar^3((\Delta_{21} - j\gamma_a/2)(\Delta_{31} - j\gamma_d/2) - (|\Omega_c|^2/4))^*} \times \sum_i \mu_{i1}\mu_{i3}\left(\frac{1}{\omega_i - \omega_a} + \frac{1}{\omega_i - \omega_b}\right)$$

Nonlinear Optical Processes

The dependence of the susceptibilities upon the detuning is plotted in Figure 3. In this plot the effects of inhomogeneous (Doppler) broadening are

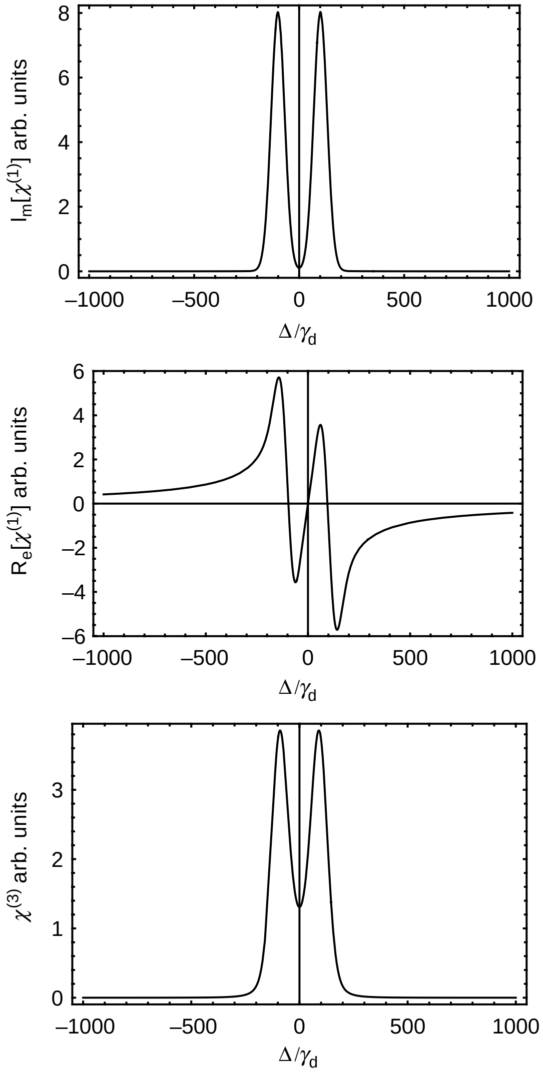

Figure 3 Electromagnetically induced transparency is shown in the case of significant Doppler broadening. The Doppler averaged values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. The Doppler width is taken to be $20\gamma_d$, $\Omega_c = 100\gamma_d$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

also incorporated by computing the susceptibilities over the inhomogeneous profile (see below for more details).

By inspection of eqn [14] we see that the absorptive loss at the minimum of $\mathrm{Im}[\chi^{(1)}]$ varies as γ_a/Ω_c^2. This dependence is a consequence of the interference discussed above. In the absence of interference (i.e., just simply Autler–Townes splitting) the minimum loss would vary as γ_d/Ω_c^2. Since $|3\rangle - |1\rangle$ is an allowed decay channel (in contrast to $|2\rangle - |1\rangle$) it follows that $\gamma_b \gg \gamma_a$ and so the absorption is much less as a consequence of EIT. $\mathrm{Re}[\chi^{(1)}]$ in eqn [15] and Figure 3 is also modified significantly. The resonant value of refractive index is equal to the vacuum value where the absorption is a minimum, the dispersion is normal in this region with a gradient determined by the strength of the coupling laser, a point we will return to shortly. For an unmodified system the refractive index is also unity at resonance but in that case there is high absorption and steep anomalous absorption. Reduced group velocities result from the steep normal dispersion that accompanies EIT. Inspection of the expression [16] and Figure 3 shows that $\chi^{(3)}$ is also modified by the coupling field. The nonlinear susceptibility depends upon $1/\Omega_c^2$ as is expected for a laser dressed system, however in this case there is not destructive but rather constructive interference, between the field-split components. This result is of great importance: it ensures that the absorption can be minimized at frequencies where the nonlinear absorption remains large.

As a consequence of constructive interference the nonlinear susceptibility remains resonantly enhanced whilst the linear susceptibility vanishes or becomes very small at resonance due to destructive interference. Large nonlinearity accompanying vanishing absorption (transparency) of course match conditions for efficient frequency mixing as a large atomic density can then be used. Moreover the dispersion (controlling the refractive index) also vanishes at resonance; this leads to perfect phase matching (i.e., zero wavevector mismatch between the fields) in the limit of a simple three-level system. As a result of these features a large enhancement of the conversion efficiency in this type of scheme can be achieved.

To compute the generated field strength Maxwell's equations must be solved using these expression for the susceptibility to describe the absorption, refraction, and nonlinear coupling in the medium. We will treat this within the slowly varying envelope approximation since the fields are cw or nanosecond pulses. To be specific we assume that the susceptibilities are time independent, i.e., that we are in the steady-state (cw) limit. We make the assumptions also that there is no pump field absorption and that we have plane waves. Under these assumptions the generated field amplitude A_d is given in terms of the other field amplitudes A_i by:

$$\frac{\partial}{\partial z}A_\mathrm{d} = \mathrm{i}\frac{\omega_\mathrm{d}}{4c}\chi^{(3)}A_\mathrm{a}^2A_\mathrm{c}\mathrm{e}^{-\mathrm{i}\Delta k_\mathrm{d}z} - \frac{\omega_\mathrm{d}}{2c}\mathrm{Im}[\chi^{(1)}]A_\mathrm{d} + \mathrm{i}\frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]A_\mathrm{d} \qquad [16]$$

where the wavevector mismatch is given by:

$$\Delta k_\mathrm{d} = k_\mathrm{d} + k_\mathrm{c} - 2k_\mathrm{a} \qquad [17]$$

The wavevector mismatch will be zero on resonance for the three-level atom considered in this treatment. In fact the contribution to the refraction from all the other atomic levels must be taken into account whilst computing Δk_d and it is implicit that these make a finite contribution to the wavevector mismatch.

We can solve this first-order differential equation with the boundary condition $A_\mathrm{d}(z=0)=0$ to find the generated intensity $I(\omega_\mathrm{d})$ after a length z:

$$I(\omega_\mathrm{d}) = \frac{3n\omega_\mathrm{d}^2}{8Z_0c^2}|\chi^{(3)}|^2|A_\mathrm{a}|^4|A_\mathrm{c}|^2 \times \frac{\left\{1+\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z} - 2\mathrm{e}^{-\frac{\omega_\mathrm{d}}{c}\mathrm{Im}[\chi^{(1)}]z}\cos\left(\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right)z\right\}}{\frac{\omega_\mathrm{d}^2}{4c^2}\mathrm{Im}[\chi^{(1)}]^2 + \left\{\Delta k + \frac{\omega_\mathrm{d}}{2c}\mathrm{Re}[\chi^{(1)}]\right\}^2}$$

where Z_0 is the impedance of free space. This expression is quantitatively correct for the case of the assumptions made. More generally the qualitative predictions and general scaling remain valid in the limit where the pulse duration is significantly longer than the time required to traverse the medium. Note that both real and imaginary parts of $\chi^{(1)}$ and $\chi^{(3)}$ play an important role, as we would expect for resonant frequency mixing. The influence of that part of the medium refraction which is due to other levels is contained in the terms with Δk. In the case of a completely transparent medium this becomes a major limit to the conversion efficiency.

Propagation and Wave-Mixing in a Doppler Broadened Medium

Doppler shifts arising from the Maxwellian velocity distribution of the atoms in the medium lead to a corresponding distribution in the detunings for the various members of the atomic ensemble. The response of the medium, as characterized by the susceptibilities, must therefore include the Doppler effect by performing a weighted sum over possible detunings. The weighting is determined by the Gaussian form of the Maxwellian velocity distribution. From this operation the effective values of the susceptibilities at a given frequency are obtained, and these quantities can be used to calculate the generated field. This step is of considerable practical importance as in most up-conversion schemes it is not possible to achieve Doppler-free geometries and the use of laser cooled atoms, although in principle possible, limits the atomic density that can be employed. The interference effects persist in the dressed profiles providing the coupling laser Rabi frequency is comparable to or larger than the inhomogeneous width. This is because the Doppler profile follows a Gaussian distribution which falls off much faster in the wings of the profile than the Lorentzian profile due to the natural broadening.

In the case considered with weak probe field, excited state populations and coherences remain small. The two-photon transition need not be strongly driven (i.e., a small two-photon Rabi frequency can be used) but a strong coupling laser is required. The coupling laser must be intense enough that its Rabi frequency is comparable to or exceeds the inhomogenous widths in the system (i.e., Doppler width), and a laser intensity of above 1 MW cm^{-2} is required for a typical transition. This is trivially achieved even for unfocused pulsed lasers, but does present a serious limit to cw lasers unless a specific Doppler-free configuration is employed. The latter is not normally suitable for a frequency up-conversion scheme if a significant up-conversion factor is required, e.g., to the vacuum ultraviolet (VUV); however recent experiments report significant progress in cw frequency up-conversion using EIT and likewise a number of other possibilities, e.g., laser-cooled atoms and standing-wave fields, have been proposed.

A transform-limited single-mode laser pulse is essential for the coupling laser field since a multimode field will cause an additional dephasing effect on the coherence, resulting in a deterioration of the quality of the interference. In contrast, whilst it is advantageous for the field driving the two-photon transition to be single mode (in order to achieve optimal temporal and spectral overlap with the EIT hole induced by the dressing laser), this is not essential since this field does not need to drive the coherence responsible for interference.

When a pulsed laser field is used additional issues must be considered. The group velocity is modified for pulses propagating in the EIT large reductions, e.g., by factors down to $<c/100$, in the group velocity have been observed. Another consideration beyond that found in the simple steady-state case is that the medium can only become transparent if the pulse contains enough energy to dress all the atoms in the interaction volume. The minimum pulse energy to prepare a transparency is:

$$E_{\text{preparation}} = \frac{f_{13}}{f_{23}} NL\hbar\omega \qquad [18]$$

where f_{ij} are the oscillator strengths of the transitions and NL the product of the density and the length. Essentially the number of photons in the pulse must exceed the number of atoms in the interaction volume to ensure all atoms are in the appropriate dressed state. This puts additional constraints on the laser pulse parameters.

Up-conversion to the UV and vacuum UV has been enhanced by EIT in a number of experiments. Only pulsed fields have so far been up-converted to the VUV with EIT enhancement. The requirements on a minimum value of $\Omega_c > \Delta_{\text{Doppler}}$ constrains the conversion efficiency that can be achieved because the $1/\Omega_c^2$ factor in eqn [17] ultimately leads to diminished values of $\chi^{(3)}$. The use of gases of higher atomic weight at low temperatures is therefore highly desirable in any experiment utilizing EIT for enhancement of four-wave mixing to the VUV. Conversion efficiencies, defined in terms of the pulse energies by E_d/E_a or E_d/E_c of a few percent have been achieved using the EIT enhancement technique. It is typically most beneficial to maximize the conversion efficiency defined by the first ratio since ω_a is normally in the UV and is the lower energy of the two applied pulses.

Nonlinear Optics with a Pair of Strong Coupling Fields in Raman Resonance

An important extension of the EIT concept occurs when two strong fields are applied in Raman resonance between a pair of states in a three-level system. Considering the system illustrated in Figure 1 we can imagine that both applied fields are now strong. Under appropriate adiabatic conditions the system evolves to produce the maximum possible value for the coherence $\varrho_{12} = 0.5$. Adiabatic evolution into the maximally coherent state is achieved by adjusting either the Raman detuning or the pulse sequence (counter to-intuitive order). The pair of fields may also be in single-photon resonance with a third level, in which case the EIT-like elimination of absorption will be important. This situation is equivalent to the formation of a darkstate, since neither of the two strong fields is absorbed by the medium. For sufficiently strong fields the single-photon condition need not be satisfied and a maximum coherence will still be achieved.

An additional field applied to the medium can participate in sum- or difference-frequency mixing with the two Raman resonant fields. The importance of the large value of coherence is that it is the source polarization that drives the new fields generated in the frequency mixing process. Complete conversion can occur over a short distance that greatly alleviates the constraints usually set by phase-matching in nonlinear optics. Recently near unity conversion efficiencies to the far-UV were reported in an atomic lead system where maximum coherence had been created. In a molecular medium large coherence between vibrational or rotational levels has also been achieved using adiabatic pulse pairs. Efficient multi-order Raman sideband generation has been observed to occur. This latter observation may lead the way to synthesizing very short duration light pulses since the broadband Raman sideband spectrum has been proved to be phase-locked.

Pulse Propagation and Nonlinear Optics for Weak CW Fields

In a Doppler-free medium a new regime can be accessed. This is shown in Figure 4 where the possibility now arises of extremely narrow transparency dips since very small values of Ω_c are now sufficient to induce EIT. The widths of these features are typically subnatural and are therefore accompanied by very steep normal dispersion, which corresponds to a much reduced group velocity. The ultraslow propagation of pulses is one of the most dramatic manifestations of EIT in this regime. Nonlinear susceptibilities are now very large as there is constructive interference controlling the value and the splitting of the two states is negligible compared to their natural width. Nonlinear optics at very low light levels, i.e., at the few-photon limit, is possible in this regime.

Propagation of pulses is significantly modified in the presence of EIT. Figure 4 shows the changes to $\mathrm{Re}[\chi^{(1)}]$ that arise. Within the transparency dip there exists a region of steep normal dispersion. In the vicinity of resonance this is almost linear and it becomes reasonable to consider the leading term only that describes the group velocity. An analysis of the refractive changes has been provided by Harris who expanded the susceptibilities (both real and imaginary parts) of the dressed atom around the resonance frequency to determine various terms in $\mathrm{Re}[\chi^{(1)}]$. The first term of the series (zero order) $\mathrm{Re}[\chi^{(1)}](\omega_{13}) = 0$ corresponds to the vanishing dispersion at resonance. The next term $\partial[\mathrm{Re}\chi^{(1)}](\omega)/\partial\omega$ gives the slope of the dispersion curve; at ω_{13} this takes the value:

$$\frac{\partial \mathrm{Re}\chi^{(1)}(\omega_{13})}{\partial \omega} = \frac{|\mu_{13}|^2 N 4(\Omega_c^2 - \gamma_a^2)}{2\pi\varepsilon_0(\Omega_c^2 + \gamma_a \gamma_d)} \qquad [19]$$

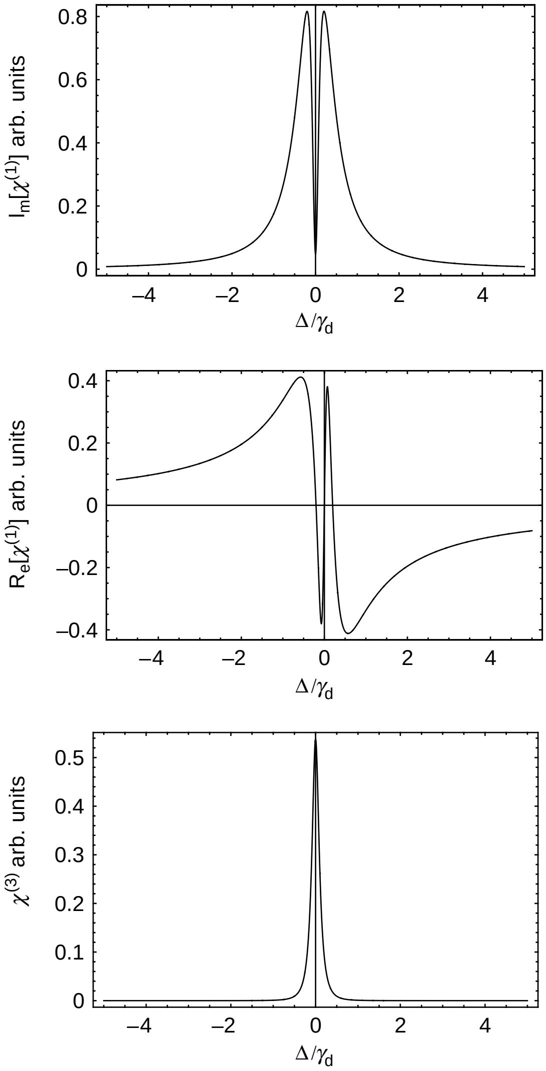

Figure 4 Electromagnetically induced transparency is shown in the case where there is no significant Doppler broadening. The values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. We take $\Omega_c = \gamma_d/5$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

The slope of the dispersion profile leads to a reduced group velocity v_g:

$$\frac{1}{v_g} = \frac{1}{c} + \frac{\pi}{\lambda}\frac{\partial \chi^{(1)}}{\partial \omega} \qquad [20]$$

From the expression for $\partial\chi/\partial\omega$ we see that this slope is steepest (and so v_g mimimum) in the case where $\Omega_c \gg \Gamma_2$ and $\Omega_c^2 \gg \Gamma_2\,\Gamma_3$ but is still small compared to Γ_3 (i.e., $\Omega_C < \Gamma_3$) and hence $\partial\chi/\partial\omega \propto 1/\Omega_c^2$. In the limit of small Ω_c the following expression for v_g therefore holds:

$$v_g = \frac{\hbar c \varepsilon_0 [\Omega_c]^2}{2\omega_d |\mu_{13}|^2 N} \qquad [21]$$

Extremely low group velocities, down to a few meters per second, are achieved in this limit using excitation of the hyperfine split ground states in either laser cooled atomic samples or Doppler-free configurations in finite temperature samples. Recently similar light slowing has been observed in solids. Storage of the optical pulse within the medium has also been achieved by adiabatically switching off the coupling field and thus trapping the optical excitation as an excitation within the hyperfine ground states for which the storage time can be very long (>1 ms) due to very low dephasing rates. Since the storage scenario should be valid even for single photons this process has attracted considerable attention recently as a means to enable quantum information storage.

Extremely low values of Ω_c are sufficient to induce complete transparency (albeit in a very narrow dip) and at this location the nonlinear response is resonantly enhanced. Very high efficiency nonlinear frequency mixing and phase conjugation at low light levels have been reported under these conditions. It is expected that high-efficiency nonlinear optical processes will persist to ultralow intensities (the few photon level) in an EIT medium of this type.

One example of highly enhanced nonlinear interactions is the predicted large values of the Kerr type nonlinearity (nonlinear refractive index). The origin of this can be seen by considering the steep dispersion profile in the region of the transparency dip in Figure 4. Imagine that we apply an additional field ω_f, perhaps a very weak one, at a frequency close to resonance between state |2> and a fourth level |4>. The ac Stark shift caused by this new field to the other three level, although small, will have a dramatic effect upon the value of the refractive index because of the extreme steepness of the dispersion profile. Essentially even a very small shift of the resonant wavelength causes a large change in the refractive index for a field applied close to this frequency. It is predicted that strong cross-phase modulations will be created in this process between the fields ω_f and ω_d, even in the quantum limit for these fields. This is predicted to lead to improved schemes for quantum nondemolition measurements of photons through the measurement of the phase-shifts they induce (through cross-phase modulation) on another quantum field. This type of measurement has direct application in quantum information processing.

See also

Nonlinear Optics, Applications: Phase Matching; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Arimondo E (1996) Coherent population trapping in laser spectroscopy. *Progress in Optics* 35: 257–354.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50: 36–42.

Harris SE and Hau LV (1999) Non-linear optics at low light levels. *Physical Review Letters* 82: 4611–4614.

Harris SE, Field JE and Imamoglu A (1990) Nonlinear optical processes using electromagnetically induced transparency. *Physical Review Letters* 64: 1107–1110.

Hau LV, Harris SE, Dutton Z and Behroozi CH (1999) Light speed reduction to 17 metres per second in an ultra-cold atomic gas. *Nature* 397: 594–598.

Marangos JP (2001) Electromagnetically induced transparency. In: Bass M, *et al.* (ed.) *Handbook of Optics*, vol. IV, ch. 23. New York: McGraw-Hill.

Merriam AJ, Sharpe SJ, Xia H, *et al.* (1999) Efficient gas-phase generation of vacuum ultra-violet radiation. *Optics Letters* 24: 625–627.

Schmidt H and Imamoglu A (1996) Giant Kerr nonlinearities obtained by electromagnetically induced transparency. *Optics Letters* 21: 1936–1938.

Scully MO (1991) Enhancement of the index of refraction via quantum coherence. *Physical Review Letters* 67: 1855–1858.

Scully MO and Suhail Zubairy M (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Sokolov AV, Walker DR, Yavuz DD, *et al.* (2001) Femtosecond light source for phase-controlled multi-photon ionisation. *Physical Review Letters* 87: 033402-1.

Zhang GZ, Hakuta K and Stoicheff BP (1993) Nonlinear optical generation using electromagnetically induced transparency in hydrogen. *Physical Review Letters* 71: 3099–3102.

ENVIRONMENTAL MEASUREMENTS

Contents

Doppler Lidar
Hyperspectral Remote Sensing of Land and the Atmosphere
Laser Detection of Atmospheric Gases
Optical Transmission and Scatter of the Atmosphere

Doppler Lidar

R M Hardesty, National Oceanic and Atmospheric Administration, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lidars developed for the measurement of atmospheric winds make use of the Doppler effect, in which radiation scattered or emitted from a moving object is shifted in frequency as a result of the movement of the particles. Atmospheric Doppler lidars irradiate a volume of atmosphere with a pulse of very narrowband, laser-produced radiation, then detect the change in frequency of the radiation backscattered from atmospheric aerosol particles or molecules present in the volume. Although Doppler lidars are conceptually similar to Doppler weather radars, lidar wavelengths are 3–4 orders of magnitude shorter than radar wavelengths, producing some important differences in propagation and scattering characteristics. Radiation at wavelengths in the optical region of the spectrum is efficiently scattered by aerosols and molecules, therefore Doppler lidars, unlike commonly used Doppler weather radars, do not require the presence of hydrometeors or insects to obtain useful results. Also, optical radiation can be tightly collimated, virtually eliminating effects of ground clutter and enabling lidar probing of small volumes to within a few meters of terrain or structures.

The primary disadvantage of lidar techniques is the severe attenuation of the optical radiation by cloud water droplets and fog. Doppler lidars do not typically probe beyond the edge of most atmospheric clouds – the one exception being tenuous ice clouds such as cirrus, which often are characterized by low optical extinction and high backscatter, making them excellent lidar targets. Doppler lidars also do not have the extended range, which can exceed 100 km, available from contemporary meteorological radars when scatterers are present.

In clear air, however, the characteristics and capabilities of Doppler lidar are well suited for obtaining detailed wind and turbulence observations for a wide variety of applications. Lidar beams can be easily scanned to characterize motions within very confined three-dimensional spaces such as shallow atmospheric boundary layers, narrow canyons, and turbulent structures. The relative compactness of Doppler lidars makes deployment on aircraft or other moving platforms extremely viable, and a Doppler lidar deployed on a satellite has been proposed as a way to obtain global measurements of atmospheric wind fields. By scanning a lidar beam from an orbiting satellite and analyzing the returns backscattered from clouds, aerosols, and molecules, a satellite-based instrument could provide important wind information for numerical forecast models.

Doppler Lidar Fundamentals

Backscattered Light

As laser light propagates through the atmosphere, some of the incident energy is scattered by the atmospheric molecules and constituents through which the light passes. In lidar applications, the light backscattered (scattered directly back toward the source) is collected and analyzed. The backscattering properties of an atmospheric particle depend on its refractive index, shape, and size. In the special case when the particle is much smaller than the wavelength of the incident radiation, as is the case for laser radiation scattered by atmospheric molecules, the scattering process is characterized as Rayleigh scattering. Because, in the Rayleigh scattering regime the energy backscattered by a particle increases proportionally as the inverse of the fourth power of the wavelength, Doppler lidar systems designed for molecular scatter operate at short wavelengths. Molecular Doppler lidar systems typically operate in the visible or ultraviolet spectral regions.

Aerosol particles, the other component of the atmosphere that scatters laser light, produce Mie scattering, which applies when the diameter of the scatterers is not orders of magnitude smaller than the incident wavelength. Aerosol particles present in the atmosphere include dust, soot, smoke, and pollen, as well as liquid water and ice. Although for Mie scattering the relationship between the power backscattered by an ensemble of aerosol particles and the incident wavelength is not simply characterized, most studies have shown that the energy backscattered increases roughly proportional to the first or second power of the inverse of the incident wavelength, depending on the characteristics of the scattering aerosol particles. In a polluted environment, such as in the vicinity of urban areas, an abundance of large particles often results in a roughly linear relationship between the inverse of the incident wavelength and the backscattered energy, while in more pristine environments, such as the free troposphere, the inverse wavelength/backscatter relationship can approach or exceed a square-law relationship.

The primary objective in Doppler lidar is to measure the Doppler frequency shift of the scattered radiation introduced by the movements of the scattering particles. Figure 1 shows a typical spectrum of the radiation collected at the lidar receiver for a volume of atmosphere irradiated by a monochromatic laser pulse. Molecular scattering produces the broadband distribution in Figure 1, where the broadening results from the Doppler shifts of the radiation backscattered from molecules moving at their thermal velocities. The width of the molecular velocity distribution in the atmosphere ranges from ~320 to 350 m s^{-1}, scaling as the square root of the temperature. In the center of the spectrum is a much narrower peak resulting from scattering of the light by aerosol particles. Because the thermal velocity of the much larger aerosol particles is very low, the width of the distribution of the aerosol return is determined by the range of velocities of particles moved about by small-scale turbulence within the scattering volume. This is typically on the order of a few m s^{-1}. Also shown in Figure 1 is an additional broadband distribution, due to scattered solar radiation collected at the receiver. If the laser source is not monochromatic, the backscattered signal spectrum is additionally broadened, with the resulting spectrum being the convolution of the spectrum shown in Figure 1 with the spectrum of the laser pulse. As seen in the figure, the entire spectrum is Doppler-shifted in frequency, relative to the frequency of the laser pulse. The object for a Doppler lidar system is to measure this shift, given by $\delta f = 2v_{\mathrm{rad}}/\lambda$, where v_{rad} is the component of the mean velocity of the particles in the direction of propagation of the lidar pulse and λ is the laser wavelength.

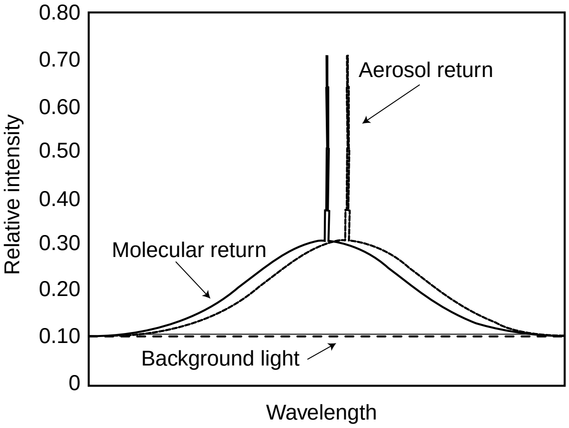

Figure 1 Distribution of wind velocities in a lidar return, showing Doppler shift resulting from wind-induced translation of the scatterers. The ratio of aerosol to molecular signals increases with increasing wavelength.

Elements of a Doppler Lidar

Doppler lidar systems can be designed primarily to measure winds from aerosol-scattered radiation, molecule-scattered radiation, or both. The type of system places specific requirements on the primary components that comprise a Doppler lidar system. A Doppler lidar consists of a laser transmitter to produce pulses of energy that irradiate the atmospheric volume of interest, a receiver which collects the backscattered photons and estimates the energy and Doppler shift of the return, and a beam pointing mechanism for directing the laser beam and receiver field of view together in various directions to probe different atmospheric volumes. Independent of whether the primary scatterers are molecules or aerosol particles, the system design criteria for a Doppler system are driven by a fundamental relationship between the error in the estimate of mean frequency shift δf_1, the bandwidth of the return f_2 (proportional to the distribution of velocities), and the number of incident backscattered photons detected N as

$$\delta f_1 \propto f_2/N^{1/2} \qquad [1]$$

Thus, the precision of the velocity estimate is improved by increasing the number of detected signal photons and/or decreasing the bandwidth of the backscattered signal. It is obvious from the equation that a significantly greater number of photons are required to achieve the same precision in a Doppler measurement from a molecular backscattered signal, characterized by higher bandwidth f_2, compared to the number required to compute velocities from aerosol returns. The improved measurement precision gained from a narrow bandwidth return also indicates that the laser transmitter in a Doppler lidar system should be characterized by narrow spectral width (considerably narrower than the spectral broadening caused by the distribution of velocities), and to increase the number of photons detected, maximum transmit energy.

Narrowband, single-frequency pulses are obtained by employing a low-power, precisely frequency-stabilized master oscillator laser, whose radiation is either used to seed a stabilized, higher-energy laser cavity (injection seeding) or amplified in one or more optical amplifiers (master-oscillator, power amplifier, (MOPA)) to produce frequency-stabilized pulses.

The lidar receiver gathers the backscattered photons and extracts the wind velocity as a function of the range to the scattering volume. This requires a telescope, to gather and focus the scattered radiation, and a receiver to detect the scattered radiation and determine the Doppler shift. Frequency analysis in a Doppler lidar receiver is carried out using one of two techniques: coherent detection (also known as heterodyne detection) or direct detection (alternately labeled incoherent detection). The techniques differ fundamentally. In a heterodyne receiver, frequency analysis is carried out on a digitized time series created by mixing laser radiation with the backscattered radiation, while in direct detection lidar an interferometer optically analyzes the backscattered radiation to produce a light pattern which contains the information on the frequency content.

Coherent (Heterodyne) Doppler Lidar

Description

Coherent or heterodyne lidar is implemented by optically mixing the backscattered laser light with radiation from a stable, continuous-wave, local oscillator (LO) laser whose frequency is precisely controlled to be a known frequency offset, typically on the order of tens of MHz, from that of the laser transmitter (Figure 2). The mixing process at the face of an optical detector generates, after high pass filtering, an electrical signal with amplitude proportional to amplitude of the backscattered electromagnetic field and frequency equal to the difference between the backscattered field frequency and the LO laser field frequency. This signal is sampled, and then digitally processed to estimate the range-dependent mean frequency shift of the backscattered signal, from which the radial wind component can be derived. A property of coherent lidar is that, because of constructive and destructive interference of the returns from individual scatterers, single pulse returns are characterized by random fluctuations in the amplitude and phase of the detected time series and Fourier spectrum. Consequently, averaging (sometimes referred to as accumulation) over multiple pulses usually is necessary to produce a sufficiently precise estimate of the signal mean frequency estimate. Because the phase of the detected time series is random within each pulse, averaging must be carried out in the power spectrum or autocorrelation function domain, rather than on the detected time series.

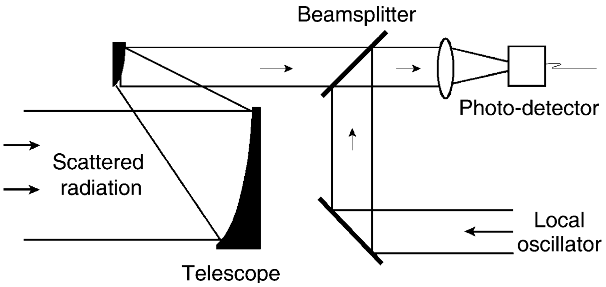

Figure 2 Schematic of coherent detection of backscattered radiation.

The optical mixing process in coherent Doppler lidar provides both benefits and design challenges. Because receiver efficiencies (equivalent fraction of incident photons entering the receiver that are converted to electrons) are quite high, wind velocities can be estimated from very weak backscattered signals. Also, signal processing is performed on a time series derived from the mixed signal, enabling the use of electrical filters to produce a narrow receiver bandwidth and effectively eliminating broadband solar background light as a major source of noise. As a result, unlike many lidars coherent Doppler lidar performance is not degraded under daylight conditions.

The primary limitation of coherent detection results from the added photon noise from the local oscillator radiation (which is usually much stronger than the backscattered radiation). The effect of local oscillator photon noise is to define a system noise level below which very weak signals cannot be practically extracted, even with substantial multiple-pulse averaging. In addition, because efficient mixing requires phase coherence of the backscattered signal field across the detector, coherent lidar performance at longer ranges is degraded by strong optical turbulence along the path of the laser pulse. These adverse effects of turbulence increase for shorter wavelength systems.

Although measurements at wavelengths near 1.06 μm have been demonstrated, coherent Doppler lidar systems used for regular atmospheric probing operate in the eye-safe, infrared portion of the spectrum at wavelengths longer than 1.5 μm. Currently, the two most common system wavelengths for coherent lidar wind systems are in atmospheric window spectral regions around 2 and 10.6 μm. Early coherent Doppler lidar measurements, beginning in the 1970s, employed CO_2 laser transmitters and local oscillators operating at wavelengths near 10.6 μm. Pulsed CO_2 laser transmitters with as much as 10 J of energy have since been demonstrated, and systems with 1 J lasers have probed the atmosphere to ranges of 30 km or more. In the late 1980s, solid state laser transmitters operating near 2 μm wavelengths were introduced into coherent lidar wind-measuring systems. The compact size and potential reliability advantages of solid-state transmitters, in which the transmitter laser is optically pumped by an array of laser diodes, provide advantages over larger CO_2 laser technology. Also, because for a given measurement accuracy the range-resolution obtainable is proportional to wavelength, 2 μm instruments have an enhanced capability to probe small-scale features. However, although development of higher energy, single-frequency coherent lidar sources operating at 2 μm region, as well as at 1.5 μm, is currently an active research area, solid state lasers with pulse energies greater than several tens of mJ have yet to be incorporated into lidar systems for atmospheric probing.

Applications of Coherent Doppler Lidar

Coherent lidars have been used to measure winds for a variety of applications, and from an assortment of platforms, such as ships and aircraft. Since these lidars operate in the infrared where aerosol scattering dominates molecular scattering, they require aerosol particles to be present at some level to obtain usable returns. Although clouds also provide excellent lidar targets, most of the more useful applications of coherent lidars involve probing the atmospheric boundary layer or lower troposphere where aerosol content is highest. Because of the capability to scan the narrow lidar beam directly adjacent to terrain, a unique application of lidar probing is the measurement of wind structure and evolution in complex terrain such as mountains and valleys. Over the past two decades, Doppler lidar studies have been used, for example, to study structure of damaging windstorms on the downwind side of mountain ranges, advection of pollution by drainage flows from valleys, and formation of mountain leeside turbulence as a potential hazard to landing aircraft. The interactions of wind flows with complex terrain can produce dangerous conditions for transportation, especially aircraft operations. Figure 3 shows a strong mountain wave measured by a Doppler lidar near Colorado Springs, during an investigation of the effects of downslope winds and turbulence on aircraft operations at the Colorado Springs airport. The mountains are on the right of the figure, where the strong wave with wavelength of about 6 km can be seen. Note the reversal in winds near the location where the wave descends to the surface. The Colorado Springs study was prompted by the crash of a passenger jet while landing at the airport during a period when the winds were down the slope of the Rocky Mountains just to the west of the airport.

A Doppler lidar was recently deployed in an operational mode for wind shear detection at the new Hong Kong International Airport. Because winds flowing over mountainous terrain just south of the

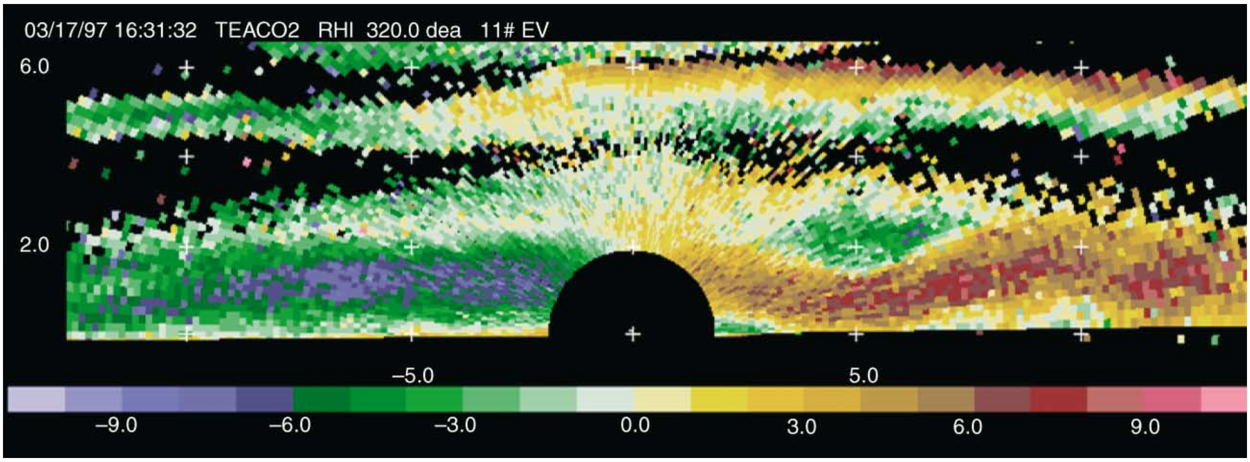

Figure 3 Vertical scan of the winds in a mountain wave measured by a 10.6 μm coherent Doppler lidar near Colorado Springs, Colorado. Courtesy L. Darby, NOAA.

airport can produce mountain waves and channeled flows, wind shear is frequently encountered during landing and takeoff operations. Figure 4 shows a gust front situated just to the west of the airport, characterized by a sharp wind change (which would produce a corresponding change in airspeed) approaching 13 m s^{-1} over a distance of less than one kilometer. Providing a capability to detect such events near the airport and warn pilots during approach or preparation for takeoff was the primary reason for deployment of the lidar. Within the first year of operation the Hong Kong lidar was credited with improving wind shear detection rates and providing more timely warnings for pilots of the potentially hazardous conditions. As of August 2003 this lidar had logged more than 10 000 hours of operation.

Because coherent Doppler lidars are well matched to applications associated with probing small-scale, turbulent phenomena, they have also been deployed at airports to detect and track wing tip vortices generated by arriving or departing aircraft on nearby parallel runways. In the future, a network of ground-based lidars could provide information on vortex location and advection speed as well as wind shear, leading to a potential decrease in congestion at major airports. Also, compact lidar systems looking directly ahead of research aircraft have shown the capability to detect wind changes associated with potentially hazardous clear-air turbulence. In the future, a Doppler lidar installed on the commercial aircraft fleet could potentially look ahead and provide a warning to passengers to fasten seat belts before severe turbulence is encountered.

The high resolution obtainable in a scanning lidar enables visualization of finely structured wind and turbulence layers. Figure 5 shows an image of turbulence associated with a nocturnal low-level jet just 50 m above the surface obtained at night over flat

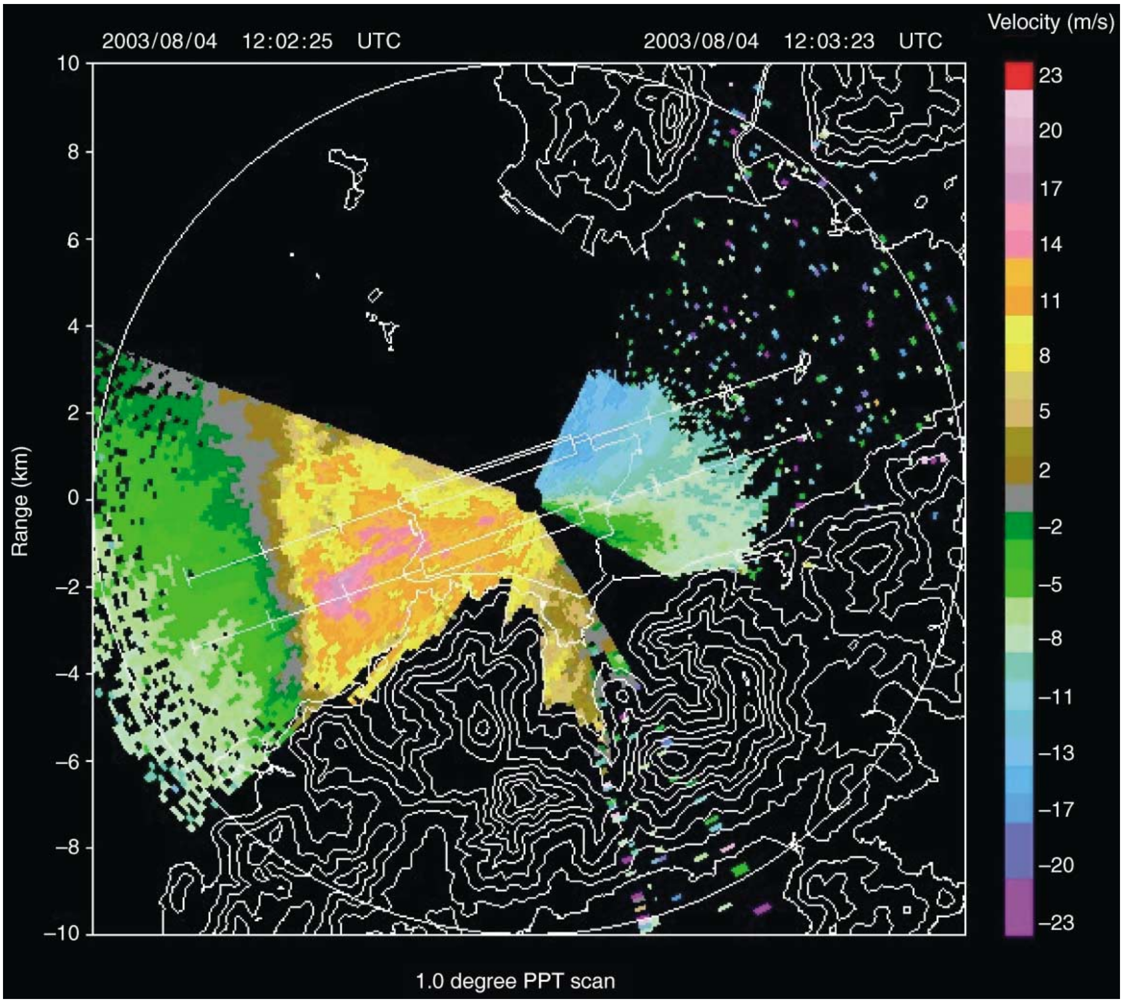

Figure 4 Horizontal depiction of the radial component of the wind field measured at Hong Kong Airport by a 2.02 μm Doppler lidar during a gust front passage, showing strong horizontal shear just west of the airport. Courtesy C. M. Shun, Hong Kong Observatory.

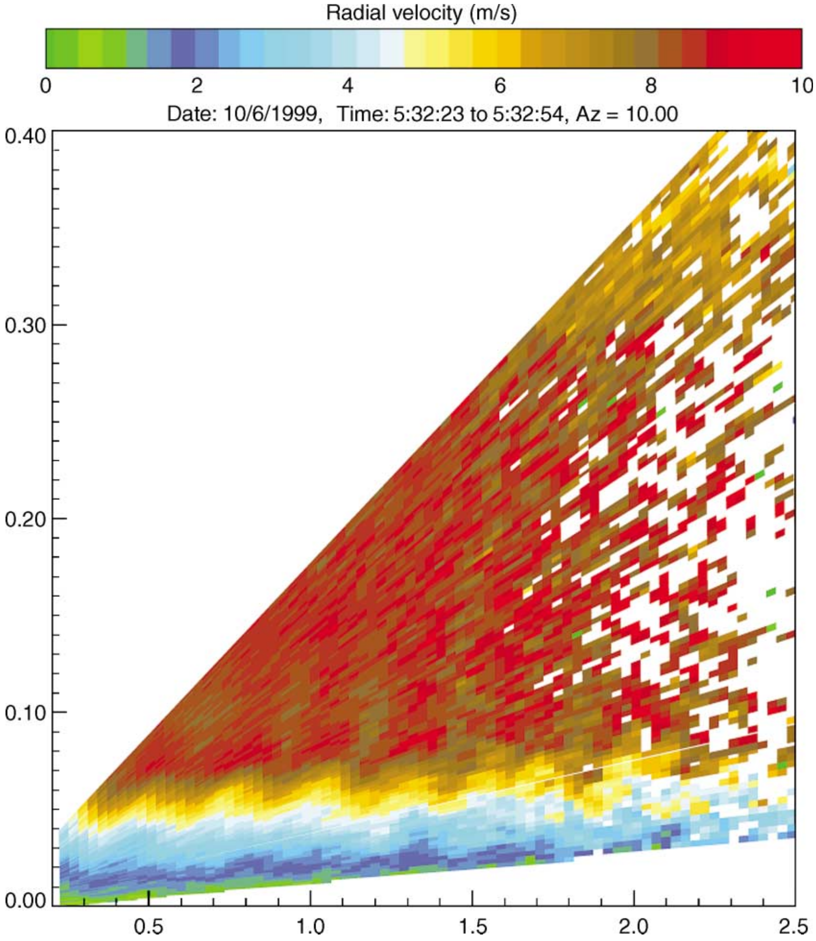

Figure 5 Turbulence structure within a nocturnal stable layer at a height of 50 m. Note the stretching of the vertical scale. Courtesy R. Banta, NOAA.

terrain in Kansas. Although a low-level jet was present on just about every evening during the experiment, similar images obtained at different times by the Doppler lidar illustrated markedly different characteristics, such as a wide variation in the observed mechanical turbulence along the interface. Such observations enable researchers to improve parameterizations of turbulence in models and better understand the conditions associated with vertical turbulent transport and mixing of atmospheric constituents such as pollutants.

By deploying Doppler lidars on moving platforms such as ships and aircraft, the spatial coverage of the measurements can be greatly increased. Although removal of platform motion and changes in orientation are not trivial, aircraft-mounted lidar can map out such features as the low-level jet and the structure of boundary layer convective plumes. Figure 6 shows horizontal and vertical motions associated with convective plumes measured along an approximately 60 km path over the southern Great Plains of the United States. From such measurements estimates of turbulence intensity, integral scale, and other boundary layer characteristics can be computed.

Direct Detection Doppler Lidar

Description

Direct detection or incoherent Doppler lidar has received significant attention in recent years as an alternative to coherent lidar for atmospheric wind measurements. In contrast to coherent lidar, in which an electrical signal is processed to estimate Doppler shift, an optical interferometer, usually a Fabry–Perot etalon, serves as the principal element in a direct detection lidar receiver for determining the frequency shift of the backscattered radiation.

One implementation of a direct-detection Doppler lidar receiver is the ‘fringe imaging’ technique. In this design, the interferometer acts as a spectrum analyzer. The backscatter radiation is directed through a

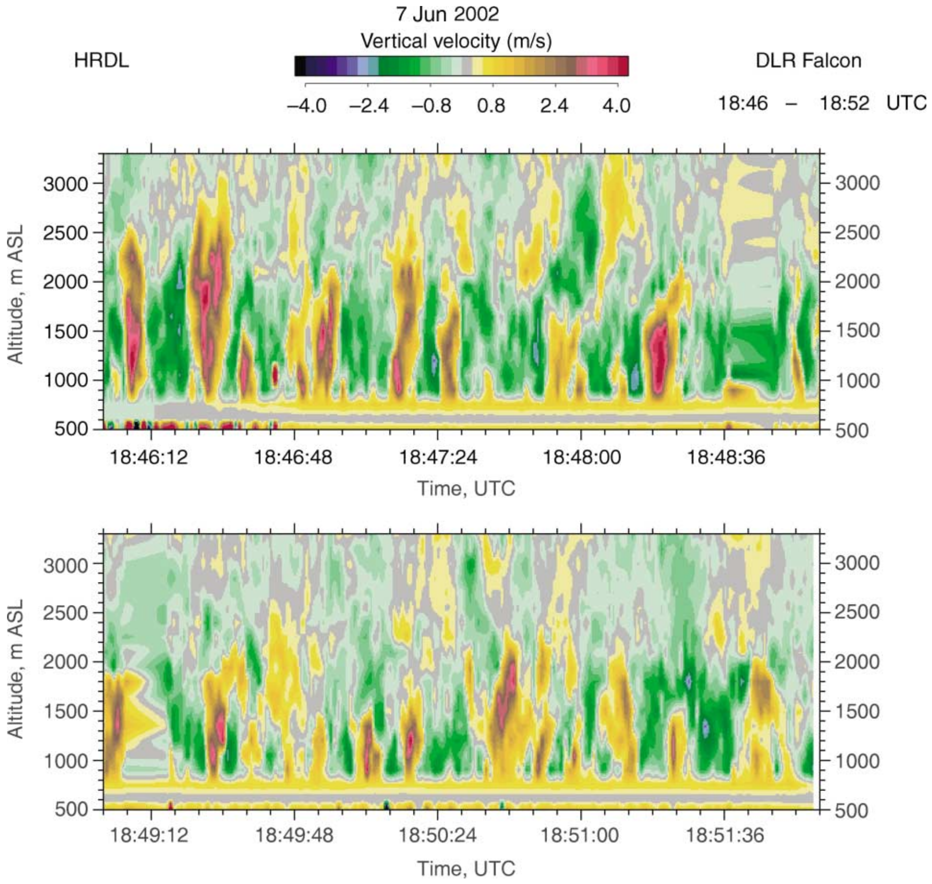

Figure 6 Vertical motion in convective plumes measured by a vertically pointing airborne coherent Doppler lidar. Total horizontal extent of the measurements is approximately 60 km, reddish colors correspond to upward motions.

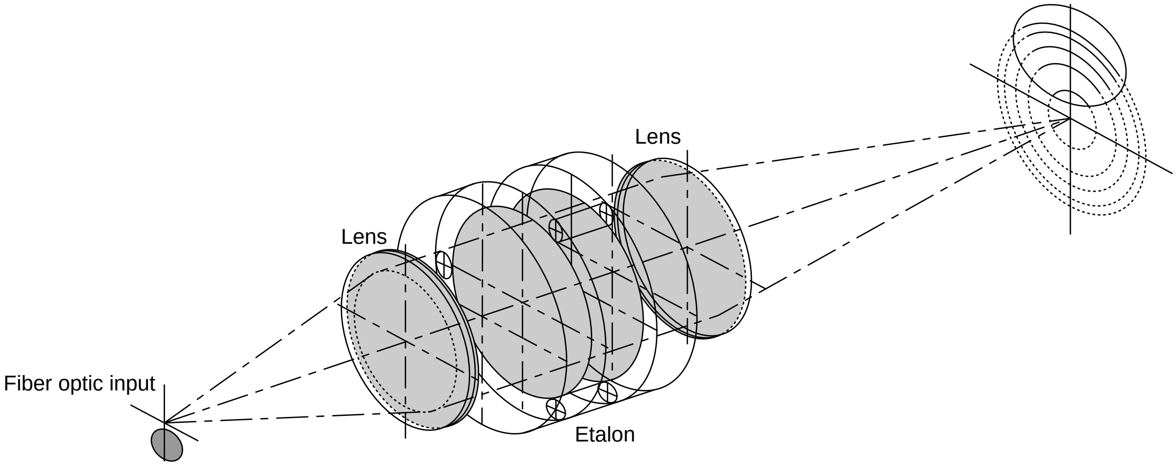

Figure 7 Schematic of Fabry–Perot etalon in a direct detection, fringe-imaging lidar receiver. Courtesy P. Hays, Michigan Aerospace Corporation.

Fabry–Perot interferometer, which produces a ring pattern in the focal plane (Figure 7). The spectral content information of the incident radiation is contained in the radial distribution of the light. Each ring corresponds to an order of the interferometer and is equivalent to a representation of the backscattered signal frequency spectrum. As the mean frequency of the backscattered radiation changes, the rings move inward or outward from the center. To extract a spectrum of the backscattered light, one or more of the rings are imaged onto a two-dimensional detector, and the resulting pattern analyzed. The rings can either be detected using a multi-element ring detector, or can be converted to a linear pattern with a circle-to-line converter optic and then imaged on to a charge-coupled device array device.

An alternate direct detection receiver configuration, generally called the double edge technique, is to employ two interferometers as bandpass filters, with the center wavelength of each filter set above and below the laser transmitter wavelength, as shown in Figure 8. The incoming radiation is split between the two interferometers, and the wavelength shift is computed by examining the ratio of the radiation transmitted by each interferometer. Both the double edge and fringe-imaging techniques have demonstrated wind measurements to heights well into the stratosphere. The major challenge associated with the double edge receiver is optimizing the instrument when both aerosol and molecular scattered radiation are present, since in general the change in transmission as a function of velocity is different for the aerosol and molecular signals. Double edge receivers optimized for both aerosol and molecular returns place the bandpass of the etalon filters at the precise wavelength where the change in transmission with change in Doppler shift is the same for both aerosol and molecular returns.

For both types of direct detection receivers described above, much of the radiation incident on the interferometer is reflected out of the system, reducing the overall efficiency of the receiver. Recently, designs that incorporate fiber optics to collect a portion of the reflected radiation and 'recycle' it back into the etalon have been demonstrated as a method to improve the transmission efficiency of the etalon in a fringe imaging receiver.

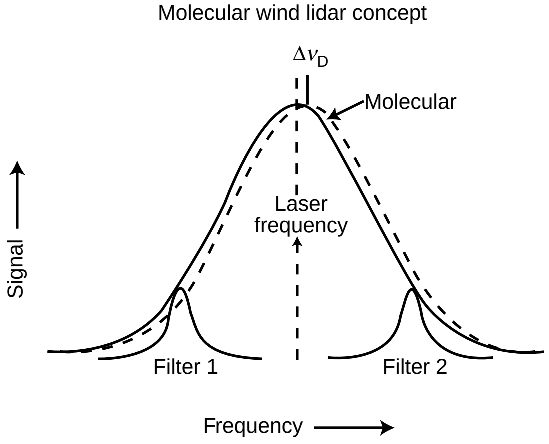

Figure 8 Spectrum of molecular lidar return showing placement of bandpass filters for a dual channel (double edge) direct detection receiver. Courtesy B. Gentry, NASA.

Doppler Wind Measurements Based on Molecular Scatter

One of the primary advantages of direct detection Doppler lidar is its capability for measurements based on scatter from atmospheric molecules. Measurement of Doppler shifts from molecular scattered radiation is challenging because of the large Doppler-broadened bandwidth of the return. Because one wants to measure a mean wind velocity with a precision of a few m s^{-1} or better, in the order of 10^5 photons are required. Some combination of multiple pulse averaging, powerful lasers, and large receiver optics is required to obtain these high photon counts from backscattered returns.

Molecular-scatter wind measurements have been demonstrated in the visible spectral region at 532 nm wavelength as well as at 355 nm in the near ultraviolet. The ultraviolet region has the dual advantages of enhanced molecular scatter and less restrictive laser eye-safety restrictions. Figure 9 shows the time series of a wind profile measured in the troposphere from both aerosols and clouds using a molecular-scatter, ground-based 355-nm wavelength Doppler fringe-imaging lidar. The figure shows measurements from receiver channels optimized for the wideband molecular signal and the narrowband aerosol return. For this measurement, backscattered photons were collected by a 0.5 m receiver aperture, averaged for 1 minute, and processed. In the absence of clouds, direct detection Doppler lidars have measured wind profiles continuously from the surface to beyond 15 km height. Figure 10 shows that the estimated wind error for the same 355 nm lidar is less than 1 m s^{-1} to about 10 km, and less than 4 m s^{-1} at 15 km height.

Heterodyne and Direct-Detection Doppler Trade-Offs

Lively debates within the lidar community have occurred over the past decade regarding the relative merits of heterodyne versus direct detection Doppler lidars. To a large extent, the instruments are complementary. Generally, heterodyne instruments are much more sensitive when significant aerosols are present. Coherent lidar processing techniques have been developed that can produce accurate wind measurement rates using only a few lidar pulses with as few as 100 detected photons, such that several wind observations per second can be obtained for nominal pulse rates. This inherent sensitivity has led to numerous applications in which a lidar beam has been scanned rapidly over a large volume to obtain time-varying, three-dimensional wind measurements.

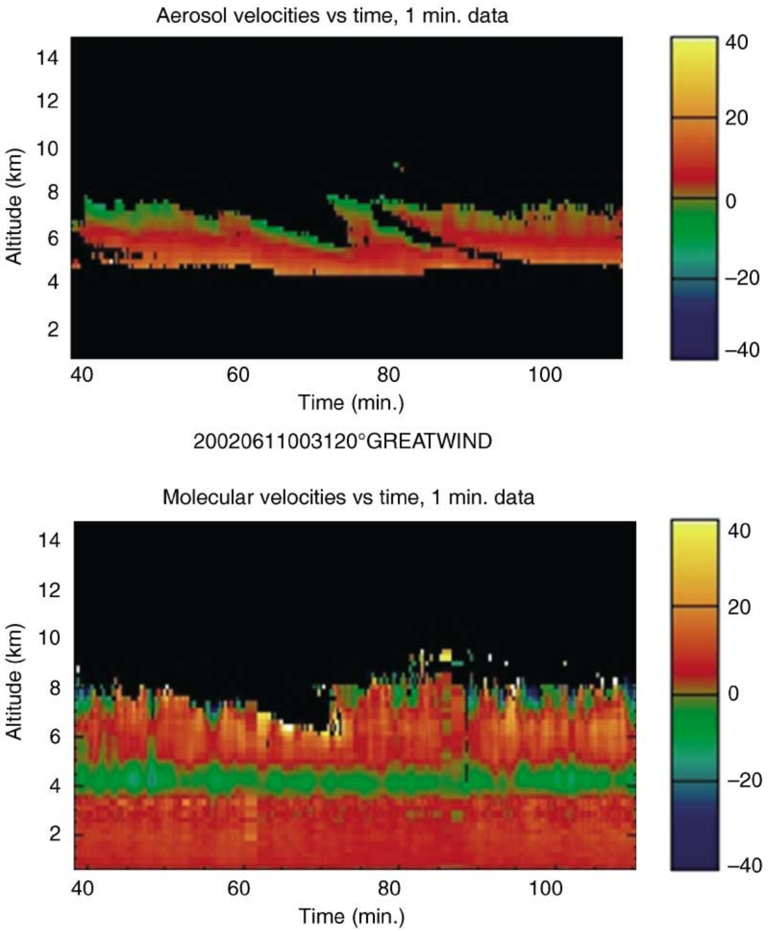

Figure 9 Time series of radial wind speed profiles measured by a 355 nm fringe-imaging direct detection lidar aerosol channel (top) and molecular channel (bottom) at Maura Loa Observatory, HI. Change in wind speed and blocking of the return by clouds is clearly seen. Courtesy C. Nardell, Michigan Aerospace Corp.

The primary advantage of direct detection instruments is their demonstrated capability to measure winds from molecular-backscattered returns in the middle and upper troposphere. In very pristine air, direct detection offers the only method for long-range wind measurements, even though significant averaging may be required. Direct-detection lidars have the additional advantage of being unaffected by atmospheric refractive turbulence. Unlike heterodyne lidars, which require a very pure laser pulse and a diffraction-limited receiver field of view matched to and precisely aligned with the transmitted beam, direct detection systems can have a wider bandwidth transmitter and a receiver field of view several times diffraction-limited. In direct detection lidar design, the field of view is usually constrained by the need to limit background light during daytime operation. A major design challenge for direct detection instruments is holding the Fabry–Perot etalon plate spacing stable over a range of temperatures and in high vibration environments.

Global Wind Measurements

A satellite-based Doppler lidar has frequently been proposed as a way to measure wind fields over most of the Earth. At present, winds are the one major meteorological variable not well-measured from orbiting platforms. Measurement of winds is especially important over regions of the Earth that are not currently well sampled, such as over Northern Hemisphere oceans, as well as over most of the tropics and Southern Hemisphere. Wind profile information is currently obtained from radiosondes and by tracking cloud and water vapor inhomogeneities using satellite imagers. Doppler lidar wind measurements would greatly augment the current data set by providing wind estimates throughout the troposphere under clear conditions, and highly height-resolved observations down to cloud tops when cloud decks are present. Observing system simulation experiments conducted in recent years indicate that satellite-based lidar global wind

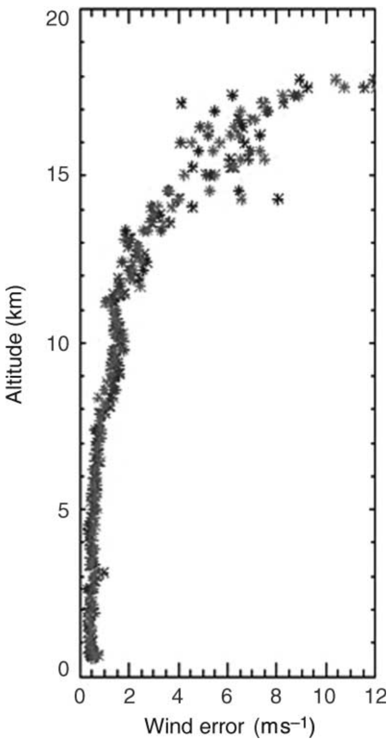

Figure 10 Estimated error in the radial velocity estimate versus altitude for a 355 nm direct detection, fringe imaging Doppler lidar system at Mauna Loa observatory on Hawaii. The lidar provides data for assimilation into local mesoscale forecast models. Courtesy C. Nardell, Michigan Aerospace Corp.

measurements could lead to a significant improvement in long-term forecast skill, provided the wind fields can be observed with sufficient accuracy and spatial resolution.

In a Doppler lidar wind mission, a satellite carrying a lidar system would orbit the Earth in a nearly polar orbit. The pulsed laser beam would be scanned conically about the nadir to obtain different components of the wind velocity. The scanning could be either continuous or 'stop and stare'. After sufficient returns are averaged at a given pointing angle to produce an acceptable estimate, the radial component of the velocity would be computed and assimilated directly into numerical analysis and forecast models.

Doppler lidar measurement of winds from space is theoretically feasible but technologically difficult. Depending on the orbital height, the scattering volume is anywhere from 450 to ~850 km from the satellite, which challenges the sensitivity of current system types. Because weight and power consumption are critical parameters for space systems, telescope diameter and laser power cannot be easily increased to obtain the necessary sensitivity. Similarly, the ability to average returns from multiple pulses is also limited by time limitations. Because a satellite moves at about 7 km s^{-1}, in order to obtain measurements over a horizontal distance of 300 km (the resolution of the radiosonde network) only about 45 seconds are available to make enough observations from multiple look angles to obtain a useful measurement. It should also be noted that, as a result of the high orbital velocity of the satellite, precise knowledge of beam pointing is extremely critical for measurements from satellites. For a lidar operating at a nadir angle of 45 degrees, an error in the knowledge of pointing angle of just 1 mrad results in an error of about 5 ms^{-1} in the measured radial component of the wind.

Despite the challenge of employing a satellite-based Doppler lidar, efforts are continuing to develop the appropriate technology and to assess the impact of the observations. The European Space Agency is planning a Doppler wind lidar demonstration mission for the late 2000s that would incorporate a nonscanning, direct-detection instrument with both aerosol and molecular channels. Doppler lidar technology research and numerical simulations aimed at satellite-based wind sensing is ongoing at several research centers within the United States, Europe, and Japan. One option currently being studied is a 'hybrid' lidar system that would combine a direct detection lidar for measurement of winds in the clear free troposphere with a low-energy coherent system that would obtain its observations from clouds and the aerosol-rich boundary. Although a hybrid instrument reduces some requirements on system power and aperture size, combining the two techniques on a single platform will likely require innovative engineering and new approaches to beam combining and scanning.

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Scattering:** Scattering Theory.

Further Reading

Baker WE, Emmitt GD, Robertson F, *et al.* (1998) Lidar-measured wind from space: a key component for weather and climate prediction. *Bulletin of the American Meteorological Society* 79: 581–599.

Banta RM, Olivier LD and Levinson DH (1993) Evolution of the Monterey Bay sea breeze layer as observed by pulsed Doppler lidar. *Journal of Atmospheric Science* 50 (24): 3959–3982.

Banta RM, Olivier LD, Gudiksen PH and Lange R (1996) Implications of small-scale flow features to modeling dispersion over complex terrain. *Journal of Applied Meteorology* 35: 330–342.

Banta RM, Darby LS, Kaufmann P, Levinson DH and Zhu C-J (1999) Wind flow patterns in the Grand Canyon as revealed by Doppler lidar. *Journal of Applied Meteorology* 38: 1069–1083.

Frehlich R (1996) Coherent Doppler lidar measurements of winds in the weak signal regime, 1997. *Applied Optics* 36: 3491–3499.

Gentry B, Chen H and Li SX (2000) Wind measurements with a 355 nm molecular Doppler lidar. *Optics Letters* 25: 1231–1233.

Grund CJ, Banta RM, George JL, *et al.* (2001) High-resolution Doppler lidar for boundary-layer and cloud research. *Journal of Atmosphere and Oceanic Technology* 18: 376–393.

Huffaker RM and Hardesty RM (1996) Remote sensing of atmospheric wind velocities using solid state and CO_2 coherent laser systems. *Proceedings of the IEEE* 84: 181–204.

Menzies RT and Hardesty RM (1989) Coherent Doppler lidar for measurements of wind fields. *Proceedings of the IEEE* 77: 449–462.

Post MJ and Cupp RE (1990) Optimizing a pulsed Doppler lidar. *Applied Optics* 29: 4145–4158.

Rees D and McDermid IS (1990) Doppler lidar atmospheric wind sensor: reevaluation of a 355 nm incoherent Doppler lidar. *Applied Optics* 29: 4133–4144.

Rothermel J, Cutten DR, Hardesty RM, *et al.* (1998) The multi-center airborne coherent lidar atmospheric wind sensor. *Bulletin of the American Meteorological Society* 79: 581–599.

Skinner WR and Hays PB (1994) Incoherent Doppler lidar for measurement of atmospheric winds. *Proceedings of SPIE* 2216: 383–394.

Souprayen C, Garnier A and Hertzog A (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. II: Mie scattering effect, theory, and calibration. *Applied Optics* 38: 2422–2431.

Souprayen C, Garnier A, Hertzog A, Hauchecorne A and Porteneuve J (1999) Rayleigh–Mie Doppler wind lidar for atmospheric measurements. I: Instrumental setup, validation, first climatological results. *Applied Optics* 38: 2410–2421.

Hyperspectral Remote Sensing of Land and the Atmosphere

W H Farrand, Space Science Institute, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral remote sensing is a true marriage of imaging technology with spectroscopy. Hyperspectral remote sensing systems (also known as imaging spectrometer systems) fully sample the optical wavelength range of interest, whether it be the reflected solar range (0.35 to 2.5 μm) or the range of thermal emission from the Earth's surface (3.0 to 14 μm). A hyperspectral sensor views the Earth's surface in a series of contiguous and spectrally narrow image bands. Figure 1 presents the concept of hyperspectral remote sensing, wherein each spatial element of an image has an associated full resolution spectrum. The calibrated reflectance or emittance spectra collected by a hyperspectral system are meant to be directly comparable with that of materials measured in the laboratory. With such high spectral resolution, it thus becomes possible to do reflectance or emittance spectroscopy of the Earth's surface from an overhead perspective.

Underlying Principles

Optical remote sensing is done over wavelength intervals, or windows, in which the atmosphere is largely transparent. For the reflective solar portion of the spectrum, the atmosphere is mostly transparent from approximately 0.35 to 2.5 μm. Water, CO_2, and other gases have absorptions of varying strength in this range. This window is subdivided into the range of human vision (the visible) and the region of slightly longer wavelengths known as the near infrared; together they are the visible/near infrared or VNIR and extend from approximately 0.35 to 1.0 μm. The region from approximately 1.0 to 2.5 μm is known as the short wavelength infrared or SWIR. In the range of emitted terrestrial radiation there are two window regions. The first extends from approximately 3 to 5 μm and is known as the medium wavelength infrared or MWIR, and the second, the long wavelength infrared or LWIR, extends from approximately 8 to 14 μm.

The light that is intercepted by the entrance aperture of a sensor is known by the quantity of radiance which is measured in units of microwatts per square centimeter per nanometer per unit of solid angle. Physical quantities related to the material properties of Earth surface materials are reflectance and emittance. Both properties are the result of ratios and so are unitless. One definition of reflectance is that it is the ratio of radiance reflected from a material, divided by the radiance reflected from an identically illuminated perfectly diffuse reflector. Likewise, a definition for emittance is the ratio of radiance emitted from a material, divided by the ratio

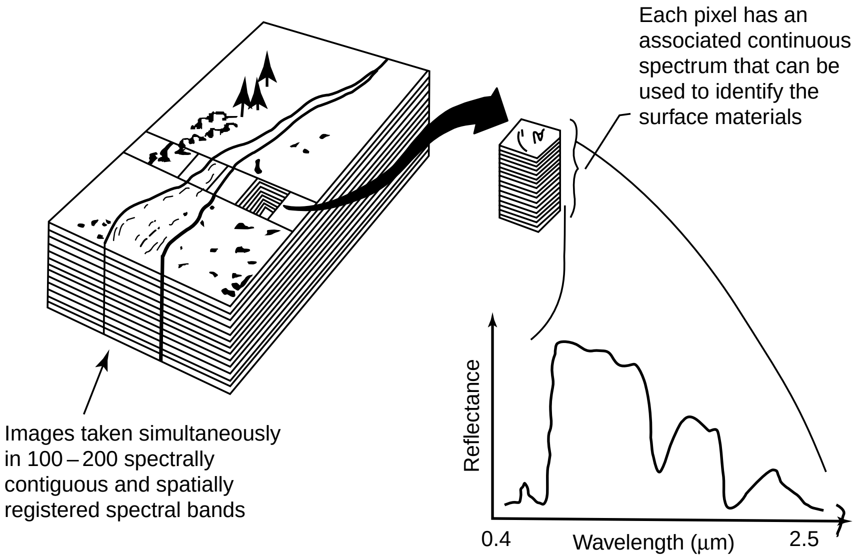
Each pixel has an associated continuous spectrum that can be used to identify the surface materials
Images taken simultaneously in 100 – 200 spectrally contiguous and spatially registered spectral bands
Reflectance
0.4
Wavelength (µm)
2.5

Figure 1 Hyperspectral remote sensing concept.

of radiance emitted from a perfectly emitting material of the same temperature.

Materials covering the Earth's surface, or gases in the atmosphere, can be identified in hyperspectral data on the basis of absorption features, or bands, in the spectra recorded by the sensor. Absorption bands result from the preferential absorption of energy in some wavelength interval. The principal types of processes by which light can be absorbed include, in order of decreasing energy of the process, electronic charge transfers, electronic crystal field effects, and molecular vibrations. The former two processes are observed in minerals and manmade materials that contain transition group metal cations, most usually, iron. Charge transfer absorptions are the result of cation-to-anion or cation-to-cation electron transfers. Crystal field absorptions are explainable by the quantum mechanical theory of atomic structure, wherein an atom's electrons are contained in orbital shells. The transition group metals have incompletely filled *d* orbital shells. When a transition metal cation, such as iron, is surrounded by anions, and potentially selects other cations, an electric field, known as the crystal field, exists. Crystal field absorptions occur when radiant energy causes that cation's orbital shell energy levels to be split by interaction with the crystal field. The reflectance spectra of different iron-bearing minerals are unique, because the iron cation has a unique positioning with respect to the anions and other cations that compose the mineral.

A material's component molecules have bonds to other molecules and as a result of interaction with radiant energy, these bonds can stretch or bend. These molecular vibrations, and overtones of those vibrations, are less energetic processes than the electronic absorption processes described above and so the resulting absorption features occur at longer wavelengths. Fundamental vibrational features occur in the MWIR and LWIR. Overtones and combination overtones of these vibrations are manifested as weaker absorption features in the SWIR. These overtone features in the SWIR include absorptions diagnostic of carbonate and certain clay minerals. Absorption features caused by vibrational overtones of C–H, O–H, and N–H stretches are also manifested in the SWIR and these absorption features are characteristic of important vegetative biochemical components such as cellulose, lignin, starch, and glucose. An example of the effect of the changes in reflected energy recorded at a sensor, due to the absorption of incident energy by green vegetation and other surface materials, is provided in Figure 2.

Laboratory spectrometers have been used for many years to analyze the diagnostic absorptions of materials caused by the above effects. More recently, spectrometers were mounted on telescopes to determine the composition of the Moon and the other solid bodies in our solar system. Technology progressed to where profiling spectrometers (instruments that measure a successive line of points on the surface) could be mounted on aircraft. The next logical step was the construction of imaging spectrometers (hyperspectral sensors) that measured spectra in two spatial dimensions. In fact, the data produced by a hyperspectral sensor are often thought of as an image cube (Figure 3) because they consist of three dimensions: two spatial and one spectral.

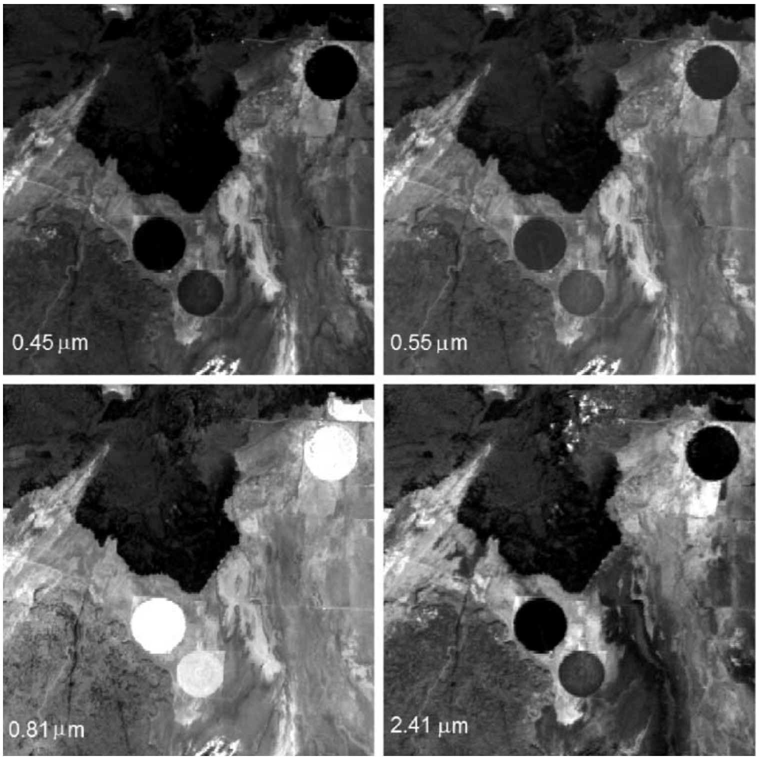

Figure 2 Portion of a scene from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) from Utah that includes a basalt flow, circularly irrigated grasses sparsely vegetated plains, and damp alluvial sediments. AVIRIS channels centered at 0.45, 0.55, 0.81, and 2.41 μm are shown. Green vegetation is brightest in the 0.81 μm band which samples the peak of the near infrared plateau (see **Figure 7**) just beyond the 'red edge'. The green vegetation and damp sediments are darkest in the 2.41 μm band where water has a very low reflectance (note also **Figure 8**). The bright patches in the basalt flow in the 2.41 μm image are occurrences of oxidized red cinders which have a high reflectance in the SWIR.

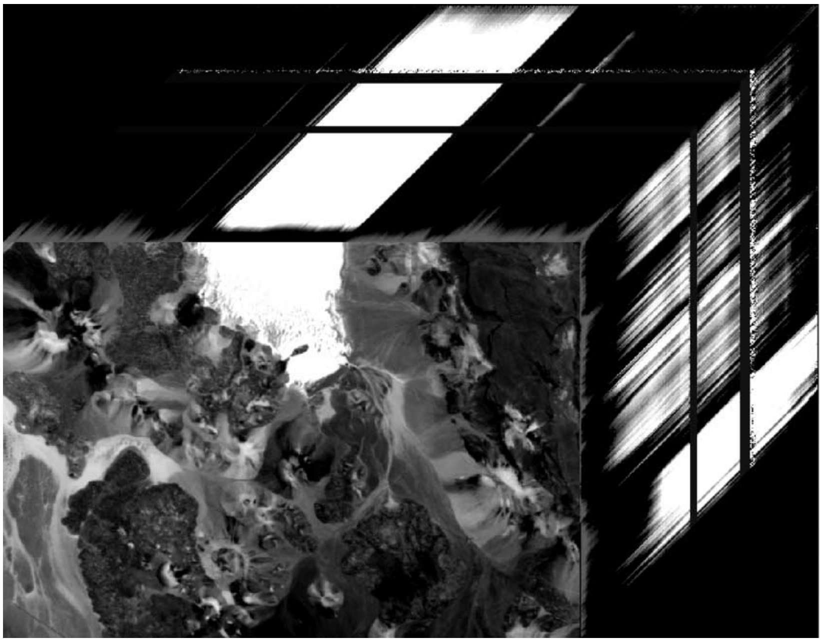

Figure 3 Visualization of a hyperspectral image cube. A 1.7 μm image band from the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over the Lunar Crater Volcanic Field, Nevada is shown with each of the 224 rightmost columns and 224 topmost lines arrayed behind it. In the lower right-hand corner, the low reflectance of ferric oxide-rich cinders in the blue and green show up as dark in the stacking of columns with the high reflectance in the infrared showing up as bright.

Advances in hyperspectral sensor technology have been accompanied by advances in associated technologies, such as computing power and the evolution of data processing algorithms.

Data Processing Approaches

An early impediment to the widespread use of hyperspectral data was the sheer volume of the data sets. For instance, a single scene from NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) is 614 samples by 512 lines by 224 bands. With its short integer storage format, such a scene takes up 140.8 Mbytes. However, since the dawn of the hyperspectral era, in the 1980s, computing power has expanded immensely and computing tasks, which once seemed prohibitive, are now relatively effortless. Also, numerous processing algorithms have been developed which are especially well suited for use with hyperspectral data.

Spectral Matching

In one class of algorithms, the spectrum associated with each spatial sample is compared against one or a group of model spectra and some similarity metric is applied. The model spectrum can be measured by a laboratory or field portable spectrometer of pure materials or can be a single pixel spectrum, or average of pixel spectra, covering a known occurrence of the material(s) to be mapped. A widely used similarity metric is the Spectral Angle Mapper (SAM). SAM is the angle, θ, obtained from the dot product of the image pixel spectrum, t, and the library spectrum, r, as expressed by

$$\theta = \cos^{-1}\left(\frac{t * r}{\|t\| * \|r\|}\right) \qquad [1]$$

SAM is insensitive to brightness variations and determines similarity based solely on spectral shape. Lower θ values mean more similar spectra.

In a related approach, matching is done, not on the entire spectrum, but rather on characteristic absorption features of the materials of interest. In Figure 4, the geometry of an absorption band is plotted. To map a given mineral, with a specific absorption feature, a high resolution laboratory spectrum is resampled to the spectral resolution of the hyperspectral sensor. The spectrum is subsampled to match the specific absorption feature (i.e., only bands from the short wavelength shoulder of the absorption to the long wavelength shoulder are included). A straight line continuum (calculated, based on a line between the two-band shoulders) is divided out from both the laboratory and the sensor spectra. Contrast between the library and the sensor spectra is mitigated through the use of an additive constant. This constant is incorporated into a set of equations which are solved through the use of standard least squares. The result of such a band-fitting algorithm is two data numbers per spatial sample of the hyperspectral scene. First the band depth (or, alternatively, band area) of the feature is calculated, and second, a goodness-of-fit parameter is calculated. These two parameters are most often combined in a band-depth times goodness-of-fit image. More sophisticated versions of the band-mapping algorithm target multiple absorption features and are, essentially, expert systems.

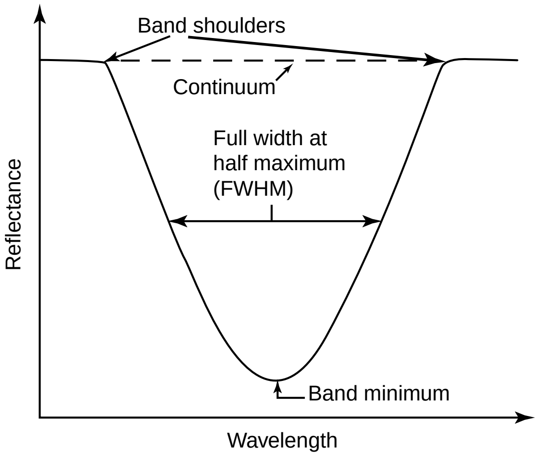

Figure 4 Parameters that describe an absorption band.

Spectral Mixture Analysis

A fundamentally different methodology for analyzing hyperspectral data sets is to model the measured spectrum of each spatial sample as a linear combination of endmember spectra. This methodology is based on the fact that the ground area corresponding to any given spatial sample likely will be covered by more than one material, with the consequence that the measured reflectance or emittance spectrum is a mixed pixel spectrum. The objective of linear SMA is to try to determine the fractional abundance of the component materials, or endmembers. The basis of linear spectral unmixing is to model the response of each spatial element as a linear combination of endmember spectra. The basis equation for linear SMA is

$$r(x, y) = \alpha\mathbf{M} + \varepsilon \qquad [2]$$

where $r(x, y)$ = the relative reflectance spectrum for the pixel at position (x, y), α = the vector of endmember abundances, $\mathbf{M}$ = the matrix of endmember spectra, and ε = the vector of residuals between the modeled and the measured reflectances. Application of SMA results in a series of fraction images for each endmember wherein the data numbers range, ideally, between 0 and 1. Fraction image pixels with a digital number (DN) of 0 are taken to be devoid of the endmember material. Fraction image pixels with a DN of 1.0 are taken to be completely covered with the endmember material.

Techniques related to SMA have been developed which also map the abundance of a target material on a per pixel basis. These include implementations highly similar to SMA such as Orthogonal Subspace Projection (OSP) wherein, as with SMA, all endmembers must be determined a priori. Other approaches such as Constrained Energy Minimization (sometimes referred to as ‘Matched Filter’) do not require a priori knowledge of the endmembers. Instead, only the reflectance or emittance spectrum of the target need be known and the undesired spectral background is estimated, and ultimately compensated for, using the principal eigenvectors of the sample correlation or covariance matrix of the hyperspectral scene.

Expert Systems and Artificial Neural Networks

Another means of extracting information from hyperspectral data sets is to use computer approaches that in some way mimic the human thought process. This can take the form of an expert system which applies a set of rules or tests as each pixel spectrum is successively analyzed. For example, if a spectrum has an absorption at 2.2 μm it could tentatively be classified as being caused by a clay mineral. Additionally, if the absorption is a doublet, the specific clay mineral is likely a kaolinite. In practice, expert system codes are lengthy and complex although excellent results can be demonstrated from their use, albeit at the expense of processing time. Another approach in this vein is the use of an artificial neural network (ANN) for data processing. The use of ANNs is motivated by their power in pattern recognition. ANN architectures are well suited for a parallel processing approach and thus have the potential for rapid data processing.

Hyperspectral Remote Sensing of the Land

Geology

Early efforts with airborne hyperspectral sensors focused on geologic remote sensing. This was, at least in part, a consequence of the mineralogic diversity of the Earth’s surface and the fact that many minerals have absorption features which are unique and diagnostic of the mineral’s identity. The expectation, which was soon borne out by results from these early sensors, was that, given exposure of the surface, mineralogic maps as detailed as the spatial resolution of the sensor could be derived from hyperspectral data. As some of the processing techniques discussed above have become more readily available, it has become possible to produce mineralogic maps from hyperspectral data far more rapidly and cost-effectively than geologic maps produced by standard means (e.g., by walking over the ground and manually noting which lithologies are underfoot).

The propensity for transition metals, most especially iron, to absorb energy through charge transfer and crystal field effects, is noted above. Fe–O and Fe^{2+}–Fe^{3+} charge transfers cause a profound absorption in the reflectance spectrum of Fe-bearing minerals shortwards of 0.4 μm. The wing of this absorption causes the low blue and green reflectance of Fe-bearing minerals. Crystal field bands cause an absorption feature in the 0.9 to 1.0 μm region. The ability to discriminate subtle differences among Fe-bearing minerals in hyperspectral data has proven extremely valuable in applications such as mineral exploration, volcanology, and in the characterization of abandoned mine lands.

In the SWIR, absorptions caused by vibrational overtones of molecular bonds within minerals. These include absorptions in the 2.2 and 2.3 μm region which are characteristic of many Al- and Mg-bearing clay minerals. Absorptions in the SWIR are narrower in width than the Fe-generated crystal field bands discussed previously. In fact, the requirement to efficiently resolve these vibrational overtone absorption bands (which have full width at half maximum (FWHM) bandwidths of 20 to 40 nm) helped to drive the selection of the nominal 10 nm bandwidth of early and most current hyperspectral sensors. Certain clay minerals can be indicators of hydrothermal activity associated with economic mineral deposits; thus, the SWIR is an important spectral region for mineral exploration. Clay minerals, by definition, have planar molecular structures that are prone to failure if subjected to shearing stresses. The ability to uniquely identify and map these minerals, using hyperspectral data, has thus been used to good effect to identify areas on volcanoes and other hydrothermally altered terrains that could be subject to landslides.

Minerals with adsorbed or molecularly bound OH^- and/or water have vibrational overtone absorptions near 1.45 and 1.9 μm, although these features are masked in remotely sensed data by prominent atmospheric water vapor bands at 1.38 and 1.88 μm. The reflectance of water and OH-bearing minerals decrease to near zero at wavelengths longwards of

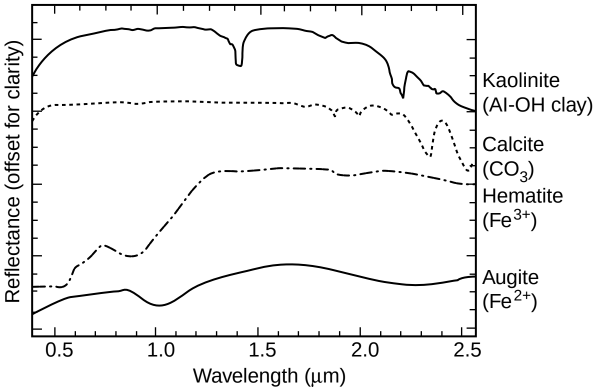

Figure 5 Reflectance spectra of representative minerals. Kaolinite represents clay minerals with its diagnostic 2.2 μm band caused by an overtone vibration of the Al–OH bond. The 2.335 μm band of calcite is caused by an overtone vibration of the carbonate molecule. The features shortwards of 1.0 μm in the hematite spectrum are caused by crystal field and charge transfer absorptions due to Fe^{3+}. The bands near 1.0 and longwards of 2.0 μm in the augite spectrum are the result of Fe^{2+} crystal field bands.

2.5 μm, due to the presence of water and OH fundamental vibrations near 3.0 μm. Minerals bearing the carbonate (CO_3^{-2}) molecule are abundant on the Earth's surface due to the ubiquitous occurrences of limestone and dolomite. Calcite and dolomite have overtone absorptions centered at 2.335 and 2.315 μm, respectively. A good test for the spectral resolution and spectral calibration of a hyperspectral sensor is its ability to successfully discriminate calcite from dolomite on the basis of the aforementioned absorption features. Figure 5 shows representative spectra of carbonate, clay, and ferric and ferrous iron-bearing minerals.

Absorptions resulting from fundamental molecular absorptions are manifested in the LWIR. These absorptions are of great geologic interest because they include absorptions resulting from vibrations of the Si–O bond and silicate minerals form the bulk of the Earth's crust. The wavelength at which the Si–O stretching feature occurs is dictated by the level of polymerization (or molecule-to-molecule bonding) of the silicate mineral. Highly polymerized framework silicate minerals, such as quartz and feldspar, have a shorter wavelength absorption than do silicate minerals, such as olivine which are composed of disconnected SiO_4 molecules. In Figure 6, laboratory emission spectra of highly and poorly polymerized silicate minerals are shown.

Vegetation and the Environment

While different minerals are, by definition, generally composed of diverse component molecules, different species of vegetation represent variations of the same general biochemical constituents (e.g., chlorophyll, proteins, lignin, cellulose, sugar, starch, etc.). In the thermal IR, vegetation has a generally flat emittance, thus it is in the reflective solar spectrum that most vegetative remote sensing is performed. The major features in the reflectance spectrum of green vegetation are shown in Figure 7. In the visible, the major absorption features are caused by the presence of chlorophyll. Chlorophyll has strong absorptions in the blue and the red, leaving a reflectance maximum (the green peak) at 0.55 μm. In the NIR, scattering in the leaf structure causes high reflectance leaving a profound absorption edge (the red edge) between

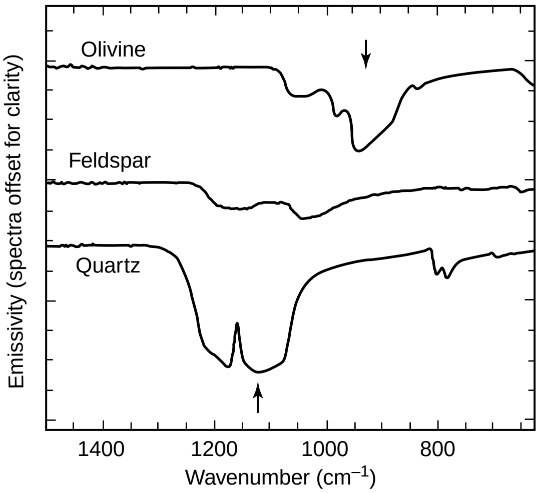

Figure 6 Emissivity spectra of SiO_4-bearing minerals showing the shift in band minimum to higher wave number with increasing polymerization.

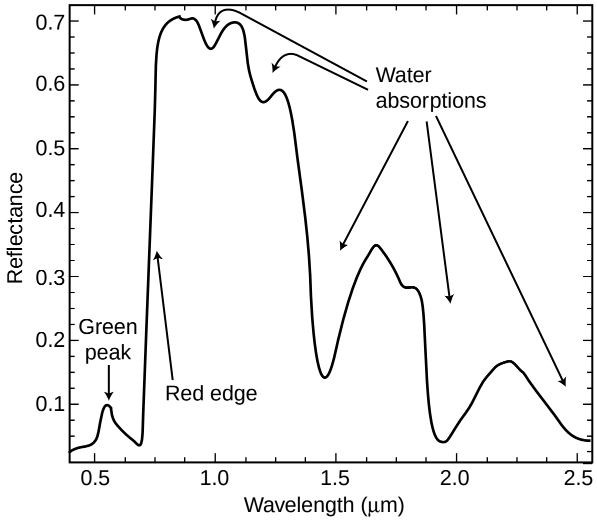

Figure 7 Major spectral features in the reflectance spectrum of green vegetation.

0.7 and 0.78 μm. Past the red edge and into the SWIR, the spectrum of green vegetation is dominated by water in the leaf, with leaf water absorptions occurring at 0.97, 1.19, 1.45, 1.93, and 2.50 μm. Studies of vegetation, using multispectral systems, have made use of a number of broadband indices. Hyperspectral systems allow for the discrimination of individual absorption features and subtle spectral shape differences in vegetation spectra which are unresolvable using broadband multispectral systems. For example, the absorptions caused by chlorophyll can be used for studies of chlorophyll content in vegetative canopies. The unique interaction of chlorophyll with leaf structures can provide a morphology to the green peak that is unique to a given species and thus is mappable using a spectral feature fitting approach.

A desired goal in the study of vegetation and ecosystem processes is to be able to monitor the chemistry of forest canopies. The foliar biochemical constituents making up forest canopies have associated absorption features in the SWIR, that result from vibrational overtones and combinations of C–O, O–H, C–H, and N–H molecular bonds. However the absorptions from these biochemical components overlap so that the chemical abundance of any one plant component cannot be directly related to any one absorption feature. An even more serious challenge is introduced by the strong influence of water in the SWIR in green vegetation spectra. Water constitutes 40 to 80% of the weight of leaves. However, the unique characteristics of high-quality hyperspectral data (e.g., excellent radiometric and spectral calibration, high signal-to-noise ratio, high spectral resolution) have been used to detect even these subtle features imprinted on the stronger water features.

The amount of water present in plant leaves is, in itself, a valuable piece of information. Hyperspectral data can be used to determine the equivalent water thickness present in vegetative canopies. Conversely, the amount of dry plant litter and/or loose wood can be estimated using spectral mixture analysis and related techniques. Taken together, all these data: vegetation species maps, determinations of leaf water content and the relative fractions of live vegetation versus litter, can be used to characterize woodlands for forest fire potential. The ability to map different species of vegetation has also proved useful in the field of agriculture, for assessing the impact of invasive weed species on farm and ranch land. Being able to assess the health of crops, in terms of leaf water content, is also important for agriculture.

Another subtle vegetative spectral feature, that provides information on plant vitality, is the aforementioned red edge. Shifts in the position of the red edge have been linked to vegetation stress. Studies have shown that vegetation stress and consequent shifts in the position of the red edge can be caused by a number of factors, including insufficient water intake or the intake of pernicious trace metals. Vegetation growing over mineral deposits can display red edge shifts and this can be used to assist in mineral exploration efforts.

Another application area, in which hyperspectral remote sensing has found great utility, is in the analysis of snow-covered areas. Over the reflective solar spectrum, the reflectance of snow varies from values near zero in the SWIR to values near one in the blue. Differences in the grain size of snow result in differences in reflectance as is illustrated in Figure 8. The spectral variability of snow makes it easily distinguishable from other Earth surface materials.

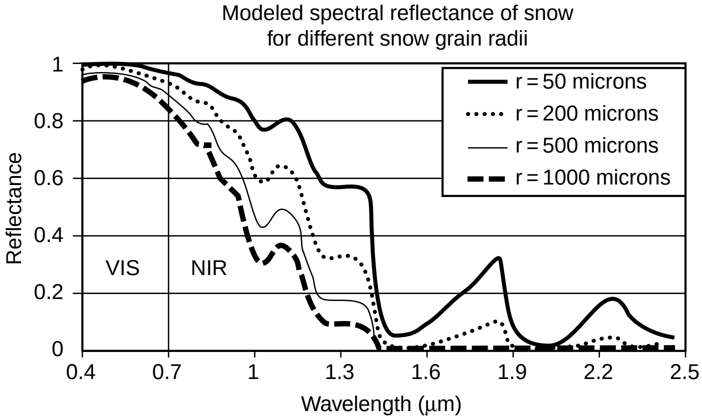

Figure 8 Differences in snow reflectance spectra resulting from differences in snow grain size. Courtesy of Dr Anne Nolin of Oregon State University.

Linear spectral unmixing, using hyperspectral data, has been shown to successfully discriminate between snow, vegetation, rock, and clouds and to accurately map snow cover fraction in mixed pixels. Accurate maps of snow cover, types of snow, and fractions of liquid water admixed with snow grains, are required for forecasting snowmelt runoff and stream discharge in watersheds dominated by snow cover.

Hyperspectral Remote Sensing of the Atmosphere

The physical quantity most often sought in land remote sensing studies is surface reflectance. However, the quantity recorded by a hyperspectral sensor is radiance at the entrance aperture of the sensor. In order to obtain reflectance, the background energy level of the Sun and/or Earth must be removed and the scattering and absorbing effects of the atmosphere must be compensated for. In the VNIR through SWIR, there are seven atmospheric constituents with significant absorption features: water vapor, carbon dioxide, ozone, nitrous oxide, carbon monoxide, methane, and oxygen. Molecular scattering (commonly called Rayleigh scattering) is strong in the blue but decreases rapidly with increasing wavelength. Above 1 μm, its effect is negligible. Scattering caused by atmospheric aerosols, or Mie scattering, also is more prominent at shorter wavelengths and decreases with increasing wavelength, but the dropoff of effects from Mie scattering is not as profound as that for Rayleigh scattering. Consequently, effects from aerosol scattering can persist into the SWIR.

There are three primary categories of methods for removing the effects of atmosphere and solar insolation from hyperspectral imagery, in order to derive surface reflectance. Methods of atmospheric corrections can be considered as being either an image-based, empirical, or model-based approach. An image-based, or 'in-scene' approach uses only data measured by the instrument. Empirical methods make use of the remotely sensed data in combination with field measurements of reflectance, to solve a simplified equation of at-sensor radiance such as eqn [3].

$$L_s = A\rho + B \qquad [3]$$

where L_s is the at-sensor radiance, ρ is the reflectance of the surface, and A and B are quantities that incorporate, respectively, all multiplicative and additive contributions to the at-sensor radiance. All the quantities expressed in eqn [3] can be considered as varying as a function of wavelength, λ. Approximations for A and B from eqn [3] constitute a set of gains and offsets derived from the empirical approach. In some empirical approaches, atmospheric path radiance is ignored and only a multiplicative correction is applied.

Model-based approaches seek to model what the at-sensor radiance should be on a pixel-by-pixel basis, including the contribution of the atmosphere. The at-sensor radiance, L_s, at any wavelength, λ, can be expressed as

$$L_s = \frac{1}{\pi}(E\rho + M_T)\tau_\theta + L_u \qquad [4]$$

where E is the irradiance at the surface of the Earth, ρ is the reflectance of the Earth's surface, M_T is the spectral radiant exitance of the surface at temperature, T, τ_θ is the transmissivity of the atmosphere at zenith angle θ, and L_u is the spectral upwelling path radiance of the atmosphere. The ability to solve eqn [4] and do atmospheric correction on a pixel-by-pixel basis is appealing, in that it negates the shortcomings of an empirical approach where a correction based on calibration targets in one part of a scene might not be appropriate for another part of the scene, due to differences in atmospheric path-length or simple atmospheric heterogeneity. Model-based approaches are also able to take advantage of the greater spectral dimensionality of hyperspectral data sets for the derivation of atmospheric properties (e.g., the amount of column water vapor, CO_2 band depths, etc.) directly from the data.

The main atmospheric component, affecting imagery in the reflective solar portion of the spectrum, is water vapor. Model-based atmospheric correction techniques determine the column water vapor for each pixel in the scene, based on the depth of the 940 and/or 1130 nm atmospheric water bands. Thus for each pixel, an appropriate amount of water can be removed. The maps of atmospheric water vapor distribution are themselves of interest to atmospheric scientists. Absorption features caused by well-mixed gases, such as the 760 nm O_2 band, can be used by model-based programs, to produce images of approximate scene topography. More advanced model-based approaches also solve for the scattering effects of aerosols on the hyperspectral data.

As noted above, atmospheric constituents, including industrial effluents, can be detected and spatially mapped by hyperspectral sensors. The fundamental vibrational absorption of gases of interest occurs in the MWIR and the LWIR. Overtones of these gaseous molecular vibrations occur in the VNIR to SWIR, but are generally too weak to be detected. The ability to detect anthropomorphically produced gases depends on a number of factors, including the temperature difference between the gas cloud and the background atmosphere, the concentration of gas within the cloud, and the size of the cloud (e.g., the path length of light through the cloud).

See also

Imaging: Infrared Imaging. **Instrumentation:** Spectrometers.

Further Reading

Clark RN (1999) Spectroscopy of rocks and minerals, and principles of spectroscopy. In: Rencz AN (ed.) *Remote Sensing for the Earth Sciences*, pp. 3–57. New York: John Wiley and Sons.

Farrand WH and Harsanyi JC (1997) Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho through the use of a Constrained Energy Minimization technique. *Remote Sensing Environment* 59: 64–76.

Gao B-C and Goetz AFH (1995) Retrieval of equivalent water thickness and information related to biochemical components of vegetation canopies from AVIRIS data. *Remote Sensing Environment* 52: 155–162.

Gao B-C, Heidebrecht KH and Goetz AFH (1993) Derivation of scaled surface reflectances from AVIRIS data. *Remote Sensing Environment* 44: 165–178.

Goetz AFH (1989) Spectral remote sensing in geology. In: Asrar G (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 491–526. New York: John Wiley & Sons.

Goetz AFH, Vane G, Solomon JE and Rock BN (1985) Imaging spectrometry for Earth remote sensing. *Science* 228: 1147–1153.

Hapke B (1993) *Theory of Reflectance and Emittance Spectroscopy*. Cambridge, UK: Cambridge University Press.

Kirkland LE, Herr KC and Salisbury JW (2001) Thermal infrared spectral band detection limits for unidentified surface materials. *Applied Optics* 40: 4852–4862.

Martin ME, Newman SD, Aber JD and Congalton RG (1998) Determining forest species composition using high spectral resolution remote sensing data. *Remote Sensing Environment* 65: 249–254.

Painter TH, Roberts DA, Green RO and Dozier J (1998) The effect of grain size on spectral mixture analysis of snow-covered area with AVIRIS data. *Remote Sensing Environment* 65: 320–332.

Pieters CM and Englert PAJ (eds) *Remote Geochemical Analysis: Elemental and Mineralogical Composition*. Cambridge, UK: Cambridge University Press.

Pinzon JE, Ustin SL, Castaneda CM and Smith MO (1997) Investigation of leaf biochemistry by hierarchical foreground/background analysis. *IEEE Transactions on Geoscience and Remote Sensing* 36: 1–15.

Roberts DA, Green RO and Adams JB (1997) Temporal and spatial patterns in vegetation and atmospheric properties using AVIRIS. *Remote Sensing Environment* 62: 223–240.

Wessman CA, Aber JD and Peterson DL (1989) An evaluation of imaging spectrometry for estimating forest canopy chemistry. *International Journal of Remote Sensing* 10: 1293–1316.

Laser Detection of Atmospheric Gases

E V Browell, W B Grant and S Ismail, NASA Langley Research Center, Hampton, VA, USA

Published by Elsevier Ltd.

Introduction

Lidar (light detection and ranging) systems are able to measure profiles of atmospheric aerosols, clouds, and gases by transmitting a pulsed laser beam into the atmosphere and collecting the backscattered radiation from aerosols and molecules in the atmosphere with a receiver located near the transmitter. The differential absorption lidar (DIAL) approach is the most widely used technique for measuring a variety of gases. In this approach, two closely spaced laser wavelengths are used: one which is absorbed by the gas of interest, and the other which is only weakly, or not at all, absorbed by the gas. A differential with respect to range and wavelength is calculated, to determine the average gas concentration along any segment of the lidar path, using the familiar Beer–Lambert law for an absorbing medium.

The DIAL equation can be expressed in its simple form as

$$N = \left(1/(2(R_2 - R_1)(\sigma_{\mathrm{on}} - \sigma_{\mathrm{off}}))\right) \times \ln\left(P_{\mathrm{off}}(R_2)P_{\mathrm{on}}(R_1)\right)/\left(P_{\mathrm{off}}(R_1)P_{\mathrm{on}}(R_2)\right) \quad [1]$$

where N is the average gas concentration, R is the range, σ_{on} and σ_{off} are the absorption cross sections at the on- and off-line wavelengths, and $P_{\mathrm{on}}(R)$ and $P_{\mathrm{off}}(R)$ are the powers received at the on- and off-line wavelengths. However, since it is performed in the atmosphere and not a laboratory, great care must be taken to ensure that the data are analyzed properly to minimize random and systematic errors. Random errors arise from noise in the backscattered signal, the solar background signal, and the inherent detector noise, and this type of error can be reduced by signal averaging. Systematic errors arise from uncompensated instrument and atmospheric effects and must be carefully considered in system design and operation and data processing.

Another approach, the Raman lidar approach, uses inelastic scattering (scattered/emitted light has different wavelength than the illuminating light) from gases where the wavelength shift corresponds to vibrational or rotational energy levels of the molecules. Any illuminating laser wavelength can be used, but since the Raman scattering cross-section varies as λ^{-4}, where λ is the wavelength, the shorter laser wavelengths, such as in the near UV spectral region, are preferred. While even shorter wavelengths in the solar blind region below ~300 nm would permit operation in daytime, the atmospheric attenuation, due to Rayleigh scattering and UV molecular absorption, limits the measurement range. High-power lasers are used due to the low Raman scattering cross-sections. The Raman lidar measurements have to be carefully calibrated against a profile measured using another approach, such as is done for water vapor (H_2O) measurements with the launch of hygrometers on radiosondes, since the lidar system constants and atmospheric extinctions are not usually well known or modeled. In order to obtain mixing ratios of H_2O with respect to atmospheric density, the H_2O signals are ratioed to the nitrogen Raman signals. However, care must be exercised in processing these data due to the spectral dependences of atmospheric extinction.

Surface-Based Lidar Systems

Surface-based lidar systems measure the temporal evolution of atmospheric profiles of aerosols and gases. The first lidar systems used to remotely measure atmospheric gases were Raman lidar systems. The Raman lidar systems were thought to be very promising since one high-power laser could be used to measure a variety of gases. They are easier to develop and operate than DIAL systems, because the laser does not have to be tuned to a particular gas absorption feature. However, the Raman scattering cross-section is low, and the weak, inelastically scattered signal is easily contaminated by daylight background radiation, resulting in greatly reduced performance during daytime. Raman lidar systems have been developed primarily for measuring H_2O and retrieving atmospheric temperature.

Surface-based UV DIAL systems, that are part of the Network for the Detection of Stratospheric Change (NDSC) have made important contributions to the understanding of stratospheric O_3. There are DIAL systems at many locations around the Earth, with sites in Ny-Ålesund, Spitzbergen (78.9°N, 11.9°E), Observatoire Haute Provence, France (43.9°N, 5.7°E), Table Mountain, California (34.4°N, 118.2°W), Mauna Loa, Hawaii (19.5°N, 155.6°W), and Lauder, New Zealand (45.0°S, 169.7°E). First established in the 1980s, these DIAL systems are strategically located so that O_3 in different latitude bands can be monitored for signs of change. In addition, they can provide some profiles for comparison with space-based O_3 measuring instruments. The parameters of a typical ground-based UV DIAL system used in the NDSC are given in Table 1.

Airborne DIAL Systems

Airborne lidar systems expand the range of atmospheric studies beyond those possible by surface-based lidar systems, by virtue of being located in aircraft that can be flown to high altitudes and to remote locations. Thus, they permit measurements at locations inaccessible to surface-based lidar systems. In addition, they can make measurements of large atmospheric regions in times that are short compared with atmospheric motions, so that the large-scale patterns are discernible. Another advantage of airborne lidar operation is that lidar systems perform well in the nadir (down) direction since the atmospheric density and aerosol loading generally increases with decreasing altitude towards the surface, which helps to compensate for the R^{-2} decrease in the lidar

Table 1 Parameters for the Jet Propulsion Laboratory's Mauna Loa DIAL systems for stratospheric O_3 measurements

L [illegible]	XeCl	Nd:YAG (3^{rd} Harmon.)
Wavelength (nm)	308 (on)	355 (off)
Pulse energy (mJ)	300	150
Pulse repetition frequency (Hz)	200	100
R [illegible]		
Area (m^2)	0.79	
Optical efficiency (%)	~40	
Wavelengths–Rayleigh (nm)	308	355
N_2 Raman (nm) (for aerosol correction)	332	387
[illegible]		
Measurement range* (km)	15–55	
Vertical resolution (km)	3 at bottom, 1 at O_3 peak, 8–10 at top	
Measurement averaging time (hours)	1.5	
Measurement accuracy	<5% at peak, 10–15% at 15 km, >40% at 45 km	

*Chopper added to block beam until it reaches 15 km in order to avoid near field signal effects.

signal with range. For the zenith direction, the advantage is that the airborne lidar system is closer to the region being measured.

Field Measurement Programs

Global O_3 Measurements

The first airborne DIAL system, which was developed by the NASA Langley Research Center (LaRC), was flown for O_3 and aerosol investigations in conjunction with the Environmental Protection Agency's Persistent Elevated Pollution Episodes (PEPE) field experiment, conducted over the east coast of the US in the summer of 1980. This initial system has evolved into the advanced UV DIAL system that will be described in the next section. Airborne O_3 DIAL systems have also been developed and used in field measurement programs by several other groups in the United States, Germany, and France.

The parameters of the current NASA LaRC airborne UV DIAL system are given in Table 2. The on-line and off-line UV wavelengths of 288.2 and 299.6 nm are used for DIAL O_3 measurements during tropospheric missions, and 301 and 310 nm are used for stratospheric missions. This system also transmits 1,064 and 600 nm beams for aerosol and cloud measurements. The time delay between the on- and off-wavelength pulses is 400 μs, which is sufficient time for the return at the first set of wavelengths to end, but short enough that the same region of the atmosphere is sampled. This system has a demonstrated absolute accuracy for O_3 measurements of better than 10% or 2 ppbv (parts per billion by volume), whichever is larger, and a measurement precision of 5% or 1 ppbv with a vertical resolution of 300 m and an averaging time of 5 minutes (about 70 km horizontal resolution at typical DC-8 ground speeds).

The NASA LaRC airborne UV DIAL systems have made significant contributions to the understanding of both tropospheric and stratospheric O_3, aerosols, and clouds. These systems have been used in 18 international and 3 national field experiments over the past 24 years, and during these field experiments, measurements were made over, or near, all of the oceans and continents of the world. A few examples of the scientific contributions made by these airborne UV DIAL systems are given in Table 3.

The NASA LaRC airborne UV DIAL system has been used extensively in NASA's Global Tropospheric Experiment (GTE) program which was started in the early 1980s, had as its primary mission the study of tropospheric chemistry in remote regions of the Earth, in part to study and gain a better understanding of atmospheric chemistry in the unperturbed atmosphere. A related goal was to document the Earth's atmosphere in a number of places during seasons when anthropogenic influences are largely absent, then return to these areas later to document the changes. The field missions have included campaigns in Africa, Alaska, the Amazon Basin,

Table 2 Parameters of the NASA LaRC airborne UV DIAL system

L[illegible]: Nd:YAG-pumped dye lasers, frequency doubled into the UV			
Pulse repetition frequency (Hz)	30		
Pulse length (ns)	8–12		
Pulse energy (mJ) at 1.06 μm	250–300		
Pulse energy (mJ) at 600 nm	50–70		
UV pulse energy (mJ)			
For troposphere at 288/300 nm	20		
For stratosphere at 301/310 nm	20		
Dimensions (l × w × h) (cm)	594 × 102 × 109		
Mass (kg)	1735		
Power requirement (kW)	30		
	Wavelength region (nm)		
R[illegible]	289–311	572–622	1064
Area (m^2)	0.086	0.086	0.864
Receiver optical efficiency (%)	30	40	30
Detector quantum efficiency (%)	26 (PMT)	8 (PMT)	40 (APD)
Field-of-view (mrad)	≤1.5	≤1.5	≤1.5
[illegible]			
Measurement range (km)	up to 10–15 (nadir and zenith)		
Vertical resolution (m)	300–1500, depending on range		
Horizontal resolution (km)	≤70		
Measurement accuracy	≤10% or 2 ppbv, whichever is greater		

Table 3 Examples of significant contributions of airborne O_3 DIAL systems to the understanding of tropospheric and stratospheric O_3 (see also **Figures 1–5**)

[illegible]:
Air mass characterizations in a number of remote regions
Continental pollution plume characterizations (see **Figures 1** and **2**)
Study of biomass burn plumes and the effect of biomass burning on tropospheric O_3 production (see **Figure 3**)
Case study of warm conveyor belt transport from the tropics to the Arctic
Study of stratospheric intrusions and tropopause fold events (see **Figure 4**)
Observation of the decay of a cutoff low
Power plant plume studies
[illegible]:
Contributions to the chemical explanation of behavior of Antarctic O_3
Quantification of O_3 depletion in the Arctic (see **Figure 5**)
Polar stratospheric clouds – particle characterizations
Intercomparison with surface-based, airborne and space-based instruments
Quantification of O_3 reduction in tropical stratosphere after the June 1991 eruption of Mount Pinatubo and characterization of the tropical stratospheric reservoir edge
Cross-vortex boundary transport

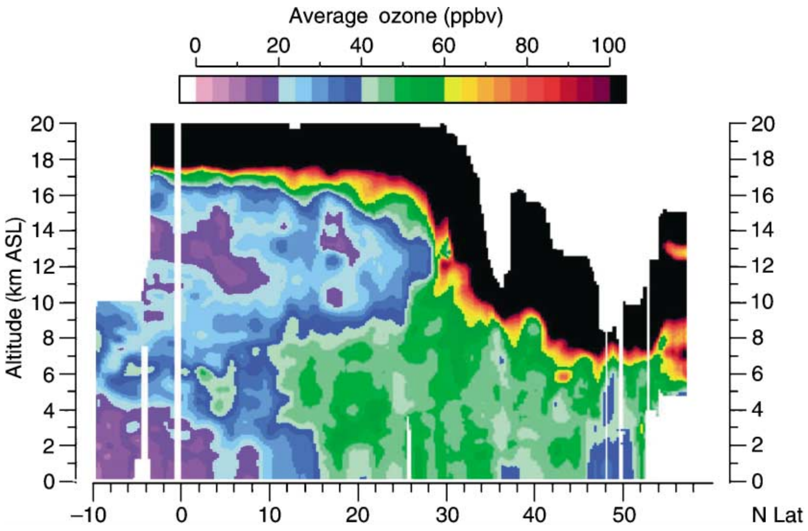

Figure 1 Latitudinal distribution of ozone over the western Pacific Ocean obtained during the second Pacific Exploratory Mission (PEM West B) in 1994.

Canada, and North America, and over the Pacific Ocean from Antarctica to Alaska, with concentrated measurements off the east coast of Asia and in the tropics. One of the bigger surprises of the two decades of field missions, was the discovery in the mid-1990s, that the tropical and South Pacific Ocean had very high tropospheric O_3 concentrations in plumes from biomass burning in Africa and South America during the austral spring. There were few indications from surface-based or space-based measurements that there were extensive biomass burn plumes in the area, primarily since the plumes were largely devoid of aerosols due to being stripped out during cloud convective lofting. Once over the ocean, horizontal transport appears to proceed relatively unimpeded unless a storm system is encountered.

The NASA LaRC airborne UV DIAL system has also been flown in all the major stratospheric O_3 campaigns starting in 1987 with the Airborne Antarctic Ozone Experiment (AAOE) to determine the cause of the Antarctic ozone hole. The UV DIAL system documented the O_3 loss across the ozone hole region. Later, when attention was turned to the Arctic and the possibility of an ozone hole there, the system was used to produce an estimate of O_3 loss during the winter season as well as to better characterize the polar stratospheric cloud (PSC) particles. The UV DIAL system was also used to study O_3 loss in the tropical stratospheric reservoir following the eruption of Mount Pinatubo in June 1991. The loss was spotted by the Microwave Limb Sounder (MLS) on the Upper Atmospheric Research Satellite (UARS), but the MLS

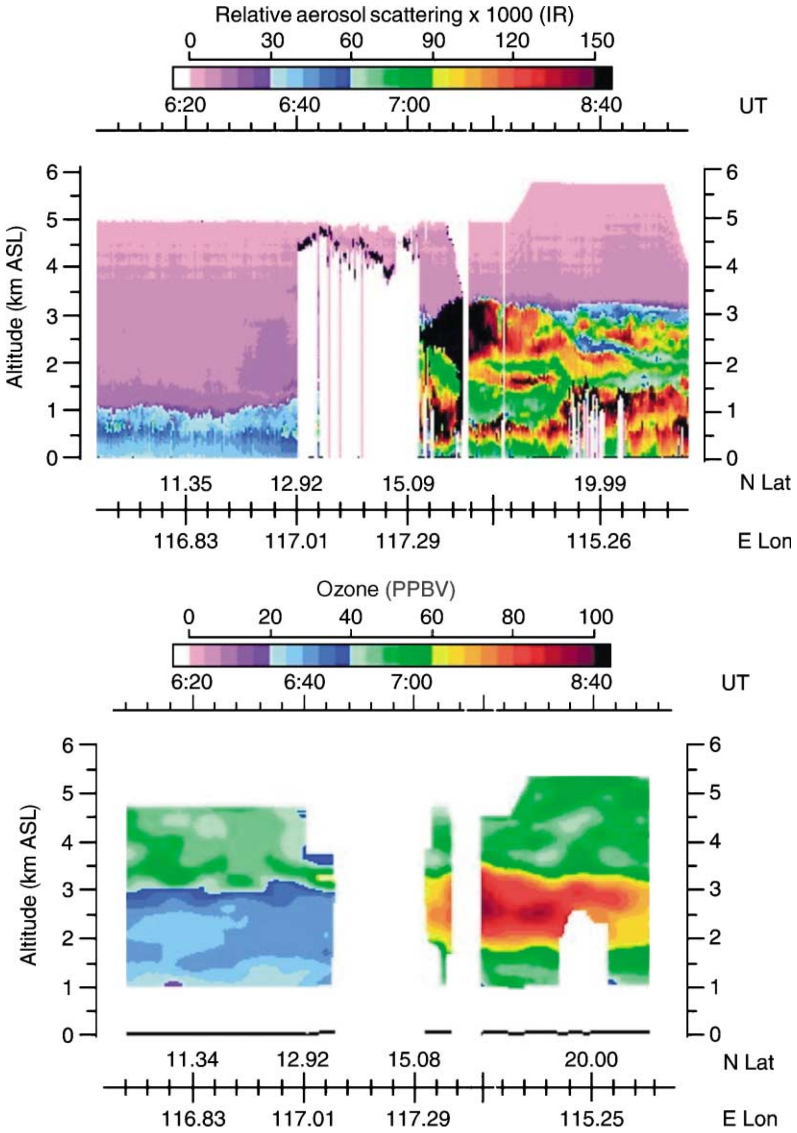

Figure 2 Pollution outflow from China over the South China Sea (right side of figure) with clean tropical air on south side of a front (left side of figure), observed during PEM West B in 1994.

was unable to study the loss in detail due to its low vertical resolution (5 km) compared to the small-scale (2–3 km) features of the O_3 loss. Other traditional space-based O_3 measuring instruments also had difficulty during this period, due to the high aerosol loading in the stratosphere following the eruption.

Space-Based O_3 DIAL System

In order to obtain nearly continuous, global distributions of O_3 in the troposphere, a space-based O_3 DIAL system is needed. A number of key issues could be addressed by a space-based O_3 DIAL system including: the global distribution of photochemical O_3 production/destruction and transport in the troposphere; location of the tropopause; and stratospheric O_3 depletion and dynamics. High-resolution airborne O_3 DIAL and other aircraft measurements show that to study tropospheric processes associated with biomass burning, transport of anthropogenic pollutants, tropospheric O_3 chemistry and dynamics, and stratosphere–troposphere exchange, a vertical profiling capability from space with a resolution of 2–3 km is needed, and this capability cannot currently be achieved using passive remote sensing satellite instruments. An example of the type of latitudinal O_3 cross-section that could be provided by a space-based O_3 DIAL system is shown in Figure 1. This figure shows many different aspects of O_3 loss and production; vertical and horizontal transport; and stratosphere–troposphere exchange that occurs from the tropics to high latitudes.

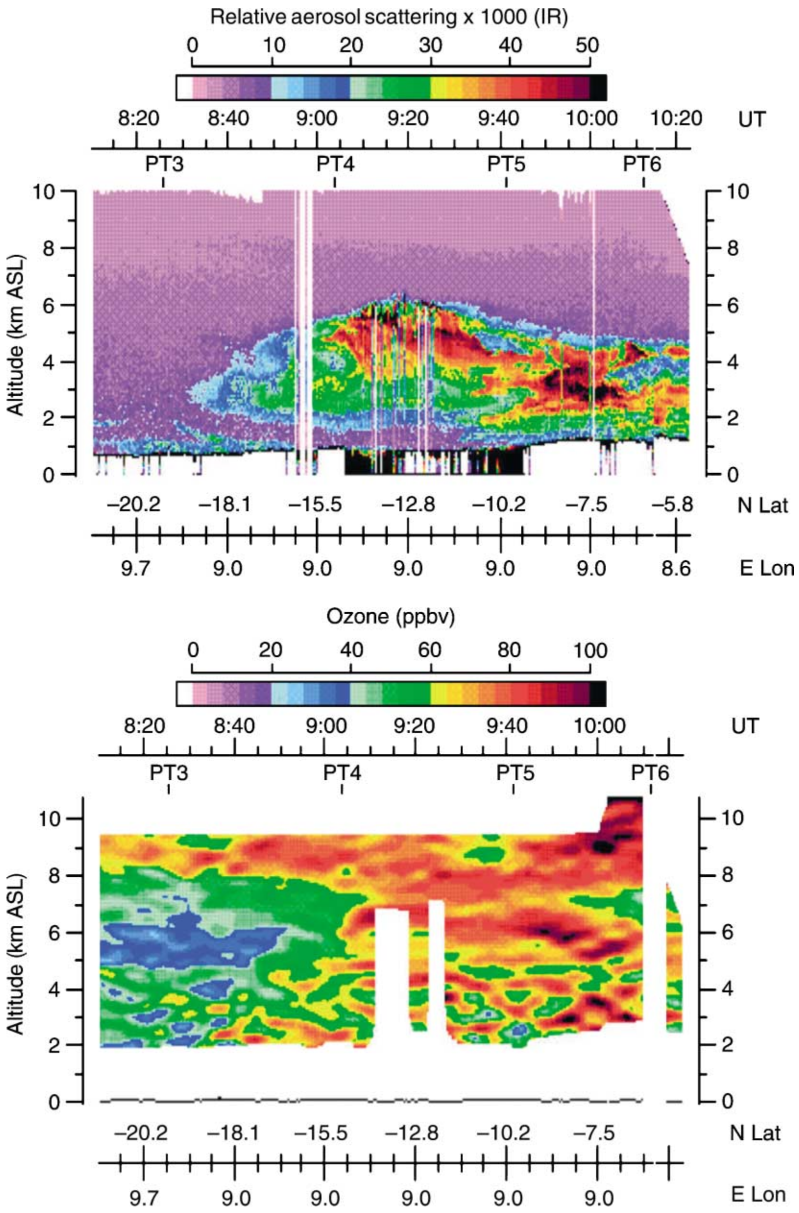

Figure 3 Biomass burning plume over the Atlantic Ocean arising from biomass burning in the central part of western Africa, observed during the TRACE-A mission in 1992.

This type of data would be available from just one pass from a space-based O_3 DIAL system. A space-based O_3 DIAL system optimized for tropospheric O_3 measurements (see system description in Table 4a,b) would also permit high-resolution O_3 measurements in the stratosphere (1 km vertical, 100 km horizontal), along with high-resolution aerosol measurements (100 m vertical, 10 km horizontal). In addition, these DIAL measurements will be useful in assisting in the interpretation of passive remote sensing measurements and in helping to improve their data processing algorithms.

Global H_2O Measurements

H_2O and O_3 are important to the formation of OH in the troposphere, and OH is at the center of most of the chemical reactions in the lower atmosphere. In addition H_2O is an excellent tracer of vertical and horizontal transport of air masses in the troposphere, and it can be used as a tracer of stratosphere–troposphere exchange. Increased aerosol sizes, due to high relative humidities, can also affect heterogeneous chemical processes and radiation budgets in the boundary layer and in cloud layers. Knowledge of H_2O is important to weather forecasting and climate

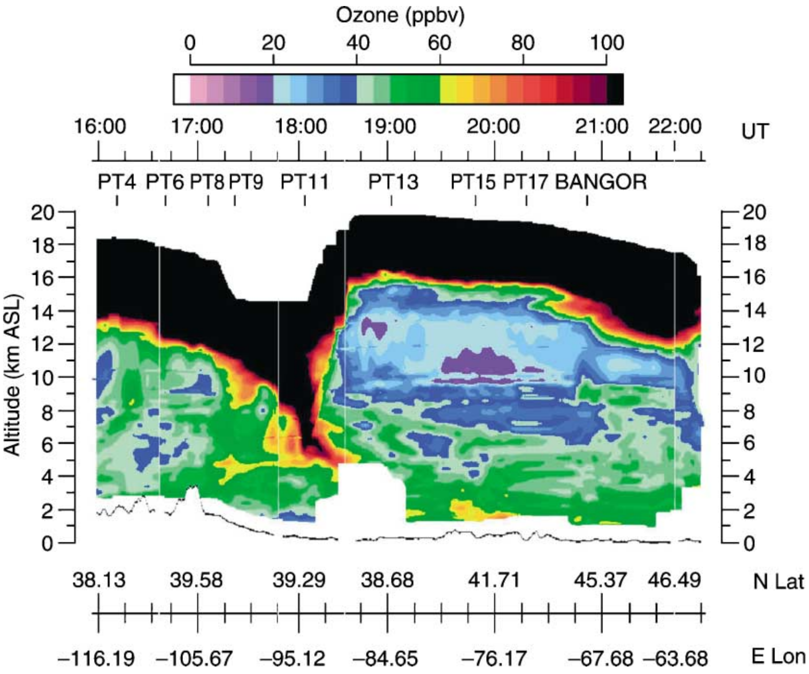

Figure 4 Ozone distribution observed on flight across US during the SASS (Subsonic Assessment) Ozone and Nitrogen Experiment (SONEX) in 1997. A stratospheric intrusion is clearly evident on left side of figure, and low ozone air from tropics transported to mid-latitudes can be seen in the upper troposphere on the right.

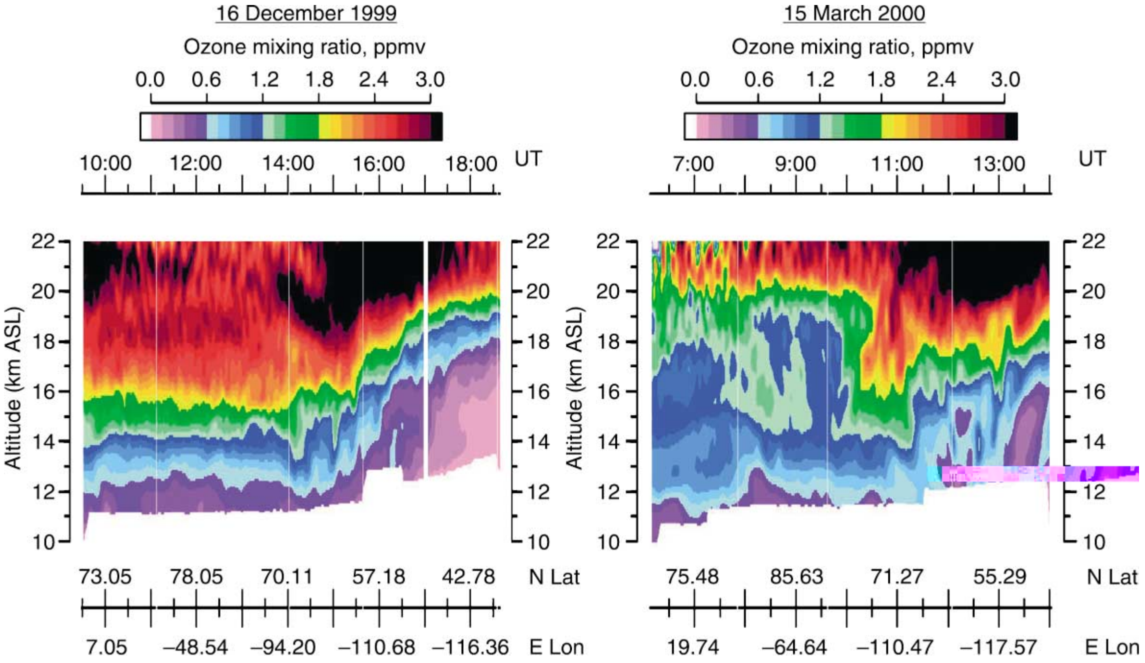

Figure 5 Ozone cross-sections in the stratosphere measured in the winter of 1999/2000 during the SOLVE mission. The change in ozone number density in the Arctic polar vortex due to chemical loss during the winter is clearly evident at latitudes north of 72°N.

predictions. Thus, H_2O distributions can be used in several different ways to better understand chemical, transport, radiation and meteorological processes in the global troposphere.

H_2O Raman Lidar Systems

The first Raman lidar measurements of H_2O were made in the late 1960s, but not much progress in using the Raman approach for H_2O was made until a system using an abandoned searchlight mirror was employed by the NASA Goddard Space Flight Center in the mid-1980s, to show that if the signal collector (telescope) was large enough, useful Raman measurements could be made to distances of several kilometers. Until recently, Raman lidar measurements of H_2O were largely limited to night-time, due to the high sunlight background interference. Measurements are now made during daytime, using very narrow telescope fields of view and very narrowband filters in the receiver. A good example of a Raman lidar system used to measure H_2O is that at the Department of Energy's Atmospheric Radiation Measurement – Cloud and Radiation Test Bed (ARM-CART) site in Oklahoma. The site is equipped with a variety of instruments aimed at studying the radiation properties of the atmosphere. The system parameters are given in Table 4a,b.

Raman lidar systems have been used for important measurements of H_2O from a number of ground-based locations. Some of the important early work dealt with the passage of cold and warm fronts. The arrival of the wedge-shaped cold front, pushing the warm air up, was one of these studies. Raman lidar have also been located at the ARM-CART site in Kansas and Oklahoma, where they could both provide a climatology of H_2O as well as provide correlative measurements of H_2O for validation of space-based instruments.

Table 4a Parameters for the US Department of Energy's surface-based Raman lidar system for measurements of H_2O

L	Nd:YAG Laser
Wavelength (nm)	355
Pulse energy (mJ)	400
Pulse repetition frequency (Hz)	30
R	
Area (m^2)	1.1
Wavelengths (nm)	
Water vapor	407
Nitrogen	387
Measurement range (km)	
Night-time	near surface to 12
Daytime	near surface to 3
Range resolution (m)	39 at low altitudes, 300 >9 km
Measurement accuracy (%)	
Night-time	<10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)
Daytime	10 <1 km; 5–15 for 1–3 km (10 min avg)

Table 4b Parameters for the NASA Goddard Space Flight Center's Scanning Raman lidar system for measurements of H_2O

	D/N	*N*
L	Nd:YAG Laser	XeF
Wavelength (nm)	355	351
Pulse energy (mJ)	300	30–60
Pulse repetition frequency (Hz)	30	400
R		
Area (m^2)	1.8	1.8
Wavelengths (nm)		
Water vapor	407	403
Nitrogen	387	382
Measurement range (km)		
Night-time	near surface to 12	
Daytime	near surface to 4	
Range resolution (m)	7.5 at low altitudes, 300 >9 km	
Measurement accuracy (%)		
Night-time	<10 <5 km (10 sec and 7.5 m range resolution) <10 <7 km (1 min avg) 10–30 at high altitudes (10–30 min avg)	
Daytime	<10 <4 km (5 min avg)	

H_2O DIAL Systems

H_2O was first measured with the DIAL approach using a temperature-tuned ruby laser lidar system in the mid-1960s. The first aircraft-based H_2O DIAL system was developed at NASA LaRC, and was flown in 1982, as an initial step towards the development of a space-based H_2O DIAL system. This system was based on Nd:YAG-pumped dye laser technology, and it was used in the first airborne H_2O DIAL atmospheric investigation, which was a study of the marine boundary layer over the Gulf Stream. This laser was later replaced with a flashlamp-pumped solid-state alexandrite laser, which had high spectral purity, i.e., little out-of-band radiation, a requirement since water vapor lines are narrow, and this system was used to make accurate H_2O profile measurements across the lower troposphere.

A third H_2O DIAL system, called LASE (Lidar Atmospheric Sensing Experiment) was developed as a prototype for a space-based H_2O DIAL system, and it was completed in 1995. This was the first fully autonomously operating DIAL system. LASE uses a Ti:sapphire laser that is pumped by a double-pulsed, frequency-doubled Nd:YAG to produce laser pulses in the 815 nm absorption band of H_2O (see Table 5). The wavelength of the Ti:sapphire laser is controlled by injection seeding with a diode laser that is frequency locked to a H_2O line using an

Table 5 Parameters of LASE H_2O DIAL system

L [illegible]	Ti:sapphire
Wavelength (nm)	813–818
Pulse energy (mJ)	100
Pulse-pair repetition frequency	5 (on- and off-line pulses separated by 300 μs)
Linewidth (pm)	<0.25
Stability (pm)	<0.35
Spectral purity (%)	>99
Beam divergence (mrad)	<0.6
Pulse width (ns)	35
R [illegible]	
Area (m^2)	0.11
Receiver optical efficiency (%)	50 (night), 35 (day)
Avalanch Photodiode (APD) detector quantum efficiency (%)	80
Field-of-view (mrad)	1.0
Noise equivalence power ($W\ Hz^{-0.5}$)	2×10^{-14}
Excess noise factor	3
[illegible]	
Measurement range (altitude) (km)	15
Range resolution (m)	300–500
Measurement accuracy (%)	5

Table 6 Examples of significant contributions of airborne H_2O DIAL systems to the understanding of H_2O distributions (see also **Figures 6–8**)

NA A L [illegible] *H_2O DIAL* [illegible] *LA E*
Study of marine boundary layer over Gulf Stream
Observation of H_2O transport at a land/sea edge
Study of large-scale H_2O distributions across troposphere (see **Figure 6**)
Correlative [illegible] and remote measurements
Observations of boundary layer development (see **Figure 7**)
Cirrus cloud measurements
Hurricane studies (see **Figure 8**)
Relative humidity effects on aerosol sizes
Ice supersaturation in the upper troposphere
Studies of stratospheric intrusions
H_2O distributions over remote Pacific Ocean
O [illegible] *H_2O DIAL* [illegible]
Boundary layer humidity fluxes
Lower-stratospheric H_2O studies

absorption cell. Each pulse pair consists of an on-line and off-line wavelength for the H_2O DIAL measurements. To cover the large dynamic range of H_2O concentrations in the troposphere (over 3 orders of magnitude), up to three line pair combinations are needed. LASE uses a novel approach of operating from more than one position on a strongly absorbing H_2O line. In this approach, the laser is electronically tuned at the line center, side of the line, and near the wing of the line to achieve the required absorption cross-section pairs (on and off). LASE has demonstrated measurements of H_2O concentrations across the entire troposphere using this 'side-line' approach. The accuracy of LASE H_2O profile measurements was determined to be better than 6% or 0.01 g/kg, whichever is larger, over the full dynamic range of H_2O concentrations in the troposphere. LASE has participated in over eight major field experiments since 1995. See Table 6 for a listing of topics studied using airborne H_2O DIAL systems (Figures 6–8).

Space-Based H_2O DIAL System

The technology for a space-based H_2O DIAL system is rapidly maturing in the areas of: high-efficiency, high-energy, high-spectral-purity, long-life lasers with tunability in the 815- and 940-nm regions; low-weight, large-area, high-throughput, high-background-rejection receivers; and high-quantum-efficiency, low-noise, photon-counting detectors. With the expected advancements in lidar technologies leading to a 1-J/pulse capability, a space-based H_2O DIAL system could be flown on a long-duration space mission this decade.

Space-based DIAL measurements can provide a global H_2O profiling capability, which when combined with passive remote sensing with limited vertical resolution, can lead to 3-dimensional measurements of global H_2O distributions. High vertical resolution H_2O (≤1 km), aerosol (≤100 m), and cloud top (≤50 m) measurements from the lidar along the satellite ground-track, can be combined with the horizontally contiguous data from nadir passive sounders to generate more complete high-resolution H_2O, aerosol, and cloud fields for use in the various studies indicated above. In addition, the combination of active and passive measurements can provide significant synergistic benefits leading to improved temperature and relative humidity measurements. There is also strong synergism with aerosol and cloud imaging instruments and with future passive instruments that are being planned or proposed for missions addressing atmospheric chemistry, radiation, hydrology, natural hazards, and meteorology.

Tunable Laser Systems for Point Monitoring

Tunable diode laser (TDL) systems are also used to measure atmospheric gases on a global scale from aircraft and balloons. TDLs are small lasers that emit extremely narrowband radiation and can be tuned in the near IR spectral region using a combination of temperature and current. TDLs can be built into very sensitive and compact systems and located on the surface or flown on aircraft or balloons and used to measure such species as CO, HCl, CH_4, and oxides of nitrogen such as N_2O and NO_2. They derive their

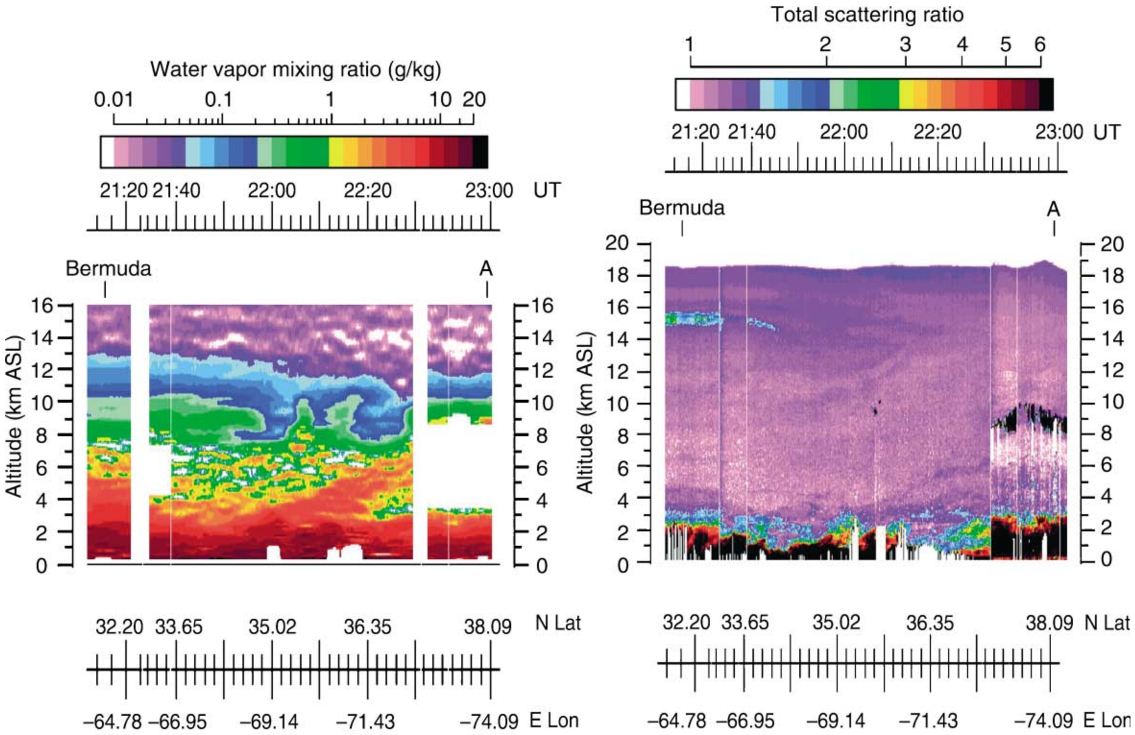

Figure 6 LASE measurements of water vapor (left) and aerosols and clouds (right) across the troposphere on an ER-2 flight from Bermuda to Wallops during the Tropospheric Aerosol Radiative Forcing Experiment (TARFOX) conducted in 1996.

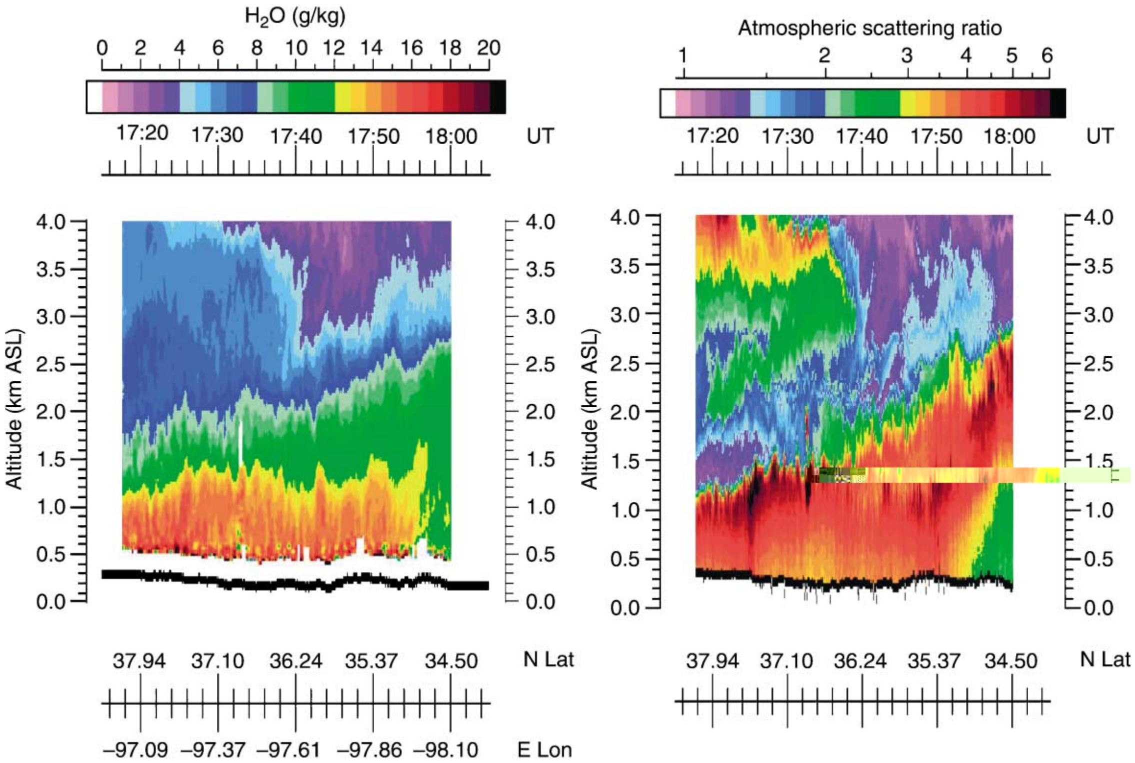

Figure 7 Water vapor and aerosol cross-section obtained on a flight across a cold front during the Southern Great Plains (SGP) field experiment conducted over Oklahoma in 1997.

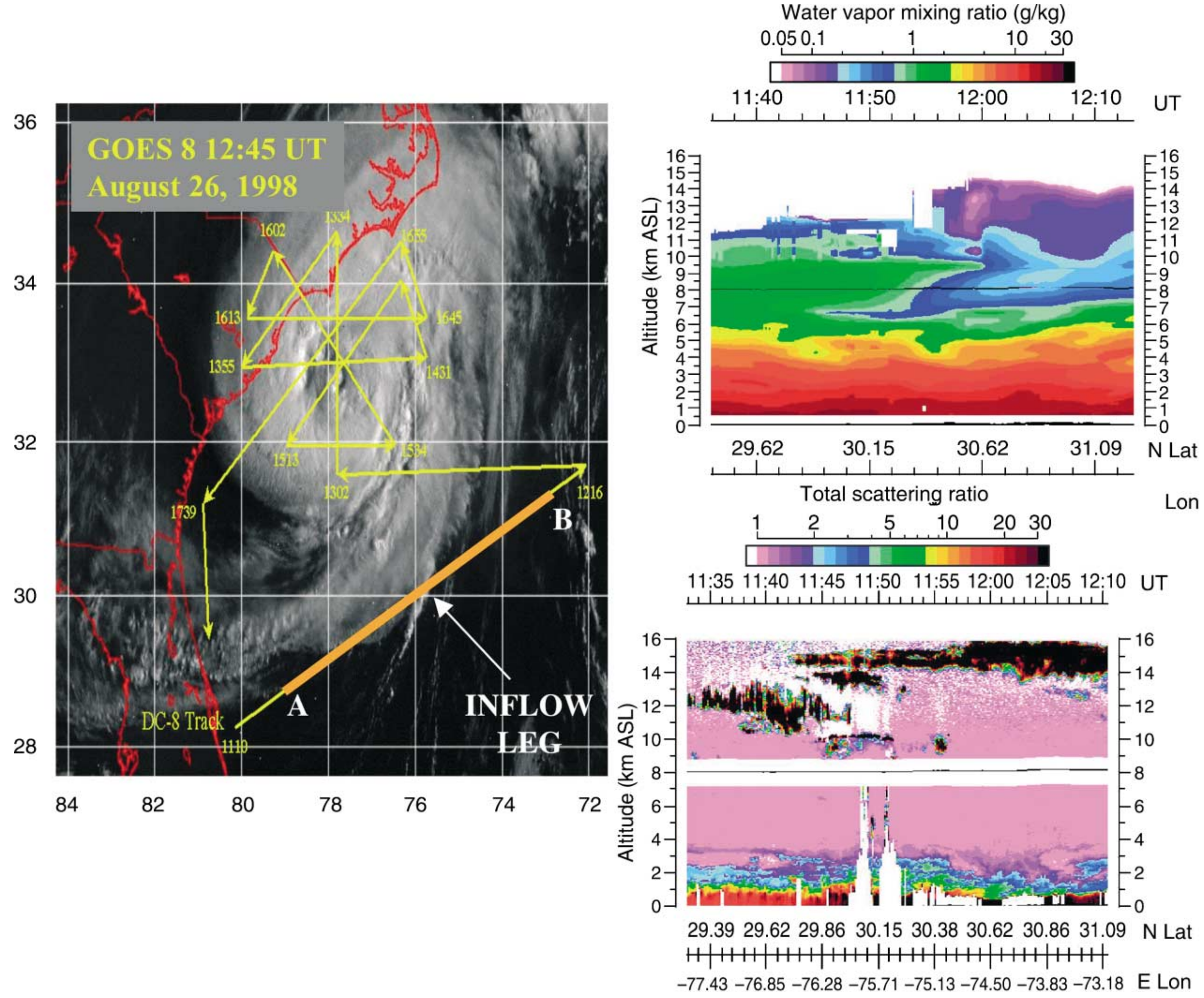

Figure 8 Measurements of water vapor, aerosols, and clouds in the inflow region of Hurricane Bonnie during 1998 Convection and Moisture Experiment (CAMEX-3). A rain band can be clearly seen at the middle of Leg-AB on the satellite and LASE cross-sections.

high sensitivity to the fact that the laser frequency is modulated at a high frequency, permitting a small spectral region to be scanned rapidly. The second- or fourth-harmonic of the scan frequency is used in the data acquisition, effectively eliminating much of the low-frequency noise due to mechanical vibrations and laser power fluctuations. In addition, multipass cells are employed, thereby generating long paths for the absorption measurements.

TDL systems have been used in a number of surface-based measurement programs. One system was mounted on a ship doing a latitudinal survey in the Atlantic Ocean and monitored NO_2, formaldehyde (HCHO), and H_2O_2. Another TDL system was located at the Mauna Loa Observatory and was used to monitor HCHO and H_2O_2 during a photochemistry experiment, finding much lower concentrations of both gases than models had predicted. The TDL system used to measure HCHO has been used in a number of additional ground-based measurement programs and has also been flown on an aircraft in a couple of tropospheric missions.

One TDL system, called DACOM (Differential Absorption CO Measurement), has been used in a large number of NASA GTE missions. It makes measurements of CO in the 4.7 μm spectral region and CH_4 in the 3.3 or 7.6 μm spectral region. This instrument is able to make measurements at a 1 Hz rate and with a precision of 0.5–2.0%, depending on the CO value, and an accuracy of ±2%. DACOM has been very useful in determining global distributions of CO due to the wide-ranging nature of the GTE missions. It has also contributed to the characterization and understanding of CO in air masses encountered during the GTE missions (Table 7).

A pair of TDL instruments, the Airborne Tunable Laser Absorption Spectrometer (ATLAS) and the Aircraft (ER-2) Laser Infrared Absorption Spectrometer (ALIAS), have been flown on several NASA missions to explore polar O_3 chemistry and atmospheric transport. The ATLAS instrument measures N_2O, and ALIAS is a 4-channel spectrometer that measures a large variety of gases, including HCl, N_2O, CH_4, NO_2, and CO and is currently configured to measure water isotopes across the tropopause.

A new class of lasers, tunable quantum-cascade (QC) lasers, are being added to the list of those available for *in situ* gas measurement systems. Using a cryogenically cooled QC laser during a series of 20 aircraft flights beginning in September 1999 and extending through March 2000, measurements were made of CH_4 and N_2O up to ~20 km in the stratosphere over North America, Scandinavia, and Russia, on the NASA ER-2. Compared with its companion lead salt diode lasers, that were also flown on these flights, the single-mode QC laser, cooled to 82 K, produced higher output power (10 mW), narrower laser linewidth (17 MHz), increased measurement precision (a factor of 3), and better spectral stability (~0.1 cm^{-1} K). The sensitivity of the QC laser channel was estimated to correspond to a minimum-detectable mixing ratio of approximately 2 ppbv of CH_4.

Laser Long-Path Measurements

Laser systems can also be used in long-path measurements of gases. The most important of such programs entailed the measurement of hydroxyl radical (OH) in the Rocky Mountains west of

Table 7 Examples of significant contributions of airborne tunable diode laser systems to the understanding of atmospheric chemistry and transport

B[illegible]
 The first [illegible] measurements of the suite NO_2, NO, O_3, and the NO_2 photolysis rate to test NO[illegible] (NO_2 + NO) photochemistry
 The first [illegible] stratospheric measurements of NO[illegible] over a full diurnal cycle to test N_2O_5 chemistry
 [illegible] measurements of NO_2 and HNO_3 over the 20–35 km region to assess the effect of Mt. Pinatubo aerosol on heterogeneous atmospheric chemistry
 Measurements of HNO_3 and HCl near 30 km
 Measurements of CH_4, HNO_3, and N_2O for validation of several satellite instruments
 Intrusions from midlatitude stratosphere to tropical stratospheric reservoir
A[illegible]
 C[illegible]
  Measurement of CO from biomass burning from Asia, Africa, Canada, Central America, and South America
  Detection of thin layers of CO that were transported thousands of miles in the upper troposphere
  Observed very high levels of urban pollution in plumes off the Asian continent
  Emission indices for many gases have been calculated with respect to CO
 M[illegible]
  Determined CH_4 flux over the Arctic tundra, which led to rethinking of the significance of tundra regions as a global source of CH_4
  Found that biomass burning is a significant source of global CH_4
  Found strong enhancements of CH_4 associated with urban plumes
A[illegible]
 [illegible]
  Extreme denitrification observed in Antarctic winter vortex from NO[illegible]:N_2O correlation study
  Observed very low N_2O in Antarctic winter vortex, which helped refute the theory that the O_3 hole is caused by dynamics
  Contributed to the study of transport out of the lower stratospheric Arctic vortex by Rossby wave breaking
  Measurement of concentrations of gases involved in polar stratospheric O_3 destruction and production
  Activation of chlorine in the presence of sulfate aerosols
 M[illegible]
  Measurements in aircraft exhaust plumes in the lower stratosphere, especially of reactive nitrogen species and CO
  Vertical profiles of CO in the troposphere and lower stratosphere
  Determination of the hydrochloric acid and the chlorine budget of the lower stratosphere
  Measurement of NO_2 for testing atmospheric photochemical models
  Trends in HCl/Cl[illegible] in the stratosphere <21 km, 1992–1998
  Gas concentration measurements for comparison with a balloon-borne Fourier transform spectrometer observing the Sun
  Near-IR TDL laser hygrometers (ER-2, WB57, DC-8) for measuring H_2O and total water in the lower stratosphere and upper troposphere
  Remnants of Arctic winter vortex detected many months after breakup
 [illegible]
  Tropical entrainment time scales inferred from stratospheric N_2O and CH_4 observations

Boulder, Colorado. OH was measured using a XeCl excimer laser operating near 308 nm and transmitting a beam to a retroreflector 10.3 km away. The measurements were quite difficult to conduct, primarily since OH abundance is very low (10^5–$10^7\,\mathrm{cm}^{-3}$), yielding very low absorption (~0.02% for an abundance of $10^5\,\mathrm{cm}^{-3}$) over the 20.6 km path. In addition, the excimer laser could generate OH from ambient H_2O and O_3, unless the laser energy density was kept low. To help ensure good measurements, a white light source was also employed during the measurements to monitor H_2O, O_3, and other gases. The measurements were eventually very successful and led to new values for OH abundances in the atmosphere.

Laser-Induced Fluorescence (LIF)

The laser-induced fluorescence approach has been used to measure several molecular and ionic species *in situ*. The LIF approach has been used to measure OH and HO_2 (HO_x) on the ground and on aircraft platforms. One airborne LIF system uses a diode-pumped Nd:YAG-pumped, frequency-doubled dye laser to generate the required energy near 308 nm. The laser beam is sent into a White cell where it can make 32–36 passes through the gas in the cell to increase the LIF signal strength. NO is used to convert HO_2 to OH. The detection limit in 1 minute of about 2–3 ppqv (10^{-15}) above 5 km altitude, which translates into a concentration of about 4×10^4 molec/cm^3 at 5 km and 2×10^4 molec/cm^3 at 10 km altitude. One of the interesting findings from such measurements is that HO_x concentrations are up to 5 times larger than model predictions based on NO_x concentrations, suggesting that NO_x emissions from aircraft could have a greater impact on O_3 production than originally thought.

NO is detected using the LIF technique in a two-photon approach: electrons are pumped from the ground state using 226 nm radiation and from that state to an excited state using 1.06 μm radiation. The 226 nm radiation is generated by frequency doubling a dye laser to 287 nm and then mixing that with 1.1 μm radiation derived from H_2 Raman shifting of frequency-mixed radiation from a dye laser and a Nd:YAG laser. From the excited level, 187–201 nm radiation is emitted. In order to measure NO_2, it is first converted to the photofragment NO via pumping at 353 nm from a XeF excimer laser. One of the interesting findings from airborne measurements in the South Pacific is that there appeared to be a large missing source for NO_x in the upper troposphere.

Summary

DIAL and Raman lidar systems have played important roles in studying the distribution of gases such as O_3 and H_2O on local, regional, and global scales, while TDL systems have played corresponding roles for such gases as CO, HCl, and oxides of nitrogen. It is anticipated that these approaches will continue to yield valuable information on these and other gases, as new and more capable systems are developed. Within the next decade, it is expected that a DIAL system will be placed in orbit to make truly global measurements of O_3, H_2O, and/or carbon dioxide.

See also

Imaging: Lidar. **Scattering:** Raman Scattering.

Further Reading

Browell EV (1994) Remote sensing of trace gases from satellites and aircraft. In: Calvert J (ed.) *Chemistry of the Atmosphere: The Impact on Global Change*, pp. 121–134. Cambridge, MA: Blackwell Scientific Publications.

Browell EV, Ismail S and Grant WB (1998) Differential absorption lidar (DIAL) measurements from air and space. *Applied Physics. B* 67: 399–410.

Dabas A, Loth C and Pelon J (eds) (2001) *Advances in Laser Remote Sensing*. Selected Papers presented at the 20th International Laser Radar Conference (ILRC), Vichy, France, 10–14 July 2000, pp. 357–360. Palaiseau Cedex, France: A. Edition d'Ecole Polytechnique.

Goldsmith JEM, Bisson SE, Ferrare RA, *et al.* (1994) Raman lidar profiling of atmospheric water vapor: simultaneous measurements with two collocated systems. *Bulletin American Meteorological Society* 75: 975–982.

Goldsmith JEM, Blair FH, Bisson SE and Turner DD (1998) Turn-key Raman lidar for profiling atmospheric water vapor, clouds, and aerosols. *Applied Optics* 37: 4979–4990.

Grant WB (1995) Lidar for atmospheric and hydrospheric studies. In: Duarte F (ed.) *Tunable Laser Applications*, pp. 213–305. New York: Marcel Dekker Inc.

Grant WB, Browell EV, Menzies RT, Sassen K and She C-Y (eds) (1997) *Laser Applications in Remote Sensing*. SPIE Milestones series, 690 pp, 86 papers.

Grisar R, Booettner H, Tacke M and Restelli G (1992) *Monitoring of Gaseous Pollutants by Tunable Diode Lasers*. Proceedings of the International Symposium held in Freiburg, Germany, 17–18 October 1991, 372 pp. Dordrecht, The Netherlands: Kluwer Academic Publishers.

McDermid IS, Walsh TD, Deslis A and White M (1995) Optical systems design for a stratospheric lidar system. *Applied Optics* 34: 6201–6210.

Mégie G (1988) Laser measurements of atmospheric trace constituents. In: Measures RM (ed.) *Laser Remote Chemical Analysis*, pp. 333–408. New York: J Wiley & Sons.

Podolske JR and Loewenstein M (1993) Airborne tunable diode laser spectrometer for trace gas measurements in the lower stratosphere. *Applied Optics* 32: 5324–5330.

Pougatchev NS, Sachse GW, Fuelberg HE, *et al.* (1999) Pacific Exploratory Mission-Tropics carbon monoxide measurements in historical context. *Journal of Geophysics Research* 104: 26,195–26,207.

Sachse GW, Collins Jr JE, Hill GF, *et al.* (1991) Airborne tunable diode laser sensor for high precision concentration and flux measurements of carbon dioxide and methane. *Proceedings of the SPIE International Society for Optical Engineering* 1443: 145–156.

Svanberg S (1994) Differential absorption lidar (DIAL). In: Segrist MW (ed.) *Air Monitoring Spectroscopic Techniques*, pp. 85–161. New York: J Wiley & Sons.

Webster CR, *et al.* (1994) Aircraft (ER-2) laser infrared absorption spectrometer (ALIAS) for in situ measurements of HCl, N_2O, CH_4, NO_2, and HNO_3. *Applied Optics* 33: 454–472.

Whiteman DN, Melfi SH and Ferrare RA (1992) Raman lidar system for the measurement of water vapor and aerosols in the Earth's atmosphere. *Applied Optics* 31: 3068–3082.

Zanzottera E (1990) Differential absorption lidar techniques in the determination of trace pollutants and physical parameters of the atmosphere. *Critical Reviews in Analytical Chemistry* 21: 279–319.

Optical Transmission and Scatter of the Atmosphere

S M Adler-Golden and A Berk,
Spectral Sciences, Inc., Burlington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to understand and model radiative transfer (RT) processes in the atmosphere is critical for remote sensing, environmental characterization, and many other areas of scientific and practical interest. At the Earth's surface, the bulk of this radiation, which originates from the sun, is found in the ultraviolet to infrared range between around 0.3 and 4 μm. This article describes the most significant RT processes for these wavelengths, which are absorption (light attenuation along the line of sight LOS) and elastic scattering (redirection of the light). Transmittance, T, is defined as one minus the fractional extinction (absorption plus scattering). At longer wavelengths (in the mid- and long-wave infrared) the major light source is thermal emission.

A few other light sources are mentioned here. The moon is the major source of visible light at night. Forest fires can be a significant source of mid-wave infrared radiation. Manmade light sources include continuum sources such as incandescent lamps and spectrally narrow sources such as fluorescent lamps and lasers. In general, spectrally narrow sources need to have a different RT treatment than continuum sources due to the abundance of narrow spectral absorption lines in the atmosphere, as is discussed below.

The challenge of atmospheric RT modeling is essentially to solve the following equation that describes monochromatic light propagation along the LOS direction:

$$\left(\frac{1}{k}\right)\frac{\mathrm{d}I}{\mathrm{d}u} = -I + J \qquad [1]$$

where I is the LOS radiance (watts per unit area per unit wavelength per steradian), k is the extinction coefficient for the absorbing and scattering species (per unit concentration per unit length), u is the material column density (in units of concentration times length), and J is the radiance source function. I is a sum of direct (i.e., from the sun) and diffusely scattered components. The source function represents the diffuse light scattered into the LOS, and is the angular integral over all directions Ω_i of the product of the incoming radiance, $I(\Omega_\mathrm{i})$, and the scattering phase function, $p(\Omega_\mathrm{o}, \Omega_\mathrm{i})$:

$$J(\Omega_\mathrm{o}) = \int_{\Omega_\mathrm{i}} p(\Omega_\mathrm{o}, \Omega_\mathrm{i}) I(\Omega_\mathrm{i}) \mathrm{d}\Omega_\mathrm{i} \qquad [2]$$

The scattering phase function describes the probability density for incoming light from direction Ω_i scattering out at angle Ω_o, and is a function of the difference (scattering) angle θ.

The direct radiance component, I', is described by eqn [1] with the source function omitted. Integrating along the LOS leads to the well-known Beer's Law equation for transmittance:

$$T = I'/I'_0 = \exp(-ku) \qquad [3]$$

where I'_0 is the direct radiance at the boundary of the LOS. The quantity $ku = \ln(1/T)$ is known as the optical depth.

Atmospheric Constituents

The atmosphere has a large number of constituents, including numerous gaseous species and suspended liquid and solid particulates. Their contributions to extinction are depicted in Figure 1. The largest category is gases, of which the most important are water vapor, carbon dioxide, and ozone. Of these, water vapor is the most variable and carbon dioxide the least, although the CO_2 concentration is gradually increasing. In atmospheric RT models, carbon dioxide is frequently taken to have a fixed and altitude-independent concentration, along with other 'uniformly mixed gases' (UMGs). The concentration profiles of the three major gas species are very different. Water vapor is located mainly in the lowest 2 km of the atmosphere. Ozone has a fairly flat profile from the ground through the stratosphere (~30 km). The UMGs decline exponentially with altitude, with a scale height of around 8 km.

Gases

Gas molecules both scatter and absorb light. Rayleigh scattering by gases scales inversely with the fourth power of the wavelength, and is responsible for the sky's blue color. For typical atmospheric conditions, the optical depth for Rayleigh extinction is approximately $0.009/\lambda^4$ per air mass (λ is in μm and air mass is defined by the vertical column from ground to space). The Rayleigh phase function has a $(1 + \cos^2\theta)$ dependence.

Absorption by gases may consist of a smooth spectral continuum (such as the ultraviolet and visible electronic transitions of ozone) or of discrete spectral lines, which are primarily rotation lines of molecular vibrational bands. In the lower atmosphere (below around 25 km altitude), the spectral shape of these lines is determined by collisional broadening and described by the normalized Lorentz line shape formula:

$$f_{\text{Lorentz}}(\nu) = \frac{\alpha_c/\pi}{\alpha_c^2 + (\nu - \nu_0)^2} \qquad [4]$$

Here ν is the wavenumber (in cm^{-1}) [$\nu = (10\,000\ \mu\text{m}/\text{cm})/\lambda$], ν_0 is the molecular line transition frequency and α_c is the collision-broadened half-width (in cm^{-1}), which is proportional to pressure. The constant of proportionality, known as the pressure-broadening parameter, has a typical value on the order of $0.06\ \text{cm}^{-1}\ \text{atm}^{-1}$ at ambient temperature. The extinction coefficient $k(\nu)$ is the product of $f_{\text{Lorentz}}(\nu)$ and the integrated line strength S, which is commonly in units of $\text{atm}^{-1}\ \text{cm}^{-2}$.

At higher altitudes in the atmosphere the pressure is reduced sufficiently that Doppler broadening

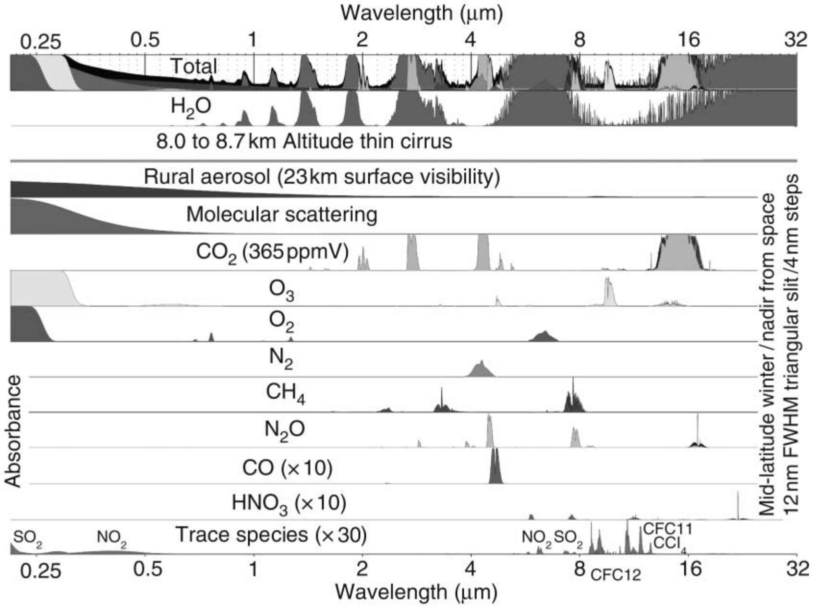

Figure 1 Spectral absorbance (1 − transmittance) for the primary sources of atmospheric extinction. The 12 nm resolution data were generated by MODTRAN for a vertical path from space with a mid-latitude winter model atmosphere.

becomes competitive with collisional broadening, and the Lorentz formula becomes inaccurate. The general lineshape for combined collisional and Doppler broadening is the Voigt lineshape, which is proportional to the real part of the complex error (probability) function, w:

$$f_{\mathrm{Voigt}}(\nu) = \frac{1}{\alpha_{\mathrm{d}}\sqrt{\pi}} \mathrm{Re}\left[w\left(\frac{\nu - \nu_0}{\alpha_{\mathrm{d}}} + \frac{\alpha_{\mathrm{c}}}{\alpha_{\mathrm{d}}} \mathrm{i} \right) \right] \qquad [5]$$

Here, α_{d} is the Doppler $1/e$ half-width.

Comprehensive spectral databases have been compiled of the transition frequencies, strengths, and pressure-broadened half-widths for atmospherically important molecules throughout the electromagnetic spectrum. Perhaps the most notable of these databases is HITRAN, which was developed by the US Air Force Research Laboratory and is currently maintained at the Harvard-Smithsonian Center for Astrophysics in Cambridge, MA.

Liquids and Solids

The larger particulates in the atmosphere (greater than a few μm in radius) typically belong to clouds. Low-altitude clouds consist of nearly spherical water droplets, while high-altitude cirrus clouds are mainly a collection of ice crystals. Other large particulates include sand dust. Their light scattering is close to the geometric limit at visible and ultraviolet wavelengths. This means that the extinction is nearly wavelength-independent, and the scattering phase function and single-scattering albedo may be reasonably modeled with ray-tracing techniques that account for the detailed size and shape distributions of the particles. However, Mie scattering theory is typically used to calculate cloud optical properties because it is exact for spherical particles of any size.

The smaller particulates in the atmosphere belong to aerosols, which are very fine liquid particles, and dusts, which are solids such as minerals and soot. These particulates are concentrated mainly in the lower 2 km or so of the atmosphere; however, they are also present at higher altitudes in smaller concentrations. Their optical properties are typically modeled using Mie theory. The wavelength dependence of the scattering is approximately inversely proportional to a low power of the wavelength, typically between 1 and 2, as befits particulates intermediate in size between molecular and geometric-limit. The scattering phase functions have a strong forward-scattering peak; values of the asymmetry parameter g (the average value of $\cos\theta$) typically range from 0.6 to 0.8 at solar wavelengths.

Solution Methods

Geometry

Atmospheric properties are primarily a function of altitude, which determines pressure, temperature and species concentration profiles. Accordingly, most RT methods define a stratified atmosphere. The most accurate treatments of transmission and scattering account for the spherical shape of the layers and refraction; however, most RT models use a plane-parallel approximation for at least some computations, such as multiple scattering.

Spectral Resolution

Optical instruments have finite, and frequently broad, wavelength responses. Nevertheless, modeling their signals requires accounting for the variation of absorption on an extremely fine wavelength scale, smaller than the widths of the molecular lines.

'Exact' monochromatic methods

The most accurate RT solution method involves explicitly solving the RT problem for a very large number of monochromatic wavelengths. This line-by-line method is used in a number of RT models, such as FASCODE. It allows Beer's law to be applied to combine transmittances from multiple LOS segments, and provides an unambiguous definition of the optical parameters. It is suitable for use with spectrally structured light sources, such as lasers. The one major drawback of this method is that it is computationally intensive, and therefore may not be practical for problems where large wavelength ranges, multiple LOS views and multiple atmospheric conditions need to be treated.

To alleviate the computational burden of monochromatic calculations, some approximate methods have been developed that model RT in finite spectral intervals, as described below.

Statistical band models

The band model method represents spectral lines in a narrow interval, $\Delta\nu$, statistically using such parameters as the total line strength, the mean pressure-broadening parameter, and the effective number of lines in the interval. An example of a popular band model-based RT algorithm is MODTRAN, which is described below.

A key to the success of band models is the availability of approximate analytical formulas for the integrated absorption for an individual molecular transition of strength S, known as the single-line total equivalent width, W_{sl}:

$$W_{\mathrm{sl}} = \int_{-\infty}^{\infty} [1 - \exp(-Suf_\nu)]\mathrm{d}\nu \qquad [6]$$

In the optically thin (small absorption) limit, W_{sl} is proportional to the molecular species column density, while in the optically thick (large absorption) limit it scales as the square root of the column density. A further assumption made by MODTRAN is that the line centers are randomly located within the interval, i.e., spectrally uncorrelated. With this assumption, the net transmittance can be expressed as the product of the transmittances for each individual line, whether the line belongs to the same molecular species or a different species.

A particular challenge in band models is treatment of the inhomogeneous path problem – that is, the variations in path properties along a LOS and their effect on the statistical line parameters, which arise primarily from differences in pressure and hence line width. The Curtis–Godson path averaging method provides a reasonable way to define 'equivalent' homogeneous path parameters for the band model. Another challenge is to define an effective extinction optical depth for each layer in order to solve the RT problem with scattering. One option in MODTRAN is to define it by computing the cumulative transmittance through successive layers in a vertical path.

Correlated-*k* model

Another well-known approximate RT algorithm for spectral intervals is the correlated-*k* method. This method starts with an 'exact' line-by-line calculation of extinction coefficients (k's) within the interval on a fine spectral grid, from which a *k*-distribution (vs. cumulative probability) is computed. A database of *k* values and probabilities summing to 1 is built from these *k*-distributions for a grid of atmospheric pressures and temperatures, and for all species contributing to the spectral interval. Inhomogeneous paths are handled by recognizing that the size order of the k's is virtually independent of pressure and typically only weakly dependent on temperature. LOS transmittances, LOS radiances, and fluxes are calculated by interpolating the database over temperature and pressure to define the k's for each LOS segment, solving the monochromatic RT equation at each fixed distribution location, and finally integrating over the distribution. The correlated-*k* method has been found to be quite accurate for atmospheric paths containing a single molecular species; however, corrections must be applied for spectral intervals containing multiple species.

Scattering Methods

When the diffuse light field is of interest, scattering methods are used to calculate the source function J of eqns [1] and [2].

Single scattering

If scattering is weak, the approximation may be made that the solar radiation scatters only once. Thus the integral over the scattering phase function, eqn [2], is straightforwardly calculated using the direct radiance component, which is given by eqn [3]. The neglect of multiple scattering (i.e., the diffuse contribution to the source function) means that the diffuse radiance is underestimated; however, single scattering is sufficiently accurate for some atmospheric problems in clear weather and at infrared wavelengths.

Multiple scattering

A number of different methods have been developed to solve the multiple scattering problem. Two-stream methods, which are the simplest and fastest, resolve the radiance into upward and downward directions. These methods generally produce reasonably accurate values of hemispherically averaged radiance, which are also referred to as horizontal fluxes or irradiances.

A much more accurate approach to the multiple scattering problem is the method of discrete ordinates. It involves expansion of the radiation field, the phase function and the surface reflectance as a series of spherical harmonics, leading to a system of linear integral equations. Evaluation of the integrals by Gaussian quadrature leads to a solvable system of linear differential equations. An important approximation called delta-M speeds up convergence of the discrete ordinates method, especially when scattering phase functions are strongly forward peaked, by representing the phase function as the sum of a forward direction δ-function and a remainder term. For most scattering problems, the solution is converged upon with a modest number (~8 to 16) of quadrature points (streams).

A very different type of multiple scattering technique, called the Monte Carlo method, is based on randomly sampling a large number of computer-simulated 'photons' as they travel through the atmosphere and are absorbed and scattered. The basic idea here is that sensor radiance can be expressed as a multiple path integral over the local source terms, and Monte Carlo methods solve integrals by sampling the integrand. The major advantage of this method is that it is flexible enough to allow for all of the complexity of a realistic atmosphere, often neglected by other methods. The major drawback is its computational burden, as a large number of photons is required for reasonable convergence; the Gaussian error in the calculation declines with the square root of the number of photons. The convergence problem is moderated to a large extent by using the physics of the problem being solved to bias the selection of photon paths toward those trajectories which contribute most; mathematically, this is equivalent to requiring that the integrand be sampled most often where its contributions are most significant.

An Example Atmospheric RT Model: MODTRAN

MODTRAN, developed collaboratively by the Air Force Research Laboratory and Spectral Sciences, Inc., is the most widely used atmospheric radiation transport model. It defines the atmosphere using stratified layering and computes transmittances, radiances, and fluxes using a moderate spectral resolution band model with IR through UV coverage. The width of the standard spectral interval, or bin, in MODTRAN is 1 cm^{-1}. At this resolution, spectral correlation among extinction sources is well characterized as random. Thus, the total transmittance from absorption and scattering of atmospheric particulates and molecular gases is computed as the product of the individual components.

Rayleigh, aerosol, and cloud extinction are all spectrally slowly varying and well represented by Beer's Law absorption and scattering coefficients on a 1 cm^{-1} grid. Calculation of molecular absorption is more complex because of the inherent spectral structure and the large number of molecular transitions contributing to individual spectral bins. As illustrated in Figure 2, MODTRAN partitions the spectral bin molecular attenuation into 3 components:

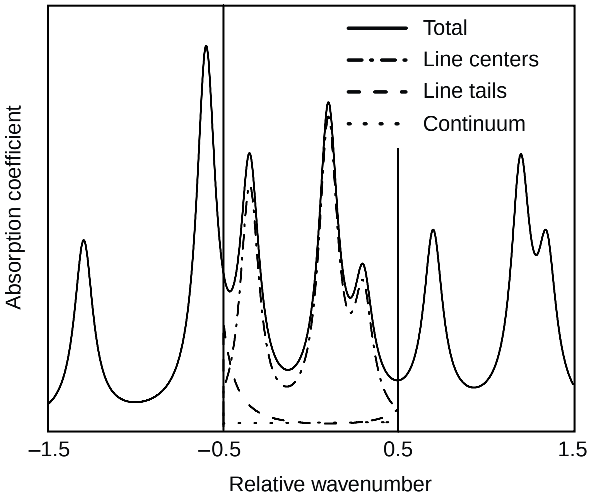

Figure 2 Components of molecular absorption. The line center, line tail and continuum contributions to the total absorption are illustrated for the central 1 cm^{-1} spectral bin.

- *Line center absorption* from molecular transitions centered within the spectral bin;
- *Line tail absorption* from the tails of molecular lines centered outside of the spectral bin but within 25 cm^{-1}; and
- H_2O and CO_2 *continuum absorption* from distant (>25 cm^{-1}) lines.

Within the terrestrial atmosphere, only H_2O and CO_2 have sufficient concentrations and line densities to warrant inclusion of continuum contributions. These absorption features are relatively flat and accurately modeled using 5 cm^{-1} spectral resolution Beer's Law absorption coefficients.

Spectral bin contributions from neighboring line tails drop off, often rapidly, from their spectral bin edge values, but the spectral curves are typically simple, containing at most a single local minimum. For MODTRAN, these spectral contributions are pre-computed for a grid of temperature and pressure values, and fit with Padé approximants, specifically the ratio of quadratic polynomials in wavenumber. These fits are extremely accurate and enable line tail contributions to be computed on an arbitrarily fine grid. MODTRAN generally computes this absorption at a resolution equal to one-quarter the spectral bin width, i.e., 0.25 cm^{-1} for the 1.0 cm^{-1} band model.

The most basic *ansatz* of the MODTRAN band model is the stipulation that molecular line center absorption can be approximated by the absorption of n identical Voigt lines randomly distributed within the band model spectral interval, $\Delta\nu$. Early in the development of RT theory, Plass derived the expression for the transmittance from these n randomly distributed lines:

$$T = \left(1 - \frac{W'_{\mathrm{sl}}}{\Delta\nu}\right)^n \qquad [7]$$

MODTRAN's evolution has resulted in a fine tuning of the methodology used to define both the effective line number n and the in-band single-line equivalent width W'_{sl}. The effective line number is initially estimated from a relationship developed by Goody, in which lines are weighted according the to square root of their strength, but MODTRAN combines nearly degenerate transitions into single lines because these multiplets violate the random distribution assumption. The initial effective line number values are refined to insure a match with higher resolution transmittance predictions degraded to the band model resolution. The in-band equivalent width is computed for an off-centered Voigt line of strength *S*. Lorentz and Doppler half-widths are determined as strength-weighted averages. The off-center distance is fixed to insure that the weak-line Lorentz equivalent width exactly equals the random line center value.

MODTRAN scattering calculations are optimally performed using the DISORT discrete ordinates algorithm developed by Stamnes and co-workers. Methods for computing multiple scattering such as DISORT require additive optical depths, i.e., Beer's Law transmittances. Since the in-band molecular transmittances of MODTRAN do not satisfy Beer's Law, MODTRAN includes a correlated-*k* algorithm option. The basic band model *ansatz* is re-invoked to efficiently determine *k*-distributions; tables of *k*-data are pre-computed as a function of Lorentz and Doppler half-widths and effective line number, assuming spectral intervals contain *n* randomly distributed identical molecular lines. Thus, MODTRAN *k*-distributions are statistical, only dependent on band model parameter values, not on the exact distribution of absorption coefficients in each spectral interval.

Applications

Among the many applications of atmospheric transmission and scattering calculations, we briefly describe two complementary ones in the area of remote sensing, which illustrate many of the RT features discussed earlier as well as current optical technologies and problems of interest.

Earth Surface Viewing

The first example is Earth surface viewing from aircraft or spacecraft with spectral imaging sensors. These include hyperspectral sensors, such as AVIRIS, which have typically a hundred or more contiguous spectral channels, and multispectral sensors, such as Landsat, which typically have between three and a few tens of channels. These instruments are frequently used to characterize the surface terrain, materials and properties for such applications as mineral prospecting, environmental monitoring, precision agriculture, and military uses. In addition, they are sensitive to properties of the atmosphere such as aerosol optical depth and column water vapor. Indeed, in order to characterize the surface spectral reflectance, it is necessary to characterize and remove the extinction and scattering effects of the atmosphere.

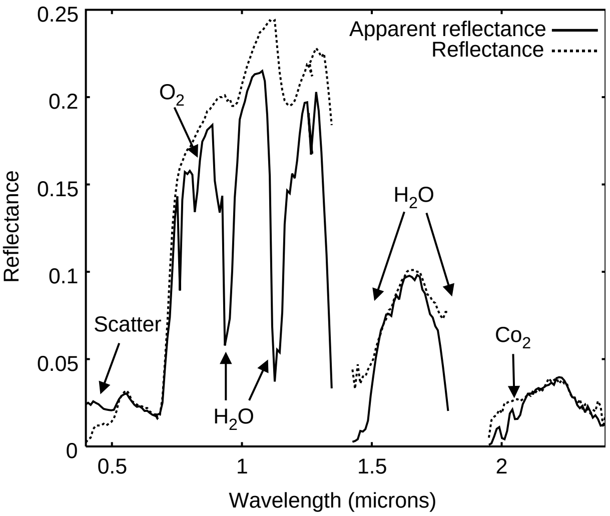

Figure 3 Vegetation spectrum viewed from 3 km altitude before (apparent reflectance) and after (reflectance) removal of atmospheric scatter and absorption.

Figure 3 also shows an example of data collected by the AVIRIS sensor at ~3 km altitude over thick vegetation. The apparent reflectance spectrum is the observed radiance divided by the Top-of-Atmosphere horizontal solar flux. Atmospheric absorption by water vapor, oxygen, and carbon dioxide is evident, as well as Rayleigh and aerosol scattering. Figure 3 shows the surface reflectance spectrum inferred by modeling and then removing these atmospheric effects (this process is known as atmospheric removal, compensation, or correction). It has the characteristic smooth shape expected of vegetation, with strong chlorophyll absorption in the visible and water bands at longer wavelengths. A detailed analysis of such a spectrum may yield information on the vegetation type and its area coverage and health.

Sun and Sky Viewing

The second remote sensing example is sun and sky viewing from the Earth's surface with a spectral radiometer, which can yield information on the aerosol content and optical properties as well as estimates of column concentrations of water vapor, ozone, and other gases. Figure 4 shows data from a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer, which measures both 'direct flux' (the direct solar flux divided by the cosine of the zenith angle) and diffuse (sky) flux in narrow wavelength bands. The plot of ln(direct signal) versus the air mass ratio is called a Langley plot, and is linear for most of the bands, illustrating Beer's Law. The extinction coefficients (slopes) vary with wavelength consistent with a combination of Mie and Rayleigh scattering. The water-absorbing 940 nm band has the lowest values and a curved

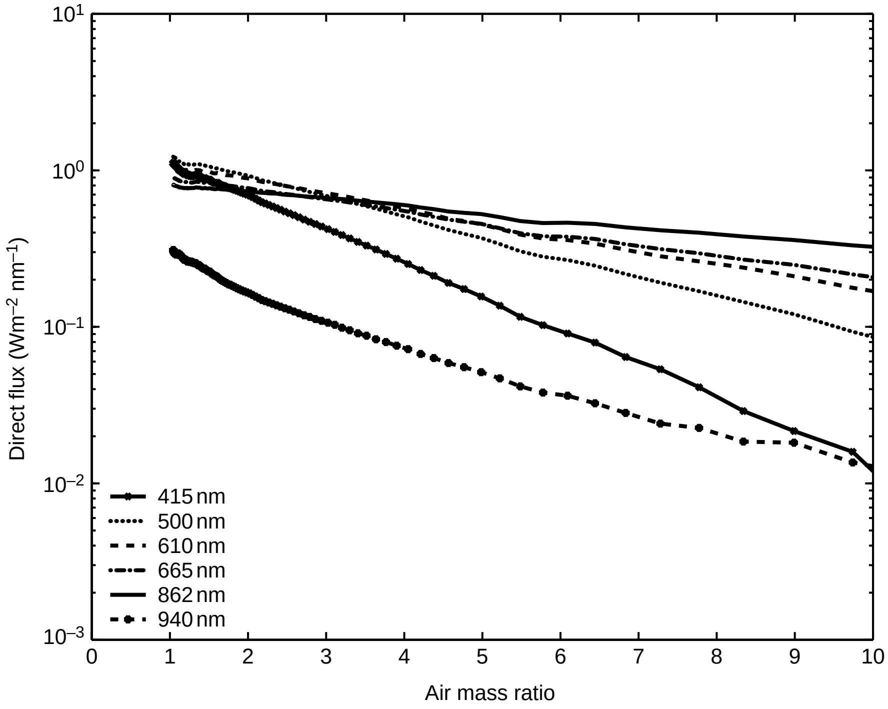

Figure 4 Direct solar flux versus air mass at the surface measured by a Yankee Environmental Systems, Inc. multi-filter rotating shadow-band radiometer at different wavelengths.

plot, in accordance with the square-root dependence of the equivalent width for optically thick lines. The diffuse fluxes, which arise from aerosol and Rayleigh scattering, have a very different dependence on air mass than the direct flux. In particular, the diffuse contributions often increase with air mass for high sun conditions.

The ratio of the diffuse and direct fluxes is related to the total single-scattering albedo, which is defined as the ratio of the total scattering coefficient to the extinction coefficient. Results from a number of measurements showing lower than expected diffuse-to-direct ratios have suggested the presence of black carbon or some other continuum absorber in the atmosphere, which would have a significant impact on radiative energy balance at the Earth's surface and in the atmospheric boundary layer.

List of Units and Nomenclature

Line shape function [cm]	f_ν
Lorentz line shape function [cm]	$f_{\mathrm{Lorentz}}(\nu)$
Voigt line shape function [cm]	$f_{\mathrm{Voigt}}(\nu)$
Scattering asymmetry parameter	g
Direct plus diffuse radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I
Direct radiance (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'
Direct radiance at the LOS boundary (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	I'_0
Source function (W cm^{-1} sr^{-1} or W cm^{-2} μm^{-1} sr^{-1})	J
Extinction coefficient (cm^{-1} atm^{-1})	k
Effective number of lines in bin	n
Scattering phase function (sr^{-1})	p
Line strength (cm^{-2} atm^{-1})	S
Transmittance	T
Column density (atm cm)	u
Complex error (probability) function	w
Single-line total equivalent width (cm^{-1})	W_{sl}
Single-line in-band equivalent width (cm^{-1})	W'_{sl}
Collision-broadened half-width at half maximum (cm^{-1})	α_{c}
Doppler half-width at 1/e (cm^{-1})	α_{d}
Spectral interval (bin) width (cm^{-1})	$\Delta\nu$
Scattering angle (radian)	θ
Wavelength (μm)	λ
Wavenumber (cm^{-1})	ν
Molecular line transition frequency	ν_0
Direction of incoming light (sr)	Ω_{i}
Direction of outgoing light (sr)	Ω_{o}

See also

Environmental Measurements: Laser Detection of Atmospheric Gases. **Instrumentation:** Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Berk A, Anderson GP, Acharya PK, Bernstein LS, Shettle EP, Adler-Golden SM, Lee J and Muratov L (2004) MODTRANS User's Manual. Airforce Research Laboratory, Hanscom AFB, MA, in press.

Chandrasekhar S (1960) *Radiative Transfer*. New York: Dover Publications, Inc., Originally published by Oxford University Press, London (1950).

Fenn RW, Clough SA, Gallery WO, Good RE, Kneizys FX, Mill JD, Rothman LS, Shettle EP and Volz FE (1985) Optical and infrared properties of the atmosphere. In: Jursa AS (ed.) *Handbook of Geophysics and the Space Environment*, chap. 18, US Air Force Geophysics Laboratory, Hanscom AFB, MA, AFGL-TR-85-0315, available from NTIS ADA 167000.

Goody RM and Yung YL (1989) *Atmospheric Radiation, Theoretical Basis*, 2nd edn. New York: Oxford University Press.

HITRAN Special Issue, J. Quant. Spectrosc. Rad. Transfer, vol. 82, Nos. 1–4, 15 November–15 December (2003).

Houghton J (2001) *The Physics of Atmospheres*, 3rd edn, Cambridge, UK: Cambridge University Press.

Killinger DK, Churnside JH and Rothman LS (1994) Atmospheric optics. In *Handbook of Optics*, 2nd edn, 2-Volume Set, chap. 44. New York: McGraw-Hill.

Liou KN (1997) *An Introduction to Atmospheric Radiation*, 2nd edn. London: Academic Press.

Smith FG (ed.) (1993) *The Infrared and Electro-Optical Systems Handbook*, vol. 2, *Atmospheric Propagation of Radiation*. Ann Arbor: Environmental Research Institute of Michigan.

Thomas GE and Stamnes K (1999) *Radiative Transfer in the Atmosphere and Ocean*. Cambridge, UK: Cambridge University Press.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover Publications, Inc., Originally published in 1957, New York: John Wiley & Sons, Inc.

F

FIBER AND GUIDED WAVE OPTICS

Contents

Overview
Dispersion
Fabrication of Optical Fiber
Light Propagation
Measuring Fiber Characteristics
Nonlinear Effects (Basics)
Nonlinear Optics
Optical Fiber Cables
Passive Optical Components

Overview

A Mickelson, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is said that some in ancient Greece knew that light could be guided inside slabs of transparent material. From this purported little-known curiosity in the ancient world, guided wave optics has grown to be the technology of the physical level of the systems which transfer most of the information about the world; the waveguide of the worldwide telecommunications network is the optical fiber. The optical fiber itself is but a passive waveguide, and guided wave optics is the technology which includes all of the passive and active components which are necessary to prepare optical signals for transmission, regenerate optical signals during transmission, route optical signals through systems and code onto optical carriers, and decode the information from optical carriers, to more conventional forms. In this article, some introduction to this rather encompassing topic will be given.

This article will be separated into four parts. In the first section, discussion will be given to fiber optics, that is, the properties of the light guided in optical waveguides which allow the light to be guided in distinctly nonrectilinear paths over terrestrial distances. The second section will then turn to the components which can be used along with the fiber optical waveguides in order to form useful systems. These components include sources and detectors as well as optical amplifiers. In the third section, we will discuss the telecommunications network which has arisen due to the availability of fiber optics and fiber optic compatible components. The closing section will discuss integrated optics, the field of endeavor which has such great promise to form the future of optical technology.

Fiber Optics

Fiber optics is a term which generally refers to a technology in which light (actually infrared, visible, or ultraviolet radiation) is transmitted through the transparent cores of small threads of composite material. These threads, or fibers as they are called, when surrounded by a cladding material and coated with polymer for environmental protection (see Figure 1) can be coiled like conventional wire and

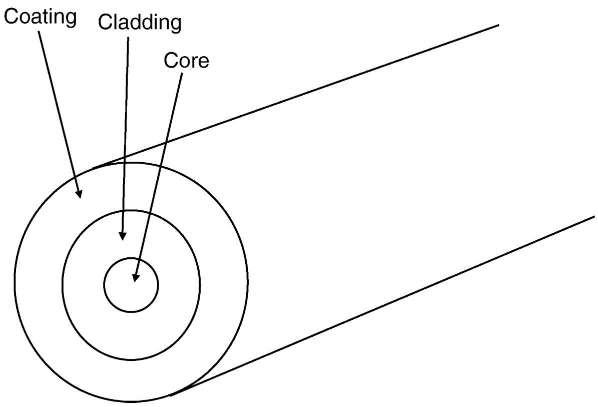

Figure 1 A depiction of the structure of an optical fiber. The innermost cylinder is the core region in which the majority of the light is confined. The next concentric region is the cladding region which is still made of a pure material but one of a lower index of refraction than the innermost core region, such that the light of the lightwave decays exponentially with extension into this region. The outermost region is a coating which protects the fused silica of the core and cladding from environmental contaminants such as water. For telecommunications fibers, the core is comprised of fused silica doped with germanium and typically has an index of refraction of 1.45, whereas the cladding differs from this index by only about 1%. The coating is often a polyimide plastic which has a higher index (perhaps 1.6) but no light guided light should see the cladding-coating interface.

when cabled can resemble (a very lightweight flexible version of) conventional transmission wire. Although most of the optical fiber in use today is fabricated by a process of gas phase chemical deposition of fused silica doped with various other trace chemicals, fiber can be made from a number of different material systems and in a number of different configurations for use with various types of sources. In what follows in this opening section, we will limit discussion to the basic properties of the light guided by the fiber and leave more technological discussion to following sections.

There are two complementary mathematical descriptions of the propagation of light in an optical waveguide. In the ray description, light incident on a fiber endface is considered to be made up of a bundle of rays. In a uniform homogeneous medium, each ray is like an arrow that exhibits rectilinear propagation from its source to its next interface with a dissimilar material. These rays satisfy Snell's laws of reflection and refraction at interfaces between materials with dissimilar optical properties that they encounter along their propagation path. That is, at an interface, a fraction of the light is reflected backwards at an angle equal to the incident angle and a portion of the light is transmitted in a direction which is more directed toward the normal to the interface when the index increases across the boundary and is directed more away from the normal when the index decreases. At the input endface of a waveguide, a portion of the energy guided by each ray is refracted due to the change in refractive index at the guide surface and then exhibits a more interesting path within the fiber. In a step index optical fiber, where the index of refraction is uniformly higher in the fiber core than in a surrounding cladding, the rays will propagate along straight paths until encountering the core cladding interface. Guided rays (see Figure 2) are totally internally reflected back into the fiber core to again be totally internally reflected at the next core cladding interface and so on. Radiating rays (see Figure 3) will be only partially reflected at the core cladding interface and will rapidly die out in propagating down the fiber when it is taken into account that typical distances between successive encounters with the boundary may be sub-millimeter and total propagation distances may be many kilometers. In graded index fibers, the refractive index within the fiber core varies continuously from a maximum somewhere within the core to a minimum at which the core ends and attaches continuously to a cladding. The ray paths within such fibers are curved and the guided rays are characterized by the fact that once in the fiber they never encounter the cladding. Radiating rays encounter the cladding and are refracted out of the fiber. This description of fiber propagation is quite simple and pleasing but does not take into account that each ray actually is carrying a clock that remembers how long it has been following its given path. When two rays come together, they can

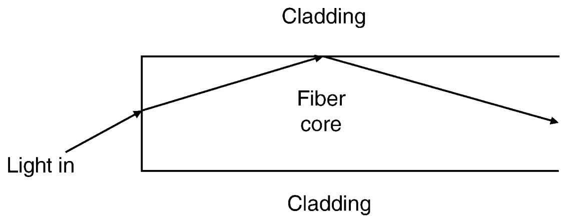

Figure 2 Schematic depiction of the geometrical optics interpretation of the coupling of light into an optical fiber. Here a ray that is incident on the center of the fiber core at an upward angle is first refracted at the fused silica–air interface into the fused silica core. The ray is then totally internally reflected at the core cladding interface such that the ray power remains in the core. The ray picture of light coupling is applicable to single rays coupling to a multimode fiber. When there are multiple rays, interference between these rays must be taken into account which is difficult to do in the ray picture. When there is only one (diffraction limited) mode in the fiber, the interference between a congruence of (spatially coherent) incident rays is necessary to describe the coupling. This description is more easily effected by a quasi-monochromatic mode picture of the propagation.

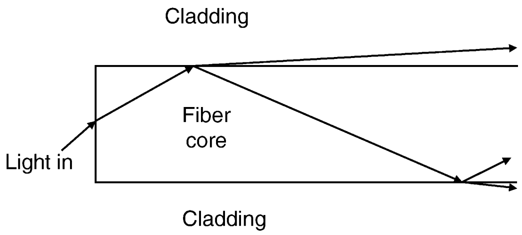

Figure 3 Schematic depiction of the existence of a cut-off angle for coupling into an optical fiber. A ray incident from air on the center of the fused silica fiber core is refracted at the air fused silica interface and then split when it comes to the core cladding interface within the fiber. Were the refracted angle of this ray zero (along the core cladding interface) we would say the ray is totally internally reflected and no power escapes. As it is, there is refracted power into the cladding and the ray will rapidly attenuate as it propagates along the fiber. Although most telecommunications fiber in this day and age is graded index (the rays will curve rather than travel along straight paths), a typical repeat period in such a fiber is a millimeter, whereas propagation distances may be 100 km. With a loss of even only 0.1% at each reflection form the cladding, the light light at cut-off will be severely attenuated in one meter, no loss, one kilometer.

either add or subtract, depending on the reading on their respective clocks. When the light is nearly monochromatic, the clocks' readings are simply a measure of the phase of the ray when it was radiated from its source and the interference between rays can be quite strong. This interference leads to conditions which only allow certain ray paths to be propagated. When the light has a randomly varying phase in both space and time (and is therefore polychromatic), the interference all but disappears. But the condition that allows coupling to a specific ray path (mode) also is obliterated by the phase randomness, and coupling to a single mode guide, for example, becomes inefficient.

When we need to preserve phase relations in order to preserve information while it propagates along a fiber path, we need to use single modes excited by nearly transform limited sources, that is sources whose time variation exhibits well-defined (nonrandom) phase variation. As a monochromatic wave carries no information, we will require a source which can be modulated. When modulation can be impressed on a carrier without loss of phase information of either information or carrier, information can be propagated at longer distances than when the frequency spectrum of the carrier is broadened by noise generated during the modulation. We refer to such a source that can be modulated without broadening as a coherent or nearly coherent source. We cannot easily adapt the ray picture to the description of propagation of such coherent radiation in an optical fiber. Instead we must resort to using the time harmonic form of Maxwell's equations, the equations which describe electromagnetic phenomena, whereas rays can be described by a simplification of the time-dependent form of these equations. In the time harmonic approach, we assume that the source is monochromatic and then solve for a set of modes of the fiber at the assumed frequency of the source. These modes have a well-defined phase progression as a function of carrier frequency and possess a given shape in the plane transverse to the direction of propagation. Information can be included in the propagation by assuming the quasi-monochromatic variation of the source, that is, one assumes that the source retains the phase relations between the various modulated frequency components even while its amplitude and phase is being varied externally. When one assumes that time harmonically (monochromatic) varying fields are propagating along the fiber axis, one obtains solutions of Maxwell's equations in the form of a summation of modes. These are guided modes, ones that propagate down the fiber axis without attenuation. There are also radiation modes that never couple into a propagating mode in the fiber. These modes are analogous to the types of rays we see in the ray description of fiber propagation. There are also another set of modes which are called evanescent modes which show up at discontinuities in the fiber or at junctions between the fiber and other components. In the modal picture of fiber propagation, each of these modes is given an independent complex coefficient (that is, a coefficient with both amplitude and phase). These coefficients are determined first by any sources in the problem and then must be recalculated at each discontinuity plane along the propagation path in the fiber. When the source(s) in the problem is not transform limited, the phases of the coefficients become smeared out and the coupling problems at discontinuities take on radically different solutions.

In some limit, these randomized solutions must appear as the ray solutions. Optical fibers are generally characterized by their numerical aperture (NA), as well as the number which characterizes its transverse dimension. The numerical aperture is essentially the sine of the maximum angle into which the guide will radiate into free space. When the guide is multimoded, then the transverse dimension is generally given as a diameter. When a guide is single-moded, the transverse dimension is generally given as a spotsize, that is, by a measure of the size of a unity magnification image of the fiber endface. Multimode guides can be excited by even poorly coherent sources so long as they radiate into the NA or capture angle at the input of the guide. Single-mode fibers require an excitation which matches the shape and size of their fundamental mode.

Active Fiber Compatible Components

Our discussion of propagation in fiber optic waveguides would be a rather sterile one if there were not a number of fiber compatible components that can generate, amplify, and detect light streams available to us in order to construct fiber optic systems. In this section, we will try to discuss some of the theory of operation of active optical components in order to help elucidate some of the general characteristics of such components and the conditions that inclusion of such components in a fiber system impose on the form that the system must take.

In a passive component such as an optical fiber, we need only consider the characteristics of the light that is guided. Operation of active components requires that the optical fields interact with a medium in such a manner that energy can be transferred from the field through an excitation of the medium to the controlling electrical stream and/or vice versa. Generally one wants the exchange to go only one way, that is from field to electrical stream or from electrical stream to field. In a detector this is not difficult to achieve, whereas in lasers and amplifiers it is quite hard to eliminate the back reaction of the field on the device. Whereas a guiding medium can be considered as a passive homogeneous continuum, the active medium has internal degrees of freedom of its own as well as a granularity associated with the distribution of the microscopic active elements. At least, we must consider the active medium as having an active index of refraction with a mind (set of differential equations anyway) of its own. If one also wants to consider noise characteristics, one needs to consider the lattice of active elements which convert the energy. The wavelength of operation of that active medium is determined by the energy spacing between the upper and lower levels of a transition which is determined by the microscopic structure of these grains, or quanta, that make up the medium. In a semiconductor, we consider these active elements to be so uniformly distributed and ideally spaced that we can consider the quanta (electron hole pairs) to be delocalized but numerous and labeled by their momentum vectors. In atomic media such as the rare earth doped optical fibers which serve as optical amplifiers, we must consider the individual atoms as the players. In the case of the semiconductor, a current, flowing in an external circuit, controls the population of the upper (electronic) and lower (hole) levels of each given momentum state within the semiconductor. When the momentum states are highly populated with electrons and holes, there is a flow of energy to the field (at the transition wavelength) and when the states are depopulated, the field will tend to be absorbed and populate the momentum states by giving up its energy to the medium. The situation is similar with the atomic medium except that the pump is generally optical (rather than electrical) and at a different wavelength than the wavelength of the field to be amplified. That is to say, one needs to use a minimum of three levels of the localized atoms in the medium in order to carry out an amplification scheme.

The composition of a semiconductor determines its bandgap, that is, the minimum energy difference between the electron and hole states of a given momentum value. A source, be it a light emitting diode (LED) or laser diode (see Figure 4) will emit light at a wavelength which corresponds to an energy slightly above the minimum gap energy (wavelength equals the velocity of light times Planck's constant divided by energy) whereas a detector can detect almost any energy above the bandedge. Only semiconductors with bandgaps which exhibit a minimum energy at a zero momentum transfer can be made to emit light strongly. Silicon does not exhibit a direct gap and although it can be used as a detector, it cannot be used as a source material. The silicon laser is and has been the 'Holy Grail' of electronics because of the ubiquity of silicon electronics. To even believe that a 'Holy Grail' exists requires a leap of faith. Weak luminescence has been

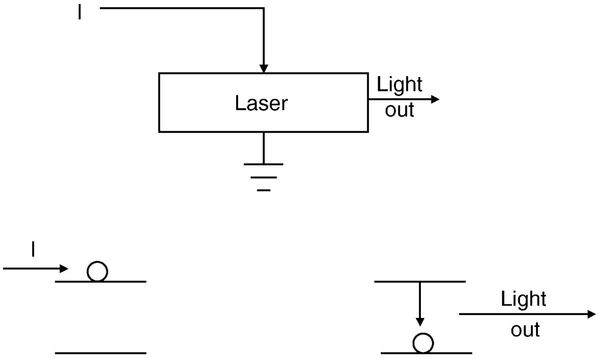

Figure 4 A schematic depiction of the workings of a semiconductor laser light source. The source is fabricated as a diode, and normal operation of this diode is in forward bias, that is, the p-side of the junction is biased positively with respect to ground which is attached to the n-side of the junction. With the p-side positively biased, current should freely flow through the junction. This is not quite true as there is a heterojunction region between the p- and n-regions in which electrons from the n-side may recombine with holes from the p-side. This recombination gives off a photon which is radiated. In light emitting diodes (LEDs), the photon simply leaves the light source as a spontaneously emitted photon. In a laser diode, the heterojunction layer serves as a waveguide and the endfaces of the laser as mirrors to provide feedback and allow laser operation in which the majority of the photons generated are generated in stimulated processes.

observed in certain silicon structures. This luminescence has been used for mid-infrared sources (where thermal effects mask strong luminescence) and in concert with rare earth dopants to make visible light emitting diodes. The silicon laser is still only a vision.

Source materials are, therefore, all made of compound semiconductors. Detectors are in essence the inverse of sources (see Figure 5) but also must detect at wavelengths above their gap energy. For primarily historical reasons, as will be further discussed below, telecommunications employs essentially three windows of wavelengths for different applications, wavelengths centered about 0.85 microns in the first window, wavelengths about 1.3 microns in the second window, and wavelengths about 1.55 microns in the third window. Materials made of layers of different compositions of AlGaAs and mounted on GaAs substrates are used in the first band, while the other bands require the increased degree of freedom allowed by the quaternary alloys of InGaAsP mounted on InP substrates. Other material mixes are possible, these above-mentioned materials are the most common. Rare earth ions exhibit many different transitions. The most useful ones have proven to be the transitions of Er in the third telecommunications window, the one that covers the spectral region to either side of 1.55 microns. Although both Nd and Pr can be made to amplify near the 1.3 micron window, amplifiers made of these materials have not proven to be especially practical.

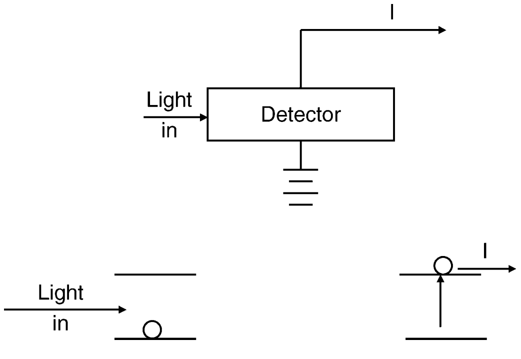

Figure 5 A schematic depiction of the workings of a semiconductor detector. As with a semiconductor source, the detector is fabricated to operate as a diode, that is a p-n junction, but in the case of the detector this diode is to be operated only in reverse bias, that is, with the n-material being biased positively with respect to the ground which is attached to the p-material's contact. In this mode of operation, no current should flow through the junction as there are no free carriers in the material, they have all been drawn out through the contacts. A photon of sufficiently high energy can change this situation by forming an electron hole pair in the material. The electron and hole are thereafter drawn to the n- and p-contacts, respectively, causing a detectable current to flow in the external circuit.

Telecommunications Technology

The physical transmission layer of the worldwide telecommunications network has come to be dominated by fiber optic technology, in particular, by wavelength division multiplexed or WDM single-mode fiber optics operating in the 1.55 micron telecommunications window. In the following, we will first discuss how this transformation to optical communications took place and then go on to discuss some of the specifics of the technology.

Already by the middle of the 1960s, it was clear that changes were going to have to take place in order that the exponential growth of the telephone system in the United States, as well as in Europe, could continue. The telephone system then, pretty much as now, consisted of a hierarchy of tree structures connected by progressively longer lines. A local office is used to connect a number of lines emanating in a tree structure to local users. The local offices are connected by trunk lines which emanate from a toll office. The lines from there on up the hierarchy are long distance ones which are termed long lines and can be regional and longer. The most pressing problem in the late 1960s was congestion in the so-called trunk lines which connect the local switching offices. The congestion was naturally most severe in urban areas. These trunk lines were generally one kilometer in length at that time. The problem was that there was no more space in the ducts that housed these lines. A solution was to use time division multiplexing or TDM to increase the traffic that could be carried by each of the lines already buried in the conduit. The problem was that the twisted pair lines employed would smear out the edges of the time varying bit streams carrying the information at the higher bitrates (aggregated rates due to the multiplexing) due to the inherent dependence of signal propagation velocity on frequency known as dispersion. After discussion and even development of a number of possible technologies, in 1975 a demonstration of a fiber optic system which employed multimode semiconductor lasers feeding multimode optical fibers all operating at 0.85 micron wavelength, proved to be the most viable model for trunk line replacement. The technology was successful and already in 1980 advances in single-mode laser and single-mode fiber technology operating at the 1.3 micron wavelength had made the inclusion of fiber into the long lines viable as well. For roughly the decade from 1985 onward, single-mode fiber systems dominated long line replacement for terrestrial as well as transoceanic links. The erbium doped fiber amplifier which operated in the 1.55 micron wavelength third telecommunication window had proven itself to be viable for extending repeater periods by around 1990. Development of the InGaAsP quaternary semiconductor system allowed for reliable lasers to be manufactured for this window as development of strained lattice laser technology in the InGaAs system allowed for efficient pump lasers for the fiber amplifiers. As the optical amplifiers could amplify across the wavelength band that could be occupied by many aggregated channels of TDM signals, the move of the long line systems to the third telecommunications window was accompanied by the adoption of wavelength division multiplexing or WDM. That is, electronics becomes more expensive as the analog bandwidth it can handle increases. Cost-effective digital rates are now limited to a few Gb/s, and analog rates to perhaps 20 GHz. Optical center frequencies are in the order of 2×10^{14} Hz and purely optical filters can be made to cleanly separate information signals spaced as closely as 100 GHz apart on optical carriers. An optical carrier can then be made to carry hundreds of channels of 10 Gb/s separated by 100 GHz and this signal can be propagated thousands of kilometers through the terrestrial fiber optics network. Such is today's Worldwide Web.

Integrated Optics as a Future of Guided Wave Optics

As electronics has been pushed to ever greater levels of integration, its power has increased and its price has dropped. A goal implicit in much of the development of planar waveguide technology has been that optics could become an integrated technology in the same sense as electronics. This has not occurred and probably will not occur. This is not to say that integrated optical processing circuits are not being developed and that they will be not employed in the future. They most definitely will.

Integrated optics was coined as a term when it was used in 1969 to name a new program at Bell Laboratories. This program was aimed at investigation of all the technologies that were possible with which to fabricate optical integrated circuits. The original effort was in no way driven by fiber optics as the Bell System only reluctantly moved to a fiber optic network solution in 1975. The original integrated optics efforts were based on such technologies as the inorganic crystal technologies that allowed for large second-order optical nonlinearities, a necessity in order that a crystal also exhibits a large electro-optic coefficient. An electro-optic coefficient allows one to change the index of refraction of a crystal by applying a low frequency or DC electromagnetic field across the crystal.

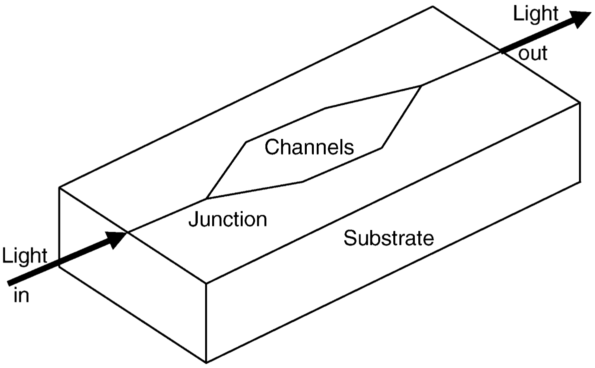

Figure 6 A schematic depiction of an integrated optical device which is often used as a high speed modulator. The device is an integrated version of a Mach–Zehnder interferometer. The basic idea behind its operation is that spatially and temporally coherent light is input into a single mode optical channel. That channel is then split into two channels by a Y-junction. Although not depicted in the figure, the two arms should not have completely equivalent propagation paths. That is to say, that if the light in one of those paths propagates a slightly longer distance (as measured in terms of phase fronts of the wave) then one cannot recombine the power from the two arms. That is, in the second Y-junction, a certain amount of the light which we are trying to combine from the two arms will be radiated out from the junction. In fact, if the light were temporally incoherent (phase fronts not well defined), exactly half of the light would be radiated form the junction. This is a statement of the brightness theorem which was thought to be a law of propagation before there were coherent sources of light. In a high speed modulator, the substrate is an electro-optic crystal and electrodes are placed over the two arms which apply the electrical signal to be impressed on the optical carrier to the channels.

Lithium niobate technology is a technology that has lasted to the present as an optical modulator technology (see Figure 6) and an optical switch technology (see Figure 7), as well as a technology for parametric wavelength conversion. Unfortunately, although low loss passive waveguide can be fabricated in lithium niobate, the crystal is not amenable to any degree of monolithic integration. Glass was also investigated as an integration technology from the early days of integrated optics, but the lack of second-order nonlinearity in glass strictly limited its applicability. Attention for a period turned to monolithic integration in the semiconductor materials which were well progressing as for use as lasers, detectors, and integrated circuits. Semiconductor crystals, however, are too pure to allow for low loss light propagation which requires the material defects to be so numerous that optical wavelengths cannot sample them. Passive waveguides in semiconductors incur huge losses that can only be mitigated by almost constant optical amplification which, unfortunately, cannot track the rapid optical field variations necessary for information transfer. Early on, attention turned to silicon

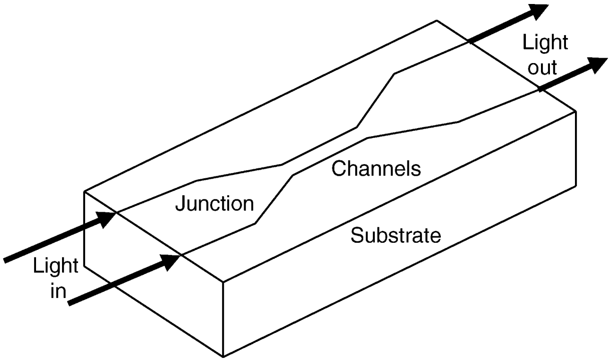

Figure 7 A schematic depiction of an integrated optical device which is often used as an optical switch. This device is interferometric but really has no free space optics counterpart as the Mach–Zhender interferometer does. In this interferometer, spatially and temporally coherent light is input into a single mode input channel. The two single optical mode input channels, which are initially so far apart from each other in terms of their mode sizes that they are effectively uncoupled, are then brought ever closer together until they are strongly coupled and their modes are no longer confined to their separate waveguides but are shared between the two channels. As the channels are symmetric, but the initial excitation, if from a single waveguide alone, is not, the pattern must evolve with propagation. Were both input channels excited, symmetry would still require a symmetry of their phases. If the phases were either equal to each other or completely out of phase, then there would be no evolution because we would have excited coupler modes. If the coupled region is the proper length, light input to one channel will emerge from the other after the channels are pulled apart. If the substrate is an electro-optic crystal and electrodes are placed over the channels, application of a voltage can affect the evolution of the interference in such a manner so as to 'decouple' the channels and switch the light back to the original channel.

optical bench technology, a technology which has presently spawned micro-electro-mechanical-system (MEMs) technology and still persists as an expensive vehicle to hybrid electro-optical integration. In silicon bench, individual components are micromounted with solder or silicon pop-up pieces on a silicon wafer which serves as a micro-optical bench. The technology is expensive and hybrid but pervasive in the highest end of cost classes of optical systems.

The tremendous success of the fiber optical network spawned a movement in which groups tried to achieve every possible system functionality in fibers themselves. A notable success has been the optical fiber amplifier. Significantly less successful in the fiber optic network has been the fiber Bragg grating. Attempts to carry out WDM functionalities in all fiber configuration leads to unwieldy configurations requiring large amounts of fiber and significant propagation delays, although, Bragg grating sensors have proven to be somewhat useful in sensing applications. WDM functions have been implemented in a large number of different hybrid configurations, involving various types of glasses, fused silica on silicon and simple hybrids of birefringent crystals.

The telecommunications network is not the driver that it once was. That this is so is perhaps the strongest driving force yet for integrated optics. The driver now is new applications that must be implemented in the most efficient manner. The most efficient manner is quite generally the one in which there is a maximum degree of integration. In the long run, polymer has always been the winning technology for optics of any kind, due to the flexibility of the technology and the drop in cost that accompanies mass production. Indeed, polymer integrated optics progresses. There are also a number of researchers involved in investigating strongly guiding optics, so-called photonic crystal optics. That is, in order to achieve low loss as well as low cost, the fiber and integrated optic waveguides up to the present have used small variations in optical properties of material to achieve guidance at the cost of having structures that are many wavelengths in size. In radio frequency (RF) and microwave technology, for example, guides are tiny fractions of a wavelength. This can lead to an impedance matching problem. RF and microwave impedance matching costs (versus circuit impedance matching costs) are generally quite high, precluding mass application. That radio frequency systems, such as cell phone transceivers link to the rest of the world by antenna, allows that the overall circuit dimension can be kept less than a single wavelength and impedance matching can be foregone in this so-called circuit limit. It is usually hard to keep an overall microwave system to less than a wavelength in overall extent except in the cell phone case or in a microwave oven. Optical miniaturization will require high index contrast guides, and will require optical impedance matching. There are few complete optical systems which will comprise less than an optical wavelength in overall extent. But then this so-called photonic crystal technology will likely be polymer compatible and there may be ways to find significant cost reduction. The future of integrated optics is unclear but bright.

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Nonlinear Optics; Optical Fiber Cables; Passive Optical Components. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1995) *Fiber-Optic Communication Systems*. New York: Wiley.

Bette, DeMarchis G and Iannone (1995) *Coherent Communications*. New York: Wiley.

Haken H (1985) *Light*. New York: Elsevier North-Holland.

Hecht E (1998) *Optics*. Reading, MA: Addison-Wesley.

Klein MV and Furtak TE (1986) *Optics*. New York: Wiley.

Marcuse D (1982) *Light Transmission Optics*. New York: Van Nostrand Reinhold.

Marcuse D (1991) *Theory of the Dielectic Optical Waveguide*. Boston, MA: Academic Press.

Mickelson AR, Lee Y-C and Basavanhally NR (1997) *Optoelectronic Packaging*. New York: Wiley.

Saleh BAA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Yariv A (1997) *Optical Electronics in Modern Communications*. New York: Van Nostrand Reinhold.

Dispersion

L Thévenaz, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Dispersion designates the property of a medium to propagate the different spectral components of a wave with a different velocity. It may originate from the natural dependence of the refractive index of a dense material to the wavelength of light, and one of the most evident and magnificent manifestations of dispersion is the appearance of a rainbow in the sky. In this case dispersion gives rise to a different refraction angle for the different spectral components of the white light, resulting in an angular dispersion of the sunlight spectrum and explicating the etymology of the term dispersion. Despite this spectacular effect, dispersion is mostly seen as an impairment in many applications, in particular for optical signal transmission. In this case, dispersion causes a rearrangement of the signal spectral components in the time domain, resulting in temporal spreading of the information and eventually in severe distortion.

There is a trend to designate all phenomena resulting in a temporal spreading of information as a type of dispersion, namely the polarization dispersion in optical fibers that should be properly named as polarization mode delay (PMD). In this article chromatic dispersion will be solely addressed, that designates the dependence of the propagation velocity on wavelength and that corresponds to the strict etymology of the term.

Effect of Dispersion on an Optical Signal

An optical signal may always be considered as a sum of monochromatic waves through a normal Fourier expansion. Each of these Fourier components propagates with a different phase velocity if the optical medium is dispersive, since the refractive index n depends on the optical frequency ν. This phenomenon is linear and causal and gives rise to a signal distortion that may be properly described by a transfer function in the frequency domain.

Let $n(\nu)$ be the frequency-dependent refractive index of the propagation medium. When the optical wave propagates in a waveguiding structure the effective wavenumber is properly described by a propagation constant β, that reads:

$$\beta(\nu) = \frac{2\pi\nu}{c_0} n(\nu) \qquad [1]$$

where c_0 is the vacuum light velocity. The propagation constant corresponds to an eigenvalue of the wave equation in the guiding structure and normally takes a different value for each solution or propagation mode. For free-space propagation this constant is simply equal to the wavenumber of the corresponding plane wave.

The complex electrical field $E(z,t)$ of a signal propagating in the z direction may be properly described by the following expression:

$$\begin{aligned} &E(z,t) = A(z,t)\, \mathrm{e}^{\mathrm{i}(2\pi\nu_0 t - \beta_0 z)} \\ &\quad \text{with} \quad \nu_0\text{: the central optical frequency} \\ &\quad \text{and} \quad \beta_0 = \beta(\nu_0) \end{aligned} \qquad [2]$$

where $A(z,t)$ represents the complex envelope of the signal, supposed to be slowly varying compared to the carrier term oscillating with the frequency ν_0. Consequently, the signal will spread over a narrow band around the central frequency ν_0 and the propagation constant β can be conveniently approximated by a limited expansion to the second order:

$$\beta(\nu) = \beta_0 + \left.\frac{\mathrm{d}\beta}{\mathrm{d}\nu}\right|_{\nu=\nu_0} (\nu - \nu_0) + \left.\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2}\right|_{\nu=\nu_0} (\nu - \nu_0)^2 \qquad [3]$$

For a known signal $A(0,t)$ at the input of the propagation medium, the problem consists in determining the signal envelope $A(z,t)$ after propagation over a distance z. The linearity and the causality of the system make possible a description using a transfer function $H_z(\nu)$ such as:

$$\tilde{A}(z,\nu) = H_z(\nu)\,\tilde{A}(0,\nu) \qquad [4]$$

where $\tilde{A}(z,\nu)$ is the Fourier transform of $A(z,t)$.

To make the transfer function $H_z(\nu)$ explicit, let us assume that the signal corresponds to an arbitrary harmonic function:

$$A(0,t) = A_0\,\mathrm{e}^{\mathrm{i}2\pi ft} \qquad [5]$$

Since this function is arbitrary and the signal may always be expanded as a sum of harmonic functions through a Fourier expansion, there is no loss of generality. The envelope identified as a harmonic function actually corresponds to a monochromatic wave of optical frequency $\nu = \nu_0 + f$ as defined by eqn [2]. Such a monochromatic wave will experience the following phase shift through propagation:

$$\begin{aligned} E(z,t) &= A_0\,\mathrm{e}^{\mathrm{i}[2\pi(\nu_0+f)t-\beta(\nu_0+f)z]} \\ &= A_0\,\mathrm{e}^{\mathrm{i}2\pi ft}\mathrm{e}^{\mathrm{i}[2\pi\nu_0 t-\beta(\nu_0+f)z]} \qquad [6] \end{aligned}$$

On the other hand, an equivalent expression may be found using the linear system described by eqns [2] and [4]:

$$\begin{aligned} E(z,t) &= A(z,t)\,\mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \\ &= A_z\,\mathrm{e}^{\mathrm{i}2\pi ft}\,\mathrm{e}^{\mathrm{i}(2\pi\nu_0 t-\beta_0 z)} \qquad [7] \end{aligned}$$

Since eqns [6] and [7] must represent the same quantities and using the definition in eqn [4], a simple comparison shows that the transfer function must take the following form:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}[\beta(\nu)z-\beta_0 z]} \qquad [8]$$

The transfer function takes a more analytical form using the approximation in eqn [3]:

$$H_z(\nu) = \mathrm{e}^{-\mathrm{i}2\pi(\nu-\nu_0)\tau_\mathrm{D}}\,\mathrm{e}^{-\mathrm{i}\pi D_\nu(\nu-\nu_0)^2 z} \qquad [9]$$

In the transfer function interpretation the first term represents a delay term. It means that the signal is delayed after propagation by the quantity:

$$\tau_\mathrm{D} = \frac{1}{2\pi}\frac{\mathrm{d}\beta}{\mathrm{d}\nu_0}z = \frac{z}{V_\mathrm{g}} \qquad [10]$$

where V_g represents the signal group velocity. This term therefore brings no distortion for the signal and thus states that the signal is replicated at the distance z with a delay τ_D.

The second term is the distortion term which is similar in form to a diffusion process and normally results in time spreading of the signal. In the case of a light pulse it will gradually broaden while propagating along the fiber, like a hot spot on a plate gradually spreading as a result of heat diffusion. The effect of this distortion is proportional to the distance z and to the coefficient D_ν, named group velocity dispersion (GVD):

$$D_\nu = \frac{1}{2\pi}\frac{\mathrm{d}^2\beta}{\mathrm{d}\nu^2} = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right) \qquad [11]$$

It is important to point out that the GVD may be either positive (normal) or negative (anomalous) and the distortion term in the transfer function in eqn [9] may be exactly cancelled by propagating in a medium with D_ν of opposite sign. It means that the distortion resulting from chromatic dispersion is reversible and this is widely used in optical links through the insertion of dispersion compensators. These are elements made of specially designed fibers or fiber Bragg gratings showing an enhanced GVD coefficient, with a sign opposite to the GVD in the fiber.

From the transfer function in eqn [9] it is possible to calculate the impulse response of the dispersive medium:

$$h_z(t) = \frac{1}{\sqrt{i|D_\nu|z}}\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu z}} \qquad [12]$$

so that the distortion of the signal may be calculated by a simple convolution of the impulse response with the signal envelope in the time domain.

The effect of dispersion on the signal can be more easily interpreted by evaluating the dispersive propagation of a Gaussian pulse. In this particular case the calculation of the resulting envelope can be carried out analytically. If the signal envelope takes the following Gaussian distribution at the origin:

$$A(0,t) = A_0\,\mathrm{e}^{-\frac{t^2}{\tau_0^2}} \qquad [13]$$

with τ_0 the $1/e$ half-width of the pulse, the envelope at distance z is obtained by convoluting the initial envelope with the impulse response $h_z(t)$:

$$\begin{aligned} A(z,t) &= h_z(t)\otimes A(0,t) \\ &= A_0\sqrt{\frac{\mathrm{i}z_0}{z+\mathrm{i}z_0}}\,\mathrm{e}^{\mathrm{i}\pi\frac{(t-\tau_\mathrm{D})^2}{D_\nu(z+\mathrm{i}z_0)}} \qquad [14] \end{aligned}$$

where

$$z_0 = -\frac{\pi\tau_0^2}{D_\nu} \qquad [15]$$

represents the typical dispersion length, that is the distance necessary to make the dispersion effect noticeable.

The actual pulse spreading resulting from dispersion can be evaluated by calculating the intensity of the envelope at distance z:

$$|A(z,t)|^2 = A_0 \frac{\tau_0}{\tau(z)} \mathrm{e}^{-\frac{2(t-\tau_D)^2}{\tau^2(z)}} \qquad [16]$$

that is still a Gaussian distribution centered about the propagation delay time τ_D, with $1/e^2$ half-width:

$$\tau(z) = \tau_0\sqrt{1 + (z/z_0)^2} \qquad [17]$$

The variation of the pulse width $\tau(z)$ is presented in Figure 1 and clearly shows that the pulse spreading starts to be nonnegligible from the distance $z = z_0$. This gives a physical interpretation for the dispersion length z_0. It must be pointed out that there is a direct formal similarity between the pulse broadening of a Gaussian pulse in a dispersive medium and the spreading of a free-space Gaussian beam as a result of diffraction. Asymptotically for distances $z \gg z_0$, the pulsewidth increases linearly:

$$\tau(z) \simeq |D_\nu| \frac{z}{\pi\tau_0} \qquad [18]$$

It must be pointed out that the width increases proportionally to the dispersion D_ν, but also inversely proportionally to the initial width τ_0. This results from the larger spectral width corresponding to a narrower pulsewidth, giving rise to a stronger dispersive effect.

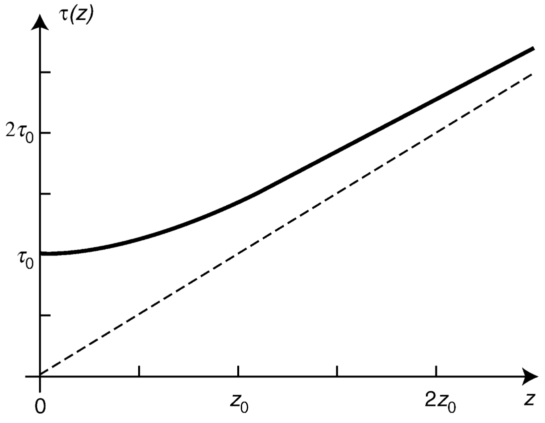

Figure 1 Variation of the $1/e^2$ half-width of a Gaussian pulse, showing the pulse spreading effect of dispersion. The dashed line shows the asymptotic linear spreading for large propagation distance.

The chromatic dispersion does not modify the spectrum of the transmitted light, as any linear effect. This can be straightforwardly demonstrated by evaluating the intensity spectrum of the signal envelope at any distance z, using the eqns [4] and [8]:

$$|\tilde{A}(z,\nu)|^2 = |H_z(\nu)\tilde{A}(0,\nu)|^2 = |\mathrm{e}^{-\mathrm{i}[\beta(\nu)z - \beta_0 z]}|^2 |\tilde{A}(0,\nu)|^2 = |\tilde{A}(0,\nu)|^2 \qquad [19]$$

It means that the pulse characteristics in the time and frequency domains are no longer Fourier-transform limited, since after the broadening due to dispersion, the spectrum should normally spread over a narrower spectral width. This feature results from a re-arrangement of the spectral components within the pulse. This can be highlighted by evaluating the distribution of instantaneous frequency through the pulse. The instantaneous frequency ω_i is defined as the time-derivative of the wave phase factor $\phi(t)$ and is uniformly equal to the optical carrier pulsation $\omega_i = 2\pi\nu_0$ for the initial pulse, as can be deduced from the phase factor at $z = 0$ by combining eqns [2] and [13]. This constant instantaneous frequency means that all spectral components are uniformly present within the pulse at the origin.

After propagation through the dispersive medium the phase factor $\phi(t)$ can be evaluated by combining eqns [2] and [14] and evaluating the argument of the resulting expression. The instantaneous frequency ω_i is obtained after a simple time derivative of $\phi(t)$ and reads:

$$\omega_i(t) = \omega_0 + \frac{2\pi z}{D_\nu(z^2 + z_0^2)}(t - \tau_D) \qquad [20]$$

For $z > 0$ the instantaneous frequency varies linearly over the pulse, giving rise to a frequency chirp. The frequency components are re-arranged in the pulse, so that the lower frequency components are in the leading edge of the pulse for a normal dispersion $D_\nu > 0$ and the higher frequencies in the trailing edge. For an anomalous dispersion $D_\nu < 0$, the arrangement is opposite, as can be seen in Figure 2. The effect of this frequency chirp can be visualized in Figure 3, showing that the effect of dispersion is equivalent to a frequency modulation over the pulse. The chirp is maximal at position z_0 and the pulsewidth takes its minimal value τ_0 when the chirp is zero. It is evident with this description that the pulse broadening may be entirely compensated through propagation in a medium of opposite

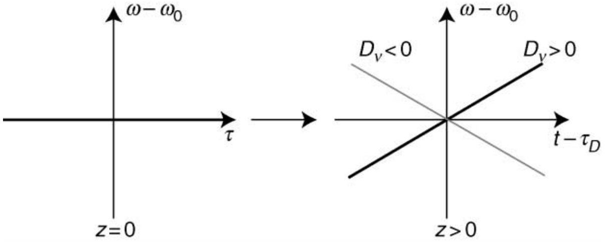

Figure 2 Distribution of the instantaneous frequency through a Gaussian pulse. At the origin the distribution is uniform (left) and the dispersion induces a frequency chirp that depends on the sign of the GVD coefficient D_ν.

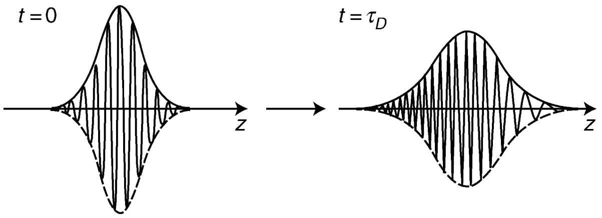

Figure 3 The dispersion results in a pulse broadening together with a frequency chirp, here for a normal GVD, that can be seen like a frequency modulation.

group velocity dispersion; this is equivalent to reversing the time direction in Figure 3. Moreover a pre-chirped pulse can be compressed to its Fourier-transform limited value after propagation in a medium with the proper dispersion sign. This feature is widely used for pulse compression after pre-chirping through propagation in a medium subject to optical Kerr effect.

The above description of the propagation of a Gaussian pulse implicitly states that the light source is perfectly coherent. In the early stages of optical communications it was not at all the case, since most sources were either light emitting diodes or multimode lasers. In this case the spectral extent of the signal was much larger than actually required for the strict need of modulation. Each spectral component may thus be considered as independently propagating the signal and the total optical wave can be identified to light merged from discrete sources emitting simultaneously the same signal at a different optical frequency. The group velocity dispersion will cause a delay $\delta\tau$ between different spectral components separated by a frequency interval $\delta\nu$ that can be simply evaluated by a first-order approximation:

$$\delta\tau = \frac{\mathrm{d}\tau_\mathrm{D}}{\mathrm{d}\nu}\delta\nu = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{z}{V_\mathrm{g}}\right)\delta\nu = D_\nu z \delta\nu \qquad [21]$$

where eqns [10] and [11] have been used. The delay $\delta\tau$ is thus proportional to the GVD coefficient D_ν, to the propagation distance z and the frequency separation $\delta\nu$. This description can be extended to a continuous frequency distribution with a spectral width σ_ν, resulting to the following temporal broadening σ_τ:

$$\sigma_\tau = |D_\nu|\sigma_\nu z \qquad [22]$$

Traditionally the spectral characteristics of a source are given in units of wavelength and the GVD coefficient is expressed in optical fibers accordingly. Following the same description as above, the temporal broadening σ_τ, for a spectral width σ_λ in wavelength units, reads:

$$\sigma_\tau = |D_\lambda|\sigma_\lambda z \qquad [23]$$

Since equal spectral widths must give equal broadening the value of the GVD in units of wavelength can be deduced from the natural definition in frequency units:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_\mathrm{g}}\right) = \frac{\mathrm{d}}{\mathrm{d}\nu}\left(\frac{1}{V_\mathrm{g}}\right)\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} = -\frac{c_0}{\lambda^2}D_\nu \qquad [24]$$

It must be pointed out that the coefficient D_λ takes a sign opposite to D_ν; in other words, a normal dispersion corresponds to a negative GVD coefficient D_λ. It is usually expressed in units of picoseconds of temporal broadening, per nanometer of spectral width and per kilometer of propagating distance, or ps/nm km. For example, a pulse showing a spectral width of 1 nm propagating through a 100 km fiber having a dispersion D_λ of $+10$ ps/nm km, will experience, according to eqn [23], a pulse broadening σ_τ of $10 \times 1 \times 100 = 1000$ ps or 1 ns.

Material Group Velocity Dispersion

Any dense material shows a variation of its index of refraction n as a function of the optical frequency ν. This natural property is called material dispersion and is the dominant contribution in weakly guiding structures such as standard optical fibers. This natural dependence results from the noninstantaneous response of the medium to the presence of the electric field of the optical wave. In other words, the polarization field $P(t)$ corresponding to the material response will vary with some delay or inertia to the change of the incident electric field $E(t)$. This delay between cause and effect generates a memory-type response of the medium that may be described using a time-dependent medium susceptibility $\chi(t)$. The relation between medium polarization at time t and incident field results from the weighted superposition of the effects of $E(t')$ at all previous times $t' < t$. This takes the form of the following convolution:

$$P(t) = \varepsilon_0 \int_{-\infty}^{+\infty} \chi(t - t')E(t')\mathrm{d}t' \qquad [25]$$

Through application of a simple Fourier transform this relation reads in the frequency domain as:

$$P(\nu) = \varepsilon_0 \chi(\nu)E(\nu) \qquad [26]$$

showing clearly that the noninstantaneous response of the medium results in a frequency-dependent refractive index using the standard relationship with the susceptibility χ:

$$n(\nu) = \sqrt{1 + \chi(\nu)} \quad \text{where} \quad \chi(\nu) = \mathrm{FT}\{\chi(t)\} \qquad [27]$$

This means that a beautiful natural phenomenon such as a rainbow, originates on a microscopic scale from the sluggishness of the medium molecules to react to the presence of light. For signal propagation, it results in a distortion of the signal and in most cases in a pulse spreading, but the microscopic causes are in essence identical. This noninstantaneous response is tightly related to the molecules' vibrations that also give rise to light absorption. For this reason it is convenient to express the propagation in an absorptive medium by adding an imaginary part to the susceptibility $\chi(\nu)$:

$$\chi(\nu) = \chi'(\nu) + \mathrm{i}\chi''(\nu) \qquad [28]$$

so that the refractive index $n(\nu)$ and the absorption coefficient $\alpha(\nu)$ reads in a weakly absorbing medium:

$$\begin{aligned} n(\nu) &= \sqrt{1 + \chi'(\nu)} \\ \alpha(\nu) &= -\frac{2\pi\chi''(\nu)}{\lambda n(\nu)} \end{aligned} \qquad [29]$$

Since the response of the medium, given by the time-dependent susceptibility $\chi(t)$ in eqn [25], is real and causal, the real and imaginary part of the susceptibility in eqn [28] are not entirely independent and are related by the famous Kramers–Kronig relations:

$$\begin{aligned} \chi'(\nu) &= \frac{2}{\pi}\int_0^{\infty} \frac{s\chi''(\nu)}{s^2 - \nu^2}\mathrm{d}s \\ \chi''(\nu) &= \frac{2}{\pi}\int_0^{\infty} \frac{\nu\chi'(\nu)}{\nu^2 - s^2}\mathrm{d}s \end{aligned} \qquad [30]$$

Absorption and dispersion act in an interdependent way on the propagating optical wave and knowing either the absorption or the dispersion spectrum is theoretically sufficient to determine the entire optical response of the medium. The interdependence between absorption and dispersion is typically illustrated in Figure 4. The natural tendency is to observe a growing refractive index for increasing frequencies in low absorption regions. In this case, the dispersion is called normal and this is the most observed situation in transparency regions of a material, which obviously offer the largest interest for optical propagation. In an absorption line the tendency is opposite, a diminishing index for increasing wavelength, and such a response is called anomalous dispersion. The dispersion considered here is the phase velocity dispersion, represented by the slope of $n(\nu)$, that must not be mistaken with the group velocity dispersion (GVD) that only matters for signal distortion. The difference between these two quantities is clarified below.

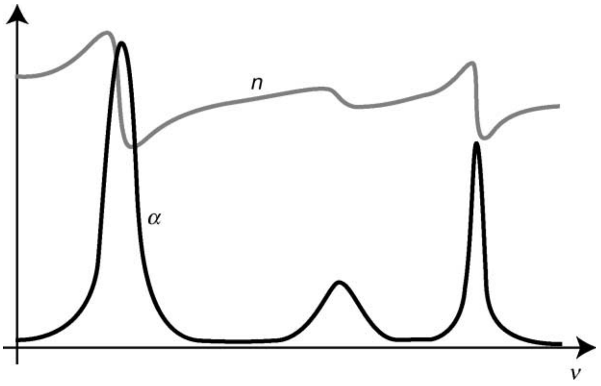

Figure 4 Typical optical response of a transparent medium, showing the spectral interdependence between the absorption coefficient α and the index of refraction n.

To demonstrate that a frequency-dependent refractive index $n(\nu)$ gives rise to a group velocity dispersion and thus a signal distortion we use eqns [1] and [10], so the group velocity V_g can be expressed:

$$V_\mathrm{g} = \frac{c_0}{N} \quad \text{with} \quad N = n + \nu\frac{\mathrm{d}n}{\mathrm{d}\nu} = n - \lambda\frac{\mathrm{d}n}{\mathrm{d}\lambda} \qquad [31]$$

N is called group velocity index and differs from the phase refractive index n only if n shows a spectral dependence. In a region of normal dispersion ($\mathrm{d}n/\mathrm{d}\nu > 0$), the group index is larger than the phase index and this is the situation observed in the great majority of transparent materials. From eqn [31] and using eqn [24] the GVD coefficient D_λ can be expressed as a function of the refractive index $n(\lambda)$:

$$D_\lambda = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{V_\mathrm{g}}\right) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{N}{c_0}\right) = -\frac{\lambda}{c_0}\frac{\mathrm{d}^2 n}{\mathrm{d}\lambda^2} \qquad [32]$$

The GVD coefficient is proportional to the second derivative of the refractive index n with respect to wavelength and is therefore minimal close to a point of inflexion of $n(\lambda)$. As can be seen in Figure 4 such a point of inflexion is always present where absorption is minimal, at the largest spectral distance from two absorption lines. It means that the conditions of low-group velocity dispersion and high transparency are normally fulfilled in the same spectral region of an optical dielectric material. In this region the phase velocity dispersion is normal at any wavelength, but the group velocity is first normal for shorter wavelengths, then is zero at a definite wavelength corresponding to the point of inflexion of $n(\lambda)$, and finally becomes anomalous for longer wavelengths. The situation in which normal phase dispersion and anomalous group dispersion are observed simultaneously is in no way exceptional.

In pure silica the zero GVD wavelength is at 1273 nm, but is subject to be moderately shifted to larger wavelengths in optical fibers, as a result of the presence of doping species to raise the index in the fiber guiding core. This shift normally never exceeds 10 nm using standard dopings; larger shifts are observed resulting from waveguide dispersion and this aspect will be addressed in the next section. The zero GVD wavelength does not strictly correspond to the minimum attenuation in silica fibers, because the dominant source of loss is Rayleigh scattering in this spectral region and not molecular absorption. This scattering results from fluctuations of the medium density as observed in any amorphous materials such as vitreous silica and is therefore a collective effect of many molecules that does not impinge on the microscopic susceptibility $\chi(t)$. It has therefore no influence on the material dispersion characteristics and this explains the reason for the minimal attenuation wavelength at 1550 nm, mismatching and quite distant from the zero material GVD at 1273 nm.

Material GVD in amorphous SiO_2 can be accurately described using a three-term Sellmeier expansion of the refractive index:

$$n(\lambda) = \sqrt{1 + \sum_{j=1}^{3} \frac{C_j \lambda^2}{\lambda^2 - \lambda_j^2}} \qquad [33]$$

and performing twice the wavelength derivative. The coefficients C_j and λ_j are found in most reference handbooks and result in the GVD spectrum shown in Figure 5. Such a spectrum explains the absence of interest for propagation in the visible region through optical fibers, the dispersion being very important in this spectral region. It also explains the large development of optical fibers in the 1300 nm region as a consequence of the minimal material dispersion there. It must be pointed out that it is quite easy to set up propagation in an anomalous GVD regime in optical fibers, since this regime is observed in the lowest attenuation spectral region. Anomalous dispersion makes possible interesting propagation features when combined with a third-order nonlinearity as the optical Kerr effect, namely soliton propagation and efficient spectral broadening through modulation instability.

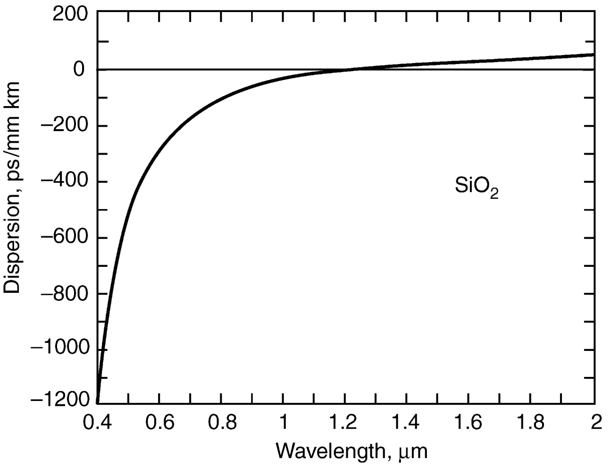

Figure 5 Material dispersion of pure silica. The visible region (0.4–0.7 µm) shows a strong normal GVD that decreases when moving into the infrared and eventually vanishes at 1273 nm. In the minimum attenuation window (1550 nm) the material GVD is anomalous.

Waveguide Group Velocity Dispersion

Solutions of the wave equation in an optical dielectric waveguide such as an optical fiber are discrete and limited. These solutions, called modes, are characterized by an unchanged field distribution along the waveguides and by a uniform propagation constant β over the wavefront. This last feature is particularly important if one recall that the field extends over regions presenting different refractive indices in a dielectric waveguide. For a given mode, the propagation constant β defines an effective refractive index n_{eff} for the propagation by similarity to eqn [1]:

$$\beta = \frac{2\pi\nu}{c_0} n_{\text{eff}} \qquad [34]$$

The value of this effective refractive index n_{eff} is always bound by the value of the core index n_1 and of the cladding index n_2, so that $n_2 < n_{\text{eff}} < n_1$. For a given mode, the propagation constant β, and so the effective refractive index n_{eff}, only depend on a quantity called normalized frequency V that essentially scales the light frequency to the waveguide optical parameters:

$$V = \frac{2\pi}{\lambda} a\sqrt{n_1^2 - n_2^2} \qquad [35]$$

where a is the core radius. Figure 6 shows the dependence of the propagation constant β of the fundamental mode LP_{01} on the normalized frequency V. The variation is nonnegligible in the single-mode region and gives rise to a chromatic dispersion, since V depends on the wavelength λ, even in the fictitious case of dispersion-free refractive indices in the core and the cladding materials. This type of chromatic dispersion is called waveguide dispersion.

To find an expression for the waveguide dispersion in function of the guiding properties, let us define another normalized parameter, the normalized phase constant b, such as:

$$\beta = \frac{2\pi}{\lambda}\sqrt{n_2^2 + b(n_1^2 - n_2^2)} \qquad [36]$$

The parameter b takes values in the interval $0 < b < 1$, is equal to 0 when $n_{\mathrm{eff}} = n_2$ at the mode cutoff, and is equal to 1 when $n_{\mathrm{eff}} = n_1$. This latter situation is never observed and is only asymptotic for very large normalized frequencies V. Solving the wave equation provides the dispersion relation $b(V)$ and it is important to point out that this relation between normalized quantities depends only on the shape of the refractive index profile. Step-index, triangular or multiple-clad index profiles will result in different $b(V)$ relations, independently of the actual values of the refractive indices n_1 and n_2 and of the core radius a.

From the definitions in eqns [10] and [35] and in the fictitious case of an absence of material dispersion, the propagation delay per unit length reads:

$$\frac{1}{V_g} = \frac{1}{2\pi}\frac{d\beta}{d\nu} = \frac{1}{2\pi}\frac{d\beta}{dV}\frac{dV}{d\nu} = \frac{\lambda}{2\pi}V\frac{d\beta}{dV} \qquad [37]$$

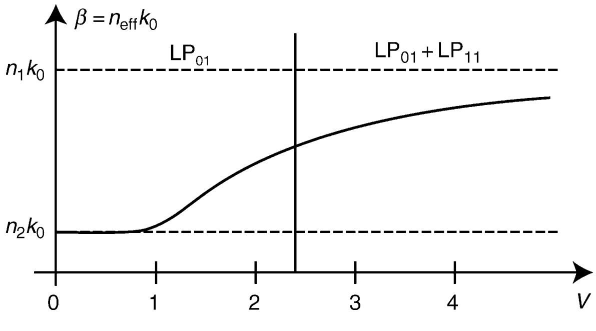

Figure 6 Propagation constant β as a function of the normalized frequency V in a step-index optical fiber. The single mode region is in the range $0 < V < 2.405$.

so that the waveguide group velocity dispersion can be expressed using the relation in eqn [24]:

$$\begin{aligned} D_\lambda^{\mathrm{w}} &= \frac{d}{d\lambda}\left(\frac{1}{V_g}\right) = \frac{d}{d\nu}\left(\frac{1}{V_g}\right)\frac{d\nu}{d\lambda} \\ &= -\frac{\nu}{\lambda}\frac{d}{d\nu}\left(\frac{1}{V_g}\right) = -\frac{1}{2\pi c_0}V^2\frac{d^2\beta}{dV^2} \end{aligned} \qquad [38]$$

The calculation of the combined effect of material and waveguide dispersions results in very long expressions in which it is difficult to highlight the relative effect of each contribution. Nevertheless, by making the assumption of weak guidance:

$$\frac{n_1 - n_2}{n_2} \simeq \frac{N_1 - N_2}{N_2} \ll 1 \qquad [39]$$

where N_1 and N_2 are the group indices in the core and the cladding, respectively, defined in eqn [31], the complete expression can be drastically simplified to obtain for the delay per unit length:

$$\frac{1}{V_g} = \frac{1}{c_0}\left[N_2 + (N_1 - N_2)\frac{d(bV)}{dV}\right] \qquad [40]$$

and for the total dispersion:

$$\begin{aligned} D_\lambda = {} & D_2 + (D_1 - D_2)\frac{d(bV)}{dV} \\ & - (N_1 - N_2)\frac{N_2}{n_1}\frac{1}{\lambda c_0}V\frac{d^2(bV)}{dV^2} \end{aligned} \qquad [41]$$

where

$$D_1 = -\frac{\lambda}{c_0}\frac{d^2 n_1}{d\lambda^2} \quad \text{and} \quad D_2 = -\frac{\lambda}{c_0}\frac{d^2 n_2}{d\lambda^2} \qquad [42]$$

are the material GVD in the core and the cladding, respectively.

The first two terms in eqn [41] represent the contribution of material dispersion weighted by the relative importance of core and cladding materials for the propagating mode. In optical fibers, the difference between D_1 and D_2 is small, so that this contribution can be often well approximated by D_2, independently of any guiding effects.

The last term represents the waveguide dispersion and is scaled by 2 factors:

- *The core-cladding index difference* $n_1 - n_2 \simeq N_1 - N_2$. The waveguide dispersion will be significantly enhanced by increasing the index difference between core and cladding.
- *The shape factor* $V(d^2(bV)/dV^2)$. This factor uniquely depends on the shape of the refractive index profile and may substantially modify the

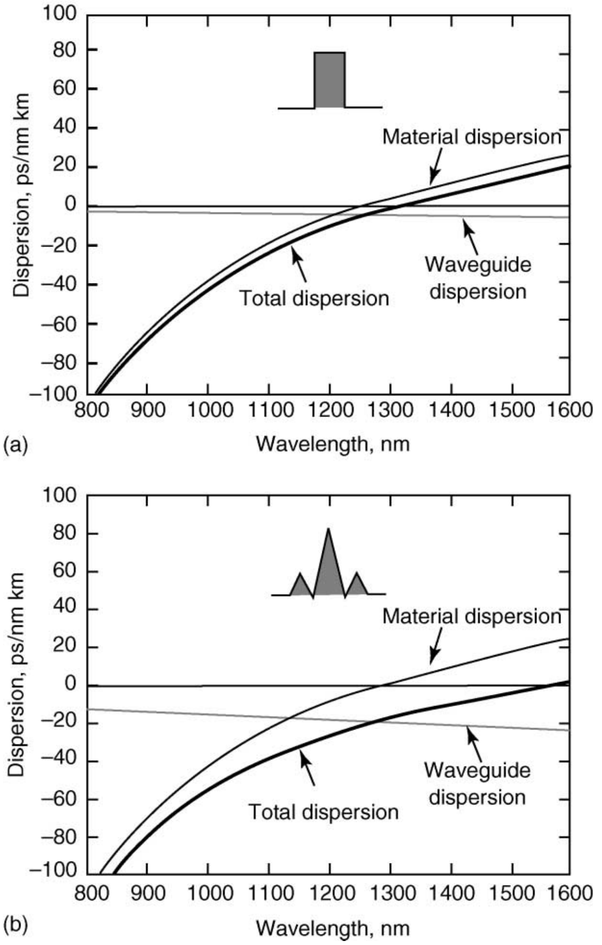

Figure 7 Material, waveguide and total group velocity dispersions for: (a) step-index fiber; (b) triangular profile fiber. The waveguide dispersion can be significantly enhanced by changing the shape of the index profile, making possible a shift of the zero GVD to the minimum attenuation window.

spectral dependence of the waveguide dispersion, making possible a great variety of dispersion characteristics.

Due to this degree of freedom brought by waveguide dispersion it is possible to shift the zero GVD wavelength to the region of minimal attenuation at 1550 nm in silica optical fibers. Figure 7a shows the total group velocity dispersion of a step-index core fiber, together with the separate contributions of material and waveguide GVD. In this case, the material GVD is clearly the dominating contribution, while the small waveguide GVD results in a shift of the zero GVD wavelength from 1273 nm to 1310 nm. A larger shift could be obtained by increasing the core-cladding index difference, but this also gives rise to an increased attenuation from the doping and no real benefit can be expected from such a modification. In Figure 7b, the core shows a triangular index profile to enhance the shape factor in eqn [41], so that the contribution of waveguide GVD is significantly increased with no impairing attenuation due to excessive doping. This makes it possible to realize the ideal situation of an optical fiber showing a zero GVD at the wavelength of minimum attenuation. These dispersion-shifted fibers (DSF) have now become successful in modern telecommunication networks.

Nevertheless, the absence of dispersion favors the efficiency of nonlinear effects and several classes of fibers are now proposed, showing a small but nonzero GVD at 1550 nm, with positive or negative sign, nonzero DSF (NZDSF). By interleaving fibers with positive and negative GVDs it is possible to propagate along the optical link in a dispersive medium and thus minimize the impact of nonlinearities, while maintaining the overall GVD of the link close to zero and canceling any pulse spreading accordingly.

List of Units and Nomenclature

Chromatic dispersion [ps nm^{-1} km^{-1}]	D_λ
Electric field [V m^{-1}]	E
Group index	N
Group velocity [m s^{-1}]	V_g
Group Velocity Dispersion GVD [s^2 m^{-1}]	D_ν
Linear attenuation coefficient [m^{-1}]	α
Medium susceptibility	χ
Normalized frequency	V
Optical frequency [s^{-1}]	ν
Propagation constant [m^{-1}]	β
Propagation delay [s]	τ_D
Polarization density field [A s m^{-2}]	P
Refractive index	n
Vacuum light velocity [m s^{-1}]	c_0
Wavelength [m]	λ

See also

Dispersion Management. **Fiber and Guided Wave Optics:** Overview. **Fourier Optics**. **Nonlinear Optics, Basics:** Kramers–Kronig Relations in Nonlinear Optics.

Further Reading

Ainslie BJ and Day CR (1986) A review of single-mode fibers with modified dispersion characteristics. *IEEE Journal of Lightwave Technology* LT-4(8): 967–979.

Bass M (ed.) (1994) *Handbook of Optics*, sponsored by the Optical Society of America, 2nd edn, vol. II, chap. 10 & 33. New York: McGraw-Hill.

Buck JA (1995) In: Goodman JW (ed.) *Fundamentals of Optical Fibers*, chap. 1, 5, 6. New York: Wiley Series in Pure and Applied Optics.

Gloge D (1971) Dispersion in weakly guiding fibers. *Applied Optics* 10: 2442–2445.

Jeunhomme LB (1989) *Single Mode Fiber Optics: Principles and Applications*. Optical Engineering Series, No. 23. New York: Marcel Dekker.

Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. Series on Quantum Electronics – Principles and Applications, chap. 2. New York: Academic Press.

Marcuse D (1980) Pulse distortion in single-mode fibers. *Applied Optics* 19: 1653–1660.

Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 2962–2974.

Marcuse D (1981) Pulse distortion in single-mode fibers. *Applied Optics* 20: 3573–3579.

Mazurin OV, Streltsina MV and Shvaiko-Shvaikovskaya TP (1983) *Silica Glass and Binary Silicate Glasses (Handbook of Glass Data; Part A)* Series on Physical Sciences Data. Amsterdam: Elseviervol. 15, pp. 63–75.

Murata H (1988) *Handbook of Optical Fibers and Cables*. Optical Engineering Series, No. 15, chap. 2. New York: Marcel Dekker.

Saleh BEA and Teich MC (1991) In: Goodman JW (ed.) *Fundamentals of Photonics*, chap. 5, 8 & 22. New York: Wiley Series in Pure and Applied Optics.

Fabrication of Optical Fiber

D Hewak, University of Southampton,
Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The drawing of optical fibers from silica preforms has, over a short period of time, progressed from the laboratory to become a manufacturing process capable of producing millions of kilometers of telecommunications fiber a year. Modern optical fiber fabrication processes produce low-cost fiber of excellent quality, with transmission losses close to their intrinsic loss limit. Today, fiber with transmission losses of 0.2 dB per kilometer of fiber are routinely drawn through a two-stage process that has been refined since the 1970s.

Although fibers of glass have been fabricated and used for hundreds of years, it was not until 1966 that serious interest in the use of optical fibers for communication emerged. At this time, it was estimated that the optical transmission loss in bulk glass could be as low as 20 dB km^{-1} if impurities were sufficiently reduced, a level at which practical applications were possible. At this time, no adequate fabrication techniques were available to synthesize glass of high purity, and fiber-drawing methods were crude.

Over the next five years, efforts worldwide addressed the fabrication of low-loss fiber. In 1970, a fiber with a loss of 20 dB km^{-1} was achieved. The fiber consisted of a titania doped core and pure silica cladding. This result generated much excitement and a number of laboratories worldwide actively began researching optical fiber. New fabrication techniques were introduced, and by 1986, fiber loss had been reduced close to the theoretical limit.

All telecommunications fiber that is fabricated today is made of silica glass, the most suitable material for low-loss fibers. Early fiber research studied multicomponent glasses, which are perhaps more familiar optical materials; however, low-loss fiber, could not be realized, partly due to the lack of a suitable fabrication method. Today other glasses, in particular the fluorides and sulfides, continue to be developed for speciality fiber applications, but silica fiber dominates in most applications.

Silica is a glass of simple chemical structure containing only two elements, silicon and oxygen. It has a softening temperature of about 2,000°C at which it can be stretched, i.e. drawn into fiber. An optical fiber consists of a high purity silica glass core, doped with suitable oxide materials to raise its refractive index (Figure 1). This core, typically on the order of 2–10 microns in diameter, is surrounded by silica glass of lower refractive index. This cladding layer extends the diameter to typically 125 microns. Finally a protective coating covers the entire structure. It is the phenomenon of total internal reflection at the core cladding interface that confines light to the core and allows it to be guided. The basic requirements of an optical fiber are as follows:

1. The material used to form the core of the fiber must have a higher refractive index than the cladding material, to ensure the fiber is a guiding structure.

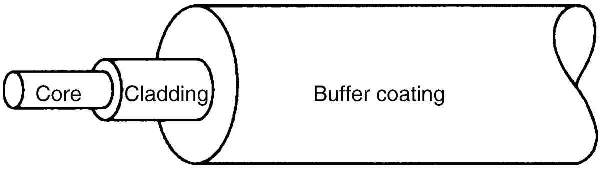

Figure 1 Structure of an optical fiber.

2. The materials used must be low loss, providing transmission with no absorption or scattering of light.
3. The materials used must have suitable thermal and mechanical properties to allow them to be drawn down in diameter into a fiber.

Silica (SiO_2) can be made into a glass relatively easily. It does not easily crystallize, which means that scattering from unwanted crystalline centers within the glass is negligible. This has been a key factor in the achievement of a low-loss fiber. Silica glass has high transparency in the visible and near-infrared wavelength regions and its refractive index can be easily modified. It is stable and inert, providing excellent chemical and mechanical durability. Moreover, the purification of the raw materials used to synthesize silica glass is quite straightforward.

In the first stage of achieving an optical fiber, silica glass is synthesized by one of three main chemical vapor processes. All use silicon tetrachloride ($SiCl_4$) as the main precursor, with various dopants to modify the properties of the glass. The precursors are reacted with oxygen to form the desired oxides. The end result is a high purity solid glass rod with the internal core and cladding structure of the desired fiber.

In the second stage, the rod, or preform as it is known, is heated to its softening temperature and stretched to diameters of the order of 125 microns. Tens to hundreds of kilometers of fiber are produced from a single preform, which is drawn continuously, with minimal diameter fluctuations. During the drawing process one or more protective coatings are applied, yielding long lengths of strong, low-loss fiber, ready for immediate application.

Preform Fabrication

The key step in preparing a low-loss optical fiber is to develop a technique for completely eliminating transition metal and OH ion contamination during the synthesis of the silica glass. The three methods most commonly used to fabricate a glass optical fiber preform are: the modified chemical vapor deposition process (MCVD); the outside vapor deposition process (OVD); and the vapor-axial deposition process (VAD).

In a typical vapor phase reaction, halide precursors undergo a high temperature oxidation or hydrolysis to form the desired oxides. The completed chemical reaction for the formation of silica glass is, for oxidation:

$$SiCl_4 + O_2 \rightarrow SiO_2 + 2Cl_2 \qquad [1]$$

For hydrolysis, which occurs when the deposition occurs in a hydrogen-containing flame, the reaction is:

$$SiCl_4 + 2H_2O \rightarrow SiO_2 + 4HCl \qquad [2]$$

These processes produce fine glass particles, spherical in shape with a size of the order of 1 nm. These glass particles, known as soot, are then deposited and subsequently sintered into a bulk transparent glass. The key to low-loss fiber is the difference in vapor pressures of the desired halides and the transition metal halides that cause significant absorption loss at the wavelengths of interest. The carrier gas picks up a pure vapor of, for example, $SiCl_4$, and any impurities are left behind.

Part of the process requires the formation of the desired core/cladding structure in the glass. In all cases, silica-based glass is produced with variations in refractive index produced by the incorporation of dopants. Typical dopants used are germania (GeO_2), titania (TiO_2), alumina (Al_2O_3), and phosphorous pentoxide (P_2O_5) for increasing the refractive index, and boron oxide (B_2O_2) and fluorine (F) for decreasing it. These dopants also allow other properties to be controlled, such as the thermal expansion of the glass and its softening temperatures. In addition, other materials, such as the rare earth elements, have also been used to fabricate active fibers that are used to produce optical fiber amplifiers and lasers.

MCVD Process

Optical fibers were first produced by the MCVD method in 1974, a breakthrough that completely solved the technical problems of low-loss fiber fabrication (Figure 2).

As shown schematically in Figure 3, the halide precursors are carried in the vapor phase by oxygen carrier gas into a pure silica substrate tube. An oxy-hydrogen burner traverses the length of the tube, which it heats externally. The tube is heated to temperatures of about 1,400°C which then oxidizes the halide vapor materials. The deposition temperature is sufficiently high to form a soot made of glassy particles which are deposited on the inside wall of the substrate tube but low enough to prevent the softened silica substrate tube from collapsing. The process usually takes place on a horizontal glass-working lathe.

During the deposition process, the repeated traversing of the burner forms multiple layers of soot. Changes to the precursors entering the tube and thus the resulting glass composition are introduced for layers which will form the cladding and then the core. The MCVD method allows germania to be doped into

Figure 2 Fabrication of an optical fiber preform by the MCVD method.

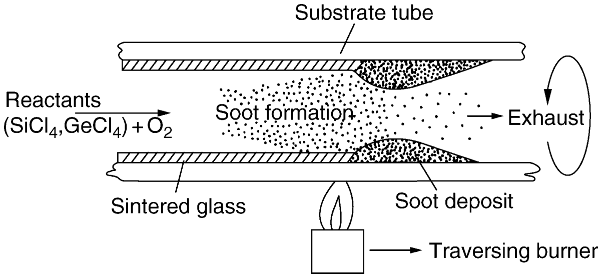

Figure 3 Schematic of preform fabrication by the MCVD method.

the silica glass and the precise control of the refractive index profile of the preform. When deposition is complete, the burner temperature is increased and the hollow, multilayered structure is collapsed to a solid rod. A characteristic of fiber formed by this process is a refractive index dip in the center of the core of the fiber. In addition, the deposition takes place in a closed system, which dramatically reduces contamination by OH^- ions and maintains low levels of other impurities. The high temperature allows high deposition rates compared to traditional chemical vapor deposition (CVD) and large performs, built up from hundreds of layers can be produced.

The MCVD method is still widely used today though it has some limitations, particularly on the preform size that can be achieved and thus the manufacturing cost. Diameter of the final preform is determined by the size of the initial silica substrate tube, which, due to the high purity required, accounts for a significant portion of the cost of the preform. A typical preform fabricated by MCVD yields about 5 km of fiber. Its success has spurred improvements to the process, in particular to address the fiber yield from a single preform.

OVD Process

The outside vapor deposition (OVD) methods, also known as outside vapor phase oxidation (OVPO), synthesize the fine glass particles within a burner flame. The precursors, oxygen, and fuel for the burner, are introduced directly into the flame. The soot is deposited onto a target rod that rotates and traverses in front of the burner. As in MCVD, the preform is built up layer by layer, though now initially by depositing the core glass and then building up the cladding layers over this. After the deposition process is complete, the preform is removed from the target rod and is collapsed and sintered into a transparent glass preform. The center hole remains, but disappears during the fiber drawing process. This technique has advantages in both size of preform which can be obtained and the fact that a high-quality silica substrate tube is no longer required. These two advantages combine to make a more economical process. From a single preform, several hundred kilometers of fiber can be produced.

VAD Process

The most recent refinement to the fabrication process was developed again to aid the mass production of high-quality fibers. In the VAD process, both core and cladding glasses are deposited simultaneously. Like OVD, the soot is synthesized and deposited by flame hydrolysis, as shown in Figure 4, the precursors are blown from a burner, oxidized, and deposited onto a silica target rod.

Burners for the VAD process consist of a series of concentrate nozzles. The first delivers an inert carrier and the main precursors $SiCl_4$, the second delivers an inert carrier glass and the dopants, the third delivers hydrogen fuel, and the fourth delivers oxygen. Gas flows are up to a liter per minute and deposition rates can be very high. With the VAD process, both core and cladding glasses can be deposited simultaneously. The main advantage is that this is a continuous process, as the soot which forms the core and cladding glasses are deposited axially onto the end of the rotating silica rod, it is slowly drawn upwards into a furnace which sinters and consolidates the soot into a transparent glass. The upward motion is such that the end at which deposition is occurring remains in a fixed position and essentially the preform is grown from this base.

The advantages of the VAD process are a preform without a central dip and, most importantly, the mass production associated with a continuous process. The glass quality produced is uniform and the resulting fibers have excellent reproducibility and low loss.

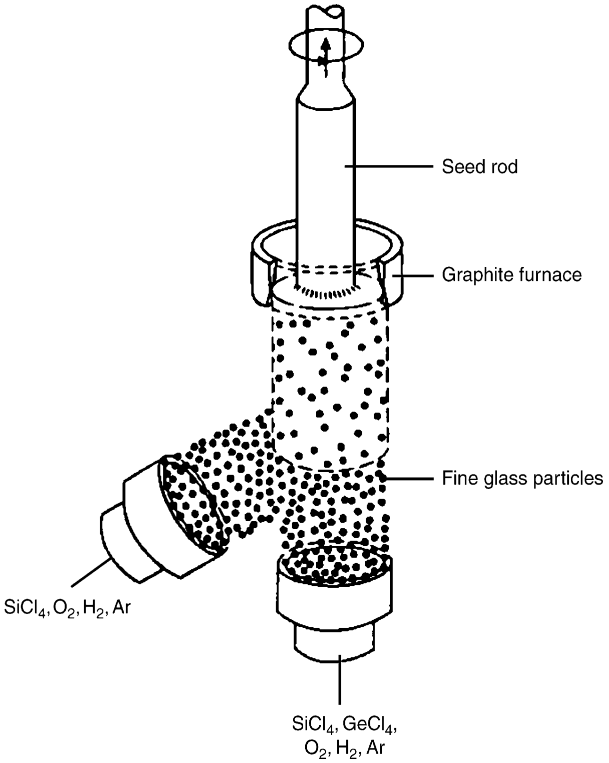

Figure 4 Schematic of preform fabrication by the VAD method.

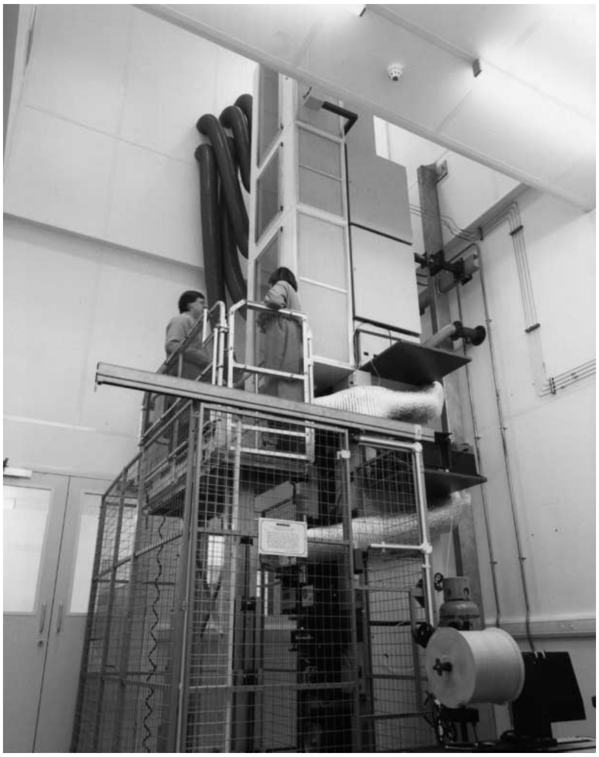

Figure 5 A commercial scale optical fiber drawing tower.

Other Methods of Preform Fabrication

There are other methods of preform fabrication, though these are not generally used for today's silica glass based fiber. Some methods, such as fabrication of silica through sol–gel chemistry could not provide the large low-loss preforms obtained by chemical vapor deposition techniques. Other methods, such as direct casting of molten glass into rods, or formation of rods and tubes by extrusion, are better suited to and find application in other glass chemistries and speciality fibers.

Fiber Drawing

Once a preform has been made, the second step is to draw the preform down in diameter on a fiber drawing tower. The basic components and configuration of the drawing tower have remained unchanged for many years, although furnace design has become more sophisticated and processes like coating have become automated. In addition, fiber drawing towers have increased in height to allow faster pulling speeds (Figure 5).

Essentially, the fiber drawing process takes place as follows. The preform is held in a chuck which is mounted on a precision feed assembly that lowers the preform into the drawing furnace at a speed which matches the volume of preform entering the furnace to the volume of fiber leaving the bottom of the furnace. Within the furnace, the fiber is drawn from the molten zone of glass down the tower to a capstan, which controls the fiber diameter, and then onto a drum. During the drawing process, immediately below the furnace, is a noncontact device that measures the diameter of the fiber. This information is fed back to the capstan which speeds up to reduce the diameter or slows down to increase the fiber diameter. In this way, a constant diameter fiber is produced. One or more coatings are also applied in-line to maintain the pristine surface quality as the fiber leaves the furnace and thus to maximize the fiber strength.

The schematic drawing in Figure 6 shows a fiber drawing tower and its key components. Draw towers are commercially available, ranging in height from 3 m to greater than 20 m, with the tower height increasing as the draw speed increases.

Typical drawing speeds are on the order of 1 m s^{-1}. The increased height is needed to allow the fiber to cool sufficiently before entering the coating applicator, although forced air-cooling of the fiber is commonplace in a manufacturing environment.

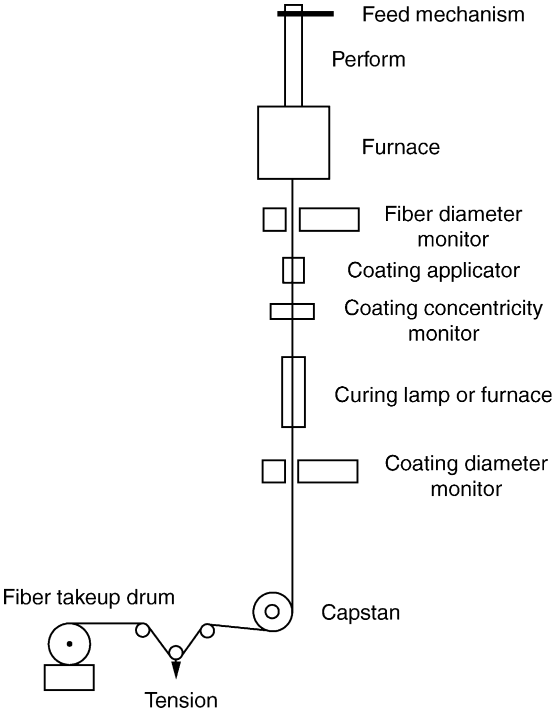

Figure 6 Schematic diagram of an optical fiber drawing tower and its components.

Furnaces

A resistance furnace is perhaps the most common and economical heating source used in a fiber drawing tower. A cylindrical furnace, with graphite or tungsten elements, provides heating to the preform by blackbody radiation. These elements must be surrounded by an inert gas, such as nitrogen or argon, to prevent oxidation. Gas flow and control is critical to prevent variations in the diameter of the fiber. In addition, the high temperature of the elements may lead to contamination of the preform surface that can then weaken the fiber.

An induction furnace provides precise clean heating through radio frequency (RF) energy that is inductively coupled to a zirconia susceptor ring. This type of furnace is cleaner than the graphite resistance type and is capable of continuous running for several months. It also has the advantage that it does not require a protective inert atmosphere and consequently the preform is drawn in a turbulence-free environment. These advantages result in high-strength fibers with good diameter control.

Diameter Measurement

A number of noncontact methods of diameter measurement can be applied to measure fiber diameter. These include laser scanning, interferometry, and light scattering techniques. To obtain precise control of the fiber diameter, the deviation between the desired diameter and the measured diameter is fed into the diameter control system. In order to cope with high drawing speeds, sampling rates as high as 1,000 times per second are used. The diameter control is strongly affected by the gas flow in the drawing furnace and is less affected by the furnace temperature variation. The furnace gas flow can be used to achieve suppression of fast diameter fluctuations. This is used in combination with drawing speed control to achieve suppression of both fast and slow diameter fluctuations. Current manufacturing processes are capable of producing several hundreds of kilometers of fiber with diameter variations of ± 1 micron.

There are two major sources of fiber diameter fluctuations: short-term fluctuations caused by temperature fluctuations in the furnace; and long-term fluctuations caused by variations in the outer diameter of the preform. Careful control of the furnace temperature, the length of the hot zone, and the flow of gas minimize the short-term fluctuations. Through optimization of these parameters, diameter errors of less than ± 0.5 microns can be realized. The long-term fluctuations in diameter are controlled by the feedback mechanism between the diameter measurement and the capstan.

Fiber Coating

A fiber coating is primarily used to preserve the strength of a newly drawn fiber and therefore must be applied immediately after the fiber leaves the furnace. The fiber coating apparatus is typically located below the diameter measurement gauge at a distance determined by the speed of fiber drawing, the tower height, and whether there is external cooling of the fiber. Coating is usually one of the limiting factors in the speed of fiber drawing. To minimize any damage or contamination of the pristine fiber leaving the furnace, this portion of the tower, from furnace to the application of the first coating, is enclosed in a clean, filtered-air chamber.

A wide variety of coating materials has been applied; however, conventional, commercial fiber generally relies on a UV curable polymer as the primary coating. Coating thickness is typically 50–100 microns. Subsequent coatings can then be applied for specific purposes. A dual coating is often used with an inner primary coating that is soft and an outer, secondary coating which is hard. This ratio of low to high elastic modulus can minimize stress on the fiber and reduce bending loss.

Alternative coatings include hermetic coatings of a low melting temperature metal, ceramics, or amorphous carbon. These can be applied in-line before the polymeric coating. Metallic coatings are applied by passing the fiber through a molten metal while ceramic or amorphous coatings utilize an in-line chemical vapor deposition reactor.

Types of Fiber

The majority of silica fiber drawn today is single mode. This structure consists of a core whose diameter is chosen such that, with a given refractive index difference between the core and cladding, only a single guide mode propagates at the wavelength of interest. With a discrete index difference between core and cladding, it is often referred to as a step index fiber. In typical telecommunications fiber, single mode operation is obtained with core diameters of 2–10 microns with a standard outer diameter of 125 microns.

Multimode fiber has core diameters considerably larger, typically 50, 62.5, 85, and 110 microns, again with a cladded diameter of 125 microns. Multimode fibers are often graded index, that is the refractive index is a maximum in the center of the fiber and smoothly decreases radially until the lower cladding index is reached. Multimode fibers find use in non-telecommunication applications, for example optical fiber sensing and medicine.

Single mode fibers, which are capable of maintaining a linear polarization input to the fiber, are known as polarization preserving fibers. The structure of these fibers provides a birefringence that removes the degeneracy of the two possible polarization modes. This birefringence is a small difference in the effective refractive index of the two polarization modes that can be guided and it is achieved in one of two ways. Common methods for the realization of this birefringence are an elliptical core in the fiber, or through stress rods, which modify the refractive index in one orientation. These fiber structures are shown in Figure 7.

While the loss minimum of silica-based fiber is near 1.55 microns, step index single-mode fiber offers zero dispersion close to 1.3 micron wavelengths and dispersion at the loss minimum is considerable. A modification of the structure of the fiber, and in particular a segmented refractive index profile in the core, can be used to shift this dispersion minimum to 1.55 microns. This fiber, illustrated in Figure 7 is known as dispersion shifted fiber. Similarly fibers, with a relatively low dispersion over a wide wavelength range, known as dispersion flattened fibers, can be obtained by the use of multiple cladding layers.

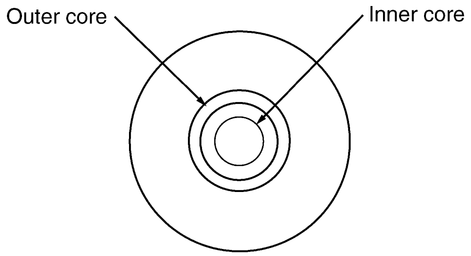

(a) Dispersion shifted fiber

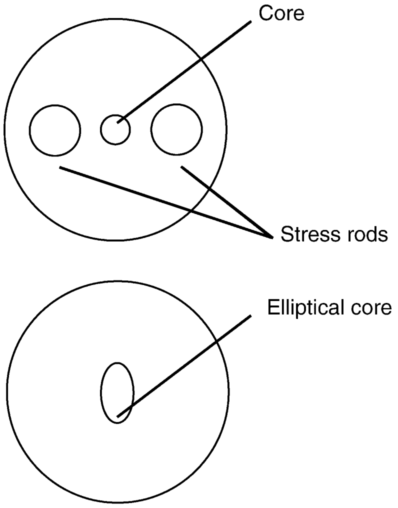

(b) Polarization preserving fiber

Figure 7 Structure of (a) dispersion shifted fiber and (b) two methods of achieving polarization preserving fiber.

List of Units and Nomenclature

Dopants	Elements or compounds added, usually in small amounts, to a glass composition to modify its properties.
Fiber drawing	The process of heating and thus softening an optical fiber preform and then drawing out a thin thread of glass.
Fiber loss	The transmission loss of light as it propagates through a fiber, usually measured in dB of loss per unit length of fiber. Loss can occur through the absorption of light in the core or scattering of light out of the core.
Glass	An amorphous solid formed by cooling from the liquid state to a rigid solid with no long range structure.
Modified chemical vapor deposition (MCVD)	A process for the fabrication of an optical preform where gases flow into the inside of a rotating tube, are heated and react

	to form particles of glass which are deposited onto the wall of the glass tube. After deposition, the glass particles are consolidated into a solid preform.
Outside vapor deposition (OVD)	A process for the fabrication of an optical preform where glass soot particles are formed in an oxy-hydrogen flame and deposited on a rotating rod. After deposition, the glass particles are consolidated into a solid preform.
Preform	A fiber blank, a bulk glass rod consisting of a core and cladding glass composite which is drawn into fiber.
Refractive index	A characteristic property of glass, which is defined by the speed of light in the material relative to the speed of light in a vacuum.
Silica	A transparent glass formed from silicon dioxide.
Vapor-axial deposition	A process similar to OVD, where the core and cladding layers are deposited simultaneously.

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Historical Development; Wavelength Division Multiplexing. **Optical Materials:** Optical Glasses.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: John Wiley & Sons.

Bass M and Van Stryland EV (eds) (2001) *Fiber Optics Handbook: Fiber, Devices, and Systems for Optical Communications*. New York: McGraw-Hill Professional Publishing.

Fujiura K, Kanamori T and Sudo S (1997) Fiber materials and fabrications. In: Sudo S (ed.) *Optical Fiber Amplifiers*, ch. 4, pp. 193–404. Boston, MA: Artech House.

Goff DR and Hansen KS (2002) *Fiber Optic Reference Guide: A Practical Guide to Communications Technology*, 3rd edn. Oxford, UK: Butterworth-Heinemann UK.

Hewak D (ed.) (1998) *Glass and Rare Earth Doped Glasses for Optical Fibres*. EMIS Datareview Series, INSPEC. London: The Institution of the Electrical Engineers.

Keck DB (1981) Optical fibre waveguides. In: Barnoski MK (ed.) *Fundamentals of Optical Fiber Communications*. New York: Academic Press.

Li T (ed.) (1985) *Optical Fiber Communications: Fiber Fabrication*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Schultz PC (1979) In: Bendow B and Mitra SS (eds) *Fiber Optics*. New York: Plenum Press.

Light Propagation

F G Omenetto, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fibers have become a mainstay of our modern way of life. From phone calls to the internet, their presence is ubiquitous. The unmatched ability that optical fibers have to transmit light is an amazing asset that, besides being strongly part of our present, is poised to shape the future.

The foundation of the unique ability of optical fibers to transmit light hinges on the well-known physical concept of total internal reflection. This phenomenon is described by considering the behavior of light traveling from one material (A) to a different material (B). When light traverses the interface of the two materials with a specific angle, its direction of propagation is altered with respect to the normal to the surface, according to the well-known Snell law which relates the incident angle (onto the interface between A and B) to the refracted angle (away from the interface between A and B). This is written as $n_A \sin i = n_B \sin r$, where i is the angle of incidence, r the angle of refraction, and n_A and n_B are the refractive indices of the materials A and B. Under appropriate conditions, a critical value of the incidence angle (i_{CR}) exists for which the light does not propagate into B and is totally reflected back into A. This value is given by $i_{CR} = \arcsin(n_A/n_B)$. It follows that total internal reflection occurs when light is traveling from a medium with a higher index of refraction into a medium with a lower index of refraction. As an approximation, this is the basic idea behind light propagation in fibers.

Optical fibers are glass strands that include a concentric core with a cladding wrapped around it. The index of refraction of the core is slightly higher than the index of refraction of the cladding, thereby creating the necessary conditions for the confinement of light within them. Such index mismatches between the core and the cladding are determined by using different materials for each, by doping the glass matrix before pulling the fiber to create an index gradient, by introducing macroscopic defects such as air-holes in the structure of the fiber cladding (such as in the recently developed photonic crystal fibers), or by using materials other than glass, such as polymers or plastic.

The real physical picture of light guiding in fibers is more complex than the conceptual description stated above. The light traveling in the fiber is, more precisely, an electromagnetic wave with frequencies that lie in the optical range, and the optical fiber itself constitutes an electromagnetic waveguide. As such, the waveguide supports guided electromagnetic modes which can take on a multitude of shapes (i.e., of energy distribution profiles), depending on the core/cladding index of refraction profiles. These profiles are often complex and determine the physical boundary conditions that influence the field distribution within the fiber. The nature of the fiber modes has been treated extensively in the literature where an in-depth description of the physical solutions can be found.

There are many distinguishing features that identify different optical fiber types. Perhaps the most important one, in relation to the electromagnetic modes that the fiber geometry supports, is the distinction that is made between single-mode fibers and multimode fibers. As the name implies, a single-mode fiber is an optical fiber whose index of refraction profile between core and cladding is designed in such a way that there is only one electromagnetic field distribution that the fiber can support and transmit. That is to say, light is propagated very uniformly across the fiber. A very common commercially used fiber of this kind is the Corning SMF-28 fiber, which is employed in a variety of telecommunication applications. Multimode fibers have larger cores and are easier to couple light into, yet the fact that many modes are supported implies that there is less control over the behavior of light during its propagation, since it is more difficult to control each individual mode.

The condition for single-mode propagation in optical fibers is determined during the design and manufacture of the fiber by its geometrical parameters and is specific to a certain operation wavelength. A quantity V is defined as $V = (2\pi/\lambda) \times a(n_{\mathrm{core}}^2 - n_{\mathrm{cladding}}^2)^{1/2}$, where a is the core radius and λ is the wavelength of light. For example, a step-index fiber (i.e., a fiber where the transition from the index of refraction of the core to the index of refraction in the cladding is abrupt), supports a single electromagnetic mode if $V < 2.405$.

Light propagation in optical fibers, however, is affected by numerous other factors besides the electromagnetic modes that the fiber geometry allows.

First, the ability to efficiently transmit light in optical fiber is dependent on the optical transparency of the medium that the fiber is made of. The optical transparency of the medium is a function of the wavelength (λ) of light that needs to be transmitted through the fiber. Optical attenuation is particularly high for shorter wavelengths and the transmission losses for ultraviolet wavelengths become considerably higher than in the near infrared. The latter region, due to this favorable feature, is the preferred region of operation for optical fiber based telecommunication applications. This ultimate physical limit is determined by the process of light being scattered off the atoms that constitute the glass, a process called Rayleigh scattering. The losses due to Rayleigh scattering are inversely proportional to the fourth power of the wavelength (losses $\propto \lambda^{-4}$), and therefore become quite considerable as λ is decreased.

The ability of optical fibers to transmit light over a certain set of wavelengths, is quantified by their optical attenuation. This parameter is defined as the ratio of the output power versus the input power and is usually measured in units of decibels (dB), i.e., attenuation (dB) $= 10 \log_{10}(\mathrm{P}_{\mathrm{out}}/\mathrm{P}_{\mathrm{in}})$.

The values that are found in the literature relate optical attenuation to distance and express the attenuation per unit length (such as dB/km). Single-mode fibers used in telecommunications are usually manufactured with doped silica glasses designed to work in the near infrared, in a wavelength range between 1.3 and 1.6 μm ($1\ \mu\mathrm{m} = 10^{-6}$ m) which provides the lowest levels of Rayleigh scattering and is located before the physical infrared absorption of silica (~1.65 μm) kicks in. Such fibers, which have been developed and refined for decades, have attenuation losses of less that 0.2 dB/km at a wavelength of 1.55 μm. Many other materials have been used to meet specific optical transmission needs. While silica remains by far the most popular material used for optical fiber manufacturing, other glasses are better suited for different portions of the spectrum: quartz glass can be particularly transmissive in the ultraviolet whereas plastic is sometimes conveniently used for short distance transmission in the visible.

The ability to collect light at the input of the fiber represents another crucial and defining parameter. Returning to total internal reflection, such an effect can only be achieved by sending light onto an interface at a specific angle. In fibers, the light needs to be coupled into a circular core and reach the core cladding interface at that specific geometrical angle. The measure of the ability to couple light in the fiber is referred to as 'coupling efficiency'. The core dimension defines the geometry of the coupling process which, in turn, determines the efficiency of the propagation of light inside the fiber. In order to be coupled efficiently and therefore guided within the fiber, the light has to be launched within a range of acceptable angles. Such a range is usually defined or derivable from the numerical aperture (NA) of the fiber which is approximately $\mathrm{NA} = (n_{\mathrm{core}}^2 - n_{\mathrm{cladding}}^2)^{1/2}$, or can be alternatively described as a function of the critical angle for total internal reflection within the fiber by $\mathrm{NA} = n_{\mathrm{core}} \sin i_{\mathrm{CR}}$. Also, light has to be focused into the core of the fiber and therefore the interplay between core size, focusability, and NA (i.e., acceptance angle) of the fiber influence the coupling efficiency into the fiber (which in turn affects transmission). Typical telecommunication fibers have a 10 μm core (a NA of ~0.2) and have very high coupling efficiencies (of the order of 70–80%). Specialized single mode fibers (such as photonic crystal fibers) can have core diameters down to a few microns, significantly lower coupling efficiencies, and high numerical apertures. Multimode fibers, in contrast, are easy to couple into given their large core size.

Other factors that contribute to losses during propagation are due to the physical path that the actual optical fiber follows. A typical optical fiber strand is rarely laid along a straight line. These types of losses are called bending losses and can be intuitively understood by thinking that when the fiber is straight, the light meets its total reflection condition at the core-cladding interface, but when the fiber bends, the interface is altered and varies the incidence conditions of light at the interface. If the bend is severe, light is no longer totally reflected and escapes from the fiber.

Another important issue that influences propagation arises when the light that travels through the fiber is sent in bursts or in pulses of light. Pulsed light is very important in fiber transmission because the light pulse is the carrier of optical information, most often representing a binary digit of data. The ability to transmit pulses reliably is at the heart of modern day telecommunication systems. The features that affect the transmission of pulses through optical fibers depend, again, largely on the structure and constituents of the fiber. One of the most important effects on pulsed light is dispersion. This phenomenon causes the broadening in time of the light pulses during their travel through the fiber. This can be easily explained by a Fourier analogy by thinking that a short pulse in time (such as the 10–20 picosecond pulse duration that is used for optical communications, $1\ \text{picosecond} = 10^{-12}$ seconds) has a large frequency content (i.e., is composed of a variety of 'colors'). Since the speed of light through bulk media, and therefore through the fiber constituent, depends on wavelength, different 'colors' will travel at different speeds, arriving at the end of the fiber slightly delayed with respect to one another. This causes a net pulse broadening and is dependent on the properties of the fiber that are depicted in the dispersion curve of the fiber. This type of dispersion is called chromatic dispersion and is dependent on the distance the pulse travels in the fiber. A measure of the pulse broadening can be calculated by multiplying the dispersion value for the fiber (measured in ps/km) at the wavelength of operation times and the distance in kilometers that the pulse travels.

There are other types of dispersion that affect pulses traveling through fibers: light with different polarization travels at a different speed, thereby causing an analogous problem to chromatic distortion. In multimode fibers, different modes travel with different properties. Dispersion issues are extremely important in the design of the next-generation telecommunication systems. As a push is made toward higher transmission capacity, pulses become shorter and these issues have to be dealt with carefully. Specifically, transmission of light in optical fibers becomes more complex because the interaction between the light and the fiber material constituents becomes nonlinear. As pulses become shorter, their energy is confined in a smaller temporal interval, making their peak power (which is equal to energy/time) increase. If the peak power becomes sufficiently high, the atoms that form the glass of which the fiber is made are excited in a nonlinear fashion adding effects that need to be addressed with caution. Energetic pulses at a certain wavelength change their shape during propagation and can become severely distorted or change wavelength (i.e., undergo a frequency conversion process), thereby destroying the information that they were meant to carry.

Several physical manifestations of optical nonlinearity affect pulsed light propagation, giving rise to new losses but also providing new avenues for efficient transmission.

Perhaps the most successful example of the positive outcome of nonlinear effects in fibers is the optical soliton. The latter is an extremely stable propagating pulse that is transmitted through the fiber undistorted, since it is a result of a peculiar balance between the linear distortion and nonlinear distortion that cancel each other by operating in a specific region of dispersion of the optical fiber (the anomalous dispersion region).

Nonlinear effects in fibers are an extremely rich area of study which, besides several drawbacks, carries immense opportunities if the nonlinearities can be controlled and exploited.

See also

Dispersion Management. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Solitons:** Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Hecht J (2001) *Understanding Fiber Optics*. New York: Prentice Hall.

Snyder AW and Love JD (1983) *Optical Waveguide Theory*. Boston: Kluwer Academic Press.

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego: Academic Press.

Measuring Fiber Characteristics

A Girard, EXFO, Quebec, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The optical fiber is divided in two types: multimode and singlemode. Each type is used in different applications and wavelength ranges and is consequently characterized differently. Furthermore, the corresponding test methods also vary.

The optical fiber characteristics may be divided into four categories:

- The optical characteristics (transmission related);
- The dimensional characteristics;
- The mechanical characteristics; and
- The environmental characteristics.

These categories will be reviewed in the following sections, together with their corresponding test methods.

Fiber Optical Characteristics and Corresponding Tests Methods

The following sections will describe the following optical characteristics:

- attenuation;
- macrobending sensitivity;
- microbending sensitivity;
- cut-off wavelength;
- multimode fiber bandwidth;
- differential mode delay for multimode fibers;
- chromatic dispersion;
- polarization mode dispersion;
- polarization crosstalk; and
- nonlinear effects.

Attenuation

The spectral attenuation of an optical fiber follows exponential power decay from the power level at a cross-section 1 to the power level at cross-section 2, over a fiber length L, as follows:

$$P_2(\lambda) = P_1(\lambda) \cdot e^{-\gamma(\lambda)L} \qquad [1]$$

$P_1(\lambda)$ is the optical power transmitted through the fiber core cross-section 1, expressed in mW; $P_2(\lambda)$ is the optical power transmitted through the fiber core cross-section 2 away from cross-section 1, expressed in mW; $\gamma(\lambda)$ is the spectral attenuation coefficient in linear units, expressed in km^{-1}; and L is the fiber length expressed in km.

Attenuation may be characterized at one or more specific wavelengths or as a function of wavelength. In the later case, attenuation is referred to spectral attenuation. Figure 1 illustrates such power decay.

Equation [1] may be expressed in relative units as follows:

$$\log_{10} P_2 = (\log_{10} P_1) \cdot - \gamma L \cdot \log_{10} e \qquad [2]$$

P is expressed in dBm units using the following definition.

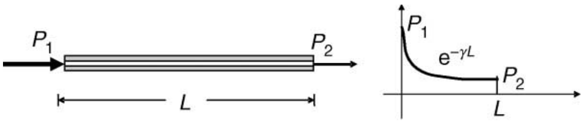

Figure 1 Power decay in an optical fiber due to attenuation.

The power in dBm is equal to 10 times the logarithm in base 10 of the power in mW; or

$$0\ \text{dBm} = 10\log_{10}(1\ \text{mW}) \qquad [3]$$

Then

$$\gamma(\lambda)[\text{km}^{-1}] = [(10\log_{10} P_1) - (10\log_{10} P_2)] \div L\log_{10} e \qquad [4]$$

$$\alpha(\lambda)[\text{dB/km}] = [P_1(\text{dBm}) - P_2(\text{dBm})]/L \qquad [5]$$

A new relative attenuation coefficient $\alpha(\lambda)$ with units of dB/km has been introduced, which is related to the linear coefficient $\gamma(\lambda)$ with units of km^{-1} as follows:

$$\alpha(\lambda) = (\log_{10} e)\cdot\gamma(\lambda) \approx 4{,}343\gamma(\lambda) \qquad [6]$$

$\alpha(\lambda)$ is the spectral attenuation coefficient of a fiber and is illustrated in Figure 2.

Test methods for attenuation

The attenuation may be measured by the following methods:

- cut-back method;
- backscattering method; and
- insertion loss method.

Cut-back method. The cut-back method is a direct application of the definition in which the power levels P_1 and P_2 are measured at two points of the fiber change of input conditions. P_2 is the power emerging from the far end of the fiber and P_1 is the power emerging from a point near the input after cutting the fiber.

The output power P_2 is recorded from the fiber under test (FUT) placed in the measurement setup. Keeping the launching conditions fixed, the FUT is cut to the cut-back length, for example 2 m from the launching point. The FUT attenuation, between the points where P_1 and P_2 have been measured, is calculated using P_1 and P_2, from the definition equations provided above.

Backscattering method. The attenuation coefficient of a singlemode fiber is characterized using bidirectional backscattering measurements. This method is also used for:

- attenuation uniformity;
- optical continuity;
- physical discontinuities;
- splice losses; and
- fiber length.

An optical time domain reflectometer (OTDR) is used for performing such characterization. Adjustment of laser pulsewidth and power is used to obtain a compromise between resolution (a shorter pulsewidth provides a better resolution but at lower power) and dynamic range/fiber length (higher power provides better dynamic range but with longer pulsewidth). An example of such an instrument is shown in Figure 3.

The measurement is applicable to various FUT configurations (e.g., cabled fiber in production or deployed in the field, fiber on a spool, etc.). Two unidirectional backscattering curves are obtained, one from each end of the fiber (Figure 4).

Each backscattering curve is recorded on a logarithmic scale, avoiding the parts at the two ends of the curves, due to parasitic reflections.

The FUT length is found from the time interval between the two ends of the backscattering loss

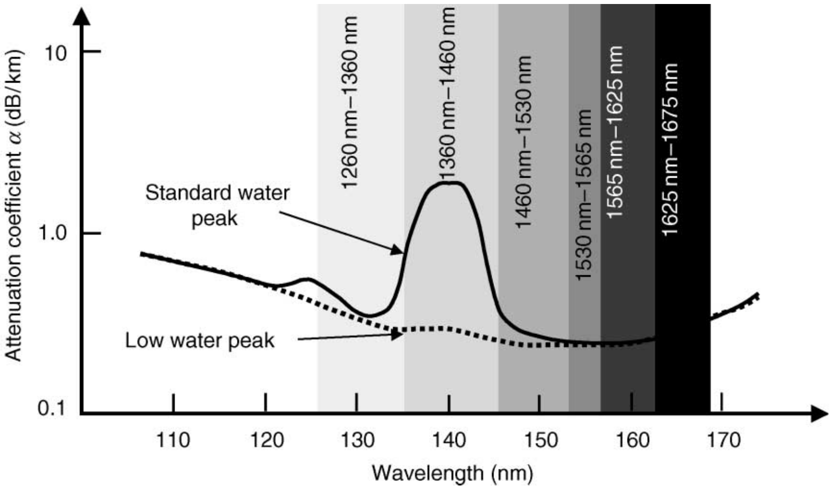

Figure 2 Typical spectral attenuation of a singlemode fiber.

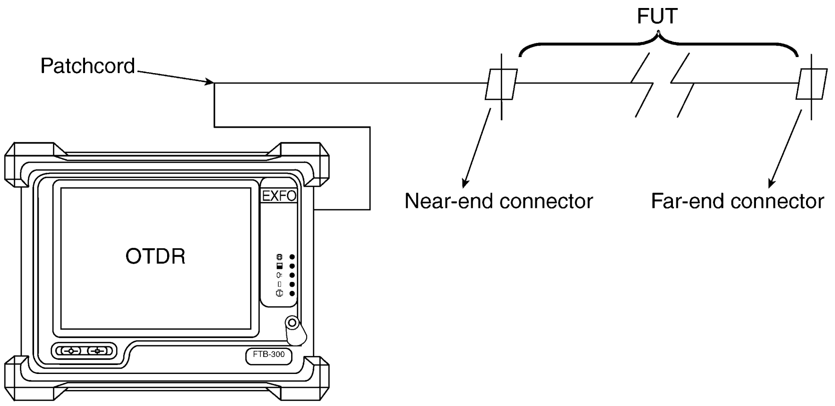
FUT
Patchcord
OTDR
EXFO
FTB-300
Near-end connector
Far-end connector

Figure 3 The backscattering method (OTDR).

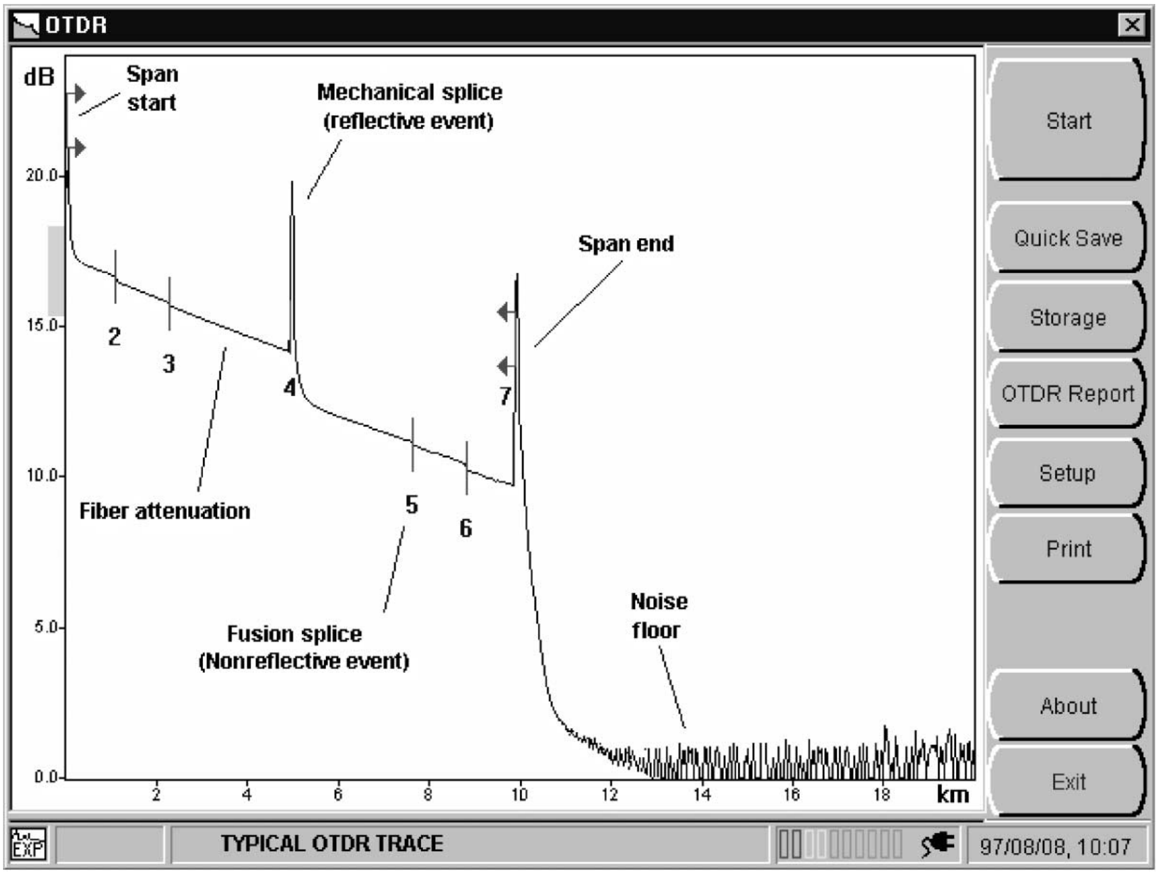
OTDR
dB
Span start
Mechanical splice (reflective event)
Span end
Fiber attenuation
Fusion splice (Nonreflective event)
Noise floor
2
3
4
5
6
7
20.0
15.0
10.0
5.0
0.0
2
4
6
8
10
12
14
16
18
km
Start
Quick Save
Storage
OTDR Report
Setup
Print
About
Exit
EXP
TYPICAL OTDR TRACE
97/08/08, 10:07

Figure 4 Unidirectional OTDR backscattering loss measurement.

curve together with the FUT group index of refraction, n_g, as:

$$L_f = c_{fs} \cdot \frac{\Delta T_f}{n_g} \qquad [7]$$

c_{fs} is the velocity of light in free space.

The bidirectional backscattering curve is obtained using two unidirectional backscattering curves and calculating the average loss between them. The end-to-end FUT attenuation coefficient is obtained from the difference between two losses divided by the difference of their corresponding distances.

Insertion loss method. The insertion loss method consists of the measurement of the power loss due to the FUT insertion between a launching and a receiving system, previously interconnected (reference condition). The powers P_1 and P_2 are evaluated in a less straightforward way than in the cut-back method. Therefore, this method is not intended for use in manufacturing.

The insertion loss technique is less accurate than the cut-back, but has the advantage of being non-destructive for the FUT. Therefore, it is particularly suitable in the field.

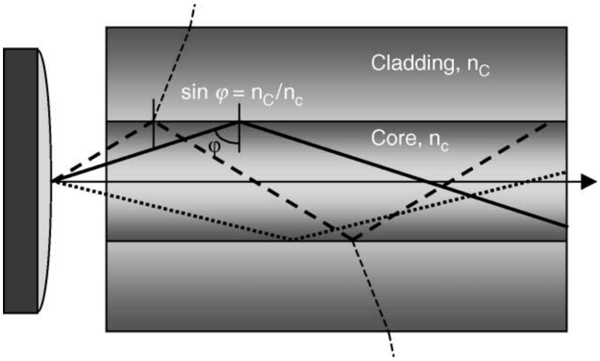
Cladding, n_C
$\sin\varphi = n_C/n_c$
Core, n_c
φ

Figure 5 Total internal reflection and macrobending effect on the light rays.

Macrobending Sensitivity

Macrobending sensitivity is the property by which there is a certain amount of light leaking (loss) into the cladding when the fiber is bent and the bending angle is such that the condition of total internal reflection is no longer met at the core-cladding interface. Figure 5 illustrates such a case.

Macrobending sensitivity is a direct function of the wavelength: the longer the wavelength and/or the smaller the bending diameter, the more loss the fiber experiences. It is recognized that 1625 nm is a wavelength that is very sensitive to macrobending (see Figure 6).

The macrobending loss is measured by the power monitoring method (OTDR, especially for field assessment, see Figure 6) or the cut-back method.

Microbending Sensitivity

Microbending is a fiber property by which the core-cladding concentricity randomly changed along

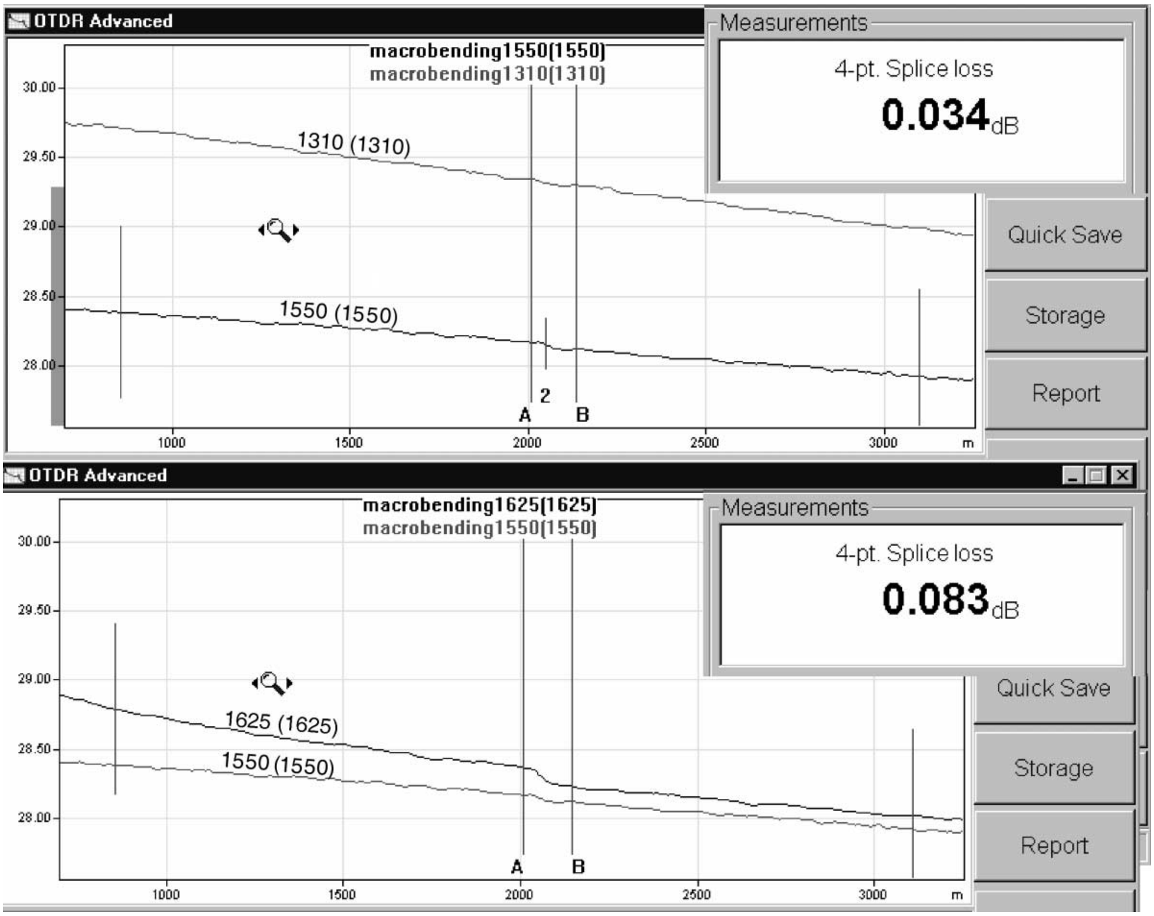
OTDR Advanced
macrobending1550(1550)
macrobending1310(1310)
30.00
29.50
29.00
28.50
28.00
1310 (1310)
1550 (1550)
2
A
B
1000
1500
2000
2500
3000
m
Measurements
Quick Save
Storage
Report
OTDR Advanced
macrobending1625(1625)
macrobending1550(1550)
30.00
29.50
29.00
28.50
28.00
1625 (1625)
1550 (1550)
A
B
1000
1500
2000
2500
3000
m
Measurements
Quick Save
Storage
Report

4-pt. Splice loss
0.034dB

4-pt. Splice loss
0.083dB

Figure 6 Macrobending sensitivity as a function of wavelength.

the fiber length. It causes the core to wobble inside the cladding and along the fiber length.

Four methods are available for characterizing microbending sensitivity in optical fibers:

- expandable drum for singlemode fibers and optical fiber ribbons over a wide range of applied linear pressure or loads;
- fixed diameter drum for step-index multimode, singlemode, and ribbon fibers for a fixed linear pressure;
- wire mesh and applied loads for step-index multimode and singlemode fibers over a wide range of applied linear pressure or loads; and
- 'basketweave' wrap on a fixed diameter drum for singlemode fibers.

The results from the four methods can only be compared qualitatively. The test is nonroutine for general evaluation of optical fiber.

Cut-off Wavelength

The cut-off wavelength is the shortest wavelength at which a single mode can propagate in a singlemode fiber. This parameter can be computed from the fiber refractive index profile (RIP). At wavelengths below the cut-off wavelength, several modes propagate and the fiber is no longer singlemode, but multimode.

In optical fibers, the change from multimode to singlemode behavior does not occur at a specific wavelength, but rather over a smooth transition as a function of wavelengths. Consequently, from a fiber-optic network standpoint, the actual threshold wavelength for singlemode performance is more critical. Thus an effective cut-off wavelength is described below.

The cut-off wavelength is defined as the wavelength greater than the ratio between the total power, in the higher-order modes and the fundamental mode, which has decreased to less than 0.1 dB. According to this definition, the second-order mode LP_{11} undergoes 19.3 dB more attenuation than the fundamental LP_{01} mode when the modes are equally excited.

Because the cut-off wavelength depends on the fiber length, bends, and strain, it is defined on the basis of the following three cases:

- fiber cut-off wavelength;
- cable cut-off wavelength; and
- jumper cable cut-off wavelength.

Fiber cut-off wavelength: Fiber cut-off wavelength λ_{fc} is defined for uncabled primary-coated fiber and is measured over 2 m with one loop of 140 mm radius loosely constrained, with the rest of the fiber kept essentially straight. The presence of a primary coating on the fiber usually will not affect λ_{fc}. However, the presence of a secondary coating may result in λ_{fc} being significantly shorter than that of the primary coated fiber.

Cable cut-off wavelength: Cable cut-off wavelength is measured prior to installation on a substantially straight 22 m cable length prepared by exposing 1 m of primary-coated fiber at both ends, the exposed ends each incorporating a 40 mm radius loop. Alternatively, this parameter may be measured on 22 m primary-coated uncabled fiber in the same configuration as for the λ_{fc} measurement.

Jumper cable cut-off wavelength: Jumper cable cut-off wavelength is measured over 2 m with one loop of 76 mm radius, or equivalent (e.g., split mandrel), with the rest of the jumper cable kept essentially straight.

Multimode Fiber Bandwidth for Multimode Fibers

The −3 dB bandwidth of a multimode optical fiber (or modal bandwidth) is defined as the lowest frequency where the magnitude of the baseband frequency response in optical power has decreased by 3 dB relative to the power at zero frequency. Modal bandwidth is also called intermodal dispersion as it takes into account the dispersion between the modes of propagation of the transmitted signal into the multimode fiber.

Various methods of reporting the results are available, but the results are typically expressed in terms of the −3 dB (optical power) frequency. Figure 7 illustrates modal bandwidth.

The bandwidth or pulse broadening may be normalized to a unit length, such as GHz · km, or ns/km.

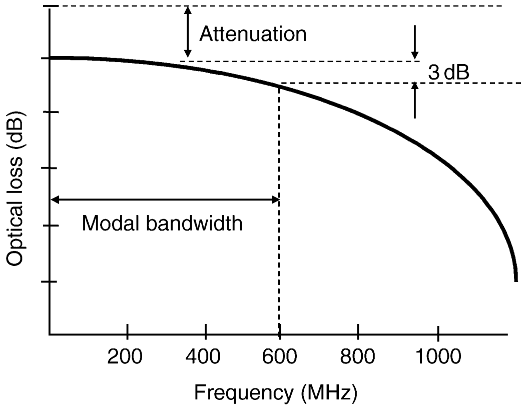

Figure 7 Determination of modal bandwidth.

Two methods are available for determining transmission capacity of multimode fibers:

- the frequency domain measurement method in which the baseband frequency response is directly measured in the frequency domain by determining the fiber response to a sinusoidally modulated light source; or
- the optical time domain measurement method (pulse distortion) in which the baseband response is measured by observing the broadening of a narrow pulse of light.

Differential Mode Delay for Multimode Fibers

Differential mode delay (DMD) characterizes the modal structure of a graded-index glass-core multimode fiber. DMD is useful for assessing the bandwidth performance of a fiber when used with short-pulse, narrow spectral-width laser sources.

The output from a singlemode probe fiber excites the multimode FUT at the test wavelength. The probe spot is scanned across the FUT endface, and the optical pulse delay is determined at specified radial offset positions between an inner and an outer limit. DMD is the difference in optical pulse delay time between the FUT fastest and slowest modes excited for all radial offset positions between and including the inner and the outer limits.

The related critical issues influencing DMD are the temporal width of the optical pulse, jitter in the timing, the finite bandwidth of the optical detector, and the mode broadening due to the source spectral width and the FUT chromatic dispersion.

The test method is commonly used in production and research facilities, but is not easily accomplished in the field. DMD may be a good predictor of the source launching conditions. DMD may be normalized to a unit length, such as ps/m.

Chromatic Dispersion

Chromatic dispersion in a singlemode fiber is a combination of material dispersion and waveguide dispersion (see Figure 8), and it contributes to pulse broadening and distortion in a digital signal.

Material dispersion is produced by the dopants used in glass and is important in all fiber types. Waveguide dispersion is produced by the wavelength dependence of the index of refraction and is critical in singlemode fibers only.

From the point of view of the transmitter, this is due to two causes:

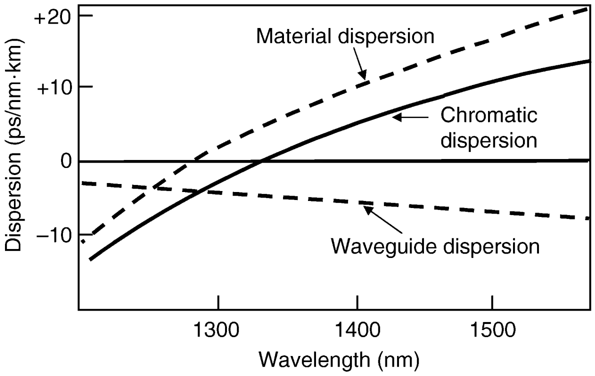

Figure 8 Contribution of the material and waveguide dispersions to the chromatic dispersion.

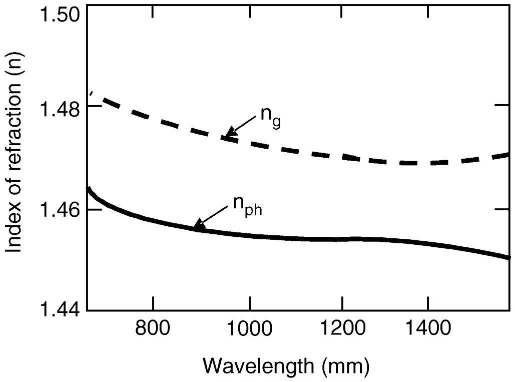

Figure 9 Difference between the phase and the group index of refraction.

- The presence of wavelengths in the source optical spectrum. Each wavelength has a different phase delay and group delay (different group velocities) along the fiber, because they travel under different index of refraction (or phase) as the index varies as a function of wavelengths, as shown in Figure 9.
- The other cause is the modulation of the source, which itself has two effects:
  - As bit-rates increase, the spectral width of the modulated signal increases and can become comparable to or exceed the spectral width of the source.
  - Chirp occurs when the source wavelength spectrum varies during the pulse. By convention, positive chirp at the transmitter occurs when the spectrum during the rise/fall of the pulse shifts towards shorter/longer wavelengths respectively. For a positive fiber dispersion coefficient, longer wavelengths are delayed relative to shorter wavelengths. Hence if the sign of the product of chirp and dispersion is positive, the two processes combine to produce pulse broadening. If the product is negative, pulse compression can occur over an initial fiber length until the pulse reaches a minimum width and then broadens again with increasing dispersion.

The electric field propagating into the FUT may be simply described as follows:

$$E(t, z) = E_0 \sin(\omega t - \beta z) \qquad [8]$$

$\omega = 2\pi c/\lambda$ [rad/s] is the angular frequency; $\beta = kn = (\omega/c)n$ is the effective index; as β has units of m^{-1} it is often referred to as wavenumber or even sometimes propagation constant; k is the propagation constant.

The group delay τ_g is then given by:

$$\mathrm{d}\beta/\mathrm{d}\omega = \beta_1 = \tau_g \qquad [9]$$

An example of the group delay spectral distribution is shown in Figure 10. Assuming n is not complex; β_1 is the first-order derivative of β.

The group velocity v_g is given by

$$v_g = (\mathrm{d}\beta/\mathrm{d}\omega)^{-1} = -(\lambda^2/2\pi c)(\mathrm{d}\beta/\mathrm{d}\lambda)^{-1} \qquad [10]$$

The FUT input–output delay is given by

$$\tau_d = L/v_g \qquad [11]$$

L is the FUT length.

The group index of refraction n_g is given by

$$n_g = c/v_g = n - \lambda(\mathrm{d}n/\mathrm{d}\lambda) \qquad [12]$$

The dispersion parameter or dispersion coefficient D (ps/nm · km) is given by

$$D = -(\omega/\lambda)(\mathrm{d}\tau_g/\mathrm{d}\omega) = -(2\pi c/\lambda^2)(\mathrm{d}^2\beta/\mathrm{d}\omega^2)$$

$$= -(\lambda/c)(\mathrm{d}^2 n/\mathrm{d}\lambda^2) \qquad [13]$$

$$\mathrm{d}^2\beta/\mathrm{d}\omega^2 = \beta_2 \qquad [14]$$

An example of the spectral distribution of D obtained from the group delay is shown in Figure 10.

β_2 (ps^2/km) is the group velocity dispersion parameter, so D may be related to β_2, as follows:

$$D = -(\omega/\lambda)\beta_2 \qquad [15]$$

When β_2 is positive then D is negative and vice-versa. The region where β_2 is positive is called normal dispersion while the negative-β_2 region is called anomalous dispersion.

At λ_0, β_1 is minimum, and $\beta_2 = 0$, then $D = 0$.

The dispersion slope S (ps/nm^2 · km), also called the differential dispersion parameter or second-order dispersion, is given by

$$S = \mathrm{d}D/\mathrm{d}\lambda = (\omega/\lambda)\beta_3 + (2\omega/\lambda^2)\beta_2 \qquad [16]$$

$$\beta_3 = \mathrm{d}\beta_2/\mathrm{d}\omega = \mathrm{d}^3\beta/\mathrm{d}\omega^3$$

At λ_0, β_1 is minimum, $\beta_2 = 0$, then $D = 0$; but S is not zero and depends on β_3. An example of the spectral distribution of S and S_0 is illustrated in Figure 10.

Overall, the general expression of β is given by

$$\beta(\omega) = \beta_0 + (\omega - \omega_0)\beta_1 + (1/2)(\omega - \omega_0)^2\beta_2$$

$$+ (1/12)(\omega - \omega_0)^3\beta_3 + \cdots \qquad [17]$$

Figure 11 illustrates the difference between dispersion unshifted fiber (ITU-T Rec. G.652), dispersion shifted fiber (ITU-T Rec. G.653) and nonzero dispersion shifted fiber (ITU-T Rec. G.655).

Test methods for the determination of chromatic dispersion

All methods measure the group delay at a specific wavelength over a range and use agreed fitting functions to evaluate λ_0 and S_0.

In the phase shift method, the group delay is measured in the frequency domain, by detecting, recording, and processing the phase shift of a sinusoidal modulating signal between a reference, a secondary fiber path, and the channel signal.

Setup variances exist and some do not require the secondary reference path. For instance, by using a reference optical filter at the FUT output it is possible to completely decouple the source from the phase-meter. With such an approach, chromatic dispersion may now be measured in the field over very long links using optical amplifiers (see Figure 12).

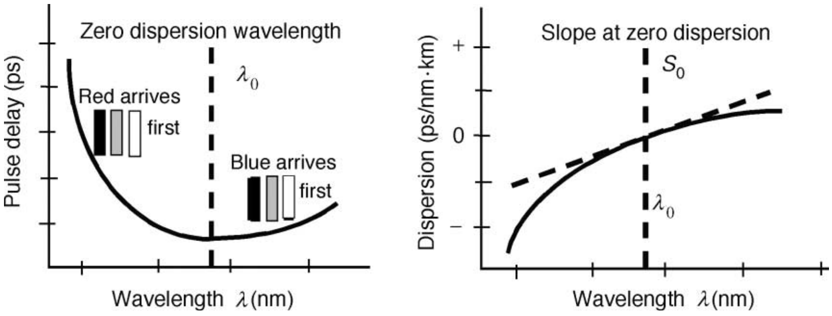

Figure 10 Relation between the pulse (group) delay and the (chromatic) dispersion.

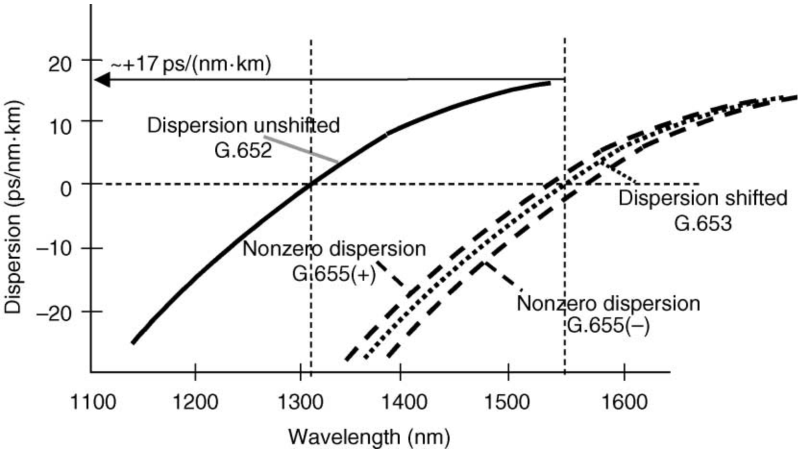

Figure 11 Chromatic dispersion for various types of fiber.

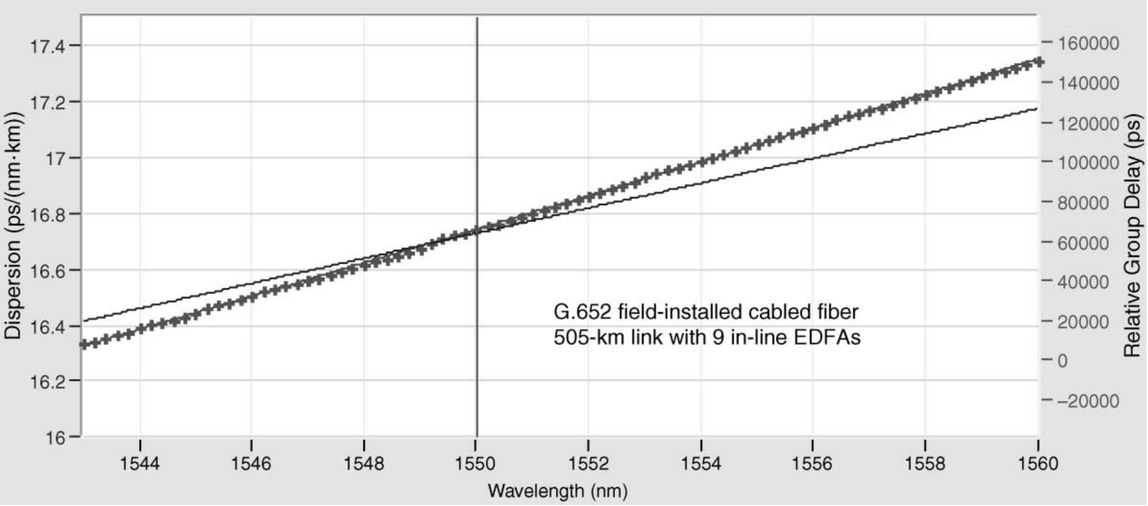

Figure 12 Test results for field-related chromatic dispersion using the phase-shift technique.

In the differential phase shift method, two detection systems are used together with two wavelength sources at the same time. In this case the chromatic dispersion may be determined directly from the two group delays. This technique usually offers faster and more reliable results but costs much more than the phase-shift technique which is usually preferred.

In the interferometric method, the group delay between the FUT and a reference path is measured by a Mach–Zehnder interferometer. The reference delay line may be an air path or a singlemode fiber standard reference material (SRM). The method can be used to determine the following characteristics:

- longitudinal chromatic dispersion homogeneity; and
- effect of overall or local influences, such as temperature changes and macrobending losses.

In the pulse delay method, the group delay is measured in the time domain, by detecting, recording, and processing the delay experienced by pulses at various wavelengths.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) causes an optical pulse to spread in the time domain and may impair the performance of a telecommunications system. The effect can be related to differential phase and group velocities and corresponding arrival time of different polarization components of the pulse signal. For a sufficiently narrowband source, the effect can be related to a differential group delay (DGD), $\Delta\tau$, between a pair of orthogonally polarized principal states of polarization (PSP) at a given wavelength or optical frequency (see Figure 13a).

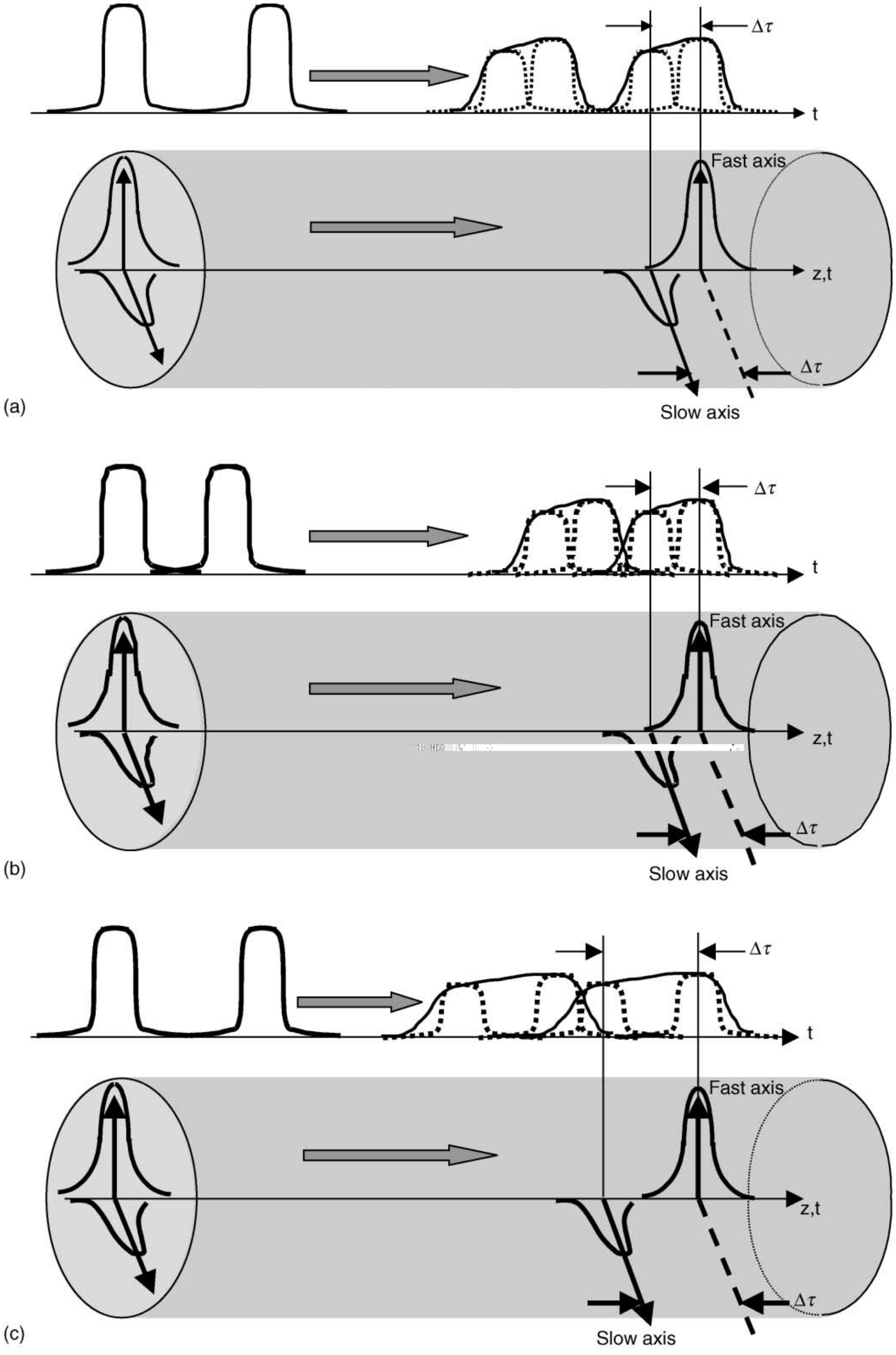
Δτ
t
Fast axis
z,t
Δτ
(a)
Slow axis
Δτ
t
Fast axis
z,t
Δτ
(b)
Slow axis
Δτ
t
Fast axis
z,t
Δτ
(c)
Slow axis

Figure 13 PMD effect on the pulse broadening and its possible pulse impairment. (a) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is too low to create an impairment. (b) The pulse is spread by the DGD $\Delta\tau$ but the bit rate is high enough to create an impairment. (c) The DGD $\Delta\tau$ is large enough even at low bit rate to make the pulse spreading and creating impairment.

In an ideal circular symmetric fiber, the two PSPs propagate with the same velocity. However:

- a real fiber is not perfectly circular;
- the core is not perfectly concentric with the cladding;
- the core may be subjected to microbending;
- the core may present localized clusters of dopants; and
- the environmental conditions may stress the deployed cable and affect the fiber.

Each time the fiber undergoes local stresses and consequently birefringence. These asymmetry characteristics vary randomly along the fiber and in time, lead to a statistical behavior of PMD.

For a deployed cabled fiber at a given time and optical frequency, there always exist two PSPs such that the pulse spreading due to PMD vanishes, if only one PSP is excited. On the contrary, the maximum pulse spread due to PMD occurs when both PSPs are equally excited, and is related to the difference in their group delays, the DGD associated with the two PSPs. For broadband transmission, the DGD statistically varies as a function of wavelengths or frequencies and result in an output pulse that is spread out in the time domain (see Figures 13a–c). In this case, the spreading can be related to the RMS (root mean square) of DGD values $\langle \Delta\tau^2 \rangle^{1/2}$. However, if a known distribution such as the Maxwell distribution may be fit to the DGD distribution probability, then a mean (or average) value of the DGD $\langle \Delta\tau \rangle$ may be correlated to the RMS value and used as a system performance predictor in particular with a maximum value of the DGD distribution associated to a low probability of occurrence. This maximum DGD may then be used to define the quality of service that would tolerate values lower than this maximum DGD.

Test methods for polarization mode dispersion

Three methods are generically used for measuring PMD. Other methods or analyses may exist but they are generally not standardized or are limited in their applications.

- Stokes parameter evaluation (SPE)
  - Jones matrix eigenanalysis (JME)
  - Poincaré sphere analysis (PSA)
- Fixed analyzer (FA)
  - Extrema counting (EC)
  - Fourier transform (FT)
- Interferometry (INTY)
  - Traditional analysis (TINTY)
  - General analysis (GINTY).

All methods use a linearly polarized source at the FUT input and are suitable for laboratory measurements of factory lengths of fiber and cable. However, the interferometric method is the only method appropriate for measurements of cabled fiber that may be moving or vibrating such as is found in the field.

Stokes parameter evaluation. SPE determines PMD by measuring a response to a change of narrowband light (from a tuneable light source with broadband detector – JME, or a broadband source with a filtered detector such as an interferometer – PSA) across a wavelength range. The Stokes vector of the output light is measured for each wavelength. The change of these Stokes vectors with angular optical frequency (wavelength), ω and with the change in input SOP (state of polarization), yields the DGD as a function of wavelength.

For both JME and PSA analyses, three distinct and known linear SOPs (orthogonal on the Poincaré sphere) must be launched for each wavelength. Figure 14 illustrates the test setup and examples of test results.

The JME and PSA method are mathematically equivalent.

Fixed analyzer. FA determines PMD by measuring a response to a change of narrowband light across a wavelength range. For each SOP, the change in output power that is filtered through a fixed polarization analyzer, relative to the power detected without the analyzer, is measured as a function of wavelength. Figure 15 illustrates a test setup and examples of test results.

The resulting measured function can be analyzed in one of two ways:

- by counting the number of peaks and valleys (EC) of the curve and application of a formula. This analysis is considered as a frequency domain approach; and
- by taking the Fourier transform (FT) of the measured function. This FT is equivalent to the pulse spreading obtained by TINTY.

Interferometry. INTY uses a broadband light source and an interferometer. The fringe pattern containing the source auto-correlation, together with the PMD related cross-correlation of the emerging electromagnetic field, is determined by the interference pattern of the output light, i.e., the interferogram. The PMD determination for the wavelength range associated with the source spectrum is based on the envelope of the fringe

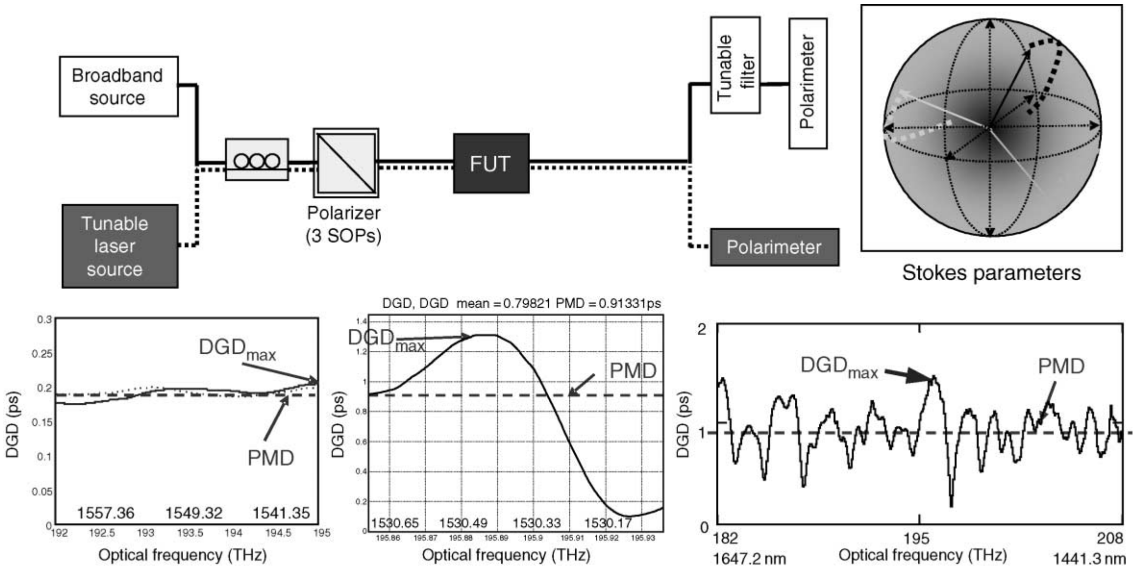
Broadband source
Tunable laser source
Polarizer (3 SOPs)
FUT
Tunable filter
Polarimeter
Polarimeter
Stokes parameters
DGD (ps)
DGD$_{max}$
PMD
1557.36 1549.32 1541.35
Optical frequency (THz)
DGD, DGD mean = 0.79821 PMD = 0.91331ps
DGD$_{max}$
PMD
1530.65 1530.49 1530.33 1530.17
Optical frequency (THz)
DGD$_{max}$
PMD
182 1647.2 nm
195
208 1441.3 nm
Optical frequency (THz)

Figure 14 PMD by Stokes parameter evaluation method.

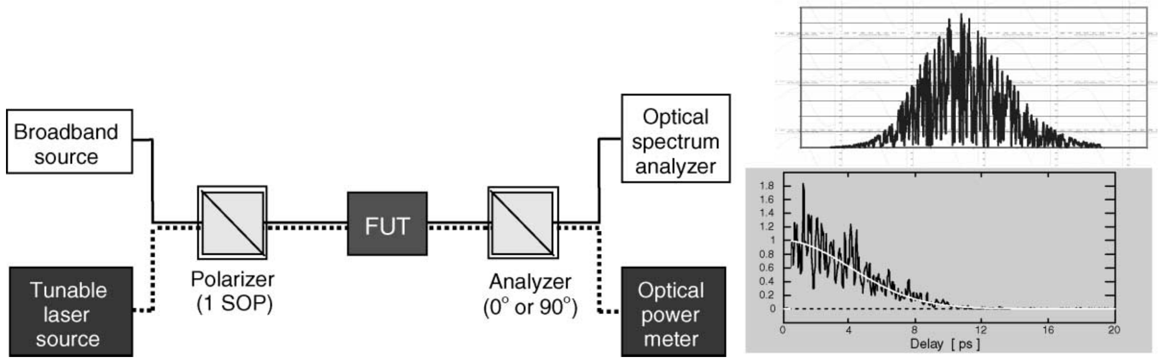
Broadband source
Tunable laser source
Polarizer (1 SOP)
FUT
Analyzer (0° or 90°)
Optical spectrum analyzer
Optical power meter
Delay [ps]

Figure 15 PMD by fixed analyzer method.

pattern interferogram. Two analyses are available to obtain the PMD:

- TINTY uses a set of specific operating conditions for its successful applications and a basic setup; and
- GINTY uses no limiting operating conditions, but in addition to the same basic setup, also using a modified setup compared to TINTY.

Figure 16 illustrates the test setup for both approaches and examples of test results.

Polarization Crosstalk

Polarization crosstalk is a characteristic of energy mixing/transfer/coupling between the two PSPs in a PMF (polarization maintaining fiber) when their isolation is imperfect. It is the measure of the strength of mode coupling or output power ratio between the PSPs within a PMF.

A PMF is an optical fiber capable of transmitting, under external perturbations, such as bending or lateral pressures, both HE_{11}^{x} and HE_{11}^{y} polarization modes whose electric field vector directions are orthogonally to each other and which have different propagation constants β_x and β_y.

Two methods are available for measuring the polarization crosstalk of PMF:

- *power ratio method*, which uses the maximum and minimum values of output power at a specified wavelength, and is applicable to fibers and connectors jointed to a PMF, and to two or more PMFs joined in series; and
- *in-line method*, which uses an analysis of the Poincaré sphere, and is applicable to single or cascaded sections of PMF, and to PMF interconnected with optical devices.

Nonlinear Effects

When the power of the transmission signal is increased to achieve longer span lengths at high bit

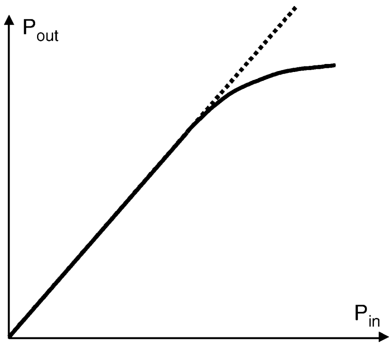

Figure 17 Power output-to-power input relationship for production of nonlinear effects.

expressed as follows:

$$n = n_0 + n_2 I \qquad [18]$$

n is the nonlinearity dependent index; n_0 is the linear part of the index; n_2 is the nonlinear index, also called the Kerr nonlinear index (2.2 to 3.4×10^{-16} cm^2/W); and I is the optical intensity inside the fiber.

The field propagation at a distance L into the fiber is described by the following equation:

$$E_{out}(L) = E_{in}(0)\exp[-a/2 + i\beta + \gamma P(L,t)/2]L \qquad [19]$$

$a/2$ is the attenuation; $i\beta$ is the phase of the wave; and $\gamma P(L,t)/2$ is the nonlinearity term;

$$\gamma = 2\pi n_2/\lambda A_{eff} \qquad [20]$$

γ is the nonlinearity coefficient and may be a complex number; A_{eff} is the fiber core effective area; $P(L,t)$ is the total power; and λ is the signal wavelength and t the time variable.

The nonlinear coefficient is defined as n_2/A_{eff}. This coefficient plays a critical role in the fiber and is closely related to system performance degradation due to nonlinearities when very high power is used.

Methods for measuring the nonlinear coefficient. Two methods are available for measuring the nonlinear coefficient:

- Continuous-wave dual-frequency (CWDF); and
- Pulsed single-frequency (PSF).

In the CWDF method, light from two wavelengths is injected into the fiber. At higher power, the light from the two wavelengths beat due to the nonlinearity and produce an output spectrum that is spread. The relationship of the power level to a particular spreading is used to calculate the nonlinear coefficient.

In the PSF method, the pulsed light from a single wavelength is injected into the fiber. Very short pulses (<1 ns) and their input peak power must be measured and related to the nonlinear spreading of the output spectrum.

Stimulated Brillouin scattering

In an intensity modulated system using a source with a narrow linewidth, significant optical power is transferred from the forward-propagating signal to a backward-propagating signal when the SBS power threshold is exceeded. At that point periodic regions of index of refraction produce a grating traveling at speed of sound away from the source. The grating reflects backward part of the incident light. The reflected sound waves (acoustic phonons) scatter light back to the source. Phase matching (or momentum conservation) dictates that the scattered light preferentially travels in the backward direction. The scattered light will be Doppler-shifted (downshifted or Brillouin-shifted) by approximately 11 GHz (at 1550 nm, for G.652 fiber). The scattered light has a very narrow spectrum (very coherent) and very close to the carrier signal and may be very detrimental.

Stimulated Raman scattering

SRS is an interaction between light and the fiber molecular vibrations as adjacent atoms vibrate in opposite directions (an 'optical phonon'). Some of the energy of the main carrier (optical pump wave) is transferred to the molecules, thereby further increasing the amplitude of their vibrations. If the vibrations become large enough, a threshold is reached at which the local index of refraction changes. These local changes then scatter light in all directions similar to Rayleigh scattering. However, unlike Rayleigh scattering, the wavelength of the Raman scattered light is shifted to longer wavelengths by an amount that corresponds to the molecular vibration frequencies and the Raman signal spreads over a large spectrum.

Self-phase modulation

SPM is the effect that a powerful pulse has on its own phase, considering that in eqn [18], $I(t)$ varies in time:

- $I(t) \rightarrow n(t) = n_0 + n_2 I(t) \rightarrow$ modulates the phase $\beta(t)$ of the pulse; and
- $dI/dt \rightarrow dn/dt \rightarrow d\beta/dt$(chirp) → broadening in the frequency domain → broadening in the time domain.

$I(t)$ peaks at the center of the pulse (peak power) and consequently increases the index of refraction. A higher index causes the wavelengths in the center of the pulse to accumulate phase more quickly than at the wings of the pulse:

- this causes wavelength stretching (shift to longer wavelengths) at pulse leading edge (risetime); and
- this causes wavelength compression (shift to shorter wavelengths) at pulse trailing edge (falltime).

The pulse will then broaden with negative (normal) dispersion and shorten with positive (anomalous) dispersion. SPM may then be used for dispersion compensation considering that self-phase modulation imposes $C > 0$ (positive chirp). It can cancel the dispersion if properly manage as a function of the sign of the dispersion. SPM is one of the most critical nonlinear effects for the propagation of soliton or very short pulses over very long distance.

Cross-phase modulation

XPM is the effect that a powerful pulse has on the phase of an adjacent pulse from another WDM system channel traveling in phase or at slightly the same group velocity. It concerns spectral interference between two WDM channels:

- the increasing $I(t)$ at the leading edge of the interfering pulse shifts the other pulse to longer wavelength; and
- decreasing $I(t)$ at the trailing edge of the interfering pulse shifts the other pulse to shorter wavelengths.

This produces spectral broadening, which dispersion converts to temporal broadening depending on the sign of the dispersion. XPM effect is similar to SPM except it depends on the channel count.

Four-wave (four-photon) mixing

FWM is the by-product production effect from two or more WDM channels. For two channels $I(t)$ modulates the phase of each signal (ω_1 and ω_2). An intensity modulation appears at the beat frequency $\omega_1 - \omega_2$.

Two sideband frequencies are created in a similar way as harmonics generation. New wavelengths are created in a number equal to $N^2(N-1)/2$, where $N =$ number of original wavelengths.

Fiber Dimension Characteristics and Corresponding Tests Methods

Table 1 provides a list of the various fiber dimensional characteristics and their corresponding test methods.

Table 1 Fiber dimensional characteristics

Attribute	*Measured parameter*
Fiber geometry	Core/cladding diameter Core/cladding noncircularity Core-cladding concentricity error
Numerical aperture	
Mode field diameter	
Coating geometry	
Length	

Fiber Geometry Characteristics

The fiber geometry is related to the core and cladding characteristics.

Core

The core center is the center of a circle which best fits the points at a constant level in the near-field intensity profile emitted from the central region of the fiber, using wavelengths above and/or below the cut-off wavelength.

The RIP can be measured by refracted near field (RNF) or transverse interferometry techniques and transmitted near field (TNF).

The core concentricity error is the distance between the core center and the cladding center. This definition applies very well for multimode fibers. The distance between the center of the near field profile and the center of the cladding is also used for singlemode fibers.

The mode field diameter (MFD) represents a measure of the transverse electromagnetic field intensity of the mode in a fiber cross-section and it is defined from the far-field intensity distribution.

The MF is the singlemode field distribution of the LP_{01} mode, giving rise to a spatial intensity distribution in the fiber.

The MF concentricity error is the distance between the MF center and the cladding center.

The core noncircularity is a measure of the core ellipticity. This parameter is one of the causes for creating birefringence in the fiber and consequently PMD.

Cladding

The cladding is the outermost region of constant refractive index in the fiber cross-section.

The cladding center is the center of a circle best fitting the outer limit (boundary) of the cladding.

The cladding diameter is the diameter of the circle defining the cladding center.

The cladding noncircularity is a measure of the difference between the diameters of the two circles defined by the cladding tolerance field divided by the nominal cladding diameter.

Coating

The primary coating is one or more layers of protective material applied to the cladding during or after the drawing process to protect the cladding surface (e.g., a 250 μm protective coating). The secondary coating is one or more layers of protective material applied over the primary coating in order to give additional protection or to provide a particular structure.

Measurement of the fiber geometrical attributes

The fiber geometry is measured by the following methods:

- TNF;
- RNF;
- Side-view technique/transverse interference;
- TNF image technique; and
- Mechanical diameter.

Test instrumentation may incorporate two or more methods such as the one shown in Figure 18.

Transmitted near-field technique. The cladding diameter, core concentricity error, and cladding noncircularity are determined from the near-field intensity distribution. Figure 19 provides a series of examples of test results from TNF measurements.

Refracted near-field technique. The RIP across the entire fiber (core and cladding) can be directly obtained from the RNF measurement, as shown in Figure 20.

The geometrical characteristics of the fiber can be obtained from the refractive index distribution using suitable algorithms:

- core/cladding diameter;
- core/cladding concentricity error;
- core/cladding noncircularity;
- maximum numerical aperture (NA); and
- index and relative index of refraction difference.

Figure 21 illustrates the core geometry.

Side-view technique/transverse interference. The side-view method is applied to singlemode fibers to determine the core concentricity error, cladding diameter and cladding noncircularity by measuring the intensity distribution of light that is refracted inside the fiber. The method is based on an interference microscope focused on the side view of an FUT illuminated perpendicular to the FUT axis. The fringe pattern is used to determine the RIP.

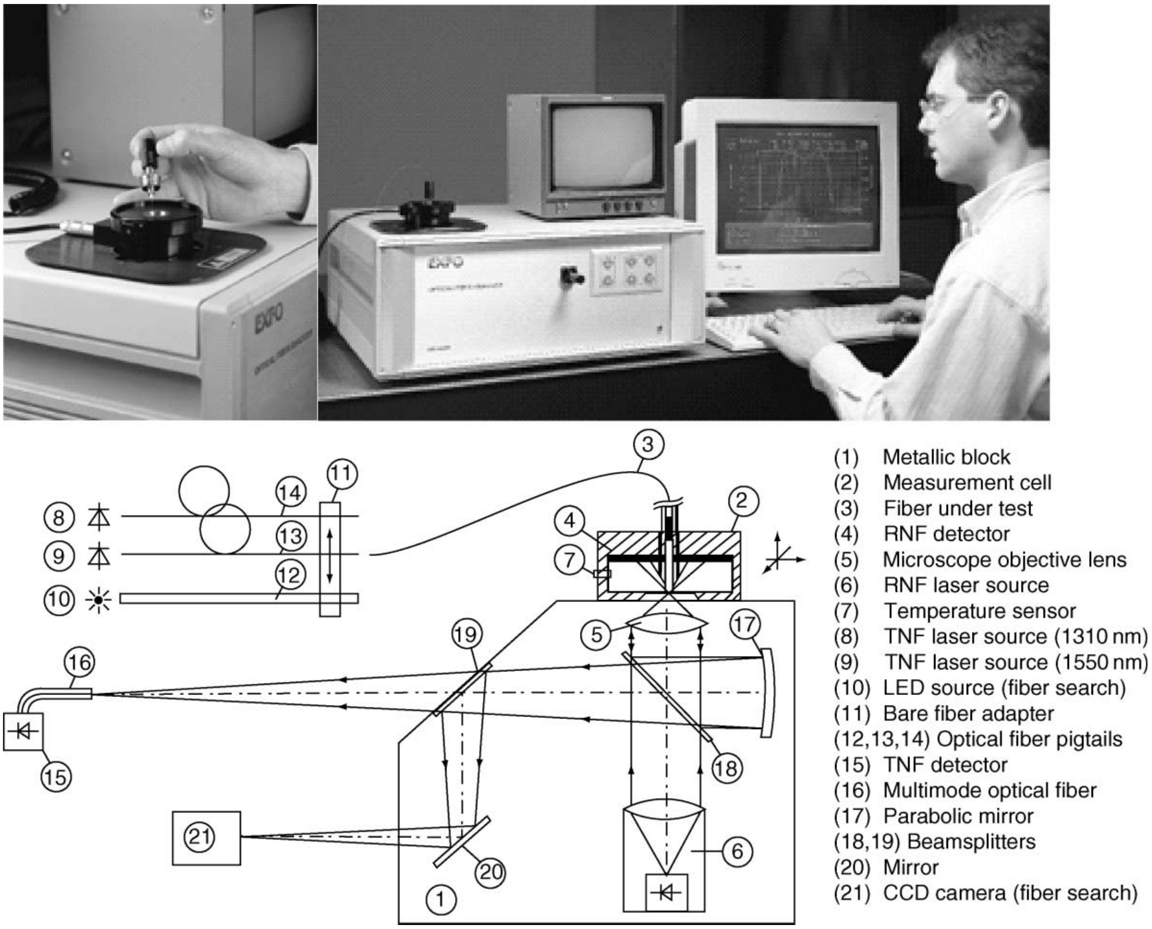

(1) Metallic block
(2) Measurement cell
(3) Fiber under test
(4) RNF detector
(5) Microscope objective lens
(6) RNF laser source
(7) Temperature sensor
(8) TNF laser source (1310 nm)
(9) TNF laser source (1550 nm)
(10) LED source (fiber search)
(11) Bare fiber adapter
(12,13,14) Optical fiber pigtails
(15) TNF detector
(16) Multimode optical fiber
(17) Parabolic mirror
(18,19) Beamsplitters
(20) Mirror
(21) CCD camera (fiber search)

Figure 18 RNF/TNF combined instrumentation.

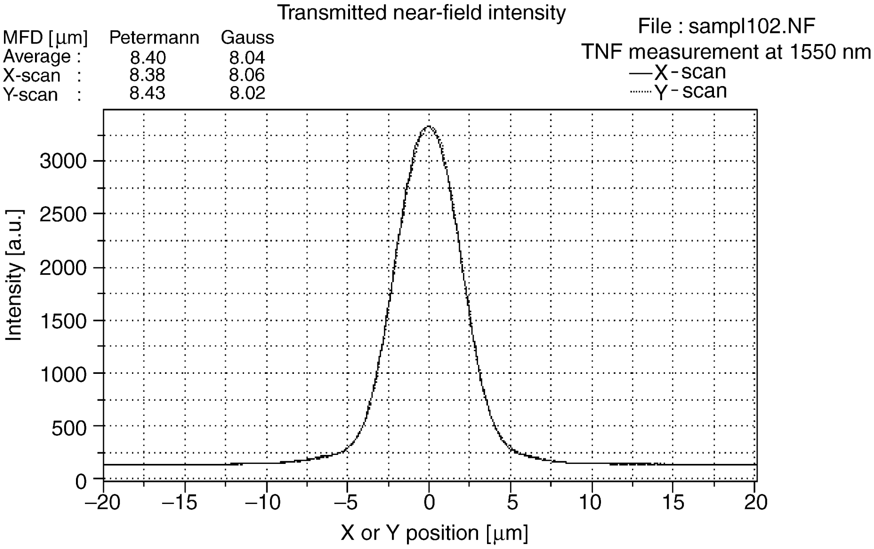

Figure 19 MFD measurement by TNF.

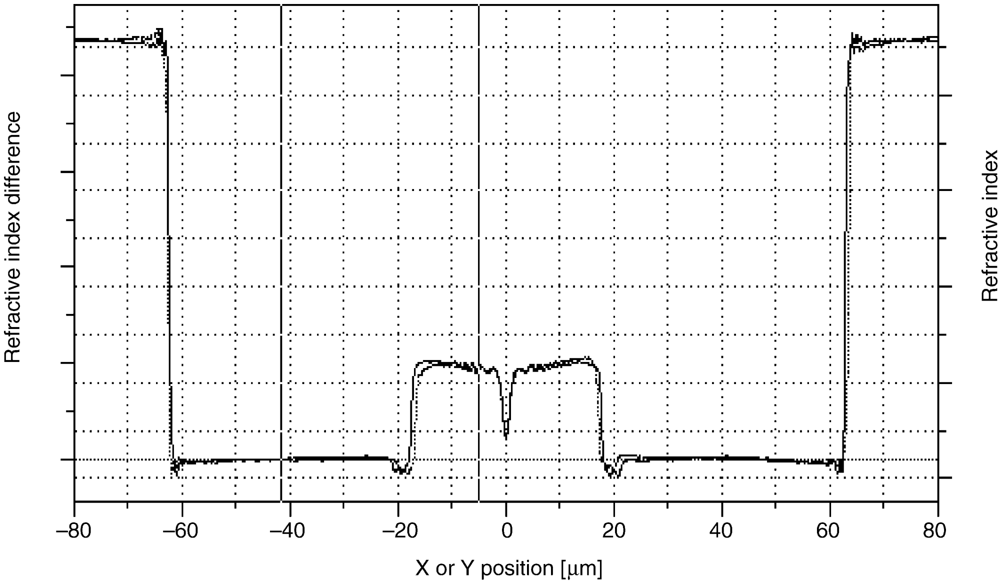

Figure 20 RIP from RNF measurement.

TNF image technique. The TNF image technique, also called near-field light distribution, is used for the measurement of the geometrical characteristics of singlemode fibers. The measurement is based on analysis of magnified images at the FUT output. Two subsets of the method are available:

- grey-scale technique which performs an x–y near-field scan using a video system; and
- Single near-field scanning technique performing a one-dimensional scan.

Mechanical diameter. This is a precision mechanical diameter measurement technique used to accurately determine the cladding diameter of silica fibers. The technique uses an electronic micrometer such as based on a double-pass Michelson interferometer. The technique is used for providing calibrated fibers to the industry as SRM.

Numerical Aperture

The NA is an important attribute for multimode fibers in order to predict their launching efficiency, joint loss at splices and micro/macrobending characteristics.

A method is available for the measurement of the angular radiant intensity distribution (far-field)

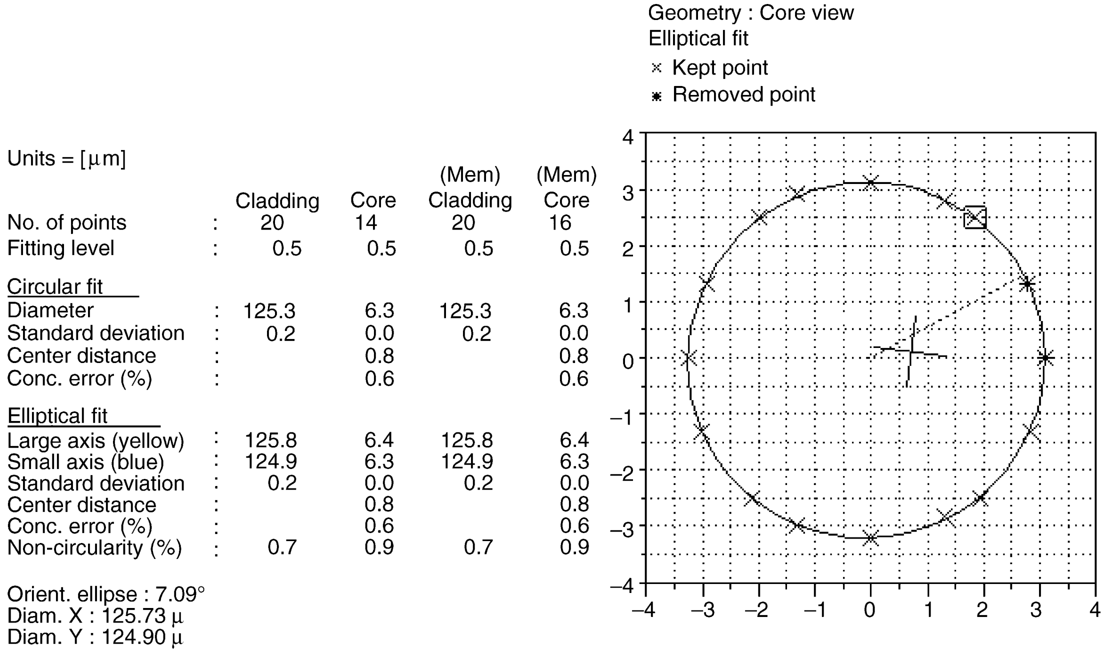

Units = [μm]

		Cladding	Core	(Mem) Cladding	(Mem) Core
No. of points	:	20	14	20	16
Fitting level	:	0.5	0.5	0.5	0.5
Circular fit					
Diameter	:	125.3	6.3	125.3	6.3
Standard deviation	:	0.2	0.0	0.2	0.0
Center distance	:		0.8		0.8
Conc. error (%)	:		0.6		0.6
Elliptical fit					
Large axis (yellow)	:	125.8	6.4	125.8	6.4
Small axis (blue)	:	124.9	6.3	124.9	6.3
Standard deviation	:	0.2	0.0	0.2	0.0
Center distance	:		0.8		0.8
Conc. error (%)	:		0.6		0.6
Non-circularity (%)	:	0.7	0.9	0.7	0.9

Orient. ellipse : 7.09°
Diam. X : 125.73 μ
Diam. Y : 124.90 μ

Figure 21 Core geometry by RNF measurement.

distribution or the RIP at the output of an FUT. NA can be determined from the analysis of the test results.

Mode Field Diameter

The definition of the MFD is given in the section describing the core center above. Four measurement methods are available:

- direct far-field scan determines MFD from the far-field intensity distribution;
- variable aperture in far field determines MFD from the complementary aperture transmission function $a(x)$, $x = d\tan\theta$ being the aperture radius and d the distance between the aperture and the FUT:

$$\mathrm{MFD} = \frac{\lambda}{\pi d}\left[\int_0^\infty a(x)\frac{x}{(x^2+d^2)^2}\,\mathrm{d}x\right]^{-1/2} \qquad [21]$$

Equation [21] is valid for small-θ approximation.

- near-field scan determines MFD from the near-field intensity distribution I_{NF}, r being the radial coordinate:

$$\mathrm{MFD} = 2\left[2\frac{\int_0^\infty rI_{\mathrm{NF}}(r)\,\mathrm{d}r}{\int_0^\infty r\left[\frac{\mathrm{d}I^{1/2}(r)}{\mathrm{d}r}\right]^2\mathrm{d}r}\right]^{1/2} \qquad [22]$$

Equation [22] is valid for small-θ approximation; and

- bidirectional backscattering uses an OTDR and bidirectional measurements to determine MFD by comparing the FUT results with a reference fiber.

Effective Area

A_{eff} is a critical nonlinearity parameter and is defined as follows:

$$A_{\mathrm{eff}} = \frac{2\pi[\int_0^\infty I(r)r\,\mathrm{d}r\,]^2}{\int_0^\infty I(r)^2r\,\mathrm{d}r} \qquad [23]$$

$I(r)$ is the field intensity distribution of the fiber fundamental mode at radius r. The integration in the equation is carried out over the entire fiber cross-section. For a Gaussian approximation:

$$I(r) = \exp - 2\left(\frac{2r}{\mathrm{MFD}}\right)^2 \qquad [24]$$

which yields:

$$A_{\mathrm{eff}} = \frac{\pi}{4}\mathrm{MFD}^2 \qquad [25]$$

Three methods are available for the measurement of A_{eff}:

- direct far-field;
- variable aperture in far-field; and
- near-field.

Mechanical Measurement and Test Methods

Table 2 describes the various mechanical characteristics and their corresponding test methods.

Proof Stressing

The proof stress level is the value of tensile stress or strain applied to a full fiber length over a short period of time. The method for fiber proof stressing is the longitudinal tension which describes procedures for applying tensile loads to a length of fiber. The fiber stress is calculated from the applied tension. The tensile load is applied over short periods of time but not too short in order for the fiber to experience proof stress.

Residual Stress

Residual stress is the stress built up by the thermal expansion difference between core and cladding during the fiber drawing process or splicing. Methods are available for measuring residual stress based on polarization effects. A light beam produced by a rotating polarizer propagates to the x-axis direction while the beam polarization is in the $y-z$ plane and the fiber longitudinal axis in the z-axis. The light experiences different phase shift in the y- and z-axis due to the FUT birefringence. The photoelastic effect gives the relationship between this phase change and residual stresses.

Stress Corrosion Susceptibility

The stress corrosion susceptibility is related to the dependence of crack growth on applied stress. It depends on the environmental conditions and static and dynamic values may be observed.

Environmental Characteristics

Table 3 lists the fiber characteristics related to the effect of the environment.

Table 2 Fiber mechanical characteristics

Attribute
Proof stress
Residual stress
Stress corrosion susceptibility
Tensile strength
Stripability
Fiber curl

Table 3 Fiber environmental characteristics

Attribute
Hydrogen aging
Nuclear gamma irradiation
Damp heat
Dry heat
Temperature cycling
Water immersion

Hydrogen Aging for Low-Water-Peak Single-Mode Fiber

Hydrogen aging on low-water peak fibers, such as G.652.C, is based on a test performed at 1.0 atmosphere of hydrogen pressure at room temperature over a period of one month. Other proportional combinations are possible.

Nuclear Gamma Irradiation

Nuclear radiation is considered on the basis of a steady state response of optical fibers and cables exposed to gamma radiation and to determine the level of related radiation-induced attenuation produced in singlemode or multimode cabled or uncabled fibers.

The fiber attenuation generally increases when exposed to gamma radiation. This is primarily due to the trapping of radiolytic electrons and holes at defect sites in the glass (i.e., the formation of 'color centers'). Two regimes are considered:

- the low dose rate suitable for estimating the effect of environmental background radiation; and
- the high dose rate suitable for estimating the effect of adverse nuclear environments.

The effects of environmental background radiation are measured by the attenuation (cut-back method). The effects of adverse nuclear environments are tested by power monitoring before, during and after FUT exposure.

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Imaging:** Interferometric Imaging. **Interferometry:** Gravity Wave Detection; Overview. **Polarization:** Introduction. **Scattering:** Scattering Phenomena in Optical Fibers.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.
Girard A (2000) *Guide to WDM Technology and Testing.* Quebec: EXFO, pp. 194.
Hecht J (1999) *Understanding Fiber Optics*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
Masson B and Girard A (2004) *FTTx PON Guide, Testing Passive Optical Networks*. Quebec: EXFO, pp. 56.
Miller JL and Friedman E (2003) *Optical Communications Rules of Thumb*. New York: McGraw-Hill, pp. 428.
Neumann E-G (1988) *Single-Mode Fibers, Fundamentals.* Berlin: Springer-Verlag.

Nonlinear Effects (Basics)

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many physical systems in various areas such as condensed matter or plasma physics, biological sciences, or optics, give rise to localized large-amplitude excitations having a relatively long lifetime. Such excitations lead to a host of phenomena referred to as nonlinear phenomena. Of the many disciplines of physics, the optics field is probably the one in which practical applications of nonlinear phenomena have been the most fruitful, in particular, since the discovery of the laser in 1960. This discovery has thus led to the advent of a new branch in optics, referred to as *nonlinear optics*. The applications of nonlinear phenomena in optics include the design of various kinds of laser sources, optical amplifiers, light converters, light-wave communication systems for data transmission purposes, to name a few.

In this article, we present an overview of some basic principles of nonlinear phenomena that result from the interaction of light waves with dielectric waveguides such as optical fibers. These nonlinear phenomena can be broadly divided into two main categories, namely, parametric effects and scattering phenomena. Parametric interactions arise whenever the state of the dielectric matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between energy levels in the medium. More fundamentally, parametric interactions originate from the electron motion under the electric field of a light wave, whereas scattering phenomena originate from the motion of heavy ions (or molecules).

Linear and Nonlinear Signatures

The macroscopic properties of a physical system can be obtained by analyzing the response of the system under an external excitation. For example, consider at time t the response of a system, such as an amplifier, to an input signal $E_1 = A\sin(\omega t)$. In the low-amplitude limit of the output signal, the response R_1 of the system is proportional to the excitation:

$$R_1 = \alpha_1 E_1 \qquad [1]$$

where α_1 is a constant. This type of behavior corresponds to the so-called linear response. In general, a physical system executes a linear response when the superposition of two (or more) input signals E_1 and E_2 yields a response which is a superposition of the output signals, as schematically represented in Figure 1:

$$R = \alpha_1 R_1 + \alpha_2 R_2 \qquad [2]$$

Now, in almost all real physical systems, if the amplitude of an excitation E_1 becomes sufficiently large, distortions will occur in the output signals. In other words, the response of the system will no longer be proportional to the excitation, and consequently, the law of superposition of states will no longer be observed. In this case, the response of the system may take the following form:

$$R = \alpha_1 E_1 + \alpha_2 E_1^2 + \alpha_3 E_1^3 + \dots \qquad [3]$$

which involves not only a signal at the input frequency ω, but also signals at frequencies 2ω, 3ω, and so on. Thus, harmonics of the input signal are generated. This behavior, called *nonlinear response*, is at the origin of a host of important phenomena in many

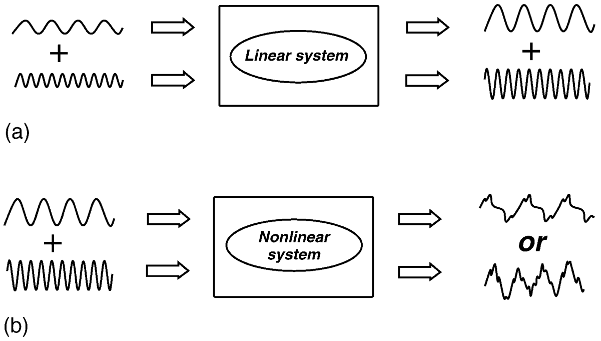

Figure 1 Schematic representation of linear and nonlinear responses of a system, to two input signals.

branches of sciences, such as in condensed matter physics, biological sciences, or in optics.

Physical Origin of Optical Nonlinearity

Optical nonlinearity originates fundamentally from the action of the electric field of a light wave on the charged particles of a dielectric waveguide. In contrast to conductors where charges can move throughout the material, dielectric media consist of *bound* charges (ions, electrons) that can execute only relatively limited displacements around their equilibrium positions. The electric field of an incoming light wave will cause the positive charges to move in the polarization direction of the electric field whereas negative charges will move in the opposite direction. In other words, the electric field will cause the charged particles to become dipoles, as schematically represented in Figure 2.

Then each dipole will vibrate under the influence of the incoming light, thus becoming a source of radiation. The global light radiated by all the dipoles represents the scattered light. When the charge displacements are proportional to the excitation (incoming light), i.e., in the low-amplitude limit of scattered radiation, the output light vibrates at the same frequency as that of the excitation. This process corresponds to Rayleigh scattering. On the other hand, if the intensity of the excitation is sufficiently large to induce displacements that are not negligible with respect to atomic distances, then the charge displacements will no longer be proportional to the excitation. In other words, the response of the medium, which is no longer proportional to the excitation, becomes nonlinear. In this case, the scattered waves will be generated not only at the excitation frequency (the Kerr effect), say ω, but also at frequencies that differ from ω (e.g., 2ω, 3ω). On the other hand, it is worth noting that when all induced dipoles vibrate coherently (that is, their relative phase does not vary randomly), their individual radiation may, under certain conditions, interfere constructively and lead to a global field of high intensity. The condition of constructive interference is the phase-matching condition.

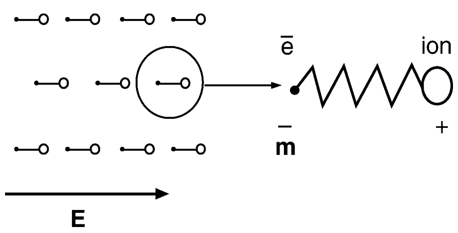

Figure 2 Electric dipoles in a dielectric medium under an external electric field.

In practice, the macroscopic response of a dielectric is given by the polarization, which corresponds to the total amount of dipole moment per unit volume of the dielectric. As the mass of an ion is much larger than that of an electron, the amplitude of the ion motion is generally negligible with respect to that of the electrons. As a consequence, the electron motion generally leads to the dominant contribution in the macroscopic properties of the medium. The behavior of an electron under an optical electric field is similar to that of a particle embedded in an anharmonic potential. A very simple model (called the Lorentz model) that provides a deep insight into the dielectric response consists of an electron of mass m and charge $-e$ connected to an ion by an elastic spring (see Figure 2). Under the electric field $E(t)$, the electron executes a displacement $x(t)$ with respect to its equilibrium position, which is governed by the following equation:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + 2\lambda\frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x + (a^{(2)}x^2 + a^{(3)}x^3 + \cdots) = -\frac{e}{m}E(t) \qquad [4]$$

where $a^{(2)}$, $a^{(3)}$, and so on, are constant parameters, ω_0 is the resonance angular frequency of the electron, and λ is the damping coefficient resulting from the dipolar radiation. When the amplitude of the electric field is sufficiently large, then the restoring force on the electrons becomes a nonlinear function of x; hence the presence of terms such as $a^{(2)}x^2$, $a^{(3)}x^3$, and so on, in eqn [4]. In this situation, the macroscopic response of the dielectric is the polarization

$$P = -\sum ex(\omega, 2\omega, 3\omega, \ldots) \qquad [5]$$

where $x(\omega, 2\omega, 3\omega, \ldots)$ is the solution of eqn [4] in the frequency domain, and the summation extends over all the dipole moments per unit volume. In terms of the electric field E the polarization may be written as

$$P = \varepsilon_0(\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots) \qquad [6]$$

where $\chi^{(1)}$, $\chi^{(2)}$, $\chi^{(3)}$, and so on, represent the susceptibility coefficients. Figure 3 (top left) illustrates schematically the polarization as a function of the electric field.

In particular, one can clearly observe that when the amplitude of the incoming electric field is sufficiently small (bottom left in Figure 3), the polarization (top right) is proportional to the electric field (bottom left), thus implying that the electric dipole radiates a wave having the same frequency as that of the incoming light (bottom right). On the other hand, Figure 4

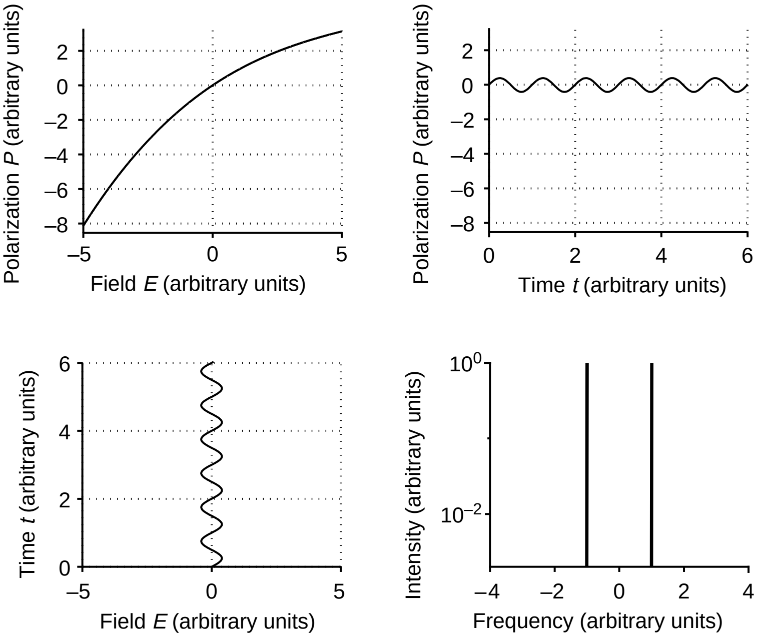

Figure 3 Polarization induced by an electric field of small amplitude. Nonlinear dependence of the polarization as a function of field amplitude (top left) and time dependence of the input electric field (bottom left). Time dependence of the induced polarization (top right) and corresponding intensity spectrum (bottom right).

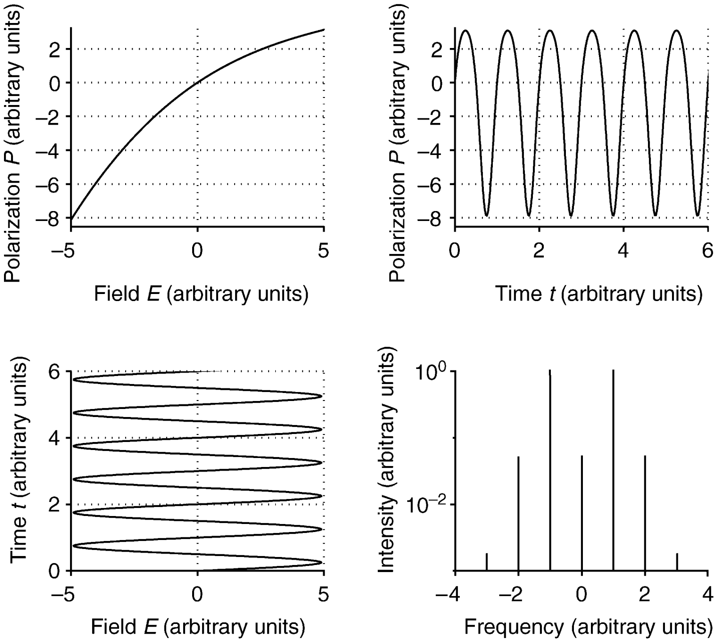

Figure 4 Polarization induced by an incoming electric field of large amplitude.

shows that for an electric field of large amplitude, the polarization is no longer proportional to the electric field, leading to radiation at harmonic frequencies (see Figure 4, bottom right).

This nonlinear behavior leads to a host of important phenomena in optical fibers, which are useful for many optical systems but detrimental for other systems.

Parametric Phenomena in Optical Fibers

In anisotropic materials, the leading nonlinear term in the polarization, i.e., the $\chi^{(2)}$ term, leads to phenomena such as the harmonic generation or optical rectification. This $\chi^{(2)}$ term vanishes in homogeneous isotropic materials such as cylindrical optical fibers, and there the leading nonlinear term becomes the $\chi^{(3)}$ term. Thus, most of outstanding nonlinear phenomena in optical fibers originate from the third-order nonlinear susceptibility $\chi^{(3)}$. Some of those phenomena are described below.

Optical Kerr Effect

The optical Kerr effect is probably the most important nonlinear effect in optical fibers. This effect induces an intensity dependence of the refractive index, which leads to a vast wealth of fascinating phenomena such as self-phase modulation (SPM), cross-phase modulation (CPM), four-wave mixing (FWM), modulational instability (MI) or optical solitons. The Kerr effect can be conveniently described in the frequency domain through a direct analysis of the polarization, which takes the following form:

$$P_{\mathrm{NL}}(\omega) = \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega)|E(\omega)|^2E(\omega) \qquad [7]$$

The constant 3/4 comes from the symmetry properties of the tensor $\chi^{(3)}$. Setting $P_{\mathrm{NL}}(\omega) = \varepsilon_0\varepsilon_{\mathrm{NL}}E(\omega)$, where $\varepsilon_{\mathrm{NL}} = \frac{3}{4}\chi^{(3)}|E|^2$ is the nonlinear contribution to the dielectric constant, the total polarization takes the form

$$P(\omega) = P_{\mathrm{L}} + P_{\mathrm{NL}} = \varepsilon_0[\chi^{(1)}(\omega) + \varepsilon_{\mathrm{NL}}]E(\omega) \qquad [8]$$

As eqn [8] shows, the refractive index n, at a given frequency ω, is given by

$$n^2 = 1 + \chi^{(1)} + \varepsilon_{\mathrm{NL}} = (n_0 + \Delta n_{\mathrm{NL}})^2 \qquad [9]$$

with $n_0^2 = 1 + \chi^{(1)}$. In practice $\Delta n_{\mathrm{NL}} \ll n_0$, and then, the refractive index is given by

$$n(\omega, |E|^2) = n_0(\omega) + n_2^e|E|^2 \qquad [10]$$

where n_2^e is the nonlinear refractive index defined by $n_2^e = 3\chi^{(3)}/(8n_0)$. The linear polarization P_{L} is responsible for the frequency dependence of the refractive index, whereas the nonlinear polarization P_{NL} causes an intensity dependence of the refractive index, which is referred to as the optical Kerr effect. Knowing that the wave intensity I is given by $I = \alpha|E|^2$, with $\alpha = \frac{1}{2}\varepsilon_0 cn_0$, the refractive index can be then rewritten as

$$n(\omega, I) = n_0(\omega) + n_2 I \qquad [11]$$

with $n_2 = n_2^e/\alpha = 2n_2^e/(\varepsilon_0 cn_0)$. For fused silica fibers one has typically: $n_2 = 2.66 \times 10^{-20}\ \mathrm{m^2\ W^{-1}}$. For example, an intensity of $I = 1\ \mathrm{GW\ cm^{-2}}$ leads to $\Delta n_{\mathrm{NL}} = 2.66 \times 10^{-7}$, which is much smaller than $n_0 \approx 1.45$.

Four-Wave Mixing

The four-wave mixing (FWM) process is a third-order nonlinear effect in which four waves interact through an energy exchange process. Let us consider two intense waves, $E_1(\omega_1)$ and $E_2(\omega_2)$, with $\omega_2 > \omega_1$, called pump waves, propagating in an optical fiber. Hereafter we consider the simplest case when waves propagate with the same polarization. In this situation the total electric field is given by

$$\begin{aligned} E_{\mathrm{tot}}(\mathbf{r}, t) &= E_1 + E_2 \\ &= A_1(\omega_1)\exp[\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r} - \omega_1 t)] \\ &\quad + A_2(\omega_2)\exp[\mathrm{i}(\mathbf{k}_2\cdot\mathbf{r} - \omega_2 t)] \end{aligned} \qquad [12]$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are the wavevectors of the fields E_1 and E_2, respectively. Equation [7], which gives the nonlinear polarization induced by a single monochromatic wave, remains valid provided that the frequency spacing between the two waves is relatively small, i.e., $|\Delta\omega| = |\omega_2 - \omega_1| \ll \omega_0 = (\omega_1 + \omega_2)/2$. In this context, eqn [7] leads to

$$P_{\mathrm{NL}} \approx \frac{3}{4}\varepsilon_0\chi^{(3)}(\omega_0)|E_{\mathrm{tot}}|^2E_{\mathrm{tot}} \qquad [13]$$

Substituting eqn [12] in eqn [13] yields

$$\begin{aligned} P_{\mathrm{NL}} = 2n_0n_2\varepsilon_0\Big[&\left(|E_1|^2 + 2|E_2|^2\right)E_1 \\ &+ \left(|E_2|^2 + 2|E_1|^2\right)E_2 + E_1^2E_2^* + E_1^*E_2^2\Big] \end{aligned} \qquad [14]$$

The term $|E_1|^2E_1$ in the right-hand side of eqn [14] represents the self-induced Kerr effect on the wave ω_1. The term $|E_2|^2E_1$, which corresponds to a modification of the refractive index seen by the wave ω_1, due to the presence of the wave ω_2, represents the cross-Kerr effect. Thus, the refractive index at frequency ω_1 depends simultaneously on the intensities of the two pumps, i.e., $n = n(\omega_1, |E_1|^2, |E_2|^2)$. Similarly the third and fourth terms in the right-hand side of eqn [14] illustrate the self-induced and cross-Kerr effects on the wave ω_2, respectively. Note that the self-induced Kerr effect is responsible for a self-phase modulation, which induces a spectral broadening of a pulse propagating through the optical fiber, whereas the cross-Kerr effect leads to a cross-phase modulation that induces a spectral broadening of one wave in the presence of a second wave. The two last terms in the right-hand side of eqn [14] correspond to the generation of new waves at frequencies $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, respectively. The wave with the smallest frequency $2\omega_1 - \omega_2 = \omega_{\mathrm{s}}$ is called the Stokes wave, whereas the wave with the highest frequency $2\omega_2 - \omega_1 = \omega_{\mathrm{as}}$ is called the anti-Stokes wave. These two newly generated waves interact

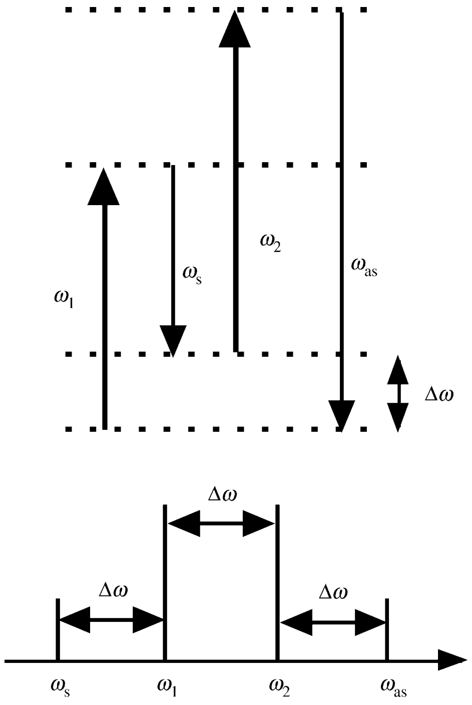

Figure 5 Schematic diagram representing a non-degenerate four-wave mixing process (top) and the corresponding frequency spectrum (bottom).

with the two pumps through an energy exchange process. This interaction is commonly referred to as a four-wave mixing process.

From a quantum mechanical point of view, four-wave mixing corresponds to a process in which two photons with frequencies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_s and ω_{as}, respectively. The new waves are generated at frequencies ω_s and ω_{as} such that

$$\omega_1 + \omega_2 = \omega_s + \omega_{as} \qquad [15]$$

which states that the total energy is conserved during the interaction. But the condition for this FWM process to occur is that the total momentum is conserved, that is,

$$\Delta \mathbf{k} = \mathbf{k}_s + \mathbf{k}_{as} - \mathbf{k}_1 - \mathbf{k}_2 = 0 \qquad [16]$$

or equivalently,

$$n(\omega_s)\omega_s + n(\omega_{as})\omega_{as} - n(\omega_1)\omega_1 - n(\omega_2)\omega_2 = 0 \quad [17]$$

Equation [16] is known as the phase-matching condition. Figure 5 displays a schematic diagram of a four-wave mixing process with the corresponding frequency spectrum.

FWM with different pump frequencies $\omega_1 \neq \omega_2$ is called 'nondegenerate FWM', whereas the case $\omega_1 = \omega_2$ is referred to as 'degenerate FWM', or more simply, as three-wave mixing.

Conclusion

Optical fibers constitute a key device for many areas of optical sciences, and in particular for ultrafast optical communications. The nonlinear phenomena that arise through the nonlinear refractive index change (induced mainly by the Kerr effect) have been largely investigated these last two decades for applications such as all-optical wavelength conversion, parametric amplification, generation of new optical frequencies or ultrahigh repetition rate pulse trains. However, in many other optical systems such as optical communication systems, these nonlinear phenomena become harmful effects, and in this case control processes are being developed in order to suppress or at least reduce their impact in the system.

List of Units and Nomenclature

Nonlinear refractive index n_2 [$\mathrm{m^2\,W^{-1}}$]
Optical intensity I [$\mathrm{GW\,cm^{-2} = 10^9\,W\,cm^{-2}}$]

CPM: cross phase modulation
FWM: four wave mixing
MI: modulational instability
SPM: self-phase modulation

See also

Fiber and Guided Wave Optics: Light Propagation; Measuring Fiber Characteristics. **Lasers:** Optical Fiber lasers. **Optical Amplifiers:** Semiconductor Optical Amplifiers. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Bloembergen N (1977) *Nonlinear Optics*. Reading, MA: Benjamin.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Butcher PN and Cotter DN (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) *Nonlinear Effects in Optical Fibres*. Switzerland: Harwood Academic Publishers.

Hellwarth R (1977) Third-order optical susceptibilities of liquids and s\olids. *Progress in Quantum Electronics* 5: 1–68.

Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Nonlinear Optics

K Thyagarajan and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the laser provided us with a light source capable of generating extremely large optical power densities (several MW/m^2). At such large power densities, matter behaves in a nonlinear fashion and we come across new optical phenomena such as second-harmonic generation (SHG), sum and difference frequency generation, intensity dependent refractive index, mixing of various frequencies, etc. In SHG, an incident light beam at frequency ω interacts with the medium and generates a new light wave at frequency 2ω. In sum and difference frequency generation, two incident beams at frequencies ω_1 and ω_2 mix with each other producing sum $(\omega_1 + \omega_2)$ and difference $(\omega_1 - \omega_2)$ frequencies at the output. Higher-order nonlinear effects, such as self-phase modulation, four-wave mixing, etc. can also be routinely observed today. The field of nonlinear optics dealing with such nonlinear interactions is gaining importance due to numerous demonstrated applications in many diverse areas such as optical fiber communications, all-optical signal processing, realization of novel sources of optical radiation, etc.

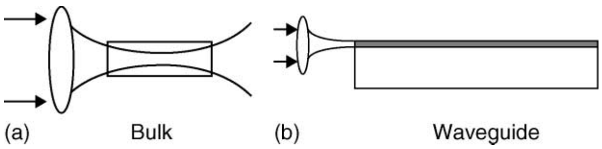

Figure 1 (a) In bulk media, tighter focusing produces larger intensities, but over shorter interaction lengths. (b) In optical waveguides diffraction effects are balanced by waveguiding and the interaction lengths are much larger.

Nonlinear optical interactions become prominent when the optical power densities are high and interaction takes place over long lengths. The usual method to increase optical intensity is to focus the light beam using a lens system. For a given optical power, the tighter the focusing, the larger will be the intensity for a given optical power; however, greater will be the divergence of the beam. Thus tighter focusing produces larger intensities, but over shorter interaction lengths (see Figure 1a). Optical waveguides, in which the light beam is confined to a small cross-sectional area, are currently being explored for realizing efficient nonlinear devices. In contrast to bulk media, in waveguides, diffraction effects are balanced by waveguiding and the beam can have small cross-sectional areas over much longer interaction lengths (see Figure 1b). A simple optical waveguide consists of a high index dielectric medium surrounded by a lower index dielectric medium so that light waves can be trapped in the high index region by the phenomenon of total internal reflection. Figure 2 shows a planar waveguide, a channel waveguide and an optical fiber. In the planar waveguide a film of refractive index n_f is deposited/diffused on a substrate of refractive index n_s and has a cover of refractive index n_c (with n_s, $n_c < n_f$). The waveguide has typical cross-sectional dimensions of a few micrometers. Unlike planar waveguides, in which guidance takes place only in one dimension, in channel waveguides the diffused region in a substrate is surrounded on all sides by lower index media. Optical fibers are structures with cylindrical symmetry and have a central cylindrical core of doped silica surrounded by a concentric cylindrical cladding of pure silica which has a slightly lower refractive index. Light guidance in all these waveguides takes place through the phenomenon of total internal reflection.

In contrast to bulk interactions requiring beam focusing, in the case of optical waveguides, the beam can be made to have extremely small cross-sectional

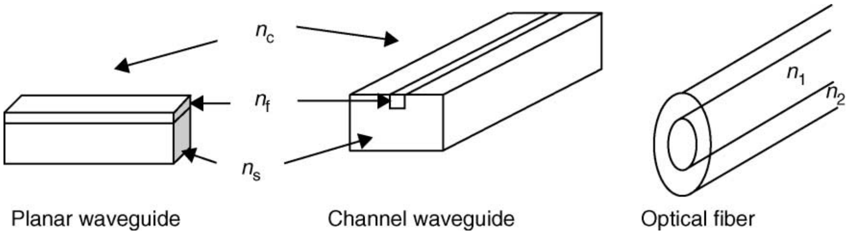

Figure 2 A planar waveguide, a channel waveguide and an optical fiber.

areas (~25 μm^2) over very long interaction lengths (~20 mm in integrated optical waveguides and tens of thousands of kilometers in the case of optical fibers). This leads to very much increased nonlinear interaction efficiencies even at moderate powers (approximately a few tens of mW).

In the following sections, we will discuss some of the important nonlinear interactions that are being studied with potential applications to various branches of science and engineering. Obviously it is impossible to cover all aspects of nonlinear optics – several books have been written in this area (see Further Reading section at the end of this article) – what we will do is to discuss the physics of some of the important nonlinear effects.

Apart from intensity and length of interaction, one of the most important requirements for many efficient nonlinear interactions is the requirement of phase matching. The concept of phase matching can be easily understood from the point of view of SHG. In SHG, the incident wave at frequency ω generates a nonlinear polarization at frequency 2ω and this nonlinear polarization is responsible for the generation of the wave at 2ω. Now, the phase velocity of the nonlinear polarization wave at 2ω is the same as the phase velocity of the electromagnetic wave at frequency ω, which is usually different from the phase velocity of the electromagnetic wave at frequency 2ω; this happens due to wavelength dispersion in the medium. When the two phase velocities are unequal, the polarization wave at 2ω (which is the source) and the electromagnetic wave at 2ω pass in and out of phase with each other as they propagate through the medium. Due to this, the energy flowing in from ω to 2ω cannot add constructively and the efficiency of second-harmonic generation is limited. If the phase velocities of the waves at ω and 2ω are matched then the polarization wave and the wave at 2ω remain in phase leading to drastically increased efficiencies. This condition is referred to as phase matching and plays a very important role in most nonlinear interactions.

Nonlinear Polarization

In a linear medium, the electric polarization P is assumed to be a linear function of the electric field E:

$$P = \varepsilon_0 \chi E \tag{1}$$

where, for simplicity, a scalar relation has been written. The quantity χ is termed as linear dielectric susceptibility. At high optical intensities (which corresponds to high electric fields), all media behave in a nonlinear fashion. Thus eqn [1] is modified to

$$P = \varepsilon_0(\chi E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ldots) \tag{2}$$

where $\chi^{(2)}, \chi^{(3)}, \ldots$ are higher order susceptibilities giving rise to the nonlinear terms. The second term on the right-hand side is responsible for SHG, sum and difference frequency generation, parametric interactions, etc. while the third term is responsible for third-harmonic generation, intensity dependent refractive index, self-phase modulation, four-wave mixing, etc. For media possessing an inversion symmetry, $\chi^{(2)}$ is zero and there is no second-order nonlinear effect. Thus silica optical fibers, which form the heart of today's communication networks, do not possess the second-order nonlinearity.

We will first discuss second-harmonic generation and parametric amplification which arises due to the second-order nonlinear term and then go on to effects due to third-order nonlinear interaction.

Second-Harmonic Generation in Crystals

The first demonstration of SHG was made in 1961 by focusing a 3 kW ruby laser pulse ($\lambda_0 = 6943$ Å) on a quartz crystal. An incident beam from a ruby laser (red color) after passing through a crystal of KDP, gets partly converted into blue light which is the second harmonic. Ever since, SHG has been one of the most widely studied nonlinear interactions.

We consider a plane wave of frequency ω propagating along the z-direction through a medium and consider the generation of the second-harmonic frequency 2ω as the beam propagates through the medium. Now, the field at ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω. Corresponding to the frequencies ω and 2ω, the electric fields are assumed to be given by

$$E^{(\omega)} = \frac{1}{2}(E_1(z)e^{i(\omega t - k_1 z)} + \text{c.c.}) \tag{3}$$

and

$$E^{(2\omega)} = \frac{1}{2}(E_2(z)e^{i(2\omega t - k_2 z)} + \text{c.c.}) \tag{4}$$

respectively; c.c. stands for the complex conjugate of the preceding quantities. The quantities:

$$k_1 = \omega\sqrt{(\varepsilon_1 \mu_0)} = (\omega/c)n^{\omega} \tag{5}$$

and

$$k_2 = 2\omega\sqrt{(\varepsilon_2 \mu_0)} = (2\omega/c)n^{2\omega} \tag{6}$$

represent the propagation vectors at ω and 2ω, respectively; ε_1 and ε_2 represent the dielectric permittivities at ω and 2ω, and n^{ω} and $n^{2\omega}$ represent the corresponding wave refractive indices. It should be noted that the amplitudes E_1 and E_2 are assumed to be z dependent – this is because at $z = 0$ (where the beam is incident on the medium) the amplitude E_2 is zero and this would develop as the beam propagates through the medium. We will now develop an approximate theory for the generation of the second harmonic.

We start with the wave equation:

$$\nabla^2 E - \varepsilon\mu_0 \frac{\partial^2 E}{\partial t^2} = \mu_0 \frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \qquad [7]$$

where P_{NL} is the nonlinear polarization. An incident wave at frequency ω generates a polarization at 2ω, which acts as a source for the generation of an electromagnetic wave at 2ω.

In order to consider SHG, we write the wave equation corresponding to 2ω with the nonlinear polarization at 2ω given by

$$P_{\mathrm{NL}}^{(2\omega)} = \frac{1}{2}\left(\tilde{P}_{\mathrm{NL}}^{(2\omega)}\, \mathrm{e}^{2\mathrm{i}(\omega t - k_1 z)} + \mathrm{c.c.}\right) \qquad [8]$$

where

$$\tilde{P}_{\mathrm{NL}}^{(2\omega)} = dE_1E_1 \qquad [9]$$

and

$$d = \frac{\varepsilon_0 \chi^{(2)}}{2} \qquad [10]$$

represents the effective nonlinear coefficient and depends on the nonlinear material, the polarization states of the fundamental and the second harmonic and also on the propagation direction. Simple manipulations give us, for the rate of change of amplitude of the second harmonic wave:

$$\frac{\mathrm{d}E_2}{\mathrm{d}z} = -\frac{i\mu_0\, dc\omega}{n^{2\omega}} E_1^2(z)\mathrm{e}^{\mathrm{i}(\Delta k)z} \qquad [11]$$

where

$$\Delta k = k_2 - 2k_1 = (2\omega/c)(n^{2\omega} - n^{\omega}) \qquad [12]$$

and we have assumed:

$$\mathrm{d}^2E_2/\mathrm{d}z^2 \ll k_2(\mathrm{d}E_2/\mathrm{d}z) \qquad [13]$$

In order to solve eqn [11] we neglect depletion of the fundamental field, i.e., $E_1(z)$ is almost a constant and the quantity E_1^2 on the right-hand side can be assumed to be independent of z. If we now integrate eqn [11], we obtain:

$$E_2(z) = -\frac{i\mu_0\, dc\omega}{n^{2\omega}} zE_1^2\mathrm{e}^{\mathrm{i}\sigma}\frac{\sin\sigma}{\sigma} \qquad [14]$$

where

$$\sigma = \frac{1}{2}(\Delta k)z = (\omega/c)(n^{2\omega} - n^{\omega})z \qquad [15]$$

Now, the powers associated with the beams corresponding to ω and 2ω are given by:

$$P_1 = \frac{n^{\omega}}{2c\mu_0}S|E_1|^2, \qquad P_2 = \frac{n^{2\omega}}{2c\mu_0}S|E_2|^2 \qquad [16]$$

where S represents the cross-sectional area of the beams. Substituting for $|E_2|^2$ from eqn [16], we get after some elementary simplifications:

$$\eta = \frac{P_2}{P_1} = \frac{2c^3\mu_0^3 d^2\omega^2}{(n^{\omega})^2 n^{2\omega}} z^2 \frac{P_1}{S}\left(\frac{\sin\sigma}{\sigma}\right)^2 \qquad [17]$$

when η represents the SHG efficiency. Note that the efficiency of SHG increases if P_1/S, the intensity of the fundamental wave increases. Also η increases if the nonlinear coefficient d increases (obviously) and if the frequency increases. However, for a given power P_1, the conversion efficiency increases if the area of the beam decreases – thus a focused beam will have a greater SHG efficiency. The most important factor is $(\sin\sigma/\sigma)^2$ which is a sharply peaked function around $\sigma = 0$, attaining a maximum value of unity at $\sigma = 0$. Thus for maximum SHG efficiency:

$$\sigma = 0 \Rightarrow n^{2\omega} = n^{\omega} \qquad [18]$$

i.e., the refractive index at 2ω must be equal to the refractive index at ω – this is known as the phase matching condition.

The phase matching condition can be pictorially represented by a vector diagram (see Figure 3). In Figure 3, $\mathbf{k}_1\omega$ and $\mathbf{k}_2\omega$ represent the wave vectors of the fundamental and the second harmonic respectively. To achieve reasonably effective SHG, it is very important to satisfy the phase-matching condition. Efficiencies under nonphase matched operation can be orders of magnitude lower than under phase-matched operation.

We see from eqn [17] that the smallest z for which η is maximum is:

$$z = L_{\mathrm{c}} = \frac{\pi}{\Delta k} = \frac{\pi c}{2\omega(n^{2\omega} - n^{\omega})} \qquad [19]$$

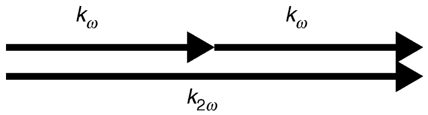

Figure 3 Vector diagram representing the phase matching condition for SHG.

where we have used eqn [15] for Δk. The length L_c is called the phase coherence length and represents the maximum crystal length up to which the second-harmonic power increases. Thus, if the length of the crystal is less than L_c, the second-harmonic power increases almost quadratically with z. For $z > L_c$, the second-harmonic power begins to reduce again.

In general, because of dispersion, it is very difficult to satisfy the phase-matching condition. However, in a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω (see Figure 4). For media with $n_e > n_o$, the direction would correspond to that along which the refractive index of the *e*-wave for ω equals the refractive index for the *o*-wave for 2ω. This can readily be understood by considering a specific example. We consider the SHG in KDP corresponding to the incident ruby laser wavelength ($\lambda_0 = 0.6943\ \mu\text{m}$, $\omega = 2.7150 \times 10^{15}$ Hz). For this frequency:

$$\left\{ \begin{array}{ll} n_o^{\omega} = 1.50502, & n_e^{\omega} = 1.46532 \\ n_o^{2\omega} = 1.53269, & n_e^{2\omega} = 1.48711 \end{array} \right\} \qquad [20]$$

The refractive index variation for the *e*-wave is given by:

$$n_e(\theta) = \left(\frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \right)^{-\frac{1}{2}} \qquad [21]$$

where θ is the angle that the wave propagation direction makes with the optic axis. Now, as can be seen from the above equations:

$$n_e < n_e(\theta) < n_o \qquad [22]$$

Since n_o at 0.6943 μm lies between n_e and n_o at 0.34715 μm, there will always exist an angle θ_m along which $n_e^{2\omega}(\theta_m) = n_o^{\omega}$. We can solve for θ_m and obtain:

$$\cos\theta_m = \left[\frac{(n_o^{\omega})^2 - (n_e^{2\omega})^2}{(n_o^{2\omega})^2 - (n_e^{2\omega})^2} \right]^{1/2} \frac{n_o^{2\omega}}{n_o^{\omega}} \qquad [23]$$

For the values given by eqn [20], we find $\theta_m = 50.5°$.

Quasi Phase Matching (QPM)

As mentioned earlier, phase matching is extremely important for any efficient nonlinear interaction. Recently the technique of quasi phase matching (QPM) has become a convenient way to achieve phase matching at any desired wavelength in any material. QPM relies on the fact that the phase mismatch in the two interacting beams can be compensated by periodically readjusting the phase of interaction through periodically modulating the nonlinear characteristics of the medium at a spatial frequency equal to the wavevector mismatch of the two interacting waves. Thus in SHG, when the nonlinear polarization at 2ω and the electromagnetic wave at 2ω have an accumulated phase difference of π, then the sign of the nonlinear coefficient is reversed so that the energy flowing from the polarization to the wave can add constructively with the existing energy (see Figure 5). Thus, by properly choosing the period of the spatial modulation of the nonlinear coefficient, one could achieve phase matching. This scheme, QPM, is being very widely studied for application to nonlinear interactions in bulk and in waveguides.

In a ferroelectric material such as lithium niobate, the signs of the nonlinear coefficients are linked to the direction of the spontaneous polarization. Thus a

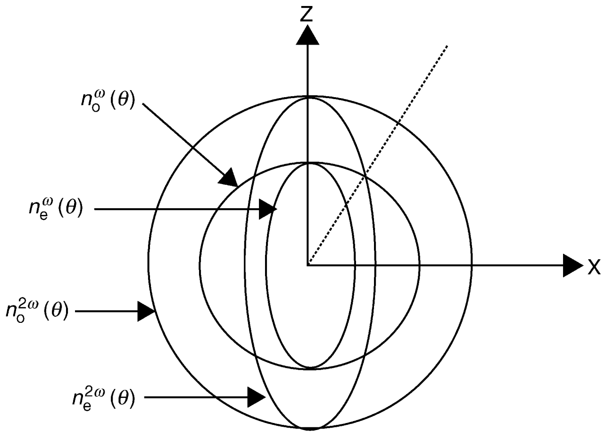

Figure 4 In a birefringent medium, for example with $n_o > n_e$, it may be possible to find a direction along which the refractive index of the *o*-wave for ω equals the refractive index of the *e*-wave for 2ω.

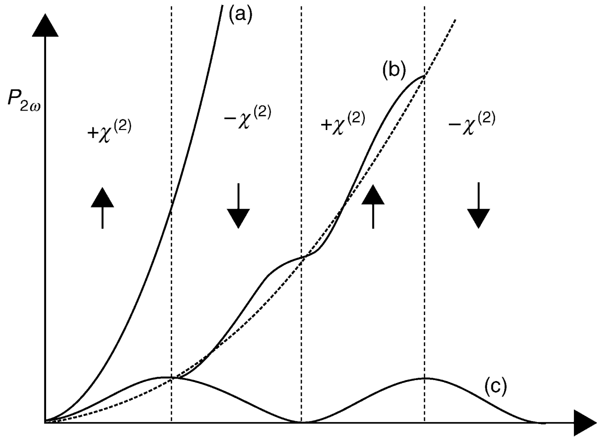

Figure 5 By reversing the sign of the nonlinear coefficient after every coherence length, the energy in the second harmonic can be made to grow. (a) Perfect phase matching; (b) Quasi phase matching; (c) Nonphase matched.

periodic reversal of the domains of the crystal can be used for achieving QPM (see Figure 6). This is the currently used technique to obtain high-efficiency SHG and other nonlinear interactions in $LiNbO_3$, $LiTaO_3$, and KTP. The most popular technique today, to achieve periodic domain reversal in $LiNbO_3$, is the technique of electric field poling. In this method, a high electric field pulse is applied to properly oriented lithium niobate crystal using lithographically defined electrode patterns to produce a permanent periodic domain reversed pattern. Such a periodically domain reversed $LiNbO_3$ crystal with the periodically reversed domains going through the entire depth of the crystal is also referred to as PPLN (periodically poled lithium niobate, pronounced 'piplin') and is now commercially available.

In order to analyze SHG in a periodically poled material, let us assume that the nonlinear coefficient d varies sinusoidally with a period Λ. In such a case we have:

$$d = d_0 \sin(Kz) \qquad [24]$$

where d_0 is the amplitude of modulation of the nonlinear coefficient and $K(= 2\pi/\Lambda)$ represents the spatial frequency of the periodic modulation. For easier understanding we are assuming the modulation to be sinusoidal; in general, the modulation will be periodic but not sinusoidal. Any periodic modulation can be written as a superposition of sinusoidal and cosinusoidal variations. Thus our discussion is valid for one of the Fourier components of the variation.

By using eqn [24], eqn [11] for the variation of the amplitude of the second harmonic becomes:

$$\frac{dE_2}{dz} = -\frac{i\mu_0 d_0 c\omega}{n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} \sin(Kz) \qquad [25]$$

which can be written as:

$$\frac{dE_2}{dz} = -\frac{\mu_0 d_0 c\omega}{2n^{2\omega}} E_1^2(z) e^{i(\Delta k)z} [e^{iKz} - e^{-iKz}] \qquad [26]$$

Using similar arguments as earlier, it can be shown that if $\Delta k - K = 0$, then only the second term within the brackets in eqn [26] contributes to the growth of the second harmonic and similarly if $\Delta k + K = 0$, then only the first term within the brackets contributes to the growth of the second harmonic. The first condition implies that:

$$2\frac{2\pi}{\lambda_0}(n^{2\omega} - n^{\omega}) - \frac{2\pi}{\Lambda} = 0 \qquad [27]$$

where $\Lambda = 2\pi/K$ represents the spatial period of the modulation of the nonlinear coefficient, and λ_0 is the wavelength of the fundamental. Thus the modulation period Λ required for QPM SHG is:

$$\Lambda = \frac{\lambda_0}{2(n^{2\omega} - n^{\omega})} = 2\,L_c \qquad [28]$$

Figure 7 shows the vector diagram corresponding to QPM SHG. The phase mismatch between the fundamental and second harmonic is compensated by the spatial frequency vector of the periodic variation of d.

In the case of waveguides, n^{ω} and $n^{2\omega}$ would represent the effective indices of the modes at the fundamental and the second-harmonic frequency. It may be noted that the index difference between the fundamental and the second harmonic is typically 0.1 and thus the required spatial periods for a fundamental wavelength of 800 nm is ~4 μm.

The main advantage of QPM is that it can be used at any wavelength within the transparency window of the material; only the period needs to be correctly chosen for a specific fundamental wavelength. One can also choose appropriate polarization to make use of the largest nonlinear optical coefficients. Another advantage is the possibility of varying the domain reversal period (chirp) to achieve specific interaction characteristics such as increased bandwidth, etc.

In general, the spatial variation of the nonlinear grating is not sinusoidal. In this case the efficiency of interaction would be determined by the Fourier component of the spatial variation at the spatial frequency corresponding to the period given by eqn [28]. Also, since the required periods are very small, it is possible to use a higher spatial period of modulation and use one of the Fourier components for the nonlinear interaction process. Thus, in the case of periodic reversal of the nonlinear coefficient

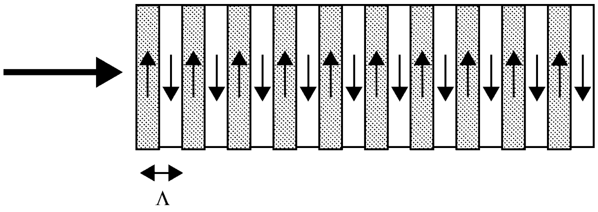

Figure 6 QPM can be achieved by a periodic reversal of the domains of a ferroelectric material. Arrows represent the direction of the spontaneous polarization of the crystal.

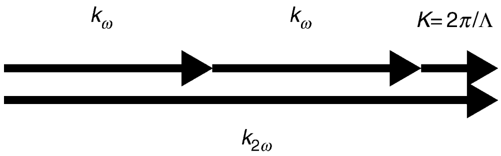

Figure 7 Vector diagram corresponding to QPM-SHG.

with a spatial period given by:

$$\Lambda_g = m\frac{\lambda_0}{2(n^{2\omega} - n^{\omega})}; \qquad m = 1, 3, 5, \ldots \qquad [29]$$

which happens to be the m^{th} harmonic of the fundamental spatial frequency required for QPM, the corresponding nonlinear coefficient that would be responsible for SHG would be the Fourier amplitude at that spatial frequency. This can be taken into account by defining an effective nonlinear coefficient:

$$d_{\text{QPM}} = \frac{2d_0}{m\pi} \qquad [30]$$

Of course the largest effective nonlinear coefficient is achieved by using the fundamental frequency with $m = 1$. Higher spatial periods are easier to fabricate but would lead to reduced nonlinear efficiencies.

As compared to bulk devices, in the case of waveguides, the interacting waves are propagating in the form of modes having specific intensity distributions in the transverse plane. Because of this, the nonlinear interaction also depends on the overlap between the periodically inverted nonlinear medium, the fields of the fundamental and second-harmonic waves. Thus if $E_\omega(x, y)$ and $E_{2\omega}(x, y)$ represent the electric field distributions of the waveguide modes at the fundamental and second-harmonic frequency, then the efficiency of SHG depends on the following overlap integral:

$$I_{\text{ovl}} = \int\int d_0(x, y)E_\omega^2(x, y)E_{2\omega}(x, y)\,\mathrm{d}x\,\mathrm{d}y \qquad [31]$$

where $d_0(x, y)$ represents the transverse variation of the nonlinear coefficient of the waveguide. Thus optimization of SHG efficiency in the case of waveguide interactions has to take account of the overlap integral.

Since QPM relies on periodic phase matching, it is highly wavelength dependent. Thus the required period Λ is different for different fundamental frequencies. Any deviation in the frequency of the fundamental would lead to a reduction in the efficiency due to deviation from QPM condition. Thus the pump laser needs to be highly stabilized at a frequency corresponding to the fabricated period.

Apart from SHG, QPM has been used for other three-wave interaction processes such as difference frequency generation, parametric amplification, etc. Among the many materials that have been studied for SHG using QPM, the important ones are lithium niobate, lithium tantalite, and potassium titanyl phosphate (KTP). Many techniques have been developed for periodic domain reversal to achieve a periodic variation in the nonlinear coefficient. This includes electric field poling, electron bombardment, thermal diffusion treatment, etc.

Third-Order Nonlinear Effects

In the earlier sections, we discussed effects arising out of second-order nonlinearity, i.e., the term proportional to E^2 in the nonlinear polarization. This nonlinear term is found only in media not possessing an inversion symmetry. Thus amorphous materials or crystals possessing inversion symmetry do not exhibit second-order effects. The lowest-order nonlinear effects in such a medium is of an order three wherein the nonlinear polarization is proportional to E^3.

Self-phase modulation (SPM), cross-phase modulation (XPM) and four-wave mixing (FWM) represent some of the very important consequences of third-order nonlinearity. These effects have become all the more important as they play a significant and important role in wavelength division multiplexed optical fiber communication systems.

Self-Phase Modulation (SPM)

Consider the propagation of a plane light wave at frequency ω through a medium having $\chi^{(3)}$ nonlinearity. The polarization generated in the medium is given by

$$P = \varepsilon_0\chi E + \varepsilon_0\chi^{(3)}E^3 \qquad [32]$$

If we consider a single frequency wave with an electric field given by

$$E = E_0\cos(\omega t - kz) \qquad [33]$$

then

$$P = \varepsilon_0\chi E_0\cos(\omega t - kz) + \varepsilon_0\chi^{(3)}E_0^3\cos^3(\omega t - kz) \qquad [34]$$

Expanding $\cos^3\theta$ in terms of $\cos\theta$ and $\cos 3\theta$, we obtain the following expression for the polarization at frequency ω:

$$P = \varepsilon_0\left(\chi + \frac{3}{4}\chi^{(3)}E_0^2\right)E_0\cos(\omega t - kz) \qquad [35]$$

For a plane wave given by eqn [33], the intensity is

$$I = \frac{1}{2}c\varepsilon_0 n_0 E_0^2 \qquad [36]$$

where n_0 is the refractive index of the medium at low intensity. Then

$$P = \varepsilon_0\left(\chi + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0}I\right)E \qquad [37]$$

The polarization P and electric field are related through the following equation:

$$P = \varepsilon_0(n^2 - 1)E \qquad [38]$$

where n is the refractive index of the medium. Comparing eqns [37] and [38], we get

$$n^2 = n_0^2 + \frac{3}{2}\frac{\chi^{(3)}}{c\varepsilon_0 n_0} I \qquad [39]$$

where

$$n_0^2 = 1 + \chi \qquad [40]$$

Since the last term in the eqn [39] is usually very small, we get

$$n = n_0 + n_2 I \qquad [41]$$

where

$$n_2 = \frac{3}{4}\frac{\chi^{(3)}}{c\varepsilon_0 n_0^2} \qquad [42]$$

is the nonlinear coefficient.

For fused silica $n_0 \approx 1.47$, $n_2 \approx 3.2 \times 10^{-20}$ m^2/W and if we consider power of 100 mW having a cross-sectional area of 100 μm^2, the resultant intensity is 10^9 W/m^2 and the corresponding change in refractive index is

$$\Delta n \approx n_2 I \approx 3.2 \times 10^{-11} \qquad [43]$$

Although this is very small, when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant.

In the case of an optical fiber, the light beam propagates as a mode having a specific transverse electric field distribution and thus the intensity is not constant across the cross-section. In such a case, it is convenient to express the nonlinear effect in terms of the power carried by the mode (rather than in terms of intensity). If the linear propagation constant of the mode is represented by β, then in the presence of nonlinearity, the effective propagation constant is given by

$$\beta_{\mathrm{NL}} = \beta + \frac{k_0 n_2}{A_{\mathrm{eff}}} P \qquad [44]$$

where $k_0 = 2\pi/\lambda_0$, P is the power carried by the mode. The quantity A_{eff} represents the effective transverse cross-sectional area of the mode and is defined by

$$A_{\mathrm{eff}} = \frac{\left[\int_0^\infty \int_0^{2\pi} \psi^2(r) r \, \mathrm{d}r \, \mathrm{d}\phi\right]^2}{\int_0^\infty \int_0^{2\pi} \psi^4(r) r \, \mathrm{d}r \, \mathrm{d}\phi} \qquad [45]$$

where $\psi(r)$ represents the transverse mode field distribution of the mode. For example, under the Gaussian approximation:

$$\psi(r) = \psi_0 \, \mathrm{e}^{-r^2/w_0^2} \qquad [46]$$

where ψ_0 is a constant and $2w_0$ represents the mode field diameter (MFD), we get:

$$A_{\mathrm{eff}} = \pi w_0^2 \qquad [47]$$

It is usual to express the nonlinear characteristic of an optical fiber by the coefficient given by

$$\gamma = \frac{k_0 n_2}{A_{\mathrm{eff}}} \qquad [48]$$

Thus, for the same input power and same wavelength, smaller values of A_{eff} lead to greater nonlinear effects in the fiber. Typically:

$$A_{\mathrm{eff}} \sim 50\text{–}80 \ \mu\mathrm{m}^2 \quad \text{and} \quad \gamma \approx 2.4 \ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \ \text{to} \ 1.5 \ \mathrm{W}^{-1}\,\mathrm{km}^{-1} \qquad [49]$$

When a light beam propagates through an optical fiber, the power decreases because of attenuation. Thus, the corresponding nonlinear effects also reduce. Indeed, the phase change suffered by a beam in propagating from 0 to L is given by

$$\phi = \int_0^L \beta_{\mathrm{NL}} \, \mathrm{d}z = \beta L + \gamma \int_0^L P \, \mathrm{d}z \qquad [50]$$

If α represents the attenuation coefficient, then

$$P(z) = P_0 \, \mathrm{e}^{-\alpha z} \qquad [51]$$

and we get

$$\phi = \beta L + \gamma P_0 L_{\mathrm{eff}} \qquad [52]$$

where

$$L_{\mathrm{eff}} = \frac{1 - \mathrm{e}^{-\alpha L}}{\alpha} \qquad [53]$$

is referred to as the effective length. For $\alpha L \gg 1$, $L_{\mathrm{eff}} \approx 1/\alpha$ and for $\alpha L \ll 1$, $L_{\mathrm{eff}} \approx L$.

The effective length represents the length of the fiber over which most of the nonlinear effects has accumulated. For a loss coefficient of 0.20 dB/km, $L_{\mathrm{eff}} \approx 21$ km.

If we consider a fiber length much longer than L_{eff}, then to have reduced impact of SPM, we must have

$$\gamma P_0 L_{\mathrm{eff}} \ll 1 \qquad [54]$$

or

$$P_0 \ll \frac{1}{\gamma L_{\mathrm{eff}}} \approx \frac{\alpha}{\gamma} \qquad [55]$$

For $\alpha = 4.6 \times 10^{-2}$ km^{-1} (which corresponds to an attenuation of 0.2 dB/km) and $\gamma = 2.4$ W^{-1} km^{-1}, we get

$$P_0 \ll 19 \ \mathrm{mW} \qquad [56]$$

Propagation of a Pulse

When an optical pulse propagates through a medium, it suffers from the following effects:

(i) attenuation;
(ii) dispersion; and
(iii) nonlinearity.

Attenuation refers to the reduction in the pulse energy due to various mechanisms, such as scattering, absorption, etc. Dispersion is caused by the fact that a light pulse consists of various frequency components and each frequency component travels at a different group velocity. Dispersion causes the temporal width of the pulse to change; in most cases it results in an increase in pulse width, however, in some cases the temporal width could also decrease. Dispersion is accompanied by chirping, the variation of the instantaneous frequency of the pulse within the pulse duration. Since both attenuation and dispersion cause a change in the temporal variation of the optical power, they closely interact with nonlinearity in deciding the pulse evolution as it propagates through the medium.

Let $E(x, y, z, t)$ represent the electric field variation of an optical pulse. It is usual to express E in the following way:

$$E(x, y, z, t) = \frac{1}{2}[A(z, t)\psi(x, y)\mathrm{e}^{\mathrm{i}(\omega_0 t - \beta_0 z)} + \text{c.c.}] \quad [57]$$

where $A(z, t)$ represents the slowly varying complex envelope of the pulse, $\psi(x, y)$ represents the transverse electric field distribution, ω_0 represents center frequency, and β_0 represents the propagation constant at ω_0.

In the presence of attenuation, second-order dispersion and third-order nonlinearity, the complex envelope $A(z, t)$ can be shown to satisfy the following equation:

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A - \beta_1\frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} - i\gamma|A|^2 A \quad [58]$$

Here

$$\beta_1 = \left|\frac{\mathrm{d}\beta}{\mathrm{d}\omega}\right|_{\omega=\omega_0} = \frac{1}{v_{\mathrm{g}}} \quad [59]$$

represents the inverse of the group velocity of the pulse, and

$$\beta_2 = \left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_0} = -\frac{\lambda_0^2}{2\pi c}D \quad [60]$$

where D represents the group velocity dispersion (measured in ps/km nm).

The various terms on the RHS of the eqn [58] represent the following:

I term: attenuation
II term: group velocity term
III term: second-order dispersion
IV term: nonlinear term

If we change to a moving frame defined by coordinates $T = t - \beta_1 z$, eqn [58] becomes

$$\frac{\partial A}{\partial z} = -\frac{\alpha}{2}A + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [61]$$

If we neglect the attenuation term, we obtain the following equation which is also referred to as the nonlinear Schrödinger equation:

$$\frac{\partial A}{\partial z} = i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} - i\gamma|A|^2 A \quad [62]$$

The above equation has a solution given by

$$A(z, t) = A_0 \operatorname{sech} \sigma T\, \mathrm{e}^{-\mathrm{i}gz} \quad [63]$$

with

$$A_0^2 = -\frac{\beta_2}{\gamma}\sigma^2, \qquad g = -\frac{\sigma^2}{2}\beta_2 \quad [64]$$

Equation [63] represents an envelope soliton and has the property that it propagates undispersed through the medium. The full width at half maximum (FWHM) of the pulse envelope will be given by $\tau_{\mathrm{f}} = 2\tau_0$, where

$$\operatorname{sech}^2\sigma\tau_0 = \frac{1}{2} \quad [65]$$

which gives the FWHM τ_{f}:

$$\tau_{\mathrm{f}} = 2\tau_0 = \frac{2}{\sigma}\ln(1 + \sqrt{2}) \approx \frac{1.7627}{\sigma} \quad [66]$$

The peak power of the pulse is:

$$P_0 = |A_0|^2 = \frac{|\beta_2|}{\gamma}\sigma^2 \quad [67]$$

Replacing σ by τ_{f}, we obtain

$$P_0\tau_{\mathrm{f}}^2 \approx \frac{\lambda_0^2}{2c\gamma}D \quad [68]$$

where we have used eqn [60]. The above equation gives the required peak for a given τ_{f} for the formation of a soliton pulse.

As an example, we have $\tau_{\mathrm{f}} = 10$ ps, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$, $\lambda_0 = 1.55\ \mu\mathrm{m}$, $D = 2$ ps/km nm and the required peak power will be $P_0 = 33$ mW.

Soliton pulses are being extensively studied for application to long distance optical fiber communication. In actual systems, the pulses have to be optically amplified at regular intervals to compensate for the loss suffered by the pulses. The amplification could be carried out using erbium doped fiber amplifiers (EDFAs) or fiber Raman amplifiers.

Spectral Broadening due to SPM

In the presence of only nonlinearity, eqn [61] becomes

$$\frac{\mathrm{d}A}{\mathrm{d}z} = -i\gamma|A|^2 A \qquad [69]$$

whose solution is given by

$$A(z,t) = A(z=0,t)\mathrm{e}^{-\mathrm{i}\gamma Pz} \qquad [70]$$

where $P = |A|^2$ is the power in the pulse. If P is a function of time, then time dependent phase term at $z = L$ becomes

$$\mathrm{e}^{\mathrm{i}\phi(t)} = \mathrm{e}^{\mathrm{i}[\omega_0 t - \gamma P(t)L]} \qquad [71]$$

We can define an instantaneous frequency as

$$\omega(t) = \frac{\mathrm{d}\phi}{\mathrm{d}t} = \omega_0 - \gamma L\frac{\mathrm{d}P}{\mathrm{d}t} \qquad [72]$$

for a Gaussian pulse:

$$P = P_0\,\mathrm{e}^{-2T^2/\tau_0^2} \qquad [73]$$

giving

$$\omega(t) = \omega_0 + \frac{4\gamma LTP_0\,\mathrm{e}^{-2T^2/\tau_0^2}}{\tau_0^2} \qquad [74]$$

Thus the instantaneous frequency within the pulse changes with time, leading to chirping of the pulse (see Figure 8). Note that since the pulsewidth has not changed, but the pulse is chirped, the frequency spectrum of the pulse has increased. Thus SPM leads to the generation of new frequencies. By Fourier transform theory, an increased spectral width implies that the pulse can now be compressed in the temporal domain by passing it through a medium with the proper sign of dispersion. This is indeed one of the standard techniques to realize ultrashort femtosecond optical pulses.

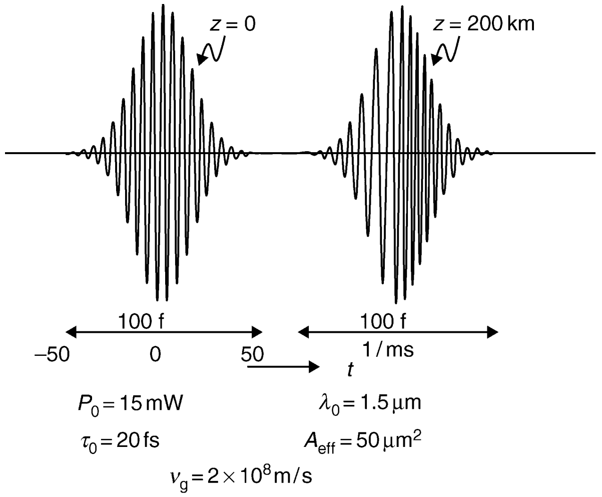

Figure 8 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse.

Cross Phase Modulation (XPM)

Like SPM, cross-phase modulation also arises due to the intensity dependence of refractive index, leading to spectral broadening. Unlike SPM, in the case of XPM, intensity variations of a light beam at a particular frequency modulate the phase of light beam at another frequency. If the signals at both frequencies are pulses, then due to difference in group velocities of the pulses, there is a walk off between the two pulses, i.e., if they start together, they will separate as they propagate through the medium. Nonlinear interaction takes place as long as they physically overlap in the medium. Smaller the dispersion, smaller will be the difference in group velocities (assuming closely spaced wavelengths) and the longer they will overlap. This would lead to stronger XPM effects. At the same time, if two pulses pass through each other, then since one pulse will interact with both the leading and the trailing edge of the other pulse, XPM effects will be nil provided there is no attenuation. In the presence of attenuation in the medium, the pulse will still get modified due to XPM. A similar situation can occur when the two interacting pulses are passing through an optical amplifier.

To study XPM, we assume simultaneous propagation of two waves at two different frequencies through the medium. If ω_1 and ω_2 represent the two frequencies, then one obtains for the variation of the amplitude A_1 of the frequency ω_1:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -i\gamma(\tilde{P}_1 + 2\tilde{P}_2)A_1 \qquad [75]$$

where $\tilde{P}_1$ and $\tilde{P}_2$ represent the powers at frequencies ω_1 and ω_2, respectively. The first term in eqn [75] represents SPM while the second term corresponds to XPM. If the powers are assumed to attenuate at the same rate, i.e.:

$$\tilde{P}_1 = P_1\,\mathrm{e}^{-\alpha z}, \qquad \tilde{P}_2 = P_2\,\mathrm{e}^{-\alpha z} \qquad [76]$$

then the solution of eqn [75] is

$$A_1(L) = A_1(0)\,\mathrm{e}^{-\mathrm{i}\gamma(P_1+2P_2)L_{\mathrm{eff}}} \qquad [77]$$

where, as before, L_{eff} represents the effective length of the medium. When we are studying the effect of power at ω_2 on the light beam at frequency ω_1, we will refer to the wave at frequency ω_2 as pump, and the wave at frequency ω_1 as probe or signal. From eqn [77] it is apparent that the phase of signal at frequency ω_1 is modified by the power at another frequency. This is referred to as XPM. Note also that XPM is twice as effective as SPM.

Similar to the case of SPM, we can now write for the instantaneous frequency in the presence of XPM as eqn [72]:

$$\omega(t) = \omega_0 - 2\gamma L_{\text{eff}} \frac{\mathrm{d}P_2}{\mathrm{d}t} \qquad [78]$$

Hence the part of the signal that is influenced by the leading edge of the pump will be down-shifted in frequency (since in the leading edge $\mathrm{d}P_2/\mathrm{d}t > 0$) and the part overlapping with the trailing edge will be up-shifted in frequency (since $\mathrm{d}P_2/\mathrm{d}t < 0$). This leads to a frequency chirping of the signal pulse just as in the case of SPM.

If the probe and pump beams are pulses, then XPM can lead to induced frequency shifts depending on whether the probe pulse interacts only with the leading edge or trailing edge or both, as they both propagate through the medium. Let us consider a case when the group velocity of pump pulse is greater than that of the probe pulse. Thus, if both pulses enter the medium together, then since the pump pulse travels faster, the probe pulse will interact only with the trailing edge of the pump. Since in this case $\mathrm{d}P_2/\mathrm{d}t$ is negative, the probe pulse suffers a blue-induced frequency shift. Similarly if the pulses enter at different instants but completely overlap at the end of the medium, then $\mathrm{d}P_2/\mathrm{d}t > 0$ and the probe pulse would suffer a red-induced frequency shift. Indeed, if the two pulses start separately and walk through each other, then there is no induced shift due to cancellation of shifts induced by leading and trailing edges of the pump. Figure 9 shows measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber.

When pulses of light at two different wavelengths propagate through an optical fiber, due to different group velocities of the pulses, they pass through each other, resulting in what could be termed as a collision. In the linear case, the pulses will pass through without affecting each other, but when intensity levels are high, XPM induces phase shifts in both pulses. We can define a parameter termed walk off length L_{wo} which is the length of the fiber required for the interacting pulses to walk off relative to each other. The walk off length is given by

$$L_{\text{wo}} = \frac{\Delta\tau}{D\Delta\lambda} \qquad [79]$$

where D represents the dispersion coefficient, and $\Delta\lambda$ represents the wavelength separation between the interacting pulses. For return to zero (RZ) pulses, $\Delta\tau$ represents the pulse duration while for nonreturn to zero (NRZ) pulses, $\Delta\tau$ represents the rise term or fall time of the pulse. Closely spaced channels will thus interact over longer fiber lengths, thus leading to greater XPM effects. Larger dispersion coefficients will reduce L_{wo} and thus the effects of XPM. Since the medium is attenuating, the power carried by the pulses decreases as they propagate and thus leading to reduced XPM effect. The characteristic length for attenuation is the effective length L_{eff}, defined by eqn [53]. If $L_{\text{wo}} \ll L_{\text{eff}}$, then over the length of interaction of the pulses, the intensity levels do not change appreciably and the magnitude of the XPM-induced effects will be proportional to the wavelength spacing $\Delta\lambda$. For small $\Delta\lambda$'s, $L_{\text{wo}} \gg L_{\text{eff}}$ and the interaction length is now determined by the fiber losses (rather than by walk off) and the XPM-induced effects become almost independent of $\Delta\lambda$. Indeed, if we consider XPM effects between a continuous wave (cw) probe beam and a sinusoidally industry modulated pump beam, then the amplitude of the XPM-induced

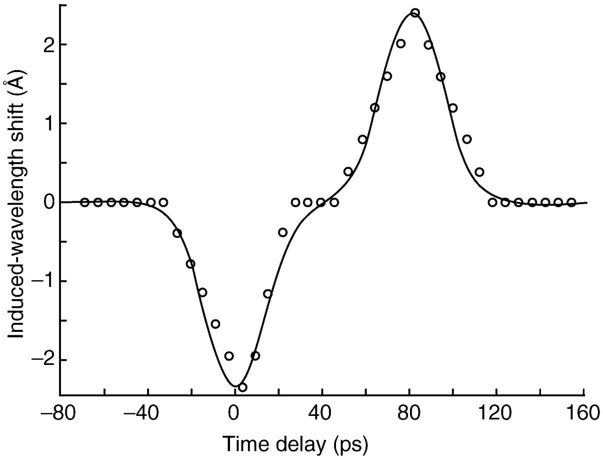

Figure 9 Measured induced frequency shift of 532 nm probe pulse as a function of the delay between this pulse and a pump pulse at 1064 nm, when they propagate through a 1 m long optical fiber. (Reproduced from Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941 with permission from the American Institute of Physics.)

phase shift ($\Delta\Phi_{\mathrm{pr}}$) in the probe beam is given by:

$$\begin{aligned} \Delta\Phi_{\mathrm{pr}} &\approx 2\gamma P_{2\mathrm{m}} L_{\mathrm{eff}} \quad \text{for} \quad L_{\mathrm{wo}} \gg L_{\mathrm{eff}} \\ \Delta\Phi_{\mathrm{pr}} &\approx 2\gamma P_{2\mathrm{m}} L_{\mathrm{wo}} \quad \text{for} \quad L_{\mathrm{wo}} \ll L_{\mathrm{eff}} \end{aligned} \qquad [80]$$

Here $P_{2\mathrm{m}}$ is the amplitude of the sinusoidal power modulation of the pump beam.

XPM-induced intensity interference can be studied by simultaneously propagating an intensity modulated pump signal and a cw probe signal at a different wavelength. The intensity modulated signal will induce phase modulation on the cw probe signal and the dispersion of the medium will convert the phase modulation to intensity modulation of the probe. Thus, the magnitude of the intensity fluctuation of the probe signal serves as an estimate of the XPM induced interference. Figure 10 shows the intensity fluctuations on a probe signal at 1550 nm, induced by a modulated pump for a channel separation of 0.6 nm. Figure 11 shows the variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. The experiment has been performed over four amplified spans of 80 km of standard single mode fiber (SMF) and nonzero dispersion shifted fiber (NZDSF). The large dispersion in SMF has been compensated using dispersion compensating chirped gratings. The probe modulation, in the case of SMF, decreases approximately linearly with $1/\Delta\lambda$ for all $\Delta\lambda$s; the modulation is independent of $\Delta\lambda$. In contrast the NZDSF shows a constant modulation for $\Delta\lambda \leq 1$ nm.

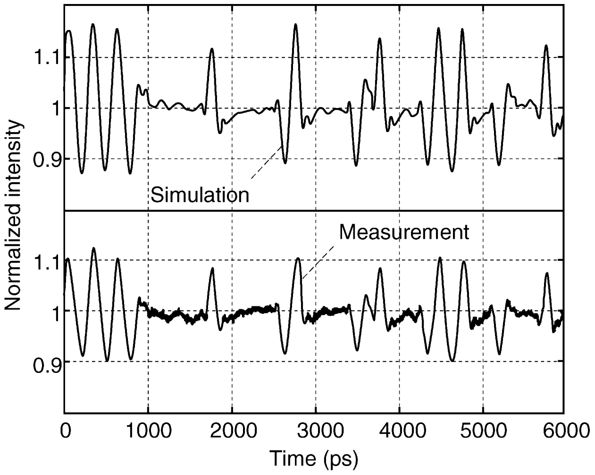

Figure 10 Intensity fluctuations induced by cross phase modulation on a probe signal at 1550 nm by a modulated pump for a channel separation of 0.6 nm. (Reproduced with permission from Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595, © 2004 IEEE.)

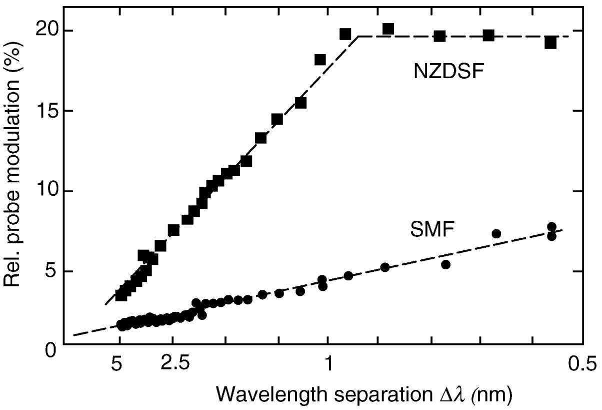

Figure 11 Variation of the RMS value of probe intensity modulation with the wavelength separation between the intensity modulated signal and the probe. (Reproduced with permission from Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90, © 2004 IEEE.)

This is consistent with the earlier discussion in terms of L_{wo} and L_{eff}.

Four-Wave Mixing (FWM)

Four-wave mixing (FWM) is a nonlinear interaction that occurs in the presence of multiple wavelengths in a medium, leading to the generation of new frequencies. Thus, if light waves at three different frequencies ω_2, ω_3, ω_4 are launched simultaneously into a medium, the same nonlinear polarization that led to intensity dependent refractive index, leads to nonlinear polarization component at a frequency:

$$\omega_1 = \omega_3 + \omega_4 - \omega_2 \qquad [81]$$

This nonlinear polarization, under certain conditions, leads to the generation of electromagnetic waves at ω_1. This process is referred to as four-wave mixing due to the interaction between four different frequencies. Degenerate four-wave mixing corresponds to the case when two of the input waves have the same frequency. Thus, if $\omega_3 = \omega_4$, then inputting waves at ω_2 and ω_3 leads to the generation of waves at the frequency

$$\omega_1 = 2\omega_3 - \omega_2 \qquad [82]$$

During FWM process, there are four different frequencies present at any point in the medium. If we write the electric field of the waves as

$$E_i = \frac{1}{2}[A_i(z)\psi_i(x,y)\mathrm{e}^{\mathrm{i}(\omega_i t - \beta_i z)} + \mathrm{c.c.}], \qquad [83]$$

$$i = 1, 2, 3, 4$$

where, as before, $A_i(z)$ represents the amplitude of the wave, $\psi_i(x, y)$ the transverse field distribution, and β_i the propagation constant of the wave. The total electric field is given by

$$E = E_1 + E_2 + E_3 + E_4 \qquad [84]$$

Substituting for the total electric field in the equation for nonlinear polarization, the term with frequency ω_1 comes out as

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{1}{2}[P_{\mathrm{NL}}^{(\omega_1)}\,\mathrm{e}^{\mathrm{i}(\omega_1 t - \beta_1 z)} + \mathrm{c.c.}] \qquad [85]$$

where

$$P_{\mathrm{NL}}^{(\omega_1)} = \frac{3\varepsilon_0}{2}\chi^{(3)}A_2^* A_3 A_4 \psi_2 \psi_3 \psi_4\,\mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [86]$$

and

$$\Delta\beta = \beta_3 + \beta_4 - \beta_2 - \beta_1 \qquad [87]$$

In writing eqn [86], we have only considered the FWM term, neglecting the SPM and XPM terms.

Substituting the expression for $P_{\mathrm{NL}}^{(\omega_1)}$ in the wave equation for ω_1 and making the slowly varying approximation (in a manner similar to that employed in the case of SPM and XPM), we obtain the following equation for $A_1(z)$:

$$\frac{\mathrm{d}A_1}{\mathrm{d}z} = -2i\gamma A_2^* A_3 A_4\,\mathrm{e}^{-\mathrm{i}\Delta\beta z} \qquad [88]$$

where γ is defined by eqn [48] with $k_0 = \omega/c$, ω representing the average frequency of the four interacting waves, and A_{eff} is the average effective area of the modes.

Assuming all waves to have the same attenuation coefficient α and neglecting depletion of waves at frequencies ω_2, ω_3 and ω_4, due to nonlinear conversion we obtain for the power in the frequency ω_1 as

$$P_1(L) = 4\gamma^2 P_2 P_3 P_4 L_{\mathrm{eff}}^2 \eta\,\mathrm{e}^{-\alpha L} \qquad [89]$$

where

$$\eta = \frac{\alpha^2}{\alpha^2 + \Delta\beta^2}\left[1 + \frac{4\mathrm{e}^{-\alpha L}\sin^2\dfrac{\Delta\beta L}{2}}{(1 - \mathrm{e}^{-\alpha L})^2}\right] \qquad [90]$$

and L_{eff} is the effective length (see eqn [53]). Maximum four-wave mixing takes place when $\Delta\beta = 0$, since in such a case $\eta = 1$. Now:

$$\Delta\beta = \beta(\omega_3) + \beta(\omega_4) - \beta(\omega_2) - \beta(\omega_1) \qquad [91]$$

Since the frequencies are usually close to each other, we can make a Taylor series expansion about any frequency, say ω_2. In such a case, we obtain

$$\Delta\beta = (\omega_3 - \omega_2)(\omega_3 - \omega_1)\left|\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right|_{\omega=\omega_2} \qquad [92]$$

In optical fiber communication systems, the channels are usually equally spaced. Thus, we assume the frequencies to be given by

$$\begin{aligned} &\omega_4 = \omega_2 + \Delta\omega, \qquad \omega_3 = \omega_2 - 2\Delta\omega \ \text{ and} \\ &\omega_1 = \omega_2 - \Delta\omega \end{aligned} \qquad [93]$$

Using these frequencies and eqn [60], eqn [92] gives us:

$$\Delta\beta = -\frac{4\pi D\lambda^2}{c}(\Delta\nu)^2 \qquad [94]$$

where $\Delta\omega = 2\pi\Delta\nu$. Thus maximum FWM takes place when $D = 0$. This is the main problem in using wavelength division multiplexing (WDM) in dispersion shifted fibers which are characterized by zero dispersion at the operating wavelength of 1550 nm, as FWM will then lead to crosstalk among various channels. FWM efficiency can be reduced by using fiber with nonzero dispersion. This has led to the development of nonzero dispersion shifted fiber (NZDSF) which have a finite nonzero dispersion of about ±2 ps/km nm at the operating wavelength.

From eqn [94], we notice that for a given dispersion coefficient D, FWM efficiency will reduce as $\Delta\nu$ increases.

In order to get a numerical appreciation, we consider a case with $D = 0$, i.e., $\Delta\beta = 0$. For such a case $\eta = 1$. If all channels were launched with equal power P_{in} then:

$$P_1(L) = 4\gamma^2 P_{\mathrm{in}}^3 L_{\mathrm{eff}}^2\,\mathrm{e}^{-\alpha L} \qquad [95]$$

Thus the ratio of power generated at ω_1, due to FWM and that existing at the same frequency, is

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} = \frac{P_1(L)}{P_{\mathrm{in}}\,\mathrm{e}^{-\alpha L}} = 4\gamma^2 P_{\mathrm{in}}^2 L_{\mathrm{eff}}^2 \qquad [96]$$

Typical values are $L_{\mathrm{eff}} = 20$ km, $\gamma = 2.4\ \mathrm{W}^{-1}\ \mathrm{km}^{-1}$. Thus:

$$\frac{P_{\mathrm{g}}}{P_{\mathrm{out}}} \approx 0.01 P_{\mathrm{in}}^2(\mathrm{mW}^2) \qquad [97]$$

Figure 12 shows the output spectrum measured at the output of a 25 km-long dispersion shifted fiber ($D = -0.2$ ps/km nm) when three 3 mW wavelengths are launched simultaneously. Notice the generation of many new frequencies by FWM. Figure 13 shows the

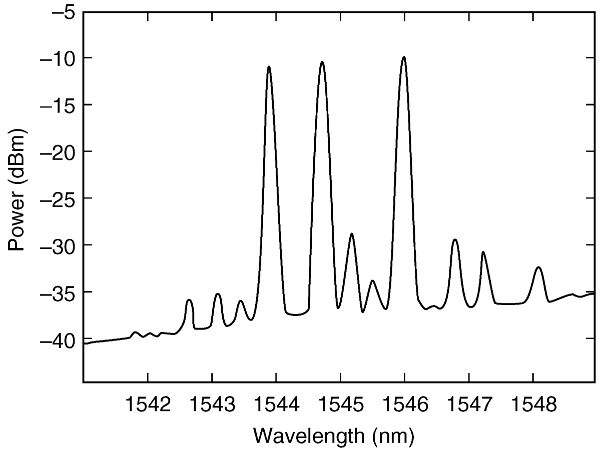

Figure 12 Generation of new frequencies because of FWM when waves at three frequencies are incident in the fiber. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

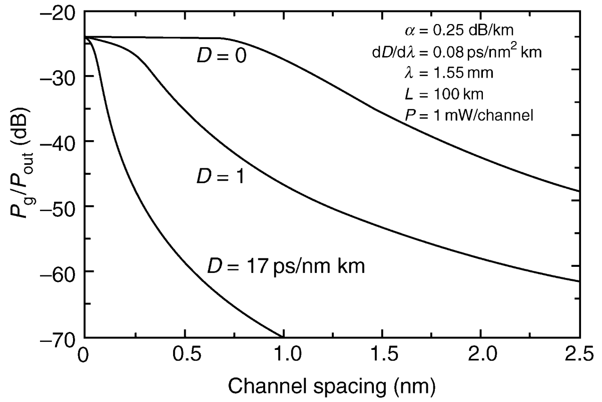

Figure 13 Ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. (Reproduced with permission from Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of lightwave Technology* 13: 841–849, © 2004 IEEE.)

ratio of generated power to the output as a function of channel spacing $\Delta\lambda$ for different dispersion coefficients. It can be seen that by choosing a nonzero value of dispersion, the four-wave mixing efficiency can be reduced. Larger the dispersion coefficient, smaller can be the channel spacing for the same crosstalk.

Since dispersion leads to increased bit error rates in fiber optic communication systems, it is important to have low dispersion. On the other hand, lower dispersion leads to crosstalk due to FWM. This problem can be resolved by noting that FWM depends on the local dispersion value in the fiber, while the pulse spreading at the end of a link depends on the overall dispersion in the fiber link. If one chooses a link made up of positive and negative dispersion coefficients, then by an appropriate choice of the lengths of the positive and negative dispersion fibers, it would be possible to achieve a zero total link dispersion while at the same time maintaining a large local dispersion. This is referred to as dispersion management in fiber optic systems.

Although FWM leads to crosstalk among different wavelength channels in an optical fiber communication system, it can be used for various optical processing functions such as wavelength conversion, high-speed time division multiplexing, pulse compression, etc. For such applications, there is a concerted worldwide effort to develop highly nonlinear fibers with much smaller mode areas and higher nonlinear coefficients. Some of the very novel fibers that have been developed recently include holey fibers, photonic bandgap fibers, or photonic crystal fibers which are very interesting since they possess extremely small mode effective areas ($\sim 2.5\ \mu m^2$ at 1550 nm) and can be designed to have zero dispersion even in the visible region of the spectrum. This is expected to revolutionize nonlinear fiber optics by providing new geometries to achieve highly efficient nonlinear optical processing at lower powers.

Supercontinuum Generation

Supercontinuum (SC) generation is the phenomenon in which a nearly continuous spectrally broadened output is produced through nonlinear effects on high peak power picosecond and subpicosecond pulses. Such broadened spectra find applications in spectroscopy, wavelength characterization of optical components such as a broadband source from which many wavelength channels can be obtained by slicing the spectrum.

Supercontinuum generation in an optical fiber is a very convenient technique since it provides a very broad bandwidth (>200 nm), the intensity levels can be maintained high over long interaction lengths by choosing small mode areas, and the dispersion profile of the fiber can be appropriately designed by varying the transverse refractive index profile of the fiber.

The spectral broadening that takes place in the fiber is attributed to a combination of various third-order effects such as SPM, XPM, FWM, and Raman scattering. Since dispersion plays a significant role in the temporal evolution of the pulse, different dispersion profiles have been used in the literature to achieve broadband SC.

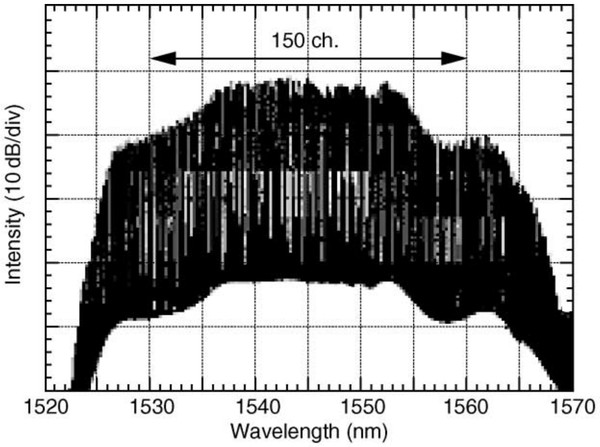

(a) SC spectrum

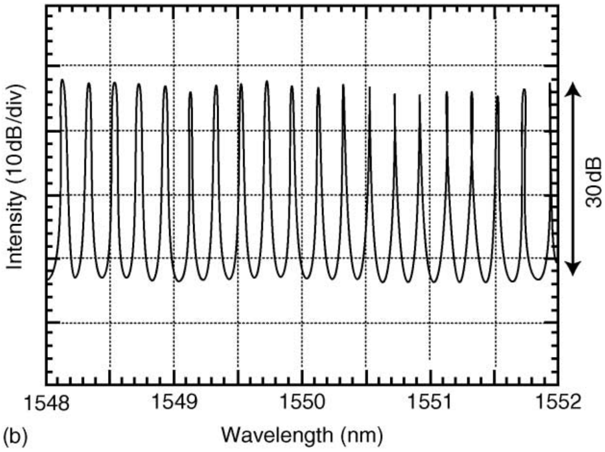

Figure 14 Super Continuum spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. (Reproduced from Yamada E, Takara H, Ohara T, Satc K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306 with permission from IEEE.)

Some studies have used dispersion decreasing fibers, dispersion flattened fibers, while others have used a constant anamolous dispersion fiber followed by a normal dispersion fiber.

Figure 14 shows the SC spectrum generated by passing 25 GHz optical pulse train at 1544 nm generated by a mode locked laser diode and amplified by an erbium doped fiber. The fiber used for SC generation is a polarization maintaining dispersion flattened fiber. The output is a broad spectrum containing more than 150 spectral components at a spacing of 25 GHz with a flat top spectrum over 18 nm. The output optical powers range from −9 dBm to +3 dBm. Such sources are being investigated as attractive solutions for dense WDM (DWDM) optical fiber communication systems.

Conclusions

The advent of lasers provided us with an intense source of light and led to the birth of nonlinear optics. Nonlinear optical phenomena exhibit a rich variety of effects and coupled with optical waveguides and fibers, such effects can be observed even at moderate optical powers. With the realization of compact femtosecond lasers, highly nonlinear holey optical fibers, quasi phase matching techniques, etc, the field of nonlinear optics is expected to grow further and lead to commercial exploitation such as in compact blue lasers, soliton laser sources and multiwavelength sources and for fiber optic communication.

List of Units and Nomenclature

Attenuation	The decrease in optical power of a propagating mode usually expressed in dB/km
Birefringent medium	A medium characterized by two different modes of propagation characterized by two different polarization states of a plane light wave along any direction
Cross-phase modulation (XPM)	The phase modulation of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by another propagating light wave
Effective length (L_{eff})	The length over which most of the nonlinear effect is accumulated in a propagating light beam
Four-wave mixing (FWM)	The mixing of four propagating light waves due to intensity dependent refractive index of the medium. This mixing leads to the generation of new frequencies from a set of two or three input light beams
Mode effective area (A_{eff})	The cross sectional area of the mode which determines the nonlinear effects on the propagating mode. This is different from the core area
Nonlinear polarization (P_{NL})	When the response of the medium to an applied electric field is not proportional to the applied electric field, then this results in a nonlinear polarization
Optical waveguides	Devices which are capable of guiding optical beams by overcoming diffraction effects using

	the phenomenon of total internal reflection
Phase coherence length	The minimum crystal length upto which the second harmonic power generated by the nonlinear interaction increases
Phase matching	Condition under with the nonlinear polarization generating an electromagnetic wave propagates at the same phase velocity as the generated electromagnetic wave
Pulse dispersion	The broadening of a light pulse as it propagates in an optical fiber
Quasi-phase matching (QPM)	Periodic modulation of the nonlinear coefficient to achieve efficient nonlinear interaction
Second-harmonic generation (SHG)	The generation of a wave of frequency 2ω from an incident wave of frequency ω through nonlinear interaction in a medium
Self-phase modulation (SPM)	The modulation of the phase of a propagating light wave due to the nonlinear change in refractive index of the medium brought about by itself
Supercontinuum generation (SC)	The phenomenon in which a nearly contiunuous spectrally broadband output is produced through nonlinear effects on high peak power picosecond and sub-picosecond pulses
Walk off length (L_{wo})	Length of the fiber required for two interacting pulses at different wavelengths to walk off relative to each other

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation; Four-Wave Mixing. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, MA: Academic Press.

Armstrong JA, Bloembergen N, Ducuing J and Pershan PS (1962) Interactions between light waves in a nonlinear dielectric. *Physics Review* 127: 1918.

Baldeck PL, Alfano RR and Agrawal GP (1998) Induced frequency shift of copropagating ultrafast optical pulses. *Applied Physics Letters* 52: 1939–1941.

Baldi P (1992) *Generation de fluorescence parametrique guidee sur niobate et tantalite de lithium polarizes periodiquement. Etudes preliminaire d'un oscillateur parmetrique optique integree*. Doctoral Thesis of the University of Nice, France.

Chiang TK, Kagi N, Marhic ME and Kazovsky LG (1996) Cross phase modulation in fiber links with multiple optical amplifiers and dispersion compensators. *Journal of Lightwave Technology* 14: 249–259.

Franken PA, Hill AE, Peter CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Ghatak AK and Thyagarajan K (1987) *Optical Electronics*. Cambridge, UK: Cambridge University Press.

Ghatak AK and Thyagarajan K (1998) *Introduction to Fiber Optics*. Cambridge, UK: Cambridge University Press.

Lim EJ, Fejer MM and Byer RL (1989) Second-harmonic generation green light in periodically poled planar lithium niobate waveguide. *Electronic Letters* 25: 174.

Lim EJ, Fejer MM, Byer RL and Kozovsky WJ (1989) *Electronic Letters* 25: 731–732.

Myers LE and Bosenberg WR (1997) Periodically poled lithium niobate and quasi phase matcher parametric oscillators. *IEEE Journal of Quantum Electronics* 33: 1663–1672.

Newell AC and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison Wesley Publishing Co.

Rapp L (1997) Experimental investigation of signal distortions induced by cross phase modulation combined with dispersion. *IEEE Photon Technical Letters* 9: 1592–1595.

Shen YR (1989) *Principles of Nonlinear Optics*. New York: Wiley.

Shtaif M, Eiselt M and Garret LD (2000) Cross-phase modulation distortion measurements in multispan WDM systems. *IEEE Photon Technical Letters* 12: 88–90.

Tkach RW, Chraplyvy AR, Forghiari F, Gnanck AH and Derosier RM (1995) Four-photon mixing and high speed WDM systems. *Journal of Lightwave Technology* 13: 841–849.

Wang QZ, Lim QD, Lim DH and Alfano R (1994) High resolution spectra of self phase modulation in optical fibers. *Journal of Optical Society of America* B-11: 1084.

Webjorn J, Siala S, Nan DW, Waarts RG and Lang RJ (1997) Visible laser sources based frequency doubling in nonlinear waveguides. *IEEE Journal of Quantum Electronics* 33: 1673–1686.

Yamada M, Nada N, Saitoh M and Watanabe K (1993) *Applied Physics Letters* 62: 435–436.

Yamada E, Takara H, Ohara T, Sato K, Morioka T, Jinguji K, Itoh M and Ishi M (2001) A high SNR, 150 Ch. Supercontinuum CW optical source with high SNR and precise 25 GHz spacing for 10 Gbit/s DWDM systems. *Electronics Letters* 37: 304–306.

Yariv A (1997) *Optical Electronics in Modern Communications*, 5th ed. New York: Oxford University Press.

Optical Fiber Cables

G Galliano, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In order to be put to practical use in the telecommunication network, optical fibers must be protected from environmental and mechanical stresses which can change their transmission characteristics and reliability.

The fibers, during manufacturing, are individually protected with a thin plastic layer (extruded during the drawing) to preserve directly their characteristics from damage due to the environment and handling. This coating surrounding the fiber cladding (with an external diameter of 125 μm) is known as the primary coating, which is obtained with a double layer of UV resin for a maximum external diameter of about 250 μm.

Primary coated fiber, at the end of the manufacturing process, is subjected to a dynamic traction test (proof-test) to monitor its mechanical strength. This test, carried out on all fibers manufactured, consists of the application of a well-defined strain (from 0.5% to 1%) for a fixed time (normally 1 s) and thus permits the exclusion of, from successive cabling, those fibers with poor mechanical characteristics.

At the end of the manufacturing process the optical fibers are not directly usable. It is necessary to surround the single fiber (or groups of fibers) with a structure which provides protection from mechanical and chemical hazards and allows for stresses that may be applied during cable manufacture, installation, and service.

Suitable protection is provided using fiber conditioning within the cable (secondary coating or loose cabling), while entire protection is obtained with some cable elements, such as strength members, core assembling, filling, and outer protections. Moreover, the cable structure and its design is heavily influenced by the application (e.g., number of fibers), installation (laying and splicing), and environmental conditions of service (underground, aerial, or submarine).

In this article, the main characteristics of optical cables (with both multimode or single mode fibers) are described and analyzed, together with suitable solutions.

Secondary Fiber Coating

The fiber, before cabling, can be protected with a secondary coating (a tight jacketing) to preserve it during cable stranding. However, fiber with only a primary coating may be cabled with a suitable cable protection.

In the first case the fiber can be directly stranded to form a cable core; in the second case, a single group of fibers must be inserted into a loose structure (tubes or grooves) with a proper extra-length. These two solutions are not always distinguishable; in fact, tight fibers inserted into loose tubes or grooves are often used.

The implementation of the secondary coating and the choice of proper materials requires particular attention, in order to avoid microbending effects on the fiber, which can cause degradation of the transmission characteristics. Microbending is caused by the pressure of the fiber on a microrough surface, or by the fiber buckling, due to the contraction of the structure (e.g., secondary coating) containing it. Microbending causes an increase in fiber attenuation at long wavelengths (1300 and 1550 nm) for both single and multimode fibers, but is particularly emphasized in multimode fibers.

Tight Jacket

In a tight protection, the primary coating fiber is surrounded by a second plastic jacket in contact with the fiber. This jacket mainly protects the fiber from lateral stresses that can cause the microbending.

The fiber can be individually protected with a single or double layer of plastic material or in a ribbon structure. In Figure 1 four different types of tight jacket are shown. In Figure 1a and b the fibers are singularly protected with a single or double layer up to an external diameter of 0.9 mm. The second type is normally preferred because the external hard plastic gives a good fiber protection and the inner soft plastic avoids microbending effects on the fiber.

In a ribbon structure, from 4 to 16 primary coated fibers are laid close and parallel and covered with plastic. Two types of ribbon are normally classified ‘encapsulated’ (Figure 1c) or ‘edge bonded’ (Figure 1d). In the first case, the fibers are assembled with a common secondary coating (one or two layers), in the second case, an adhesive is used to stick together the fibers with only the primary coating.

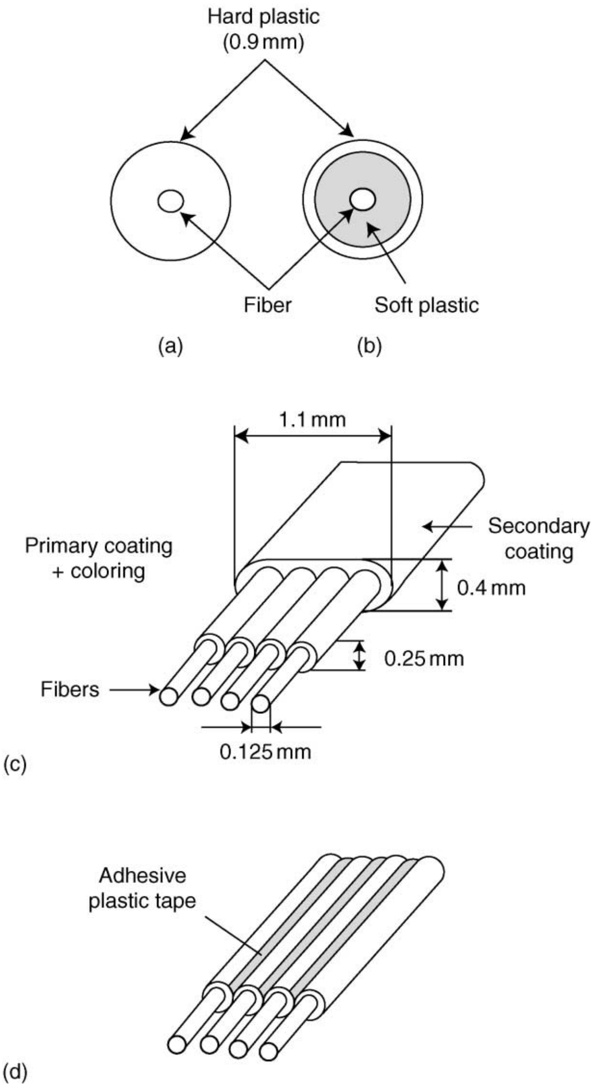

Figure 1 Different types of tight coatings.

Care should be taken in applying tight protection, in particular for a ribbon structure, to avoid an attenuation increase due to the action of the jacket on the fiber (in particular for the thermal contraction of the coating materials). The ribbon structure is designed to permit the simultaneous junction of multiple fibers; for this reason the ribbon geometry must be precision controlled.

Loose Jacket

In a loose protection the fiber can be inserted into a cable with only the primary coating. A thin layer of colored plastic may be added onto the primary coating to identify the fiber in the cable. The fibers are arranged in a tube or in a groove on a plastic core (slotted core). The diameter of the tube or the dimension of the core must be far larger than the diameter of the fiber so that the fiber is completely free inside the structure. The fibers can be inserted into a tube or into a slotted core, singularly or in groups. The space inside the tube around the fibers can be empty or filled with a suitable jelly. Some examples of loose structures are shown in Figure 2. In the tube structure, the external diameter of the jacket depends on the number of the fibers: for a single fiber up to 2 mm (Figure 2a), and for a multifiber tube (Figure 2b) up to 3 mm (for 10 fibers).

Examples of slotted core structures are shown in Figure 2c and d, respectively, for one fiber and for a group of fibers, the dimensions of grooves depending on the number of fibers.

Plastic materials with a high Young modulus are normally used for the tube that is extruded on the single or on the group of fibers. Instead, for the slotted core, plastic materials with good extrusion are used to obtain a precise profile of the grooves.

The loose structure guarantees freedom to the fibers in a defined interval of elongation and compression. Out of this interval, microbending effects can arise which increase fiber attenuation. Axial compression generally occurs when the cable is subjected to low temperatures and is generated from the thermal contraction of the cable materials that is different from that of the fibers. Axial elongation normally occurs when the cable is pulled or by the effects of high temperature.

For a correct design of a cable with a loose structure, the extra length of the fiber with respect to the tube or groove dimension and the stranding pitch must be evaluated, starting from the range of axial compression and elongation. In fact, care must be taken to prevent the fiber from being strained or forced against the tube or groove walls in the designed range.

For the tube solution, the extra length of the fibers is controlled during the tube extrusion; for the slotted core solution, the length is controlled during cabling, giving a controlled tension to the slotted core. The loose structure is often used for the cabling of fibers with tight jacketing, joining the advantages of the tight protection and loose cable. This solution is typical for ribbon cables.

Fiber Identification

To identify the fibers in a cable, a proper coloration with defined color codes must be given to the single fibers. In the tight protection, the tight buffer is directly colored during extrusion. In the loose protection, a thin colored layer is added on the primary coating of the fiber; tubes or slotted core may

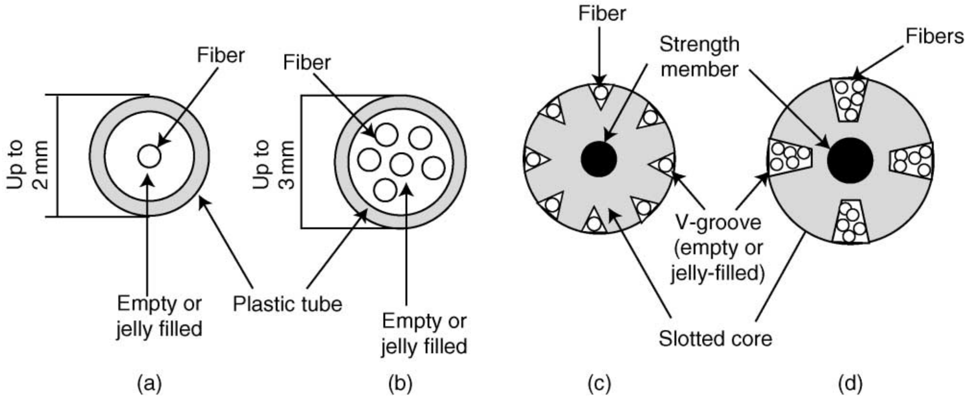

Figure 2 Different types of loose coatings.

be also colored to identify groups of fibers in high potentiality cables.

Cable Components

The fundamental components of an optical cable are: the cable core with the optical fibers, the strength member (one or more) which provide the mechanical strength of the cable, sheaths which provide the external protection, and filling materials.

Cable Core

The optical fibers with the secondary coating (tight or loose) are rejoined together in a cable core. For tight fibers or loose tubes, the cable core is obtained by stranding the fibers or the tubes around central elements that normally act as strength members too (see below).

The stranding pitch must be long enough to avoid excessive bending of the fibers and short enough to guarantee stress apportionment among the fibers in cable bends and maximum elongation of the cable without fiber strain. Normally, a helical stranding is used, but often SO stranding, that consists in an inversion of stranding direction every three of more turns, is employed. The second stranding type makes the cabling easier but may reduce its performance.

The unit core or the slotted core are made of plastic and normally incorporate a central strength member.

Strength Member

The strength member may be defined as an element that provides the mechanical strength of the cable, to withstand elongation and contraction during the installation and to ensure thermal stability. Fiber elongation, when the cable is pulled or the fiber compressed due to low temperatures, depends greatly on the characteristics of the strength member inserted in the cable.

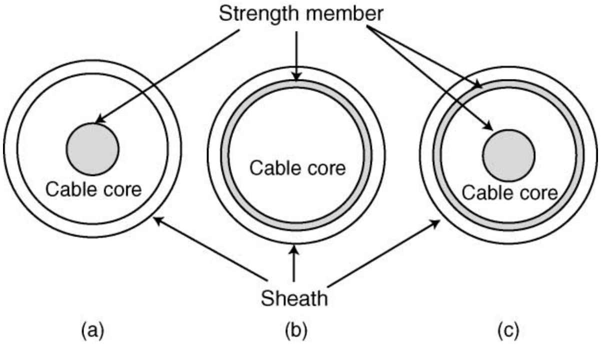

Figure 3 Different locations of the strength member in an optical cable.

As the maximum elongation values suggested for the fibers are in the range from 20 to 30% of the proof-test value, the maximum elongation allowed for a tight cable is ~0.1% instead of ~0.3% for a loose cable. To warrant these performances, a material with a high Young modulus is normally used as a strength member. Also, the strength member must be light (to avoid excessive cable weight) and flexible: these properties make it easier when laying the cable in ducts.

The strength member may be metallic or non-metallic and may be inserted in the core (Figure 3a) or in the outer part of the cable (Figure 3b), normally under the external sheath. Often two strength members are placed in an optical cable: the first one in the core and the second one under the sheath (Figure 3c).

Metallic strength members, generally one or more wires of steel, are inserted in the center or in the outer part of the cable. The steel, for its low thermal expansion coefficient and high Young modulus, behaves as a good strength member, particularly when it is arranged in the central core. Steel is a nonexpensive material, but needs protection against corrosion and electrical-induced voltages, such as by lightning or current flow in the ground.

If cables with high flexibility are requested, steel is not suitable and aramid fiber yarn or glass-fiber reinforced plastic (G-FRP) are used, particularly when the strength member is placed in the outer part of the cable. Glass fiber elements are generally used as a compressive strength member in the central cable core to avoid contraction at low temperatures, but they may be used as strength elements when the traction strength is low.

Aramid yarns are not active as resisting compression elements but are normally arranged in the outer part of the optical cable. Aramid yarns have a high Young modulus (as with steel) and low specific weight: in commercial form they consist of a large number of filaments gathered into a layer. Normally two layers stranded in opposite directions are used to avoid cable torsion during the installation.

Metal-free cables are protected by G-FRP and aramid yarns. This solution results in optical cables which are insensitive to electrical interferences.

Sheaths

The functions of the sheaths are to protect an optical cable during its life, in particular from mechanical and environmental stresses. During the installation or when the cable is in service, it may be subjected to mechanical stresses, for example, presence of humidity or water and chemical agents. The sheaths act as a barrier to avoid degradations of the fibers in the internal cable core.

The plastic sheaths (one or more, depending on the cable type and structure) are extruded over the core. This extrusion is a delicate operation in cable manufacturing because the cable core may be subjected to tension, while during the cooling the thermal compression of the plastic produces a stress on all the cable elements. Examples of cable sheaths are shown in Figure 4.

Different plastic materials, such as polyvinylchloride (PVC), polyethylene, and polyurethane are normally used. PVC has excellent mechanical properties, flexibility, and is flame retardant but is permeable to humidity. On the other hand, high-density polyethylene has a permeability lower than PVC, with good mechanical properties and low coefficient of friction, but it is more flammable than PVC and less flexible. Finally, polyurethane, for its softness and high flexibility, is often used for inner sheaths. When low toxicity is requested (e.g., in indoor cables), halogen-free materials are employed.

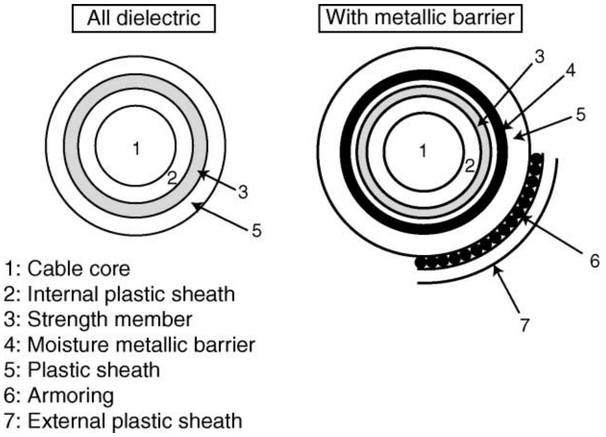

Figure 4 Examples of optical cable sheaths.

Often a metallic barrier is used under the external sheath to prevent moisture entering the cable core. The type most widely used is an aluminum ribbon (of a few tenths of a millimeter) bonded into the external polyethylene sheath, during the sheath extrusion. This structure is normally indicated as a LAP (laminated aluminum polyethylene) sheath.

When the cable is directly buried, a metallic armor (corrugated steel tape or armoring wires) is generally inserted. The metallic armor protects the cable core against radial stresses but it is also used as protection against rodents and insects.

Filling Compounds and Other Components

Fillings are used in optical cables to avoid the presence of moisture and water propagation in the cable core should a failure occur in the cable sheath. Suitable jelly compounds, with constant viscosity at a wide range of temperatures and chemical stabilities, are generally inserted in the cable core and, sometimes, in the outer protection layer. In loose structures, the jelly is in direct contact with the fibers: the compound must be compatible with them and guarantee the fiber freedom. Special compounds may be used to adsorb and permanently fix any hydrogen present in the cable core. The presence of hydrogen gas, that may be generated from the internal materials of the cable (mainly metals), is harmful to the fibers because it causes permanent attenuation increases and thus damage to them.

With the generic terminology ‘other components’, some other components used in the optical cables are included. They are often similar to those used in conventional cables. The most important are: insulated conductors, plastic fillers, and cushion layers and tapes around the fiber core.

Insulated copper conductors may be incorporated in the cable core, for example, to carry power supply or as service channels, or to detect the presence of moisture in the cable (in this case, the insulation is missing or partial). The conductors may cause problems due to the voltages induced by power lines or lighting.

Plastic fillers are often used to complete the cable core geometry. Cushion layers are used to protect the cable core against radial compression. Normally, they are based on plastic materials wound around the core. Tapes are wound around the cable core to hold the assemblies together and secondly to provide a heat barrier during the sheath extrusion. Mylar tapes are usually employed.

Optical Cable Structures, Types, and Applications

Cable Structures

The structure of an optical cable is strictly dependent on the construction method and the application type that determines the design and the grade of external protection. The main core (or inner) structures of an optical cable can be classified as: stranded structures (tight and loose); slotted core cable; or ribbon cable. In this section, a few examples of cable structures are presented. Below, separate sections are devoted to submarine cables and to special application cables, for their particular structures and features.

In the stranded tight structure, a low number of tight or loose fibers is stranded around a strength member to form a cable unit. Some of this cable unit may be joined to constitute a cable with a medium/high number of fibers (Figure 5a). Alternatively the fibers may be arranged in a multilayer structure (Figure 5b).

The main advantages of a tight stranded structure are the small dimensions of the cable core, even when a large number of fibers are involved, and also the facility of handling. Care must be taken in the design of the strength member: in fact, the fibers are immediately subjected to tension when the cable is pulled into the ducts or compressed when subjected to low temperatures.

In the stranded loose structures, two categories of cables are defined, one fiber per tube and a group of fibers (up to 20) per tube (Figure 5c). As already mentioned, in the loose structures, the fibers are free inside the tube and for this reason the cable is less critical than the tight cable. The external dimensions of loose cables are larger than tight cables, but the solution with multiform tubes allows for cables with a high number of fibers in relatively small dimensions.

Slotted core cables (Figure 2c,d) contain fibers with only the primary coating and may be divided, as in the stranded loose structure, into one fiber per groove or into a group of fibers per groove. The cable design and the choice of a suitable plastic material for the slotted core joined to a proper strength member, make it possible to obtain cables with fiber stability at a wide range of temperature. As in the loose stranded solutions, the slotted core solution is adopted to obtain cables with a high number of fibers in small dimensions.

In the ribbon solution, the fibers (from 4 to 16) are placed in a linear ribbon array. Many ribbons are stacked together to constitute the optical unit. The ribbon cable normally has a loose structure (loose tube or slotted core). Three structures are shown in Figure 6. In the first example (classical AT & T cable), 12 ribbons of 12 fibers are directly stacked and stranded in a tube (Figure 6a). In the second and third examples, ribbons are inserted into tubes or grooves (Figure 6b,c). Several tubes or slotted cores can be assembled together to give a cable with very high fiber density. The ribbon approach may be efficient when a large number of fibers must be

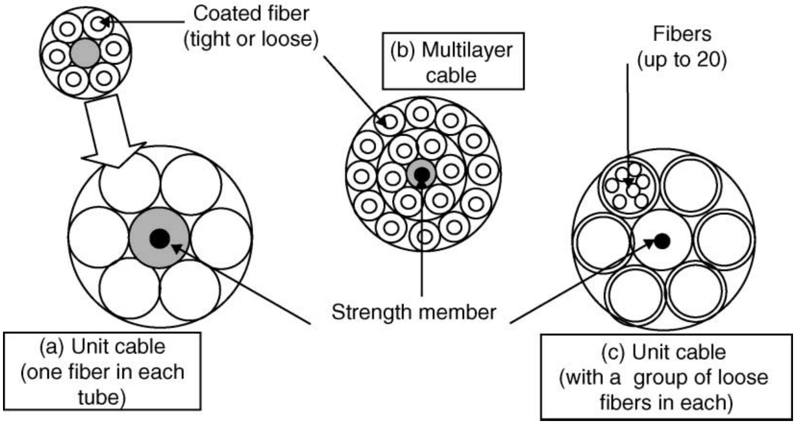

Figure 5 Structures of stranded core cables.

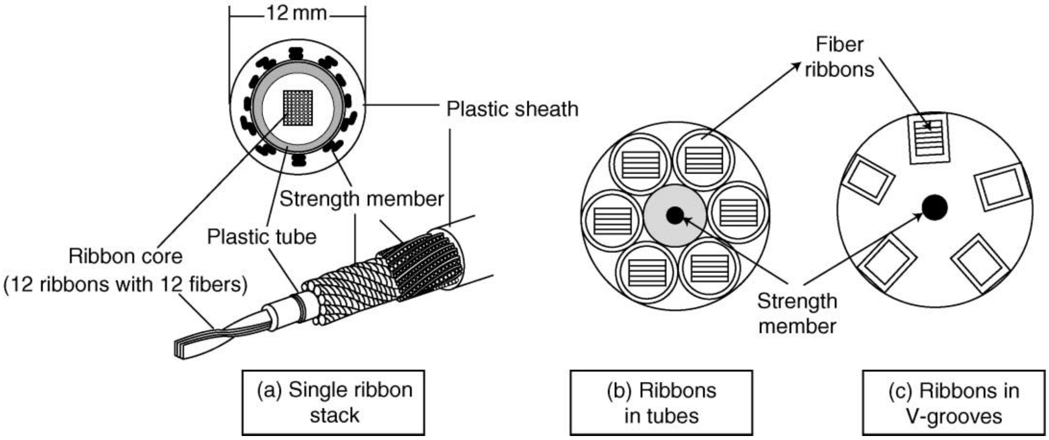

Figure 6 Structures of ribbon cables.

simultaneously handled, spliced, and terminated at connectors.

Types and Applications

The choice of an optical cable is heavily established by many factors, but mainly on the type of installation. On this basis, some different typology of cables may be identified: aerial cables, duct cables, direct buried cables, indoor cables, underwater cables, etc.

Aerial cables can be clasped onto a metal strand or can be self supporting. In the first case, the cable is not subjected to a particular tension but requires good thermal and mechanical performances. This solution is chosen when the cable must withstand strong ice and wind and when long distances between the poles are required. All the structures mentioned above may be adopted. In the second case, the cable is normally exposed to high mechanical stresses during its life and high tensile strength must be guaranted to maintain the fiber elongation at a safety level. Generally, loose structures are preferred for their greater fiber freedom. Often, in self-supporting cables, the strength member is external to the cable core. An example is shown in Figure 7.

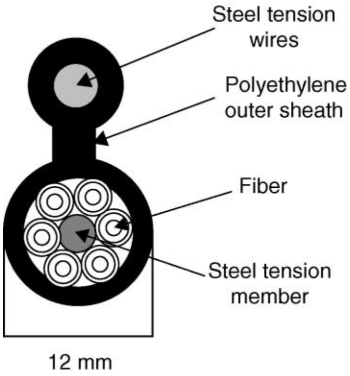

Figure 7 Self-supporting aerial cable.

Cables pulled in ducts must be resistant to pulling and torsion forces, and light and flexible to permit the installation of long sections. The cable must also protect the fibers against water and moisture that may be present in ducts or manholes. Usually the cable is filled with a jelly compound, a metallic barrier is normally used and, often, an armoring is added when protection against rodents is necessary. All the structures described above may be employed in duct cables.

The characteristics of direct buried cables are very similar to those of duct cables, but additional protections, such as metallic armoring, are required to avoid the risk of damages from digging and other earthmoving work.

Cables for indoor applications normally require a smaller number of fibers. Their main characteristics are: small dimensions, flexibility, small curvature radius, and ease in handling and jointing. For indoor applications, the sheath must not permit flame propagation and must have a low emission of toxic gases and dark smoke. Tight cables are preferably used for indoor applications due to their compactness and small dimensions.

Underwater cables are employed in the crossing of rivers, in lakes and lagoons, and for shallow water applications in general. Such cables must have stringent requirements such as resistance to moisture and water penetration and propagation, pulling resistance (during the installation, and recovery in event of failure), and high resistance to static pressure and core structure compatible with land cables (when it is part of a ground link). In addition, due to the presence of metallic structures, attention should be paid to hydrogen effects.

Submarine Cables

Submarine cable is manufactured for laying in deep-sea conditions. In designing a submarine cable, it is necessary to provide high reliability that the mechanical and transmission characteristics of the optical fibers will be stable over a long period of time. Different cables are used for submarine applications, but generally follow the basic scheme of Figure 8.

The central core (with a few fibers – up to 12 – in a slotted core or in a tight structure) is surrounded by a double layer of steel wires that act as strength members against tensile action and water pressure. Metal pipes at the inner or outer sides of the strength member, act as a water barrier and as power supply conductors (for the supply of the undersea regenerators).

The outer polyethylene insulating sheath is the ultimate protection for the ordinary deep-sea cable. For cables requiring special protection, steel wire armoring (anti-attack from fishes), and further polyethylene are added.

The cable must have stringent mechanical requirements, such as resistance to traction, torsion, crushing, impact, and to shark attack. The cable must be suitable for installation using standard cable-laying ships. Furthermore the cable materials must have low content of hydrogen and emissions. If necessary, a hydrogen absorber may be included in the cable core.

Special Cables

This category comprises cables not specifically used in ordinary telecommunication networks, but cables used for specific applications and according to specified requirements. Cables for military use are normally employed for temporary plant restoration. They must be light but crush resistant and with good mechanical characteristics. Cables for mobile applications (robots, elevators, aircrafts, submarines, etc.) generally require high flexibility and tensile strength. Special cables may be installed in particular environments, where they must be insensitive, for instance, to nuclear radiation, chemical agents, and high temperatures.

Optical cables can be installed in power lines, together with the power conductors, or into the ground wires of high-voltage overhead power lines. In this last type of cable, the central stranded steel wire, making up the ground wire, is replaced with a metal tube containing the optical unit (normally a small groups of optical fibers inserted in a tube or in a slotted core).

Blown Fiber

This technology consists of the installation of a single fiber or of small fiber bundles (or cables) into pre-installed tubes or ducts using a flux of compressed air. Initially, cables with empty tubes are installed and successively, at the time of need, the tubes can be filled by blowing fibers (singly, or in bundles) or cables. The installation time is normally fast; some compressors can blow tens of meters in a minute. The typical maximum blowing distance is about 1 km for horizontal planes, but it is reduced

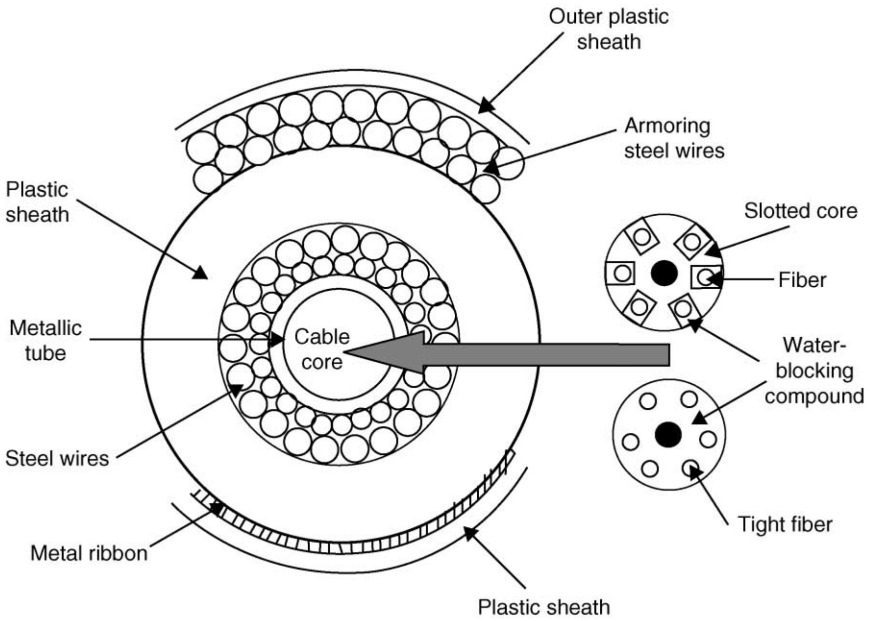

Figure 8 Basic structure of a submarine cable.

along winding roads. The advantage of this technique is to produce variable and flexible structures, increasing the number of fibers to fit the network need. Besides, substitutions or change of fibers can be quickly carried out without substitution of the cable.

See also

Fiber and Guided Wave Optics: Dispersion. **Light Emitting Diodes**.

Further Reading

Griffioen W (1989) The installation of conventional fiber-optic cables in conduits using the viscous flow of air. *Journal of Lightwave Technology* 7(2): 297–302.

Hughes H (1997) *Telecommunications Cables. Design, Manufacture and Installation.* : John Wiley and Sons.

Keiser G (1986) *Optical Fiber Communications*. McGraw-Hill.

Murata H (1988) *Handbook of Optical Fibers and Cables*. Marcel Dekker.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook Second Edition.* : TAB Books.

Passive Optical Components

D Suino, Telecom Italia Lab, Torino, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern telecommunications networks are based on fiber optics as signal transporters. In these networks the signals travel in the form of light confined within the fibers. As light power travels through the fibers from the source apparatus to the receiver, a number of components are employed to manage the optical signal. All other components that are not used to manage the signal are called ‘passive optical components’. They serve several functions, such as to join two different fibers to distribute the signal onto more optical branches to limit the optical power to select particular wavelengths, and so on.

These components can be located anywhere in the network, depending on their function and on the network architecture. However, their main location is near the apparatus in which it is necessary to manage the signal before it is sent through the network or at the junction in the network where it is sent on to the receiver.

The main types of passive components that are found in a network can be grouped as follows:

- Joint devices:
  - Optical connectors;
  - Mechanical splices.
- Branching devices:
  - Nonwavelength-selective branching devices;
  - WDM.
- Attenuators;
- Filters;
- Isolators;
- Circulators.

Optical fibers are also divided into several groups (multimode 50/125, multimode 62.5/125, singlemode, dispersion shifted, NZD, and so on). There are also components of different typology that match up with each particular fiber. However, the general concept in terms of functionality, technology, and characteristic parameters, are the same among passive optical components of the same type (connectors, branching devices, or attenuators) but of different groups (single-mode, multimode, etc.).

Optical Joint Devices

The function of an optical joint is to perform a junction between two fibers, giving optical continuity to the line. Two different classes of joint can be considered: the connectors and the splices. The joints made by using connectors, are flexible points in the network, which means they can be opened and reconnected several times in order, for example, to reconfigure the network plan. However, a splice is a fixed joint and usually cannot be opened and re-mated. There are two type of splice: mechanical splices, holding joints to fibers by glue or mechanical crimp, and fusion splices fusing the two fiber heads with a suitable fusion splicer.

General Considerations

Considering that fibers can be described as pipes, in the core of which flows light, continuity between two fibers means that the two cores must be aligned in a stable and accurate way. Obviously this alignment must be carried out with minimum power loss in the junction. The main factors that affect the power loss in a joint point between two fibers are:

- x = Lateral offset;
- z = Longitudinal offset;
- θ = Angular misalignment; and
- w_T/w_R = Mode field diameter ratio (in the single mode fiber) or numerical aperture ratio (in the multimode fiber).

These geometrical mismatching between the two fibers is represented in Figure 1.

The function that describes the joint attenuation is different for single-mode and multimode connectors, but the parameters that contribute to it are the same. The general formulation for these functions are complicated and, in particular, for the multimode connector it involves an integral on the shape distribution of the light in the fiber core. These functions become simpler in particular conditions: for example, if we consider a single-mode joint without a gap between the two fibers (that is the more usual condition in the modern connectors) the z parameter is nil, and in this case the function that describes the attenuation of the connector is

$$A(\mathrm{d}B) = -10\log\left\{\left(\frac{2w_T w_R}{w_T^2 + w_R^2}\right)^2 \times \exp\left[-2\frac{(x\lambda)^2 + (\pi n w_T w_R \sin(\vartheta))^2}{\lambda^2(w_T^2 + w_R^2)}\right]\right\} \qquad [1]$$

where n is the fiber core refractive index and λ is the wavelength of the light.

Figure 2 shows the trend of eqn [1], where the parameter ranges in the graphs are typical for the most common products on the market. It is evident that the main contribution to attenuation arises from the lateral offset.

For connectors of good quality, both for single-mode and multimode fibers, the attenuation values are in the range of a few tenths of dB (typically less than 0.3 dB). Another important factor for a joint, in particular for connectors used in networks for high speed or analog signals, is the return loss (RL). This is a parameter that gives a measure of the quantity of light back-reflected onto the optical discontinuity of the connector, due to the Fresnel effect. The RL is defined as the ratio in dB between the incident light power (P_{inc}) and the reflected light power (P_{ref}):

$$\mathrm{RL} = -10\log\left(\frac{P_{\mathrm{ref}}}{P_{\mathrm{inc}}}\right) \qquad [2]$$

Fresnel reflection arises when the light passes between two media with different refractive indices. In the case of connectors in which the two fibers are not in contact, the light passes from the silica of the fiber

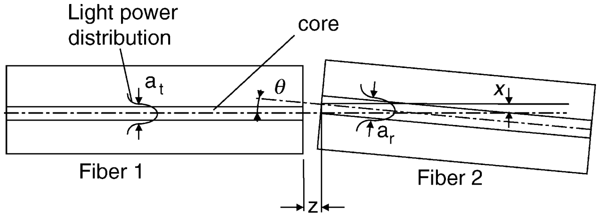

Figure 1 Main parameter that affects the joint power loss.

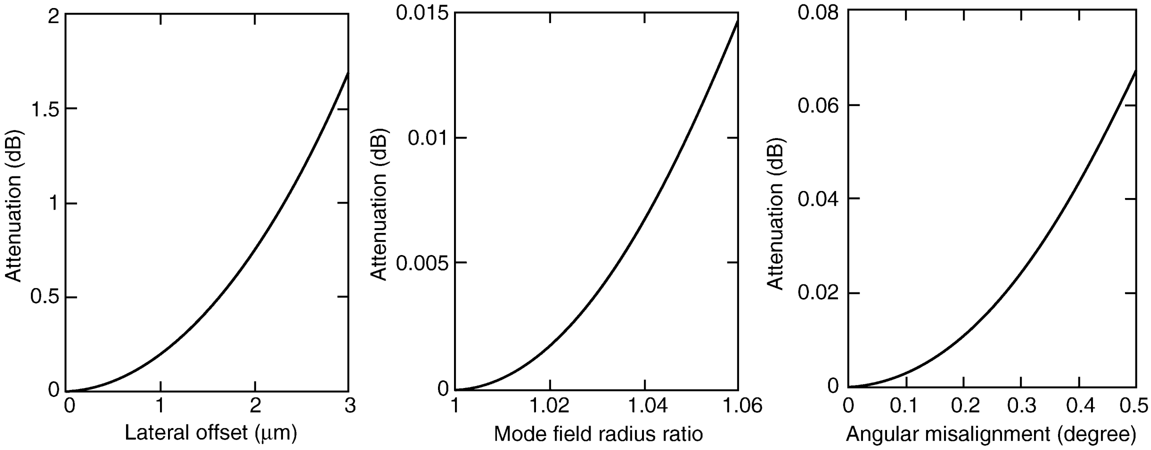

Figure 2 Single mode connector attenuation as a function of the geometrical parameters. Each curve is plotted keeping the other parameter constant.

core to the air in the gap. This causes the reflection of 4% of the incident light that, as RL, is about 14 dB.

To improve the RL performances of an optical joint, it is necessary to reduce the refractive index difference. This can be achieved either by index matching material between the two fibers that reduces the differences between the refractive index of the fiber and the gap, or by performing a physical contact between the fiber heads. The first technique is typically used in mechanical splices but, due to the problem related with the pollution contamination of the index matching, it is not a good solution for connectors that are to be opened and reconnected several times.

The physical contact solution is the most frequently adopted for these connectors. Physical contact is obtained by polishing the ferrule and the fiber end-faces in a convex shape, with the fiber core positioned in the apex (Figure 3). With this technique, the typical values of return loss for a single mode connector are in the range from 35 dB to 55 dB, depending on the polishing grade. The residual back-reflected light is generated in the thin layer of the fiber end surface, because of slight changes in the refractive index due to the polishing process.

Higher return loss values (small quantity of reflected light) are achieved with the angled physical contact (APC) technique. This is obtained by polishing the convex ferrule end face angled with respect to the fiber axis. In this way, the light is not back reflected into the fiber core, but into the cladding, and is thus eliminated (Figure 4). In APC connectors, typical return loss values are greater than 60 dB.

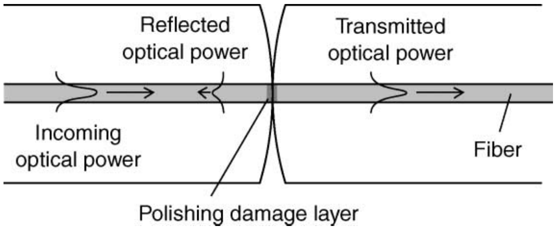

Figure 3 Back reflection effects in PC connectors.

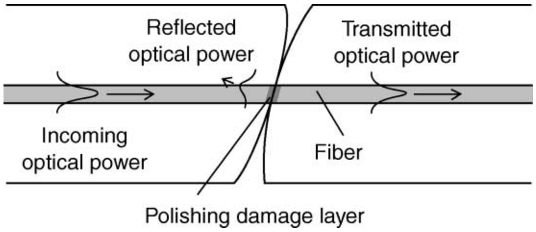

Figure 4 Back reflection effects in APC connectors.

Optical Connectors

On the market are several types of optical connectors. It is possible to divide them into three main groups:

- Connectors with direct alignment of bare fiber;
- Connectors with alignment by means of a ferrule; and
- Multifiber connectors.

Connectors with direct alignment of bare fiber

In this type of connector the bare fibers are aligned on an external structure that is usually a V-groove (Figure 5). Two configurations can be performed: plug and socket or plug–adapter–plug. In the first case, the fiber in the socket is fixed and the one in the plug aligns to it. In the plug–adapter–plug configuration, two identical plugs containing the bare fibers are locked onto an adapter in which the fibers are placed on the aligned structure. These connectors are cheaper than the connectors with a ferrule, (described below), but this is unreliable. In fact, the bare fibers are delicate and brittle and they can be affected by the external mechanical or environmental stresses.

Connectors with alignment by means ferrule

In this type of connector, the fibers are fixed in a secondary alignment structure, usually of cylindrical shape, called a ferrule. The fiber alignment is obtained by inserting the two ferrules containing the fibers into a sleeve. Usually this type of connector is in the plug–adapter–plug configuration (Figure 6). With this technique, the precision of the alignment is moved from the V-groove precision to the dimensional tolerance of the ferrule. Typical tolerance values on the parameters for a cylindrical ferrule with a diameter of 2.5 mm, are in the range of a tenth of a micrometer, and must be verified on the diameter, eccentricity between the hole (through which the fiber is fixed) and cylindricity of the external surface.

Multifiber connectors

In some types of optical cable, the fibers are organized in multiple structures, in which many fibers are glued together to form a ribbon. The ribbon can have

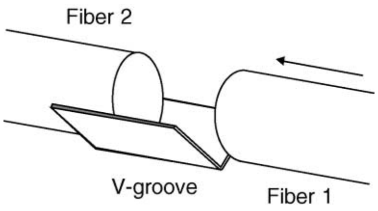

Figure 5 Scheme of the bare fibers alignment on a V-groove.

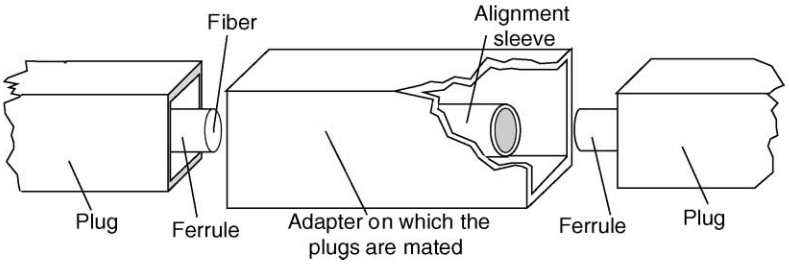

Figure 6 Scheme of a plug–adapter–plug connector with alignment by ferrules and sleeve.

2, 4, 8, 12, or 16 fibers, for example. When these cables are used, it is convenient to maintain the ribbon fiber structure in the connectors (until the point in which it is necessary to use the fibers singularly). Multifiber connectors are typically based on a secondary alignment on a plastic ferrule of rectangular shape in which the fibers maintain the same geometry that they have in the ribbon. The rectangular ferrules of the two plugs are normally aligned by means of two metallic pins placed in two hole in line with the fiber. In Figure 7, a scheme of a multifiber connector is shown.

Mechanical Splices

Mechanical splices are yet another way to join together two fibers. All the main considerations for the connectors are true for the mechanical splices. The difference is that the mechanical splice is a fixed joint and, normally, it cannot be opened and re-mated. Mechanical splices, both for single fiber and for multifiber structures, are the most usually manufactured. In a mechanical splice, the fibers are aligned directly in a V-groove and fixed by glue or a mechanical holder.

These components are cheaper than fusion splices but their optical, mechanical, and environmental functions are lower than the fusion splices. Figure 8, shows a mechanical splice for multifibers in which the fibers are crimped on a V-groove array. Usually, to avoid high optical reflection in the junction point, this type of joint is filled with index matching material.

Branching Devices

Branching devices are passive components devoted to distributing the optical power from an input fiber to two or more fibers. These components can be classified in two broad classes, on the basis of the wavelength dependence. The nonwavelength-selective branching devices, called also couplers or splitters, share the input power to two or more fibers independently from the light wavelength.

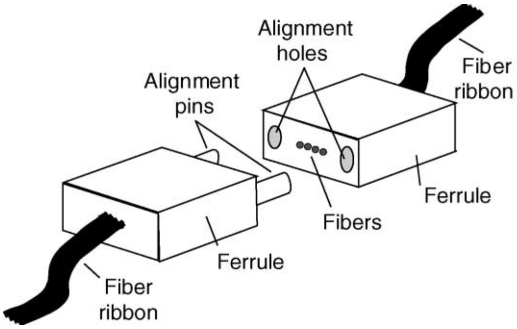

Figure 7 Multifiber connector.

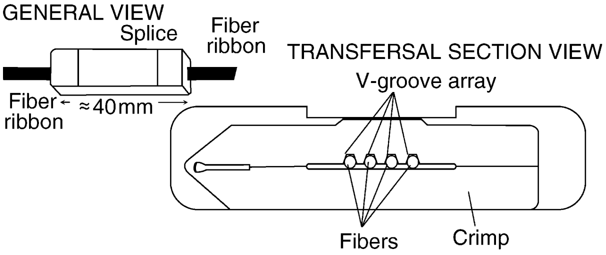

Figure 8 Example of a multifiber mechanical splice.

The components that perform the light distribution on more fibers on the basis of the different wavelengths of the input light are called a WDM (wavelength division (de)multiplexer).

Branching devices have three or more ports, bare fiber or connectors, for the input and/or output of optical power, and share optical power among these ports. A generic branching device can be represented as a multiport device having N input ports and M output ports, as shown in the Figure 9.

The optical characteristics of a branching device can be represented by a square transfer matrix of $n \times n$ coefficients, where n is the total number of the component ports. Each coefficient t_{ij} in the matrix is the fractional optical power transferred among designated ports, that is the ratio of the optical power P_{ij} transferred out of port j with respect to input

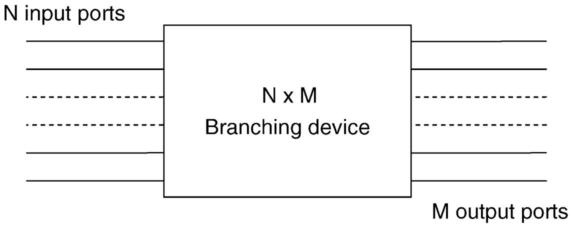

Figure 9 Scheme of a generic N × M branching device.

power P_i into port i:

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & & & \\ \vdots & & t_{ij} & \vdots \\ & & \ddots & \\ t_{n1} & & & t_{nn} \end{bmatrix} \qquad [3]$$

where:

$$t_{ij} = \frac{P_{ij}}{P_i} \qquad [4]$$

So, according to the relationship between two ports, the coefficients represent several parameters. For example, if the port i is an input port and j is an output port, t_{ij} is the transferred power signal through the component, while if they exist as two isolated ports the coefficients represent the crosstalk between the two ports.

The three main techniques to obtain this type of component are:

(i) All fiber components;
(ii) Micro-optics; and
(iii) Integrated optics.

Nonwavelength-Selective Branching Devices

A nonwavelength-selective branching device is a component having three or more ports, which shares an input signal among the output ports in a predetermined fashion. Usually these components are completely reversible and so they also work as a combiner of signals from more input ports onto a single output port.

There are several types of couplers for different applications:

- Y-coupler with one input port and two output ports;
- X-coupler with two input ports and two output ports; and
- Star coupler with more than two ports in input and/or in output.

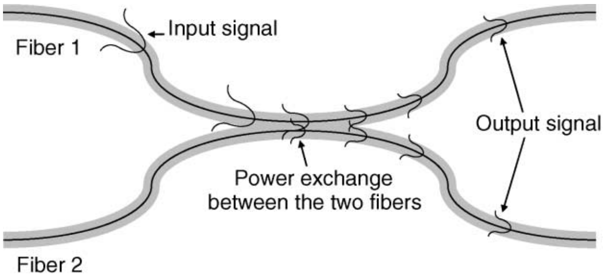

Figure 10 Scheme of a fusion 2 × 2 coupler.

Moreover the couplers can be symmetrical, and in this case they divide the input power light onto n equal output flows, or asymmetrical devices in which the quantity of power light in the output ports is shared in a predefined nonuniform way.

The techniques used to carry out this type of device are mainly the fusion technique or planar technique. The first one is based on the phenomenon of evanescent wave coupling: when the core of two optical guides (as fibers) are located in close proximity, a power exchange from a guide to another is possible. This configuration can be obtained by stretching the fibers under controlled fusion (see Figure 10).

The planar technique, that is the simpler integrated optic application, is based on the optical guides with appropriate shapes being placed directly onto a silicon wafer by means of lithographic processes (Figure 11).

The fusion technique has the advantage of being 'all in fiber' so the component is ready to be inserted into the optical networks by means of usual joint techniques. But for a high number of ports, the coupling ratio among the branches cannot be controlled in a precise way and the cost of the fusion device is proportional to the number of the branches. On the contrary, a device obtained with the planar technology must be coupled with fibers normally used and this operation may be difficult due to the different geometry of the planar guide (rectangular) and the fiber (circular). On the other hand, this technology allows a high precision in the optical characteristics of the component and the cost of the device is independent of the branch number.

WDM

A wavelength-selective branching device, usually called WDM (Wavelength Depending Multiplexer),

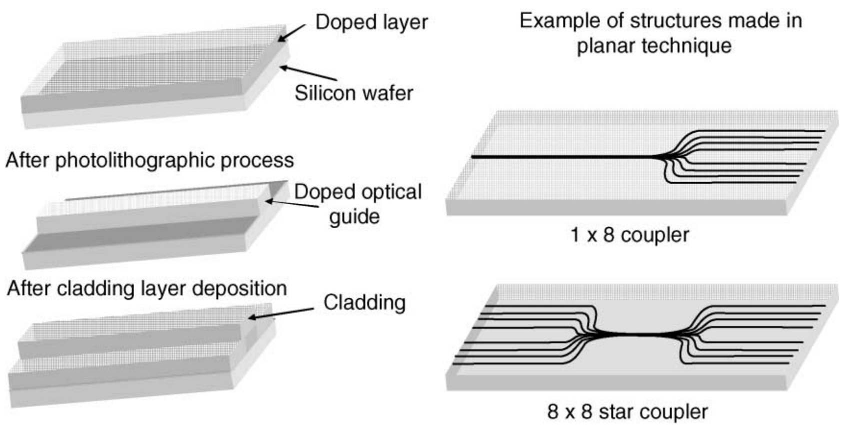
Doped layer
Silicon wafer
After photolithographic process
Doped optical guide
After cladding layer deposition
Cladding
Example of structures made in planar technique
1 x 8 coupler
8 x 8 star coupler

Figure 11 Scheme of the process to obtain a planar coupler and an example of a device.

is a component having three or more ports which shares an input signal among the output ports in a predetermined fashion, depending on the wavelength, so that at least two different wavelength ranges are nominally transferred between two different couples of ports.

WDM devices may be divided in three categories on the basis of the bandwidth of the channels (the spacing between two adjacent wavelength discriminated by the device):

(i) DWDM (Dense WDM) device operates for channel spacing equal or less than 1000 GHz (about 6–8 nm in the range of typical optical wavelength used in telecommunications systems, 1310 or 1550 nm);
(ii) CWDM (Coarse WDM) device operates for channel spacing less than 50 nm and greater than 1000 GHz; and
(iii) WWDM (Wide WDM) device operates for channel spacing equal or greater than 50 nm.

These components are used to combine, at the input side of an optical link in only one fiber, more optical signals with different wavelengths and separate them at the receiving end. This technology allows to send along a fiber, more communication channels, thus increasing the total bit rate transmitted. These types of components can be obtained using filters to select single a wavelength (as a diffractive grating that resolves the light into its monochromatic components) or using the phenomena of evanescent field coupling occurring between two adjacent fiber cores (as described for coupler devices), controlling in a precise way the coupling zone; this is, in fact, dependent on the wavelength (Figure 12).

Using wavelength filters, as they select a singular wavelength, it is necessary to perform a cascade filter structure for separating more than two wavelengths.

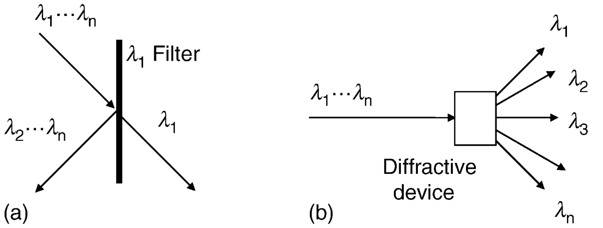
$\lambda_1 \cdots \lambda_n$
λ_1 Filter
$\lambda_2 \cdots \lambda_n$
λ_1
(a)
$\lambda_1 \cdots \lambda_n$
Diffractive device
λ_1
λ_2
λ_3
λ_n
(b)

Figure 12 (a) Scheme of the filter technique and (b) diffractive device technique.

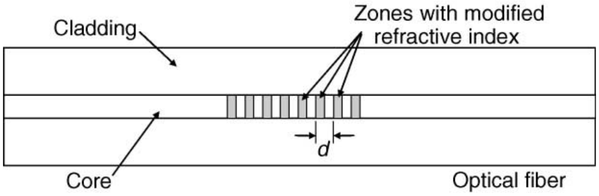
Cladding
Zones with modified refractive index
d
Core
Optical fiber

Figure 13 Scheme of FBG.

Micro-optic technologies are generally used for multimode fiber due to the critical collimation problem. For the single-mode fiber the evanescent field coupling technique is applied directly to fiber or to planar optical guides.

Another efficient technique to create WDM devices is with a fiber Bragg grating (FBG) wavelength filter directly applied to the fiber. FBGs consist in periodic modulation (see Figure 13) of the refractive index along the core of the fiber, and they act as reflection filters.

During the fabrication process (see Figure 14) the core of the optical fiber is exposed to UV radiation, and this exposure increases the refractive index of the core in defined zones, with a periodic shape obtained by the interference fringe created by crossing two coherent beams.

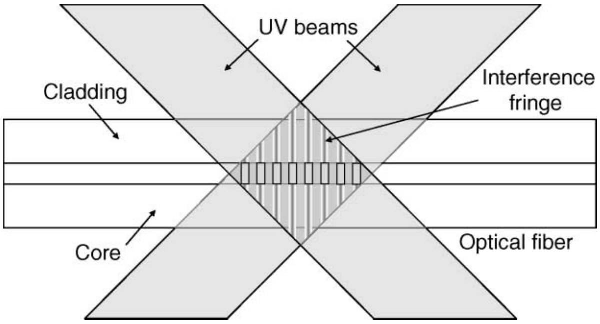

Figure 14 Scheme of FBG production.

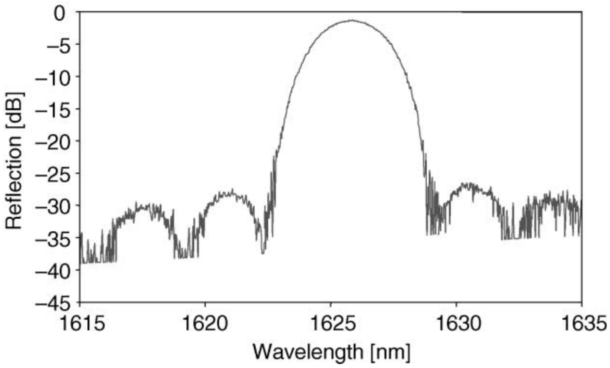

Figure 15 Transmission characteristic of FBG.

Figure 15 shows a representative spectral curve of a FBG filter centered at 1625 nm. The central wavelength, the width, and the value of reflection depend on the grating period (*d*) and on the deposited UV energy in the fiber core.

Other Passive Optical Components

In an optical network, besides the devices described above, there are a number of other passive optical components that perform particular functions on the optical signal; the main ones are:

Attenuators These are devices used to attenuate the optical power, for example, in front of a receiver when the optical line is short and the power of the received signal is still too high for the detector, or when it is necessary to equalize the signals arriving from different optical lines. It exists both in tunable and fixed versions. The first ones allow adjustment of the attenuation value, however, the second ones have a fixed and predefined value. The fixed optical attenuators can be obtained by inserting a filter or stressing a fiber in a controlled and permanent way. The tunable attenuators are based on variable optical parameters as, for example, the distance or the lateral offset in an optical connection. They exist both as connect lengths of fiber (attenuated optical jumpers) or as compact components similar to a double connector (plug stile attenuator).

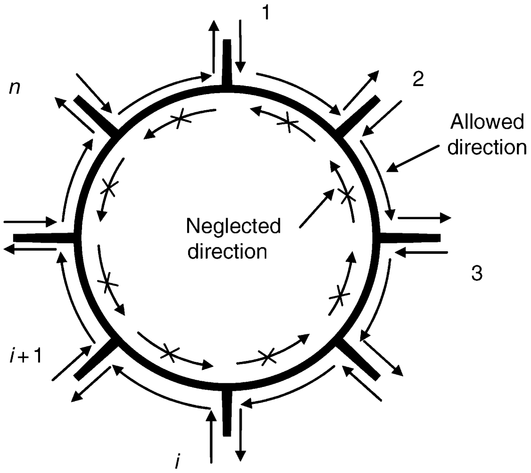

Figure 16 Principle scheme of an optical circulator.

Filters These are components with two ports employed to select or to filter some fixed wavelengths. They can be divided into the following categories:

- short-wave pass (only wavelengths lower than or equal to a specified value are passed);
- long-wave pass (only wavelengths greater than or equal to a specified value are passed);
- band-pass (only an optical window is allowed); and
- notch (only an optical window is inhibited).

It is also possible to have a combination of the above categories. The optical filtering can be obtained by inserting one or more filtering devices (as interferential films) into a fiber or creating a filter directly on the fiber by the FBG described above.

Isolators These are nonreciprocal two-port optical device in which the light flows in the forward direction, and not in the reverse one. The isolators are used to suppress backward reflections along an optical fiber transmission line, while having minimum insertion loss in the forward direction. Fiber optic isolators are commonly used to avoid back reflections into laser diodes and optical amplifiers, which can make the laser and amplifiers oscillate unstably, and cause noise in the fiber optic transmission system.

Circulators These are passive components having three or more ports numbered sequentially $(1, 2, \ldots, n)$ through which optical power is transmitted nonreciprocally from port 1 to port 2,…, from port (i) to port $(i+1), \ldots$ and from port n to port 1 (Figure 16).

See also

Diffraction: Fresnel Diffraction. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Adam MJ (1981) *An Introduction to Optical Waveguides*. New York: John Wiley & Sons.

Blonder GE, *et al.* (1989) Glass waveguides on silicon for hybrid optical packaging. *IEEE Journal of Lightwave Technology* 7: 1530.

Digonnet M and Shaw HJ (1983) Wavelength multiplexing in single-mode fiber couplers. *Applied Optics* 22: 484.

Jeunhomme LB (1983) *Single-Mode Fiber Optics – Principles and Applications*. New York: Marcel Dekker.

Kashima N (1995) *Passive Optical Components for Optical Fiber Transmission*. Norwood, MA: Artech House.

Kashima N (1993) *Optical Transmission for the Subscriber Loop*. Norwood, MA: Artech House.

Miller CM (1986) *Optical Fiber Splices and Connectors – Theory and Methods*. New York: Marcel Dekker.

SC86B IEC (1999) *IEC 60874-1: Connectors for Optical Fibres and Cables – Part 1: Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 61073-1: Optical Fibres and Cables – Mechanical Splices and Fusion Splice Protectors – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (1999) *IEC 60869-1: Fibre Optic Attenuators – Part 1: Generic Specification*, 3rd edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 60875-1: Non-wavelength Selective Fibre Optic Branching Devices – Part 1:Generic Specification*, 4th edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 62077-1: Fibre Optic Circulators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61977-1 Ed. 1.0: Fibre Optic Filters – Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

SC86B IEC (2000) *IEC 61202-1: Fibre Optic Isolators – Part 1: Generic Specification*, 1st edn. Geneva: International Electrotechnical Commission.

Technical Staff of CSELT (1990) *Fiber Optic Communications Handbook*, 2nd edn, TAB Books. New York: McGraw-Hill.

Yokohama I, *et al.* (1988) Novel mass-fabrication of fiber couplers using arrayed fiber ribbons. *Electronic Letters* 24: 1147.

FIBER GRATINGS

P S Westbrook, OFS Laboratories, Murray Hill, NJ, USA

B J Eggleton, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Fiber gratings are periodic index variations inscribed into the core of an optical fiber, and are important devices for manipulating fiber guided light. Since their first fabrication in the late 1970s, fiber grating use and manufacture has increased significantly and they are now employed commercially as optical filters, reflectors and taps in telecommunications systems, and as strain and temperature sensors in various industries. We review optical fibers and the fundamentals of fiber grating characteristics as well as fabrication. Finally we consider several of the major industrial and scientific applications of fiber gratings.

Fiber Modes

A fiber grating couples light between the modes of an optical fiber, and therefore a knowledge of these modes is essential to understanding the characteristics of fiber gratings. In what follows, we emphasize single-mode fibers used near 1550 nm because these are very important in most practical applications that use fiber gratings. A single-mode fiber can be viewed as a thin cylinder of silica with a central raised index region of small enough diameter so that only a single mode is guided through total internal reflection at the core boundary. The electric field (E-field) of the fundamental mode propagating in the core is described by a characteristic transverse E-field profile and a propagation constant k:

$$\begin{aligned} E &= E(r, \varphi)\mathrm{e}^{-\mathrm{i}(\omega t - kz)} \\ k &= 2\pi n_{\mathrm{eff}}/\lambda \end{aligned} \qquad [1]$$

Here r and φ are the radial and azimuthal directions, z is measured longitudinally along the fiber, ω is the radial frequency, and λ is the vacuum wavelength, and the E-field points in the same direction everywhere. As the second equation indicates, the fiber mode propagates in the fiber as though it has an effective refractive index n_{eff}.

The silica cladding acts to isolate the core mode from the outside, hence the core mode penetrates very little into the cladding. However, if the outer cladding surface allows total internal reflection, the cladding may also guide light. Such 'cladding modes' are also important in understanding fiber gratings, since the grating couples the core mode with the cladding modes. As shown in Figure 1, the cladding modes extend outside the core into the cladding region, and their effective index is therefore always less than that of the core mode. We also note that in the case when the outer coating of the fiber does not support total internal reflection, 'leaky' cladding modes result. In the limit that there is no reflection at the cladding boundary a continuum of 'radiation modes' results. Although these modes are not guided, they are exploited in a variety of applications and can be carefully tailored by a suitably designed fiber grating. Note that the radiation and cladding modes exist regardless of the presence of a grating; a grating simply couples power from the core mode to the cladding/radiation modes; the grating is typically only a small perturbation on the index profile of the waveguide and the corresponding mode fields.

Fiber Grating Theory

Fiber gratings are longitudinal periodic variations in the refractive index (or, more generally, the dielectric function) of the core and/or cladding of an optical fiber. The scattering from any grating may be understood simply by considering the scattering off each successive period and adding these contributions while taking the phase of the E-field into account. Figure 2 illustrates this analysis in the case of a fiber Bragg reflector. When the reflections from each period of the grating add constructively, a strong back-reflection results. This is known as Bragg reflection. The condition for Bragg reflection is given by:

$$\lambda_{\mathrm{Bragg}} 2 n_{\mathrm{eff}} = \Lambda_{\mathrm{grating}} \qquad [2]$$

Here λ_{Bragg} is the vacuum wavelength of the light (equivalent to its frequency, $\omega = 2\pi c/\lambda$), n_{eff} is the effective index of the mode, and $\Lambda_{\mathrm{grating}}$ is the period of the grating refractive index modulations. This Bragg reflection is evident in a narrow dip in the

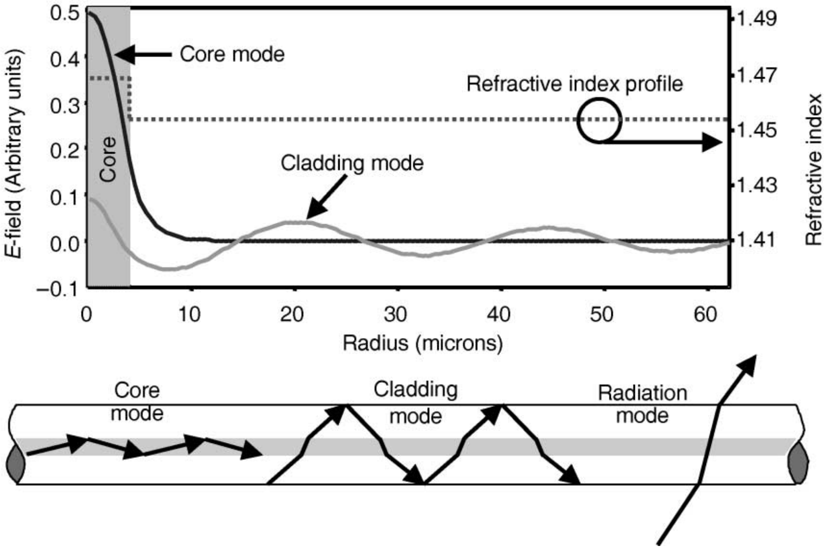

Figure 1 Fiber modes. Refractive index profile of a step index, single-mode fiber (dashed line). Radial profiles of mode E-fields of the fundamental (LP01) core mode, guided within the raised index region defined by the core, and one of the many higher-order cladding modes guided by the outer, air–silica interface. Also shown is a fiber indicating the ray paths for core, cladding, and radiation modes.

transmission spectrum of the grating (see Figure 3a). The ability of a fiber grating to selectively reflect the core light at one wavelength is one of its most important practical properties since it can be used as a narrowband filter or reflector. As a fiber component, the fiber grating has the advantage of compactness and provides in-line wavelength filtering.

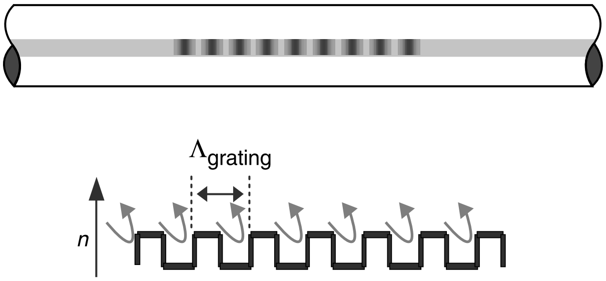

Figure 2 Bragg reflection. Each period of the grating reflects a small fraction of the incident light. When the period of the grating is adjusted so that these reflections add coherently, a strong Bragg reflection results and a narrow dip is observed in the transmission spectrum of the fiber (see Figure 3a). As shown in Figures 3 and 4, Bragg reflection can also occur into cladding modes.

The condition for Bragg reflection may be expressed more generally as a phase-matching condition. Phase matching refers to the condition when the phases of all the grating-scattered waves are the same and there is constructive interference. More specifically, this condition expresses the relationship between the spatial frequencies of the incident light, the scattered light, and the grating modulation, necessary for constructive interference in the scattered light. For a grating, the phase matching condition may be expressed in terms of the grating wavevector, $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}}$ and of the two mode propagation constants, k_{incident} and $k_{\text{scattered}}$ (defined in eqn [1])

$$k_{\text{incident}} - K_{\text{grating}} = k_{\text{scattered}} \qquad [3]$$

Using phase matching, we can understand the various transmission spectra that result from fiber

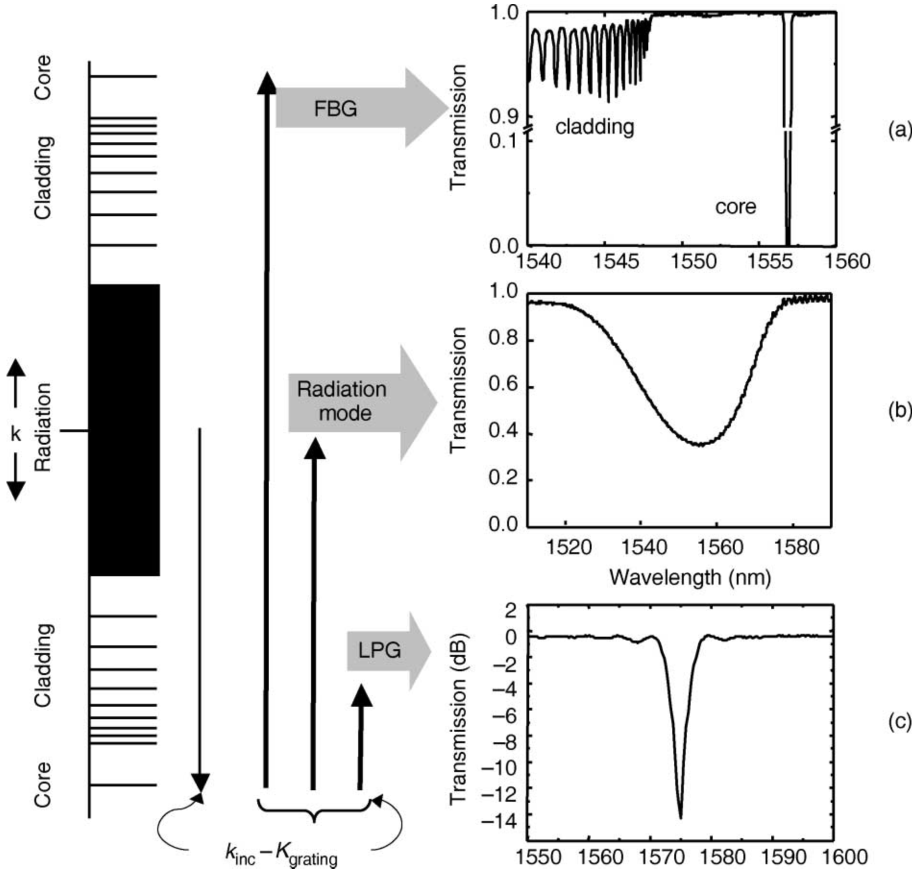

Figure 3 Phase matching. Fiber grating resonances occur when the phase matching condition is satisfied: $k_{\text{inc}} - K_{\text{grating}} = k_{\text{scat}}$. Propagation constants ($k = 2\pi n_{\text{eff}}/\lambda$) of the fiber modes are indicated as hash marks on the vertical line. The region indicates the regime of unguided radiation modes. The fiber grating is represented by its longitudinal spatial frequency (wavevector) $K_{\text{grating}} = 2\pi/\Lambda_{\text{grating}}$, where Λ_{grating} is the period of the grating. Assuming an incident forward-going core mode, the grating wavevector then phase matches to (a) counter-propagating core and cladding modes (fiber Bragg grating, FBG), (b) continuum of radiation modes, or (c) a copropagating cladding mode (long-period grating, LPG).

gratings. Figure 3 shows the transmission characteristic observed for core guided light in a single-mode fiber containing three different types of gratings: Bragg reflector, radiation mode coupler, and copropagating mode coupler. Each grating has a different type of phase matching condition. Figure 3 also depicts the phase matching condition in eqn [3] graphically. The vertical line contains a mark representing the propagation constants, k, of modes of a single mode fiber (both positive and negative for the two propagation directions) and a solid black region representing the radiation mode continuum. Bragg reflection into backward-propagating core and cladding modes occurs for short grating periods ($K_{\text{grating}} \sim 2k_{\text{core}}$). Such gratings are known as fiber Bragg gratings (FBGs) and a typical transmission spectrum with a strong Bragg reflection and several cladding mode resonances is shown in Figure 3a (a typical Bragg reflection spectrum is shown in Figure 8 below; the FBG grating period was ~500 nm). Core-cladding mode reflection results in loss resonances for wavelengths shorter than the core–core Bragg reflection. When the grating period is very long (in general, $K_{\text{grating}} < k_{\text{core}}(n_{\text{cladding}} - 1)/n_{\text{cladding}}$; typically Λ_{grating} is ~1000 times that of an FBG), phase matching to copropagating cladding modes occurs. Such gratings are known as long-period gratings (LPGs). A typical LPG spectrum is shown in Figure 3c. The LPG resonance corresponds to coupling between the core mode and a single copropagating cladding mode. For wavevectors in between these two regimes, coupling to the continuum of radiation modes is possible. An example of such a grating spectrum is shown in Figure 3b. We note that fiber cladding modes can be either well guided (resulting in the sharp grating resonances of Figure 3a) or poorly guided (resulting in a broad continuum of Figure 3b) depending on the fiber surface. The radiation mode approximation is valid in the limit that the cladding mode is very poorly guided (i.e., has high loss) within the grating region.

While phase matching yields the resonance wavelengths of a fiber grating, the strengths of these resonances must be derived from a full solution of Maxwell's equations. The most common approximation used in these calculations is known as coupled mode theory (CMT). In CMT, the longitudinal refractive index variation is assumed to be a small perturbation to the waveguide modes, which results only in a change in the propagation constant of the modes but does not change the transverse dependence of the mode E-fields (see eqn [1]). CMT then assumes that the E-field amplitude of each mode varies slowly (compared to wavelength) along the grating. A pair of coupled equations relating the two mode amplitudes as a function of position along the grating can then be derived, and these allow for computation of the grating reflection and transmission. The coupling between two modes is determined by an overlap integral κ_{12} between the product of the E-fields of the two modes (E_1 and E_2) and the periodic refractive index variation:

$$\kappa_{12} \propto \int\!\!\int_{\text{Area}} E_1(r,\phi)E_2(r,\phi)\delta n(r,\phi)\,\mathrm{d}A \qquad [4]$$

where $\delta n(r,\phi)$ is derived from the refractive index modulation:

$$\delta n(r,\phi,z) = \delta n(r,\phi)\cos(K_{\text{grating}}z) \qquad [5]$$

This overlap integral is useful because it allows for a rough comparison of the strength of resonances for different modes without the need to obtain a full CMT solution. In the simple case of Bragg reflection of the core mode in a standard single-mode fiber (see Figure 1):

$$\kappa = \frac{\pi \delta n \eta}{\lambda}, \qquad [6]$$

where $\eta \approx 0.8$ is the overlap integral of the core mode intensity with the core.

To understand fiber grating spectra we first discuss a 'uniform' Bragg grating, of length L. Such a grating has uniform index modulation along its length and is shown in Figure 4a. Depending on the length and refractive index modulation of the grating, it may be classified as either weak or strong. The division between these two regimes may be understood by considering the light scattered by each successive period of the grating. The fraction of light scattered from each period can be approximated by $\delta n/2n$ using the Snell's law reflectance formula. The boundary between a strong and a weak grating occurs when the grating is either long enough or strong enough that the sum of the scattering from all periods is of order 1. This division may also be expressed in terms of κ and L: $\kappa L \sim 1$.

In the weak limit, the incident light propagates through the grating with very little scattering. In this case, both the LPG and FBG resonances have transmission spectra that are approximately related to the Fourier transform of the grating index profile. (For LPGs, this may not hold if modal dispersion is large; see the discussion below.) In the case of a uniform grating profile, this results in a $1 - \text{sinc}^2$ wavelength dependence of the transmission spectrum and a spectral width $\delta\lambda/\lambda \sim \Lambda_{\text{grating}}/L = 1/N_{\text{periods}}$ as shown in Figure 4a. An important feature of the uniform grating is the side lobes. The side lobes result from the sudden 'turn on' of the grating. In the FBG,

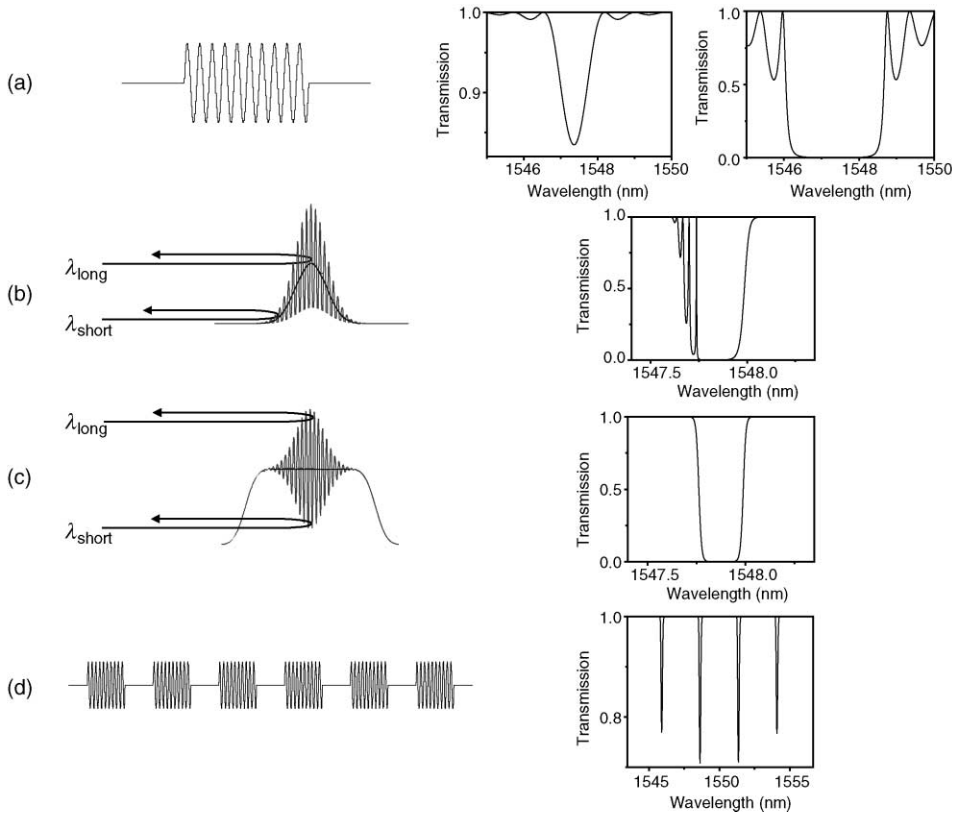

Figure 4 Fiber Bragg grating refractive index profiles and spectra. (a) A uniform profile with transmission spectra for both strong and weak index modulations. (b) An 'unapodized' Gaussian grating profile showing Fabry–Perot-like structure within the grating resonance. (c) A '100% dc apodized' Gaussian grating profile showing a symmetric transmission spectrum. (d) A superstructure grating exhibiting four resonance peaks.

they can be understood as resonances of the effective Fabry–Perot cavity defined by the sharp grating boundaries.

In the strong limit, a substantial fraction of light is scattered out of the incident mode, and the FBG and LPG transmission spectra become different. As shown in Figure 4a, the FBG spectrum widens to a width determined only by δn: $\delta\lambda/\lambda = \delta n/n$. The incident light penetrates only a short distance ($\sim \lambda/\delta n$) into the grating before being completely reflected. The strong LPG, on the other hand, becomes 'overcoupled' as the second mode is reconverted into the incident mode. The result is a higher transmission for the incident core mode (provided that the scattered mode propagates without loss).

While uniform gratings are easy to model and manufacture, most practical fiber gratings have a nonuniform profile. In a nonuniform grating profile, both the refractive index modulation amplitude, $\delta n_{\mathrm{AC}}(z)$, and the average effective index value, $\delta n_{\mathrm{DC}}(z)$, may vary along the fiber grating length:

$$\delta n(z) = \delta n_{\mathrm{DC}}(z) + \delta n_{\mathrm{AC}}(z)\cos(K_{\mathrm{grating}} z) \qquad [7]$$

By controlling the ac and dc index change, a greatly improved filter may be fabricated. Figure 4b shows the index profile and the grating transmission spectrum of a fiber Bragg grating with an 'unapodized' Gaussian profile. Note that due to the smooth increase in modulation amplitude, the sidebands of the uniform grating have been eliminated. However, because the dc index (and hence Bragg condition) is not constant within the grating, sharp resonant structure is observed within the grating resonance. This structure can be understood as Fabry–Perot resonances in the cavity formed by the nonuniform Bragg condition within the grating. Figure 4c shows a grating with a '100% dc apodized' Gaussian profile. This grating has a constant dc refractive index within the grating and therefore the resonance is smooth and symmetric. Most useful bandpass filters require a profile close to 100% apodization. Although not shown in Figure 4, a smoothly varying index profile also eliminates the sidelobes in LPG resonances.

Other important types of nonuniform fiber gratings include chirped fiber gratings and superstructure fiber gratings. In a chirped fiber grating the grating period slowly changes along the length of the grating. As discussed below, chirped fiber gratings may be used as chromatic dispersion compensators. They may also be used as broadband reflectors. LPGs may also be chirped to produce a broadband response. Superstructure gratings have oscillations in the period and/or amplitude of the refractive index modulations. The resulting grating can have several resonance peaks (see Figure 4d). The properties of nonuniform gratings may be computed using CMT and also understood intuitively using 'band-diagrams' and an effective index model of the fiber grating response.

Grating spectra showing radiation mode coupling are unlike FBGs or LPGs because there is no conversion of the radiation modes back to the incident core mode. The sharp resonances smooth out into a broad continuum. The grating transmission spectrum depends less on the refractive index profile of the grating than on the overlap of the core mode and grating with the radiation modes. Radiation mode coupling is greatly affected by tilting the grating planes with respect to the axis of the fiber. As discussed below, such tilted gratings may be designed as loss filters and optical fiber taps. Grating tilt has also been applied to decrease core mode back-reflection.

LPG resonances (and to some degree FBGs) may also exhibit strong dependence on the modal dispersion (or variation of the propagation constant, k, with wavelength). This wavelength dependence can dominate the wavelength dependence arising from CMT, making the resonance more or less narrow than predicted by CMT without modal dispersion. In an extreme case, very broadband LPG resonances (40 nm) may be produced between the core mode and a specially designed higher-order mode whose propagation constant varies such that it is resonant with the core mode over a large bandwidth.

Fiber Grating Fabrication

Photosensitivity

The key discovery leading to the study and practical application of fiber gratings was the phenomenon of UV-induced photosensitivity in the conventional germanosilicate optical fibers used in telecommunications systems. UV irradiation of the germanosilicate core region can produce refractive index changes in the range 0.01 to 0.0001. Such refractive index changes are large enough to produce useful filters and reflectors. The refractive index change is most efficiently produced by irradiation close to 242 nm or around 193 nm, where germanosilicate glass has strong absorptions. Numerous other dopants have been shown to be photosensitive at wavelengths ranging from 484 nm to 155 nm. These include P, Ce, Pb, Sn, N, Sn-Ge, Sb, and Ge-B. Other dopants, such as Al and F, show very little photosensitivity.

A second major discovery enabling practical manufacture of fiber gratings was the effect of 'hydrogen loading'. The fiber is placed in a pressurized hydrogen atmosphere until molecular hydrogen has diffused into the core region to a level of a few mole percent (a level similar to that of the photosensitive dopants). UV irradiation of the resulting fiber produces index changes an order of magnitude larger than in the unloaded fiber. Hydrogen loading allows gratings to be easily imprinted in conventional single-mode fibers, whose Ge content is too low to allow for strong gratings. The increase in photosensitivity is true for most photosensitive dopants. Index changes in excess of 10^{-2} have been produced with hydrogen loading. In order to avoid increased OH absorption in the telecommunications band (1520–1630 nm), deuterium (D_2) is normally used in place of hydrogen.

The origins of UV photosensitivity are not fully understood; however, at least two mechanisms are agreed to contribute to the index change. Firstly, as is evident from Figure 5, the UV absorption spectrum is modified during UV exposure. This bleaching corresponds to a change in refractive index in the infrared which may be computed using the Kramers–Kronig relations. In particular, the absorption band at

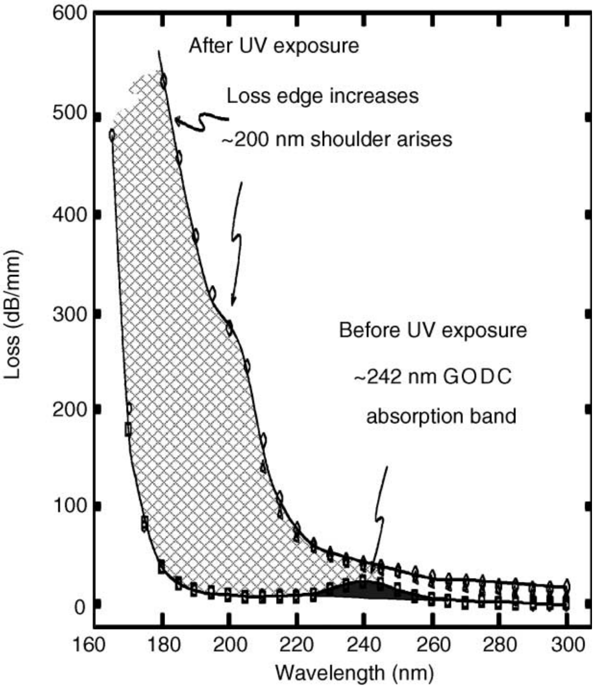

Figure 5 UV absorption spectrum of Ge-doped silica before and after irradiation with 242 nm light, showing bleaching of the Ge-oxygen deficient defect feature at 242 nm.

242 nm resulting from germanium oxygen deficient centers is bleached, indicating that these defects are modified during exposure. The infrared index change results both from this bleaching and from the change in the UV absorption edge at ~190 nm. Secondly, absorption of UV causes the silica matrix to contract. This compaction and the resulting stress distribution also change the refractive index.

Both mechanisms are enhanced by the presence of hydrogen in the silica lattice. The hydrogen reacts with oxygen, thus increasing the number of UV-absorbing germanium oxygen-deficient centers. As a result, in hydrogen loaded fibers, more of the Ge atoms participate in the index change, explaining the large increase in photosensitivity. The number of oxygen-deficient defects may also be increased during fiber manufacture by collapsing the fiber preform in a low-oxygen atmosphere. Such fibers show increased photosensitivity without the need for hydrogen loading. Co-doping a Ge-doped fiber with boron has also been shown to increase the level of photosensitivity of a fiber for a given level of index change in the core.

An important aspect of UV-induced index changes is that they require thermal annealing to ensure long-term stability. Sufficient exposure to elevated temperatures can completely erase the refractive index change in some Ge-doped gratings. Gratings can be completely erased by several hours of heating to greater than 500 °C. Thermal stability has been described by an activation energy model which assumes a distribution of energies for the defects giving rise to the refractive index change. Therefore, at a given anneal temperature a fixed fraction of the index change will quickly decay, leaving only stable defects that easily survive at lower temperatures. By annealing at several temperatures for a fixed time, a relationship between temperature and time required for a given refractive index decay may then be derived and used in accelerated aging tests of fiber gratings. For 20-year grating reliability of ~1% decrease in UV-induced refractive index change, a typical Ge-doped grating requires annealing at 140 °C for 12 hours.

Grating Inscription

The first method used in UV grating inscription is known as 'internal writing'. In this method light propagating in the fiber causes a standing wave to form within the core, and UV photosensitivity from a two-photon process results in permanent grating formation. Both LPG and FBG resonances have been fabricated using this method. The technique is inflexible, however, because the grating parameters are not easily controlled.

Another important advance that made possible the widespread use of fiber gratings is the 'transverse side-writing' technique. In this method, a UV interference pattern is projected through the side of the fiber onto the core region. The first demonstration of this technique employed a free space interferometer to produce the UV fringes. A significant improvement on this technique employs a zero-order nulled phase mask to generate the interference pattern. The phase mask acts to split the incident beam, and these two then form the interference pattern. The method is shown in Figure 6. The advantage of the phase mask

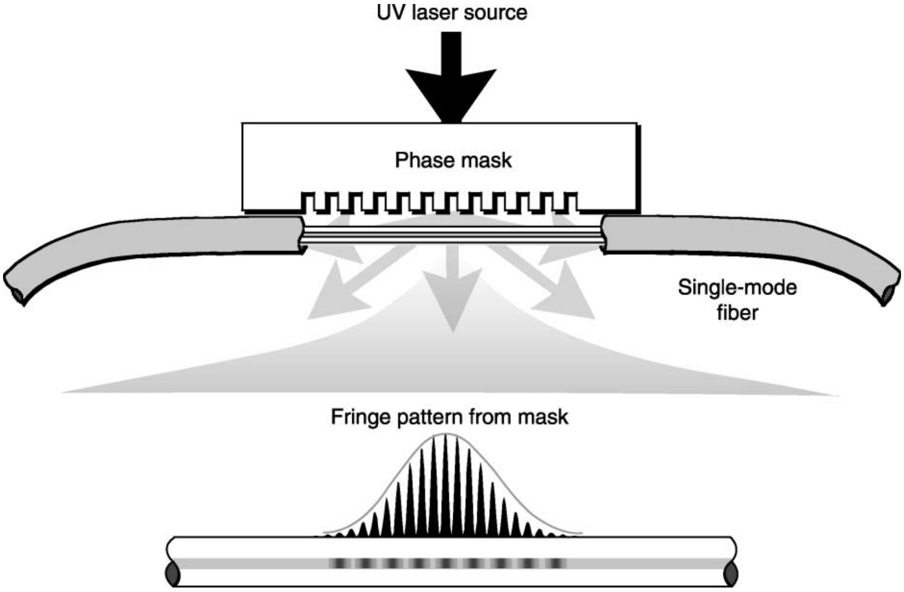

Figure 6 Phase mask fabrication of a fiber grating.

technique is that the fringes are very stable, since the UV beams propagate only a short distance. Another important technique for writing fiber gratings is the ‘point-by point’ writing method. In a point-by-point technique the UV interference pattern is modulated and translated along a fiber in such a way that a long grating may be formed. Although complex and sensitive, such methods can yield gratings as long as a few meters and also allow for the fabrication of complex grating profiles. Point-by-point methods have been used to fabricate broadband chirped dispersion compensating gratings and Bragg reflectors with very low dispersion.

LPGs may also be fabricated with UV irradiation. An amplitude mask is typically used. Requirements for beam coherence are greatly reduced because of the long period (typically a few hundred microns). Other techniques may also be used to fabricate LPGs. The fiber may be periodically deformed by heating or with periodic microbends. Dynamic LPGs may also be produced by propagating acoustic waves along the fiber.

Applications

Fiber gratings are key enabling technologies in a number of important applications. In this section we review these, highlighting the unique functionalities that can be provided by fiber gratings.

Fiber Grating Strain and Temperature Sensitivity

Fiber gratings are sensitive to both strain and temperature and may therefore be used as sensors. Both LPG and FBG resonances change with strain and temperature. LPGs show greater but more complicated sensitivity and typically FBGs are used in such applications. Strain sensitivity arises from the fact that gratings are extended in length. Because fibers are very thin (diameter ~125 μm), significant extension is possible. A smaller strain sensitivity arises from the stress optic effect. Typical sensitivity at 1550 nm is ~1 pm/1 $\mu\varepsilon$ in FBGs. Temperature sensitivity arises from temperature dependence of the refractive index and gives rise to a change order 0.011 nm/°C of temperature change of the FBG. FBG strain sensors have been applied to map stress distributions and can be employed in a distributed manner with many Bragg reflectors in one fiber being examined with one detector.

These sensitivities may also be used to make fiber grating filters tunable. Temperature changes are more stable but typically give 1–2 nm of change in an FBG resonance. Strain tuning is more difficult to package commercially but can give up to 4 nm in extension and more than 10 nm in compression. Passive thermal stabilization of grating resonances may also be achieved by designing packages in which strain and temperature variations cancel each other.

Fiber Grating Stabilized Diode Lasers and Fiber Lasers

One of the first important commercial applications of fiber gratings was in stabilization of fiber coupled diode lasers. Normally these lasers operate on several cavity modes making the output unstable. If feedback is provided in the form of a narrowband reflection then the laser output will be only in this narrow bandwidth. The fiber grating is implemented in the fiber pigtail of the diode laser and typically provides a few percent reflectivity over a bandwidth of ~1 nm. The application is depicted schematically in Figure 7.

More generally, fiber gratings have been used to realize fiber laser cavities and fiber distributed feedback lasers. One important example is the cascaded Raman resonator fiber laser, in which several FBGs recycle Raman pump light to produce high-power, fiber coupled light at any design wavelength.

Optical Add/Drop Multiplexers

Another important commercial FBG application is the optical add/drop multiplexer (see Figure 8). These are key components in wavelength division multiplexed (WDM) networks. WDM networks transmit large amounts of data through a single fiber by modulating several optical carrier wavelengths and sending all the wavelengths through a single fiber. The optical add/drop multiplexer allows individual wavelengths to be added or dropped from a given fiber.

Fiber Bragg gratings are useful in this application because of the very narrow resonances that are achievable (typically a few tenths of a nm at 1550 nm). For wavelengths close to the Bragg wavelength an FBG strongly reflects. The reflectivity and bandwidth are determined by the depth of the index modulation and grating length, as discussed earlier. These degrees of freedom make fiber gratings ideal filters to be used in wavelength division multiplexed lightwave networks, in which multiple channels are transmitted along long lengths of optical fiber. Figure 8 illustrates schematically the principle of

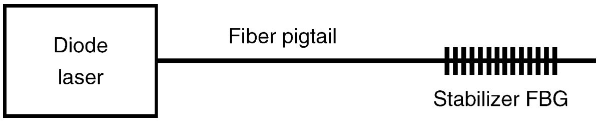

Figure 7 Fiber Bragg grating stabilized diode laser. The fiber grating reflector in the laser pigtail provides a small level of feedback to stabilize the laser output power and wavelength.

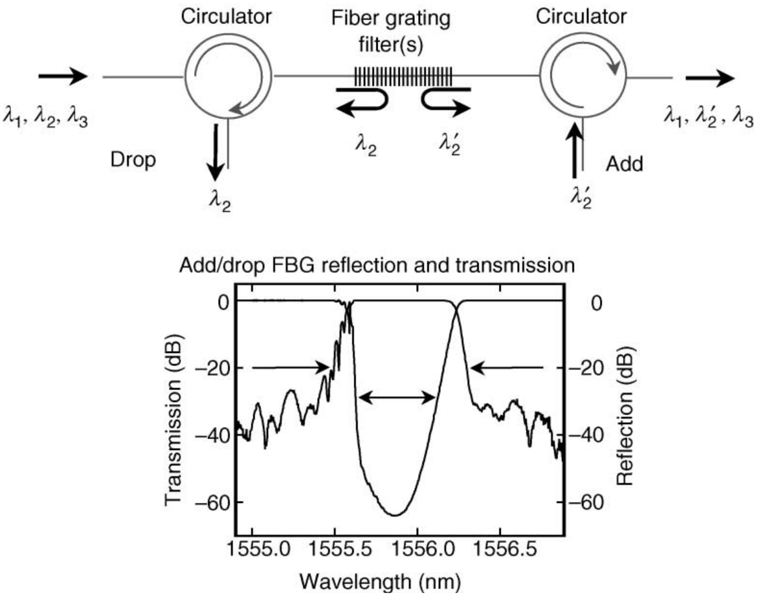

Figure 8 Top: Add/drop node employing fiber grating reflector. Bottom: characteristic reflection and transmission spectrum of an add/drop fiber grating filter.

a fiber grating based add/drop multiplexer. The grating is located between two three-port optical circulators. (A circulator is a bulk optic element that allows transmission only from port n to port $n + 1$.) Different channels are incident upon the grating from the left and only one particular channel undergoes reflection at the grating. Note that by carefully designing the grating properties the reflectivity in the adjacent channel can be made to be negligible. This one channel is reflected and re-routed by the three-port circulator. The remainder of the channels are transmitted through the grating. The grating also enables the 'add' functionality. In this case light originating from a different part of the optical network enters the link again through an optical circulator and is reflected off the fiber grating.

Gain-Flattening Loss Filters

A primary strength of fiber gratings is the diversity of filter shapes that is achievable through modification of the grating profile. Gain-flatteners are an important application requiring such filter design. Gain-flattening filters are important in WDM telecommunications systems in order to smooth the wavelength dependence of fiber amplifiers (such as Er or Raman amplifiers). Both LPGs and FBGs may be employed as gain flattening filters. The desired loss spectrum is translated into a corresponding grating profile. LPGs have the advantage of low return loss, but the LPG resonance wavelengths and strengths are very sensitive to the fiber geometry, making the computation of the grating profile less accurate. FBGs may also be used in both tilted and untilted form; however they typically have back-reflections that require careful fiber grating growth or an isolator to suppress back-reflections. Figure 9 shows an example of a gain equalizing loss filter using LPGs.

Dispersion Compensator

Chromatic dispersion can limit the transmission distances and capacities of long-haul fiber links. Chromatic dispersion arises from the variation in propagation velocity with wavelength. An initially sharp pulse is thus dispersed and begins to overlap with adjacent pulses resulting in degraded signal quality. A key technology for long-haul fiber links is therefore 'dispersion compensation'. The most common solution has been 'dispersion compensating fiber' (DCF), which is simply fiber in which the chromatic dispersion has been engineered to be the exact opposite of the dispersion in the fiber link. Limitations of DCF include nonlinearities, higher-order dispersion, and lack of tunability. Fiber gratings offer an alternative approach that overcomes the problems.

Fiber Bragg gratings can provide strong chromatic dispersion when operated in reflection with a three-port circulator (see Figure 10). In this case the grating period is made to vary along the length of the grating. This means that different wavelengths in a pulse reflect at different positions in the grating. This gives rise to wavelength-dependent group delay and therefore chromatic dispersion. The advantage of the fiber grating compared to DCF is that the device can be made to be very compact, with potentially lower

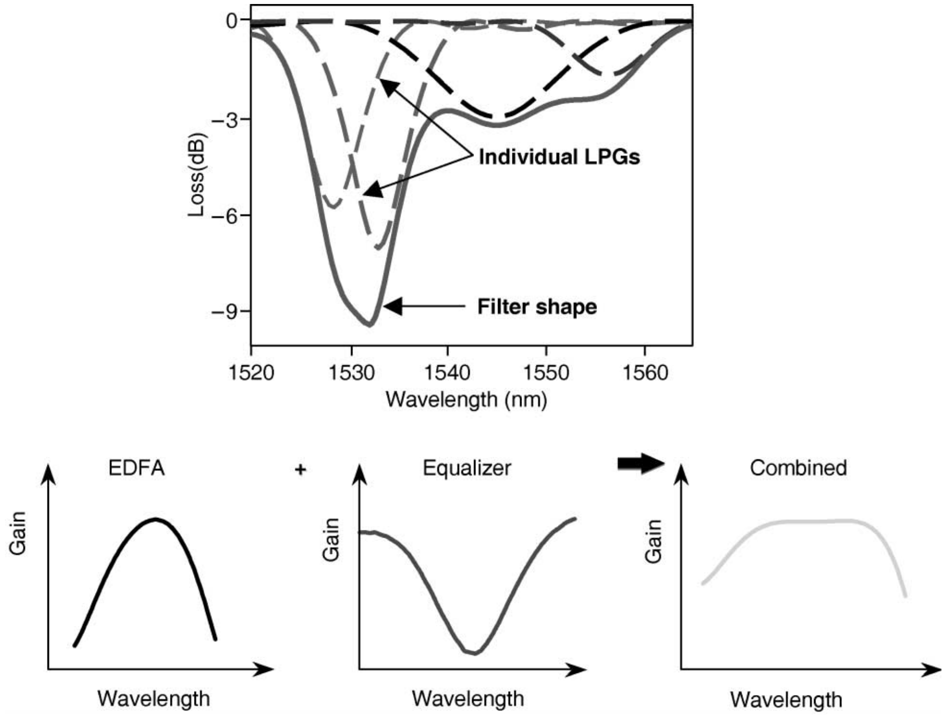

Figure 9 Long-period grating gain equalizer. Several individual gratings (dashed line) are combined to produce the overall shape (solid line).

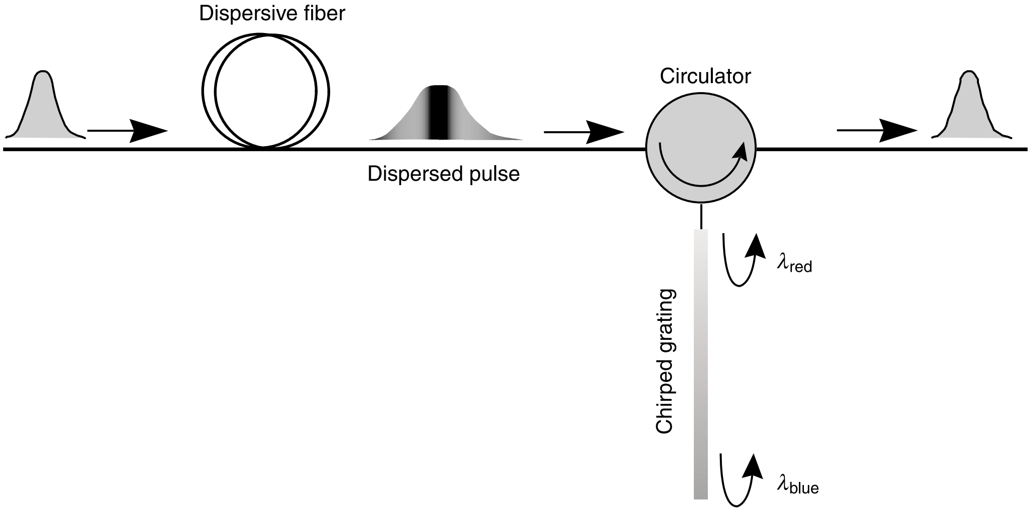

Figure 10 Chirped fiber grating dispersion compensator.

insertion loss, reduced optical nonlinearity, and controllable optical characteristics. In particular fiber gratings offer the possibility of engineering the chromatic dispersion to allow for compensation of dispersion slope and other degradations such as polarization mode dispersion.

Optical Monitors

Fiber gratings may also be employed as compact optical taps coupling core guided light to the radiation modes propagating outside the fiber (see Figure 3b). When a conventional Bragg grating is tilted with respect to the fiber axis it can scatter light out of the fiber in a highly directional manner. The angle of scatter is also wavelength dependent because of the phase matching condition discussed above. This wavelength dependence may form the basis of a compact wavelength monitor. The fiber grating is used to simultaneously couple light out of the fiber and separate the wavelengths. A simple

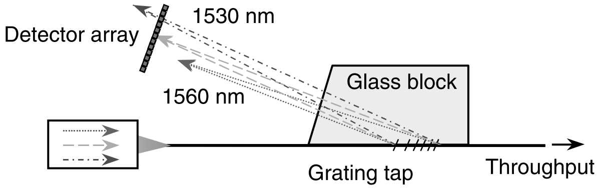

Figure 11 Fiber grating wavelength monitor. The fiber grating acts to scatter light out of fiber and simultaneously disperse the light.

detector array can then measure a spectrum of all the channels in a typical WDM system (see Figure 11).

Fiber grating taps may also scatter light in a polarization sensitive manner. This may be achieved by tilting the fiber grating by 45° with respect to the fiber axis. The resulting broadband taps may be employed in compact polarization monitoring schemes in which detectors are placed on the outside of the fiber to detect light at specific polarizations.

Advanced Topic in Fiber Gratings

Fiber Gratings in Air–Silica Microstructured Optical Fibers

Air–silica microstructure fibers (ASMFs), also known as 'holey' fibers, are fibers with air regions running along their length and allow for more complicated modal properties than conventional silica-doped fiber. It has been shown that fiber gratings can be UV written into photosensitive regions of various types of microstructure fibers. ASMF gratings open new possibilities for fiber grating devices. In addition, they are useful in studying the guidance properties of ASMFs, because they can selectively excite higher-order modes of the fiber. The phase matching condition may be used to derive the effective indices of the higher-order modes and mode profiles may be recorded by imaging a cleaved surface. Such measurements agree well with standard beam propagation computations. Figure 12 shows an example of a mode field image generated with a FBG.

One important application of ASMF gratings is a widely tunable LPG filter. Such filters employ a 'hybrid waveguide' formed by infusing a polymer into the air regions. Because the polymer index is much more sensitive to temperature than silica, the guidance properties of the cladding modes may be tuned. Thus, widely tunable LPG filters may be formed. Moreover, because the cladding modes are confined within an 'inner cladding' such filters are insensitive to external index, allowing for instance, the use of on-fiber, thin film heaters for power-efficient temperature tuning.

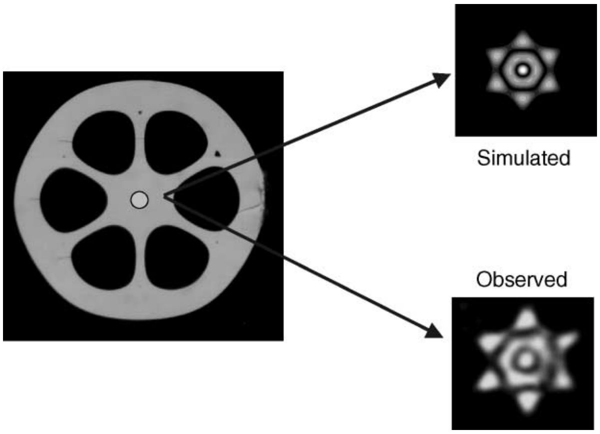

Figure 12 Cross-section of an air–silica microstructure fiber along with measured and simulated modes guided by the air regions.

Nonlinear Optics in Fiber Gratings

The linear properties of fiber gratings are well understood and as explained earlier lead to a number of interesting phenomena and applications. When the peak intensity of the optical field is very high a new class of nonlinear effects can occur. These optical nonlinearities arise from the intrinsic nonlinearity of the material system in which the grating is inscribed. In the case of optical fibers, the dominant nonlinearity is third-order nonlinearity which manifests itself in the refractive index of the material depending on the peak intensity of the optical field: $n = n_0 + n_2 I$. This nonlinearity gives rise to a number of dramatic nonlinear effects in fiber gratings, including all-optical switching, bistability, multistability, soliton generation, pulse compression, and wavelength conversion.

To understand these effects we consider an intense optical field incident upon a fiber grating with a central wavelength tuned inside the stopband of the grating where the reflectivity is high. At low intensities the pulse is reflected and only a small amount of light leaks through the grating. As the intensity of the incident optical field is increased the optical nonlinearity becomes significant and results in shifting of the local Bragg wavelength of the grating. This in turn can lead to an increase in transmission of light through the grating and bistable operation.

Theoretical studies of the continuous wave nonlinear properties of periodic media show that such media can exhibit switching as well as bistability and multistability. This prediction has been confirmed experimentally in a variety of geometries, such as in semiconductors, and in colloidal crystals. Experiments studying nonlinear effects in gratings can also be performed in optical fiber gratings. In these experiments short optical pulses are typically used in order to obtain the very high intensities. This has several key advantages. The first of these is that long, high-quality Bragg gratings can be permanently imprinted into the core of optical fibers by side illumination with interfering ultraviolet beams (see above). This allows for very long interaction lengths (compared to picosecond pulse lengths) and enables nonlinear propagation phenomena to be studied in detail. These include a number of propagation phenomena, such as the formation and propagation of Bragg solitons, modulational instability, and pulse compression. The second advantage is that although the nonlinearity is weak, it has a response time of the order of a few femtoseconds, and thus acts essentially instantaneously on the scale of all but the shortest optical pulses. Finally, the ability to engineer the fiber grating offers a number of unique geometries and potential applications of nonlinear fiber gratings.

A key idea in the interpretation of nonlinear optics in fiber Bragg gratings is the Bragg soliton. Briefly, in these solitons the enormous dispersion provided by the fiber Bragg grating is balanced by the fiber nonlinearity. This is analogous to the optical soliton that has been demonstrated in standard single-mode fiber where the dispersion is predominantly associated with the intrinsic material properties of silica. Since the dispersion of a Bragg grating can be five orders of magnitude larger than that of unprocessed fiber, Bragg solitons can be observed in gratings of only a few centimeters in length, thus rivaling the compactness of other soliton geometries, such as spatial solitons. Of course this implies that the required optical intensities to generate Bragg solitons are correspondingly higher than in uniform media.

Figure 13 shows an example of experimental results when the peak intensity of a pulse incident upon a fiber grating was approximately 10 GW/cm^2 and the incident pulse is tuned to the anomalous dispersion regime of the grating. In particular, Figure 13 shows the transmitted intensity versus time for different values of detuning. By varying the strain on the grating, we can change the detuning of the input pulse, and thus tune to different points on the dispersion curve. Note also from Figure 13 that far from the photonic bandgap edge, where the grating-induced dispersion is negligible, the pulse is unaffected and propagates through the grating at the speed of light. Closer to the edge of the photonic bandgap the pulse is substantially compressed due to the formation of a Bragg soliton. In addition to this compression, the transmitted pulse is substantially delayed in time as a result of the reduced group velocity of the pulse propagating through the grating.

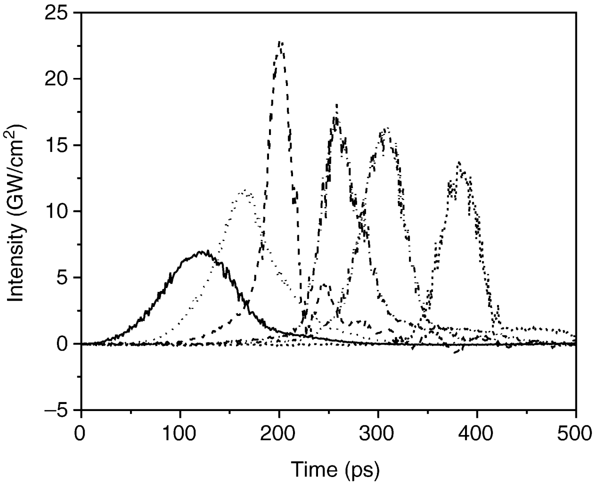

Figure 13 Pulse compression associated with nonlinear pulse propagation through a fiber Bragg grating. Compression is associated with the formation of a Bragg soliton when the incident pulse is tuned just outside of the photonic bandgap formed by the grating.

List of Units and Nomenclature

cw	Continuous wave
FBG	Fiber Bragg grating
LPG	Long-period grating

See also

Detection: Fiber Sensors. **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber; Light Propagation; Measuring Fiber Characteristics; Nonlinear Optics; Optical Fiber Cables. **Lasers:** Optical Fiber Lasers. **Optical Amplifiers:** Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts. **Optical Materials:** Optical Glasses. **Photon Picture of Light**.

Further Reading

Adams MJ (1981) *An Introduction to Optical Waveguides*. Chichester, UK: Wiley.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Eggleton BJ, Westbrook PS, Windeler RS, Spalter S and Strasser TA (1999) Grating resonances in air-silica microstructured optical fibers. *Optical Letters* 24: 1460–1462.

Eggleton BJ, Lenz G and Litchinitser NM (2000) Optical pulse compression schemes that use nonlinear Bragg gratings. *Fiber Integrated Optics* 19: 383–421.

Espindola RP, Windeler RS, Abramov AA, *et al.* (1999) External refractive index insensitive air-clad long period fiber grating. *Electronic Letters* 35: 327–328.

Hill KO, Fujii Y, Johnson DC and Kawasaki BS (1978) Photosensitivity in optical fiber waveguides: applications to reflection filter fabrication. *Applied Physics Letters* 32: 647–649.

IEEE Journal of Lightwave Technology (1997) Special issue on fiber gratings, photosensitivity and poling 15(8).

Kashyap R (1999) *Fiber Bragg Gratings*. New York: Academic Press.

Kogelnik H (1990) Theory of optical waveguides. In: Tamir T (ed.) *Guided-Wave Optoelectronics*. Berlin: Springer-Verlag.

Lemaire PJ, Atkins RM, Mizrahi V and Reed WA (1993) High-pressure H_2 loading as a technique for achieving ultrahigh UV photosensitivity and thermal sensitivity in GeO_2 doped optical fibers. *Electronic Letters* 29: 1191–1193.

Marcuse D (1974) *Theory of Dielectric Optical Waveguides*. New York: Academic Press.

Meltz G, Morey WW and Glenn WH (1989) Formation of Bragg gratings in optical fibres by transverse holographic method. *Optical Letters* 14: 823–825.

Othonos A and Kalli K (1999) *Fiber Bragg Gratings: Fundamentals and Applications in Telecommunications and Sensing*. Norwood, MA: Artech House.

Polladian L (1993) Graphical and WKB analysis of nonuniform Bragg gratings. *Physical Review E* 48: 4758–4767.

Westbrook PS, Eggleton BJ, Windeler RS, *et al.* (2001) Cladding mode loss in hybrid polymer-silica microstructured optical fiber gratings. *IEEE Photonic Technology Letters* 12: 495–497.

FOURIER OPTICS

S Jutamulia, University of Northern California, Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Coherent optical information processing is almost entirely based on the Fourier-transform property of a lens. A Fourier-transform lens is actually an ordinary lens. If the input transparency is placed in the front focal plane of the lens and illuminated with coherent collimated light (planewave), the amplitude function in the back focal plane of the lens will be the Fourier transform of the input transparency as shown in Figure 1. A coherent optical processor usually consists of two lenses, as shown in Figure 2, thus it performs two Fourier transforms successively. The first lens transforms the input function from space domain into frequency domain, and the second lens transforms back the frequency function from frequency domain to the output function in space domain.

Fourier Transform Property of Lens

Recently, the Fourier-transform property of lenses has been explained, simply using the concept of linear system and geometrical optics. In contrast, the Fourier-transform property of lenses is traditionally analyzed, based on the Fresnel–Kirchhoff diffraction theory involving quadratic phase terms (eqns [2] and [3]). In this article, we adopt the simple geometrical theory based on geometrical optics.

We start our discussion by looking at Figure 1. Consider that a coherent point source is at the center of the front focal plane of the lens, which is denoted by plane (x, y). The point source generates a collimated beam parallel to the optical axis. The amplitude function at the back focal plane of the lens, which is denoted by plane (u, v), can be described as unity. Consequently, we can express mathematically:

$$f(x, y) = \delta(x, y) \quad [1]$$

$$F(u, v) = 1 \quad [2]$$

where $f(x, y)$ is the amplitude function in the front focal plane, and $F(u, v)$ is the amplitude function in the back focal plane.

If the point source shifts a distance a along the y-axis, the amplitude function in the front focal plane is

$$f(x, y) = \delta(x, y + a) \quad [3]$$

The off-axis point source generates a tilted collimated beam. If we place a screen in the plane (u, v), we will see uniform intensity across the screen similar to that which would be seen when the source is centered on the axis. Mathematically, it means:

$$|F(u, v)|^2 = 1 \quad [4]$$

However, the phase of the amplitude function is not uniform, since the collimated beam is not parallel to

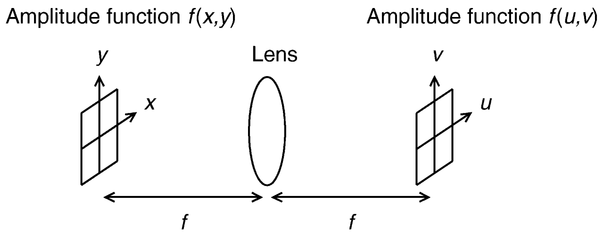
Amplitude function f(x,y)
Amplitude function f(u,v)
y
Lens
v
x
u
f
f

Figure 1 A lens Fourier-transforms amplitude function $f(x, y)$ in the front focal plane to amplitude function $F(u, v)$ in the back focal plane.

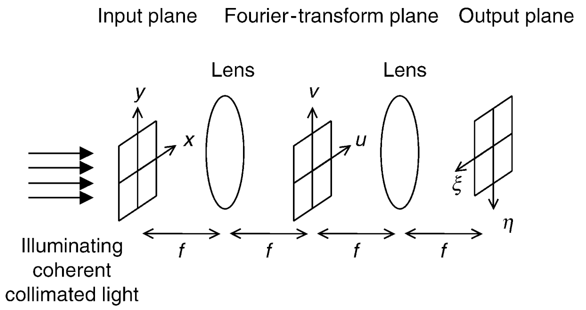
Input plane
Fourier-transform plane
Output plane
Lens
Lens
y
x
v
u
ξ
η
f
f
f
f
Illuminating
coherent
collimated light

Figure 2 A pair of lenses form 4-f coherent optical processor.

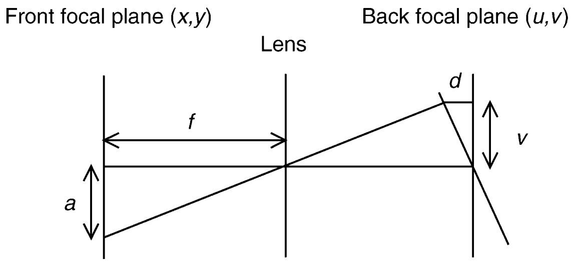
Front focal plane (x,y)
Back focal plane (u,v)
Lens
d
f
v
a

Figure 3 Calculation of phase function in the back focal plane (u, v) based on geometrical optics.

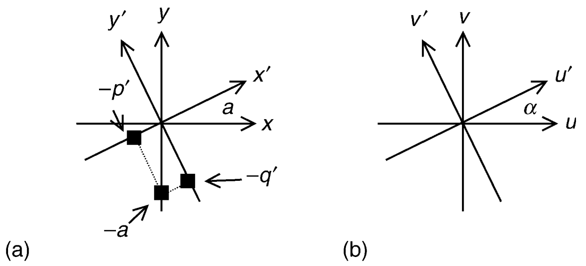
y′
y
x′
−p′
a
x
−q′
−a
(a)
v′
v
u′
α
u
(b)

Figure 4 (a) Front focal plane coordinate system (x, y) is rotated by α. (b) Similarly, back focal plane coordinate system (u, v) is rotated by α.

the optical axis. In other words, plane (u, v) is not parallel to the wavefront.

If the phase at the center of plane (u, v) is considered zero, the phase at any point at plane (u, v) can be calculated referring to Figure 3. Using simple geometry, we know that

$$d(u, v) = \frac{a}{f} v \qquad [5]$$

where $d(u, v)$ is the optical path difference. Correspondingly, the phase is

$$\phi(u, v) = \frac{2\pi}{\lambda f} av \qquad [6]$$

where λ is the wavelength of light.

In complex representation, the amplitude function is expressed as

$$F(u, v) = A(u, v) \exp[i\phi(u, v)] \qquad [7]$$

where $A(u, v)$ and $\phi(u, v)$ are magnitude and phase of the amplitude function, respectively. From eqn [4], we know that

$$A(u, v) = |F(u, v)| = 1 \qquad [8]$$

Substituting eqns [8] and [6] into [7] yields

$$F(u, v) = \exp\left(i \frac{2\pi}{\lambda f} av \right) \qquad [9]$$

Equation [9] is the amplitude function in the back focal plane of the lens, when the amplitude function in the front focal plane of the lens is given by eqn [3].

Consider now that coordinate (x,y) is rotated by angle α to new coordinates (x',y') as shown in Figure 4a. The amplitude function in the front focal plane now becomes

$$f(x', y') = \delta(x' + p', y' + q') \qquad [10]$$

where

$$p' = a \sin \alpha \qquad [11a]$$

$$q' = a \cos \alpha \qquad [11b]$$

Correspondingly, coordinate (u, v) is also rotated by the same angle α to new coordinates (u', v') as shown in Figure 4b. The old coordinates u and v can be described in terms of new coordinates u' and v' as follow:

$$u = u' \cos \alpha - v' \sin \alpha \qquad [12a]$$

$$v = u' \sin \alpha + v' \cos \alpha \qquad [12b]$$

Substituting eqn [12b] into eqn [6] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f} a(u' \sin \alpha + v' \cos \alpha) \qquad [13]$$

Further, substituting eqns [11a] and [11b] into [13] yields

$$\phi(u', v') = \frac{2\pi}{\lambda f}(p'u' + q'v') \qquad [14]$$

Thus, the amplitude function in the back focal plane of lens is

$$F(u', v') = \exp\left[i\frac{2\pi}{\lambda f}(p'u' + q'v') \right] \qquad [15]$$

For convenience, we now change parameters x' and y' to x and y, u' and v' to u and v, p' and q' to $-p$ and $-q$. Equations [10] and [15] become

$$f(x, y) = \delta(x - p, y - q) \qquad [16]$$

$$F(u, v) = \exp\left[-i\frac{2\pi}{\lambda f}(pu + qv) \right] \qquad [17]$$

We may simply consider that $f(x, y)$ is an input function and $F(u, v)$ is an output function of a system, since they are both physically the same amplitude functions. If there are two mutually coherent point sources at $(0, 0)$ and $(0, -a)$ with different amplitudes k and l in the front focal plane, the input function is

$$f(x, y) = k\delta(x, y) + l\delta(x, y + a) \qquad [18]$$

These two point sources will generate two collimated beams. Since they are mutually coherent, two collimated beams will interfere, and the resultant amplitude function in the back focal plane is simply the superposition of two beams, which is

$$F(u, v) = k + l \exp\left(i\frac{2\pi}{\lambda f} av \right) \qquad [19]$$

Thus, the lens is a linear system when the light is coherent. Furthermore, if input function $f(x, y)$ is a group of mutually coherent point sources with different amplitudes (i.e., an optical transparency illuminated with a planewave), it can be expressed as follows:

$$f(x, y) = \iint_{-\infty}^{\infty} f(p, q)\delta(x - p, y - q)\mathrm{d}p\, \mathrm{d}q \qquad [20]$$

Since the lens is a linear system, output function will be the same superposition as follows:

$$F(u,v) = \iint_{-\infty}^{\infty} f(p,q)\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv) \right]\mathrm{d}p\mathrm{d}q \qquad [21]$$

By again changing parameters p and q to x and y, respectively, eqn [21] becomes

$$F(u,v) = \iint_{-\infty}^{\infty} f(x,y)\exp\left[-i\frac{2\pi}{\lambda f}(xu+yv) \right]\mathrm{d}x\mathrm{d}y \qquad [22]$$

which is the optical Fourier transform of the input function $f(x,y)$.

4-f Coherent Optical Processor

If the amplitude function in the front focal plane of a lens is denoted by $f(x, y)$, then the amplitude function in the back focal plane is simply $F(u, v)$ given in eqn [22], which is the Fourier transform of $f(x, y)$. Mathematically, $f(x, y)$ can also be derived from $F(u, v)$ (eqn [4]). After neglecting the normalization factor, it can be expressed as follows:

$$f(x, y) = \iint_{-\infty}^{\infty} F(u, v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv) \right]\mathrm{d}u\, \mathrm{d}v \qquad [23]$$

The normalization factor $(1/\lambda f)^2$ is caused by the coordinate scale factor λf in the Fourier-transform plane (u,v).

Equations [22] and [23] are Fourier transform and inverse Fourier transform, respectively. The operation of Fourier transform is represented by a notation $F\{\cdot\}$, and inverse Fourier transform is $F^{-1}\{\cdot\}$.

Two Fourier transform lenses can be combined into a two-lens processor, which is better known as a 4-f optical processor, as shown in Figure 2. Figure 2 illustrates the principle of a 4-f optical processor, in which $f(x, y)$, $F(u, v)$, and $f(\xi, \eta)$ are amplitude functions in input, Fourier transform, and output planes, respectively. Since $F(u, v)$ is the function of spatial frequency, Fourier-transform plane is also called a frequency plane. Since a lens performs only Fourier transform, to perform inverse Fourier transform, the output plane must be rotated by 180°. Thus, orientations of ξ and η correspond to $-x$ and $-y$.

In practice, $f(x, y)$ is represented by an optical mask, which can be a film transparency or a spatial light modulator. The optical mask with $f(x, y)$ amplitude transmittance is illuminated with collimated coherent light to produce the amplitude function $f(x, y)$. The Fourier transform $F(u, v)$ is physically formed in the frequency plane. The frequency content of the input function $f(x, y)$ can be directly altered by placing an optical mask in the frequency plane. Finally, an image with amplitude function $f(\xi, \eta)$ is formed in the output plane. Notice that all detectors, including our eyes, detect intensity that is the square of amplitude.

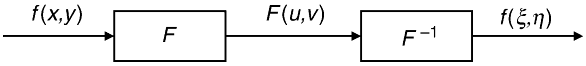

Figure 5 Block diagram of 4-f coherent optical processor.

According to eqn [22], $F(u, v)$ can be expressed in terms of $f(x, y)$ as

$$F(u, v) = F\{f(x, y)\} \qquad [24]$$

Similarly, the amplitude function in the output plane must be

$$f(x', y') = F\{F(u, v)\} \qquad [25]$$

where (x', y') is the coordinate of the output plane having the same orientation as (x, y) in the input plane. This function can be expressed in terms of inverse Fourier transform as follows:

$$f(-x', -y') = F^{-1}\{F(u, v)\} \qquad [26]$$

A rotated coordinate (ξ, η), shown in Figure 2, substitutes $(-x', -y')$, such that

$$f(\xi, \eta) = F^{-1}\{F(u, v)\} \qquad [27]$$

Thus, the successive operations can be simply illustrated by the diagram shown in Figure 5.

Spatial Filtering

The origin of optical information processing can be traced as far back as the work of Abbe of 125 years ago. His interest at that time concerned the theory of microscopy in order to establish a scientific approach to microscope design. He found that a larger aperture resulted in higher resolution. An object under the microscope, whose size was of the order of the wavelength of light, diffracted the illuminating light, i.e., the straight light was bent. If the aperture of the objective lens was not large enough to collect all of the diffracted light, the image did not correspond exactly to that object. Rather, it related to a fictitious object whose complete diffracted light pattern matched the one collected by the objective lens. These lost portions are associated with the higher spatial frequency. Their removal will result in a loss in image sharpness and resolution. The basic ideas of Abbe's theory were later clarified in a series of experiments by Porter in the early 1900s.

Since the late 1940s, the possibility of applying coherent optical techniques to image processing had been explored. In 1960, Cutrona *et al.* presented an optical system using lenses to perform Fourier transforms. Thus, every Fourier transformation in the processor can be realized by a lens. The power of this architecture is its capability of forming Fourier spectra and also being cascaded.

Consider an image formed by an optical system such as a camera; because of the error or limitation of the system, the perfect point in the object may smear. If we consider that the original object is the collection of a large number of very fine points, the formed image will be the collection of the same number of blurred spots, instead of fine points. Thus, the image formation can be expressed by convolution. The blurred spot is represented by the impulse response $h(x, y)$, which is the characteristic of the optical system.

When the input is a point (an impulse), its amplitude function is a delta function $\delta(x, y)$, and the output amplitude function is $h(x, y)$. The observed image is the intensity function $|h(x, y)|^2$. Since it indicates the spread of a point, $|h(x, y)|^2$ is called the point spread function.

The output image function $o(x, y)$ is the convolution of input function $f(x, y)$ and the impulse response $h(x, y)$ as follows:

$$o(x, y) = \int\int_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\,\mathrm{d}q \qquad [28]$$

Notice that $h(x - p, y - q)$ is the blurred spot centered at (p, q). $H(u, v)$, $F(u, v)$, and $O(u, v)$ are Fourier transforms of $h(x, y)$, $f(x, y)$, and $o(x, y)$, respectively. Substitution of

$$h(x-p, y-q) = \int\int_{-\infty}^{\infty} H(u,v) \times \exp\left[i\frac{2\pi}{\lambda f}[(x-p)u+(y-q)v]\right]\mathrm{d}u\,\mathrm{d}v \qquad [29]$$

then

$$F(u,v) = \int\int_{-\infty}^{\infty} f(p,q)\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv)\right]\mathrm{d}p\,\mathrm{d}q \qquad [30]$$

yields

$$o(x,y) = \int\int_{-\infty}^{\infty} F(u,v)H(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}u\,\mathrm{d}v \qquad [31]$$

Since

$$o(x,y) = \int\int_{-\infty}^{\infty} O(u,v)\exp\left[i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}u\,\mathrm{d}v \qquad [32]$$

we obtain

$$O(u,v) = H(u,v)F(u,v) \qquad [33]$$

The function $H(u,v)$, which is the Fourier transform of the impulse response $h(x,y)$, is called transfer function. Actually, $H(u,v)$ changes the spectrum of input function $F(u,v)$ to become $O(u,v)$ and the spectrum of the output function becomes $H(u,v)F(u,v)$.

The first application of spatial filtering may be the restoration of a photograph. The image formed by a camera is a collection of spots. If the image is in focus, the size of the spots is very fine, and thus the image looks sharp. If the image is out of focus, the spot size is fairly large and the image is blurred. Interestingly, an out-of-focus image can be restored using spatial filtering technique.

Consider that the original image is $f(x,y)$. Because it is out of focus, a point in the image becomes a spot expressed by $h(x,y)$. If the image is in focus, $h(x,y)$ is a delta function. If the in-focus image is $f(x,y)$, its frequency function is $F(u,v)$. The defocused image $o(x,y)$ is expressed by eqn [28], and its frequency function is $F(u,v)H(u,v)$. To produce the in-focus image, the frequency function must be $F(u,v)$. Thus $F(u,v)H(u,v)$ must be multiplied by the filter function of $1/H(u,v)$ to produce $F(u,v)$ only. The filter of $1/H(u,v)$ can be made from a photographic plate to provide its magnitude, and a glass plate having two values of thickness to provide +1 and −1 signs. The glass plate is essentially a phase filter. The photographic plate with varying density function is an amplitude filter. By passing frequency function $F(u,v)H(u,v)$ through the combination of amplitude and phase filters, it will result in $F(u,v)$ only. When $F(u,v)$ is again Fourier transformed by a lens, an in-focus image $f(x,y)$ will be produced.

Complex Matched Spatial Filtering

The restoration of a photograph mentioned above, utilizes the filter function $1/H(u,v)$, which has positive and negative real values. We will see an application in which the filter function is a complex function. A complex function filter can be produced by a single hologram, instead of using a combination of a phase filter and an amplitude filter. The complex matched spatial filter can perform a correlation, as first demonstrated by Vander Lugt. The complex filter is actually a Fourier transform hologram proposed by Leith and Upatnieks.

The synthesis of a complex filter is shown in Figure 6. First, a function $g(x,y)$ is displayed in the input plane by a transparency or spatial light modulator and illuminated with a collimated coherent light beam. Its Fourier-transform $G(u,v)$ is generated in the back focal plane of the lens. The hologram is made in the Fourier-transform plane by recording the interference pattern of $G(u,v)$ and the reference beam $R(u,v)$. $R(u,v)$ is an oblique collimated beam, having magnitude of unity, which is expressed mathematically as

$$R(u,v) = \exp(iau) \qquad [34]$$

and

$$a = \frac{2\pi \sin\alpha}{\lambda} \qquad [35]$$

where α is the angle of the reference beam and λ is the wavelength of light.

The transmittance of the recorded hologram will be

$$\begin{aligned} T(u,v) &= |G(u,v) + R(u,v)|^2 \\ &= |G(u,v)|^2 + 1 + G(u,v)\exp(-iau) \\ &\quad + G^*(u,v)\exp(iau) \end{aligned} \qquad [36]$$

where * denotes the complex conjugate. The hologram, which is called a complex matched spatial filter, is then placed back in the Fourier-transform plane, as shown in Figure 7. A new input function $f(x,y)$ is displayed in the input plane. The Fourier-transform of input function, $F(u,v)$ will be generated in the Fourier-transform plane. The amplitude of light,

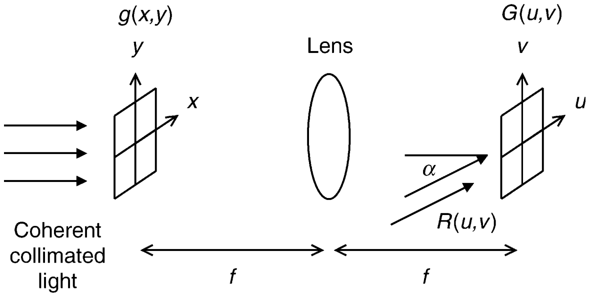

Figure 6 Synthesis of complex matched spatial filter is the recording of Fourier-transform hologram.

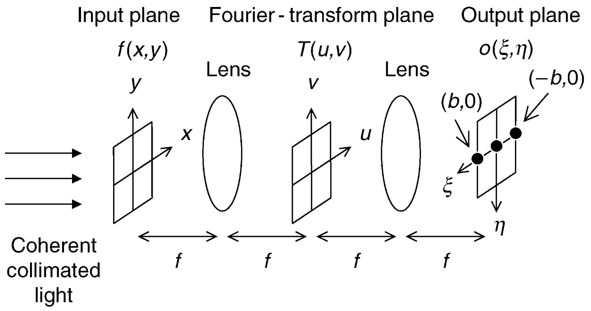

Figure 7 Optical processing using complex matched spatial filter in a 4-f processor.

immediately after the complex matched spatial filter, is

$$\begin{aligned}F(u,v)T(u,v) = {} & (|G(u,v)|^2+1)F(u,v)\\ & + F(u,v)G(u,v)\exp(-iau)\\ & + F(u,v)G^*(u,v)\exp(iau)\end{aligned} \qquad [37]$$

The second lens then produces the inverse Fourier transform of $F(u,v)T(u,v)$ in the output plane:

$$\begin{aligned}F^{-1}\{F(u,v)T(u,v)\} = {} & F^{-1}\{(|G(u,v)|^2+1)F(u,v)\}\\ & + F^{-1}\{F(u,v)G(u,v)\exp(-iau)\}\\ & + F^{-1}\{F(u,v)G^*(u,v)\exp(iau)\}\end{aligned} \qquad [38]$$

The first term will be at the center of the output plane. The second and the third terms will appear at $(b,0)$ and $(-b,0)$ in (ξ,η) plane, respectively, where

$$b = f\sin\alpha \qquad [39]$$

and f is the focal length of lens.

We are now in a position to analyze the inverse Fourier transform of $F(u,v)G(u,v)$ and $F(u,v)G^*(u,v)$. Referring to eqns [28]–[33], the inverse Fourier transform of $F(u,v)G(u,v)$ is the convolution of $f(\xi,\eta)$ and $g(\xi,\eta)$. Thus, the complex matched spatial filtering can also be applied to the restoration of photography mentioned previously without using the combination of amplitude and phase filters.

The inverse Fourier transform of $F(u,v)G^*(u,v)$ is derived as follows:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} F(u,v)G^*(u,v) \times\exp\left[i\frac{2\pi}{\lambda f}(\xi u+\eta v)\right]\mathrm{d}u\,\mathrm{d}v \qquad [40]$$

Substitute

$$F(u,v) = \iint_{-\infty}^{\infty} f(\alpha,\beta) \times\exp\left[-i\frac{2\pi}{\lambda f}(\alpha u+\beta v)\right]\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [41]$$

and

$$G^*(u,v) = \iint_{-\infty}^{\infty} g^*(-p,-q) \times\exp\left[-i\frac{2\pi}{\lambda f}(pu+qv)\right]\mathrm{d}p\,\mathrm{d}q \qquad [42]$$

into eqn [40], one obtains

$$\begin{aligned}o_3(\xi,\eta) = {} & \iint_{-\infty}^{\infty} f(\alpha,\beta)\iint_{-\infty}^{\infty} g^*(-p,-q)\\ & \times\iint_{-\infty}^{\infty}\exp\left[i\frac{2\pi}{\lambda f}(\xi-\alpha-p)u\right.\\ & \left. +(\eta-\beta-q)v\right]\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}p\,\mathrm{d}q\,\mathrm{d}\alpha\,\mathrm{d}\beta\end{aligned}$$

$$\begin{aligned}o_3(\xi,\eta) = {} & \iint_{-\infty}^{\infty} f(\alpha,\beta)\iint_{-\infty}^{\infty} g^*(-p,-q)\\ & \times\delta(\xi-\alpha-p,\eta-\beta-q)\mathrm{d}p\,\mathrm{d}q\,\mathrm{d}\alpha\,\mathrm{d}\beta\end{aligned}$$

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)g^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\,\mathrm{d}\beta \qquad [43]$$

Equation [43] is the correlation of $f(x,y)$ and $g(x,y)$. It is apparent that the correlation output $o_3(\xi,\eta)$ will achieve the maximum when $f(x,y)$ is identical to $g(x,y)$. If $f(x,y)$ and $g(x,y)$ are different, the correlation function will be a flat function with minor random fluctuation. Optical correlators using complex matched spatial filters have been applied to automatic pattern recognition.

Joint Transform Correlator

Alternatively, the joint transform correlator can perform a correlation without recording a complex matched spatial filter in advance. In addition to the correlation, the joint transform correlator can also be used as a general optical processor.

A transparency with amplitude transmittance $f(x,y)$ is centered at $(-a,0)$, and a second transparency with amplitude transmittance $g(x,y)$, is centered at $(a,0)$. The two transparencies are illuminated by a coherent collimated beam and jointly Fourier transformed by a single lens as shown in Figure 8. In the back focal plane of the lens, the amplitude function $S(u,v)$ is

$$S(u,v) = \iint_{-\infty}^{\infty}[f(x+a,y)+g(x-a,y)] \times\exp\left[-i\frac{2\pi}{\lambda f}(xu+yv)\right]\mathrm{d}x\,\mathrm{d}y \qquad [44]$$

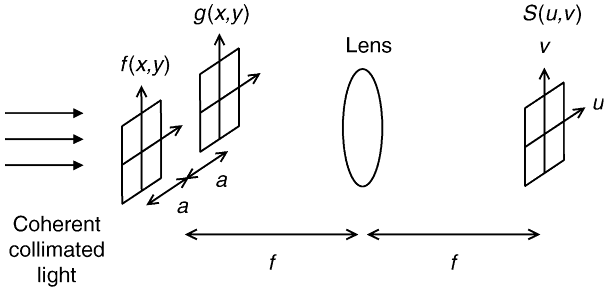

Figure 8 Joint Fourier transform of two input functions.

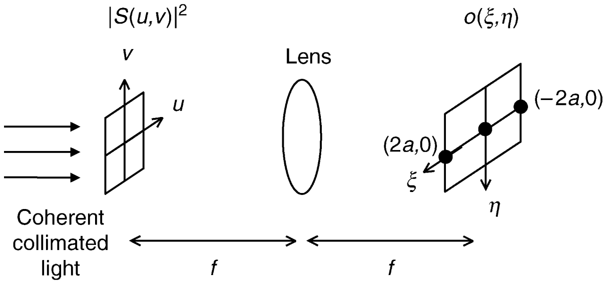

Figure 9 Fourier transform of the joint power spectrum.

$S(u,v)$ is the joint Fourier spectrum of $f(x,y)$ and $g(x,y)$. It can be written as

$$S(u,v) = F(u,v)\exp\left[i\frac{2\pi}{\lambda f}au\right] + G(u,v)\exp\left[-i\frac{2\pi}{\lambda f}au\right] \quad [45]$$

where $F(u,v)$ and $G(u,v)$ are Fourier transforms of $f(x,y)$ and $g(x,y)$, respectively.

If a square-law detector is placed in the back focal plane of the lens, the recorded pattern of joint power spectrum $|S(u,v)|^2$ is

$$\begin{aligned}|S(u,v)|^2 = {} & |F(u,v)|^2 \\ & + F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right] \\ & + F^*(u,v)G(u,v)\exp\left[-i\frac{4\pi}{\lambda f}au\right] \\ & + |G(u,v)|^2 \end{aligned} \quad [46]$$

where * denotes the complex conjugate. A photographic plate is a square-law detector and can be used to record $|S(u,v)|^2$. When the developed plate is placed in the input plane, as shown in Figure 9, the input transmittance is $|S(u,v)|^2$ as given in eqn [46].

The inverse Fourier transform of the joint power spectrum $|S(u,v)|^2$, in the back focal plane of the lens, is

$$o(\xi,\eta) = \iint_{-\infty}^{\infty} |S(u,v)|^2 \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [47]$$

which consists of four terms, according to eqn [46]. We now analyze this equation term by term. The first term is

$$o_1(\xi,\eta) = \iint_{-\infty}^{\infty} |F(u,v)|^2 \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [48]$$

which is the auto-correlation of $f(x,y)$

$$o_1(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)f^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\,\mathrm{d}\beta \quad [49]$$

The second term is

$$o_2(\xi,\eta) = \iint_{-\infty}^{\infty} F(u,v)G^*(u,v)\exp\left[i\frac{4\pi}{\lambda f}au\right] \times \exp\left[i\frac{2\pi}{\lambda f}(\xi u + \eta v)\right]\mathrm{d}u\,\mathrm{d}v \quad [50]$$

which is

$$o_2(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta)g^*[\alpha-(\xi+2a),\beta-\eta]\mathrm{d}\alpha\,\mathrm{d}\beta \quad [51]$$

Equation [51] is the correlation between $f(x,y)$ and $g(x,y)$, shifted by $-2a$ in the ξ-axis direction.

Similarly, the third term will be the correlation between $g(x,y)$ and $f(x,y)$, shifted by $2a$ in the ξ-axis direction:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} g(\alpha,\beta)f^*[\alpha-(\xi-2a),\beta-\eta]\mathrm{d}\alpha\,\mathrm{d}\beta \quad [52]$$

And the fourth term is

$$o_4(\xi,\eta) = \iint_{-\infty}^{\infty} g(\alpha,\beta)g^*(\alpha-\xi,\beta-\eta)\mathrm{d}\alpha\,\mathrm{d}\beta \quad [53]$$

which is the auto-correlation of $g(x,y)$ and overlaps with the first term at the center of the output plane.

From the analysis given above, it is understood that the correlation of $f(x,y)$ and $g(x,y)$, and the correlation of $g(x,y)$ and $f(x,y)$, are shifted by $-2a$ and $2a$, respectively, along the ξ-axis. The correlation outputs are centered at $(-2a,0)$ and $(2a,0)$ in the back focal plane of the lens. If $f(x,y)$ and $g(x,y)$ are the same, the two correlation outputs are identical. The first and fourth terms are the autocorrelations of $f(x,y)$ and $g(x,y)$, centered at $(0,0)$. Thus the second and third terms, which are correlation signals, can be isolated. A versatile real-time joint transform correlator, using inexpensive liquid crystal televisions, is widely used in laboratory demonstration.

See also

Information Processing: Coherent Analogue Optical Processors. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Abbe E (1873) Beitrage zur Theorie des Mikroskops und der mikroskopichen Wahrnehmung. *Archiv Mikroskopie Anatomie* 9: 413–468.

Bracewell RN (1965) *The Fourier Transform and Its Applications*. New York: McGraw-Hill.

Cutrona LJ, Leith EN, Palermo CJ and Porcello LJ (1960) Optical data processing and filtering systems. *IRE Transactions on Information Theory* IT-6: 386–400.

Goodman JW (1968) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Jutamulia S (1992) Joint transform correlators and their applications. *Proceedings of the SPIE* 1812: 233–243.

Jutamulia S and Asakura T (2002) Optical Fourier-transform theory based on geometrical optics. *Optical Engineering* 41: 13–16.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal of the Optical Society of America* 52: 1123–1130.

Porter AB (1906) On the diffraction theory of microscopic vision. *Philosophical Magazine* 11: 154–166.

Rau JE (1966) Detection of differences in real distributions. *Journal of the Optical Society of America* 56: 1490–1494.

Tsujiuchi J (1963) Correction of optical images by compensation of aberrations and by spatial frequency filtering. *Progress in Optics* 2: 133–180.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10: 139–145.

Weaver CS and Goodman JW (1966) A technique for optically convolving two functions. *Applied Optics* 5: 1248–1249.

Yu FTS (1983) *Optical Information Processing*. New York: Wiley.

Yu FTS and Ludman JE (1988) Joint Fourier transform processor. *Microwave Optical Technology Letters* 1: 374–377.

Yu FTS, Jutamulia S, Lin TW and Gregory DA (1987) Adaptive real-time pattern recognition using a liquid crystal TV based joint transform correlator. *Applied Optics* 26: 1370–1372.

GEOMETRICAL OPTICS

Contents

Lenses and Mirrors
Aberrations
Prisms

Lenses and Mirrors

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lenses are carefully shaped pieces of transparent material used to form images. The term ‘transparent’ does not necessarily mean that the lens is transmitting visible light. The element germanium, well-known in the transistor industry, is a light-gray opaque crystal which forms images with infrared radiation, and provides the optics for night-vision instruments. The lenses used in cameras, telescopes, microscopes, and other well-known applications, are made from special glasses or – more recently – from plastics. Hundreds of different optical glass formulas are listed in the manufacturers’ catalogues. Plastics have come into use in recent years because it is now possible to mold them with high precision. The way in which lenses form images has been studied for several hundred years, resulting in an enormous and very complicated literature. In this article, we show how this theory can be simplified by introducing the restriction that lenses are perfect. This limitation provides an easy introduction to the image forming process. A companion article **Geometrical Optics:** Aberrations, will deal with imperfect components, showing how their defects degrade images and how corrections are made.

The Cardinal Points and Planes of an Optical System

The location, size and orientation of an image are items of information which a designer needs to know. The nature of the image can be determined by a simple model of an optical system, based on two experimental observations. The first is well-known and is pictured in Figure 1. Parallel rays of light passing through a double convex lens (a simple magnifying glass, for example) should meet at the focal point or focus, designated as F'. For rays from the right side of the lens, another such point F exists at an equal distance from the lens. The other experiment uses the lens to magnify (Figure 2a). The amount of magnification decreases as the lens is brought closer to ‘ISAAC NEWTON’ (Figure 2b), and when the lens touches the paper, object and image are approximately the same size. The locations of object and image corresponding to equality in size are called the unit or principal points H and H', respectively, and the set of four points F, F', H, and H' are the cardinal points. Therefore, the planes through these points normal to the axis of the lens are the cardinal planes. Figure 3 shows an object whose image we can now determine. The coordinate orientation in this figure may appear strange at first. It is customary in optics to designate the lens axis as OZ, and the other two axes form a right-handed system. The x-axis then lies in the

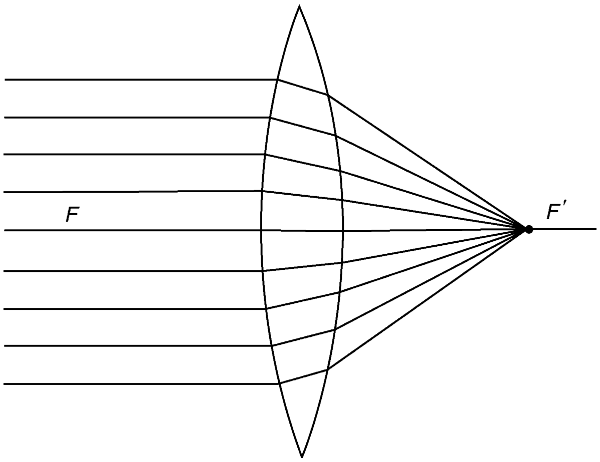

Figure 1 Focal points of a double convex lens.

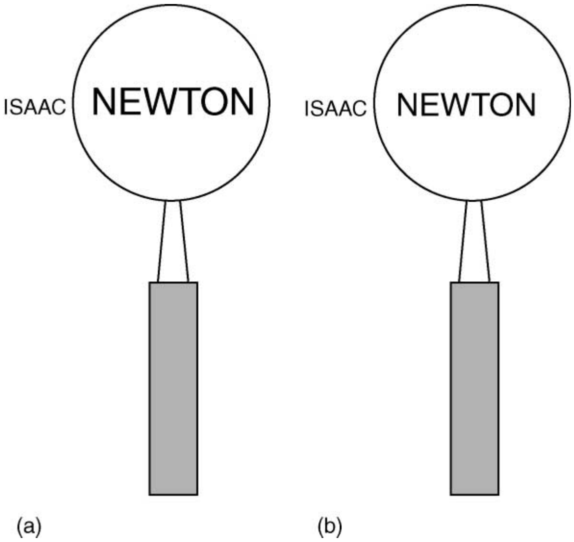

Figure 2 Magnification by a double convex lens. The magnification in (a) is reduced when the lens is brought closer to the image (b).

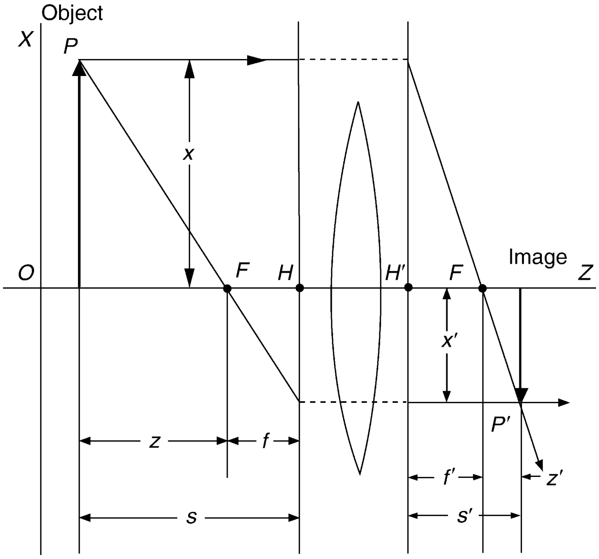

Figure 3 Ray tracing using cardinal points and planes.

plane of the figure; the y-axis, coming out of the paper and perpendicular to it, is not shown. The object, of height x, is placed to the left of the focal point F. The cardinal points occur in the order F, H, H', F', since we expect – based on the experiment corresponding to Figure 2 – that the unit points are very close to the lens. Another fact to be demonstrated later is that H and H' are actually inside the lens. Two rays are shown leaving the point P at the top of the object. The ray which is parallel to the axis strikes the unit plane through H and continues to the right, completely missing the lens. The ray then meets the unit plane through H' at a point which must be a distance x from the axis. To understand why, we recognize that if the object were relocated so as to lie on the object-space unit plane, the ray emerging at the other unit plane will be at a distance x from the axis; as required by the definition of these planes. We then assume that the ray is bent or refracted at the unit plane at H' and proceeds to, and beyond, the focal point F'. The second ray, from P through the focal point F, is easily traced since the lens is indifferent to the direction of the light. Assume temporarily that we know the location of the image point P'. Let a parallel ray leave this point and travel to the left. The procedure just given indicates that this ray will strike the unit plane at H', emerge at the other unit plane, be refracted so as to pass through F and reach the object point P. Then reversing the direction of this ray, it will start at P, emerge at the unit plane H' and travel parallel to the axis, intersecting the other ray at P'. The distance between F and H is called the object space focal length f, with f' being the corresponding image space quantity. Simple geometry, making use of the similar triangles and the distances indicated in this diagram, can be used to derive the Newton lens equation or the equivalent Gauss lens equation. Detailed derivations of these two equations will be found in either of the sources listed in the Further Reading at the end of this article. It is not usually made clear in physics texts that these equations apply only to the special case of a single, thin lens, and we shall not apply them here. The image in this figure is real, inverted, and reduced. That is, it is smaller than the object, oriented in the opposite direction, and can be projected onto a screen located at the image position. This result can be verified by using an ordinary magnifier to form the image of a distant light source on a sheet of white paper.

Paraxial Matrices

To consider how light rays behave in optical systems, we introduce the index of refraction n of a material medium, defined as the ratio of the velocity of light c in a vacuum to its velocity v in that medium, or:

$$n = \frac{c}{v} \qquad [1]$$

The velocity of light in air is approximately the same as it is in a vacuum. A light ray crossing the interface between air and a denser medium such as glass will be bent towards the normal to the surface. This is the phenomenon of refraction and the amount of bending is governed by Snell's law, which has the form:

$$n \sin\theta = n' \sin\theta' \qquad [2]$$

where n and n' are the indices of refraction of the two media and θ and θ' are the angles as measured from the normal. Figure 4 shows a ray of light leaving an object point P and striking the first surface of a lens at point P_1. It is refracted there, and proceeds to the second surface. All rays shown lie in the z, x-plane or meridional plane, which passes through the symmetry axis OZ. The amount of refraction is specified by Snell's law, eqn [2], which we shall now simplify. The sine of an angle can be approximated by the value of the angle itself, if this value is small when expressed in radians (less than 0.1 rad), and Snell's law simplifies to

$$n_1\theta_1 = n'_1\theta'_1 \qquad [3]$$

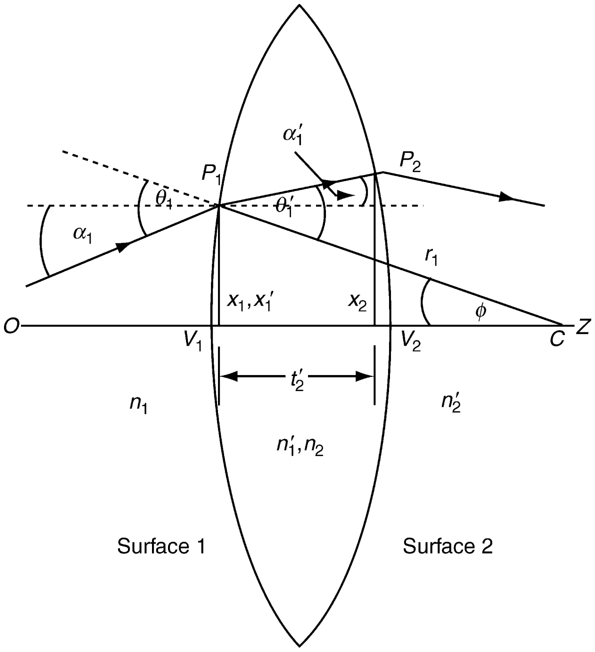

Figure 4 Ray passing through a double convex lens.

This is called the paraxial form of Snell's law; the word paraxial means 'close to the axis'. It turns out, however, that the Snell's law angles are not convenient to work with, and we eliminate them with the terms:

$$\theta_1 = \alpha_1 + \phi, \qquad \theta'_1 = \alpha'_1 + \phi \qquad [4]$$

where α_1 is the angle which the incident ray makes with the axis OZ, α'_1 is the corresponding angle for the refracted ray, and ϕ is the angle which the radius r_1 (the line from C to P_1) of the lens surface makes at the center of curvature C. This particular angle can be specified as

$$\sin\phi = \frac{x_1}{r_1} \qquad [5]$$

but using the paraxial approximation simplifies this to

$$\phi = \frac{x_1}{r_1} \qquad [6]$$

Substituting for the angles gives

$$n'_1\alpha'_1 = \frac{n_1 - n'_1}{r_1} + n_1\alpha_1 \qquad [7]$$

Notice that the distance from the point P_1 to the axis is labeled as either x_1 or x'_1. This strange notation leads to the trivial relation

$$x_1 = x'_1 \qquad [8]$$

and this can be combined with eqn [7] to obtain the matrix equation:

$$\begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [9]$$

where the constant k_1 is called the refracting power of surface 1 and is defined as

$$k_1 = \frac{n'_1 - n_1}{r_1} \qquad [10]$$

To review the procedure for matrix multiplication, the first element $n'_1\alpha'_1$ in the one-column product matrix is calculated by multiplying the upper left-hand corner of the 2×2 matrix with the first element of the one-column matrix to its right, obtaining $1(n_1\alpha_1)$, and to this is added the product $-k_1x_1$ of the upper right-hand element in the 2×2 matrix and the second member of the one-column matrix. This square matrix is called the refraction matrix R_1 for surface 1, defined as

$$R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} \qquad [11]$$

Next, we look at what happens to the ray as it travels from surface 1 to surface 2. As it goes from P_1 to P_2, its distance from the axis becomes

$$x_2 = x'_1 + t'_1 \tan \alpha'_1 \qquad [12]$$

or using the paraxial approximation for small angles:

$$x_2 = x'_1 + t'_1 \alpha'_1 \qquad [13]$$

Another approximation we shall use is to regard t'_1 as being equal to the distance between the lens vertices V_1 and V_2. Using the identity:

$$\alpha_2 = \alpha'_1 \qquad [14]$$

leads to a second matrix equation:

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} \qquad [15]$$

where the 2×2 translation matrix T_{21} is defined as

$$T_{21} = \begin{pmatrix} 1 & 0 \\ t'_1/n'_1 & 1 \end{pmatrix} \qquad [16]$$

To get an equation which combines a refraction followed by a translation, we take the one-column matrix on the left side of eqn [9] and substitute it into eqn [15] to obtain

$$\begin{pmatrix} n_2\alpha_2 \\ x_2 \end{pmatrix} = T_{21}R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [17]$$

Note that the 2×2 matrices appear in right to left order and that the multiplication of two square matrices is an extension of the rule given above. This equation gives the location and slope of the ray which strikes surface 2 after a refraction and a translation. To continue the ray trace at surface 2, we introduce a second refraction matrix R_2 by expressing k_2 in terms of the two indices at this surface and the radius. Then eqn [17] is extended to give the relation:

$$\begin{pmatrix} n'_2\alpha'_2 \\ x'_2 \end{pmatrix} = R_2T_{21}R_1 \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [18]$$

This process can obviously be applied to any number of components. The product of the three 2×2 matrices in the above equation is known as the system matrix S_{21} and it completely specifies the effect of the lens on the incident ray passing through it. It is also written as

$$S_{21} = R_2T_{21}R_1 = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \qquad [19]$$

where the four quantities a, b, c, and d are known as the Gaussian constants. They are used extensively in specifying the behavior of a lens or an optical system. We now state the sign and notation conventions which are necessary for the application of paraxial matrix methods:

1. Light normally travels from left to right.
2. The optical systems we deal with are symmetrical about the z-axis. The intersections of the refracting or reflecting surfaces with this axis are the vertices and are designated in the order encountered as V_1, V_2, etc.
3. Positive directions along the axes are measured from the origin in the usual Cartesian fashion, so that horizontal distances (that is, along the z-axis) are positive if measured from left to right. Angles are positive when measured up from the z-axis.
4. Quantities associated with the incident ray are unprimed; those for the refracted ray are primed.
5. A subscript denotes the associated surface.
6. If the center of curvature of a surface is to its right, the radius is positive, and vice versa.
7. There is a special rule for mirrors to be explained below.

Using the Gaussian Constants

Figure 5 shows the double convex lens of Figure 4 with the assumed locations of the cardinal points indicated. These positions, which we shall now determine accurately, are at distances designated as l_F, l'_F, l_H, and l'_H and are measured from the associated vertex. The object position can be called t, a positive quantity which is measured to the right from object to the first vertex, or it can be called t_1 if measured in the opposite direction. For the first choice, the matrix specifying the translation from object to lens will have the quantity t/n_1 in its lower left-hand corner. However, it is both logical and convenient to use the first vertex as the reference point. Hence, we replace t/n_1 in the translation matrix with the quantity $-t_1/n_1$, and remember to specify t_1 as a negative number when calculating the image position. The equation connecting object and image is obtained by starting with this matrix, multiplying it by the system matrix of eqn [19], and finally by a translation matrix

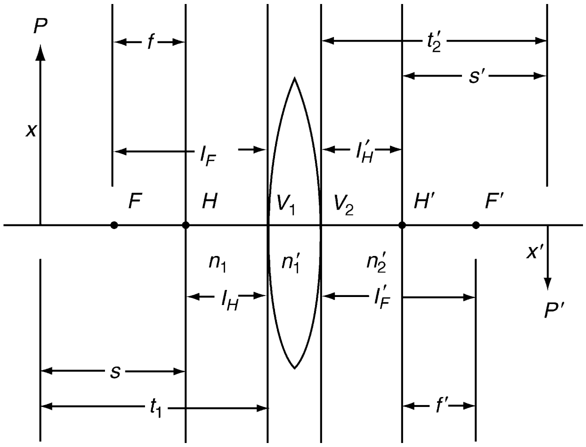

Figure 5 Definitions of cardinal plane locations.

corresponding to the translation t'_2 from lens to image to obtain:

$$\begin{pmatrix} n'_2\alpha'_2 \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'_2/n'_2 & 1 \end{pmatrix}\begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ -t_1/n_1 & 1 \end{pmatrix}\begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [20]$$

where α or α_1 is the angle that the ray from the top of the object makes with the z-axis. This equation takes the initial value of the inclination α and of the position x of the ray leaving the object, and determines the final values, α' and x' at the image. The product of the three 2×2 matrices on the right-hand side can be consolidated into a single matrix called the object-image matrix $S_{P'P}$. Then eqn [20] can be written as

$$\begin{pmatrix} n'_2\alpha'_2 \\ x' \end{pmatrix} = S_{P'P}\begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [21]$$

and this matrix has the complicated form:

$$S_{P'P} = \begin{pmatrix} b + \dfrac{at_1}{n_1} & -a \\ \dfrac{bt'_2}{n'_2} + \dfrac{at_1t'_2}{n_1n'_2} - d - \dfrac{ct_1}{n_1} & c - \dfrac{at'_2}{n'_2} \end{pmatrix} \qquad [22]$$

If we put this matrix into the previous equation, we obtain an unsatisfactory result: the value of x' will depend on the angle α_1 made by the incident ray, as can be seen by multiplying the two matrices on the right side. The ratio of x' to x is called the magnification m; that is

$$m = \frac{x'}{x} \qquad [23]$$

A perfect image can be formed only if the magnification is determined solely by the object distance and the constants of the lens. To eliminate this difficulty, the lower left-hand element in the matrix, eqn [22] is required to be equal to zero, and the magnification is then:

$$m = \frac{x'}{x} = c - \frac{at'_2}{n'_2} \qquad [24]$$

The determinant of this matrix is unity, since it is the product of three matrices whose individual determinants are unity. It follows that

$$\frac{1}{m} = b + \frac{at_1}{n_1} \qquad [25]$$

so that

$$\begin{pmatrix} n'_2\alpha'_2 \\ x' \end{pmatrix} = \begin{pmatrix} 1/m & -a \\ 0 & m \end{pmatrix}\begin{pmatrix} n_1\alpha_1 \\ x \end{pmatrix} \qquad [26]$$

Using the fact that the lower left-hand element of the matrix eqn [22] is zero shows that

$$\frac{t'_2}{n'_2} = \frac{d + ct_1/n_1}{b + at_1/n_1}, \qquad \frac{t_1}{n_1} = \frac{d - bt'_2/n'_2}{-c + at'_2/n'_2} \qquad [27]$$

This equation connects the object distance t_1 with the image distance t'_2, both quantities being measured from the appropriate vertex. This relation is known as the generalized Gauss lens equation. It applies to all lens systems, no matter how complicated.

We can determine the location of the unit planes by letting $m = 1$. This t'_2 in eqn [27] is the location l'_H of the image space unit plane, and it becomes

$$l'_H = \frac{n'_2(c-1)}{a} \qquad [28]$$

This relation expresses the location of the unit plane on the image side as the distance from V_2 to H'. Similarly, the location of H with respect of V_1 is

$$l_H = \frac{n_1(1-b)}{a} \qquad [29]$$

To locate the focal planes, consider a set of parallel rays coming from infinity and producing a point image at F'. The terms containing t_1 in eqn [27] are much larger than d or b, and this relation becomes

$$l'_F = \frac{n'_2 c}{a} \qquad [30]$$

Letting t'_2 become infinite in the right-hand half of eqn [27], the location of F is given by the equation

$$l_F = \frac{-n_1 b}{a} \qquad [31]$$

The focal lengths f and f' shown in Figure 5 were defined earlier as the distances between their respective focal and unit planes. Hence

$$f' = l'_F - l'_H = \frac{n'_2}{a} \qquad [32]$$

and

$$f = l_F - l_H = \frac{-n_1}{a} \qquad [33]$$

If we let $n_1 = n'_2 = 1$ for air, these become

$$-f = f' = \frac{1}{a} \qquad [34]$$

The Gaussian constant a is thus the reciprocal of the focal length in air. Even when the lens is not in air, it is still true that $f' = -f$ if the lens is in a single medium. It is also true regardless of whether or not the lens is symmetric.

It is customary in optics to work with a consistent set of units, such as centimeters or millimeters, and the sign conventions used for x and y are the standard Cartesian rules. As mentioned above, curved surfaces which open to the right have a positive radius and vice versa. As an example, let a double convex lens have radii of 2.0 and 1.0, respectively, an index of refraction of 1.5, and a thickness of 0.5. These quantities are then denoted as

$$\begin{aligned} &r_1 = +2.0, \qquad r_2 = -1.0, \qquad t'_1 = t_2 = +0.5 \\ &n_1 = 1.0, \qquad n'_1 = 1.5 = n_2, \qquad n'_2 = 1.0 \end{aligned} \qquad [35]$$

By eqn [10]:

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{1.5 - 1.0}{2.0} = 0.25 \qquad [36]$$

and

$$k_2 = \frac{1.0 - 1.5}{-1.0} = 0.50 \qquad [37]$$

The system matrix S_{21} is

$$\begin{aligned} S_{21} &= R_2 T_{21} R_1 \\ &= \begin{pmatrix} 1 & -0.50 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0.5/1.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -0.25 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.83 & -0.71 \\ 0.33 & 0.92 \end{pmatrix} \end{aligned} \qquad [38]$$

and we note that the determinant is

$$(0.83)(0.92) - (0.71)(-0.33) = 1.00 \qquad [39]$$

This property of the system matrix provides a very useful check on the accuracy of matrix multiplications. The locations of the four cardinal points can be determined from the formulas given above and their locations are shown in the scale drawing of Figure 6. Since we know how to locate the cardinal points of a lens, we are in a position to ray trace in a precise manner, using the method of Figure 3 and adding a third ray. After locating F, H, H', and F', we no longer need the lens; the cardinal planes with the proper separation are fully equivalent to the lens they replace. This is very much like what electrical engineers do when they replace a complicated circuit with a black box; this equivalent circuit behaves just like the original. Ray tracing becomes more complicated when we deal with diverging lenses (Figure 7), but the procedure gives above still applies. Parallel rays spread apart and do not come to a focus on the right-hand side of the lens. However, if the refracted

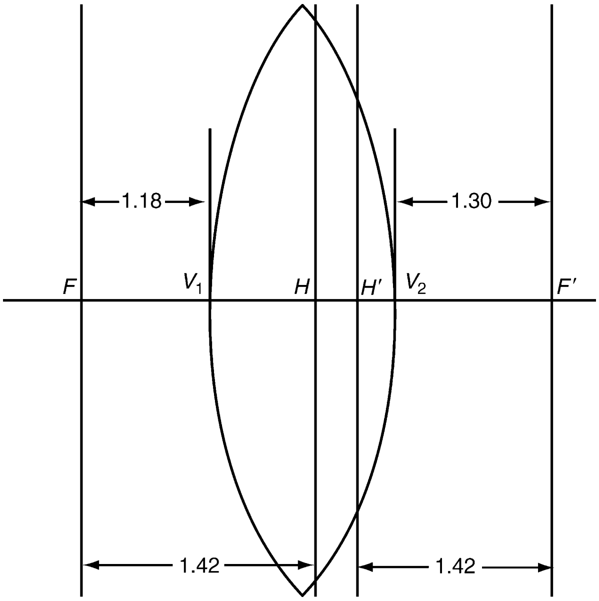

Figure 6 Positions of cardinal planes for an asymmetric lens.

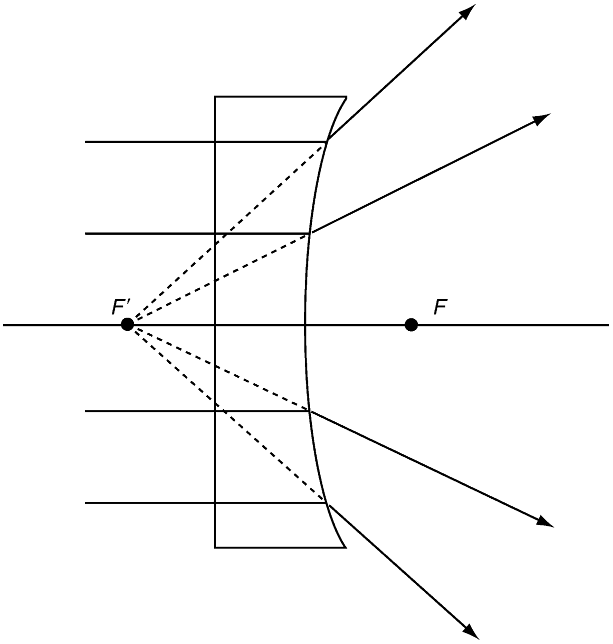

Figure 7 Ray tracing for a planar-concave lens.

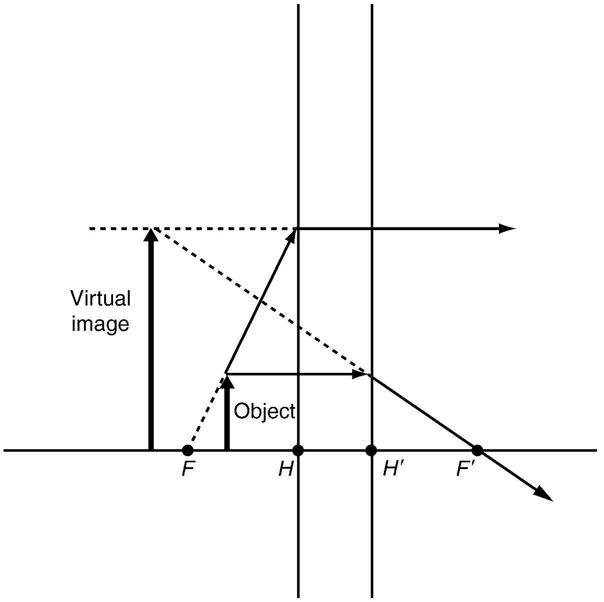

Figure 8 Formation of a virtual image by a convex lens.

rays are extended backwards, these extensions will meet as indicated. This behavior can be verified quantitatively by using the formulas developed above. It will be found that F and F' have exchanged places, but H and H' retain their original positions inside the lens. If an object is placed between the focal point F and the first vertex of a convex lens (Figure 8), then our usual procedure produces two rays which do not intersect in image space. The ray from the object which is parallel to the axis is refracted downward so that it does not intersect the other ray in image space. However, the extensions to the left of these two rays meet to form an image which is erect, magnified, and virtual. This is the normal action of a magnifying lens; the image can be seen but cannot be projected onto a screen, unlike the real, inverted image of Figure 3. This ray tracing diagram confirms what would be seen when a double convex lens is used to magnify an object lying close to the lens. For the double concave lens of Figure 9, a parallel ray leaves an object point P and is refracted upward at the unit plane H' in such a way that its extension, rather than the ray itself, passes through F'. The ray headed for F is refracted at H before it can reach the object-space focal plane and becomes parallel to the axis. The third ray, going from P to H, emerges at H' parallel to its original direction and its extension to the left passes through the image point P' already determined by the intersection of the other two rays. The resulting virtual image is upright and reduced. All three rays in this diagram behave exactly like the corresponding rays in Figure 3; the only change is the use of the extensions to locate the image. The eye receives the diverging rays from P and believes that they are coming from P'.

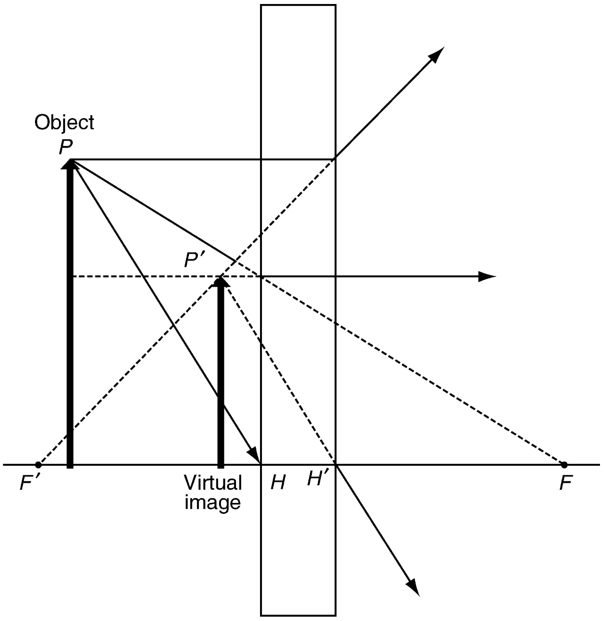

Figure 9 Formation of a virtual image by a concave lens.

Nodal Points and Planes

Let a ray leave a point on the z-axis at an angle α_1, and have an angle α'_2 when it reaches image space.

Since $x = 0$ at the starting point, eqn [26] shows that

$$n_2'\alpha_2' = n_1\alpha_1/m \qquad [40]$$

The ratio of the final angle to the initial angle in this equation is the angular magnification μ, which is then

$$\mu = n_1/mn_2' \qquad [41]$$

The locations of object and image for unit angular magnification are called the nodal points, labeled as N and N'. The above definition shows that the linear and angular magnification are reciprocals of one another when object and image are in air or the same medium. The procedure that located the unit points will show that the nodal points have positions given by

$$\frac{l_N}{n_1} = \frac{(n_2'/n_1) - b}{a} \qquad [42]$$

and

$$\frac{l_N'}{n_2'} = \frac{c - (n_1/n_2')}{a} \qquad [43]$$

When the two indices are identical, these relations are identical to those for the points H and H'. That is why the rays from P to H and H' to P' in Figure 3 are parallel.

Compound Lenses

A great advantage of the paraxial matrix method is the ease of dealing with systems having a large number of lenses. To start simply, consider the cemented doublet of Figure 10: a pair of lenses for which surface 2 of the first element and surface 1 of the second element match perfectly. The parameters of this doublet are given in the manner shown below. It is understood that the first entry under r is r_1, the second entry is r_2, and so on. Note that this doublet has only three surfaces; if the two parts were separated by an air space, producing an air-spaced doublet, then there would be four surfaces to specify. The values of the indices n' and the spacings t' are placed between the values of r. The constants of the doublet are then

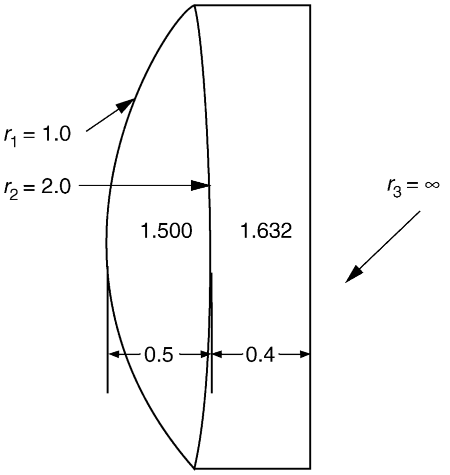

Figure 10 Specification for a cemented doublet.

r	n'	t'
1.0		
	1.500	0.5
−2.0		
	1.632	0.4
∞		

[44]

Note that surface 3, which is flat, has a radius of infinity. Numbering the vertices in the usual manner, the system matrix is

$$S_{31} = R_3T_{32}R_2T_{21}R_1 \qquad [45]$$

where

$$k_2 = (n_2' - n_2)/r_2 = (1.632 - 1.500)/-2.0 \qquad [46]$$

Approximations for Thin Lenses

An advantage of paraxial matrix optics is the ease of obtaining useful relations. For example, combining the several expressions for the Gaussian constant a, we obtain:

$$a = -\frac{n_1}{f} = \frac{n_2'}{f'} = k_1 + k_2 - \frac{k_1k_2t_1'}{n_1'} \qquad [47]$$

and using the definitions of k_1 and k_2 with the lens in air leads to

$$\frac{1}{f'} = (n_1' - 1)\left[\frac{1}{r_1} - \frac{1}{r_2} + \frac{(n_1' - 1)t_1'}{n_1'r_1r_2}\right] \qquad [48]$$

which is the lensmakers' equation. It tells how to find the focal length of a lens from a knowledge of its material and geometry. This derivation is much more direct than what you will usually find. This equation becomes simpler when we are dealing with thin lenses – those for which the third term in brackets can be neglected by assuming that the lens thickness is approximately zero. Then

$$\frac{1}{f'} = (n_1' - 1)\left[\frac{1}{r_1} - \frac{1}{r_2}\right] \qquad [49]$$

which is the form often seen in texts. We then realize that the original form is valid for lenses of any thickness, but the thin lens form can be used if the spacing is much less than the two radii. It is also seen that when the thickness is approximately zero, then

$$b = c = 1 \qquad [50]$$

and the unit planes have locations

$$l_H = l_H' = 0 \qquad [51]$$

Thus, they are now coincident at the center of the lens. A ray from any point on the object would pass undeviated through the geometrical center of the lens, as is often shown in the elementary books. These new values of the Gaussian constants give a system matrix of the form:

$$S_{21} = \begin{pmatrix} 1 & -1/f' \\ 0 & 1 \end{pmatrix} \qquad [52]$$

This matrix contains a single constant, the focal length of the lens, so that the preliminary design of an optical system is merely a matter of specifying the focal length of the lenses involved and using the resulting matrices to determine the object-image relationship. We also note that this matrix looks like a refraction matrix; the non-zero element is in the upper right-hand corner. This implies that the refraction–translation–refraction procedure that actually occurs can be replaced, for a thin lens, by a single refraction occurring at the coinciding unit planes. Ophthalmologists take advantage of the thin lens approximation by specifying focal lengths in units called diopters. The reciprocal of the focal length in meters determines its refraction power in diopters. For example, a lens with $f' = 10$ cm will be a 10 diopter lens. If two lenses are placed in contact, the combined power is the sum of the individual powers, for multiplying matrices of the above form gives the upper right-hand element as $(-1/f_1' - 1/f_2')$.

The Paraxial System for Design or Analysis

The material given in this article can serve as the basis for an organized way of taking the specifications of an optical system and using them to gain a full understanding of this system. We shall demonstrate it with a practical example which has some interesting features. Chemical engineers have long known that the index of refraction of a liquid can be determined by observing the empty bore of a thick glass tube with a telescope having a calibrated eyepiece and then measuring the magnification of the bore when the liquid flows through it. Consider a liquid with an index of 1.333 and a tube with bore radius of 3 cm, outer radius of 4 cm, and glass of index equal to 1.500. The first step is to specify the parameters of the optical system. These are the radii $r_1 = -3$, $r_2 = -4$ and the thickness $t_1' = 1$, regarding the tube as a concentric lens with an object to its left; the indices $n_1 = 4/3$ (fractions are convenient), $n_1' = 3/2$, $n_2' = 1$. Calculating the elements of the two refraction matrices and the translation matrix, the system matrix is

$$S_{21} = \begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 2/3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1/18 \\ 0 & 1 \end{pmatrix} \qquad [53]$$

Multiplying, the Gaussian constants are

$$a = 2/27, \quad b = 11/12, \quad c = 28/27, \quad d = -2/3 \qquad [54]$$

An important check is to verify that $bc - ad = 1$, as is the case here. The expressions given previously for the positions of the cardinal points then lead to the values

$$\begin{aligned} &l_H = 3/2, \qquad l_H' = 1/2, \qquad l_F = -16\ 1/2, \\ &l_F' = 14, \qquad l_N = -3, \qquad l_N' = -4 \end{aligned} \qquad [55]$$

These points are shown in Figure 11. Although the same scale has been used for both horizontal and vertical distances, this is usually not necessary; the vertical scale, which merely serves to verify the magnification, can be arbitrary. The entire liquid column is the object in question, but it is simpler to use a vertical radius as the object. Then the image, generated as described in all the previous examples,

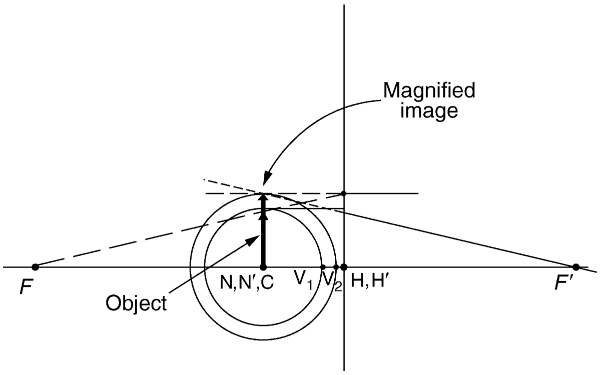

Figure 11 Ray tracing diagram for a thick-walled glass tube.

coincides with the object and it is enlarged, erect, and virtual. We confirm the image location shown by using the generalized Gaussian lens equation, giving

$$\begin{aligned} t' &= [(-2/3) + (28/27)(-3)/(4/3)]/[(11/12) \\ &\quad + (2/27)(-3)/(4/3)] \\ &= -4 \end{aligned} \qquad [56]$$

and this agrees with the sketch. The denominator of this expression was shown to be the reciprocal of the magnification $1/m$ with a value of 3/4, and m itself can be calculated from t' as $c - at' = 4/3$; this is another important check on the accuracy of the calculation; the sketch obeys this conclusion, and the purpose of this analysis is confirmed.

Reflectors

To show how reflecting surfaces are handled with matrices, consider a plane mirror located at the origin and normal to the z-axis (Figure 12). The ray which strikes it at an angle α_1 leaves at an equal angle, resulting in specular reflection. Since $k = 0$ for a flat surface, the refraction matrix reduces to the unit matrix and by eqn [9]:

$$n'_1 \alpha'_1 = n_1 \alpha_1 \qquad [57]$$

This contradicts Figure 12, since α_1 and α'_1 are opposite in sign.

If, however, we specify that

$$n'_1 = -n_1 \qquad [58]$$

then the difficulty is removed. Hence, reflections involve a reversal of the sign on the index of refraction, and two successive reflections restore the original sign. For a mirror, the refraction power is

$$k_1 = \frac{n'_1 - n_1}{r_1} = \frac{-1-(1)}{r_1} = -\frac{2}{r_1} \qquad [59]$$

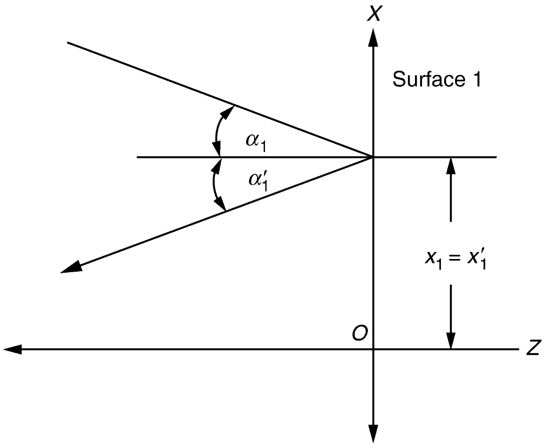

Figure 12 Specular reflection.

The system matrix for a single refracting surface thus becomes

$$S_{11} = R_1 = \begin{pmatrix} 1 & -k_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2/r_1 \\ 0 & 1 \end{pmatrix} \qquad [60]$$

with

$$a = -2/r_1, \qquad b = c = 1, \qquad d = 0 \qquad [61]$$

The connection between object and image can be expressed as

$$\begin{pmatrix} n'_1\alpha'_1 \\ x'_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ t'/n'_1 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -t/n_1 & 1 \end{pmatrix} \begin{pmatrix} n_1\alpha_1 \\ x_1 \end{pmatrix} \qquad [62]$$

and using the same procedure that was applied to eqns [28]–[31], the locations of the six cardinal points are

$$\begin{aligned} &l_F = -n_1 b/a, \quad l'_F = n'_1 c/a \\ &l_H = n_1(1-b)/a, \quad l'_H = n'_1(c-1)/a \qquad [63] \\ &\frac{l_N}{n_1} = \frac{(n'_1/n_1) - b}{a}, \quad \frac{l'_N}{n'_1} = \frac{c - (n_1/n'_1)}{a} \end{aligned}$$

so that for the spherical mirror:

$$l_F = \frac{(-1)(1)}{-2/r_1} = \frac{r_1}{2} = l'_F \qquad [64]$$

$$l_H = l'_H = 0 \qquad [65]$$

and

$$l_N = \frac{-1-1}{-2/r_1} = r_1 \qquad [66]$$

but

$$\frac{l'_N}{-1} = \frac{1-(-1)}{-2/r_1}, \quad l'_N = r_1 \qquad [67]$$

Since r_1 is negative, these equations show that the unit points lie on the vertex, the foci coincide halfway between the center of curvature and the

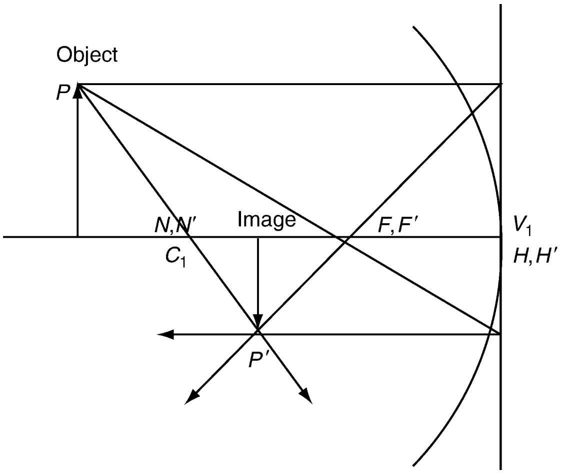

Figure 13 Ray tracing for a concave mirror.

vertex, and the nodal points are at this center. The perfect focusing is a consequence of the paraxial approximation. The true situation is presented in **Geometrical Optics:** Aberrations. Ray tracing for this mirror uses the procedure developed for lenses, which applies to optical systems of any complexity. Figure 13 shows an object to the left of the center of curvature. The ray from P parallel to the axis goes to H' and then through F', while the ray through F goes to H and then becomes parallel to the axis. Their intersection at P' produces a real, inverted image. The third set of rays represents something different: it requires a knowledge of the nodal point locations. The ray from P to N should be parallel to the ray from N' to P', by definition; in this example, they meet this requirement by being colinear.

As an example of the power of matrix optics to simplify the design or analysis of an optical system involving lenses and mirrors, consider an object that is 15 units to the left of a converging lens, with focal length $f' = 10$. A concave mirror, of radius $r_1 = 16$, is 20 units to the right of the lens. The refraction power of the mirror is $k_1 = (-1 - 1)/-16 = 1/8$ and using the thin lens form of the system matrix, the combined matrix is

$$\begin{pmatrix} 1 & -1/8 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 20 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1/10 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -3/2 & 1/40 \\ 20 & -1 \end{pmatrix} \qquad [68]$$

This gives the constants a, b, c, and d, from which we find that the image is $-4\ 4/9$ units to the left of the mirror, its magnification is $-8/9$, and it is inverted. Compare the simplicity of this procedure with the usual way of doing this calculation, which involves first finding the image in the lens and then using that image as a virtual object.

Conclusion

It has been shown that the use of paraxial matrices provides a simple approach to an understanding of the kind of optical systems that provide perfect images. Further details and derivations of expressions given here will be found in the Further Reading below.

See also

Geometrical Optics: Aberrations.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Aberrations

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The article **Geometrical Optics:** Lenses and Mirrors, describes image formation by components which are optically perfect. The images may be magnified or reduced and they may be erect or inverted, but they are otherwise faithful replicas of the object. This behavior is a consequence of ray tracing based on an approximate or paraxial form of Snell's law which we write as

$$n\theta = n'\theta' \qquad [1]$$

where n and n' are the indices of refraction on either side of an interface and θ and θ' are the angles of incidence and refraction. If, however, the following exact form of Snell's law:

$$n \sin\theta = n' \sin\theta' \qquad [2]$$

governs the ray behavior, as is the case when the rays have large inclinations with respect to the symmetry axis, then it is necessary to devise a more involved ray

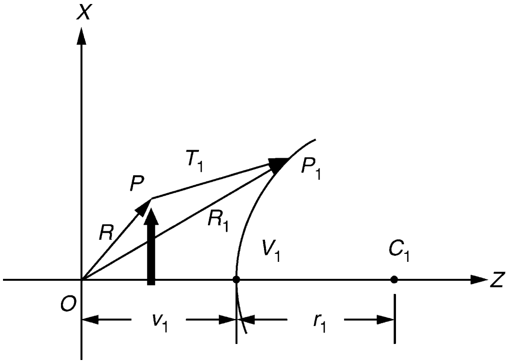

Figure 1 Ray tracing in the meridional plane.

tracing procedure. Figure 1, which shows the first surface of a lens, will be used to explain how this procedure works. A ray leaves the object point P and strikes the lens surface at the point P_1. Rays which lie completely in the plane ZOX are said to be meridional; this is the plane which passes through the symmetry axis, just as meridians on the Earth's surface are determined by planes through the geographic axis. By regarding the various distances in the figure as vectors, it is possible to start with the initial coordinates of the ray and with the direction-cosines which specify its slope and find the coordinates of the point P' where the ray strikes the lens surface. A derivation of the equation governing this process will be found in the text of A Nussbaum (see Further Reading). This derivation involves only elementary algebra and vector analysis, and will be found simpler than the usual textbooks treatment of ray tracing procedures. The equation is

$$T_1 = \frac{F}{-E + \sqrt{E^2 - c_1 F}} \qquad [3]$$

where

$$E = Bc_1/2 = c_1[(z - v_1)N + xL] - N \qquad [4]$$

and

$$F = Cc_1 = c_1\left[(z - v_1)^2 + x^2\right] - 2(z - v_1) \qquad [5]$$

Knowing the coordinates z and x of the starting point P of the ray, as well as its starting cosines N and L, plus the curvature c_1 or radius $r_1 = 1/c_1$ of the lens surface and the location v_1 of the surface's vertex, we can readily calculate the length T_1 of the ray. Then its components on the two axes give the coordinates of P'. Next, we need to find the slope of the ray after it is refracted at the lens surface, and this can be calculated from Snell's law combined with the behavior of the incident and refracted rays regarded as vectors. Again we invoke the reference cited above for a derivation of the equations giving the value of the direction-cosines after refraction; these expressions are

$$n_1' N_1' = K_1/c_1 + n_1 N_1 - K_1 z_1 \qquad [6]$$

and

$$n_1 L_1' = n_1 L_1 - K_1 x_1 \qquad [7]$$

where K_1 is known as the refracting power and has the definition

$$K_1 = c_1(n_1' \cos\theta' - n_1 \cos\theta) \qquad [8]$$

The ray can now be traced to the next surface knowing the lens thickness, and then to the image plane. These two sets of equations, when incorporated into a computer program of about two dozen lines, will trace meridional rays to any desired accuracy through an optical system containing an arbitrary number of lenses and spherical reflectors. As an example, consider a symmetric double convex lens with radii of 50 units at surface 1, −50 units at surface 2, a thickness of 15 units, and an index of 1.50. Rays parallel to the axis – one at 0.2 units above the axis, a second one at 2 units above the axis, and a third one at 20 units above the axis – are started well to the left and traced to the paraxial focal plane, which is at 47.368 units from the second surface. The ray that is 0.2 units above the axis will cross the focal plane at the axis; this is a paraxial ray, passing through the paraxial focal point F'. The ray which is 2 units above the axis crosses the focal plane slightly below the focal point, and is therefore not quite paraxial, while the ray which starts at 20 units above the axis falls well short of the focal point, intersecting the axis before it reaches the focal plane. This behavior indicates that the lens possesses a defect known as spherical aberration and these calculations imply that spherical aberration should be defined as the failure of nonparaxial rays to conform to the behavior of paraxial rays. This definition does not appear in optics texts, and the few definitions that are given appear to be incorrect. As clearly indicated in **Geometrical Optics:** Lenses and Mirrors, the focal point of a lens is strictly a paraxial concept and it is a constant, obtainable from the parameters of the lens or lens system. It is meaningless to speak of a focal point as defined by a pair of nonparaxial rays; this infinite set of ray-pairs determines a corresponding number of intersection points, all of which fall short of the true focal point F'. This situation is nicely illustrated by the next figure, which shows a large number of parallel rays, the central ones being paraxial and the others behaving differently. The rays which are very close to the axis will meet as expected at the focal point F'. The rays a little farther out – the intermediate rays – will cross the axis just a little to the left of F', and those near the edge of the lens – the *marginal* rays – will fall very short, as indicated. Rotating this diagram about the axis, we realize that spherical aberration produces a very blurry circular image at the paraxial plane. The three-dimensional envelope of the group of rays produced by this rotation is known as the caustic surface and the narrowest cross-section of this surface, slightly to the left of the paraxial focal plane, is called the circle of least confusion. If this location is chosen as the image plane, then the amount of spherical aberration can be reduced somewhat. Other methods of improving the image will be considered below. Figure 2 was obtained by adding a graphics step to the computer program described above. Figure 3 shows how to define spherical aberration quantitatively. Two kinds of aberration are specified in this figure. The place where the ray crosses the axis, lying to the left of F', is at a distance, measured from the paraxial focus, called the longitudinal spherical aberration (LSA), while the distance below the axis at the focal plane is the transverse spherical aberration (TSA). To reduce this aberration, note in Figure 2 that the amount of refraction at the two surfaces of the lens for marginal rays is not the same. Equalizing the refraction at the two surfaces will improve the situation. Suppose we alter the lens while keeping the index, thickness, and focal length constant; only the radii will change. This can be done if the new radii satisfy the paraxial equation giving the focal length in terms of the lens constants. The process of varying the shape of lens, while holding the other parameters constant, is called *bending the lens*. This can be easily accomplished by a computer program, obtaining what are known as optimal shape lenses, which are commercially available. To study their effect on spherical aberration, define the shape factor σ of a lens as

$$\sigma = \frac{r_2 + r_1}{r_2 - r_1} \qquad [9]$$

from which $\sigma = 0$ for a symmetric lens ($r_1 = -r_2$). Figure 4a shows the monitor screen for a symmetric lens, and Figure 4b illustrates the result of lens bending. This lens has had the curvature of the first surface raised and the second surface has been flattened to keep the focal length constant. Let us now consider what lens designers can do about spherical aberration. Brouwer in his book (see Further Reading) gives the specifications for a cemented doublet consisting of a double convex lens followed by a lens with two concave surfaces. The longitudinal spherical aberration of the front lens alone varies as shown in Figure 5. The way that the designer arranged for the marginal rays to meet at the paraxial focus, rather than falling short, was to recognize that a diverging lens following the front element would refract the rays away from – rather than closer to – the axis. When this second component is added, we would expect the behavior shown in Figure 6. Note that the value of LSA for the doublet has been enormously reduced, as indicated by the horizontal scale change. Spherical aberration can be reduced in more elaborate systems as part of the design process. Telescopes, in particular, make use of corrector plates which are placed in front of the large mirrors.

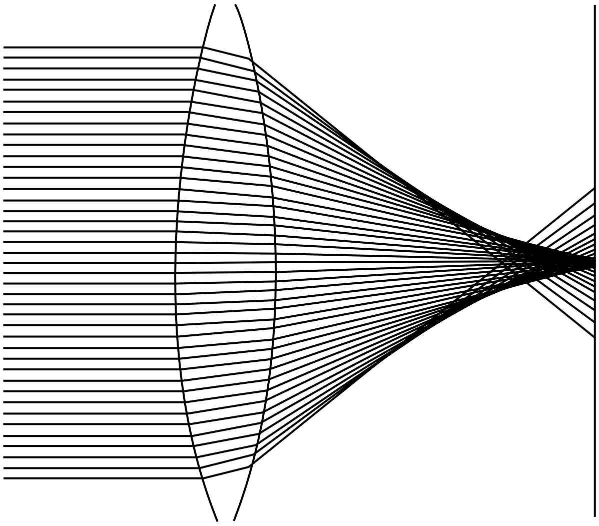

Figure 2 Spherical aberration for parallel meridional rays.

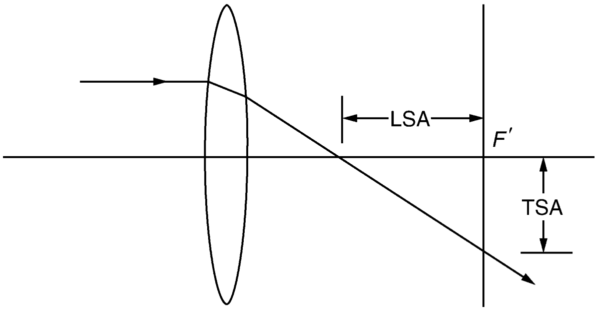

Figure 3 Definitions of transverse and longitudinal spherical aberration.

Coma

We have seen how to trace nonparaxial, meridional rays and found that spherical aberration appears because the intermediate and marginal rays do not behave like the central rays. We now wish to deal

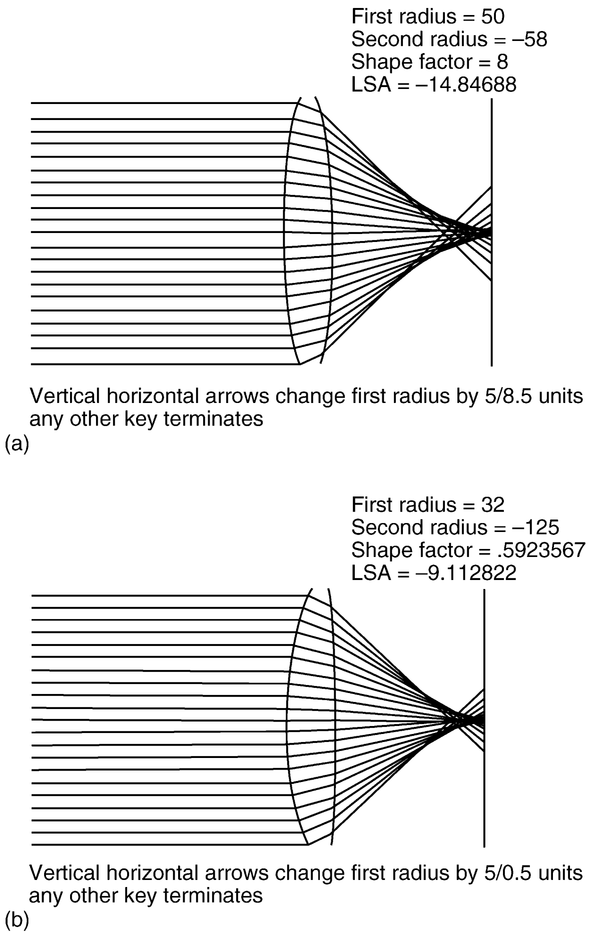
First radius = 50
Second radius = –58
Shape factor = 8
LSA = –14.84688
Vertical horizontal arrows change first radius by 5/8.5 units
any other key terminates
(a)
First radius = 32
Second radius = –125
Shape factor = .5923567
LSA = –9.112822
Vertical horizontal arrows change first radius by 5/0.5 units
any other key terminates
(b)

Figure 4 Lens bending as performed numerically.

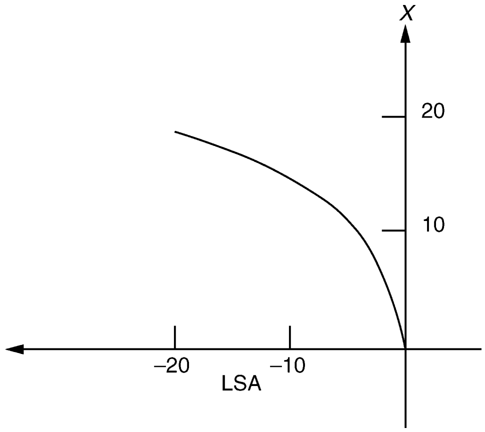

Figure 5 Spherical aberration for a single lens.

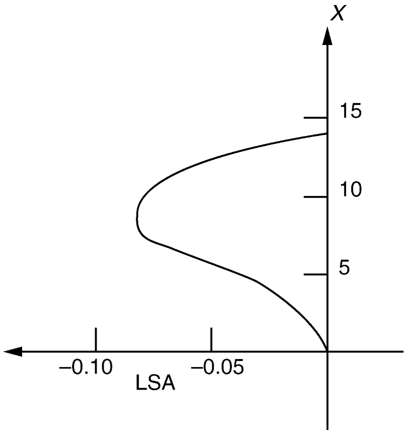

Figure 6 Reduction of spherical aberration by a doublet.

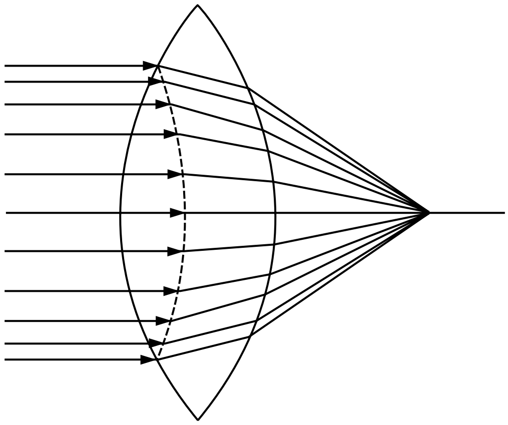

Figure 7 Configuration used to illustrate coma.

with rays that do not lie in a meridional plane; these are nonmeridional or skew rays. Calculating the behavior of such a ray would appear to be a complicated problem, but it turns out to be a simple extension of what has already been accomplished. Let a skew ray start at an arbitrary point in space with coordinates x, y, and z and let it have direction cosines L, M, and N with respect to OX, OY, and OZ. The points P and P_1 in Figure 1 are now out of the plane ZOX and when the vectors in this figure are expressed as the sum of three, rather than two, components, the equations given above for meridional rays acquire terms in y and M which have the same form as those for x and L. Now trace a set of skew rays chosen to lie on a cylinder (Figure 7) which is centered about the z-axis of a lens, so that all the rays are meridional. These rays are fairly far from the axis, so that they are also nonparaxial, but they all meet at a common image point, forming a cone whose apex is this image point. Then give this cylinder a downward displacement while holding fixed the intersection point of each ray with the dotted circle on the front of the lens. The tilting of the cylinder changes all its rays – except the top and the bottom rays – from meridional to skew. In addition, the cone on the image side will then tilt upwards; it should not change in any other way if the skew rays continue to meet at a well-defined apex. But this is not what happens in a simple lens, as will now be explained.

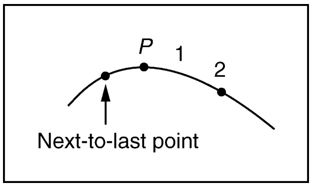

Figure 8 Starting a coma plot.

Let the top and bottom meridional rays meet at a point *P* in image space (Figure 8) and use this point to determine the location of an image plane. The skew ray just below the uppermost ray in Figure 7 will pass through this plane fairly close to the intersection point. Let us guess (and confirm later) that it will pass through a point slightly to the right of *P* and a little below it, as we observe this plane from the lens position. This is point 1 in Figure 8. The next skew ray on the front of the lens will strike the image plane still farther below and to the right, producing point 2 in the figure. As we continue to trace these rays down the front half of the circle, we realize that a ray almost at the bottom of Figure 7 will have to be very close to the lower meridional ray and that its image point is just below but to the left of *P*; this is the point labeled 'next-to-last' in Figure 8. All of these points join to form a closed curve. The back halves of the dotted circle and of the ray cylinder are not shown in the figure, but since we are again starting at *P* in Figure 8, the pattern will repeat itself. That is, we get a second closed curve which – by symmetry – should be identical to the first curve. This is a most unusual effect; double-valued phenomena are quite rare. To confirm this prediction, we use the skew ray equations, and trace the pattern formed by a set of concentric cylinders of different sizes. The appearance of the coma pattern generated by the computer is shown in Figure 9. Each half-circle on the front side of the lens produces a closed figure (an approximation to a circle) as an image. If coma were the only aberration (when it is called ideal coma), the plots would be the perfect, coinciding circles of Figure 10. The comet-like shape explains where the name comes from. We can picture the generation of these circles in the way shown in this figure. The ray through the geometrical center of the lens produces a point in the paraxial focal plane. The top and bottom rays (the meridional rays) from the circles of increasing size produce points which define image planes that are successively closer to the lens; this is a consequence of spherical aberration. Moving the top ray parallel to itself and along the half-circle generates the image plane pattern. The lines tangential to the coma circles make an angle of 60° with each other, as can be seen by measurement on the computer output. Most optics books look at coma in a way different from that given here. Coma has been defined by WJ Smith (see Further Reading) as the variation of magnification with aperture, analogous to the definition of spherical aberration as the variation of focal length with aperture. And as was previously mentioned, focal length and magnification are purely paraxial concepts, not applicable to the description of aberrations. The definition that comes from the arguments

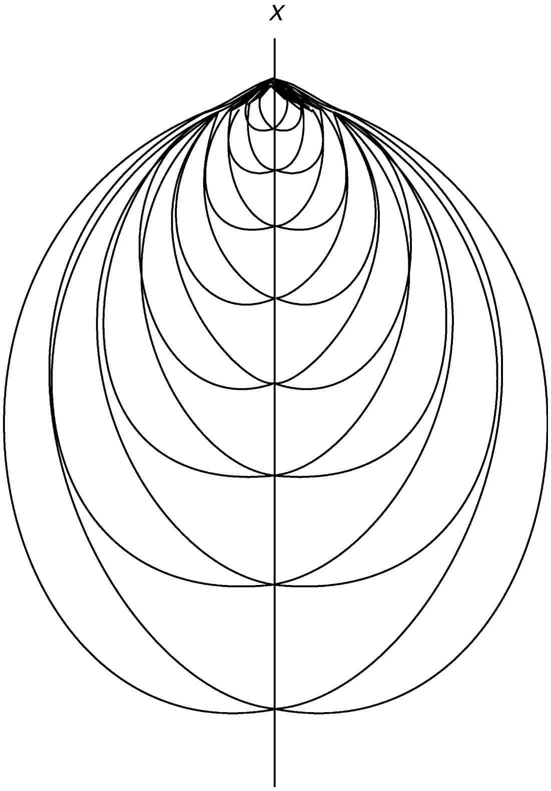

Figure 9 Coma as calculated numerically.

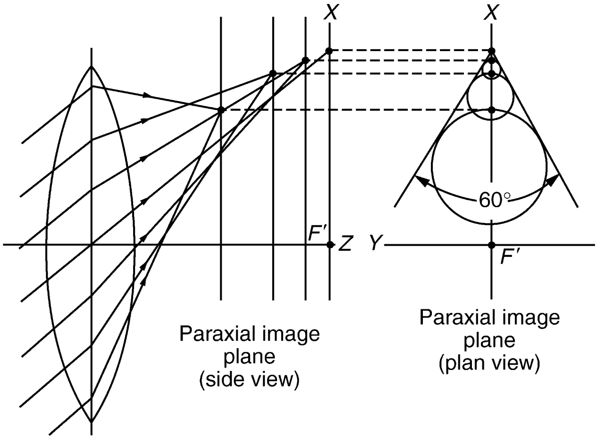
X
X
F′
Z
Y
F′
60°
Paraxial image plane (side view)
Paraxial image plane (plan view)

Figure 10 Customary way of illustrating coma.

given above is that coma is the failure of skew rays to match the behavior of meridional rays, just as spherical aberration is the failure of meridional rays to match the behavior of paraxial rays. Most texts show coma in its ideal form.

Astigmatism

We have shown that spherical aberration is the failure of meridional rays to obey the paraxial approximation and coma is the failure of skew rays to match the behavior of meridional rays. To see the connection between these two aberrations, consider the object point *P* of Figure 11. This very complicated diagram can actually be very helpful in understanding the third of the five geometrical aberrations. Let *P* be the source of a meridional fan: this is the group of rays with the top ray labeled PA and the bottom ray is PB. If the lens is completely corrected for spherical aberration, this fan will have a sharp image point P'_T lying directly below the z-axis. Now let *P* also be the source of a fan bounded by the rays PC and PD. This fan is at right-angles to the other one; we can think of these rays as having the maximum skewness possible. If the lens has no coma, this fan also will produce a sharp image point P'_s which, for a converging lens, will be farther from the lens than the tangential focal point. In other words, even though the lens has been fully corrected for both aberrations, the two corrections will not necessarily produce a common image. In the figure, the meridional fan is called a tangential fan and the skew fan is called a sagittal fan; these are the terms commonly used by lens designers. The failure of the sagittal and tangential rays to produce a single image in a lens corrected for both spherical aberration and coma is known as astigmatism. Astigmatism, coma, and spherical aberration are the point aberrations that an optical system can have. We have already noted that spherical aberration is the only one that can be associated with a point on the z-axis; the others require an off-axis object. Astigmatism is very common in the human eye; to see how it shows up, look at the sagittal image point P'_s of Figure 11. This point is a distance z_s, from the x,y-plane. A plane through this point will have a line image on it, the sagittal line, due to the tangential fan, which has already come to a focus and is now diverging. The converse effect, producing a tangential line, occurs at the other image plane. If we locate an image plane halfway between the two, then both fans contribute to the image and in the ideal case it will be a circle. As the image plane is moved forwards or backwards, these images become ellipses and eventually reduce to a sagittal or a tangential line, as shown in Figure 12. Because there are two different

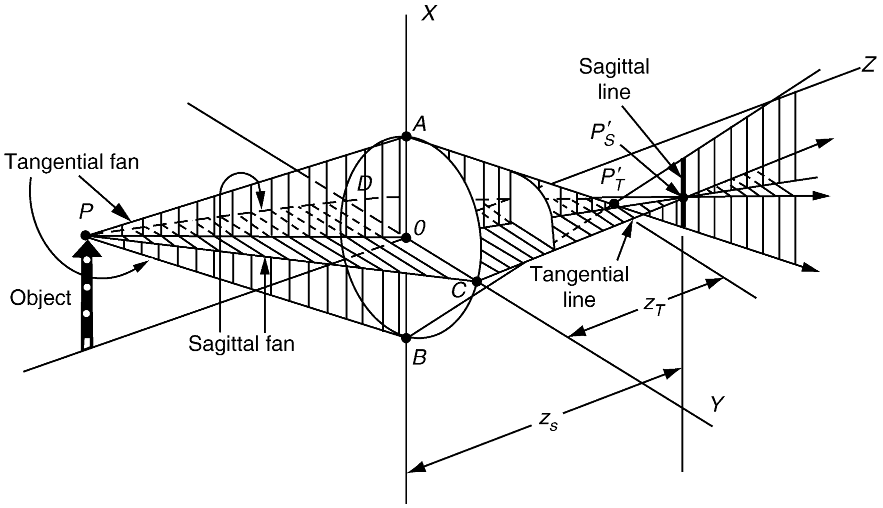

Figure 11 Tangential and sagittal fans displaying astigmatism.

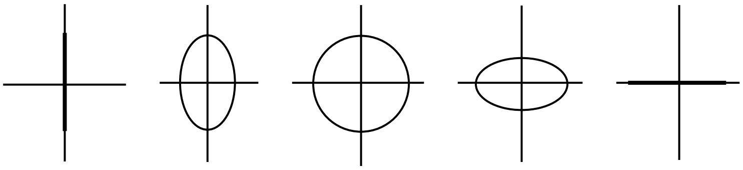

Figure 12 Astigmatic image patterns.

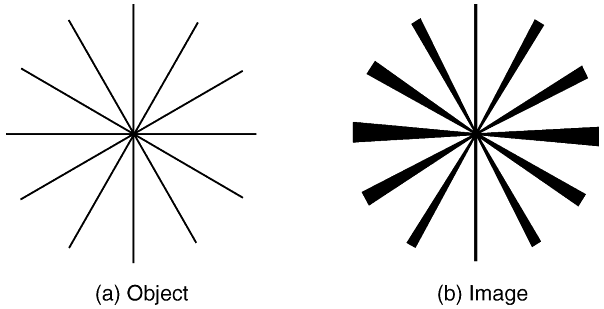

Figure 13 Image degraded by astigmatism.

image planes, an object with spokes (Figure 13) will have an image for which the vertical lines are sharp and the others get gradually poorer as they become more horizontal, or vice versa, depending on which image plane is chosen. Eye examinations detect astigmatism if the spokes appear to go from black to gray. This example of the effect of astigmatism was often based on a radial pattern of arrows, which explains the origin of the word sagittal, derived from the Latin ‘sagitta’ for arrow.

Curvature of Field and Distortion

Having covered the three point aberrations, there are two lens defects associated with extended objects. If we move the object point P in Figure 11 closer to or farther from the z-axis, we would expect the positions of the tangential and sagittal focal planes to shift, for it is only when the paraxial approximation holds that these image points are independent of x. Hence, we obtain the two curves of Figure 14, which shows what astigmatism does to the image of a two-dimensional object. If the astigmatism could be eliminated, the effect would be to make these curved image planes coincide, but we have no guarantee that the common image will be flat, or paraxial. The resulting defect is called Petzval curvature or curvature of field. For a single lens, the Petzval surface can be flattened by a stop in the proper place, and this is usually done in inexpensive cameras. Petzval curvature is associated with the z-axis. If we take the object in Figure 11 and move it along the y-axis, then all rays leaving it are skew and this introduces distortion, the aberration associated with the coordinates normal to the symmetry axis. Distortion is what causes vertical lines to bulge outward (barrel distortion) or inward (pincushion distortion) at the image plane, as shown in Figure 15. A pinhole camera will have no distortion, since there are no skew rays, and a single thin lens with a small aperture will have very little. Placing a stop near the lens to reduce astigmatism and curvature of field introduces distortion because, as shown in Figure 16a, the rays for object points far from the axis are limited to off-center sections of the lens. The situation in this figure corresponds to barrel distortion; placing the stop on the other side of the lens produces pincushioning. Distortion can be

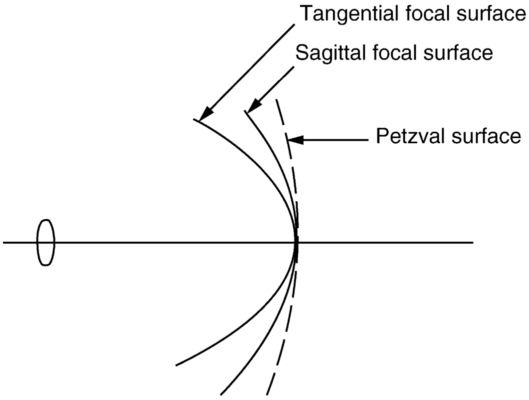

Figure 14 Petzval of field curvature.

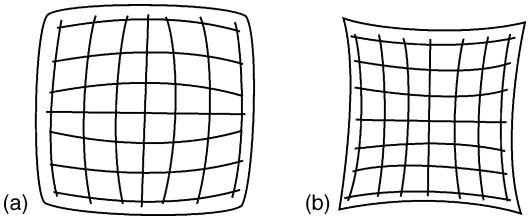

Figure 15 Barrel and pincushion distortion.

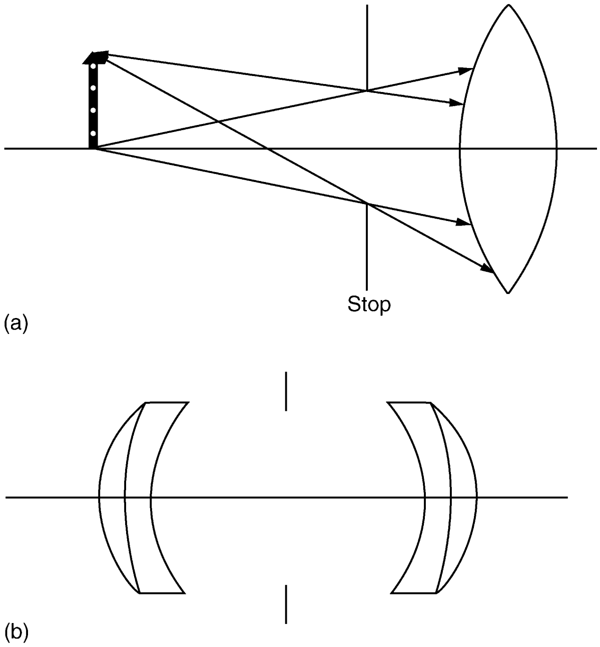

Figure 16 (a) Distortion-creating stop. (b) Symmetric-component lens that reduces distortion.

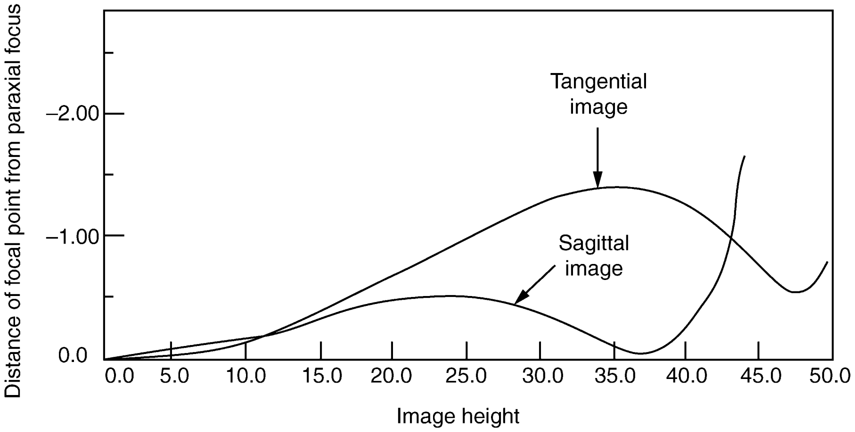

Figure 17 Reduction of astigmatism through design.

therefore reduced by a design which consists of two identical groupings (Figure 16b), with the iris diaphragm placed between them; the distortion of the front half cancels that of the back half. Figure 17 shows the output for a calculation using approximate formulas rather than exact ray tracing. Note that the designer has arranged for astigmatism to vanish at the lens margin. This example indicates the nature of astigmatism in a system for which the coma and the spherical aberration are low but not necessarily zero; it is the failure of the tangential and sagittal image planes to coincide.

Nonspherical Surfaces

So far, we have considered optical systems that use only spherical or plane reflecting and refracting surfaces. However, many astronomical telescopes use parabolic and hyperbolic surfaces. In recent years, nonspherical or aspheric glass lenses have become more common because they can be mass produced by computer-controlled machinery and plastic lenses can be molded with very strict tolerances. Ray tracing procedures given above can easily be extended to the simplest kind of aspheric surfaces: the conic sections. Iterative procedures may be applied to surfaces more complicated than the conic sections. A well-known example is the curved mirror; a spherical reflecting surface will have all the aberrations mentioned here, but a parabolic mirror will bring parallel rays to a perfect focus at the paraxial focal point, thus eliminating the spherical aberration. Headlamp reflectors in automobiles are a well-known example. A plano-convex lens with this defect corrected will be one whose first surface is flat and whose second surface is a hyperbola with its eccentricity equal to the index of refraction of the glass composing the lens. One form of astronomical telescope, which has a large primary concave mirror to collect the light and reflect it back to a smaller secondary convex mirror, is the Ritchey–Chrétien design of 1924. Both mirrors are hyperbolic in shape, completely eliminating spherical aberration. This is the design of the Hubble space telescope. The two-meter diameter primary mirror, formed by starting with a spherical mirror and converted to a hyperbola by computer-controlled polishing, was originally fabricated incorrectly and caused spherical aberration. From a knowledge of the computer calculations, it was possible to design, build, and place in orbit a large correcting optical element which removed all the aberrations from the emerging image.

Conclusion

The three-point aberrations and the two extended object aberrations have been defined and illustrated. Simple ways of reducing them have been mentioned, but the best way of dealing with aberrations is in the process of lens design. In addition, there is a completely separate defect known as chromatic aberration. All the above calculations are based on ray tracing for a single wavelength (or color) of visible light. However, the index of refraction of a transparent material varies with the wavelength of the light passing through it, and a design good for one value of the wavelength will give an out-of-focus image for any other wavelength. It has been known for many years that an optical system composed of two different kinds of glass will provide a complete correction for two colors – usually red and blue – and other colors will be partially corrected. Such lenses are said to be achromatic. For highly demanding applications, such as faithful color reproductions

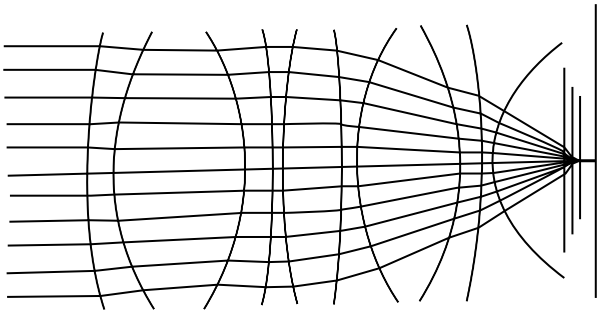

Figure 18 High-power high-resolution microscope objective.

of works of art, it is possible to design elaborate and expensive lenses which correct for yellow light as well. This kind of optical system is said to be apochromatic. With perfect imaging at the two ends and the middle of the visible spectrum, the color error elsewhere is virtually nonexistent. As an example of a design which is free of all aberrations – geometric and chromatic – Figure 18 shows the ray trace for a high-power aberration-free and apochromatic microscope objective. Note that the trace shows parallel rays brought to a focus inside the cover glass; in use, the light source would be a point on the slide and the light would converge at the eyepiece.

See also

Geometrical Optics: Lenses and Mirrors; Prisms.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Laikin M (1990) *Lens Design*. New York: Marcel Dekker.

Nussbaum A (1998) *Optical System Design*. Upper Saddle River, NJ: Prentice-Hall.

Nussbaum A and Phillips RA (1976) *Contemporary Optics for Scientists and Engineers*. Englewoods Cliffs, NJ: Prentice Hall.

Smith WJ (1990) *Modern Optical Engineering*, 2nd edn. New York: McGraw-Hill Book Co.

Prisms

A Nussbaum, University of Minnesota, Minneapolis, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Prisms are solid structures made from a transparent material (usually glass). Figure 1 shows the simplest type of prism – one that has many uses. The vertical sides are two identical rectangles at right angles to each other and a larger third rectangle is at 45° to the other two. The top and bottom are 45°–45°–90° triangles. The most common applications of prisms are:

1. To bend light in a specified direction
2. To fold an optical system into a smaller space
3. To provide proper image orientation
4. To combine or split optical beams, using partially reflecting surfaces
5. To disperse light in optical instruments such as spectrographs.

Single Prisms as Reflectors

A prism like that of Figure 1 can act as a very effective reflector; it is more efficient than a mirror since there is no metal coating to absorb light. Figure 2 shows a light ray entering the prism at right angles to one of the smaller sides, reflected at the larger side, and emerging at right angles to the other small side. The path shown is determined by Snell's law. This law has the form:

$$n \sin \theta = n' \sin \theta' \qquad [1]$$

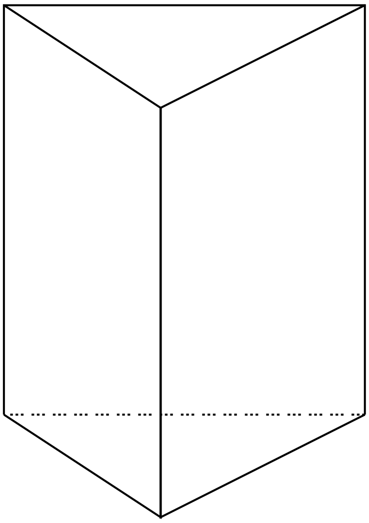

Figure 1 A 45–45–90 degree prism.

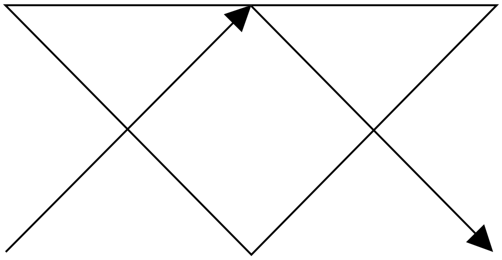

Figure 2 Total internal reflection by a prism.

This indicates that a ray of light making an angle θ to the normal at the interface between media of indices of refraction n and n', will be refracted and the angle to the normal becomes θ'. Notice that the light ray strikes the surface where the reflection takes place at an angle of 45° to the normal; this angle is too large to permit it to emerge from the prism after refraction. To understand why, we recognize that when the ray is able to leave the glass, Snell's law requires that the emerging angle be larger than the incident angle. The largest possible incident angle – known as the critical angle – will correspond to an emerging angle of 90°; that is, this ray will be parallel to the interface. For an index of 1.50, as an example, the critical angle θ_c satisfies Snell's law in the form:

$$1.5 \sin \theta_c = 1.0 \sin 90° \qquad [2]$$

from which $\theta_c = 41.8°$. The ray shown in the figure will therefore remain in the glass. Furthermore, since the indices appearing on both sides of Snell's law are now those of the glass, then the incident and refracted angles must be identical, and the prism reflects the ray in the same way that a mirror would. This phenomenon is known as total internal reflection.

This prism has an important application as an optical component of a fingerprint machine, whose action depends on modifying the total internal reflection which normally occurs. Figure 3 shows the prism and the reflecting behavior of Figure 2, with a person's finger pressed against the long side. When the ray strikes a place on the surface corresponding to a valley in the fingerprint pattern, the index at this region is that of air, and the ray is reflected in the usual manner. But a ray striking a portion of the prism which lies immediately below a raised section of the skin will make an angle with the normal which is now smaller than the critical angle, since the index is much greater than that of air, so that the ray will cross the interface and be absorbed by the finger. The reflected beam, with the fingerprint pattern sharply reproduced, is sent to a detector and the resulting signal goes to a computer, where it is processed and used to drive a printer. The fingerprint cards generated by this kind of equipment are sharper than those done by the traditional smeared-ink method. It is possible to observe the phenomenon just described by looking at your fingers while holding a glass of water – the fingerprints will show up in an enhanced manner. In addition to the better quality, the manufacturer of one version of this equipment has incorporated a feature in their computer program which can detect attempts by the person being fingerprinted to degrade the image.

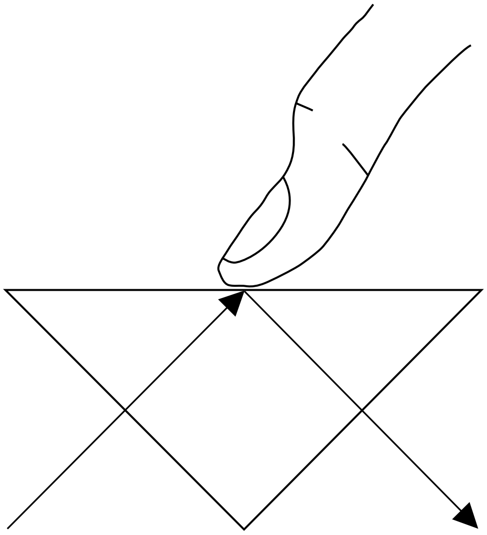

Figure 3 Use of a prism to produce a fingerprint.

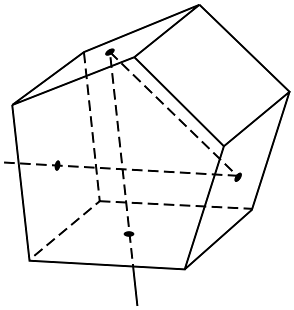

Figure 4 A pentaprism.

A more elaborate single prism – one with five sides – is the pentaprism of Figure 4, used extensively in single lens reflex cameras. This prism does not take advantage of the total internal reflection process, as described above. Below the bottom face is a mirror at a 45° angle. Light from the camera's lens is reflected upward and emerges from the left-hand face where the scene being photographed is observed by the eye.

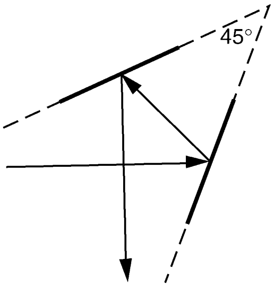

Figure 5 Action of coated faces of a pentaprism.

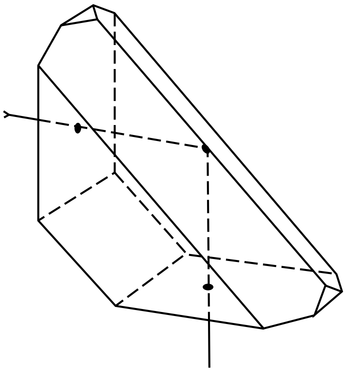

Figure 6 Amici prism.

Since there are two reflections, the reversal caused by a single mirror is eliminated. The camera's mirror is hinged and when the exposure button is pressed, it flies up and the incoming light can reach the film. The two reflecting faces, which lie at an angle of 45° to one another (Figure 5), are coated. To see why this is necessary, note that the two angles in the isosceles triangle formed by the light ray and the face extensions must equal (180° − 45°)/2 or 67.5°. This ray must then lie at an angle of 90° − 67.5° to the normal, or 22.5°, which is much less than the critical angle needed for reflection.

There are at least a dozen other forms of prism which have been designed and used. For example, the Amici prism (Figure 6) is a truncated right-angle prism with a roof section added to the hypotenuse face. It splits the image down the middle and thereby interchanges the right and left portions. These prisms must be very carefully made to avoid producing a double image. One use is in telescopes to correct the reversal of an image.

Double Prism Reflectors

Another extensive application of prisms is their use in pairs as components of telescopes and binoculars. The most widely used is a pair of Porro prisms: two right angle prisms arranged as shown in Figure 7. The double change in direction has two effects; it eliminates image reversal and it shortens the spacing between objective and eyepiece lenses, making the binoculars lighter and more compact. It is possible to calculate how light rays behave in Porro prisms with a ray tracing procedure. This method involves the use of 4 × 4 matrices and its development is rather complicated. A simple understanding of the way light passes through this optical component can be based on Figure 8, which shows the two prisms separated for clarity. An arbitrary ray (the heavy line) coming from the object will be reflected at a 45° angle at each of the four faces that it strikes. Using two line segments at right angles as an object, the ray from the top of the vertical arrow and the ray from the end of the horizontal arrow will behave as indicated. The image is then identical to the object but rotated by 180°. Since the optics of the binoculars – the combined effect of the objective and the eyepiece – result in an inverted image, the Porro prism corrects this problem.

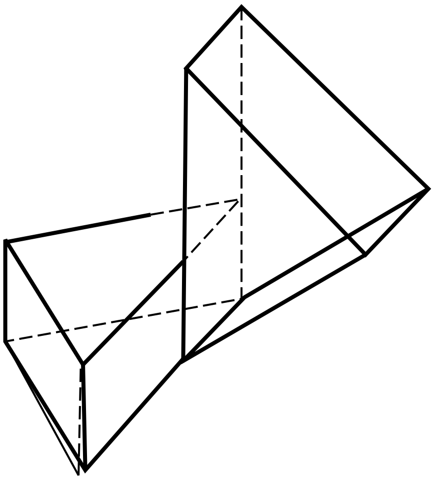

Figure 7 A pair of Porro prisms.

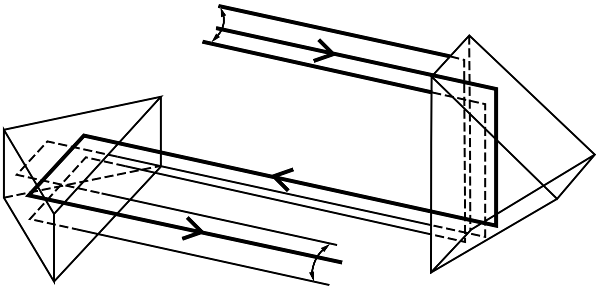

Figure 8 Action of Porro prisms.

Prisms as Instruments

Prisms are the crucial optical element in many laboratory instruments. To show how they can be used to measure the index of refraction of transparent materials – solids or liquids – consider the prism of Figure 9, with apex angle A. Light comes to the prism at an incident angle i, enters the prism at the angle of refraction r, crosses the prism and is refracted a second time when it emerges, the associated angles being designated as r' and i'. When the ray leaves the prism, it is traveling in a direction which represents an angular deviation δ from its original orientation. This deviation is the sum of the individual changes in direction at each face, or

$$\delta = (i - r) + (r' - i') \qquad [3]$$

For the triangle formed by the projected normals, which meet at the angle A:

$$r = r' + A \qquad [4]$$

Hence

$$\delta = i + i' - A \qquad [5]$$

The minimum value of the deviation is an important property of the prism; we find it by differentiation to obtain:

$$\mathrm{d}\delta = \mathrm{d}i + \mathrm{d}i' = 0 \qquad [6]$$

Differentiation of the form of Snell's law valid at each interface, namely:

$$\sin i = n \sin r, \quad \sin i' = n \sin r' \qquad [7]$$

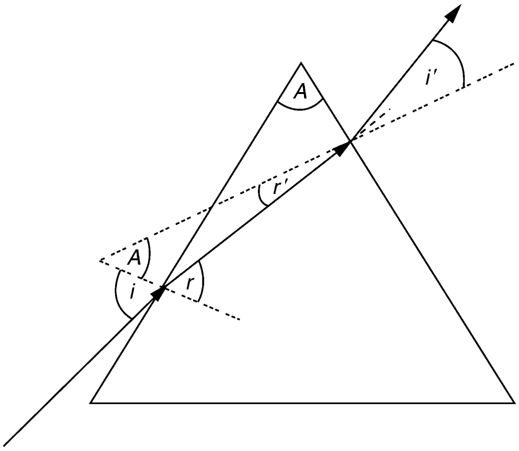

Figure 9 Calculation of prism deviation.

gives

$$\mathrm{d}i = n \cos r \, \mathrm{d}r / \cos i \qquad [8]$$

$$\mathrm{d}i' = n \cos r' \, \mathrm{d}r' / \cos i'$$

$$= n \cos r' \, \mathrm{d}r / \cos i' \qquad [9]$$

where $\mathrm{d}r'$ can be replaced by $\mathrm{d}r$, as obtained from the equation connecting r, r', and A. It now follows that:

$$\mathrm{d}\delta = n\Big[\big\{\cos r/\sqrt{(1 - n^2 \sin^2 r)}\big\} - \big\{\cos(A - r)/\sqrt{(1 - n^2 \sin^2(A - r)}\big\}\Big]\mathrm{d}r = 0 \qquad [10]$$

from which

$$\cos^2 r\{1 - n^2 \sin^2(A - r)\} = \cos^2(A - r)\{1 - n^2 \sin^2 r\} \qquad [11]$$

or:

$$\cos^2 r = \cos^2(A - r) \qquad [12]$$

and finally:

$$r = A/2, \quad r' = A/2, \quad i' = i \qquad [13]$$

The differentiation thus leads to the conclusion that the ray enters and leaves the prism in a symmetrical manner. The amount of deviation is now:

$$\delta = 2i - A \qquad [14]$$

or:

$$i = (\delta + A)/2 \qquad [15]$$

and:

$$n = \sin i / \sin r = \sin\{(\delta + A)/2\} / \sin(A/2) \qquad [16]$$

Using this expression to calculate n as a function of δ, it is found that δ is a minimum, so that a measurement of the minimum deviation produced by a prism gives the value of its index of refraction in a direct and accurate way. This method can also be used for liquids; a triangular glass cell with walls of reasonable thickness is made very accurately and used to measure δ. The refraction due to the walls does not affect the measurement, since the refraction occurring at the entrance wall is equal but opposite to that at the exit wall.

The description of prism behavior given so far assumes that we are dealing with monochromatic light; that is light of a single wavelength or single

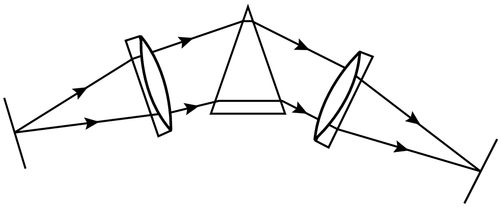

Figure 10 A prism spectrometer.

color in the visible spectrum. However, the index *n* of a prism varies with wavelength and white light passing through a prism will undergo dispersion – each component of a beam will emerge at a different angle and can be observed. The prisms that are part of expensive glass chandeliers display such an effect. This property of a prism is used in spectrometers – instruments which can measure the wavelength of light. Figure 10 shows a schematic diagram of this kind of instrument of high quality. Note that the light is brought into and taken out of the spectrometer with doublets. As shown in **Geometrical Optics:** Aberrations, doublets can be designed to be achromatic – that is, they focus light throughout the spectrum in approximately the same way. Various kinds of spectrometers have been put into use, with the ability to resolve closely spaced wavelengths being a crucial factor in the design.

See also

Geometrical Optics: Aberrations; Lenses and Mirrors.

Further Reading

Brouwer W (1964) *Matrix Methods in Optical Instrument Design*. New York: WA Benjamin.

Longhurst RS (1967) *Geometrical and Physical Optics*, 2nd edn. New York: John Wiley & Sons.

MIL-HDBK-141 (1962) *Military Standardization Handbook*. Washington, DC: US Government Printing Office.

H

HOLOGRAPHY, APPLICATIONS

Contents

Art Holography
High-Resolution Holographic Imaging and Subsea Holography
Holographic Recording Materials and Their Processing

Art Holography

A Pepper, Shearwater Foundation, Nottingham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography captures the phase and amplitude of light reflected from, or transmitted through, an object. In visual terms it captures 'true' three-dimensional space and any objects which are placed within that space. This is very different from the various optical illusions, including perspective drawing and painting, which 'trick' the eye into thinking it is seeing a dimensional image on a flat surface. Not surprisingly, artists have been intrigued by display holography since reports about it became more public in the early 1960s.

Since much of the information about the holographic process was originally published in professional scientific and optics journals, which the artistic community rarely read, or had access to, the transfer of practical information from optical scientist to visual artist was slow. As less technical reports about holography began to appear in publications, such as *Scientific American*, and art journals, such as *Leonardo*, more artists became aware of the possibilities holography offered. It is now over 50 years since its invention by Dennis Gabor, and art holography can be found in public and private collections around the world, ranging from major museums, through academic art galleries to large individual collections.

For the purpose of this article, holography can be defined as a section of a broad optical field, encompassing display holograms produced by artists, creative individuals, and groups. These works are often produced through independent or self-driven research, but can also be the result of a commissioned project. Holographic art works tend not to be produced specifically for commercial gain (although some do sell for large amounts and can be editioned in the same way prints or photographs, to increase their salability or make them more accessible to a wider audience). It is incredibly difficult to assess the number of artists working in this field today, as no specific research has been carried out, but it can be estimated as between 50–100 individuals worldwide, which includes those who have shown work in curated group and solo exhibitions. This relatively small number has made an enormous impact and includes artists who have made holography their main medium of expression, as well as those, more established in other media, who have extended their creative vocabulary by embracing the holographic process. Salvador Dali for example, had holograms made of his sculptural compositions in the early 1970s, and Bruce Nauman used the process to capture a series of self-portraits in the 1960s. Apart from these well known and established contemporary artists, much of the pioneering work has been driven by individuals who have concentrated on holography as their medium of choice.

Some artists build their own optical studio in which to experiment in a hands-on environment, some collaborate with scientific and research establishments, where the work, facilities and expertise is shared, and others design and supervise the production of their work which is produced for them by a commercial company. All of them produce a unique contribution to the visual arts.

Early Exploration

Holography offers a unique method of recording and reproducing three dimensions on a flat surface without the need for special glasses (for stereoscopic viewing), lenses (such as lenticular viewing) or linear perspective which, since the Renaissance, has been our main method of tricking the eye into believing that an image recedes into the distance and displays three dimensions.

Many of the very early works produced by artists concentrated on the obvious still life nature of the medium and explored the unusual phenomenon of objects 'frozen' in holographic space. Almost all of these early works were recorded using continuous wave helium neon lasers, either in scientific/research facilities, which artists had arranged to use, or in collaboration with established scientists and optical imaging researchers.

As early as 1969, British artist Margaret Benyon was experimenting with this new medium. Her laser transmission hologram 'Picasso', is a recording of images taken from Pablo Picasso's painting 'Demoiselles D'Avignon'. Here Margaret Benyon took a copy of the painting, cut out the characters, placed them in a three-dimensional spatial arrangement, with real objects, and then made a hologram of this 'scene'. Picasso attempted to show, through many of his paintings, multiple views of the same subject on a two-dimensional surface, one of the main thrusts of Cubism in the late twentieth century. Benyon achieved this in 'real space' by recording a hologram of the flat reproductions. By doing so, she demonstrated that holography automatically achieved what Picasso, and many other key contemporary visual artists, had been trying to do for years.

Holographic Black Holes

It has been known, from the very early days of holographic research that, when using a continuous wave laser to record a hologram, if any object in the scene moves more than half the wavelength of light, no interference pattern will be produced and therefore no recording of the object will take place. Unlike photography, where if an object moves during the exposure, it will become blurred, in holography it will simply not be recorded. Clearly, if no recording can be made, optical researchers working with holography, while noting the phenomenon, would not necessarily feel the need to make holograms which would fail. Several artists, on the other hand, not confined by research deadlines, have acknowledged this strange phenomenon and exploited this limitation of the holographic process. One of the very early examples is 'Hot Air' (Figure 1), a laser transmission hologram by Margaret Benyon and now in the Australian National Gallery Collection. Produced in 1970, it shows a three-dimensional composition of a jug, cup, and the artist's hand. The jug and cup were filled with hot water and during the recording of the holographic plate, with a continuous wave laser, the heat from the hot water disturbed the air above these objects. This tiny amount of movement resulted in the area above the jug and cup not being recorded. Similarly, the artist's hand moved slightly during the exposure and therefore did not record. Strangely, when viewing the resulting laser transmission hologram, the hand is visible. What is displayed is a holographic hand-shaped black hole, the three-dimensional recording of the space once occupied by the hand; the dimensional recording of the absence of an object.

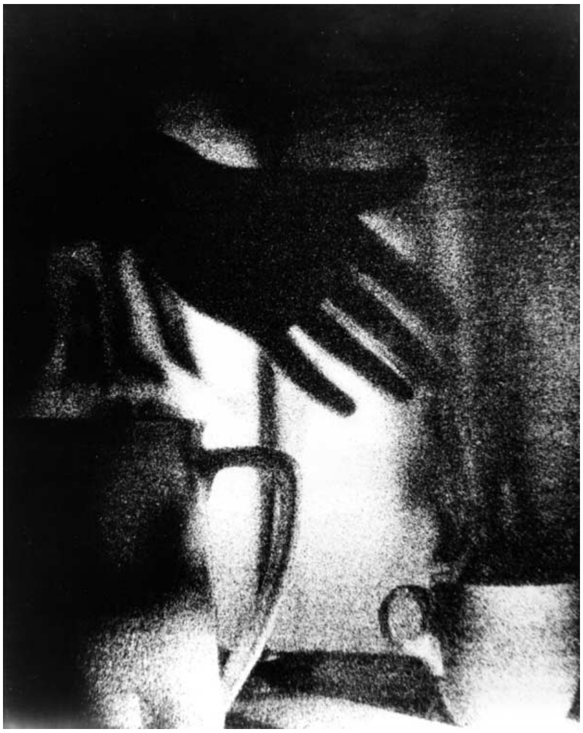

Figure 1 'Hot Air', Margaret Benyon 1970. Laser transmission hologram. Copyright © 1970 M. Benyon. Reproduced with permission.

> Artists in display holography place continual demands on the technology. Their natural tendency is to break rules. I believe that most successful holographic artists belong to this 'suck it and see' category of experimentalists. On hearing about the 'rules' of holography, they break them, just to see what happens. (Margaret Benyon, PhD Thesis, 1994).

Benyon started by breaking the rules and has continued to develop from these early pioneering laser transmission works, which helped establish a visual 'vocabulary' including many important social and political issues.

Other artists have exploited the 'limitations' of the holographic process to produce visually stunning works which question how we look at the three-dimensional world and question what is 'real'. Rick Silberman (USA) developed a shadowgram technique where the hologram is made of a laser illuminated ground glass screen. Objects are placed in front of this luminous surface and their three-dimensional silhouette is recorded. The viewer of the hologram is looking at the space where the object used to be. A pivotal work in this field is 'The Meeting' 1979 (Figure 2), a white light reflection hologram which had been recorded in such a way as to cause the three-dimensional shadow of a wine glass to protrude out of the holographic plate and be viewable several centimeters in front of it, between the observer and the holographic emulsion. The actual wine glass used in the recording has been broken and what remains (the base, stem, and small part of the bowl of the glass) is placed on a platform attached to the holographic plate. When viewing the work head on, the missing parts of the broken glass are recreated, in three dimensions, by the holographic shadow. When viewed obliquely, from the side for example, only the broken (real), glass is visible. Not only does this piece have an element of performance in the recording, breaking, mounting, and reconstruction of the wine glass, it also explores questions of visual perception and concepts of reality.

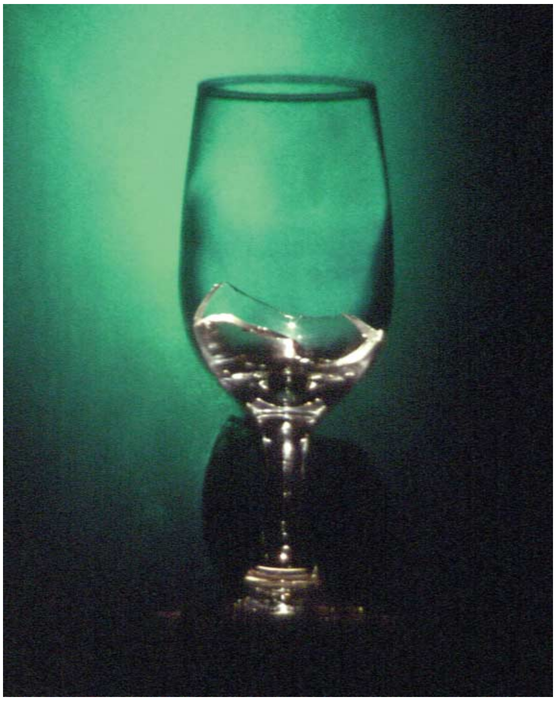

Figure 2 'The Meeting', Rick Silberman 1979. White light reflection hologram (shadowgram). Copyright © 1980 ATP. Reproduced with permission.

The recording of the shadow of an object, or objects, has now become commonplace in creative holography, with many artists employing techniques to mix different objects together, change their color and rearrange them in space. One advantage of this technique is that dependence on absolute stability of the objects used is reduced. The objects are 'encouraged' to move during the recording process and so create 'negative' images. Steve Weinstock (USA), Marie-Andrée Cossette (Canada), Michael Wenyon and Susan Gamble (USA/UK) have all contributed to this area in some of their early works.

Public Display

As the technology and optical processes of holography have developed, so greater flexibility, and public accessibility, have emerged. The single most important aspect has been the ability to display holograms in white light.

Emmett Leith and Uris Upatnieks invented laser transmission holography in 1964. They demonstrated the practicalities of Gabor's original invention, by utilizing laser light, to produce three-dimensional images. To display these recorded images, laser light needed to be shone through the holographic plate, at the original angle used by the recording reference beam (off-axis). The results were breathtaking, with full visual parallax in all directions. Although it was the laser light which made the recording and reconstruction of these unusual images possible, the laser was also a limiting factor. Each time a hologram was displayed, a laser was required. This relegated the technique to displays in scientific conferences, laboratories, or other locations with optical equipment. What was needed was a technique allowing holograms to be displayed without the need for this expensive and dangerous laser light. White light holography was the answer.

Although a laser is still required to record the original hologram, by using a copying technique, a restricted aperture or one step process, the resulting holograms can be reconstructed using a small source of white light such as a spotlight or even direct sunlight. As expensive lasers were no longer needed to display holograms, large public exhibitions could be mounted. Gallery and museum spaces could include holography in their exhibitions and individuals could buy and display holography in domestic locations. This technique also benefited the commercial display market, where holograms could now be used for advertising and promotional events.

Stephen Benton (USA) (1942–2003), inventor of white light transmission (rainbow) holography in 1968, made a major contribution to the use of the medium by artists. His technique not only allowed holograms to be viewed in white light, but they were extremely bright and offered the possibility of creating multicolor images. His 'Crystal Beginnings' 1977, (Figure 3), shows just how much space can be created with very simple points of light. Benton even taught master classes for artists so that he could explain how his technique functioned and explain the mathematics involved. He also worked with several artists to help them realize their individual projects.

Harriet Casdin-Silver (USA) was one of the first pioneering artists to work with Benton and produced several key works now seen as integral to the art of holography. Their early collaboration 'Cobweb Space', 1972, sculpted pure light and was one of the very first art holograms to be displayed using Benton's rainbow hologram technique. A later work, 'Equivocal Forks', 1977, exists in different formats; as a laser transmission hologram showing the pseudoscopic display of domestic folks which protrude, inside out, from the holographic plate, and as a number of white light transmission holograms, of the same image, which were shown in an outdoor installation mounted by The Center for Advanced Visual Studies, part of the Massachusetts Institute of Technology. This demonstrated to a wide, nonscientific, audience, that holograms could be accessible, easily illuminated and weatherproof. Casdin-Silver has continued to produce art holography and has, most recently, worked with large-scale portraiture.

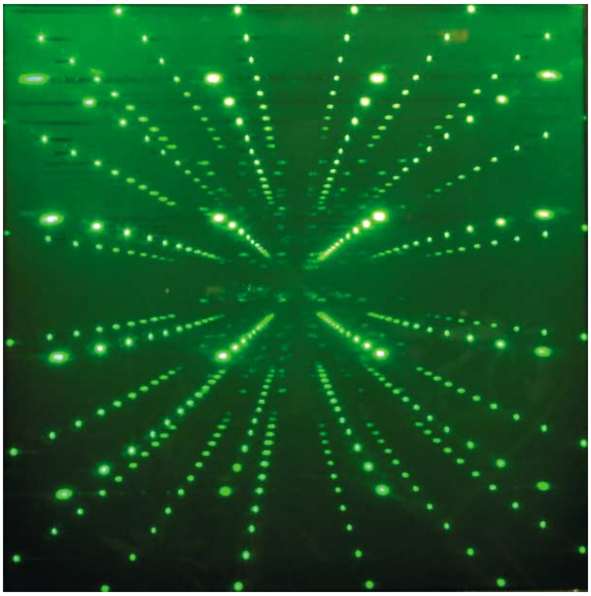

Figure 3 'Crystal Beginnings' Stephen Benton 1977, White light transmission (rainbow) hologram. Jonathan Ross Collection, London, UK. One of the best known works by the inventor of the rainbow hologram. Copyright © 2004 J. Ross. Reproduced with permission.

Although Benton's process increased the accessibility of holography, an earlier technique, invented by Yuri Denisyuk (Russia) in 1962, allowed holograms to be recorded in a single step and then displayed using white light. The political situation between the USSR and the rest of the world during the early 1960s meant that information and results from much of this early work was not openly accessible, and communication between Denisyuk and other key researchers in holography was limited. However, as soon as details of the Denisyuk technique became more widely available in the 1970s, artists began to embrace it as a practical alternative.

The work of Nicholas Phillips (UK), whose research improved chemical processing of display holograms during the late 1970s, helped artists and the rest of the holographic field to achieve much better results when making reflection holograms, a technique favored in Europe. He was also heavily involved with the rock band, 'The Who', and their Holoco company which mounted two very successful exhibitions of display holography at the Royal Academy of Art, London, at this time. Artists who might not have come across the holographic phenomenon became aware of its potential due to its presence in one of the UK's more traditional art venues. Other large exhibitions in the late 1970s and early 1980s in USA, Sweden, and Japan also highlighted the immense possibilities holography offered to a nonscientific community of gallery visitors.

As interest in all aspects of holography increased, dedicated spaces opened to sell, display, and archive the output from artists, scientists, and commercial producers. One of the most influential of these was the Museum of Holography, located in the SoHo art district of Lower Manhattan, New York (Figure 4). From 1976 through to 1992, it regularly mounted group and solo exhibitions by artists, giving many of

Figure 4 Dr Dennis Gabor, inventor of holography with, Posy Jackson, Director of the Museum of Holography, during his visit to the Museum in March, 1977. Copyright © 1977 Paul D. Barefoot. Reproduced with permission.

them their first major exposure to a museum-going public. There were displays relating to the scientific, commercial, and industrial usage of the medium, as well as exceptionally popular traveling exhibitions. ‘Through the Looking Glass’, which traveled worldwide, allowed early display holography to be experienced by thousands of people who would not normally have come into contact with the medium. In the Summer of 1981, over 300,000 visitors saw ‘Through the Looking Glass’ during the 11 weeks it was on view at the Israel Museum, Jerusalem.

After the Museum of Holography closed in 1992, its entire collection was purchased by the MIT Museum, Boston, USA, and now forms part of the largest museum collection of creative and scientific holography in the world. The New York museum was not an isolated venue. During the late 1970s and early 1980s, museums of holography opened in Chicago, Paris, and Pulheim (near Cologne). Galleries were also founded in many major cities to sell art and commercial holograms.

Stability Solutions

Although being able to display holograms using inexpensive sources of white light allowed the medium to become much more accepted as a display process, there has always been the cost and complexity of the optical laboratories needed to record the holograms in the first place. Few artists were able to equip their studios with ‘off the shelf’ optical components and precision-built isolation tables most often found in scientific and research labs. The solution was to replace the commercially built isolation table with much less expensive sand tables.

Sand box holography revolutionized the field for independent artists. By building a wood or brick box filled with sand, resting on car inner tubes, artists were able to hand-make their own isolation tables. Inexpensive optics, mounted onto plastic pipes were simply thrust into the sand, allowing complex optical recording setups to be constructed. Surprisingly, many key artworks have been produced using this technique, which opened the field to anyone with basic construction skills and a laser.

The laser has remained an expensive element in the holography studio but, most recently, the prospect of replacing this with laser diodes has offered a much more economic alternative. Not only can light from these diodes provide more compact optical setups to record holograms, they can also be used to display them. Significant progress has been made in both diode recording and display, which is opening up new opportunities for the visual artist. Almost all of the early practitioners in the field were attracted to holography because of the laser transmission holograms they saw. They have a depth and dimensional quality which is extremely different from the white light viewable holograms we tend to see today. Now, with the possibility of displaying holograms illuminated by inexpensive laser diodes, the laser transmission hologram is set, once again, to be a viable medium for artists.

Object Holography

Many artists were attracted to holography because of its unusual three-dimensional properties. However, they did not have the resources to make holograms of large objects or scenes, so scaled these down and constructed small sculptural settings for recording. One of the pioneering artists in this area was Dan Schweitzer (1946–2001). Originally an actor and familiar with theatrical staging, many of his holograms are populated by tiny figures placed in specially built ‘theatrical’ settings. Holography not only recorded these specially constructed miniature ‘rooms’, but the resulting images were realistic enough to suggest to the viewer that they were looking at a scaled-down life-size event. We do this regularly when looking at photographs and accept that they are a small-scale version of a larger scene. Traditional holography does not offer this ‘scaling down’ in the same way as photography. One celebrated example is Schweitzer’s ‘Thendara’, 1978, where a sculpted, miniature figure sits and looks towards a window frame. He also produced a companion piece in which the physical model, used to make the earlier hologram, is on display as a piece of sculpture. Here the miniature window is the frame for a hologram displaying a previously recorded model of a landscape. As Schweitzer’s technical skills

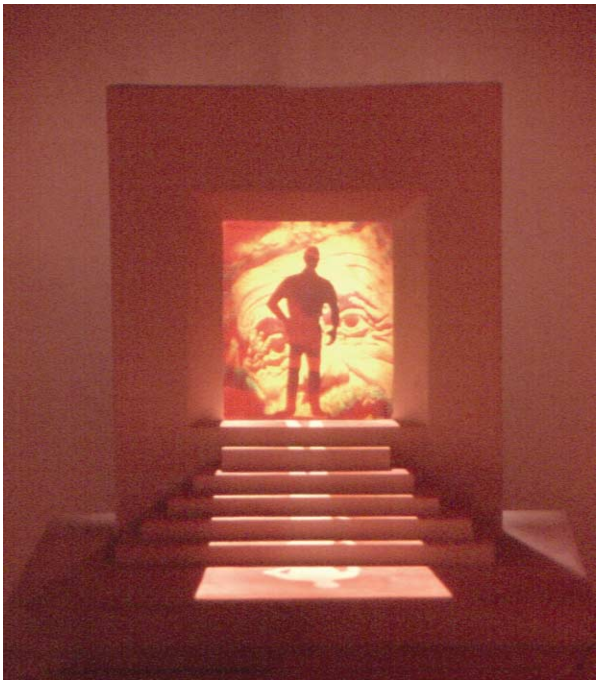

Figure 5 'Od Vent', Dan Schweitzer 1980, White light transmission hologram in sculptural mount. Copyright © 1980 ATP. Reproduced with permission.

developed, he began to produce complex multilayered works incorporating still-life models, hologram fragments, and optical animation to construct visual spaces which could not exist in the real world.

In his work 'Od Vent' (1980) (Figure 5), Schweitzer framed an exceptionally complex white light transmission hologram into a theatrical model of a doorway. The miniature stairs lead up to an opening, a portal into another world, in which the hologram is mounted. Illuminated from behind, the hologram displays multiple exposures, optical animation, and several superimpositions of objects as an observer moves from left to right in front of the piece. In this complex holographic world, the silhouette of the figure, standing in the door, casts a bright (negative) shadow down the miniature stairs. Not only are we able to look into an illusionary world through the holographic doorway, where at a certain point Einstein looks back at us, but the peculiarities of this holographic world extend into our own world as seen by the 'real' (negative) shadow cast down onto the stairs.

Another artist, who combined the use of multiple exposed objects displayed in a single space, is John Kaufman (USA). He established a reputation by recording objects (often rocks) collected from around his Californian holography studio. These found objects were captured in three dimensions and were then, through a multiple exposure and emulsion swelling technique, combined into a single still-life composition. He became expert at masking out areas of his objects so that they could appear in different, intense colors, and in many cases used the holographic process and multiple exposures to mix colors directly in the holographic space he was working with. In 'Stone Head' (1982) (Figure 6), for example, this relatively 'normal' stone is transformed by the use of intense color and the three-dimensional volume it occupies.

When considering how the hologram should be framed and displayed, several artists have utilized the mount in which the hologram sits as an integral part of a much more complex work. Frithioff Johansen (Denmark) recorded a hologram of smoke, freezing it in time and space using a pulsed laser (see below). This was then converted into a white light transmission hologram and mounted in a frame of granite. 'Hardware-Software 1' 1987–1988, allows the gallery visitor to not only view the massiveness of the granite structure, but also the delicate sculptural solidity of smoke 'frozen' in space.

Rebecca Deem (USA) recorded what appears to be an extremely simple hologram of a luminous tube, mounted onto a swivel at the end of an elegant metal support. In 'W/Hole Message' viewers can look into the small holographic plate and see the image of the tube extending away from them into the volume behind the plate. By flipping the hologram, on its swivel, to turn it over, the image in the hologram is also flipped and the tube appears in the viewer's space, between them and the plate. This three-dimensional pseudoscopic image, one of the attractive and unusual aspects of holographic display, is inside out, back to front, and upside down. The gallery visitor really does have a chance to view the whole message.

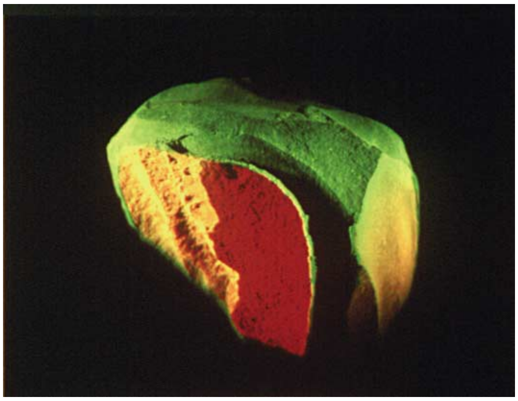

Figure 6 'Stone Head', John Kaufman 1982, White light reflection hologram. Copyright © 1982 John Kaufman. Reproduced with permission.

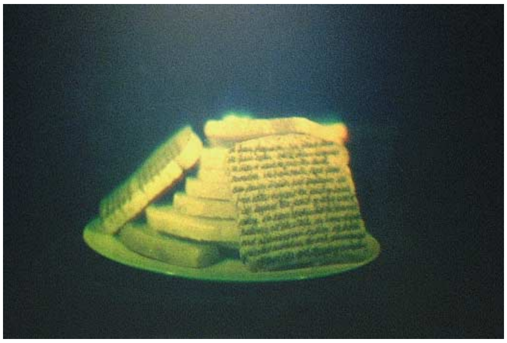

Figure 7 'Breakfast', Pearl John 2002. Laser etched toast displayed in a white light reflection hologram. Copyright © 2002 Pearl John. Reproduced with permission.

Text has become an important element in holography, as it is often incorporated in, onto, or around holographic displays. Eduardo Kac (Brazil) has produced several holograms, which display poems, words, and letters floating within the holographic space he has created.

Pearl John (UK) most recently produced a series of still-life holograms using a pulsed laser but with the added subtlety of writing onto the surface of her objects. In 'Breakfast' 2002 (Figure 7), for example, the slices of toast, making up the composition, have been laser etched with text which appears 'burnt' into the surface of the bread.

Pulsed Lasers

Continuous wave lasers placed limits on the holographic recording process and dictated that no movement in the recorded object could take place. We have seen that this 'problem' was exploited by artists at an early stage, but there was also a desire to capture living objects, in the form of the human figure. Pulsed laser holography was the solution. Because this type of laser generates a bright flash of coherent laser light, the intensity and duration of the pulse is enough to 'freeze' any physical movement during the recording process, allowing the recording of people, animals, or anything which moves and could be brought into an holography studio. Again, the cost and accessibility of this specialist equipment was prohibitive to artists, but they found ways of persuading scientists with pulsed lasers to help them. Early adopters of this technique were artists such as Ana Maria Nicholson (USA), Melissa Crenshaw and Sydney Dinsmore (Canada), Margaret Benyon (UK), Anait (USA), and Alexander (USA/UK). Recordings could now be made of live figurative compositions and individual portraits. Many of the artists, not content with producing the holographic version of a traditional photographic portrait, began to manipulate these recordings, distorting, multiple exposing and inverting these unusual three-dimensional images of people.

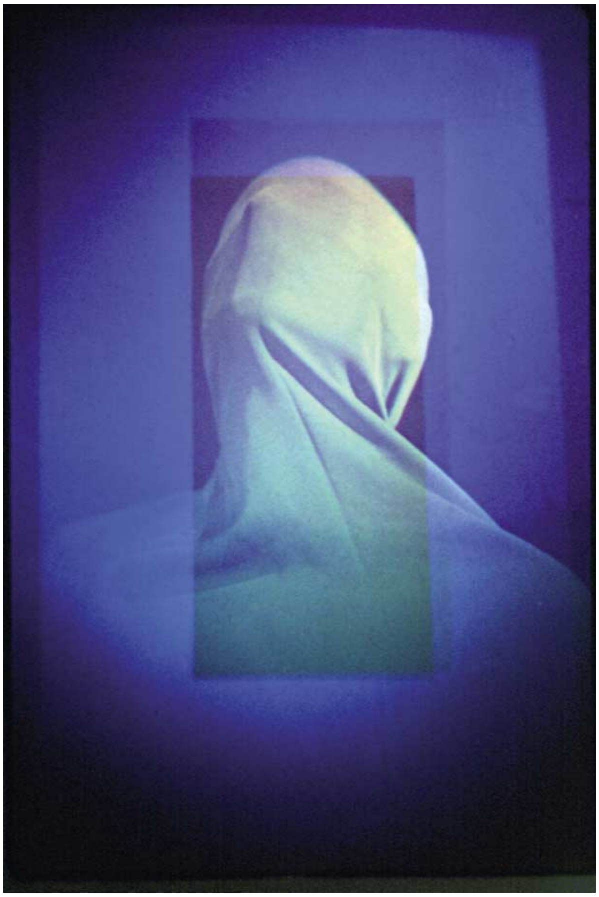

Figure 8 'Awakening', Ana Maria Nicholson, 1993. White light (pulsed) reflection hologram. Copyright © 1993 Ana Maria Nicholson. Reproduced with permission.

Holographic portraits do have an unusual quality, which painters or photographers have found difficult to deal with but art holographers have embraced. Some, like Ana Maria Nicholson (USA), have combined subtle processing techniques, to introduce color 'layers' into the portraits, combined with nonstandard views or compositions (see 'Awakening' 1993 (Figure 8)). Other artists, who emerged in the 1980s, produced complex, highly staged compositions, such as those from Patrick Boyd (UK), or they offered us a different view of the 'self portrait' like those by Martin Richardson (UK). Others, such as Shu-Min Lin (Taiwan), used the combination of multiple views of human figures recorded onto the same holographic plate. By walking past these displays, an animated effect could be experienced.

So much interest was generated in pulsed holography that several studios emerged for use by artists, including independent facilities and those within educational establishments. No longer was this the domain of the well funded scientific research facility. Several of these 'studios' still exist and are offering artists access to equipment through specialized artist-in-residence programs.

Abstraction

Although holography is an exceptional process for recording every detail of an object and the volume it occupies, many artists have wanted to work with the abstract and kinetic possibilities the process offers. Light can be recorded and displayed in three dimensions, it can be animated, inverted, compressed, sculpted, and projected. Douglas Tyler (USA), in his In the 'Dream Point Series' 1983 (Figure 9), recorded small laser transmission holograms but displayed them using normal white light, producing a smearing effect of the three-dimensional object. He then mounted these onto transparent surfaces (Plexiglas) and drew on the surface of the plastic to incorporate etched lines and shapes. Rudie Berkhout (Netherlands/USA) began his holographic exploration with stunning abstract landscapes made of luminous, multicolor, animated objects which took advantage of his mastery of the white light transmission technique. Most recently he has been recording reflection holograms which display undulating colored 'waves' of light as an observer moves past the hologram. These have been included in specially produced paintings for large-scale architectural commissions.

Sam Moree (USA) combines multiple images, exposures, diffraction gratings, and smaller holograms into a single work. The finished holograms display a complex series of image fragments and colors within the depth of the holographic space he has constructed. Ruben Nuñez (Venezuela), originally approached holography from a background in glass production and used the ability of the holographic process to 'capture' light reflecting and passing through his glass objects. The resulting white light transmission holograms display an abstract 'photonic' display of form and color. Fred Unterseher (USA), uses holography to record optical distortions and diffraction gratings which would display abstract patterns of light. Many of these early works were mandala images which displayed an intense kinetic effect as an observer walked past the surface of the holographic plate.

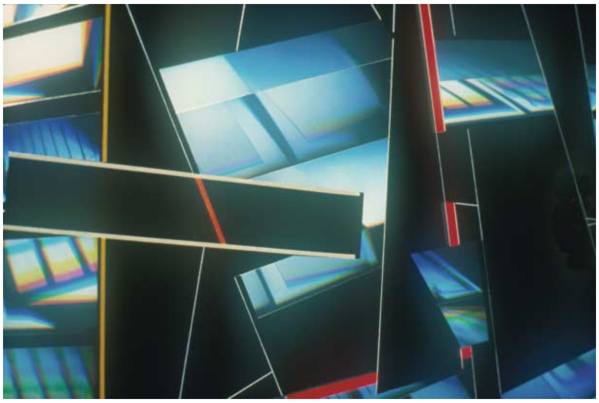

Figure 9 'Dream Passages': 32″ × 138″ (3 panels 32″ × 46″). Douglas E. Tyler 1983, Laser transmission holograms, tape and Plexiglas, illuminated with white light. Detail of center panel. Copyright © 1983 D.E. Tyler. Reproduced with permission.

Pseudo Holograms

In 1973, Lloyd Cross invented the multiplex (integral) hologram technique which used sequential two-dimensional images recorded on movie film combined with white light (rainbow) hologram techniques. The result was a holographic stereogram, which gave viewers the illusion of seeing animated images in three dimensions. The final holographic sheet was displayed either curved or as a complete cylinder, which was often motorized so that viewers could stand in front of the unit and watch the animation take place as the hologram rotated. Such a display does give the impression that an observer can walk all the way round the image, enhancing its perceived three-dimensional effect.

As the original images are recorded photographically, there are several advantages – anything which can be recorded on inexpensive black and white movie film can be converted into a holographic display. Short animated sequences can be presented while accepted film techniques such as zoom, pan, tilt, and transitions can be incorporated into the finished sequence of images. Although very popular for advertizing and promotional campaigns during the 1980s, particularly in the USA, several artists used this technique. They could make the movie film sequences themselves and then send this film to a commercial producer for a hologram to be made for them. This offered an interesting balance in the creative process. The artist could concentrate on the filmic image, using facilities and photographic techniques they were familiar with and not have to learn the holographic process or build a holographic laboratory themselves. Artists such as Anait (USA) produced several experiments in the area, one of which showed her inside the holographic display,

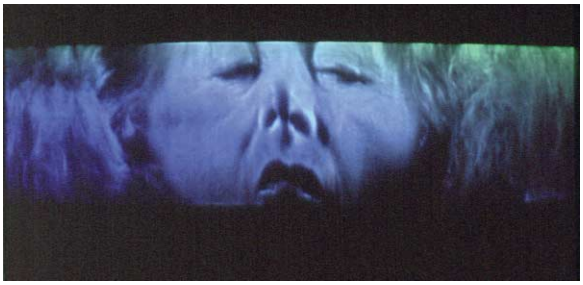

Figure 10 'Woman' Harriet Casdin-Silver 1979. Multiplex hologram (white light transmission). Copyright © 1980 ATP. Reproduced with permission.

writing, with spray cream, seemingly onto the holographic surface. Dieter Jung (Germany), used this technique to display a specially written poem which appeared to spiral around inside of the display.

Harriet Casdin-Silver used the technique to produce an unusual self-portrait. 'Woman', 1979 (Figure 10), is a multiplex hologram of the artist's head. Black and white film images were recorded of Casdin-Silver's head, through 360 degrees, and then each frame of film was converted into a thin white light transmission hologram. The artist chose not to display the resulting image in a normal cylindrical display, but to 'flatten' it out so that the front, sides, and back of her head could be seen, distorted, in a single head-on view.

As video and digital imaging processes have developed over the years they too have been combined with this pseudo holographic technique, opening up a wide avenue of possibilities. Video footage, computer generated modeling, or any process which can produce sequential 2-D images, can now be converted into an animated 3-D hologram. As digital video cameras and computer 3-D modeling software drop in price, and increase in sophistication, a whole new area of display has become available to the digital and video artist.

Multi Media

As in other areas of the visual arts, more than one medium or process are often combined. Doris Vila (USA) has produced complex environments combining large-scale holographic images which, rather than being illuminated with a single source of light, are reconstructed with multiple spotlights or via video projectors. The lighting is activated by sensors in the exhibition space, which are computer controlled and made to trigger certain sequences dependent on the input they receive from the movement of visitors.

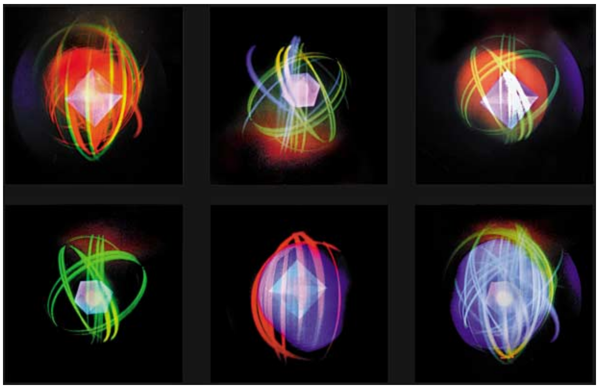

Figure 11 'Neuro Hologram', Ikuo Nakamura, 1993. White light transmission hologram with multiple, computer controlled, lighting. Copyright © 1993 Ikuo Nakamura. Reproduced with permission.

Traditional illumination of a hologram involves a single point source of light shining onto, or through, the holographic plate. This will reconstruct a single image. If multiple lights are shone onto the hologram, multiple reconstructions of the single image will result. By using many independently controlled lights on one hologram, Vila has been able to create animation and kinetic effects as the resulting multiple images overlap. Using light from a video projector also allows the artist to combine a reconstructed holographic image with two-dimensional animated footage from the video projector.

Other artists to combine holograms with structures, interactive lighting and environments, sound, and installational displays include: Georges Dyens (Canada), Marie-Christiane Mathieu (Canada), Matthew Schreiber (USA), Philippe Boissonnet (Canada) Ana MacArthur (USA) and Shunsuke Mitamura (Japan).

Ikuo Nakamura (Japan/USA) has used multiple lighting of a single computer generated hologram to produce animated effects controlled by the viewer. In his 'Neuro Hologram' 1993 (Figure 11), brain waves from the viewer, wearing a special set of sensors on their head, are measured and converted into control signals for a number of lights behind the hologram. By relaxing, or becoming agitated, the viewer can change their brain waves and actively control the appearance of the hologram, causing multiple views of the holographic image to fade in and out, giving an impression of abstract animation.

Architectural Scale

One of the great limitations of holography has been the size of the holographic plates on which the light sensitive emulsion is coated. These tend to be in the region of 20 × 23 cm, 30 × 40 cm, or meter-square formats. Generally, the larger the finished hologram, the larger the laboratory needed to produce it. This impacts a great deal on artists who do not have the facilities, or finances, to produce ‘large’ holograms. A solution has been to combine smaller holographic plates together, as a mosaic, to give the size and scale needed. In recent years, several artists have completed public commissions for large-scale interior and exterior holographic installations.

Setsuko Ishii, (Japan) has produced several installations in public spaces where individual holographic plates are mounted together to form a larger composite structure. Many of these pieces use holograms recorded on dichromate gelatin plates which produce an exceptionally bright image, ideal for public spaces, where visibility, from multiple angles, is important.

Michael Bleyenberg (Germany) produced an entire exterior holographic wall (13 × 5 meters) during 2002 (Figure 12), for the extension of a research building in Bonn. The installed graphic wall, ‘EyeFire’ (AugenFeuer), is made up of 26 smaller panels, each constructed from tiny holographic ‘pixels’ (holographic optical elements, HOE) to give an intense color effect. Working in collaboration with the Institute for Light and Building Technology, Cologne, Bleyenberg was able to design his graphic panels on a computer screen and have this information translated into the thousands of tiny holographic pixels which, when seen together, make up the complex and high intensity colour design.

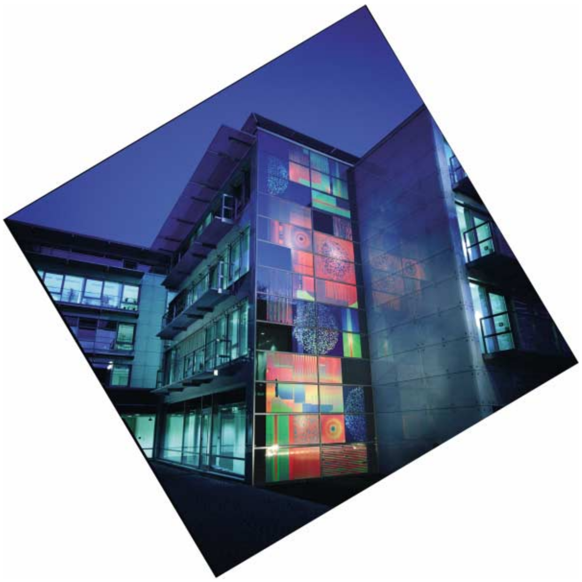

Figure 12 ‘EyeFire/AugenFeuer’, Michael Bleyenberg 2000. White light transmission holograms mounted onto building exterior. Copyright © 2000 M. Bleyenberg. Reproduced with permission.

Dieter Jung (Germany) has produced commissioned art works for large interior architectural spaces. In ‘Perpetuum Mobile’ 2001, located at the European Patent Office, The Hague, holograms are placed on the floor, which can be walked over, and an extensive holographic mobile structure hangs overhead. The light, and color, from the abstract holograms in the mobile sections, can be reflected in the abstract holographic floor and combine to produce an interior lightscape impossible to produce with more traditional media.

Holograms can be mass produced using a mechanical embossing process which transfers the optical contents of the original master hologram onto plastic sheeting for duplication in large runs. One novel use of this material has been the interior of the Canadian pavilion at Expo ‘92 (Figure 13). Melissa Crenshaw (Canada/USA) used huge quantities of embossed holographic material to cover the 3000 square foot interior walls of the specially built pavilion. Lighting was installed to illuminate this optical diffracting wall and the entire structure was covered in glass, over which water flowed. The resulting interior appeared to be a wall of liquid light which changed color and hue depending on where it was viewed from.

Sally Weber (USA) used holographic embossed material for several outdoor commissions. Structures were clad in embossed holographic sheets and made water resistant. Sunlight illuminated the installations to produce surfaces of constantly changing color. In direct (bright) sunlight, intense rainbow colors would be diffracted by the holographic surfaces. On an overcast day, pastel, more muted,

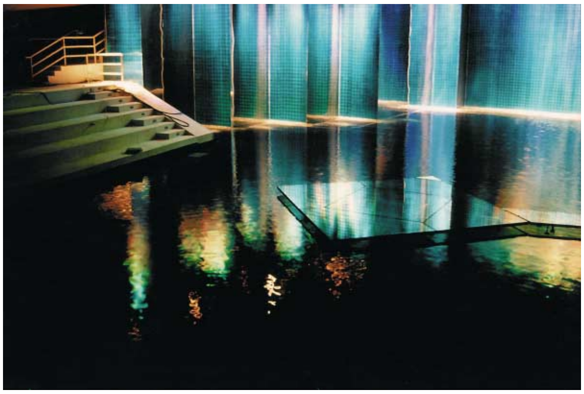

Figure 13 Interior of the Canadian Pavilion Expo ‘92, Spain. Holographic wall water feature by Melissa Crenshaw. White light transmission (embossed) hologram foil mounted into pavilion walls. Copyright © 1992 M. Crenshaw. Reproduced with permission.

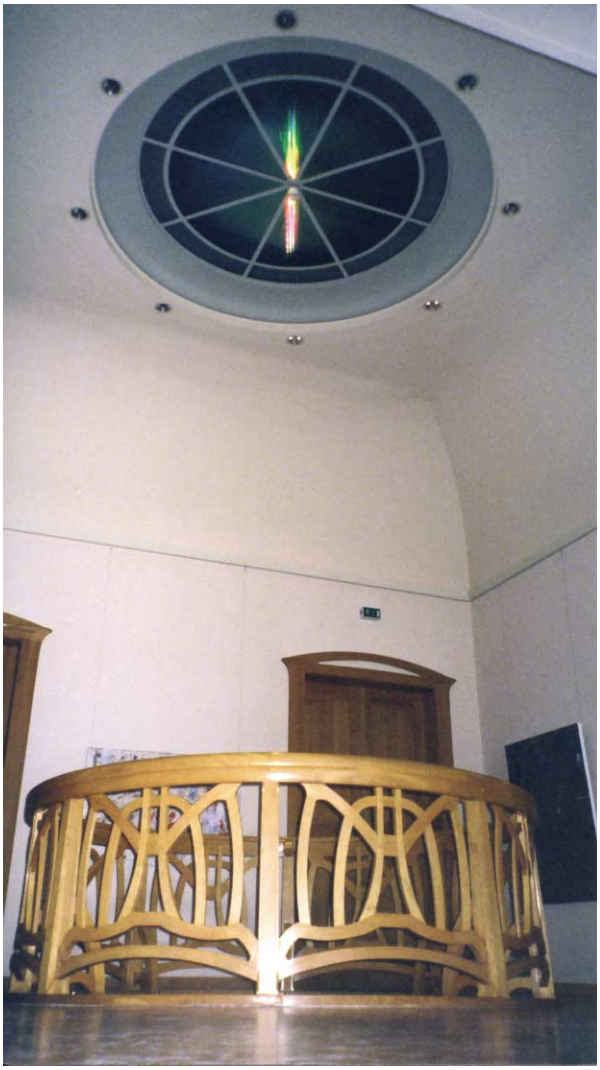

Figure 14 Signature of Source, Sally Weber 1997. Site specific window installation at the Karl Ernst Osthaus-Museum, Hagen, Germany. Copyright © 1997 Sally Weber. Reproduced with permission.

colors would appear. Although she is often known for these large-scale outdoor embossed works, one of her permanent commissions is housed in the Karl Ernst Osthaus-Museum, Hagen, Germany, where specially made holographic diffraction gratings are incorporated into an existing building. 'Signature of Source' 1997, (Figure 14), is a permanent, site-specific installation producing a line of light which enters the museum space from the skylight it is built into. This abstract line of pure color becomes a sculptural element within the space of the museum and is generated by the action of light from outside the building being manipulated and 'formed' as it passes through the holographic window.

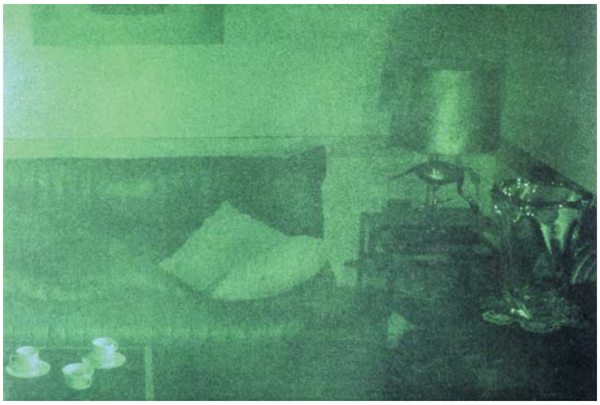

Figure 15 'There's no Place Like Home', Paula Dawson 1979–1980. Laser transmission hologram of an entire room. Copyright © 1980 Paula Dawson. Reproduced with permission.

One of the most celebrated and very early attempts to incorporate large-scale holograms into art works was the installation 'There's No Place Like Home' by Paula Dawson (Australia) in 1979–1980 (Figure 15). A laser transmission hologram of an entire room was recorded. The room was specially constructed, life size, and recorded onto a single holographic plate. The resulting hologram was then built into the window of a small 'house' constructed in the gallery. Visitors were able to approach the house, look into the window and 'see' the interior with its furniture, fixtures, and fittings. They would then enter the house and walk past the space where the room appeared to exist. When viewed from this vantage point there was nothing there except an empty space. Artists have often questioned and researched the concepts of 'inside and outside', voyeurism, reality, normality, and 'existence', but few have had the tenacity to produce an installation with such visual and conceptual impact.

Twenty-First Century Art

Art holography has developed and matured over the past 50 years to become a viable and significant element in the contemporary visual arts. Critics are still asking 'is it art?' in the same way they used to worry about whether photography was art. While they consider this problem, practitioners in the medium are continuing to find ways to push past its technical and visual boundaries (Figure 16). They continue to collaborate with the scientific and research world and use holography as a medium to comment on life and society in the early twenty-first century.

> ...I hope, that one day I will be able to give you, artists, something really powerful; panoramic holograms,

MANIFESTO

.A code.
.Spatial.
.Didactic.
.Photonic.
.Futuristic.
.A discipline.
.Illumination.
.A photograph.
.Science↔Art.
.A precision tool.
.Illusion concrete.
.Pure chiaroscuro.
.An automatic Eden.
.An indivisible whole.
.Painting↔Sculpture.
.3D on 2D intimating nD.
.Material and immaterial.
.Writing the whole picture.
.The presence of absence.
.Making the invisible visible.
.An artificial 3D imaging system.
.Re-evaluating our visual traditions.
.Part of an innate human grammar.
.A bridge into the future with a past.
.The smile without the Cheshire Cat.
.Neither geometric nor organic in form.
.The reconstruction of light wave-fronts.
.Shattering the illusionist paradigm in art.
.A perceptual screen from the 'real world'.
.A potential mass-communications medium.
.A visible medium and a subvisible principle.
.A model for the complexities of brain function.
.Not art, but a condition in which art can happen.
.Mirrored in the life process, our own physiology.
.A mirror to our evolutionary progress from objects to systems.
.Realizing space time/time reversed/back as front/inside as out.
.A model for non-fragmentary thinking, language and behaviour.
.A diminutive shadow-theatre in which metaphors are strongly lit.
.A possible visual model for quantum theory and relativity theory.
.The most complete form of reproductive imaging that exists to date.
.A means of making statements about the world, or presenting pieces of it.
.The fulfillment of a longstanding cultural vision of what imaging ought to be like.
.Joined with high technology to raise the ceiling on humanity, or record Doomsday.
.The automatic execution of the ultimate cubist notion of a simultaneous view at every given point.
.An expression for mystical aphorisms such as 'the universe in a grain of sand' and 'we are all one'.

HOLOGRAPHY AS ART

Figure 16 'Holography as Art', Margaret Benyon 1980. Manifesto. Copyright © 1980 M. Benyon. Reproduced with permission.

extending to infinity, in natural colors. Dennis Gabor, Inventor of holography in a letter to artist Margaret Benyon 28th February, 1973.

See also

Coherence: Coherence and Imaging. **Diffraction:** Diffraction Gratings. **Holography, Techniques:** Overview. **Imaging:** Multiplex Imaging. **Interferometry:** White Light Interferometry. **Introductory Article:** Early Development of the He–Ne Laser (Title TBC).

Further Reading

Benyon M (1994) *How is Holography Art?* PhD thesis, Royal College of Art, London.

Brill LM (ed.) (1989) Holography as an art medium, Leonardo. *Journal of the International Society of Arts, Sciences and Technology*, 22 (3 and 4).

Coyle R and Hayward P (1995) *Apparition, Holographic Art in Australia*. University of Sydney, NSW, Australia: Power Publications.

Casdin-Silver H (1998) *The Art of Holography Exhibition Catalogue*. Lincoln, MA, USA: DeCordova Museum and Sculpture Park.

Holographische Visionen, Bilder durch Licht zum Leben erweckt (1991) Exhibition Catalogue, Museum für Holographie und neue Visuelle Medien, Pulheim, Germany: Edition Braus.

Holosphere, Formerly published by the Museum of Holography. New York: Boston, USA: Archieved at the MIT Museum.

Jackson R and Burns J (1976) *Holografi: Det 3-Dimensionella Mediet, Exhibition Catalogue*. Museum of Holography, New York: Stockholm Cultural Center.

Jung D (ed.) (2003) *Holographic Network*. Germany: Rasch Verlag, Bramsach.

Orazem V and Schepers J (1992) *Olografia, Holografie, Holography, Exhibition Catalogue*. Rocca, Paolina, Italy: Centro Espositivo.

Lipp A and Zec P (1985) *Mehr Licht/More Light, Exhibition Catalogue*. Kabel-Verlag: Hamburger Kunsthalle.

Pepper AT (ed.) (1992) *The Creative Holography Index: The International Catalogue for Holography*. Germany: Monand Press.

Pepper AT (ed.) (1996) *Art in Holography 2*. International Symposium on Art in Holography. Nottingham, UK: University of Nottingham.

Margaret Benyon M (1989) *Phases: Exhibition Catalogue*. New York, USA: Museum of Holography.

Tyler DE (ed.) (1990) *Report, International Congress on Art in Holography*. Notre Dame, Indiana, USA: Saint Mary's College.

Tyler DE (ed.) (2003) *Leading Lights, Women in Holography, Exhibition Catalogue*. Notre Dame, Indiana, USA: Saint Mary's College.

High-Resolution Holographic Imaging and Subsea Holography

J Watson, University of Aberdeen, Aberdeen, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Holography is well-known to most of us: there can be few who are not familiar with the white-light display hologram seen in exhibitions and shops and on credit cards. The impact of holography, though, spreads far and wide into areas as diverse as vibration analysis, image processing, holographic-optical elements, and optical computers. One aspect often overlooked is its suitability for high-precision imaging and accurate dimensional measurement. It is the unique ability of holography to recreate, from a flat photographic film, a three-dimensional image that is optically indistinguishable from the original which sets it apart from other imaging techniques. In particular, holography comes into its own when recording objects in a hazardous environment or where access is difficult, such as underwater or in nuclear fuel element inspection.

We describe here the use of holography for imaging, identification, and measurement of aquatic organisms. A detailed knowledge of the distribution and dynamics of marine and freshwater organisms and particles, such as plankton, is crucial to our understanding of such systems and how they affect our environment. Traditional methods of gathering *in situ* data on aquatic particles, such as net-collection and electronic counting or photography, are not usually suited to observing precise spatial relationships and can destroy the more delicate organisms. The overriding benefit of holography is that it permits nonintrusive and nondestructive observation and analysis of organisms in their natural environment, while preserving their relative spatial distribution. Holography also offers a permanent archive. The use of a short-duration pulsed laser is essential, in order to reduce the effects of vibration or movement in any of the optical components or objects being recorded. Consequently, the object scene is effectively 'frozen' at the recording instant, thereby allowing fast moving particles to be imaged. Compared to a photograph, a hologram captures vastly more data. Sequential holograms can record changes within the volume over a defined period of time and the wide recording dynamic range afforded by a hologram allows images to be captured that would otherwise be lost in noise.

Methodology of Holographic Recording and Replay

Of the many possible recording methods of holography two in particular, 'in-line' and 'off-axis', find use for high-resolution *in situ* imaging underwater. The concepts and fundamental principles outlined here are generally applicable wherever high-precision measurement is needed from a hologram.

In an in-line reference beam hologram (ILH), a single laser beam (Figure 1) is directed through the

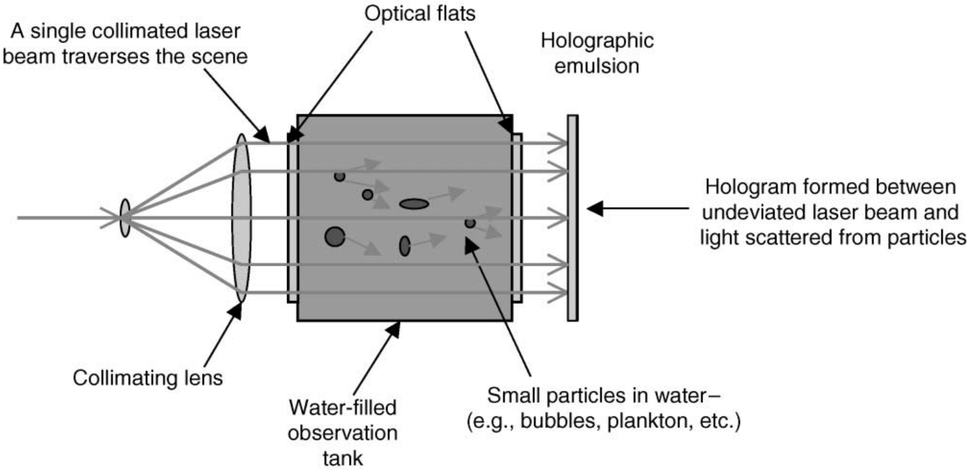
A single collimated laser beam traverses the scene
Optical flats
Holographic emulsion
Hologram formed between undeviated laser beam and light scattered from particles
Collimating lens
Water-filled observation tank
Small particles in water– (e.g., bubbles, plankton, etc.)

Figure 1 Recording an in-line hologram.

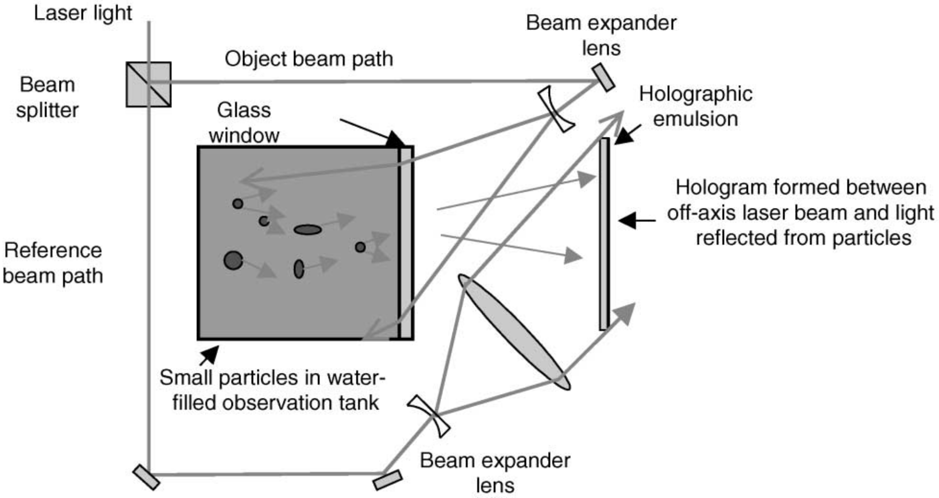
Laser light
Beam splitter
Object beam path
Beam expander lens
Glass window
Holographic emulsion
Hologram formed between off-axis laser beam and light reflected from particles
Reference beam path
Small particles in water-filled observation tank
Beam expander lens

Figure 2 Recording an off-axis hologram.

sample volume towards the holographic plate and records the optical interference between light diffracted by the object and the undeviated portion of the illuminating beam. The subject volume is illuminated in transmission and the scene needs an overall transparency of about 80% so that speckle noise does not seriously degrade image quality. Particles down to around 5 μm dimension can be readily recorded and identified. The particular benefits of ILH are its geometric simplicity, minimization of laser coherence and energy requirements, and high resolution over a large depth of field.

For recording an off-axis reference beam hologram (OAH), a two beam geometry is utilized: one beam illuminates the scene and the other directly illuminates the holographic film at an oblique incidence angle (Figure 2). Interference occurs between diffuse light reflected from the scene and the angularly separated reference beam. On replay, the real and virtual images are similarly angularly separated which makes their interrogation easier. OAH is primarily applied to opaque subjects of large volume. Although there is no real upper limit to the size of particles that can be recorded (this is determined by available energy and coherence of the laser) the practical lower limit to the resolution that can be achieved is around 50 to 100 μm. The scene can be front, side or back illuminated (or all three combined) allowing high concentrations of large particles to be recorded, provided that they reflect sufficient light to the film, and allowing subject visibility to be optimized. OAH provides a more complete record than ILH, in the sense that it gives information on the surface structure of the object, while ILH records only a silhouette.

Holograms are replayed or reconstructed in a variety of ways, depending on the recording parameters and subsequent processing of the recording medium. In scientific and engineering applications the requirements of high resolution and image fidelity demand that they are replayed with a laser, and viewed with some auxiliary optical system and (usually) a video camera. Depending on the orientation of the hologram, either the virtual image, which is apparently located behind the hologram and seen through it, or the real image, which is formed in space in front of the hologram, will be visible.

Considering first the off-axis hologram in the virtual image mode, the processed hologram is replaced in the position in which it was recorded and illuminated with an exact duplicate of the original reference wave, in terms of its wavelength, curvature, and beam angle. In this mode, which might be termed 'normal' viewing of an off-axis hologram, a virtual image (Figure 3) is observed as if viewing the scene through a window. Other than the colour, this image is almost optically indistinguishable from the original scene; it retains features such as parallax and three-dimensionality and does not suffer from perspective distortion or defocusing of the image.

For data extraction and measurement from a hologram, creation of the virtual image is not the best method of observation. Here the projected real image (Figure 4) is most useful. If we illuminate an off-axis hologram from behind, with a wave which is the exact conjugate of the original (i.e., one which possesses the same wavelength as the original but with the opposite direction and curvature), we generate an image that appears to float in the real space in front of the observer. Often a collimated reference beam is used in recording, since reconstruction is then a simple case of turning the hologram around. No lens is needed to form this image, and it is optically identical to the original wave that it is reversed left-to-right and back-to-front (pseudoscopic). Planar sections of the real image can be optically interrogated by directly projecting onto a video camera (often with the lens removed) which can be moved throughout the 3D volume. It is this 'optical sectioning' feature of a holographic recording that is so useful when nondestructive (noninvasive) evaluation must be made, at high resolution, of collections of objects which extend over a considerable depth.

A series of images taken from the real image of one hologram and captured on a video camera (Figure 5) illustrates how the viewing plane can be moved sequentially through the reconstruction to show firstly one marine organism and then another. The axial separation of the organisms is about 35 mm.

To replay an in-line hologram, it is also placed in a conjugate of the recording beam (Figure 6). The replayed hologram simultaneously forms two images, one virtual, the other real, which are located on the optic axis on opposite sides of the holographic plate. These images are co-axial, and located (for the typical case of collimated recording and replay beams) at equal distances in front of and behind the hologram plane. Although a true 3D real image is produced it has almost no parallax, and cannot easily be seen by the unaided eye. This image has inverted depth perspective and is seen in darkfield against the out-of-focus virtual image.

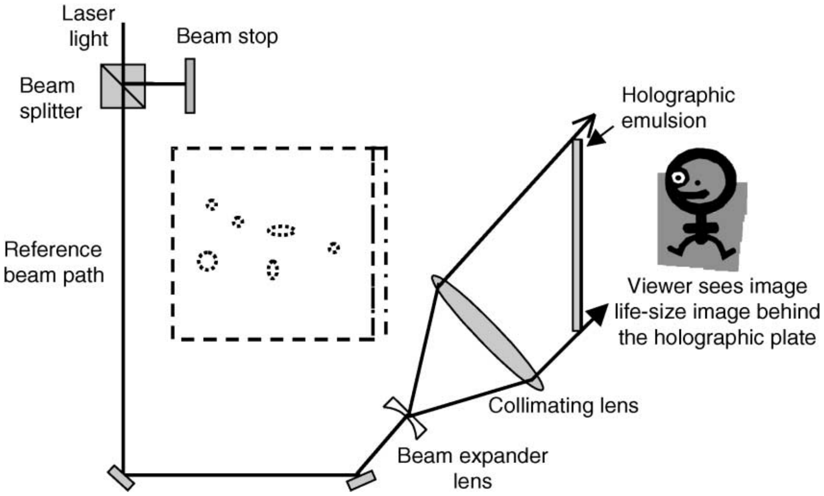

Figure 3 Replay of an off-axis hologram in virtual image mode.

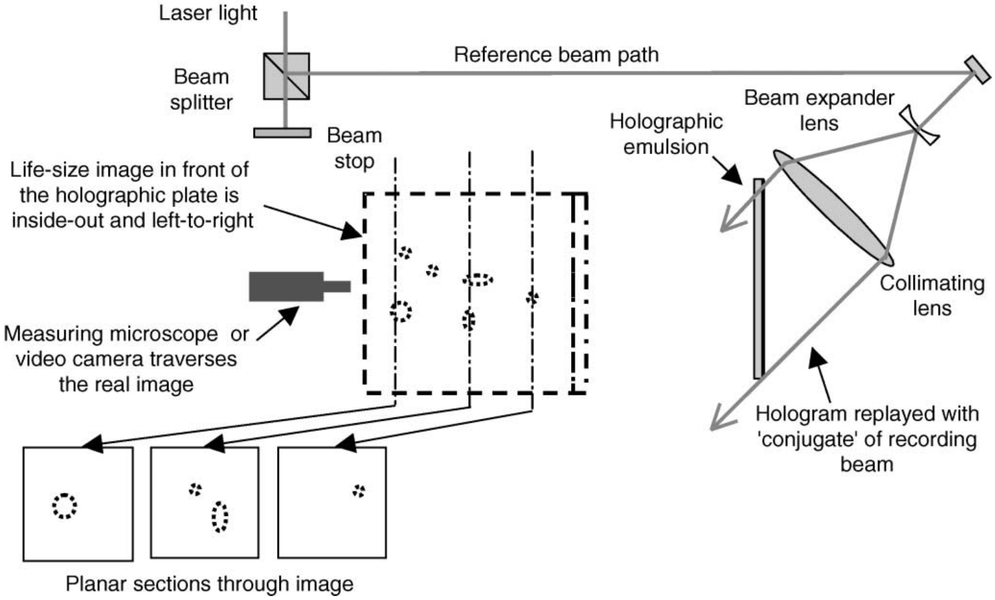
Laser light
Beam splitter
Reference beam path
Beam stop
Beam expander lens
Holographic emulsion
Life-size image in front of the holographic plate is inside-out and left-to-right
Measuring microscope or video camera traverses the real image
Collimating lens
Hologram replayed with 'conjugate' of recording beam
Planar sections through image

Figure 4 Replay of an off-axis hologram in real image mode.

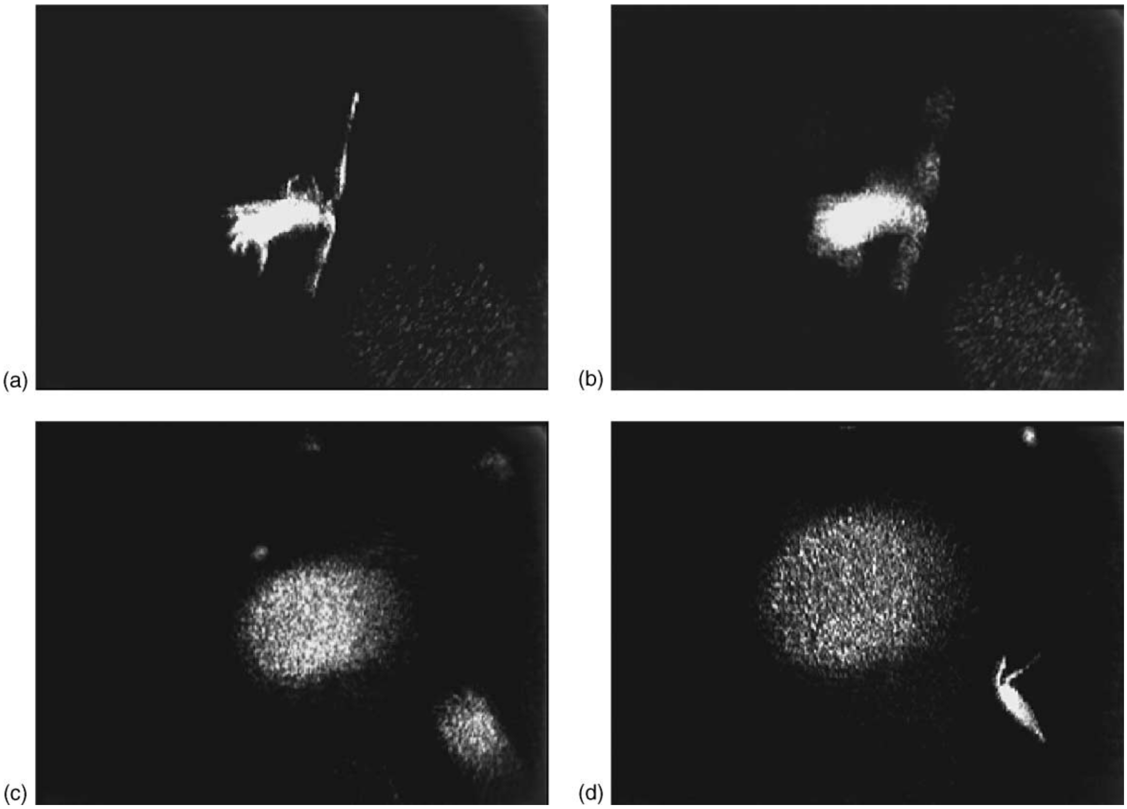

Figure 5 A series of images taken from a single hologram showing the concept of moving the viewing plane through the real image.

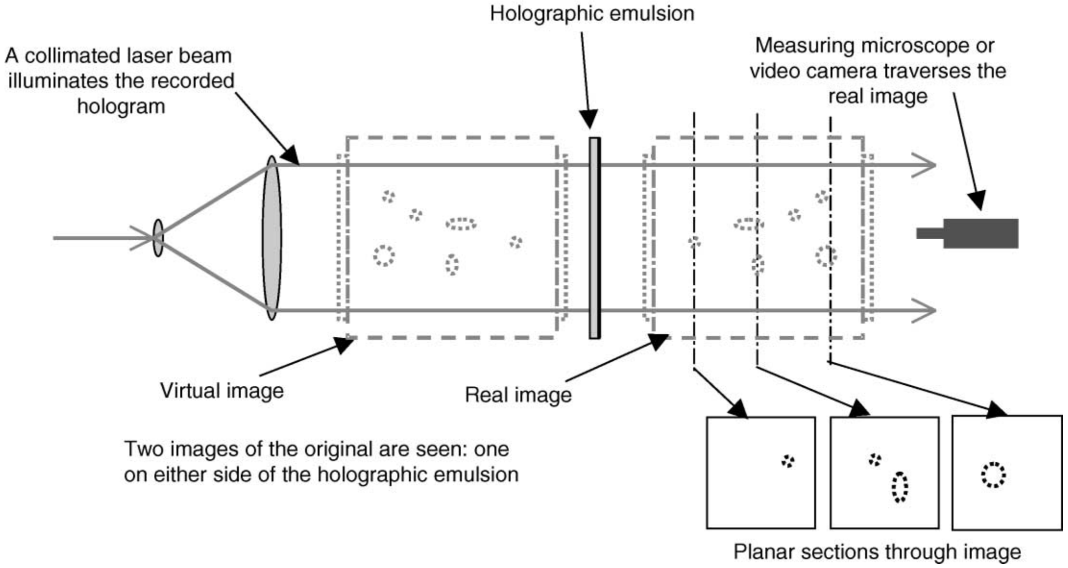
Holographic emulsion
A collimated laser beam illuminates the recorded hologram
Measuring microscope or video camera traverses the real image
Virtual image
Real image
Two images of the original are seen: one on either side of the holographic emulsion
Planar sections through image

Figure 6 Replay of an in-line hologram.

Holographic Image Quality

In general, the usefulness of holography for accurate inspection and measurement is dependent on its ability to reproduce an image of the object which is low in optical aberrations and high in image resolving power. Regardless of whether or not the hologram is recorded underwater, high resolution, high contrast, and low noise are the dominant requirements, rather than image brightness. In practice, loss of resolution and image degradation can occur at any stage in the recording and replay of a hologram. All the primary monochromatic aberrations to be found in any optical system (spherical aberration, astigmatism, coma, distortion, and field curvature), may be present in the holographic image. The most significant factors affecting this are departure from conjugacy of the reconstruction beam; positional mismatch and misalignment between reference beam and film; degradation of the reconstruction beam due to poor optical quality of the film substrate or optical components; and, distortion of the fringe pattern recorded in the holographic film as a result of chemical processing. For precise and accurate reconstruction of the real image, the reconstruction reference beam should be the exact phase conjugate of that used in recording. Under these conditions, the lateral, longitudinal, and angular magnifications of the real image with all equal unity and aberrations will be reduced to a minimum.

Diffraction-limited resolution of a holographically produced real image, in the absence of all aberrations, is usually defined from the ability to distinguish two points in the image separated by a distance r. The resolution as defined in this way is dependent upon reconstruction wavelength, λ_c, the separation between hologram and reconstructed image, s, and the effective aperture, D, of the hologram:

$$r = 1.22 s \lambda_c / D \qquad [1]$$

This relationship is well-known in imaging optics: it defines the ideal resolving power of a lens and is equally applicable to holography. In holography, though, the presence of speckle effects introduced by the coherence of the light and the finite aperture of the viewing system influence the resolving power. In practice, the minimum resolution is increased by a factor of two to three to take account of speckle. For off-axis holograms the choice of film size (effective aperture) has a considerable bearing on image resolution. Typically, for 100 mm diameter film the theoretical image resolution (neglecting speckle effects) is around 6 μm (150 lp/mm) with a 532 nm laser at a target distance of 1 m. For in-line holograms (with collimated illumination) the film size also defines the transverse dimensions of the volume recorded.

Wavelength matching between recording and reconstruction beams require that, if possible, the same type of laser be used in both replay and recording. Of course, this is not possible in field situations when the use of a pulsed laser is essential to ensure that the fringe pattern is not blurred by any movement in the system. In such applications, laboratory replay of the hologram may require the use of a tuneable laser to match wavelengths precisely, although in practice an argon laser provides a useful selection of wavelengths. A mismatch between record and replay wavelengths will lead to distortion in the image, because longitudinal and lateral magnifications will no longer have the same value.

In contrast to off-axis holograms, the interference patterns on an in-line hologram are localized about the geometric shadow of the objects. The lack of a spatially separate reference beam leads to a requirement that there be a large fraction (>80%) of the beam unobscured. This limits the size of particles that can be recorded and sets an upper limit on the recordable concentration of particles (around 40 particles cm^{-3} for 20 μm diameter particles and a recorded depth of 1 m). There is also a need to balance the size of the particles with the object-to-hologram distance. The range of object size recordable on an in-line hologram has an upper limit set by the requirement to record in the far-field of the object, so that the conditions of Fraunhofer recording are satisfied:

$$z > d^2/\lambda \qquad [2]$$

where z is the object-to-film distance and d is the maximum dimension of the object to be recorded. In-line holograms are relatively easy to replay as the only requirement is to place the hologram normal to the replay beam. Following this criterion, the upper limit for good size measurement is about 2 to 3 mm particle size over a recording distance of a meter. Unlike off-axis holograms which are extremely sensitive to misalignment, the in-line geometry minimizes the effects of film rotation about its axes. A collimated recording beam is by far the easiest to duplicate upon replay. Small deviations from collimation can significantly affect conjugacy and hence resolution, introducing aberrations (mainly spherical) and a depth-dependent magnification (e.g., a cuboid replays as a trapezoidal prism).

In general, the interference field is captured on photographic emulsions; these may be laid on acetate or polyester film bases or on glass plates. If film is used, it is essential that it be held as flat as possible between thin flat glass plates or under vacuum, and should not be stretched or put under any strain during exposure. An in-line hologram does not require an extremely high-resolution recording medium. The fringe spacing is not as fine as the typical carrier fringe frequency used in off-axis holograms, and holograms can be made on conventional slow photographic emulsions. In practice, however, ultra-fine grained holographic film is almost always used since the intensity of the forward scattered light (a function of grain size) is then reduced: this is important in increasing the signal-to-noise ratio, as the replayed image is viewed against a background of the attenuated replay beam. Since the proportion of the illuminating beam that goes into the images is very low, this lack of angular separation requires that the average optical density of the hologram is high. The high optical density attenuates the undiffracted light, thus the image appears bright against a dark background, resulting in contrast reversal.

The exposed film is chemically processed to render the interference field permanent. This process is a crucial step in the holographic procedure, and especially critical for good off-axis holograms. Some of the factors, which have to be considered, include image brightness, image resolution, reconstruction wavelength, emulsion shrinkage, humidity, and noise level. It has been well established that for bright off-axis holograms on silver halide film, bleaching of the emulsion with its consequent conversion into a phase hologram is essential to maximize diffraction efficiency. Some forms of bleaching produce a nonuniform shrinkage of the emulsion which will give rise to astigmatism in the reconstructed image. Bleaching significantly increases the background noise levels of the hologram and introduces a high degree of variability into the result. In practice, the best repeatability will be obtained with amplitude processing if the low brightness can be tolerated.

Underwater Holography

When holography is extended to the inspection of underwater scenes, the hologram is recorded in water but the image is replayed, in the laboratory, in air. Because of the refractive index difference between recording and replay spaces, optical aberrations will be introduced into the reconstructed image, and will seriously impair the potential for precision measurement, unless some method of aberration reduction is incorporated. In ILH, only spherical aberration is significant since both reference and object beam angles are normal to the recording plane. However, in OAH the dominant aberrations are astigmatism and coma, which increase with the field angle of the subject in the reconstructed image. These limit resolution and introduce uncertainty in co-ordinate location, since the focal ‘point’ is now distributed over a finite region of image space. A key point here is an appreciation that the additional aberrations introduced are entirely due to the refractive index mismatch between object and image spaces and are unconnected with the holographic process itself: they are in essence the same distortions seen when looking into a fish-tank. Furthermore, the water itself may be expected to influence the quality of the images produced. An increase in the overall turbidity of the water will adversely affect both in-line and off-axis techniques and would be expected to create a background noise that will reduce image fidelity.

Techniques, such as immersing the holographic plate in water during recording or direct replay back into water, have been used to minimize underwater aberrations, but such methods are generally impractical in the field. One practical solution, unique to holography, compensates the change of effective wavelength of light as it passes through water, by allowing a deliberate mismatch of recording and replay reference beams to balance the refractive index mismatch.

For holograms recorded and replayed in air, a usual prerequisite is that the reconstruction wavelength should remain unchanged from that used in recording. By noting the dependence of wavelength on the refractive index, we can apply the more general condition that it is the ratio λ/n that must remain constant. This relationship suggests that a hologram, immersed in water, and recorded at a wavelength λ_a will produce an aberration-free image in air when replayed at a reconstruction wavelength, λ_c, which is itself equivalent to the wavelength of light when passing through water, λ_w, i.e.:

$$\lambda_c = \lambda_w = \lambda_a n_a / n_w \qquad [3]$$

where n_a and n_w are the respective refractive indices of air and water. Thus, in principle, the aberrations may be completely eliminated by the introduction of an appropriate wavelength change between recording and replay. If a green (532 nm) laser is used in recording, the ideal replay wavelength is about 400 nm (i.e., 532 nm/1.33). However, no convenient laser lines exist at exactly this wavelength. Furthermore, complete correction assumes that the entire recording system be located in water. Since this is both impractical and undesirable, holograms are usually recorded with the holographic film in air behind a planar glass window. The additional glass and air paths affect the compensation of aberrations. However, third-order aberration theory shows that if the window-to-air path length ratio is appropriately chosen for a specific replay wavelength, then aberration balancing occurs and residual aberrations are reduced to a minimum over a wide range of field angles and object locations. For an air gap of 120 mm and a BK7 window of 30 mm thickness, a blue replay wavelength of 442 nm (HeCd laser) achieves a good performance over a full field angle of about 40°.

In most underwater applications the objects (or the camera) are in motion during the exposure. The effect of this motion is to blur out the finer fringes, and thus reduce resolution and contrast. In-plane motion is the most severe, and adopting the experimentally verified criterion that the maximum allowable motion is less than one-tenth of the minimum required fringe spacing of the smallest object, then for in-line holograms the maximum object motion must be less than one-tenth of the object's dimension. For particles of 10 μm dimension, and a typical Q-switched YAG laser pulse duration of 20 ns, a high transverse velocity of up to 50 m s^{-1} can be tolerated. Off-axis holograms are more demanding in their requirements and the maximum allowed velocity is reduced to about 3 m s^{-1}. This is, however, more than adequate for most field applications of the technique.

Underwater Holographic Cameras

The first-known use of holography for recording living marine plankton was by Knox in 1966, who recorded in-line holograms (ILH), using a short coherence length ruby laser, of a variety of living marine plankton species in a tank. However, to avoid problems with refractive index and wavelength changes, the holographic plate was immersed directly in water, which is completely impractical for field use. It was not until the late 1970s that quantitative data from replayed holograms of marine particles recorded *in situ* were obtained by Carder and his team. Small volumes (10 or 20 cm^3) were recorded using ILH and a low-power HeNe laser, which limited the technique to slowly moving objects. Later, Heflinger used the off-axis geometry with both pulsed argon-ion (514 nm) and pulsed xenon (535 nm) lasers to record plankton on 35 mm film. In order to compensate for the aberrations introduced by replaying the real image in air, the real image was projected back into the water tank using a conjugate beam from a continuous-wave argon laser. The sides of their tank were flexible so that the real image could be placed close to the exit window and viewed via an external microscope. Knox later developed a ship-borne in-line holographic camera which was deployed from the *RV Ellen B Scripps*. A pulsed xenon laser was used to record water volumes of up to 10^5 cm^3 to a water depth of 100 m.

It was not until the late 1990s that significant use of underwater holography for plankton recording came to the fore again. Joseph Katz at Johns Hopkins University, Baltimore, USA developed a camera based around a ruby laser to record in-line holograms on rolls of film, in a 63 mm diameter sample volume that can be varied from 100 to 680 mm in length. The camera housing utilizes a two-tube design with one tube holding the laser and another containing film holder and power units. The recording volume is situated 890 mm above the camera to minimize disturbance to flow. The camera is designed as a 'Lagrangian drifter' to move with the water current and was successfully deployed in Chesapeake Bay, Baltimore, and Solomon's Island, Maryland. Holograms were recorded of various species of phytoplankton and zooplankton and represented a significant advance in the deployment of sub-sea holography.

From the early 1980s, a team led by Aberdeen University, Scotland, investigated the use of underwater holography for offshore applications. Primary interest was in inspection of subsea pipelines for corrosion and damage, and much of the work concentrated on an analysis of the optical aberrations introduced by recording in water and replaying in air and methods of compensating for such aberrations. However, this work was later adapted for holography of plankton and led to the consequent development of the HoloMar system. The HoloMar camera was the first to incorporate simultaneous ILH and OAH recording.

More recent developments have utilized digital or electronic recording of holograms in the in-line mode, and it is to be expected that much more use will be made of 'e-holography' in the future.

For field applications of holography, it is essential to record the holograms using a pulsed laser in order to create stable interference fringes. For subsea purposes, a laser with a wavelength in the blue-green region of the spectrum will match the oft-quoted peak transmission window of seawater. In practice, a frequency-doubled Nd-YAG laser with a wavelength of 532 nm is a good choice, although ruby lasers operating at 694 nm are also acceptable if transmission distance is not an issue. Usually OAH requires more energy and places stricter constraints on coherence properties than ILH. Beam quality is governed by quality of the optical components in the beam path and also by the output beam of the laser itself. Because off-axis holography uses diffuse front/side illumination, energies of up to several hundred millijoules are often required for large (up to a meter in diameter) poorly reflective objects, and coherence lengths of greater than a meter are needed to record large scene depths. To obtain a long coherence length, and thus enable large-volume objects to be recorded, it is necessary that the laser operate in a single longitudinal cavity mode so that the bandwidth of the laser line is kept small. To enable smooth even illumination of both subject and holographic emulsion, good spatial beam quality is needed. This usually dictates laser output beams with either a Gaussian (TEM_{00}) mode profile or a flat profile, and spatial filtering during beam expansion. Compared to off-axis holography, the demands placed upon lasers for ILH are minimal. The only essential requirement is for a good spatial mode structure. Energy requirements are moderate: as little as 10 mJ is enough since little light is lost during the passage of the beam through the recording volume. The laser coherence length can be quite short, and only depends on the

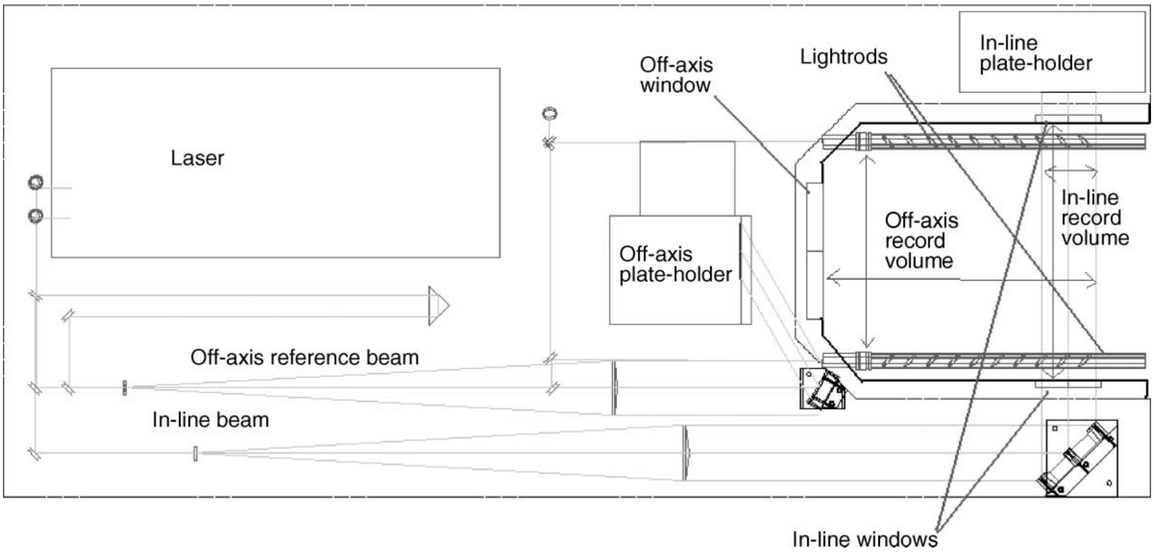

Figure 7 The HoloMar camera configuration showing the separate in-line and off-axis beam paths.

effective hologram area, thus lasers operating on more than one longitudinal mode are usually quite suitable.

The HoloMar System

The HoloMar system (shown diagrammatically in Figure 7) was developed to record simultaneous in-line and off-axis holograms of partially overlapping, but orthogonal, views of the water column. Although the recording views are orthogonal to each other, the arrangement provides an element of cross-correlation between holograms. It is designed around a Q-switched, frequency-doubled Nd-YAG laser with an output energy of 700 mJ in a single pulse of less than 10 ns duration and a coherence length in excess of 2 m. The laser output is split into two beams: one of 100 mJ and the other of 600 mJ energy. The 100 mJ beam is further split into two equal energy beams and expanded and collimated to a diameter of about 90 mm. One path forms the illuminating beam for the in-line mode and the other path forms the reference beam for the off-axis mode. The in-line path records a water column of 470 mm by 90 mm diameter (9,000 cm^3) at 400 mm from the front face of the housing, whereas the off-axis hologram records almost the entire volume (50,000 cm^3) delineated by the front window to the end of the arms that house the in-line system (see Figure 7). Three 'lightrods', encompassing hollow perspex tubes with 10 glass beam-splitter plates distributed equally along the length, positioned at 45° to the main axis, provide roughly even side illumination to the off-axis volume. The front face of the camera is shown in Figure 8; the lightrods are positioned on either side of the recording volume (two on one side, one on the other). The exit window for the in-line beam path and the off-axis window can also be seen.

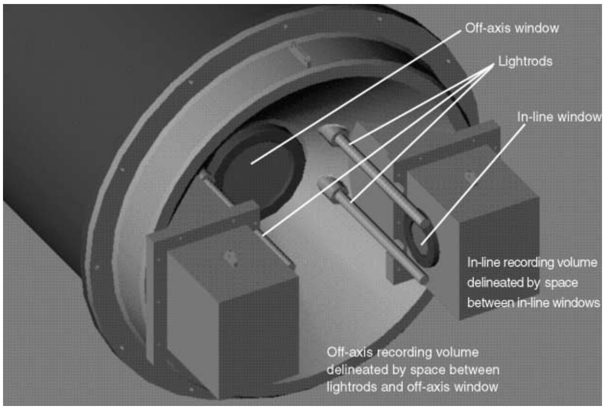
Off-axis window
Lightrods
In-line window
In-line recording volume delineated by space between in-line windows
Off-axis recording volume delineated by space between lightrods and off-axis window

Figure 8 Front-end of the HoloMar camera showing the 'lightrods' and in-line and off-axis windows.

Two motor-driven plate holders and transport mechanisms hold up to 20 plates for in-line and 24 for off-axis holograms. One pair of plates can be exposed every 10 seconds, or they can be exposed separately. The camera is remotely operated from the surface via a control console (PC). A network of micro-controllers serves as the backbone of the in-camera control system and communication between topside and camera modules is implemented using a Controller Area Network (CAN) bus via an umbilical cable. The entire camera is enclosed in a water-tight stainless steel housing of 1,000 mm diameter by 2,200 mm length. The prototype system is capable of either ship deployment or attachment to a fixed buoy and allows recording down to a water depth of 100 m.

The HoloMar camera is shown, in Figure 9, being lowered into Loch Etive, a sea loch on the west coast of Scotland, on its maiden dive. Figures 10 and 11 show holographic images of plankton recorded on this cruise. The first shows a view from a virtual

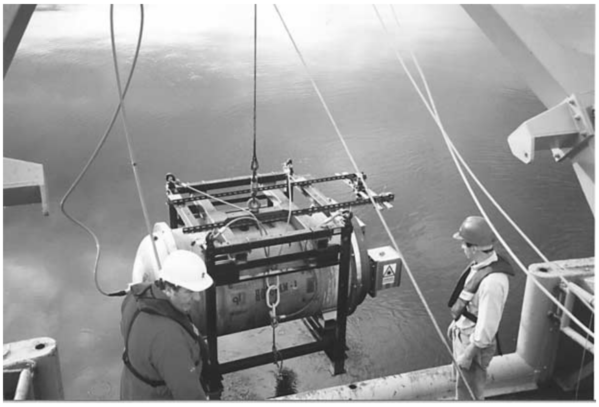

Figure 9 The HoloMar camera being lowered into Loch Etive, Scotland.

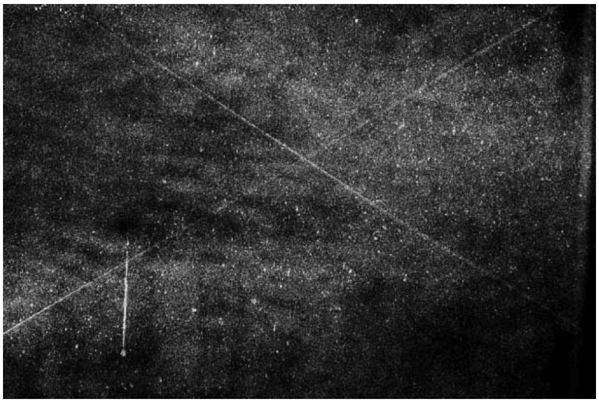

Figure 10 A photograph taken from a virtual image replay of an off-axis hologram recorded in Loch Etive.

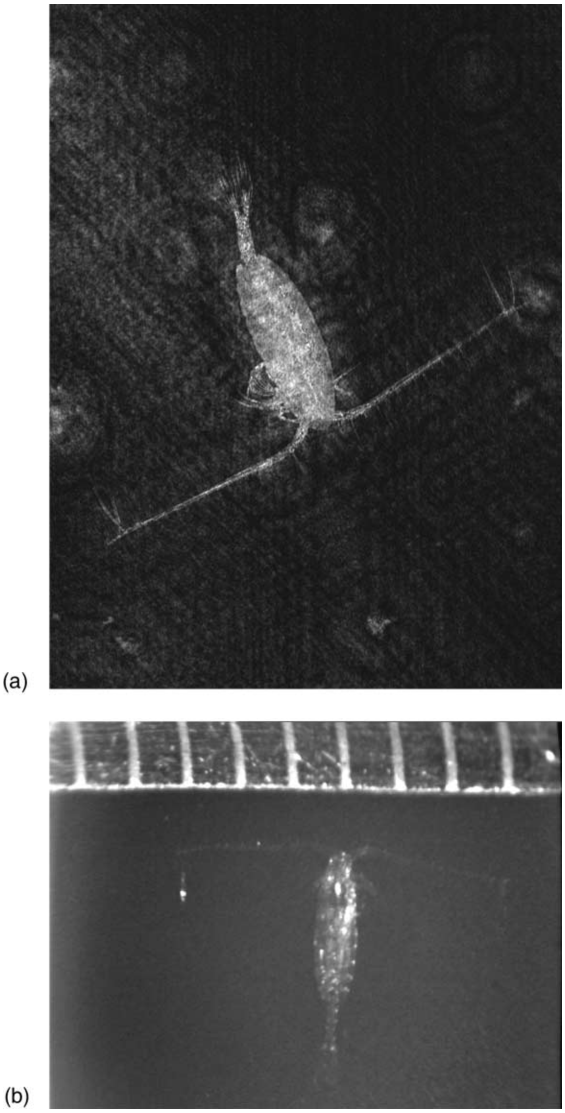

Figure 11 Images of calanoid copepods from (a) an in-line hologram recorded at 70 m depth and (b) an off-axis hologram recorded at 60 m depth, the millimeter scale is added at the replay stage. Each organism is approximately 4 mm long.

replay of an off-axis hologram at a wavelength of 514 nm: this is what a viewer would see when looking from the inside of the holocamera into the laser-illuminated water. Each bright dot signifies an object of interest; two fiducial wires can also be seen fixed diagonally across the arms at the end of the camera, some 300 mm from the off-axis window. In the lower left of the image a long 'string-like' object about 40 mm long is probably 'floc', the dead organic matter that forms the food-stuff of many zooplankton. Objects like this would be destroyed in net-collection. Figure 11 shows images taken from the real image projection of in-line and off-axis holograms. The holograms are replayed using a collimated beam at wavelengths of 514, 488, 457, or 442 nm, depending on location in the image volume. Some of the contrasting features between in-line and off-axis holograms are apparent: the in-line image shows clearly the outline of the organism with detail resolvable down to about 10 μm, whereas the off-axis hologram, although not achieving this resolution, reveals the body structure of the organism.

List of Units and Nomenclature

d	dimension of object to be recorded
D	effective aperture of hologram
n	refractive index
n_a	refractive index of air
n_w	refractive index of water
r	minimum resolvable separation between two points on the hologram
s	distance between hologram and image location
z	distance between hologram and object location
λ	wavelength
λ_a	in-air wavelength of recording beam
λ_c	reconstruction wavelength
λ_w	effective wavelength of recording beam in water

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Computer-Generated Holograms. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Beers JR, Knox C and Strickland JDH (1970) A permanent record of plankton samples using holography. *Limnology and Oceanography* 15: 967–970.

Carder KL (1979) Holographic microvelocimeter for use in studying ocean particle dynamics. *Optical Engineering* 18: 524–525.

Hariharan P (1984) *Optical Holography*, 2nd edn. Cambridge, UK: Cambridge University Press.

Heflinger LO, Stewart GL and Booth CR (1978) Holographic motion pictures of microscopic plankton. *Applied Optics* 17: 951–954.

Hobson PR and Watson J (2002) The principles and practice of holographic recording of plankton. *Journal of Optics A: Pure and Applied Optics* 4: S34–S49.

Katz J, Donathay PL, Zhang J, King S and Russell K (1999) Submersible holocamera for detection of particle characteristics and motions in the ocean. *Deep-Sea Research* 46: 1455–1481.

Kilpatrick JM and Watson J (1993) Underwater hologrammetry: reduction of aberrations by index compensation. *Journal of Physics D: Applied Physics* 26: 177–182.

Knox C (1966) Holographic microscopy as a technique for recording dynamic microscopic subjects. *Science* 153: 989–990.

Malkiel E, Alquaddoomi O and Katz J (1999) Measurements of plankton distribution in the ocean submersible holography. *Measurement Science and Technology* 10: 1141–1152.

Owen RB and Zozulya AA (2000) In-line digital holographic sensor for monitoring and characterizing marine particulates. *Optical Engineering* 39: 2187–2197.

Stewart GL, Beers JR and Knox C (1970) Application of holographic techniques to the study of marine plankton in the field and the laboratory. *Proceedings of SPIE* 41: 183–188.

Sun H, Dong H, Player MA, *et al.* (2002) In-line digital holography for the study of erosion processes in sediments. *Measurement Science and Technology* 13: L1–L6.

Watson J, Alexander S, Craig G, *et al.* (2001) Simultaneous in-line and off-axis subsea holographic recording of plankton and other marine particles. *Measurement Science and Technology* 12: L9–L15.

Holographic Recording Materials and Their Processing

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The first off-axis transmission holograms in the USA were recorded on Kodak 649F spectroscopic silver halide material (now discontinued). For the first reflection holograms, which were recorded in Russia, a very fine-grain emulsion (Lippmann emulsion) had to be prepared to resolve the fine interference pattern associated with such holograms. Since then various types of recording materials for holography have been developed and manufactured, including improved silver halide emulsions.

The following provides a brief description of the holographic recording process. For recording holograms and holographic optical elements (HOEs), dichromated gelatin, photopolymer materials, and photoresist materials are employed, in addition to silver halide materials. Existing holographic recording materials are described, as well as suitable processing methods to obtain high-quality holograms and HOEs.

Holographic Recording Materials

To produce a hologram, a micropattern, caused by interference between the object beam and the reference beam, is recorded in a light-sensitive high-resolution material. The quality of a holographic image depends on a number of factors, such as: the geometry and stability of the recording setup; the coherence of the laser light; the reference and object beam ratio; the type of hologram produced; the size of the object and its distance from the recording material; the recording material and the emulsion substrate used; the processing technique applied as well as the reconstruction conditions. The material must comply with certain requirements to be suitable for this purpose. The most important of these concerns the resolving power of the material. The recording material must be able to resolve the highest spatial frequencies of the interference pattern created by the maximal angle θ between the reference and the object beams in the recording setup (Figure 1).

If λ is the wavelength of the laser light used for the recording of a hologram and n the refractive index of the emulsion, then the closest separation d_e between the fringes in the interference pattern created by the angle θ between the reference and the object beams in the recording setup is

$$d_e = \frac{\lambda}{2n \sin(\theta/2)} \quad [1]$$

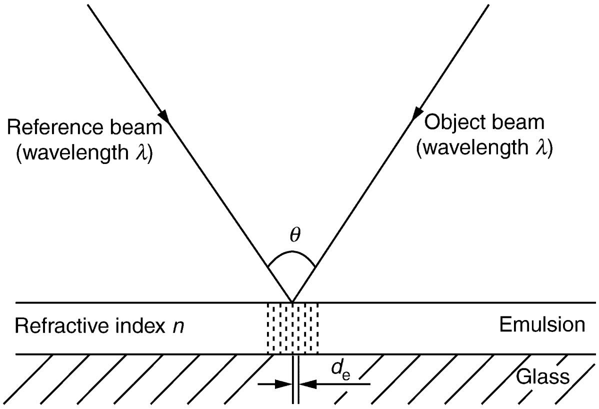

Figure 1 Demand on resolution for recording a hologram.

For high-quality holograms the resolution limit of the material must be higher than the minimum value obtained according to the above formula. For a reflection hologram recorded in blue light ($\lambda = 400\,\text{nm}$) with an angle of 180° between the beams, a minimum resolving power of 7600 lines $(\text{mm})^{-1}$ is required.

In holography, the resolution of the holographic image and the resolving power of the recording material are not directly related to the way in which they are in photography.

Equation [1] gives the minimum resolving power needed to record a hologram. The resolution of the image depends on the area (diameter D) of the recording material, the recording laser wavelength (λ), and the distance (L) between the recording material and the object (Figure 2).

Theoretically, the resolution of the holographic image is the diffraction-limited resolution that can be obtained when the information is collected over an aperture equal to the size of the recording holographic plate. In principle, the larger the holographic plate the better the resolution will be. To obtain high image resolution, it is important that the recorded interference fringes will not change or be distorted during processing. To prevent that from happening, a stable support for the emulsion (like a glass plate) is needed and the processing method applied must not affect the recorded fringe pattern in the emulsion. In coherent imaging systems laser speckles are present which may also have an effect upon image resolution.

There are two main types of holograms: amplitude and phase holograms. In a pure amplitude hologram only the absorption varies with the exposure (after processing), whereas in a pure phase hologram, either the refractive index or the emulsion thickness varies with the exposure. If the hologram is thin, $d \approx 0$, phase variations are then caused by surface relief variations only. If the hologram is thick and has a negligible surface relief ($\Delta d = 0$), phase variations are caused by index variations only. In many cases, the phase modulation in a hologram is a combination of the two different extreme types (a complex hologram).

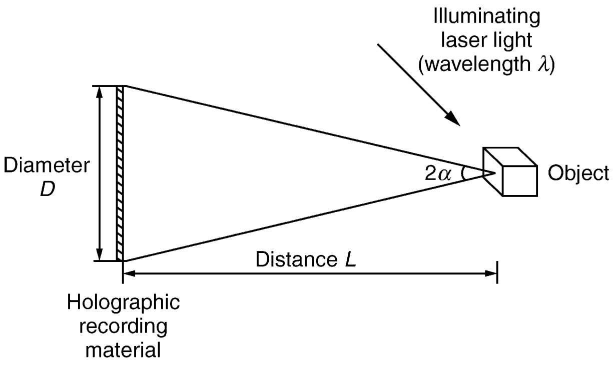
Illuminating laser light (wavelength λ)
Diameter D
2α
Object
Distance L
Holographic recording material

Figure 2 Resolution of a holographic image depends on the area (diameter D) of the recording material, the recording laser wavelength λ, and the distance L between the recording material and the object.

For amplitude holograms, amplitude transmission T_a against exposure or log exposure is used (Figure 3). For a phase hologram, the corresponding curve is the phase shift against the log exposure relation (Figure 4).

One of the most important hologram characteristics is its diffraction efficiency η. It is defined as the diffracted intensity E_i of the wanted diffraction order of the hologram in relation to the incident

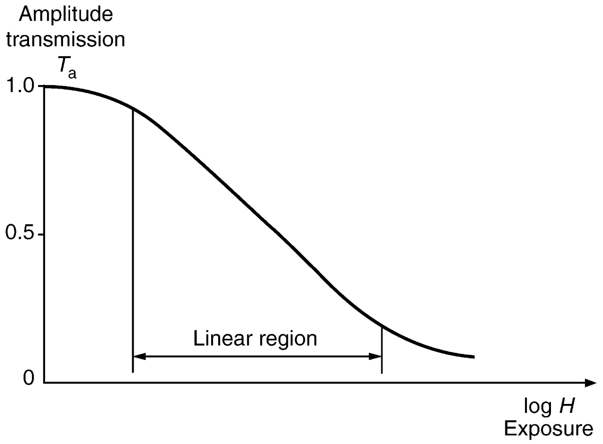

Figure 3 Amplitude transmission T_a versus the logarithm of exposure H used for the characterization of amplitude transmission holograms.

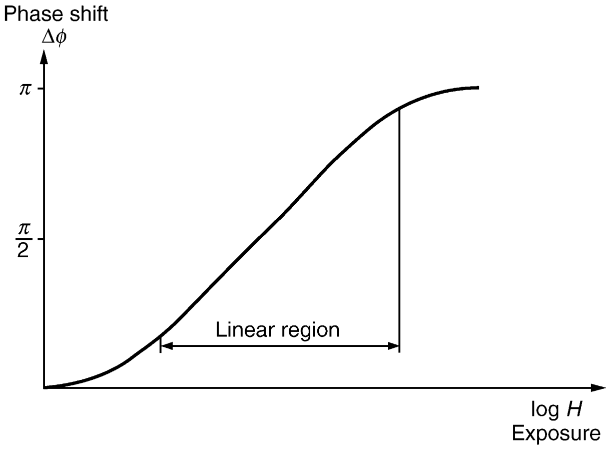

Figure 4 Phase shift $\Delta\phi$ versus the logarithm of exposure H used for the characterization of phase transmission holograms.

intensity of the reconstruction beam E_r : $\eta = E_i/E_r$. In the definition of η, the incidental light losses caused by surface reflection and base absorption should be subtracted, but are often ignored. Other important hologram characteristics are the signal-to-noise ratio (SNR) and the dynamic range of the recording material.

The resolution capability of an image reproduction process is normally described by the Modulation Transfer Function (MTF). Briefly, a test pattern containing a sinusoidal variation in illuminance combined with a continuous variation in spatial frequency along one direction is recorded. The modulation M of the pattern in the test target is

$$M = (H_{\mathrm{MAX}} - H_{\mathrm{MIN}})(\mathrm{H}_{MAX} + \mathrm{H}_{MIN})^{-1} \qquad [2]$$

where **H** is the exposure incident on the photographic material. When this pattern is recorded in the material, light scattering will take place in the emulsion, which will reduce the original contrast of the pattern. Therefore, the modulation of the pattern will be decreased, in particular at high spatial frequencies. The effective exposure modulation M' will then be

$$\mathrm{M}' = (H'_{\mathrm{MAX}} - H'_{\mathrm{MIN}})(H'_{\mathrm{MAX}} + H'_{\mathrm{MIN}})^{-1} \qquad [3]$$

where **H**$'$ is exposure in the emulsion.

The original modulation M is constant and accurately known; it is also independent of the spatial frequency. After the tested emulsion has been processed, the corresponding 'exposed' modulation is obtained from the density variations. The ratio between modulation M' in the emulsion and modulation M of the incident exposure is called the modulation transfer factor, also called response

$$R = M(M')^{-1} \qquad [4]$$

If the response is plotted as the function of spatial frequency, this curve will then be the modulation transfer function of the material. The MTF for both a holographic emulsion and a conventional photographic emulsion are illustrated in Figure 5.

There are many ways of categorizing holograms but here we are mainly interested in definitions related to the recording material and its processing. One of the valid criteria is the thickness of the recording layer (as compared to the interference fringe spacing within the layer), i.e., the layer coated on the material substrate. Holograms can thus be classified into 'thin' or 'thick' (sometimes also called 'plane' or 'volume' holograms, respectively). To distinguish between the two types, the Q-parameter is normally used; it is defined in the following way:

$$Q = 2\pi\lambda d(n\Lambda^2)^{-1} \qquad [5]$$

where λ is the wavelength of the illuminating light, d is the thickness of the layer, n is the refractive index of the emulsion, and Λ the spacing between the recorded fringes. A hologram is considered thick if $Q \geq 10$, and thin if $Q \leq 1$. Holograms with Q-values between 1 and 10 can be treated as either thin or thick.

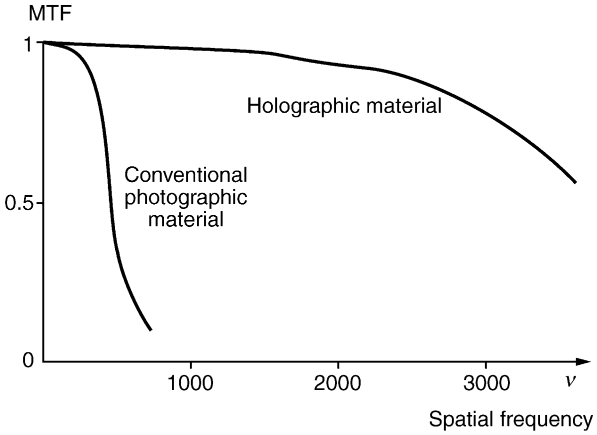

Figure 5 Graphical representation of the modulation transfer function (MTF) curves for both a conventional emulsion and a holographic emulsion are shown.

Recording Materials

The most common holographic materials used for recording holograms are the following: silver halide emulsions; dichromated gelatin emulsions; photopolymer materials; photoresist materials; thermoplastic materials; bacteriorhodopsin; and photorefractive crystals.

The following materials can also be considered for holographic recordings, but not many practical applications have been reported so far: chalcogenide glasses; ferroelectric-photoconductors; liquid crystals; magneto-optic films; metal and organic dye ablative films; photochromic and photodichroic materials; photothermorefractive glass; and transparent electrophotographic films.

Sensitivity of Photographic and Holographic Materials

The time of exposure for a given material will depend on the sensitivity of the material used, as well as on the intensity of the interference pattern. Some holographic materials must be processed after exposure in a specific way to obtain a hologram. The recorded intensity variations are converted during this processing step to local variations in optical density or refractive index and/or thickness of the recording layer.

Exposure H is defined as the incident intensity E times the time t of exposure of the recording material. If the intensity is constant during the whole exposure time, which is usually the case, then

$$H = Et \qquad [6]$$

Holographic materials are usually characterized using radiometric units. The radiometric equivalent of illuminance is irradiance. The unit of irradiance is $\mathrm{W(m)^{-2}}$ and the exposure will then be expressed in $\mathrm{J(m)^{-2}}$. The sensitivity of a holographic emulsion is most often expressed in $\mu\mathrm{J(cm)^{-2}}$ or $\mathrm{mJ(cm)^{-2}}$. Knowing the sensitivity of the material used and having measured the irradiance at the position of the holographic plate, the exposure time can be calculated using the above formula, i.e.,

$$\text{Exposure time} = \text{sensitivity/irradiance}$$

Holographic materials are sensitized in such a way that they are optimized for laser wavelengths commonly used in holography.

Silver Halide Materials

A silver halide recording photographic material is based on one type, or a combination of silver halide crystals embedded in a gelatin layer. The emulsion is coated on a flexible or stable substrate material. Silver-halide grain sizes vary from about 10 nanometers for the ultra-fine-grain holographic emulsions to a few micrometers for highly sensitive photographic emulsions (Table 1).

Silver salts are only sensitive to UV light, violet, and deep blue light. Therefore special sensitizers (dyes) must be added to the emulsion to make it sensitive to other parts of the spectrum. Orthochromatic emulsions are sensitive to green light. If the material has been sensitized to both green and red light, it is said to be panchromatic.

Table 1 Silver halide emulsion grain sizes

Type of emulsion	*Average grain diameter [nm]*
Ultra-fine-grain holographic emulsion	10–30
Fine-grain holographic emulsion	30–50
Fast holographic emulsion	50–100
Chlorobromide paper emulsion	200
Lithographic emulsion	200–350
Fine-grained photographic emulsion	350–700
Fast photographic emulsion	1000–2000
Fast medical X-ray emulsion	2500

In the following, silver halide materials manufactured by photographic and holographic companies are described. The main difference between the new materials and the previous emulsions from Agfa, Ilford, etc., is that for the current silver halide emulsions, the grain sizes are smaller. The present holographic silver halide emulsion manufacturing process is more or less based on the Russian ultrafine-grain emulsion technology.

Substrates for Holographic Emulsions

The material on which the emulsion is coated has a strong bearing on the final quality of the hologram. The best choice is often a glass plate as it is mechanically stable and optically inactive. High resolution imaging, hologram interferometry, holographic optical elements (HOEs) and spatial filters are a few examples where a very stable emulsion support is important. Mass produced holograms are mainly recorded on film substrates. However, the use of film has many advantages as compared to that of glass (breakage, weight, cost, size, etc.). Film substrates are used exclusively in the production of large-format holograms. Film substrates are of mainly two types: a polyester (polyethylene terephthalate); or a cellulose ester, commonly triacetate (cellulose triacetate) or acetate–butyrate. Polyester is mechanically more stable than triacetate film and it is also less sensitive to humidity. Because of the higher tensile strength, the polyester film can be made thinner than the triacetate film. On the other hand, polyester is birefringent, which can cause problems when recording holograms. In Table 2 most current commercial silver halide materials for holography are listed.

Processing of Silver Halide Emulsions

In silver halide emulsions a latent image is formed during exposure, which later is developed to a silver image. Processing (development, fixing, and bleaching) silver halide emulsions requires some expertise to obtain high-quality holograms. With the exception of DCG emulsions, other holographic materials are easier to process and more straightforward. Here it is not possible to describe all techniques employed and recipes used for silver halide processing. The reader is referred to the Further Reading section for more details.

Table 2 Commercial holographic recording materials

Material	*Thickness [μm]*	*Spectral sensitivity [nm]*	*Sensitivity [μJ $(cm)^{-2}$] at*				*Resolv. power [lp $(mm)^{-1}$]*	*Grain size [nm]*
			442	*514*	*663*	*694*		
SILVER HALIDE EMULSIONS:								
Slavich								
Red PFG-01	7	<700	–	–	80	–	>3000	35–40
Red PFG-03M	7	<700	–	–	1500	–	>5000	10–20
Green VRP-M	7	<550	–	80	–	–	>3000	35–40
Pan PFG-03C	9	400–700	1000	2000	1000	–	>5000	10–20
Colourholographic								
Red BB-700	7	<700	–	–	50	150	>2500	50–60
Red BB-640	7	<650	–	–	150	–	>4000	20–25
Green BB-520	7	<540	150	150	–	–	>4000	20–25
Blue BB-450	7	<470	150	–	–	–	>4000	20–25
Kodak								
Red 131PX	9	<650	2	–	0.5	–	>1250	70
Red 131CX	9	<650	2	–	0.5	–	>1250	70
Red 120PX	6	<750	60	–	40	40	>2500	58
Red 120CX	6	<750	60	–	40	40	>2500	50
FilmoTec-ORWO								
Red HF65	10	<650	–	–	1500	–	5000	30
Green HF53	10	<550	–	700	–	–	5000	30
Ultimate								
Ultimate 15	7	<700	–	150	150	150	>5000	15
Ultimate 08	7	<650	120	200	200	–	>7000	8
DICHROMATED GELATIN EMULSIONS								
Slavich								
Blue PFG-04	16	<515	1×10^5	2.5×10^5	–	–	10000	–
FilmoTec-ORWO								
Blue GF 40	6/20	*	*	–	–	–	–	–
* delivered: non-sensitized								
Holotec								
Custom orders only								
THERMOPLASTIC MATERIALS								
Tavex America								
Pan TCC-2	–	<800	1	1	1	1	1500	–
MD Diffusion								
Custom orders only								
PHOTORESIST MATERIALS								
Towne Technologies								
UV-Blue Shipley 1800	1.5–2.4	<450	1.5×10^5	–	–	–	1000	–
Hoya Corporation								
Custom orders only								
BACTERIORHODOPSIN MATERIALS								
MIB GmbH								
BR-WT B-type	30–100	<650	–	80×10^3	–	–	5000	–
BR-D96N M-type	30–100	<650	30×10^3	–	–	–	5000	–

The developing agent is the most important constituent in a holographic developer. In addition to the developing agent, a developer consists of the following components: a preservative (or antioxidant); a weak silver-solvent agent; an accelerator (or activator); a restrainer; and a solvent which is usually water. Employing a developer which contains a silver-solvent agent, solution-physical development or semi-physical development can be performed. Such a developer is particularly suitable for processing the ultrafine-grain emulsions.

Holographic Developers

Silver halide materials for holography are all of the fine-grain type. Such materials require the application of special processing techniques to achieve the best possible results. For the processing of amplitude holograms a high-contrast developer is required, such as Kodak D-19. For holograms which are going to be bleached (phase holograms), developers based on ascorbic acid, hydroquinone, metol, pyrocatechol, or pyrogallol are frequently used. The following developers are recommended for the new silver halide materials: D-19; CW-C2; AAC; MAA-3; PAAP; and for pulsed laser recordings, the SM-6 developer. Colloidal development, employing a highly diluted developer which contains silver solvents, is recommended for the ultrafine-grain emulsions. Colloidal silver grains formed in the emulsion during such development are very small which means that very little absorption occurs. The method produces holograms with high diffraction efficiency combined with very low light scattering noise. The Russian GP-2 developer is recommended for colloidal development. Recipes for common holographic developers are found in Table 3.

Bleach Baths

Bleaching is used to convert a developed hologram (containing a silver image) into a phase hologram, to obtain a higher diffraction efficiency than is possible for amplitude holograms. During bleaching the silver image is converted into a water-soluble or an insoluble silver compound. Holographic bleaching techniques can be divided into three categories:

1. Conventional or direct (rehalogenating) bleaching;
2. Fixation-free rehalogenating bleaching;
3. Reversal (complementary) or solvent bleaching.

In conventional bleaching, the developed silver image is, after fixing, converted into a transparent silver halide compound, i.e., after the unexposed silver halide crystals have been removed. This method is not recommended for ultrafine-grain emulsions. Instead the fixation-free rehalogenating bleaching is preferred. The hologram is bleached directly after development, without fixing, leaving the unexposed silver halide crystals in the emulsion. A rehalogenating bleach bath consists of: an oxidizing agent; an alkali halide (often KBr); and a buffer. Among the rehalogenating bleaches, the following bleach baths work well: PBQ-bleach; Ferric EDTA; and the PBU-bleaches (PBU-amidol or the PBU-metol).

In reversal bleaching, the developed silver image is not fixed. Instead it is converted into a soluble silver complex which is removed from the emulsion during bleaching, leaving only the original unexposed silver halide grains in the emulsion. A reversal bleach bath consists of an oxidizing agent and a buffer. For ultrafine-grain silver halide emulsions, the reversal method is not recommended. The very small unexposed silver halide grains are also affected (partly dissolved) during reversal bleaching, resulting in low diffraction efficiency. For materials with larger grains, a more diluted version of the dichromate Kodak R-9 bleach can be

Table 3 Developers for silver halide emulsions

Constituents	*CWC2*	*GP-2*	*MAA-3*	*PAAP*	*AAC*	*SM6*
Methylphenidone		0.2 g				
Catechol	10 g					
Ascorbic acid	5 g		10 g	18 g	16 g	18 g
Hydroquinone		5 g				
Metol			2.5 g			
Phenidone				0.5 g		6 g
Sodium sulfite	5 g	100 g				
Sodium phosphate (diabasic)				28.4 g		28.4 g
Urea	50 g					
Potassium hydroxide		5 g				
Sodium hydroxide				12 g		12 g
Sodium carbonate	30 g		55.6 g		60 g	
Ammonium thiocyanate		12 g				
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l
Dilution:	no	15 ml dev. + 400 ml water	no	no	no	no

Table 4 Bleach baths for silver halide emulsions

Constituents	*Ferric-EDTA (rehal.)*	*PBQ (rehal.)*	*PBU-(amidol/metol) (rehal.)*	*Copper (rehal.)*	*Chrome (reversal)*	*PSSB (reversal)*
Amidol *or* metol*			1 g			
Cupric bromide			1 g			
Potassium persulfate			10 g			20 g
Potassium dichromate					4 g	
p-Benzoquinone		2 g				
Ferric III sulfate	30 g					30 g
Copper sulfate				35 g		
Disodium EDTA	30 g					
Boric acid		1.5 g				
Citric acid			50 g			
Potassium bromide	30 g	30 g	20 g	30 g		
Sodium hydrogen sulfate	30 g			30 g		30 g
Sulfuric acid					4 ml	
Distilled water	1 l	1 l	1 l	1 l	1 l	1 l

*To be added after all other constituents have been dissolved.

used. Recipes for common holographic bleach baths are found in Table 4.

Commercial Silver Halide Emulsions

In Table 2 most of the current silver halide holographic emulsions are listed. The main manufacturer of holographic silver halide emulsions is the Micron branch of the SLAVICH Joint Stock Company photographic company located outside Moscow (2 pl. Mendeleeva, 152140 Pereslavl-Zalessky, Russia, www.slavich.com). Both film (triacetate) and glass plates are manufactured. Many different sizes are produced, including 115 cm by 10 m film rolls. Recommended holographic developers: CW-C2, SM-6. Many rehalogenating holographic bleach baths work well and, in particular, the PBU-type bleaches are recommended.

Colourholographic Ltd is a new company based in England (Braxted Park, Great Braxted, Witham CM8 3BX, England; www.colourholographic.com). The Colourholographic materials are based on the HRT emulsion, previously manufactured in Germany by Dr Birenheide. At the present time only glass plates are manufactured in England. Recommended processing solutions for the new HRT materials are a sodium-sulfite-free ascorbic acid–metol developer followed by a rehalogenating ferric EDTA or PBU bleaches. For reflection holograms, a pyrogallol developer and the ferric-EDTA bleach can be used.

Eastman Kodak's holographic materials are produced in America (343 State Street, Rochester, New York 14650, USA; www.kodak.com/go/PCB products). These plates are usually made to order only. Materials can be ordered with (marked -01) or without (marked -02) anti-halation layer. The 120 emulsion can be used for recording holograms using a pulsed ruby laser. Kodak recommends the developer Kodak D-19 for the processing.

FilmoTec is located in Germany (Chemie Park Bitterfeld-Wolfen, Röntgenstrasse, D-06766 Wolfen, Germany; www.filmotec.de). The company manufactures ORWO holographic emulsions which are coated only on triacetate film, 104 cm wide in 10 and 30 m lengths.

In France, M. Yves Gentet has started a small-scale emulsion manufacturing company: Atelier de Création d'Art en Holographie Yves Gentet (50, rue Dubourdieu, F-33800 Bordeaux, France; www.perso.wanadoo.fr/holographie) Monochromatic and panchromatic ultrafine-grain emulsions are produced called 'Ultimate'. Plates and film up to 60 cm by 80 cm are produced. An ascorbic acid–metol developer is recommended followed by a Ferric EDTA bleach. In addition, the Russian GP-2 developer can be used. Also a special developer can be ordered from the company.

Konica in Japan is manufacturing a green-sensitive emulsion for the use in the American holographic VOXEL 3D medical imaging system and, in addition, for the Japanese market only.

Dichromated Gelatin Materials

Dichromated gelatin (DCG) is an excellent recording material for volume phase holograms and HOEs. The grainless material has its highest sensitivity in the UV region, but extends into the blue and green parts of the spectrum. There are possibilities to sensitize the emulsion in the red part of the spectrum as well. However, DCG is most often exposed with blue laser wavelengths. Depending on processing parameters, diffraction efficiency and bandwidth can be controlled. It is easy to obtain high diffraction efficiency combined with a large SNR.

During the exposure of a DCG emulsion to UV or blue light, the hexavalent chromium ion (Cr^{6+}) is photo-induced to trivalent chromium ion (Cr^{3+}) which causes cross-linking between neighboring gelatin molecules. The areas exposed to light are hardened and become less soluble than the unexposed areas. Developing consists of a water wash which removes the residual or unreacted chemical compounds. Dehydration of the swollen gelatin follows after the material has been immersed in isopropanol, which causes rapid shrinkage resulting in voids and cracks in the emulsion, thus creating a large refractive index modulation. The underlying mechanism is not completely clear because high modulation can also be caused by the binding of isopropanol molecules to chromium atoms at the cross-linked sites. The DCG material has a rather low sensitivity of about $100\,\mathrm{mJ(cm)^{-2}}$. In many cases DCG plates are prepared by fixing an unexposed silver halide emulsion and then sensitizing it in a dichromate solution. It is also, for example, possible to make a 10 μm thick emulsion on a 8″ by 10″ glass plate using 1 g ammonium dichromate mixed with 3 g photographic grade gelatin in 25 ml deionized water. Then the emulsion is spin-coated on the glass substrate or the emulsion is applied by the doctor blade coating technique. Many HOE manufacturers prefer to produce their own emulsions but there are also commercial DCG emulsions available on the market.

Processing of DCG holograms are performed in graded alcohol (propanol) baths starting with water–propanol solutions of high water content and ending with pure propanol (cold or hot) baths. Depending on the wanted HOE characteristics, one has to carefully control the propanol/water mixtures as well as the temperature of the different baths. Cold baths produce better uniformity and lower noise. Warm baths can yield high index modulation but often with increased noise. The bandwidth can be controlled by the processing temperature and the ratio of propanol/water mixtures.

Commercial DCG Materials

Slavich in Russia is one manufacturer of presensitized dichromated plates for holography. The DCG emulsion is marked PFG-04. Plates up to a size of 30 by 40 cm can be ordered.

The FilmoTec DCG emulsions are only available on 190 μm triacetate film, 104 cm wide in 10 and 30 m lengths; emulsion thickness 6 or 20 μm. Note that the film needs to be sensitized in dichromate solution before recording. This means that the company only supplies large-format gelatin-coated film.

Holotec is a German-based company which offers presensitized DCG coated on both plates and film (Jülicher Strasse 191, D-52070 Aachen, Germany; www.holotec.de). The Holotec emulsion is the high-quality DCG emulsions developed by Professor Stojanoff in Aachen. The company can supply large-format DCG glass plates (m^2 size) or film (PET-Polyethylenterephtalat), pre-sensitized and ready to use.

Photopolymer Materials

Photopolymer materials have become popular for recording phase holograms and HOEs, in particular for mass production of holograms since some photopolymer materials only require dry processing techniques.

A photopolymer recording material, such as the DuPont material, consists of three parts: a photopolymerizable monomer; an initiator system (initiates polymerization upon exposure to light); and a polymer (the binder). First, an exposure is made to the information-carrying interference pattern. This exposure polymerizes a part of the monomer. Monomer concentration gradients, formed by variation in the amount of polymerization due to the variation in exposures, give rise to diffusion of monomer molecules from the regions of high concentration to the regions of lower concentration. The material is then exposed to regular light of uniform intensity until the remaining monomer is polymerized. A difference in the refractive index within the material is obtained. The DuPont material requires only a dry processing technique (exposure to UV light and a heat treatment) to obtain a hologram. The DuPont photopolymer materials have a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5″ width on a 14″ wide Mylar® polyester base which is .002″ thick. The film is protected with a .00092″ thick Mylar® polyester cover sheet.

The recording of a hologram on DuPont polymer is rather simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication a laser line scanning technique can provide the highest production rate.

The photopolymer material needs an exposure of about 10 mJ(cm)$^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 mJ(cm)$^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image. The process is simple and very suitable for machine processing using, for example, a baking scroll oven.

Commercial Pholymer Materials

Currently the only manufacturer of holographic photopolymer materials is E.I. du Pont de Nemours & Co. (DuPont Holographics, POB 80352, Wilmington, DE 19880, USA; www.dupont.com/holographics). In the past it was possible to obtain photopolymer materials from DuPont, but in 2002 the company changed its policy and today it is producing materials mainly for DuPont Holographics, but approved customers may still be able to buy material from DuPont. The reason for the restrictions is that many applications of mass-produced holograms are in the field of document security, where DuPont's photopolymers are used to produce optical variable devices (OVDs).

Photoresist Materials

Photoresist materials in holography are used for making masters for display and security holograms. The recorded relief image in the photoresist plate is then used to make the nickel shim needed as the tool for embossing holograms into plastic materials.

The exposure to actinic radiation produces changes in the photoresist layer that result in a solvency differentiation as a function of exposure. Processing is done with a suitable solvent dissolving either the unexposed or exposed regions, depending on whether the resist is of the negative or the positive type. For optical recording, positive photoresist (exposed resist removed during development) is preferred to the negative type because of the higher resolving power and low scatter. On the resulting surface, relief pattern particles of nickel are deposited by electrolysis to make a mold, which can then be used as an embossing tool. The photoresist process can be used for making transmission holograms only. If an embossed hologram is mirror-backed by using, for example, an aluminum coating process, it can be used in the reflection reconstruction mode as well. Photoresists are UV and deep-blue sensitive. The material is spincoated on glass substrates to obtain a thickness of between 0.5 and 2 μm, then it is baked at about 75 °C for 15 minutes. A typical photoresist for holography (e.g., Shipley Microposit 1350) has a sensitivity of about 10 mJ(cm)$^{-2}$. The most common laser wavelengths for recording photoresist holograms are the following wavelengths: 413 nm (Krypton-ion), 442 nm (Helium–Cadmium), and 458 nm (Argon-ion). At higher wavelengths, e.g., 488 nm, the sensitivity is rather low. Development of Shipley resist is performed in Shipley 303A developer, diluted 1 part developer + 5 parts de-ionized water. Normally only a 10 second development time is needed (under agitation). After that the plate has to be rinsed in de-ionized water for 2 minutes. The plate needs to be dried quickly with a high pressure steam of compressed nitrogen or dry, filtered compressed air.

Commercial Resist Materials

The main American manufacturer of photoresist is Shipley Co. (1457 Macarthur Rd., #101, Whitehall, PA 18052, USA). There are a few companies which make plates based on Shipley resist.

Towne Technologies, Inc. (6–10 Bell Ave., Sommerville, NJ 08876, USA; www.townetech.com) produces spin-coated plates using Shipley resist (Shipley S-1800 series). Towne plates have a sublayer of iron oxide which enhances adhesion of the resist during electroplating operations. In addition, the sublayer eliminates unwanted backscatter from the glass substrate. The plates are developed in Shipley 303A developer.

Hoya Corporation (7-5 Naka-Ochiai 2-chome, Shinjuku-ku, Tokyo, Japan; www.hoya.co.jp) is a producer of mask blanks coated with 0.5 μm thick Shipley S-1800 resist (SLW plates). Different types of glass as well as quartz plates are produced. Hoya supply plates for holography, however, custom orders only.

Thermoplastic Materials

The holographic thermoplastic is a multilayer structure coated on glass or film. The substrate is first coated with a conducting layer, e.g., evaporated gold, then a photoconductor, e.g., poly-n-vinyl carbazole, PVK, that has been sensitized with, e.g., 2,4,7-trinitro-9-fluorenone (TNF), and on top of this layer a thermoplastic coating is deposited (usually a styrene-methacrylate material). The recording of a hologram starts with a uniform charging of the surface of the thermoplastic material using a corona charger. The charge is divided between the photoconductor and the thermoplastic layer. Exposure and the consequent photogeneration in the photoconductor cause charges of opposite signs to migrate to the interface with the thermoplastic layer and the substrate. This will not change the charge, but only decrease the surface potential. Before the image can be developed, the material has to be recharged with the uniform corona charger. This step adds additional charges to the exposed areas, in proportion to the reduced potential which is also proportional to the exposure. The material is then heated to a softening temperature of the thermoplastic layer which will be deformed due to the electrostatic forces acting on it. The material is then cooled and the image is fixed as a relief surface pattern in the thermoplastic layer. The processing time after the exposure necessary to obtain the image is between ten seconds and half a minute. If the material is heated to a somewhat higher temperature, the pattern disappears and the image is erased. The thermoplastic material can now be used for the recording of another hologram, repeating the recording procedure. Such a cycling technique can be repeated hundreds of times, without any serious effect on the quality of the image. That, and the fast dry processing made such materials popular for the use in holographic cameras for nondestructive testing. The sensitivity is between 10 and 100 $\mu J(cm)^{-2}$ over the whole visible electromagnetic spectrum.

Commercial Thermoplastic Materials

In the past, several companies produced thermoplastic materials and recording equipment for holography. Since there is very little demand for such materials today, there is only one company which can deliver thermoplastic equipment for holography: Tavex America Inc. (14 Garden Road, Natick, MA 01760, USA; www.tavexamerica.com). Tavex America manufactures a thermoplastic camera, TCC-2 in which 40 mm by 40 mm plates are used. For obsolete systems, thermoplastic film can still be obtained from MD Diffusion in France (93 rue d'Adelshoffen, F67300 Schiltigheim, France).

Bacteriorhodopsin

Photochromic materials, a class to which bacteriorhodopsin (BR) belongs, have not been used much in holography. Photochromics are real-time recyclable materials which need no processing for development. They can easily be erased and reused over and over again. In holography the only successful material has been the BR film which was introduced only a few years ago. BR is a biomolecule with several applications in photonics, comprising a light-driven molecular proton pump with exceptional photoelectric and photochromic properties. BR is a photochromic protein, found in the photosynthetic system of a salt-marsh bacterium, termed purple membrane (PM), of the cell membrane of *Halobacterium salinarum*. BR thin films are made from polymer suspensions of the isolated purple membrane. Some of the desirable properties of BR films are real-time writing and erasing in microsecond time-scales, durability allowing millions of write/erase cycles without degradation, very high spatial resolution of $5000\ lp(mm)^{-1}$, reasonably good light sensitivity of approximately $10\ mJ(cm)^{-2}$, the encoding of both amplitude and phase patterns, and the use of visible laser light for writing and reading. The composition of the films may also be altered by chemical means to control the optical characteristics, and the BR molecules themselves may be mutagenically changed for performance optimization. Particular properties that can be enhanced in these ways are image lifetimes and the diffraction efficiency of holograms recorded in the films. One advantage of this material is that it is reversible. Therefore it has been considered to be used in holographic storage systems and for real-time holographic interferometry applications. BR film can be used for recording polarization holograms because of BR films inherent photo-inducable optical anisotropy. The material is also possible to expose with pulsed lasers, which makes it suitable for fast holographic recording.

The BR film works in the following way. When illuminated with green light, the BR molecule undergoes an absorption resulting in a maximum shift of about 160 nm towards the red and its color changes from purple to yellow. The yellow image can be erased by illuminating the BR material with blue light. The recording–erasing cycle can be repeated about a million times. Suitable BR-films for holography are obtained by embedding purple membranes into inert matrices like polyvinylalcohol or polyacrylamide. For hologram recording and erasure two photochemical conversions $B \rightarrow M$ (B-type holograms) and $M \rightarrow B$ (M-type holograms) can be employed. After excitation by light, BR cycles through a sequence of spectrally distinguishable intermediates and returns to the initial state. The time associated with this thermal relaxation process depends on several aspects, for example, type of BR (wildtype or variant), pH value of film-matrix, film-temperature, etc. For example, the M intermediate is populated from the B state (absorption spectrum peaked at 570 nm) by a green pumping beam and information is recorded with blue light, which initiates the photochemical $M \rightarrow B$ transition. The M state has its absorption peak at 410 nm. In BR variants in which the thermal relaxation of the M state has been prolonged, photocontrolled switching between these two states can be achieved. Only when the pumping beam is incident upon the BR film can information be recorded with blue light. The pumping beam can be used as a readout beam for the recorded hologram.

Commercial BR Materials

A holographic camera based on a BR film is manufactured in Germany. The holographic system, FringeMaker-*plus*, developed at the University of Marburg is marketed by Munich Innovation Biomaterials (MIB) GmbH. (Hans-Meerwein-Str., D-35032 Marburg, Germany; www.mib-biotech.de). MIB is also a manufacturer of both wildtype and genetically modified BR film for holography and other scientific/technical applications. BR-films offered by MIB can be used in two different recording configurations: the transition from the initial B to M-state (induced with yellow light) serves as basis for optical data recording and processing (B-type recording); or the photochemically induced transition from M to B state (M-type recording). In the latter case, yellow light is used to control the population of the M-state and blue light is used to write the information. MIB offers BR-films with normal and slow thermal relaxation characteristics. Typical relaxation time ranges (RTR), after which the photochemically excited BR molecules have returned to the original state, are between 0.3 and 80 s. The BR film is sealed between two windows of high-quality optical glass. The BR film layer has a thickness of 30–100 μm, dependent on the optical density of the film. The diameter of the film device is 25 mm and the clear aperture is 19 mm.

Photorefractive Crystal Materials

Photorefractive crystals are electro-optic materials which are suitable for recording volume phase holograms in real-time. These crystals are important components in holographic storage systems and are seldom used in other holographic image applications. Holograms recorded in crystals, such as: lithium niobate ($LiNbO_3$); barium titanate ($BaTiO_3$); potassium tantalate niobate (KTN); barium sodium neobate (SBN); bismuth silicon oxide (BSO); and bismuth germanium oxide (BGO), consist of bulk space charge patterns. An interference pattern acting upon a crystal will generate a pattern of electronic charge carriers that are free to move. Carriers are moving to areas of low optical illumination where they get trapped. This effect will form patterns of net space charge, creating an electrical field pattern. As these crystals have strong electro-optic properties, the electrical field pattern will create a corresponding refractive index variation pattern, which means a phase hologram. These holograms can be immediately reconstructed with a laser beam differing from the beams creating the hologram. A hologram can also be erased and a new hologram can be recorded in the crystal. Holograms can be fixed by converting the charge patterns to the patterns of ions which are not sensitive to light. In some materials the hologram can be fixed by heating the crystal to above 100 °C based on thermally activated iconic conductivity. The resulting iconic is frozen upon cooling the crystal back to room temperature. More details about these materials and storage applications are found in the Further Reading.

See also

Holography, Techniques: Overview.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. Heidelberg: Springer-Verlag.

Bjelkhagen HI (ed.) (1996) *Selected Papers on Holographic Recording Materials*. SPIE Milestone Series, vol. MS130. Bellingham, WA, USA.

Lessard RA (ed.) (1995) *Selected Papers on Photopolymers, Physics, Chemistry, and Applications*. SPIE Milestone Series, vol. MS114. Bellingham, WA, USA.

Sincerbox GT (ed.) (1994) *Selected Papers on Holographic Storage*. SPIE Milestone Series, vol. MS95. Bellingham, WA, USA.

Smith HM (ed.) (1977) *Holographic Recording Materials*, Springer Topics in Applied Physics, vol. 20. Berlin: Springer-Verlag.

HOLOGRAPHY, TECHNIQUES

Contents

Overview
Color Holography
Computer-Generated Holograms
Digital Holography
Holographic Interferometry
Sandwich Holography and Light in Flight

Overview

C Shakher and A K Ghatak, Indian Institute of Technology, New Delhi, India

© 2005, Elsevier Ltd. All Rights Reserved.

Major Milestones

1948:	Essential concept for holographic recording by Dennis Gabor.
1960:	The first successful operation of a laser device by Theodore Maiman.
1962:	Off-axis technique of holography by Leith and Upatnieks.
1962:	Yu N Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.
1964:	Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.
1969:	S A Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications.

Holography is the science of recording an entire optical wavefront, both amplitude and phase information, on appropriate recording material. The record is called a hologram. Unlike conventional photography, which records a three-dimensional scene in a two-dimensional format, holography records true three-dimensional information about the scene.

Holography was invented by Gabor in 1948 and his first paper introduced the essential concept for holographic recording – the reference beam. Holography is based on the interference between waves and, it provides us with a way of storing all the light information arriving at the film in such a way that it can be regenerated later.

The use of the reference beam is utilized because the physical detectors and recorders are sensitive only to light intensity. The phase is not recorded but is manifest only when two coherent waves of the same frequency are simultaneously present at the same location. In that case, the waves combine to form a single wave whose intensity depends not only on intensities of the two individual waves, but also on the phase difference between them. This is key to holography. The film record, or hologram, can be considered as a complicated diffraction grating. Holograms bear no resemblance to conventional photographs in that an image is not actually recorded. In fact, the interferometric fringes which are recorded on the recording material are not visible to an unaided eye because of extremely fine interfringe spacing (~0.5 micrometer). The fringes which are visible on the recording material are the result of dust particles in the optical system used to produce the hologram.

Gabor's original technique is now known as in-line holography. In this arrangement, the coherent light source as well as the object, which is a transparency containing small opaque details on a clear background, is located along the axis normal to the photographic plate. With such a system, an observer focusing on one image observes it superposed on the out-of-focus twin image as well as a strong coherent background. This constitutes the most serious problem of Gabor's original technique.

The first successful technique for separating the twin images was developed by Leith and Upatnieks in 1962. This is called the off-axis or side band technique of holography. We shall consider mainly the off-axis technique here.

Basic Holography Principle

Light from a laser or any light beam is characterized by spatial coherence (l_c) and temporal coherence (τ_c), which is discussed in all textbooks on optics. For typical laboratory hologram recording, the required degree of coherence of laser radiation is determined by the type of object and geometrical arrangement being used for the recording. The following condition must hold:

$$L \ll c\tau_c \qquad [1]$$

where τ_c is the coherence time, and L the maximum path length difference between the two waves chosen to record the hologram. To begin recording, two wavefronts are derived from the same laser source. One wavefront is used to illuminate the object and another is used as a reference wavefront. The wavefront derived by illuminating the object is superimposed with the reference wave and the interference pattern is recorded. These object and reference waves must:

- be coherent (derived from the same laser); and
- have a fixed relative phase at each point on the recording medium.

If the above conditions are not met, the fringes will move during the exposure and the holographic record will be smeared. Therefore, to record the hologram one should avoid air currents and vibrations. The recording medium must have sufficient resolution to record the fine details in the fringe pattern. This will be typically of the order of the wavelength. To create a three-dimensional image from the holographic process one then has to record and reconstruct the hologram. The object and reference waves are derived from the same laser.

Holograms record an interference pattern formed by interference of object and reference wavefronts, as explained above. We may mention here that for the two plane waves propagating at an angle θ between them, the spacing of the interference fringes is given by

$$\alpha = \frac{\lambda_0}{2\sin(\theta/2)} \qquad [2]$$

where λ_0 is the free space wavelength.

Processing of the hologram (developing, fixing, and washing of the recording material) yields a plate with alternating transparent and opaque parts, variation of refractive index, or variation of height corresponding to intensity variation in the fringe pattern. Such a plate can be regarded as a complicated diffraction grating. In this process, the hologram is illuminated with monochromatic, coherent light. The hologram diffracts this light into wavefronts which are essentially indistinguishable from the original waves which were diffracted from the object. These diffracted waves produce all the optical phenomena that can be produced by the original waves. They can be collected by a lens and brought into focus, thereby forming an image of the original object, even though the object has since been removed. If the reconstructed waves are intercepted by the eye of an observer, the effect is exactly as if the original waves were being observed; the observer sees what appears to be the original object in true three-dimensional form. As the observer changes his viewing position, the perspective of the image scene changes; parallactic effects are evident and the observer must refocus when the observation point is changed from a near to a distant object in the scene. Assuming that both the construction and reconstruction of the hologram are made with the same monochromatic light source, there is no visual test which can be made to distinguish between the real object and the reconstructed image of the object. It is as if the hologram were a window through which the apparent object is viewed.

Hologram of a Point Object

Consider a hologram recorded with a collimated reference wave normal to the recording plate and a point object inclined at a certain angle. If the hologram is illuminated once again with the same collimated reference wave, it reconstructs two images, one virtual true image and the other real image. However, the two images differ in one very important respect.

While the virtual image is located in the same position as the object and exhibits the same parallax properties, the real image is formed at the same distance from the hologram but in front of it. Corresponding points on the real and virtual images are located at equal distances from the plane of the hologram; the real image has the curious property that its depth is inverted. Such an image is not formed with a normal optical system; it is therefore called a pseudo image as opposed to a normal or orthoscopic image.

This depth inversion results in conflicting visual clues, which make viewing of the real image psychologically unsatisfactory. Thus, if O_1 and O_2 are two elements in the object field, and if O_1 blocks the light scattered by O_2 at a certain angle, the hologram records information only on the element O_1 at this angle and records no information about this part of O_2. An observer viewing the real image from the corresponding direction then cannot see this part of O_2, which, contrary to normal experience, is obscured by O_1, even though O_2 is in front of O_1.

Production of an Orthoscopic Real Image

An orthoscopic real image of an object can be produced by recording two holograms in succession. In the first step, a hologram is recorded of the object with a collimated reference beam. When this hologram is illuminated once again with the collimated reference beam that was used to record it, it reconstructs two images of the object at unit magnification, one of them being an orthoscopic virtual image, while the other is a pseudoscopic real image. A second hologram is then recorded of this real image with a second collimated reference beam.

When the second hologram is illuminated with a collimated beam it reconstructs a pseudoscopic virtual image located in the same position as the real image formed by the second hologram is an orthoscopic image. Since a collimated reference beam is used throughout, the final real image is the same size as the original object and free from aberrations.

In addition to these characteristic linked intensities with the three-dimensional nature of the reconstruction, the holographic recording has several other properties. Each portion of the hologram can reproduce the entire image scene. If a smaller and smaller portion of the hologram is used for reconstruction, there is loss of image intensity and reconstruction. When a hologram is reversed, such as in contact printing processes, it will still reconstruct a positive image indistinguishable from the image produced by the original hologram.

Simple Mathematical Description of Holography

Let the object and reference waves be given by

$$U_o = O(x, y)e^{i\phi(x,y)} \qquad [3]$$

and

$$U_r = R\, e^{iky\sin\theta} \qquad [4]$$

U_r describes the reference (plane wave) propagating at angle θ to the z-axis.

The intensity at the hologram plane is

$$\begin{aligned} I &= |U_r + U_o|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \end{aligned} \qquad [5]$$

As can be seen from the above equation, the amplitude and phase of the wave are encoded as the amplitude and phase modulation of a set of interference fringes. The material used to record the patterns of interference fringes described above is assumed to provide linear mapping of the intensity incident during the reconstruction process into amplitude transmitted by or reflected from the recorded material. Usually both light detection and wavefront modulation are performed by photographic plate/film. We assume the amplitude transmission properties of the plate/film after processing to be described by

$$T = T_o - b(E - E_o) \qquad [6]$$

where the exposure E, at the film is $E = I\tau$; here τ is the exposure time. T_0 is the transmittance of the unexposed plate.

If the hologram is reilluminated by the reference wave, the transmitted wave amplitude will be

$$U_t = U_r(T_o - b\tau(|U_o|^2 + U_r^* U_o + U_r U_o^*)) \qquad [7]$$

The first term is reference wave times a constant. The second term is the reference wave modulated by $|U_o|^2 = O(x, y)^2$; it gives small-angle scattering about the reference wave direction. The third term is proportional to U_o. It is the same as the original object wave (note this is only so if the reconstruction wave is identical to the reference wave). The fourth term is

$$-b\tau U_r^2 U_0^* = -b\tau R^2\, e^{iky2\sin\theta} O(x, y)e^{-i\phi(x,y)} \qquad [8]$$

This is essentially a wave traveling in a direction $\sin^{-1}(2\sin\theta)$ to the z-axis, with the correct object amplitude modulation but its phase reversed in sign, producing a conjugate wave.

Types of Holograms

Primarily, the holograms are classified as thin and thick, based on the thickness of the recording medium. When the thickness of the recording medium is small compared with the average spacing of the interference fringes, the hologram can be treated as a thin hologram. Such holograms are characterized by spatially varying complex amplitude transmittance:

$$\mathbf{t}(x, y) = [\mathbf{t}(x, y)]\exp[-\mathrm{i}\phi(x, y)] \qquad [9]$$

Thin holograms can be further categorized as thin amplitude holograms or thin phase holograms. If amplitude transmittance of the hologram is such that $\phi(x, y)$ is constant while $\mathbf{t}(x, y)$ varies over the hologram, the hologram is termed an amplitude hologram. For a lossless phase hologram, $|\mathbf{t}(x, y)| = 1$, so that the complex amplitude transmittance is caused by variation in phase.

When the thickness of the hologram recording material is large compared to the fringe spacing of the interference fringes, the holograms may be considered as volume holograms. These may be treated as a three-dimensional system of layers corresponding to a periodic variation of absorption coefficient or refractive index, and the diffracted amplitude is at a maximum when Bragg's condition is satisfied. In general, the behavior of thin and thick holograms corresponds to Raman Nath and Bragg diffraction regimes. The distinctions between two regimes is made on the basis of a parameter Q, which is defined by the relation:

$$Q = \frac{2\pi\lambda_o d}{n_o \Lambda^2} \qquad [10]$$

where

Λ = spacing of fringe on hologram, measured normal to the surface;
d = thickness of recording medium;
n_o = mean refractive index of the medium; and
λ_o = wavelength of light.

Small values of Q ($Q < 1$) correspond to thin gratings, while large values of Q ($Q > 10$) correspond to volume gratings. For values of Q between 1 and 10, intermediate properties are exhibited. However, this criterion is not always adequate. The boundaries between the thin and thick holograms are given by the value of

$$P = \frac{\lambda_o^2}{\Lambda^2 n_o n_1} \qquad [11]$$

where P^{-2} is the relative power diffracted into higher orders, and $P < 1$ for thin holograms and $P > 10$ for thick holograms. For P having values between 1 and 10, the hologram may be thin or thick, depending on the other parameters involved.

Holograms can also be classified based on whether they are reconstructed in transmitted light or reflected light. For transmission holograms, during the recording stage, two interfering wavefronts make equal but opposite angles to the surfaces of recording medium and are incident on it from the same side. Reflection holograms reconstruct images in reflected light. They are recorded such that the interfering wavefronts are symmetrical with respect to the surface of the recording medium but are incident on it from opposite sides. When the angle between the interfering wavefronts is maximum (180°), the spacing between the fringe planes of the recording medium is minimum. Under such conditions, reflection holograms may have wavelength sensitivity high enough to be reconstructed, even with white light.

An important development in the field of holography was the invention of rainbow holograms by Benton in 1969. This provides a method for utilizing white light for illumination when viewing the holograms. The technique does so by minimizing the blur introduced by color dispersion in transmission holograms, at the price of giving up parallax information in one dimension.

Recording Materials

An ideal recording medium for holography should have a well matching spectral sensitivity corresponding to available laser wavelengths, linear transfer characteristics, high resolution, and low noise. It should also be either inexpensive or indefinitely recyclable. Toward achieving the above-mentioned properties, several materials have been studied, but none has been found so far that meets all the requirements. Materials investigated include: silver halide photographic emulsion; dichromated gelatin plates/films; silver halide sensitized gelatin plates/films; photo-resists; photo polymer systems; photochromics; photo thermoplastics; and ferro-electric crystals. Recently, the use of storage and processing capabilities of computers, together with CCD cameras, has been used for recording holograms.

Silver halide photographic emulsions are a commonly used recording material for holography, mainly because of relatively high sensitivity and easy availability. Manufacture, use, and processing of these emulsions have been well standardized. The need for wet processing and drying may constitute a major drawback. Dichromated gelatin can be considered an almost ideal recording material for volume phase holograms, as it has a large refractive index modulation capability, high resolution, low absorption and scattering.

Because of these features, dichromated gelatin has been extensively investigated. The most significant disadvantage is its comparatively low sensitivity. Pennington *et al.* developed an alternative technique, which combines the advantage of silver halide emulsions (high sensitivity) with that of dichromated gelatin (low absorption scattering and high stability). It involves exposing silver halide photographic emulsion and then processing it, so as to obtain a volume phase hologram consisting solely of hardened gelatin.

Thin phase holograms can be recorded on photoresists, which are light-sensitive organic films yielding relief image after exposure and development. They offer advantages of easy replication using thermoplastics but are slow in response and undergo nonlinear effects at diffraction efficiency greater than ~0.05. Shipley AZ-1350 is a widely used photoresist, with maximum sensitivity in the ultraviolet, dropping rapidly for longer wavelengths towards the blue.

Photopolymers are being keenly investigated because they offer advantages such as ease of handling, low cost, and real-time recording for the application of holography and nonlinear optics. However, they have low sensitivity and short shelf life.

Thin phase surface relief holograms can be recorded in a thin layer of thermoplastics. These materials have a reasonably high sensitivity over the whole visible spectrum, fairly high diffraction efficiency, and do not require wet processing. Their application in holographic interferometry, optical information processing, and in making compact holographic devices has been widely reported.

Photorefractive crystals such as $Bi_{12}SiO_{20}$, $LiNbO_3$, $Bi_{12}GeO_{20}$, $BaTiO_3$, $LiTaO_3$, etc., as recording materials, offer high-angular sensitivity and provide capability to read and write volume holographic data in real time. Besides the materials discussed here, several other materials have been investigated for recording holograms; these include photocromics, elastomere devices, magneto-optic materials, etc.

Application of Holography

Holography can be constructed not only with the light waves of lasers, but also with sound waves, microwaves, and other waves in the electromagnetic spectrum of radiation. Holograms made with ultraviolet light or X-rays can record images of objects/particles smaller than the wavelength of visible light, e.g., atoms or molecules. Acoustical holography uses sound waves to see through solid objects. Holography has a vast scope of practical applications, which have been classified into two major categories:

1. applications requiring three-dimensional images for visual perception; and
2. applications in which holography is used as a measuring tool.

The unique ability of holography – to record and reconstruct both electromagnetic and sound waves – makes it a valuable tool for education, science, industry, and business. Below are some of the important applications:

1. Holographic interferometry (HI) is one of the most powerful measuring tools. In HI, two states of an object, i.e., initial and deformed state are recorded on the same photographic plate. After reconstruction of the light wave corresponding to two states of an object, interferences and deformations are displayed in terms of the interference pattern. The change of distance of one tenth of a micron, or lower, can be resolved. HI provides scientists/engineers with crucial data for design of critical machine parts of power-generating equipment, in the aircraft industry, automobile industry, and nuclear installations (say, for example, in the design of containers used to transport nuclear materials, improve the design of aircraft wings and turbine blades, etc.). Presently, HI is being widely used in mechanical engineering, acoustics, and aerodynamics, for nondestructive testing, to investigate oscillation in diaphragms and flow around various objects, respectively.
2. Microwave holography can detect objects deep within spaces, by recording the radio waves they emit.
3. Another important application of holography is the design of optical elements, which possess special properties. A holographic recording of a concave mirror behaves in much the same way as the mirror itself, i.e., it can focus the light. In some cases chromatism can be introduced in the design of elements so that a location of the point, where the beams are focused, depends on the wavelength. This can be achieved by accurately choosing the recording arrangement of the focusing elements, these elements are, in fact, diffraction gratings. These can have low noise levels, freedom from astigmatism, and have other useful properties. Holographic optical elements have found applications in supermarket scanners to read barcodes, headup displays in fighter aircraft to observe critical cockpit instruments, etc.
4. A telephone credit card used in Europe and other developed countries has embossed surface holograms which carry a monetary value. When the card is inserted into the telephone, a card reader discerns the amount due and deducts the appropriate amount to cover the cost of the call.

5. Holography is having applications in analog and digital computers, offering remarkable opportunities to realize various logical operators, devices for identifying images based on matched filtering, and in computer memory units. The basic advantage of holographic memory is that a relatively large volume of information can be stored and that there are a limited number of ways to change the record. The arrival of the first prototype of optical computers, which use holograms as storage material for data, could have a dramatic impact on the holographic market. The yet-to-be-unveiled optical computers will be able to deliver trillions of bits of information faster than the current generation of computers.
6. Optical tweezers are going to become an important tool for the study of micro-organisms/bacteria, etc.
7. Holograms can be used to locate and retrieve information without knowing its location in the storage medium, only needing to know some of its content.
8. The links between computer science and holography are now well-established and are developing, with at least two aspects making computer-generated holograms extremely interesting. First, such holograms enable us to obtain visual 3-dimensional reconstructions of imagined objects. For example, one can reconstruct three dimensions of a model of an object still in the design stage. Second, computer-generated holograms can be used to reconstruct lightwaves with specified wave fronts. This means specially computed and manufactured holograms may function as optical elements that transform the incident light wave into desired wavefronts.
9. Another important applications of holography is its utilization to compensate for the distortion that occurs when viewing objects through optically heterogeneous mediums. It can be achieved based on the principle of beam reversal by the hologram.
10. Finally, let us consider the application of holography to art. The development of holography gives very effective ways of creating qualitative three-dimensional images. Thus, a new independent area of holographic creative work representational/artistic holography has appeared. The art of holographic depiction has developed along two major routes. The first is creation of the view hologram, used as holograms of natural objects that are to be displayed in exhibitions and museums; these are also known as artistic holograms. Portrait holography is also classified under this category. The progress in portrait holography is hampered partly because of imperfection of pulsed lasers and partly because of the deterioration of photographic material when exposed to pulsed electromagnetic radiation.

The principle behind the creation of elusion by using composite holograms is also very convincing for the display of objects. To synthesize the composite holograms, the photographs of various aspects of a scene are printed onto the photographic plate. The synthesis techniques for preparing the composite hologram are very complicated, while the images created by holograms are still far from perfect. However, there is no doubt that composite holography opens holography up as an artistic technique. Recently rainbow holography has been very popular to display such objects.

See also

Holography, Applications: Art Holography. **Holography, Techniques:** Color Holography; Computer-Generated Holograms; Digital Holography; Holographic Interferometry. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Benton SA (1969) On a method for reducing the information content of holograms. *Journal of the Optical Society of America* 59: 1545.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Sov. Phys.-Dokl.* 7: 543.

Denisyuk YuN (1963) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt. I. *Optical Spectroscopy* 15: 279.

Denisyuk YuN (1965) On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt II. *Optical Spectroscopy* 18: 152.

Gabor SD (1948) A new microscopic principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society* A197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceedings of the Physical Society* B64: 449.

Hariharan P (1983) *Optical Holography: Principle, Techniques and Applications*. Cambridge, UK: Cambridge University Press.

Leith EN and Upatnieks J (1962) Reconstructed wavefronts and communication theory. *Journal Optical of the Society of America* 52: 1123.

Leith EN and Upatnieks J (1963) Wavefront reconstruction with continuous-tone objects. *Journal of the Optical Society of America* 53: 1377.

Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295.

Pennington KS and Lin LH (1965) Multicolor wavefront reconstruction. *Applied Physics Letters* 7: 56.

Stroke GW and Labeyrie AE (1966) White-light reconstruction of holographic images using the Lippmann-Bragg diffraction effect. *Physics Letters* 20: 368.

Color Holography

H I Bjelkhagen, De Montfort University, Leicester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the appearance of the first laser-recorded monochromatic holograms in the 1960s, the possibilities of recording full-color high-quality holograms have now become a reality. Over many years holographic techniques have been created to produce holograms with different colors, referred to as pseudo- or multicolor holograms. For example, in the field of display holography, it is common to make multiple-exposed reflection holograms using a single-wavelength laser and emulsion thickness manipulation to obtain an image with different colors. In this way many beautiful pseudocolor holograms have been made by artists. Applying Benton's rainbow hologram technique (for example, the one used for embossing holograms on credit cards), it has been possible to record transmission holograms as well as mass-produce embossed holograms with 'natural' colors. However, a rainbow hologram has a very limited viewing position from which a correct color image can be seen.

This article will acquaint the reader with the topic of color holography (full-color or true-color holography). This describes a holographic technique to obtain a color 3D image of an object where the color rendition is as close as possible to the color of the real object. A brief review of the development of color holography, the current status and the prospects of this new 3D imaging technique, are presented. Also included are recording techniques using three lasers providing red, green and blue (RGB) laser light as well as computer-generating techniques for producing full parallax 3D color images.

The Development of Color Holography

In theory, the first methods for recording color holograms were established in the early 1960s. Already Leith and Upatnieks (1964) proposed multicolor wavefront reconstruction in one of their first papers on holography. The early methods concerned mainly transmission holograms recorded with three different wavelengths from a laser or lasers, combined with different reference directions to avoid cross-talk. Such a color hologram was then reconstructed (displayed) by using the original laser wavelengths from the corresponding reference directions. However, the complicated and expensive reconstruction setup prevented this technique from becoming popular. More suitable for holographic color recording is reflection holography, which can be reconstructed and viewed in ordinary white light.

Most likely, the three-beam transmission technique might eventually become equally applicable, provided inexpensive multicolor semiconductor lasers become widely available. Such RGB lasers can be used to display a very realistic, deep, and sharp holographic image of, for example, a room.

Relatively few improvements in color holography were made during the 1960s and 1970s. Not until the end of the 1970s did a few examples of color holograms appear. During the 1980s, new and improved techniques were introduced. A review of various transmission and reflection techniques for color holography was published by Hariharan (1983). From 1990 until today, many high-quality color holograms have been recorded, mainly due to the introduction of new and improved panchromatic recording materials. On the market are ultra-fine-grain silver halide emulsions (manufactured by Slavich in Russia) as well as photopolymer materials (manufactured by E. I. du Pont de Nemours & Co. in the USA).

Color reflection holography presents no problems with regard to the geometry of the recording setup, but the final result is highly dependent on the recording material used and the processing techniques applied. There are some problems associated with recording color holograms in general and specifically in silver halide emulsions:

- Scattering occurring in the blue part of the spectrum typical of many holographic coarse-grain silver halide emulsions makes them unsuitable for the recording of color holograms.

- Multiple storage of interference patterns in a single emulsion reduces the diffraction efficiency of each individual recording. The diffraction efficiency of a three-color recording in a single-layer emulsion is lower than a single wavelength recording in the same emulsion.
- During wet processing, emulsion shrinkage frequently occurs, causing a wavelength shift. White-light-illuminated reflection holograms normally show an increased bandwidth upon reconstruction, thus affecting the color rendition.
- The fourth problem, related to some extent to the recording material itself, is the selection of appropriate laser wavelengths and their combination in order to obtain the best possible color rendition of the object.

In the beginning, when no suitable panchromatic emulsion existed, the sandwich technique was used to make color reflection holograms. Two plates were sandwiched together, in which, for example, two different types of recording materials were used. The most successful demonstration of the sandwich recording technique was made by Kubota (1986) in Japan. He used a dichromated gelatin (DCG) plate for the green (515 nm) and the blue (488 nm) components, and an Agfa 8E75 silver halide plate for the red (633 nm) component of the image. Kubota's sandwich color hologram of a Japanese doll recorded in 1986, demonstrated the potential of high-quality color holography for the first time. Color holograms have also been recorded in red-sensitized DCG materials.

Extensive work in the field of reflection color holography was performed by Hubel and Solymar (1991). Primarily they employed Ilford silver halide materials for the recording of color holograms applying the sandwich technique.

Very important is the research on recording materials by Usanov and Shevtsov (1990) in Russia. Their work is based on the formation of a microcavity structure in gelatin, by applying a special processing technique to silver halide emulsions. Such holograms recorded in a silver halide emulsion have a high diffraction efficiency and exhibit a typical DCG hologram quality.

Not until panchromatic ultra-fine-grain silver halide emulsions were introduced in Russia in the early 1990s, was it possible to record high-quality color holograms in a single emulsion layer as demonstrated by the present author. Although it is now possible to produce color holograms, they need to be displayed correctly. An important improvement would be to find an edge-illuminating technique to make color holograms easier to display. The display problem still remains the main obstacle preventing color holograms, as well as monochrome holograms, being more widely used.

Recording of Color Holograms

Currently, most transmission color holograms are of the rainbow type. Large-format holographic stereograms made from color-separated movie or video recordings have been produced. There are also some examples of embossed holograms in which a correct color image is reproduced. In order to generate a true color rainbow hologram, a special setup is required in which the direction of the reference beam can be changed in between the recordings of the color-separated RGB primary images. However, as already mentioned, holographic color images of the rainbow type can reconstruct a correct color image only along a horizontal line in front of the film or plate and, thus, of less interest for serious color holography applications. Reflection holography can offer full parallax, large field of view 3D color images where the colors do not change when observed from different directions.

Silver Halide Materials

To be able to record high-quality color reflection holograms it is necessary to use extremely low light-scattering recording materials. This means, for example, the use of ultra-fine-grain silver halide emulsions (grain size about 10 nm). Currently, the only producer of a commercial holographic panchromatic ultra-fine-grain silver halide material is the Micron branch of the Slavich photographic company located outside Moscow. The PFG-03c emulsion comes in standard sizes from 63 mm × 63 mm format up to 300 mm × 406 mm glass plates and is also available on film. Some characteristics of the Slavich material are presented in Table 1.

By using suitable processing chemistry for the PFG-03c emulsion it has been possible to obtain high-quality color holograms. These holograms are recorded in a single-layer emulsion, which greatly simplifies the recording process as compared with many earlier techniques. The aforesaid single-layer technique is described below.

Table 1 Characteristics of the Slavich panchromatic emulsion

Silver halide material	PFG-03C
Emulsion thickness	7 μm
Grain size	12–20 nm
Resolution	~10,000 lp(mm)$^{-1}$
Blue sensitivity	~1.0–1.5 × 10^{-3} J(cm)$^{-2}$
Green sensitivity	~1.2–1.6 × 10^{-3} J(cm)$^{-2}$
Red sensitivity	~0.8–1.2 × 10^{-3} J(cm)$^{-2}$
Color sensitivity peaked at	633 nm and 530 nm

Photopolymer Materials

The color holography photopolymer material from DuPont is another alternative recording material for color holograms. In particular, this type of material is suitable for mass production of color holograms. Although, being less sensitive than the ultra-fine-grain silver halide emulsion, it has special advantages of easy handling and dry processing (only UV-curing and baking). The DuPont color photopolymer material has a coated film layer thickness of about 20 μm. The photopolymer film is generally coated in a 12.5-inch width on a 14-inch wide Mylar® polyester base which is 0.002-inch thick. The film is protected with a 0.00092-inch thick Mylar® polyester cover sheet. However, the panchromatic polymer material is only supplied to specially selected and approved hologram producers. The reason for this is that some applications of color holograms are in the field of document security where optically variable devices (OVDs) are produced.

The recording of a color hologram on DuPont polymer is simple. The film has to be laminated to a piece of clean glass or attached to a glass plate using an index-matching liquid. To obtain the right color balance, the RGB sensitivity depends on the particular material, but typically red sensitivity is lower than green and blue sensitivities. It is difficult to obtain high red-sensitivity of photopolymer materials. Simultaneous exposure is the best recording technique for photopolymer materials. Holograms can be recorded manually, but in order to produce large quantities of holograms, a special machine is required. For hologram replication the scanning technique can provide the highest production rate. In this case, three scanning laser lines are needed, which can be adjusted in such a way that all three simultaneously can scan the film. The color photopolymer material needs an overall exposure of about 10 mJ(cm)$^{-2}$.

After the exposure is finished, the film has to be exposed to strong white or UV light. DuPont recommends about 100 mJ(cm)$^{-2}$ exposure at 350–380 nm. After that, the hologram is put in an oven at a temperature of 120 °C for two hours in order to increase the brightness of the image.

Laser Wavelengths for Color Holograms

The problem of choosing optimal primary laser wavelengths for color holography is illustrated in the 1976 CIE chromaticity diagram (Figure 1) which indicates suitable laser wavelengths for color holograms.

It may seem that the main aim of choosing the recording wavelengths for color holograms would be to cover as large an area of the chromaticity diagram as possible. However, there are many other

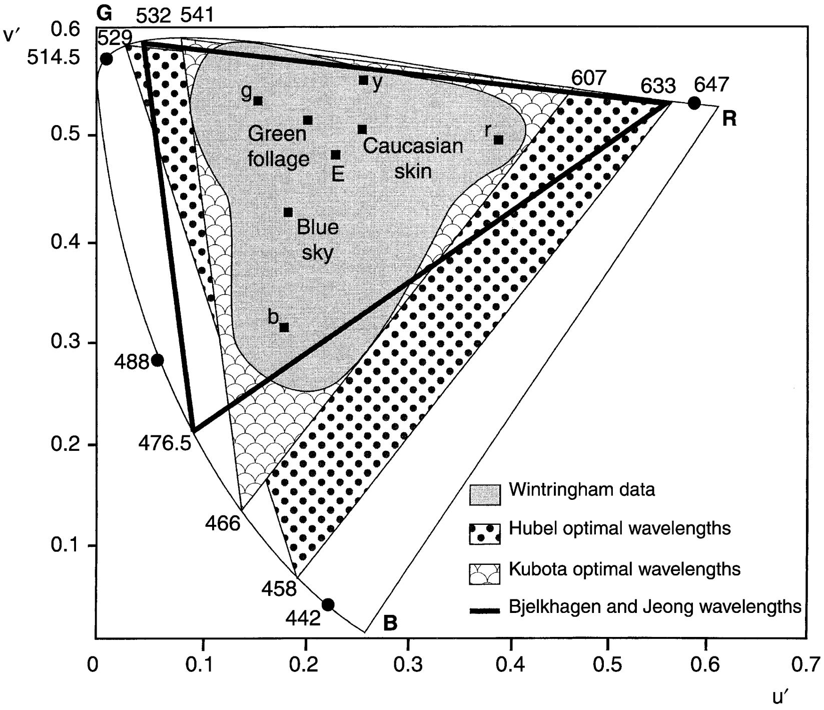

Figure 1 The 1976 CIE uniform scales chromaticity diagram shows the gamut of surface colors and positions of common laser wavelengths. Optimal color-recording laser wavelengths are also indicated.

considerations that must be taken into account when choosing the wavelengths for color holograms. One of these important considerations is the question of whether three wavelengths are really sufficient for color holography. The wavelength selection problem for color holography has been treated in several papers, for example by Peercy and Hesselink (1994).

Hubel and Solymar (1991) provided a definition of color recording in holography:

> A holographic technique is said to reproduce 'true' colors if the average vector length of a standard set of colored surfaces is less than 0.015 chromaticity coordinate units, and the gamut area obtained by these surfaces is within 40% of the reference gamut. Average vector length and gamut area should both be computed using a suitable white light standard reference illuminant.

An important factor to bear in mind when working with silver halide materials is that a slightly longer blue wavelength than the optimal one might give higher-quality holograms (better signal-to-noise ratio) because of reduced Rayleigh scattering during the recording. Another important factor to consider is the reflectivity of the object at primary spectral wavelengths. It has been shown that the reflectivity of an object at three wavelength bands, peaked at 450, 540, and 610 nm, has an important bearing on color reconstruction. These wavelengths can also be considered optimal for the recording of color holograms even though the laser lines are very narrow-band. A triangle larger than necessary can be considered in order to compensate for color desaturation (color shifting towards white) that takes place when reconstructing color reflection holograms in white light. According to Hubel's color rendering analysis, the optimal wavelengths are 464, 527, and 606 nm for the sandwich silver halide recording technique. If the calculations are performed to maximize the gamut area instead, the following set of wavelengths is obtained: 456, 532, and 624 nm. However, when approached from a different viewpoint, the optimal trio of wavelengths based on the reconstructing light source of 3400 K, a 6 μm thick emulsion with a refractive index of 1.63 and an angle of 30° between the object and the reference beam the following wavelengths were obtained: 466.0, 540.9, and 606.6 nm. Peercy and Hesselink (1994) discussed wavelength selection by investigating the sampling nature of the holographic process. During the recording of a color hologram the chosen wavelengths point-sample the surface-reflectance functions of the object. This sampling on color perception can be investigated by the tristimulus value of points in the reconstructed hologram which is mathematically equivalent to integral approximations for the tristimulus integrals. Peercy and Hesselink used both Gaussian quadrature and Riemann summation for the approximation of the tristimulus integrals. In the first case they found the wavelengths to be 437, 547, and 665 nm. In the second case the wavelengths were 475, 550, and 625 nm. According to Peercy and Hesselink, the sampling approach indicates that three monochromatic sources are almost always insufficient to preserve all of the object's spectral information accurately. They claim that four or five laser wavelengths are required.

Only further experiments will show how many wavelengths are necessary and which combination is the best for practical purposes. Another factor that may influence the choice of the recording wavelengths is the availability of cw lasers currently in use in holographic recording, for example, argon ion, krypton ion, diode-pumped solid state (DPSS) frequency-doubled Nd:YAG, helium neon, and helium cadmium lasers (Table 2).

The recent progress in DPSS laser technology has made available both red and blue DPSS lasers. These lasers are air-cooled, small, and each laser requires less than a hundred watts of electric power to operate. Usually, a set of three DPSS lasers will be the best choice of cw lasers for color holography in the future.

Setup for Recording Color Holograms

A typical reflection hologram recording setup is illustrated in Figure 2.

Table 2 Wavelengths from cw lasers

Wavelength [nm]	*Laser type*	*Single line power [mW]*
442	Helium cadmium	<100
457	DPSS blue	<500
458	Argon ion	<500
468	Krypton ion	<250
476	Krypton ion	<500
477	Argon ion	<500
488	Argon ion	<2000
497	Argon ion	<500
502	Argon ion	<400
514	Argon ion	<5000
521	Krypton ion	<100
529	Argon ion	<600
531	Krypton ion	<250
532	DPSS green	<3000
543	Green neon	<10
568	Krypton ion	<100
633	Helium neon	<80
647	Krypton ion	<2000
656	DPSS red	<1000

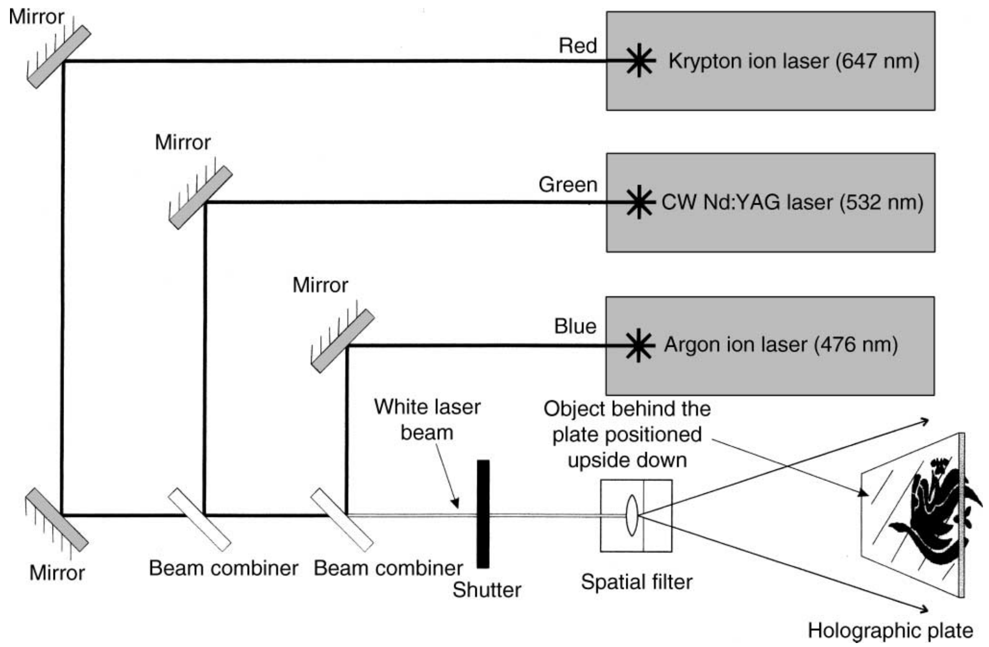

Figure 2 The setup for recording color reflection holograms.

The different laser beams necessary for the exposure of the object pass through the same beam expander and spatial filter. A single-beam Denisyuk arrangement is used, i.e., the object is illuminated through a recording holographic plate. The light reflected from the object constitutes the object beam of the hologram. The reference beam is formed by the three expanded laser beams. This 'white' laser beam illuminates both the holographic plate and the object itself through the plate. Each of the three primary laser wavelengths forms its individual interference pattern in the emulsion, all of which are recorded simultaneously during the exposure. In this way, three holographic images (a red, a green, and a blue image) are superimposed upon one another in the emulsion.

Three laser wavelengths are employed for the recording: 476 nm, provided by an argon ion laser, 532 nm, provided by a cw frequency-doubled Nd:YAG laser, and 647 nm, provided by a krypton laser. Two dichroic filters are used for the combining of the three laser beams. The 'white' laser beam goes through a spatial filter, illuminating the object through the holographic plate.

By using the dichroic filter beam combination technique, it is possible to perform simultaneous exposure recording, which makes it possible to control independently the RGB ratio and the overall exposure energy in the emulsion. The RGB ratio can be varied by individually changing the output power of the lasers, while the overall exposure energy is controlled solely by the exposure time. The overall energy density for exposure is about 3 $\mathrm{mJ(cm)^{-2}}$.

In the initial experiments performed by the author in 1993, a specially designed test object consisting of the 1931 CIE chromaticity diagram, a rainbow ribbon cable, pure yellow dots, and a cloisonné elephant, was used for the color balance adjustments and exposure tests. Later, another test target was employed – the Macbeth ColorChecker® chart, which was used for color rendering tests.

Color reflection holograms can also be produced using copying techniques so that projected real images can be obtained, however, normally associated with a restricted field of view. For many display purposes, the very large field of view obtainable in a Denisyuk color hologram is often more attractive.

Processing of Color Holograms

The dry processing of color holograms recorded on photopolymer materials has already been described. The process is simple and very suitable for machine processing using, for example, a baking scroll oven. The processing of silver halide emulsions is more difficult and critical. The Slavich emulsion is rather soft, and it is important to harden the emulsion before the development and bleaching takes place. Emulsion shrinkage and other emulsion distortions caused by the active solutions used for the processing must be avoided. In particular, when recording master color holograms intended for photopolymer replication, shrinkage control is extremely important. The processing steps are summarized in Table 3.

Table 3 Color holography processing steps

1.	Tanning in a formaldehyde solution	6 min
2.	Short rinse	5 s
3.	Development in the CWC2 developer	3 min
4.	Wash	5 min
5.	Bleaching in the PBU-amidol bleach	~5 min
6.	Wash	10 min
7.	Soaking in acetic acid bath (printout prevention)	1 min
8.	Short rinse	1 min
9.	Washing in distilled water with wetting agent added	1 min
10.	Air drying	

It is very important to employ a suitable bleach bath to convert the developed silver hologram into a phase hologram. The bleach must create an almost stain-free clear emulsion so as not to affect the color image. In addition, no emulsion shrinkage can be permitted, as it would change the colors of the image. Washing and drying must also be done so that no shrinkage occurs. Finally, to prevent any potential emulsion thickness variation by humidity variations, the emulsion needs to be protected by a glass plate sealed onto the hologram plate.

Evaluation of Color Holograms

Recorded color holograms of the two test targets are presented here. The illuminating spotlight to reconstruct the recorded color holograms was a 12-Volt 50-Watt halogen lamp. This type of spotlight is suitable for displaying color holograms. The selection of a suitable lamp for the reconstruction of color holograms is much more important than the selection of lamps for monochrome hologram display. The color balance for the recording of a color hologram must be adjusted to the type of spotlight that is going to be used for the display of the finished hologram. Figure 3 shows a typical normalized spectrum obtained from a white area of the color test target hologram.

This means that the diffraction efficiency of each color component is obtained assuming a flat spectrum of the illuminating source. One should note the high diffraction efficiency in blue, needed to compensate for the low blue light emission of the halogen spotlight. The noise level, mainly in the blue part of the spectrum, is visible and low. The three peaks are exactly at the recording wavelengths; i.e., 647, 532, and 476 nm.

In Table 4 some results of the Macbeth ColorChecker® investigation are presented. The 1931 CIE *x* and *y* coordinates are measured at both the actual target and the holographic image of the target. The measured fields are indicated in the table by color and the corresponding field number.

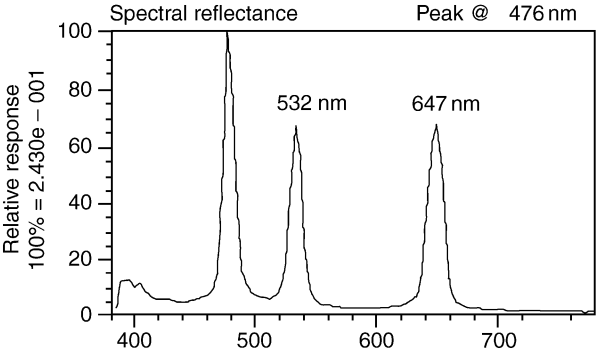

Figure 3 Normalized spectrum from a white area of a color test target hologram.

Color reproductions of two color holograms by the author are presented. A photograph of the 100 mm by 120 mm hologram of the CIE test object is shown in Figure 4. Featured in Figure 5 is a 200 mm by 250 mm hologram of a large Russian egg.

Computer-Generated Color Holograms

Today it is not possible to obtain a computer-generated hologram (CGH) with the same high information content as the laser-recorded ones of real objects. What may become possible in the future is a technique to compute and record the interference pattern that is stored in a Denisyuk color hologram which, upon illumination, can produce an image like the laser-recorded ones of real objects. Therefore, the best compromise is to use holographic stereograms which can provide high-quality computer-generated images. A holographic stereogram is created by using a series of 2D photographic images or 2D images displayed on a LCD screen from which the hologram is recorded. In order to obtain a high-quality holographic image with a large field of view, many 2D images are needed. As mentioned earlier, some color CGHs of the rainbow transmission type have been produced. However, image color changes as a function of viewer position in that type of hologram, making them less attractive than computer-generated reflection holograms. Over the last few years there has been rapid progress in color CGHs of the reflection type. Remarkable results have been achieved by Klug *et al.* (1997) at Zebra Imaging Inc. A new technique, to record very large full-parallax color reflection CGHs, has been developed. Color holograms can be produced, having both vertical and horizontal parallax, with a 100° field of view. The generation of a holographic hardcopy of either digitized images or computer graphics models is based on the following technique. The 'object'

Table 4 Chromaticity coordinates from color hologram recording tests using the Macbeth ColorChecker®

Object	*White #19*	*Blue #13*	*Green #14*	*Red #15*	*Yellow #16*	*Magenta #17*	*Cyan #18*
CIE *x y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*	*x*/*y*
Target	0.435/0.405	0.295/0.260	0.389/0.514	0.615/0.335	0.517/0.450	0.524/0.322	0.285/0.380
Image	0.354/0.419	0.335/0.362	0.337/0.449	0.476/0.357	0.416/0.437	0.448/0.338	0.295/0.366

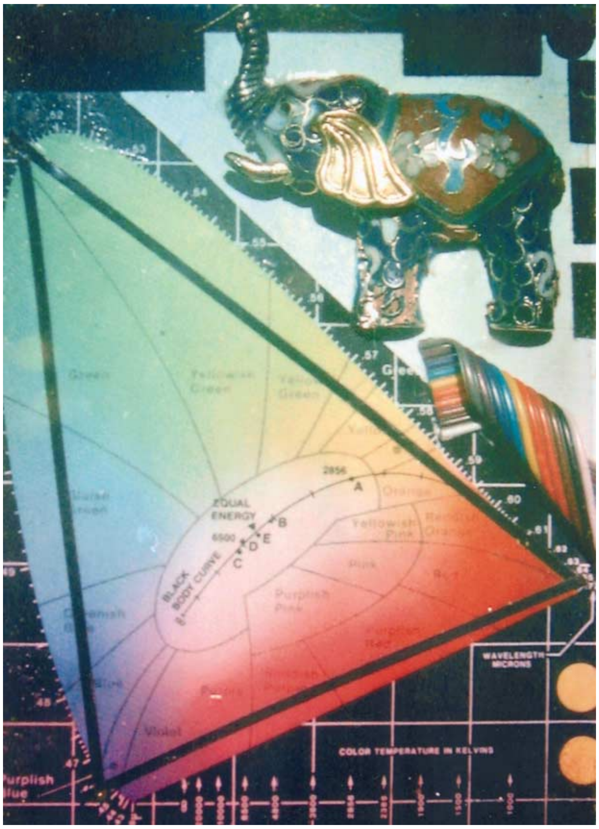

Figure 4 Hologram of the CIE test object recorded with 476, 532, and 647 nm laser wavelengths. Notice the rainbow ribbon cable, the pure yellow dots, the full range of colors, and a balanced white in the center of the CIE color test target.

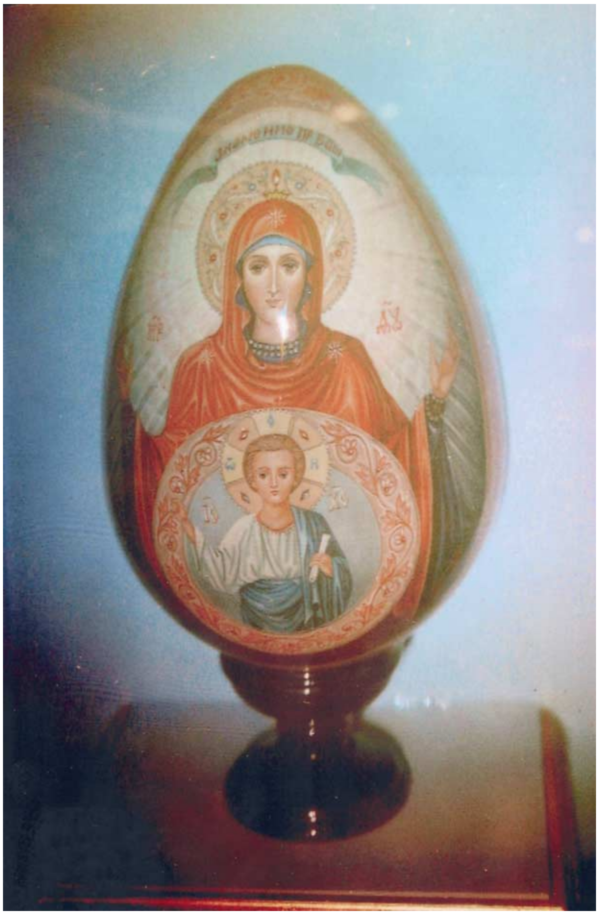

Figure 5 Color hologram of a large Russian egg, size 100 mm by 120 mm.

sub-beam is directed through a sequence of digital images on a liquid-crystal screen. Each resulting exposure, about two millimeters square, is called a 'hogel'. The full-color hogels are the holographic building blocks of a finished CGH image. In an automated step-and-repeat process, 9×10^4 hogels are formed on a flat square tile of DuPont panchromatic photopolymer film by simultaneous RGB laser exposures. The 60 cm by 60 cm tile itself is the finished hologram, or to obtain larger holograms, the 3D image is made in 60 cm by 60 cm tiled increments. So far, the largest image created was of Ford's P2000 Prodigy concept car in which ten such holograms tiled together made up the large color reflection hologram (1.2 m by 3 m) which is reproduced in Figure 6.

This technique has opened the door to real 3D computer graphics. However, generating such a large hologram is a very time-consuming process. The Ford P2000 hologram took almost two weeks (300 hours) to produce, since each individual panel requires a 24-hour recording time.

The Future of Color Holography

The manufacturing of large-format color reflection holograms is now possible. Mass production of color holograms on photopolymer materials has started in Japan. Although good color rendition can be obtained, some problems remain to be solved, such

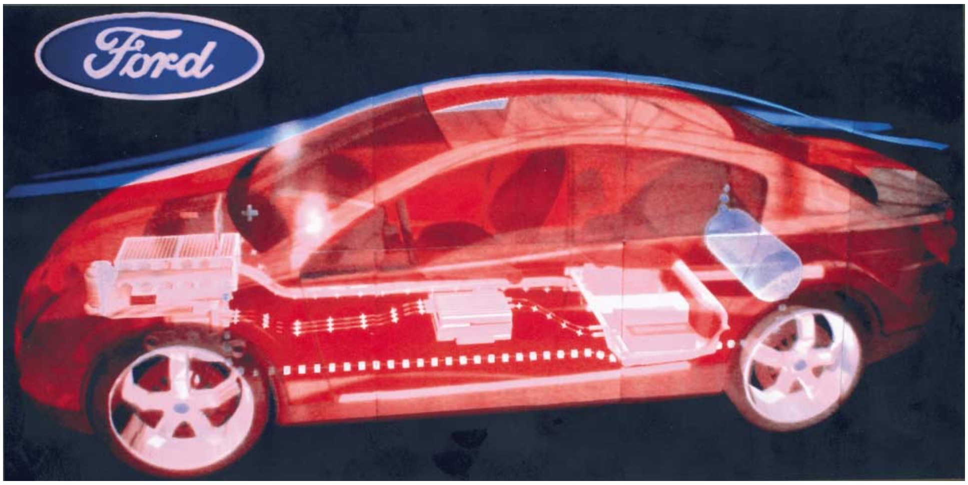

Figure 6 Computer-generated color hologram of Ford's P2000 Prodigy concept car. A ten tile full parallax Mosaic hologram produced by Zebra Imaging, Inc., size 1.2 m by 3 m.

as desaturation, image resolution, signal-to-noise ratio and dynamic range. Other limitations concerning holographic color recording include the fact that some colors we see are the result of fluorescence, which cannot be recorded in a hologram. There are some differences in the recorded colors in one of the test charts shown here. However, color rendition is a very subjective matter. Different ways of rendition may be preferable for different applications, and different people may have different color preferences.

At the present time, research on processing color reflection holograms recorded on ultra-fine-grain materials is still in progress. Work is carried out on microcavity structure holograms in order to produce low-noise reflection holograms with high diffraction efficiency. For producing large quantities of color holograms, the holographic photopolymer materials are most suitable and will be further improved in the future.

Another important field of research is the development of a three-wavelength pulsed laser. Employing such a laser, dynamic events, as well as portraits, can be recorded in a holographic form. Currently, research work is in progress at the French German Research Institute ISL, Saint Louis, France, as well as at the Geola company in Lithuania.

The virtual color image behind a color holographic plate represents the most realistic image of an object that can be obtained today. The extensive field of view adds to the illusion of beholding a real object rather than an image of it. The wavefront reconstruction process recreates accurately the three-wavelength light scattered off the object during the recording of the color hologram. This 3D imaging technique has many obvious applications, in particular, in displaying unique and expensive artifacts. There are also many potential commercial applications of this new feature of holographic imaging, provided that the display problem can be solved using some sort of integrated lighting.

Today, it is technologically possible to record and replay acoustical waves with very high fidelity. Hopefully, holographic techniques will be able to offer the same possibility in the field of optical waves, wavefront storage, and reconstruction. Further development of computer-generated holographic images will make it possible to display extremely realistic full-parallax 3D color images of nonexisting objects, which could be applied in various spheres, for example, in product prototyping, as well as in other applications in 3D visualization and three-dimensional art. Eventually it will become possible to generate true-holographic 3D color images in real-time, as computers become faster and possess greater storage capacity. However, the most important issue is the development of extremely high-resolution electronic display devices which are necessary for electronic holographic real-time imaging.

See also

Holography, Techniques: Computer-Generated Holograms.

Further Reading

Bjelkhagen HI (1993) *Silver Halide Recording Materials for Holography and Their Processing*. Springer Series in Optical Sciences, vol. 66. New York: Springer-Verlag.

Bjelkhagen HI, Jeong TH and Vukičević D (1996) Color reflection holograms recorded in a panchromatic ultrahigh-resolution single-layer silver halide emulsion. *Journal of Imaging Science* & *Technology* 40: 134–146.

Bjelkhagen HI, Huang Q and Jeong TH (1998) Progress in color reflection holography. In: Jeong TH (ed.) *Sixth International Symposium on Display Holography*, Proc. SPIE, vol. 3358, pp. 104–113. Bellingham, WA: The International Society for Optical Engineering.
Hariharan P (1983) Colour holography. In: Wolf E (ed.) *Progress in Optics*, vol. 20, pp. 263–324. Amsterdam: North-Holland.
Hubel PM and Solymar L (1991) Color reflection holography: theory and experiment. *Applied Optics* 30: 4190–4203.
Klug MA, Klein A, Plesniak W, Kropp A and Chen B (1997) Optics for full parallax holographic stereograms. In: Benton SA and Trout TJ (eds) *Practical Holography XI and Holographic Materials III*, Proc. SPIE, vol. 3011, pp. 78–88. Bellingham, WA: The International Society for Optical Engineering.
Kubota T (1986) Recording of high quality color holograms. *Applied Optics* 25: 4141–4145.
Leith EN and Upatnieks J (1964) Wavefront reconstruction with diffused illumination and three-dimensional objects. *Journal of the Optical Society of America* 54: 1295–1301.
Peercy MS and Hesselink L (1994) Wavelength selection for true-color holography. *Applied Optics* 33: 6811–6817.
Usanov YuE and Shevtsov MK (1990) Principles of fabricating micropore silver-halide-gelatin holograms. *Optics and Spectroscopy (USSR)* 69: 112–114.
Watanabe M, Matsuyama T, Kodama D and Hotta T (1999) Mass-produced color graphic arts holograms. In: Benton SA (ed.) *Practical Holography XIII*, Proc. SPIE, vol. 3637, pp. 204–212. Bellingham, WA: The International Society for Optical Engineering.

Computer-Generated Holograms

W J Dallas, University of Arizona, Tucson, AZ, USA

A W Lohmann, University of Erlangen-Nürnberg, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A hologram is a tool for shaping the amplitude and the phase of a light wave. Shaping the amplitude distribution is easy, one way is by a photographic plate, another way is by binary pixels as used in ‘halftone printing’. The phase distribution can be shaped by RE-fraction on a thin film of variable thickness, or by DIF-fraction on a diffraction grating that has been deliberately distorted.

Both kinds of holograms can be manufactured based on a computer design, the refractive version being called a ‘kinoform’, and the diffractive version being known as a computer-generated hologram (CGH). We will explain the fundamentals of computer holography with emphasis on the ‘Fourier–CGH’, and will also highlight the large variety of applications.

From the Classical Hologram to the Computer-Generated Hologram: CGH

Holography is a method to form images, consisting of two steps: recording and reconstruction. In the recording step, some interference fringes are recorded on a photographic plate (Figure 1a). The two interfering waves originate from the object and from a reference light source. After being developed by the usual photo-chemical treatment the photographic plate is called a hologram. In the reconstruction step, the hologram acts as a diffracting object that is illuminated by a replica of the reference light. Due to diffraction, the light behind the hologram is split into three parts (Figure 1b). One part proceeds to the observer who perceives a virtual object where the genuine object had been originally during the recording step.

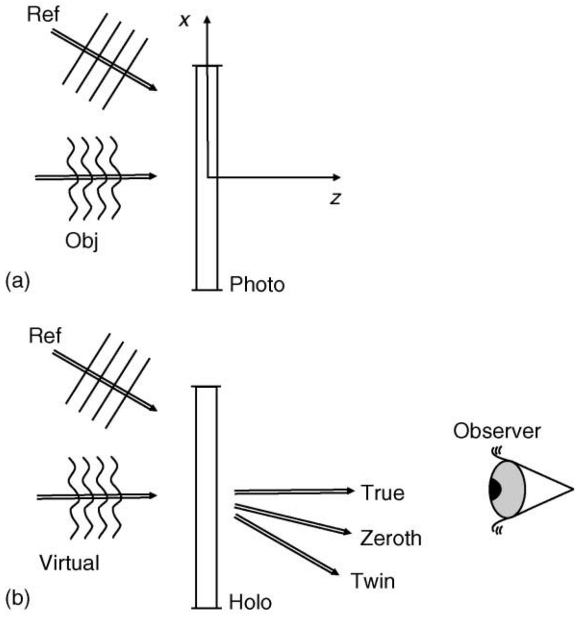

Figure 1 Holographic image formation in two steps: (a) recording the hologram, (b) reconstructing the object wave.

This was a brief description of 'classical-holography'. In 'computer holography' the first step, recording, is synthetic. In other words, the propagation of the light from the object to the photographic plate is simulated digitally. The simulation includes the addition of the object wave to the reference wave and the subsequent modulus square process, which describes the transition from the two complex amplitudes to the hologram irradiance:

$$|u_0(x) + u_R(x)|^2 = I_H(x) \qquad [1]$$

The amplitude transmittance of the hologram is

$$T_H(x) = c_0 + c_1 I_H(x) = c_1 u_0 u_R^* + \cdots \qquad [2]$$

The coefficients c_0 and c_1 represent the photochemical development. In the reconstruction process (Figure 1b) the transmittance T_H is illuminated by the reference beam. Hence, we get

$$u_R(x) T_H(x) = c_1 u_0(x) |u_R(x)|^2 + \cdots \qquad [3]$$

If the reference intensity $|u_R(x)|^2$ is constant, this term represents a reconstruction of the object wave $u_0(x)$. The omitted parts in eqns [2] and [3] describe those beams that are ignored by the observer (Figure 1b). The hologram transmittance $T_H(x)$ is an analog signal:

$$0 \le T_H \le 1 \qquad [4]$$

Before the 1960s, when computer holograms were invented, the available plotters could produce only binary transmittance patterns. Hence, it was necessary to replace the analog transmission $T_H(x)$ by a binary transmission $B_H(x)$, which is possible without loss of image quality. In addition, binary computer holograms have better light efficiency and a better signal-to-noise ratio. They are also more robust when being copied or printed.

Another advantage of a CGH is that the object does not have to exist in reality, which is important for some applications. The genuine object may be difficult to manufacture or, if it is three-dimensional, difficult to illuminate. The CGH may be used to implement an algorithm for optical image processing. In that case, the term 'object' becomes somewhat fictitious. We will discuss more about applications, towards the end of this article.

From a Diffraction Grating to a Fourier CGH

Now we will explain, in some detail, the 'Fourier type' CGH. Fourier holograms are more popular than two other types: 'image-plane' CGHs and 'Fresnel-type' CGHs. The Fresnel-type will be treated in the section on 'software' because of some interest in 3-D holograms. Certain hardware issues will also be described.

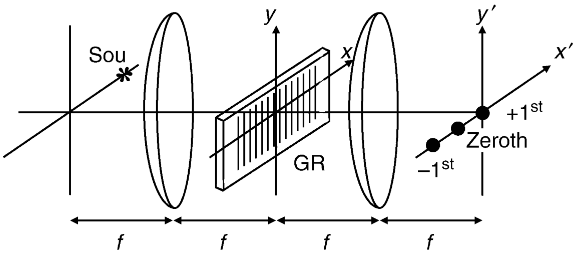

Figure 2 Grating diffraction, but with an off-axis source.

The explanation of Fourier CGHs begins with grating diffraction (Figure 2). The only uncommon feature in Figure 2 is the off-axis location of the source. It is arranged such that the plus-first diffraction order will hit the center of the output plane.

We will now convert a simple Ronchi grating (Figure 3a) into a Fourier CGH. The Ronchi grating transmittance can be expressed as the Fourier series:

$$G(x) = \sum_{(m)} C_m \exp\left(\frac{2\pi i m x}{D}\right); \quad C_0 = \frac{1}{2};$$

$$C_m = \frac{\sin(\pi m/2)}{\pi m} \qquad [5]$$

Now we shift the grating bars by an amount PD. Then we modify the width from $D/2$ to WD, as shown in Figure 3b. The Fourier coefficients are now

$$C_m = \frac{\exp(2\pi i m P)\sin(\pi m W)}{\pi m} \qquad [6]$$

The C_{+1} coefficient is responsible for the complex amplitude at the center of the readout plane:

$$C_{+1} = \frac{\exp(2\pi i P)\sin(\pi W)}{\pi} \qquad [7]$$

We are able now to control the beam magnitude, $A = (\frac{1}{\pi})\sin(\pi W)$, by the relative bar width W and the phase $\phi = 2\pi P$ by the relative bar shift P. This particular kind of phase shift is known as 'detour phase'.

So far, the light that ends up at the center of the exit plane, leaves the grating as a plane wave with a wave vector parallel to the optical axis. Now we distort the grating on purpose (Figure 3c):

$$W \rightarrow W(x, y); \qquad P \rightarrow P(x, y) \qquad [8]$$

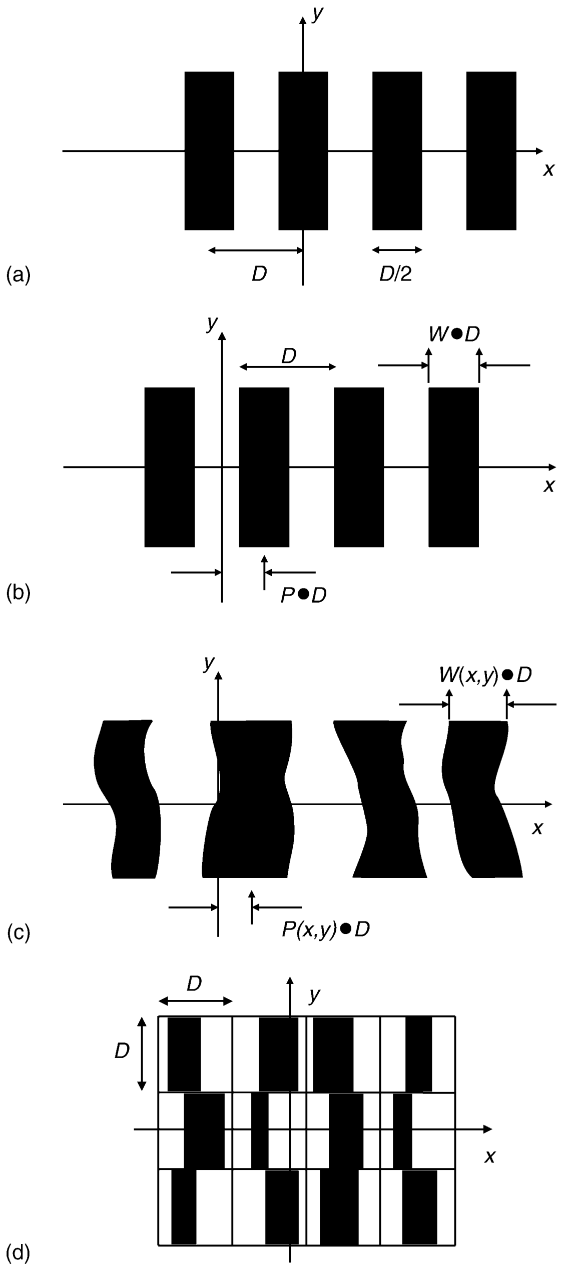

Figure 3 Modifying a grating into a hologram in 3 steps: (a) Ronchi grating, (b) Modified width and location of the grating bars, (c) Distorted grating bars, (d) as in Figure 3c, but now discrete distortions.

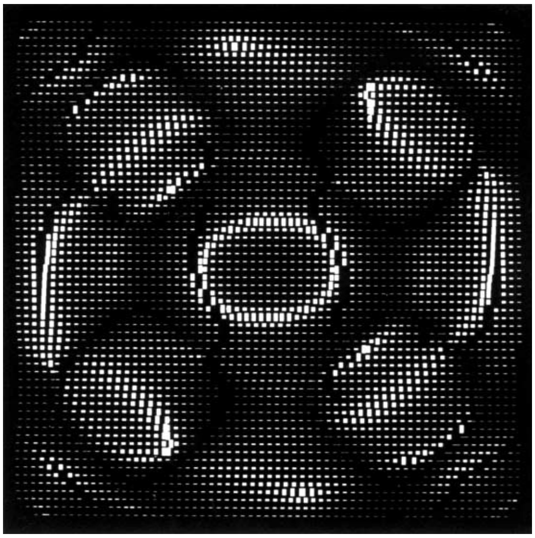

Figure 4 CGH with 64 × 64 cells.

These distortions should be mild enough that the plus-first order light behind the distorted grating can be described by the complex amplitude:

$$\left(\frac{1}{\pi}\right)\sin[\pi W(x,y)]\exp[2\pi i P(x,y)] = u_{\mathrm{H}}(x,y) \qquad [9]$$

This is a generalization of eqn [7]. The phase $2\pi P(x,y)$ covers the range from $-\pi$ to $+\pi/2$ if the shift is bounded by $|P| \leq \frac{1}{2}$. That is enough to cover the range of complex amplitudes within the circle of $|u_{\mathrm{H}}| \leq \frac{1}{\pi}$. We will return shortly to the condition 'mild-enough distortions' in the context of Figure 4.

Most plotters move their pens only in the x-direction and y-direction and printers place dots on a grid. Therefore it is convenient to replace the continuous (x,y) variations of W and P by piecewise constant variations, as in Figure 3d. That restriction turns out to be acceptable, as demonstrated by the first computer holograms. The theoretical proof involves the sampling theorem.

Recall that the light propagation through a 2-f system (Figure 2, between grating and output plane) can be described by a Fourier transformation:

$$\iint u_{\mathrm{H}}(x,y)\exp\left[\frac{-2\pi i(xx'+yy')}{\lambda f}\right]\mathrm{d}x\,\mathrm{d}y = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f},\frac{y'}{\lambda f}\right) \qquad [10]$$

Suppose now that we wish to see a particular image: $v(x',y')$. Hence, we request that

$$v(x',y') = \tilde{u}_{\mathrm{H}}\left(\frac{x'}{\lambda f},\frac{y'}{\lambda f}\right)\mathrm{RECT}\left(\frac{x'}{\lambda f}\right) \qquad [11]$$

The RECT function indicates that only the PLUS first diffraction order is of interest. As a consequence we have to select the grating distortions $W(x,y)$ and $P(x,y)$ such that the hologram $u_{\mathrm{H}}(x,y)$ is the

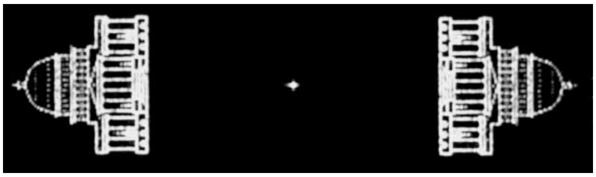

Figure 5 A reconstruction from a 1024 × 1024 CGH. The real image and the symmetrical twin image appear. The zeroth-order is blocked.

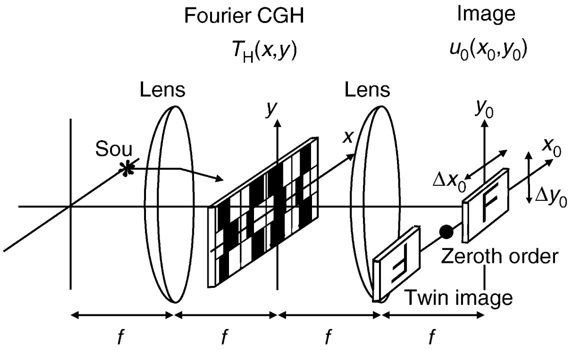

Figure 6 Reconstruction setup for a Fourier CGH.

Fourier transform of $v(x', y')$:

$$u_H(x,y) = \iint v(x',y') \exp\left[\frac{2\pi i(xx'+yy')}{\lambda f}\right] \mathrm{d}x' \, \mathrm{d}y' \qquad [12]$$

The condition 'mild enough distortions' has been treated in the literature. There is not enough space here to present that part of the CGH theory. But we show an actual CGH with 64 × 64 cells (Figure 4). This CGH is no longer a grating, but, it shows enough resemblance to our heuristic derivation. Hence, the CGH theory appears to be trustworthy. A CGH looks less regular if the image $v(x', y')$ contains a random phase that simulates a diffuser. The diffuser spreads light across the CGH and so levels out unacceptably large fluctuations in the amplitude.

With 512 × 512 or 1024 × 1024 cells in a CGH one can obtain an image of the quality seen in Figure 5. The reconstruction setup is shown in Figure 6, here with the letter F as the image.

About Some CGH Algorithms

The computational effort for getting the desired hologram amplitude of a Fourier hologram u_H is reasonable due to the powerful FFT algorithm. However, to use it the data must be discrete. In other words, the CGH consists of cells, centered at $x_m = mD$, $y_n = nD$ as shown in Figure 3d. The questions now are: 'What is the proper cell size D?' and 'How many cells are needed?' The answers to those questions depend on the parameters of the image. If the zero order (see Figure 6) should be at the distance Δx_0, then the quasi-grating period D should obey:

$$\frac{\lambda f}{D} \geq \Delta x_0 \qquad [13]$$

And if the size of an image pixel should be as fine as δx_0, the size Δx_H of the hologram should obey:

$$\frac{\lambda f}{\Delta x_H} \leq \delta x_0 \qquad [14]$$

This condition is plausible because the finite size Δx_H of the hologram acts like a resolution-limiting aperture. The combination of eqns [13] and [14] yields

$$\frac{\Delta x_0}{\delta x_0} \leq \frac{\Delta x_H}{D} \qquad [15]$$

This means that the number of image pixels $\Delta x_0/\delta x_0$ is bounded by the number of CGH cells $\Delta x_H/D$. The generalization to two dimensions is straightforward. Again, Fourier CGHs benefit from the fact that the light propagation from the virtual object to the hologram plane can be described by a Fourier transformation. Hence, the FFT can be used.

And now we move to the synthesis of a Fresnel hologram. A Fresnel hologram is recorded at a finite distance z from the object. Hence, we have to simulate the wave propagation in free space from the object (at $z = 0$) to the hologram at a distance z. This free space propagation can be described (in the paraxial approximation) as a convolution of object and quasi-spherical wave:

$$u_0(x) \rightarrow \int_{-\infty}^{\infty} u_0(x') \exp\left[\frac{-i\pi(x - x')^2}{\lambda z}\right] \mathrm{d}x' = u(x,z) \qquad [16]$$

In the Fourier domain the corresponding operation is:

$$\tilde{u}_0(\mu) \rightarrow \tilde{u}_0(\mu) \exp[-i\pi\lambda z\mu^2] = \tilde{u}(\mu, z) \qquad [17]$$

The indirect execution of eqn [16], based on eqn [17], is easy: a Fourier transformation of the object converts $u_0(x)$ into $\tilde{u}_0(\mu)$, which is then multiplied by the quadratic phase factor, yielding $\tilde{u}(\mu, z)$. An inverse Fourier transformation produces $u(x, z)$. The two Fourier transformations consist typically of $4N \log N$ multiply/add operations. N is the number of pixels: image size Δx_0, divided by the pixel size δx_0.

In the case of a rectangular object $u_0(x, y)$, the number of pixels is $\Delta x_0 \Delta y_0 / \delta x_0 \delta y_0$.

Now to describe the algorithmic effort for computing the convolution, the $u_0(x)$ has N pixels, and, the lateral size of the exponential is essentially $M = z/\delta z$, where δz represents 'depth of focus'. M describes the lateral size of the diffraction cone, measured in pixel size units δx_0, and the convolution involves MN multiply/adds. A comparison of these two algorithms is dictated by the ratio

$$\frac{M}{4 \log N} \qquad [18]$$

The direct convolution is preferable if the distance z is fairly small. The integration limits of eqn [16] depend on the oscillating phase $\pi x^2/\lambda z$. Effectively, nothing is contributed to this integral, when this phase varies by more than 2π over the length, δx_0, of an object pixel. Note that the pixel size of $u(x, z)$ does not change with the distance z, because the bandwidth $\Delta \mu(z) = \Delta \mu_0$ is z-independent. This is so because the power spectrum is z-invariant:

$$|\tilde{u}(\mu, z)|^2 = |\tilde{u}_0(\mu)|^2 \qquad [19]$$

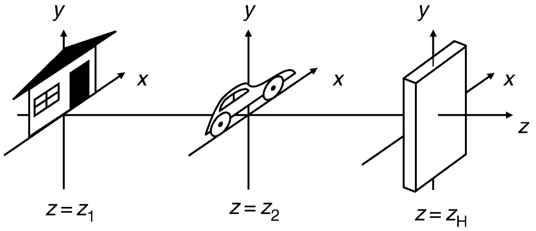

Figure 7 Synthesis of a hologram for a three-dimensional object.

It is easy to proceed from here to the CGH synthesis of a 3-D object. One starts with an object detail $u_1(x)$ at z_1, for example, the house as in Figure 7. Then, one propagates to the next object detail at z_2. This second detail may act as a multiplication (transmission) and as addition (source elements). This process is repeated as the wave moves towards the plane of the hologram. There, a reference wave has to be added, as in the Fourier case.

A very powerful algorithm for the CGH synthesis is 'IFTA' (iterative Fourier transform algorithm). Suppose, we want to design a Fourier CGH, which looks like a person and whose reconstructed image shows the signature of that person (Figure 8a,b). In other words, the two intensities

$$|u_H(x, y)|^2 \text{ and } |\tilde{u}_H(\mu, \nu)|^2 \qquad [20]$$

are determined. But, the phase distributions of u_H and $\tilde{u}_H$ are still at our disposal. The IFTA algorithm will yield those phases in many cases at the expense of ten, hundred or more Fourier transformations.

The IFTA algorithm is sometimes called 'Fourier Ping-Pong' because it bounces back and forth between the (x, y) and the (μ, ν) domain. The amplitudes $|u_H|$ and $|u_0|$ are enforced at every opportunity. But the associated phases are free to vary and to converge, eventually. If so, u_H and u_0 are completely known. The IFTA will not always converge, for example not if

$$|u_H(x, y)| = \delta(x, y) \text{ and } |\tilde{u}_H(\mu, \nu)| = \delta(\mu, \nu) \qquad [21]$$

But the chances are fairly good if both amplitudes are often close to one and seldom near zero.

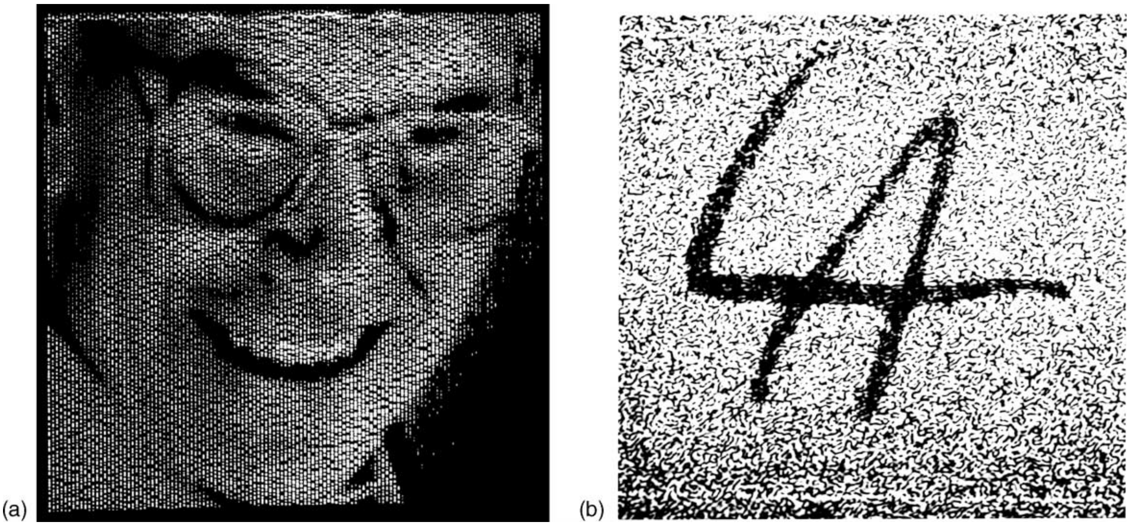

Figure 8 IFTA-CGH: (a) The Fourier CGH. (b) The optical reconstruction thereof. (Courtesy of H.O. Bartelt.)

Finally, an image plane CGH is trivial from an algorithmic point of view. The simulated propagation from object plane to image plane is usually an identity, apart from the finite bandwidth of the image-forming system. Nevertheless, image plane holograms can be quite useful, for example for interferometric testing of a nonspherical mirror of an astronomical telescope. The CGH is laid out to provide a wavefront that is suitable for an interferometric ‘null test’. The CGH serves as a ‘synthetic prototype’.

A fourth kind of CGH, to be named near-field CGH may emerge. It operates in the near-field and it employs evanescent waves as they occur, if object details are comparable in size with, or even smaller than, the wavelength. Conceivable topics are ‘super resolution’ and ‘sub-lambda lithography’. Fundamentals and history of classical evanescent holography are reported in the Further Reading.

Some Hardware Aspects

We now describe some hardware aspects of the CGHs. Remembering the CGH cell structure in Figure 3d, we saw the amplitude encoded as relative width W (eqn [6]). Instead of the width, one may use the relative height $H \leq 1$ as the amplitude parameter. More efficient, but more difficult to manufacture are saw tooth-shaped hologram cells. The saw tooth prism may be approximated by a phase stair, which is possible with today’s lithographic technology. The corresponding theory is called phase quantization.

Some CGH Applications

Some applications had been mentioned already before, e.g., 3-D display (see Figure 7). The data volume can be very high, hence, shortcuts such as eliminating the vertical perspective, are necessary. The largest CGHs for visual light are used for testing astronomical telescopes by means of interferometry. The IFTA algorithm has been used, among other things, for designing a CGH which acts as an optimal diffuser. Speckles and light efficiency are the critical issues in diffuser design. Another project, that also benefits from IFTA, is beam shaping. That is a broad topic with aims at, for example, homogenizing the output of laser, beam structuring for welding, cutting, and scanning.

A CGH can also be used as a spatial filter for image processing. For example, an input image may be Hilbert-transformed in the setup shown in Figure 9a. The filter (Figure 9b) is a grating, of which one sideband is shifted by one-half of the grating period. Such a geometric shift generates the π-phase shift that is needed for the Hilbert transformation. The same setup can also be used to implement Zernike’s phase contrast. Now the grating is distorted such that the zero-frequency region at the optical axis is reduced in amplitude by (typically) a factor of 0.2. In addition, the phase is shifted by $\pi/2$ due to a fringe shift of a quarter period (Figure 10a). In the image plane one can see an ordinary image on-axis, and two phase-contrast images in the plus-minus first diffraction orders. One of them has positive phase contrast, the

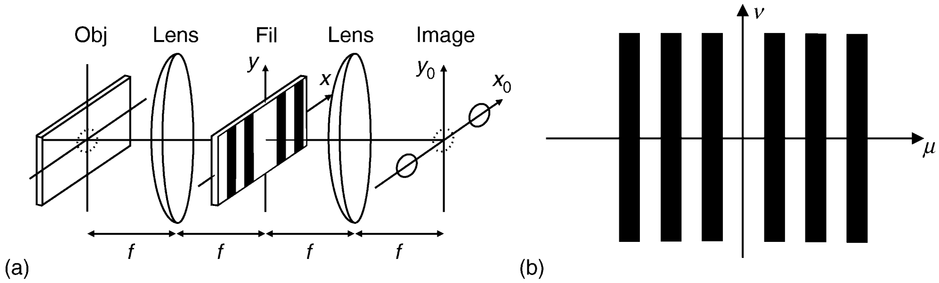

Figure 9 (a) Hilbert setup. (b) Hilbert filter.

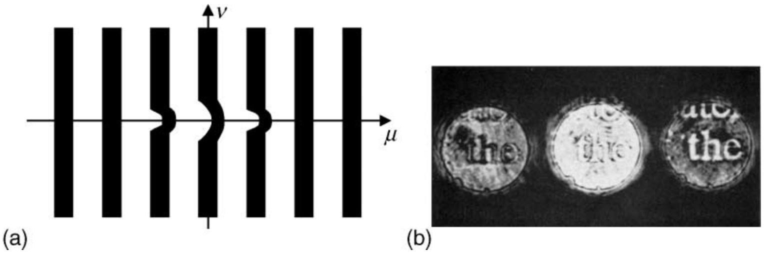

Figure 10 (a) Phase contrast filter. (b) Phase contrast output.

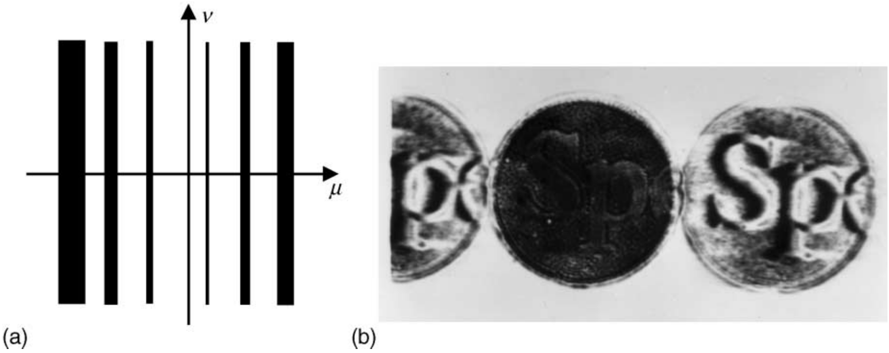

Figure 11 (a) Differentiation filter. (b) Differentiation output.

other one has negative phase contrast (Figure 10b). Another simple spatial filter (Figure 11a) produces a derivative of the input $u(x,y) : \partial u(x,y)/\partial x$ (Figure 11b). Equations [22]–[24] explain this differential filtering experiment.

$$u(x,y) \rightarrow \frac{\partial u(x,y)}{\partial x} = v(x,y) \qquad [22]$$

$$\tilde{u}(\mu,\nu) \rightarrow 2\pi i\mu\cdot\tilde{u}(\mu,\nu) = \tilde{v}(\mu,\nu) \qquad [23]$$

$$\tilde{p}(\mu,\nu) = 2\pi|\mu|\cdot\begin{Bmatrix} +i(\mu > 0) \\ -i(\mu > 0) \end{Bmatrix} \qquad [24]$$

The phase portion of the filter (Figure 11a) is the same as for the Hilbert filter, but, the filter amplitude now enhances the higher frequencies. The object used for getting the results of Figure 10b and 11b was a phase-only object: a bleached photograph.

A very famous spatial filtering experiment is 'matched filtering', which can be used for pattern recognition. The original matched filters were classical Fourier holograms of the target pattern. Computer hologram's can do the same job, even with spatially incoherent objects, this work being initiated by Katyl.

Fourier holograms are also attractive for data storage, due to their robustness against local defects. However, a local error in the Fourier domain spreads out all over the image domain. The computed hologram has an advantage over classical holograms because of the freedom to incorporate any error detection and correction codes for reducing the bit error rate even further.

One of the most recent advances is in matched filtering with totally incoherent light. This step allows moving spatial filtering out of the well-guarded laboratory into hostile outdoor environments.

Finally, a word about terminology; several terms are used in place of 'CGH': synthetic hologram, digital hologram, holographic optical element (HOE), diffractive optical element (DOE) or digital optical element.

CGHs have also been made for other waves, such as acoustical waves, ocean waves, and microwaves. Digital antenna steering represents a case of computer holography.

See also

Diffraction: Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms; Holographic Interferometry; Overview. **Imaging:** Information Theory in Imaging; Inverse Problems and Computational Imaging; Three-Dimensional Field Transformations. **Information Processing:** Coherent Analogue Optical Processors; Free-Space Optical Computing; Incoherent Analog Optical Processors. **Optical Communication Systems:** Free-Space Optical Communications. **Optical Processing Systems**. **Phase Control:** Wavefront Coding.

Further Reading

Andrés P (1999) *Opt. Lett.* 245: 1331.
Brown BR and Lohmann AW (1966) *Appl. Opt.* 5: 967.
Brown BR and Lohmann AW (1969) *IBM J. R&D* 13: 160.
Bryngdahl O (1973) *Progr. Opt.* 11: 168.
Bryngdahl O and Wyrowski F (1990) *Progr. Opt.* 28: 1–86.
Dallas WJ (1971) *Appl. Opt.* 10: 673.
Dallas WJ (1973) *Appl. Opt.* 12: 1179.
Dallas WJ (1980) *Topics in Appl. Phys.* 40: 291–366.
Dallas WJ (1999–2003) http://www.radiology.arizona.edu/dallas/CGH.htm. University of Arizona, Tucson, AZ.
Dresel T, Beyerlein M and Schwider J (1996) *Appl. Opt.* 35: 4615.
Fienup JR (1981) *Proc. SPIE* 373: 147.
Fienup JR (1982) *Appl. Opt.* 21: 2758.
Gerchberg RW and Saxton WO (1972) *Optik* 35: 237.
Goodman JW and Silvestri AM (1970) *IBM J. R&D* 14: 478.

Hansler RL (1968) *Appl. Opt.* 7: 1863.
Hauk D and Lohmann AW (1958) *Optik* 15: 275.
Ichioka Y and Lohmann AW (1972) *Appl. Opt.* 11: 2597.
Katyl RH (1972) *Appl. Opt.* 11: 1255.
Kozma A and Kelly DL (1965) *Appl. Opt.* 4: 387.
Lee WH (1978) *Progr. Opt.* 16: 219–232.
Lesem LB, Hirsch PM and Jordan JA (1969) *IBM J. R&D* 13: 150.
Lohmann AW and Paris DP (1967) *Appl. Opt.* 6: 1746.
Lohmann AW and Paris DP (1968) *Appl. Opt.* 7: 651.
Lohmann AW and Sinzinger S (1995) *Appl. Opt.* 34: 3172.
Lohmann AW and Werlich HW (1971) *Appl. Opt.* 10: 670.
McGovern AJ and Wyant JC (1971) *Appl. Opt.* 10: 619.
Pe'er A, Wang D, Lohmann AW and Friesem AA (1999) *Opt. Lett.* 24: 1469; (2000) 25: 776.
Sinzinger S and Jahns J (1999) *Micro Optics*. New York: Wiley.
Vander Lugt AB (1964) *IEEE, IT* 10: 139.
Wyrowski F and Bryngdahl O (1991) *Rep. Prog. Phys.* 1481–1571.
Yaroslavskii LP and Merzlyakov NS (1980) *Methods in Digital Holography*. New York: Plenum Press.

Digital Holography

W Osten, University of Stuttgart, Stuggart, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The sinusoidal nature of light, its ability to interfere, and the precise knowledge of its wavelength make it possible to measure several metrological quantities, such as distances and displacements, with high accuracy. However, the high frequency of light results in the difficulty that the primary quantity to be measured, i.e., the phase of the light wave, cannot be observed directly. All quadratic sensors (CCD, CMOS, photo plates) are only able to record intensities. Therefore, one makes use of a trick: the transformation of phase changes into recordable intensity changes as the basic principle of interferometry and holography. Holography provides the possibility to record and to reconstruct the complete information of optical wave fields. Numerous potential applications such as 3D-imaging of natural objects, fabrication of diffractive optical elements, and interferometric investigation of complex technical components become possible. Further progress has been made by recording holograms directly with electronic targets and not in photographic emulsions. In this way the independent numerical reconstruction of phase and intensity can be accomplished by the computer. One important consequence is that interferometric techniques, such as 3D-displacement analysis and 3D-contouring, can be implemented easily and with high flexibility. Furthermore, digitally recorded holograms can be transmitted via telecommunication networks and optically reconstructed at any desired place using an appropriate spatial light modulator. Consequently, the complete optical information of complex objects can be transported between different places with high velocity. This opens the possibility to compare nominal identical but physically different objects (master sample comparison) available at different locations with interferometric sensitivity. We call this remote metrology.

Direct Phase Reconstruction by Digital Holography

In digital speckle pattern interferometry (DSPI), the object is focused onto the target of an electronic sensor. Thus, an image plane hologram is formed as a result of the interference with an inline reference wave. In contrast with DSPI, a digital hologram is recorded without imaging. The target records the superposition of the reference and the object wave in the near-field region – a so-called Fresnel hologram. The basic optical setup in digital holography for recording holograms is the same as in conventional holography (Figure 1a). A laser beam is divided into two coherent beams, one illuminating the object and forming the object wave, the other entering the target directly and forming the reference wave. On this basis, very compact solutions are possible. Figure 1b shows a holographic camera where the interferometer, containing a beamsplitter and some wavefront shaping components, is mounted in front of the camera target.

For a description of the principle of digital Fresnel holography, we use a simplified version of the optical setup (Figure 2a). The object is modeled by a plane rough surface that is located in the (x, y)-plane and illuminated by laser light. The scattered wavefield forms the object wave $u(x, y)$. The target of an electronic sensor (e.g., a CCD or a CMOS) used for recording the hologram is located in the (ξ, η)-plane at the distance d from the object. Following the basic principles of holography, the hologram $h(\xi, \eta)$ originates from the interference of the object wave $u(\xi, \eta)$ and the reference wave $r(\xi, \eta)$

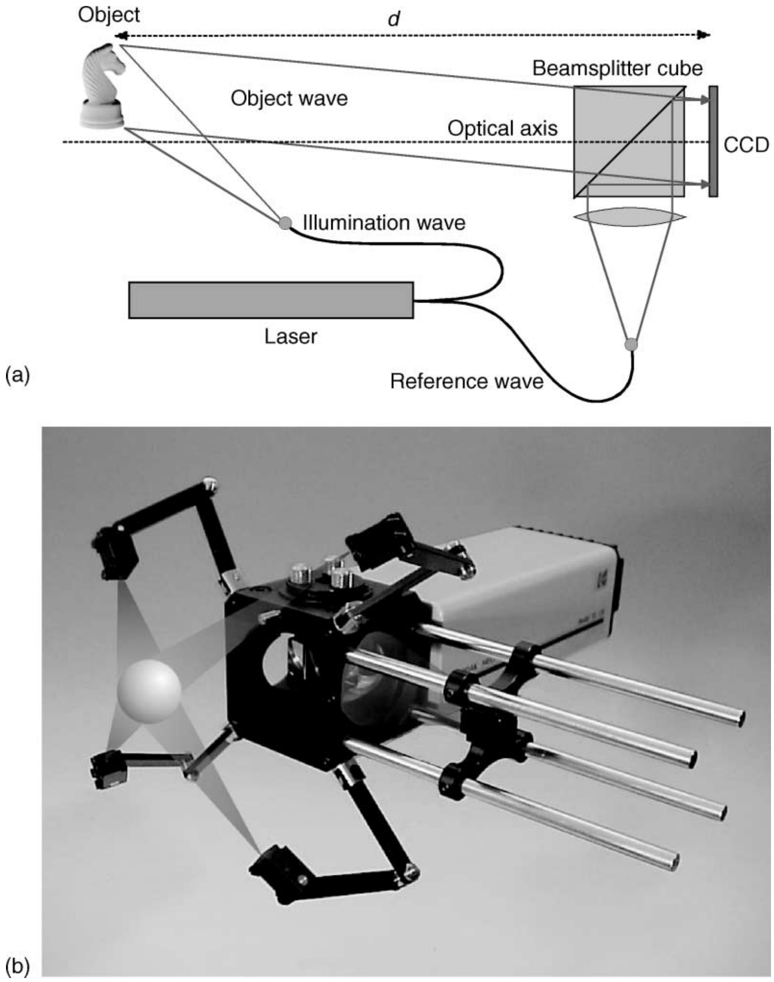

Figure 1 Setup for recording digital holograms. (a) Schematic setup for recording a digital hologram onto a CCD target. (b) Implementation of a holographic camera by mounting a compact holographic interferometer in front of a CCD-target (4 illumination directions are used). Reproduced from Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of microcomponents using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126, copyright © 2001, with permission from Elsevier.

in the (ξ, η)-plane:

$$\begin{aligned} h(\xi, \eta) &= |u(\xi, \eta) + r(\xi, \eta)|^2 \\ &= rr^* + ru^* + ur^* + uu^* \end{aligned} \qquad [1]$$

The transformation of the intensity distribution into a gray value distribution, that is stored in the image memory of the computer, is considered by the characteristic function t of the sensor. This function is generally only approximately linear:

$$T = t[h(\xi, \eta)] \qquad [2]$$

Because the sensor has a limited spatial resolution, the spatial frequencies of the interference fringes in the hologram plane – the so-called micro-interferences – have to be considered. The fringe spacing g and the spatial frequency f_x, respectively, are determined by the angle β between the object and the reference wave (Figure 2b):

$$g = \frac{1}{f_x} = \frac{\lambda}{2\sin(\beta/2)} \qquad [3]$$

with λ as the wavelength. If we assume that the discrete sensor has a pixel pitch (distance between adjacent pixels) $\Delta\xi$, the sampling theorem requires at least two pixels per fringe for a correct reconstruction of the periodic function:

$$2\Delta\xi < \frac{1}{f_x} \qquad [4]$$

Consequently, we obtain for small angles β:

$$\beta < \frac{\lambda}{2\Delta\xi} \qquad [5]$$

Modern high-resolution CCD or CMOS chips have a pitch, $\Delta\xi$, of about 4 μm. In this case, a maximum angle between the reference and the

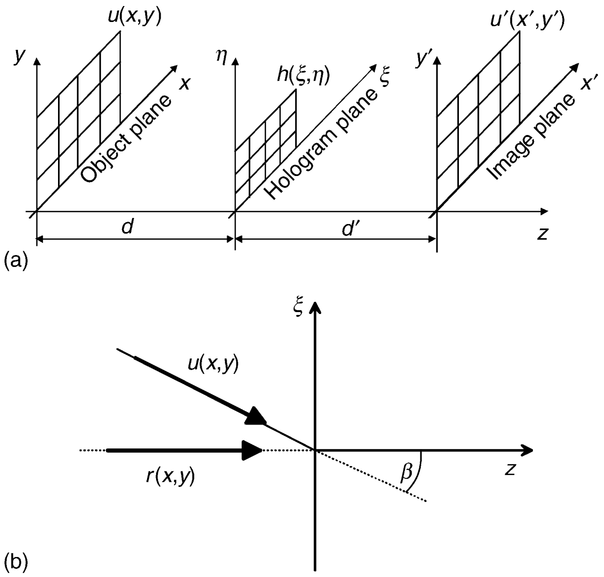

Figure 2 Geometry for recording and reconstruction of digital holograms. (a) Schematic setup for Fresnel holography. (b) Interference between the object and reference wave in the hologram plane.

object wave of only 4° is acceptable. The practical consequence of the restricted angle resolution in digital holography is a limitation of the effective object size that can be stored holographically by an electronic sensor. However, this is only a technical handicap. Larger objects can be placed at a sufficient distance from the hologram, or reduced optically by imaging with a negative lens.

The reconstruction is done by illuminating the hologram with a so-called reconstruction wave $c(\xi, \eta)$:

$$u'(x', y') = t[h(\xi, \eta)]c(\xi, \eta) \qquad [6]$$

The hologram, $t[h(\xi, \eta)]$, diffracts the wave, $c(\xi, \eta)$, in such a way that images of the object wave are reconstructed. Usually four terms are reconstructed if the wave, $u'(x', y')$, propagates in space. An assumed linear characteristic function, $t(h) = \alpha h + t_0$, delivers:

$$u' = Tc = t(h)c \qquad [7]$$

and

$$u' = \alpha\lfloor cu^2 + cr^2 + cur^* + cru^* \rfloor + ct_0 \qquad [8]$$

with two relevant image terms, $[cur^*]$ and $[cru^*]$. The appearance of the image terms depends on the concrete shape of the reconstruction wave. Usually the reference wave, $c = r$, or its conjugated version, $c = r^*$, is applied. In the case of the conjugated reference wave, a direct or real image will be reconstructed due to a converging image wave that can be imaged on a screen at the location of the original object.

However, in digital holography the reconstruction of the object wave in the image plane, $u'(x', y')$, is done by numerical simulation of the physical process, as shown in Figure 3a. The reconstruction wave with a well-known shape equal to the reference wave, $r(\xi, \eta)$, propagates through the hologram, $h(\xi, \eta)$. Following the Huygens principle, each point, $P(\xi, \eta)$, on the hologram, acts as the origin of a spherical elementary wave. The intensity of these elementary waves is modulated by the transparency, $h(\xi, \eta)$. In a given distance, $d' = d$, from the hologram, a sharp real image of the object can be reconstructed as the superposition of all elementary waves. For the reconstruction of a virtual image, $d' = -d$ is used.

Consequently, the calculation of the wavefield, $u'(x', y')$, in the image plane starts with the pointwise multiplication of the stored and transformed intensity values, $t[h(\xi, \eta)]$, with a numerical model of the reference wave, $r(\xi, \eta)$. If the hologram is assumed to be uniformly illuminated by a normally incident, monochromatic plane wave of amplitude 1 the reference wave can be modeled by $r(\xi, \eta) = 1$. After the multiplication the resulting field in the hologram plane is propagated in free space. In the distance, d', the diffracted field, $u'(x', y')$, can be found by solving the well-known Rayleigh–Sommerfeld diffraction formula, that is also known as the

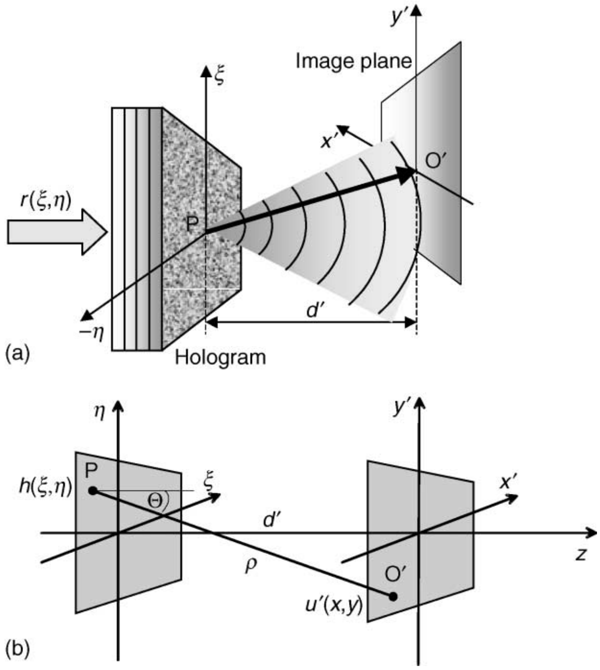

Figure 3 Reconstruction of digital holograms. (a) Principle of wave front reconstruction. (b) Light propagation by diffraction (Huygens–Fresnel principle).

Huygens–Fresnel principle:

$$u'(x',y')=\frac{1}{\mathrm{i}\lambda}\int\!\!\int_{-\infty\cdots\infty}t[h(\xi,\eta)]r(\xi,\eta)\times\frac{\exp(\mathrm{i}k\rho)}{\rho}\cos\theta\,\mathrm{d}\xi\,\mathrm{d}\eta \qquad [9]$$

with

$$\rho(\xi-x',\eta-y')=\sqrt{d'^2+(\xi-x')^2+(\eta-y')^2} \qquad [10]$$

as the distance between a point $\mathrm{O}'(x',y',z=d')$ in the image plane and a point $\mathrm{P}(\xi,\eta,z=0)$ in the hologram plane, and

$$k=\frac{2\pi}{\lambda} \qquad [11]$$

as the wave number. The obliquity factor, $\cos\theta$, represents the cosine of the angle between the outward normal and the vector joining P to O′. This term is given as:

$$\cos\theta=\frac{d'}{\rho} \qquad [12]$$

and therefore the Huygens–Fresnel principle can be rewritten:

$$u'(x',y')=\frac{d'}{\mathrm{i}\lambda}\int\!\!\int_{-\infty\cdots\infty}t[h(\xi,\eta)]r(\xi,\eta)\frac{\exp(\mathrm{i}k\rho)}{\rho^2}\mathrm{d}\xi\mathrm{d}\eta \qquad [13]$$

This numerical reconstruction delivers the complex amplitude of the wavefront. Consequently, the phase distribution, $\phi(x',y')$, and the intensity, $I(x',y')$, can be calculated directly from the reconstructed complex function, $u'(x',y')$:

$$\phi(x',y')=\arctan\frac{\mathrm{Im}|u'(x',y')|}{\mathrm{Re}|u'(x',y')|} \qquad [-\pi,\pi] \qquad [14]$$

$$I(x',y')=u'(x',y')u'^*(x',y') \qquad [15]$$

The direct approach to the phase yields several advantages for imaging and metrology applications that are discussed later.

Reconstruction Principles in Digital Holography

Different numerical reconstruction principles have been investigated: the Fresnel approximation, the convolution approach, the lens-less Fourier approach, the phase shifting approach, and the phase-retrieval approach. In this section the main techniques are briefly described.

The Fresnel Approximation

If the distance, d, between the object and hologram plane, and equivalently, $d'=d$, between the hologram and image plane, is large compared to $(\xi-x')$ and $(\eta-y')$, then the denominator of eqn [13] can be replaced by d'^2 and the parameter ρ in the numerator can be approximated by a binomial expansion for the square root (10) where only the first two terms are considered:

$$\rho\approx d'\left[1+\frac{(\xi-x')^2}{2d'^2}+\frac{(\eta-x')^2}{2d'^2}\right] \qquad [16]$$

The resulting expression for the field at (x',y') becomes

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\int\!\!\int_{-\infty\cdots\infty}t[h(\xi,\eta)]r(\xi,\eta)\times\exp\left[\frac{\mathrm{i}k}{2d'}\left\{(\xi-x')^2+(\eta-y')^2\right\}\mathrm{d}\xi\,\mathrm{d}\eta\right] \qquad [17]$$

Equation [17] is a convolution integral that can be expressed as

$$u'(x',y)=\int\!\!\int_{-\infty\cdots\infty}t[h(\xi,\eta)]r(\xi,\eta)H\times(\xi-x',\eta-y')\mathrm{d}\xi\,\mathrm{d}\eta \qquad [18]$$

with the convolution kernel:

$$H(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\exp\left[\frac{\mathrm{i}k}{2d'}(x'^2+y'^2)\right] \qquad [19]$$

Another notation of eqn [17] is found if the term $\exp[(\mathrm{i}k/2d')(x'^2+y'^2)]$ is taken out of the integral:

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda\mathrm{d}'}\exp\left[\mathrm{i}\frac{k}{2d'}(x'^2+y'^2)\right]\int\!\!\int_{-\infty\cdots\infty}\times\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[\mathrm{i}\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\}\times\exp\left[-\mathrm{i}\frac{2\pi}{\lambda d'}(\xi x'+\eta y')\right]\mathrm{d}\xi\,\mathrm{d}\eta \qquad [20]$$

or

$$u'(x',y')=\frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda d'}\exp\left[\mathrm{i}\frac{\mathrm{k}}{2d'}(x'^2+y'^2)\right]\times\mathrm{FT}_{\lambda d'}\left\{t[h(\xi,\eta)]r(\xi,\eta)\exp\left[\mathrm{i}\frac{k}{2d'}(\xi^2+\eta^2)\right]\right\} \qquad [21]$$

where $\mathrm{FT}_{\lambda d'}$ indicates the 2-dimensional Fourier transform that has been modified by a factor $1/(\lambda d')$. Equation [21] makes clear that the diffracted wavefront consists of the Fourier transform (FT) of the digitally stored hologram, $t[h(\xi,\eta)]$, multiplied by the reference wave, $r(\xi,\eta)$, and the chirp function, $\exp\{(i\pi/\lambda d')(\xi^2+\eta^2)\}$. This Fourier transform is scaled by the constant factor, $1/(id'\lambda)$, and a phase factor that is independent of the processed hologram.

The discrete sampling in digital processing requires the transformation of the infinite continuous integral in eqn [20] into a finite discrete sum. This results in the finite Fresnel transform:

$$u'(m,n) = \sum_{j=0}^{M-1}\sum_{l=0}^{N-1} t[h(j\Delta\xi, l\Delta\eta)]r(j\Delta\xi, l\Delta\eta) \times \exp\left[\frac{i\pi}{d'\lambda}(j^2\Delta\xi^2 + l^2\Delta\eta^2)\right] \times \exp\left\{-i2\pi\left(\frac{jm}{M}+\frac{ln}{N}\right)\right\} \quad [22]$$

where constants and pure phase factors preceding the sums have been omitted. The main parameters are the pixel number $M \times N$ and the pixel pitches $\Delta\xi$ and $\Delta\eta$ in the two directions, which are defined by the used sensor chip.

The discrete phase distribution, $\phi(m,n)$, of the wavefront and the discrete intensity distribution, $I(m,n)$, on the rectangular grid of $M \times N$ sample points can be calculated from the reconstructed complex function, $u'(m,n)$, by using eqns [14] and [15], respectively. The wrapped phase distribution is computed directly without the need of additional temporal or spatial phase modulation such as phase shifting or spatial heterodyning.

For interferometric applications with respect to displacement and shape measurement the phase difference, $\delta(m,n)$, of two reconstructed wave fields, $u_1'(m,n)$ and $u_2'(m,n)$, needs to be calculated. The algorithm can be reduced to the following equation:

$$\delta(m,n) = \phi_1(m,n) - \phi_2(m,n) \quad [23]$$

Figure 4 shows the principle of displacement measurement by digital holography. Besides the investigation of surface displacements in the region of the wavelength, digital holography can be applied advantageously for the measurement of the shape of complex objects. In Figure 5, a turbine blade was investigated with the 2-wavelength contouring method. Both, the digitally reconstructed mod 2π-phase and its demodulated version after-phase unwrapping, are presented. Phase unwrapping is necessary in digital holography, since the fringe-counting problem still remains. However, there are some efficient approaches to overcome the difficulties of this process.

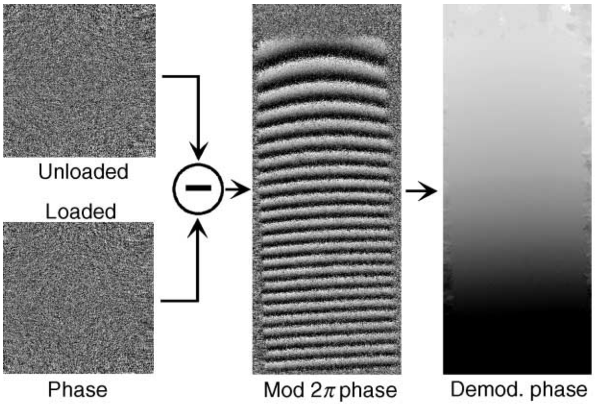

Figure 4 The principle of digital holographic interferometry demonstrated on an example of a loaded small beam. The interference phase is the result of the subtraction of the two phases which correspond to the digital holograms of both object states.

Following eqn [22] the pixel spacing in the reconstructed image is

$$\Delta x' = \frac{d'\lambda}{M\Delta\xi} \quad \text{and} \quad \Delta y' = \frac{d'\lambda}{N\Delta\eta} \quad [24]$$

Additional to the real image, a blurred virtual image and a bright dc-term, the zero-order diffracted field, is reconstructed. This term can effectively be eliminated by preprocessing the stored hologram or the different terms can be separated by using the off-axis instead of the in-line technique. However, the spatial separation between the object and the reference field requires a sufficient space-bandwidth product of the used CCD-chip, as discussed in the section on direct phase reconstruction above.

Numerical Reconstruction by the Convolution Approach

In the section above, we have already mentioned that the connection between the image term, $u'(x',y')$, and the product, $t[h(\xi,\eta)]r(\xi,\eta)$, can be described by a linear space-invariant system. The diffraction formula [17] is a superposition integral with the impulse response:

$$H(x'-\xi, y'-\eta) = \frac{\exp(ikd')}{i\lambda d'} \times \exp\left[\frac{ik}{2d'}\{(\xi-x')^2+(\eta-y')^2\}\right] \quad [25]$$

A linear space-invariant system is characterized by a transfer function G, that can be calculated as the

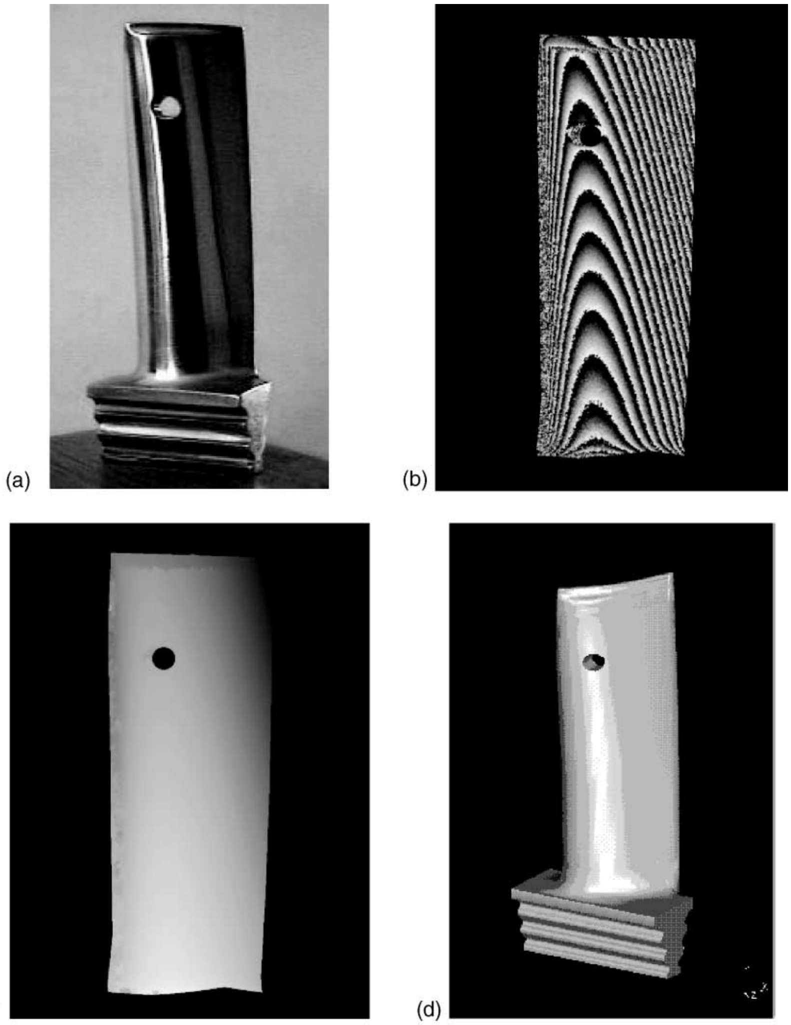

Figure 5 Phase reconstruction by digital holographic interferometry on an example of 2-wavelength contouring of a turbine blade. (a) Image of the turbine blade. (b) Mod-2π phase distribution. (c) Unwrapped phase distribution. (d) CAD reconstruction of the turbine blade.

Fourier-transform of the impulse response:

$$G(f_x, f_y) = \mathrm{FT}\{H(\xi, \eta, x', y')\} \qquad [26]$$

with the spatial frequencies f_x and f_y. Consequently, the convolution theorem can be applied, which states that the Fourier transform of the convolution, $t[h(\xi,\eta)]r(\xi,\eta)$, with H is the product of the individual Fourier transforms, $\mathrm{FT}\{t[h(\xi,\eta)]r(\xi,\eta)\}$ and $\mathrm{FT}\{H\}$. Thus $u'(x', y')$ can be calculated by the inverse Fourier transform of the product of the Fourier-transformed convolution partners:

$$u'(x', y') = \mathrm{FT}^{-1}\{\mathrm{FT}[t(h)r]\mathrm{FT}[H]\} \qquad [27]$$

$$u'(x', y') = [t(h)r] \otimes H \qquad [28]$$

with $\otimes$ as the convolution symbol. The computing effort is comparatively high: two complex multiplications and three Fourier transforms. The main difference to the Fresnel transform is the different pixel size in the reconstructed field:

$$\Delta x' = \Delta\xi \qquad \text{and} \qquad \Delta y' = \Delta\eta \qquad [29]$$

Numerical Reconstruction by the Lensless Fourier Approach

We have already mentioned that the limited spatial resolution of the sensor restricts the angle resolution of the digital hologram. The sampling theorem requires that the angle between the object beam and the reference beam at any point of the electronic sensor be limited in such a way that the microinterference spacing is larger than double the pixel size. In general, the angle between the reference beam and the object beam varies over the sensor's surface, and so does the maximum spatial frequency. Thus, for most holographic setups, the full spatial bandwidth of the sensor cannot be used. However, it is important to use the entire spatial bandwidth of the sensor because the lateral resolution of the reconstructed image depends on a complete evaluation of the entire information obtained from the sensor.

Even if the speed of digital signal processing is increasing rapidly, algorithms should be as simple and as fast to compute as possible. For the Fresnel approach and the convolution approach, several fast Fourier transforms and complex multiplications are necessary. Therefore, a more effective approach, such as the subsequently described algorithm, seems promising.

The lensless Fourier approach is the fastest and most suitable algorithm for small objects. The corresponding setup is shown in Figure 6. It allows us to choose the lateral resolution in a range from a few microns to hundreds of microns without any additional optics. Each point, (ξ, η), on the hologram is again considered as a source point of a spherical elementary wavefront (Huygen's principle). The intensity of these elementary waves is modulated by $t[h(\xi, \eta)]$ – the amplitude transmission of the hologram. The reconstruction algorithm for lensless Fourier holography is based on the Fresnel reconstruction. Here again, $u(x, y)$ is the object wave in the object plane, $h(\xi, \eta)$ the hologram, $r(\xi, \eta)$ the reference wave in the hologram plane, and $u'(x', y')$ the reconstructed wavefield.

For the specific setup of lensless Fourier holography, a spherical reference wave is used with an origin at the same distance from the sensor as the object itself. In the case that $d' = -d$, $x = x'$, and $y = y'$, both virtual images are reconstructed and the reconstruction algorithm then reads:

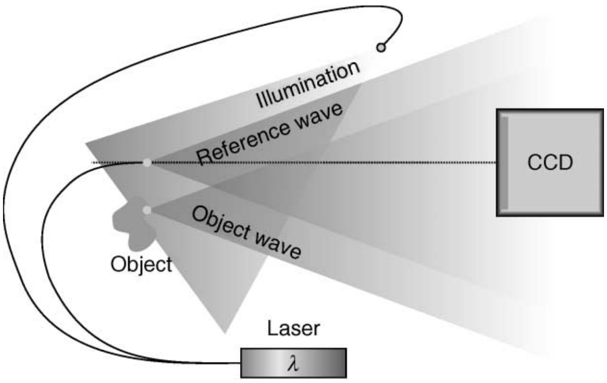

Figure 6 Scheme of a setup for digital lensless Fourier holography of diffusely reflecting objects.

$$u'(x', y') = \frac{\exp(\mathrm{i}kd)}{\mathrm{i}\lambda d} \mathrm{e}^{-\mathrm{i}\frac{\mathrm{k}}{2d}(x^2+y^2)} \times \mathrm{FT}_{\lambda d'}\left\{t[h(\xi, \eta)] \cdot r(\xi, \eta) \cdot \mathrm{e}^{-\mathrm{i}\frac{\mathrm{k}}{2d}(\xi^2+\eta^2)}\right\} \quad [30]$$

with $\mathrm{FT}_{\lambda d}$ as the 2-dimensional Fourier transformation which has been scaled by a factor $1/\lambda d$. In this recording configuration, the effect of the spherical phase factor associated with the Fresnel diffraction pattern of the object is eliminated by the use of a spherical reference wave, $r(\xi, \eta)$, with the same average curvature:

$$r(x, y) = \text{const.} \times \exp\left(\mathrm{i}\frac{\pi}{\lambda d}(\xi^2 + \eta^2)\right) \quad [31]$$

This results in a more simple reconstruction algorithm which can be described by

$$u'(x', y') = \text{const.} \times \exp\left(-\mathrm{i}\frac{\pi}{\lambda d}(x^2 + y^2)\right) \times \mathrm{FT}_{\lambda d}\{h(\xi, \eta)\} \quad [32]$$

Besides the simpler and faster reconstruction algorithm (only one Fourier transformation has to be computed), the Fourier algorithm uses the full space-bandwidth product (SBP) of the sensor chip because it adapts the curvature of the reference wave to the curvature of the object wave.

Influences of Discretization

For the estimation of the lateral resolution in digital holography, three different effects related to discretization have to be considered: averaging, sampling, and the limited sensor size. We assume a quadratic sensor with $N \times M$ pixels of the size $\Delta\xi \times \Delta\xi$. Each pixel has a light-sensitive region with a side length $\gamma\Delta\xi = \Delta\xi_{\mathrm{eff}}$. γ^2 is the so-called fill factor $(0 \leq \gamma \leq 1)$, that indicates the active area of the pixel. The sensor averages the incident light which has to be considered because of possible intensity fluctuations over this area. The continuous expression for the intensity $I(\xi, \eta) = t[h(\xi, \eta)]$, registered by the sensor, has to be integrated over the light-sensitive area. This can be expressed mathematically by the convolution of the incident intensity $I(\xi, \eta)$, with a rectangle function:

$$I^1(\xi, \eta) \propto I \otimes \mathrm{rect}_{\Delta\xi_{\mathrm{eff}} . \Delta\xi_{\mathrm{eff}}} \quad [33]$$

The discrete sampling of the light field is modeled by the multiplication of the continuous assumed hologram with the 2-dimensional comb-function:

$$I^2(\xi, \eta) \propto I^1(\xi, \eta) \times \mathrm{comb}_{\Delta\xi.\Delta\xi} \qquad [34]$$

Finally, the limited sensor size requires that the comb-function has to be truncated at the borders of the sensor. This is considered by the multiplication with a 2-dimensional rect-function of the size $N \times \Delta\xi$:

$$I^3(\xi, \eta) \propto I^2 \times \mathrm{rect}_{N\Delta\xi N.\Delta\xi} \qquad [35]$$

The consequences are amplitude distortions, aliasing and speckling that have to be considered in the reconstruction procedure.

Advantages of Digital Holography

Besides the electronic processing and the direct access to the phase, some further advantages recommend digital holography for several applications such as metrology and microscopy:

- The availability of the independently reconstructed phase distributions of each individual state of the object and interferometer, respectively, offers the possibility to record a series of digital holograms with increased load amplitude. In the evaluation process, the convenient states can be compared interferometrically. Furthermore, a series of digital holograms with increasing load can be applied to unwrap the mod 2π-phase temporally. In this method, the total object deformation is subdivided into many measurement steps, in which the phase differences are smaller than 2π. By adding up those intermediate results, the total phase change can be obtained without any further unwrapping. This is an important feature, since it is essential to have an unwrapped phase to be able to calculate the real deformation data from the phase map. Figure 7 shows an example of such a measurement. The left image shows the wrapped deformation phase for a clamped coin which was loaded with heat. The right image shows the temporal unwrapped phase which has been obtained by dividing the total deformation into 85 sub-measurements. Thus, the displacement of the object can be observed almost in real time during the loading process.
- The independent recording and reconstruction of all states also gives a new degree of freedom for optical shape measurement. In the case of multiwavelength contouring, each hologram can be stored and reconstructed independently with its corresponding wavelength. This results in a drastic decrease of aberrations and makes it possible to use larger wavelength differences for the generation of shorter synthetic wavelengths.
- Because all states of the inspected object can be stored and evaluated independently, only seven digital holograms are necessary to measure the 3D-displacement field and shape of an object under test: one hologram for each illumination direction before and after the loading, respectively, and one hologram with a different wavelength (or a different source-point of illumination) which can interfere with one of the other holograms to perform two-wavelength-contouring (or two-source-point-contouring) for shape measurement. If four illumination directions are used, nine holograms are necessary.
- The direct access to all components of the wavefield makes it possible to correct wavefront aberrations effectively by computer.
- The digital hologram containing all information of the wavefront can be transmitted via the internet to any location and can be reconstructed digitally by computer or as an analogue version by a spatial light modulator. Consequently, remote access to wavefronts, that are generated at distant places, is possible.
- Finally, the possibility to miniaturize complex holographic sensor setups and to use this method for remote comparative interferometry, makes digital holography a versatile tool for the solution of numerous inspection and measurement problems.

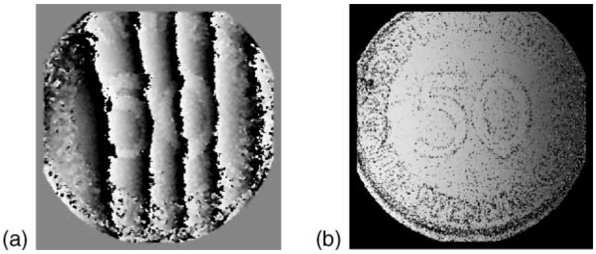

Figure 7 Temporal unwrapping by digital holography: (a) Wrapped phase of a thermal loaded coin. (b) Temporally unwrapped phase.

See also

Diffraction: Fresnel Diffraction. **Fourier Optics**. **Holography, Techniques:** Holographic Interferometry.

Further Reading

Butters JN and Leendertz JA (1971) Holographic and video techniques applied to engineering measurement. *Journal of Measurement Control* 4: 349–354.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Demetrakopoulos T and Mittra T (1974) Digital and optical reconstruction of images from suboptical diffraction patterns. *Applied Optics* 13: 665–670.

Denisyuk YuN (1962) Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Doklady Akademii Nauk SSSR* 144: 1275–1279.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceedings of the Royal Society, A* 197: 454–487.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw Hill Comp. Inc.

Goodman JW and Lawrence RW (1967) Digital image formation from electronically detected holograms. *Applied Physics Letters* 11: 77–79.

Kolenovic E, Osten W, Klattenhoff R, Lai S, Kopylow C and Jüptner W (2003) Miniaturized digital holography sensor for distal three-dimensional endoscopy. *Applied Optics* 42(25): 5167–5172.

Kreis T (1996) *Holographic Interferometry*. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) Principles of digital holography. In: Jüptner W and Osten W (eds) *Proc. Fringe'97, 3rd International Workshop on Automatic Processing of Fringe Patterns*, pp. 353–363. Berlin: Akademie Verlag.

Kreis Th and Jüptner W (1997) The suppression of the dc-term in digital holography. *Optical Engineering* 36: 2357–2360.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*. Bristol and Philadelphia: IOP Publishing Ltd.

Lee W-H (1978) Computer generated holograms: techniques and applications. In: Wolf E (ed.) *Progress in Optics*, vol. XVI, pp. 121–230. Amsterdam: North Holland.

Leith EN and Upatnieks J (1961) New techniques in wavefront reconstruction. *JOSA* 51: 1469–1473.

Osten W (2000) Holography and virtual 3D-testing. Proc. International Berlin Workshop HoloMet 2000, New Prospects of Holography and 3D-Metrology, *Strahltechnik* 14: 14–17.

Osten W (2003) Active metrology by digital holography. *Proceedings of SPIE* 4933: 96–110.

Osten W and Kujawinska M (2000) Active phase measuring metrology. In: Rastogi PK and Inaudi D (eds) *Trends in Optical Nondestructive Testing and Inspection*, pp. 45–69. Amsterdam: Elsevier Science B.V.

Osten W, Jüptner W and Seebacher S (1998) *The Qualification of Large Scale Approved Measurement Techniques for the Testing of Microcomponents*. Proc. 18th Symposium on Seebacher, Experimental Mechanics of Solids, Jachranka, pp. 43–55.

Osten W, Baumbach T, Seebacher S and Jüptner W (2001) Remote shape control by comparative digital holography. *Proc. Fringe 2001*, Bellingham: Elsevier Science, pp. 373–382.

Osten W, Seebacher S and Jüptner W (2001) The application of digital holography for the inspection of microcomponents. *Proceedings of SPIE* 4400: 1–15.

Osten W, Baumbach T and Jüptner W (2002) Comparative digital holography. *Optics Letters* 27(20): 1764–1766.

Pedrini G, Schedin S and Tiziani HJ (2001) Aberration compensation in digital holographic reconstruction of microscopic objects. *Journal of Modern Optics* 48: 1035–1041.

Powell RL and Stetson K (1965) Interferometric vibration analysis of three-dimensional objects by wavefront reconstruction. *JOSA* 55: 1593–1598.

Schnars U (1994) Direct phase determination in hologram interferometry with use of digitally recorded holograms. *Journal of the Optical Society of America A* 11: 2011–2015.

Schnars U and Jüptner W (1994) Direct recording of holograms by a CCD target and numerical reconstruction. *Applied Optics* 33(2): 179–181.

Seebacher S (2001) Application of digital holography for 3d-shape and deformation measurement of micro components. PhD thesis, University of Bremen.

Seebacher S, Osten W and Jüptner W (1997) 3-D deformation analysis of micro-components using digital holography. *Proceedings of SPIE* 3098: 382–391.

Seebacher S, Osten W, Baumbach Th and Jüptner W (2001) The determination of material parameters of micro-components using digital holography. *Optics and Lasers in Engineering* 36(2): 103–126.

Takeda M, Taniguchi K, Hirayama T and Kogho H (1996) Single transform Fourier–Hartley fringe analysis for holographic interferometry. In: Füzessy Z, Jüptner W and Osten W (eds) *Simulation and Experiment in Laser Metrology*, pp. 67–73. Berlin: Akademie Verlag.

Wagner C, Seebacher S, Osten W and Jüptner W (1999) Digital recording and numerical reconstruction of lensless Fourier holograms in optical metrology. *Applied Optics* 38(22): 4812–4820.

Wagner C, Osten W and Seebacher S (2000) Direct shape measurement by digital wavefront reconstruction and multi-wavelength contouring. *Optical Engineering* 39(1): 79–85.

Yamaguchi I and Zhang T (1979) Phase-shifting digital holography. *Optical Letters* 22: 1268–1270.

Yaroslavsky LP and Merzlyakov NS (1980) *Methods of Digital Holography*. New York: Consultants Bureau.

Zhang Y, Pedrini G, Osten W and Tiziani H (2003) Image reconstruction for in-line holography with the Yang-Gu algorithm. *Applied Optics* 42(32): 6452–6457.

Holographic Interferometry

P Rastogi, Swiss Federal Institute of Technology, Lausanne, Switzerland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The development of holographic interferometry has redefined our perception of optical interferometry. The unique property of holographic interferometry to bring a wavefront, generated at some earlier time, stored in a hologram and released at a later time, to interfere with a comparison wavefront, has made possible the interferometric comparison of a rough surface, which is subject to stress, with its normal state. Assuming the surface deformations to be very small, the interference of the two speckled wavefronts forms a set of interference fringes that are indicative of the amount of displacement and deformation undergone by the diffuse object.

Holographic interferometry is an important and exciting area of research and development. It has established itself as a highly promising noninvasive technique in the field of optical metrology. Techniques within its folds for the measurement of static and dynamic displacements, topographic contours, and flow fields have been developed and demonstrated with success in a wide range of problems. Significant and extensive contributions to the development of holographic interferometry have been realized both from the theoretical and experimental perspectives. A wide range of procedures have been developed which make it not only possible to measure surface displacements and deformations of engineering structures to an accuracy of a fraction of a micrometer but also to detect material flaws and inhomogeneities having escaped the manufacturing process. The methodology of quantitative analysis of holographic interferograms has undergone extensive development during the last decade. This article provides a brief review of the holographic techniques with emphasis placed on bringing out its relevance in experimental mechanics and nondestructive testing. The series of books given in the Further Reading section at the end of this article treat the field in detail.

Basic Methods in Holographic Interferometry

The ability of holography to store a wavefront and release it for reconstruction at a later time offers access to the possibility to compare wavefronts, which have albeit existed at different times. There are several schemes that have been established to obtain the interferometric comparison of the wavefronts.

Double-Exposure Holographic Interferometry

In frozen form of wavefront comparison, two holograms of the object are recorded on a photographic plate. The first hologram is made with the object in its initial, unstressed state and the second hologram is recorded with the object in its final, stressed state. Upon reconstruction, the hologram releases the two stored wavefronts, one corresponding to the object in its unstressed state and the other corresponding to the object in its stressed state (Figure 1). The virtual image is overlaid by a set of

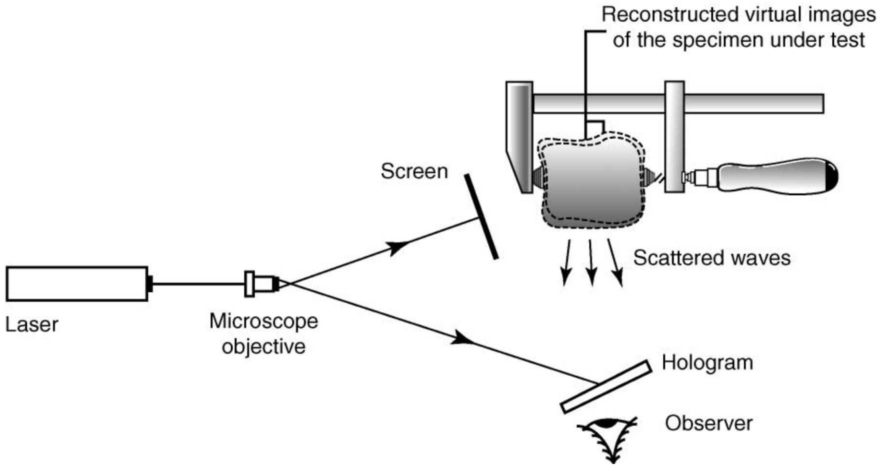

Figure 1 Layout for observing a double exposure holographic interferogram.

bright and dark interference fringes, which are due to the displacement of the object between the two exposures.

The fringe pattern is permanently stored on the holographic plate. The interference fringes denote the loci of points that have undergone an equal change of optical path between the light source and the observer. The image intensity of the holographic interferogram is

$$I(x,y) = I_0(x,y)\{1 + V(x,y)\cos(\varphi(x,y) - \varphi'(x,y))\} \qquad [1]$$

where $I_0(x,y)$ is the background intensity, $V(x,y)$ is the fringe contrast, and φ and φ' are the phases of the waves from the object in its initial and deformed states, respectively.

A bright fringe is produced whenever

$$\varphi(x,y) - \varphi'(x,y) = 2n\pi \qquad n = 0, 1, 2, \ldots \qquad [2]$$

where n is the fringe number. The wave scattered by a rough surface shows rapid variations of phase, which have little correlation across the wavefront. The phase difference varies randomly over the ray directions contained within the aperture of the viewing system, which would normally prevent the observation of fringes with large aperture optics near such points. However, a surface may exist at a distance from the object at which the value of phase difference is stationary over the cone of ray pairs defined by the viewing aperture. Interference fringes are localized in the region where the variation in phase difference is minimum over the range of viewing directions. This approach has been widely used to compute fringe localization for any object displacement and for any arbitrary illumination and observation geometry.

Real-Time Holographic Interferometry

A hologram is made of an arbitrarily shaped rough surface. After development, the hologram is placed back in exactly the same position. Upon reconstruction, the hologram produces the original wavefront. A person looking through the hologram sees a superposition of the original object and its reconstructed image (Figure 2). The object wave interferes with the reconstructed wave to produce a dark field due to destructive interference. If the object is now slightly deformed, interference fringes are produced which are related to the change in shape of the object. A dark fringe is produced, whenever:

$$\varphi(x,y) - \varphi'(x,y) = 2n\pi \qquad n = 0, 1, 2, \ldots \qquad [3]$$

The method is very useful for determining the direction of the object displacement, and for compensating on the interferogram the influence of the rigid body motion of the object when subjected to a stress field.

The optical setup for real-time holography uses an immersion tank, which is mounted on a universal stage and contains the photographic plate. The holographic plate is followed by a vidicon camera, which visualizes the interferograms directly onto the TV monitor screen connected to a video-recording tape to memorize the information. The use of thermoplastic plates as recording material has gained in importance in recent years for implementing real-time and double exposure holographic interferometry setups, as it can be processed rapidly *in situ* using electrical and thermal processing.

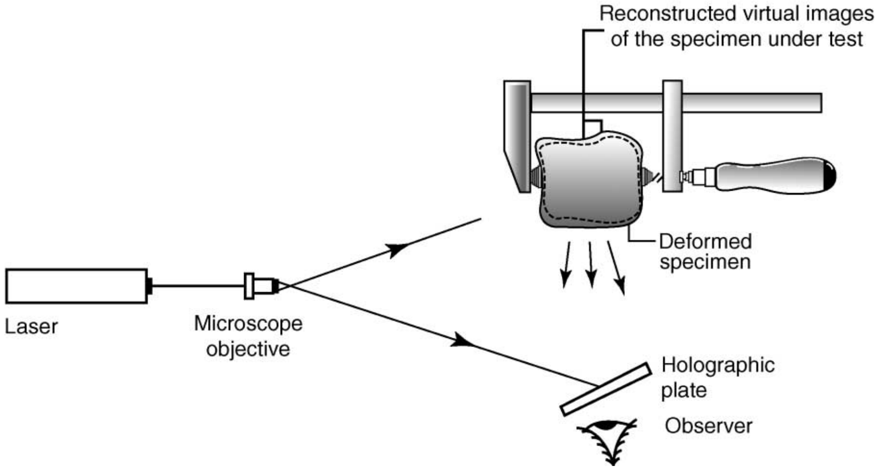

Figure 2 Layout for observing a holographic interferogram in real-time.

The plate is also cost-effective as it is erasable and reusable at least several hundred times. In thermoplastic recording the information is recorded as a thickness variation corresponding to a charge intensity pattern deposited on the thermoplastic layer. Self-developing or photorefractive crystals are other interesting alternatives to the use of film as the recording medium.

Digital Holographic Interferometry

The method consists of recording two holograms corresponding to undeformed and deformed states of an object using a CCD camera. These images are stored in a digital image processing system. The two holograms are then reconstructed separately by numerical methods from the digitaly stored holograms. The advantage of digital holography is that it allows for computing both the intensity and the phase of a holographically stored wavefront during the numerical reconstruction. The contours of constant optical phase change, due to deformation, are obtained by subtracting the reconstructed phases of the undeformed from the deformed object wavefield.

A hologram can be recorded on a CCD as long as the sampling theorem is fulfilled; that is to say, each period of the spatial variation of the hologram intensity is sampled by at least two pixels of the CCD array. The condition imposes a limit on the angle that the object beam can make with the reference beam, which must necessarily be small given that the resolution of CCDs is relatively low. The applications of the method are thus limited mainly by the pixel size in CCD array. The maximum spatial frequency $f_{\max}$ that must be resolved by the recording medium is determined by the maximum angle $\Delta\theta_{\max}$ between the reference and the object wave:

$$f_{\max} = \frac{2}{\lambda}\sin\frac{\Delta\theta_{\max}}{2} \qquad [4]$$

where λ is the wavelength of the laser source. Since the CCD cameras have resolutions of around 100 lines/mm, the maximum angle between the object and reference waves is limited to a few degrees.

Calculation of Phase Change on Object Loading

A diffusely reflecting object is deformed such that a point P on the surface moves to P′ as shown in Figure 3. In double-exposure or real-time holographic interferometry the two wavefronts related to the states of a surface before and after deformation are reconstructed simultaneously and compared. The change of optical path length due to displacement d is

$$\Delta\varphi = \frac{2\pi d}{\lambda}(\cos\psi_1 + \cos\psi_2) \qquad [5]$$

where ψ_1 and ψ_2 are the angles which the illumination and observation waves make with respect to the direction of displacement; eqn [5] can be written in the form:

$$\Delta\varphi = \frac{2\pi d}{\lambda}\cos\eta\cos\psi \qquad [6]$$

where ψ is the angle of bisection between the incident and viewing directions, and η is the angle which the bisector makes with the direction of displacement. The term $d\cos\eta$ being the resolved part of the displacement PP′ in the direction of the bisector, implies that the fringe pattern provides the measure of the displacement component along the bisector of the angle between the incident and viewing directions.

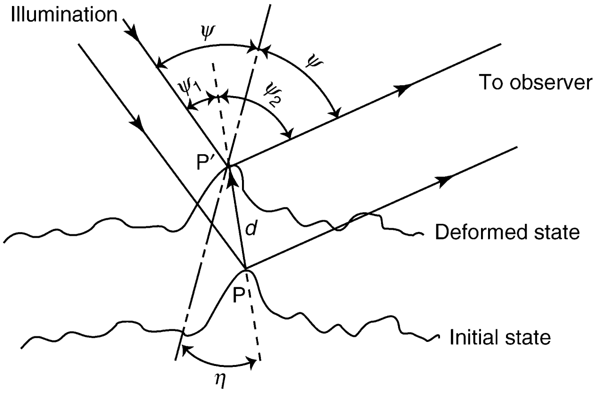

Figure 3 Calculation of the change in optical phase due to surface displacement.

Digital Phase Measurement

The last decade has seen a rapid development of techniques for the automatic and precise reconstruction of phases from fringe patterns. These are based on the concepts of fringe tracking, Fourier transform, carrier frequency, and phase shift interferometry. Quantitative measurement of wavefront phase, for example, can be obtained in a few seconds using phase shift interferometry. With this technique one records a series of holograms by introducing artificially known steps of phase differences in the interference image in eqn [1]:

$$I_i(x,y) = I_0(x,y)\{1 + V(x,y)\cos(\Delta\varphi(x,y) + \Delta\psi_i)\}; \quad i = 0,1,2,\ldots \qquad [7]$$

where $\Delta\psi_i$ is the phase shift which is produced, for example, by shifting the phase of the reference wave. A minimum of three intensity patterns is required to calculate the phase at each point on the object surface. Known phase shifts are commonly introduced using a piezoelectric transducer, which serves to shorten or elongate the reference beam path by a fraction of wavelength; the phase-shifter is placed in one arm of the interferometer. Numerous phase shifting techniques have been developed and which can be incorporated to holographic interferometry to generate a phase map, $\Delta\varphi(x, y)$, corresponding to the object displacement. The calculated phase $\Delta\varphi(x, y)$ is independent of the terms $I_0(x, y)$ and $V(x, y)$, and considerably reduces the dependence of the accuracy of measurements on the fringe quality.

The introduction of phase shift interferometry not only provides accurate phase measurements but also eliminates the problem of phase sign ambiguity of the interference fringes. Dynamic phase shifting techniques allow for producing time sequences of deformation maps for studying objects subjected to time-varying loads. The concept is based on considering the normal phase variations caused by a dynamic phenomenon as equivalent to the phase shifting that is introduced in a holographic interferometer in order to perform the phase evaluation task. This method extends the possibility of applying the method to continuous deformation measurements.

Basic Interferometers for Static Displacement Measurement

Equation [6] shows that observation along a single direction yields information only about the resolved part of the surface displacement in one particular direction. The approach to measuring three-dimensional vector displacement would be to record holograms from three different directions in order to obtain three independent resolved components of displacement. The computation of the resulting system of equations would then allow for determining the complete displacement vector.

Out-of-Plane Displacement Component Measurement

The out-of-plane component of displacement is usually measured by illuminating and viewing the object surface in a near normal direction. The fringe equation corresponding to the line of sight displacement component is given by

$$w = \frac{n\lambda}{2} \qquad [8]$$

where w is the out-of-plane component of displacement. An example of a fringe pattern depicting the out-of-plane displacements in a rectangular aluminum plate, clamped along the edges and drawn out at the center by means of a bolt, subjected to three-point loading, is shown in Figure 4.

In-Plane Displacement Component Measurement

The in-plane component of displacement can be measured by illuminating the object along two directions symmetric to the surface normal and observing along a common direction. This is schematized in Figure 5. The in-plane displacements, in a direction containing the two beams, are mapped as a moiré between the interferograms due to each illumination beam. The moiré appears as a family of fringes:

$$u = \frac{n\lambda}{2\sin\theta_e} \qquad [9]$$

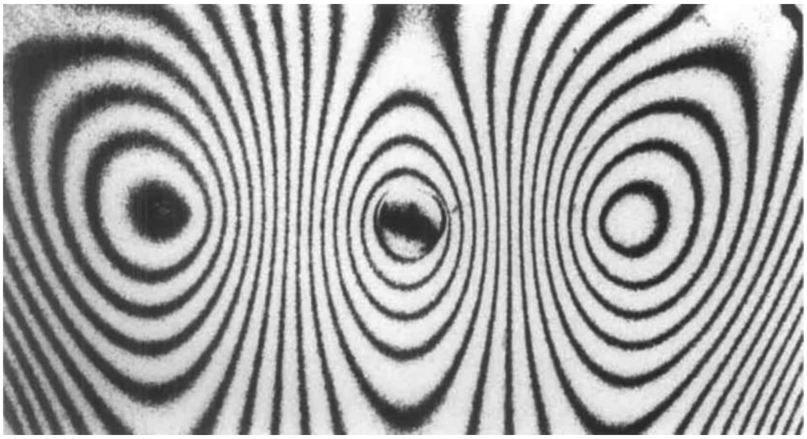

Figure 4 Example of interferogram corresponding to the whole field distribution of out-of-plane component of displacement, *w*, on an aluminum plate subjected to three-point bending.

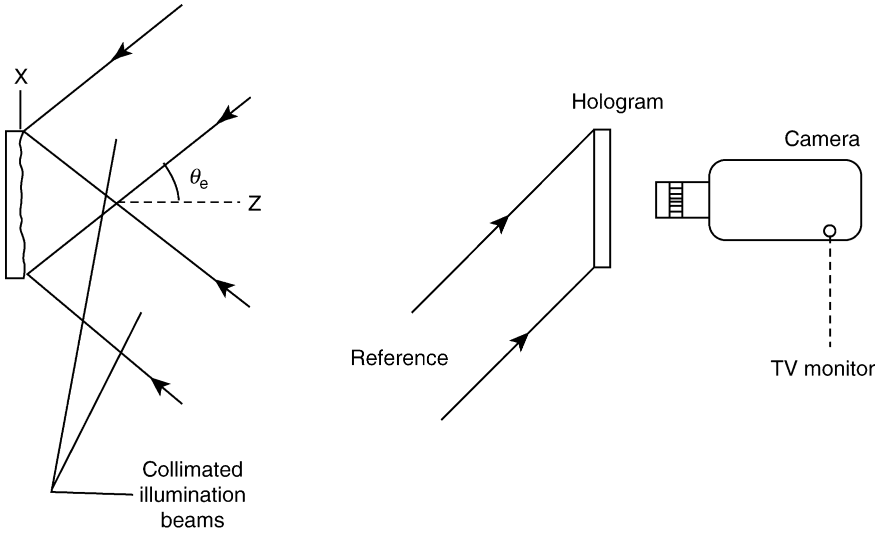

Figure 5 Schematic of the geometrical configuration for measuring the in-plane component of displacement, *u*.

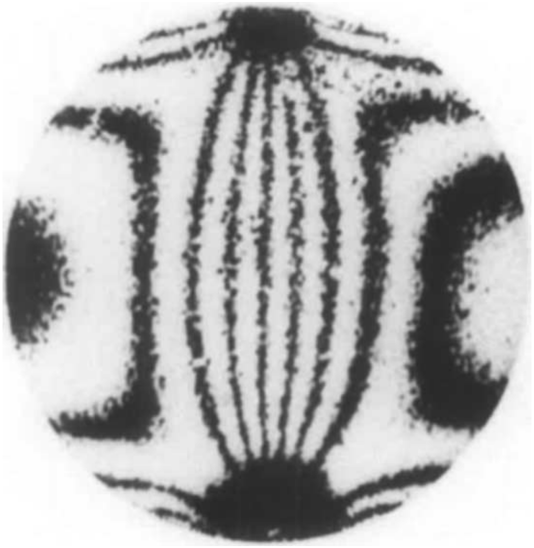

Figure 6 Fringe contours showing the in-plane displacement response of a disk under diametral compression.

where u is the displacement component in the x-direction and θ_e is the angle which the illumination beams make with the surface normal. Figure 6 shows the u pattern corresponding to a disk under diametral compression. The moiré pattern is obtained after optically filtering the interference image.

Comparative Holographic Interferometry

Another technique of significant interest in nondestructive testing is comparative holography that provides the contours of path variations related to the difference in displacements or shapes of two objects. The method provides information about the resolved part of the difference in displacement vector along the direction bisecting the illumination and observation rays. In the case of illumination and observation normal to the object surface, the fringe equation becomes

$$\Delta w = \frac{n\lambda}{2} \qquad [10]$$

where Δw is the difference in displacement component along z direction. The technique provides a tool to compare the mechanical responses of two nominally identical specimens subjected to the same loading and also the shapes of two nominally identical specimens. These features are useful in detecting anomalies in a test specimen with respect to the flaw-free master specimen. One such example is shown in Figure 7a, which displays the presence of a defect in a test specimen. In this case two plates, one flaw-free and the other with a flaw, are subjected to identical loadings. The fringe map in Figure 7b is related to the difference in displacements of test and master specimens subjected to unequal loadings. The presence of flaws is detected in both interferograms.

Basic Interferometers for Dynamic Displacement Measurement

Another important application of holographic interferometry is in the study of the resonant modes of vibration of an object.

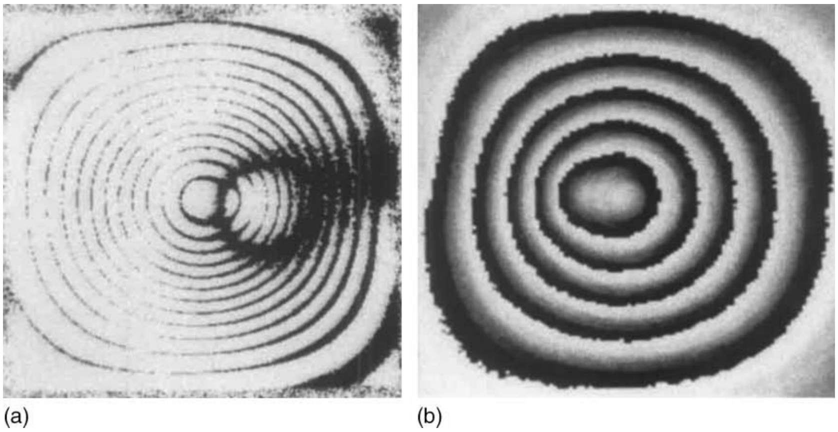
(a) (b)

Figure 7 Comparative holography reveals the presence of defects in aluminum plates using (a) identical and (b) unequal loadings.

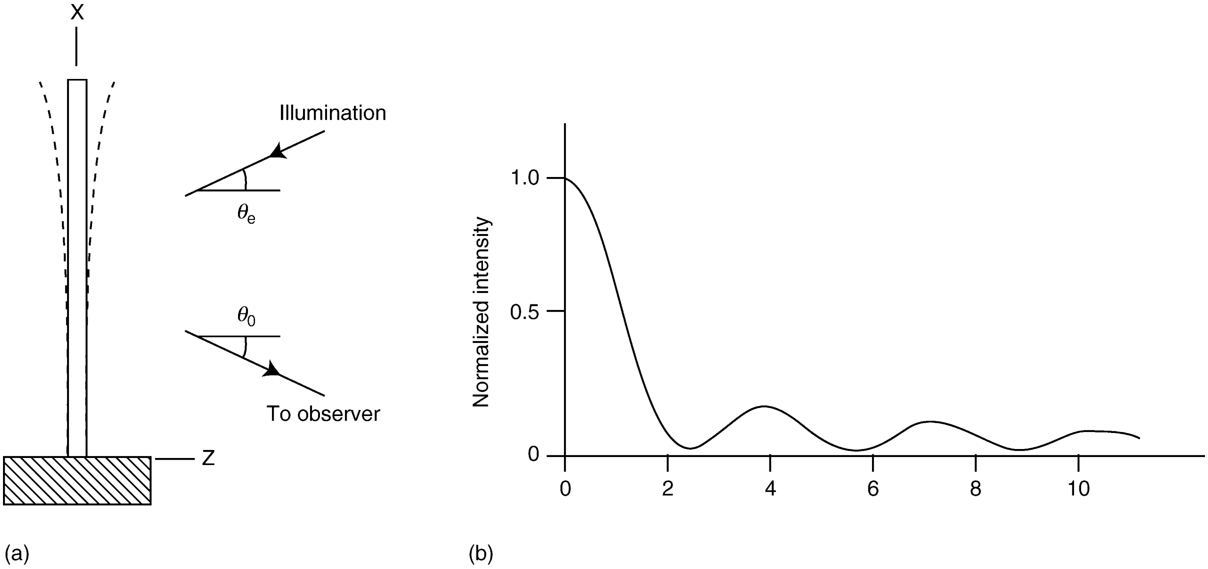

Figure 8 (a) Schematic of the geometrical configuration of a vibrating cantilever beam; (b) plot of fringe function for time-average holography.

Frozen Time-Average Holographic Interferometry

Consider that a hologram is recorded of an object, which is vibrating sinusoidally in a direction normal to its surface (Figure 8a). The exposure time is supposed to be much longer than the period of vibration. The intensity distribution of the image reconstructed by this hologram is

$$I(x, y) = I_o(x, y) J_o^2\left[\frac{2\pi}{\lambda} d(x, y)(\cos\theta_e + \cos\theta_o)\right] \quad [11]$$

where J_o is the zero-order Bessel function of the first kind. The virtual image is modulated by the $J_o^2(\xi)$ function. The dark fringes correspond to the zeros of the function $J_o^2(\xi)$. A plot of the function, shown in Figure 8b, is characterized by a comparatively brighter zero-order fringe, which corresponds to the nodes, a decreasing intensity, and an unequal spacing between the successive zeros. The interferogram in Figure 9 shows the vibrational modes of a helicopter component. The nodes that represent zero motion are clearly seen as the brightest areas in the time average reconstructed pattern.

Real-Time Time-Average Holographic Interferometry

The possibility of studying the response of vibrating objects in real-time extends the usefulness of the technique to identify the resonance states (resonant vibration mode shapes and resonant frequencies) of the object. A single exposure is made of the object in its state of rest. The plate is processed, returned to its original condition and reconstructed. The observer looking through the hologram at the sinusoidally

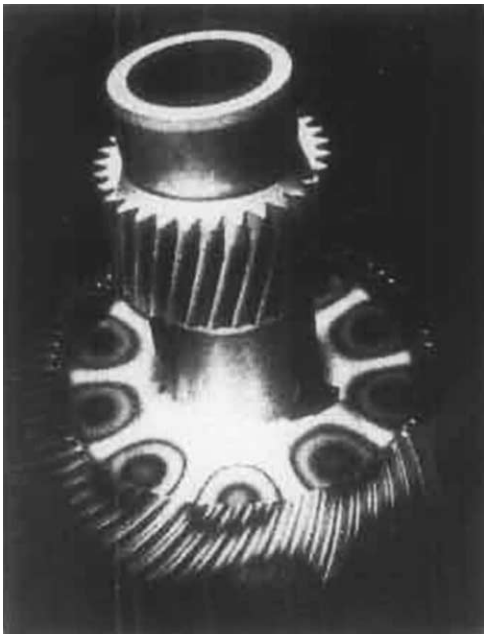

Figure 9 Fringe contours showing the mode shape of a vibrating helicopter component.

vibrating object sees the time-averaged intensity:

$$I(x,y) = I_o(x,y)\left[1 - J_o\left\{\frac{2\pi}{\lambda}d(x,y)(\cos\theta_e + \cos\theta_o)\right\}\right] \qquad [12]$$

Stroboscopic Holographic Interferometry

Stroboscopic holography is another interesting variant in the study of vibrations. The hologram is recorded by exposing the photographic plate twice for short time intervals during a vibration cycle. The pulsed exposures are synchronized with the vibrating surface, which is equivalent to making the surface virtually stationary during the recording. Reconstruction of the hologram yields cosinusoidal fringes, which are characteristic of double exposure holography. In real-time stroboscopic holography, a hologram of a nonvibrating object is first recorded. If the vibrating object is illuminated stroboscopically and viewed through the hologram, the reconstructed image from the hologram interferes directly with the light scattered from the object to generate live fringes.

Flow Measurement

Application of holographic interferometry to flow visualization and the measurement of spatial refractive index, density or temperature distributions have led to an advancement of understanding in areas such as aerodynamics, plasma diagnostics, and heat transfer. In the holographic interferometer shown in Figure 10 two consecutive exposures are made, usually the first exposure without flow and the second in the presence of a flow field. Double pulsed holography is used if the flow field is changing rapidly. The optical phase change due to flow between the exposures is

$$\Delta\varphi(x,y) = \frac{2\pi}{\lambda}\int_0^t \{n'(x,y,z) - n'_o\}dz \qquad [13]$$

where $n'(x,y,z)$ is the refractive index distribution during the second exposure, n'_o is the uniform refractive index during the first exposure, and t is the length of the test section. Assuming that the properties of the flow are constant in the z direction, the expression for phase change can be expressed as

$$\Delta\tau(x,y) = KL\{\rho(x,y) - \rho_o\} \qquad [14]$$

where K is the Gladstone–Dale constant, ρ is the density of the gas, and ρ_o is the density in a no-flow situation. The interference pattern contours the change in the density field of the flow. The change in density per fringe is given by λ/Kt. An example of application related to temperature measurements in an axisymmetric premixed flame is illustrated in Figure 11.

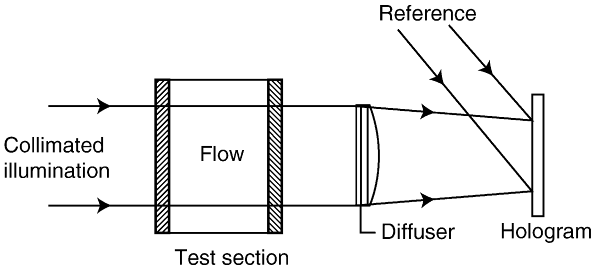

Figure 10 Schematic of a configuration for studying flow.

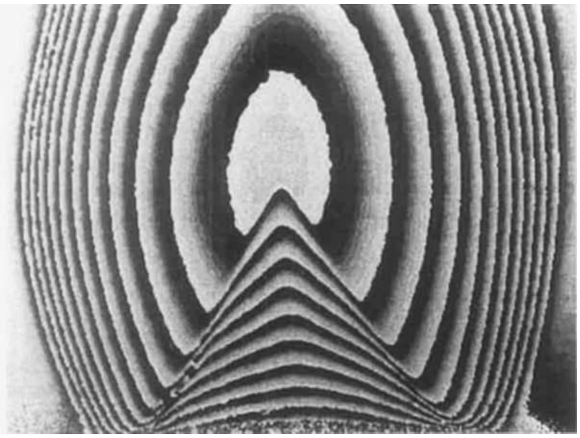

Figure 11 Fringe contours showing the phase distribution of a test flame.

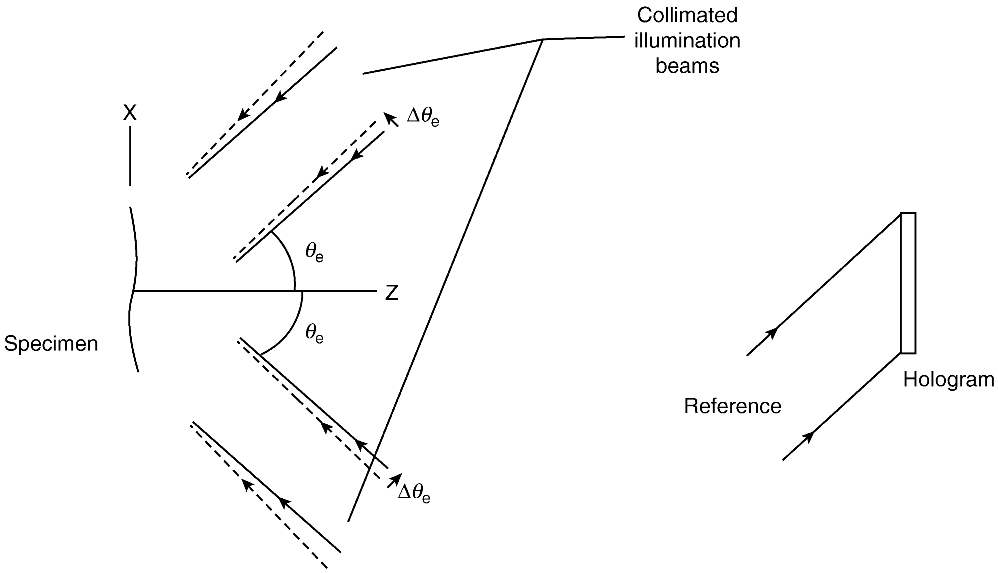

Figure 12 Schematic diagram of a two-beam method for shape measurement.

Surface Contouring

A practical way to display the shape of an object is to obtain contour maps showing the intersection of the object with a set of equidistant planes orthogonal to the line of sight.

The Dual Refractive Index Method

The method requires placing the model in a glass tank filled with a liquid (or gas) of refractive index n'_1. A hologram of the object is recorded, and the liquid contained in the tank is replaced by one of refractive index n'_2. An observer looking through the hologram sees the object surface modulated by a set of interference fringes arising from the change of optical phase in the light rays traversing the two liquids. The contour interval is given by

$$\Delta z = \frac{\lambda}{2(n'_1 - n'_2)} \qquad [15]$$

The Dual Illumination Method

In this method, the object is illuminated obliquely by collimated beams along two directions symmetric to the surface normal. The scattered waves are recorded on a holographic plate and the object and its reconstructed image are viewed through the hologram. Two sets of parallel equidistant fringes are projected onto the object surface by introducing small tilts, $\Delta\theta_e$, to the illumination beams (Figure 12). The contour sensitivity per fringe is given by

$$\Delta z = \frac{\lambda}{2\sin\theta_e \sin\Delta\theta_e} \qquad [16]$$

The sensitivity of the method can be tuned in a wide range. An example of topographic contour pattern is shown in Figure 13. The distance between two adjacent moiré contour planes is 172 μm.

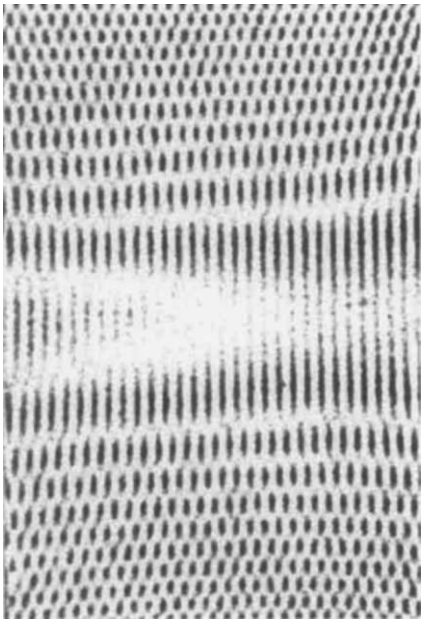

Figure 13 Example of a moiré topographic contour pattern obtained using the two-beam method.

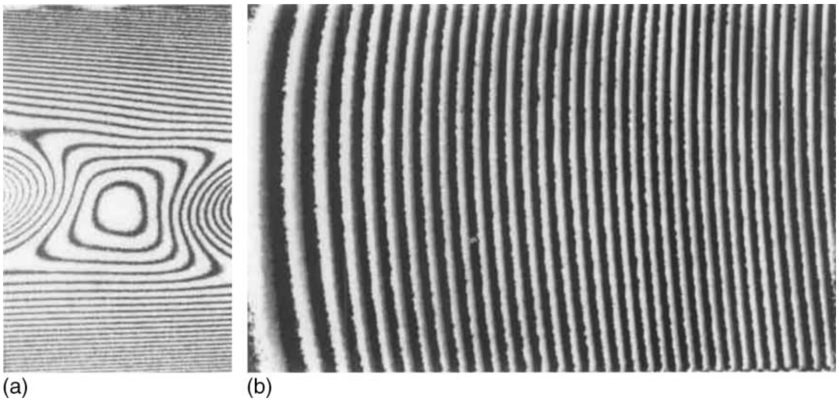

Figure 14 Examples of (a) shape measurement: the distance between two adjacent contour planes is 208 μm; and, (b) very large out-of-plane deformation measurement on a cantilever beam. The distance between two adjacent contour planes is 19 μm.

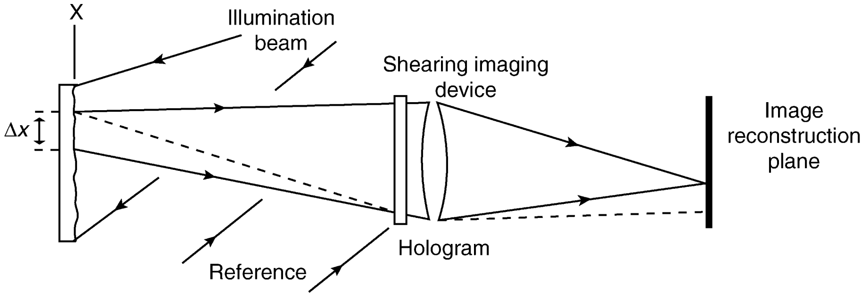
X
Illumination beam
Shearing imaging device
Image reconstruction plane
Δx
Hologram
Reference

Figure 15 Schematic diagram of the principle used in holographic shearing interferometry.

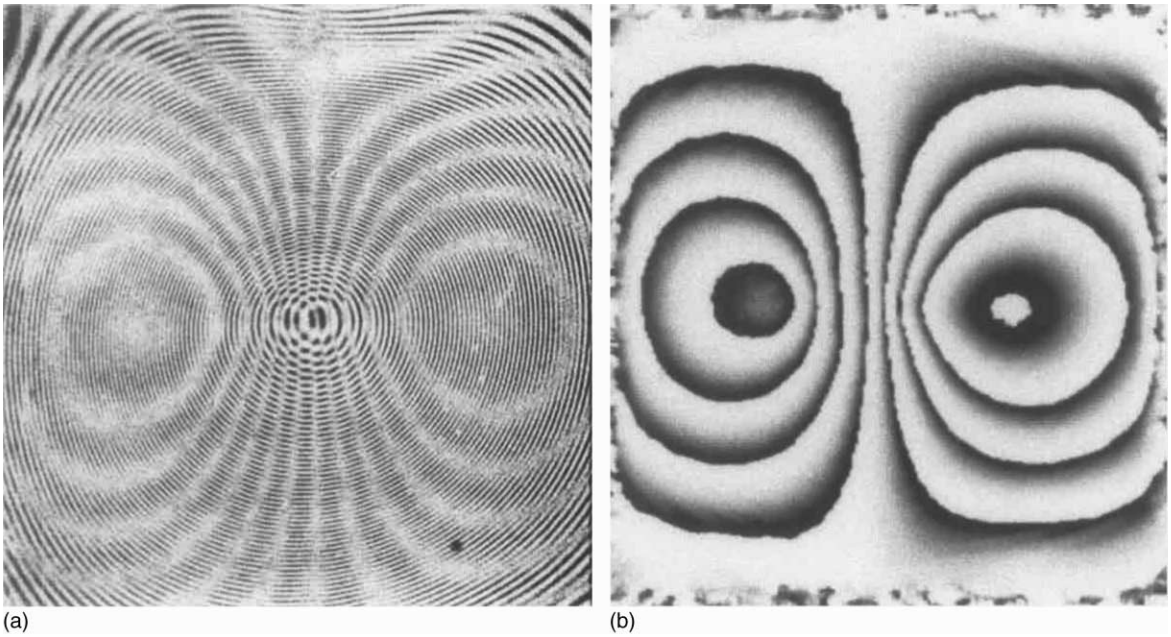

Figure 16 Examples of (a) moiré fringe contour, and (b) phase distribution corresponding to slope change produced on a centrally loaded aluminum plate clamped along its boundary.

In another method a collimated beam obliquely illuminates the object and a set of parallel equidistant fringes is projected onto the object surface by introducing a small tilt, $\Delta\theta_e$, to the illumination beam. Providing an appropriate reference beam rotation and a holographic plate translation generates contouring surfaces normal to the line of sight. A typical result using one beam illumination is shown in Figure 14a. The distance between two adjacent moiré contour planes is 208 μm.

This approach has been used to obtain measurement of large out-of-plane deformations undergone by a deformed object. The fringe pattern in Figure 14b corresponds to the out-of-plane displacement of a cantilever beam subjected to load at its free end. The distance between two adjacent contour planes is 19 μm.

Holographic Shearing Interferometry

Holographic shearing interferometry directly provides the patterns of slope change contours by laterally shearing wavefronts diffracted from the object surface. There are numerous ways to achieve shearing, such as, by introducing a Michelson type configuration, an inclined glass plate device, or a split-lens assembly in a holographic interferometer. The role of the shearing device is to enable the observation of a point on the object along two distinct neighboring directions. As a result of lateral wavefront shearing in the interferometer, a point in the image plane receives contributions from two different points on the object (Figure 15). Assuming that the illumination beam lies in the $x-z$ plane and makes an angle θ_e with the z-axis, and the observation is carried along the direction normal to the object surface, the change of optical phase due to deformation is

$$\Delta\varphi = \frac{2\pi}{\lambda}\left[\frac{\partial u}{\partial x}\sin\theta_e + \frac{\partial w}{\partial x}(1+\cos\theta_e)\right]\Delta x \qquad [17]$$

where Δx is the object shear along the x-direction. For $\theta_e = 0$, eqn [11] reduces to

$$\frac{\partial w}{\partial x} = \frac{n\lambda}{2\Delta x} \qquad [18]$$

Equation [18] displays a family of holographic moiré fringes corresponding to the contours of constant slope change. The fringe pattern shown in Figure 16a illustrates a moiré corresponding to contours of constant slope change for a centrally loaded square aluminum plate clamped along its boundary. An example of phase distribution

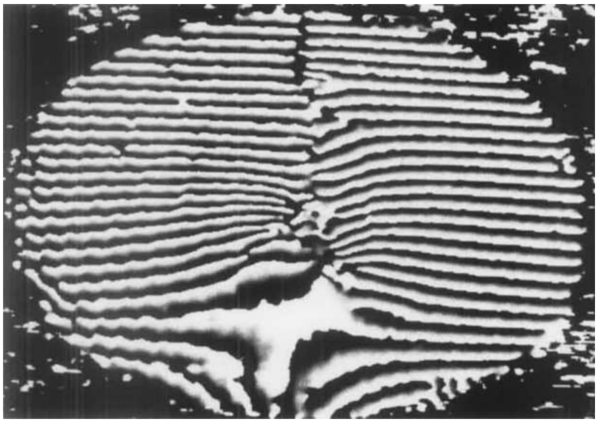

Figure 17 Example taken from the study of the fracture process zone in concrete: phase distribution corresponds to in-plane displacements on a notched concrete specimen subjected to a wedge splitting test.

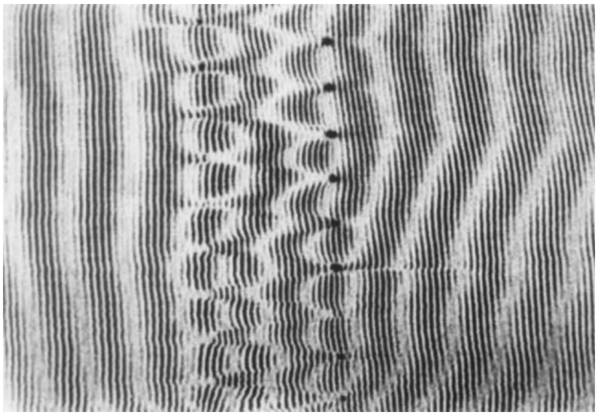

Figure 18 Example of application to nondestructive testing on a fingertip-joint wood specimen under tension using an in-plane sensitive holographic moiré setup.

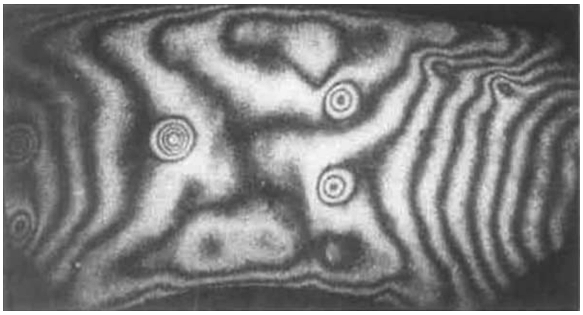

Figure 19 Application to the detection of defects in a tyre. Courtesy of H. Steinbichler, Steinbichler GmbH.

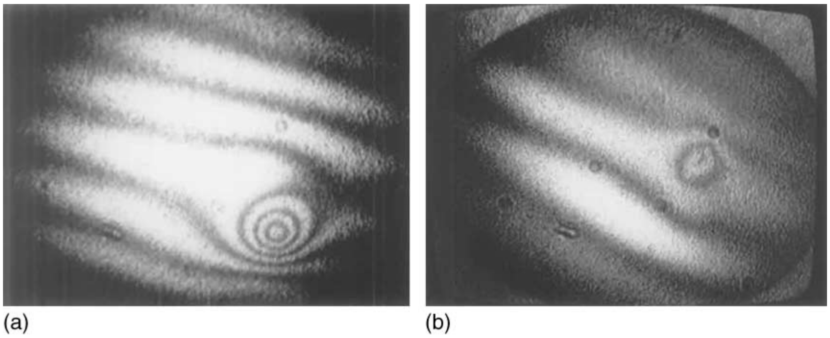

Figure 20 Fringe contours showing (a) a specific limb movement of a chick embryo during incubation, and (b) the hatching behavior. The point on the shell receiving blows is clearly visible.

corresponding to slope change along x-direction, is shown in Figure 16b.

Examples of Applications

Holographic interferometry has firmly established itself as an attractive and useful class of technique in nondestructive testing. Defects in a specimen are made apparent by the presence of anomalies in the interference fringe distribution of the deformed state of the specimen. The presence of flaws such as voids, disbonds, delaminations, and cracks weakens the strength of the specimen. The resulting strain concentration areas produced by the presence of flaws in the specimen are detected through their effect on the surface deformation map. The flaws in the specimen manifest themselves as distortion of the fringes, pockets of fringe concentrations, or localized discontinuities of the fringes in the full-field display of the deformation field of a test specimen in response to the applied load.

Figure 17 shows phase distribution corresponding to in-plane displacements of a notched fiber-reinforced concrete specimen subjected to a wedge-splitting test. The application is taken from the study of a fracture process zone in fiber-reinforced concrete. Another example of the application of holography in nondestructive testing is shown in Figure 18. The fringe pattern corresponds to in-plane displacements on a fingertip joint wood specimen under tension. Figure 19 shows an example of an application of holography in the tyre industry. The interferogram reveals, through the changes in the fringe pattern, defects in the tyre. The picture in Figure 20a is taken from the study of the embryonic behavior of embryos during incubation. The photograph shows a specific limb movement near the air space. The method also allows for a lively display of the hatching behavior of the chick. The point on the shell receiving blows is clearly visible in Figure 20b.

Conclusions

The aim of this article is to introduce readers to holographic techniques available for the noninvasive and whole-field investigation of the time-evolution and spatial distribution of the displacement and refractive index fields. Diverse techniques to obtain contourgrams of three-dimensional objects are also presented. The method offers strong potential in nondestructive testing applications, such as in the detection of flaws, in the conception of prototypes, in the evaluation of the performances of structural components, and in the verification of the reliability of finite-element computer codes.

See also

Holography, Techniques: Digital Holography; Overview. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Erf RK (1974) *Holographic Non-Destructive Testing*. New York: Academic.

Ostrovsky YI, Butusov MM and Ostrovskaya GV (1980) *Interferometry by Holography*. Berlin: Springer-Verlag.

Ostrovsky YI, Schepinov VP and Yakovlev VV (1991) *Holographic Interferometry in Experimental Analysis*. Berlin: Springer-Verlag.

Rastogi PK (1994) *Holographic Interferometry – Principles and Methods*. Berlin: Springer-Verlag.

Schumann W and Dubas M (1979) *Holographic Interferometry*. Berlin: Springer-Verlag.

Schumann W, Zürcher JP and Cuche D (1985) *Holography and Deformation Analysis*. Berlin: Springer-Verlag.

Vest CM (1979) *Holographic Interferometry*. New York: Interscience.

Sandwich Holography and Light in Flight

N Abramson, Royal Institute of Technology, Stockholm, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to Holographic Interferometry

Holography is a photographic principle that produces 3D images that are so real that it is possible to take measurements inside this image with a precision of a fraction of a thousandth of a millimeter. Contours are formed within the holographic image that, like the relief lines of a map, indicate the displacements of every point on the object surface. The method is named holographic interferometry (*see* **Holography, Techniques:** Holographic Interferometry) and it can be used to measure deformation, vibration and also, with slightly lower resolution, dimension.

In real time holography the deformed object is directly compared to the holographic image of its undeformed state. In double exposure holography the comparison is made between two holographic recordings on the same plate, while in sandwich holography the two states are recorded on two different plates that are sandwiched together for comparison.

Sandwich holography is useful in industrial environments, because it can measure much larger deformations and also compensate for unwanted movement of the object or the setup between the two exposures. Further on in the process it will produce any sign of deformation and comparison can be made at any time without referring back to original setup. However, processing of the two plates is needed before the comparison can be made and the plates have to be positioned with high accuracy.

Interferometric Sensitivity

To calculate the interferometric sensitivity let us name (A) the point from which the laser light is emitted, (B) the point behind the hologram plate from which the observation is made, and (C) the displaced point on the object surface to be measured. Draw an ellipsoid through (C) with (A) and (B) as focal points. If (C) is displaced along the surface of the ellipsoid, the path length ACB will remain unchanged and no interference fringes are formed, while a displacement normal to this surface produces maximal interference. A diagram, the holodiagram, is based on these ellipsoids and used for the making and evaluation of holograms.

The bisector to ACB will be the normal to the ellipsoids. If the path length ACB is repeatedly increased by δL a diagram like Figure 1 is formed where the separation (s) of the ellipses, and thus the interferometric sensitivity, varies over the diagram. If δL is the wavelength of the laser light, one interference fringe is formed each time an object point has moved the distance (s) because of a deformation of the object between the two exposures. The displacement (d) can be calculated:

$$d = ns \tag{1}$$

$$s = k0.5\delta L \tag{2}$$

$$k = 1/\cos\alpha \tag{3}$$

where n is number of fringes between fixed and studied point, k is constant along arcs of circles through A and B, and α is half the angle ACB.

When two slightly dissimilar patterns are placed one on top of the other, a moire pattern is formed that shows up the difference. The moire analogy to holographic interferometry is based on the fact that one fringe is formed for every ellipse that is crossed by an object point displaced between the two exposures.

If δL is the wavelength (λ) of the laser light the displacement (d) is

$$d = nk0.5\lambda \tag{4}$$

If two-wavelength holography is used to measure the 3D shape of an undisturbed object, it will be seen intersected by ellipses where the separation (s) is

$$s = k0.5\lambda_1\lambda_2/(\lambda_1 - \lambda_2) \tag{5}$$

If δL is the coherence length or the pulse length of the laser source, the object will be seen intersected by one single bright fringe, the thickness (s) of which is

$$s = k\delta L \tag{6}$$

Everything outside that ellipse will be dark during reconstruction. This phenomenon can be used either to optimize the use of a limited coherence length or to explain the results of 'light-in-flight recordings', as described later.

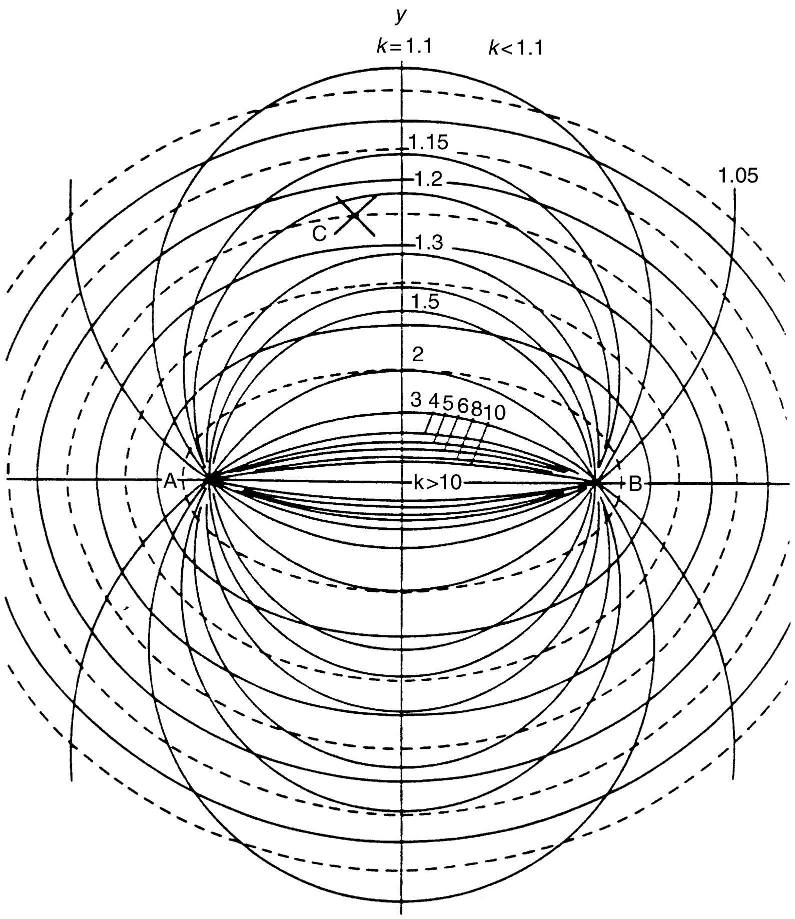

Figure 1 The holodiagram used for the evaluation of holograms.

The Principle of Sandwich Holography

In the following discussion on sandwich holography, we will only study the simplified example where the directions of illumination and observation are normal to the studied displacement and therefore α is equal to zero. The reflected light from an object point combined with the reference light produces a set of hyperbolas (Young's fringes) on the hologram plate.

When the object point is displaced between exposures on two different hologram plates, the hyperbolas will be positioned differently on the plates. There is a one-to-one relationship between the ellipsoids and the hyperbolas. Thus one point on the hologram plate will be crossed by one hyperbola if the corresponding object point is moved so that it crosses one of the ellipses of the holodiagram. Thus, the moire effect of the displaced hyperbolas at the hologram plate represents the moire effect of the ellipses in the object space. The interference pattern in the reconstructed object image can therefore be manipulated by moving one plate in relation to the other during reconstruction.

We have found that the simplest way to do this is to glue the two plates together separated by the glass thickness of one plate. To make the situation identical during exposure and reconstruction, the plates also have to be recorded in pairs. Thus, the first exposure can be made with one front plate and one back plate in contact in the hologram holder (Figure 2). The second exposure is made in exactly the same position with two new plates. After processing the front plate of the second exposure is glued in front of the back plate of the first exposure.

Then a reconstruction beam similar to the reference beam, is used to reconstruct the sandwich hologram. The interference fringes seen on the image of the object are manipulated by tilting the sandwich hologram, which results in an addition or subtraction of new fringes as in the following examples:

1. If the object is rotated by the small angle ϕ_1, the number of fringes can, during reconstruction, be

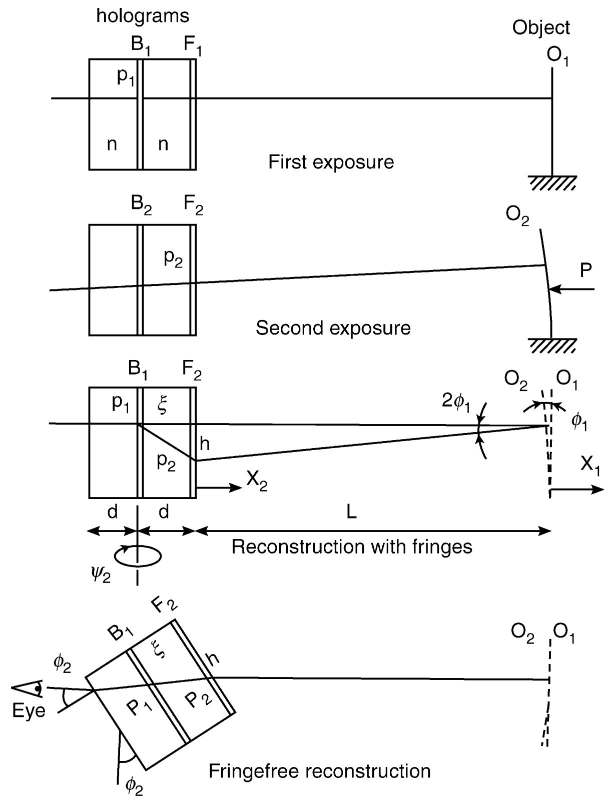

Figure 2 The exposures and reconstructions of a sandwich hologram.

eliminated by rotating the sandwich in the same direction by the much larger angle ϕ_2.

$$\phi_1 = \left[\frac{d}{2Ln}\right]\phi_2 \qquad [7]$$

2. If the sandwich during reconstruction is rotated around an axis perpendicular to that of the object rotation, new fringes are introduced at the angle γ, which is analogous to, but much larger than, the object rotation ϕ_1.

$$\phi_1 = \left[\frac{d}{2Ln}\right]\varphi_2 tg\gamma \qquad [8]$$

3. When the sandwich is rotated as described in (2) fringes are formed, the radius of curvature (r_2) of which are much larger than those of the bent object (r_1).

$$r_1 = \left[\frac{2Ln}{d}\right]\left[\frac{1}{\varphi_2}\right]r_2 \qquad [9]$$

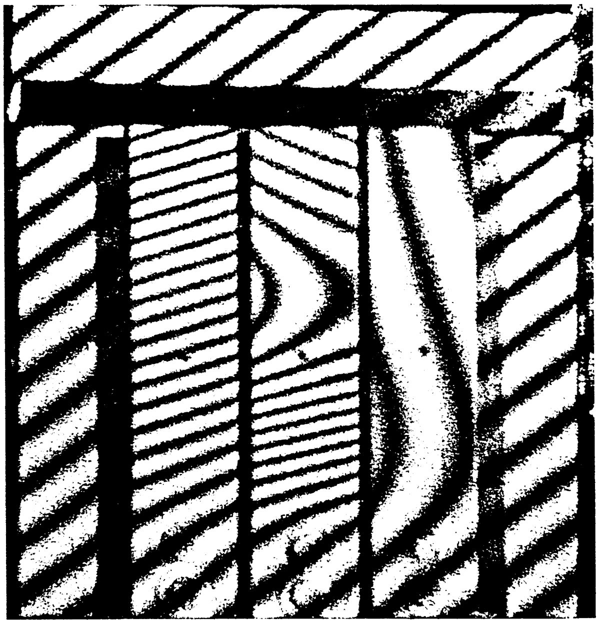

Figure 3 The movements of three steel bars in relation to a frame are to be measured by the fringes of sandwich holography. However, the unwanted motion of the total set-up makes the evaluation impossible.

Examples

The object consisted of three vertical steel bars (Figure 3) that were fixed by screws at their lower end to a rigid frame that surrounded all the bars. The middle bar was also supported at the upper end. Forces were applied at the middle of the bars. The right bar is deformed away from the observer while the other two are deformed towards him. After deformation of the three objects the second pair of plates were exposed. Then all four plates were processed.

The back plate of the first exposure (B_1 in Figure 2) was repositioned in the plate holder in front of the plate of the second exposure (F_2). The two plates were bonded together to form a sandwich hologram. O_1 and O_2 represent the two positions of one object, which between the two exposures was bent at angle ϕ_1 toward the hologram. (L) is the distance between the front hologram and the object (~1 m); (L) is also the distance between the point of illumination and the object. (D) is the distance between the point source of the reference beam and the hologram plate ($D = 2L$); (d) is the thickness of one hologram plate (~1.3 mm); (n) is the refractive index of the plates (~1.5 mm).

When the two plates were properly repositioned in the plate holder, fringes were seen everywhere, both on the steel bars and on the supposedly fixed frame which apparently had made an accidental movement (Figure 3). By tilting the plate holder in different directions, the fringes on the frame could be eliminated and the hologram could be evaluated as if the frame had remained fixed (Figure 4).

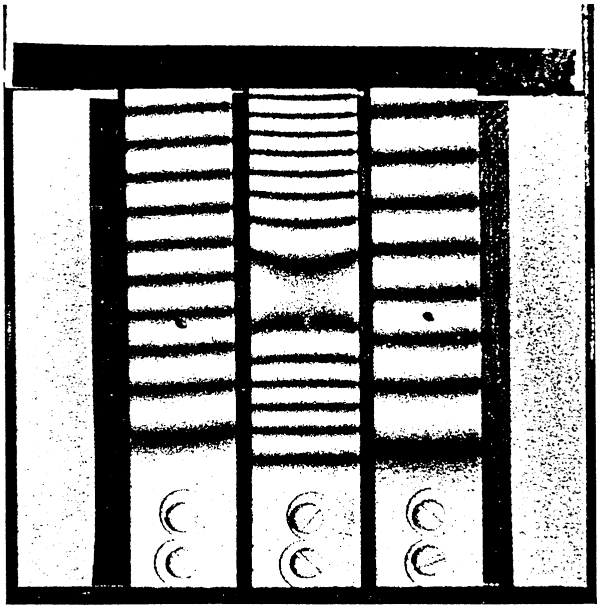

Figure 4 The unwanted fringes caused by the motion of the total set-up have been eliminated by a tilt of the sandwich hologram.

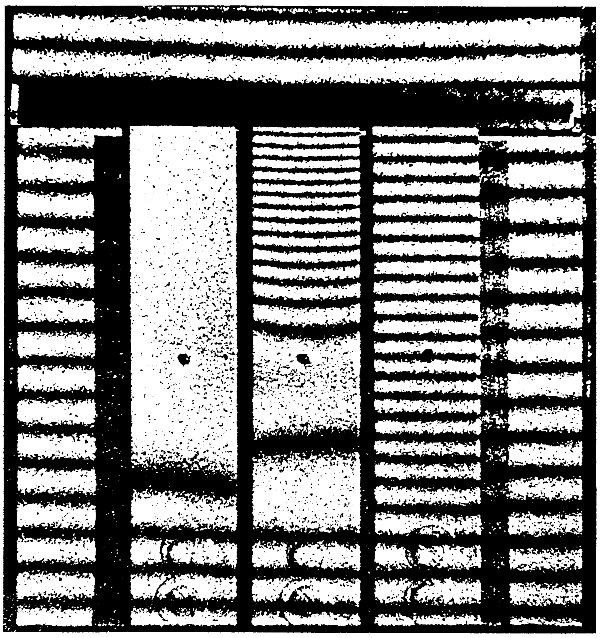

Figure 5 Another tilt of the sandwich hologram reveals that the left and the right bar have been deformed in opposite directions.

The plate holder was then tilted backwards at angle ϕ_2 and horizontal fringes were found on the frame while the top of the left bar became fringe-free (Figure 5). The number of fringes on the right bar increased, proving that it had been tilted in the opposite direction. If the sandwich were instead rotated around a vertical axis at angle φ_2, a set of vertical fringes were formed on the frame, while the bars became covered by fringes, the inclination and curvature of which represent a magnified view of the deformed bars as seen in profile (Figure 6).

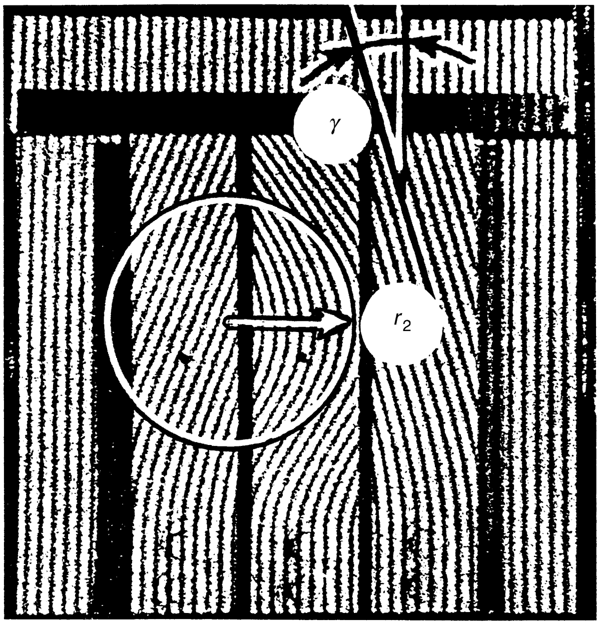

Figure 6 A rotation of the sandwich hologram around a vertical axis produces fringes that reveal direction of tilt and bending radius of all steel bars.

Introduction to Light-in-flight Recording by Holography

Recording of light pulses in flight have been possible since the beginning of the 1990s, using two-photon fluorescence or ultrafast Kerr cells driven by laser pulses. Very fast optical phenomena, such as refractive-index changes in laser-produced plasmas, have been recorded by holography using short illumination pulses.

In 1968, Staselko *et al.* published a method of studying the time coherence pattern of wave trains from pulsed lasers by the use of autocorrelation in which the unknown pulse is compared to itself. The method we name 'light-in-flight recording by holography' (LIF) is based on studying both the temporal and the spatial shape of pulses by correlation, which means that we study the change in shape of a pulse before and after it has passed through the transformation we want to study. For this process, we want to start with as short a pulse (or short coherence length) as possible and then compare the transformed pulse to the original pulse.

In a hologram only those parts of an object will be recorded for which the path length from the laser to the holographic plate via the object does not differ from the path length of the reference beam by more than the pulse length (or coherence length) of the laser light used for the recording. If that length is short, during reconstruction a large object will be seen intersected by an imaginary surface in the form of an ellipsoid, one of its two foci being the point from which the spherical wavefronts of illumination are emitted (spatial filter (A) in Figure 1) and the other being the point of observation (B). A spherical wavefront appears distorted into an ellipsoid if it is divergent, and into a hyperboloid if it is convergent, while a flat wavefront appears distorted into a paraboloid. The general rule is that the apparent wavefront has the shape of a mirror that focuses light from (A) to (B).

By careful planning, it is possible to optimize the geometry of the holographic setup in such a way that the utilized portions of the ellipsoids and hyperboloids can be approximated into flat surfaces. Thus, the object is seen intersected by one flat imaginary interference surface, the depth of which can be altered during reconstruction by changing the point of observation at the hologram plate.

Light of short duration, or short coherence length, can, in this way, be used in a holographic method for finding depth contours. If, however, the 3D shape of the object is known a priori, e.g., if it is a flat surface, this method can just as well be used to study the 3D shape of the illuminating wavefront. As mentioned before, the intersections of the ellipsoids can be approximated by intersections of spherical wave fronts of illumination, if the distance between object and hologram plate is long and if the flat surface is perpendicular to the line of sight.

Examples of Wavefront Studies

As an example we will describe an experiment that consists of an opaque white diffusely reflecting flat screen, that is illuminated from the left at an oblique angle, as seen in Figure 7. A mirror is fixed to the screen so that some of the light is reflected upwards. During recording of the hologram, the reference pulse is reflected by the two mirrors (M) so that it arrives at the middle of the hologram plate at the same time as the object pulse, in the form of diffuse light, arrives from the middle of the screen. As the reference pulse passes over the hologram plate from left to right, it works like a curtain shutter (focal-plane shutter) recording early object light to the left and later light to the right. This 'light shutter' moves across the plate with a velocity faster than light which is possible as it is just an intersection between pulsefront and plate. The difference in time of the recordings made at the two ends of the hologram plate is the time it takes for the reference pulse to move across the plate, or about 800 picoseconds. During reconstruction with a continuous laser the early arriving object light is seen through the left part of the plate, while the later arriving light is seen through its right part. If the eye, or a TV camera scans from left to right, the pulse is seen moving continuously over the object screen, with some of it reflected upwards by the mirror (Figure 8a–d). The corresponding pulse length is 100 ps (~30 mm), and to minimize pulsefront distortions the screen is perpendicular to the line of sight.

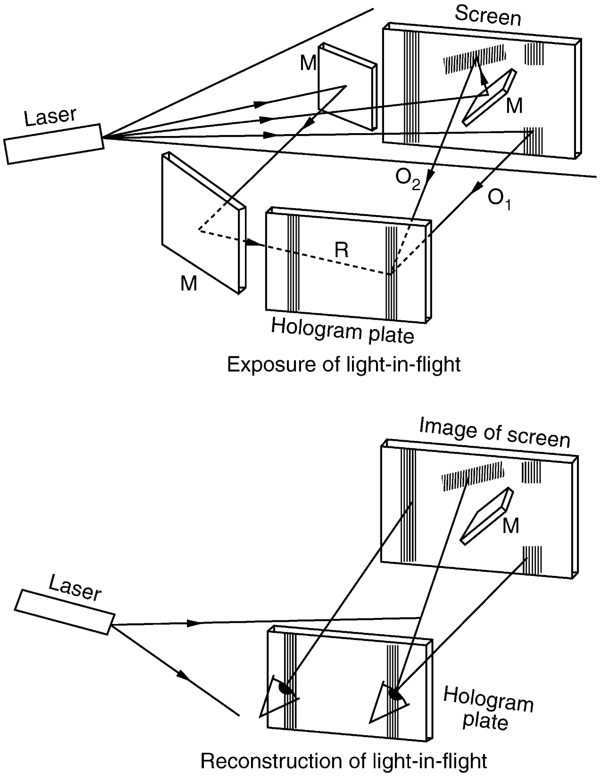

Figure 7 During exposure of the light-in-flight hologram, the short pulse from the laser is used to obliquely illuminate both the object and the hologram plate. Only those parts will be recorded where path length for object light and reference light are equal.

The photo in Figure 9 shows the result of another experiment. A spherical wave expands from a point source at left and moves to the right. The spherical wavefront intersects a flat observation screen consisting of a white-coated aluminum plate of 20 × 30 cm. A cylindrical lens is fixed, with its axis normal to the screen, so that it focuses the arriving light almost parallel to the screen surface. Figure 8 is a composite of several photographs taken through different parts

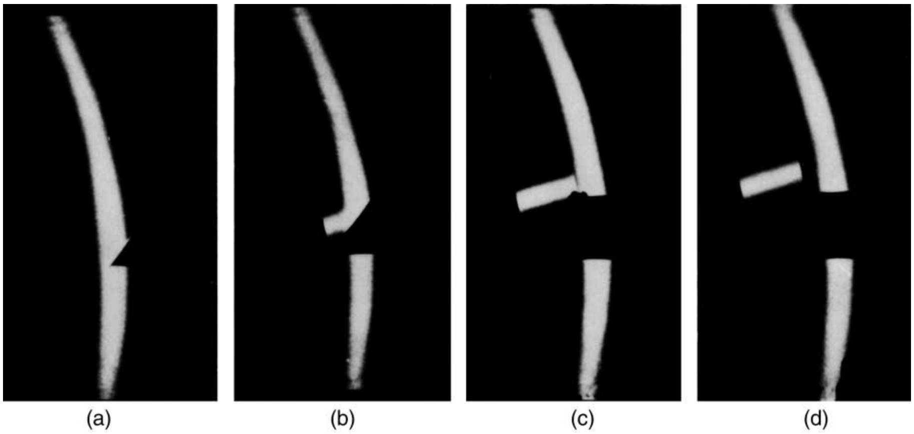

Figure 8 Resulting reconstructed images of the light reflected by a mirror at about 45°. These four photos were taken during reconstruction while the camera behind the hologram plate was moved from left to right in between each of the exposures.

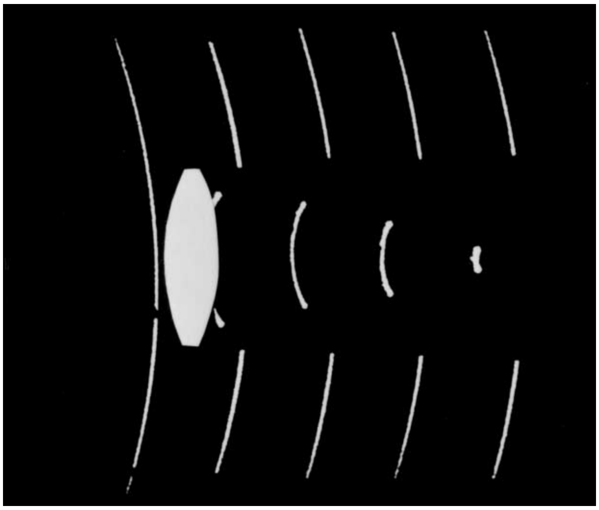

Figure 9 Light focused by a lens. A composite of several photos taken from different parts of the hologram plate. An image of the lens is included to make the picture clear.

of the hologram plate. The pulse length was in this case 10 ps (~3 mm).

Measuring the Shape of 3D Objects

The wavefront (pulse front) of a short pulse forms a thin sheet of light. The intersection of this sheet by a 3D body produces on its surface a thin line representing a cross-section. It is advantageous if this cross-section represents a flat surface, which will be approximately the case if the illumination and observation beams are collimated or arrive from points at large distances.

By moving the point of observation along the hologram plate from left to right, the intersecting sheet of light can be moved in depth forward or backward. Determination of fringe order number (by finding the zero fringe) no longer represents any problem because only the zero order fringe exists. Therefore, this way of contouring has great advantages, especially when the observations and calculations are made by computer. Figure 10 shows a set of spheres intersected by a thin light sheet.

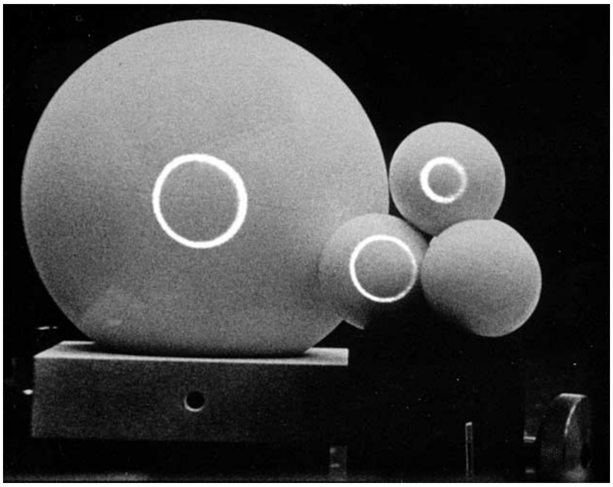

Figure 10 Spheres intersected by a thin sheet of light.

The intersecting light sheet can be manipulated in yet another way, i.e., it can be rotated. If the point of observation is moved away from the plate, the right part of an object will be seen through the right part of the plate were the recording made later than at the left part. During the time interval between left and right records, the intersecting light sheet has moved so that it intersects the object further away, resulting in a tilted cross-section of the object. This way it has been possible to rotate the cross-section up to 90 degrees, so that if, for example, a face is illuminated from the front by a flat sheet of light, the reconstructed intersection will show the face in profile.

In a limited way it is also possible to produce results similar to those of LIF by using a CCD camera instead of the holographic plate. The great advantage is its compatibility with a PC, while the main disadvantage is a lower resolution.

Other Uses

Like all methods of 'gated viewing', LIF can be used to look through scattering media, e.g., to look through human tissue. By observing only the first light the influence of the later arriving scattered light is eliminated. Thus, LIF can be used as a diagnostic method based on the idea that the scattering, absorption, or refractive index of cancer is different from that of healthy tissue. Another use of LIF is to image just a thin section of a gas or liquid filled with particles. In this way it is possible to observe the distribution and velocity in a certain localization without influence of the scattering from particles in front of or behind the studied section.

Up to now we have discussed gated viewing by LIF in the laboratory using picosecond light pulses and distances of up to a few meters. Instead we could look at astronomical objects with light pulses of days (e.g., a supernova) and distances of many light years. Even in these cases we will see particles in space illuminated by ellipsoids with (A) being the supernova and (B) a telescope on Earth. While each ellipsoid in the laboratory may represent a pathlength of some nanoseconds, each ellipsoid in space might represent light years.

However, the methods we have shown here are not limited to the study of the apparent shape of spheres of light. They can also be used to study other objects moving at ultrahigh velocities. If an exploding supernova throws out particles in all directions at a rather slow velocity, it can be seen surrounded by an expanding sphere. If, instead, this sphere expanded with the speed of light (which is not possible) it would appear to us just like an ellipsoid of the holodiagram. If, however, it expanded at a very high velocity, but lower than that of light, the sphere would appear to us in the shape of an egg with its pointed end towards the observer. The velocity of expansion could, in this case, appear superluminal (higher than the speed of light). If the velocity is (v), its approaching velocity will to a first approximation appear to be the true velocity divided by $(1 - v/c)$.

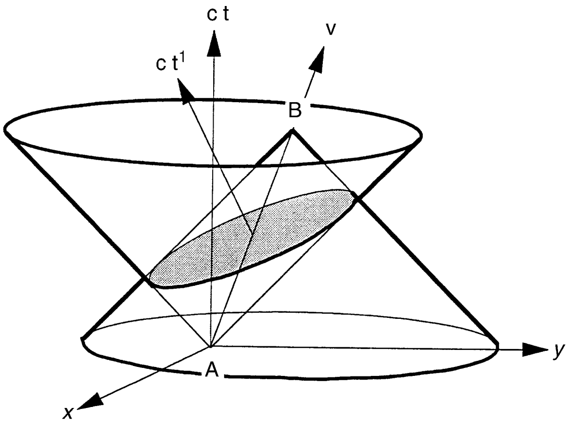

Figure 11 A modified Minkowski diagram consisting of one cone of illumination (A) and one cone of observation (B). The separation in time and space between (A) and (B) could either be static as in LIF or caused by a relativistic velocity of the observer from (A) to (B).

Finally, we will show a relationship between LIF and special relativity, by looking at a modification of the 'Minkowski light cones'. Figure 11 shows a diagram where (x) and (y) represent space coordinates, while (z) represents time multiplied by (c). A sphere from a short pulse of light that expands from the point A is represented by a cone with its apex at (A). A sphere from a short observation that converges to (B) is represented by a cone with its apex at (B). Only those object points, that are positioned at the elliptic intersection of the two cones, can be seen illuminated. This ellipse is identical to one of the ellipses of the holodiagram. The distance in time and space separating (A) and (B) could be either caused by the fixed distance between source of laser light and hologram plate in holographic LIF, or by a high velocity of the observer. Thus, the Lorentz contraction in special relativity could be explained as a measuring error because the high velocity of the observer traveling from (A) to (B) causes his sphere of observation to be transformed into an ellipsoid of observation. The tilt of the elliptic intersection indicates which observations that appear to be simultaneous depending on the separation of (A) and (B) along the y-axis, both in LIF and in relativity.

See also

Holography, Techniques: Holographic Interferometry.

Further Reading

Abramson N (1971) Moiré patterns and hologram interferometry. *Nature* 231: 65–67.

Abramson N (1981, 1986) *The Making and Evaluation of Holograms*. London: Academic Press.

Abramson N (1996) *Light in Flight, or, The Holodiagram – The Columbi Egg of Optics*. USA: SPIE Press, vol. PM 27.

An W and Carlsson T (1999) Digital measurement of three-dimensional shapes using light-in-flight speckle interferometry. *Optical Engineering* 38: 1366–1370.

Dugay MA and Mattick AT (1971) Ultrahigh speed photography of picosecond light pulses and echoes. *Applied Optics* 10: 2162–2170.

Garrington S (1992) Invisible jet sheds new light. *Nature* 355: 771–772.

Giordmaine J and Rentzepis P (1967) Two-photon excitation of fluorescence by picosecond light pulses. *Applied Physics Letters* 11: 216.

Hinrichs H and Hinsch KD (1997) Light-in-flight holography for flow visualization and velocimetry in three-dimensional flows. *Optics Letters* 22: 828–830.

Malin D and Allen D (1990) Echoes of the supernova. *Sky and Telescopes* 22–25, January.

Powell R and Stetson K (1965) Interferometric analysis by wavefront reconstruction. *Journal of the Optical Society of America* 55: 1593–1598.

Spears K and Serafin J (1989) Chrono-coherent imaging for medicine. *IEEE Transactions on Medical Engineering* 36: 1210–1220.

Staselko DI, Denisyuk YN and Smirnow AG (1969) Holographic recording of the time-coherence pattern of a wave train from a pulsed laser source. *Optics and Spectroscopy* 26: 413.

I

IMAGING

Contents

Information Theory in Imaging
Inverse Problems and Computational Imaging
Adaptive Optics
Hyperspectral Imaging
Imaging Through Scattering Media
Infrared Imaging
Interferometric Imaging
Lidar
Multiplex Imaging
Photon Density Wave Imaging
Three-Dimensional Field Transformations
Volume Holographic Imaging
Wavefront Sensors and Control (Imaging Through Turbulence)

Information Theory in Imaging

F O Huck and C L Fales, NASA Langley Research Center, Hampton, VA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Information theory is an intuitively attractive mathematical approach for combining observation with communication, or sensing with processing. So far, however, this approach has remained largely in the realm of signal processing and transmission, as first formalized by Shannon and Wiener over 50 years ago. Recent extensions of this approach to imaging have mostly addressed the interpretation of images, without explicitly accounting for the critical limiting factors that constrain image gathering and restoration. Yet, to be useful in the characterization of imaging, information theory must account for these constraints and lead to a close correlation between predicted and actual performance. The characterization that is outlined here traces the rate of information from the scene to the observer; and it pairs this rate, first, with the theoretical minimum data rate to assess the efficiency with which information can be transmitted, and, second, with the maximum-realizable fidelity to assess the quality with which images can be restored. Further details, especially for the mathematical development of figures of merit and for the design of imaging systems, can be found in the Further Reading.

Model of Imaging

Description

Modern imaging systems usually have the following three stages, as depicted in Figure 1:

1. *Image gathering*, to capture the radiance field that is either reflected or emitted by the scene and transform this field into a digital signal;

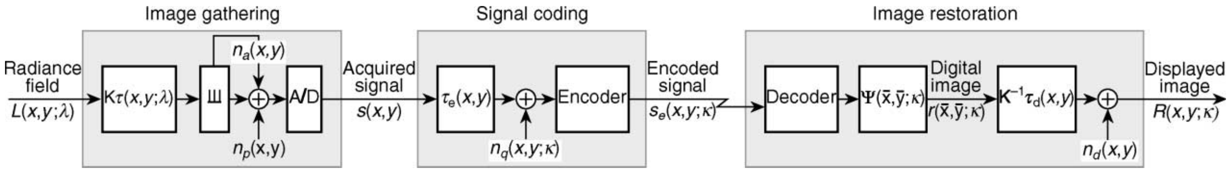

Figure 1 Model of digital imaging systems.

2. *Signal coding*, to encode the acquired signal for the efficient transmission and/or storage of data and decode the received signal for the restoration of images. The encoding may be either lossless (without loss of information) or lossy (with some loss of information), if a higher compression is required;
3. *Image restoration*, to produce a digital representation of the scene and transform this representation into a continuous image for the observer.

Human vision may be included in this model as a fourth stage, to characterize imaging systems in terms of the information rate that the eye of the observer conveys from the displayed image to the higher levels of the brain.

These stages are similar to each other in one respect: each contains one or more transfer functions, either continuous or discrete, followed by one or more sources of noise, also either continuous or discrete. The continuous transfer function of the image-gathering device (and of the human eye) is followed by sampling that transforms the captured radiance field into a discrete signal with analog magnitudes. In most devices the photodetection mechanism conveys this signal serially into an analog-to-digital (A/D) converter to produce a digital signal. However, analog signal processing in a parallel structure, akin to that in human vision, has many advantages that are increasingly emulated by neural networks.

This model of imaging differs from the classical model of communication that Shannon and Wiener addressed in two fundamental ways:

1. *The continuous-to-discrete transformation in image gathering*. Whereas communication is constrained critically only by bandwidth and noise, image gathering is constrained also by the compromise between blurring and aliasing, due to limitations in the response of optical apertures and in the sampling density of photodetection mechanisms. This additional constraint requires the rigorous treatment of insufficiently sampled signals throughout the imaging system.
2. *The sequence of image gathering followed by signal coding*. Whereas the characterization of communication addresses the perturbations that occur at the encoder and decoder or during transmission, the characterization of imaging systems must also address the perturbations in the image-gathering process prior to coding. These additional perturbations require a clear distinction between the information rate of the encoded signal and the associated theoretical minimum data rate or, more generally, between information and entropy.

The mathematical model of imaging that is outlined here accounts for the appropriate deterministic and stochastic (random) properties of natural scenes and for the critical limiting factors that constrain the performance of imaging systems. The corresponding formulations of the figures of merit are sufficiently general to account for a wide range of radiance field properties, photodetection mechanisms, and signal processing structures. However, the illustrations emphasize: (a) the salient properties of natural scenes that dominate the captured radiance field in the visible and near-infrared region under normal daylight conditions; and (b) the photodetector arrays that are used most frequently as the photodetection mechanism of image-gathering devices.

Radiance Field

The radiance field $L(x, y; \lambda)$, reflected by natural scenes, may be modeled approximately as

$$L(x, y; \lambda) = \frac{1}{\pi} I(\lambda)\rho(\lambda)\rho(x, y) \qquad [1]$$

where $I(\lambda)$ and $\rho(\lambda)$, respectively, represent the spectral irradiance and reflectance properties as a function of the wavelength λ, and $\rho(x, y)$ represents the random reflectance properties as a function of the spatial coordinates (x, y). Figure 2 illustrates a model of $\rho(x, y)$ for which the power spectral density (PSD) $\hat{\Phi}_\rho(\upsilon, \omega)$ corresponds closely to that typical of natural scenes, as given by

$$\hat{\Phi}_\rho(\upsilon, \omega) = \frac{2\pi\mu_\rho^2\sigma_\rho^2}{[1 + (2\pi\mu_\rho\zeta)^2]^{3/2}} \qquad [2]$$

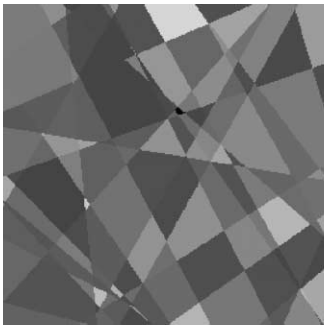

Figure 2 Model of reflectance $\rho(x, y)$.

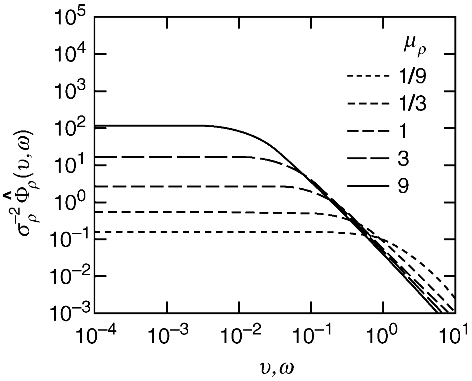

Figure 3 Normalized PSD of the reflectance $\rho(x, y)$.

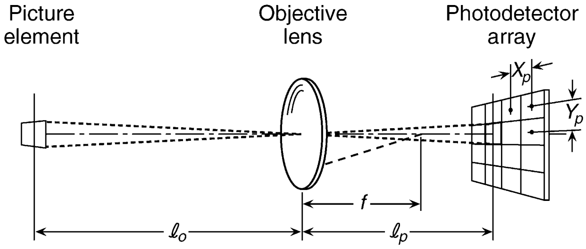

Figure 4 Optical geometry of image-gathering device.

where $\zeta^2 = \upsilon^2 + \omega^2$ and (υ, ω) are the spatial frequency coordinates of the Fourier-transform domain (see Figure 3). This model consists of random polygons, whose boundaries are distributed according to Poisson probability with a mean separation of μ_p, and whose magnitudes are distributed according to independent zero-mean Gaussian statistics of variance σ_ρ^2. The constraints associated with eqn [1] are satisfied by letting the mean reflectance $\bar{\rho} = 0.5$ and the standard deviation $\sigma_\rho \leq 0.2$.

Image Gathering

Figure 4 depicts the basic optical geometry of an image-gathering device, in which the objective lens forms an image of the captured radiance field across the photodetector array, and the array transforms this image into a discrete signal. The sampling lattice (X, Y), which determines the resolution of this device, is formed in angular units by the separation (X_p, Y_p) of the photodetector-array apertures and the distance ℓ_p as given for one of the dimensions by

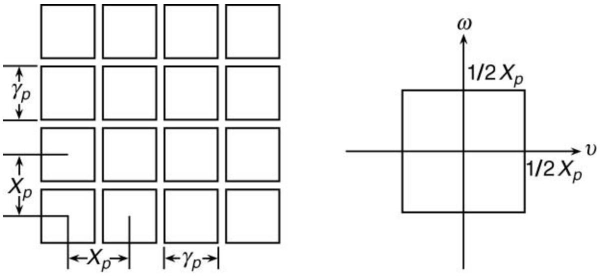

(a) Geometry (b) Sampling passband

Figure 5 Photodetector-array geometry and sampling passband.

$$X = 2\tan^{-1}\frac{X_p}{2\ell_p} \approx \frac{X_p}{\ell_p}\text{rad} = \frac{57.3X_p}{\ell_p}\text{deg} \qquad [3]$$

Normally, the sampling lattice is square, as shown in Figure 5, so that $X_p = Y_p$. Then the corresponding area of the photodetector apertures is $A_p = \gamma_p^2$, and the solid angle Ω of the instantaneous field of view (IFOV) is $\Omega \approx A_p/\ell_p^2$ ster.

The sampling lattice projected onto the scene at the distance ℓ_o has the dimensions $(X_o, Y_o) \approx (X\ell_o, Y\ell_o)$. Because these dimensions represent the finest spatial detail that normally can be restored, it is convenient to measure the mean spatial detail of the scene relative to the sampling lattice, as given by $\mu = \mu_\rho(X_oY_o)^{-1/2}$, which reduces to $\mu = \mu_\rho X_o^{-1}$ for $X = Y$. Hence, $\mu = 1$ indicates that the sampling interval is equal to the mean spatial detail.

The image-gathering process that transforms the continuous radiance field $L(x, y; \lambda)$ into the discrete signal $s(\mathbf{x}, \mathbf{y})$ may be modeled by the expression

$$s(\mathbf{x}, \mathbf{y}) = K\rho(\mathbf{x}, \mathbf{y}) * \tau(\mathbf{x}, \mathbf{y}) + n_p(\mathbf{x}, \mathbf{y}) \qquad [4]$$

where the discrete coordinates $(\mathbf{x}, \mathbf{y}) = (mX, nY)$, and the symbol $*$ denotes continuous spatial convolution. This model accounts, in the simplest form, for the reflectance-to-signal conversion gain K, the reflectance $\rho(x, y)$ of the scene, the spatial response (SR) $\tau(x, y)$ of the image-gathering device, the continuous-to-discrete transformation by the photodetector-array lattice, and the discrete noise $n_p(\mathbf{x}, \mathbf{y})$ of the photodetectors. The discrete coordinates are defined by the sampling function

$$\underline{\text{Ш}} \equiv \underline{\text{Ш}}(\mathbf{x}, \mathbf{y}) = XY \sum_{\mathbf{x},\mathbf{y}} \delta(x - \mathbf{x}, y - \mathbf{y}) \qquad [5]$$

where $\delta(x, y)$ is the Dirac delta function; and the gain is given by

$$K = \frac{1}{\pi} A_\ell \Omega \int I(\lambda)\rho(\lambda)r(\lambda)\mathrm{d}\lambda \qquad [6]$$

where $A_\ell = \pi D^2/4$ is the area of the objective lens with diameter D, and $r(\lambda)$ is the responsivity of the photodetectors. This gain relates the magnitude of the signal $s(\mathbf{x}, \mathbf{y})$ directly to that of the reflectance $\rho(x, y)$; and, together with the variances σ_ρ^2 and σ_p^2 of the reflectance and the photodetector noise, respectively, it also determines the rms-signal-to-rms-noise ratio (SNR) $K\sigma_\rho/\sigma_p$.

The DFT of $s(\mathbf{x}, \mathbf{y})$ is the periodic spectrum $\tilde{s}(\upsilon, \omega)$ given by

$$\tilde{s}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) * \hat{\underline{\text{Ш}}} + \tilde{n}_p(\upsilon, \omega) \qquad [7]$$

where $\hat{\rho}(\upsilon, \omega)$ is the spectrum of the reflectance, $\hat{\tau}(\upsilon, \omega)$ is the spatial frequency response (SFR) of the image-gathering device, and $\tilde{n}_p(\upsilon, \omega)$ is the spectrum of the photodetector noise. The symbol $\hat{\underline{\text{Ш}}}$ is the Fourier transform of the sampling function $\underline{\text{Ш}}$. It accounts for the sidebands generated by the sampling process and is given by

$$\hat{\underline{\text{Ш}}} \equiv \hat{\underline{\text{Ш}}}(\upsilon, \omega) = \sum_{\mathbf{v},\mathbf{w}} \delta(\upsilon - \mathbf{v}, \omega - \mathbf{w}) \qquad [8]$$

where $(\mathbf{v}, \mathbf{w}) = (m/X, n/Y)$. The corresponding sampling passband is

$$\hat{B} = \left\{(\upsilon, \omega);\ |\upsilon| < \frac{1}{2X},\ |\omega| < \frac{1}{2Y}\right\} \qquad [9]$$

The above expression for $\tilde{s}(\upsilon, \omega)$ can be rewritten more conveniently as

$$\tilde{s}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) + \hat{n}(\upsilon, \omega) \qquad [10]$$

where

$$\hat{n}(\upsilon, \omega) = \hat{n}_\mathrm{a}(\upsilon, \omega) + \tilde{n}_p(\upsilon, \omega) \qquad [11]$$

is the spectrum of the total noise in the acquired signal. The first term $\hat{n}_\mathrm{a}(\upsilon, \omega)$ represents the aliased signal components caused by insufficient sampling, as given by

$$\hat{n}_\mathrm{a}(\upsilon, \omega) = K\hat{\rho}(\upsilon, \omega)\hat{\tau}(\upsilon, \omega) * \hat{\underline{\text{Ш}}}_\mathrm{s} \qquad [12]$$

where

$$\hat{\underline{\text{Ш}}}_\mathrm{s} \equiv \hat{\underline{\text{Ш}}}_\mathrm{s}(\upsilon, \omega) = \sum_{\mathbf{v},\mathbf{w} \neq (0,0)} \delta(\upsilon - \mathbf{v}, \omega - \mathbf{w}) \qquad [13]$$

If, for spatial coordinates at either the scene or the photodetector array, the unit of (x, y) is meters, then the unit of (υ, ω) is cycles m^{-1}; and if, for angular coordinates centered at the objective lens, the unit of (x, y) is radians, then the unit of (υ, ω) is cycles rad^{-1}. Either way, it is convenient to normalize the sampling intervals to unity (i.e., $X = Y = 1$). Then the area of the sampling passband is $|\hat{B}| = 1$, the unit of (x, y) is samples, and the unit of (υ, ω) is cycles sample^{-1}.

The SFR $\hat{\tau}(\upsilon, \omega)$ of the image-gathering device is

$$\hat{\tau}(\upsilon, \omega) = \hat{\tau}_\ell(\upsilon, \omega)\hat{\tau}_p(\upsilon, \omega) \qquad [14]$$

where $\hat{\tau}_\ell(\upsilon, \omega)$ and $\hat{\tau}_p(\upsilon, \omega)$ are the SFRs of the objective lens and photodetector apertures, respectively. This SFR may be modeled approximately by the circularly symmetric Gaussian shape

$$\hat{\tau}(\upsilon, \omega) = \exp[-(\zeta/\zeta_\mathrm{c})^2] \qquad [15]$$

where, as shown in Figure 6, the optical-response index ζ_c controls the trade-off between the blurring that is caused by the decrease of the SFR inside the sampling passband and the aliasing that is caused by the extension of the SFR beyond this passband.

Signal Coding

Signal coding may be divided into two stages, as depicted in Figure 1. The first stage, which is represented by the digital operator $\tau_\mathrm{e}(\mathbf{x}, \mathbf{y})$ and the quantization noise $n_q(\mathbf{x}, \mathbf{y}; \kappa)$, accounts for manipulations of the acquired signal that seek to extract those

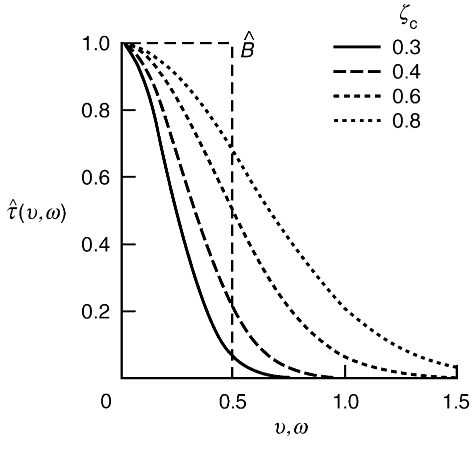

Figure 6 SFRs $\hat{\tau}(\upsilon, \omega)$ of the image-gathering device relative to the sampling passband $\hat{B}$ for unit sampling intervals (i.e., $X = Y = 1$).

features of the captured radiance field considered to be perceptually most significant, which usually entails some loss of information. The second stage, which is represented by the encoder, accounts for manipulations that seek to reduce redundancies in the acquired signal, without loss of information. The resultant encoded signal $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ may be expressed in the same form as eqn [4] for the acquired signal $s(\mathbf{x}, \mathbf{y})$ as

$$s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa) = s(\mathbf{x}, \mathbf{y}) \otimes \tau_{\mathrm{e}}(\mathbf{x}, \mathbf{y}) + n_q(\mathbf{x}, \mathbf{y}; \kappa) \qquad [16]$$

where the symbol $\otimes$ denotes discrete spatial convolution, and $n_q(\mathbf{x}, \mathbf{y}; \kappa)$ is the quantization noise. If the data transmission and/or storage are error-free, as is assumed here, then the signal that leaves the decoder remains identical to the signal that enters the encoder.

The DFT of $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ is the periodic spectrum $\tilde{s}_{\mathrm{e}}(\upsilon, \omega; \kappa)$ given in the same form as eqn [10] by

$$\tilde{s}_{\mathrm{e}}(\upsilon, \omega; \kappa) = K\hat{\rho}(\upsilon, \omega)\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega) + \hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa) \qquad [17]$$

where

$$\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega) \qquad [18]$$

is the throughput SFR from image gathering to transmission, and

$$\hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa) = \hat{n}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega) + \tilde{n}_q(\upsilon, \omega; \kappa) \qquad [19]$$

is the accumulated noise in the encoded signal. For η-bit quantization, where $\eta = \log \kappa$ and κ is the number of uniformly spaced quantization levels, $\tilde{n}_q(\upsilon, \omega; \kappa)$ may be treated (for $\eta \geq 4$) as independent Gaussian noise with the PSD

$$\tilde{\Phi}_q(\upsilon, \omega; \kappa) = \sigma_q^2 = (\sigma_{s_{\mathrm{e}}}/\kappa)^2 \qquad [20]$$

where $\sigma_{s_{\mathrm{e}}}^2$ is the variance given approximately by

$$\sigma_{s_{\mathrm{e}}}^2 \approx K^2 \int_{\hat{B}}\!\!\int \hat{\Phi}_\rho(\upsilon, \omega)|\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega)|^2 \mathrm{d}\upsilon \mathrm{d}\omega \qquad [21]$$

Image Restoration

Image restoration, as depicted in Figure 1, transforms the decoded signal $s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ into the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$, as modeled by the expression

$$r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) = \overline{|||}s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa) \otimes \Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa) \qquad [22]$$

The operation indicated by $\overline{|||}s_{\mathrm{e}}(\mathbf{x}, \mathbf{y}; \kappa)$ (called zero padding) prepares the decoded signal for image restoration with interpolation, and the operator $\Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ produces the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ on the interpolation lattice $\overline{|||}$ that is Z times denser than the sampling lattice $\underline{|||}$. If the interpolation in the image restoration is sufficiently dense (normally, for $Z \geq 4$), then the blurring and raster effects of the image-display process are suppressed entirely. This condition leads not only to the best realizable image quality but also to the simplest expression for the continuous displayed image, as given in the same form as eqns [10] and [17] by

$$\hat{R}(\upsilon, \omega; \kappa) = \hat{\rho}(\upsilon, \omega)\hat{\Gamma}_r(\upsilon, \omega; \kappa) + \hat{n}_r(\upsilon, \omega; \kappa) \qquad [23]$$

where

$$\hat{\Gamma}_r(\upsilon, \omega; \kappa) = \hat{\tau}(\upsilon, \omega)\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)\hat{\Psi}(\upsilon, \omega; \kappa) \qquad [24]$$

is the throughput SFR from image gathering to display, and

$$\begin{aligned}\hat{n}_r(\upsilon, \omega; \kappa) &= K^{-1}\hat{n}_{\mathrm{e}}(\upsilon, \omega; \kappa)\hat{\Psi}(\upsilon, \omega; \kappa)\\ &= \hat{n}_{\mathrm{ar}}(\upsilon, \omega; \kappa) + \hat{n}_{pr}(\upsilon, \omega; \kappa)\\ &\quad + \hat{n}_{qr}(\upsilon, \omega; \kappa)\end{aligned} \qquad [25]$$

is the accumulated noise in the restored image. This noise may be expressed explicitly by substituting the noise components given above. The noise due to the granularity in image-display mediums is not accounted for here.

The Wiener filter, that minimizes the mean-square error due to the perturbations in image gathering and coding, is given by

$$\hat{\Psi}(\upsilon, \omega; \kappa) = \frac{\hat{\Phi}'_\rho(\upsilon, \omega)\hat{\Gamma}^*_{\mathrm{e}}(\upsilon, \omega)}{\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\Gamma}_{\mathrm{e}}(\upsilon, \omega)|^2 + \hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa)} \qquad [26]$$

where $\hat{\Phi}'(\upsilon, \omega) = \sigma_\rho^{-2}\hat{\Phi}_\rho(\upsilon, \omega)$ is the normalized PSD of the radiance field and

$$\begin{aligned}\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa) = (K\sigma_\rho)^{-2}\Big[&\hat{\Phi}_n(\upsilon, \omega)|\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)|^2\\ &+ \tilde{\Phi}_q(\upsilon, \omega; \kappa)\Big]\end{aligned}$$

is the normalized PSD of the accumulated noise in the encoded signal. If the photodetector noise is white so that the PSD $\tilde{\Phi}_p(\upsilon, \omega)$ is equal to its variance σ_p^2 and the PSD $\tilde{\Phi}_q(\upsilon, \omega; \kappa)$ is equal to its variance σ_q^2, then the PSD of the noise simplifies to

$$\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega; \kappa) = \hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega) + \tilde{\Phi}'_q(\upsilon, \omega; \kappa) \qquad [27]$$

where

$$\begin{aligned}\hat{\Phi}'_{n_{\mathrm{e}}}(\upsilon, \omega) = \Big[&\hat{\Phi}'_\rho(\upsilon, \omega)|\hat{\tau}(\upsilon, \omega)|^2 * \hat{\underline{|||}}_{\mathrm{s}}\\ &+ (K\sigma_\rho/\sigma_p)^{-2}\Big]|\tilde{\tau}_{\mathrm{e}}(\upsilon, \omega)|^2\end{aligned} \qquad [28]$$

and

$$\hat{\Phi}'_q(\upsilon, \omega; \kappa) = (K\sigma_\rho/\sigma_{s_{\mathrm{e}}})^{-2}\kappa^{-2} \qquad [29]$$

These simplifying conditions permit the Wiener filter given above and the figures of merit given below to be expressed as a function of the SNR $K\sigma_\rho/\sigma_p$, which can be specified without accounting explicitly for either the gain K or the variances σ_ρ^2 and σ_p^2.

Figures of Merit

Information Rate $\mathcal{H}$

The information rate $\mathcal{H}$ (in bits sample^{-1}) of the encoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ is defined by

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \hat{\rho}(v, \omega) : (v, \omega) \in \hat{B}] \qquad [30]$$

where the first term, $\mathcal{E}[\cdot]$, represents the entropy of $\tilde{s}_e(v, \omega; \kappa)$ and the second term, $\mathcal{E}[\cdot|\cdot]$, represents the conditional entropy of this signal when the reflectance spectrum $\hat{\rho}(v, \omega)$ within the sampling passband $\hat{B}$ is known. For the assumptions that have been made about image gathering and coding, the conditional entropy becomes the entropy of the accumulated noise $\hat{n}_e(v, \omega; \kappa)$, so that $\mathcal{H}$ simplifies to

$$\begin{aligned}\mathcal{H} &= \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\hat{n}(v, \omega; \kappa) : (v, \omega) \in \hat{B}] \\ &= \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\hat{\Phi}'_{n_e}(v, \omega; \kappa)} \right] \mathrm{d}v \mathrm{d}\omega \qquad [31]\end{aligned}$$

The theoretical upper bound of $\mathcal{H}$ is Shannon's channel capacity $\mathcal{C}$ for a bandwidth-limited system, with white photodetector noise and an average power limitation (here the variance σ_ρ^2), as given by

$$\mathcal{C} = \frac{1}{2} \log \left[1 + (K\sigma_\rho/\sigma_p)^2 \right] \qquad [32]$$

$\mathcal{H}$ can reach $\mathcal{C}$ only when

$$\hat{\Phi}_\rho(v, \omega) = \begin{cases} \sigma_\rho^2, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases}$$

$$\hat{\tau}(v, \omega) = \begin{cases} 1, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \qquad [33]$$

However, neither of these two conditions can occur in practice because both the PSD $\hat{\Phi}_\rho(v, \omega)$ of natural scenes and the SFR $\hat{\tau}(v, \omega)$ of image-gathering devices decrease gradually with increasing spatial frequency, as depicted in Figures 3 and 6, respectively. Hence, $\mathcal{H}$ can never reach $\mathcal{C}$.

Theoretical Minimum Data Rate $\mathcal{E}$

The theoretical minimum data rate $\mathcal{E}$ (also in bits sample^{-1}) that is required to convey the information rate $\mathcal{H}$ is defined by

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \tilde{s}(v, \omega)] \qquad [34]$$

where the second term, $\mathcal{E}[\cdot|\cdot]$ is the uncertainty of $\tilde{s}_e(v, \omega; \kappa)$, when $\tilde{s}(v, \omega)$ is known. For the assumptions that have been made about the coding process, the conditional entropy becomes the entropy $\mathcal{E}[\tilde{n}_q(v, \omega; \kappa)]$ of the quantization noise, so that $\mathcal{E}$ simplifies to

$$\begin{aligned}\mathcal{E} &= \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{n}_q(v, \omega; \kappa)] \\ &= \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2 + \hat{\Phi}'_{n_e}(v, \omega)}{\tilde{\Phi}'_q(v, \omega; \kappa)} \right] \mathrm{d}v \mathrm{d}\omega \qquad [35]\end{aligned}$$

This expression for $\mathcal{E}$ represents the entropy of completely decorrelated data.

Information Efficiency $\mathcal{H}/\mathcal{E}$

The information efficiency of completely decorrelated data is defined by the ratio $\mathcal{H}/\mathcal{E}$. This ratio reaches its upper bound $\mathcal{H}/\mathcal{E} = 1$ when $\hat{\Phi}'_{n_e}(v, \omega) \ll \hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2$ and $\hat{\Phi}'_{n_e}(v, \omega) \ll \tilde{\Phi}'_q(v, \omega; \kappa)$, because then both eqn [31] for $\mathcal{H}$ and eqn [35] for $\mathcal{E}$ reduce to the same expression

$$\mathcal{H} = \mathcal{E} = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\tilde{\Phi}'_q(v, \omega; \kappa)} \right] \mathrm{d}v \mathrm{d}\omega \qquad [36]$$

This upper bound can be approached with a small loss in $\mathcal{H}$ only when the aliasing and photodetector noise are small. Hence, the electro-optical design that optimizes $\mathcal{H}$ also optimizes $\mathcal{H}/\mathcal{E}$. However, a compromise always remains between $\mathcal{H}$ and $\mathcal{H}/\mathcal{E}$ in the selection of the number of quantization levels: $\mathcal{H}$ favors fine quantization, whereas $\mathcal{H}/\mathcal{E}$ favors coarse quantization.

Maximum-Realizable Fidelity $\mathcal{F}$

The maximum-realizable fidelity $\mathcal{F}$ of the images restored with the Wiener filter $\hat{\Psi}(v, \omega; \kappa)$ can be expressed as a function of the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$ given by eqn [24] as

$$\mathcal{F} = \int \int \hat{\Phi}'_\rho(v, \omega) \hat{\Gamma}_r(v, \omega; \kappa) \, \mathrm{d}v \mathrm{d}\omega \qquad [37]$$

or, equivalently, as a function of the integrand $\hat{\mathcal{H}}(v, \omega; \kappa)$ of the information rate $\mathcal{H}$ given by eqn [31] as

$$\mathcal{F} = \int\int \hat{\Phi}'_\rho(v, \omega)\left[1 - 2^{-\hat{\mathcal{H}}(v,\omega;\kappa)}\right] \mathrm{d}v\mathrm{d}\omega \qquad [38]$$

This relationship between fidelity and information rate can be extended to other goals. For example, the enhancement of spatial features (e.g., edges and boundaries), for robotic vision, can be accommodated by letting $\tilde{\tau}_e(v, \omega)$ be the desired feature-enhancement filter.

Performance and Design Trade-offs

Electro-optical Design

Figure 7 presents the information rate $\mathcal{H}$ (for lossless encoding) as a function of the optical-response index ζ_c and the mean spatial detail μ. The information rate $\mathcal{H}$ versus ζ_c is given for several SNRs $K\sigma_\rho/\sigma_p$ and $\mu = 1$, and $\mathcal{H}$ versus μ is given for the three informationally optimized designs specified in Table 1.

The curves show that $\mathcal{H}$ can reach its highest possible value only when both of the following conditions are met:

1. The relationship between the SFR $\hat{\tau}(v, \omega)$ and the sampling passband $\hat{B}$ is chosen to optimize $\mathcal{H}$ for the available SNR $K\sigma_\rho/\sigma_p$. This design is said to be informationally optimized.
2. The relationship between the sampling passband $\hat{B}$ and the PSD $\hat{\Phi}_\rho(v, \omega)$ is chosen to optimize $\mathcal{H}$. For the PSDs $\hat{\Phi}_\rho(v, \omega)$ typical of natural scenes, the best match occurs when the sampling interval is near the mean spatial detail of the scene (i.e., when $\mu \approx 1$).

These conditions are consistent with those for which the information rate $\mathcal{H}$ reaches Shannon's channel capacity $\mathcal{C}$. However, blurring and aliasing constrain $\mathcal{H}$ to values that are smaller than $\mathcal{C}$ by a factor of nearly two (i.e., $\mathcal{H} \lesssim \mathcal{C}/2$). Hence, these conditions replace the role of white and band-limited signals in classical communication theory, to establish an upper bound on the information rate in imaging systems.

The above two conditions appeal intuitively when the problem of image restoration is considered:

1. The result that the SFR $\hat{\tau}(v, \omega)$ that optimizes $\mathcal{H}$ depends on the available SNR is consistent with the observation that, in one extreme, when the SNR is low, substantial blurring should be avoided because the photodetector noise would constrain the enhancement of fine spatial detail even if aliasing were negligible. In the other extreme, when the SNR is high, substantial aliasing should be avoided so that this enhancement is relieved from any constraints except those that the sampling passband inevitably imposes.

Table 1 Informationally optimized designs

Design	$K\sigma_\rho/\sigma_p$	ζ_c	$\mathcal{H}^*$
1	256	0.3	4.4
2	64	0.4	3.3
3	16	0.5	2.2

$^*\mu = 1$.

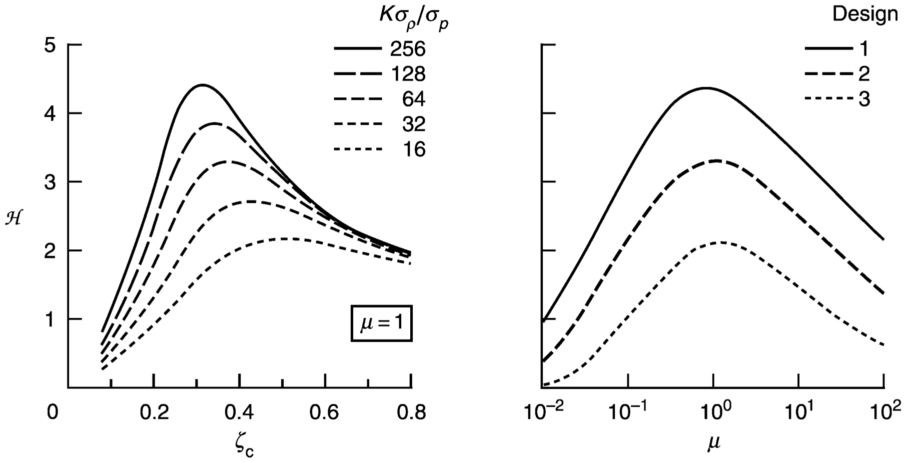

Figure 7 Information rate $\mathcal{H}$ versus the optical-response index ζ_c and the mean spatial detail μ (relative to the sampling interval). The designs are specified in Table 1, and the SFRs $\hat{\tau}(v, \omega)$ for ζ_c are shown in Figure 6.

2. The result that $\mathcal{H}$ reaches its highest value when the sampling interval is near the mean spatial detail of the scene is consistent with the observation that, ordinarily, it would not be possible to restore spatial detail that is finer than the sampling interval, whereas it would be possible to restore much coarser detail from fewer samples.

Information Rate and Data Rate

Figure 8 presents an information-entropy $\mathcal{H}(\mathcal{E})$ plot that characterizes the relationship between the information rate $\mathcal{H}$ of the encoded signal and the associated theoretical minimum data rate, or entropy, $\mathcal{E}$ for η-bit quantization. The encoder SFR $\hat{\tau}_e(v, \omega)$ is assumed to be unity. The plot illustrates the trade-off between $\mathcal{H}$ and $\mathcal{E}$ in the selection of the number of quantization levels for encoding the acquired signal. It shows, in particular, that the design that realizes the highest information rate $\mathcal{H}$ also enables the encoder to realize the highest information efficiency $\mathcal{H}/\mathcal{E}$. This result appeals intuitively, because the perturbations due to aliasing and photodetector noise that interfere with the image restoration can also be expected to do so with the signal decorrelation.

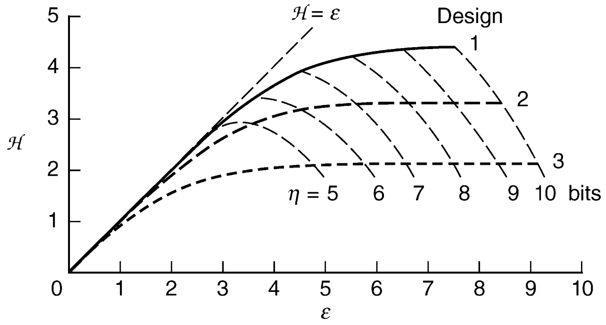

Figure 8 The information-entropy $\mathcal{H}(\mathcal{E})$ plot of the information rate $\mathcal{H}$ versus the associated theoretical minimum data rate $\mathcal{E}$ for η-bit quantization. The designs are specified in Table 1.

Throughput SFR

Figure 9 illustrates the SFRs $\hat{\tau}(v, \omega)$ and $\hat{\Psi}(v, \omega; \kappa)$ of the image-gathering device and Wiener filter, respectively, and of their product, the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$, for the three designs specified in Table 1. As the accumulated noise becomes negligible, $\hat{\Gamma}_r(v, \omega; \kappa)$ approaches the ideal SFR of the classical communication channel, which is unity inside the sampling passband and zero outside.

Information Rate, Fidelity and Robustness

The PSD $\hat{\Phi}_\rho(v, \omega)$ of a scene is seldom known when images are restored. Moreover, even if the PSD were known to be the one given by eqn [2], then the mean

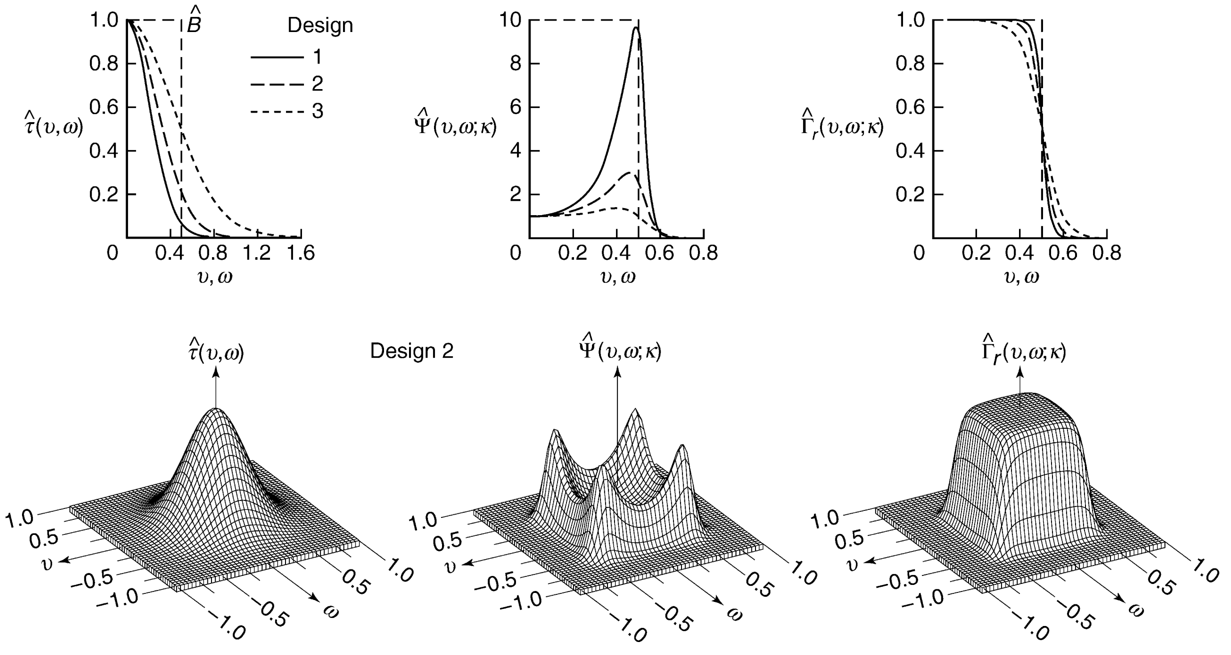

(a) Image gathering (b) Wiener filter (c) Throughput

Figure 9 SFRs of image gathering, restoration, and throughput. The designs are specified in Table 1.

spatial detail μ would usually remain uncertain because it depends, not only on the properties of the scene, but also on the angular resolution and viewing distance of the image-gathering device. In practice, therefore, image restorations have to rely on estimates of the statistical properties of scenes. The tolerance of the quality of the restored image to errors in these estimates is referred to as the robustness of the image restoration.

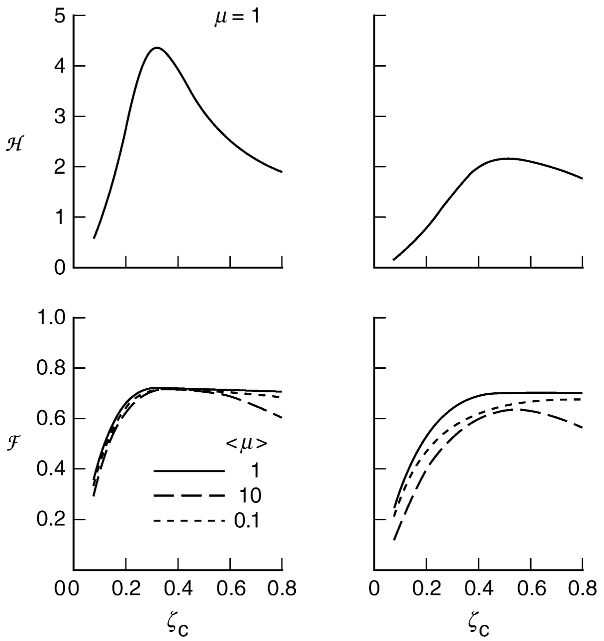

(a) $K\sigma_\rho/\sigma_p = 256$ (b) $K\sigma_\rho/\sigma_p = 16$

Figure 10 Information rate $\mathcal{H}$ and maximum-realizable fidelity $\mathcal{F}$ versus the optical-response index ζ_c for two SNRs $K\sigma_\rho/\sigma_p$. $\mathcal{F}$ is given for the matched ($\langle\mu\rangle = 1$) and two mismatched Wiener restorations.

Figure 10 presents the information rate $\mathcal{H}$ and the corresponding maximum-realizable fidelity $\mathcal{F}$ for matched and mismatched Wiener restorations. These restorations represent the digital image $R(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \kappa)$ or, equivalently, the continuous image $R(x, y; \kappa)$ displayed on a noise-free medium. The matched restorations use the correct value of μ, whereas the mismatched restorations use wrong estimates of μ. These curves reveal that the informationally optimized designs produce the highest fidelity and robustness, and that both improve with increasing $\mathcal{H}$. Moreover, these curves reveal that, ordinarily, one cannot go far wrong in Wiener restorations by assuming that the mean spatial detail is equal to the sampling interval (i.e., $\mu = 1$). This observation appeals intuitively because spatial detail that is much smaller cannot be resolved, and detail that is larger is not degraded substantially by blurring and aliasing.

Information Rate and Visual Quality

Figures 11 and 12 present Wiener restorations and their three components, as given by eqn [23], where $\rho(x, y) * \Gamma_r(x, y; \kappa)$ accounts for blurring, $n_{ar}(x, y; \kappa)$ accounts for aliasing, and $n_{pr}(x, y; \kappa)$ accounts for

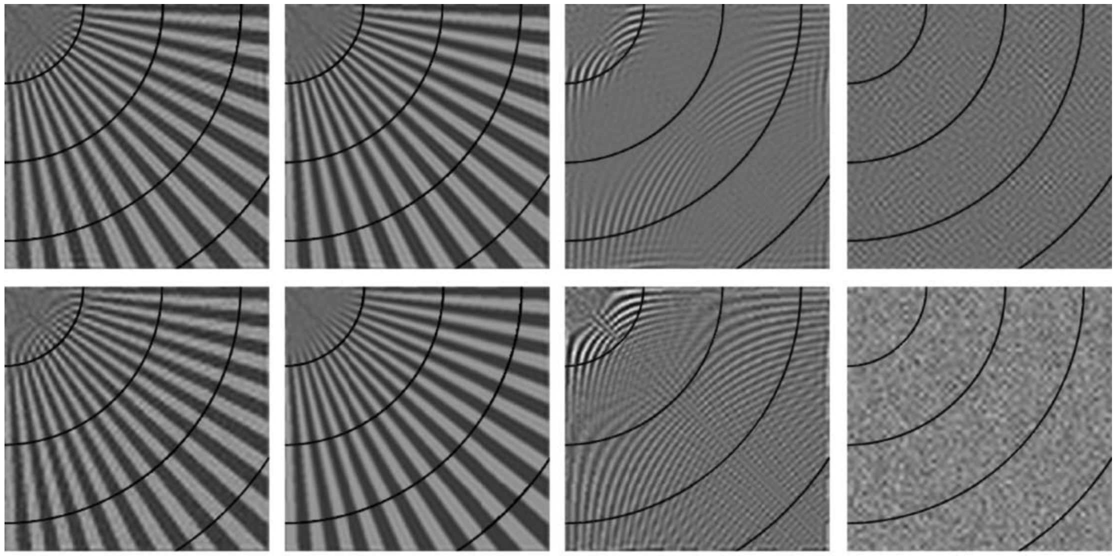

(a) $R(x,y;\kappa)$ (b) $\rho(x,y) * \Gamma_r(x,y;\kappa)$ (c) $4n_{ar}(x,y;\kappa)$ (d) $4n_{pr}(x,y;\kappa)$

Figure 11 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

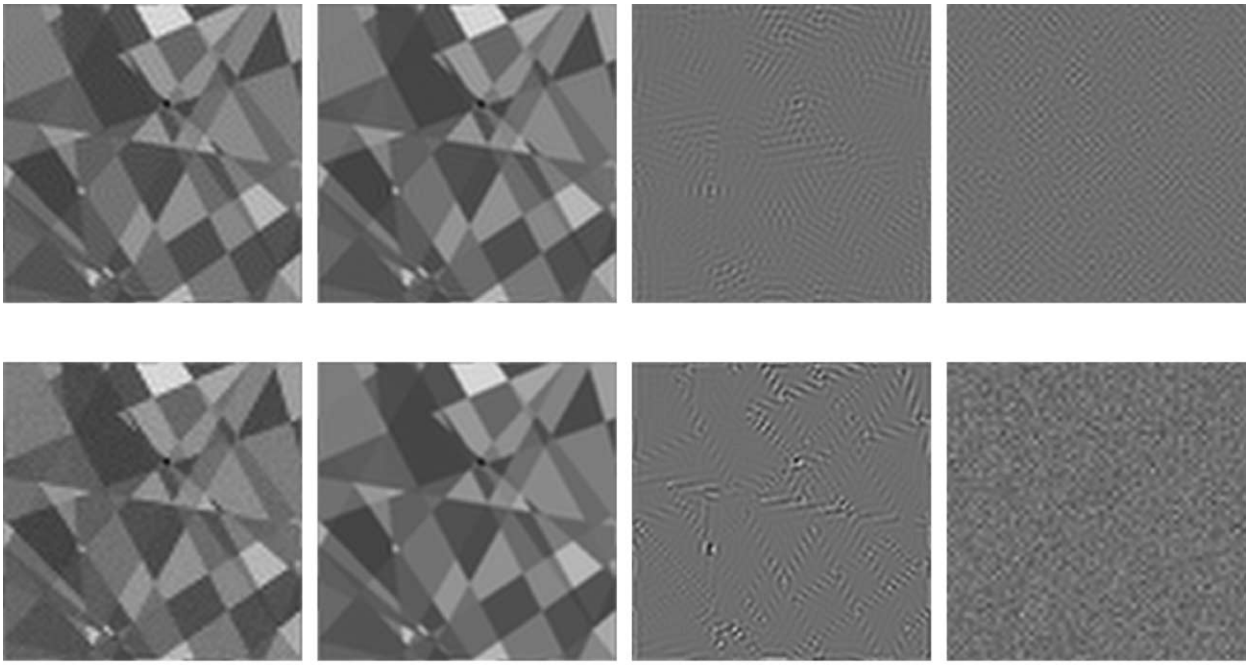

(a) $R(x,y;\kappa)$ (b) $\rho(x,y)*\Gamma_r(x,y;\kappa)$ (c) $4n_{ar}(x,y;\kappa)$ (d) $4n_{pr}(x,y;\kappa)$

Figure 12 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

photodetector noise. The contrast of the illustrations of the two noise components has been increased by a factor of four to improve their visibility. The extraneous structure that aliasing produces in the images of the resolution wedges is commonly referred to as Moiré pattern. The corresponding distortion in the images of the random polygons emerges more subtly as jagged (or staircase) edges.

The images are formed from a set of 64 × 64 samples with a $Z = 4$ interpolation and have a 4 × 4 cm format. If these images were displayed in a smaller format, then the jagged edges could not be resolved by the observer and would appear to be the result of blurring or photodetector noise instead of aliasing. Aliasing has, therefore, often been overlooked as a significant source of image degradation.

The Wiener restoration produces images in which fine spatial detail near the limit of resolution normally has a high contrast approaching that in the scene. However, these images also tend to contain visually annoying defects, due to the accumulated noise and the ringing near sharp edges (Gibbs phenomenon). Hence, it is often desirable to combine this restoration with an enhancement filter that gives the user some control over the trade-off among fidelity, sharpness and clarity.

Figure 13 presents images for which the Wiener restorations given in Figures 11 and 12 are enhanced for visual quality by suppressing some of the defects due to aliasing and ringing. This enhancement improves clarity at the cost of a small loss in sharpness. The images are produced again from 64 × 64 pixels, but they are displayed in three different formats. The smallest format, with a density of 64 pixels/cm, corresponds to a display of 256 × 256 pixels in a 4 × 4 cm area, which is typical of the images found in the prevalent digital image-processing literature. As can be observed, this small format leads to images that are indistinguishable from each other because they hide many of the distortions that larger formats clearly reveal.

In general, if the images are displayed in a format that is large enough to reveal the finest spatial detail near the limit of resolution of the imaging system, then distortions due to the perturbations in this system become also visible. When no effort is spared to reduce these distortions without a perceptual loss of resolution and sharpness, then it becomes strongly evident that the best visual quality with which images can be restored improves with increasing information rate, even after the fidelity has essentially reached its upper bound. This improvement continues until it is gradually ended by the unavoidable compromise among sharpness, aliasing, and ringing as well as by the granularity of the image display.

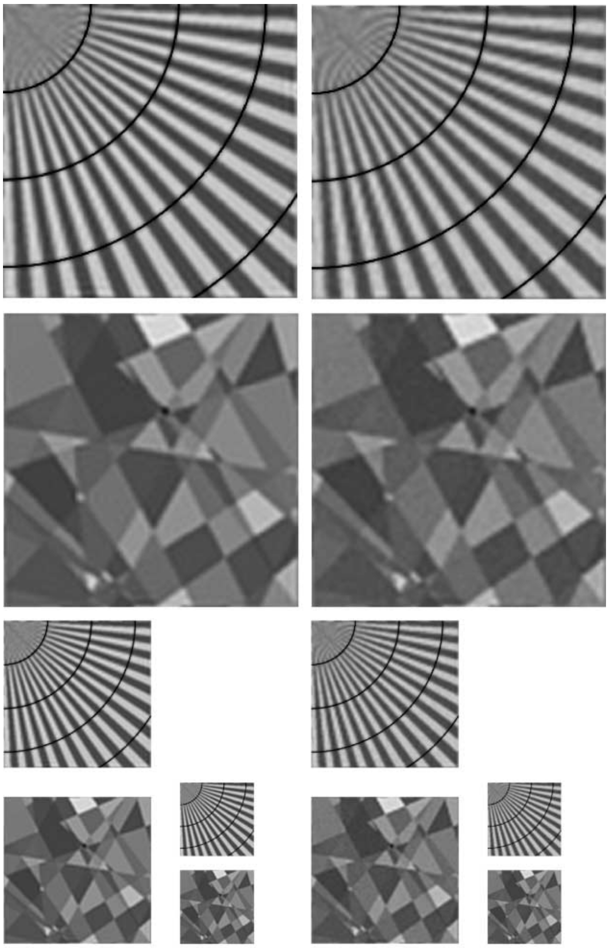
(a) Design 1
(b) Design 3

Figure 13 Images restored with the Wiener filter and enhanced for visual quality. The large, medium, and small formats contain 16, 32 and 64 pixels/cm, respectively.

Concluding Remarks

The performance of image-gathering devices and the human eye is constrained by the same critical limiting factors. When these factors are accounted for properly, then it emerges that the design of the image-gathering device that is optimized for the highest realizable information rate corresponds closely, under appropriate conditions, to the design of the human eye that evolution has optimized for viewing the real world. This convergence of information theory and evolution toward the same design clearly supports the extension of information theory to the characterization of imaging systems.

Numerous other information-theoretic characterizations of imaging are now emerging. However, none of these characterizations accounts for the critical limiting factors that constrain image gathering and restoration, as outlined here. Hence, it is not possible for them to address the efficiency and accuracy with which images can be conveyed. Moreover, none of these characterizations combines observation with communication, or sensing with processing, in a general formalization. Instead, each of them still deals with a particular measurement or imaging problem. Hence, the role of information theory in imaging is far from mature. Many challenges remain: to develop general formalizations, to account for physical phenomena, and to correlate predictions with measurements.

See also

Information Processing: Optical Digital Image Processing. **Modulators:** Electro-optics.

Further Reading

Fales CL, Huck FO, Alter-Gartenberg R and Rahman Z (1996) Image gathering and digital restoration. *Philosophical Transactions of the Royal Society of London* A354: 2249–2287.

Huck FO and Fales CL (2001) Characterization of image systems. *Encyclopedia of Imaging Science and Technology*. New York: John Wiley & Sons.

Huck FO, Fales CL and Rahman Z (1996) An information theory of visual communication. *Philosophical Transactions of the Royal Society of London* A354: 2193–2248.

Huck FO, Fales CL and Rahman Z (1997) *Visual Communication: An Information Theory Approach*. Boston, CA: Kluwer Academic Publishers.

Lesurf JCG (1995) *Information and Measurement*. Bristol, UK: Institute of Physics Pub.

O'Sullivan JA, Blahut RE and Snyder DL (1998) Information-theoretic image formation. *IEEE Transactions Information Theory* 44: 2094–2123.

Shannon CE (1948) A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 28: 623–656.

Shannon CE and Weaver W (1964) *The Mathematical Theory of Communication*. Urbana, IL: U. Illinois Press.

Wiener N (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons.

Inverse Problems and Computational Imaging

M Bertero and P Boccacci, University of Genova, Genova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A digital image is an array (2D image) or a cube (3D image) of quantized numbers, providing a representation of a physical object or a map of some distributed property of the object (density of matter, energy, etc.). It can be obtained by digitizing the original image (for instance, a photograph) or can be directly formed by the outputs of the imaging system (digital camera, microscope, telescope, etc.). When visualized on a suitable image display device, a digital image must provide a representation of the physical object which can be easily interpreted and analyzed. In many cases, however, this condition is not satisfied. We give two typical examples.

The first is that of an image degraded by blurring which can be introduced by defects of the optical systems (such as aberrations) or by image motion, image defocusing, etc. The representation of the object does not have the quality which should be desirable for its interpretation.

The second case is even more important and arises when the output of the imaging system is not directly related to the physical quantity to be imaged. For instance, in X-ray tomography, the output of a detector is the attenuation of an X-ray pencil crossing the body while the quantity to be imaged is the density of the body.

In the situations described above, the desired image can be obtained by solving a number of mathematical problems. If we denote f as the object, namely the target of the imaging system, and g as the image of f provided by the system, then the problems to be solved can be summarized as follows:

- develop a physical model of the system to obtain a mathematical relationship between the object f and the image g; the computation of g from a given f is usually called the solution of the direct problem;
- solve the problem of obtaining the physical quantity f from given outputs g of the system; this second step is usually called the solution of the inverse problem.

The digital image of f, obtained by solving an inverse problem, is a computed one, generated by the computer where the algorithm for the solution of the inverse problem has been implemented.

It may be convenient to formulate both the direct and the inverse problem in terms of functions rather than arrays, cubes, etc., because such a formulation allows the use of powerful tools of functional analysis for their investigation. Next, the results obtained in the continuous case can be used for understanding the features of the corresponding discrete problem. In most important applications, for instance, the relationship between f and g is linear so that, in the discrete version of the direct problem, one obtains g by applying a suitable matrix A to f while, in the discrete version of the inverse problem, one obtains f from g by solving a linear algebraic system. However, the solution of the inverse problem is not so simple and the difficulties can be understood by looking at the corresponding continuous problem.

The basic reason relies on the fact that inverse problems are ill-posed in the sense of Hadamard: the solution may not exist; even if it exists it may not be unique; and, even if it exists and is unique, it may not depend continuously on the data. The last statement means that a small perturbation of the data, such as that caused by noise or any kind of experimental error, can completely modify the solution.

Inverse problems are ill-posed because imaging systems do not transmit complete information about the physical object. Therefore, the problem is to extract the useful information contained in the data or, in mathematical terms, to look for approximate solutions which are stable against noise. This is the purpose of a mathematical theory introduced in its general form by the Russian mathematician, Tikhonov and known as regularization theory.

We illustrate the general framework outlined above by means of a specific example – image deconvolution, also known as image deblurring, image restoration, etc. The object f is the image which should be recorded in the absence of degradation, while g is the image corrupted by aberrations or other causes of blurring. Both f and g are functions of 2D (or 3D) space variables $\mathbf{x}$ which are the coordinates of a point in the image domain. If the imaging system is isoplanatic, then it is described by a space-invariant point spread function (PSF) $K(\mathbf{x})$, which is the image of a point source located in the center of the image domain. The PSF provides the response of the system to any point source wherever it is located. On the other hand, its Fourier transform, the transfer function $\hat{K}(\omega)$ (the hat denotes Fourier transform and ω, the spatial frequencies associated to the space variables $\mathbf{x}$), tells us how a signal of a fixed frequency is propagated through the linear system, so that the blurring can also be viewed as a sort of frequency filtering.

Under these assumptions the image g is the convolution product of object f and PSF K:

$$g(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')\mathrm{d}\mathbf{x}' \qquad [1]$$

and the solution of the direct problem is just the computation of a convolution product.

The image g, as given by eqn [1], is the so-called noise-free image. The actually recorded image g_r is affected by a noise term and, in general, it can be written in the following form:

$$g_r(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \qquad [2]$$

where $n(\mathbf{x})$ is a function describing noise contamination. It is a random function and, therefore, it is unknown even if one understands its statistical properties. We point out that eqn [2] does not imply that the noise is additive or signal independent. For instance, in microscopy and astronomy, the contamination of the image is due both to photon noise, which satisfies Poisson statistics, and read-out noise, which is basically Gaussian and white. Therefore the noise term in eqn [2] must only be intended as the difference between the noisy and the noise-free image.

In a first attempt at approaching the inverse problem, it seems quite natural to ignore the noise term $n(\mathbf{x})$ and to solve eqn [1] for $f(\mathbf{x})$ with $g(\mathbf{x})$ replaced by $g_r(\mathbf{x})$. Then, by taking the Fourier transform of both sides of this equation and using the well-known convolution theorem, one finds the following relationship:

$$\hat{g}_r(\omega) = \hat{K}(\omega)\hat{f}(\omega) \qquad [3]$$

which relates the Fourier transform of f and g to the transfer function of the imaging system.

The solution of this equation looks elementary. However, a first problem is due to the fact that an optical system is generally band-limited; its band Ω is the bounded domain of spatial frequencies where the transfer function is different from zero. Since the noise is not band-limited or, at least, has a band much broader than that of the optical system, outside Ω the right-hand side of eqn [3] is zero while the left-hand side is not; in other words, the solution of the deconvolution problem, as formulated above, does not exist because no object f satisfies eqn [3] everywhere. In this particular case, one can circumvent the problem by applying a suitable filter to $g_r(\mathbf{x})$, in order to suppress the out-of-band noise.

The second problem is the nonuniqueness of the solution of the problem with the filtered data. This is due to the so-called invisible objects, namely objects whose Fourier transform is zero on Ω, so that the corresponding images are exactly zero. If we find a solution of the deconvolution problem, by adding an arbitrary invisible object to this solution, we find another solution of the same problem. A well-defined solution can be obtained by requiring its Fourier transform to be zero outside Ω. In other words, we set to zero what is not transmitted by the imaging system.

The standard way for approaching the previous questions is to look for solutions in the least-squares sense, namely for objects which minimize the functional:

$$\varepsilon^2(f) = \|K * f - g_r\|^2 \qquad [4]$$

where $*$ denotes the convolution product, as defined in eqn [1], and the right-hand side is the square of the L^2-norm (which can also be interpreted as energy) of the discrepancy between the recorded image g_r and the computed image $K * f$. If one looks for a least-squares solution with minimal energy then one automatically re-obtains the solution discussed above, namely an object whose Fourier transform in Ω is given by

$$\hat{f}_{\mathrm{inv}}(\omega) = \frac{\hat{g}_r(\omega)}{\hat{K}(\omega)} \qquad [5]$$

and is zero outside Ω. Such a solution is that provided by the so-called inverse filter.

However, there is an additional difficulty in that the transfer function usually tends to zero at the boundary of the band while the Fourier transform of the noise, hence that of g_r, does not, or tends to zero in a different way. It follows that $\hat{f}_{\mathrm{inv}}(\omega)$ is divergent or very large at the boundary of the band. We conclude that its inverse Fourier transform does not exist or, if it exists, is affected by large artifacts due to noise propagation. The latter is the typical situation in the case of digital images. In Figure 1 we give the result of a numerical simulation showing the typical effect produced by the inverse filter. The object is a beautiful picture of a galaxy recorded by the Hubble Space Telescope (HST), while the PSF is a computed one describing the blurring of HST before installation of the optics correction. The blurred image is obtained by convolving the object with the PSF and by adding white noise. The restored image produced by the inverse filter is completely corrupted by noise.

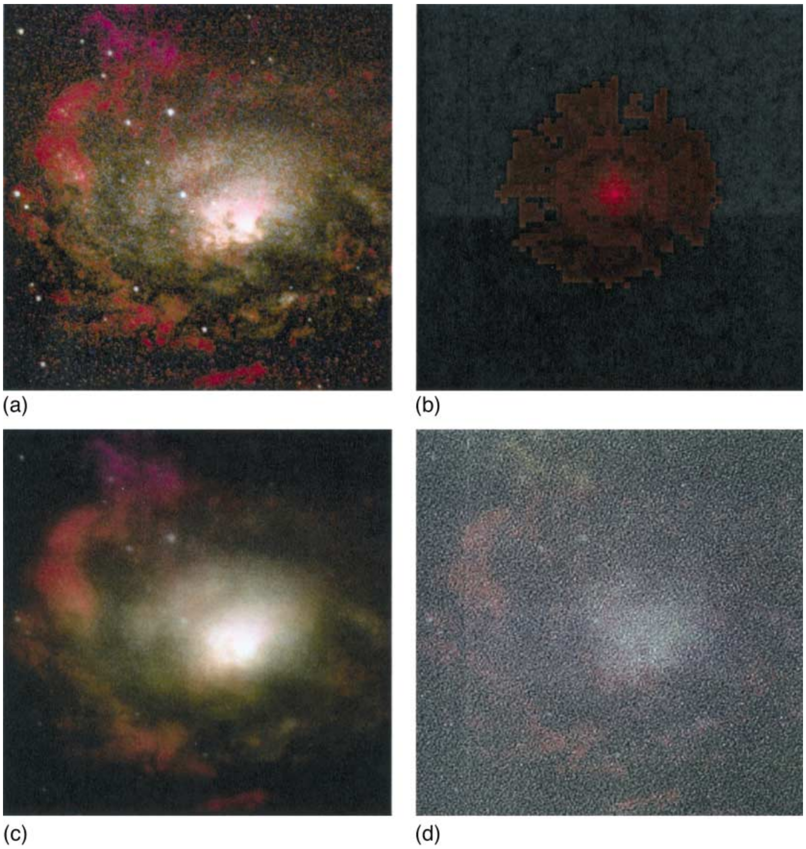

Figure 1 Illustrating the effect of the inverse filter: (a) Image of the Circinus galaxy taken on 10 April 1999, with the Wide Field Planetary Camera 2 of the Hubble Space Telescope; (b) the PSF used for blurring the RGB components of the image (here shown in red); (c) the blurred image obtained by convolving the components of the image in (a) with the corresponding PSFs and adding noise; (d) the restoration obtained by applying the inverse filter to the three components of (c). Part (a) courtesy of NASA/Space Telescope Science Institute.

The inverse filter provides a solution which fits the data in the best possible way and therefore also fits the noise in the frequency domains where the signal is weakly transmitted by the imaging system. This can be achieved at the cost of an energy (L^2-norm) of the solution which is too large. Therefore, such a conclusion suggests that one must search for a reasonable compromise between data fitting and energy of the solution. This is the basic idea of regularization theory, at least in its most simple form. The mathematical formulation is obtained by looking for objects which minimize the functional:

$$\Phi_\mu(f) = \|K * f - g_r\|^2 + \mu\|f\|^2 \qquad [6]$$

where μ, the so-called regularization parameter, is a parameter controlling the trade-off between data fitting and energy of the solution: when μ is small, the data fitting is good and the energy is large while, when μ is large, the data fitting is poor and the energy is small.

For a given μ, the function f_μ, which minimizes the functional of eqn [6] is called the regularized solution of the problem. It is easy to show that its Fourier transform is given by

$$\hat{f}_\mu(\omega) = \frac{\hat{K}^*(\omega)}{|\hat{K}(\omega)|^2 + \mu}\hat{g}_r(\omega) \qquad [7]$$

The regularized solutions form a one-parameter family of functions. When μ is small these functions are strongly contaminated by noise whose effect is gradually reduced when μ increases. For too large values of μ, one re-obtains blurred versions of the original object.

Such behavior is illustrated by the sequence of regularized solutions given in Figure 2 and corresponding to the example of Figure 1. The regularization parameter increases starting from left to right and from top to bottom. Its values are $\mu = 0$ (inverse filter), 0.001, 0.01, 0.05, 0.1, and 0.5 (the noise is of the order of a few percent and the PSF is normalized in such a way that the sum of all its value is 1). The sequence makes evident the well-established theoretical result of the existence of an optimal value of the regularization parameter. Such an optimal value can

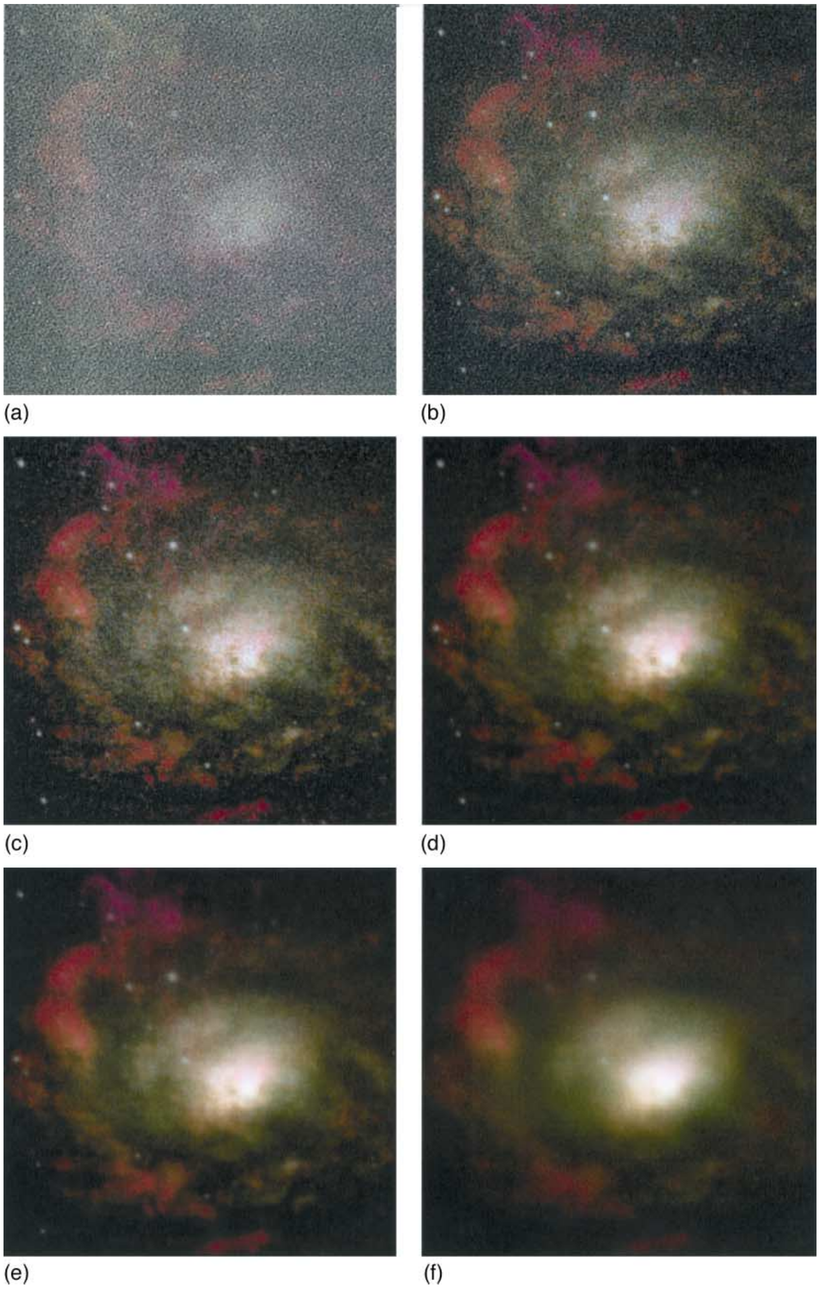
(a) (b) (c) (d) (e) (f)

Figure 2 Illustrating the effect of the regularization parameter. The blurred image is that shown in **Figure 1c**. The sequence is obtained by increasing the value of the regularization parameter: (a) $\mu = 0$ (inverse filter); (b) $\mu = 0.001$; (c) $\mu = 0.01$; (d) $\mu = 0.05$; (e) $\mu = 0.1$; (f) $\mu = 0.5$. It is evident that the best restoration is that shown in (c).

be determined in the case of numerical simulations as that represented in Figure 2. However, its estimation in the case of real images is more difficult. Several methods for the selection of μ have been proposed. We only mention the so-called discrepancy principle. It consists of determining the value of μ, such that the value of the discrepancy functional at f_μ, $\varepsilon^2(f_\mu)$ coincides with an estimate of the energy of the noise. In other words, the data are fitted with an accuracy comparable with their uncertainty.

As shown by eqn [7], the regularized solution has the same band Ω of the imaging system, because $\hat{K}^*(\omega)$ is zero outside Ω. As is known, a band-limited function can be represented in terms of its samples taken at a rate which is roughly proportional to the size of the band. This is the content of the famous Shannon sampling theorem (more precisely its 2D extension). On the other hand, the sampling distance can be taken as a measure of the resolution provided by the restored image so that one concludes that the restoration method discussed above does not produce an improvement of the resolution of the imaging system. It certainly produces an improvement of the quality of the image which makes possible a visual verification of that resolution.

In many circumstances, the term super-resolution is used for describing methods which allow an improvement of resolution beyond the limit provided by the sampling theorem. Therefore, a super-resolving method must produce a restored image with a band broader than that of the imaging system. Therefore, super-resolution is related to the problem of out-of-band extrapolation.

Such an approach was already proposed in the 1960s, when it was observed that the Fourier transform of an object with a finite spatial extent (all objects have this property) is analytic; then the theorem of unique analytic continuation implies that the Fourier transform of the object can be determined everywhere from its values on the band of the imaging system.

Unfortunately this problem is ill-posed, so that out-of-band extrapolation is not feasible in practice. Only recently it has been recognized that a considerable amount of super-resolution is possible when the spatial extent of the object is not much greater than the resolution distance of the instrument. For example, super-resolving methods could be used for detecting unresolved binary star in astronomical images.

The important point is that a super-resolving method must implement explicitly the finite extent of the object: the domain of the object must be estimated and the method must search for a solution which is zero outside this domain. This introduces an important question in object restoration and, more generally, in the theory of inverse problems, namely the design of regularized solutions satisfying additional conditions (constraints). This question is also important from the theoretical point of view: the problem is ill-posed because of insufficient information on the object as transmitted by the imaging system; the use of additional constraints reduces the class of the solutions which are compatible with the data and therefore can improve the restoration. This is the use of a priori information in the solution of inverse problems.

A constraint which has been widely investigated and used is the positivity of the solution, i.e., all the values of the restored image must be positive or zero. The physical meaning of the constraint is obvious. Its beneficial effect is to reduce ringing artifacts which affect the regularized solutions previously discussed, in cases where the object contains bright spots over a black background or sharp intensity variations from zero to some positive value. The ringing is essentially related to the well-known Gibbs effect in the truncation of the Fourier series.

The requirement of positivity is the requirement of a particular lower bound on the values of the solution. Therefore, one can consider more general constraints in requiring a given lower and/or upper bound on the values of the solution, eventually combined with a constraint on the domain to produce, for instance, a super-resolved positive solution. The most general form of these constraints requires that the solution belongs to a given closed and convex set $\mathcal{C}$ in some functional space. A general method producing regularized solutions in a convex set is an iterative method which is essentially a gradient method for the minimization of the least-squares functional, with a projection on the set $\mathcal{C}$ at each iteration. In the case of object deconvolution, the least-squares functional is that given in eqn [4] and the method has the following form: if f_k is the result of the k-th iteration, then f_{k+1} is given by

$$f_{k+1} = P_{\mathcal{C}}[f_k + \tau K^{\mathrm{T}} * (g_r - K * f_k)] \qquad [8]$$

where $P_{\mathcal{C}}$ is the convex projection onto the set $\mathcal{C}$, τ is a relaxation parameter and $K^{\mathrm{T}}(\mathbf{x}) = K(-x)$. In general, the algorithm, known as the projected Landweber method, is initialized with $f_0 = 0$. An important property is that it has a regularization effect, in the sense that iterations must not be pushed to convergence but stopped after a suitable number of iterations to avoid strong noise contamination. In Figure 3 we compare, in a specific example, the result obtained by means of the linear regularization method and that obtained by means of the projected iterative method with a lower and upper bound on the solution. The reduction of ringing is evident.

Among the methods producing positive solutions we must mention the most popular one in astronomy, the so-called Richardson–Lucy method, which is an iterative method for Maximum-Likelihood estimation. If the PSF has been normalized in such a way that the sum of all its values is one, it takes the following form:

$$f_{k+1} = f_k\left(K^{\mathrm{T}} * \frac{g_r}{K * f_k}\right) \qquad [9]$$

The method is normally initialized with a uniform image and it must not be pushed to convergence to avoid noise amplification. Another algorithm suggested by statistical arguments and producing positive solution is the so-called Maximum Entropy Method (MEM).

The image restoration methods described above have wide applications both in microscopy and

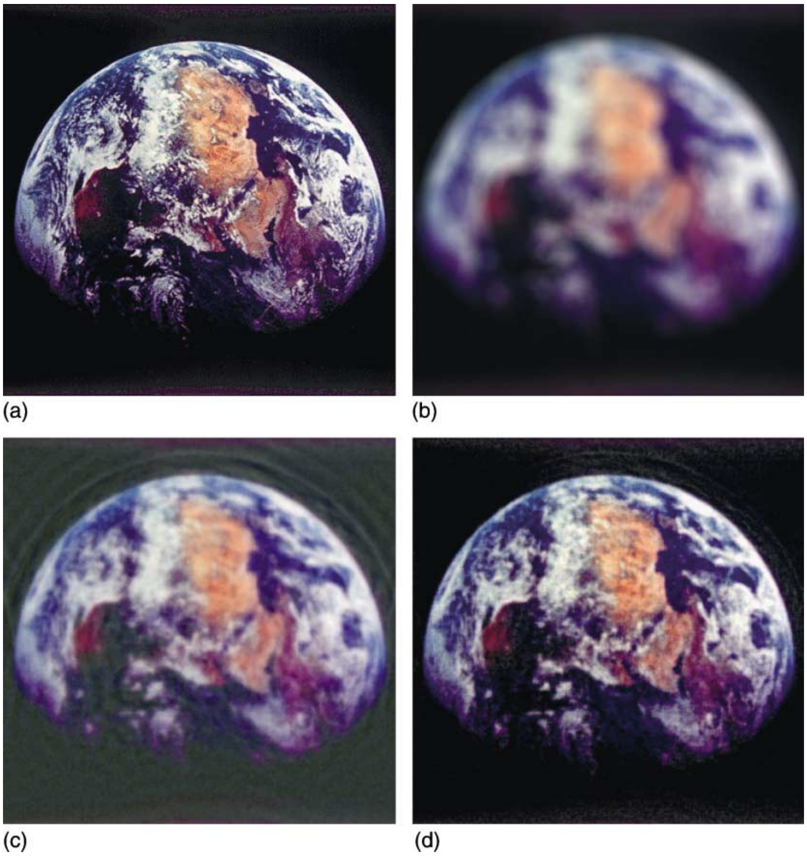

Figure 3 Illustrating the effect of constraints on image restoration: (a) Image of the Earth taken during the Apollo 11 mission; (b) blurred image of (a) obtained by assuming out-of-focus blur and adding noise; (c) optimal restoration provided by the regularization method; the ringing around the boundary of the Earth is evident; (d) optimal restoration obtained by means of the projected Landweber method with a lower and an upper bound on the solution at each iteration. Part (a) courtesy of NASA.

astronomy. For instance, the minimization of the functional eqn [6], with the additional constraint of positivity, has been used for deconvolving images in microscopy to reduce the effect of the missing cone. On the other hand, the Richardson–Lucy method was widely used for deconvolving the images of HST before installation of corrective optics.

Nowadays, image deconvolution is becoming more important for the ground-based telescopes equipped with adaptive optics. Indeed, even if adaptive optics is able to provide a considerable compensation of the atmospheric blur, a further improvement can be obtained by deconvolving the detected images. The PSF is provided by the image of a suitable guide star.

Finally, it is worth mentioning the Large Binocular Telescope, in construction on the Mount Graham in Arizona, because this instrument requires image restoration methods for a full exploitation of its imaging properties. The telescope (Figure 4) consists of two 8 m mirrors on a common mount: both mirrors are equipped with adaptive optics and are combined interferometrically to reach the resolution

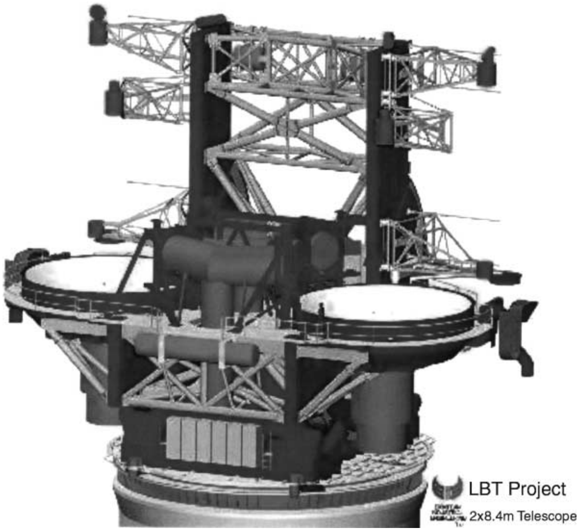

Figure 4 Picture of the Large Binocular Telescope (LBT) in construction on the Mount Graham in Arizona. It consists of two mirrors on the same mount. The images of the two mirrors are combined interferometrically to produce a high-resolution image in the direction of the line joining the centers of the mirrors (baseline).

of a 22.8 m mirror in the direction of the baseline. The telescope can be rotated to record images of the same astronomical target with different orientations of the baseline, i.e., the line joining the centers of the two mirrors. Finally the images must be processed by means of multiple images deconvolution methods to obtain a unique high-resolution image, equivalent to that produced by a 22.8 m mirror.

The applications of inverse problems with major social impact are in the area of medical imaging. The spectacular success of X-ray computerized tomography (CT), invented by GH Hounsfield at the beginning of the 1970s, has stimulated the development of other imaging modalities, based essentially on the same principle, such as positron emission tomography (PET) and single photon emission computerized tomography (SPECT). Computed images are also provided by magnetic resonance imaging (MRI). In such a case, however, the computational problem is rather simple since it consists essentially in Fourier transform inversion.

The basic principle of X-ray CT is described as a finely collimated source S (see Figure 5a) emitting a pencil of X-rays which propagates through the body along a straight line L up to a well collimated receiver R. If we know both the intensity I_0 of the source and the intensity I detected by the receiver, the logarithm of the ratio I_0/I is the integral of the linear attenuation function (roughly proportional to the density function of the body) along the line L. If the source and the receiver are moved along two parallel lines, having direction θ and defining the plane Π (in practice a planar slice of the body under consideration), one obtains the integrals of the linear attenuation function $f(\mathbf{x})$ along all parallel lines orthogonal to θ. The scanning variables in the plane Π are defined in Figure 5b: φ is the angle formed by the unit vector θ with the x-axis and s is the signed distance between the origin and the integration line L. The integrals along all parallel lines orthogonal to θ define a function of s which is called the projection of $f(\mathbf{x})$ in the direction θ:

$$P_\theta(s) = \int_{-\infty}^{+\infty} f(s\cos\varphi - u\sin\varphi, s\sin\varphi + u\cos\varphi)\mathrm{d}u \qquad [10]$$

The set of the projections of $f(\mathbf{x})$ for all possible θ is called the Radon transform of $f(\mathbf{x})$, in honor of the mathematician Johann Radon, who first investigated the problem of recovering a function of two variables from its line integrals. Today, such a problem is known as Radon transform inversion or object restoration from projections. Sampled values of the Radon transform, i.e., sampled values of the projections for a number of possible directions between 0 and π, are the outputs of the acquisition system of the medical equipment known as CT-scanners (Figure 6). Their data-processing system contains the implementation of an algorithm for Radon transform inversion and the final result is a set of computed images providing maps of slices of the human body.

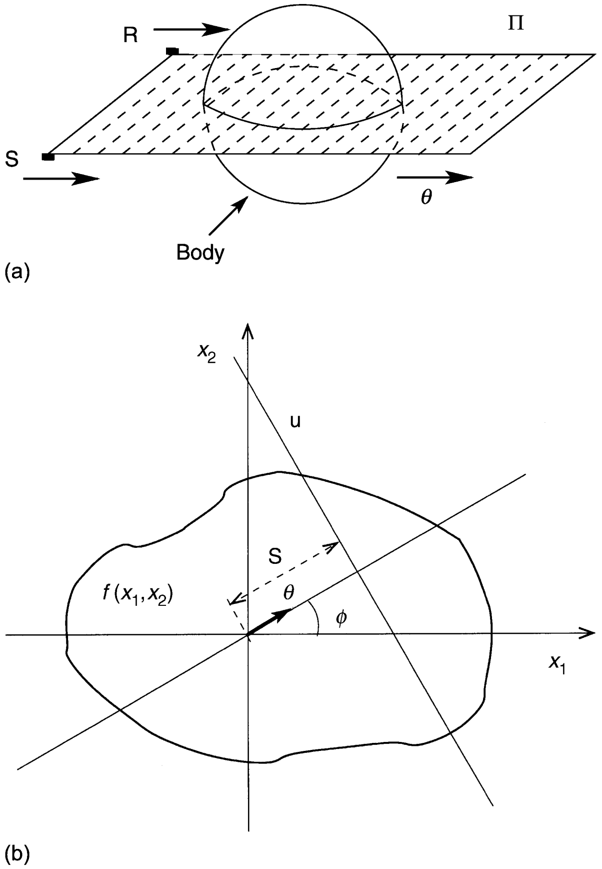

Figure 5 (a) Scheme of the scanning procedure in X-ray tomography: the source(S)–receiver(R) pair is moved along a direction θ orthogonal to the line S–R, defining the plane Π to be imaged; when the scanning has been completed the system is rotated by a certain angle, the scanning procedure is repeated, and so on. (Reproduced with permission from Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging.* Bristol, UK: IOP Publishing.) (b) Definition of the variables used in the representation of the Radon transform; φ is the angle formed by the S–R line with the x_1-axis; s is the signed distance between the origin and the integration line (orthogonal to θ); and u is the coordinate on the integration line.

The plot of the Radon transform of f in the (s, φ)-plane, obtained by representing its values as gray levels, is called the sinogram of f. In Figure 7, we give the image of the sinogram of a brain slice. The horizontal rows provide representations of the projections of the phantom. In the image, the angle φ defining the direction θ takes values between 0 and 2π

while the variable s takes values between $-a$ and a, if a is the radius of a disc containing the phantom. As is evident, the sinogram is symmetric with respect to the point $(0, \pi)$. Each point in the rectangle corresponds to a straight line crossing the phantom and the straight lines through a fixed point of the phantom describe a sinusoidal curve in the (s, φ)-plane. This property has given rise to the name sinogram.

The sinogram is, in a sense, the image of the body slice under investigation, as provided directly by a CT-scanner. However, an inspection of this image does not easily provide information about the body. Therefore, it must be processed to obtain a more significant image. Radon transform inversion is an example of inverse and ill-posed problem and its solution must be treated with extra care.

The crucial point in Radon transform inversion is the Fourier slice theorem, whose content is the following: the Fourier transform of the projection in the direction θ of the function f, is the Fourier transform of f on the straight line passing through the origin and having direction θ:

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\theta) \qquad [11]$$

This theorem clarifies the information content of CT data: each projection provides the Fourier transform of the function f along a well-defined straight line in the plane of the spatial frequencies. It is also the starting point for deriving the basic algorithm of data inversion in tomography, the filtered back-projection (FBP). It is obtained from Fourier transform inversion in polar coordinates and is a two-step algorithm: the first step is a filtering of the projections by means of a ramp-filter; the second is the back-projection of the filtered sinogram. More precisely the two steps are as follows:

- *Filtering*: for each θ the projection $P_\theta(s)$ is replaced by a filtered projection given by

  $$Q_\theta(s) = \int_{-\infty}^{+\infty} |\omega| \hat{P}_\theta(\omega) e^{i\omega s}\, \mathrm{d}\omega \qquad [12]$$

  where the multiplication by the ramp filter $|\omega|$ derives from the Jacobian of polar coordinates;
- *Back-projection*: this operation is the dual of the projection operation: indeed the projection assigns to a straight line L with coordinates (s, φ), hence to a point of the domain of definition of the Radon transform, the integral of f along L; on the other hand, the back-projection assigns the value of the Radon transform at (s, φ) to all points of the straight line L with the same coordinates (a pictorial representation of projection and back-projection is given in Figure 8); as a consequence the back-projection operator $R^\#$ assigns to each

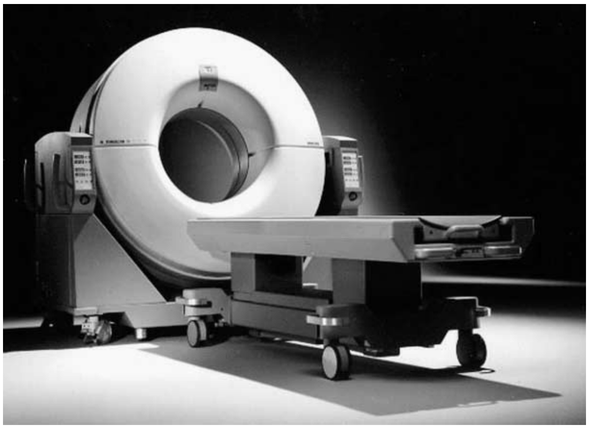

Figure 6 Picture of a scanner for X-ray tomography. The annular part contains the acquisition and scanning systems.

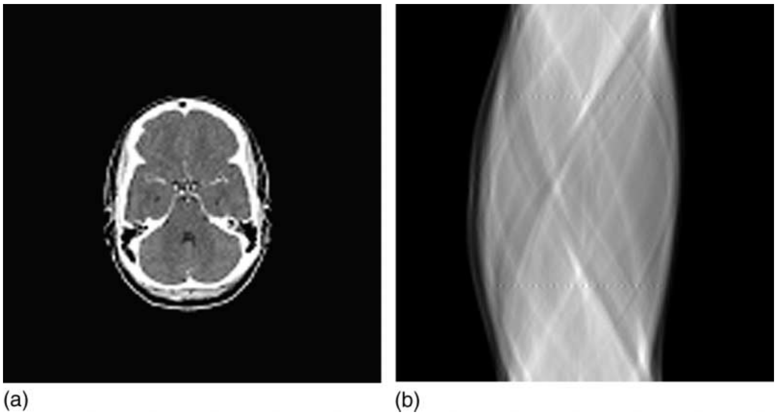

(a)

(b)

Figure 7 (a) Picture of a brain slice; (b) the corresponding sinogram. As explained in the text, the sinogram is a gray-level representation of the Radon transform in the s, φ plane. Each horizontal line in (b) corresponds to a projection of (a).

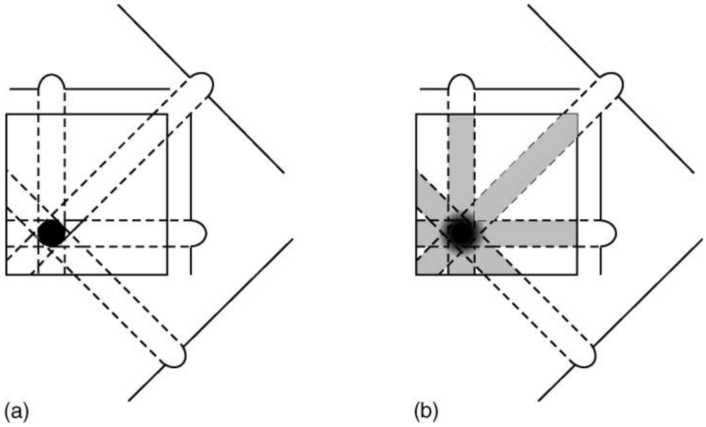

Figure 8 (a) The effect of the projection operation is shown by drawing the projections of a circular phantom corresponding to three directions. (b) shows the result obtained by applying the back-projection operation to the three projections given in (a). It is evident that the picture provided by this operation has a maximum on the domain of the circular phantom.

point **x** the sum (integral) of all values of the Radon transform corresponding to straight lines passing through **x**:

$$R^{\#}P_{\theta}(\mathbf{x}) = \int_0^{2\pi} P_{\theta}(x\cos\varphi + y\sin\varphi)\mathrm{d}\varphi \qquad [13]$$

The back-projection operator, when applied directly to the projections as in eqn [13], provides a blurred image of f; in order to get a satisfactory restoration it must be applied to the filtered projections $Q_{\theta}(s)$. This point is illustrated in Figure 9.

We point out that the ramp filter introduced in eqn [12] amplifies the high frequency components of the projections hence the noise corrupting these components. This remark clarifies that Radon transform inversion is an ill-posed problem and requires some kind of regularization. This is obtained by introducing an additional low-pass filter which reduces the effect of the high frequencies. The most frequently used combination of ramp and low-pass filter is the so-called Shepp–Logan filter.

X-ray tomography is also called transmission computerized tomography (TCT) because the image is obtained by detecting the X-rays transmitted by the body. It provides information about anatomical details of human organs because the map of the linear attenuation function is essentially the map of the density of the tissues.

A different type of information is obtained by the so-called emission computerized tomography (ECT) which is based on the administration of radionuclide-labeled agents known as radio-pharmaceuticals. Their distribution in the body of the patient depends on factors such as blood flow, metabolic processes, etc. Then a map of this distribution is obtained by detecting the γ-rays produced by the decay of the radio-nuclides. Therefore, ECT yields functional information, in the sense that the images produced by ECT show the function of the tissues of the organs.

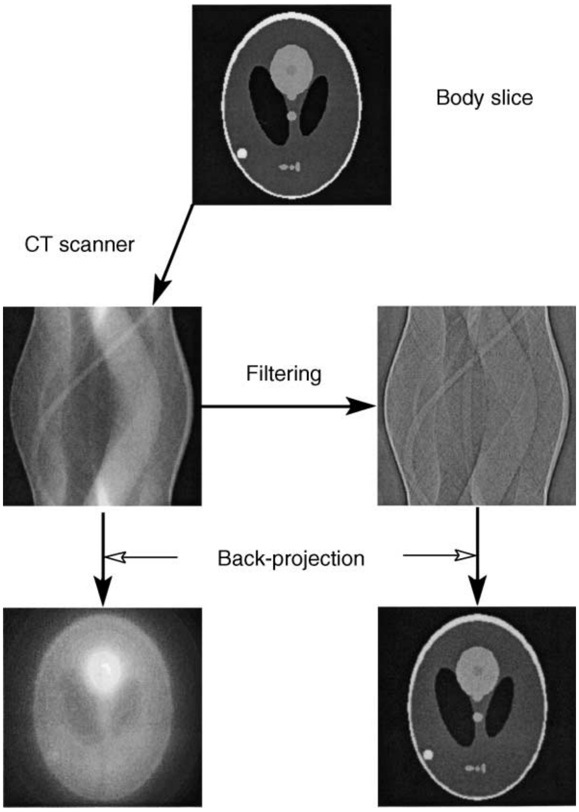

Figure 9 Illustrating the effect of the filtered back-projection. Shown here is: a slice of the so-called Shepp–Logan phantom; the corresponding sinogram (indicated by the arrow CT scanner); the filtered sinogram (indicated by the arrow filtering; the restorations obtained by applying the back-projection operation both to the original sinogram and to the filtered sinogram.

Two different modalities of ECT are usually considered:

- single photon emission computerized tomography (SPECT), which makes use of radio-isotopes where a single γ-ray is emitted per nuclear disintegration;
- positron emission tomography (PET), which makes use of β^+-emitters, where the final result of a nuclear disintegration is a pair of γ-rays, propagating in opposite directions, produced by the annihilation of the emitted positron in the tissues.

In both cases it is necessary to detect the γ-rays coming from well-defined regions of the body and, to this purpose, different methods are used in each case.

In SPECT, the discrimination of the γ-rays is obtained by means of a collimator, consisting of holes in a large lead slab covering the crystal detector face. In PET the collimation is obtained by means of pairs of detectors in coincidence. This technique, which is also called electronic collimation, is more accurate than the physical collimation used in SPECT and allows the design of systems with a great efficiency.

The basic reconstruction method in TCT, namely FBP, is often used in the reconstruction of SPECT and PET data. However, several corrections are necessary in practice. The principal ones are due to the collimator blur and to the scattering of photons in the body. If one develops a more accurate model of data acquisition by taking into account these effects, then the restoration problem implies the inversion of a very large matrix which is sparse and ill-conditioned. To this purpose iterative regularization methods are used such that the basic operation required at each step is matrix-vector multiplication. Other imaging modalities, which have been proposed for medical applications, are electrical impedance and microwave tomography. The underlying inverse problems are basically nonlinear so that the computational cost of the methods for their solution is, in general, too high for clinical applications. However the research in these areas is very active and, therefore, the situation could be improved in the near future.

See also

Tomography: Tomography and Optical Imaging.

Further Reading

Bertero M (1989) Linear inverse and ill-posed problems. In: Hawkes PW (ed.) *Advances in Electronics and Electron Physics*. vol. 75, pp. 1–120. New York: Academic Press.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.

Bertero M and De Mol C (1996) Super-resolution by data inversion. In: Wolf E (ed.) *Progress in Optics*. vol. XXXVI, pp. 129–178. Amsterdam, Elsevier.

Engl HW, Hanke M and Neubauer A (1996) *Regularization of Inverse Problems*. Dordrecht, Germany: Kluwer.

Groetsch CW (1993) *Inverse Problems in Mathematical Sciences*. Braunsschweig, Germany: Vieweg.

Herman GT (ed.) (1979) *Image Reconstruction from Projections*. Berlin: Springer.

Herman GT (1980) *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography.* New York: Academic Press.

Huang TS (ed.) (1975) *Picture Processing and Digital Filtering*. Berlin: Springer.

Kak AC and Slaney M (1988) *Principles of Computerized Tomographic Imaging*. New York: IEEE Press.

Krestel E (ed.) (1990) *Imaging Systems for Medical Diagnostics*. Berlin: Siemens Aktiengesellschaft.

Natterer F (1986) *The Mathematics of Computerized Tomography*. New York: Wiley.

Tikhonov AN and Arsenin VY (1977) *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley.

Adaptive Optics

C Pernechele, Astronomical Observatory of Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The image quality of a perfect optical system is ideally limited by its diffraction figure. In the real world there are many effects which degrade that physical limit, resulting in poorer image quality at the focal plane. These effects may be static (e.g., time-independent, as, for example, error in the alignment of the optical train) or dynamic (e.g., variable with time, such as optical misalignment introduced by thermo-mechanical distortion). Active and adaptive optics, the topics of this article, are technologies able to correct for the optical aberrations induced by dynamical effects.

There are a variety of causes liable for dynamical image degradation, among these time-variable optical misalignments, pointing and tracking errors (for motorized instruments), and atmospheric perturbations. In practice, both atmospheric turbulence and misaligned instrument introduce random spatial and temporal perturbations on the optical beam, distorting the spherical (ideal) wavefront incoming on to the focal plane. This leads to loss of both the diffraction limited condition and image resolution. A typical situation for astronomical application is shown in Figure 1. A target (a star in this case) emits light whose wavefront, being far from the observing system, can be considered plane. When this wavefront passes through the atmosphere it is distorted with a temporal frequency of the order of hundreds of Hertz. This dynamically distorted wavefront fills the aperture of the optical system and will once more be distorted by the dynamic misalignment of the optical system itself. The final wavefront will be deformed with respect to the ideal (spherical) wavefront.

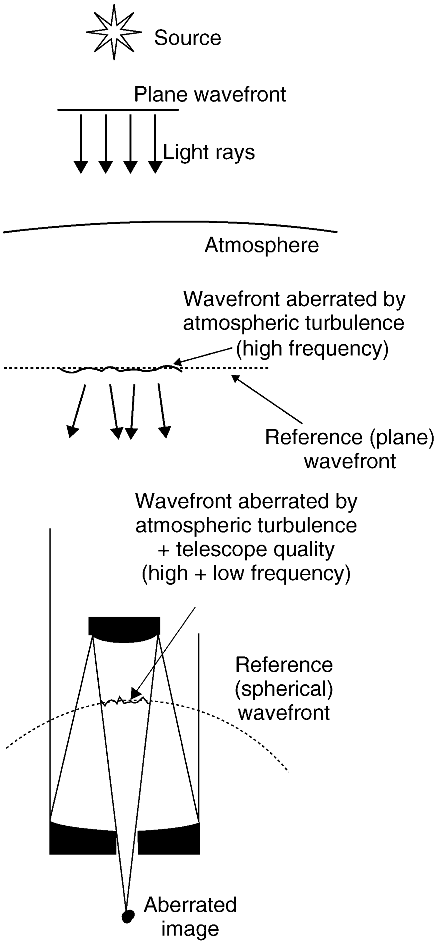
Source
Plane wavefront
Light rays
Atmosphere
Wavefront aberrated by atmospheric turbulence (high frequency)
Reference (plane) wavefront
Wavefront aberrated by atmospheric turbulence + telescope quality (high + low frequency)
Reference (spherical) wavefront
Aberrated image

Figure 1 Dynamical wavefront aberration.

Dynamic optical misalignment may depend on gravitational and thermal deformation, while pointing error may appear as very slow (drift), slow (wander), or fast (jitter) tilt of the image on the focal plane. Atmospheric perturbations are produced by air turbulence and may be natural (as in the astronomical or satellite tracking cases) or induced by the experiment itself (thermal blooming in high power laser applications). All these effects appear on very different time-scales, as shown in Figure 2. As noticeable, atmospheric turbulence introduces aberrations of all orders, at a frequency above about 1 Hz. Pointing and tracking errors introduce an image tilt with different time-scale. Gravitational effects induce mechanical deformation of the optics, mainly in astronomical applications because of the massive components (the telescope primary mirror in particular). Thermal deformation produces optical misalignment among optical components mainly resulting in defocus. Figure 2 is purely illustrative, because some effects may produce wavefront aberrations at different frequencies than indicated there. For example, drift may be faster than indicated in the figure, and thermal blooming (a thermal effect) introduces aberrations at a higher frequency, and so on.

The correction of the wavefront errors is usually made by means of devices able to change the phase wave somewhere in the light path. Though the concept for correcting aberrated wavefront at slow or high frequency is the same, for example, the use of one or more deformable elements in the optical path, the technologies developed for each area are quite

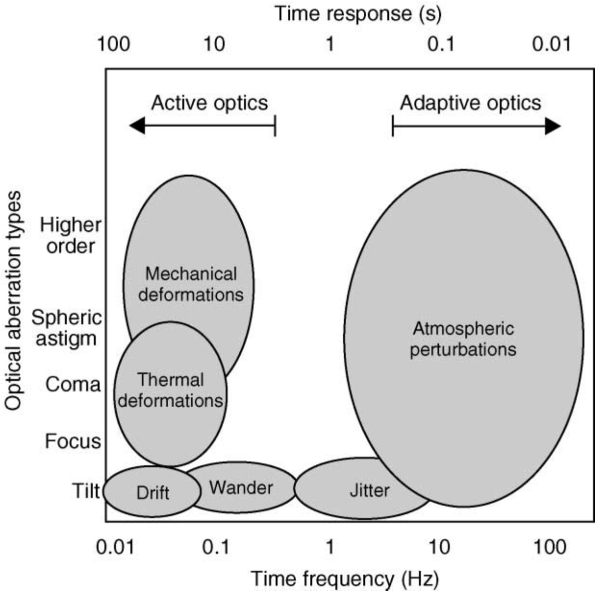

Figure 2 Wavefront aberration time-scale and associated aberrations.

different: variations at slow frequency (approximately below 0.1 Hz in Figure 2) are indicated as quasi-static errors and the techniques developed to overcome them are properly called active optics, while adaptive optics indicates the techniques to overcome the unwanted effects at high frequency (above about 1 Hz in frequency). The frequency value dividing the active from adaptive optics is not well defined and depends on the field of application and its scientific engineering communities. The important point is that a correct use of a combination of the two techniques may ultimately lead to bringing the imaging system close to its diffraction limit.

Slow Corrections: Active Optics

The analysis of the wavefront is required before its correction, and one or more deformable/movable mirrors and their control system, to operate the corrections. A typical system for low frequency correction is shown in Figure 3. A telescope contains three mirrors (M1, M2, and M3) and an instrument to detect its scientific target, usually placed on its optical axis. The dynamic deformations present in this kind of instrument are mainly third-order spherical (thermal deformation of M1), third-order astigmatism (a saddle-horse deformation of M1 due to gravity when the telescope is pointing away from the zenith), and third-order coma (due to gravitational decentering between M1 and M2). Since a telescope has a tracking system some pointing error may be present, causing motion of the image in the focal plane. The observation of an off-axis target (typically a star for astronomical applications) is made to measure the deformations of the incoming wavefront by means of a wavefront sensor (WS) system. This step is completed using some of the wavefront sensing technique described elsewhere in this encyclopedia. The WS must measure the information on the wavefront distortion with adequate spatial resolution to compensate for aberrations across the full aperture of the system. Once the wavefront distortion has been measured by the WS, the needed corrections are sent to the control system which apply the appropriate mirror deformation/movements. In the example of Figure 3 only the primary mirror (entrance pupil) is deformable and it is used for correcting spherical and astigmatism aberrations. M2 is used to compensate the decentering coma by means of lateral displacement. M3 may be tilted to correct for low-frequency pointing errors. In astronomical applications active optics is essential for telescopes with primary-mirror diameters larger than approximately four meters which cannot be manufactured rigidly at reasonable cost. Active optics requires a thin deformable or a segmented primary mirror maintained on the correct shape at every elevation. Using this method, low-frequency correction has been achieved to date in telescopes with primary mirrors of the order of ten meters.

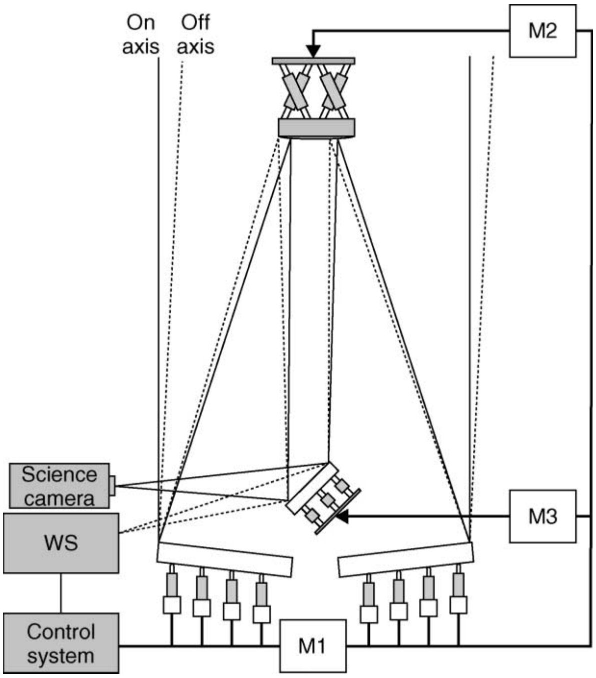

Figure 3 A typical system for correcting low-frequency dynamical aberrations (active optics).

Fast Corrections: Adaptive Optics

It is clear that corrections at high frequency are impossible using the approach discussed above, because of the large dimensions and weight of the mirror and the low response time required. For an adaptive correction it is necessary to image a smaller system pupil in the optical train and make the active correction of the wavefront. A typical configuration is shown in Figure 4. The light from the telescope focal plane (corrected for low frequency errors by means of active optics) is collimated and a pupil is located in this beam. Behind the active mirror a beamsplitter is used to send part of the light into the WS and measure wavefront distortions. The analysis results are sent to the control system which performs the active mirror deformation for the wavefront compensation. If the WS and the control system are sufficiently fast (typically faster than the time-variations of the wavefront distortion), this system is able to produce a plane wavefront which can be focused at the diffraction limit of the system. For this reason

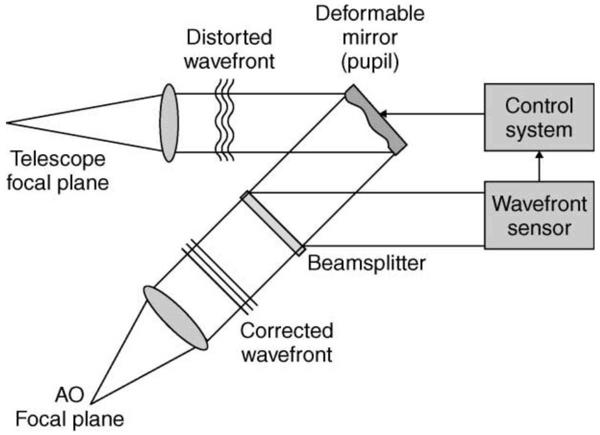
Deformable mirror (pupil)
Distorted wavefront
Control system
Telescope focal plane
Wavefront sensor
Beamsplitter
Corrected wavefront
AO Focal plane

Figure 4 A typical setup for correcting high-frequency dynamical aberrations (adaptive optics).

adaptive optics are often referred to as real-time system.

The seminal paper on the adaptive optics concept was written by H.W. Babcock of the Mount Wilson and Palomar observatories in 1953. Since then a great improvement has been made, especially in military and astronomical fields. In recent years, however, many applications of adaptive optics have developed in fields different from traditional ones, such as precision manufacturing by means of high-power lasers, telecommunications, aero-optics, medical physics, and ophthalmology.

We describe adaptive optics with particular reference to astronomical applications. We provide a brief description of other applications at the end of this article.

Imaging Through the Atmosphere

The optical effects of the atmospheric turbulence are due to local temperature variations that induce fluctuations in the refractive index of air. The refractive index changes depend on the air density as well as the range of the temperature variations. Air density is highest at sea level and decreases exponentially with altitude. Optical effects of turbulence decrease with the altitude of the site, the reason for locating astronomical telescopes on mountains. The effects of turbulence on a point source image are depicted in Figure 5. Hereafter we consider an ideal telescope, for example, one without need for active correction at low frequencies. We consider the source 1 (continous line). The radiation emitted by the source is a spherical wavefront that can be treated as a plane, because of the very long distance to the Earth's atmosphere.

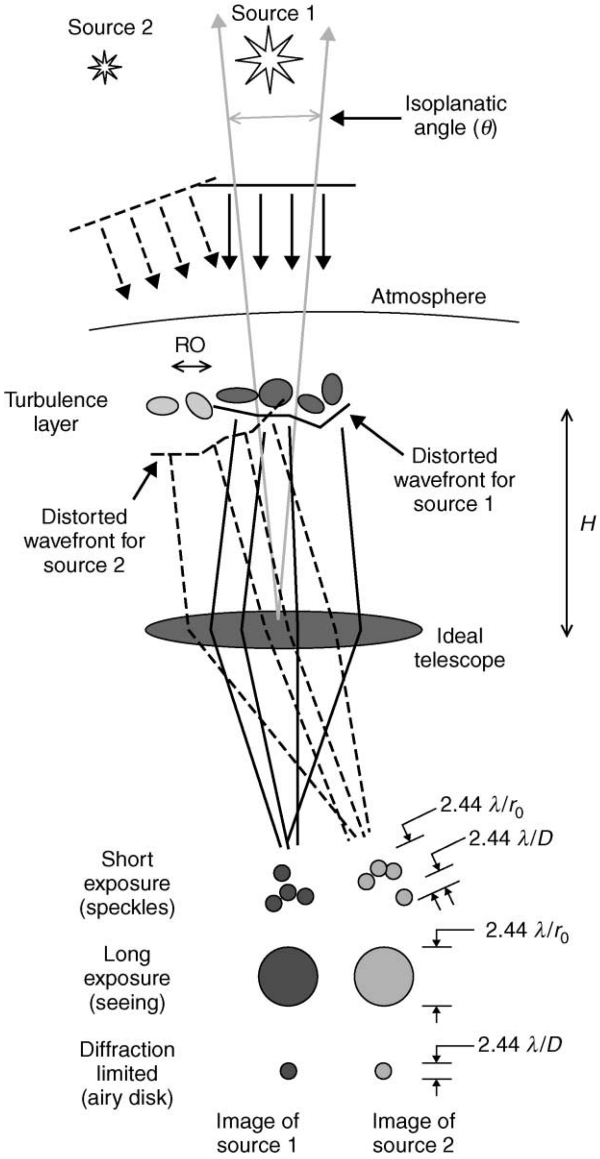
Source 2
Source 1
Isoplanatic angle (θ)
Atmosphere
RO
Turbulence layer
Distorted wavefront for source 1
Distorted wavefront for source 2
H
Ideal telescope
$2.44\ \lambda/r_0$
$2.44\ \lambda/D$
Short exposure (speckles)
$2.44\ \lambda/r_0$
Long exposure (seeing)
Diffraction limited (airy disk)
$2.44\ \lambda/D$
Image of source 1
Image of source 2

Figure 5 Imaging through the atmosphere.

In absence of wavefront distortion, the angular diameter of the image (Airy disk), is approximately $2.44\ \lambda/D$, with λ the wavelength of the observation and D the telescope aperture (the mirror diameter). In this case the image is said to be diffraction limited, and it is the best possible performance for the observation. This is the performance of a space-based ideal telescope.

In the presence of the atmosphere the incoming wavefront is distorted by the turbulence and when it reaches the telescope aperture it is no longer a plane (see in Figure 5). The so-called Fried's parameter r_0 (also called atmospheric coherence length) is the most common parameter used to quantify the distortion effects of atmospheric turbulence. It is defined as the distance over which the wavefront phase is highly correlated (in this way it provides a measure of the turbulence scale size). For an undistorted plane wave, r_0 is infinity, while for atmospheric distortion it typically varies between 5 cm (bad sites) and 20 cm (excellent sites). We depict Fried's parameter schematically in Figure 5, as bubbles of dimension r_0 in the turbulent layer (referred to as Fried's cells). r_0 is a statistical parameter, which depends on time and wavelength ($r_0 = r_0(t, \lambda^{6/5})$). Its variability with time may be considerable, sometimes changing by a factor of 2 within a few seconds. A small telescope, with a diameter less than r_0, will always operate at this diffraction limit. When the r_0 is smaller than the telescope aperture, the angular size of the image is different if we consider long- or short-time exposures. If the exposure time is very short (less than hundredths of a second) it is possible to image a structure in the focal plane (see short exposure image in Figure 5). This structure, which varies with time, is due to high-order wavefront distortions that occur within the aperture. The short exposure image is broken into a large number of speckles of diameter 2.44 λ/D, which are randomly distributed over a circular region of approximately 2.44 λ/r_0 in angular diameter. A lower-order wavefront distortion is correlated with the size of the largest disturbances produced by atmospheric turbulence (the so-called outer scale) which varies considerably (from 1 m to over 1 km). Such disturbances produce a time-dependent tilt of the overall wavefront that causes fluctuations in the angle of arrival of the light from a star and randomly displace the image. For D of the order of the Fried parameter r_0, image motion dominates, and a simple tip-tilt mirror (not just deformable) provides useful correction. For D greater than r_0, speckles dominate and a deformable mirror is necessary to account for the wavefront distortion. For long exposures the dimension of the image is determined by the ratio λ/r_0 rather than λ/D. The angular resolution is diminished due to the atmospheric turbulence (see Figure 5). For a 4 m telescope, atmospheric distortion degrades the spatial resolution by more than one order of magnitude compared to the diffraction limit, and the intensity at the center of the star image is lowered by a factor of 100 or more. Even at the best astronomical sites, ground-based telescopes observing at visible wavelengths cannot, because of atmospheric turbulence alone, achieve an angular resolution better than telescopes of about 20 cm diameter.

Most of the distortions imposed by the atmosphere are in the wave-phase and an initially plane wavefront traveling 20 km through the turbulent atmosphere accumulates, across the diameter of a large telescope, phase errors of a few micrometers.

Turbulence is distributed along the propagation path through the atmosphere, and then the wavefront errors become uncorrelated as the field angle increases. This effect is known as angular isoplanatism and is one of the major limitations to the astronomical adaptive optics. The structure becomes completely uncorrelated out of the isoplanatic angle, approximately equal to r_0/H, where H is the altitude of the turbulence layer. This decorrelation is sketched in Figure 5, where the short exposure images for objects 1 and 2 have different speckles structure.

In Earth's atmosphere significant turbulence occurs at altitudes up to 20 km, often with large peaks in the vicinity of the tropopause, at around 10 km there exist, in practice, more than one parameter H in Figure 5, one for each turbulence layer. This three-dimensional distribution of the turbulence considerably restricts the angle over which adaptive compensation is effective and leads to the design of multiconjugate adaptive optics, as we next explain.

Adaptive System Design

As stated above, an active system is composed of at least one wavefront sensor and one deformable mirror. Typically, the wavefront is corrected by making independent tip, tilt, and piston movements of the deformable mirror surface, as shown in Figure 6. As a rule of thumb, a deformable mirror used for atmospheric correction should have one actuator for each Fried's cell, such that the total number of actuators should be equal to the number of Fried's cells filling the telescope aperture, i.e., D^2/r_0^2. For instance, a near-perfect correction for an observation in visible light (0.6 μm) with an 8 m telescope placed in a good astronomical site ($r_0 = 20$ cm), would require a deformable mirror with approximately 6400 actuators in its adaptive optics system. Recalling that r_0 is strongly dependent on the wavelength (i.e., proportional to $\lambda^{6/5}$), similar performance in the infrared at 2 μm requires only 250 actuators. This is one reason why adaptive correction is easier in the infrared. A large number

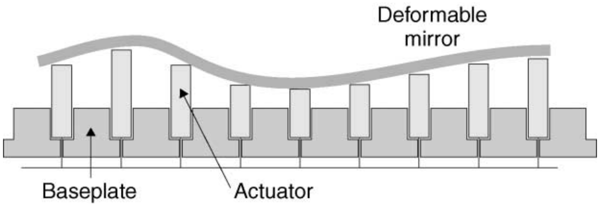

Figure 6 A sketch of a deformable mirror.

of Fried's cells requires a similarly large number of subapertures in the wavefront sensor. Commonly used WSs are based on modifications of the classic Hartmann screen test or of curvature type. In the former case the adaptive device is made with individual piezoelectric actuators (see Figure 6), while in the latter the correction is achieved with a bimorph adaptive mirror, made of two bonded piezoelectric plates. With both methods, wavefront sensing is done on a reference off-axis object (Figure 3), or even on the observed object itself (Figure 4) if it is bright enough. In the former case the angular distance of the reference object must fall inside the isoplanatic angle and must be sufficiently luminous. In typical astronomical sites the isoplanatic angle at a wavelength of 0.6 μm is ~5 arcsec, while it can be ~20 arcsec at 2 μm. This is an additional reason why the adaptive optics system is most efficient in the infrared range. Correction in the visible requires a reference star approximately 25 times brighter than in the infrared. Most current astronomical systems are designed to provide diffraction-limited images in the near-infrared (1 to 2 μm) with the capability for partial correction in the visible. However, some military systems for satellite observations do provide full correction in the visible on at least 1 m class telescopes.

The control system is generally a specialized computer that calculates, from the WS measurements, the commands to send the actuators of the deformable mirror. The calculation must be performed quickly (within 0.5 to 1 ms), otherwise the state of the atmosphere may have changed making the wavefront correction inaccurate. The required computing power can exceed several hundred million operations for each command set sent to a 250 actuator deformable mirror. As in active optics systems, zonal or modal control methods are used. In zonal control, each zone or segment of the mirror is controlled independently by wavefront signals that are measured for the subaperture of that zone. In modal control, the wavefront is expressed as the best-fit linear combination of polynomial modes of the wavefront distorted by the atmospheric perturbations.

Future Development

As explained above, a convenient method for measuring the wavefront distortion is to look for an off-axis reference star (also known as natural guide star, NGS) within the isoplanatic angle. However, it is generally impossible to find a bright reference star close to an arbitrary astronomical target. Conditions are far better in the infrared than in the visible because atmospheric turbulence (and in particular its high spatial frequencies) is, for a given AO correction, less pronounced at longer wavelengths. The spatial and temporal sampling of the disturbed wavefront can be reduced, which in turn can permit the use of fainter reference stars. Coupled with the larger isoplanatic angle in the infrared, this gives a much better outlook for AO correction than in the visible.

Nevertheless, even for observations at 2.2 μm, the sky coverage achievable by this technique (equal to the probability of finding a suitable reference star in the isoplanatic patch around the chosen target) is of the order of 0.5 to 1%. It is therefore unsurprising that most scientific applications of AO have to date been limited to the study of objects which provide their own reference object. The most promising way to overcome the isoplanatic angle limitation is by the use of artificial reference stars, also referred to as laser guide stars (LGS). These are patches of light created by the back-scattering of pulsed laser light of sodium atoms in the high mesosphere or by molecules and particles located in the low stratosphere. The laser beam is focused at an altitude of about 90 km in the first case (sodium resonance) and 10 to 20 km in the second case (Rayleigh diffusion). Such an artificial reference star can be created as close to the astronomical target as desired, and a wavefront sensor measuring the scattered laser light is used to correct the wavefront aberrations on the target object.

Military laboratories have reported the successful operation of adaptive optics devices at visible wavelengths using laser guide star on both a 60 cm telescope and a 1.5 meter telescope, achieving an image resolution of 0.15 arcsec. Nevertheless, there are still physical limitations of a LGS. The first problem, focus anisoplanatism, or cone-effect, was realized long ago. Since the artificial star is created at a relatively low altitude, back-scattered light collected by the telescope forms a conical beam, which does not cross the same turbulence-layer areas as light from the astronomical target. This leads to a phase estimation error, which in principle can be solved by the simultaneous use of several laser guide stars surrounding the targeted object. A more significant problem is image motion or tilt determination. Since the paths of the outgoing and returning light rays are the same, the centroid of the artificial light appears to be stationary in the sky, while the apparent position of an astronomical source suffers lateral motions (also known as tip/tilt). The simplest solution is to supplement the AO system and LGS with a tip/tilt corrector set on a (generally) faint close NGS. Performance is then limited by poor photon statistics in the tip/tilt correction.

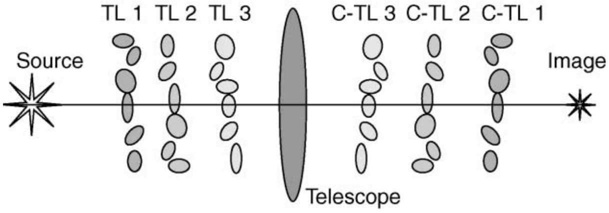

Figure 7 Conjugated planes of turbulent layers.

Adaptive optics with a multicolor laser probe is another concept that has been studied to solve the tilt determination problem of laser beacon based AO. Only applicable to sodium resonant scattering at 90 km, it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength. Its main drawback is the limited returned flux, due to the saturation of mesospheric sodium layer.

The quality of a site for adaptive optics is determined not only by the median seeing, but also by the height and thickness of the turbulent layers (often only a few thin layers dominate). Some adaptive-optics systems can be adapted to correct for turbulence originating at a particular height by using several reference stars. This technique is known as multiconjugate adaptive optics (MCAO) and is still in his infancy. This concept deals with a 3-D view of the atmosphere, and for this reason it is also called turbulence tomography. Several wavefront sensors are needed for tomography. Turbulence tomography is also useful to correct for the cone effect when using LGS. When several deformable mirrors are placed on the appropriate conjugate planes, as shown in Figure 7, anisoplanatism reduces and the size of the corrected field of view increases. The situation is roughly depicted in Figure 7, where each turbulent layer (indicated as TL) has a proper conjugated plane (C-TL).

Another challenging optical solution is one of utilizing large secondary adaptive mirrors. This application seems especially interesting at thermal wavelengths, where any additional mirror raises the significant thermal background detected by instruments.

Other Adaptive Optics Applications

In recent years the adaptive optics technique has been developed in many applications beside astronomy and the military, for example, as in high-power laser welding or cutting, in ophthalmology and in aero-optics.

In high-power laser applications, thermal blooming takes place when laser light is absorbed by the gases and particles that make up the atmosphere. Since energy is conserved, the energy lost by the laser (its intensity diminishes) is gained by the atmosphere that heats up. This inhomogeneity at the boundary of the laser beam generates wavefront distortions by turbulence, as shown in Figure 8. This means that laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. The net effect is that, in the case of a Gaussian laser beam, light following a wavefront which passed at an earlier time, ‘sees’ a concave lens that deflects the successive rays out of the beam (away from the axis) causing the beam to appear to grow in radius or ‘bloom’. Adaptive optics techniques for real-time compensation of blooming in laser beams were developed in the late 1960s. Given the high coherence and monochromaticity of a laser, the technique is more straightforward than for more common white-light applications of the astronomical applications.

Density fluctuations in a compressible shear layer are sufficiently high to cause a time varying index of refraction of the layer. This situation is common in a turbulent flowfield, and the study of the wavefront aberrations induced by this flowfield, when it is of relatively short propagation length in the near-field of the optical system, is referred to as aero-optics. Turbulent flowfield induces various types of degradations in the performances of an optical system. For example the jitter and defocusing of laser-based weapons cause a reduction in the amount of energy delivered to the target. Tracking systems experience bore-site errors and imaging systems experience blurring. Recent work suggests that adaptive optics compensation may successfully overcome these problems.

In ophthalmology the adaptive technique finds application in studying the retina of the eye. In this case the eye itself produces wavefront aberration, and the retina is seen as blurred. Techniques in ophthalmic laser surgery are being developed to use adaptive optics to correct for detected distortions.

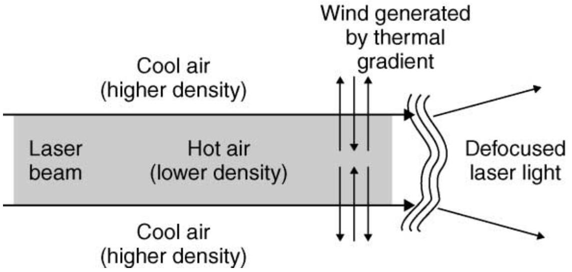
Wind generated by thermal gradient
Cool air (higher density)
Laser beam
Hot air (lower density)
Defocused laser light
Cool air (higher density)

Figure 8 Thermal blooming effect in high-power laser applications.

In the field of telecommunications, studies are in progress for using adaptive techniques in data transmission between ground-based stations and satellites.

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Instrumentation:** Astronomical Instrumentation; Telescopes. **Phase Control:** Wavefront Coding; Wavefront Sensors and Control (Imaging Through Turbulence).

Further Reading

Beckers JM (1993) Adaptive optics for astronomy: principles, performance and application. *ARAA* 31: 13–62.

Fried DL (1965) Statistics of a geometric representation of wavefront distortion. *JOSA* 55: 1427–1435.

Hardy JW (1998) *Adaptive Optics for Astronomical Telescopes*. New York: Oxford University Press.

Jumper EJ and Fitzgerald EJ (2001) Recent advances in aero-optics. *Progress in Aerospace Sciences* 37: 299–339.

Liang J, Williams DR and Miller DT (1997) Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSA A* 14(11): 2884–2892.

Lukin VP (1995) *Atmospheric Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Ragazzoni R, Marchetti E and Valente G (2000) Adaptive-optics corrections available for the whole sky. *Nature* 403(6765): 54–56.

Roddier F (1999) *Adaptive Optics in Astronomy*. Cambridge: Cambridge University Press.

Roggeman M and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Tyson RK (2000) *Introduction to Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.

Welford W (1974) *Aberrations of the Symmetrical Optical System*. New York: Academic Press.

Wilson RW and Jenkins CR (1996) Adaptive optics for astronomy: theoretical performance and limitations. *MNRAS* 268: 39.

Hyperspectral Imaging

M L Huebschman, R A Schultz and H R Garner,
University of Texas, Dallas, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hyperspectral imaging techniques on a microscopic level have paralleled the hyperspectral remote astronomical sensing which has developed over the years. It is especially apparent in remote sensing of our own planet's land masses and atmosphere, such as in the LANDSAT satellite program. In biology, biochemistry, and medicine, hyperspectral imaging and multispectral imaging are often interchanged, although there is a distinct difference in the two techniques.

A multispectral imager produces images of each of a number of discrete but relatively wide spectral bands. Multispectral imagers employ several wide bandpass filters to collect the data for the images. A hyperspectral imager produces images over a contiguous range of spectral bands and, thus, offers the potential for spectroscopic analysis of data. Hyperspectral imagers may use gratings or interferometers to produce the contiguous range of spectral bands to record.

Fluorescent dyes (also referred to as fluorophores or fluorochromes) and recently quantum dots, as markers, are a very powerful tool for analyzing organic tissues for researchers in cytology (study of cell function), cytogenetics (study of the human chromosomes), histology (study of structure and function of tissues), microarray scanning (analysis of concentrations of multiple varied probes), and biological material science. Without the correlation of the spectral and spatial data in biological images, there would be little useful information for the researcher. Therefore, hyperspectral imaging in the biological sciences must combine imaging, spectroscopy, and fluorescent probe techniques. Such imaging requires high spatial resolution to present a true physical picture, and the spectroscopy must also have high spectral resolution in order to determine the quantities of biochemical probes present. The correlation must provide the color-position relation for identification of meaningful information.

This description will begin with an initial review of multispectral imaging to lead into a discussion of hyperspectral imaging. The discussion will examine aspects of hyperspectral imaging based on an interferometric method and on a grating dispersion method similar to the methods in multispectral imaging.

Image Parameters

The analysis of fluorophores involves the excitation of the dye with a shorter wavelength and thus higher energy than the emission spectrum. The use of the appropriate excitation wavelength will yield a maximum emission intensity at a characteristic wavelength. The quality of a fluorescent image generally combines contrast and brightness. Contrast depends on the ratio of emission intensity to that of background light. Consequently, the use of filters to block the excitation wavelength and prevent it from overwhelming the field is critical in providing high contrast for the detection of fluorophore emission. However, the use of narrow filters to restrict the background can severely limit the brightness of the fluorescent image. This is due to the barrier filters allowing only a portion of the actual emission spectrum to pass. Brightness is an intrinsic property of each individual fluorophore, but is also dependent on the ability to illuminate with the peak excitation wavelength for that fluorophore.

The balance between contrast and brightness may be further complicated by applications designed to facilitate the evaluation of multiple fluorophores. The selection of appropriate excitation and emission filters within a single set to allow for the simultaneous viewing of multiple fluorophores can be very inhibiting. Although quad-band beamsplitters are currently available, commercial advances to date have yielded only triple-pass combinations for simultaneous detection, such as one of which permits the simultaneous detection of the widely used fluorophore combination fluorescein-5-isothiocyanate (FITC), Texas Red and 4′,6-diamidino-2-phenylindole (DAPI) (e.g., Chroma Technology Corp.).

Although the capture of digital images has propelled this field, efforts to collect multicolored images as digital files also suffer from technical limitations. Color CCD cameras often utilize filters or liquid crystals to segregate the red, green, and blue wavelengths onto the imaging chip. The employment of these cameras is restricted to resolving colors as discriminated by the filters used. For scientific applications, then, the greatest sensitivity and selectivity has been achieved through black and white CCD cameras with optimized filters for a specific application. While this approach is widely used, it suffers from the inherent limits of the spectra of data that can be analyzed since it is defined by the characteristics of the optical filters employed. Thus, the system is restricted to the use of fluorophores that can be effectively resolved from each other by filters, and the filters selected limit the data collected to a portion of the total emission spectrum for any one fluorophore.

Multispectral Imaging

Multispectral imaging systems in biomedical research are composed of a microscope, filters and/or interferometer for spectral selection, a charged coupled device (CCD) camera to record the data, and computer control for acquisition, analysis, and display (Figure 1a). The introduction of CCD cameras has offered a means to document results and capture images as digital computer files. This has facilitated the analysis of results with more and more complex probe sets and has allowed the use of more sophisticated computer algorithms for data analysis. The other major component of the multispectral imaging system is a light source to excite the sample's fluorescent dyes whose emissions are selected by filters or an interferometer for recording on the CCD camera.

An example of the use of multispectral imaging is in the identification of the 24 human chromosomes. A scheme was presented for the analysis of human chromosomes by differentially labeling each chromosome with a unique combination of 5 fluorophore-labeled nucleotides (building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group). These complex probes are simultaneously hybridized to metaphase chromosomes and analyzed to determine which of the fluorophores are present. Combinations of one, two, or three of the five fluorophores are enough to distinguish a particular chromosome from the others.

An interferometric multispectral system used to perform this type of analysis is the Spectral Karyotyping (chromosome typing) or SKY System (Figure 1a). This technology does not represent hyperspectral imaging, but is presented for comparison. Illumination light passes through a triple band-pass dichroic filter which allows for the simultaneous excitation of all the fluorochromes in, for instance, Speicher's scheme. Modern multi-band pass filters add little to image distortion while effectively blocking unwanted autofluorescence or excitation illumination. The fluorescence emissions from the sample are split into two separate beams by the interferometer. The beams are then recombined with an optical path difference (OPD) between them such that the two beams are no longer in register and interfere with each other. By adjusting the OPD, a particular wavelength band can be selected for, which results in constructive interference and thus can be collected by the CCD camera pixels. After adjusting the OPD to record

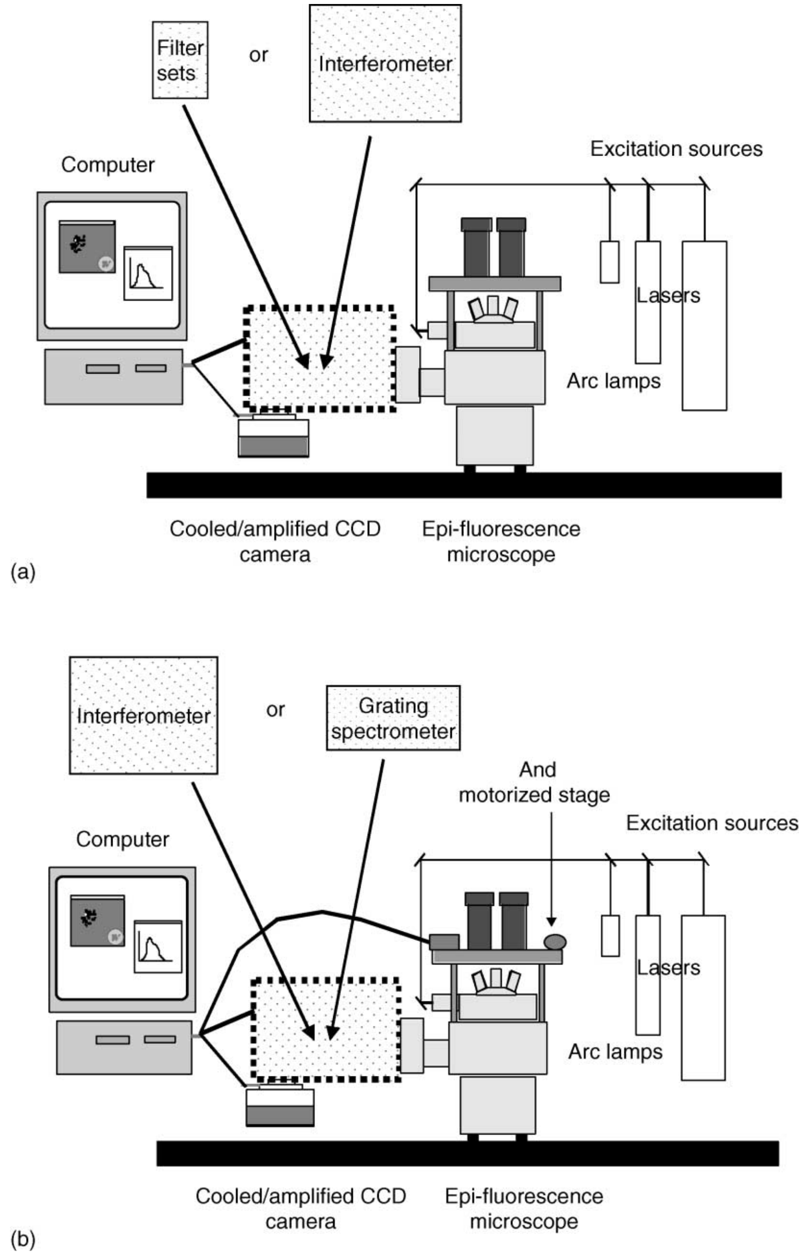
Filter sets
or
Interferometer
Computer
Excitation sources
Lasers
Arc lamps
Cooled/amplified CCD camera
Epi-fluorescence microscope
(a)
Interferometer
or
Grating spectrometer
And motorized stage
Computer
Excitation sources
Lasers
Arc lamps
Cooled/amplified CCD camera
Epi-fluorescence microscope
(b)

Figure 1 The instruments for multispectral and hyperspectral imaging are similar. (a) Multispectral imagers use interferometers or bandpass filters to collect noncontiguous emission wavelengths. (b) Hyperspectral imagers employ interferometers or gratings to collect a contiguous wavelength band. Computers, microscopes, CCD cameras, and excitation light sources are common to all the techniques. The systems which use filter sets or interferometers use motorized filter placement systems or automated interferometers. The gratings systems employ a motorized stage for the sample movement. Adapted with permission from Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

the 5 dye wavelengths, each pixel's 5 records are then compared to the known probe-labeling scheme to identify the corresponding chromosome material. The image of the metaphase chromosome spread can be karyotyped using specific pseudo-colors for maximal visual discrimination of the chromosomes.

Another nonhyperspectral method is the Multi-Fluorescence *In Situ* Hybridization or M-FISH System. It uses five single-band-pass excitation/emission filter sets where each filter set is specific for one of the five fluorochromes used in the combinatorial labeling process (Figure 1a). Again, five separate exposures are taken (one exposure for each filter set) and combined to create a multicolor image. Most M-FISH systems now have motorized filter set selectors which minimize any image registration shifts that may occur between exposures. For each pixel, the presence/absence of signal from each of the five fluorochromes is assessed and then matched to the probe-coding scheme to determine which chromosome material is present. As with the SKY System, an image of the metaphase spread is converted into a karyotype and specific pseudo-colors are used to maximize visual discrimination of the chromosomes.

Hyperspectral Imaging

As has been stated, the main difference between multi- and hyper-spectral imaging is the collection of a broad contiguous band of wavelengths. The apparatus and methods for hyperspectral imaging are similar to multispectral imaging (Figure 1). Instead of collecting sections of the emission spectrum using filters or interferometers, hyperspectral imaging collects a continuous emission spectrum by dispersing the wavelength of light using gratings and measuring the wavelength intensity, or by dispersing Fourier components of light using interferometers and measuring interference intensity (Figure 1b). Multispectral images are limited to measuring total intensity within the selected wavelength band. Determining the abundance of more than one dye per pixel is less accurate when only a small segment of each dyes' spectrum is passed by the emission filter. This has not been considered a limitation in the past because most research depended on just identifying the presence and not the abundance of several spectrally broad and overlapping dyes which also occupied relatively large spatial areas. With the advent of smaller and smaller featured microarrays, the desire to use more dyes in identification schemes, and the use of quantum-dot markers which can be used to indicate both genetic makeup and quantity, the limitations of multispectral imaging has led to a move to hyperspectral imaging.

In contrast to the fluorescence multispectral imaging approaches, a hyperspectral image is fully three-dimensional. Every XY spatial pixel of the image has an arbitrarily large number of wavelengths allied with it. An intensity cube of data is collected. Two of the axes of the cube are spatial and one axis is wavelength (see Figure 2). At each of the 3D points, there is an associated intensity of emitted energy. Therefore, it is a four-dimensional information space. The individual intensity signature of differing emitting bodies can then be separated to give a complete view. Spectral decomposition is a mathematical/algorithmical process by which individual emission spectra of the dyes can be uniquely separated from a composite (full spectral) image by knowing the signature of each dye emission spectra component measured separately. Spectral decomposition may be performed using standard algorithms with codes which are readily available.

Interferometric Hyperspectral Imaging

The system architecture of the interferometric hyperspectral imager resembles that of the SKY System (Figure 1). But, instead of just sequencing the interferometer to select a set of specific wavelengths to record as in the SKY System, the interferometer in the hyperspectral imager is sequenced through

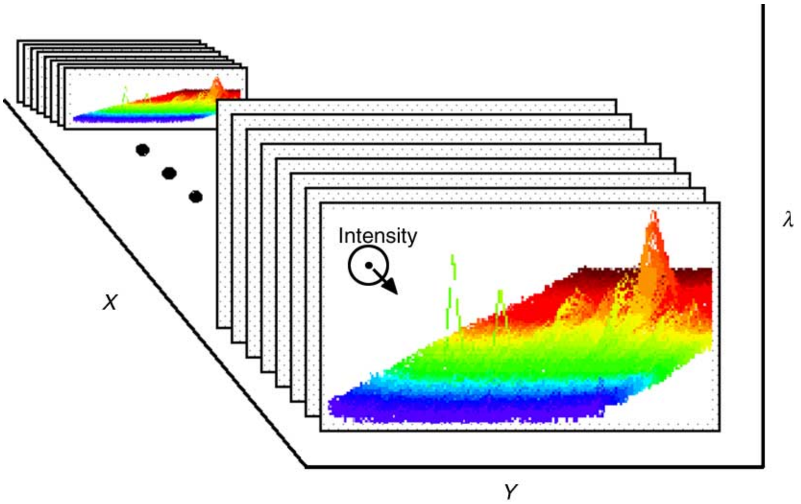

Figure 2 Intensity data cube. In this data cube, each CCD exposure captures a measurement in intensity (number of counts) by wavelength for a small spatial area in *Y* and Δ*X*. Many displacements in *X* with subsequent CCD acquisitions yield a 'cube' of data in *X*, *Y*, and λ. The intensity counts may be thought of as another dimension, *I*(*x*,*y*,λ). Another form of the data cube is with each CCD acquisition comprising intensity counts for a spatial *X* and *Y* area for a small Δλ and multiple displacements in λ for each acquisition.

discrete optical path differences to record the intensity at each pixel for each OPD. Fourier spectroscopy techniques are used to transform the OPD intensities at each pixel into the wavelength intensities at each pixel. Each pixel may be thought of as a Fourier spectrometer which is not dependent on any other pixel. Then, spectral decomposition algorithms can be run to determine which dyes are present (see the next section).

Like the multifilter methods, the interferometer approach provides a real two-dimensional image that can be observed, focused, and optimized prior to measurement. But like the filtering methods, the excitation source must be filtered for sufficient contrast to view the fluorescent emissions. Interferometers have high optical throughput, relative to other methods, and can have high and variable spectral resolution. The spectral range can cover from the ultraviolet to the infrared. The high and variable spectral resolution allows the user to trade off sensitivity, spectral resolution, and acquisition time in a way that is not possible with any other technique. And interferometers do not cause polarization of light as a filter may.

However, interferometric instruments are susceptible to image registration and lateral coherence difficulties. They are limited by the extreme precision required of the optical components which results in the continuous need for monitoring and adjustment, since aberrations are not tolerated in interferometric instruments. The throughput is reduced by loss due to the beamsplitter design and from the mirrors. Additionally, the Fourier transform instrument deconvolutes based on a two-step process; the correlation of the light passed by an interferometer followed by an inverse Fourier transform in software.

To mitigate the extreme precision and sensitivity of the most interferometric systems, Sagnac interferometers have been employed because of their common path intrinsic stability. The common path compensates any misalignment, vibration, or temperature affect. The OPD is provided by the Sagnac affect. When the Sagnac interferometer is rotated, the path traveled in one direction will be different from that of the light on the common path in the other direction. Although, this interferometer method is trading one difficulty for another, it is somewhat easier to manage the rotation of a physical system than to maintain the exacting precision of mirror alignments in a standard interferometer.

Spatial Resolution

The spatial resolution of a Sagnac Hyperspectral Imager is set by the size of the pixels of the CCD camera and the microscope system magnification, since the interferometric hyperspectral viewer sees the entire field of view at once. Let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M, then the spatial resolutions in the x and y directions are $2\delta x/M$ and $2\delta y/M$, respectively. For CCD pixel sizes of 10 microns square, binning together two pixels in the x- and two pixels in the y- direction, and an overall magnification of 60, the x and y resolutions are both 0.66 microns.

Field of View

For N pixel rows and C pixel columns, the field of view will be $N\delta y/M$ and $C\delta x/M$ for the length of the area along the y- and x- axis, respectively. For a CCD array of 1536 columns by 1024 rows, and the parameters above, the field of view is 171 microns by 256 microns, x and y, respectively.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution will depend on the ability to vary the OPD and the Fourier transform. While the interferometer is used basically as a filter in the multispectral imager to select a wavelength, in the hyperspectral imager it is used to vary the OPD for all the wavelengths whose energy is integrated at the CCD camera pixels.

The electric field, E, for emissions at one wavelength, λ, or wave number, $k = 1/\lambda$, radial frequency, ω, and amplitude, A, can be written as

$$E(k) = A(k)\, \mathrm{e}^{-\mathrm{i}(2\pi kx - \omega t)} \qquad [1]$$

and the intensity which is measured is the magnitude squared of the electric field:

$$I = E^{*}E \qquad [2]$$

Emissions from fluorescent sources are usually noncoherent and composed of a range of wavelengths, even with only one dye present. The intensity measured at a pixel is the superposition of many wave numbers. After traveling a certain OPD, L, the intensity integrated over all wave numbers may be written as

$$I_{\mathrm{pxl}}(L) = \frac{1}{2}\left[\int I(k)\mathrm{d}k + \int I(k)\cos(2\pi kL)\mathrm{d}k\right] \qquad [3]$$

or

$$I_{\mathrm{pxl}}(L) = C_0 + C_1 \cdot \mathrm{Re}\{\mathrm{FT}[I(k)]\} \qquad [4]$$

where FT stands for Fourier transform.

The inverse transform is

$$\int I_{\text{pxl}}(L)\, \mathrm{e}^{-\mathrm{i}2\pi k'L}\, \mathrm{d}L = C_0 \int \mathrm{e}^{-\mathrm{i}2\pi k'L}\, \mathrm{d}L$$
$$+ C_1 \int\int I(k)\mathrm{e}^{-\mathrm{i}2\pi kL'}\, \mathrm{d}k \cdot \mathrm{e}^{\mathrm{i}2\pi k'L}\, \mathrm{d}L \qquad [5]$$

$$\int I_{\text{pxl}}(L)\mathrm{e}^{-\mathrm{i}2\pi k'L}\, \mathrm{d}L = C_3 + C_4 I(k)$$

The integration constant, C_3, can be ignored for large wave numbers, as in the case of the visible range, since C_3 will be proportional to the inverse of the wave number. The intensity as a function of wave number is then

$$I(k) = C_5 \int I_{\text{pxl}}(L)\mathrm{e}^{-\mathrm{i}2\pi kL}\, \mathrm{d}L \qquad [6]$$

For absolute quantitative measurements, the constant C_5 can be determined by calibration with a known emission. For discrete measurements made by changing the optical path length, then the working formula is

$$I_{\text{pxl}}(k) = \sum_{L_i} I_{\text{pxl}}(L_i)\mathrm{e}^{-i2\pi kL_i}\, \Delta L \qquad [7]$$

This is the discrete Fourier transform which must be preformed for each pixel to determine the intensity as a function of wavelength. Per the Nyquist sampling tenets:

$$\Delta L = \frac{1}{(N/2)\Delta k} \quad \text{or} \quad \Delta k = \frac{2}{N\Delta L} \qquad [8]$$

where Δk is the discrete wave number interval and N the number of different OPDs. Sagnac interferometers can make rotations which give at least a change of $\Delta L = \lambda/4$, so substituting in eqn [8]:

$$\Delta k = k_1 - k_0 = \frac{1}{\lambda_1} - \frac{1}{\lambda_0} = \frac{\Delta\lambda}{\lambda^2 + \lambda\Delta\lambda} = \frac{8}{N\lambda} \qquad [9]$$

$$\frac{\Delta\lambda}{\lambda^2} \approx \frac{8}{N\lambda} \quad \text{or} \quad \frac{\Delta\lambda}{\lambda} \approx \frac{8}{N}$$

A 5% resolution in wavelength (or 20 nm at 400 nm) requires $N = 160$ and, of course, a 2.5% resolution requires $N = 360$. Therefore, about 400 OPD changes and acquisitions of the intensity fringes will give calculated resolutions of less than 10 nm in wavelength.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Software techniques for displaying data will be described in the next section.

Grating-Dispersion Hyperspectral Imaging

A hyperspectral imaging microscope collects the complete visible emission spectrum from a microscope slide by using a grating spectrometer to provide the wavelength dispersion. This microscope correlates spatial and spectral information with minimal use of optical filters. A hyperspectral microscope system is composed of a standard epi-fluorescence microscope (Olympus IX70) equipped with objectives lens (including 1.25 × and 10 × Plan Fluorite, 40 × Plan Apochromat Dry, and 60 ×, and 100 × Plan Apochromat Oil), as well as a set of standard filter cubes (Chroma Technology) to allow for traditional visualization of dyes through the eyepiece. (Prototype typical components are listed in parentheses.) Multiple excitation sources can be utilized including standard 100 W mercury and xenon arc lamps, tungsten brightfield illumination, a 532 nm solid state laser (Brimrose), a helium neon laser (Uniphase), an argon ion laser (Uniphase), and a pulsed-doubled nitrogen dye laser (Laser Science, Inc.) (Figure 1b).

One side port of microscope is optically coupled to a spectrograph (Acton Research SpectraPro® 556i). A narrow entrance slit into the spectrograph allows for only one line (ΔX) of the image to be captured at a time. A 50-micron slit permits 80% of maximum throughput for the spectrometer. The spectrograph separates the polychromatic light illuminated from this single line into its light components. Normal operation uses a grating with 50 grooves/mm and a 600 nm-blaze wavelength, which allows wavelengths of approximately 400–780 nm to be measured. Different spectrograph gratings can be used for special applications from near UV to near IR. The center wavelength and grating can be shifted under computer control. The output of the spectrograph is then captured with an air/liquid-cooled CCD camera (Photometrics-Quantix) and an optional image intensifier (Video Scope International VS4-1845). The camera has a resolution of 1536 × 1024 pixels with a depth of 12 bits per pixel. The 1536 pixels are aligned in the spatial direction to allow the greatest spatial resolution. During normal operation the wavelength resolution is 'binned' to 128 pixels, sufficient to resolve ~8 different spectra in a single excitation scan. Higher resolution can be employed for special applications.

A line of excitation light can be provided by a line generator when laser excitation is used or by a combination of slits and cylindrical lenses when using traditional lamps or burners (mercury or xenon). This line of light illuminates the sample in the same position that maps the resulting emission light through the imaging spectrograph and onto the camera. Some advantages of illumination with a line of light include decreased sample bleaching, reduced background due to scattering, and efficient use of available excitation power. A traditional bandpass, a low-pass dichroic filter, or a linear variable filter (Raynard Corp.) may be employed to select or restrict the excitation wavelength during specific applications.

Linear Variable Filter (LVF)

The linear variable filter (LVF) (Raynard Corporation) can be used to replace the standard excitation filter cube to provide a capability for continuous excitation wavelength band selection. The LVF is $1'' \times 2.25''$ and is coated such that it continuously filters light from ~400 nm to ~800 nm along its length. Custom filters that span other wavelengths are available. The coating and therefore wavelength is uniform along a short distance, making it an ideal method for selecting an excitation wavelength band to paint across the slide, excite the slide's contents, and map the contents' emission into the slit of the imaging spectrograph. The bandwidth passed is about 21 nm wide. The LVF introduces ~50% attenuation of the intensity.

Spatial Resolution

The spatial resolution of a grating spectrograph Hyperspectral Imager is set by the size of the pixels of the CCD camera in the y- direction and the microscope system magnification. However, differing from the interferometric system, the spatial resolution in the x- direction depends on the spectrometer slit width and the microscope system magnification. As before, let the x- and y- dimensions of the CCD pixel be δx and δy, respectively, and the magnification be M. Let w_s be the slit width of the spectrometer, and note than the slit width is always going to be larger than a few δx. Then the spatial resolution in the y-direction is again $2\delta y/M$, but the spatial resolution in the x- direction is now $2w_s/M$. Again, for CCD pixel sizes of 10 microns square, binning together two pixels in the y-direction, and an overall magnification of 60, the y- resolution is 0.66 microns, as it was for the interferometer method. For a slit width of 50 microns, the x- resolution is 1.67 microns, independent of the number of pixels binned in the x-direction. The control software moves the stage the right number of steps to minimize the overlap or skipped areas. The slit width can be reduced, but at a loss of throughput or increased exposure time. The control software can compensate for different spatial resolution to display the image as 1:1.

Field of View

For N pixel rows (y- dimension) and C pixel columns (x- dimension), the field of view will be $N\delta y/M$ for the length of the area along the y-axis. If the spectrometer slit length is less than the height of the CCD, then the length is just h_s/M, where h_s is the slit length. We orient our CCD array so that the 1536-pixel dimension is along the y-axis. The y-view is thus 256 microns as in the interferometric example. To achieve the same field of view as the Sagnac method imager, we would need to take 103 stage movements and acquisitions. This highlights the difference in the two hyperspectral methods. The grating method collects the complete spectrum on each acquisition with no spectrometer adjustment but needs to step down the slide to produce the full spatial image. The interferometer method collects the complete spatial image on each acquisition with no position adjustment but needs to step the interferometer through a range of optical path differences to acquire multi-spectral content.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution depends on spectrometer instrument function which includes the slit width, the dispersion of the grating, the CCD pixel size, and number of pixels binned together. The dispersion for normal operation was determined to be ~40 nm/mm or 0.4 nm/pixel in the 10 micron CCD pixel example. For example, binning of 8 pixels yields a spectral resolution of 3.2 nm at all wavelengths. To achieve this spectral resolution, the interferometer method would require 1000 OPD increments and acquisitions.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Although the spatial resolutions differ considerably in the two methods, as long as the dye peaks are separated by the spectral resolution spacing, the spectral decomposition algorithms will find all dyes within the spatial resolution area. We have demonstrated that the analysis algorithms can resolve 5 dyes with peaks separated by more than 6 nm in various combinations and known quantity ratios against the dye standards. The program successfully identifies the dyes present with a 91% accuracy in the quantities.

Software

Software (*Hyper*Scope, a 'C' language program) controls all the hardware components for acquisition.

It includes features for analyzing and displaying the resultant $X-Y$ images and spectral information for the grating spectrometer Hyperspectral Imager. During a scan, individual pictures are taken in the $Y-\lambda$ plane, while the stage is incremented in the X-direction. A collection of such $Y-\lambda$ scan files, which can total from 50–1000, is merged to build an image cube (Figure 2). Another advantage of the single-line acquisition mode is that time-dependent features can be monitored continuously. The software was designed to permit repeated collection of a single line at variable time points. The utility of this feature has been demonstrated by capturing images of fluorescent objects moving through UV-transparent capillaries.

The software was designed to display the resulting $X-Y$ plane extracted in various ways from the image cube. A graph of the emission spectrum for any pixel within that image may be viewed by clicking on any pixel of interest in the $X-Y$ image (Figure 3c). A standard linear curve-fitting algorithm is used to determine the contribution of individual dyes to the measured emission spectrum. The software also features a windowing technique (integration of the emission spectra over a fixed window) which emulates standard filtering. An overlay feature within the software allows the curve fitting results to be displayed in false color and compared by viewing the composite for multiple fluorophores in a single image. Additionally, it permits the contribution of a fluorophore to be emphasized for visualization purposes by either displaying the contribution of a single fluorescent signature in the absence of the others that were present in the image or by scaling individual display intensity values up or down. Overlapping pseudo-colors assigned to represent each fluorophore (either by curve fitting or by windowing) may be displayed as added, averaged, maximum or minimum values. For each new fluorophore–excitation source combination, the characteristic emission spectra is acquired and added to a library of available spectra that can be used as components in the spectral decomposition. This permits the user to introduce new fluorophores at any time. Moreover, background spectra (defined as a region of interest by the user) can be acquired and used as one element of the decomposition. This background component can be turned off in the false color view (as any of the spectral components may be) resulting in enhanced noise reduction and visual contrast improvement.

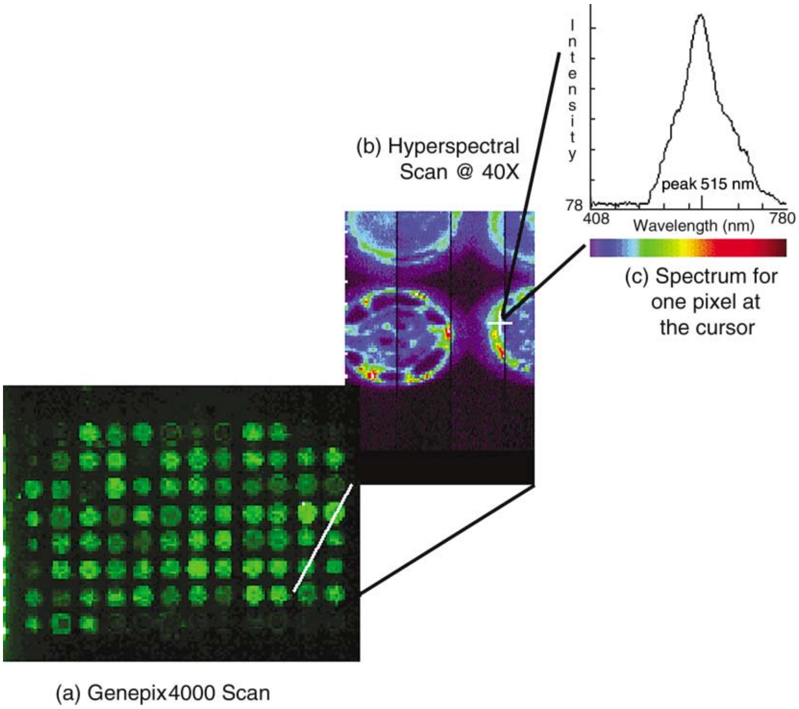

Figure 3 The benefit of spectral imaging is demonstrated in a spatial high-resolution scan of an expression microarray. (a) A portion of the scan of an expression microarray on a microscope slide using the Genepix4000 scanner. It records the intensity emitted by the Cy3 dye when excited with a 532 nm laser. (b) A portion of the scan of the same microarray with the Hyperspectral Imaging Microscope using a 40X objective and Hg excitation lamp. False color is displayed to indicate the intensity variations. (c) A *Hyper*Scope computer program display of the spectrum recorded when the cursor is placed over one pixel in a scan.

Description of Acquisition

In order to acquire a full image, a slide is placed on the microstage mover (Ludl Electronics) that allows for precise movement of the sample across the image acquisition plane. The user can position the stage via a joystick or via software to move to predetermined coordinates. The image is then scanned, with the number of scans being selected at the time of the image acquisition. Furthermore, the sample can be scanned at one magnification or resolution and then rescanned at another. Each $Y-\lambda$ image is saved as a single $Y-\lambda$ file for each increment along the sample. Thus, each image file corresponds to the excitation line of light and the emission collection region that maps through the imaging spectrograph to the CCD camera. Saving individual files permits unlimited $Y-\lambda$ image acquisition at the same Y coordinate, valuable for time-dependent studies. This feature also provides some robustness by enabling a partial $X-Y$ image to be displayed should acquisition of one or more Y scans fail for any reason. After acquiring each image, the stage is moved automatically, one or more steps in the X direction, to acquire emissions at the next line and generate the next $Y-\lambda$ file. Therefore, only the fluorescent probes in a line at a single X-location are examined at each acquisition time. Step sizes can be adjusted by the software to give: (i) the best resolution; (ii) the best 1:1 image; (iii) the minimum overlap; or (iv) the fast acquisition time.

After a variable number of rows of associated emission have been captured (typically 50–1000 scans), an image cube is loaded into memory comprising all the $Y-\lambda$ data files. A graph of the spectrum from any pixel (or bin) in the $X-Y$ plane may be displayed in a separate window by clicking on the cube (Figure 3c). A typical acquisition scan of 250 $Y-\lambda$ files, which are used to construct a $250 \times 768 \times 128$ ($\times$ 12 bit) image cube, consists of approximately 100 Mb of information.

MicroArray Scanning and Customized Excitation

Another use of this technology is to analyze spotted microarrays. The Hyperspectral Imager has been compared to commercial two-color scanners with comparable sensitivity at approximately 0.5 fluors per square micron (Figure 3). Increased sensitivity (500–10 000 gain) can be provided by using a microchannel plate amplifier (Model VS4-1845, Videoscope International, Ltd.) installed between the CCD camera and the imaging spectrograph.

Since the hyperspectral imaging microscope collects images that reflect only the $Y-\lambda$ plane, it contains the potential to use a single line of light as a source for image illumination. Therefore, lights systems being developed, which are capable of generating a line of light composed of variable or complex wavelengths, could be employed to precisely control excitation. The incorporation of such an excitation source in hyperspectral imaging microscopy could be useful in many clinical and basic research applications.

See also

Environmental Measurements: Hyperspectral Remote Sensing of Land and the Atmosphere. **Imaging:** Interferometric Imaging.

Further Reading

Anguiano A (2000) Fluorescence in situ hybridization (FISH): Overview and medical applications. *Journal of Clinical Ligand Assay* 23: 33–42.

Berland K, Jacobson K and French T (1998) Electronic cameras for low-light microscopy. *Methods in Cell Biology* 56: 19–44.

Bernheim A, Vagner-Capodano AM and Couturier J (2000) From cytogenetics to cytogenomics of tumors. *S-Medical Science* 16: 528–539.

Bezrookove V, Hansson K, van der Burg M, *et al.* (2000) Individuals with abnormal phenotype and normal G-banding karyotype: improvement and limitations in the diagnosis by the use of 24-colour FISH. *Human Genetics* 106: 392–398.

Boland MV and Murphy RF (1999) Automated analysis of patterns in fluorescence-microscope images. *Trends in Cell Biology* 9: 201–202.

Cardullo RA and Alm EJ (1998) Introduction to image processing. *Methods in Cell Biology* 56: 91–115.

Garini Y, Katzir N, Cabib D, Buckwald RA, Soenksen D and Malik Z (1996) Spectral bio-imaging. In: Wang XF and Herman B (eds) *Fluorescence Imaging Spectroscopy and Microscopy*, pp. 87–122. New York: John Wiley and Sons.

Garini Y, Macville MV, du Manoir S, *et al.* (1996) Spectral Karyotyping. *Bioimaging* 4: 65–72.

Gothot A, Grosdent J and Palulus J (1996) A strategy for multiple immunophenotyping by image cytometry: Model studies using latex microbeads labeled with seven Streptaviden-bound fluorochromes. *Cytometry* 24: 214–225.

Gribble SM, Roberts I, Grace C, Andrews KM, Green AR and Nacheva EP (2000) Cytogenetics of the chronic myeloid leukemia-derived cell line K562: Karyotype clarification by multicolor fluorescence *in situ* hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenetics* 118: 1–8.

Han H, Gao X, Su J and Nie S (2001) Quantum-dot-tagged microbeads for multiplexed optical coding of bimolecules. *Nat Biotech.* 19: 631–635.

Huebschman ML, Schultz RA and Garner HR (2002) Characteristics and capabilities of the hyperspectral imaging microscope. *Eng in Med and Bio*. 21: 104–117.

Inoue T and Gliksman N (1998) Techniques for optimizing microscopy and analysis through digital image processing. *Methods in Cell Biology* 56: 63–90.

Jain AK (1997) *Fundamental of Digital Image Processing*, pp. 80–128. New Jersey: Prentice-Hall.

Jalal SM and Law ME (1999) Utility of multicolor fluorescent in situ hybridization in clinical cytogenetics. *Genet Med*. 1: 181–186.

Johannes C, Chudoba I and Obe G (1999) Analysis of X-ray-induced aberrations in human chromosome 5 using high-resolution multicolour banding FISH (mBAND). *Chromosome Research* 7: 625–633.

Kruse F (1993) The Spectral Image Processing System (SIPS) – Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environment* 44: 145–163.

Lu YJ, Morris JS, Edwards PAW and Shipley J (2000) Evaluation of 24-color multifluor-fluorescence *in-situ* hybridization (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47. *Chromosome Research* 8: 127–132.

Malik Z, Cabib D, Buckwald RA, Talmi A, Garini Y and Lipson SG (1996) Fourier transform multipixel spectroscopy for quantitative cytology. *Journal of Microscopy* 182: 133–140.

Press WH, Flannery BP, Teukolsky SA and Vetterling WT (1986) *Numerical Recipes, the Art of Scientific Computing*, pp. 546. New York: Cambridge University Press.

Raap AK (1998) Advances in fluorescence in situ hybridization. *Mutat. Res. Fundam. Mol. Mech. Mutagen* 400: 287–298.

Sakamoto M, Pinkel D, Mascio L, *et al.* (1995) Semiautomated DNA probe mapping using digital imaging microscope: II. system performance. *Cytometry* 19: 60–69.

Schrock E, duManoir S, Veldman T, *et al.* (1996) Multicolor spectral karyotyping of human chromosomes. *Science* 273: 494–497.

Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspecteral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

Schwartz S (1999) Molecular cytogenetics: show me the colors. *Genet Med*. 1: 178–180.

Speicher MR, Ballard SG and Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12: 368–375.

Wang YL (1998) Digital deconvolution of fluorescence images for biologists. *Methods in Cell Biology* 56: 305–315.

Zhao L, Hayes K and Glassman A (2000) Enhanced detection of chromosomal abnormalities with the use of RxFISH multicolor banding technique. *Cancer Genet Cytogenetics* 118: 108–111.

Imaging Through Scattering Media

A C Boccara, Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We are often faced with the problem of seeing through scattering media such as smoke, fog, clothes, blood, tissues, ceramics, walls, etc. Tremendous effort has been made to introduce techniques which will allow us to overcome the intrinsic difficulty of selecting ‘good’ photons which are able to provide valuable information on structures which appear fuzzy or, for most of the time, completely hidden.

Although a number of concepts and techniques that we will discuss in this article could be applied to a large number of scattering samples, we will restrict ourselves to biological tissues.

Studying light propagation through biological tissues is becoming more and more popular among physicists, biologists, engineers, and doctors. Indeed this research domain appears to be growing rapidly as seen by the number of published articles, conferences, and international meetings. Optical techniques are attractive and are starting to compete with other well-established techniques (magnetic resonance imaging, X-ray imaging, ultrasonic imaging, positron emission tomography, etc.), because they are nonionizing, nondestructive, they can reach high resolution and are usually much cheaper. Optical techniques reveal an optical contrast which is valuable in providing morphologic as well as functional information.

Most biological tissues are weakly absorbing in the deep red and near infrared region of the spectrum; the main difficulty in performing imaging arises from the high level of scattering. Figure 1 shows the picture of a small flash light bulb placed in the mouth: it appears fuzzy and red.

In Figure 2 we consider various photon trajectories between a short pulsed source S and a fast detector D.

Unscattered photons are called 'ballistic' photons, their number decreases exponentially, and they are completely damped after one or two millimeters.

Then there are scattered photons which propagate close to the ballistic trajectory reaching the detector; these are named 'snake-like' photons and can be used for imaging purposes.

Most of the energy which reaches D is related to the scattered photons which may exhibit very long trajectories (typically ~10 times the distance S–D).

The description of various experimental methods used to overcome the difficulty introduced by scattering will be associated with the various photon families that we have mentioned: ballistic, snake-like, and scattered photons. We will first give a few orders of magnitude of the parameters which will dictate the way light interacts with the tissues. Let us emphasize that these parameters also carry the information related to the contrast mechanism. The ultimate goal is to realize an 'optical biopsy', noninvasive but as close as possible to the regular biopsy followed by histopathology.

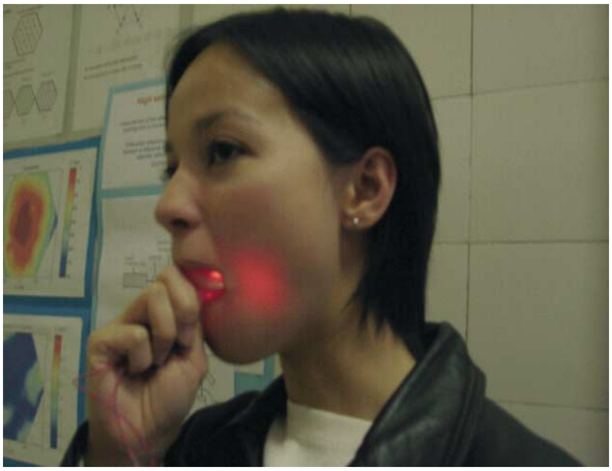

Figure 1 Biological tissues are highly scattering but the absorption is low in the red (as seen in this digital visible photography) and the near infrared.

Optical Parameters

Scattering Cross-Section

As shown in Figure 3, biological tissues are constituted of structures which exhibit various characteristic scales: from a few nanometers in membranes to about ten micrometers for an entire cell.

As we know from scattering theories these various scales correspond to various scattering properties. For small particles ('small' means much smaller than the optical wavelength λ) we are in the Rayleigh regime which is based on the evaluation of a time-varying dipole induced by a uniform time-varying (at the light frequency scale) electric field. This dipole radiates in the far field and its scattering cross-section, which is the ratio of the scattered power to the incoming irradiance for a sphere of radius r, is given by

$$\sigma = \frac{\langle P \rangle}{I} = \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{8\pi}{3} r^6 k^4 \qquad [1]$$

n being the refractive index of the sphere divided by the index of the surrounding medium and $k = 2\pi/\lambda$. The scattering diagram for such particles is isotropic if we use unpolarized light. This cross-section is much smaller than the 'geometrical' section S of the particle; for instance, if $n = 1.5$, $a = 10$ nm, $\lambda = 0.6$ μm, $\sigma = 1510^{-8}$ μm^2, which is much smaller than the actual section (3×10^{-4} μm^2).

For larger particles the sphere (or the ellipse) interacting with a plane polarized wave is rather complicated because we are not restricted to the dipolar approximation (uniform field), but induced multipolar moments and their radiating fields have to be taken into account. This theory was proposed by Mie and can be found in optics textbooks. As an example, Figure 4 shows the scattering cross-section of a micron-size particle as a function of the

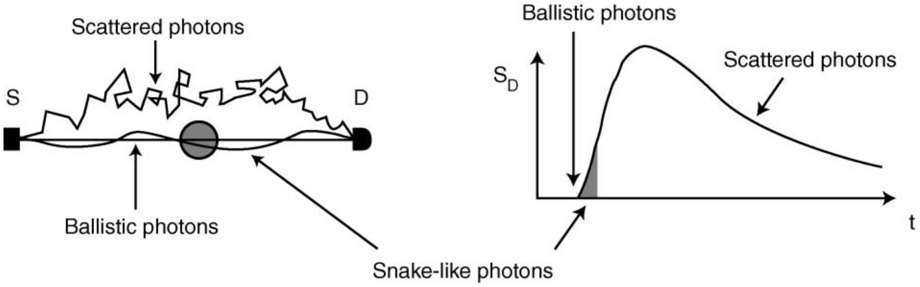

Figure 2 Between an ultrafast light source S and a detector D can be distinguished various classes of photons which can be used for imaging.

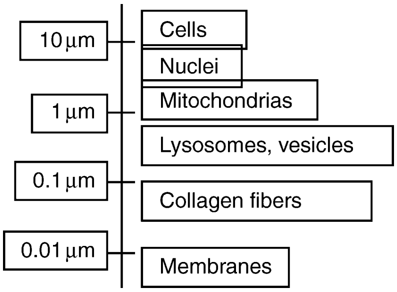

Figure 3 Various sizes of scatterers may be found in cell structures.

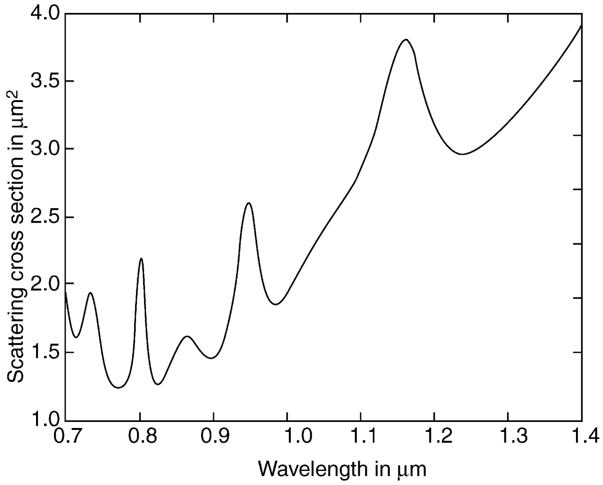

Figure 4 In the Mie scattering regime resonances are observed when varying the wavelength (here a 1 μm sphere, refractive index: 2).

wavelength: a number of 'resonances' can be observed.

The scattering diagram is no longer isotopic, it exhibits lobes, and forward scattering is much higher than back-scattering. Scattering is characterized by a scattering coefficient, μ_s, which is the inverse mean free path l_s of photons between scattering events:

$$l_s = 1/\mu_s = \sigma\rho \qquad [2]$$

where σ is the scattering cross-section and ρ the number of scatterers by unit volume.

The scattering coefficients μ_s of tissues are found within the range 100–1000 cm^{-1}. After a path length L, the number of ballistic photons will be damped by a factor $\exp(-L/l_s)$, so tissue imaging cannot rely only on ballistic photons as they survive only to a path length larger than about one mm. One needs to consider in more detail the scattering properties of the scatterers.

The dominant sizes of the scatterers in tissue (cells) demand consideration of the scattering in the Mie regime. Such a scattering event does not result in isotropic scattering angles. Instead, the scattering in tissue is biased in the forward direction.

Anisotropic scattering is quantified in a coefficient, g, which is defined as the mean cosine of the scattering angle θ, $p(\theta)$ being the probability of a particular scattering angle:

$$g \equiv \langle\cos(\theta)\rangle \equiv \frac{\int_0^{\pi} p(\theta)\cos(\theta)\sin(\theta)\mathrm{d}\theta}{\int_0^{\pi} p(\theta)\sin(\theta)\mathrm{d}\theta} \qquad [3]$$

For isotropic scattering, $g = 0$. For complete forward scattering, $g = 1$. In tissues, g varies from 0.7 to 0.95.

Photon migration, especially in the so-called diffusion regime, is often based on isotropic scattering such as for phonons (heat) or charge carriers scattering. For diffusion-like models, it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ_s', and obtain equivalent results using:

$$\mu_s' = \mu_s(1 - g) \qquad [4]$$

One can consider $l_s' = 1/\mu_s'$ as the length over which the incoming photon loses the memory of its initial direction, μ_s' is usually called the transport-corrected scattering coefficient.

Absorption Cross-Section

The absorption coefficient, μ_a (cm^{-1}) represents the inverse mean path length of photons before absorption. $1/\mu_a$ is also the distance in a medium at which intensity is attenuated by a factor of $1/e$ (Beer's Lambert law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In the visible range, the blood absorption is relatively high, whereas it is much weaker in the near infrared. Water absorption is low in the visible and NIR region and increases above 1 μm, with tissues turning completely opaque above 2 μm. Thus, for greatest penetration of light in tissue, wavelengths in the 700–1300 nm spectrum are used. This region of the spectrum is called 'the therapeutic window'. Different spectra of chromophores allow one to separate the contribution of varying functional species in tissue such as quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation.

Finally, when light is propagating in scattering media of thickness L, the damping of the incoming energy involves a combination of its absorption and scattering properties, and is given by

$$\exp(-L/a) \qquad \text{with} \qquad a^2 = 1/3\mu_a(\mu_s' + \mu_a) \qquad [5]$$

This formula reflects the fact that the photon's pathlength is increased by scattering, leading to an increase of the damping, which is related to absorption.

These few examples provide a 'static' view of light behavior in scattering media. Studying in a more detailed way, the spatial and temporal distribution of light, is more difficult.

Photon Migration in Scattering Media

Photon migration models have been inspired by other fields, such as astrophysics, soft matter physics, and neutronics, where physical media are random in both space and time. To avoid difficulties linked to the microscopic complexity such investigations rely on a statistical basis.

The most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla}\cdot(D\vec{\nabla}\Phi(\vec{r},t)) - \mu_a\Phi(\vec{r},t) = \frac{1}{c}\frac{\partial\Phi(\vec{r},t)}{\partial t} - S(\vec{r},t) \qquad [6]$$

where $\vec{r}$ and t are spatial and temporal coordinates, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients by the formula:

$$D = \frac{1}{3[\mu_a + \mu_s']} \qquad [7]$$

The quantity $\Phi(\vec{r},t)$ is called the fluence, defined as the power divided by the unit area. This equation does not reflect the scattering angular dependence; the use of μ_s' instead of μ_s ensures that isotropic scattering has been reached. Indeed, for the use of the diffusion theory which implies anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient; $S(\vec{r},t)$ is the source term.

The diffusion equation has been solved analytically for different types of measurements such as reflection and transmission modes, assuming that the optical properties remain invariant throughout the tissue. To incorporate the finite boundaries, the method of images has been used.

The diffusion approximation equation in the frequency-domain is the Fourier transform of the time-domain equation. This leads to the equation:

$$\vec{\nabla}\cdot(D\vec{\nabla}\Phi(\vec{r},\omega)) - \left[\mu_a + \frac{i\omega}{c}\right]\Phi(\vec{r},\omega) + S(\vec{r},\omega) = 0 \qquad [8]$$

Here the time variable is replaced by the frequency ω which is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the so-called 'photon density wave' in addition to a DC component:

$$\begin{aligned}\Phi(\vec{r},\omega) &= \Phi_{AC}(\vec{r},\omega) + \Phi_{DC}(\vec{r},0)\\ &= I_{AC}\exp(i\theta) + \Phi_{DC}(\vec{r},0) \qquad [9]\end{aligned}$$

where θ is the phase shift of the diffusing wave whose wavelength is:

$$2\pi\sqrt{\frac{2c}{3\mu_s'\omega}} \qquad [10]$$

Although analytical techniques provide a rapid understanding of most phenomena involved in static or dynamic situations, real biological samples are much more complex than homogeneous isotropic media. In these cases, direct and inverse algorithms map the spatially varying optical properties.

Monte Carlo simulations are very popular nowadays; a number of useful programs are available on the websites of various groups. Results may include time or polarization dependence of the photons along their paths.

The other method used in tissue optics is the random walk theory (RWT) on a lattice. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Imaging in the Ballistic Regime: Taking Advantage of Various Coherence Properties

Shallow Tissue Imaging Through Selection of Ballistic Photons

When imaging thin (less than 2 mm) tissue samples, it is possible to use photons which have not been scattered, called the ballistic photons which can be used to form high-resolution images in the same manner as if no scattering had taken place.

Their number, however, decreases exponentially with propagation distance and this ballistic signal is usually hidden by multiply scattered photons that obscure the image. For relatively shallow tissue depths it is possible to use various spatial filtering techniques to block the multiply scattered photons and there is much current research that aims to extend optical imaging and biopsy to depths of a few mm.

Confocal microscopy rejects much of the scattered light through spatial filtering and has been shown to form 3-D images at tissue depths of up to 200–300 μm, both *in vitro* and *in vivo*, which corresponds to about 50 dB of damping for the ballistic photons.

Two-photon fluorescence microscopy is particularly interesting in this regime since it ensures that all the detected fluorescence photons originate from the desired image plane. This technique also ensures a better penetration (using IR excitation) and a better resolution (nonlinear response) than single photon confocal microscopes.

Imaging to much greater tissue depths does not appear to be practical using optical microscopy and spatial filtering alone and more sophisticated techniques must also be employed to discriminate in favor of the ballistic photons. These techniques exploit either the fact that the ballistic signal photons retain spatial coherence of incident light or the fact that the ballistic photons keep the optical phase memory that multiply diffused photons have lost.

One of the most successful ballistic light-imaging techniques based on the latter filtering approach is optical coherence tomography (OCT) which uses coherence gating using low temporal coherence light and an interferometric detection. Coherent detection is used to discriminate against the multiply scattered photons that scatter back into the trajectories of ballistic photons. The ballistic light signal is measured by interference with a reference beam and the resulting pattern is detected with high sensitivity using homodyne or heterodyne detection. By using low coherence length sources and matching the time-of-flight of the reference beam signal to that of the desired ballistic light, OCT can acquire depth-resolved images in scattering media such as biological tissues. For powers in the mW range for tissue irradiation in the infrared, the unscattered ballistic (or more precisely single back scattered) component of the light will be limited by the shot noise detection limit after propagating through approximately 25 scattering mean free paths (MFP) of a scattering medium (or more than 100 dB). This corresponds to a 'typical' tissue thickness of about 1 mm tissue depth for the reflection geometry which is potentially useful for clinical applications such as screening for skin cancer, or for retinal examination.

OCT has proved clinically useful as a means of acquiring depth-resolved images in the eye and is being developed for application in strongly scattering biological tissue. Although the image acquisition requires pixel-by-pixel scanning, the use of new high-power superluminescent diodes in the 100 mW range or of ultrafast lasers to provide high average power, low coherence length radiation (sub-two cycle lasers) has resulted in an *in vivo* OCT system providing real-time imaging.

There are a number of other approaches to coherent imaging, including whole-field 2-D acquisition techniques which remove the need to scan transversely pixel by pixel.

Whole-field imaging using a temporally and spatially incoherent source and a well-balanced interferometric microscope intrinsically provides higher image acquisition rates and can take advantage of inexpensive spatially incoherent, broadband sources such as high-power LEDs, and white light thermal sources.

As an example, Figure 5 shows Linnik-type white light interferometer with optical path modulation and Figure 6 an image of an *ex-vivo* rat retina where one can recognize the main feature expected from histology.

Holography, which works in the Fourier space rather than in the object or image space, is also an interferometric technique able to discriminate between ballistic and scattered photons. Since the development of electronic holography, which takes advantage of available arrays of detectors with multimillion pixels, this approach is rapidly growing.

Real-time 3-D imaging systems based on photorefractive holography in multi-quantum-wells devices, which can potentially acquire depth-resolved images at thousands of frames per second, have been demonstrated.

These ballistic light techniques can extend the depth of tissue imaging to the mm range when

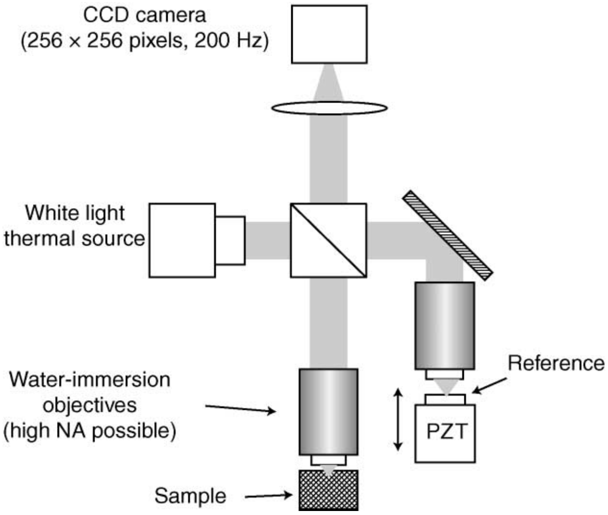

Figure 5 The full-field OCT setup based on a Linnik interferometer. The PZT actuator is used to modulate the path difference between the two arms.

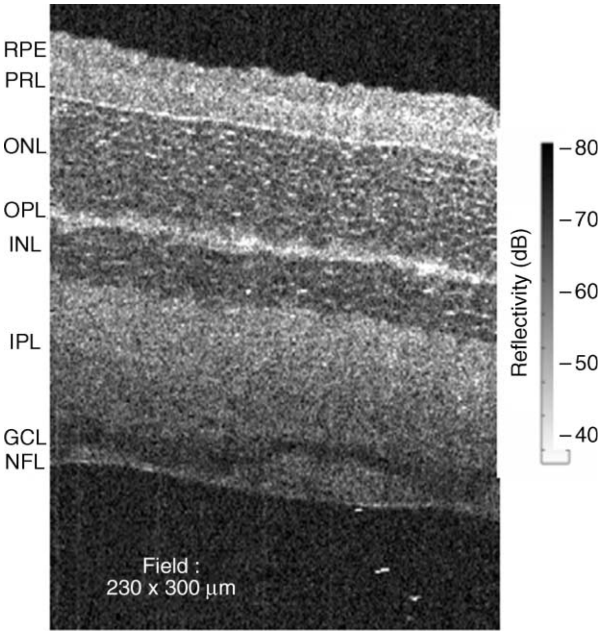

Figure 6 OCT rat retinal image: retinal pigmented epithelium. PRL: photoreceptor layer. ONL: outer nuclear layer. OPL: outer plexiform layer. INL: inner nuclear layer. IPL: inner plexiform layer. GCL: ganglion cell layer. NFL: nerve fiber layer.

imaging in reflection. When it is necessary to image through cms, rather than mms of tissue, however, there will be no detectable ballistic light signal and one must extract useful information from the scattered photons.

Imaging in the Snake-Like Regime: Taking Advantage of Time (or Frequency) Gating

The first approach is to use time gating to select the earliest arriving scattered light, which will provide the most useful image information.

For moderate tissue depths one can exploit the fact that biological tissue is highly forward scattering, with most photons being only slightly deviated from their original direction upon each scattering event. After a few cm of tissue, there can be a significant number of photons that have followed reasonably well-defined ‘snake-like’ paths about the original direction through the tissue, arriving at the detector after the ballistic light but before the fully diffuse photons whose trajectories are effectively randomized. This early arriving light can be selectively gated using ultrafast lasers and a variety of detectors such as streak cameras, time-gated optical image intensifiers, and time-gated fast photon counting systems, which provide picoseconds time gates.

An alternative, sometimes cheaper, approach to temporal gating is the use of (high) frequency signal discrimination (which is the Fourier transform of the pulse experiment). The periodic solution of the diffusion equation with a periodic source term gives rise to the so-called ‘photon density waves’. Such diffusive waves, like thermal waves, are used in the near-field regime (their amplitudes are damped by $\exp(-2\pi)$ after propagating over a single wavelength path) and they can reveal subsurface structures through their refractive and diffusive behavior. Coherence gating and polarization gating can be used when the early arriving light still retains some degree of coherence of the incident light. Indeed polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, from the photons which experience multiple scattering resulting in a complete loss of memory of their initial polarization state.

As an example, Figure 7 shows the image of sub-mm light sources (LEDs) taken through 1 cm of polystyrene solution ($\mu_s' \sim \mu_s \sim 10\ \text{cm}^{-1}$) by selecting photons which kept the memory of their initial circular polarization (here a few less than 1%). One can see the sources quite clearly.

For many important biomedical applications, such as mammography or imaging brain function, it is necessary to penetrate through 5–10 cm of tissue, after which all the detected signal is diffuse. Detecting earlier arriving photons can provide more information but, if the time window is too narrow, the detected signal becomes too weak to use with acceptable data acquisition times.

This problem is, however, being addressed with some success using statistical models of photon transport, with different degrees of approximation ranging from full Monte Carlo simulation of photon propagation to the diffusion equation. The approach is to address the inverse problem, i.e. to measure the scattered light signal as comprehensively as is practical and to calculate what distribution of absorption and scattering properties would have produced the measured signal. This provides a means to quantify the optical properties of biological tissue, averaged over a volume corresponding to a particular distribution of photon paths, and to form relatively low-resolution tomographic images.

Multiplying the number of sources and detectors and correlating the various intensity distributions over the whole structure under examination gives, as expected, better precision in localizing the size and position of the local structure to be identified.

Despite the multiple input and multiple output approaches, real-life situations with the previously

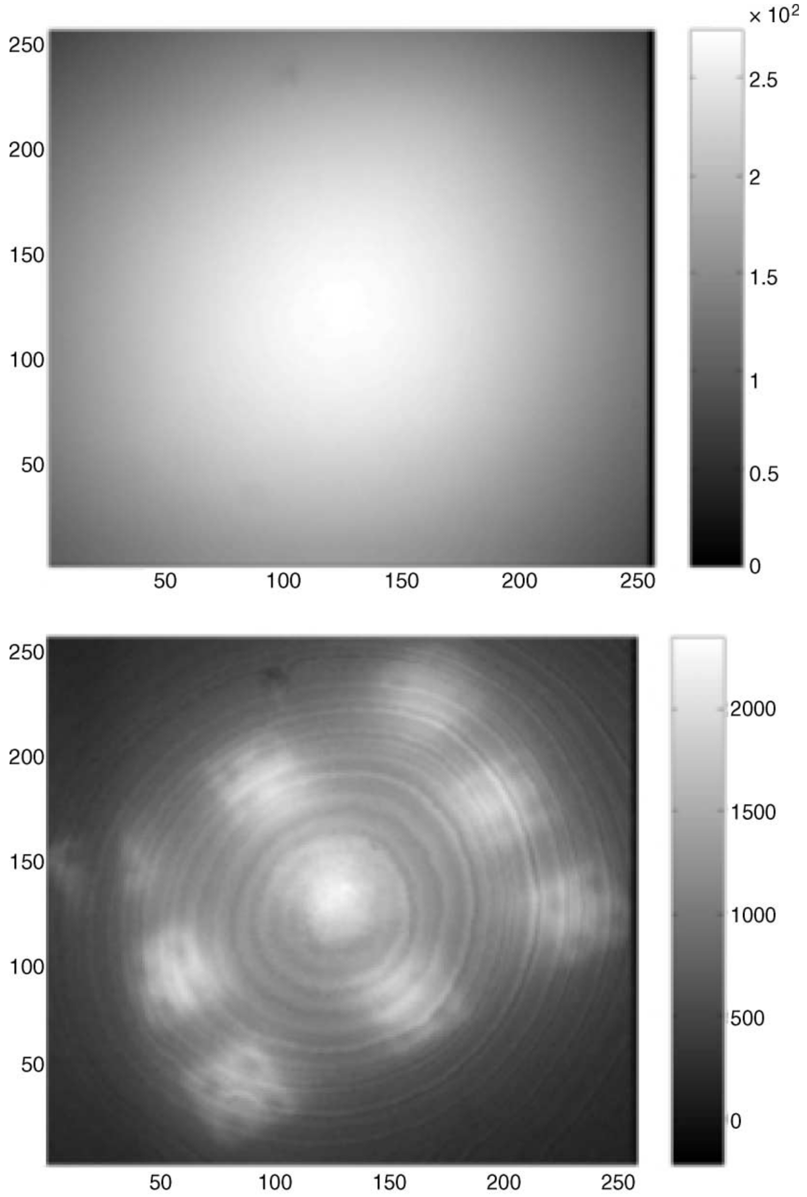

Figure 7 Polarization imaging: here the degree of circular polarization is measured in the upper image. The lower image shows the field without polarization filtering.

mentioned problem of the inhomogeneity of the optical properties of human tissues means that the results are still poor in term of resolution and subcentimeter structures can hardly be observed in breast imaging, for instance.

One important exception to this limited success in finding precisely the position, the size, and the optical properties of hidden structures is the case of 'activation' studies. This corresponds to a 'differential' situation in which an unknown structure is locally changing with time. Here the goal is not to reach a full description of the unknown structure but only to characterize the local changes induced by the activation.

The main field of application of this approach is certainly in brain activation studies. A matrix of light emitters (usually at two wavelengths which differentiate oxy- and deoxy-hemoglobin, for instance 780 and 840 nm) is coupled to a matrix of light detectors. The signals are usually balanced and any local change which breaks the symmetry of the scattered photon paths induces a differential signal, easy to detect, because there is no background signal before activation takes place.

Moreover it has been demonstrated that for weak perturbations (weak variations of the optical parameters), the inverse problem can be rigorously solved. This geometrical selection provides a cheap and simple setup with rather spectacular results. Sometimes a modulation is added and the phase shift of the photon density wave is enhanced by the differential approach.

Imaging in the Multiple Scattering Regime: Taking Advantage of all the Photons

Light Only: Transillumination and Fluorescence Imaging

The optical systems used for CW measurements are rather simple and require only the alignment of a number of light sources and detectors. Compared to time-gated techniques due to multiple scattering a much stronger path length distribution is present in the signal and the result is a loss of localization and resolution. In the so-called transillumination geometry, collimated detection is used to isolate multiply scattered photons close to the normal emergence angle.

Direct imaging with the required resolution using CW techniques in very thick tissues, such as breast tissue, has not been established even with sophisticated multi-sources/multi-detectors combinations (typically a few tens of sources and detectors at three different wavelengths).

Despite rapid progress in the modeling of light behavior in scattering media it has not yet been demonstrated that CW imaging can provide images with suitable resolution in thick tissue and time-dependent measurement techniques discussed earlier are dominantly used.

For fluorescence imaging, one has to account for the strong attenuation of light as it passes through tissue. Fluorescent dyes (porphyrin, indo-cyanine, and more recently quantum dots) have been developed that are excited and re-emit in the ‘biological window’ at near-infrared (NIR) wavelengths, where scattering and absorption coefficients are lower, which is favorable for deep fluorescence imaging. Fluorescent intensity in deep tissues depends on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, and on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. In order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon path lengths which depend on all these different parameters. The amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore.

Using fluorescence contrast has the unique feature of carrying two spectroscopic contrasts related to both excitation and fluorescent photons.

The next two techniques take advantage of another way to handle the optical information by coupling optics and acoustics.

Coupling Light and Sound: Acousto-Optic and Photo-Acoustic

Although the two approaches that we will now describe are different in their basic principles as well as in the experimental setups, their names are similar – opto-acoustics (sometimes called photo-acoustics) and acousto-optics.

Opto- (or Photo-)Acoustics

In this technique a pulsed (or modulated) electromagnetic beam irradiates the structure under examination. Despite their tortuous paths (about 10 to 100 cm) photons propagate through the volume in a few nanoseconds. Local absorbing centers then absorb the electromagnetic energy, they experience a fast thermal expansion, and become sources of acoustic waves.

The optical problem of tomographic reconstruction is then transformed into a simpler problem of acoustic reconstruction of source distribution because the speed of the sound is much slower than the speed of the light. More precisely we can compute the signal generated by a short light pulse impinging on a scattering medium which contains an absorber (thickness e) as can be seen on Figure 8, when $e\mu_a \ll 1$.

The amount of power per unit surface which is deposited in the absorbing slice is $E\mu_a$ (E is the fluence in W/m^2). When the light pulse duration τ is short enough, despite the propagation in the scattering medium ($\tau \ll e/v$ where v is the speed of the sound),

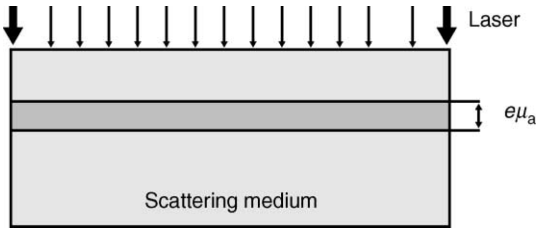

Figure 8 Photo acoustic signal generation: 1-D model.

we find a variation of the mass by volume unit:

$$\Delta\rho \sim \rho\beta\Delta T \sim \rho\frac{\beta E\mu_{\mathrm{a}} e}{C_{\mathrm{v}} e} \qquad [11]$$

The corresponding pressure is:

$$\Delta\rho = \frac{v^2}{\gamma}\Delta\rho \qquad \text{or} \qquad \Delta\rho \sim \rho\frac{v^2\beta E}{C_\rho}\mu_{\mathrm{a}} \qquad [12]$$

Although this approach is new, compared to the purely optical transillumination approach, it has led to a large number of studies in which time gating brings new information to the accurate determination of the US emitter depth. It has been applied to breast tumor examination to complement standard imaging techniques such as echography and X-ray mammography. Recently this technique has demonstrated submillimeter resolution for *in vivo* blood distribution in rat brain imaging.

Acousto-Optic

Here a DC single frequency laser is used to illuminate the sample. For a typical biological sample a few cm thick the trajectories are distributed between the sample thickness value and more than ten times this value. Due to the good temporal coherence of the source all the scattered photons are likely to interfere when they emerge from the sample volume. These interferences are seen as a speckle field which can be observed at a suitable distance from the sample surface.

As seen in Figure 9, adding an ultrasonic field will mainly induce a small (smaller than the optical wavelength) displacement of the scatterers (it also induces local variation of the refractive index) which will cause a speckle modulation at the ultrasonic frequency.

If the zone irradiated by the ultrasonic field is optically absorbing, less photons will emerge from this zone than from a nonabsorbing one and thus less modulation will be seen when scanning across the sample volume. This modulation can be detected by selecting a single speckle grain and averaging over a number of uncorrelated speckle distributions to obtain a suitable average value (for instance with latex particles). This is not always possible, in particular because the overall geometry is rather stable in a semisolid tissue. Using a single detector which receives many speckle grains gives less contrast. A multiple detector and a parallel processing of many speckle grains improve greatly the speed and the signal-to-noise ratio: a typical image of absorbing spheres, embedded in a scattering phantom, is shown in Figure 10.

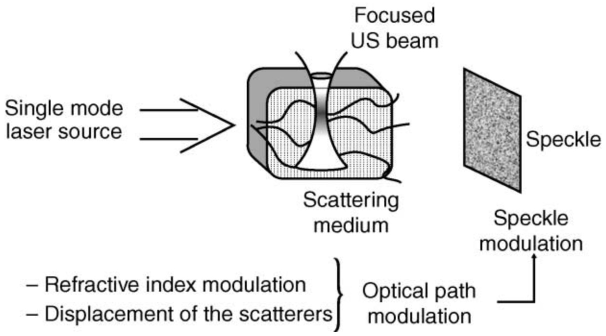

Figure 9 Acousto-optic imaging principle: photons are tagged by the acoustic field.

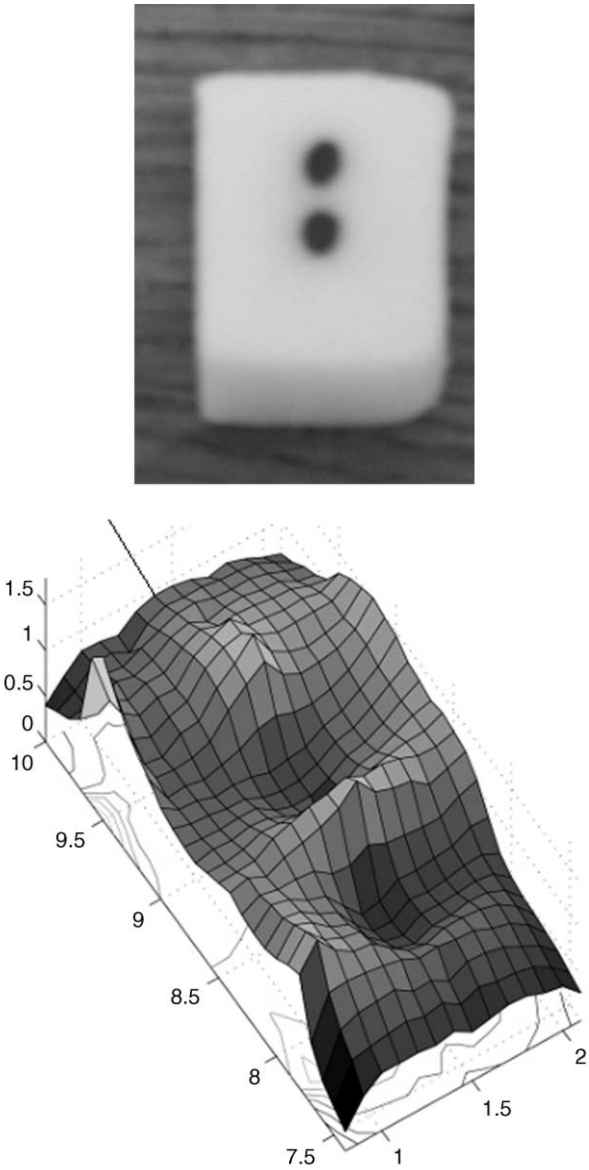

Figure 10 Acousto-optic imaging (lower image) of two absorbing spheres embedded in a scattering medium (upper image shows a section of the gel sample).

Conclusion

The complexity associated with multiple scattering in turbid media, such as tissues of the human body, prevents an easy use of light as a routine tool for in-depth examination, whereas superficial examinations have always been sources of diagnostics.

Optical techniques will obviously progress to better characterization of tissues' optical properties, improvement of laser properties and detection techniques, as well as new mathematical approaches of the inverse problem.

We believe that this field of research is still open to new ideas, and new experimental schemes leading to new breakthroughs in the delicate problem of using light in highly scattering media.

Acknowledgements

I would like to warmly thank all the research staff (Arnaud Dubois, Benoit Forget, François Ramaz, Vincent Loriette) and PhD students (Michael Atlan, Kate Grieve, Gael Moneron, Juliette Selb) at ESPCI whose results have been used to illustrate this article. I also acknowledge discussions with my colleagues Pr Paul French (Imperial College London) and Dr Amir Gandbjbakhche (NIH Bethesda).

See also

Coherence: Coherence and Imaging. **Holography, Techniques:** Digital Holography. **Imaging:** Interferometric Imaging. **Interferometry:** White Light Interferometry. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Boccara AC and Fink M (eds) (2001) *Optical and Acoustical Imaging of Biological Media*. Compte Rendu de l'Académie des Sciences, Physics and Astrophysics série IV tome 2, no. 8. Paris: Elsevier.

Born M and Wolf E (1997) *Principles of Optics*. New York: Oxford University Press.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Hecht E and Zajac A (1979) *Optics*. Reading, MA: Addison-Wesley.

Sebbah P (ed.) (2001) *Waves and Imaging through Complex Media*. Dordrecht: Kluwer Academic Publishers.

Van de Hulst HC (1981) *Light Scattering by Small Particles*. New York: Dover.

Van Tiggelen B and Skipetrov S (eds) (2003) *Wave Scattering in Complex Media: From Theory to Applications*. Dordrecht: Kluwer Academic Publishers.

Infrared Imaging

K Krapels and R G Driggers, US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

Published by Elsevier Ltd.

Introduction

All objects above 0 Kelvin emit electromagnetic radiation associated with the thermal activity on the surface of the object. For terrestrial temperatures (around 300 Kelvin), objects emit a good portion of the electromagnetic flux in the infrared part of the electromagnetic spectrum. The visible band spans wavelengths from 0.4 to 0.7 micrometers (infrared engineers typically use micrometer/μm for wavelength rather than the nanometer or wavenumber more commonly used in other fields). Infrared imaging devices convert energy in the infrared portion of the electromagnetic spectrum into images that can be created by the human eye. The unaided human eye cannot image infrared energy because the lens of the eye is opaque to infrared radiation.

The infrared spectrum begins at the red end of the visible spectrum where the eye can no longer sense energy. It spans from 0.7 μm to 100 μm. The infrared spectrum is, by common convention, broken into five different bands (this may vary according to the application). The bands are typically defined as: near infrared (NIR) from 0.7 to 1.0 μm; short wavelength infrared (SWIR) from 1.0 to 3.0 μm; mid-wavelength infrared (MWIR) from 3.0 to 5.0 μm; long wavelength infrared (LWIR) from 8.0 to 14.0 μm; and far infrared (FIR) from 14.0 to 100 μm. These bands are depicted graphically in Figure 1, which shows the atmospheric transmission for a 1-kilometer horizontal ground path for a 'standard' day in the United States. These types of transmission graphs can be tailored for any conditions, using sophisticated atmospheric models such as MODTRAN (from www.ontar.com). Note that

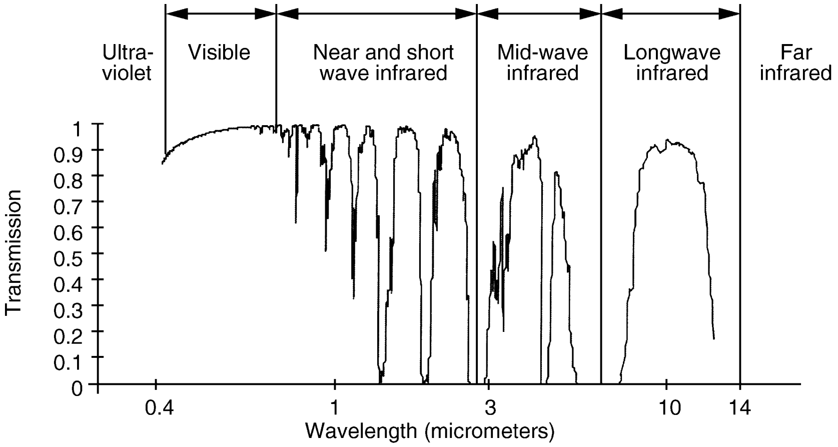

Figure 1 Visible and infrared spectra.

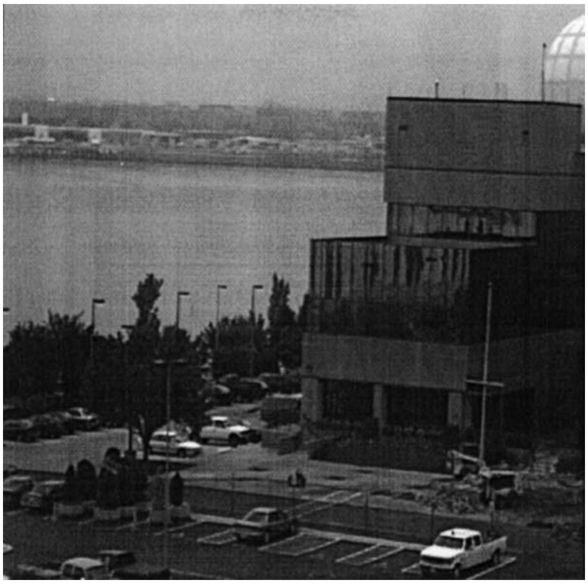

Figure 2 Visible image–reflected energy.

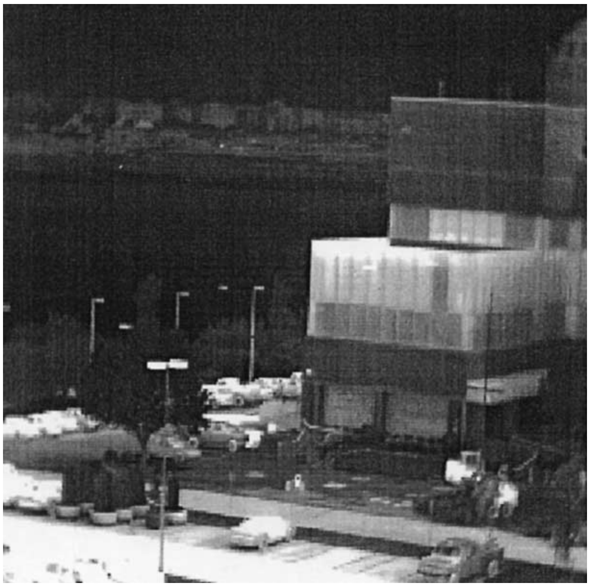

Figure 3 Infrared image–self-emitted energy.

there are many atmospheric 'windows', such that an imager designed with such a band selection can see through the atmosphere.

The primary difference between a visible spectrum camera and an infrared imager is the physical phenomonology of the radiation from the scene being imaged. The energy used by a visible camera is predominantly reflected solar, or some other illuminating energy, in the visible spectrum. The energy imaged by infrared imagers (Forward Looking Infrared, commonly known as FLIRs) in the MWIR and LWIR bands, is primarily self-emitted radiation. From Figure 1, the MWIR band has an atmospheric window in the 3 to 5 μm region and the LWIR band has an atmospheric window in the 8 to 12 μm region. The atmosphere is opaque in the 5 to 8 μm region, so it would be pointless to construct a camera that responds to this waveband.

Figures 2 and 3 are images that show the difference in the source of the radiation sensed by the two types of cameras. The visible image (Figure 2) is light that was provided by the Sun, propagated through Earth's atmosphere, reflected off the objects in the scene, traversed through a second atmospheric path to the sensor, and then imaged with a lens and a visible band detector. A key here is that the objects in the scene are represented by their reflectivity characteristics. The image characteristics can also change by any change in atmospheric path or source characteristic change. The atmospheric path characteristics from the Sun to the objects change frequently, because the Sun angle changes throughout the day and also weather and cloud conditions change. The visible imager characterization model is therefore a multipath problem that is extremely difficult.

The LWIR image given in Figure 3 is obtained primarily by the emission of radiation by objects in the scene. The amount of electromagnetic flux depends on the temperature and emissivity of the objects. A higher temperature and a higher emissivity correspond to a higher flux. The image shown is 'white-hot' or the whiter a point in the image the higher the flux leaving the object. It is interesting to note that trees have a natural self-cooling process since a high temperature can damage foliage. Objects that have absorbed a large amount of solar energy are hot and are emitting large amounts of infrared radiation. This is sometimes called 'solar loading'.

The characteristics of the infrared radiation emitted by an object are described by Planck's black-body law in terms of spectral radiant emittance

$$M_\lambda = \varepsilon(\lambda)\frac{c_1}{\lambda^5(e^{c_2/\lambda T}-1)}\ \mathrm{W/cm^2\ \mu m} \qquad [1]$$

where c_1 and c_2 are constants of 3.7418×10^4 W μm^4/cm^2 and 1.4388×10^4 μm K. The wavelength, λ, is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A black-body source is defined as an object with an emissivity of 1.0, so it is a perfect emitter. Source emissions of black-bodies at typical terrestrial temperatures are shown in Figure 4. Often, in modeling, the terrestrial background temperature is assumed to be 300 K. The source emittance curves are shown for other temperatures for comparison purposes. One curve corresponds to an object colder than the background and two curves correspond to temperatures hotter than the background.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to the left (shorter wavelengths) as the body temperature increases. If the temperature of a black-body is increased to that of the Sun (5900 Kelvin), the peak of the spectral shape would decrease to 0.55 μm or green light. This peak wavelength is described by

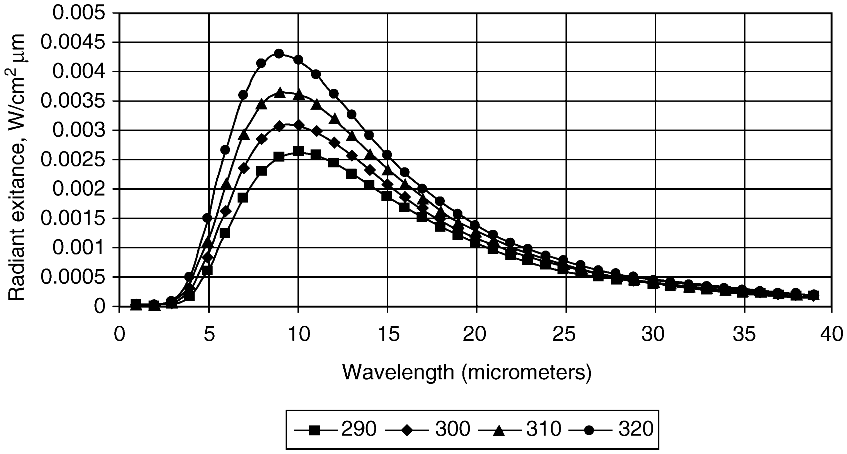

Figure 4 Planck's blackbody radiation curves for four temperatures from 290 K to 320 K.

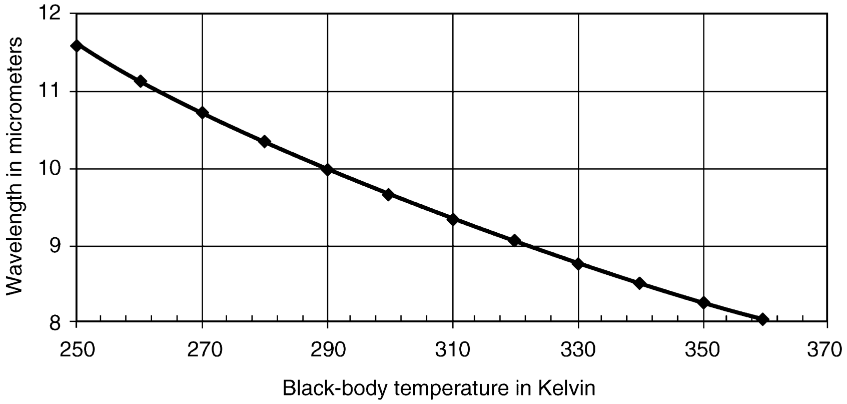

Figure 5 Location of peak of black-body radiation, Wien's Law.

Wien's displacement law

$$\lambda_{\max} = 2898/T, \ \mu\text{m} \qquad [2]$$

Figure 5 shows the peak wavelength as a function of temperature through the longwave region. It is important to note that the difference between the black-body curves is the 'signal' in the infrared bands. For an infrared sensor, if the background is at 300 K and the target is at 302 K, the signal is the difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. However, in the visible band this is not the case. For example, consider the case of a white target on a black background. The black background is generating no signal, whereas the white target is generating a maximum signal (given that the sensor gain has been adjusted). Dynamic range may be fully utilized in a visible sensor. In the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux never has a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors. Table 1 summarizes this characteristic numerically for the 300 K background and targets shown in Figure 4.

A typical infrared imaging system scenario is depicted in Figure 6. The scene consists of two major components, the target and the background. In an IR scene, the majority of the energy is remitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. All of these contributors, which are not the target of interest, essentially reduce one's ability to discriminate the target. So, at the entrance aperture for the sensor, the target signature is reduced from the values observed at the target. Then, the signal is degraded by the optics and scanner (as applicable) of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for straightforward modifications to a performance model for changes in the sensor or environment when performing trade-off analyses.

Table 1 Signal/dynamic range limitation in IR bands

Blackbody temperature	*Exitance in 8–12 μm band (w/cm^{-2})*	*Contrast*
290 ($-10\Delta T$ K target)	1.25E − 02	−8.48%
300 (background)	1.48E − 02	
310 ($+10\Delta T$ target)	1.74E − 02	7.97%
320 ($+20\Delta T$ target)	2.02E − 02	15.39%

History of Infrared Imaging

Night vision systems began to be developed extensively during World War II. Satisfying the requirement to image at night was approached with two different methodologies. The first method was image intensification (I^2). This method involves amplification of any small light that was available and displaying it directly to the eye. Typically I^2 devices are sensitive to visible light and a small portion of the short-wave IR (SWIR, 0.7–3.0 μm) band. They are often classified (along with visible band imagers) as electro-optical (EO) imagers. With an I^2 imager, there must be some source of illumination for them to function well (as little as starlight is sufficient

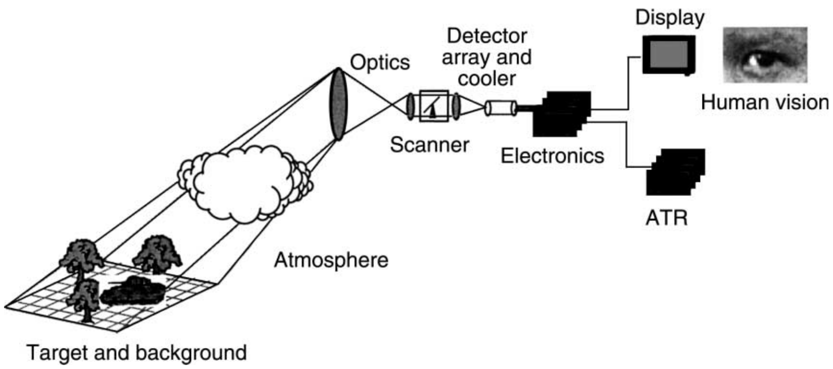

Figure 6 Typical infrared imaging scenario.

Table 2 Applications of infrared imaging

Category	*User*	*Application*
Military	Intelligence analyst	Intelligence surveillance and reconnaissance
	Pilot, driver	Navigation/pilotage
	Gunner, weapons officer	Targeting/fire control
	Air defense	Search and track
Commercial	*Civil government*	
	Police	Surveillance, fugitive tracking
	Fire	Rescue, hot-spot location
	Environmental	
	EPA	Emission tracking
	Interior department	Resource management
	NIMA\USGS	Mapping
	NOAA\MWS	Weather forecasting
	Industrial	
	Manufacturers	Machine vision
	Maintenance	Non-destructive testing
	Medical	
	Doctors	Diagnostic imaging

for approx 20/40 acuity with current systems). The second type of night vision device is the infrared or thermal imager (also known as FLIRs). FLIRs are sensitive to the radiation in either the mid-wave IR (MWIR, 3–5 μm) or long-wave IR (LWIR, 8–14 μm) bands.

The first infrared cameras used photographic film that was sensitive to infrared wavelengths. Intelligence and reconnaissance imaging from aircraft drove the infrared system development as a night augmentation to existing visible cameras. These imagers used either continuous moving film or stationary film, like an ordinary camera. A slit aperture was used in the continuous moving film cameras and aircraft motion provided the scan of the scene along the film to build a continuous image. The introduction of television, as electronic imaging in the visible spectrum, led to the development of electronic imagers in the infrared band. Electronic infrared imagers were not restricted to looking down and using aircraft motion for scan, hence these devices came to be known as FLIRs (Forward Looking InfraRed). Today, an electronic infrared imager may be called a FLIR or a thermal imager, with the trend being to use FLIR to describe military applications.

Early FLIRs were often accompanied by illumination devices. They were not very sensitive compared to modern FLIRs so active illumination was necessary. Infrared searchlights or illuminators were used to get higher signal-to-noise ratios. These types of FLIRs are called 'active' IR imagers, while FLIRs that do not use illuminators are 'passive' imagers. Generally, active imagers are used in civil applications and are declining in military use. Broadly, applications for thermal imagers fall into two categories; military and commercial. Table 2 lists some of the purposes and users for modern thermal imagers. The design and performance criteria vary widely for some of the applications.

Types of Infrared Imagers

Infrared imagers are classified by different characteristics: scan type, detector material/cooling requirements, and detector physics. The scan type refers to the method used to construct the electronic image. The camera may use a single detector which is raster scanned over the input scene to build an image. Alternatively, a parallel scan uses a linear array of detectors scanned across a scene to build an image. The latest advances in materials have led to staring arrays of detectors. In a staring system, a detector is present in two dimensions to represent each image pixel and no mechanical motion of the focal plane is necessary to construct the image. There are some hybrid FLIR types that combine the different imaging techniques. Usually this combination leads to an improvement in signal-to-noise ratio or to reduce undersampling.

The second classification, by detector material/cooling requirements, usually describes FLIRs built using differing materials. For example, a typical LWIR FLIR material is mercury cadmium telluride (HgCdTe or MCT). In order to achieve high sensitivity, these devices are cooled to decrease dark current. Usually the cooling is based on liquid nitrogen or a cryogenic cooler, and the detectors operate at 77 K. Another common detector material is indium antimonide (InSb), which is used for MWIR FLIRs and is also cooled. A newer class of infrared cameras is not cooled, being referred to as 'uncooled' FLIRs. These uncooled FLIRs are microbolometers (resistive elements) or pyrometers (capacitive elements) and have less sensitivity than cooled imagers. Typically, the cooled imagers are comprised of photon detectors while the uncooled imagers are based on thermal detectors, the uncooled FLIRs being used in lower-performance applications.

Infrared Imager Performance

There are three general categories of infrared sensor performance characterizations. The first is sensitivity and the second is resolution. When end-to-end, or human-in-the-loop (HITL), performance is required, the third type of performance characterization describes the visual acuity of an observer through a sensor. The former two are both related to the hardware and software that comprises the system, while the latter includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene/environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis.

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range/quantization of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In infrared systems, noise equivalent temperature difference (NETD) is often a first-order description of the system sensitivity. The 3D noise model describes more detailed representations of sensitivity parameters (see Further Reading).

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in infrared systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution.

Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity. Typically, visible band systems have plenty of sensitivity and are resolution-limited, whereas infrared imagers have been more sensitivity-limited. With staring infrared sensors, the sensitivity has seen significant improvements.

Often metrics, such as NETD and MTF, are considered to be separable. However, in an actual sensor, sensitivity and resolution performance are not independent. As a result, minimum resolvable temperature difference (MRT or MRTD) has become the primary performance metric for infrared systems. MRT is a quantitative performance measure in terms of both sensitivity and resolution and a simple MRT curve is shown in Figure 7. The performance is bounded by the sensor's limits and the observer's limits. The temperature difference (or thermal contrast) required to image smaller details in a scene increases with detail size. The inclusion of observer performance yields a single sensor performance characterization, MRT. It describes sensitivity as a function of resolution, and includes the human visual system.

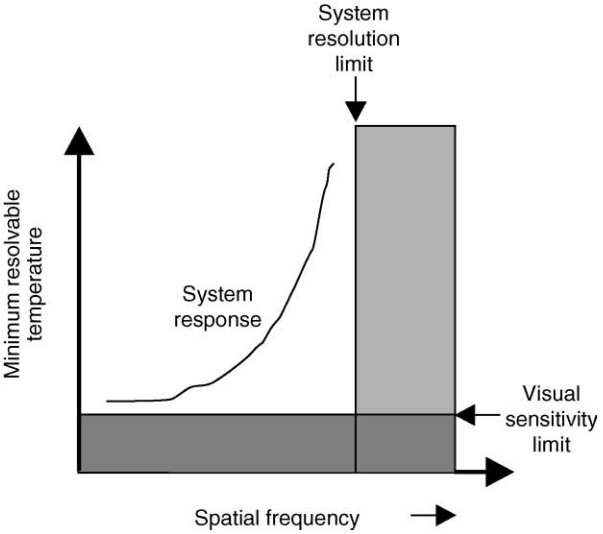

Figure 7 IR imager performance.

Many different sensor characteristics may be considered to increase the fidelity of a sensor model. A list of typical infrared imaging system parameters is given in Figure 8. When considering performance in the system sense, the groupings are often blocks in the system diagram, each with separate transfer functions. This approach works well unless the linear shift invariance assumption is not valid.

General Characteristics of Infrared Imagers

There are a number of imaging characteristics that make infrared systems a little different than conventional visible imaging systems. These include source flux levels, detector charge storage and sensitivity, detector size, diffraction blur, sampling, and uniformity characteristics.

First, we consider the source flux levels. The daytime and night-time flux levels (in photons per second per square centimeter) on Earth in the visible (0.3 to 0.7 μm) is 1.5×10^{17} and around 1×10^{10}, respectively. In the mid-wave (3 to 5 μm), the daytime and night-time flux levels are 4×10^{15} and 2×10^{15} μm where the flux is a combination of

- Illumination
  - Spectral irradiance
- Target
  - Reflectivity
  - Emissivity
  - Size and spatial characteristics
  - Temperature
- Atmosphere
  - Weather conditions
  - Obscurants
  - Transmission
  - Optical transfer function
  - Scattering
- Sensor optics
  - Lens transmissions
  - Mirror reflectivity
  - Filter transmission
  - Aberrations
  - Diffraction
  - Aperture size and shape
- Sensor detector
  - Responsivity
  - Detectivity
  - Noise characteristics
  - Detector angular substense
- Sensor electronics
  - Temporal characteristics
  - Digital and/or analog
  - Filter characteristics
- Display
  - Resolution
  - Brightness and contrast
- Human psychophysics
  - Temporal response
  - Image transfer function
  - Brightness dependence
- ATR or AiTR response
  - Discrimination probability
  - False alarm rate

Figure 8 Factors in infrared sensor modeling.

emitted and solar reflected flux. In the longwave, the flux is primarily emitted where both day and night yield a 2×10^{17} μm frequency, the AMOP drops rapidly to zero. The effect is to extend the prediction of the MRT beyond the half sample rate and introduce a parameter called minimum temperature difference perceived (MTDP) to replace MRT in the range performance estimate and also in the lab to characterize an undersampled system.

The concept of MTDP follows from observation of standard 4-bar images as the pattern frequency passes the imager half sample rate. The image changes from four bars to three, two, and finally one (can be perceived). The phase of the target image on the detector array may need to be adjusted to observe this progression. The standard MRT measurement requires that all four bars be resolvable by the observer during the measurement process. For use in the lab, the MTDP is defined as the minimum temperature difference at which two, three, or four bars can be resolved in the image of the standard 4-bar test pattern by an observer, with the test pattern positioned at optimum phase. The optimum phase is the phase at which two, three, or four bars are 'best perceived'. TRM3 uses the standard definition of MRT as for adequately sampled imagers. The TRM3 Approach Model is implemented if the imager is considered undersampled as defined when the prefilter MTF at the half-sample frequency is larger than 10%. The MTDP equation is given by

$$\mathrm{MTDP}(f) = \frac{\frac{\pi}{2}\mathrm{SNR}_{\mathrm{th}}\Psi(f)}{\mathrm{AMOP}(f)} \qquad [3]$$

where $\mathrm{SNR}_{\mathrm{th}}$ is a threshold signal-to-noise ratio and Ψ is the total system filtered noise. MTDP is used in the same manner as MRTD in range calculations.

Another approach is the triangle orientation discrimination (TOD) threshold. In the TOD, the test pattern is a (nonperiodic) equilateral triangle in four possible orientations (apex up, down, left, or right), and the measurement procedure is a robust 4AFC (four alternative forced-choice) psychophysical procedure. In this procedure, the observer has to indicate which triangle orientation he sees, even if he is not sure. Variation of triangle contrast and/or size leads to a variation in the percentage correct between 25% (complete guess) and 100%, and by interpolation the exact 75% correct threshold can be obtained. A complete TOD curve (comparable to an MRTD curve) is obtained by plotting the contrast thresholds as a function of the reciprocal of the triangle angular size (Figure 9).

The TOD method has a large number of theoretical and practical advantages: it is suitable for undersampled and well-sampled electro-optical and optical

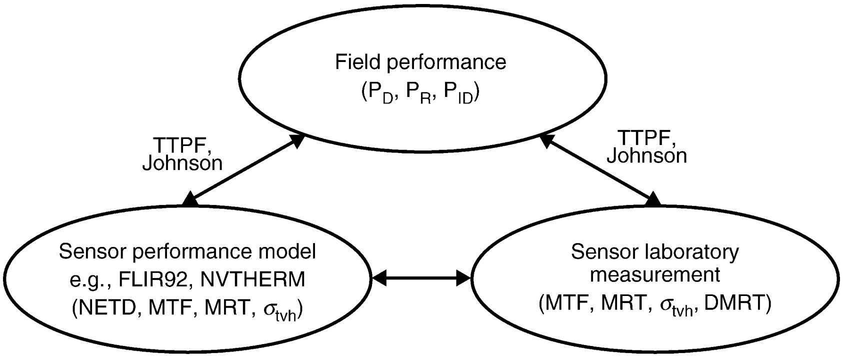

Figure 9 Modeling, testing and performance triangle.

imaging systems in both the thermal and visual domains, it has a close relationship to real target acquisition, and the observer task is easy. The results are free from observer bias and allow statistical significance tests. The method may be implemented in current MRTD test equipment with little effort, and the TOD curve can be used easily in a target acquisition (TA) model such as ACQUIRE. Two validation studies with real targets show that the TOD curve predicts TA performance for under-sampled and well-sampled imaging systems very well. Currently, a theoretical sensor model to predict the TOD (comparable to NVTherm or TRM3) is under development. The lab measurement and field performance appear sound, but the model is not yet available.

Testing Infrared Imagers: NETD, MTF, and MRTD (Sensitivity, Resolution and Acuity)

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. For infrared imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. The system intensity transfer function (SITF) can be used to estimate the NETD, which is the system noise rms voltage over the noise differential output. NETD is the smallest measurable signal produced by a large target (extended source).

Equation [4] describes NETD as a function of noise voltage and the system intensity transfer function. The measured values are determined from a line of video stripped from the image of a test target, as depicted in Figure 10. A square test target is placed before a black-body source. The delta T is the difference between the black-body temperature and the mask. This target is then placed at the focal point of an off-axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Figure 10b. The SITF slope for the scan line in Figure 10c is the $\Delta S/\Delta T$ where ΔS is the signal measured for a given ΔT. The N_{rms} is the background signal on the same line.

$$\mathrm{NETD} = \frac{N_{\mathrm{rms}}[\mathrm{volts}]}{\mathrm{SITF\_Slope}[\mathrm{volts/K}]} \qquad [4]$$

Resolution is a general term that describes the size of resolvable features in an imager's output. With infrared imaging systems, the resolution is described by the system modulation transfer function (MTF). Consider Figure 11 for the determination of MTF, from either point spread function or edge spread function (psf or esf). In Figure 11a, a test target is placed at the focal point of a collimator. The target may be a point, edge, or line as shown in Figure 11b. The IR system images this target. The output of a detector scan, or a line of detectors in the staring array case, is taken for analysis, in Figure 11c. For the edge spread function, the spatial derivative is taken to get the psf. The Fourier transform is then calculated to give the system MTF as shown in Figure 11d. The point spread function is really the impulse response of the system, so a smaller psf is desirable, where a wide MTF is desirable. Such a psf/MTF combination gives a higher resolution.

The MRTD of an infrared system is defined as the differential temperature of a four bar target that makes the target just resolvable by a particular observer. It is a measure of observer visual acuity when a typical observer is using the infrared imager. It results in a descriptor, which is a function not just of a single value. It is a plot of sensitivity as a function of spatial frequency.

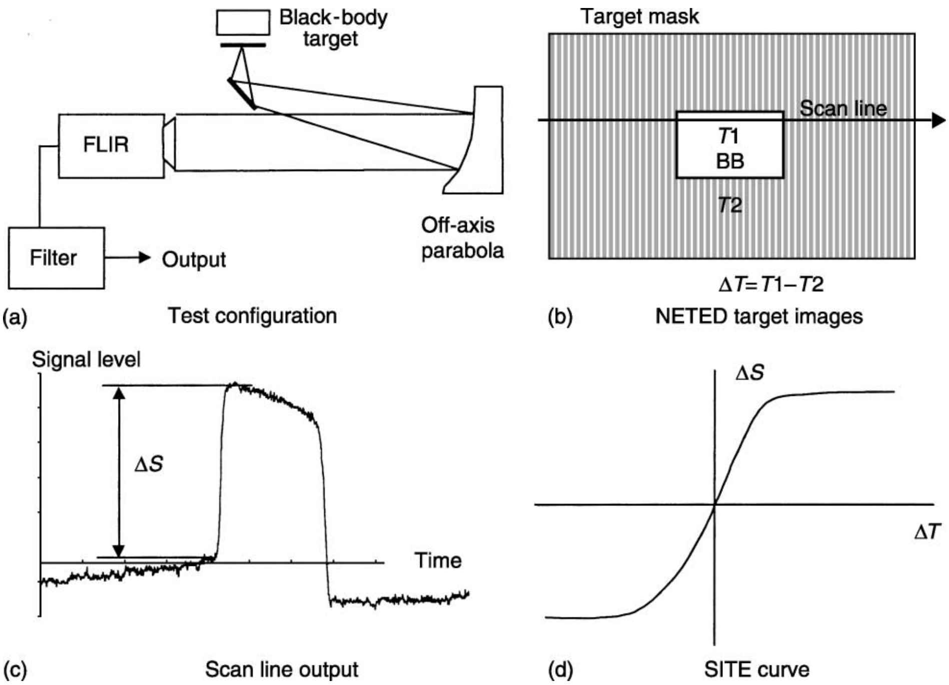
Black-body target
FLIR
Filter
Output
Off-axis parabola
(a) Test configuration
Target mask
Scan line
T1
BB
T2
ΔT=T1−T2
(b) NETED target images
Signal level
ΔS
Time
(c) Scan line output
ΔS
ΔT
(d) SITE curve

Figure 10 NETD measurement.

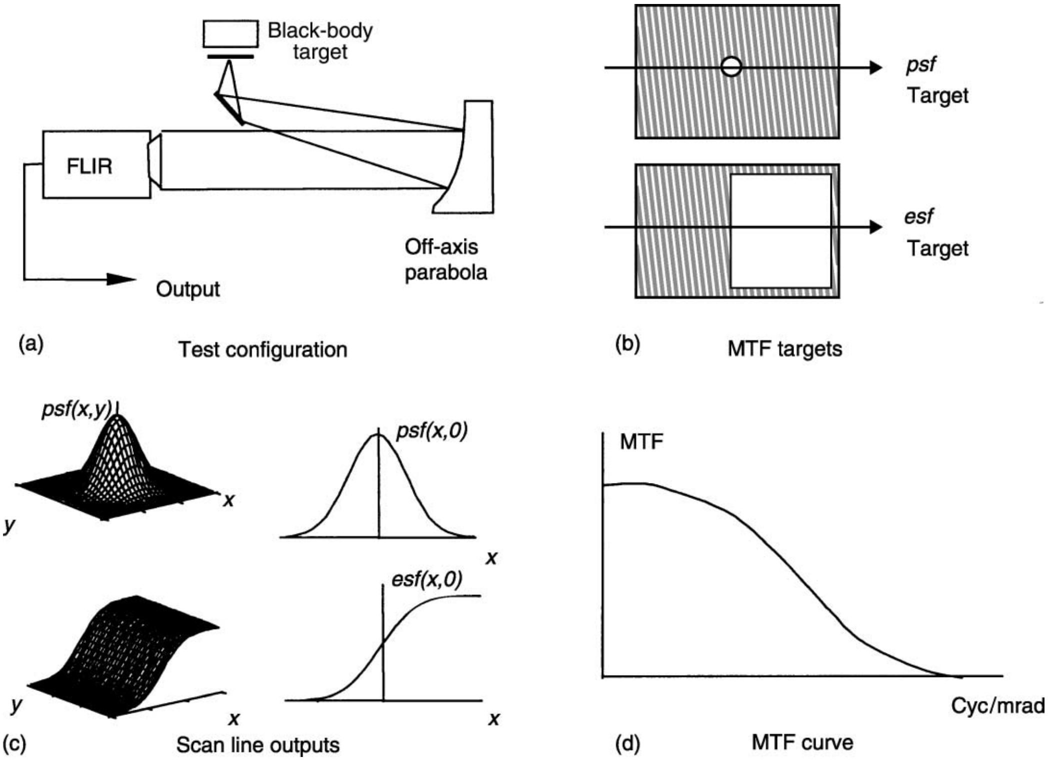
Black-body target
FLIR
Off-axis parabola
Output
(a) Test configuration
psf Target
esf Target
(b) MTF targets
psf(x,y)
x
y
psf(x,0)
x
esf(x,0)
y
x
x
(c) Scan line outputs
MTF
Cyc/mrad
(d) MTF curve

Figure 11 MTF measurement.

This parameter can be extended to field performance using Johnson's criteria.

As a laboratory measurement, it uses a 7:1 aspect ratio bar target. The procedure is shown in Figure 12. A bar target mask, with a black-body illuminator, is placed at the focal plane of a collimator (Figure 12a and 12b). The temperature is increased from a small value until the bars are just resolvable to a trained

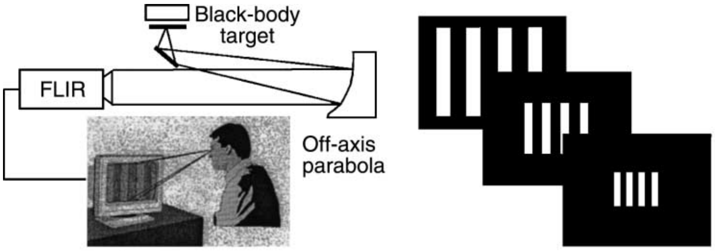

(a) Test configuration (b) MRT targets

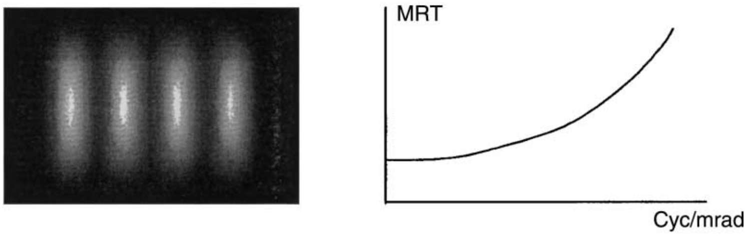

(c) MRT target image (d) MRT curve

Figure 12 MRTD measurement.

observer (Figure 12c). Then the temperature difference is decreased until the bars are no longer visible. These are averaged for a particular target. The same procedure is performed with a negative contrast bar target. The average of the positive and negative values is the MRT for a particular spatial frequency. These data points are plotted and a curve may be fitted to interpolate/extrapolate performance at other than the discrete spatial frequencies of the targets, as in Figure 12d.

Summary

We have provided a general description of infrared imaging systems in terms of characteristics, modeling, field performance, and performance measurement. The characteristics of infrared imagers continue to change. Significant changes have occurred in the past five years, to include the development of higher performance uncooled imagers and ultra-narrow field of view photon systems. Large format detector arrays are commercially available in the mid-wave and are becoming more available in the longwave. Still in the research phase are dual-band focal plane level. At first sight, it appears that the longwave flux characteristics are as good as a daytime visible system; however, there are two other factors limiting performance. First, the energy bandgaps of infrared photons are much smaller than those photons in the visible, so the detectors suffer from higher dark current. The detectors are usually cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%. In the infrared, all terrestrial objects are emitting with an emission temperature of around 300 Kelvin. Typically, a two-degree equivalent black-body difference in photon flux is considered a high contrast target to background flux. The flux difference between two black-bodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in Figure 4. This is the difference in signal that provides an image, so note the difference in signal compared to the ambient background flux. In the longwave, the signal is three percent of the mean flux and in the mid-wave, it is six percent of the mean flux. This means that there is a large flux pedestal associated imaging in the infrared. Unfortunately, there are two problems accompanying this large pedestal. First, photon noise is the dominant noise and is determined by the mean of the pedestal and is compared to the small signal differences. Second, the charge well storage in infrared detectors is limited to around 10^7 charge carriers. A longwave system in a hot desert background would generate 10^{10} charge carriers in a 33 millisecond integration time. Smaller F-numbers, spectral bandwidths, and integration times are used so that the charge well does not saturate. This results in SNR of 10 to 30 times below the ideal. It has been suggested that some on-chip compression may be a solution to the well pedestal problem. In many scanning FLIR systems, the pedestal is eliminated by AC-coupling the detector signals. Infrared focal plane array (IRFPA) readout circuits have been previously designed and fabricated to perform charge skimming and charge partitioning.

Another major difference in infrared systems, from that of visible systems, is the size of the detector and diffraction blur. For infrared photon detectors, typical sizes range from 25 to 50 μm, where visible systems can be fabricated with 3 μm detectors. Also, the diffraction blur for the longwave is more than ten times larger than the visible blur and mid-wave blur is eight times larger than visible blur. Therefore, the image blur, due to diffraction and detector size, is much larger in an infrared system than a visible system. Also, it is common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance.

Finally, both infrared photon and thermal detectors in staring arrays have responsivities that vary dramatically and it is current practice to correct for the resulting nonuniformity (nonuniformity correction or NUC). The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise (especially in static imaging or stabilized target acquisition imaging).

Modeling Infrared Imagers

Otto Schade Sr. developed the earliest imaging system models and performance measures. His work describing television in the 1950s and 1960s, pioneered the way for the characterizations of imaging sensors used by engineers today. His performance measure for still pictures, moving pictures and television systems was based on an observer resolving a three bar periodic target.

John Johnson developed the technique of relating the acquisition of military targets in the field to laboratory measurements of resolvable bar targets. He divided the discrimination tasks into four categories: detection, orientation, recognition, and identification. This technique was termed an ‘equivalent bar pattern approach’. It related performance on a simple test target to performance with complex object targets. Johnson viewed scale models and bar targets in the laboratory against a bland background. The smallest discernible barchart target yielded the maximum resolvable bar pattern frequency. These results, tabulated in Table 3 as cycles across the minimum dimension, were the basis for the discrimination methodology in widespread use today. Note that detection took only one cycle, but as the tasks got more specific or difficult, the requirement went as high as 8 cycles across the target minimum dimension.

Table 3 Johnson’s criteria

Target	*Resolution per minimum dimension*			
Broadside view	*Detection*	*Orientation*	*Recognition*	*Identification*
Truck	0.9	1.25	4.5	8
M-48 Tank	0.75	1.2	3.5	7
Stalin Tank	0.75	1.2	3.5	6
Centurion Tank	0.75	1.2	3.5	6
Half-track	1	1.5	4	5
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier (standing)	1.5	4.8	3.8	8
105 howitzer	1	1.5	4.8	6
Average	1 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Table 4 Moser’s data

Discrimination task	Line pairs/critical dimension
Detect ship	Aperiodic treatment
Classify as combatant	4
Recognize type	10

Recognizing that typical vehicle aspect ratios were somewhat limited, Johnson and Lawson conducted further experiments. Paul Moser conducted some of his own, reanalyzed Johnson and Lawson’s, and developed the concept that this task was related to the average or critical dimension instead of the minimum dimension. This led to the conversion of line pairs to resolution elements, introducing the second dimension. Additionally, he performed a similar experiment on marine targets, which is summarized in Table 4. Lloyd and Sendall developed the MRT metric, which combines both resolution and sensitivity characteristics. This measurement of MRT is described in detail in a following article, but there is a theoretical model that is used to evaluate new sensor designs and concepts. It relates the minimum temperature difference between bar pairs at which they are just resolvable as a function of spatial frequency. MRT describes the infrared imager sensitivity as a function of resolution. More specifically, MRT is a measure of thermal contrast sensitivity as a function of spatial frequency.

Ratches *et al.* developed the NVL (US Army Night Vision and Electronic Sensors Directorate formerly known as the Night Vision Laboratory) Static Performance Model, which predicted the end-to-end performance. It started with the target signature and carried through to the observer. The MRT, a laboratory measurement or a modeled value, was related to the probabilities of discrimination through the target transfer probability function, cumulative percentages related to Johnson’s criteria.

D'Agostino and Webb created the 3D noise model, both spatial and temporal considerations, for analyzing system noise. This technique divided the noise components into manageable and understandable components. It also simplified the integration of these effects into models.

The introduction of focal plane array (FPA) imagers has created a requirement to improve the model performance to account for the difference between scanning arrays and staring arrays. This change has made sampling an important area for improvement. A new model called NVTHERM is a result of the NVESD model improvement effort, which addresses undersampled imaging system performance.

Sensor characterization is seen in three ways: theoretical models, field performance, and laboratory measurements (Figure 9). Theoretical models are developed that describe sensor sensitivity, resolution, and human performance for the purpose of evaluating new conceptual sensors. Acquisition models, and other models, are developed to relate the theoretical models to field performance. This link allows theoretical models to be converted to field performance quantities (e.g., probabilities of detection, recognition, and identification). Field performance is measured in the field so that the theoretical models can be refined and become more accurate with advanced sensor developments. Since field performance activities are so expensive, methods for the direct measurement of sensor performance are developed for the laboratory. Field performance testing of every infrared sensor built, including buy-off, acceptance, and life-cycle testing, is ridiculous and out of the question. Laboratory measurements of sensor performance are developed such that, given these measurements, the field performance of a system can be predicted. Sensor characterization programs require accurate sensor models, field performance estimates with acquisition models, and repeatable laboratory measurements.

There are a few alternatives to the US NVTherm model. One candidate approach to undersampled imager modeling is Germany's TRM3, or the MDTP approach. The impact of undersampling in the image of a 4-bar target can readily be seen by simply observing the change in the image as a function of spatial frequency and phase. The spatial frequency is defined as line pairs per angular extent of the target, and the phase specifies the relative location of the target image to the detector array raster. These effects are obviously not independent of each other, but for each target there can be found an optimal phase where the observer can see the maximum amplitude modulation in the image of the target. MRT measurements in the past have utilized this variation with phase by allowing the observer to optimize the displayed image through target or system line-of-sight changes during the measurement process. For undersampled imagers, TRM3 addresses the problem of the MRT calculation's inability to predict beyond the half sample rate of the sensor, by replacing the MTF in the denominator of the MRT equation with an appropriately scaled term called the average modulation at optimum phase (AMOP). AMOP is the average signal difference in the image of the 4-bar standard pattern, with the test pattern positioned at optimum phase. AMOP oscillates between the presample MTF and the bar modulation. Beyond a frequency of 0.8 times the sample rate or 1.6 times the half sample arrays and quantum well detector systems. These systems will find their place in applications to include both military and commercial sectors.

Further Reading

Bijl P and Valeton MJ (1998) Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* 37(7): 1976–1983.

D'Agostino J and Webb C (1991) 3-D analysis framework and measurement methodology for imaging system noise. *Proceedings of SPIE* 1488: 110–121.

Driggers RG, Cox P and Edwards T (1998) *Introduction to Infrared and Electro-Optical Systems*. Boston, MA: Artech House, p. 8.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 347.

Holst GC (1995) *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 432.

Johnson J (1958) Analysis of image forming systems. *Proceedings of IIS* 249–273.

Johnson J and Lawson W (1974) Performance modeling methods and problems. *Proceedings of IRIS*.

Ratches JA (1973) *NVL Static Performance Model for Thermal Viewing Systems*. USA Electronics Command Report ECOM 7043, AD-A011212.

Schade OH (1948) Electro-optical characteristics of television systems. *RCA Review* IX(1–4).

Sendall R and Lloyd JM (1970) Improved specifications for infrared imaging systems. *Proceedings of IRIS* 14(2): 109–129.

Vollmerhausen RH and Driggers RG (1999) NVTHERM: next generation night vision thermal model. *Proceedings of IRIS Passive Sensors* 1.

Wittenstein W (1998) Thermal range model TRM3. *Proceedings of SPIE* 3436.

Interferometric Imaging

D L Marks, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Interferometric imaging is image formation by measuring the interference between electromagnetic signals from incoherently emitting or illuminated sources. To learn about interferometric image formation using coherent light, *see* **Holography, Techniques:** Overview. To understand what optical coherence is, *see* **Coherence:** Overview. To learn what kinds of sources produce incoherent light, *see* **Incoherent Sources:** Lamps. To learn about interferometric instruments, *see* **Interferometry:** Overview. To learn about interferometric instruments utilizing incoherent light, *see* **Interferometry:** White Light Interferometry. To better understand the effect of the coherence of light sources on image formation, *see* **Coherence:** Coherence and Imaging.

Interferometric imaging attempts to improve the resolution of images by interferometrically combining the signals from several apertures. By doing so, a resolution can be achieved that is superior to that achievable through standard image formation by any of the individual apertures. This is especially useful in astronomy, where resolving more distant and smaller stellar objects is always desirable. It is also finding increasing use in microscopy. It has several advantages over conventional noninterferometric imaging methods. The method can combine the light gathered from several imaging instruments to form an image superior to that which can be formed by any one of the instruments. It can also obviate the need to produce a single very large aperture to achieve the equivalent resolution, which in many cases is of impractical size. In addition, because it can produce a phase-resolved measurement, the data can be processed more flexibly to form a computed image estimate. There are also considerable disadvantages to interferometric imaging compared to conventional imaging methods. The amount of signal gathered is far less than could be gathered with an equivalent single aperture. It is difficult to mechanically or feedback stabilize the time delay between two widely separated apertures to obtain an accurate phase measurement. Often only a small bandwidth of the source light can be utilized, further reducing the available signal. The interference component of the combined signal can be very small. Typically an interferometric image must be computed from the data, rather than being directly measured on photographic film or an electronic focal plane array. Even with these disadvantages, interferometric imaging is frequently more feasible than building a single large aperture to achieve the highest available resolution images.

In contrast to interferometric imaging, a standard image forming instrument such as a telescope has a resolution limited by the telescope mirror aperture in the absence of aberrations and atmospheric turbulence. As telescope mirrors become very large, they become bulky, extremely heavy, and difficult to manipulate, as well as deform under their own weight and temperature gradients. However, adaptive optics is being more frequently utilized to dynamically correct for instrument aberrations as well as atmospheric turbulence. The images captured by telescopes are typically intensity-only images, which are amenable to image processing and deconvolution, but nevertheless lack phase information. Interferometric imaging is an alternative, where the resolution of a large aperture can be attained by measuring the interference between individual points of the electromagnetic field within an area equivalent to the large aperture extent. The interferometric combining of light from sub-apertures to achieve the resolution of a larger aperture is called aperture synthesis.

Unlike image formation by a lens, interferometric images are never physically formed. Conventional images are formed when diverging spherical waves emanating from sources are focused by a lens to converging spherical waves on a photographic film or an electronic sensor. In this case, the image information is directly contained in the exposure at each position on the sensor. Interferometric measurements contain the image information as statistical correlations between the electromagnetic fields at pairs of spatial points. The statistical properties of electromagnetic waves are modeled by optical coherence theory. Interferometric imaging works by measuring the statistical correlations between various pairs of points in the electromagnetic field, and then uses optical coherence theory to infer what the image of the source of the radiation is.

The light that emanates from most natural and many artificial sources is incoherent. Incoherent radiation is optical random noise. It is analogous to the sound of static heard on a radio receiving no signal. Because it is random, the fluctuations of the electromagnetic field produced by an incoherent source are completely unrelated to the fluctuations of any other source. The origin of the randomness is usually from microscopic processes such as the thermal motions of electrons or the spontaneous emission of an excited atom. Because these microscopic processes tend to act independently of one another, the light produced by different sources tends to be mutually incoherent. Because incoherent sources do not radiate a deterministic field, the average amount of power they radiate is usually specified rather than the field itself, which is random.

However, as the fields propagate away from their sources, the fields can become partially coherent. To see this, consider the light waves propagating away from a single point-like incoherent source. The field produced at the source point consists of random fluctuations in time, but these fluctuations travel outward in a spherical wave at the speed of light. The fluctuations of the field of any two points an equal distance from the source will be the same, or perfectly correlated, because they arrive from the source at the same time. This is illustrated in Figure 1a, where a random wave is emanated from a point incoherent source. Because of the complete correlation, we can say that these two points are fully spatially coherent.

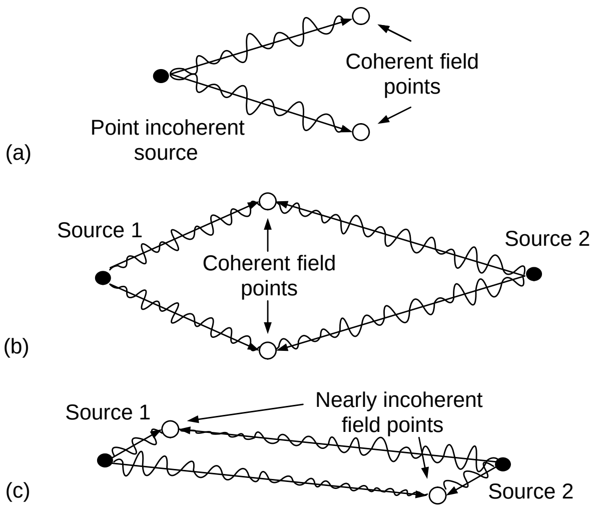

Figure 1 The distance between incoherent sources and two field points determines the coherence between the field points. (a) Because there is only one source, the field everywhere is determined by the radiation of that source. Therefore, the same fluctuations of the wave arrive at two points an equal distance from the source at the same time, and produce identical fields at both points, and so are completely coherent. (b) Two field points are now the same distance from two incoherent sources. Because both sources contribute identical fields to both points (because the fluctuations of the wave from both sources arrive at both points at the same time), the superposition of the fields from both sources at both field points are identical, and so the fields are coherent. (c) Two field points are now near two sources. Because each field point receives radiation mostly from its nearby source and very little from the other source, the fields at each field point are almost completely mutually incoherent.

So while the source itself produces random noise, the fluctuations of the field at points distant from the source become correlated because they originate from the same source.

Partially coherent waves occur when the measured field contains a superposition of the fields from more than one incoherent source. If there are now two point sources instead of one, each field point in space will receive radiation from both sources. The amount of the contribution from each source to each point in space depends on the distance between the field point and the source. In general, because both sources can be different distances from the two field points, the fluctuations at the field points will not be the same. For example, two field points that are the same distances from both sources will be perfectly correlated, or coherent, because they receive the same combination of fields from both sources. This situation is illustrated in Figure 1b. However, if one considers two points that are near both respective sources, the points will usually receive light from the nearby source. Because the sources are incoherent, and each point is mostly receiving radiation from its closer source, these points will be nearly uncorrelated, as shown in Figure 1c.

A spatial distribution of incoherent sources radiating various amounts of power will create a pattern of correlations in the field remote from the source. By using interferometry, the correlation between pairs of field points remote from the sources can be measured. The interference between two points in the electromagnetic field is achieved by relaying the two fields (e.g., with mirrors) to a common location that has an equal travel time for the fields from both points, where the fields are superimposed. The power of the superimposed beams is then detected, by a photocathode, for example. If the two signals being combined are incoherent, then no interference will take place, and the total power measured on the photodetector is simply the sum of the power of the two signals. However, if the two signals are partially coherent, there will be a deviation in the measured power from the sum of the power of the two signals. The magnitude of this deviation relative to the power of the constituent signals indicates the degree of partial coherence of the signal.

An interferometer, such as depicted in Figure 2, gathers the light from two areas of the electromagnetic field remote from an object and combines them together at the same point to form an interferogram. In this example, two telescopes of a known separation gather light from a star. The primary mirror of each telescope is of insufficient size to resolve the star as anything but a point. However, a telescope is used in practice to collect enough light to produce a

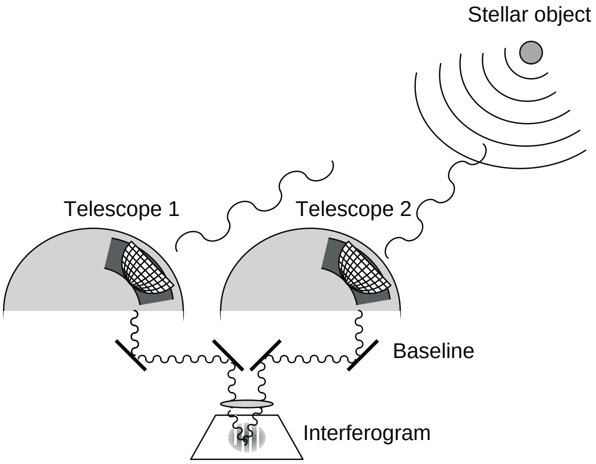

Figure 2 Two telescopes receive radiation from a stellar object. The light is collimated and directed down the baseline to a point halfway between the telescopes, where the two images are interfered together onto a focal plane.

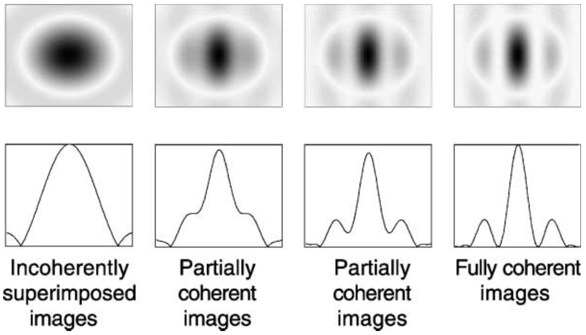

Figure 3 The image of two interfering wavefronts from a point object with various states of coherence.

measurable interference signal. Each telescope primary mirror forms a point image of the star. At the image plane of each telescope a mirror directs the point image to a measurement plane equidistant between the two telescopes. This line that joins the two telescopes is called the baseline, and it is the baseline length rather than the individual telescope mirror size that determines the minimally resolvable size. The image from each telescope is superimposed by mirrors (or perhaps beamsplitters) and imaged onto a focal plane, to get an image similar to that of Figure 3. If the light arriving from one telescope is blocked, then the image formed looks like an Airy ring pattern formed by an image, a point-like object, which is the incoherent image of Figure 3. However, when both signals are allowed to interfere, the image will have vertical stripes superimposed on it. If these stripes are very prominent, so that the image is fully darkened inside the stripes, then an image similar to the rightmost image of Figure 3 is obtained, complete intereference is occurring and the two field points at which the telescopes are situated are coherent or completely correlated. If the stripes disappear, then no interference is occurring and the two field points at which the telescopes reside are incoherent. These examples are shown in Figure 3, with images of two superimposed points with varying degrees of coherence on the top row, and a line plot of the measured field through the center of the images in the bottom row. The degree to which the contrast of the stripes is enhanced by coherence is referred to as the 'modulation' and indicates the degree or magnitude of partial coherence.

In addition to having a degree of coherence, the correlation between two electromagnetic field points also has a relative phase shift. This phase shift is given by the relative position of the interference maxima, as shown in Figure 4. As the relative phase between the

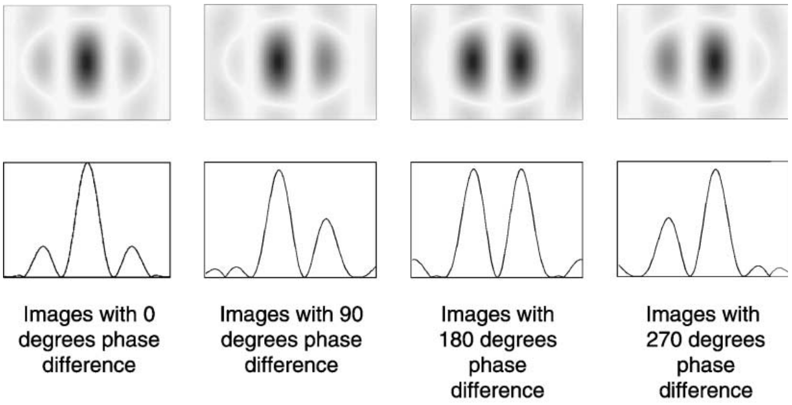

Figure 4 The image of two interfering wavefronts from a point object with various phase differences between the two wavefronts.

fields collected by the telescopes changes (which is a function of the path length difference between the light arriving at the two telescopes), the peaks of the interferogram between the two telescope images shift position. By observing the depth of the formed interference fringes and the position of the fringes, both the amplitude and the phase of the correlation between the two field points can be simultaneously measured. The amplitude of the modulation and the phase of the interference are used to compute the complex degree of coherence that is used in the computation of interferometric images.

In practice, the phase of the complex degree of coherence is much more difficult to measure than its magnitude. This is because while the amplitude tends to change very slowly as the positions of the two correlated points change, the phase is sensitive to changes in relative path length on the order of an optical wavelength. Even though a long baseline produces better resolution, the baseline length must be known to a small fraction of a wavelength to obtain a meaningful phase measurement. Minute vibrations in the positions of the mirrors on the baseline can easily cause fluctuations in the path length greater than a wavelength. This is observed as wild oscillations in the position of the fringes on the image plane. In addition, atmospheric turbulence can also produce large random phase shifts. Achieving a stable phase measurement is one of the most challenging engineering aspects of interferometric imaging.

To model realistic sources, we must determine the coherence produced by sources of a finite size. Finite size incoherent radiators can be regarded as an arrangement of an infinite number of point-like incoherent sources radiating various amounts of power. When an interferogram is made of a field which has two superimposed incoherent sources, the power of the interferograms that each would make alone is added together, rather than the fields of the interferograms. This is because when the fields of the two sources interfere, the phases of both sources are varying randomly over time. The two fields will randomly vary between constructively and destructively interfering with each other. Over a long time, the constructive and destructive interference averages out to no interference. We can then calculate what the interferogram of many incoherent sources combined is by computing the interferogram of a point source and adding the power of many such interferograms together.

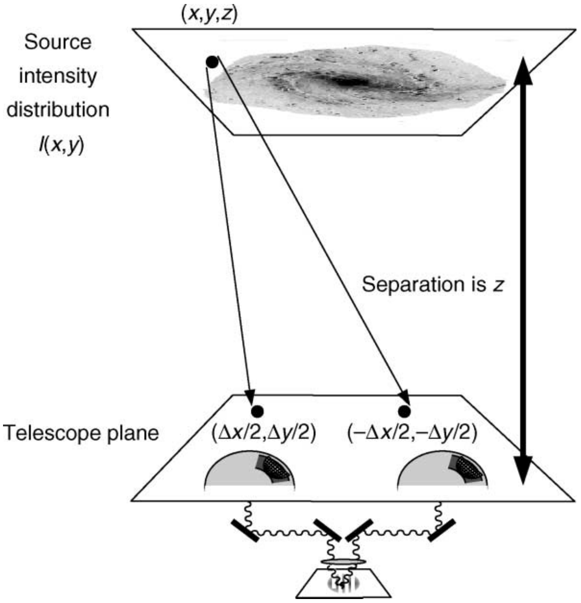

Figure 5 Diagram of source and telescope plane used for derivation of Van Cittert–Zernike theorem. A source at a distance z from the telescope plane has a power density $I(x,y)$. The telescope plane has two movable apertures symmetric about the origin of the telescope plane separated by $(\Delta x,\Delta y)$. The light from both apertures is directed to the center of the baseline at the origin of the telescope plane, where it is interfered on the sensor.

To figure out what the power of an interferogram of an arbitrary object is, first consider what the interferogram of a single point source is. Consider a situation depicted in Figure 5, but with a single point illuminator at the source plane positioned at (x,y,z) with radiant power I. We will have a receiving plane at a distance z from the source plane with two telescopes located symmetrically about the origin at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The fields received at the telescopes from the point will be interfered together at the center of the baseline, which is at the origin. We will assume that the point source produces a narrow band of radiation centered at wavelength λ. The point source produces a spherical wave that is received at both telescopes. The interference power is proportional to the two intensities measured at each telescope with the relative phase at the two telescopes causing constructive or destructive interference:

$$P \propto I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\phi \qquad [1]$$

where I_1 and I_2 are the intensities of the field at each telescope, and ϕ is the phase difference between the two fields at each telescope. We will assume that the point (x,y,z) is very far away from the telescopes, and so the intensities at the two telescopes are equal and proportional to the point intensity I, so that $P \propto I(1+\cos\phi)$. The phase ϕ is given by the difference in distance from the the source point at (x,y,z) to the receivers at $(\Delta x/2,\Delta y/2,0)$ and $(-\Delta x/2,-\Delta y/2,0)$. The phase is given by the difference in the Euclidean distances:

$$\phi = \frac{2\pi}{\lambda}\left(\sqrt{z^2 + (x + \Delta x/2)^2 + (y + \Delta y/2)^2} - \sqrt{z^2 + (x - \Delta x/2)^2 + (y - \Delta y/2)^2}\right) \approx \frac{2\pi(x\Delta x + y\Delta y)}{z\lambda} \qquad [2]$$

where the second term gives the difference in the distance to a first-order approximation in x and y. The interference intensity relation is then $P \propto I(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda))$.

If the source is now of a finite site, we can regard it as radiating an intensity (power per unit area) $I(x, y)$. To compute the total intensity of the interferogram, we integrate over all the sources in the source plane:

$$P(\Delta x,\Delta y) \propto \int I(x,y)(1+\cos(2\pi(x\Delta x+y\Delta y)/z\lambda))\mathrm{d}x\mathrm{d}y = \int I(x,y)\mathrm{d}x\mathrm{d}y + \int I(x,y)\cos(2\pi(x\Delta x+y\Delta y)/z\lambda)\mathrm{d}x\mathrm{d}y \qquad [3]$$

This result can be interpreted as follows. The intensity of the interference when the two telescopes are separated by a distance $(\Delta x,\Delta y)$ is proportional to the real part of the two-dimensional Fourier transform of the intensity distribution $I(x,y)$ evaluated at spatial frequencies $(\Delta x/z\lambda,\Delta y/z\lambda)$, plus a constant intensity. This result, related to the Van Cittert–Zernike theorem, is a commonly used result in interferometric imaging. It is covered in more detail in **Coherence:** Overview.

By moving the telescopes to various horizontal and vertical separations and measuring the interference intensity, the real part of the Fourier transform of the intensity is inferred. In practice the imaginary part can also be found by varying the baseline optical path length difference by a quarter wavelength. With these samples of the Fourier transform, the inverse Fourier transform computes the image intensity distribution $I(x,y)$. This is how simple computational image formation occurs with interferometric data.

One application of the Van Cittert–Zernike theorem and interferometric imaging is the measurement of the diameter of stars. When observing coherence, the correlations between field points tend to decrease as the separation between the points grows. By observing this decrease of the coherence, the angular size of a star can be measured. This method was pioneered by Albert Michelson in 1890 with the invention of the Michelson Stellar Interferometer. It is needed because the aperture size required to measure the angular diameter of most nearby stars in the Milky Way is larger than the size of most telescope apertures. Stars usually have a circular profile that produces a circularly symmetric pattern of correlations. The coherence between field points as the distance between them is increased does not decrease monotonically, but disappears at certain separations, and has many revivals. The first separation distance at which interference disappears (and the degree of coherence is zero) for a circular profile object is $\mathrm{d}_0 = 0.61\lambda/\alpha$, where λ is the wavelength of the light measured from the star, α is the angular diameter (in radians), and d_0 is the first separation from zero at which the interference disappears. In practice, each wavelength produces a null in interference at a different separation, so all but a small bandwidth of the star's light is filtered out before measurement to achieve a suitably strong interference null. For example, for the star Sirius, which has an angular diameter of 0.0068 arc seconds viewed from the Earth, measuring at a wavelength of $\lambda = 500$ nm (blue-green in the visible spectrum), the first separation distance at which no interference will be observed is at 18.5 m. This is larger than the largest available telescope aperture, and therefore an image cannot be directly formed that can resolve the size of Sirius.

Modern examples of stellar interferometry include the European Southern Observatory in Paranal, Chile, the CHARA Array at Mt Wilson, California, USA, and the NASA Jet Propulsion Laboratory Keck Interferometer in Hawaii, USA. The ESO main telescope system is the Very Large Telescope Interferometer (VLTI) consisting of four 8.2 m diameter telescopes as well as several smaller ones that can be moved independently, and will achieve milliarcsecond resolution through the interferometric combination of the telescopes. These telescopes are connected by underground tunnels through which the collected light from each telescope is combined. To measure the phase of the interference, extreme positioning precision is required in the relative delay between the light collected by the telescopes along the baseline, of 50 nm over 120 m, or 1 part in 2400 million, equivalent to 1.6 cm over the circumference of the Earth. Achieving this requires combinations of electronic and mechanical feedback systems with laser stabilized interferometric position measurement. The CHARA array has a very large 330-meter baseline, with 1-meter telescope apertures. The Keck Interferometer consists of two 10 m diameter telescopes with a baseline separation of 85 m. One of the main goals of achieving such high resolution is to be able to observe planets and planet formation around distant stars, to identify possible signs of life in other solar systems.

A similar principle is used for interferometric image formation in radio astronomy. The chief differences are the methods used to collect, detect, and correlate the electromagnetic radiation. Radio astronomical observatories, such as the Very Large Array in New Mexico, USA, and MERLIN at the Jodrell Bank Observatory in Cheshire, UK, use collection dishes and radio antennas to directly measure the electric field, which is not feasible at optical frequencies. These fields are converted to electrical voltages, which can then be directly correlated using electronic mixers, rather than detecting the correlation indirectly through interference. Because the wavelength is much larger, the stability requirements are greatly relaxed. Radio and optical frequency interferometric images are formed in essentially the same way, using the Van Cittert–Zernike theorem.

Interferometric imaging methods have also become of interest for microscopy applications. Aperture synthesis is not needed in microscopy, because the objects of interest are very small and therefore single apertures can be used to achieve diffraction-limited resolution. However, interferometric detection of partially coherent light affords greater flexibility because an image may be computed from the data rather than being directly detected. Typically this is achieved by creating an incoherent hologram of the object, typically through shearing interferometry. For example, by using a Rotational Shearing Interferometer, one can measure an incoherent hologram that can be used to produce a representation of an object with a greatly increased depth-of-field. In addition, direct detection of the partial coherence of sources can allow the methods of computed tomography to be used to produce three-dimensional images of volumes of incoherently emitting objects. Because of the versatility that interferometric detection can provide, interferometric imaging will surely be increasingly used outside of astronomy.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Overview. **Incoherent Sources:** Lamps. **Interferometry:** Overview; White Light Interferometry.

Further Reading

Baldwin JE and Haniff CA (2002) The application of interferometry to optical astronomical imaging. *Philosophical Transactions of the Royal Society of London A* 360: 969–986.

Beckers JM (1991) Interferometric imaging with the very large telescope. *Journal of Optics* 22: 73–83.

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill Inc.

Goodman JW (2000) *Statistical Optics*. New York: Wiley-Interscience.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. New York: Cambridge University Press.

Marks DL, Stack RA, Brady DJ, Munson DC Jr and Brady RB (1999) Visible cone-beam tomography with a lensless interferometric camera. *Science* 284: 2164–2166.

Monnier JD (2003) Optical interferometry in astronomy. *Reports on Progress in Physics* 66: 789–857.

Potuluri P, Fetterman MR and Brady DJ (2001) High depth-of-field microscopic imaging using an interferometric camera. *Optics Express* 8: 624–630.

Thompson AR, Moran JM and Swenson GW (2001) *Interferometry and Synthesis in Radio Astronomy.* New York: Wiley-Interscience.

Lidar

M L Simpson and D P Hutchinson, Oak Ridge National Laboratory, Oak Ridge, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light detection and ranging (LIDAR) refers to an optical system that measures any or all of a variety of target parameters including range, velocity, and chemical constituents. Usually a LIDAR system consists of a transmitting laser source, transmitter optics, receiver optics, and a detector. Several configurations of LIDAR systems are possible, based on the type of laser source and detection scheme selected. Basically, sources can be continuous wave (CW), pulsed, or chirped. Sources that are CW are most often used to measure velocity or chemical constituents. Pulsed and chirped sources provide the additional capability of measuring range. LIDAR receivers are either incoherent (direct detection – amplitude only) or coherent (heterodyne detection – amplitude and frequency). Generally, for velocity detection, a coherent receiver with a local oscillator (LO) laser is required. A coherent receiver is also needed if a chirped laser source is used, for instance, in frequency modulation continuous wave (FM-CW) LIDAR.

For the special case of imaging LIDAR, the detector has multiple elements. In an imaging LIDAR system, the detector, or focal plane array (FPA), and the receiver optics are designed to image independent volumes in object space onto individual elements in the FPA. Each detector element, therefore, contains information for a specific volume in space and can be thought of as an independent LIDAR system. The advantage of an imaging LIDAR system is that multiple measurements can be made on extended target(s) at the same time. Thus, characterization of rapidly changing events (like wake vortices), speckle reduction by spatial averaging, synthetic aperture radar, multiple target tracking, and target identification are all possible without the sophisticated algorithms and hardware platforms required in single channel scanning systems. In addition because the multiple pixels allow parallel image capture, a true imaging LIDAR can acquire data much more rapidly than a scanning system.

In this chapter, for clarity, the emphasis will be on imaging LIDAR systems applied to hard body targets for the measurement of range and velocity. The basic configurations presented here, though, are extensible to distributed targets and measurements of wind profiles and atmospheric constituents. Also, since general textbooks are available that cover LIDAR systems and incoherent imaging systems, the concentration will be on architectures, optics, detectors, and design considerations that are specifically applicable to coherent imaging LIDAR.

Frequency Modulated Continuous Wave (FM-CW) LIDAR

One LIDAR technique, which is particularly suited for imaging due to the simplicity of the processing, is FM-CW LIDAR. FM-CW LIDAR simultaneously measures range and velocity (Doppler shift) by using a frequency chirped waveform. Referring to Figure 1, assume that a continuous wave (CW) laser frequency is linearly modulated (chirped) by a triangular wave about an offset frequency, such that the modulated signal is given as:

$$f(t) = f_o + \alpha\tau \qquad [1]$$

where f_o is the subcarrier offset frequency and α is the linear modulation rate in frequency/second, and τ is a time variable. The transmitted beam is focused by the optical system onto the target surface, reflected (or absorbed or scattered) and arrives at the receiver after τ seconds, where

$$\tau = 2\frac{R}{c} \qquad [2]$$

and R is the distance to the target surface, and c is the speed of light. The returning signal is optically mixed with the LO on the detector and produces a sum/difference beat frequency (Δf). Figure 1 shows the incident wave with its linear or chirped frequency characteristics and the delayed version returned from the target. Only the difference frequency can be optically detected since only it falls within the detector's bandwidth. Assuming the target is not moving, the result is

$$\Delta f = f(t+\tau) - f(t) = f_o + \alpha(t+\tau) - (f_o + \alpha t) = \alpha\tau \qquad [3]$$

where $\alpha = B/T$, the chirp rate of the modulator in units of Hz/s, B is the modulation bandwidth, and T is the period of the chirp. Substituting into eqn [3] for

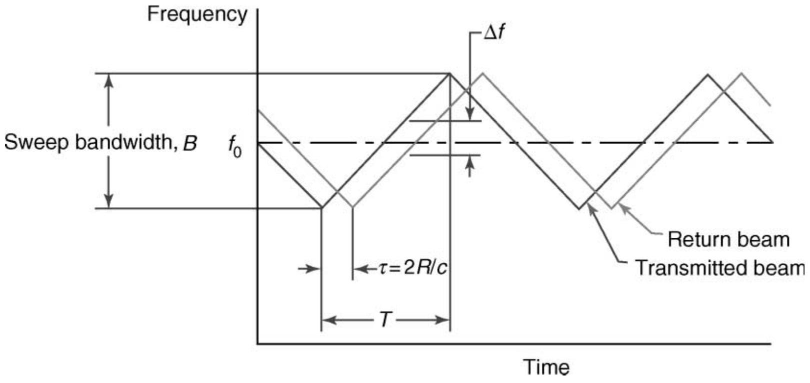

Figure 1 Triangular-wave modulation Envelope on the transmitted beam (black) and return beam (gray).

τ and α we have

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) \quad [4]$$

Finally, one can solve for the distance to the target, R, as a function of the beat frequency and chirp rate:

$$R = \frac{cT}{2B}\Delta f \quad [5]$$

The difference frequency, Δf, is the same for the up and down ramp of the modulation signal. Figure 2 demonstrates the return signal for the case of a target moving toward the LIDAR system. The difference frequency now contains a component proportional to the target induced Doppler shift given by

$$f_{\mathrm{d}} = 2\frac{v}{\lambda} \quad [6]$$

where v is the line-of-sight target velocity and λ is the laser wavelength. From Figure 2, the Doppler shift causes the difference frequency on the up ramp to be less than the range induced frequency shift given by eqn [5] and the net difference frequency is greater on the down slope. In this case, the difference frequency is given by

$$\Delta f = \frac{B}{T}\left(2\frac{R}{c}\right) + f_{\mathrm{d}} \quad [7]$$

The range and Doppler contributions to the difference frequency may be separated by measuring the frequency difference over both the up and down chirp of the waveform. The Doppler frequency $f_{\mathrm{d}} = \Delta f(\mathrm{up}) - \Delta f(\mathrm{down})$ and the range is given by the average of Δf(up) and Δf(down) over a complete cycle of the modulating triangular wave. Therefore, substituting in eqns [5] and [6], the range may be determined from the equation

$$R = \frac{cT}{2B}\left\{\Delta f(\mathrm{up}) - \Delta f(\mathrm{down})\right\} \quad [8]$$

and the line-of-sight (LOS) velocity is

$$v = \frac{\lambda}{2}\left\{\frac{\Delta f(\mathrm{up}) + \Delta f(\mathrm{down})}{2}\right\} \quad [9]$$

The range resolution of a FM-CW LIDAR system is the range for which the difference frequency equals 1 cycle in the sweep time interval of T or $\Delta f = 1/T$. Then, from eqn [5], the range resolution, ΔR, is given by

$$\Delta R = \frac{cT}{2B}\frac{1}{T} = \frac{c}{2B} \quad [10]$$

Therefore, the range resolution is only a function of the sweep bandwidth and not the laser wavelength. However, the Doppler shift is dependent on the laser frequency. The accuracy with which the difference frequency may be determined is a function of the number of cycles over which the measurement is made. The number of cycles is maximized for the case of $\tau = T/2$ or T should be set for twice the time delay for the maximum range of interest.

Numerical Example

A range resolution of 30 cm requires a chirp bandwidth, B, of 500 MHz. For a total range of interest of 10 km, the chirp time, T, should be 133.33 μs or the chirp rate should be 3750 Hz (1 up and 1 down ramp per cycle). The difference frequency for a 10 km range target will be 250 MHz, which demonstrates the need for wideband detectors to capture high-resolution range images (Table 1).

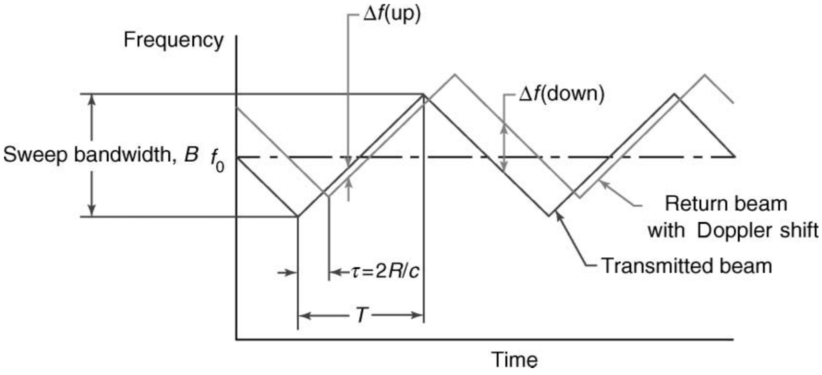

Figure 2 Lidar return with Doppler shift proportional to target velocity.

Table 1 Summary of FM-CW LIDAR characteristics

Parameter	*Value*
Range resolution (m)	0.3
Chirp rate (Hz)	3750
Chirp bandwidth (Hz)	500×10^6
Maximum range (km)	10

Optical Configuration for Imaging LIDAR

Imaging LIDAR systems can be constructed for virtually any wavelength range with appropriate sources, optical components, and detectors. The design of an optical receiver system capable of recording coherent images requires additional considerations over those necessary for a video camera or single-channel receiver. First, the collection optics must have resolution capabilities consistent with both the detector grid dimensions and the coherence properties of the source. Next, the local oscillator (LO) power must be applied to each array element. For a typical infrared (HgCdTe) detector, one to ten milliwatts of LO power might be required to reach the shot noise limit. For large arrays, this may result in unacceptable heat loads, even for momentary application. Thus, detector arrays with substantial spacing between detector elements benefit from the application of individual LO 'beamlets' to shield the insensitive portions of the array substrate. These beamlets can be generated by illuminating a mask containing an array of holes with the LO Gaussian laser beam. This scheme is illustrated in Figure 3.

In the configuration shown in the figure, the signal beam (from the target) and the LO beamlets are combined at the beamsplitter prior to being focused onto the detector plane by the objective lens. The detector plane coincides with the far-field image plane of the objective lens, and any additional optical elements utilized to enhance the camera's resolution are not shown. An expanded LO beam illuminates an aperture array located at the far-field object plane of the ancillary lens. For instance, if the objective lens and the ancillary lens are identical, the LO mask is imaged onto the detector plane with unity magnification. The detector surface is assumed to be flat with uniform video and heterodyne quantum efficiency.

Several aspects of this design are discussed below. We first discuss how the optical parameters may be adjusted to give nearly-Gaussian LO profiles at each detector position. The receiver antenna array is then shown to consist of nonoverlapping sub-beams spaced with the detector array geometry. Signal considerations for both incoherent and coherent sources are then discussed, and it is found that optimal detector size and array-spacing is dependent on the coherence properties of the intended object field.

Aperture Array

As an alternative to holographic LO beam array generation, the LO mask design presented here consists of an array of apertures with size and spacing equal to the detector dimensions. The LO is expanded to coherently illuminate the entire mask with approximately uniform intensity. The mask is located at the front focal plane of the ancillary lens and its image is located at the back focal plane (detector plane) of the objective lens. The coherent image of the mask formed at the detector plane by the ancillary-objective lens combination can be described by a two-stage Fraunhofer diffraction process.

Finite dimensions of the ancillary pupil results in spatial filtering which can affect the image profile of the LO mask. If desired, the ancillary pupil may be chosen to provide nearly-Gaussian LO beamlet profiles with beam waists at each detector location.

It is useful to first consider the coherent image of a single aperture uniformly illuminated with plane LO wavefronts. In one dimension, this wave train diffracts from the aperture to form a far-field diffraction pattern at the Fraunhofer plane of the ancillary lens given by

$$E_\mathrm{F}(\beta) = E_\mathrm{F}\frac{\sin(\beta)}{\beta} \qquad [11]$$

where $\beta = \frac{1}{2}kb\sin(\theta)$, $k = 2\pi/\lambda$, λ is the wavelength, b is the aperture width, and θ is measured from the optical axis. For small θ, $\beta = \pi b\nu/(\lambda f)$ where ν measures a point on the Fraunhofer plane and f is the focal length of the ancillary lens. The finite diameter D (determined by an aperture stop in the Fraunhofer plane) of the ancillary lens spatially filters the diffraction pattern and this limits the range of β values in eqn [11]. We define an aperture-limited pupil as one which limits β to the range $-\pi < \beta < \pi$; this will be the case when $f/D = b/2\lambda$.

Consider now a linear aperture array consisting of N individual apertures of width b and separation $a > b$. Assume that each aperture is illuminated by plane, uniform, mutually coherent LO wavefronts of equal intensity. In this case, the LO profile at the Fraunhofer plane is characterized by the diffraction pattern for a coherently illuminated linear diffraction grating:

$$E_\mathrm{F}(\theta) = E_0\mathrm{sinc}\,(\beta)\frac{\sin(N\alpha)}{N\sin(\alpha)} \qquad [12]$$

where $\alpha = \frac{1}{2}ka\sin\theta$ and where β, θ, and k are defined above.

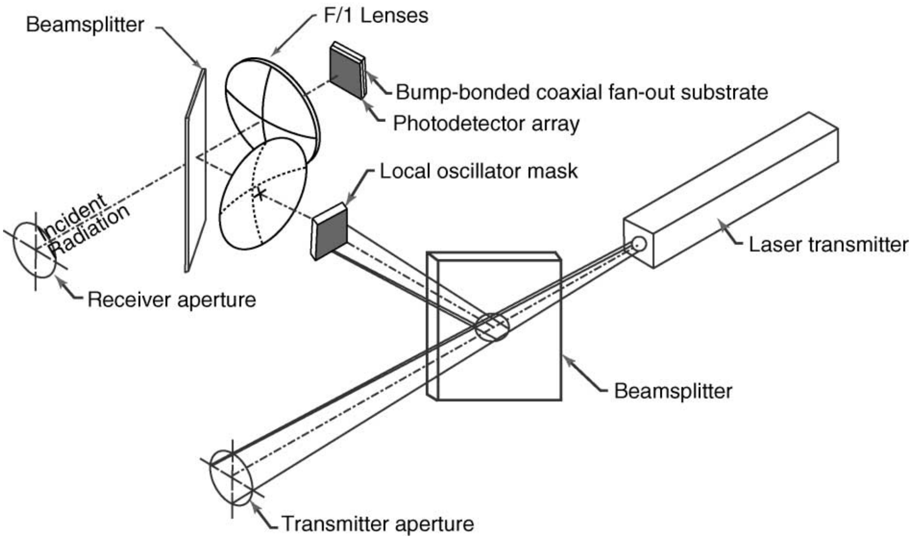

Figure 3 The individual beamlets are generated by the local oscillator or aperture mask. Reproduced with permission from Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Propagation of the LO from the ancillary lens Fraunhofer plane to the detector plane can be described by applying the Fraunhofer diffraction integral to eqn [12] with integration limits appropriate to the ancillary pupil dimensions. The larger pupil gives increased illumination at the detector edges while the aperture-limited pupil gives a nearly Gaussian LO profile with 1/*e* points close to the detector edges.

Receiver Antenna Array

The requirement of wavefront matching between the LO field and the signal field leads to the notion of an 'antenna beam' or equivalently, 'backward propagating LO' (BPLO) to define the field of view of an individual heterodyne receiver element. This beam is formed by allowing a reverse projection of the local oscillator field to diffract from the detector surface, through all collection optics in the signal beam path to a point on the object field of the imaging system. A receiver array will produce an array of such BPLO beams which must maintain the detector array geometry in the object field if the LO and signal beams are to mix efficiently.

As above, spatial filtering by the receiver pupil (assumed to be located at the Fraunhofer plane of the objective lens) will affect the BPLO profile. In the Fraunhofer plane of the objective lens, the uniformly illuminated single-detector BPLO profile is given by eqn [11], and we define an aperture-limited detector as one which limits the range of β to $\pm\pi$. Similarly, eqn [12] characterizes the BPLO profile resulting from a detector array uniformly illuminated with an extended uniform LO wavefront. The uniform illumination BPLO profiles in the object field differ from the detector plane profiles only in horizontal scale which now is consistent with the magnified detector dimensions.

For our purposes, we may define the far field of the objective lens according to the condition for maximum Gaussian beam collimation:

$$\omega_d \omega_0 = \frac{f\lambda}{\pi} \quad [13]$$

where ω_0 is an effective BPLO Gaussian beam waist at the Fraunhofer plane and ω_d is an effective BPLO beam waist at the detector. Under this circumstance, the BPLO Rayleigh range is $Z_D = \pi\omega_0^2/\lambda = f(\omega_0/\omega_d)$. Identifying ω_0/ω_d as the magnification (notice that for the BPLO, the magnification is greater than unity) gives $d_o = (d_i/f)Z_D \sim Z_D$ where d_o is the object distance and d_i is the image distance. Thus the far-field of the objective lens is at least one BPLO Rayleigh range away.

The object field BPLO profiles for an aperture-limited detector array consists of nearly Gaussian sub-beams with 1/*e* points located approximately at the magnified detector edge locations. Since the far-field object distance exceeds Z_D, the divergence of each BPLO sub-beam is constant and remains distinct and nonoverlapping for all $d_o > Z_D$.

Imaging Applications Involving Incoherent Sources

For an object plane in the far field of aperture-limited collection optics, the nearly Gaussian-profile BPLO propagates from the receiver aperture according to the condition for maximum Gaussian beam collimation. The waist is at the Fraunhofer plane where the BPLO wavefront is nearly plane, while at the object field the BPLO wavefronts are spherical with radius equal to

$$R(d_o) = d_o\left[1 + \left(\frac{Z_D}{d_o}\right)^2\right] \qquad [14]$$

where we have set $d_o - f \approx d_o$.

It will generally be the case that an imaging system will be used to analyze extended objects whose size exceeds the resolution limit. Following convention, we define a spatially coherent source as one which is illuminated with spatially coherent light with appreciable structure and roughness of depth within the coherence length of the illuminating light, and where the target motion does not appreciably perturb the speckle field during the measurement process. The speckle effects, which so seriously degrade the information content in images formed of spatially coherent objects, are not present in images formed of incoherent objects such as thermal emitters. Speckle may be reduced by time-averaging speckle fluctuations resulting from target motion, atmospheric effects, and independent transmitter pulses, and in the limit such objects become incoherent pseudothermal sources. Similar remarks apply to spatial averages over detector subarrays.

Independent radiators on the object field produce disturbances which can diffract to illuminate the entire receiver aperture as well as a finite coherence area in the image field A_{ci} given by

$$A_{ci} \gg \frac{\lambda^2}{\Omega_{ci}} \qquad [15]$$

where Ω_{ci} is the solid angle subtended by the receiver aperture at the image field. Alternatively, we may interpret A_{ci} as the area of an aperture in the object plane which, if illuminated with wavefronts characterized by eqn [14] (reciprocal illumination), would just fill the receiver aperture with a central diffraction order. Thus the combination of independent disturbances within the detector's field of view on the object plane, combine to form a spatially coherent disturbance over the receiver aperture and within A_{ci}.

According to the Van Cittert–Zernike theorem, the complex degree of spatial coherence μ_s is a unity amplitude wavetrain that propagates exactly as a signal beam resulting from the reciprocal coherent illumination of an object-field aperture of area A_{ci}. Thus, μ_s and the BPLO are Helmholtz-reciprocal pairs whose overlap determines the mixing efficiency with which heterodyne signal is generated from radiation collected from an extended incoherent source. Following convention, we refer to the overlap of μ_s and the BPLO as the effective receiver aperture.

Detectors larger than the aperture-limit, whose area exceeds A_{ci}, exhibit reduced receiver apertures and a corresponding decrease in heterodyne mixing efficiency. For example, varying the receiver pupil, while maintaining constant LO illumination, results in a heterodyne signal proportional to the received power for detector diameters much less than the aperture-limit, while detector diameters much greater than the aperture-limit produce heterodyne signal only in proportion to the square-root of the received power. Thus, for constant LO illumination and incoherent sources, increasing the entendue past the aperture limit will always increase the heterodyne signal, but less than for direct detection.

Optimal Detector Size

As noted above, an aperture-limited system is one where detectors of diameter d are matched with an objective lens with $f/D = d/2\lambda$. Commercially available detectors for use at 10 μm are commonly 100 μm wide suggesting the use of a $f/5$ lens, however aspheric lenses as fast as $f/1$ are also commercially available. Apart from aberrations, it would be advantageous to utilize a 5×5 subarray of 20 μm detectors with an $f/1$ lens in place of a single 100 μm detector for applications involving coherent sources. Applications involving incoherent thermal sources would generally favor a wide detector/fast lens combination since thermal sources provide weak heterodyne signals.

Noise Analysis and Performance

Radar Range Equation

The laser power requirements for the LIDAR system depend on the signal-to-noise ratio of the FPA, the i.f. amplifier performance, atmospheric transmission, and the optics system design. The following calculations are used to determine the performance of the LIDAR system for a single pixel of the FPA. The radar range equation is

$$P_r = P_t\left(\frac{G_t}{4\pi R^2}\right)\left(\frac{\sigma_{pt}}{4\pi R^2}\right)\left(\frac{\pi D^2}{4}\right)\eta_{atm}\eta_{sys} \qquad [16]$$

where P_r = received power; P_t = transmitted power; G_t = antenna gain; σ_{pt} = radar cross-section; R = range; D = transmitting and receiving aperture diameter; η_{atm} = atmospheric transmission; and η_{sys} = optical system efficiency.

The beam divergence of the transceiver is

$$\theta_t = \left[\left(\frac{D}{R}\right)^2 + \left(\frac{\lambda}{D}\right)^2\right]^{0.5} \qquad [17]$$

where λ = wavelength and the antenna gain is

$$G_t = \frac{4\pi}{\theta_t^2} \qquad [18]$$

The radar cross-section for a lambertian diffuse point target is $\sigma_{pt} = 4\rho_t dA$. The target reflectivity is ρ_t and dA is the target area, which is effectively the size of a pixel of the FPA projected onto the target. For purposes of the following calculations, a surface reflectivity of 0.1 is assumed. The pixel size at the target (assuming Gaussian beam optics) is given by

$$dA = \pi\left(\frac{D}{3}\right)^2\left[1 + \left[\frac{\lambda R}{\pi}\left(\frac{3}{D}\right)^2\right]^2\right] \qquad [19]$$

where D = transmitter and receiver mirror diameter and it is assumed that the target is larger than dA.

Finally, the signal-to-noise (SNR) ratio for a coherent receiver is given by

$$\text{SNR} = \frac{\eta P_r}{h\nu B_{if}} \qquad [20]$$

where η = detector quantum efficiency; P_r = received power from eqn [16]; h = Planck's constant; ν = frequency of laser; and B_{if} = post-detection bandwidth of signal processor.

Numerical example

Assuming an atmospheric transmission efficiency, η_{atm}, of 98%, an optical system efficiency, η_{sys}, of 30%, and a detector efficiency, η, of 10%, the SNR was calculated as a function of the transmitting/receiving telescope (assumed to be a mirror) diameter, D, with target range as a parameter and plotted in Figure 4. This calculation assumed that the pixel field-of-view (FOV) at the target was illuminated with 1 watt of power. Even at a range of 10 km, the 32 × 32 element imaging LIDAR receiver with a SNR of 100 for each pixel will require only 75 watts (approximately 74-milliwatts per pixel) of total transmitted power for a mirror diameter of 30 cm. From Figure 4, it appears that the telescope mirror size could be reduced to a diameter of 20 cm and still maintain an excellent SNR. However, the SNR is not the only consideration on the transmitting telescope diameter. The cross-range target resolution is determined by the projected pixel size at the target. This target pixel size is plotted versus the telescope mirror diameter in Figure 5. Diffraction from the transmitting and receiving optic sets the limiting resolution that can be achieved with the LIDAR system and is only a function of the diameter and wavelength. A transmit/receive telescope aperture of 30 cm seems to represent a good trade-off between cross-range resolution and SNR.

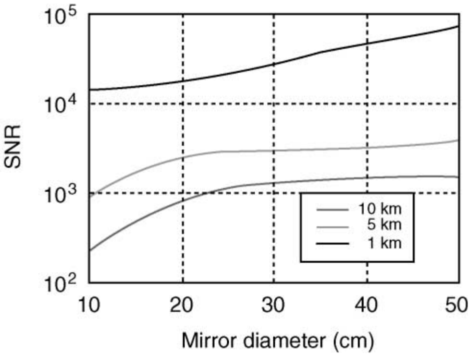

Figure 4 The SNR saturates for a mirror diameter greater than 30 cm at long range.

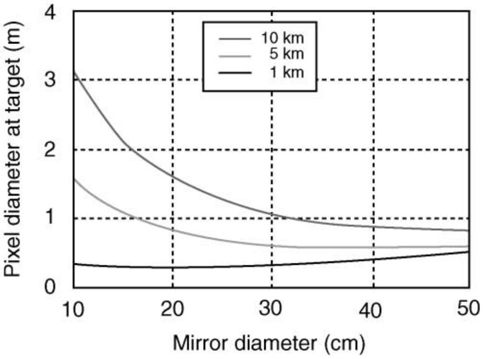

Figure 5 The pixel size determines the target cross-range image resolution.

Applications and Future Directions

The remote sensing community has recognized the advantages of coherent receivers over conventional direct IR detection with respect to increased measurement sensitivity and multidimensional information capability. These advantages, however, come at a price of added system complexity, power, size, weight, and cost. In addition, choices of components, particularly in the long wavelength infrared (LWIR) region of the spectrum, have been limited to liquid nitrogen-cooled nonimaging detectors and table-top-sized gas lasers. Thus, implementations of imaging LIDAR systems have been primarily single detector, single source scanning systems which require high pulse repetition frequency lasers, complex tracking systems, and platforms, and augmentation of automated target recognition input to compensate for multiple, simultaneous measurements on the same target. New technology developments in quantum well infrared photodetector (QWIP) detectors and arrays, and in HgCdZnTe technology are now paving the way toward thermo-electrically cooled LWIR imaging array detectors. Also waveguide, folded cavity, laser resonator designs are providing compact, lightweight packages for CO_2 lasers. Coherent receivers, that use the new focal plane arrays and compact sources, provide the capability of making parallel LIDAR measurements without the added hardware and software overhead of scanning systems. In the following section, several examples are provided that demonstrate the versatility of imaging LIDAR systems that use detector arrays for measurement applications such as chemical plumes, velocity gradients, and speckle reduction.

Wideband heterodyne receivers are useful in measuring chemical species whose absorption peaks are not well aligned with the laser source line. If the species absorption peak corresponds directly with the laser line, an absorption measurement can be made using direct detection (amplitude only). For off-line (off-resonance) measurements, heterodyning is needed, where the bandwidth of the heterodyne receiver dictates how far away from the laser line an absorption peak can be measured. For example, the CO_2 laser line spacing in the 9 to 12 micron region of the spectrum averages around 50 GHz. With the new QWIP detectors theoretically having heterodyne bandwidths of 30 GHz or more, full high-resolution spectral coverage for the long-wave IR is within grasp. For chemical detection with heterodyne receivers, there are two modes of operation. The first is termed passive heterodyne radiometry where the source is a black-body radiator and the absorption is measured for chemical species located between the blackbody and the receiver. The second mode of operation is thermoluminescence where a laser transmitter is used to excite a chemical species in the atmosphere and the resulting emission (luminescence) is measured with the heterodyne receiver. The advantage of imaging in both cases is the ability to capture chemical plumes in the presence of wind and other anomalies.

To illustrate the ability of heterodyne imaging receivers to spectroscopically resolve chemical plumes, Figure 6 shows a series of 10 × 10 pixel images of a small bottle of concentrated ammonium hydroxide against a 212 °C blackbody recorded in both passive heterodyne radiometry and direct detection modes. Figure 6a shows a heterodyne image of the bottle with the top in place where, in Figure 6b the top is removed and the absorbing plume is clearly imaged. The heterodyne images in Figures 6a and 6b were recorded with the LO adjusted to emit the 9R(30) line which is known to

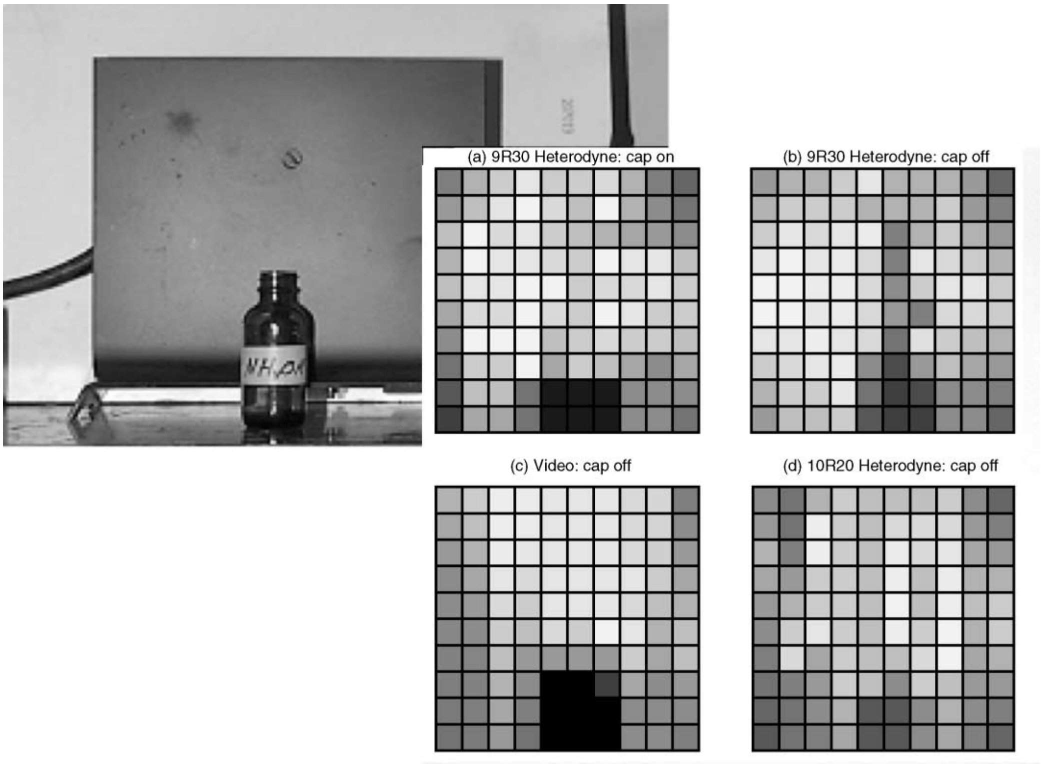
NH4OH
(a) 9R30 Heterodyne: cap on
(b) 9R30 Heterodyne: cap off
(c) Video: cap off
(d) 10R20 Heterodyne: cap off

Figure 6 Ammonia plume characterization.

be absorbed strongly by NH_3. In Figure 6c, the same image was recorded by direct detection and fails to detect the ammonia plume. Similarly, in Figure 6d, the LO was adjusted to emit a nearby line known not to be absorbed by NH_3 and the plume is again not detected. Both the objective and ancillary lens were *f*/1 38 mm, and the object field was about three meters away. The top of the bottle was approximately 2 cm in diameter and is located at the bottom of the field of view. Measurements utilizing hot ammonium hydroxide imaged against a cold background resulted in similar images (albeit at substantially reduced signal to noise ratios) where the plume was bright and the background was dark.

Another use of imaging LIDAR is the measurement of velocity profiles. Applications include wake vortex measurements for aircraft, wind shear detection, and vibration signature measurements for identification. In another experiment, active mode images were recorded by illuminating the target with a portion of the LO laser beam which had been shifted in frequency by 40 MHz using an acousto-optic modulator. The resulting Doppler image is recorded using an 8 × 8 HgCdTe array with discrete electronics for each pixel element in the array. Figure 7 shows the active image of a vertical squirrel cage fan where each pixel is rendered to represent the peak Doppler shift measured with the detector at the corresponding location in the image plane. The moving target scatters incoherently and thus speckle effects do not degrade the image. The measured Doppler shifts are consistent with the known rotational velocity of the target.

Speckle effects associated with coherent mixing of laser light reflected from an extended target (constructive and destructive interference) are a major source of noise in LIDAR measurements. These speckle effects can be reduced by signal averaging. For single detector LIDAR, averaging over many transmitter pulses can reduce speckle but this technique also reduces temporal resolution. The alternative considered here is to sacrifice spatial resolution by averaging over an array of detectors, all of which record the signal returned from a single transmitter pulse. Detectors separated by a coherence area diameter will record signals that approach statistical independence, and in this case speckle effects can be reduced by an amount approaching the square-root of the number of detectors in the subarray.

The following laboratory experiment demonstrates speckle reduction by spatial averaging with a fixed focal plane array. The receiver used in this experiment consists of a 3 × 3 array of HgCdTe detectors. The 50-micron diameter detectors are arranged in a square pattern, with 100 micron center-to-center spacing. Custom electronics multiplex the output of each detector for subsequent processing. As with the velocity profile experiment, LIDAR illumination is produced by frequency shifting a portion of the LO beam by 40 MHz with an acousto-optic modulator prior to illuminating the object field. Polarization of the illumination beam in the experiment is parallel to the LO prior to scattering from the target. The measured receiver bandwidth, limited mostly by the dewar design, is about 100 MHz, which permits the acquisition of the 40 MHz heterodyne signal. The collection optic consists of a 38 mm *f*/2 asphere which produces a 2 mm diameter pixel image at a 3 m object distance.

The target for this experiment is a 30 cm diameter disk coated with an aluminum powder. The particle size of this powder ranges from 1 to 100 microns as determined by electron microscopy. The disk is attached to a stepper motor that is incremented between frames to provide statistically independent speckle fields. Figure 8 shows the distribution of

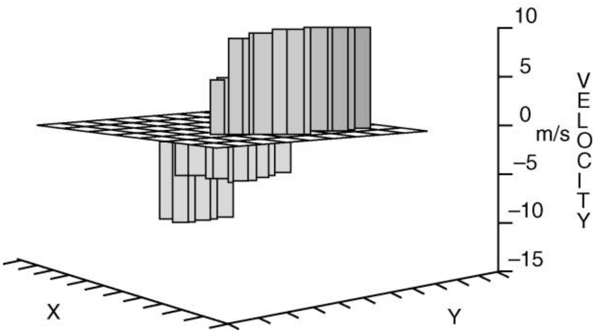

Figure 7 Doppler image of a rotating squirrel cage fan.

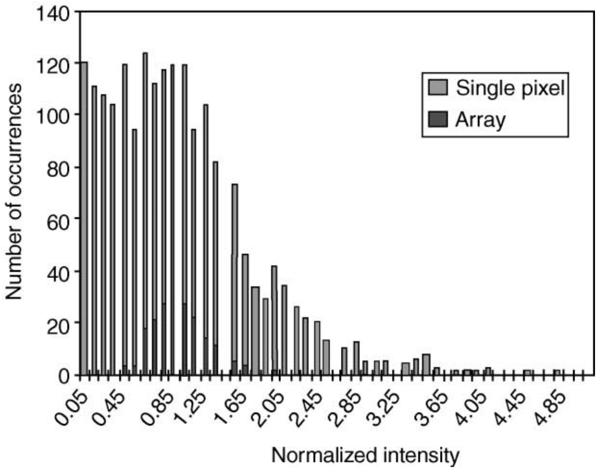

Figure 8 Speckle reduction using spatial averaging over a coherent array.

measurements taken with a single pixel superimposed on the distribution taken by averaging over the focal plane array. A reduction in speckle noise by a factor of the square root of the number of averaged pixels is obtained as predicted.

In this chapter, an overview of imaging LIDAR systems has been presented. Imaging LIDAR system design, performance, and application were discussed. With new technology developments in wide-bandwidth QWIP detector arrays, thermo-electrically cooled detectors, and compact sources, substantial potential exists for imaging LIDAR systems in the next generation in remote sensing. Due to the parallel, multifunction measurement capability of these new receivers a single sensor system can provide three-dimensional information, including range, shape, velocity, and chemical composition. In addition to the sensor fusion advantages, snapshot imaging allows the capture of high-speed, rapidly evolving events that elude present scanning systems.

See also

Environmental Measurements: Doppler lidar. **Imaging:** Infrared Imaging.

Further Reading

Born M and Wolf E (1964) *Principles of Optics*. London: Pergamon.

Forrestor AT (1956) On coherence properties of light waves. *American Journal of Physics* 24: 192–196.

Goodman JW (1975) Statistical properties of laser speckle patterns. In: Dainty JC (ed.) *Laser Speckle and Related Phenomena*, Chap. 2, pp. 9–75. New York: Springer-Verlag.

Haner AB and Isenor NR (1970) Intensity correlations from pseudothermal light sources. *American Journal of Physics* 38: 748–750.

Holzhauser E and Massig JH (1978) An analysis of optical mixing in plasma scattering experiments. *Plasma Physics* 20: 867–877.

Jelalian AV (1992) *Laser Radar Systems*, pp. 3–5. Norwood, WA: Artech House.

Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1566.

Kusters JF, Rye BJ and Walker AC (1989) Spatial weighting in laboratory light scattering experiments. *Applied Optics* 28: 657–664.

Martienssen W and Spiller E (1964) Coherence and fluctuations in light beams. *American Journal of Physics* 32: 919–926.

Measures RM (1984) *Laser Remote Sensing Fundamentals and Applications*. New York: John Wiley.

Rye BJ (1979) Antenna parameters for incoherent backscatter heterodyne lidar. *Applied Optics* 18: 1390–1398.

Shannon RR (1997) *The Art and Science of Optical Design*. New York: Cambridge University Press.

Siegman AE (1966) The antenna properties of optical heterodyne receivers. *Applied Optics* 5: 1588–1594.

Simpson ML, Bennett CA, Emery MS, *et al.* (1997) Coherent imaging with two-dimensional focal-plane arrays: design and applications. *Applied Optics* 36: 6913–6920.

Veldkamp WB (1983) Holographic local oscillator beam mutiplexing for array heterodyne detection. *Applied Optics* 22: 891–900.

Veldkamp WB and Van Allen EJ (1983) Binary holographic LO beam multiplexer for IR imaging detector arrays. *Applied Optics* 22: 1497–1506.

Zernike F (1938) The concept of degree of coherence and its application to optical problems. *Physica* V8: 785–795.

Multiplex Imaging

A Lacourt, Université de Franche-Comté, Besancon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ability to project an image onto a screen in a darkened room has been known since antiquity and Leonardo da Vinci (1515) precisely described principles of the Camera oscura. When one exposes a paper sheet directly to sunlight, it is uniformly lighted. However, if one pierces a little hole through the paper sheet then an image of the sun can be displayed on a screen beyond the hole (Figure 1). Similarly, the shape of the light spot we can see in a dark room, from the rays passing through a small aperture of a closed shutter, is not that of the aperture. Generally, the spot has a circular shape, as an image of the sun. Actually, it is neither a shadow of the hole nor an image of the sun but a combination of both. Light incident on an aperture contains the entire information about the light distribution of the object, but this information has to be 'deblurred', or decoded. The aperture acts as an elementary spatial 2D-filter that selects and preserves the direction and intensity of light rays coming from each point of the scene. The projection on the screen of each pencil of light passing through the hole is a small light spot, the geometry of which is the same as the aperture, and

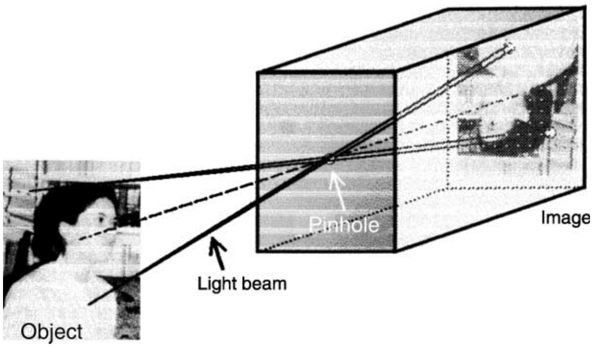

Figure 1 Principle of darkroom.

the intensity is proportional to that of the emitting point. The smaller the hole, the more detail the image exhibits, but with less luminosity. Thus, we must adopt a compromise between the resolution and the illumination of the image.

Now, if the mask has several holes, the amount of light reaching the screen, being proportional to the whole aperture area, is extensively increased. On the other hand, as each hole displays an image on the screen, the overlapping of these multiple images gives a new but blurred image. However, the information exists and we should be able to retrieve it if we know the characteristics of the complex aperture. This is the basic concept of multiplex imaging. The crucial question is: what is a good aperture design and a good way to reconstruct images having high resolution and with high illumination levels? The suitable masks with this aim are named synthetic apertures.

Porta (1589) and Kepler (1611) found the perfect solution, by setting up a lens in front of a large hole. Indeed, one can regard a lens as a wide aperture made of continuously juxtaposed microprisms (Figure 2). The lens obtains such refractive properties that the image displayed by each of these microprisms is moved toward a center. Moreover, as the angle of the microprisms varies continuously, it focuses on a single point, all the rays coming from a given object point. Thus, the blurring is suppressed and the illumination is increased. The way was opened toward the invention of optical instruments and photography and the 'pinhole' has fallen into oblivion for a long time.

Two new problems awake interest in synthetic apertures, around 1960. First, the domain of imaging has extended out of visible light: infrared and ultraviolet light, γ-rays, X-rays, ultrasound waves, and others. Materials suitable for making lenses or mirrors, operating with such waves, are expensive or often do not exist. For example, most of the transparent materials have a refractive index close to 1 for γ-rays and X-rays. However, some are opaque enough to enable the use of masks.

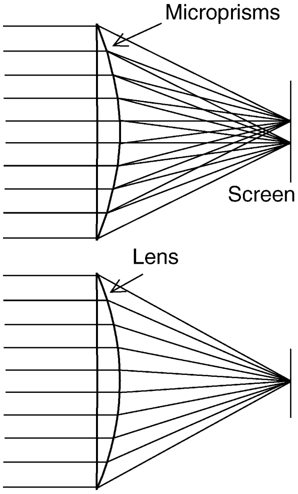

Figure 2 A lens can be modeled as a juxtaposition of microprisms with continously variable angles.

The second challenge appeared with progress in astronomy. In order to explore the universe more, further from Earth and closes to the Big Bang, astrophysicists needed telescopes providing very high luminosity and high resolution, since distant stars, nebulae, and other galaxies are extremely weak and small luminous objects. This exploration can be achieved by increasing the diameter of the primary mirror of telescopes. Indeed, due to diffraction of light by the pupil of the primary mirror, the theoretical limit of resolution of a telescope (i.e., the smaller distance between two points that one can discriminate), is inversely proportional to the diameter of the mirror. In addition, the illumination is proportional to its area. However, the working of large mirrors is a technical prowess that is very expensive and has high risks. Defect of a fraction of micrometer when polishing the surface can have dramatic consequences, e.g., Hubble. Also, such mirrors, being heavy, can bend out of shape under their own weight. Nevertheless, outstandingly large telescopes have been implemented, based on Adaptive optics: the deformations of the mirror are measured in real time by an interference method and compensated by activating a mosaic of jacks (New Technology Telescope – NTT – European program, 1989). Another way consists in replacing the solid-state mirror by a mosaic of 36 mirrors accurately directed (Keck telescope, USA). However, in either case, the mirror diameter does not exceed 8 to 10 meters. Synthetic apertures could be a solution to that difficulty, as earlier demonstrated by Fizeau (1868) and Michelson (1890).

In the first part of this article, any fundamentals are briefly recalled about properties of light and imaging.

The second part describes recent experimental methods and results obtained in synthetic aperture imaging. The last part reports the important programs in progress today relating to stellar interferometry.

Background

Electromagnetic Waves Propagation

Light power, which is involved in image formation, is carried by the electric field of an electromagnetic wave (Figure 3). The electric field $\vec{E}$ has properties of a vector orthogonal to the direction of propagation $\vec{N}$ of the light ray. It is characterized by its amplitude $\vec{E}_0$, the angular time frequency of vibration ω, and the wavevector $\vec{k}$, which specifies the propagating direction of the ray. Let $P(\vec{d})$ be a current point on a plane wavefront, defined by the position vector $\vec{d}$ from the origin of the reference frame. The electric field propagation of a monochromatic wave is expressed as:

$$\vec{E}(\vec{d}, t) = \vec{E}_0 \exp[\mathrm{i}(\vec{k}\cdot\vec{d} - \omega t)] \qquad [1]$$

The light power carried by the electromagnetic field is proportional to $I = \|\vec{E}_0\|^2 = \vec{E}\cdot\vec{E}^*$ where $\vec{E}^*$ is the complex conjugate of $\vec{E}$. The time frequency $\nu = \omega/2\pi$, or the wavelength $\lambda = c/\nu$, where c is the electromagnetic wave celerity, characterize the hue of the light. The wavevector $\vec{k} = \vec{N}\omega/c$, where $\|\vec{N}\| = 1$ is an angular space frequency vector, as $\vec{n} = \vec{N}/\lambda = \vec{N}\nu/c$ is a space frequency vector. We have $\vec{k} = 2\pi\vec{n}$. As λ is a constant for a given monochromatic light, $\vec{n}$ defines the direction of propagation of a ray.

Spatially Coherent Light

Consider a very small source stated as the point source S_1, emitting uniform lightwaves in any direction (Figure 4). The wave reaching a farfield point, P, from S_1, propagates along a direction defined by the wavevector $\vec{k}_1$. The vectors $\vec{k}_1$ and $\vec{d}_1 = \overrightarrow{S_1P}$ are collinear. Another point source, S_2 is characterized from P by the wavevector $\vec{k}_2$, such that $\vec{k}_2$ and $\vec{d}_2 = \overrightarrow{S_2P}$ are collinear. When the distances d_1 and d_2 are close to each other, the amplitudes $\vec{E}_{01}$ and $\vec{E}_{02}$ of the waves at P are proportional to those of the corresponding emitting sources, S_1 and S_2. According to the principle of superposition, the resulting field at the point P, $\vec{E} = \vec{E}_1 + \vec{E}_2$, and the instant wavevector $\vec{k}$, are unique vectors.

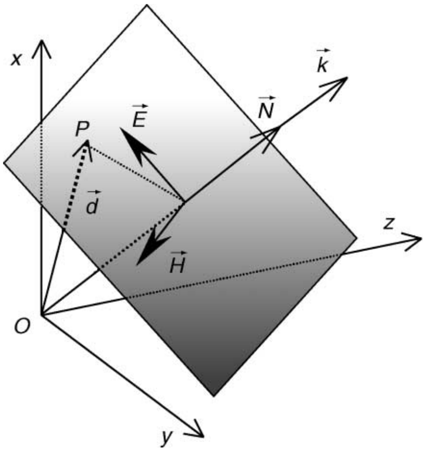

Figure 3 Electromagnetic wave propagation.

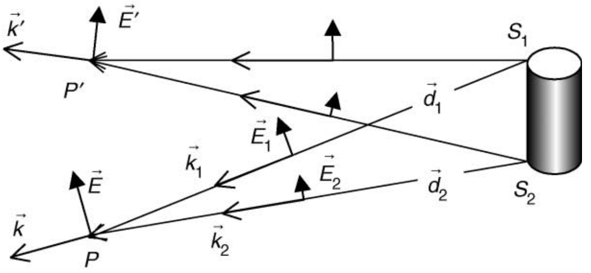

Figure 4 Spatially coherent light.

If S_1 and S_2 are two synchronous sources, the amplitude $\vec{E}_0$ of $\vec{E}$ and the wavevector $\vec{k}$ are constant versus time at any point P. This is still the case when an extended object is made of an arrangement of synchronous points. That situation defines a spatially coherent object. Thus, one cannot recover multiple data about the object's distribution, from the field framework at a single point P, since the relationship between the two is multi-unequivocal. Nevertheless, due to diffraction, one can retrieve them by knowing the complex amplitude distribution of light on a set of wavevectors available over an extensive area of a screen. Let $\vec{E}_0(\vec{r})$ be the amplitude of the electric field at the object point $M(\vec{r})$ and let $\vec{E}_0(\vec{n})$ be the one diffracted by the object to farfield toward the direction corresponding to the space frequency $\vec{n}$. Fourier transforms give the connection between $\vec{E}_0(\vec{r})$, and $\vec{E}_0(\vec{n})$:

$$\left\{\begin{aligned} \vec{E}_0(\vec{n}) &= \int \vec{E}_0(\vec{r})\exp[\mathrm{i}2\pi\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r}, \\ \vec{E}_0(\vec{r}) &= \int \vec{E}_0(\vec{n})\exp[-\mathrm{i}2\pi\vec{n}\cdot\vec{r}]\mathrm{d}\vec{n}. \end{aligned}\right\} \qquad [2]$$

This is the situation encountered when an object is illuminated by a laser or by a point source such as a single star.

Spatially Incoherent Light

When the point sources S_1 and S_2 are independent – spatially incoherent – both wavevectors $\vec{k}_1$ and $\vec{k}_2$

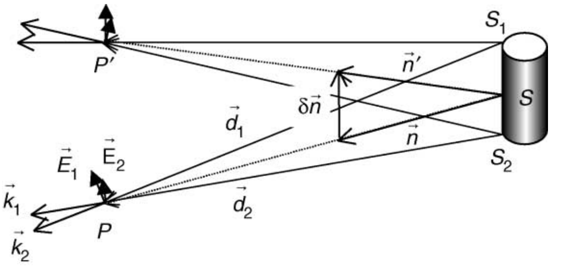

Figure 5 Spatially incoherent light.

coexist at the point P, each characterizing the angular position of S_1 and S_2 (Figure 5). If S_1 and S_2 are part of an extended intensity distribution $I_0(\vec{r})$, fans of wavevectors and fields are mixed at P, each coming from an independent point source S. Together, they carry the whole data on the object distribution. Those data are partially uncorrelated at any two points P and P' on a screen, related to the different optical paths from any point S to P and P' and to the finite time of coherence of light. The correlation between data available to farfield in two directions $\vec{n}$ and $\vec{n}'$ from the object, is given by the mutual intensity function, $\Gamma(\delta\vec{n}) = \int \vec{E}(\vec{n})\cdot\vec{E}^*(\vec{n}+\delta\vec{n})\mathrm{d}\vec{n}$, where $\delta\vec{n} = \vec{n}' - \vec{n}$. Fourier transforms give the relation between $I_0(\vec{r})$ and $\Gamma(\delta\vec{n})$:

$$\left\{\begin{array}{l} \Gamma(\delta\vec{n}) = \displaystyle\int I(\vec{r})\exp[\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}]\mathrm{d}\vec{r}, \\ I(\vec{r}) = \displaystyle\int \Gamma(\delta\vec{n})\exp[-\mathrm{i}2\pi\delta\vec{n}\cdot\vec{r}]\mathrm{d}\delta\vec{n} \end{array}\right\} \qquad [3]$$

Therefore, by drawing a map of $\Gamma(\delta\vec{n})$, one can recover the object distribution $I_0(\vec{r})$ by a Fourier transformation. Most of time, one uses the spatial degree of coherence function, $\gamma(\delta\vec{n}) = \Gamma(\delta\vec{n})/\Gamma(\vec{0})$. When the object illuminates a screen, at a large distance d, the degree of coherence of light between two points P and P' on the screen is $\gamma(\delta\vec{\rho}/\lambda d)$, with $\delta\vec{\rho} = \overrightarrow{PP'}$. The 2D extension of $\gamma(\delta\vec{\rho}/\lambda d)$ defines the area of coherence of light on the screen. The width of the area of coherence, $\Delta\rho$, and that of the object distribution, Δr, are inversely proportional:

$$\Delta\rho\cdot\Delta r \sim \lambda d \qquad [4]$$

If a mask perforated by two pinholes, separated by a distance $\delta\vec{\rho}$, replaces the screen, one can observe Young fringes of interference beyond the mask, the visibility of them being precisely $\gamma(\delta\vec{\rho}/\lambda d)$. Therefore, one can built a map of $\gamma(\delta\vec{\rho}/\lambda d)$ versus $\delta\vec{\rho}$ from a set of 2D measurements of visibilities of fringes, then reconstruct the object distribution by a Fourier transformation.

Notice that the entire data of the object distribution is available from a single coherence area. That is, a pinhole covering exactly one area of coherence would select the entire available data from the object and would display to farfield the most accurate image possible. Generally, such a pinhole is so small that practically no light passes through it. However, we can regard each area of coherence as an independent channel of information. Then an imaging device based on a juxtaposition of such reduntant channels can be considered as a multiplex imaging device.

Self-luminous objects such as extended or double stars, and artificial thermic sources are spatially incoherent sources.

Multiplex Imaging

Methods of multiplex – or lensless – imaging are performed either with coherent of incoherent light. The very different properties of the two types of light leads to completely different methods, even if all are related to holography. Holography with coherent light is an interference method for local recording of both the direction and modulus of the wavevectors and the amplitude of the field contained in many juxtaposed grains of speckle, scattered by the object. Each grain is an independent channel of imaging.

Synthetic Apertures with Spatially Incoherent Objects

As interferences are impossible by this method with incoherent light, we must improve other methods of imaging. Synthetic aperture is a concept consisting of using a redundant multichannel system to increase the luminosity of images without loss of resolution. The basic principle is that of a multiple holes darkroom. The processing is in two steps: recording and reading the image. The recording step consists in illuminating a photosensitive receptor with the object through a mask (Figure 6). Generally, the receptor is a

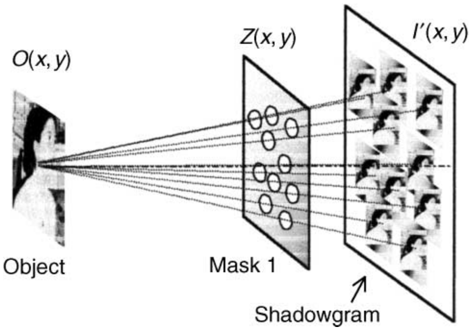

Figure 6 Principle of synthetic aperture: recording.

photographic plate. Obviously, the light distribution on the plate is blurred. Indeed, any elementary transparencies on the mask project on the receptor inverted images of the object located the shadow of the mask. Such a distribution is expressed by a convolution. Let $O(x, y)$ be the conical projection of the object on the receptor and $Z(x, y)$ that of the mask. The shadows distribution on the receptor is (the variable being normalized):

$$\begin{aligned} I'(x', y') &= Z(x, y) * O(x, y) \\ &= \int Z(x, y) O(x' - x, y' - y) \mathrm{d}x \, \mathrm{d}y \end{aligned} \qquad [5]$$

The reading step is a decorrelation processing from the recorded intensity (Figure 7). It consists in achieving a new correlation of I' with an arbitrary function T:

$$\begin{aligned} I''(x, y) &= T(x, y) * [Z(x, y) * O(x, y)] \\ &= [T(x, y) * Z(x, y)] * O(x, y) \end{aligned} \qquad [6]$$

$R[x, y] = [T(x, y) * Z(x, y)]$ is the system point spread function (SPSF) of the processing. It expresses the image distribution of a single-point. The perfect image is obtained when $R(x, y)$ is a 2D *Dirac* – or *delta*– distribution, $\delta(x, y)$:

$$R(x, y) = \delta(x, y), \quad \text{as} \quad \int \delta(x, y) O(x' - x, y' - y) \times \mathrm{d}x \, \mathrm{d}y = O(x', y') \qquad [7]$$

Fresnel zone plates

First Mertz and Young used Fresnel zone plate (FZP) as a synthetic aperture for X-ray imaging purposes. FZPs are made of alternate transparent and opaque concentric crowns of equal area. The radius of the nth circle is $r_1\sqrt{n}$, r_1 being the smallest. When illuminated by a cylindrical light beam, the FZP is a multiple foci diffractive lens (Figure 8).

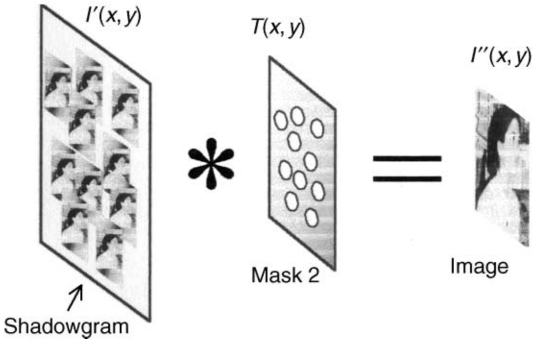

Figure 7 Synthetic aperture: reading.

The reading step is either analogical or numerical. Through the analogical method, we consider the recorded plate as a pseudohologram (Figure 9). When the 'shadowgram' is lighted with a parallel beam of laser light, the FZPs recorded around each point concentrates the light on its foci. Then an image of the object is reconstructed on the focal planes; a method that is simple and elegant. Nevertheless, the processing is not linear since we record an intensity distribution at the recording step, while the reading step by diffraction involves amplitude. Moreover, the S/N ratio decreases as the amount of data increases, recorded on the photographic plate because of its finite dynamic.

Mertz and Young had the idea of using a second FZP to perform the deconvolution processing by optical methods. In this case, the SPSF, $R(x, y)$, is the autocorrelation function of the FZP transparency (Figure 10). It presents a narrow central peak, emerging from a pyramidal ground (Figure 11). The optical setup is the same as the recording step.

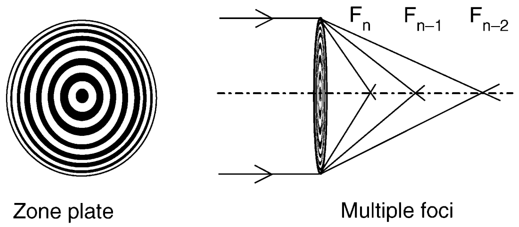

Figure 8 Properties of FZP.

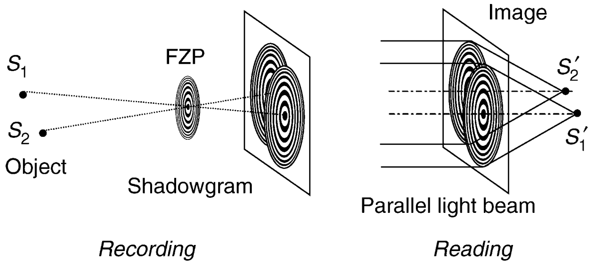

Figure 9 Recording and analogical reading with a FZP.

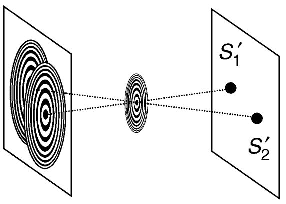

Figure 10 Optical decorrelation.

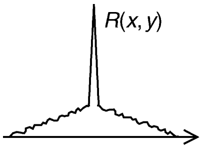

Figure 11 System point spread function with a Fresnel zone plate.

The FZP scale factor only must be adjusted to compensate for the magnification ratio generated by the conical projection at the recording step.

In either digital or analogical deconvolution, the image reconstructed is accurate, with good sharpness, if the number of zones is sufficient. Ceglio *et al.* obtained relevant images of 8 μm targets with FZP having 100 zones. Nevertheless, the multiple foci of FZP induces decreasing of the S/N ratio which is an important limitation. Besides, we pointed out above that the larger the object, the more the decrease in the S/N ratio.

Random and uniformly redundant arrays

Dicke and Ables substituted a random array to the FZP. The mask is made of randomly distributed holes and the expected advantage is doubled. First, one can reduce the transparent area of the mask by less than 50% in order to increase the S/N ratio when the object is wide. Second, the autocorrelation function of such an array is theoretically a Dirac distribution when its width is infinite. Thus, one can hope to narrow the system point response. However, when the width is finite, the pyramidal ground of the point response mentioned above is present and the advantage is not clearly demonstrated. In order to correct this effect, Chris Brown used a 'mismatched' function, for the postprocessing deconvolution. This function is calculated, first by achieving the complement of the recording mask function, then subtracting from it. Thus, the 1 transparencies result in +1 while the 0 transparencies are changed by −1. Brown shows images digitally reconstructed of X-ray sources, 30 sec. of arc wide. Fenimore and Cannon conceived a uniformly redundant array. The center random motif of the mask is half replicated on each side with only the center area being used for the recording step. The deconvolution is calculated by taking into account the whole area in a mismatched way. The result is an exact delta function. The authors show images of 200 μm-wide microspheres with a similar resolution but a signal 7100 times stronger than with a comparable pinhole camera.

Theta Rotating Interferometer

The theta shearing interferometer enables us to record complete data from the Fourier transform of a spatially incoherent object (Figure 12). The source can be any natural one, but monochromatic. This is easily practicable using a good filter or monochromator.

The device is based on a Michelson interferometer, the plane mirrors being replaced by two roof prisms. One of the roof prisms is fixed so the other can rotate around the axis of the interferometer. As a result, we have two images of the object. Let θ be the angle between the two edges of the prisms. The angle between the images is 2θ. Since the original object is noncoherent, and two homologous points being mutually coherent, the light they emit interferes, resulting in fringes.

The intensity, space frequency, and direction of the fringes characterize the intensity and the position of the point source. One shows that the farfield visibility distribution of fringes $V(\delta\vec{n})$ and the object distribution $O(\vec{r})$ are related by Fourier transforms:

$$\left\{\begin{aligned} V(\delta\vec{n}) &= \int O(\vec{r})\exp\left[\mathrm{i}2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]\frac{\mathrm{d}\vec{r}}{2\sin\theta}, \\ O(\vec{r}) &= \int V(\delta\vec{n})\exp\left[-\mathrm{i}2\pi\frac{\delta\vec{n}\cdot\vec{r}}{2\sin\theta}\right]\mathrm{d}\delta\vec{n} \end{aligned}\right\} \quad [8]$$

Then, one can see that the visibility of the interference pattern is an image of the degree of coherence $\gamma(\delta\vec{n})$, with a scale factor $1/2\sin\theta$.

Such a pattern recorded on a photographic plate is a Fourier hologram of the noncoherent object. An image can be reconstructed by diffraction at infinity or near a focus (see Figure 13, examples performed in 1972 by the author). The technique was successfully applied to the determination of the modulation transfer function of optical systems. C Roddier and F Roddier applied the method to astronomy.

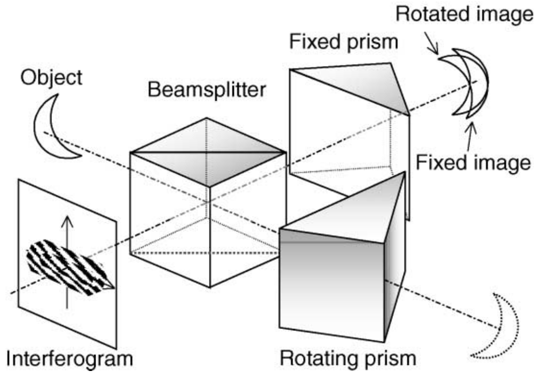

Figure 12 Theta shearing interferometer.

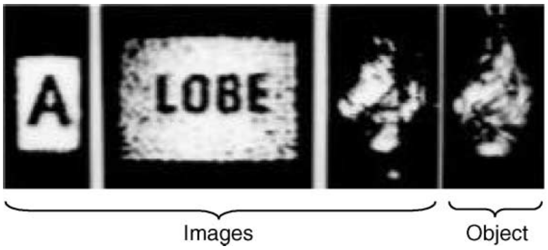

Figure 13 Incoherent light holograms.

For instance, they examined Alpha Orionis and discovered the evolution of its dust envelope and the possible existence of a companion.

High Resolution Imaging in Astronomy

In astronomy, the limit of resolution depends chiefly on two factors. The first is a technical one. Adaptive optics and very large telescopes (VLT) brought spectacular progress to clarity and resolution of telescopes. The 8.2 m VLT of Kueyen set up in 1989, shows images easily resolved of HIC 69495 Centauri, separated by 0.12 sec. of arc. Nevertheless, working of large mirrors for telescopes reaches its limits. Therefore, their clarity remains too weak. On the other hand, atmospheric turbulences make the refractive index of air unstable and the position of the images in the focal plane of telescopes unsettled. Thus, the point response is a granular spot named 'speckle'. Short exposure times relieve that inconvenience, but they prohibit the observation of weak stars. In 1970, Labeyrie proposed a method of speckle interferometry consisting of a statistical analysis of a set of speckles. The result is a display of a diffraction-limited image. He resolved many tens of single and double stars. Only Hubble, above the atmosphere, is able to do it by direct imaging. However, the primary mirror of Hubble being only 2.40 m wide, has clarity much lower. Stellar interferometry is a very promising alternative solution.

Principles of Stellar Interferometry

Stellar interferometry is based on eqn [3]. The light emitted by double or wide stars is noncoherent, so that one can access an image of them trough 2D measurements of the degree of coherence of light. We showed above that one could achieve this by using interference devices such as Young holes. Eqn [3] shows that if we measure the visibility of Young fringes at increasing distances $\delta\rho$ between the pinholes, the distance $\Delta\rho$ from which the visibility is zero gives the angular width of the star, $\Delta\alpha$, under the formula:

$$\Delta\rho \cdot \Delta\alpha \sim \lambda \qquad [9]$$

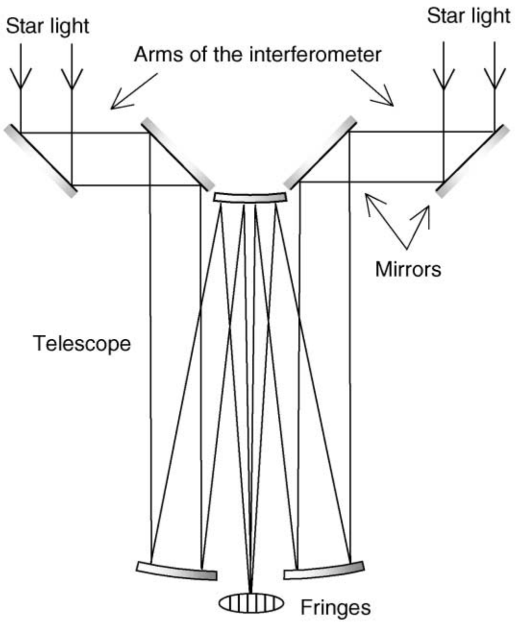

Figure 14 Stellar interferometer of Michelson.

where $\Delta\alpha = \Delta r/d$. In the case of a double star, the visibility varies periodically versus $\delta\rho$ when the pinholes are aligned along the axis of two stars. Let $\Delta\alpha$ be the angular distance between the stars: the smallest distance $\Delta\rho$ for which the visibility is zero is given by:

$$\Delta\rho \cdot \Delta\alpha = \lambda/2 \qquad [10]$$

The larger the uttermost distance between the holes, the more the limit of resolution of the processing is potentially sharp.

In order to exploit this idea, Stephan placed and rotated a mask with two holes in front of the objective of a refractive telescope. He established that the diameter of the stars he observed was much less than 0.16 sec. of arc. Then, Michelson built up, on top of a 30 cm refractive telescope, a stellar interferometer consisting of two arms catching the light at 7 m of distance from each other (Figure 14). He measured correctly the diameter of the satellites of Jupiter. Hale carried the distance of the base up to 15 m in 1920.

The Two Telescopes Interferometer of Labeyrie

In 1975, Labeyrie improved the idea by setting up two separate 25 cm diameter Cassegrain-coudé telescopes on a 12 m baseline, oriented north–south (Figure 15). The setting and mechanical stability of the device are extremely sensitive. The tolerances do not exceed any μm. A mobile table near the common focal plane of the two telescopes compensates for the optical path differences due to stars tracking. In spite

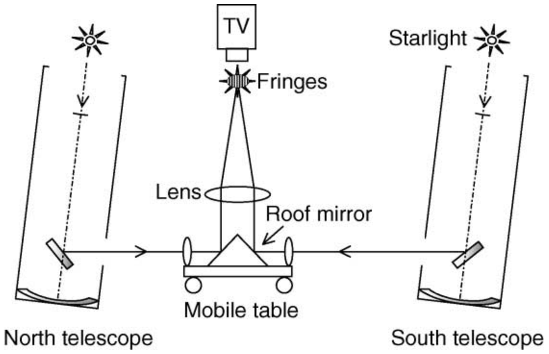

Figure 15 The two telescopes interferometer of Labeyrie.

of the acute difficulties, he succeeded in observing fringes on Vega. He obtained a resolution of 0.01 sec. of arc versus 0.5 sec. with a single 25 cm telescope. Later, one telescope was rendered mobile bringing the base width variable between 4 and 67 m. The resolution reaches up to 0.00015 sec. of arc. Note that the rotation of Earth also enables east–west analysis with resolution of 0.005 sec., and today, the 25 cm telescopes are replaced by 1.5 m ones.

The Very Large Telescope Interferometer (VLTI)

The two telescopes interferometer of Labeyrie is a first step toward a very ambitious project: the telescope at synthetic aperture and very large base. Proofs of reliability of such a principle was already given by the very large array (VLA) experimented in radio astronomy. Labeyrie with Lena, and the team of the ESO (Ecole Supérieure d'Optique, Paris), imagined a very large telescope interferometer (VLTI) made of 6 telescopes of 2 m, in a 2D structure on a base of 100 m. The European VLTI was built at Cerro Paranal. It is made of four 8.2 m telescopes corrected in real time by adaptive optics (NTT technology) and three 1.8 m telescopes. The fourth 8.2 m telescope was put in service in 2001. Some days, each of the 8.2 m telescopes used separately gives a resolution close to that of Hubble. However, the clarity is much better because of the larger primary mirror diameter (8.2 m against 2.4 m). In May 2003, the team of J-G Cuby observed a galaxy the farthest away ever seen, only 900 millions years old since the Big Bang. Obviously, a drastic gain of resolution (by a factor 10) is expected from the interference configuration.

See also

Coherent Lightwave Systems. **Diffraction:** Fresnel Diffraction. **Imaging:** Adaptive Optics; Interferometric imaging. **Instrumentation:** Telescopes. **Interferometry:** Overview.

Further Reading

Labeyrie A (1976) High resolution techniques in optical astronomy. In: Wolf E (ed.) Progress in Optics, vol. XIV, pp. 49–85. New York: North-Holland.

Mertz L (1965) *Transformations in Optics*. Amsterdam, Oxford: John Wiley & Sons, Inc.

Photon Density Wave Imaging

V Toronov, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photon density wave imaging (PDWI) employs diffused light, usually in the near-infrared band, for imaging inhomogeneous structures in a highly scattering medium. The other names for PDWI are diffuse optical tomography and photon migration imaging. The term photon density wave (PDW) refers to the wave of light density spreading in the highly scattering medium from the source whose power changes in time. Therefore, rigorously speaking, photon density wave imaging uses intensity-modulated light sources, such as pulsed lasers, or sources whose power is harmonically modulated. However, many applications employ steady-state light sources. Most applications of PDWI are in the field of imaging of biological tissues, including medicine and cognitive neuroscience.

As in the other imaging methods, such as x-ray, computed tomography (CT) or ultrasound imaging, the problem of PDWI consists in the reconstruction of the internal structure of the medium from the measurements made outside the medium (on its surface). In the case of PDWI the internal structure of the medium is represented by the spatial distribution of its optical properties. The reconstruction procedure is essentially an iterative mathematical algorithm, which includes the modeling of the light propagation inside the medium assuming a certain distribution of the optical properties, and the correction of this distribution to match the model prediction with the actual measurement. The modeling part of the imaging problem is usually referred to as the forward problem, and the correction procedure is known as an inverse problem.

Forward Problem: Photon Density Waves

Unlike optically transparent media, in a highly scattering medium the coherence of light from the coherent source very rapidly decays with distance. This happens due to the interference of many randomly and multiply scattered light waves at every point of the medium that is distanced from the source much farther than the photon mean free path l, which is the mean distance at which a light wave undergoes a single scattering event. For biological tissues, where light scattering occurs on cell membranes and other heterogeneities of the index of refraction, l is typically about 1 mm. Although high scattering eliminates the coherence of the electromagnetic waves at the short time intervals corresponding to oscillation at optical frequencies ($\sim 10^{-15}$ s), in the case of an intensity-modulated light source, the macroscopic wave of the density of light energy, called 'photon density wave', forms in the highly scattering medium. The properties of PDW depend on the type of the source modulation and on the local optical properties of the medium, which are the scattering coefficient μ_s [cm^{-1}] and the absorption coefficient μ_a [cm^{-1}].

Because of the complexity and incoherence of the macroscopic electromagnetic field in highly scattering media, i.e. media with $\mu_a \ll \mu_s$, the fundamental Maxwell equations of electrodynamics do not provide a convenient approach to the solution of the problem of light propagation in such media at distances much larger than l. Therefore, statistical approaches have been developed, such as the photon transport model and its diffusion approximation. In the photon transport model the light is characterized by the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-2} steradian^{-1}] (where the steradian is the unit solid angle), which is the power of radiation near the point $\mathbf{r}$ propagating within a small solid angle around the direction $\hat{\Omega}$ at time t. The photon transport model accounts for the possible nonisotropy of the radiation flux. For the given spatio-temporal distribution of the optical properties, μ_a and μ_s, and for the given spatial and angular distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-3} steradian^{-1}] the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ can be found from the linear radiation transfer equation, also called the photon transport equation, or simply the transport equation (TE):

$$\frac{1}{v}\frac{\partial L(\mathbf{r}, \hat{\Omega}, t)}{\partial t} + \nabla \cdot L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega} + (\mu_s + \mu_a)L(\mathbf{r}, \hat{\Omega}, t) = \mu_s \int L(\mathbf{r}, \hat{\Omega}', t) f(\hat{\Omega}, \hat{\Omega}')\mathrm{d}\hat{\Omega}' + S(\mathbf{r}, \hat{\Omega}, t) \qquad [1]$$

where $f(\hat{\Omega}, \hat{\Omega}')$ is the phase function, which represents the portion of light energy scattered from the direction $\hat{\Omega}$ to the direction $\hat{\Omega}'$, and v is the speed of light in the medium. TE is equivalent to the Boltzmann transport equation describing the transport of particles undergoing scattering.

Equation [1] is difficult to solve even numerically. Therefore, a hierarchy of approximations to the PTM was developed. This hierarchy, known as the P_N approximation, is based on the expansion of the radiance in a series of spherical harmonics $Y_{nm}(\hat{\Omega})$ truncated at $n = N$. In the P_1 approximation the coefficients of the series are proportional to the radiation fluence

$$\Phi(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t)\mathrm{d}\hat{\Omega}, \qquad [2]$$

and the flux

$$J(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t)\hat{\Omega}\mathrm{d}\hat{\Omega}, \qquad [3]$$

both having units of [W cm^{-2}]. By substituting the P_1 series for $L(\mathbf{r}, \hat{\Omega}, t)$ and $S(\mathbf{r}, \hat{\Omega}, t)$ into eqn [1] and assuming isotropic light sources and sufficiently slow change of source intensity, one can obtain the equation that includes only $\Phi(\mathbf{r}, t)$ as the unknown. Usually, this equation

$$v\mu_a U(\mathbf{r}, t) + \frac{\partial U(\mathbf{r}, t)}{\partial t} - D\Delta U(\mathbf{r}, t) = Q(\mathbf{r}, t) \qquad [4]$$

is written in terms of the value $U(\mathbf{r}, t) = \frac{1}{v}\Phi(\mathbf{r}, t)$ [J cm^{-3}], which has units of energy density, and is called the photon density. In eqn [4] $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s] is the light diffusion coefficient, and $\mu_s' = \mu_s(1 - g_1)$ [cm^{-1}] is the reduced scattering coefficient; g_1 is the average cosine of the scattering angle. The source term $Q(\mathbf{r}, t)$ [J cm^{-3} s^{-1}] in the right-hand side of the equation [4] describes the spatial distribution of light sources (in the isotropic approximation, i.e., neglecting the dipole and higher momenta of $S(\mathbf{r}, \hat{\Omega}, t)$) and the temporal modulation of source intensity.

Equation [4] is mathematically equivalent to the equation describing the density of the medium consisting of the particles undergoing linear diffusion. Therefore, it is called the diffusion equation (DE), and the P_1 approximation is also known as the diffusion approximation. The diffusion approximation is valid under the following three conditions: (a) the light sources are isotropic or the distance between the source and the detector is much larger than l; (b) characteristic source modulation frequencies ν are low enough that $v{\mu_s}'/(2\pi\nu) \gg 1$; and (c) $\mu_a \ll \mu_s'$. In the biomedical applications of PDWI the condition (a) is fulfilled when the source–detector distance is larger than 1 cm. The condition (b) is valid when the characteristic modulation frequencies of the light source are less than 1 GHz. The condition (c) is fulfilled in most biological tissues, for which the scattering is highly isotropic, i.e. $g_1 \ll 1$. Thus, the forward problem of PDWI, i.e., the modeling of light propagation, can be solved using the diffusion approximation in a large variety of situations.

PDWs described by equation [4] are the scalar damped traveling waves, characterized by the coherent changes of light energy (photon density) in different locations on time intervals from microseconds to nanoseconds. Usually pulsed or harmonically modulated light sources are used to generate PDWs. The properties of the PDWs from a harmonic source are fundamental, because the PDW from the source of any type of modulation can be considered as a superposition of harmonic waves from sources having different frequencies using the Fourier transform.

In order to get an idea of the basic properties of PDW, one can consider the analytical solution of [4] for the case of a harmonically modulated point-like source placed in an infinite homogeneous scattering medium. This solution is

$$U(r,t) = U_{\mathrm{dc}}(r) + U_{\mathrm{ac}}(r)\exp[-\mathrm{i}\omega t + \mathrm{i}\phi(r)] \qquad [5]$$

where U_{ac}, U_{dc}, and ϕ are the ac, dc and phase, respectively, of the wave measured at the distance r from the source, and $\omega = 2\pi\nu$. U_{ac}, U_{dc}, and ϕ depend on the medium optical properties, source intensity Q_0, modulation amplitude A_0, and initial phase ψ_0 as

$$U_{\mathrm{ac}}(r) = \frac{Q_0 A_0}{4\pi D} \times \frac{\exp\left[-r\left(\frac{v\mu_{\mathrm{a}}}{2D}\right)^{1/2}((1+x^2)^{1/2}+1)^{1/2}\right]}{r} \qquad [6a]$$

$$U_{\mathrm{dc}}(r) = \frac{Q_0}{4\pi D}\;\frac{\exp\left[-r\left(\frac{v\mu_{\mathrm{a}}}{D}\right)^{1/2}\right]}{r} \qquad [6b]$$

$$\phi(r) = r\left(\frac{v\mu_a}{2D}\right)^{1/2}\sqrt{(1+x^2)^{1/2}-1} \qquad [6c]$$

where $x = \omega/(v\mu_{\mathrm{a}})$. Equations [5]–[6] describe a spherical wave whose wavelength at $\mu_{\mathrm{a}} = 0.1\,\mathrm{cm}^{-1}$ and $\mu_s' = 8\,\mathrm{cm}^{-1}$ (typical for biological tissues) and at modulation frequency $\nu = 100$ MHz, is close to 35 cm, so that the phase of the wave changes approximately by 10° per cm. In the same time, the wave decays exponentially at a distance of about 3 cm, so that PDWs are strongly damped. A similar analytical solution in the form of a spherical wave can be obtained for the case of a semi-infinite homogeneous medium and point source placed on the surface of the medium (see Figure 1). These analytical solutions are employed to measure μ_{a} and μ_s' of the homogeneous medium. The method is based on fitting measured ac, dc, and phase of the PDW as functions of source–detector distance by the analytical solution.

Inhomogeneities in the optical properties of a turbid medium cause distortion of the PDW from sphericity (see Figure 1). PDWs in homogeneous media of limited size or with strong surface curvature also usually have complex structure. For a limited number of configurations one can obtain solutions to the DE using analytical methods of mathematical physics, such as the method of Green's functions. However, in practice the complexity of the inhomogeneities and surfaces usually requires numerical solution of the DE. This can be achieved using numerical methods for the solution of partial differential equations, such as the finite element method and the finite difference method. The finite difference method can also be used to solve the TE. This may be necessary in situations when the DE is not valid, for example to accurately account for the influence of the cerebrospinal fluid (which has relatively low scattering) on light transport in the human skull. However, the solution of the TE requires formidable computational resources, which increase proportionally to the volume of the medium. Therefore, hybrid DE-TE methods were developed, in

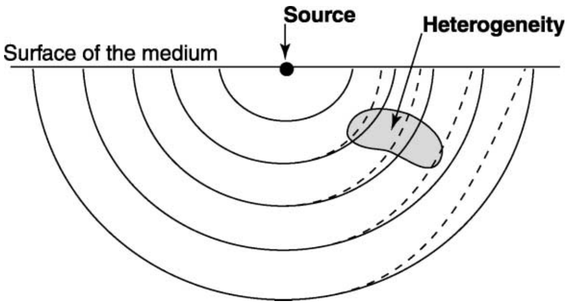

Figure 1 Equiphase surfaces of the PDW induced by a point source in a semi-infinite homogeneous medium (solid lines) and in a medium including heterogeneity (dashed lines).

which DE is used to describe light propagation in most of the medium except small low-scattering areas described by the TE.

One should note that although the term 'photon' suggests quantum properties of light, no consistent application of quantum theory to problems related to PDWI has been developed so far. Also, it should be noted that the notion of a photon as a particle of light having certain energy, momentum, and localization at the same time is not completely consistent with quantum electrodynamics. However, since the theoretical models of light transport in a scattering medium (such as TE and DE) are mathematically equivalent to the models describing systems of randomly moving particles, the notion of a photon as a 'particle of light' may be accepted as adequate and is widely used in the limits of PDWI.

Inverse Problem and Imaging Algorithm

Mathematically the forward problem can be expressed in the form of a general nonlinear operator F:

$$G = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}] \qquad [7]$$

where G is the set of measured values (ac, dc, and phase), $\mathbf{r}$ is the coordinate vector of a medium element (voxel), and $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ denotes the set of tissue optical properties corresponding to whole collection of voxels. The problem of finding the spatial distribution of medium optical properties can be formalized as

$$\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} = F^{-1}[G] \qquad [8]$$

Neither the transport model nor the diffusion model allow inversion in a closed analytical form for a general boundary configuration and spatial distribution of optical properties. The basic reason for the complexity of the inverse problem is in the complexity of the light paths in the highly scattering medium. This situation is opposite to that in CT, where the x-ray radiation propagates through the medium along straight lines almost without scattering, which significantly simplifies the forward problem and allows its analytical inversion. From the mathematical point of view, the complexity of the inverse problem of PDWI is due to its nonlinearity and ill-posedness. Only approximate methods of inversion have been developed so far.

The simplest method is called backprojection by analogy with the method of the same name used in CT. The method attempts to reconstruct the actual spatial distribution of the optical properties μ_a and/or μ_s' from the 'measured' bulk values μ_a^m and $\mu_s'^m$ obtained at a variety of locations assuming homogeneity of the medium. The measured values are considered as a result of the volumetric 'projection'

$$\{\mu_a^m(\mathbf{r}), \mu_s'^m(\mathbf{r})\} = \int \{\mu_a(\mathbf{r}'), \mu_s'(\mathbf{r}')\} h(\mathbf{r} - \mathbf{r}')\mathrm{d}^3\mathbf{r}' \qquad [9]$$

of the actual distribution, where $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the measurement and reconstruction, respectively. A weight function $h(\mathbf{r})$ accounts for the contribution of the particular region to a measured value and should be derived from the forward model. Different versions of the backprojection method use different weight functions $h(\mathbf{r})$. In the backprojection method the values of μ_a and/or μ_s' are obtained by inverting the convolution [9]. In practice the convolution integral is usually estimated by a sum over a limited number of voxels for a limited number of measurements. Therefore, the inversion of the convolution reduces to the solution of the set of algebraic equations corresponding to [9]. The advantage of the method is high speed, which allows real-time imaging of the dynamic processes, for example, functional hemodynamic changes in the brain. However, the backprojection method provides a phenomenological rather than a self-consistent inversion of the forward problem, and its quantitative accuracy is low.

A more consistent method is based on the perturbation approach to the solution of the inverse problem [8]. If we have an estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ that is close to the ideal solution, then $G_e = F[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]$ (see eqn [7]) is close to S. Then the difference $G - G_e$ can be written as a Taylor series:

$$G - G_e = J[\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e]\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} + \ldots \qquad [10]$$

where $J[\ldots]$ is the Jacobian, and $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$. Neglecting terms after the first one in [10], the problem of finding the correction $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\}$ to the estimate $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ reduces to the calculation of the Jacobian and to the solution of a linear algebraic equation.

This approach provides a basis for the PDWI algorithm, which is the following. First one makes an initial guess at the optical properties of the medium. Then one calculates the ac, dc, and phase of the PDW using a model of light propagation through the tissue and the geometry of the source–detector configuration. The calculated values are compared with the measured ones. If the error is above a set tolerance level, one calculates corrections to the current values of μ_a and μ_s' by solving eqn [10]. The algorithm is completed when the required error level is achieved. Thus, the PDWI algorithm iteratively applies both the forward and inverse problems. Since at each of many iterations one should solve the forward problem, which in the simplest case is a partial differential equation, the algorithm typically requires considerable computational resources.

One should note that the success of the method is largely dependent on the closeness of the initial guess to the actual values of the optical properties. In many studies the ‘first guess’ of the tissue optical properties was obtained by measuring the bulk absorption and reduced scattering coefficients. However, because of the ill-posedness of the inverse problem, homogeneous initial conditions did not always lead to the correct reconstruction. A more productive approach to the problem of the initial conditions is to assign approximate initial values of the optical properties known for example from *in vitro* measurements to different structures revealed by nonoptical imaging techniques, such as magnetic resonance imaging (MRI) or CT.

Applications

PDWI was proposed and developed as a noninvasive biomedical imaging technique. Although the spatial resolution of PDWI cannot compete with that of CT or MRI, PDWI has the advantages of low cost, high temporal resolution, and high biochemical specificity. The latter is based on the optical spectroscopic approach. Biological tissues are relatively transparent to light in the near-infrared band between 700 and 1000 nm (a near-infrared window). The major biological chromophores in this window are oxy- and deoxyhemoglobin, and water. One can separate contributions of these chromophores into light absorption by employing light sources at different wavelengths. Multi-wavelength PDWI can also be used to measure the redox state of the cytochrome oxidase, which is the terminal electron acceptor of the mitochondrial electron transport chain and responsible for 90% of cellular oxygen consumption.

One of the hot topics in cognitive neuroscience is the hemodynamic mechanism of the oxygen supply to neurons and removal of products of cerebral metabolism. High temporal resolution and biochemical specificity make PDWI a unique noninvasive tool for studies of functional cerebral hemodynamics, although light can penetrate no deeper than the very surface of the adult brain. Figure 2 displays a typical arrangement of light sources and detectors used in PDWI of functional cerebral hemodynamics, and the images of the activated area in the brain.

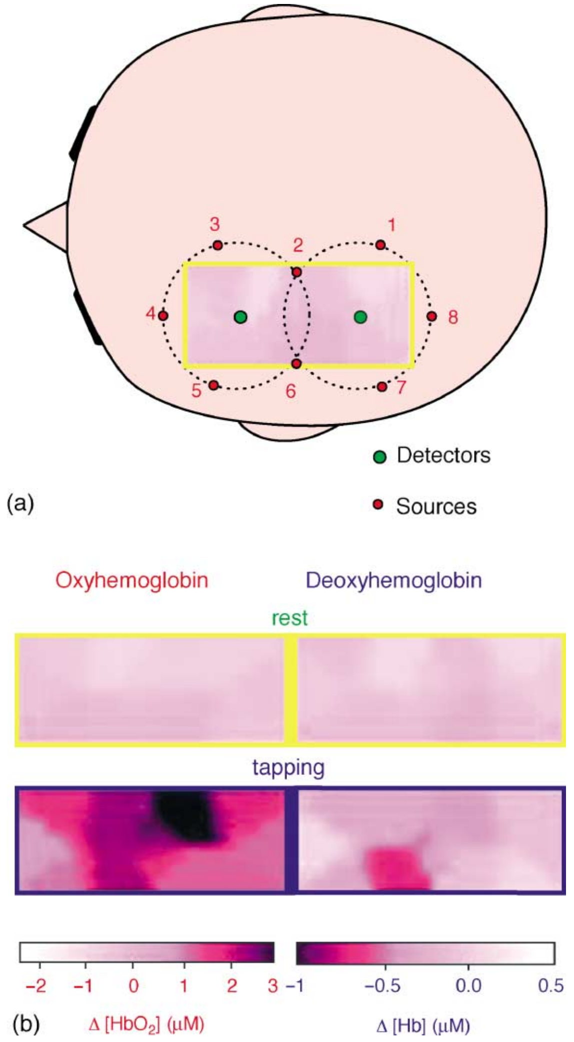

Figure 2 (a) Geometrical arrangement of the 16 sources and two detectors on the head of a human subject. Each one of the eight small red circles (numbered 1–8) represents one pair of source fibers at 758 and 830 nm. Two larger green circles represent detector fiber bundles. The rectangle inside the head outline is the $4 \times 9\ \text{cm}^2$ imaged area. (b) Maps of oxyhemoglobin ([HbO$_2$], left panels) and deoxyhemoglobin ([Hb] right panels) concentration changes at rest (top) and at the time of maximum response during the third tapping period (bottom). The measurement protocol involved hand-tapping with the right hand (contralateral to the imaged brain area). Adapted with permission from Franceschini MA, Toronov V, Filiaci M, Gratton E, Fantini S (2000) On-line optical imaging of the human brain with 160-ms temporal resolution. *Optics Express* **6**(3), 49–57 (Figures 1 and 4).

Because of its low cost, noninvasiveness, and compact size of the optical sensors, PDWI is also considered to be a promising tool for mammography and structural imaging of anomalies in the brain. PDWI is an especially promising tool for neonatology, because the infant tissues have low absorption.

PDWI instruments employ lasers, light-emitting diodes and gas-discharge tubes as sources of light. Among the detectors there are photomultiplier tubes, avalanche photodiodes, and CCD cameras. Both sources and detectors can be coupled with the tissue by means of optical fibers. A number of commercially produced near-infrared instruments for PDWI are available on the market. According to the type of source modulation, there are three groups of instruments: frequency domain (harmonically modulated sources), time domain (pulsed sources), and steady state. Steady-state instruments are simpler and cheaper than frequency-domain and time-domain ones, but their capabilities are much lower. Time-domain instruments can theoretically provide a maximum of information about the tissue structure, but technically it is difficult to build a time-domain instrument with high temporal resolution and high signal-to-noise ratio. Frequency-domain instruments provide temporal resolution up to several milliseconds and have high quantitative accuracy in the determination of optical properties.

Nomenclature

Absorption coefficient, μ_a [cm^{-1}]
Alternating part of PDW, alternating current (ac)
Average cosine of the scattering angle, g_1
Circular modulation frequency, ω [s^{-1}]
Computed tomography (CT)
Difference between the estimate and actual optical properties, $\{\Delta\mu_a(\mathbf{r}), \Delta\mu_s'(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
Diffusion coefficient, $D = v/(3\mu_a + 3\mu_s')$ [cm^2/s]
Diffusion equation (DE)
Distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-3} steradian^{-1}]
Divergence operator, ∇ [cm^{-1}]
Estimate of optical properties, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}_e$ [cm^{-1}]
General nonlinear operator, F
Jacobian, J
Laplace operator, Δ [cm^{-2}]
Magnetic resonance imaging (MRI)
Measured bulk optical properties (absorption and reduced scattering coefficients) $\{\mu_a^m \mu_s'^m\}$ [cm^{-1}]
Modulation frequency, ν [s^{-1}]
Non-alternating part of PDW, direct current (dc)
Optical properties of the medium (absorption and reduced scattering coefficients) as functions of spatial coordinate, $\{\mu_a(\mathbf{r}), \mu_s'(\mathbf{r})\}$ [cm^{-1}]
Phase as a function of source–detector distance $\phi(r)$
Phase function, $f(\hat{\Omega}, \hat{\Omega}')$
Photon density, $U(\mathbf{r}, t)$ [J cm^{-3}]
Photon density wave imaging (PDWI)
Photon density wave (PDW)
Photon mean free path, l [cm]
Radiance, $L(\mathbf{r}, \hat{\Omega}, t)$ [W cm^{-2} steradian^{-1}]
Radiation fluence, $\Phi(\mathbf{r}, t)$ [W cm^{-2}]
Radiation flux, $J(\mathbf{r}, t)$ [W cm^{-2}]
Reduced scattering coefficient, μ_s' [cm^{-1}]
Scattering coefficient, μ_s [cm^{-1}]
Set of measured values (ac, dc, and phase), G
Set of predicted values (ac, dc, and phase), G_e
Source term $Q(r, t)$ [J cm^{-3} s^{-1}]
Source–detector distance, r [cm]
Spatial coordinate, $\mathbf{r}$ [cm]
Speed of light in medium, v [cm/s]
Spherical harmonics $Y_{nm}(\hat{\Omega})$
Transport equation (TE)
Weight function, $h(\mathbf{r})$
$x = \omega/(v\mu_a)$

See also

Coherence: Coherence and Imaging. **Fiber and Guided Wave Optics:** Light Propagation. **Imaging:** Imaging Through Scattering Media. **Optical Materials:** Heterogeneous Materials. **Scattering:** Scattering Theory.

Further Reading

Arridge SR (1999) Optical tomography in medical imaging. *Inverse Problems* 15: R41–R93.

Boas DA (1996) *Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media: Theory and Biomedical Applications*. PhD Thesis, University of Pennsylvania.

National Research Council (US), Institute of Medicine (US) (1996) *Mathematics and Physics of Emerging Biomedical Imaging*. Washington, DC: National Academy Press.

O'Leary MA (1996) *Imaging with Diffuse Photon Density Waves*. PhD Thesis, University of Pennsylvania.

Special Issue (1997) Near-infrared spectroscopy and imaging of living systems. *Philosophical Transactions of the Royal Society B* 352(1354).

Tuchin V (2002) *Handbook of Optical Biomedical Diagnostics*. Bellingham, WA: SPIE Press.

Tuchin VV, Luo Q and Ulyanov SS (2001) *Imaging of Tissue Structure and Function*. Proceedings of SPIE 4427. Bellingham, WA: SPIE Press.

Villringer A and Chance B (1997) Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience* 20: 435–442.

Villringer A and Dirnagl U (1993) *Optical Imaging of Brain Function and Metabolism*. New York: Plenum Press.

Villringer A and Dirnagl U (1997) *Optical Imaging of Brain Function and Metabolism 2: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. New York: Plenum Press.

Three-Dimensional Field Transformations

R Piestun, University of Colorado, Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Conventional optical imaging is understood as an isomorphic mapping from an object space onto an image space. The object space is composed of scattering or emitting objects. The scattering objects can be opaque, transparent, or partially absorbing, while emitting objects can be as diverse as stars in astronomical imaging or fluorescent molecules in microscopic imaging. The image space is an optical field representation of the object.

In most classical applications of optical imaging the goal is to optimize the mapping between one plane in the object space and one plane in the image space. Modern optical systems, though, are aimed at acquiring three-dimensional information regarding the object. The image is thus a three-dimensional field representation of the object.

While many imaging systems are aimed at obtaining visually appealing images, many modern imaging systems register the field information in some type of sensitive device such as film or CCD cameras for later processing. This processing is typically performed electronically by a computer and the final image is ultimately a mathematical digital representation of the object. This type of imaging is called computational imaging. An example of computational imaging is optical tomography, which involves the registration of a set of images that do not necessarily resemble the object, and off-line digital processing to obtain a three-dimensional representation of the object.

In this article we will focus on imaging in the conventional sense but will also focus on the three-dimensional aspects of the image.

Geometrical Optics Transformations

In many optical problems it is appropriate to consider the wavelength of the electromagnetic wave as infinitely small. This approximation is the origin of the field of geometrical optics in which the propagation of light is formulated in terms of geometrical laws. Accordingly, the energy is transported along light rays and their trajectories are determined by certain differential or integral equations that can be directly derived from Maxwell's equations. One of these formulations uses Fermat's principle that states the principle of the shortest optical path. The optical path between two points P_1 and P_2 is defined as:

$$OP = \int_{P_1}^{P_2} n(P)\mathrm{d}s \qquad [1]$$

where n is the refractive index of the medium. Fermat's principle states that the optical path of a real ray is always a local minimum.

In ideal optical imaging systems, every point P_o of the so-called object space will generate a sharp image point P_i, that is a point where infinite rays emerging from P_o pass through P_i. Perfect imaging occurs when every curve defined by object points is geometrically similar to its image. Perfect imaging can occur between three-dimensional spaces or between surfaces.

An interesting result of geometrical optics is Maxwell's theorem. This theorem states that if an optical instrument produces a sharp image of a three-dimensional domain, the optical length of any curve in the object is equal to the optical length of its image. More explicitly, if the object space has index n_o and the image space index n_i, the theorem states that:

$$\int_{C_\mathrm{o}} n_\mathrm{o}\,\mathrm{d}s_\mathrm{o} = \int_{C_\mathrm{i}} n_\mathrm{i}\,\mathrm{d}s_\mathrm{i} \qquad [2]$$

where C_i is the image of the object curve C_o.

An important consequence of Maxwell's theorem is a theorem by Carathéodory which states that any imaging transformation has to be a projective transformation, an inversion, or a combination of them. Another consequence of this theorem is that if the two media are homogeneous, perfect imaging can only occur if the magnification is equal to the ratio of the refractive indices in the object and image space. If the media have the same index, the imaging is trivial with unit magnification. The simplest example of a perfect imaging instrument is a planar mirror. Therefore, if we want to achieve nontrivial imaging between homogeneous spaces with equal index, we must give up either the perfect imaging (similarity) or the sharpness of the image.

The geometrical optics approximation is very useful in the design of optical imaging systems. However, it fails to provide a complete description of wave optical effects, notably diffraction. Therefore, the following sections provide a description of field transformations based on wave optics.

Systems Approach to 3D Imaging

In this article we deal with linear imaging systems. Hence the relation between three-dimensional (3D) input (object) signals and 3D output (image) signals can be completely specified by the 3D impulse response or point spread function (PSF) of the system $h(x, y, z; x'y', z')$. This function specifies the response of the system at image space coordinates (x, y, z) to a δ-function input (point source) at object coordinates (x', y', z'). The optical system object and image signals are related by the superposition integral:

$$\mathrm{i}(x,y,z) = \int\int\int_{\Omega} \mathrm{o}(x',y',z')h(x,y,z;x'y',z')\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}z' \tag{3}$$

where Ω is the object domain. It is clear that for a generic linear imaging system the image signal is completely specified by the response to input impulses located at all possible points of the image domain.

If the system is shift invariant, the PSF due to a point source located anywhere in the object domain will be a linearly shifted version of the response due to a reference point source. This approximation is in general not valid and the system's PSF is a function of the position of the point source in the object domain. Shift invariance is traditionally divided into radial and axial shift invariance. Radial shift variance is the change in the image of a point source as a function of its radial distance from the optical axis. Radial shift variance is caused by lens aberrations. Axial shift variance is the change in the image of a point source as a function of its axial position, z', in the object space. Axial shift variance is inherent to any rotationally symmetric lens system because they can only have one pair of conjugate surfaces, i.e., object–image corresponding surfaces.

In most practical cases the image is detected at a single plane, $z = z_1$ transverse to the optical axis Oz. In those situations, we need to specify the impulse response on this plane:

$$h(x, y, z_1; x'y', z') = h_{z'}(x, y; x'y') \tag{4}$$

The condition for lateral shift invariance is then expressed as $h_{z'}(x, y; x'y') = \tilde{h}_{z'}(x - x', y - y')$. In this case, the superposition integral is a two-dimensional (2D) convolution and it is possible to define the transfer function of the system that is the 2D Fourier transform of the impulse response $\tilde{h}_{z'}(x - x', y - y')$.

Point Spread Function for Coherent, Incoherent, and Partially Coherent Illumination

The input–output relations of an optical imaging system relate different magnitudes according to the type of illumination. If the illumination is coherent, the system is linear in the complex amplitude, while if the illumination is fully incoherent the system is linear in intensity. The 3D impulse response of the incoherent (h_{Inc}) and coherent (h_{Coh}) systems are related as follows:

$$h_{\mathrm{Inc}}(x, y, z; x'y', z') = \left|h_{\mathrm{Coh}}(x, y, z; x'y', z')\right|^2 \tag{5}$$

Notwithstanding, many optical systems cannot be considered fully coherent or fully incoherent. In these situations the system is said partially coherent. If the optical path differences within the optical system are much shorter than the coherence length of the illumination, the light effectively has complete temporal coherence and the magnitude of interest is the mutual intensity. The mutual intensity is defined as the statistical correlation of the complex wavefunction $U(\mathbf{r}, t)$ between different points in space:

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1, t)U(\mathbf{r}_2, t)\rangle \tag{6}$$

where $\mathbf{r}_i = (x_i, y_i, z_i), i = 1, 2$. The illumination is then called quasi-monochromatic. For the case of partially coherent quasi-monochromatic illumination, the system is linear in the mutual intensity. The partially coherent impulse response (h_{PC}) is also related to the coherent impulse response as follows:

$$h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) = h^*_{\mathrm{Coh}}(\mathbf{r}_1; \mathbf{r}'_1)h_{\mathrm{Coh}}(\mathbf{r}_2; \mathbf{r}'_2) \tag{7}$$

The 3D superposition integral that relate object and image mutual intensities is now:

$$J_{\mathrm{image}}(\mathbf{r}_1, \mathbf{r}_2) = \int\int\int_{\Omega} h_{\mathrm{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) \times J_{\mathrm{object}}(\mathbf{r}'_1, \mathbf{r}'_2)\mathrm{d}\mathbf{r}'_1\,\mathrm{d}\mathbf{r}'_2 \tag{8}$$

Therefore, we conclude that knowledge of the coherent PSF is enough to characterize the optical system for illumination of various degrees of coherence.

Spatial Frequency Content of the 3D Point Spread Function

As previously stated, the coherent impulse response of the optical system completely defines its imaging characteristics. Therefore, it is justified to wonder what the limitations are in the specification of this response. In this section we analyze the frequency domain of coherent and incoherent impulse responses.

The impulse response of the optical system will satisfy the corresponding wave equations regardless of the particulars of the optical system. In particular, assuming a homogeneous isotropic medium in the imaging domain, the coherent response $h(\mathbf{r})$ to an impulse at $\mathbf{r}'$ will satisfy the homogeneous Helmholtz equation in the image domain:

$$\nabla^2 h(\mathbf{r}) + k^2 h(\mathbf{r}) = 0 \qquad [9]$$

where $h(\mathbf{r})$ is the complex amplitude and $k = 2\pi/\lambda$, with λ the wavelength in the medium. If the field at a given origin of coordinates $z = 0$, $f(x, y)$, is known, $h(\mathbf{r})$ can be calculated using the angular spectrum representation as follows:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} F(f_x, f_y) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\,\mathrm{d}f_y \qquad [10]$$

where $f_z = \sqrt{\lambda^{-2} - (f_x^2 + f_y^2)}$ and $F(f_x, f_y) = \mathrm{FT}\{f(x, y)\}$; FT represents a 2D Fourier transformation and f_x, f_y are the spatial frequencies along the x and y axes respectively.

Equation [10] can be written as a 3D integral as follows:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} F(f_x, f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\,\mathrm{d}f_y\,\mathrm{d}f_z \qquad [11]$$

where δ represents the dirac impulse. Next we represent the 3D impulse response $h(x, y, z)$ as a 3D Fourier decomposition:

$$h(x, y, z) = \int\int_{-\infty}^{\infty} H(f_x, f_y, f_z) \times \exp[i2\pi(f_x x + f_y y + f_z z)]\mathrm{d}f_x\,\mathrm{d}f_y\,\mathrm{d}f_z \qquad [12]$$

where $H(f_x, f_y, f_z) = \mathrm{FT}^{3\mathrm{D}}\{h(x, y, z)\}$. Comparing eqns [11] and [12] we can infer that:

$$H(f_x, f_y, f_z) = F(f_x, f_y)\cdot\delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \qquad [13]$$

This relation implies that the domain of the Fourier transform of the impulse response is confined to a spherical shell of radius $1/\lambda$. Alternatively, $H(f_x, f_y, f_z)$ is zero everywhere, except from the surface of the sphere.

The previous derivation did not consider any limitation in the range of spatial frequencies allowed through the system. We now assume a finite aperture that will introduce an effective radial cutoff frequency $f_c < 1/\lambda$ and will further limit the spatial spectral domain of $h(x, y, z)$. This is depicted in Figure 1. Note that the transverse and longitudinal spatial frequencies are related, hence the longitudinal spatial frequencies are located within a bandpass of width $\Delta f_z = 1 - \sqrt{\lambda^{-2} - f_c^2}$.

Let us now consider the frequency content of possible incoherent PSFs. From eqn [5] we derive the 3D Fourier transform of the incoherent impulse response:

$$\begin{aligned} H_{\mathrm{Inc}}(f_x, f_y, f_z) &= \mathrm{FT}\left\{h_{\mathrm{Inc}}(x, y, z; x'y', z')\right\} \\ &= \mathrm{FT}\left\{|h_{\mathrm{Coh}}(x, y, z; x'y', z')|^2\right\} \\ &= H \otimes H(f_x, f_y, f_z) \end{aligned} \qquad [14]$$

where $\otimes$ represents a 3D correlation. The domain of H_{Inc} now becomes a volume of revolution as shown in Figure 2. Moreover, because h_{Inc} is real, H_{Inc} is conjugate symmetric, i.e., $H_{\mathrm{Inc}} = H_{\mathrm{Inc}}^*$. Now, the transverse cutoff frequency is $f_{\mathrm{c}}^{\mathrm{Inc}} = 2f_{\mathrm{c}}$ while the longitudinal cutoff frequency is $f_{\mathrm{cl}}^{\mathrm{Inc}} = 2\Delta f_z$.

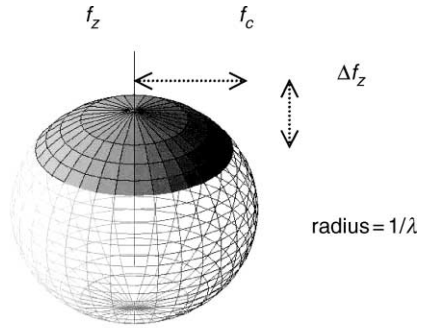

Figure 1 Three-dimensional spatial-frequency domain of the coherent point spread function.

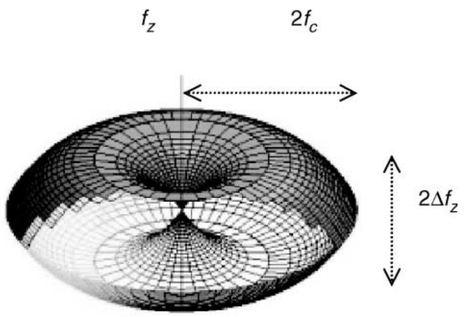

Figure 2 Three-dimensional spatial-frequency domain of the incoherent point spread function.

Diffraction Calculation of the 3D Point Spread Function

The 3D-image field created by a point source in the object space can be predicted by proper modeling of the optical system. In the absence of aberrations, the distortion of the PSF is originated in the diffraction created by the limiting aperture of the optical system. Diffraction is thus the ultimate limitation to the resolution of the optical system. Under the design conditions of the system, i.e., assuming all aberrations are corrected, the image of a point source is the diffraction pattern produced when a spherical wave converging to the (geometrical) image point is diffracted by the limiting aperture of the system. This is called a diffraction limited image.

There are many formulations to describe the 3D diffraction pattern obtained with a converging wave through an aperture. Here we review the Kirchhoff formulation. Let us consider the aperture **A** and the field at a point P_d originated from a converging spherical wave focused at point P_F (see Figure 3). We define the vectors **r** joining a generic point on the aperture P_0 and P_F, and **s** joining P_0 and P_d. The Kirchhoff formulation states that the complex amplitude $U(P_\mathrm{d}/P_\mathrm{F})$ is the integral over the aperture of the complex amplitude at each point propagated to P_d by a spherical wave $\exp(iks)/s$ modified by the directional factor $\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})$. $\hat{\mathbf{n}}$ is the unit vector normal to the aperture:

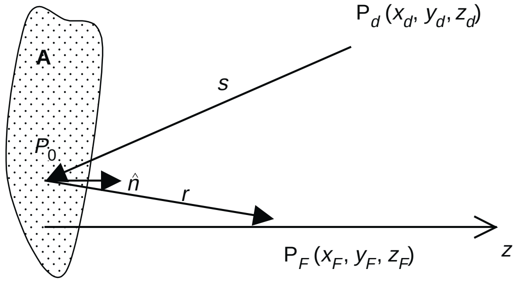

Figure 3 Diffraction of a spherical wave by a planar aperture.

$$U(P_\mathrm{d}/P_\mathrm{F}) = \frac{-iC}{2\lambda} \iint_\mathbf{A} \frac{\exp[-ik(r-s)]}{rs} \times [\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})]\mathrm{d}S \qquad [15]$$

where $r = \|\mathbf{r}\|, s = \|\mathbf{s}\|$, and C is the amplitude of the spherical wave.

The directional factor varies slowly and in many situations can be approximated by a constant equal to 2 for small image-field angles. The amplitude over the aperture is also slowly varying and can be approximated by a constant, i.e., $C/r \approx C_0$. Different formulations of the Kirchhoff integral differ in the estimates of the factor $1/s$ and the phase $k(r-s)$. For example, one such formulation leads to the following expression for the 3D diffraction pattern for a focal point on the axis of an optical system with circular aperture:

$$U(P_\mathrm{d}/P_\mathrm{F}) = \frac{-ika^2C_0}{z_d} \exp\left[ik\left(z_\mathrm{d} - z_\mathrm{F} + \frac{r_\mathrm{d}^2}{2z_\mathrm{d}}\right)\right] \times \int_0^1 J_0\left(\frac{ka\rho r_\mathrm{d}}{z_\mathrm{d}}\right) \exp\left[\frac{-ika^2\rho^2(z_\mathrm{d} - z_\mathrm{F})}{2z_\mathrm{d}z_\mathrm{F}}\right] \rho \mathrm{d}\rho \qquad [16]$$

where a is the radius of the aperture, r_d is the radial coordinate of P_d, and ρ is the radial coordinate of P_0.

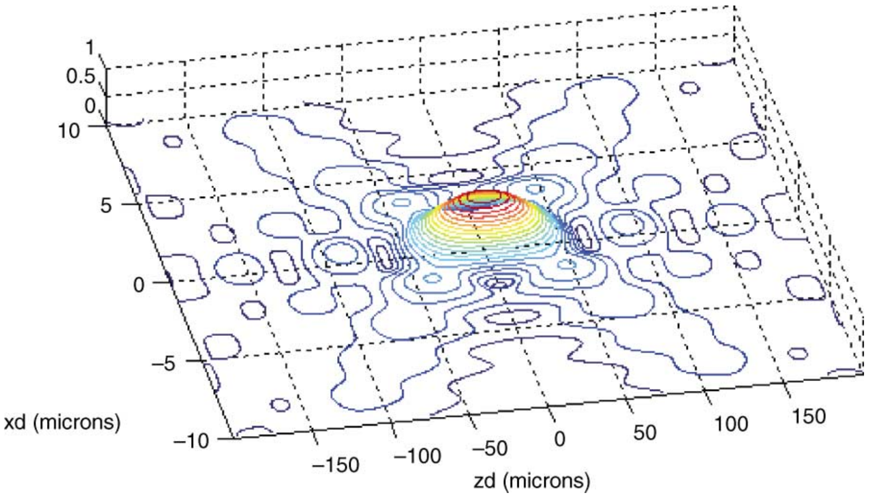

Figure 4 Lines of equal intensity in a meridional plane near the focus of a converging spherical wave diffracted at a circular aperture ($\lambda = 0.5\ \mu$, $a = 0.5$ cm, $z_\mathrm{F} = 4$ cm).

To illustrate the use of eqn [16], we calculated the intensity around the focus of a spherical wave diffracted through a circular aperture of radius $a = 0.5$ cm, focusing at distance $z_F = 4$ cm from the screen. We assumed the wavelength to be $\lambda = 0.5\ \mu$. Figure 4 depicts the magnitude of the complex amplitude in a meridional plane near the focus of the converging spherical wave.

Models that take into account off-axis focal points have also been developed. For the most accurate calculations, it is necessary to consider the full electromagnetic nature of the optical field.

Conclusions

Classical imaging theory can be applied to the 3D imaging transformation from object to image. The systems approach to the imaging problem implies that knowledge of the 3D coherent PSF suffices to predict the 3D response of the system under illumination of varying degrees of coherence. The PSF response can be predicted using well-known formulations of diffraction.

See also

Coherent Lightwave Systems. **Fourier Optics**.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. New York: Cambridge University Press.

Gibson SF and Lanni F (1989) Diffraction by a circular aperture as a model for three-dimensional optical microscopy. *Journal of the Optical Society of America A* 6: 1357–1367.

Goodman JW (1996) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Li Y and Wolf E (1984) Three-dimensional intensity distributions near the focus in systems of different Fresnel numbers. *Journal of the Optical Society of America A* 1: 801–808.

Piestun R and Shamir J (2002) Synthesis of 3D light fields and applications. *Proceedings of the IEEE* 90: 222–244.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Stamnes JJ (1986) *Waves in Focal Regions*. Bristol: Hilger.

Volume Holographic Imaging

G Barbastathis, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Published by Elsevier Ltd.

Volume holographic imaging (VHI) refers to a class of imaging techniques which use a volume hologram in at least one location within the optical path. The volume hologram acts as a ‘smart lens,’ which processes the optical field to extract spatial information in three dimensions (lateral as well as longitudinal) and spectral information.

Figure 1 is a generic VHI system. The object is either illuminated by a light source (e.g., sunlight or a pump laser) as shown in the figure, or it may be self-luminous. Light scattered or emitted by the object is first transformed by an objective lens and then illuminates the volume hologram. The role of the objective is to form an intermediate image which serves as input to the volume hologram. The volume hologram itself is modeled as a three-dimensional (3D) modulation $\Delta\varepsilon(\mathbf{r})$ of the dielectric index within a finite region of space. The light entering the hologram is diffracted by $\Delta\varepsilon(\mathbf{r})$ with efficiency η, defined as

$$\eta = \frac{\text{Power diffracted by the volume hologram}}{\text{Power incident to the volume hologram}} \qquad [1]$$

We assume that diffraction occurs in the Bragg regime. The diffracted field is Fourier-transformed by the collector lens, and the result is sampled and measured by an intensity detector array (such as a CCD or CMOS camera).

Intuitively, we expect that a fraction of the illumination incident upon the volume hologram is Bragg-matched and is diffracted towards the Fourier-transforming lens. The remainder of the incident illumination is Bragg-mismatched, and as result is transmitted through the volume hologram un-diffracted. Therefore, the volume hologram acts as a filter which admits the Bragg-matched portion of the object and rejects the rest. When appropriately designed, this ‘Bragg imaging filter’ can exhibit very rich behavior, spanning the three spatial dimensions and the spectral dimension of the object.

To keep the discussion simple, we consider the specific case of a transmission geometry volume

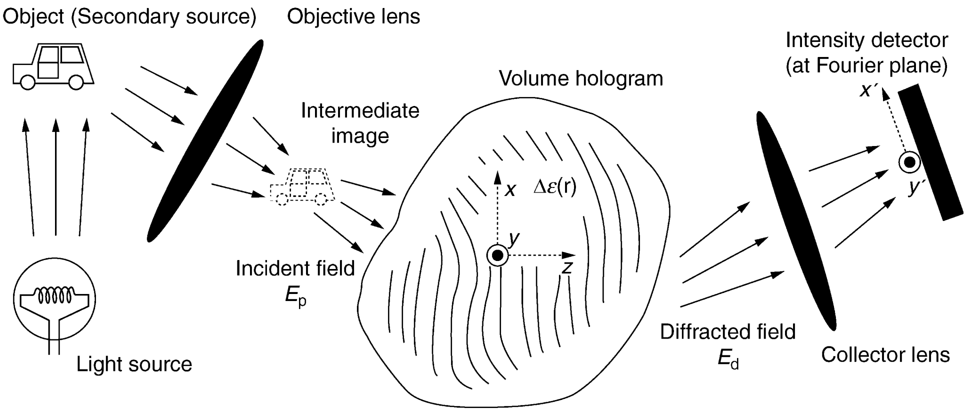

Figure 1 Volume holographic imaging (VHI) system.

hologram, described in Figure 2. The volume hologram is created by interfering a spherical wave and a planewave, as shown in Figure 2a. The spherical wave originates at the coordinate origin. The planewave is off-axis and its wavevector lies on the xz plane. As in most common holographic systems, the two beams are assumed to be at the same wavelength λ, and mutually coherent. The volume hologram results from exposure of a photosensitive material to the interference of these two beams.

First, assume that the object is a simple point source. The intermediate image is also approximately a point source, that we refer to as 'probe,' located somewhere in the vicinity of the reference point source. Assuming the wavelength of the probe source is the same as that of the reference and signal beams, volume diffraction theory shows that:

(i) If the probe point source is displaced in the y direction relative to the reference point source, the image formed by the volume hologram is also displaced by a proportional amount. Most common imaging systems would be expected to operate this way.
(ii) If the probe point source is displaced in the x direction relative to the reference point source, the image disappears (i.e., the detector plane remains dark).
(iii) If the probe point source is displaced in the z direction relative to the reference point source, a defocused and faint image is formed on the detector plane. 'Faint' here means that the fraction of energy of the defocused probe transmitted to the detector plane is much smaller than the fraction that would have been transmitted if the probe had been at the origin.

Now consider an extended, monochromatic, spatially incoherent object and intermediate image, as in Figure 2f. According to the above description, the volume hologram acts as a 'Bragg slit' in this case: because of Bragg selectivity, the volume hologram transmits light originating from the vicinity of the y axis, and rejects light originating anywhere else. For the same reason, the volume hologram affords depth selectivity (like the pinhole of a confocal microscope). The width of the slit is determined by the recording geometry, and the thickness of the volume hologram. For example, in the transmission recording geometry of Figure 2a, where the reference beam originates a distance z_0 away from the hologram, the planewave propagates at angle θ with respect to the optical axis z (assuming $\theta \ll 1$ radian), and the hologram thickness is L (assuming $L \ll z_0$), the width of the slit is found to be

$$\Delta x \approx \frac{\lambda z_0}{L\theta} \qquad [2]$$

The imaging function becomes richer if the object is polychromatic. In addition to its Bragg slit function, the volume hologram exhibits then dispersive behavior, like all diffractive elements. In this particular case, dispersion causes the hologram to image simultaneously multiple Bragg slits, each at a different color and parallel to the original slit at wavelength λ, but displaced along the z axis. Light from all these slits finds itself in focus at the detector plane, thus forming a 'rainbow image' of an entire slice through the object, as shown in Figure 2g.

To further exploit the capabilities of volume holograms, we recall that in general it is possible to 'multiplex' (superimpose) several volume gratings

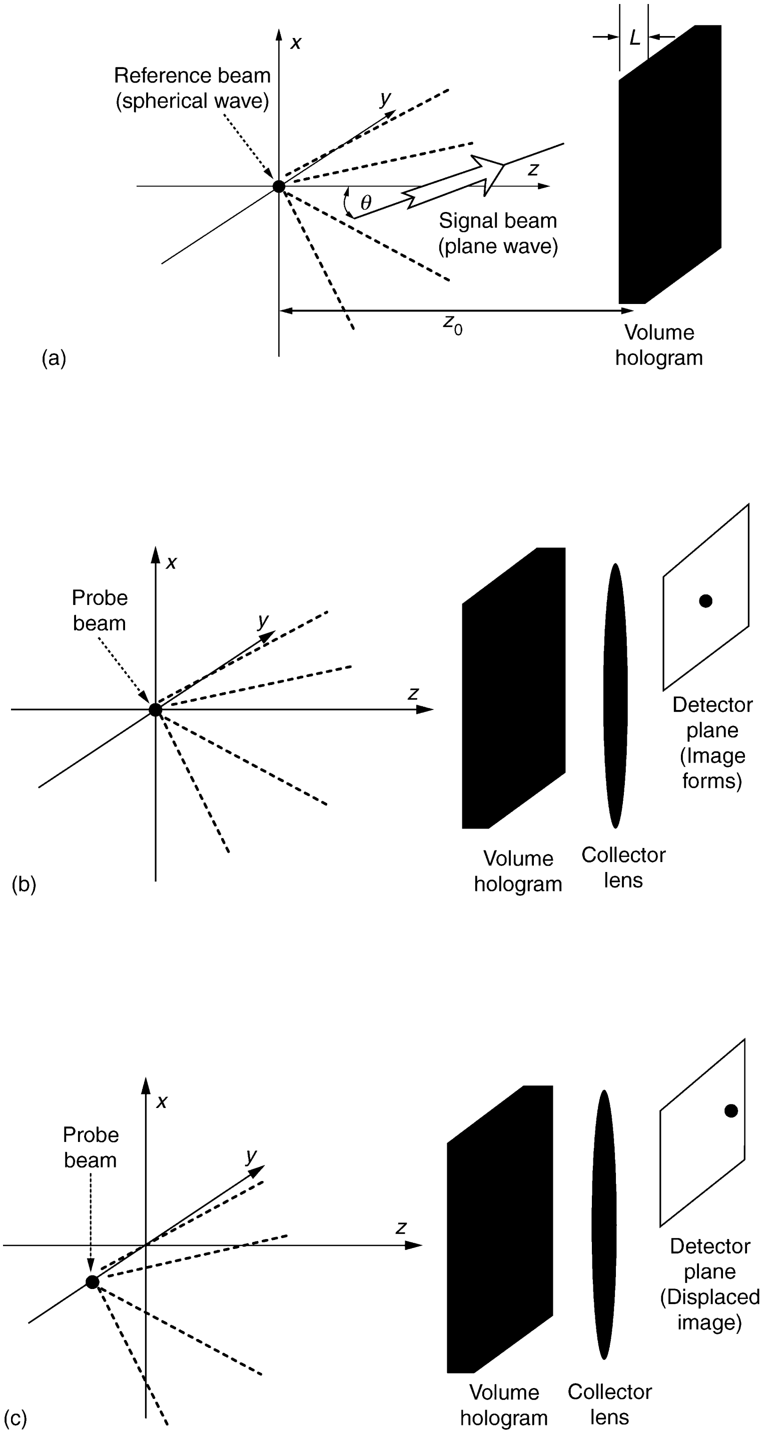
x
Reference beam (spherical wave)
y
z
θ
Signal beam (plane wave)
L
z_0
Volume hologram
(a)
x
Probe beam
y
z
Volume hologram
Collector lens
Detector plane (Image forms)
(b)
x
Probe beam
y
z
Volume hologram
Collector lens
Detector plane (Displaced image)
(c)

Figure 2 (a) Recoding of a transmission-geometry volume hologram with a spherical wave and an off-axis plane wave with its wave-vector on the *xz* plane. (b) Imaging of a probe point source that replicates the location and wavelength of the reference point source using the volume hologram recorded in part (a). (c) Imaging of a probe point source at the same wavelength but displaced in the *y* direction relative to the reference point source. (d) Imaging of a probe point source at the same wavelength but displaced in the *x* direction relative to the reference point source. (e) Imaging of a probe point source at the same wavelength but displaced in the *z* direction relative to the reference point source. (f) 'Bragg slitting:' Imaging of an extended monochromatic, spatially incoherent object using a volume hologram recorded as in (a). (g) Joining Bragg slits from different colors to form rainbow slices: Imaging of an extended polychromatic object using a volume hologram recorded as in (a). (h) Multiplex imaging of several slices using a volume hologram formed by multiple exposures.

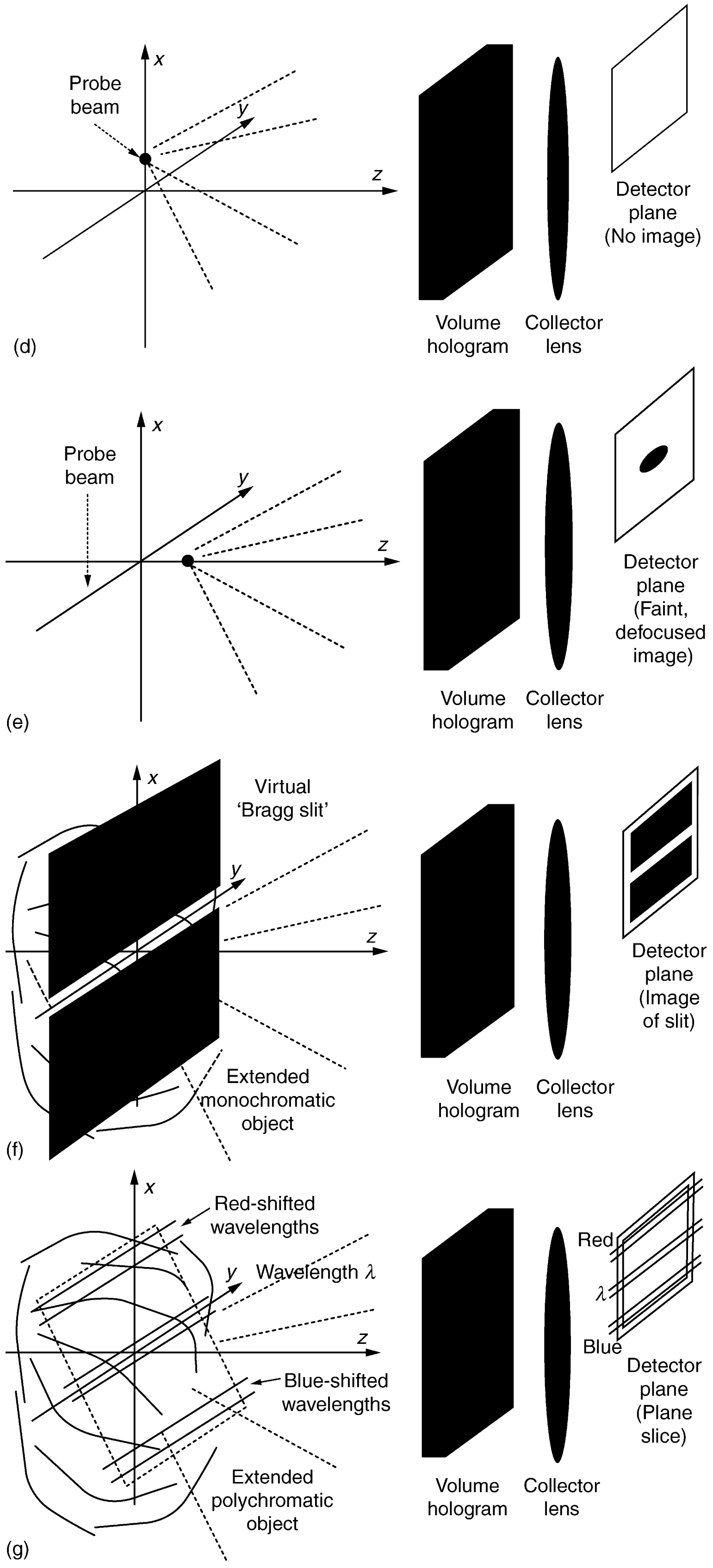
(d)
x
y
z
Probe beam
Volume hologram
Collector lens
Detector plane (No image)
(e)
x
y
z
Probe beam
Volume hologram
Collector lens
Detector plane (Faint, defocused image)
(f)
x
y
z
Virtual 'Bragg slit'
Extended monochromatic object
Volume hologram
Collector lens
Detector plane (Image of slit)
(g)
x
y
z
Red-shifted wavelengths
Wavelength λ
Blue-shifted wavelengths
Extended polychromatic object
Volume hologram
Collector lens
Red
λ
Blue
Detector plane (Plane slice)

Figure 2 *Continued.*

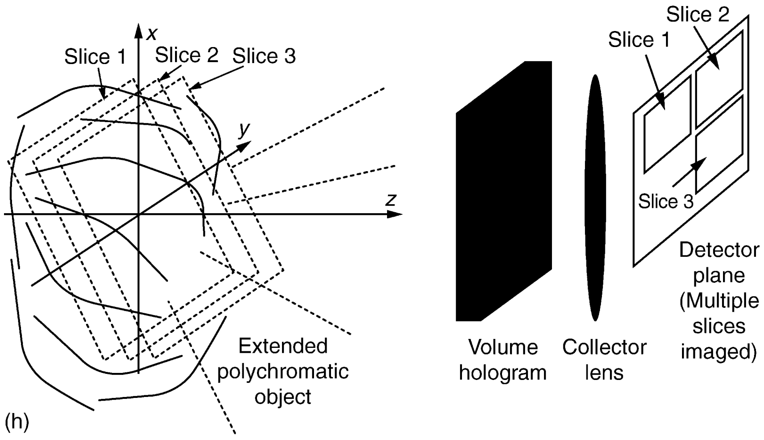

Figure 2 *Continued.*

within the same volume by successive exposures. In the imaging context, suppose that we multiplex several gratings similar to the grating described in Figure 2a but with spherical reference waves originating at different locations and plane signal waves at different orientations. When the multiplexed volume hologram is illuminated by an extended polychromatic source, each grating forms a separate image of a rainbow slice, as described earlier. By spacing appropriately the angles of propagation of the plane signal waves, we can ensure that the rainbow images are formed on nonoverlapping areas on the detector plane, as shown in Figure 2h. This device is now performing true four-dimensional (4D) imaging: it is separating the spatial and spectral components of the object illumination so that they can be measured independently by the detector array. Assuming the photon count is sufficiently high and the number of detector pixels is sufficient, this 'spatio-spectral slicing' operation can be performed in real time, without need for mechanical scanning.

The complete theory of VHI in the spectral and spatial domains is given in the Further Reading section below. Experimental demonstrations of VHI have been performed in the context of a confocal microscope where the volume hologram performs the function of a 'Bragg pinhole' as well as a real-time 4D imager. The limited diffraction efficiency η of volume holograms poses a major concern for VHI systems. Holograms, however, are known to act as matched filters. It has been shown that the matched filtering nature of volume holograms as imaging elements is superior to other filtering elements (e.g., pinholes or slits) in an information-theoretic sense. Efforts are currently underway to strengthen this property by design of volume holographic imaging elements which are more elaborate than described here.

See also

Holography, Techniques: Overview. **Imaging:** Three-Dimensional Field Transformations.

Further Reading

Barbastathis G and Brady DJ (1999) Multidimensional tomographic imaging using volume holography. *Proceedings of the IEEE* 87(12): 2098–2120.

Barbastathis G and Sinha A (2001) Information content of volume holographic images. *Trends in Biotechnology* 19(10): 383–392.

Barbastathis G, Balberg M and Brady DJ (1999) Confocal microscopy with a volume holographic filter. *Optics Letters* 24(12): 811–813.

Barbastathis G, Sinha A and Neifeld MA (2001) Information-theoretic treatment of axial imaging. In: *OSA Topical Meeting on Integrated Computational Imaging Systems*, pp. 18–20. Albuquerque, NM.

Liu W, Psaltis D and Barbastathis G (2004) Real time spectral imaging in three spatial dimensions. *Optics Letters* in press.

van-der Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10(2): 139–145.

Wavefront Sensors and Control (Imaging Through Turbulence)

C L Matson, Air Force Research Laboratory,
Kirtland, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In many situations, distant objects are imaged using optical or near-infrared imaging systems. Examples include terrestrial surveillance from space, tactical surveillance from airborne imaging systems, and ground-based astronomical imaging. Ideally, the resolution in these images is limited only by the wavelength of light λ used to create the images divided by the imaging system's effective diameter D. The ratio λ/D is called the diffraction limit of the imaging system. Unfortunately, when the imaging system is located in the Earth's atmosphere, it is often atmospheric turbulence fluctuations that limit the level of detail in the long-exposure images, not the diffraction limit of the telescope. For example, the Keck telescopes, located on top of Mauna Kea in Hawaii, are 10 m diameter telescopes that have a diffraction-limited angular resolution of 0.2 μrad for an imaging wavelength of 2 μm. However, atmospheric turbulence limits the achievable resolution in long-exposure images to the order of 1–5 μrad at this wavelength.

The important discovery by Anton Labeyrie in 1970 that short-exposure images of objects retain meaningful diffraction-limited information about the object, where 'short' means that the exposure time is less than the atmospheric turbulence coherence time, has revolutionized imaging through turbulence. Atmospheric coherence times depend upon a number of factors, including wind speeds, wind altitudes, and observation wavelength. At astronomical observatories for visible observation wavelengths, a rule of thumb is that atmospheric turbulence stays essentially constant for exposure times of one to ten milliseconds under average clear-sky weather conditions. A variety of techniques that exploit this discovery have been developed. These techniques can be grouped into three different categories: techniques that remove blurring from the detected intensity images (post-detection compensation), techniques that modify the optical field prior to detecting the light (pre-detection compensation), and hybrid techniques that use both post-detection and pre-detection compensation. In this article, a variety of these techniques are presented that are representative of the myriad of methods that have been developed during the past decades. To simplify the discussion, it is assumed that the blurring due to atmospheric turbulence is the same at every point in the image (isoplanaticity), that the images are formed using incoherent light as emitted by the sun and stars, and that a single filled aperture is used to create the images. Interferometric techniques that use multiple apertures to estimate the mutual coherence function of the optical field and calculate an image from it, are discussed in a separate article (*see* **Imaging:** Interferometric Imaging).

The emphasis in this article is on identifying and removing atmospheric blurring in images. As part of this process, regularization techniques must be included in the image reconstruction algorithms in order to obtain the best possible resolution for an acceptable level of noise. Although regularization techniques are not discussed in this article, information can be found in the Further Reading section at the end of this article (*see* **Imaging:** Inverse Problems and Computational Imaging).

Imaging and Atmospheric Turbulence

In Figure 1, a conceptual diagram illustrates the impact that atmospheric turbulence has upon resolution in classical imaging systems. A distant star is being imaged by a telescope/detector system. In the absence of atmospheric turbulence, the field emitted by the star has a planar wavefront over the extent of the lens since the star is a long distance from the imaging system and thus unresolved by the telescope. When imaged by the telescope, the diffraction pattern of the telescope is obtained, resulting in a diffraction-limited image (Figure 2). However, in the presence of atmospheric turbulence near the telescope, the planar wavefront is randomly distorted by the turbulence, resulting in a corrugated wavefront with a spatial correlation scale given by the Fried parameter r_o. For visible wavelengths, at good astronomical observing sites, r_o ranges from 5 cm to 20 cm. When this corrugated wavefront is converted to an image by the telescope, a distorted image is obtained (Figure 3). Because atmospheric turbulence is time-varying, the distortions in the image also change over time. If the imaging shutter on the telescope remains open for many occurrences of atmospheric turbulence, the resulting image will be an average of all the individually distorted images, resulting in what is known as the seeing disk (Figure 4). It is the seeing disk that typically limits spatial resolution in long-exposure images, not the diffraction pattern of the

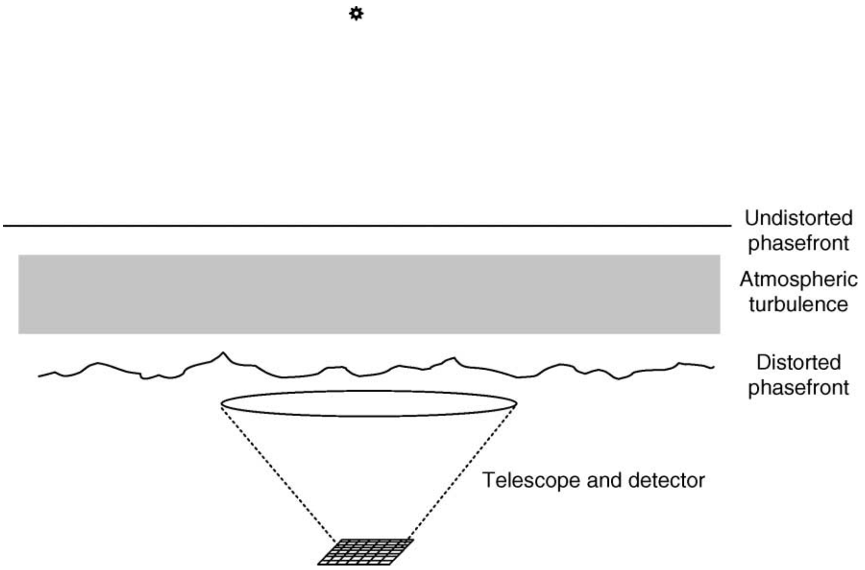

Figure 1 Conceptual diagram of imaging through atmospheric turbulence.

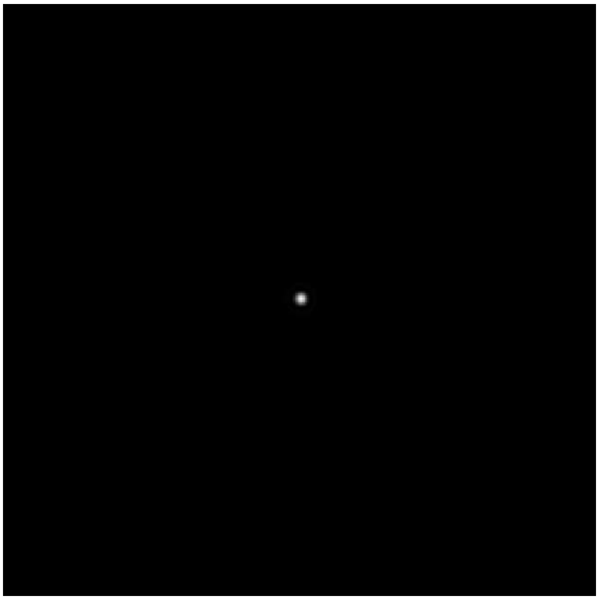

Figure 2 Computer-simulated diffraction-limited image of a point source.

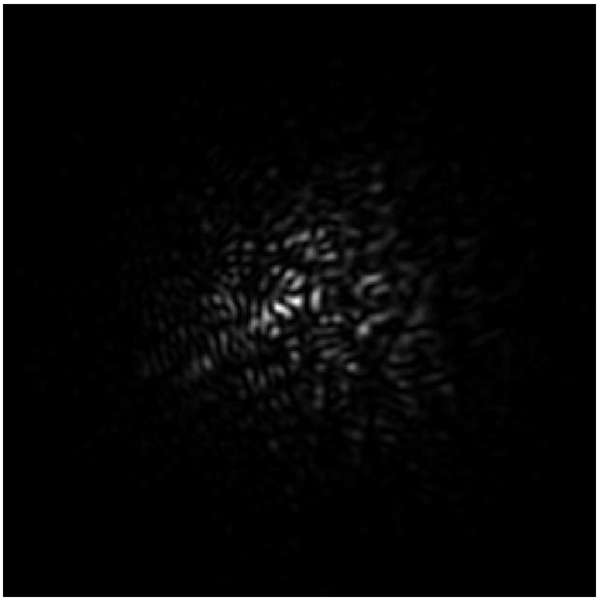

Figure 3 Computer-simulated short-exposure image of a point source through atmospheric turbulence.

telescope. As a result, the effective spatial resolution in long-exposure images in the presence of atmospheric turbulence is given by λ/r_o, not λ/D.

Atmospheric turbulence distorts a field's wavefront by delaying it in a way that is spatially and temporally random. The origin of these delays is the random variation in the index of refraction of the atmosphere as a result of turbulent air motion brought about by temperature variations. A Fourier optics imaging model is useful in characterizing the effects of atmospheric turbulence on spatial resolution in images. Let $i(\vec{x})$ be an image of an object $o(\vec{x})$, where $\vec{x}$ is a two-dimensional spatial vector, and let $I(\vec{f})$ and $O(\vec{f})$ be their spatial Fourier transforms, where $\vec{f}$ is a two-dimensional spatial frequency vector. In the absence of atmospheric turbulence, the diffraction-limited Fourier transform $I(\vec{f})$ is given by

$$I_d(\vec{f}) = O(\vec{f})H_d(\vec{f}) \qquad [1]$$

where the subscript 'd' denotes the diffraction-limited case and where $H_d(\vec{f})$ is the diffraction-limited

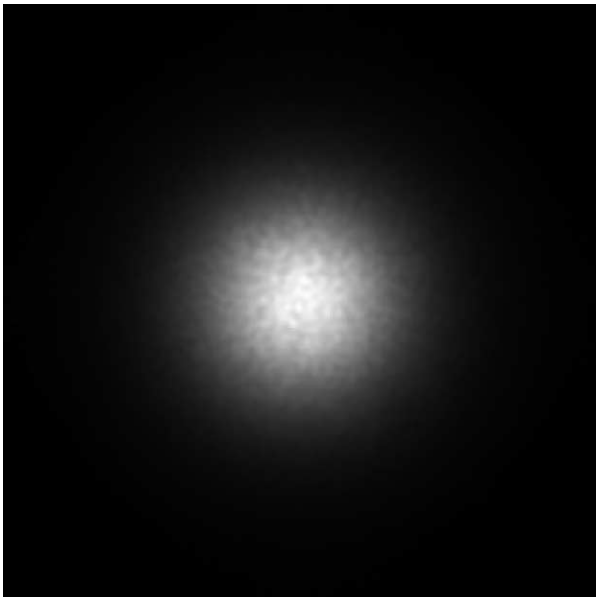

Figure 4 Computer-simulated long-exposure image of a point source through atmospheric turbulence.

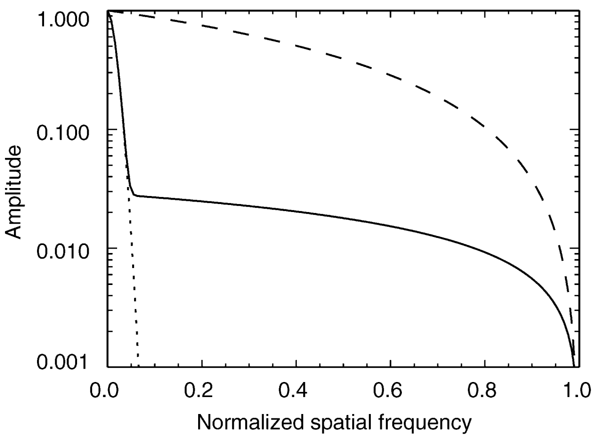

Figure 5 Plots of the Fourier amplitudes of the diffraction-limited transfer function (dashed line), Fourier amplitudes of the long-exposure transfer function (dotted line), and the square root of the average short exposure energy spectrum. The spatial frequency axis is normalized to one at the diffraction limit of the telescope.

telescope transfer function. In the presence of atmospheric turbulence, the Fourier transform of a long exposure image $i_{\mathrm{le}}(\vec{x})$ is given by

$$I_{\mathrm{le}}(\vec{f}) = O(\vec{f})H_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) = I_{\mathrm{d}}(\vec{f})H_{\mathrm{le}}(\vec{f}) \qquad [2]$$

where the subscript 'le' denotes long exposure. If the turbulence satisfies Kolmogorov statistics, as is typically assumed, then:

$$H_{\mathrm{le}}(\vec{f}) = \exp\left[-3.44\left(\frac{\lambda d|\vec{f}|}{r_{\mathrm{o}}}\right)^{5/3}\right] \qquad [3]$$

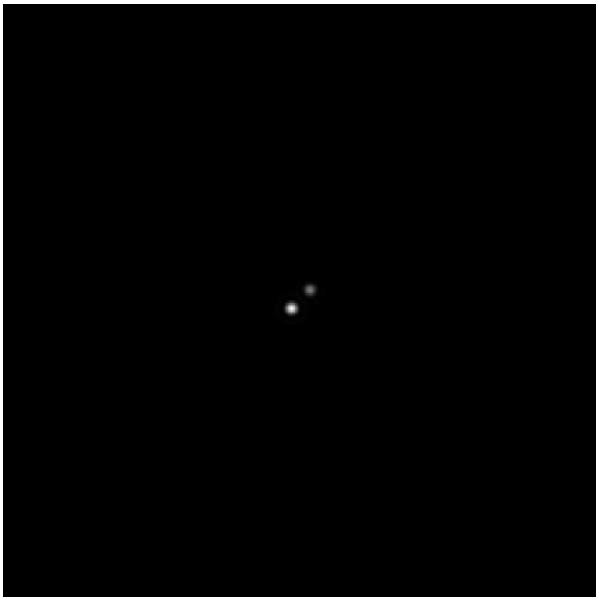

Figure 6 Computer-simulated diffraction-limited image of a binary star.

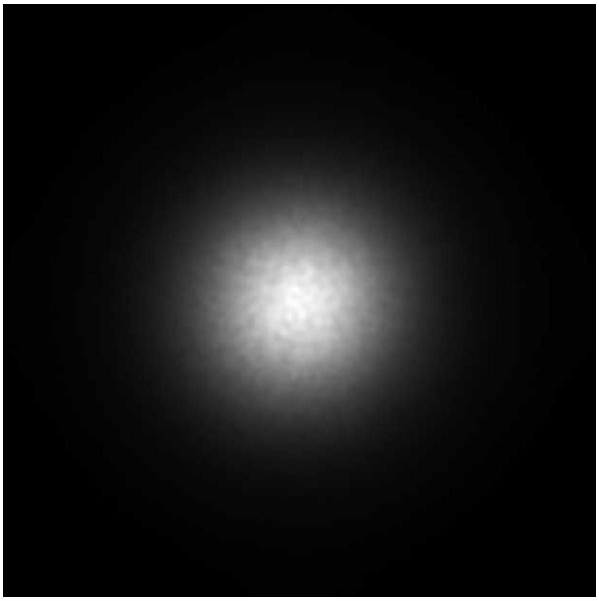

Figure 7 Computer-simulated long-exposure image of a binary star.

where d is the distance between the imaging system's exit pupil and detector plane. As can be seen from eqn [2], the effect of atmospheric turbulence on a long-exposure image is a multiplication of the diffraction-limited image's Fourier transform by a transfer function due to atmospheric turbulence. In Figure 5, an amplitude plot of a diffraction-limited transfer function is displayed along with an amplitude plot of a long-exposure transfer function for a 2.3 m diameter telescope, an imaging wavelength of 0.5 μm, and an r_o value of 10 cm. Notice that the long exposure transfer function significantly reduces the amount of high spatial frequency information in an image. Diffraction-limited and long-exposure turbulence-limited images of a binary star are shown in Figures 6 and 7, illustrating the significant loss of spatial resolution brought on by atmospheric turbulence.

Post-Detection Compensation

Speckle Imaging

Labeyrie's 1970 discovery was motivated by his analysis of short-exposure star images. In these images (see Figure 3), a number of speckles are present, where the individual speckle sizes are roughly the size of the diffraction-limited spot of the telescope. From this, he realized that high spatial frequency information is retained in short-exposure images, albeit distorted. He then showed that the energy spectra $E_I(\vec{f})$ of the individual short-exposure images can be averaged to obtain an estimate of the average energy spectrum that retains high signal-to-noise ratio (SNR) information at high spatial frequencies. The square root of the average energy spectrum (the amplitude spectrum) of a star, calculated using Labeyrie's method (also known as speckle interferometry), is shown in Figure 5, along with the diffraction-limited and long-exposure amplitude spectra. Notice that meaningful high spatial frequency information is retained in the average amplitude spectrum out to the diffraction limit of the telescope. Because the atmospherically corrupted amplitude spectrum estimate is significantly lower than the diffraction-limited amplitude, a number of short exposure images must be used to build up the SNRs at the higher spatial frequencies.

To reconstruct an image, both the Fourier amplitudes and Fourier phases of the image must be known. In fact, it is well known that the Fourier phase of an image carries most of the information in the image. Because only the Fourier amplitudes are available using Labeyrie's method, other techniques were subsequently developed to calculate the Fourier phase spectrum from a sequence of short-exposure images. Techniques that combine amplitude spectrum estimation using Labeyrie's method along with phase spectrum estimation are called speckle imaging techniques. The two most widely used phase spectrum estimation methods are the Knox–Thompson (or cross spectrum) and bispectrum algorithms. Both of these algorithms calculate Fourier-domain quantities that retain high spatial frequency information about the Fourier phase spectrum when averaged over multiple short-exposure images. The cross spectrum and bispectrum techniques calculate the average cross spectrum $C(\vec{f}, \Delta\vec{f})$ and bispectrum $B(\vec{f}_1, \vec{f}_2)$, respectively, that are given by

$$C(\vec{f}, \Delta\vec{f}) = I(\vec{f})I^*(\vec{f} + \Delta\vec{f}) \qquad [4]$$

$$B(\vec{f}_1, \vec{f}_2) = I(\vec{f}_1)I(\vec{f}_2)I^*(\vec{f}_1 + \vec{f}_2) \qquad [5]$$

where the superscript $*$ denotes complex conjugate. Although the cross spectrum is defined for all values of $\vec{f}$ and $\Delta\vec{f}$, only for values of $\Delta\vec{f}$ or $\vec{f}$ less than r_o/λ are the cross spectrum elements obtained with sufficiently high SNRs to be useful in the reconstruction process. The equivalent constraint on the bispectrum elements is that either $\vec{f}_1, \vec{f}_2$, or $\vec{f}_1 + \vec{f}_2$ must be less than r_o/λ. Once the average cross spectrum or bispectrum is calculated from a series of short-exposure images, the phase spectrum $\phi_I(\vec{f})$ of the underlying image can be calculated from their phases using one of a number of algorithms based upon the following equations:

$$\phi_I(\vec{f} + \Delta\vec{f}) = \phi_I(\vec{f}) - \phi_C(\vec{f} + \Delta\vec{f}) \qquad [6]$$

for the cross spectrum and:

$$\phi_I(\vec{f}_1 + \vec{f}_2) = \phi_I(\vec{f}_1) + \phi_I(\vec{f}_2) - \phi_B(\vec{f}_1 + \vec{f}_2) \qquad [7]$$

for the bispectrum. Because the phases are only known for units of 2π, phasor versions of eqns [6] and [7] must be used. For the cross spectrum technique, each short-exposure image must be centroided prior to calculating its cross spectrum and adding it to the total cross spectrum. The bispectrum technique is insensitive to tilt, so no centroiding is necessary; however, the bispectrum technique requires setting two image phase spectrum values to arbitrary values to initialize the phase reconstruction process. Typically, the image phase spectrum values at $\vec{f} = (0, 1)$ and $\vec{f} = (1, 0)$ are set equal to zero, which centroids the reconstructed image but does not affect its morphology.

To obtain an estimated Fourier transform $I_e(\vec{f})$ of the image from the energy and phase spectrum estimates, the square root of the average energy spectrum is combined with the average phase spectrum calculated from either the bispectrum or cross spectrum. The result is given by

$$I_e(\vec{f}) = [E_I(\vec{f})]^{1/2} \exp[j\phi_I(\vec{f})] \qquad [8]$$

where $j = \sqrt{-1}$. The result in eqn [8] includes atmospheric attenuation of the Fourier amplitudes, so it must be divided by the amplitude spectrum of the atmosphere to obtain the desired object spectrum estimate. The atmospheric amplitude spectrum estimate is obtained by using Labeyrie's technique on a sequence of short-exposure images of a nearby unresolved star. No phase spectrum correction is necessary because it has been shown theoretically and verified experimentally that atmospheric turbulence has a negligible effect upon $\phi_I(\vec{f})$.

Blind Deconvolution

Speckle imaging techniques are effective as well as easy and fast to implement. However, it is necessary to obtain a separate estimate of the atmospheric energy spectrum using an unresolved star in order to reconstruct a high-resolution image. Usually, the star is not co-located with the object and the star images are not collected simultaneously with the images of the object. As a result, the atmospheric energy spectrum estimate may be inaccurate. At the very least, collecting the additional star data expends valuable observation time. In addition, there is often meaningful prior knowledge about the object and the data, such as the fact that measured intensities are positive (positivity), that many astronomical objects are imaged against a black background (support), and that the measured data are band-limited, among others. Speckle imaging algorithms do not exploit this additional information. As a result, a number of techniques have been developed to jointly estimate the atmospheric blurring effects as well as an estimate of the object being imaged using only the short-exposure images of the object and available prior knowledge. These techniques are called blind deconvolution techniques. Two of the most common techniques are known as multi-frame blind deconvolution (MFBD) and projection-based blind deconvolution (PBBD).

The MFBD algorithm is based upon maximizing a figure of merit to determine the best object and atmospheric short-exposure point spread function (PSF) estimates given the data and available prior knowledge. This figure of merit is created using a maximum likelihood problem formulation. Conceptually, this technique produces estimates of the object and the atmospheric turbulence short-exposure PSFs that most likely generated the measured short-exposure images. Mathematically, the MFBD algorithm obtains these estimates by maximizing the conditional probability density function (pdf) $p[i_m(\vec{x})|o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M]$, where $i_m(\vec{x})$ is the mth of M short exposure images and is given by

$$i_m(\vec{x}) = h_m(\vec{x}) * o(\vec{x}) + n_m(\vec{x}) \qquad [9]$$

where $h_m(\vec{x})$ and $n_m(\vec{x})$ are the atmospheric/system blur and detector noise functions for the mth image, respectively, and the asterisk denotes convolution. The noise is assumed to be zero mean with variance $\sigma^2(\vec{x})$. The conditional pdf $p[i_m(\vec{x})|o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M]$ is the pdf of the noise with a mean equal to $h_m(\vec{x}) * o(\vec{x})$ and, as a function of $o(\vec{x})$ and $h_m(\vec{x})$, it is known as a likelihood function. For spatially independent Gaussian noise, as is the case for camera read noise, the likelihood function is given by

$$L_G[o(\vec{x}); h_m(\vec{x}),\ m = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{1}{\sqrt{2\pi\sigma^2(\vec{x})}} \times \exp\left\{ -\left[i(\vec{x}) - (o * h_m)(\vec{x})\right]^2 / 2\sigma^2(\vec{x}) \right\} \qquad [10]$$

For Poisson noise, the likelihood function is given by

$$L_P[o(\vec{x}); h_m(\vec{x}), \mathrm{m} = 1, \ldots, M] = \prod_m \prod_{\vec{x}} \frac{\left[(o * h_m)(\vec{x})\right]^{i(\vec{x})} \exp\left\{ -\left[(o * h_m)(\vec{x})\right]\right\}}{i(\vec{x})!} \qquad [11]$$

In practice, logarithms of eqns [10] and [11] are taken prior to searching for the estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that maximize the likelihood function.

In general, there are an infinite number of estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that reproduce the original dataset when convolved together. However, when additional prior knowledge is included to constrain the solution space, such as positivity, support, and the knowledge that $h_m(\vec{x})$ is band-limited, along with regularization, the MFBD algorithm almost always converges to the correct estimates of $o(\vec{x})$ and $h_m(\vec{x})$.

The PBBD technique also uses prior knowledge constraints to jointly estimate $o(\vec{x})$ and $h_m(\vec{x})$. However, instead of creating and maximizing a likelihood function, the PBBD technique uses the method of convex projections to find estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that are consistent with the measured short-exposure images and all the prior knowledge constraints. The data consistency constraint and the prior knowledge constraints are mathematically formulated as convex sets and the PBBD technique repeatedly takes the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ and projects them onto the convex constraint sets. As long as all the sets are truly convex, the algorithm is guaranteed to converge to a solution that is in the intersection of all the sets as long as the algorithm doesn't stagnate at erroneous solutions called traps. These traps are the projection equivalent of local minima in cost function minimization algorithms.

A key component of this algorithm is the ability to formulate the measured short-exposure images and prior knowledge constraints into convex sets. As an example, consider the convex set $C_{\text{pos}}[u(\vec{x})]$ that corresponds to the positivity constraint. This set is given by

$$C_{\text{pos}}[u(\vec{x})] \triangleq \{u(\vec{x}) \text{ such that } u(\vec{x}) \geq 0\} \qquad [12]$$

It is straightforward to show that $C_{\text{pos}}[u(\vec{x})]$ is a convex set. Other constraints can be converted into convex sets in a similar manner. Although all constraints can be formulated into sets, most but not all of these sets are convex.

The projection process is straightforward to implement. For example, the projections of the current estimates of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto $C_{\text{pos}}[u(\vec{x})]$ are carried out by zeroing out the negative values in these estimates. Similarly, the projections of $o(\vec{x})$ and $h_{\text{m}}(\vec{x})$ onto the convex sets that correspond to their support constraints are carried out by zeroing out all their values outside of their supports. In practice, some versions of PBBD algorithms have been formulated to use minimization routines as part of the projection process in order to minimize the possibility of stagnation.

Deconvolution from Wavefront Sensing

A key distinguishing feature of post-detection compensation algorithms is the type of additional data required to estimate and remove atmospheric blurring in the measured short-exposure images. Blind deconvolution techniques need no additional data. Speckle imaging techniques estimate average atmospheric blurring quantities and need an estimate of the average atmospheric energy spectrum. Because average quantities are calculated when using speckle imaging techniques, only an estimate of the average atmospheric energy spectrum is needed, not the specific energy spectra that corrupted the measured short-exposure images. As a result, the additional data need not be collected simultaneously with the short-exposure images. The technique described in this section, deconvolution from wavefront sensing (DWFS), seeks to deconvolve the atmospheric blurring in each short-exposure image by estimating the specific atmospheric field phase perturbations that blur each short-exposure image. Thus, the additional data must be collected simultaneously with the short-exposure images. The atmospheric field phase perturbations are calculated from data obtained with a wavefront sensor using a different region of the optical spectrum than used for imaging so that no light is taken from the imaging sensor. Most commonly, the wavefront sensor data consists of the first derivatives of the phasefront (phase differences) as obtained with Shack–Hartmann or shearing interferometer wavefront sensors, but curvature sensing wavefront sensors are also used that produce data that are proportional to the second derivatives of the phasefront.

The original algorithm to implement the DWFS technique produced an estimate $\hat{O}(\vec{f})$ of the Fourier transform $O(\vec{f})$ of the object being imaged using the following estimator:

$$\hat{O}(\vec{f}) = \frac{\sum_m I_{\text{m}}(\vec{f})\tilde{H}^*_{\text{m}}(\vec{f})}{\sum_m |\tilde{H}_{\text{m}}(\vec{f})|^2} \qquad [13]$$

where $\tilde{H}_{\text{m}}(\vec{f})$ is the overall system transfer function (including the telescope and atmospheric effects) for the *m*th short-exposure image estimated from the wavefront sensor data. To create $\tilde{H}_{\text{m}}(\vec{f})$, the atmospheric field phase perturbations corrupting the *m*th image must be calculated from the wavefront sensor measurements. This topic is discussed in more detail in **Imaging**: Adaptive Optics. Then $\tilde{H}_{\text{m}}(\vec{f})$ is obtained using standard Fourier optics techniques. Although eqn [13] is easily implemented given an existing system that includes a wavefront sensor, it suffers from the fact that it is a biased estimator. An unbiased version of eqn [13] is given by

$$\hat{O}(\vec{f}) = \frac{\sum_m I_{\text{m}}(\vec{f})\tilde{H}^*_{\text{m}}(\vec{f})}{\sum_n H_{\text{n,ref}}(\vec{f})\tilde{H}^*_{\text{n,ref}}(\vec{f})} \qquad [14]$$

where $\tilde{H}_{\text{n,ref}}(\vec{f})$ and $H_{\text{n,ref}}(\vec{f})$ are estimates of the overall system transfer function obtained from data collected on a reference star using wavefront sensor data and image-plane data, respectively. Thus, to obtain an unbiased estimator, two more datasets must be collected. The number of frames in the reference star dataset need not be the same as for the imaging dataset. As for speckle imaging, the atmospheric statistics must be the same for the reference star and imaging datasets.

Another technique that uses at least one additional data channel along with the measured image is called phase diversity. In the additional data channels, images are also recorded, but with known field phase aberrations intentionally introduced to add known diversity to the unknown field phase corrupting the measurements. The most commonly introduced field phases aberration in an additional channel is defocus due to its ease of implementation. In the case of defocus, the additional data channel along with the imaging channel can be viewed as samples of the focal volume of the imaging system. Both the noise-free object and the unknown field phase aberrations can be recovered from the measured data using optimization techniques to enforce data consistency along with other prior knowledge constraints. These types of systems can use higher-order aberrations as well. The phase diversity concept has been generalized to an approach known as computational imaging, where an imaging system is designed by taking into account its intended use, the expected image degradation mechanisms, and the fact that the image will be the product of both an optical system and a postprocessing algorithm.

Pre-Detection Compensation

The primary pre-detection-only compensation technique in use today employs dynamical electro-optical systems to sense and remove the phases in the impinging wavefront due to atmospheric turbulence. This technique, known as adaptive optics, provides much higher-quality data than post-detection compensation techniques, when there is enough light from the object being imaged or a guidestar (artificial or natural) to drive the adaptive optics system at a bandwidth high enough to keep up with the temporal changes in atmospheric turbulence. It consists of a wavefront sensor that measures the portion of the phasefront in the pupil of the telescope that is due to atmospheric turbulence, a wavefront reconstructor that generates the deformable mirror drive signals from the wavefront sensor data, and a deformable mirror that applies the conjugate of the atmospheric phase distortions to a reimaged pupil. Some of the light reflected off the deformable mirror is then sent to the imaging sensor. Adaptive optics systems are discussed in much greater detail in **Imaging**: Adaptive Optics. Here, it suffices to say that no adaptive optics system produces imagery that is diffraction limited in quality, although, under the right conditions, the images can be near diffraction limited. However, atmospheric seeing conditions fluctuate significantly over a night as well as seasonally. Therefore, virtually all adaptive optics images benefit from applying post-compensation techniques, as is discussed next.

Hybrid Techniques

Adaptive Optics and Post-Processing

Often, an adaptive optics system does not fully correct for atmospheric turbulence. This can occur when atmospheric conditions become worse than the system is designed to handle. Also, an adaptive optics system may be intentionally designed to only partially correct for atmospheric turbulence in order to lower the cost of the system. As a result, there still may be significant atmospheric blurring in systems that use adaptive optics. For this reason, additional post-processing of adaptive optics imagery can result in significant improvement in image quality. All of the techniques discussed in the section on post-detection compensation can be applied to adaptive optics imagery. Applying these techniques should never degrade image quality, and will improve the quality if the adaptive optics system doesn't fully correct for atmospheric turbulence. Generally speaking, the amount of improvement decreases as the performance of the adaptive optics system increases. Because the adaptive optics system's performance parameters can vary as atmospheric conditions vary, the blind deconvolution algorithms are especially applicable to processing adaptive optics imagery since they estimate the overall system transfer function from the imaging data. Because speckle imaging techniques require a separate estimate of the overall system transfer function, it can be difficult and time-consuming to obtain all of the additional data necessary to apply these techniques to adaptive optics data. Since these post-detection compensation techniques require short-exposure images, the adaptive optics system needs to have a camera capable of collecting short-exposure images. Finally, even if an adaptive optics system fully corrects for the atmosphere, deconvolution techniques that remove the diffraction-limited transfer function amplitude attenuation can improve image quality.

Partially Redundant Pupil Masking

Although adaptive optics systems can function very well, currently they are expensive and require significant maintenance. The primary benefit of adaptive optics systems is that they remove atmospheric turbulence phase distortions in the pupil of the telescope, thus permitting light from all areas of the pupil to interfere in phase when creating an image. An interferometric view of imaging reveals that multiple regions of the light in the pupil contribute to a single spatial frequency in the image. When the phasefront in the pupil is distorted by the atmosphere, these multiple regions tend to not interfere in phase, reducing the amplitude of the image's Fourier component at that spatial frequency. This amplitude reduction effect occurs for all spatial frequencies greater than the limit imposed by the atmosphere, r_0/λ. Therefore, techniques have been developed to modify the wavefront by masking part of the pupil prior to measuring the field with an imaging sensor. This class of techniques has various names including optical aperture synthesis, pupil masking, and aperture masking. Here this class of techniques will be referred to as partially redundant pupil masking (PRPM).

PRPM techniques function by masking the pupil in such a way as to minimize the amount of out-of-phase interference of light on the imaging detector. Early efforts used pupil masks that were opaque plates with r_0-sized holes in them. Because the atmospheric coherence length is r_0, all the light from an r_0-sized hole is essentially in phase. By choosing the holes in the mask so that only two holes contribute to a single spatial frequency in the image, out-of-phase interference problems are removed. This type of mask is called a nonredundant pupil mask. Of course, a significant amount of light is lost due to this masking, so only bright objects benefit from this particular implementation. Often, the pupil mask needs to be rotated in order to obtain enough spatial frequencies in the image to be able to reconstruct a high-quality estimate of the object. To increase the light levels in the image, masks have been made to allow some out-of-phase interference for a given spatial frequency, thus trading off the amplitude of the spatial frequencies allowed in the image with the amount of light in the image. Masks with this property are called partially redundant pupil masks.

One version of PRPM is called the variable geometry pupil (VGP) technique. It is envisioned to work in conjunction with an adaptive optics system that does not fully correct the wavefront for atmospheric phase distortions. It employs a spatial light modulator or a micro-electromechanical mirror between the deformable mirror and the imaging sensor. Either of these two devices is driven by signals from the wavefront sensor in the adaptive optics system in such a way that the portions of the pupil where the phasefront has a sufficiently large second derivative is blocked, while the remainder of the phasefront is transmitted.

All PRPM techniques benefit from post-compensation processing. For the case of nonredundant pupil masks, post-processing techniques are required to convert the discrete-spatial-frequency information into an image with reasonable quality. These techniques are essentially the same techniques as used in interferometric imaging, not those described in this article. When partially redundant pupil masks are used (including VGP masks), the image may include enough spatial frequency content that the methods described in the post-detection compensation section can be applied.

See also

Imaging: Adaptive Optics; Interferometric Imaging; Inverse Problems and Computational Imaging.

Further Reading

Ayers GR, Northcott MJ and Dainty JC (1988) Knox–Thompson and triple-correlation imaging through atmospheric turbulence. *Journal of the Optical Society of America A* 5: 963–985.

Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: Institute of Physics Publishing.

Craig IJD and Brown JC (1986) *Inverse Problems in Astronomy*. Bristol, UK: Adam Hilger Ltd.

Dainty JC (1984) Stellar speckle interferometry. In: Dainty JC (ed.) *Topics in Applied Physics, vol. 9: Laser Speckle and Related Phenomena*, 2nd edn, pp. 255–320. Berlin: Springer-Verlag.

Goodman JW (1985) *Statistical Optics*. New York: John Wiley and Sons.

Haniff CA and Buscher D (1992) Diffraction-limited imaging with partially redundant masks I: Infrared imaging of bright objects. *Journal of the Optical Society of America A* 9: 203–218.

Katsaggelos AK and Lay KT (1991) Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In: Katsaggelos AK (ed.) *Digital Image Restoration*, pp. 143–176. Berlin: Springer-Verlag.

Nakajima T and Haniff CA (1993) Partial adaptive compensation and passive interferometry with large ground-based telescopes. *Publications of the Astronomical Society of the Pacific* 105: 509–520.

Roggemann MC and Welsh B (1996) *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.

Schulz TJ (1993) Multiframe blind deconvolution of astronomical images. *Journal of the Optical Society of America A* 10: 1064–1073.

Tyson RK (1991) *Principles of Adaptive Optics*. San Diego, CA: Academic Press.

Youla DC and Webb H (1982) Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Transactions on Medical Imaging* M1-1: 81–94.

INCOHERENT SOURCES

Contents

Lamps
Synchrotrons

Lamps

V Roberts, Roberts Research and Consulting, Inc.,
Burnt Hills, NY, USA

© 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

Introduction

Electric lamps are by far the most common source of light used on Earth, other than our sun. These sources are composed of a very large number of individual radiating elements, all acting independently. This independence causes the light generated by electric lamps to have a very low degree of spatial and temporal coherence. When compared to lasers, lamps are usually considered to be incoherent sources.

Electric lamps are widely used for general lighting, such as residential, commercial and industrial space, and task lighting. A much less frequent use of incoherent sources, but perhaps of more interest to the reader, is as wavelength, intensity, and total light output standard used to calibrate various coherent light sources. We will discuss the general characteristics of incandescent, gas discharge, and solid state lamps, with sources in the Further Reading section at the end of this article, which are used as calibration standards for coherent sources.

Definitions

Before discussing individual lamp technologies, we need to define some of the key terms that are used for describing the performance of electric lamps. Most of these definitions are based on or abstracted from the IESNA Lighting Handbook:

- *Efficacy*: The total light output of a lamp, measured in lumens, divided by the power input in watts. Efficacy has the units of lumens per watt, or $\mathrm{l\,m\,W^{-1}}$, and can be greater than 100. The efficacy of common light sources varies from $10\,\mathrm{l\,m\,W^{-1}}$ to over $200\,\mathrm{l\,m\,W^{-1}}$.
- *Correlated Color Temperature* (*of a light source*) (*CCT*): the absolute temperature of a blackbody whose chromaticity most nearly resembles that of the light source. CCT is measured in Kelvin (K). The CCT of typical light sources varies from 2000 K to 6000 K, with the most popular sources being in the range from 2700 K to perhaps 4500 K.
- *Color Rendition Index* (*of a light source*) (*CRI*): a measure of the degree of color shift objects undergo when illuminated by the light source as compared with those same objects when illuminated by a reference source of comparable color temperature. The highest possible CRI is 100. The lowest theoretical CRI is less than zero, but CRI less than 50 is generally considered to have no value in determining the color quality of a light source.
- *Photopic Luminous Efficiency Function* (V_λ): a measure of the sensitivity of the human eye in lumens per watt of radiation received, as a function of wavelength, for luminance levels higher than $\sim 3\,\mathrm{cd\,m^{-2}}$. V_λ is based on relatively few observations, and individuals may differ significantly from the standard V_λ function. V_λ provides the basis for measuring the efficacy of all light sources.
- *Visible Light*: generally accepted to mean radiation between 380 nm and 780 nm.

Incandescent Lamps

Incandescent lamps are quasi-blackbody thermal radiators. A true blackbody radiator would generate light according to Planck's Law, as shown in eqn [1] and Figure 1, for various temperatures. Most incandescent lamps use tungsten filaments, and the emissivity of tungsten, the ratio of its radiative output to that of a blackbody at the same temperature, is about 0.44 in the visible region of the spectrum and about 0.33 in the infrared. The radiative output of a tungsten incandescent lamp is, therefore, less than half that of a blackbody of the same temperature and the intensity versus wavelength curve does not exactly follow Plank's law, being slightly favorable to visible

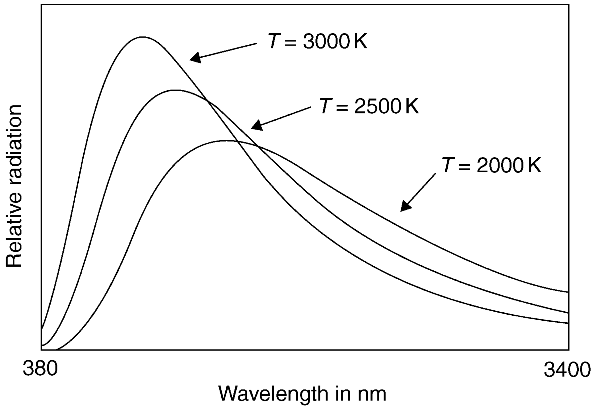

Figure 1 Radiation from blackbody source at various temperatures. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

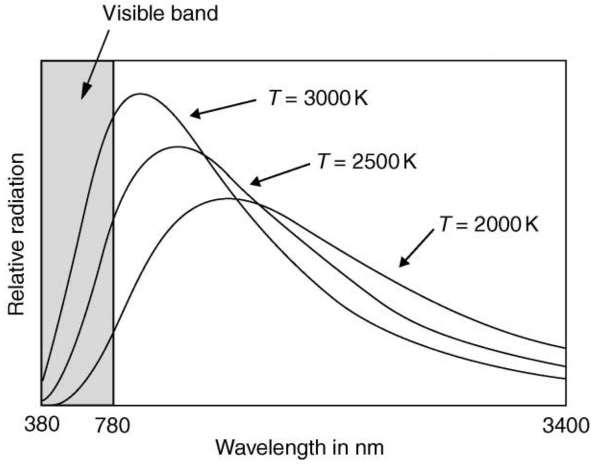

Figure 2 Radiation from blackbody source at various temperatures with visible band. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

radiation:

$$E(\lambda) = \frac{2\pi h c^2 10^{-9}}{\lambda^5(\exp^{hc/k\lambda T} - 1)} \qquad [1]$$

where:

E = power per square meter of surface area per nm of wavelength
λ = wavelength in meters
h = Plank's constant, 6.63×10^{-34} Joule seconds
c = speed of light, 3×10^{8} meters sec^{-1}
κ = Boltzman's constant, 1.38×10^{-23} Joule K^{-1}
T = temperature of the source in K

Tungsten lamps have excellent color quality, reflected in a perfect CRI of 100, are inexpensive, available in a great variety of sizes, shapes, and operating voltages, have a small source size that permits excellent light control with relatively small reflectors and lenses, run equally well on AC or DC power, can be easily dimmed, can have very long life, are insensitive to operating position, are relatively insensitive to ambient temperature, contain no hazardous materials, and light virtually instantly, though not as fast as LEDs. However, incandescent lamps have one major disadvantage; they are very inefficient. Most of the energy consumed by incandescent lamps is radiated in the infrared, while only 5% to 10% is radiated in the visible portion of the spectrum. The reason for this is shown in Figure 2, which shows the visible region of the spectrum superimposed on the blackbody radiation curves.

There are a number of techniques that have been used or are being investigated to increase the efficacy of incandescent lamps. First, as can be seen in Figure 2, increasing the filament temperature shifts the wavelength for peak emission toward lower wavelengths, and will result in a larger fraction of the radiated energy being in the visible region. Unfortunately, increasing filament temperature also increases the rate of tungsten evaporation and therefore shortens the life of tungsten lamps, since evaporation of tungsten is the dominant incandescent lamp failure mode. In a similar manner, the life of incandescent lamps can be increased to arbitrarily long times by decreasing filament temperature. This technique, which is the means used in far too many 'long life' incandescent lamps, unfortunately leads to a decrease in the already low efficacy of incandescent lamps. The approximate dependences of life, efficacy, and color temperature on operating voltage, for common incandescent lamps operating between 90% to 110% of rated voltage, are shown in eqns [2], [3], and [4]:

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{VOLTS}_2}{\mathrm{VOLTS}_1}\right)^{13} \qquad [2]$$

$$\frac{\mathrm{EFFICACY}_1}{\mathrm{EFFICACY}_2} = \left(\frac{\mathrm{VOLTS}_1}{\mathrm{VOLTS}_2}\right)^{1.9} \qquad [3]$$

$$\frac{\mathrm{COLOR\_TEMPERATURE}_1}{\mathrm{COLOR\_TEMPERATURE}_2} = \left(\frac{\mathrm{VOLTS}_1}{\mathrm{VOLTS}_2}\right)^{0.42} \qquad [4]$$

Combining eqns [2] and [3] yields eqn [5]:

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{EFFICACY}_2}{\mathrm{EFFACACY}_1}\right)^{6.84} \qquad [5]$$

Under the assumption that the color temperature of an incandescent lamp is approximately the same as the operating temperature of the filament, we can combine eqns [2] and [4] to obtain eqn [6], and combine eqns [3] and [4] to obtain eqn [7]:

$$\frac{\mathrm{LIFE}_1}{\mathrm{LIFE}_2} = \left(\frac{\mathrm{TEMPERATURE}_2}{\mathrm{TEMPERATURE}_1}\right)^{30.95} \qquad [6]$$

$$\frac{\mathrm{EFFACACY}_1}{\mathrm{EFFICACY}_2} = \left(\frac{\mathrm{TEMPERATURE}_1}{\mathrm{TEMPERATURE}_2}\right)^{4.52} \qquad [7]$$

One way to allow operation with higher filament temperatures, while not decreasing filament life, is to fill the lamp with a high molecular weight and/or high pressure gas, as either will retard tungsten evaporation. Lamps using krypton or xenon in place of the normal argon fill allow for a moderate increase in filament temperature. Better yet is a fill of halogen gas, such as HBr, CH_3Br, or CH_2Br^2 plus Kr or Xe at total pressures that range from 2 to 15 atmospheres. The high total pressure inhibits tungsten evaporation from the filament, and the halogen fill gas reacts with tungsten that deposits on the wall of the lamp to remove this tungsten and maintain light output over the life of the lamp. The tungsten removed from the wall of the lamp is redeposited on the tungsten filament, increasing lamp life under some circumstances. Due to the high gas pressures involved, tungsten halogen incandescent lamps are constructed in small ‘filament tubes’ to minimize both the amount of gas needed and the total explosive energy. The filament tubes are manufactured from quartz or special high temperature glass. For general purpose applications, the quartz or high temperature glass filament tube is enclosed in a larger glass envelope to protect the user should the filament tube explode and to protect the filament tube from environmental contaminants, such as salts from human fingers.

Since 90% to 95% of the energy consumed by an incandescent lamp is radiated as infrared energy, it is easy to see that the potential exists to dramatically increase the efficacy of incandescent lamps by reflecting some of this infrared energy back onto the filament, thus allowing the filament to be maintained at its normal operating temperature with far less electrical power. The reflector must have very low absorption in the visible region of the spectrum, high reflectivity in the IR and a reasonably sharp transition between the two, because the greatest IR emission is just below the visible band. The reflector and filament must also be precisely positioned so that the reflected IR radiation hits the filament, and the reflector must also be low enough in cost to be commercially viable. Successful IR reflectors have been made from multilayer dielectric films. These are similar to the coatings used to make high reflectivity mirrors for He–Ne lasers, but they have much broader bandwidth and lower reflectivity. The reflector-filament alignment problem is most easily solved using the geometry of a tungsten halogen filament tube as the carrier for the IR reflector. The typical efficacy gain is about 50%: a 60-watt IR-halogen lamp produces about as much light as a 90-watt halogen lamp.

Other methods being investigated to increase the efficacy of incandescent lamps involve modification of the filament to decrease its emissivity in the IR and/or increase its emissivity in the visible, and hence turn it into more of a ‘selective radiator’ than natural tungsten. Three general methods are under investigation: one involves creating a dense pattern of submicron size cavities on the surface of the tungsten to inhibit IR radiation; a second involves coating the tungsten with a material that is itself a selective radiator that favors the visible portion of the spectrum, while a third involves replacing the tungsten filament with a material that is a natural selective radiator. No products exist that employ these techniques, but active R&D projects are underway that may yield a more efficient incandescent lamp in the future.

Incandescent lamps can be used as standards for total flux and color temperature if carefully constructed and operated. The National Institutes for Standards and Technology (NIST) in the US sells 1000-watt tungsten halogen lamps calibrated to an accuracy of 0.6% in luminous intensity and 8 K in color temperature at a color temperature of 2856 K. Similar service are provided by standards organizations in other countries.

Discharge Lamps

The most efficient lamps produced today are based on electric discharges in gasses. These include ‘low pressure’ lamps such as fluorescent lamps and low pressure sodium lamps; and ‘high pressure’ or high intensity discharge (HID) lamps such as metal halide, high pressure sodium, high pressure mercury, and sulfur lamps. These lamps are all more efficient than incandescent lamps because they are ‘selective radiators’ instead of blackbody sources and have been designed to generate as much radiant energy as possible in the visible band while minimizing energy produced outside the visible band. The best white light discharge lamps have conversion efficiencies from electric power to visible light of 25% to 30%, making them almost five times as efficient as the typical incandescent lamp.

Fluorescent Lamps

Fluorescent lamps, as shown in Figure 3, are generally constructed in a cylindrical soda lime glass envelope with a length to diameter ratio significant greater than one. Most contain thermionic electrodes at each end to generate an electric field that establishes and sustains the discharge. The lamps are filled with about 2 Torr of Argon and/or another rare gas, and a few mg of liquid mercury, while the inside surface of the tube is coated with a phosphor that converts UV radiation generated by excited mercury atoms into visible light.

As with any other low pressure gas discharge, various discharge regions are aligned along the axis of the tube:

- Aston Dark Space
- Cathode Glow
- Cathode Dark Space
- Negative Glow
- Faraday Dark Space
- Positive Column
- Anode Glow
- Anode Dark Space

The axial lengths of all these regions, except for the positive column, are fixed by the properties of the cathode and anode, the fill gas type and pressure, and the discharge current. The length of the positive column is variable and fills the space between the Faraday Dark Space and the Anode Glow. The cathode and anode regions of the discharge have a total axial length of about twice the diameter of the discharge tube. Therefore, if the length-to-diameter ratio of the discharge tube is substantially greater than one, the majority of the tube is filled by the Positive Column, as shown in Figure 3. The positive column is the part of the discharge that creates the UV photons that excite the phosphor and cause it to generate visible light. The remainder of this discussion will therefore focus on the positive column.

The basic operation of a fluorescent lamp positive is shown in Figure 4. Free electrons are accelerated by the electric field established between the two electrodes. These electrons collide with both the mercury and rare gas atoms, but selectively excite and ionize mercury atoms, since the excitation and ionization energies of mercury are significantly lower than that of any rare gas used in the lamps. Ionization of mercury creates more free electrons to sustain the discharge. The gas density in the lamp is low enough to require that most of the recombination of mercury ions and free electrons takes place on the wall of the tube, since this recombination requires a three-body collision. The excited mercury atoms generate both UV and visible photons, and the UV photons generate visible light once they reach the phosphor coated on the inner surface of the lamp.

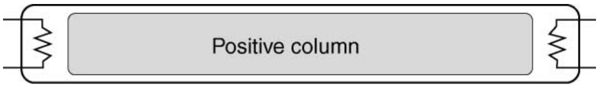

Figure 3 Linear fluorescent lamp. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

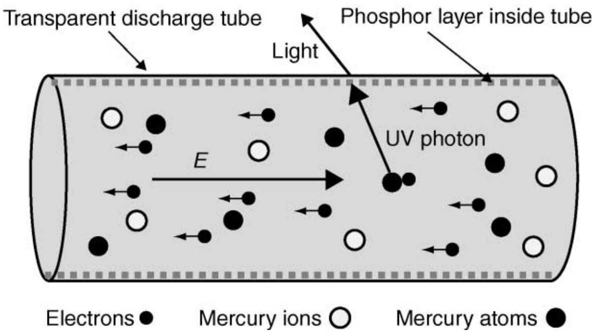

Figure 4 Operation of fluorescent lamp positive column. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

The high efficiency of fluorescent lamps is directly related to the fact that 60% to 65% of the electric power consumed by the positive column of the fluorescent lamp discharge is converted into 245 nm UV photons as a result of the resonance transition from the 6^3P_1 state of mercury at 4.86 eV to the 6^1S_0 ground state. The resonance transition from the 6^1P_1 state of mercury generates 185 nm UV photons that carry about 10% of the energy. About 3% of the positive column energy ends up in mercury visible lines at 405 nm, 436 nm, 546 nm, 577 nm, and 579 nm.

The 245 nm and 185 nm photons generated by the excited mercury atoms travel only a fraction of a mm before they are absorbed by a ground-state mercury atom. However, due to the relative absence of other de-excitation mechanisms, a large fraction of these absorbed photons are re-emitted by the absorbing atoms, and most of the UV photons generated in the core of the discharge eventually reach the phosphor coated interior wall of the discharge tube where they are converted into visible light (and waste heat).

The phosphors used in modern fluorescent lamps are a blend of three rare-earth phosphors:

- Blue: $BaMg_2Al_{16}O_{27}:Eu^{2+}$
- Green: $CeMgAl_{11}O_{19}(Ce^{3+}):Tb^{3+}$
- Red: $Y_2O_3:Eu^{3+}$

Rare earth phosphors provide a combination of high efficiency, high color rendition index, and high resistance to degradation that far exceeds the performance of the halophosphate phosphors used in older generations of fluorescent lamps.

However, due to the Stokes Shift, there is a substantial energy loss converting the UV photons produced by excited mercury atoms into visible light, even using a 'perfect' phosphor that has quantum efficiency of 1.0. The average energy of a visible photon is 2.43 eV, using the 380 nm to 780 nm limits for visible radiation discussed at the start of this article. When one 245 nm photon that has 4.885 eV of energy is converted to one visible photon with an average energy of 2.43 eV, a bit over 50% of the photon's energy is lost. The situation is worse for the 185 nm photons, which each have 6.707 eV of energy. They lose 64% of their energy when they are converted to visible photons, so it is good that the fluorescent lamp discharge generates six to seven 245 nm photons for each 185 nm photon. The Stokes shift is responsible for the largest single energy loss mechanism in fluorescent lamps. Work is underway to develop phosphors with a quantum efficiency greater than 1.0, so less energy will be lost in the UV-to-visible light conversion process, but no practical phosphors have been developed to date.

The mercury pressure for optimum UV production is about 6×10^{-3} Torr, which is the vapor pressure of mercury over a liquid pool at 40 °C. At higher mercury pressure there is too much absorption by ground state mercury atoms, and at lower pressures there is an insufficient number of mercury atoms for excitation by the electrons. The diameter of most linear lamps is chosen so that the coolest spot on the interior of the lamp wall will be at 40 °C when the lamp is operated at its design power in an ambient temperature of 25 °C. Usually this 'cold spot' will be near the middle of the lamp, but in some higher power linear lamps the wall temperature is well over 40 °C. In these lamps a cold spot is often created at the end of the lamp by use of a longer electrode mount and a heat shield mounted between the electrode and the end of the lamp. Obviously, the performance of fluorescent lamps designed in this manner is highly sensitive to ambient temperature. In some smaller, high-power density lamps, most commonly compact fluorescent lamps (CFLs), it is not possible to create a region with a temperature as low as 40 °C. In these lamps the mercury pressure is controlled by amalgamating the mercury with Pb, Bi, or Sn, creating a compound that both reduces the mercury vapor pressure at any given temperature and also reduces the variability of mercury vapor pressure with temperature. These mercury amalgams have the secondary advantage of making the fluorescent lamp less sensitive to ambient temperature.

Mercury, which is the key ingredient in fluorescent lamps, is considered to be a hazardous material by regulatory agencies in many countries. At least as early as the mid-1980s researchers began to study ways to reduce or eliminate mercury from fluorescent lamps.

Development of a mercury-free fluorescent lamp is very difficult and the only mercury-free fluorescent lamp that exists, as of mid-2003 has an efficacy of about 10 l mW $^{-1}$, one-tenth the efficacy of the best mercury-based fluorescent lamp (Osram Sylvania LINEX® Linear Excimer Lamp System, Model LINEX A3-10W40). There are three key problems that have to be solved to develop an efficient mercury-free fluorescent lamp system:

(i) Identify a nontoxic material that generates UV photons and has its optimum gas density at a convenient and safe temperature. The most promising candidate is xenon, used as either an atomic radiator or an excimer.

(ii) Develop a phosphor that produces high-quality white light with high efficiency when excited by UV photons from the chosen mercury-replacement material. This has proven a major challenge. The resonance line of atomic Xe has a wavelength of 147 nm, while the output from a xenon excimer discharge is at 172 nm. The Stokes shift loss converting each of these to visible light is 71% for atomic xenon and 66% for xenon excimers. To reduce the Stokes shift loss, a quantum-splitting phosphor, one that generates more than one visible photon for each UV photon, is needed. Most of the effort to develop mercury-free fluorescent lamps has been concentrated in this area. No viable phosphor has been developed, but work continues.

(iii) Develop a 'reservoir' for xenon or the other working gas. The mass of 'active' gas needed for a fluorescent lamp is small; a 4-foot, 1.5-inch diameter lamp needs 0.1 mg of mercury in the gas phase to operate properly. However, during life, some of this gas is consumed through physical attachment or chemical reaction and must be replaced. The drop of liquid mercury or small mass of mercury amalgam used in conventional fluorescent lamps provides a reservoir that maintains the mercury gas pressure in the presence of these consumption mechanisms.

In contrast to the failure to develop a mercury-free fluorescent lamp, there has been great success in reducing the amount of mercury used in each lamp. A typical 4-foot, 1.5-inch diameter fluorescent lamp produced in the 1970s would have 50 mg to 100 mg of mercury. This high amount of mercury was mostly the result of the type of manufacturing equipment used, but mercury consumption required that each lamp had 15 mg to 20 mg of mercury in order to survive for its rated 20 000-hour life. The development of coatings and other materials that reduce mercury consumption, when combined with new manufacturing technology, has allowed the amount of mercury used in modern 4-foot fluorescent lamps to be reduced to 3 mg to 5 mg, and further reductions are possible with improved technology.

As stated above, the length-to-diameter ratio of most fluorescent lamps is significantly greater than one. The reason for this is that there is a 10 to 15 volt drop at the electrodes, and the power lost in these regions does not produce a significant amount of light. A large length-to-diameter ratio increases the voltage and power of the positive column, where the bulk of the UV photons are generated, relative to the power lost in the electrode regions. When the energy crunch hit in the mid-1970s, there was strong interest in developing a fluorescent lamp that could replace the inefficient incandescent lamp. However, the need for high length-to-diameter ratios meant that it would not be practical to make a fluorescent lamp shaped like an incandescent lamp. This need created the technology to manufacture CFLs which are constructed of long tubes bent or coiled into shapes that allow them to fit into spaces occupied by incandescent lamps.

Fluorescent lamps and other discharge lamps have a negative incremental voltage vs. current (V/I) characteristic, which means that their operating voltage decreases as the operating current increases. Due to this, such discharges need to be operated from a current limited power source, commonly known as a 'ballast' in the lighting industry. Ballasts were originally simple series inductors and lamps operated at the power line or mains frequency, though many line frequency ballasts are now considerable more complex than an inductor. In the 1940s, it was discovered that the efficacy of fluorescent lamps could be increased by operation at frequencies of a few kHz. It was not until the 1970s, however, until the cost of power electronics came down to the point where it was possible to make affordable high frequency electronic ballasts. Electronic ballasts operating in the 20 kHz to 100 kHz range are now used with almost all CFLs and in most commercial applications using linear fluorescent lamps.

The desire for fluorescent lamps shaped like incandescent lamps also sparked the development of electrodeless fluorescent lamps. Since the need for high length-to-diameter ratios is driven by the losses at the lamp electrodes, researchers realized they could make fluorescent lamps in short, wide shapes if they eliminated the electrodes. As shown in Figure 5, an electrodeless lamp is just like a transformer, except it has a single-turn discharge secondary instead of a multiturn copper or aluminum secondary. An open core version of an electrodeless fluorescent lamp with integral ballast is shown in Figure 6. The electric field is created by a time-varying magnetic field. Because of the relative high voltage-per-turn imposed by the discharge secondary, these lamps operate only at high frequency. There are three electrodeless fluorescent lamps on the market. The closed core design

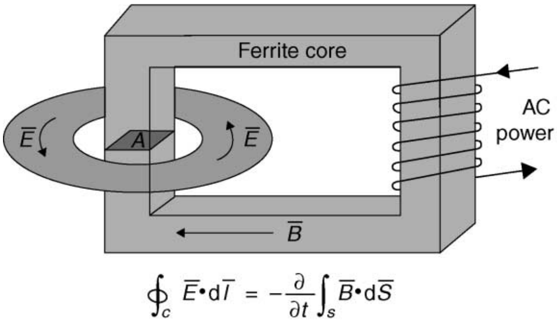

$$\oint_c \overline{E}\bullet d\overline{l} = -\frac{\partial}{\partial t}\int_s \overline{B}\bullet d\overline{S}$$

Figure 5 Basic electrodeless fluorescent lamp with closed ferrite core. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

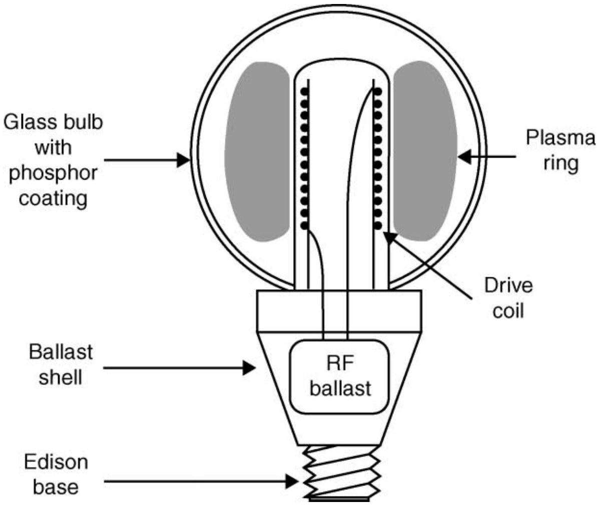

Figure 6 Open core electrodeless lamp with integral ballast. © 2005 Roberts Research and Consulting, Inc. Published by Elsevier Ltd. All rights reserved.

shown in Figure 7 operates near 250 kHz with an external ballast, while the open core design shown in Figure 8 operates near 2.5 MHz. Figure 9 shows an open core design with an integral ballast that also operates near 2.5 MHz. Some people have suggested that electrodeless fluorescent lamps are fundamentally different from normal fluorescent lamps and the high frequency field drives the phosphor to emit light. This is incorrect. Other than the manner in which the electric field is created, and the new shapes this allows, electrodeless fluorescent lamps operate in the same manner as normal fluorescent lamps.

Due to the strong dependence of fluorescent performance on mercury temperature, these lamps are not useful as flux standards. However, the five mercury visible lines listed above, plus the 365 nm line of mercury, make the ordinary fluorescent lamp an ideal source for wavelength calibration of spectrometers. Germicidal lamps, which do not use phosphor but are constructed in quartz envelopes and use the same type of mercury-rare gas discharge as fluorescent lamps, are a convenient source of 254 nm radiation that can be used to calibrate UV spectrometers. Care should be taken when using these lamps, as the 254 nm radiation can damage human eyes and skin.

High Intensity Discharge Lamps

If the mercury pressure in a discharge lamp is raised from 6×10^{-3} Torr to an atmosphere or

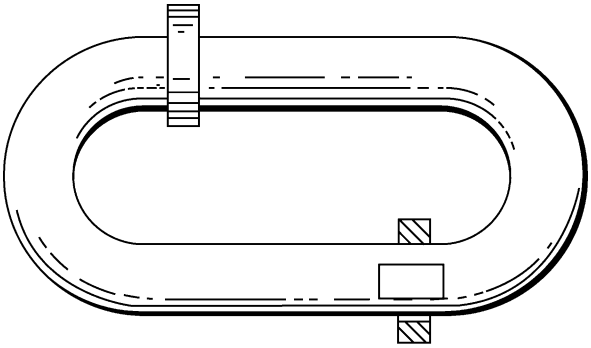

Figure 7 Closed core electrodeless fluorescent lamp, Osram Sylvania Icetron/Endura Lamp, from US Patent 5,834,905.

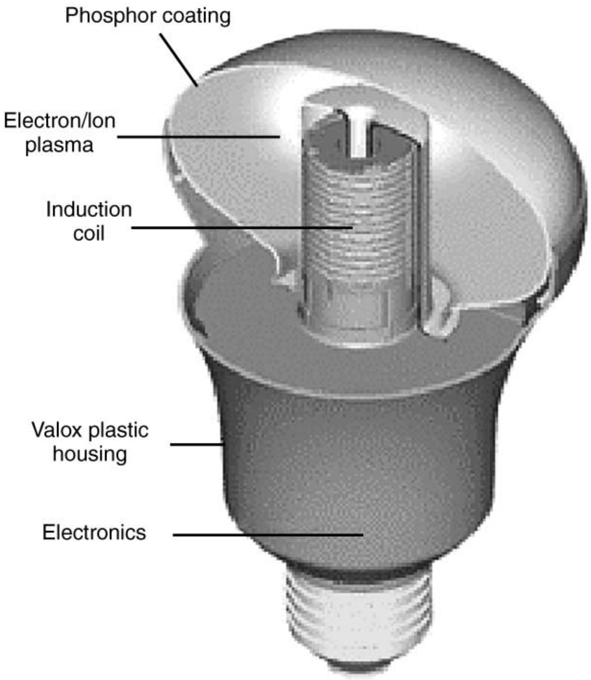

Figure 9 GE Genura® open core electrodeless fluorescent lamp with integral ballast. Courtesy of General Electric Company.

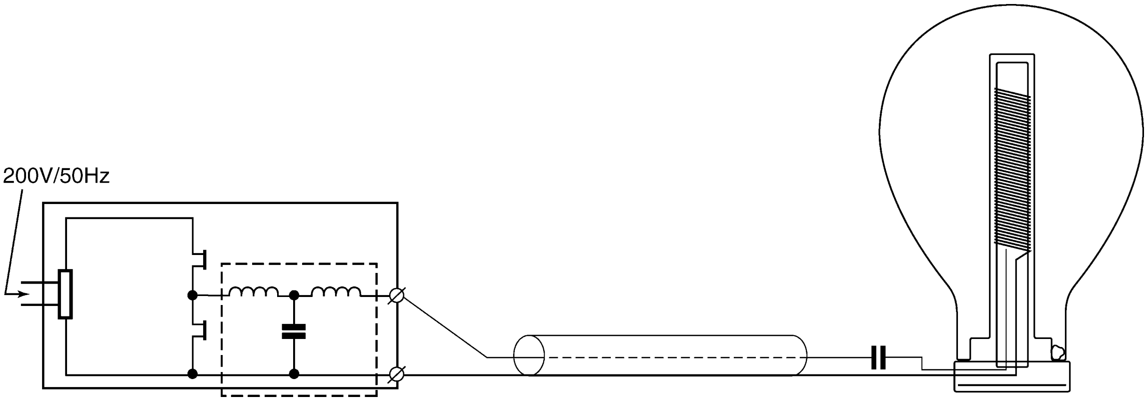

Figure 8 Open core electrodeless fluorescent lamp with separate ballast, Philips QL lamp, from US Patent 6,373,198 B1.

greater, the 245 nm and 185 nm resonance lines are trapped via absorption by ground-state mercury atoms and the radiation output from the discharge shifts to the five optically thin visible lines listed in the discussion of fluorescent lamps, plus the 365 nm near-UV line and a number of lines near 300 nm. The result is a high pressure mercury lamp that operates with an efficacy of about $55\,\mathrm{l\,m\,W^{-1}}$ and has a CRI of about 50. The high-mercury vapor pressure requires a wall temperature of at least 360 °C, which in turn requires that the arc tube be constructed from quartz and operated at high power density. A 400-watt high-pressure mercury arc tube has a diameter of 22 mm and length of about 100 mm. In contrast, a 32-watt fluorescent lamp has a diameter of 25.4 mm and a length of about 1200 mm. This accounts for the classification of these high pressure lamps as high intensity discharge (HID) lamps. Due to the high temperature of the arc tube, most HID lamps are constructed with the arc tube mounted inside a hard glass outer jacket that protects both the user and the arc tube. The high-pressure mercury lamp is a direct visible radiator, but the arc tube does generate a small amount of radiation in the 254 nm to 365 nm region, that radiation is trapped by the outer glass jacket. Some high-pressure mercury lamps have a phosphor coating on the inside of the outer jacket to convert this UV radiation into visible radiation, mostly in regions of the spectrum where the high-pressure mercury arc is deficient in radiation, thus improving the color of the lamp.

Most high-pressure mercury lamps operate at a mercury pressure of about four atmospheres. Because their spectrum is dominated by mercury line emission, the output of these lamps is deficient in the red portion of the spectrum and the lamps have poor color rendition. In 1992, Philips received a US patent for a new type of ultra-high pressure mercury lamp that they call UHP for 'Ultra High Performance'. This lamp operates at a mercury pressure of 200 atmospheres or more. The output spectrum is characterized by a significant amount of continuum radiation from quasi-molecular states of mercury and this continuum includes energy in the red portion of the spectrum that is missing from conventional high pressure mercury lamps. UHP lamps also include a small amount of halogen, such as bromine, which creates a tungsten-halogen cycle similar to that discussed earlier in relation to tungsten-halogen incandescent lamps. During lamp operation, tungsten is sputtered and evaporated from the electrodes. In the absence of the halogen fill, this tungsten rapidly deposits on the walls, leading to very short lamp life. The halogen combines with most of the tungsten before it has a chance to reach the wall of the arc tube. The halogen will also combine with any tungsten that has deposited on the wall of the arc tube. The tungsten-halogen compound formed is gaseous at the operating temperature of the lamp and will diffuse back toward the core of the discharge where the high temperature of the arc dissociates the tungsten-halogen compound. The tungsten is deposited back on the electrodes, thereby greatly extending the life of the lamp. Due to their very high pressures, UHP lamps are constructed in small discharge tubes and have a very short discharge length, typically 1 to 2 mm, making these lamps a virtual point source. Due to their excellent spectral characteristics and small source size, UHP lamps have found application in various projection applications, including LCD-based electronic projectors.

The performance of high-pressure mercury lamps can be substantially improved by adding metals such as sodium and scandium to the high-pressure mercury arc. The sodium emits near the peak of the eye sensitivity curve, providing high efficacy, while the scandium has many transitions in the visible, which fill out the spectrum and provide high CRI. If metallic elements are added directly to the high-pressure mercury arc they would deposit on or react with the wall of the arc tube. To prevent this they are added as iodides, such as NaI or $\mathrm{ScI_3}$. This creates a regenerative cycle in which the NaI and $\mathrm{ScI_3}$ are heated and dissociated in the 5000 K core of the arc, allowing the free metals to radiate. As the metals diffuse toward the cooler sections of the arc near the walls, they recombine with the iodine to reform the iodides. These metal halide lamps have produce light with an efficacy of about $90\,\mathrm{l\,m\,W^{-1}}$ to $100\,\mathrm{l\,m\,W^{-1}}$ and a CRI of about 65 to 75.

The high-pressure sodium (HPS) lamp is an HID lamp that uses hot sodium vapor to generate a golden white light. The lamp uses a mixture of sodium, mercury, and rare gas such as xenon. The core temperature of the sodium-mercury-xenon discharge is typically about 4000 K, while the temperature at the edge of the discharge and the walls of the ceramic arc tube is typically about 1500 K. At low pressure, emission from sodium is dominated by the resonance lines at 589.0 and 589.6 nm, so the light is virtually monochromatic. However, in the HPS lamp, the sodium pressure is typically 75 Torr, which is high enough to almost completely trap emission from the resonance lines, forcing radiant energy to escape in the wings of these lines, thereby increasing the color quality of the light. Since sodium will attack quartz, this mixture is contained in a translucent alumina (aluminum oxide) ceramic arc tube. As with other HID lamps, the alumina arc tube is contained in a hard glass outer bulb. HPS lamps are the most efficient generators of 'white light' available. A typical 400 watt HPS lamp produces light with an efficacy of 125 $\mathrm{l\,m\,W^{-1}}$, this high efficacy being due to the fact that the emission matches the peak of the eye sensitivity curve. The light has a CRI of only 22; however, the HPS lamp is widely used for roadway and other outdoor lighting applications.

Solid State Sources

The newest incoherent sources to be used for general purpose lighting are solid state devices: light emitting diodes (LEDs), and organic light emitting diodes (OLEDs). LEDs are very similar to laser diodes but without the two mirrors used to form the laser cavity in laser diods. Photons are generated as a result of electron-hole recombination as electrons move across the energy gap of a biased p-n junction. By appropriate selection of materials, photons can be created at visible wavelengths. Low brightness infrared and red LEDs have been available for quite some time. It was not until the mid-1980s that high brightness red LEDs were developed. These were followed by increasingly more challenging yellow, green, and then blue high brightness LEDs, developed in 1994 by Shuji Nakamura, of Nichia Chemical Industries. Nakamura added phosphor to his high brightness blue LEDs in 1996 and developed the first 'white' LED, thus triggering the current level of excitement concerning the use of LEDs to replace incandescent and discharge lamps. Monochromatic high brightness LEDs have already replaced color-filtered incandescent lamps in many applications, such as traffic lights, vehicle signal lights, and signage. In these applications, LEDs are the clear winner in life and efficacy. However, when compared to unfiltered sources, the best white LEDs are about as efficient as tungsten halogen incandescent lamps and far less efficient than fluorescent and metal halide lamps. They also cost far more per lumen than even the most expensive discharge lamp. LEDs have high internal quantum efficiency, so the main technical challenges are to fabricate devices to minimize internal light trapping in order to improve device efficiency and reduce the cost per lumen produced.

Organic light emitting diodes (OLEDs) generate light by the passage of electric current through special organic molecules, using a process known as electrophosphorescence. The electrophosphorescent material is usually coated on a transparent and perhaps flexible substrate, and overcoated with a transparent conductor. OLEDs currently have lower efficacy and shorter life than LEDs, but they also have lower manufacturing costs. If LEDs are the 'new' point sources that replace incandescent lamps and perhaps CFLs, then OLEDs are the 'new' diffuse sources that some say will replace liner fluorescent lamps.

See also

Light Emitting Diodes.

Further Reading

Cobine JD (1958) *Gaseous Conductors: Theory and Engineering Applications*. New York: Dover Publications.

de Groot JJ and van Vliet JAJM (1986) *The High Pressure Sodium Lamp*. Philips Technical Library. Deventer–Antwerprn: MacMillian Education.

Elenbass W (1951) *The High Pressure Mercury Vapor Discharge*. Amsterdam: North-Holland Publishing Company.

Fisher HE (1996) *High Pressure Mercury Discharge Lamp*, U.S. Patent, 5,497,049.

Fisher HE and Hörster H (1998) *High-Pressure Mercury-Vapor Discharge Lamp*, U.S. Patent 5,109,181.

Godyak VA (2002) Bright Idea: Radio-frequency light sources. *IEEE Industry Applications Magazine* 8: 42–49.

Hirsh MN and Oskam HJ (1978) *Gaseous Electronics, Volume 1, Electrical Discharges*. New York: Academic Press.

Mitchel AGC and Zemansky MW (1971) *Resonance Radiation and Excited Atoms*, 2nd edn. London: Cambridge University Press.

Nichia's Shuji Nakamura: Dream of the Blue Laser Diode, ScienceWatch, January/February 2000. Available at: http://www.sciencewatch.com/jan-feb2000/sw_jan-feb2000_page3.htm

Pekarski P, Hechtfischer U, Heusler G, Koerber A, Moonch H, Retsch JP, Weichmann U and Kitz A (2003) *UHP Lamps for Projection Systems*, XXVI International Conference on Phenomena in Ionized Gasses, Greifswald, Germany. Available at: http://www.icpig.uni-greifswald.de/proceedings/data/Pekarski_1

Raizer YP (1991) *Gas Discharge Physics*. Berlin: Springer-Verlag.

Rea MS (2000) *IESNA Lighting Handbook*, 9th edn. New York: Illuminating Engineering Society of North America.

Roberts VD (2003) *Radio-Frequency Lighting*. McGraw-Hill Yearbook of Science & Technology, pp. 349–352. New York: The McGraw-Hill Companies, Inc.

Schewe PF and Stein B (1999) *Physics News Update*, Number 437 (Story #2), The American Institute of Physics Bulletin of Physics News. Available at: http://www.aip.org/enews/physnews/1999/split/pnu437-2.htm

Waymouth JF (1971) *Electric Discharge Lamps*. Cambridge, Massachusetts: MIT Press.

Wharmby DO (1993) Electrodeless Lamps for Lighting: A Review. *IEE Proceedings-A* 140(6): 465–473.

Synchrotrons

R Clarke, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their discovery by Röntgen in 1895, X-rays have been used in a wide variety of applications ranging from medical and industrial radiography to crystallographic studies of the structure of biomolecules. The short wavelength ($\lambda \sim 0.1$ nm) and high penetrating power of hard X-rays makes them an ideal probe of the internal structure and composition of condensed matter. Indeed, Watson and Crick's elucidation of the double helix structure of DNA in the early 1950s, based on the X-ray diffraction patterns of Rosalind Franklin, has revolutionized biology and provided an atomistic view of the stuff of life.

The quest for ever higher spatial resolution, and the desire to obtain high-quality X-ray data on shorter time-scales, have both been prime motivations for the development of high intensity hard X-ray sources since the mid-1970s. This evolutionary trend has led to the design and construction of large synchrotron radiation facilities, serving many thousands of users and enabling forefront experimental studies across a broad range of disciplines. One such example of a national facility, dedicated to high-brightness X-ray synchrotron radiation, is the Advanced Photon Source (APS), located at Argonne National Laboratory in the United States (Figure 1). Facilities of a similar scope and purpose have also been constructed in Europe and Japan, and a number of new high brightness sources are currently under construction in several countries around the world (see Table 1).

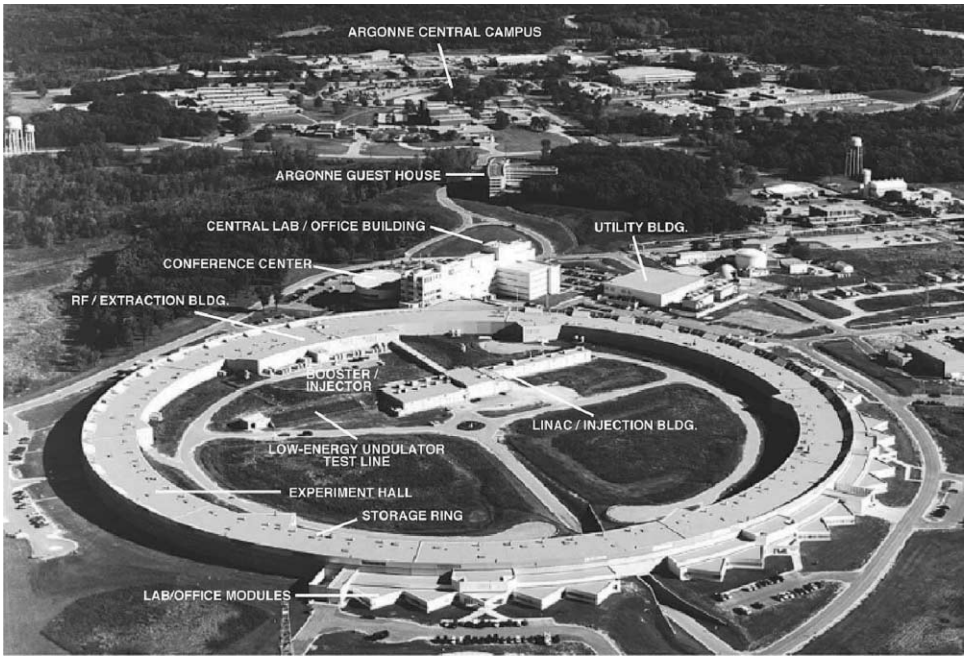

Figure 1 Aerial view of the Advanced Photon Source, Argonne National Laboratory. Courtesy of Advanced Photon Source.

Table 1 Third-generation synchrotron radiation sources

Location	*Facility*	*Electron energy (GeV)*
Grenoble, France	European Synchrotron Radiation Facility (ESRF)	6
Argonne, Illinois, USA	Advanced Photon Source (APS)	7
Nishi Harima, Japan	8-GeV Super Photon Ring (SPring-8)	8
Berkeley, CA, USA	Advanced Light Source (ALS)	1.5–1.9
Canberra, Australia	Boomerang[a]	3
Didcot, UK	DIAMOND[a]	3
Orsay, France	SOLEIL (LURE)	2.75
Karlsruhe, Germany	Angströmquelle-Karlsruhe (ANKA)	2.5
Hamburg, Germany	PETRA III[b]	6
Indore, India	INDUS II[a]	2.5
Villigen, Switzerland	Swiss Light Source (SLS)	2.4
Saskatoon, Canada	Canadian Light Source[a]	2.9
Stanford, CA, USA	SPEARS	3
Upton, New York, USA	NSLS II[b]	2.5–2.8
Trieste, Italy	Elettra	2.0–2.4

[a] Under construction as of March 2004.
[b] Planning stage.

In this article we will draw on examples from APS to illustrate the characteristics and performance of the current generation of synchrotron sources in the hard X-ray region.

From the earliest days of particle physics studies in the 1930s and 40s, it was realized that copious amounts of X-rays could be produced by accelerating charged particles in a curved trajectory, as in a cyclotron or a synchrotron. For particle physics such radiation is highly detrimental because it limits the ultimate energy of particle accelerators and forces their circumference to become inordinately large in order to reduce radiation losses.

Originally considered to be a major source of concern in particle accelerator design, synchrotron radiation soon began to emerge as a boon to X-ray science. This came about in the early 1970s after the capability was developed to produce and store significant circulating currents of high-energy charged particles (electrons or positrons) for colliding beam experiments. At that time X-ray scientists were tolerated as 'parasitic' experimenters at such storage rings, taking X-ray radiation from the bending magnets which maintained the stored current in an overall circular trajectory. The PETRA (Positron–Electron Tandem Ring Accelerator) facility at HASYLAB/DESY in Hamburg, Germany, and the Cornell Electron Storage Ring (CESR)/Cornell High-Energy Synchrotron Source (CHESS) at Cornell University, Ithaca, New York, are examples of accelerator facilities that were primarily intended for particle physics experiments and then became partly dedicated to X-ray production. Such shared synchrotron facilities are referred to as 'first-generation sources'.

The first generation sources demonstrated several attractive features of synchrotron radiation, including a high degree of collimation in the vertical direction with a divergence half-angle given approximately by the relativistic factor $\gamma^{-1} = m_0c^2/E$, where m_0c^2 is the rest-mass energy (0.51 MeV) of the electron (or positron) and E is the energy of the stored particle beam. For a bending magnet source with $E = 2-3$ GeV the beam would have a vertical collimation of ~0.3 mrad (17 mdeg). This would produce an X-ray beam with a vertical extent of ~3 mm at a distance of 10 m from the bending magnet source. The horizontal fan of bend magnet radiation is broad and is usually limited by a mask or a slit to a few mrad. The most useful characteristic of bend magnet radiation is its large bandwidth whereby the X-ray spectrum is continuous and increasing in intensity (see Figure 2) until some critical energy above which the intensity begins to fall off. Bending magnet radiation is therefore useful for X-ray measurements requiring a 'white' beam, for example, Laue X-ray diffraction, energy dispersive diffraction, and Extended X-ray Absorption Fine Structure (EXAFS) analysis.

Beginning in the 1980s, several second-generation storage rings were constructed in various locations around the world. The design of these facilities for the first time was optimized for the production of bright beams of X-ray and XUV (soft X-ray to ultraviolet) and their operation was dedicated exclusively to synchrotron radiation science. The increase in brightness provided by these new storage ring sources, compared to standard laboratory X-ray sources (e.g., rotating anode generators), was dramatic (5 or 6 orders of magnitude) and this opened up many new experiments that were not previously feasible. However, it was soon realized that the capabilities of such sources could be extended to

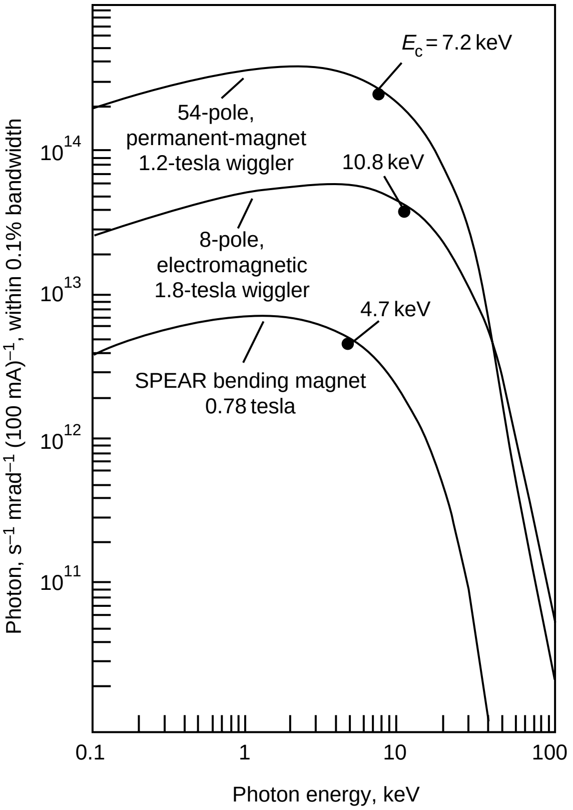

Figure 2 Comparison of bending magnet and wiggler sources. With permission from Stanford Synchrotron Radiation Laboratory, Stanford Linear Accelerator Center, Menlo Park, CA.

even higher levels of intensity and to significantly higher X-ray energies.

One of the limitations of bending magnet sources relates to the X-ray critical energy alluded to above:

$$E_c = 3\hbar c\gamma^3/4\pi\rho \quad [1]$$

where ρ is the radius of curvature imposed on the particle beam trajectory by the bending magnet. For a given particle acceleration energy, E, eqn [1] leads to a critical X-ray energy given by:

$$E_c\ (\text{keV}) = 2.2\,E^3\ (\text{GeV})/\rho\ (\text{m}) \quad [2]$$

For a typical second-generation machine with $E = 2.5$ GeV and $\rho = 7$ m (limited by the magnetic field generated by the dipole electromagnet), the critical energy is only about 5 keV. There is useable flux well above this energy, but X-ray photons become scarce in the energy range above 10 keV for a bending magnet source with these characteristics. Since E is effectively fixed by the design parameters of the storage ring, the only possibility to achieve higher photon energies is to install additional magnetic devices which will introduce perturbations of the trajectory with effectively smaller radius of curvature in the denominator of eqn [2]. Such a device, known as a 'wiggler', shifts the effective critical X-ray energy to higher values in proportion to the strength of the magnetic field B (see Figure 2). Using wigglers based on superconducting magnets ($B \sim 5$ Tesla), E_c can be raised to >20 keV, a value which would require accelerator energies of more than 6 GeV using conventional bending magnets alone. In addition to shifting the output spectrum to higher energies, it is possible to enhance the intensity at a given energy by producing multiple wiggles with no net displacement of the particle trajectory. Multipole wigglers are able to produce significant increases in flux in the hard X-ray range. The 27-pole hybrid wiggler installed on beamline X25 at the National Synchrotron Light Source, a second generation facility at Brookhaven National Laboratory, Upton, New York, is one of the most successful designs of its kind, providing $\sim 6 \times 10^{14}$ photons/sec per 1% bandwidth at 8 keV. When focused with a toroidal X-ray mirror, this beam is about an order of magnitude brighter than the standard bending magnet sources at NSLS. For reasons that will become clear later in this article, the radiation from multiple wiggles in such a device adds incoherently.

Insertion Devices

The idea of using wigglers to enhance the characteristics of synchrotron radiation sources has led to the design and construction of storage rings with many straight sections to accommodate wigglers and other types of 'insertion devices'. Such facilities are referred to as third-generation sources. Coming on-line in the mid-1990s, several such sources are now in mature operation and are supporting a large international community of X-ray scientists.

One of the most important types of insertion device to emerge from the development of third-generation facilities is the undulator. Conceptually, the idea behind an undulator is similar to that of the wiggler, i.e., it provides additional tightly curved transverse displacements in the particle beam. There are two important differences: in an undulator the displacements are smaller and they are more numerous, being produced by a linear periodic array of many short alternating dipoles (Figure 3). Under these conditions the synchrotron radiation emitted from each 'undulation' of the particle beam trajectory adds coherently and therefore the peak intensity from such a device can be enhanced by a factor of N^2 where N is the number of dipole magnet pairs, or periods.

A simple theoretical description of the basic operation of an undulator can be discussed in terms of radiation from a charged particle moving at

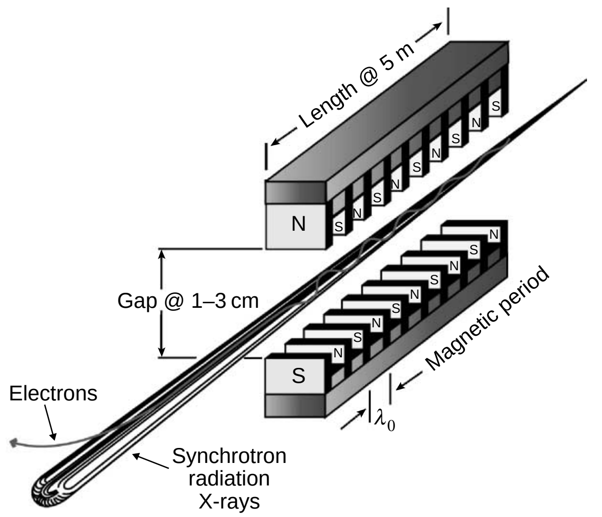

Figure 3 Schematic of undulator. Courtesy of Advanced Photon Source.

relativistic speeds through an alternating magnetic field. The magnetic field in the vertical direction (y) is taken to be periodic:

$$B_y = B_0 \sin[(2\pi/\lambda_u)z] \qquad [3]$$

where z is in the direction of average particle motion.

The transverse particle trajectory is described by:

$$x = (\lambda_u \phi_{\max}/2\pi)\sin[(2\pi/\lambda_u)z] \qquad [4]$$

where λ_u is the magnet period, and $\phi_{\max}$ is the maximum particle deflection angle:

$$\phi_{\max} = (eB_0\lambda_u c)/2\pi m_0 c^2\gamma = K/\gamma \qquad [5]$$

where

$$K = 93.4B_0\ (\text{Tesla})\lambda_u\ (\text{meters}) \qquad [6]$$

The field strength parameter K serves as a useful quantity for discussing the characteristics of various insertion device designs. For $K \leq 1$, which defines the undulator regime, the transverse angular deflections are less than or on the order of the angular spread of the emitted radiation and in this case the radiation from each of the deflections adds coherently. This is distinguished from the case of a multipole wiggler where $K \gg 1$ leading to incoherent superposition.

Most of the interesting features of undulator radiation are revealed by considering the trajectory of the particle in a Lorentz frame moving at speed v^* with respect to the laboratory frame. In this frame the particle will execute oscillatory motion with the period of oscillations T' given by the Lorentz transformation:

$$T' = \gamma^*(T - \beta^*\lambda_u/c) = \lambda_u/(v^*\gamma^*) \qquad [7]$$

with a corresponding oscillation frequency $\omega' = 2\pi/T' = (2\pi c/\lambda_u)\gamma^*$.

For $K \ll 1$, the motion in the co-moving frame is purely transverse and nonrelativistic. However, for $K \sim 1$, the oscillatory motion in this frame can have a longitudinal component and in this case the transverse motion becomes nonharmonic with frequency components at odd multiples of the fundamental frequency ω'. This nonharmonic motion is a consequence of the fact that in the laboratory frame the particle is moving with constant speed $\approx c$ but its path length is slightly longer than λ_u, owing to the transverse motion. This constrains the motion in the co-moving frame to be nonharmonic.

Undulator Spectrum

It now remains to transform ω' into the laboratory frame so that we can obtain the spectral distribution of the output of the undulator. Invoking the Doppler transformation:

$$\omega = \gamma^*\omega'(1 + \beta^* \cos\theta') \qquad [8a]$$

$$\tan\theta = (\sin\theta')/[\gamma^*(\cos\theta' + \beta^*)] \qquad [8b]$$

where θ is the angle at which the radiation is observed:

$$\omega_n(\theta) = (4\pi c n\gamma^2)/(\lambda_u[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2]) \qquad [9a]$$

or in terms of wavelength:

$$\lambda_n(\theta) = (\lambda_u/2n\gamma^2)[1 + \tfrac{1}{2}K^2 + (\gamma\theta)^2] \qquad [9b]$$

Equation [9] is the central expression describing the spectral output of an undulator. Notice how the macroscopic period of the undulator λ_u is transformed into a microscopic wavelength $\lambda_n(\theta)$ by the relativistic factor γ^2. For third-generation hard X-ray sources with particle energies in the 6–8 GeV range, $\gamma \sim 14\,000$, taking cm scales down to the angstrom level.

Equation [9] describes a spectrum of monochromatic peaks, for on-axis radiation ($\theta = 0$), with wavelengths that increase with increasing K. Since $K \sim B_0\lambda_u$ (eqn [6]) any particular harmonic can be tuned either by changing B_0 or λ_u. In most undulator designs, λ_u is fixed by the construction of the permanent magnet array, so the tuning is conveniently accomplished by varying B_0. In practice this is done by changing the vertical distance between the poles

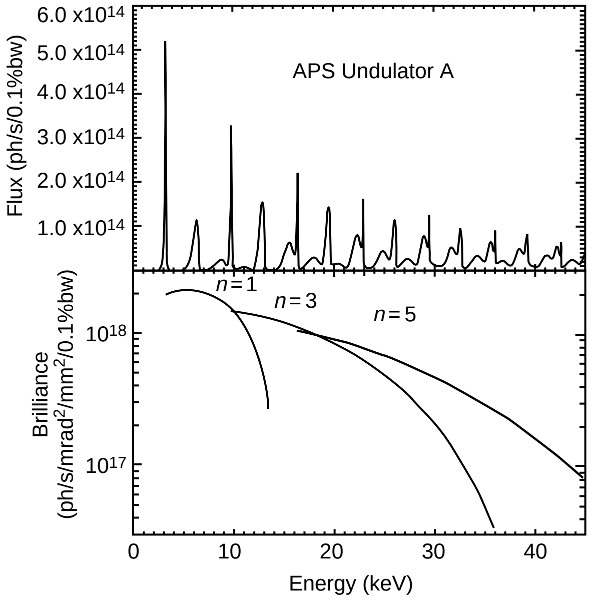

Figure 4 Undulator spectrum (upper panel) and tuning range for first, third and fifth harmonics (lower panel) of Advanced Photon Source Undulator A.

of the magnets (the 'undulator gap') similar to opening the jaws of a vise. Typical tuning ranges for the $n = 1$, 3, and 5 harmonics are illustrated in Figure 4.

Since the number of undulator oscillations of the particle beam is finite, the nth harmonic is not strictly monochromatic and has a spectral broadening of order $1/nN$, where N is the number of oscillator periods. In practice with N typically ~100, a bandwidth of ~1% is expected for the first harmonic. Additional broadening of a similar magnitude might also be expected from the θ-dependence in eqn [9], depending on the acceptance angle of the optics that is used in a particular experiment. This quasi-monochromatic radiation is sometimes referred to as 'pink beam'.

The storage ring energy of the APS was carefully chosen so that the tuning ranges of the different harmonics would overlap in energy, thus providing a continuous tuning of the X-ray energy over a wide range (Figure 4). Note also that the discrete spectral characteristics of the undulator can be deliberately smeared out to mimic the broadband (white beam) characteristics of a bend magnet (with consequent reduction in brightness). This is done simply by tilting the jaws of the undulator with respect to each other so that B_0 is made to vary along the length of the undulator. The ability to taper the undulator in this way is a standard feature of most designs.

The radiation from an undulator is sharply peaked in the forward direction in both the vertical and horizontal directions. The half-angle of the emission cone is given by:

$$\theta_{1/2} = (1/\gamma)[(1+\tfrac{1}{2}K^2)/2Nn]^{1/2} = [\lambda_n/L]^{1/2} \qquad [10]$$

where L is length of the undulator. For a typical third-generation source undulator, such as the Type A undulator at the Advanced Photon Source, $L = 2.4$ m, giving $\theta_{1/2} \approx 6$ μrad for the first harmonic tuned to 12 keV. This corresponds to an X-ray beam size of ~180 μm at a distance 30 m from the source.

The very tight collimation of undulator radiation (eqn [10]) highlights the importance of distinguishing between the brightness and the 'brilliance' of the beam. The latter term takes into account the natural collimation of the beam, as reflected in its units of photons/sec/mm^2/mrad^2/0.1% bandwidth. Undulator radiation is well matched to the acceptance angle of ideal crystal monochromators and is ~10 000 times more brilliant than the radiation from a bend magnet source. The high brilliance of undulator X-ray beams is extremely advantageous in X-ray imaging applications, X-ray microprobe measurements, and in experiments where the sample size is very small (e.g., high-pressure diamond anvil cell measurements).

In practice the emission cone of the source might be broadened a little when the undulator characteristics are convolved with those of the storage ring. The largest effect is in the horizontal direction where the photon beam divergence is dominated by the electron beam divergence. This is a consequence of the particular conditions under which the storage ring is operated in order to stabilize the particle beam orbit.

Equation [10] demonstrates that preserving the exquisite performance capabilities of an undulator insertion device places severe demands on the stability and precision of the storage ring orbit. An important measure of this precision is characterized by the 'emittance' of the particle beam defined in terms of the area of phase space (position and angular divergence) occupied by the stored beam. In general, storage rings are designed to achieve very low emittances and in current third-generation sources they are close to being diffraction limited. Table 2 summarizes the characteristics of the APS storage ring operated in 'high-emittance' and 'low-emittance' modes.

The large power densities resulting from the extreme brightness of undulator sources (up to 160 kW/mrad^2 for APS Undulator A) necessitates careful design of the X-ray optical components that are used to condition the beam. For example, if a double-bounce Silicon crystal pair is used as the first optical component to monochromatize the undulator

Table 2 Advanced photon source storage ring parameters

Parameter	*Low emittance*	*High emittance*
Storage ring energy, E	7.0 GeV	7.0 GeV
Storage ring current, I	100 mA	100 mA
Beam energy spread, $\delta E/E$	0.096%	0.096%
Horizontal emittance, ε_x	3.5×10^{-9} m-rad	7.7×10^{-9} m-rad
Vertical emittance, ε_y	3.5×10^{-11} m-rad	7.7×10^{-11} m-rad
Coupling constant	1%	1%
Horizontal beta function, β_x	14.4 m	16.1 m
Vertical beta function, β_y	4.0 m	4.3 m
Dispersion function, η_x	0.124 m	0.0 m
Horizontal beam size, σ_x	254 μm	351 μm
Vertical beam size, σ_y	12 μm	18 μm
Horizontal beam divergence, $\sigma_{x'}$	15.6 μrad	21.8 μrad
Vertical beam divergence, $\sigma_{y'}$	3.0 μrad	4.2 μrad

beam, the first crystal is normally cooled with liquid nitrogen to dissipate the heat. Alternately, high thermal conductivity diamond crystals are sometimes used as monochromators.

Helical Undulators

The use of undulators to produce extremely bright, tunable X-ray beams has enabled many new applications of X-ray science, as well as offering significant improvements to existing techniques in terms of quality of data and the time required to accumulate a high degree of statistical accuracy. One of the areas that has benefited significantly from novel undulator design is research on magnetic materials and other spin-dependent phenomena. It is often advantageous to probe such systems with circularly polarized X-rays, for example to study magnetic X-ray circular dichroism (MXCD). Early MXCD measurements utilized the left and right elliptical polarization of bending magnet radiation above and below the plane of the ring. Such measurements are difficult because the intensity falls off dramatically out of the plane. The problem to produce high-brightness beams of circularly polarized X-rays is effectively solved by designing an undulator which produces a transverse periodic helical magnetic field of constant magnitude. Several helical undulators are now in operation at third-generation sources. The soft X-ray region (500–1500 eV) is particularly interesting because it covers the L-edges of the 3d transition elements.

Coherence

While this article describes what would be nominally considered as 'incoherent' sources in the parlance of conventional optics, it should be pointed out that undulators are in fact excellent sources of coherent X-rays. This is by virtue of the high degree of collimation inherent in the emission of radiation from the undulations of a highly relativistic charged particle (eqn [10]). The spatial (transverse) coherence angle is given by:

$$\lambda_n/2d \qquad [11]$$

where d is the source size. This is the angle within which the beam is spatially coherent.

For APS undulator A, the source size is ~15 μm (vertical) × 240 μm (horizontal), giving a coherence angle in the vertical direction of 6.5 μrad and 0.4 μrad in the horizontal direction. The X-ray beam divergence is approximately 4 μrad (vertical) by 15 μrad (horizontal). This means that the beam has complete spatial coherence in the vertical direction and is coherent over about 3% of the horizontal fan. At a distance of 30 m from the source, an on-axis pinhole of ~12 micron diameter will be completely illuminated by coherent X-rays. Transversely coherent X-ray beams are finding novel applications in X-ray photon correlation spectroscopy, coherent X-ray scattering, and phase reconstruction (lensless) imaging.

Timing Structure

X-ray storage rings produce pulsed X-rays. In the Advanced Photon Source, one or more long trains of electron bunches (15 ns long) are accelerated by the linac to 325 MeV. The electron bunches are injected into a booster that ramps the single bunch from 325 MeV to 7 GeV in about 225 ms. At the end of the booster cycle, the bunch is extracted and injected into the storage ring. The bunch can be injected into any of 1296 'buckets' that are spaced by 2.842 ns, forming any arbitrary pattern after several injection cycles.

The APS storage ring rf frequency of 351.927 MHz determines the bucket spacing of 2.842 ns. The circumference of 1104 m gives a revolution frequency of 271.5 kHz which allows for 1296 buckets in 3683 ns. The storage ring is filled with a selected bunch pattern depending on the requirements for a particular operations cycle. The standard 'singlets' bunch pattern at APS is 102 mA of average current in a train of 23 bunches (4.43 mA per bunch) each spaced by 153 ns (54 buckets). There is a gap of 306 ns (108 buckets) in this pattern, useful as a timing guide. The singlets pattern has a lower lifetime than patterns with less charge per bunch, due to natural internal scattering processes. A 'top-up' mode has been devised to overcome the problem of overall intensity fall-off due to the exponential decay of the stored current. In top-up mode, new bunches are injected into the storage ring every few minutes to maintain the stored current at 102 mA.

A second fill pattern, known as the 'hybrid' pattern, is again 102 mA but distributed in one bunch of 5 mA, plus 8 groups of 7 consecutive bunches (1.73 mA/bunch), spaced by 68 ns (24 buckets). The 8 groups are diametrically opposite the 5 mA bunch, allowing for a 1.5 μs gap on both sides of the 5 mA bunch. This hybrid mode permits a single bunch to be isolated using a mechanical chopper or a fast framing camera. Since the RMS width of the single bunch is ~36 ps, relatively fast time-resolved experiments can be performed stroboscopically using this pulsed X-ray beam. In another implementation of time-resolved experiments, the singlet pattern is used, isolating each bunch with a fast avalanche photodiode detector, or an X-ray streak camera. With the streak camera it is possible to achieve time resolutions of a few picoseconds. This approach has been used to study impulsive phonon excitations and ultrafast strain wave propagation in solids.

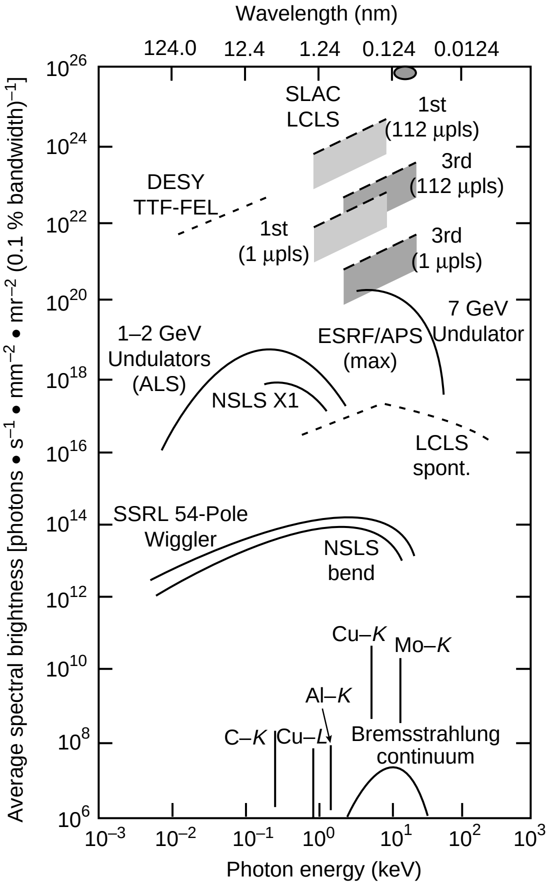

Figure 5 Comparison of spectral brilliance of X-ray sources. Courtesy of Linear Coherent Light Source, Stanford Linear Acceleration Center.

The Future

This article highlights current (third-generation) storage ring sources based on undulator-type insertion devices. These sources will provide high-brightness tunable X-rays in support of a wide range of experiments for many years to come. The brightness of these sources will increase steadily into the future as continual improvements are made, such as further reducing the emittance of the particle beam, improved undulator design, increasing the average storage ring current, and making changes to the storage ring lattice. These incremental improvements will be important to maintain the vitality and usefulness of national user facilities based on third-generation sources. New storage ring sources with designs and specifications optimized for low emittance and new types of undulators (e.g., in-vacuum undulators) will come online and will undoubtedly lead to new technical advances, supporting new and exciting science. It seems likely that linear accelerator (linac) based undulator sources will provide a promising route towards higher brilliance in the future. The superconducting Energy Recovery Linac (ERL) proposed by CHESS is an interesting concept which uses regeneration of the electron beam to overcome the high energy cost of single pass particle acceleration.

A new generation of X-ray sources (the 'fourth generation') is envisioned to take on the task of providing vastly brighter beams (>10 orders more peak brilliance than at present, see Figure 5) with much shorter pulses (~100 fs). These facilities will also be based on linear accelerators. Two such *Free Electron Laser* sources are currently being designed and should be operational towards 2010: the Linac Coherent Light Source (LCLS), to be built at the Stanford Linear Accelerator Center (SLAC), and the X-ray Free Electron Laser (XFEL) Source, to be constructed by the TESLA collaboration at HASYLAB, Hamburg.

See also

Coherence: Overview. **Physical Applications of Lasers:** Free Electron Lasers in Physics.

Further Reading

Brown G and Lavender W (1991) Synchrotron radiation spectra. In: Brown GS and Moncton DE (eds) *Handbook on Synchrotron Radiation*, vol. 3. Amsterdam: North Holland.

Duke PJ (1993) Synchrotron radiation sources. In: Michette AG and Buckley CJ (eds) *X-Ray Science and Technology*. Bristol: Institute of Physics Publishing Ltd.

Duke PJ (2000) *Synchrotron Radiation: Production and Properties*. Oxford, UK: OUP.

Mills DM (ed.) (2002) *Third Generation Hard X-ray Synchrotron Radiation Sources*. Berlin: Wiley, VCH.

INFORMATION PROCESSING

Contents

All-Optical Multiplexing/Demultiplexing
Coherent Analog Optical Processors
Free-Space Optical Computing
Incoherent Analog Optical Processors
Optical Bit-Serial Computing
Optical Digital Image Processing
Optical Neural Networks

All-Optical Multiplexing/Demultiplexing

Z Ghassemlooy, Northumbria University, Newcastle upon Tyne, UK

G Swift, Sheffield Hallam University, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The ever-increasing aggregate demand of electrically based time division multiplexing systems should have coped with the steady growth rate of voice traffic. However, since 1990, the explosive growth of the Internet and other bandwidth-demanding multimedia applications, has meant that long-haul telecommunication traffic has been increasingly dominated by data, not voice traffic. Such systems suffer from a bandwidth bottleneck due to speed limitations of the electronics. This limits the maximum data rate to considerably less than the THz bandwidth offered by an optical fiber. Optical technology is proposed as the only viable option and is expected to play an ever increasing role in future ultrahigh-speed links/networks. There are a number of multiplexing techniques, such as space division multiplexing (SDM), wavelength division multiplexing (WDM), and optical time division multiplexing (OTDM), that are currently being applied to effectively utilize the bandwidth of optical fiber as a means to overcome the bandwidth bottleneck imposed by electrical time division multiplexing (TDM). In SDM a separate optical fiber is allocated to each channel, but this is the least preferred option for increasing channel numbers. In WDM, a number of different data channels are allocated to discrete optical wavelengths for transmission over a single fiber. Dense WDM technology has been improving at a steady rate in recent years, with the latest systems capable of operating at a data rate of >1 T bps, using a large number of wavelengths over a single fiber link. However, there are a number of problems associated with the WDM systems such as:

- Performance of WDM is highly dependent on the nonlinearities associated with fiber, i.e.:
  - *Stimulated Raman scattering*: degrades the signal-to-noise (SNR) as the number of channels increases;
  - *Four-wave mixing*: limits the channel spacing;
  - *Cross-phase modulation*: limits the number of channels.
- Relatively static optical paths, thus offering no fast switching with high performance within the network;

- Switching is normally carried out by separating each wavelength of each fiber onto different physical outputs. Space switches are then used to spatially switch the separated wavelengths, an extremely inefficient way of utilizing network resources;
- The need for amplifiers with high gain and flat spectra.

In order to overcome these problems, OTDM was introduced that offers the following:

- Flexible high bandwidth on demand (>1 Tbit/s compared to the bit rates of 2.5–40 Gbit/s per wavelength channel in WDM systems);
- The total bandwidth offered by a single channel network is equal to DWDM;
- In a network environment, OTDM provides potential improvements in:
  - Network user access time, delay and throughput, depending on the user rates and statistics.
  - Less complex end node equipment (single-channel versus multichannels).
- Self-routing and self-clocking characteristics;
- Can operate at second- and third-transmission windows:
  - 1500 nm (like WDM) due to Erbium doped fiber amplifier (EDFA);
  - 1300 nm wavelengths.
- Offers both broadcast and switched based networks.

Principle of OTDM

Figure 1 show the generic block diagram of a point-to-point OTDM transmission link, where N optical data channels, each of capacity M Gbps, are multiplexed to give an aggregate rate of $N \times M$ Gbps. The fundamental components are a pulsed light source, an optical modulator, a multiplexer, channel, add/drop unit, and a demultiplexer. The light source needs to have good stabilities and be capable of generating ultrashort pulses (<1 ps). Direct modulation of the laser source is possible but the preferred method is based on external modulation where the optical signal is gated by the electronic data. The combination of these techniques allows the time division multiplexed data to be encoded inside a subnanosecond time slot, which is subsequently interleaved into a frame format. Add/drop units provide added versatility (see Figure 2) allowing the 'adding' and 'dropping' of selected OTDM channels to interchange data at chosen points on the link. At the receiving end, the OTDM pulse stream is demultiplexed down to the individual channels at the initial M Gbps data rate. Data retrieval is then within the realm of electronic devices and the distinction between electronic and optical methods is no longer relevant. Demultiplexing requires high-speed all optical switching and can be achieved using a number of methods, which will be discussed in more detail below.

The optical multiplexing (or interleaving) can be carried out at the bit level (known as bit interleaving) or at the packet level (known as packet interleaving), where blocks of bits are interleaved sequentially. This is in accord with the popular conception of packet switched networks. The high data rates required and consequently narrow time slots necessitate the need for strict tolerances at processing nodes, e.g., switches. As such, it is important that the duration

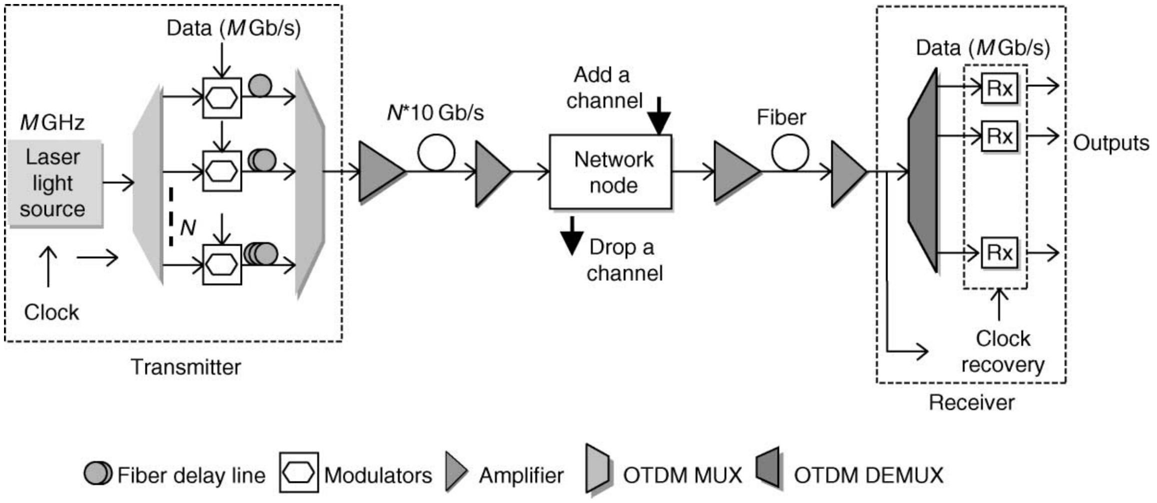

Figure 1 Block diagram of a typical OTDM transmission system.

of the optical pulses is chosen to be significantly shorter than the bit period of the highest multiplexed line rate, in order to reduce the crosstalk between channels.

Bit Interleaved OTDM

A simple conceptual description of a bit interleaved multiplexer is shown in Figure 3. It uses a number of different length optical fiber delay lines (FDL) to interleave the channels. The propagation delay of each FDL is chosen to position the optical channel in its corresponding time slot in relation to the aggregate OTDM signal. Prior to this, each optical pulse train is modulated by the data stream. The output of the modulators and an undelayed pulse train, labeled the framing signal, are combined, using a star coupler or combiner, to produce the high bit rate OTDM signal (see Figure 3b). As shown in Figure 3b, the framing pulse has a higher intensity for clock recovery purpose. At the demultiplexer, the incoming OTDM pulse train is split into two paths (see Figure 4a). The lower path is used to recover the framing pulse by means of thresholding, this is then delayed by an amount corresponding to the position of the *i*th (wanted) channel (see Figure 4b). The delayed framing pulse and the OTDM pulse stream are then passed through an AND gate to recover the *i*th channel. The AND operation can be carried out all optically using, for example, a nonlinear loop mirror, a terahertz optical asymmetric demultiplexer, or a soliton-trapping gate.

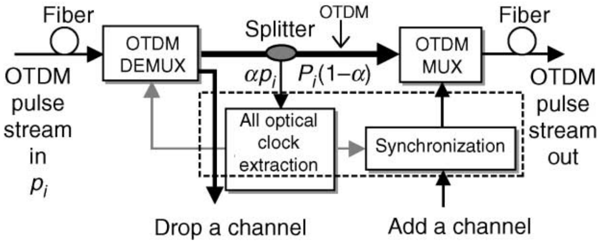
Fiber
OTDM DEMUX
Splitter
OTDM
OTDM MUX
Fiber
OTDM pulse stream in p_i
αp_i
$P_i(1-\alpha)$
All optical clock extraction
Synchronization
OTDM pulse stream out
Drop a channel
Add a channel

Figure 2 Add/drop unit in an OTDM network node.

Packet Interleaved OTDM

Figure 5a shows the block diagram of a system used to demonstrate packet interleaving. Note that the time interval between successive pulses now needs to be much less than the bit interval *T*. This is achieved by passing the low bit rate output packet of the modulator into a compressor, which is based on a feed forward delay line structure. The feed forward delay lines are based on a cascade of passive M–Z interferometers with unequal arm lengths. This configuration generates the delay times required, i.e., $T-\tau, 2(T-\tau)\ldots(2^{n-1})(T-\tau)$, etc., where $n = \log_2 k$ is the number of stages, T and τ are the

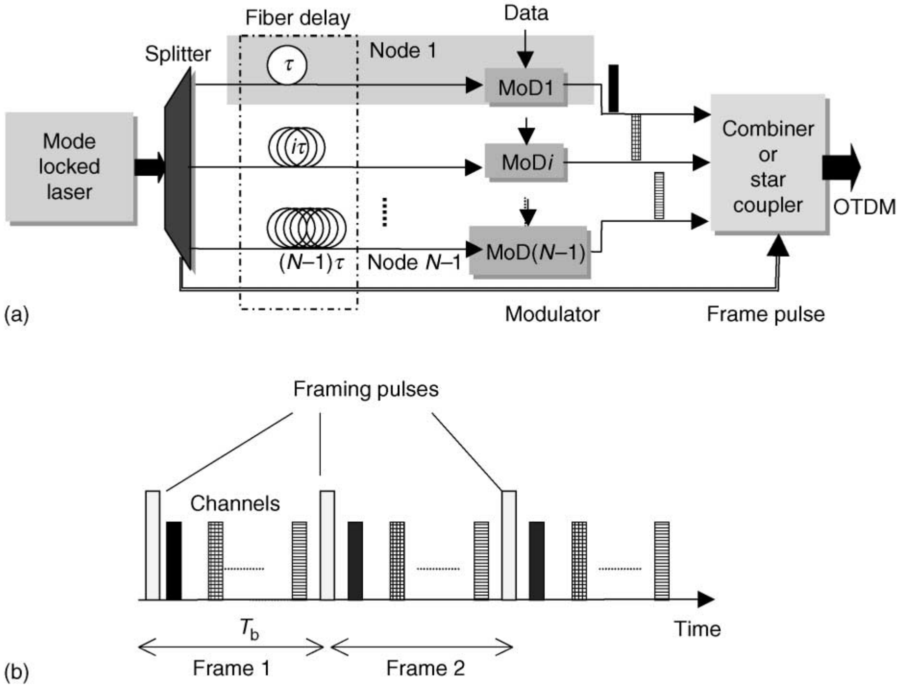
Fiber delay
Data
Splitter
τ
Node 1
MoD1
Mode locked laser
$(i\tau)$
MoD*i*
Combiner or star coupler
OTDM
$(N-1)\tau$
Node *N*–1
MoD(*N*–1)
(a)
Modulator
Frame pulse
Framing pulses
Channels
T_b
Time
Frame 1
Frame 2
(b)

Figure 3 Bit interleaved OTDM; (a) block diagram and (b) timing waveforms.

incoming bit duration and the outgoing bit duration, respectively, and k is the packet length in bits. Pulse i's location at the output is given by $(2^n - 1)(T - t) + (i - 1)\tau$. The signal at the input of the compressor is:

$$I_{\text{in}}(t) = \sum_{i=0}^{k-1} \delta(t - iT)A_i \qquad [1]$$

Where $\delta(\cdot)$ is the optical carrier pulse shape, and A_i is the ith data bit in the packet.

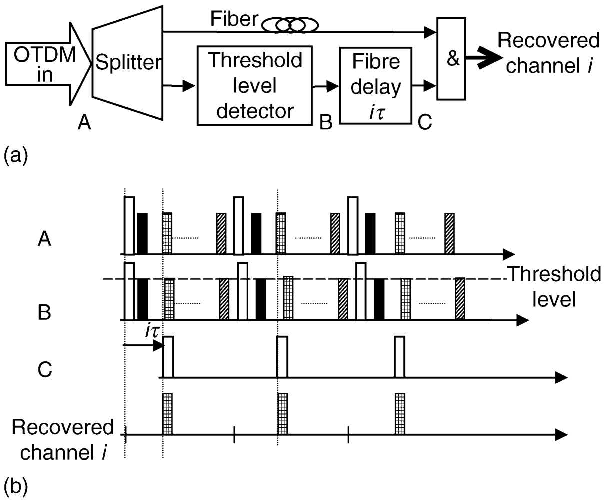
OTDM in
Splitter
Fiber
Threshold level detector
Fibre delay $i\tau$
&
Recovered channel i
A
B
C
(a)
A
B
C
Threshold level
$i\tau$
Recovered channel i
(b)

Figure 4 Bit interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

As shown in Figure 5b, each bit is split, delayed, and combined to produce the four bit packets O_i:

$$O_i(t) = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} I_i[t - i(T - \tau)] \qquad [2]$$

Note the factor of $1/(2^{n-1})$ is due to the signal splitting at the 2 × 2 3-dB coupler at the input of each stage.

The combined signal is shown in Figure 5b, and is defined as:

$$I_{\text{out}} = \sum_{i=0}^{k-1} O_i = \frac{1}{2^{n+1}} \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} I_i[t - (i+j)T + j\tau]A_i \qquad [3]$$

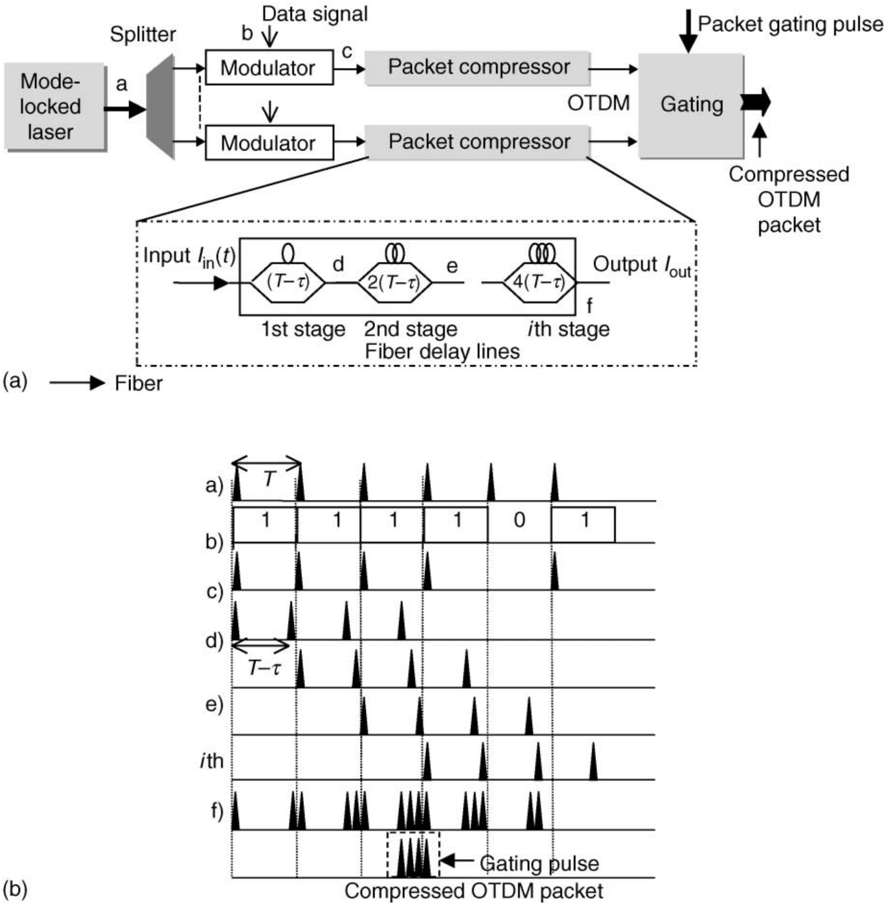
Mode-locked laser
a
Splitter
Data signal
b
c
Modulator
Modulator
Packet compressor
Packet compressor
Packet gating pulse
OTDM
Gating
Compressed OTDM packet
Input $I_{\text{in}}(t)$
$(T-\tau)$
d
$2(T-\tau)$
e
$4(T-\tau)$
f
Output I_{out}
1st stage
2nd stage
ith stage
Fiber delay lines
(a)
Fiber
a)
T
b)
1
1
1
1
0
1
c)
d)
$T-\tau$
e)
ith
f)
Gating pulse
Compressed OTDM packet
(b)

Figure 5 Packet interleaving multiplexer: (a) block diagram and (b) typical waveforms.

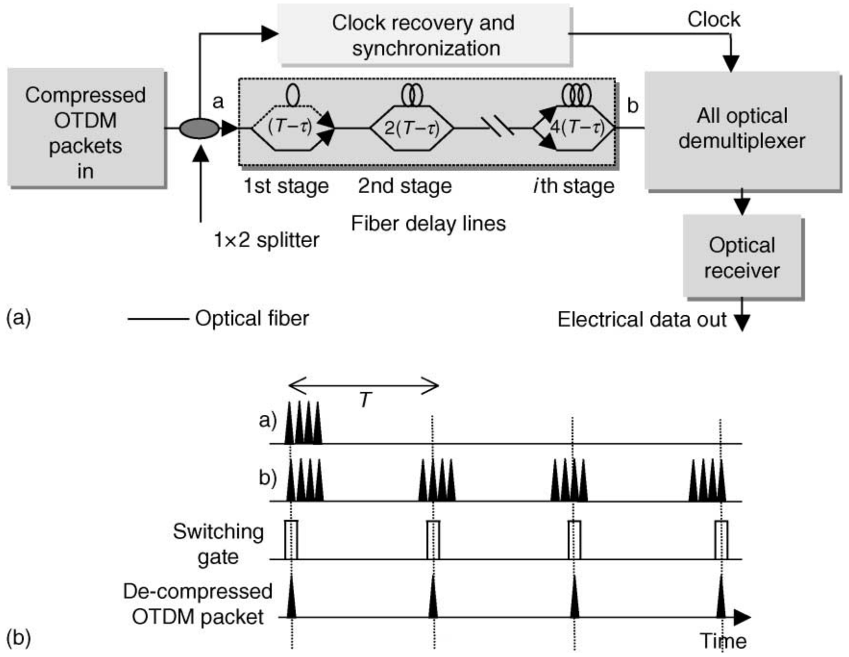

Figure 6 Packet interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

For a packet delay of $(i + j)T$ packets are being built up. For the case $(i+j) = 3$, four bits are compressed into a packet (see Figure 5b). An optical gating device is used to select the only complete usable compressed copy of the incoming packets from the unwanted copies.

At the receiving end, demultiplexing is carried out by decompressing the OTDM packet stream using the same delay line structure (see Figure 6). Multiple copies of the compressed packet are generated, and each is delayed by $(T - \tau)$. An optical demultiplexer placed at the output of the delay lines generates a sequence of switching window at the rate of $1/T$. By positioning the switching pulses at the appropriate position, a series of adjacent packet (channel) bits from each of the copied compressed packets is extracted (see Figure 6b). The output of the demultiplexer is the decompressed packet signal, which can now be processed at a much slower rate using electronic circuitry.

Components of an OTDM System

Optical Sources

In an ultrahigh speed OTDM system, it is essential that the optical sources are capable of generating transform-limited subnanosecond pulses having low duty cycles, tuneability, and a controllable repetition rate for synchronization. A number of suitable light sources are: gain-switch distribute feedback laser (DFB), active mode locked lasers (MLL) (capable of generating repetitive optical pulses), and harmonic mode-locking Erbium-doped fiber (EDF) lasers. Alternative techniques include super-continuum pulse generation in a dispersion shifted fiber with EDF pumping, adiabatic soliton compression using dispersion-flattened dispersion decreasing fiber, and pedestal reduction of compressed pulses using a dispersion-imbalanced nonlinear optical loop mirror. As shown in Figure 1, the laser light source is power split to form the pulse source for each channel. These are subsequently modulated with an electrical data signal (external modulation). External modulation is preferred in an OTDM system as it can achieve narrow carrier linewidth, thus reducing the timing jitter of the transmitted pulse. For ultrahigh bit rate OTDM systems, the optical pulses emerging from the external modulators may also need compressing. One option is to frequency-chirp the pulses and pass them through an anomalous dispersive medium. Using this approach, it is essential that the frequency chirp be linear throughout the duration of the pulse, in addition the midpoint of the linear-frequency chirp should coincide with the center of the pulse. When a frequency-chirped pulse passes through a dispersive medium, different parts of the pulse travel at different speeds, due to a temporal variation of frequency. If the trailing edge travels faster than the leading edge, the result would be pulse compression. An optical fiber cable or a semiconductor laser amplifier (SLA) can be used to create the frequency chirp.

Multiplexers

Multiplexing of the pulses generated by the optical sources can be implemented either passively or actively. The former method is commonly implemented using a mono-mode optical fiber. This method has the advantage of being simple and cost-effective. The latter method uses devices more complex in nature, for example, electro-optic sampling switches, semiconductor optical amplifiers, and integrated optics.

An integrated active multiplexer can be made by integrating semiconductor laser amplifiers (SOA) into a hybrid planar light-wave circuit (PLC). If the optical path length is kept short then temperature control is made easier. An alternative to this approach is to use lithium niobate (LiNb) technology. Periodically poled LiNb (PPLN) based OTDM multiplexers offer compact size and low noise (due to the absence of amplifier spontaneous emission noise and pattern effect, which is a feature of SOA based devices). Figure 7 is a schematic of a PLC-based OTDM multiplexer composed of two PLCs (R and L), composed of 1 × 8 and 8 × 1 couplers, eight 2 × 1 couplers, and eight different path length PPLN waveguides. The input clock pulse is split into eight by the 1 × 8 coupler in PLC-L, and are combined with the modulated optical pulse trains using the 2 × 1 couplers. The outputs of the 2 × 1 couplers are then passed through a PPLN waveguide to generate a return-to-zero (RZ) optically modulated signal. These are then combined, using the 8 × 1 coupler in PLC-R. When the path-length difference between the waveguides is set to one time slot (equal to 1/8 of the base data rate), then the output of the 8 × 1 coupler is the required high-bit rate OTDM pulse.

Demultiplexers

In contrast to multiplexing, demultiplexing must be performed as an active function. It can be implemented electro-optically or optically. The former method needs to complete demultiplexing of all channels in order to extract a single channel (see Figure 8). The demultiplexer in Figure 8 uses two LiNb Mach–Zehender (M–Z) modulators in tandem. The first and second modulators are driven with a sinusoidal signal of amplitudes $2V_\pi$ and V_π, respectively, to down-covert the N-bit rate to $N/2$ and $N/4$, respectively. Channels can be selected by changing either the DC-bias V + to the M–Zs, or the electrical phase delay. At ultrahigh speed implementation of an electro-optics demultiplexer becomes increasingly difficult due to the higher drive voltage requirement by the M–Z. An alternative is to use all optical demultiplexing based on the nonlinear effect in a fiber and optical active devices offering switching resolution in the order of picoseconds. There are a number of methods available to implement all optical demultiplexing. The most popular methods that use fast phase modulation of an optical signal are based on M–Z and Sagnac interferometers. Four-wave mixing is another popular method.

Mach–Zehnder (M–Z) Interferometers

The key to interferometric switching is the selection of an appropriate material for the phase modulation. Semiconductor materials, often in the form of an SLA, are suitable devices for this purpose.

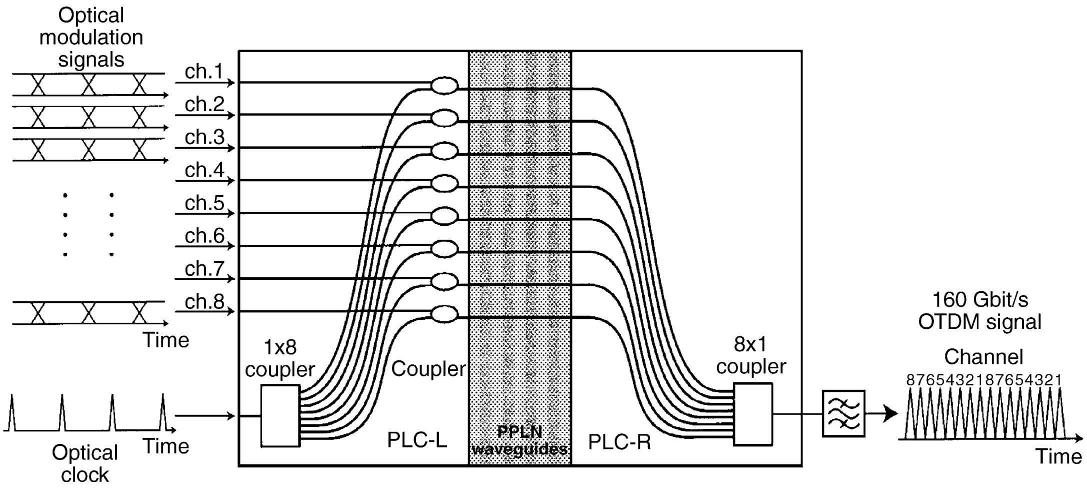

Figure 7 Configuration of a PLC-OTDM-MUX. Reproduced with permission from Ohara T, Takara H and Shake I, *et al.* (2003). 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304. © 2003 IEEE.

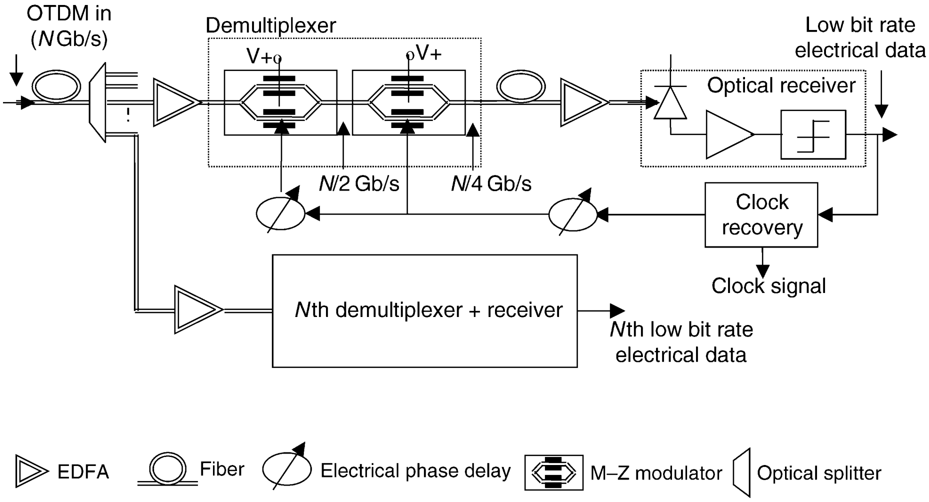

Figure 8 Electro-optics demultiplexer and receiver system block diagram.

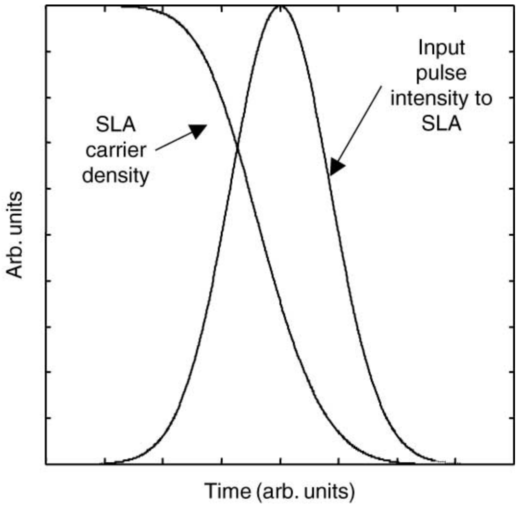

Figure 9 SLA carrier density response to input pulse.

The refractive index of an SLA is a function of the semiconductor carrier density, which can be modulated optically resulting in fast modulation. If a high-intensity pulse is input to an SLA the carrier density changes nonlinearly, as shown in Figure 9. Phase modulation is affected via the material chirp index (dN/dn), which represents the gradient of the refractive index carrier density curve for the material. The phase modulation is quite strong and, in contrast to the intensity-based nonlinearity in an optical fiber, (see Kerr effect below), is a consequence of a resonant interaction between the optical signal and the material. The nonlinearity is relatively long-lived and would be expected to limit the switching speed when using semiconductors. Semiconductors, when in excitation from an optical field, tend to experience the effect quite quickly (ps) with a slow release (hundreds of picoseconds) time. Advantage can be taken of this property by allowing the slow recovery to occur during the time between channels of an OTDM signal, as this time may be of the order of hundreds of pico-seconds or more. A Mach–Zehnder configuration using two SLAs, one in each arm of the interferometer placed asymmetrically, is shown in Figure 10. An optical control pulse entering the device via port 3 initiates the nonlinearity. The transmission equation relating the input signal (port 1) to the output port is given by

$$\frac{I_{\mathrm{out}}(t)}{I_{\mathrm{in}}(t)} = 0.25\Big(G_{\mathrm{SLA1}}(t) + G_{\mathrm{SLA2}}(t) \pm 2\sqrt{G_{\mathrm{SLA1}}(t)G_{\mathrm{SLA2}}(t)}\cos\Delta\phi(t)\Big) \qquad [4]$$

where $G_{\mathrm{SLA1}}(t)$ and $G_{\mathrm{SLA2}}(t)$ refer to the gain profiles of the respective SLAs, and $\Delta\phi(t)$ is the time-dependent phase difference between them. Assuming that the gain and phase profiles of an excited amplifier are given by $G(t)$ and $\phi(t)$, respectively, then the signals passing through the upper and lower arms experience optical properties of the material given by:

$$G(t),\ G(t - T_{\mathrm{d}}),\ \phi(t) \text{ and } \phi(t - T_{\mathrm{d}}) \qquad [5]$$

where T_{d} is given by $2L_{\mathrm{d}}N_{\mathrm{SLA}}/c$, L_{d} is the distance between SLA1 and SLA2, N_{SLA} is the SLA index, and c is the speed of light in a vacuum. As the width of the $\Delta\phi(t)$ profile is dependent on the distance between the SLAs, placing them in close proximity allows high-resolution switching. Less emphasis has been placed on the SLA gain as this is considered to be less effective when phase differences of π are reached. It only remains for the gain and phase

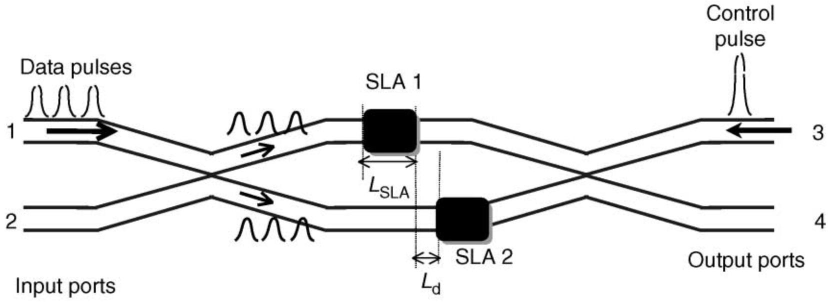

Figure 10 Asymmetric TWSLA Mach–Zehnder devices.

modulation to recover, and as indicated previously, this is allowable over a time-scale commensurate with tributary rates.

Sagnac Interferometers

There are two main types; the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD).

NOLM

In this method of switching the inherent nonlinearity of an optical fiber known as the Kerr effect is used. The phase velocity of any light beam passing through the fiber will be affected by its own intensity and the intensity of any other beams present. When the intrinsic third-order nonlinearity of silica fibers is considered via the intensity-dependent nonlinear refraction component, then the signal phase shift is

$$\Delta\phi_{\text{signal}} = \frac{2\pi}{\lambda_{\text{ds}}} n_2 L I_{\text{ds}} + 2\frac{2\pi}{\lambda_{\text{ds}}} n_2 L I_{\text{c}} \qquad [6]$$

where n_2 is the Kerr coefficient, I_{ds} is the intensity of the data signal to be switched, I_{c} is the intensity of the control signal used to switch the data signal, λ_{ds} is the data signal wavelength, and L is the fiber length. The optical loop mirror consists of a long length of mono-mode fiber formed into a fiber coupler at its free ends (see Figure 11). The input to the loop comprises the high-frequency data stream plus a control pulse at the frame rate. The data split at the coupler and propagate around the loop in contra directions (clockwise E_{CW} and counter-clockwise E_{CCW}) recombining back at the coupler. In the absence of a control pulse, the pulse exits via port 1. If a particular pulse in the loop (in this example E_{CW}) is straddled by the control pulse (see Figure 12), then that pulse experiences cross-phase modulation, according to the second term on the right-hand side of eqn [6], and undergoes a phase change relative to E_{CW}. The difference in the phase between E_{CW} and E_{CCW} causes the pulse to exit via port 2. The phase shift profile experienced by the co-propagating pulse is

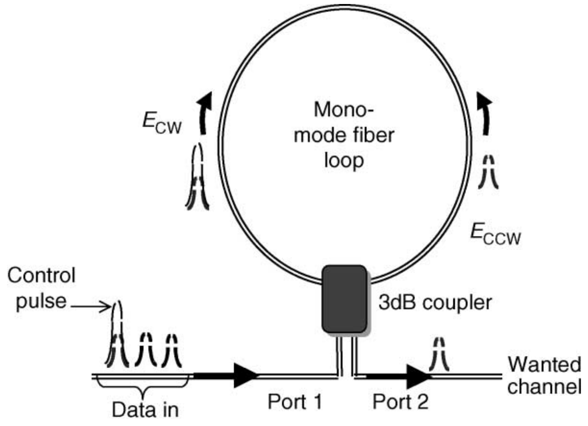

Figure 11 Nonlinear optical loop mirror.

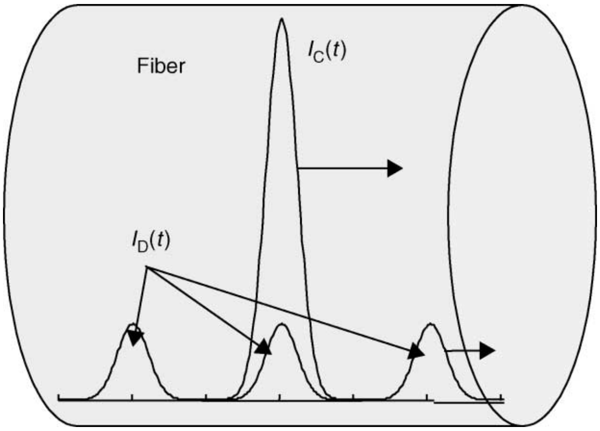

Figure 12 Control and data pulses propagation in the fiber.

$$\Delta\phi(t) = \frac{4\pi}{\lambda_{\text{ds}}} n_2 \int_0^L I_{\text{c}}(t)\,\mathrm{d}x \qquad [7]$$

Assuming unity gain around the loop, the transmittance of the NOLM is

$$\begin{aligned} T_{\mathrm{x}} &= 1 - \cos^2\left(\frac{\Delta\phi}{2}\right) = \Delta\phi(t) \\ &= 1 - \cos^2\left(\frac{2\pi}{\lambda_{\mathrm{ds}}} n_2 \int_0^L I_{\mathrm{C}}(t)\,\mathrm{d}x\right) \end{aligned} \qquad [8]$$

As it stands, the switching resolution is determined by the width of the control pulse; however, it is possible to allow the signal and control to 'walk off' each other, allowing the window width to be increased by an amount determined by the 'walk off'. The phase shift is now

$$\Delta\phi(t) = 2\frac{2\pi}{\lambda_{\mathrm{ds}}} n_2 \int_0^L I_{\mathrm{C}}(t - T_{\mathrm{w}}x)\mathrm{d}x \qquad [9]$$

where the parameter T_{w}, is the walk off time per unit length between control and signal pulses. The increased window width is accompanied by a lower peak transmission (see Figure 13).

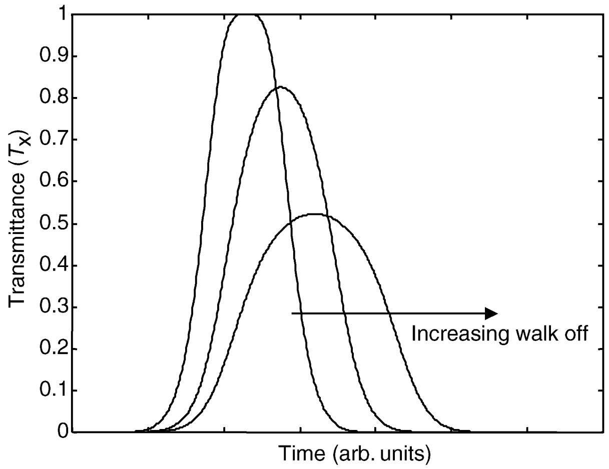

Figure 13 Transmittance profiles with the walk-off time as a parameter.

TOAD

The TOAD uses a loop mirror architecture that incorporates an SLA and only needs a short length of fiber loop (see Figure 14a). The SLA carrier density is modulated by a high-intensity control pulse, as in the M–Z-based demultiplexer. The operation of the TOAD is similar to the NOLM, where the loop transmittance is determined by the phase difference between CW and CCW traveling pulses. Strictly speaking, the gain of the SLA must also be taken into account; however, without any loss of generality the effect is adequately described by considering only the phase property. Figure 14a shows the timing diagram associated with a TOAD demultiplexer. The control pulse, shown in Figure 14b(1), is incident at the SLA at a time t_1; the phase profiles for the CCW and CCW data pulses are shown in Figures 14b(2) and (3), respectively. The resulting transmission window is shown in Figure 14b(4). The window width is given by $t_{\mathrm{asy}} = 2\Delta x/c_{\mathrm{f}}$, where Δx is the SLA offset from the loop center and c_{f} is the speed of light in the fiber. As in NOLM, if the phase difference is of sufficient magnitude, then data can be switched to port 2. The switch definition is, in principle, determined by how close the SLA is placed to the loop center when the asymmetry is relatively large. However, for small asymmetries the switching window is asymmetric, which is due to the CW and CCW gain profiles being different (see Figure 15). Assuming the phase modulation dominates the transmission then the normalized transmission for a small asymmetry loop is as depicted in Figure 16. The gain response (not shown) would have a similar shape and temporal position as the phase response.

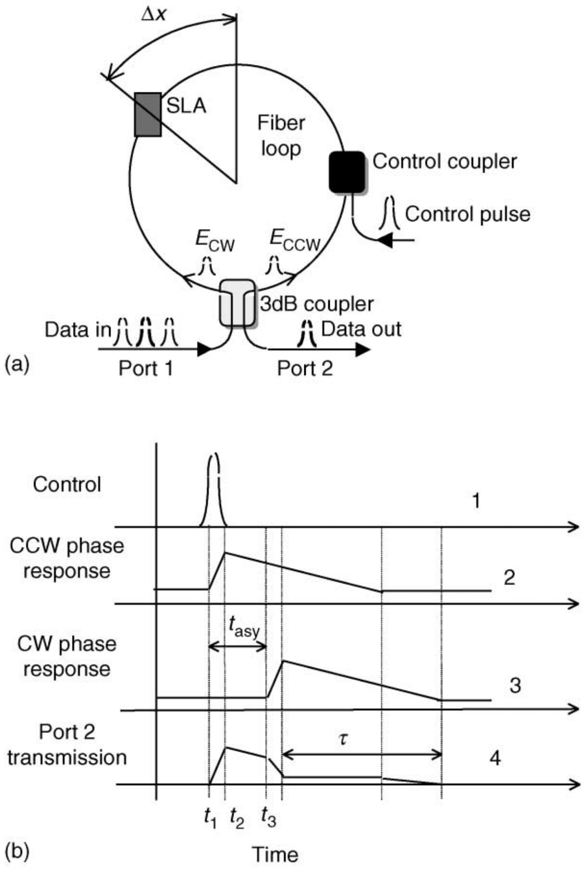

Figure 14 TOAD: (a) architecture and (b) timing diagrams.

Four-Wave Mixing (FWM)

FWM demultiplexing uses a concept whereby two optical signals of different wavelengths are mixed together in a nonlinear medium to produce harmonic components. The nonlinearity in this case arises from the third-order nonlinear susceptibility $\chi^{(3)}$ of an optical fiber, such that the polarization P induced on a pair of electric fields propagating through the fiber is

$$P = \varepsilon_0 \chi^{(3)} (E_{\mathrm{ds}} + E_{\mathrm{c}})^3 \qquad [10]$$

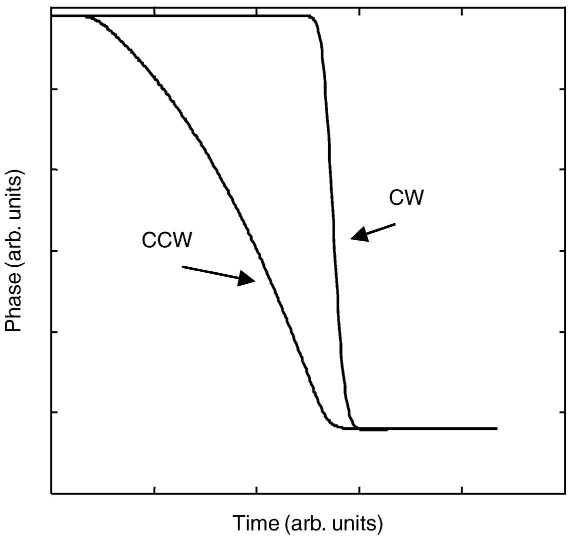

Figure 15 Phase responses for CW and CCW components of TOAD.

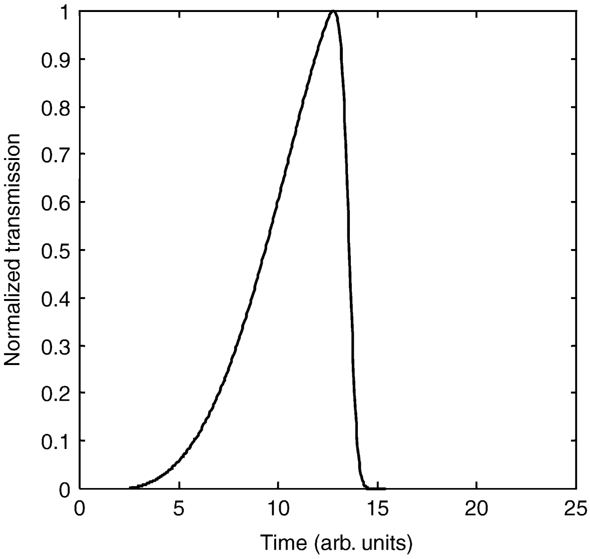

Figure 16 TOAD transmission window profile for small asymmetry loop.

where ε_0 is the vacuum permittivity of the medium, E_{ds} and E_{c} are the electric fields of data and control signals, respectively. The mixing of the two signals takes place in a long length of fiber in which the control signal propagates alongside the channel to be demultiplexed (see Figure 17). The control signal is of sufficient intensity to cause the fiber to operate in the nonlinear regime. The nonlinear relationship causes a number of frequencies to be generated, with the ones of interest having a frequency given by

$$f_{\mathrm{FWM}} = 2f_{\mathrm{c}} - f_{\mathrm{ds}} \qquad [11]$$

An optical filter is then used to demultiplex the required channel from the composite signal. The nonlinearity is detuned from the resonant frequency of the fiber glass and as such tends to be weak, requiring long lengths of fiber to give a measurable effect. The power in the demultiplexed signal, for given data and control signal wavelengths and fiber material, depends on the input power and the fiber length according to

$$P_{\mathrm{FWM}} = kP_{\mathrm{ds}}P_{\mathrm{c}}L_{\mathrm{E}}^2 \qquad [12]$$

where k is a constant, P_{ds} is the signal launch power, P_{c} is the control signal launch power, and L_{E} is the fiber effective length. FWM is essentially an inefficient method as power is wasted in the unused frequency components of the four-wave signal. More in line with integrated structures, the nonlinear properties of a semiconductor laser amplifier can be used. Here a relatively high-power optical signal is input to the SLA (see Figure 18). This enables saturation and operation in the nonlinear regime

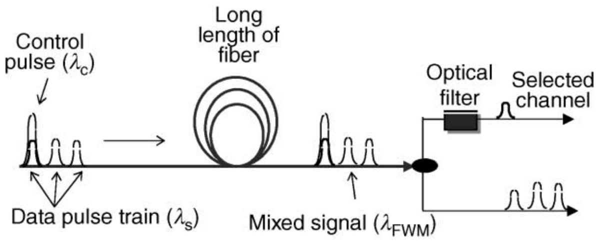
Control pulse (λ_c)
Long length of fiber
Optical filter
Selected channel
Data pulse train (λ_s)
Mixed signal (λ_{FWM})

Figure 17 Block diagram of FWM demultiplexer.

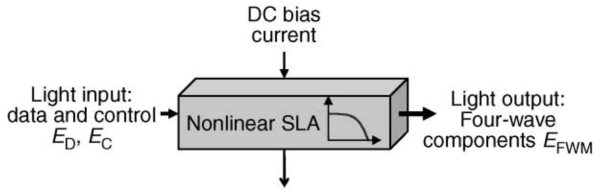
DC bias current
Light input: data and control E_D, E_C
Nonlinear SLA
Light output: Four-wave components E_{FWM}

Figure 18 Four-wave mixing in SOA.

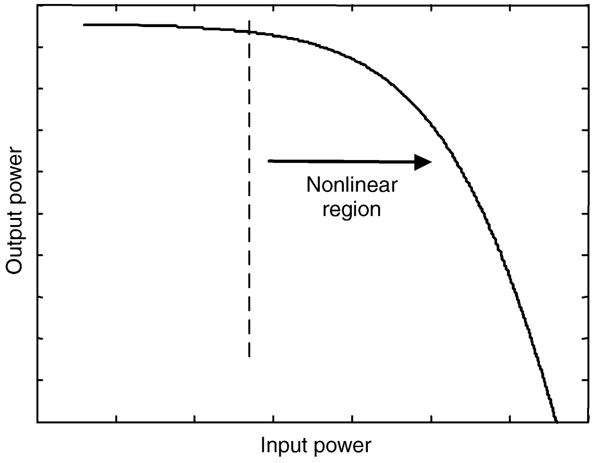

Figure 19 SLA output–input power curve.

(see Figure 19). Operating the SLA in saturation allows the nonlinear effects to produce the FWM components as in the fiber method.

Clock and Data Synchronization in OTDM

In common with electronically based TDM systems, clock recovery is fundamental to the recovery of data in ultrahigh-speed OTDM systems. Two main methods are proposed: (i) clock signal transmitted with the OTDM signal (i.e., multiplexed); and (ii) extraction of clock signal from the incoming OTDM pulse stream. In a packet-based system, synchronization between the clock and the data packet is normally achieved by sending a synch pulse with each packet. However, techniques based on optical-phased locked loops are proving popular and remove the need for a separate clocking signal.

Clock Multiplexing

(i) *Space division multiplexing*: This is conceptually the simplest to implement, where the clock signal is carried on a separate fiber from the data. However, it is susceptible to any differential delay between the different paths taken by clock and data due to temperature variation. It is difficult to justify in systems where the installed fiber base is at a premium.

(ii) *Wavelength division multiplexing*: Here different wavelengths are allocated to the clock, and payload. It is only really practical for predetermined path lengths between nodes in single-hop networks such as point–point links or broadcast-and-select star networks. It also suffers from random delays between the clock and the payload, which is problematic in an asynchronous packet switched based network, where the optical path length a packet may take is nondeterministic.

(iii) *Orthogonal polarization*: This is suitable for small links, where separate polarizations are used for the clock and data. However, in large networks it is quite difficult to maintain the polarization throughout the transmission link due to polarization mode dispersion and other nonlinear effects.

(iv) *Intensity division multiplexing*: This uses higher-intensity optical clock pulses to differentiate it from the data pulses as discussed above. However, in long-distance transmission links, it is difficult to maintain both the clock intensity and its position, due to the fiber nonlinearity.

(v) *Time division multiplexing*: In this scheme a single clock pulse, which has the same wavelength, polarization, and amplitude as the payload pulses, is separated in time, usually ahead of the data pulses.

Synchronization – Optical Phased Locked Loops (PLL)

The PLL is a common technique used for clock recovery in electronic TDM systems. However, the speed of conventional electronic PLLs tends to be limited by the response of the phase comparators used. There are a number of approaches based on opto-electronic PLL. Opto-electronic PLLs based on four-wave mixing in a traveling wave laser amplifier are complex and can suffer from frequency modulation in the recovered clock. However, others based on balanced photo-detectors, result in low timing jitter and good phase stability. In contrast, a number of all optical methods clock recovery scheme exist, one technique based on the TOAD (see above) is depicted in Figure 20. The high-speed data stream enters the TOAD at a rate $n \times R$ where n is the number of channels per OTDM frame, and R is the frame rate. A pulse generator, such as a mode locked fiber laser (MLFL) clocked by a local oscillator (LO), is used as the TOAD control input (MLFL-C). The OTDM data is switched by the TOAD at a frequency of say, $R + \Delta f$ Hz. Thus, the switching window samples data pulses at a rate higher or lower than $n \times R$ Hz and uses this signal for cross correlation in the PLL unit. The output of the phase comparator is used to regulate a voltage controlled oscillator (VCO)

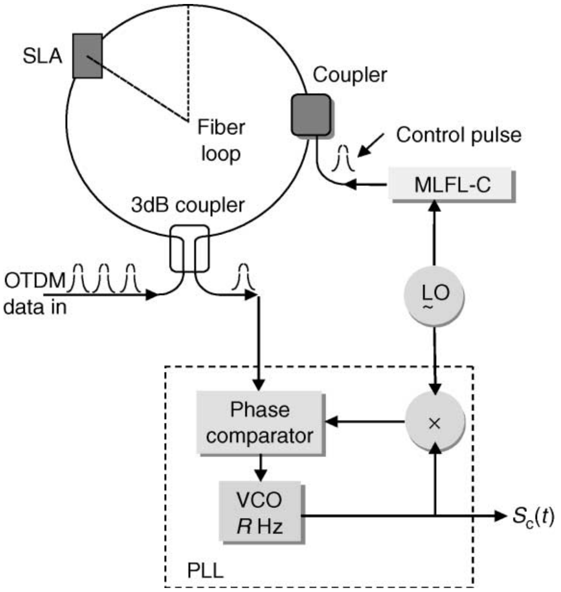
SLA
Fiber loop
Coupler
Control pulse
MLFL-C
3dB coupler
OTDM data in
LO ~
Phase comparator
×
VCO R Hz
$S_c(t)$
PLL

Figure 20 Ultrafast clock recovery using TOAD.

running at R Hz, which in turn feeds the control signal. The PLL circuit locks into the clock frequency, generating the clock signal $S_c(t)$.

OTDM Bit-Error Rate (BER) Performance

Figure 21 shows a typical block diagram of a high-speed optical receiver for an OTDM system, composed of a NOLM or a TOAD demultiplexer, an optical pre-amplifier, an optical bandpass filter, and a conventional optical receiver using a PIN photodiode. Due to the crosstalk introduced in the demultiplexing process, the demultiplexed optical signal contains not only the target channel but may also contain nontarget channels with reduced amplitude (see the inset in Figure 21). The intensity of the demultiplexed optical signal is boosted by the optical pre-amplifier (EDFA). Amplified spontaneous emission (ASE) from the optical preamplifier adds a wide spectrum of optical fields onto the demultiplexed optical signal. Although an optical bandpass filter (BPF) can reduce the ASE, it still remains one of the major noise sources. The function of the optical filter is to reduce the excess ASE to within the range of the signal spectrum. The PIN photodiode converts the received optical power into an equivalent electrical current, which is then converted to a voltage signal by an electrical amplifier. Finally, the output voltage is sampled periodically by a decision circuit for estimating the correct state (mark/space) of each bit.

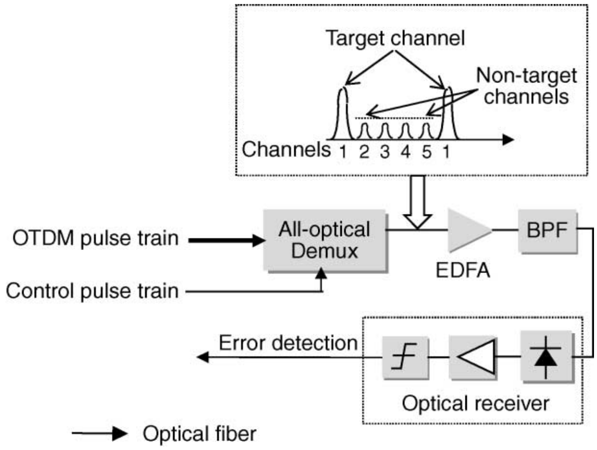
Target channel
Non-target channels
Channels 1 2 3 4 5 1
OTDM pulse train
Control pulse train
All-optical Demux
EDFA
BPF
Error detection
Optical receiver
Optical fiber

Figure 21 Block diagram of OTDM receiver.

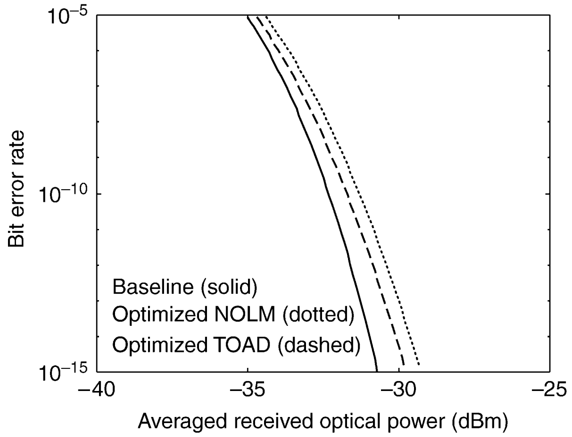

Figure 22 BER versus the average received optical power for 100 Gb/s (10 channels) OTDM system.

The BER performance of an OTDM system deteriorates because of the noise and crosstalk introduced by the demultiplexer and is:

$$\mathrm{BER} = \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-Q^2/2\right)}{Q} \qquad [13]$$

where Q is defined as:

$$Q = \frac{\bar{I}_\mathrm{m} - \bar{I}_\mathrm{s}}{\sqrt{\sigma^2_{\mathrm{RIN,m}} + \sigma^2_{\mathrm{RIN,s}} + \sigma^2_{\mathrm{amp,m}} + \sigma^2_{\mathrm{amp,s}} + \sigma^2_{\mathrm{rec,m}} + \sigma^2_{\mathrm{rec,s}}}} \qquad [14]$$

and $\bar{I}_\mathrm{m}$ and $\bar{I}_\mathrm{s}$ are the average photocurrents for a mark and a space, respectively. $\sigma_{x,i}$ are the variances of the relative intensity noise (RIN), further optical pre-amplifier and receiver for a mark and a space.

For 100 Gb/s (10 channels) OTDM system, the BER against average received optical power for optimized NOLM and TOAD demultiplexers, is shown in Figure 22.

List of Units and Nomenclature

A_i	ith data bit in a packet
c	Speed of light in a vacuum
c_{f}	Speed of light within the fiber
E_{c}	Electric fields of control signals
E_{ds}	Electric fields of data signals
f_{FWM}	Frequency due to four-wave mixing
$G_{\mathrm{SLA}}(t)$	Gain profiles of SLA
I_{c}	Intensity of the control signal
I_{D}	Intensity of the data signal
$\bar{I}_{\mathrm{m}}$	Average photo current for a mark
$\bar{I}_{\mathrm{s}}$	Average photo current for a space
k	Packet length in bits
L	Fiber length
L_{d}	Distance between two SLAs
L_{E}	Fiber effective length
n	Number of stages
n_2	Kerr coefficient
N_{SLA}	SLA index
P	Polarization
P_{c}	Control signal launch power
P_{ds}	Data signal launch power
t_{asy}	The window width is given by $t_{\mathrm{asy}} = 2\Delta x/c_f$
T	Incoming bit interval
T_{w}	Walk off time per unit length between control and data pulses
T_{x}	Transmittance of the NOLM
$\delta(\cdot)$	Optical carrier pulse shape
$\Delta\phi(t)$	Time-dependent phase difference
Δx	SLA offset from the fiber loop center
ε_0	Vacuum permittivity of the medium
λ_{s}	Data signal wavelength
$\sigma_{\mathrm{amp,m}}$	Variances for optical pre-amplifier for a mark
$\sigma_{\mathrm{amp,s}}$	Variances for optical pre-amplifier for a space
$\sigma_{\mathrm{rec,m}}$	Variances for receiver for a mark
$\sigma_{\mathrm{rec,s}}$	Variances for receiver for a space
$\sigma_{\mathrm{RIN,m}}$	Variances of RIN for a mark
$\sigma_{\mathrm{RIN,s}}$	Variances of RIN for a space
τ	Outgoing bit interval
$\phi(t)$	Phase profile of SLA
$\chi^{(3)}$	Third-order non-linear susceptibility of an optical fiber

See also

Interferometry: Overview. **Nonlinear Optics, Basics**: Four-Wave Mixing. **Optical Communication Systems**: Optical Time Division Multiplexing; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (2002) *Fiber-Optic Communication Systems*, 3rd edn. New York: J Wiley and Sons Inc.

Chang K (ed.) (2003) *Handbook of Optical Components and Engineering*, 2nd edn. New York: Wiley and Sons Inc.

De Marchis G and Sabella R (1999) *Optical Networks Design and Modelling*. Dordrecht: Kluwer Academic.

Hamilton SA, Robinson BS, Murphy TE, Savage SJ and Ippen EP (2002) 100 Gbit/s optical time-division multiplexed networks. *Journal of Lightwave Technology* 20(12): 2086–2100.

Jhon YM, Ki HJ and Kim SH (2003) Clock recovery from 40 gbps optical signal with optical phase-locked loop based on a terahertz optical asymmetric demultiplexer. *Optics Communications* 220: 315–319.

Kamatani O and Kawanishi S (1996) Ultrahigh-speed clock recovery with phase locked loop based on four wave mixing in a traveling-wave laser diode maplifier. *Journal of Lightwave Technology* 14: 1757–1767.

Nakamura S, Ueno Y and Tajima K (2001) Femtosecond switching with semiconductor-optical-amplifier-based symmetric-Mach–Zehnder-type all-optical switch. *Applied Physics Letters* 78(25): 3929–3931.

Ohara T, Takara H, Shake I, *et al.* (2003) 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304.

Ramaswami R and Sivarajan KN (2002) *Optical Network; a Practical Perspective*. Morgan Kaufman.

Reman E (2001) Trends and evolution of optical networks and technology. *Alcatel Telecommun. Rev. 3rd Quarter* 173–176.

Sabella R and Lugli P (1999) *High Speed Optical Communications*. Dordrecht: Kluwer Academic.

Seo SW, Bergman K and Prucnal PR (1996) Transparent optical networks with time division multiplexing. *Journal of Lightwave Technology* 14(5): 1039–1051.

Sokolloff JO, Prucnal PR, Glesk I and Kane M (1993) A terahertz optical asymmetric demultiplexer (TOAD). *IEEE Photonics Technology Letters* 5(7): 787–790.

Stavdas A (ed.) (2001) *New Trends in Optical Network Design and Modelling*. Dordrecht: Kluwer Academic.

Ueno Y, Takahashi M, Nakamura S, Suzuki K, Shimizu T, Furukawa A, Tamanuki T, Mori K, Ae S, Sasaki T and Tajima K (2003) Control scheme for optimizing the interferometer phase bias in the symmetric-Mach–Zehnder all-optical switch (invited paper). Joint special issue on recent progress in optoelectronics and communications. *IEICE Trans. Electron.* E86-C(5): 731–740.

Coherent Analog Optical Processors

H H Arsenault, S Roy and D Lefebvre, University of Laval, Quebec, PQ, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Linear filtering is a useful tool commonly used in image processing. The correlation operation can be defined as:

$$f(x,y)*h(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)h(x-x',y-y')\mathrm{d}x'\mathrm{d}y' \qquad [1]$$

where $*$ denotes the correlation. The operation of correlation and the Fourier transform

$$G(\mu,\nu)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)\mathrm{e}^{-j2\pi(\mu x+\nu y)}\mathrm{d}x\mathrm{d}y \qquad [2]$$

are intimately related through the correlation theorem of Fourier theory, which states that the correlation between two functions $f(x,y)$ and $h(x,y)$ is equal to the inverse Fourier transform of the product of the Fourier transforms:

$$f(x,y)*h(x,y)=FT^{-1}[F(\mu,\nu)H^{*}(\mu,\nu)] \qquad [3]$$

where $F(\mu,\nu)$ and $H(\mu,\nu)$ are the Fourier transforms of $f(x,y)$ and of $h(x,y)$ respectively, and where H^* designates the complex conjugate of H. The ability to carry out and display Fourier transforms in two dimensions, by means of optical systems, is at the root of optical information processing.

One of the main advantages of correlation is that the operations do not require segmentation of the regions of interest. This means that the locations of objects do not have to be segmented from the scene and that the detection is shift-invariant, that is the correlation peaks appear at locations in the scene corresponding to locations of the targets. These advantages are clearer for detection in uncontrolled conditions, for example for detecting vehicles in natural scenes. But they are also useful in security applications, such as identity verification or for detecting defects in assembly lines.

The Optical Fourier Transform

The Fourier transform of eqn [2] is widely used in the field of information processing. The inverse Fourier transform is simply defined as:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}G(\mu,\nu)\mathrm{e}^{i2\pi(\mu x+\nu y)}\mathrm{d}\mu\mathrm{d}\nu \qquad [4]$$

We will now discuss how to perform Fourier transforms by means of four different optical setups.

FT Without a Lens

If we illuminate a transparency $U(x_0,y_0)$ with a coherent plane wave, the diffraction pattern we observe at a point d from the input plane is given by the following relation, known as the Fresnel approximation for diffraction (Figure 1):

$$U(x_1,y_1)=\frac{\exp\left(i\frac{2\pi d}{\lambda}\right)}{j\lambda d}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}U(x_0,y_0)\times\exp\left\{i\frac{\pi}{\lambda d}[(x_0-x_1)^2+(y_0-y_1)^2]\right\}\mathrm{d}x_0\mathrm{d}y_0 \qquad [5]$$

where λ is the wavelength of the light and d is the distance between the diffracting screen and the observed pattern.

In order to use the Fresnel approximation we need only to satisfy the geometrical optics paraxial conditions. If we add an additional approximation, we can further simplify this equation to yield the Fourier transform of $U(x_0,y_0)$. The extra condition is known as the Fraunhofer or far-field approximation, which is only satisfied when the diffracted image is observed far away from the diffracting screen or when the extent of $U(x_0,y_0)$ is very small. Mathematically this condition is expressed by:

$$\frac{\pi}{\lambda d}(x_0^2+y_0^2)\ll 1 \qquad [6]$$

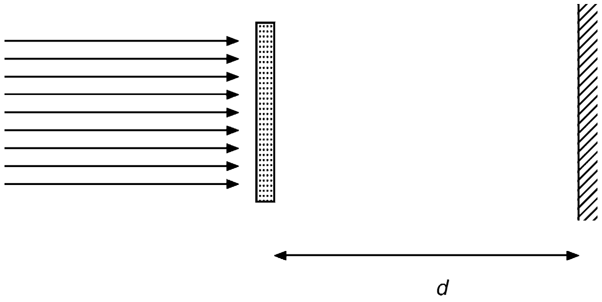

Figure 1 Fresnel diffraction without a lens.

After expanding the squares in the exponentials, there will be three terms: one can go outside the integral and one will be approximately equal to unity so the diffracted pattern becomes

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi d}{\lambda}\right)}{i\lambda d} \times \exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda d}(x_0 x_1 + y_0 y_1)\right] \mathrm{d}x_0 \mathrm{d}y_0 \qquad [7]$$

This is the Fourier transform of $U(x_0, y_0)$, scaled by a factor λd, with an extra quadratic phase factor

$$\exp\left[\frac{i\pi}{\lambda d}(x_1^2 + y_1^2)\right]$$

The factor $\exp(i2\pi d/\lambda)/i$ is usually of no importance since it is a phase factor distributed across the output plane. In some applications we only want to know the power spectrum (squared modulus) of the Fourier transform. In such cases the quadratic phase factor disappears because the squared modulus of any complex exponential is equal to 1. Performing a Fourier transform without a lens is possible within the Fraunhofer approximation. But if it is required to obtain the Fourier transform of an image which extends about 1 cm from the optical axis using a coherent He–Ne laser ($\lambda = 632.8$ nm), the condition to be satisfied is $d \gg 496$ m: the required observation distance d is at least a few kilometers. This kind of optical setup is usually impractical, but can fortunately be alleviated by means of lenses.

Object on the Lens

Instead of having only the transparency in the input plane, we place a lens immediately after it as shown in Figure 2. For mathematical simplicity the lens and input plane are superimposed (more realistic cases are considered below). The lens acts as a phase factor

$$\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right]$$

where f is the focal length of the lens. Now in the input plane the object $U(x_0, y_0)$ is multiplied by the phase factor of the lens:

$$U'(x_0, y_0) = U(x_0, y_0)\exp\left[-i\frac{\pi}{\lambda f}(x_0^2 + y_0^2)\right] \qquad [8]$$

Inserting this expression into the integral formula for Fresnel diffraction yields the diffraction pattern observed in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left(i\frac{2\pi f}{\lambda}\right)}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_0 x_1 + y_0 y_1)\right] \mathrm{d}x_0 \mathrm{d}y_0 \qquad [9]$$

The lens exactly compensates for the other phase factor that was in the integral. This expression for the Fraunhofer approximation is almost the same as before, but the focal length f of the lens replaces the distance d. So by placing a lens immediately after the input plane we can have a much more compact setup since lenses are available with a wide range of focal lengths, usually from a few centimeters to a few meters. Another way to look at this is to consider that the effect of a lens is to bring the pattern at infinity to the focal plane of the lens – except for the phase factor.

Object Before the Lens

Another setup used to perform the Fourier transform is shown in Figure 3, where the lens is placed somewhere between the input and the output planes. Using the Fresnel formula for diffraction it can be

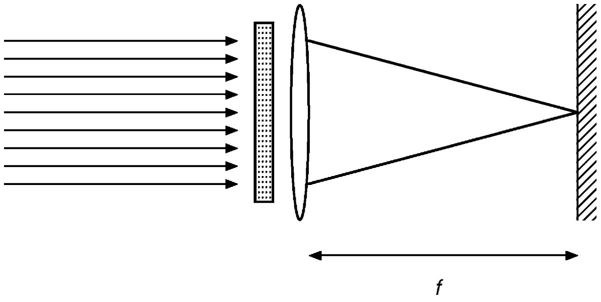

Figure 2 Diffraction with the object on the lens.

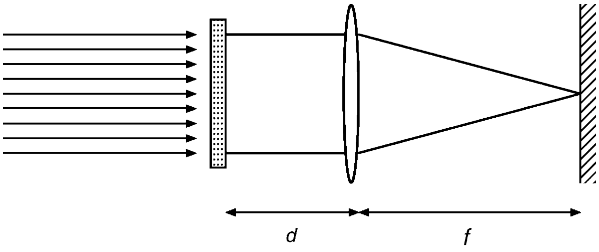

Figure 3 Diffraction with the object before the lens.

shown that the diffraction pattern in the focal plane is:

$$U(x_1, y_1) = \frac{1}{i\lambda f} \times \exp\left[i\frac{\pi}{\lambda f}\left(1 - \frac{d}{f}\right)(x_1^2 + y_1^2)\right] \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[- i\frac{2\pi}{\lambda f}(x_0x_1 + y_0y_1)\right] \mathrm{d}x_0\mathrm{d}y_0 \quad [10]$$

where d is the distance between the input plane and the lens. If we set d equal to the focal length f of the lens, the quadratic phase factor in front in the integral vanishes and we have an exact Fourier transform. This setup with the object and filter in the front and back focal planes of the lens respectively is the one most frequently seen in scientific articles and is known as the *f*-*f* Fourier transform system.

Object After the Lens

Still another configuration is the one shown in Figure 4, where the lens is placed at a distance d before the input plane. For this configuration the Fresnel diffraction yields the following pattern in the focal plane of the lens:

$$U(x_1, y_1) = \frac{\exp\left[i\frac{\pi}{\lambda d}(x_1^2 + y_1^2)\right]}{i\lambda d} \frac{f}{d} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x_0, y_0) \times \exp\left[- i\frac{2\pi}{\lambda d}(x_0x_1 + y_0y_1)\right] \mathrm{d}x_0\mathrm{d}y_0 \quad [11]$$

Once again the Fourier transform of $U(x_0, y_0)$ is multiplied by a quadratic phase factor. It is interesting to note that, in this case, the Fourier transform is scaled by a factor λd. This can be used to control the scale of the Fourier transform by changing the distance between the lens and the object.

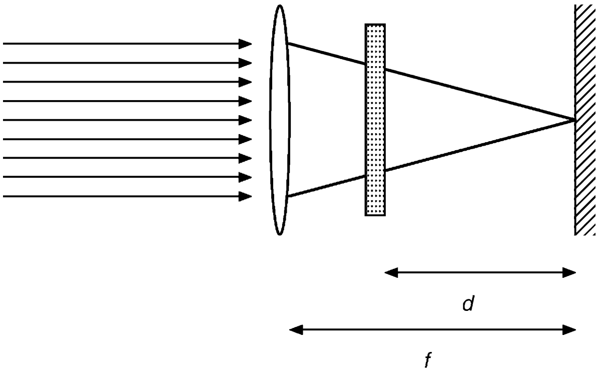

Figure 4 Diffraction with the object after the lens.

Note that although two of the three methods yield Fourier transforms that are accompanied by a quadratic phase factor, the latter is not necessarily an impediment to correlation, contrary to popular belief. Indeed all three configurations have been used for correlation pattern recognition, and each configuration has its advantages.

Optical Correlation

The last section showed how to obtain Fourier transforms by means of an optical setup. The proposed optical setups cannot perform the inverse Fourier transform. Because the Fourier transform and its inverse are very similar, the Fourier transform is performed instead, yielding the mirror image of the correlation:

$$(f * h)(-x, -y) = \mathrm{FT}[F(\mu, \nu)H^*(\mu, \nu)] \quad [12]$$

The correlation can be considered as the result of three operations: a Fourier transform; a product; and an additional Fourier transform. The Fourier transform of a 2D function can be obtained by means of coherent optics. The equivalent of multiplication in optics is transmissivity, so multiplication is accomplished by sending light successively through two superimposed transparencies.

There are two main categories of optical correlators, serial correlators and joint transform correlators.

Serial Correlators

A serial correlator functions in two stages. The first stage performs the Fourier transform of the input image $f(x, y)$, which is multiplied in the Fourier plane with a transparency $H(\mu, \nu)$. The filter $H(\mu, \nu)$ can be a hologram, a spatial light modulator (SLM) or a simple transparency. A second Fourier transform displays in the output plane the correlation of the input image with the impulse response $h(x, y)$, which is the Fourier transform of the filter function $H(\mu, \nu)$.

Serial correlators directly implement the two Fourier transforms and the multiplication operation in a pipelined manner. The input plane and the correlation plane are conjugate planes of the system, that is in the absence of a filter, the output plane displays an inverted image of the input $f(x, y)$. Between those conjugate planes, at the conjugate plane of the source with respect to the first lens, is the Fourier plane where the spatial frequency filter is placed. When the input is illuminated with parallel light, the frequency plane is at the back focal plane of the first lens. A SLM can be placed in the input plane to feed the images to be correlated; a second SLM can be placed in the Fourier plane to display the filter. The filter is often calculated by a digital computer, and usually takes the form of a hologram if phase information must be used in the filter.

The 4-*f* correlator

The 4-*f* correlator is the most intuitive optical correlator system. We showed above that an *f*-*f* system can be used to optically obtain a Fourier transform (Figure 5). If $f(x, y)$ is in the front focal plane, its Fourier transform $F(\mu, \nu)$ appears at the back focal plane of the first lens. If a filter function $H^*(\mu, \nu)$ is inserted at this point, by means of a SLM or a hologram, then a second *f*-*f* system can perform a second Fourier transform to yield the correlation plane at the back focal plane of the second lens.

Convergent beam correlator

This kind of correlator consists of two conjugate systems. This time, the Fourier transform is obtained by placing the input object in a convergent beam after the first lens. The second lens is placed so that the output plane is conjugate to the input plane, and as before, the filter is placed at the point of convergence of the light after the first lens. It is important to note that this is not the focal plane of the lens, but the conjugate plane of the source of light. In order to have an imaging system between the input and output planes, the distances must satisfy the conjugate relation:

$$\frac{1}{s_0} + \frac{1}{s_i} = \frac{1}{f_2} \qquad [13]$$

where s_i and s_0 are defined in Figure 6. As previously discussed for convergent light, the Fourier transform will have an additional parabolic phase factor. This must be taken into account when constructing the filter H.

This configuration has some advantages over the 4-*f* correlator. By moving the object between the first lens and the filter, it is possible to control the scale of the Fourier transform. And because the first lens is imaging only a single point on the optical axis, only spherical aberration needs to be compensated for. In 4-*f* systems, the first lens must be corrected for all aberrations for the whole input plane, and for a given extent of the input object, the first lens must be larger than that for the convergent system in order to avoid space variance effects. Unfortunately, specially designed expensive lenses are required to correct for those aberrations, so well-corrected Fourier transform lenses for 4-*f* systems are expensive, whereas a simple doublet is sufficient to correct for on-axis spherical aberration, which is all that is required for the convergent light system. In addition, the convergent light system does not require an expensive collimator that is required in the 4-*f* system to obtain the parallel illuminating beam. Aberration correction requirements for the second lens are less stringent

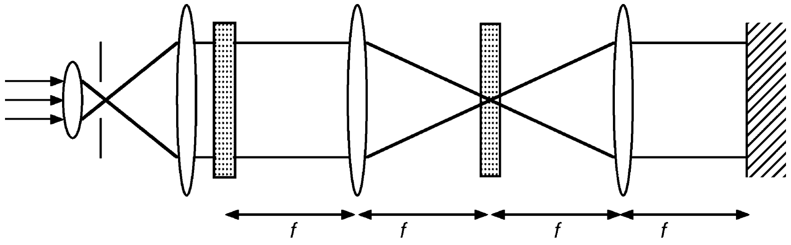

Figure 5 4-*f* correlator setup.

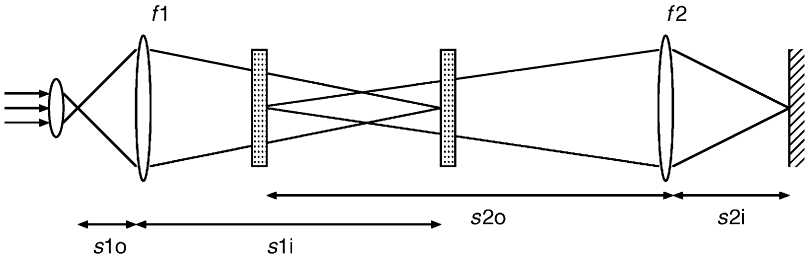

Figure 6 Convergent beam correlator setup.

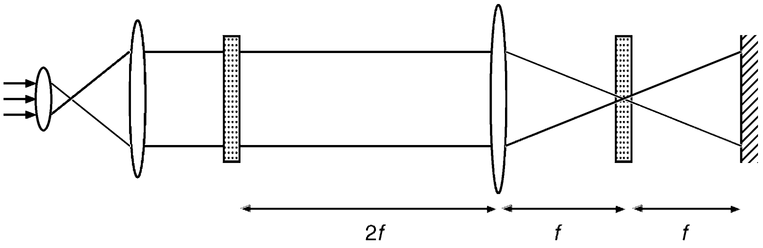

Figure 7 Single lens correlator setup.

than for the first, because in most cases, a slight deformation of the output plane can be acceptable, whereas comparable distortions in the Fourier plane can cause misregistrations between the Fourier transform and the filter that can seriously degrade the correlation.

Divergent beam correlators

Instead of placing the input object in the convergent beam, as in the previous case, another configuration that has been used places the input object in the divergent beam before the first lens. This configuration is rarely seen today, because it exacerbates the space variance and aberration problems of the 4-*f* system, and the control it allows over the scale of the Fourier transform can be achieved with fewer penalties by means of the convergent light system.

Single lens correlator

Consider an imaging system using only one lens illuminated with a parallel beam of coherent light (Figure 7). The input and output planes are located at distances $2f$ on either side of the lens – the shortest conjugate distances possible. If a filter $H(\mu, \nu)$ is placed at the focal plane of the lens on the object's Fourier transform, the correlation will appear on the output plane. This system requires satisfying impractical conditions: aberration corrections for both the Fourier transform and for the imaging operation, so it is rarely seen.

Filters

Many types of filters can be used for different applications. Table 1 lists some commonly used filters for pattern recognition. The classical matched filter yields the optimum signal to noise ratio in the presence of additive noise. The phase-only filter (POF) is frequently used because of its high discrimination capability. The inverse filter is sometimes used because it theoretically yields very sharp correlation peaks, but is very sensitive to noise, and when the spectrum $F(\mu, \nu)$ has some zeros, the realization of $1/F(\mu, \nu)$ can be impossible. A close relative of the inverse filter is the Wiener filter, which avoids the problem of the inverse filter by introducing a quantity $S(\mu, \nu)$ in the denominator of the filter function. The function $S(\mu, \nu)$ can sometimes be a constant. $H(\mu, \nu)$ is usually chosen to minimize the effect of noise or to optimize some parameter such as the peak to sidelobe ratio, or the discrimination ability of the system.

Table 1 Commonly used filters for image processing

Name	*Filter*
Matched filter	$H_{\mathrm{CMF}}(\mu, \nu) = F(\mu, \nu)$
Phase-only filter	$H_{\mathrm{POF}}(\mu, \nu) = \dfrac{F(\mu, \nu)}{\lvert F(\mu, \nu)\rvert}$
Inverse filter	$H_{\mathrm{INV}}(\mu, \nu) = \dfrac{1}{F(\mu, \nu)}$
Wiener filter	$H_{\mathrm{W}}(\mu, \nu) = \dfrac{F^{*}(\mu, \nu)}{\lvert F(\mu, \nu)\rvert^{2} + S(\mu, \nu)}$

Joint Transform Correlators

In serial transform correlators, the filter is generally a complex function. Spatial light modulators are usually conceived with only one degree of freedom for each pixel, which can be optimized for amplitude modulation or for phase modulation. In order to achieve a fully complex filter, two modulators are required, or some spatial bandwidth of a modulator is sacrificed to encode a computer-generated hologram.

The joint-transform correlator (JTC) does not require a fully complex filter. The filter is stored in the spatial domain so there is usually no need for the computation of a filter in the spectral domain. In the JTC, the correlation of a scene with a reference image is contained in the power spectrum of an input plane consisting of the scene placed side by side with the reference. The correlation terms are revealed when the Fourier transform of the power spectrum is computed. The joint transform correlation can be summarized in three steps:

1. The scene and the reference are located side by side in the input plane.
2. The Fourier transform of the input plane is computed or carried out optically by one of the methods previously described. The transform intensity is measured and constitutes the joint power spectrum. Some optional enhancements, based on spatial filtering and/or nonlinear processing on the power spectrum, can also be performed.
3. The Fourier transform of this joint power spectrum is carried out, yielding the correlation plane containing the correlation of the reference with the scene.

A mathematical analysis is required to understand the location of the correlation peaks. The side-by-side input scene $s(x, y)$ and reference image $r(x, y)$ are

$$i(x, y) = r(x - x_{\mathrm{r}}, y - y_{\mathrm{r}}) + s(x - x_{\mathrm{s}}, y - y_{\mathrm{s}}) \qquad [14]$$

The joint power spectrum $E(\mu, \nu)$ is equal to the modulus squared of the Fourier transform of the input plane:

$$\begin{aligned} I(\mu, \nu) &= R(\mu, \nu)\mathrm{e}^{-i2\pi(\mu x_{\mathrm{r}} + \nu y_{\mathrm{r}})} + S(\mu, \nu)\mathrm{e}^{-i2\pi(\mu x_{\mathrm{s}} + \nu y_{\mathrm{s}})} \\ E(\mu, \nu) &= \|I(\mu, \nu)\|^2 = \\ &\quad \|R(\mu, \nu)\|^2 + \|S(\mu, \nu)\|^2 \\ &\quad + R(\mu, \nu)S^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_{\mathrm{r}} - x_{\mathrm{s}}) + \nu(y_{\mathrm{r}} - y_{\mathrm{s}})]} \\ &\quad + R^*(\mu, \nu)S(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_{\mathrm{s}} - x_{\mathrm{r}}) + \nu(y_{\mathrm{s}} - y_{\mathrm{r}})]} \end{aligned} \qquad [15]$$

The joint power spectrum contains the Fourier transforms of the autocorrelation of the input scene, the autocorrelation of the reference, the correlation of the scene with the reference and its complex conjugate. The correlation peaks are obtained from the Fourier transform of the joint power spectrum:

$$\begin{aligned} C(x', y') = {} & R_{\mathrm{ss}}(x', y') + R_{\mathrm{rr}}(x', y') \\ & + R_{\mathrm{sr}}\big(x' - (x_{\mathrm{r}} - x_{\mathrm{s}}),\ y' - (y_{\mathrm{r}} - y_{\mathrm{s}})\big) \\ & + R_{\mathrm{sr}}\big(x' + (x_{\mathrm{r}} - x_{\mathrm{s}}),\ y' + (y_{\mathrm{r}} - y_{\mathrm{s}})\big) \end{aligned} \qquad [16]$$

where

$$R_{\mathrm{ab}}(x', y') = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(\xi, \zeta)b^*(\xi - x', \zeta - y')\mathrm{d}\xi\mathrm{d}\zeta \qquad [17]$$

Multiple targets and multiple references

The input scene of the JTC may contain multiple targets. Each target in the scene will correlate with the other targets. The additional correlation peaks can clutter the correlation plane, so some care must be taken when positioning the scene with respect to the reference. Let $s(x, y)$ contain multiple targets $s_i(x, y)$:

$$s(x, y) = \sum_{i=1}^{N} s_i(x - x_i, y - y_i) \qquad [18]$$

The joint power spectrum becomes:

$$\begin{aligned} E(\mu, \nu) = {} & \|R(\mu, \nu)\|^2 + \sum_{i=1}^{N} \|S_i(\mu, \nu)\|^2 + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} \\ & \times S_i(\mu, \nu)S_k^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_i - x_k) + \nu(y_i - y_k)]} + \mathrm{C.C.} \\ & + \sum_{i=1}^{N} R(\mu, \nu)S_i^*(\mu, \nu)\mathrm{e}^{-i2\pi[\mu(x_{\mathrm{r}} - x_i) + \nu(y_{\mathrm{r}} - y_i)]} + \mathrm{C.C.} \end{aligned} \qquad [19]$$

where C.C. denotes the complex conjugate of the preceding term. The correlation plane will now be:

$$\begin{aligned} C(x', y') = {} & R_{\mathrm{rr}}(x', y') + \sum_{i=1}^{N} R_{\mathrm{s}_i\mathrm{s}_i}(x', y') \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{\mathrm{s}_i\mathrm{s}_k}\big(x' - (x_i - x_k),\ y' - (y_i - y_k)\big) \\ & + \sum_{i=1}^{N} R_{\mathrm{rs}_i}\big(x' - (x_{\mathrm{r}} - x_i),\ y' - (y_{\mathrm{r}i} - y_i)\big) \\ & + \sum_{i=1}^{N} \sum_{\substack{k=1 \\ i \neq k}}^{N} R_{\mathrm{s}_k\mathrm{s}_i}\big(x' + (x_i - x_k),\ y' + (y_i - y_k)\big) \\ & + \sum_{i=1}^{N} R_{\mathrm{s}_i\mathrm{r}}\big(x' + (x_{\mathrm{r}} - x_i),\ y' + (y_{\mathrm{r}i} - y_i)\big) \end{aligned} \qquad [20]$$

The second lines of eqns [9] and [20] are the cross-correlation terms between the targets. In order to avoid confusion between reference–target correlations and target–target correlations, special care must be taken in positioning the reference with respect to the scene. All the target–target correlations will be inside a centered box that has twice the length and twice the width of the input scene, so the reference should be placed at least at twice the length or width of the scene.

Multiple targets in the input scene can be viewed as multiple references, and the correlation of multiple references with multiple targets are all computed simultaneously. This property has led to applications in multiclass pattern recognition problems and in optical neural networks, where each neuron of a layer is coded as a reference object. This is one of the advantages of the JTC compared to serial correlators, since serial correlators are limited to one filter at a time.

Example

Here is an example of joint transform correlation. The input plane is shown in Figure 8a. The plane is composed of the input scene in the top third of the input plane and the reference is placed in the center. This input scene contains two targets. The joint power spectrum is shown in Figure 8b with its pixels mapped to a logarithmic intensity scale. The correlation plane is shown in Figure 8c and d with the central peak truncated. The central peak contains the autocorrelation terms R_{ss} and R_{rr}. On each side of the autocorrelation peak are the cross-correlations between the different targets of the input scene. The correlation between the reference image and the input scene is located in the top part of the correlation plane. The correlation plane is symmetrical with respect to the origin, so the correlation peaks are reproduced on the bottom half of the output plane.

Joint transform correlator architectures

Different architectures have been used for JTCs. Three setups are presented here. This is not an exhaustive list. The simplest optical setup shown in Figure 9 is an optoelectronic implementation of a power spectrum machine. A convergent beam

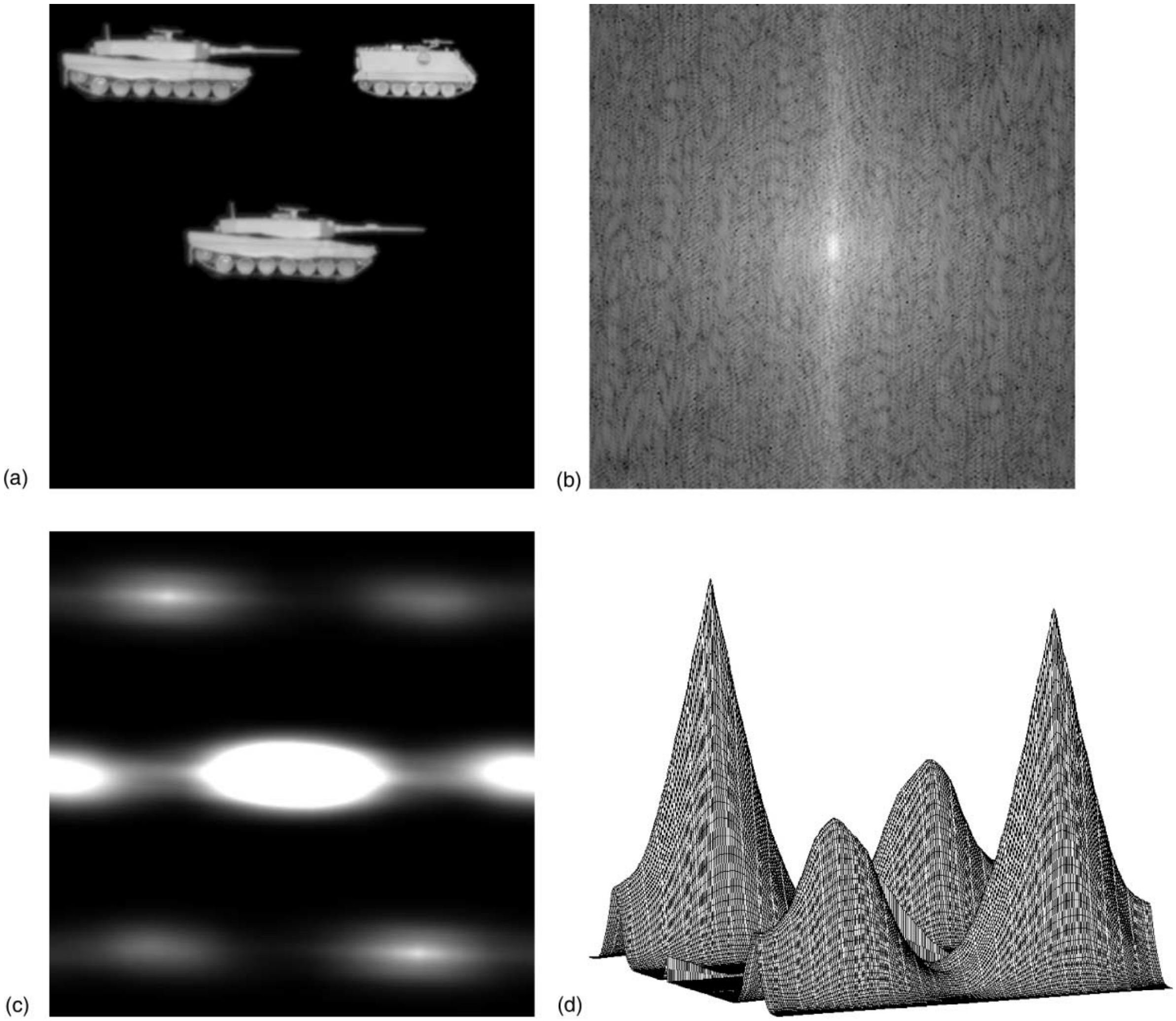

Figure 8 Joint transform correlation example. (a) Input scene; (b) joint power spectrum; (c) correlation plane (top view); (d) correlation plane (mesh view).

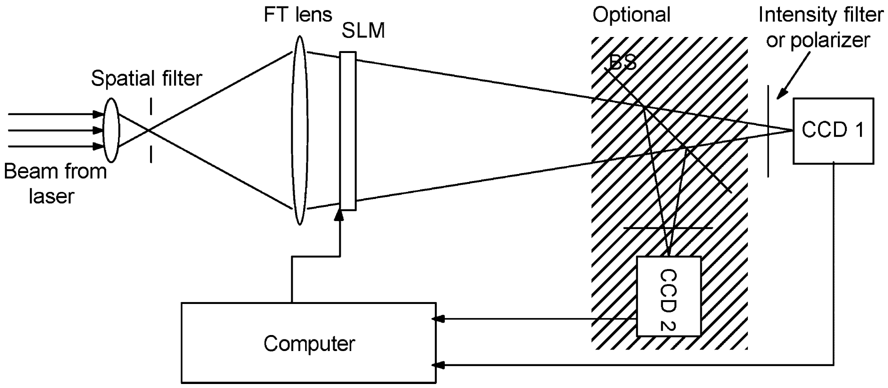

Figure 9 Joint transform correlator implementation as an optical power spectrum machine.

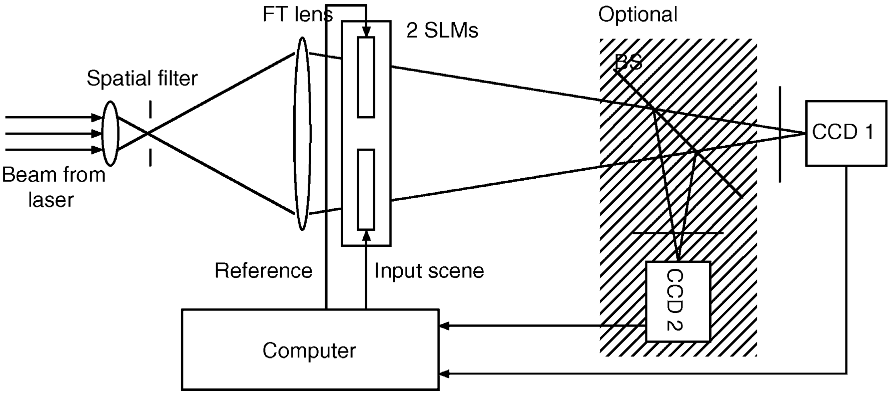

Figure 10 Optical JTC using two SLMs in the input plane.

optical Fourier transform system is used to produce the Fourier transform of an image displayed on a single spatial light modulator (SLM). A CCD camera is placed in the Fourier plane to measure the joint power spectrum. The joint transform correlation is a two-step process: the input plane containing the scene and the reference is first displayed on the SLM, and the camera simultaneously captures the joint power spectrum. In the second step, the joint power spectrum is displayed on the SLM, so the correlation plane is now on the camera. An optional second camera with intensity filters can be added to this setup. In this case, one camera serves to digitize the joint power spectrum during step 1 and the second camera captures the correlation plane at step 2 and transmits it to a display device. This optional configuration is useful to keep both the correlation plane and the joint power spectrum within the dynamic range of their specific cameras.

The setup shown in Figure 9 has two main disadvantages: the correlation is performed in two steps and requires some processing to construct the input plane, and only half of the spatial light modulator can be used to display the input scene. There are alternative setups to alleviate those drawbacks, but they require the use of two or three SLMs. Two modulators can be used in the input plane to simulate a full modulator for the scene, as shown in Figure 10. The original setup can also be doubled with one optical Fourier transform system for the joint power spectrum and one for the correlation plane (Figure 11). With three modulators, it is possible to perform the joint transform correlation in a single step with a full modulator for the input scene.

Comparison to serial transform correlators

The joint transform correlator has some advantages over the serial transform correlator. Since the phase of the Fourier transform is destroyed in the JTC, there are much fewer constraints on the design. So it is advantageous to use an optical convergent beam setup to obtain the Fourier transform. As mentioned before, there is no need for a fully complex spatial light modulator in the joint transform correlator.

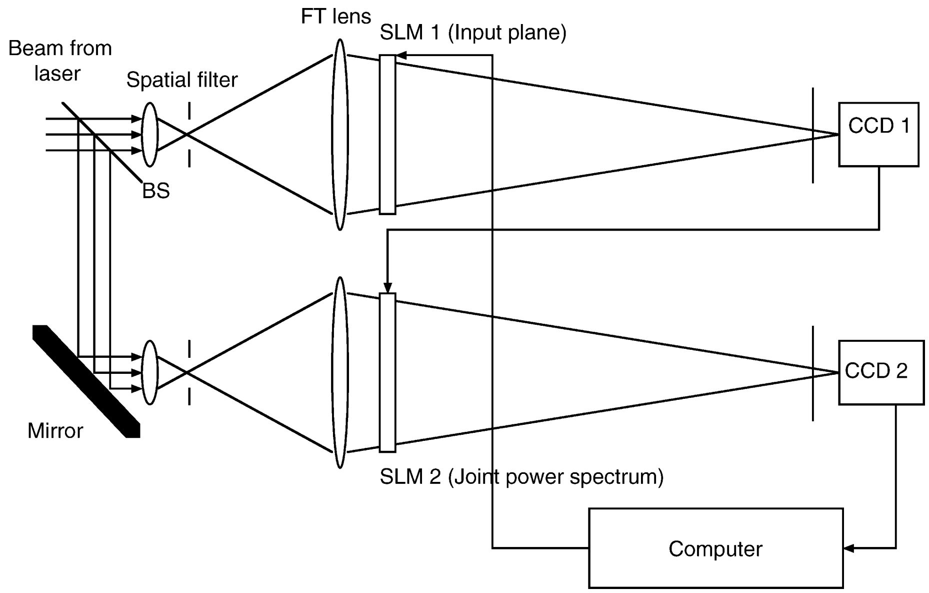

Figure 11 JTC implementation using two power spectrum machines.

This allows the use of commercially available spatial light modulators primarily designed for projectors or for displays, where phase modulation is not required. Since there is a wider market for such devices, they are cheaper.

For pattern recognition, the joint transform correlator has some properties that a serial transform correlator cannot reproduce. Some pattern recognition techniques require a linear combination of multiple amplitude correlations of the scene with different references. This is impossible with serial correlators because only the intensity of the correlation plane is detected. In a JTC, the joint power spectrum is linked to the correlation plane by the Fourier transform. So it is possible to obtain linear combinations of correlation planes by performing the corresponding operations on the joint power spectra. Some recent pattern recognition techniques such as morphological correlation and the sliced orthogonal nonlinear generalized (SONG) correlation, have been optically implemented using a JTC, but the required operations could not be carried out by means of serial optical correlators.

Whereas optical serial transform correlators always carry out linear filtering operations, optical joint transform correlators can be made to carry out nonlinear filtering. This is achieved by imposing a nonlinear characteristic curve on the joint power spectrum, instead of the usual squaring operation inherent to the detection process. The nonlinearity can be intrinsic to the components (saturation, dynamic range) or obtained by introducing electronics between the CCD camera and the second SLM, or by using an optically addressed spatial light modulator. Nonlinear processing can enhance the weights of the high frequencies in the correlation, resulting in sharper correlation peaks and improved discrimination. To illustrate this effect, the joint power spectrum of the example of Figure 8 has been modified with a logarithmic transformation:

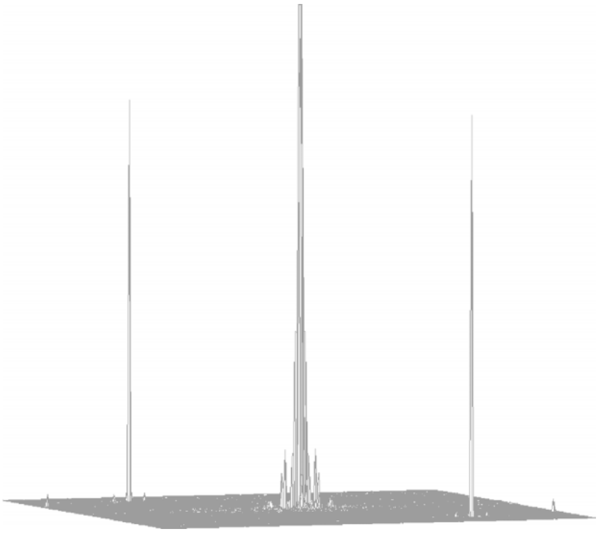

Figure 12 Correlation plane with a logarithmic transformation of the joint power spectrum.

$$E(\mu, \nu) = \ln(1 + \beta \| I(\mu, \nu) \|^2) \qquad [21]$$

with $\beta = 0.25$. The correlation plane for the transformed joint power spectrum is shown in Figure 12.

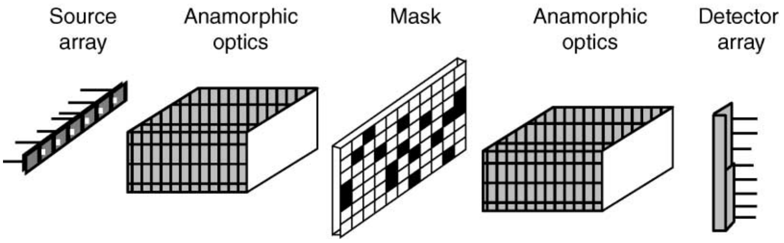

Figure 13 Optical vector–matrix multiplier.

It is clear that the discrimination has improved over that of the previous case.

Optical Vector–Matrix Multiplier

An operation that is frequently required in information processing is the product of a matrix and a vector. This can be accomplished by means on anamorphic optics as shown in Figure 13.

Here the input vector is a set of vertical fans of light from the source array, that are spread out, each over one row of the matrix mask by means of anamorphic optics. The inverse operation is carried out in the second half, so that the vertical array of spots in the output contains the vector–matrix product between the intensities of the source array and the transmissivities of the mask array.

A vector–matrix product can also be carried out by means of a conventional optical correlator as shown in Figure 14.

The input vector **a** is placed in a parallel beam at the focal plane of a conventional 2D correlator. A hologram in the Fourier plane contains a vector array **H**. The vector–matrix product **aH** appears along the axis in the correlation plane.

The above figure shows an input vector $f(x,b)\delta(y-b)$ containing the vector components $f(x1, x2, \ldots, xN)$ arrayed along the horizontal axis x, and shifted a distance a from the horizontal axis. If the Fourier transform of the mask in the filter plane is $h(x,y)$, then the output plane has the distribution:

$$g(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mu,v)\delta(\mu-b)h(x+\mu,y+v)\mathrm{d}\mu\mathrm{d}v \qquad [22]$$

$$g(x,y)=\int_{-\infty}^{\infty} f(\mu,a)h(x+\mu,b-y)\mathrm{d}\mu \qquad [23]$$

$$g(x,b)=\int_{-\infty}^{\infty} f(\mu,b)h(x+\mu,b)\mathrm{d}\mu \qquad [24]$$

$$g(0,b)=\int_{-\infty}^{\infty} f(\mu,b)h(\mu,b)\mathrm{d}\mu \qquad [25]$$

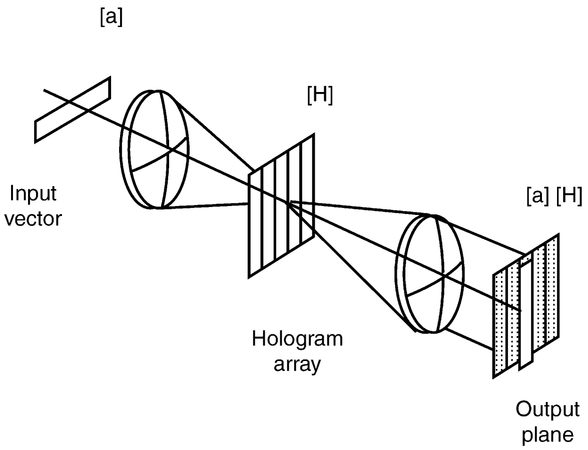

Figure 14 A vector–matrix product using a conventional 2D optical correlator.

So the output value along the y-axis at a distance a from the horizontal axis is equal to the scalar product of the vector $f(x,b)$ with the vector $h(x,b)$. Now if input vectors are stacked up one over the other in the input plane, each one will yield along the output y-axis a value that is the scalar product of the corresponding vector with the vector **h**. The set of values along the y-axis is therefore the vector–matrix product of the matrix constituted by the set of vectors in the input plane, with the vector **h** that is the impulse response of the filter. The cost for this operation is some considerable loss of light, since only the light along the y-axis is used in the output.

Applications of the optical vector–matrix product are optical crossbar, numerical matrix operations such as the solving of linear equations, optical neural networks, spectrum analysis and discrete Fourier transforms.

Conclusion

It has been estimated that serial coherent optical correlators should be able to attain data rates of up to 10^{12} operations per second, using the fastest spatial light modulators available. Joint transform correlators are somewhat slower, due to the input plane being divided between the object and the reference, to geometrical constraints in placement of targets, and due to the need to use electronics to carry out some of the operations. Until now, speeds have been limited by the data rates at the input, but as the speeds of optically addressed SLMs increase, the problem could be shifted to the output, where the lack of optical high spatial bandwidth threshold devices could cause a bottleneck.

List of Units and Nomenclature

A	Vector or matrix
CCD	Charge coupled device
$F(\mu, \nu)$	Fourier transform of $f(x, y)l$
FT	Fourier transform
FT^{-1}	Inverse Fourier transform
H^*	Complex conjugate of H
JTC	Joint-transform correlator
POF	Phase-only filter
R_{ab}	Correlation between functions *a* and *b*
(x, y)	Spatial coordinates
λ	Wavelength of light
(μ, ν)	Spatial frequencies
$*$	Correlation
SLM	Spatial light modulator
SONG	Sliced orthogonal nonlinear generalized correlation

See also

Coherence: Coherence and Imaging. **Diffraction:** Fraunhofer Diffraction; Fresnel Diffraction. **Fourier Optics**. **Geometrical Optics:** Aberrations. **Holography, Techniques:** Computer-Generated Holograms. **Imaging:** Information Theory in Imaging. **Information Processing:** Incoherent Analog Optical Processors; Optical Neural Networks.

Further Reading

Arsenault HH, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. New York: Academic Press.

Bartelt H (1987) Unconventional correlators. In: Horner JL (ed.) *Optical Signal Processing*, pp. 97–128. New York: Academic Press.

Garcia-Martinez P, Arsenault HH and Ferreira C (2001) Time sequential nonlinear correlations for optical pattern recognition. Optoelectronic Information processing: optics for information systems. *Proceedings of SPIE* CR81-12: 186–207.

Goodman JW (1988) *Introduction to Fourier Optics*. New York: McGraw-Hill.

Javidi B (1994) Nonlinear joint transform correlators. In: Javidi B and Réfrégier P (eds) *Workshop on Optical Pattern Recognition*. Bellingham, WA: SPIE – The International Society for Optical Engineering, PM-12.

Javidi B and Horner JL (eds) (1994) *Real-Time Optical Information Processing*. New York: Academic Press.

Javidi B and Johnson KM (eds) (1996) *Optoelectronic Devices and Systems for Processing*. Bellingham, WA: SPIE – The International Society for Optical Engineering, CR65.

Lu XL, Yu FTS and Gregory D (1990) Comparison of Vander Lugt and joint transform correlators. *Applied Physics B* 51: 153–164.

McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley & Sons.

Preston K (1972) *Coherent Optical Computers*. New York: McGraw-Hill.

Vander Lugt AB (1992) *Optical Signal Processing*. New York: John Wiley & Sons.

Free-Space Optical Computing

A A S Awwal, Lawrence Livermore National Laboratory, Livermore, CA, USA

M Arif, Infineon Technologies, Longmont, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical computing encompasses a vast field, from digital and discrete to analog and hybrid optical computing. This article gives an overview of various algorithms, such as shadow-casting, symbolic substitution to nonlinear optics-based processors. Some of the core devices, such as modulators and nonlinear logic elements, are discussed. The optical computing concepts, originated with nonlinear optics and analog optical computing, are more commonly known as optical signal processing. Now they have culminated into digital optical computing, opening the door to highly specialized parallel computing architecture.

Parallel computation is a direct approach to enhance the speed of digital computations which are otherwise inhibited by the sequential processing bottleneck. Physical limitations – such as limited interconnections, interactions amongst electrons, with the consequent inability to share a data bus simultaneously, and constant propagation delay of electronic circuits – place an upper bound on the speed of the fastest operation possible with an electronic computer, without a modification to the computing architecture. Optics provides an architectural advantage over electronics, due to the fact that the most natural way to use optics lends itself to a parallel implementation. Optical signals propagate in parallel, cross each other without interference (whereas charged electrons interact heavily with each other), have the lowest possible propagation delay of all signals, and can provide million channel free-space interconnections with simple lenses (whereas an unrealistic number of connections are necessary for electrons). Thus, optics is an excellent candidate for implementing parallel computing machines using the 3D space connectivity. While electronic architecture is naturally 2D, optical free space architecture is 3D, exploiting the third dimension of space. The communication bottleneck forces electronic computers to update the computation state space sequentially, whereas the use of optics will empower one to change the entire computation space in parallel.

Historical Perspective

Historically, digital optical computing was proposed as a natural extension of nonlinear optics. Just as transistors, which are nonlinear electronic devices, led to the realization of the computing logic using transistors, it was envisioned that nonlinear optical effects would give rise to optical transistors which would lead to development of optical logic gates. One of the limitations of nonlinear optical device-based computation is the high laser power requirement to demonstrate any nonlinear optical effect.

One of the effects used to demonstrate nonlinearity is to change the refractive index with the incident intensity. In other words, most materials will behave linearly with low optical energy. The refractive index expressed as $n = n_0 + n_2 I$ is nominally linear because the nonlinear co-efficient n_2 is usually very small. In the presence of high power, the $n_2 I$ is significant enough to cause an optical path change by altering the refractive index. When such a medium is placed inside a Fabry–Perot resonator, it will allow one to control the transmission of the filter by changing the incident intensity. Thus, at high power, it shows a nonlinear behavior or bistability. The term optical bistability was coined, which meant the presence of two stable states of output for the same intensity value.

Another development took place at the University of Michigan by the development of a precision optical processor; this was an optical processor which could process radar signals using optics, this is the example of an analog optical computing processor.

A third area which is in between the digital and the analog are the discrete level devices which most often take the form of matrix-based processors. Free space optical matrix vector multipliers are examples of such processors.

The latest field of optical computing research is in the algorithms and architecture area. The motivation here is to develop new algorithms which take advantage of optics. The thing that is lacking is the presence of appropriate device technology that can implement the algorithm in a given architecture. In terms of practical realization, analog optical computing is the most successful one. Adaptive optic processors have found success both in astronomy as well as vision sciences, and more recently in optical imaging through the atmosphere.

Development of free space optical computing has two main driving forces: the technology (device) base and the algorithm base. In this classification, since the field is in its infancy, both device-based classes and architectural (algorithm) classes are listed side-by-side as separate entities, ignoring the natural overlap that may be present in some instances. For example, nonlinear optical devices could be implemented into a variety of architecture that is listed below. Holographic processor is another technology-based class while some other architectures, such as, symbolic substitution may involve use of holograms, but it does not exclude it being a class by itself. The following list gives a classification of optical free space computing:

1. Digital Optical Computing:
    (a) Pattern coded computing or cellular logic:
        (i) Optical shadow-casting processor;
        (ii) Symbolic substitution;
        (iii) Theta logic based computation.
    (b) Parallel arithmetic circuits:
    (c) Hybrid optical computing:
    (d) Nonlinear optical device based:
    (e) Threshold logic.
2. Discrete optical computing:
    (a) Optical fuzzy logic:
    (b) Optical matrix processors:
        (i) Vector matrix multiplier;
        (ii) Differential equation solver;

        - (iii) Linear algebra processor;
        - (iv) DFT processor;
        - (v) Programmable logic array;
        - (vi) Associative memory processor.
    - (c) Optical neural network.
3. Analog optical computing:
    - (a) Adaptive optics processor;
    - (b) Partially coherent processing (white light processing):
    - (c) Fourier transform based processor:
    - (d) Nonlinear contact screen based:
    - (e) Holographic processors:
    - (f) Linear optical processors (image differentiation).

We start by discussing some of the device based classes and their underlying concepts: (a) nonlinear optics; (b) SLMs; and (c) hologram based.

Core Device Technologies Used for Free-Space Optical Computing

One of the most difficult challenges of optical computing (OC) is inventing devices that can realize aforementioned key advantages of optics. There exist, however, several key technology areas, where a breakthrough in any one of these will pave the way to real-world optical computing. We discuss here some of the major device technologies that relate to OC.

Spatial light modulator (SLM) converts information from different forms to the optical domain. As a result, SLM is used as an input device to an optical computing system. High contrast ratio, very high speed, and parallel addressing scheme are desirable properties of an SLM as an input device. For example, the SLM could be used to input a 2-dimensional signal or an image into a Fourier transform-based processor. Several applications of SLM are possible:

1. *Optical encoding*: Pattern coded logic can perform logic operations if inputs can be encoded using certain pattern;
2. *Encoding complex filter*: It could encode phase, amplitude or both; and
3. *As phase modulator*: for adaptive optics.

Two types of SLMs are liquid-crystal based and micro-electromechanical (MEM) based. The former uses liquid crystals to change the phase of transmitted or reflected light leading to modulation of phase. Although the most successful commercial application of MEMs is its use in consumer electronics, such as projection TVs or video projectors, in free-space optical computing it can be used as:

(a) an interconnection device;
(b) an SLM for matrix; or
(c) a phase modulating device in AO close loop control of optical phase aberrations or phase only filter.

The heart of the MEMs is tiny electrically controllable mirrors which can steer the incoming beam in a different direction or bend to create a relative optical path difference for the reflected beam.

Algorithms, where an SLM plays a major role, are pattern coded computing, optical matrix processors, adaptive optic processors, partially coherent processing, Fourier transform based processors, and linear optical processors.

Nonlinear Optical Element

A nonlinear optical element is an architecture or device structure which allows for utilization of the nonlinearity of the optical material that produces detectable changes of optical intensity. Figure 1 shows the basic block diagram for an optical computing basic element. To achieve an optical logic gate, one needs a scheme where light can be controlled by light, as current is controlled by current in a transistor. An optical device showing optical bistability can be used for this purpose. Optical bistability is defined as the situation where two different output intensities are possible for a given input intensity.

For example, as shown in Figure 2 below, at the intensity level of C, there are two stable states, one HIGH and one LOW. Just as the transistor has two stable states ON and OFF, the optical bistable device exhibits two states ON and OFF. The fact that such nonlinear behavior could implement logic behavior

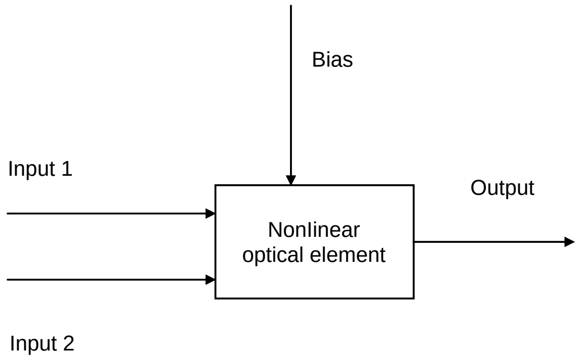

Figure 1 Block diagram of an optical logic gate using nonlinear optical element.

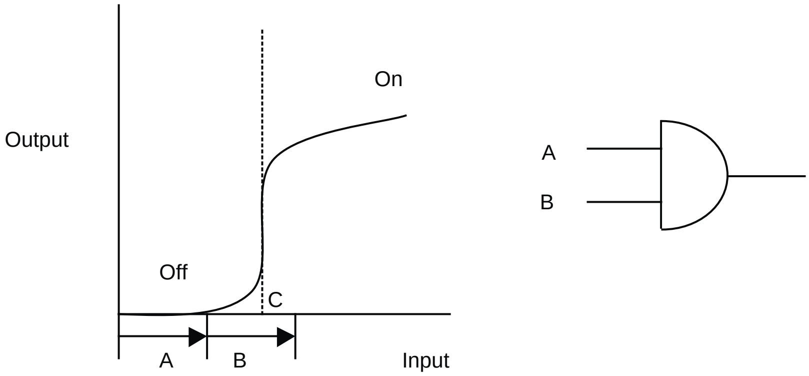

Figure 2 Bistable behavior, a typical AND behavior with inputs A and B.

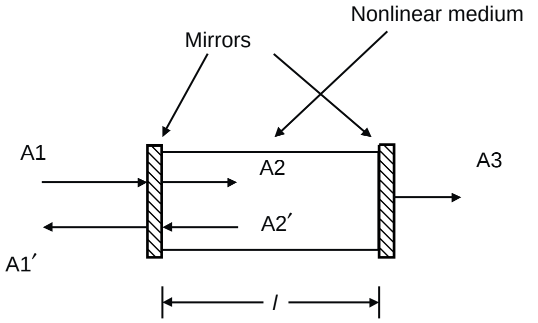

Figure 3 Bistable optical device, nonlinear medium in a Fabry–Perot cavity.

can be explained as follows. Assume that a two input device possesses a nonlinear characteristic as shown in Figure 2. Assume that the input intensity value of 1 is indicated by an intensity level equal to A. When both A and B are equal to 1, then the combined intensity switches the device to HIGH output state. However, if only one input is present, the output will be LOW. Thus, this bistable device acts as a logical AND gate.

To achieve optical bistability, two features are required, namely, nonlinearity and feedback. One way to achieve this is by placing a nonlinear medium in a Fabry–Perot cavity. Figure 3 shows a schematic of these devices. In Figure 3, A_1 is the incident field, A_1' is the reflected field, A_2 and A_2' are the forward and backward fields inside the cavity, respectively, and l is the cavity length.

The coupled wave equation for this system is given by

$$A_2' = \rho A_2 e^{2ikl-\alpha l} \qquad [1a]$$

$$A_2 = \tau A_1 + \rho A_2' \qquad [1b]$$

where, ρ is the amplitude reflectivity and τ is the amplitude transmissivity. The wave vector $k = n\omega/c$ and α is the absorption coefficient. In deriving the above equations, it was assumed that k and α are spatially invariant and real. Solving for A_2, from eqns [1a] and [1b] yields

$$A_2 = \frac{\tau A_1}{1 - \rho^2 e^{2ikl-\alpha l}} \qquad [2]$$

Equation [2] is the so-called Airy's equation which describes the properties of the Fabry–Perot cavity. From eqn [2] it can be seen that if k or α is a sufficiently nonlinear function, this device will show bistability. In general, both k and α can show nonlinearity. Hence optical bistability can be either absorptive or dispersive.

Absorptive Bistability

If the absorption coefficient α depends nonlinearly on the intensity, assuming k is constant, after some manipulation, eqn [2] can be written as

$$I_2 = \frac{TI_1}{[1 - R(1 - \alpha l)]^2} \qquad [3]$$

where, $I_i \propto |A_i|^2$, $T = |\tau|^2$ and $R = |\rho|^2$. In deriving eqn [3], it was assumed that the mirror separation is such that the system is in resonance and $\alpha l \ll 1$. If α depends on the intensity, it can be assumed that it follows the relationship of two level saturable absorbers, for simplicity, which is

$$\alpha = \frac{\alpha_0}{1 + I/I_s} \qquad [4]$$

where, α_0 denotes the unsaturated absorption coefficient, I is the local value of intensity, and I_s is the saturation intensity. Using eqn [4] in eqn [3] yields

$$I_1 = TI_2\left(1 + \frac{C_0}{1 + 2I_2/I_s}\right)^2 \qquad [5]$$

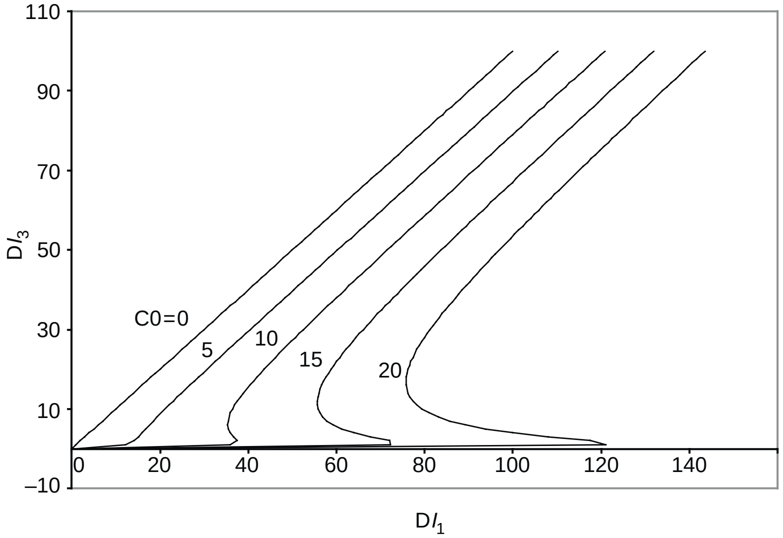

Figure 4 The input–output relationship of absorptive bistability predicted by eqns [5] and [6].

where, $C_0 = R\alpha_0 l/(1-R)$. Also the output intensity I_3 is related to I_2 by

$$I_3 = TI_2 \qquad [6]$$

Using eqns [5] and [6], it can be shown that there can be multiple solutions to the above input–output relationship which predicts the optical bistability. Figure 4 shows a plot of DI_1 versus DI_3, where $D = 2/TI_s$, as a function of various C_0. As can be seen from Figure 4, for a high enough value of C_0, the system possesses multiple solutions.

Dispersive Bistability

The intensity dependence of the refractive index can be written as

$$n = n_0 + n_2 I \qquad [7]$$

If the refractive index n varies nonlinearly with intensity, assuming $\alpha = 0$ and using eqn [7], eqn [2] can be written as

$$A_2 = \frac{\tau A_1}{1-\rho^2 e^{2ikl}} = \frac{\tau A_1}{1 - Re^{i\delta}} \qquad [8]$$

where, $\delta = \delta_0 + \delta_2$. δ_0 and δ_2 are given by

$$\delta_0 = \varphi + 2n_0\frac{\omega c}{l}$$
$$\delta_2 = 2n_2 I\frac{\omega c}{l} \qquad [9]$$

where, φ is the phase associated with ρ and the fact that $k = n\omega/c$ is used.

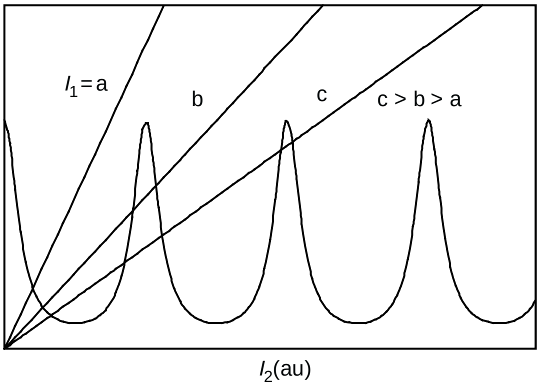

Figure 5 Plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the straight lines represent the left-hand side with increasing values.

After some manipulation, eqn [8] can be rewritten in terms of intensity as

$$\frac{I_2}{I_1} = \frac{1/T}{1+(4R/T^2)\sin^2(\delta/2)} \qquad [10]$$

and

$$\delta = \delta_0 + \left(4n_2\frac{\omega}{c}l\right)I_2 \qquad [11]$$

Equation [10] can be solved graphically to show the introduction of bistable behavior as a function of input intensity. Figure 5 shows the plots of both sides of eqn [10] as a function of I_2 and for increasing values of input intensity I_1. The oscillatory curve represents the right-hand side of eqn [10] and the straight lines represent the left-hand side with increasing values of I_1. It can be readily seen from Figure 5 that at sufficiently high values of I_1, multiple solutions are possible for eqn [10]. This again will give rise to the bistable behavior.

Nonlinear Interference Filters

The optical bistability from the nonlinear interference filters (or etalons) can be used to achieve optical logic gates. The transmission property of these etalons can be modified by another optical beam, known as a bias beam. A typical logic gate and the transmission property are shown in the following figure. If the bias beam is such that, $P_{\text{switch}} - (P_a + P_{\text{bias}}) < P_b$, then the transmitted beam shown in Figure 6b readily follows an AND gate. The truth table for such configurations is shown in Figure 6d.

Four-Wave Mixing

Four Wave Mixing (FWM) can be used to implement various computing functions. Using the photorefractive effect and FWM in a third-order nonlinear optical material, optical phase conjugate beams can be obtained. A schematic of FWM and phase conjugation is shown in Figure 7. The forward pump beam interferes with the probe beam to produce a grating in the photorefractive material. The backward pump beam is then diffracted to produce a beam which retraces the path of the probe beam but has an opposite phase. It can be seen from Figure 7, that the optical phase conjugation can be readily used as an AND logical operator. By using the four-wave mixing, an optical counterpart of transistors has been introduced. FWM have also been used in AND-based optical symbolic substitution (OSS) operations. FWM is also used to produce matched filter-based optical correlators.

Another device showing the optical bistability is the self electro-optic effect device (SEED). SEEDs were used to implement optical logic circuits. The SEEDs are produced by putting multiple quantum wells (MQW) in the intrinsic region of a p-i-n structure. The MQWs are created by placing alternating thin layers of high (barriers) and low (wells) bandgap materials. An electric field is applied to the SEEDs externally. The absorption coefficient is a function of applied bias voltage. When a light beam is incident on these devices, it creates a photocurrent, consequently lowering the bias voltage across the MQW. This, in turn, increases the absorption and when the input optical power is high enough, the system sees a peak absorption and suddenly switches to a lower state and thus shows the bistable behavior. A schematic of SEED is shown in Figure 8.

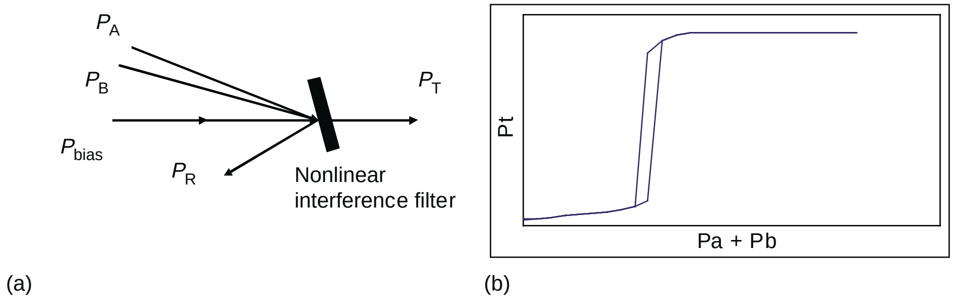

Truth table for an AND function from the above setup from the transmitted beam

P_A= input 1	P_B = input 2	P_T = output
0	0	0
0	1	0
1	0	0
1	1	1

(d)

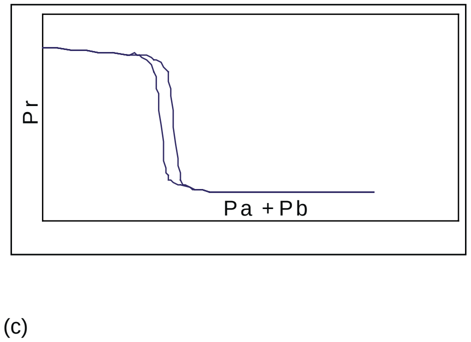

Figure 6 A typical optical logic gate configuration using nonlinear interference filter: (a) schematic; (b) the transmitted power output; (c) reflected power output (both as functions of inputs); and (d) the truth table for AND operation with beams A and B as inputs.

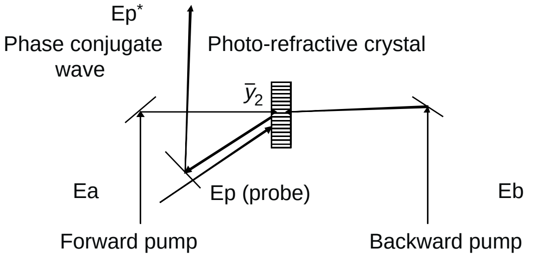

Figure 7 Four-wave mixing setup.

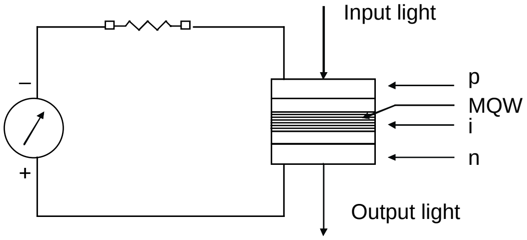

Figure 8 Schematic of self electro-optic effect devices.

Replacing the resistive load by a photodiode, a diode-based SEED (D-SEED) can be constructed, which is analogous to the way diode transistor logic was developed from resistor transistor logic in electronics. Electrically connecting two MQW p-i-n structure symmetric SEEDs (S-SEED) are constructed, which are less sensitive to optical power fluctuations and effectively provide isolation between input and output.

Pattern Coded Optical Computing: Symbolic Substitution or Cellular Logic

This is a keyword which is equivalent to simple digital logic operation. In digital logic, a truth table can describe the relationship between input and output variable. In symbolic substitution logic, this relationship is expressed as a 2D pattern rule. In any specific rule (equivalent to a row in the truth table) of SS logic, a specific 2D pattern is identified and replaced by another pattern. For example, in the case of binary logic, we can express the addition of two binary bits with the following truth table:

Rule number	*Input*		*Output*	
	A	*B*	*C*	*S*
1	0	0	0	0
2	0	1	0	1
3	1	0	0	1
4	1	1	1	0

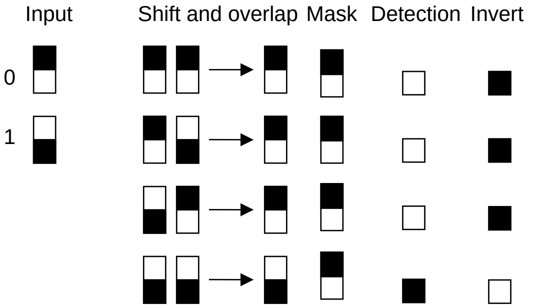

Figure 9 Symbolic substitution example, the two inputs are coded using dual rails. The second column shows the four combinations of inputs (00, 01, 10, and 11). They are shifted and overlapped creating the fourth column. A mask detects the output as a dark pixel appearing through the transparent opening. The dark is inverted to create a detection output.

Here A and B are the two digits to be added, C and S are the sum and the carry bit, respectively. To convert this operation to an SS operation, we can say that a two bit binary pattern is identified and replaced by a sum bit and a carry bit. If we now express the above in terms of symbolic substitution, we will say let 0 and 1 be represented by the following 2D symbol. Then the rule for the carry bit generation is shown in Figure 9.

Note that each of these rules has to be implemented individually. However, one can implement these by only generating the 1 output or the zero output, whichever leads to the minimum rule. Here implementing the last rule is convenient, one can make a copy of the whole pattern and then overlap it with its left shift version. This will result in a dark pixel at the origin. This is inverted and then produces a detection output. Natural extension of SS is in signed digit computing, where carry free addition can be achieved, allowing parallel implementation of numerical algorithm using free space optics.

Optical Shadow Casting (OSC)

All digital logic is nonlinear. The nonlinearity of the digital logic is converted into nonlinearity of coding. The OSC system originally proposed by Tanida and Ichioka uses two-dimensional spatially encoded patterns as data input and light emitting diodes (LEDs) as light sources. In the original shadow-casting system, the inputs were represented by vertical and horizontal stripes of opaque and transparent bars. With this system, programmability was attained by changing the LED pattern and it was possible to realize 16 logical operations between two

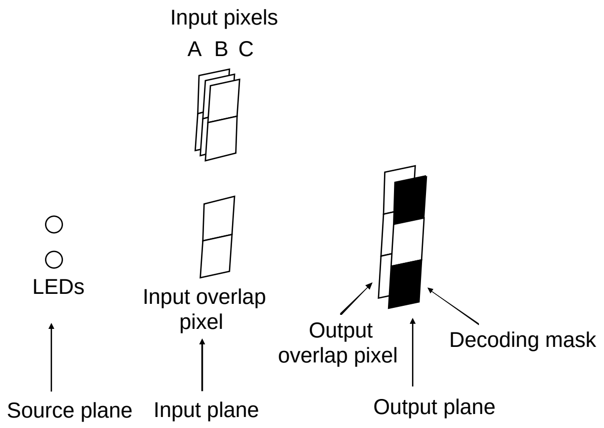

Figure 10 An optical shadow-casting system.

variables as well as a half addition as an example of arithmetic operation. Later, polarized codes were introduced into shadow-casting systems. This improved technique, called the polarization-encoded optical shadow casting (POSC), enabled complex combinational logic units like binary and trinary full adders, as well as image processing applications, to be realized.

The lensless OSC system, as shown in Figure 10, consists of spatially encoded 2D binary pixel patterns as the inputs. The input patterns are kept in close contact at the input plane and the resulting input overlapped pixel pattern is illuminated by a set of LEDs from the source plane. The light originating from each of the LEDs located at the source plane produces a shadow of the input overlap pixel pattern at the output plane. The overlap of these relatively displaced shadows results in an output overlap pixel pattern at the output plane. A decoding mask placed at the output plane is then used to spatially filter and detect the logical output. To prevent crosstalk, specific spacing among the different processing elements must be maintained.

Discrete Processors: Optical Matrix Processor

Most of the discrete optical processing schemes revolve around optical matrix–vector multiplication. Many numerically useful operations such as: (a) linear algebraic operations; (b) signal processing algorithms; (c) Boolean logic; (d) digital multiplication by analog convolution; and (e) neural networks, can be formulated as matrix–vector multiplication. An example of the digital multiplication by analog convolution (DMAC) can be understood by the following numerical example:

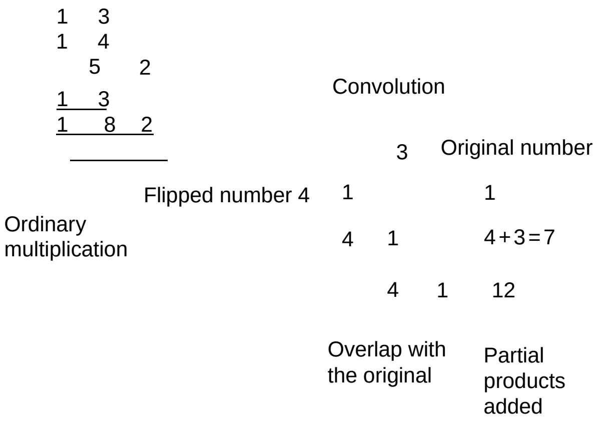

$$1\ \ 7\ \ 12 = 1\times100+7\times10+12\times1=182$$

The two numbers are convolved as shown above. This is done by reversing one of the numbers (4 1), while keeping the other number in the same order (1 3), then sliding the reversed number to the right, and collecting the element by element products. The second partial products are 4 × 1 and 1 × 3 which are then added to give the convolution results. Surprisingly, the weighted number in mixed radix (since numbers that have digits higher than 9, are no longer decimal) has the same value as the product shown in the upper left. It is possible to have fast parallel convolution optically, resulting in fast parallel multiplier. However, an efficient electronic post processor is necessary to convert the mixed radix numbers to the radix of the to-be-multiplied numbers.

A basic optical matrix–vector multiplier is shown in Figure 11. Here the vector, represented by an expanding column of light, is projected onto the 2D matrix, possibly represented by a glass plate or electronically addressable spatial light modulator. As the light representing the vector passes through the transmitting medium, each element of the vector (**I**) gets multiplied by a column of matrix represented by the transmittance of the matrix (T). The product (**I**T)

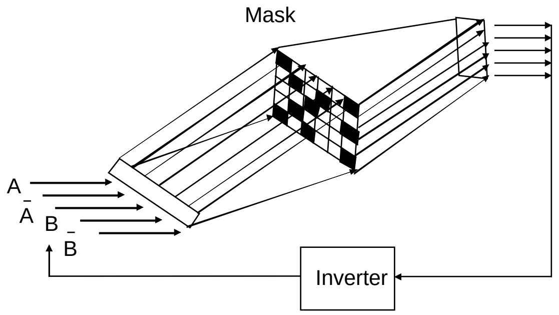

Figure 11 A matrix–vector multiplier.

is integrated horizontally using a cylindrical lens. This leads to an addition operation of the partial products, resulting in the generation of the output product vector. Mathematically:

$$y = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}b_1 + a_{12}b_2 + a_{13}b_3 + a_{14}b_4 \\ a_{21}b_1 + a_{22}b_2 + a_{23}b_3 + a_{24}b_4 \\ a_{31}b_1 + a_{32}b_2 + a_{33}b_3 + a_{34}b_4 \\ a_{41}b_1 + a_{42}b_2 + a_{43}b_3 + a_{44}b_4 \end{bmatrix} \qquad [12]$$

The above can be achieved if the vector b is expanded as

$$\begin{bmatrix} b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \\ b_1 & b_2 & b_3 & b_4 \end{bmatrix}$$

and a point-by-point multiplication is performed between the expanded vector and the original matrix. Then the addition is performed by the cylindrical lenses.

Another simple application of vector matrix multiplier is in digital logic, known as programmable logic array (PLA). PLA is an electronic device that consists of a set of AND gates followed by a set of OR gates, which can be organized in AND–OR format suitable for realizing any arbitrary general purpose logic operation. AND logic is performed when light representing a logic level is passed through a transparency representing the other variable; OR is achieved by converging light to a common detector.

$$\text{Assume a function } F = AC + \bar{A}BD + A\bar{B}C \qquad [13]$$

where $\bar{A}$ represents the logical inverse of A. Using DeMorgans' theorem in digital logic, this can be expressed as

$$F = \overline{(\bar{A} + \bar{C})} + \overline{(A + \bar{B} + \bar{C})} + \overline{(\bar{A} + B + \bar{C})}$$

$$= \bar{y}_1 + \bar{y}_2 + \bar{y}_2 \qquad [14]$$

Now the y_is can be generated by the vector matrix product, as shown below:

$$\text{where} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} A \\ \bar{A} \\ B \\ \bar{B} \\ C \\ \bar{C} \\ D \\ \bar{D} \end{pmatrix} \qquad [15]$$

Note here that the AND has been converted to OR, which needs to be inverted and then ORed to generate the function F. A setup shown in **Figure 11** can be used to accomplish this. In step 1, the y_is are generated, and then in step 2 they are inverted and summed by a second pass through the same system with a different mask. Other applications of vector–matrix multipliers are in neural networks, where the interconnection weights are represented by the matrix and the inputs are the vector.

Analog Optical Computing

Correlation is a way to detect the presence of signal in additive noise. Optics offer a fast way of performing correlations. The practical motivation of this area comes from the fact that a lens performs a Fourier transform of a two-dimensional signal at its focal plane. The most famous analog optical computing is known as optical matched filter. The theoretical basis of an optical matched filter is in the Fourier theorem of matched filtering which states that the cross correlation between two signals can be formed by multiplying the Fourier transform of the signal and the complex conjugate of the second signal and followed by a subsequent Fourier transform operation. Mathematically:

$$\text{Correlation}(f, g) = \mathrm{F}^{-1}[\mathrm{F}\{f\}^{*}\text{conjugate}(\mathrm{F}\{g\})] \qquad [16]$$

where F represents a Fourier transform operation and F^{-1} represents an inverse transform. Since the lens can easily perform the Fourier transform, and multiplication can be performed by light passing through a medium, the only problem that has to be

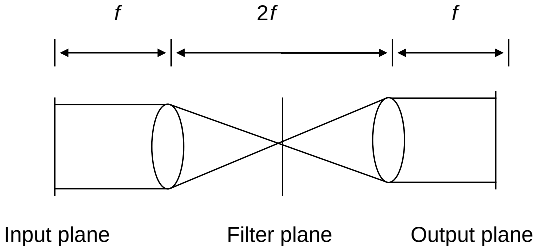

Figure 12 An optical 4-*f* set up. Lenses perform Fourier transform at their focal plane.

solved is to represent the signal and the filter in the optical domain. Since an SLM can represent both the real signal and the complex filter, thus the whole operation can be performed in a 4-*f* setup as shown in Figure 12. The input is displayed in the input plane by an SLM, the first lens performs a Fourier transform. This filter (conjugate F{*g*}) is encoded using another SLM and placed at the filter plane. When light from the F{*f*} passes through the filter, it performs the required multiplication. The second lens performs a second Fourier transform and produces the correlation output at the output plane. Variations of the basic setup exist as a joint transform correlator (JTC) or binary phase-only filter (BPOF). While the JTC achieves the product of the complex domain by adding the two signals and squaring, the BPOF is achieved by simplifying the complex domain filter by numerically approximating it with the binarized phase. Other applications of linear optical processing can be performed in the basic 4-*f* optical setup where filters such as low-pass or high-pass filters, can be set up in the correlation plane.

The hologram base computing treats holograms as a legitimate way of signal transformation from one form to another, just as a lens transforms from one domain to another. Holograms are complex representations that can act on a signal. A lens can be approximated by a hologram, in reality functions of many optical elements, such as lens, grating, prism, and other transformation such as beamsplitting, it can be combined into single computing element known as computer-generated hologram (CGH). The smartness of this type of system depends on how many operations have been replaced by a single CGH.

It is envisioned that optical computing is better suited for special purpose computing rather than general purpose computing. However, the future of optical computing lies in the marriage of appropriate device technology with the right architecture at the right applications. The development of quantum computing may hold new promises for optical computing.

List of Units and Nomenclature

DMAC	Digital multiplication by analog convolution
JTC	Joint transform correlator
OC	Optical computing
PLA	Programmable logic array
QC	Quantum computing
SLM	Spatial light modulator

See also

Nonlinear Optics, Basics: Four-Wave Mixing; Nomenclature and Units. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arsenault H, Szoplik T and Macukow B (eds) (1989) *Optical Processing and Computing*. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*. San Diego, CA: Academic Press.

Gaskill J (1978) *Linear Systems Fourier Transform and Optics*. New York: Wiley.

Gibbs HM (ed.) (1982) *Optical Bistabiligy II*. New York: Plenum Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw Hill.

Ichioka Y and Tanida J (1984) Optical parallel logic gates using a shadow-casting system for optical digital computing. *Proceedings of the IEEE* 72: 787.

Karim MA and Awwal AAS (1992) *Optical Computing: An Introduction*. New York: Wiley.

Koren I (1993) *Computer Arithmetic Algorithms*. New Jersey: Prentice Hall.

McAulay A (1992) *Optical Computing Architecture*. New York: Wiley.

Reynolds GO, DeVelis JB, Parrent GB and Thompson BJ (eds) (1989) *Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, CA: SPIE Optical Engineering Press.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

VanderLugt A (1992) *Optical Signal Processing*. New York: Wiley.

Yu FTS and Jutamulia S (1992) *Optical Signal Processing, Computing and Neural Networks*. New York: Wiley.

Incoherent Analog Optical Processors

S Jutamulia, University of Northern California, Petaluma, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

After low-cost lasers became widely available, it seemed that incoherent analog optical processing was no longer important. In general, incoherent optical processing has no apparent significant advantages, as compared with its coherent optical counterpart or digital computer processing. However, although nobody now attempts to build an incoherent optical processor, such as an incoherent correlator, most instruments are still operating using incoherent or natural light. These incoherent instruments relate to imaging, including cameras, microscopes, telescopes, projection displays, and lithographic equipment. From this point of view, it is still very important to study the principles of incoherent analog optical processing, in order to further progress or improve the incoherent image.

Incoherent Image Formation

The image formation in a coherent optical system can be explained using the linear system. A perfect point $\delta(x, y)$ in the input plane is imaged in the output plane as the impulse response of the system $h(x, y)$. When the input is only a single point, the output amplitude is simply $h(x, y)$ and the output intensity is $|h(x, y)|^2$. Accordingly, the function of $|h(x, y)|^2$ is called point spread function (PSF).

If we consider that the input object is the collection of a large number of very fine points, under coherent light illumination, the output image will be the collection of the same number of blurred spots. The shape of each blurred spot is $h(x, y)$. If the amplitude function of input object is $f(x, y)$, the amplitude function of output image is

$$o(x, y) = \iint_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\,\mathrm{d}q \qquad [1]$$

which is a convolution of $f(x, y)$ and $h(x, y)$. The intensity function of output image is

$$|o(x, y)|^2 = \left| \iint_{-\infty}^{\infty} f(p, q)h(x - p, y - q)\mathrm{d}p\,\mathrm{d}q \right|^2 \qquad [2]$$

According to the convolution theorem, in the Fourier transform domain or frequency domain, we have

$$O(u, v) = F(u, v)H(u, v) \qquad [3]$$

where $O(u, v)$, $F(u, v)$, and $H(u, v)$ are the Fourier transforms of $o(x, y)$, $f(x, y)$, and $h(x, y)$, respectively. $H(u, v)$ is the transfer function of the coherent imaging system. To distinguish it from the incoherent imaging system, it is also called coherent transfer function (CTF).

The same optical imaging system is now illuminated with incoherent light instead of coherent light. The amplitude impulse response is still the same $h(x, y)$, and the point spread function is also the same $|h(x, y)|^2$. Similarly, the input object is the collection of perfect points, and the output image is the collection of blurred images of each point. However, since the illuminating light is incoherent, light from any point in the input object is not coherent with light from any other points. Thus, there is no interference with light from different points. The output image is simply the addition of intensity patterns generated by each point in the input object. The amplitude function of the output image is not computable. However, the intensity function of the output image is

$$|o(x, y)|^2 = \iint_{-\infty}^{\infty} |f(p, q)|^2 |h(x - p, y - q)|^2 \mathrm{d}p\,\mathrm{d}q \qquad [4]$$

In the frequency domain, we now have

$$O_I(u,v) = F_I(u,v)H_I(u,v) \qquad [5]$$

where $O_I(u,v)$, $F_I(u,v)$, $H_I(u,v)$ are Fourier transforms of $|o(x,y)|^2$, $|f(x,y)|^2$, and $|h(x,y)|^2$, respectively. The intensity impulse response $|h(x,y)|^2$ is the PSF. Note that the index I of O_I, F_I, and H_I denotes intensity. The function $H_I(u,v)$ is now the transfer function of the incoherent imaging system, which is called optical transfer function (OTF).

Referring to Fourier analysis, if $H(u, v)$ is the Fourier transform of $h(x, y)$, the Fourier transform of $|h(x, y)|^2$ is the autocorrelation of $H(u, v)$ as follows:

$$H_I(u, v) = \iint_{-\infty}^{\infty} H(p, q)H^*(p - u, q - v)\mathrm{d}p\,\mathrm{d}q \qquad [6]$$

The OTF (the function $H_I(u, v)$) is the autocorrelation of CTF (the function $H(u, v)$), which is a complex function in general. It can be written as

$$\mathrm{OTF} = |\mathrm{OTF}|\exp(i\phi) \qquad [7]$$

We can further write

$$\mathrm{MTF} = |\mathrm{OTF}| \qquad [8]$$

where MTF stands for modulation transfer function, and

$$\mathrm{PTF} = \phi \qquad [9]$$

where PTF stands for phase transfer function.

Measurement of MTF

The MTF of an incoherent imaging system can be measured empirically using an input that is a cosine grating having intensity transmittance function:

$$|f(x)|^2 = 1 + \cos(2\pi a x) \qquad [10]$$

where a is the frequency of grating. The output or the image intensity function will be

$$|o(x)|^2 = 1 + m\cos(2\pi a x + \varphi) \qquad [11]$$

where m, the contrast, can be measured by

$$m = \frac{|o(x)|^2_{\max} - |o(x)|^2_{\min}}{|o(x)|^2_{\max} + |o(x)|^2_{\min}} \qquad [12]$$

In the frequency domain, the Fourier transform of the input function is

$$F_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a) + \tfrac{1}{2}\delta(u+a) \qquad [13]$$

The Fourier transform of the output function will be

$$O_I(u) = F_I(u)\mathrm{OTF}(u) \qquad [14]$$

Substitution of eqn [13] yields

$$O_I(u) = \delta(u)\mathrm{OTF}(0) + \tfrac{1}{2}\delta(u-a)\mathrm{OTF}(a) + \tfrac{1}{2}\delta(u+a)\mathrm{OTF}(-a) \qquad [15]$$

Notice that

$$\mathrm{OTF}(0) = 1 \qquad [16]$$

due to normalization, and

$$\mathrm{OTF}(a) = \mathrm{OTF}(-a) \qquad [17]$$

Because OTF is an autocorrelation function, it must be a symmetric function. Substitution of eqns [16] and [17] into eqn [15] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\mathrm{OTF}(a) + \tfrac{1}{2}\delta(u+a)\mathrm{OTF}(a) \qquad [18]$$

Remembering that OTF is a complex function given in eqn [7], we can write

$$\mathrm{OTF}(a) = \mathrm{MTF}(a)\exp(i\phi(a)) \qquad [19]$$

Substitution of eqn [19] into eqn [18] yields

$$O_I(u) = \delta(u) + \tfrac{1}{2}\delta(u-a)\mathrm{MTF}(a)\exp(i\phi(a)) + \tfrac{1}{2}\delta(u+a)\mathrm{MTF}(a)\exp(i\phi(a)) \qquad [20]$$

The inverse Fourier transform of eqn [20] is

$$|o(x)|^2 = 1 + \mathrm{MTF}(a)\cos(2\pi a x + \phi(a)) \qquad [21]$$

By comparing eqn [21] and eqn [11], we find

$$\mathrm{MTF}(a) = m \qquad [22]$$

and the PTF

$$\phi(a) = \varphi \qquad [23]$$

Therefore, the MTF at frequency $u = a$ of the incoherent imaging system can be measured using eqn [12]. To get the complete MTF(u), the measurement is repeated using cosine gratings with different frequencies. The PTF can also be measured at the same time.

Incoherent Spatial Filtering

In the 4f coherent optical processor, the first lens performs the Fourier transform, and the second lens performs the inverse Fourier transform, as shown in Figure 1. The Fourier transform domain or frequency domain is materialized in the frequency plane that is the back focal plane of the first lens.

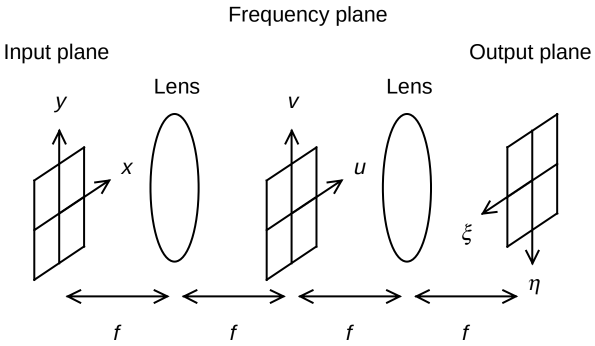

Figure 1 4f coherent optical processor consists of two Fourier transform lenses.

Thus, the Fourier transform of the object can be visualized in the frequency plane. For example, if the object is a grating, we will see its frequency components in the frequency plane. However, if the coherent illumination (e.g., from a laser) is replaced with incoherent illumination (e.g., from a lightbulb), the pattern of Fourier transform in the frequency plane disappears. Does it mean that we can no longer perform spatial filtering? No, we still can perform spatial filtering, although the pattern of Fourier transform is no longer visualized.

Consider that we have a spatial filter in the frequency plane whose physical form can be expressed by the function $H(u,v)$. For illustration, this spatial filter is simply a small one-dimensional window $H(u)$, as shown in Figure 2a. If the 4f optical processor is illuminated with coherent light, the CTF of the system is $H(u)$ itself as shown in Figure 2b. However, if the 4f optical processor is illuminated with incoherent light, we will use OTF instead of CTF in the analysis of linear system. The OTF is the autocorrelation of CTF, which is graphically depicted in Figure 2c. In this example, since the OTF is real and positive, the MTF is identical to the OTF.

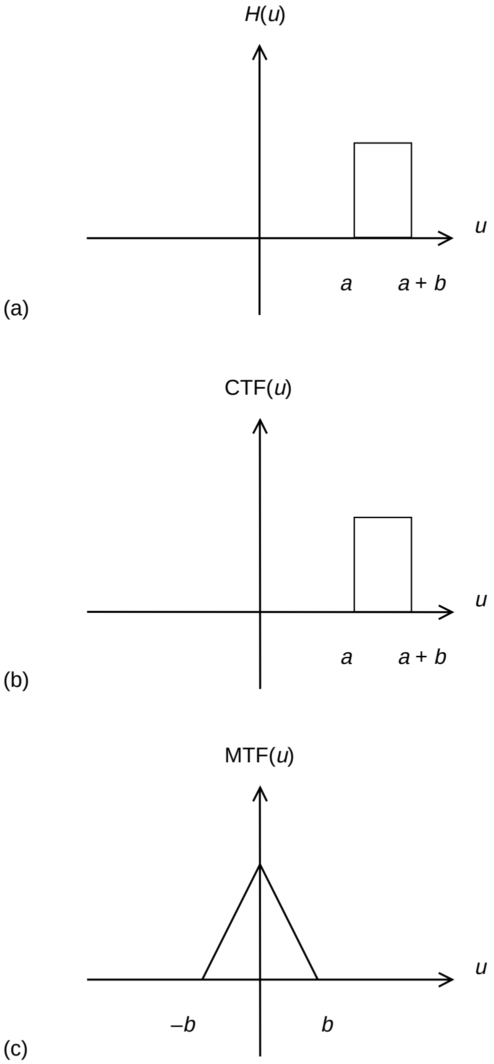

Figure 2 (a) One-dimensional window in frequency plane functions as a spatial filter. (b) CTF and (c) OTF or MTF produced by spatial filter (a).

For simplicity, we analyze and show a one-dimensional filter only in a two-dimensional graph. It certainly can be extended to the two-dimensional filter. And we need a three-dimensional graph to show the two-dimensional autocorrelation.

It is important to note that, although the CTF is a high pass filter that passes frequencies in the band of $u = a$ to $u = a + b$, the MTF is a low pass filter that passes frequencies from $u = -b$ to $u = b$, independent of a. This indicates that an incoherent processor cannot enhance high frequency. It is always a low pass processor. It is interesting to note that the OTF is independent of the location of the filter $u = a$, but it is determined by the width of the filter $\Delta u = b$ only.

Accordingly, an aperture made up of randomly distributed pinholes behaves as a low pass filter in an incoherent imaging system. The cut-off frequency of such a filter can be derived from the diameter of the pinhole, which is equivalent to Figure 2b. The procedure for using this filter is simple. If a camera is used to take a picture, the random pinhole filter can be simply attached to the camera lens. The picture taken will not consist of high-frequency components, because the filter acts like a low pass filter. The low pass filtering can also be done by reducing the aperture of camera. However, by doing so, the light entering the camera is also reduced. The use of a random pinhole filter removes high frequencies without reducing the light intensity. When the random pinhole filter is placed in the frequency plane of a 4f coherent optical processor, it can be used to reduce the speckle noise. Note that the speckle noise is the main drawback in coherent optical processing.

Incoherent Complex Matched Spatial Filtering

Lohmann and Werlich pointed out that the Vander Lugt correlator can also be operated with incoherent light. In the coherent Vander Lugt correlator the correlation term is:

$$o_3(\xi,\eta) = \iint_{-\infty}^{\infty} f(\alpha,\beta) g*(\alpha-\xi,\beta-\eta) \mathrm{d}\alpha \mathrm{d}\beta \qquad [24]$$

Its intensity distribution is

$$|o_3(\xi,\eta)|^2 = \left| \int\int_{-\infty}^{\infty} f(\alpha,\beta) g*(\alpha-\xi,\beta-\eta) \mathrm{d}\alpha \mathrm{d}\beta \right|^2 \quad [25]$$

However, when the input object is illuminated with incoherent light instead of coherent light, the intensity distribution of the correlation term becomes

$$|o_3(\xi,\eta)|^2 = \int\int_{-\infty}^{\infty} |f(\alpha,\beta)|^2 |g(\alpha-\xi,\beta-\eta)|^2 \mathrm{d}\alpha \mathrm{d}\beta \quad [26]$$

Equation [26] shows the correlation of $|f(x,y)|^2$ and $|g(x,y)|^2$. Therefore, when $f(x,y)$ is identical to $g(x,y)$, the correlation peak always appears regardless of whether the illumination is coherent or incoherent. However, the detailed structures of the coherent and incoherent correlation outputs are different. Note that the joint transform correlator cannot be operated using incoherent light, since no joint Fourier spectra will be formed.

Depixelization of Projection Display

Recently, a technique to remove the pixel structure from an image projected by liquid crystal projection display has been reported. A liquid crystal projection display must have a pixel structure to form an image. Can we remove this pixel structure from the projected image to obtain a movie-like image? It is well known that a single hole in the frequency plane can perform a coherent spatial filtering, such that the pixel structure will disappear from the output image.

Because of the pixel structure, we get multiple Fourier spectra in the frequency plane. By passing only one spectrum order through the hole in the frequency plane, the pixel structure can be removed as shown in Figure 3.

It is interesting to note that selecting any one of the spectra will produce the image at the same location.

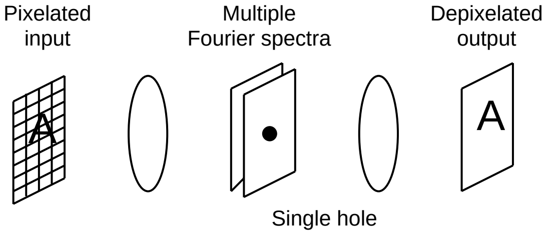

Figure 3 Removal of pixel structure using a single hole in frequency plane. Most of the energy is wasted and the output intensity is dim.

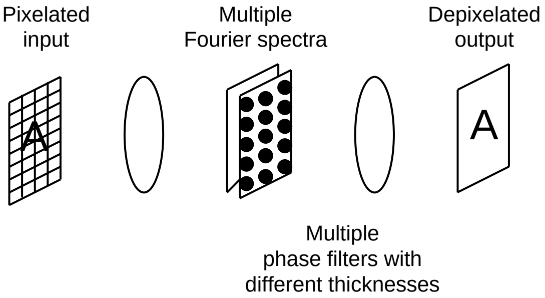

Figure 4 Removal of pixel structure using multiple phase filters with different thickness to cover each Fourier spectrum order. No energy is wasted and the output intensity is very bright.

However, if two spectra are passed, on the top of the produced image, Young interference fringes are also produced. If all spectra are passed, the produced interference fringes are in fact the pixel structure itself. By passing only one spectrum order and blocking most spectrum orders, we lose most of the energy. The projected image will be very dim, although the pixel structure is removed.

To overcome this problem, we may cover each spectrum order with a transparent material having different thickness as shown in Figure 4. Thus, every spectrum order is passed and delayed by a phase filter with a different thickness. If the delay produced by the phase filter is larger than the coherent length of light, every spectrum order is no longer coherent to each other. In other words, the resultant image is the sum of the intensities of all images. As a result, the pixel structure does not appear, and no intensity is lost. For a white light source, since the frequency bandwidth is large, the coherent length is typically on the order of several tens of μm.

This technique can significantly improve the quality of the liquid crystal projection display. Figure 5 shows the experimental result of a depixelated projection image. Figure 5a shows the projected image of an input with pixel structure when no phase filter is applied in the frequency plane. Figure 5b shows the projected image of the same input with pixel structure when phase filters are applied in the frequency plane. The pixel structure is successfully removed as shown in Figure 5b.

Computed Tomography

The computed tomography or CT using X-ray is usually considered beyond optical information processing. However, it is interesting to review the principle of the X-ray CT, since the X-ray source is incoherent, and the CT image reconstruction utilizes Fourier transformation that is commonly used in

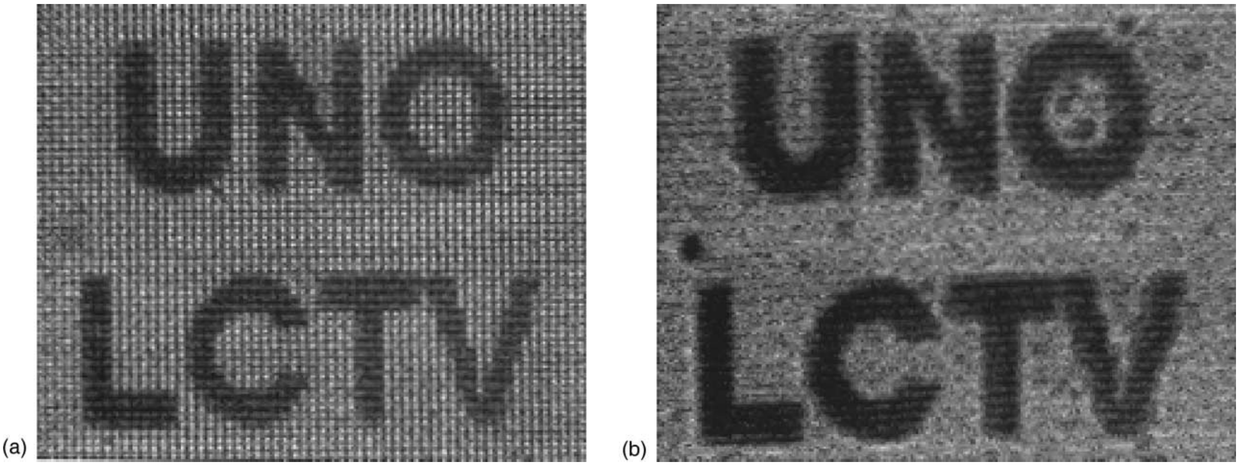

Figure 5 (a) Projected image consists of pixel structure when no phase filter is applied in the frequency plane. (b) Depixelated image is produced when phase filters are applied in the frequency plane.

optical information processing. The X-ray CT is used for taking the cross-section picture of the human body. Note that a conventional X-ray picture only provides a projection instead of a cross-section picture.

Consider the case when an X-ray penetrates through an object characterized by its attenuation coefficient or density function $f(x,y)$ as shown in Figure 6. The detected X-ray intensity is

$$I = I_0 \exp - \left\lfloor \int f(x,y) \mathrm{d}s \right\rfloor \qquad [27]$$

where I_0 is the original intensity. Equation [27] can be written as

$$\int f(x,y)\mathrm{d}s = -\ln \frac{I}{I_0} \qquad [28]$$

Apparently, we cannot directly measure $f(x,y)$, although we can obtain $\int f(x,y)\mathrm{d}s$ from the measurement of I and I_0. However, we want to get the cross-section picture which is the density function $f(x,y)$. The basic idea is to compute $f(x,y)$ from its Fourier transform $F(u,v)$, and $F(u,v)$ is derived from eqn [28]. $F(u,v)$ can be expressed as

$$F(u,v) = \int\int_{-\infty}^{\infty} f(x,y) \exp[-\mathrm{i}2\pi(ux+vy)]\mathrm{d}x\,\mathrm{d}y \qquad [29]$$

For $v = 0$, eqn (29) becomes

$$F(u,0) = \int\int_{-\infty}^{\infty} f(x,y) \exp[-\mathrm{i}2\pi ux]\mathrm{d}x\,\mathrm{d}y \qquad [30]$$

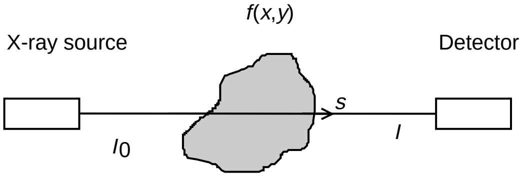

Figure 6 X-ray passes through object having density function $f(x,y)$.

which can be written

$$F(u,0) = \int_{-\infty}^{\infty} p(x) \exp[-\mathrm{i}2\pi ux]\mathrm{d}x \qquad [31]$$

where

$$p(x) = \int_{-\infty}^{\infty} f(x,y)\mathrm{d}y \qquad [32]$$

The projection function $p(x)$ given by eqn [32] can be obtained using an X-ray parallel beam as shown in Figure 7. Then $F(u,0)$ can be computed using computer based on eqn [31]. $F(u,0)$ is the Fourier transform along the u-axis. To get other data, we rotate the coordinate (u,v) to (u',v') by a small angle α. Correspondingly, the coordinate (x,y) is also rotated to (x',y') by the same angle α while the object is not rotated as shown in Figure 8.

From the measurement shown in Figure 8, we can get $p(x')$, which, in turn, provides $F(u',v'=0)$. By increasing the rotation angle α, we will get other data of $F(u'',v''=0)$. The completion of rotation of 180 degrees will give us the data in the frequency domain (u,v) as shown in Figure 9.

After the data $F(u,v)$ is collected, as shown in Figure 9, $F(u,v)$ can be inverse Fourier transformed to

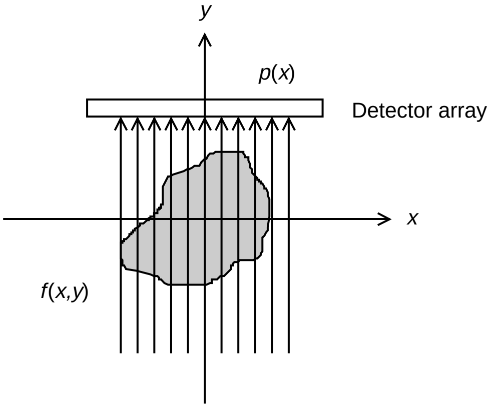

Figure 7 X-ray parallel beam passes through the object to produce projection function $p(x)$.

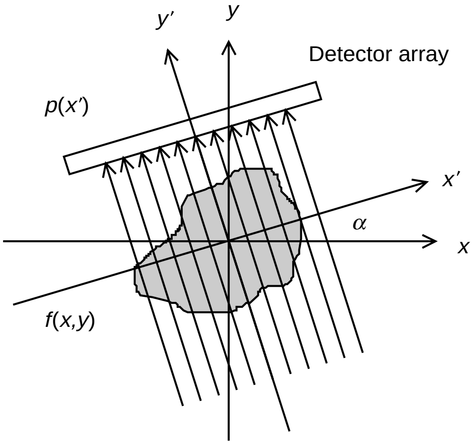

Figure 8 Rotated X-ray parallel beam produces the rotated projection function $p(x')$.

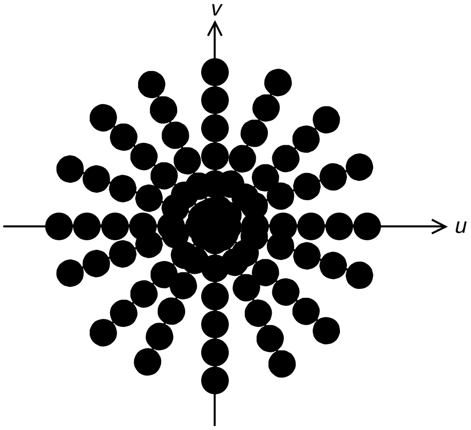

Figure 9 $F(u, v)$ computed from a set of rotated projection functions.

produce $f(x, y)$. There are two approaches in the inverse Fourier transform. The first approach is to interpolate the computed value $F(u, v)$ such that the value of $F(u, v)$, on a regular grid structure, can be defined before an inverse Fourier transform can be taken.

The second approach is to apply polar coordinates, instead of Cartesian coordinates, to the inverse Fourier transform, such that no interpolation in frequency domain (u, v) is required. The inverse Fourier transform in Cartesian coordinates is as follows:

$$f(x, y) = \int\int_{-\infty}^{\infty} F(u, v) \exp[i2\pi(ux + vy)]\mathrm{d}u\,\mathrm{d}v \quad [33]$$

It can be written in polar coordinates as:

$$f(x,y) = \int_0^{2\pi}\int_0^{\infty} F(\rho\cos\phi, \rho\sin\phi) \times \exp[i2\pi\rho(x\cos\phi + y\sin\phi)]|\rho|\mathrm{d}\rho\,\mathrm{d}\phi \quad [34]$$

where

$$u = \rho\cos\phi \quad [35]$$

$$v = \rho\sin\phi \quad [36]$$

Therefore $f(x,y)$ can be obtained from the collected $F(u,v)$, as shown in Figure 9, without further interpolation using eqn [34]. Note that eqns [31], [33], and [34] are computed using a computer program.

See also

Spectroscopy: Fourier Transform Spectroscopy.

Further Reading

Bracewell RN (1965) *The Fourier Transform and its Applications*. New York: McGraw-Hill.

Hounsfield GN (1973) Computerized transverse axial scanning (tomography): Part I. *British Journal of Radiology* 46: 1016–1022.

Iizuka K (1985) *Engineering Optics*. New York: Springer-Verlag.

Jutamulia S, Toyoda S and Ichihashi Y (1995) Removal of pixel structure in liquid crystal projection display. *Proceedings of SPIE* 2407: 168–176.

Klein MV (1970) *Optics*. New York: Wiley.

Lohmann AW and Werlich HW (1971) Incoherent matched filtering with Fourier holograms. *Applied Optics* 10: 670–672.

Reynolds GO, DeVelis JB, Parrent GB Jr and Thompson BJ (1989) *The New Physical Optics Notebook: Tutorials in Fourier Optics*. Bellingham, WA: SPIE.

Vander Lugt A (1964) Signal detection by complex spatial filtering. *IEEE Trans. Inf. Theo.* IT-10: 139–145.

Yang X, Jutamulia S and Li N (1996) Liquid-crystal projection image depixelization by spatial phase scrambling. *Applied Optics* 35: 4577–4580.

Yu FTS and Wang EY (1973) Speckle reduction in holography by means of random spatial sampling. *Applied Optics* 12: 1656–1659.

Optical Bit-Serial Computing

A D McAulay, Lehigh University, Bethlehem, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Why Optical Bit-Serial Computing?

Optical bit-serial computing will lower cost and improve dataflow rates in optical telecommunications and associated computing. Moreover, the electromagnetic immunity of optics will enhance security for computing as well as for communications. Also, absence of sparking with optics enables safe operation, even in hazardous environments.

Currently, optical bit-serial communication involves passing terabits of data per second (10^{12} bps) – thousands of encyclopedias per second – along an optical fiber by switching light of multiple colors on and off, representing ‘1’ and ‘0’ bits respectively. Replacing electronic computing functions, constrained to tens of gigabits per second (10^9 bps), with integrated optics, ones of much higher speed, enables faster flow rates – allowing continuation of the annual doubling of capacity-distance that has occurred since the 1970s. Cost is reduced by eliminating the demultiplexing and multiplexing for optic-electronic-optic (OEO) conversions. Two illustrative areas of optical bit-serial computing for replacing electronic computing functions are described: computing at up to 40 Gbps with semiconductor optical amplifiers (SOAs) and computing at higher bit rates with integrated optic components. Several books address the relevant technical background: free-space optical information processing, optical computer architectures, optical fiber communications, and integrated optics.

Computing with Semiconductor Optical Amplifiers

Semiconductor optical amplifiers, SOAs, allow nonlinear operations (switching, logic) at low power (mW) because of amplification – typically up to 1000. An SOA is a laser diode with antireflection coatings in place of reflecting facets. Light pulses at bit rates in excess of 40 Gbps are amplified by absorbing power from an electronic pump. SOAs are used in the following, at bit rates up to 40 Gbps, for nonlinear operations in cross-gain, cross-phase, or cross-polarization.

SOA-XGM for Logic

Figure 1 shows an SOA in cross-gain modulation (XGM) performing a NOR function. Either A or B inputs at a ‘1’ level, driving the SOA through couplers into saturation, lowering its gain. A continuous wave at C, having a different frequency λ_3, from that of A, λ_1 and B, λ_2, experiences the decrease in gain, resulting in an output ‘0’ level at frequency λ_3. The filter in Figure 1 can be avoided by counter-propagating A and B through the SOA from right to left. XOR gates use interference between two incoming streams of bits (locked to identical frequency, phase, polarization) to correlate headers on internet protocol (IP) address headers – similar to matching zip codes in postal mail routing machines. Bit-serial-adders are constructed from combinations of XOR and NOR gates; subsequently the adders are used to construct ripple-carry-adders for word addition. In addition, NOR gates are used to construct flip-flops; hence registers which form short-term memory.

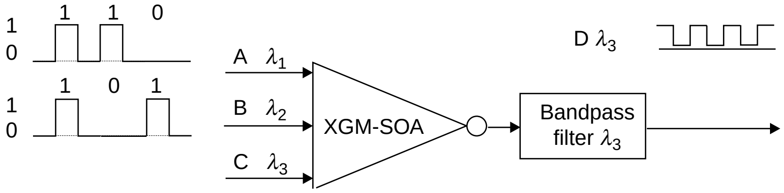

Figure 1 Optical XGM-SOA NOR gate.

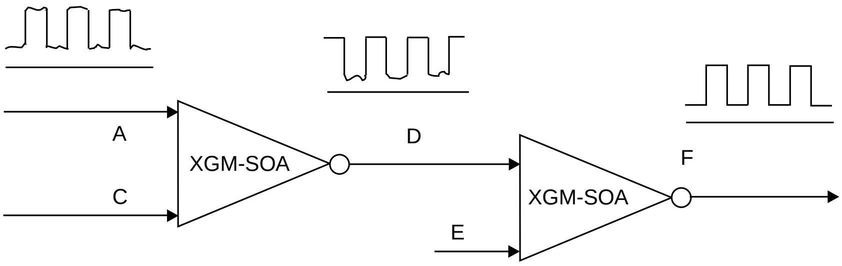

Figure 2 Optical XGM-SOA 2R signal regeneration.

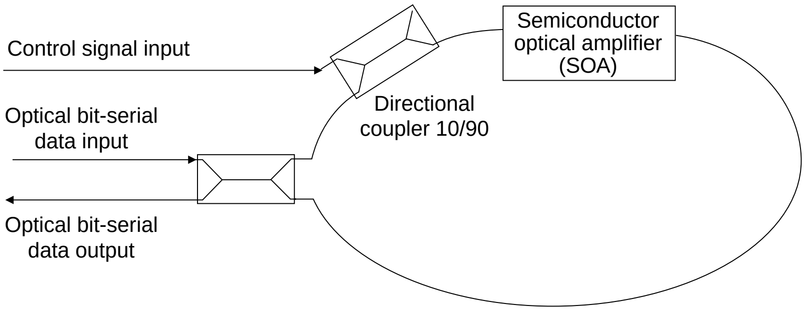
Control signal input
Semiconductor optical amplifier (SOA)
Directional coupler 10/90
Optical bit-serial data input
Optical bit-serial data output

Figure 3 XPM-SOA in Sagnac interferometer for optical manipulation of bits at high data rates.

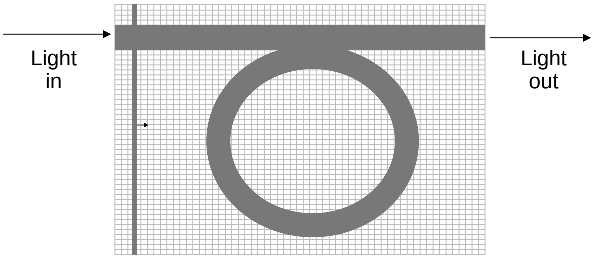

Figure 4 Integrated optic Optiwave layout for microring resonator.

SOA-XGM for Frequency Conversion

Ignoring input B, the bit information in Figure 1 has been transferred from carrier frequency A onto the carrier frequency C, with inversion. This enables switching of signals in wavelength division multiplexing (WDM), for example, color reuse is simplified and a color red is routed to the location where blue was previously routed; this is equivalent to time slot interchangers for switching in time division multiplexing (TDM).

SOA-XGM for 2-R Signal Regeneration

Figure 2 show how two SOAs may be used in sequence to restore signal levels distorted by propagation or computation; reamplification and reshaping (2-R). The ‘1’ level at the output of the first SOA is clipped by gain saturation, removing noise, this becoming the ‘0’ level at the output of the second SOA. The second SOA clips the ‘1’ level at the output.

SOA-XPM with Interferometer

In cross-phase modulation (XPM), the signal at A, Figure 1, changes the phase of carrier for signal C passing through the SOA (signal levels may be less than for XGM). Inclusion of an SOA in an interferometer, Mach–Zehnder or Sagnac, converts phase to intensity for frequency conversion, switching or high-speed bit manipulation. Figure 3 shows a control pulse entering a coupler at the left and providing a control input to the SOA at one frequency. The bits circulating round the loop at a different frequency can be switched on and off by the control pulse, to synchronize or manipulate high bit rate signals or for frequency conversion.

Computing with Integrated Optics

Combining several discrete optical components into a single optical chip is more beneficial than in electronic very large-scale integration (VLSI) because optical connections require greater precision; moreover bit rates are higher because photons, unlike electrons, have no mass or charge (only spin).

Integrated Optic Microring Resonator

Figure 4 shows a microring resonator in an integrated optic chip for optical filtering, light entering at the left couples into the loop. For loop lengths that are a multiple of wavelength, light cycling the loop is in phase and resonance occurs. At resonant frequencies,

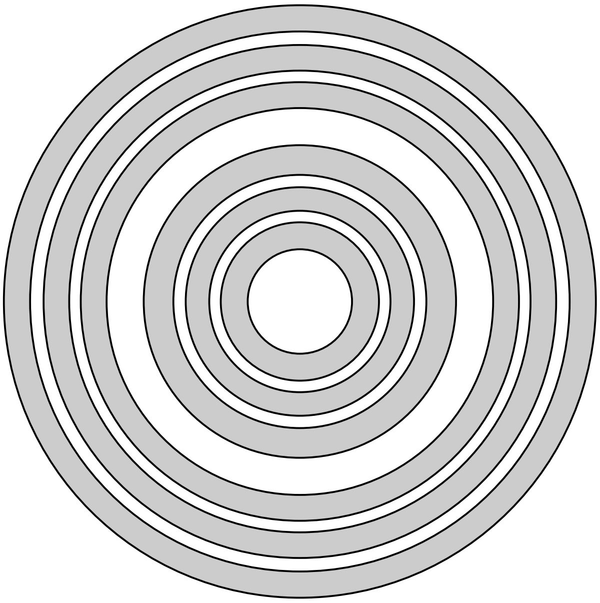

Figure 5 Integrated optic Optiwave layout for Bragg microring resonator.

light exiting the loop is out of phase with that in the linear guide, causing cancellation of the input. As frequency varies away from resonance, light passes and the device acts as a notch filter to block light at the resonant frequencies. With nonlinear Kerr material, low-level light intensity may be used to change a resonant frequency to make a switch.

Bragg Integrated Optic Microring Resonator

Performance is improved by using a Bragg microring resonator, Figure 5, which confines light in the central light colored ring by dielectric mirror reflection from Bragg gratings on either side. The refractive index is less in the ring than in the surrounding regions, in contrast to a conventional waveguide (used in Figure 4) that guides by total internal reflection. The higher reflectivity, smaller ring radius and lower refractive index – these help to reduce power loss, volume, and travel time around the ring. This increases filter selectivity, drives up peak power for enhanced nonlinearity for switching, decreases stored energy for fast response, and increases frequency range (free spectral range, *FSR*). In this case, an equivalent photonic crystal may be more expensive because it requires higher resolution lithography and greater refractive index contrast.

Photonic Crystal Mach–Zehnder Switch

Recently developed photonic crystal technology allows a reduction in integrated optic chip size from centimeters to submillimeters. The 2D or 3D periodic

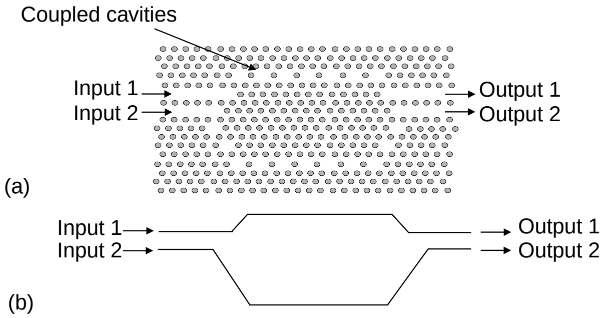

Figure 6 Proposed photonic crystal Mach–Zehnder interferometer.

structure of dielectric variations emulates crystals and provides dielectric mirror reflections in multiple dimensions for lossless waveguide bends, high Q resonators, and slowing of wave propagation to enhance effective nonlinearity of materials. Figure 6 shows a proposed 2D photonic crystal Mach–Zehnder interferometer switch in which light is split into two paths by a coupler and propagation velocity is slowed for enhancing nonlinearity by using coupled-resonator optical waveguides (CROWs), in which alternate posts of different dielectric constant remain in the waveguides to form coupled cavities.

Conclusion

Optical bit-serial computing can reduce cost and increase flow-rate for expanding telecommunication network capacity, while providing enhanced security from electromagnetic sabotage and tampering. Advances in integrated optics (microring resonators), progress in optoelectronic components (SOAs), improved materials and fabrication techniques – these technologies enable cost-effective optical bit-serial computing solutions for future telecommunication evolution.

See also

All-Optical Signal Regeneration. **Information Processing:** All-Optical Multiplexing/Demultiplexing.

Further Reading

Agrawal G (2002) *Fiber-Optic Communication Systems*. New York: Wiley.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic Publishers.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill.

Heuring VP, Jordan HF and Pratt JP (1992) A bit serial architecture for optical computing. *Applied Optics* 31(17): 3213–3224.

Hunsperger RG (2002) *Integrated Optics*, 5th edn. New York: Springer.

Joannopoulios JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.

McAulay AD (1991) *Optical Computer Architectures*. New York: Wiley.

McAulay AD (2004) Novel all-optical flip-flop using semiconductor optical amplifiers in innovating frequency-shifting inverse-threshold pairs. *Optical Engineering* 43(5): 1115–1120.

Scheuer J and Yariv A (2003) Two-dimensional optical ring resonators based on radial Bragg resonance. *Optics Letters* 28(17): 1528–1530.

Xu L, Glesk I, Baby V and Prucnal PR (2004) All-optical wavelength conversion using SOA at nearly symmetric position in a fiber-based Sagnac interferometric loop. *IEEE Photonics Technical Letters* 16(2): 539–541.

Yariv A (1997) *Optical Electronics in Modern Communications*, 5th edn. New York: Oxford University Press.

Optical Digital Image Processing

B L Shoop, United States Military Academy, West Point, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Optical approaches to image processing provide a natural method by which to obtain the inherent advantages associated with optics including massive parallelism, high-speed processing, and the inherent compatibility with image formats. Early applications of optics employed analog processing techniques with the most common being the optical Fourier transform, matrix-vector processors, and correlators. Later, advances in digital signal processing (DSP), digital storage, and digital communication systems demonstrated the potential for higher resolution, improved flexibility and functionality, and increased noise immunity over analog techniques and quickly became the preferred method for accurate signal processing. Consequently, optical processing architectures, that incorporated digital processing techniques, were explored. Advances in very large-scale integration (VLSI) electronic circuitry and optoelectronic devices later enabled smart pixel technology, which integrates the programmability and processing of electronic circuitry with the two-dimensional (2D) nature of an optical architecture and made possible a particular realization of optical digital image processing. This chapter describes an application of optical digital image processing, based on smart pixel technology, called digital image halftoning.

Natural images are, by definition, continuous in intensity and color. Halftoning is the process by which a continuous-tone, gray-scale image is converted to one containing only binary output pixels. Halftoning can be thought of as an image compression technique whereby a high-resolution image is transformed to a low-resolution image containing only black-and-white pixels. The transformation from a continuous-tone, gray-scale image to one containing only binary-valued pixels is also similar, in principle, to low-resolution analog-to-digital (A/D) conversion.

An early manual approach to printing continuous-tone images, using only the presence or absence of ink, is the Mezzotint. Ludwig von Siegen (1609–1680), an amateur Dutch printmaker, first invented the mezzotint, or ‘half tone process’, in 1640. The process later came into prominence in England during the early eighteenth century. Mezzotints are produced on copper plates where the surface of the plate is roughed with a tool called the Mezzotint Rocker, shaped like a wide chisel with a curved and serrated edge. By rocking the toothed edge backwards and forwards over the plate, a rough burr is produced which holds the ink. The dark regions of an image were roughed in a random fashion, while the areas to be lightened were scraped and burnished. This process was found to be especially useful for the reproduction of paintings, due to its ability to capture the subtlest gradations of tone from rich, velvety blacks to glowing highlights. The mezzotint is, therefore, an early artistic predecessor to the modern day halftone.

Optical halftoning has been in use by commercial printers for over 100 years. Commercial halftone screens are based on a discovery made by William Henry Fox Talbot (1800–1877), in 1852. He demonstrated the feasibility of optical halftoning by photographing an image through a loosely woven fabric or ‘screen’. In the 1890s, this process came into practical use when the halftone screen, consisting of two ruled glass plates cemented together, became commercially available. Commercial halftone screens produce an effect of variably sized dots on the photographic plate that gives the illusion of a continuous tone image.

Digital image halftoning, sometimes referred to as spatial dithering, is the process of converting an electronic continuous-tone, gray-scale image to one containing only binary-valued pixels, for the purpose of displaying, printing, or storing the image. The underlying concept is to provide the viewer of the image with the illusion of viewing a continuous-tone image when, in fact, only black and white pixel values are used in the rendering. This process is particularly important in applications such as laser printing, bilevel displays, xerography, and more recently, facsimile.

There are a number of different methods by which this digital image halftoning can be accomplished, which are generally classified as either point or neighborhood processes. A point process is one which computes the output pixel value based strictly on some characteristic of the corresponding input pixel. In contrast, a neighborhood process computes a single output pixel based on a number of pixels in a neighborhood or region of the input image. Ordered dither, which is considered a point process, produces an output by comparing a single continuous-tone input value against a deterministic periodic array of threshold values. If the value of the input pixel under consideration is greater than the corresponding threshold value, the corresponding output pixel is rendered white. If the intensity is less that the threshold, it is rendered black. Dispersed-dot ordered dither is a subset of ordered dither in which the halftone dots are of a fixed size while clustered-dot ordered dither uses variable-sized dots and simulates the variable-sized dots of a printer's halftone screen in the rendering. Among the advantages of point process halftoning, in general and ordered dither halftoning specifically, are simplicity and speed of implementation. The primary disadvantage is that ordered dither produces patterns in the halftoned image, which are visually undesirable.

In contrast, halftoning using the error diffusion algorithm, first introduced by Floyd and Steinberg in 1975, employs neighborhood operations and is currently the most popular neighborhood process. In this approach to halftoning, the output pixel value is determined, not solely by the value of the input pixel and some deterministic pattern, but instead by a weighted average of values from pixels in a neighborhood surrounding the specific pixel being computed. Here, the error of the quantization process is computed and spatially distributed or diffused within a local neighborhood in order to influence future pixel quantization decisions within that neighborhood and thereby improve the overall quality of the halftoned image. In classical unidirectional error diffusion, the image is processed sequentially, proceeding from the upper-left to the lower-right. Starting in the corner of the image, the first pixel is thresholded and the quantizer error is calculated. The error is then diffused to neighboring pixels that have not yet been processed according to the scaling and interconnect defined by the error diffusion filter. The remainder of the pixels in the image are subsequently processed with each quantizer input, now being the algebraic sum of the corresponding original input pixel intensity and the weighted error from previously processed pixels. While error diffusion produces halftone images of superior quality, the unidirectional processing of the algorithm continues to introduce visual artifacts that are directly attributable to the algorithm itself.

In an effort to overcome the implementation constraints of serial processing and improve overall halftone image quality, a number of different parallel architectures have been investigated, including several based on neural network algorithms. While neural network approaches can provide distinct advantages in improved halftone image quality, they also present challenges to hardware implementations, including speed of convergence and the physical interconnect requirements. These challenges, in addition to the natural compatibility with imaging media, provide the motivation for developing optical image processing architectures for these types of applications.

The Error Diffusion Algorithm

Figure 1 shows a block diagram of the error diffusion architecture used in digital image halftoning. The input image is represented as $x_{m,n}$, $w_{m,n}$ is the impulse response of a 2D causal, unity gain error diffusion filter, and the output quantized image is described by $y_{m,n} \in \{-1, 1\}$. The quantizer $q[u_{m,n}]$ provides the one-bit thresholding functionality necessary to convert each analog input pixel to a low-resolution digital output pixel. The quantizer error $\varepsilon_{m,n}$ is computed as the difference between the output and input to the quantizer and distributed to adjacent pixels according to the weighting and interconnection specified by the error diffusion filter. The unity gain constraint ensures

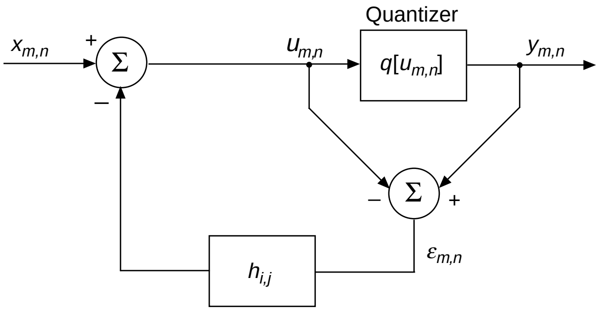

Figure 1 Block diagram of recursive error diffusion architecture.

that no amplification or attenuation of the error signal occurs during the error diffusion process and preserves the average intensity of the image.

In this architecture, the error associated with the quantizer decision at spatial coordinates (m, n) is diffused within a local 2D region to influence adjacent quantization decisions. All of the state variables in this architecture are scalar values since this architecture represents a recursive algorithm in which each pixel in the image is processed sequentially. In the general case, where $w_{m,n}$ is assumed to be a 2D finite impulse response (FIR) filter with coefficients that span a region of support defined by $R_{m,n}$, the quantizer input state $u_{m,n}$ can be written as

$$u_{m,n} = x_{m,n} - \sum_{i,j \in R_{m,n}} w_{i,j} \varepsilon_{m-i,n-j} \qquad [1]$$

Here, $w_{i,j} = 0$, for all $i = j$, since the error signal is diffused spatially within the region of support $R_{m,n}$ and does not influence the pixel under consideration. In this unidirectional error diffusion, the 2D error diffusion filter is therefore noncausal. Figure 2 shows the impulse response of several popular error diffusion filters that have been used in rectangular grid architectures. Figure 2a shows the original error diffusion filter weights developed by Floyd and Steinberg. They argued that the four weights were the smallest number of weights that would produce good halftone images. Figure 2b is a filter that contains a larger region of support developed by Jarvis, Judice, and Ninke in 1976. Finally, the filter in Figure 2c, developed by Stucki, contains the same region of support with coefficients that are a multiple of 2 for digital compatibility and computational efficiency. Here, '•' represents the pixel being processed and the weights describe the local region over which the error is diffused. The normalization factors ensure that the filter coefficients sum to one and therefore meet the unity gain criterion.

To understand the effect of error diffusion and its impact on the quantizer error, it is instructive to describe the effect of error diffusion mathematically. Consider the relationship between the quantizer error $\varepsilon_{m,n} \equiv y_{m,n} - u_{m,n}$ and the overall quantization error $\hat{\varepsilon}_{m,n} \equiv y_{m,n} - x_{m,n}$ in the frequency domain. If we assume that the error is uncorrelated with the input and has statistical properties consistent with a white process, z-transform techniques can be applied to show that the feedback architecture provides the following noise-shaping characteristics:

$$H_{\text{ns}}(z_1, z_2) \equiv \frac{\hat{E}(z_1, z_2)}{E(z_1, z_2)} = 1 - W(z_1, z_2) \qquad [2]$$

Here, $\hat{E}(z_1, z_2)$, $E(z_1, z_2)$, and $W(z_1, z_2)$ represent the z-transforms of $\hat{\varepsilon}_{m,n}$, $\varepsilon_{m,n}$, and $w_{m,n}$, respectively. Equation [2] shows that the noise-shaping characteristic of the error diffusion architecture is directly related to the spectral characteristics of $W(z_1, z_2)$. Appropriate selection of $W(z_1, z_2)$ can spectrally shape the quantizer noise in such a way that minimizes the effect of the low-resolution quantization process on the overall halftoning process. An error diffusion filter with low-pass spectral characteristics produces an overall noise-shaping characteristic $1 - W(z_1, z_2)$, which has the desired high-pass frequency characteristics. This noise-shaping function suppresses the quantization noise within those frequencies occupied by the image and spectrally shapes the noise to high frequencies, which are less objectionable to the human visual system. Circular symmetry is another important characteristic in the error diffusion filter since the human visual system is particularly sensitive to directional artifacts in the image.

$$\left(\tfrac{1}{16}\right)\times \quad \begin{array}{ccc} & \bullet & 7 \\ 3 & 5 & 1 \end{array}$$

(a)

$$\left(\tfrac{1}{48}\right)\times \quad \begin{array}{ccccc} & & \bullet & 7 & 5 \\ 3 & 5 & 7 & 5 & 3 \\ 1 & 3 & 5 & 3 & 1 \end{array}$$

(b)

$$\left(\tfrac{1}{42}\right)\times \quad \begin{array}{ccccc} & & \bullet & 8 & 4 \\ 2 & 4 & 8 & 4 & 2 \\ 1 & 2 & 4 & 2 & 1 \end{array}$$

(c)

Figure 2 Three common error diffusion filters used in rectangular grid architectures. (The '•' represents the origin.)

In the following qualitative description of digital image halftoning, a 348 × 348 natural image of the Cadet Chapel at West Point was selected as the test image. This image provides a variety of important image characteristics, with regions of uniform grayscale, edges, and areas with fine detail. This image was scanned from a high-resolution black and white photograph at 150 dots per inch (dpi) and then printed using a 300 dpi laser printer.

Figure 3 shows the test image of the Cadet Chapel rendered using dispersed-dot ordered dither and results in 256 gray-levels at 300 dpi. Figure 4 shows the halftone image of the Cadet Chapel using the error diffusion algorithm and the Floyd–Steinberg filter coefficients shown in Figure 2a. The unidirectionality of the processing and the causality of the diffusion filter result in undesirable visual artifacts in the halftone image. These include directional hysteresis, which is manifested as 'snakes' running from upper-left to lower-right, and transient behavior near boundaries, which appears as 'shadows' below and to the right of sharp-intensity changes. The directional hysteresis is particularly objectionable in uniform gray intensity areas such as the cloud structure in the upper-left of Figure 4. Similar artifacts are also present in images halftoned using the error diffusion algorithm and other causal diffusion kernels. The logical conclusion to draw from this limited, qualitative analysis is that if we could diffuse the error symmetrically and simultaneously process the entire image in parallel, we could reduce some of these visual artifacts and thereby improve overall halftone image quality.

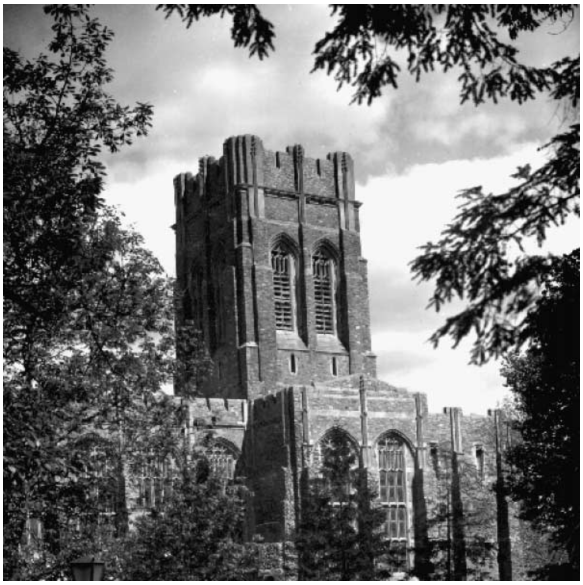

Figure 3 Original gray-scale image of the Cadet Chapel rendered using dispersed-dot ordered dither.

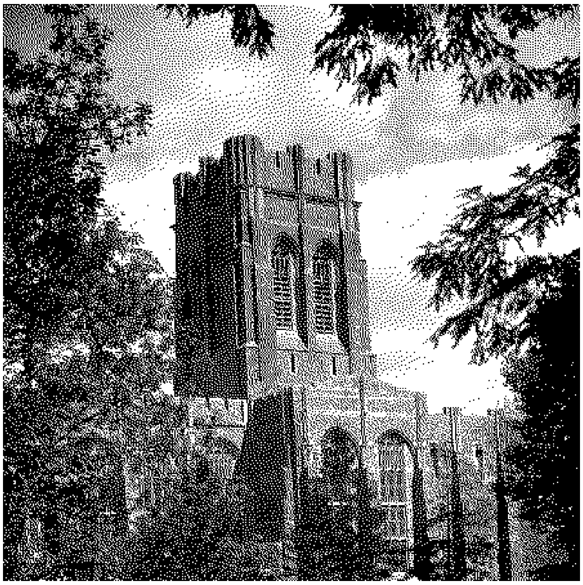

Figure 4 Halftone image of the Cadet Chapel produced using the error diffusion algorithm and the Floyd–Steinberg weights.

The Error Diffusion Algorithm and Neural Networks

The popularity of the neural network-based approach to signal and image processing lies in the ability to minimize a particular performance metric associated with a highly nonlinear system of equations. Specifically, the problem of creating a halftone image can be cast in terms of a nonlinear quadratic optimization problem where the performance measure to be minimized is the difference between the original and the halftone images.

The Hopfield-Type Neural Network

Tank and Hopfield first proposed a mathematical description of the functionality of the neural processing network for signal processing applications. Figure 5 shows an electronic implementation of a four-neuron architecture. Here, a single neuron is comprised of both a standard and inverting amplifier and the synapses or neural interconnections are represented by the physical connections between the input and output of the amplifiers. If the input to amplifier i is connected to the output of amplifier j by a resistor with resistance $R_{i,j}$, the amplitude of the connection $T_{i,j}$ is the conductance $1/R_{i,j}$.

The dynamic behavior of an N-neuron Hopfield-type neural network can be described by the following system of N nonlinear differential equations:

$$c\frac{\mathrm{d}u_i(t)}{\mathrm{d}t} = -u_i(t) + \sum_j T_{i,j}\Im[u_j(t)] + x_i \qquad [3]$$

where $i = 1, 2, \ldots N$, $\Im[\cdot]$ is a monotonically increasing sigmoid function, x_i is an input vector containing N

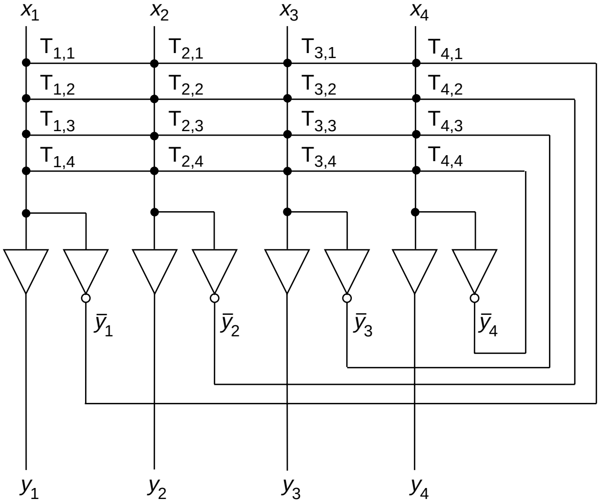

Figure 5 Electronic realization of a four-neuron Hopfield-type neural network.

elements, and c is a scaling factor. In equilibrium, eqn [3] implies that

$$u_i = x_i + \sum_j T_{i,j}\Im[u_j] \qquad [4]$$

Hopfield showed that when the matrix of interconnection weights $\mathbf{T}$ is symmetric with zero diagonal elements and the high-gain limit of the sigmoid $\Im[\cdot]$ is used, the stable states of the N functions $y_i(t) = \Im[u_j]$ are the local minima of the energy function:

$$E = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{T}\mathbf{y} - \mathbf{x}^{\mathrm{T}}\mathbf{y} \qquad [5]$$

where $\mathbf{y} \in \{-1, 1\}$ is the N-vector of quantized states. As a result, if the neural network can be shown to be stable, as the network converges the energy function is minimized. In most neural network applications, this energy function is designed to be proportional to a performance metric of interest and therefore as the network converges, the energy function and consequently the performance metric is also minimized.

The Error Diffusion Neural Network

An understanding of the operation and characteristics of the Hopfield-type neural network can be applied to the development and understanding of a 2D extension of the error diffusion algorithm. Figure 6 shows an electronic implementation of a four-neuron error diffusion-type neural network, where the individual neurons are represented as amplifiers and the synapses by the physical connections between the input and output of the amplifiers.

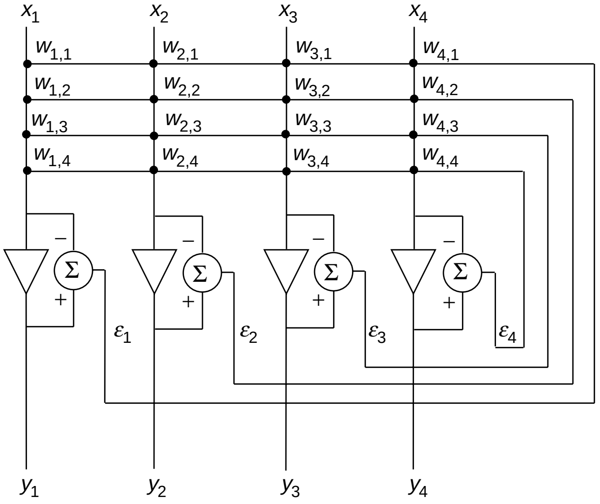

Figure 6 Four-neuron electronic implementation of the error diffusion neural network.

In equilibrium, the error diffusion neural network can be shown to satisfy

$$\mathbf{u} = \mathbf{W}(\mathbf{y} - \mathbf{u}) + \mathbf{x} \qquad [6]$$

Here, $\mathbf{u}$ is the state vector of neuron inputs, $\mathbf{W}$ is a matrix containing the interconnect weights, $\mathbf{x}$ is the input state vector, and $\mathbf{y}$ is the output state vector. For an $N \times N$ image, $\mathbf{W}$ is an $N^2 \times N^2$ sparse, circulant matrix derived from the original error diffusion weights $w_{m,n}$. If we define the coordinate system such that the central element of the error diffusion kernel is $(i, j) = (0, 0)$, then the matrix $\mathbf{W}$ is defined as $W(i, j) = -w[(j - i) \text{ div } N, (j - i) \bmod N]$ where:

$$x \text{ div } y = \left\lfloor \frac{x}{y} \right\rfloor \text{ if } xy \geq 0 \text{ or } \left\lceil \frac{x}{y} \right\rceil \text{ if } xy < 0 \qquad [7]$$

An equivalence to the Hopfield network can therefore be described by

$$\mathbf{u} = \mathbf{A}(\mathbf{W}\mathbf{y} + \mathbf{x}) \qquad [8]$$

where $\mathbf{A} = (\mathbf{I} + \mathbf{W})^{-1}$. Effectively, the error diffusion network includes a pre-filtering of the input image $\mathbf{x}$ by the matrix $\mathbf{A}$ while still filtering the output image $\mathbf{y}$ but now with a new matrix, $\mathbf{AW}$. Recognizing that $\mathbf{AW} = \mathbf{I} - \mathbf{A}$ and adding the arbitrary constant $k = \mathbf{y}^{\mathrm{T}}\mathbf{y} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, we can write the energy function of the error diffusion neural network as

$$\hat{E}(x, y) = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \qquad [9]$$

This energy function is a quadratic function, which can be factored into

$$\hat{E}(x, y) = [\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}]^{\mathrm{T}}[\mathbf{B}\underbrace{(\mathbf{y} - \mathbf{x})}_{\text{error}}] \qquad [10]$$

where $\mathbf{A} = \mathbf{B}^{\mathrm{T}}\mathbf{B}$. From eqn [10] we find that as the error diffusion neural network converges and the energy function is minimized, so too is the error between the output and input images.

If the neurons update independently, the convergence of the error diffusion network is guaranteed if

$$\forall k : [\mathbf{AW}]_{k,k} \geq 0 \qquad [11]$$

We find in practice, that even in a synchronous implementation, the halftoned images converge to a solution, which results in significantly improved halftone image quality over other similar halftoning algorithms.

The Error Diffusion Filter

The purpose of the error diffusion filter is to spectrally shape the quantization noise in such a way that the quantizer error is distributed to higher spatial frequencies, which are less objectionable to the human visual system. In this application, a feedback filter for the error diffusion neural network was designed using conventional 2D filter design techniques resulting in the 7 × 7 impulse response shown in Figure 7. Figure 8 shows the same image of the Cadet Chapel halftoned using the error diffusion neural network algorithm and the 2D error diffusion filter shown in Figure 7. There is clear improvement in halftone image quality over the image produced using sequential error diffusion in Figure 4. Notice particularly the uniform distribution of pixels in the cloud formation in the upper-left of Figure 8. Also noteworthy is the improvement around the fine detail portions of the tree branches and next to the vertical edges of the chapel.

0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0068	0.0328	0.0766	•	0.0766	0.0328	0.0068
0.0051	0.0248	0.0583	0.0766	0.0583	0.0248	0.0051
0.0019	0.0103	0.0248	0.0328	0.0248	0.0103	0.0019
0.0003	0.0019	0.0051	0.0068	0.0051	0.0019	0.0003

Figure 7 Impulse response of a 7 × 7 symmetric error diffusion filter. ('•' represents the origin.)

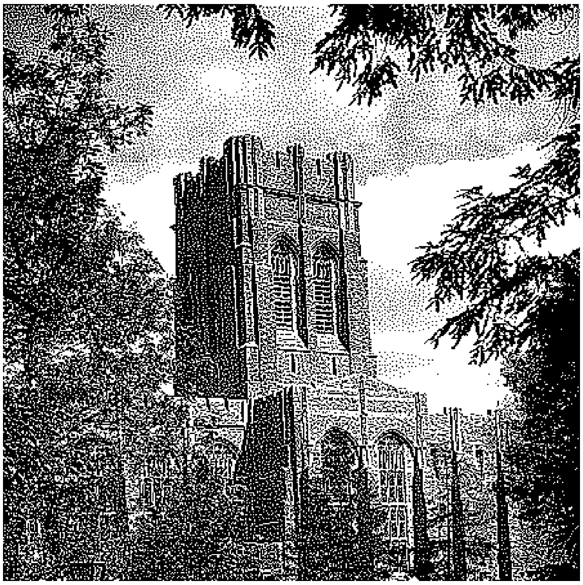

Figure 8 Halftone image of the Cadet Chapel using the error diffusion neural network and the 7 × 7 symmetric error diffusion filter shown in Figure 7.

Smart Pixel Technology and Optical Image Processing

The concept underlying smart pixel technology is to integrate electronic processing and individual optical devices on a common chip, to take advantage of the complexity of electronic processing circuits and the speed of optical devices. Arrays of these smart pixels then allow the advantage of parallelism that optics provides. There are a number of different approaches to smart pixel technology that differ primarily in the kind of opto-electronic devices, the type of electronic circuitry, and the method of integration of the two. Common opto-electronic devices found in smart pixels include: light-emitting diodes (LEDs); laser diodes (LDs); vertical cavity surface emitting lasers (VCSELs); multiple quantum well (MQW) modulators; liquid crystal devices (LCDs); and photodetectors (PDs). The most common types of electronic circuitry are silicon-based semiconductors such as complementary metal oxide semiconductors (CMOS) and compound semiconductors such as gallium arsinide (GaAs)-based circuitry. There are a number of different approaches to smart pixels, which generally differ in the way in which the electronic and optical devices are integrated. Monolithic integration, direct epitaxy, and hybrid integration are the three most common approaches in use today.

Smart pixel technology provides a natural methodology by which to implement optical image processing architectures. The opto-electronic devices provide a natural optical interface while the electronic circuitry provides the ability to perform either analog or digital computation. It is important to understand that any functionality that can be implemented in electronic circuitry can be integrated into a smart pixel architecture. The only limitations arise from physical space constraints imposed by the integration of the opto-electronic devices with the electronic circuitry. While smart pixels can be fabricated with digital or analog circuitry, the smart pixel architecture described subsequently uses mixed-signal circuitry.

A Smart Pixel Implementation of the Error Diffusion Neural Network

A smart pixel hardware implementation of the error diffusion neural network provides the potential to simultaneously achieve the computational complexity of electronic circuitry and the parallelism and high-speed switching of optics. The specific smart pixel architecture described here integrates MQW modulators called self electro-optic effect devices (SEEDs) with silicon CMOS

VLSI circuitry, using a hybrid integration approach called flip-chip bonding. This type of smart pixel is commonly referred to as CMOS-SEED smart pixel technology.

To provide an example of the application of smart pixel technology to digital image halftoning, a 5 × 5 CMOS-SEED smart pixel array was designed and fabricated. The CMOS circuitry was produced using a 0.5 μm silicon process and the SEED modulators were subsequently flip-chip bonded to silicon circuitry using a hybrid integration technique. The central neuron of this new smart pixel array consists of approximately 160 transistors while the complete 5 × 5 array accounts for over 3,600 transistors. A total of 50 optical input/output channels are provided in this implementation.

Figure 9 shows the circuitry associated with a single neuron of the smart pixel error diffusion neural network. All state variables in each circuit of this architecture are represented as currents. Beginning in the upper-left of the circuit and proceeding clockwise, the input optical signal incident on the SEED is continuous in intensity and represents the individual input analog pixel intensity. The input SEED at each neuron converts the optical signal to a photocurrent and subsequently, current mirrors are used to buffer and amplify the signal. The width-to-length ratio of the metal oxide semiconductor field effect transistors (MOSFETs) used in the CMOS-SEED circuitry provide current gain to amplify the photocurrent. The first circuit produces two output signals: $+I_u$ which represents the state variable $u_{m,n}$ as the input to the quantizer and $-I_e$ which represents the state variable $-\varepsilon_{m,n}$ as the input to the feedback differencing node. The function of the quantizer is to provide a smooth, continuous threshold functionality for the neuron that produces the output signal I_{out}, corresponding to the output state variable $y_{m,n}$. This second electronic circuit is called a modified wide-range transconductance amplifier and produces a hyperbolic tangent sigmoidal function when operated in the sub-threshold regime. The third circuit takes as its input I_{out}, produces a replica of the original signal, and drives

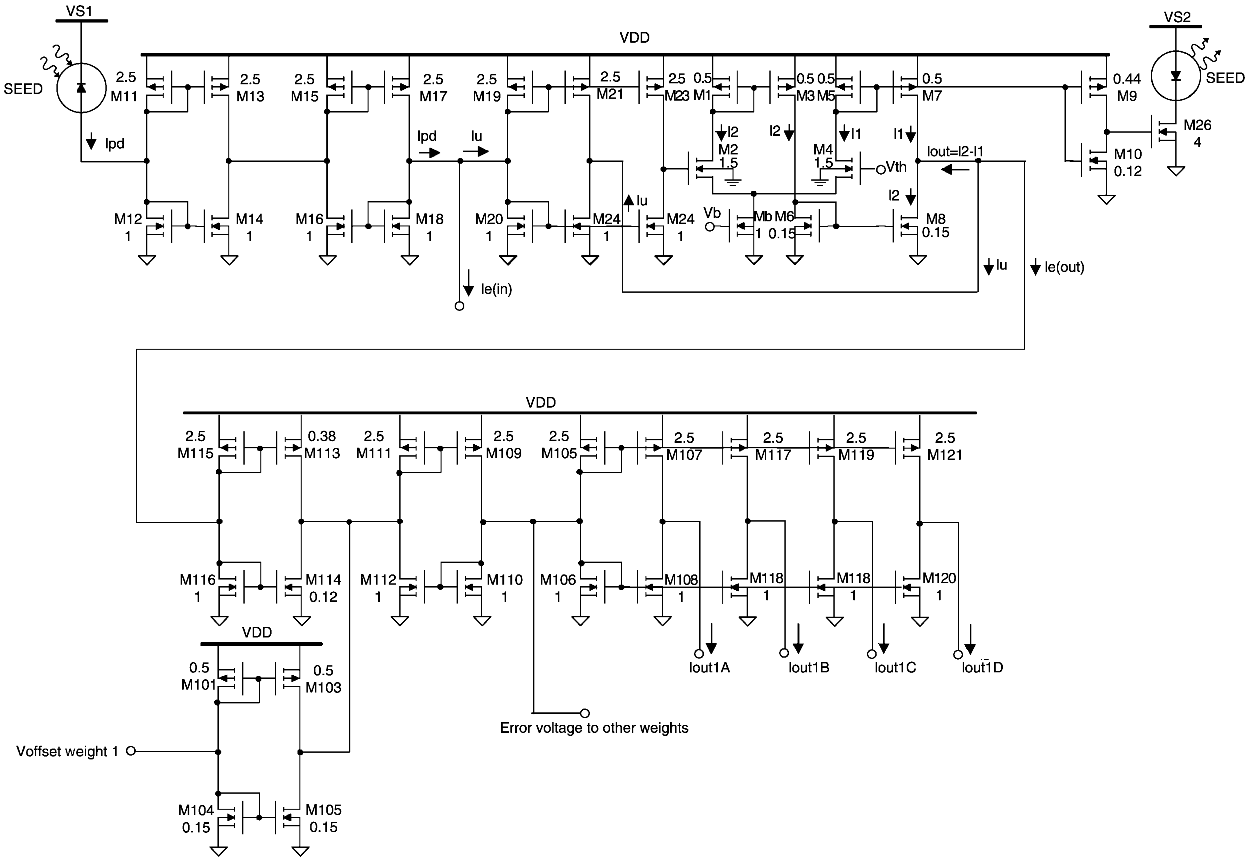
VS1
SEED
2.5 M11
Ipd
2.5 M13
2.5 M15
2.5 M17
2.5 M19
2.5 M21
2.5 M23
0.5 M1
0.5 M3
0.5 M5
0.5 M7
0.44 M9
VS2
SEED
VDD
Ipd
Iu
I2
M2 1.5
I2
I1
M4 1.5
Vth
I1
Iout=I2-I1
M26 4
M10 0.12
Iu
I2
M12 1
M14 1
M16 1
M18 1
M20 1
M24 1
M24 1
Vb
Mb 1
M6 0.15
M8 0.15
Iu
Ie(out)
Ie(in)
VDD
2.5 M115
0.38 M113
2.5 M111
2.5 M109
2.5 M105
2.5 M107
2.5 M117
2.5 M119
2.5 M121
M116 1
M114 0.12
M112 1
M110 1
M106 1
M108 1
M118 1
M118 1
M120 1
VDD
0.5 M101
0.5 M103
Voffset weight 1
M104 0.15
M105 0.15
Iout1A
Iout1B
Iout1C
Iout1D
Error voltage to other weights

Figure 9 Circuit diagram of a single neuron and a single error weight of the 5 × 5 error diffusion neural network based on a CMOS-SEED-type smart pixel architecture. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.6, p. 222. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

the output optical SEED. In this case, the output optical signal is a binary quantity represented as the presence or absence of light. When the SEED is forward-biased, light is generated through electroluminescence. The last circuit at the bottom of the schematic implements a portion of the error weighting and distribution function of the error diffusion filter. The individual weights are implemented by scaling the width-to-length ratio of the MOSFETs to achieve the desired weighting coefficients. The neuron-to-neuron interconnections are accomplished using the four metalization layers of the 0.5 μm silicon CMOS process. In this design, the error diffusion filter was limited to a 5 × 5 region of support because of the physical constraints of the circuitry necessary to implement the complete error diffusion neural network in silicon circuitry. The impulse response of the 5 × 5 filter used in this particular smart pixel architecture is shown in Figure 10. The error weighting circuitry at the bottom of Figure 9 represents only the largest weight (0.1124) with interconnects to its four local neighbors ($I_{\mathrm{outA}} - I_{\mathrm{outD}}$).

0.0023	0.0185	0.0758	0.0185	0.0023
0.0185	0.0254	0.1124	0.0254	0.0185
0.0758	0.1124	•	0.1124	0.0758
0.0185	0.0254	0.1124	0.0254	0.0185
0.0023	0.0185	0.0758	0.0185	0.0023

Figure 10 Error diffusion filter coefficients used in the 5 × 5 CMOS-SEED error diffusion architecture. ('•' represents the origin.)

Figure 11 shows a photomicrograph of a 5 × 5 error diffusion neural network, which was fabricated using CMOS-SEED smart pixels. Figure 12 shows a photomicrograph of a single neuron of the 5 × 5 CMOS-SEED neural network. The rectangular features are the MQW modulators while the silicon circuits are visible between and beneath the modulators. The MQW modulators are approximately 70 μm × 30 μm and have optical windows, which are 18 μm × 18 μm.

Simulations of this network predict individual neuron switching speeds of less than 1 μs, which corresponds to network convergence speeds capable of providing real-time digital image halftoning. Individual component functionality and dynamic operation of the full 5 × 5 neural array were both experimentally characterized. Figure 13 shows CCD images of the operational 5 × 5 CMOS-SEED smart pixel array. The white spots show SEED MQW modulators emitting light. Figure 13a shows the fully functioning array while Figure 13b shows the 5 × 5 CMOS-SEED neural array under 50% gray-scale input. Here 50% of the SEED modulators are shown to be in the *on*-state demonstrating correct operation of the network and the halftoning architecture.

Both the simulations and the experimental results demonstrate that this approach to a smart pixel implementation of the error diffusion neural network provides sufficient accuracy for the digital halftoning application. The individual neuron switching speeds also demonstrate the capability for this smart pixel hardware implementation to provide real-time halftoning of video images.

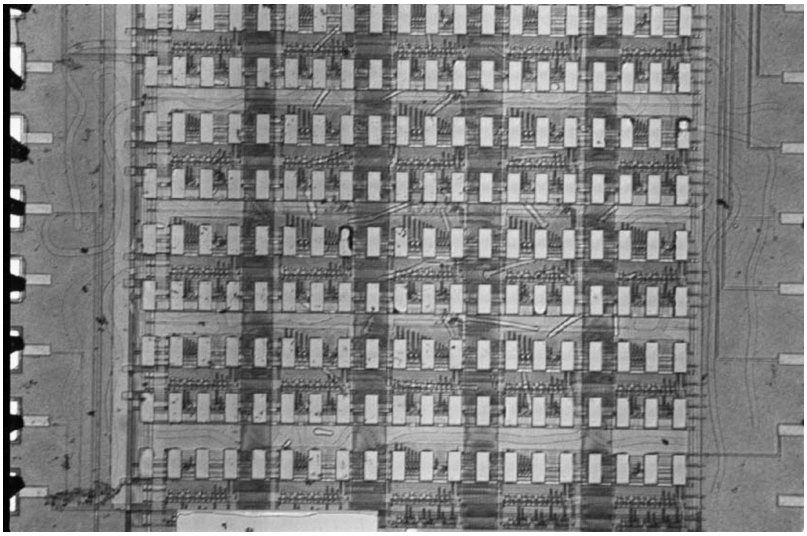

Figure 11 Photomicrograph of a smart pixel implementation of a 5 × 5 CMOS-SEED error diffusion neural network for digital image halftoning. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.5, p. 221. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

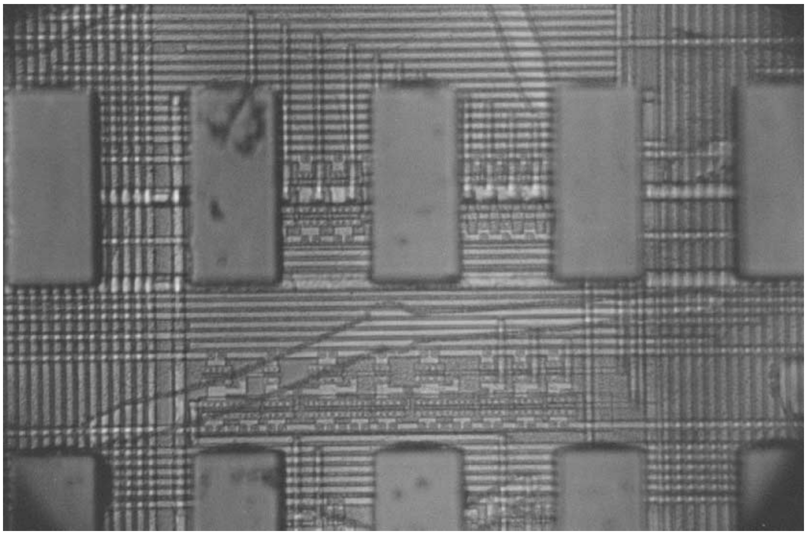

Figure 12 Photomicrograph of a single neuron of the 5 × 5 error diffusion neural network. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Fig. 8.4, p. 220. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

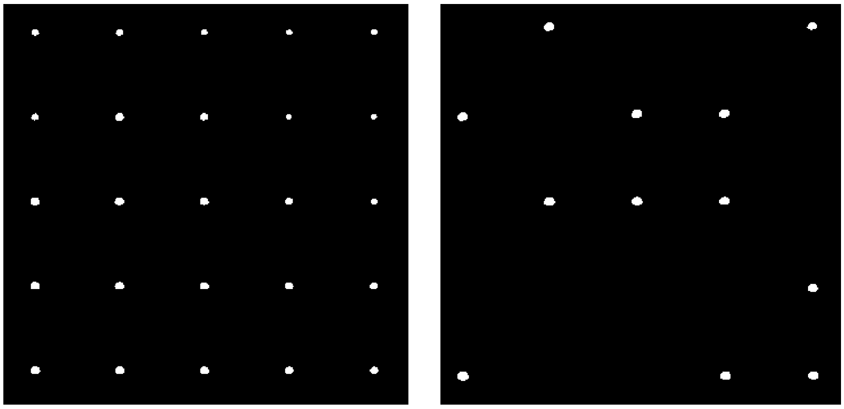

(a) (b)

Figure 13 CCD images of the 5 × 5 CMOS-SEED array. (a) Fully-operational array, and (b) under 50% gray scale illumination. (Reprinted with permission from Shoop BL, *Photonic Analog-to-Digital Conversion*, Springer Series in Optical Sciences, Vol. 81 [1], Figs. 8.19 and 8.20, p. 236. Copyright 2001, Springer-Verlag GmbH & Co. KG.)

Image Processing Extensions

Other image processing functionality is also possible by extending the fundamental concepts of 2D, symmetric error diffusion to other promising image processing applications. One important application includes color halftoning while other extensions include edge enhancement and feature extraction. In the basic error diffusion neural network, the error diffusion filter was specifically designed to produce visually pleasing halftoned images. Care was taken to ensure that the frequency response of the filter was circularly symmetric and that the cutoff frequency was chosen in such a way as to preserve image content. An analysis of the error diffusion neural network shows that the frequency response of this filter directly shapes the frequency spectrum of the output halftone image and therefore directly controls halftone image content. Other filter designs with different spectral responses can provide other image processing features such as edge enhancement, which could lead to feature extraction and automatic target recognition applications.

See also

Detection: Smart Pixel Arrays. **Information Processing:** Optical Neural Networks. **Optical Processing Systems**.

Further Reading

Dudgeon DE and Mersereau RM (1984) *Multidimensional Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.
Lau DL and Arce GR (2001) *Modern Digital Halftoning*. New York: Marcel Dekker.
Lim JS (1990) *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice Hall.
Mead CA (1989) *Analog VLSI and Neural Systems*. New York: Addison-Wesley.
Poon T-C and Banerjee PP (2001) *Contemporary Optical Image Processing with Matlab*. New York: Elsevier.
Russ JC (1998) *The Image Processing Handbook*. Boca Raton, FL: CRC Press.
Shoop BL (2001) *Photonic Analog-to-Digital Conversion*. Springer Series in Applied Optics, vol. 81. Berlin: Springer-Verlag.
Shoop BL, Sayles AH and Litynski DM (2002) New devices for optoelectronics: smart pixels. In: De Cusatis (ed.) *Fiber Optic Data Communication: Technological Trends and Advances*, pp. 352–421. San Diego, CA: Academic Press.
Ulichney R (1987) *Digital Halftoning*. Cambridge, MA: The MIT Press.
Wax C (1990) *The Mezzotint: History and Technique*. New York: Harry N. Abrahms, Inc.

Optical Neural Networks

H J Caulfield, Fisk University, Nashville, TN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

An artificial neural network (ANN) is a nonlinear signal processing system based on the neural processes observed in animals. Usually they have multiple inputs and often multiple outputs also. Conventionally, each input sends its signal to many neurons, and each neuron receives signals from many inputs. The neuron forms an intermediate sum of weighted inputs and transforms that sum, according to some nonlinearity, to form an output signal. Often the output signal from one neuron is used as the input signal for another in feed forward, feedback, or mixed mode complex neural networks.

ANNs are of value in helping us understand how biological neural networks perform and in a variety of technical applications, most of which involve pattern association. An auto-associative neural network, that associates patterns with themselves, can have useful noise cleanup or pattern completion properties. A hetero-associative neural network is often used in pattern recognition to associate input patterns with preselected class labels.

Neural Networks – Natural and Artificial

The most primitive one-celled organisms have no nerve cells, but they do exhibit various chemical means for interacting with their environments, within their own cells, and among individuals of the same or different types. As organisms increased in size, a common chemical environment became impractical, and nerves arose as a means for long-distance communication within larger organisms. One thing seen in evolution is that widespread useful adaptations are often modified but seldom discarded. Thus, it should not be surprising to find that chemical interactions play a major role in the most complex thing we know of – the human brain. Chemicals play a huge role in making us who we are, setting our moods, and so forth. Realizing this, in the late twentieth century, psychiatrists started to make serious progress in treating mental diseases through the use of chemistry.

There are a variety of processes that we use as humans, but the primary means is neural. Only some animals and no plants have neurons, but soon after neurons evolved, they began to concentrate in special regions of the body that developed into primitive brains. Not only were there many neurons, but also they were densely and complexly interconnected to form neural networks.

Neural networks did not replace chemical interactions, they used and supplemented them. Chemical communication across the synaptic cleft dominates communication between networks, but the neural networks also control chemical production and distribution within the brain, and those chemicals change the behavior of the neurons that cause their secretion. The functioning of real brains is incredibly complex, possibly irreducibly complex in terms of current mathematics and logic.

These things are noted to avoid a confusion that is distressingly common. Technologists produce simplified ANNs and then try to interpret the behavior of brains in those terms. Quite often, ANNs yield useful insight into brain behavior but seldom offer a detailed account of them. Reality is far more complex than any ANN. Current mathematics provides no means to describe continuous sets of events, each of which both causes and is caused by the others.

ANNs themselves have grown much more complicated as well. Starting out as simple, feed-forward, single-layer systems, they rapidly became bidirectional and even multilayered as well. Higher-order neural networks processed the input signals before inserting them into the neural network. Pulse coupled neural networks (PCNNs) arose that more closely modeled the pulsations in real brains. Learning was broken into learning with instruction, following example, and self-organizing, the latter leading to self organized maps and adaptive resonance theory, and many related subjects. These are but crude pointers to the wealth and variety of concepts now part of the ANN field.

The history of ANNs is both complicated and somewhat sordid. This is not the occasion to retrace the unpleasant things various members of this community have done to others in their quests for stature in the field. One result of that history is a dramatically up-and-down pattern of interest in and support for ANNs. The first boom followed the development of a powerful single-layer neural network called the Perceptron by its inventor Rosenblatt. Perceptrons were biologically motivated, fast, and reasonably effective; so interest in the new field of ANNs was high. That interest was destroyed by a book on the field arguing that Perceptrons were only linear discriminants and most interesting problems are not linearly discriminable. Ironically, the same argument has been made about optical Fourier correlators, and it is true there as well. But optical processing folk largely ignored the problem and kept working on it. Sociologists of science might enjoy asking why reactions differed so much in the two fields. Funding and interest collapsed and the field went into hibernation.

It was widely recognized that multilayer ANNs would comprise nonlinear discriminants, but training them presented a real problem. The simple rewards and punishment of weights depending on performance was easy when each output signal came from a known set of weighted signals that could be rewarded or punished jointly, according to how the output differed from the sought-after output. With a multilayer perceptron, credit assignment of inner layers is no longer simple. It was not until Werbos invented 'backward error propagation' that a solution to this problem was available and the field started to boom again. Also key to the second boom were enormously popular articles by Hopfield on rather simple auto- and hetero-associative neural networks. Basically, those networks are almost never used, but they were of great value in restarting the field. In terms of research and applications, we are still in that second boom, that began during the 1980s.

> An ANN is a signal processor comprised of a multiplicity of nodes which receive one or more input signals and generate one or more output signals in a nonlinear fashion wherein those nodes are connected with one or two way signaling channels.

Note also that biological inspiration is no longer a part of the definition. Most ANNs used for association are only vaguely related to their biological counterparts.

Optical Processors

Optics is well-suited for massive parallel interconnections, so it seems logical to explore optical ANNs. The first question to address is "Why bother to use optical processors at all?" as electronic digital computers seem to have everything:

- Between the writing of this article (on an electronic computer) and its publication, the speed of electronic computers will almost double and they will become even cheaper and faster.
- They can have arbitrary dynamic range and accuracy.
- Massively improved algorithms also improve their speed. For instance, the wonderful fast Fourier transform (FFT) that opticists have used to simulate optics for decades has been replaced by an even faster fastest Fourier transform in the West (FFTW). And for wavelets, the fast wavelet transform (FWT) is even faster.

Optical processors, on the other hand, have severely limited analog accuracy and few analog algotectures or archirithms. They are specialized systems (equivalent in electronics, not to computers but to application specific integrated circuits (ASICs)). They are usually larger and clumsier than electronic systems and always cost more.

There are several reasons why optics may still have a role in some computations. First, many things on which we wish to perform computations are inherently optical. If we can do the computations optically before detection, we can avoid some time and energy and sensitivity penalties inflicted, if we first detect the signals then process them electronically, and then (sometimes) convert them back to optics. Examples include spectroscopy and optical communication.

Second, pure optics is pure quantum mechanics. Input of data and setup of the apparatus is experiment preparation. Readout of data is classical measurement. Nature does the rest free of charge. There is a unitary transformation from input to output that requires no human intervention and is, in fact, destroyed by human intervention. Those intermediate (virtual) computations require no energy and no time, beyond the time taken for light to propagate through the system. Speed and power consumption are limited only by the input and output. This applies for everything from Fourier optics to digital optics to logic gates. Thus, optics can have speed and power consumption advantages over electronics.

Third, if quantum computers involving entanglement ever emerge, they may well be optical. Such computers would be used for nondeterministic solutions of hard problems. That is, they would explore all possible paths at once and see which one does best. This is something we can already do electronically or optically for maze running (using classical, not quantum entanglement) means.

Fourth, most optical processors are analog, which is sometimes superior to digital. In those cases, we would still have to choose between analog optics and analog electronics, of course.

So, optical processing (like all other processing) is a niche market. But some of the niches may be quite large and important. It is not a universal solution, but neither is digital electronics.

The Major Branches of Optical Processing and Their Corresponding Neural Networks

Every branch of optical processing leads to its own special purpose optical neural networks, so those branches become a good place to start an article such as this one.

Optical Linear Algebra

Many neural network paradigms use linear algebra followed by a point-by-point nonlinearity in each of multiple stages. The linear operation is a matrix–vector multiplication, a matrix–matrix multiplication, or even a matrix–tensor multiplication. In any case, the key operations are multiplication and addition – operations very amenable to performance with analog optics. With effort, it is possible to combine simple electronics with analog optics operating on non-negative signals (the amount of light emitted from a source, the fraction of light passing through a modulator, etc.) to expand to real or even complex numbers. With even more effort, we can use multiple analog non-negative signals to encrypt a digital signal, allowing digital optical processors. Almost all optical neural networks using linear algebra use only real numbers. In most of these, the real number is encoded as a positive number and a negative number. Data are read in parallel using spatial light modulators (SLMs), source arrays, acousto-optic modulators, etc. They are operated upon by other SLMs, acousto-optic systems, and so forth. Finally, they are detected by individual detectors or detector arrays. The required nonlinearities are then done electronically. It is a theory that holds for all linear algebraic processes up to computational complexity $O(N^4)$ that they can be performed with temporal complexity $O(N^0)$ if the extra dimensions of complexity are absorbed in space, up to $O(N^2)$ through 2D arraying and in fanin/fanout, the ability of multiple beams of light to be operated upon in parallel by the same physical modulator – this has the same complexity as the 2D array of modulators, namely $O(N^2)$. Thus, we can invert a matrix, an order $O(N^4)$ for an $N \times N$ matrix, $O(N^0)$ in time, that is, independently of the size of N, as long as that matrix is accommodated by the apparatus. This offers a speed electronics cannot match.

Most of the optical neural networks of the mid-1980s had optical vector–matrix multipliers at their heart. Most were a few layers of feed forward systems, but by the 1990s feedback had been incorporated as well. These are typically N inputs connected to N outputs through an $N \times N$ matrix, with N varying from 100 to 1000. With diffuser interconnection and a pair of SLMs, it is possible to connect an $N \times N$ input with an $N \times N$ output using N^4 arbitrary weights, but that requires integrating over a time up to N^2 intervals. Most of the N^4-dimensional matrices can be approximated well by fewer than N^4 terms using singular value decomposition (SVD) and (as this version of the interconnection uses outer products), that reduces the number of integration times considerably in most cases. We have shown, for instance, that recognizing M target images can be accomplished well with only M terms in the SVD.

Coherent Optical Fourier Transformation

It took opticists several years, during the early 1960s, to realize that if Fourier mathematics was useful in describing optics, optics might be useful in performing Fourier mathematics. The first applications of optical Fourier transforms were in image filtering and particularly in pattern recognition. Such processors are attractive not only for their speed but also for their ability to locate the recognized pattern in 1D, 2D, or even 3D. It was soon obvious that a Fourier optical correlator was a special case of a single layer perceptron. Of course, it inherits the weaknesses of a single layer perceptron too. Soon, workers had added feedback to create a winner-takes-all type of auto-associative optical neural network, a system that associated images with themselves. This proved useful in restoring partial or noisy images. It was not until 2003 that we learned how to make powerful hetero-associative networks with Fourier optical systems. This creates a nonlinear discrimination system that maintains the target location and generalizes well from a few samples of multiple classes to classify new data quite accurately.

Imaging

Imaging is the central act of optics and might seem to be of little value in optical processing. But multiple imaging (with lenslets, self-focusing fibers, holograms, and the like) has been the basis of numerous optical neural networks. Multiple imaging is a nonrandom way of distributing information from one plane to the next. Diffusers have also been used to produce fully interconnected optical neural networks. If each broad image point contains the whole image (something easily arranged with a diffuse hologram), then we have an auto-associative memory. Different such auto-associative networks can be superimposed in a single hologram. Then feedback after nonlinear image filtering can be used to make a general auto-associative memory.

Even deliberately poor imaging has been used. A blurred or out-of-focus image can accomplish the lateral signal spreading needed to facilitate a PCNN. These networks are substantially more like real brain networks than those discussed so far. It starts with a 2D array of integrate-and-fire units. The firing rate of each depends only in the incident power (how long integration to the thresholding firing takes). But if the fired signal is spread to its neighbors, the firing of a neighbor can hasten any unit's firing. This causes firings to synchronize and results in 'autowaves' of firing patterns moving across the 2D plane. Thus an incident 2D image is converted to a 3D signal – the two spatial dimensions and time. This results, for all but some very simple cases, in a chaotic signal occupying a strange attractor. It is interesting but not very revealing in itself. But suppose we 'integrate out' either the two spatial dimensions or the time dimension. If we integrate out the spatial dimensions, we are left with a time signal. Its strange attractors are remarkably useful in pattern recognition, because they are more syntactic than statistical (in pattern recognition terms). They describe the shape as similar, independently of the size, rotation, perspective, or even shading of the pattern! If we integrate out the time dimension, we restore the original image – almost. In fact, we can use this aspect of PCNNs for optimum blind segmentation of images. The autowaves can be made to 'leave tracks' so they can be reversed, which allows them to solve maze problems nondeterministically, because they can take all the paths. When the first wave exits, we simply retrace it to find the shortest path through the maze – a prototypical nondeterministic polynomial operation.

Holograms

Holograms are the most general, complex, and flexible means of light manipulation so far invented. Naturally, they have many applications in optical neural networks. One such application is the interconnection of each of an $N \times N$ array of inputs to each of an $N \times N$ array of outputs with its own unique connection strength (weight in the language of neural networks). The practical limit seems to be around $N = 1000$, yielding 10^{12} weights, roughly the same number as in the human brain! The brain, however, is not fully interconnected, so this produces an entirely different type of neural network. The resulting network would have fixed weights, but it could perform a variety of useful operations and be both forward and backward connected.

Special on-axis holograms, called diffractive optical elements, can be mass manufactured very effectively and inexpensively to perform tasks such as multiple imaging.

Photorefractive Neural Networks

A very complex thing happens when a pattern of fringes, such as those that form an interference pattern for hologram recording, strike materials such as lithium niobate, strontium barium niobate, gallium arsenide, and many others that fall into this category of 'photorefractive' materials. Of course, 'photo' refers to light and 'refractive' refers to the speed of light in a material – the speed of light in a vacuum divided by the index of refraction. These materials are photoconductive, so the light releases charge carriers in what would otherwise be an electrical insulator. The periodic charge pattern sets up a periodic space charge within the material. But the material is also electro-optic, that is, its index of refraction changes with applied electric field, thus generating a periodic index of refraction grating in the material. Finally, that index of refraction pattern (a phase hologram) diffracts the incoming light. Usually, but not always, a steady-state situation is reached after time that depends on the material and the photon density, although chaos can arise under some conditions. This makes for a dynamic hologram, one that can change or adapt over time as the input pattern changes. Once a satisfactory photorefractive hologram is formed, it can be stabilized (fixed) if desired. Most optical neural network uses envision a dynamic situation with changing interconnections.

What has not been discussed above is that the index of refraction varies with the light's electric field vector cubed, so photorefractives provide a convenient way to accomplish such nonlinear operations as winner-takes-all.

Sometimes, we can record holograms with photorefractives that cannot be recorded by conventional holograms in conventional materials. To explain this, a hologram is a transducer between two optical wavefronts – usually called the reference and object beams. If the reference beam is incident on the hologram, it is (fully or partially) converted into the object beam. If the precisely reversed (phase conjugated) reference beam is incident on it, the precisely reversed (phase conjugated) object beam is derived from it. Likewise, the object wavefront (or its phase conjugate) produces the reference wavefront (or its phase conjugate). The hologram is made by recording the interference pattern between the two wavefronts: the reference and object beams. The ability of two wavefronts to form a temporally stable interference pattern at some point in space is called the mutual coherence of those wavefronts. Usually, both wavefronts are derived from the same laser beam to make achieving high mutual coherence possible. And, of course, both beams must be present to form an interference pattern. In some cases of photorefractive holograms, all of those conditions can be violated. The two wavefronts can derive from two sources that are different in wavelength, polarization, etc., and need not be simultaneously present. This phenomenon, called mutual phase conjugation, converts each wavefront into the phase conjugate of the other, but for effects due to any differences in wavelength and polarization. This allows new freedoms that have been exploited in optical neural networks.

Photorefractive materials also tend to be piezoelectric. It has been found that up to about 20 totally independent reflective holograms can be stored with different applied electric field in lithium niobate. Changing electric fields so changes the fringe spacings and indices of refraction, that only one of those holograms has any appreciable diffraction efficiency at any one electric field value. The result is an electrically selectable interconnection pattern. The same hologram can contain the information needed for many neural networks or multiple layers of fewer of them. There is no need, in general, to have more than an input layer, an output layer, and two intermediate or hidden layers.

Others have used photorefractive material to implement learning in neural networks. In biological neural networks the synaptic strengths (the weights between neurons) continue to adapt. With photorefractive holograms, we can achieve this in an optical ANN.

As previously discussed, a key development in the history of ANNs was a way to train multilayer neural networks and thus accomplish nonlinear discrimination. The first, and still most popular, way to do this is called backward error propagation. Several groups have been implemented in photorefractive optical neural networks.

Conclusions

Essentially every major type of neural network has been implemented optically. Every major tool of optics has found its way into these optical neural networks. Yet, it remains the case that almost every neural network used for practical applications is electronic. The mere existence proof of a use for optical neural networks may be what the field needs to move forward from the demonstration stage to the application stage. The tools exist and there are niches where optics seems more appropriate than electronics. Hopefully, the next review of the field will include a number of commercial successes. Nothing spurs science better than the funds that follow successful applications.

List of Units and Nomenclature

Adaptive resonance theory (ART)	Another self-organized neural net categorizer that works by mutual adjustment between bottom-up and top-down neural connections.
Computational complexity	The measure of how the number of required calculations scales with the problem size, usually denoted by N. The scaling is represented by the Big O form of (N), where $f(n)$ shows how it scales asymptotically (large N). Problems that scale as $O(N^P)$ are said to be of polynomial complexity. Unfortunately, some of the most important problems have exponential complexity.

Fanin/Fanout	Beams of light (unlike beams of electrons or holes) do not interfere with one another, so many of them can be modulated by the same physical elements quite independently, so long as they can be identified. This is usually done by sending the different beams in at different angles, giving rise to the words fanin and fanout, as suggested by the drawing below: 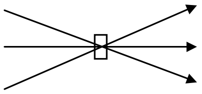
Self organized maps (Kohonen networks)	A means of self-organization of inputs into (normally 2D) maps wherein every concept falls into some category and the categories themselves develop out of application of the Kohonen's algorithm and user-selected parameters applied to the given data.
Singular value decomposition (SVD)	Any matrix can be broken into (decomposed) a weighted sum of simpler, rank-one matrices. The simplest such decomposition (in the sense of fewest number of terms) is the SVD. The weights are the singular values and are indicators of the importance of that term in the decomposition.
Spatial light modulator (SLM)	A 1D or 2D array of elements under electronic or optical control that change some property of the light incident on them. Usually, it is the index of refraction in some direction that is changed. That change can be used to modulate the amplitude, the polarization, or the phase of the light.

See also

Fourier Optics. **Holography, Techniques:** Overview. **Nonlinear Optics, Basics:** Photorefraction. **Optical Communication Systems:** Local Area Networks. **Optical Processing Systems**.

Further Reading

Abu-Mostafa YS and Psaltis D (1987) Optical neural computers. *Scientific American* 256(3): 88–95.

Anderson DZ and Erie MC (1987) Resonator memories and optical novelty filters. *Optical Engineering* 26: 434–444.

Brady D, Psaltis D and Wagner K (1988) Adaptive optical networks using photorefractive crystals. *Applied Optics* 27: 1752–1759.

Caulfield HJ, Kinser J and Rogers SJ (1987) Optical neural networks. *Proc. IEEE* 77: 1573–1583.

Chavel P, Lalanne P and Taboury J (1989) Optical inner-product implementation of neural network models. *Applied Optics* 28: 377–385.

Collins SA, Ahalt SC, Krishnamurthy AK and Stewart DF (1993) Optical implementation and applications of closest-vector selection in neural networks. *Applied Optics* 32: 1297–1303.

Farhat NH, Psaltis D, Prata A and Paek E (1985) Optical implementation of the Hopfield model. *Applied Optics* 24: 1469–1475.

Galstyan T, Frauel Y, Pauliat G, Villing A and Roosen G (1997) Topological map from a photorefractive self-organizing neural network. *Optics Communications* 135: 179–188.

Goodman JW, Dias AR and Woody JM (1978) Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Optics Letters* 2: 1–3.

Johnson JL (1994) Pulse-coupled neural nets: Translation, rotation, scale, distortion, and intensity signal invariance for images. *Applied Optics* 33: 6239–6253.

Jutamulia S (1994) *Selected Papers on Optical Neural Networks*. Bellingham: SPIE Press.

Kinser JM, Caulfield HJ and Shamir J (1988) Design for a massive all-optical bidirectional associative memory: the big BAM. *Applied Optics* 27: 3442–3444.

Neifeld MA and Psaltis D (1993) Optical implementations of radial basis classifiers. *Applied Optics* 32: 1370–1379.

Owechko Y, Dunning GJ, Marom E and Soffer BH (1987) Holographic associative memory with nonlinearities in the correlation domain. *Applied Optics* 26: 1900–1910.

Paek EG and Psaltis D (1987) Optical associative memory using Fourier-transform holograms. *Optical Engineering* 26: 428–433.

Psaltis D and Farhat NH (1985) Optical information processing based on associative-memory model of neural nets with thresholding and feedback. *Optics Letters* 10: 98–100.

Soffer BH, Dunning GJ, Owechko Y and Marom E (1986) Associative holographic memory with feedback using phase-conjugate mirrors. *Optics Letters* 11(2): 118–120.

Yu FTS (1993) Optical neural networks – architecture, design and models. In: Wolf E (ed.) *Progress in Optics*, II, pp. 63–144. Amsterdam: North-Holland.

INSTRUMENTATION

Contents

Astronomical Instrumentation
Ellipsometry
Photometry
Scatterometry
Spectrometers
Telescopes

Astronomical Instrumentation

J Allington-Smith, University of Durham, Durham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the requirements placed on astronomical instruments and the methods used to achieve them, making reference to generic techniques described elsewhere in this encyclopedia. It is limited in scope to electromagnetic radiation with wavelengths from the atmospheric cutoff in the ultraviolet (0.3 μm) to the point in the near-infrared where the technology changes radically (5 μm), taking in the transition from uncooled to cooled instruments at ~1.7 μm. Thus, it excludes the extreme UV and X-ray regimes, where standard optical techniques become ineffective (e.g., requiring grazing incidence optics) and all mid-infrared, sub-mm, THz, and radio-frequency instruments. Terrestrial telescopes are the main focus but space-borne telescopes are subject to similar design considerations, except that the diffraction limit of the optics may be achieved at short wavelengths. Adaptive optics (AO), of increasing importance in astronomy and almost essential for the next generation of extremely large telescopes (ELTs), is covered elsewhere in this encyclopedia.

Astronomical Data

The purpose of astronomical instruments is to measure the following properties of photons arriving on or near to the Earth and brought to a focus by an astronomical telescope:

(i) direction of arrival;
(ii) energy;
(iii) polarization state; and
(iv) time of arrival.

In practice, these properties cannot be measured simultaneously, requiring the following generic classes of instrument:

- *Imagers* to measure (i) and, if equipped with ultra-low-noise detectors, (iv);
- *Spectrographs* to measure (ii), although some can be configured to do a good job on (i) as well. *Integral field spectrographs* can measure (i) and (ii) simultaneously; and
- *Polarimeters* for (iii) plus (i) and *spectropolarimeters for* (iii) plus (ii).

In principle, energy-resolving detectors, such as arrays of superconducting tunnel-junction devices (STJs) can replace (i), (ii), and (iv).

These must interface to the telescope which provides a focal surface with an image scale which is approximately constant over the field. Currently, almost all telescopes are based on the Cassegrain or Ritchy-Chretien reflecting configurations, comprising a large curved primary mirror (M1) which defines the collecting aperture (diameter D_T) and a curved secondary mirror (M2). This feeds the Cassegrain focus via a hole in M1 in line with the central obstruction caused by M2, or the Nasmyth focus if a tertiary (usually plane) mirror is added to direct light from M2 along the elevation axis to a location in which the gravity vector changes only in rotation about the optical axis. For the Cassegrain focus, the full range of change in the gravity vector is experienced as the telescope tracks objects during long exposures or slews between targets. It is also possible to dispense with M2 and mount the

Table 1 Optical interfaces to representative types of telescope and achievable angular resolution

Aperture diameter D_T (m)	*Primary focus*			*Final focus*			*Diffraction limit $2.44\lambda/D_T$(mas)*	
	F/#	*Image scale (μm/″)*	*FOV (′)*	*F/#*	*Image scale (μm/″)*	*FOV (′)*	*$\lambda = 0.5$ μm*	*$\lambda = 5.0$ μm*
2.4^{i}	2.3			24	279	25	101	1007
4.2^{ii}	2.8	60	40	11	221	15	60	600
8.0^{iii}	1.8	(70)	(45)	16	621	10	32	315
30^{iv}	1–1.5	~180	~20	15–19	~2500	~20	8.4	84
100^{v}	~1.4			~6.5	~3000	~10	2.5	25
			For enclosed telescopes				500–1500	
All	Obtainable angular resolution for terrestrial telescopes		Free-atmosphere seeing at good site		Best		250	160
					Median		420	270
			With adaptive optics (approx)				50–200	*Diff. limit*[a]

[a] The diffraction limit listed in the column above for $\lambda = 5.0$ μm applies.

Exemplars: (i) Hubble Space Telescope; (ii) Herschel Telescope, La Palma; (iii) Gemini telescopes, Hawaii/Chile; (iv) CELT/GSMT (USA); (v) OWL (Europe).

instrument directly at the focus of M1 (prime focus), although this imposes severe restrictions on the mass and size of the instrument. An alternative mounting for instruments is via optical fibers. The input, with any fore-optics required to couple the telescope beam into the fiber, is mounted at the appropriate focus while the output, possibly several tens of meters distant, is mounted where the instrument is not subjected to variations in the gravity vector.

The useful field of view (FOV) of the telescope is determined by how rapidly the image quality degrades with radius in the field; restrictions on the size of the Cassegrain hole and the space set aside for the star trackers required for accurate guiding and auxiliary optics involved in target acquisition, AO, etc.

Typical telescope parameters are shown in Table 1, of which the last two entries are speculative examples of ELTs. The table also lists the diffraction limit of the telescopes and the limits imposed by seeing, with and without the effects of the telescope structure and enclosure, and approximate limits for performance with AO which approaches the diffraction limit in the infrared.

Instrument Requirements

Efficiency

Almost all astronomical observations are of very faint objects or require very high signal-to-noise (SNR) to extract the desired subtle information. The former, for example, might reveal the initial phase of galaxy or star formation, while the latter could reveal the mix of elements in the big bang or uncover changes in fundamental constants on cosmological timescales ($\sim 10^{10}$ years). The term ‘efficiency’ has several components:

(i) Throughput of the optical system, including the efficiency of the optical components (lenses or mirrors), dispersers, filters, and also detectors and also includes losses due to transmission/reflection, diffraction, scattering, and vignetting by internal obstructions;
(ii) Noise from all sources: the detector, including time-independent (read-noise) and time-dependent (dark-current) components; thermal emission from the instrument, telescope, enclosure and sky; straylight (see above), and electronic interference;
(iii) Efficient use of the instrument and telescope to maximize the fraction of the time spent accumulating useful photons instead of, say, calibrating; reconfiguring the instrument; or acquiring targets;
(iv) Parallelizing the observations so that the same instrument records data from many objects at the same time. This is discussed further below.

Angular Resolution

The required angular resolution is summarized in Table 2 for key features in galaxies. The resolutions currently achievable with AO systems (with severe restrictions on sky coverage pending routine availability of laser beacons), or without, are shown in bold. Finer resolution is possible using interferometric techniques, but is limited to a handful of bright stars.

Table 2 Typical requirements for angular resolution and single-object field of view

Target	*Distance*[a]	*Active nucleus*		*Star cluster (10 pc)*	*Visible extent*
		Central engine (10 AU)	*Dynamical core (0.5 pc)*		
Milky Way	8 kpc	1 mas	10″	4′	
Local group galaxy (M31)	800 kpc	0.01 mas	100 mas	3″	100″
Distant galaxy	>1 Gpc	10^{-5} mas	0.1 mas	2 mas	0.5–2″
Cluster of galaxies	>1 Gpc				3′
Cosmological structure	>1 Gpc				~1°

[a] 1 pc = 3.1×10^{16} m.

Field of View and Multiplexing

The FOV of an individual observation must be sufficient to cover the angular extent of the object with the minimum number of separate pointings of the telescope. As well as minimizing the total observation time, this reduces photometric errors caused by changes in the ambient conditions (e.g., flexure, thermal effects, and changes in the sky background). (Photometry is the measurement of the flux density received from an astronomical object, usually done with the aid of an imager employing filters with standardized broad passbands. Spectrophotometry is the same but in the much narrower passbands obtained by dispersing the light.) The last column of Table 2 indicates the required size of FOV. It is also highly desirable to observe many objects simultaneously since most studies require large statistically homogenous samples. The precise requirement depends on the surface number density of potential targets and the sampling strategy adopted to create the sample. Cosmological studies of the large-scale distribution of galaxies and clusters emphasize the size of the field ($\geq 0.5°$), since the targets may be sparsely sampled, while studies of clustered objects (galaxies or stars) aim to maximize the surface density of targets (≤ 1000 arcmin^{-2}) within a modest FOV ($\sim 5'$).

Spectral Resolution

The required energy resolution is specified in terms of the spectral resolving power, $R \equiv \lambda/\delta\lambda$, where $\delta\lambda$ is the resolution in wavelength defined usually in terms of the Rayleigh criterion. One of the main data products required in astrophysics is the velocity along the line of sight (radial velocity) obtained through measurement of the wavelength of a known spectral feature (emission or absorption line). This may be determined to an accuracy of:

$$\delta v = \frac{Qc}{R} \qquad [1]$$

where Q is the fraction of the width of the spectral resolution element, $\delta\lambda$, to which the centroid can be determined. This is limited by either the SNR of the observation of the line via $Q \approx 1/\mathrm{SNR}$, or by the stability of the spectrograph via:

$$Q = \frac{\Delta x}{\chi}\left(\frac{\mathrm{d}\chi}{\mathrm{d}x}\right) \qquad [2]$$

where χ is the angular width of the slit measured on the sky and $\delta\chi/\mathrm{d}x$ is the image scale at the detector. Δx is the amount of uncorrectable flexure in units of the detector pixel size. This is the part of the motion of the centroid of the line which is not susceptible to modeling with the aid of frequent wavelength calibration exposures. Typically, $\Delta x \approx 0.2$ pixels is achievable (roughly 3 μm at the detector or 10 μm at the slit for an 8 m telescope) implying $Q \approx 1/30$ for a typical $(\mathrm{d}\chi/\mathrm{d}x) = 0.1$–$0.2''$/pixel and $\chi = 0.5$–$1''$. Table 3 gives typical requirement for R and the value of Q required to achieve it, given the spectral resolution obtainable for different types of current spectrograph. Extrasolar planet studies require very high stability.

Some spectral observations require very accurate measurements of the shape of the spectral features to infer the kinematics of ensembles of objects which are individually not resolvable (e.g., stars in distant galaxies, gas in blackhole accretion disks). Other applications require accurate measurements of the relative flux of various spectral features within the same spectrum (e.g., to reveal abundances of elements or different types of star).

Optical Principles

Imaging

The simplest type of imager consists of a detector, comprising a 2D array of light-sensitive pixels (Table 4) placed at a telescope focal surface without any intervening optics except for the filters required for photometry. Although maximizing throughput, this arrangement removes flexibility in changing the image scale since the physical size of pixels (typically 15–25 μm) is fixed. Thus the Gemini telescopes

Table 3 Typical requirements for spectral resolution and stability

Target	*δv or required data product*	*Required λ/δλ*	*Spectral regime*	*Available R*	*Stability required = 1/Q*	*Simultaneous wavelength range*
Cosmological parameters	1000 km/s (line centroid)	300	*Low dispersion*	300–3000	1	≳1 octave
Galaxies						
Intra-cluster	100 km/s (line centroid)	3000			1–10	100 nm
Element abundance	Ratio of line fluxes	1000–3000				1 octave
Stellar orbits	10 km/s (line shapes)	3×10^4	*Medium*	$1–3 \times 10^4$	(1)	10 nm
Individual stars	1 km/s (line centroid)	3×10^5			10–30	10 nm
Extrasolar planets	<1 m/s (line centroid)	$>3 \times 10^8$	*High*	$\sim 10^5$	>3000	<1 nm

Table 4 Detectors in current use for astronomy

Wavelength (μm)	*Material/ type*	*Pixel size (μm)*	*Format*		*Read noise (e-/pix)*	*Dark current (e-/pix/hr)*	*QE range (%)*
			Single	*Mosaic*			
		Typical	*State-of-the-art*				
0.3–1.0	CCD	10–20	4k × 4k	12k × 8k	2	Negligible	10–90[a]
	CMOS	5–10	4k × 4k		5–30		
1–2.5 +	HgCdTe	15–20					
0.6–5	InSb	20–30	2k × 2k	4k × 4k	10	0.01–0.1	70–90

[a] Quantum efficiency varies strongly over quoted wavelength range.

($D_T = 8$ m) provides an uncorrected image scale of 25–40 mas/pixel at its sole F/16 focus. Although well-suited to high-resolution observations with AO, this is inappropriate for most observations of faint galaxies where 100–200 mas/pixel is required. This arrangement also provides limited facilities for defining the wavelength range of the observations since there is no intermediate focal surface at which to place narrowband filters, whose performance is sensitive to field angle and are, therefore, best placed at an image conjugate. There is also no image of the telescope pupil at which to place a cold stop to reject the external and instrumental thermal radiation which is the dominant noise source at $\lambda > 1.5$ μm.

The alternative is to use a focal-reducer consisting of a collimator and camera. The auxiliary components (filters, cold-stops, Fabry–Perot etalons, etc.) then have an additional image and pupil conjugate available to them and the image scale may be chosen by varying the ratio of the camera and collimator focal lengths. The extra complexity of optics inevitably reduces throughput but, through the use of modern optical materials, coatings such as Sol–Gel and the avoidance of internally obstructed reflective optics, losses may be restricted to 10–20% except at wavelength shorter than ~0.35 μm. An imaging focal reducer is a special case of the generic spectrograph described in the next section.

Spectroscopy

Basic principles

Astronomical spectrographs generally employ plane diffraction gratings. The interference between adjacent ray paths as they propagate through the medium is illustrated in Figure 1. From this, we obtain the grating equation:

$$m\rho\lambda = n_1 \sin\alpha + n_2 \sin\beta \equiv G \qquad [3]$$

where α and β are the angles of incidence and diffraction respectively, $\rho = 1/a$ is the ruling density, λ is the wavelength, and m is the spectral order. The refractive indices in which the incident and diffracted rays propagate are n_1 and n_2 respectively.

The layout of a generic spectrograph employing a plane diffraction grating in a focal reducer arrangement is shown in Figure 2. The spectrograph re-images the slit onto the detector via a collimator and camera. The disperser is placed in the collimated beam close to a conjugate of the telescope pupil.

A field lens placed near the focus is often incorporated in the collimator for this purpose.

The geometrical factor, G, defined in the equation is constrained by the basic angle of the spectrograph:

$$\psi = \alpha - \beta \qquad [4]$$

which is the (normally fixed) angle between the optical axes of the collimator and camera. $\psi = n\pi$, corresponds to the Littrow configuration (for integer n).

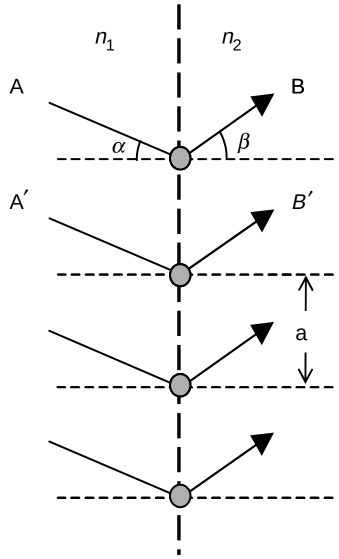

Figure 1 Derivation of the grating equation by consideration of the optical path difference between neighboring ray paths AB and A′B′.

Considering the case of a diffraction grating in vacuo, differentiation of the grating equation with respect to the diffracted angle yields the angular dispersion:

$$\frac{d\lambda}{d\beta} = \frac{\cos\beta}{m\rho} \qquad [5]$$

from which the linear dispersion is obtained by considering the projection of $d\beta$ on the detector, dx:

$$\frac{d\lambda}{dx} = \frac{d\lambda}{d\beta}\frac{d\beta}{dx} = \frac{\cos\beta}{m\rho f_2} \qquad [6]$$

where f_2 is the focal length of the camera (see Figure 2).

In the diffraction-limited case, where the slit is arbitrarily narrow, the resolving power is given simply by the total number of rulings multiplied by the spectral order:

$$R^* = m\rho W \qquad [7]$$

where w is the length of the disperser as defined below. But in astronomy, it is usually determined by the width of the image of the slit, s, projected on the detector, s'. By conservation of Etendue:

$$s' = s\frac{F_2}{F_1} \qquad [8]$$

where F_1 and F_2 are the collimator and camera focal ratios, respectively. The spectral resolution of the spectrograph is simply the width of this expressed in

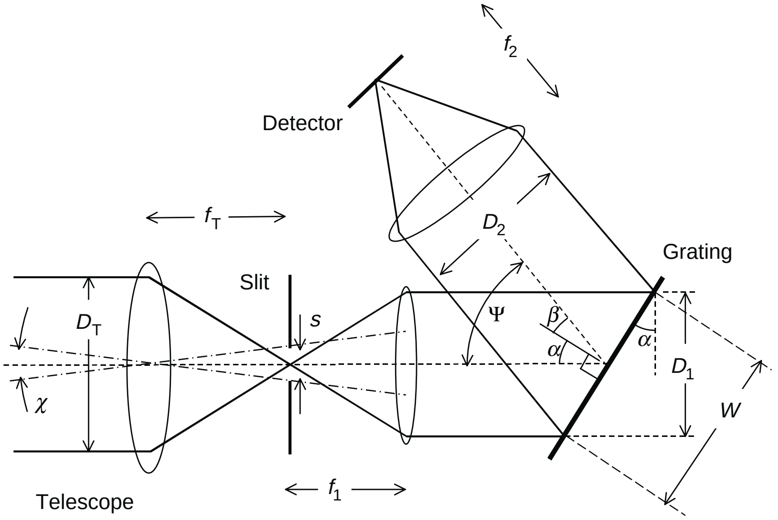

Figure 2 Generic spectrograph employing a plane diffraction grating. Note that according to the sign convention implicit in Figure 1, the diffraction angle, β, is negative in this configuration.

wavelength units:

$$\delta\lambda = \left(\frac{d\lambda}{dx}\right)s' = \frac{\cos\beta}{m\rho f_2} s \frac{F_2}{F_1} = \frac{s D_1 \cos\beta}{m\rho D_2 f_1} \qquad [9]$$

The length of the intersection between the collimated beam and the plane of the grating (not necessarily the actual physical length of the grating) is

$$W = \frac{D_2}{\cos\beta} = \frac{D_1}{\cos\alpha} \qquad [10]$$

Using this to eliminate cos β, the spectral resolution becomes

$$\delta\lambda = \frac{s}{m\rho F_1 W} \qquad [11]$$

and the resolving power is

$$R \equiv \frac{\lambda}{\delta\lambda} = \frac{m\rho\lambda F_1 W}{s} = G\frac{F_1 W}{s} \qquad [12]$$

where G is the quantity defined in eqn [3]. Note that R is now independent of the details of the camera. This expression is useful for the laboratory since it is expressed in terms of the parameters of the experimental apparatus; but, for astronomy, it is more useful to express the resolving power in terms of the angular slit width, χ, and the telescope aperture diameter, D_T, via:

$$s = \chi f_T = \chi F_T D_T = \chi F_1 D_T \qquad [13]$$

since the collimator and telescope focal ratios are the same if the spectrograph is directly beam-fed from the telescope. Note that even if the telescope focal surface is re-imaged onto the slit, the expression for the resolving power still holds due to the conservation of Etendue in the re-imaging optics. Thus the resolving power is

$$R = \frac{m\rho\lambda W}{\chi D_T} = \frac{GW}{\chi D_T} \qquad [14]$$

Note that the resolving power obtained with a nonzero slit width is always less than the theoretical maximum, $R \le R_*$, for wavelengths:

$$\lambda < \lambda_* = \chi D_T \qquad [15]$$

Thus, long-wavelength applications may approach the theoretical limit; in which case they are said to be diffraction-limited. If so, the resolving power is independent of the slit width, simplifying the interface to different telescopes. For a nondiffraction-limited spectrograph, the resolving power obtained depends on the aperture of the telescope to which it is fitted.

Optical configuration

The non-Littrow configuration shown in Figure 2 introduces anamorphism into the recorded spectra such that a monochromatic image of the slit mask is compressed in the dispersion direction by a factor:

$$A = \frac{D_2}{D_1} = \frac{\cos\beta}{\cos\alpha} \qquad [16]$$

which varies, depending on the choice of angles required to place the chosen wavelength on the detector using eqns [3] and [4].

Practical configurations that maximize R, do so by increasing $W = D_1/\cos\alpha$. This implies maximizing cos β, since $\psi = \alpha - \beta$, which results in $A > 1$. This is known as the grating-to-camera configuration since the grating normal points more towards the camera than to the collimator. The grating-to-collimator configuration may also be used at the penalty of lower R but does permit a smaller camera. Thus the camera aperture must be oversized in the dispersion direction since $D_2 > D_1$. The Littrow configuration is rarely used with reflective designs when large FOV is required because of internal vignetting and consequent light loss. However, it can be used effectively with transmissive dispersers.

If the disperser is replaced by a plane mirror (or if the disperser is removed in a transmissive system), the system acts as a versatile imager, with $A = 1$. This also facilitates target acquisition which requires that the telescope attitude be adjusted until the targets are aligned with apertures cut in an opaque mask placed at the input of the spectrograph. These apertures (normally slits) are usually oversized in the direction perpendicular to the dispersion to allow the spectrum of the sky to be recorded directly adjacent to that of the target to permit accurate subtraction of the background.

The slit width is chosen by trading off the desire to maximize the amount of light entering the instrument from the target (but not so much as to admit excessive background light) and the need to maximize spectral resolving power, R, by reducing the slit width, χ, according to eqn [14].

Since $|G| \le 2$, the maximum attainable resolving power, according to eqn [14], is determined chiefly by the telescope aperture and angular slitwidth. The only important parameter of the spectrograph is the illuminated grating length, W. Therefore, maintaining the same resolving power as the telescope aperture increases, requires W to increase in direct proportion. Because of limits placed on the geometry by the need to avoid obstructions, etc., this means that both the collimated beam diameter and the size of the disperser must increase. This is one of the major problems in devising instruments for ELTs. Current instruments on 8 m telescopes feature collimated beam diameters of $D_1 \approx 150$ mm. For ELTs, the implied sizes are 0.5–2 m, requiring huge optics, accurately co-aligned mosaics of dispersing elements, and enormous support structures to provide the required stability. A simple calculation indicates that the ruling stylus for a classically ruled diffraction grating would have to cover a thousand kilometers without significant wear unless the disperser was made from smaller pieces.

Choice of dispersing element

Astronomical spectrographs mostly use surface-relief diffraction (SL) gratings. These are ruled with groove densities $50 < \rho < 3000\ \text{mm}^{-1}$.

Blazing diffraction gratings. The intensity of the multibeam interference pattern from N rulings of finite width is given by

$$I = \left(\frac{\sin^2 N\phi}{\sin^2 \phi}\right)\left(\frac{\sin^2 \theta}{\theta^2}\right) \qquad [17]$$

where 2ϕ is the phase difference between the center of adjacent rulings and θ is the phase difference between the center and edge of a single ruling. To be useful for astronomy, the peak in this pattern, determined by the second term, must coincide with a useful order, such as $m = 1$, rather than zero order. This can be done by introducing a phase shift into the second term, which represents the diffraction of a single ruling, such that:

$$\theta = \frac{\pi \cos \gamma}{\rho\lambda}(\sin i - \sin r) \qquad [18]$$

where the groove angle, γ, is the angle by which the facets are tilted with respect to the plane of the grating. With the aid of Figure 3, it can be seen that $i = \alpha - \gamma$ and $r = \gamma - \beta$ so that the maximum efficiency, which occurs when $\theta = 0$, corresponds to $i = r$, which is equivalent to simple reflection from the facets, and $\alpha + \beta = 2\gamma$.

Thus, through this process of blazing, the grating eqn [3], at the blaze wavelength, λ_B, becomes:

$$\rho m \lambda_B = 2 \sin \gamma \cos \frac{\psi}{2} \qquad [19]$$

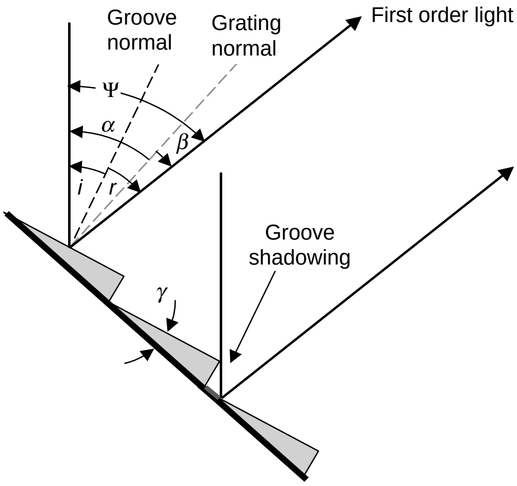

Figure 3 Blazing of ruled diffraction grating.

where ψ, is given by eqn [4] using the identity: $\sin x + \sin y \equiv 2 \sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x+y}{2}\right)$. Thus, the resolving power at blaze is obtained from eqns [10], [14], and [19] as:

$$R_B = \frac{2D_1}{\chi D_T}\frac{\sin\gamma \cos(\psi/2)}{\cos(\gamma + \psi/2)} \qquad [20]$$

Use of grisms. An alternative configuration uses transmission gratings. The advantage here is that the collimator and camera can share the same optical axis in a straight-through configuration with unit anamorphism, $A = 1$, which does not require oversizing of the camera. The same phase-shift can be applied as before by making the facets into prisms with the required groove angle. However, this would only allow zero-order to propagate undeviated into the camera so an additional prism is required with vertex angle $\phi = \gamma$ to allow first-order light to pass undeviated into the camera. This composite of blazed transmission grating and prism is called a grism (Figure 4).

The grating equation (eqn [3]) is modified for a grism in the blaze condition to $m\rho\lambda_B = (n-1)\sin\phi$ since $\alpha = -\beta = \phi$ and $n_1 \equiv n, n_2 = 1$. Noting that $\psi = 0$, the resolving power at blaze is then:

$$R_B = \frac{D_1}{\chi D_T}(n-1)\tan \phi \qquad [21]$$

where n is the refractive index of the prism and grating material. Due to problems such as groove shadowing (illustrated for reflection gratings in Figure 3), these lack efficiency when $\phi \gtrsim 30°$ restricting their use to low-resolution spectroscopy.

Use of volume phase holographic gratings. A newer alternative to surface-relief (SR) dispersers are volume phase holographic (VPH) gratings in which the interference condition is provided by a variation of refractive index of a material such as dichromated gelatine, n_g, which depends harmonically on position inside the grating material as:

$$n_g(x,z) = \bar{n}_g + \Delta n_g \cos[2\pi\rho_g(x\sin\gamma + z\cos\gamma)] \qquad [22]$$

where ρ_g is the density of the lines of constant n_g and $\bar{n}_g$ and Δn_g are the mean refractive index and its modulation amplitude, respectively (Figure 5). These are described by the same grating (eqn [3]), as before leading to an identical expression for the resolving power at blaze, where $m\rho\lambda_B = 2\sin\alpha$, to that of a blazed plane diffraction grating in the Littrow configuration (eqn [20]):

$$R_B = \frac{D_1}{\chi D_T} 2\tan\alpha \qquad [23]$$

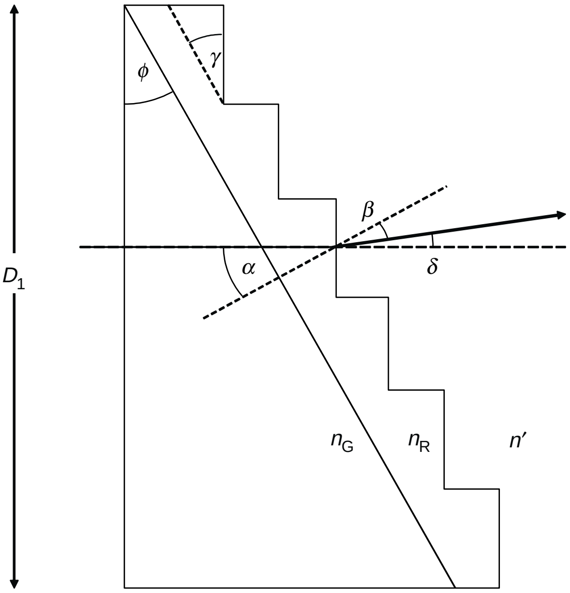

Figure 4 Typical configuration of a grism. The prism vertex and groove angles are chosen for maximum efficiency on the assumption that refractive indices, $n_G = n_R = n$ and $n' = 1$. The blaze condition occurs when $\delta = 0$. In this configuration, $\phi = \gamma = \alpha$.

Unlike SR gratings, the blaze condition of VPH gratings may be varied by changing the basic angle of the spectrograph, $\psi = 2(\gamma + \alpha)$, although this is often mechanically inconvenient. They may also be sandwiched between prisms to provide the necessary incident and diffraction angles while retaining the advantage of a straight-through configuration, to form an analogue of a grism, sometimes termed a vrism. Unlike grisms, these are not constrained by groove shadowing and so can be efficient at relatively high dispersion, subject to the constraint imposed by the size of the device.

Use of echelle gratings. Another option to increase the resolving power is to use a coarse grating in high order via the use of an echelle format. The resolving power in the near-Littrow configuration in which it is usually operated, is given by eqn [20] with $\psi = 0$.

$$R_B = \frac{D_1}{\chi D_T} 2\tan\gamma \qquad [24]$$

where γ is the groove angle.

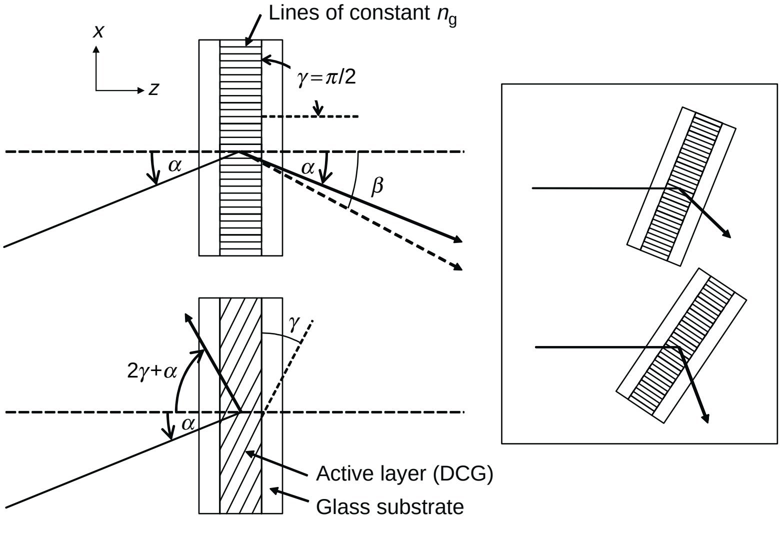

Figure 5 Various configurations of volume phase holographic gratings. The blaze condition is when $\beta = \alpha$. The inset shows how the blaze condition may be altered by changing the basic angle of the spectrograph.

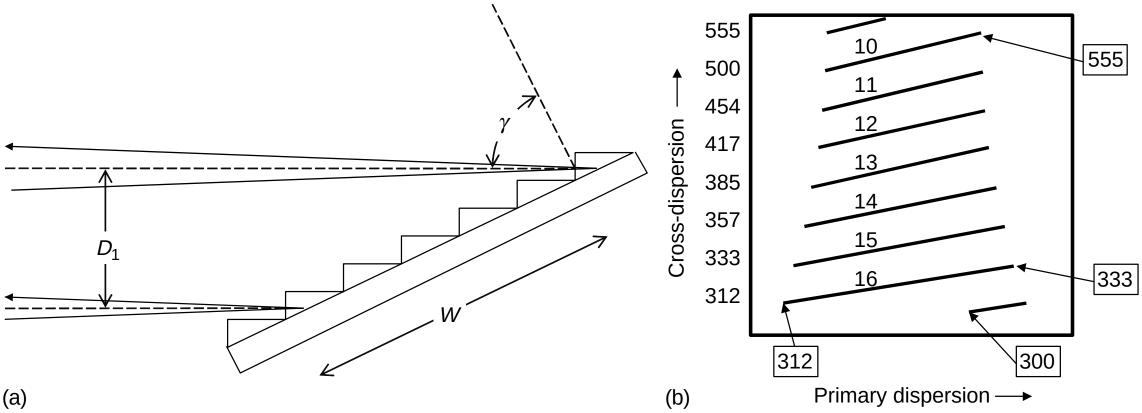

Figure 6 (a) Basic layout of an Echelle grating used in a near-Littrow configuration. (b) Illustrative cross-dispersed spectrogram showing a simplified layout on the detector. The numbers 10–16 label the different spectral orders. The numbers labeling the vertical axis are the wavelength (nm) at the lowest end of each complete order. Other wavelengths are labeled for clarity. For simplicity the orders are shown evenly spaced in cross-dispersion.

In the configuration shown in Figure 6, R is maximized by increasing W. This also means that $G = m\rho\lambda$ is large. In order to avoid excessively high ruling densities, the grating may be used in high orders. However, this presents the problem of order overlap since the wavelength in order m occurs in the same place as wavelength $\lambda_n = (m/n)\lambda_m$ in order n. Limiting the observation to a single order via the use of order-sorting filters is one solution while another is to use cross-dispersion via a chain of prisms or a grating with dispersion perpendicular to that of the echelle to separate the orders out. This option severely curtails the use of such spectral formats in multi-object spectroscopy.

Increasing resolving power using immersed gratings. All the methods of using diffraction gratings discussed so far are subject to various geometric constraints which ultimately limit the maximum obtainable resolving power by placing limits on W (e.g., in eqn [14]). These may be partly overcome by optically coupling the dispersers to immersing prisms. This can lead to a doubling of the resolving power in some cases.

Use of prisms. The final method of dispersion is to use a prism. This has the advantage that, since it does not work by multibeam interference, the light is not split into separate orders which removes the problem of order-overlap and improves efficiency. However, it does not produce the high spectral resolution required for some applications and the dependency of dispersion on λ is often markedly nonlinear.

From a consideration of Fermat's principle, it can be shown that the resolving power of a prism is:

$$R = \frac{\lambda t}{\chi D_{\mathrm{T}}}\left(\frac{\mathrm{d}n}{\mathrm{d}\lambda}\right) \qquad [25]$$

where t is the baselength of the prism. Using modern materials, $R \lesssim 300$ may be obtained. The problem of the nonlinearity in dispersion can be alleviated by the use of composite materials with opposite signs of dispersion.

Multiobject spectroscopy (MOS)

As discussed above, there is great value in increasing the multiplex gain. This can be done in two ways:

(i) *Increasing the number of apertures in the slit mask (Multislits).* This requires that each slit is carefully cut in the mask because, although the telescope attitude (translation and rotation) may be adjusted to suit one aperture, errors in the relative positions of slits cannot be compensated since the image scale is fixed. Each aperture produces a spectrum on the detector and the mask designer must ensure that the targets and their matching slits are chosen to avoid overlaps between spectra and orders. This has the effect of limiting the surface density of targets which can be addressed in one observation. Passband filters can be used to limit the wavelength range and reduce the overlap problem. This option requires superior optics able to accommodate both a wide FOV and large disperser.

(ii) *Using optical fibers (Multifibers).* A number of fibers may be deployed at the telescope focus to direct light from the various targets to a remote spectrograph whose input focus consists of one or more continuous pseudoslits. These are made up of the fiber outputs arranged in a line. The continuous nature of the slit means that spectral overlap is avoided without restricting the surface density of addressable targets; although there is a limit imposed by the distance of closest approach of the fibers. The method of deploying the fiber inputs may be a plugplate consisting of pre-cut holes into which encapsulated fibers are manually plugged, or a pick-and-place robot which serially positions the fiber inputs at the correct location on a magnetized field plate. The latter is highly versatile but mechanically complex with a significant configuration time that may erode the actual on-target integration time.

The two systems have contrasting capabilities (Figure 7). The multislit approach provides generally better SNR since the light feeds into the spectrograph directly, but is compromised in terms of addressable surface density of targets, by the need to avoid spectral overlap and in FOV by the difficulty of the wide-field optics. A current example of this type of instrument, GMOS, is shown in Figure 8 and described in Table 5. The multifiber approach is limited in SNR by modal noise in the fibers and attendant calibration uncertainties and the lack of contiguous estimates of the sky background. However, it is easier to adapt to large fields since the spectrograph field can be much smaller than the field over which the fiber inputs are distributed. In summary, multislit systems are best for narrow-but-deep surveys while the multifiber systems excels at wide-but-shallow surveys. Fiber-fed instruments may further prove their worth in ELTs where technical problems may require that these bulky instruments are mounted off the telescope (see below for further discussion).

Of paramount importance in MOS is the quality of the background subtraction as discussed above. Traditionally, this requires slits which sample the sky background directly adjacent to the object. An alternative is nod-and-shuffle (va-et-vient) in which nearby blank regions are observed alternately with the main field using the same slit mask by moving the telescope (the nod). In the case of CCDs, the photo-generated charge from the interleaved exposures is stored temporarily on the detector by moving it to an unilluminated portion of the device (the shuffle). After many repeats on a timescale less than that of variations in the sky background (a few minutes), the accumulated charge is read out, incurring the read-noise penalty only once. Although requiring an effective doubling of the exposure time and an

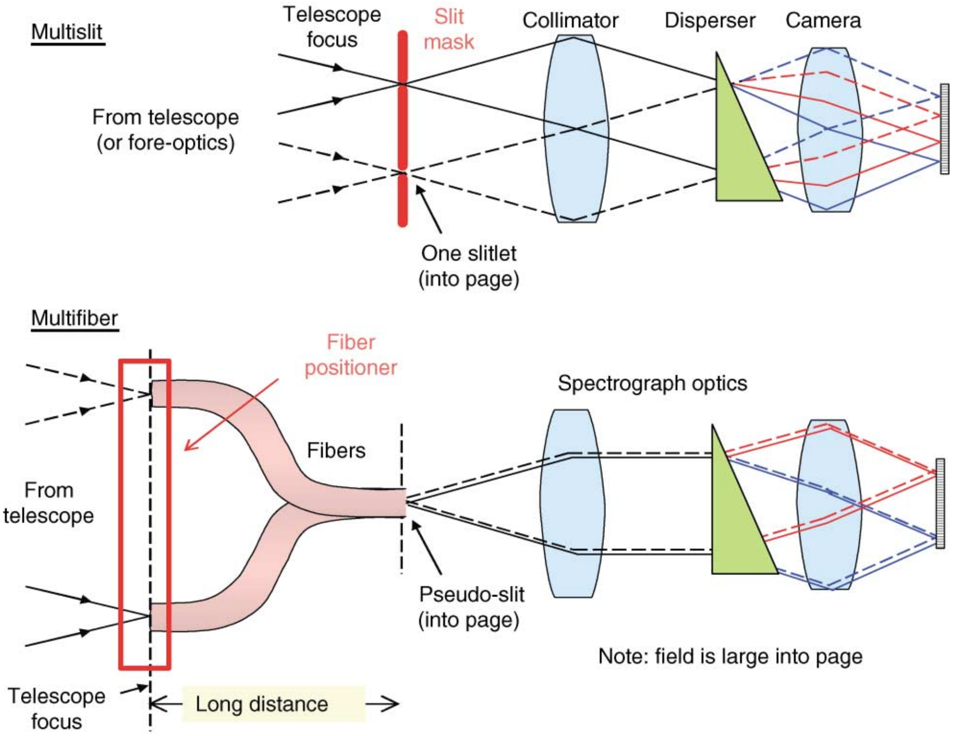

Figure 7 Illustration of the difference between the multislit and multifiber approaches to multi-object spectroscopy.

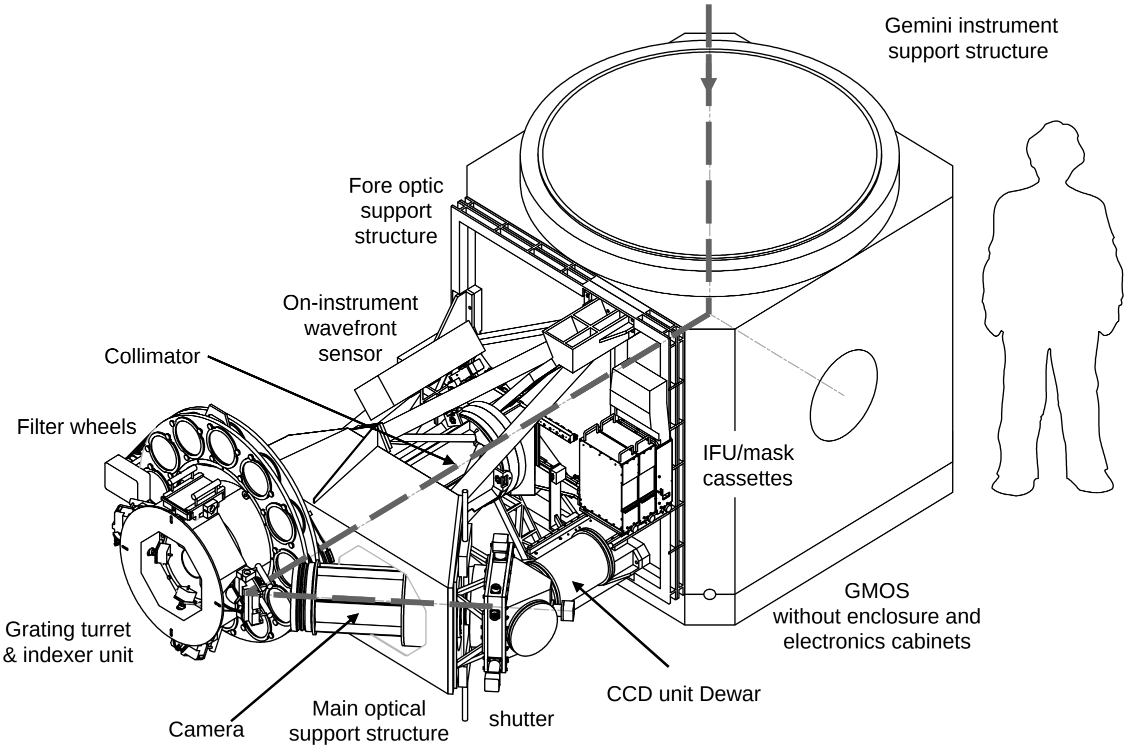
Gemini instrument support structure
Fore optic support structure
On-instrument wavefront sensor
Collimator
Filter wheels
IFU/mask cassettes
GMOS without enclosure and electronics cabinets
Grating turret & indexer unit
CCD unit Dewar
Camera
Main optical support structure
shutter

Figure 8 The Gemini Multiobject Spectrograph (GMOS). One of two built for the two Gemini 8-m telescopes by a UK–Canadian consortium. It includes an integral field capability provided by a fiber-lenslet module built by the University of Durham. The optical configuration is the same as shown in Figure 2. It is shown mounted on an instrument support unit which includes a mirror to direct light from the telescope into different instruments. The light path is shown by the red dashed line. The slit masks are cut by a Nd-YAG laser in 3-ply carbon fiber sheets. See Table 5 for the specification.

Table 5 Main characteristics of the Gemini multiobject spectrographs

Image scale	72 mas/pixel	FOV	$5.5' \times 5.5'$
Detector	CCD with 13.5 μm pixels Format: $3 \times (4608 \times 2048)$	Slit/mask configuration	$\leq$few $\times$ 100 slits of width $\geq 0.2''$
Wavelength range	Total: 0.4–1.0 μm Simultaneous: $\leq$ 1 octave	Integral field unit	$1500 \times 0.2''$ samples in 50 sq.$''$
Spectral resolving power	$R \leq 5000$ with $0.5''$ slit	Dispersion options	3 gratings + mirror for imaging

increase in the size of the detector, this technique allows the length of the slits to be reduced since no contiguous sky sample is required, thereby greatly increasing the attainable multiplex gain if the number density of potential targets is sufficiently high.

Integral field spectroscopy (IFS)

IFS provides a spectrum of each spatial sample simultaneously within a contiguous FOV. Other approaches provide the same combination of imaging and spectroscopy but require a series of nonsimultaneous observations. Examples include imaging through a number of narrow passband filters with different effective wavelength and spectroscopy with a single slit which is stepped across the object. Other such techniques are Fabry–Perot, Fourier-transform and Hadamard spectroscopy. Since the temporal variation of the sky background is a major limitation in astronomy of faint objects, IFS is the preferred technique for terrestrial observations of faint objects, but nonsimultaneous techniques are preferable in certain niche areas, and relatively more important in space where the sky background is reduced.

The main techniques (Figure 9) are as follows:

(i) *Lenslets*. The field is subdivided by placing an array of lenslets at the telescope focus (or its conjugate following re-imaging fore-optics) to form a corresponding array of micropupils. These are re-imaged on the detector by a focal

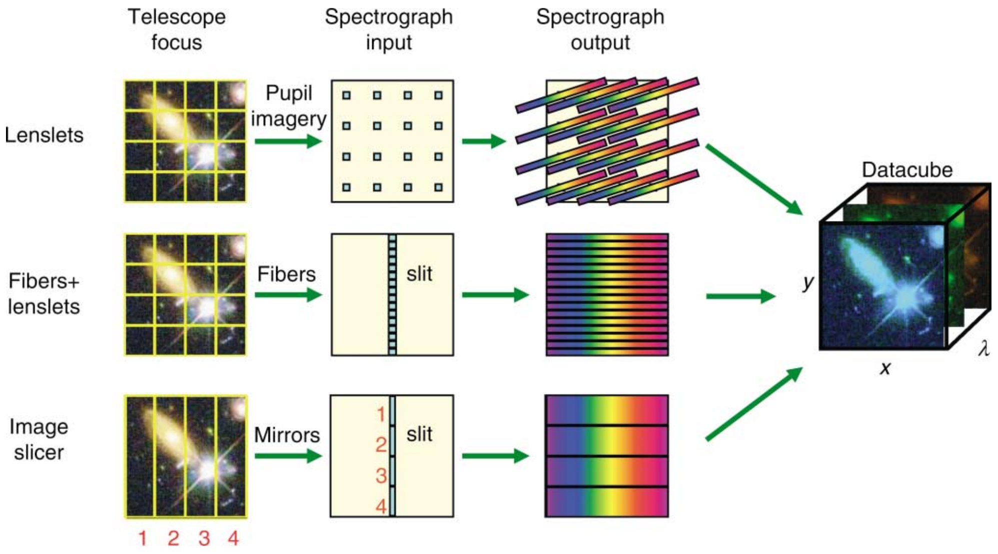

Figure 9 Main techniques of integral field spectroscopy.

reducer using a conventional disperser. The spectra are, therefore, arranged in the same pattern as that of the lenslet array. Spectral overlap is reduced by angling the dispersion direction away from the symmetry axes of the lenslet array and by the fact that the pupil images are smaller than the aperture of the corresponding lenslet.

(ii) *Fibers + lenslets*. The field is subdivided as in (i) but the pupil images are relayed to a remote spectrograph using optical fibers which reformat the field into linear pseudoslits. This avoids the problem of overlap, allowing a greater length of spectrum than in (i), but is more complex since the arrays of fibers and lenslets must be precisely co-aligned and are subject to modal noise in the fibers.

(iii) *Image slicer*. The field is subdivided in only one dimension by a stack of thin slicing mirrors placed at a conjugate of the telescope focus (Figure 10). Each mirror is angled so as to direct light to its own pupil mirror which re-images the field to form part of a linear pseudoslit. Thus the slices into which the image is divided are rearranged end-to-end to form a continuous slit at the spectrograph input. Unlike the other techniques, this retains spatial information along the length of each slice. This is dispersed by conventional means via a focal-reducer. An additional optic is often required at the image of each slice at the slit to reimage the micropupil images produced by the slicing mirrors onto a common pupil inside the spectrograph. In principle this is the most efficient of the three methods since the one-dimensional slicing produces fewer diffraction losses in the spectrograph than the two-dimensional division used by the others, and minimizes the fraction of the detector surface which must be left unilluminated in order to avoid cross-talk between noncontiguous parts of the field. However the micromirrors are difficult to make since they require diamond-turning or grinding in metal or glass with a very fine surface finish (typically with RMS ~ 1 nm for the optical and ~ 10 nm in the infrared). This sort of system is, however, well-matched to cold temperatures, since the optical surfaces and mounts may be fabricated from the same material (e.g., Al) or from materials with similar thermal properties.

The data are processed into a datacube whose dimensions are given by the two spatial coordinates, x, y, plus wavelength. The datacube may be sliced up in ways analogous to tomography to understand the physical process operating in the object.

Spectropolarimetry and Polarimetry

Spectropolarimetry and polarimetry are analogous to spectroscopy and imaging, where the polarization state of the light is measured instead of the total intensity. Like IFS, spectropolarimetry is a photon-starved (i.e., limited in SNR by photon noise) area which benefits from the large aperture of current and projected telescopes.

The Stokes parameters of interest are I, Q, and U. V is generally very small for astronomical objects.

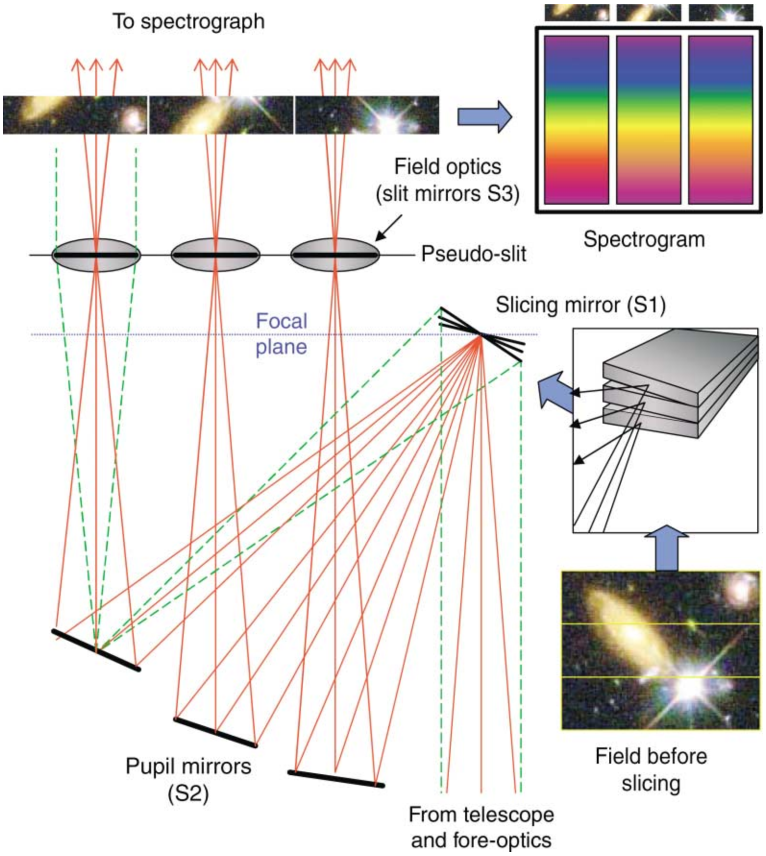

Figure 10 Principle of the advanced image slicer. Only three slices are shown here. Real devices have many more, e.g., the IFU for the *Gemini Near-Infrared Spectrograph* has 21 to provide a 5″ × 3″ field with 0.15″ × 0.15″ sampling. Reproduced from Content R (1997) A new design for integral field spectroscopy with 8 m telescopes. *Proceedings SPIE* 2871: 1295–1305.

From these, the degree and azimuthal angle of linear polarization are obtained as:

$$\pi_{\mathrm{L}} = \left(\frac{Q^2 + U^2}{I^2}\right) \qquad \psi = \frac{1}{2}\arctan\left(\frac{U}{Q}\right) \qquad [26]$$

The Stokes parameters may be estimated using a modified spectrograph with a rotatable achromatic half-wave retarder, characterized by the angle of rotation about the optical axis, θ, placed before the instrument and a polarizing beamsplitter which separates the incident beam into orthogonal polarization states (Figure 11). The two states are recorded simultaneously on different regions of the detector. The separation is achieved usually through the angular divergence produced by a Wollaston prism placed before the disperser or through the linear offset produced by a calcite block placed before the slit. The intensity as a function of the waveplate rotation angle, $S(\theta)$, is recorded for each wavelength and position in the spectrum or image. The Stokes parameters for each sample are then given by

$$\begin{aligned}
\frac{Q}{I} &= \frac{S(0) - S(\pi/2)}{S(0) + S(\pi/2)} \\
&= \left\langle \frac{A_1 - \omega_{AB}B_1}{A_1 + \omega_{AB}B_1}, \frac{A_2 - \omega_{AB}B_2}{A_2 + \omega_{AB}B_2} \right\rangle \\
\frac{U}{I} &= \frac{S(\pi/4) - S(3\pi/4)}{S(\pi/4) + S(3\pi/4)} \\
&= \left\langle \frac{C_1 - \omega_{CD}D_1}{C_1 + \omega_{CD}D_1}, \frac{C_2 - \omega_{CD}D_2}{C_2 + \omega_{CD}D_2} \right\rangle \\
I &= S(0) + S(\pi/2) \\
&= \left\langle \frac{A_1 + \omega_{AB}B_1}{\eta_{A1}}, \frac{A_2 + \omega_{AB}B_2}{\eta_{A2}} \right\rangle \text{ etc.}
\end{aligned} \qquad [27]$$

where A and B are the signals (functions of position and/or λ) recorded with the waveplate at $\theta = 0$ and $\theta = \pi/2$ and the subscripts 1 and 2 refer to the two polarization states produced by the beamsplitter.

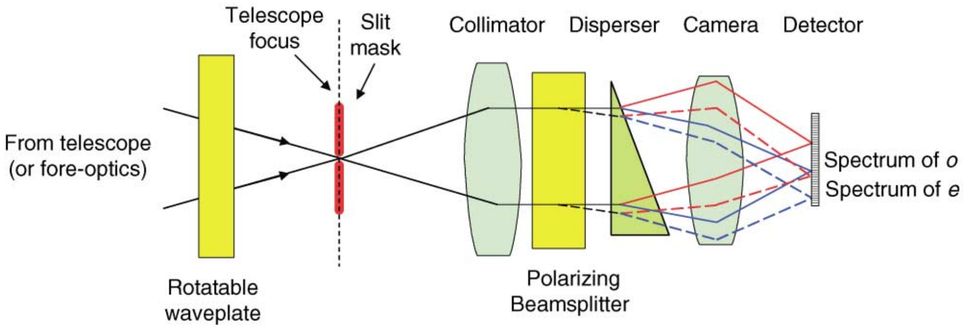

Figure 11 Schematic layout of spectropolarimeter using a polarizing beamsplitter in the collimated beam (e.g., a Wollaston prism producing *e*- and *o*-states) to produce separate spectra for orthogonal polarization states on the detector. Alternative configurations in which the two polarizations have a linear offset (e.g., a calcite block) may be used in which case the beamsplitter is placed before the slit.

Likewise, C and D are a pair of observations taken with $\theta = \pi/4$ and $\theta = 3\pi/4$. The angled brackets indicate a suitable method of averaging the quantities inside. The factors ω_{AB} and ω_{CD} are estimated as

$$\omega_{AB} = \sqrt{\frac{A_1 A_2}{B_1 B_2}} \qquad \omega_{CD} = \sqrt{\frac{C_1 C_2}{D_1 D_2}} \qquad [28]$$

and η_{A1} is a single calibration factor for polarization state 1 of observation A, etc. One problem is instrumental polarization caused by oblique reflections from telescope mirrors upstream of the polarizing waveplate, such as the fold mirror external to GMOS (see Figure 8). This must be carefully accounted for by taking observations of stars with accurately known polarization. The dispersing element is also likely to be very strongly dependent on the polarization state of the incident light with widely different efficiency as a function of wavelength. Although this effect is cancelled out using the procedure described above, it introduces a heavy toll in terms of SNR reduction.

Technology Issues

Use of Optical Fibers in Astronomy

As discussed above, optical fibers are commonly used in astronomy for coupling spectrographs to wide, sparsely sampled FOVs in the case of MOS, or small contiguously sampled FOVs in the case of IFS. For both applications, the most important characteristics are:

(i) *Throughput*. Near-perfect transmission is required for wavelength intervals of ~1 octave within the overall range of 0.3–5 μm. For MOS, a prime focus fiber feed to an off-telescope spectrograph implies fiber runs of several tens of meters for 8 m telescopes, increasing proportionally for ELTs. For IFS, runs of only ~1 m are required for self-contained par-focal IFUs. For $\lambda < 2.5$ μm, fused silica is a suitable material, although it is not currently possible to go below ~0.35 μm, depending on fiber length. Newer techniques and materials may improve the situation. For $\lambda > 2.5$ μm, alternative materials are required, such as ZrF4 and, at longer wavelengths, chalcogenide glasses, but these are much more expensive and difficult to use than fused silica.

(ii) *Conservation of Etendue*. Fibers suffer from focal ratio degradation, a form of modal diffusion, which results in a speeding up of the output beam with respect to the input. Viewed as an increase in entropy, it results in a loss of information. In practice, this implies either reduced throughput, as the fiber output is vignetted by fixed-size spectrograph optics, or a loss in spectral resolution if the spectrograph is oversized to account for the faster beam from the fibers. This effect may be severe at the slow focal ratios pr oduced by many telescopes ($F > 10$) but is relatively small for fast beams ($F < 5$). Thus, its effect may be mitigated by speeding up the input beam by attaching lenslets directly to the fibers, either individually (for MOS) or in a close-packed array (for IFS). If the spectrograph operates in both fiber- and beam-fed modes (e.g., Figure 8), it is also necessary to use lenslets at the slit to convert the beam back to the original speed.

(iii) *Efficient coupling to the telescope*. This requires that the fibers are relatively thick to match the physical size of the sampling at the telescope focus which is typically $\chi_s = 0.1''$ for IFS and $\chi_s = 2''$ for MOS. By conservation of Etendue, the physical size of the fiber aperture is

$$d_f = \chi_s D_T F_f \qquad [29]$$

where F_f is the focal ratio of the light entering the fiber and D_T is the diameter of the telescope aperture. Using the above values with $F_f = 4$, as required to maintain Etendue, implies $15 < d_f < 300$ μm for an 8 m or $60 < d_f < 1200$ μm for a 30 m telescope. Except at the lower end, this implies the use of multimode (step-index) fibers. However, the need in IFS to oversize the fibers to account for manufacturing errors and current requirements of $\chi_s \geq 0.2''$ to produce sufficient SNR in a single spatial sample gives a practical limit of $d_f > 50$ μm which is in the multimode regime.

Recently, the use of photonic crystal fibers in astronomy has been investigated. These, together with other improvements in material technology may improve the performance of optical fibers, which will be of special relevance to ELTs where, for example, there will be a need to couple very fast beams into fibers.

Cooling for the Near-Infrared

As shown in Figure 12, it is necessary to cool instruments for use in the infrared. This is to prevent thermal emission from the instrument becoming the dominant noise source. The minimum requirement is to place a cold stop at a conjugate of the telescope pupil. However, much of the rest of the instrument requires cooling because of the nonzero emissivity of the optics. The telescope also requires careful design. For example, the Gemini telescopes are optimized for the NIR by undersizing M2 so that the only stray light entering the final focal surface is from the cold night sky.

Irrespective of the thermal background and wavelength, all detectors (see Table 4), require cooling to reduce internally generated noise to acceptable levels. For the CCDs used in the optical, the cryogen is liquid nitrogen, but for the infrared, cooling with liquid helium is required.

Structure

As discussed above, the necessity for instruments to scale in size with the telescope aperture cannot be achieved by simply devising a rigid structure, since the required materials are generally not affordable (instruments on 8 m telescopes cost roughly €/$ 5–15 M, including labor). Furthermore, the stability also depends critically on the mounting of the individual optical elements, where the choice of material is constrained by the need for compliance with the optical materials. The solution adopted for GMOS was to design the collimator and camera mounting so that variations in the gravity vector induce a translation in each component orthogonal to the optical axis without inducing tilts or defocus. Therefore, the only effect of instability is a translation of the image of the slit on the detector. Since the slit area is rigidly attached to the telescope interface, the image of a celestial object does not move with respect

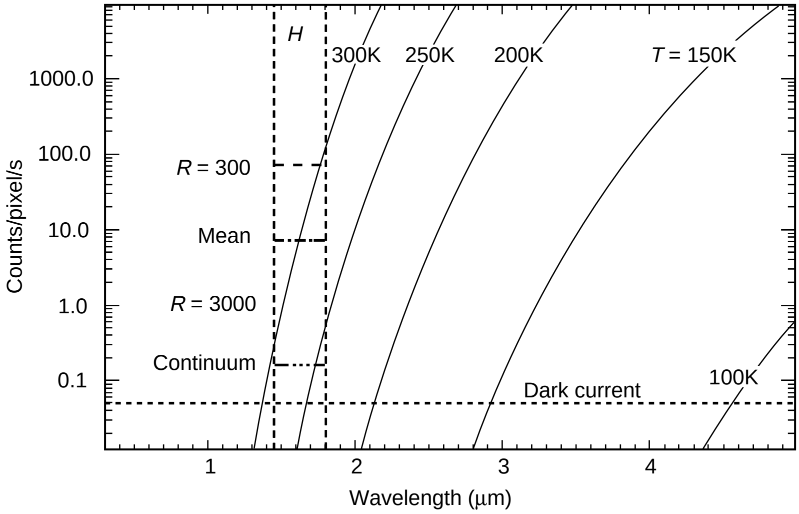

Figure 12 Illustration of the need for cooling in NIR instruments. This is modeled for the case of a spectrograph on an 8 m telescope with $0.2'' \times 0.2''$ sampling with both throughput and total emissivity of 50%. The curves indicate the cumulative thermal background for wavelengths shortward of that shown on the scale. For comparison, a typical dark current is shown for NIR detectors (see Table 4). Also shown is the background from the night sky for the case of the *H*-band, for two values of spectral resolution, *R*. The mean sky background level is shown for both values of *R*. For $R = 3000$, the continuum level found between the strong, narrow OH-lines which make up most of the signal is also shown. To observe in the *H*-band, cooling to −40 °C is required for high spectral resolution, but is unnecessary at low resolution. At longer wavelengths, fully cryogenic cooling with liquid nitrogen is generally required.

to the slit providing that the telescope correctly tracks the target. The motion of the image of the slit on the detector is then corrected by moving the detector orthogonal to the optical axis to compensate. Taking account of nonrepeatable motion in the stucture and the detector mounting, it proved possible to attain the desired stability in open-loop by measuring the flexure experienced with the control system turned off, making a look-up table to predict the required detector position for each attitude setting (elevation and attitude) and applying the correction by counting pulses sent to the stepper motors which control the detector translation.

Some instruments require much greater stability than GMOS does. The flexure-control outlined above may be augmented by operation in closed-loop so that nonrepeatable movements may be accounted for. However, this ideally requires a reference light source which illuminates the optics in the same way as the light from the astronomical target and which is recorded by the science detector without adversely impacting the observation. Alternatively, partial feedback may be supplied by repeated metrology of key optical components such as fold-mirrors.

One strategy to measure radial velocities to very small uncertainities (a few meters per second) is to introduce a material (e.g., iodine) into the optical path, which is present throughout the observations. This absorbs light from the observed object's continuum at one very well-defined wavelength. Thus, instrument instability can be removed by measuring the centroid of the desired spectral line in the object relative to the fixed reference produced by the absorption cell. However, flexure must still be carefully controlled to avoid blurring the line during the course of a single exposure.

Mounting of instruments via a fiber-feed remotely from the telescope, at a location where they are not subjected to variations in the gravity vector, is a solution for applications where great stability is required. But, even here, care must be taken with the fore-optics and pickoff system attached to the telescope, and to account for modal noise induced by changes in the fiber configuration as the telescope tracks.

Finally, the instrument structure has other functions. For cooled instruments, a cryostat is required inside which most of the optical components are mounted. For uncooled instruments, careful control of temperature is needed to avoid instrument motion due to thermal expansion. This requires the use of an enclosure which not only blocks out extraneous light, but provides a thermal buffer against changes in the ambient temperature during observations and reduces thermal gradients by forced circulation of air.

See also

Diffraction: Diffraction Gratings. **Fiber and Guided Wave Optics**: Fabrication of Optical Fiber. **Imaging**: Adaptive Optics. **Instrumentation**: Telescopes; Spectrometers. **Spectroscopy**: Fourier Transform Spectroscopy; Hadamard Spectroscopy and Imaging.

Further Reading

Allington-Smith JR, Murray G, Content C, *et al.* (2002) Integral field spectroscopy with the Gemini Multiobject Spectrograph. *Publications of the Astronomical Society of the Pacific* 114: 892.

Bjarklev A, Broeng J and Sanchez Bjarklev AS (2003) *Photonic Crystal Fibers*. Kluwer.

Bowen IS (1938) The Image-Slicer: a device for reducing loss of light at slit of stellar spectrograph. *Astrophysical Journal* 88: 113.

Carrasco E and Parry I (1994) A method for determining the focal ratio degradation of optical fibres for astronomy. *Monthly Notices of the Royal Astronomical Society* 271: 1.

Courtes G (1982) *Instrumentation for Astronomy with Large Optical Telescope, IAU Colloq, 67*. In: Humphries CM (ed.) Astrophysics and Space Science Library, vol. 92, pp. 123. Dordrecht: Reidel.

Glazebrook K and Bland-Hawthorn J (2001) Microslit nod-shuffle spectrocopy: a technique for achieving very high densities of spectra. *Publications of the Astronomical Society of the Pacific* 113: 197.

Lee D, Haynes R, Ren D and Allington-Smith JR (2001) Characterisation of microlens arrays for astronomical spectroscopy. *PASP* 113: 1406.

McLean I (1997) *Electronic Imaging in Astronomy.* Chichester, UK: John Wiley and Sons Ltd.

Palmer C (2000) *Diffraction Grating Handbook*. Richardson Grating Laboratory, 4th edn. Rochester, NY.

Ridgway SH and Brault JW (1984) Astronomical Fourier transform spectroscopy revisited. *Ann. Rev. Astron. Astrophys.* 22.

Saha SK (2002) Modern optical astronomy: technology and impact of interferometry. *Review of Modern Physics* 74: 551.

Schroeder D (2000) *Astronomical Optics*, 2nd edn. San Diego, CA: Academic Press.

van Breugel W and Bland-Hawthorn J (eds) *Imaging the Universe in Three Dimensions*, ASP Conference Series, vol. 195.

Weitzel L, Krabbe A, Kroker H, *et al.* (1996) 3D: The next generation near-infrared imaging spectrometer. *Astron. & Astrophys. Suppl.* 119: 531.

Ellipsometry

J N Hilfiker, J A Woollam & Co., Inc., Lincoln, NE, USA

J A Woollam, University of Nebraska, Lincoln, NE, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ellipsometry measures a change in polarization as light reflects from or transmits through a material structure. The polarization-change is represented as an amplitude ratio, Ψ, and a phase difference, Δ. The measured response is dependent on optical properties and thickness of each material. Thus, ellipsometry is primarily used to determine film thickness and optical constants. However, it is also applied to the characterization of composition, crystallinity, roughness, doping concentration, and other material properties associated with a change in optical response.

Interest in ellipsometry has grown steadily since the 1960s as it provided the sensitivity necessary to measure nanometer-scale layers used in microelectronics. Today, the range of applications has spread to basic research in physical sciences, semiconductor, data storage, flat panel display, communication, biosensor, and optical coating industries. This widespread use is due to increased dependence on thin films in many areas and the flexibility of ellipsometry to measure most material types: dielectrics, semiconductors, metals, superconductors, organics, biological coatings, and composites of materials.

This article provides a fundamental description of ellipsometry measurements along with the typical data analysis procedures. The primary applications of ellipsometry are also surveyed.

Light and Materials

Ellipsometry measurements involve the interaction between light and material.

Light

Light can be described as an electromagnetic wave traveling through space. For ellipsometry, it is adequate to discuss the electric field behavior in space and time, also known as polarization. The electric field of a wave is always orthogonal to the propagation direction. Therefore, a wave traveling along the z-direction can be described by its x- and y-components. If the light has completely random orientation and phase, it is considered unpolarized. For ellipsometry, we are interested in the case where the electric field follows a specific path and traces out a distinct shape at any point. This is known as polarized light. When two orthogonal light waves are in-phase, the resulting light will be linearly polarized (Figure 1a). The relative amplitudes determine the resulting orientation. If the orthogonal waves are 90° out-of-phase and equal in amplitude, the resultant light is circularly polarized (Figure 1b). The most general polarization is 'elliptical', which combines orthogonal waves of arbitrary amplitude and phase (Figure 1c). This is how ellipsometry gets its name.

Materials

Two values are used to describe optical properties, which determine how light interacts with a material.

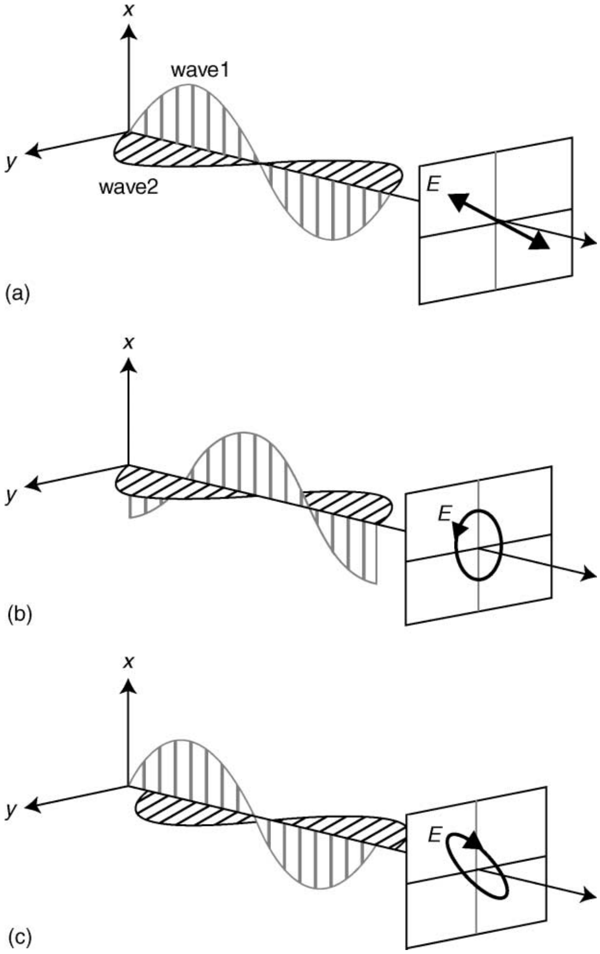

Figure 1 Orthogonal waves combine to demonstrate polarization: (a) linear; (b) circular; and (c) elliptical.

They are generally represented as a complex number. The complex refractive index ($\tilde{n}$) consists of the index (n) and extinction coefficient (k):

$$\tilde{n} = n + \mathrm{i}k \qquad [1]$$

Alternatively, the optical properties can be represented as the complex dielectric function:

$$\tilde{\varepsilon} = \varepsilon_1 + \mathrm{i}\varepsilon_2 \qquad [2]$$

with the following relation between conventions:

$$\tilde{\varepsilon} = \tilde{n}^2 \qquad [3]$$

The index describes the phase velocity for light as it travels through a material compared to the speed of light in vacuum, c:

$$v = \frac{c}{n} \qquad [4]$$

Light slows as it enters a material with higher index. Because frequency remains constant, the wavelength will shorten. The extinction coefficient describes the loss of wave energy to the material. It is related to the absorption coefficient, α, as:

$$\alpha = \frac{4\pi k}{\lambda} \qquad [5]$$

Light loses intensity in an absorbing material, according to Beer's Law:

$$I(z) = I(0)\,\mathrm{e}^{-\mathrm{i}\alpha z} \qquad [6]$$

Thus, the extinction coefficient relates how quickly light vanishes in a material. These concepts are demonstrated in Figure 2, where a light wave travels from air into two different materials of varying properties.

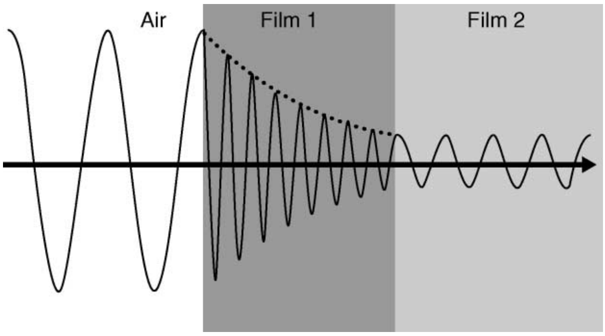

Figure 2 Wave travels from air into absorbing Film 1 and then transparent Film 2. The phase velocity and wavelength change in each material depending on index of refraction (Film 1: $n = 4$, Film 2: $n = 2$).

The optical constants for TiO_2 are shown in Figure 3 from the ultraviolet (UV) to the infrared (IR). The optical constants are wavelength dependent with absorption ($k > 0$) occurring in both UV and IR, due to different mechanisms that take energy from the light wave. IR absorption is commonly caused by molecular vibration, phonon vibration, or free-carriers. UV absorption is generally due to electronic transitions, where light provides energy to excite an electron to an elevated state. A closer look at the optical constants in Figure 3 shows that real and imaginary optical constants are not independent. Their shapes are mathematically coupled together through Kramers–Krönig consistency. Further details are covered later in this article.

Interaction Between Light and Materials

Maxwell's equations must remain satisfied when light interacts with a material, which leads to boundary conditions at the interface. Incident light will reflect and refract at the interface, as shown in Figure 4. The angle between the incident ray and sample normal

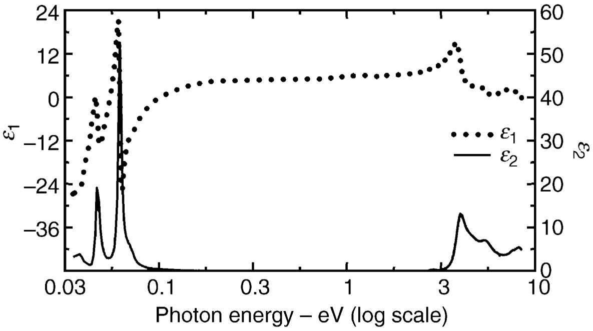

Figure 3 Complex dielectric function for TiO_2 film covering wavelengths from the infrared (small eV) to the ultraviolet (high eV).

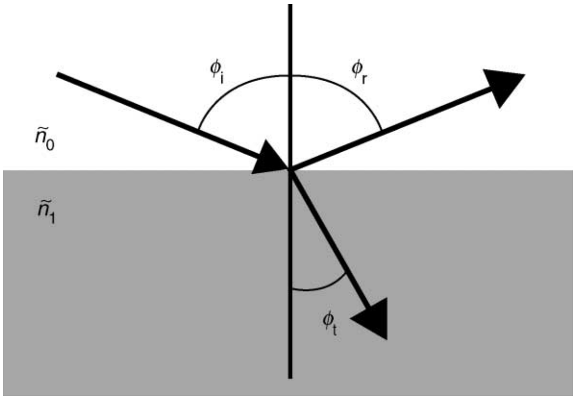

Figure 4 Light reflects and refracts according to Snell's law.

(ϕ_i) will be equal to the reflected angle, (ϕ_r). Light entering the material is refracted to an angle (ϕ_t) given by:

$$n_0 \sin(\phi_i) = n_1 \sin(\phi_t) \qquad [7]$$

The same occurs at every interface where a portion reflects and the remainder transmits at the refracted angle. This is illustrated in Figure 5. The boundary conditions provide different solutions for electric fields parallel and perpendicular to the sample surface. Therefore, light can be separated into orthogonal components with relation to the plane of incidence. Electric fields parallel and perpendicular to the plane of incidence are considered p- and s-polarized, respectively. These two components are independent for isotropic materials and can be calculated separately. Fresnel described the amount of light reflected and transmitted at an interface between materials:

$$r_s = \left(\frac{E_{0r}}{E_{0i}}\right)_s = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \qquad [8a]$$

$$r_p = \left(\frac{E_{0r}}{E_{0i}}\right)_p = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_i \cos\theta_t + n_t \cos\theta_i} \qquad [8b]$$

$$t_s = \left(\frac{E_{0t}}{E_{0i}}\right)_s = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \qquad [8c]$$

$$t_p = \left(\frac{E_{0t}}{E_{0i}}\right)_p = \frac{2n_i \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \qquad [8d]$$

Thin film and multilayer structures involve multiple interfaces, with Fresnel reflection and transmission coefficients applicable at each. It is important to track the relative phase of each light component to correctly determine the overall reflected or transmitted beam. For this purpose, we define the film phase thickness as:

$$\beta = 2\pi\left(\frac{t_1}{\lambda}\right) n_1 \cos\theta_1 \qquad [9]$$

The superposition of multiple light waves introduces interference that is dependent on the relative phase of each light wave. Figure 5 illustrates the combination of light waves in the reflected beam and their corresponding Fresnel calculations.

Ellipsometry Measurements

For ellipsometry, primary interest is measurement of how p- and s-components change upon reflection or transmission relative to each other. In this manner, the reference beam is part of the experiment. A known polarization is reflected or transmitted from the sample and the output polarization is measured. The change in polarization is the ellipsometry measurement, commonly written as:

$$\rho = \tan(\Psi)\, e^{i\Delta} \qquad [10]$$

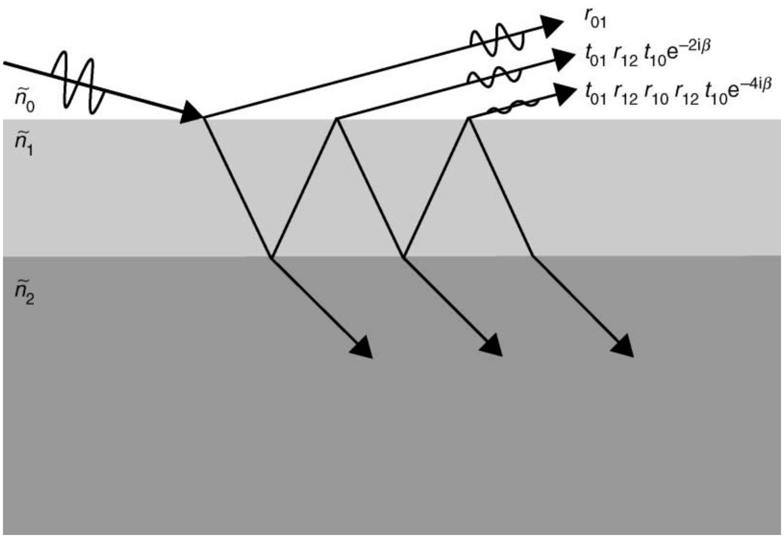
r_{01}
$t_{01} r_{12} t_{10} e^{-2i\beta}$
$t_{01} r_{12} r_{10} r_{12} t_{10} e^{-4i\beta}$
$\tilde{n}_0$
$\tilde{n}_1$
$\tilde{n}_2$

Figure 5 Light reflects and refracts at each interface, which leads to multiple beams in a thin film. Interference between beams depends on relative phase and amplitude of the electric fields. Fresnel reflection and transmission coefficients can be used to calculate the response from each contributing beam.

An example ellipsometry measurement is shown in Figure 6. The incident light is linear with both p- and s-components. The reflected light has undergone amplitude and phase changes for both p- and s-polarized light, and ellipsometry measures their changes.

The primary methods of measuring ellipsometry data all consist of the following components: light source, polarization generator, sample, polarization analyzer, and detector. The polarization generator and analyzer are constructed of optical components that manipulate the polarization: polarizers, compensators, and phase modulators. Common ellipsometer configurations include rotating analyzer (RAE), rotating polarizer (RPE), rotating compensator (RCE), and phase modulation (PME).

The RAE configuration is shown in Figure 7. A light source produces unpolarized light, which is sent through a polarizer. The polarizer passes light of a preferred electric field orientation. The polarizer axis is oriented between the p- and s-planes, such that both arrive at the sample surface. The linearly polarized light reflects from the sample surface, becoming elliptically polarized and travels through a continuously rotating polarizer (referred to as the analyzer). The amount of light allowed to pass will depend on the polarizer orientation relative to the electric field 'ellipse' coming from the sample. The detector converts light from photons to electronic signal to determine the reflected polarization. This information is used with the known input polarization to determine the polarization change caused by the sample reflection. This is the ellipsometry measurement of Ψ and Δ.

Single Wavelength Ellipsometry (SWE)

SWE uses a single frequency of light to probe the sample. This results in a pair of data, Ψ and Δ, used to determine up to two material properties. The optical design can be simple, low-cost, and accurate. Lasers are an ideal light source with well-known wavelength and significant intensity. The optical elements can be optimized to the single wavelength. However, there is relatively low information content (two measurement values). SWE is excellent for quick measurement of

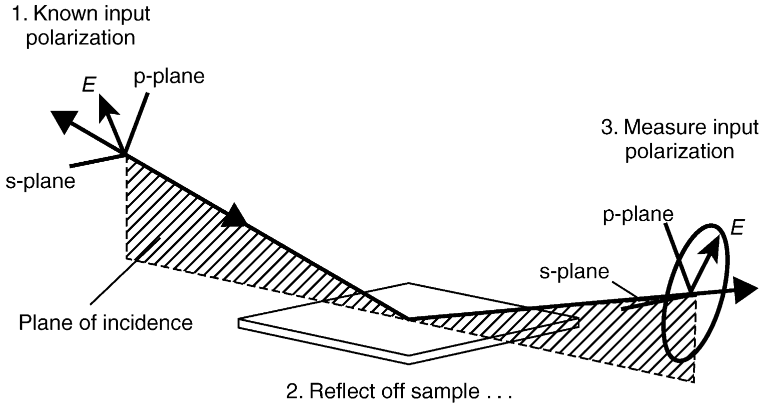
1. Known input polarization
E
p-plane
s-plane
Plane of incidence
3. Measure input polarization
p-plane
E
s-plane
2. Reflect off sample . . .

Figure 6 Typical ellipsometry configuration, where linearly polarized light is reflected from the sample surface and the polarization change is measured to determine the sample response.

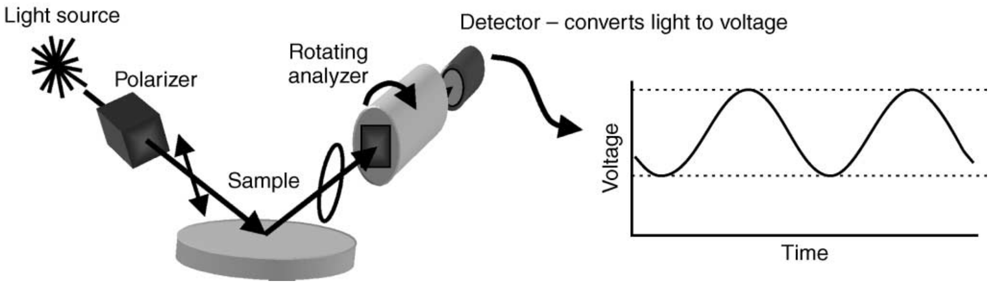

Figure 7 Rotating analyzer ellipsometer configuration uses a polarizer to define the incoming polarization and a rotating polarizer after the sample to analyze the outgoing light. The detector converts light to a voltage with the time-dependence leading to measurement of the reflected polarization.

nominally known films, like SiO_2 on Si. Care must be taken when interpreting unknown films, as multiple solutions occur for different film thickness. The data from a transparent film will cycle through the same values as the thickness increases. From Figure 5, it can be seen that this is related to interference between the multiple light beams. The refracted light may travel through different thicknesses to emerge in-phase with the first reflection, but is the returning light delayed by one wavelength or multiple wavelengths? Mathematically, the thickness cycle can be determined as:

$$D_\phi = \frac{\lambda}{2\sqrt{\tilde{n}_1^2 - \tilde{n}_0^2 \sin^2\phi}} \qquad [11]$$

This is demonstrated in Figure 8, where the complete thickness cycle is shown for SiO_2 on Si at 75° angle of incidence and 500 nm wavelength. A bare silicon substrate would produce data at the position of the star. As the film thickness increases, the data will move around this circle in the direction of the arrow. From eqn [11], the thickness cycle is calculated as 226 nm. Thus the cycle is completed and returns to the star when the film thickness reaches 226 nm. Any data point along the cycle represents a series of thicknesses which depend on how many times the cycle has been completely encircled. For example, the square represents a data point for 50 nm thickness. However, the same data point will represent 276 and 502 nm thicknesses. A second data point at a new wavelength or angle would help determine the correct thickness.

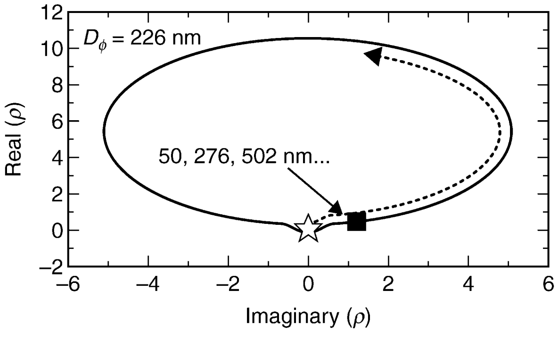

Figure 8 Data from a single wavelength will cycle through the same points as the film thickness increases. The star represents the starting point, with no SiO_2 film on Si. As the thickness increases, the data follow the cycle in the direction of the arrow. After the cycle is complete (thickness = D_ϕ), the data repeat the cycle. Thus, any point along the curve represents multiple possible thicknesses – example shown by a square at 50 nm thickness, which is also equal to 276 and 502 nm.

Spectroscopic Ellipsometry (SE)

Spectroscopic measurements solve the 'period' problem discussed for SWE. Data at surrounding wavelengths insure the correct thickness is determined, even as there remain multiple solutions at one wavelength in the spectrum. This is demonstrated in Figure 9, showing the spectroscopic data for the first three thickness results. The SE data oscillations clearly distinguish each thickness. Thus, a single thickness solution remains to match data at multiple wavelengths.

SE provides additional information content for each new wavelength. While the film thickness will remain constant regardless of wavelength, the optical constants will change across the spectrum. The optical constant shape contains information regarding the material microstructure, composition, and more. Different wavelength regions will provide the best information for each different material property. For this reason, SE systems have been developed to cover very wide spectral regions from the UV to the IR.

Spectroscopic measurements require an additional component, a method of wavelength selection. The most common methods include the use of monochromators, Fourier-transform spectrometers, and gratings or prisms with detection on a diode array. Monochromators are typically slow as they sequentially scan through wavelengths. Spectrometers and diode arrays allow simultaneous measurement at multiple wavelengths. This has become popular with the desire for high-speed SE measurements.

Variable Angle Ellipsometry

Ellipsometry measurements are typically performed at oblique angles, where the largest changes in polarization occur. The Brewster angle, defined as:

$$\tan(\phi_B) = \frac{n_1}{n_0} \qquad [12]$$

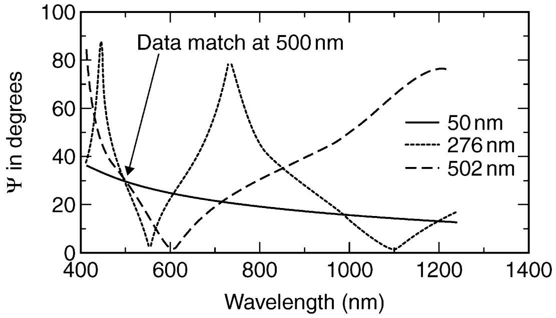

Figure 9 Three spectroscopic ellipsometry measurements that match at 500 nm wavelength, but are easily distinguishable as a function of wavelength.

gives the angle where reflection of p-polarized light goes through a minimum. Ellipsometry measurements are usually near this angle, which can vary from 55° for low-index dielectrics to 75° or 80° for semiconductors and metals.

It is also common to measure at multiple angles of incidence in ellipsometry. This allows additional data to be collected for the same material structure under different conditions. The most important change introduced by varying the angle is the change in path length through the film as the light refracts at a different angle. Multiple angles do not always add new information about a material structure, but the extra data builds confidence in the final answer.

In Situ

Within the last decade, it has become common to employ optical diagnostics during processing. *In situ* ellipsometry allows the optical response to be monitored in real-time. This has also led to feedback control for many processes. While *in situ* is generally restricted to a single angle of incidence, there is a distinct advantage to collecting data at different 'times' during processing to get unique 'glimpses' of the sample structure as it changes.

In situ SE is commonly applied to semiconductors. Conventional SE systems measure from the UV to NIR, which matches the spectral region where semiconductors absorb due to electronic transitions. The shape and position of the absorption can be very sensitive to temperature, composition, crystallinity, and surface quality, which allows SE to closely monitor these properties. *In situ* SE has also been used to monitor numerous material types, including metals and dielectrics. A promising *in situ* application is for multilayer optical coatings, where thickness and index can be determined in real-time to allow correction of future layers, to optimize optical performance.

Data Analysis

Ellipsometry measures changes in polarization. However, it is used to determine the material properties of interest, such as film thickness and optical constants. In the case of a bulk material, the equations derived for a single reflection can be directly inverted to provide the 'pseudo' optical constants from the ellipsometry measurement, ρ:

$$\langle \tilde{\varepsilon} \rangle = \sin^2(\phi)\left[1 + \tan^2(\phi)\left(\frac{1-\rho}{1+\rho}\right)\right] \qquad [13]$$

This equation assumes there are no surface layers of any type. However, there is typically a surface oxide or roughness for any bulk material and the direct inversion would incorporate these into the bulk optical constants. The more common procedure used to deduce material properties from ellipsometry measurements follows the flowchart in Figure 10. Regression analysis is required because an exact equation cannot be written. Often the answer is overdetermined with hundreds of experimental data points for a few unknowns. Regression analysis allows all of the measured data to be included when determining the solution.

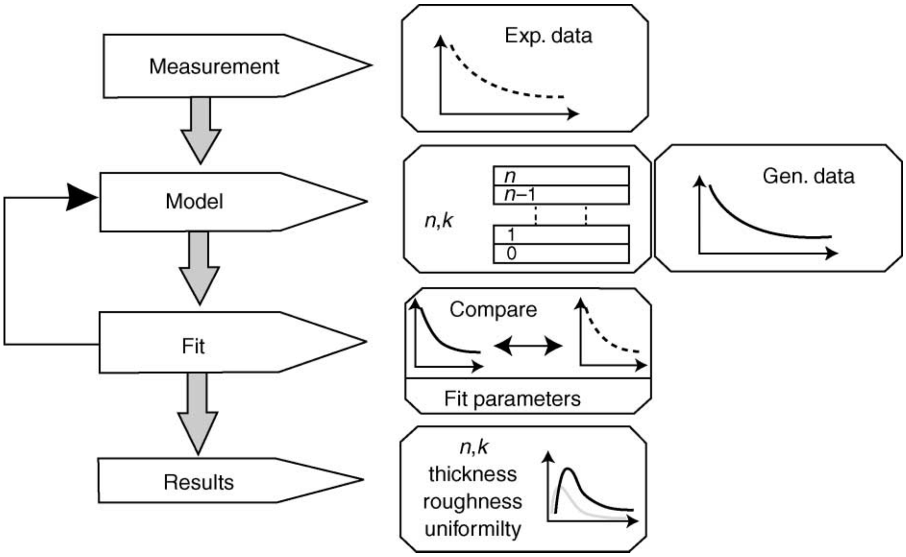

Figure 10 Flowchart for ellipsometry data analysis.

Data analysis proceeds as follows. After the sample is measured, a model is constructed to describe the sample. The model is used to calculate the predicted response from Fresnel's equations. Thus, each material must be described with a thickness and optical constants. If these values are not known, an estimate is given for the purpose of the preliminary calculation. The calculated values are compared to experimental data. Any unknown material properties can then be varied to improve the match between experiment and calculation. The number of unknown properties should not exceed the information content contained in the experimental data. For example, a single-wavelength ellipsometer produces two data points (Ψ,Δ) which allows a maximum of two material properties to be determined. Finding the best match between the model and experiment is typically done through regression. An estimator, like the mean squared error (MSE), is used to quantify the difference between curves. The unknown parameters are allowed to vary until the minimum MSE is reached.

Care must be taken to ensure the best answer is found, corresponding to the lowest MSE. For example, Figure 11a shows the MSE curve versus film thickness for a transparent film on silicon. There are multiple 'local' minima, but the lowest MSE value occurs at a thickness of 749 nm. This corresponds to the correct film thickness. It is possible that the regression algorithm will mistakenly fall into a 'local' minima, depending on the starting thickness and the MSE structural conditions. Comparing the results by eye for the lowest MSE and a local minima easily distinguish the true minima (see Figures 11b and c).

Ellipsometry Characterization

The two most common material properties studied by ellipsometry are film thickness and optical constants. In addition, ellipsometry can characterize material properties that affect the optical constants. Crystallinity, composition, resistivity, temperature, and molecular orientation can all affect the optical constants and in turn be measured by ellipsometry. This section details many of the primary applications important to ellipsometry.

Film Thickness

The film thickness is determined by interference between light reflecting from the surface and light traveling through the film. Depending on the relative phase of the rejoining light to the surface reflection, there can be constructive or destructive interference.

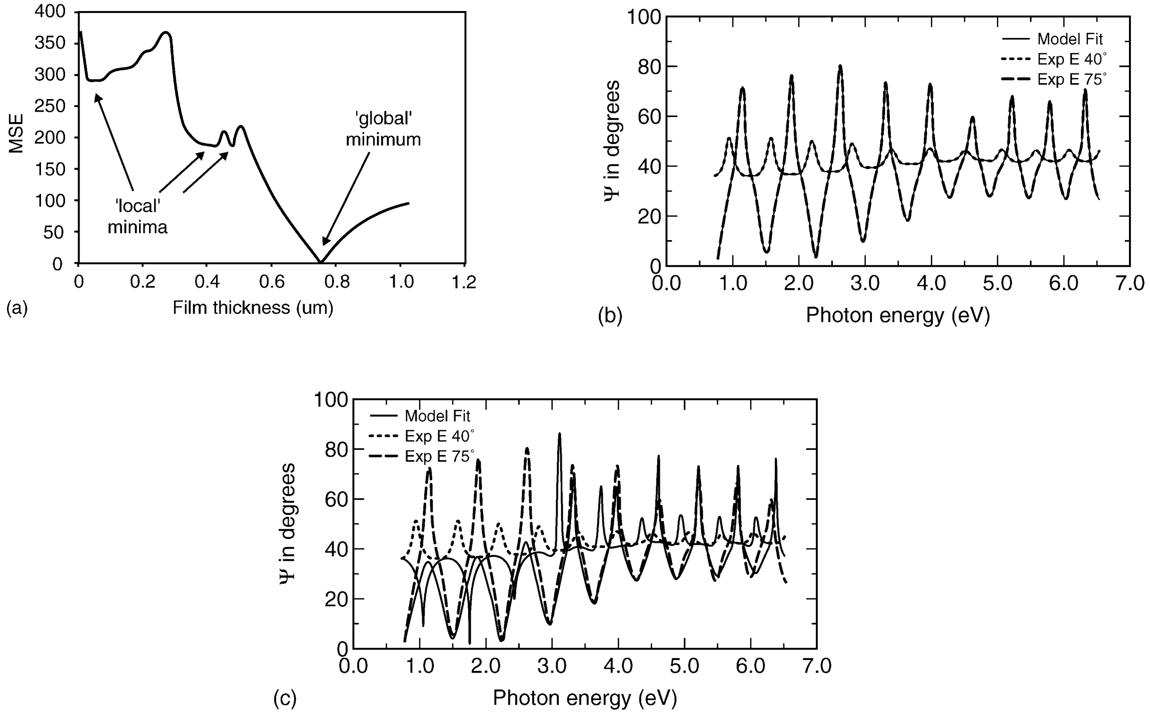

Figure 11 (a) MSE curve versus thickness shows the 'global' minimum where the best match between model and experiment occurs, and 'local' minima that may be found by the regression algorithm, but do not give the final result. (b) The experimental data and corresponding curves generated for the model at the 'global' minimum. (c) Similar curves at the 'local' minimum near 0.45 microns thickness is easily distinguishable as an incorrect result.

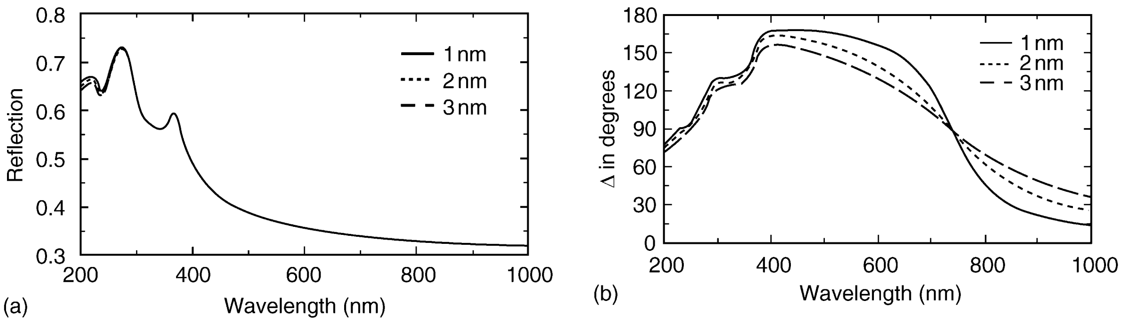

Figure 12 (a) Reflected intensity and (b) ellipsometric delta for three thin oxides on silicon show the high sensitivity of Δ to nanometer-scale films not available from the intensity measurement.

The interference involves both amplitude and phase information. The phase information from Δ is very sensitive to films down to submonolayer thickness. Figure 12 compares reflected intensity and ellipsometry for the same series of thin SiO_2 layers on Si. There are large variations in Δ, while the reflectance for each film is nearly the same.

Ellipsometry is typically used for films with thickness ranging from sub-nanometers to a few microns. As films become greater than several tens of microns thick, it becomes increasingly difficult to resolve the interference oscillations, except with longer infrared wavelengths, and other characterization techniques become preferred.

Thickness measurements also require a portion of the light to travel through the entire film and return to the surface. If the material is absorbing, thickness measurements by optical instruments will be limited to thin, semi-opaque layers. This limitation can be circumvented by measuring in a spectral region where there is lower absorption. For example, an organic film may strongly absorb UV and IR light, but remain transparent at mid-visible wavelengths. For metals, which strongly absorb at all wavelengths, the maximum layer for thickness determination is typically ~100 nm.

Optical Constants

Thickness measurements are not independent of the optical constants. The film thickness affects the path length of light traveling through the film, but the index determines the phase velocity and refracted angle. Thus, both contribute to the delay between surface reflection and light traveling through the film. Both *n* and *k* must be known or determined along with the thickness to get the correct results from an optical measurement.

The optical constants for a material will vary for different wavelengths and must be described at all wavelengths probed with the ellipsometer. A table of optical constants can be used to predict the response at each wavelength. However, it is less convenient to adjust unknown optical constants on a wavelength-by-wavelength basis. It is more advantageous to use all wavelengths simultaneously. A dispersion relationship often solves this problem, by describing the optical constant shape versus wavelength. The adjustable parameters of the dispersion relationship allow the overall optical constant shape to match the experimental results. This greatly reduces the number of unknown 'free' parameters compared to fitting individual *n*, *k* values at every wavelength.

For transparent materials, the index is often described using the Cauchy or Sellmeier relationship. The Cauchy relationship is typically given as:

$$n(\lambda) = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \qquad [14]$$

where the three terms are adjusted to match the refractive index for the material. The Sellmeier relationship enforces Kramers–Krönig (KK) consistency, which ensures the optical dispersion retains a realistic shape. The Cauchy is not constrained by KK consistency and can produce unphysical dispersion. The Sellmeier relationship can be written as:

$$\varepsilon_1 = \frac{A\lambda^2\lambda_0^2}{(\lambda^2 - \lambda_0^2)} \qquad [15]$$

Absorbing materials will often have a transparent wavelength region that can be modeled with the Cauchy or Sellmeier relationships. However, the absorbing region must account for both real and imaginary optical constants. Many dispersion relationships use oscillator theory to describe the absorption for materials, which include the Lorentz, Harmonic, and Gaussian oscillators. They all share similar attributes, where the absorption features are described with an amplitude, broadening, and center energy (related to frequency of light).

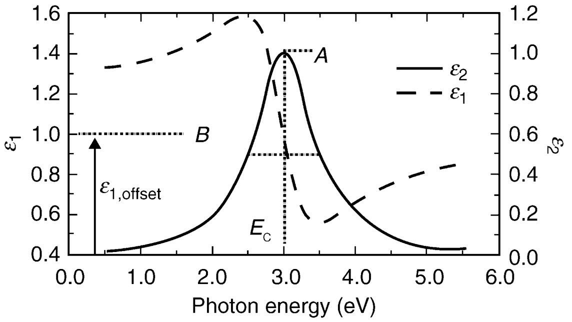

Figure 13 Lorentz oscillator illustrating the primary oscillator parameters: amplitude (*A*), broadening (*B*), and center energy (E_c) to describe the imaginary dielectric function shape and an offset ($\varepsilon_{1,\text{offset}}$) to help match the real component after KK transformation has defined its shape.

Kramers–Krönig consistency is used to calculate the shape of the real component after the imaginary behavior is described by the oscillator. An offset to the real component is added to account for extra absorption outside the measured spectral region. The Lorentz oscillator can be written as:

$$\tilde{\varepsilon} = \varepsilon_{1,\text{offset}} + \frac{AE_c}{E_c^2 - E^2 - iBE} \qquad [16]$$

where the parameters for amplitude (A), broadening (B), center energy (E_c), and offset ($\varepsilon_{1,\text{offset}}$) are also shown in Figure 13 for a typical Lorentz oscillator. The energy, E, is related to the frequency of a wave, ν:

$$E = h\nu \cong \frac{1240}{\lambda_{\text{nm}}} \qquad [17]$$

where h is Planck's constant and the wavelength, λ, is given in nanometers. More advanced dispersion models, like the Tauc–Lorentz and Cody–Lorentz, will include terms to describe the bandgap energy.

Mixing Materials

When two or more materials are mixed on a microscopic level, an effective medium approximation (EMA) may be used to determine the resulting optical constants. In the case of the Bruggemann EMA, which is commonly used, the mixture optical constants ($\tilde{\varepsilon}_{\text{eff}}$) relate to those of the individual materials, as:

$$f_a \frac{\varepsilon_a - \varepsilon_{\text{eff}}}{\varepsilon_a + 2\varepsilon_{\text{eff}}} + f_b \frac{\varepsilon_b - \varepsilon_{\text{eff}}}{\varepsilon_b + 2\varepsilon_{\text{eff}}} = 0 \qquad [18]$$

This can be interpreted as small particles of material A suspended in host material B. The length-scale for mixed particles must satisfy certain electromagnetic equations: typically smaller than one tenth the wavelength of light. In practice, EMA theory is useful for studying very thin surface roughness or interfacial intermixing of materials. These cases are both generally approximated by mixing the two surrounding materials in equal portions. In the case of surface roughness, the second material is void ($n = 1$). EMA theory has also been extended for application to porous materials where the directional dependence of the inclusions is handled mathematically.

Crystallinity

Semiconductors such as Si, are widely used materials, but their properties depend strongly on crystallinity. The UV absorption features in crystalline silicon are broadened and shifted as the material becomes more amorphous. This change in optical properties related to the degree of crystallinity has been used to advantage with ellipsometry measurements to monitor semiconductors and other films. Polysilicon films are used in both the display and semiconductor industry. The degree of crystallinity varies for different process conditions and can be monitored optically to ensure consistent material properties.

Composition

The composition of many alloys will affect the optical constants. The strongest changes often occur in the absorbing region, with shifts in position and amplitude of absorption features. For example, the electronic transitions in $Hg_{1-x}Cd_xTe$ move to higher energy with Cd concentration increase (Figure 14). This material is used for IR detectors that require precise control of composition. Spectroscopic ellipsometry performs this task in real-time with instant feedback to correct the composition during processing. Other applications include AlGaN and InGaN for optoelectronics.

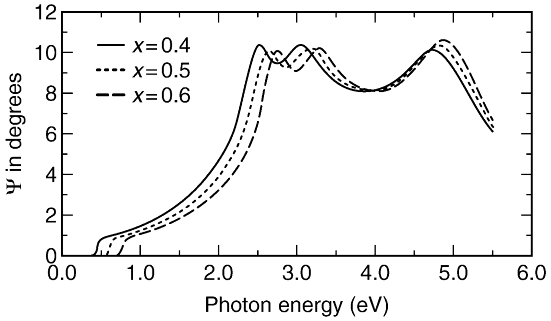

Figure 14 Optical properties for $Hg_{1-x}Cd_xTe$ vary with changes in composition.

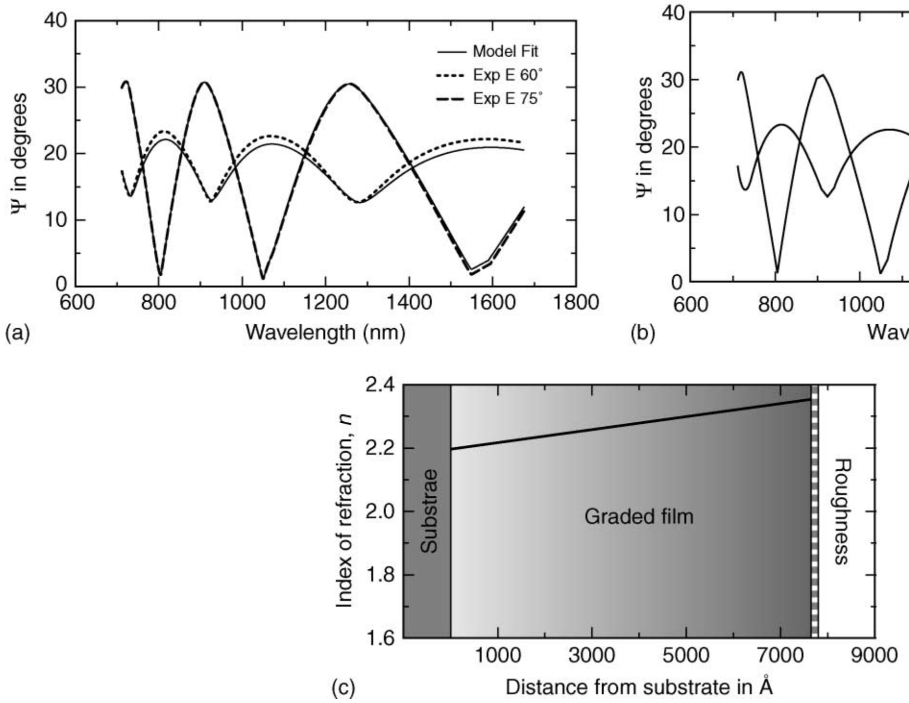

Figure 15 Experimental ellipsometry data and corresponding fit when the model is described as (a) a homogeneous single-layer and (b) a graded film with index variation through the film. (c) The best model to match subtleties in experimental data allows the index to increase toward the surface of the film, with a thin roughness layer on the surface.

Doping Concentration

Dopants in a material will introduce absorption in the infrared due to free-carriers. Spectroscopic ellipsometry measurements at long wavelengths can characterize this optical absorption, thus characterizing the doping concentration. This is common for highly doped semiconductors and transparent conductors such as indium tin oxide (ITO).

Optical Variation (Grading)

Many thin film properties change vertically throughout the film (along the direction perpendicular to the surface). This is most often induced by processing conditions, whether intentional or unintentional. Figure 15a shows the fit to experimental spectroscopic ellipsometry data taken from a single-layer film when it is modeled as a homogeneous layer. To improve the agreement between experimental and model-generated curves, the index was allowed to vary in a linear manner throughout the film. The best fit is shown in Figure 15b, where the model includes the index variation and a thin surface roughness layer. The sample structure is demonstrated in Figure 15c.

Optical Anisotropy

Many materials are optically anisotropic; i.e., their optical properties vary in different film directions. Typical ellipsometry measurements assume no cross-coupling between p- and s-polarizations. This cannot be assumed with anisotropic materials, which has led to 'generalized' ellipsometry measurements. Generalized ellipsometry measures additional information regarding the p- to s- and s- to p- conversion upon reflection. This allows characterization of anisotropic materials, which contain directionally dependent optical constants.

Conclusions

Ellipsometry is a common optical technique for measuring thin films and bulk materials. It relies on the polarization changes due to reflection or transmission from a material structure to deduce material properties, like film thickness and optical constants. This technique continues to develop as the requirement for thin film characterization increases.

See also

Geometrical Optics: Lenses and Mirrors; Prisms. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids.

Polarization: Introduction; Matrix Analysis. **Semiconductor Physics:** Band Structure and Optical Properties. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Aspnes DE (1985) The accurate determination of optical properties by ellipsometry. In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, pp. 89–112. Orlando FL: Academic Press.

Azzam RMA and Bashara NM (1987) *Ellipsometry and Polarized Light*. Amsterdam, The Netherlands: Elsevier Science B.V.

Boccara AC, Pickering C and Rivory J (eds) (1993) *Spectroscopic Ellipsometry*. Amsterdam: Elsevier Publishing.

Collins RW, Aspnes DE and Irene EA (eds) (1998) *Proceedings from the Second International Conference on Spectroscopic Ellipsometry*. Thin Solid Films, vols. 313–314.

Gottesfeld S, Kim YT and Redondo A (1995) Recent applications of ellipsometry and spectroellipsometry in electrochemical systems. In: Rubinstein I (ed.) *Physical Electrochemistry: Principles, Methods, and Applications*. New York: Marcel Dekker.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 425–479. San Diego, CA: Academic Press.

Johs B, Woollam JA, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part II: Advanced applications. *Optical Metrology*, vol. CR72, pp. 29–58. Bellingham, Washington: SPIE, Bellingham.

Johs B, Hale J, Ianno NJ, *et al.* (2001) Recent developments in spectroscopic ellipsometry for *in situ* applications. In: Duparré A and Singh B (eds) *Optical Metrology Roadmap for the Semiconductor, Optical, and Data Storage Industries II*, vol. 4449, pp. 41–57. Bellingham, Washington: SPIE.

Roseler A (1990) *Infrared Spectroscopic Ellipsometry*. Berlin: Akademie-Verlag.

Rossow U and Richter W (1996) Spectroscopic ellipsometry. In: Bauer G and Richter W (eds) *Optical Characterization of Epitaxial Semiconductor Layers*, pp. 68–128. Berlin: Springer-Verlag.

Tompkins HG (1993) *A User's Guide to Ellipsometry*. San Diego, CA: Academic Press.

Tompkins HG and Irene EA (eds) (in press) *Handbook of Ellipsometry*. New York: William Andrew Publishing.

Tompkins HG and McGahan WA (1999) *Spectroscopic Ellipsometry and Reflectometry*. New York: John Wiley & Sons, Inc.

Woollam JA (2000) Ellipsometry, variable angle spectroscopic. In: Webster JG (ed.) *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 109–116. New York: John Wiley & Sons.

Woollam JA and Snyder PG (1992) Variable angle spectroscopic ellipsometry. In: Brundle CR, Evans CA and Wilson S (eds) *Encyclopedia of Materials Characterization: Surfaces, Interfaces, Thin Films*, pp. 401–411. Boston, MA: Butterworth-Heinemann.

Woollam JA, Johs B, Herzinger CM, *et al.* (1999) Overview of variable angle spectroscopic ellipsometry (VASE), Part I: Basic theory and typical applications. *Optical Metrology*, vol. CR72, pp. 3–28. Washington: SPIE, Bellingham.

Photometry

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fundamentals of Photometry

Photometry is defined by the International Commission on Illumination (internationally known as CIE from the abbreviation of its French name: Commission Internationale de l'Eclairage) as 'measurement of quantities referring to radiation as evaluated according to a given spectral efficiency function, e.g., $V(\lambda)$ or $V'(\lambda)$.' A note to the above definition states that in many languages it is used in a broader sense, covering the science of optical radiation measurement. We will restrict our treatise to the above fundamental meaning of photometry.

Under spectral efficiency function we understand the spectral luminous efficiency function of the human visual system. The internationally agreed symbols are $V(\lambda)$, for photopic vision (daylight conditions) and $V'(\lambda)$ for scotopic vision (nighttime conditions). CIE standardized the $V(\lambda)$ function in 1924 and the $V'(\lambda)$ function in 1951; their spectral distribution is shown in Figure 1. The $V(\lambda)$ function was determined using mainly the so-called flicker photometric technique, where the light of a reference stimulus and of a test stimulus of varying wavelengths are alternatively shown to the human observer at a frequency at which the observer is already unable to perceive the difference in color (due to the different wavelength of the two radiations), but still perceives a

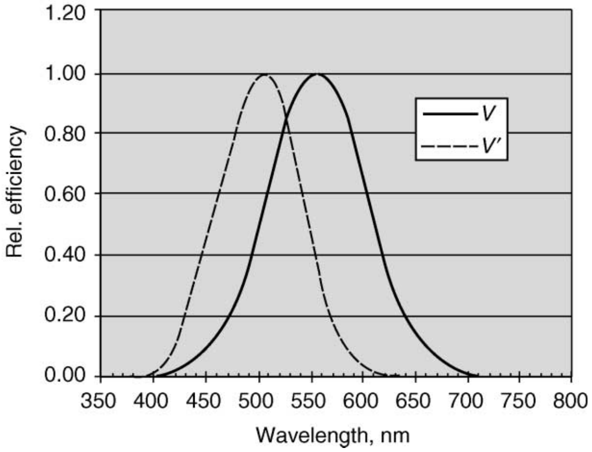

Figure 1 Spectral efficiency functions of the human eye under photopic $V(\lambda)$ and scotopic $V'(\lambda)$ conditions (see CIE The basis of physical photometry. CIE 18.2:1983).

flicker sensation if the luminances of the two stimuli are different. Adjusting the radiance of the test stimulus, we can reach the situation of minimum flicker sensation. In this case we state that we set the two stimuli to equal luminance. Luminance is the quantity that the human observer perceives as brightness (in the case of near white stimuli, see below); radiance is the physical counterpart measured in $\mathrm{W\,m^{-2}\,sr^{-1}}$. For the $V'(\lambda)$ function one can just project the two stimuli side by side, as under scotopic adaptation we cannot distinguish colors, only brightness differences are perceived, thus one can set equal scotopic luminance for the two stimuli by adjusting the brightness of the test stimulus until it agrees with that of the reference stimulus.

Illuminating engineering was mainly interested in describing the effect of near white light, and therefore a psychophysical correlation was selected that described the perceived brightness sensation reasonably well. For near white stimuli, additivity holds, i.e., if A, B, C, and D are four stimuli, and the A stimulus matches the B stimulus and the C stimulus matches the D stimulus, then the superposition of the A and C stimuli matches the superposition of the B and D stimuli. The 'match' word is used here to describe stimuli that produce the same perception. It could be shown that for these stimuli, if the radiance is weighted with the $V(\lambda)$ function, the constructed luminances will be equal, i.e:

if $\quad L_A = L_B$ and $L_C = L_D$

then $\quad L_A + L_C = L_B + L_D$, or $L_A + L_D = L_B + L_C$

where

$$L_x = K_m \int_{380\ \mathrm{nm}}^{780\ \mathrm{nm}} S_{\lambda x} V(\lambda) \mathrm{d}\lambda \qquad [1]$$

and x refers to A, or B, or C, or D, and $S_{\lambda x}$ is the spectral radiance distribution that produces the stimulus

$$S_\lambda = \frac{\mathrm{d}S(\lambda)}{\mathrm{d}\lambda}$$

where $S(\lambda)$ is the spectral radiance. K_m is the maximum value of the luminous efficacy of radiation, its value is $683\ \mathrm{lm\,W^{-1}}$ (see the discussion of luminous flux in the sub-section on photometric quantities). To be precise the integration should go from 0 nm to infinity, but it is usual to define the lower and upper wavelength limits of the visible spectrum as 380 nm and 780 nm. $V(\lambda)$ is defined between 360 nm and 830 nm, and $V'(\lambda)$ between 380 nm and 780 nm, see Figure 1.

We have to stress that the concept of luminance – and the entire system of the present-day photometry – is not to quantify brightness perception; it is only a reasonable approximation for near white stimuli. For colored lights a brightness–luminance discrepancy exists (called the Helmhotz–Kohlrausch effect: saturated colors look brighter than predicted by luminance). Luminance is, however, a good description for the visibility of fine details, thus it is a good concept for lighting calculations.

Photometry has a unique situation in the SI system of units: the candela (cd) is a base unit of the SI system, the only one that is connected to psychophysical phenomena. By 1979, in the definition of the 16th General Conference of Weights and Measures, the candela was traced back to radiation quantities; nevertheless it was kept as the base unit of photometry:

> The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of $1/683\ \mathrm{W\,sr^{-1}}$. ($540 \times 10^{12}$ Hz corresponds approximately to 555 nm.)

To be able to calculate a photometric quantity of radiation of other wavelengths, one of the psychophysical functions $V(\lambda)$ or $V'(\lambda)$ has to be used and eqn [1] applied. We will see later that besides $V(\lambda)$ and $V'(\lambda)$ in modern applications, a number of further spectral luminous efficiency functions might be used.

Photometric Quantities

As discussed in the previous section, from the point of view of vision, the most important quantity is luminance. Looking, however, at the definition of the base unit, it is obvious that from the physical point of view the definition of a quantity corresponding to radiant power, measured in watts, can help to bridge the gap between photometry and radiometry: luminous flux, measured in lumens (lm), with the symbol Φ, is defined by eqn [1], with $S_\lambda(\lambda)$ inserted in W m^{-1}.

Based on this more practical quantity of luminous flux and its unit, lm, the different quantities used in photometry can be built up as follows:

$$\textit{Luminous flux}: \quad \Phi = K_m \int_0^\infty \frac{\mathrm{d}\Phi_\mathrm{e}(\lambda)}{\mathrm{d}\lambda} V(\lambda)\mathrm{d}\lambda \qquad [2]$$

where $\Phi_\mathrm{e}(\lambda)$ is the radiant flux measured in W,

$$\Phi_{\mathrm{e},\lambda} = \frac{\mathrm{d}\Phi_\mathrm{e}(\lambda)}{\mathrm{d}\lambda}$$

is the spectral distribution (or spectral concentration) of the radiant flux (W m^{-1}). $K_m = 683$ lm W^{-1}, is the maximum value of the luminous efficacy of radiation for $\lambda_m \approx 555$ nm.

Similar equations can be written for scotopic luminous flux, with $V(\lambda)$ exchanged with $V'(\lambda)$, where $K'_m = 1700$ lm W^{-1} for $\lambda_m \approx 507$ nm. All further quantities can be defined both for photopic and scotopic conditions. Here we write them only as photopic quantities.

$$\textit{Luminous intensity}: \quad I = \frac{\mathrm{d}\Phi}{\mathrm{d}\Omega} \qquad [3]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, assuming a point source (see Figure 2). The unit of luminous intensity is the candela (cd = lm sr^{-1}).

$$\textit{Luminance}: \quad L = \frac{\partial^2\Phi}{\partial A \cos\Theta\, \partial\Omega} \qquad [4]$$

where $\mathrm{d}\Phi$ is the luminous flux traveling in an elementary solid angle $\mathrm{d}\Omega$, $\mathrm{d}A$ is the elementary surface area emitting the radiation, and Θ is the angle between the normal of $\mathrm{d}A$ and the direction of luminance measurement (see Figure 3). The unit of luminance is cd m^{-2} (in some older literature called nit).

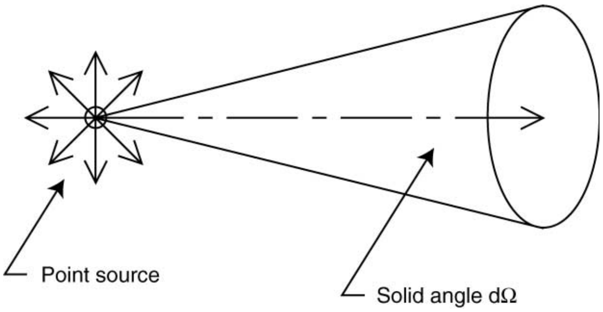

Figure 2 Concept of point source, solid angle, and luminous intensity.

$$\textit{Illuminance}: \quad E = \frac{\mathrm{d}\Phi}{\mathrm{d}A} \qquad [5]$$

where $\mathrm{d}\Phi$ is the luminous flux incident on the $\mathrm{d}A$ element of a surface (see Figure 4). The unit of illuminance is the lux (lx = lm m^{-2}).

A remark on the use of SI units and related quantities

No prefixes can be added to the SI units, thus irrespective whether one measures luminance based

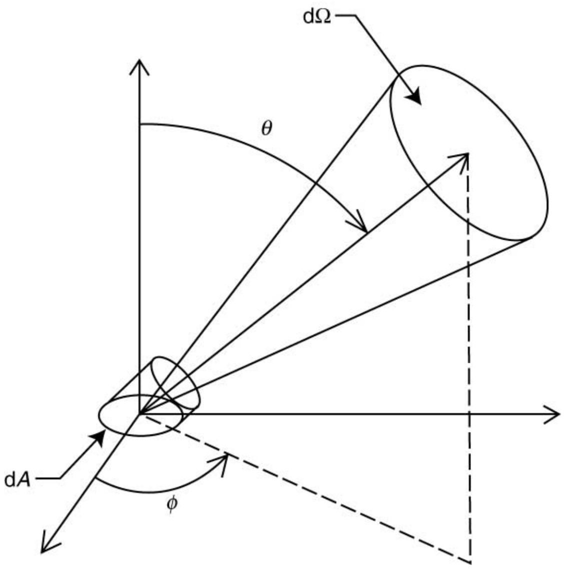

Figure 3 Geometry for the definition of luminance.

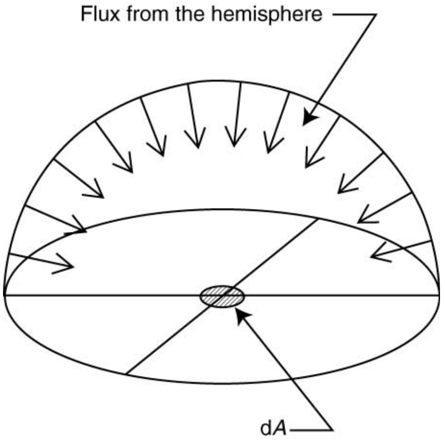

Figure 4 Illuminance is the total luminous flux per unit area incident at a point coming from a hemispherical solid angle.

on the photopic $V(\lambda)$ function, or the scotopic $V'(\lambda)$ function, the so-determined photopic luminance or scotopic luminance is measured in $\mathrm{cd\,m^{-2}}$ (no photopic or scotopic lumen, candela, etc., exist!). This often creates confusion, because only for a 555 nm monochromatic radiation will a $1\,\mathrm{cd\,m^{-2}}$ photopic or scotopic luminance produce equal visual sensation, for every other wavelength the two are not commensurable. One can only relate photopic measurement results to photopic ones, and scotopic measurement results to scotopic ones.

Concepts of Advanced Photometry

The quantities and units, as described in the previous section, are the quantities and units internationally agreed by the Meter Convention and the International Standards Organization (ISO). Modern photometric applications need, however, some further quantities and weighting functions. Thus CIE, the international organization for the development of standards in the field of optical radiation measurements, has defined a series of further weighting functions and quantities. The most important ones are the following:

$V_M(\lambda)$ function

In the blue part of the spectrum (below 460 nm) the values of the $V(\lambda)$ function turned out to be too low. For decades this was only of concern for the vision research community, but with the introduction of blue LEDs and other short wavelength emitting sources (e.g., blue channel of displays) this shortcoming of the $V(\lambda)$ function became of practical importance. The 'CIE 1988 2° spectral luminous efficiency function for photopic vision' corrects this anomaly (see Figure 5).

$V_{10}(\lambda)$ function

The official $V(\lambda)$ function is valid only for foveal vision (i.e., for targets that are smaller than 4° of visual angle). The foveal area of the retina is covered with a yellow pigmented layer (macula lutea) that absorbs in the blue part of the spectrum, at larger visual angles this screening is not effective anymore. The $V_{10}(\lambda)$ function was determined for a visual angle of 10° (see Figure 5). Its international recommendation is still under consideration at the time of writing this article and if accepted, it will be recommended for targets seen at approximately 10° off-axis, e.g., for measuring the photometric properties of traffic signs and signals, where the driver has to observe the information in the periphery of his or her eye.

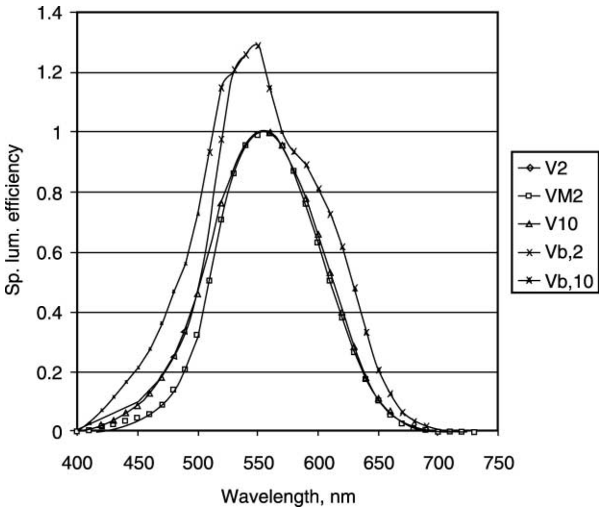

Figure 5 Spectral luminous efficiency (SPL) functions defined or under consideration for international adoption: V2: standard $V(\lambda)$ function; VM2: CIE 1988 $V_M(\lambda)$ function, this is equivalent to the brightness SPL for a point source; V10: 10° visual field SPL, Vb,2: 2° visual field brightness SPL, $V_{b,2}(\lambda)$; Vb,10: 10° visual field brightness SPL, $V_{b,10}(\lambda)$.

Brightness matching functions

As already mentioned in the Introduction, luminance is not a good correlate of brightness, which is a human perception. To find a better correlation of brightness the problem to be addressed is that brightness is a nonadditive phenomenon, i.e., in eqn [1] one cannot add (integrate) the monochromatic radiations to get a brightness correlation for a nonmonochromatic radiation. Brightness evaluating spectral luminous efficiency functions can be used only to compare monochromatic radiations. The CIE has compiled such functions for point sources, 2° and 10° visual field sizes. Figure 5 also shows these functions.

Due to the fact that the brightness perception is nonadditive in respect of the stimuli that produce it, no brightness photometry can be built that uses equations, as shown in eqn [1]. For brightness description we have to rely on the concept of equivalent luminance, a term the definition of which has recently been updated.

Equivalent luminance

Of a field of given size and shape, for a radiation of arbitrary relative spectral distribution L_{eq} : Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape which may be different from that of the field considered.

To build an instrument that measures this quantity is a real challenge for the future.

Advanced Use of Photometry

Based on above newly defined quantities, several attempts are under way to extend the usefulness of photometry in designing the human visual environment. The eye is an optical system and as in every such system, the depth of focus and the different aberrations of the system will decrease with decreasing pupil size. Pupil size will decrease with increasing illumination, and in the case of constant luminance with higher content of short wavelength radiation. Thus, there exists a school of researchers who advocate that increased blue content in the light has beneficial effects on vision, and one should extend the classical photopic-based photometry with a scotopic-based one to properly describe the visual effect of lighting.

Other investigations are concerned about the visibility at low light levels, the range used in street lighting (from about a few thousands of a candela per square meter to about a few candelas per square meter, according to one definitions: 10^{-3}cd m^{-2}–3 cd m^{-2}). In this mesopic range both rods and cones are contributing to vision, and this changes with lighting level and direction of view (for foveal vision, i.e., looking straight ahead, photopic photometry seems to hold even at low light levels). For peripheral visual angles brightness perception and the perception of an object (a signal, sign or obstacle in a nighttime driving situation) seem to have different spectral responsivity. In driving situations the necessary reaction time of the driver is an important parameter, thus experiments are going on to define a photometric system based on reaction time investigations.

In indoor situations apart from the necessary level of illumination, the observed glare is a contributor whether an environment will be accepted as pleasing or annoying. Illuminating engineering distinguishes between two types of glare: disability glare reduces visibility, discomfort glare is just an annoying experience without influencing the short-term task performance. An interesting question is the spectral sensitivity to discomfort glare, as it can influence not only indoor but also outdoor activity. Preliminary experiments seem to show that luminance sensitivity and discomfort glare sensitivity have different spectral distribution; glare sensitivity seems to peak at shorter wavelengths.

The above might be related to a further question, lying already at the boundaries of photometry, but that has to be considered in photometric design and measurement: the human daily and yearly rhythm (circadian and seasonal rhythm) of human activity coupled to hormone levels. They are influenced by light, as, for example, the hormone melatonin production is influenced by light exposure. Physiological investigations showed that melatonin production suppression has a maximum around 460 nm and might be coupled to a radiation sensitive ganglion cell in the retina. Whether discomfort sensation is mediated via the same neural pathway or via a visual one has not yet been decided. But photometry has to take these also into consideration and in the future, measurement methods and instruments to determine them, will have to be developed.

Advances in Photometric Measurements

Primary Standards

The main concern in photometry is that the uncertainty of photometric measurements is still much higher than that in other branches of physics. This is partly due to the higher uncertainty in radiometry and partly to the increase in uncertainty within the chain of uncertainty propagation from the National Laboratory to the workshop floor measurement.

In National Standards Laboratories, very sophisticated systems are used to determine the power of the incoming radiation and then elaborated spectroradiometric techniques are used to evaluate the radiation in the form of light, i.e., perform photometric measurements (e.g., NIST CIRCUS equipment, where multiple laser sources are used as power sources, and highly sophisticated methods to produce a homogeneous nonpolarized radiation field for the calibration of secondary photometric detectors).

There is, however, also a method to supply end users with absolute detectors for the visible part of the spectrum. Modern high-end Si photovoltaic cells have internal quantum efficiencies in the visible part of the spectrum of over 99.9%. Reflection losses at the silicon surface are minimized by using three or more detectors arranged in a trap configuration, where the light reflected from one detector is fed to the second one, from there to the third one, and eventually to some further ones. In a three detector configuration, as shown in Figure 6, the light from the third detector is reflected back to the second and from there to the first one. As every detector reflects only a small amount of radiation, by the fifth reflection practically

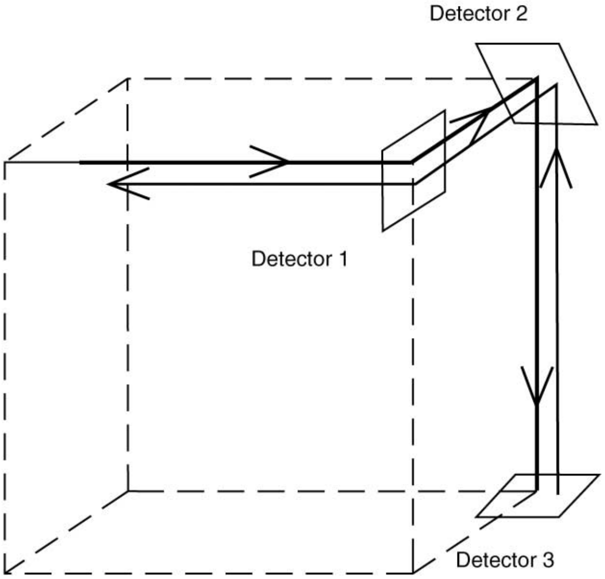

Figure 6 Schematic layout of a three Si-cell trap detector: light comes in from the left, is first partly absorbed, partly reflected on Detector 1, then on Detector 2, then on Detector 3, from where it is reflected back to Detector 2 and 1.

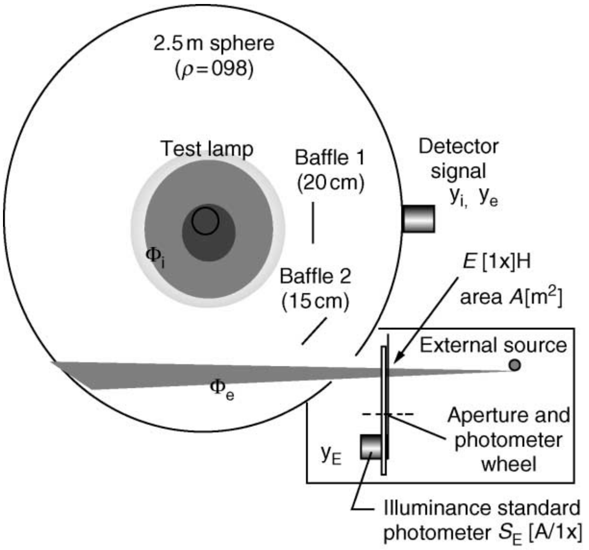

Figure 7 Arrangement of the absolute integrating sphere system developed at NIST for the detector-based total luminous flux calibration. By permission of IESNA from Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J. IES Summer* 21–26.

all the radiation is absorbed and contributes to the electric signal. Such trap detectors have an almost 100% quantum efficiency in the visible part of the spectrum, and can be used as photometric detectors if a well designed color correcting filter is applied in front of the detector.

Secondary Type Measurements

In practical photometry the three most important quantities to be measured are the total luminous flux of different lamps, the illuminance in a plane and the luminance.

Light source measurement

Total luminous flux. The two methods to measure the total luminous flux is to use a goniophotometer or a photometer (Ulbicht) sphere. In goniophotometry, recent years have not brought major breakthroughs, the automation of the systems got better, but the principles are unchanged.

The integrating sphere photometer (a sphere with inner white diffuse coating, where the lamp is in the middle of the sphere) used to be a simple piece of equipment to compare total luminous flux lamps against flux standards. In recent years a new technique has been developed at NIST–USA. This enables the absolute measurement of luminous flux from illuminance measurement, the fundamentals of this new arrangement being shown in Figure 7: the test lamp is as usual in the middle of the sphere, but now light from an external source is introduced into the sphere. An illuminance meter measures the flux entering from this source. The sphere detector compares the two signals (y_i and y_e). Knowing the absolute characteristics of the sphere (a difficult measurement) one can determine the total luminous flux of the test lamp using the absolute illuminance value. As illuminance is easily determined from luminous intensity (from eqns [3] and [5] one gets with $d\Omega = dA/r^2$ that $E = I/r^2$, where r is the distance between the source and the illuminated surface, supposed to be perpendicular to the direction to the source), this technique permits us to derive the total luminous flux scale from illuminance or luminous intensity measurement using an integrating sphere.

Luminous intensity of LEDs. An other major break-through achieved during the past years was the unified measurement of LED intensity. Light-emitting diodes became, in recent years, important light sources for large scale signaling and signing, and it is foreseen that they will become important contributors in every field of light production (from car headlamps to general illumination). The most fundamental parameter of the light of an LED is its luminous intensity. The spatial power distribution of LEDs is usually collimated, but the LEDs often squint, as seen in Figure 8. In the past, some manufacturers measured the luminous intensity in the direction of maximum emission, others used the direction of the optical axis for this quantity. The highly collimated character of the radiation made measurements in far field rather difficult. Therefore, CIE recommended a new term and measuring geometry: average LED intensity can be measured under two measuring conditions, as shown in Figure 9. The detector has to be set in the direction of the LED mechanical axis (discussions are still going on as to what the reference direction should be with modern surface mounted LEDs, as with those the mechanical axis is ill-defined, the normal to the base-plane could be a better reference direction). The detector has to have an exactly 1.00 cm^2 circular aperture, and the distance between this aperture and the tip of the LED is for condition A, $d = 0.316$ m, and for Condition B, $d = 0.100$ m (these two distances with the 1.00 cm^2 detector area provide 0.001 sr (steradians) and 0.01 sr opening angles). Recent international round-robins have shown that, based on the new recommendations, agreement between different laboratories decreased from the 10 to 20% level to 1 to 2%. The remaining difference is mainly due to the fact that the LEDs emit in narrow wavelength bands, and the transfer of the calibration value for the 100 mm^2 detector from a white (CIE Standard Illuminant A color temperature) incandescent lamp to the narrow band LED emission is still uncertain, mainly due to stray light effects in the spectral responsivity and emission measurement.

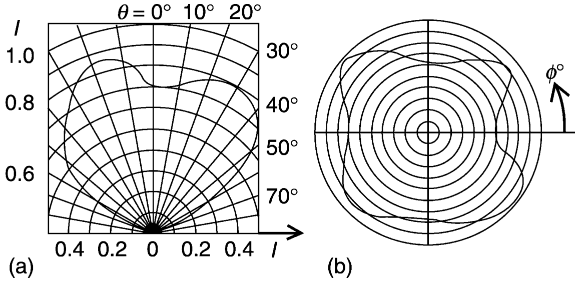

Figure 8 Spatial light distribution of an LED, Figure 8a shows the distribution of a 'squinting' LED in a plane including the optical axis, Figure 8b shows light distribution in a plane perpendicular to the optical axis. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

Luminance distribution measurement

The human observer sees luminance (and color) differences. Thus for every illuminating engineering design task the luminance distribution in the environment is of utmost importance. Traditionally this was measured using a spot-luminance meter, aiming the device into a few critical directions. The recent development of charge coupled device (CCD) two-dimensionally sensitive arrays (and other, e.g., MOS-FET, CMOS, Charge injection device (CID), charge imaging matrix (CIM) systems: as for the time being the CCD technology provides best performance, we will refer to two-dimensional electronic image capture devices as to CCD cameras) opened the possibility of using an image-capturing camera for luminance distribution measurements. Such measurements are badly needed in display calibration, near-field photometry (photometry in planes nearer than in which the inverse square law holds), testing of car headlamp light distribution, glare, and homogeneity studies indoors and outdoors, etc.

Solid-state cameras have the big advantage over older type vacuum-tube image capturing devices, that the geometric position and alignment of the single pixels is well-defined and stays constant. Nowadays, CCD image detector chips are mass produced, and one can get a variety of such devices and cameras starting with some small resolution

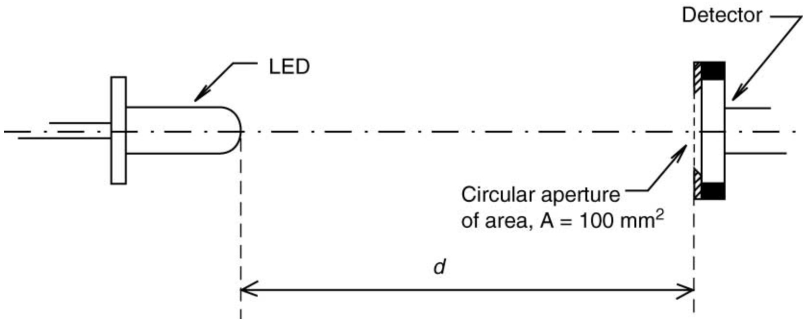

Figure 9 Schematic diagram of CIE Standard Conditions for the measurement of Average LED Intensity. Distance $d = 0.316$ m for Condition A and $d = 0.100$ m for Condition B. By permission of the Commission Internationale de l'Eclairage, from the Publication "Measurement of LEDs" CIE 127-1997; CIE Publications are obtainable from the CIE Central Bureau: Kegelgasse 27, A-1033 Wien, Austria.

(few thousand pixels) devices up to cameras with tens of mega pixel resolution. Detectors are now available with internal intensification enabling measurements down to a few photons per second intensity levels.

Main problems with these detectors are:

- spectral and absolute nonuniformities of the single pixels (see Figure 10), where a spatial homogeneity map of a CCD two-dimensional array detector is shown; the irregular 3% sensitivity change on the surface of the detector is negligible for imaging purposes, but has to be corrected in cases of photometric measurements. Even if the receptor chip would have an absolutely homogeneous sensitivity, there would be a drop in response from the middle of the imaging area to the boarders: Light reaching the edges of the detector reach the detector at an oblique angle and this produces a decrease of sensitivity with α^4, where α is the angle of incidence, measured from the middle of the lens to the given pixel of the detector and the surface normal of the detector.
- aliasing effects if the pixel resolution is not large enough to show straight lines as such when they are not in line with a pixel row or column.
- nonlinearity and cross-talk among the adjacent pixels. Figure 11 shows the so-called 'inverse gamma' characteristic of a CCD camera. The digital electronic output of the camera shows a $Y = E^{-\gamma}$ type function, where Y is the output DAC (digital-analog converter) values, E is the irradiance of the pixel, and γ is the exponent (this description comes from the film industry, where the film density depends exponentially on the irradiation; display devices have usually a nonlinear input (DAC value) – output (luminance) characteristic, and the camera inverse gamma value corrects for this output device characteristic. This is, however, not required if the camera is used for photometric measurements; this built-in nonlinearity has to be corrected in the evaluating software if the camera is intended for photometric measurements.
- to be able to perform photometric measurements the camera has to have a spectral responsivity corresponding to the CIE $V(\lambda)$-function. Many cameras have built-in filters to do this – eventually also red and blue filters to be able to capture color – but the color correction of cameras, where the correction is made by small adjacent filter chips, is usually very poor. Figure 12 shows the spectral

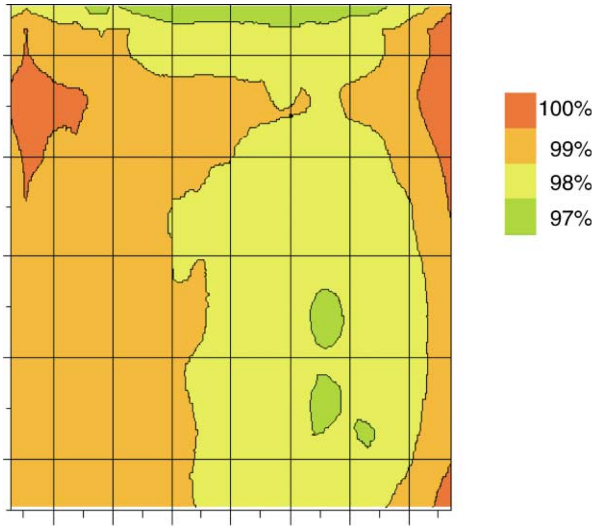

Figure 10 Spatial homogeneity of a two-dimensional CCD array.

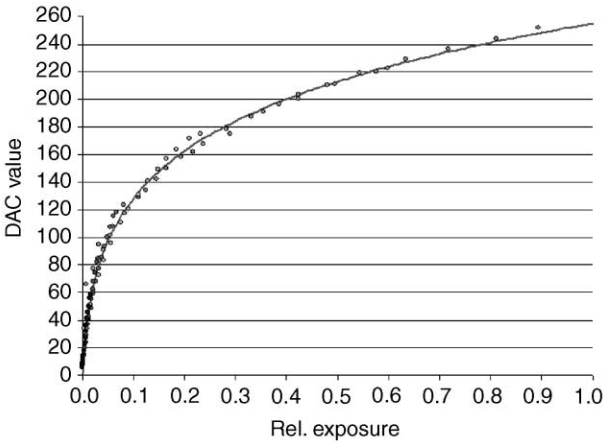

Figure 11 'Inverse gamma' characteristic of a commercial digital photographic camera, measurement points are shown at different speed settings, curve is a model function representative of the camera response.

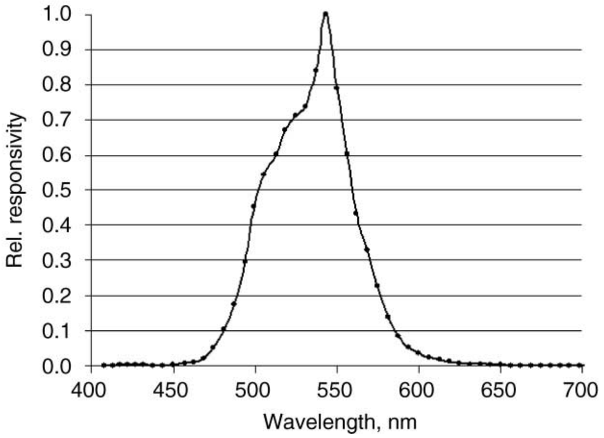

Figure 12 Spectral sensitivity of a commercial digital photographic camera.

sensitivity curve of a digital photographic camera; the output signal is produced by an internal matrix transformation of the signals produced by adjacent pixels equipped with different color filters. Figure 13 shows the spectral sensitivity of a CCD camera specially designed for photometric measurements. Naturally, meaningful photometric measurements can be made only with such a camera. Very often, however, an approximate luminance distribution is enough, and then a picture captured by a digital photographic camera, plus one luminance measurement of a representative object for absolute calibration, suffices.

Above nonspectral systematic errors can be corrected by appropriate soft-ware, so a CCD camera photometer, as shown schematically in Figure 14, is well suited to measure display characteristics, indoor and outdoor luminance distributions. The challenge for illuminating engineering is at present how the many millions of luminance values can be evaluated to get to meaningful light measurement data.

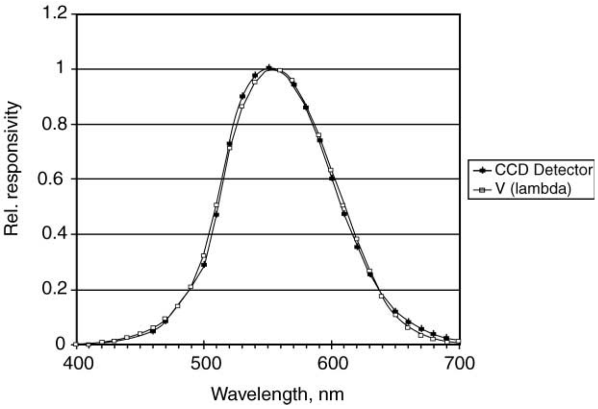

Figure 13 Spectral sensitivity of a CCD camera specially designed for photometric measurements. Kindly supplied by INPHORA Inc.

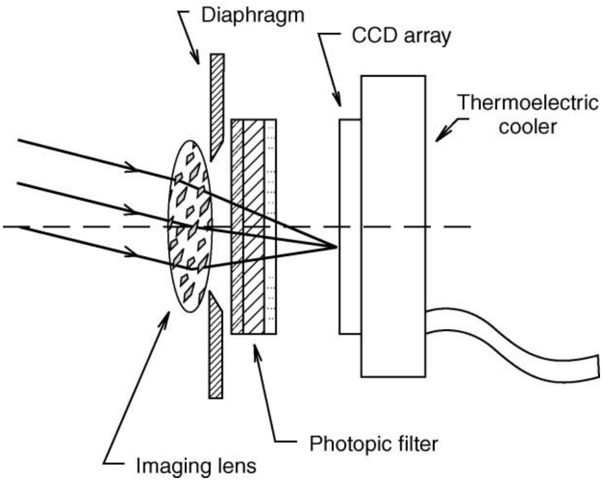

Figure 14 Cross-section of a photometric CCD camera.

The real challenge will come if visual science provides better hints how the human visual system evaluates the illuminance distribution on the retina, and instrument manufacturers will be able to capture signals corresponding to those produced by the receptor cells and provide the necessary algorithms our brain uses to get to brightness, lightness, color, luminance contrast, and glare type of output information.

Concluding Remarks

Advances in optical instrumentation, both in the field of light sources and detectors – coupled with the possibilities modern digital computation (eventually in the future increasing the use of neural networks) – provide already many new measurement technical solutions and further ones are certainly underway.

The use of LEDs needs the rethinking of many classical illuminating engineering concepts, from visibility and glare evaluation, evenness of illumination to color rendering. All of them are coupled with problems in basic photometry. Thus, there is a need to re-evaluate concepts used in design techniques. The new area-sensitive detectors provide methods of determining classical photometric quantities of entire visual fields in one shot, but already foreshadow the development of new quantities that correlate better with visual perceptions.

List of Units and Nomenclature

Terms with an $*$ refer to definitions published by the CIE in the International Lighting Vocabulary, CIE 17.4:1986, where further terms and definitions related to light and lighting are to be found.

Candela*: SI unit of luminous intensity: The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of 1/683 watt per steradian. (16th General Conference of Weights and Measures, 1979).

$$1 \text{ cd} = 1 \text{ lm sr}^{-1}$$

Equivalent luminance*: Luminance of a comparison field in which monochromatic radiation of frequency 540×10^{12} Hz has the same brightness as the field considered under the specified photometric conditions of measurement; the comparison field must have a specified size and shape which may be different from that of the field considered.

$$\text{unit} : \text{cd m}^{-2}$$

Notes:

1. Radiation at a frequency of 540×10^{12} Hz has a wavelength in standard air of 555.016 nm.
2. A comparison field may also be used in which the radiation has any relative spectral distribution, if the equivalent luminance of this field is known under the same conditions of measurement.

Far field photometry: Photometry where the inverse square law is valid.

Flicker photometer*: Visual photometer in which the observer sees either an undivided field illuminated successively, or two adjacent fields illuminated alternately, by two sources to be compared, the frequency of alteration being conveniently chosen so that it is above the fusion frequency for colours but below the fusion frequency for brightnesses.

Fovea*: Central part of the retina, thin and depressed, which contains almost exclusively cones and forming the site of most distinct vision.

Note: The fovea subtends an angle of about 0.026 rad (1.5 degree) in the visual field.

Goniophotometer*: Photometer for measuring the directional light distribution characteristics of sources, luminaires, media or surfaces.

Illuminance*: Quotient of the luminous flux $d\Phi_v$ incident on an element of the surface containing the point, by the area dA of that element.

Equivalent definition. Integral, taken over the hemisphere visible from the given point, of the expression $L_v \cos\theta \, d\Omega$, where L_v is the luminance at the given point in the various directions of the incident elementary beams of solid angle $d\Omega$, and θ is the angle between any of these beams and the normal to the surface at the given point.

$$E_v = \frac{d\Phi_v}{dA} = \int_{2\pi sr} L_v \cos\theta \, d\Omega$$

$$\text{unit} : \text{lx} = \text{lm m}^{-2}$$

Inverse square law: The illumination at a point on a surface varies directly with the luminous intensity of the source, and inversely as the square of the distance between the source and the point if the source is seen as a point source.

Lumen*: SI unit of luminous flux: Luminous flux emitted in unit solid angle (steradian) by a uniform point source having a luminous intensity of 1 candela. (9th General Conference of Weights and Measures, 1948).

Equivalent definition. Luminous flux of a beam of monochromatic radiation whose frequency is 540×10^{12} hertz and whose radiant flux is 1/683 watt.

Luminance*: Quantity defined by the formula

$$L = \frac{\partial^2\Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the luminous flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{cd m}^{-2}$$

Luminous intensity*: Quotient of the luminous flux $d\Phi_v$ leaving the source and propagated in the element of solid angle $d\Omega$ containing the given direction, by the element of solid angle:

$$I_v = \frac{d\Phi_v}{d\Omega}$$

$$\text{unit} : \text{cd} = \text{lm sr}^{-1}$$

Lux*: SI unit of illuminance: Illuminance produced on a surface of area 1 square meter by a luminous flux of 1 lumen uniformly distributed over that surface.

$$1 \text{ lx} = 1 \text{ lm m}^{-2}$$

Note: Non-metric unit: **lumen per square foot** (lm ft^{-2}) or **footcandle** (fc) (USA) = 10.764 lx.

Near-filed photometry: photometry made in the vicinity of an extended source, so that the inverse square law is not valid.

Photometer (or Ulbicht) sphere*: A hollow sphere, whitened inside. Owing to the internal reflexions in the sphere, the illumination on any part of the sphere's inside surface is proportional to the luminous flux entering the sphere, or produced inside the sphere by a lamp. The illuminance of the internal sphere wall is measured via a small window.

Pixel: The individual picture elements of an image or elements in a display that can be addressed individually.

Radiance*: Quantity defined by the formula

$$L = \frac{\partial^2\Phi}{\partial A \cos\theta \, \partial\Omega}$$

where $\partial\Phi$ is the radiant flux transmitted by an elementary beam passing through the given point and propagating in the solid angle $d\Omega$ containing the given direction; dA is the area of a section of that beam containing the given point, θ is the angle between the normal to that section and the direction of the beam.

$$\text{unit} : \text{W m}^{-2} \text{ srad}$$

Retina*: Membrane situated inside the back of the eye that is sensitive to light stimuli; it contains photoreceptors, the cones and the rods, and nerve cells that interconnect and transmit to the optic nerve the signals resulting from stimulation of the photoreceptors.

Spectral*: An adjective that, when applied to a quantity X pertaining to electromagnetic radiation, indicates:

- either that X is a function of the wavelength λ, symbol: $X(\lambda)$,
- or that the quantity referred to is the spectral concentration of X, symbol: $X_\lambda \equiv \mathrm{d}X/\mathrm{d}\lambda$.

X_λ is also a function of λ and in order to stress this may be written $X_\lambda(\lambda)$ without any change of meaning.

Spectral luminous efficiency function: for photopic vision, $V(\lambda)$; for scotopic vision $V'(\lambda)$*: Ratio of the radiant flux at wavelength λ_m to that at wavelength λ such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_m is chosen so that the maximum value of this ratio is equal to 1.

Unless otherwise indicated, the values used for the spectral luminous efficiency in photopic vision are the values agreed internationally in 1924 by the CIE (*Compte Rendu* 6^e session, p.67), completed by interpolation and extrapolation (*Publications CIE* No.18 (1970), p.43 and No. 15 (1971), p.93), and recommended by the International Committee of Weights and Measures (CIPM) in 1972. For *scotopic vision*, the CIE in 1951 adopted, for young observers, the values published in *Compte Rendu* 12^e session, Vol. 3, p. 37, and ratified by the CIPM in 1976. These values define respectively the $V(\lambda)$ or $V'(\lambda)$ functions.

Total luminous flux: luminous flux of a source emitted into 4π steradians.

Trap detector: Detector array prepared from detectors of high internal quantum efficiency, where the reflected radiation of the first detector is directed to the second one, and so on so that practically all the radiation is absorbed by one of the detectors (the radiation is “trapped” in the detector system).

See also

Displays. **Incoherent sources:** Lamps.

Further Reading

Commission Internationale de l’Eclairage: *The basis of physical photometry* (1983) CIE 18.2.

Commission Internationale de l’Eclairage: Proceedings of the CIE Symposium 2002 on *Temporal and Spatial Aspects of Light and Colour Perception and Measurement*, August 2002: Publ.CIE x025:2003.

Cromer CL, Eppledauer G, Hardis JE, Larason TC and Parr AC (1993) National Institute of Standards and Technology detector-based photometric scale. *Applied Optics* 32: 2936–2948.

Decusatis C (ed.) (1997) *Handbook of Applied Photometry.* Woodbury, New York: American Institute of Physics.

Rea MS (ed.) (2000) Illuminating Engineering Society of North America. *The IESNA Lighting Handbook: Reference and Application*, 9th edn.

Kaiser PK and Boynton RM (1966) *Human Color Vision*, 2nd edn. Washington, DC: Optical Society of America.

McCluney WR (1994) *Introduction to radiometry and photometry.* Boston, London: Artech House.

Ohno Y and Bergman R (2003) Detector-referenced integrating sphere photometry for industry. *J IES Summer* 21–26.

Walsh JWT (1953) *Photometry.* London: Constable.

Scatterometry

J C Stover, The Scatter Works, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scattered light is a limiting source of optical noise in many advanced optical systems, but it can also be a sensitive indicator of optical component quality. Consider the simple case of a telescope successfully used to image a dim star against a dark background; however, if light from a bright source (such as the moon located well out of the field of view) enters the telescope, it will scatter from the interior walls and the imaging optics themselves. Some of this light eventually reaches the detector and creates a dim background haze that washes out the image of the distant star. A good telescope design accounts for these effects by limiting potential scatter propagation paths and by requiring that critical elements in the optical system meet scatter specifications. This means doing a careful system analysis and a means to quantify the scattering properties of the telescope components. This article discusses modern techniques for quantifying, measuring, and analyzing scattered light, and reviews their development.

Like many scientific advances moving scatterometry from an art to a reliable metrology was done in a series of small hops (not always in the same direction), rather than a single leap. It started in 1961 when a paper by Hal Bennett and Jim Porteous reported measurements made by gathering most of the light scattered from front surface mirrors and normalizing this signal by the much larger specular reflection. They defined this ratio as the total integrated scatter (TIS), and using a scalar diffraction theory result drawn from the radar literature, related it to the surface root mean square (rms) roughness. By the mid-1970s, several angle-resolved scatterometers had been built as research tools in university, government, and industry labs. Unfortunately, instrument operation and data manipulation were generally poor, and meaningful comparison measurements were virtually impossible due to instrument differences, sample contamination, and confusion over what parameters should be compared. Analysis of scatter data, to characterize sample surface roughness, was the subject of many publications. A derivation of what is commonly called 'BRDF' (bidirectional reflectance distribution function) was published by Nicodemus and co-workers at the National Bureau of Standards (now the National Institute of Science and Technology or NIST) in 1970, but did not gain common acceptance as a way to quantify scatter measurements until the late 1980s when the advent of small computers, combined with inspection requirements for defense-related optics, dramatically stimulated the development of scatter metrology. Commercial laboratory instrumentation became available that could measure and analyze as many as 50 to 100 samples a day, and the number (and sophistication) of measurement facilities increased dramatically. The first ASTM Standards were published (TIS in 1987 and BRDF in 1991), but it was still several years before most publications correctly used these quantifying terms. Government defense funding decreased dramatically in the early 1990s, following the end of the Cold War, but the economic advantages of using scatter metrology and analysis for space applications and in the rapidly advancing semiconductor industry, continued state-of-the-art advancements.

The following sections detail how scatter is quantified when related to area (roughness) and local (pit/particle) generating sources. Instrumentation and the use of scattering models are also briefly reviewed.

Quantifying Scattered Light

Scatter signals can be easily quantified as scattered light power per unit solid angle (in watts per steradian); however, in order to make the results more meaningful, these signals are usually normalized, in some fashion, by the light incident on the scatter source. The three ways commonly employed to do this are defined below.

If the scattering feature in question is uniformly distributed across the illuminated spot on the sample (such as surface roughness), then it makes sense to normalize the collected scattered power in watts/steradian by the incident power. This simple ratio, which has units of inverse steradians, was commonly referred to as 'the scattering function.' Although this term is occasionally still found in the literature, it has been generally replaced by the closely related BRDF, which is defined by the differential ratio of the sample radiance normalized by its irradiance. After some simplifying assumptions are made, this reduces to the original scattering function with a cosine of the polar scattering angle in the denominator. The BRDF, defined in this manner, has become the standard way to report angle-resolved scatter from features that uniformly fill the illuminated spot. The cosine term results from the fact that NIST used radiometric terms to define BRDF:

$$\mathrm{BRDF} = \frac{\mathrm{Ps}/\Omega}{\mathrm{Pi}\cos\theta_s} \qquad [1]$$

The scatter function is often referred to as the 'cosine corrected BRDF' and is simply equal to the BRDF multiplied by the cosine of the polar scattering angle. Figure 1 gives the geometry for the situation, and defines the polar and azimuthal angles (θ_s and ϕ_s), as well as the solid collection angle (Ω). Other common abbreviations are BSDF, for the more generic bidirectional scatter distribution function, and BTDF for quantifying transmissive scatter.

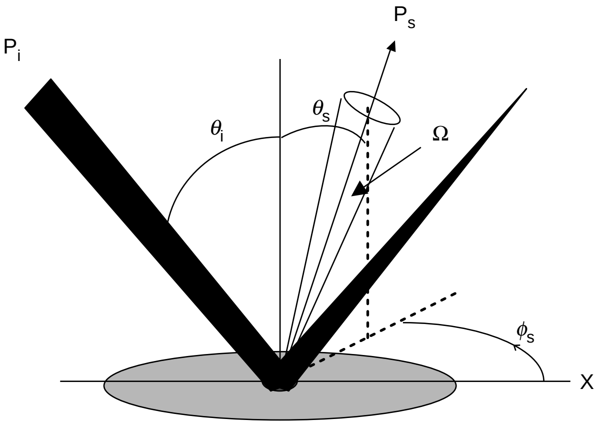

Figure 1 Scatter analysis uses standard spherical coordinates to define terms.

Integration of the scatter signal over much of the scattering hemisphere allows calculation of TIS, as the ratio of the scatter signal to the reflected specular power. This integration is usually carried out experimentally in such a way that both the incident beam and reflected specular beam are excluded. In the most common TIS situation, the beam is incident at a small angle near surface normal, and the integration is done from small values of θ_s to almost 90 degrees. If the fraction of light scattered from the specular reflection is small and if the scatter is caused by surface roughness, then it can be related to the rms surface roughness of the reflecting surface. As a ratio of powers, the TIS is a dimensionless quantity. The normalization is by Pr (instead of Pi) because reductions in scatter caused by low reflectance do not influence the roughness calculation. The pertinent relationships are given below, where σ is the rms roughness and λ is the light wavelength:

$$\mathrm{TIS} = \mathrm{Ps}/\mathrm{Pr} \cong (4\pi\sigma/\lambda)^2 \qquad [2]$$

Of course, all scatter measurements are integrations over a detector collection aperture, but the TIS designation is reserved for situations where the aim is to gather as much scattered light as possible, while ‘angle resolved’ designs are created to gain information from the distribution of the scattered light. Notice that TIS values become very large when measured from a diffuse surface, where the specular reflection is very small. Although TIS can be measured for any surface, the diffuse reflectance (equal to Ps/Pi) would often be more appropriate for diffuse surfaces. The various restrictions associated with relating TIS to rms roughness are detailed below.

Scatter from discrete features, such as particles and pits, which do not completely fill the illuminated spot, must be treated differently. This is because changes in spot size, with no corresponding change in total incident power, will change the incident intensity (watts/unit area) at the feature and thus also change the scatter signal (and BRDF) without any corresponding changes in the scattering feature. Clearly this is unacceptable if the object is to characterize the defect with scatter measurements. The solution is to define another quantification term, known as the differential scattering cross-section (DSC), where the normalization is the incident intensity at the feature (the units for DSC are area/steradian). Because this was not done in terms of radiometric units at the time it was defined, the cosine of the polar scattering angle is not in the definition. The same geometrical definitions, found in Figure 1, also apply for the DSC:

$$\mathrm{DSC} = \frac{\mathrm{Ps}/\Omega}{Ii} \qquad [3]$$

If the DSC is integrated over the solid angle associated with a collection aperture then the value has units of area. Because relatively small area focused laser beams are often used as a source, area is most commonly given in micrometers squared.

These three scatter parameters, the BRDF, the TIS, and the DSC, are obviously functions of system variables such as geometry, scatter direction (both in and out of the incident plane), incident wavelength and polarization, as well as feature characteristics. It is the dependence of the scatter signal on these system parameters that makes the scatter models useful for optimizing instrument designs. It is their dependence on feature characteristics that makes scatter measurement a useful metrology tool.

A key point needs to be stressed. When applied appropriately, TIS, BRDF, and DSC are absolute terms, not relative terms. The DSC of a 100 nm PSL in a given direction for a given source is a fixed value, which can be repeatedly measured and even accurately calculated from models. The same is true for TIS and BRDF values associated with surface roughness of known statistics. Scatter measuring instruments, such as particle scanners or lab scatterometers, can be calibrated in terms of these quantities. As has already been pointed out, the user of a scanner will almost always be more interested in characterizing defects than in the resulting scatter values, but the underlying instrument calibration can always be expressed in terms of these three quantities. This is true even though designers and users may find it convenient to use other metrics (such as polysterene latex (PSL) spheres) as a way to relate calibration.

Angle Resolved Scatterometers

The diagram in Figure 2 shows the most common scatterometer configuration. The source is fixed and the sample is rotated to the desired incident angle. The receiver is then rotated about the sample during scatter measurement. Most commonly, scatterometers operate just in the plane of incidence; however, instruments capable of measuring at virtually any location in either the reflective or transmissive hemispheres have been built. Although dozens of instruments have been built following this general design, other configurations are in use. For example, the source and receiver may be fixed and the sample rotated so that the scatter pattern moves past the receiver. This is easier mechanically than moving the receiver at the end of an arm, but complicates analysis because the incident angle and the observation angle change simultaneously. Another combination is to fix the source and sample together, at constant incident angle, and rotate this unit (about the point of illumination on the sample) so that the scatter pattern moves past a fixed receiver. This has the advantage that a long receiver/sample distance can be used without motorizing a long (heavy) receiver arm. It has the disadvantage that heavy (or multiple) sources are difficult to deal with. Other configurations, with everything fixed, have been designed that employ several receivers to merely sample the BSDF and display a curve fit of the resulting data. This is an economical solution if the BSDF is relatively uniform without isolated diffraction peaks. The goniometer section of a real instrument, similar to that of Figure 2, is shown in Figure 3.

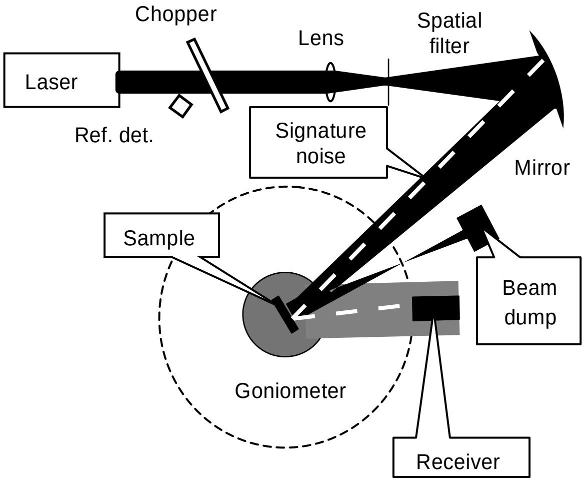

Figure 2 Basic elements of an incident plane scatterometer are shown.

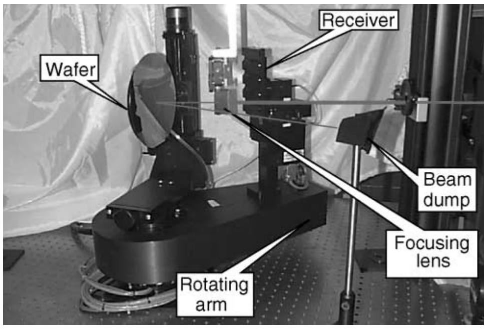

Figure 3 The author's scatterometer, which is similar to the diagram of Figure 2 is shown. In this case a final focusing lens is introduced to produce a very small illuminated spot on the silicon wafer sample. The white background was introduced to make the instrument easier to view.

Computer control of the measurement is essential to maximize versatility and minimize measurement time. The software required to control the measurement plus the display and analysis of the data can be expected to be a significant portion of total instrument development cost. The following reviews typical design features (and issues) associated with the source, sample mount and receiver components.

The source in Figure 2 is formed by a laser beam that is chopped, spatially filtered, expanded, and finally brought to a focus on the receiver path. The beam is chopped to reduce both optical and electronic noise. This is usually accomplished through the use of lock-in detection in the electronics package which suppresses all signals except those at the chopping frequency. Low noise, programmable gain electronics are essential to reducing system noise. The reference detector is used to allow the computer to ratio out laser power fluctuations and, in some cases, to provide the necessary timing signal to the lock-in electronics. Polarizers, wave plates, and neutral density filters are also commonly placed prior to the spatial filter. The spatial filter removes source scatter from the laser beam and presents a point source which is imaged by the final focusing element, in this case a mirror, to the detector zero position. Focusing the beam at this location allows near specular scatter to be more easily measured. Lasers are convenient sources, but are not necessary. Broadband sources are often required to meet a particular application or to simulate the environment where a sample will be used. Monochromators and filters can be used to provide scatterometer sources of arbitrary wavelength. The noise floors with these tunable incoherent sources increases as the spectral bandpass is narrowed, but they have the advantage that the scatter pattern does not contain laser speckle.

The sample mount can be very simple or very complex. In principal, six degrees of mechanical freedom are required to fully adjust the sample. The order in which these stages are mounted affects the ease of use (and cost) of the sample holder. In practice, it often proves convenient to either eliminate, or occasionally duplicate, some of these degrees of freedom. In addition, some of these axes may be motorized to allow the sample area to be raster-scanned to automate sample alignment or to measure reference samples. As a general rule, the scatter pattern is insensitive to small changes in incident angle but very sensitive to small angular deviations from specular. Instrumentation should be configured to allow location of the specular reflection (or transmission) very accurately. Receiver designs vary, but changeable entrance apertures, bandpass filters, lenses, and field stops are generally positioned in front of the detector.

A serious measurement problem is getting light scattered by the instrument, called instrument signature, confused with light scattered by the sample. An example of instrument signature is shown by the dotted line in Figure 2, which represents scatter from the final mirror. The signature is often measured in the straight through (transmission) direction, multiplied by the measured specular reflectance and then compared to the measured sample BRDF. Another issue is the fact that the measured BRDF is really the convolution of the receiver aperture with the actual (incremental) BRDF. When the scatter signal varies slowly across the aperture the measurement is virtually identical to the true BRDF. Near the specular reflection, or at diffraction peaks, the differences between the measurement (or convolution) and the actual (incremental) BRDF can be huge. Measurements made using invisible sources and measurements of curved samples present additional problems. These problems and the issues of calibration and accuracy are covered in the Further Reading section at the end of this article.

TIS Instruments

The two common methods of making TIS measurements are shown in Figures 4 and 5. The first one is based on a hemispherical mirror (or Coblentz sphere) to gather scattered light from the sample and image it onto the scatter detector. The specular beam enters and leaves the hemisphere through a small circular hole. The diameter of that hole defines the near specular limit of the instrument. The reflected beam (not the incident beam) should be centered in the hole because the BRDF will be symmetrical about it. Alignment of the hemispherical mirror is critical, and not trivial, in this approach. The second approach involves the use of an integrating sphere. A section of the sphere is viewed by a recessed detector. If the detector field of view (FOV) is limited to a section of the sphere that is not directly illuminated by scatter from the sample, then the signal will be proportional to total scatter from the sample. Again, the reflected beam should be centered on the exit hole. The Coblentz sphere method presents more signal to the detector; however, some of this signal is incident on the detector at very high angles. Thus, this approach tends to discriminate against high-angle scatter (which is not a problem for many samples). The integrating sphere is easier to align, but has a lower signal to noise ratio (less signal on the detector) and is more difficult to build in the IR where uniform diffuse surfaces are harder to obtain. A common mistake with TIS measurements is to assume that for near normal incidence, the orientation between source polarization and sample orientation is not an issue. TIS measurements made with a linearly polarized source on a grating at different orientations will quickly demonstrate this dependence.

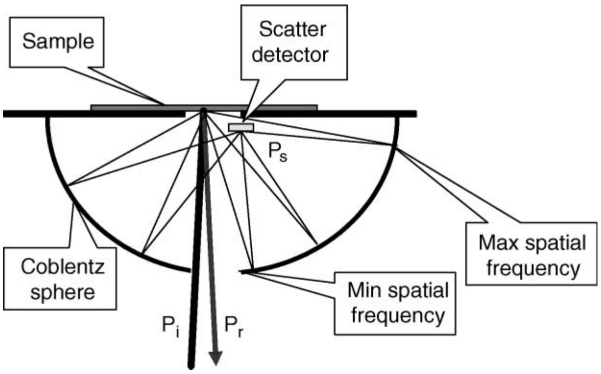

Figure 4 A diagram showing the Coblentz sphere approach to TIS measurements used in the early development of scatter instrumentation.

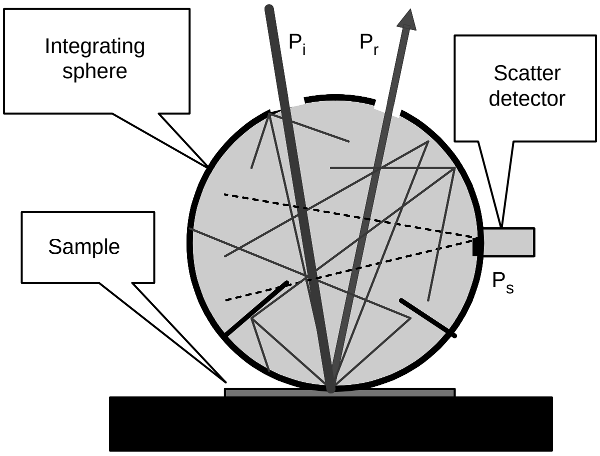

Figure 5 More modern TIS instruments make use of an integrating sphere approach which is easier to align and does not suffer from problems associated with measuring high-angle scatter from the sample.

TIS measurements can be made very near the specular reflection by utilizing a diffusely reflecting plate with a small hole in it. A converging beam is reflected off the sample and through the hole. Scatter is diffusely reflected from the diffuse plate to a receiver designed to uniformly view the plate. The reflected power is measured by moving the plate so the specular beam misses the hole and then taking that measurement. The ratio of the two scatter measurements gives the TIS. Measurements starting closer than 0.1 degrees from specular can be made in this manner and it is an excellent way to check incoming optics or freshly coated optics for low scatter.

Analyzing Scatter from Surface Roughness

The preceding sections have concentrated on obtaining and quantifying accurate scatter data, but that leaves the question of what to do with it once you have it. In rare situations you may be given a scatter (BRDF) specification – such as, the BRDF from the mirror must be less than 10^{-4} sr^{-1} 10 degrees from specular when measured at a wavelength of 633 nm incident at 5 degrees with an S polarized source. Unfortunately this is very uncommon. If the issue is limiting scatter as a noise source, you will probably have to generate your own specification based on specific system requirements. More difficult, and often of more economic value, is the situation where scatter measurements are being used as a metrology to learn something about the sample characteristics – like roughness or defect size and/or type.

The relationship between the measured BRDF and reflector roughness statistics was a subject of intense interest from the mid-1970s through the early 1990s. Dozens of papers, and even some books, have been written on the subject, and it can only be outlined here. The relatively easy case of scatter from roughness on a clean, optically smooth, front surface reflector was first published in 1975; however it was several years later before confirming experiments were completed. The deceptively simple relationship, based on vector perturbation theory, is shown below.

$$\text{BRDF} = (16\pi^2/\lambda^4) \cos\theta_\text{i} \cos\theta_\text{s}\, QS(f_x, f_y) \qquad [4]$$

Q is the polarization factor and is determined by the material constants of the reflector, as well as the system geometry. In many cases, it is numerically about equal to the specular reflectance and this approximation is often justified. Exact expressions are available in the literature. $S(f_x, f_y)$ is the surface power spectral density function (or PSD). It may be thought of as roughness power (surface height variations squared) per unit spatial frequency (undulations per unit distance instead of per unit time). Integration of the PSD over spatial frequency space results in the mean square roughness over that band of frequencies. Taking the square root gives the root mean square roughness (or rms). Frequencies in both the x and y directions on the surface are involved and they are defined by the well-known grating equations as:

$$f_x = (\sin\theta_\text{s} \cos\phi_\text{s} - \sin\theta_\text{i})/\lambda \qquad \text{and}$$

$$f_y = (\sin\theta_\text{s} \sin\phi_\text{s})/\lambda \qquad [5]$$

Thus eqn [4] becomes a model for surface roughness that allows BRDF measurement to be used to find and/or verify surface roughness specifications. Of course, there are exceptions. If the 'clean, optically smooth, front surface reflector' limitations are violated, then the surface will have more than just roughness as a source of scatter and the PSD found from eqn [4] will be too large. Obtaining the same PSD from BRDF measurements made at different wavelengths, or different polarizations, is an indication that the surface is scattering 'topographically' and the PSD can be found using this technique. Because some scatter measurements can be made very rapidly, there are industry situations where scatter metrology offers a very fast means of monitoring surface quality.

A little study of eqns [4] and [5] makes it clear that ranges of spatial frequencies in the PSD correspond directly to angular ranges of the BRDF. Scatter from a single spatial frequency corresponds to a single scatter direction.

Following the above discussion, it becomes clear why the pioneering integrated (TIS) scatter measurements could be used to produce surface rms values as indicated by eqn [2]. Unfortunately, the result of eqn [4] was not available in 1961. Instead eqn [2] was derived for the special case of a smooth surface with Gaussian statistics. Nobody was thinking about spatial bandwidths and angular limits. When striking differences were found between roughness measurements made by TIS and profilometer, the Gaussian assumption became the 'whipping boy,' and TIS scatter measurements took an underserved hit. In fact, integration of eqn [4] results in the TIS result given in eqn [2] under the generally true small angle assumption that most of the scatter is close to the specular reflection. The differences make sense when the concept of spatial frequency bandwidths (or appropriate angular limits) is introduced. It becomes clear that different wavelengths, incident angles and scatter collection angles will also generate different rms values (for the same surface) and why so much confusion resulted following the definition of TIS in terms of rms roughness. There is no such thing as a unique rms roughness for a surface, anymore than there is a single spatial bandwidth for the PSD, or a single set of angles over which to integrate the BRDF. TIS measurements and rms measurements should always be given with enough information to determine bandwidth limits.

Measuring and Analyzing Scatter from Isolated Surface Features

Understanding scatter from discrete surface features has led to big changes in the entertainment business (CDs, DVDs, digitized music, and films, etc.) as well as providing an important source of metrology for the semiconductor industry as they develop smaller faster chips for a variety of modern uses. Thus, just about everybody in the modern world utilizes our understanding of scatter from discrete surface features.

On the metrology side, the roughness signals described in the last section are a serious source of background noise that limits the size of the smallest defects that can be found. Discrete surface features come in a dazzling array of types, sizes, materials, and shapes – and they all scatter differently. A 100 nm silicon particle scatters a lot differently than a 100 nm silicon oxide particle (even if they have the same shape), and a 100 nm diameter surface pit will have a different scatter pattern. Models describing scatter from a variety of discrete surface features have been developed. Although many of these are kept confidential for competitive reasons, NIST offers some models publicly through a web site.

Confirming a model requires knowing exactly what is scattering the light. In order to accomplish this, depositions of PSLs of known size are made on the surface. Scatter is measured from one or more spheres. A second measurement of background scatter is then subtracted and the net BRDF is converted to DSC units using the known (measured) illuminated spot size. Measurements of this type have been used to confirm discrete feature scatter models. The model is then used to calculate scatter from different diameters and materials. Combined with the model for surface roughness to evaluate system noise, this capability allows signal to noise evaluation of different defect scanners designs.

Practical Industrial Instrumentation

Surface defects and particles, smaller than 100 nm are now routinely found on silicon wafers using scatter instruments known as particle scanners. Thousands of these instruments (with price tags approaching a million dollars each depending on type and use) are in daily use. These instruments are truly amazing; they inspect 200 mm wafers at a rate of one every thirty seconds reporting feature location, approximate size, and in some cases even type (pit or particle) by analyzing scatter signals that last only about 100 nanoseconds. Similar systems are now starting to be used in the computer disk industry and flat panel display inspection of surface features will follow.

Scanners are calibrated with PSLs of different sizes. Because it is impossible to identify all defect types and because diameter has little meaning for irregularly shaped objects, feature 'size' is reported in 'PSL equivalent diameters.' This leads to confusing situations for multiple detector systems, which (without some software help) would report different sizes for real defects that scatter much differently than PSLs. These difficulties have caused industry confusion similar to the 'Gaussian statistics/bandwidth limited' issues encountered before roughness scatter was understood. The publication of international standards relating to scanner calibration has reduced the level of confusion.

Scatter Related Standards

Early scatter related standards for BRDF, TIS, and PSD calculations were written for ASTM (the American Society for Testing Materials), but as the industrial need for these documents moved to the semiconductor industry, there was pressure to move the documents to SEMI (Semiconductor Equipment and Materials Inc.), which is an international organization. By 2004, the process of rewriting the ASTM documents in SEMI format was well underway in the Silicon Wafer Committee. Topics covered include: surface defect specification (M35), defect capture rate (M50), scanner specifications (M52), scanner calibration (M53), particle deposition testing (M58), BRDF measurement (ME1392), and PSD calculation (MF1811). This body of literature is probably the only place where all aspects of these related problems are brought together in one place.

Conclusion

The bottom line is that scatter measurement and analysis has moved from an art to a working metrology. Dozens of labs around the world can now take the same sample and get about the same measured BRDF from it. Thousands of industrial surface scanners employing scattered light signals are in use every day on every continent. In virtually every house in the modern world there is at least one entertainment device that depends on scatter signals. In short – scatter works.

See also

Scattering: Scattering Theory.

Further Reading

Bennett JM and Mattsson L (1995) *Introduction to Surface Roughness and Scattering*, 2nd edn. Washington, DC: Optical Society of America.

Bennett HE and Porteus JO (1961) Relation between surface roughness and specular reflectance at normal incidence. *Journal of the Optical Society of America* 51: 123.

Cady FM, Bjork DR, Rifkin J and Stover JC (1989) BRDF error analysis. *Proceedings of SPIE* 1165: 154–164.

Church EL and Zavada JM (1975) Residual surface roughness of diamond-turned optics. *Applied Optics* 14: 1788.

Doicu A, Eremin Y and Wriedt T (2000) *Acoustic & Electromagnetic Scattering Analysis Using Discrete Sources*. San Diego, CA: Academic Press.

Gu ZH, Dummer RS, Maradudin AA and McGurn AR (1989) Experimental study of the opposition effect in the scattering of light from a randomly rough metal surface. *Applied Optics* 28(3): 537.

Nicodemus FE, Richmond JC, Hsia JJ, Ginsberg IW and Limperis T (1977) *Geometric Considerations and Nomenclature for Reflectance*. NBS Monograph 160, US Dept. of Commerce.

Scheer CA, Stover JC and Ivakhnenko VI (1998) Comparison of models and measurements of scatter from surface-bound particles. *SPIE Proceedings* 3275.

Schiff TF, Stover JC and Bjork DR (1992) Mueller matrix measurements of scattered light. *Proceedings of SPIE* 1753: 269–277.

Stover JC (1975) Roughness characterization of smooth machined surfaces by light scattering. *Applied Optics* 14(8): 1796.

Stover JC (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE.

Stover JC (2001) Calibration of particle detection systems. In: Alain Diebold (ed.) *Handbook of Silicon Semiconductor Metrology*, chap. 21. New York: Marcel Dekker.

Wolfe WL and Bartell FO (1982) Description and limitations of an automated scatterometer. *Proceedings of SPIE* 362-30.

Spectrometers

K A More, SenSyTech, Inc. Imaging Group, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sp7ectrometers were developed after the discovery that glass prisms disperse light. Later, it was discovered that diffraction from multiple, equally spaced, wires or fibers also dispersed light. About 100 years ago, Huygens proposed his wave theory of light and Fraunhofer developed diffraction theory, which allowed scientific development of diffraction gratings.

These discoveries then led to the development of spectrometers. Light theory was sufficiently developed such that spectrometer designs, developed over 100 years ago, are still being used today. These theories and designs are briefly described along with comments on how current technology has improved upon these designs. This is followed by some examples of imaging spectrometers which have wide spectral coverage from 450 nm to 14 μm and produce images of more than 256 spectral bands.

The basic elements of a spectroscopic instrument are shown in Figure 1. The source, or more usually an image of the source, fills an entrance slit and the radiation is collimated by either a lens or mirror. The radiation is then dispersed, by either a prism or a grating, so that the direction of propagation of the radiation depends upon its wavelength. It is then brought to a focus by a second lens or mirror and the spectrum consists of a series of monochromatic images of the entrance slit. The focused radiation is detected, either by an image detector such as a photographic plate, or by a flux detector such as a photomultiplier, in which case the area over which the flux is detected is limited by an exit slit. In some cases the radiation is not detected at this stage, but passes through the exit slit to be used in some other optical system. As the exit slit behaves as a monochromatic source, the instrument can be regarded as a wavelength filter and is then referred to as a monochromator.

Prisms

The wavelength dependence of the index of refraction is used in prism spectrometers. Such an optical

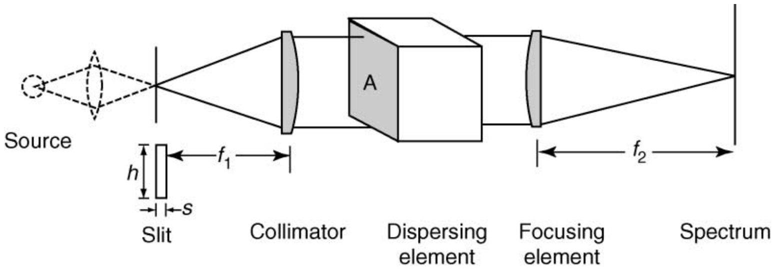

Figure 1 The basic elements of a spectroscopic instrument. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

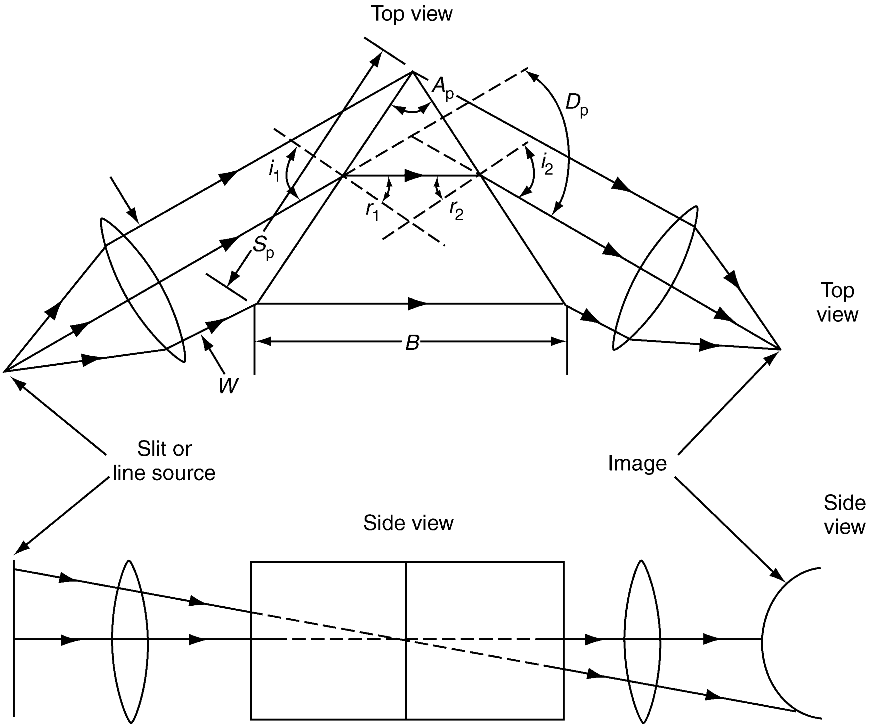

Figure 2 Elementary prism spectrometer schematic. *W* is the width of the entrance beam; S_p is the length of the prism face; and *B* is the prism base length. Reproduced with permission from *The Infrared Handbook* (1985). Ann Arbor, MI: Infrared Information Analysis Center.

element disperses parallel rays or collimated radiation into different angles from the prism according to wavelength. Distortion of the image of the entrance slit is minimized by the use of planewave illumination. Even with planewave illumination, the image of the slit is curved because not all of the rays from the entrance slit can traverse the prism in its principal plane. The prism is shown in the position of minimum angular deviation of the incoming rays in Figure 2. At minimum angular deviation, maximum power can pass through the prism. For a prism adjusted to the position of minimum deviation:

$$r_1 = r_2 = A_p/2 \qquad [1]$$

and

$$i_1 = i_2 = (D_p + A_p)/2 \qquad [2]$$

where

D_p = angle of deviation
A_p = angle of prism
r_1 and r_2 = internal angles of refraction
i_1 and i_2 = angles of entry and exit.

The angle of deviation, D_p, varies with wavelength. The resulting angular dispersion is defined as $\mathrm{d}D_p/\mathrm{d}\lambda$, while the linear dispersion is $\mathrm{d}x/\mathrm{d}\lambda = F\,\mathrm{d}D_p/\mathrm{d}\lambda$, where F is the focal length of the camera or imaging lens and x is the distance across the image plane.

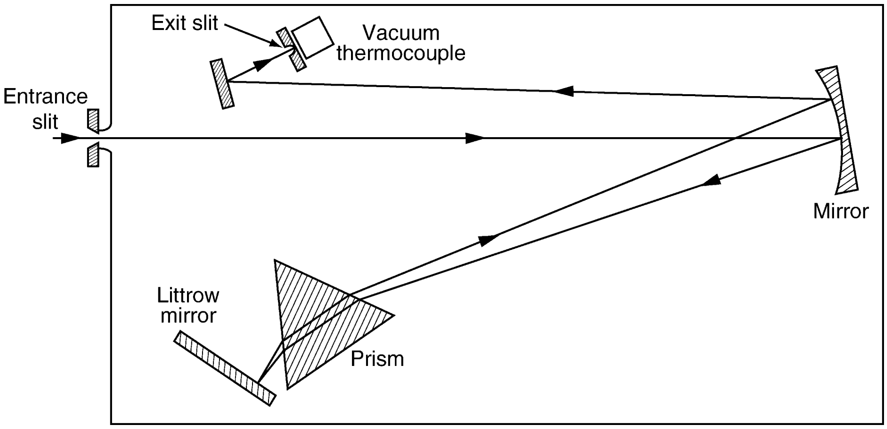

Figure 3 Infrared spectrograph of the Littrow-type mount with a rock salt prism. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

It can be shown that:

$$\mathrm{d}D_\mathrm{p}/\mathrm{d}\lambda = [B/W][\mathrm{d}n/\mathrm{d}\lambda] = [\mathrm{d}D_\mathrm{p}/\mathrm{d}n][\mathrm{d}n/\mathrm{d}\lambda] \qquad [3]$$

where

B = base length of the prism,
W = width of the illumination beam
n = index of refraction

while

$$\mathrm{d}x/\mathrm{d}\lambda = F[B/W][\mathrm{d}n/\mathrm{d}\lambda] \qquad [4]$$

One may define the resolving power, RP, of an instrument as the smallest resolvable wavelength difference, according to the Rayleigh criterion, divided into the average wavelength in that spectral region. Thus:

$$\mathrm{RP} = \lambda/\Delta\lambda = [\lambda/\mathrm{d}D_\mathrm{p}][\mathrm{d}D_\mathrm{p}/\mathrm{d}\lambda]$$
$$= [\lambda/\mathrm{d}D_\mathrm{p}][B/W][\mathrm{d}n/\mathrm{d}\lambda] \qquad [5]$$

The limiting resolution is set by diffraction due to the finite beamwidth, or effective aperture of the prism, which is rectangular. Thus:

$$\mathrm{RP} = [\lambda/(\lambda/W)][B/W][\mathrm{d}n/\mathrm{d}\lambda] \qquad [6]$$

so that:

$$RP = B[\mathrm{d}n/\mathrm{d}\lambda] \qquad [7]$$

If the entire prism face is not illuminated, then only the illuminated base length must be used for B.

Littrow showed that aberrations would be minimized by making the angle of incidence equal to the angle of refraction β (also known as the Littrow Configuration). Littrow used a plane mirror behind the prism for a double pass through the prism, as shown in Figure 3.

Gratings

Rowland is credited with the development of a grating mount that reduced aberrations in the spectrogram. He found that a grating ruled on a concave surface of radius R and locating the entrance and exit slits on the same radius would give the least aberrations, as shown in Figure 4. $r = R\cos\alpha$ is the distance to the entrance slit *(r)* and $r_1 = R\cos\beta$ is the distance to the focal point for the exit slit (r_1). Rowland showed that:

$$\cos\alpha/R - \cos^2\alpha/r + \cos\beta/R - \cos^2\beta/r_1 = 0 \qquad [8]$$

One solution to this equation is for α and β to each be zero and

$$r = R\cos\alpha$$

and

$$r_1 = R\cos\beta$$

This condition is met if r and r_1 lie on the Rowland circle.

There are various ways in which the Rowland circle condition may be satisfied and some of them are shown in Figure 5. The simplest mounting of all is that due to Paschen and Runge (Figure 5a) in which the entrance slit is positioned on the Rowland circle and a photographic plate (or plates) is constrained to fit the Rowland circle. Alternatively, for photoelectric detection, a series of exit slits is arranged around the Rowland circle each having its own detector. In the latter case, the whole spectrum is not recorded, only a series of predetermined wavelengths, but when used in this ‘polychromator’ form, it is very rugged and convenient for applications such as the routine analysis of samples of metals and alloys. In this case, slits and detectors are set up to measure the light in various spectral lines, each characteristic of a particular component or trace element.

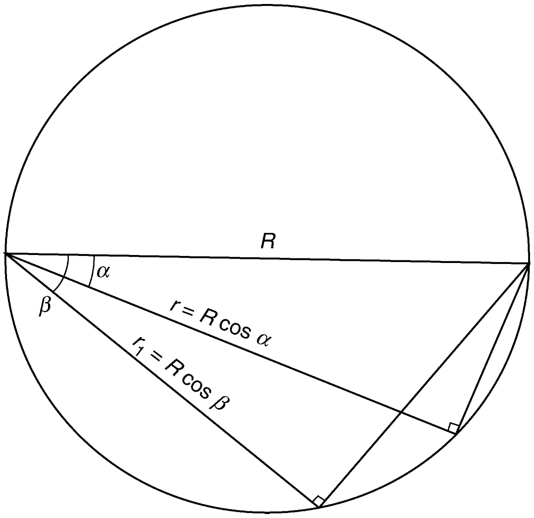

Figure 4 The construction of the Rowland Circle. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

The oldest concave grating mount is that designed by Rowland himself and which bears his name (Figure 5b). In this case, the grating and photographic plates are fixed at opposite ends of the diameter of the Rowland circle by a moveable rigid beam. The entrance slit remains fixed above the intersection of two rails at right angles to each other and along which the grating plate holder (or the exit slit) is free to move. In this way, this entrance slit, grating, and plate holder are constrained always to lie on the Rowland circle and it has the advantage that the dispersion is linear, which is useful in the accurate determination of wavelengths. Unfortunately, this mounting is rather sensitive to small errors in the position of the entrance slit and in the orthogonality of the rails, and is now very rarely used.

A variation on this mounting was devised by Abney, who again mounted the grating and the plate holder on a rigid bar at opposite ends of the diameter of the Rowland circle (Figure 5c). The entrance slit is mounted on a bar of length equal to the radius of the Rowland circle. In this way, the slit always lays on the Rowland circle, but it had to rotate about its axis in order that the jaws should remain perpendicular to the line from the grating to

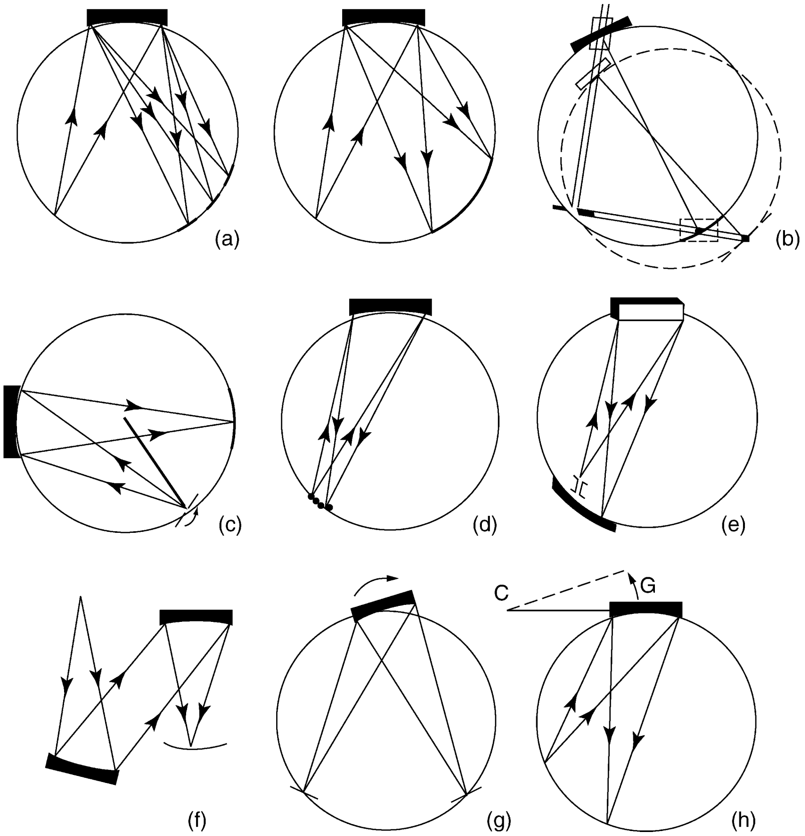

Figure 5 Mountings of the concave grating: (a) Paschen–Runge; (b) Rowland; (c) Abney; (d) and (e) Eagle; (f) Wadsworth; (g) Seya–Namioka; (h) Johnson–Onaka. With permission from Hutley MC (1982) *Diffraction Gratings*, pp. 57–232. London: Elsevier.

the slit. It also had the disadvantage that the source must move with the exit slit, which could be inconvenient.

In the Eagle mounting (Figures 5d and 5e), the angles of incidence and diffraction are made equal, or very nearly so, as with Littrow mounting for plane gratings. Optically, this system has the advantage that the astigmatism is generally less than that of the Paschen–Runge or Rowland mounting, but on the other hand, the dispersion is nonlinear. From a mechanical point of view, it has the disadvantage that it is necessary with great precision both to rotate the grating and to move it nearer to the slits in order to scan the spectrum. However, it does have the practical advantage that it is much more compact than other mountings, and this is of particular importance when we bear in mind the need to enclose the instrument in a vacuum tank. Ideally, the entrance slit and exit slit or photographic plate should be superimposed if we are to set $\alpha = \beta$. In practice, of course, the two are displaced either sideways, in the plane of incidence as shown in Figure 5, or out of the plane of incidence, in which case the entrance slit is positioned below the meridonal plane and the plate holder just above it, as shown in Figure 5. The out-of-plane configuration is generally used for spectrographs and the in-plane system for monochromators. The penalty incurred in going out of plane is that coma is introduced in the image, and slit curvature becomes more important. This limits the length of the ruling that can effectively be used.

The second well-known solution to the Rowland equation is the Wadsworth mounting (Figure 5f), in which the incident light is collimated, so r is set at infinity and the focal equation reduces to $r_1 = R\cos^2\beta/(\cos\alpha + \cos\beta)$. One feature of this mounting is that the astigmatism is zero when the image is formed at the center of the grating blank, i.e., when $\beta = 0$. It is a particularly useful mounting for applications in which the incident light is naturally collimated (for example, in rocket or satellite astronomy, spectroheliography and in work using synchrotron radiation). However, if the light is not naturally collimated, the Wadsworth mount requires a collimating mirror, so one has to pay the penalty of the extra losses of light at this mirror. The distance from the grating to the image is about half that for a Rowland circle mounting which makes the instrument more compact and, since the grating subtends approximately four times the solid angle, there is a corresponding increase in the brightness of the spectral image.

Not all concave grating mountings are solutions to the Rowland equation. In some cases, other advantages may compensate for a certain defect of focus. A particularly important example of this is the Seya–Namioka mounting (Figure 5g), in which the entrance slit and exit slit are kept fixed and the spectrum is scanned by a simple rotation of the grating (Figure 5g). In order to achieve the optimum conditions for this mounting, we set $\alpha = \varphi + \theta$ and $\beta = \theta - \varphi$, where 2φ is the angle subtended at the grating by the entrance and exit slit and θ is the angle through which the grating is turned. The amount of defocus is given by:

$$F(\theta, \varphi, r, r_1) = [\cos^2(\theta + \varphi)/r] + [\cos(\theta + \varphi)/R] + [\cos^2(\theta - \varphi)/r_1] + [\cos(\theta - \varphi)/R] \quad [9]$$

and the optimum conditions are those for which $F(\theta,\varphi,r,r_1)$ remains as small as possible as θ is varied over the required range. Seya set F and three derivatives of F with respect to θ equal to zero for $\theta = 0$ and obtained the result:

$$\varphi = \sin^{-1}(1/\sqrt{3}) = 35°15' \quad [10]$$

and

$$r = r_1 = R\cos\varphi \quad [11]$$

which corresponds to the Rowland circle either in zero order or for zero wavelength. In practice, it is usual to modify the angle slightly so that the best focus is achieved in the center of the range of interest rather than at zero wavelength.

The great advantage of the Seya–Namioka mounting is its simplicity. An instrument need consist only of a fixed entrance and exit slit, and a simple rotation of the grating is all that is required to scan the spectrum. It is, in fact, simpler than instruments using plane gratings. Despite the fact that at the ends of the useful wavelength range the resolution is limited by the defect of focus and the astigmatism is particularly bad, the Seya–Namioka mounting is very widely used, particularly for medium-resolution rather than high-resolution work.

A similar simplicity is a feature of the Johnson–Onaka mounting (Figure 5h). Here again, the entrance and exit slits remain fixed, but the grating is rotated about an axis which is displaced from its center; in this way, it is possible to reduce the change of focus that occurs in the Seya–Namioka mounting. The system is set up so that at the center of the desired wavelength range, the slits and grating lie on the Rowland circle, as shown in Figure 5h. The optimum radius of rotation, i.e., the distance GC, was found by

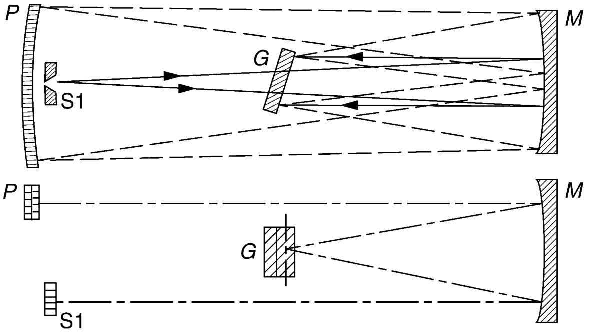

Figure 6 Ebert mounting of the plane grating designed by Fastie. 'Sl' is the entrance slit; *G* is the grating; *M* is the concave mirror; and *P* is the photographic plate. The horizontal section is at the top and the vertical section is at the bottom. Reproduced with permission from *The Infrared Handbook* (1985) Ann Arbor, MI: Infrared Information Analysis Center.

Onaka to be:

$$\mathrm{GC}_{\mathrm{opt}} = [R\sin\tfrac{1}{2}(\alpha+\beta)]/[1-\tfrac{1}{2}\tan(\alpha+\beta) \times (\tan\beta - \tan\alpha)] \qquad [12]$$

Another non-Rowland spectrometer is the Ebert–Fastie mount, which mounts the slit, flat grating, and a detector array in the arrangement shown in Figure 6. Ebert first developed the design using two separate concave mirrors, one for collimating the incident beam and the second to focus the diffracted spectrum. Fastie used a single but larger concave mirror, which simplified the mounting structure and produced a rugged, compact spectrometer that has been used in rocket flights and space satellite observatories for astronomical and upper atmospheric applications. The Czerny–Turner mount is similar to the Ebert mount, except the flat grating is located in the same plane that contains the entrance and exit slits.

Advanced Spectrometers

While the above spectrometer designs are still used, major advances in implementation are now available. Ray tracing allows the designer to quantify the aberrations and determine solutions to remove them. Aspheric optic elements can now be fabricated to correct aberrations. Gratings can be ruled on aspherical surfaces to not only disperse the light beam, but also be an element in the optical design. Holography has been developed to etch gratings for use over a wide spectral range and to not only disperse the light, but also work as an optical element. Because holographic gratings are chemically etched, there are no machine burrs to scatter light and the hologram is free of periodic differences in groove widths that create ghosts in ruled gratings. Linear and array photodetectors have replaced film. An advantage of film was its ability to fit curved focal planes. Photodetectors are etched into flat wafers of the photodiode material. To adapt flat arrays to a curved focal plane, fiber optic face plate couplers have been ground on one side to match the focal plane curvature and the flat back side is either optically coupled to the detector or closely coupled for proximity focusing. The photodetector arrays are available in multiple materials to cover the spectral range, from soft X-rays to the thermal infrared. Often cooling is required to obtain low noise. They are now available with sensitivities never reached with film and the ease of coupling the output data to a computer for real-time analysis has taken much of the labor out of analyzing the spectrograph. Another advance is precision motor drives with precision motion sensors, some using laser interferometry, to provide feedback on automated movement of gratings, slits, and detectors as a spectrometer cycles through its wavelength range.

Imaging Spectrometers

Spectrometers that image the spectral characteristics in each pixel of the image to form a data cube, as shown in Figure 7, are named imaging spectrometers, and if the spectra have high resolution and blocks of consecutive neighbors, then the data are called hyperspectral. Imaging spectrometers have several applications that range from medical to remote sensing of land use and environment. This article will cover examples for remote sensing imaging spectrometers which require spectral coverage in all of the atmospheric windows from the UV to thermal IR.

Imaging spectrometers require a combination of spectrometers, light collecting optics, and scan mechanisms to scan the instantaneous field of view of the spectrometer over a scene. Remote sensing of the Earth applications requires an aerial platform. Either a helicopter, an aircraft, or orbital space satellites are used. The platform motion is used as part of the scanning process so that the optics image a single point on the ground so that a scanner (called a line scanner) scans a long line that is cross tracked to the platform motion. Or, the optics image a slit that is parallel to the platform track that covers many scan lines and a scanner moves the slit cross track to the platform motion. This scanner is called a whiskbroom scanner. Or, the optics image a large slit so no scan mechanism is needed other than the platform

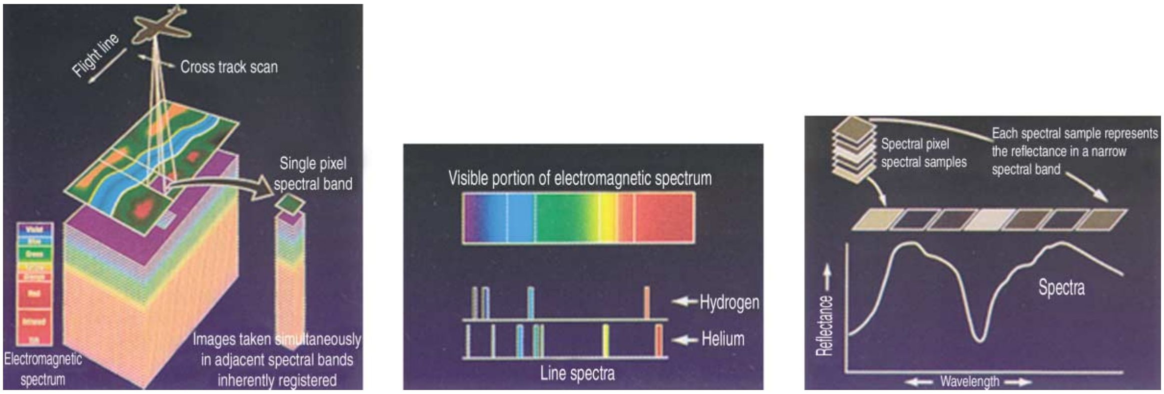
Flight line
Cross track scan
Single pixel spectral band
Electromagnetic spectrum
Images taken simultaneously in adjacent spectral bands inherently registered
Visible portion of electromagnetic spectrum
Hydrogen
Helium
Line spectra
Spectral pixel spectral samples
Each spectral sample represents the reflectance in a narrow spectral band
Spectra
Reflectance
Wavelength

Figure 7 Hyperspectral Data Cube. Hyperspectral imagers divide the spectrum into many discrete narrow channels. This fine quantization ofn spectral information on a pixel by pixel basis enables researchers to discriminate the individual constituents in an area much more effectively. For example, the broad spectral bands of a multispectral sensor allow the user only to coarsely discriminate between areas of deciduous and coniferous forest, plowed fields, etc., whereas a hyperspectral imager provides characteristic signatures which can be correlated with specific spectral templates to help determine the individual constituents and possibly even reveal details of the natural processes which are affecting them.

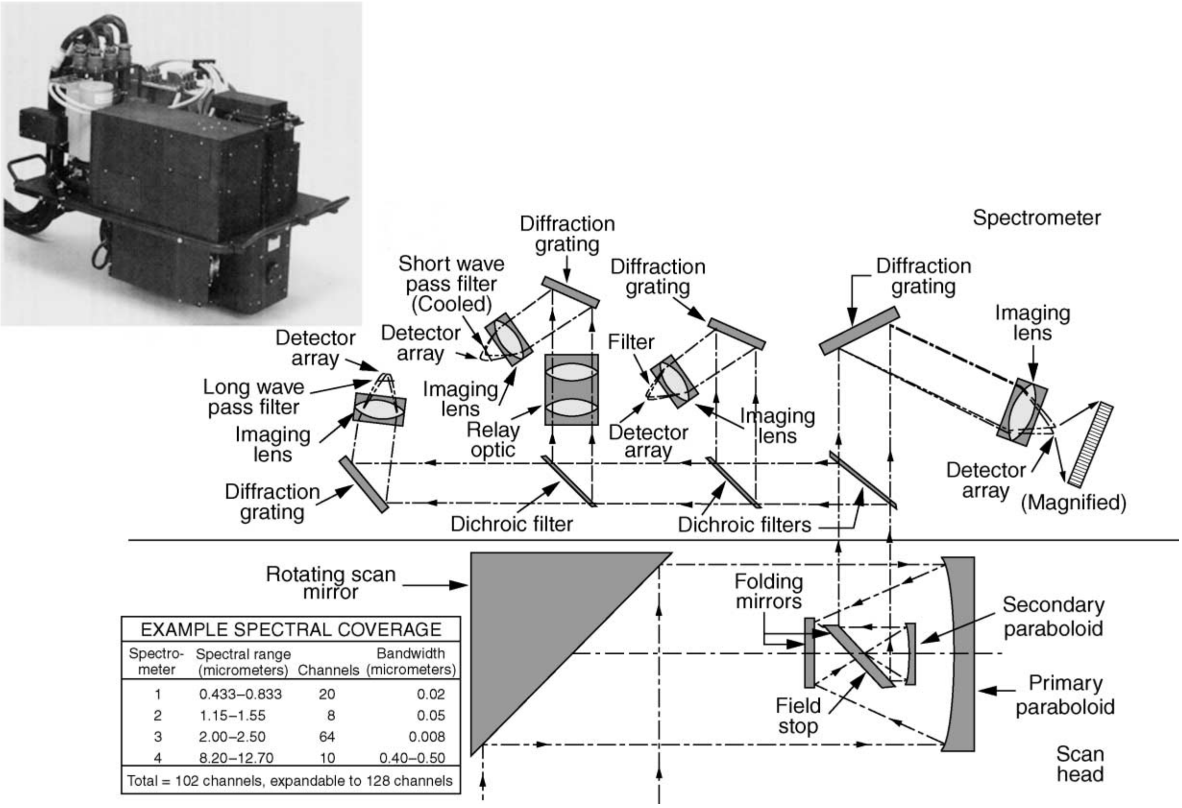
Spectrometer
Diffraction grating
Short wave pass filter (Cooled)
Diffraction grating
Diffraction grating
Detector array
Detector array
Filter
Imaging lens
Long wave pass filter
Imaging lens
Imaging lens
Relay optic
Detector array
Imaging lens
Detector array
(Magnified)
Diffraction grating
Dichroic filter
Dichroic filters
Rotating scan mirror
Folding mirrors
Secondary paraboloid
Primary paraboloid
Field stop
Scan head

EXAMPLE SPECTRAL COVERAGE

Spectro-meter	Spectral range (micrometers)	Channels	Bandwidth (micrometers)
1	0.433–0.833	20	0.02
2	1.15–1.55	8	0.05
3	2.00–2.50	64	0.008
4	8.20–12.70	10	0.40–0.50

Total = 102 channels, expandable to 128 channels

Figure 8 Multispectral infrared and visible imaging spectrometer optical schematic.

motion to form an image. This scanner is called a pushbroom scanner.

One important requirement is that all spectral measurements of a pixel be coregistered. Most airborne imaging spectrometers use a common aperture for a line scanner, or a slit for whisk and pushbroom scanners, so that platform instability from pitch roll and yaw and from inability to fly in a straight line do not compromise the coregistration of spectral data on each pixel. The image data may require geometric correction, but the spectral data are not compromised.

There is one other type of imaging spectrometer that uses a linear variable filter (also known as a wedge filter) over a 2D array of photodetectors. Each image contains a different spectral band over each row of pixels so that each frame images a scene with a different spectral band over each row of pixels. The array is oriented so the rows of spectral bands are perpendicular to the flight track. After the platform motion moves over one ground pixel, the whole array is read out and the frame is shifted one row of pixels so the second frame adds a second spectral band to each row imaged in the first frame. This frame stepping is carefully timed to the platform velocity and is repeated until each row of pixels is imaged in all spectral bands. This type of imaging spectrometer has been flown in aircraft and satellites. In aircraft, the platform motion corrupts the coregistration of the spectrum in each pixel. Extensive ground processing is required to geometrically correct each frame to improve spectral coregistration. This is a difficult task and this type of imaging spectrometer has lost favor for airborne use. Stabilized satellites have been a better platform and the wedge type of imaging spectrometer has been used successfully in space.

As examples of airborne and ground-based imaging spectrometers, a line scanner imaging

Table 1 MIVIS physical properties

	Height		*Width*		*Depth*[a]		*Weight*	
	in	*cm*	*in*	*cm*	*in*	*cm*	*lbs*	*kg*
Scan head	26.5	67.0	20.6	52.0	28.1	71.0	220	100
Electronics	40.0	102.0	19.0	48.3	24.0	61.0		
Total system weight (approx.)							460	209

[a] Not including connectors or cable bends.

Table 2 MIVIS spectral coverage (μm)

OPTICAL PORT 1			*OPTICAL PORT 3*						*OPTICAL PORT 4*		
CH #	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*		*CH #*	*Band Edges*	
	Lower	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*		*Lower*	*Upper*
1	0.43	0.45	29	2.000	2.008	61	2.250	2.258	93	8.20	8.60
2	0.45	0.47	30	2.008	2.016	62	2.258	2.266	94	8.60	9.00
3	0.47	0.49	31	2.016	2.023	63	2.266	2.273	95	9.00	9.40
4	0.49	0.51	32	2.023	2.031	64	2.273	2.281	96	9.40	9.80
5	0.51	0.53	33	2.031	2.039	65	2.281	2.289	97	9.80	10.20
6	0.53	0.55	34	2.039	2.047	66	2.289	2.297	98	10.20	10.70
7	0.55	0.57	35	2.047	2.055	67	2.297	2.305	99	10.70	11.20
8	0.57	0.59	36	2.055	2.063	68	2.305	2.313	100	11.20	11.70
9	0.59	0.61	37	2.063	2.070	69	2.313	2.320	101	11.70	12.20
10	0.61	0.63	38	2.070	2.078	70	2.320	2.328	102	12.20	12.70
11	0.63	0.65	39	2.078	2.086	71	2.328	2.336			
12	0.65	0.67	40	2.086	2.094	72	2.336	2.344			
13	0.67	0.69	41	2.094	2.102	73	2.344	2.352			
14	0.69	0.71	42	2.102	2.109	74	2.352	2.359			
15	0.71	0.73	43	2.109	2.117	75	2.359	2.367			
16	0.73	0.75	44	2.117	2.125	76	2.367	2.375			
17	0.75	0.77	45	2.125	2.133	77	2.375	2.383			
18	0.77	0.79	46	2.133	2.141	78	2.383	2.391			
19	0.79	0.81	47	2.141	2.148	79	2.391	2.398			
20	0.81	0.83	48	2.148	2.156	80	2.398	2.406			
OPTICAL PORT 2			49	2.156	2.164	81	2.406	2.414			
21	1.15	1.20	50	2.164	2.172	82	2.414	2.422			
22	1.20	1.25	51	2.172	2.180	83	2.422	2.430			
23	1.25	1.30	52	2.180	2.188	84	2.430	2.438			
24	1.30	1.35	53	2.188	2.195	85	2.438	2.445			
25	1.35	1.40	54	2.195	2.203	86	2.445	2.453			
26	1.40	1.45	55	2.203	2.211	87	2.453	2.461			
27	1.45	1.50	56	2.211	2.219	88	2.461	2.469			
28	1.50	1.55	57	2.219	2.227	89	2.469	2.477			
			58	2.227	2.234	90	2.477	2.484			
			59	2.234	2.242	91	2.484	2.492			
			60	2.242	2.250	92	2.492	2.500			

spectrometer and a ground-based scanned push broom scanner are described below.

MIVIS (multispectral infrared and visible imaging spectrometers) hyperspectral scanner developed by SenSyTech Imaging Group (formerly Daedalus Enterprises) for the CNR (Consiglio National Researche) of Italy is a line scanner imaging spectrometer. The optical schematic is shown in Figure 8. A rotating 45° mirror scans a line of pixels on the ground. Each pixel is collimated by a parabolic mirror in a Gregorian telescope mount. The collimated beam is reflected by a Pfund assembly to an optical bench that houses four spectrometers. An aperture pinhole in the Pfund assembly defines a common instantaneous field of view for each pixel. The collimated beam is then split off with either thin metallic coated or dielectric coated dichroics to four spectrometers. The thin metallic mirrors reflect long wavelengths and transmit short wavelengths. The multiple dielectric layers cause interference such that light is reflected at short wavelengths and transmitted at long wavelengths. After splitting off four wide bands (visible, near infrared, mid-wave infrared, and thermal infrared) each band is dispersed in its own spectrometers.

The wavelength coverage for each spectrometer is based on the wavelength sensitivity of different photodetector arrays. The visible spectrometer uses a silicon photodiode array; the near infrared spectrometer, an InGaAs array, the mid-infrared a InSb array and the thermal infrared, a MCT (mercury doped cadmium teluride) photo-conductor array. Note the beam expander in spectrometer 3 for the

Table 3 LAFS technical specifications

Spatial resolution	1.0 mrad IFOV vertical and horizontal (square pixels). Optical lens for 0.5 mrad IFOV
Spatial coverage	15° TFOV horizontal, nominal. A second, 7.5° TFOV, is an option
Focus range	5′ to infinity
Spectral resolution	5.0 nm per spectral channel, sampled at 2.5 nm interval
Spectral range	400–1100 nm
Dynamic range	8 bit (1 part in 256)
Illumination	Solar illumination ranging from 10:00 a.m. to 2:00 p.m. under overcast conditions to full noon sunshine
Acquisition time	1.5 sec for standard illuminations. Option for longer times for low light level conditions
Viewfinder	Near real time video view of the scene
Calibrations	Flat fielding to compensate for CCD variations in responsivity. Special calibration
Data displays	Single band imagery, selectable Waterfall Chart (one spatial line by spectral, as acquired by the CCD). Single pixel spectral display
Data format	Convertible to format compatible with image processors
Data storage	Replaceable hard disc, 14 data cubes/disk

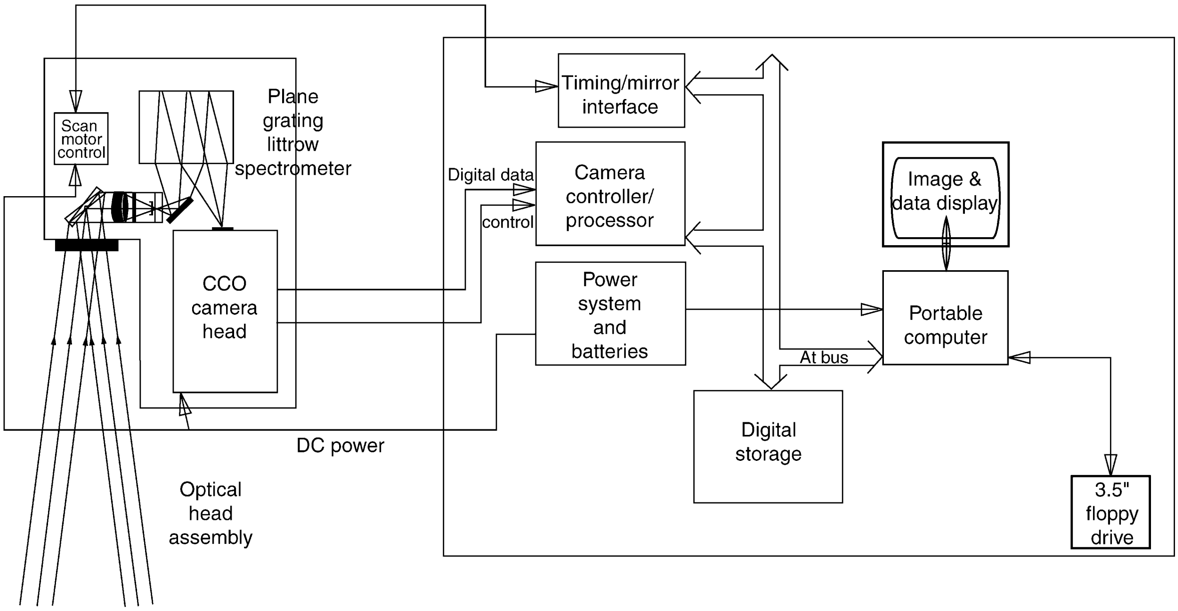

Figure 9 LAFS system block diagram.

mid-infrared (IR). Since wavelength resolution increases with the size of the illuminated area of a diffraction grating a beam expander was needed to meet the specified mid-IR resolution.

The physical characteristics of MIVIS are given in Table 1 and the spectral coverage in Table 2.

An example of a push broom imaging spectrometer is the large area fast spectrometer (LAFS), which is a field portable, tripod mounted, imaging spectrometer. LAFS uses a push broom spectrometer with a galvanometer driven mirror in front of the entrance slit. The mirror is stepped to acquire a 2D image × 256 spectral bands data cube.

LAFS was developed for the US Marine Corps under the direction of the US Navy Coastal Systems Station of the Dahlgren Division. The technical specifications are shown in Table 3. Various concepts for the imaging spectrometer were studied and the only concept that could meet the above specification was a Littrow configuration grating spectrometer that imaged one line in the scene and dispersed the spectra perpendicular to the line image onto a CCD array. A galvanometer mirror stepped the line image over the scene to generate a spectral image of the scene. A block diagram of LAFS is shown in Figure 9 and a photograph of the prototype in Figure 10. Figure 11 is a photograph of the optical head. Data from a single frame are stored in computer (RAM) memory then transferred to a replaceable hard disk for bulk storage. Data are collected in a series of 256 CCD frames and each CCD frame has a line of 256 pixels along one axis and 256 spectral samples along the second axis. The 256 CCD frames constitute a data cube as shown in Figure 12. This does not allow field viewing of an image in a single spectral band. Earth View™ software is used in the field portable computer

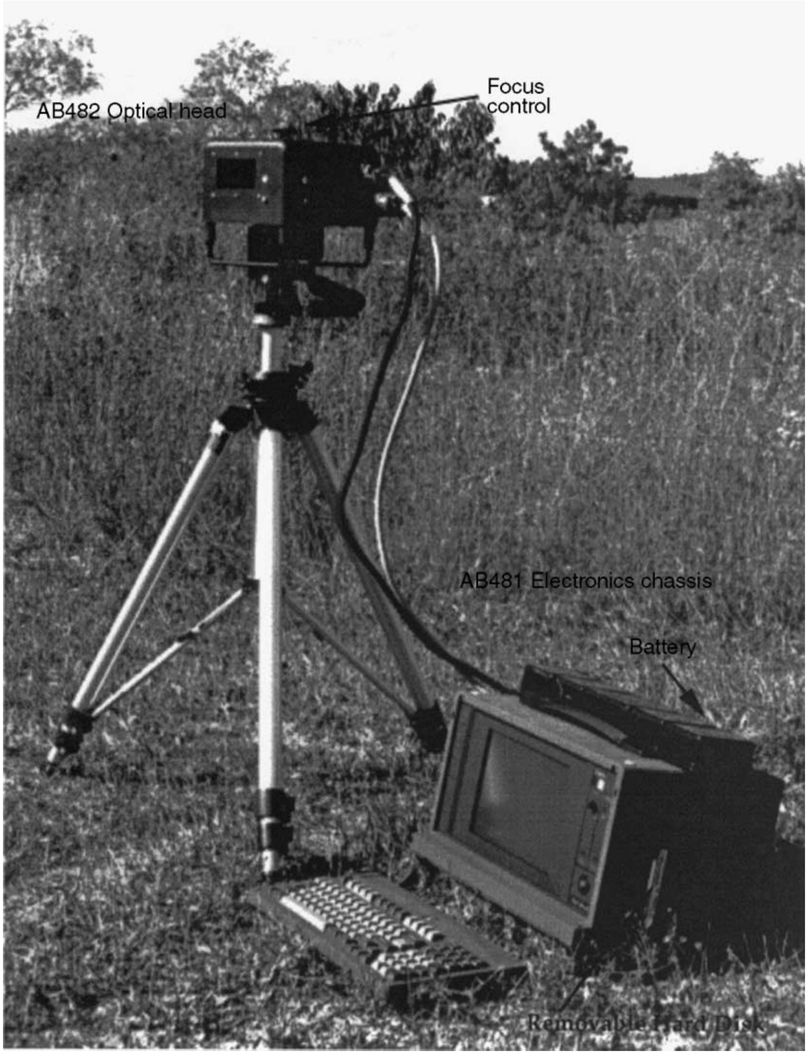

Figure 10 LAFS.

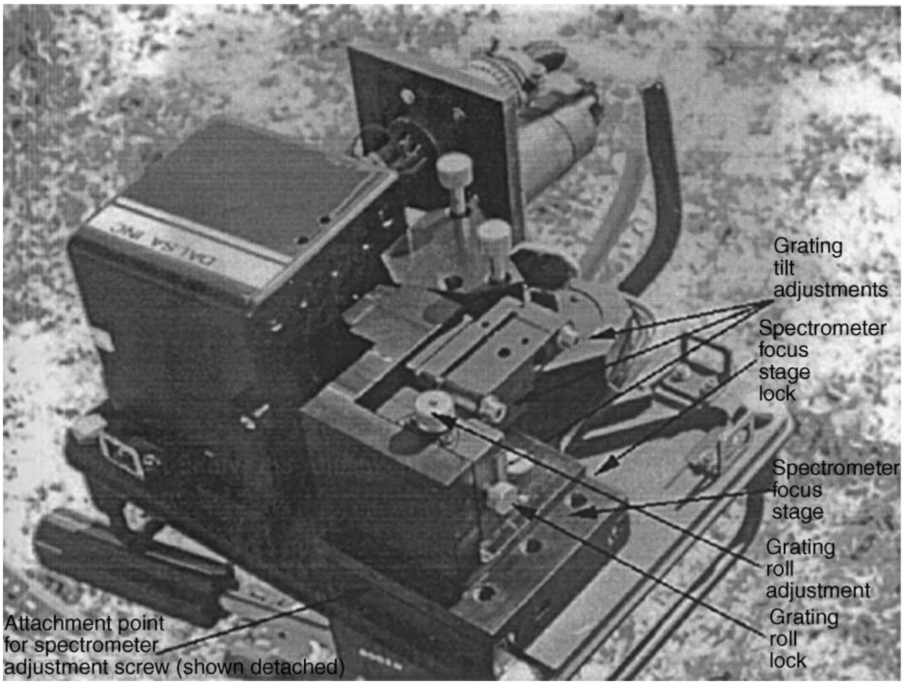
Grating
tilt
adjustments
Spectrometer
focus
stage
lock
Spectrometer
focus
stage
Grating
roll
adjustment
Grating
roll
lock
Attachment point
for spectrometer
adjustment screw (shown detached)
DALSA INC

Figure 11 LAFS optical head.

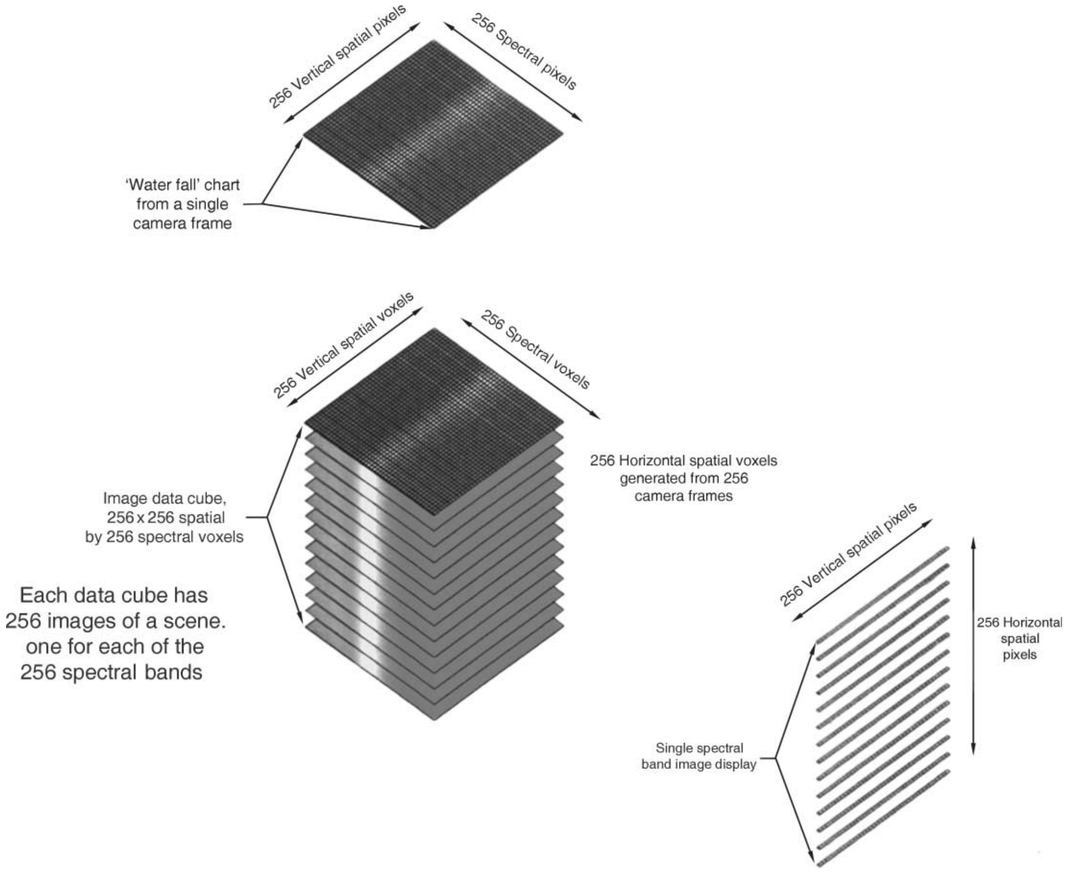
256 Vertical spatial pixels
256 Spectral pixels
'Water fall' chart
from a single
camera frame
256 Vertical spatial voxels
256 Spectral voxels
256 Horizontal spatial voxels
generated from 256
camera frames
Image data cube,
256 x 256 spatial
by 256 spectral voxels
Each data cube has
256 images of a scene.
one for each of the
256 spectral bands
256 Vertical spatial pixels
256 Horizontal
spatial
pixels
Single spectral
band image display

Figure 12 Arrangement of image memory and construction of image displays.

Table 4 LAFS physical properties

Dimensions and weight	*Size (inches)*	*weight (lbs)*
Optical head	7 × 9 × 8	11
Electronics	12 × 15.5 × 9	30
Battery	4.5 × 16.5 × 2	7
		48 + cables

Power: Less than 150 W at 12 VDC; Battery time – 45 min/charge.

to reorder the data cube into 256 spatial images, one for each spectral band as shown in Figure 12.

The prototype is packaged in two chassis, a compact optical head and a portable electronic chassis. The size, weight, and power are shown in Table 4.

LAFS was designed to also work as an airborne imaging spectrometer. The framing mirror can be locked so it views a scene directly below the aircraft (push broom mode) or the mirror can be programmed to sweep a 256 lines cross track to the aircraft flight path (whisk broom mode). In the whisk broom mode, the scan mirror rotation arc can be increased for a wider field of view than in the push broom mode. LAFS illustrates the decrease in size, weight, and power of a push broom imaging spectrometer that results from a 256 × 256 pixel array rather than single pixel line arrays of a line scanner. The array increases integration time of a line by a factor of 256, which allows longer detector integration dwell time on each pixel and thus smaller light collecting optics.

List of Units and Nomenclature

Data cube	Multiple images of a scene in many spectral bands. The spectra of each pixel can be obtained by plotting the spectral value on the same pixel in each spectral band image.
Pixel	An image point (or small area) defined by the instantaneous-field-of-view of either the optics, entrance aperture or slit, and the detector area, or by some combination of these components. A pixel is the smallest element in a scene that is resolved by the imaging system.
Instantaneous-field-of-view (IFOV)	The IFOV is used in describing scanners to define the size of the smallest field-of-view (usually in milliradians or microradians) that be resolved by a scanner system.
Field-of-view	The total size in angular dimensions of a scanner or imager.
Hyperspectral	A data cube with many (>48) spectral bands.
Line scanners	An optical/mechanical system that scans one pixel at a time along a line of a scene.
Whisk broom	An optical/mechanical system that scans two or more lines at a time.
Push broom	An optical system that images a complete line at a time.

See also

Diffraction: Diffraction Gratings; Fraunhofer Diffraction. **Fiber Gratings**. **Geometrical Optics:** Prisms. **Imaging:** Hyperspectral Imaging; Interferometric Imaging. **Interferometry:** Overview. **Modulators:** Acousto-Optics. **Optical Materials:** Color Filters and Absorption Glasses. **Semiconductor Materials:** Dilute Magnetic Semiconductors; Group IV Semiconductors, Si/SiGe; Large Gap II–VI Semiconductors; Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Spectroscopy:** Fourier Transform Spectroscopy; Raman Spectroscopy.

Further Reading

Baselow R, Silvergate P, Rappaport W, *et al.* (1992) *HYDICE (Hyperspectral Digital Imagery Collection Experiment Instrument Design*. International Symposium on Spectral Sensing Research.

Bianchi, R, Marino, CM and Pignatti, S *Airborne Hyperspectral Remote Sensing in Italy.* Pomezia (Roma), Italy: CNR, Progetto LARA, 00040.

Borough HC, Batterfield OD, Chase RP, Turner RW and Honey FR (1992) *A Modular Wide Field of View Airborne Imaging Spectrometer Preliminary Design Concept*. International Symposium on Spectral Sensing Research.

Carmer DC, Horvath R and Arnold CB (1992) *M7 Multispectral Sensor Testbed Enhancements*. International Symposium on Spectral Sensing Research.

Hackwell JA, Warren DW, Bongiovi RP, *et al.* (1996) LWIR/MWIR Imaging Hyperspectral Sensor for Airborne and Ground-Based Remote Sensing. In: Descour M and Mooney J (eds) *Imaging Spectrometry II*. Proc. SPIE 2819, pp. 102–107.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

James JF and Sternberg RS (1969) *The Design of Optical Spectrometers*. London: Chapman and Hall.

Kulgein NG, Richard SP, Rudolf WP and Winter EM (1992) *Airborne Chemical Vapor Detection Experiment*. International Symposium on Spectral Sensing Research.

Lucey PG, Williams T, Horton K, Hinck K and Buchney C (1994) *SMIFTS: A Cryogenically Cooled, Spatially Modulated, Imaging Fourier Transform Spectrometer for Remote Sensing Applications*. International Symposium on Spectral Sensing Research.

Mourales P and Thomas D (1998) Compact Low-Distortion Imaging Spectrometer for Remote Sensing. *SPIE Conference on Imaging Spectrometry IV* 3438: 31–37.

Pagano TS (1992) *Design of the Moderate Resolution Imaging Spectrometer – NADIR (MODIS-N)*. American Society of Photogrametry & Remote Sensing, pp. 374–385.

Stanich CG and Osterwisch FG (1994) *Advanced Operational Hyperspectral Scanners MIVIS and AHS*. First International Airborne Remote Sensing Conference and Exhibition, Strasbourg, France: Environmental Research Institute of Michigan.

Sun X and Baker JJ (2001) *A Hight Definition Hyperspectral Imaging System*. Proceedings, Fifth International Airborne Remote Sensing Conference and Exhibition, Veridian.

Torr MR, Baselow RW and Torr DG (1982) Imaging Spectrascopy of the Thermosphere from the Space Shuttle. *Applied Optics* 21: 4130.

Torr MR, Baselow RW and Mounti J (1983) An imaging spectrometric observatory for spacelab. *Astrophys. Space Science*. 92: 237.

Vane G, Green RO, Chrien TG, Enmark HT, Hansen EG and Porter WM (1993) The Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). *Remote Sensing of the Environment* 44: 127–143.

Warren CP, Anderson RD, Velasco A, Lin PP, Lucas JR and Speer BA (2001) *High Performance VIS and SWIR Hyperspectral Imagers for Airborne Applications*. Fifth International Conference and Exhibition, Veridian.

Wolfe WL (1997) *Introduction to Imaging Spectrometers*. SPIE Optical Engineering Press.

Wolfe WL and Zisfis GJ (eds) (1985) *The Infrared Handbook*, revised edition. Washington, DC: Office of Naval Research, Department of the Navy.

Zywicki RW, More KA, Holloway J and Witherspoon N (1994) *A Field Portable Hyperspectral Imager: The Large Area Fast Spectrometer.* Proceedings of the International Symposium on Spectral Sensing Research.

Telescopes

M M Roth, Astrophysikalisches Institut Potsdam, Potsdam, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In a broad sense, telescopes are optical instruments which provide an observer with an improved view of a distant object, where improved may be defined in terms of magnification, angular resolution, and light collecting power. Historically, the invention of the telescope was a breakthrough with an enormous impact on fundamental sciences (physics, astronomy), and, as a direct consequence, on philosophy; but also on other technological, economical, political, and military developments. It seems that the first refractive telescopes were built in the Netherlands near the end of the sixteenth century by Jan Lipperhey, Jakob Metius, and Zacharias Janssen. Based on reports of those first instruments, Galileo Galilei built his first telescope, which was later named after him, and made his famous astronomical observations of sunspots, the phases of Venus, Jupiter's moons, the rings of Saturn, and his discovery of the nature of the Milky Way as an assembly of very many stars. The so-called astronomical telescope was invented by Johannes Kepler, who in 1611 also published in *Dioptrice* the theory of this telescope. The disturbing effects of spherical and chromatic aberrations were soon realized by astronomers, opticians, and other users, but it was not before Moor Hall (1792) and John Dollond (1758) that at least the latter flaw was corrected for by introducing the achromat. Later, the achromat was significantly improved by Peter Dollond, Jesse Ramsden, and Josef Fraunhofer. The catadioptric telescope was probably introduced by Lenhard Digges (1571) or Nicolas Zucchius (1608). In 1671, Sir Isaac Newton was the first to use a reflector for astronomical observations. Wilhelm Herschel improved this technique and began in 1766 to build much larger telescopes with mirror diameters up to 1.22 m. When, in the nineteenth century, the conventional metal mirror was replaced by glass, it was Léon Foucauld (1819–1868) who applied a silver coating in order to improve the reflectivity of its surface.

Modern refractors can be built with superb apochromatic corrections, and astronomical reflectors have advanced to mirror diameters of up to 8.4 m (monolithic mirror), and 10 m (segmented mirror).

While the classical definition of a telescope involves an observer, i.e., the eye, in modern times often electronic detectors have replaced human vision. In the following, we will, therefore, use a more general definition of a telescope as an optical instrument, whose purpose is to image a distant object, either in a real focal plane, or in an afocal projection for the observation by eye.

Basic Imaging Theory

In the most simple case, a telescope may be constructed from a single lens or a single mirror, creating a real focal plane. Let us, therefore, introduce some basic principles of imaging using optical elements with spherical surfaces.

Refraction at a Spherical Surface

Figure 1 shows an example of refraction at a spherical surface. A ray emerging from object O hits the surface between two media of refractive index n and n' at point P. After refraction, the ray continues at a deflected angle ε' towards the normal and intersects the optical axis at point O′. In the ideal case, all rays emerging from O at different angles σ are collected in O′, thus forming a real image of object point O in O′. In the paraxial approximation, we have $\sin\sigma = \tan\sigma = \sigma$, and $p \approx 0$. In triangle (O, O′, P): $\sigma' - \sigma = i - i'$. Using Snell's law $ni = n'i'$, we obtain:

$$\sigma' - \sigma = i'\frac{n' - n}{n} \qquad [1]$$

Since $\phi = \sigma' + i'$, and in the paraxial approximation:

$$\sigma' n' - \sigma n = \phi(n' - n) \qquad [2]$$

$$\frac{h}{s'}n' - \frac{h}{s}n = \frac{h}{r}(n' - n) \qquad [3]$$

we obtain finally:

$$\frac{n'}{s'} - \frac{n}{s} = \frac{n' - n}{r} \qquad [4]$$

For an object at infinity, where $s = \infty$, s' becomes the focal distance f':

$$f' = r\frac{n'}{n' - n} \quad \text{and} \quad \tilde{f} = -r\frac{n}{n' - n} \qquad [5]$$

Lenses

Combining two or more refractive surfaces, as in the preceding section, allows us to describe a single lens or more complex optical systems with several lenses in series. For a single lens in air, we just give the lens equation (without deriving it in detail):

$$s_2'(F') = \frac{1}{n - 1}\,\frac{r_2[nr_1 - (n - 1)d]}{(n - 1)d + n(r_2 - r_1)} \qquad [6]$$

where d is the lens thickness, measured between the vertices of its surfaces, r_1, r_2 are the radii of the surfaces, n the index of refraction of the lens material, and $s_2'(F')$ the distance of the object image from the vertex of the last surface, facing the focal plane.

The back focal distance is given by:

$$f' = \frac{1}{n - 1}\,\frac{nr_1r_2}{(n - 1)d + n(r_2 - r_1)} \qquad [7]$$

An important special case is where the lens thickness is small compared to the radii of the lens surfaces, i.e., when $d \ll |r_2 - r_1|$. In this case we can neglect the term of $(n - 1)d$ and write:

$$f' = \frac{1}{n - 1}\,\frac{r_1r_2}{(r_2 - r_1)} \qquad [8]$$

Reflection at a Spherical Surface

Let us now consider the other simple case of imaging object O into O′ by reflection from the concave spherical surface, as depicted in Figure 2. In the paraxial case, we have:

$$\sigma = y/s, \quad i = \phi - \sigma, \quad \sigma' = y/s', \quad i' = \phi - \sigma', \quad \phi = y/r \qquad [9]$$

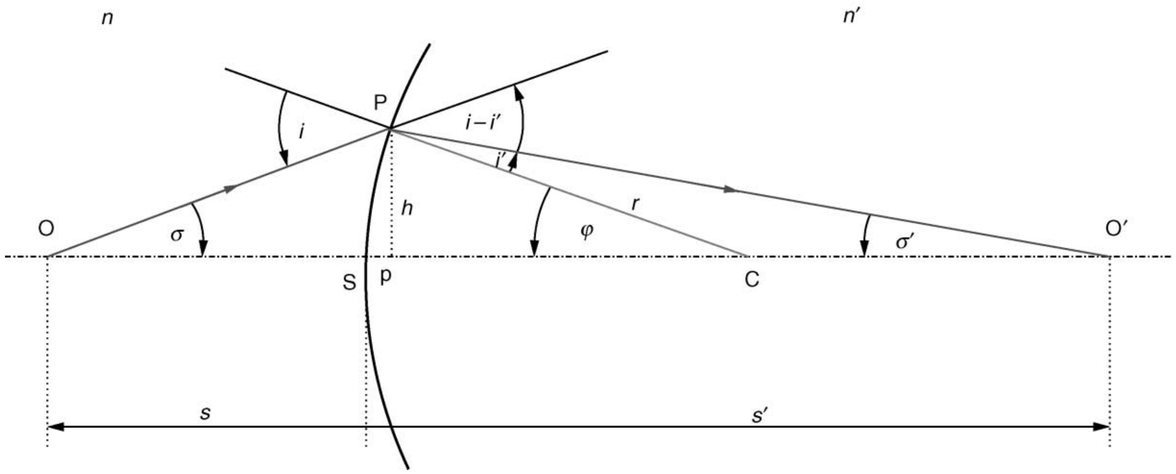

Figure 1 Refraction at a spherical surface between media with refractive indices n and n'.

With $i = -i'$, we obtain:

$$\frac{1}{s'} - \frac{1}{s} = -\frac{2}{r} \qquad [10]$$

Again, with $s = \infty$, the focal length becomes:

$$f' = -\frac{r}{2} \qquad [11]$$

In what follows, we will now assume that the distance of the object is always large (∞), which, in general, is the situation when a telescope is employed. The image is formed at a surface defined by O′, ideally a focal plane (paraxial case). We will consider deviations from the ideal case below.

Simple Telescopes

As we have observed above, a telescope is in principle nothing but an optical system, which images an object at infinity. In the most simple case, a lens with positive power or a concave mirror will satisfy this condition. The principle of a refractor and a reflecting telescope is shown in Figures 3 and 4. In both cases, parallel light coming from the distant object is entering through the entrance pupil, experiences refraction (reflection) at the lens (mirror), respectively, and is converging to form a real image in the focal plane.

Two point sources, for example two stars in the sky, separated by an angle δ, will be imaged in the focal plane as two spots with separation Δy. When f is the focal length of the telescope, we therefore have:

$$\tan\delta = \frac{\Delta y}{f} \qquad [12]$$

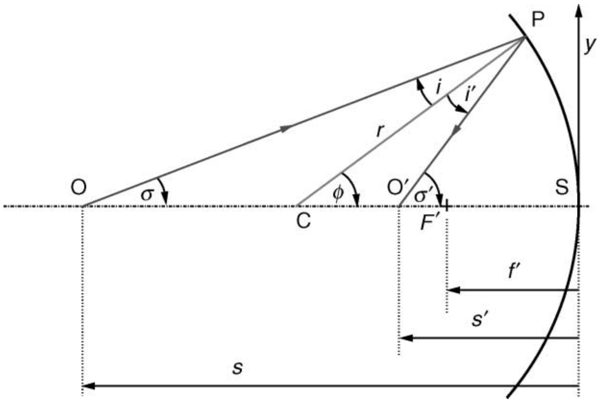

Figure 2 Reflection at a spherical surface with radius r.

The so-called plate scale m of the telescope in units of arcsec/mm is given by:

$$m = 3600\,\mathrm{atan}\frac{1}{f} \quad (f \text{ in mm}) \qquad [13]$$

In the case of the refractor, we have also shown an eyepiece, illustrating that historically the instrument was invented to enhance the human vision. The eye is located at the position of the exit pupil, thus receiving parallel light, i.e., observing an object at infinity, however (i) with a flux which is increased by a factor A, given by the ratio of the area of the telescope aperture with diameter D to the area of the eye's pupil d with $A = (D/d)^2$, and (ii) with a magnification M of the apparent angle between two separate objects at infinity, which is given by the ratio of the focal lengths of the objective and the eyepiece:

$$\Gamma = \frac{f_{\mathrm{obj}}}{f_{\mathrm{ocl}}} \qquad [14]$$

The magnification can also be expressed as the ratio of the diameters of the exit pupil and the entrance pupil:

$$\Gamma = \frac{D_{\mathrm{expup}}}{D_{\mathrm{enpup}}} \qquad [15]$$

The first astronomical telescopes with mirrors were equipped with an eyepiece for visual observation. The example in Figure 4 shows a configuration which is more common nowadays for professional astronomical observations, using a direct imaging detector in the focal plane (photographic plate, electronic camera).

Aberrations

Reflector Surfaces

We shall now study the behavior of a real telescope beyond the ideal in the case of a reflector. Figure 5 shows the situation of a mirror, imaging an on-axis

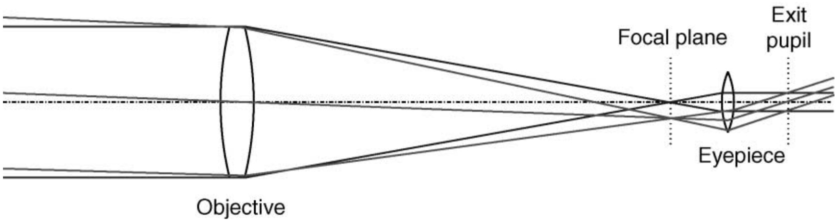

Figure 3 Refractor, astronomical telescope for visual observations with eyepiece (Kepler telescope).

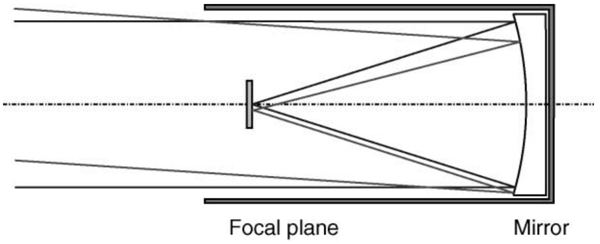

Figure 4 Reflector, astronomical telescope with prime focus imager.

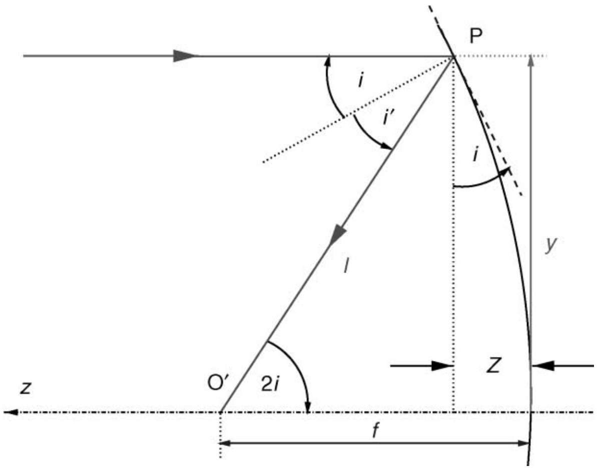

Figure 5 Imaging an object at infinity with reflector.

object at infinity into focus at point O′. With Fermat's principle, the optical paths of all rays forming the image must have identical length l. For all j, we must satisfy:

$$Z_j + f = l \tag{16}$$

From the geometry in Figure 5, we find also:

$$y_i^2 + (f - Z_j)^2 = l^2 \tag{17}$$

Combining eqns [16] and [17], we obtain the equation for a surface which satisfies the condition for perfect on-axis imaging:

$$y^2 = -4fz \tag{18}$$

Obviously only a paraboloid is capable of fulfilling the condition for all rays parallel to the optical axis at whatever separation y. Also, we see that spherical mirrors are less than ideal imagers: rays at increasing distance y from the axis will suffer from increasing path length differences. The effect is called spherical aberration. We will investigate aberrations more quantitatively further below.

Elementary axisymmetrical surfaces can be expressed as follows, with parameter C being the conic constant:

$$y^2 - 2Rz + (1 + C)z^2 = 0 \tag{19}$$

and the following characteristics:

$C > 0$	ellipsoidal, prolate
$C = 0$	spherical
$-1 < C < 0$	ellipsoidal, oblate
$C = -1$	paraboloidal
$C < -1$	hyperboloidal

Let us now leave the paraxial approximation and determine the focal length of rays which are parallel to, but further away at a distance y from the optical axis. Figure 5 shows the relevant geometrical relations. The intersection of the reflected ray with the optical axis at distance f from the vertex can be found from:

$$f = z + (f - z) \quad \text{where} \quad \frac{y}{f - z} = \tan(2i) \tag{20}$$

We use the slope of the tangent in point P:

$$\frac{\mathrm{d}z}{\mathrm{d}y} = \tan i \tag{21}$$

differentiate eqn [19] to get:

$$\frac{\mathrm{d}z}{\mathrm{d}y} = \frac{y}{R - (1 + C)z'} \tag{22}$$

use the trigonometrical relation:

$$\tan 2\phi = \frac{2\tan\phi}{1 - \tan^2\phi} \tag{23}$$

and substitute for tan(2i) in eqn [20] to obtain:

$$f = \frac{R}{2} + \frac{(1 - C)z}{2} - \frac{y^2}{2(R - (1 + C)z)} \tag{24}$$

In order to express eqn [24] solely in terms of y, we rewrite eqn [19]:

$$z = \frac{R}{1 + C}\left(1 - \sqrt{1 - \frac{y^2}{R^2}(1 + C)}\right) \tag{25}$$

Expanding z in a series and neglecting any orders higher than 6:

$$z = \frac{y^2}{2R} + (1 + C)\frac{y^4}{8R^3} + (1 - C)^3\frac{y^6}{16R^5} \tag{26}$$

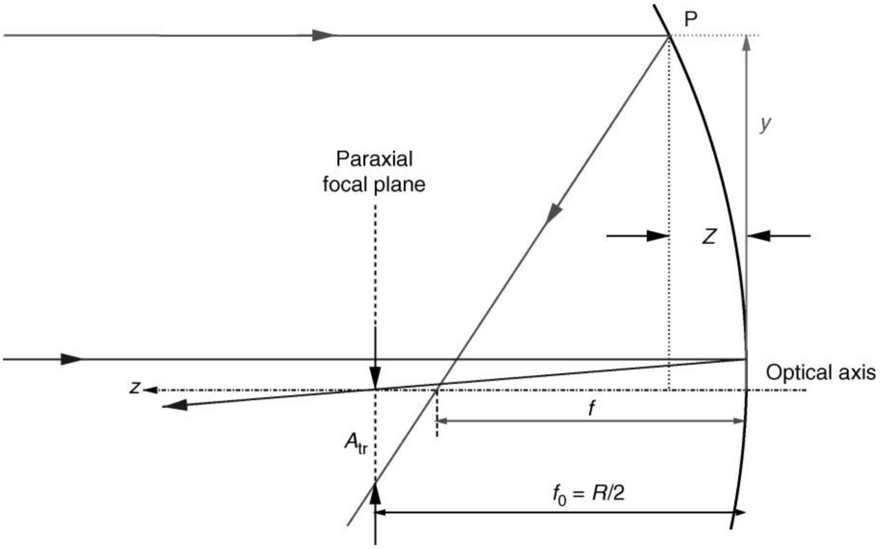

Figure 6 Transverse spherical aberration.

we obtain finally:

$$f = \frac{R}{2} - \frac{1+C}{4R}y^2 - \frac{(1+C)(3+C)}{16R^3}y^4 \qquad [27]$$

.

Transverse Spherical Aberration

Obviously, for any $C \neq -1$, rays of different distance y to the optical axis are focused at different focal lengths f (spherical aberration). The net effect for a distant point source is that the image is blurred, rather than forming a sharp spot. The lateral deviation of nonparaxial rays intersecting the nominal focal plane at $f_0 = R/2$ is described by the transverse spherical aberration A_{tr} (Figure 6):

$$A_{\text{tr,sph}} = \frac{y}{f-z}(f_0 - f) \qquad [28]$$

Using eqns [19] and [27], one obtains:

$$A_{\text{tr,sph}} = -\frac{1+C}{2R^2}y^3 + \frac{3(1+C)(3+C)}{8R^4}y^5 + \cdots \qquad [29]$$

Angular Spherical Aberration

It is sometimes more convenient to consider the angular spherical aberration, since a telescope is naturally measuring angles between objects at infinity. To this end, we compare an arbitrary reflector surface according to eqn [19] with a reference paraboloid, which is known to be free of on-axis spherical aberration (Figure 7). For any given ray, we observe how the reflected ray under study is tilted towards the reference ray, which would be reflected from the ideal surface of the paraboloid. While the former is measured under an exit angle of $2i$ towards the optical axis (see Figure 5), we designate the latter $2i_{\text{p}}$. The angular spherical aberration is defined as the difference of these two angles:

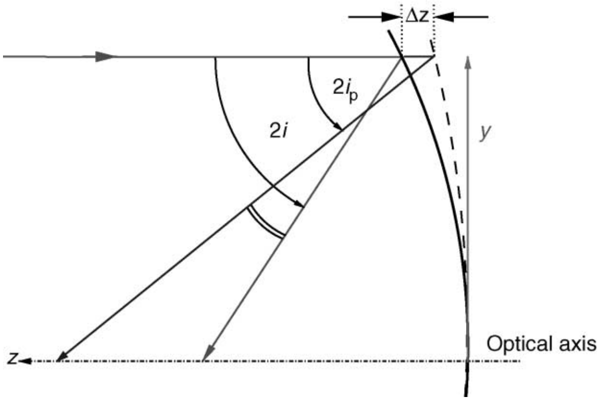

Figure 7 Angular aberration determined with reference to a paraboloid.

$$A_{\text{ang,sph}} = 2i - 2i_{\text{p}} = 2(i - i_{\text{p}}) \qquad [30]$$

Again, we can use the slope of the tangent to determine the change of angles $i \rightarrow i_{\text{p}}$:

$$A_{\text{ang,sph}} = 2\left(\frac{\mathrm{d}z}{\mathrm{d}y} - \frac{\mathrm{d}z_{\text{p}}}{\mathrm{d}y}\right) = 2\frac{\mathrm{d}}{\mathrm{d}y}\Delta z \qquad [31]$$

It is, therefore, sufficient to consider the difference in z between the surface under study and the reference paraboloid. From eqn [26], we obtain as an approximation to third order:

$$A_{\mathrm{ang,sph}} = -(1 + C)\frac{y^3}{R^3} \qquad [32]$$

Aberrations in the Field

Since the use of a telescope exclusively on-axis would be somewhat limited, let us finally investigate the behavior under oblique illumination, i.e., when an object at infinity is observed under an angle Θ towards the optical axis of the system. This will allow us to assess aberrations in the field. For simplicity, we are considering a paraboloid, remembering that on-axis the system is free of spherical aberration. Again, we are employing an imaginary reference paraboloid which is pointing on-axis towards the object at angle Θ (Figure 8). Its coordinate system (O', z', y') is displaced and tilted with regard to the system under study (O, z, y). One can show that the offset Δz, which is used to derive the angular aberration in the same way as above, is given by:

$$2\Delta z = a_1 \frac{y^3 \Theta}{R^2} + a_2 \frac{y^2 \Theta^2}{R} + a_3 y \Theta^3 \qquad [33]$$

The angular aberration in the field is then:

$$A_{\mathrm{ang,f}} = 3a_1 \frac{y^2 \Theta}{R^2} + 2a_2 \frac{y \Theta^2}{R} + a_3 \Theta^3 \qquad [34]$$

The coefficients are called:

a_1: coma
a_2: astigmatism
a_3: distortion

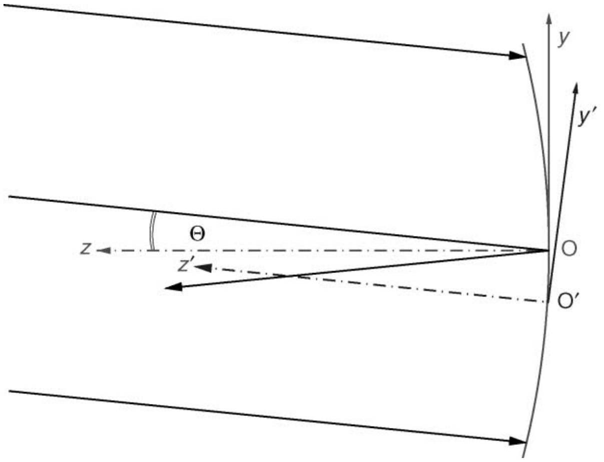

Figure 8 Aberration in the field, oblique rays.

Chromatic Aberration

So far we have only considered optical aberrations of reflecting telescopes. Spherical aberration in the case of a lens is treated in analogous ways and will not be discussed again. However, the wavelength dependence of the index of refraction in optical media (dispersion) is an important factor and gives rise to chromatic aberrations which are only encountered in refractive optical systems. We can distinguish two principal effects in the paraxial approximation: (i) longitudinal chromatic aberration, resulting in different axial image positions as a function of wavelength; and (ii) lateral chromatic aberration, which can be understood as an aberration of the principal ray as a function of wavelength, leading to a varying magnification as a function of wavelength (plate scale in the case of a telescope).

As an example, let us determine the wavelength dependence of the focal length of a single thin lens. Opticians are conventionally using discrete wavelengths of spectral line lamps of different chemical elements, e.g., the Mercury e line at 546.1 nm, Cadmium F' 480.0 nm, or Cadmium C' 643.8 nm, covering more or less the visual wavelength range of the human eye. Differentiation of eqn [8] yields for the wavelength of e:

$$\frac{f'}{\mathrm{d}n} = -\frac{1}{(n_e - 1)^2}\frac{r_1 r_2}{r_2 - r_1} \qquad [35]$$

Substituting

$$\frac{r_1 r_2}{r_2 - r_1} = f'_e(n_e - 1) \qquad [36]$$

we get:

$$\mathrm{d}f' = -\frac{\mathrm{d}n}{n_e - 1} f'_e \qquad [37]$$

Replacing the differential by the differences $\Delta f' = f'_{F'} - f'_{C'}$ and $\Delta n = n_{F'} - n_{C'}$, then:

$$f'_{F'} - f'_{C'} = -\frac{n_{F'} - n_{C'}}{n_e - 1} f'_e = -\frac{f'_e}{\nu_e} \qquad [38]$$

ν_{e} is called Abbe's Number:

$$\nu_e = \frac{n_e - 1}{n_{F'} - n_{C'}} \qquad [39]$$

It is a tabulated characteristic value of dispersion for any kind of glass and allows one to quickly estimate the difference in focal length at the extreme wavelengths F' and C' when the nominal focal length of a lens is known as e.

Angular Resolution

In the previous sections, we have exclusively employed geometrical optics for the description of telescope properties. With geometrical optics and in the absence of aberrations, there is no limit to the angular resolution of a telescope, i.e., there is no limiting separation angle δ_{lim}, below which two distant point sources can no longer be resolved. However, real optical systems experience diffraction, leading to the finite width of the image of a point source (Airy disk), whose normalized intensity distribution in the focal plane as a function of radius r' from the centroid for the case of a circular aperture of diameter D can be conveniently described by a Bessell function (λ: wavelength, f': focal length of the objective):

$$\frac{I(r)}{I_0} = \left[\frac{2J_1(\nu)}{\nu}\right]^2 \quad \text{where} \quad \nu = \frac{\pi D r'}{\lambda f'} \qquad [40]$$

The intensity distribution is plotted in Figure 9. The first four minima of the diffraction ring pattern are listed in Table 1. Two point sources are defined to be at the resolution limit of the telescope when the maximum of one diffraction pattern happens to coincide with just the first diffraction minimum of the other one (Figure 9, right). As a rule of thumb, the angular resolution is of a telescope with diameter D is given by:

$$\delta_{\min} = 1.22\frac{\lambda}{D} \rightarrow \frac{138}{D} \qquad [41]$$

the latter in units of arcsec for a wavelength of 550 nm and D in mm.

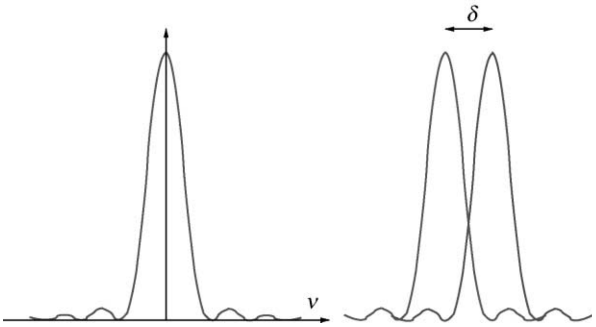

Figure 9 Left: radial intensity distribution of diffraction pattern (Airy disk). Right: two overlapping point source images with same intensity at limiting angular separation δ.

Table 1 Diffraction minima of airy disk

n	ν_n	$r'D/2f'\lambda$
1	3.83	0.61
2	7.02	1.12
3	10.2	1.62
4	13.3	2.12

Types of Telescopes

There are very many different types of telescopes which were developed for different purposes. In order to obtain an impression, we will only briefly introduce a nonexhaustive selection of refractors and reflectors, which by no means is representative of the whole. For a more complete overview, the reader is referred to the literature.

Refractor Objectives

Refractors are first of all characterized by their objectives. Figure 10 shows a sequence of lenses, indicating the progress of lens design developments from left to right. All lenses are groups 2 or 3, designed to compensate for chromatic aberration through optimized combinations of crown and flint glasses. The first lens is a classical C achromat, e.g., using BK7 and SF2 glass, which is essentially the design which was introduced by Hall and Dollond. It is still a useful objective for small telescopes, where a residual chromatic aberration is tolerable. The second lens is the Fraunhofer E type, which is characterized by an air gap between the two lenses (BK7, F2), which helps to optimize the image quality of the system, reducing spherical aberration and coma. As a disadvantage, the two additional glass–air interfaces lead to ghost images (reflected light) and Fresnell losses. The objective is critical with regard to the alignment of tip and tilt. The third example is an improved AS type achromat (KzF2, BK7), which is similar to the E type in

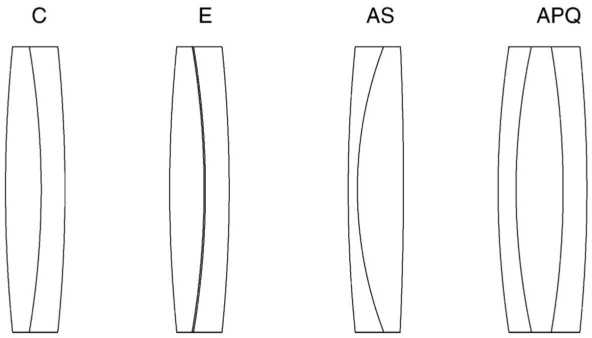

Figure 10 Different types of refractor objectives, from left to right: classical C achromat, Fraunhofer E achromat with air gap, AS achromat, APQ apochromat.

that it also possesses an air gap, however with an improved chromatic aberration correction in comparison to the Fraunhofer objective. The lenses show a more pronounced curvature and are more difficult to manufacture. The beneficial use of KzF2 glass is an example of the successful research for the development of new optical media, which was pursued near the end of the nineteenth century in a collaboration between Otto Schott and Ernst Abbé. The last example is a triplet, employing ZK2, CaF_2, and ZK2. This APQ type of objective is an apochromat, whose chromatic aberration vanishes at 3 wavelengths. It is peculiar for its use of a crystal (calcium fluoride, CaF_2) as the optical medium of the central lens. The manufacture of the CaF_2 lens is critical and expensive, but it provides for an excellent correction of chromatic and spherical aberration of this high quality objective. The triplet is also difficult to manufacture because of the coefficient of thermal expansion, which is roughly different by a factor of two between CaF_2 and the glass. In order to avoid disrupture of the cemented lens group, some manufacturers have used oil immersion to connect the triplet. Since CaF_2 not only has excellent dispersion properties to correct for chromatic aberration, but also exhibits an excellent transmission beyond the visual wavelength regime, this type of lens is particularly useful for applications in the blue and UV.

Finally, Figure 11 shows the complete layout of an astrograph, which is a telescope for use with photographic plates. The introduction of this type of instrument meant an important enhancement of astrophysical research by making available a significantly more efficient method, compared to the traditional visual observation of individual stars, one by one. Using photographic plates, it became possible to measure position and brightness of hundreds and thousands of stars, when previously one could investigate tens. Typical astrographs were built with apertures between 60 and 400 mm, and a field-of-view as large as 15 degrees. Despite the growing importance of space-based observatories for astrometry, some astrographs are still in use today.

Basic Oculars

While the general characteristics of a refractor are dominated by the objective, the eyepiece nevertheless plays an important role for the final appearance to the observer, with an effect on magnification, image quality, chromatism, field of view, and other properties of the entire system. From a large variety of different ocular types, let us briefly consider a few simple examples.

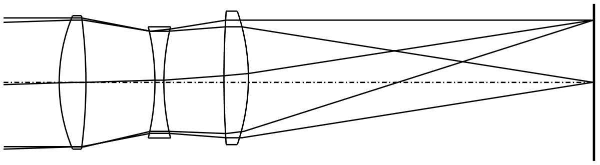

Figure 11 Astrograph.

The first eyepiece in Figure 12 is the one of the Kepler telescope, which was already introduced in Figure 3. It has a quite simple layout with a single lens of positive power. The lens is located after the telescope focal plane at a distance which is equal to the ocular focal length f_oc. The telescope and the ocular focal planes are coinciding, and as a result, the outcoming beam is collimated, forming the system exit pupil after a distance f_oc. This is where ideally the observer's eye is located. The focal plane is creating a real image, which is a convenient location for a reticle. The Kepler telescope is, therefore, a good tool for measurement and alignment purposes. We note that the image is inverted, which is relatively unimportant for a measuring telescope and astronomy, but very annoying for terrestrial use.

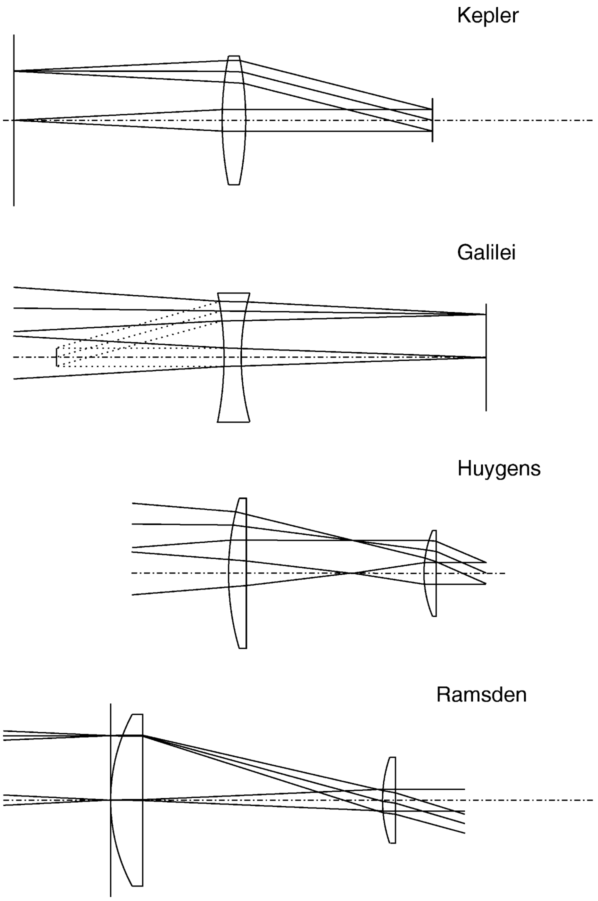

Figure 12 Elementary eyepieces after Kepler, Galilei, Huygens, and Ramdsen.

This latter property has been avoided in the next example, which is the ocular of the Galilean telescope, which was built by Galileo and used for his famous observations of planets and moons in the solar system. It uses a single lens, but contrary to the Kepler type with negative power. The lens is located in the telescope beam in front of the focal plane of the objective, which means that the exit pupil is virtual and not accessible to the observer (indicated by dotted lines in Figure 12). The overall length of the telescope is somewhat shorter than Kepler's, however at the expense of the lack of a real focal plane, so no reticle can be applied.

The third example is the 2-lens-ocular of Christian Huygens, who was able to prove with his design that it is possible to correct paraxially the lateral chromatic aberration. The eyepiece is still being built today.

Another elementary ocular is shown in Figure 12, which is the eyepiece after Ramsden. This system is also corrected for lateral chromatic aberration. The real image is formed on the front surface of the second lens, providing a location for a reticle or a scale for measuring purposes. The exit pupil is interior to the ocular. In addition to these selected basic types, there are many more advanced multilens oculars in use today, with excellent correction and large field of view (up to 60°).

Reflectors

Let us now describe several basic types of reflectors (see Figure 13). The first example is the Newton reflector, which uses a flat folding mirror to create an accessible exterior focus. It is conceptually simple and still in use today for amateur astronomy. When the main mirror is a paraboloid, the spherical aberration is corrected on the axis of the system.

The second example represents an important improvement over the simple 1-mirror design of Newton's telescope, in that the combination of a parabolic first mirror (main mirror, M1) with a hyperbolic secondary mirror (M2) eliminates spherical aberration, but not astigmatism and coma. The Cassegrain telescope has a curved focal surface which is located behind the primary mirror. Due to the onset of coma and astigmatism, the telescope has a modest useful field-of-view. Besides the correction properties, the system has significant advantages in terms of mechanical size and weight, playing an important role for the construction of large telescopes. The basic type of this design has practically dominated the construction of the classical large astronomical telescopes with M1 diameters of up to 5 m, and, more recently, the modern 8–10 m class telescopes. On the practical side, the external focus of the Cassegrain on the bottom of the mirror support cell is a convenient location for mounting focal plane instruments (e.g., CCD cameras, spectrographs, and others).

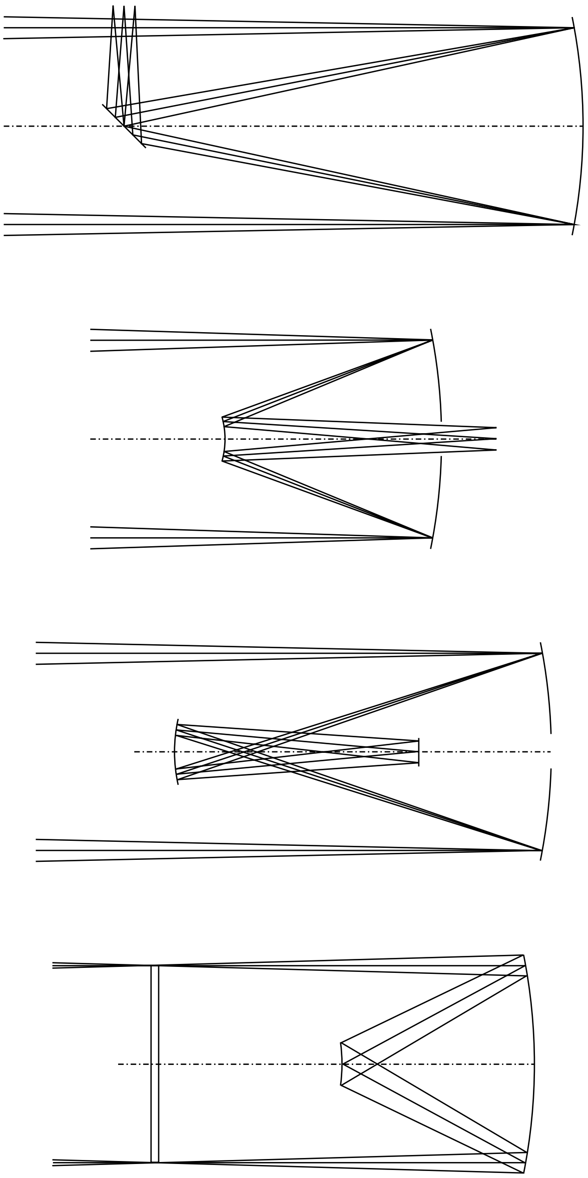

Figure 13 Reflectors, from top to bottom: Newton, Cassegrain, Gregory, Schmidt.

The more common variant of the Cassegrain type, which was actually used for the design of these telescopes, is the layout of Ritchey (1864–1945) and Cretien (1879–1956), which combines two hyperbolic mirrors for M1 and M2 (RCC), but otherwise has the same appearance as the Cassegrain. Besides spherical aberration, it also corrects for coma, giving a useful field-of-view of up to 1°. Like in the case of the Cassegrain telescope, the RCC exhibits some curvature of the focal plane.

Our third example is the Gregory reflector, which also is a 2-mirror system, however with a concave elliptical secondary mirror. The correction of this telescope is similar, but somewhat inferior to the one of the Cassegrain. The use of a concave M2 is advantageous for manufacture and testing. Nevertheless, the excessive length of the system has precluded the use of this layout for large astronomical telescopes, except for stationary beams, for example, in solar observatories.

The last example is the design put forward by Bernhard Schmidt (1879–1935) who wanted to create a telescope with an extremely wide field-of-view. The first telescope of his new type achieved a field-of-view of 16° – something which had not been feasible with previous techniques. The basic idea is to use a fast spherical mirror, and to correct for spherical aberration by means of an aspherical corrector plate in front of the mirror, which also forms the entrance pupil of the telescope. The focal surface is strongly curved, making it necessary to employ a detector with the same radius of curvature (bent photographic plates), or to compensate the curvature by means of a field lens immediately in front of the focus. The presence of the corrector plate and, if used, the field flattener, gives rise to significant chromatism. Nevertheless, the Schmidt telescope was a very successful design and dominated wide-angle survey work in astronomy for many decades.

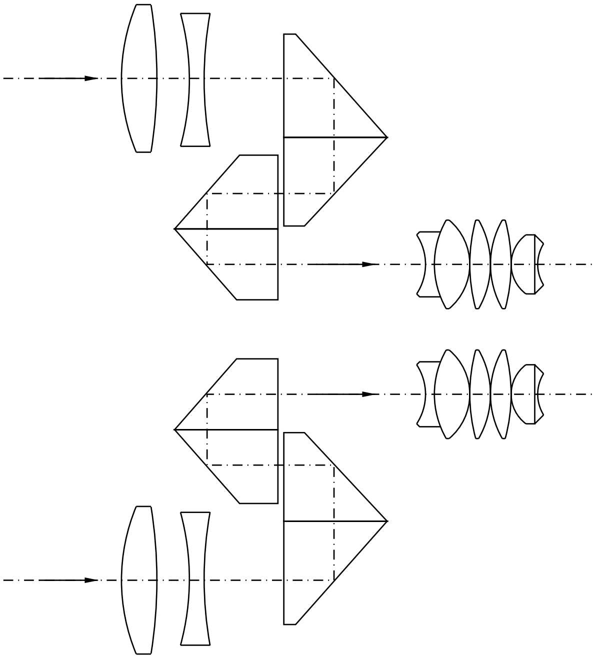

Figure 14 Principle of prism binoculars with folded beams.

Applications

General Purpose Telescopes

According to the original meaning of the Greek 'tele–skopein', the common conception of the telescope is the one of an optical instrument, which provides an enlarged view of distant objects. In addition, a large telescope aperture, compared to the eye, improves night vision and observations with poorly illuminated scenes. There are numerous applications in nautical, security, military, hunting, and other professional areas. For all of these applications, an intuitive vision is important, essentially ruling out the classical astronomical telescope (Kepler telescope, see Figure 3), which produces an inverted image.

Binoculars with Porro prisms are among the most popular instruments with upright vision, presenting a compact outline due to the folded beam geometry (Figure 14), but also single tube telescopes with re-imaging oculars are sometimes used.

Metrology

As we have realized in the introduction, telescopes are instruments which perform a transformation from angle of incidence in the aperture to distances from the origin in the focal plane. If the focal length of the objective is taken sufficiently long, and a high-quality eyepiece is used, very small angles down to the arcsec regime can be visually observed. Telescopes are, therefore, ideal tools to measure angles to very high precision.

In Figure 15 we see a pair of telescopes, facing each other, where one instrument is equipped with a light source, and the other one is used to observe the first one. The first telescope is also called a collimator. It possesses a lamp, a condensor system, a focal plane mask (usually a pinhole, a crosshair, or some other useful pattern imprinted on a transparent plate), and the telescope objective. The focal plane mask is projected to infinity. It can thus be observed with the

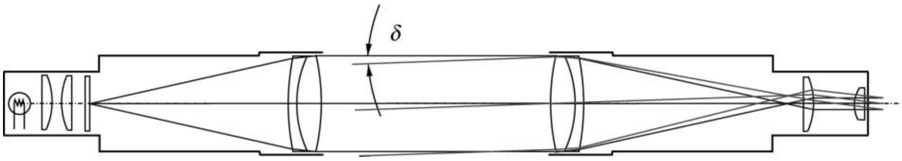

Figure 15 Collimator and telescope.

measuring telescope, such that an image of the collimator mask appears in the telescope focal plane. Also the telescope is equipped with a mask in its focal plane, usually with a precision scale or another pattern for convenient alignment with the collimator mask. If now the collimator is tilted by a small angle δ against the optical axis of the telescope, the angle of incidence in the measuring telescope changes by this same amount, giving rise to a shift of the image by $\Delta x = f \tan \delta$. Note that the measurement is completely insensitive to parallel shifts between the telescope and the collimator, since it is parallel light which is transmitted between the two instruments.

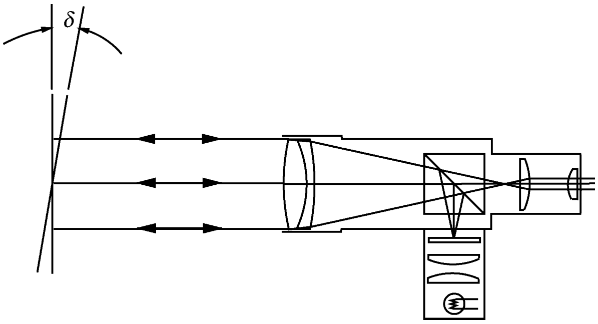

Figure 16 Autocollimator.

Another special device, which combines the former two instruments into one, is the Autocollimator. The optical principle is shown in Figure 16: in principle, the autocollimator is identical to a measuring telescope, except that a beamsplitter is inserted between the focal plane and the objective. The beamsplitter is used to inject light from the same kind of light source device as in the normal collimator. The axis of the folded light source beam and of the illuminated collimator mask are aligned to match the focal plane mask of the telescope section: the optical axis of the emerging beam is identical to the optical axis of the telescope. Under this condition, the reflected light from a test object will be exactly centered on the telescope focal plane mask, when it is perfectly aligned to normal incidence. Normally, a high-quality mirror is attached to the device under study, but sometimes also the test object itself possesses a plane surface of optical quality, which can be used for the measurement.

Figure 17 shows schematically an example for such a measurement, where a mirror is mounted on a linear stage carriage for the purpose of measuring the flexure of the stage to very high precision. With a typical length of ≈ 1 m, this task is not a trivial problem for a mechanical measurement.

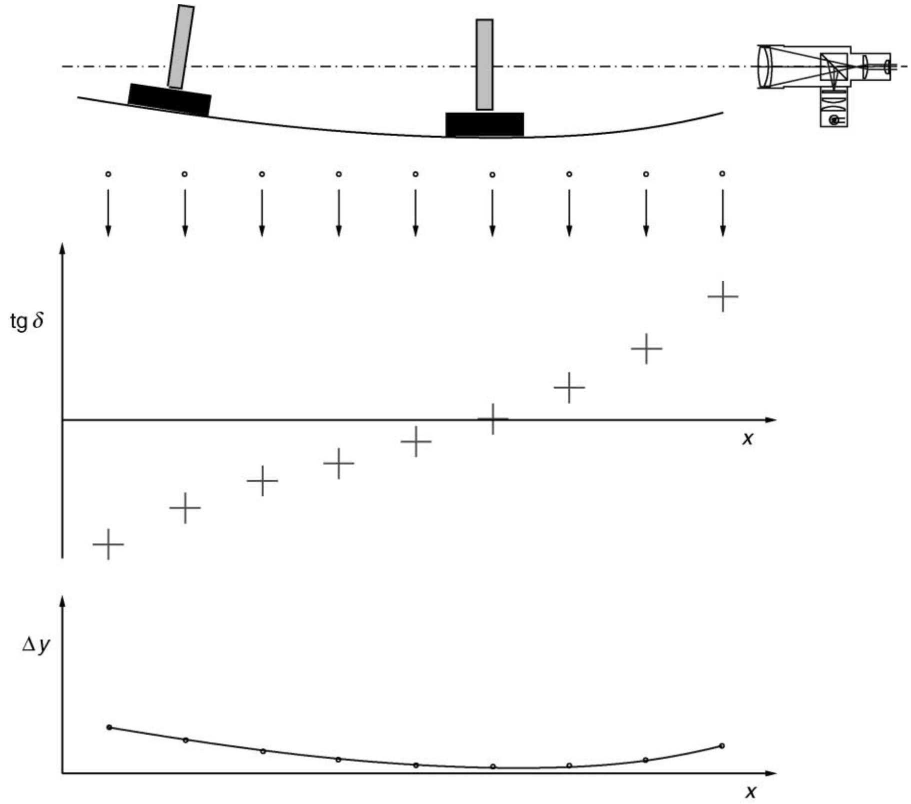

Figure 17 Using an autocollimator to measuring the flexure of a linear stage.

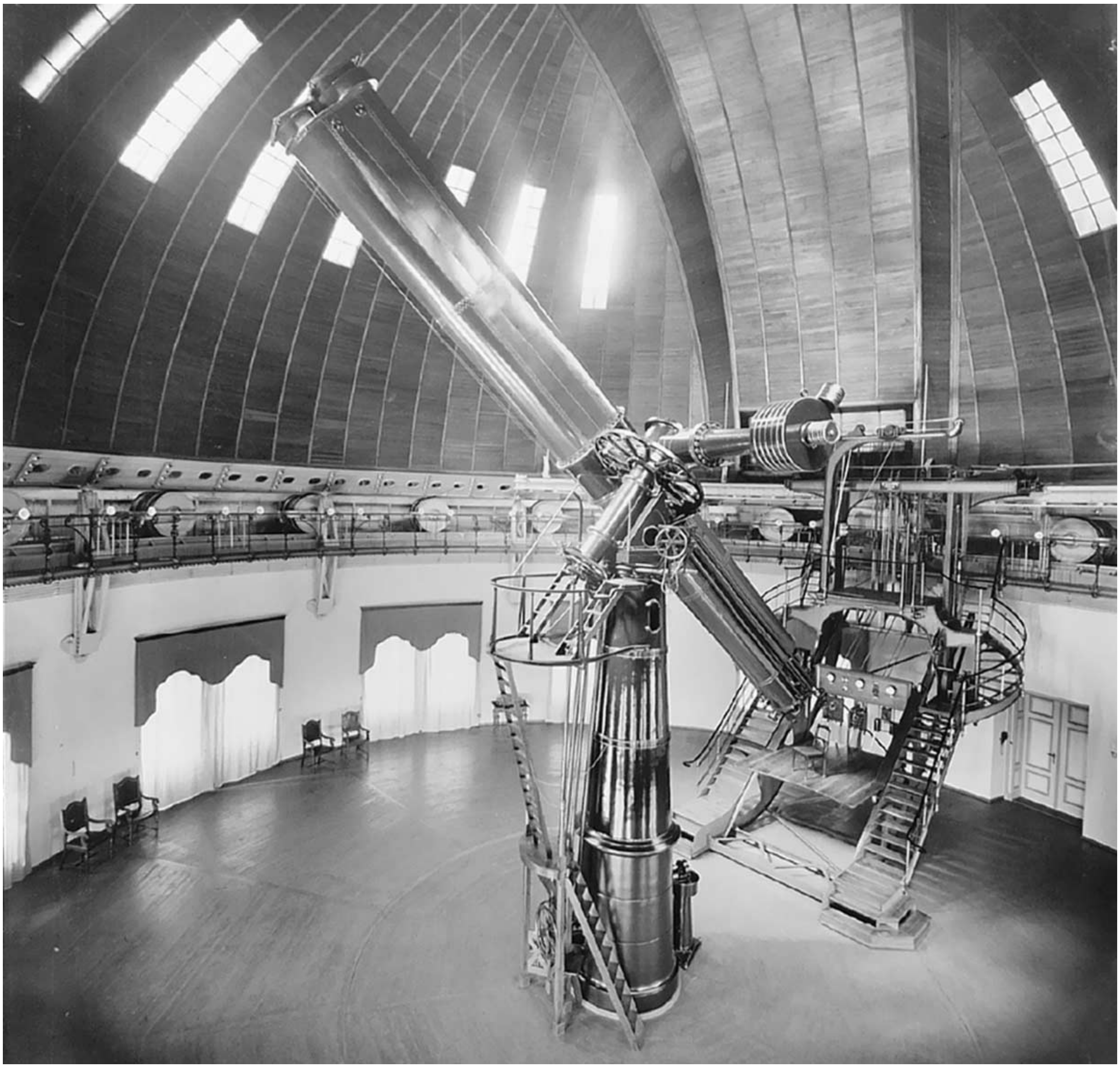

Figure 18 The 80 cm + 60 cm 'Great Refractor' at the Astrophysical Observatory Potsdam. The correction of the 80 cm objective of this telescope led Johannes Hartmann to the development of the 'Hartmann Test', which was the first quantitative method to measure aberrations. This method is still in use today, and has been further developed to become the 'Shack-Hartmann' technique of wavefront sensing. Courtesy of Astrophysikalisches Institut Potsdam.

By measurement with the autocollimator the tilt of the mirror consecutively for a sufficiently large number of points x_i along the rail, we can now plot the slope of the rail at each point x_i as tan d:

$$\frac{dy}{dx} = \tan\delta \qquad [42]$$

Integrating the series of slope values reproduces the shape of the rail $y = f(x)$.

Other applications include multiple autocollimation, wedge measurements, and right angle measurements.

Multiple autocollimation is achieved by inserting a fixed plane-parallel and semi-transparent mirror between the autocollimator and the mirror under study. Light coming from the test mirror will be partially transmitted through the additional mirror towards the autocollimator, and partially reflected back to the test mirror, and so forth. After k reflections, a deflection of $k\delta$ is measured, when δ is the tilt of the test mirror.

The wedge angle α of a plane-parallel plate with refractive index n_w is easily measured by observing the front and back surface reflections from the plate, keeping in mind that the beam bouncing back from the rear will also experience refraction in the plate. If δ is the measured angle difference between the two front and rear beams, then $\alpha = \delta/2n_w$.

Two surfaces of an object can be tested for orientation at right angles by attaching a mirror to surface 1 and observe the mirror with the fixed autocollimator. In a second step, an auxiliary pentaprism is inserted into the beam, deflecting the light towards surface 2, where the mirror is attached next. If a precision pentaprism is used, any deviation from right angles between the two surfaces is seen as an offset between step 1 and 2.

These are only a few examples of the very many applications for alignment and angle measurements in optics and precision mechanics.

Without going into much detail, we are just mentioning that also the theodolite and sextant make use of the same principle of angular measurements, however without the need of a collimator or autocollimation.

Astronomical Telescopes

Despite its importance as a measurement and testing device for optical technology and manufacture as described in the preceeding paragraph, it is the original invention and subsequent improvement of the telescope for *Astronomy* which is probably the most fascinating application of any optical system known to the general public. The historical development of the astronomical telescope spans from the 17th century Galileian refractor with an aperture of just a few centimeters in diameter, over the largest refractors built for the Lick and Yerkes observatories with objectives of up to ≈1 m in diameter (see also Figures 18–20), the 4 m-class reflectors of the second half of the 20th century like e.g. the Hale Observatory 5m, the Kitt Peak and Cerro Tololo 4 m, the La Silla 3.6 m, or the La Palma 4.2 m telescopes (to name only a few), to the current state-of-the-art of 8–10 m class telescopes, prominent examples being the two Keck Observatory 10 m telescopes on Mauna Kea (Hawaii), or the ESO Very Large Observatory (Paranal, Chile), consisting of four identical 8.2 m telescopes (Figures 21–24).

This development has always been driven by the discovery of fainter and fainter celestial objects, whose quantitative study is the subject of modern astrophysics, involving many specialized disciplines such as atom/quantum/nuclear physics, chemistry, general relativity, electrodynamics, hydrodynamics, magneto-hydrodynamics, and so forth. In fact, cosmic objects are often referred to as ‘laboratories’

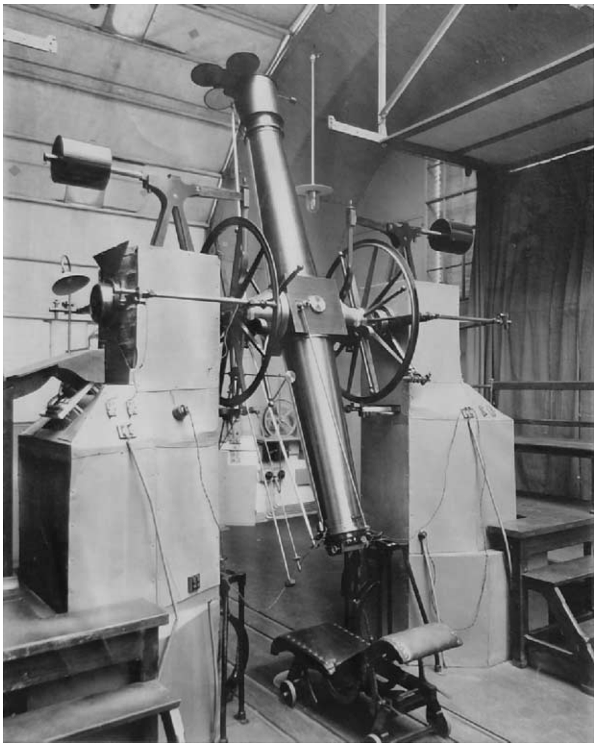

Figure 19 Pistor & Martin Meridian Circle at the former Berlin Observatory in Potsdam Babelsberg. The instrument was one of the finest art telescopes with ultra-high precision for astrometry, built in 1868. Courtesy of Astrophysikalisches Institut Potsdam.

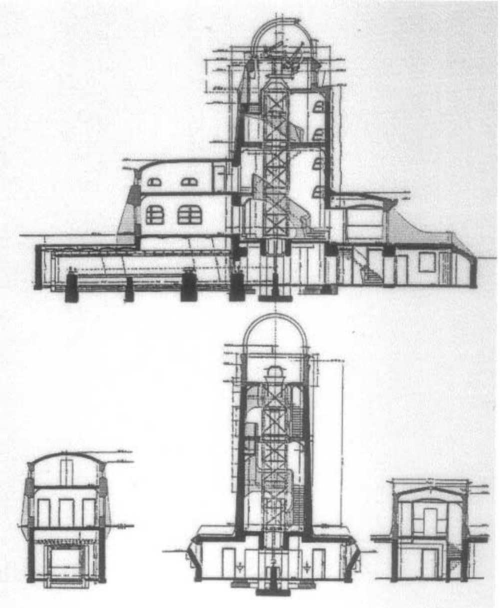

Figure 20 Architect’s sketch of historical ‘Einsteinturm’ solar telescope, Astrophysical Observatory Potsdam. The telescope objective is mounted in the tower structure, facing upward a siderostat which is mounted inside the dome. The light is focused in the basement after deflection from a folding mirror on the slit of a bench-mounted high resolution spectrograph. The telescope represented state-of-the-art technology at the beginning of the twentieth century. Courtesy of Astrophysikalisches Institut Potsdam.

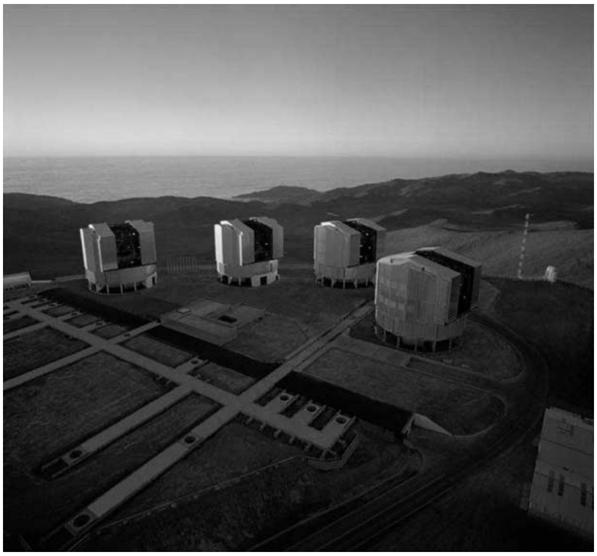

Figure 21 The Very Large Telescope Observatory (VLT) on Mt Paranal, Chile, with VLT Interferometer (VLTI). The four large enclosures are containing the active optics controlled VLT unit telescopes, which are operated either independently of each other, or in parallel to form the VLTI. The structures on the ground are part of the beam combination optics/delay lines. Courtesy of European Southern Observatory.

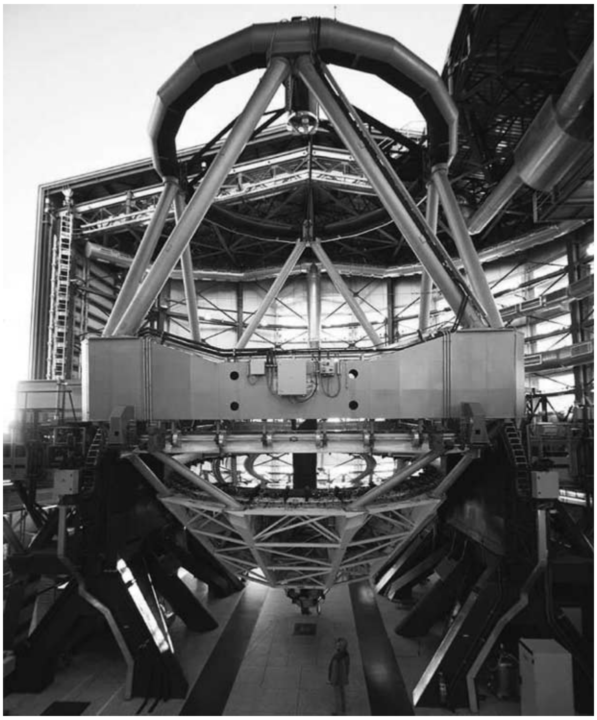

Figure 22 Total view of Kueyen, one of four unit telescopes of the VLT observatory. Courtesy of European Southern Observatory.

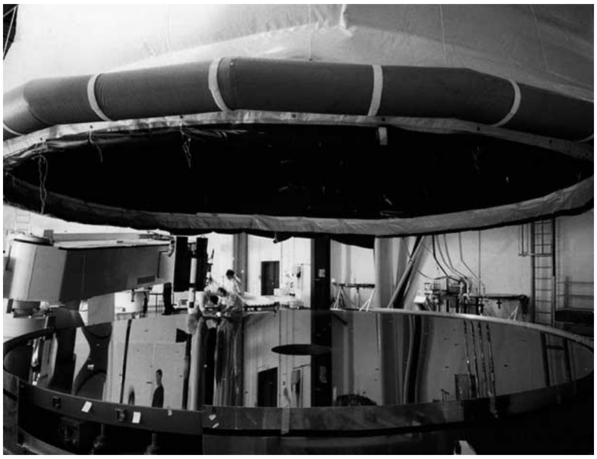

Figure 23 8.2 m thin primary mirror for the VLT. The VLT design is based on thin meniscus mirrors, made out of ZERODUR, which have low thermal inertia and which are always kept at ambient temperature, thus virtually removing the effect of thermal turbulence (mirror seeing). Seeing is the most limiting factor of ground-based optical astronomy. The optimized design of the VLT telescopes has introduced a significant leap forward to a largely improved image quality from the ground, and, consequently, improved sensitivity. The thin mirrors are constantly actively controlled in their mounting cells, using numerous piston actuators on the back surface of the mirror. Another key technology under development is 'adaptive optics', a method to compensate atmospheric wavefront deformations in real-time, thus obtaining diffraction limited images. Courtesy of European Southern Observatory.

with extreme conditions of density, temperature, magnetic field strength, etc., which would be impossible to achieve in any laboratory on Earth. Except for particles (neutrinos, cosmic rays), gravitational waves, and the physical investigation of meteorites and other solar system material collected by space probes, telescopes (for the entire wavelength range of the electromagnetic spectrum) are the *only* tool to accomplish quantitative measurements of these cosmic laboratories.

In an attempt to unravel the history of the birth and evolution of the universe, to discover and to measure the first generation of stars and galaxies, and to find evidence for traces of life outside of planet Earth, plans are currently underway to develop yet another new generation of giant optical telescopes with apertures of 30–100 m in diameter (ELT: extremely large telescope). The technological challenges in terms of precision and stability are outstanding.

Along with the impressive growth of light collecting area, we must stress the importance of angular resolution, which has also experienced a significant evolution over the history of astronomical telescopes: from ≈5 arcsec of the Galilei telescope, which already provided an order of magnitude improvement over the naked eye, over roughly 1 arcsec of conventional telescopes in the 20th century, a few hundredths of an arcsec for the orbiting Hubble Space

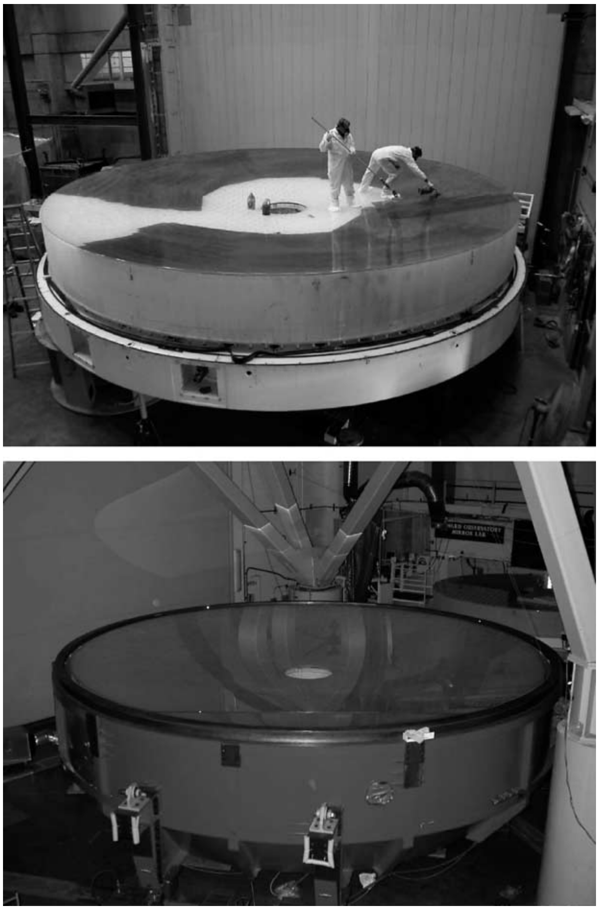

Figure 24 Top: spin-cast 8.4 m honeycomb mirror #1 for the Large Binocular Telescope, the largest monolithic mirror of optical quality in the world. Shown in the process of applying a protective plastic coating to the end-polished surface. Bottom: LBT mirror #1, finally mounted in mirror cell. Courtesy of Large Binocular Telescope Observatory.

Telescope (Figure 25), down to milli-arcsec resolution of the Keck and VLT Interferometers. The development of the aforementioned ELTs is thought to be only meaningful if operated in a diffraction-limited mode, using the emerging technique of *adaptive optics*, which is a method to correct for wavefront distortions caused by atmospheric turbulence (image blurr) in real-time.

The combination of large optical/infrared telescopes with powerful detectors and focal plane instruments (*see* **Instrumentation:** Astronomical Instrumentation) has been a prerequisite to make astrophysics one of the most fascinating disciplines of fundamental sciences.

See also

Geometrical Optics: Aberrations. **Imaging:** Adaptive Optics. **Instrumentation:** Astronomical Instrumentation.

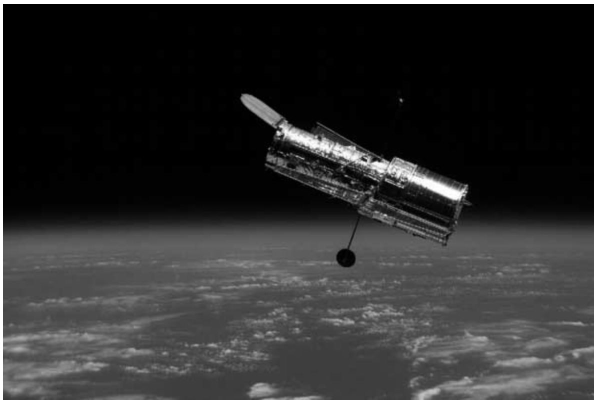

Figure 25 Hubble Space Telescope (HST) in orbit. This 2.4 m telescope is probably the most productive telescope which was ever built. Due to the absence of the atmosphere, the telescope delivers diffraction-limited, extremely sharp images, which have revealed unprecedented details of stars, star clusters, nebulae and other objects in the Milky Way and in other galaxies. Among the most spectacular results, HST has provided the deepest look into space ever, revealing light from faint galaxies that was emitted when the universe was only at 10% of its present age. Courtesy of Space Telescope Science Institute, Baltimore.

Further Reading

Bely PY (ed.) (2003) *The Design and Construction of Large Optical Telescopes*. New York: Springer.

Born M and Wolf E (1999) *Principles of Optics*, 7th ed. Cambridge: Cambridge University Press.

Haferkorn H (1994) *Optik*, 3rd ed. Leipzig: Barth Verlagsgesellschaft.

Korsch D (1991) *Reflective Optics*. San Diego: Academic Press.

Laux U (1999) *Astrooptik*, 2nd ed. Heidelberg: Verlag Sterne und Weltraum.

McCray WP (2004) *Giant Telescopes: Astronomical Ambition and the Promise of Technology*. Cambridge: Harvard University Press.

Osterbrock DE (1993) *Pauper and Prince: Richey, Hale, and Big American Telescopes*. Tucson: University of Arizona Press.

Riekher R (1990) *Fernrohre und ihre Meister*, 2nd ed. Berlin: Verlag Technik GmbH.

Rutten HGJ and van Venrooij MAM (1988) *Telescope Optics*. Richmond, VA: Willman-Bell.

Schroeder DJ (1987) *Astronomical Optics*. San Diego: Academic Press.

Welford WT (1974) *Aberrations of the Symmetrical Optical System*. London: Academic Press.

Wilson RN (1996) *Reflecting Telescope Optics I: Basic Design Theory and its Historical Development*. Berlin: Springer.

Wilson RN (1999) *Reflecting Telescope Optics II: Manufacture, Testing, Alignment, Modern Techniques*. Berlin: Springer.

INTERFEROMETRY

Contents

Overview
Gravity Wave Detection
Phase-Measurement Interferometry
White Light Interferometry

Overview

J C Wyant, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interferometers are a powerful tool used in numerous industrial, research, and development applications, including measuring the quality of a large variety of manufactured items such as optical components and systems, hard disk drives and magnetic recording heads, lasers and optics used in CDs and DVD drives, cameras, laser printers, machined parts, components for fiber optics systems, and so forth. Interferometers can also be used to measure distance, spectra, and rotations, etc. The applications are almost endless.

There are many varieties of interferometers. Some interferometers form interference fringes using two beams and some use multiple beams. Some interferometers are common path; some use lateral shear; some use radial shear; and some use phase-shifting techniques. Examples of the various types are given below, but first we must give the basic equation for two-beam interference.

Two-Beam Interference

When two quasi-monochromatic waves interfere, the irradiance of the interference pattern is given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos[\phi] \qquad [1]$$

where the I's represent the irradiance of the individual beams and ϕ is the phase difference between the two interfering beams. The maximums of the interference occur when $\phi = 2m\pi$, where m is an integer. In most applications involving interferometers, ϕ is the quantity of interest since it is related to the quantity being measured. Figure 1 shows typical two-beam interference fringes.

There are many two-beam interferometers and we will now look at a few of them.

Fresnel Mirrors

Fresnel Mirrors are a very simple two beam interferometer. A beam of light reflecting off two mirrors, set at a slight angle to one another, will produce an interference pattern. The fringes are straight and equally spaced. The fringes become more finely spaced as the angle between the two mirrors increases, d increases, and D becomes smaller (Figure 2).

Plane Parallel Plate

When a monochromatic source illuminates a plane parallel plate, as shown in Figure 3, the interference

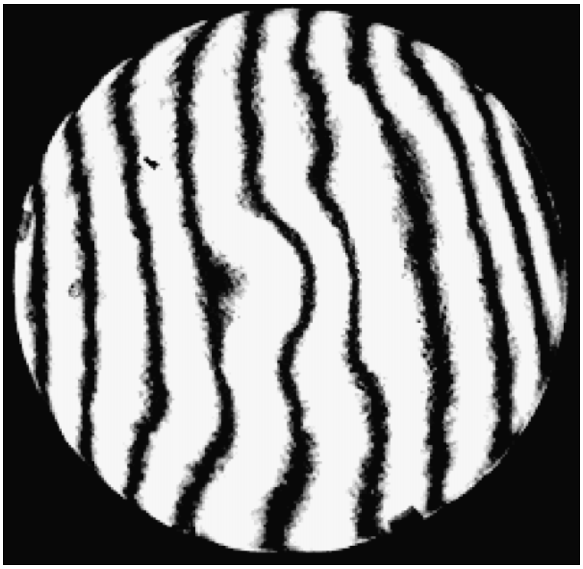

Figure 1 Typical two-beam interference fringes.

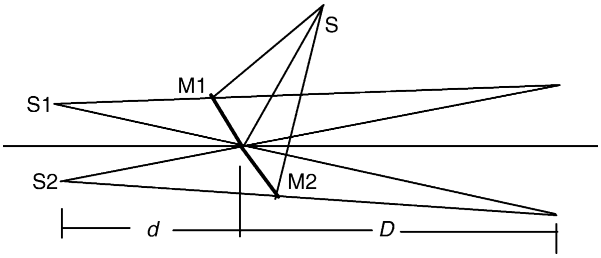

Figure 2 Fresnel mirrors.

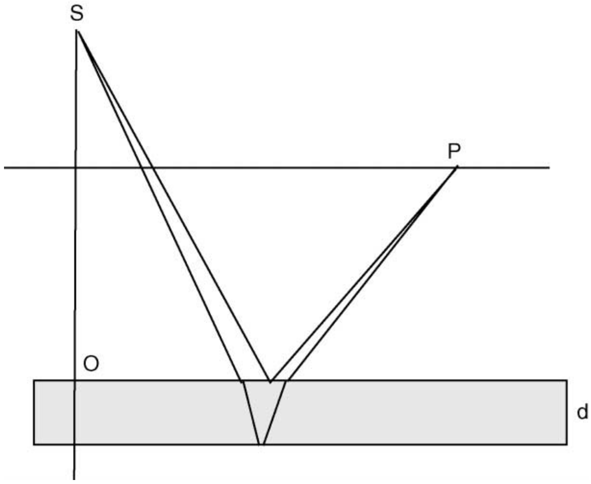

Figure 3 Plane parallel plate interferometer.

fringes are circles centered on the normal to plate. The fringes are called Haidinger fringes, or fringes of equal inclination since, for a given plate thickness, they depend on the angle of incidence. If the maximum or minimum occurs at the center, the radii of the fringes are proportional to the square root of integers. If the plate has a slight variation in thickness and it is illuminated with a collimated beam then the interference fringes are called Fizeau fringes of equal thickness and they give the thickness variations in the plate.

Michelson Interferometer

A Michelson interferometer is shown in Figure 4. Light from the extended source is split into two beams by the beamsplitter. The path difference can be viewed as the difference between the mirror M1 and the image of mirror M2, M2′. With the M1 and M2′ parallel, the fringes are circular and localized at infinity. With M1 and M2′ at a slight angle, the interference fringes are straight lines parallel to the equivalent intersection of the mirrors and localized approximately at the intersection. When the mirrors are only a few wavelengths apart, white light fringes appear, and can be used to determine their coincidence.

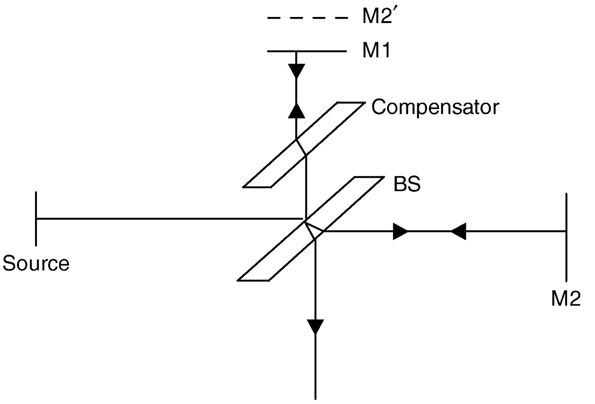

Figure 4 Michelson interferometer.

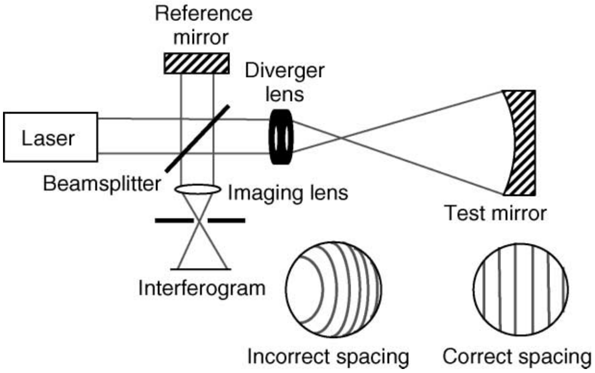

Figure 5 Twyman–Green interferometer for testing a concave spherical mirror.

Twyman–Green Interferometer

If a point source is used in a Michelson interferometer it is generally called a Twyman–Green interferometer. Twyman–Green interferometers are often used to test optical components such as flat mirrors, curved mirrors, windows, lenses, and prisms. Figure 5 shows a Twyman–Green interferometer for the testing of a concave spherical mirror. The interferometer is aligned such that the focus of the diverger lens is at the center of curvature of the spherical mirror. If the mirror under test is perfect, straight equi-spaced fringes are obtained. Figure 6 shows two interference fringes and the relationship between surface height error for the mirror being tested and fringe deviation. λ is the wavelength of the light.

Fizeau Interferometer

The Fizeau interferometer is a simple device used for testing optical surfaces, especially flats and spheres

(Figure 7). The fringes are fringes of equal thickness. It is sometimes useful to tilt the reference surface a little to get several nearly straight interference fringes. The relationship between surface height error and fringe deviation from straightness is the same as for the Twyman–Green interferometer.

Mach–Zehnder Interferometer

A Mach–Zehnder interferometer is sometimes used to look at samples in transmission. Figure 8 shows a typical Mach–Zehnder interferometer for looking at a sample in transmission.

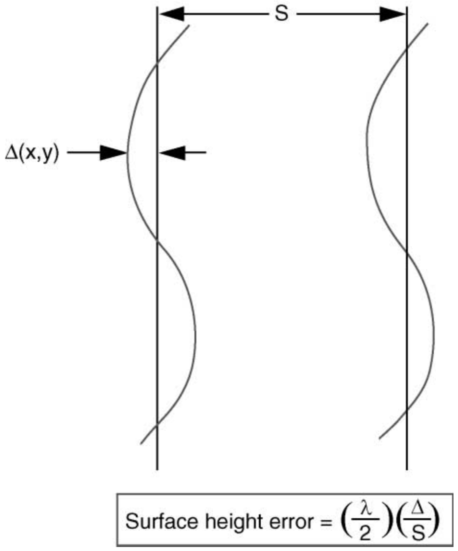
S
Δ(x,y)
Surface height error = $\left(\frac{\lambda}{2}\right)\left(\frac{\Delta}{S}\right)$

Figure 6 Relationship between surface height error and fringe deviation.

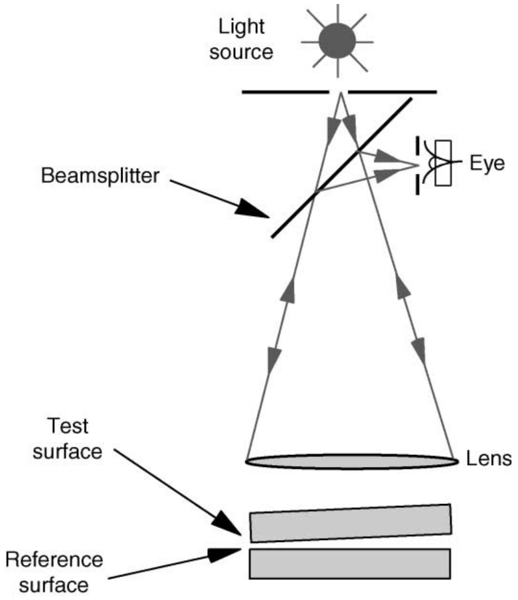
Light source
Beamsplitter
Eye
Test surface
Lens
Reference surface

Figure 7 Fizeau interferometer.

Murty Lateral Shearing Interferometer

Lateral shearing interferometers compare a wavefront with a shifted version of itself. While there are many different lateral shear interferometers, one that works very well with a nearly collimated laser beam is a Murty shearing interferometer shown in Figure 9. The two laterally sheared beams are produced by reflecting a coherent laser beam off a plane parallel plate. The Murty interferometer can be used to measure the aberrations in the lens or it can be used to determine if the beam leaving the lens is collimated. If the beam incident upon the plane parallel plate is collimated a single fringe results, while if the beam is not perfectly collimated straight equi-spaced fringes result where the number of fringes gives the departure from collimation.

Radial Shear Interferometer

Radial shear interferometers compare a wavefront with a magnified or demagnified version of itself. While there are many different radial shear

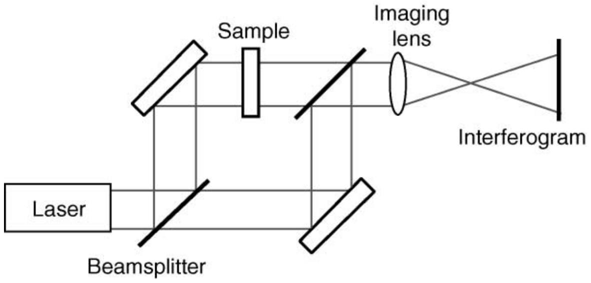
Sample
Imaging lens
Interferogram
Laser
Beamsplitter

Figure 8 Mach–Zehnder interferometer.

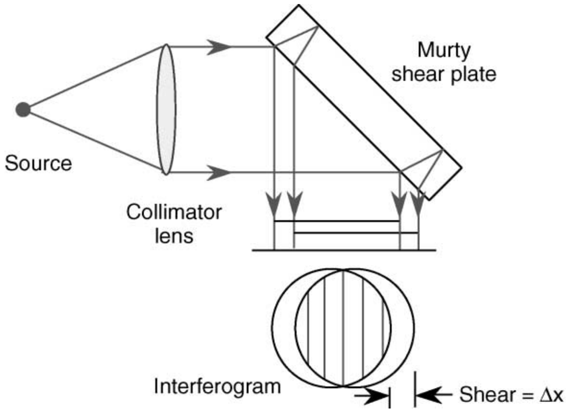
Murty shear plate
Source
Collimator lens
Interferogram
Shear = Δx

Figure 9 Murty lateral shear interferometer.

interferometers, one that works very well with a nearly collimated beam is shown in Figure 10. It is essentially a Mach–Zehnder interferometer where, in one arm, the beam is magnified and in the other arm the beam is demagnified. Interference fringes result in the region of overlap. The sensitivity of the interferometer depends upon the amount of radial shear. If the two interfering beams are approximately the same size there is little sensitivity, while if the two beams greatly differ in size there is large sensitivity. Sometimes radial shear interferometers are used test the quality of optical components.

Scatterplate Interferometer

A scatterplate interferometer, invented by Jim Burch in 1953, is one of the cleverest interferometers. Almost any light source can be used and no high-quality optics are required to do precision interferometry. The critical element in the interferometer is the scatterplate which looks like a piece of ground glass, but the important item is that the scattering points are arranged so the plate has inversion symmetry as shown in Figure 11. The light that is scattered the first time through the plate, and unscattered the second time through the plate, interferes with the light unscattered the first time and scattered the second time. The resulting interference fringes show errors in the mirror under test. This type of interferometer is insensitive to vibration because it is what is called a common path interferometer, in that the test beam (scattered–unscattered) and the reference beam (unscattered–scattered) travel along almost the same paths in the interferometer.

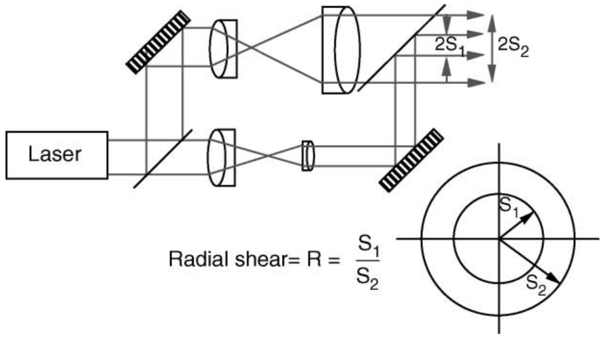

Figure 10 Radial shear interferometer.

Smartt Point Diffraction Interferometer

Another common path interferometer used for the testing of optics is the Smartt point diffraction interferometer (PDI) shown in Figure 12. In the PDI the beam of light being measured is focused onto a partially transmitting plate containing a pinhole. The pinhole removes the aberration from the light passing through it (reference beam), while the light passing through the plate (test beam) does not have the aberration removed. The interference of these two beams gives the aberration of the lens under test.

Sagnac Interferometer

The Sagnac interferometer has beams traveling in opposite directions, as shown in Figure 13. The interferometer is highly stable and easy to align. If the interferometer is rotated with an angular velocity there will be a delay between the transit times of the clockwise and the counterclockwise beams. For this reason, the Sagnac interferometer is used in laser gyros.

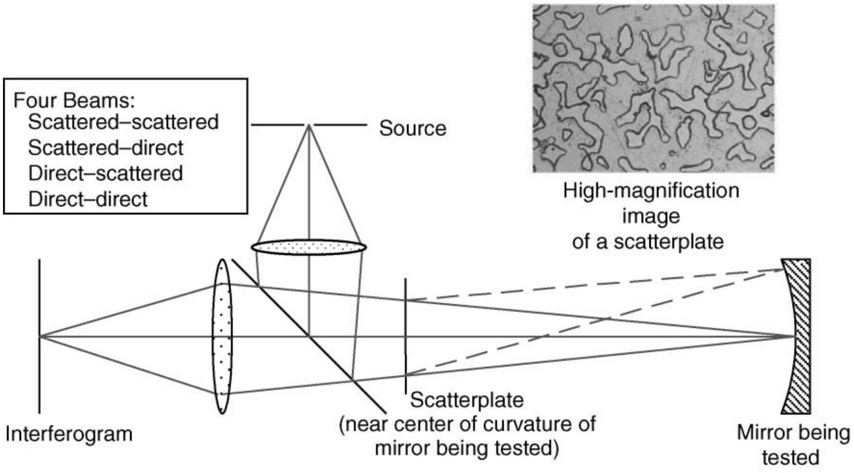

Figure 11 Scatterplate interferometer.

Fiber Interferometers

Fiber interferometers were first used for rotation sensing by replacing the ring cavity in the Sagnac interferometer with a multiloop made of a single-mode fiber. Fiber-interferometer rotation sensors are attractive as rotation sensors because they are small and low-cost.

Since the optical path in a fiber is affected by its temperature and it also changes when the fiber is stretched, fiber interferometers can be used as sensors for mechanical strains, temperature, and pressure. They can also be used for the measurement of magnetic fields by bonding the fiber to a magneto-restrictive element. Electric fields can be measured by bonding the fiber to a piezoelectric film.

Multiple Beam Interferometers

In general, multiple beam interferometers provide sharper interference fringes than a two beam interferometer. If light is reflected off a plane parallel plate there are multiple reflections. If the surfaces of the plane parallel plate have a low reflectivity, the higher-order reflections have a low intensity and the multiple reflections can be ignored and the resulting interference fringes have a sinusoidal intensity profile, but if the surfaces of the plane parallel plate have a high reflectivity, the fringes become sharper. Figure 14 shows the profile of the fringes for both the reflected and the transmitted light for different values of surface reflectivity. Multiple beam interference fringes are extremely useful for measuring the spectral content of a source. One such multiple beam interferometer, that is often used for measuring spectral distributions of a source, is a Fabry–Perot interferometer that consists of two plane-parallel plates separated by an air space, as shown in Figure 15.

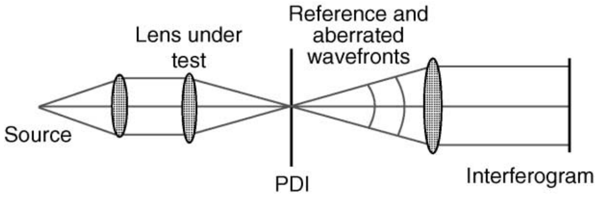

Figure 12 Smartt point diffraction interferometer.

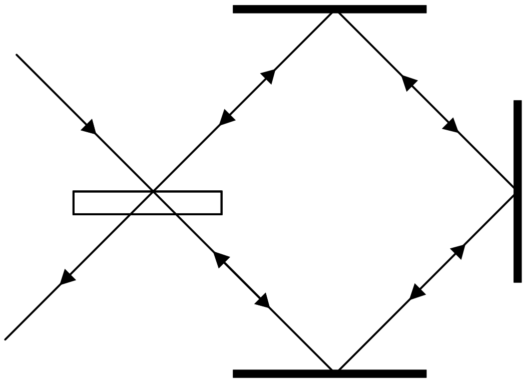

Figure 13 Sagnac interferometer.

Phase-Shifting Interferometry

The equation for the intensity resulting from two-beam interference, contains three unknowns, the two individual intensities and the phase difference between the two interfering beams. If three or more measurements are made of the intensity of the two-beam interference as the phase difference between the two beams is varied in a known manner, all three unknowns can be determined. This technique, called phase-shifting interferometry, is an extremely powerful tool for measuring phase distributions. Generally the phase difference is

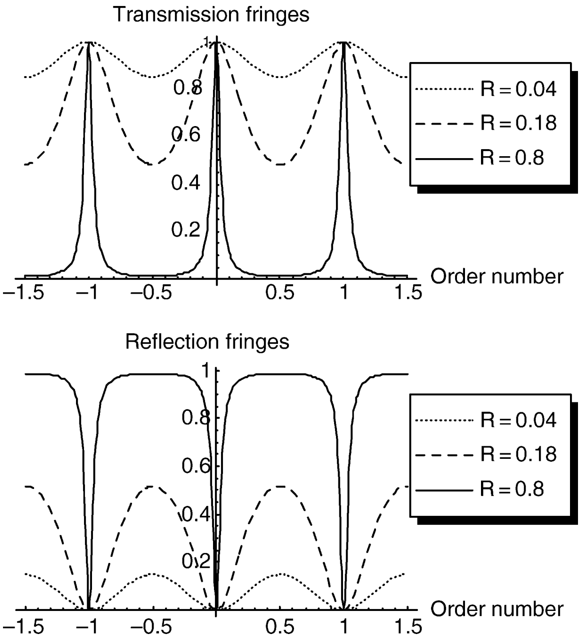

Figure 14 Multiple beam interference fringes for the transmitted and reflected light for a plane parallel plate.

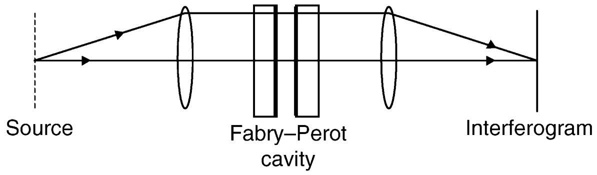

Figure 15 Fabry–Perot interferometer.

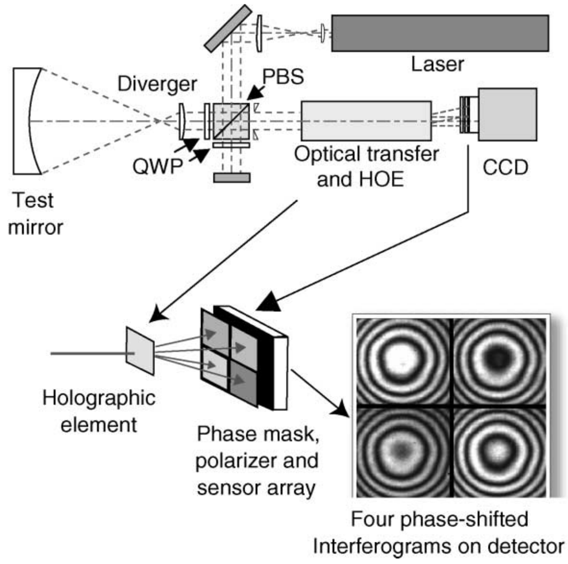
Laser
Diverger
PBS
QWP
Optical transfer and HOE
CCD
Test mirror
Holographic element
Phase mask, polarizer and sensor array
Four phase-shifted Interferograms on detector

Figure 16 Vibration-insensitive phase-shifting interferometer.

changed by 90 degrees between consecutive measurements of the interference intensity in which case the three intensities can be written as

$$I_a(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) - 90°)$$
$$I_b(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y))$$
$$I_c(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\phi(x,y) + 90°)$$

and the phase distribution is given by

$$\phi(x,y) = \mathrm{ArcTan}\left(\frac{I_a(x,y) - I_c(x,y)}{-I_a(x,y) + 2I_b(x,y) - I_c(x,y)}\right)$$

Phase-shifting interferometry has greatly enhanced the use of interferometry in metrology since it provides a fast low noise way of getting interferometric data into a computer.

Vibration-Insensitive Phase-Shifting Interferometer

While phase-shifting interferometry has greatly enhanced the use of interferometry in metrology, there are many applications where it cannot be used because of the environment, especially vibration. One recent phase-shifting interferometer that works well in the presence of vibration is shown in Figure 16. In this interferometer, four phase-shifted frames of data are captured simultaneously. The test and reference beams have orthogonal polarization. After the test and reference beams are combined they are passed through a holographic element to produce four identical beams. The four beams are then transmitted through wave retardation plates to cause 0, 90, 180, and 270 phase difference between the test and reference beams. After passing through a polarizer the four phase-shifted interferograms fall on the detector array. Not only can the effects of vibration be eliminated, but by making short exposures to freeze the vibration, the vibrational modes can be measured. Movies can be made showing the vibration. Likewise, flow fields can be measured.

Combining modern electronics, computers, and software with old interferometric techniques, provides for very powerful measurement capabilities.

List of Units and Nomenclature

I	irradiance
R	intensity reflectance
λ	wavelength
ϕ	phase

See also

Coherence: Speckle and Coherence. **Detection:** Fiber Sensors. **Holography, Techniques:** Computer–Generated Holograms; Digital Holography; Holographic Interferometry. **Imaging:** Adaptive Optics; Interferometric Imaging; Wavefront Sensors and Control (Imaging Through Turbulence). **Interferometry:** Phase Measurement Interferometry; White Light Interferometry; Gravity Wave Detection. **Microscopy:** Interference Microscopy. **Tomography:** Optical Coherence Tomography.

Further Reading

Candler C (1951) *Modern Interferometers*. London: Hilger and Watts.

Francon M (1966) *Optical Interferometry*. New York: Academic Press.

Hariharan P (1991) Selected papers on interferometry. *SPIE*, vol. 28.

Hariharan P (1992) *Basics of Interferometry*. San Diego: Academic Press.

Hariharan P (2003) *Optical Interferometry*. San Diego: Academic Press.

Hariharan P and Malacara D (1995) Selected papers on interference, interferometry, and interferometric metrology. *SPIE*, vol. 110.

Malacara D (1990) Selected papers on optical shop metrology. *SPIE*, vol. 18.

Malacara D (1992) *Optical Shop Testing*. New York: Wiley.

Malacara D (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Decker.

Robinson D and Reid G (1993) *Interferogram Analysis*. Brisol, UK: Institute of Physics.

Steel WH (1985) *Interferometry*. Cambridge: Cambridge University Press.

Tolansky S (1973) *An Introduction to Interferometry*. New York: Longman.

Gravity Wave Detection

N Christensen, Carleton College, Northfield, MN, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the later part of the nineteenth century Albert Michelson performed extraordinary experiments that shook the foundations of the physics world. Michelson's precise determination of the speed of light was an accomplishment that takes great skill to reproduce today. Edward Morley teamed up with Michelson to measure the velocity of the Earth with respect to the aether. The interferometer that they constructed was exquisite, and through amazing experimental techniques the existence of the aether was disproved. The results of Michelson and Morley led to a revolution in physics, and provided evidence that helped Albert Einstein to develop the general theory of relativity. Now the Michelson interferometer may soon provide dramatic confirmation of Einstein's theory through the direct detection of gravitational radiation.

An accelerating electric charge produces electromagnetic radiation – light. It should come as no surprise that an accelerating mass produces gravitational light, namely gravitational radiation. In 1888 Heinrich Hertz had the luxury to produce and detect electromagnetic radiation in his laboratory. There will be no such luck with gravitational radiation because gravity is an extremely weak force.

Albert Einstein postulated the existence of gravitational radiation in 1916, and in 1989 Joe Taylor and Joel Weisberg indirectly confirmed its existence through observations of the orbital decay of the binary pulsar 1913 + 16. Direct detection of gravity waves will be difficult. For humans to have any chance of detecting gravity waves there must be extremely massive objects accelerating up to relativistic velocities. The only practical sources are astrophysical: supernovae, pulsars, neutron star or black hole binary systems, black hole formation or even the Big Bang. The observation of these types of events would be extremely significant for contributing to knowledge in astrophysics and cosmology. Gravity waves from the Big Bang would provide information about the Universe at its earliest moments. Observations of supernovae will yield a gravitational snapshot of these extreme cataclysmic events. Pulsars are neutron stars that can spin on their axes at frequencies in the hundreds of hertz, and the signals from these objects will help to decipher their characteristics. Gravity waves from the final stages of coalescing binary neutron stars could help to accurately determine the size of these objects and the equation of state of nuclear matter. The observation of black hole formation from these binary systems would be the *coup de grace* for the debate on the existence of black holes, and the ultimate triumph for general relativity.

Electromagnetic radiation has an electric field transverse to the direction of propagation, and a charged particle interacting with the radiation will experience a force. Similarly, gravitational radiation will produce a transverse force on massive objects, a tidal force. Explained via general relativity it is more accurate to say that gravitational radiation will deform the fabric of space-time. Just like electromagnetic radiation there are two polarizations for gravity waves. Let us imagine a linearly polarized gravity wave propagating in the z-direction, $h(z,t) = h_{0+}e^{i(kz-\omega t)}$. The fabric of space is stretched due to the strain created by the gravity wave. Consider a length L_0 of space along the x-axis. In the presence of the gravity wave the length oscillates like

$$L(t) = L_0 + \frac{h_{0+}L}{2}\cos\omega t$$

hence there is a change in its length of

$$\Delta L_x = \frac{h_{0+}L}{2}\cos\omega t$$

A similar length L_0 of the y-axis oscillates, like

$$L(t) = L_0 - \frac{h_{0+}L}{2}\cos\omega t$$

or

$$\Delta L_y = \frac{-h_{0+}L}{2}\cos\omega t$$

One axis stretches while the perpendicular one contracts, and then vice versa, as the wave propagates through. The other polarization ($h_{0\times}$) produces a strain on axes 45° from (x,y). Imagine some astrophysical event produces a gravity wave that has amplitude h_{0+} on Earth; in order to detect a small distance displacement ΔL one should have a detector that spans a large length L. A supernova within our own Galaxy might possibly produce a gravity wave of size $h \sim 10^{-18}$ with characteristic frequencies around

1 kHz, but the occurrence of an event such as this will be extremely rare. More events will come from novae in other galaxies, but due to the great distances and the fact that the magnitude of the gravity wave falls of as $1/r$, such events will be substantially diminished in magnitude.

A Michelson interferometer, with arms aligned along the x- and y- axes, can measure small phase differences between the light in the two arms. Therefore, this type of interferometer can turn the length variations of the arms produced by a gravity wave into changes in the interference pattern of the light exiting the system. This was the basis of the idea from which modern laser interferometric gravitational radiation detectors have evolved. Imagine a gravity wave of amplitude h is incident on an interferometer. The change in the arm length will be $h \sim \Delta L/L_0$, so in order to optimize the sensitivity it is advantageous to make the interferometer arm length L_0 as large as possible. Detectors coming on-line are attempting to measure distance displacements that are of order $\Delta L \sim 10^{-18}$ m or smaller, less that the size of an atomic nucleus! If the interferometers can detect such distance displacements, and directly detect gravitational radiation, it will be one of the most spectacular accomplishments in experimental physics. The first implementation of a laser interferometer to detect a gravity wave was by Forward who used earphones to listen to the motion of the interference signal. The engineering of signal extraction for modern interferometers is obviously far more complex.

Numerous collaborations are building and operating advanced interferometers in order to detect gravitational radiation. In the United States there is the Laser Interferometer Gravity Wave Observatory (LIGO), which consists of two 4-km interferometers located in Livingston, Louisiana and Hanford, Washington. In addition, there is an additional 2-km interferometer within the vacuum system at Hanford. An Italian–French collaboration (VIRGO) has a 3-km interferometer near Pisa, Italy. GEO, a German–British collaboration, is a 600-m detector near Hanover, Germany. TAMA is the Japanese 300-m interferometer in Tokyo. The Australians (AIGO) are constructing a 500-m interferometer in Western Australia. All of the detectors will be attempting to detect gravity waves with frequencies from about 50 Hz to 1 kHz.

As will be described below, there are a number of terrestrial noise sources that will inhibit the performance of the interferometric detectors. The sensitivity of detection increases linearly with interferometer arm length, which implies that there could be advantages to constructing a gravity wave detector in space. This is the goal of the Laser Interferometer Space Antenna (LISA) collaboration. The plan is to deploy three satellites in a heliocentric orbit with a separation of about 5×10^6 km. The launch for LISA is planned for around 2015, and the detector will be sensitive to gravity waves within the frequency range of 10^{-3} to 10^{-1} Hz. Due to the extremely long baseline, LISA is not strictly an interferometer, as most light will be lost as the laser beams expand while traveling such a great distance. Instead, the phase of the received light will be detected and used to lock the phase of the light that is re-emitted by another laser.

Joe Weber pioneered the field of gravitational radiation detection in the 1960s. He used a 1400-kg aluminum cylinder as an antenna. Gravitational waves would hopefully excite the lowest-order normal mode of the bar. Since the gravity wave is a strain on space-time, experimentalists win by making their detectors longer in length. This provides a long-term advantage for laser interferometers, which can scale in length relatively easily. However, as of 2002, interferometers and bars (now cooled to 4 K) have comparable sensitivities.

Interferometer Styles

The Michelson interferometer is the tool to be used to detect a gravitational wave. Figure 1 shows a basic system. The beamsplitter and the end mirrors would be suspended by wires, and effectively free to move in the plane of the interferometer. The arms have lengths L_1 and L_2 that are roughly equal on a kilometer scale. With a laser of power P and wavelength λ incident on the beamsplitter, the light

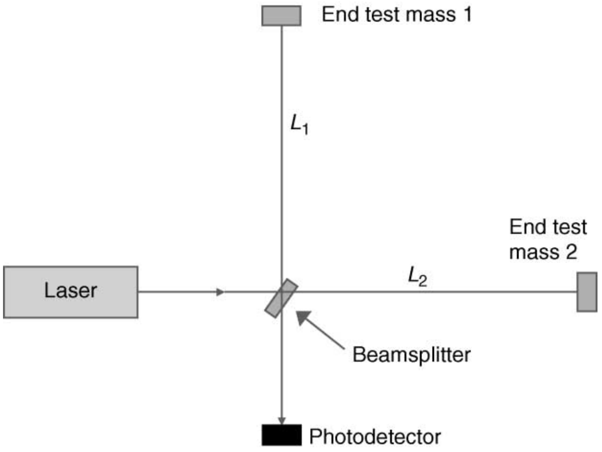

Figure 1 A basic Michelson interferometer. The photodetector receives light exiting the dark port of the interferometer and hence the signal.

exiting the dark port of the interferometer is

$$P_{\text{out}} = P \sin^2\left[\frac{2\pi}{\lambda}(L_1 - L_2)\right]$$

The interferometer operates with the condition that in the absence of excitation the light exiting the dark port is zero. This would be the case for a simple and basic interferometer. However, interferometers like those in LIGO will be more sophisticated, and will use a heterodyne detection strategy. If E_0 is the amplitude of the electric field from the laser, and assuming the use of a 50–50 beamsplitter, the electric field (neglecting unimportant common phase shifts) for the light incident on the photodetector would be

$$E_{\text{out}} = \frac{E_0}{2}\left(e^{i\delta\phi_1} - e^{i\delta\phi_2}\right) \approx i\frac{E_0}{2}(\phi_1 - \phi_2)$$
$$= iE_0\frac{2\pi}{L}(L_1 - L_2)$$

The laser light will be phase modulated at a frequency in the MHz regime. As such, the deconvolved current from the photodiode that detects light at the interferometer output will be proportional to the phase acquired by the light, namely $I \propto \frac{2\pi}{\lambda}(L_1 - L_2)$.

A gravity wave of optimal polarization normally incident upon the interferometer plane will cause one arm to decrease in length while the other increases. The Michelson interferometer acts as a gravity wave transducer; the change in arm lengths results in more light exiting the interferometer dark port. The mirrors in the interferometer are suspended via wires so that they are free to move under the influence of the gravity wave.

An interferometer's sensitivity increases with arm length, but geographical and financial constraints will limit the size of the arms. If there could be some way to bounce the light back and forth to increase the effective arm length it would increase the detector performance. Fabry–Perot cavities on resonance light have a storage time of $2L/[c(1 - \sqrt{R_1R_2})]$. Figure 2 shows the system of a Michelson interferometer with Fabry–Perot cavities. This gravity wave interferometer design was proposed and tested in the late 1970s by Ron Drever. The far mirror R_2 has a very high reflectivity ($R_2 \sim 1$) in order to ultimately direct the light back towards the beamsplitter. The front mirror reflectivity R_1 is such that LIGO's effective arm length increases to $L \sim 300$ km. The optical properties of the mirrors of the Fabry–Perot cavities must be exquisite in order to achieve success. LIGO's mirrors were tested, and the root mean squared surface uniformity is less than 1 nm, scattered light is less than 50 parts per million (ppm), absorption is less than 2 ppm, and the radii of curvature for the mirrors are matched to less than 3%. A LIGO test mass (and therefore a Fabry–Perot mirror) can be seen in Figure 3.

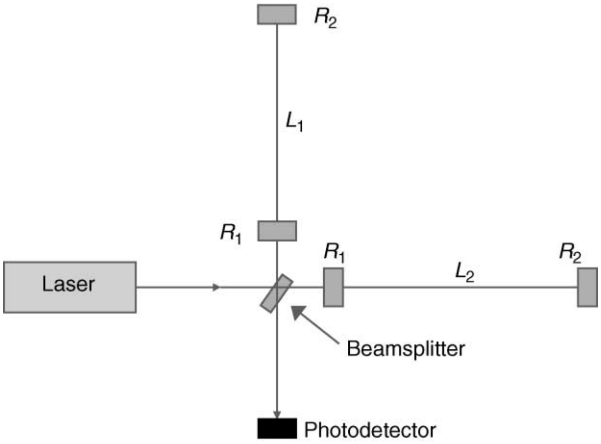

Figure 2 A Michelson interferometer with Fabry–Perot cavities in each arm. The front cavity mirrors have reflectivity R_1 while the end mirrors have $R_2 \sim 1$. By using Fabry–Perot cavities LIGO will increase the effective arm length by a factor of 75.

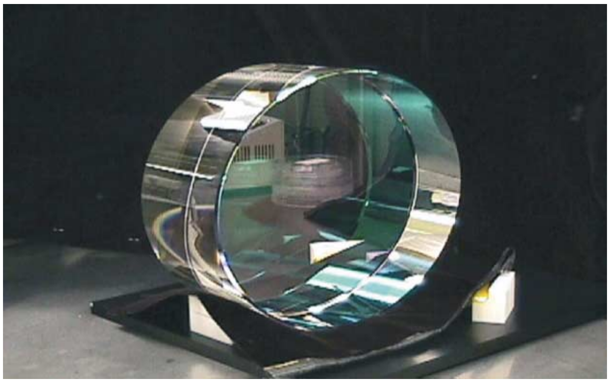

Figure 3 A picture of a mirror and test mass for LIGO. The fused silica component is 10.7 kg in mass and 25 cm in diameter. Photograph courtesy of Caltech/LIGO.

In 1888 Michelson and Morley, with their interferometer, had a sensitivity that allowed the measurement of 0.02 of a fringe, or about 0.126 radian. Prototype interferometers constructed by the LIGO science team have already demonstrated a phase noise spectral density of $\phi(f) = 10^{-10}$ radian/$\sqrt{\text{Hz}}$ for frequencies above 500 Hz. Assuming a 1 kHz signal with 1 kHz bandwidth this implies a phase sensitivity of $\Delta\phi = 3.2 \times 10^{-9}$ radian. This is about the phase sensitivity that LIGO hopes to accomplish in the 4-km Fabry–Perot system. There has been quite an evolution in interferometry since Michelson's time.

The noise sources that inhibit the interferometer performance are discussed below. However, let us

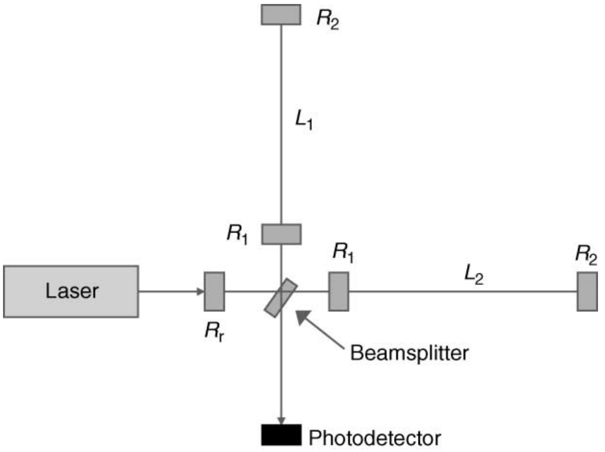

Figure 4 A power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light would exit the interferometer through the light port and head back to the laser. Installation of the recycling mirror with reflectivity R_r sends the light back into the system. A Fabry–Perot cavity is formed between the recycling mirror and the first mirror (R_1) of the arms. For LIGO this strategy will increase the power circulating in the interferometer by a factor of 50.

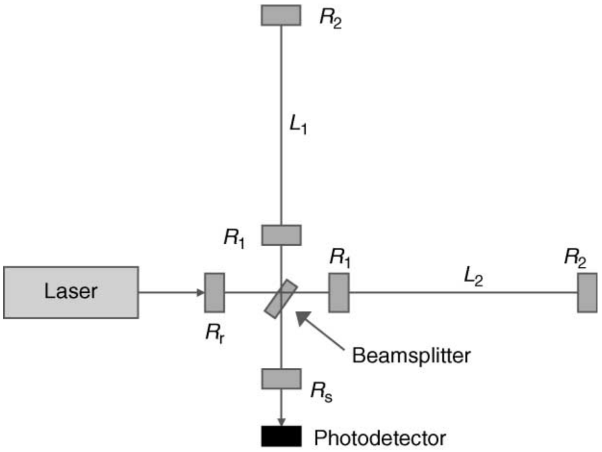

Figure 5 A signal recycled and power recycled Michelson interferometer with Fabry–Perot cavities in each arm. Normally light containing the gravity wave signal would exit the interferometer through the dark port and head to the photodetector. Installation of the signal recycling mirror with reflectivity R_s sends the light back into the system. The phase of the light acquired from the gravity wave will build up at a particular frequency determined by the reflectivity R_s.

consider one's ability to measure the relative phase between the light in the two arms. The Heisenberg uncertainty relation for light with phase ϕ and photon number N is $\Delta\phi\Delta N \sim 1$. For a measurement lasting time τ using laser power P and frequency f, the photon number is $N = P\lambda\tau/hc$, and with Poisson statistics describing the light $\Delta N = \sqrt{N} = \sqrt{P\lambda\tau/hc}$. Therefore

$$\Delta\phi\Delta N = \frac{2\pi}{\lambda}\Delta L\sqrt{P\lambda\tau/hc} = 1$$

implies that

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

With more light power the interferometer can measure smaller distance displacements and achieve better sensitivity. LIGO will use about 10 W of laser power, and will eventually work towards 100 W. However, there is a nice trick one can use to produce more light circulating in the interferometer, namely power recycling. Figure 4 displays the power recycling interferometer design. The interferometer operates such that virtually none of the light exits the interferometer dark port, and the bulk of the light returns towards the laser. An additional mirror, R_r in Figure 4, recycles the light. For LIGO, recycling will increase the effective light power by another factor of 50. The higher circulating light power therefore improves the sensitivity.

There is one additional modification to the interferometer system that can further improve sensitivity, but only at a particular frequency. A further Fabry–Perot system can be made by installing what is called a signal recycling mirror; this would be mirror R_s in Figure 5. Imagine light in arm 1 on the interferometer that acquires phase as the arm expands due to the gravity wave. The traveling gravity wave's oscillation will subsequently cause arm 1 to contract while arm 2 expands. If the light that was in arm 1 could be sent to arm 2 while it is expanding, then the beam would acquire additional phase. This process could be repeated over and over. Mirror R_s serves this purpose, with its reflectivity defining the storage time for light in each interferometer arm. The storage time defined by the cavity formed by the signal recycling mirror, R_s, and the mirror at the front of the interferometer arm cavity, R_1, determines the resonance frequency. Signal recycling will give a substantial boost to interferometer sensitivity at a particular frequency, and will eventually be implemented in all the main ground-based interferometric detectors.

The LIGO interferometers are infinitely more complex than the relatively simple systems displayed in the figures of this paper. Figure 6 presents an aerial view of the LIGO site at Hanford, Washington. The magnitude of the 4-km system is apparent.

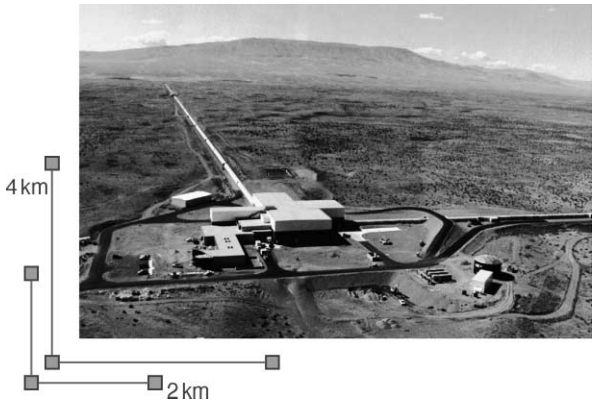

Figure 6 Arial view of the LIGO Hanford, Washington site. The vacuum enclosure at Hanford contains both 2-km and 4-km interferometers. Photograph courtesy of Caltech/ LIGO.

Noise Sources and Interferometer Sensitivity

If the interferometers are to detect distance displacements of order $\Delta L \sim 10^{-18}$ m then they must be isolated from a host of deleterious noise sources. Seismic disturbances should not shake the interferometers. Thermal excitation of components will affect the sensitivity of the detector and should be minimized. The entire interferometer must be in an adequate vacuum in order to avoid fluctuations in gas density that would cause changes in the index of refraction and hence modification of the optical path length. The laser intensity and frequency noise must be minimized. The counting statistics of photons influences accuracy. If ever there was a detector that must avoid Murphy's law this is it; little things going wrong cannot be permitted if such small distance displacements are to be detected. The expected noise sensitivity for the initial LIGO interferometers is displayed in Figure 7.

In the best of all worlds the interferometer sensitivity will be limited by the counting statistics of the photons. A proper functioning laser will have its photon number described by Poisson statistics, or shot noise; if the mean number of photons arriving per unit time is N then the uncertainty is $\Delta N = \sqrt{N}$, which as noted above implies an interferometer displacement sensitivity of

$$\Delta L = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P\tau}}$$

or a noise spectral density of

$$\Delta L(f) = \frac{1}{2\pi}\sqrt{\frac{hc\lambda}{P}} \quad \text{(in units of m}/\sqrt{\text{Hz}})$$

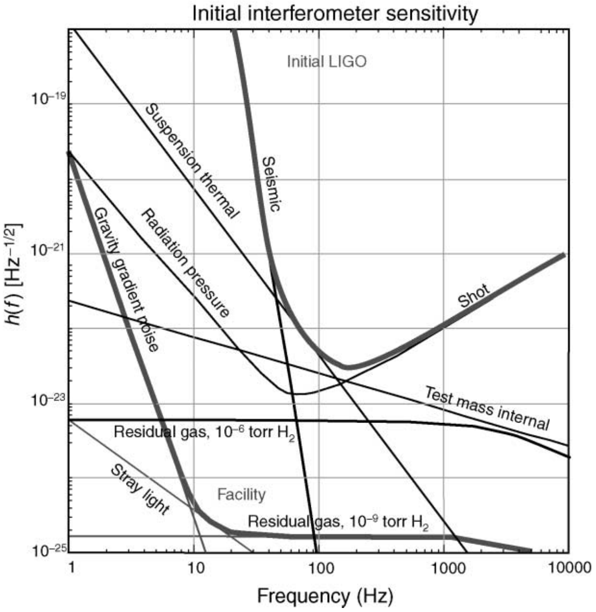

Figure 7 The target spectral density of the noise for the initial LIGO system. LIGO will be dominated by seismic noise at low frequencies (10–100 Hz), thermal noise (from the suspension system and internal modes within the mirrors) in the intermediate regime (100–300 Hz), and photon shot noise thereafter. Other sources of noise are also noted, specifically gravity gradients (gravitational excitation of masses from the seismic motion of the ground), radiation pressure of photons on the mirrors, stray (scattered) light, and index of refraction fluctuations from residual gas in the vacuum.

Note also that the sensitivity increases as the light power increases. The reason for this derives from the statistics of repeated measurements. The relative lengths of the interferometer arm lengths could be measured, once, by a photon. However, the relative positions are measured repeatedly with every photon from the laser, and the variance of the mean decreases as $\sqrt{N}$, where N is the number of measurements (or photons) involved. The uncertainty in the difference of the interferometer arm lengths is therefore inversely proportional to photon number, and hence the laser's power. In terms of strain sensitivity this would imply

$$h(f) = \frac{1}{2\pi L}\sqrt{\frac{hc\lambda}{P}} \quad \text{(in units of } 1/\sqrt{\text{Hz}})$$

This assumes the light just travels down the arm and back once. With Fabry–Perot cavities the light is stored, and the typical photon takes many trips back and forth before exiting the system. In order to maximize light power the end mirrors ($R_2 \sim 1$) and the strain sensitivity is improved to

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}$$

where $L' = 4L/(1 - R_1)$.

As the frequency of gravity waves increases the detection sensitivity will decrease. If the gravity wave causes the interferometer arm length to increase, then decrease, while the photons are still in the arm cavity, then the phase acquired from the gravity wave will be washed away. This is the reason why interferometer sensitivity decreases as frequency increases, and explains the high-frequency behavior seen in Figure 7. Taking this into account, the strain sensitivity is

$$h(f) = \frac{1}{2\pi L'}\sqrt{\frac{hc\lambda}{P}}\left(1+\left(\frac{2\pi L' f}{c}\right)^2\right)^{1/2}$$

where $L' = 4L/(1 - R_1)$ and f is the frequency of the gravity wave.

If the gravitational wave is to change the interferometer arm length then the mirrors that define the arm must be free to move. In systems like LIGO wires suspend the mirrors; each mirror is like a pendulum. While allowing the mirrors to move under the influence of the gravity wave is a necessary condition, the pendulum itself is the first component of an elaborate vibration isolation system. Seismic noise will be troublesome for the detector at low frequencies. The spectral density of the seismic noise is about $(10^{-7}/f^2)\mathrm{m}/\sqrt{\mathrm{Hz}}$ for frequencies above 1 Hz. A simple pendulum, by itself, acts as a motion filtering device. Above its resonance frequency the pendulum filters motion with a transfer function like $T(f) \propto (f_0/f)^2$. Detectors such as LIGO will have a pendulum with resonant frequencies of about $f_0 \approx 1$ Hz, thus providing an isolation of 10^4 when looking for signals at $f = 100$ Hz. The various gravity wave detector collaborations have different vibration isolation designs. The mirrors in these interferometers will be suspended in elaborate vibration isolation systems, which may include multiple pendulum and isolation stacks. Seismic noise will be the limiting factor for interferometers seeking to detect gravity waves in the tens of hertz range, as can be seen in the sensitivity curve presented in Figure 7.

Due to the extremely small distance displacements that these systems are trying to detect it should come as no surprise that thermal noise is a problem. This noise enters through a number of components in the system. The two most serious thermal noise sources are the wires suspending the mirrors in the pendulum, and the mirrors themselves. Consider the wires; there are a number of modes that can oscillate (i.e. violin modes). At temperature T each mode will have energy of $k_B T$, but distributed over a band of frequencies determined by the quality factor (or Q) of the material. Low-loss (or high-Q) materials work best; for the violin modes of the wires there will be much noise at particular frequencies (in the hundreds of hertz). The mirror is a cylindrical object, which will have normal modes of oscillation that can be thermally excited. The first generation of LIGO will have these masses composed of fused silica, which is typical for optical components. The Qs for the internal modes are greater than 2×10^6. A switch may eventually be made to sapphire mirrors, which have better thermal properties. The limitation to the interferometers' sensitivity due to the thermal noise internal to the mirrors can be seen in Figure 7, and will be a worrying noise source for the first-generation LIGO in the frequency band around 100–400 Hz.

The frequency noise of the laser can couple into the system to produce length displacement sensitivity noise in the interferometer. With arm lengths of 4 km, it will be impossible to hold the length of the two arms absolutely equal. The slightly differing arm spans will mean that the light sent back from each of the two Fabry–Perot cavities will have slightly differing phases. As a consequence, great effort is made to stabilize the frequency of the light entering the interferometer. The LIGO laser can be seen in Figure 8. The primary laser is a Lightwave Model 126 nonplanar ring oscillator. High power is generated from this stabilized laser through the use of optical amplifiers. The beam is sent through four optical amplifiers, and then retro-reflected back

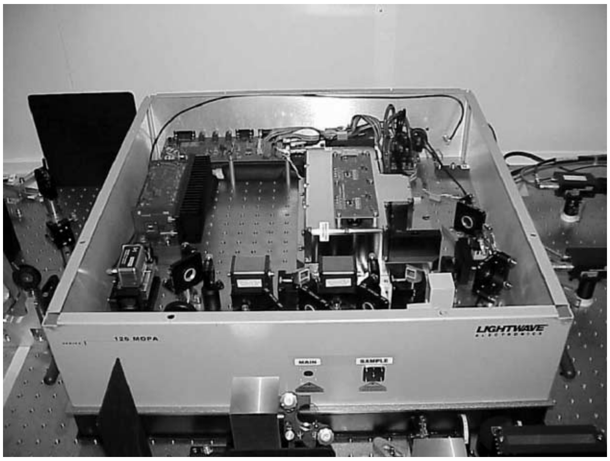

Figure 8 A picture of the Nd:YAG laser and amplifier system that produces 10 W of light for LIGO. Photograph courtesy of Caltech/LIGO.

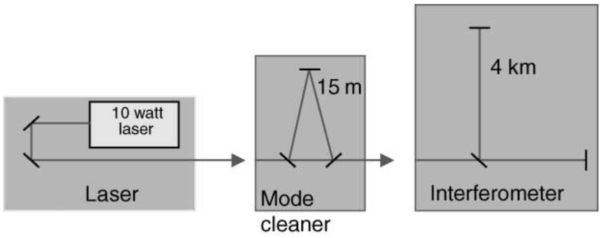

Figure 9 The optical system for the LIGO system. Light moves from the frequency-stabilized laser to a mode cleaner, and then to the interferometer. Presently 8 W of TEM_{00} laser power is delivered to the interferometer.

through the amplifiers again. For LIGO, the laser is locked and held to a specific frequency by use of signals from a reference cavity, a mode cleaner cavity, and the interferometer. For low-frequency stabilization the temperature of the ring oscillator is adjusted. At intermediate frequencies adjustment is made by signals to a piezo-electric transducer within the ring oscillator cavity. At high frequencies the noise is reduced with the use of an electro-optic crystal. The LIGO lasers currently have a frequency noise of $2 \times 10^{-2}\,\mathrm{Hz}/\sqrt{\mathrm{Hz}}$ at frequencies above 1 kHz.

It will prove important to worry about the stability of the laser power for the interferometric detectors. The hope is to be shot noise limited at frequencies above a few hundred hertz. The Nd:YAG power amplifiers used are pumped with an array of laser diodes, so the light power is controlled through feedback to the laser diodes. The LIGO requirements for the fluctuations on the power P are $\Delta P/P < 10^{-8}/\sqrt{\mathrm{Hz}}$. The spatial quality of the light is ensured through the use of a mode-cleaning cavity. LIGO uses a triangular array of mirrors separated by 15 m. LIGO's input optical system can be seen in Figure 9. The current LIGO optical system yields 8 W of 1.06 μm light in the TEM_{00} mode.

Conclusion

The attempt to measure gravitational radiation with laser interferometers could possibly be the most difficult optical experiment of our time. Over a hundred years ago Michelson succeeded in carrying off experiments of amazing difficulty as he measured the speed of light and disproved the existence of the aether. Gravity wave detection is an experiment worthy of Michelson, and there are hundreds of physicists striving to make it a reality.

Great success has already been achieved. The TAMA 300-m interferometer is operational and has achieved a sensitivity $h(f) = 5 \times 10^{-21}/\sqrt{\mathrm{Hz}}$ in the 700 Hz to 1.5 kHz frequency band. The LIGO interferometers are now operating, and scientists are presently de-bugging the system in order to achieve the target sensitivity. The first scientific data taking for LIGO commenced in 2002, and is continuing at the time of writing.

This will be a new telescope to peer into the heavens. With every new means of looking at the sky there has come unexpected discoveries. Physicists do know that there will be signals that they can predict. Binary systems of compact objects (neutron stars or black holes) will produce chirp signals that may be extracted by matched filtering techniques. A supernova will produce a burst that will hopefully rise above the noise. Pulsars, or neutron stars spinning about their axes at rates sometimes exceeding hundreds of revolutions per second, will produce continuous sinusoidal signals that can be seen by integrating for sufficient lengths of time. Gravity waves produced by the Big Bang will produce a background stochastic noise that can possibly be extracted by correlating the outputs from two or more detectors. These are exciting physics results that will come through tremendous experimental effort.

See also

Imaging: Interferometric Imaging. **Interferometry:** Overview.

Further Reading

Barish BC (1997) Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 155–161. Tokyo: Universal Academic.

Blair DG, Taniwaki M, Zhao CN, *et al.* (1997) Progress in the development of technology for advanced laser interferometric gravitational wave detectors. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 75–94. Tokyo: Universal Academic.

Brillet A (1997) Virgo status report. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 163–173. Tokyo: Universal Academic.

Danzmann K (1997) LISA–A gravitational wave observatory in heliocentric orbit. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 111–116. Tokyo: Universal Academic.

Drever RWP, Ford GM, Hough J, *et al.* (1980) A gravity-wave detector using optical cavity sensing. In: Schmutzer E (ed.) *Proceedings of the Ninth International Conference on General Relativity and Gravitation*, pp. 265–267. Berlin: Springer.

Forward RL (1978) Wideband laser-interferometer gravitational-radiation experiment. *Physical Review* D17: 379.

Hough J and the GEO 600 Team (1997) GEO 600 Current status and some aspects of the design. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 175–182. Tokyo: Universal Academic.

Robertson NA (2000) Laser interferometric gravitational wave detectors. *Classical Quantum Gravity* 17: R19–R40.

Saulson PR (1994) Fundamentals of Interferometric Gravitational Wave Detectors. Singapore: World Scientific.

Taylor JH and Weisberg JM (1989) Further experimental tests of relativistic gravity using the binary pulsor PSR 1913 + 16. *Astrophysical Journal* 345: 434–450.

Thorne KS (1987) Gravitational radiation. In: Hawking S and Isreal W (eds) *300 Years of Gravitation*, pp. 330–458. Cambridge, UK: Cambridge University Press.

Tsubono K and the TAMA collaboration (1997) TAMA Project, Gravitational wave detection. In: Tsubono K, Fujimoto MK and Kurodo K (eds) *Gravitational Wave Detection*, pp. 183–191. Tokyo: Universal Academic.

Weiss R (1972) Electromagnetically coupled broadband gravitational antenna. *MIT Quarterly Progress Report* no. 105.

Phase-Measurement Interferometry

K Creath, Optineering and University of Arizona, Tucson, AZ, USA

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase-measurement interferometry is a way to measure information encoded in interference patterns generated by an interferometer. Fringe patterns (fringes) created by interfering beams of light are analyzed to extract quantitative information about an object or phenomenon. These fringes are localized somewhere in space and require a certain degree of spatial and temporal coherence of the source to be visible. Before these techniques were developed in the late 1970s and early 1980s, fringe analysis was done either by estimating fringe deviation and irregularity by eye or by manually digitizing the centers of interference fringes using a graphics tablet. Digital cameras and desktop computers have made it possible to easily obtain quantitative information from fringe patterns. The techniques described in this article are independent of the type of interferometer used.

Interferometric techniques using fringe analysis can measure features as small as a micron wide or as large as a few meters. Measurable height ranges vary from a few nanometers up to 10s of mm, depending upon the interferometric technique employed. Measurement repeatability is very consistent, and typically repeatability of 1/100 rms of a fringe is easily obtainable while 1/1000 rms is possible. Actual measurement precision depends on what is being measured and how the technique is implemented, while accuracy depends upon comparison to a sanctified standard.

There are many different types of applications for fringe analysis. For example, optical surface quality can be determined using a Twyman–Green or Fizeau interferometer. In addition, wavefront quality measurements of a source or an optical system can be made in transmission, and the index of refraction and homogeneity of optical materials can be mapped out. Many nonoptical surfaces can also be measured. Typically, surface topography information at some specific spatial frequency scale is extracted. These measurements are limited by the resolution of the optical system and the field of view of the imaging system. Lateral and vertical dimensions can also be measured. Applications of nonoptical surfaces include disk and wafer flatness, roughness measurement, distance and range sensing. Phase measurement techniques used in holographic interferometry, TV holography, speckle interferometry, moiré, grating interferometry, and fringe projection are used for nondestructive testing to measure surface structure as well as displacements due to stress and vibration. This article outlines the basics of phase measurement interferometry (PMI) techniques as well as the types of algorithms used. The bibliography lists a number of references for further reading.

Background

Basic Parts of a Phase-Measuring Interferometer

A phase-measuring interferometer consists of a light source, an illumination system (providing uniform illumination across the test surface), a beamsplitter (usually a cube or pellicle so both beams have the same optical path), a reference surface (needs to be good because these techniques measure the difference between the reference and test surface), a sample fixture, an imaging system (images a plane in space where the surface is located onto the camera), a camera (usually a monochrome CCD), an image digitizer/frame grabber, a computer system, software (to control the measurement process and calculate the surface map), and often a spatial or temporal phase shifter to generate multiple interferograms.

Steps of the Measurement Process

To generate a phase map, a sample is placed on a sample fixture and aligned, illumination levels are adjusted, the sample image is focused onto the camera, the fringes are adjusted for maximum contrast, the phase shifter or fringe spacing is adjusted or calibrated as necessary, a number of images is obtained and stored with the appropriate phase differences, the optical path difference (OPD) is calculated as the modulo 2π phase and then is unwrapped at each pixel to determine the phase map.

To make consistent measurements, some interferometers need to be on vibration isolation systems and away from heavy airflows or possible acoustical coupling. It helps to cover any air paths that are longer than a few millimeters. Consideration needs to be made for consistency of temperature and humidity.

The human operating the interferometer is also a factor in the measurement. Does this person always follow the same procedure? Are the measurements sampled consistently? Are the test surfaces clean? Is the sample aligned the same and correctly? Many different factors can affect a measurement. To obtain repeatable measurements it is important to have a consistent procedure and regularly verify measurement consistency.

Common Interferometer Types

One of the most common interferometers used in optical testing is the Twyman–Green interferometer (Figure 1). Typically, a computer controls a mirror pushed by a piezo-electric transducer (PZT). The test surface is imaged onto the camera and the computer has a frame grabber that takes frames of fringe data.

The most common commercially available interferometer for optical testing is the Fizeau interferometer (Figure 2). This versatile instrument is very insensitive to vibrations due to the large common path for both interfering wavefronts. The reference surface (facing the test object) is moved by a PZT to provide the phase shift.

Interference microscopes (*see* **Microscopy:** Interference Microscopy. **Interferometry:** White Light Interferometry) are used for looking at surface roughness and small structures (see Figure 3). These instruments can employ Michelson, Mirau, and Linnik interference objectives with a laser or white light source, or Fizeau-type interference objectives that typically use a laser source because of the unequal paths in the arms of the interferometer. The phase shift is accomplished by moving the sample, the reference surface, or parts of the objective relative to the sample. Figure 3 shows a schematic of a Mirau-type interferometric microscope for phase measurement.

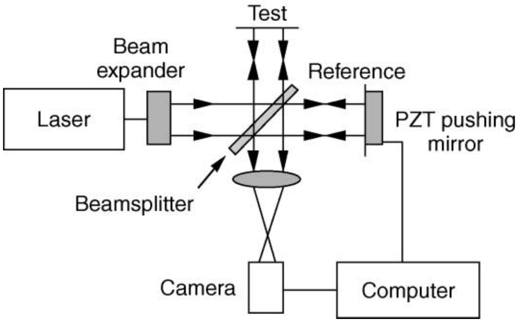

Figure 1 Twyman–Green interferometer.

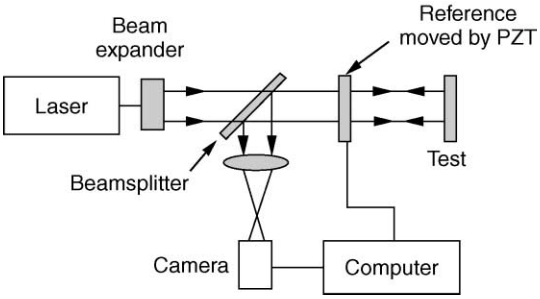

Figure 2 Fizeau interferometer.

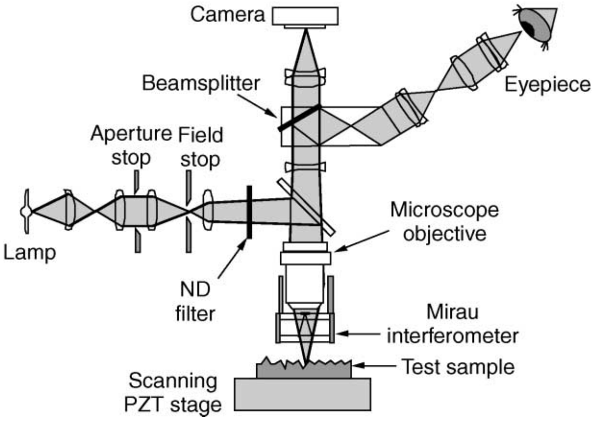

Figure 3 Mirau interference microscope.

Determination of Phase

The Interference Equation

Interference fringes from a coherent source (e.g., a laser) are theoretically sinusoidal (Figure 4a), while fringes from an incoherent source (e.g., white light) are localized in a wavepacket at a point in space (Figure 4b) where the optical paths of the arms of the interferometer are equal and marked on Figure 4 as scanner positions $z = 0$ (*see* **Interferometry:** White Light Interferometry). For generality, this analysis considers the determination of phase within the wave packet for fringes localized in space.

Interference fringes at any point in the wavepacket can be written in the following form:

$$I(x, y) = I_0\{1 + \gamma(x, y, z)\cos[\phi(x, y)]\} \qquad [1]$$

where I_0 is the dc irradiance, γ is the fringe visibility (or contrast), $2I_0\gamma$ is the modulation (irradiance amplitude or the ac part of the signal), and ϕ is the phase of the wavefront as shown in Figure 5. For simplicity, this drawing assumes that the interference fringe amplitude is constant.

Fringe visibility can be determined by calculating

$$\gamma = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \qquad [2]$$

where $I_{\max}$ is the maximum value of the irradiance for all phase values and $I_{\min}$ is the minimum value. The fringe visibility has a real value between 0 and 1 and varies with the position along the wavepacket. The fringe visibility, as defined here, is the real part of the complex degree of coherence.

Types of Phase Measurement Techniques

Phase can be determined from either a number of interference fringe patterns or from a single interferogram with appropriate fringe spacing. Temporal techniques require an applied phase shift between the test and reference beams as a function of time while multiple frames of interference fringe data are obtained. Spatial techniques can obtain data from a single interferogram that requires a carrier pattern of almost straight fringes to either compare phases of adjacent pixels or to separate orders while performing operations in the Fourier domain. Spatial techniques may also simultaneously record multiple interferograms with appropriate relative phase shift differences separated spatially in space. Multiple frame techniques require more data than single frame techniques. Temporal techniques require that the interference fringes be stable over the time period it takes to acquire the number of images. While single-frame spatial techniques require less data and can be done with a single image, they generally have a reduced amount of resolution and less precision than temporal techniques.

There are literally hundreds of algorithms and techniques for extracting phase data from interference fringe data. The references listed in the Further Reading offer more details of these techniques.

Temporal Phase Measurement

Temporal techniques use data taken as the relative phase between the test and reference beams is modulated (shifted). The phase (or OPD) is calculated at each measured point in the interferogram. As the phase shifter is moved, the phase at a single point in the interferogram changes. The effect looks like the fringes are moving across the interferogram, and because of these techniques are sometimes called fringe scanning or fringe shifting techniques. However, the fringes are not really moving; rather the irradiance at a single detector point is changing (hopefully sinusoidally) in time (see Figure 6). A 180° or π phase causes a bright fringe to become a dark fringe.

Phase Modulation Techniques

There are many ways to introduce a phase modulation (or shift). These include moving a mirror or the

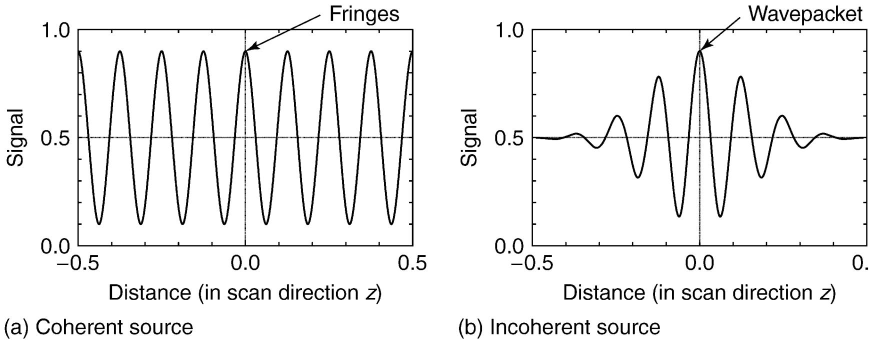

(a) Coherent source

(b) Incoherent source

Figure 4 Interference fringes for coherent and incoherent sources as observed at point $x, y, z = 0$ corresponds to equal optical path lengths of the reference and object beams.

sample, tilting a glass plate, moving a diffraction grating, rotating a polarizer, analyzer, or half-wave plate, using a two-frequency (Zeeman) laser source, modulating the source wavelength, and switching an acousto-optic Bragg cell, or magneto-optic/electro-optic cell. While any of these techniques can be used for coherent sources, special considerations need to be made for temporally incoherent sources (*see* **Interferometry:** White Light Interferometry). Figure 7 shows how moving a mirror introduces a relative phase shift between object and reference beams in a Twyman–Green interferometer.

Extracting Phase Information

Including the phase shift, the interference equation is written as

$$I(x,y) = I_0(x,y)\{1 + \gamma_0(x,y)\cos[\phi(x,y) + \alpha(t)]\} \quad [3]$$

where $I(x,y)$ is the irradiance at a single detector point, $I_0(x,y)$ is the average (dc) irradiance, $\gamma_0(x,y)$ is the fringe visibility before detection, $\phi(x,y)$ is the phase of the wavefront being measured, and $\alpha(t)$ is the phase shift as a function of time.

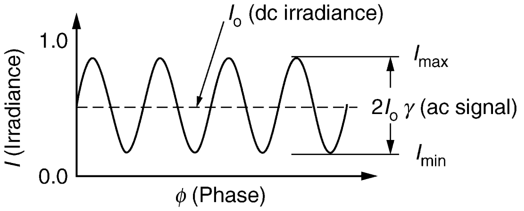

Figure 5 Variable definitions for interference fringes.

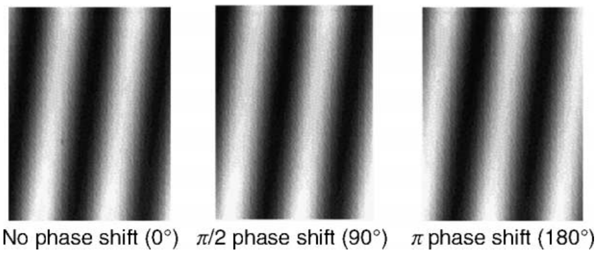
No phase shift (0°) $\pi/2$ phase shift (90°) π phase shift (180°)

Figure 6 Interference fringe patterns corresponding to different relative phase shifts between test and reference beams.

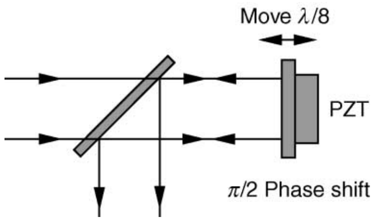

Figure 7 Moving a mirror to shift relative phase by $\lambda/4$.

Since the detector has to integrate for some finite time, the detected irradiance at a single point becomes an integral over the integration time Δ (Figure 8) where the average phase shift for the ith frame of data is α_i:

$$I_i(x,y) = \frac{1}{\Delta}\int_{\alpha_i-(\Delta/2)}^{\alpha_i+(\Delta/2)} I_0(x,y)\{1 + \gamma_0(x,y) \times \cos[\phi(x,y) + \alpha(t)]\}\mathrm{d}\alpha(t) \quad [4]$$

After integrating over $\alpha(t)$, the irradiance of the detected signal becomes

$$I_i(x,y) = I_0(x,y)\left\{1 + \gamma_0(x,y)\operatorname{sinc}\left(\frac{\Delta}{2}\right) \times \cos[\phi(x,y) + \alpha_i]\right\} \quad [5]$$

where $\operatorname{sinc}(\Delta/2) = \sin(\Delta/2)/(\Delta/2)$, which reduces the detected visibility.

Ramping Versus Stepping

There are two different ways of shifting the phase; either the phase shift can be changed in a constant and linear fashion (ramping) (see Figure 8) or it can be stepped in increments. Ramping provides a continuous smooth motion without any jerking motion. This option may be preferred if the motion does not wash out the interference fringes. However, ramping requires good synchronization of the camera/digitizer and the modulator to get the correct OPD (or phase) changes between data frames. Ramping allows faster data taking but requires electronics that are more sophisticated. In addition, it takes a finite time for a mass to move linearly. When ramping, the first frame or two of data usually needs to be discarded because the shift is not correct until the movement is linear.

The major difference between ramping and stepping the phase shift is a reduction in the modulation of the interference fringes after detection (the sinc($\Delta/2$) term in eqn [5]). When the phase shift is stepped ($\Delta = 0$), the sinc term has a value of one. When the phase shift is ramped ($\Delta = \alpha$) for a phase

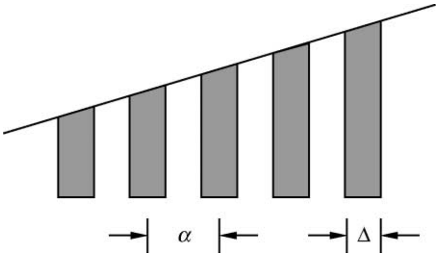

Figure 8 Change in OPD integrating over a Δ phase shift with sample separation α.

shift of $\alpha = 90° = \pi/2$, this term has a value of 0.9. Therefore, ramping slightly reduces the detected fringe visibility.

Signal Modulation

Phase measurement techniques assume that the irradiance of the interference fringes at the camera covers as much of the detector's dynamic range as possible and that the phase of the interference fringe pattern is modulating at each individual pixel as the phase between beams is modulated. If the irradiance at a single detector point does not modulate as the relative phase between beams is shifted, the height of the surface cannot be calculated. Besides the obvious reductions in irradiance modulation due to the detector sampling and pixel size or a bad detector element, scattered light within the interferometer, and defects or dirt on the test object, can also reduce signal modulation. Phase measurement techniques are designed to take into account the modulation of the signal at each pixel. If the signal does not modulate enough at a given pixel, then the data at that pixel are considered unusable, flagged as 'bad' and are often left blank. Phase values for these points may be interpolated from surrounding pixels if there are sufficient data.

Three-Frame Technique

The simplest method to determine phase uses three frames of data. With three unknowns, three sets of recorded fringe data are needed to reconstruct a wavefront providing a phase map. Using phase shifts of $\alpha_i = \pi/4$, $3\pi/4$, and $5\pi/4$, the three fringe measurements at a single point in the interferogram may be expressed as

$$\begin{aligned} I_1 &= I_0\left[1+\gamma\cos\left(\phi+\frac{\pi}{2}\right)\right] \\ &= I_0\left[1+\frac{\sqrt{2}}{2}\gamma(\cos\phi-\sin\phi)\right] \end{aligned} \qquad [6]$$

$$\begin{aligned} I_2 &= I_0\left[1+\gamma\cos\left(\phi+\frac{3\pi}{2}\right)\right] \\ &= I_0\left[1+\frac{\sqrt{2}}{2}\gamma(-\cos\phi-\sin\phi)\right] \end{aligned} \qquad [7]$$

$$\begin{aligned} I_3 &= I_0\left[1+\gamma\cos\left(\phi+\frac{5\pi}{2}\right)\right] \\ &= I_0\left[1+\frac{\sqrt{2}}{2}\gamma(-\cos\phi+\sin\phi)\right] \end{aligned} \qquad [8]$$

Note that (x, y) dependencies are still implied. The choice of the specific phase shift values is to make the math simpler. The phase at each detector point is

$$\phi = \tan^{-1}\left(\frac{I_3 - I_2}{I_1 - I_2}\right) \qquad [9]$$

In most fringe analysis techniques, we are basically trying to solve for terms such that we end up with the tangent function of the phase. The numerator and denominator shown above are respectively proportional to the sine and cosine of the phase. Note that the dc irradiance and fringe visibility appear in both the numerator and denominator. This means that variations in fringe visibility and average irradiance from pixel to pixel do not affect the results. As long as the fringe visibility and average irradiance at a single pixel is constant from frame to frame, the results will be good. If the different phase shifts are multiplexed onto multiple cameras, the results will be dependent upon the gain of corresponding pixels.

Bad data points with low signal modulation are determined by calculating the fringe visibility at each data point using:

$$\gamma = \frac{\sqrt{(I_3 - I_2)^2 + (I_1 - I_2)^2}}{\sqrt{2}I_0} \qquad [10]$$

It is simpler to calculate the ac signal modulation $(2I_0\gamma)$ and then set the threshold on the modulation to a typical value of about 5–10% of the dynamic range. If the modulation is less than this value at any given data point, the data point is flagged as bad. Bad points are usually caused by noisy pixels and can be due to scratches, pits, dust and scattered light.

Although three frames of data are enough to determine the phase, this and other 3-frame algorithms are very sensitive to systematic errors due to nonsinusoidal fringes or nonlinear detection, phase shifter miscalibration, vibrations and noise. In general, the larger the number of frames of data used to determine the phase, the smaller the systematic errors.

Phase Unwrapping

The removal of phase ambiguities is generally called phase unwrapping, and is sometimes known as integrating the phase. The phase ambiguities owing to the modulo 2π arctangent calculation can simply be removed by comparing the phase difference between adjacent pixels. When the phase difference between adjacent pixels is greater than π, a multiple of 2π is added or subtracted to make the difference less than π. For the reliable removal of

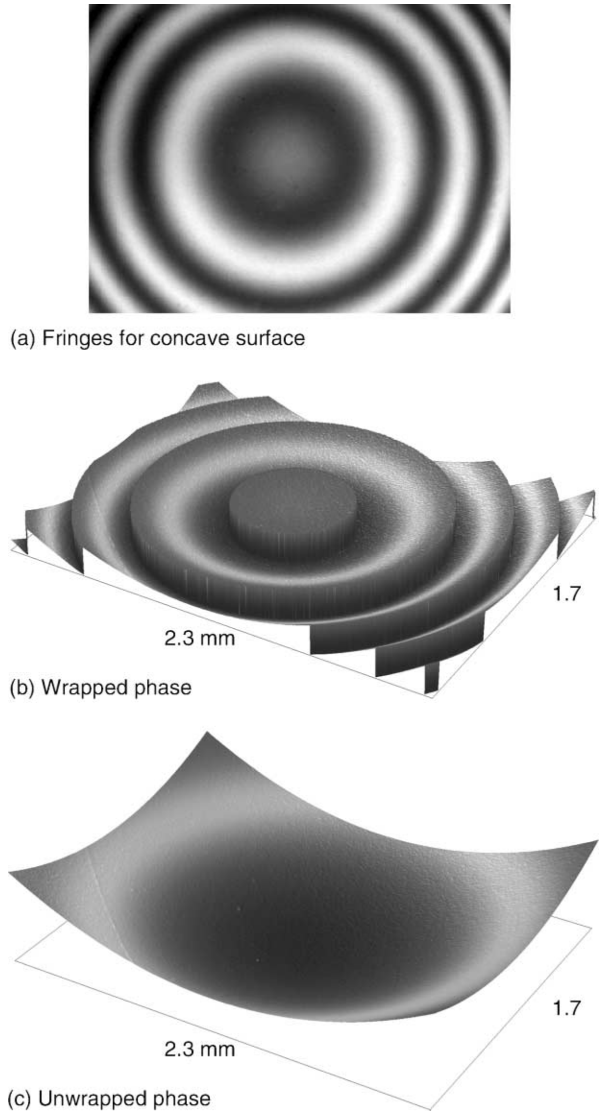

Figure 9 Fringes, wrapped, and unwrapped phase maps.

discontinuities, the phase must not change by more than π ($\lambda/2$ in optical path (OPD)) between adjacent pixels. Figure 9 shows an example of wrapped and unwrapped phase values.

Given a trouble-free wrapped phase map, it is enough to search row by row (or column by column) for phase differences of more than π between neighboring pixels. However, fringe patterns usually are not perfect and are affected by systematic errors. Some of the most frequently occurring error sources are noise, discontinuities in the phase map, violation of the sampling theorem and invalid data points (e.g., due to holes in the object or low modulation regions). Some phase maps are not easily unwrapped. In these cases different techniques like wave packet peak sensing in white light interferometry are used.

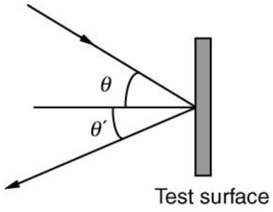

Figure 10 Angles of illumination and viewing.

From Wavefront to Surface

Once the phase of the wavefront is known, surface shape can be determined from the phase map. Surface height H of the test surface relative to the reference surface at a location (x, y) is given by

$$H(x,y) = \frac{\phi(x,y)\lambda}{2\pi(\cos\theta + \cos\theta')} \quad [11]$$

where λ is the wavelength of illumination, and θ and θ' – the angles of illumination and viewing with respect to the surface normal – are shown in Figure 10. For interferometers (e.g., Twyman–Green or Fizeau) where the illumination and viewing angles are normal to the surface ($\theta = \theta' = 0$), the surface height is simply:

$$H(x,y) = \frac{\lambda}{4\pi}\phi(x,y) \quad [12]$$

Since the wavefront measured represents the relative difference between the interfering reference and test wavefronts, this phase map only directly corresponds to the surface under test when the reference wavefront is perfectly flat. In practice, the shape of the reference surface needs to be accounted for by measuring it using a known test surface and subtracting this reference measurement from subsequent measurements of the test surface.

Phase Change on Reflection

Phase shifting interferometry measures the phase of reflected light to determine the shape of objects. The reflected wavefront will represent the object surface (within a scaling factor) if the object is made of a single material. If the object is comprised of multiple materials that exhibit different phase changes on reflection, the measured wavefront needs to be corrected for these phase differences (*see* **Interferometry:** White Light Interferometry).

Overview of Phase Measurement Algorithms and Techniques

There are literally hundreds of published algorithms and techniques. The optimal algorithm depends on the application. Most users prefer fast algorithms using a minimal amount of data that are as accurate and repeatable as possible, immune to noise, adaptable, and easy to implement. In practice, there are trade-offs that must be considered when choosing a specific algorithm or technique. This section provides an overview of the types of algorithms to aid the reader in sifting through published algorithms.

Synchronous Detection

One of the first techniques for temporal phase measurement utilized methods of communication theory to perform synchronous detection. To synchronously detect the phase of a noisy sinusoidal signal, the signal is first correlated (or multiplied) with sinusoidal and cosinusoidal reference signals (signals in quadrature) of the same frequency and then averaged over many periods of oscillation. This method of synchronous detection as applied to phase measurement can be extracted from the least squares estimation result when the phase shifts are chosen such that N measurements are equally spaced over one modulation period. With phase shifts α_i such that:

$$\alpha_i = \frac{2\pi}{N} i, \qquad \text{with } i = 1, \ldots, N \qquad [13]$$

the phase can be calculated from

$$\phi(x, y) = \tan^{-1}\left[\frac{\sum I_i(x, y)\sin\alpha_i}{\sum I_i(x, y)\cos\alpha_i}\right] \qquad [14]$$

Note that N can be any number of frames (or samples). The more frames of data, the smaller the systematic errors. This technique does not take large amounts of memory for a large number of frames, because only the running sums of the fringes multiplied by the sine and cosine of the phase shift need to be remembered. The 4-frame algorithm from Table 1 is an example of a direct adaptation of synchronous detection where simple values of 1, −1 or 0 for every $\pi/2$ phase shift can be assigned to the sine and cosine functions.

Algorithm Design

In the last ten years, a lot of work has been done to generalize the derivation of fringe analysis algorithms. This work has enabled the design of algorithms for specific applications, which are insensitive to specific systematic error sources.

Most algorithms use polynomials for the numerator and denominator. Given fringe data:

$$I_i = I_0[1 + \gamma\cos(\phi + \alpha_i)] \qquad [15]$$

the phase is calculated using

$$\phi = \tan^{-1}\left[\frac{\sum_i n_i I_i}{\sum_i d_i I_i}\right] \qquad [16]$$

Table 1 Sampling function weights for a few selected algorithms

N frames	*Phase shift*	*Coefficients*
3	$\frac{\pi}{2}$	$\frac{1, -1, 0}{0, 1, -1}$
4	$\frac{\pi}{2}$	$\frac{0, -1, 0, 1}{1, 0, -1, 0}$
5	$\frac{\pi}{2}$	$\frac{0, -2, 0, 2, 0}{1, 0, -2, 0, 1}$
7	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, 1, 1, 0, -1, -1, 0]}{-1, -1, 1, 2, 1, -1, -1}$
8	$\frac{\pi}{2}$	$\frac{1, 5, -11, -15, 15, 11, -5, -1}{1, -5, -11, 15, 15, -11, -5, 1}$
12	$\frac{\pi}{3}$	$\frac{\sqrt{3}[0, -3, -3, 3, 9, 6, -6, -9, -3, 3, 3, 0]}{2, 1, -7, -11, -1, 16, 16, -1, -11, -7, 1, 2}$

The numerator and denominator of the arctangent argument are both polynomials. The numerator is a sum proportional to the sine (imaginary part) and the denominator is a sum proportional to the cosine (real part),

$$\text{num} = 2kI_0\gamma\sin\alpha_i = \sum_i n_i I_i \qquad [17]$$

$$\text{den} = 2kI_0\gamma\cos\alpha_i = \sum_i d_i I_i \qquad [18]$$

where the constant k depends on the values of coefficients. From this the fringe visibility is given by

$$\gamma = \frac{\sqrt{(\text{num})^2 + (\text{den})^2}}{2kI_0} \qquad [19]$$

The coefficient vectors for the numerator and denominator are window functions. For an algorithm such as the 4-frame technique, the weights of all samples are equal $[1, 1, 1, 1]$. This makes the coefficients all equal. For other algorithms, such as the 5-frame technique, the weights are larger on the middle frames than on the outer frames. The weights for the 5-frame technique are $[1, 2, 2, 2, 1]$. A property of the coefficient vectors is that the sum of the coefficients for each, the numerator and denominator, should be zero. Examples of coefficient vectors for a few selected algorithms are given in Table 1.

Heterodyne Interferometry

Historically, heterodyne techniques were developed and used before temporal phase-shifting techniques. These techniques generally determine the phase electronically at a single point by counting fringes and fractions of fringes. Areas are analyzed by scanning a detector.

Phase shifts are usually obtained using two slightly different frequencies in the reference and test beams. The beat frequency produced by the interference between the reference and test beams is compared to a reference sinusoidal signal, which may be produced either optically or electronically. The time delay (or distance traveled) between the crossing of the zero phase points of the test and reference sinusoidal signals is a measure of the phase. Every time the test signal passes through another zero in the same direction as the test surface is moved, another fringe can be counted. This is how fringe orders are counted. If the beam is interrupted as the detector is scanned across the interferogram, the fringe count is corrupted and the measurement needs to be started again. Frequency multiplication (harmonics) can also be used to determine fractions of fringes. Today, heterodyne techniques are used mainly in distance measuring interferometers. The precision and accuracy of distance measuring interferometers is at least on the order of 1 part in 10^6.

Fourier-Transform Technique

The Fourier-transform technique is a way to extract phase from a single interferogram. It is used a lot in nondestructive testing and stellar interferometry where it is difficult to get more than a single interferogram. The basic technique is shown schematically in Figure 11. The recorded interferogram distribution is Fourier transformed, and one order (usually the + or − first order) is either isolated and shifted to zero frequency or filtered out using a rectangular window. After an inverse Fourier transform, the result is the phase.

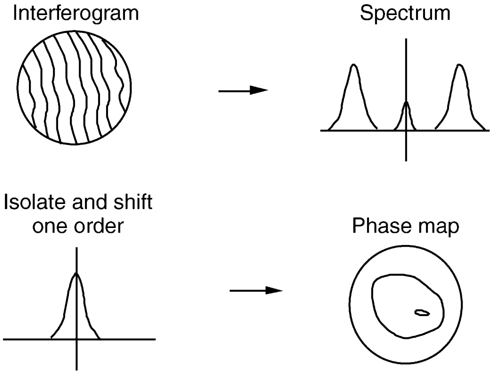

Figure 11 Fourier transform technique.

To illustrate this technique mathematically, the interference equation is rewritten as

$$I(x, y) = I_0(x, y) + c(x, y)\exp(\mathrm{i}2\pi f_0 x) + c^*(x, y)\exp(-\mathrm{i}2\pi f_0 x) \qquad [20]$$

where $c(x, y) = I_0(x, y)\gamma(x, y)\exp[\mathrm{i}\phi(x, y)]$ and the $*$ indicates a complex conjugate. The term $c(x, y)$ contains the phase information we wish to extract. After performing a one-dimensional Fourier transform:

$$I(\xi, y) = I_0(\xi, y) + c(\xi - f_0, y) + c^*(\xi - f_0, y) \qquad [21]$$

where ξ is the spatial frequency in the x direction, and italics indicate Fourier transforms. The next step is to filter out and isolate the second term, and then inverse Fourier transform to yield $c(x, y)$. The wavefront modulo 2π is then given by

$$\phi(x, y) = \tan^{-1}\left\{\frac{\mathrm{Im}[c(x, y)]}{\mathrm{Re}[c(x, y)]}\right\} \qquad [22]$$

where Re and Im refer to the real and imaginary part of the function.

This technique has limitations. If the fringes are nonsinusoidal, there will not be a simple distribution in the frequency space; there will be many orders. Another problem is overlapping orders. There needs to be a carrier frequency present that ensures that the orders are separated in frequency space. This carrier frequency is produced by adding tilt fringes until the orders are separated. This means that the aberration (fringe deviation) has to be less than the fringe spacing. Another problem is aliasing. If the interferogram is not sampled sufficiently, there will be aliasing and it will not be possible to separate the orders in frequency space. Finally, large variations in average fringe irradiance and fringe visibility across the interferogram can also cause problems.

Spatial Carrier-Frequency Technique

This is essentially the equivalent of the Fourier transform technique but is performed in the spatial domain. It is also used when there is only one interferogram available and its major applications include nondestructive testing and measurement of large optics.

These techniques relate closely to the temporal phase-measurement methods; however, instead of using a number of interferograms they can obtain all the information from a single interferogram. As an example, let's assume that the fringes are vertical and parallel to the columns on the detector array.

The carrier frequency (i.e., the number of tilt fringes) is adjusted so that there is an α phase change from the center of one pixel to the next. As long as there is not much aberration (deviation), the phase change from pixel to pixel across the detector array will be approximately constant.

When the fringes are set up this way, the phase can be calculated using adjacent pixels (see Figure 12). If one fringe takes up 4 pixels, the phase shift α between pixels will be 90°. An algorithm such as the three-frame, four-frame, or five-algorithm can be used with adjacent pixels as the input. Therefore, 3, 4, or 5 pixels in a row will yield a single-phase point. Then, the analysis window is shifted sideways one pixel and phase is calculated at the next point. This technique assumes that the dc irradiance and fringe visibility do not change over the few pixels used to calculate each phase value.

Spatial Multichannel Phase-Shift Techniques

These techniques detect all phase maps simultaneously and multiplex the phase shift using static optical elements. This can be done by using either separate cameras as illustrated below or by using different detector areas to record each of the interferograms used to calculate the phase. The phase is usually calculated using the same techniques that are used for the temporal phase techniques.

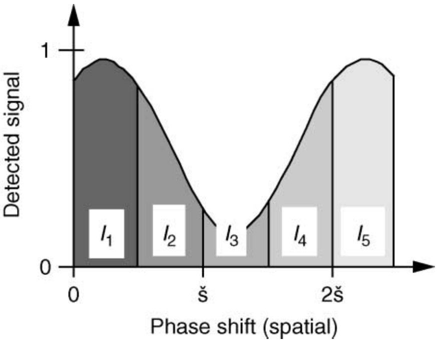

Figure 12 Spatial phase shifting. Different relative phase shifts are on adjacent detector elements.

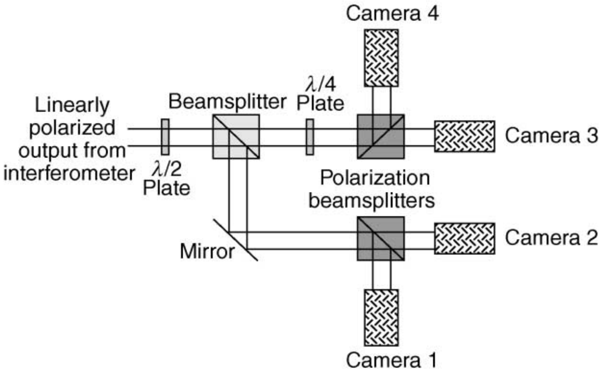

Figure 13 Spatial phase shifting using four cameras.

As an example, a four-channel interferometer can be made using the setup shown in Figure 13 to record four interferograms with 90° phase shifts between them. Camera 1 will yield fringes shifted 180° with respect to camera 2, and cameras 3 and 4 will have phase shifts of 90° and 270°. The optical system may also utilize a holographic optical element to split the beam to multiplex the four phase shifts on four quadrants of a single camera.

Signal Demodulation Techniques

The task of determining phase can be broadened by looking toward the field of signal processing. For communication via radio, radar, and optical fibers electrical engineers have developed a number of ways of compressing and encoding a signal as well as decompressing a signal and decoding it. An interference fringe pattern looks a lot like an am radio signal. Thus, it can be demodulated in similar ways. In recent years many new algorithms have been developed by drawing on techniques from communication theory and applying them to interferogram processing. Many of these use different types of transforms such as Hilbert transforms for straight fringes or circular transforms for closed fringes.

Extended Range Phase Measurement Techniques

A major limitation of phase measurement techniques is that they cannot determine surface discontinuities larger than $\lambda/4$ ($\lambda/2$ in optical path (OPD)) between adjacent pixels. One obvious solution is to use longer wavelength sources in the infrared where optically rough surfaces look smooth and their shape can be measured. An alternative is two-wavelength interferometry where two measurements at different wavelengths are taken and the measurable height limitation is now determined by the equivalent wavelength:

$$\lambda_{\mathrm{eq}} = \frac{\lambda_1 \lambda_2}{|\lambda_1 - \lambda_2|} \qquad [23]$$

Another method that allows for measurement of smooth surfaces with large height discontinuities that is limited only by the working distance of the objective combines a white light and a phase measurement interferometric technique in one long scan. The position of the wave packet resolves 2π ambiguities that result from the arctangent function (see Phase unwrapping section above). The phase of the fringe close to the wave packet maximum is determined without ambiguity. Sometimes it is not the step height discontinuity that is a problem but rather the high slope of the measured surface. If we know that the surface is continuous, then the unwrapping procedure can take advantage of this a priori information to look for a continuous.

Techniques for Deformation Measurement

Some of the techniques for deformation measurement have already been described earlier (see sections on Single frame techniques and Spatial multichannel phase-shift techniques). However, fringes in any interferometer can be analyzed not only in the x, y planes but also in the x, z, or y, z planes for which the carrier frequency of the fringes is introduced. Analysis of fringes in any plane has the same restrictions.

If the initial static object shape is measured using conventional phase measurement techniques, not only can the motion of each object point be determined but also the deformation of the whole object in time. If the motion of the object is periodic, then the object motion can be 'frozen' by using stroboscopic illumination of the same frequency as the object's motion. Once the motion is frozen, any temporal technique can be used to measure object shape at some phase of its motion. Changing the time offset between stroboscopic illumination and the periodic signal driving the motion, the object can be 'frozen' and thus measured at different phases of its periodic motion.

Systematic Errors

Noise Sensitivity

Measurement noise mostly arises from random fluctuations in the detector readout and electronics. This noise reduces precision and repeatability. Averaging multiple measurements can reduce effects of these random fluctuations.

Phase Shifter Errors

Phase shifter errors can be due to both miscalibration of the system and nonlinearities in the phase shift. It is possible to purchase very linear phase shifters. It is also possible to correct nonlinearities by determining the voltage signal making the phase shifter provide a linear phase shift.

Linear phase shifter errors (miscalibration) of the phase shift have an error signature that is at twice the frequency of the fringes. If there are two fringes across the field of view, the error signature will have four across the field of view. Figure 14 shows the difference between a calibrated and an uncalibrated phase shifter as well as the difference in error for two different phase measurement algorithms. Some algorithms are obviously more sensitive than others to this type of error.

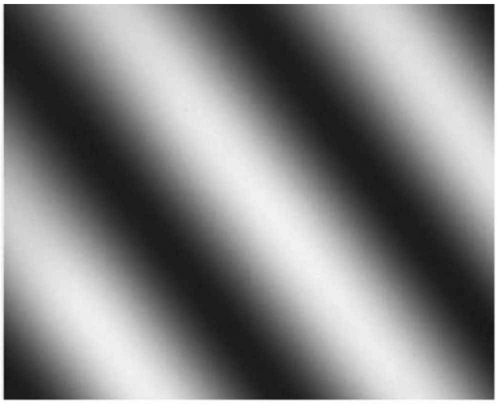

(a) Fringes on flat

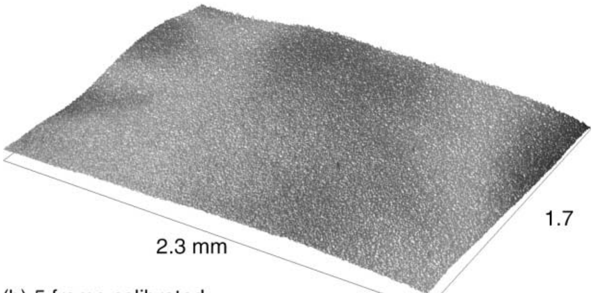

(b) 5 frame calibrated

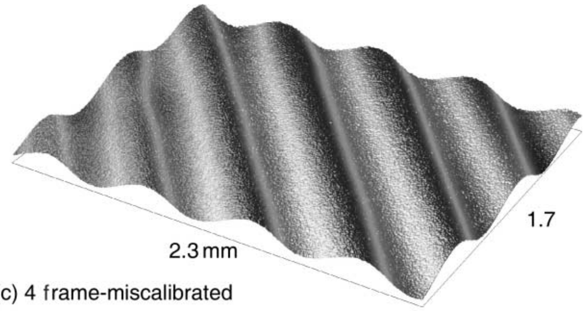

(c) 4 frame-miscalibrated

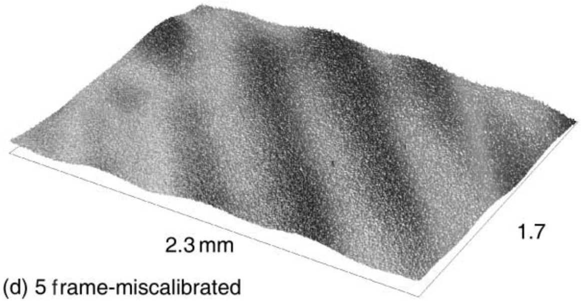

(d) 5 frame-miscalibrated

Figure 14 Comparison of phase maps for calibrated and uncalibrated phase shifts using two different phase measurement algorithms.

Other Types of Systematic Errors to Consider

Other types of errors to consider are detector nonlinearities, quantization errors due to analog-to-digital converters, and dissimilar materials (see Phase change upon reflection). For temporal phase measurement techniques errors due to vibration and

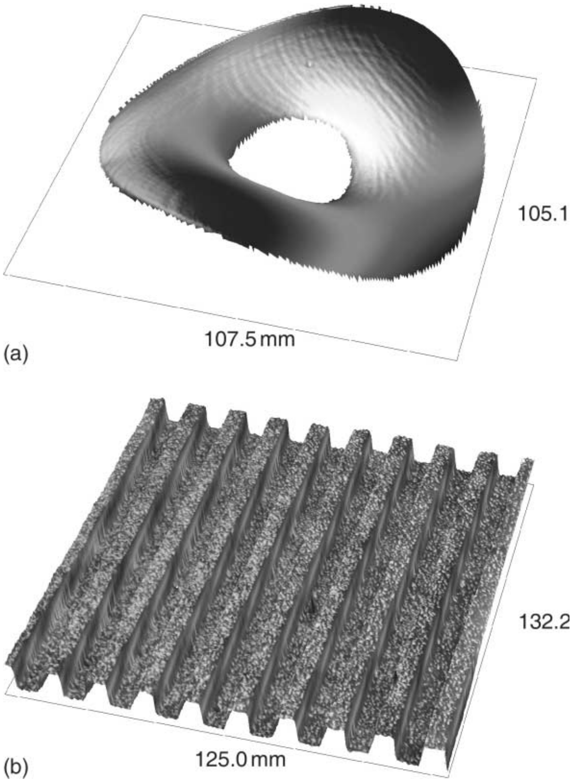

Figure 15 Measurements of (a) hard disc substrate with a vertical range is 2.2 micrometers and (b) binary grating roughness standard.

air turbulence need to be considered as well. Spatial phase measurement techniques are sensitive to miscalibrated tilt (wrong carrier frequency), unequally spaced fringes, and sampling and windowing in Fourier transform techniques.

Choosing an Algorithm or Technique

Each algorithm and type of measurement technique is sensitive to different types of systematic errors. Choosing a proper algorithm or technique for a particular type of measurement depends on the specific conditions of the test itself and reducing the systematic errors for a particular type of measurement. This is the reason that so many algorithms exist. The references in the bibliography will help the reader determine what type of algorithm will work best for a specific application.

Examples of Applications

Phase shifting interferometry can be used for measurements such as hard disk flatness, quality of optical elements, lens curvature, dimensions and quality of air-bearing surfaces of magnetic read/write heads, cantilevers, and semiconductor elements. Figure 15 shows results for measurements of a hard disk substrate and a roughness grating.

Conclusions

Phase measurement interferometry techniques have increased measurement range and precision enabling the production of more complex and more precise components. As work continues on development of interferometric techniques, phase measurement techniques will continue to become more robust and less sensitive to systematic errors. Anticipated advances will enable measurements of objects that were unimaginable 20 or 30 years ago.

See also

Interferometry: White Light Interferometry. **Microscopy:** Interference Microscopy.

Further Reading

Bruning JH (1978) Fringe scanning interferometers. In: Malacara D (ed.) *Optical Shop Testing*, pp. 409–438. New York: Wiley.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. 26, pp. 349–393. Amsterdam: Elsevier.

Creath K (1993) Temporal phase measurement methods. In: Robinson DW and Reid GT (eds) *Interferogram Analysis*, pp. 94–140. Bristol: IOP Publishing.

Creath K and Schmit J (1996) N-point spatial phase-measurement techniques for non-destructive testing. *Optics and Lasers in Engineering* 24: 365–379.

Ghiglia DC and Pritt MD (1998) *Two-Dimensional Phase Unwrapping*. New York: Wiley.

Greivenkamp JE and Bruning JH (1992) Phase shifting interferometry. In: Malacara D (ed.) *Optical Shop Testing*, 2nd edn, pp. 501–598. New York: Wiley.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Kerr D (ed.) (2004) FASIG special issue on fringe analysis. *Optics and Lasers in Engineering* 41: 597–686.

Kujawinska M (1993) Spatial phase measurement methods. In: Robinson DW and Reid GT (eds). *Interferogram Analysis*, pp. 141–193. Bristol, UK: IOP Publishing.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of Optical Society of America A* 13: 832–842.

Malacara D (1992) *Optical Shop Testing*, 2nd edn. New York: Wiley.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Robinson DW and Reid GT (1993) *Interferogram Analysis: Digital Processing Techniques for Fringe Pattern Measurement*. Bristol, UK: IOP Publishing.

Schmit J and Creath K (1996) Window function influence on phase error in phase-shifting algorithms. *Applied Optics* 35: 5642–5649.

Surrel Y (2000) Fringe analysis. *Photo-Mechanics: Topics in Applied Physics* 77: 55–102.

White Light Interferometry

J Schmit, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1960 the development of the laser injected new possibility into an old discipline, interferometry (*see* **Interferometry**: Overview). The excellent spatial and temporal coherence (ideally a single point and a single wavelength source) of the laser allowed for the formation of nonlocalized interference fringes. With lasers, researchers could easily generate good contrast fringes practically anywhere where the beams overlapped. These excellent-quality fringes from laser interferometers enabled high-resolution distance and displacement measurements on the order of meters and made possible noncontact surface probing with nanometer and subnanometer resolution. This kind of precision, which was supported by the advent of computers and the advancement of detectors, was previously unimaginable.

However, interference phenomena were used in metrology long before lasers and can be observed without any complicated device or source. On a rainy day we can observe the colorful interference patterns that have formed on a thin layer of gasoline in a puddle. The relationship between colors and the small distance between reflecting surfaces, observed as far back as 1665 by Hook, is still used to approximate layer thicknesses. The colors in interference patterns have been used to determine the spectral components of beams. By the end of the nineteenth century, Michelson and Benoit were using interference phenomena to determine distances, long before the invention of the computer and the laser.

While the application of lasers in interferometry has advanced the science, white light interferometry (WLI), which uses a spatially and temporally incoherent source and creates fringes localized over a few microns in space, has also benefited from technological advancements. The development of solid state detectors, fast computers, electronic signal processing, and precise scanning stages has allowed for incredibly fast analysis of white light interferograms. In fact, WLI is used in many applications from film thickness and surface measurement, through spectroscopy to astronomy.

This chapter focuses on the formation of white light fringes. We first examine the existence of white light fringes in the everyday world. A description of different interferometric setups and their uses follows. The bulk of this article details the formation of white light fringes and examines their different applications with an emphasis on the analysis of object structure. Suggestions for further reading, both within this encyclopedia and from outside sources, can be found at the end of this article. A basic theory of interferometry and coherence is to be found in corresponding articles of this encyclopedia.

White Light Interference in Everyday Life

The effects of interference display spectacular colors when observed in white light, such as the sun. We often see these effects in oil or gasoline spills, in soap bubbles, or between two smooth pieces of glass in contact (separated by a very thin layer of air film) – which are the simplest everyday white light interferometers. The colors we see in these displays are interference colors, and their origin is interference rather than dispersion as in the colors of a rainbow or light reflected from a CD. The picture in Figure 1 is of a gasoline spill under cloudy sky illumination; we see beautiful interference colors that are hard to find in a rainbow: iron-gray, magenta, grayish blue, whitish green, and brownish yellow. Interference colors will differ not only with the thickness of the layer but also with the layer's absorption and dispersion, the relative indices of refraction of the film, the surrounding media, and the illumination. Interference colors can be observed for layers from a fraction of a micron to a few microns thick.

Interferometers for White Light Observation

In a white light interferometer, either the colors or the intensity distribution of the fringes is typically analyzed to retrieve the necessary encoded information, such as the film thickness, birefringence, index of refraction, dispersion, spectral properties, and surface structure. White light interference can be observed only in interferometric designs where the optical paths in both arms of the interferometer are (nearly) equal and the system is (nearly) compensated for dispersion.

Interference in Thin Film

Beams of light of different wavelengths incident on a transparent thin film (such as the ones in the puddle) are partially reflected from the top air/film interface

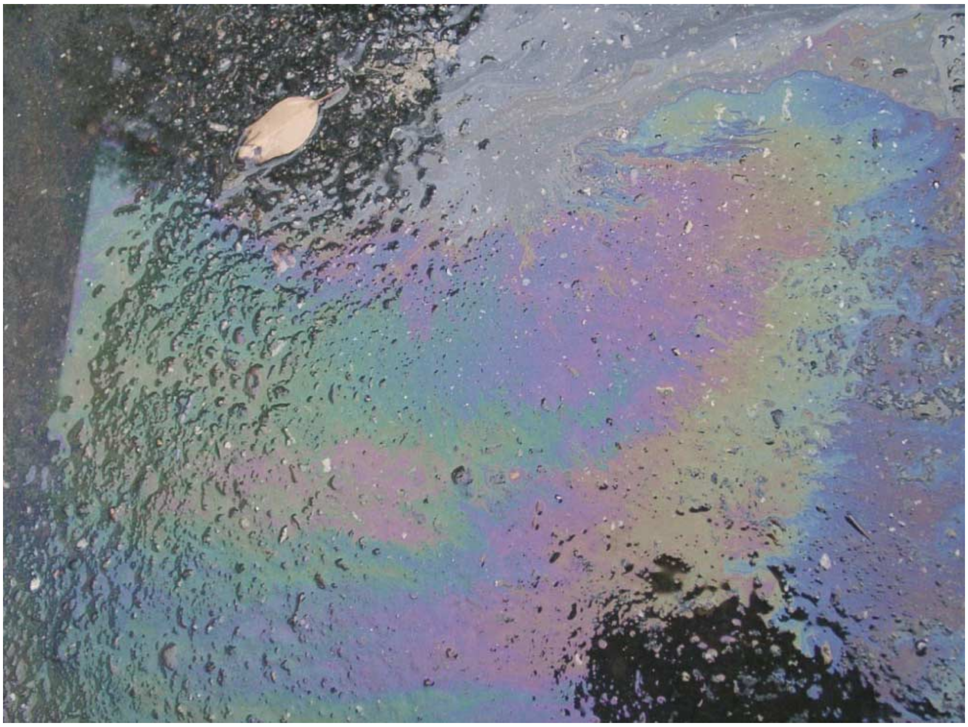

Figure 1 Interference colors in a gasoline spill in a puddle. This picture would closely resemble a picture from an interference microscope, if not for the leaf in the top left corner.

and partially from the bottom film/substrate interface and then beams of the same wavelength interfere with each other. The optical path difference (OPD) traveled by the interfering beams is related to the film thickness, the index of refraction for the given wavelength and the angle of incidence.

Fringes of equal thickness will be observed if the film thickness (distance between interfering wavefronts) varies. These fringes represent points of equal distance between the two surfaces of the film and are formed close to the film (see Figure 2). This situation is typical in any interferometer since rarely are wavefronts perfectly plane and parallel.

For any individual wavelength, λ_i, where the OPD is equal to $m\lambda_i$, and where m is an integer, a bright fringe of the color of the wavelength, λ_i, will be observed due to constructive interference. For white light illumination, the color at a given point will be dominated by the color of the wavelength for which the interference will be constructive. This color can be used to estimate the optical thickness of the film.

An additional factor needs to be taken into account when color is used for estimating optical thickness, namely the phase change on reflection. Different colors will be observed if the ratio of indices of refraction for both film interfaces is <1 (or >1) as opposed to when the ratio of indices is <1 for one interface and >1 for the other.

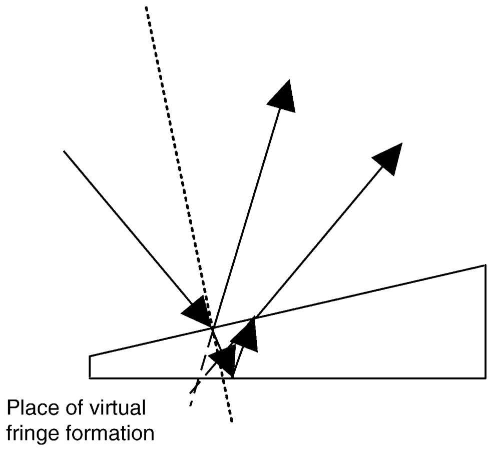

Figure 2 Formation of fringes of equal thickness for film with wedge.

Polarization Microscope

Interference colors that correspond to equal thickness fringes can represent the birefringence of an object; these colors are often observed using a polarization microscope (*see* **Microscropy:** Interference Microscopy). The polarization microscope is a conventional microscope with a set of polarizers with crossed polarization axes, one placed above and the other placed below the tested birefringent object.

Interference colors have been tabulated for many years and for many different purposes. Newton devised his color scale to describe interference fringes for two pieces of glass in contact with each other. Michel-Lévy and Lacroix in 1889, created a color scale to help recognize different rock forming minerals. For more information about colors in white light interference see sections on *White Light Interference* and *Spectral Interference* below.

Michelson Interferometer

A Michelson interferometric setup, shown in Figure 3, is often used to analyze white light interference, but the intensity of the fringes is observed as one mirror is scanned rather than the color of fringes that is usually observed.

Two plane mirrors (equivalent to the top and bottom surfaces of a thin film) return the beams to the beamsplitter, which recombines parts of the returned beams and directs them towards the detector where the interference is observed. Beam S1 travels through the parallel plate three times while beam S2 passes through the plate only once causing the system to be not well compensated (balanced) for dispersion. Thus, for observation of white light interference, an identical plate is placed in the path of beam S2. As one of the plane mirrors moves along the optical axis to change the OPD, the detector collects the irradiance, the intensity. Fringes will be observed if the system is well-compensated for dispersion and for optical path lengths. Any spectral changes or changes in optical path lengths in the interferometer affect the shape or position of the fringes, and the interferometer measures these changes.

Common-path interferometers, like the Mach–Zender or Jamin interferometers, are naturally compensated interferometers and can be used for measurement of dispersion in gases. Other interferometers, such as the Twyman–Green or Fizeau interferometers, make use of their unequal interferometer arms and their nonlocalized fringes from laser sources for testing different optical elements and systems in reflection and transition (*see* **Interferometry:** Phase Measurement Interferometry).

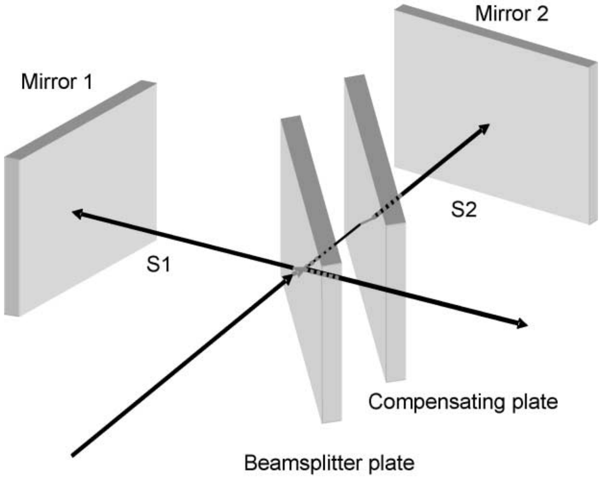

Figure 3 Michelson interferometer.

White Light Interference

A white light source consists of a wide spectrum of wavelengths in visible spectrum, from about 380 up to 750 (violet to red) nanometers. However, the principles of WLI described in this article basically apply when any low coherence source is used. Low coherence refers to a source of not only low temporal but also low spatial coherence; WLI can also be referred to as low coherence interferometry. We will be concerned mainly with temporal effect of the source, the source spectrum.

Since different wavelengths from the source spectrum are mutually incoherent, we will first look at interference between two waves of a selected monochromatic illumination with wave number $k = 2\pi/\lambda$, where λ is the wavelength. The intensity of the interference fringes at point x, y (these coordinates are omitted in all equations), as one of the mirrors is scanned along the optical axis, z can be described as

$$I(k,z) = I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma(z)|\cos(kz) \qquad [1]$$

or can be written in the form:

$$I(k,z) = I_0\left[1 + \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma(z)|\cos(kz)\right] \qquad [2]$$

where I_1 and I_2 are intensities of each of the beams and $I_0 = I_1 + I_2$, $|\gamma(z)|$ is the modulus of the complex mutual coherence function (*see* **Coherence:** Overview) assumed here to equal 1 (perfect coherence for each wavelength). The optical path difference z equals $z_1 - z_2$ where z_1 and z_2 are the optical path lengths that the interfering waves have traveled. The difference in the traveled paths, z, corresponds to the position of the scanning mirror.

White light interference (polychromatic interference) is the overlaying of all the monochromatic fringes created for each wavelength of the source spectrum (Figure 4a). A detector observes the sum of all the fringe intensities.

In mathematical form, this interference can be described as the integral of all the fringes $I(k,z)$ for all

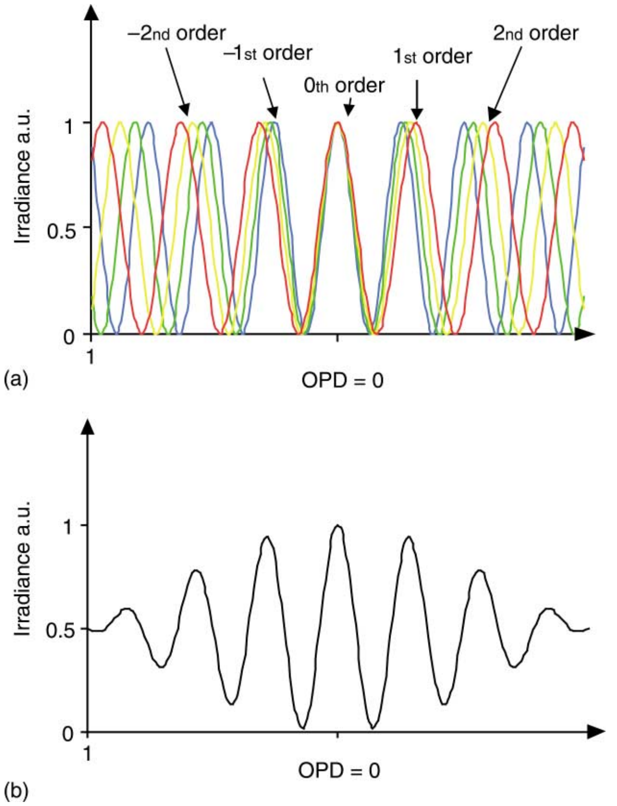

Figure 4 Formation of white light fringes: (a) fringes for individual wavelengths; (b) fringes for white light.

wavenumbers k:

$$I(z) = \int_{k_1}^{k_2} S(k)D(k)I(k,z)\mathrm{d}k \qquad [3]$$

where $S(k)$ is the spectral distribution of the light source, with $S(k) = 0$ outside of k_1 through k_2 range of wave numbers, and $D(k)$ is a spectral response of the detector. We assume that the detector's response $D(k)$ equals 1 over the whole spectrum. Because the spacing of the fringes for each wavelength emitted by the white light source is different, the maxima of fringes will align around only one point for a well-compensated interferometer, as shown in Figure 4a. This alignment around a single point, occurs because in a well-compensated interferometer there is one point for which the OPD is zero for all wavelengths. Away from this point, the observed sum of the intensities quickly falls off, as shown in Figure 4b. The maximum fringe, the fringe that marks the zero OPD, is called the zero-order fringe, and each next fringe of smaller amplitude on either side is called $+1$ and -1, $+2$ and -2 order fringe, and so on.

Because white light fringes are localized and can only be found within microns or tens of microns of the zero OPD, white light interferometers are excellent distance and 3D position sensors. However, this same characteristic makes them more difficult to align than interferometers with nonlocalized fringes.

Envelope of Fringes Due to Source Spectrum

The resultant intensity of interference from a broad spectrum source (see eqn [3]) can be described in general form as

$$I(z) = I_0[1 + V(z)\cos(k_0 z)] \qquad [4]$$

where $I_0 = I_1 + I_2$ is the background intensity, $V(z)$ is the fringe visibility function or coherence envelope, and $k_0 = 2\pi/\lambda_0$ is the central wave number for fringes under the envelope. The $V(z)$ is proportional to the modulus of the Fourier transform of the source spectrum $S(k)$. Generally, if the light source has a Gaussian spectrum $S(k)$, then the envelope of the fringes can be described also as a Gaussian function $V(z)$. If the spectrum of the source is rectangular, then the envelope of the fringes will be a sinc function. The wider that the spectrum of the source is the narrower the width of the envelope will be, as shown in Figure 5. The width of the fringe envelope determines the coherence length of the source (*see* **Coherence:** Overview); for a white light source this width is in the order of 1–2 microns.

Different white light sources, such as tungsten-halogen, incandescent, or arc lamps, have different spectra and thus create different coherence envelopes, as shown in Figure 5. The spectra of semiconductor light sources, such as LEDs and SLDs, are similar in shape to a Gaussian function.

The fact that the coherence envelope is related to the spectrum of the source by its Fourier transform is commonly used in Fourier transform spectroscopy, where the Fourier transform of the detected fringes is calculated to find the spectral components of the beams.

Position of Fringes Under Envelope Due to Reflection of Dielectric

Thus far we have assumed that in WLI the interferometer is compensated for all wavelengths, and thus, for this position the maximum of the fringes aligns with the maximum of the envelope, namely where there is a zero phase shift $\varphi_0 = 0$ between fringe and envelope maxima.

If there is an odd number of reflections from dielectric surfaces in one arm of the interferometer and an even number in the other, the fringes will be shifted by $\varphi_0 = 180°$ under the coherence envelope and the minimum of the fringe will align with the maximum of the coherence envelope. Thus eqn [4]

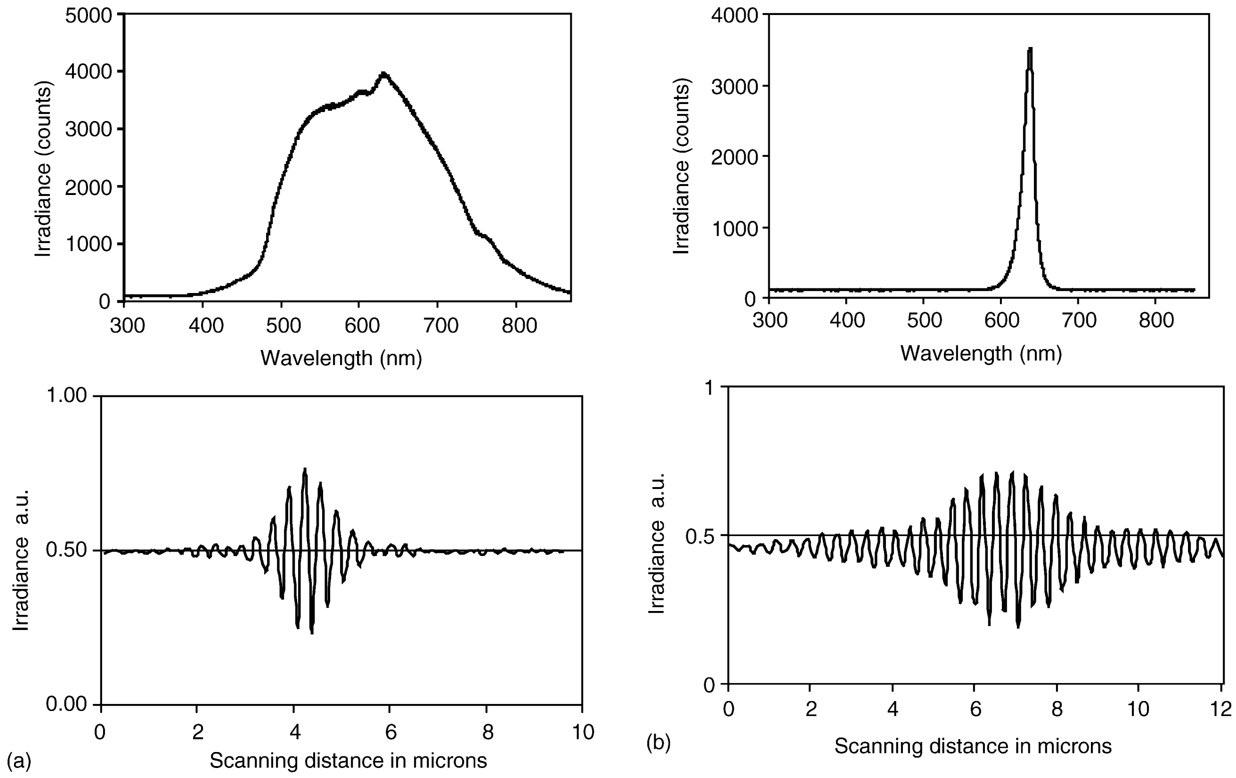

Figure 5 Spectrum and interferogram for (a) tungsten-halogen lamp and for (b) red LED sources.

can be expressed in a more general form:

$$I(z) = I_0[1 + V(z)\cos(k_0 z + \varphi_0)] \qquad [5]$$

Figure 6 shows color fringes for such a case; the dark fringe marks the zero OPD, and this fringe is surrounded by the greenish blue colors of shorter wavelengths. In contrast, the bright fringe at the maximum envelope position or marking the zero OPD would be surrounded by the reddish colors of longer wavelengths.

In a real system the fringes may be shifted with respect to the envelope by any amount of φ_0, and this shift may be due to any number of factors, such as the phase shift on reflection from nondielectric surfaces and dispersion, which we consider next.

Changes in Envelope and Fringes Due to Reflection of Nondielectric Materials

The relative position of the envelope maximum and fringe position will be different if the beam is reflected from different nondielectric materials. This difference exists because the phase change on reflection from a test surface, like metals or heavily doped semiconductors, varies with each wavelength. This variance in fringe and peak coherence position may be predicted and corrected for in surface height calculations. The linear dependence of the phase change on reflection versus wave number shifts the location of the coherence envelope peak position of fringes by different amounts. The constant phase change on reflection and higher-order terms only shift the fringes underneath the coherence envelope. As long as the change on reflection has a small dependence on the second- and higher-order of the wave number, the shape of the coherence envelope is preserved.

Changes in Envelope and Fringes Due to Dispersion

Dispersion in an interferometer that is not balanced, perhaps because a dispersive element was placed in one arm or the compensating plate has an incorrect thickness, will influence fringe formation. The phase delay between interference patterns for individual wavelengths is proportional to the product of the geometrical path and the index of refraction equal to $d \times n(k)$.

The intensity may be described as:

$$I(z) = \int_{k_1}^{k_2} \langle 1 + V(z)\cos\{kz - kd[n(k)]\}\rangle \, \mathrm{d}k \qquad [6]$$

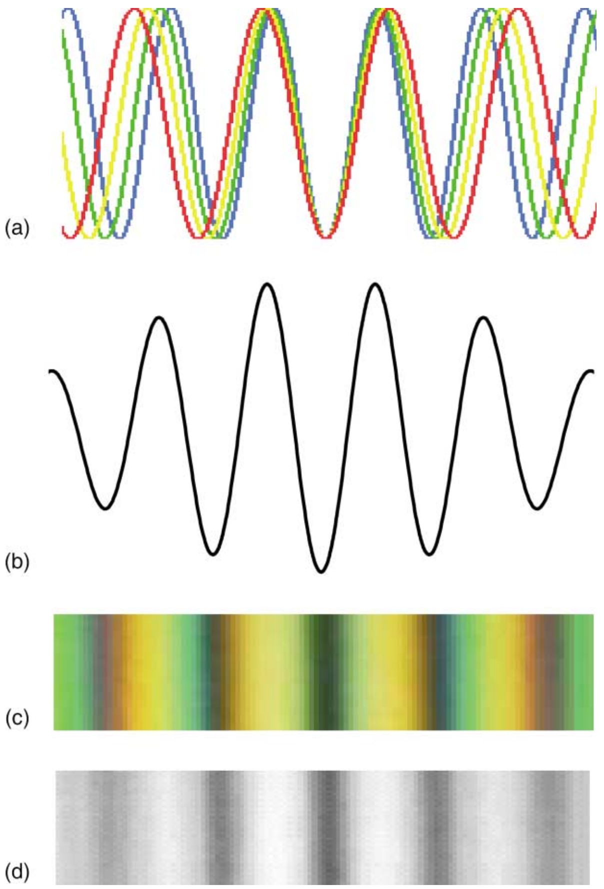

Figure 6 Formation of white light fringes with destructive interference for OPD = 0.

The dependence of the refractive index on the wave number k can be described as a linear expansion:

$$n(k) = n(k_0) + \frac{\mathrm{d}n}{\mathrm{d}k}(k - k_0) \qquad [7]$$

The linear dispersion shifts the envelope by the group index of refraction times the thickness of the dispersive element; this dispersion also shifts the fringes under the envelope slightly. In other words, the linear phase delay for a range of wavelengths causes a group delay of the whole envelope (wave packet). Higher-order dispersion, absorption of the elements, or effects due to thin films can cause the envelope to widen or even become asymmetrical, the position of the fringes under the envelope to shift, the fringes to lose contrast, and the period of the fringes to change or vary with the z position. Dispersion effects will be stronger for sources that have a wider spectrum; however, the observed changes will be different for different shapes of spectra.

The phase of the fringes under the envelope and the position of the envelope itself are parameters often used in astronomy. Highly accurate white light fringe estimation, using the optical path-length delay between the two arms of the interferometer, is a cornerstone of stellar interferometry.

Fourier Analysis of White Light Interferograms

Wavelength-dependent changes in a white light interferogram can be more easily analyzed in the spectrum rather than the spatial domain. The Fourier transform of a white light interferogram yields two symmetrical side lobes at the mean wavelength of the interferogram; analysis of just one of these side lobes is sufficient. The spectral amplitude of the side lobe contains information about the spectral components of the interfering beams, while the spectral phase in regions with appreciable spectral amplitude supplies information about any unbalanced dispersion in the interferometer, as shown in Figure 7. Fourier transform analysis is extensively used in Fourier transform spectrometry (*see* **Spectroscopy:** Fourier Transform Spectroscopy).

For a dispersion-balanced interferometer, the interferogram is symmetrical around the zero OPD position. For a symmetrical interferogram, the spectral phase will be zero if the zero OPD position of the sampled interferogram is in the middle of sampling range; otherwise a linear factor appears in the spectral phase. This linear term, while useful to surface profiling, because it determines the position of the coherence signal with respect to scanner sampling, is unwanted in Fourier spectroscopy and needs to be corrected for. Dispersion and possible thin-film effects will commonly introduce a nonlinearity in the spectral phase.

Controlled Phase Shift of Fringes Under the Envelope – Geometric Phase Shift

Many techniques in interferometry depend on shifting the phase of the interfering wavefronts. Mechanical shifters used in white light interferometry will introduce the same shift measured in nanometers for all wavelengths; however, when measured in degrees or radians, the shift will vary for different wavelengths. Geometric phase shifters (achromatic phase shifters operating on the principle of geometric phase) introduce for all wavelengths the same shift when measured in degrees and are based on polarization elements like a rotating wave plate in circularly polarized light or a rotating polarizer in a circularly polarized beam. Fringes underneath the coherence envelope shift, as shown in Figure 8, while the coherence envelope stays in place.

The advantage to these techniques is that only the phase of fringes changes, not the fringe contrast

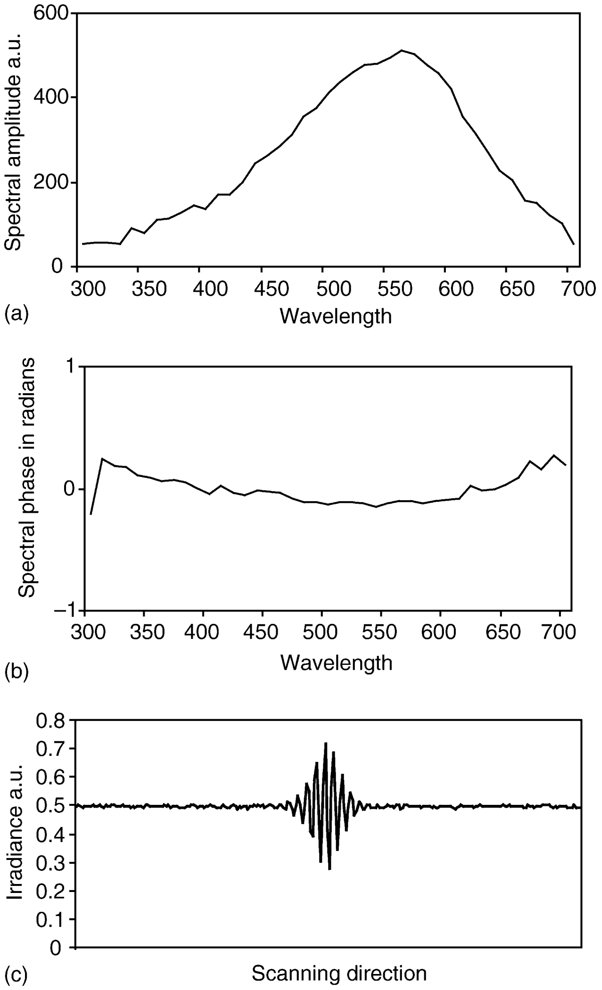

Figure 7 Spectral amplitude (a) and spectral phase (b) for input from unbalanced interferometer (c).

(*see* **Interferometry:** Phase Measurement Interferometry). A geometrical phase shifter can be very useful in polarization microscopes (Figure 9), white light shearing interferometers, or any system where the phase of the white light fringes needs to be measured.

Spectral Interference

White light fringes, because they are made up of fringes of many wavelengths, are observed only over very small path differences. If we filter only a narrow band from the white light spectrum, fringes would be visible over a much larger scan. For different wavelengths we would observe different color and frequency fringes; this is simply the reverse process of the description in the white light interference section.

If we place a spectrometer in the observation plane of the white light interferometer (Figure 10), we will observe fringes with continuously changing wavelengths in dispersed light. These fringes are called fringes of chromatic order or channeled spectrum fringes and find their application in film thickness measurement or absolute distance measurement in the range up to 1 mm. Channeled spectra were used for analysis of dispersion, thin film, and spectroscopic measurements. The number of observed fringes for a given wavelength is directly proportional to the measured optical path difference. The optical path difference can be determined if the difference in fringe numbers is determined for two well-known wavelengths (this is equivalent to two-wavelength interferometry). The optical path difference can be also quickly estimated from the frequency of fringes for a given wavelength; the larger the optical path difference, the denser the fringes.

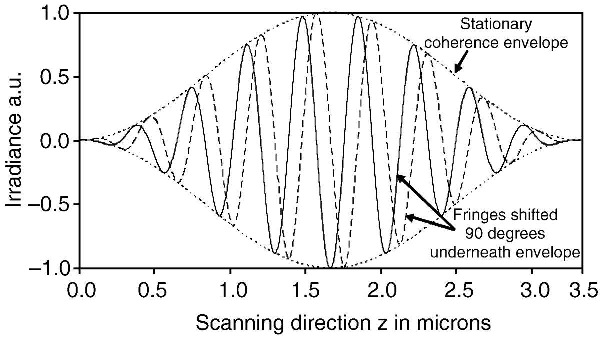

Figure 8 Fringes for two different geometric phase shifts.

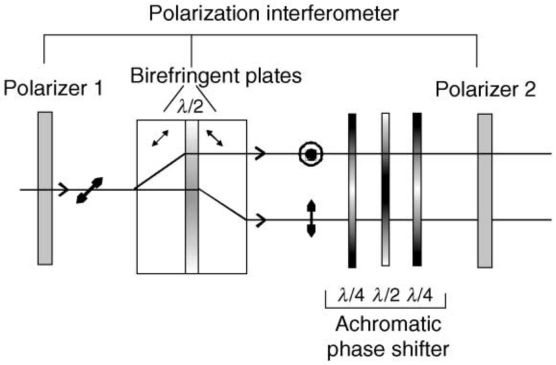

Figure 9 Polarization interferometer with geometric phase shift.

Surface Topography and Object Structure Measurement

Although WLI has many applications, this section focuses on white light interference as applied to surface topography measurement. Interference microscopes that use white light illumination are often based on the Michelson setup shown in Figure 11.

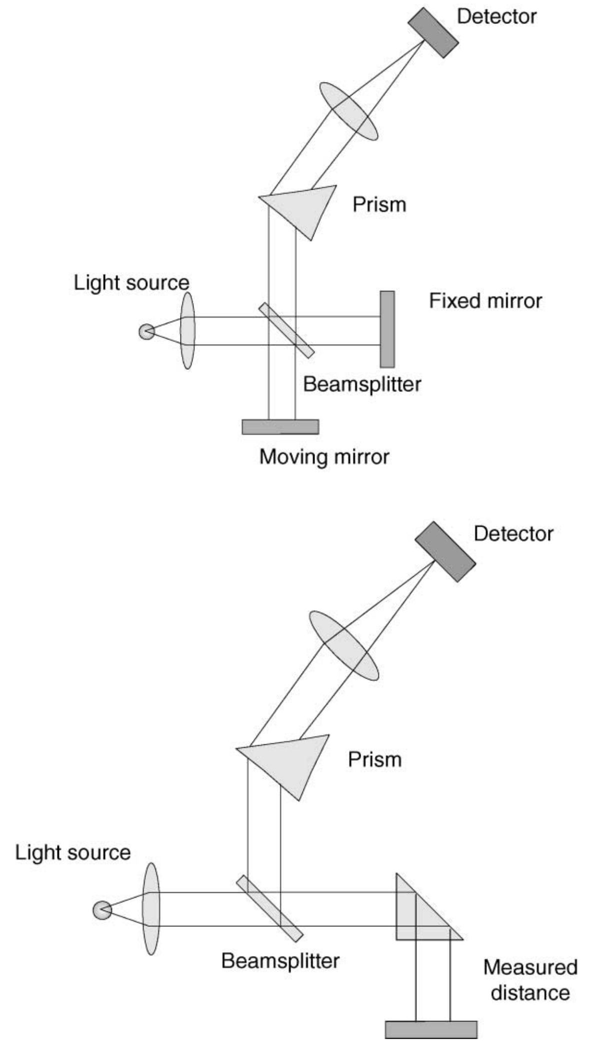

Figure 10 Interferometers with prism spectrometer to observe fringes of equal chromatic order for OPD measurement.

A beamsplitter cube is placed underneath the bright field objective, and one mirror is placed to the side at the focus plane of the objective while the other mirror is replaced by the measured surface with respect to which the whole interference objective is scanned. The narrowly localized fringes at the best focus position for each point of the surface corresponding to a pixel on the CCD camera is observed during the scan. Because of the design of the Michelson objective, in order to accommodate higher magnification objectives constraints, other interference setups like Mirau and Linnik were developed (*see* **Microscopy:** Interference Microscopy). However, these designs are still based on the Michelson interferometer, being also compensated interferometers with equal lengths and amounts of glass in each arm.

These interference microscopes typically have two complementary modes of operation; one mode uses monochromatic illumination and the other employs a white light source. Monochromatic illumination provides excellent vertical resolution but is limited in its range. It cannot correctly assign fringe order to discontinuous surfaces larger than 160 nanometers. Figure 12a shows monochromatic fringes for the surface of the profile shown in Figure 13. We see that with the monochromatic illumination the height of the grooves remains unknown because it is impossible to assign order numbers to these fringes.

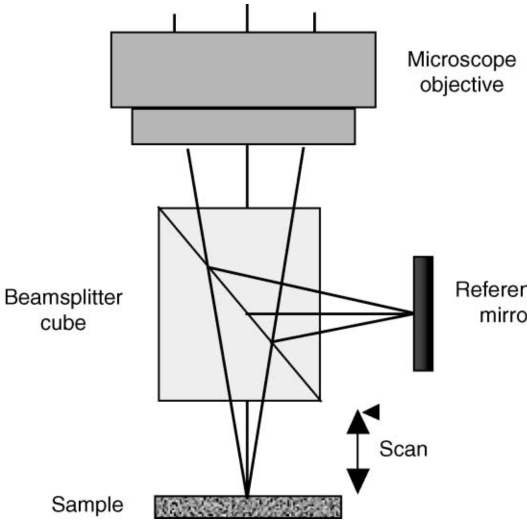

Figure 11 Michelson interferometric objective.

To resolve this ambiguity, white light illumination is employed because it allows for easy identification of the zero order fringe. Figure 12b shows white light fringes created for the same surface. We see the zero-order fringe for the zero OPD as well as the next orders due to the decreasing contrast of the fringes. The position of the zero-order fringe can be followed and approximate heights can be determined visually. Thus, white light illumination permits the measurement of a broader range of surfaces that are rough or have large height variations up to a few millimeters.

The white light interferometer is a great focus detector (Figure 14) at each point of the field of view. The principle behind WLI is finding these individual focus positions using localized fringes observed during the surface scan. WLI interferometers provide high lateral resolution and large vertical range in the order of hundreds of nanometers or micrometers (Figure 15). White light interferometers are commonly used to measure magnetic heads, MEMS devices, binary optic, and machined surfaces.

Signal Processing of White Light Interferograms

For WLI topographic measurements, each pixel registers an interferogram whose position varies

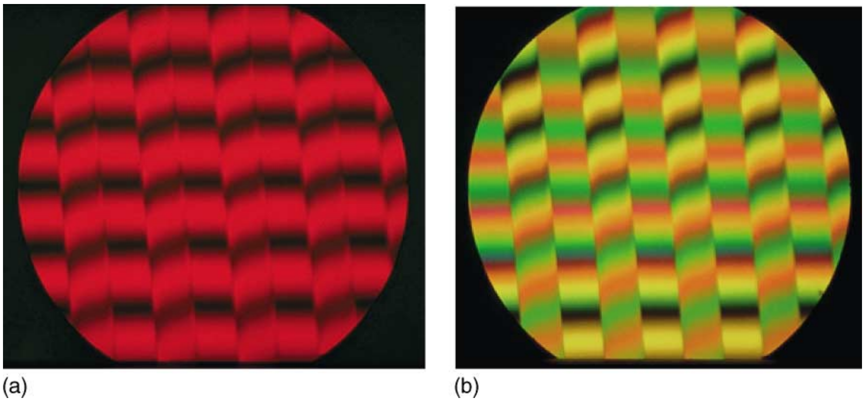

Figure 12 Fringes in quasimonochromatic and white light for object similar to one as presented in Figure 13.

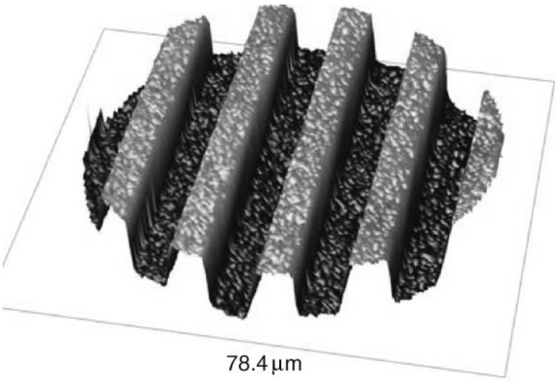

Figure 13 3D object profile – binary grating.

with the surface height. Figure 16 shows two interferograms for two positions in the field of view of a measured step height surface; the first shows the top surface of the step height and the second shows the bottom surface of the step height.

Algorithms have been developed to analyze the signal in different ways. Some focus on finding the fringe envelope position and others examine the position of the fringe underneath the envelope. These algorithms work under the assumption that white light fringes can be described by the same function over the whole field of view of the interferometer. Algorithms are applied to a signal so as to find the position of the coherence envelope for each pixel. This process is called coherence gating, and all algorithms perform coherence gating.

Some algorithms look for the position of the envelope's maximum in a process whereby first the coherence envelope of the fringes is found using signal filtering. Then, the curve is fit to a few points around the envelope's maximum, and finally the position of the envelope's maximum is found. Other algorithms calculate the center of mass of the coherence signal, as described in eqn [8]:

$$H(x,y) = \frac{\sum_{n=1}^{N-1} \left(I_i(x,y) - I_{i+1}(x,y)\right)^2 z_i}{\sum_{n=1}^{N-1} \left(I_i(x,y) - I_{i+1}(x,y)\right)^2} \qquad [8]$$

This method is very fast and computationally efficient. Center of mass calculations are equivalent to calculations of the maximum of the envelope position, but only for a symmetrical signal. For an asymmetrical signal a piston is introduced for each point; however, this piston does not affect the whole measurement. Still another method calculates the spectral phase slope in the Fourier domain to determine the position of the coherence envelope in a process that is equivalent to finding the center of mass. Finally, a different group of algorithms tracks the position of the bright or dark fringe close to the envelope's maximum.

Scanner Speed – Sampling Rate

A practical sampling rate of the white light interference signal is around four pixels per fringe. Because in most algorithms it is not the position of the fringes that we want to determine but rather the position of the envelope of the fringes, sampling the envelope of the fringes with four samples is usually sufficient. However, this sufficiency comes at the expense of lower repeatability (higher noise). The advantage to this sampling rate is short measurement time; higher sampling rates increase measurement time. Measurement speed can be increased by up to 25 times when the envelope of the fringes is widened by filtering light of a bandwidth about 20–40 nm from the white light source.

Scanner Nonlinearity

Scanner motion is assumed to be linear, but nonlinearity in this motion impairs measurement

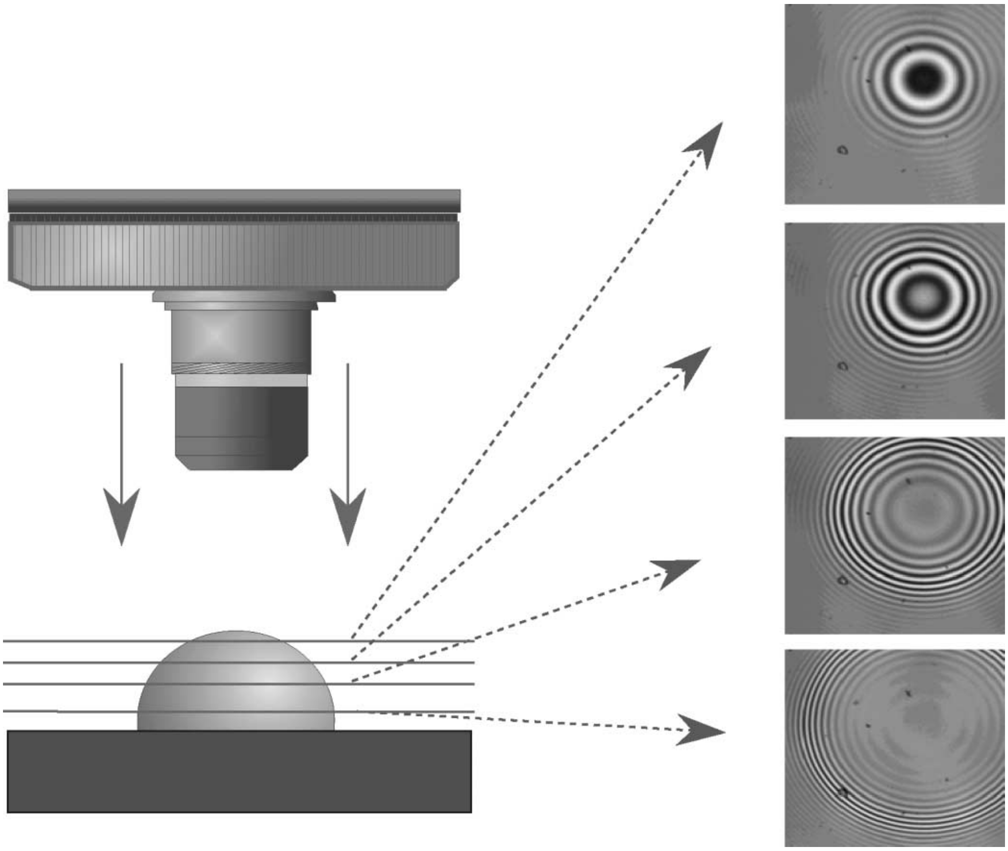

Figure 14 Operation of white light interferometer for surface topography measurement.

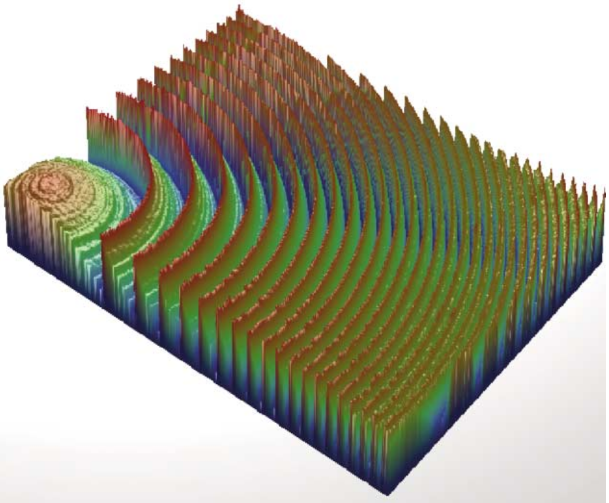

Figure 15 WLI measurement of Fresnel lens in a high-resolution mode. Courtesy of Glimmerglass.

accuracy. To account for this, simple algorithms can calculate the average steps along each interferogram in scanning direction (Figure 17) and measured steps can be used in coherence detection algorithms.

In order to have continuous information about the scan steps, the fringes should be visible somewhere in the field of view at each moment of the scan. This task may require introducing a large tilt for samples with large discontinuities such as stepped surfaces. For large discontinuities where introducing sample tilt may not be sufficient, the envelope of the fringes can be extended by increasing the spatial or temporal coherence of the light source. An alternate solution involves measuring the scanner motion with a distance-measuring interferometer or triggering the camera to collect intensity data at every equal scan step, which can be determined by using the zero-crossing technique, as is commonly done in Fourier transform spectroscopy. Equal scan steps are more suitable for techniques based on transform techniques that assume equal sampling. Other peak detection algorithms, such as center of mass calculations, make use of measured but not necessarily equal steps directly in the algorithm.

Observing the phase along the interferogram can also provide information that is important in optical fiber sensing. These observed changes, which assume that sampling rates of the interferogram are known,

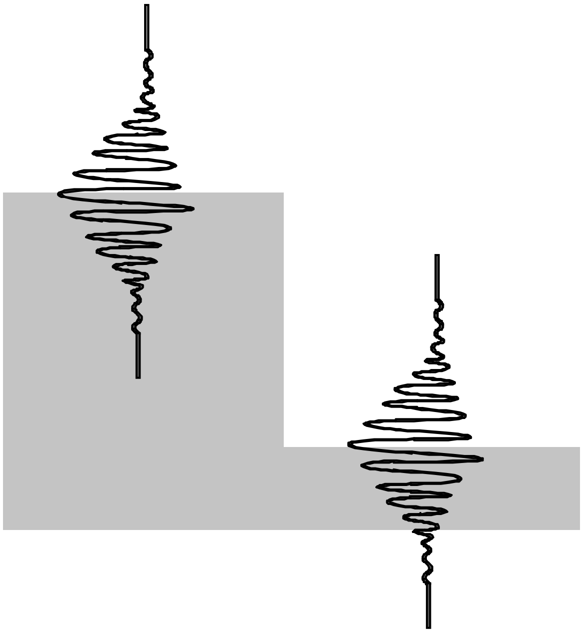

Figure 16 White light fringes as seen by two pixels during scan through focus for object in form of step.

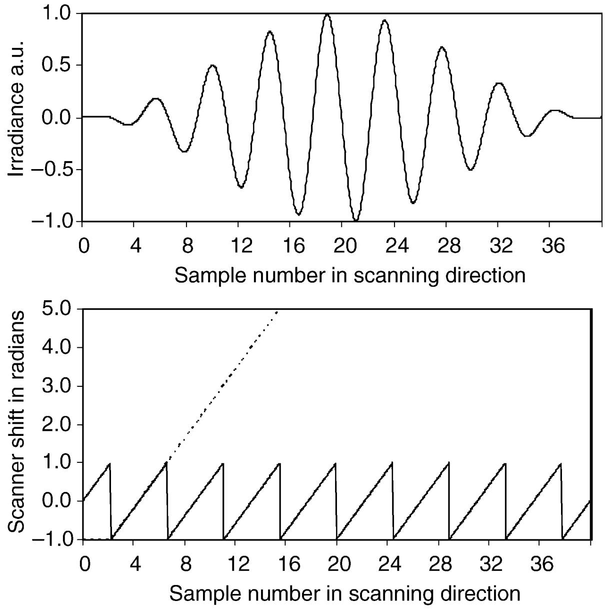

Figure 17 Phase calculated along the white light interferogram. This calculated phase can be used to determine scanner motion.

enable correction for changes in wave number value k_0 which can be due to such things as changes in the working voltage of bulb, an introduced higher-order dispersion, or a large tilt of the sample.

Increased Resolution White Light Interferometry

Interferometric methods that employ a monochromatic light source to detect the phase of the fringes (*see* **Interferometry:** Phase Measurement Interferometry) can achieve about 10 times better (0.3 nanometers versus 3 nanometers) vertical resolution than the WLI methods described so far. Combining coherence position detection from WLI to determine fringe order with the phase detection of phase shifting techniques allows for the measurement of samples with height discontinuities larger than 160 nanometers with the resolution and accuracy of phase shifting interferometry (PSI). This combination is particularly well-suited for determining the shape of smooth surfaces with large height differences such as binary diffractive optics or micro-electromechanical systems (MEMS) (see Figure 15).

Using this combined method we obtain both a lower-resolution map of the envelope position and a higher-resolution map of the phase (position) of the zero order fringe. These maps may differ slightly due to effects similar to those discussed in sections *Changes in Envelope and Fringes Due to Reflection of Nondielectric* and *Changes in Envelope and Fringes Due to Dispersion* above. In interference microscopes, shifts in the envelope and fringe position may be introduced by field-dependent and chromatic aberrations of the system and the system's effective numerical aperture. These shifts can vary for different points on the tested surface, but for simplicity are assumed to be constant over the field. Some correction for these effects can be applied.

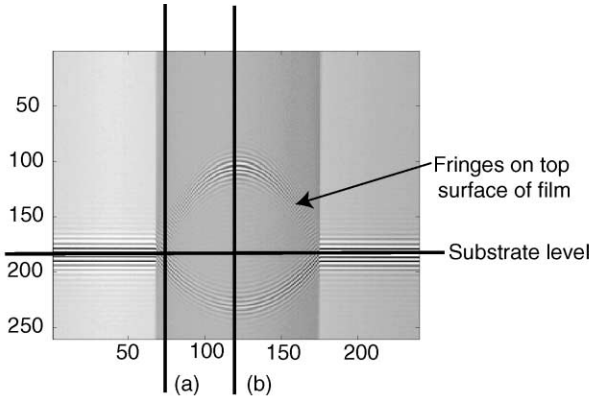

Figure 18 Fringes for (a) thin and (b) thick film.

Film Thickness Measurement

White light interferograms that are obtained in an optical profiler can be used to measure transparent film thicknesses because the character of the interferogram changes with the thickness of the film (Figure 18). Two different techniques, a thin or thick film technique, are used depending of the range of the film thickness. A thick film technique is used if two

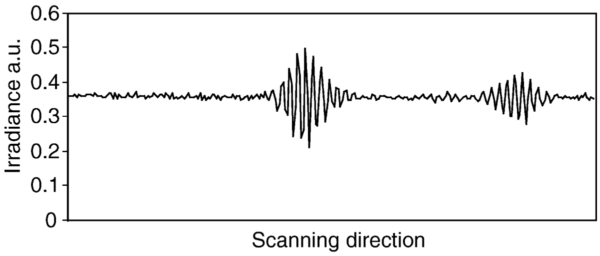

Figure 19 Thick film (see **Figure 18b**) irradiance.

sets of best contrast white light fringes from each interface are clearly separated, meaning that no interference occurs between the wavefronts reflected from the top and bottom surfaces of the film. A thin film technique is employed when the two sets of fringes overlap.

Thick film measurement

A simple technique for finding the relative position of the peaks of the fringe envelopes can be implemented to find the thickness of a film. Figure 19 shows two clearly separated sets of fringes formed for the air/film and film/substrate interfaces. Whether these two sets of fringes are separated or not depends mainly on the geometrical thickness of the film and its group index of refraction, which is determined by the dispersion of the film. The typical range of measurable film thicknesses runs from 3 to 150 microns, depending on the dispersion of the film. This measurement allows for the detection of flows on the surface and interface of the film. Similar principles based on finding the position of the coherence envelopes are used for distance sensing, thickness measurement of plates, and the cornea of the eye, in low coherence reflectometry and structure measurements in biological samples in optical coherence tomography.

Thin film measurement

For thicknesses of five microns down to tens of nanometers, the white light interferogram is created from the interference between the beams reflected from the reference mirror and the two beams reflected from the thin film layer. Once the interferogram is registered while the objective is scanned vertically with a constant velocity, the spectral phase is calculated by applying a Fourier transform to the measured signal at each pixel as described in the section *Fourier Analysis of White Light Interferograms* above. The phase slope is subtracted and the dispersion of the system needs to be known. The spectral phase for the thin film interference has the form of a polynomial; thus, the polynomial for the chosen film model (n and k) is fitted, and regression analysis is used to find the best fit and, therefore, the film thickness (Figure 20).

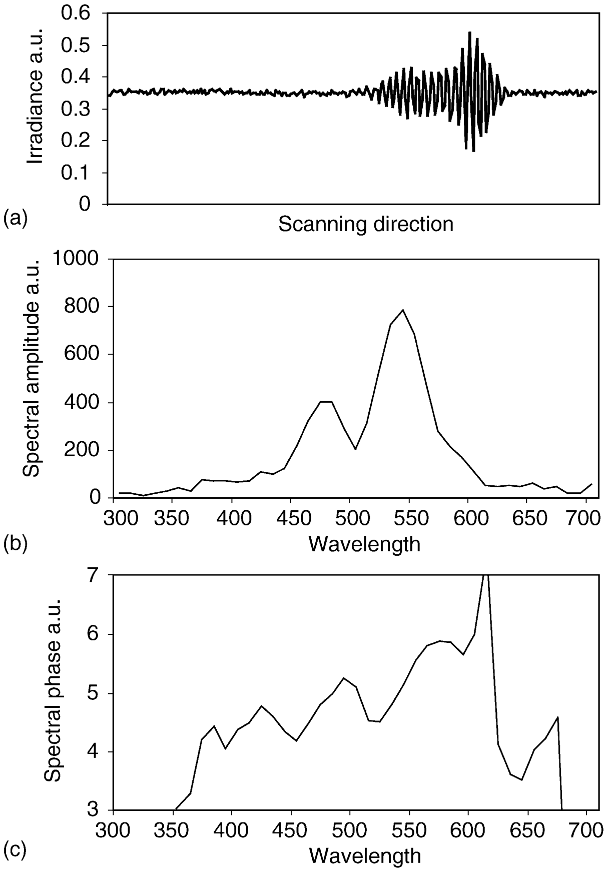

Figure 20 Thin film (see **Figure 18a**): (a) irradiance, (b) spectral amplitude and (c) spectral phase.

Spatial Coherence Effects in the Interference Microscope

So far we have been discussing temporal coherence effects, but everywhere that the source has a wavelength bandwidth, spatial coherence may play an important role. In an interference microscope for surface topography measurement, the size of the aperture of the condenser is on the order of the aperture of the objective that illuminates the surface with a wide range of angles. For large angles, which are determined by the numerical aperture (NA) of the objective, where NA = 0.5–0.95, the combination of the influence of the spatial and temporal coherence is clearly visible. The additional spatial effects include a reduction of the fringe envelope width and a decrease in fringe spacing corresponding to k_0. Calibration of fringe spacing is typically done on the system and accounts for spatial effects as well as for any uncertainty of k_0 due to the source spectra, the working temperature of the source, the spectral response of the detector and the influence of other factors like the intensity distribution in the illuminating aperture. The spatial influences can be reduced by stopping down the condenser, which causes an increase in the contrast of the fringes.

Rough Surfaces

Rough surfaces are difficult to measure using interferometric techniques, but under certain coherence conditions white light interference can do the job. For rough surfaces, if the microstructure of the object is not resolved by the imaging system, speckles, rather than fringes, are observed. Each speckle has a random phase which is approximately constant in the whole speckle area. If the rough surface is scanned through focus, each individual speckle exhibits the intensity modulation that is typical for WLI. These speckles enable the measurement, but they also introduce noise proportional to the roughness of the measured surface.

Despite the noise that speckles introduce into the WLI measurement, WLI has an advantage because it rejects the light that has undergone scattering outside of a small sample volume, thus allowing precise noninvasive measurement of object structure, even in dense media.

Applications

WLI is used in many disciplines and instruments such as:

- Fourier transform spectroscopy – source and material properties;
- Michelson stellar interferometer – angular size of star, binary stars, delay measurement in optical paths of interferometer;
- Shearing interferometry – structure measurement;
- DIC Nomarski interferometry – structure measurement;
- Speckle interferometry – structure measurement;
- Holography – structure measurement;
- Optical sensors – temperature, pressure, distance;
- Optical coherence tomography – structure measurement.

See also

Coherence: Coherence and Imaging; Overview. **Holography, Techniques:** Holographic Interferometry. **Interferometry:** Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Born M and Wolf E (1997) *Principles of Optics*. New York: Cambridge University Press.

Creath K (1997) *Sampling requirements for white light interferometry*. Fringe'97 Proceedings of the 3rd International Workshop on Automatic Processing of Fringe Patterns, pp. 52–59.

Danielson BL and Boisbert CY (1991) Absolute optical ranging using low coherence interferometry. *Applied Optics* 30: 2975–2979.

Davis SP, Abrams MC and Brault JW (2001) *Fourier Transform Spectrometry*. London: Academic Press.

de Groot P and Deck L (1995) Surface profiling by analysis of white light interferograms in the spatial frequency domain. *Journal of Modern Optics* 42: 389–401.

Dresdel T, Hausler G and Venzke H (1992) Three dimensional sensing of rough surfaces by coherence radar. *Applied Optics* 31(7): 919–925.

Harasaki A, Schmit J and Wyant JC (2001) Offset envelope position due to phase change on reflection. *Applied Optics* 40: 2102–2106.

Hariharan P (1992) *Basics of Interferometry*. New York: Academic Press.

Hariharan P, Larkin KG and Roy M (1994) The geometric phase: interferometric observation with white light. *Journal of Modern Optics* 41: 663–667.

Larkin KG (1996) Neat nonlinear algorithm for envelope detection in white light interferometry. *Journal of the Optical Society of America A* 13: 832–842.

Michelson AA and Benoit JR (1895) Détermination expérimentale de la valeur du métre en longueurs d'ondes lumineuses. *Travaux et Mémoires des Bureau International des Poids et Mesures* 11: 1.

Olszak AG and Schmit JG (2003) High stability white light interferometry with reference signal for real time correction of scanning errors. *Optical Engineering* 42(1): 54–59.

Park M-C and Kim S-W (2000) Direct quadratic polynomial fitting for fringe peak detection of white light scanning interferograms. *Optical Engineering* 39: 952–959.

Pavlicek P and Soubusta J (2004) Measurement of the influence of dispersion on white light interferometry. *Applied Optics* 43: 766–770.

Schmit JG and Olszak A (2002) High precision shape measurement by white light interferometry with real time scanner error correction. *Applied Optics* 41(28): 5943–5950.

Sheppard CJR and Larkin KG (1995) Effects of numerical aperture on interference fringe spacing. *Applied Optics* 34: 4731–4734.

Sheppard CJR and Roy M (2003) Low-coherence interference microscopy. In: Torok P and Kao F-J (eds) *Optical Imaging and Microscopy*, pp. 257–273. Berlin, Heidelberg, New York: Springer-Verlag.

Steel W (1987) *Interferometry*. New York: Cambridge University Press.

Turyshew SG (2003) Analytical modeling of the white light fringe. *Applied Optics* 42(1): 71–90.

L

LASERS

Contents

Carbon Dioxide Laser
Dye Lasers
Edge Emitters
Excimer Lasers
Free Electron Lasers
Metal Vapor Lasers
Noble Gas Ion Lasers
Optical Fiber Lasers
Organic Semiconductors and Polymers
Planar Waveguide Lasers
Semiconductor Lasers
Up-Conversion Lasers

Carbon Dioxide Laser

C R Chatwin, University of Sussex, Brighton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article gives a brief history of the development of the laser and goes on to describe the characteristics of the carbon dioxide laser and the molecular dynamics that permit it to operate at comparatively high power and efficiency. It is these commercially attractive features and its low cost that has led to its adoption as one of most popular industrial power beams. This outline also describes the main types of carbon dioxide laser and briefly discusses their characteristics and uses.

Brief History

In 1917 Albert Einstein developed the concept of stimulated emission which is the phenomenon used in lasers. In 1954 the MASER (Microwave Amplification by Stimulated Emission of Radiation) was the first device to use stimulated emission. In that year Townes and Schawlow suggested that stimulated emission could be used in the infrared and optical portions of the electromagnetic spectrum. The device was originally termed the optical maser, this term being dropped in favor of LASER, standing for: Light Amplification by Stimulated Emission of Radiation. Working against the wishes of his manager at Hughes Research Laboratories, the electrical engineer Ted Maiman created the first laser on the 15 May 1960. Maiman's flash lamp pumped ruby laser produced pulsed red electromagnetic radiation at a wavelength of 694.3 nm. During the most active period of laser systems discovery Bell Labs made a very significant contribution. In 1960, Ali Javan, William Bennet and Donald Herriot produced the first Helium Neon laser, which was the first continuous wave (CW) laser operating at 1.15 μm. In 1961, Boyle and Nelson developed a continuously operating Ruby laser and in 1962, Kumar Patel, Faust, McFarlane and Bennet discovered five noble gas lasers and lasers using oxygen mixtures. In 1964, C.K.N. Patel created the high-power carbon dioxide laser operating at 10.6 μm. In 1964, J.F. Geusic and R.G. Smith produced the first Nd:Yag laser using neodymium doped yttrium aluminum garnet crystals and operating at 1.06 μm.

Characteristics

Due to its operation between low lying vibrational energy states of the CO_2 molecule, the CO_2 laser has a high quantum efficiency, ~40%, which makes it extremely attractive as a high-power industrial materials processing laser (1 to 20 kW), where energy and running costs are a major consideration. Due to the requirement for cooling to retain the population inversion, the efficiency of electrical pumping and optical losses – commercial systems have an overall efficiency of approximately 10%. Whilst this may seem low, for lasers this is still a high efficiency. The CO_2 laser is widely used in other fields, for example, surgical applications, remote sensing, and measurement. It emits infrared radiation with a wavelength that can range from 9 μm up to 11 μm. The laser transition may occur on one of two transitions: $(00^01) \rightarrow (10^00)$, $\lambda = 10.6$ μm; $(00^01) \rightarrow (02^00)$, $\lambda = 9.6$ μm, see Figure 1. The 10.6 μm transition has the maximum probability of oscillation and gives the strongest output; hence, this is the usual wavelength of operation, although for specialist applications the laser can be forced to operate on the 9.6 μm line.

Figure 1 illustrates an energy level diagram with four vibrational energy groupings that include all the significantly populated energy levels. The internal relaxation rates within these groups are considered to be infinitely fast when compared with the rate of energy transfer between these groups. In reality the internal relaxation rates are at least an order of magnitude greater than the rates between groups.

Excitation of the upper laser level is usually provided by an electrical glow discharge. However, gas dynamic lasers have been built where expanding a hot gas through a supersonic nozzle creates the population inversion; this creates a nonequilibrium region in the downstream gas stream with a large population inversion, which produces a very high-power output beam (135 kW – Avco Everett Research Lab). For some time the gas dynamic laser was seriously considered for use in the space-based Strategic Defence Initiative (SDI-USA). The gas mixture used in a CO_2 laser is usually a mixture of carbon dioxide, nitrogen, and helium. The proportions of these gases varies from one laser system to another, however, a typical mixture is 10%-CO_2;

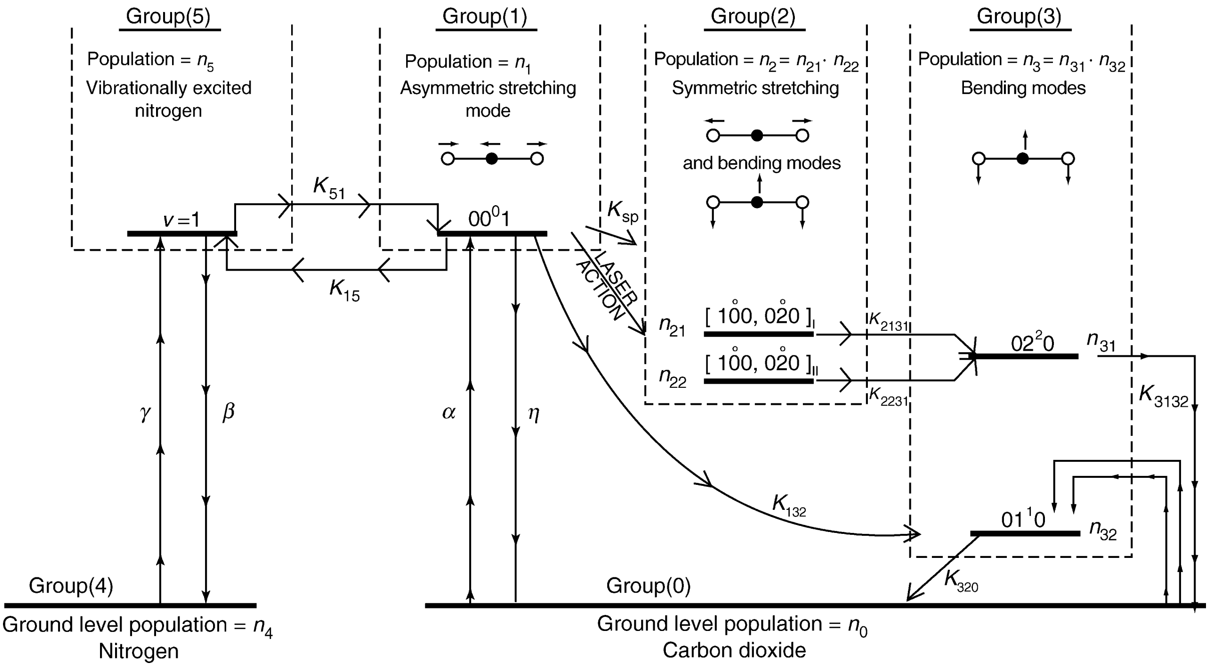
Group(5)
Population = n_5
Vibrationally excited nitrogen
Group(1)
Population = n_1
Asymmetric stretching mode
Group(2)
Population = $n_2 = n_{21} \cdot n_{22}$
Symmetric stretching
and bending modes
Group(3)
Population = $n_3 = n_{31} \cdot n_{32}$
Bending modes
K_{51}
$v=1$
00^01
K_{sp}
K_{15}
LASER ACTION
n_{21}
$[1\overset{\circ}{0}0, 0\overset{\circ}{2}0]_I$
K_{2131}
02^20
n_{31}
n_{22}
$[1\overset{\circ}{0}0, 0\overset{\circ}{2}0]_{II}$
K_{2231}
K_{3132}
γ
β
α
η
K_{132}
01^10
n_{32}
K_{320}
Group(4)
Group(0)
Ground level population = n_4
Nitrogen
Ground level population = n_0
Carbon dioxide

Figure 1 Six level model used for the theoretical description of CO_2 laser action.

10%-N_2; 80%-He. Helium plays a vital role in the operation of the CO_2 laser in that it maintains the population inversion by depopulating the lower laser level by nonradiative collision processes. Helium is also important for stabilization of the gas discharge; furthermore it greatly improves the thermal conductivity of the gas mixture, which assists in the removal of waste heat via heat exchangers.

Small quantities of other gases are often added to commercial systems in order to optimize particular performance characteristics or stabilize the gas discharge; for brevity we only concern ourselves here with this simple gas mixture.

Molecular Dynamics

Direct Excitation and De-excitation

It is usual for excitation to be provided by an electrical glow discharge. The direct excitation of carbon dioxide (CO_2) and nitrogen (N_2) ground state molecules proceeds via inelastic collisions with fast electrons. The rates of kinetic energy transfer are α and γ, respectively, and are given by eqns [1] and [2]:

$$\alpha = \frac{F_{\mathrm{CO_2}} \times \mathrm{IP}}{E_{00^01} \times n_0} (\mathrm{sec}^{-1}) \qquad [1]$$

$$\gamma = \frac{F_{\mathrm{N_2}} \times \mathrm{IP}}{E_{v=1} \times n_4} (\mathrm{sec}^{-1}) \qquad [2]$$

where:

$F_{\mathrm{CO_2}}$ = Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}, n_0 is the CO_2 ground level population density;

$F_{\mathrm{N_2}}$ = Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$; and n_4 is the N_2 ground level population density.

The reverse process of the above occurs when molecules lose energy to the electrons and the electrons gain an equal amount of kinetic energy; the direct de-excitation rates are given by η and β, eqns [3] and [4], respectively:

$$\eta = \alpha \times \exp\left(\frac{E_{00^01}}{E_e}\right) (\mathrm{sec}^{-1}) \qquad [3]$$

$$\beta = \gamma \times \exp\left(\frac{E_{v=1}}{E_e}\right) (\mathrm{sec}^{-1}) \qquad [4]$$

where E_e is the average electron energy in the discharge.

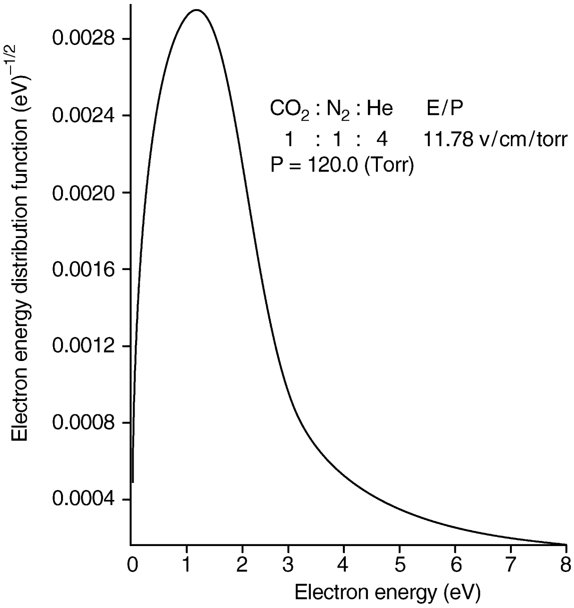

Figure 2 Electron energy distribution function.

E_e, $F_{\mathrm{CO_2}}$, and $F_{\mathrm{N_2}}$ are obtained by solution of the Boltzmann transport equation (BTE); the average electron energy can be optimized to maximize the efficiency ($F_{\mathrm{CO_2}}$, $F_{\mathrm{N_2}}$) with which electrical energy is utilized to create a population inversion. Hence, the discharge conditions required to maximize efficiency can be predicted from the transport equation. Figure 2 shows one solution of the BTE for the electron energy distribution function.

Resonant Energy Transfer

Resonant energy transfer between the CO_2 (00^01) and N_2 ($v = 2$) energy levels (denoted 1 and 5 in Figure 1) proceeds via excited molecules colliding with ground state molecules. A large percentage of the excitation of the upper laser level takes place via collisions between excited N_2 molecules and ground state CO_2 molecules. The generally accepted rate of this energy transfer is given by eqns [5] and [6]:

$$K_{51} = 19\,000\, P_{\mathrm{CO_2}}\ (\mathrm{sec}^{-1}) \qquad [5]$$

$$K_{15} = 19\,000\, P_{\mathrm{N_2}}\ (\mathrm{sec}^{-1}) \qquad [6]$$

where $P_{\mathrm{CO_2}}$ and $P_{\mathrm{N_2}}$ are the respective gas partial pressures in Torr.

Hence, CO_2 molecules are excited into the upper laser level by both electron impact and impact with excited N_2 molecules. The contribution from N_2 molecules can be greater than 40% depending on the discharge conditions.

Collision–Induced Vibrational Relaxation of the Upper and Lower Laser Levels

The important vibrational relaxation processes are illustrated by Figure 1 and can be evaluated from eqns [7–10]; where the subscripts refer to the rate between energy levels 1 and 32; 21 and 31; 22 and 31; 32 and 0, respectively:

$$K_{132} = 367\, P_{\mathrm{CO_2}} + 110\, P_{\mathrm{N_2}} + 67\, P_{\mathrm{He}}\ (\mathrm{sec}^{-1}) \qquad [7]$$

$$K_{2131} = 6 \times 10^5\, P_{\mathrm{CO_2}}\ (\mathrm{sec}^{-1}) \qquad [8]$$

$$K_{2231} = 5.15 \times 10^5\, P_{\mathrm{CO_2}}\ (\mathrm{sec}^{-1}) \qquad [9]$$

$$K_{320} = 200\, P_{\mathrm{CO_2}} + 215\, P_{\mathrm{N_2}} + 3270\, P_{\mathrm{He}}\ (\mathrm{sec}^{-1}) \qquad [10]$$

K_{132}, K_{2131}, and K_{2231} are vibration/vibration transfer rates and K_{320} is a vibration/translation transfer rate. Note the important effect of helium on eqn [10]; helium plays a major role in depopulating the lower laser level, thus enhancing the population inversion. P_{He} is the partial pressure of helium in Torr.

Radiative Relaxation

Spontaneous radiative decay is not a major relaxation process in the CO_2 laser but it is responsible for starting laser action via spontaneous emission. The Einstein '*A*' coefficient for the laser transition is given by eqn [11]:

$$A = (\tau_{\mathrm{sp}})^{-1} = 0.213\ (\mathrm{sec}^{-1}) \qquad [11]$$

Gain

The gain (g) is evaluated from the product of the absorption coefficient (σ) and the population inversion, eqn [12]:

$$g = \sigma\left(n_{00^01} - \frac{g_1}{g_2} n_{10^00}\right)\ \mathrm{cm}^{-1} \qquad [12]$$

For most commercial laser systems the absorption coefficient is that for high-pressure collision-broadening ($P > 5.2$ Torr) where the intensity distribution function describing the line shape is Lorentzian. The following expression describes the absorption coefficient, eqn [13]:

$$\sigma = \frac{692.5}{T n_{\mathrm{CO_2}}\left(1 + 1.603\dfrac{n_{\mathrm{N_2}}}{n_{\mathrm{CO_2}}} + 1.4846\dfrac{n_{\mathrm{He}}}{n_{\mathrm{CO_2}}}\right)}(\mathrm{cm}^2) \qquad [13]$$

where T is the absolute temperature and n refers to the population density of the gas designated by the subscript.

This expression takes account of the different constituent molecular velocity distributions and different collision cross-sections for $CO_2 \rightarrow CO_2$, $N_2 \rightarrow CO_2$ and $He \rightarrow CO_2$ type collisions. Equation [13] also takes account of the significant line broadening effect of helium.

Neglecting the unit change in rotational quantum number, the energy level degeneracies g_1 and g_2 may be dropped. n_{10^00} is partitioned such that $n_{10^00} = 0.1452 n_2$ and eqn [12] can be re-cast as eqn [14]:

$$g = \sigma(n_1 - 0.1452 n_2)\ \mathrm{cm}^{-1} \qquad [14]$$

where n_1 and n_2 are the population densities of energy groups '1' and '2' respectively.

Stimulated Emission

Consider a laser oscillator with two plane mirrors, one placed at either end of the active gain medium, with one mirror partially transmitting (see Figure 4). Laser action is initiated by spontaneous emission that happens to produce radiation whose direction is normal to the end mirrors and falls within the resonant modes of the optical resonator. The rate of change of photon population density (I_{p}) within the laser cavity can be written as eqn [15]:

$$\frac{\mathrm{d}I_{\mathrm{P}}}{\mathrm{d}t} = I_{\mathrm{p}} c g - \frac{I_{\mathrm{p}}}{T_0} \qquad [15]$$

where the first term on the right-hand side accounts for the effect of stimulated emission and the second term represents the number of photons that decay out of the laser cavity, T_0 is the photon decay time, given by eqn [16], and is defined as the average time a photon remains inside the laser cavity before being lost either through the laser output window or due to dispersion; if dispersion is ignored, I_{p}/T_0, is the laser output:

$$T_0 = \frac{2L}{c \log_{\mathrm{e}}\left(\dfrac{1}{R_{\mathrm{B}} R_{\mathrm{F}}}\right)} \qquad [16]$$

where L is the distance between the back and the front mirrors, which have reflectivities of R_{B} and R_{F}, respectively. The dominant laser emission occurs on a rotational–vibrational P branch transition $P(22)$, that is $(J = 21) \rightarrow (J = 22)$ line of the $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\ \mu\mathrm{m}$ transition, where J is the rotational quantum number. The rotational level relaxation rate is so rapid that equilibrium is maintained between rotational levels so that they feed all their energy through the $P(22)$ transition. This model simply assumes constant intensity, basing laser performance on the performance of an average unit volume. By introducing the stimulated emission term into the molecular rate equations, which describe the rate of transfer of molecules between the various energy levels illustrated in Figure 1, a set of molecular rate eqns [17–21] can be written that permit simulation of the performance of a carbon dioxide laser:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = \alpha n_0 - \eta n_1 + K_{51}n_5 - K_{15}n_1 - K_{\mathrm{sp}}n_1 - K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_{\mathrm{e}} n_{32}\right) - I_{\mathrm{p}}cg \qquad [17]$$

$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = K_{\mathrm{sp}}n_1 + I_{\mathrm{p}}cg - K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) - K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) \qquad [18]$$

$$\frac{\mathrm{d}n_3}{\mathrm{d}t} = 2K_{2131}\left(n_{21} - \left[\frac{n_{21}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) - 2K_{2231}\left(n_{22} - \left[\frac{n_{22}}{n_{31}}\right]_{\mathrm{e}} n_{31}\right) + K_{132}\left(n_1 - \left[\frac{n_1}{n_{32}}\right]_{\mathrm{e}} n_{32}\right) - K_{320}\left(n_{32} - \left[\frac{n_{32}}{n_0}\right]_{\mathrm{e}} n_0\right) \qquad [19]$$

$$\frac{\mathrm{d}n_5}{\mathrm{d}t} = \gamma n_4 - \beta n_5 - K_{51}n_5 + K_{15}n_1 \qquad [20]$$

$$\frac{\mathrm{d}I_{\mathrm{p}}}{\mathrm{d}t} = I_{\mathrm{p}}cg - \frac{I_{\mathrm{p}}}{T_0} \qquad [21]$$

The terms in square brackets ensure that the system maintains thermodynamic equilibrium; subscript ‘e’ refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium. The set of five simultaneous differential equations can be solved using a Runge–Kutta method. They can provide valuable performance prediction data that is helpful in optimizing laser design, especially when operated in the pulsed mode. Figure 3a illustrates some simulation results for the transverse flow laser shown in Figure 9. The results illustrate the effect of altering the gas mixture and how this can be used to control the gain switched spike that would result in unwelcome work piece plasma generation if allowed to become too large. Figure 3b shows an experimental laser output pulse from the high pulse repetition frequency (prf–5kHz) laser illustrated in Figure 9. This illustrates that even a quite basic physical model can give a good prediction of laser output performance.

Optical Resonator

Figure 4 shows a simple schematic of an optical resonator. This simple optical system consists of two mirrors (the full reflector is often a water cooled gold coated copper mirror), which are aligned to be orthogonal to the optical axis that runs centrally along the length of the active gain medium in which there is a population inversion. The output coupler is a partial reflector (usually dielectrically coated zinc selenide – ZnSe–that may be edge cooled) so that some of the electromagnetic radiation can escape as an output beam. The ZnSe output coupler has a natural reflectivity of about 17% at each air–solid interface. For high power lasers (2 kW) 17% is sufficient for laser operation; however, depending on the required laser performance the inside surface is often given a reflective coating. The reflectivity of the inside face depends on the balance between the gain (eqn [14]), the output power and the power stability requirements. The outside face of the partial reflector must be anti-reflection (AR) coated.

Spontaneous emission occurs within the active gain medium and radiates randomly in all directions; a fraction of this spontaneous emission will be in the same direction as the optical axis, perpendicular to the end mirrors, and will also fall into a resonant mode of the optical resonator. Spontaneous emission photons interact with CO_2 molecules in the excited upper laser level, excited state (00^01), which stimulates these molecules to give up a quanta of vibrational energy as photons via the radiative transition $(00^01) \rightarrow (10^00)$, $\lambda = 10.6\,\mu\mathrm{m}$. The radiation given up will have exactly the same phase and direction as the stimulating radiation and thus will be coherent with it. The reverse process of absorption also occurs, but so long as there is a population inversion there will be a net positive output. This process is called light amplification by stimulated emission of radiation (LASER). The mirrors continue to redirect the photons parallel to the optical axis and so long as the population inversion is not depleted, more and more photons are stimulated by stimulated emission which dominates the process and also dominates the spontaneous emission, which is important to initiate laser action.

Light emitted by lasers contains several optical frequencies, which are a function of the different modes of the optical resonator; these are simply the standing wave patterns that can exist within the resonator structure. There are two types of

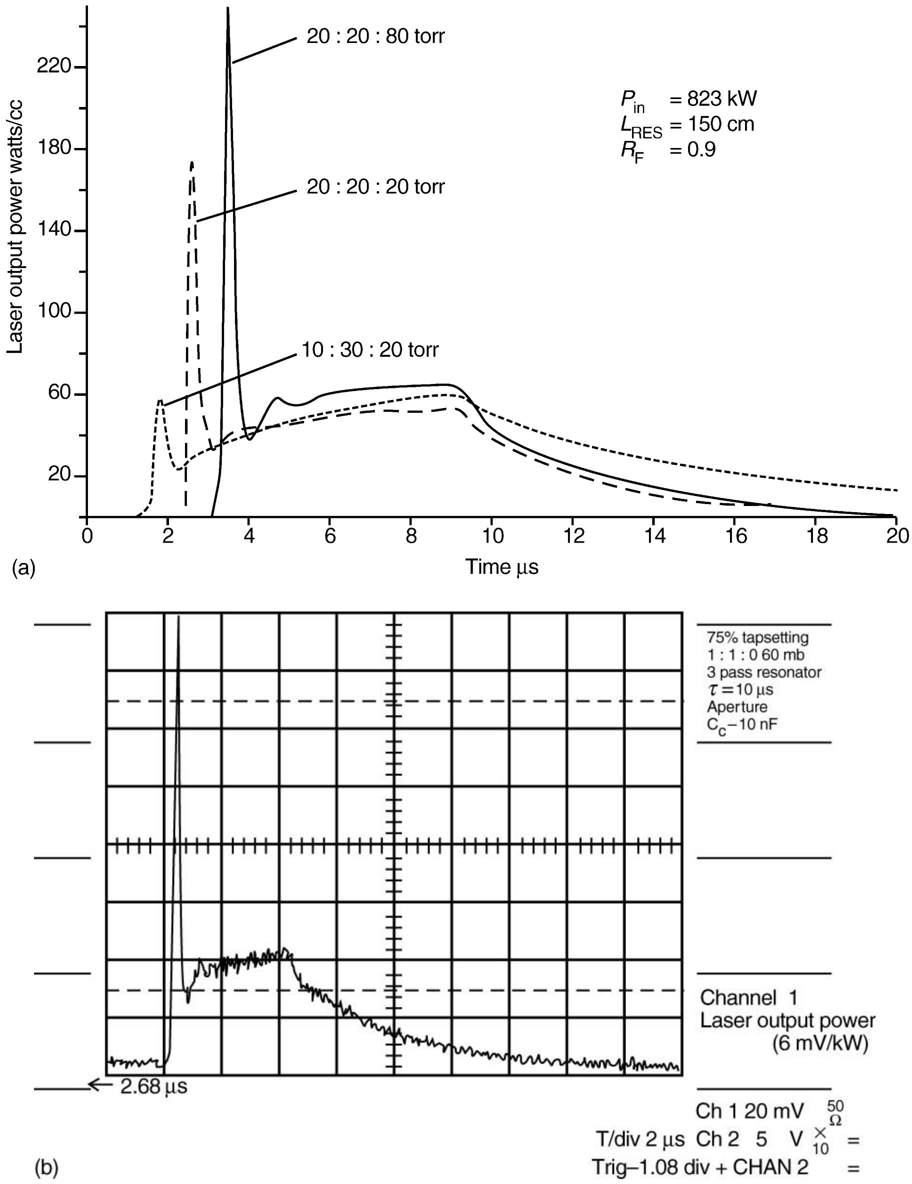

Figure 3 (a) Predicted output pulses for transverse flow CO_2 laser for different gas mixtures, (b) Experimental output pulse from transverse flow CO_2 laser.

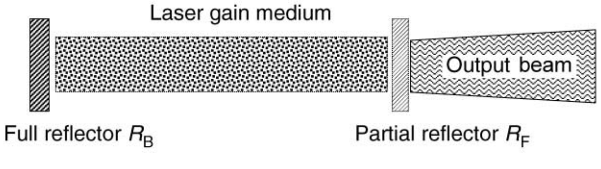

Figure 4 Optical resonator.

resonator modes: longitudinal and transverse. Longitudinal modes differ from one another in their frequency of oscillation whereas transverse modes differ from one another in their oscillation frequency and field distribution in a plane orthogonal to the direction of propagation. Typically CO_2 lasers have a large number of longitudinal modes; in CO_2 laser applications these are of less interest than the transverse modes, which determine the transverse beam intensity and the nature of the beam when focused. In cylindrical coordinates the transverse modes are labelled TEM_{pl}, where subscript '*p*' is the number of radial nodes and '*l*' is the number of angular nodes. The lowest order mode is the TEM_{00}, which has a Gaussian-like intensity profile with its maximum on the beam axis. A light beam emitted from an optical resonator with a Gaussian profile is said to be operating in the 'fundamental mode' or the TEM_{00} mode. The decrease in irradiance (I) with distance '*r*' from the axis (I_0) of the Gaussian beam is described by eqn [22]:

$$I(r) = I_0 \exp(2r^2/w^2) \qquad [22]$$

where w is the radial distance, where the power density is decreased to $1/e^2$ of its axial value. Ideally a commercial laser should be capable of operation in the fundamental mode as, with few exceptions, this results in the best performance in applications. Laser cutting benefits from operation in the fundamental mode; however, welding or heat treatment applications may benefit from operation with higher-order modes. Output beams are usually controlled to be linearly or circularly polarized, depending upon the requirements of the application. For materials processing applications the laser beam is usually focused via a water cooled ZnSe lens or, for very high power lasers, a parabolic gold coated mirror. Welding applications will generally use a long focal length lens and cutting applications will use a short focal length, which generates a higher irradiance at the work piece than that necessary for welding. The beam delivery optics are usually incorporated into a nozzle assembly that can deliver cooling water and assist gases for cutting and anti-oxidizing shroud gases for welding or surface engineering applications.

Laser Configuration

CO_2 lasers are available in many different configurations and tend to be classified on the basis of their physical form and the gas flow arrangement, both of which greatly affect the output power available and beam quality. The main categories are: sealed off lasers, waveguide lasers, slow axial flow, fast axial flow, diffusion cooled, transverse flow, transversely excited atmospheric lasers, and gas dynamic lasers.

Sealed-Off and Waveguide Lasers

Depopulation of the lower laser level is via collision with the walls of the discharge tube, so the attainable output power scales with the length of the discharge column and not its diameter. Output powers are in the range 5 W to 250 W. Devices may be constructed from concentric glass tubes with the inner tube providing the discharge cavity and the outer tube acting to contain water-cooling of the inner discharge tube. The inner tube walls act as a heat sink for the discharge thermal energy (see Figure 5). The DC electrical discharge is provided between a cathode and anode situated at either end of the discharge tube. A catalyst must be provided to ensure regeneration of CO_2 from CO. This may be accomplished by adding about 1% of H_2O to the gas mixture, or alternatively, recombination can be achieved via a hot (300 °C) Ni cathode, which acts as a catalyst. RF-excited all metal sealed-off tube systems can deliver lifetimes greater than 45 000 hours.

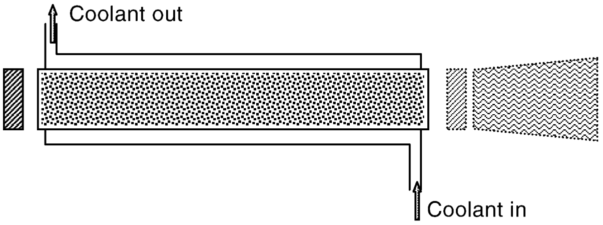

Figure 5 Schematic of sealed-off CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

Diffusion-cooled slab laser technology will also deliver reliable sealed operation for 20 000 hrs. Excitation of the laser medium occurs via RF excitation between two water-cooled electrodes. The water-cooled electrodes dissipate (diffusion cooled) the heat generated in the gas discharge. An unstable optical resonator provides the output coupling for such a device (see Figure 6). Output powers are in the range 5 W to 300 W and can be pulsed from 0 to 100 kHz. These lasers are widely used for marking, rapid prototyping, and cutting of nonmetals (paper, glass, plastics, ceramics) and metals.

Waveguide CO_2 lasers use small bore tubes (2–4 mm) made of BeO or SiO_2 where the laser radiation is guided by the tube walls. Due to the small tube diameter, a gas total pressure of 100 to 200 Torr is necessary, hence the gain per unit length is high. This type of laser will deliver 30 W of output power from a relatively short (50 cm long) compact sealed-off design; such a system is useful for microsurgery and scientific applications. Excitation can be provided from a longitudinal DC discharge or from an RF source that is transverse to the optical axis; RF excitation avoids the requirement for an anode and cathode and results in a much lower electrode voltage.

Slow Axial Flow Lasers

In slow flow lasers the gas mixture flows slowly through the laser cavity. This is done to remove the products of dissociation that will reduce laser efficiency or prevent it from operating at all, and the main contaminant is CO. The dissociated gases (mainly CO and O_2) can be recombined using a catalyst pack and then reused via continuous recirculation. Heat is removed via diffusion through the walls of the tube containing the active gain medium. The tube is frequently made of Pyrex glass with a concentric outer tube to facilitate water-cooling of the

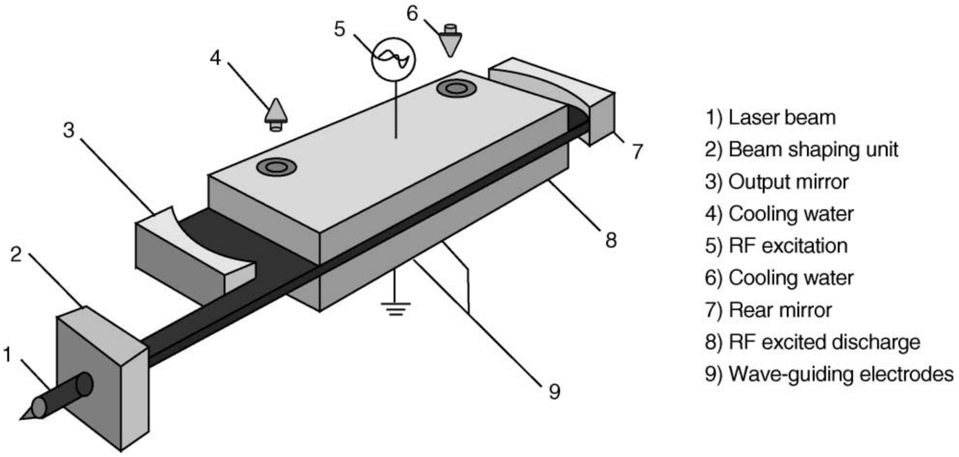

Figure 6 Schematic for sealed-off slab laser and diffusion cooled laser with RF excitation (courtesy of Rofin).

laser cavity (see Figure 7). Slow flow lasers operate in the power range 100 W to 1500 W, and tend to use a longitudinal DC electrical discharge which can be made to run continuously or pulsed if a thyratron switch is build into the power supply; alternatively, electrical power can be supplied via transverse RF excitation. The power scales with length, hence high power slow flow lasers have long cavities and require multiple cavity folds in order to reduce their physical size.

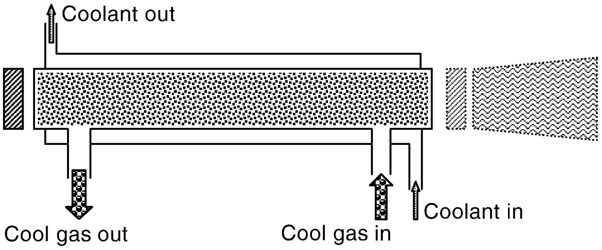

Figure 7 Slow flow CO_2 laser, approximately 100 W per meter of gain length, gas cooled by diffusion to the wall.

Fast Axial Flow Lasers

The fast axial flow laser, Figure 8, can provide output powers from 1 kW to 20 kW; it is this configuration that dominates the use of CO_2 lasers for industrial applications. Industrial lasers are usually in the power range 2–4 kW. The output power from these devices scales with mass flow, hence the gas mixture is recycled through the laser discharge region at sonic or supersonic velocities. Historically this was achieved using Rootes blowers to compress the gas upstream of the laser cavity. Rootes compressors are inherently inefficient and the more advanced laser systems utilize turbine compressors, which deliver greater efficiency and better laser stability. Rootes compressors can be a major source of vibration. With this arrangement heat exchangers are required to remove heat after the laser discharge region and also after the compressor stage, as the compression process heats up the laser gases. Catalyst packs are used to regenerate gases but some gas replacement is often required. These laser systems have short cavities and use folded stable resonator designs to achieve higher output powers with extremely high-quality beams that are particularly suitable for cutting applications. They also give excellent results when used for welding and surface treatments.

Fast axial flow lasers can be excited by a longitudinal DC discharge or a transverse RF discharge. Both types of electrical excitation are common. For materials processing applications it is often important to be able to run a laser in continuous wave (CW) mode or as a high pulse repetition rate (prf) pulsed laser and to be able to switch between CW and pulsed in real time; for instance, laser cutting of accurate internal corners is difficult using CW operation but very easy using the pulsed mode of operation. Both methods of discharge excitation can provide this facility.

Diffusion Cooled Laser

The diffusion-cooled slab laser is RF excited and gives an extremely compact design capable of delivering 4.5 kW pulsed from 8 Hz to 5 kHz prf or CW with good beam quality (see Figure 6). The optical resonator is formed by the front and rear mirrors and two parallel water cooled RF-electrodes. Diffusion cooling is provided by the RF-electrodes, removing the requirement for conventional gas recirculation via Rootes blowers or turbines.

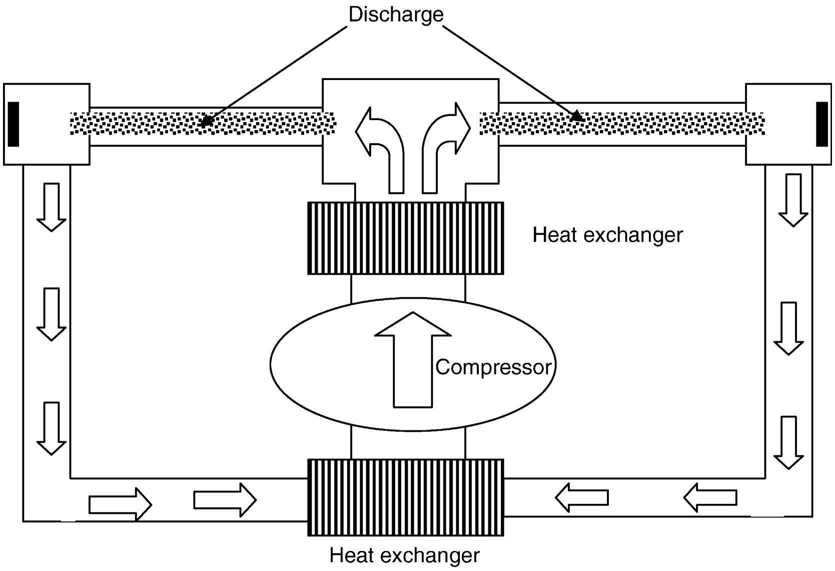

Figure 8 Fast axial flow carbon dioxide laser.

This design of laser results in a device with an extremely small footprint that has low maintenance and running costs. Applications include: cutting, welding, and surface engineering.

Fast Transverse Flow Laser

In the fast transverse flow laser (Figure 9a) the gas flow, electrical discharge, and the output beam are at right angles to each other (Figure 9b). The transverse discharge can be high voltage DC, RF, or pulsed up to 8 kHz (Figure 9c). Very high output power per unit discharge length can be obtained with an optimal total pressure (*P*) of ~100 Torr; systems are available delivering 10 kW of output power, CW or pulsed (see Figures 3a and b). The increase in total pressure requires a corresponding increase in the gas discharge electric field, *E*, as the ratio *E*/*P* must remain constant, since this determines the temperature of the discharge electrons, which have an optimum mean value (optimum energy distribution, Figure 2) for efficient excitation of the population inversion. With this high value of electric field, a longitudinal-discharge arrangement is impractical (500 kV for a 1 m discharge length); hence, the discharge is applied perpendicular to the optical axis. Fast transverse flow gas lasers provided the first multikilowatt outputs but tend to be expensive to maintain and operate. In order to obtain a reasonable beam quality, the output coupling is often obtained using a multipass unstable resonator. As the population inversion is available over a wide rectangular cross-section, this is a disadvantage of this arrangement and beam quality is not as good as that obtainable from fast axial flow designs. For this reason this type of laser is suitable for a wide range of welding and surface treatment applications.

Transversely Excited Atmospheric (TEA) Pressure

If the gas total pressure is increased above ~ 100 Torr it is difficult to sustain a stable glow discharge, because above this pressure instabilities degenerate into arcs within the discharge volume. This problem can be overcome by pulse excitation; using submicrosecond pulse duration, instabilities do not have sufficient time to develop; hence, the gas pressure can be increased above atmospheric pressure and the laser can be operated in a pulsed mode. In a mode locked format optical pulses shorter than 1 ns can be produced. This is called a TEA laser and with a transverse gas flow is capable of producing short high power pulses up to a few kHz repetition frequency. In order to prevent arc formation, TEA lasers usually employ ultraviolet or e-beam preionization of the gas discharge just prior to the main current pulse being applied via a thyratron switch. Output coupling is usually via an unstable resonator. TEA lasers are used for marking, remote sensing, range-finding, and scientific applications.

Conclusions

It is 40 years since Patel operated the first high power CO_2 laser. This led to the first generation of lasers which were quickly exploited for industrial laser materials processing, medical applications, defense,

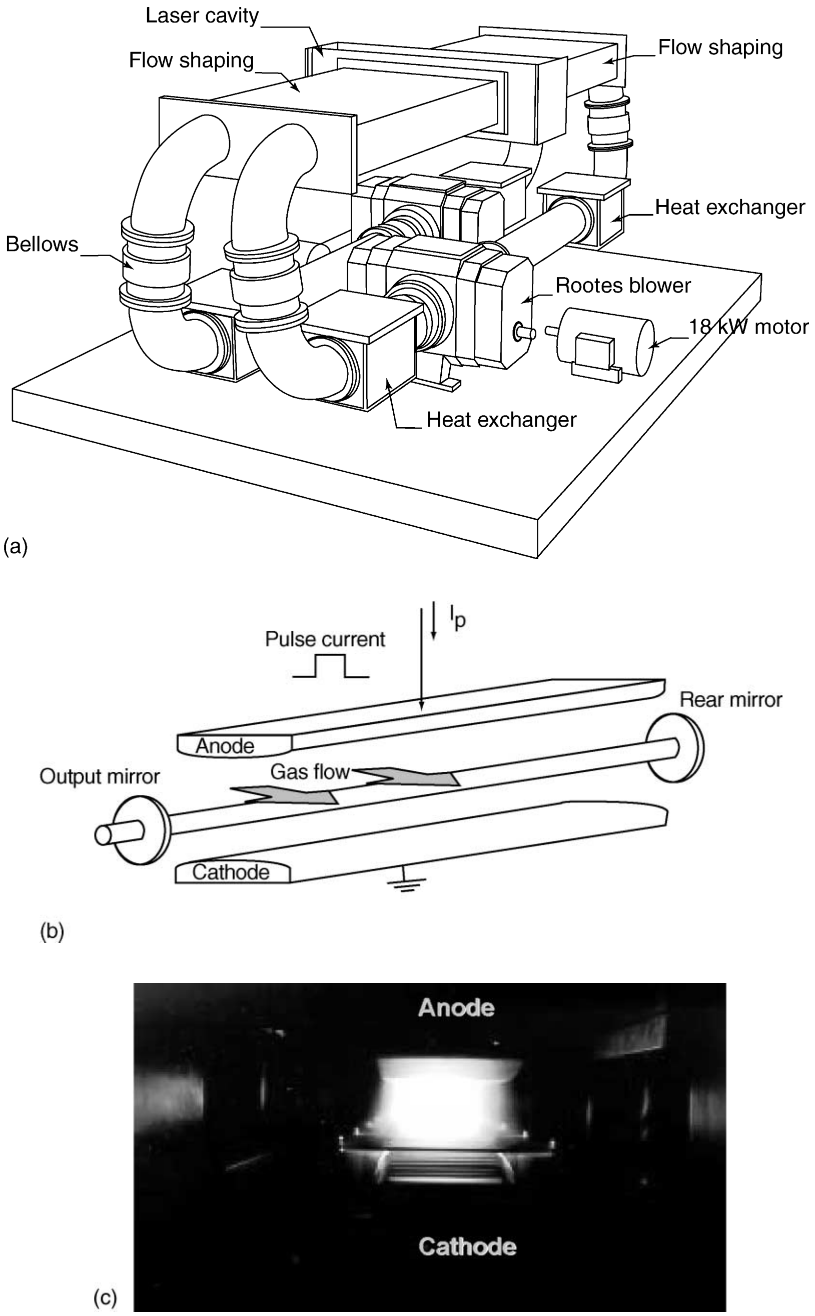
Laser cavity
Flow shaping
Flow shaping
Heat exchanger
Bellows
Rootes blower
18 kW motor
Heat exchanger
(a)
I_p
Pulse current
Rear mirror
Anode
Output mirror
Gas flow
Cathode
(b)
Anode
Cathode
(c)

Figure 9 (a) Transverse flow carbon dioxide laser gas recirculator, (b) Transverse flow carbon dioxide electrodes, (c) Transverse flow carbon dioxide gas discharge as seen from the output window.

and scientific research applications; however, the first generation of lasers were quite unreliable and temperamental. After many design iterations, the CO_2 laser has now matured into a reliable, stable laser source available in many different geometries and power ranges. The low cost of ownership of the latest generation of CO_2 laser makes them a very attractive commercial proposition for many industrial and scientific applications. Commercial lasers incorporate many novel design features that are beyond the scope of this article and are often peculiar to the particular laser manufacturer. This includes gas additives and catalysts that may be required to stabilize the gas discharge of a particular laser design; it is this optimization of the laser design that has produced such reliable and controllable low-cost performance from the CO_2 laser.

List of Units and Nomenclature

α	The direct excitation of carbon dioxide (CO_2) ground state molecules (sec^{-1})
β	The direct de-excitation of nitrogen (N_2) (sec^{-1})
γ	The direct excitation of nitrogen (N_2) ground state molecules (sec^{-1})
η	The direct de-excitation of carbon dioxide (CO_2) (sec^{-1})
λ	wavelength (m)
σ	absorption coefficient (cm^2)
τ	Electrical current pulse length (μs)
τ_{sp}	Spontaneous emission life time of the upper laser level (sec)
$A = (\tau_{sp})^{-1}$ $= 0.213(sec^{-1})$	The Einstein '*A*' coefficient for the laser transition
c	velocity of light ($cm\ sec^{-1}$)
C_c	Coupling capacitance (nF)
CO_2	Carbon dioxide
e	subscript 'e' refers to the fact that the populations in the square brackets are the values for thermodynamic equilibrium
E	Electric Field ($V\ cm^{-1}$)
F_{CO_2}	Fraction of the input power (IP) coupled into the excitation of the energy level E_{00^01}
F_{N_2}	Fraction of the input power (IP) coupled into the excitation of the energy level $E_{v=1}$
g	gain (cm^{-1})
g_1 and g_2	energy level degeneracy's of levels 1 and 2
He	Helium
I	beam irradiance ($W\ cm^{-2}$)
I_0	beam irradiance at the center of a Gaussian laser beam ($W\ cm^{-2}$)
I_p	Photon population density ($photons\ cm^{-3}$)
$\mathrm{I_p}$	Input current (A)
IP	Electrical input power ($W\ cm^{-3}$)
J	the rotational quantum number
K_{51}, K_{15}	Resonant energy transfer between the CO_2 (00^01) and $N_2(v = 2)$ energy levels proceeds via excited molecules colliding with ground state molecules (sec^{-1})
K_{132}, K_{2131}, K_{2231}	are vibration/vibration transfer rates (sec^{-1}) between energy levels 1 and 32; 21 and 31; 22 and 31, respectively (see Figure 1)
K_{320}	is a vibration/translation transfer rate (sec^{-1}) between energy levels 32 and 0 (see Figure 1)
K_{sp}, A	Spontaneous emission rate (sec^{-1})
L	The distance between the back and the front mirrors, which have reflectivity's of R_B and R_F (cm)
n	molecular population (molecules cm^{-3})
N_2	Nitrogen
P	Pressure (Torr)
P_{CO_2}, P_{He} and P_{N_2}	The respective gas partial pressures (Torr)
P_{in}	Electrical input power (kW)
r	radius of laser beam (cm)
R_B	Back mirror reflectivity
R_F	Front mirror reflectivity
t	time (sec)
T	Temperature (deg K)
T_0	the photon decay time (sec)
w	the radial distance where the power density is decreased to $1/e^2$ of its axial value

See also

Fiber and Guided Wave Optics: Overview. **Lasers:** Noble Gas Ion Lasers.

Further Reading

Anderson JD (1976) *Gasdynamic Lasers: An Introduction*. New York: Academic Press.

Chatwin CR, McDonald DW and Scott BF (1991) Design of a High P.R.F. Carbon Dioxide Laser for Processing High Damage Threshold Materials. Selected Papers on Laser Design, SPIE Milestone Series, pp. 425–433. Washington: SPIE Optical Engineering Press.

Cool AC (1969) Power and gain characteristics of high speed flow lasers. *Journal of Applied Physics* 40(9): 3563.

Crafer RC, Gibson AF, Kent MJ and Kimmit MF (1969) Time-dependent processes on CO_2 laser amplifiers. *British Journal of Applied Physics* 2(2): 183.

Gerry ET and Leonard AD (1966) Measurement of 10.6-μ CO_2 laser transition probability and optical broadening cross sections. *Applied Physics Letters* 8(9): 227.

Gondhalekar A, Heckenberg NR and Holzhauer E (1975) The mechanism of single-frequency operation of the hybrid CO_2 laser. *IEEE Journal of Quantum Electronics* QE-11(3): 103.

Gordiets BF, Sobolev NN and Shelepin LA (1968) Kinetics of physical processes in CO_2 lasers. *Soviet Physics JETP* 26(5): 1039.

Herzberg G (1945) *Molecular Spectra & Molecular Structure*. Infra-red and Raman Spectra of Polyatomic Molecules, vol. 2. New York: Van Nostrand.

Hoffman AL and Vlases GC (1972) A simplified model for predicting gain, saturation and pulse length for gas dynamic lasers. *IEEE Journal of Quantum Electronics* 8(2): 46.
Johnson DC (1971) Excitation of an atmospheric pressure CO_2–N_2–He laser by capacitor discharges. *IEEE Journal of Quantum Electronics* Q.E.-7(5): 185.
Koechner W (1988) *Solid State Laser Engineering*. Berlin: Springer Verlag.
Kogelnik H and Li T (1966) Laser beams and resonators. *Applied Optics* 5: 1550–1567.
Levine AK and De Maria AJ (1971). *Lasers*, vol. 3. New York: Marcel Dekkar, Chapter 3.
Moeller G and Ridgen JD (1965) *Applied Physics Letters* 7: 274.
Moore CB, Wood RE, Bei-Lok Hu and Yardley JT (1967) Vibrational energy transfer in CO_2 lasers. *Journal of Chemical Physics* 11: 4222.
Patel CKN (1964) *Physics Review Letters* 12: 588.
Siegman A (1986) *Lasers*. Mill Valley, California: University Science.
Smith K and Thomson RM (1978) *Computer Modeling of Gas Lasers*. New York and London: Plenum Press.
Sobolev NN and Sokovikov VV (1967) CO_2 lasers. *Soviet Physics USPEKHI* 10(2): 153.
Svelto O (1998) *Principles of Lasers*, 4th edn. New York: Plenum.
Tychinskii VP (1967) Powerful lasers. *Soviet Physics USPEKHI* 10(2): 131.
Vlases GC and Money WM (1972) Numerical modelling of pulsed electric CO_2 lasers. *Journal of Applied Physics* 43(4): 1840.
Wagner WG, Haus HA and Gustafson KT (1968) High rate optical amplification. *IEEE Journal of Quantum Electronics* Q.E.-4: 287.
Witteman W (1987) *The CO_2 Laser*. Springer Series in Optical Sciences, vol. 53.

Dye Lasers

F J Duarte, Eastman Kodak Company, New York, NY, USA

A Costela, Consejo Superior de Investigaciones Cientificas, Madrid, Spain

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Background

Dye lasers are the original tunable lasers. Discovered in the mid-1960s these tunable sources of coherent radiation span the electromagnetic spectrum from the near-ultraviolet to the near-infrared (Figure 1). Dye lasers spearheaded and sustained the revolution in atomic and molecular spectroscopy and have found use in many and diverse fields from medical to military applications. In addition to their extraordinary spectral versatility, dye lasers have been shown to oscillate from the femtosecond pulse domain to the continuous wave (cw) regime. For microsecond pulse emission, energies of up to hundreds of joules per pulse have been demonstrated. Further, operation at high pulsed repetition frequencies (prfs), in the multi-kHz regime, has provided average powers at kW levels. This unrivaled operational versatility is summarized in Table 1.

Dye lasers are excited by coherent optical energy from an excitation, or pump, laser or by optical energy from specially designed lamps called flash-lamps. Recent advances in semiconductor laser

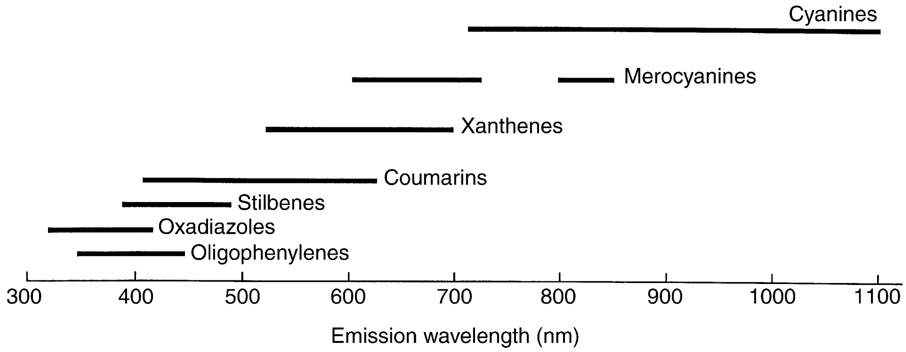

Figure 1 Approximate wavelength span from the various classes of laser dye molecules. Reproduced with permission from Duarte FJ (1995) *Tunable Laser Handbook*. New York: Academic Press.

Table 1 Emission characteristics of liquid dye lasers

Dye laser class	*Spectral coverage*[a]	*Energy per pulse*[b]	*Prf*[b]	*Power*[b]
Laser-pumped pulsed dye lasers	350–1100 nm	800 J[c]	13 kHz[d]	2.5 kW[d]
Flashlamp-pumped dye lasers	320–900 nm	400 J[e]	850 Hz[f]	1.2 kW[f]
CW dye lasers	370–1000 nm			43 W[g]

[a] Approximate range.
[b] Refers to maximum values for that particular emission parameter.
[c] Achieved with an excimer-laser pumped coumarin dye laser by Tang and colleagues, in 1987.
[d] Achieved with a multistage copper-vapor-laser pumped dye laser using rhodamine dye by Bass and colleagues, in 1992.
[e] Reported by Baltakov and colleagues, in 1974. Uses rhodamine 6G dye.
[f] Reported by Morton and Dragoo, in 1981. Uses coumarin 504 dye.
[g] Achieved with an Ar^+ laser pumped folded cavity dye laser using rhodamine 6G dye by Baving and colleagues, in 1982.

technology have made it possible to construct very compact all-solid-state excitation sources that, coupled with new solid-state dye laser materials, should bring the opportunity to build compact tunable laser systems for the visible spectrum. Further, direct diode-laser pumping of solid-state dye lasers should prove even more advantageous to enable the development of fairly inexpensive tunable narrow-linewidth solid-state dye laser systems for spectroscopy and other applications requiring low powers. Work on electrically excited organic gain media might also provide new avenues for further progress.

The literature of dye lasers is very rich and many review articles have been written describing and discussing traditional dye lasers utilizing liquid gain media. In particular, the books *Dye Lasers*, *Dye Laser Principles*, *High Power Dye Lasers*, and *Tunable Lasers Handbook* provide excellent sources of authoritative and detailed description of the physics and technology involved. In this article we offer only a survey of the operational capabilities of the dye lasers using liquid gain media in order to examine with more attention the field of solid-state dye lasers.

Brief History of Dye Lasers

1965: Quantum theory of dyes is discussed in the context of the maser (R. P. Feynman).
1966: Dye lasers are discovered (P. P. Sorokin and J. R. Lankard; F. P. Schäfer and colleagues).
1967: The flashlamp-pumped dye laser is discovered (P. P. Sorokin and J. R. Lankard; W. Schmidt and F. P. Schäfer).
1967–1968: Solid-state dye lasers are discovered (B. H. Soffer and B. B. McFarland; O. G. Peterson and B. B. Snavely).
1968: Mode-locking, using saturable absorbers, is demonstrated in dye lasers (W. Schmidt and F. P. Schäfer).
1970: The continuous-wave (cw) dye laser is discovered (O. G. Peterson, S. A. Tuccio, and B. B. Snavely).
1971: The distributed feedback dye laser is discovered (H. Kogelnik and C. V. Shank).
1971–1975: Prismatic beam expansion in dye lasers is introduced (S. A. Myers; E. D. Stokes and colleagues; D. C. Hanna and colleagues).
1972: Passive mode-locking is demonstrated in cw dye lasers (E. P. Ippen, C. V. Shank, and A. Dienes).
1972: The first pulsed narrow-linewidth tunable dye laser is introduced (T. W. Hänsch).
1973: Frequency stabilization of cw dye lasers is demonstrated (R. L. Barger, M. S. Sorem, and J. L. Hall).
1976: Colliding-pulse-mode locking is introduced (I. S. Ruddock and D. J. Bradley).
1977–1978: Grazing-incidence grating cavities are introduced (I. Shoshan and colleagues; M. G. Littman and H. J. Metcalf; S. Saikan).
1978–1980: Multiple-prism grating cavities are introduced (T. Kasuya and colleagues; G. Klauminzer; F. J. Duarte and J. A. Piper).
1981: Prism pre-expanded grazing-incidence grating oscillators are introduced (F. J. Duarte and J. A. Piper).
1982: Generalized multiple-prism dispersion theory is introduced (F. J. Duarte and J. A. Piper).
1983: Prismatic negative dispersion for pulse compression is introduced (W. Dietel, J. J. Fontaine, and J-C. Diels).
1987: Laser pulses as short as six femtoseconds are demonstrated (R. L. Fork, C. H. Brito Cruz, P. C. Becker, and C. V. Shank).
1994: First narrow-linewidth solid-state dye laser oscillator (F. J. Duarte).
1999–2000: Distributed feedback solid-state dye lasers are introduced (Wadsworth and colleagues; Zhu and colleagues).

Molecular Energy Levels

Dye molecules have large molecular weights and contain extended systems of conjugated double bonds. These molecules can be dissolved in an adequate organic solvent (such as ethanol, methanol, ethanol/water, and methanol/water) or incorporated into a solid matrix (organic, inorganic, or hybrid). These molecular gain media have a strong absorption generally in the visible and ultraviolet regions, and exhibit large fluorescence bandwidths covering the entire visible spectrum. The general energy level diagram of an organic dye is shown in Figure 2. It consists of electronic singlet and triplet states with each electronic state containing a multitude of overlapping vibrational–rotational levels giving rise to broad continuous energy bands. Absorption of visible or ultraviolet pump light excites the molecules from the ground state S_0 into some rotational–vibrational level belonging to an upper excited singlet state, from where the molecules decay nonradiatively to the lowest vibrational level of the first excited singlet state S_1 on a picosecond time-scale. From S_1 the molecules can decay radiatively, with a radiative lifetime on the nanosecond time-scale, to a higher-lying vibrational–rotational level of S_0. From this level they rapidly thermalize into the lowest vibrational–rotational levels of S_0. Alternatively, from S_1, the molecules can experience nonradiative relaxation either to the triplet state T_1 by an intersystem crossing process or to the ground state by an internal conversion process. If the intensity of the pumping radiation is high enough a population inversion between S_1 and S_0 may be attained and stimulated emission occurs. Internal conversion and intersystem crossing compete with the fluorescence decay mode of the molecule and therefore reduce the efficiency of the laser emission. The rate for internal conversion to the electronic ground state is usually negligibly small so that the most important loss process is intersystem crossing into T_1 that populates the lower metastable triplet state. Thus, absorption on the triplet–triplet allowed transitions could cause considerable losses if these absorption bands overlap the lasing band, inhibiting or even halting the lasing process. This triplet loss can be reduced by adding small quantities of appropriate chemicals that favor nonradiative transitions that shorten the effective lifetime of the T_1 level. For pulsed excitation with nanosecond pulses, the triplet–triplet absorption can be neglected because for a typical dye the intersystem crossing rate is not fast enough to build up an appreciable triplet population in the nanosecond time domain.

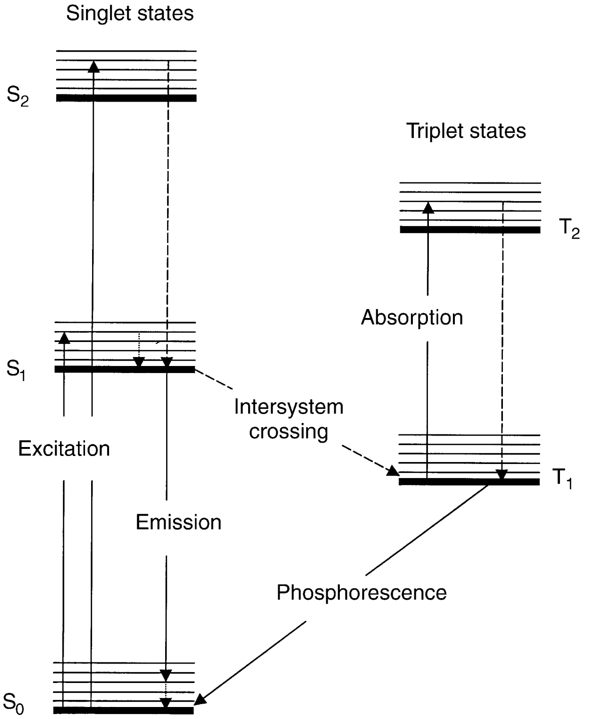

Figure 2 Schematic energy level diagram for a dye molecule. Full lines: radiative transitions; dashed lines: nonradiative transitions; dotted lines: vibrational relaxation.

Dye molecules are large (a typical molecule incorporates 50 or more atoms) and are grouped into families with similar chemical structures. A survey of the major classes of laser dyes is given later. Solid-state laser dye gain media are also considered later.

Liquid Dye Lasers

Laser-Pumped Pulsed Dye Lasers

Laser-pumped dye lasers use a shorter wavelength, or higher frequency, pulsed laser as the excitation or pump source. Typical pump lasers for dye lasers are gas lasers such as the excimer, nitrogen, or copper lasers. One of the most widely used solid-state laser pumps is the frequency doubled Nd:YAG laser which emits at 532 nm.

In a laser-pumped pulsed dye laser the active medium, or dye solution, is contained in an optical cell often made of quartz or fused silica, which provides an active region typically some 10 mm in length and a few mm in width. The active medium is then excited either longitudinally, or transversely, via a focusing lens using the pump laser. In the case of transverse excitation the pump laser is focused to a beam ~10 mm in width and ~0.1 mm in height. Longitudinal pumping requires focusing of the excitation beam to a diameter in the 0.1–0.15 mm range. For lasers operated at low prfs (a few pulses per second), the dye solution might be static. However, for high-prf operation (a few thousand pulses per second) the dye solution must be flowed at speeds of up to a few meters per second in order to dissipate the heat. A simple broadband optically pumped dye laser can be constructed using just the pump laser, the active medium, and two mirrors to form a resonator. In order to achieve tunable, narrow-linewidth, emission, a more sophisticated resonator must be employed. This is called a dispersive tunable oscillator and is depicted in Figure 3. In a dispersive tunable oscillator the exit side of the cavity is comprised of a partial reflector, or an output coupler, and the other end of the resonator is composed of a multiple-prism grating assembly. It is the dispersive characteristics of this multiple-prism grating assembly and the dimensions of the emission beam produced at the gain medium that determine the tunability and the narrowness, or spectral purity, of the laser emission.

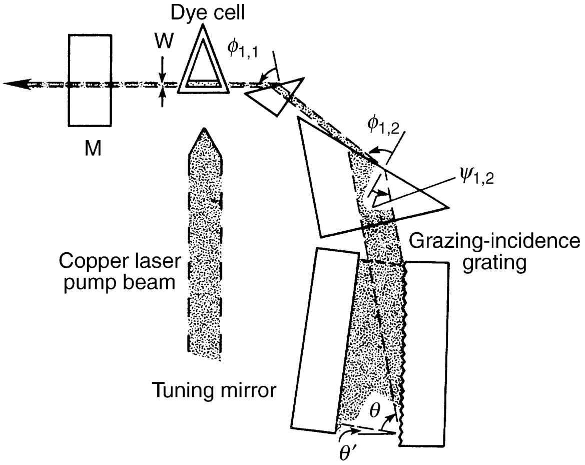

Figure 3 Copper-vapor-laser pumped hybrid multiple-prism near grazing incidence (HMPGI) grating dye laser oscillator. Adapted with permission from Duarte FJ and Piper JA (1984) Narrow linewidth high prf copper laser-pumped dye-laser oscillators. *Applied Optics* 23: 1391–1394.

In order to selectively excite a single vibrational–rotational level of a molecule such as iodine (I_2), at room temperature, one needs a laser linewidth of $\Delta\nu \approx 1.5$ GHz (or $\Delta\lambda \approx 0.0017$ nm at $\lambda = 590$ nm). The hybrid multiple-prism near-grazing-incidence (HMPGI) grating dye laser oscillator illustrated in Figure 4 yields laser linewidths in the 400 MHz $\leq \Delta\nu \leq$ 650 MHz range at 4–5% conversion efficiencies whilst excited by a copper-vapor laser operating at a prf of 10 kHz. Pulse lengths are ~10 ns at full-width half-maximum (FWHM). The narrow-linewidth emission from these oscillators is said to be single-longitudinal-mode lasing because only one electromagnetic mode is allowed to oscillate.

The emission from oscillators of this class can be amplified many times by propagating the tunable narrow-linewidth laser beam through single-pass amplifier dye cells under the excitation of pump lasers. Such amplified laser emission can reach enormous average powers. Indeed, a copper-vapor-laser pumped dye laser system at the Lawrence Livermore National Laboratory (USA), designed for the laser isotope separation program, was reported to

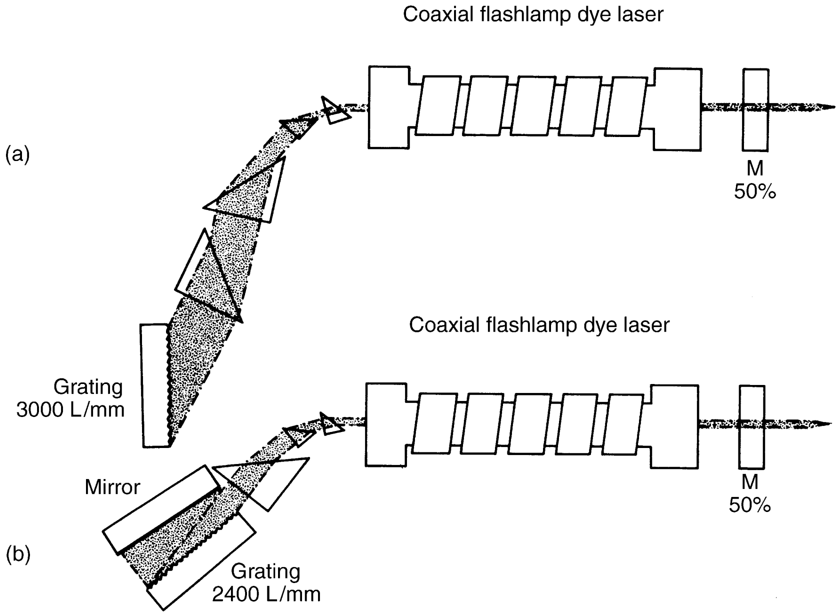

Figure 4 Flashlamp-pumped multiple-prism grating oscillators. From Duarte FJ, Davenport WE, Ehrlich JJ and Taylor TS (1991) Ruggedized narrow-linewidth dispersive dye laser oscillator. *Optics Communications* 84: 310–316. Reproduced with permission from Elsevier.

yield average powers in excess of 2.5 kW, at a prf of 13.2 kHz, at a better than 50% conversion efficiency as reported by Bass and colleagues in 1992. Besides high conversion efficiencies, excitation with copper-vapor lasers, at $\lambda = 510.554$ nm, has the advantage of inducing little photodegradation in the active medium thus allowing very long dye lifetimes.

Flashlamp-Pumped Dye Lasers

Flashlamps utilized in dye laser excitation emit at black-body temperatures in the 20 000 K range, thus yielding intense ultraviolet radiation centered around 200 nm. One further requirement for flashlamps, and their excitation circuits, is to deliver light pulses with a fast rise time. For some flashlamps this rise time can be less than a few nanoseconds.

Flashlamp-pumped dye lasers differ from laser-pumped pulsed dye lasers mainly in the pulse energies and pulse lengths attainable. This means that flashlamp-pumped dye lasers, using relatively large volumes of dye, can yield very large energy pulses. Excitation geometries use either coaxial lamps, with the dye flowing in a quartz cylinder at the center of the lamp, or two or more linear lamps arranged symmetrically around the quartz tube containing the dye solution. Using a relatively weak dye solution of rhodamine-6G (2.2×10^{-5} M), a coaxial lamp, and an active region defined by a quartz tube 6 cm in diameter and a length of 60 cm, Baltakov and colleagues, in 1974, reported energies of 400 J in pulses 25 μs long at FWHM.

Flashlamp-pumped tunable narrow-linewidth dye laser oscillators described by Duarte and colleagues, in1991, employ a cylindrical active region 6 mm in diameter and 17 cm in length. The dye solution is made of rhodamine 590 at a concentration of 1×10^{-5} M. This active region is excited by a coaxial flashlamp. Using multiple-prism grating architectures (see Figure 4) these authors achieve a diffraction limited TEM_{00} laser beam and laser linewidths of $\Delta\nu \approx 300$ MHz at pulsed energies in the 2–3 mJ range. The laser pulse duration is reported to be $\Delta t \approx 100$ ns. The laser emission from this class of multiple-prism grating oscillator is reported to be extremely stable. The tunable narrow-linewidth emission from these dispersive oscillators is either used directly in spectroscopic, or other scientific applications, or is utilized to inject large flashlamp-pumped dye laser amplifiers to obtain multi-joule pulse energies with the laser linewidth characteristics of the oscillator.

Continuous Wave Dye Lasers

CW dye lasers use dye flowing at linear speeds of up to 10 meters per second which are necessary to remove the excess heat and to quench the triplet states. In the original cavity reported by Peterson and colleagues, in 1970, a beam from an Ar^+ laser was focused on to an active region which is contained within the resonator. The resonator comprised dichroic mirrors that transmit the blue-green radiation of the pump laser and reflect the red emission from the dye molecules. Using a pump power of about 1 W, in a TEM_{00} laser beam, these authors reported a dye laser output of 30 mW. Subsequent designs replaced the dye cell with a dye jet, an introduced external mirror, and integrated dispersive elements in the cavity. Dispersive elements such as prisms and gratings are used to tune the wavelength output of the laser. Frequency-selective elements, such as etalons and other types of interferometers, are used to induce frequency narrowing of the tunable emission. Two typical cw dye laser cavity designs are described by Hollberg, in 1990, and are reproduced here in Figure 5. The first design is a linear three-mirror folded cavity. The second one is an eight-shaped ring dye laser cavity comprised of mirrors M_1, M_2, M_3, and M_4. Linear cavities exhibit the effect of *spatial hole burning* which allows the cavity to lase in more than one longitudinal mode. This problem can be overcome in ring cavities (Figure 5b) where the laser emission is in the form of a traveling wave.

Two aspects of cw dye lasers are worth emphasizing. One is the availability of relatively high powers in single longitudinal mode emission and the other is the

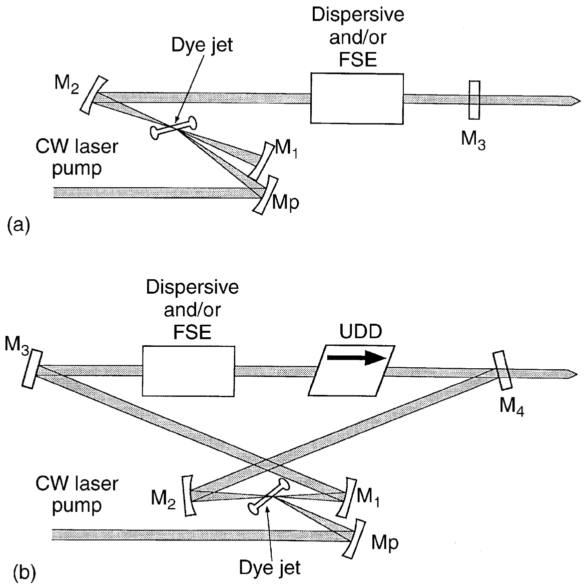

Figure 5 CW laser cavities: (a) linear cavity and (b) ring cavity. Adapted from Hollberg LW (1990) CW dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 185–238. New York: Academic Press. Reproduced with permission from Elsevier.

demonstration of very stable laser oscillation. First, Johnston and colleagues, in 1982, reported 5.6 W of stabilized laser output in a single longitudinal mode at 593 nm at a conversion efficiency of 23%. In this work eleven dyes were used to span the spectrum continuously from ~400 nm to ~900 nm. In the area of laser stabilization and ultra narrow-linewidth oscillation it is worth mentioning the work of Hough and colleagues, in 1984, that achieved laser linewidths of less than 750 Hz employing an external reference cavity.

Ultrashort-Pulse Dye Lasers

Ultrashort-pulse, or femtosecond, dye lasers use the same type of technology as cw dye lasers configured to incorporate a saturable absorber region. One such configuration is the ring cavity depicted in Figure 6. In this cavity the gain region is established between mirrors M_1 and M_2 whilst the saturable absorber is deployed in a counter-propagating arrangement. This arrangement is necessary to establish a collision between two counter-propagating pulses at the saturable absorber thus yielding what is known as colliding-pulse mode locking (CPM) as reported by Ruddock and Bradley, in 1976. This has the effect of creating a transient grating, due to interference, at the absorber thus shortening the pulse. Intracavity prisms were incorporated in order to introduce negative dispersion, by Dietel and colleagues in 1983, and thus subtract dispersion from the cavity and ultimately provide the compensation needed to produce femtosecond pulses.

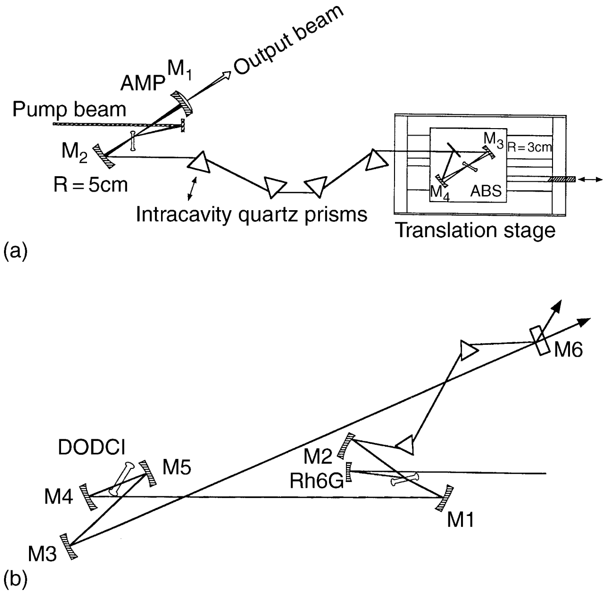

Figure 6 Femtosecond dye laser cavities: (a) linear femtosecond cavity and (b) ring femtosecond cavity. Adapted from Diels J-C (1990) Femtosecond dye lasers. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 41–132. New York: Academic Press. Reproduced with permission from Elsevier.

The shortest pulse obtained from a dye laser, 6 fs, has been reported by Fork and colleagues, in 1987, using extra-cavity compression. In that experiment, a dye laser incorporating CPM and prismatic compensation was used to generate pulses that were amplified by a copper-vapor laser at a prf of 8 kHz. The amplified pulses, of a duration of 50 fs, were then propagated through two grating pairs and a four-prism sequence for further compression.

Solid-State Dye Laser Oscillators

In this section the principles of linewidth narrowing in dispersive resonators are outlined. Albeit the discussion focuses on multiple-prism grating solid-state dye laser oscillators, in particular, the physics is applicable to pulsed high-power dispersive tunable lasers in general.

Multiple-Prism Dispersion Grating Theory

The spectral linewidth in a dispersive optical system is given by

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda\theta)^{-1} \qquad [1]$$

where $\Delta\theta$ is the light beam divergence, $\nabla_\lambda = \partial/\partial\lambda$, and $\nabla_\lambda\theta$ is the overall dispersion of the optics. This identity can de derived either from the principles of geometrical optics or from the principles of generalized interferometry as described by Duarte, in 1992.

The cumulative single-pass generalized multiple-prism dispersion at the mth prism, of a multiple-prism array as illustrated in Figure 7, as given by Duarte and Piper, in 1982,

$$\begin{aligned}\nabla_\lambda\phi_{2,m} &= H_{2,m}\nabla_\lambda n_m + (k_{1,m}k_{2,m})^{-1}\\ &\quad \times\left(H_{1,m}\nabla_\lambda n_m \pm \nabla_\lambda\phi_{2,(m-1)}\right)\end{aligned} \qquad [2]$$

In this equation

$$k_{1,m} = \cos\psi_{1,m}/\cos\phi_{1,m} \qquad [3a]$$

$$k_{2,m} = \cos\phi_{2,m}/\cos\psi_{2,m} \qquad [3b]$$

$$H_{1,m} = \tan\phi_{1,m}/n_m \qquad [4a]$$

$$H_{2,m} = \tan\phi_{2,m}/n_m \qquad [4b]$$

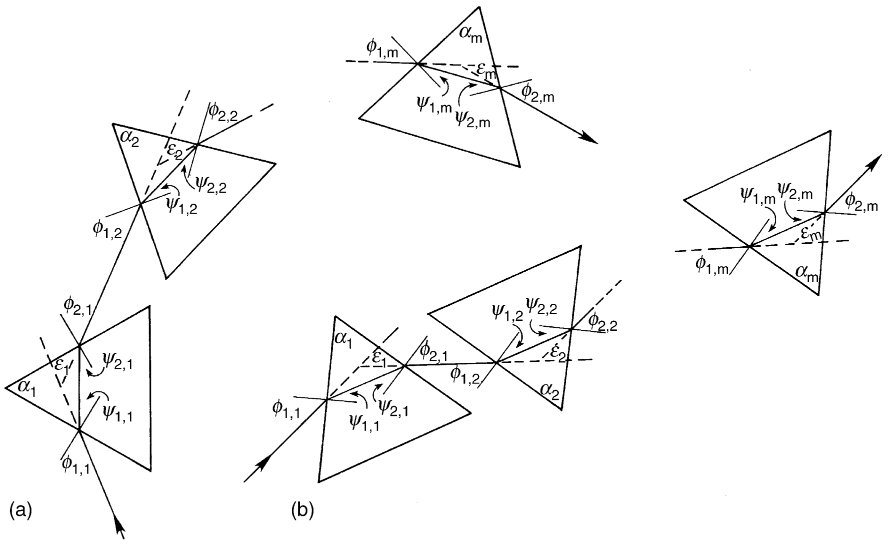

Figure 7 Generalized multiple-prism arrays in (a) additive and (b) compensating configurations. From Duarte FJ (1990) Narrow-linewidth pulsed dye laser oscillators. In: Duarte FJ and Hillman LW (eds) *Dye Laser Principles*, pp. 133–183. New York: Academic Press. Reproduced with permission from Elsevier.

Here, $k_{1,m}$ and $k_{2,m}$ represent the physical beam expansion experienced by the incident and the exit beams, respectively.

Equation [2] indicates that $\nabla_\lambda \phi_{2,m}$, the cumulative dispersion at the mth prism, is a function of the geometry of the mth prism, the position of the light beam relative to this prism, the refractive index of the prism, and the cumulative dispersion up to the previous prism $\nabla_\lambda \phi_{2,(m-1)}$.

For an array of r identical isosceles, or equilateral, prisms arranged symmetrically, in an additive configuration, so that the angles of incidence and emergence are the same, the cumulative dispersion reduces to

$$\nabla_\lambda \phi_{2,r} = r\nabla_\lambda \phi_{2,1} \qquad [5]$$

Under these circumstances the dispersions add in a simple and straightforward manner. For configurations incorporating right-angle prisms, the dispersions need to be handled mathematically in a more subtle form.

The generalized double-pass, or return-pass, dispersion for multiple-prism beam expanders was introduced by Duarte, in 1985:

$$\nabla_\lambda \Phi_P = 2M_1M_2\sum_{m=1}^{r}(\pm 1)H_{1,m}\left(\prod_{j=m}^{r}k_{1,j}\prod_{j=m}^{r}k_{2,j}\right)^{-1}\nabla_\lambda n_m + 2\sum_{m=1}^{r}(\pm 1)H_{2,m}\left(\prod_{j=1}^{m}k_{1,j}\prod_{j=1}^{m}k_{2,j}\right)\nabla_\lambda n_m \qquad [6]$$

Here, M_1 and M_2 are the beam magnification factors given by

$$M_1 = \prod_{m=1}^{r} k_{1,m} \qquad [7a]$$

$$M_2 = \prod_{m=1}^{r} k_{2,m} \qquad [7b]$$

For a multiple prism expander designed for an orthogonal beam exit, and Brewster's angle of incidence, eqn [6] reduces to the succinct expression given by Duarte, in 1990:

$$\nabla_\lambda \Phi_P = 2\sum_{m=1}^{r}(\pm 1)(n_m)^{m-1}\nabla_\lambda n_m \qquad [8]$$

Equation [6] can be used to either quantify the overall dispersion of a given multiple-prism beam expander or to design a prismatic expander yielding zero dispersion, that is, $\nabla_\lambda \Phi_P = 0$, at a given wavelength.

Physics and Architecture of Solid-State Dye-Laser Oscillators

The first high-performance narrow-linewidth tunable laser was introduced by Hänsch in 1972. This laser yielded a linewidth of $\Delta\nu \approx 2.5$ GHz (or $\Delta\lambda \approx 0.003$ nm at $\lambda \approx 600$ nm) in the absence of an intracavity etalon. Hänsch demonstrated that the laser linewidth from a tunable laser was narrowed significantly when the beam incident on the tuning grating was expanded using an astronomical telescope. The linewidth equation including the intracavity beam magnification factor can be written as

$$\Delta\lambda \approx \Delta\theta(M\nabla_\lambda\Theta_G)^{-1} \qquad [9]$$

From this equation, it can be deduced that a narrow $\Delta\lambda$ is achieved by reducing $\Delta\theta$ and increasing the overall intracavity dispersion $(M\nabla_\lambda\Theta_G)$. The intracavity dispersion is optimized by expanding the size of the intracavity beam incident on the diffractive surface of the tuning grating until it is totally illuminated.

Hänsch used a two-dimensional astronomical telescope to expand the intracavity beam incident on the diffraction grating. A simpler beam expansion method consists in the use of a single-prism beam expander as disclosed by several authors (Myers, 1971; Stokes and colleagues, 1972; Hanna and colleagues, 1975). An extension and improvement of this approach was the introduction of multiple-prism beam expanders as reported by Kasuya and colleagues, in 1978, Klauminzer, in 1978, and Duarte and Piper, in 1980. The main advantages of multiple-prism beam expanders, over two-dimensional telescopes, are simplicity, compactness, and the fact that the beam expansion is reduced from two dimensions to one dimension. Physically, as explained previously, prismatic beam expanders also introduce a dispersion component that is absent in the case of the astronomical telescope. Advantages of multiple-prism beam expanders over single-prism beam expansion are higher transmission efficiency, lower amplified spontaneous emission levels, and the flexibility to either augment or reduce the prismatic dispersion.

In general, for a pulsed multiple-prism grating oscillator, Duarte and Piper, in 1984, showed that the return-pass dispersive linewidth is given by

$$\Delta\lambda = \Delta\theta_R(MR\nabla_\lambda\Theta_G + R\nabla_\lambda\Phi_P)^{-1} \qquad [10]$$

where R is the number of return-cavity passes. The grating dispersion in this equation, $\nabla_\lambda\Theta_G$, can be either from a grating in Littrow or near grazing-incidence configuration. The multiple-return-pass equation for the beam divergence was given by Duarte, in 2001:

$$\Delta\theta_R = (\lambda/\pi w)\left(1 + (L_R/B_R)^2 + (A_R L_R/B_R)^2\right)^{1/2} \qquad [11]$$

Here, $L_R = (\pi w^2/\lambda)$ is the Rayleigh length of the cavity, w is the beam waist at the gain region, while A_R and B_R are the corresponding multiple-return-pass elements derived from propagation matrices.

At present, very compact and optimized multiple-prism grating tunable laser oscillators are found in two basic cavity architectures. These are the multiple-prism Littrow (MPL) grating laser oscillator, reported by Duarte in 1999 and illustrated in Figure 8, and the hybrid multiple-prism near grazing-incidence (HMPGI) grating laser oscillator, introduced by Duarte in 1997 and depicted in Figure 9. In early MPL grating oscillators the individual prisms integrating the multiple-prism expander were deployed in an additive configuration thus adding the cumulative dispersion to that of the grating and thus contributing to the overall dispersion of the cavity. In subsequent architectures the prisms were deployed in compensating configurations so as to yield zero dispersion and thus allow the tuning characteristics of the cavity to be

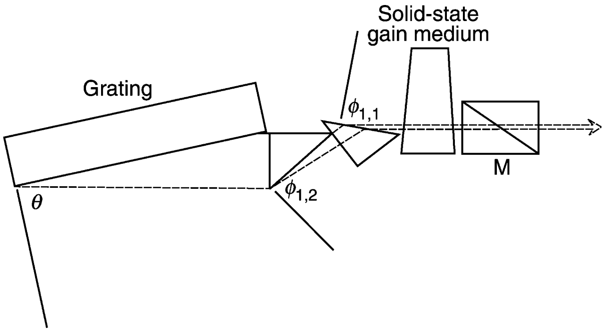

Figure 8 MPL grating solid-state dye laser oscillator: optimized architecture. The physical dimensions of the optical components of this cavity are shown to scale. The length, of the solid-state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1999) Multiple-prism grating solid-state dye laser oscillator: optimized architecture. *Applied Optics* 38: 6347–6349.

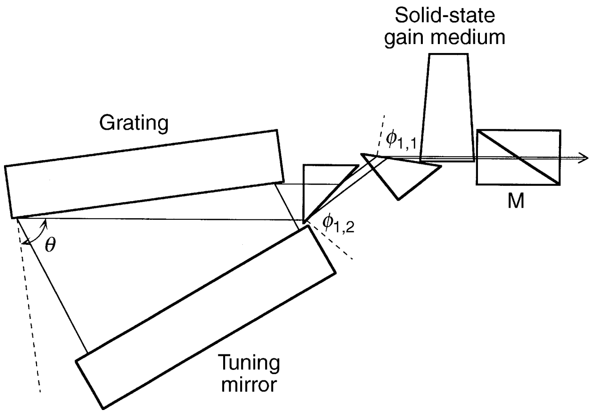

Figure 9 HMPGI grating solid-state dye laser. Schematics to scale. The length of the solid state dye gain medium, along the optical axis, is 10 mm. Reproduced with permission from Duarte FJ (1997) Multiple-prism near-grazing-incidence grating solid-state dye laser oscillator. *Optics and Laser Technology* 29: 513–516.

determined by the grating exclusively. In this approach the principal role of the multiple-prism array is to expand the beam incident on the grating thus augmenting significantly the overall dispersion of the cavity as described in eqns [9] or [10]. In this regard, it should be mentioned that beam magnification factors of up to 100, and beyond, have been reported in the literature. Using solid-state laser dye gain media, these MPL and HMPGI grating laser oscillators deliver tunable single-longitudinal-mode emission at laser linewidths 350 MHz $\leq \Delta\nu \leq$ 375 MHz and pulse lengths in the 3–7 ns (FWHM) range. Long-pulse operation of this class of multiple-prism grating oscillators has been reported by Duarte and colleagues, in 1998. In these experiments laser linewidths of $\Delta\nu \approx 650$ MHz were achieved at pulse lengths in excess of 100 ns (FWHM) under flashlamp-pumped dye laser excitation.

The dispersive cavity architectures described here have been used with a variety of laser gain media in the gas, the liquid, and the solid state. Applications to tunable semiconductor lasers have also been reported by Zorabedian, in 1992, Duarte, in 1993, and Fox and colleagues, in 1997.

Concepts important to MPL and HMPGI grating tunable laser oscillators include the emission of a single-transverse-mode (TEM_{00}) laser beam in a compact cavity, the use of multiple-prism arrays, the expansion of the intracavity beam incident on the grating, the control of the intracavity dispersion, and the quantification of the overall dispersion of the multiple-prism grating assembly via generalized dispersion equations. Sufficiently high intracavity dispersion leads to the achievement of return-pass dispersive linewidths close to the free-spectral range of the cavity. Under these circumstances single-longitudinal-mode lasing is readily achieved as a result of multipass effects.

A Note on the Cavity Linewidth Equation

So far we have described how the cavity linewidth equation

$$\Delta\lambda \approx \Delta\theta(\nabla_\lambda\theta)^{-1}$$

and the dispersion equations can be applied to achieve highly coherent, or very narrow-linewidth, emission. It should be noted that the same physics can be applied to achieve ultrashort, or femtosecond, laser pulses. It turns out that intracavity prisms can be configured to yield negative dispersion as described by Duarte and Piper, in 1982, Dietel and colleagues, in 1983, and Fork and colleagues in 1984. This negative dispersion can reduce significantly the overall dispersion of the cavity. Under those circumstances, eqn [1] predicts broadband emission which according to the uncertainty principle, in the form of,

$$\Delta\nu\Delta t \approx 1 \qquad [12]$$

can lead to very short pulse emission since $\Delta\nu$ and $\Delta\lambda$ are related by the identities

$$\Delta\lambda \approx \lambda^2/\Delta x \qquad [13]$$

and

$$\Delta\nu \approx c/\Delta x \qquad [14]$$

Distributed-Feedback Solid-State Dye Lasers

Recently, narrow-linewidth laser emission from solid-state dye lasers has also been obtained using a distributed-feedback (DFB) configuration by Wadsworth and colleagues, in 1999, and Zhu and colleagues, in 2000. As is well known, in a DFB laser no external cavity is required, the feedback being provided by Bragg reflection from a permanently or dynamically written grating structure within the gain medium as reported by Kogelnik and Shank, in 1971. By using a DFB laser configuration where the interference of two pump beams induces the required periodic modulation in the gain medium, laser emission with linewidth in the 0.01–0.06 nm range have been reported. Specifically, Wadsworth and colleagues reported a laser linewidth of 12 GHz (0.016 nm at 616 nm) using Perylene Red doped PMMA at a conversion efficiency of 20%.

It should also be mentioned that the DFB laser is also a dye laser development that has found wide and extensive applicability in semiconductor lasers employed in the telecommunications industry.

Laser Dye Survey

Cyanines

These are red and near-infrared dyes which present long conjugated methine chains (!CH = CH!) and are useful in the spectral range longer than 800 nm, where no other dyes compete with them. An important laser dye belonging to this class is 4-dicyanomethylene-2-methyl-6-(*p*-dimethylamino-styryl)-4H-pyran (DCM, Figure 10a), with laser emission in the 600–700 nm range depending on the pump and solvent, liquid or solid, used.

Xanthenes

These dyes have a xanthene ring as the chromophore and are classified into rhodamines, incorporating

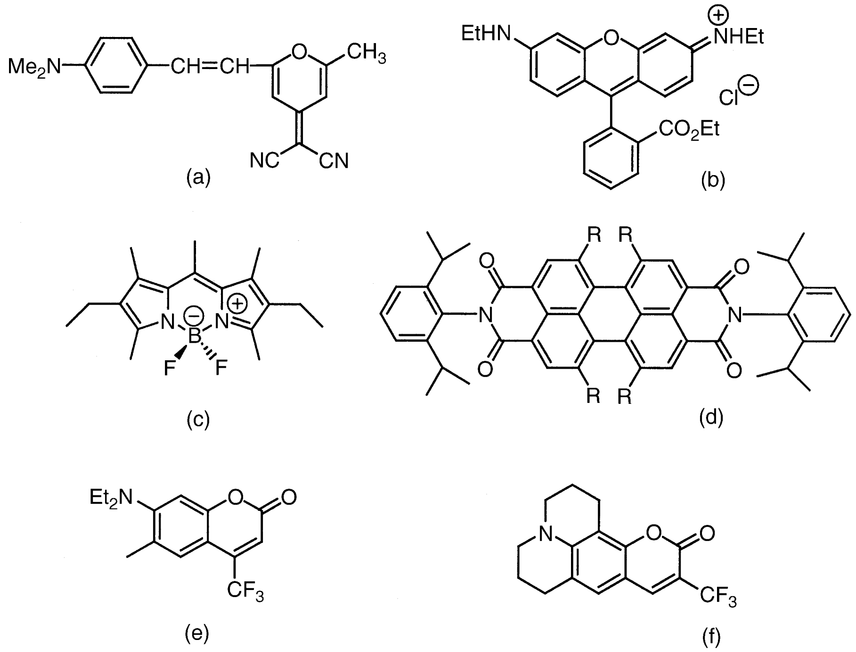

Figure 10 Molecular structures of some common laser dyes: (a) DCM; (b) Rh6G; (c) PM567; (d) Perylene Orange (R = H) and Perylene Red (R = C_4H_6O); (e) Coumarin 307 (also known as Coumarin 503); (f) Coumarin 153 (also known as Coumarin 540A).

amino radicals substituents, and fluoresceins, with hydroxyl (OH) radical substituents. They are generally very efficient and chemically stable and their emission covers the wavelength region from 500 to 700 nm.

Rhodamine dyes are the most important group of all laser materials, with rhodamine 6G (Rh6G, Figure 10b), also called rhodamine 590 chloride, being probably the best known of all laser dyes. Rh6G exhibits an absorption peak, in ethanol, at 530 nm and a fluorescence peak at 556 nm, with laser emission typically in the 550–620 nm region. It has been demonstrated to lase efficiently both in liquid and solid solutions.

Pyrromethenes

Pyrromethene.BF_2 complexes are a new class of laser dyes synthesized and characterized more recently as reported by Shah and colleagues, in 1990, Pavlopoulos and colleagues, in 1990, and Boyer and colleagues, in 1993. These laser dyes exhibit reduced triplet–triplet absorption over their fluorescence and lasing spectral region while retaining a high quantum fluorescence yield. Depending on the substituents on the chromophore, these dyes present laser emission over the spectral region from the green/yellow to the red, competing with the rhodamine dyes and have been demonstrated to lase with good performance when incorporated into solid hosts. Unfortunately, they are relatively unstable because of the aromatic amine groups in their structure, which render them vulnerable to photochemical reactions with oxygen as indicated by Rahn and colleagues, in 1997. The molecular structure of a representative and well-known dye of this class, 1,3,5,7,8-pentamethyl-2,6-diethylpyrromethene-difluoroborate complex (pyrromethene 567, PM567) is shown in Figure 10c. Recently, analogs of dye PM567 substituted at position 8 with an acetoxypolymethylene linear chain or a polymerizable methacryloyloxy polymethylene chain have been developed. These new dipyrromethene.BF_2 complexes have demonstrated improved efficiency and better photostability than the parent compound both in liquid and solid state.

Conjugated Hydrocarbons

This class of organic compounds includes perylenes, stilbenes, and *p*-terphenyl. Perylene and perylimide dyes are large, nonionic, nonpolar molecules (Figure 10d) characterized by their extreme photostability and negligible singlet–triplet transfer, as discussed by Seybold and Wagenblast and colleagues, in 1989, which have high quantum efficiency because of the absence of nonradiative relaxation. The perylene dyes exhibit limited solubility in conventional solvents but dissolve well in acetone, ethyl acetate, and methyl methacrylate, with emission wavelengths in the orange and red spectral regions.

Stilbene dyes are derivatives of uncyclic unsaturated hydrocarbons such as ethylene and butadiene, with emission wavelengths in the 400–500 nm range. Most of them end in phenyl radicals. Although they are chemically stable, their laser performance is inferior to that of coumarins. *p*-Terphenyl is a valuable and efficient *p*-oligophenylene dye with emission in the ultraviolet.

Coumarins

Coumarin derivatives are a popular family of laser dyes with emission in the blue-green region of the spectrum. Their structure is based on the coumarin ring (Figure 10e,f) with different substituents that strongly affect its chemical characteristics and allow covering an emission spectral range between 420 and 580 nm. Some members of this class rank among the most efficient laser dyes known, but their chemical photobleaching is rapid compared with xanthene and pyrromethene dyes. An additional class of blue-green dyes is the tetramethyl derivatives of coumarin dyes introduced by Chen and colleagues, in 1988. The emission from these dyes spans the spectrum in the 453–588 nm region. These coumarin analogs exhibit improved efficiency, higher solubility, and better lifetime characteristics than the parent compounds.

Azaquinolone Derivatives

The quinolone and azaquinolone derivatives have a structure similar to that of coumarins but with their laser emission range extended toward the blue by 20–30 nm. The quinolone dyes are used when laser emission in the 400–430 nm region is required. Azaquinolone derivatives, such as 7-dimethylamino-1-methyl-4-metoxy-8-azaquinolone-2 (LD 390), exhibit laser action below 400 nm.

Oxadiazole Derivatives

These are compounds which incorporate an axodiazole ring with aryl radical substituents and which lase in the 330–450 nm spectral region. Dye 2-(4-biphenyl)-5-(4-*t*-butylphenyl)-1,3,4-oxadiazole (PBD), with laser emission in the 355–390 nm region, belongs to this family.

Solid-State Laser Dye Matrices

Although some first attempts to incorporate dye molecules into solid matrices were already made in the early days of development of dye lasers, it was not until the late 1980s that host materials with the required properties of optical quality and high damage threshold to laser radiation began to be developed. In subsequent years, the synthesis of new high-performance dyes and the implementation of new ways of incorporating the organic molecules into the solid matrix resulted in significant advances towards the development of practical tunable solid-state dye lasers. A recent detailed review of the work done in this field has been given by Costela *et al.*

Inorganic glasses, transparent polymers, and inorganic–organic hybrid matrices have been successfully used as host matrices for laser dyes. Inorganic glasses offer good thermal and optical properties but present the difficulty that the high melting temperature employed in the traditional process of glass making would destroy the organic dye molecules. It was not until the development of the low-temperature sol-gel technique for the synthesis of glasses that this limitation was overcome. The sol-gel process, based on inorganic polymerization reactions performed at about room temperature and starting with metallo-organic compounds such as alkoxides or salts, as reported by Brinker and Scherrer, in 1989, provides a safe route for the preparation of rigid, transparent, inorganic matrix materials incorporating laser dyes. Disadvantages of these materials are the possible presence of impurities embedded in the matrix or the occurrence of interactions between the dispersed dye molecules and the inorganic structure (hydrogen bonds, van der Waals forces) which, in certain cases, can have a deleterious effect on the lasing characteristics of the material.

The use of matrices based on polymeric materials to incorporate organic dyes offers some technical and economical advantages. Transparent polymers exhibit high optical homogeneity, which is extremely important for narrow-linewidth oscillators, as explained by Duarte, in 1994, good chemical compatibility with organic dyes, and allow control over structure and chemical composition making possible the modification, in a controlled way, of relevant properties of these materials such as polarity, free volume, molecular weight, or viscoelasticity. Furthermore, the polymeric materials are amenable to inexpensive fabrication techniques, which facilitate miniaturization and design of integrated optical systems. The dye molecules can be either dissolved in the polymer or linked covalently to the polymeric chains.

In the early investigations on the use of solid polymer matrices for laser dyes, the main problem to be solved was that of the low resistance to laser radiation exhibited by the then existing materials. In the late 1980s and early 1990s new modified polymeric organic materials began to be developed with a laser-radiation-damage threshold comparable to that of inorganic glasses and crystals as reported by Gromov and colleagues, in 1985, and Dyumaev and colleagues, in 1992. This, together with the above-mentioned advantages, made the use of polymers in solid-state dye lasers both attractive and competitive.

An approach that intends to bring about materials in which the advantages of both inorganic glasses and polymers are preserved but the difficulties are avoided is that of using inorganic–organic hybrid matrices. These are silicate-based materials, with an inorganic Si–O–Si backbone, prepared from organosilane precursors by sol-gel processing in combination with organic cross-linking of polymerizable monomers as reported by Novak, in 1993, Sanchez and Ribot, in 1994, and Schubert, in 1995. In one procedure, laser dyes are mixed with organic monomers, which are then incorporated into the porous structure of a sol-gel inorganic matrix by immersing the bulk in the solution containing monomer and catalyst or photoinitiator as indicated by Reisfeld and Jorgensen, in 1991, and Bosch and colleagues, in 1996. Alternatively, hybrids can be obtained from organically modified silicon alkoxides, producing the so-called ORMOCERS (organically modified ceramics) or ORMOSILS (organically modified silanes) as reported by Reisfeld and Jorgensen, in 1991. An inconvenient aspect of these materials is the appearance of optical inhomogeneities in the medium due to the difference in the refractive index between the organic and inorganic phases as well as to the difference in density between monomer and polymer which causes stresses and the formation of optical defects. As a result, spatial inhomogeneities appear in the laser beam, thereby decreasing its quality as discussed by Duarte, in 1994.

Organic Hosts

The first polymeric materials with improved resistance to damage by laser radiation were obtained by Russian workers, who incorporated rhodamine dyes into modified poly(methyl methacrylate) (MPMMA), obtained by doping PMMA with low-molecular weight additives. Some years later, in 1995, Maslyukov and colleagues demonstrated lasing efficiencies in the range 40–60% with matrices of MPMMA doped with rhodamine dyes pumped longitudinally at 532 nm, with a useful lifetime (measured as a 50% efficiency drop) of 15 000 pulses at a prf of 3.33 Hz.

In 1995, Costela and colleagues used an approach based on adjusting the viscoelastic properties of the material by modifying the internal plasticization of the polymeric medium by copolymerization with appropriate monomers. Using dye Rh6G dissolved in a copolymer of 2-hydroxyethyl methacrylate (HEMA) and methyl methacrylate (MMA) and transversal pumping at 337 nm, they demonstrated laser action with an efficiency of 21% and useful lifetime of 4500 pulses (20 GJ/mol in terms of total input energy per mole of dye molecule when the output energy is down to 50% of its initial value). The useful lifetime increased to 12 000 pulses when the Rh6G chromophore was linked covalently to the polymeric chains as reported by Costela and colleagues, in 1996.

Comparative studies on the laser performance of Rh6G incorporated either in copolymers of HEMA and MMA or in MPMMA were carried out by Giffin and colleagues, in 1999, demonstrating higher efficiency of the MPMMA materials but superior normalized photostability (up to 240 GJ/mol) of the copolymer formulation.

Laser conversion efficiencies in the range 60–70% for longitudinal pumping at 532 nm were reported by Ahmad and colleagues, in 1999, with PM567 dissolved in PMMA modified with 1,4-diazobicyclo[2,2,2] octane (DABCO) or perylene additives. When using DABCO as an additive, a useful lifetime of up to 550 000 pulses, corresponding to a normalized photostability of 270 GJ/mol, was demonstrated at a 2 Hz prf.

Much lower efficiencies and photostabilities have been obtained with dyes emitting in the blue-green spectral region. Costela and colleagues, in 1996, performed some detailed studies on dyes coumarin 540A and coumarin 503 incorporated into methacrylate homopolymers and copolymers, and demonstrated efficiencies of at most 19% with useful lifetimes of up to 1200 pulses, at a 2 Hz prf, for transversal pumping at 337 nm. In 2000, Somasundaram and Ramalingam obtained useful lifetimes of 5240 and 1120 pulses, respectively, for coumarin 1 and coumarin 490 dyes incorporated into PMMA rods modified with ethyl alcohol, under transversal pumping at 337 nm and a prf of 1 Hz.

A detailed study of photo-physical parameters of R6G-doped HEMA-MMA gain media was performed by Holzer and colleagues, in 2000. Among the parameters determined by these authors are quantum yields, lifetimes, absorption cross-sections, and emission cross-sections. A similar study on the photophysical parameters of PM567 and two PM567 analogs incorporated into solid matrices of different acrylic copolymers and in corresponding mimetic liquid solutions was performed by Bergmann and colleagues, in 2001.

Preliminary experiments with rhodamine 6G-doped MPMMA at high prfs were conducted using a frequency-doubled Nd:YAG laser at 5 kHz by Duarte and colleagues, in 1996. In those experiments, involving longitudinal excitation, the pump laser was allowed to fuse a region at the incidence window of the gain medium so that a lens was formed. This lens and the exit window of the gain medium comprised a short unstable resonator that yielded broadband emission. In 2001, Costela and colleagues demonstrated lasing in rhodamine 6G- and pyrromethene-doped polymer matrices under copper-vapor-laser excitation at prfs in the 1.0–6.2 kHz range. In these experiments, the dye-doped solid-state matrix was rotated at 1200 rpm. Average powers of 290 mW were obtained at a prf of 1 kHz and a conversion efficiency of 37%. For a short period of time average powers of up to 1 W were recorded at a prf of 6.2 kHz. In subsequent experiments by Abedin and colleagues the rpf was increased to 10 kHz by using as pump laser a diode-pumped, *Q*-switched, frequency doubled, solid state Nd:YLF laser. Initial average powers of 560 mW and 430 mW were obtained for R6G-doped and PM567-doped polymer matrices, respectively. Lasing efficiency was 16% for R6G and the useful lifetime was 6.6 min (or about 4.0 million shots).

Inorganic and Hybrid Hosts

In 1990, Knobbe and colleagues and McKiernan and colleagues, from the University of California at Los Angeles, incorporated different organic dyes, via the sol-gel techniques, into silicate (SiO_2), aluminosilicate (Al_2O_3–SiO_2), and ORMOSIL host matrices, and investigated the lasing performance of the resulting materials under transversal pumping. Lasing efficiencies of 25% and useful lifetimes of 2700 pulses at 1 Hz repetition rate were obtained from Rh6G incorporated into aluminosilicate gel and ORMOSIL matrices, respectively. When the dye in the ORMOSIL matrices was coumarin 153, the useful lifetime was still 2250 pulses. Further studies by Altman and colleagues, in 1991, demonstrated useful lifetimes of 11 000 pulses at a 30 Hz prf when rhodamine 6G perchlorate was incorporated into ORMOSIL matrices. The useful lifetime increased to 14 500 pulses when the dye used was rhodamine B.

When incorporated into xerogel matrices, pyrromethene dyes lase with efficiencies as high as the highest obtained in organic hosts but with much higher photostability: a useful lifetime of 500 000 pulses at 20 Hz repetition rate was obtained by Faloss and colleagues, in 1997, with PM597 using a pump energy of 1 mJ. The efficiency of PM597 dropped to 40% but its useful lifetime increased to over 1 000 000 pulses when oxygen-free samples were prepared, in agreement with previous results that documented the strong dependence of laser parameters of pyrromethene dyes on the presence of oxygen. A similar effect was observed with perylene dyes: deoxygenated xerogel samples of Perylene Red exhibited a useful lifetime of 250 000 pulses at a prf of 5 Hz and 1 mJ pump energy, to be compared with a useful lifetime of only 10 000 pulses obtained with samples prepared in normal conditions. Perylene Orange incorporated into polycom glass and pumped longitudinally at 532 nm gave an efficiency of 72% and a useful lifetime of 40 000 pulses at a prf of 1 Hz as reported by Rhan and King, in 1998.

A direct comparison study of laser performance of Rh6G, PM567, Perylene Red, and Perylene Orange in organic, inorganic, and hybrid hosts was carried out by Rahn and King in 1995. They found that the nonpolar perylene dyes had better performance in partially organic hosts, whereas the ionic rhodamine and pyrromethene dyes performed best in the inorganic sol-gel glass host. The most promising combinations of dye and host for efficiency and photostability were found to be Perylene Orange in polycom glass and Rh6G in sol-gel glass.

An all-solid-state optical configuration, where the pump was a laser diode array side-pumped, Q-switched, frequency-doubled, Nd:YAG slab laser, was demonstrated by Hu and colleagues, in 1998, with dye DCM incorporated into ORMOSIL matrices. A lasing efficiency of 18% at the wavelength of 621 nm was obtained, and 27 000 pulses were needed for the output energy to decrease by 10% of its initial value at 30 Hz repetition rate. After 50 000 pulses the output energy decreased by 90% but it could be recovered after waiting for a few minutes without pumping.

Violet and ultraviolet laser dyes have also been incorporated into sol-gel silica matrices and their lasing properties evaluated under transversal pumping by a number of authors. Although reasonable efficiencies have been obtained with some of these laser dyes, the useful lifetimes are well below 1000 pulses.

More recently, Costela and colleagues have demonstrated improved conversion efficiencies in dye-doped inorganic–organic matrices using pyrromethene 567 dye. The solid matrix in this case is poly-trimethyl-silyl-methacrylate cross-linked with ethylene glycol and copolymerized with methyl methacrylate. Also, Duarte and James have reported on dye-doped polymer-nanoparticle laser media where the polymer is PMMA and the nanoparticles are made of silica. These authors report TEM_{00} laser beam emission and improved dn/dT coefficients of the gain media that results in reduced $\Delta\theta$.

Dye Laser Applications

Dye lasers generated a renaissance in diverse applied fields such as isotope separation, medicine, photochemistry, remote sensing, and spectroscopy. Dye lasers have also been used in many experiments that have advanced the frontiers of fundamental physics.

Central to most applications of dye lasers is their capability of providing coherent narrow-linewidth radiation that can be tuned continuously from the near ultraviolet, throughout the visible, to the near infrared. These unique coherent and spectral properties have made dye lasers particularly useful in the field of high-resolution atomic and molecular spectroscopy. Dye lasers have been successfully and extensively used in absorption and fluorescence spectroscopy, Raman spectroscopy, selective excitations, and nonlinear spectroscopic techniques. Well documented is their use in photochemistry, that is, the chemistry of optically excited atoms and molecules, and in biology, studying biochemical reaction kinetics of biological molecules. By using nonlinear optical techniques, such as harmonic and sum frequency and difference frequency generation, the properties of dye lasers can be extended to the ultraviolet.

Medical applications of dye lasers include cancer photodynamic therapy, treatment of vascular lesions, and lithotripsy as described by Goldman in 1990.

The ability of dye lasers to provide tunable sub-GHz linewidths in the orange-red portion of the spectrum, at kW average powers, made them particularly suited to atomic vapor laser isotope separation of uranium as reported by Bass and colleagues, in 1992, Singh and colleagues, in 1994, and Nemoto and colleagues, in 1995. In this particular case the isotopic shift between the two uranium isotopes is a few GHz. Using dye lasers yielding $\Delta\nu \leq 1$ GHz the ^{235}U can be selectively excited in a multistep process by tuning the dye laser(s) to specific wavelengths, in the orange-red spectral region, compatible with a transition sequence leading to photoionization. Also, high-prf narrow-linewidth dye lasers, as described by Duarte and Piper, in 1984, are ideally suited for guide star applications in astronomy. This particular application requires a high-average power diffraction-limited laser beam at $\lambda \approx 589$ nm to propagate for some 95 km, above the surface of the Earth, and illuminate a layer of sodium atoms. Ground detection of the fluorescence from the sodium atoms allows measurements on the turbulence of the atmosphere. These measurements are then used to control the adaptive optics of the telescope thus compensating for the atmospheric distortions.

A range of industrial applications is also amenable to the wavelength agility and high-average power output characteristics of dye lasers. Furthermore, tunability coupled with narrow-linewidth emission, at large pulsed energies, make dye lasers useful in military applications such as directed energy and damage of optical sensors.

The Future of Dye Lasers

Predicting the future is rather risky. In the early 1990s, articles and numerous advertisements in commercial laser magazines predicted the demise and even the oblivion of the dye laser in a few years. It did not turn out this way. The high power and wavelength agility available from dye lasers have ensured their continued use as scientific and research tools in physics, chemistry, and medicine. In addition, the advent of narrow-linewidth solid-state dye lasers has sustained a healthy level of research and development in the field. Today, some 20 laboratories around the world are engaged in this activity.

The future of solid-state dye lasers depends largely on synthetic and manufacturing improvements. There is a need for strict control of the conditions of fabrication and a careful purification of all the compounds involved. In the case of polymers, in particular, stringent control of thermal conditions during the polymerization step is obligatory to ensure that adequate optical uniformity of the polymer matrix is achieved, and that intrinsic anisotropy developed during polymerization is minimized. From an engineering perspective what are needed are dye-doped solid-state media with improved photostability characteristics and better thermal properties. By better thermal properties is meant better dn/dT factors. To a certain degree progress in this area has already been reported. As in the case of liquid dye lasers, the lifetime of the material can be improved by using pump lasers at compatible wavelengths. In this regard, direct diode laser pumping of dye-doped solid-state matrices should lead to very compact, long-lifetime, tunable lasers emitting throughout the visible spectrum. An example of such a device would be a green laser-diode-pumped rhodamine-6G doped MPMMA tunable laser configured in a multiple-prism grating architecture.

As far as liquid lasers are concerned, there is a need for efficient water-soluble dyes. Successful developments in this area were published in the literature with coumarin-analog dyes by Chen and colleagues, in 1988, and some encouraging results with rhodamine dyes have been reported by Ray and colleagues, in 2002. Efficient water-soluble dyes could lead to significant improvements and simplifications in high-power tunable lasers.

See also

Coherent Control: Experimental. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information. **Relativistic Nonlinear Optics**.

Further Reading

Costela A, García-Moreno I and Sastre R (2001) In: Nalwa HS (ed.) *Handbook of Advanced Electronic and Photonic Materials and Devices*, vol. 7, pp. 161–208. San Diego, CA: Academic Press.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. New York: Academic Press.

Demtröder W (1995) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1991) *High Power Dye Lasers*. Berlin: Springer-Verlag.

Duarte FJ (ed.) (1995) *Tunable Lasers Handbook*. New York: Academic Press.

Duarte FJ (ed.) (1995) *Tunable Laser Applications*. New York: Marcel Dekker.

Duarte FJ (2003) *Tunable Laser Optics*. New York: Elsevier Academic.

Duarte FJ and Hillman LW (eds) (1990) *Dye Laser Principles*. New York: Academic Press.

Maeda M (1984) *Laser Dyes*. New York: Academic Press.

Schäfer FP (ed.) (1990) *Dye Lasers*, 3rd edn. Berlin: Springer-Verlag.

Radziemski LJ, Solarz RW and Paisner JA (eds) (1987) *Laser Spectroscopy and its Applications*. New York: Marcel Dekker.

Weber MJ (2001) *Handbook of Lasers*. New York: CRC.

Edge Emitters

J J Coleman, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The semiconductor edge emitting diode laser is a critical component in a wide variety of applications, including fiber optics telecommunications, optical data storage, and optical remote sensing. In this section, we describe the basic structures of edge emitting diode lasers and the physical mechanisms for converting electrical current into light. Laser waveguides and cavity resonators are also outlined. The power, efficiency, gain, loss, and threshold characteristics of laser diodes are presented along with the effects of temperature and modulation. Quantum well lasers are outlined and a description of grating coupled lasers is provided.

Like all forms of laser, the edge emitting diode laser is an oscillator and has three principal components; a mechanism for converting energy into light, a medium that has positive optical gain, and a mechanism for obtaining optical feedback. The basic edge emitting laser diode is shown schematically in Figure 1. Current flow is vertical through a pn junction and light is emitted from the ends. The dimensions of the laser in Figure 1 are distorted, in proportion to typical real laser diodes, to reveal more detail. In practical laser diodes, the width of the stripe contact is actually much smaller than shown and the thickness is smaller still. Typical dimensions are (stripe width × thickness × length) $2 \times 100 \times 1000\ \mu\text{m}$. Light emission is generated only within a few micrometers of the surface, which means that 99.98% of the volume of this small device is inactive.

Energy conversion in diode lasers is provided by current flowing through a forward-biased pn junction. At the electrical junction, there is a high density of injected electrons and holes which can recombine in a direct energy gap semiconductor material to give an emitted photon. Charge neutrality requires equal numbers of injected electrons and holes. Figure 2 shows a simplified energy versus momentum (E–k) diagrams for (a) direct, and (b) indirect semiconductors. In a direct energy gap material, such as GaAs, the energy minima have the same momentum vector, so recombination of an electron in the conduction band and a hole in the valence band takes place directly. In an indirect material, such as silicon, momentum conservation requires the participation of a third particle – a phonon. This third-order process is far less efficient than direct recombination and, thus far, too inefficient to support laser action.

Optical gain in direct energy gap materials is obtained when population inversion is reached.

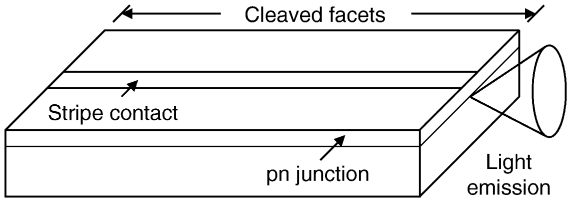

Figure 1 Schematic drawing of an edge emitting laser diode.

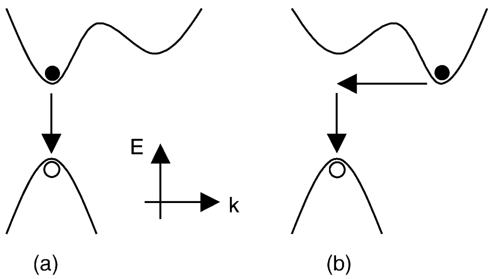

Figure 2 Energy versus momentum for (a) direct, and (b) indirect semiconductors.

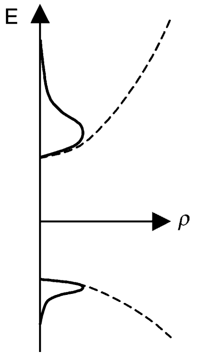

Figure 3 Density of states versus energy for conduction and valence bands.

Population inversion is the condition where the probability for stimulated emission exceeds that of absorption. The density of excited electrons n, in the conduction band is given by

$$n = \int_{E_c}^{\infty} \rho_n(E) f_n(E) \mathrm{d}E \text{ cm}^{-3} \qquad [1]$$

where $\rho_n(E)$ is the conduction band density of states function and $f_n(E)$ is the Fermi occupancy function. A similar equation can be written for holes in the valence band. Figure 3 shows the density of states ρ versus energy as a dashed line for conduction and valence bands. The solid lines in Figure 3 are the $\rho(E)f(E)$ products, and the areas under the solid curves are the carrier densities n and p.

The Fermi functions are occupation probabilities based on the quasi-Fermi levels E_{Fn} and E_{Fp}, which are, in turn, based on the injected carrier densities, δn and δp. Population inversion implies that the difference between the quasi-Fermi levels must exceed the emission energy which, for semiconductors, must be greater than the bandgap energy:

$$E_{\mathrm{Fn}} - E_{\mathrm{Fp}} \geq \hbar\omega \sim E_g \qquad [2]$$

Charge neutrality requires that $\delta n = \delta p$. Transparency is the point where these two conditions are just met and any additional injected current results in optical gain. The transparency current density J_o, is given by

$$J_o = \frac{q n_o L_z}{\tau_{\mathrm{spon}}} \qquad [3]$$

where n_o is the transparency carrier density ($\delta n = \delta p = n_o$ at transparency), L_z is the thickness of the optically active layer, and τ_{spon} is the spontaneous carrier lifetime in the material (typically 5 nsec).

A resonant cavity for optical feedback is obtained simply by taking advantage of the natural crystal formation in single crystal semiconductors. Most III–V compound semiconductors form naturally into a zincblende lattice structure. With the appropriate choice of crystal planes and surfaces, this lattice structure can be cleaved such that nearly perfect plane–parallel facets can be formed at arbitrary distances from each other. Since the refractive index of these materials is much larger (~3.5) than that of air, there is internal optical reflection for normal incidence of approximately 30%. This type of optical cavity is called a Fabry–Perot resonator.

Effective lasers of all forms make use of optical waveguides to optimize the overlap of the optical field with material gain and cavity resonance. The optical waveguide in an edge emitting laser is best described by considering it in terms of a transverse waveguide component, defined by the epitaxial growth of multiple thin heterostructure layers, and a lateral waveguide component, defined by conventional semiconductor device processing methods. One of the simplest, and yet most common, heterostructure designs is shown in Figure 4. This five-layer separate confinement heterostructure (SCH) offers excellent carrier confinement in the active layer as well as a suitable transverse waveguide.

Clever choice of different materials for each layer results in the energy band structure, index of refraction

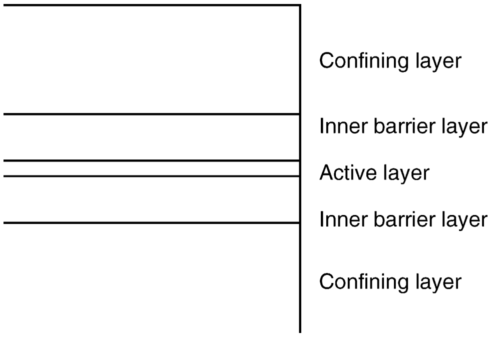

Figure 4 Schematic cross-section of a typical five-layer separate confinement heterostructure (SCH) laser.

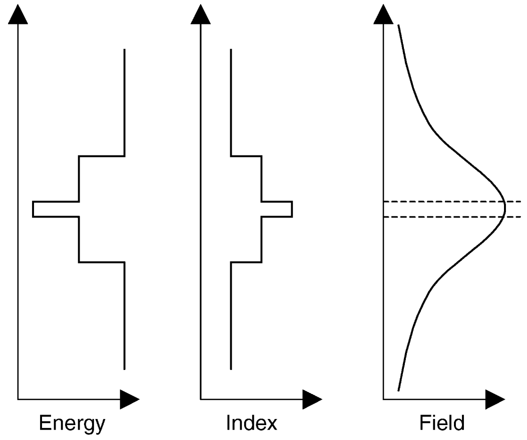

Figure 5 Energy band structure, index of refraction profile, and optical field mode profile of the SCH laser of Figure 4.

profile, and optical field mode profile shown in Figure 5. The outer confining layers are the thickest (~1 μm) and have the largest energy bandgap and lowest refractive index. The inner barrier layers are thinner (~0.1 μm), have intermediate bandgap energy and index of refraction, and serve both an optical role in defining the optical waveguide and an electronic role in confining carriers to the active layer. The active layer is the narrowest bandgap, highest index material, and is the only layer in the structure designed to provide optical gain. It may be a single layer or, as in the case of a multiple quantum well laser, a combination of layers. The bandgap energy, E, of this active layer plays a key role in determining the emission wavelength, λ, of the laser:

$$E = \frac{hc}{\lambda} \qquad [4]$$

where h is Planck's constant and c is the speed of light.

The bandgap energy of the SCH structure is shown in Figure 5. The lowest energy active layer collects injected electrons and holes and the carrier confinement provided by the inner barrier layers allows the high density of carriers necessary for population inversion and gain. The layers in the structure also have the refractive index profile shown in Figure 5 which forms a five-layer slab dielectric waveguide. With appropriate values for thicknesses and indices of refraction, this structure can be a very strong fundamental mode waveguide with the field profile shown. A key parameter of edge emitting diode lasers is Γ, the optical confinement factor. This parameter is the areal overlap of the optical field with the active layer, shown as dashed lines in the figure and can be as little as a few percent, even in very high performance lasers.

Lateral waveguiding in a diode laser can be accomplished in a variety of ways. A common example of an index guided laser is the ridge waveguide laser shown in Figure 6. After the appropriate transverse waveguide heterostructure sample is grown, conventional semiconductor processing methods are used to form parallel etched stripes from the surface near, but not through, the active layer. An oxide mask is patterned such that electrical contact is formed only to the center (core) region between the etched stripes, the core stripe being only a few microns wide. The lateral waveguide that results arises from the average index of refraction (effective index) in the etched regions outside the core being smaller than that of the core. Figure 7 shows the asymmetric near field emission profile of the ridge waveguide laser and cross-sectional profiles of the laser emission and effective index of refraction.

The optical spectrum of an edge emitting diode laser below and above threshold is the superposition of the material gain of the active layer and the resonances provided by the Fabry–Perot resonator. The spectral variation of material gain with injected carrier density (drive current) is shown in Figure 8. As the drive is increased, the intensity and energy of peak gain both increase.

The Fabry–Perot resonator has resonances every half wavelength, given by

$$L = \frac{m}{2}\frac{\lambda}{n} \qquad [5]$$

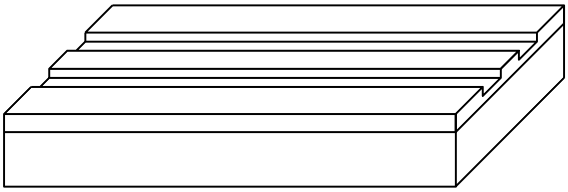

Figure 6 Schematic diagram of a real index guided ridge waveguide diode laser.

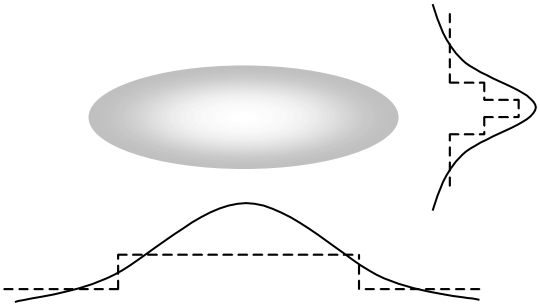

Figure 7 Near field emission pattern, cross-sectional emission profiles (solid lines), and refractive index profiles (dashed lines) from an index guided diode laser.

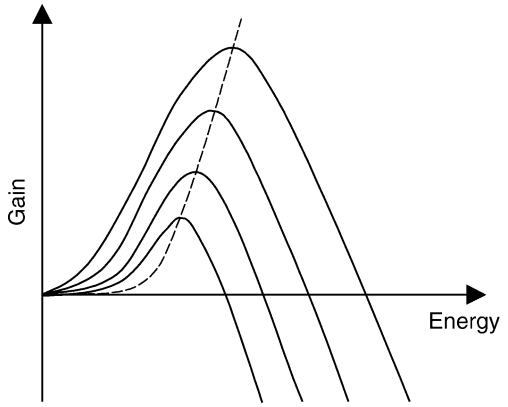

Figure 8 Gain versus emission energy for four increasing values of carrier (current) density. Peak gain is shown as a dashed line.

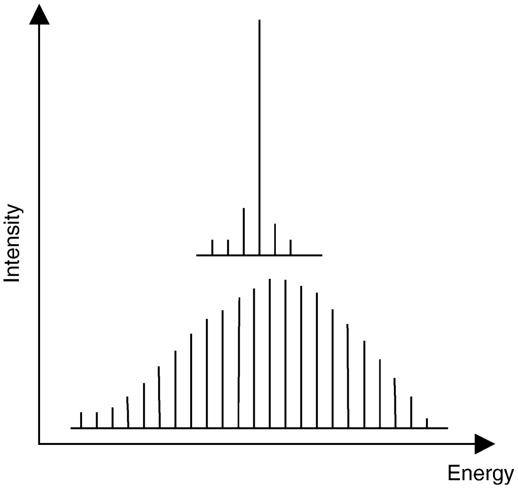

Figure 9 Emission spectra from a Fabry–Perot edge emitting laser structure at, and above, laser threshold.

where L is the cavity length, λ/n is the wavelength in the medium (n is the refractive index), and m is an integer. For normal length edge emitter cavities, these cavity modes are spaced only 1 or 2 Å apart. The result is optical spectra, near laser threshold and above, that look like the spectra of Figure 9. Many cavity modes resonate up to threshold, while above threshold, the superlinear increase in emission intensity with drive current tends to greatly favor one or two of the Fabry–Perot longitudinal cavity modes.

Laser threshold can be determined from a relatively simple analysis of the round trip gain and losses in the Fabry–Perot resonator. If the initial intensity is I_o, after one round trip the intensity will be given by

$$I = I_o R_1 R_2 e^{(g-\alpha_i)2L} \qquad [6]$$

where R_1 and R_2 are the reflectivities of the two facets, g is the optical gain, α_i are losses associated with optical absorption (typically 3–15 cm^{-1}), and L is the cavity length. At laser threshold, $g = g_{th}$ and $I = I_o$, i.e., the round trip gain exactly equals the losses. The result is

$$g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [7]$$

where the second term represents the mirror losses α_m. Of course, mirror losses are desirable in the sense that they represent useable output power. Actually, the equation above does not take into account the incomplete overlap of the optical mode with the gain in the active layer. The overlap is defined by the confinement factor Γ, described above, and the equation must be modified to become

$$\Gamma g_{th} = \alpha_i + \frac{1}{2L} \ln \frac{1}{R_1 R_2} \qquad [8]$$

For quantum well lasers, the material peak gain, shown in Figure 8, as a function of drive current is given approximately by

$$g_t = \beta J_o \ln \frac{J}{J_o} \qquad [9]$$

where J_o is the transparency current density. The threshold current density J_{th}, which is the current density where the gain exactly equals all losses, is given by

$$J_{th} = \frac{J_o}{\eta_i} \exp \frac{\alpha + \frac{1}{2L} \ln \frac{1}{R_1 R_2}}{\Gamma \beta J_o} \qquad [10]$$

Note that an additional parameter η_i has appeared in this equation. Not all of the carriers injected by the drive current participate in optical processes; some are lost to other nonradiative processes. This is accounted for by including the internal quantum efficiency term η_i, which is typically at least 90%. The actual drive current, of course, must include the geometry of the device and is given by

$$I_{th} = wLJ_{th} \qquad [11]$$

where L is the cavity length, and w is the effective width of the active volume. The effective width is basically the stripe width of the core but also includes current spreading and carrier diffusion under the stripe.

The power–current characteristic (P–I) for a typical edge emitting laser diode is shown in Figure 10. As the drive current is increased

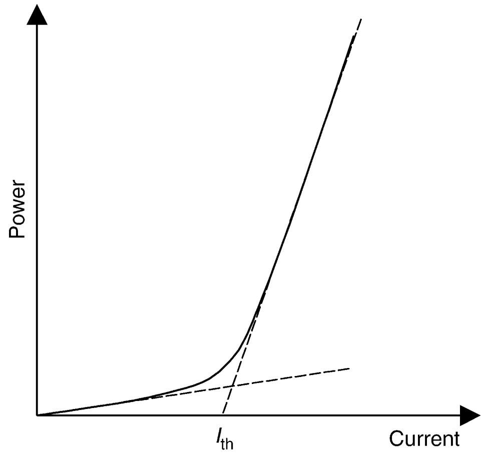

Figure 10 Power versus current (P–I) curve for a typical edge emitting laser diode.

from zero, the injected carrier densities increase and spontaneous recombination is observed. At some point the gain equals the internal absorption losses, and the material is transparent. Then, as the drive current is further increased, additional gain eventually equals the total losses, including mirror losses, and laser threshold is reached. Above threshold, nearly all additional injected carriers contribute to stimulated emission (laser output). The power generated internally is given by

$$P = wL(J - J_{\text{th}})\eta_{\text{i}}\frac{h\nu}{q} \qquad [12]$$

where wL is the area and $h\nu$ is the photon energy. Only a fraction of the power generated internally is extracted from the device

$$P_{\text{o}} = wL(J - J_{\text{th}})\eta_{\text{i}}\frac{h\nu}{q}\frac{\alpha_{\text{m}}}{\alpha_{\text{i}} + \alpha_{\text{m}}} \qquad [13]$$

Clearly a useful design goal is to minimize the internal optical loss α_{i}. The slope of the P–I curve of Figure 10 above threshold is described by an external differential quantum efficiency η_{d}, which is related to the internal quantum efficiency by

$$\eta_{\text{d}} = \eta_{\text{i}}\left[\frac{\alpha_{\text{m}}}{\alpha_{\text{i}} + \alpha_{\text{m}}}\right] \qquad [14]$$

Other common parameters used to characterize the efficiency of laser diodes include the power conversion efficiency η_{p}

$$\eta_{\text{p}} = \frac{P_{\text{o}}}{V_{\text{F}}I} \qquad [15]$$

and the wall plug efficiency η_{w}

$$\eta_{\text{w}} = \frac{P_{\text{o}}}{I} \qquad [16]$$

The power conversion efficiency η_{p}, recognizes that the forward bias voltage V_{F} is larger than the photon energy by an amount associated with the series resistance of the laser diode. The wall plug efficiency η_{w}, has the unusual units of WA^{-1} and is simply a shorthand term that relates the common input unit of current to the common output unit of power. In high performance low power applications, the ideal device has minimum threshold current since the current used to reach threshold contributes little to light emission. In high power applications, the threshold current becomes insignificant and the critical parameter is external quantum efficiency.

Temperature effects may be important for edge emitting lasers, depending on a particular application. Usually, concerns related to temperature arise under conditions of high temperature operation ($\text{T} > 50\,^{\circ}\text{C}$), where laser thresholds rise and efficiencies fall. It became common early in the development of laser diodes to define a characteristic temperature T_{o} for laser threshold, which is given by

$$I_{\text{th}} = I_{\text{th}o}\exp\left(\frac{T}{T_{\text{o}}}\right) \qquad [17]$$

where a high T_{o} is desirable. Unfortunately, this simple expression does not describe the temperature dependence very well over a wide range of temperatures, so the appropriate temperature range must also be specified. For applications where the emission wavelength of the laser is important, the change in wavelength with temperature $\text{d}\lambda/\text{d}T$ may also be specified. For conventional Fabry–Perot cavity lasers, $\text{d}\lambda/\text{d}T$ is typically 3–5 $\text{Å}\,\text{C}^{-1}$.

In order for a diode laser to carry information, some form of modulation is required. One method for obtaining this is direct modulation of the drive current in a semiconductor laser. The transient and temporal behavior of lasers is governed by rate equations for carriers and photons which are necessarily coupled equations. The rate equation for carriers is given by

$$\frac{\text{d}n}{\text{d}t} = \frac{J}{qL_{\text{z}}} - \frac{n}{\tau_{\text{sp}}} - (c/n)\beta(n - n_{\text{o}})\varphi(E) \qquad [18]$$

where J/qL_{z} is the supply, n/τ_{sp} is the spontaneous emission rate, and the third term is the stimulated emission rate $(c/n)\beta(n - n_{\text{o}})\varphi(E)$ where β is the gain coefficient and $\varphi(E)$ is the photon density.

The rate equation for photons is given by

$$\frac{d\varphi}{dt} = (c/n)\beta(n - n_o)\varphi + \frac{\theta n}{\tau_{sp}} - \frac{\varphi}{\tau_p} \qquad [19]$$

where θ is the fraction of the spontaneous emission that couples into the mode (a small number), and τ_p is the photon lifetime (~1 psec) in the cavity. These rate equations can be solved for specific diode laser materials and structures to yield a modulation frequency response. A typical small signal frequency response, at two output power levels, is shown in Figure 11. The response is typically flat to frequencies above 1 GHz, rises to a peak value at some characteristic resonant frequency, and quickly rolls off. The characteristic small signal resonant frequency ω_r, is given by

$$\omega_r^2 = \frac{(c/n)\beta\bar{\varphi}}{\tau_p} \qquad [20]$$

where $\bar{\varphi}$ is the average photon density. Direct modulation of edge emitting laser diodes is limited by practicality to less than 10 GHz. This is in part because of the limits imposed by the resonant frequency and in part by chirp. Chirp is the frequency modulation that arises indirectly from direct current modulation. Modulation of the current modulates the carrier densities which, in turn, results in modulation of the quasi-Fermi levels. The separation of the quasi-Fermi levels is the emission energy (wavelength). Most higher-speed lasers make use of external modulation schemes.

The discussion thus far has addressed aspects of semiconductor diode edge emitting lasers that are common to all types of diode lasers irrespective of the choice of materials or the details of the active layer structure. The materials of choice for efficient laser devices include a wide variety of III–V binary compounds and ternary or quaternary alloys. The particular choice is based on such issues as emission wavelength, a suitable heterostructure lattice match, and availability of high-quality substrates. For example, common low-power lasers for CD players ($\lambda \sim 820$ nm) are likely to be made on a GaAs substrate using a GaAs–$Al_xGa_{1-x}As$ heterostructure. Lasers for fiberoptic telecommunications systems ($\lambda \sim 1.3$–1.5 µm) are likely to be made on an InP substrate using an InP–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Red laser diodes ($\lambda \sim 650$ nm) are likely to be made on a GaAs substrate using a GaAs–$In_xGa_{1-x}As_yP_{1-y}$ heterostructure. Blue and ultraviolet lasers, a relatively new technology compared to the others, make use of $Al_xGa_{1-x}N$–GaN–$In_yGa_{1-y}N$ heterostructures formed on sapphire or silicon carbide substrates.

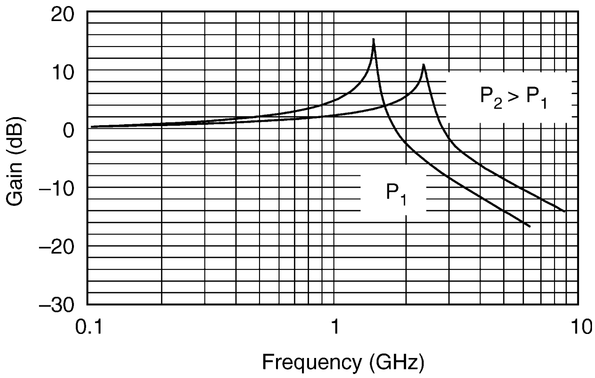

Figure 11 Direct current modulation frequency response for an edge emitting diode laser at two output power levels.

An important part of many diode laser structures is a strained layer. If a layer is thin enough, the strain arising from a modest amount of lattice mismatch can be accommodated elastically. In thicker layers, lattice mismatch results in an unacceptable number of dislocations that affect quantum efficiency, optical absorption losses, and, ultimately, long-term failure rates. Strained layer GaAs–$In_xGa_{1-x}As$ lasers are a critical component in rare-earth doped fiber amplifiers.

The structure of the active layer in edge emitting laser diodes was originally a simple double heterostructure configuration with a single active layer of 500–1000 Å in thickness. In the 1970s however, advances in the art of growing semiconductor heterostructure materials led to growth of high-quality quantum well active layers. These structures, having thicknesses that are comparable to the electron wavelength in a semiconductor (<200 Å), revolutionized semiconductor lasers, resulting in much lower threshold current densities, higher efficiencies, and a broader range of available emission wavelengths. The energy band diagram for a quantum well laser active region is shown in Figure 12.

This structure is a physical realization of the particle-in-a-box problem in elementary quantum mechanics. The quantum well yields quantum states in the conduction band at discrete energy levels, with odd and even electron wavefunctions, and breaks the degeneracy in the valence band, resulting in separate energy states for light holes and heavy holes. The primary transition for recombination is from the $n = 1$ electron state to the $h = 1$ heavy hole state and takes place at a higher energy than the bulk bandgap energy. In addition, the density of states for quantum wells becomes a step-like function and optical gain is enhanced. The advantages in practical edge

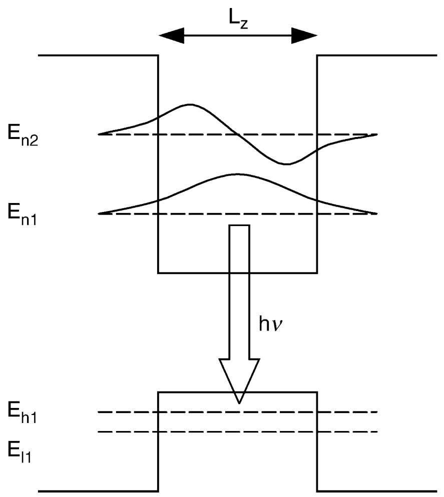

Figure 12 Energy band diagram for a quantum well heterostructure.

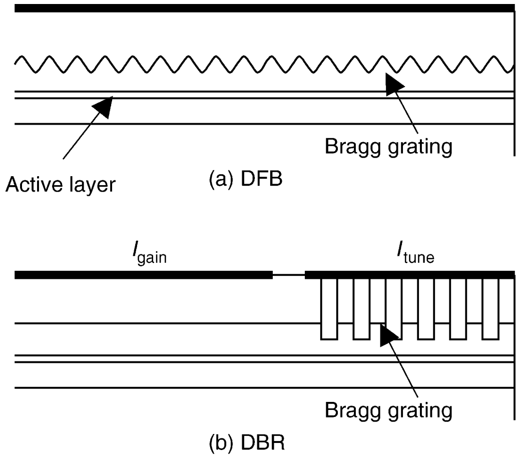

Figure 13 Schematic diagram of distributed feedback (DFB) and distributed Bragg reflector (DBR) edge emitting laser diodes.

emitting lasers that are the result of one-dimensional quantization are such that virtually all present commercial laser diodes are quantum well lasers. These kinds of advantages are driving development of laser diodes with additional degrees of quantization, including two dimensions (quantum wires) and three dimensions (quantum dots).

The Fabry–Perot cavity resonator described here is remarkably efficient and relatively easy to fabricate, which makes it the cavity of choice for many applications. There are many applications, however, where the requirements for linewidth of the laser emission or the temperature sensitivity of the emission wavelength make the Fabry–Perot resonator less desirable. An important laser technology that addresses both of these concerns involves the use of a wavelength selective grating as an integral part of the laser cavity. A notable example of a grating coupled laser is distributed feedback laser shown schematically in Figure 13a. This is similar to the SCH cross-section of Figure 4 with the addition of a Bragg grating above the active layer.

The period Λ, of the grating is chosen to fulfill the Bragg condition for mth-order coupling between forward- and backward-propagating waves, which is

$$\Lambda = \frac{m\lambda}{2n} \qquad [21]$$

where λ/n is the wavelength in the medium. The index step between the materials on either side of the grating is small and the amount of reflection is also small. Since the grating extends throughout the length of the cavity, however, the overall effective reflectivity can be large. Another variant is the distributed Bragg reflector (DBR) laser resonator, shown in Figure 13b. This DBR laser has a higher index contrast, deeply etched surface grating located at one or both ends of the otherwise conventional SCH laser heterostructure. One of the advantages of this structure is the opportunity to add a contact for tuning the Bragg grating by current injection.

Both the DBR and DFB laser structures are particularly well suited for telecommunications or other applications where a very narrow single laser emission line is required. In addition, the emission wavelength temperature dependence for this type of laser is much smaller, typically 0.5 $\text{Å}\,\text{C}^{-1}$.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

Further Reading

Agrawal GP (ed.) (1995) *Semiconductor Lasers Past, Present, and Future*. Woodbury, NY: American Institute of Physics.

Agrawal GP and Dutta NK (1993) *Semiconductor Lasers*. New York: Van Nostrand Reinhold.

Bhattacharya P (1994) *Semiconductor Optoelectronic Devices*. Upper Saddle River, NJ: Prentice-Hall.

Botez D and Scifres DR (eds) (1994) *Diode Laser Arrays*. Cambridge, UK: Cambridge University Press.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: John Wiley & Sons.

Coleman JJ (1992) *Selected Papers on Semiconductor Diode Lasers*. Bellingham, WA: SPIE Optical Engineering Press.

Einspruch NG and Frensley WR (eds) (1994) *Heterostructures and Quantum Devices*. San Diego, CA: Academic Press.

Iga K (1994) *Fundamentals of Laser Optics*. New York: Plenum Press.

Kapon E (1999) *Semiconductor Lasers I and II*. San Diego, CA: Academic Press.

Nakamura S and Fasol G (1997) *The Blue Laser Diode*. Berlin: Springer.

Verdeyen JT (1989) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Zory PS Jr (ed.) (1993) *Quantum Well Lasers*. San Diego, CA: Academic Press.

Excimer Lasers

J J Ewing, Ewing Technology Associates, Inc., Bellevue, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We present a summary of the fundamental operating principles of ultraviolet, excimer lasers. After a brief discussion of the economics and application motivation, the underlying physics and technology of these devices is described. Key issues and limitations in the scaling of these lasers are presented.

Background: Why Excimer Lasers?

Excimer lasers have become the most widely used source of moderate power pulsed ultraviolet sources in laser applications. The range of manufacturing applications is also large. Production of computer chips, using excimer lasers as the illumination source for lithography, has by far the largest commercial manufacturing impact. In the medical arena, excimer lasers are used extensively in what is becoming one of the world's most common surgical procedures, the reshaping of the lens to correct vision problems. Taken together, the annual production rate for these lasers is a relatively modest number compared to the production of semiconductor lasers. Current markets are in the range of $0.4 billion per year (Figure 1). More importantly, the unique wavelength, power, and pulse energy properties of the excimer laser enable a systems and medical procedure market, based on the excimer laser, to exceed $3 billion per year. The current average sales price for these lasers is in the range of $250 000 per unit, driven primarily by the production costs and reliability requirements of the lasers for lithography for chip manufacture. Relative to semiconductor diode lasers, excimer lasers have always been, and will continue to be, more expensive per unit device by a factor of order 10^4. UV power and pulse energy, however, are considerably higher.

The fundamental properties of these lasers enable the photochemical and photophysical processes used in manufacturing and medicine. Short pulses of UV light can photo-ablate materials, being of use in medicine and micromachining. Short wavelengths enable photochemical processes. The ability to provide a narrow linewidth using appropriate resonators enables precise optical processes, such as lithography. We trace out some of the core underlying properties of these lasers that lead to the core utility. We also discuss the technological problems and

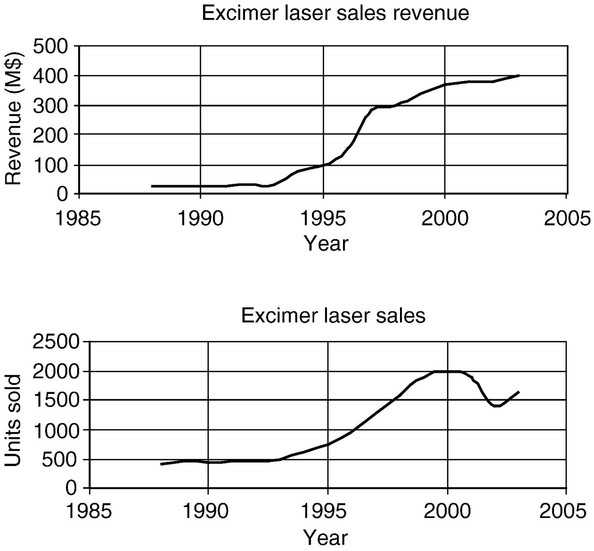

Figure 1 The sales of excimer lasers have always been measured in small numbers relative to other, less-expensive lasers. For many years since their commercial release in 1976 the sales were to R&D workers in chemistry, physics and ultimately biology and medicine. The market for these specialty but most useful R&D lasers was set to fit within a typical researcher's annual capital budget, i.e., less than $100K. As applications and procedures were developed and certified or approved, sales and average unit sales price increased considerably. Reliance on cyclical markets like semiconductor fabrication led to significant variations in production rates and annual revenues as can be seen.

solutions that have evolved to make these lasers so useful.

The data in Figure 1 provide a current answer to the question of 'why excimer lasers?' At the dawn of the excimer era, the answer to this question was quite different. In the early 1970s, there were no powerful short wavelength lasers, although IR lasers were being scaled up to significant power levels. However, visible or UV lasers could in principle provide better propagation and focus to a very small spot size at large distances. Solid state lasers of the time offered very limited pulse repetition frequency and low efficiency. Diode pumping of solid state lasers, and diode arrays to excite such lasers, was a far removed development effort. As such, short-wavelength lasers were researched over a wide range of potential media and lasing wavelengths. The excimer concept was one of those candidates.

We provide, in Figure 2, a 'positioning' map showing the range of laser parameters that can be obtained with commercial or developmental excimer lasers. The map has coordinates of pulse energy and pulse rate, with diagonal lines expressing average power. We show where both development goals and current markets lay in a map. Current applications are shown by the boxes in the lower right-hand corner. Excimer lasers have been built with clear apertures in the range of a fraction of 0.1 to 10^4 cm^2, with single pulse energies covering a range of 10^7. Lasers with large apertures require electron beam excitation. Discharge lasers correspond to the more modest pulse energy used for current applications. In general, for energies lower than ~5 J, the excitation method of choice is a self-sustained discharge, clearly providing growth potential for future uses, if required.

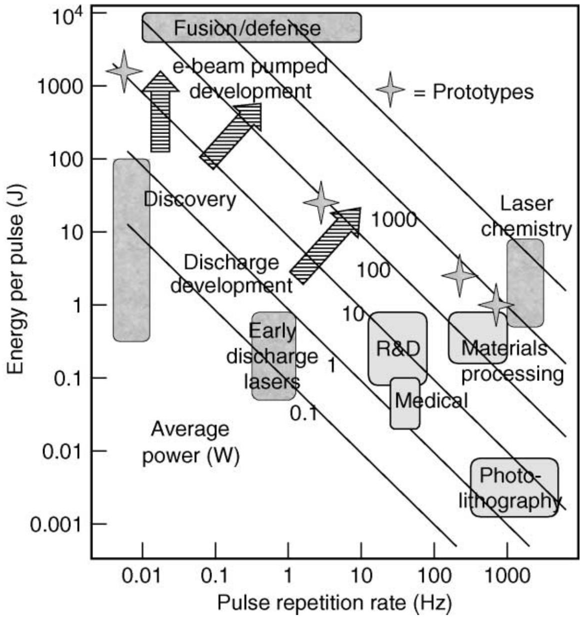

Figure 2 The typical energy and pulse rate for certain applications and technologies using excimer lasers. For energy under a few J, a discharge laser is used. The earliest excimer discharge lasers were derivatives of CO_2 TEA (transversely excited atmospheric pressure) and produced pulse energy in the range of 100 mJ per pulse. Pulse rates of order 30 Hz were all that the early pulsed power technology could provide. Early research focused on generating high energy. Current markets are at modest UV pulse energy and high pulse rates for lithography and low pulse rates for medical applications.

The applications envisaged in the early R&D days were in much higher energy uses, such as in laser weapons, laser fusion, and laser isotope separation. The perceived need for lasers for isotope separation, laser-induced chemistry, and blue-green laser-based communications from satellites, drove research in high pulse rate technology and reliability extension for discharge lasers. This research resulted in prototypes, with typical performance ranges as noted on the positioning map that well exceed current market requirements. However, the technology and underlying media physics developed in these early years made possible advances needed to serve the ultimate real market applications.

The very important application of UV lithography requires high pulse rates and high reliability. These two features differentiate the excimer laser used for lithography from the one used in the research laboratory. High pulse rates, over 2000 Hz in current practice, and the cost of stopping a computer chip production line for servicing, drive the reliability requirements toward 10^{10} shots per service interval. In contrast, the typical laboratory experiments are more often in the range of 100 Hz and, unlike a lithography production line, do not run 24 hours a day, 7 days a week. The other large market currently is in laser correction of myopia and other imperfections in the human eye. For this use the lasers are more akin to the commercial R&D style laser in terms of pulse rate and single pulse energy. Not that many shots are needed to ablate tissue to make the needed correction, and shot life is a straightforward requirement. However, the integration of these lasers into a certified and accepted medical procedure and an overall medical instrument drove the growth of this market.

The excimer concept, and the core technology used to excite these lasers, is applicable over a broad range of UV wavelengths, with utility at specific wavelength ranges from 351 nm in the near UV to 157 nm in the vacuum UV (VUV). There were many potential excimer candidates in the early R&D phase (Table 1). Some of these

Table 1 Key excimer wavelengths

Excimer emitter	*λ (nm)*	*Comment*
Ar_2	126	Very short emission wavelength, deep in the Vacuum UV; inefficient as laser due to absorption, see Xe_2.
Kr_2	146	Broad band emitter and early excimer laser with low laser efficiency. Very efficient converter of electric power into Vacuum UV light.
F_2	157	Molecule itself is not an excimer, but uses lower level dissociation; candidate for next generation lithography.
Xe_2	172	The first demonstrated excimer laser. Very efficient formation and emission but excited state absorption limits laser efficiency. Highly efficient as a lamp.
ArF	193	Workhorse for corneal surgery and lithography.
KrCl	222	Too weak compared to KrF and not as short a wavelength as ArF.
KrF	248	Best intrinsic laser efficiency; numerous early applications but found major market in lithography. Significant use in other materials processing.
XeI	254	Never a laser, but high formation efficiency and excellent for a lamp. Historically the first rare gas halide excimer whose emission was studied at high pressure.
XeBr	282	First rare gas halide to be shown as a laser; inefficient laser due to excited state absorption, but excellent fluorescent emitter; a choice for lamps.
Br_2	292	Another halogen molecule with excimer like transitions that has lased but had no practical use.
XeCl	308	Optimum medium for laser discharge excitation.
Hg_2	335	Hg vapor is very efficiently excited in discharges and forms excimers at high pressure. Subject of much early research, but as in the case of the rare gas excimers such as Xe_2 suffer from excited state absorption. Despite the high formation efficiency, this excimer was never made to lase.
I_2	342	Iodine UV molecular emission and lasing was first of the pure halogen excimer-like lasers demonstrated. This species served as kinetic prototype for the F_2 'honorary' excimer that has important use as a practical VUV source.
XeF	351, 353	This excimer laser transition terminates on a weakly bound lower level that dissociates rapidly, especially when heated. The focus for early defense-related laser development; as this wavelength propagates best of this class in the atmosphere.
XeF	480	This very broadband emission of XeF terminates on a different lower level than the UV band. Not as easily excited in a practical system, so has not had significant usage.
HgBr	502	A very efficient green laser that has many of the same kinetic features of the rare gas halide excimer lasers. But the need for moderately high temperature for the needed vapor pressure made this laser less than attractive for UV operation.
XeO	540	This is an excimer-like molecular transition on the side of the auroral lines of the O atom. The optical cross-section is quite low and as a result the laser never matured in practice.

candidates can be highly efficient, >50% relative to deposited energy, at converting electrical power into UV or visible emission, and have found a parallel utility as sources for lamps, though with differing power deposition rates and configurations. The key point in the table is that these lasers are powerful and useful sources at very short wavelengths. For lithography, the wavelength of interest has consistently shifted to shorter and shorter wavelengths as the printed feature size decreased. Though numerous candidates were pursued, as shown in Table 1, the most useful lasers are those using rare gas halide molecules. These lasers are augmented by dye laser and Raman shifting to provide a wealth of useful wavelengths in the visible and UV.

All excimer lasers utilize molecular dissociation to remove the lower laser level population. This lower level dissociation is typically from an unbound or very weakly bound ground-state molecule. However, halogen molecules also provide laser action on excimer-like transitions, that terminate on a molecular excited state that dissociates quickly. The vacuum UV transition in F_2 is the most practical method for making very short wavelength laser light at significant power. Historically, the rare gas dimer molecules, such as Xe_2, were the first to show excimer laser action, although self absorption, low gain, and poor optics in the VUV limited their utility. Other excimer-like species, such as metal halides, metal rare gas continua on the edge of metal atom resonance lines, and rare gas oxides were also studied. The key differentiators of rare gas halides from other excimer-like candidates are strong binding in the excited state, lack of self absorption, and relatively high optical cross-section for the laser transition.

Excimer Laser Fundamentals

Excimer lasers utilize lower-level dissociation to create and sustain a population inversion. For example, in the first excimer laser demonstrated, Xe_2, the lasing species is one which is comprised of two atoms that do not form a stable molecule in the ground state, as xenon dimers in the ground state are not a stable molecule. In the lowest energy state, that with neither of the atoms electronically excited, the interaction between the atoms in a collision is primarily one of repulsion, save for a very weak 'van der Waals' attraction at larger internuclear separations. This is shown in the potential energy diagram of Figure 3. If one of the atoms is excited, for example by an electric discharge, there can be a chemical binding in the electronically excited state of the molecule relative to the energy of one atom with an electron excited and one atom in the ground state. We use the shorthand notation * to denote an electronically excited atomic molecular species, e.g., Xe^*. The excited dimer, excimer for short, will recombine rapidly at high pressure from atoms into the Xe_2^* electronically excited molecule. Radiative lifetimes of the order 10 nsec to 1 μsec are typical before photon emission returns the molecule to the lower level. Collisional quenching often reduces the lifetime below the radiative rate. For Xe_2^* excimers, the emission is in the vacuum ultraviolet (VUV).

There are a number of variations on this theme, as noted in Table 1. Species, other than rare gas pairs, can exhibit broadband emission. The excited state binding energy can vary significantly, changing the fraction of excited states that will form molecules. The shape of the lower potential energy curve can vary as well. From a semantic point of view heteronuclear, diatomic molecules, such as XeO, are not excimers, as they are not made of two of the same atoms. However, the more formal name 'exciplex' did not stick in the laser world; excimer laser being preferred. Large binding energy is more efficient for capturing excited atoms into excited excimer-type molecules in the limited time they have before emission. Early measurements of the efficiency of converting electrical energy deposited into fluorescent radiation showed that up to 50% of the energy input could result in VUV light emission. Lasers were not this efficient due to absorption.

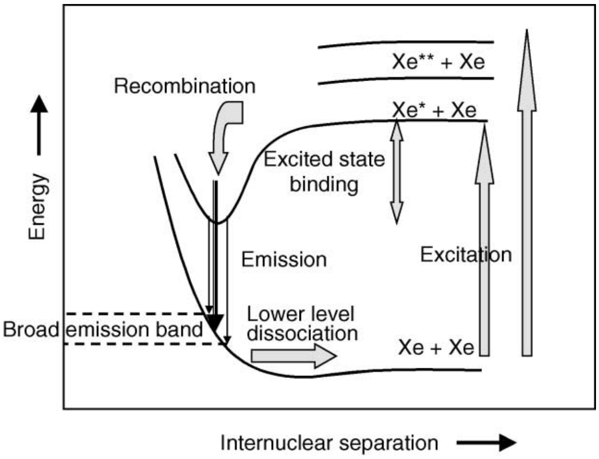

Figure 3 A schematic of the potential energy curves for an excimer species such as Xe_2. The excited molecule is bound relative to an excited atom and can radiate to a lower level that rapidly dissociates on psec timescales since the ground state is not bound, save for weak binding due to van der Waals forces. The emission band is intrinsically broad.

The diagram shown in Figure 3 is simplified because it does not show all of the excited states that exist for the excited molecule. There can be closely lying excited states that share the population and radiate at a lower rate, or perhaps absorb to higher levels, also not shown in Figure 3. Such excited state absorption may terminate on states derived from higher atomic excited states, noted as Xe^{**} in the figure, or yield photo-ionization, producing the diatomic ion-molecule plus an electron, such as:

$$Xe_2^* + h\nu \rightarrow Xe_2^+ + e \qquad [1]$$

Such excited state absorption limited the efficiency for the VUV rare gas excimers, even though they have very high formation and fluorescence efficiency.

Rare gas halide excimer lasers evolved from 'chemi-luminescence' chemical kinetics studies which were looking at the reactive quenching of rare gas metastable excited states in flowing after glow experiments. Broadband emission in reactions with halogen molecules was observed in these experiments, via reactions such as:

$$Xe^* + Cl_2 \rightarrow XeCl^* + Cl \qquad [2]$$

At low pressure the emissions from molecules, such as $XeCl^*$, are broad in the emissions bandwidth. Shortly after these initial observations, the laser community began examination of these species in high-pressure, electron-beam excited mixtures. The results differed remarkably from the low-pressure experiments and those for the earlier excimers such as Xe_2^*. The spectrum was shifted well away from the VUV emission. Moreover, the emission was much sharper at high pressure, though still a continuum with a bandwidth of the order of 4 nm. A basis for understanding is sketched in the potential energy curves of Figure 4. In this schematic we identify the rare gas atoms by Rg, excited rare gas atoms by

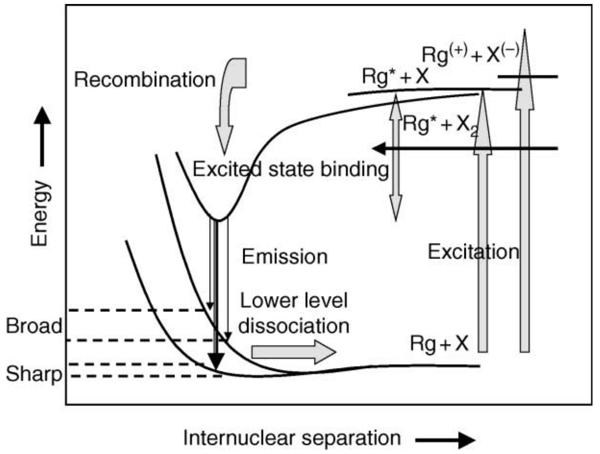

Figure 4 The potential energy curves for rare gas halides are sketched here. Both relatively sharp continuum emission is observed along with broad bands corresponding to different lower levels. The excited ion pair states consist of 3 distinct levels, two that are very close in energy and one that is shifted up from the upper laser level (the B state) by the spin orbit splitting of the rare gas ion.

Rg^*, and halogen atoms by X. Ions play a very important role in the binding and reaction chemistry, leading to excited laser molecules.

The large shift in wavelength from the VUV of rare gas dimer excimers, ~308 nm in $XeCl^*$ versus 172 nm in Xe_2^*, is due to the fact that the binding in the excited state of the rare gas halide is considerably stronger. The sharp and intense continuum at high pressure ($> \sim 100$ torr total pressure) is due to the fact that the 'sharp' transition terminates on a 'flat' portion of the lower-level potential energy curve. Indeed, in XeCl and XeF, the sharp band laser transition terminates on a slightly bound portion of the lower-level potential energy curve. Both the sharp emissions and broad bands are observed, due to the fact that some transitions from the excited levels terminate on a second, more repulsive potential energy curve.

An understanding of the spectroscopy and kinetic processes in rare gas plus halogen mixtures is based on the recognition that the excited state of a rare gas halide molecule is very similar to the ground state of the virtually isoelectronic alkali halide. The binding energy and mechanism of the excited state is very close to that of the ground state of the most similar alkali halide. Reactions of excited state rare gases are very similar to those of alkali atoms. For example, Xe^* and Cs are remarkably similar, from a chemistry and molecular physics point of view, as they have the same outer shell electron configuration and similar ionization potentials. The ionization potential of Xe^* is similar to that of cesium (Cs). Cs, and all the other alkali atoms, react rapidly with halogen molecules. Xenon metastable excited atoms react rapidly with halogen molecules. The alkali atoms all form ionically bonded ground states with halogens. The rare gas halides are effectively strongly bonded ion pairs, $Xe^{(+)}X^{(-)}$, that are very strongly bonded relative to the excited states that yield them. The difference between, for example, $XeCl^*$ and CsCl, is that $XeCl^*$ radiates in a few nanoseconds while CsCl is a stable ground-state molecule. The binding energy for these ionic bonded molecules is much greater than the more covalent type of bonding that is found in the rare gas excimer excited states. Thus XeI^*, which is only one electron different (in the I) than Xe_2^*, emits at 254 nm instead of 172 nm.

The first observed and most obvious connection to alkali atoms is in formation chemistry. Mostly, rare gas excited states react with halogen molecules rapidly. The rare gas halide laser analog reaction sequence is shown below, where some form of electric discharge excitation kicks the rare gas atom up to an excited state, so that it can react like an alkali atom, eqn [3]:

$$e_{\text{fast}} + Kr \rightarrow Kr^* + e_{\text{slow}} \qquad [3]$$

$$Kr^* + F_2 \rightarrow KrF^*(v, J)_{\text{high}} + F \qquad [4]$$

There are subtle and important cases where the neutral reaction, such as that shown in eqn [2] or [4], is not relevant as it is too slow compared to other processes. Note that the initial reaction does not yield the specific upper levels for the laser but states that are much higher up in the potential well of the excited rare gas halide excimer molecule. Further collisions with a buffer gas are needed to relax the states to the vibrational levels that are the upper levels for the laser transition. Such relaxation at high pressure can be fast, which leads the high-pressure spectrum to be much sharper than those first observed in flowing after-glow experiments.

The excited states of the rare gas halides are ion pairs bound together for their brief radiative lifetime. This opens up a totally different formation channel, indeed the one that often dominates the mechanism, ion–ion recombination. Long before excimer lasers were studied, it was well known that halogen molecules would react with 'cold' electrons (those with an effective temperature less than a few eV) in a process called attachment. The sequence leading to the upper laser level is shown below for the Kr/F_2 system, but the process is generically identical in other rare gas halide mixtures.

KrF^{*} formation kinetics via ion channel:

$$\text{Ionization:} \qquad \mathrm{e} + \mathrm{Kr} \rightarrow \mathrm{Kr}^{+} + 2\mathrm{e} \qquad [5]$$

$$\text{Attachment:} \qquad \mathrm{e} + \mathrm{F}_2 \rightarrow \mathrm{F}^{(-)} + \mathrm{F} \qquad [6]$$

Ion–ion recombination:

$$\mathrm{Kr}^{+} + \mathrm{F}^{(-)} + \mathrm{M} \rightarrow \mathrm{KrF}^{*}(v,J)_{\mathrm{high}} + \mathrm{M} \qquad [7]$$

Relaxation:

$$\mathrm{KrF}^{*}(v,J)_{\mathrm{high}} + \mathrm{M} \rightarrow \mathrm{KrF}^{*}(v,J)_{\mathrm{low}} + \mathrm{M} \qquad [8]$$

Finally, there is one other formation mechanism, displacement, that is unique to the rare gas halides. In this process, a heavier rare gas will displace a lighter rare gas ion in an excited state to form a lower-energy excited state:

$$\mathrm{ArF}^{*} + \mathrm{Kr} \rightarrow \mathrm{KrF}^{*}(v,J)_{\mathrm{high}} + \mathrm{Ar} \qquad [9]$$

In general, the most important excimer lasers tend to have all of these channels contributing to excited state formation for optimized mixtures.

These kinetic formation sequences express some important points regarding rare gas halide lasers. The source of the halogen atoms for the upper laser level disappears via both attachment reactions with electrons and by the reaction with excited states. The halogen atoms eventually recombine to make molecules, but at a rate too slow for sufficient continuous wave (cw) laser gain. For excimer-based lamps, however, laser gain is not a criterion and very efficient excimer lamps can be made. Another point to note is that forming the inversion requires transient species (and in some cases halogen fuels) that absorb the laser light. Net gain and extraction efficiency are a trade-off between pumping rate, an increase in which often forms more absorbers and absorption itself. Table 2 provides a listing of some of the key absorbers. Finally, the rare gas halide excited states can be rapidly quenched in a variety of processes. More often than not, the halogen molecule and the electrons in the discharge react with the excited states of the rare gas halide, removing the contributors to gain. We show, in Table 3, a sampling of some of the types of formation and quenching reactions with their kinetic rate constant parameters. Note that one category of reaction, three-body quenching, is unique to excimer lasers. Note also that the three-body recombination of ions has a rate coefficient that is effectively pressure dependent. At pressures above ~3 atm, the diffusion of ions toward each other is slowed and the effective recombination rate constant decreases.

Though the rare gas halide excited states can be made with near-unity quantum yield from the primary excitation, the intrinsic laser efficiency is not this high. The quantum efficiency is only ~25% (recall rare gas halides all emit at much longer wavelengths than rare gas excimers); there is inefficient extraction due to losses in the gas and windows. In practice there are also losses in coupling the power into the laser, and wasted excitation during the finite start-up time. The effect of losses is shown in Table 4 and Figure 5 where the pathways are charted and losses identified. In KrF, it takes ~20 eV to make a rare gas ion in an electron beam excited mixture,

Table 2 Absorbers in rare gas halide lasers

Species	*XeF*	*XeCl*	*KrF*	*ArF*
Laser λ (nm)	351, 353	308	248	193
F_2 absorption σ (cm^2)	8×10^{-21}	NA	1.5×10^{-20}	
Cl_2 absorption σ (cm^2)	NA	1.7×10^{-19}	NA	NA
HCl absorption σ (cm^2)	NA	Nil	NA	NA
$\mathrm{F}^{(-)}$ absorption σ (cm^2)	2×10^{-18}	NA	5×10^{-18}	5×10^{-18}
$\mathrm{Cl}^{(-)}$ absorption σ (cm^2)	NA	2×10^{-17}	NA	NA
$\mathrm{Rg}_2^{(+)}$ diatomic ion absorption σ (cm^2)^a	Ne: $\sim 10^{-18}$ Ar: $\sim 3 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-17}$	Ne: $\sim 10^{-17}$ Ar: $\sim 4 \times 10^{-17}$ Xe: $\sim 3 \times 10^{-18}$	Ne: $\sim 2.5 \times 10^{-17}$ Ar: $\sim 2 \times 10^{-18}$ Kr $< 10^{-18}$	Ne: $\sim 10^{-18}$ Ar: NIL
Other transient absorbers	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}	Excited states of rare gas atoms and rare gas dimer molecules $\sigma \sim 10^{-19}$–10^{-17}
Other	Lower laser level absorbs unless heated	Weak lower level absorption $\sigma \sim 4 \times 10^{-16}$	Windows and dust can be an issue	Windows and dust can be an issue

[a] Note that the triatomic rare gas halides of the form $\mathrm{Rg}_2\mathrm{X}$ will exhibit similar absorption as the diatomic rare gas ions as the triatomic rare gas halides of this form are essentially ion pairs of an $\mathrm{Rg}_2^{(+)}$ ion and the halide ion.

Table 3 Typical formation and quenching reactions and rate coefficients

Formation	
Rare gas plus halogen molecule	$Kr^* + F_2 \rightarrow KrF^*(v, J) + F$ $k \sim 7 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $KrF^* \sim 1$ $Ar^* + F_2 \rightarrow ArF^*(v, J) + F$ $k \sim 6 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $ArF^* \sim 60\%$ $Xe^* + F_2 \rightarrow XeF^*(v, J) + F$ $k \sim 7 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $XeF^* \sim 1$ $Xe^* + HCl(v = 0) \rightarrow Xe + H + Cl$ $k \sim 6 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $XeCl^* \sim 0$ $Xe^* + HCl(v = 1) \rightarrow XeCl^* + H$ $k \sim 2 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $XeCl^* \sim 1$
Ion–ion recombination	$Ar^{(+)} + F^{(-)} + Ar \rightarrow ArF^*(v, J)$ $k \sim 1 \times 10^{-25}\ \mathrm{cm^6/sec}$ at 1 atm and below $k \sim 7.5 \times 10^{-26}\ \mathrm{cm^6/sec}$ at 2 atm; Note effective 3-body rate constant decreases further as pressure goes over ~2 atm $Kr^{(+)} + F^{(-)} + Kr \rightarrow KrF^*(v, J)$ $k \sim 7 \times 10^{-26}\ \mathrm{cm^6/sec}$ at 1 atm and below; rolls over at higher pressure
Displacement	$ArF^*(v, J) + Kr \rightarrow KrF^*(v, J) + Ar$ $k \sim 2 \times 10^{-10}\ \mathrm{cm^3/sec}$; Branching ratio to $KrF^* \sim 1$
Quenching	
Halogen molecule	$XeCl^* + HCl \rightarrow Xe + Cl + HCl$ $k \sim 5.6 \times 10^{-10}\ \mathrm{cm^3/sec}$; $KrF^* + F_2 \rightarrow Kr + F + F$ $k \sim 6 \times 10^{-10}\ \mathrm{cm^3/sec}$; All halogen molecule quenching have very high reaction rate constants
3-body inert gas	$ArF^* + Ar + Ar \rightarrow Ar_2F^* + Ar$ $k \sim 4 \times 10^{-31}\ \mathrm{cm^6/sec}$ $KrF^* + Kr + Ar \rightarrow Kr_2F^* + Ar$ $k \sim 6 \times 10^{-31}\ \mathrm{cm^6/sec}$ $XeCl^* + Xe + Ne \rightarrow Xe_2Cl^* + Ne$ $k \sim 1 \times 10^{-33}\ \mathrm{cm^6/sec}$; $4 \times 10^{-31}\ \mathrm{cm^6/sec}$ with Xe as 3rd body These reactions yield a triatomic excimer at lower energy that can not be recycled back to the desired laser states
Electrons	$XeCl^* + e \rightarrow Xe + Cl + e$ $k \sim 2 \times 10^{-7}\ \mathrm{cm^3/sec}$ In 2-body quenching by electrons, the charge of the electron interacts with the dipole of the rare gas halide ion pair at long range; above value is typical of others
2-body inert gas	$KrF^* + Ar \rightarrow Ar + Kr + F$ $k \sim 2 \times 10^{-12}\ \mathrm{cm^3/sec}$ $XeCl^* + Ne \rightarrow Ne + Xe + Cl$ $k \sim 3 \times 10^{-13}\ \mathrm{cm^3/sec}$ 2-body quenching by rare gases is typically slow and less important than the reactions noted above

somewhat less in an electric discharge. The photon release is ~5 eV; the quantum efficiency is only 25%. Early laser efficiency measurements (using electron beam excitation) showed laser efficiency in the best cases slightly in excess of 10%, relative to the deposited energy. The discrepancy relative to the quantum efficiency is due to losses in the medium. For the sample estimate in Table 4 and Figure 5 we show approximate density of key species in the absorption chain, estimated on a steady-state approximation of the losses in a KrF laser for an excitation rate of order 1 MW/cc, typical of fast discharge lasers. The net small signal gain is of the order of 5%/cm. The key absorbers are the parent halogen molecule, F_2, the fluoride ion $F^{(-)}$, and the rare gas excited state atoms. A detailed pulsed kinetics code will provide slightly different results due to transient kinetic effects and the intimate coupling of laser extraction to quenching and absorber dynamics. The ratio of gain to loss is usually in the range of 7 to 20, depending on the actual mixture used and the rare gas halide wavelength. The corresponding extraction efficiency is then limited to the range of ~50%. The small signal gain, g_o, is related to the power deposition

Table 4 Typical ground and excited state densities shown for KrF at $P_{in} \sim 1$ MW/cm^3

Species	*Amount particles/cm^3*	*Absorption σ (cm)^2*	*Loss %/cm*
F_2	$\sim 1.2 \times 10^{17}$	1.5×10^{-20}	0.18
Kr	$\sim 3 \times 10^{18}$	–	–
Buffer	4×10^{19}	–	–
$F^{(-)}$	$\sim 10^{14}$	5×10^{-18}	0.05
Rg^*, Rg^{**}	1×10^{15}	$\langle 10^{-18} \rangle$ depends on ratio of Rg^{**}/Rg^*	0.1
KrF (before dissociating)	$\sim 10^{12}$	2×10^{-16}	0.02
Rg_2F^*	8×10^{14}, Note this species density decreases as flux grows	1×10^{-18}	0.08
KrF^* (B)	2.5×10^{14}	2×10^{-16}	5 (gain)

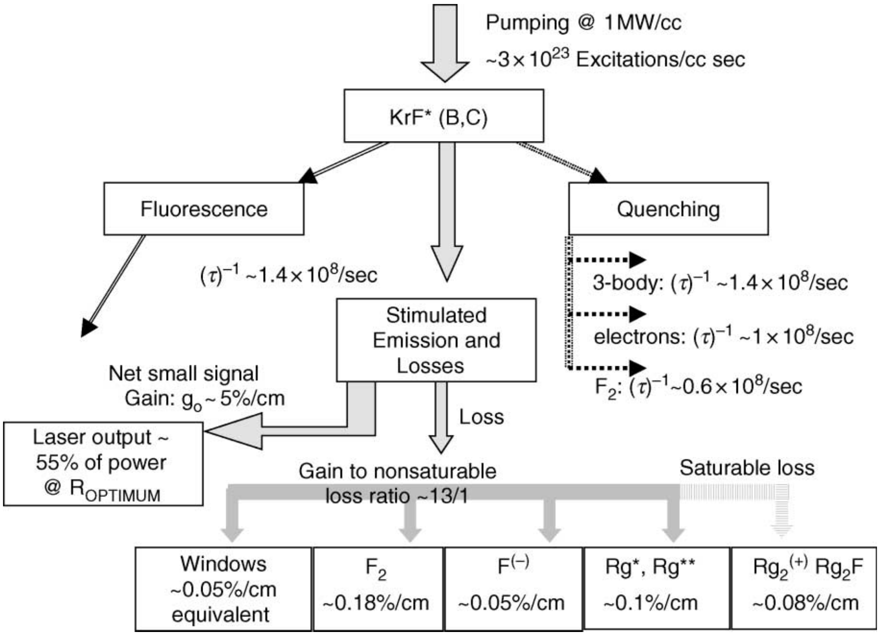
Pumping @ 1MW/cc
~3×10^{23} Excitations/cc sec
KrF* (B,C)
Fluorescence
Quenching
$(\tau)^{-1}$ ~1.4×10^8/sec
3-body: $(\tau)^{-1}$ ~1.4×10^8/sec
Stimulated Emission and Losses
electrons: $(\tau)^{-1}$ ~1×10^8/sec
Net small signal Gain: g$_o$~5%/cm
F$_2$: $(\tau)^{-1}$~0.6×10^8/sec
Loss
Laser output ~ 55% of power @ R$_{OPTIMUM}$
Gain to nonsaturable loss ratio ~13/1
Saturable loss
Windows ~0.05%/cm equivalent
F$_2$ ~0.18%/cm
F$^{(-)}$ ~0.05%/cm
Rg*, Rg** ~0.1%/cm
Rg$_2$$^{(+)}$ Rg$_2$F ~0.08%/cm

Figure 5 The major decay paths and absorbers that impact the extraction of excited states of rare gas halide excimers. Formation is via the ion–ion recombination, excited atom reaction with halogens and displacement reactions, not shown. We note for a specific mixture of KrF the approximate values of the absorption by the transient and other species along with the decay rates for the major quenching processes. Extraction efficiency greater than 50% is quite rare.

rate per unit volume, $P_{\text{deposited}}$, by eqn [10].

$$P_{\text{deposited}} = g_0 E^* (\tau_{\text{upper}} \eta_{\text{pumping}} \eta_{\text{branching}} f \sigma)^{-1} \qquad [10]$$

E^* is the energy per excitation, ~20 eV or 3.2×10^{-18} J/excitation, f is the fraction of excited molecules in the upper laser level, ~40% due to KrF* in higher vibrational states and the slowly radiating 'C' state. The laser cross-section is σ, and the η terms are efficiencies for getting from the primary excitation down to the upper laser level. The upper state lifetime, τ_{upper}, is the sum of the inverses of radiative and quenching lifetimes. When the laser cavity flux is sufficiently high, stimulated emission decreases the flow of power into the quenching processes.

The example shown here gives a ratio of net gain to the nonsaturating component of loss of 13/1. Of the excitations that become KrF*, only ~55% or so will be extracted in the best of cases. For other less ideal rare gas halide lasers, the gain to loss ratio, the extraction efficiency, and the intrinsic efficiency are all lower. Practical discharge lasers have lower practical wall plug efficiency due to a variety of factors, discussed below.

Discharge Technology

Practical discharge lasers use fast-pulse discharge excitation. In this scheme, an outgrowth of the CO_2 TEA laser, the gas is subjected to a rapid high-voltage pulse from a low-inductance pulse forming line or network, PFL or PFN. The rapid pulse serves to break down the gas, rendering it conductive with an electron density of $> \sim 10^{14}$ electrons/cm^3. As the gas becomes conductive, the voltage drops and power is coupled into the medium for a short duration, typically less than 50 nsec. The discharge laser also requires some form of 'pre-ionization', usually provided by sparks or a corona-style discharge on the side of the discharge electrodes. The overall discharge process is intrinsically unstable, and pump pulse duration needs to be limited to avoid the discharge becoming an arc.

Moreover, problems with impedance matching of the PFN or PFL with the conductive gas lead to both inefficiency and extra energy that can go into post-discharge arcs, which both cause component failure and produce metal dust which can give rise to window losses. A typical discharge laser has an aperture of a few cm^2 and a gain length of order 80 cm and produces outputs in the range of 100 mJ to 1 J. The optical pulse duration tends to be ~20 nsec, though the electrical pulse is longer due to the finite time needed for the gain to build up and the stimulated emission process to turn spontaneously emitted photons into photons in the laser cavity.

The pre-ionizer provides an initial electron density on the order of 10^8 electrons/cm^3. With spark pre-ionization or corona pre-ionization, it is difficult to make a uniform discharge over an aperture greater than a few cm^2. For larger apertures, one uses X-rays for pre-ionization. X-ray pre-ionized discharge excimer lasers have been scaled to energy well over 10 J per pulse and can yield pulse durations over 100 nsec. The complexity and expense of the X-ray approach, along with the lack of market for lasers in this size range, resulted in the X-ray approach remaining in the laboratory.

Aside from the kinetic issues of the finite time needed for excitation to channel into the excimer upper levels and the time needed for the laser to 'start up' from fluorescent photons, these lasers have reduced efficiency due to poor matching of the pump source to the discharge. Ideally, one would like to have a discharge such that when it is conductive the voltage needed to sustain the discharge is less than half of that needed to break the gas down. Regrettably for the rare gas halides (and in marked distinction to the CO_2 TEA laser) the voltage that the discharge sustains in a quasi-stable mode is ~20% of the breakdown voltage, perhaps even less, depending on the specific gases and electrode shapes.

Novel circuits, which separate the breakdown phase from the fully conductive phase, can readily double the efficiency of an excimer discharge laser, though this approach is not necessary in many applications. The extra energy that is not dissipated in the useful laser discharge often results in arcing or 'hot' discharge areas after the main laser pulse, leading to electrode sputtering, dust that finds its way to windows, and the need to service the laser after ~10^8 pulses. When one considers the efficiency of the power conditioning system, along with the finite kinetics of the laser start-up process in short pulses, and the mismatch of the laser impedance to the PFN/PFL impedance, lasers that may have 10% intrinsic efficiency in the medium in e-beam excitation may only have 2% to 3% wall plug efficiency in a practical discharge laser.

A key aspect of excimer technology is the need for fast electrical circuits to drive the power into the discharge. This puts a significant strain on the switch that is used to start the process. While spark gaps, or multichannel spark gaps, can handle the required current and current rate of rise, they do not present a practical alternative for high-pulse rate operation. Thyratrons and other switching systems, such as solid state switches, are more desirable than a spark gap, but do not offer the desired needed current rate of rise. To provide the pumping rate of order 1 MW/cm^3 needed for enough gain to turn the laser on before the discharge becomes unstable, the current needs to rise to a value of the order of 20 kA in a time of ~20 nsec; the current rate of rise is ~10^{12} A/sec. This rate of rise will limit the lifetime of the switch that drives the power into the laser gas. As a response to this need, magnetic switching and magnetic compression circuits were developed so that the lifetime of the switch can be radically increased. In the simplest cases, one stage of magnetic compression is used to make the current rate of rise to be within the range of a conventional thyratron switch as used in radar systems. Improved thyratrons and a simple magnetic assist also work. For truly long operating lifetimes, multiple stages of magnetic switching and compression are used, and with a transformer one can use a solid state diode as the start switch for the discharge circuit. A schematic of an excimer pulsed power circuit, that uses a magnetic assist with a thyratron and one stage of magnetic compression, is shown in Figure 6. The magnetic switching approach can also be used to provide two separate circuits to gain high efficiency by having one circuit to break down the gas (the spiker) and a second, low-impedance 'sustainer' circuit, to provide the main power flow. In this approach, the laser itself can act as part of the circuit by holding off the sustaining voltage for a μ-sec or so during the charge cycle and before the spiker breaks down the gas. A schematic of such a circuit is shown in Figure 7. Using this type of circuit, the efficiency of a XeCl laser can be over 4% relative to the wall plug. A wide variety of technological twists on this approach have been researched.

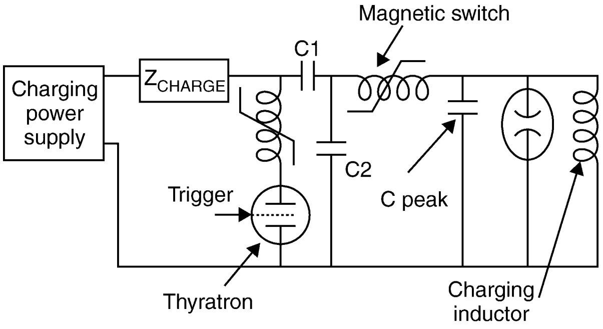

Figure 6 One of the variants on magnetic switching circuits that have been used to enhance the lifetime of the primary start switch. A magnetic switch in the form of a simple 1 turn inductor with a saturable core magnetic material such as a ferrite allows the thyratron to turn on before major current flows. Other magnetic switches may be used in addition to peak up the voltage rate of rise on the laser head or to provide significant compression. For example one may have the major current flow in the thyratron taking place on the 500 nsec time scale while the voltage pulse on the discharge is in the 50 nsec duration. A small current leaks through and is accommodated by the charging inductor.

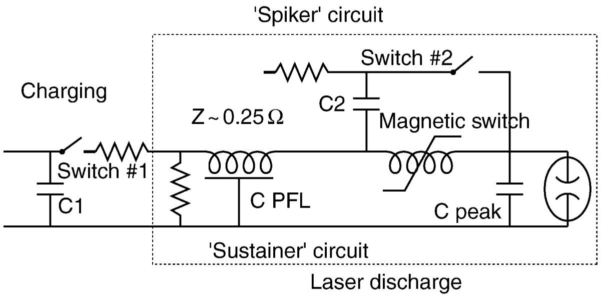

Figure 7 For the ultimate in efficiency the circuit can be arranged to provide a leading edge spike that breaks the gas down using a high impedance 'spiker' and a separate circuit that is matched to the discharge impedance when it is fully conductive. High coupling efficiency reduces waste energy that can erode circuit and laser head components via late pulse arcs. The figure shows a magnetic switch as the isolation but the scheme can be implemented with a second laser head as the isolation switch, a rail gap (not appropriate for long life) or a diode at low voltage.

Excimer lasers present some clear differences in resonator design compared to other lasers. The apertures are relatively large, so one does not build a TEM_{00} style cavity for good beam quality and expect to extract any major portion of the energy available. The typical output is divergent and multimode. When excimer lasers are excited by a short pulse discharge, ~30 nsec, the resulting gain duration is also short, ~20 nsec, limiting the number of passes a photon in a resonator can have in the gain. The gain duration of ~20 nsec only provides ~3.5 round trips of a photon in the cavity during the laser lifetime, resulting in the typical high-order multimode beam. Even with an unstable resonator, the gain duration is not long enough to collapse the beam into the lowest-order diffraction-limited output. If line narrowing is needed, and reasonable efficiency is required, an oscillator amplifier configuration is often used with the oscillator having an appropriate tuning resonator in the cavity. If both narrow spectral linewidth and excellent beam quality are needed one uses a seeded oscillator approach where the narrowband oscillator is used to seed an unstable resonator on the main amplifier. By injecting the seed into the unstable resonator cavity, a few nsec before the gain is turned on, the output of the seeded resonator can be locked in wavelength, bandwidth, and provide good beam quality. Long pulse excimer lasers, such as e-beam excited lasers or long pulse X-ray pre-ionized discharge devices can use conventional unstable resonator technology with good results.

Excimer lasers are gas lasers that run at both high pressure and high instantaneous power deposition rates. During a discharge pulse, much of the halogen bearing 'fuel' is burned out by the attachment and reactive processes. The large energy deposition per pulse means that pressure waves are generated. All of these combine with the characteristic device dimensions to require gas flow to recharge the laser mixture between pulses. For low pulse rates, less than ~200 Hz, the flow system need not be very sophisticated. One simply needs to flush the gas from the discharge region with a flow having a velocity of the order of 5 m/sec for a typical discharge device. The flow is transverse to the optical and discharge directions. At high pulse rates, the laser designer needs to consider flow in much more detail to minimize the power required to push the gas through the laser head. Circulating pressure waves need to be damped out so that the density in the discharge region is controlled at the time of the next pulse. Devices called acoustic dampers are placed in the sides of the flow loop to remove pressure pulses. A final subtlety occurs in providing gas flow over the windows. The discharge, post pulse arcs, and reactions of the halogen source with the metal walls and impurities creates dust. When the dust coats the windows, the losses in the cavity increase, lowering efficiency. By providing a simple gas flow over the window, the dust problem can be ameliorated. For truly long life, high-reliability lasers, one needs to take special care in the selection of materials for electrodes and insulators and avoid contamination, by assembling in clean environments.

See also

Lasers: Carbon Dioxide Lasers. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics,**

Applications: Pulse Compression via Nonlinear Optics; Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Ballanti S, Di Lazzaro P, Flora F, *et al.* (1998) Ianus the 3-electrode laser. *Applied Physics B* 66: 401–406.

Brau CA (1978) Rare gas halogen lasers. In: Rhodes CK (ed.) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer Verlag.

Brau CA and Ewing JJ (1975) Emission spectra of XeBr, XeCl, XeF, and KrF. *Journal of Chemical Physics* 63: 4640–4647.

Ewing JJ (2000) Excimer laser technology development. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 1061–1071.

Levatter J and Lin SC (1980) Necessary conditions for the homogeneous formation of pulsed avalanche discharges at high pressure. *Journal of Applied Physics* 51: 210–222.

Long WH, Plummer MJ and Stappaerts E (1983) Efficient discharge pumping of an XeCl laser using a high voltage prepulse. *Applied Physics Letters* 43: 735–737.

Rhodes CK (ed.) (1979) *Excimer Lasers, Topics of Applied Physics*, vol. 30. Berlin: Springer-Verlag.

Rokni M, Mangano J, Jacob J and Hsia J (1978) Rare gas fluoride lasers. *IEEE Journal of Quantum Electronics* QE-14: 464–481.

Smilanksi I, Byron S and Burkes T (1982) Electrical excitation of an XeCl laser using magnetic pulse compression. *Applied Physics Letters* 40: 547–548.

Taylor RS and Leopold KE (1994) Magnetic-spiker excitation of gas discharge lasers. *Applied Physics B, Lasers and Optics* 59: 479–509.

Free Electron Lasers

A Gover, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The Free Electron Laser (FEL) is an exceptional kind of laser. Its active medium is not matter, but charged particles (electrons) accelerated to high energies, passing in vacuum through a periodic undulating magnetic field. This distinction is the main reason for the exceptional properties of FEL: operating at a wide range of wavelengths – from mm-wave to X-rays with tunability, high power, and high efficiency.

In this article we explain the physical principles of FEL operation, the underlying theory and technology of the device and various operating schemes, which have been developed to enhance performance of this device.

The term 'Free Electron Laser' was coined by John Madey in 1971, pointing out that the radiative transitions of the electrons in this device are between free space (more correctly – unbound) electron quantum states, which are therefore states of continuous energy. This is in contrast to conventional atomic and molecular lasers, in which the electron performs radiative transition between bound (and therefore of distinct energy) quantum states. Based on these theoretical observations, Madey and his colleagues in Stanford University demonstrated FEL operation first as an amplifier (at $\lambda = 10.6\ \mu\text{m}$) in 1976, and subsequently as an oscillator (at $\lambda = 3.4\ \mu\text{m}$) in 1980.

From the historical point of view, it turned out that Madey's invention was essentially an extension of a former invention in the field of microwave-tubes technology – the Ubitron. The Ubitron, a mm-wave electron tube amplifier based on a magnetic undulator, was invented and developed by Philips and Enderbry who operated it at high power levels in 1960. The early Ubitron development activity was not noticed by the FEL developers because of the disciplinary gap, and largely because its research was classified at the time. Renewed interest in high-power mm-wave radiation emission started in the 1970s, triggered by the development of pulsed-line generators of 'Intense Relativistic Beams' (IRB). This activity, led primarily by plasma physicists in the defense establishment laboratories of Russia (mostly IAP in Gorky – Nizhny Novgorod) and the US (mostly N.R.L. – DC) led to development of high-gain high-power mm-wave sources independently of the development of the optical FEL. The connection between these devices and between them to conventional microwave tubes (as Traveling Wave Tubes – TWT) and other electron beam radiation schemes, like Cenenkov and Smith-Purcell radiation that may also be considered FELs, was revealed in the mid-1970s, starting with the theoretical works of P. Spangle, A. Gover and A. Yariv who identified that all these devices satisfy the same dispersion equation as the TWT derived by John Pierce in the 1940s. Thus, the optical FEL could be conceived as a kind of immense electron tube, operating with a high-energy electron beam in the low gain regime of the Pierce TWT dispersion equation.

The extension of the low-gain FEL theory to the general 'electron-tube' theory is important because it led to development of new radiation schemes and new operating regimes of the optical FEL. This was exploited by physicists in the discipline of accelerator physics and synchrotron radiation, who identified, starting with the theoretical works of C. Pellegrini and R. Bonifacio in the early 1980s, that high-current, high-quality electron beams, attainable with further development of accelerators technology, could make it possible to operate FELs in the high-gain regime, even at short wavelengths (vacuum ultra-violet – VUV and soft X-ray) and that the high-gain FEL theory can be extended to include amplification of the incoherent synchrotron spontaneous emission (shot noise) emitted by the electrons in the undulator. These led to the important development of the 'self (synchrotron) amplified spontaneous emission (SASE) FEL', which promised to be an extremely high brightness radiation source, overcoming the fundamental obstacles of X-ray lasers development: lack of mirrors (for oscillators) and lack of high brightness radiation sources (for amplifiers).

A big boost to the development of FEL technology was given during the period of the American 'strategic defense initiative – SDI' (Star-Wars) program in the mid-1980s. The FEL was considered one of the main candidates for use in a ground-based or space-based 'directed energy weapon – DEW', that can deliver megawatts of optical power to hit attacking missiles. The program led to heavy involvement of major American defense establishment laboratories (Lawrence–Livermore National Lab, Los-Alamos National Lab) and contracting companies (TRW, Boeing). Some of the outstanding results of this effort were demonstration of the high-gain operation of an FEL amplifier in the mm-wavelength regime, utilizing an Induction Linac (Livermore, 1985), and demonstration of enhanced radiative energy extraction efficiency in FEL oscillator, using a 'tapered wiggler' in an RF-Linac driven FEL oscillator (Los-Alamos, 1983). The program has not been successful in demonstrating the potential of FELs to operate at the high average power levels needed for DEW applications. But after the cold-war period, a small part of the program continues to support research and development of medical FEL application.

Principles of FEL Operation

Figure 1 displays schematically an FEL oscillator. It is composed of three main parts: an electron accelerator, a magnetic wiggler (or undulator), and an optical resonator.

Without the mirrors, the system is simply a synchrotron undulator radiation source. The electrons in the injected beam oscillate transversely to their propagation direction z, because of the transverse magnetic Lorenz force:

$$\mathbf{F}_\perp = -ev_z\hat{\mathrm{e}}_z \times \mathbf{B}_\perp \quad [1]$$

In a planar (linear) wiggler, the magnetic field on axis is approximately sinusoidal:

$$\mathbf{B}_\perp = B_\mathrm{w}\hat{e}_y \cos k_\mathrm{w} z \quad [2]$$

In a helical wiggler:

$$\mathbf{B}_\perp = B_\mathrm{w}(\hat{\mathrm{e}}_y \cos k_\mathrm{w} z + \hat{\mathrm{e}}_x \sin k_\mathrm{w} z) \quad [3]$$

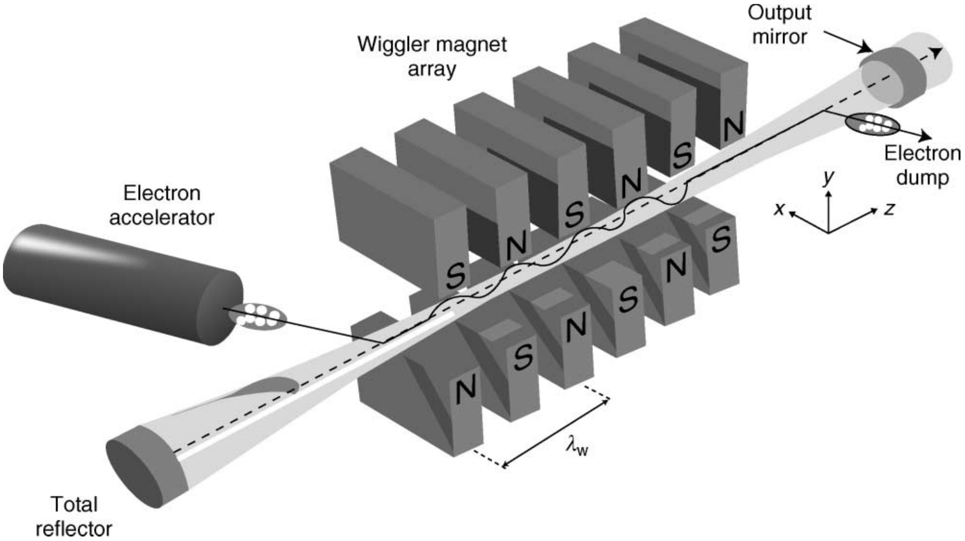

Figure 1 Components of a FEL-oscillator. (Reproduced from Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25. Illustration by Jaynie Martz.)

In either case, if we assume constant (for the planar wiggler – only on the average) axial velocity, then $z = v_z t$. The frequency of the transverse force and the mechanical oscillation of the electrons, as viewed transversely in the laboratory frame of reference, is:

$$\omega_{0s} = k_w v_z = 2\pi \frac{v_z}{\lambda_w} \qquad [4]$$

where $\lambda_w = 2\pi/k_w$ is the wiggler period.

The oscillating charge emits an electromagentic radiation wavepacket. In a reference frame moving with the electrons, the angular radiation pattern looks exactly like dipole radiation, monochromatic in all directions (except for the frequency-line-broadening due to the finite oscillation time, i.e., the wiggler transit time). In the laboratory reference-frame the radiation pattern concentrates in the propagation ($+z$) direction, and the Doppler up-shifted radiation frequency depends on the observation angle Θ relative to the z-axis:

$$\omega_0 = \frac{\omega_{0s}}{1 - \beta_z \cos\Theta} \qquad [5]$$

On axis ($\Theta = 0$), the radiation frequency is:

$$\omega_0 = \frac{c k_w \beta_z}{1 - \beta_z} = (1 + \beta_z)\beta_z \gamma_z^2 c k_w \cong 2\gamma_z^2 c k_w \qquad [6]$$

where $\beta_z \equiv v_z/c$, $\gamma_z \equiv (1 - \beta_z^2)^{-1/2}$ are the axial (average) velocity and the axial Lorenz factor, respectively, and the last part of the equation is valid only in the (common) highly relativistic limit $\gamma_z \gg 1$.

Using the relations $\beta_z^2 + \beta_\perp^2 = \beta^2$, $\beta_\perp = a_w/\gamma$, one can express γ_z:

$$\gamma_z = \frac{\gamma}{1 + a_w^2/2} \qquad [7]$$

(this is for a linear wiggler, in the case of a helical wiggler the denoninator is $1 + a_w^2$):

$$\begin{aligned} \gamma \equiv (1 - \beta^2)^{-1/2} &= 1 + \frac{\varepsilon_k}{mc^2} \\ &= 1 + \varepsilon_k\,[\text{MeV}]/0.511 \end{aligned} \qquad [8]$$

and a_w – (also termed K) ‘the wiggler parameter’ is the normalized transverse momentum:

$$a_w = \frac{eB_w}{k_w mc} = 0.093 B_w\,[\text{KGauss}]\lambda_w\,[\text{cm}] \qquad [9]$$

Typical values of B_w in FEL wigglers (undulators) are of the order of Kgauss’, and λ_w of the order of CMs, and consequently $a_w < 1$. Considering that electron beam accelerator energies are in the range of MeV to GeV, one can appreciate from eqns [6]–[8], that a significant relativistic Doppler shift factor $2\gamma_z^2$, in the range of tens to millions, is possible. It, therefore, provides incoherent synchrotron undulator radiation in the frequency range of microwave to hard X-rays.

Synchrotron undulator radiation was studied in 1951 and since then has been a common source of VUV radiation in synchrotron facilities. From the point of view of laser physics theory, this radiation can be viewed as ‘spontaneous synchrotron radiation emission’ in analogy to spontaneous radiation emission by electrons excited to higher bound-electron quantum levels in atoms or molecules. Alternatively, it can be regarded as the classical shot noise radiation, associated with the current fluctuations of the randomly injected discrete charges comprising the electron beam. Evidently this radiation is incoherent, and the fields it produces average in time to zero, because the wavepackets emitted by the randomly injected electrons interfere at the observation point with random phase. However, their energies sum up and can produce substantial power.

Based on fundamental quantum-electrodynamical principles or Einstein’s relations, one would expect that any spontaneous emission scheme can be stimulated. This principle lies behind the concept of the FEL, which is nothing but stimulated undulator synchrotron radiation. By stimulating the electron beam to emit radiation, it is possible, as with any laser, to generate a coherent radiation wave and extract more power from the gain medium, which in this case is an electron beam, that carries an immense amount of power. There are two kinds of laser schemes which utilize stimulation of synchrotron undulator radiation:

(i) *A laser amplifier.* In this case the mirrors in the schematic configuration of Figure 1 are not present, and an external radiation wave at frequencies within the emission range of the undulator is injected at the wiggler entrance. This requires, of course, an appropriate radiation source to be amplified and availability of sufficiently high gain in the FEL amplifier.

(ii) *A laser oscillator.* In this case an open cavity (as shown in Figure 1) or another (waveguide) cavity is included in the FEL configuration. As in any laser, the FEL oscillator starts building up its radiation from the spontaneous (synchrotron undulator) radiation which gets trapped in the resonator and amplified by stimulated emission along the wiggler. If the threshold condition is satisfied (having single path gain higher than the round trip losses), the oscillator arrives to saturation and steady state coherent operation after a short transient period of oscillation build-up.

Because the FEL can operate as a high-gain amplifier (with a long enough wiggler and an electron beam of high current and high quality), also a third mode of operation exists: self amplified spontaneous emission (SASE). In this case, the resonator mirrors in Figure 1 are not present and the undulator radiation generated spontaneously in the first sections of the long undulator is amplified along the wiggler and emitted at the wiggler exit at high power and high spatial coherence.

The Quantum-Theory Picture

A free electron, propagating in unlimited free space, can never emit a single photon. This can be proven by examining the conservation of energy and momentum conditions:

$$\varepsilon_{k_\mathrm{i}} - \varepsilon_{k_\mathrm{f}} = \hbar\omega \qquad [10]$$

$$\mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{f} = \mathbf{q} \qquad [11]$$

that must be satisfied, when an electron in an initial free-space energy and momentum state $(\varepsilon_{k_\mathrm{i}}, \hbar k_\mathrm{i})$ makes a transition to a final state $(\varepsilon_{k_\mathrm{f}}, \hbar k_\mathrm{f})$, emitting a single photon of energy and momentum $(\hbar\omega, q)$. In free space:

$$\varepsilon_k = c\sqrt{(\hbar \mathbf{k})^2 + (mc^2)^2} \qquad [12]$$

$$\mathbf{q} = \frac{\omega}{\mathrm{c}}\hat{\mathrm{e}}_q \qquad [13]$$

and eqns [10]–[13] have only one solution, $\omega = 0$, $\mathbf{q} = 0$. This observation is illustrated graphically in the energy–momentum diagram of Figure 2a in the framework of a one-dimensional model. It appears that if both eqns [10] and [11] can be satisfied, then the phase velocity of the emitted radiation wave $v_\mathrm{ph} = \omega/q$ (the slope of the chord) will equal the electron wavepacket group velocity $v_\mathrm{g} = v_\mathrm{z}$ at some intermediate point $k^* = p^*/\hbar$:

$$v_\mathrm{ph} = \frac{\omega}{q} = \frac{\varepsilon_{k\mathrm{i}} - \varepsilon_{k\mathrm{f}}}{\hbar(k_\mathrm{i} - k_\mathrm{f})} = \left.\frac{\partial \varepsilon}{\partial p}\right|_{\mathrm{p}^*} = v_\mathrm{g} \qquad [14]$$

For a radiation wave in free space (eqn [13]), this results in $c = v_\mathrm{g}$, which contradicts special relativity.

The reason for the failure to conserve both energy and momentum in the transition is that the photon momentum $\hbar q$ it too small to absorb the large momentum shift of the electron, as it recoils while releasing radiative energy $\hbar\omega$. This observation leads to ideas on how to make a radiative transition possible:

(i) *Limit the interaction length*. If the interaction length is L, the momentum conservation

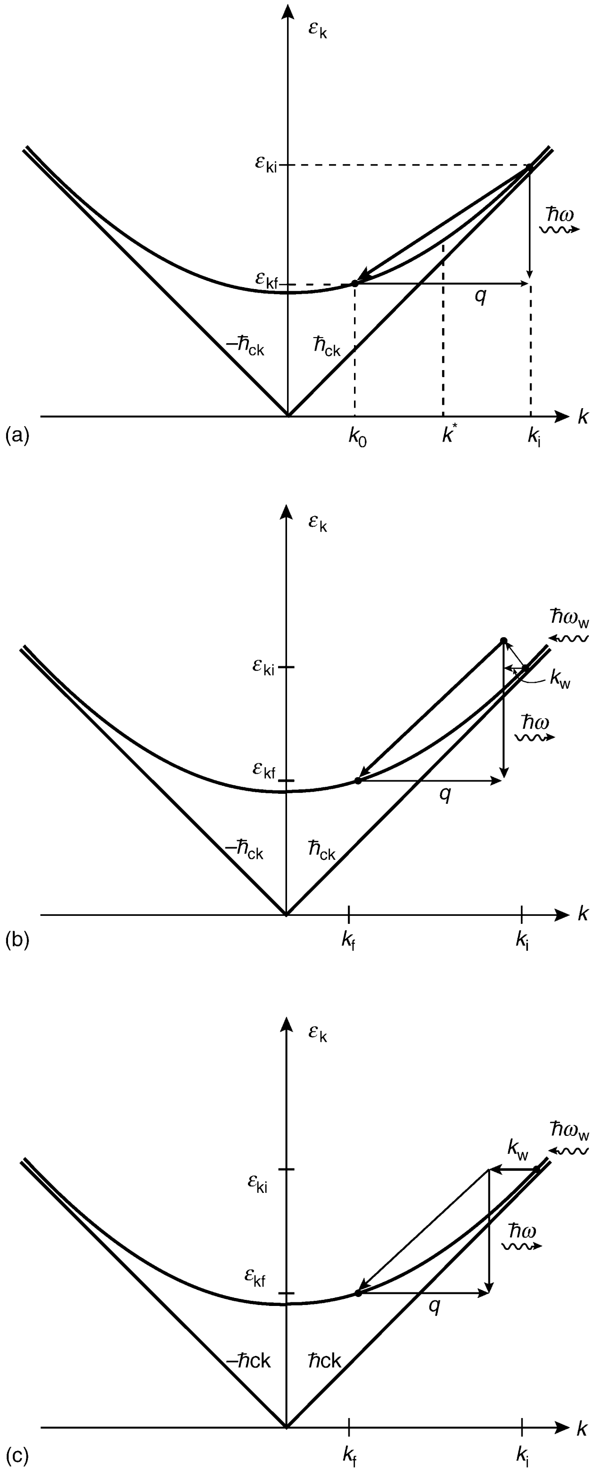
ε_k
ε_ki
$\hbar\omega$
ε_kf
q
$-\hbar\mathrm{ck}$
$\hbar\mathrm{ck}$
k
(a)
k_0
k^*
k_i
ε_k
$\hbar\omega_\mathrm{w}$
ε_ki
k_w
$\hbar\omega$
ε_kf
q
$-\hbar\mathrm{ck}$
$\hbar\mathrm{ck}$
k
(b)
k_f
k_i
ε_k
$\hbar\omega_\mathrm{w}$
k_w
ε_ki
$\hbar\omega$
ε_kf
q
$-\hbar\mathrm{ck}$
$\hbar\mathrm{ck}$
k
(c)
k_f
k_i

Figure 2 Conservation of energy and momentum in forward photon emission of a free electron: (a) The slope of the tangent to the curve at intermediate point k^*, $\partial\varepsilon_k(k^*)/\partial k$ may be equal to the slope of the chord $\hbar\omega/q$ which is impossible in free space. (b) electron radiative transition made possible with an electromagnetic pump (Compton Scattering). (c) The wiggler wavenumber – k_w conserves the momentum in electron radiative transition of FEL.

condition in eqn [11] must be satisfied only within an uncertainty range $\pm\pi/L$. This makes it possible to obtain radiative emission in free electron radiation effects like 'Transition Radiation' and in microwave tubes like the Klystron.

(ii) *Propagate the radiation wave* in a 'slow wave' structure, where the phase velocity of the radiation wave is smaller than the speed of light, and satisfaction of eqn [14] is possible. For example, in the Cerenkov effect, charged particles pass through a medium (gas) with index of refraction $n > 1$. Instead of eqn [13], $\mathbf{q} = n(\omega)(\omega/c)\hat{\mathbf{e}}_q$, and consequently $q_z = n(\omega)(\omega/c)\cos\Theta_q$, where we assume radiative emission at an angle Θ_q relative to the electron propagation axis z. Substitution in eqn [14] results in the Cerenkov radiation condition $v_g n(\omega)\cos\Theta_q = 1$.

Another example for radiation emission in a slow-wave structure is the Traveling Wave Tube (TWT). In this device, a periodic waveguide of periodicity λ_w permits (via the Floquet theorem) propagation of slow partial waves (space harmonics) with increased wavenumber $q_z + mk_w$ ($m = 1, 2, \ldots$), and again eqn [14] can be satisfied.

(iii) *Rely on a 'two-photon' radiative transition.* This can be 'real photon' Compton scattering of an intense radiation beam (electromagnetic pump) off an electron beam, or 'virtual photon' scattering of a static potential, as is the case in bremsstrahlung radiation and in synchrotron–undulator radiation. The latter radiation scheme may be considered as a 'magnetic brehmsstrahlung' effect or as 'zero frequency pump' Compton scattering, in which the wiggler contributes only 'crystal momentum' $\hbar k_w$, to help satisfy the momentum conservation condition in eqn [11]. The Compton scattering scheme is described schematically for the one-dimensional (back scattering) case in Figure 3, and its conservation of energy and momentum diagram is depicted in Figure 2b (a 'real photon' (ω_w, k_w) free-space pump wave is assumed with $k_w = \omega_w/c$). The analogous diagram of a static wiggler ($\omega_w = 0$, $k_w = 2\pi/\lambda_w$) is shown in Figure 2c. It is worth noting that the effect of the incident scattered wave or the wiggler is not necessarily a small perturbation. It may modify substantially the electron energy-dispersion diagram of the free electron and a more complete 'Brillouin diagram' should be used in Figure 2c. In this sense, the wiggler may be viewed as the analogue of a one-dimensional crystal, and its period λ_w analogous to the crystal lattice constant. The momentum conservation during a radiation transition, with the aid of the wiggler 'crystal momentum' $\hbar k_w$ is quite analogous to the occurrence of vertical radiative transitions in direct bandgap semiconductors, and thus the FEL has, curiously enough, some analogy to microscopic semiconductor lasers.

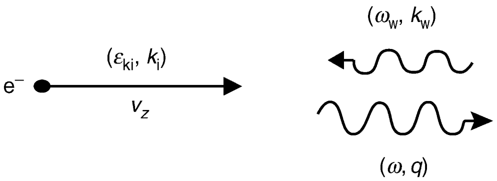

Figure 3 The scheme of backward scattering of an electromagnetic wave off an electron beam (Doppler shifted Compton scattering).

All the e-beam radiation schemes already mentioned can be turned into stimulated emission devices, and thus may be termed 'free electron lasers' in the wide sense. The theory of all of these devices is closely related, but most of the technological development was carried out on undulator radiation (Magnetic brehmsstrahlung) FELs, and the term FEL is usually reserved for this kind (though some developments of Cerenkov and Smith–Purcell FELs are still carried out).

When considering a stimulated emission device, namely enhanced generation of radiation in the presence of an external input radiation wave, one should be aware, that in addition to the emission process described by eqns [10] and [11] and made possible by one of the radiation schemes described above, there is also a stimulated absorption process. Also, this electronic transition process is governed by the conservation of energy and momentum conditions, and is described by eqns [10] and [11] with k_i and k_f exchanged.

Focusing on undulator-radiation FEL and assuming momentum conservation in the axial (z) dimension by means of the wiggler wavenumber k_w, the emission and absorption quantum transition levels and radiation frequencies are found from the solution of equations:

$$\varepsilon_{k_{zi}} - \varepsilon_{k_{zf}} = \hbar\omega_e \qquad [15a]$$

$$k_{zi} - k_{zf} = q_{ze} + k_w \qquad [15b]$$

$$\varepsilon_{k_{za}} - \varepsilon_{k_{zi}} = \hbar\omega_a \qquad [16a]$$

$$k_{za} - k_{zi} = q_{za} + k_w \qquad [16b]$$

For fixed k_w, fixed transverse momentum and given e-beam energy $\varepsilon_{k_{zi}}$ and radiation emission angle Θ_q ($q_z = (\omega/c)\cos\Theta_q$), eqns [15] and [16] have separately

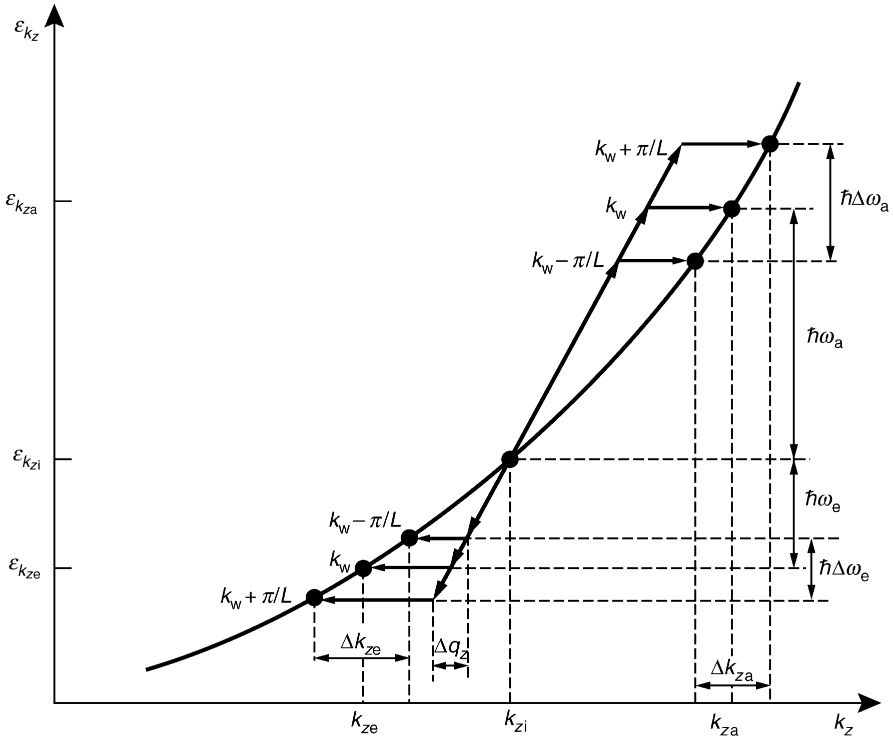

Figure 4 The figure illustrates that the origin of difference between the emission and absorption frequencies is the curvature of the energy dispersion line, and the origin of the homogeneous line broadening is momentum conservation uncertainty $\pm\pi/L$ in a finite interaction length. (Reproduced with permission from Friedman A, Gover A, Ruschin S, Kurizki G and Yariv A (1988) Spontaneous and stimulated emission from quasi-free electrons. *Reviews of Modern Physics* 60: 471–535. Copyright (1988) by the American Physical Society.)

distinct solutions, defining the electron upper and lower quantum levels for radiative emission and absorption respectively. The graphical solutions of these two set of equations are shown in Figure 4, which depicts also the 'homogeneous' frequency-line broadening $\hbar\Delta\omega_e$, $\hbar\Delta\omega_a$ of the emission and absorption lines due to the uncertainty in the momentum conservation $\pm\pi/L$ in a finite interaction length. In the quantum limit of a cold (monoenergetic) e-beam and a long interaction length L, the absorption line center ω_a is larger than the emission line center ω_e, and the linewidths $\Delta\omega_e \cong \Delta\omega_a = \Delta\omega_L$ are narrower than the emission and absorption lines spacing $\omega_a - \omega_e$, as shown in Figure 5a. The FEL then behaves as a 3-level quantum system, with electrons occupying only the central level, and the upper level is spaced apart from it more than the lower level (Figure 4).

In the classical limit $\hbar \rightarrow 0$, one can Taylor-expand ε_{kz} around k_{zi}. Using:

$$v_z = \frac{1}{\hbar}\frac{\partial \varepsilon_{kz}}{\partial k_z}, \qquad \frac{1}{\gamma_z^2 \gamma m} = \frac{1}{\hbar^2}\frac{\partial^2 \varepsilon_{kz}}{\partial k_z^2}$$

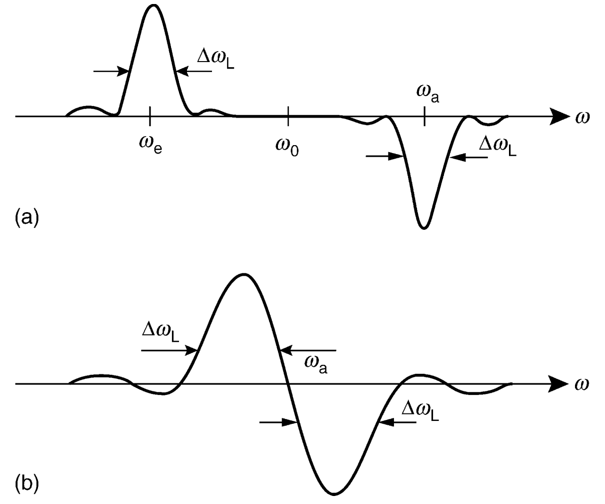

Figure 5 Net gain emission/absorption frequency lines of FEL: (a) in the quantum limit: $\omega_a - \omega_e \gg \Delta\omega_L$, (b) in the classical limit: $\omega_a - \omega_e \ll \Delta\omega_L$.

one obtains:

$$\omega_e \cong \omega_a \cong \omega_0 = v_z(q_{z0} + k_w) \qquad [17]$$

which for $q_{z0} = (\omega/c)\cos\Theta_q$ reproduces the classical synchronism condition in eqn [5]. The homogeneous broadening linewidth is found to be:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_w} \qquad [18]$$

where $N_w = L/\lambda_w$ is the number of wiggler periods.

The classical limit condition requires that the difference between the emission and absorption line centers will be smaller than their width. This is expressed in terms of the 'recoil parameter ε' (for $\Theta_q = 0$):

$$\varepsilon = \frac{\omega_a - \omega_e}{\Delta\omega_L} = \frac{1+\beta}{\beta_z^2}\frac{\hbar\omega_0}{\gamma mc^2}N_w \ll 1 \qquad [19]$$

This condition is satisfied in all practical cases of realizable FELs. When this happens, the homogeneous line broadening dominates over the quantum-recoil effect, and the emission and absorption lines are nearly degenerate (Figure 5b). The total quantum-electrodynamic photonic emission rate expression:

$$\frac{d\nu_q}{dt} = \Gamma_{sp}[(\nu_q + 1)F(\omega - \omega_e) - \nu_q F(\omega - \omega_a)] \qquad [20]$$

reduces then into:

$$\frac{d\nu_q}{dt} = \nu_q\Gamma_{sp}\varepsilon\Delta\omega_L\frac{d}{d\omega}F(\omega - \omega_e) + \Gamma_{sp}F(\omega - \omega_0) \qquad [21]$$

Here ν_q is the number of photons in radiation mode q, Γ_{sp} – the spontaneous emission rate, and $F(\omega - \omega_0)$ is the emission (absorption) lineshape function. Figure 5b depicts the transition of the net radiative emission/absorption rate into a gain curve which is proportional to the derivative of the spontaneous emission lineshape function (first term in eqn [21]). Equation [21] presents a fundamental relation between the spontaneous and stimulated emission of FELs, which was observed first by John Madey (Madey's theorem). It can be viewed as an extension of Einstein's relations to a classical radiation source.

The Classical Picture

The spontaneous emission process of FEL (synchrotron undulator radiation) is nothing but dipole radiation of the undulating electrons, which in the laboratory frame of reference is Doppler shifted to high frequency. The understanding of the stimulated emission process requires a different approach.

Consider a single electron, following a sinusoidal trajectory under the effect of a planar undulator magnetic field in eqn [2] (Figure 1):

$$v_x = v_w \cos k_w z_e(t) \qquad [22]$$

$$x = x_w \sin k_w z_e(t) \qquad [23]$$

where $v_w = ca_w/\gamma$ and $x_w = v_w/(v_z k_w)$. An electromagnetic wave $E_x(z,t) = E_0\cos(\omega t - k_z z)$ propagates collinearly with the electron. Figure 6 displays the electron and wave 'snap-shot' positions as they propagate along one wiggler period λ_w.

If the electron, moving at average axial velocity v_z, enters the interaction region $z = 0$ at $t = 0$, its axial position is $z_e(t) = v_z t$, and the electric force it experiences is $-eE_x(z_e(t), t) = -eE_0\cos(\omega - k_z v_z)t$. Clearly, this force is (at least initially at $t = 0$) opposite to the transverse velocity of the electron $v_x = v_w\cos(k_w v_z)t$ (imply in deceleration) and the power exchange rate $-ev_e \cdot E = -ev_x E_x$ corresponds to transfer of energy into the radiation field on account of the electron kinetic energy. Because the phase velocity of the radiation mode is larger than the electron velocity, $v_{ph} = \omega/k_z > v_z$, the electron phase $\varphi_e = (\omega - k_z v_z)t$ grows, and the power exchange rate $-ev_x E_x$ changes. However, if one synchronizes the electron velocity, so that while the electron traverses one wiggler period ($t = \lambda_w/v_z$), the electron phase advances by $2\pi : (\omega - k_z v_z)\cdot\lambda_w/v_z = 2\pi$, then the power exchange rate from the electron to the wave remains non-negative through the entire interaction length, because then the electron transverse velocity

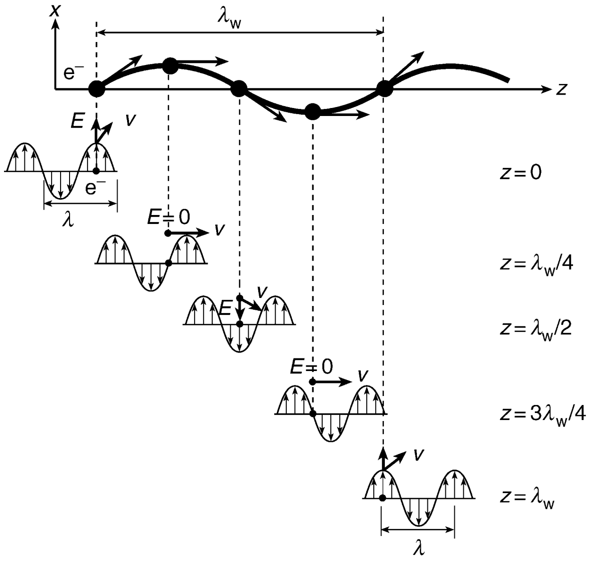

Figure 6 'Snapshots' of an electromagnetic wave period slipping relative to an undulating electron along one wiggler period λ_w. The energy transfer to the wave – $e\mathbf{v}\cdot\mathbf{E}$ remains non-negative all along.

v_x and the wave electric field E_x reverse sign, at each period exactly at the same points ($\lambda_w/4, 3\lambda_w/4$). This situation is depicted in Figure 6, which shows the slippage of the wave crests relative to the electron at five points along one wiggler period. The figure describes the synchronism condition, in which the radiation wave slips one optical period ahead of the electron, while the electron goes through one wiggle motion. In all positions along this period, $v \cdot E \geq 0$ (in a helical wiggler and a circularly polarized wave this product is constant and positive $v \cdot E > 0$ along the entire period). Substituting $\lambda_w = 2\pi/k_w$, this phase synchronism condition may be written as:

$$\frac{\omega}{v_z} = k_z + k_w \qquad [24]$$

which is the same as eqns [17] and [5].

Figure 6 shows that a single electron (or a bunch of electrons of duration smaller than an optical period) would amplify a co-propagating radiation wave, along the entire wiggler, if it satisfies the synchronism condition in eqn [24] and enters the interaction region ($z = 0$) at the right (decelerating) phase relative to the radiation field. If the electron enters at the opposite phase it accelerates (on account of the radiation field energy which is then attenuated by 'stimulated absorption'). Thus, when an electron beam is injected into a wiggler at the synchronism condition with electrons entering at random times, no net amplification or absorption of the wave is expected on the averages. Hence, some more elaboration is required, in order to understand how stimulated emission gain is possible then.

Before proceeding, it is useful to define the 'pondermotive force' wave. This force originates from the nonlinearity of the Lorenz force equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma m \mathbf{v}) = -e(\mathbf{E} \times \mathbf{v} \times \mathbf{B}) \qquad [25]$$

At zero order (in terms of the radiation fields), the only field force on the right-hand side of eqn [25] is due to the strong wiggler field (eqns [2] and [3]), which results in the transverse wiggling velocity (eqn [22] for a linear wiggler). When solving next eqn [25] to first order in terms of the radiation fields:

$$\begin{aligned} \mathbf{E}_s(r, \mathbf{t}) &= \mathrm{Re}\left[\tilde{\mathbf{E}}_s\, \mathrm{e}^{\mathrm{i}(k_z z - \omega t)}\right] \\ \mathbf{B}_s(r, \mathbf{t}) &= \mathrm{Re}\left[\tilde{\mathbf{B}}_s\, \mathrm{e}^{\mathrm{i}(k_z z - \omega t)}\right] \end{aligned} \qquad [26]$$

the cross product $\mathbf{v} \times \mathbf{B}$ between the transverse components of the velocity and the magnetic field generates a longitudinal force component:

$$\mathrm{F}_{\mathrm{pm}}(z, t) = \mathrm{Re}\left[\hat{\mathrm{e}}_z \tilde{\mathrm{F}}_{\mathrm{pm}} \mathrm{e}^{\mathrm{i}(k_z + k_w)z - \mathrm{i}\omega t}\right] \qquad [27]$$

that varies with the beat wavenumber $k_s + k_w$ at slow phase velocity ($v_{\mathrm{ph}} = \omega/(k_z + k_w) < c$). This slow force-wave is called the pomdermotive (PM) wave. Assuming the signal radiation wave in eqn [26] is polarization-matched to the wiggler (linearly polarized or circularly polarized for a linear or helical wiggler respectively), the PM force amplitude is given by:

$$|\tilde{F}_{\mathrm{pm}}| = e|\tilde{E}_{s\perp}|a_w/\gamma\beta_z \qquad [28]$$

With large enough k_w, it is always possible to slow down the phase velocity of the pondermotive wave until it is synchronized with the electron velocity:

$$v_{\mathrm{ph}} = \frac{\omega}{k_z + k_w} = v_z \qquad [29]$$

and can apply along the interaction length a decelerating axial force, that will cause properly phased electrons to transfer energy to the wave on account of their longitudinal kinetic energy.

This observation is of great importance. It reveals that even though the main components of the wiggler and radiation fields are transverse, the interaction is basically longitudinal. This puts the FEL on an equal footing with the slow-wave structure devices as the TWT and the Cerenkov–Smith–Purcell FELs, in which the longitudinal interaction takes place with the longitudinal electric field component of a slow TM radiation mode. The synchronism condition in eqn [29] between the pondermotive wave and the electron, which is identical with the phase-matching condition in eqn [24], is also similar to the synchronism condition between an electron and a slow electromagnetic wave (eqn [14]).

Using the pondermotive wave concept, we can now explain the achievement of gain in the FEL with a random electron beam. Figure 7 illustrates the interaction between the pondermotive wave and electrons, distributed at the entrance ($z = 0$) randomly within the wave period. Figure 7a shows 'snap-shots' of the electrons in one period of the pondermotive wave $\lambda_{\mathrm{pm}} = 2\pi/(k_z + k_w)$ at different points along the wiggler, when it is assumed that the electron beam is perfectly synchronous with the pondermotive wave $v_{\mathrm{ph}} = v_0$. As explained before, some electrons are slowed down, acquiring negative velocity increment Δv. However, for each such electron, there is another one, entering the wiggler at an accelerating phase of the wave, acquiring the same positive velocity increment Δv. There is then no net change in the energy of the e-beam or the wave, however, there is clearly an effect of 'velocity-bunching' (modulation), which turns along the wiggler into 'density-bunching' at the same

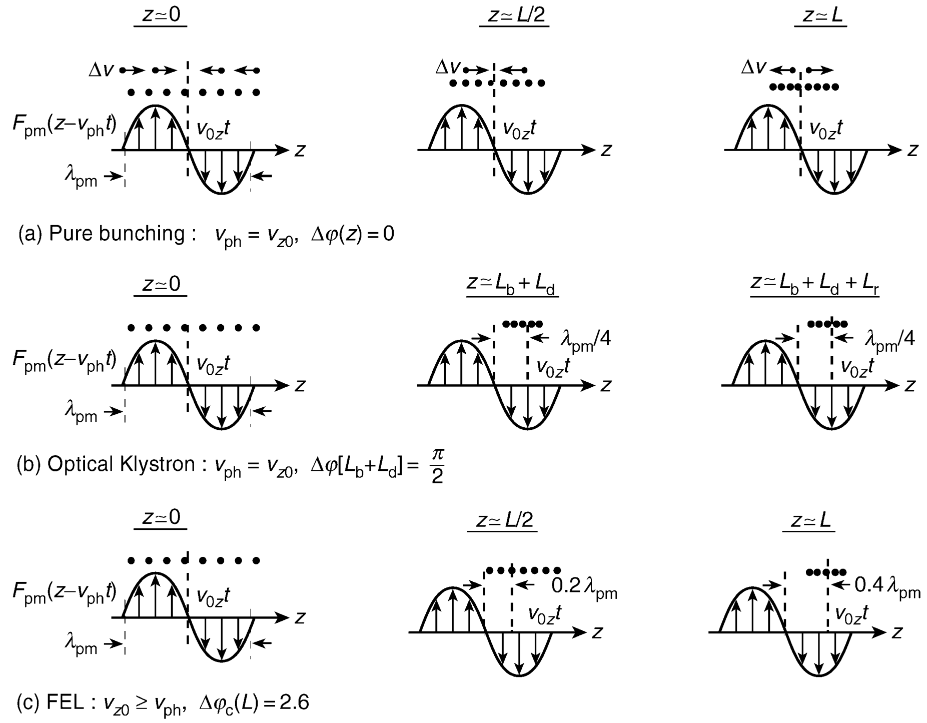

z≃0
z≃L/2
z≃L
Δv
F_pm(z−v_ph t)
v_0z t
z
λ_pm
(a) Pure bunching : v_ph = v_z0, Δφ(z) = 0
z≃0
z≃L_b + L_d
z≃L_b + L_d + L_r
λ_pm/4
(b) Optical Klystron : v_ph = v_z0, Δφ[L_b+L_d] = π/2
z≃0
z≃L/2
z≃L
0.2λ_pm
0.4λ_pm
(c) FEL : v_z0 ≥ v_ph, Δφ_c(L) = 2.6

Figure 7 'Snapshots' of a pondermotive wave period interacting L with an initially uniformly distributed electron beam taking place respectively along the interaction length $0 < z < L_b$. (a) Exact synchronism in a uniform wiggler (bunching). (b) Energy bunching, density bunching and radiation are in the energy buncher, dispersive magnet $L_b < z < L_b + L_d$ and radiating wiggler $L_b + L_d < z < L_b + L_d + L_r$ sections of an Optical–Klystron. (c) Slippage from bunching phase to radiating phase at optimum detuning off synchronism in a uniform wiggler FEL.

frequency ω and wavenumber $k_z + k_w$ as the modulating pondermotive wave. The degree of density-bunching depends on the amplitude of the wave and the interaction length L. In the nonlinear limit the counter propagating (in the beam reference frame) velocity modulated electrons may over-bunch namely cross over, and debunch again.

Bunching is the principle of classical stimulated emission in electron beam radiation devices. If the e-beam had been prebunched in the first place, we would have injected it at a decelerating phase relative to the wave and obtained net radiation gain right away (super radiant emission). This is indeed the principle behind the 'Optical-Klystron' (OK) demonstrated in Figure 7b. The structure of the OK is described ahead in Figure 19. The electron beam is velocity (energy) modulated by an input electromagnetic radiation wave in the first 'bunching-wiggler section' of length L_b. It then passes through a drift-free 'energy-dispersive magnet section' (chicane) of length L_d, in which the velocity modulation turns into density bunching. The bunched electron beam is then injected back into a second 'radiating-wiggler section', where it co-propagates with the same electromagnetic wave but with a phase advance of $\pi/2 - m2\pi$, $m = 1, 2, \ldots$ (spatial lag of $\lambda_{pm}/4 - m\lambda_{pm}$ in real space) which places the entire bunch at a decelerating phase relative to the PM-wave and so amplifies the radiation wave.

The principle of stimulated-emission gain in FEL, illustrated in Figure 7c, is quite similar. Here the wiggler is uniform along the entire length L, and the displacement of the electron bunches into a decelerating phase position relative to the PM-wave is obtained by injecting the electron beam at a velocity v_{z0}, slightly higher than the wave v_{ph} (velocity detuning). The detailed calculation shows that detuning corresponding to a phase shift of $\Delta\Psi(L) = [(\omega/v_{z0}) - (k_z + k_w)]L = -2.6$ (corresponding to spatial bunch advance of $0.4\lambda_{pm}$ along the wiggler length), provides sufficient synchronism with the PM-wave in the first half of the wiggler to obtain bunching, and sufficient deceleration-phasing of the created bunches in the second part of the wiggler to obtain maximum gain.

Principles of FEL Theory

The 3D radiation field in the interaction region can be expanded in general in terms of a complete set of free-space or waveguide modes $\{\varepsilon_q(x,y)\mathcal{H}_q(x,y)\}$:

$$E(r,t) = \mathrm{Re}\Big[\sum_q c_q(z)\varepsilon_q(x,y)\mathrm{e}^{\mathrm{i}(k_{zq}z-\omega t)}\Big] \qquad [30]$$

The mode amplitudes $C_q(z)$ may grow along the wiggler interaction length $0 < z < L$, according to the mode excitation equation:

$$\frac{\mathrm{d}}{\mathrm{d}z}C_q(z) = -\frac{1}{4\mathcal{P}}\mathrm{e}^{-\mathrm{i}k_{zq}}\int\!\!\int \mathrm{J}(x,y,z)\cdot\varepsilon_q^*(x,y)\mathrm{d}x\,\mathrm{d}y \qquad [31]$$

where $\mathcal{P} = -\frac{1}{2}\mathrm{Re}\int\int \varepsilon_q \times \mathcal{H}_q\cdot\hat{e}_z\,\mathrm{d}x\,\mathrm{d}y$ is the mode normalization power, and $\tilde{\mathrm{J}}$ is the bunching current component at frequency ω, that is phase matched to the radiation waves, and needs to be calculated consistently from the electron force equations.

The FEL Small Signal Regime

We first present the basic formulation of FEL gain in the linear (small signal) regime, namely the amplified radiation field is assumed to be proportional to the input signal radiation field, and the beam energy loss is negligible. This is done in the framework of a one-dimensional (single transverse radiation mode) model.

The electron beam charge density, current density, and velocity modulation are solved in the framework of a one-dimensional plasma equations model (kinetic or fluid equations). The longitudinal PM-force in eqn [27] modulates the electron beam velocity via the longitudinal part of the force eqn [25]. This brings about charge modulation $\rho(z,t) = \mathrm{Re}[\tilde{\rho}(k_z+k_\mathrm{w},\omega)\mathrm{e}^{\mathrm{i}(k_z+k_\mathrm{w})z-\mathrm{i}\omega t}]$ and consequently, also longitudinal space-charge field $\tilde{E}_\mathrm{sc}(k_{zq}+k_\mathrm{w},\omega)$ and longitudinal current density modulation $\tilde{J}_z(k_z+k_\mathrm{w},\omega)$, related through the Poison and continuity equations:

$$\mathrm{i}(k_z+k_\mathrm{w})\tilde{E}_\mathrm{sc}(k_z+k_\mathrm{w},\omega) = \tilde{\rho}(k_z+k_\mathrm{w},\omega)/\varepsilon \qquad [32]$$

$$(k_z+k_\mathrm{w})\tilde{J}_z(k_z+k_\mathrm{w},\omega) = \omega\tilde{\rho}(k_z+k_\mathrm{w},\omega) \qquad [33]$$

Solving the force eqn [25] for a general longitudinal force $F_z(z,t) = \mathrm{Re}[\tilde{F}_z(k_z,\omega)\mathrm{e}^{\mathrm{i}(k_zz-\omega t)}]$ results in a linear longitudinal current response relation:

$$\tilde{J}_z(k_z,\omega) = -\mathrm{i}\omega\chi_\mathrm{p}(k_z,\omega)\tilde{F}_z(k_z,\omega)/(-e) \qquad [34]$$

where $\chi_\mathrm{p}(k_z,\omega)$ is the longitudinal susceptibility of the electron beam 'plasma'.The beam charge density in the FEL may be quite high, and consequently the space charge field $\tilde{E}_\mathrm{sc}$, arising from the Poison eqn [32], may not be negligible. One should take into consideration then that the total longitudinal force $\tilde{F}_z$ is composed of both the PM-force of eqn [27] and an arising longitudinal space-charge electric force $-e\tilde{E}_\mathrm{sc}$. Thus, one should substitute in eqn [34]:

$$\tilde{F}_z(k_z+k_\mathrm{w},\omega) = -e[\tilde{E}_\mathrm{pm}(k_z+k_\mathrm{w},\omega) + \tilde{E}_\mathrm{sc}(k_z+k_\mathrm{w},\omega)] \qquad [35]$$

and solve it self-consistently with eqns [32] and [33] to obtain the 'external-force' response relation:

$$\tilde{J}_z(k_z+k_\mathrm{w},\omega) = \frac{-\mathrm{i}\omega\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)}{1+\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon}\tilde{E}_\mathrm{pm}(k_z+k_\mathrm{w},\omega) \qquad [36]$$

where we defined the PM 'field': $\tilde{E}_\mathrm{pm} = \tilde{F}_\mathrm{pm}/(-e)$.

In the framework of a single-mode interaction model, we keep in the summation of eqn [30] only one mode q (usually the fundamental mode, and in free space – a Gaussian mode). The transverse current density components in eqn [31] $\tilde{J}_\perp = \frac{1}{2}\tilde{\rho}\,\tilde{v}_w$ are found using eqns [22], [33], and [36]. Finally, substituting $C_q(z) = \tilde{C}_q\,\mathrm{e}^{\mathrm{i}\delta kz}$ (where $\delta k \equiv k_z - k_{qz}$ and $k_z \cong k_{qz}$ is the wavenumber of the radiation wave modified by the interaction with the electrons) results in the general FEL dispersion relation:

$$(k_z-k_{zq})[1+\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon] = \kappa\chi_\mathrm{p}(k_z+k_\mathrm{w},\omega)/\varepsilon_0 \qquad [37]$$

Equation [37] is a general expression, valid for a wide variety of FELs, including Cerenkov–Smith–Purcell and TWT. They differ only in the expression for κ. For the conventional (magnetic wiggler) FEL:

$$\kappa = \frac{1}{4}\frac{A_\mathrm{e}}{A_\mathrm{em}}\frac{a_\mathrm{w}^2}{\gamma^2\beta_z^2}\frac{\omega}{c}A_\mathrm{JJ}^2 \qquad [38]$$

where A_e is the cross-section area of the electron beam, and $A_\mathrm{em} \equiv \mathcal{P}_q/[\frac{1}{2}\sqrt{\varepsilon_0/\mu_0}|\varepsilon_{q\perp}(0,0)|^2]$ is the effective area of the interacting radiation-mode q, and it is assumed that the electron beam, passing on axis $(x_\mathrm{e},y_\mathrm{e}) = (0,0)$, is narrow relative to the transverse mode variation $A_\mathrm{e}/A_\mathrm{em} \ll 1$. The 'Bessel-functions coefficient' A_JJ is defined for a linear wiggler only, and is given by:

$$A_\mathrm{JJ} = J_0\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2+2)}\right] - J_1\left[\frac{a_\mathrm{w}^2}{2(a_\mathrm{w}^2+2)}\right] \qquad [39]$$

In a helical wiggler, $A_\mathrm{JJ} \equiv 1$. Usually $a_\mathrm{w} \ll 1$, and therefore $A_\mathrm{JJ} \cong 1$.

The Pierce Dispersion Equation

The longitudinal plasma response susceptibility function $\chi_\mathrm{p}(k_z,\omega)$ has been calculated, in any plasma formulation, including fluid model, kinetic model, or even quantum-mechanical theory. If the electron beam axial velocity spread is small enough (cold beam), then the fluid plasma equations can be used. The small signal longitudinal force equation derived from eqn [25], together with eqn [33] and the small signal current modulation expression:

$$\tilde{J}_z \cong \rho_0 \tilde{v}_z + v_z \tilde{\rho} \qquad [40]$$

result in:

$$\chi_{\mathrm{p}}(k_z, \omega) = -\frac{\omega_{\mathrm{p}}'^2}{(\omega - k_z v_z)^2}\varepsilon \qquad [41]$$

where $\omega_{\mathrm{p}}' = (e^2 n_0/\gamma\gamma_z^2 \varepsilon m)^{1/2}$ is the longitudinal plasma frequency, n_0 is the beam electrons density ($\rho_0 = -en_0$), v_z is the average axial velocity of the beam.

In this 'cold-beam' limit, the FEL dispersion eqn [37] reduces into the well-known 'cubic dispersion equation' derived first by John Pierce in the late 1940s for the TWT:

$$\delta k(\delta k - \theta - \theta_{\mathrm{pr}})(\delta k - \theta + \theta_{\mathrm{pr}}) = Q \qquad [42]$$

where $\delta k = k_z - k_{zq}$, θ is the detuning parameter (off the synchronism condition of eqn [24]):

$$\theta \equiv \frac{\omega}{v_z} - k_{zq} - k_{\mathrm{w}} \qquad [43]$$

$$\theta_{\mathrm{p}} = \frac{\omega_{\mathrm{p}}}{v_z} \qquad [44]$$

$$Q = \kappa \theta_{\mathrm{p}}^2 \qquad [45]$$

Here $\theta_{\mathrm{pr}} = r_{\mathrm{p}}\theta_{\mathrm{p}}$, where $r_{\mathrm{p}} < 1$ is the plasma reduction factor. It results from the reduction of the longitudinal space-charge field $\tilde{E}_{\mathrm{sp}}$ in a beam of finite radius r_{b} due to the fringe field effect ($r_{\mathrm{p}} \to 1$ when the beam is wide relative to the longitudinal modulation wavelength: $r_{\mathrm{b}} \gg \lambda_{\mathrm{pm}} = 2\pi/(k_{zq} + k_{\mathrm{w}})$).

The cubic equation has, of course, three solutions δk_i ($i = 1, 2, 3$), and the general solution for the radiation field amplitude and power is thus:

$$C_q(z) = \sum_{j=1}^{3} A_j \, e^{i\delta k_j z} \qquad [46]$$

$$P(z) = |C_q(z)|^2 \mathcal{P}_q \qquad [47]$$

The coefficients A_i can be determined from three initial conditions of the radiation and e-beam parameters $C_q(0)$, $\tilde{v}(0)$, $\tilde{i}(0)$, and can be given as a linear combination of them (here $\tilde{i} = A_{\mathrm{e}}\tilde{J}_z$ is the longitudinal modulation current):

$$A_j = A_j^{\mathrm{E}}(\omega) C_q^{\omega}(0) + A_j^{v}(\omega)\tilde{v}^{\omega}(0) + A_j^{i}(\omega)\tilde{i}^{\omega}(0) \qquad [48]$$

Alternatively stated, the exit amplitude of the electromagnetic mode can in general be expressed in terms of the initial conditions:

$$C_q(L) = H^{\mathrm{E}}(\omega) C_q(\omega, 0) + H^{\mathrm{v}}(\omega)\tilde{v}_q(\omega, 0) + H^{\mathrm{i}}(\omega)\tilde{i}(\omega, 0) \qquad [49]$$

where

$$H^{(\mathrm{E,v,i})}(\omega) = \sum_{j=1}^{3} A_j^{(\mathrm{E,v,i})}(\omega) e^{i\delta k_j L} \qquad [50]$$

In the conventional FEL, electrons are injected in randomly, and there is no velocity prebunching ($\tilde{v}(\omega, 0) = 0$) or current prebunching ($\tilde{i}(\omega, 0) = 0$) (or equivalently $\tilde{n}(\omega, 0) = 0$). Consequently, $C_q(z)$ is proportional to $C_q(0)$ and one can define and calculate the FEL small-signal single-path gain parameter:

$$G(\omega) \equiv \frac{P(L)}{P(0)} = \frac{|C_q(\omega, L)|^2}{|C_q(\omega, 0)|^2} = |H^{\mathrm{E}}(\omega)|^2 \qquad [51]$$

The FEL Gain Regimes

At different physically meaningful operating regimes, some parameters in eqn [42] can be neglected relative to others, and simple analytic expressions can be found for δk_{i}, A_{i}, and consequently $G(\omega)$. It is convenient to normalize the FEL parameters to the wiggler length: $\bar{\theta} = \theta L$, $\bar{\theta}_{\mathrm{pr}} = \theta_{\mathrm{pr}} L$, $\bar{Q} = QL^3$. An additional figure of merit parameter is the 'thermal' spread parameter:

$$\bar{\theta}_{\mathrm{th}} = \frac{v_{z\mathrm{th}}}{v_z}\frac{\omega}{v_z} L \qquad [52]$$

where $v_{z\mathrm{th}}$ is the axial velocity spread of the e-beam (in a Gaussian velocity distribution model: $f(v_z) = \exp[(v_z - v_{z0})/v_{z\mathrm{th}}]/\sqrt{\pi} v_{z\mathrm{th}})$. The axial velocity spread can result out of beam energy spread or angular spread (finite 'emittance'). It should be small enough, so that the general dispersion relation of eqn [37] reduces to eqn [42] (the practical 'cold beam' regime).

Assuming now a conventional FEL ($\tilde{v}_z(0) = 0$, $\tilde{i}(0) = 0$), the single path gain in eqn [51] can be calculated. We present next this gain expression in the different regimes. The maximum values of the gain expression in the different regimes are listed in Table 1.

Low gain

This is the regime where the differential gain in a single path satisfies $G - 1 = [P(L) - P(0)]/P(0) \ll 1$. It is not useful for FEL amplifiers but most FEL oscillators operate in this regime.

Table 1 The gain regimes maximum gain expressions

	Gain regime	*Parameters domain*	*Max. gain expression*
I	Tenuous beam low-gain	$\bar{Q},\ \bar{\theta}_{\rm pr},\ \bar{\theta}_{\rm th} < \pi$	$\dfrac{P(L)}{P(0)} = 1 + 0.27\bar{Q}$
II	Collective low-gain	$\bar{\theta}_{\rm pr} > \dfrac{\bar{Q}}{2},\ \bar{\theta}_{\rm th},\ \pi$	$\dfrac{P(L)}{P(0)} = 1 + \bar{Q}/2\bar{\theta}_{\rm pr}$
III	Collective high-gain	$\bar{Q}/2 > \bar{\theta}_{\rm pr} > \bar{Q}^{1/3},\ \bar{\theta}_{\rm th},\ \bar{Q} > \pi$	$\dfrac{P(L)}{P(0)} = \dfrac{1}{4}\exp(\sqrt{2\bar{Q}/\bar{\theta}_{\rm pr}})$
IV	Strong coupling high-gain	$\bar{Q}^{1/3} > \bar{\theta}_{\rm pr},\ \bar{\theta}_{\rm th},\ \bar{Q} > \pi$	$\dfrac{P(L)}{P(0)} = \dfrac{1}{9}\exp(\sqrt{3}\bar{Q}^{1/3})$
V	Warm beam	$\bar{\theta}_{\rm th} > \bar{\theta}_{\rm pr},\ \bar{Q}^{1/3},\ \pi$	$\dfrac{P(L)}{P(0)} = \exp(3\bar{Q}/\bar{\theta}_{\rm th}^2)$

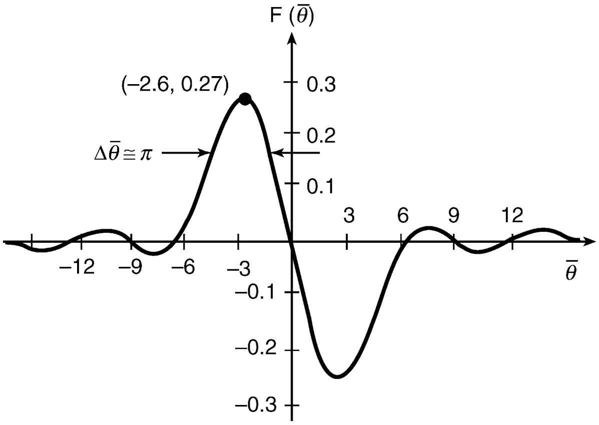

Figure 8 The low-gain cold-beam small-signal gain curve of FEL as a function of the detuning parameter $\bar{\theta}(\omega)$.

The three solutions of eqn [42] – namely the terms of eqn [46] – are reminiscent of the three eigenwaves of the uncoupled system ($\kappa = Q = 0$) : the radiation mode and the two plasma (space-charge) waves of the e-beam (the slow and fast waves, corresponding respectively to the forward and backward propagating plasma-waves in the beam rest reference-frame). In the low-gain regime, all three terms in eqn [46] are significant. Calculating them to first order in κ, results in analytical gain expressions in the collective ($\bar{\theta}_{\rm pr} \gg \pi$) and tenuous-beam ($\bar{\theta}_{\rm pr} \ll \pi$) regimes (note that $\bar{\theta}_{\rm pr}/2\pi = f'_{\rm pr}L/v_z$ is the number of plasma oscillations within the wiggler transit time L/v_z).

In most practical situations the beam current density is small enough, and its energy high enough, to limit operation to the tenuous-beam regime. The gain curve function is then:

$$G(\omega) - 1 = \bar{Q}F(\bar{\theta}(\omega)) = \bar{Q}\frac{\rm d}{{\rm d}\bar{\theta}}{\rm sinc}^2(\bar{\theta}/2) \qquad [53]$$

$$\bar{\theta}(\omega) \equiv \theta(\omega)L = 2\pi\frac{\omega - \omega_0}{\Delta\omega_L} \qquad [54]$$

where ${\rm sinc}(u) \equiv (\sin u)/u$, and in free space (no waveguide) propagation ($k_{zq} = \omega/c$), the FWHM frequency bandwidth of the ${\rm sinc}^2(\bar{\theta}/2)$ function is:

$$\frac{\Delta\omega_L}{\omega_0} = \frac{1}{N_{\rm w}} \qquad [55]$$

The small signal gain curve is shown in Figure 8. There is no gain at synchronism – $\omega = \omega_0$. Maximum gain – $G - 1 = 0.27\bar{Q}$, is attained at a frequency slightly smaller than ω_0, corresponding to $\bar{\theta} = -2.6$. The small gain curve bandwidth is $\Delta\omega_{\rm SG} \cong \Delta\omega_L/2$, namely:

$$\frac{\Delta\omega_{\rm SG}}{\omega_0} = \frac{1}{2N_{\rm w}} \qquad [56]$$

High gain

This is the regime where the FEL gain in a single path satisfies $G = P(L)/P(0) \gg 1$. It is useful, of course, when the FEL is used as an amplifier.

Since the coefficients of the cubic eqn [42] are all real, the solutions δk_i ($i = 1, 2, 3$) must be either all real, or composed of one real solution δk_3 and two complex solutions, which are complex conjugate of each other: $\delta k_1 = \delta k_2^*$. In the first case, all terms in eqn [46] are purely oscillatory, there is no exponential growth, and the FEL operates in the low gain regime. In the second case, assuming ${\rm Im}(\delta k_1) < 0$, ${\rm Im}(\delta k_2) > 0$, the first term grows exponentially, and if L is long enough it will dominate over the other decaying ($j = 2$) and oscillatory ($j = 3$) terms, and result in an exponential gain expression:

$$G(\omega) = \left(\frac{A_1}{A_1 + A_2 + A_3} \right)^2 \mathrm{e}^{2(\mathrm{Im}\ \delta k_1)L} \qquad [57]$$

If we focus on the tenuous-beam strong coupling (high–gain) regime $\theta_{\mathrm{pr}} \ll |\delta k|$, then the cubic eqn [42] gets the simple form:

$$\delta k[(\delta k)^2 - \theta^2] = \Gamma^3 \qquad [58]$$

where

$$\Gamma = Q^{1/3} = \left(\pi \frac{a_{\mathrm{w}}^2}{\gamma^3 \gamma_z^2 \beta_z^5} \frac{\omega/c}{A_{\mathrm{em}}} \frac{I_{\mathrm{b}}}{I_{\mathrm{A}}} A_{\mathrm{JJ}}^2 \right)^{1/3} \qquad [59]$$

and $I_{\mathrm{A}} = 4\pi\varepsilon_0 m_{\mathrm{e}} c^3/e \cong 17$ kA is the Alfven current. The solution of eqn [58] near synchronism ($\theta \cong 0$) is:

$$\delta k_1 = \frac{1 - \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad \delta k_2 = \frac{1 + \sqrt{3}\mathrm{i}}{2}\Gamma, \qquad \delta k_3 = -\Gamma \qquad [60]$$

resulting in:

$$H^{\mathrm{E}}(\omega) = \frac{C_q(z)}{C_q(0)} = \frac{1}{3}\left[\mathrm{e}^{\frac{\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{\frac{-\sqrt{3}+\mathrm{i}}{2}\Gamma z} + \mathrm{e}^{-\mathrm{i}\Gamma z} \right] \qquad [61]$$

and for $\Gamma L \gg 1$:

$$G \cong \frac{1}{9} \mathrm{e}^{\sqrt{3}\Gamma L} \qquad [62]$$

The FEL gain is then exponential and can be very high. The gain exponential coefficient is characterized then by its third-order root scaling with the current, $\alpha I_{\mathrm{b}}^{1/3}$. The high-gain frequency detuning curve (found by solving eqn [58] to second order in θ) is:

$$G \cong \frac{1}{9} \mathrm{e}^{\sqrt{3}\Gamma L} \exp(-\theta^2/\Gamma L 3^{3/2}) \equiv \frac{1}{9} \mathrm{e}^{\sqrt{3}\Gamma L} \exp\left[-\frac{(\omega - \omega_0)^2}{\Delta\omega_{\mathrm{HG}}^2} \right] \qquad [63]$$

where $\Delta\omega_{\mathrm{HG}}$ is the 1/e half-width of the gain curve:

$$\frac{\Delta\omega_{\mathrm{HG}}}{\omega_0} = \frac{3^{3/4}}{2\pi} \frac{\lambda_{\mathrm{w}}}{(L/\Gamma)^{1/2}} \qquad [64]$$

Super-Radiance, Spontaneous-Emission and Self Amplified Spontaneous Emission (SASE)

Intense coherent radiation power can be generated in a wiggler or any other radiation scheme without any input radiation signal ($C_q(\omega, 0) = 0$) if the electron beam velocity or current (density) are prebunched. Namely, the injected e-beam has a frequency component $\tilde{v}(\omega)$ or $\tilde{i}(\omega)$ in the frequency range where the radiation device emits. In the case of pure density bunching ($\tilde{v}(\omega) = 0$), the coherent power emitted is found from eqns [46, 47, 49]:

$$(P_q)_{\mathrm{SR}} = P_q |H^i(\omega)|^2 |\tilde{i}(\omega, 0)|^2 \qquad [65]$$

A 'prebunched-beam FEL' emits coherent radiation based on the process of Super-radiant Emission (in the sense of Dike). Because all electrons emit in phase radiation wavepackets into the radiation mode, the resultant field amplitude is proportional in this case to the beam current I_{b} and the radiation. By contrast, spontaneous emission from a random electron beam (no bunching) is the result of incoherent superposition of the wavepackets emitted by the electrons and its power is expected to be proportional to the current I_{b}.

When the current to radiation field transfer function $H^i(\omega)$ is known, eqn [65] can be used to calculate the superradiant power, and in the high-gain regime also the amplified-superradiant power. The latter is the amplification of the superradiant radiation in the downstream sections of a long wiggler. Such unsaturated gain is possible only when the beam is partly bunched $\tilde{i}(\omega)(I_{\mathrm{b}})$ (because the FEL gain process requires enhanced bunching).

The expressions for the current to field transfer function, in the superradiant gain and the high-gain amplified superradiance limits respectively, are:

$$|H^i(\omega)| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{I_{\mathrm{b}}} \mathrm{sinc}(\theta L/2) \qquad [66]$$

$$\left| H^i(\omega) \right| = \frac{(P_{\mathrm{pb}}/P_q)^{1/2}}{3\Gamma L I_{\mathrm{b}}} \mathrm{e}^{(\sqrt{3}/2)\Gamma L}\ \mathrm{e}^{-(\omega-\omega_0)^2/2(\Delta\omega)^2_{\mathrm{HG}}} \qquad [67]$$

where

$$P_{\mathrm{pb}} = \frac{I_{\mathrm{b}}^2 Z_q}{32} \left(\frac{a_{\mathrm{w}}}{\gamma\beta_z} \right)^2 \frac{L^2}{A_{\mathrm{em}}} \qquad [68]$$

Z_q is the radiation mode impedance (in free-space $Z_q = \sqrt{\mu_0/\varepsilon_0}$). From these expressions one can calculate the power and spectral power of both coherent (superradiant) and partially coherent (spontaneous emission) radiation of FEL in the negligible gain and high gain regimes. The corresponding super-radiant power is in the negligible superradiance gain limit:

$$P_{\mathrm{SR}} = P_{\mathrm{pb}} \left| \frac{\tilde{i}(\omega)}{I_{\mathrm{b}}} \right|^2 \mathrm{sinc}^2(\theta L/2) \qquad [69]$$

(proportional, as expected, to the modulation current squared) and in the high-gain amplified superradiance limit (assuming initial partial bunching $|i(\omega)/I\mathrm{b}| \ll 1$):

$$P_{\mathrm{SR}} = P_{\mathrm{pb}}\left|\frac{\tilde{i}(\omega)}{I_{\mathrm{b}}}\right|^2 \frac{1}{9(\Gamma L)^2}\, \mathrm{e}^{\sqrt{3}\Gamma L}\, \mathrm{e}^{-(\omega-\omega_0)^2/(\Delta\omega)^2_{\mathrm{HG}}} \qquad [70]$$

The discussion is now extended to incoherent (or partially coherent) spontaneous emission. Due to its particulate nature, every electron beam has random frequency components in the entire spectrum (shot noise). Consequently, incoherent radiation power is always emitted from an electron-beam passing through a wiggler, and its spectral-power can be calculated through the relation:

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{2}{\pi}\mathcal{P}_q|H^i(\omega)|^2\frac{\langle|\tilde{i}(\omega)|^2\rangle}{T} \qquad [71]$$

Here $\tilde{i}(\omega)$ is the Fourier transform of the current of randomly injected electrons $i(t) = -e\sum_{j=1}^{N_T}\delta(t - t_{oj})$, where N_T is the average number of electrons in a time period T, namely, the average (DC) current is $I_{\mathrm{b}} = -eN_T/T$. For a randomly distributed beam, the shot noise current is simply $\langle|i(\omega)|^2\rangle/T = eI_{\mathrm{b}}$, and therefore the spontaneous emission power of the FEL, which is nothing but the 'synchrotron-undulator radiation', is given by (see eqn [66]):

$$\frac{\mathrm{d}P_q}{\mathrm{d}\omega} = \frac{1}{16\pi}eI_{\mathrm{b}}Z_q\frac{L^2}{A_{\mathrm{em}}}\left(\frac{a_{\mathrm{w}}}{\gamma\beta_z}\right)^2\mathrm{sinc}^2(\theta L/2) \qquad [72]$$

If the wiggler is long enough, the spontaneous emission emitted in the first part of the wiggler can be amplified by the rest of it (SASE). In the high-gain limit (see eqn [67]), the amplified spontaneous emission power within the gain bandwidth of eqn [64] is given by:

$$P_q = \frac{2}{\pi}\mathcal{P}_q eI_{\mathrm{b}}\int_0^\infty |H^i(\omega)|^2\mathrm{d}\omega = \frac{1}{9}P_{\mathrm{sh}}\,\mathrm{e}^{\sqrt{3}\Gamma L} \qquad [73]$$

where P_{sh} is an 'effective shot-noise input power':

$$P_{\mathrm{sh}} = \frac{2}{\sqrt{\pi}}\frac{eP_{\mathrm{pb}}}{I_0(\Gamma L)^2}(\Delta\omega)_{\mathrm{HG}} \qquad [74]$$

Saturation Regime

The FEL interaction of an electron with an harmonic electromagnetic (EM) wave is essentially described by the longitudinal component of the force in eqn [25], driven by the pondermotive force of eqns [27] and [28]:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma_i m v_{zi}) = |\tilde{F}_{\mathrm{pm}}|\cos[\omega t - (k_z + k_{\mathrm{w}})z_i] \qquad [75]$$

$$\mathrm{d}z_i/\mathrm{d}t = v_{zi} \qquad [76]$$

As long as the interaction is weak enough (small signal regime), the change in the electron velocity is negligible – $v_{zi} \cong v_{z0}$, and the phase of the force-wave, experienced by the electron, is linear in time $\Psi_i(t) = [\omega - (k_z + k_{\mathrm{w}})v_{z0}](t - t_{0i}) + \omega t_{0i}$. Near synchronism condition $\theta \cong 0$ (eqn [24]), eqn [75] results in bunching of the beam, because different acceleration/deceleration forces are applied on each electron, depending on their initial phase $\Psi_i(0) = \omega t_{0i}(-\pi < \Psi_i(0) < \pi)$ within each optical period $2\pi/\omega$ (see Figure 7). Taylor expansion of v_{zi} around v_{z0} in eqns [75] and [76], and use of conservation of energy between the e-beam and the radiation field, would lead again to the small signal gain expression eqn [53] in the low gain regime.

When the interaction is strong enough (the nonlinear or saturation regime), the electron velocities change enough to invalidate the assumption of linear time dependence of Ψ_i and the nonlinear set of eqns [75] and [76] needs to be solved exactly.

It is convenient to invert the dependence on time $z_i(t) = \int_{t_{0i}}^{t} v_{zi}(t')\mathrm{d}t'$, and turn the coordinate z to the independent variable $t_i(z) = \int_0^z \mathrm{d}z'/v_{zi}(z') + t_{0i}$. This, and direct differentiation of $\gamma_i(v_{zi})$, reduces eqns [75] and [76] into the well-known pendulum equation:

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}z} = K_{\mathrm{s}}^2 \sin\Psi_i \qquad [77]$$

$$\frac{\mathrm{d}\Psi_i}{\mathrm{d}z} = \theta_i \qquad [78]$$

where

$$\Psi_i = \int_0^z (\omega/v_{zi}(z') - k_z - k_{\mathrm{w}})\,\mathrm{d}z' \qquad [79]$$

$$\theta_i = \frac{\omega}{v_{zi}} - k_z - k_{\mathrm{w}} \qquad [80]$$

are respectively the pondermotive potential phase and the detuning value of electron i at position z.

$$K_{\mathrm{s}} = \frac{k\sqrt{a_{\mathrm{w}}a_{\mathrm{s}}A_{\mathrm{JJ}}}}{\gamma_0\gamma_{z0}\beta_{z0}^2} \qquad [81]$$

is the synchrotron oscillation wavenumber, where a_{w} is given in eqn [9], $a_{\mathrm{s}} = e|\tilde{E}|\omega/mc$, and $\gamma_0 = \gamma(0)$, $\gamma_{z0} = \gamma_z(0)$, and $\beta_{z0} = \beta_z(0)$ are the initial parameters of the assumed cold beam.

The pendulum eqns [77] and [78] can be integrated once, resulting in:

$$\tfrac{1}{2}\theta_i^2(z) - K_{\mathrm{s}}^2\cos\Psi_i(z) = C_i \qquad [82]$$

and the integration constant is determined for each electron by its detuning and phase relative to the pondermotive wave at the entrance point ($z = 0$): $C_i = \frac{1}{2}\theta_i^2(0) - K_s^2\cos\Psi_i(0)$.

The $\theta(z)$, $\Psi(z)$ phase-space trajectories of eqn [82] are shown in Figure 9 for various values of C_i (corresponding to the initial conditions $\theta_i(0)$, $\Psi_i(0)$). The trajectories corresponding to $|C_i| > K_s^2$ are open; namely electrons on these trajectories, while oscillating, can slip-off out of the pondermotive–potential wave period to adjacent periods, ahead or backward, depending on the value of their detuning parameter $\bar{\theta}$. The trajectories corresponding to $|C_i| < K_s^2$ are closed, namely the electrons occupying these trajectories are 'trapped', and their phase displacement is bound to a range $|\Psi_i(z) - n\pi| < \Psi_{\text{im}} \equiv \arccos(|C_i|/K_s^2) < \pi$ within one pondermotive-wave period. The trajectory $C_i = K_s^2$ defines the 'separatrix':

$$\theta_i(z) = \pm 2K_s\cos(\Psi_i/2) \qquad [83]$$

which is sometimes referred to as the 'trap' or 'bucket'. Every electron within the separatrix stays trapped, and the ones out of it are free (untrapped). The height of the separatrix (maximum detuning swing) is $\Delta\theta_{\text{trap}} = 4K_s$. The oscillation frequency of the trapped electrons can be estimated for deeply trapped electrons ($\Psi_m \ll 2\pi$). In this case the physical pendulum eqns [77] and [78] reduce to the mathematical pendulum equation with an oscillation frequency K_s, in the z coordinate. This longitudinal oscillation, called 'synchrotron oscillation', takes place as a function of time at the 'synchrotron frequency' $\Omega_s = K_s v_z$.

Differentiation of $\theta_i(v_{zi})$ and $v_{zi}(\gamma_i)$ permits to describe the phase-space dynamics in terms of the more physical parameters $\delta v_{zi} = v_{zi} - v_{\text{ph}}$ and $\delta\gamma_i = \gamma_i - \gamma_{\text{ph}}$, where:

$$v_{\text{ph}} = \frac{\omega}{k_z + k_w} \qquad [84]$$

is the phase velocity of the pondermotive wave and $\gamma_{\text{ph}} \equiv (1 - \beta_{\text{ph}}^2)^{-1/2}$:

$$-\theta_i = \frac{\omega}{c^2\beta_{z0}^2}\delta v_{zi} = \frac{k}{\beta_{z0}^3\gamma_{z0}^2\gamma_0}\delta\gamma_i \qquad [85]$$

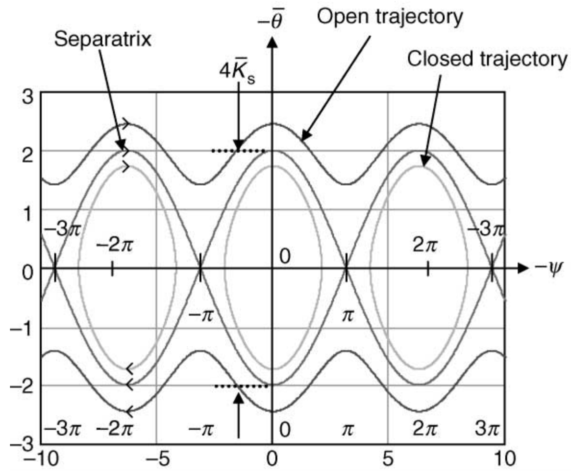

Figure 9 The (θ–Ψ) phase-space trajectories of the pendulum equation.

Figure 10 displays a typical dynamics of electron beam phase-space (γ, Ψ) evolution for the case of a cold beam of energy $\gamma(0)$ entering the interaction region at $z = 0$ with uniform phase distribution (random arrival times t_{0i}). The FEL is assumed to operate in the low-gain regime (typical situation in an FEL oscillator), and, therefore, the trap height (corresponding to $\Delta\theta_{\text{trap}} = 4k_s$):

$$\Delta\gamma_{\text{trap}} = 8\beta_z^2\gamma_z^2\gamma K_s/k \qquad [86]$$

remains constant along the interaction length. Figure 10a displays the e-beam phase-space evolution in the small signal regime. The uniform phase distribution evolves along the wiggler into a bunched distribution (compare to Figure 7c), and its average kinetic energy goes down $(\Delta E_k) = [\langle\gamma_i(L)\rangle - \gamma(0)]mc^2 < 0$, contributing this energy to the field of the interacting radiation mode, $\Delta P_q = (\Delta E_k)I_0/e$. In this case (corresponding in an FEL oscillator to the early stages of oscillation build-up), the electrons remain free (untrapped) along the entire length L.

Figure 10b displays the e-beam phase-space evolution in the large signal (saturation) regime (in the case of an oscillator – at the steady-state saturation stage). Part of the electrons are found inside the trap, immediately upon entering the interaction region ($z = 0$), and they lose energy of less than (but near) $mc^2\Delta\gamma_{\text{trap}}$ as they pass through the interaction region ($z = L$). A portion of the electrons remain outside the traps upon entrance. They follow open trajectories and lose less energy or may even become accelerated due to their interaction with the wave.

It can be appreciated from this discussion that a good design strategy in attempting to extract maximum power from the electron beam in the FEL interaction, is to set the parameters determining the synchrotron oscillation frequency K_s in eqn [81] so that only half a synchrotron oscillation period will be performed along the interaction length:

$$K_s L = \pi \qquad [87]$$

This is controlled in an amplifier by keeping the input radiation power $P_q(0)$ (and consequently a_s) small enough, so that K_s will not exceed the value set

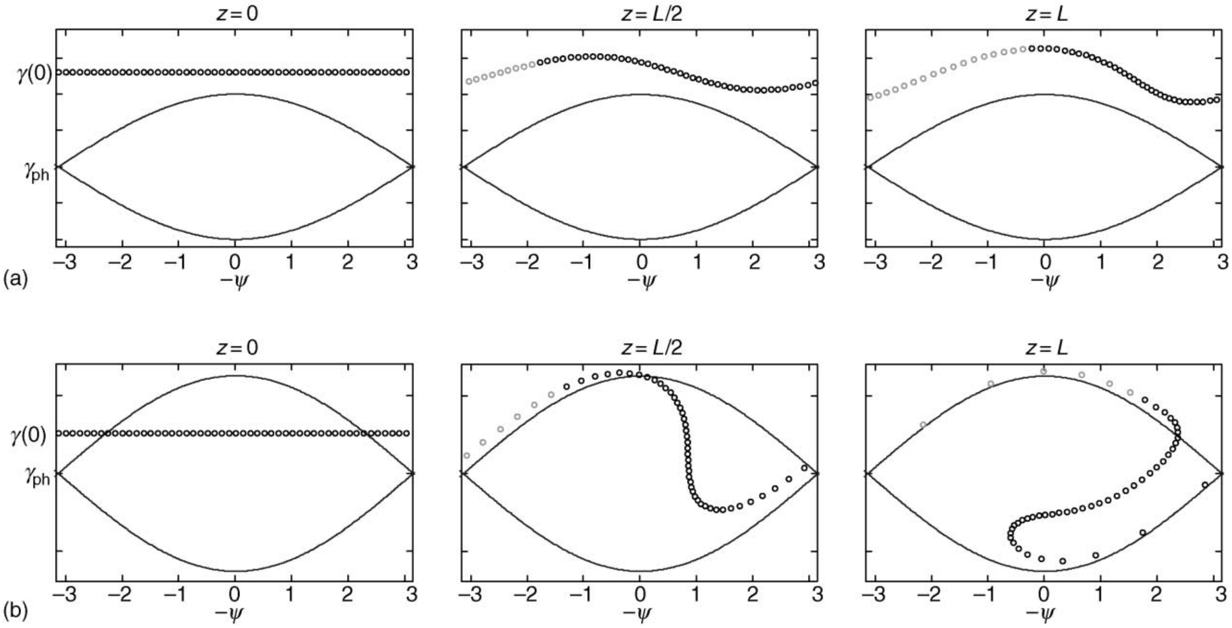

Figure 10 'Snapshots' of the ($\gamma-\Psi$) phase-space distribution of an initially uniformly distributed cold beam relative to the PM-wave trap at three points along the wiggler (a) Moderate bunching in the small-signal low gain regime. (b) Dynamics of electron beam trapping and synchrotron oscillation at steady state saturation stage of a FEL oscillator ($K_sL = \pi$).

by eqn [87]. In an oscillator, this is controlled by increasing the output mirror transmission sufficiently, so that the single path incremental small signal gain G-1 will not be much larger than the round trip loss, and the FEL will not get into deep saturation. When the FEL is over-saturated ($K_sL > \pi$), the trapped electrons begin to gain energy as they continue to rotate in their phase-space trajectories beyond the lowest energy point of the trap, and the radiative energy extraction efficiency drops down.

A practical estimate for the FEL saturation power emission and radiation extraction efficiency can be derived from the following consideration: the electron beam departs from most of its energy during the interaction with the wave, if a significant fraction of the electrons are within the trap and have positive velocity δv_{zi} relative to the wave velocity v_{ph} at $z = 0$, and if at the end of the interaction length ($z = L$), they complete half a pendulum swing and reverse their velocity relative to the wave $\delta v_{zi}(L) \cong -\delta v_{zi}(0)$. Correspondingly, in the energy phase-space diagram (Figure 10b) the electrons perform half a synchrotron oscillation swing and $\delta\gamma_i(L) = \gamma_i(L) - \gamma_{ph} = -\delta\gamma_i(0)$. In order to include in this discussion also the FEL amplifier (in the high gain regime), we note that in this case the phase velocity of the wave v_{ph} in eqn [84], and correspondingly γ_{ph}, are modified by the interaction contribution to the radiation wavenumber – $k_z = k_{z0} + \mathrm{Re}(\delta k)$, and also the electron detuning parameter (relative to the pondermotive wave) θ_i in eqn [80] differs from the beam detuning parameter θ in eqn [43]: $\theta_i = \theta - \mathrm{Re}(\delta k)$. Based on these considerations and eqn [85], the maximum energy extraction from the beam in the saturation process is:

$$\Delta\gamma = 2\delta\gamma_i(0) = 2\beta_{z0}^3\gamma_{z0}^2\gamma_0\frac{\mathrm{Re}\,\delta k - \theta}{k} \qquad [88]$$

where θ is the initial detuning parameter in eqn [43].

In an FEL oscillator, operating in general in the low-gain regime, $|\mathrm{Re}\,\delta k| \ll |\theta|$, oscillation will start usually at the resonator mode frequency, corresponding to the detuning parameter $\theta(\omega) = -2.6/L$, for which the small signal gain is maximal (see Figure 8). Then the maximum radiation extraction efficiency can be estimated directly from eqn [88]. It is, in the highly relativistic limit ($\beta_{z0} \cong 1$):

$$\eta_{\mathrm{ext}} = \frac{\Delta\gamma}{\gamma_0} \cong \frac{1}{2N_{\mathrm{w}}} \qquad [89]$$

In an FEL amplifier, in the high-gain regime $\mathrm{Re}\,\delta k = \Gamma/2 \gg |\theta|$, and consequently in the same limit:

$$\eta_{\mathrm{ext}} \cong \frac{\Gamma\lambda_{\mathrm{w}}}{4\pi} \qquad [90]$$

It may be interpreted that the effective wiggler length for saturation is $L_{\mathrm{eff}} = 2\pi/\Gamma$.

Equation [90], derived here for a coherent wave, is considered valid also for estimating the saturation efficiency also in SASE-FEL. In this context, it is also called 'the efficiency parameter' $-\rho$.

FEL Radiation Schemes and Technologies

Contrary to conventional atomic and molecular lasers, the FEL operating frequency is not determined by natural discrete quantum energy levels of the lasing matter, but by the synchronism condition of eqn [24] that can be predetermined by the choice of wiggler period, $\lambda_w = 2\pi/k_w$, the resonator dispersion characteristics $k_{zq}(\omega)$, and the beam axial velocity v_z.

Because the FEL design parameters can be chosen at will, its operating frequency can fit any requirement, and furthermore, it can be tuned over a wide range (primarily by varying v_z). This feature of FEL led to FEL development efforts in regimes where it is hard to attain high-power tunable conventional lasers or vacuum-tube radiation sources – namely in the sub-mm (far infrared or THz) regimes, and in the VUV down to soft X-ray wavelengths.

In practice, in an attempt to develop short wavelength FELs, the choice of wiggler period λ_w is limited by an inevitable transverse decay of the magnetic field away from the wiggler magnets surface (a decay range of $\approx k_w^{-1}$) dictated by the Maxwell equations. To avoid interception of electron beam current on the walls or on the wiggler surfaces, typical wiggler periods are made longer than $\lambda_w > 1$ cm. FELs (or FEMs – free electron masers) operating in the long wavelengths regime (mm and sub-mm wavelengths) must be based on waveguide resonators to avoid excessive diffraction of the radiation beam along the interaction length (the wiggler). This determines the dispersion relation $k_{zq}(\omega) = (\omega^2 - \omega_{coq}^2)^{1/2}/c$ where ω_{coq} is the waveguide cutoff frequency of the radiation mode q. The use of this dispersion relation in eqn [24] results in an equation for the FEL synchronism frequency ω_0. Usually the fundamental mode in an overmoded waveguide is used (the waveguide is overmoded because it has to be wide enough to avoid interception of electron beam current). In this case ($\omega_0 \gg \omega_{co}$) and certainly in the case of an open resonator (common in FELs operating in the optical regime) $k_{zq} = \omega/c$, and the synchronism condition in eqn [24] simplified to the well-known FEL radiation wavelength expression in eqn [6]:

$$\lambda = (1 + \beta_z)\beta_z \gamma_z^2 \lambda_w \cong 2\gamma_z^2 \lambda_w \qquad [91]$$

where γ_z, a_w are defined in eqns [7]–[9].

To attain strong interaction, it is desirable to keep the wiggler parameter a_w large (eqn [38]), however, if $a_w > 1$, this will cause reduction in the operating wavelength according to eqns [7] and [91]. For this reason, and also in order to avoid harmonic frequencies emission (in case of a linear wiggler), $a_w < 1$ in common FEL design. Consequently, considering the practical limitations on λ_w, the operating wavelength eqn [91] is determined primarily by the beam relativistic Lorentz factor γ (eqn [8]).

The conclusion is that for a short wavelength FEL, one should use an electron beam accelerated to high kinetic energy E_k. Also, tuning of the FEL operating-wavelength can be done by changing the beam energy. Small-range frequency tuning can be done also by changing the spacing between the magnet poles of a linear wiggler. This varies the magnetic field experienced by the e-beam, and effects the radiation wavelength through change of a_w (see eqns [7] and [91]).

Figure 11 displays the operating wavelengths of FEL projects all over the world versus their e-beam energy. FELs were operated or planned to operate over a wide range of frequencies, from the microwave to X-ray – eight orders of magnitude. The data points fall on the theoretical FEL radiation curve eqns [7], [8], and [91].

FEL Accelerator Technologies

The kind of accelerator used is the most important factor in determining the FEL characteristics. Evidently, the higher the acceleration energy, the shorter is the FEL radiation wavelength. However, not only the acceleration beam energy determines the shortest operating wavelength of the FEL, but also the e-beam quality. If the accelerated beam has large energy spread, energy instability, or large emittance (the product of the beamwidth with its angular spread), then it may have large axial velocity spread v_{zth}. At high frequencies, this may push the detuning spread parameter $\bar{\theta}_{th}$ (eqn [52]) to the warm beam regime (see Table 1), in which the FEL gain is diminished, and FELs are usually not operated.

Other parameters of the accelerator determine different characteristics of the FEL. High current in the electron beam enables higher gain and higher power operation. The e-beam pulse shape (or CW) characteristics, affect, of course, the emitted radiation waveform, and may also affect the FEL gain and saturation characteristics. The following are the main accelerator technologies used for FEL construction. Their wavelength operating-regimes (eqn [91]) (determined primarily by their beam acceleration energies), are displayed in Figure 12.

Modulators and pulse-line accelerators

These are usually single pulse accelerators, based on high voltage power supplies and fast discharge stored

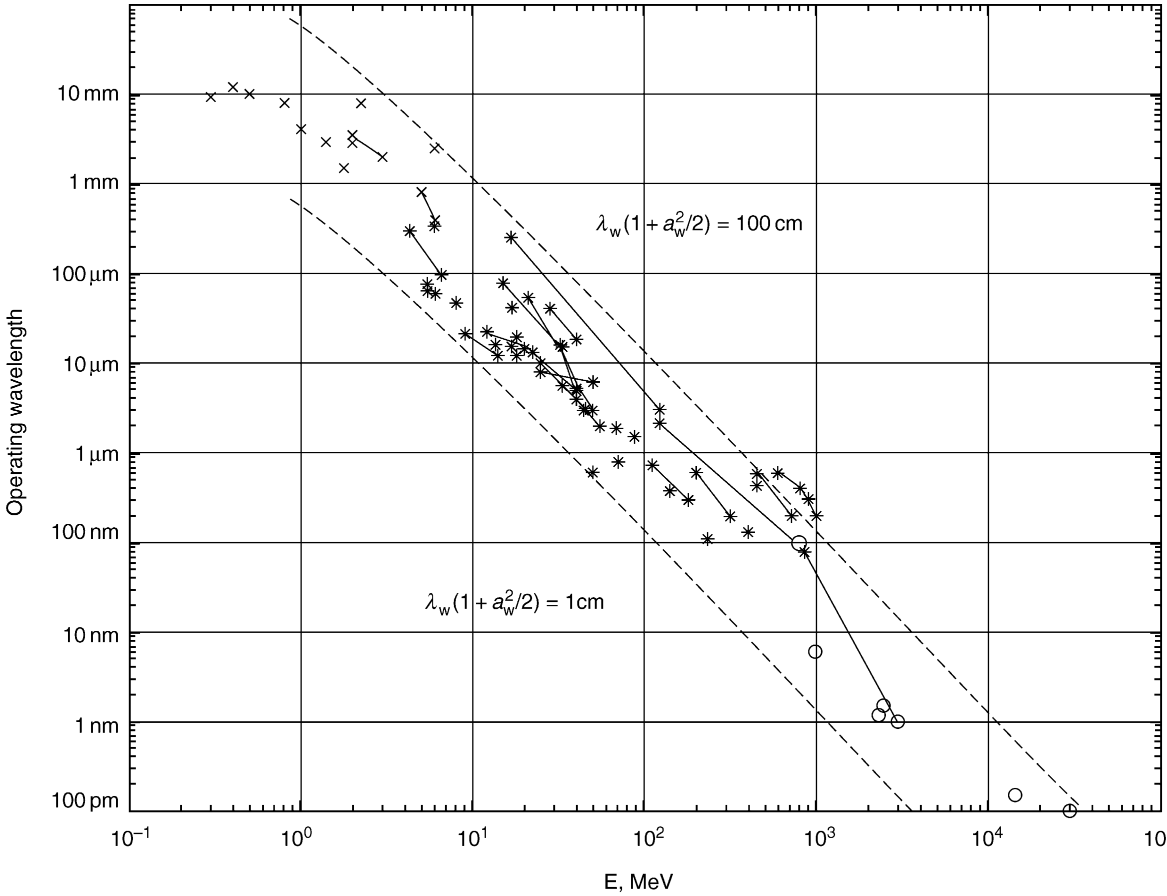

Figure 11 Operating wavelengths of FELs around the world vs. their accelerator beam energy. The data points correspond in ascending order of accelerator energy to the following experimental facilities: NRL (USA), IAP (Russia), KAERI (Korea), IAP (Russia), JINR/IAP (Russia), INP/IAP (Russia), TAU (Israel), FOM (Netherlands), KEK/JAERI (Japan/Korea), CESTA (France), ENEA (Italy), KAERI-FEL (Korea), LEENA (Japan), ENEA (Italy), FIR FEL (USA), mm Fel (USA), UCSB (USA), ILE/ILT (Japan), MIRFEL (USA), UCLA-Kurchatov (USA/Russia), FIREFLY (GB), JAERI-FEL (Japan), FELIX (Netherlands), RAFEL (USA), ISIR (Japan), UCLA-Kurchatov-LANL (USA/RU), ELSA (France), CLIO (France), SCAFEL (GB), FEL (Germany), BFEL (China), KHI-FEL (Japan), FELI4 (Japan), iFEL1 (Japan), HGHG (USA), FELI (USA), MARKIII (USA), ATF (USA), iFEL2 (Japan), VISA (USA), LEBRA (Japan), OK-4 (USA), UVFEL (USA), iFEL3 (Japan), TTF1 (Germany), NIJI-IV (Japan), APSFEL (USA), FELICITAI (Germany), FERMI (Italy), UVSOR (Japan), Super-ACO (France), TTF2 (Germany), ELETTRA (Italy), Soft X-ray (Germany), SPARX (Italy), LCLS (USA), TESLA (Germany). X, long wavelengths; $*$, short wavelengths; ○, planned short wavelengths SASE-FELs. Data based in part on H. P. Freund, V. L. Granatstein, Nucl. Inst. and Methods In Phys. Res. A249, 33 (1999), W. Colson, Proc. of the 24th Int. FEL conference, Argone, Ill. (ed. K. J. Kim, S. V. Milton, E. Gluskin). The data points fall close to the theoretical FEL radiation condition expression (91) drawn for two practical limits of wiggler parameters.

electric energy systems (e.g., Marx Generator), which produce short pulse (tens of nSec) Intense Relativistic Beam (IRB) of energy in the range of hundreds of keV to few MeV and high instantaneous current (order of kAmp), using explosive cathode (plasma field emission) electron guns. FELs (FEMs), based on such accelerators, operated mostly in the microwave and mm-wave regimes. Because of their poor beam quality and single pulse characteristic, these FELs were, in most cases, operated only as Self Amplified Spontaneous Emission (SASE) sources, producing intense radiation beams of low coherence at instantaneous power levels in the range of 1–100 MW. Because of the high e-beam current and low energy, these FEMs operated mostly in the collective high-gain regime (see Table 1).

Some of the early pioneering work on FEMs was done in the 1970s and 1980s in the US (NRL, Columbia Univ., MIT), Russia (IAP), and France (Echole Politechnique), based on this kind of accelerators.

Induction linacs

These too are single pulse (or low repetition rate) accelerators, based on induction of electromotive potential over an acceleration gap by means of an electric-transformer circuit. They can be cascaded to high energy, and produce short pulse (tens to hundreds

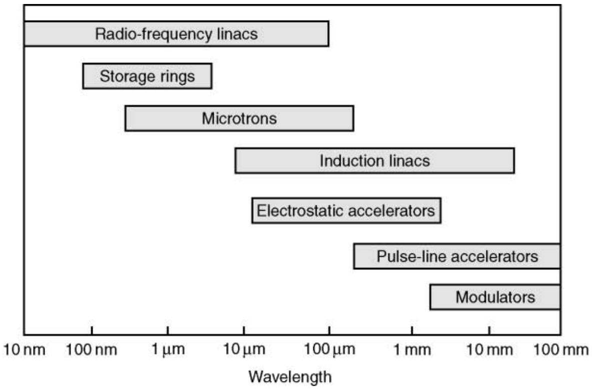

Figure 12 Approximate wavelength ranges accessible with FELs based on current accelerator and wiggler technologies.

of nSec), high current (up to 10 kA) electron beams, with relatively high energy (MeV to tens of MeV). The interest in FELs, based on this kind of accelerator technology, stemmed in the 1980s primarily from the SDI program, for the propose of development of a DEW FEL. The main development of this technology took place on a 50 MeV accelerator – ATA (for operating at 10 μm wavelength) and a 3.5 MeV accelerator – ETA (for operating at 8 mm wavelength). The latter experiment, operating in the high-gain regime, demonstrated record high power (1 GW) and energy extraction efficiency (35%).

Electrostatic accelerators

These accelerators are DC machines, in which an electron beam, generated by a thermionic electron-gun (typically 1–10 Amp) is accelerated electrostatically. The charging of the high voltage terminal can be done by mechanical charge transport (Van de Graaff) or electrodynamically (Crockford–Walton accelerator, Dynamitron). The first kind can be built at energies up to 25 MeV, and the charging current is less than mAmp. The second kind have terminal voltage less than 5 MeV, and the charging current can be hundreds of mAmps.

Because of their DC characteristics, FELs based on these kinds of accelerators can operate at arbitrary pulse shape structure and in principle – continuously (CW). However, because of the low charging current, the high electron beam current (1–10 Amp), required for FEL lasing must be transported without any interception along the entire way from the electron gun, through the acceleration tubes and the FEL wiggler, and then decelerated down to the voltage depressed beam-collector (multistage collector), closing the electric circuit back to the e-gun (current recirculation). The collector is situated at the e-gun potential, biased by moderate voltage high current power supplies, which deliver the current and power needed for circulating the e-beam and compensates for its kinetic energy loss in favor of the radiation field in the FEL cavity. This beam current recirculation is, therefore, also an ‘Energy retrieval’ scheme, and can make the overall energy transfer efficiency of the electrostatic-accelerator FEL very high.

In practice, high-beam transport efficiency in excess of 99.9% is needed for CW lasing, and has not been demonstrated yet. To avoid HV-terminal voltage drop during lasing, electrostatic-accelerator FELs are usually operated in a single pulse mode. Few FELs of this kind have been constructed. The first and main facility is the UCSB FEL shown in Figure 13. It operates in the wavelength range of 30 μm to 2.5 mm (with three switchable wigglers) in the framework of a dedicated radiation user facility. This FEL operates in the negatively charged terminal mode, in which the e-gun and collector are placed in the negatively charged HV-terminal inside the pressurized insulating gas tank, and the wigglers are situated externally at ground potential. An alternative operating mode of positively charged terminal internal cavity electrostatic-accelerator FEM was demonstrated in the Israeli Tandem–Accelerator FEM and the Dutch F.O.M. Fusion-FEM projects. This configuration enables operating with long pulse, high coherence, and very high average power. Linewidth of $\Delta\omega/\omega \cong 10^{-5}$ was demonstrated in the Israeli FEM and high power (730 kW over few microseconds) was demonstrated in the Dutch FEM, both at mm-wavelengths. The goal of the latter development project (which was not completed) was quasicontinuous operation at 1 MW average power for application in fusion plasma heating.

Radio-frequency (RF) accelerators

RF-accelerators are by far the most popular electron-beam sources for FELs. In RF accelerators, short electron beam bunches (bunch duration 1–10 pSec) are accelerated by the axial field of intense RF radiation (frequency about 1 GHz), which is applied in the acceleration cavities on the injected short e-beam bunches, entering with the accelerating-phase of the RF field. In microtrons, the electron bunches perform circular motion, and get incremental acceleration energy every time they re-enter the acceleration cavity. In RF-LINACs (linear accelerator), the electron bunches are accelerated in a sequence of RF cavities or a slow-wave structure, which keep an accelerating-phase synchronization of the traversing electron bunches with the RF field along a long linear acceleration length. The bunching of the electrons, prior to the acceleration step, is traditionally performed by bunching RF-cavities and a dispersive magnet (chicane) pulse compression system.

The FEL small signal gain, must be large enough to build-up the radiation field in the resonator from noise to saturation well within the macropulse duration.

RF-Linacs are essential facilities in synchrotron radiation centers, used to inject electron beam current into the synchrotron storage ring accelerator from time to time. Because of this reason, many FELs based on RF-LINACs were developed in synchrotron centers, and provide additional coherent radiation sources to the synchrotron radiation center users.

Figure 15 displays FELIX – a RF-LINAC FEL which is located in one of the most active FEL radiation user-centers in FOM – Holland.

Storage rings

Storage rings are circular accelerators in which a number of electron (or positron) beam bunches (typically of 50–500 pS pulse duration and hundreds of ampere peak current) are circulated continuously by means of a lattice of bending magnets and quadrupole lenses. Typical energies of storage ring accelerators are in the hundreds of MeV to GeVs range. As the electrons pass through the bending magnets, they lose a small amount of their energy due to emission of synchrotron radiation. This energy is replenished by a small RF acceleration cavity placed in one section of the ring. The electron beam bunch dimensions, energy spread, and emittance parameters are set in steady state by a balance between the electrons oscillations within the ring lattice and radiation damping due to the random synchrotron emission process. This produces high-quality (small emittance and energy spread) continuous train of electron beam bunches, that can be used to drive a FEL oscillator placed as an insertion device in one of the straight sections of the ring between two bending magnets.

Demonstrations of FEL oscillators, operating in a storage ring, were first reported by the French (LURE-Orsay) in 1987 (at visible wavelengths) and the Russians (VEPP-Novosibirsk) in 1988 (in the ultra-violet). The short wavelength operation of storage-ring FELs is facilitated by the high energy, low emittance and low energy spread parameters of the beam.

Since storage ring accelerators are at the heart of all synchrotron radiation centers, one could expect that FEL would be abundant in such facilities as insertion devices. There is, however, a problem of interference of the FEL operating as an insertion device in the normal operation of the ring itself. The energy spread increase, induced in the electron beam during the interaction in a saturated FEL oscillator, cannot be controlled by the synchrotron radiation damping process, if the FEL operating power is too high.

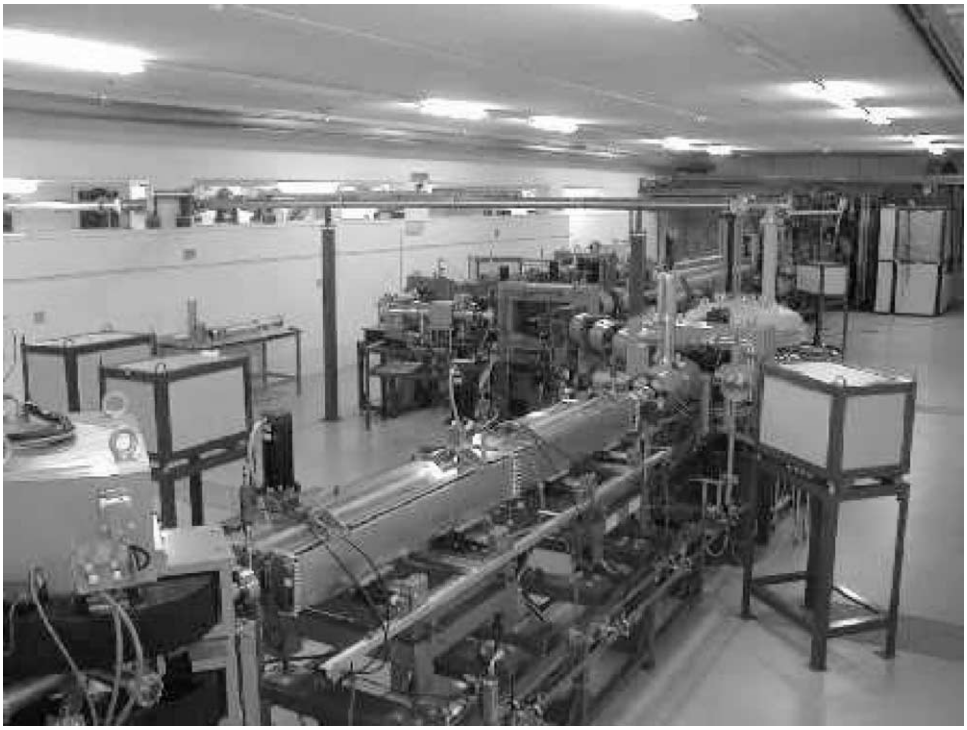

Figure 15 The FELIX RF-Linac FEL operating as a radiation users center in F.O.M. Netherlands. (Courtesy of L. van der Meer, F.O.M.)

This limits the FEL power to be kept as a fraction of the synchrotron radiation power dissipation all around the ring (the 'Renieri Limit'). The effect of the FEL on the e-beam quality, reduces the lifetime of the electrons in the storage ring, and so distrupts the normal operation of the ring in a synchrotron radiation user facility.

To avoid the interference problems, it is most desirable to operate FELs in a dedicated storage ring. This also provides the option to leave long enough straight sections in which long enough wigglers provide sufficient gain for FEL oscillation. Figure 16 displays the Duke storage ring FEL, which is used as a unique radiation user facility, providing intense coherent short wavelength radiation for applications in medicine, biology, material studies, etc.

Superconducting (SC) RF-LINACS

When the RF cavities of the accelerator are superconducting, there are very low RF power losses on the cavity walls, and it is possible to maintain continuous acceleration field in the RF accelerator with a moderate-power continuous RF source, which delivers all of its power to the electron beam kinetic energy. Combining the SC-RF-LINAC technology with an FEL oscillator, pioneered primarily by Stanford University and Thomas Jefferson Lab (TJL) in the US and JAERI Lab in Japan, gave rise to an important scheme of operating such a system in a current recirculating energy retrieval mode. This scheme revolutionized the development of FELs in the direction of high-power, high-efficiency operation, which is highly desirable, primarily for industrial applications (material processing, photochemical production, etc.).

In the recirculating SC-RF-LINAC FEL scheme the wasted beam emerging out of the wiggler after losing a fraction of only few percents (see eqn [89]) out of its kinetic energy, is not dumped into a beam-dump, as in normal cavity RF accelerators, but is re-injected, after circulation, into the SC-RF accelerator. The timing of the wasted electron bunches re-injection is such that they experience a deceleration phase along the entire length of the accelerating cavities. Usually, they are re-injected at the same cell with a fresh new electron bunch injected at an acceleration phase, and thus the accelerated fresh bunch receives its acceleration kinetic energy directly from the wasted beam bunch, that is at the same time decelerated. The decelerated wasted beam bunches are then dumped in the electron beam dump at much lower energy than without recirculation, at energies that are limited primarily just by the energy spread induced in the beam in the FEL laser-saturation process. This scheme, not only increases many folds the over-all energy transformation efficiency from wall-plug to radiation, but would solve significant heat dissipation and radioactivity activation problems in a high-power FEL design.

Figure 17 displays the TJL Infrared SC-RF-LINAC FEL oscillator, that demonstrated for the first time

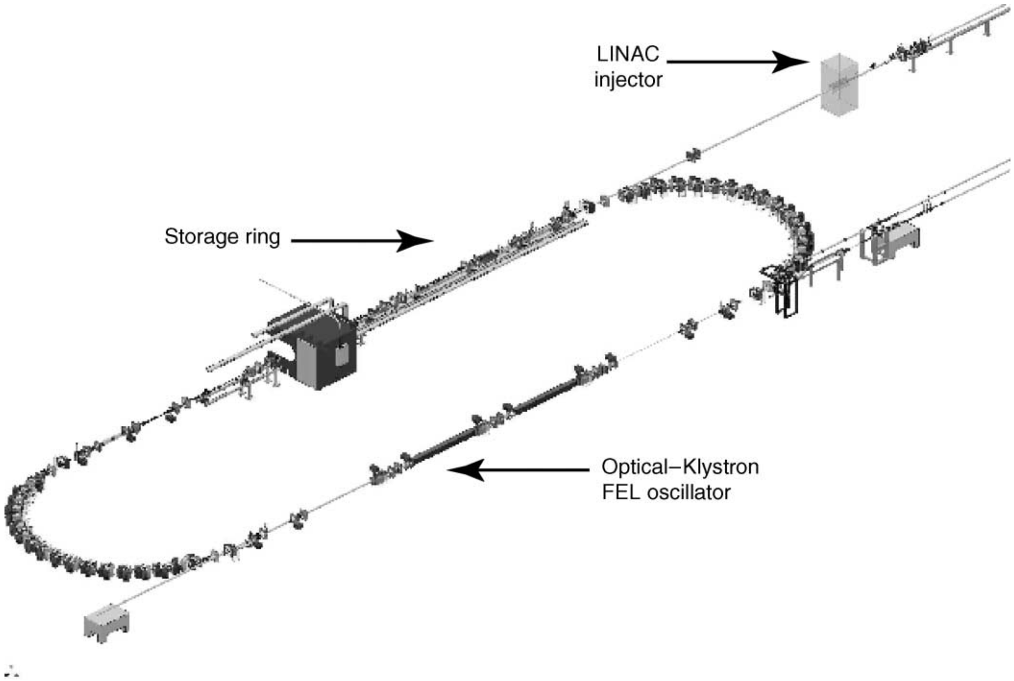

Figure 16 The Duke – University Storage Ring FEL operating as a radiation-users center in N. Carolina, USA. (Mendening: Matthew Busch, courtesy of Glenn Edwards, Duke FEL Lab.)

record high average power levels – nearly 10 kWatt at optical frequencies (1–14 μm). The facility is in upgrade development stages towards eventual operation at 100 kWatt in the IR and 1 kWatt in the UV. It operates in the framework of a laser material processing consortium and demonstrates important material processing applications, such as high-rate micromachining of hard materials (ceramics) with picoSecond laser pulses.

The e-beam current recirculation scheme of SC-RF-LINAC FEL has a significant advantage over the e-beam recirculation in a storage ring. As in electrostatic accelerators, the electrons entering the wiggler are 'fresh' cold-beam electrons from the injector, and not a wasted beam corrupted by the laser saturation process in a previous circulation through the FEL. This also makes it possible to sustain high average circulating current despite the disruptive effect of the FEL on the e-beam. This technological development has given rise to a new concept for a radiation-user facility light-source 4GLS (fourth-generation light source), which is presently in a pilot project development stage at Daresbury Lab in the UK (see Figure 18). In such a scheme, IR and UV FEL oscillators and XUV SASE-FEL can be operated together with synchrotron magnet dipole and wiggler insertion devices without disruptive interference. Such a scheme, if further developed, can give rise to new radiation-user, light-source facilities, that can provide a wider range of radiation parameters than synchrotron centers of previous generation.

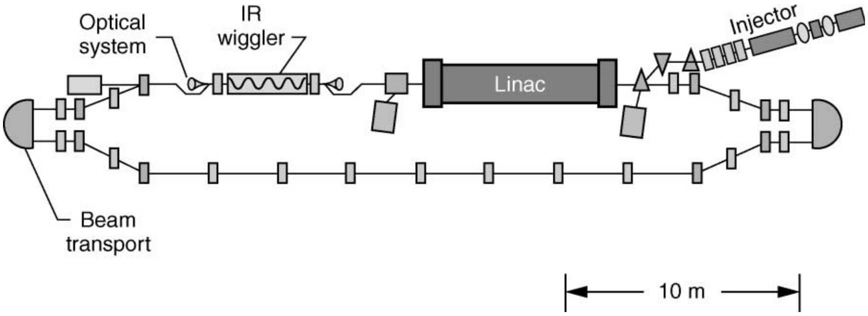
Optical system
IR wiggler
Injector
Linac
Beam transport
10 m

Figure 17 The Thomas Jefferson Lab. recirculating beam-current superconducting Linac FEL operating as a material processing FEL-user center in Virginia USA (Courtesy of S. Benson, Thomas Jefferson Laboratory).

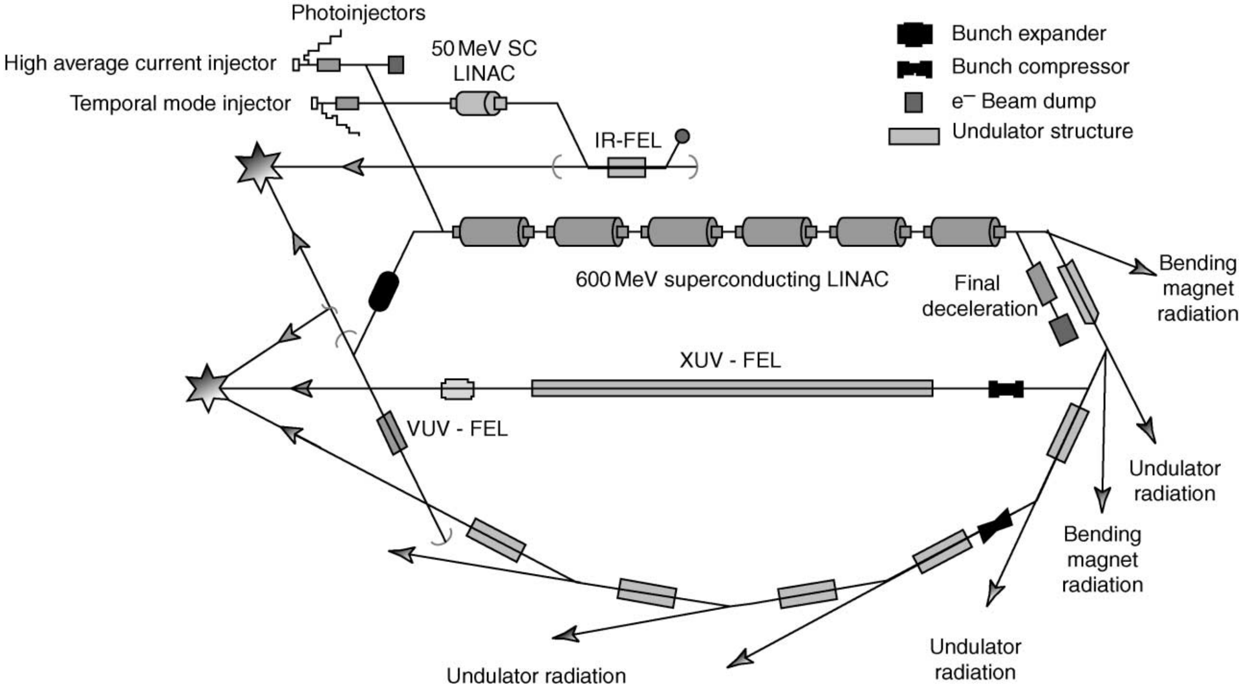
Photoinjectors
High average current injector
Temporal mode injector
50 MeV SC LINAC
IR-FEL
Bunch expander
Bunch compressor
e⁻ Beam dump
Undulator structure
600 MeV superconducting LINAC
Final deceleration
Bending magnet radiation
XUV - FEL
VUV - FEL
Undulator radiation
Bending magnet radiation
Undulator radiation
Undulator radiation

Figure 18 The Daresbury Fourth Generation Light-Source concept (4GLS). The circulating beam-current superconducting Linac includes SASE-FEL, bending magnets and wigglers as insertion devices. (Courtesy of M. Poole, Damesbury Laboratory)

Magnetic Wiggler Schemes

The optical klystron

The stimulated emission process in FEL (see Figure 7c) is based on velocity (energy) bunching of the e-beam in the first part of the wiggler, which turns into density bunching along the central part of the wiggler, and then the density-bunched electron beam performs ‘negative work’ on the radiation wave and emits radiative energy in the last part of the wiggler. In the OK, these steps are carried out in three separate parts of the wiggler: the energy bunching wiggler section, the dispersive magnet density buncher, and the radiating wiggler section (see Figure 7b).

A schematic of the OK is shown in Figure 19. The chicane magnetic structure in the dispersive section brings all electrons emerging from the bunching wiggler back onto the axis of the radiating wiggler, but provides variable delay $\Delta t_{\mathrm{d}i} = \int_{L_\mathrm{b}}^{L_\mathrm{b}+L_\mathrm{d}} (v_{zi}^{-1} - v_{\mathrm{ph}}^{-1})\,\mathrm{d}z = [\mathrm{d}(\Delta t_\mathrm{d})/\mathrm{d}\gamma]\delta\gamma_i$ relative to the pondermotive wave phase to different electrons, which acquired different energy modulation increments $\delta\gamma_i = \gamma_i - \gamma_{\mathrm{ph}}$ in the final section. The radiation condition is satisfied whenever the bunch-center phase satisfies $\Delta\varphi_\mathrm{d} = \omega\Delta t_\mathrm{d} = \pi/2 - 2m\pi$ (see Figure 7b). However, because the energy dispersion coefficient $\mathrm{d}(\Delta t_\mathrm{d})/\mathrm{d}\gamma$, is much larger in the chicane than in a wiggler of the same length, the density bunching amplitude, and consequently the OK gain, are much larger than in a uniform wiggler FEL of the same length.

The OK was invented by Vinokurov and Skrinsky in 1977 and first demonstrated in 1987 at visible wavelengths in the ACO storage ring of LURE in Orsay, France, and subsequently in 1988 at UV wavelengths, in the VEPP storage ring in Novosibirsk, Russia. The OK is an optimal FEL configuration, if used as an insertion device in a storage ring, because it can provide sufficient gain to exceed the high lasing threshold at the short operating wavelengths of a high-energy storage-ring FEL, and still conform with the rather short straight sections available for insertion devices in conventional synchrotron storage rings. It should be noted that the OK is equivalent to a long wiggler FEL of length L_{eff} of equal gain and therefore its axial velocity spread acceptance is small (this is determined from the cold beam limit $\bar{\theta}_{\mathrm{th}} \ll \pi$ with L_{eff} used in eqn [52]). This too is consistent with storage ring accelerators, which are characterized by small energy spread and emittance of the electron beam.

Radiation emission at harmonic frequencies

In a linear wiggler (eqn [2]), the axial velocity:

$$\beta_z = [\beta^2 - (a_\mathrm{w}/\gamma)^2 \cos^2 k_\mathrm{w} z]^{1/2} \qquad [92]$$

is not constant. It varies with spatial periodicity $\lambda_\mathrm{w}/2$, and in addition to its average value $\bar{\beta}_z = [\beta^2 - a_\mathrm{w}^2/2\gamma^2]^{1/2}$, contains Fourier components of spatial frequencies $2mk_\mathrm{w}$ $(m = 1, 2, \ldots)$. When $a_\mathrm{w} \gg 1$, the axial oscillation deforms the sinusoidal trajectory of the electrons in the wiggler (eqns [22] and [23]), and in a frame of reference moving at the average velocity $\bar{\beta}_z$, the electron trajectories in the wiggling $(x-z)$ plane forms an figure 8 shape, rather than a pure transverse linear motion. In the laboratory frame this leads to synchrotron undulator emission in the forward direction at all odd harmonic frequencies of ω_0, corresponding to substitution of $k_\mathrm{w} \to (2m+1)k_\mathrm{w}$ $(m = 1, 2, 3, \ldots)$ in eqn [6]:

$$\omega_{2m+1} = (2m+1)\omega_0 \cong 2\gamma_z^2 c(2m+1)k_\mathrm{w} \qquad [93]$$

All the stimulated emission gain expressions, presented earlier for the fundamental harmonic, are valid with appropriate substitution of

$$\theta_{2m+1} = \frac{\omega}{v_z} - k_z - (2m+1)k_\mathrm{w} \qquad [94]$$

instead of eqn [43], and substitution of the harmonic-weight Bessel-function coefficient of

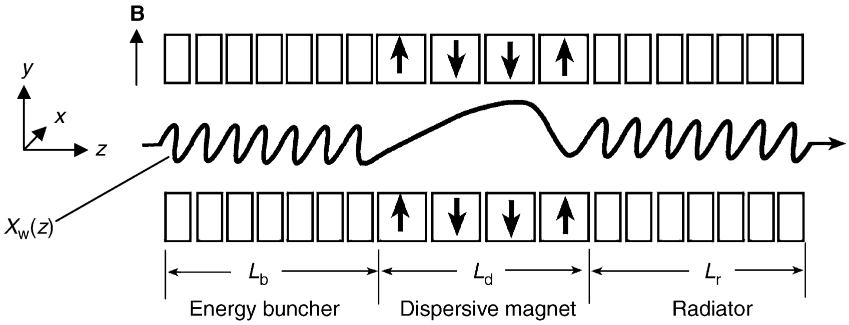

Figure 19 Schematics of the Optical–Klystron, including an energy bunching wiggler, a dispersive magnet bunching section and a radiating wiggler.

synchronism with the beam. Slowing down the PM wave can be done by the gradual increase of the wiggler wavenumber $k_w(z)$ (or decrease of its period $\lambda_w(z)$), so that eqns [29] or [91] keep being satisfied for a given frequency, even if v_z (or γ_z) goes down.

A more correct description of the nonlinear interaction dynamics of the electron beam in a saturated tapered-wiggler FEL is depicted in Figure 20: the electron trap synchronism energy $\gamma_{ph}(z)$ tapers down (by design) along the wiggler, while the trapped electrons are forced to slow down with it, releasing their excess energy by enhanced radiation. An upper limit estimate for the extraction efficiency of such a tapered wiggler FEL would be:

$$\eta_{ext} = \frac{\gamma_{ph}(0) - \gamma_{ph}(L)}{\gamma_{ph}(0)} \qquad [99]$$

and the corresponding radiative power generation would be: $\Delta P = \eta_{ext} I_b E_k / e$. In practice, the phase-space area of the tapered wiggler separatrix is reduced due to the tapering, and only a fraction of the electron beam can be trapped, which reduces correspondingly the practical enhancement in radiative extraction efficiency and power.

An alternative wiggler tapering scheme consists of tapering the wiggler field $B_w(z)$ (or wiggler parameter amplitude $a_w(z)$). If these are tapered down, the axial velocity and axial energy (eqn [7]) can still keep constant (and in synchronism with the PM wave) even if the beam energy γ goes down. Thus, in this scheme, the excess radiative energy extracted from the beam comes out of its transverse (wiggling) energy.

Efficiency and power enhancement of FEL by wiggler tapering have been demonstrated experimentally both in FEL amplifiers (first by Livermore, 1985) and oscillators (first by Los-Alamos, 1983). This elegant way to extract more power from the beam still has some limitations. It can operate efficiently only at a specified high radiation power level for which the tapering was designed. In an oscillator, a long enough untapered section must be left to permit sufficient small signal gain in the early stages of the laser oscillation build-up process.

FEL Oscillators

Most FEL devices are oscillators. As in any laser, in order to turn the FEL amplification process into an oscillation process, one provides a feedback mechanism by means of an optical resonator. In steady state saturation, $GR_{rt} = 1$, where R_{rt} is the round trip reflectivity factor of the resonator and $G = P(L)/P(0)$ is the saturated single-path gain coefficient of the FEL. To attain oscillation, the small signal (unsaturated) gain, usually given by the small gain expression in eqn [53], must satisfy the lasing threshold condition $G > 1/R_{rt}$, as in any laser.

When steady state oscillation is attained, the oscillator output power is:

$$P_{out} = \frac{T}{1 - R_{rt}} \Delta P_{ext} \qquad [100]$$

where $\Delta P_{ext} = \eta_{ext} I_0 (\gamma_0 - 1) mc^2/e$ and η_{ext} is the extraction efficiency, usually given by eqn [89] (low-gain limit).

Usually, FEL oscillators operate in the low-gain regime, in which case $1 - R_{rt} = L + T \ll 1$ (where L is the resonator internal loss factor). Consequently, then $P_{out} \cong \Delta P_{ext} T/(L + T)$, which would give a maximum value, depending on the saturation level of the oscillator. In the general case, one must solve the nonlinear force equations together with the resonator feedback relations of the oscillating radiation mode, in order to maximize the output power (eqn [100]) or efficiency by choice of optimal T for given L.

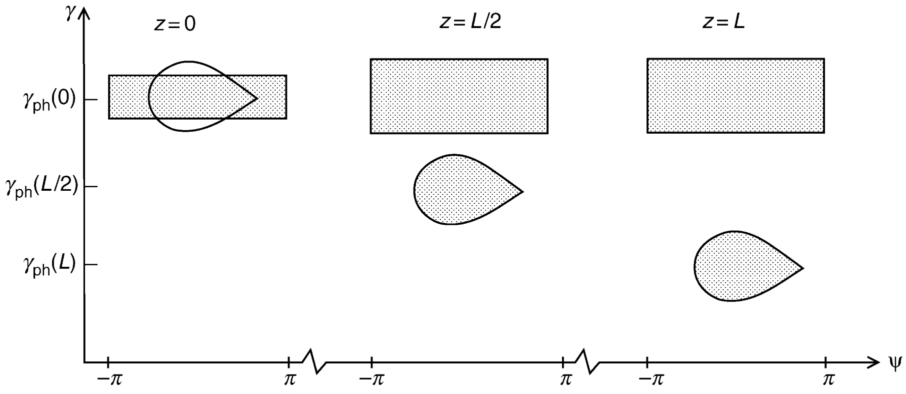

Figure 20 'Snapshots' of the trap at three locations along a tapered wiggler FEL.

In an FEL oscillator operating with periodic electron bunches (as in RF-acclerator based FEL), the solution for the FEL gain and saturation dynamics requires extension of the single frequency solution of the electron and electromagnetic field equations to the time domain. In principle, the situation is similar to that of a mode-locked laser, and the steady state laser pulse train waveform constitutes a superposition of the resonator longitudinal modes that produces a self-similar pulse shape with the highest gain (best overlap with the e-beam bunch along the interaction length). Because the e-beam velocity v_{z0} is always smaller (in an open resonator) than the group velocity of the circulating radiation wavepacket, the radiation wavepacket slips ahead of the electron bunch one optical period λ in each wiggling period (Slippage effect). This reduces the overlap between the radiation pulse and the e-beam bunch along the wiggler (see Figure 14) and consequently decreases the gain. Fine adjustment of the resonator mirrors (as shown in Figure 14) is needed to attain maximal power and optimal radiation pulse shape. The pulse-slippage gain reduction effect is negligible only if the bunch length is much longer than the slippage length $N_w\lambda$, which can be expressed as:

$$\tau_p \gg 2\pi/\Delta\omega_L \qquad [101]$$

where $\Delta\omega_L$ is the synchrotron undulator radiation frequency bandwidth (eqn [55]). This condition is usually not satisfied in RF-accelerator FELs operating in the IR or lower frequencies, and the

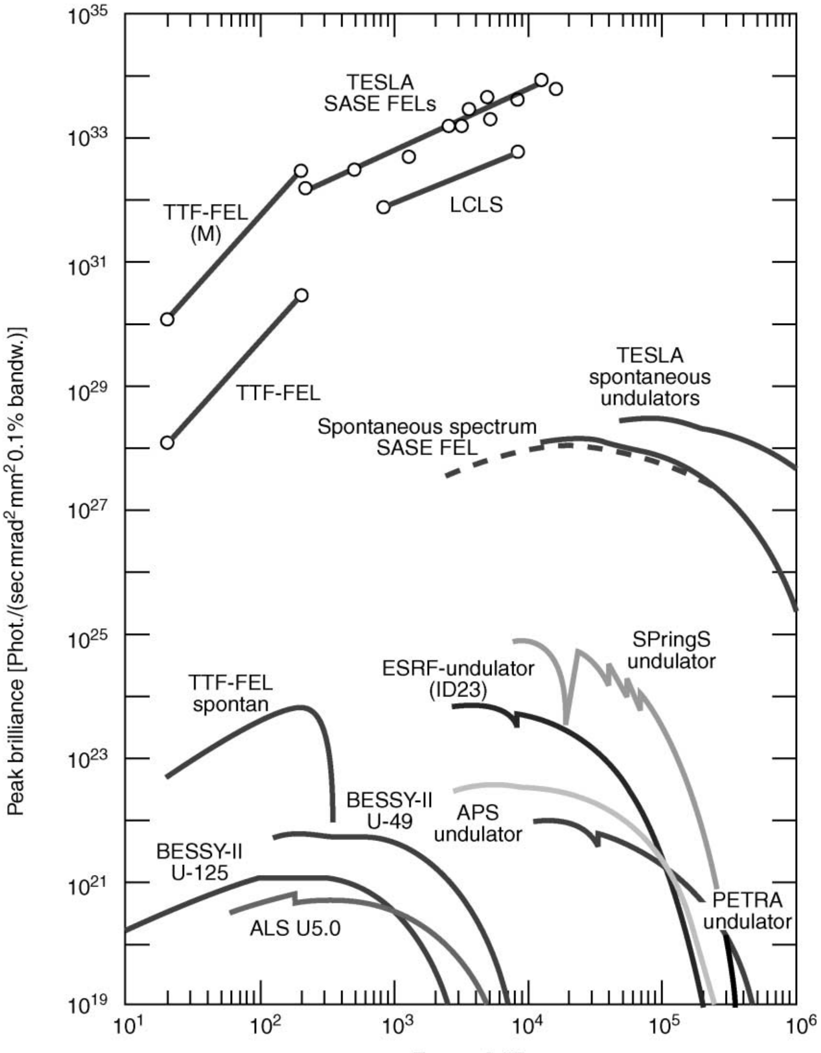

Figure 21 Anticipated peak brightness of SASE FELs (TTF-DESY, LCLS-SLAC) in comparison to the undulators in present third generation Synchrotron Radiation sources. Figure courtesy of DESY, Hamburg, Germany.

slippage effect gain reduction must be then taken into account.

An FEL operating in the cold-beam regime constitutes an 'homogeneous broadening' gain medium in the sense of conventional laser theory. Consequently, the longitudinal mode competition process that would develop in a CW FEL oscillator, leads to single-mode operation and high spectral purity (temporal coherence) of the laser radiation. The minimal (intrinsic) laser linewidth would be determined by an expression analogous to the Schawlow–Towns limit of atomic laser:

$$(\Delta f)_{\mathrm{int}} = \frac{(\Delta f_{1/2})^2}{I_b/e} \qquad [102]$$

where $\Delta f_{1/2}$ is the spectral width of the cold resonator mode. Expression [102] predicts extremely narrow linewidth. In practice, CW operation of FEL was not yet attained, but Fourier transform limited linewidths in the range of $\Delta f/f_0 \cong 10^{-6}$ were measured in long-pulse electrostatic accelerator FELs. In an FEL oscillator, based on a train of e-beam bunches (e.g., an R.F. accelerator beam), the linewidth is very wide and is equal to the entire gain bandwidth (eqn [56]) in the slippage dominated limit, and to the Fourier transform limit $\Delta\omega \cong 2\pi/\tau_p$ in the opposite negligible-slippage limit (eqn [101]). Despite this slippage, it was observed in RF-LINAC FEL that the radiation pulses emitted by the FEL oscillator are phase corrected with each other, and therefore their total temporal coherence length may be as long as the

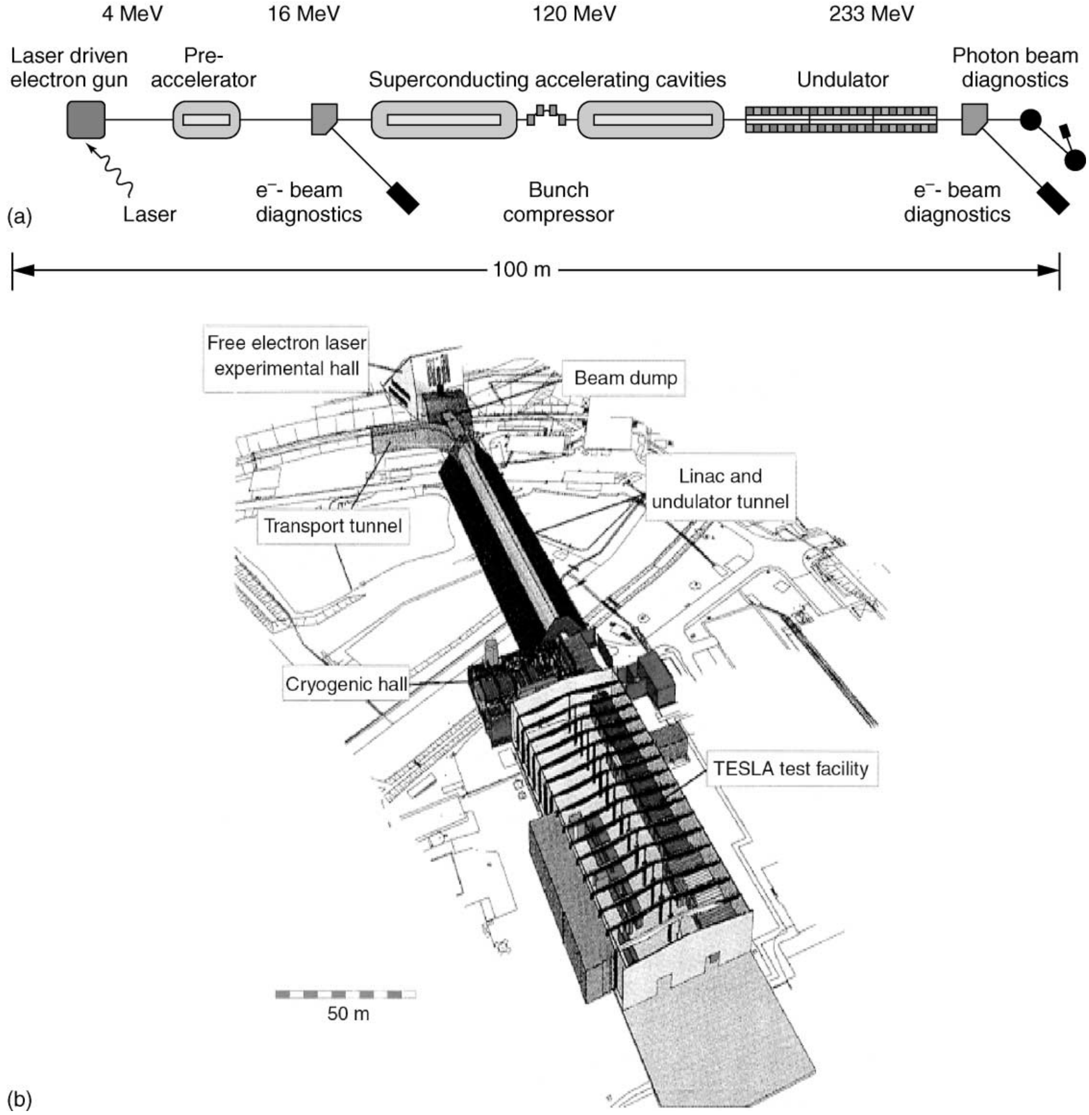
4 MeV
16 MeV
120 MeV
233 MeV
Laser driven electron gun
Pre-accelerator
Superconducting accelerating cavities
Undulator
Photon beam diagnostics
(a)
Laser
e⁻- beam diagnostics
Bunch compressor
e⁻- beam diagnostics
100 m
Free electron laser experimental hall
Beam dump
Linac and undulator tunnel
Transport tunnel
Cryogenic hall
TESLA test facility
50 m
(b)

Figure 22 Phase 1 of the SASE FE L (TTF VUV-FEL1): (a) Accelerator layout scheme; (b) General view of the TESLA test facility. Figure courtesy of DESY, Hamburg, Germany.

Metal Vapor Lasers

D W Coutts, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Metal vapor lasers form a class of laser in which the active medium is a neutral or ionized metal vapor usually excited by an electric discharge. These lasers fall into two main subclasses, namely cyclic pulsed metal vapor lasers and continuous-wave metal ion lasers. Both types will be considered in this article, including basic design and construction, power supplies, operating characteristics (including principal wavelengths), and brief reference to their applications.

Self-Terminating Resonance-Metastable Pulsed Metal Vapor Lasers

The active medium in a self-terminating pulsed metal vapor laser consists of metal atoms or ions in the vapor phase usually as a minority species in an inert buffer gas such as neon or helium. Laser action occurs between a resonance upper laser level and a metastable lower laser level (Figure 1). During a fast pulsed electric discharge (typically with a pulse duration of order 100 ns) the upper laser level is preferentially excited by electron impact excitation because it is strongly optically connected to the ground state (resonance transition) and hence has a large excitation cross-section. For a sufficiently large metal atom (or ion) density, the resonance radiation becomes optically trapped, thus greatly extending the lifetime of the upper laser level such that decay from the upper laser level is channelled through the emission of laser radiation to the metastable lower laser level. Lasing terminates when the electron temperature falls to a point such that preferential pumping to the upper laser level is no longer sustained, and the build-up of population in the metastable lower laser level destroys the population inversion. Therefore after each excitation pulse, the resulting excited species in the plasma (in particular the metastable lower laser levels which are quenched by collisions with cold electrons) must be allowed sufficient time to relax and the plasma must be allowed to partially recombine before applying the next excitation pulse. The relaxation times for self-terminating metal vapor lasers correspond to operating pulse repetition frequencies from 2 kHz to 200 kHz.

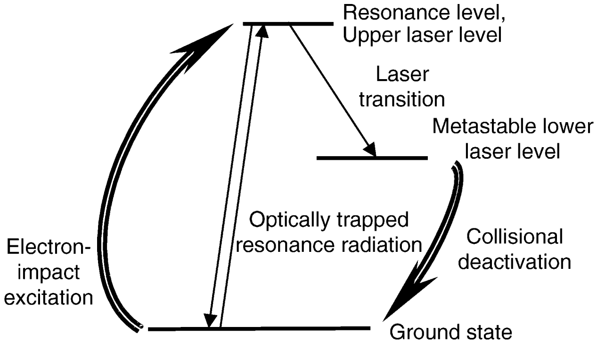

Figure 1 Resonance-metastable energy levels for self-terminating metal vapor lasers.

Many metal vapors can be made to lase in the resonance-metastable scheme and are listed together with their principal wavelengths and output powers in Table 1. The most important self-terminating pulsed metal vapor laser is the copper vapor laser and its variants which will be discussed in detail in the following sections. Of the other pulsed metal vapor lasers listed in Table 1, only the gold vapor laser (principal wavelengths 627.8 nm and 312.3 nm), and the barium vapor laser which operates in the infrared (principal wavelengths 1.5 μm and 2.55 μm) have had any commercial success. All the self-terminating pulsed metal vapor lasers have essentially the same basic design and operating characteristics as exemplified by the copper vapor laser.

Copper Vapor Lasers

Copper vapor lasers (CVLs) are by far the most widespread of all the pulsed metal vapor lasers. Figure 2 shows the energy level scheme for copper. Lasing occurs simultaneously from the $^2P_{3/2}$ level to the $^2D_{5/2}$ level (510.55 nm) and from the $^2P_{1/2}$ level to the $^2D_{3/2}$ level (578.2 nm). Commercial devices are available with combined outputs of over 100 W at 510.55 nm and 578.2 nm (typically with a green-to-yellow power ratio of 2:1).

A typical copper vapor laser tube is shown in Figure 3. High-purity copper pieces are placed at intervals along an alumina ceramic tube which typically has dimensions of 1–4 cm diameter and 1–2 m long. The alumina tube is surrounded by a solid fibrous alumina thermal insulator, and a glass or quartz vacuum envelope. Cylindrical electrodes made of copper or tantalum are located at each end of the plasma tube to provide a longitudinal discharge

Table 1 Principal self-terminating resonance-metastable metal vapor lasers

Metal	*Principal wavelengths (nm)*	*Powers*		*Total efficiencies*	*Pulse repetition frequency (kHz)*	*Technological development*
		Typical (W)	*Maximum (W)*			
Cu	510.55	2–70	2500 total	1%	4–40	Highly developed and commercially available
	578.2	1–50				
Au	627.8	1–8	20	0.13%	2–40	Commercially available
	312	0.1–0.2	1.2		1–8	
Ba	1500	2–10	12	0.5%	5–15	Have been produced commercially, but largely experimental
	2550	1	1.5			
	1130	0.5	1.0			
Pb	722.9		4.4	0.15%	10–30	Experimental
Mn	534.1 (>50%)		12 total	0.32%	~10	Experimental
	1290					

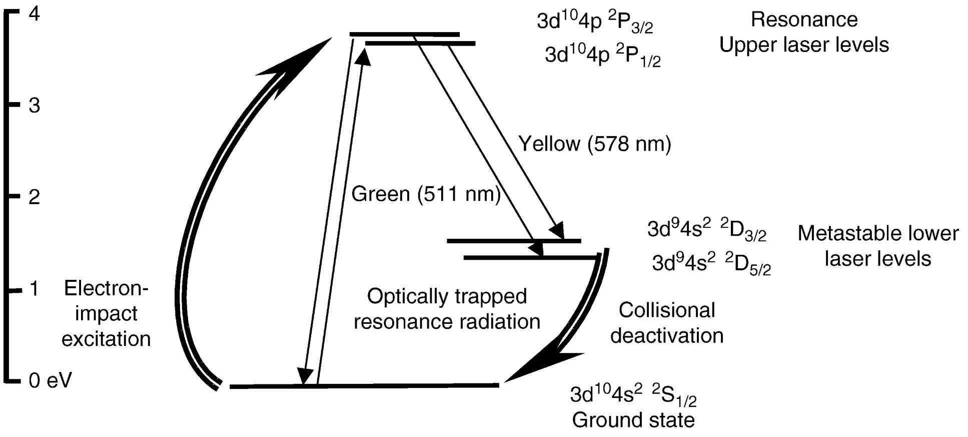
4
3
2
1
0 eV
$3d^{10}4p\ ^2P_{3/2}$
$3d^{10}4p\ ^2P_{1/2}$
Resonance
Upper laser levels
Yellow (578 nm)
Green (511 nm)
$3d^94s^2\ ^2D_{3/2}$
$3d^94s^2\ ^2D_{5/2}$
Metastable lower
laser levels
Electron-
impact
excitation
Optically trapped
resonance radiation
Collisional
deactivation
$3d^{10}4s^2\ ^2S_{1/2}$
Ground state

Figure 2 Partial energy level scheme for the copper vapor laser.

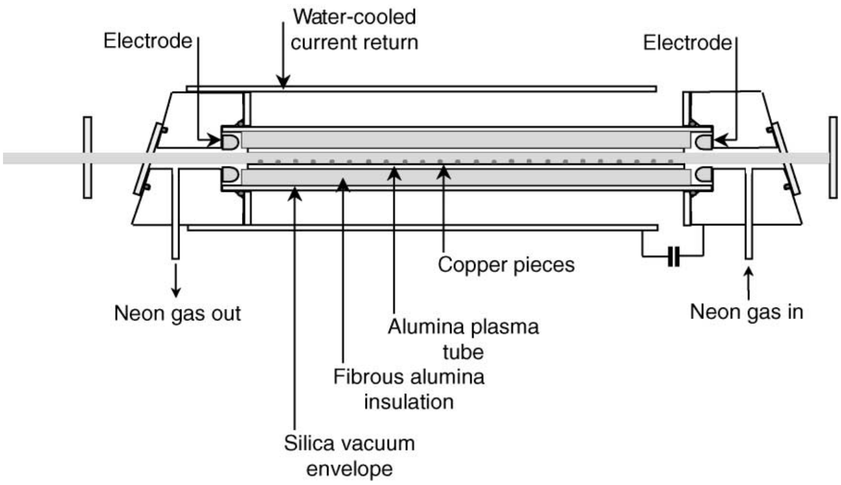
Electrode
Water-cooled
current return
Electrode
Copper pieces
Neon gas out
Neon gas in
Alumina plasma
tube
Fibrous alumina
insulation
Silica vacuum
envelope

Figure 3 Copper vapor laser tube construction.

arrangement. The cylindrical electrodes and silica laser end windows are supported by water-cooled end pieces. The laser windows are usually tilted by a few degrees to prevent back reflections into the active medium. The laser head is contained within a water cooled metal tube to provide a coaxial current return for minimum laser head inductance. Typically a slow flow (~5 mbar l min^{-1}) of neon at a pressure of 20–80 mbar is used as the buffer gas with an approximately 1% H_2 additive to improve the afterglow plasma relaxation. The buffer gas provides a medium to operate the discharge when the laser is cold and slows diffusion of copper vapor out of the ends of the hot plasma tube. Typical copper fill times are of order 200–2000 hours (for 20–200 g copper load). Sealed-off units with lifetimes of order 1000 hours have been in production in Russia for many years.

During operation waste heat from the repetitively pulsed discharge heats the alumina tube up to approximately 1500 °C at which point the vapor pressure of copper is of approximately 0.5 mbar which corresponds to the approximate density required for maximum laser output power. Typical warm-up times are therefore relatively long at around one hour to full power.

One way to circumvent the requirement for high temperatures required to produce sufficient copper density by evaporation of elemental copper (and hence also reduce warm-up times) is to use a copper salt with a low boiling point located in one or more side-arms of the laser tube (Figure 4). Usually copper halides are used as the salt, with the copper bromide laser being the most successful. For the CuBr laser, a temperature of just 600 °C is sufficient to produce the required Cu density by dissociation of CuBr vapor in the discharge. With the inclusion of 1–2% H_2 in the neon buffer gas, HBr is also formed in the CuBr laser, which has the additional benefit of improving recombination in the afterglow via dissociative attachment of free electrons: $HBr + e^- \rightarrow H + Br^-$, followed by ion neutralization: $Br^- + Cu^+ \rightarrow Br + Cu^*$. As a result of the lower operating temperature and kinetic advantages of HBr, CuBr lasers are typically twice as efficient (2–3%) as their elemental counterparts. Sealed-off CuBr systems with powers of order 10–20 W are commercially produced.

An alternative technique for reducing the operating temperature of elemental CVLs is to flow a buffer gas mixture consisting of ~5% HBr in neon at approximately 50 mbar l min^{-1} and allow this to react with solid copper metal placed within the plasma tube at about 600 °C to produce CuBr vapor *in situ*. The so-called Cu HyBrID (hydrogen bromide in discharge) laser has the same advantages as the CuBr laser (e.g., up to 3% efficiency) but at the cost of requiring flowing highly toxic HBr in the buffer gas, a requirement which has so far prevented commercialization of Cu HyBrID technology.

The kinetic advantages of the hydrogen halide in the CuBr laser discharge can also be applied to a conventional elemental CVL through the addition of small partial pressure of HCl to the buffer gas in addition to the 1–2% H_2 additive. HCl is preferred to HBr as it is less likely to dissociate (the dissociation energy of HCl at 0.043 eV is less than HBr at 0.722 eV). Such kinetic enhancement leads to a doubling in average output power, a dramatic increase in beam quality through improved gain characteristics, and shifts the optimum pulse repetition frequency for kinetically enhanced CVLs (KE-CVLs) from 4–10 kHz up to 30–40 kHz.

Preferential pumping of the upper laser levels requires an electron temperature in excess of the 2 eV range, hence high voltage (10–30 kV), high current (hundreds of A), short (75–150 ns) excitation pulses are required for efficient operation of copper vapor lasers. To generate such excitation pulses CVLs are typically operated with a power supply incorporating a high-voltage thyratron switch. In the most basic configuration, the charge-transfer circuit shown

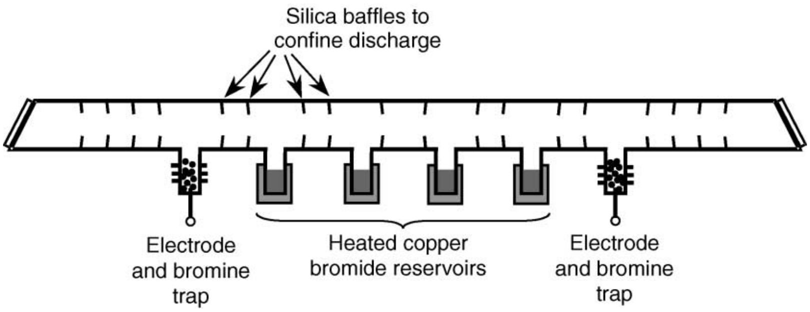
Silica baffles to confine discharge
Electrode and bromine trap
Heated copper bromide reservoirs
Electrode and bromine trap

Figure 4 Copper bromide laser tube construction.

in Figure 5a, a dc high-voltage power supply resonantly charges a storage capacitor (C_S, typically a few nF) through a charging inductor L_C, a high-voltage diode and bypass inductor L_B, up to twice the supply voltage V_S in a time of order 100 μs. When the thyratron is triggered, the storage capacitor discharges through the thyratron and the laser head on a time-scale of 100 ns. Note that during the fast discharge phase, the bypass inductor in parallel with the laser head can be considered to be an open circuit. A peaking capacitor C_P ($\sim 0.5\ C_S$) is provided to increase the rate of rise of the voltage pulse across the laser head. Given the relatively high cost of thyratrons, more advanced circuits are now often used to extend the service lifetime of the thyratron to several thousand hours. In the more advanced circuit (Figure 5b) an LC inversion scheme is used in combination with magnetic pulse compression techniques and operates as follows. Storage capacitors C_S are resonantly charged in parallel to twice the dc supply voltage V_S as before. When the thyratron is switched, the charge on C_{S1} inverts through the thyratron and transfer inductor L_T. This LC inversion drags the voltage on the top of C_{S2} down to $-4V_S$. When the voltage across the first saturable inductor L_{S1} reaches a maximum ($-4V_S$) L_{S1} saturates, and allows current to flow from the storage capacitors (now charged in series) to the transfer capacitor ($C_T = 0.5C_{S2}$) thereby transferring the charge from C_{S1} and C_{S2} to C_T in a time much less than the initial LC inversion time. At the moment when the charge on transfer capacitor C_T reaches a maximum (also $-4V_S$), L_{S2} saturates, and the transfer capacitor is discharged through the laser tube, again with a peaking capacitor to increase the voltage rise time. By using magnetic pulse compression the thyratron switched voltage can be reduced by 4 and the peak current similarly reduced (at the expense of increased current pulse duration) thereby greatly extending the thyratron lifetime. Note that in both circuits a 'magnetic assist' L_A saturable inductor is provided in series with the thyratron to delay the current pulse through the thyratron until after the thyratron has reached high conductivity thereby reducing power deposition in the thyratron.

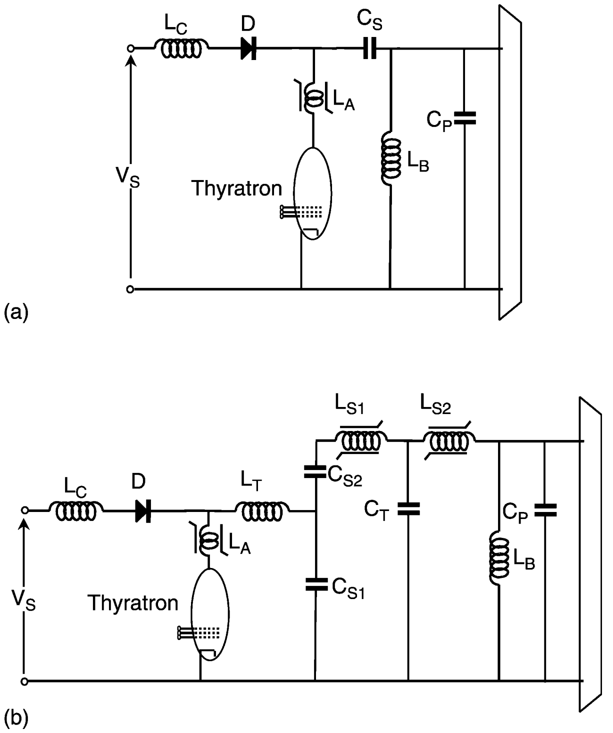

Figure 5 Copper vapor laser excitation circuits. (a) Charge transfer circuit; (b) LC inversion circuit with magnetic pulse compression.

Copper vapor lasers produce high average powers (2–100 W available commercially, with laboratory devices producing average powers of over 750 W) and have wall-plug efficiencies of approximately 1%. Copper vapor lasers also make excellent amplifiers due to their high gains, and the amplifiers can be chained together to produce average powers of several kW. Typical pulse repetition frequencies range from 4 to 20 kHz, with a maximum reported pulse repetition frequency of 250 kHz. An approximate scaling law states that for an elemental device with tube diameter D (mm) and length L (m) the average output power in watts will be of order $D \times L$. For example, a typical 25 W copper vapor laser will have a 25 mm diameter by 1 m long laser tube, and operate at 10 kHz corresponding to 2.5 mJ pulse energy and 50 kW peak power (50 ns pulse duration). Copper vapor lasers have very high single-pass gains (greater than 1000 for a 1 m long tube), large gain volumes and short gain durations (20–80 ns, sufficient for the intracavity laser light to make only a few round trips within the optical resonator). Maximum output power is therefore usually obtained in a highly 'multimode' (spatially incoherent) beam by using either a fully stable or a plane–plane resonator with a low reflectivity output coupler (usually Fresnel reflection from an uncoated optic is sufficient). To obtain higher beam quality a high magnification unstable resonator is required (Figure 6). Fortunately, copper vapor lasers have sufficient gain to operate efficiently with unstable resonators with magnifications ($M = R_1/R_2$) up to 100 and beyond. Resonators with such high magnifications impose very tight geometric constraints on propagation of radiation on repeated round-trips within the resonator, such that after two round-trips the divergence is typically diffraction-limited. Approximately half the stable-resonator

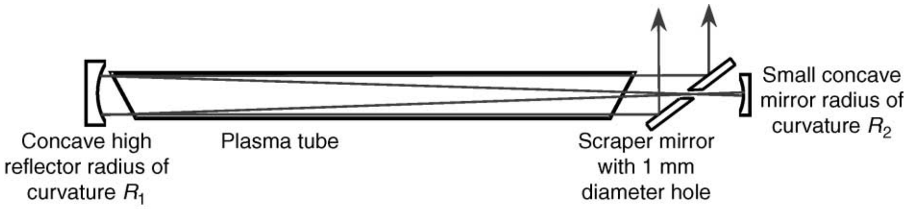
Concave high reflector radius of curvature R_1
Plasma tube
Scraper mirror with 1 mm diameter hole
Small concave mirror radius of curvature R_2

Figure 6 Unstable resonator configuration often used to obtain high beam quality from copper vapor lasers.

output power can therefore be obtained with near diffraction-limited beam quality by using an unstable resonator. Often, a small-scale oscillator is used in conjunction with a single power amplifier to produce high output power with diffraction-limited beam quality. Hyperfine splitting combined with Doppler broadening lead to an inhomogeneous linewidth for the laser transitions of order 8–10 GHz corresponding to a coherence length of order 3 cm.

The high beam quality and moderate peak power of CVLs allows efficient nonlinear frequency conversion to the UV by second harmonic generation (510.55 nm → 255.3 nm, 578.2 nm → 289.1 nm) and sum frequency generation (510.55 nm + 578.2 nm → 271.3 nm) using β-barium borate β-BaB_2O_4 (BBO) as the nonlinear medium. Typically average powers in excess of 1 W can be obtained at any of the three wavelengths from a nominally 20 W CVL. Powers up to 15 W have been obtained at 255 nm from high-power CVL master-oscillator power-amplifier systems using cesium lithium borate; $CsLi\ B_6O_{10}$ (CLBO) as the nonlinear crystal.

Key applications of CVLs include pumping of dye lasers (principally for laser isotope separation) and pumping Ti:sapphire lasers. Medically, the CVL yellow output is particularly useful for treatment of skin lesions such as port wine stain birth marks. CVLs are also excellent sources of short-pulse stroboscopic illumination for high-speed imaging of fast objects and fluid flows. The high beam quality, visible wavelength and high pulse repetition rate make CVLs very suited to precision laser micromachining of metals, ceramics and other hard materials. More recently, the second harmonic at 255 nm has proved to be an excellent source for writing Bragg gratings in optical fibers.

Afterglow Recombination Metal Vapor Lasers

Recombination of an ionized plasma in the afterglow of a discharge pulse provides a mechanism for achieving a population inversion and hence laser output. The two main afterglow recombination metal vapor lasers are the strontium ion vapor laser (430.5 nm and 416.2 nm) and calcium ion vapor laser (373.7 nm and 370.6 nm) whose output in the violet and UV spectral regions extends the spectral coverage of pulsed metal vapor lasers to shorter wavelengths.

A population inversion is produced by recombination pumping where doubly ionized Sr (or Ca) recombines to form singly ionized Sr (or Ca) in an excited state: $Sr^{++} + e^- + e^- \rightarrow Sr^{+*} + e^-$. Note that recombination rates for a doubly ionized species are much faster than for singly ionized species hence recombination lasers are usually metal ion lasers. Recombination (pumping) rates are also greatest in a cool dense plasma, hence helium is usually used as the buffer gas as it is a light atom hence promotes rapid collisional cooling of the electrons. Helium also has a much higher ionization potential than the alkaline-earth metals (including Sr and Ca) which ensures preferential ionization of the metal species (up to 90% may be doubly ionized).

Recombination lasers can operate where the energy level structure of an ion species can be considered to consist of two (or more) groups of closely spaced levels. In the afterglow of a pulsed discharge electron collisional mixing within each group of levels will maintain each group in thermodynamic equilibrium yielding a Boltzmann distribution of population within each group. If the difference in energy between the two groups of levels (5 eV >> kT) is sufficiently large then thermal equilibrium between the groups cannot be maintained via collisional processes. Given that recombination yields excited singly ionized species (usually with a flux from higher to lower ion levels), it is possible to achieve a population inversion between the lowest level of the upper group and the higher levels within the lower group. This is the mechanism for inversion in the Sr^+ (and analogous Ca^+) ion laser as indicated in the partial energy level scheme for Sr^+ shown in Figure 7.

Strontium and calcium ion recombination lasers have similar construction to copper vapor lasers described above. The lower operating temperatures (500–800 °C) mean that minimal or no thermal insulation is required for self-heated devices.

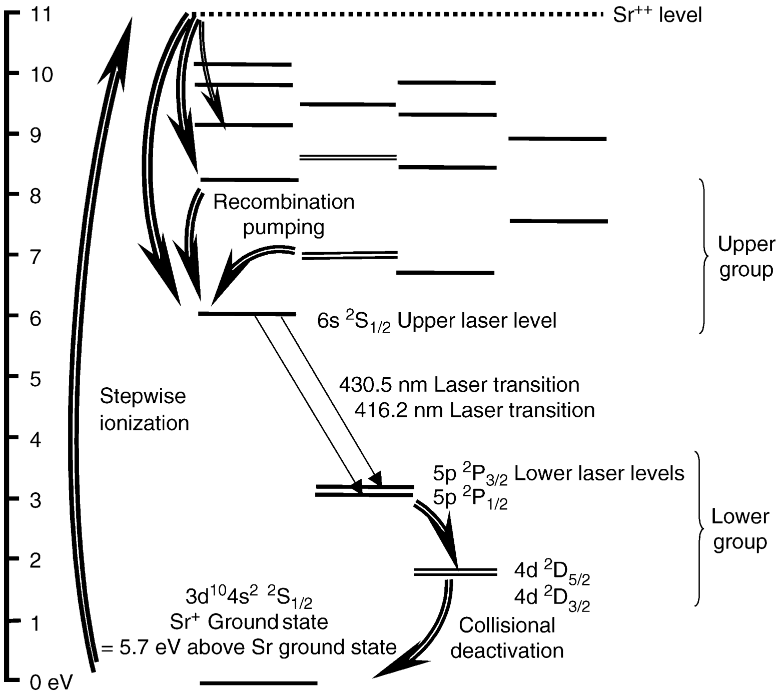
11
10
9
8
7
6
5
4
3
2
1
0 eV
Sr^{++} level
Recombination pumping
Upper group
$6s\ ^2S_{1/2}$ Upper laser level
Stepwise ionization
430.5 nm Laser transition
416.2 nm Laser transition
$5p\ ^2P_{3/2}$ Lower laser levels
$5p\ ^2P_{1/2}$
Lower group
$4d\ ^2D_{5/2}$
$4d\ ^2D_{3/2}$
$3d^{10}4s^2\ ^2S_{1/2}$
Sr^+ Ground state
= 5.7 eV above Sr ground state
Collisional deactivation

Figure 7 Partial energy level scheme for the strontium ion laser.

For good tube lifetime BeO plasma tubes are required due to the reactivity of the metal vapors. Usually helium at a pressure of up to one atmosphere is used as the buffer gas. Typical output powers at 5 kHz pulse repetition frequency are of order 1 W (~0.1% wall plug efficiency) at 430.5 nm from the He Sr^+ laser and 0.7 W at 373.7 nm from the He Ca^+ laser. For the high specific input power densities (10–15 W cm^{-3}) required for efficient lasing, overheating of the laser gas limits aperture scaling beyond 10–15 mm diameter. Slab laser geometries have been used successfully to aperture-scale strontium ion lasers. Scaling of the laser tube length beyond 0.5 m is not practical as achieving high enough excitation voltages for efficient lasing becomes problematic. Gain in strontium and calcium ion lasers is lower than in resonance-metastable metal vapor lasers such as the CVL, hence optimum output coupler reflectivity is approximately 70%. The pulse duration is also longer at around 200–500 ns resulting in moderate beam quality with plane–plane resonators.

For both strontium and calcium ion lasers the two principal transitions share upper laser levels and hence exhibit gain competition such that without wavelength-selective cavities usually only the longer wavelength of the pair is produced. With wavelength-selective cavities up to 60% (Sr^+) and 30% (Ca^+) of the normal power can be obtained at the shorter wavelength.

Many potential applications exist for strontium and calcium ion lasers, given their ultraviolet (UV)/violet wavelengths. Of particular importance is fluorescence spectroscopy in biology and forensics, treatment of neonatal jaundice, stereolithography, micromachining and exposing photoresists for integrated circuit manufacture. Despite these many potential applications, technical difficulties in power scaling mean that both strontium ion and calcium ion lasers have only limited commercial availability.

Continuous-Wave Metal Ion Lasers

In a discharge excited noble gas, there can be a large concentration of noble gas atoms in excited metastable states and noble gas ions in their ground states. These species can transfer their energy to a minority metal species (M) via two key processes: either charge transfer (Duffendack reactions) with the noble gas ions (N^+):

$$M + N^+ \rightarrow M^{+*} + N + \Delta E$$

(where M^{+*} is an excited metal ion state)

or Penning ionization with a noble gas atom in an excited metastable state (N^*):

$$M + N^* \rightarrow M^{+*} + N + e^-$$

In a continuous discharge in a mixture of a noble gas and a metal vapor, steady generation of excited metal ions via energy transfer processes can lead to a steady state population inversion on one or more pairs of levels in the metal ion and hence produce cw lasing.

Several hundred metal ion laser transitions have been observed to lase in a host of different metals. The most important such laser is the helium cadmium laser, which has by far the largest market volume by number of unit sales of all the metal vapor lasers. In a helium cadmium laser, Cd vapor is present at a concentration of about 1–2% in a helium buffer gas which is excited by a dc discharge. Excitation to the upper laser levels of Cd (Figure 8) is primarily via Penning ionization collisions with He^+ 2^3S_1 metastable ions produced in the discharge. Excitation via electron-impact excitation from the ion ground state may also play an important role in establishing a population inversion in Cd^+ lasers. Population inversion can be sustained continuously because the $^2P_{3/2}$ and $^2P_{1/2}$ lower laser levels decay via strong resonance transitions to the $^2S_{1/2}$ Cd^+ ground state, unlike in the self-terminating metal vapor lasers. Two principal wavelengths can be produced, namely 441.6 nm (blue) and 325.0 nm (UV) with cw powers up to 200 mW and 50 mW available respectively from commercial devices.

Typical laser tube construction (Figure 9) consists of a 1–3 mm diameter discharge channel typically 0.5 m long. A pin anode is used at one end of the laser tube together with a large-area cold cylindrical cathode located in a side-arm at the other end of the tube. Cadmium is transported to the main discharge tube from a heated Cd reservoir in a side-arm at 250–300 °C. With this source of atoms at the anode end of the laser a cataphoresis process (in which the positively charged Cd ions are propelled towards the cathode end of the tube by the longitudinal electric field) transports Cd into the discharge channel. Thermal insulation of the discharge tube ensures that it is kept hotter than the Cd reservoir to prevent condensation of Cd from blocking the tube bore. A large-diameter Cd condensation region is provided at the cathode end of the discharge channel. A large-volume side arm is also provided to act as a gas ballast for maintaining correct He pressure. As He is lost through sputtering and diffusion through the Pyrex glass tube walls, it is replenished from a high-pressure He reservoir by heating a permeable glass wall separating the reservoir from the ballast chamber which allows He to diffuse into the ballast chamber. Typical commercial sealed-off Cd lasers have operating lifetimes of several thousand hours. Overall laser construction is not that much more

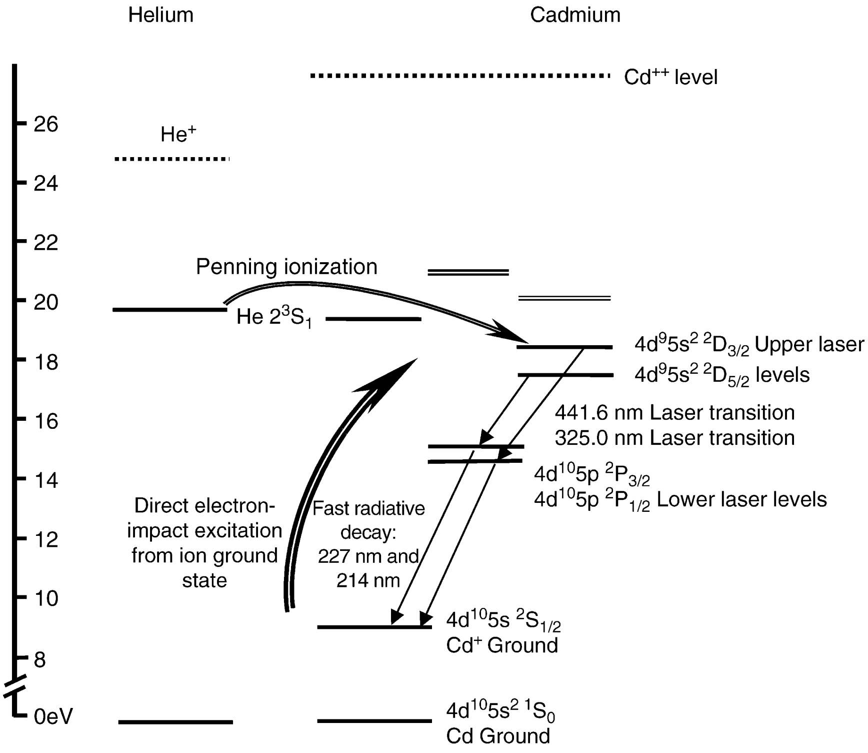
Helium
Cadmium
Cd⁺⁺ level
26
24
22
20
18
16
14
12
10
8
0eV
He⁺
Penning ionization
He 2³S₁
4d⁹5s² ²D₃/₂ Upper laser
4d⁹5s² ²D₅/₂ levels
441.6 nm Laser transition
325.0 nm Laser transition
4d¹⁰5p ²P₃/₂
4d¹⁰5p ²P₁/₂ Lower laser levels
Direct electron-impact excitation from ion ground state
Fast radiative decay: 227 nm and 214 nm
4d¹⁰5s ²S₁/₂
Cd⁺ Ground
4d¹⁰5s² ¹S₀
Cd Ground

Figure 8 Partial energy level diagram for helium and cadmium giving HeCd laser transitions.

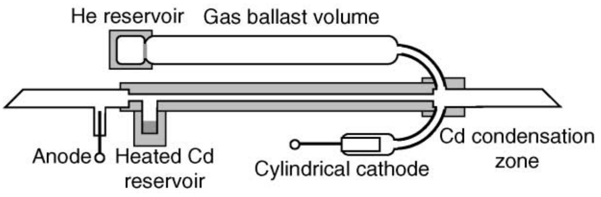

Figure 9 Helium cadmium laser tube construction.

complex than a HeNe laser, hence unit costs are considerably lower than low-power argon ion lasers which also provide output in the blue. Usually Brewster angle windows are provided together with a high Q stable resonator (97–99% reflectivity output coupler) to provide polarized output.

With their blue and UV wavelengths and relatively low cost (compared to low-power argon ion lasers), HeCd lasers have found wide application in science, medicine and industry. Of particular relevance is their application for exposing photoresists where the blue wavelength provides a good match to the peak photosensitivity of photoresist materials. A further key application is in stereolithography where the UV wavelength is used to cure an epoxy resin. By scanning the UV beam in a raster pattern across the surface of a liquid epoxy, a solid three-dimensional object may be built up in successive layers.

List of Units and Nomenclature

Boltzmann constant	[eV K^{-1}]	k
Excited metal ion		M^{+*}
Excited noble gas atom		N^*
Energy difference	[eV]	ΔE
Electron		e^-
Gas flow	[mbar l min^{-1}]	
Mirror curvatures	[m]	R_1, R_2
Metal atom		M
Noble gas atom		N
Noble gas ion		N^+
Pulse duration	[ns]	
Pulse repetition frequency	[kHz]	
Quality factor		Q
Temperature	[K]	T
Tube diameter	[mm]	D
Tube length	[m]	L
Wavelength	[nm], [μm]	
Unstable resonator magnification		M

See also

Nonlinear Sources: Harmonic Generation in Gases.

Further Reading

Little CE (1999) *Metal Vapor Lasers*. Chichester, UK: John Wiley.

Ivanov IG, Latush EL and Sem MF (1996) *Metal Vapor Ion Lasers, Kinetic Processes and Gas Discharges*. Chichester, UK: John Wiley.

Little CE and Sabotinov NV (eds) (1996) *Pulsed Metal Vapor Lasers*. Dordrecht, The Netherlands: Kluwer.

Lyabin NA, Chursin AD, Ugol'nikov SA, Koroleva ME and Kazaryan MA (2001) Development, production, and application of sealed-off copper and gold vapor lasers. *Quantum Electronics* 31: 191–202.

Petrash GG (ed.) (1989) *Metal Vapor and Metal Halide Vapor Lasers*. Commack, NY: Nova Science Publishers.

Silfvast WT (1996) *Laser Fundamentals*. New York: Cambridge University Press.

Noble Gas Ion Lasers

W B Bridges, California Institute of Technology, Pasadena, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History

The argon ion laser was discovered in early 1964, and is still commercially available in 2004, with about $70 million in annual sales 40 years after this discovery. The discovery was made independently and nearly simultaneously by four different groups; for three of the four, it was an accidental result of studying the excitation mechanisms in the mercury ion laser (historically, the first ion laser), which had been announced only months before. For more on the early years of ion laser research and development, see the articles listed in the Further Reading section at the end of this article.

The discovery was made with pulsed gas discharges, producing several wavelengths in the blue and green portions of the spectrum. Within months, continuous operation was demonstrated, as well as oscillation on many visible wavelengths in ionized krypton and xenon. Within a year, over 100 wavelengths were observed to oscillate in the ions of neon, argon, krypton, and xenon, spanning the spectrum from ultraviolet to infrared; oscillation was also obtained in the ions of other gases, for example, oxygen, nitrogen, and chlorine. A most complete listing of all wavelengths observed as gaseous ion lasers is given in the Laser Handbook cited in the Further Reading section. Despite the variety of materials and wavelengths demonstrated, however, it is the argon and krypton ion lasers that have received the most development and utilization.

Continuous ion lasers utilize high current density gas discharges, typically 50 A or more, and 2–5 mm in diameter. Gas pressures of 0.2 to 0.5 torr result in longitudinal electric fields of a few V/cm of discharge, so that the power dissipated in the discharge is typically 100 to 200 W/cm. Such high-power dissipation required major technology advances before long-lived practical lasers became available. Efficiencies have never been high, ranging from 0.01% to 0.2%. A typical modern ion laser may produce 10 W output at 20 kW input power from 440 V three-phase power lines and require 6–8 gallons/minute of cooling water. Smaller, air-cooled ion lasers, requiring 1 kW of input power from 110 V single-phase mains can produce 10–50 mW output power, albeit at even lower efficiency.

Theory of Operation

The strong blue and green lines of the argon ion laser originate from transitions between the 4p upper levels and 4s lower levels in singly ionized argon, as shown in Figure 1. The 4s levels decay radiatively to the ion ground state. The strongest of these laser lines are listed in Table 1. The notation used for the energy levels is that of the L–S coupling model. The ion ground state electron configuration is $3s^23p^5(^2P^o_{3/2})$. The inner ten electrons have the configuration $1s^22s^22p^6$, but this is usually omitted for brevity. The excited states shown in Figure 1 result from coupling a 4p or 4s electron to a $3s^23p^4(^3P)$ core, with the resulting quantum numbers S (the net spin of the electrons), L (the net angular momentum of the electrons), and J (the angular momentum resulting from coupling S to L) are represented by $^{2S+1}L_J$, where $L = 0, 1, 2, \ldots$ is denoted S, P, D, F, … The superscript ‘o’ denotes an odd level, while even levels omit a superscript. Note that some weaker transitions involving levels originating from the $3s^23p^4(^1D)$ core configuration also oscillate. Note also that the quantum mechanical selection rules for the L–S coupling model are not rigorously obeyed, although the stronger laser lines satisfy most or all of these rules. The selection rule $|\Delta J| = 1$ or 0, but not $J = 0 \rightarrow J = 0$ is always obeyed. All transitions shown in Figure 1 and Table 1 belong to the second spectrum of argon, denoted Ar II. Lines originating from transitions in the neutral atom make up the first spectrum, Ar I; lines originating from transitions in doubly ionized argon are denoted Ar III, and so forth for even more highly ionized states.

Much work has been done to determine the mechanisms by which the inverted population is formed in the argon ion laser. Reviews of this extensive research

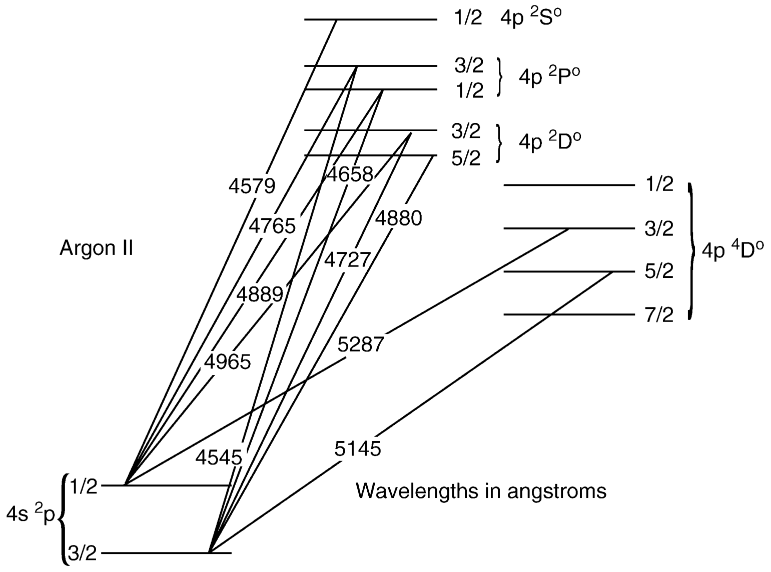

Figure 1 4p and 4s doublet levels in singly ionized argon, showing the strongest blue and green laser transitions.

Table 1 Ar II laser blue-green wavelengths

Wavelength (nanometers)	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
454.505	4p $^2P^o_{3/2} \rightarrow$ 4s $^2P_{3/2}$	0.8
457.935	4p $^2S^o_{1/2} \rightarrow$ 4s $^2P_{1/2}$	1.5
460.956	(^1D)4p $^2F^o_{7/2} \rightarrow$ (^1D)4s $^2D_{5/2}$	–
465.789	4p $^2P^o_{1/2} \rightarrow$ 4s $^2P_{3/2}$	0.8
472.686	4p $^2D^o_{3/2} \rightarrow$ 4s $^2P_{3/2}$	1.3
476.486	4p $^2P^o_{3/2} \rightarrow$ 4s $^2P_{1/2}$	3.0
487.986	4p $^2D^o_{5/2} \rightarrow$ 4s $^2P_{3/2}$	8.0
488.903	4p $^2P^o_{1/2} \rightarrow$ 4s $^2P_{1/2}$	[c]
496.507	4p $^2D^o_{3/2} \rightarrow$ 4s $^2P_{1/2}$	3.0
501.716	(^1D)4p $^2D^o_{5/2} \rightarrow$ 3d $^2D_{3/2}$	1.8
514.179	(^1D)4p $^2F^o_{7/2} \rightarrow$ 3d $^2D_{5/2}$	[c]
514.532	4p $^4D^o_{5/2} \rightarrow$ 4s $^2P_{3/2}$	10
528.690	4p $^4D^o_{3/2} \rightarrow$ 4s $^2P_{1/2}$	1.8

[a] All levels are denoted in $L-S$ coupling with the (^3P) core unless otherwise indicated. Odd parity is denoted by the superscript ‘o’.

[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2080-25S argon ion laser.

[c] These lines may oscillate simultaneously with the nearby strong line, but are not resolved easily in the output beam.

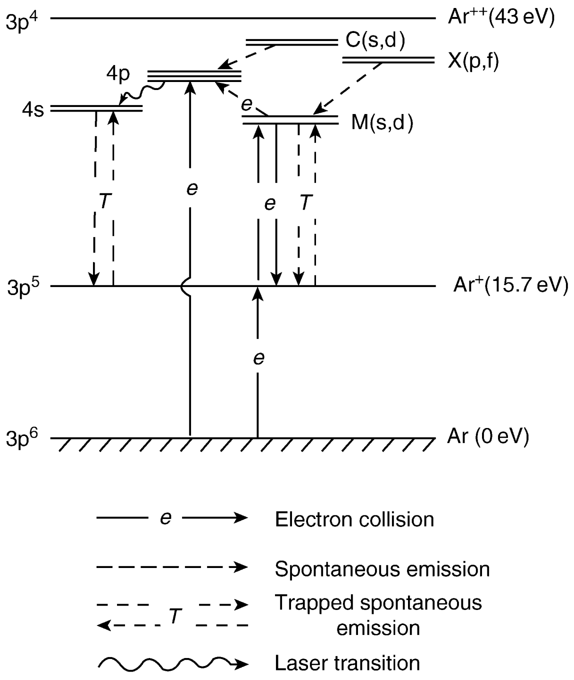

Figure 2 Schematic representation of energy levels in neutral and singly ionized argon, indicating alternative pathways for excitation and de-excitation of the argon ion laser levels.

are found in the Further Reading section. While a completely quantitative picture of argon ion laser operation is lacking to this day, the essential processes are known. Briefly, the 4p upper levels are populated by three pathways, as illustrated in Figure 2:

(i) by electron collision with the $3p^6$ neutral ground state atoms. This ‘sudden perturbation process’ requires at least 37 eV electrons, and also singles out the 4p $^2P^o_{3/2}$ upper level, which implies that only the 476 and 455 nm lines would oscillate. This is the behavior seen in pulsed discharges at very low pressure and very high axial electric field. This pathway probably contributes little to the 4p population under ordinary continuous wave (cw) operating conditions, however.

(ii) by electron collision from the lowest-lying s and d states in the ion, denoted M(s,d) in Figure 1. This only requires 3–4 eV electrons. These states have parity-allowed transitions to the ion ground state, but are made effectively metastable by radiation trapping (that is, there is a high probability that an emitted photon is re-absorbed by another ground state ion before it escapes the discharge region), or by requiring $|\Delta J|$ to be 2 to make the transition (forbidden by quantum selection rules). Thus, these levels are both created and destroyed primarily by electron collision, causing the population of the M(s,d) states to follow the population of the singly ionized ground state, which, in turn, is approximately proportional to the discharge current. Since a second electron collision is required to get from the M(s,d) states to the 4p states, a quadratic variation with current for the laser output power would be expected, and that is what is observed over some reasonable range of currents between threshold and saturation. Note that radiative decay from higher-lying opposite-parity p and f states, denoted X(p,f), can also contribute to the population of M(s,d), but the linear variation in population of the M(s,d) with discharge current is assured by electron collision creation and destruction.

(iii) by radiative decay from higher-lying opposite-parity s and d states, denoted C(s,d). These states are populated by electron collision with the $3p^5$ ion ground states, and thus have populations that also vary quadratically with discharge current. The contribution of this cascade process has been measured to be 20% to 50% of the 4p upper laser level population.

Note that it is not possible to distinguish between processes (ii) and (iii) by the variation of output power with discharge current; both give the observed quadratic dependence.

The radiative lifetimes of the 4s ^2P levels are sufficiently short to depopulate the lower laser levels by radiative decay. However, radiation trapping greatly lengthens this decay time, and a bottleneck can occur. In pulsed ion lasers, this is exhibited by the laser pulse terminating before the excitation current pulse ends, which would seem to preclude continuous operation. However, the intense discharge used in continuous operation heats the ions to the order of 2300 K, thus greatly Doppler broadening the absorption linewidth and reducing the magnitude of the absorption. Additionally, the ions are attracted to the discharge tube walls, so that the absorption spectrum is further broadened by the Doppler shift due to their wall-directed velocities. The plasma wall sheath gives about 20 V drop in potential from the discharge axis to the tube wall, so most ions hit the wall with 20 eV of energy, or about ten times their thermal velocity. A typical cw argon ion laser operates at ten times the gas pressure that is optimum for a pulsed laser, and thus the radiation trapping of the $4s \rightarrow 3p^5$ transitions is so severe that it may take several milliseconds after discharge initiation for the laser oscillation to begin.

As the discharge current is increased, the intensities of the blue and green lines of Ar II eventually saturate, and then decrease with further current. At these high currents, there is a buildup in the population of doubly ionized atoms, and some lines of Ar III can be made to oscillate with the appropriate ultraviolet mirrors. Table 2 lists the strongest of these lines, those that are available in the largest commercial lasers. Again, there is no quantitative model for the performance in terms of the discharge parameters, but the upper levels are assumed to be populated by processes analogous to those of the Ar II laser. At still higher currents, lines in Ar IV can be made to oscillate as well.

Much less research has been done on neon, krypton, and xenon ion lasers, but it is a good assumption that the population and depopulation processes are the same in these lasers. Table 3 lists both the Kr II and Kr III lines that are available from the largest commercial ion lasers. Oscillation on lines in still-higher ionization states in both krypton and xenon have been observed.

Operating Characteristics

A typical variation of output power with discharge current for an argon ion laser is shown in Figure 3. This particular laser had a 4 mm diameter discharge in a water-cooled silica tube, 71 cm in length, with a 1 kG axial magnetic field. The parameter is the argon pressure in the tube before the discharge was struck. Note that no one curve is exactly quadratic, but that the envelope of the curves at different filling pressures is approximately quadratic. At such high discharge current densities (50 A in the 4 mm tube is approximately 400 A/cm^2) there is substantial pumping of gas out of the small-bore discharge region. Indeed, a return path for this pumped gas must be provided from anode to cathode ends of the discharge to keep the discharge from self-extinguishing. The axial electric field in this discharge was 3–5 V/cm, so the input power was of the order of 10 to 20 kW, yielding an efficiency of less than 0.1%, an unfortunate characteristic of all ion lasers.

Table 2 Ultraviolet argon ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
275.392	III	$(^2D^o)4p\ ^1D_2 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.3
275.6	?	?	0.02
300.264	III	$(^2P^o)4p\ ^1P_1 \rightarrow (^2P^o)3d\ ^1D^o_2$	0.5
302.405	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.5
305.484	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)4s\ ^3P^o_1$	0.2
333.613	III	$(^2D^o)4p\ ^3F_4 \rightarrow (^2D^o)4s\ ^3D^o_3$	0.4
334.472	III	$(^2D^o)4p\ ^3F_3 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.8
335.849	III	$(^2D^o)4p\ ^3F_2 \rightarrow (^2D^o)4s\ ^3D^o_1$	0.8
350.358	III	$(^2D^o)4p\ ^3D_2 \rightarrow (^2D^o)4s\ ^3D^o_2$	0.05
350.933	III	$(^4S^o)4p\ ^3P_0 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.05
351.112	III	$(^4S^o)4p\ ^3P_2 \rightarrow (^4S^o)4s\ ^3S^o_1$	2.0
351.418	III	$(^4S^o)4p\ ^3P_1 \rightarrow (^4S^o)4s\ ^3S^o_1$	0.7
363.789	III	$(^2D^o)4p\ ^1F_3 \rightarrow (^2D^o)4s\ ^1D^o_2$	2.5
379.532	III	$(^2P^o)4p\ ^3D_3 \rightarrow (^2P^o)3d\ ^3P^o_2$	0.4
385.829	III	$(^2P^o)4p\ ^3D_2 \rightarrow (^2P^o)3d\ ^3P^o_1$	0.15
390.784	III	$(^2P^o)4p\ ^3D_1 \rightarrow (^2P^o)3d\ ^3P^o_0$	0.02
408.904	IV?	?	0.04
414.671	III	$(^2D^o)4p\ ^3P_2 \rightarrow (^2P^o)4s\ ^3P^o_2$	0.02
418.298	III	$(^2D^o)4p\ ^1P_1 \rightarrow (^2D^o)4s\ ^1D^o_2$	0.08

[a] All levels are denoted in $L-S$ coupling with the core shown in ( ). Odd parity is denoted by the superscript ‘o’.

[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2085-25S argon ion laser.

Table 3 Krypton ion laser wavelengths

Wavelength (nanometers)	*Spectrum*	*Transition*[a] *(upper level) → (lower level)*	*Relative strength*[b] *(Watt)*
337.496	III	$(^2P^o)5p\ ^3D_3 \rightarrow (^2P^o)5s\ ^3P^o_2$	–
350.742	III	$(^4S^o)5p\ ^3P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.5
356.432	III	$(^4S^o)5p\ ^3P_1 \rightarrow (^4S^o)5s\ ^3S^o_1$	0.5
406.737	III	$(^2D^o)5p\ ^1F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.9
413.133	III	$(^4S^o)5p\ ^5P_2 \rightarrow (^4S^o)5s\ ^3S^o_1$	1.8
415.444	III	$(^2D^o)5p\ ^3F_3 \rightarrow (^2D^o)5s\ ^1D^o_2$	0.3
422.658	III	$(^2D^o)5p\ ^3F_2 \rightarrow (^2D^o)4d\ ^3D^o_1$	–
468.041	II	$(^3P)5p\ ^2S^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.5
476.243	II	$(^3P)5p\ ^2D^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
482.518	II	$(^3P)5p\ ^4S^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.4
520.832	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	–
530.865	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^4P_{3/2}$	1.5
568.188	II	$(^3P)5p\ ^4D^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	0.6
631.024	III	$(^2D^o)5p\ ^3P_2 \rightarrow (^2P^o)4d\ ^3D_1$	0.2
647.088	II	$(^3P)5p\ ^4P^o_{5/2} \rightarrow (^3P)5s\ ^2P_{3/2}$	3.0
676.442	II	$(^3P)5p\ ^4P^o_{1/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	0.9
752.546	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)5s\ ^2P_{1/2}$	1.2
793.141	II	$(^1D)5p\ ^4F^o_{7/2} \rightarrow (^3P)4d\ ^2F_{5/2}$	0.3
799.322	II	$(^3P)5p\ ^4P^o_{3/2} \rightarrow (^3P)4d\ ^4D_{1/2}$	

[a] All levels are denoted in *L–S* coupling with the core shown in ( ). Odd parity is denoted by the superscript 'o'.
[b] Relative strengths are given as power output from a commercial Spectra-Physics model 2080RS ion laser.

Technology

With such high input powers required in a small volume to produce several watts output, ion laser performance has improved from 1964 to the present only as new discharge technologies were introduced. The earliest laboratory argon ion lasers used thin-walled (≈1 mm wall thickness) fused silica discharge tubes, cooled by flowing water over the outside wall of the tube. The maximum input power per unit length of discharge was limited by thermal stresses in the silica walls caused by the temperature differential from inside to outside. Typically, ring-shaped cracks would cause the tube to fail catastrophically. Attempts to make metal–ceramic structures with alumina (Al_2O_3) discharge tubes to contain the plasma were made early on (1965) but were not successful. Such tubes invariably failed from fracture by thermal shock as the discharge was turned on. Later, successful metal–ceramic tubes were made with beryllia (BeO), which has a much higher thermal conductivity than silica or alumina and is much more resistant to thermal shock. Today, all of the lower power (less than 100 mW output) are made with BeO discharge tubes. Some ion lasers in the 0.5 to 1 W range are also made with water-cooled BeO discharge tubes.

A typical low-power air cooled argon ion laser is shown in Figure 4. The large metal can (2) on the right end of the tube contains an impregnated-oxide hot cathode, heated directly by current through ceramic feed-through insulators (3). The small (≈1 mm diameter) discharge bore runs down the center of the BeO ceramic rod (1), and several smaller diameter gas return path holes run off-axis parallel to the discharge bore to provide the needed gas equalization between cathode can and anode region. A copper honeycomb cooler is brazed to the cathode can, two more to the outer wall of the BeO cylinder, and one to the anode (4) at the left end of the tube. The laser mirrors (5) are glass-fritted to the ends of the tube, forming a good vacuum seal. Note that in this small laser, the active discharge bore length is less than half of the overall length.

The very early argon ion lasers were made with simple smooth dielectric tubes, allowing the continuous variation in voltage along the length of the tube required by the discharge longitudinal electric field. However, this variation can be step-wise at the discharge walls, and still be more or less smooth along the axis. Thus, the idea arose of using metal tube segments, insulated one from another, to form the discharge tube walls. The first version of this idea (1965) used short (≈1 cm) metal cylinders supported by metal disks and stacked inside a large diameter silica envelope, with each metal cylinder electrically isolated from the others. The tubes and disks were made of molybdenum, and were allowed to heat to incandescence, thus radiating several kilowatts of heat through the silica vacuum envelope to a water-cooled collector outside. While this eliminated the problems of thermal shock and poor thermal conductivity inherent in dielectric wall discharges, it made another

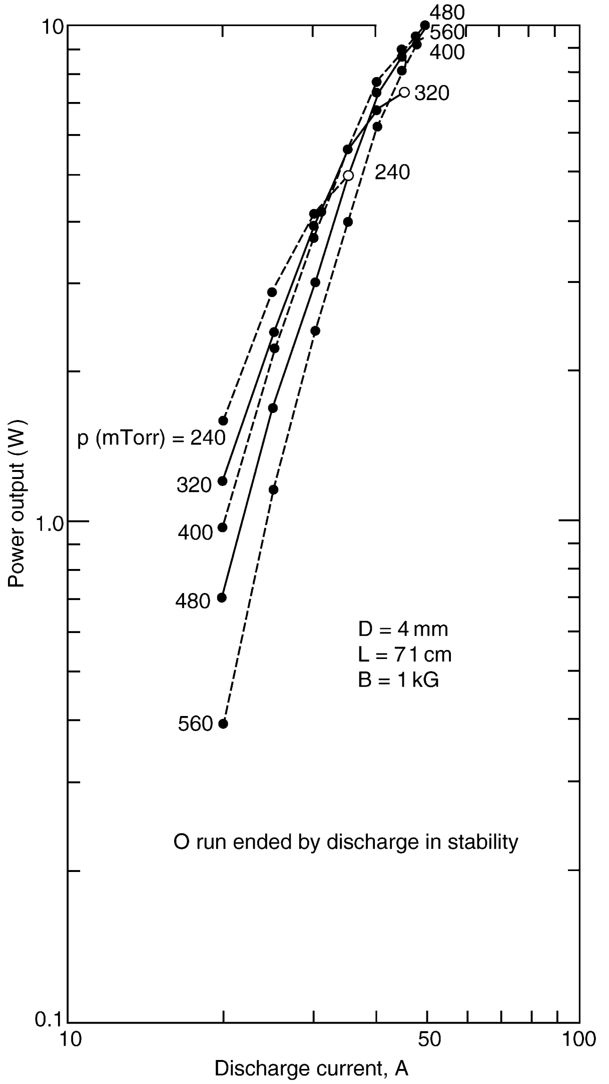

Figure 3 Laser output power (summed over all the blue and green laser lines) versus discharge current for a laser discharge 4 mm in diameter and 71 cm long, with a 1 kilogauss longitudinal magnetic field. The parameter shown is the argon fill pressure in mTorr prior to striking the discharge. The envelope of the curves exhibits the quadratic variation of output power with discharge current.

problem painfully evident. The intense ion bombardment of the metal tube walls sputtered the wall material, eventually eroding the shape of the metal cylinders and depositing metal films on the insulating wall material, thus shorting one segment to another.

Many different combinations of materials and configurations were investigated in the 1960s and 1970s to find a structure that would offer good laser performance and long life. It was found that simple thin metal disks with a central hole would effectively confine the discharge to a small diameter (on the order of the hole size) if a longitudinal d-c magnetic field of the order of 1 kiloGauss were used. The spacing between disks can be as large as 2–4 discharge diameters. Of the metals, tungsten has the lowest sputtering yield for argon ions in the 20 eV range (which is approximately the energy they gain in falling to the wall across the discharge sheath potential difference). An even lower sputtering yield is exhibited by carbon, and graphite cylinders contained within a larger diameter silica or alumina tube were popular for a while for ion laser discharges. Unfortunately, graphite has a tendency to flake or powder, so such laser discharge tubes became contaminated with 'dust' which could eventually find its way to the optical windows of the tube. Beryllia and silica also sputter under argon ion bombardment, but with still lower yields than metals or carbon; however, their smaller thermal conductivities limit them to lower-power applications.

The material/configuration combination that has evolved for higher-power argon ion lasers today is to use a stack of thin tungsten disks with 2–3 mm diameter holes for the discharge. These disks, typically 1 cm in diameter, are brazed coaxially to a larger copper annulus (actually, a drawn cup with a 1 cm hole on its axis). A stack of these copper/tungsten structures is, in turn, brazed to the inside wall of a large diameter alumina vacuum envelope. The tungsten disks are exposed to and confine the discharge, while the copper cups conduct heat radially outward to the alumina tube wall, which, in turn, is cooled by fluid flow over its exterior. Thus, the discharge is in contact only with a low-sputtering material (tungsten) while the heat is removed by a high thermal conductivity material (copper). Details differ among manufacturers, but this 'cool disk' technology seems to have won out in the end. A photo of a half-sectioned disk/cup stacked assembly from a Coherent Innova™ ion laser is shown in Figure 5. The coiled impregnated-tungsten cathode is also shown.

Sputtering of the discharge tube walls is not uniform along the length of the gas discharge. The small diameter region where the laser gain occurs is always joined to larger diameter regions containing cathode and anode electrodes (as shown in Figure 5, for example). A plasma double sheath (that is, a localized increase in potential) forms across the discharge in the transition region between large and small diameter regions (the discharge 'throats,' which may be abrupt or tapered). This sheath is required to satisfy the boundary conditions between the plasmas of different temperatures in the different regions. Such a double sheath imparts additional energy to the ions as they cross the sheath, perhaps

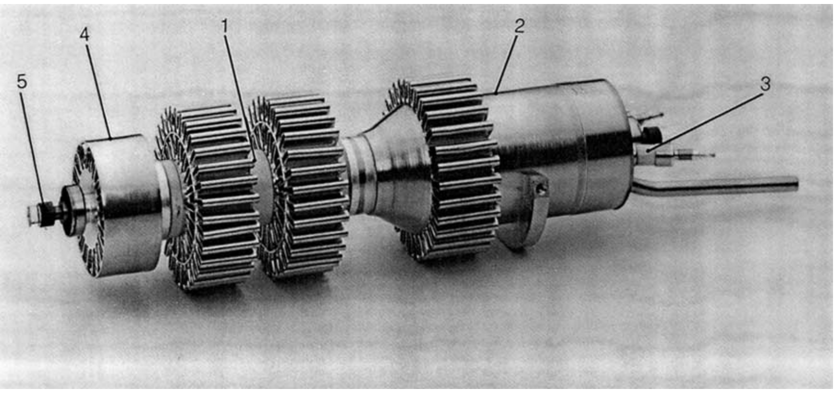

Figure 4 A typical commercial low-power, air-cooled argon ion laser. The cathode can (2) is at the right and the beryllia bore (1) and anode (4) is at the left. The cathode current is supplied through ceramic vacuum feed-throughs (3). Mirrors (5) are glass fritted onto the ends of vacuum envelope. (Photo courtesy of JDS Uniphase Corp.)

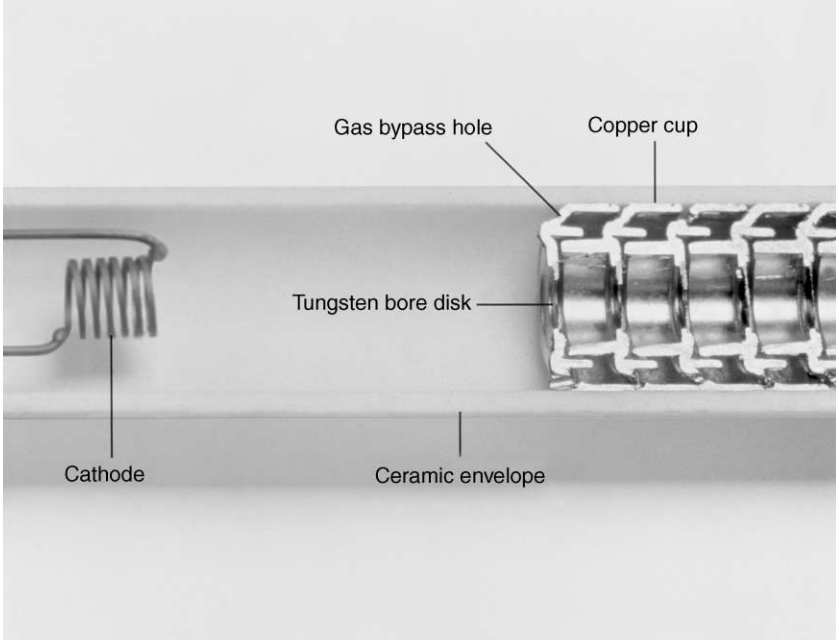

Figure 5 Discharge bore structure of a typical commercial high-power argon ion laser, the Coherent Innova™. Copper cups are brazed to the inner wall of a ceramic envelope, which is cooled by liquid flow over its outer surface. A thin tungsten disk with a small hole defining the discharge path is brazed over a larger hole in the bottom of the copper cup. Additional small holes in the copper cup near the ceramic envelope provide a gas return path from cathode to anode. Also shown is the hot oxide-impregnated tungsten cathode. (Photo courtesy of Coherent, Inc.)

an additional 20 eV. When these ions eventually hit the discharge walls near the location of the sheath, they have 40 eV of energy rather than the 20 eV from the normal wall sheath elsewhere in the small diameter discharge. Sputtering yield (number of sputtered wall atoms per incident ion) is exponentially dependent on ion energy in this low ion energy region, so the damage done to the wall in the vicinity of the 'throat' where the double sheath forms may be more than ten times that elsewhere in the small diameter bore region. This was the downfall of high-power operation of dielectric discharge bores, even BeO; while sputtering was acceptable elsewhere in the discharge, the amount of material removed in a small region near the discharge throat would cause catastrophic bore failure at that point. This localized increase in sputtering in the discharge throat is common to all ion lasers, including modern cooled tungsten disk tubes. Eventually, disks near the throat are eroded to larger diameters, and more material is deposited on the walls nearby. Attempts to minimize localized sputtering by tapering the throat walls or the confining magnetic field have proven unsuccessful; a localized double sheath always forms somewhere in the throat. Because of the asymmetry caused by ion flow, the double sheath is larger in amplitude in the cathode throat than the anode throat, so that the localized wall damage is larger at the cathode end of the discharge than at the anode end.

Another undesirable feature of sputtering is that the sputtered material 'buries' some argon atoms when it is deposited on a wall. This is the basis of the well-known Vac-Ion® vacuum pump. Thus, the operating pressure in a sealed laser discharge tube will drop during the course of operation. In low-power ion lasers, this problem is usually solved by making the gas reservoir volume large enough to satisfy the desired operating life (for example, the large cathode can in Figure 4). In high-power ion lasers, the gas loss would result in unacceptable operating life even with a large reservoir at the fill pressure. Thus, most high-power ion lasers have a gas pressure measurement system and dual-valve arrangement connected to a small high-pressure reservoir to 'burp' gas into the active discharge periodically to keep the pressure within the operating range. Unfortunately, if a well-used ion laser is left inoperative for months, some of the 'buried' argon tends to leak back into the tube, with the gas pressure becoming higher than optimum. The pressure will gradually decrease to its optimum value with further operation of the discharge (in perhaps tens of hours). In the extreme case, the gas pressure can rise far enough so that the gas discharge will not strike, even at the maximum power supply voltage. Such a situation requires an external vacuum pump to remove the excess gas, usually a factory repair.

In addition to simple dc discharges, various other techniques have been used to excite ion lasers, primarily in a search for higher power, improved efficiency and longer operating life. Articles in the Further Reading section give references to these attempts. Radio-frequency excitation at 41 MHz was used in an inductively coupled discharge, with the laser bore and its gas return path forming a rectangular single turn of an air-core transformer. A commercial product using this technique was sold for a few years in the late 1960s. Since this was an 'electrode-less' discharge, ion laser lines in reactive gases such as chlorine could be made to oscillate, as well as the noble gases, without 'poisoning' the hot cathode used in conventional dc discharge lasers. A similar electrode-less discharge was demonstrated as a quasi-cw laser by using iron transformer cores and exciting the discharge with a 2.5 kHz square wave. Various microwave excitation configurations at 2.45 GHz and 9 GHz also resulted in ion laser oscillation. Techniques common to plasma fusion research were also studied. Argon ion laser oscillation was produced in Z-pinch and Θ-pinch discharges and also by high-energy (10–45 keV) electron beams. However, none of these latter techniques resulted in a commercial product, and all had efficiencies worse than the simple dc discharge lasers.

It is interesting that the highest-power output demonstrations were made in less than ten years after the discovery. Before 1970, 100 W output on the argon blue-green lines was demonstrated with dc discharges two meters in length. In 1970, a group in the Soviet Union reported 500 W blue-green output from a two-meter discharge with 250 kW of dc power input, an efficiency of 0.2%. Today, the highest output power argon ion laser offered for sale is 50 W output.

Manufacturers

More than 40 companies have manufactured ion lasers for sale over the past four decades. This field has now (2004), narrowed to the following:

Coherent, Inc.	http://www.coherentinc.com
INVERSion Ltd.	http://inversion.iae.nsk.su
JDS Uniphase	http://www.jdsu.com
Laser Physics, Inc.	http://www.laserphysics.com
Laser Technologies GmbH	http://www.lg-lasertechnologies.com
LASOS Lasertechnik GmbH	http://www.LASOS.com
Lexel Laser, Inc.	http://www.lexellaser.com
Melles Griot	http://lasers.mellesgriot.com
Spectra-Physics, Inc.	http://www.spectraphysics.com

Further Reading

Bridges WB (1979) Atomic and ionic gas lasers. In: Marton L and Tang CL (eds) *Methods of Experimental Physics, Vol 15: Quantum Electronics, Part A*, pp. 31–166. New York: Academic Press.

Bridges WB (1982) Ionized gas lasers. In: Weber MJ (ed.) *Handbook of Laser Science and Technology, Vol. II, Gas Lasers*, pp. 171–269. Boca Raton, FL: CRC Press.

Bridges WB (2000) Ion lasers – the early years. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 885–898.

Davis CC and King TA (1975) Gaseous ion lasers. In: Goodwin DW (ed.) *Advances in Quantum Electronics*, vol. 3, pp. 169–469. New York: Academic Press.

Dunn MH and Ross JN (1976) The argon ion laser. In: Sanders JH and Stenholm S (eds) *Progress in Quantum Electronics*, vol. 4, pp. 233–269. New York: Pergamon.

Weber MJ (2000) *Handbook of Laser Wavelengths*. Boca Raton, FL: CRC Press.

Optical Fiber Lasers

G E Town, Macquarie University, NSW, Australia

N N Akhmediev, Australian National University, ACT, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Optical fiber lasers were first demonstrated in the 1960s, and since then have developed to become versatile optical sources with many desirable properties. Aided by developments in associated technologies, such as fiber design and fabrication methods, semiconductor pump diode technology, and fiber-coupled and in-fiber components such as Bragg grating filters, optical fiber lasers now compete with other laser technologies in many applications, from telecommunications to materials processing.

An optical fiber laser is fundamentally an optical oscillator, which converts input pump power to coherent optical output power at one or more well-defined wavelengths. Optical oscillators require two basic elements: optical gain, and optical feedback. For sustained oscillation to occur the round-trip gain in the laser cavity must be unity, and the round-trip phase a multiple of 2π. In optical fiber lasers the optical gain is provided by an optical fiber amplifier using one or a combination of fundamental physical processes, e.g., stimulated emission, stimulated scattering, or nonlinear parametric processes. Early fiber amplifiers used stimulated Raman scattering to produce optical gain; however, rare-earth-doped fiber amplifiers, in which gain is provided by stimulated emission, are now more common. The reader is referred to **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems, for more details on doped fiber amplifiers, and to **Scattering:** Stimulated Scattering, **Nonlinear Optics, Applications:** Raman Lasers, **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed), and **Scattering:** Raman Scattering, for more details on stimulated scattering and parametric gain processes. Optical feedback may be provided in two fundamental ways, e.g., by using a closed ring of fiber, or from reflections from nonuniformities and discontinuities in the waveguide such as in Bragg grating filters or at fiber ends. The reader is referred to **Fiber Gratings** for more details on fiber Bragg gratings.

The main advantages of optical fiber lasers are derived from the confinement of the pump and signal in a small optical waveguide. In contrast with bulk lasers, the pump intensity in the fiber waveguide is largely independent of the laser length, resulting in large amplifier gain and low laser threshold, even for gain media with small absorption and emission cross-sections. The large gain enables lossy elements such as optical fiber-coupled devices and bulk optical elements to be incorporated into the laser cavity, providing additional control over the optical signal being generated. To obtain a large gain optical fiber amplifiers must usually be relatively long (i.e., from several centimeters to many meters), and so the linear and nonlinear properties of the fiber waveguide can have a significant influence on the optical signal being generated, resulting in some interesting and useful phenomena, e.g., as in soliton fiber lasers. The reader is referred to **Fiber and Guided Wave Optics:** Optical Fiber Cables, for fiber-based components, to **Fiber and Guided Wave Optics**: Dispersion, Light Propagation, and Nonlinear Effects (Basics), for a review of the linear and nonlinear properties of optical fibers, and to **Solitons:** Soliton Communication Systems, and Temporal Solitons, for the theory and applications of temporal solitons.

Whilst optical fibers may be fabricated in a range of materials, including polymer and crystalline materials, most fibers and fiber lasers are currently made of silica containing rare-earth ion dopants – see **Fiber and Guided Wave Optics:** Fabrication of Optical Fiber, for details. Lasers in silica optical fiber have much in common with other glass-host lasers, including a wide range of potential pump and lasing wavelengths, and broadband absorption and gain due to homogeneous and inhomogeneous broadening of the lasing energy levels in the amorphous glass host. For example, erbium-doped amplifiers in alumino-silicate glass fibers at room temperature have lasing transitions with homogeneous and

Table 1 Commonly used pump and signal transitions in rare-earth doped fiber lasers

Rare earth dopant	*Pump wavelengths [nm]*	*Signal wavelengths [nm]*	*Excited state lifetime [ms]*	*Energy levels*
Praseodymium	480, 585	885, 1080	0.1	4
Neodymium	590, 800	920 (900–950)	0.5	3
		1060 (1055–1140)		4
Samarium	488	651	1.5	4
Holmium	455, 650, 1150	2040	0.6	3
Erbium	800, 975, 1480	1550 (1530–1610)	10	3
Thulium	790, 1210 (1060–1320)	1480 (1460–1520)		
		1850 (1650–2050)	0.3	3
Ytterbium	920 (840–980)	975	0.8	3
		1040 (1010–1160)		4

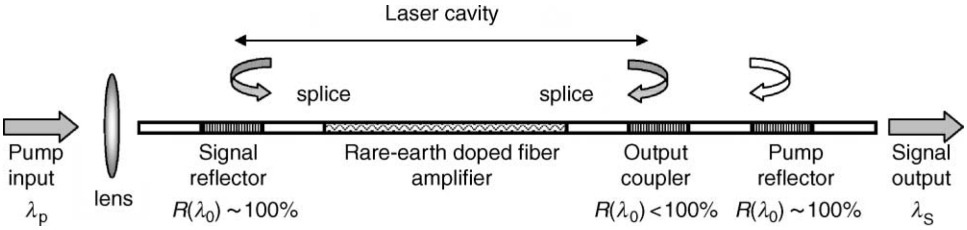
Laser cavity
splice
splice
Pump input λ_p
lens
Signal reflector $R(\lambda_0) \sim 100\%$
Rare-earth doped fiber amplifier
Output coupler $R(\lambda_0) < 100\%$
Pump reflector $R(\lambda_0) \sim 100\%$
Signal output λ_S

Figure 1 Example of a simple all-fiber optical oscillator. The rare-earth doped fiber provides optical gain, and the Bragg gratings provide narrowband optical feedback.

inhomogeneous linewidths of several nanometers, and can be optimized to provide around 30 dB gain over an 80 nm bandwidth centered around 1570 nm. The large gain-bandwidth is useful for achieving either wide tunability of the lasing wavelength and/or ultrashort pulse generation. Table 1 summarizes the most commonly used pump and signal transitions for a range of rare-earth dopants in silicate optical fibers.

Other desirable properties of silica fiber lasers include high efficiency, excellent thermal dissipation, substantial energy storage, high power capability, and compatibility with a wide range of optical fiber devices and systems. The characteristics of optical fiber lasers can be optimized for some very different applications, ranging from narrow linewidth and low-noise lasers to broadband pulsed lasers with high energy and/or high peak power. In the following sections the fundamentals of laser theory as applied to fiber lasers will be reviewed, highlighting the reasons underlying such versatility.

Fiber Laser Fundamentals

An example of a simple all-fiber laser is shown schematically in Figure 1. The laser contains the basic elements of optical gain (e.g., a rare-earth doped fiber amplifier) and optical feedback (e.g., a pair of Bragg grating filters). In the laser shown, the two Bragg gratings provide optical feedback only over a narrow range of wavelengths, which further restricts the range of possible lasing wavelengths. One grating must have a reflectivity less than unity to allow a proportion of the lasing signal to be coupled out of the laser cavity. At threshold, the gain in the amplifier exactly compensates for the loss in the output coupler.

The optical pump, typically from a semiconductor laser diode, is absorbed as is propagates along the fiber amplifier, e.g., by rare-earth ion dopants, which are raised to an excited state. The absorbed energy is stored and eventually emitted at a longer wavelength by either spontaneous or stimulated emission. Spontaneous emission produces unwanted noise and is associated with a finite excited state lifetime, whilst stimulated emission produces optical gain provided the gain medium is ‘inverted’, i.e., with more active ions in the upper lasing energy level than the lower lasing energy level. In practice some energy is also lost to nonradiative (thermal) emissions, e.g., in the transfer of energy between the pump and upper lasing energy levels. In an ideal gain medium the nonradiative transfer rate between the pump and upper lasing energy levels is much larger than the spontaneous or stimulated emission rates, so the inversion may usually be assumed to be independent of nonradiative transfer rates. Energy level diagrams for three- and four-level atoms are shown in Figure 2.

The local optical gain coefficient (i.e. power gain per unit length, in nepers^2) in a fiber amplifier is proportional to the local inversion, i.e., $g(z) \sim \sigma_S \Delta N(z)$, where σ_S is the emission cross-section at the

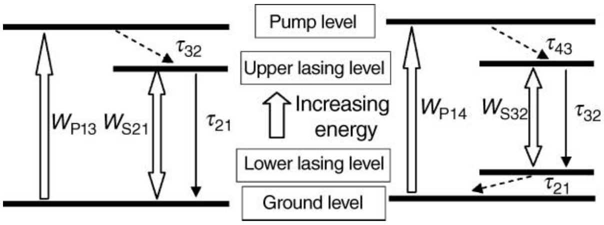

Figure 2 Energy level diagrams for three- and four-level laser transitions. *W* represents the probability of an optical transition per unit time associated with stimulated absorption or emission, and $1/\tau$ is the rate of spontaneous emission. Dotted lines represent nonradiative transitions (e.g., between the pump and upper lasing level).

signal wavelength, and $\Delta N(z)$ the difference in population density between the upper and lower lasing energy levels ($\Delta N > 0$ if the gain medium is inverted). The local inversion can be calculated by solving the rate equations including all stimulated and spontaneous transitions between the laser energy levels. The end-to-end amplifier power gain, $G = P_S(L_A)/P_S(0)$, in length L_A of amplifying fiber is then

$$G = \exp\left[\int_0^{L_A} g(z)\,dz\right] \qquad [1]$$

which in decibels is

$$G = 4.34\int_0^{L_A} g(z)\,dz$$

A large gain coefficient is desirable for amplification with minimal added noise (i.e., low noise figure), and a large gain per unit pump power is desirable for low threshold lasing. For a given gain medium the only factor usually accessible to control the inversion and hence gain is the pump rate, i.e., the rate at which pump energy is absorbed by the gain medium, which for an unbleached amplifier is proportional to pump intensity. For example, in an ideal four-level unsaturated laser medium with uniform pump intensity, the small signal gain is approximately

$$\int_0^{L_A} g(z)\,dz \approx \sigma_S \frac{P_{\mathrm{Pabs}}\tau}{h\nu_P A} \qquad [2]$$

in which τ is the lifetime of the upper lasing level, P_{Pabs} is the pump power absorbed in the fiber amplifier with cross-sectional area A, h is Planck's constant, and ν_P is the pump frequency.

For a three-level laser under the same conditions, the small signal gain is

$$\int_0^{L_A} g(z)\,dz \approx -\sigma_S N L_A + \sigma_S \frac{2P_{\mathrm{Pabs}}\tau}{h\nu_P A} \qquad [3]$$

in which the first term corresponds to ground state absorption of the signal, in which N is the dopant ion density, and the factor of 2 in the second term is due to the fact that in an ideal three-level system every absorbed pump photon increases the inversion, ΔN, by two. Equation [3] can be approximated by multiplying [2] by the factor $(P_{\mathrm{Pabs}} - P_{\mathrm{Psat}})/(P_{\mathrm{Pabs}} + P_{\mathrm{Psat}})$, where $P_{\mathrm{Psat}} = h\nu_P A/(\sigma_P \tau)$ is the pump power which must be absorbed to reduce the pump absorption coefficient to half its unpumped value (or the signal absorption coefficient to zero), and σ_P is the pump absorption cross-section. Unlike four-level lasers in which gain is available as soon as pump power is applied, three-level lasers require half the active ions to be excited before the inversion becomes positive and gain is produced.

The main advantages of optical fiber amplifiers are now clear; in an optical fiber waveguide the light is confined in a very small cross-sectional area, typically less than ten microns diameter. Consequently even a small coupled pump power can have a large intensity and produce a large gain coefficient, or low laser threshold. Even if the pump absorbed per unit length in the fiber amplifier is small (e.g., due to low dopant concentration or low absorption coefficient), this can often be compensated by using a long length of fiber with small intrinsic loss. Furthermore, rare-earth dopants have relatively long metastable state lifetimes, τ, which further assists in producing a large inversion throughout the fiber, and hence large overall gain or low lasing threshold, even at low pump powers. For example, a typical four-level neodymium doped fiber amplifier providing gain at $\lambda_S = 1.06\ \mu\mathrm{m}$, pumped at $\lambda_P = 0.8\ \mu\mathrm{m}$, with core diameter $7\ \mu\mathrm{m}$, $\tau = 0.5$ ms, $\sigma_S = 1.4 \times 10^{-20}\ \mathrm{cm}^2$, and $\sigma_P = 23 \times 10^{-21}\ \mathrm{cm}^2$, could theoretically provide a small signal gain of 0.3 dB for every milliwatt of absorbed pump power.

Whilst the above analysis highlights the main advantages of fiber amplifiers, it is approximate in a number of respects. For example, eqns [2] and [3] assume both pump and signal have a uniform intensity distribution across the gain medium, and completely overlap. In single mode optical fibers the intensity distribution across the fiber core is not uniform but approximately Gaussian, with a width and effective area that depends on the numerical aperture of the waveguide and the wavelength. Consequently the overlap of the pump and signal beams with the doped fiber core and with each other is less than 100%, and the pump and signal intensities are higher in the center of the fiber than at the core-cladding boundary. The nonuniform distribution of pump and signal beams slightly modifies both the gain efficiency and gain saturation behavior (i.e., the dependence of gain on signal power) in fiber amplifiers. More accurate analysis would also take into account factors such as the rare-earth dopant distribution (which may not be uniform), the variation of pump and signal intensities along the fiber, spectral variation in the absorption and emission cross-sections, spatial and spectral hole-burning in the gain medium, degeneracies in the energy levels, excited state absorption and other loss mechanisms, temperature, and spontaneous emission noise.

Continuous Wave Fiber Lasers

The ideal continuous wave (cw) laser converts input pump power with low coherence to a highly coherent optical output signal which is constant in amplitude and wavelength, with the spectral purity of the output (i.e., linewidth) limited only by cavity losses. Practical continuous wave lasers may be characterized by four parameters: the pump power required for the onset of oscillation (i.e., threshold); the conversion efficiency of pump to signal power above threshold; the peak wavelength of the optical output; and the spectral width of the optical output. Continuous wave fiber lasers often have a low threshold, high efficiency, and narrow linewidth relative to other types of laser. Output powers approaching 100 watts have been achieved using specialized techniques.

The threshold power is determined by the total cavity loss and gain efficiency (i.e., gain per unit of absorbed pump power) of the fiber amplifier. For example, for the laser shown in Figure 1 the internal losses are minimal, and the pump power required to reach threshold may be calculated by setting the product of the round-trip small-signal gain, and the output mirror reflectivity, G^2R, equal to unity. For the laser configuration shown in Figure 1, and using eqns [1] and [2] for a four-level laser, the pump power which must be absorbed in the amplifier for the laser to reach threshold is

$$P_{\mathrm{th}} = \alpha_{\mathrm{cav}} L_{\mathrm{C}} \frac{h\nu_{\mathrm{P}} A}{\sigma_{\mathrm{S}} \tau} \qquad [4]$$

in which L_{C} is the cavity length, and $\alpha_{\mathrm{cav}} = \alpha_{\mathrm{int}} + \frac{1}{2} L_{\mathrm{C}} \ln(1/R)$ is the total cavity loss per unit length, comprising both internal losses, α_{int}, and outcoupling losses through a mirror with reflectivity R. For example, if the internal losses were negligible and the transmittance, $T = 1 - R$, of the output reflector in the laser shown in Figure 1 was 50% (i.e. 3 dB), and the other fiber parameters were the same as above (i.e. $G = 0.3$ dB/mW), then the laser threshold would be 5 mW. In three-level fiber lasers ground state absorption of the signal is usually the dominant loss mechanism which must be overcome before a net gain is produced, and hence the threshold of three-level lasers is given approximately by eqn [4] with $\alpha_{\mathrm{int}} = \sigma_{\mathrm{S}} N L_{\mathrm{A}} 2 L_{\mathrm{C}}$, and increases with amplifier length. In either case the pump power required at the input of the laser to reach threshold, $P_{\mathrm{P}}(0)$, may be approximated using the relation $P_{\mathrm{Pabs}} = P_{\mathrm{P}}(0) - P_{\mathrm{P}}(L_{\mathrm{A}}) \approx P_{\mathrm{P}}(0)[1 - \exp(-\sigma_{\mathrm{P}} N L_{\mathrm{A}})]$.

For either three- or four-level lasers pumped above threshold, the inversion of the gain medium remains at its threshold level as the internal round-trip gain must remain unity; however a coherent optical output signal at the peak emission wavelength grows from noise until its amplitude is limited by saturation of the amplifier gain. Provided the output coupling loss is not too large (e.g., $T < 60\%$) such that the total signal intensity is constant along the whole length of the fiber amplifier, it can be shown that changes in pump power above threshold cause a proportional change in the signal power coupled out of the Fabry–Perot cavity in one direction, P^{+}_{Sout}, i.e.

$$P^{+}_{\mathrm{Sout}} = \frac{(1-R)}{\delta_0} \frac{\nu_{\mathrm{S}}}{\nu_{\mathrm{P}}} (P_{\mathrm{Pabs}} - P_{\mathrm{th}}) \qquad [5]$$

in which $\delta_0 = 2\alpha_{\mathrm{int}} L + \ln 1/R$ represents the total round-trip loss at threshold. The relationship between absorbed pump power and output signal power is shown schematically in Figure 3. The constant of proportionality is known as the slope or conversion efficiency, defined as $\eta_{\mathrm{S}} = P_{\mathrm{Sout}}/(P_{\mathrm{Pabs}} - P_{\mathrm{th}})$. In both three- and four-level fiber lasers with small intrinsic losses $\alpha_{\mathrm{int}} \approx 0$, and small output coupling (i.e., $R \approx 1$), the slope efficiency can approach the intrinsic quantum efficiency, $\nu_{\mathrm{S}}/\nu_{\mathrm{P}}$.

Sustained oscillation can only occur if the optical path length around the cavity is an integral number of wavelengths. Wavelengths satisfying this condition are called ‘modes’ of the cavity, determined by

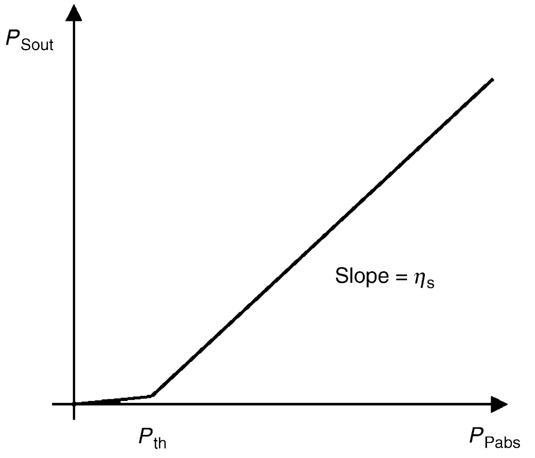

Figure 3 Representative plot of laser output power versus absorbed pump power.

$2nL_C = m\lambda_0$, where n is the effective refractive index seen by light propagating in the cavity, L_C is the cavity length, m is an integer, and λ_0 is the free-space wavelength. In Figure 1, the cavity length comprises the amplifying fiber, and two Bragg grating filters which may be regarded as having an effective length (i.e., less than the actual grating length) which depends on the grating characteristics. In fiber lasers the cavity is typically between several centimeters and many meters in length, hence there are usually very many potential lasing modes within the gain-bandwidth of the amplifying medium. For example, in the Fabry–Perot cavity of Figure 1, if the cavity length was 5 meters, the mode spacing would be $\nu_{m+1} - \nu_m = c/(2nL_C) = 20$ MHz, i.e. orders of magnitude less than the bandwidth of the optical amplifier.

Ideally the first mode to reach threshold would determine the laser output wavelength and the spectral width, or linewidth, of the laser output would be $\Delta\nu_L = h\nu/(2\pi\tau_C^2 P_{\mathrm{Sout}})$, in which τ_C is the lifetime of photons in the cavity and determined by cavity losses, including outcoupling, and given by $\tau_C = t_R/\varepsilon$, where t_R is the cavity round trip time and ε the fractional energy loss in the cavity per round trip. In practice, the linewidth is usually significantly larger than the latter theoretical limit due to transient lasing of different modes within the bandwidth of the optical amplifier and environmental perturbations to the fiber cavity. Fiber amplifiers typically have a large homogeneously broadened gain-bandwidth, and being long and flexible are susceptible to acoustic and thermal perturbations, hence care must be taken to physically stabilize the laser cavity to minimize both mode-hopping and the output linewidth.

Whilst the CW laser shown in Figure 1 has the advantage of simple construction, it would not be ideal for narrow linewidth CW generation. The reason is that counterpropagating waves in the Fabry–Perot cavity form a standing wave, i.e., in which the local signal intensity varies due to interference between forward and backward propagating waves. The standing wave can cause spatial hole-burning in the gain medium (i.e., a spatial variation in gain, linked to the spatial variation in intensity) which in turn reduces the average gain for the lasing mode and promotes mode-hopping and hence spectral broadening of the output.

A preferable arrangement for many fiber lasers is a traveling wave laser, which can be realized using a ring configuration, as shown in Figure 4. In this configuration the optical isolator ensures unidirectional lasing and avoidance of spatial hole-burning. Additional components in the cavity can include a wavelength selective coupler (WSC), i.e., to couple the pump wavelength into the laser cavity whilst not coupling the signal wavelength out, and a narrowband filter (e.g. a small Fabry–Perot resonator) to further stabilize and/or narrow the lasing linewidth. Rare-earth doped fiber lasers with similar cavity configurations have been realized with linewidths of approximately 10 kHz, limited by environmental (e.g., acoustic) perturbations.

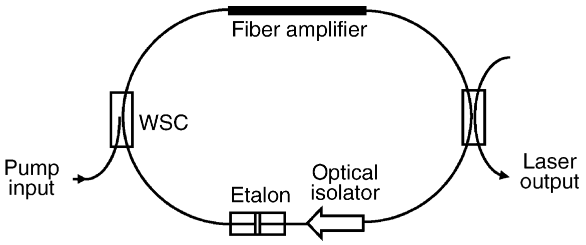

Figure 4 Schematic of a traveling wave laser, useful for low noise continuous wave oscillation. The wavelength selective coupler (WSC) is required to couple pump power into the laser cavity without coupling signal out.

Pulsed Fiber Lasers

There are two main types of fiber laser useful for generating short high-power pulses: *Q*-switched lasers, and mode-locked lasers. *Q*-switched fiber lasers are useful for generating large energy pulses (e.g., microjoules) with very high peak power (e.g., kilowatts) with relatively long pulse duration (e.g., tens of nanoseconds), whilst mode-locked fiber lasers are typically capable of generating ultrashort pulses (e.g., sub-picosecond) with moderate energy (e.g., nanojoules) and moderate-to-high peak power (e.g., tens of watts). Pulsed lasers typically produce less average output power than CW lasers (e.g., up to about 10 W). Both types of pulsed laser typically contain an element for loss modulation within the cavity, however the pulse generation mechanisms are very different.

Q-switched lasers operate by rapid switching of the cavity loss. Whilst the intracavity loss is high and the laser below threshold, energy is transferred from the pump to the lasing medium. The long excited state lifetime and small emission cross-sections of rare-earth dopants assists greatly in the latter process, so that a significant amount of energy can be stored in long fiber amplifiers before amplification of spontaneous emission noise begins to deplete the inversion. After the gain medium is fully inverted the cavity loss is suddenly reduced, and the laser is taken well above threshold, resulting in a rapid build-up of noise and the formation of a pulse which ideally extracts all the stored energy within a few round-trips in the laser cavity. The situation is shown schematically in Figure 5.

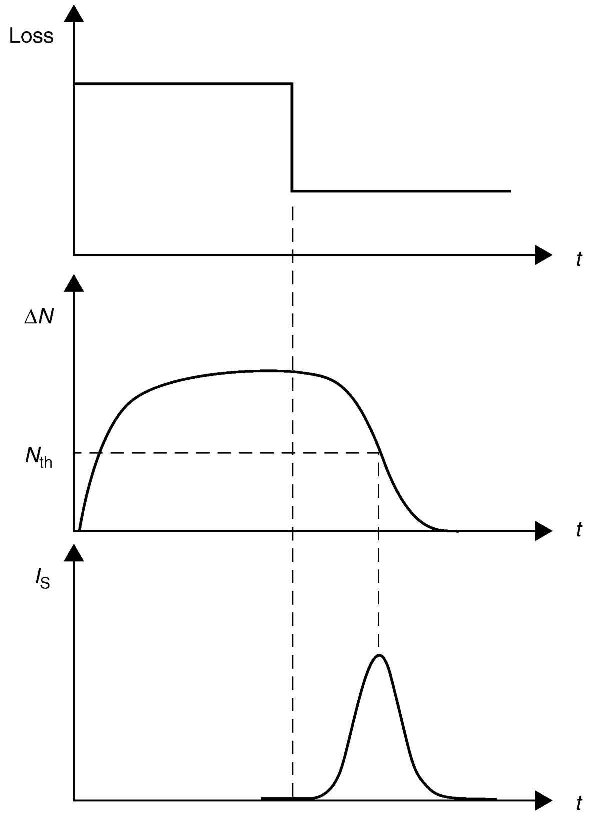

Figure 5 Pulse generation in a Q-switched laser. Initially the intracavity loss is high, the laser remains below threshold, and the inversion of the gain medium increases. When the intracavity loss is suddenly reduced, the laser goes well above threshold and spontaneous emission noise is rapidly amplified into a large pulse which extracts much of the energy previously stored in the gain medium.

An upper limit on the pulse energy obtainable from a Q-switched laser may be determined by assuming all active ions in the laser cavity are excited before the pulse is generated (i.e., at the time the cavity Q is switched high), that no ions remain excited immediately after the pulse is generated, and that the pulse generated is much shorter than the time required to pump all the ions into their excited state. The energy stored in the gain medium, and hence the maximum conceivable pulse energy, would then be $E = h\nu_S NV$, where N is the density of active ions and V is the cavity volume. For example, a typical erbium-doped fiber laser of length 10 m and dopant density $N = 10^{19}$ ions/cm in a single-mode fiber with core diameter 7 μm could ideally store up to 500 μJ of energy.

Two things limit the maximum inversion and energy storage achievable in practice: spontaneous emission noise (due to the finite excited-state lifetime) which is amplified and depletes the inversion, and unwanted feedback (e.g., due to the finite extinction ratio of the Q-switching element) which results in cw lasing which limits the inversion. Following from eqns [1] and [2], the unidirectional gain per unit of stored energy in a fiber laser may be expressed as $G = 4.34/(A_{core}F_{sat}) = 4.34/E_{sat}$ decibels per joule, where $F_{sat} = h\nu_S/\sigma_S$ is the saturation fluence, and $E_{sat} = A_{sat}F_{sat}$ is the saturation energy. For example, if the maximum gain achievable before the onset of cw lasing was 30 dB (i.e., even when the Q-switch introduces a large loss), then the maximum energy storable in the gain medium would be $E = 6.9E_{sat}$. For the same erbium-doped fiber parameters as used in the previous paragraph, the maximum storable energy would be $E \approx 70$ μJ.

Only a proportion of the energy stored in the gain medium may usually be converted into a Q-switched output pulse. The actual Q-switched pulse energy, duration, and peak power can be calculated by solving the rate equations for the photon flux and the population difference between the lasing energy levels in the gain medium. The equations must usually be solved numerically, however assuming the Q changes much more rapidly than the pulse build-up time, which is in turn much shorter than the time taken for a significant change in inversion to either pumping or spontaneous emission, the pulse parameters may be approximated analytically and are found to be determined by only two parameters, i.e., the population inversion (or stored energy) just before Q-switching, and the lifetime of photons within the cavity, τ_C, defined previously.

Under the conditions described above, standard analysis gives the Q-switched pulse duration as

$$T_P \approx \frac{\tau_C(r-1)}{(r-1-\ln r)} \qquad [6]$$

where r is the ratio of the inversion just before Q-switching (low Q) to the threshold inversion for cw lasing after Q-switching (Q high). In fiber lasers the gain can be very high, hence it is common for $r \gg 1$ and almost all available energy stored in the gain medium is extracted as a Q-switched laser pulse. The Q-switched pulse duration is then determined primarily by the cavity lifetime; shorter pulses can be obtained only by shortening the laser cavity and/or increasing the cavity loss. For example, for a 10 m long ring cavity in which the sum of cavity and outcoupling losses is 10 dB and $r \gg 1$, the pulse duration would be $T_P\tau_C = 22$ ns.

The exact proportion of energy stored in the gain medium that is converted to a Q-switched pulse depends on the number of energy levels in the lasing medium, the degeneracy of those levels, and the rate of nonradiative energy transfer between those levels. For example, for an ideal three-level gain medium (e.g., erbium dopant) with no internal cavity losses the maximum available *Q*-switched pulse energy is half the stored energy, or $E_{\text{pulse}} = E/2$. It can be shown that the peak of the output pulse power occurs as the inversion decreases through threshold, and may be approximated by dividing the pulse energy by the pulse duration. Even for moderate pulse energies the peak power is potentially very high. For the erbium-doped fiber laser parameters used previously, the peak pulse power would be approximately 1600 W, i.e., sufficient to cause significant spectral broadening of the output wavelength due to stimulated Raman scattering and other nonlinear effects within the optical fiber.

Mode-locked lasers generate optical pulses by forcing many cavity modes to be locked in phase. As per Fourier theory, the pulse repetition rate is related to the spacing of the lasing modes, and the more modes that add coherently, the greater the pulse bandwidth and the shorter the transform-limited pulse duration. For example, the relationship between mode spacing and pulse repetition rate, and between the spectral and temporal envelopes of a train of Gaussian pulses is shown in Figure 6.

In mode-locked lasers a nonlinear or time-varying element must be present which synchronously modulates the optical signal circulating in the cavity in amplitude or phase at a frequency which is a harmonic of the cavity mode spacing. There are two main methods of mode locking:

(i) active mode locking (e.g., by an electro-optic modulator), in which amplitude or phase modulation is driven by an external source; and
(ii) passive mode locking (e.g., by a fast or slow saturable absorber), in which an element with nonlinear gain or loss responds to the optical signal in the laser cavity itself.

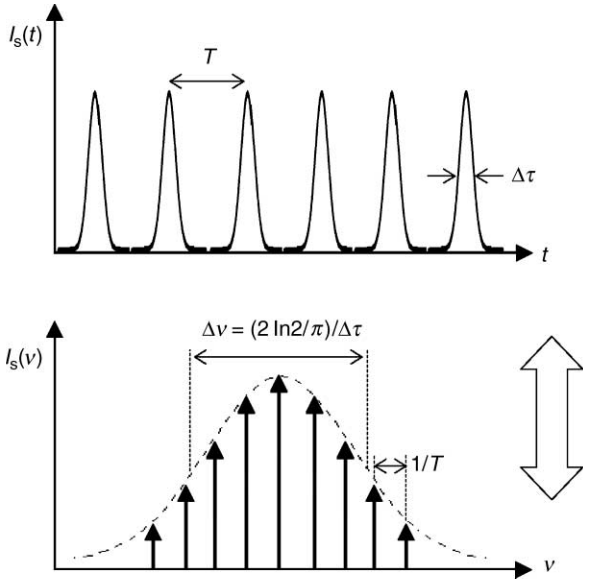

Figure 6 Schematic showing the Fourier-transform relationship between an infinite train of Gaussian pulses and the lasing modes in a mode-locked laser. The temporal and spectral widths are measured full width at half the maximum intensity.

In both cases the modulation provided by the mode-locking element distributes energy between neighboring modes, effectively injection-locking them in phase. A variation of mode locking techniques often present in fiber lasers is soliton mode-locking, in which the active or passive modulation is weak, and a balance between nonlinearity and dispersion is the dominant process underlying pulse formation. Hybrid mode-locking techniques can also be used to combine the best aspects of different pulse-generation methods.

Whilst the latter frequency-domain view of mode locking is helpful, nonlinear effects are almost always significant in mode-locked fiber lasers, and hence the linear concept of a mode is not appropriate for detailed analysis. Hence it is usually preferable to describe pulse generation in mode-locked fiber lasers in the time domain; for stable operation the pulse at any point in the cavity must self-replicate after each round-trip, with the net effects of temporal and spectral loss, gain, dispersion and nonlinearity all canceling.

An optical pulse propagating in a fiber laser is usually described by a slowly varying complex envelope, $\Psi(z,t)$, where $|\Psi(z,t)|^2$ is the instantaneous power in a frame of reference moving at the pulse group velocity, and $\arg[\Psi(z,t)]$ is the instantaneous phase relative to the optical carrier. The various elements in the laser cavity, such as the passive fiber, optical fiber amplifier, filters, and mode-locking devices, each modify the complex pulse envelope as it propagates around the laser cavity.

For example, in the absence of other effects, group velocity dispersion (GVD) causes different wavelengths to travel at different velocities along the fiber, resulting in 'chirping' and temporal broadening of optical pulses by an amount proportional to the distance propagated. The effect can be described mathematically as the accumulation of a phase shift which is quadratic in frequency, i.e., $\Psi(z,\omega) = \Psi(0,\omega)\exp[i\beta_2\omega^2 z/2]$. Taking the derivative with respect to propagation distance, and using $\partial/\partial t = -i\omega$, the incremental change in the complex envelope with propagation distance is

$$\frac{\partial \Psi}{\partial z} = -\mathrm{i}\frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} \quad [7]$$

If the GVD parameter, β_2, is positive then long wavelengths propagate faster than shorter wavelengths, and the dispersion is called 'normal'. If the reverse is true, the dispersion is 'anomalous'. In standard single-mode step index optical fibers the dispersion passes through zero at wavelengths around 1300 nm, and is anomalous for longer wavelengths. It is often necessary to control the net dispersion in a fiber laser, and this can be done by incorporating specially designed fibers with wavelength-shifted or otherwise modified dispersion characteristics, or by using chirped Bragg grating reflectors.

Whilst silica is not commonly regarded as a nonlinear medium (i.e., in which the optical properties are a function of the intensity of light), in ultrashort pulsed fiber lasers the peak powers are usually such that nonlinear effects cannot be ignored. In silica the main nonlinearity is associated with the Kerr effect, and is readily observed as an intensity-dependent refractive index, i.e. $n = n_0 + n_2 I$, where the nonlinear index, n_2, is typically about 3×10^{-20} m^2/W in silica fibers. Other nonlinear effects related to the Kerr nonlinearity which can sometimes be important are cross-phase-modulation and four-wave mixing. In the absence of other effects, the nonlinear index results in self-phase modulation (SPM), or a phase shift in the optical carrier which is proportional to the instantaneous intensity of the signal. Mathematically, $\Psi(z,t) = \Psi(0,t)\exp[\mathrm{i}\gamma|\Psi(0,t)|^2 z]$, in which $\gamma = n_2\omega_0/(cA)$ is the nonlinear coefficient of the fiber, ω_0 is the radian frequency of the optical carrier, and A is the effective core area for the guided mode. The incremental change per unit length in the complex pulse envelope induced by self-phase modulation is then

$$\frac{\partial \Psi}{\partial z} = \mathrm{i}\gamma|\Psi|^2\Psi \quad [8]$$

For example, self-phase-modulation is used to realize a nonlinear switching element in the passively mode-locked laser cavity shown schematically in Figure 7. For obvious reasons the laser is called a 'figure-8' laser, though it is actually a ring laser into which a device known as a nonlinear amplifying loop-mirror (NALM) has been incorporated. The NALM acts as a fast saturable absorber, with transmittance dependent on the instantaneous optical intensity. The NALM operates as follows: light propagating through the optical isolator is divided equally into co- and counter-propagating waves in the loop-mirror on the right-hand side. The light propagating in a clockwise direction is amplified immediately, and propagates around the loop with high intensity, whilst the counter-clockwise propagating signal has a low intensity for most of its time in the loop. The difference in intensity means that the co- and counter-propagating waves experience different nonlinear phase shifts. When they are recombined at the coupler, high-intensity light is transmitted back into the ring, whilst low-intensity light is reflected back to the isolator, and lost. In this way the NALM provides both gain and a nonlinear transmittance which depends on the instantaneous power of the input, described by

$$T \approx G\left\{1 - \frac{1}{2}\left[1 + \cos\left(\frac{1}{2}\gamma L P_{\mathrm{in}} G\right)\right]\right\} \quad [9]$$

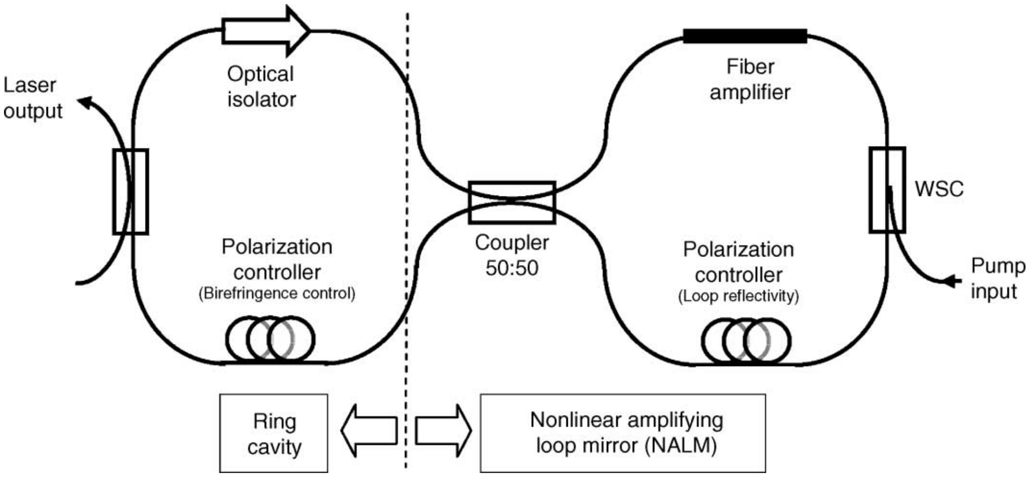

Figure 7 Schematic of a 'figure-8' fiber laser. The laser is passively mode-locked by the nonlinear amplifying loop mirror on the right, which acts as a fast saturable absorber, transmitting and amplifying high-power optical signals whilst reflecting low optical power signals.

in which G is the linear gain of the amplifier, and L the loop length, P_{in} is the input power, and T the ratio of transmitted power to input power. The nonlinear transmittance of the NALM promotes pulse formation and shortening in the laser cavity, and hence passively mode-locks the laser. The polarization controllers are to compensate for unwanted birefringence in the optical fiber, and to control the transmittance of the loop-mirror at low input powers. The fiber dispersion and nonlinearity, together with other factors, further influence the pulses generated in the laser, as described in the following paragraphs.

The combined effects of dispersion and self-phase-modulation can give rise to some interesting and useful physical phenomena. In lossless optical fibers, or fiber lasers in which loss is balanced by gain, pulse propagation is described by the nonlinear Schrödinger equation (NLSE), i.e., a nonlinear wave equation including the effects of both group velocity dispersion and intensity-dependent refractive index. The NLSE may be derived by combining eqns [8] and [9], i.e.

$$\mathrm{i}\frac{\partial \Psi}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 \Psi}{\partial t^2} + \gamma|\Psi|^2\Psi = 0 \qquad [10]$$

It can be shown that if the dispersion is anomalous ($\beta_2 < 0$) then for a particular shape of pulse envelope the distributed effects of dispersion and nonlinearity can exactly cancel, so $\partial\Psi/\partial z = 0$, and the pulse propagates without changing shape. Such pulses are called fundamental solitons, the solution for which is $\Psi(z,t) = \Psi_0 \operatorname{sech}(t/T_0)\exp[\mathrm{i}z/(2L_{\mathrm{D}})]$, in which the peak amplitude is $\Psi_0 = 1/\sqrt{\gamma L_{\mathrm{D}}}$, T_0 is the pulse width parameter, and $L_{\mathrm{D}} = T_0^2/\beta_2$ is known as the dispersion length. For example, in a standard silica fiber at $\lambda = 1.55\ \mu\text{m}$, $\beta_2 \approx -20\ \text{ps}^2/\text{km}$ and $\gamma \approx 10\ \text{W}^{-1}\text{km}^{-1}$, hence a soliton with $T_0 = 1$ ps would have peak power 2 W and energy $E_0 = 2|\beta_2|/(\gamma T_0) = 4$ pJ. The energy of a soliton is quantized, in that pulses with initial energy between 0.5 and 1.5 times E_0 can evolve into a fundamental soliton during propagation. Consequently solitons behave somewhat like particles, and are robust even in the presence of significant perturbations. The pulses which form spontaneously in the figure-8 fiber laser have many of the properties of solitons.

Mode-locked fiber lasers, such as that shown in Figure 7, can often be described by a generalized nonlinear Schrödinger equation, in which extra terms are included in eqn [10] to account for the average effects of an excess linear gain or loss, nonlinear gain and loss (e.g., due to a saturable absorber), spectral filtering, and temporal modulation, etc. Assuming the effects of the added elements can be regarded as distributed over the length of the cavity rather than lumped, such that the envelope of any pulse generated in the laser does not change shape significantly as it propagates around the cavity, then the master differential equation describing propagation of an optical signal within a fiber laser which is passively mode-locked by a fast saturable absorber can be written

$$\begin{aligned}\mathrm{i}\frac{\partial U}{\partial \zeta} + \left(\frac{D}{2} - \mathrm{i}\beta\right)\frac{\partial^2 U}{\partial \tau^2} + (1 - \mathrm{i}\varepsilon)|U|^2 U \\ = \mathrm{i}\delta U - (\nu - \mathrm{i}\mu)|U|^4 U \qquad [11]\end{aligned}$$

Equation [11] is known as the Ginzburg–Landau equation, which describes propagation of the normalized pulse envelope, $U(\zeta,\tau)$, as a function of delayed time, τ, and normalized distance, ζ, and under the combined influence of group velocity dispersion ($D = \pm 1$ for anomalous or normal group velocity dispersion, respectively), dispersion due to band-limiting by amplification in a homogeneously broadened gain medium ($\beta > 0$), nonlinear refractive index, nonlinear gain or loss (ε), linear gain (δ), and quintic terms, μ and ν, which if less than zero relate to saturation of the nonlinear gain and refractive index, respectively. Note that if the right-hand side of the equation is set to zero, and $\varepsilon = \beta = 0$, then the equation reduces to the normalized form of the nonlinear Schrödinger equation. Stable pulses generated within the laser are solutions of the above equation with $\partial U/\partial\zeta = 0$, and hence the effect of all terms must balance. Consequently it is possible for stable pulses to be generated even with normal dispersion in the cavity, though the pulses are then generally strongly chirped, and have an envelope which is approximately Gaussian shaped rather than the hyperbolic secant shape expected with anomalous dispersion. Stable soliton-like pulses can be found by solving the Ginzburg–Landau equation for specific ranges of parameters. There are many numerical solutions, and some analytical solutions, but in general the dynamics of pulse formation and propagation in passively mode-locked fiber lasers can be quite complex, and for some sets of parameters remarkable and unexpected numerical solutions can be found. One such example which has been observed experimentally is the 'exploding' soliton, which unlike the usual soliton is periodically unstable, as shown in the plot of pulse evolution in Figure 8.

Other Fiber Lasers

The lasers discussed in the previous sections are indicative of the main types of fiber laser in terms of their output characteristics; however it should be

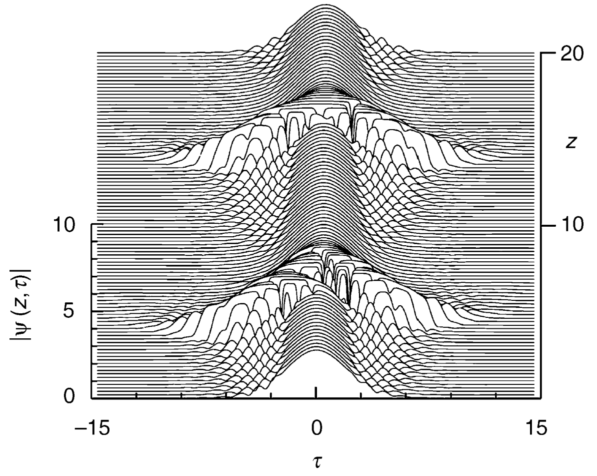

Figure 8 Two periods of the evolution of an exploding soliton found by numerical solution of the quintic Ginzburg–Landau equation with normalized parameters $\varepsilon = 1.0$, $\delta = 0.1$, $\beta = 0.125$, $\mu = 0.1$, and $\nu = 0.6$. The process never repeats itself exactly in successive 'periods'; however the pulse always returns to the same shape. Such solutions cannot be found in analytic form; nevertheless they are as common as stationary pulse solutions and exist for a wide range of parameters.

noted that there is a wide range of other fiber laser configurations with similar behaviors incorporating a variety of other fiber and/or bulk elements. For example, specially designed fibers in which the pump and/or signal has a large cross-sectional area can be used to realize fiber lasers with a high output power, comparable to that in many bulk laser systems. Linear and/or nonlinear processing of the output of fiber lasers can be used to increase their peak output power and/or pulse energy, or to change their output wavelength (i.e., by harmonic generation) from that achievable in the laser cavity itself. In mode-locked fiber lasers it is now common to include passive semiconductor devices which provide a slow saturable absorber action to initiate and stabilize soliton mode-locking.

It should also be noted that there are a number of fiber lasers which generate optical outputs using different or modified physical processes compared to those described in previous sections. For example, it is possible to realize up-conversion fiber lasers in which the output wavelength is shorter than the pump wavelength, multiwavelength fiber lasers (pulsed and cw) which produce outputs at several wavelengths simultaneously, and lasers with limited output coherence such as chaotic fiber lasers and broadband superfluorescent sources, etc. Lastly, whilst the rare-earth doped fiber lasers dealt with in previous sections have some very desirable properties, it should be recognized that optical fiber lasers based on other optical gain mechanisms, such as stimulated scattering and parametric effects, can also be realized (e.g., as in fiber Raman lasers, and fiber Brillouin lasers). In the latter cases the lasing wavelength is limited only by available pump wavelengths. The reader is referred to the references in the Further Reading for information on the wide range of fiber lasers that are possible in practice.

List of Units and Nomenclature

Absorption cross-section	σ_a	$[\mathrm{m}^{-2}]$
Amplifier length	L_A	[m]
Area	A	$[\mathrm{m}^2]$
Cavity length	L_C	[m]
Complex amplitude	Ψ	$[\mathrm{W}^{0.5}]$
Dispersion length	L_D	[m]
Distance	z	[m]
Emission cross-section	σ_e	$[\mathrm{m}^{-2}]$
Frequency	v	[Hz]
Gain coefficient	g	$[\mathrm{m}^{-1}]$
Group velocity dispersion	β_2	$[\mathrm{s}^2\mathrm{m}^{-1}]$
Inversion ratio	r	
Lifetime	τ	[s]
Loss coefficient	α	$[\mathrm{m}^{-1}]$
Nonlinear coefficient	γ	$[\mathrm{m}^{-1}\mathrm{W}^{-1}]$
Nonlinear refractive index	n_2	$[\mathrm{m}^2\mathrm{W}^{-1}]$
Number density	N	$[\mathrm{m}^{-3}]$
Power gain	G	
Pump power	P_P	[W]
Pump wavelength	λ_P	[m]
Reflectivity	R	
Saturation energy	E_{sat}	[J]
Saturation fluence	F_{sat}	$[\mathrm{J.m}^{-2}]$
Signal power	P_S	[W]
Signal wavelength	λ_S	[m]
Threshold power	P_{th}	[W]

See also

Fiber and Guided Wave Optics: Dispersion; Fabrication of Optical Fiber; Light Propagation; Nonlinear Effects (Basics); Optical Fiber Cables. **Fiber Gratings**. **Lasers:** Up Conversion Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Ultra-Fast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Agrawal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Akhmediev NN and Ankiewicz A (1997) *Solitons*. London: Chapman and Hall.

Becker PC, Olsson NA and Simpson JR (1999) *Erbium-Doped Fiber Amplifiers: Fundamentals and Technology*. San Diego, CA: Academic Press.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Digonnet MJF (ed.) (2001) *Rare-Earth-Doped Fiber Lasers and Amplifiers,* 2nd edn. New York: Marcel Dekker.

Duling IN (ed.) (1995) *Compact Sources of Ultrashort Pulses*. Cambridge, UK: Cambridge University Press.

France PW (ed.) (1991) *Optical Fibre Lasers and Amplifiers*. Glasgow, UK: Blackie.

Jeunhomme LB (1990) *Single-Mode Fiber Optics: Principles and Applications*, 2nd edn. New York: Dekker.

Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.

Siegman AE (1986) *Lasers*. Oxford, UK: Oxford University Press.

Weber MJ (ed.) (1995) *CRC Handbook of Laser Science and Technology*. Boca Raton, FL: CRC Press.

Organic Semiconductors and Polymers

G A Turnbull, University of St Andrews, Fife, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Organic semiconductors, and conjugated polymers in particular, are emerging as an interesting new class of visible laser media. The polymers are electroluminescent, broadband visible emitters and exhibit large optical gain coefficients. As plastics, they are amenable to simple processing, and can be shaped into complex structures. This combination of photonic, electronic, and processing properties make these materials extremely well suited to forming novel microlasers, with future prospects as inexpensive, plastic diode lasers operating throughout the visible spectrum.

Conjugated Polymers

Plastics are widely used in low-cost optical components such as lenses, diffraction gratings, optical fibers, and adhesives. Their common use derives from a combination of qualities: good optical transmission, a wide range of mechanical properties and simple fabrication via molding: extrusion or solution processing. Functional optical polymers, including liquid crystals and photorefractives, have also been widely applied to polarization control and for nonlinear- and electro-optics. Conjugated polymers are a remarkable class of functional plastics that exhibit some unusual properties. Firstly, they are plastics that can conduct electricity; and in fact are electrically semiconducting. Secondly, they are plastics that can be stimulated either optically or electrically to emit visible light.

These two unusual properties derive from a particular chemical structure specific to this family of plastics. In all conjugated polymers, the backbone of the molecule is a chain of carbon atoms joined by alternating single and double covalent bonds. The simplest example is illustrated in Figure 1a, which shows the chemical structure of the polymer poly(acetylene). Structural integrity of the polymer is provided by a string of single σ-bonds between adjacent carbon atoms. The additional bonding electrons in the double bonds are located in π-orbitals oriented perpendicular to the polymer chain. Adjacent π-orbitals overlap, and their electrons can be widely delocalized along the polymer backbone. The delocalized electrons can move relatively freely along the chain, giving the electronic properties of the material. The π-orbitals can be transformed between bonding and anti-bonding states by the absorption or emission of light, leading to strong optical transitions. While poly(acetylene) is a very inefficient light emitter, other conjugated polymers, notably those that include phenyl rings in the backbone (Figure 1b–d), can be highly emissive. Photoluminescence quantum

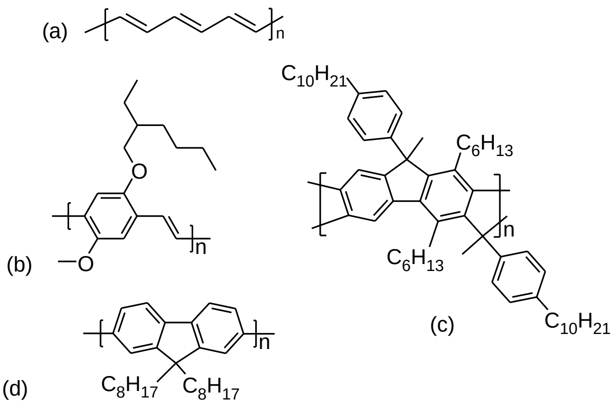

Figure 1 Chemical structures of several conjugated polymers. (a) poly[acetylene]; (b) poly[2-methoxy-5-(2′-ethylhexyloxy)-1,4-phenylene vinylene], (MEH-PPV); (c) ladder-type poly[1,4-phenylene], (MeLPPP); (d) poly[9,9-dioctylfluorene], (PFO).

efficiencies, approaching unity in solution and as high as 60% in the semiconducting solid state, have been realized.

Organic Semiconductor Lasers

While organic dye lasers have been widely used since the mid-1960s, the history of organic semiconductor lasers is much more recent. Other than a few early observations of lasing in molecular crystals in the 1970s by Karl and others, the field of organic semiconductor lasers began in earnest following the related discovery in 1990 by Burroughes *et al.* of electroluminescence in a conjugated polymer. In 1992, Moses demonstrated lasing for the first time in a solution of the conjugated polymer MEH-PPV (Figure 1b). However, it was another four years – once material synthesis had been sufficiently refined – before stimulated emission and then true laser action were observed in polymers in the semiconducting solid state. In the following year the first demonstrations were made of dye-doped small molecular semiconductor lasers. Since that time, there has been a great growth in interest in the field with an average of ~30 research papers per year. Further details of the development of the field of organic semiconductor lasers can be found in the Further Reading.

So what are the particular properties of conjugated polymers that have stimulated interest in developing organic semiconductor lasers? Firstly these materials are broadband visible emitters. Through appropriate chemical design, their emission can be tuned throughout the visible spectrum, from 400 nm to 700 nm (Figure 2). The polymers exhibit large optical gain coefficients, because the optical transitions are dipole-allowed. Compared with small molecular dyes, polymers exhibit a relatively low degree of self-absorption of the emitted light, and much reduced quenching of the luminescence with increased concentration. This means that it is possible to work with much higher chromophore densities than conventional laser dyes, which is advantageous for achieving very low threshold lasing in extremely compact resonators. Furthermore, because these materials are semiconducting, they have the potential for direct electrical excitation. This could lead to semiconductor diode lasers operating throughout the visible spectrum and, in particular, in the blue and green spectral bands where currently there are few commercially viable inorganic semiconductor sources. Finally, these materials are plastics and so are amenable to simple processing, and can be shaped and structured into complex feedback resonators.

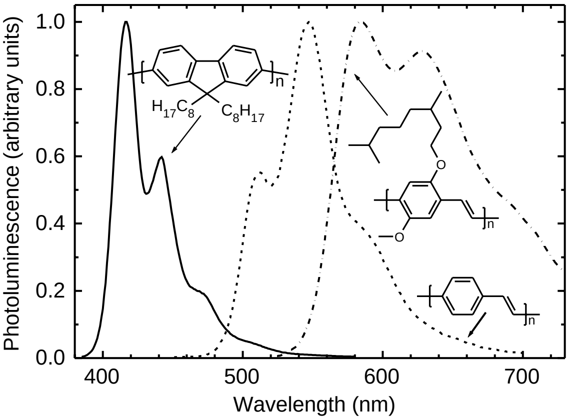

Figure 2 Photoluminescence spectra of three commonly used conjugated polymers; chemical structures inset.

Organic Semiconductor Gain Materials

A great advantage of organic laser media is that a very wide range of materials can be readily synthesized, and through minor changes in the molecular structure one can relatively simply tune the emission properties of the material. Many different organic semiconductors have been applied as lasers, though they fall into three major categories. These are conjugated polymers, dye-doped small molecular semiconductors, and molecular single crystals. Conjugated polymers make up the most diverse materials set, most notably including derivatives of poly(paraphenylene-vinylene), poly(paraphenylene), and poly(fluorene) (Figure 1b–d). The second category comprises small molecular semiconductors that are co-evaporated under vacuum with a dopant laser dye. In this blend of materials, the organic semiconductor functions simply as an energy transfer host for the emitting dye molecules. In the third category are organic single crystals, including materials such as polyacenes and thiophene oligomers. These materials have the advantage of having the highest electrical mobilities of any organic semiconductors, which is important for electrical excitation. However, they can be demanding to grow into high-quality single crystals, and so lack the key processing advantages of the other families of organic semiconductors.

Figure 3a shows a typical exciton energy level structure of a semiconducting polymer. Similar to a conventional laser dye, the energy levels comprise a manifold of singlet states and a manifold of triplet states, each with vibrational sublevels. The laser transition occurs from the bottom of the first excited singlet state S_1 to the vibronic overtones of the ground

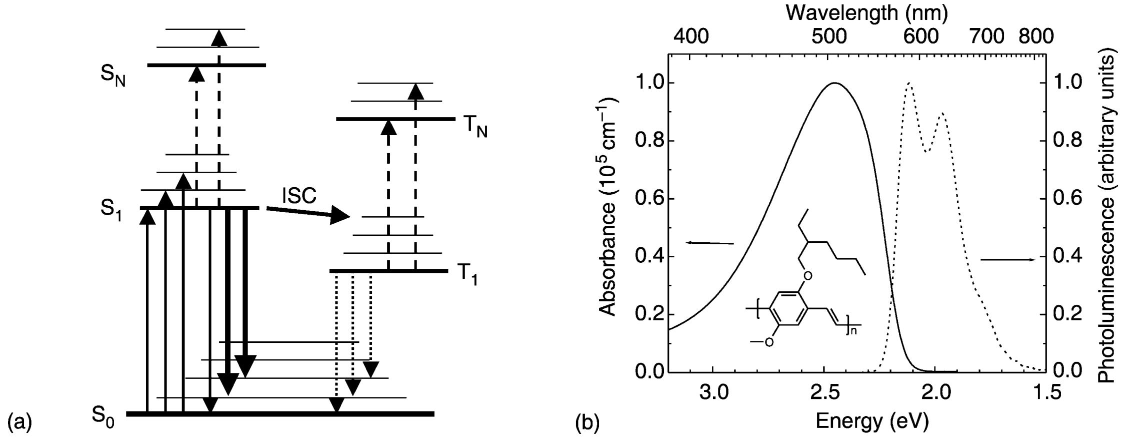

Figure 3 (a) Generic energy level structure and optical transitions of a conjugated polymer. S_N refers to the *N*th singlet state; T_N refers to the *N*th triplet state; ISC: intersystem crossing. (b) Absorption and photoluminescence spectra of the polymer MEH-PPV.

state S_0, so forming a classic 4-level laser system. The optical transition is dipole-allowed and so has a high oscillator strength. As a result, conjugated polymers typically exhibit simulated emission cross-sections in the range 10^{-16} cm^2 to 10^{-15} cm^2, while the population inversion lifetime is as short as a few hundred picoseconds.

There are, however, several other transitions that can compete with the stimulated emission process. There can be excited-state absorption from the S_1 singlet state into higher-lying singlet states. Additionally intersystem crossing from the S_1 state to the lowest excited triplet state T_1 can subsequently lead to excited triplet absorption. Other processes in the polymer can also impede the stimulated emission. Firstly, charged excitations can exhibit absorption bands that overlap with the emission. Secondly, at high excitation densities, exciton–exciton annihilation can significantly deplete the excited states available for stimulated emission. Ultimately, singlet–singlet annihilation is the limiting factor for the maximum excitation density possible, and while polymer films can have a repeat-unit density of $\sim 10^{21}$ cm^{-3}, the excited state density is typically limited to around 10^{18} cm^{-3}. In many materials this is not restrictive on achieving optical gain because stimulated emission can occur at densities of 10^{17} cm^{-3} and below.

An important factor that distinguishes the photophysics of semiconducting polymers from conventional laser dyes is an internal process of exciton migration between the absorption and emission of a photon. Excitations are confined to short straight sections of the polymer between kinks or twists in the chain. These straight sections, or 'sites', are typically a few repeat units in length, though there is a statistical spread of site lengths in any given material. When formed, excitons tend to hop to the longest, lowest-energy, sites before emitting a photon. The exciton migration process has two effects. First, it means the emission takes place from a subset of the chain segments, resulting in a lower degree of inhomogeneous broadening of the emission spectrum (vibronic peaks are better resolved than in absorption (see Figure 3b)). Second, the peaks of the absorption and emission spectra tend to be widely separated in wavelength, reducing the level of self-absorption at the lasing wavelength.

As mentioned above, blending materials is a particularly attractive approach towards further reducing self-absorption in organic semiconductor lasers. In such systems, a narrow bandgap dopant chromophore is blended with a concentration of 1–2% in a wider bandgap semiconducting host. Emission from the host is strongly quenched via nonradiative Förster energy transfer to the guest molecule. The blended material exhibits an absorption profile that is dominated by the host absorption, while the emission profile is almost completely that of the guest dopant. By strongly separating the emission peak from the absorption edge of the host, it is possible to achieve a substantially reduced (by perhaps an order of magnitude or more) residual absorption coefficient at the lasing wavelength. In doing so, the combined material acts much more like an ideal 4-level laser medium and can thereby have substantially lower pump thresholds for stimulated emission.

Measuring Gain

The dynamics of the excited state population in semiconducting polymers have been widely studied using femtosecond transient absorption experiments.

This is a pump–probe technique in which the transmission of the broadband probe pulse is compared in the presence and absence of a pump pulse that excites molecules into the first excited state. Such measurements have the time resolution of the pump pulses used and so can typically be ~100 fs. By systematically delaying the arrival of the broadband probe pulse, it is possible to map out the evolution of the gain dynamics across the emission spectrum of the material with very high time resolution.

Figure 4 shows a series of snapshots of the absorption induced by a pumping pulse in a poly(thiophene) derivative. This figure shows the changing absorption of the sample between 200 fs and 400 ps after excitation. Within the absorption band of the material, one finds that the level of absorption is slightly reduced because the pump pulse has partly depleted the ground state. In the emission band there is also a reduction in the absorption that mimics in shape the photoluminescence spectrum of the polymer. This negative absorption – or optical gain – shows direct evidence for stimulated emission through much of the emission band. In the low energy tail of the emission, however, there is an increase in absorption that arises due to the excited-state absorption process mentioned earlier. Clearly, for a good laser material, the spectra of excited-state and charge-induced absorptions should overlap as little as possible with the photoluminescence.

Many other investigations of gain in semiconducting polymer films have used a phenomenon known as spectral line narrowing to explore indirectly the amplifying properties of the medium. The experimental procedure for such studies is as follows. A thin film of the semiconducting material, of typically 100 nm thickness, is excited by a pulsed pump laser focused to a stripe of dimensions ~100 μm wide by a few millimeters in length. A significant fraction of the light emitted by the material is trapped in the film by total internal reflection at the polymer–air and polymer–substrate interfaces, and is thereby waveguided along the length of the excitation stripe. This waveguided spontaneous emission can be amplified by stimulated emission before being emitted out the edge of the film.

One finds that, above a particular threshold of pump energy, there is a superlinear increase in the output energy with pump intensity. Associated with this is a dramatic narrowing of the broad emission spectrum to a linewidth of typically less than 10 nm, as illustrated in Figure 5. These two features are signatures of gain narrowing via the process of amplified spontaneous emission. The line-narrowed feature usually appears at the first vibronic overtone of the optical transition, where the gain exceeds the losses in the material by the greatest amount. This particular wavelength experiences the greatest amplification and, above a particular pumping density, undergoes a runaway effect in which most of the light is stimulated to emit at that wavelength at the expense of the rest of the spectrum.

In the case of amplified spontaneous emission or mirrorless lasing as it is also known, the spontaneous emission guided along the length of the stripe acts as

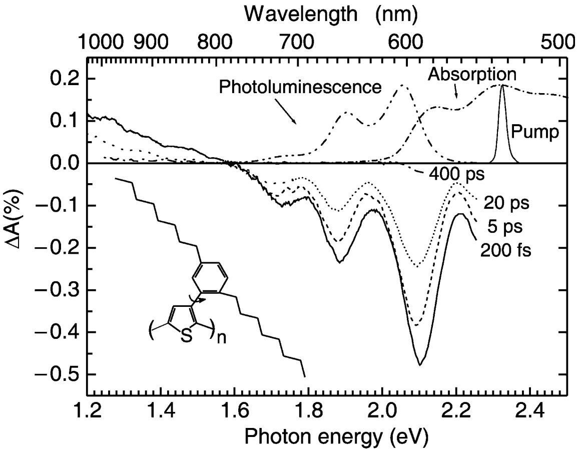

Figure 4 Transient absorption spectra of poly[3-(2,5-dioctylphenyl)-thiophene]; chemical structure inset. (Reproduced from Ruseckas A, Theander M, Valkunas L, Andersson MR, Ingänas O and Sundström V (1998) Energy transfer in a conjugated polymer with reduced inter-chain coupling. *Journal of Luminescence* 76 & 77: 474–477. Copyright (1998), with permission from Elsevier.)

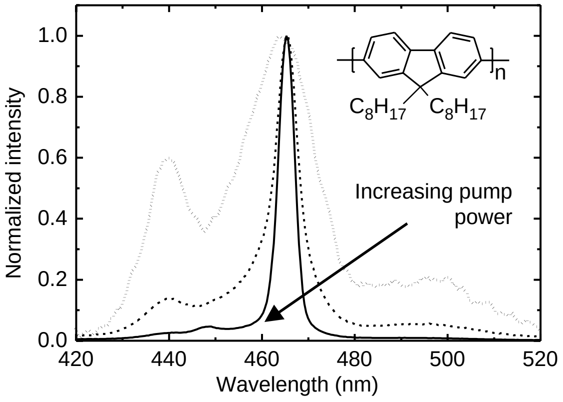

Figure 5 The change in spectral shape with increasing excitation density (from 10 to 133 μJ cm^{-2}) of the edge emission from a PFO film – the normalized spectra show a significant spectral line narrowing at high excitation densities. (Adapted from Heliotis G, Bradley DDC, Turnbull GA and Samuel IDW (2002) Light amplification and gain in polyfluorene waveguides. *Applied Physics Letters* 81: 415–417. Copyright (2002), with permission from American Institute of Physics.)

the 'probe pulse'. In these experiments it is therefore necessary to extract information indirectly about the net gain in the material. By integrating the spontaneous and stimulated emission along the length of the stripe, one finds that the wavelength-dependent output intensity $I(\lambda)$ of the spectral line narrowed light is given by the relationship

$$I(\lambda) \propto \frac{1}{g(\lambda)}[\exp[g(\lambda)l] - 1] \qquad [1]$$

where I_p is pumping intensity, l is the length of the stripe, and $g(\lambda)$ is the net gain coefficient. Therefore, by monitoring the intensity of the line-narrowed emission as a function of the length of the pump stripe, one can extract the value of the net gain for the material. Similarly, by progressively moving the stripe further and further away from the edge of the film, it is possible to evaluate the waveguide losses of light propagating through an unpumped region of the waveguide. Waveguide losses in conjugated polymers typically lie in the range of 3–50 cm^{-1}. Net gains can be as high as 60 cm^{-1} or 260 dB cm^{-1} at modest pumping densities of 4 kW cm^{-2}. These substantial gains, in very small propagation lengths, therefore highlight the attraction of waveguide-based resonators for polymer lasers. In the next section we will discuss the most commonly used feedback structures that have been used for polymer lasers.

Polymer Laser Resonators

The first demonstration of cavity lasing in a conjugated polymer was made by Moses in 1992. The resonator consisted simply of two flat dielectric mirrors on either side of a cuvette containing the polymer in solution. Lasing occurred at the peak wavelength of the gain profile. Subsequent developments in solid-state polymer lasers have involved more sophisticated resonators that exploit the capacity of these materials for very simple processing. For example, solution-processing methods such as spin-casting or dip-coating, make it possible to fabricate very high-quality optical waveguides. Wavelength-scale microstructures can also be readily formed, allowing the straightforward fabrication of laser microresonators.

Many different resonators have been explored for solid-state polymer lasers. Some of the key types are shown in Figure 6, including different geometries of both microcavities and distributed feedback structures. The first resonator used, by Tessler *et al.*, was a planar microcavity (Figure 6a), in which the polymer film was sandwiched between two mirrors. The structure supported standing-wave resonances perpendicular to the plane of the film, at three wavelengths in the emission band of the polymer. Above laser threshold, light was preferentially stimulated into only one of these modes. Microcavities of other geometries have also been studied, including annular and spherical microresonators (Figure 6b,c), in which the optical field is confined in waveguide modes, whispering gallery modes, or a superposition of the two. Such structures provide a longer interaction length in the gain medium for stimulated emission; though tend to lase on many longitudinal modes and have no directional output beam.

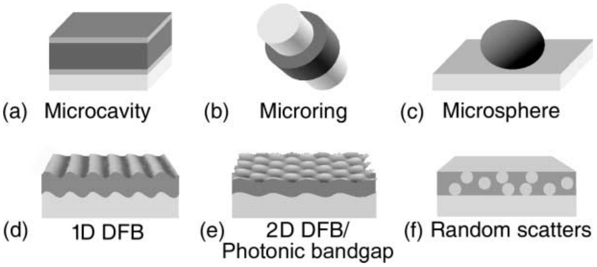

Figure 6 Resonators used for conjugated polymer lasers. (a) Planar microcavity; (b) annular microcavity; (c) spherical microcavity; (d) distributed feedback resonator; (e) 2D DFB/ photonic crystal resonator; (f) random laser.

The final key group of resonators are distributed feedback (DFB) structures, which researchers have progressively favored to attain the low pump threshold densities that will be required to realize electrically pumped lasing. Figure 6d shows a typical DFB laser structure. The device is based on a thin polymer film that is typically 100 nm in thickness deposited on top of a silica substrate. The silica–polymer–air structure forms an asymmetric slab waveguide that supports only the lowest order transverse electric mode TE_0. To provide the feedback, the substrate is modulated with a wavelength-scale periodic grating structure. These corrugations give rise to Bragg scattering of the waveguide mode. Through careful selection of the grating period, one may arrange that optical wavelengths close to the peak of the polymer gain will be Bragg scattered from the propagating TE_0 mode to the counter-propagating TE_0 mode. This is achieved when the Bragg equation

$$2n_{\text{eff}}\Lambda = m\lambda_{\text{B}} \qquad [2]$$

is satisfied, where Λ is the grating period, n_{eff} is the effective refractive index of the waveguide, m is an integer, and λ_{B} is the Bragg wavelength. In so doing, one may arrange that two counter-propagating waveguide modes can be coupled together, thereby providing the distributed resonant feedback for the laser process. Commonly in semiconducting polymer lasers, a second-order (i.e., $m = 2$) Bragg grating has been favored to provide the feedback for the laser, so that the first-order scattering provides a surface-emitted output coupling, perpendicular to the plane of the guide. This surface-emitting geometry is particularly attractive in polymer lasers because it is relatively difficult to achieve a high optical quality edge to the waveguide. DFB lasers have the advantage over simple microcavity resonators in that they can combine long interaction lengths, limited by the scattering length of the Bragg grating, with excellent wavelength selection, which is controlled by the grating period.

In addition to conventional 1D DFB gratings, a number of more exotic 2D and 3D DFB, and photonic crystal resonators have been explored in semiconducting polymers. Even randomly located scattering sites in a polymer film have been used to provide a weak closed-loop cavity feedback. Examples of 2D gratings are given in Figure 7, which shows atomic force micrographs of an eggbox structure, formed from two orthogonal gratings, and a circular DFB grating. Such 2D DFB structures can significantly improve the operation of surface-emitting polymer lasers compared with 1D gratings. Improved feedback yields lower oscillation thresholds, combined with higher slope efficiencies and significantly improved beam quality. Indeed, such surface emitting DFB lasers can produce nearly diffraction limited laser beams.

Typical operating characteristics of an egg-box DFB polymer laser are shown in Figure 8. This laser, which is based on the material MEH-PPV, is transversely pumped with a pulsed pump laser focused onto a small region of the waveguide; up to 90% of the incident pump light is absorbed in the 100 nm thick film. The laser has a threshold pump energy of 4 nJ and an external slope efficiency of >7%, and operates on a single frequency close to the Bragg wavelength of the periodic waveguide. Lasing does not occur exactly at the Bragg wavelength, because at this wavelength the propagation of light is suppressed by the periodic structure. This leads to a photonic stopband within which the emission is reduced. At the edges of this stopband the microstructure can support standing wave electric field patterns of the same spatial period. At these wavelengths there can be a strong interaction between the electric field and excited states in the gain medium, which leads to low-threshold band edge lasing as observed in the figure.

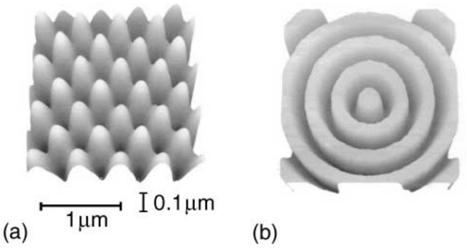

Figure 7 Atomic force microscope images of (a) 'egg-box' DFB grating, and (b) circular DFB grating; grating period is 400 nm in each structure.

Although DFB resonators have advantages over microcavity devices, the complicated lithographic and etching steps involved in their fabrication can detract from the key material advantage of simple processing. To address this issue, several different techniques for simply replicating the complicated microstructures have been explored. These include UV embossing of flexible plastic substrates, and hot embossing and micromolding of the active

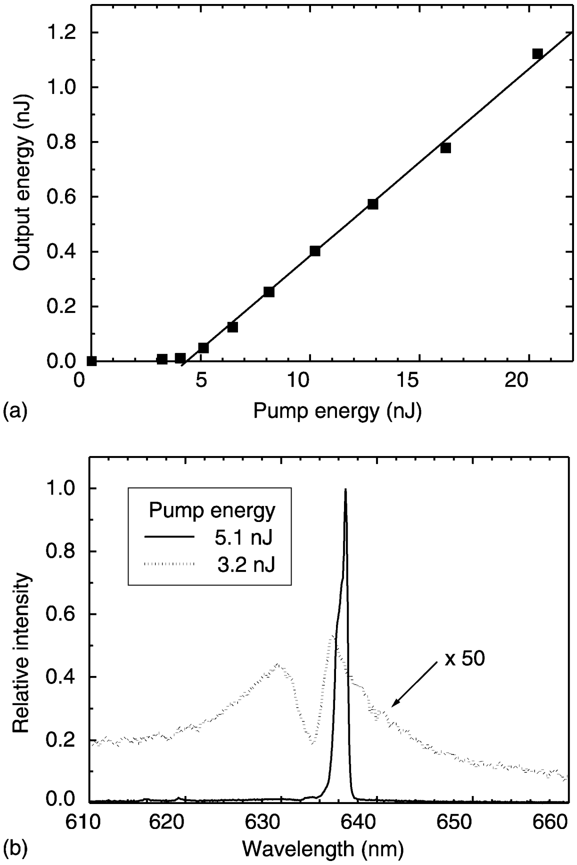

Figure 8 (a) Energy characteristic of an egg-box DFB laser; (b) normalized emission spectra above and below lasing threshold, showing lasing in the edge of the photonic stopband. (Reprinted from Turnbull GA, Andrew P, Barnes WL and Samuel IDW (2002) Operating characteristics of a semiconducting polymer laser pumped by a microchip laser. *Applied Physics Letters* 82: 313–315. Copyright (2003). With permission from American Institute of Physics.)

polymer itself. Such techniques can very readily reproduce structures with feature sizes as small as 100 nm and could allow future mass production of photonic microstructures in these materials.

Towards Plastic Diode Lasers

A typical experimental configuration for characterising optically pumped polymer lasers is shown in Figure 9. The laser is mounted in a vacuum chamber in order to isolate the polymer from oxygen and water. A pulsed visible or ultraviolet pump laser is attenuated and then focused onto the waveguide to transversely pump the polymer laser. The emission is collected using a CCD spectrograph for spectral measurements and energy meter and beam profilers for other characteristics. Typical pulsed pump lasers, which have been used, include nitrogen laser pumped dye lasers, and spectral harmonics of regeneratively amplified Ti^{3+}-sapphire lasers or flashlamp pumped Nd^{3+}:YAG lasers. Recent advances have shown the thresholds to be low enough to be successfully pumped with a modest Nd^{3+} pulsed microchip laser.

So while optically pumped polymer lasers are now proving to be compact, functional light sources, a major motivation for research in this field remains the prospect of electrically driven plastic diode lasers. Such lasers could be very low-cost, flexible light sources that would access any wavelength throughout the visible spectrum. In particular, there are exciting prospects of producing blue and green diode lasers in spectral regions currently difficult to access with inorganic semiconductors. The lowest threshold densities reported for optically pumped organic semiconductor lasers are of the order of 100 W cm^{-2}. To achieve similar excitation densities electrically, one requires pump current densities of ~200 A cm^{-2}. This is substantially more than the few mA cm^{-2} commonly used in polymer LEDs for display applications. Presently, such high current densities are infeasible with continuous excitation, but can be exceeded with short current pulses of a few tens of nanoseconds duration.

Electroluminescence in conjugated polymers was first achieved in 1990 by Burroughes and co-workers. Since that time there has been remarkable progress made in the science and technology of organic LEDs. Commercial products are now on the market, and the feasibility of both flexible and full-color light-emitting displays, in which the individual pixels have been ink-jet printed, have been demonstrated in the laboratory. Despite such rapid progress in the field of organic LEDs, there have been no successful demonstrations to date of electrically pumped lasing in any organic semiconductor. In this section we highlight some of the challenges that must be faced if electrically pumped lasing is to be achieved. To understand the problems associated with electrical driving of organic semiconductor lasers, it is useful first to review the generic structure of organic light-emitting diodes.

A generic organic LED structure is shown in Figure 10. The device consists of one or more organic layers sandwiched between two electrical contacts. Usually a transparent anode is used, commonly made from ITO, which allows light to be emitted from the front face of the LED. A low work-function metal, such as calcium or aluminum, is used as the cathode. The organic part of the device may consist of several layers. These include one or more charge transport layers, plus a light-emitting layer in which the electrons and holes combine to form excitons before emitting light. Typically, the combined thickness of these organic layers is as little as 100 nm. Such thin devices are necessary to keep operating voltages low, to overcome potential barriers for injection, and because charge mobilities in organic semiconductors are very low (10^{-5} to 10^{-2} cm^2 V^{-1} s^{-1}).

The inclusion of electrical contacts into organic semiconductor laser structures can be problematic since they need to be situated very close to the emitting region. Metallic contacts can quench the luminescence and introduce losses to the waveguide mode, thereby increasing the oscillation thresholds. Furthermore, the ITO anode has a high refractive index and so can reduce the spatial overlap of the

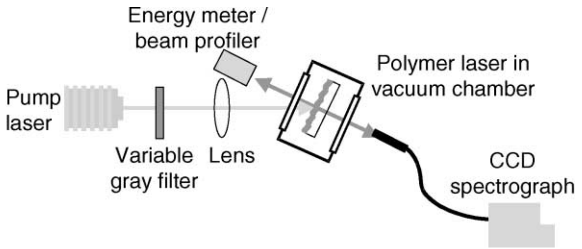

Figure 9 Schematic of a typical experimental configuration for characterizing polymer lasers.

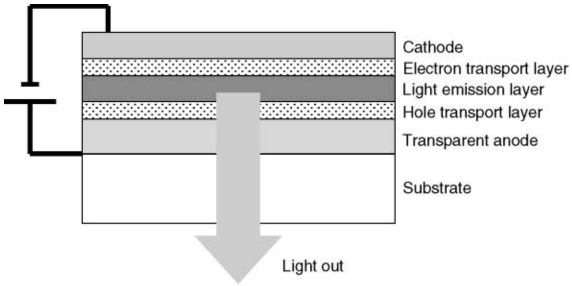

Figure 10 Schematic structure of a generic organic light-emitting diode.

guided mode with the active medium. Such issues indicate that the design of organic LEDs needs to be somewhat modified for achieving electrically pumped lasing. In particular, there needs to be a trade-off between the optimum electronic and photonic designs. To address these issues, various device structures have been proposed that distance the optical mode from metallic contacts, while reducing the thickness of the ITO layer to optimize the mode confinement in the active material.

A more demanding obstacle to electrically pumped lasing is the additional excited state absorption associated with the electrical pumping scheme. As many as three-quarters of the generated excitons in an organic LED are formed in the triplet state which is nonemissive and can lead to excited-state absorption, albeit principally at lower energies than the lasing wavelength. Additionally, injected charges that have not yet formed excitons, exhibit strong absorption, which overlaps significantly with the peak of the material gain. Indeed, this charge-induced absorption has so far prevented the demonstration of injection lasing in conjugated polymers, even in optimized device structures mentioned above. A possible alternative route for efficient pumping of polymer lasers could be through indirect electrical pumping. For example, an inorganic diode laser or LED, or even an organic LED, could be used as an electrically efficient optical pump for the semiconducting polymer laser. Recent advances in GaN-based LEDs and lasers provide interesting prospects for such alternative systems.

Research on semiconducting polymer lasers continues, both to reduce the excitation densities needed for threshold and to investigate new material systems that could overcome the obstacle of charge-induced absorption. By building on recent developments in resonator geometries, material blends, and compact pumping schemes, semiconducting polymer laser systems, whether optically or electrically pumped, should soon become practical, visible light sources for a range of applications.

List of Units and Nomenclature

Charge mobility	[$\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$]
Current density	[A cm^{-2}]
Energy density	[$\mu\text{J cm}^{-2}$]
Excited state densities	[cm^{-3}]
Gain coefficient α ($\exp(\alpha z)$)	[cm^{-1}]
Intensity/power density	[kW cm^{-2}]
Optical gain	[dB cm^{-1}]
Stimulated emission cross-section	[cm^2]
Waveguide propagation loss coefficient γ ($\exp(-\gamma z)$)	[cm^{-1}]
Wavelength	[nm]

See also

Diffractive Systems: Diffractive Laser Resonators. **Lasers:** Dye Lasers; Planar Waveguide Lasers. **Optical Amplifiers:** Basic Concepts. **Optical Materials:** Plastics. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Burroughes JH, Bradley DDC, Brown AR, *et al.* (1990) Light-emitting-diodes based on conjugated polymers. *Nature* 347: 539–541.

Friend RH, Gymer RW, Holmes AB, *et al.* (1999) Electroluminescence in conjugated polymers. *Nature* 397: 121–128.

Karl N (1972) Laser emission from an organic molecular crystal. *Physica Status Solidi (a)* 13: 651–655.

Kranzelbinder G and Leising G (2000) Organic solid-state lasers. *Reports on Progress in Physics* 63: 729–762.

McGehee MD and Heeger AJ (2000) Semiconducting (conjugated) polymers as materials for solid-state lasers. *Advanced Materials* 12: 1655–1668.

Miyata S and Nalwa HS (eds) (1997) *Organic Electroluminescent Materials and Devices*. Amsterdam: OPA.

Moses D (1992) High quantum efficiency luminescence from a conducting polymer in solution – a novel polymer laser-dye. *Applied Physics Letters* 60: 3215–3216.

Pope M and Swenberg CE (1999) *Electronic Processes in Organic Crystals and Polymers*. New York: Oxford University Press.

Samuel IDW (2000) Polymer electronics. *Philosophical Transactions of the Royal Society of London A* 358: 193–210.

Scherf U, Riechel S, Lemmer U and Mahrt RF (2001) Conjugated polymers: lasing and stimulated emission. *Current Opinion Solid State and Materials Science* 5: 143–154.

Tessler N (1999) Lasers based on semiconducting organic materials. *Advanced Materials* 11: 363–370.

Tessler N, Denton GJ and Friend RH (1996) Lasing from conjugated-polymer microcavities. *Nature* 382: 695–697.

Planar Waveguide Lasers

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers have evolved to become true marvels of technology. From their invention in the early 1960s, laser light has changed the way society operates, its application stretching to virtually every aspect of human activity. The versatility of lasers is derived from the fact that there are several different types, each with a unique set of advantages. Solid-state lasers, made from specific semiconducting materials, are commonly used in the optical networks that underlie the greatest machine on Earth – the Internet. High power gas lasers, such as CO_2 lasers, are used in industrial machining, whereas rare-earth doped and dye lasers have found applications ranging from eye surgery to tattoo removal.

Planar waveguide lasers are a relatively new class of lasers made possible by advances in the field of optical waveguides. Over the past decade, with the aim to miniaturize optical components, planar waveguide technology has been developed to a high degree of sophistication. Likewise, Er and other rare-earth doping was studied extensively owing to its importance in all-optical amplification. These elements turned out to be the building blocks for the development of fiber lasers and subsequently planar waveguide lasers.

This chapter will discuss the basic concepts underlying this category of lasers, the variations therein, and the main characteristics. It is important to understand the basic features of waveguides and lasers, which is included as a prelude. We will then discuss planar waveguide lasers with emphasis on the essential elements: the laser materials, the various designs, their performance, and the resultant advantages as well as the applications.

Waveguides

A waveguide is a general conduit for efficient propagation of an electromagnetic field. It is usually defined by making a region of higher refractive index within a dielectric medium. This behaves as an attractive potential well that confines light in the form of bound modes, much like a potential well that confines electrons to a bound state. The propagating field is referred to as the lightwave and is governed by Maxwell's equation. For a weakly guiding approximation, as is the case when the refractive index contrast (Δn) is small, Maxwell's equation can be reduced to a scalar wave equation for the transverse electric field. For a given wavelength (λ) and polarization, the scalar wave equation in a nonmagnetic medium is

$$\nabla^2 E(x) + \frac{4\pi^2}{\lambda^2} n(x)^2 E(x) = 0 \qquad [1]$$

where E is the electric field. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$ where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y) = (\bar{n})^2\Phi(x, y) \qquad [2]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$. The effective index dictates the occurrence of bound modes – these states occur when $\bar{n}$ is between n_{core} and n_{cladding}. For smaller values of $\bar{n}$, there is a continuum of unbound and radiation modes. These are leaky modes that are dispersed quickly over a short distance. Only bound modes are guided in the core of the waveguide structure.

Planar Waveguides

Waveguides formed on a flat substrate are called planar waveguides. These are typically made by stepwise deposition of films of dielectric materials (typically glass). The waveguide core is defined by one of several methods, the most common of which uses lithography and etching. In this case, a film of an appropriately higher index material is deposited on an 'under-cladding' film and then selectively etched to define a core. This is generally covered with a suitable 'upper-cladding' layer so that the index contrast is controlled in all directions. Other methods of forming the core include patterned ion exchange or ion implantation or increasing the index of the medium using a femtosecond laser, etc. In all these cases, the end result is the formation of an embedded higher-index region whose locus defines the optical path for the lightwave. A circuit made from these waveguides is known as planar lightwave circuits (PLC).

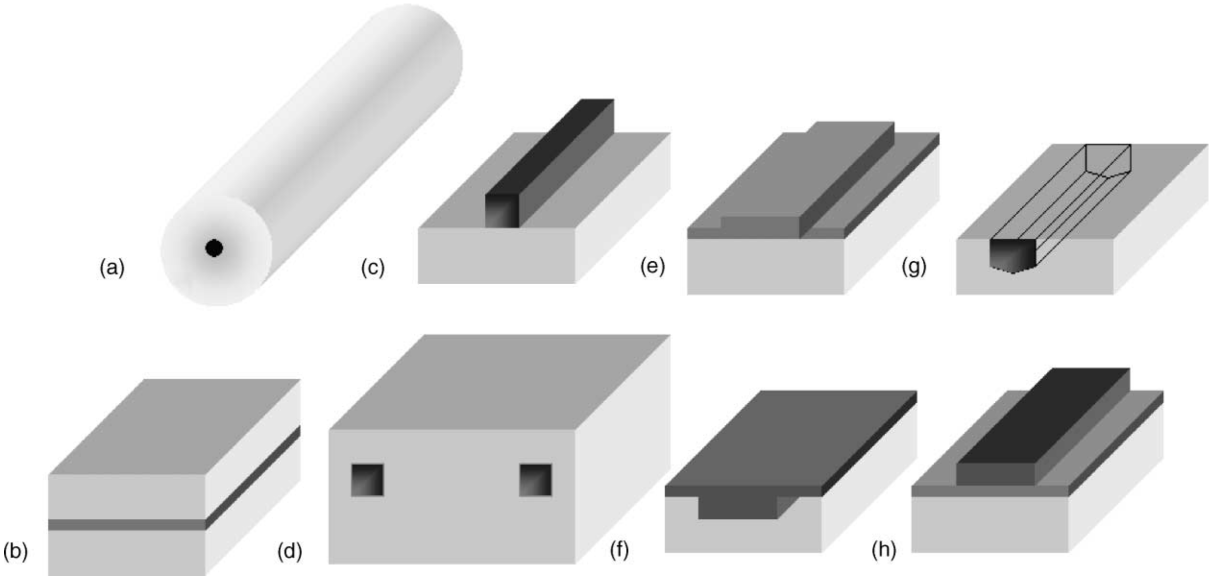

Figure 1 Range of different waveguide designs, with the core represented by the darker regions: (a) Optical fiber; (b) slab waveguide; (c) channel waveguide; (d) embedded channel waveguides with a layer of the overcladding; (e) rib waveguide; (f) inverted rib; (g) ion diffused; and (h)strip loaded, comprising of a strip of a different higher index material (that serves to confine light) on a core layer.

Waveguide cores may be designed in several different configurations, as shown in Figure 1. These designs allow for different fabrication processes as well as the control over the modal properties of the lightwave. For a given index contrast, the number of modes traveling in the waveguide is primarily determined by the dimensions of the waveguide. The aspect ratio of the waveguide affects the propagation of the transverse electric (TE) and the transverse magnetic (TM) modes, due to different propagation constant k_z for each mode.

Lasers

The generation of laser light requires four basic conditions to be satisfied:

- a source capable of light emission;
- an energy supply source – typically electrical, optical, or chemical;
- population inversion to the excited state in the lasing medium; and
- absence of parasitic effects that would absorb the emitted light.

Laser light is often a result of either electronic or vibrational excitation. This is because these excitations involve transfer between energy states that coincide with the radiation of UV, visible, or infrared light. Population inversion is an important requirement in lasers since the emission is stimulated. This involves having a greater fraction of the atoms or molecules in the excited state so that a stimulating photon can cause a radiative decay to the lower energy level. The stimulated light from most lasing media is weak, which requires unacceptably long lengths to generate sufficient power. Instead, as shown in Figure 2, lasers are made with two parallel mirrors with an axial light-generating source in between. This configuration is known as a resonant cavity, a resonator, or an oscillator. The back mirror is usually fully reflective but a partial front mirror allows a portion of the light in the cavity to pass through. The two mirrors cause the light to bounce to and fro, forcing it into making multiple passes through the cavity. Light emanating from this structure is thus amplified and is also coherent (that is the lightwaves are in phase

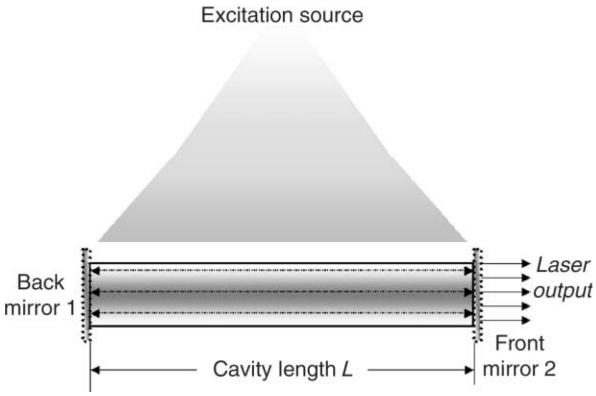

Figure 2 Conceptual representation of a simple Fabry–Perot laser.

with each other spatially and temporally). This has come to represent the signature output of a laser.

Planar Waveguide Lasers (PWL)

By way of an example, let us consider a solid state, rare-earth doped laser, such as an Nd-doped YAG laser. All rare-earth ions are able to fluoresce, some (e.g., Nd) more efficiently than others, and hence can be light emitters. These ions have to be dissolved (doped) into a solid host, say, a glass or a single crystal such as YAG (Yttrium Aluminum Garnet). The excitation of these ions is done optically with light of a suitably higher energy (or lower wavelength). When this 'pump' light is made incident on a rod of the rare earth doped material, it is absorbed by the rare earth ions and used to reach an excited electronic state. Amongst all the possible higher energy states, only certain ones are sustainable because the electron resides in that state for an acceptably long period of time (called the lifetime). Hence, population inversion can be more readily achieved in these longer lifetime states and stimulated processes such as amplification and lasing are readily possible. The laser light generated in the doped rod is emitted across its entire cross-section.

Planar waveguide lasers are designed to miniaturize the kinds of stand-alone solid state lasers describe above. In principle, a PWL is a laser where the light emitting resonant cavity is a planar waveguide. This topic of our interest is a relatively recent development that renders some unique advantages. These can be broadly classified into two categories: one relating to making efficient high-power lasers and the other, to the integration of lasers with other optical components.

An example of the first type is a Nd-YAG laser configured into a planar geometry so that the laser light is now contained within the waveguide. One design that has evolved to achieve this well is shown in Figure 3. The light is guided by the waveguiding structure setup by the sapphire planes. Since both the excitation (or pump) and the lasing modes are well confined, high modal overlap is achieved. These PWLs can be edge-pumped or face-pumped and high powers (measured in the order of Watts) can be achieved.

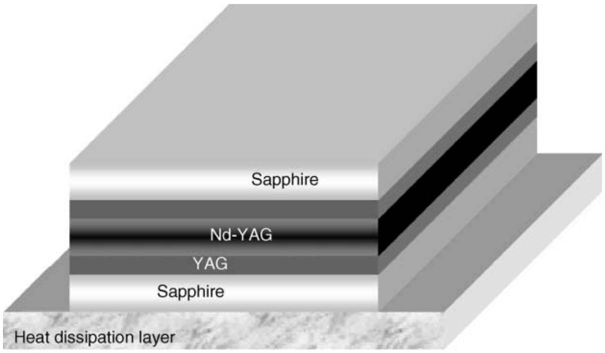

Figure 3 End view of a planar waveguide form of a Nd-YAG laser made using sapphire (single crystal Al_2O_3) for cladding.

In the above design, the cavity is defined by two parallel reflectors bracketing the light-generating medium and is known as a Fabry–Perot etalon. The distance between the two mirrors sets up a wavelength filtering mechanism whereby only resonant wavelengths are sustained. These are related to the intra-mirror distance by the simple equation shown below:

$$L_c = \frac{a\lambda}{2n} \qquad [3]$$

where L_c is the length of the cavity, a is an integer, n is the refractive index of the medium, and λ is the wavelength in vacuum. As the cavity length increases, the number of possible resonant wavelengths increases and the spacing between wavelengths is reduced. These are called the longitudinal modes (Figure 4a). Likewise, lateral modes (Figure 4b) can be generated in the cavity as well, tending to make the operation of the laser multimoded, as shown in Figure 5. This is characteristic of a Fabry–Perot laser, which can be overcome by embellishing the basic etalon design.

As stated above, another big advantage here is that the platform for fabricating PWLs can be chosen to be the same as to make PLCs. Waveguide comprizing the laser can be continued further and configured into other desirable components, such as splitters, couplers, etc. Thus the laser does not have to be separately coupled to the waveguide. This is a major advantage as this connecting process is time-consuming and incurs loss of light. In the field of photonics, the ability to integrate active components with passive ones onto the same chip is considered to be the first step toward photonic chips, analogous to the integrated circuit chip that has revolutionalized our world. This is the driver behind continued research and development of active elements such as PWLs that can be coupled with other actives, such as thin film electro-optic modulators or acousto-optic filters, etc.

Waveguide Laser Materials

The waveguiding structure is comprised of materials that are either dielectrics or semiconductors, both inorganic and organic. For their homogeneity and broad transmittance, high silica glasses are ideal

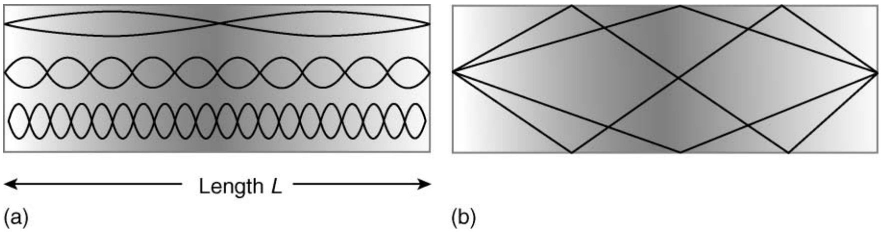

Figure 4 Generation of longitudinal (a) and lateral modes (b) in a Fabry–Perot laser.

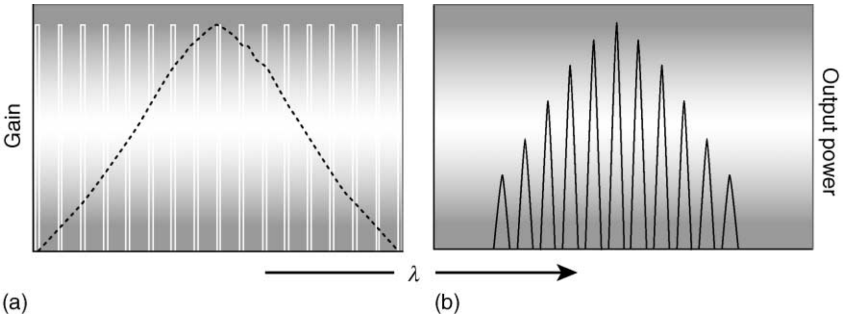

Figure 5 The output of a Fabry–Perot laser: (b) is governed by the number of modes within the gain spectrum of the lasing medium shown by the dotted lines (a). The spacing between the modes can be altered changing the length of the cavity. The smallest mode in output spectrum (b) needs to have enough gain to overcome the lasing threshold.

materials for waveguiding. The chief advantages of high silica waveguides are easy index and mode matching to standard optical fibers, low propagation loss, and good durability. When crystalline materials are used, they have to be single-crystalline to avoid grain boundary scattering (for instance, Ti-diffused waveguide in single-crystal $LiNbO_3$). If semiconductors are used, the bandgap of the material needs to be sufficiently high so as to avoid absorption of the photons. Silicon waveguide (bandgap = 1.1 eV) on a silica cladding (SOI) is commonly used system in PLCs used for telecommunication applications.

The PWLs described above are fabricated in rare-earth doped glass waveguides. Waveguides have to be much shorter than fiber, so the challenge of incorporating the rare-earth ions at high concentration in the host material has to be overcome. Planar waveguide lasers are a compressed version fiber lasers that have the unique advantage that they can be directly integrated with other components on the same chip using very large-scale integration (VLSI)-style processing. The concentration of Er (or the lasing ion in general) in a PWL needs to be many times that in fiber amplifiers (~100 ppm). The need for high gain in a short length requires high pump powers and high lasing ion concentration, both of which lead to nonideal behavior. Since rare earth ions tend to have multiple excited electronic states, there can be absorption at the excited state level as well. When a higher energy state is separated from the upper emission level by the energy in one laser photon, the ions can undergo excited state absorption or ESA (Figure 6). Rare-earth ions in close proximity tend to undergo cooperative upconversion processes, where two excited ions transfer energy between each other so that neither ion may fluoresce at the desired wavelength. These processes can negatively affect the

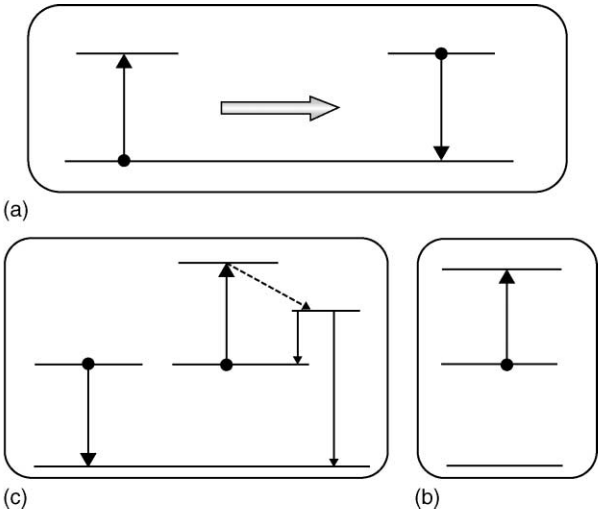

Figure 6 The energy level diagram for a representative two level system. The inset (a) shows the excitation and emission process. Excited state absorption is shown in (b) and cooperative upconversion is shown in (c).

useful gain of the system. Ensuring high spatial dispersion of the Er atoms is critical to minimize this effect. Special glass compositions, such as phosphate glasses, are used here to ensure that the Er is not phase separated or clustered. These glasses also provide a high induced cross-section for gain. Glasses with low phonon energy are also desirable to minimize phonon-mediated nonradiative decays in the rare earths. Light emitting semiconductors and polymers containing fluorescent dyes have also been used to make PWLs.

Now that we have gained an overall perspective of PWLs, we can discuss different manifestations. The following discussion uses a Er doped system as a model since there is a substantial body of work on this system.

Distributed Bragg Reflector (DBR) PWL

In planar waveguide systems, it is difficult to construct efficient mirrors or reflective facets perpendicular to the waveguide. However, mirrors can be replaced by Bragg gratings, which be fabricated relatively easily. In-line Bragg gratings result in wavelength specific reflection as per eqn [4]:

$$\Lambda = \frac{m\lambda}{2\bar{n}} \qquad [4]$$

where Λ is the grating pitch, m is the order of the grating, and λ is the central Bragg wavelength. Gratings have another important benefit in that they ensure that unwanted frequencies outside the limited reflection spectrum are dispersed before they reach the lasing threshold.

A DBR laser is similar in principle to a Fabry–Perot design but uses a Bragg grating on either end of planar Er-doped waveguide. This is shown schematically in Figure 7. Gratings can be made in a number of different ways. One popular way is to exploit the photosensitivity of glasses using 193 nm or 244 nm UV light to periodically alter the refractive index of a glass waveguide. Photosensitivity occurs by a combination of molecular changes (such as the formation of oxygen deficiency sites or color centers) as well as changes in the material strain in the region exposed to the UV light (Figure 7c). Alternatively, gratings can also be formed using etching to make periodic grooves in the waveguide. These are called corrugated or surface relief gratings (Figure 7a and 6b).

The back Grating 1 has a high reflectance of almost 100%, whereas Grating 2 is a weaker grating. The reflectance strength of Grating 2 is determined by taking into account the gain coefficient of the material and the output power required from the laser. The reflection intensity can be altered according to the following equation:

$$R = \tanh^2\left[\frac{\pi L_G \Delta n_{\text{eff}} \eta(V)}{\lambda_B}\right] \qquad [5]$$

where L_G is the grating length, Δn_{eff} is the index modulation of the grating, λ_B is the Bragg wavelength, $\eta(V)$ is the confinement factor, a function of the waveguide parameter V that represents the fraction of the integrated mode intensity contained in the core. Increasing the reflectance strength also widens the linewidth of reflected light. The overlap of the spectral width between the back (Grating 1) and the front grating (Grating 2) becomes the effective output window within the gain spectrum of the lasing material (Figure 8).

Not unexpectedly, a DBR structure generates longitudinal and lateral modes. These modes are, however, limited to the overlapping section of the reflection spectrum of the two gratings, as opposed to the entire gain spectrum as in Fabry–Perot lasers (Figure 9). Yet, this can lead to problems such as spectral hole burning and mode hopping, whereby the dominant mode hops from one wavelength to another. The separation between the modes needs to

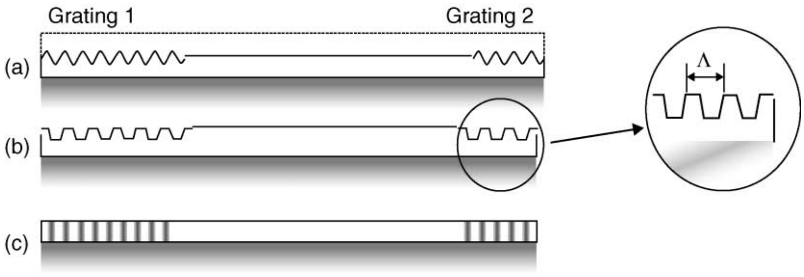

Figure 7 DBR planar waveguide laser. The PWLs shown in (a) and (b) have sinusoidal and trapezoidal surface-relief gratings respectively etched in their cores. An in-line grating of the kind seen in (c) could be fabricated using the photosensitivity of the glass. The top cladding is shown only in (a).

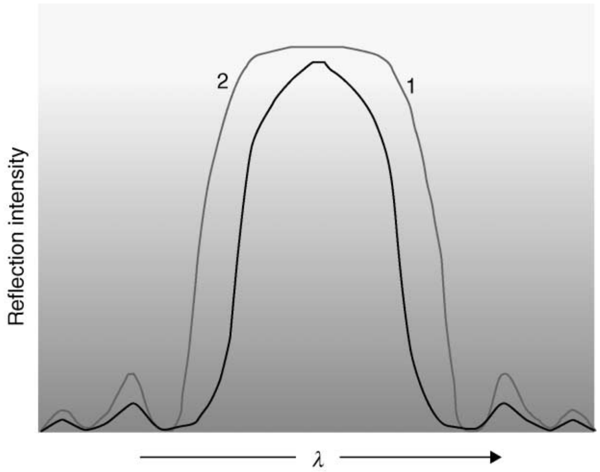

Figure 8 The reflection spectrum of the back (1) and the front (2) gratings. As the intensity of the reflected light is made to increase, so does the linewidth. Hence, the narrower spectrum of partially reflecting grating 2 tends to dictate the operating wavelength range of a DBR.

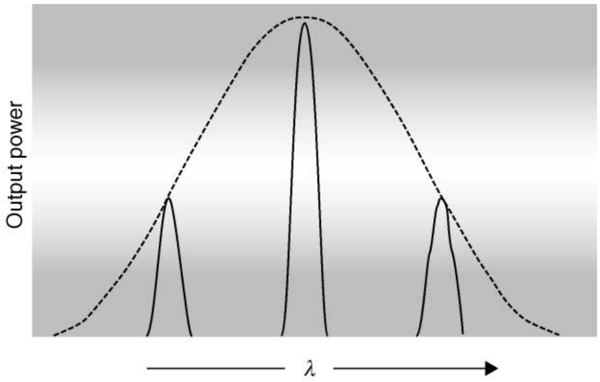

Figure 9 A sample output from a DBR laser, shown in this case with 3 modes. The dotted line shows the effective overlap between the two gratings. The existence of the modes depends on several factors, in particular the pump intensity.

be made sufficiently high compared to the spectral overlap so that single-mode operation can be achieved. But in practice the length of the cavity (L_c in eqn [3]) is determined by factors such as power and space and may not be independently altered to achieve this. However, the reflection spectrum of the back and the front gratings could be slightly shifted in opposite directions to realize a smaller overlap within which only one or a few modes may exist (Figure 10). The occurrence of a single mode also depends on the gain profile of the lasing material and the optical pump power. In Er-doped PWLs, modes other than the primary one can be suppressed by using a pump power below a threshold limit and stable operation can be thus achieved.

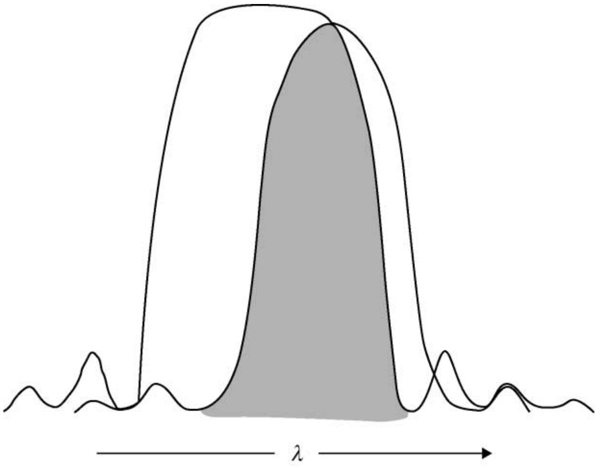

Figure 10 Shifting the front and back gratings is one means of controlling the width of the overlap region, shown in gray. This way a single mode operation can be achieved even for a longer cavity with many modes.

Distributed Feedback (DFB) PWL

The modal problems listed above can be addressed by having a single, distributed grating that is placed along the entire length of the lasing waveguide itself (Figure 11). Alternatively, the gratings may be also placed in the upper or the lower claddings to minimize loss of optical power due to scattering because of the longer grating. In this case, it is the evanescent portion of the light that interacts with the grating. In all cases, the periodicity of the grating leads to a condition for a single lasing wavelength based on constructive interference. The strongest mode occurs when the period Λ of the Bragg grating satisfies the primary Bragg condition (for first order grating or $m = 1$) in eqn [4]. The device will also function if the grating pitch is equal to small integer multiples of $(\lambda/2\bar{n})$.

The grating needs to be made strong enough to generate sufficient feedback (reflections), which also widens the linewidth of the spectrum. To ensure a single narrow linewidth output, a quarter wave phase shift is introduced in the grating, as seen in Figure 11. This phase shift creates a transmission fringe in the spectrum (Figure 12) that significantly narrows the linewidth of the output signal. This design is considered to be standard for the DFB lasers of today. Figure 13 shows a typical output of a DFB laser.

Any of the geometries seen in Figure 1 can be used to make the waveguide for the laser. Due to processing constraints, rib (c) and ion diffused (d) configurations are commonly used. Waveguide losses, characteristically due to scattering from sidewall roughness, have to be low. Otherwise, this adds a significant overhead onto the pump power and lowers the output of the PWL. Birefringence caused by thermal expansion coefficient differences and excessive sidewall roughness also need to be minimized as they can substantially alter the loss of the TE mode compared to TM.

Though their spectral properties are ideal as sources for long haul transmission, DFB lasers tend to be less powerful than the corresponding DBR designs. But the DBR PWLs are very sensitive to reflections that can cause a broadening of the laser linewidth or can act as a source of noise. In both cases, single mode lasers with very narrow linewidths in the kHz range with low radiance and noise, have been demonstrated. But neither is known to have been commercialized yet.

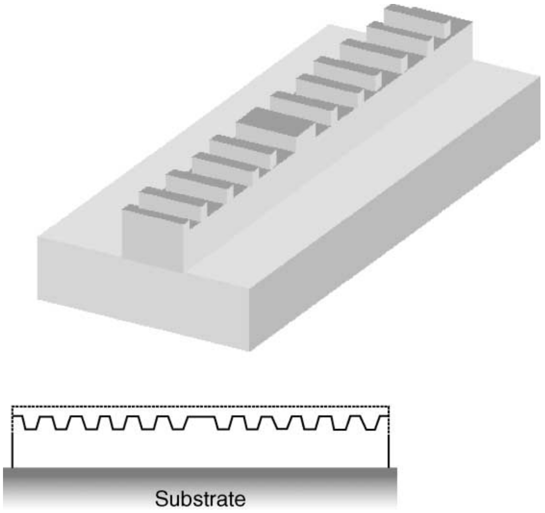

Figure 11 A DFB planar waveguide laser.

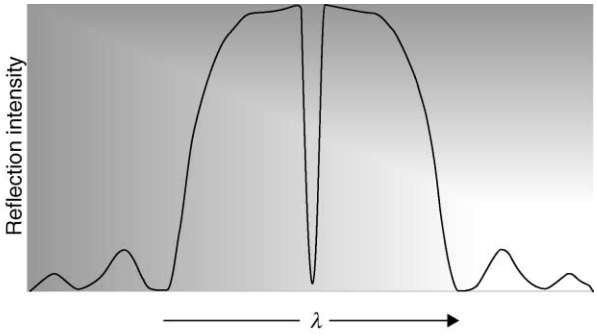

Figure 12 The reflection spectrum of a grating with a central λ/4 shift in the pitch, showing the transmission fringe.

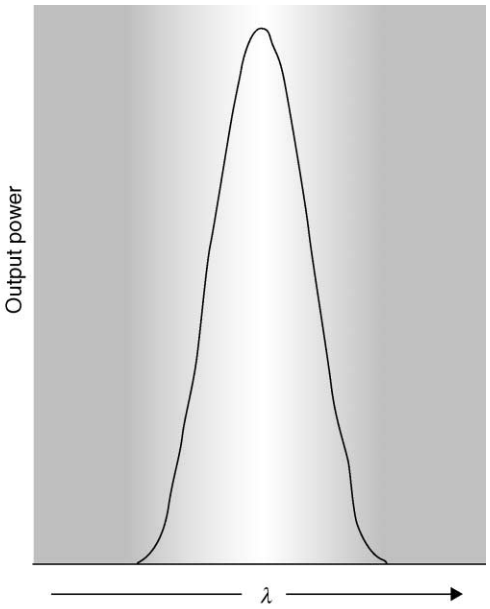

Figure 13 The typical output of a DFB PWL. The linewidth can be made to be as narrow as a few hundred kHz.

PWL Arrays

One of the unique advantages of the PWLs is that they are often processed using VLSI techniques. One important implication of this is the ability to make a PWL array where multiple lasers are formed simultaneously on the same chip, each with a unique output (Figure 14). Such arrays are likely to be very useful in future dense wave division multiplexing (DWDM) optical network systems where as many as 128 channels (and more) are planned. Since packaging of individual lasers can be up to 75% of the total cost, making chips with multiple lasers is very attractive. The lasers may be DBRs or DFBs whose output wavelength is controlled by changing the grating parameters. To vary the individual wavelengths either the effective index and/or the pitch of the grating can be changed. Thus, any output spectrum can be designed and executed merely by making the appropriate lithography mask.

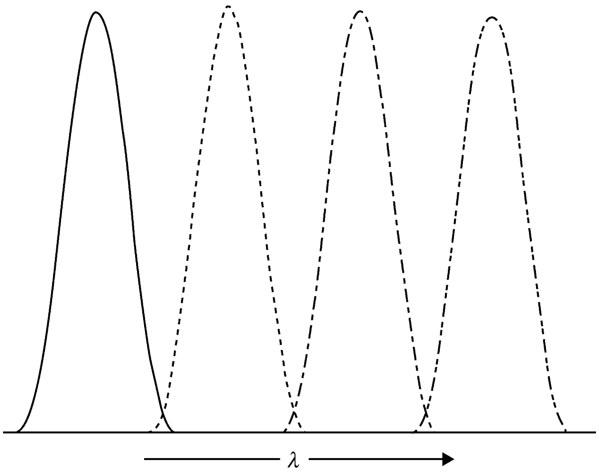

Figure 14 Idealized output from a 4-PWL array.

The gratings are often written holographically, where the wafer is exposed to an interference pattern caused by two coherent light beams. The wafer underneath can be sensitized with a photosensitive glass layer or a film a photoresist. It is not efficient to do multiple holographic exposures to individually change each of the gratings in the array. One way of doing this with a single holographic exposure is by constructing the waveguides at different relative angles. Then the component of the index modulation, resolved along the axis of the waveguide, would vary for each waveguide depending on the relative waveguide angle (Figure 15); this would lead to differences in $\bar{n}$. Alternatively, the waveguide array can be made of the same height but of different widths. This is easily done since waveguide height is usually deposited uniformly but the width is a function of the mask layout. Here the waveguide and the grating are orthogonal to each other and the variation in wavelengths is achieved by the fact that the effective propagation constant for each of the waveguides is different.

Arrays of PWLs have been made on a silicon substrate using deposition of an Er containing glass that is subsequently patterned into waveguide, and ultimately laser, arrays. Another approach has been to start with a phosphate glass as the host and the substrate. A phosphate glass is a superior host for rare-earth ions such as Er since the sensitization efficiency is nearly unity and much high doping levels are possible before concentration quenching effects set in. Waveguides are created by ion-exchange by immersion in a hot bath containing suitable ions (Na for Li, for instance). When starting with a uniformly doped glass, one issue in the assimilation of the laser array with other components is the fact that Er-doped region does not start and end with the waveguide laser section. Recently, the Er has been deposited selectively in the laser waveguide in phosphate glass using methods such as ion implantation, bringing the concept of integration closer to reality.

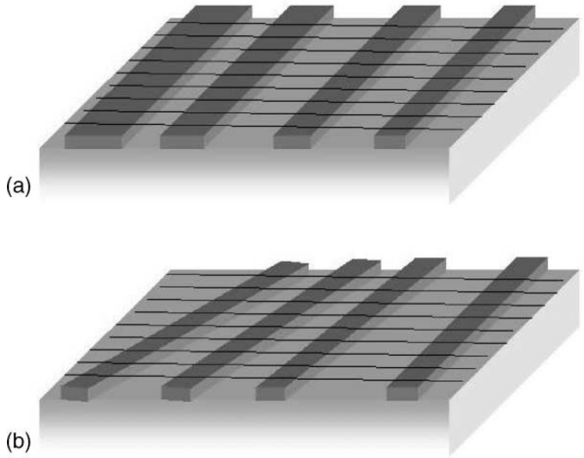

Figure 15 Two different means of making planar waveguide laser array with a single holographic exposure to write the gratings. The darker rectangular blocks represent the core channels or ribs. In (a) the width and hence the effective index of the waveguide is varied whereas the relative angle between the waveguides is altered in (b). The lines represent the direction of the gratings, which is orthogonal only to the left most waveguide in (b).

Pumping

One of the key issues here is the optical pumping of the rare-earth PWLs to generate acceptably high output power. Individual diode lasers or an array of diode lasers all on a single chip (called a diode bar) have been utilized as pumps. Both single and multimode diode lasers are available and have been used. There are pros and cons for either choice: whereas single mode lasers are less powerful (in the range of few 100 mW), it is difficult to convert the multimoded output of their counterparts to useful power. The choice is determined by the waveguide design, which needs to be optimized for high overlap of the lasing mode with the pump. Carefully tapered waveguides can be used distribute the pump power from a diode bar to the various lasers in an array without significant loss.

PWL Stability

Stable operation of a PWL is one of the big challenges. In DBR PWLs, it is seen that there is a threshold for pump power only below which single moded operation can be achieved. Additionally, any PWL die needs to be made immune to ambient temperature changes. These changes can affect many of the critical output parameters of the PWL: wavelength, linewidth, power and in the case of an array, the wavelength separation. This occurs primarily due to the thermal shift in the grating pitch as well as the thermo-optic effect. It is overcome by supplanting the laser die onto a heat spreader chip made from conducting material such as Cu or BeO, followed by a thermoelectric cooler. Ironically, the laser's output wavelength and linewidth are fine-tuned and 'fixed' using temperature to overcome any drifts due to processing errors or other reasons.

Reliability is another important issue that needs to be addressed, especially for telecommunication applications where these types of components have to pass high quality standards such as the Telecordia qualification in the USA.

Other PWL Systems

Many of the PWLs to date have been developed using Er doping, which emits light in the range of interest for telecommunications (the C-band and even the L-band). Here, Yb can be added as a co-dopant to improve power conversion efficiency. Other fluorescing ions have also been used: Nd, Tm, Ho, and Cr. Several research teams have successfully demonstrated Nd-YAG PWLs, in particular. Ridge waveguide lasers have been made using semiconductor materials InGaAs/InP heterojunctions in both Fabry–Perot and DFB configurations. These systems are electrically pumped. DFB PWLs have also been made using dye doped polymers.

It must be mentioned that there are other ways to use waveguides to make lasers. The primary example is a configuration where the waveguide resonator is external to the active region. Designs, such as ring resonator lasers and an external cavity lasers, fall in this category.

Future research and development of PWLs would undoubtedly utilize photonic crystal structures. These structures have a periodic fluctuation in the index of refraction in two or three dimensions. Owing to the resultant Bragg reflections, specific allowed and forbidden states are created which can be used to control the confinement of light. Hitherto, these structures have been used to form mirrors or hexagonal ring resonators. These more sophisticated laser designs are likely to enable new functions such as tunability of the output spectrum.

Summary

Planar waveguide lasers are a special class of laser where light is confined to a waveguide. They have distinctive advantages that include high optical gains, low laser thresholds, narrow linewidths in the kHz range, and optimal thermal management. Significantly, they afford a platform that can be readily expanded to include an array of lasers, each with an unique output, and can ultimately be combined with multiple optical components on the same chip. Such compact integrated devices are bound to take photonics to a new high level of capability.

List of Units and Nomenclature

a	integer denoting number of laser modes
E	electric field
k_z	propagation constant for travel in z direction
L_c	length of laser cavity
L_G	length of grating
m	order of the grating, an integer
n	refractive index of the medium
$\bar{n}$	modal effective refractive index
R	reflectance
V	normalized frequency or the V parameter of the waveguide
Δn_{eff}	modulation of the effective index of the grating
η	confinement factor or fraction of the light confined in the region of interest
λ	wavelength of the lightwave
λ_B	resonant wavelength of a Bragg grating or Bragg wavelength
Λ	grating pitch
Φ	electric field distribution

See also

Lasers: Carbon Dioxide Lasers; Dye Lasers. **Optical Amplifiers:** Erbium Doped Fiber Amplifiers for Lightwave Systems.

Further Reading

Carroll JE, Whiteaway JEA, Plumb RGS and Plumb D (1998) *Distributed Feedback Semiconductor Lasers*. London: IEE Publishing.

Ghafouri-Shiraz H (2003) *Principles of Semiconductor Laser Diodes and Amplifiers: Analysis and Transmission Line Laser*. London: Imperial College.

Hecht J (2001) *Understanding Lasers: An Entry Level Guide*. New York: Wiley Interscience.

Lee JR, Baker HJ, Friel JG, Hilton GJ and Hall DR (2002) High-average-power Nd:YAG planar waveguide laser, face-pumped by 10 laser diode bars. *Optical Letters* 27: 524–526.

Milonni PW and Eberly JH (1988) *Lasers*. New York: John Wiley and Sons.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: John Wiley & Sons.

Shepherd DP, Hettrick SJ, Li C, *et al.* (2001) High-power planar dielectric waveguide lasers. *Journal of Physics D: Applied Physics* 34: 2420–2432.

Siegman AE (1986) *Lasers*. Harendon, Virginia: University Science Books.

Svelto O (ed.) (1998) *The Principles of Lasers*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Veasey DL, Funk DS and Sanford A (1999) Arrays of distributed-Bragg-reflector waveguide lasers at 1536 nm in Yb/Er codoped phosphate glass. *Applied Physics Letters* 74(6): 789–791.

Young M (2001) *Optics and Lasers: Including Fibers and Optical Waveguides*. NewYork: Springer-Verlag New York, Inc.

Semiconductor Lasers

S W Koch, Philipps-University, Marburg, Germany

M R Hofmann, Ruhr-Universität Bochum, Bochum, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article discusses the principle of operation and the basic material engineering aspects for the design of semiconductor lasers. The physics of the gain medium is analyzed in detail and the role of many-body effects due to the Coulomb interaction of the carriers for the optical properties of the active semiconductor material is demonstrated. The two main types of laser resonators for diode lasers, Fabry–Perot resonators and vertical cavity structures, are described. Finally, the dynamics of semiconductor lasers, including the principles of short pulse generation and nonequilibrium gain dynamics, is addressed.

Basic Principles

The optical emission from semiconductor lasers arises from the radiative recombination of charge carrier pairs, i.e., electrons and holes in the active area of the device. The conduction-band electron fills the valence-band hole by simultaneously transferring the energy difference to the light field. This is called radiative recombination. Hence, the frequency of the laser light is mainly determined by the bandgap of the semiconductor laser material. In order to achieve lasing, one needs carrier inversion, i.e., a sufficient number of electrons in the conduction band so that the probability of light absorption is lower than the probability of emission.

It has been shown that a large variety of semiconductor materials is suited for semiconductor lasers. The most stringent requirement for the material is that it has a direct bandgap. In other words, the maximum of the valence band and the minimum of the conduction band have to be at the same position in *k*-space (in the center of the Brillouin zone). The elemental semiconductors silicon and germanium do not fulfill this condition and therefore cannot be used for semiconductor lasers.

In order to achieve carrier inversion, it is necessary to pump the semiconductor laser, i.e., to excite a sufficient number of electrons from the valence band into the conduction band. Even though this pumping in semiconductors, as in other solid state laser materials, can be done optically, the large technological impact of semiconductor lasers is at least in part due to the possibility of electrical pumping with a few tens of milli-amperes at low voltages.

The original version of such a semiconductor laser device is shown in Figure 1, The laser structure consists of a GaAs-based p-n junction which is biased in the forward direction. Electrons are injected from the n region and holes from the p region, respectively. Due to the diode characteristics of the p-n junction, semiconductor lasers are often also called laser diodes. In the device shown in Figure 1, the laser mirrors are formed by the cleaved facets of the semiconductor crystal. Because of the high refractive index of GaAs, the semiconductor–air interface provides a reflectivity of about 32% which is sufficient for laser emission.

However, the early type of laser, the so-called homojunction device, was not very efficient because of the low carrier density in the active area and because of the weak overlap between the inverted region and the optical mode. Considerable improvement towards room-temperature continuous wave (cw) operation was achieved by the introduction of the so-called double heterostructure. Its principle is shown in Figure 2.

The basic idea behind the double heterostructure laser is that the active material (e.g., GaAs) is embedded in another semiconductor with a slightly larger bandgap (e.g., (AlGa)As). This causes a potential well in which electrons and holes are confined in order to achieve high carrier densities in the active area. Moreover, the smaller refractive index of the surrounding high-bandgap material helps to guide the optical field within the region of highest inversion. Additional ridge-shaped lateral structuring is frequently used to form a complete optical waveguide.

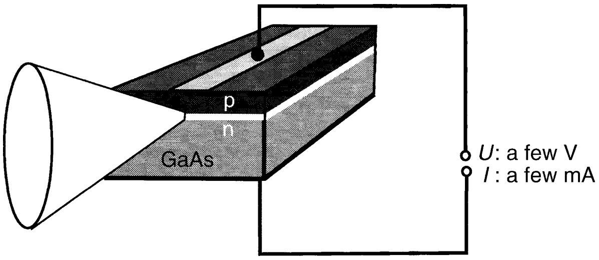

Figure 1 Principle of operation of a semiconductor laser.

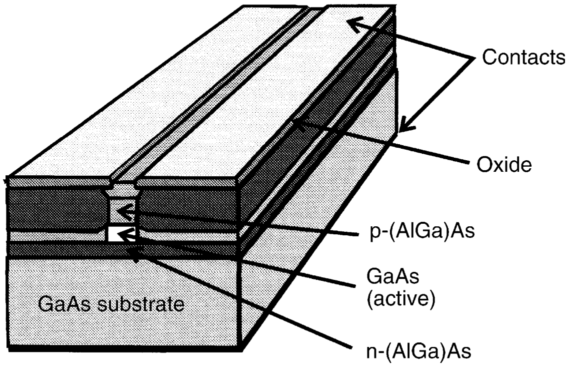

Figure 2 Double heterostructure laser diode.

Nowadays, even more advanced laser structures can be fabricated, as modern crystal growth techniques such as molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD) have become available. These techniques allow the grower not only to determine the composition of the semiconductors with remarkable precision, but also to define the shape virtually on an atomic scale. In particular, it is now possible to grow microstructures so small that their electronic and optical properties deviate substantially from those of the corresponding bulk materials. So-called quantum-well (QW) structures, in particular, turned out to be superior to bulk materials in many respects and are currently the active medium in most commercial semiconductor lasers. The gain medium in QW laser structures consists of one or more films with a thickness of a few atomic monolayers. These are embedded in a barrier material with a larger bandgap that serves to confine the carriers to the wells. In QWs the quantum confinement restricts the free carrier motion to the two dimensions within the plane of the films, since the well thickness is comparable to or less than the thermal wavelength of the carriers. Hence, one speaks of quasi-two-dimensional structures where no free motion perpendicular to the films is possible. The corresponding energy states of the carriers are quantized, i.e., only discrete values are allowed for that part of the carrier kinetic energy which is related to the confined direction.

A major consequence of this restriction of the carrier motion to two dimensions is a change in the (single particle) density of states. This density of states is altered in QWs in comparison to bulk material. The density of states for bulk material is zero at the bandgap energy and rises with the square root of the energy for higher energies. In contrast, the density of states for a two-dimensional carrier system in a QW is a step-like function with a density of states different from zero at the QW-well bandgap (the bandgap energy of the active material plus the quantization energies of electrons and holes in the well). This increase of the density of states at the bandgap promises important advantages of QW structures over bulk material. Most importantly, the number of states that have to be occupied before inversion is obtained is considerably reduced in QW structures. This reduces the threshold current and increases the modulation speed of the laser structures. Accordingly, most of the modern laser structures are QW structures.

These arguments have encouraged many researchers to develop structures with further increased carrier confinement in two dimensions (quantum wires) and three dimensions (quantum dots). The single-particle density of states is even more favorable in quantum wires and quantum dots than in QWs: in both cases the single-particle density of states peaks at the effective bandgap energy. This argument seems to promise even better performance of quantum wire and quantum dot lasers as compared to QW devices. However, the single-particle density of states description completely neglects the Coulomb interaction of the carriers which strongly influences the optical and electronic properties of semiconductor structures. The absorption is modified by excitonic effects that lead to additional resonances below the bandgap and to changes of the continuum absorption. These changes are often quantified by the so-called Sommerfeld factor, which depends on the dimensionality of the system, i.e., it differs for bulk, QW, quantum wire and quantum dot structures. For example, the Sommerfeld factor completely suppresses the singularity of the density of states at the bandgap of quantum wire structures. Accordingly, with inclusion of Coulomb effects, the benefit of the quantum confinement is much less than expected from the single-particle density of states.

The use of quantum wire and quantum dot structures as a semiconductor laser gain medium is still in its infancy even though they can be grown with various degrees of precision. However, quantum dot structures, especially, hold great promise for the future, since they can be grown in large numbers using techniques similar to those already used for QWs. The possibility of layer growth under strained conditions, such as QWs between barrier material with a slightly different atomic lattice constant, also results in a greater flexibility to design the laser gain medium with the desired emission wavelength.

In principle, the emission wavelength of a semiconductor laser can be roughly designed by the choice of the material. Different semiconductors have different bandgaps between the valence band and conduction band and the bandgap energy determines the spectral range of the emission. One can tune the bandgap and thus the spectral range of the emission by fabricating semiconductor alloys as, e.g., (GaIn)As or (AlGa)As. Presently, a large variety of different materials is commercially available covering, for example, the red ((AlGaIn)P), the near infrared ((AlGa)As, (GaIn)As), the telecom ((GaIn)(AsP)) and even the blue–ultraviolet range ((GaIn)N).

Great care has to be taken that the crystal used for a semiconductor laser is of almost perfect quality. Dislocations in the crystal, for example, have to be avoided as well as any kinds of defects since they act as nonradiative recombination centers. Nonradiative recombination competes with the radiative recombination that is essential for the laser process. Nonradiative recombination effects include Auger recombination and recombination via defects. Auger recombination is the process where an electron and a hole recombine and completely transfer the energy difference to another carrier (electron or hole) which is excited to a higher state. This is an intrinsic effect that depends mainly on the band structure and the relevant values of the transition probabilities (transition matrix elements). The probability for Auger recombination is temperature- and carrier-density-dependent and can be minimized, e.g., by properly cooling the device and by keeping the carrier densities as low as possible.

Nonradiative recombination via defects is an extrinsic effect that depends on the material quality and can thus be minimized by optimized growth. However, trying to avoid dislocations limits the choice of materials for laser structures severely. The growth processes require 'host' crystals, so-called substrates, on which the real structure can be grown. GaAs, InP, SiC and sapphire wafers are available in sufficient quality as standard substrates for optoelectronic applications. The epitaxially grown material adapts to the crystal structure of the substrate which implies that the lattice constant of the substrate and structure material should be very similar. If the lattice constants do not agree, i.e., if there is a structural mismatch, a certain amount of strain develops which increases with increasing lattice mismatch. Too much strain leads to the formation of dislocations which are detrimental to laser applications. Control of the dislocation density is currently one of the critical issues for the wide-gap III–V materials, such as GaN, because no substrate is available that matches the GaN lattice constant. However, a small amount of strain is often even desirable as long as it does not cause dislocations. Strain influences the band structure of the material in a predictable way and can therefore be used to tailor the laser emission properties.

Thus laser materials for a certain desired wavelength range have to fulfill two important requirements: first, the bandgap of the material has to fit the desired photon energy; and second, and more crucial, an appropriate substrate with the right lattice constant has to be available. This second condition excludes many material compositions from laser applications. In many cases, both conditions can only be simultaneously met with quaternary semiconductor alloys (for example (GaIn)(AsP) on InP substrates).

Physics of the Gain Medium

From a material physics point of view, the important structural aspects of a certain laser material are summarized in the band structure of the gain medium, i.e., in the quantum mechanically allowed energy states of the material electrons. The calculation of the band structure of a particular gain medium is therefore a crucial aspect of semiconductor laser modelling. Additionally, however, one has to understand the relevant properties of the excited electron–hole plasma in the active region of the laser.

In its ground state, i.e., without excitation and at very low temperatures, a perfect semiconductor is an insulator where no electrons are in the conduction-band states. The presence of a population inversion, i.e., occupied conduction band and empty valence band states in a certain range of frequencies, is, however, a requirement for obtaining optical gain and light amplification from a semiconductor. Sufficiently strong pumping leads to the generation of an electron–hole plasma, where the term 'electron' specifically refers to conduction-band electrons and 'holes' to the missing electrons in the originally full valence bands. Holes are also quasiparticles, just as the electrons. Electrons in a solid are quasiparticles with an effective mass that is considerably different from the free-electron mass in vacuum and that is determined by the band structure, i.e., the interaction with all the ions of the solid. The properties of the missing electrons, such as a positive charge for the missing negative electron charge, a spin opposite to that of the missing electron, etc., are attributed to the holes. As a consequence of their electrical charges, the excited electrons and holes interact via the Coulomb potential which is repulsive for equal charges (electron–electron, hole–hole) and attractive for opposite charges (electron–hole). Furthermore, electrons and holes obey Fermi–Dirac statistics and the Pauli exclusion principle that forbids two Fermions from occupying the same quantum state. The consequences of the Pauli principle are often referred to as 'band-filling effects' in the physics of semiconductor lasers.

The theory therefore has to account not only for the band-structure but also for this band filling, i.e., the detailed population of the electron and hole states. Taking into account only the band structure and band filling leads to a so-called free-carrier theory. But the electron–hole plasma in a semiconductor gain medium is an interacting many-body system because of the carrier interactions and the exclusion principle. In a many-body system the properties of any one carrier are influenced by all the other carriers in their respective quantum states. Generally, advanced many-body techniques are needed to systematically analyze such an interacting electron–hole plasma and to compute the resulting optical properties, such as gain, absorption or refractive index, all of which are changing in a characteristic way with the changing electron–hole density.

It is often helpful to investigate the characteristic time-scales and the most direct consequences of the various interaction effects even though the consequences of the different many-body effects are in general not additive. The fastest interaction, under typical laser conditions, is the carrier–carrier Coulomb scattering that acts on scales of femto- to picoseconds to establish Fermi–Dirac distributions of the carriers within their bands. These distributions are often referred to as 'quasi-equilibrium distributions' since they describe electrons and holes in the conduction and valence bands which may be characterized by an 'electronic temperature' or 'plasma temperature.' The plasma temperature is a measure for the mean kinetic energy of the respective carrier ensemble and relaxes towards the lattice temperature as a consequence of electron–phonon (= quantized lattice vibration) coupling, i.e., the carriers can emit or absorb phonons. The coupling to longitudinal optical phonons in polar materials is especially strong with scattering times in the picosecond range. The corresponding times for the electron–acoustic phonon scattering are in the nanosecond range. Generally, the carrier–phonon interaction provides an exchange of kinetic energy of the carriers with the crystal lattice in addition to the relaxation of the initial nonequilibrium distribution. Deviations between plasma and lattice temperature may occur, especially when lasers are subject to strong pumping or rapid dynamic changes.

The Coulomb interaction not only causes rapid carrier scattering but also directly influences the laser gain spectrum. Gain, i.e., negative absorption, occurs in the spectral region of population inversion, in other words, where the net probability for a test photon to be absorbed is lower than the probability to be amplified by stimulated electron–hole recombination. Energetically, the gain region is bounded from below by the effective (or renormalized) semiconductor bandgap and from above by the electron–hole chemical potential. Here, the chemical potential is the energy at which neither absorption nor amplification occurs, i.e., where the semiconductor medium is effectively transparent.

The effective bandgap energy for elevated carrier densities is significantly below the fundamental absorption edge of an unexcited semiconductor. This 'bandgap renormalization' is a consequence of the fact that the Pauli exclusion principle reduces the repulsive Coulomb interaction between carriers of equal charge. Furthermore, the Coulomb screening, i.e., the density-dependent weakening of the effective Coulomb interaction potential, contributes to the reduction of the bandgap.

Another many-body effect originates in the Coulomb attraction between an electron and a hole, and leads to an excitonic or Coulomb enhancement of the interband transition probability. All these effects contribute to the detailed characteristics of the optical spectra of a semicondcutor gain medium. Figure 3 shows an example of calculated spectra in comparison to experimental data.

The example demonstrates that the microscopic theory correctly accounts for the changes in the gain spectrum with changing carrier density by means of a consistent treatment of band structure, band-filling, and many-body Coulomb effects.

The Resonator

In addition to a gain medium, a laser needs a positive feedback process for the amplified stimulated emission. Hence, the inverted medium has to be

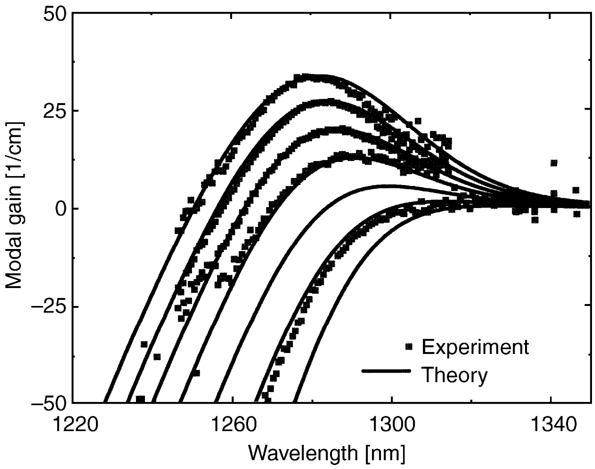

Figure 3 Comparison of experimental and theoretical gain spectra of a (GaIn)(Nas)/GaAs laser.

embedded into a resonator. One of the conceptually simplest versions of such a resonator is a so-called Fabry–Perot resonator which consists of two parallel plane mirrors. A double heterostructure Fabry–Perot laser is shown in Figure 2. Such Fabry–Perot resonators are very easy to fabricate for semiconductor lasers: the semiconductor crystal is cleaved perpendicular to the optical waveguide. The cleaved edges provide a reflectivity of about 32% due to the high refractive index of the semiconductor. This is sufficient to provide enough feedback for laser operation.

However, despite its conceptual simplicity, the cleaved-edge Fabry–Perot concept also introduces a number of problems. First, the light output occurs on two sides of the laser, whereas for most applications all the output power should be emitted out of one facet only. For practical applications, this problem can be solved by depositing coatings onto the cleaved facets of the structure. For example, deposition of a 10% reflectivity coating on one facet and of a 90% reflectivity coating on the other facet concentrates the laser emission almost completely to the facet of lower reflectivity. Antireflection coatings with reflectivities below 10^{-4} are used when laser diodes are coupled to external cavities, e.g., in tunable lasers.

A second problem with Fabry–Perot lasers is that the emission wavelength is often not well defined. A Fabry–Perot resonator of length L (optical length nL) has transmission maxima – so-called modes – at all multiples of $c/2nL$ where c is the speed of light and n the refractive index. That means that for typical devices with a length of 500 μm, hundreds of modes are present even within the limited gain bandwidth of a semiconductor. These modes often compete with each other resulting in multimode emission which may become particularly complex and uncontrollable when the laser is modulated for high-speed operation.

This problem can be overcome in so-called distributed feedback (DFB) lasers. In these devices, the refractive index is periodically modulated along the waveguide. In a simplified picture, this periodic pattern causes multiple reflections at different locations within the waveguide. For certain wavelengths, these multiple reflections interfere constructively to provide enough 'distributed' feedback for laser operation. With the DFB concept, single-mode emission can be realized even for high-speed modulation of the lasers. Alternatively, the region of distributed feedback can be separated from the region of amplification in so-called distributed Bragg reflector (DBR) lasers. These are devices with multiple sections where one section is responsible for the optical gain and another section (the DBR section) is responsible for the optical feedback.

So far, we have discussed so-called edge emitting lasers where the light emission is in the plane of the active area of the device. The lengths of these lasers are typically hundreds of microns and thus correspond to a few hundreds of optical wavelengths. This causes one of the major disadvantages of edge emitters – the multimode emission. Another disadvantage is the poor elliptic beam profile due to diffraction at the small rectangular output aperture. Finally, the fabrication procedure of edge emitters is complex since it requires multiple steps of cleaving before it is even possible to test the laser.

A novel alternative semiconductor laser concept was realized in the early 1990s: the vertical-cavity surface-emitting laser (VCSEL). A typical VCSEL structure is shown in Figure 4. In VCSELs, the light emission is parallel to the growth direction, i.e. perpendicular to the epitaxial layers of the structure. The cavity length of a VCSEL is in the order of a few (1–5) multiples of half the wavelength of the emitted light in the material. These small dimensions imply that the interaction length between the active medium and the light field in the resonator is very short. Therefore, the mirrors of a VCSEL have to be of very good quality; reflectivities of more than 99% are required in most configurations.

Very high reflectivities can be achieved with dielectric multilayer structures, so-called Bragg reflectors. Bragg reflectors consist of a series of pairs of layers of different refractive indices, where each layer has the thickness of a quarter of the emission wavelength. Most preferably, Bragg reflectors are epitaxially grown together with the active area of the device in only one growth process.

Small VCSEL resonators with Bragg reflectors at both ends of the cavity are called microcavities. They are designed in such a way that only one longitudinal mode is present in the region of the semiconductor gain spectrum. Accordingly, VCSELs are well suited to provide single-mode characteristics. Moreover, a round aperture for light emission can be realized with a diameter of a few micrometers so that

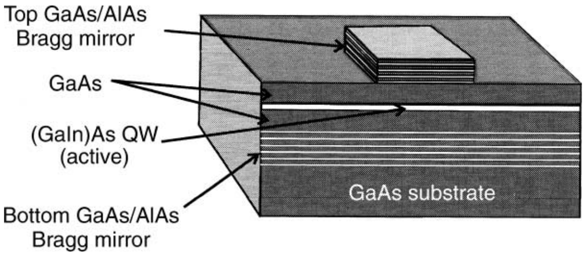

Figure 4 Vertical-cavity surface-emitting laser (VCSEL).

diffraction is weak and the beam profile is almost perfect.

VCSELs can be tested on wafer without further processing and can be arranged in two-dimensional arrays. A major disadvantage of VCSELs is their temperature dependence. The bandgap of the active material and the cavity mode depend on temperature in such a way that the cavity mode shifts away from the region of maximum gain when the temperature is varied. This severely limits the temperature range of operation for VCSELs.

Semiconductor Laser Dynamics

The dynamical behavior of semiconductor lasers is rather complex because of the coupled electron–hole and light dynamics in the active material. Like other laser materials, semiconductor lasers exhibit a typical threshold behavior when the pump intensity is increased. For weak pumping, the emission is low and spectrally broad; the device operates below threshold and shows (amplified) spontaneous emission. At threshold, the optical gain compensates the losses in the resonator (i.e., internal losses and mirror losses) and lasing action starts. Ideally the output power increases linearly with the injection current above threshold. The derivative of this curve is proportional to the differential quantum efficiency. The differential quantum efficiency can reach values close to unity and even the overall efficiency of the diode laser, i.e., the conversion rate from electrical power into optical power, can be as high as 60% which is far higher than for any other laser material.

Ideally, the laser emission is single mode above threshold. However, as discussed above, edge-emitting diode lasers tend to emit on multiple longitudinal modes and the competition of these modes can lead to complex dynamical behavior which is significantly enhanced as soon as the laser receives small amounts of optical feedback. Semiconductor lasers are extremely sensitive to feedback and tend to show dynamically unstable, often even chaotic, behavior under certain feedback conditions.

Like other lasers, semiconductor lasers respond with relaxation oscillations when they are suddenly switched on or disturbed in stationary operation. Relaxation oscillations are the damped periodic response of the total carrier density and the optical output intensity as a dynamically coupled system. Both are important for many aspects of semiconductor laser dynamics. The relaxation oscillation frequency determines the maximum modulation frequency of diode lasers. This frequency is strongly governed by the differential gain in the active material, i.e., by the variation of the optical gain with carrier density.

An important application of relaxation oscillations is short pulse generation by gain switching of laser diodes. The idea of gain switching is to inject a short and intense current pulse or an optical pump pulse into the laser structure. This pulse initiates relaxation oscillations but it is so short that already the second oscillation maximum is not amplified. As a consequence, only one intense pulse is emitted. The shortest pulse widths achieved with this scheme so far are in the few picoseconds range. For short pulse generation, a further complexity of the diode laser comes into play. The emitted pulses are observed to be strongly chirped, i.e., the center frequency changes during the pulse. This complex self-phase-modulation is a result of the strong coupling of the real and the imaginary parts of the susceptibility in the semiconductor. If the gain (which is basically given by the imaginary part of the susceptibility) changes during a pulse emission, the refractive index (basically the real part of the susceptibilty) also changes considerably. This leads to a self-phase-modulation if a strong optical pulse is emitted. This strong phase–amplitude coupling is also responsible for the high feedback sensitivity of diode lasers, for complex spatio-temporal dynamics (e.g., self-focusing, filamentation), and for a considerable broadening of the emission linewidth. The phase–amplitude coupling is often quantified with the so-called linewidth enhancement or α factor. This parameter, however, is a function of carrier density, photon wavelength, and temperature.

Diode lasers are also suited for the generation or amplification of extremely short pulses with durations below 1 picosecond. The concept used for such extremely short pulse generation is called mode-locking. The idea of mode-locking is to synchronize the longitudinal modes of the laser so that their superposition is a sequence of short pulses. The most successful concept to achieve mode-locking of diode lasers is to introduce a saturable absorber into the laser cavity and to additionally modulate the injection current synchronously to the cavity roundtrip frequency. A saturable absorber exhibits nonlinear absorption characteristics: the absorption is high for weak intensities and weak for high intensities so that high peak power pulses are supported. Such absorbers can be integrated as a separate section into multiple-section diode lasers which have already been successfully used for subpicosecond pulse generation. Furthermore, electro-absorption modulators, passive waveguide sections, and tunable DBR segments for wavelength tuning can also be integrated into such multiple-section devices.

However, when subpicosecond pulses are generated or amplified in semiconductor laser amplifiers the dynamics of the pulse generation or pulse amplification is strongly governed by complex nonequilibrium carrier dynamics in the semiconductor. The most prominent effects are spectral hole burning and carrier heating. Spectral hole burning occurs if an intense laser field removes carriers from a certain area in k-space faster than carrier–carrier scattering can compensate for. This leads to nonthermal carrier distributions and a gain suppression in certain energy regimes. As mentioned above, a nonthermal carrier distribution relaxes quickly (100 femtoseconds) into a thermal (quasi-equilibrium) distribution if the intense laser field is switched off, but this carrier distribution may have a plasma temperature much higher than the lattice temperature. This effect, which is referred to as carrier heating, also leads to a transient reduction of the gain on a few picoseconds timescale until the carrier distributions have cooled down to the lattice temperature by interactions with phonons. These ultrafast effects are of course particularly important for amplification and generation of short pulses but they generally influence the whole dynamical behavior of diode lasers including, e.g., the modulation behavior.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Chow WW and Koch SW (1999) *Semiconductor-Laser Fundamentals: Physics of the Gain Materials*. Berlin: Springer-Verlag.

Chuang SL (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Delfyett PJ, Dienes A, Heritage JP, Hong MY and Chang YH (1994) Femtosecond hybrid mode-locked semiconductor laser and amplifier dynamics. *Applied Physics* B 58: 183–195.

Haug H and Koch SW (2004) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific Publisher.

Hofmann M (1999) Gain and emission dynamics of semiconductor lasers. *Recent Research Developments, in Applied Physics* 2: 269–290.

Koch SW, Jahnke F and Chow WW (1995) Physics of semiconductor microcavity lasers, review article. *Semiconductor Science and Technology* 10: 739–751.

Osinski M and Buus J (1987) Linewidth broadening factor in semiconductor lasers – an overview. *IEEE Journal of Quantum Electronics* QE-23: 9–29.

Up-Conversion Lasers

A Brenier, University of Lyon, Villeurbanne, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under photoluminescence excitation, a luminescent center located inside a transparent material usually emits at a longer wavelength than the excitation. This is because of the existence of de-excitation losses inside the material. Thus, efficient down-conversion lasers can be built, for example, the commercial Nd^{3+}:$Y_3Al_5O_{12}$ laser which works at 1064 nm, under usual pumping near 800 nm. More surprisingly, it has been demonstrated that fluorescence can also be emitted at shorter wavelength than the excitation. This is possible if the losses are reduced and overcome by so-called 'up conversion mechanisms'. Generally speaking, a laser emitting at a shorter wavelength than the excitation is classified as an up-conversion laser.

These lasers are able to generate coherent light in the red, green, blue, and uv spectral ranges. A point of practical importance is that some of them can be pumped in the near-infrared range (around 800 nm or 980 nm) by commercially available laser diodes. They can be based on all-solid-state technology and share its advantages: compactness of devices, low maintenance, and efficiency. They are needed for a variety of applications: color displays, high density optical data storage, laser printing, biotechnology, submarine communications, gas-laser replacement, etc.

Because the emitted photons have higher energy than the excitation ones, more than one pump photon is needed to generate a single laser photon, so the physical processes involved in the up-conversion laser are essentially nonlinear. Two kinds of laser devices can be considered. In the first one, the excitation of the luminescent center up-migrates through the electronic levels and reaches the initial laser level. This step is due to some luminescence mechanisms. Then, in a second step, lasing at a short wavelength occurs through a transition towards the final laser level. In the second kind of device, the pump excitation activates a down-conversion lasing, but the laser wave is maintained inside the cavity, cannot escape, and is up-converted in short wavelengths by the second-order nonlinear susceptibility of the crystal host. This requires bifunctional nonlinear optical and laser crystals. The device is thus a self-frequency doubling laser or a self-sum frequency mixing laser. If the nonlinearity is not an intrinsic property of the laser crystal but is due to a second nonlinear optical crystal located inside or outside the cavity, the device operates intracavity or extracavity up conversion.

Up-Conversion from Luminescence Mechanisms

The numerous $^{2S+1}L_J$ energy levels ($4f^n$ configuration) of the trivalent rare earth ions, inserted in crystals or glasses, offer many possibilities for up conversion, particularly from long-lived intermediate levels that can be populated with infrared pumping. Crystal field induced transitions between them are comprised of sharp lines, because the $4f^n$ electrons are protected from external electric fields by the $5s^2p^6$ electrons. Their fluorescence spectra can exhibit cross-sections high enough (10^{-19} cm^2) to lead to laser operation in the visible range. For up-conversion purposes, the most popular ions are Er^{3+}, Pr^{3+}, Tm^{3+}, Ho^{3+}, and Nd^{3+}. Up conversion in Er^{3+} was used for the first time in 1959, to detect infrared radiation.

Energy Transfers

Two ions in close proximity interact electrostatically (Coulomb interaction) and can exchange energy. The ion which gives energy is called the sensitizer S or donor, and the energy receiving ion is the activator A or acceptor. Two examples discussed below are shown in Figure 1. The most typical case of such nonradiative energy transfer in trivalent rare earth ions in insulating materials is due to electrical dipole–dipole interaction. For the latter, the transition probability $W(\mathrm{s}^{-1})$ takes the form:

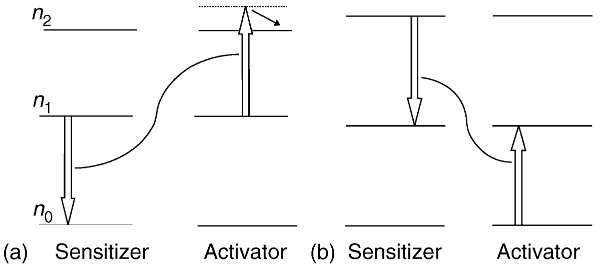

Figure 1 (a) Up-conversion cross-relaxation energy transfer. (b) Down-conversion cross-relaxation energy transfer. The example in (a) is phonon assisted, the example in (b) is resonant.

$$W = \frac{3\hbar^4 c^4}{4\pi n^4}\frac{1}{R^6}\frac{Q_A}{\tau_{\mathrm{rad}}}\int \frac{f_A(E) f_S(E)}{E^4}\,\mathrm{d}E \qquad [1]$$

where R is the S–A distance, E is the photon energy of the transition operated by each ion, n the refractive index, and Q_A is the integrated absorption cross-section of the A-transition:

$$Q_A = \int \sigma_A(E)\mathrm{d}E, \qquad f_A = \frac{\sigma_A}{Q_A} \qquad [2]$$

In eqn [1], τ_{rad} is the radiative emission probability of S for the involved transition calculated from the emission probability A_S:

$$\frac{1}{\tau_{\mathrm{rad}}} = \int A_S(E)\mathrm{d}E, \qquad f_S = \tau_{\mathrm{rad}} A_S \qquad [3]$$

The integral in eqn [1] is called the overlap integral and shows that the transfer is efficient if the luminescence spectrum of S is resonant with the absorption spectrum of A (corresponding to the S and A transitions involved in the nonradiative transfer). The transitions are often not in perfect resonance but remain efficient enough to be used in practice. Then the energy mismatch is compensated by phonons, and the energy transfer is said to be phonon-assisted.

The energy transfer visualized in Figure 1a is an up-conversion cross-relaxation. It is often met in the rare earth ions cited above when they are sufficiently concentrated in the material. The S and A ions can be of the same or of different chemical species. Successive energy transfers can promote the activator from its ground state towards higher energy levels, up to the one from which lasing can occur. This is the so-called 'Addition de Photons par Transferts d'Energie', sometimes appointed 'APTE'. Yb^{3+} is the most used ion as sensitizer and leads to the following transfers in examples of Er^{3+} or Ho^{3+} co-doping:

$$\begin{aligned} &{}^2F_{5/2}(\mathrm{Yb}) + {}^4I_{15/2}(\mathrm{Er}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^4I_{11/2}(\mathrm{Er}) \\ &{}^2F_{5/2}(\mathrm{Yb}) + {}^4I_{11/2}(\mathrm{Er}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^4F_{7/2}(\mathrm{Er}) \end{aligned} \qquad [4]$$

$$\begin{aligned} &{}^2F_{5/2}(\mathrm{Yb}) + {}^5I_8(\mathrm{Ho}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^5I_6(\mathrm{Ho}) \\ &{}^2F_{5/2}(\mathrm{Yb}) + {}^5I_6(\mathrm{Ho}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^5S_2(\mathrm{Ho}) \end{aligned} \qquad [5]$$

If several sensitizers give their energy simultaneously to a single activator, promoting it directly from its ground state towards a high energy level without intermediate levels, the sensitization is said to be 'cooperative'.

The energy transfer shown in Figure 1b is a down-conversion cross-relaxation. It is the inverse of the previous one. At first sight it is, of course, undesirable in an up-conversion laser. However, we notice that one ion in the upper level will lead to two ions in the intermediate level. This fact can be exploited in special conditions (see below) and be beneficial for up conversion.

The dynamical study of the energy transfers, starting from the microscopic point of view contained in eqn [1], is very complicated because of the random locations of the ions or because of other processes, such as energy migration among sensitizers and back transfers. A popular method leading to useful predictions in practice is the rate equation analysis dealing with average parameters. The time evolution of the population densities n_i of the various levels i are described by first-order coupled equations, solved with adequate initial conditions and including the pumping rate. As an example, the rate equations for levels 1 and 2 in Figure 1a take the form:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = -\frac{n_1}{\tau_1} - 2kn_1^2 + \frac{\beta_{21}}{\tau_2}n_2 + W \qquad [6]$$

$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = -\frac{n_2}{\tau_2} + kn_1^2 \qquad [7]$$

$$n_0 + n_1 + n_2 = 1 \qquad [8]$$

where $\tau_{1,2}$ are the lifetimes of levels 1 and 2, k is the up-conversion transfer rate, β_{21} is the branching ratio for the $2 \rightarrow 1$ transition, and W is the pump rate (not shown in Figure 1).

Sequential Two-Photon Absorption

Pumping the ground state of a luminescent ion (Figure 2a) with a wave at angular frequency ω, leads to population with density n_1 of an excited state in a first step. Because the rare earth ions have numerous energy levels, it is possible for the pump to also be resonant with the transition from level 1 towards a higher energy level 2, and is further absorbed. Such a process leading to n_2 up-conversion population is a sequential two-photon absorption. If there is no $1 \rightarrow 2$ resonant transition for the ω frequency, it is possible to use a second pump wave at a different angular frequency ω' resonant with the $1 \rightarrow 2$ transition (Figure 2b). Of course, in practice, it is advantageous when a single pump beam is required for up conversion.

Let us show the relevant terms of the rate equation analysis describing the process in Figure 2a:

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2}n_2 \qquad [9]$$

$$\frac{\mathrm{d}n_2}{\mathrm{d}t} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 \qquad [10]$$

$$n_0 + n_1 + n_2 = 1 \qquad [11]$$

where I is the photon density of the pump (photons $\mathrm{s}^{-1}\,\mathrm{cm}^{-2}$) and σ_0 and σ_1 are the absorption cross-sections (cm^2) of the $0 \rightarrow 1$ and $1 \rightarrow 2$ transitions, respectively. The n_2 population density shows a nonlinear I dependence, most often close to I^2.

Photon Avalanche

In rare earth doped materials, an up-conversion emission, intense enough for visible lasing, can result from a more complex mechanism. In this case, absorption of the pump is not resonant with a transition from the ground state (weak ground-state absorption), but on the contrary, the photon pump energy matches the gap between two excited states (high excited state absorption). These levels are labeled 1 and 2 in Figure 3. Level 1 generally has a rather long lifetime in order to accumulate

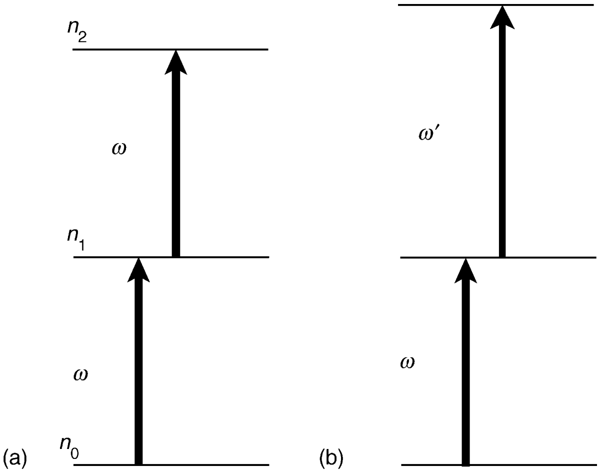

Figure 2 Two-photon absorption up-conversion.

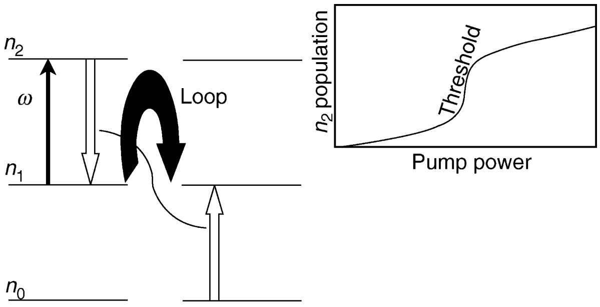

Figure 3 Photon avalanche up-conversion.

population and level 2 is the initial up-conversion laser level. When the pump light is turned on, the populations of levels 1 and 2 first grow slowly, and then can accelerate exponentially. This is generally referred to as 'photon avalanche'. The term 'looping mechanism' also applies, because we can see in Figure 3 that the excited state absorption 1 → 2 is followed by a cross-relaxation energy transfer (quantum efficiency up to two) that returns the system back to level 1. So after one loop, the n_1 population density has been amplified, and is further increased after successive loops. This pumping scheme was discovered in 1979, with $LaCl_3$:Pr^{3+} crystal used as an infrared photon counter.

The rate equation analysis applied to the three-level system in Figure 3 leads to the following time evolutions:

$$\frac{dn_1}{dt} = -\frac{n_1}{\tau_1} + I\sigma_0 n_0 - I\sigma_1 n_1 + \frac{\beta_{21}}{\tau_2} n_2 + 2kn_0 n_2 \quad [12]$$

$$\frac{dn_2}{dt} = -\frac{n_2}{\tau_2} + I\sigma_1 n_1 - kn_0 n_2 \quad [13]$$

$$n_0 + n_1 + n_2 = 1 \quad [14]$$

where the parameters have the same meaning as in eqns [6]–[11].

The steady state limit of the populations at infinite time, upon continuous excitation starting at $t = 0$, can be obtained from canceling the time derivative in the left-hand side of the equation. Let us point out the interesting result. Two different dynamical regimes and analytical expressions for $n_2(\infty)$ are obtained. In the usual case of low pumping rate in the ground state and if the parameters of the systems satisfy: $k < \frac{1-\beta_{21}}{\tau_2}$, then the first regime is observed, whatever the pumping rate $I\sigma_1$ in the excited state.

If the parameters satisfy:

$$k > \frac{1 - \beta_{21}}{\tau_2} \quad [15]$$

the first regime is still observed at low pumping rate $I\sigma_1$ in the excited state, more precisely at pumping rate low enough such that

$$I\sigma_1 < \frac{\left(k + \frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k - \frac{(1 - b_{21})}{\tau_2}}$$

But interestingly, the second regime, the so-called photon avalanche, is obtained at high pumping rate $I\sigma_1$ in the excited state, that is to say if:

$$I\sigma_1 > \frac{\left(k + \frac{1}{\tau_2}\right)\frac{1}{\tau_1}}{k - \frac{(1 - b_{21})}{\tau_2}} \quad [16]$$

So the right-hand side of eqn [16] appears to be a pump threshold separating two dynamical regimes when the condition in eqn [15] is satisfied, as the insert in Figure 3 shows.

Another way to distinguish between the two regimes is given by linearization of the system of eqns [12]–[14] ($n_0 \cong 1$). Then the description will only be valid close to $t = 0$. We know from standard mathematical methods that the solutions of eqns [12]–[14] are then a combination of $\exp(\alpha t)$ terms. If the threshold condition in eqn [16] is satisfied, a parameter α in the combination becomes positive, so the population of level 2 increases exponentially (photon avalanche) at early times and the dynamics of the system is unstable.

The physical meaning of eqns [15] and [16] is clarified by introducing the yield η_{CR} of the cross-relaxation process:

$$\eta_{CR} = \frac{k}{k + 1/\tau_2} \quad [17]$$

and the yield η of the 2 → 1 de-excitation:

$$\eta = \frac{b_{21}/\tau_2}{k + 1/\tau_2} \quad [18]$$

Then, the threshold condition in eqn [16] can be written as:

$$\left(\frac{I\sigma_1}{I\sigma_1 + 1/\tau_1}\right)(2\eta_{CR} + \eta) > 1 \quad [19]$$

Starting with one ion in level 1, the first parenthesis in eqn [19] is the yield with which it will be promoted to level 2, and the second parenthesis is the yield of the backward path 2 → 1. So, the left-hand side of eqn [19] is the new number of ions after a looping path 1 → 2 → 1, which has to be higher than 1 for the avalanche to occur. If we consider a lot of successive loops, the origin of the exponential behavior is clear. The role of the first ion in level 1 (and the weak pumping rate in the ground state $I\sigma_0$) is limited to initialize the process. Note that the value of the first parenthesis in eqn [19] is less than 1, so the second parenthesis has to be higher than 1. This condition is equivalent to eqn [15].

Materials: The Energy Gap Law

As we can see from the above mechanisms, up conversion needs intermediate energy levels and is not efficient if these cannot accumulate populations due to a short lifetime. The spontaneous rate of de-excitation of an electronic level is the sum of a radiative part W^{rad} and of a nonradiative part W^{nrad}. The radiative part can be obtained through the absorption cross-section corresponding to the transition or from the Judd–Ofelt theory which also exploits the absorption spectrum. The nonradiative rate is then obtained by subtracting W^{rad} from the measured fluorescence decay rate of the level. Based on the measurements in many hosts for the different trivalent rare earth ions, at different temperatures, the nonradiative decay rate of a J-level towards the next lower J′-level was found to have the form:

$$W_{\text{JJ}'}^{\text{nrad}} = B\exp(-\beta\Delta E_{\text{JJ}'})\left(1-\exp\left(-\frac{\omega_{\max}}{kT}\right)\right)^{-p} \qquad [20]$$

where $\Delta E_{\text{JJ}'}$ is the energy gap between the two levels, $p = \Delta E_{\text{JJ}'}/\omega_{\max}$, $\omega_{\max}$ being the energy of an effective optical phonon of the host. The values of B, β and $\omega_{\max}$ are found by fitting with experimental data. An example of such a fit is represented in Figure 4.

Equation [20] is the well-known and familiar 'energy gap law'. We have represented it for several oxide and nonoxide hosts at $T = 300$ K in Figure 5. It is clear that chloride, bromide, fluoride will be favorable hosts for up conversion due to their low phonon energy ($LaBr_3$: $\omega_{\max} = 175\ \text{cm}^{-1}$, $LiYF_4$: $\omega_{\max} = 400\ \text{cm}^{-1}$, $Y_3Al_5O_{12}$: $\omega_{\max} = 700\ \text{cm}^{-1}$, silicate glass: $\omega_{\max} = 1100\ \text{cm}^{-1}$). Let us mention also a fluorozirconate glass, ZBLAN, widely used in up-conversion fiber lasers.

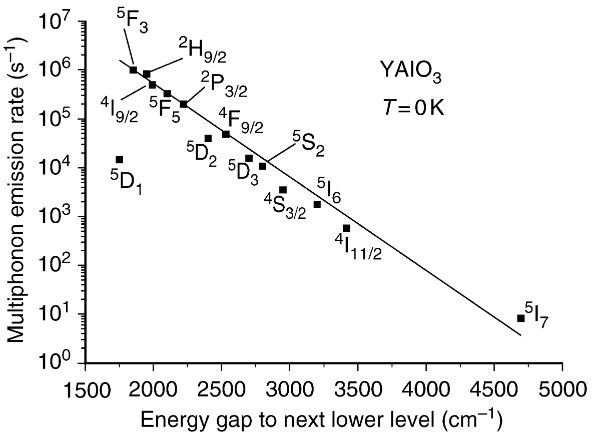

Figure 4 Energy-gap dependence of nonradiative decay rates in $YAlO_3$.

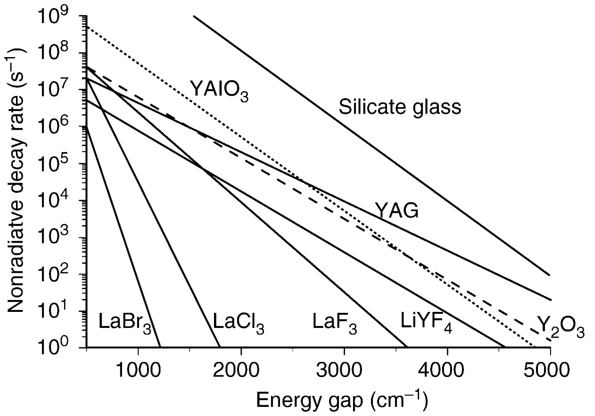

Figure 5 Energy-gap law for different hosts at 300 K.

Up-Conversion from Second-Order Optical Nonlinearity

The inter-atomic electric field acting on the electrons inside a medium has a magnitude of $\sim 10^8$ V/cm and derives from a potential that is anharmonic. The electric field **E** of an electromagnetic wave propagating through the medium drives the electrons beyond the quadratic region of the potential in the case of high **E**. So, the electron response and the associated polarization **P** take the form of a function of **E**:

$$\mathbf{P}(\mathbf{E}) = \mathscr{E}_0(\chi^{(1)} : \mathbf{E} + \chi^{(2)} : \mathbf{E}^2 + \chi^{(3)} : \mathbf{E}^3 + \cdots) \qquad [21]$$

The different terms in eqn [21] are responsible for many optical phenomena because in Maxwell electromagnetic theory, the polarization is the source of the waves. We have selected in Table 1 the terms relevant for linear effects and for the purpose of frequency up conversions: second-harmonic generation (SHG) and sum-frequency mixing, due to second-order terms. It should be noticed that the latter have a nonzero value only in noncentrosymmetric materials.

Figure 6 visualizes the up conversion from SHG: two photons of the fundamental field at angular frequency ω_1 are annihilated while one photon at twice the angular frequency is created. A similar picture could be drawn for sum-frequency up conversion: SHG is the degenerate case where $\omega_1 = \omega_2$.

The second-order nonlinear processes have an efficiency given by an effective coefficient d_{eff} given hereafter:

$$d_{\text{eff}} = \sum_{ijk} e_i(\omega_3) d_{ijk} e_j(\omega_1) e_k(\omega_2) \qquad [22]$$

where $e_i(\omega_l)$ is the ith component of the unit vector of the electric field of the wave l at angular frequency ω_l and $d_{ijk} = (1/2)\chi_{ijk}$.

Table 1 Relevant terms for frequency up-conversion

Optical process	*Term*	*New wave*
Linear		
Refractive index, absorption, stimulated emission	$\chi^{(1)}(-\omega;\omega)$	
Second order nonlinear		
Second harmonic generation	$\chi^{(2)}(-\omega_3;\omega_1,\omega_1)$	$\omega_3 = 2\omega_1$
Sum-frequency mixing	$\chi^{(2)}(-\omega_3;\omega_1,\pm\omega_2)$	$\omega_3 = \omega_1 + \omega_2$

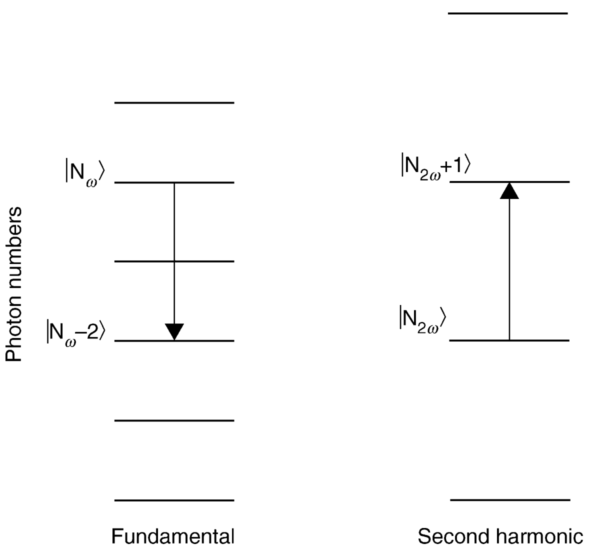

Figure 6 Up-conversion from second harmonic generation.

Moreover, the wave propagation equations in the slow varying envelope approximation, given hereafter for the E_3 electric field of the sum-frequency mixing wave:

$$\frac{\mathrm{d}E_3}{\mathrm{d}z} = \frac{\mathrm{i}\omega_3 d_{\mathrm{eff}}}{n_3 c} E_1 E_2 \quad \exp(\mathrm{i}(k_1 + k_2 - k_3)z) \qquad [23]$$

$$(\text{with } k_j = \omega_j n_j / c)$$

impose that the frequency conversion is efficient in practice, only if the following phase matching condition is satisfied:

$$\omega_1 n_1 \uparrow (\theta,\varphi) + \omega_2 n_2 \uparrow (\theta,\varphi) = \omega_3 n_3 \downarrow (\theta,\varphi) \qquad [24]$$

where $n_i \uparrow$ and $n_i \downarrow$ are respectively the upper and lower refractive index in the direction of propagation (θ,φ) (n_i in eqn [23] are the same with a simplified notation). Equation [24] is restricted here for simplicity to collinear type I (waves 1 and 2 and the same polarization) and is the expression of photon momentum conservation. It is usually achieved from the birefringence of the crystals. When this is not possible, another technique can be used, namely quasi-phase matching, which involves reversing periodically the sign of the nonlinear optical coefficient.

In bifunctional crystals, the laser effect and the $\chi^{(2)}$ interaction occur simultaneously inside the same crystal. In the case of the self-frequency doubling laser, the angular frequency $\omega_1 = \omega_2$ in eqn [24] is that of the fundamental laser wave. In the case of the self-sum frequency mixing laser, ω_1 corresponds to the fundamental laser wave and ω_2 to the pump wave. As we can see from eqns [22]–[24], the polarization of the laser emission is crucial in order for the device to work. So, the laser stimulated emission cross-section σ for propagation in the phase matching direction (θ,ϕ) in the adequate polarization has to be evaluated. This can be performed with the formula:

$$\sigma(\theta,\varphi) = \frac{\sigma_X \sigma_Y \sigma_Z}{\sqrt{\sigma_Y^2 \sigma_Z^2 e_X^2 + \sigma_X^2 \sigma_Z^2 e_Y^2 + \sigma_X^2 \sigma_Y^2 e_Z^2}} \qquad [25]$$

where $\sigma_{X,Y,Z}$ are the emission cross-sections for X, Y, Z-polarizations and $e_{X,Y,Z}$ are the components of the unit vector or the electric field of the laser wave.

To summarize, the conditions necessary for a crystal to be a bi-functional material are: (i) it must be noncentrosymmetric; (ii) it must accept fluorescent doping (usually with Nd^{3+} or Yb^{3+}); and (iii) it must be phase matchable for its laser emission. In practice, only a few crystals satisfy these requirements and have been tried with some success. The main ones are $LiNbO_3$, $LaBGeO_5$, $Ba_2NaNb_5O_{15}$, β'-$Gd_2(MoO_4)_3$, $YAl_3(BO_3)_4$, $GdAl_3(BO_3)_4$, $CaY_4(BO_3)_3O$, and $CaGd_4(BO_3)_3O$.

Up-Conversion Lasers Based on Luminescence Mechanisms in Materials

Energy Transfers

Up-conversion lasers from energy transfers have the advantage of using a single pump laser. This latter populates firstly a long-lived intermediate level. A typical example, based on $LiYF_4$:Er^{3+} crystal, is provided by Figure 7.

The pump at 802 nm from a GaAlAs diode laser, or at 969 nm, populates efficiently at the 8 ms lifetime $^4I_{11/2}$ level. The green laser, at 551 nm, corresponds to the $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition and it can be shown that energy transfer is an essential part of the mechanism by stopping abruptly the pump laser: green lasing continues to occur for several hundred microseconds because the $^4S_{3/2}$ level is still fed (an excited-state absorption would stop immediately). The involved energy transfer is:

$$^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \qquad [26]$$

and leads to a 20 mW laser threshold and 2.3 mW green output power at 50 K temperature upon 95 mW of incident power.

By using mirrors with high transmission in the green range but having high reflectivity at blue wavelengths, blue laser emission at 470 nm can be sustained, corresponding to the $^2P_{3/2} \rightarrow {}^4I_{11/2}$ transition. It originates from a mechanism involving three up-conversion energy transfers:

$$\begin{aligned} &^4I_{11/2} + {}^4I_{11/2} \rightarrow {}^4F_{7/2} + {}^4I_{15/2} \\ &^4I_{11/2} + {}^4I_{13/2} \rightarrow {}^4F_{9/2} + {}^4I_{15/2} \qquad [27] \\ &^4S_{3/2} + {}^4F_{9/2} \rightarrow {}^4I_{15/2} + {}^2K_{13/2} \end{aligned}$$

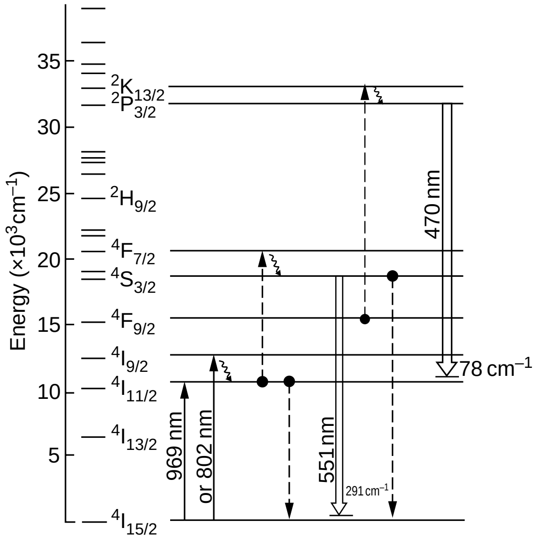

Figure 7 Energy level diagram of LiYF$_4$:Er^{3+}. The dashed lines indicate the energy transfers responsible for the green and blue up-conversion lasing. Reproduced with permission of American Institute of Physics from Macfarlane RM, Tong F, Silversmith AJ and Lenth W (1988) Violet CW neodymium upconversion lasers. *Applied Physics Letters* 52(16): 1300.

The illustration in Figure 7 shows up-conversion energy transfers between ions of the same chemical species, but co-doping the laser host with two ions of different chemical species provides the opportunity of separating their roles: one species is devoted to pump absorption (this is the sensitizer) and the other one is devoted to lasing (this is the activator). Due to the development of efficient semiconductor InGaAs diodes emitting near 980 nm, the most widely used sensitizer is Yb^{3+}. Table 2 provides a list of several up-conversion lasers based on energy transfer mechanisms, operating at room temperature. We can see that among the most efficient lasers, are those based on rare-earth ions doped glass fibers. The reason is that in fibers, the pump and laser waves remain confined inside a core (typically 5 μm diameter) and have high energy densities over a long length (tenths of cms), which is favorable for all up-conversion mechanisms.

Sequential Two-Photon Absorption

An example of an up-conversion laser pumped by a two-photon absorption mechanism is represented in Figure 8. The material is LaF$_3$:Nd^{3+} and violet lasing at 380 nm corresponds to the $^4D_{3/2} \rightarrow {}^4I_{11/2}$ transition.

Transition from the ground state up to the $^4F_{5/2}$ level is firstly obtained from an infrared beam at 790 nm in order to feed the $^4F_{3/2}$ intermediate level after a fast nonradiative de-excitation. The $^4F_{3/2}$ level has a rather long lifetime: 700 μs, so it accumulates

Table 2 Main room temperature up-conversion lasers operating from energy transfers

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY$_{1.4}$Yb$_{0.59}$Ho$_{0.01}$F$_8$	670	1540 + 1054	
BaYYb$_{0.998}$Tm$_{0.002}$F$_8$	649	1054	1%
BaYYb$_{0.99}$Tm$_{0.01}$F$_8$	799–649–510–455	960	
BaYYb$_{0.99}$Tm$_{0.01}$F$_8$	348	960	
LiYF$_4$:Er(1%):Yb(3%)	551	966	37 mW
LiY$_{0.89}$Yb$_{0.1}$Tm$_{0.01}$F$_4$	810–792	969	80 mW
LiY$_{0.89}$Yb$_{0.1}$Tm$_{0.01}$F$_4$	650	969	5 mW
LiKYF$_5$:Er (1%)	550	808	150 mW
Fiber:Yb:Pr	635	849	20 mW
Fiber:Yb:Pr	635	1016	6.2 mW
Fiber:Yb:Pr	521	833	0.7 mW
Fiber:Yb:Pr	635	860	4 mW

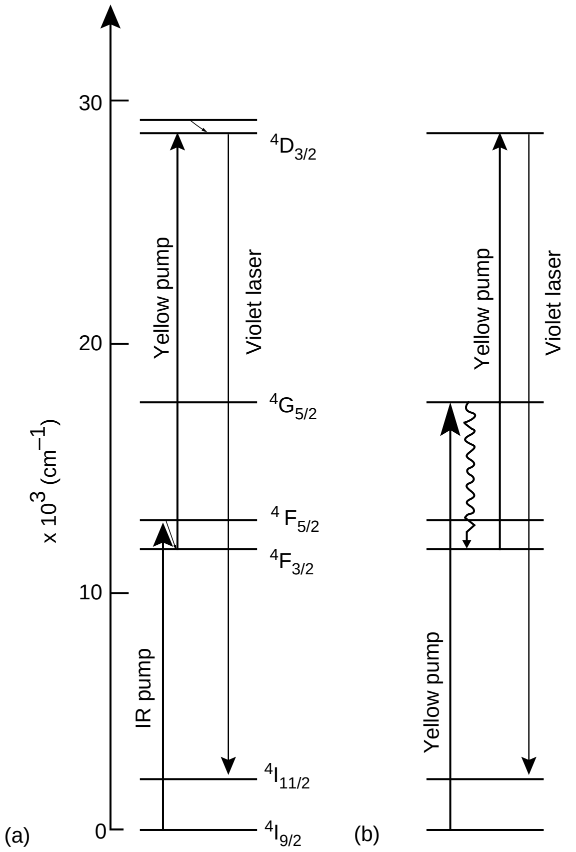

Figure 8 Simplified energy level diagram of $LaF_3:Nd^{3+}$ and up-conversion lasing from two-photon absorption. (a) Two different pump wavelengths, (b) doubly resonant single pump wavelength. Reproduced with permission of American Institute of Physics from Lenth W, Silversmith AJ and Macfarlane RM (1988) Green infrared pumped erbium up conversion lasers. In: TAM AC, Gole JL and Stwalley WC (eds) *Advances in Laser Science – III*. AIP Conference Proceedings n°172: 8–12.

enough population that a second yellow pumping at 591 nm (Figure 8a) can efficiently populate the $^4D_{3/2}$ initial laser level. With 110 mW of infrared pump and 300 mW of yellow pump, 12 mW of violet output were obtained at 20 K temperature.

A simpler device is obtained if a single-pump beam is used. This is the case in Figure 8b, where a yellow pump at 578 nm provides both ground state and excited state absorption. This doubly resonant scheme is less efficient in $LaF_3:Nd^{3+}$ than the previous one but in Figure 9, we show another example: $LiYF_4:Er^{3+}$, working at room temperature. The lasing $^4S_{3/2} \rightarrow {}^4I_{15/2}$ transition at 551 nm delivers 20 mW upon 1600 mW pumping at 974 nm.

The simplified Er^{3+} level scheme of Figure 9 contains four main levels with population densities: $n_0 = [^4I_{15/2}]$, $n_1 = [^4I_{13/2}]$, $n_2 = [^4I_{11/2}]$, $n_3 = [^4S_{3/2}]$. The rate equations of populations of this system can

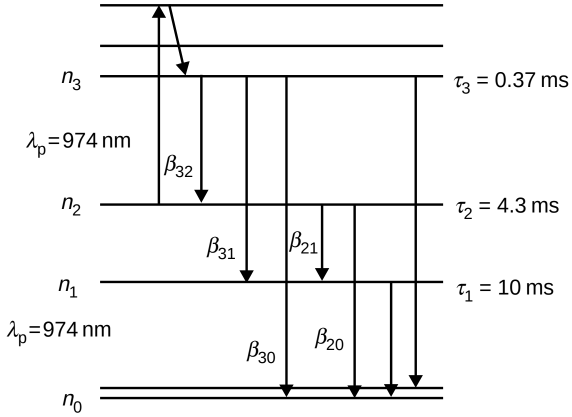

Figure 9 Simplified energy level diagram of $LiYF_4:Er^{3+}$ and up-conversion lasing from two-photon absorption (doubly resonant). Reproduced with permission of Institute of Physics Publishing from Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*, Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing, 19.

Table 3 Main room temperature up-conversion lasers operating from two-photon absorption

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
$Y_3Al_5O_{12}$:Er 1%	561	647 + 810	
$LiYF_4$:Er 1%	551	647 + 810	0.95 mJ
$LiYF_4$:Er 1%	551	810	40 mW
$LiYF_4$:Er 1%	551	974	45 mW
$LiYF_4$:Er	551	974	20 mW
KYF_4:Er 1%	562	647 + 810	0.95 mJ
$LiLuF_4$:Er	552	974	70 mW
$LiLuF_4$:Er	552	970	213 mW
$LiYF_4$:Tm 1%	453–450	781 + 649	0.2 mJ
Fiber:Tm	480	1120	57 mW
Fiber:Tm	480	1100–1180	45 mW
Fiber:Tm	455	645 + 1064	3 mW
Fiber:Tm	803–816	1064	1.2 W
Fiber:Tm	482	1123	120 mW
Fiber:Yb:Tm	650	1120	
Fiber:Ho	540–553	643	38 mW
Fiber:Ho	544–549	645	20 mW
Fiber:Er	548	800	15 mW
Fiber:Er	546	801	23 mW
Fiber:Er	544	971	12 mW
Fiber:Er	543	800	
Fiber:Pr	635	1010 + 835	180 mW
Fiber:Pr	605	1010 + 835	30 mW
Fiber:Pr	635	1020 + 840	54 mW
Fiber:Pr	520	1020 + 840	20 mW
Fiber:Pr	491	1020 + 840	7 mW
Fiber:Nd	381	590	74 μW
Fiber:Nd	412	590	500 μW

be written and solved because all the implied parameters (lifetimes, branching ratios, ground, and excited states absorption cross-sections) have been measured in this material. Predictive models of the up-conversion green laser are successful, matching experimental measurements, and confirm that the doubly resonant mechanism is realistic at low doping concentration.

Table 3 summarizes the performances of the main up-conversion lasers based on sequential photon absorption and working at room temperature.

Photon Avalanche

Photon avalanche up-conversion was observed mainly in Nd^{3+}, Tm^{3+}, Er^{3+}, and Pr^{3+} doped materials. As an illustration of this mechanism, let us consider the case of the continuous wave Pr^{3+}–Yb^{3+}-doped fluorozirconate fiber laser (3000 ppm Pr, 20 000 ppm Yb). With two Ti:sapphire pump lasers operating at 852 and 826 nm (Figure 10), an output power at 635 nm as high as 1.02 W was obtained with an incident pump power of 5.51 W (19% slope efficiency).

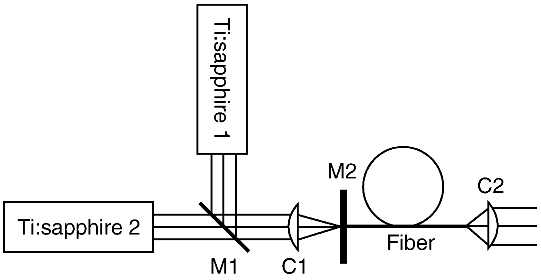

Figure 10 Experimental set-up for Yb–Pr-doped fiber up-conversion laser. M1: dichroic mirror, C1, C2: lenses, M2 dichroic input mirror. Reproduced with permission of Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

This remarkable result can be explained quantitatively by modeling the photon avalanche with a five-level diagram and restricted in Figure 11 to a single laser pump at 850 nm. The laser emission corresponds to the $^3P_0 \rightarrow {}^3F_2$ transition. The five relevant levels have the population densities: $n_0 = [^3H_4(\mathrm{Pr})]$, $n_1 = [^1G_4(\mathrm{Pr})]$, $n_2 = [^3P_0(\mathrm{Pr})]$, $n_3 = [^2F_{7/2}(\mathrm{Yb})]$, $n_4 = [^2F_{5/2}(\mathrm{Yb})]$.

The two components of the mechanism are:

(i) the excited state absorption of the pump by the spin allowed transition: $^1G_4 \rightarrow {}^1I_6$

(ii) the process leading to the doubling of the $^1G_4$ population, in this instance through two energy transfers:

$$\begin{gathered} {}^1I_6(\mathrm{Pr}) + {}^2F_{7/2}(\mathrm{Yb}) \rightarrow {}^1G_4(\mathrm{Pr}) + {}^2F_{5/2}(\mathrm{Yb}) \\ {}^2F_{5/2}(\mathrm{Yb}) + {}^3H_4(\mathrm{Pr}) \rightarrow {}^2F_{7/2}(\mathrm{Yb}) + {}^1G_4(\mathrm{Pr}) \end{gathered} \qquad [28]$$

The rate equation analysis predicts that the avalanche threshold occurs at about 1 W pump and the long fluorescence rise time is somewhat shortened at higher pump power, as observed experimentally.

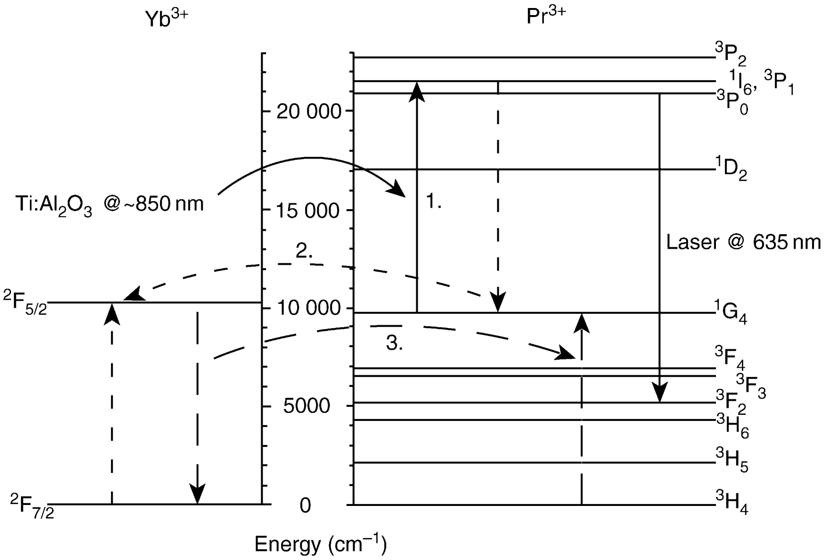

Figure 11 Energy level diagram of fluorozirconate: Yb^{3+}:Pr^{3+} fiber. The dashed lines indicate the energy transfers. Reproduced with permission of the Optical Society of America from Sandrock T, Scheife H, Heumann E and Huber G (1997) High power continuous wave up-conversion fiber laser at room temperature. *Optics Letters* 22(11): 809.

Table 4 Main room temperature up-conversion lasers operating from photon avalanche

Laser material	*Laser wavelength (nm)*	*Pump wavelength (nm)*	*Output power*
BaY_2F_8:Yb:Pr	607.5	822	55 mW
BaY_2F_8:Yb:Pr	638.7	841	26 mW
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	720	830	1%
$LiY_{0.89}Yb_{0.1}Pr_{0.01}F_4$	639.5	830	
Fiber:Yb:Pr	635–637	780–880	300 mW
Fiber:Yb:Pr	605–622	780–880	44 mW
Fiber:Yb:Pr	517–540	780–820	20 mW
Fiber:Yb:Pr	491–493	780–880	4 mW
Fiber:Yb:Pr	635	850	1.02 W
Fiber:Yb:Pr	635	850	2.06 W
Fiber:Yb:Pr	520	850	0.3 W

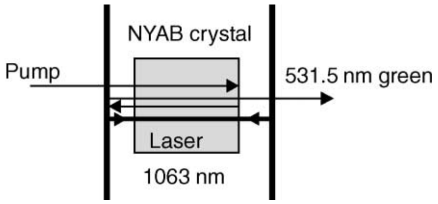

Figure 12 Scheme of a self-frequency doubling laser. Note that in reality the three beams are superimposed.

Table 4 summarizes the performances of the main up-conversion lasers based on photon avalanche mechanisms.

Up-Conversion Lasers Based on Bi-Functional Crystals

The laser stimulated emission in a bi-functional crystal was first obtained in 1969 with Tm^{3+} in $LiNbO_3$ but the most used ion is, of course, Nd^{3+} working in the two channels:

$$^4F_{3/2} \rightarrow {}^4I_{11/2} \quad [29]$$

$$^4F_{3/2} \rightarrow {}^4I_{13/2} \quad [30]$$

and more recently Yb^{3+} working in the channel:

$$^2F_{5/2} \rightarrow {}^2F_{7/2} \quad [31]$$

Channels in eqns [29] and [31] operate around 1060 nm wavelength and the channel in eqn [30] near 1338 nm.

The Self-Frequency Doubling Laser

A typical laser scheme is shown in Figure 12.

Table 5 Main results of Nd^{3+} and Yb^{3+} green self-frequency doubling lasers near 530 nm (eqns [29] and [31])

Crystal	*Input power (mW)*	*Output power (mW)*	*Pumping*
$LiNbO_3$:MgO:Nd	215	1	Dye laser
$YAl_3(BO_3)_4$:Nd	870	10	Laser diode
$YAl_3(BO_3)_4$:Nd	280	3	Laser diode
$YAl_3(BO_3)_4$:Nd	400	69	Laser diode
$YAl_3(BO_3)_4$:Nd	1380	51	Laser diode
$YAl_3(BO_3)_4$:Nd	369	35	Laser diode
$LiNbO_3$:MgO:Nd	850	18	Dye laser
$LiNbO_3$:Sc_2O_3:Nd	65	0.14	Ti:sapphire laser
$LiNbO_3$:MgO:Nd	100	0.2	Laser diode
$YAl_3(BO_3)_4$:Nd	1600	225	Laser diode
$YAl_3(BO_3)_4$:Nd	2200	450	Ti:sapphire laser
$(Y,Lu)Al_3(BO_3)_4$:Nd	880	24	Laser diode
$LiNbO_3$:ZnO:Nd	430	0.65	Ti:sapphire laser
$Ba_2NaNb_5O_{15}$:Nd	270	46	Ti:sapphire laser
$CaY_4(BO_3)_3O$:Nd	900	62	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1600	192	Ti:sapphire laser
$CaGd_4(BO_3)_3O$:Nd	1250	115	Laser diode
$CaGd_4(BO_3)_3O$:Nd	1560	225	Ti:sapphire laser
$YAl_3(BO_3)_4$:Yb	11 000	1100	Laser diode
$GdAl_3(BO_3)_4$:Nd	2.8 mJ/pulse	0.12 mJ/pulse	Pulsed dye laser

The laser beam cannot escape from the cavity and dichroic mirrors are used. The input mirror has high transmission at the pump wavelength and is highly reflective at laser- and second-harmonic wavelengths. The output mirror is highly reflective at laser wavelength and has high transmission for second harmonic. Channels in eqns [29] and [31] generate green light near 530 nm and the channel in eqn [30] generates red light near 669 nm.

The most exploited bi-functional crystal is $YAl_3(BO_3)_4:Nd^{3+}$, well-known as NYAB. It is a negative uniaxial trigonal crystal (extraordinary index lower than the ordinary one) and its laser emission is easily observed in ordinary polarization with a high cross-section: $2 \times 10^{-19}\ cm^2$ at 1063 nm. Type I phase matching occurs at a polar angle $\theta = 30.7°$ and 26.8° for channels in eqns [29] and [30] respectively. The effective nonlinear optical coefficient d_{eff} is close to 1.4 pm V^{-1} at azimuthal angle $\phi = 0°$.

The Yb^{3+} ion has some advantages over the Nd^{3+} ion due to its very simple energy level scheme: there is no excited state absorption at the laser wavelength, no up-conversion losses, no concentration quenching, and no absorption in the green. The small Stokes shift between pump and laser emission reduces the thermal loading of the material during laser operation. A self-doubling laser based on $YAl_3(BO_3)_4:Yb^{3+}$ of crystal has produced 1.1 W green power upon 11 W diode pumping.

The drawback of $YAl_3(BO_3)_4$ is that it is rather difficult to grow because it is not congruent. This difficulty was overcome by the discovery at Ecole Nationale Supérieure de Chimie de Paris of the rare-earth calcium oxoborate $CaGd_4(BO_3)_3O$ which can be grown to a large size by the Czokhralski method. It is a monoclinic biaxial crystal and doped with Nd^{3+}, it was firstly exploited (channel in eqn [29]) in the *XY* principal plane at $\theta = 90°$, $\phi = 46°$ with ($d_{eff} = 0.5$ pm V^{-1}). It was soon recognized that the optimum phase matching direction ($d_{eff} = 1.68$ pm × V^{-1}) occurred out of the principal planes, in the direction $\theta = 66.8°$, $\phi = 132.6°$, and high green power can be obtained: 225 mW under 1.56 W pump.

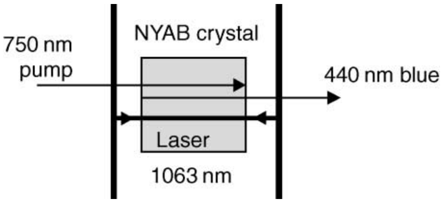

Figure 13 Scheme of a self-sum frequency mixing laser. Note that in reality the three beams are superimposed.

Table 5 summarizes the main self-frequency doubling results obtained with the channel in eqn [29].

The Self-Sum Frequency Mixing Laser

The example in Figure 13 gives the main features of such a laser. The input mirror has high transmission at the pump wavelength and both mirrors are highly reflective at laser wavelength. The output mirror has high transmission at sum frequency mixing wavelength.

The laser wave at angular frequency ω_1 in Table 1 corresponds to the channel in eqns [29] or [30] in the case of Nd^{3+}. The wave at angular frequency ω_2 in Table 1 has two different roles: first it excites the Nd^{3+} laser center and secondly its nonabsorbed part is up converted by the second-order nonlinear process. ω_2 (corresponding to the wavelength λ_2) is then chosen to match the main Nd^{3+} absorption lines:

$$\begin{aligned} &{}^4I_{9/2} \rightarrow {}^4G_{5/2} - {}^2G_{7/2}\ (\lambda_2 = 590\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{7/2} - {}^4S_{3/2}\ (\lambda_2 = 750\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{5/2} - {}^2H_{9/2}\ (\lambda = 800\ \text{nm}) \\ &{}^4I_{9/2} \rightarrow {}^4F_{3/2}\ (\lambda_2 = 880\ \text{nm}) \end{aligned} \qquad [32]$$

The most efficient self-frequency mixing lasers based on channels in eqns [29]–[31] are gathered in Table 6.

See also

Materials Characterization Techniques: $\chi^{(2)}$. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation.

Table 6 Main results of Nd^{3+} self-sum frequency mixing lasers based on eqn [29]

Crystal	*Pump wavelength (nm)*	*Generated wavelength (nm)*	*Output power*
$YAl_3(BO_3)_4$:Nd	740–760	436–443	0.16 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	740–760	436–443	0.403 mJ/pulse
$CaGd_4(BO_3)_3O$:Nd	811	465	1 mW
$YAl_3(BO_3)_4$:Nd	585–600	377–383	0.25 mJ/pulse
$GdAl_3(BO_3)_4$:Nd	587–597	378–382	0.105 mJ/pulse
$YAl_3(BO_3)_4$:Nd	488, 515	330, 380	0.2 mW

Further Reading

Brenier A (2000) The self-doubling and summing lasers: overview and modeling. *Journal of Luminescence* 91: 121–132.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Huber G (1999) Visible cw solid-state lasers. *Advances in Lasers and Applications*. Bristol and Philadelphia: Scottish Universities Summer School in Physics & Institute of Physics Publishing.

Joubert MF (1999) Photon avalanche up-conversion in rare earth laser materials. *Optical Materials* 11: 181–203.

Kaminskii AA (1996) *Crystalline Lasers: Physical Processes and Operating Schemes*. CRC Press, Inc.

Lenth W and Macfarlane RM (1992) Upconversion lasers. *Optics & Photonics News* 3: 8–15.

Powell RC (1998) *Physics of Solid-State Laser Materials*. New York: Springer-Verlag.

Risk WP, Gosnell T and Nurmikko AV (2003) *Compact Blue-Green Lasers*. Cambridge, UK: Cambridge University Press.

Weber MJ (2001) *Handbook of Lasers*. Boca Raton, FL: CRC Press.

Yariv A (1995) *Quantum Electronics,* Third Edition. John Wiley & Sons.

LASER-INDUCED DAMAGE OF OPTICAL MATERIALS

A J Glass and A H Guenther, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Shortly after the invention of the laser, in the early 1960s, it became apparent that laser performance was limited by the ability of optical materials to withstand high energy and power densities. Broadly speaking, laser damage arises from two causes: (i) thermal effects, which are associated with high average power; and (ii) dielectric breakdown, which is associated with high values of the local electric field. Early researchers quickly recognized that most of the observed damage effects arose from extrinsic causes, such as impurities and inclusions in the bulk material, or particles or defects on exposed optical surfaces. At that time, optical materials had not been developed for high-power use, and lacked the quality needed for laser applications. In fact, in many situations, the laser provided the inspection light source used to assess the properties of the materials, such as absorption, defects, scattering centers, surface topography, and dynamic behavior. Thus, the field of laser damage studies was broadened to include many aspects of material preparation and characterization. At high-power densities, nonlinear effects become important, often limiting performance without creating any irreversible change in the material's properties. The most general definition of laser-induced damage includes all effects arising in optical materials and components that limit or degrade the performance of laser systems.

The pursuit of improved performance at high values of power and energy densities was motivated by several factors and applications. In general, improved tolerance to high flux and fluence leads to smaller, lighter, and more efficient systems. These are particularly important for military and energy applications. Since the 1970s, the laser fusion program alone has produced an extensive body of research on laser-induced damage in high-power nanosecond-scale systems, which has enabled orders of magnitude improvement in the damage resistance and extraction efficiency of glass lasers. More recently, with the development of fiber laser systems producing and transporting a kilowatt of average power in a fiber of less than 100 μm diameter, laser damage phenomena on an entirely different spatial scale have become important.

Since the early 1970s, the literature of laser damage research has been collected in a series of annual publications of the Boulder Damage Symposium, initiated by the authors of this article, which was first held in 1969. Recently the papers presented at the Damage Symposium in the years 1969 to 1998 were collected and published in CD form by the SPIE. A CD containing the proceedings of the next five years (1999–2003) will be released by SPIE this year. Also in the late 1960s, in the Soviet Union, Academician Aleksei Bonch-Bruevich launched a comprehensive research project on the nonresonant interaction of laser radiation with matter, opening up a new branch of photonics research, known in the Soviet Union as power optics. In 1969, he founded the All-Union Conference on Interaction of Optical Radiation with Matter (AURICOM), which was held every two to three years until 1990. After 1990, the conference was renamed Nonresonant Laser-Matter Interaction (NLMI), and chaired by Professor Mikhail Libenson. NLMI was held in St Petersburg in 1996, 2000, and

2003. The proceedings of the 1996 AURICOM and subsequent NLMI conferences have been published by the SPIE. We are saddened to note that Professor Libenson died in February 2004.

The Boulder Damage Symposium grew out of a one-day Symposium convened in Boulder in 1969 under the auspices of the ASTM Subcommittee on Lasers, to establish standards for laser materials. In their summary of the first Symposium, the conference organizers remarked:

> The entire question of the nature of the standards remains to be addressed. Should they take the form of energy density at which catastrophic damage is likely to occur, or should they be specified in terms of a mean number of shots at a given level of energy density before certain degradation of performance is measured? The latter would seem to express the kind of information that the buyer of laser materials would find most useful. Once a standard is agreed upon, a meaningful test configuration must be established. It is likely that this will be a well-controlled oscillator amplifier chain with the test specimen an active element of the system, but this is not certain. Another point to be determined is that of the accuracy to which its specifications should be written. This depends in part on the kind of quality control the producers of laser glass believe is feasible.

The questions raised in this first Symposium remain at the core of laser damage studies today. How do we define damage, in terms of the physical effect on the material or in terms of the degradation of performance of the system? How do we measure damage thresholds or values? How do we specify the properties and preparation of the sample, including the relevant experimental and material variables? How can we ensure that damage values are reproducible, unless we make the measurements in a very well controlled and characterized system? With what precision should damage levels be stated, in order to be useful? Extensive research was required in order to answer these questions, and to elucidate the fundamental nature of damage mechanisms. Many of the results of this research are documented in the 35 year record of the Boulder Damage Symposium and the corresponding Russian-based conferences.

From the beginning of this area of research, investigators devoted a significant effort to establishing the 'first causes' of damage phenomena. However, they quickly realized that in most cases, damage occurred at surfaces and material interfaces, or was mediated by impurities or inclusions in bulk materials. As material fabrication processes improved, and higher-quality materials became available, observed damage thresholds approached the intrinsic limits of the pure material. This was particularly true of silicate glass, metal reflectors, and in some cases, polymers. In fact, silicate glass has become one of the purest noncrystalline materials available in bulk quantities. However, for practical purposes, most laser systems are limited by damage at surfaces, coatings, or impurities. Impurities aggregate at grain boundaries in crystalline materials, and coalesce into inclusions in glasses, resulting in absorption, scattering, and field enhancement.

The coherent nature of laser light makes any scattering phenomena potentially damaging, because the scattered light interferes constructively with the laser beam, leading to the formation of intensity hot spots. Every interface in an optical system is potentially a source of constructive interference between the reflected and incident waves. Multilayer dielectric coatings are particularly vulnerable to damage, due to the presence of residual stress as well as field enhancement from Fresnel reflection. This has led to attempts to mitigate the problem by use of specific designs, such as avoiding AR coatings by using Brewster-angle surfaces, or use of nonquarter wave layers to move the peak field away from the vicinity of boundaries where impurities tend to collect. Other damage-sensitive design considerations can be found in the literature. Multiple reflections are the basis of photonic crystal optics, and applications of these devices to high-power density systems will undoubtedly be limited by field-enhanced damage phenomena.

In CW operation, of course, thermal effects dominate. They may be irreversible, creating permanent damage sites in the medium, or reversible, such as color centers that can be bleached out. As laser pulse durations vary from nanosecond to picosecond and femtosecond durations, the nature of the dominant damage mechanisms changes as well. As a rule of thumb, the scaling of damage thresholds as $\tau^{1/2}$, where τ is the pulse duration, was established empirically, and was found to be valid over a wide range of pulse durations. This scaling implies the presence of diffusion effects, either of heat or of electrons, in the vicinity of a damage center. As might be expected, this scaling does not apply for very short pulses (femtoseconds or less), where the extremely high field values can lead to instantaneous breakdown of the material.

Whether the damage is due to field enhancement or is impurity-mediated, it is difficult to establish a reproducible value for a damage threshold, unless the sample is macroscopic in extent, and the sampling laser beam is reproducibly of high spatial and temporal quality, preferably truly single-mode. Damage measurements on microscopic areas are unlikely to be representative of the entire sample. Damage will occur first at a point where the field is strongest or the material is weakest, and a sufficient area must be tested to ensure that it includes such sites. Also, unless the beam quality of the laser used for testing is well controlled and well characterized, it will not be possible to reproduce the test results. It is equally important to characterize the samples, including their method of preparation, cleaning, and test environment. For damage research, as opposed to simple threshold testing, postmortem examination is also required, to elucidate the mechanism of the observed damage effect. However, we now have several techniques, such as photothermal deflection, that can provide useful pre-catastrophic indicators of the onset of laser damage, and thus are appropriate for quality control.

Professor Roger Wood had published a series of reviews and textbooks on laser damage in optical materials, which provide an overview of the field, with appropriate emphasis on measurement techniques, characterization of the test setup and sample, and implications for system design. He has also edited a collection of seminal publications in the field, published by the SPIE in 1990.

It is gratifying to those working in this field to observe how much of the research carried out to elucidate and, if possible, avoid the deleterious effects of laser damage has been applied to the constructive purpose of laser-assisted material processing and manufacturing. The introduction of low-cost, reliable femtosecond lasers has stimulated activity in this field, since it enables the experimenter to control to exquisite precision the power and energy deposited on a material. With femtosecond pulses, one can ablate material from the surface of an object while minimizing any collateral damage to the interior material. Surface layers of dirt can be removed from priceless art works such as oil paintings, without concern for damage to the object itself. Objects with arbitrary shape can be built up from metal particles, using laser additive manufacturing techniques. Thin films and photonic crystal structures can be created using laser-assisted deposition. Laser irradiation provides a power tool for surface cleaning, conditioning, and annealing. Laser interaction with materials and laser-assisted manufacturing techniques comprise a vital area of research in the materials science community.

In both the US and Russia, study of the interaction of intense coherent light with optical materials began with a disarmingly simple objective; to understand the fundamental mechanisms of interaction. In the review and summary of the 1969 ASTM Symposium, the editors wrote:

> In view of the number of problems remaining to be resolved, it is suggested that another Symposium on laser damage be held in 1970. Hopefully, at that time a better understanding of the nature of damage in laser glass will have been obtained. Higher threshold values will have been reproducibly achieved, and some agreement can be arrived at regarding useful and realistic standards.

Thirty-five years later, damage levels in optical materials have been significantly improved, and realistic and useful standards have been developed under the auspices of the International Organization for Standardization (ISO), and are now accepted globally. However, as the operating regimes of lasers have moved to femtosecond temporal scale and micrometer dimensions, new phenomena have arisen that require further research. An entirely new technology of laser-enabled manufacturing and materials conditioning has developed, based on the same physics as laser damage. Combining the disciplines of materials science and applied optics, the interaction of intense light with materials remains a scientifically challenging and economically important field of research.

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Bonch-Bruevich AM, Konov VI and Libenson MN (eds) (1990) *Optical Radiation Interaction with Matter*. SPIE Proceedings, vol. 1440.

Glass AJ and Guenther AH (eds) (1969) *Damage in Laser Glass Symposium: Review and Summary*. ASTM Special Technical Publication 469.

ISO Online Catalogue, Section 31.260, http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList

Laser-Induced Damage in Optical Materials, Collected Papers 1969–1998, (CD-ROM), SPIE, Bellingham, Washington (1999).

Laser-Induced Damage in Optical Materials, Collected Papers 1999–2003, (CD-ROM), SPIE, Bellingham, Washington (2004) (To be published).

Libenson MN (ed.) (1997) *Non-Resonant Laser-Matter Interaction (NLMI-9)*. SPIE Proceedings, vol. 3093.

Libenson MN (ed.) (2001) *Non-Resonant Laser-Matter Interaction (NLMI-10)*. SPIE Proceedings, vol. 4423.

Libenson MN (ed.) (2004) *Non-Resonant Laser-Matter Interaction (NLMI-11)*. SPIE Proceedings, vol. 5506, (To be published).

Wood RM (ed.) (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

Wood RM (ed.) (1990) *Selected Papers on Laser Damage in Optical Material*. SPIE Milestone Series, vol. MS24. SPIE Optical Engineering Press.

Wood RM (ed.) (2003) *Laser-Induced Damage in Optical Materials*. Bristol, UK: Institute of Physics.

Wood RM (ed.) (2003) *The Power- and Energy-Handling Capability of Optical Materials, Components, and Systems*. SPIE Press Tutorial Texts in Optical Engineering, vol. TT60.

LIGHT EMITTING DIODES

J Schanda, University of Veszprém, Veszprém, Hungary

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Short Historic Overview

During the second half of the nineteenth century, electric lighting became possible. In 1879, Edison demonstrated his incandescent lamp, and although some electric arc lamps competed with the incandescent lamp bulb for a few years, the first 50 years of electrical lighting was dominated by the incandescent lamp. Gas discharge lamps (both low- and high-pressure lamps) became more widely distributed only by the end of the third decade of the twentieth century, and widespread distribution of the fluorescent lamp, that still dominates our offices and many other lighting applications today, was generally accepted only after the Second World War.

The early incandescent lamps had an efficacy of a few lm/W, but by 1920 this could be increased by up to 10–15 lm/W. Gas discharge lamps started with approximately 30 lm/W, but by 1960 their efficacy increased to about 70 lm/W. Fluorescent lamps still provide only 70 to 100 lm/W; high-pressure lamp efficacy could be increased above this value. The different lamp types can be used, however, in different applications, depending on the size and the unit power of the lamp type. Nonincandescent light generation in solid-state materials dates back to the 1920s: a faint glow of a SiC crystal was observed when current passed through it at a point contact. In the 1930s, electroluminescence of ZnS powder layers was detected (Destriaux effect). By the mid-1960s, it became clear that by using the then available technology, good luminous efficiency could be expected only in III–V compounds (GaAs, GaP, and their ternary and quaternary compounds using Al, In, and N in the compounds). Pure GaAs has a bandgap in the infrared (IR), thus one can build good IR-emitters using this material, but no visible light emitter. However, GaAs proved to be an excellent coherent light-emitting material and enabled the construction of IR-emitting semiconductor lasers.

First visible light-emitting diodes were fabricated using a GaAsP alloy. Figure 1 shows how the efficacy of III–V compound LEDs can be increased by introducing new compositions, and later, new technologies and better light extraction techniques. In the figure we also show the efficacy values of average incandescent, high-pressure Hg- and fluorescent lamps. The first LEDs emitted in the red part of the spectrum, and only by 1975 was the green LED invented. By 1985, practically any color between red and green became possible, and the efficacy of these lamps reached a level where for monochromatic light applications, their efficacy could compete favorably with filtered incandescent light. LEDs emit in a relatively narrow wavelength band (20 nm–30 nm bandwidth), and if one has to filter from the continuous spectrum of an incandescent lamp such a narrowband, the loss becomes considerable. LEDs needed only a few volts to function and it became general practice to build LEDs with approximately a 20 mA current load. These LEDs became very popular for interior signaling (on instruments, household appliances, and in similar applications). The goal was, however, to be able to produce white light, but the blue-emitting LED was still missing. Only after 1990 could the problem of producing p–n junctions from GaN, that had a bandgap large enough to emit blue light, be solved. Different ternary and quaternary compounds, as shown in Figure 1, enabled efficacies comparable to incandescent lamp efficacy, and the race for producing white light began.

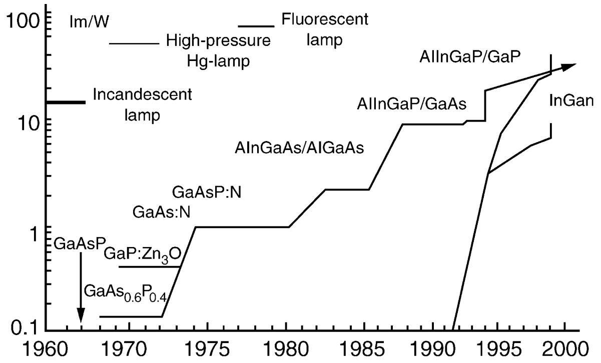

Figure 1 Increase of luminous efficacy of LEDs during the past 40 years. Courtesy of Lumileds Lighting.

Physical Fundamentals

Figure 2 shows a forward-biased p–n junction. As injected free electrons and holes are driven towards the junction, they can radiatively recombine, emitting a photon. The probability of photon emission versus nonradiative recombination will depend on the band

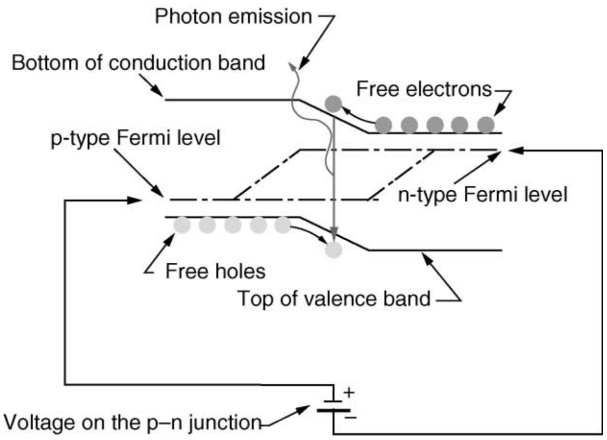

Figure 2 Forward biased p–n junction, light generation via recombination of free charge carriers.

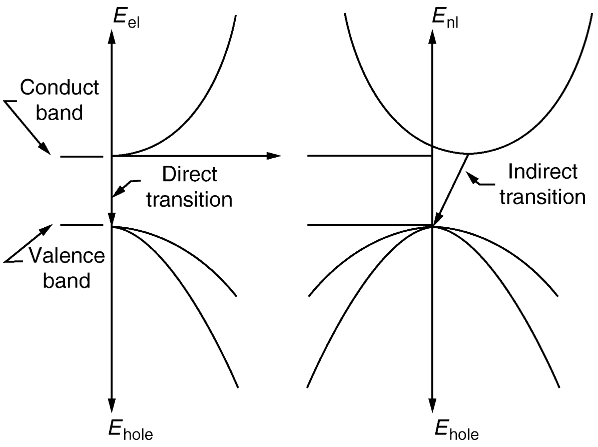

Figure 3 Band to band recombination in direct and indirect band semiconductor.

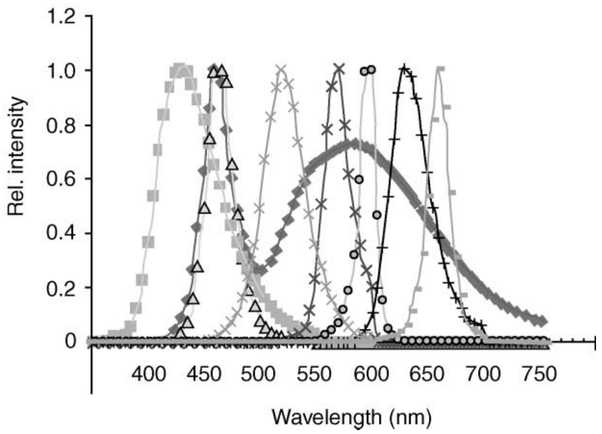

Figure 4 Relative spectral power distribution of some LED structures.

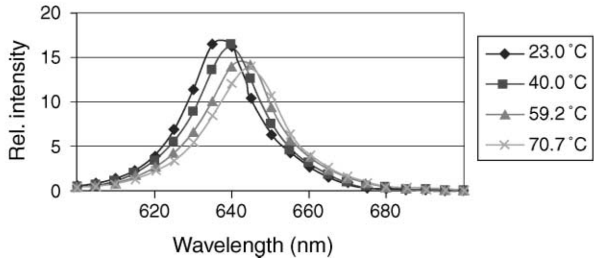

Figure 5 Spectral power distribution of a red LED at different temperatures.

structure of the semiconductor. Figure 3 shows the two possibilities of a direct and an indirect bandgap semiconductor. GaAs is, for example, a direct bandgap material. In this material free electrons at the bottom of the conduction band can recombine with free holes at the top of the valence band, as the momentum of the two ‘particles’ is the same and a photon emission takes care of the principle of energy conservation. GaP is an indirect bandgap material where the momentum of the free electrons at the bottom of the conduction band differs from the momentum of the free holes at the top of the valence band, and thus the recombination needs not only the fulfilment of the conservation of energy, but also of momentum. Thus, besides photon emission, phonon emission or absorption has to take place. Thus, the probability of the transition will decrease and competing nonradiative processes will become important. Using N doping produces very shallow donor states and from these an efficient recombination can take place. This enabled the production of early relatively efficient green-emitting LEDs.

GaN is ideal for producing LEDs, as it is a direct gap material and has a high bandgap. The technological difficulties are, however, enormous. Modern high-efficiency LEDs use (Al)InGaN-based materials for shorter wavelength LEDs, and AlInGaP-based materials for longer wavelength LEDs. Figure 4 shows a number of LED spectra of different wavelengths including a white LED. Its spectrum consists of a blue maximum produced by the LED emission and a yellow broader band produced by the luminescent material excited by the blue emission of the LED.

LEDs are semiconductor materials, and as with every semiconductor the LED characteristics are also temperature dependent. Not only the efficiency of the current carrier–photon emission process depends on the temperature, but the bandgap of the base material also depends on the temperature, and thus both the efficacy and the relative spectral distribution of the emission change with temperature. Figure 5 shows the spectral power distribution of a red LED at different temperatures.

Modern High Brightness LEDs

Besides the high quantum efficiency of the semiconductor junction, two further requirements to achieve high luminous efficacy are the efficient light extraction from the junction and keeping the temperature of the semiconductor chip low.

All semiconductors for producing LEDs have a high refractive index, thus light reflection at the semiconductor–air interface is relatively high, and for obliquely impinging rays the critical angle for total internal reflection is soon reached. Different manufacturers shaped the semiconductor chip in different forms to secure maximum light extraction. Figure 6 shows the cross-section of an LED, where the side planes have been cut obliquely to secure good light extraction.

The other problem that has to be solved is good heat conduction from the chip to a heat sink. At the time of writing, highest single chip power loads were in the order of 5 W. The structure of such a construction is seen in Figure 7. The GaN-based LED is grown on a visible light transparent sapphire (Al_2O_3) substrate. In modern LEDs the structure itself is quite complicated; in the vicinity of the p–n junction some nanometer thick layers are deposited forming quantum wells (QW(s)). The p-type original upper layer is metalized and soldered to the submount serving as heat-sink. The n-type layer gets a second soldered heat conducting current leading to the submount. As seen, the original structure grown on the sapphire substrate is inverted, and now the light can be extracted through the sapphire layer. The metalization on the p-type layer serves at the same time as a mirror. High brightness LEDs are now available in package sizes of over 100 lumen.

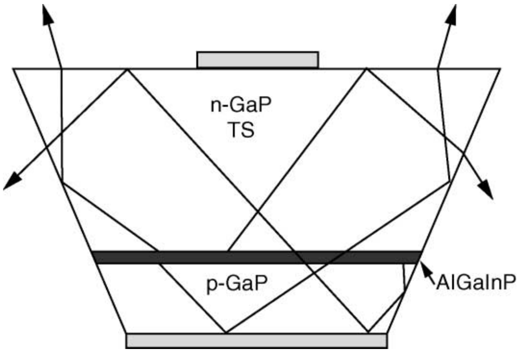

Figure 6 Truncated pyramid structure to secure good light extraction from the LED chip. Courtesy of Lumileds Lighting.

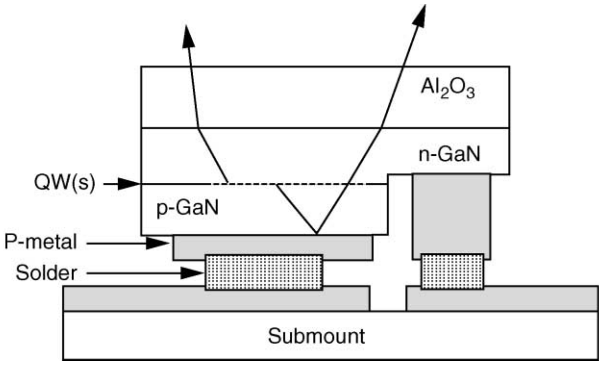

Figure 7 Inverted, flip-chip structure for good heat-sinking of the LED. Courtesy of Lumileds Lighting.

White LED Structures

The advantages of LEDs (low drive voltage, extreme long lifetime of 50000 to 100000 hours, robust shock-resistant construction) can open up a new era of lighting engineering, if white light emission can be obtained. The LED structure in itself is a narrowband emitter, thus – with inorganic LEDs – it is not possible to produce white light within one chip. Organic LEDs (OLEDs) promise the production of broadband and white light emission, but these are still in an experimental phase. At present they are used in smaller, portable displays, such as those used on portable phones, cam-coders, etc. To get further information, go to the Further Reading section at the end of this article.

At present three techniques to generate white light get thorough investigation: first, mixing the light of a red, a green, and a blue LED. This should theoretically give the highest efficiency; the problem is, however, that the three LEDs have different temperature dependence and aging characteristics, and thus it is difficult to keep the chromaticity of the mixed color constant (the three colored LEDs can be used in LED-based displays to produce full-color reproduction). At present the following technique is widely used. A blue LED excites a yellow phosphor, so that the mixture of the blue and yellow lights produce a white appearance. The color rendering of such white LEDs is not too good, and therefore alternative techniques using a green and a red phosphor are receiving increasing attention. The third method uses the same principle, but with an LED emitting in the near ultraviolet. This has the advantage that it provides a greater selection of phosphors, but as the energy difference between the UV photon and the photon emitted by the phosphor is larger than in the case of a blue-emitting LED, the theoretical efficiency is lower.

The phosphor layer is deposited on the chip surface, and for blue-emitting diodes special care has to be taken that the LED light and the phosphor light are properly mixed so that the emitted light shows no directional color shift. Early white LEDs suffered much from the fact that the spectral distribution of the light changed considerably with viewing angle.

The problems with phosphor-coated white LEDs are how a suitable correlated color temperature and color rendering can be obtained. Early white LEDs had very bluish chromaticity, their correlated color temperature was very high, above 7000 K. Preferred white light has different chromaticity in different cultural regions, in the Far East 6000 K to 6500 K is preferred, in Europe and North America for general lighting a correlated color temperature of 4000 K is well accepted and for social and home lighting even lower correlated color temperatures, 2800 K to 3200 K are preferred. Recently, white LEDs based on blue chips and special phosphor blends have been produced that meet these requirements.

A further problem is the color rendering of the white LEDs. The Commission Internationale d'Eclairage (CIE), the International Commission on Illumination, provides standards, recommendations, and technical reports in the field of light and lighting, including vision, color, and image processing. Photometry and colorimetry are based on the standards and recommendations of the CIE.

The CIE established a method of calculating a color-rendering index in 1965 and updated it in 1974. Despite several attempts, the lighting community has been unable to agree on a revision of this outdated document since the 1970s. Thus, color-rendering indices are still calculated based on this method, although several investigations have shown that the calculated indices do not correspond with the visual observation. CIE, therefore, established a technical committee (TC) dealing only with the question of LED color rendering. It is hoped that this TC will be able to come up with some recommendations soon and, therefore, that LED manufacturers can tailor the emission spectra of their LEDs for optimal efficacy and color rendering. Until this happens, however, the published color rendering indices, especially the General Color Rendering Index (R_a) have to be handled carefully. The color appearance of objects under LED illumination might deviate considerably from what one expects based on the R_a index.

Application of LEDs

LEDs are the solution for practically every signaling application. Figure 8 shows an LED traffic light. Here it is also shown that, despite the fact that LEDs have extremely long life expectancy, the failure of the electronic circuitry can lead to unpleasant dark spots and rows in the display.

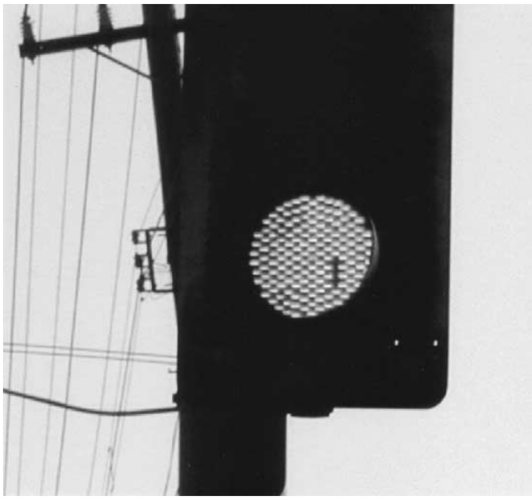

Figure 8 Green LED traffic light. Due to an electronic failure some LEDs stay dark.

LEDs are increasingly used in large active posters and displays. In such applications the highly saturated color of LED light is an advantage: LED displays can be very bright and emit saturated colors.

LEDs have an advantage in every situation where colored light has to be produced. Thus, they become the preferred automotive stoplights, dashboard lights, and are widely used in other signaling applications on the roads, at airfields, and in other traffic situations; much decorative lighting is now using LEDs.

Experiments are underway to find the best applications for white LEDs. Their drawback at the moment is still that their single unit power is below ten watts, as we often need ten to a hundred times larger units. For such applications, some LED clusters have been built, but these are still not competitive with other light sources. At places where the ruggedness of the LEDs is of an advantage, white LED lights are used – for example, in aircraft reading lights. One area where we will probably soon find higher-intensity white LEDs is car headlamps. Both the efficacy and the luminous flux output of the single LED units is increasing continuously. Applications, unbelievable only a few years ago, can now be realized efficiently with LEDs. Thus it is very difficult to foresee in which applications – beyond signaling – LEDs will first supersede other light sources, but it is certain that within the next decade we will see many new applications of LED sources.

List of Units and Nomenclature

Chromatic adaptation	Adaptation by stimuli in which the dominant effect is that of different relative spectral distributions.
Chromaticity	Partial description of the color stimulus, describing only chromatic aspects and not the 'intensity' (the luminance) of the stimulus. In light sources, chromaticity is an important descriptor, as the level of illumination can be changed by changing the distance between the source and the illuminated object. The luminance of the object will depend on the illumination and the reflection (transmission) characteristics of the object.
Color	In color science one distinguishes between color (perception), characteristic of visual perception that can be described by attributes of hue, brightness (or lightness) and colorfulness (or saturation or chroma), and color stimulus, a specification of the optical stimulus in terms of operationally defined values, such as three tristimulus values, or luminance, and chromaticity (or dominant wavelength and saturation).
Color rendering	Measure of the degree to which the psychophysical color of an object illuminated by the test illuminant conforms to that of the same object illuminated by the reference illuminant, suitable allowance having been made for the state of chromatic adaptation.
Correlated color temperature	The temperature of the Planckian radiator whose perceived color most closely resembles that of a given stimulus at the same brightness and under specified viewing conditions.
Efficacy (luminous efficacy of a source)	Quotient of the luminous flux emitted by the power consumed by the source.
Luminous efficiency	Ratio of radiant flux weighted according to $V(\lambda)$, the spectral luminous efficiency of the human eye, to the corresponding radiant flux.
Spectral luminous efficiency (of the human observer)	Ratio of the radiant flux at wavelength λ_m to that at wavelength λ, such that both radiations produce equally intense luminous sensations under specified photometric conditions and λ_m is chosen so that the maximum value of this ratio is equal to 1.

See also

Incoherent Sources: Lamps. **Lasers:** Semiconductor Lasers.

Further Reading

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Boston, MA: Artech House.

Rudden MN and Wilson J (1993) *Elements of Solid State Physics*. Chichester, UK: Wiley.

Sze SM (1985) *Semiconductor Devices: Physics and Technology*. New York: Wiley.

Wenckebach WT (1999) *Essentials of Semiconductor Physics*. Chichester, UK: Wiley.

Žukauskas A, Shur MS and Gaska R (2002) *Introduction to Solid-State Lighting*. New York: Wiley.

M

MAGNETO-OPTICS

Contents

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping
Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping

H Pascher, Universität Bayreuth, Bayreuth, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A complete understanding of the band structure of a semiconductor close to the Fermi energy requires information on the fundamental energy gap, the effective masses and the effective g-factors of the valence and conduction band. Magneto-optical methods in principle allow one to determine all these quantities; they are particularly well suited to measure the spin splittings of the energy bands.

Faraday rotation is the rotation of the plane of polarization of linearly polarized light propagating through a sample parallel to an external magnetic field. The magneto-optical Kerr effect is the analog for reflection. Though in general it yields much smaller angles of rotation than the Faraday effect, it has to be applied in spectral ranges where the investigated sample is not transparent.

If a semiconductor is excited by circularly polarized light with photon energies not too high above the energy gap a spin polarization in the conduction band can be obtained (optical pumping). A small transverse magnetic field reduces the degree of this spin polarization (the Hanle effect). Owing to the spin polarization of the photo-excited conduction electrons the photoluminescence radiation is partly circularly polarized. Thus mounting the sample in a microwave resonator and applying a longitudinal magnetic field, electron spin resonance transitions can be detected optically by monitoring the degree of polarization of the photoluminescence (ODCESR: optically detected conduction electron spin resonance). Via a 'nuclear Hanle effect' mediated by the hyperfine interaction, nuclear magnetic resonance transitions can also be detected optically (ODNMR). ODCESR and ODNMR are summarized under ODMR (optically detected magnetic resonance).

In modern optics nonlinear optical effects have gained particular importance. Since electron spin resonances in semiconductors are Raman-active transitions, coherent Raman methods such as CARS (coherent anti-Stokes Raman scattering) or CSRS (coherent Stokes Raman scattering) are of great importance for the determination of spin splittings of energy bands and the investigation of magnetic structures in semiconductors.

In the present contribution the physical basis for these methods will be discussed and typical examples for their application will be presented.

For Faraday rotation and CARS the systems from which these examples will be taken are magnetic or dilute magnetic semiconductors. In these materials paramagnetic ions like Mn^{++} or Eu^{++} are incorporated. There are two types of exchange interactions: (i) between the magnetic ions which in most cases causes an antiferromagnetic coupling; and (ii) between the ions and free carriers which causes a strong modification of the spin splittings of the energy bands with respect to nonmagnetic materials. Thus the Faraday effect or CARS are particularly well-suited methods for the investigation of such systems.

As examples of optical pumping and ODMR, results obtained with III–V semiconductors will be presented, which show the most important applications of these experimental methods.

Faraday Rotation and Magneto-optical Kerr Effect

Basic Principles

As already pointed out in the Introduction, the Faraday effect is the rotation of the plane of polarization of linearly polarized light traversing a sample parallel to a magnetic field. The angle of rotation Θ_{Far} is proportional to the length l of the sample and to the strength B of the magnetic field:

$$\Theta_{\mathrm{Far}} = VlB \qquad [1]$$

where V is called Verdet's constant.

For a phenomenological understanding of the effect one can consider plane polarized light as a superposition of right- and left-hand circularly polarized components. If the indices of refraction are different for both components, then at the end of the sample, when the two components are recombined, a phase shift occurs which manifests itself as a rotation of the polarization plane. The angle resulting from this consideration is

$$\Theta_{\mathrm{Far}} = \frac{l\omega}{2c}(n_- - n_+) \qquad [2]$$

where ω is the frequency, c the velocity of light, and n_- and n_+ are the refractive indices for left- and right-hand polarized light, respectively. If one neglects ellipticity of the light behind the sample only the real parts of the complex refractive indices $\tilde{n}_\pm$ are important,

$$\tilde{n}_\pm = n_\pm - \mathrm{i}\kappa_\pm \qquad [3]$$

The imaginary parts, however, are needed to understand the magneto-optical Kerr effect which occurs in the light reflected from the sample surface. According to Fresnel's formula for vertical incidence the field amplitude of the reflected beam is:

$$\tilde{E}^{\mathrm{r}}_\pm = \frac{1 - \tilde{n}_\pm}{1 + \tilde{n}_\pm}\tilde{E}^{\mathrm{i}}_\pm = \tilde{r}_\pm \tilde{E}^{\mathrm{i}}_\pm \qquad [4]$$

where the superscript 'r' refers to the reflected beam and 'i' to the incident beam. The complex reflection coefficient can be split into an amplitude and a phase factor:

$$\tilde{r}_\pm = r_\pm \exp(\mathrm{i}\psi_\pm) \qquad [5]$$

By comparing eqns [4] and [5] the phase shifts during reflection can be expressed in terms of the refractive index:

$$\psi_\pm = \arctan\frac{2\kappa_\pm}{1 - n_\pm^2 - \kappa_\pm^2} \qquad [6]$$

Finally the Kerr angle of rotation is obtained from the phase shifts in an analogous way to the Faraday rotation:

$$\Theta_{\mathrm{Kerr}} = \tfrac{1}{2}(\psi_- - \psi_+) \qquad [7]$$

It is seen that both effects can be understood in terms of the refractive index. In a simple oscillator model the refractive index has the form given in eqn [8]:

$$\tilde{n}_\pm^2 = \tilde{\varepsilon}_\pm(\omega, B) = \varepsilon_\infty + \sum_{i,j}\frac{F_{i,j}^\pm}{\hbar\omega_{i,j}^\pm - \hbar\omega - \frac{1}{2}\mathrm{i}\Gamma_{i,j}^\pm} \qquad [8]$$

with $\hbar\omega_{i,j}^\pm$ representing the transition energies between the energy levels i and j in σ^+ and σ^- polarization, respectively, and $F_{i,j}^\pm$ and $\Gamma_{i,j}^\pm$ are the respective oscillator strength and linewidth parameter.

From eqn [8] it is seen that when tuning the frequency, ω, resonant enhancements of the rotation angles are expected whenever the frequency corresponds to an allowed electronic transition, e.g., an exciton in a wide-gap semiconductor. This is the reason why the Faraday and magneto-optical Kerr effect can yield information on interband transition energies. In the case of such a resonance, the refractive index will be dominated by the particular term in the sum, corresponding to that transition. By a Taylor expansion one can then show that the difference of the refractive indices for σ^+ and σ^- polarization and thus the angles of rotation (see eqns [2] and [7]) are proportional to the difference of the transition energies for right- and left-hand circularly polarized light, and this difference depends on the spin splittings of the valence and conduction band. So we understand why the most important information obtained from Faraday or magneto-optical Kerr effect studies is on the spin splittings of the energy bands.

A quantum mechanical description of the relation between the material properties and the magneto-optical effects beyond the simple oscillator model given here is published by Boswarva *et al.* and the reader is referred to this paper for further studies. A matrix formalism to describe the effects for a multilayer system has been developed by Nies and Kessler.

Examples of Experimental Results

The straightforward method to measure Faraday rotation is to put the sample between polarizers and sweep the magnetic field. Then the transmitted intensity is proportional to a $\cos^2$ function. Subsequent maxima demonstrate a rotation by 180°. There have been numerous successful investigations done in this way, but the possibilities of detecting small angles of rotation are very limited in this method. Therefore differential methods have been developed either using a polarizing beam splitter, or a photo-elastic modulator which periodically switches the state of polarization. Rotation angles of the order of 10^{-4} rad can be detected in this way.

Figure 1 shows the Faraday rotation of an epitaxial film of $Pb_{0.933}Eu_{0.067}Te$ (thickness 11 μm) for two temperatures. The plot demonstrates behavior typical for a diluted magnetic semiconductor in its paramagnetic phase. The spin splittings of the valence and conduction band and accordingly the Faraday rotation are proportional to the magnetization of the sample which behaves like a Brillouin function. (The lines without noise are fitted Brillouin functions.)

Examples of the application of the magneto-optical Kerr effect to the investigation of semiconductors, particularly multilayer systems, are published, e.g., by Pohlt *et al.* and Krenn *et al.*

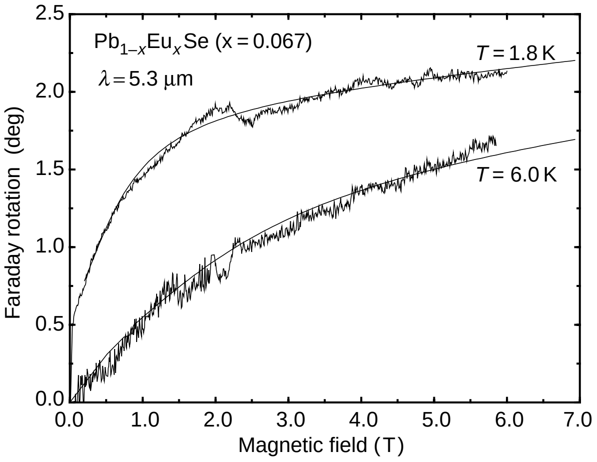

Figure 1 Faraday rotation as a function of magnetic field of a PbEuSe epitaxial film for two temperatures. Lines with noise: experimental recordings; lines without noise: fitted Brillouin functions.

Coherent Raman Scattering

Basic Principles

A more direct method for the determination of spin splittings of energy bands is provided by Raman scattering. There is a huge number of investigations of spin transitions in semiconductors using spontaneous Raman scattering (for a review of spontaneous Raman scattering in dilute magnetic semiconductors see the article by Ramdas and Rodriguez). However, there are several advantages in the spectroscopic application of the nonlinear Raman techniques compared to spontaneous Raman scattering:

(i) The scattered radiation is emitted as a collimated beam which is more easily detectable than the light spontaneously scattered into the entire solid angle. This advantage is particularly important for the investigation of narrow-gap semiconductors which have to be investigated by infrared light, a spectral range where the detectors are less sensitive than in the visible.
(ii) Fluorescence can be strongly suppressed by spatial filtering.
(iii) The spectral resolution is determined by lasers instead of a classical spectrometer.
(iv) Raman scattering cross-sections can be calculated from the intensity of a collimated beam without taking care of the aperture of the detecting system.

The coherent Raman techniques are special cases of optical four-wave mixing where two laser beams with frequencies ω_L and ω_S ($\omega_L > \omega_S$) are superimposed in the sample. Due to nonlinear interactions radiation with new frequencies is generated, among them:

$$\omega_{CARS} = 2\omega_L - \omega_S \qquad [9]$$

If one identifies ω_L with the frequency of the pump laser and ω_S with that of the scattered Stokes radiation of a spontaneous Raman experiment, then ω_{CARS} is the frequency of the anti-Stokes radiation. For the intensity of the radiation generated at ω_{CARS} one finds:

$$I_{CARS} = \frac{9}{16 c_0^4 \varepsilon_0^2} \frac{\omega_{CARS}^2}{n_{CARS} n_S n_L^2} \left|\chi^{(3)}\right|^2 l^2 I_S I_L^2 \times \left[\frac{\sin(\overrightarrow{\Delta k}\,\vec{l}/2)}{\overrightarrow{\Delta k}\,\vec{l}/2}\right]^2 \qquad [10]$$

where c_0 is the velocity of light, l is the length of the sample, I denotes the intensity, and n the refractive index at the respective frequency. $\chi^{(3)}_{kjkj}$ is the third-order nonlinear susceptibility of the scattering medium and

$$\overrightarrow{\Delta k} = 2\vec{k}_{\mathrm{L}} - \vec{k}_{\mathrm{S}} - \vec{k}_{\mathrm{CARS}} \qquad [11]$$

is the phase mismatch. $\overrightarrow{\Delta k} = 0$ can be achieved by noncollinear alignment of the laser beams. However, for very thin epitaxial layers with $l \ll \pi/\Delta k$ the phase factor in eqn [10] can be approximated by 1.

The information on the investigated material is contained in the nonlinear susceptibility. It can be calculated by time-dependent perturbation theory taking into account a ground state $|a>$ and intermediate states $|b>$, $|c>$ and $|d>$. It turns out that $\chi^{(3)}$ is proportional to a sum over all states of the system as intermediate states; in the sum 24 terms like the one given in eqn [12] have to be added:

$$\frac{\mu_{ijkl}}{(\omega_{ba} - \omega_{\mathrm{L}} - \mathrm{i}\Gamma_{ba})(\omega_{ca} - \omega_{\mathrm{L}} + \omega_{\mathrm{S}} - \mathrm{i}\Gamma_{ca})(\omega_{da} - \omega_{\mathrm{CARS}} - \mathrm{i}\Gamma_{da})} \qquad [12]$$

μ_{ijkl} is a product of four dipole matrix elements between states $|a>:|b>$; $|b>:|c>$; $|c>:|d>$ and $|d>:|a>$ with field polarizations as given by the indices. $\omega_{\alpha\beta} = (E_\alpha - E_\beta)/\hbar$ and $\Gamma_{\alpha\beta}$ is the linewidth of the respective transition. Resonances of the susceptibility are found by inspection of the denominators. $\chi^{(3)}$ exhibits Raman-like resonances whenever $\hbar(\omega_{\mathrm{L}} - \omega_{\mathrm{S}})$ is equal to the excitation energy of a Raman allowed transition (see the second factor in the denominator of eqn [12]). If in this case the laser beams are polarized according to the Raman selection rules then μ_{ijkl} is not zero. For spin transitions in a semiconductor in a magnetic field (free electron spin flip as well as transitions within the magnetic ions) this is the case if in the Voigt configuration ($\vec{k}_{\mathrm{L,S}} \perp \vec{B}$) the beams with frequencies ω_{L} and ω_{S} are polarized perpendicular to one another. In the case of a spin transition $\omega_{ca} = g^*\mu_{\mathrm{B}}B$ (g^* being the effective Landé-factor and μ_{B} the Bohr magneton). Thus holding the frequencies ω_{L} and ω_{S} fixed and tuning the magnetic field a resonance in I_{CARS} will be observed from the position of which the effective g-factor can be obtained. Using tunable dye lasers, of course, the laser frequencies can be tuned as well.

Close to the resonance the real part of the susceptibility becomes small and the imaginary part is not negligible. Furthermore, one-photon resonances may occur if the laser frequency ω_{L} corresponds to the energy of a real electronic transition, e.g., an allowed interband transition. If one-photon and Raman-like resonances occur simultaneously one talks about resonance CARS.

As discussed in the literature $\chi^{(3)}$ can be split into a resonant part χ_{r}, summing up all the resonant terms, and a nonresonant part χ_{nr}. In a semiconductor there are contributions to χ_{nr} due to free carriers and a background due to bound valence electrons. The background contribution in general is real. As can be seen from eqn [10],

$$I_{\mathrm{CARS}} \propto \left|\chi^{(3)}_{\mathrm{CARS}}\right|^2 = |\chi_{\mathrm{nr}}|^2 + |\chi'|^2 + |\chi''|^2 + 2\cdot\chi_{\mathrm{nr}}\chi' \qquad [13]$$

χ' and χ'' being the real and imaginary parts of the resonant contribution. Depending on the relative strength of resonant and nonresonant contributions to $\chi^{(3)}$ the last term in eqn [13] may cause complicated lineshapes which make it difficult to find the correct transition energy in spectroscopic applications of CARS. If more than one resonance is not completely resolved the lineshape is even more complicated and a careful analysis has to be performed to get the correct resonance positions and linewidths of the transitions.

Examples of Experimental Results

Figure 2 shows a scheme of a CARS setup suitable to measure spin transitions in a semiconductor. The scheme is drawn for backscattering geometry; forward detection is accomplished by directing the incoming beams onto the sample from the opposite side to that plotted.

Two dye lasers (frequencies ω_{L} and ω_{S}) are pumped by a frequency-doubled Nd:YAG laser. A Glan–Thomson prism combines the two beams. The monochromator is set to $2\omega_{\mathrm{L}} - \omega_{\mathrm{S}}$. The iris and the polarizer PZ3 reduce stray light. The experiment is performed either at fixed magnetic field and tuning one of the dye lasers (with the monochromator swept accordingly) or by keeping the laser frequencies fixed and sweeping the magnetic field. The latter method is also applicable in the mid-infrared using step-tunable CO or CO_2 lasers.

As an example, in Figure 3 a spectrum obtained from a CdMnTe alloy crystal is reproduced. There are two peaks observed due to the paramagnetic resonance of the Mn 3d electrons and due to the spin resonance of photo-excited free electrons in the conduction band. The inset shows the paramagnetic resonance on an extended scale. The linewidth of 0.5 cm^{-1} is not limited by the experimental resolution, which in the present case is about a factor of 10 better.

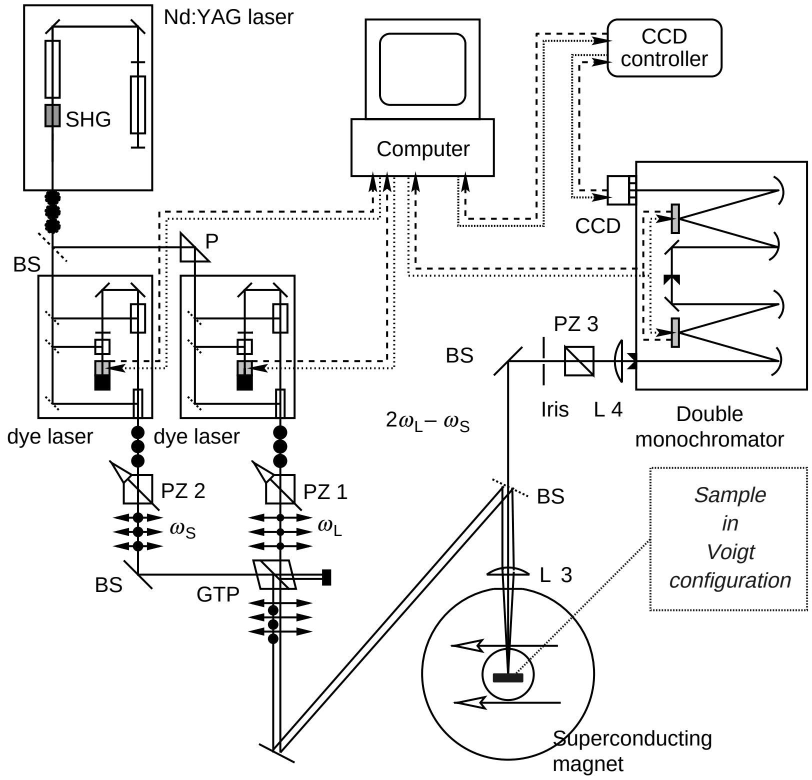
Nd:YAG laser
SHG
Computer
CCD controller
CCD
P
BS
PZ 3
BS
Iris
L 4
Double monochromator
dye laser
dye laser
$2\omega_L - \omega_S$
PZ 2
PZ 1
ω_S
ω_L
BS
BS
GTP
Sample in Voigt configuration
L 3
Superconducting magnet

Figure 2 Scheme of the experimental arrangement for CARS. Bold full lines: light beams; dashed and dotted lines: data connections; BS: beam splitter; PZ1,2: Pockels cells for rotation of linear polarization; PZ3: polarizer; GTP: Glan–Thomson prism for superposition of beams polarized orthogonally to each other; L: lenses.

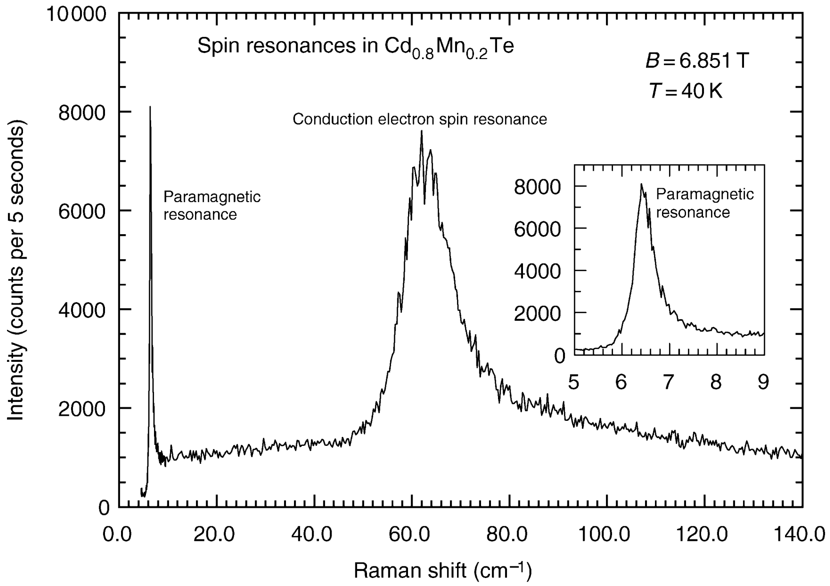

Figure 3 Coherent Raman intensity of a CdMnTe bulk crystal as a function of frequency difference $\omega_L - \omega_S$. Inset: Region of paramagnetic resonance in an extended scale. Adapted from Rupprecht R, Müller B, Pascher H, *et al.* (1998) Coherent Raman scattering from magnetic excitations in diluted magnetic semiconductors: Bulk crystals of CdMnTe. *Physical Review B* 58: 16123. Copyright (1998) by the American Physical Society.

Optical Pumping

The basic principles of optical pumping have been developed by Kastler. A material (gas or solid) is excited by absorption of circularly polarized light. Lampel was the first to demonstrate the possibility of optical spin orientation of free electrons in a semiconductor. The method has the most important applications in zincblende materials. Thus it is explained here with the example of GaAs.

Principles of Optical Spin Orientation

Most of the III–V and II–VI semiconductors have direct gaps at the center of the Brillouin zone. The conduction band (cb) is doubly degenerate with spin. The valence band consists of the heavy hole band (hh; $m_j = \pm 3/2$), the light hole band (lh; $m_j = \pm 1/2$) and the split-off band (so; $m_j = \pm 1/2$), each doubly degenerate. In bulk material hh and lh are degenerate at $k = 0$. The transition probabilities for interband transitions between hh and cb are three times larger than for transitions between lh and cb. Due to the selection rules, e.g., for the absorption of σ^+ circularly polarized light, transitions $-3/2 \rightarrow -1/2$ and $-1/2 \rightarrow +1/2$ are allowed, the first having three times higher probability. The electronic spin polarization P is defined as:

$$P = \frac{n_+ - n_-}{n_+ + n_-} \qquad [14]$$

where n_+ and n_- are the densities of electrons with the two spin orientations with respect to the direction of propagation of light. These considerations yield $P_\mathrm{i} = -0.5$ as the initial spin polarization under excitation with σ^+ circularly polarized light with a photon energy between the fundamental gap and the energy difference from the spin–orbit split valence band to the conduction band. Due to relaxation processes the actual electronic steady state polarization P is smaller than P_i:

$$P = P_\mathrm{i} \frac{T_1}{T_1 + \tau} \qquad [15]$$

with T_1 being the spin relaxation time, which in zero magnetic field is equal to the transverse relaxation time T_2, and τ is is the lifetime of conduction electrons.

The spin polarization of the electrons can be detected by analyzing the degree of circular polarization of the luminescence light which is defined as:

$$\Pi = \frac{I_+ - I_-}{I_+ + I_-} \qquad [16]$$

where I_+ and I_- are the intensities of σ^+ and σ^- polarized light, respectively. Since for emission due to free-electron/free-hole or free-exciton recombination the same transition probabilities as for absorption are obtained, one finds:

$$\Pi = 0.5\,P = 0.25 \frac{T_1}{T_1 + \tau} \qquad [17]$$

In the case of quantum wells, the lh-hh degeneracy is lifted due to different confinement energies. Then the maximum polarization of cb → hh as well as for cb → lh luminescence light is 0.5.

Of course these considerations are only true for intrinsic or p-type materials where the number of photo-excited electrons is high compared to the equilibrium number of electrons.

Hyperfine Interaction of Spin-Polarized Electrons and Atomic Nuclei

In GaAs three isotopes exist with spin 3/2, namely ^{69}Ga, ^{71}Ga and ^{75}As. Owing to the small energy separation of the nuclear spin levels in moderate magnetic fields, in thermal equilibrium no significant nuclear spin polarization exists even at a temperature of 2 K. However, the hyperfine interaction between electrons and nuclei allows so-called flip-flop processes, that is a simultaneous reversal of electronic and nuclear spins. This causes a considerable dynamic spin polarization of the nuclei if the electronic spin system is polarized due to optical pumping. Under most experimental conditions and when nuclear spin relaxation mechanisms other than via the hyperfine interaction are negligible, the mean nuclear spin is approximately:

$$\langle I \rangle = \frac{I(I+1)}{S(S+1)} \frac{P}{2} \qquad [18]$$

where I and S are the operators for nuclear and electron spin. These dynamically polarized nuclei act on the electrons and shift their spin resonance magnetic field by an effective nuclear field (the Overhauser shift). This shift is proportional to $\langle I \rangle$ and to the probability of finding the electrons at the nuclear sites, and is inversely proportional to g^*. In GaAs due to the small value of g^* this field can reach several tesla.

Effect of a Constant Transverse Magnetic Field – The Hanle Effect

In a transverse (with respect to the direction of propagation of light) magnetic field the spin-oriented electrons precess around the field with an angular frequency $\Omega = g^* \mu_B B/\hbar$. If the spin lifetime T_S $(1/T_S = 1/T_2 + 1/\tau)$ is long enough, this precession diminishes the average spin polarization and consequently the degree of circular polarization of the photoluminescence according to:

$$\Pi(B) = \Pi(0) \frac{1}{1 + \Omega^2 T_S^2} \qquad [19]$$

This is a Lorentzian with a width at half height of $\Delta B = \hbar/(g^* \mu_B T_S)$.

It is seen that by measuring $\Pi(B)$ the spin lifetime can be determined if g^* is known. Together with eqn [17] the spin and interband lifetime can be determined in a steady-state experiment. If, on the other hand, T_S is known from a time-resolved experiment, g^* can be found from a Hanle effect measurement.

Optically Detected Conduction Electron Spin Resonance (ODCESR)

Optical spin orientation of conduction electrons is the presupposition for ODCESR. Suppose now, a longitudinal magnetic field $\mathbf{B}_0$ is applied on the sample ($\mathbf{B}_0$ parallel to $\mathbf{k}$, the direction of propagation of light, which is parallel to the direction of the optically oriented spins). If the field is not too high the considerations of optical spin orientation remain unchanged. In higher fields the spin polarization may be modified due to the energy splitting of the spin levels. In this case the spin polarization is enhanced or diminished, depending on the direction of the field parallel or antiparallel to $\mathbf{k}$. If now an additional transverse microwave field B_1 of frequency ω_0, which fulfills the resonance condition $\hbar\omega_0 = g^* \mu_B B_0$, is applied, transitions between the two spin levels will occur which resonantly decrease the spin polarization. This decrease manifests itself in a decrease of the degree of circular polarization of the photoluminescence.

The method has several advantages compared to conventional ESR. Microwave photons with frequency of about 10^{10} Hz change the state of polarization of optical photons with frequency of about 10^{14} Hz. Thus the sensitivity is that of the optical detection which is much higher than in conventional ESR. Even electrons in a single quantum well or heterostructure can be detected by the optical method. There is also the possibility to distinguish between electrons in different quantized levels in heterostructures due to the different spectral position of their photoluminescence. Another advantage is that the electrons are photo-excited; the sample does not need to be n-doped. High-purity samples can be investigated which exhibit narrower linewidths.

Optically Detected Nuclear Magnetic Resonance (ODNMR)

With the presuppositions provided in the preceding sections the basic principle of ODNMR is easily understood. As in ODCESR we assume that an external magnetic field $\mathbf{B}_0$ is applied parallel to $\mathbf{k}$. Perpendicular to it a radio-frequency field $\mathbf{B}_2$ with frequency ω is applied. The NMR resonance frequency is $\omega_0 = \gamma_n \cdot B_0$ with γ_n being the gyromagnetic ratio of the nucleus under consideration. In the frame rotating with ω around the direction of $\mathbf{B}_0$ an effective field $\mathbf{B}_{\text{eff}} = \mathbf{B}_0 - \boldsymbol{\omega}/\gamma_n + \mathbf{B}_2$ acts on the nuclei. The nuclear magnetization is oriented parallel to $\mathbf{B}_{\text{eff}}$. It is seen that in resonance $\mathbf{B}_{\text{eff}}$ is parallel to $\mathbf{B}_2$, which means perpendicular to $\mathbf{k}$. Via the hyperfine interaction this nuclear field acts on the electrons, diminishing their spin polarization via the Hanle effect. So the nuclear magnetic resonance can be detected as a reduction of the degree of circular polarization of the photoluminescence.

The simplified model given above does not take into account several effects which influence the lineshape, position and intensity of the observed resonance lines.

Examples of Experimental Results for Optical Pumping and Optically Detected Magnetic Resonances

These experiments can be performed at liquid helium temperature in a standard photoluminescence setup. In addition one has to provide the possibility of applying transverse (Hanle effect) or longitudinal (ODCESR, ODNMR) magnetic fields. For ODCESR the sample is mounted in a microwave resonator with optical access; for ODNMR a radio-frequency coil is wound around the sample. The state of circular polarization of the exciting laser light has to be controlled. The measured quantity is the sign and degree of circular polarization of the luminescence light at a certain wavelength. The sensitivity can be enhanced by periodically σ^+/σ^--modulating either the exciting light or the detection. Which one of these possibilities is used depends on whether or not a high nuclear spin polarization is wanted.

As pointed out above the dynamically polarized nuclei, particularly in GaAs, strongly affect the

ODCESR resonant field. To get reliable values for the effective g-factor, the nuclear spin polarization must be completely suppressed. This can be achieved by σ^+/σ^--modulation of the exciting light with a modulation period much shorter than the nuclear spin relaxation time. But due to imperfections, e.g., if the intensities of σ^+ and σ^- light or their time intervals are not equal, residual optical fields may remain. To get precise values for g^* in addition the NMR has to be saturated by a suitable radio-frequency field. On the other hand, the strong Overhauser shift of the CESR can be useful for the determination of nuclear spin relaxation times. To perform such experiments the nuclei are first polarized by optical pumping with fixed circular polarization and afterwards, for the detection, σ^+/σ^--modulation is switched on.

For the optical detection of NMR a high nuclear spin polarization is favorable. Thus for those experiments one uses fixed polarization of the exciting beam and modulates the detected light.

Figure 4 shows examples of results obtained with a p-channel AlGaAs/GaAs single heterostructure. In this sample due to the doping profile a triangular potential well at the interface in the valence band of the GaAs layer is built, in which quantized hole levels exist. The luminescence band shown in Figure 4a between 821 nm and 825 nm is due to recombination of electrons with holes in these quantized states. For the experiments shown in the other panels of the figure the spectrometer is set at a position within this band. The other lines are due to excitons and impurities and are not interesting for what follows.

In Figure 4b the results of a Hanle experiment are reproduced. The line without noise is the fitted Lorentzian. The fit yielded $\tau = 8.1$ ns and $T_1 = 4.7$ ns.

In Figure 4c a typical ODNMR spectrum of the same sample is plotted. One observes resonances due to ^{75}As at 1.36 MHz, ^{69}Ga at 1.9 MHz and ^{71}Ga at 2.4 MHz. In the inset the ^{75}As resonance is plotted in an extended scale. A quadrupole splitting into three components is clearly resolved. This result demonstrates that the tetrahedral symmetry at the nuclear site is broken either due to local defects or due to the (rather small) lattice mismatch close to the AlGaAs/GaAs interface.

Finally Figure 4d shows the ODCESR spectrum. For the experiment the exciting beam was σ^+/σ^--modulated and in addition during the sweep the NMR of all three nuclei was permanently saturated to make sure that no Overhauser shift

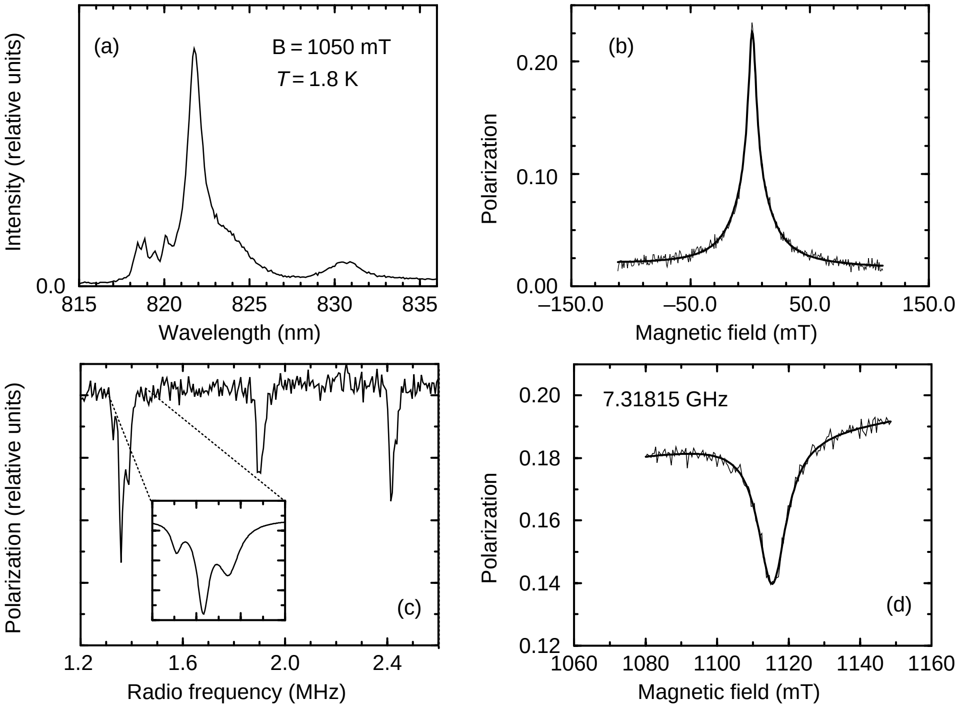

Figure 4 (a) Photoluminescence spectrum of a p-channel AlGaAs/GaAs heterostructure; (b) Hanle curve (luminescence polarization as a function of transverse magnetic field); (c) ODNMR recording; (d) ODCESR recording. Adapted from Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected magnetic resonance (ODMR) experiments. *Solid State Communications* 102: 715. Copyright (1997), with permission from Elsevier.

affects the result. The g-factor calculated from this plot is $|g^*| = 0.468$, which is larger than that found for bulk GaAs. This value was also used for fitting the Hanle curve.

List of Units and Nomenclature

Angle	[rad]
Angular frequency ω	[rad s^{-1}]
Bohr magneton μ_B	9.2742×10^{-24} J T^{-1}
Degree of circular polarization of light Π	
Electric field E	[V m^{-1}]
Gyromagnetic ratio γ	[T^{-1} s^{-1}]
Imaginary unit i	
Intensity I	[W m^{-2}]
Magnetic displacement B	[T]
Permittivity of free space ε_0	8.8542×10^{12} V s A^{-1} m^{-1}
Planck's constant $\hbar$	1.0546×10^{-34} J s
Refractive index n	
Speed of light in vacuum c_o	2.9979×10^{8} m s^{-1}
Spin polarization P	
Susceptibility χ	

CARS	coherent anti-Stokes Raman scattering
CSRS	coherent Stokes Raman scattering
ODCESR	optically detected conduction electron spin resonance
ODMR	optically detected magnetic resonance
ODNMR	optically detected nuclear magnetic resonance

Acknowledgments

The author thanks W. Herbst, B. Müller, R. Rupprecht and M. Schreiner for making unpublished results available.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Physics:** Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Boswarva IM, Howard RE and Lidiard AB (1964) Faraday effect in semiconductors. *Proceedings of the Royal Society* A269: 125–141.

Dyakonov MI and Perel VI (1984) Theory of optical spin orientation of electrons and nuclei in semiconductors. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 11–71. Amsterdam: North-Holland.

Eesley GL (1981) *Coherent Raman Spectroscopy*. Oxford: Pergamon Press.

Furdyna JK (1988) Diluted magnetic semiconductors. *Journal of Applied Physics* 64: R29–R64.

Geist F and Pascher H (1996) Resonance CARS in narrow gap semiconductors. *Journal of Raman Spectroscopy* 27: 289–295.

Geist F, Herbst W, Mejia-Garcia C, *et al.* (1997) Magneto-optical investigations of Eu-based diluted magnetic lead chalcogenides. *Physical Review B* 56: 13042.

Hermann C and Weisbuch C (1984) Optical detection of conduction electron spin resonance in semiconductors and its application to $\mathbf{k} \cdot \mathbf{p}$ perturbation theory. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 463–508. Amsterdam: North-Holland.

Krenn H, Herbst W, Pascher H, *et al.* (1999) Interband Faraday and Kerr rotation and magnetization of $Pb_{1-x}Eu_xTe$ in the concentration range $0 \times \sim 1$. *Physical Review B* 60: 8117.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*. San Diego: Academic Press.

Nies R and Kessler FR (1989) Coherent and incoherent magnetooptical reflection and transmission effects of multilayer structures. *Phys. Stat. Sol. (a)* 111: 639–654.

Oh E, Bartholomew DU, Ramdas AK, Furdyna JK and Debeska U (1988) Interband Faraday effect in CdMnSe. *Physical Review B* 38(18): 13183–13190.

Paget D and Berkovits VL (1984) Optical investigation of hyperfine coupling between electronic and nuclear spins. In: Meier F and Zakharchenya BP (eds) *Optical Orientation*, pp. 404ff. Amsterdam: North-Holland.

Pascher H, Geist F, Springholz G, *et al.* (1997) Coherent anti-Stokes Raman scattering in diluted magnetic IV–VI epilayers and superlattices. In: Shen SC, Tang DY, Zheng GZ and Bauer G (eds) *Narrow Gap Semiconductors*, pp. 269–278. Singapore: World Scientific.

Pohlt M, Herbst W, Pascher H, Faschinger W and Bauer G (1998) Magneto-optical Kerr effect in short-period CdTe/MnTe superlattices. *Physical Review B* 57(16): 9988–9994.

Ramdas AK and Rodriguez S (1988) Raman scattering in diluted magnetic semiconductors. In: Furdyna JK and Kossut J (eds) *Diluted Magnetic Semiconductors, Semiconductors and Semimetals*, vol. 25, pp. 345–412. London: Academic Press.

Schreiner M, Pascher H, Denninger G, *et al.* (1997) Nuclear spin relaxation in AlGaAs/GaAs heterostructures observed via optically detected resonance (ODMR) experiments. *Solid State Communications* 102: 715–720.

Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The addition of a magnetic field to an optical study of a semiconductor generally improves the specificity of the experiment, but in particular affects primarily the electrons rather than the lattice so that the magneto-optical properties are determined to a large degree by the electrons alone. The major interest has been in the determination of the electronic energy band structure of semiconductors, both in their bulk and low-dimensional (two-, one-, and zero-dimensional) quantum forms. The magneto-optical experiments fall mainly into two categories according to whether the effects are associated with intraband (within a single conduction or valence band) or interband (valence to conduction band) transitions; there are also hybrid (i.e., higher-order) effects such as spin-flip Raman scattering or two-photon absorption where both inter- and intraband transitions are involved in a two-stage process. Interband transitions occur typically in the visible to mid-infrared (MIR) region, and intraband transitions in the MIR to far-infrared (FIR) region.

Theoretical Background

The most common band structure involved in semiconductor magneto-optics is the so-called quasi-Ge model, which gives an accurate description of group IV, and zincblende (III–V and II–VI) semiconductors, so it is convenient to describe the theory in terms of this model. In this model the small terms associated with inversion asymmetry are neglected to a first approximation. At the Brillouin zone center (i.e., at $k = 0$) the conduction band has s-like symmetry and the triply degenerate valence band p-like symmetry. In the presence of strong spin–orbit coupling, part of this degeneracy is raised resulting in the doubly degenerate (at $k = 0$) heavy and light mass valence bands and a so-called spin–orbit split-off band. The reader should go to the general reviews quoted under ‘Further Reading’ for the equivalent coverage of other common systems such as the wurzite II–VI and the lead salt systems.

The dynamics of a free electron subject to an external magnetic field have been given by a variety of authors, but the most universally used treatment is that of Landau. For the motion of an electron in a semiconductor we treat first the case of a simple parabolic energy band. For the case of the magnetic field, B, along the z-direction, k_z remains a good quantum number, but k_x and k_y are subsumed into the new magnetic orbital (Landau) quantum number, n. The original three-dimensional energy band, $E = \hbar k^2/2m^*$, is coelesced into a series of one-dimensional Landau sub-bands in the k_z-direction (where the motion along k_z remains unchanged):

$$E_n^{\pm}(k_z) = (n + 1/2)\,\hbar\omega_c + \frac{\hbar k_z^2}{2m^*} \qquad [1]$$

The cyclotron frequency is $\omega_c = eB/m^*$ and μ_{B} is the Bohr magneton. The resulting electron energy eigenvalues at $k_z = 0$ in an external magnetic field (i.e., Landau sub-band extrema) are referred to as Landau levels, LLs. In practice each state is doubly degenerate with spin which results in an additional term $\pm g^*\mu_{\mathrm{B}}B$ on the right-hand side, where g^* is the effective g-factor. (Note that in the presence of strong spin–orbit coupling and small m^*, there is a corresponding g^* which becomes negative and substantially different from the free-electron g-factor; the largest value for a binary elemental compound is that for InSb, where $g^* \approx -50$.) The situation for a simple parabolic conduction band near $k = 0$ is shown in Figure 1a,b. Intraband optical absorption associated with direct transitions between successive Landau sub-bands is referred to as cyclotron resonance (CR) absorption. There are corresponding Landau sub-bands in the valence band, and direct transitions between valence and conduction sub-bands give rise to the so-called interband magneto-optical (IMO) effects.

Expressions for the density of states (DOS) in three, two, and one dimensions in the absence of a magnetic field are given elsewhere (‘semiconductor basics’ – Pidgeon). In the presence of a magnetic field the three-dimensional energy band is reduced to one-dimensional sub-bands so that the energy dependence of the DOS has the one-dimensional form with the energy given by eqn [1]. Performing this substitution and multiplying by a term for the degeneracy of the Landau levels (LLs) gives the following expression for the total DOS per unit volume:

$$g(E) = \frac{(2m^*)^{1/2}}{4\pi^2\hbar}\,\frac{eB}{\hbar}\sum_n \left[E - (n + 1/2)\,\hbar\omega_c\right]^{-1/2} \qquad [2]$$

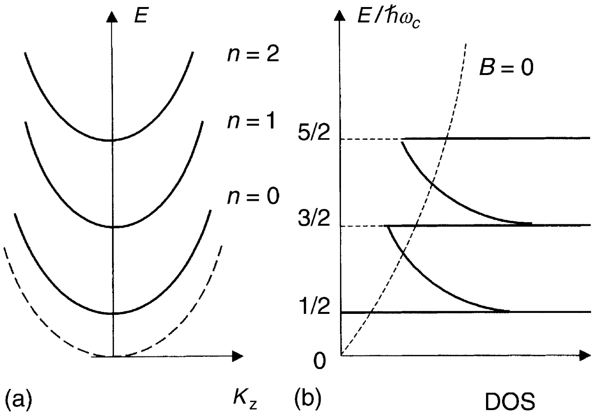

Figure 1 (a) Landau sub-bands according to eqn [1] with the magnetic field applied in the *z*-direction. The dashed line shows the parabolic conduction band for *k* along k_z and $B = 0$. (b) Schematic outline of the density of states (DOS) for free electrons in a magnetic field. It comprises the sum of identical contributions from each Landau sub-band. The density of states with $B = 0$ is shown dashed.

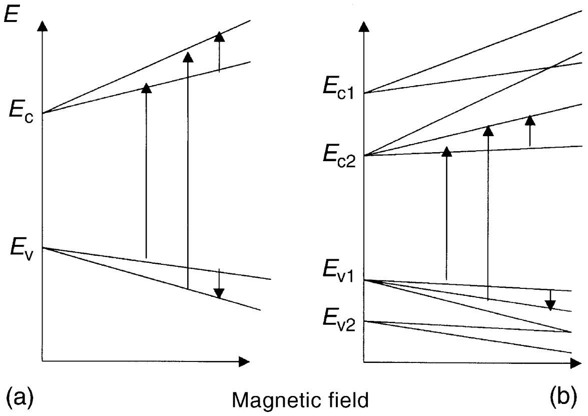

Figure 2 Fan plots of the conduction and valence band Landau levels versus magnetic field for simple parabolic bands: (a) a three-dimensional semiconductor; (b) quasi-two-dimensional systems. Some CR and interband magneto-optical transitions are shown by vertical arrows.

Thus, the DOS for Landau sub-bands is quite different from the zero-field DOS, and is a series of square-root singularities with divergences at the energies $(n + 1/2)\hbar\omega_c$. The two cases are illustrated schematically in Figure 1b. Since the optical absorption strength for both interband and intraband transitions is proportional to the joint density of states for the initial and final states, this result shows the important advantage of the use of a magnetic field in semiconductor spectroscopy because it means that the absorption lines for CR or IMO transitions are strongly peaked at the sub-band extrema (i.e., at $k_z = 0$). This allows accurate measurement of LL energy differences for transitions observed in a single experiment, and hence, by comparison to theoretical models, leads to the best determination of fundamental energy-band parameters. The basic LL selection rule can be shown straightforwardly from time-dependent perturbation theory to be $\Delta n = 0$, $\Delta m_s = 0$ for IMO transitions and $\Delta n = 1$, $\Delta m_s = 0$ for CR absorption, where m_s is the spin quantum number (i.e., $\pm\frac{1}{2}$, for spin up and spin down). The cyclotron resonance absorption occurs only for one sense of circularly polarized light, but the IMO effect can occur for all polarizations. The choice then of magnetic field direction, crystal orientation, and polarization of the optical radiation allows extremely detailed determination of the energy band parameters for particular crystal symmetries.

The data are often summarized in a so-called fan diagram of the energy levels, or transition energies, versus magnetic field. Such a plot is shown for hypothetical conduction and valence band LLs in Figure 2a, together with interband and CR transitions. The states are labeled by the Landau quantum number, with selection rules $\Delta n = 0$ and $\Delta n = 1$ for interband and CR absorption transitions, respectively (or $\Delta n = -1$ for CR emission). If the energies of the transitions themselves are plotted versus the magnetic field, then the intercept of the fan plots at $B = 0$ gives the zero-field energy gap. Note that in the presence of excitonic effects (see below) the intercept will give the energy gap plus the exciton binding energy.

For a strictly two-dimensional system the DOS per unit area is independent of energy: $g(E) = m^*/\pi\hbar^2$. In the case of quasi two-dimensional confinement in the *z*-direction the motion in the *z*-direction is quantized, and the total energy becomes (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures):

$$E = E_i + \frac{\hbar^2}{2m^*}(k_x^2 + k_y^2) \qquad [3]$$

Equation [3] describes a set of confinement sub-bands (parabolas induced by geometric, as opposed to magnetic, confinement), the bottoms of which are at the energies E_i, where i is the sub-band index. The energies, E_i, depend on the shape of the confining potential. The DOS per unit area is in this case a series of step functions of height $m^*/\pi\hbar^2$. From the point of view of magneto-optics, the main effect of the quantum confinement is to suppress the k_z motion, so that in the presence of an external magnetic field along the *z*-direction, the motion is completely quantized. The energy eigenvalues are given by:

$$E_{n,i} = E_i + (n + 1/2)\,\hbar\omega_c \qquad [4]$$

The total DOS becomes:

$$g(E) = \frac{m^* \omega_c}{\pi \hbar} \delta[E - (n + 1/2)\, \hbar\omega_c] \qquad [5]$$

a series of delta-functions located at the energy positions of the LLs. The DOS is obviously more singular in two dimensions than in three dimensions, and leads to substantially sharper magneto-optical and magneto-transport effects. For confined structures the selection rules for interband and intersub-band transitions are identical to the bulk case for the electromagnetic (em) radiation polarized in the *xy* plane provided that the external magnetic field is along the confinement (*z*-) direction. The only difference is that the conservation of momentum in the *z*-direction ($\Delta k_z = 0$) is replaced by a selection rule on the sub-band index, $\Delta i = 0$. A fan plot for this configuration is shown schematically in Figure 2b. For each confinement sub-band transition there is a set of LL transitions. When the energy positions are plotted against *B*, this leads to a set of Landau 'fans', each of which extrapolates to the sum of a valence plus conduction confinement sub-band energy at $B = 0$. Allowed transitions are shown for the lowest two confinement sub-bands in Figure 2b. Again this picture may be complicated in the presence of excitonic effects at low temperature.

Turning again to the bulk semiconductor picture, eqns [1]–[5] apply for a single parabolic energy band and are valid for the conduction band in the quasi-Ge model. In actual semiconductors further complications arise. Most importantly the triply degenerate p-like valence band (heavy hole, light hole, and split-off band) must be treated by so-called degenerate perturbation theory. In the quasi-Ge model the conduction and three valence bands are treated together as one coupled band system using the so-called $\mathbf{k} \cdot \mathbf{p}$ method (described elsewhere by Zawdzki). With spin this results in an 8×8 matrix equation which has to be diagonalized for the eigenvalues and eigenvectors of the system. The spin–orbit coupling results in so-called complex, mixed-spin valence bands and anomalously spaced low quantum number LLs (the so-called quantum effects). The overall energy band scheme for InSb, near the fundamental direct gap in the center of the Brillouin zone for conduction and (heavy and light mass) valence bands is shown schematically in Figure 3, for no magnetic field and for an applied magnetic field. Each LL is doubly degenerate with spin, the a-set representing the so-called spin-up and the b-set the spin-down states.

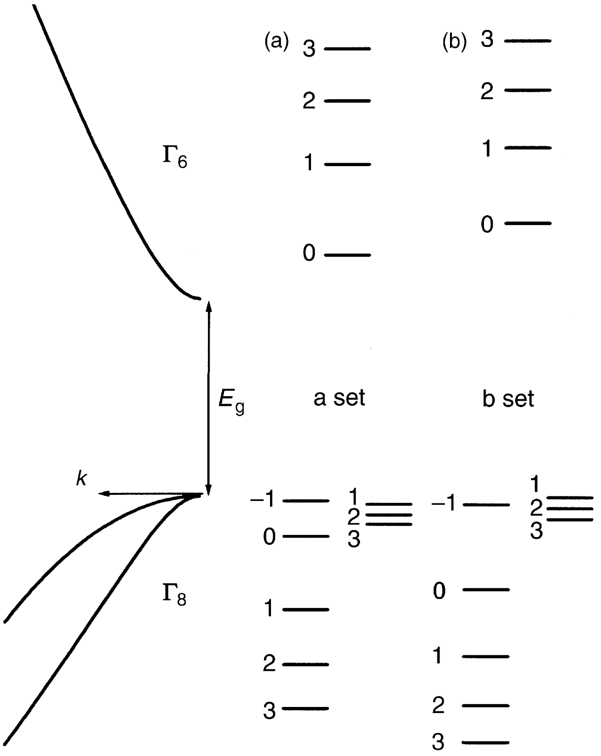

Figure 3 (a) Schematic diagram of the conduction and valence bands for InSb (symmetry labels Γ_6 and Γ_8 respectively) at $B = 0$. (b) Computed Landau levels for the conduction and valence bands of InSb at 4 K ($E_g = 0.23$ eV) with $B = 5$ T along the [100] direction, grouped according to the so-called a- and b-spin sets (see text).

Experimental

The first serious use of magneto-optics to study effective mass and energy band structure of semiconductors was in the microwave cyclotron resonance (i.e., intraband) investigations of Ge and Si in the early 1950s. Classically it can be shown that the ratio of the radiation power absorbed, *P*, to power absorbed in zero magnetic field, P_0, is given by:

$$\frac{P}{P_0} = \frac{1 + (\omega^2 + \omega_c^2)\tau^2}{[1 + (\omega_c^2 - \omega^2)\tau^2]^2 + 4\omega^2\tau^2} \qquad [6]$$

where τ is the collision time of the carriers. This expression for P/P_0 describes cyclotron resonance absorption and is plotted as a function of frequency in Figure 4 for different values of the parameter $\omega_c\tau$. Clearly the peak absorption occurs at $\omega = \omega_c$, namely at cyclotron resonance. Also evident is the very important requirement for observation of a clear resonance that $\omega_c\tau > 1$; this condition corresponds physically to a travel time between collisions of the particle of at least $1/2\pi$ of a revolution so that the particle can interact for a sufficiently long time

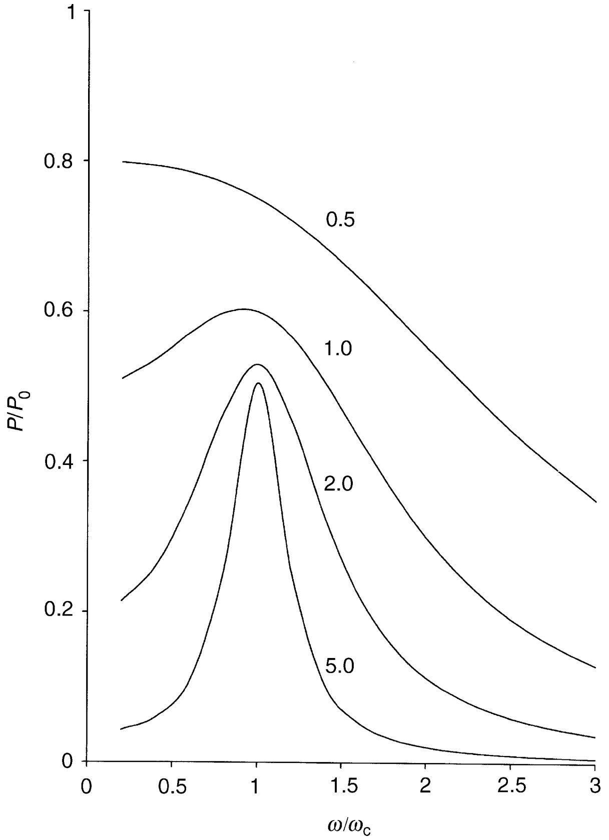

Figure 4 Ratio of the radiation power absorbed to the power absorbed at zero magnetic field, P/P_0, plotted as a function of ω/ω_c according to eqn [6], for different values of $\omega_c\tau$ from 0.5 to 5.0.

with the electromagnetic wave to absorb energy. For pure materials and low temperatures (where the condition $\omega_c\tau > 1$ is achievable) this is the best technique for obtaining the details of the electron and hole band parameters of a semiconductor near $k = 0$. The most thorough and successful studies of this type have been achieved in pure Si, Ge, and InSb where (with the aid of optical excitation to populate the levels shown in Figure 3) all possible electron, light and heavy hole cyclotron resonances have been resolved. In this way the valence band quantum effects, described above, have been observed in detail. The technique is equally applicable in low-dimensional systems where CR of both electrons and holes has been resolved in a large number of n- and p-type semiconductor heterostructures and quantum wells.

In interband – either absorption or photoluminescence (PL) – spectroscopy at low temperatures, the photo-excited electrons and holes interact and form excitons (*see* **Semiconductor Physics:** Excitons). In many cases the excitonic state may be described in the so-called envelope function approximation as a hydrogenic system, in which the electron and hole are bound by their mutual Coulomb attraction. The energy spectrum of such a hydrogenic exciton is given by

$$E_n = \frac{\mu e^4}{2\hbar^2\varepsilon^2 n^2} \qquad [7]$$

where n is the principal quantum number ($n = 1, 2, 3, \ldots$, but note that the symbol n is also used for the Landau quantum number above), μ is the electron–hole reduced effective mass, and ε is the background dielectric constant. In low-dimensional (i.e., confined) semiconductor systems of the type described above the degeneracy of the valence bands at $k = 0$ is removed, and two types of excitons are observed; these are referred to as 'heavy' and 'light' hole excitons in the literature, where heavy and light refer to the effective masses along the confinement direction. These excitons can move freely (in the remaining two unconfined directions) or be bound to an impurity or defect. In these systems exciton absorption may be observed involving not only the ground-state excitons, but also excitons associated with higher confinement sub-bands. This is shown for the so-called asymmetric quantum well system $\mathrm{Al}_y\mathrm{Ga}_{1-y}\mathrm{As/In}_x\mathrm{Ga}_{1-x}\mathrm{As/GaAs}$ in Figure 5a. The data are taken both by the direct PL method (where the exciting beam has an energy much larger than the bandgap and the electrons and holes de-excite instantaneously to the bottom of the bands before recombining as exiton recombination radiation), and so-called PL excitation spectroscopy (where the exciting beam is a tunable laser, tuned into resonance with the successive exciton resonances).

The interband Landau transitions can be observed either in PL or as excitations in a magneto-absorption or magnetoreflection experiment. Excitons are affected differently by the application of an external magnetic field depending whether the ratio $\gamma = \hbar\omega_c/2E_B$ (where E_B is the exciton binding energy with $n = 1$ above) is less than or greater than one. In the so-called magnetic limit where $\gamma \gg 1$, the excitonic ground state moves approximately as $\frac{1}{2}\hbar\omega_c$, i.e., linearly with B. In this regime the magnetic field energy is the dominant term in the exciton Hamiltonian. The various states of the exciton (ground and excited states) become associated with the interband LL transition, and exhibit an energy dependence that is nearly identical to the free carrier LL transitions. The energies of the various excitonic transitions are shifted down with respect to the corresponding LL transitions by an amount equal to the binding energy of the excitonic state to that LL. Typical PL and PLE (photoluminescence excitation)

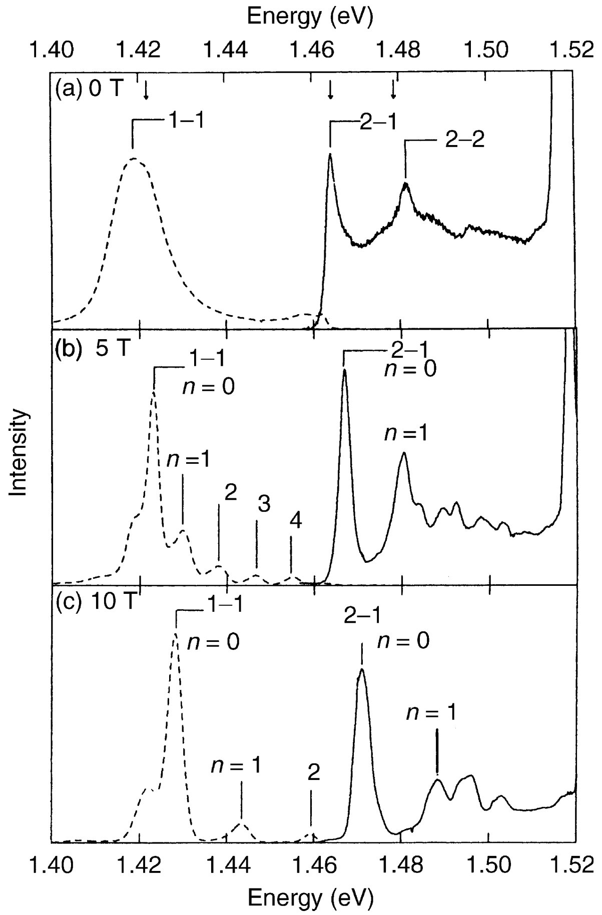

Figure 5 Interband transition data obtained in zero magnetic field (a) and fields of 5 T (b) and 10 T (c) from asymmetric quantum wells. The dashed lines indicate data obtained from PL and the solid lines data from PLE. The arrows in (a) indicate the calculated interband transition energies 1 → 1, 2 → 1, and 2 → 2, the latter being clearly resolved in PLE (the asymmetric wells mean that transitions with $\Delta i \neq 0$ are allowed). In (b) and (c) the development of the spectra into LL transitions is clearly seen, although beyond 1.48 eV the analysis is obviously extremely complicated by the overlap of LL transitions associated with all three of the above sub-bands.

spectra for interband magneto-optical transitions are shown in Figure 5b,c for a fixed applied magnetic field in the asymmetric quantum well system $Al_yGa_{1-y}As/In_xGa_{1-x}As/GaAs$.

Spin Flip Raman Scattering

In addition to the interband transitions discussed above, Raman scattering of intersub-band transitions between conduction or valence band LLs can also be studied. It provides a versatile means of probing band edge parameters of electrons and holes in semiconductors since the excitation (normally a laser pump) is not confined to the resonance frequencies under investigation, i.e., any pumping energy below the band edge will do. Electronic Raman scattering is then allowed between LLs via virtual interband transitions. In the presence of spin–orbit coupling strong spin-flip processes via electric dipole transitions become allowed giving rise to the so-called spin-flip Raman (SFR) cross-section. This is a single-particle process in which electrons or holes interacting with the radiation change their spin state. This process can be made to go stimulated particularly easily when the pump radiation is resonant with the bandgap of the semiconductor, forming the spin-flip Raman laser. The most important practical case has probably been that of InSb, resonantly pumped near 5 μm with the CO laser. In this case the small effective mass of a narrow-gap semiconductor in the presence of substantial spin–orbit coupling leads to both an extremely large anomalous effective g-factor ($g^* \sim -50$) – giving a large tuning range, $\Delta E = \hbar\omega_{SF} = g^*\beta B$ – and a strong cross-section which is larger than that of the free electron by a factor $(m_0/m^*)^2$. Efficient cw operation is possible with threshold powers less than 50 mW.

SFR scattering has enabled the observation of a great variety of excitations in nonmagnetic and dilute magnetic semiconductors, including: free electrons and holes, electrons bound to donors and holes bound to acceptors, scattering from bound magnetic polarons, and scattering due to SF transitions within the Zeeman multiplets of paramagnetic ions. In addition to pure SF transitions, combined resonances with combinations such as $\omega_{CR} \pm \omega_{SF}$ have also been detected. The process has also been used in a similar manner to electron spin resonance to study coherent processes, including Raman photon echo spectroscopy in semiconductors.

See also

Magneto-Optics: Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Häfele HG (1991) Spin-flip raman scattering. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 207–275.

Petrou A and McCombe BD (1991) Magnetospectroscopy of confined semiconductor systems. In: Landwehra G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 679–775.

Pidgeon CR (1980) Free carrier optical properties of semiconductors. In: Balkanski M (ed.) *Optical Properties of Solids. Handbook of Semiconductors*, pp. 223–328. Amsterdam, The Netherlands: North-Holland.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba WI (eds) *Landau Level Spectroscopy in Modern Problems in Condensed Matter Sciences*, 27.1, 483–512.

MATERIALS CHARACTERIZATION TECHNIQUES

Contents

$\chi^{(2)}$
$\chi^{(3)}$

$\chi^{(2)}$

R C Eckardt, Cleveland Crystals, Inc., Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Techniques used to measure second-order susceptibility $\chi^{(2)}$ or nonlinear optical coefficients are described in this article, and some measured values are quoted. These are the parameters that describe the property of a material that is responsible for second-harmonic generation and other three-wave nonlinear optical interactions. These other interactions include sum- and difference-frequency generation in which two waves combine to generate a third wave of a frequency that is the sum or difference of the first two waves. Also included are optical parametric generation and optical parametric oscillation, in which a single pump wave generates two resulting waves with frequencies that sum to the frequency of the pump wave. The measurement techniques involve measuring the efficiency of nonlinear frequency conversion by these processes.

The measurements require a great deal of care and precision, and it is advisable to perform a comparison with previous measurements and obtain correlation with accepted standards whenever possible. A great deal of effort continues to be expended by researchers to establish correlations between measurements and provide additional values that are accepted as standards. A number of techniques have been used for these measurements sometimes yielding results that disagree. There are, however, values derived from early measurements that have proven accurate. In these measurements it is necessary to accurately control and characterize the polarization, power, spectral distribution, and spatial and temporal intensity distribution of the incident radiation and accurately measure the frequency-converted light, which can be many orders of magnitude smaller.

The development and availability of single-longitudinal-mode and nearly diffraction-limited lasers has greatly reduced the ambiguity and simplified increased accuracy of measurements of second-order nonlinear optical coefficients. The improved optical quality of nonlinear-optical crystals has also been an important component of the increased accuracy of measurement. Measurement techniques have also improved, providing absolute measurements of optical intensity over wide dynamic ranges at different wavelengths. Until the early 1990s, there was a range of measured values for nonlinear optical coefficients of common nonlinear materials, and numerous definitions were in use. Caution is required when consulting early publications on this topic, and some of this ambiguity is carried over to present time. Different definitions vary by factors of 2 and the permittivity of free space, $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N × m^2. It is important that the definition of a nonlinear optical coefficient is clear before a value is used. This can be done through a statement of the nonlinear electric polarization or the coupled equations for nonlinear frequency conversion.

The standard techniques that have been used to measure nonlinear optical coefficients are covered in detail in publications listed in the Further Reading at the end of this article. Brief descriptions of some of these techniques, such as Maker-fringe and wedge, optical parametric fluorescence, and phase-matched second-harmonic generation, will be presented in this article. Each of these techniques has a unique set of advantages and disadvantages. There was some lively controversy over the technique that provided the most accurate measurements and those measurements that should be dismissed. Measurements based on phase-matched second-harmonic generation had the advantage of being most similar to practical application of nonlinear optical frequency conversion. However, the large divergence and multiple-longitudinal-mode operation of early lasers made this method unreliable. Even relative comparisons of nonlinear optical coefficients between different materials were inaccurate because highly structured laser pulses had different conversion characteristics due to differences in bandwidth acceptance, group velocity walkoff, and birefringent walkoff in different materials.

Nonlinear optical coefficients of improved accuracy were required for analysis of optical harmonic conversion in the large, well-characterized laser systems used in inertial-confinement fusion research. The development of these systems provided a standardization on the value of $d_{36} = 0.39\ \mathrm{pm\,V^{-1}}$ ($1\ \mathrm{pm} = 10^{-12}\ \mathrm{m}$) for second-harmonic generation of neodymium laser radiation in the material potassium dihydrogen phosphate (KH_2PO_4 or KDP). The development of injection-seeded Q-switched lasers provided single-longitudinal-mode laser output that allowed the phase-matched second-harmonic techniques to be extended to provide more accurate absolute measurements and relative comparisons on KDP and a number of other materials. Correlation was achieved on a larger number of relative and absolute measurements of nonlinear optical coefficients to provide some systematic agreement. It is not well understood why the earlier measurements referenced to optical parametric fluorescence produced values that were higher by about 60%. Measurement techniques have been further refined, and more exacting measurements are leading agreement, using different methods of measurement on an increasing number of nonlinear optical materials.

Definitions

The definition of the nonlinear optical coefficient can be specified in a number of ways. Examples given here are how the coefficient relates to the nonlinear electric polarization in a material, the coupled equations for nonlinear frequency conversion, and expressions describing second-harmonic generation with a monochromatic plane-wave fundamental.

The second-order optical nonlinearity is reported as both the second-order susceptibility $\chi^{(2)}$ and the second-order nonlinear optical coefficient d, which usually differ in definition by a factor of two. These are third-rank tensors that relate products of components of the electric field vector E_j and E_k to the vector components of the nonlinear polarization $P_\mathrm{i}^{\mathrm{NL}}$. The second-order susceptibility typically is used to express instantaneous values:

$$P_\mathrm{i}^{\mathrm{NL}}(t) = \varepsilon_0 \sum_{j,k} \chi_{ijk}^{(2)} E_j(t) E_k(t) \qquad [1]$$

whereas the nonlinear optical coefficient more often is used to relate the vector amplitudes of frequency components of these quantities. The electric field at the fundamental angular frequency ω can be expressed as

$$E_j(\omega, \mathbf{r}, t) = \frac{1}{2}\{E_j(\omega) \exp \mathrm{i}(\mathbf{k}_\omega \cdot \mathbf{r} - \omega t) + c.c.\} \qquad [2]$$

and the nonlinear polarization at the harmonic frequency 2ω can be expressed as

$$P_\mathrm{i}^{\mathrm{NL}}(2\omega,\mathbf{r},t)=\frac{1}{2}\left\{P_\mathrm{i}^{\mathrm{NL}}(2\omega)\exp\mathrm{i}(2\mathbf{k}_\omega\cdot\mathbf{r}-2\omega t)+c.c.\right\} \qquad [3]$$

The intensity is given by $I=(nc\varepsilon_0/2)|E|^2$. The relationship between the complex amplitudes expressed in terms on the nonlinear optical coefficient tensor is

$$P_\mathrm{i}^{\mathrm{NL}}(2\omega)=\varepsilon_0\sum_{j,k} d_{ijk}(-2\omega;\omega,\omega)E_j(\omega)E_k(\omega) \qquad [4]$$

When waves of three different frequencies are involved, the expression becomes

$$P_\mathrm{i}^{\mathrm{NL}}(\omega_3)=2\varepsilon_0\sum_{j,k} d_{ijk}(-\omega_3;\omega_2,\omega_1)E_j(\omega_2)E_k(\omega_1) \qquad [5]$$

Expressing the coefficient as $d_{ijk}(-\omega_3;\omega_1,\omega_2)$ allows for dispersion and indicates that fields at frequencies ω_1 and ω_2 interact to produce a nonlinear polarization at frequency ω_3, the sum of the first two frequencies. The relationship between the second-order electrical susceptibility and the second-order nonlinear optical coefficient is

$$\chi_{ijk}^{(2)}=2d_{ijk} \qquad [6]$$

The dispersion of nonlinear optical coefficients is approximated by the use of Miller's delta:

$$\delta_{ijk}=\frac{d_{ijk}(-\omega_3;\omega_2,\omega_1)}{(n_\mathrm{i}^2(\omega_3)-1)(n_j^2(\omega_2)-1)(n_k^2(\omega_1)-1)} \qquad [7]$$

The indices of refraction in the direction of the electric field components $E_m(\omega_n)$ are $n_m(\omega_n)$ for $m=i,j,k$ and $n=1,2,3$. There is some theoretical and experimental justification for treating the components of Miller's delta as constants. Earlier definitions of Miller's delta included multiplication by ε_0, the permittivity of free space.

The nonlinear optical coefficient is usually reduced to a scalar quantity, the effective nonlinear coefficient or d_{eff}, that relates only the specific vector components of the fields involved in the interaction. The effective nonlinear optical coefficient will be discussed in more detail later. The coupled equations for second-harmonic generation using d_{eff} are

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_{2\omega}c}E_\omega(z)E_\omega(z)\exp(-\mathrm{i}\,\Delta kz) \qquad [8]$$

$$\frac{\mathrm{d}E_\omega(z)}{\mathrm{d}z} = \mathrm{i}\frac{\omega d_{\text{eff}}}{n_\omega c}E_\omega(z)E^*_\omega(z)\exp(\mathrm{i}\,\Delta kz) \qquad [9]$$

Here ω is the angular frequency of the fundamental wave; z is distance in the direction of propagation, $n_{2\omega}$ and n_ω are the refractive indices of the harmonic and fundamental respectively; and c is the speed of light in free space. The complex electric field amplitudes of the harmonic and fundamental, $E_{2\omega}(z)$ and $E_\omega(z)$, respectively, are assumed to be slowly varying functions of position compared with the oscillation of the carrier waves $\exp \mathrm{i}(\mathbf{k}_{2\omega}\cdot\mathbf{r} - 2\omega t)$ and $\exp \mathrm{i}(\mathbf{k}_\omega\cdot\mathbf{r} - \omega t)$, with $\mathbf{k}_{2\omega}$ and $\mathbf{k}_\omega$ the wavevectors of harmonic and fundamental; the asterisk indicates complex conjugate. The quantity Δk is the wavevector mismatch and is given by $\Delta k = k_{2\omega} - 2k_\omega$. These are simplified expressions that apply to nearly monochromatic plane waves. Two solutions to the coupled equations are

$$I_{2\omega}(z) = \frac{2\omega^2 d^2_{\text{eff}} l^2 I^2_\omega(0)}{n_{2\omega}n^2_\omega c^3\varepsilon_0}\,\frac{\sin^2(\Delta kz/2)}{(\Delta kz/2)^2} \qquad [10]$$

for insignificant depletion of the fundamental amplitude and

$$I_{2\omega}(z) = I_\omega(0)\tanh^2\sqrt{\frac{2\omega^2 d^2_{\text{eff}} z^2 I_\omega(0)}{n_{2\omega}n^2_\omega c^3\varepsilon_0}} \qquad [11]$$

for the case of perfect phase matching $\Delta k = 0$ and allowing depletion of the fundamental. The quantity $I_{2\omega}(z)$ is the harmonic intensity after propagating a distance z in the nonlinear material, and $I_\omega(0)$ is the fundamental intensity at the input surface of the nonlinear material where $I_{2\omega}(0) = 0$.

The above relationships for nonlinear polarization, the coupled equations for second-harmonic generation, or harmonic conversion of a monochromatic plane wave, each determine the definition of the second-order nonlinear optical coefficient.

Symmetry

Symmetry has an important role in the measurement of nonlinear optical coefficients. Indeed the second-order nonlinear optical coefficient must be identically zero if the material has inversion symmetry. This restricts materials to noncentrosymmetric crystals or isotropic materials in which inversion symmetry is removed by such techniques as application of an electric field or polymeric materials in which alignment of molecules is achieved by some poling process. The discussion here is limited to noncentrosymmetric crystals.

The nonlinear optical coefficient is a tensor of rank 3. Representation of the tensor in a rotated coordinate system is described by

$$d'_{ijk} = \sum_{l,m,n} R_{il}R_{jm}R_{kn}d_{lmn} \qquad [12]$$

where R_{il} is the matrix that describes rotation of the initial coordinate system xyz to the new coordinate system $x'y'z'$. The representation of a vector r_l in the original coordinate system becomes $r'_\mathrm{i} = \sum_l R_{il}r_l$ in the new coordinate system. A required symmetry property of the nonlinear optical tensor, that exchange of the second two indices of a component does not change the value of the component $d_{ijk} = d_{ikj}$, allows the third-rank tensor to be written as a 3 by 6 matrix:

$$\begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} \end{pmatrix} \qquad [13]$$

with the components of the matrix related to the tensor components by

$$\begin{array}{ll} d_{i1} = d_{i11} & d_{i4} = d_{i23} = d_{i32} \\ d_{i2} = d_{i22} & d_{i5} = d_{i13} = d_{i31}, \ i = 1,2,3 \\ d_{i3} = d_{i33} & d_{i6} = d_{i12} = d_{i21} \end{array} \qquad [14]$$

A further property, tensor components with any permutation of the indices are equal $d_{ijk} = d_{kij}$, called Kleinman symmetry or Kleinman's conjecture, is often applied if there is no absorption at any of the wavelengths involved. Kleinman symmetry assures that the amount of energy depleted from the fundamental is equal to the amount of energy gained by the second harmonic. This reduces the number of independent components of the nonlinear optical matrix to 10: d_{11}; $d_{12} = d_{26}$; $d_{13} = d_{35}$; $d_{14} = d_{25} = d_{36}$; $d_{15} = d_{31}$; $d_{16} = d_{21}$; d_{22}; $d_{23} = d_{34}$; $d_{24} = d_{32}$; and d_{33}. Sometimes components equal, according to Kleinman symmetry, are measured individually, either as a check of the measurement or to check the validity of the conjecture.

Crystal symmetry places additional constraints on the components of the nonlinear optical matrix. These constraints determine the components that must be zero and determine additional relationships between the nonzero components. Matrices for selected crystal point groups are given in Table 1. Effective nonlinear optical coefficients can be found by a coordinate frame transformation in which the rotated axes align with the direction of propagation and the eigen polarizations. Effective nonlinear coefficients for a few crystal point groups are given in Table 2. More complete listings of both the reduced matrices and effective nonlinear optical coefficients are given elsewhere.

Table 1 Nonlinear coefficient matrices for selected crystal point groups

Orthorhombic *mm*2	Hexagonal 6
$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & d_{14} & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & -d_{14} & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$
Tetragonal $\bar{4}2m$	Hexagonal 6*mm*
$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{36} \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & 0 \\ 0 & 0 & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$
Trigonal 3*m* $m \perp$ X	Cubic $\bar{4}3m$
$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & -d_{22} \\ -d_{22} & d_{22} & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{14} & 0 \\ 0 & 0 & 0 & 0 & 0 & d_{14} \end{pmatrix}$

Measurements

Usually the symmetry of a nonlinear crystal is determined by techniques such as X-ray diffraction or surface morphology. Crystal symmetry determines the nonzero components of the second-order nonlinear optical tensor. Nonphase-matched second-harmonic-generation techniques, such as Maker-fringe and wedge measurements, allow selection of polarizations and propagation directions relative to the crystal lattice that isolates individual components of the tensor. Phase-matched measurement techniques usually require propagation in a direction for which an effective nonlinear optical coefficient applies and can involve more than one tensor component. Parametric fluorescence and sum- and difference-frequency-generation measurements are phase-matched processes that involve three wavelengths. All of the measurement techniques have been used with pulsed and continuous-wave laser output. In all of these methods, except parametric fluorescence, it is necessary to accurately characterize the incident radiation. Measurements relative to a standard material avoid some of the concerns of characterization of the fundamental radiation. Relative measurements are important in that they allow correlation between different sets of measurements. Measurement sets, that include both relative and absolute measurements on a number of materials, are useful in developing correlations that achieve more accurate compilations of values.

It is desirable to have an incident beam with a Gaussian-like transverse intensity distribution and close to diffraction limited propagation. Gaussian beams have advantages that they are most easily characterized and their propagation is described with relatively simple analytical methods. Accurate temporal characterization of the incident light is also essential. Here it is desirable to have stable single-mode laser operation to produce narrow bandwidth light that is well within phase-matching acceptance bandwidths and that has temporal structure slow enough to be easily characterized and slow enough

Table 2 Effective nonlinear optical coefficient for a few symmetries

*mm*2, a point group with biaxial birefringence		
$\phi = 0$, two waves $E \| Y$, one wave $E \| XZ$	$d_{\text{eff}} = d_{32} \sin\theta$	
$\phi = 90°$, two waves $E \| XY$, one wave $E \| Z$	$d_{\text{eff}} = d_{31} \cos^2\phi + d_{32} \sin^2\theta$	
$\phi = 90°$, two waves $E \| X$, one wave $E \| YZ$	$d_{\text{eff}} = d_{31} \sin\theta$	
Some crystal symmetry groups with uniaxial birefringence		
point group	two e rays and one o ray	two o rays and one e ray
$\bar{4}2m$	$d_{\text{eff}} = d_{14} \sin^2 2\theta \cos 2\phi$	$d_{\text{eff}} = -d_{14} \sin\theta \sin 2\phi$
3*m*	$d_{\text{eff}} = d_{22} \cos^2\theta \cos 3\phi$	$d_{\text{eff}} = d_{15} \sin\theta - d_{22} \cos\theta \sin 3\phi$
6 and 6*mm*	$d_{\text{eff}} = 0$	$d_{\text{eff}} = d_{15} \sin\theta$

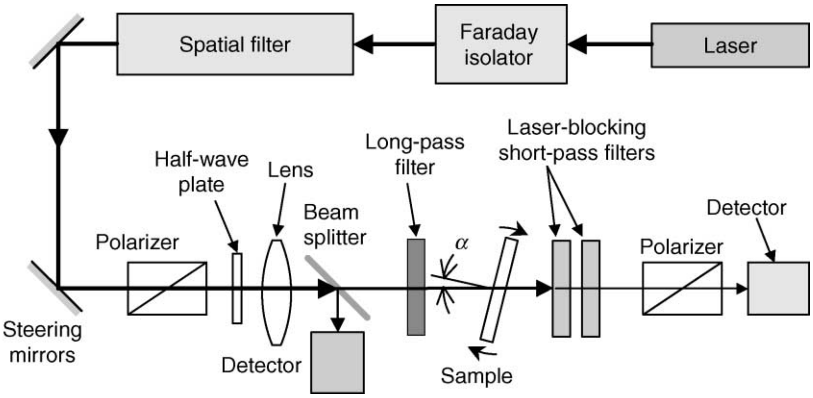

Figure 1 Schematic illustration of an experimental setup to record Maker-fringe measurements.

that group-velocity mismatch will be of no concern. Generating such laser output usually requires special techniques, such as injection-seeded pulse operation or stabilized single-frequency cw operation. Often spatial filters are required to improve the beam quality on the laser output. Even though these topics are only mentioned here, the techniques are important and demand much effort.

The analysis of measurements can also be complicated. Consider the Maker-fringe measurement technique (Figure 1). In this method a focused fundamental beam is incident on a thin slab of oriented nonlinear optical material for nonphase-matched second-harmonic generation. The thickness of the slab is large enough that the function $\sin(\Delta kz/2)$ has gone through many cycles. As the crystal is rotated both the path length in the material l and the wave vector mismatch Δk change, and $\sin^2(\Delta k \cdot l/2)$ oscillates between maxima and minima (Figure 2). The coherence length $L_c = \pi/\Delta k$ is normally small enough that conversion remains in the region of small depletion of the fundamental. Knowledge of the material dispersion and birefringence, required to calculate Δk, or L_c, can be obtained from the fringe spacing. Ideally, the incident fundamental beam will be focused to a spot size w_0, small enough to provide measurable harmonic conversion but large enough that the near-field approximation, in which diffraction is not significant, can be used for analysis of harmonic generation. In this case integration over the incident fundamental intensity distribution $I_\omega(r, z=0) = I_{\omega,0} \exp(-2r^2/w_0^2)$ yields an expression for the harmonic power:

$$W_{2\omega}(\alpha) = \frac{8\omega^2 d_{\text{eff}}^2(\alpha) W_\omega^2 L_c^2(\alpha)}{\pi^3 w_0^2 n^3 c^3 \varepsilon_0} \sin^2\left(\frac{\pi l(\alpha)}{2L_c(\alpha)}\right) \times \text{correction factors} \qquad [15]$$

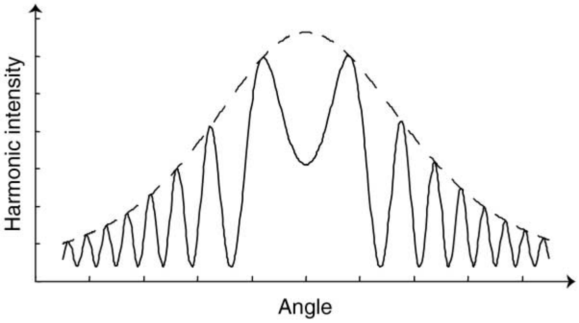

Figure 2 The envelope and spacing of the fringes allow determination of the second-order nonlinear optical coefficient and the coherence length.

where α is the angle of incidence of the fundamental radiation, $W_{2\omega}(\alpha)$ is the second harmonic power, and W_ω is the incident fundamental power. This equation applies to cw radiation. It is necessary to integrate over the temporal distributions for pulsed fundamental light. The effective nonlinear coefficient $d_{\text{eff}}(\alpha)$, the path length in the crystal $l(\alpha)$, and the coherence length $L_c(\alpha)$ change with the angle of incidence α. The correction factors include Fresnel losses at the crystal surfaces, harmonic generation at the surfaces, multiple reflections between the surfaces, absorption at either fundamental or harmonic wavelengths, and walkoff due to non-normal incidence and birefringence.

The wedge technique (Figures 3 and 4) also uses nonphase-matched second-harmonic generation. A wedged sample is translated through a focused fundamental beam, and the above equation applies with the angle α constant. The crystal length l varies with transverse position. It is important that the wedge angle and the fundamental beam spot size are chosen so that the entire beam fits well within a maximum of conversion and the beam is large enough to avoid divergence due to diffraction. Both the Maker-fringe

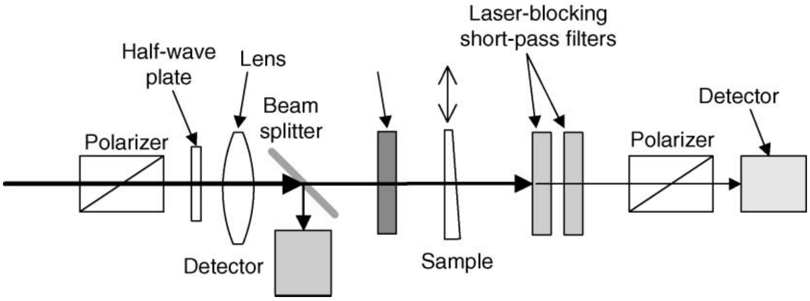

Figure 3 The setup for wedge measurements is similar to that used for Maker-fringe measurements but with a wedged sample translated in the fundamental beam to change path length.

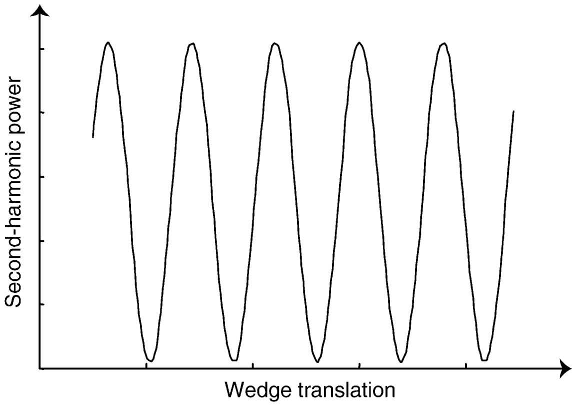

Figure 4 The harmonic intensity cycles between a minimum and maximum when the path is changed by one coherence length.

and wedge measurements have the advantage of being able to isolate individual components of the nonlinear optical tensor. Harmonic power can be low with cw beams requiring chopped, lock-in detection methods or even photon-counting techniques. Even with pulsed fundamental radiation, peak harmonic power may only be at the level of milliwatts. It is necessary to isolate the harmonic radiation from the fundamental radiation that could have peak power in the range of 10^5 W for a Q-switch laser.

Phase-matched second harmonic generation (Figure 5) offers a measurement technique that is similar to practical applications. Harmonic signals are large, thus simplifying detection. However, it is helpful to keep conversion levels below a few percent to allow the use of small depletion approximations in data analysis. Good optical quality of the nonlinear optical material and good quality and accurate characterization of the fundamental radiation are important in this measurement technique. The analysis is simplified for low levels of conversion, spectral distribution well within the phase-matching bandwidth, no absorption or other losses, and no significant diffraction or walkoff. If the fundamental intensity is a pulse with distribution of $I_\omega(r,t) = I_{\omega,0} \exp(-t^2/\tau^2)\exp(-2r^2/w_0^2)$, integration yields second-harmonic energy:

$$U_{2\omega} = \int_{-\infty}^{\infty} \mathrm{d}t \int_0^{\infty} 2\pi r\, \mathrm{d}r I_{2\omega}(r, z = l) \approx \frac{2\sqrt{2}\omega^2 d_{\mathrm{eff}}^2 l^2 U_\omega^2}{n^3 c^3 \varepsilon_0 \pi^{3/2} \tau w_0^2} \frac{\sin^2(\Delta k l/2)}{(\Delta k l/2)^2} \quad [16]$$

where $U_{2\omega}$ is the second harmonic energy and U_ω is the fundamental energy, r is the radial coordinate, t is time, and l is the length of the material. Harmonic conversion can be measured as Δk and varied, either by crystal rotation or temperature change, to confirm that optical quality of the nonlinear material and incident fundamental radiation are a level of tuning that is accurately predicted by theory (Figure 6). Even this simplified equation shows the importance of accurate characterization of the fundamental radiation and second-harmonic conversion in obtaining values of the nonlinear optical coefficient. Conversion efficiency is calculated more accurately with detailed analytical techniques or numerical methods, provided accurate and complete characterization of the fundamental radiation is available.

Parametric fluorescence, in principle, offers an experimental simplification over second-harmonic generation techniques. Parametric fluorescence (Figure 7) is a phase-matched technique in which a pump beam of angular frequency ω_p produces a signal beam and an idler beam of angular frequencies $\omega_s + \omega_i = \omega_p$ by the process of spontaneous parametric generation. It is only necessary to measure the ratio of the generated signal power W_s and the incident pump power W_p. The signal power is given by

$$W_s = \frac{2(h/2\pi)\omega_s^4 \omega_i d_{\mathrm{eff}}^2 n_s l W_p \Omega}{(2\pi)^2 \varepsilon_0 c^5 n_i n_p |\partial \Delta k/\partial \omega_s|_{\omega p}} \quad [17]$$

where h is Planck's constant, Ω is the solid angle of detection (inside the crystal), and n_s, n_i, and n_p are

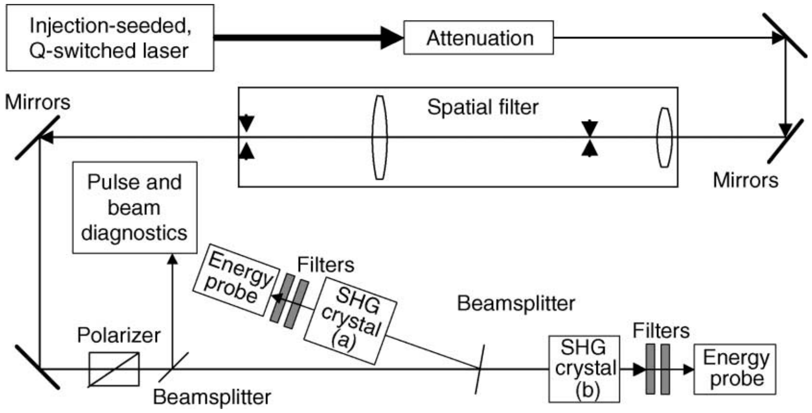

Figure 5 A setup for a relative measurement of second-order nonlinear optical coefficient based on phase-matched second-harmonic generation is shown.

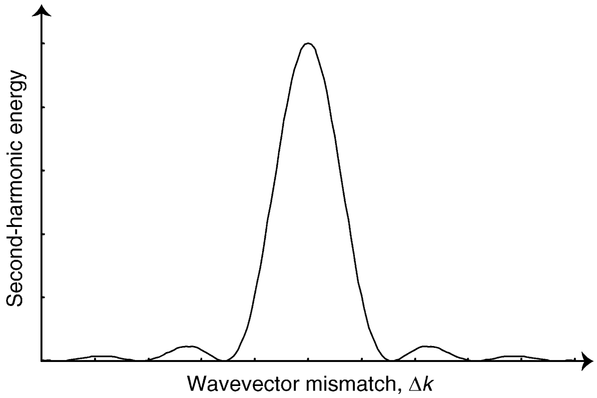

Figure 6 Observation of a theoretically predicted phase-matching tuning curve confirms required optical quality.

the indices of refraction of the signal, idler, and pump respectively. The absolute value of the partial derivative of $\Delta k = k_p - k_s - k_i$, with respect to signal frequency with the pump frequency held constant, is

$$|\partial \Delta k / \partial \omega_s|_{\omega p} = \frac{1}{c}\left| n_s - n_i + \omega_s \frac{dn_s}{d\omega_s} - \omega_i \frac{dn_i}{d\omega_i} \right| \quad [18]$$

It is still not well understood why the carefully performed earlier measurements by this method produced nonlinear optical coefficient values larger than those obtained by other techniques.

Tabulations of measurements of nonlinear optical coefficients are included in handbooks. The Internet posting ‘SNLO’ listed in the Further Reading is a good source of up-to-date values. Table 3 lists some nonlinear optical coefficients obtained from several sources. Signs of these coefficients are not always well determined, and they are reproduced where given and may only indicate relative polarity. Point group symmetry is given with some comments about axis assignments for the *mm*2 point group. The fundamental wavelength is given for second-harmonic measurements and twice the shortest wavelength is given for three-wave processes. Many of the values obtained by Shoji *et al.* are based on carefully performed wedge measurements with single frequency cw lasers and taking detailed account of multiple reflections in the analysis. Their measurements for congruent $LiNbO_3$ reported agreement between Maker-fringe and wedge measurements, and phase-matched difference-frequency generation measurements, and finally agreement with parametric fluorescence. The measurements of Alford and Smith combined optical-parametric amplification and second-harmonic generation at a number of different wavelengths and correlation with values previously reported in the literature to examine the validity of using a constant Miller’s delta, and reported best-fit values for the second-order nonlinear coefficient. Constant Miller’s delta provided reasonably accurate wavelength scaling, with good agreement for most materials in recent measurements.

The measurements of the nonlinear optical coefficients of $KNbO_3$, listed in Table 3 are of particular interest. These measurements were performed by the separated-beam wedge technique (Figure 8). With this technique, a large wedge angle is used in a relatively long crystal and with a larger fundamental beam than in the conventional wedge technique. The beam is large enough that many fringes are present in the second-harmonic at the exit surface. The fringe structure causes focused harmonic radiation to diffract into two clearly resolved spots. Integration of an earlier equation, with constant fundamental field amplitude $E_\omega(0)$, provides a simplified expression of the second-harmonic field $E_{2\omega}(z, t)$ in terms of a free wave with propagation constant

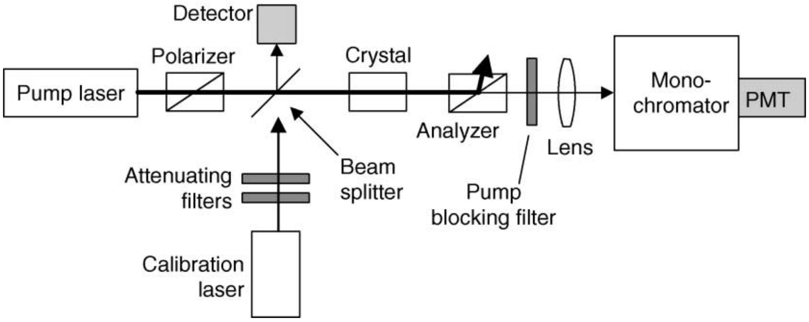

Figure 7 Spontaneous optical parametric fluorescence requires measurement of the ratio of the power of the signal fluorescent to the power of the pump radiation. The calibration laser is used to measure the signal transmission of the filtering system used to block the pump radiation from reaching the photomultiplier tube (PMT).

Table 3 Selected nonlinear optical coefficients

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
$AgGaS_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	11.2	10.6 μm	[a]
		$d_{14} = d_{24} = d_{36}$	17.5	1064 nm	[a]
$AgGaSe_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	33	10.6 μm	[a]
BaB_2O_4 (BBO)	$3m$	d_{22}	2.3	1064 nm	[a–f]
		$d_{16} = d_{21} = -d_{22}$	−2.3	1064 nm	[a]
		$d_{15} = d_{31} = d_{24} = d_{32}$	0.04	1064 nm	[d]
		d_{33}	0.04	1064 nm	[d]
CdS	$6mm$	d_{33}	32	10.6 μm	[a]
		$d_{15} = d_{31} = d_{24} = d_{32}$	−16	10.6 μm	[a]
		d_{33}	19.1	1064 nm	[e]
		$d_{15} = d_{31}$	10.1	1064 nm	[e]
		$d_{24} = d_{32}$	10.7	1064 nm	[e]
CdSe	$6mm$	d_{33}	36	10.6 μm	[a]
		$d_{15} = d_{31} d_{24} = d_{32}$	−18	10.6 μm	[a]
CdTe	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	109	1064 nm	[e]
		$d_{14} = d_{24} = d_{36}$	168	10.6 μm	[c]
$CsLiB_6O_{10}$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.74	1064 nm	[f]
GaAs	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	83	10.6 μm	[a]
			170	1064 nm	[e]
GaP	$\bar{4}3m$	$d_{14} = d_{24} = d_{36}$	37	10.6 μm	[a]
			70.6	1064 nm	[e]
KD_2PO_4 (KD^*P)	$\bar{4}2m$	d_{36}	0.37	1064 nm	[a,b]
KH_2PO_4 (KDP)	$\bar{4}2m$	d_{36}	0.39	1064 nm	[a,b,e,g,h]
$KNbO_3$	$mm2$	d_{33}	21.9	1064 nm	[h]
$Z = Z_L = x = l$		d_{15}	13	1064 nm	[h]
$X = Y_L = z = s$		d_{31}	12.4	1064 nm	[h]
$Y = X_L = -y = m$		d_{24}	9.2	1064 nm	[h]
		d_{32}	8.9	1064 nm	[h]
$KTiOAO_4$ (KTA)	$mm2$	$d_{15} = d_{31}$	4.2	1064 nm	[e]
$Z = Z_L = z = m$		d_{15}	2.91	2128 nm	[g]
$X = Y_L = -y = s$		$d_{24} = d_{32}$	2.8	1064 nm	[e]
$Y = X_L = x = l$		d_{33}	16.2	1064 nm	[e]
$KTiOPO_4$ (KTP)	$mm2$	d_{33}	14.6	1064 nm	[e]
$Z = Z_L = z = m$		d_{15}	3.7	1064 nm	[e]
$X = Y_L = -y = s$		d_{31}	3.7	1064 nm	[e]
$Y = X_L = x = l$		d_{24}	1.9	1064 nm	[e]
		d_{32}	2.2	1064 nm	[e]
LiB_3O_5 (LBO)	$mm2$	d_{33}	±0.04	1064 nm	[a,b]
$Z = Z_L = y = s$		$d_{15} = d_{31}$	±0.85	1064 nm	[a,b]
$X = Y_L = z = m$		d_{15}	1.04	1064 nm	[g]
$Y = X_L = x = l$		$d_{24} = d_{32}$	∓0.67	1064 nm	[a,b]
$LiIO_3$	6	d_{33}	4.5	1064 nm	[a,b]
		$d_{15} = d_{31}$	4.4	1064 nm	[a,b]
congruent	$3m$	d_{33}	−27.2	1064 nm	[a,b]
$LiNbO_3$		$d_{31} = d_{15} = d_{32} = d_{24}$	−4.35	1064 nm	[a,b]
		$d_{22} = -d_{21} = -d_{16}$	2.1	1064 nm	[a,b]

Table 3 (*Continued*)

Material	*Point group*	*Nonlinear optical coefficient (pm V^{-1})*		*Wavelength*	*Source*
Agreement of measurements of d_{31} by Maker fringe and wedge SHG (4.6 pm V^{-1}) and parametric fluorescence and parametric amplification (4.3 pm V^{-1})					
				1064 nm	[e]
5%MgO:LiNbO$_3$	3*m*	d_{33}	25	1064 nm	[e]
		d_{31}	4.4	1064 nm	[e]
LiTaO$_3$	3*m*	d_{33}	13.8	1064 nm	[e]
		d_{31}	0.85	1064 nm	[e]
NH$_4$H$_2$PO$_4$ (ADP)	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	0.47	1064 nm	[a,e]
αSiO$_2$	32	d_{11}	0.3	1064 nm	
ZnGeP$_2$	$\bar{4}2m$	$d_{14} = d_{24} = d_{36}$	69	10.6 μm	[a,b]

[a] Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.

[b] Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

[c] Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.

[d] Shoji I, Nakamura H, Ohdaira K, *et al.* (1999) Absolute measurement of second-order nonlinear-optical coefficients of β-BaB$_2$O$_4$ for visible to ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 16: 620–624.

[e] Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.

[f] Shoji I, Nakamura H, Ito R, *et al.* (2001) Absolute measurement of second-order nonlinear-optical coefficients of CsLiB$_6$O$_{10}$ for visible-to-ultraviolet second-harmonic wavelengths. *Journal of the Optical Society of America B* 18: 302–307.

[g] Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of KNbO$_3$, KTiOPO$_4$, KTiOAsO$_4$, LiNbO$_3$, LiIO$_3$, β-BaB$_2$O$_4$, KH$_2$PO$_4$, and LiB$_3$O$_5$ crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.

[h] Pack MV, Armstrong DJ and Smith AV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

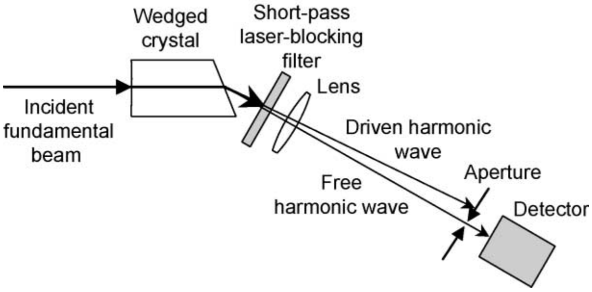

Figure 8 In the separated-beam wedge measurement technique, a fundamental beam of large transverse size generates harmonic radiation that is separated into a free harmonic wave and a driven harmonic wave after exiting the crystal.

$k_{2\omega}$ and a driven wave with propagation constant $2k_\omega$:

$$E_{2\omega}(z,t) = \frac{\omega d_{\text{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k}\{\exp i(k_{2\omega} z - 2\omega t) + c.c.\} - \frac{\omega d_{\text{eff}} E_\omega^2(0)}{n_{2\omega} c} \frac{1}{\Delta k}\{\exp i(2k_\omega z - 2\omega t) + c.c.\} \quad [19]$$

The two beams with different propagation constants are refracted at different angles at the wedged exit surface. The detailed analysis must include losses due to absorption and reflections at the surfaces and also nonlinear effects at the surfaces. Precision again is required in characterizing the fundamental beam and incident and generated intensities or energies as well as the geometry of the measurement. The large wedge eliminates problems of multiple reflections, and it is no longer necessary to carefully choose the beam spot size to fit within one fringe maximum, yet be large enough to allow use of the near field approximation. The accuracy of this measurement is demonstrated when the coefficients are expressed as Miller-delta values: $\delta_{24} = 0.153\,\text{pm}\,\text{V}^{-1}$, $\delta_{32} = 0.150\,\text{pm}\,\text{V}^{-1}$, $\delta_{15} = 0.195\,\text{pm}\,\text{V}^{-1}$, and $\delta_{31} = 0.192\,\text{pm}\,\text{V}^{-1}$. These values satisfy Kleinman symmetry, $\delta_{24} = \delta_{32}$ and $\delta_{15} = \delta_{31}$, well within the reproducibility of the measurement, which was about ±3%.

The nonlinear optical coefficients of BBO d_{22}(1064 nm) = 2.3 pm V^{-1} and KDP d_{36}(1064 nm) = 0.39 pm V^{-1} are gaining agreement from a number of independent measurements. The apparent disparity in the values for GaAs d_{36}(10.6 μm) = 83 pm V^{-1} and d_{36}(1064 nm) = 170 pm V^{-1}, is resolved when the coefficients are converted to Miller's delta values: $\delta_{36} = 0.089\,\text{pm}\,\text{V}^{-1}$ and $\delta_{36} = 0.086\,\text{pm}\,\text{V}^{-1}$, respectively. The same thing can be done for GaP, for which d_{36}(10.6 μm) = 37 pm V^{-1} and d_{36}(1064 nm) = 70.6 pm V^{-1} convert to $\delta_{36} = 0.085\,\text{pm}\,\text{V}^{-1}$ and $\delta_{36} = 0.077\,\text{pm}\,\text{V}^{-1}$, respectively. However, the older and newer values for other materials such as CdTe and CdS do not achieve this agreement. Other materials such as CLBO and LBO need additional investigation to resolve differences. The measurement of nonlinear optical coefficients is a continuing process.

Other techniques for measurement involving different nonlinear-optical processes are possible. Parametric amplification was mentioned above. Generally more complicated processes are less well suited for measurement of fundamental properties. Integrated-optical devices such as waveguides are even more complex. Often nonlinearity is measured in fractional conversion per unit of pump power to characterize overall device performance. There are survey techniques for evaluating materials such as grinding the material to a powder, illuminating with intense laser radiation, and looking for harmonic conversion. A surprising amount of information, such as a rough estimate of nonlinearity and the possibility of birefringent phase matching, can be gleaned by sorting the power and measuring the scattered harmonic radiation. Second-harmonic reflections from surfaces and thin film can be used in some instances. Electric fields can be use to orient molecules in polymers and solute molecules in solution to measure the hyperpolarizability of individual molecules.

Reporting Frames

Standardized axis specifications avoid confusion, but nonstandard axes can prove convenient in some applications provided care is used. The crystal $KNbO_3$ has *mm*2 point group; it could also be *m*2*m* or 2*mm* depending on the choice of axes to describe the crystal properties. The axis assignment for the orthorhombic class *mm*2, as it is used in Table 3, follows the IEEE/ANSI Std. 176-1987: Z is the polar axis, X is parallel to the shorter of the remaining unit cell dimensions, and Y is chosen to complete a right-handed coordinate frame. Polarity of the axes can be assigned by pyroelectric properties and piezoelectric properties of the crystal. The common assignments of the axes in the literature are given by X_L, Y_L, Z_L in the table. Reference to the biaxial indices of refraction are given by the lower case x, y, z with $n_x < n_y < n_z$. Finally the axes are identified by the unit cell dimensions $s < m < l$. Either a plus or minus rotation about the Y axis will bring the crystal into an xyz frame. However, a $+90$ degree rotation will give coefficients that are opposite in sign from a -90 degree rotation. The signs of the nonlinear optical coefficients are often not well determined because sign is not obtained from conversion efficiency in a single crystal. However, coefficient signs can be compared in the same crystal or between two crystals to determine if they are the same or opposite.

The crystals KTP or KTA are used here as an example. A 90° rotation about the Z-axis gives a new coordinate system $X'Y'Z'$ aligned with the xyz frame determined by the indices of refraction. In the new frame, the component of the nonlinear optical tensor, in terms of the standard frame, are $d_{zxx} = d_{32}$ and $d_{zyy} = d_{31}$. All that is needed here is to exchange d_{32} and d_{31} to change the standard frame to the one commonly used in the literature, X_L, Y_L, Z_L. When the largest index of refraction is not aligned with the polar axis, as is the case for LBO and $KNbO_3$, the specification of tensor components can become confusing if a standard frame is not used.

Often crystallographers and crystal manufacturers can identify axes by surface morphology or by poling processes of ferroelectric crystals. With other crystal symmetries a change in the reporting frame that does not change linear optical properties may change the sign of some of the nonlinear optical tensor components while leaving others unchanged. For a crystal of *3m* point group, a 180° rotation about Z will change the sign of d_{22}, d_{21}, and d_{16} and leave the other coefficients unchanged. The xyz frame with $n_x < n_y < n_z$ is convenient for calculations of optical propagation, and the standard XYZ frame is unambiguous for reporting of tensor properties.

Conclusion

There is still a range of values for nonlinear optical coefficients of many materials. Standardization is evolving on more commonly used materials. Greater precision in measurement and extension of measurement techniques are providing better accuracy. However, caution is still required when accessing values. It is necessary to scrutinize the sources of values to obtain the highest accuracy. Accurate measurements need to be extended to difficult spectral ranges of the ultraviolet and infrared. The ambiguity of the sign of tensor components needs to be resolved. In recent measurements, carefully executed wedge measurements are producing more reliable results. The separated beam wedge technique appears to be particularly promising. Phase-matched second-harmonic generation and other practical applications of nonlinear optical materials will continue to serve as techniques to verify accuracy. Current analytical and numerical techniques provide a method for precise modeling of nonlinear optical processes and the need for more accurate material characterization.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation.

Further Reading

Alford WJ and Smith AV (2001) Wavelength variation of the second-order nonlinear coefficients of $KNbO_3$, $KTiOPO_4$, $KTiOAsO_4$, $LiNbO_3$, $LiIO_3$, β-BaB_2O_4, KH_2PO_4, and LiB_3O_5 crystals, a test of Miller wavelength scaling. *Journal of the Optical Society of America B* 18: 524–533.

Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

Gehr RJ (1998) Separated-beam non-phase-matched second-harmonic method of characterizing nonlinear optical crystals. *Journal of the Optical Society of America B* 15: 2109–2116.

Kurtz SK (1975) Measurement of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise, vol. 1, part A*, pp. 209–281. New York: Academic Press.

Pack MV (2003) Measurement of the $\chi^{(2)}$ tensor of the potassium niobate crystal. *Journal of the Optical Society of America B* 20: 2109–2116.

Roberts DA (1992) Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions. *IEEE Journal of Quantum Electronics* 28: 2057–2074.

Shoji I (1997) Absolute scale of second-order nonlinear-optical coefficients. *Journal of the Optical Society of America B* 14: 2268–2294.

SNLO nonlinear optics code available from Smith AV Sandia National Laboratories, Albuquerque, NM 87185-1423. (an Internet publication) http://www.sandia.gov/imrl/XWEB1128/xxtal.htm.

Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn, chs 4 and 5. New York: Marcel Dekker.

$\chi^{(3)}$

P P Banerjee, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Over the past 30 years, several methods have been used to measure and characterize optical nonlinearities. With increasing interest in nonlinear optics beyond the traditional harmonic generation, third-order nonlinear processes have been shown to be responsible for self-focusing and defocusing of optical beams, phase conjugation, optical bistability, etc. Third-order nonlinear processes in Kerr materials arise from the cubic dependence of the polarization $P = P^{\mathrm{L}} + P^{\mathrm{NL}}$, $P^{\mathrm{NL}} = \chi^{(3)} \vdots E{\cdot}E{\cdot}E$ in a material on the optical field E, through the nonlinear susceptibility tensor $\chi^{(3)}$. The cubic nonlinearity in turn results in a nonlinear refractive index coefficient n_2 defined through:

$$n = n_0 + n_2 |E_{\mathrm{p}}^{\omega_0}|^2 \qquad [1]$$

where $E_{\mathrm{p}}^{\omega_0}$ is the phasor part of $E : E = Re[E_{\mathrm{p}}^{\omega_0} \exp \times \mathrm{j}\omega_0 t]$. The relation between $\chi^{(3)}$ and n_2 can be expressed as:

$$n_2 = (3/8n_0)\chi^{(3)}_{1111}(-\omega_0; \omega_0, \omega_0, -\omega_0) \qquad [2]$$

In some materials, an effective $\chi^{(3)}$ or an effective n_2 can be defined as well. For instance, in a second-order nonlinear material, an effective $\chi^{(3)}$ is established through cascading of $\chi^{(2)}$s, and in the presence of phase mismatch between the fundamental and the second harmonic. In other materials, such as liquid crystals, an effective n_2 can be defined due to the nonlinear dependence of the reorientation of the director axis on the optical field. In photorefractive materials, the induced nonlinearity stems from the nonlinear dependence of the induced space charge distribution in the material on the optical field. Some of the techniques used to measure the third-order nonlinearity in materials include degenerate four-wave mixing (DFWM) or phase conjugation, interferometry, ellipse rotation, and beam self-refraction and/or distortion and bending.

For instance, in DFWM, two contradirectional propagating pumps and the probe beam having the same frequency ω_0 interact to give rise to the conjugate, whose amplitude depends on the magnitude of the cubic nonlinearity coefficient. The conservation of energy and momentum describes a parametric interaction of the form $\omega_0 + \omega_0 - \omega_0 \rightarrow \omega_0$ (in frequency space) and, $\mathbf{k}_{\mathrm{probe}} + \mathbf{k}_{\mathrm{pump1}} - \mathbf{k}_{\mathrm{pump2}} \rightarrow \mathbf{k}_{\mathrm{conj}}$ (in $\mathbf{k}$-space). The conjugate beam is clearly distinguishable by its spatial separation from the other interacting waves. However, alignment of the three waves is critical to obtain a good phase conjugate; furthermore, the sign of the nonlinearity cannot be determined.

In the interferometric method, a laser beam is split into two parts, one serving as the reference, and the other being incident on the nonlinear sample. As in the case of interferometers, if the two beams are then recombined and introduced through a lens, an interference pattern forms at the focal plane of the lens. While the information about the magnitude of the nonlinearity can be found from the fringes and their change when the incident power is varied, information of the sign of the nonlinearity is still not readily inferable.

In this article, we will concentrate on self-refraction, particularly in light of the fact that it results in a sensitive and at the same time simple method called the *z*-scan©, requiring the propagation of a single beam through the nonlinear sample. We comment that self-diffraction does also cause self-bending of an asymmetric beam due to the asymmetric phase change induced through the cubic nonlinearity; however, extracting the information of the sign and magnitude of the nonlinearity can be effectively done only if the sample is thin, implying that there is negligible linear diffraction of the beam as it passes through the sample. In the rest of the article, we will therefore concentrate on the *z*-scan method, since it is an elegant and simple single (symmetrical) beam technique that yields information on the magnitude and sign of the cubic nonlinearity, as well as its real and imaginary parts. We will describe in detail how the *z*-scan can measure the real part of the cubic nonlinearity, for both thin and thick samples, as well as showing its application in determining the reorientational nonlinearity of liquid crystals and the nonlinearity of photorefractive materials, where photovoltaic effect and diffusion of charge carriers contribute to the effective nonlinearity.

Since the mid-1990s *z*-scan has evolved as a standard method for determining effective nonlinear refractive index coefficients of nonlinear materials. The name *z*-scan is derived from the fact that the nonlinear test sample is scanned along the direction of propagation (z) of a beam about the back focal plane of an external lens illuminated by an incident Gaussian beam from a laser. The far-field diffraction pattern yields information about the nonlinearity of the test material. The *z*-scan technique has been used to measure and characterize the nonlinear refractive index coefficient and nonlinear absorption coefficient in materials such as CS_2, ZnSe, Cs atoms, multiphoton absorption in GaN, organic dyes, thermal nonlinearities in liquids, liquid crystals, photorefractive materials, etc., to name a few. *z*-scan measurements have been made using non-Gaussian incident beams, and *z*-scans have been performed using both CW and pulsed laser sources.

The first extensive series of work on the *z*-scan technique, including analysis and application to the measurement of the cubic nonlinearity, was performed by the group of Van Stryland at the Center of Electro-Optics and Lasers (CREOL) at the University of Central Florida. Other investigators later performed similar measurements and characterization during the study of nonlinear optical properties of a solution of chlorophyll in acetone, and Chinese tea.

Theory of the Thin Sample *z*-Scan

We will first analyze propagation of a (circularly symmetric) Gaussian beam which is initially focused by an external lens and is then incident on a 'thin' slice of a cubically nonlinear Kerr medium. This case is important because, as explained below, it provides a basis for the measurement of the nonlinear refractive index coefficient n_2.

Figure 1 shows a simple ray diagram to demonstrate what happens if the cubically nonlinear material is thin. The propagating Gaussian beam induces its own lens, which is positive for a material having a positive n_2, for reasons that will become clear to the reader later on in this section. Note that if the position of the induced lens is at the back focal plane of the external lens, the former does not affect the propagation of the beam behind the nonlinear material. The situation is thus the same as the propagation of the externally focused Gaussian beam in a linear diffraction-limited environment. However, if the separation between the external lens and the sample is less than the focal length f of the external lens, the beam in the far field is wider than the linear diffraction-limited case. Conversely, if the lens-sample separation is larger than f, the beam width is smaller. This is the basis of the *z*-scan: for a material with a positive n_2, the scanning of a thin nonlinear sample through the back focal plane of the external lens, starting from a position to the left of the focus, yields far-field profiles which are initially larger and eventually smaller than the width in the linear diffraction-limited case. The opposite is true for materials with a negative n_2. For very large displacements of the sample from the back focal plane of the

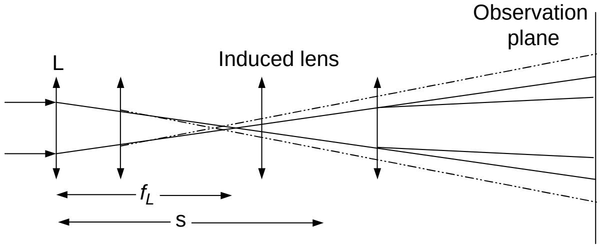

Figure 1 Geometrical optics picture of *z*-scan setup for nonlinear refractive index n_2. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

external lens, the beam power is small enough that nonlinear lensing is negligible, and hence the beam widths return to the linear diffraction-limited case. The variation of the on-axis intensity or transmittance as a function of scan distance is complementary to the variation of the width, due to conservation of power, to be explained further below. A sketch of the on-axis transmittance during a z-scan for a positive n_2 material, is shown in Figure 2.

The simpler ray picture described above can be augmented by using a q parameter approach to Gaussian beam propagation. For the sake of simplicity, we prefer to introduce the q parameter approach here, instead of the more rigorous Gaussian decomposition of the induced phase as computed by Van Stryland's group. The q of a Gaussian beam in air is defined through the relation $1/q = 1/R - \mathrm{i}(2/k_0 w^2)$, where R and w denote the radius of curvature and the beamwidth, respectively, and k_0 is the propagation constant in air. Assume a Gaussian beam with a waist w_0 (implying initially planewave fronts) incident on the external lens. The Gaussian can be described by $q = q_0 = \mathrm{i}z_{R0}$, where $z_{R0} = k_0 w_0^2/2$ is called the Rayleigh length of the incident Gaussian. The laws of q transformation of a Gaussian beam are:

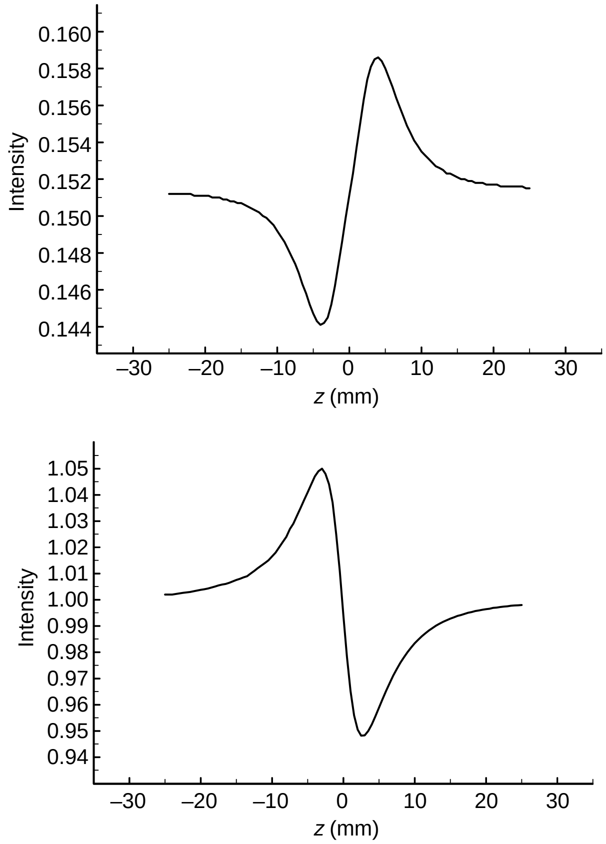

Figure 2 Typical z-scan plots for positive (top) and negative (bottom) cubic nonlinearities. The vertical axis is in arbitrary units.

(i) for translation through a distance d, the q changes to $q + d$; and
(ii) for a Gaussian with q-parameter q incident on a lens, the q immediately behind the lens, is given by $(1/q - 1/f)^{-1}$.

Using these laws, the $q = q_{\mathrm{obs}}$ of the Gaussian at a distance d from the nonlinearly induced lens of the focal length f_{ind}, which is located at a distance $s = f + \Delta s$ behind the external lens, can be found as:

$$q_{\mathrm{obs}} = \frac{f_{\mathrm{ind}}[q_0 f + (f - q_0)(f + \Delta)]}{f_{\mathrm{ind}}(f - q_0) - [q_0 f + (f - q_0)s]} + d \qquad [3]$$

In deriving eqn [3], we have used the laws of q transformation on the initial Gaussian beam of q-parameter q_0 passing through the external lens, being translated a distance $f + \Delta s$, passing through the induced lens, and finally being translated a distance d to the observation plane.

To determine $f_{\mathrm{ind}}(\Delta s)$, assume a Gaussian of the form:

$$|E_{\mathrm{p}}^{\omega_0}(x, y, z)| = E(z)\exp(-(x^2 + y^2)/w^2(z)) \qquad [4]$$

Since

$$n = n_0 + n_2|E_{\mathrm{p}}^{\omega_0}|^2 \approx n_0 - 2n_2E^2(z)(x^2 + y^2)/w^2(z) \qquad [5]$$

the nonlinearly induced phase change is

$$\Delta\phi = k_0 n\Delta z \approx k_0[n_0 - 2n_2E^2(z)(x^2 + y^2)/w^2(z)]\Delta z \qquad [6]$$

so that

$$f_{\mathrm{ind}}(z) = n_0 w^2(z)/4n_2 a^2(z)\Delta z \qquad [7]$$

Note that $f_{\mathrm{ind}}(z)$ is inversely proportional to $n_2E^2(z)$ and Δz. The longer the sample length, the more the nonlinear induced phase, implying a stronger lensing (smaller f_{ind}). Also, the stronger the nonlinearity parameter n_2 and/or on-axis intensity ($E^2(z)$), the stronger the focusing since the change in n is larger.

We must point out that the focal length f_{ind} of the induced lens is also a function of Δs, since it depends on the beam intensity at the location of the nonlinear sample. In fact, the focal length f_{ind} of the nonlinearly induced lens can be expressed in terms of the beam power P, the beam width w at the position Δs, f (the focal length of the external lens), the linear refractive index n_0, the nonlinear refractive index coefficient n_2 defined through eqn [1] where $|E_{\mathrm{p}}^{\omega_0}| = E$ is the optical field in the nonlinear material of length L, assumed 'thin' for now. The term 'thin' means that the sample length is smaller than the depth of focus of the externally focused Gaussian beam. In this case, we can assume that the beam width is a constant through the sample, and the induced focal length is then only directly dependent on the total thickness L of the sample. The focal length of a nonlinearly induced 'thin' Kerr lens can be written as

$$f_{\mathrm{ind}}(\Delta s) = \frac{n_0 w^2(\Delta s)}{4n_2|E|^2 L} \approx \frac{\pi f^4 n_0}{k_0^4 w_0^4 \eta_0 P L n_2} \qquad [8]$$

for small Δs, where η_0 is the (linear) characteristic impedance of the nonlinear medium. Note that in eqn [8], the beam width w is given by

$$w^2(\Delta s) = w_{\mathrm{f}}^2\left[1 + \frac{4(\Delta s)^2}{k_0 w_{\mathrm{f}}^4}\right] \text{ and } w_{\mathrm{f}} \cong \frac{2f}{k_0 w_0} \qquad [9]$$

In eqn [9], w_{f} denotes the waist size of the Gaussian beam at the back focal plane of the external lens. In deriving eqn [8], one has to use the power conservation relation $|E|^2 w^2 = (4/\pi)\eta P$ to re-express the dependence of the induced focal length on the field amplitude in terms of the corresponding width at the same point.

Our final interest is to find the imaginary part of $1/q_{\mathrm{obs}}(= -2/k_0 w_{\mathrm{obs}}^2)$, in order to ascertain the width w_{obs} of the Gaussian beam on the observation plane. Simple but tedious algebra leads us to the relation:

$$\frac{k_0 w_{\mathrm{obs}}^2}{2} = \frac{[(ff_{\mathrm{ind}}d - f^2 d + f^2 f_{\mathrm{ind}}) + f(f_{\mathrm{ind}} - d)\Delta s]^2 + z_{R0}^2[(f_{\mathrm{ind}} - d)\Delta s + f_{\mathrm{ind}}d]^2}{z_{R0} f^2 f_{\mathrm{ind}}^2} \qquad [10]$$

It can be shown simply that when $\Delta s = 0$, and for $|f_{\mathrm{ind}}| \gg f$, $w_{\mathrm{obs}} = w_0 d/f$ for large d. In other words, in the linear case, i.e., when $f_{\mathrm{ind}} = \infty$, as well as in the nonlinear case when $\Delta s = 0$, the far-field width of the Gaussian beam is what is predicted by geometrical optics. Furthermore, from eqn [10], the rate of change of the width with Δs around $\Delta s = 0$, (assuming $z_{R0} \gg f$, and for large d) can be found as

$$\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} = \frac{-2d^2 z_{R0}}{k_0 w_{\mathrm{obs}} f^2 f_{\mathrm{ind}}} = -\frac{w_0 d}{f f_{\mathrm{ind}}} \qquad [11]$$

provided we invoke the relation for w_{obs} derived above. In eqn [11], f_{ind} refers to the value of the induced focal length (eqn [8]) with $\Delta s = 0$. Equation [11] may be written as

$$\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} = -\frac{k_0^4 w_0^5 P L n_2}{2\pi n_0 f^5} \qquad [12]$$

This means that in a material with a positive (negative) n_2, implying a positive (negative) induced lens, the rate of change of the observed width with the position of the nonlinear sample about the focal plane of the external lens will be negative (positive), in agreement with the results from geometrical optics. Furthermore, the rate of change is proportional to the angle of convergence w_0/f of the externally focused Gaussian, and to the position d of the observation plane.

In z-scan, it is often important to relate the on-axis transmittance in the far field to the scan distance s for measurement purposes. Since for power conservation, the intensity is inversely proportional to the width of the beam in the far field, a positive n_2 would imply a positive slope as a function of scan distance with increasing s, and a negative slope for negative n_2. The value of the slope of the on-axis transmittance could give the information of the n_2 of the material. Note that the z-scan graph in this case is an odd function of the scan distance, measured about the back focal plane of the external lens.

In practice, a detector with a finite aperture area is placed on-axis, and the detected power monitored as a function of scan distance. The transmitted power through the aperture is obtained by

$$\begin{aligned} P_{\mathrm{T}} &= \frac{1}{2}\int_0^{2\pi}\int_0^{r_{\mathrm{a}}} \frac{E_{\mathrm{a}}E_{\mathrm{a}}^*}{\eta_0}\,\mathrm{d}A \\ &= \frac{\pi w_0}{2}\left(\frac{E_0^2}{2\eta_0}\right)\left(1 - \exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\right) \\ &= P\left(1 - \exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\right) \end{aligned} \qquad [13]$$

where r_{a} is the radius of aperture. The rate of change of P_{T} with Δs can be found as

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s} = \frac{-4r_{\mathrm{a}}^2 P}{w_{\mathrm{obs}}^3}\exp\left(-\frac{2r_{\mathrm{a}}^2}{w_{\mathrm{obs}}^2}\right)\frac{\mathrm{d}w_{\mathrm{obs}}}{\mathrm{d}\Delta s} \qquad [14]$$

Substituting eqn [12] into eqn [14] and using the approximation $w_{\mathrm{obs}} = w_0 d/f$, eqn [14] can be rewritten as

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s} \cong \frac{4r_{\mathrm{a}}^2 f^2 P}{f_{\mathrm{ind}} d^2 w_0^2}\exp\left(-\frac{2r_{\mathrm{a}}^2 f^2}{d^2 w_0^2}\right) \qquad [15]$$

from which

$$\frac{\mathrm{d}P_{\mathrm{T}}}{\mathrm{d}\Delta s}(\Delta s \approx 0) \cong \frac{4k_0^4 w_0^2 \eta_0 r_{\mathrm{a}}^2 P^2 L}{\pi f^2 d^2 n_0}\exp\left(-\frac{2r_{\mathrm{a}}^2 f^2}{d^2 w_0^2}\right) n_2 \qquad [16]$$

When r_a approaches 0;

$$\frac{dP_T}{d\Delta s}(\Delta s \approx 0) = 4\left(\frac{k_0^2 w_0 r_a P}{fd}\right)^2 \frac{\eta_0 L}{\pi n_0} n_2 \qquad [17]$$

Therefore, n_2 can be estimated as

$$n_2 = \left(\frac{\pi n_0}{\eta_0 L}\right)\left(\frac{d}{2k_0^2 w_0 P}\right)^2 \left(\frac{f}{r_a}\right)^2 \frac{dP_T}{d\Delta s} \qquad [18]$$

Thus, if one knows the slope of the z-scan graph around $\Delta s \approx 0$, the value of n_2 can be found. Another way to state this result is to say that the n_2 is related to the peak-to-valley change in the on-axis transmittance ΔT_{p-v}, which is proportional to $dP_T/d\Delta s$.

In passing, we comment that the z-scan method can be also adapted to determine the nonlinear absorption coefficient α_2 of a material, defined through the relation $\alpha = \alpha_0 + \alpha_2 I$, where α is the total absorption coefficient and I denotes the intensity. In order to achieve this, the aperture is assumed to be completely open, in order to collect all the light behind the nonlinear absorptive sample. For $\alpha_2 > 0$, maximum absorption is experienced when the nonlinear sample, assumed thin, is at the back focal plane of the lens since the optical intensity is a maximum due to the external focusing. The z-scan graph is an even function of scan distance, measured about the back focal plane of the external lens.

Thick Sample *z*-Scan

Contrary to the thin sample case, a thick sample is defined as one whose thickness may be comparable to the depth of focus of the externally focused Gaussian beam. In this case, the beamwidth changes during propagation through the sample, so that the induced focal length also changes with propagation. This makes the problem substantially more involved. Indeed, to track the far-field profile in this case, one needs to describe the propagation of the Gaussian beam inside the nonlinear sample using the q parameter formalism. Since both the width and the radius of curvature of the beam at the exit face of the sample are required to predict the far-field behavior, one can write a simple differential equation describing the variation of the q through the nonlinear sample. This equation can be composed from the laws of q-transformation described earlier, and the effects of propagation and nonlinear lensing can be added assuming each is small over an infinitesimal distance of propagation, much like the numerical method called the split step beam propagation method (BPM), often used to analyze such problems. We will not present details of the computation, rather show calculation results for a typical thick sample case, and point out differences with the thin sample predictions for z-scan.

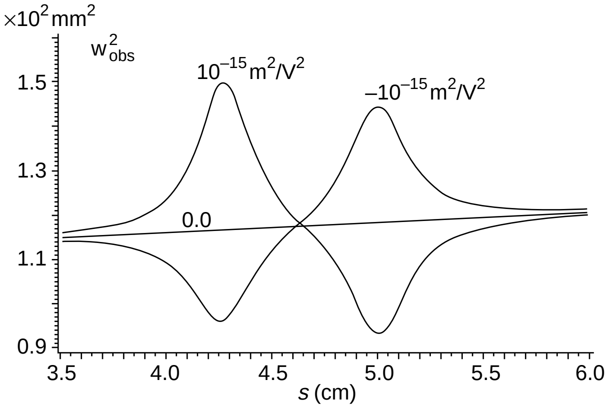

Figure 3 Plot of width of the beam in the far field as a function of the scan distance s. The light wavelength is 640 nm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

Figure 3 shows the beamwidth as a function of scan distance for the case where the observation plane is at a distance $d = 1$ m from the exit face of a 1 cm thick. Assume that a hypothetical sample with $n_0 = 1$, $n_2 = \pm 10^{-15}$ m^2/V^2 is scanned about the back focal plane of an external lens with focal length $f = 5$ cm. For comparison, the case of a linear sample is also superposed on the same graph. Note that with the variation of s, the far-field width initially increases, and then decreases before starting to increase again, for a material with a positive n_2. Note that around the point where the variations intersect the graph for the linear case, the slope of the curves are negative and positive for materials with positive and negative n_2, respectively. The graphs for the variation of on-axis transmittance are opposite to the variation of the widths, as mentioned earlier. This is also in agreement with the results predicted from geometrical optics and the q formalism for a thin sample, described earlier.

How do thick and thin sample z-scans differ? Important differences are: (i) the position of the points where the graphs for positive and negative, n_2s, intersect the graph for the linear case; and (ii) the slope of the graphs, predicted from thin and thick sample calculations. For instance, for the thin sample approximation, the value of the slope around $s = f$ is about -4 for a material with a $n_2 = 10^{-15}$ m^2/V^2. Recall from the previous section, the z-scan plot for a thin sample always intersects the graph for the linear case at the back focus of the external lens, and therefore, it is easy to determine the slope of the z-scan at this point. From the graph in the thick sample case, it is perhaps difficult to ascertain at exactly which point one must determine the slope. A figure of merit would be the minimum value, which occurs for a value of s slightly smaller than $f = 5$ cm. A conservative estimate yields a slope of approximately -1 for this particular example. Note, however, that in the thick lens analysis, we have taken a paraxial approximation to the Gaussian at every stage of its propagation through the nonlinear sample. Strictly speaking, this is not true, and to alleviate this problem, Van Stryland's group included a correction factor with the width of the Gaussian beam.

We would like to mention that values of $n_2 \approx -10^{-15}$ m^2/V^2 can indeed be measured for thermal nonlinearities in liquids. In fact, this is a typical value that we measured for our solution of chlorophyll in ethanol during our first experiments on z-scan. Values of effective $n_2 \approx 10^{-14}$ m^2/V^2 can be measured for liquid crystals with orientational nonlinearities, and will be described below.

Example of *z*-Scan: Effective Cubic Nonlinearity of Liquid Crystals

The induced refractive index in a nematic liquid crystal, due to reorientational nonlinearity, is a function of position in the crystal and depends on the intensity of the incident optical plane wave, its angle with respect to the direction of propagation, and on the applied voltage. Because of the fact that the nonlinearity depends on the direction of propagation of the planewave and the director axis of the crystal, modeling Gaussian beam propagation is rather complicated. Moreover, the director axis has a certain distribution in the material due to the applied voltage, and hence, the electrostatic field in the material, as well as the optical field. This poses a level of complexity a degree higher than that discussed above for the simple thick sample case. Since the induced refractive index is not a constant, and more importantly, a function of the spatial frequency of the optical beam (which determines the angle between a planewave component of the Gaussian beam spectrum and the director axis distribution under external field), one is compelled to analyze optical propagation using split-step BPM. A typical set of simulated z-scan plots is shown in Figure 4, assuming one transverse dimension for simplicity, and the same parameters as in the experiment described below. Actual experimental results from a z-scan of a biased liquid crystal sample is shown in Figure 5 for transmitted power versus sample position in a z-scan experiment for a liquid crystal sample for applied voltages 0.8, 1.0, 1.2, 1.8, and 2.2 V. In this case, the focal length of the external lens is approximately 50 mm and aperture diameter on the observation plane is 1 mm. The scan range is 0.15 inches or approximately 4 mm. The scan step, ΔZ, is varied from about 0.002 inches to about 0.01 inch, depending on the position of the sample. This is done so as to get the maximum information around the back focal plane of the external lens where the measurement values are to be used to find the slope of the z-scan graph, and hence the effective nonlinear refractive index coefficient.

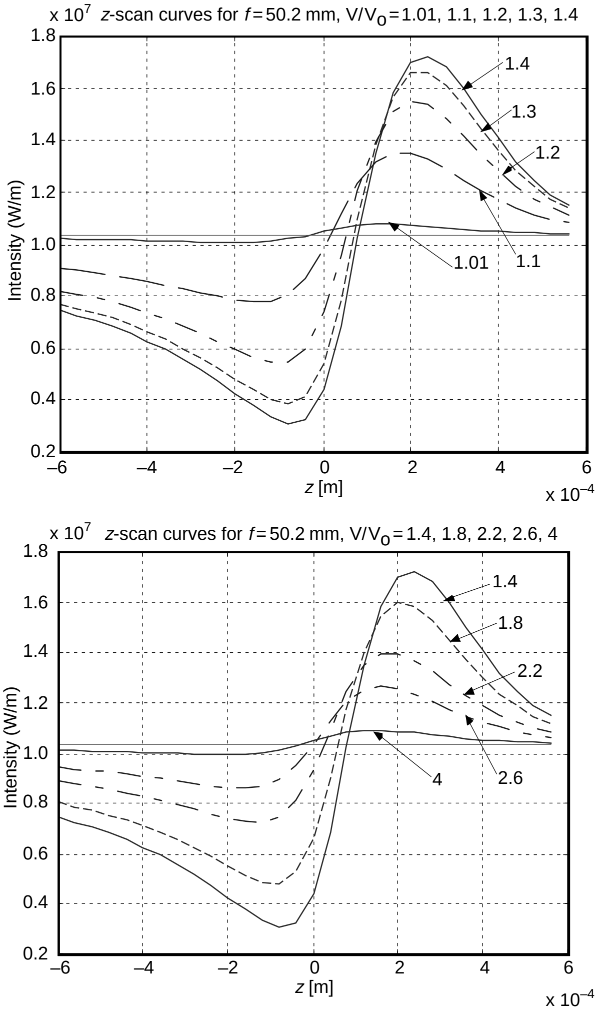

Figure 4 z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.01$, 1.1, 1.2, 1.3, 1.4 with $E_{op} = 2 \times 10^5$ V/m. z-scan simulation results for $f = 50.2$ mm and $V/V_o = 1.4$, 1.8, 2.2, 2.6, 4 with $E_{op} = 2 \times 10^5$ V/m. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

The plots of the effective $n_2 = n_{2\text{eff}}$ (simulated and measured) are shown in Figure 6a and b, respectively. As seen from both plots, the $n_{2\text{eff}}$ increases and

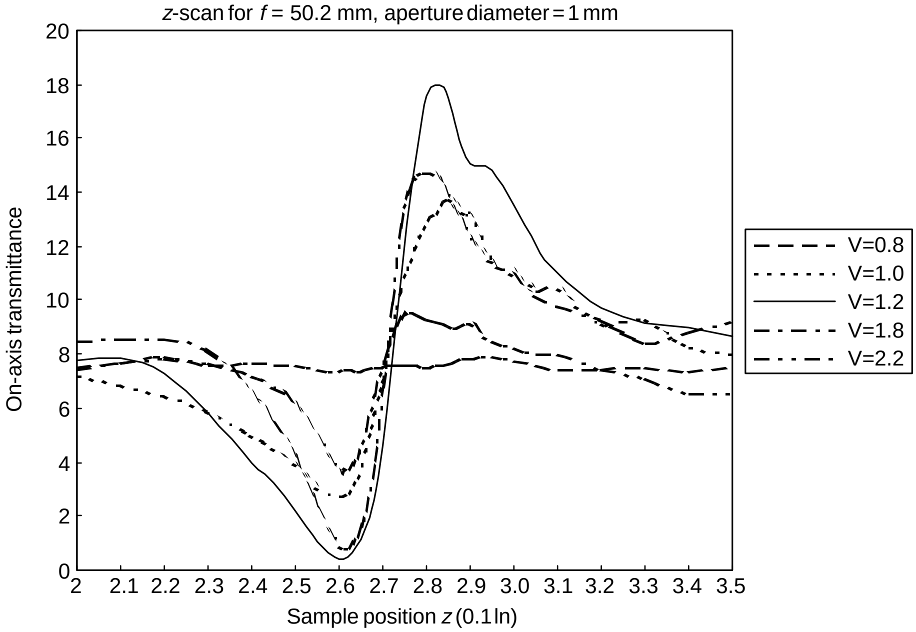

Figure 5 Measured *z*-scan graph for $f = 50.2$ mm and aperture diameter = 1 mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

reaches a maximum at $V/V_o = 1.4$, before starting to decrease. Simulated and experimental $n_{2\mathrm{eff}}$ have similar trends. Furthermore, there is reasonable agreement between the calculated and measured values of $n_{2\mathrm{eff}}$.

Discussion: Effective n_2 of Other Materials

The *z*-scan is a powerful tool in the measurement and characterization of effective cubic nonlinearities of different materials. The method is simple and easy to set up, and can be readily automated. We provide two other examples of *z*-scans below, as examples. For instance, photorefractive materials are commonly used for dynamic holography, in order to record and reconstruct volume holograms. Photorefractive materials are doped light-sensitive electro-optic materials in which optical radiation creates mobile charge carriers which redistribute themselves to create an induced space-charge field. This electrostatic field in turn modulates the refractive index which phase modulates the optical beam. A figure of merit for the effectiveness of such a material, e.g., lithium niobate, for holographic storage, is the donor to acceptor ratio in such materials. It turns out that this ratio, along with the photovoltaic coefficient, can be related to an effective n_2 in lithium niobate. This is also responsible in producing an elliptic beam in the far field by starting out from a circularly symmetric Gaussian beam focused onto the sample. The ellipticity stems from the tensor nature of the electro-optic coefficients. Typical simulated *z*-scan plots for lithium niobate are shown in Figure 7, and carries the signature variation of the scan in the variation of the beam ellipticity with scan distance. From this plot, we are able to determine the effective n_2 and hence, the donor to acceptor concentration ratio.

In other photorefractive materials, such as barium titanate, in which diffusion of charge carriers plays a dominant effect, the induced change in the refractive index is proportional to the gradient of the intensity, rather than the intensity itself, as was the case in all previous examples. In this case, as seen in Figure 8, it can be shown that the *z*-scan plot for on-axis transmittance with a small finite aperture on the observation plane is an even function of scan distance. This, however, should not be confused with the open aperture *z*-scan that is also even, but which yields information of the nonlinear absorption coefficient, as described earlier. Experiments with a new material, manganese aluminum oxide, which is thought to have both photovoltaic and diffusive contribution to space charge field distribution, yields a *z*-scan that has both odd and even components. Through careful analysis of the *z*-scan results, one is able to determine the relative contributions of each of these effects. Care should be taken in the interpretation of the results in this case, since in an actual experiment, the aperture size cannot be infinitely small, and thus, an optimum aperture size must be determined that allows for sufficient detected power,

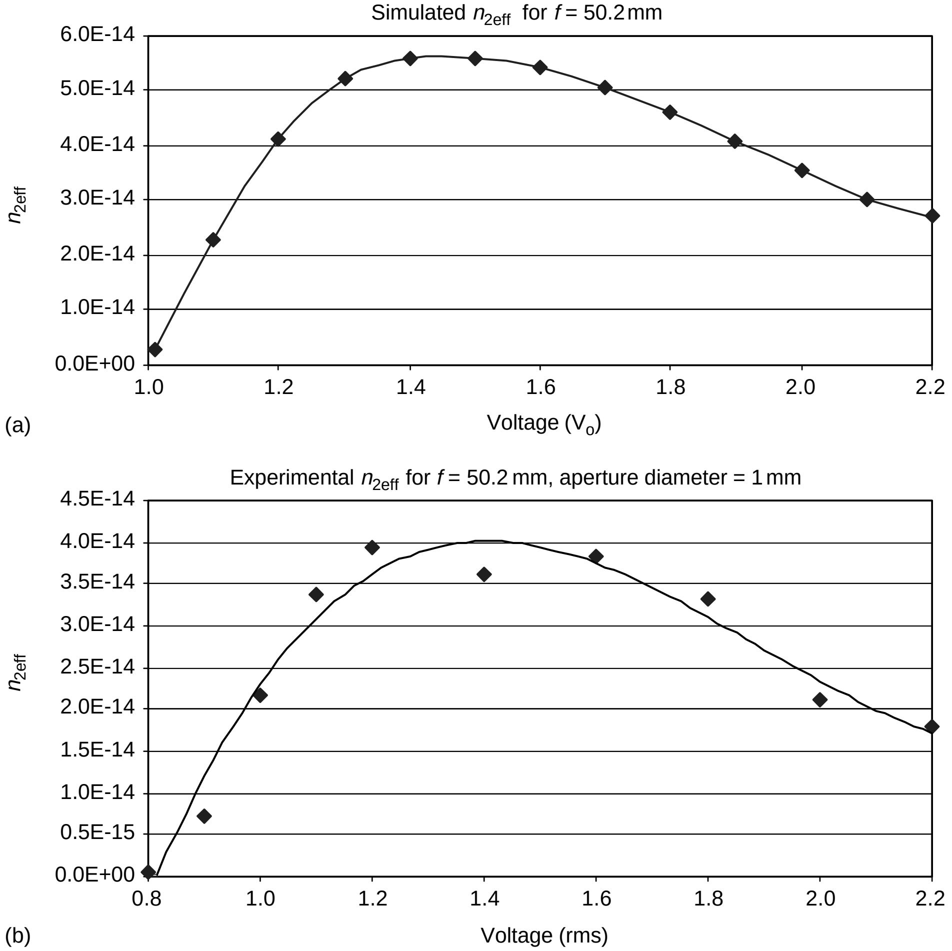

Figure 6 (a) Calculated $n_{2\mathrm{eff}}$ for $f = 50.2$ mm with $E_{\mathrm{op}} = 2 \times 10^5$ V/m; (b) Experimental $n_{2\mathrm{eff}}$ for $f = 50.2$ mm with aperture diameter = 1 mm. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

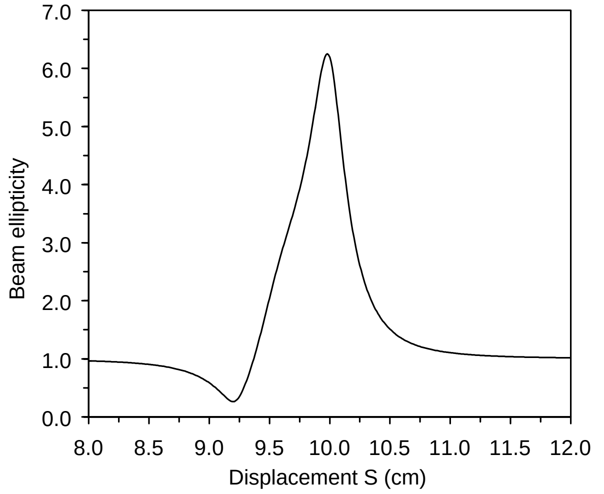

Figure 7 Typical simulated *z*-scan plot showing variation of beam ellipticity as a function of scan length. Reproduced with permission from Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications.* New York: Marcel Dekker.

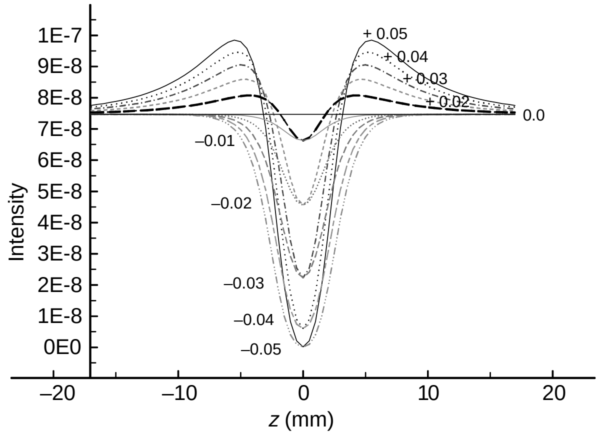

Figure 8 Typical *z*-scan plots of a diffusion dominated photorefractive material, with $\Delta n = C\nabla I$. Different values of *C* are written on the graph.

but at the same time allows one to neglect any effects of nonlinear absorption.

In summary, it is hoped that this article illuminates readers on the history of the *z*-scan, its application to thin and thick samples, its effectiveness in measurement and characterization of effective third-order optical nonlinearities for a wide class of nonlinear materials, such as $\chi^{(3)}$ materials, materials with cascaded second-order nonlinearities, liquid crystals, photorefractives, etc., and its potential for automation, while at the same time pointing out areas where care should be taken in the interpretation of the measured data. With appropriate modification in the theory, it can also be applied to characterization of linear lenses, possibly voltage controllable lenses and lenslet arrays, suitable for dynamic imaging in biomedical applications.

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Four-Wave Mixing; $\chi^{(3)}$–Third-Harmonic Generation.

Further Reading

Banerjee PP (2004) *Nonlinear Optics: Theory, Numerical Modeling, and Applications*. New York: Marcel Dekker.

Banerjee PP, Danilieko A, Hudson T and McMillen D (1998) P-scan analysis of induced inhomogeneous optical linearities. *Journal of the Optical Society of America* 15: 2446.

Banerjee PP, Misra RM and Maghraoui M (1991) Theoretical and experimental studies of propagation of beams through a finite sample of a cubically nonlinear material. *Journal of the Optical Society of America* 8: 1072.

Boyd R (2002) *Nonlinear Optics*, 2nd edn. New York: Academic Press.

Butcher P and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chapple PB, Staromlynska J and McDuff RG (1994) *Z*-scan studies in the thin- and the thick-sample limits. *Journal of the Optical Society of America B* 11: 975.

Noginov M, Helzer SW, Loutts GB, Banerjee PP, Morrisey M and Kim Y (2003) Study of photorefraction response and diffraction efficiency in Mn:YalO_3 crystals. *Journal of the Optical Society of America B* 20: 1233.

Sheik-Bahae M, Said AA and Van Stryland EW (1989) High-sensitivity, single beam n_2 measurements. *Optics Letters* 14: 955.

Sheik-Bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1990) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yeh P (1993) *Introduction to Photorefractive Nonlinear Optics*. New York: Wiley.

MATERIALS FOR NONLINEAR OPTICS

Contents

Liquid Crystals for NLO
Organic Nonlinear Materials

Liquid Crystals for NLO

I C Khoo, Pennsylvania State University, University Park, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Liquid crystals are complex organic molecular systems that exhibit various mesophases characterized by the degree of order. In this review, we shall focus mainly on thermotropic liquid crystals which exhibit various mesophases in the temperature range between the crystalline and the isotropic liquid states. In particular, we shall discuss the extraordinarily large optical nonlinearities of nematic liquid crystals, as well as the electronic nonlinearities of isotropic-phase liquid crystals. Recent studies have shown that these materials are promising candidates for advanced photonic applications.

Nematic liquid crystals (NLC) possess directional ordering, i.e., when placed in thin cells with surface

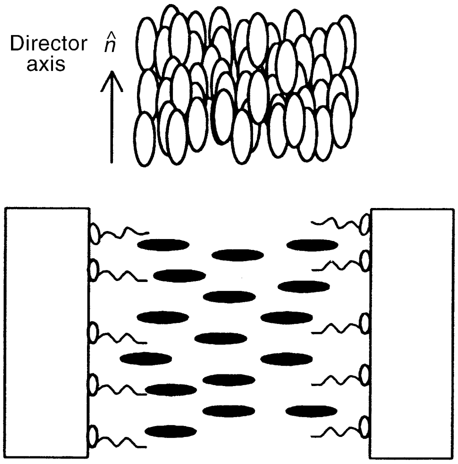

Figure 1 Nematic liquid crystals (NLC) which exhibit directional ordering characterized by the director axis. The lower figure shows how the NLC molecules are aligned in a typical cell with surfactant coated windows.

treated boundaries, the NLC molecules will self-assemble into a birefringent crystal characterized by a so-called director axis $\hat{n}$, cf. Figure 1. For light polarized parallel to $\hat{n}$, the refractive index is n_e, while for light polarized perpendicular to $\hat{n}$, the refractive index is n_0. Typically n_e is ~1.7 and n_0 is ~1.5. This large birefringence $\Delta n = n_e - n_0$ of NLC holds throughout the near UV to the far infrared spectral regime, cf. Figure 2. Perhaps the most important characteristics of NLCs is the easy susceptibility of the birefringent director (crystalline) axis orientation to externally applied fields, which has led to their ever increasing widespread use in various opto-electronics information and image display and processing devices. In analogy to dc and ac field effects, the director can easily be reoriented by an optical field, and thus cause a large optically induced index of refraction change. By doping the NLC with photosensitive or photo-charge producing agents, this process of refractive index change can be enhanced considerably. The refractive indices of NLC are also very strongly temperature dependent.

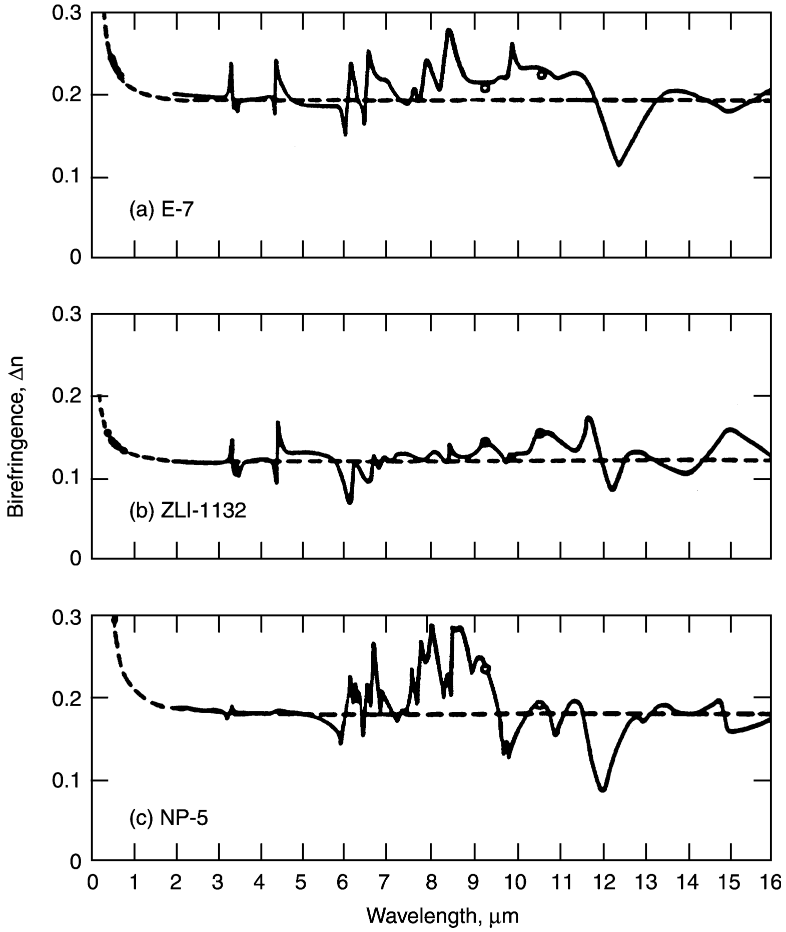

Figure 2 Broadband birefringence of nematic liquid crystals from the near ultraviolet to far infrared.

Accordingly, a large refractive index change can also be generated with laser absorption induced density and order parameter changes. In the following sections, we will broadly classify these nonlinear optical processes originating from the collective response of NLC to external fields as *nonelectronic nonlinearities*.

On the other hand, individual NLC molecules also respond to an optical field through the electronic dipolar interaction. A typical electronic energy level structure of liquid crystal molecules is shown in Figure 3. The dipolar interaction gives rise to the usual nonlinear polarization characterized by the second, third and higher-order nonlinear susceptibilities $\chi^{(2)}, \chi^{(3)}, \ldots$ etc. The complex imaginary parts of these linear and nonlinear susceptibilities account for the photonic absorption processes. Most liquid crystalline molecules possess small linear absorption in the visible region (400–700 nm), and begin to absorb near the UV region (<350 nm), cf. Figure 4. This absorption spectrum translates into a (usually) large nonlinear two-photon absorption spectrum in the visible. In some organic molecules, the intermediate absorption cross-section σ_i or the excited state absorption cross-section σ_e is much larger than the ground state linear absorption σ_g, giving rise to reverse saturable absorption (RSA) and excited state

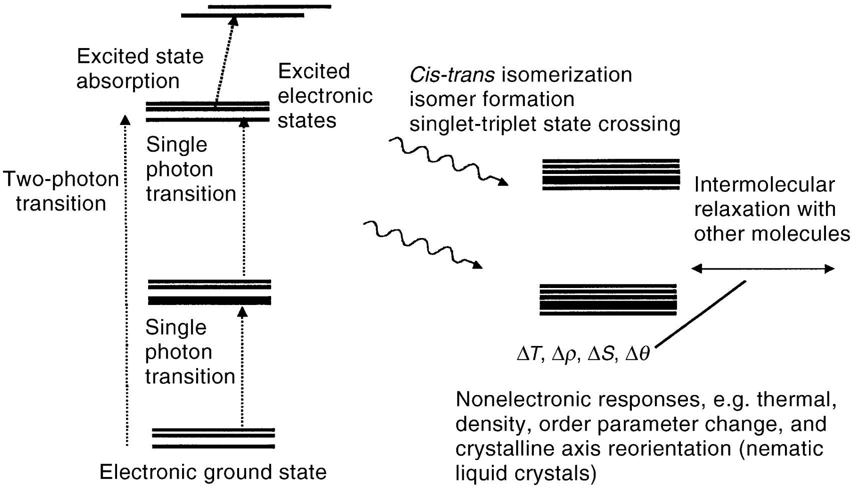
Excited state absorption
Excited electronic states
Two-photon transition
Single photon transition
Single photon transition
Electronic ground state
Cis-trans isomerization
isomer formation
singlet-triplet state crossing
Intermolecular relaxation with other molecules
ΔT, $\Delta\rho$, ΔS, $\Delta\theta$
Nonelectronic responses, e.g. thermal, density, order parameter change, and crystalline axis reorientation (nematic liquid crystals)

Figure 3 Schematic depiction of the typical energy level structure of liquid crystal molecules and various linear and nonlinear optical processes.

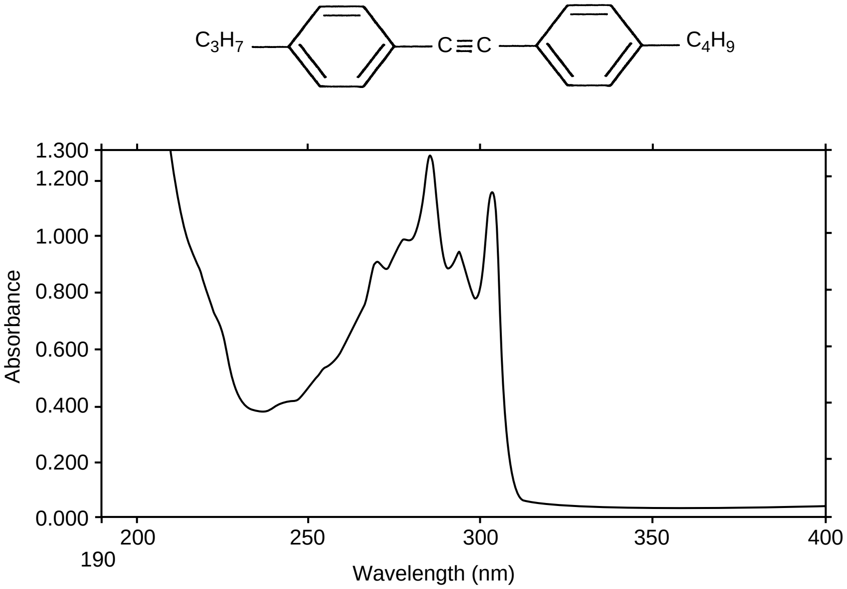

Figure 4 Molecular structure and absorption spectrum of typical isotropic liquid crystal molecules.

absorption (ESA), respectively. As we will see presently, these so-called *electronic nonlinearities* are quite useful for optical limiting applications.

Following photo-absorption, the liquid crystal molecules could undergo intersystem crossing, *cis-trans* or other conformational changes, and form isomers or charge transfer complexes. These processes will contribute to the nonelectronic nonlinearities such as reorientational, order parameter changes and thermal effects.

Orientational Nonlinearities

Purely Optically Induced Orientation

Figure 5 depicts schematically various processes that could occur in an aligned nematic film upon irradiation by an optical field. The reorientation of the director axis $\hat{n}$ is governed by the system's tendency to minimize the overall free energy, which consists of several bulk and surface interaction terms, and the interaction energy with the applied field. In purely dielectric nonabsorbing nematic liquid crystals, director axis reorientation is caused by the torque Γ_{opt} between the anisotropic nematic and the optical electric field,

$$\Gamma_{\text{opt}} = \left(\frac{\Delta\varepsilon}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^*)(\hat{n} \times \mathbf{E}) \qquad [1]$$

which competes with the elastic restoring torque Γ_{el} of the surrounding NLC molecules:

$$\Gamma_{\text{el}} = \frac{(K_1\sin^2\theta + K_3\cos^2\theta)\mathrm{d}^2\theta}{\mathrm{d}z^2} + \left[(K_1 - K_3)\sin\theta\cos\theta\right]\left(\frac{\mathrm{d}\theta}{\mathrm{d}z}\right)^2 \qquad [2]$$

where the K's are the elastic constants, and $\Delta\varepsilon$ is the (optical) dielectric anisotropy.

As a result of the director axis reorientation, an optical wave propagating through the medium will 'see' a refractive index change. A frequently employed method for measuring the refractive index change in nematic liquid crystal films is the grating diffraction setup, as shown in Figure 6a. Two polarized coherent writing beams are overlapped on a nematic film, which is probed by a polarized reading beam. If the writing beams are copolarized, they impart an optical intensity grating onto the nematic film $I = I_0(1 + \cos qx)$, where q is the grating wavevector, cf. Figure 6b. On the other hand, if the writing beams are orthogonally polarized, the resulting intensity is uniform over the illuminated area, and instead a polarization grating is created, cf. Figure 6c. The polarization state of the light in the overlapped region varies from circular, through elliptical to linear in accordance to the phase difference between the two coherent writing beams. Using a linearly polarized probe beam, this setup allows one to distinguish various intensity and reorientation dependent effects, and assess their contribution to the induced refractive index change Δn.

The magnitude of the induced reorientation angle, and therefore the extraordinary index change, depends on whether the interaction is transient or steady state. In the steady state case, to the first order of approximation, Δn is proportional to the optical intensity I, i.e., $\Delta n = n_2^{\text{SS}} I$. Under some experimental conditions, one can also show that the transient nonlinear index coefficient n_2 (τ_{p}) is given by n_2 $(\tau_{\text{p}}) \sim n_2^{\text{SS}}\tau_{\text{p}}/\tau_{\text{r}}$, where $\tau_{\text{r}} = \gamma d^2/K\pi^2$ is the orientational relaxation time constant. For $d = 100$ μm, $\gamma = 0.1$ poise and $K = 10^{-6}$ dynes, we have $\tau_{\text{r}} \sim 1$ s. Typical observed values of $n_2^{\text{SS}}(\theta)$ are

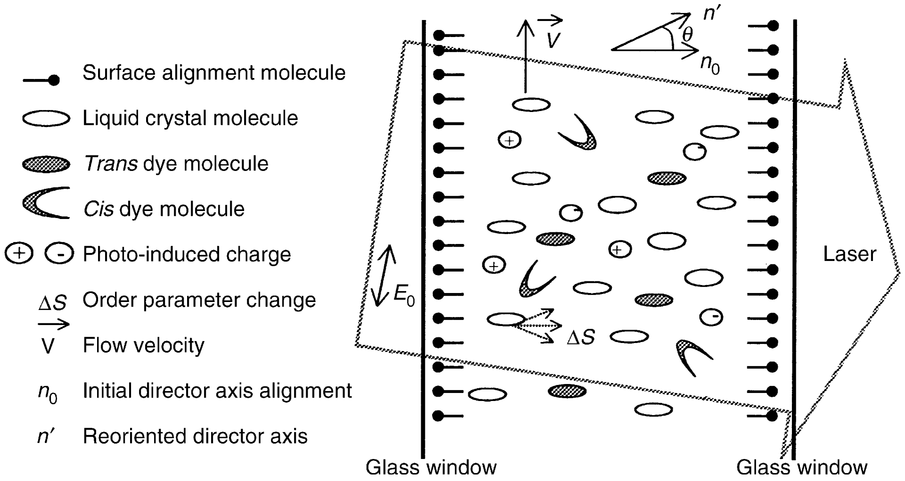

Figure 5 Schematic depiction of various laser-induced optical processes in a nematic liquid crystal film.

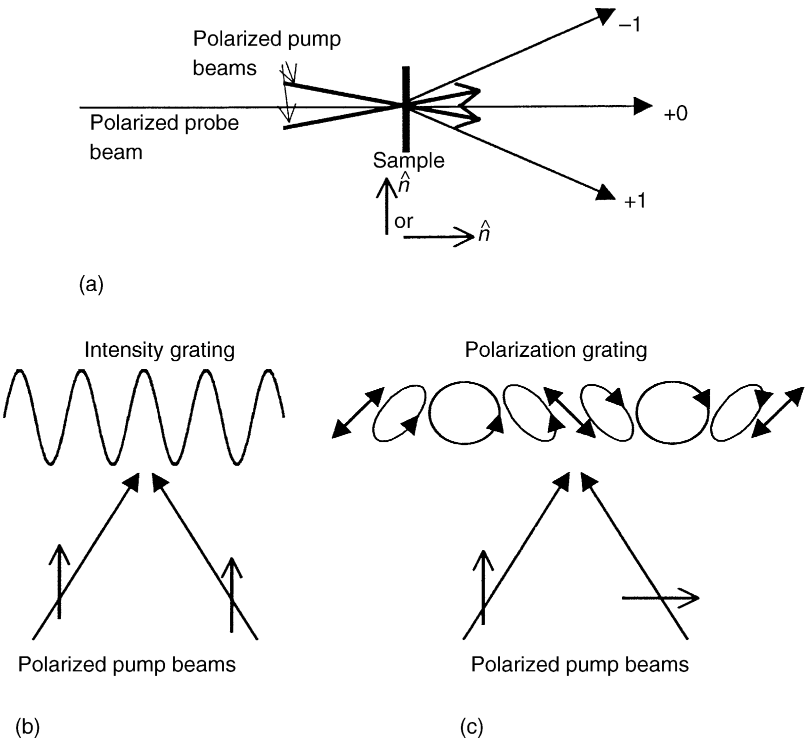

Figure 6 (a) Schematic depiction of grating formation and probe diffraction from the index grating induced in a nematic film. (b) Optical intensity grating formed by interfering two coherent copolarized writing beams. (c) Optical polarization grating formed by interfering two coherent crossed-polarized writing beams.

in the range of 10^{-5}–10^{-4} cm^2 W^{-1}. In the transient case, if $\tau_p = 1$ ns, then n_2 $(\tau_p) \sim 10^{-12}$ cm W^{-1} [$\chi^{(3)} \sim 2 \times 10^{-11}$ esu].

Excited Dye-Dopant Induced Intermolecular Torque

Studies of nematic liquid crystals doped with anthraquinone or dichroic dyes have shown that the excited dye molecules could exert an intermolecular torque Γ_{mole} that could be much larger in magnitude than Γ_{op}. By considering the interaction energies between ground and excited dyes with the surrounding LC molecules, one can show that Γ_{mole} may be written as:

$$\Gamma_{\text{mole}} = \zeta \Gamma_{\text{op}} \qquad [3]$$

where the proportionality constant ζ can be much larger than unity. The nonlinear index coefficients n_2 obtained in these films are on the order of 10^{-3} cm^2 W^{-1}. The response times for the director axis reorientation and relaxation are similar to the optically induced reorientation process discussed in the preceding section.

Photorefractive Effect

Another mechanism, namely, orientational photorefractivity, could also result in a very large optical nonlinearity. This effect occurs in nematic liquid crystals doped with photo-charge producing agents such as fullerene C_{60} or rhodamine 6G dye, in conjunction with a small applied dc field. For a 25 μm thick film with a dc bias voltage of ~1 V, the observed nonlinear coefficient n_2 is in the range of 10^{-3}–10^{-2} cm^2 W^{-1}. Figure 7 illustrates the orientational photorefractive effect. An incident optical field excites the doped liquid crystals to form charge transfer complexes (CTCs). The CTCs subsequently dissociate and create dc space charge fields through ionic diffusion, migration and other electrodynamical processes. The space charge fields, in combination with the applied dc field, create a dc field induced torque of the form:

$$\Gamma_{\text{dc}} = \left(\frac{\Delta\varepsilon_{\text{dc}}}{8\pi}\right)(\hat{n}\cdot\mathbf{E}^{*}_{\text{int}})(\hat{n} \times \mathbf{E}_{\text{int}}) \qquad [4]$$

Here the total internal dc field $\mathbf{E}_{\text{int}}$ contains the applied dc field $\mathbf{E}_{\text{dc}}$ and the generated dc space charge fields $\mathbf{E}_{\text{sc}}$. This causes director axis reorientation and refractive index change.

The space charge fields consist of one photorefractive and two anisotropy components. For an incident optical intensity grating function of the form $I_{\text{op}} = I_0(1 + m \cos q\xi)$, the photorefractive space charge

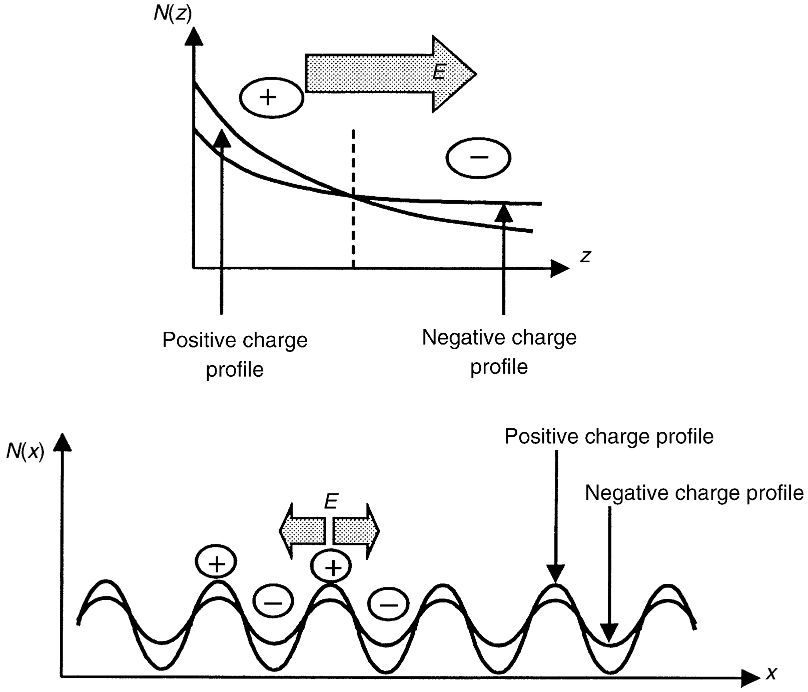

Figure 7 Schematic depiction of photo-induced space charge field formation in a nematic liquid crystal illuminated by a monotonic (upper curve) or a sinusoidal (lower curve) laser intensity function.

field is given by:

$$E_{\mathrm{ph}} = q\nu\, m k_{\mathrm{B}} T\left(\frac{\sigma - \sigma_{\mathrm{d}}}{2e\sigma}\right)\cos\left(\frac{q\xi - \pi}{2}\right) \qquad [5a]$$

where k_{B} is the Boltzmann constant, σ is the illuminated conductivity, σ_{d} is the dark state conductivity, $\nu = (D_+ - D_-)/(D_+ + D_-)$, where D_+ and D_- are the diffusion constants for the positively and negatively charged ions, respectively; m is the optical intensity modulation factor, $q = 2\pi/\Lambda$ is the magnitude of the grating wavevector, and ξ is the coordinate along q.

Two other sources of space charge fields are also at work, and they arise from the conductivity and dielectric anisotropies of the nematic under the action of the dc bias field E_{dc}:

$$E_{\Delta\sigma} = E_{\mathrm{dc}}\left(\frac{(\sigma_{//} - \sigma_{\perp})\sin\theta\cos\theta}{\sigma_{//}\sin 2\theta + \sigma_{\perp}\cos 2\theta}\right) \qquad [5b]$$

$$E_{\Delta\varepsilon} = E_{\mathrm{dc}}\left(\frac{(\varepsilon_{//} - \varepsilon_{\perp})\sin\theta\cos\theta}{\varepsilon_{//}\sin 2\theta + \varepsilon_{\perp}\cos 2\theta}\right) \qquad [5c]$$

where $(\sigma_{//} - \sigma_{\perp})$ is the low-frequency **conductivity anisotropy** and $(\varepsilon_{//} - \varepsilon_{\perp})$ the **dielectric anisotropy** of the liquid crystal, and θ is the reorientation angle of the director axis.

In addition to these space charge field induced effects, the **flows** of the charged ions and complexes under the applied dc field and the generated space charge fields also contribute significantly to reorienting the director axis. For an interaction geometry similar to Figure 5, the flow creates a shear torque Γ_{shear} of magnitude proportional to the velocity gradient

$$\Gamma_{\mathrm{shear}} = \frac{\mathrm{d}v(\zeta)}{\mathrm{d}\zeta}(\alpha_2\sin^2\theta - \alpha_3\cos^2\theta) \qquad [6]$$

where the α's are the viscosity coefficients, and $\mathbf{v}(\zeta)$ is the flow velocity. The combined action of the above space charge fields and the flow effects on director axis reorientation have been quantitatively formulated and shown to describe experimental results very well, including, in particular, the observation that maximal diffraction is obtained for a grating constant $\lambda_{\mathrm{g}} \sim 2d$. Furthermore, the flow-reorientation and adsorption of the excited dopant molecules on the surfaces are believed to be responsible for quasi-permanent and permanent director axis reorientation in rhodamine 6G- and C_{60}-doped nematic films, respectively.

Supra-Nonlinear Methyl-Red Doped NLC – Observed Phenomena

Recent studies in methyl-red doped nematic liquid crystals (MRNLCs) have ushered in the era of supra-nonlinearity. Unlike photorefractive nematic liquid crystals, no external bias field is needed. The observed n_2 values could be as large as $10\,\mathrm{cm^2\,W^{-1}}$. In MRNLC, a large photovoltage drop across the sample is also observed upon irradiation. The polarity of the photovoltage is reversed if the light impinges from the opposite cell window. The extremely nonlinear optical response of MRNLC is attributed to nematic director axis reorientation by the optically induced space charge fields, similar to the effect discussed in the previous section. The unusually large response is possibly due to several factors, including the higher photo-charge producing abilities of methyl red, and the large difference in the positive and negative photo-ion diffusion constants. Another cause of director axis reorientation is the intermolecular torque exerted by the photo-excited dye molecules on the NLC. This is particularly evident in experiments using an input polarization grating on planar aligned samples. Because the laser excited dye molecules are adsorbed and strongly 'anchored' on the cell walls, they cause permanent realignment of the nematic liquid crystal. This process usually occurs upon prolonged illumination of the liquid crystal cells with a low-power laser, and can be sped up with higher laser power.

Trans–Cis Isomerization and Order Parameter Change

Some nematic liquid crystals are azo-compounds, i.e., their molecular structures are analogous to the typical one depicted in Figure 8. An azo-compound when irradiated at its absorption wavelength will change from the *trans* to the *cis* form, cf. Figure 8. *Trans–cis* isomerization of these azo-nematic liquid crystals or azo-dopants in NLC's could also give rise to a refractive index change through modification of the NLC's order parameter and/or director axis reorientation. The *trans* species are oblong in shape, and align parallel to the director axis. On the other hand, the 'bent' structure of the *cis* species is incongruous to the nematic order, and thus causes an order parameter change, and therefore a refractive index modification. Such refractive index changes owing to order parameter modification are similar to thermal effects discussed in the next section. The response [on-] time for these *trans–cis* isomerization induced order-parameter changes can be quite fast (as fast as nanoseconds in some cases). The relaxation time (typically in the millisecond regime) is decided by the director axis relaxation rate as discussed earlier.

The other index-changing mechanism is the tendency of the *trans* molecule to align with its transition moment perpendicular to the optical polarization in order to minimize the interaction (absorption) energies. This in turn causes the NLC director axis to realign through the intermolecular torque effect. Both mechanisms are characterized by a nonlinear index coefficient n_2 on the order of $10^{-3}\,\mathrm{cm^2\,W^{-1}}$.

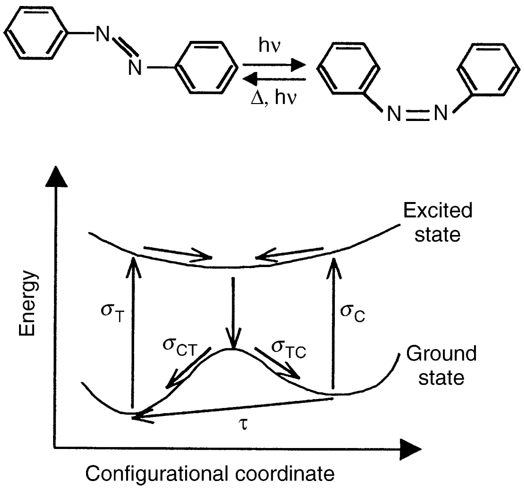

Figure 8 *Trans–cis* isomerization of azo-molecules upon optical illumination. The upper diagram depicts molecular structural changes; the lower diagram shows the excitation energy versus the molecular coordinate associated with these photo-excited processes.

Thermal and Density Effect

Laser induced temperature and density changes are described by time-dependent coupled hydrodynamics equations. In grating diffraction experiments, cf. Figure 6, the problem can be simplified to a one-dimensional one, i.e., along the grating wavevector direction, if the grating spacing is smaller than the nematic film thickness and the beam size. In that case, the thermal and Brillouin (sound wave) decay time constants are given by, respectively:

$$\tau_T = \frac{\rho_0 C_v}{\lambda_T q^2} \qquad [7a]$$

$$\tau_B = \frac{2\rho_0}{\eta q^2} \qquad [7b]$$

where $q = 2\pi/\lambda_q$ is the grating wavevector, ρ_0 the unperturbed density, C_v the specific heat, λ_T the thermal conductivity, and η the viscosity. The values of these parameters depend on the director axis alignment relative to the optical fields, thermal diffusion and sound propagation directions. For typical values of $\eta = 7 \times 10^{-2}\,\mathrm{kg\,m^{-1}\,s^{-1}}$, $\rho_0 = 10^3\,\mathrm{kg\,m^{-3}}$, $D = \lambda_T/\rho_0$, $C_v \sim 1.6 \times 10^{-7}\,\mathrm{m^2\,s^{-1}}$ and a grating spacing $\lambda_q = 20\,\mu\mathrm{m}$, we have $\tau_T \sim 50\,\mu\mathrm{s}$, whereas $\tau_B \sim 200\,\mathrm{ns}$. Such a vast difference between the relaxation times allows one to distinguish the contribution from thermal and density origins.

Table 1 Nonelectronic nonlinearities of nematic liquid crystals

Nonlinear processes	*Typical magnitude*	*Time-scale*	*Liquid crystal*
Orientational nonlinearities			
Purely optically induced n_2	10^{-4} cm^2 W^{-1}	ms	pure 5CB
Excited dopant assisted n_2	10^{-3} cm^2 W^{-1}	ms	dyed 5CB
Photorefractive	10^{-3} cm^2 W^{-1}	ms	5CB
Photorefractive + Methyl-red doped 5CB	10 cm^2 W^{-1}	ms	
Trans–Cis and order parameter modification	10^{-4} cm^2 W^{-1}	ms	dyed 5CB
Thermal nonlinearities			
n_2	10^{-6} cm^2 W^{-1}	ms	E7
n_2	10^{-11} cm^2 W^{-1}	ns	E7
Density effect			
electrostrictive n_2	10^{-11} cm^2 W^{-1}	ns	E7

In the steady state, i.e., when the laser–nematic interaction time is much longer than τ_T and τ_B, the nonlinear index coefficient is given by:

$$n_2^{SS}(\Delta T) = \left(\frac{dn}{dT}\right)\frac{\alpha\tau_r}{\rho_0 C_V} \qquad [8]$$

where dn/dT is the appropriate thermal index coefficient. The absorption constant α of liquid crystals varies widely in the visible–infrared spectrum. Near the phase transition temperature, $dn_{//,\perp}/dT$ also changes dramatically. The magnitude of $n_2^{SS}(\Delta T)$ could therefore vary by several orders of magnitude. For $\alpha = 100\ \text{cm}^{-1}$, $dn_{//,\perp}/dT \sim 10^{-3}\ \text{K}^{-1}$, and $\lambda_q = 20\ \mu\text{m}$, we have $n_2^{SS}(\Delta T) \sim 10^{-6}\ \text{cm}^2\ \text{W}^{-1}$. This is the typical value observed in wave mixing experiments with infrared [CO_2] or visible lasers.

In the transient regime ($\tau_p \ll \tau_r, \tau_B$), the nonlinear index coefficient is proportional to the laser pulse duration:

$$n_2(\tau_p) = n_2^{SS}(\Delta T)\tau_p/\tau_r \qquad [9]$$

With $\tau_r \sim 0.5 \times 10^{-4}$ s and $\tau_p = 1$ ns, $n_2(\tau_p) \sim 5 \times 10^{-11}\ \text{cm}^2\ \text{W}^{-1}$. Such nonlinearities have been observed in transient dynamic grating diffraction studies involving nanosecond laser pulses. In these studies, it is found that the contribution from the electrostrictive effect (movement of the nematic due to the optical intensity gradient) to the refractive index change can be as large as the absorptive thermal component.

In Table 1, we give a concise summary of these nonelectronic optical nonlinearities observed in nematic liquid crystals.

Electronic Nonlinearities

Second and Third Order Nonlinear Susceptibilities

Electronic interaction involves the perturbation of the electronic wavefunctions by the impinging optical fields. The induced dipole moment of a molecule is given by:

$$\mathbf{d} = \boldsymbol{\alpha}:\mathbf{E} + \boldsymbol{\beta}:\mathbf{EE} + \boldsymbol{\gamma}:\mathbf{EEE} + \cdots \qquad [10]$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are the linear, second-order and third-order nonlinear polarizability tensors, **E** is the optical electric field and : denotes tensorial operation. The macroscopic polarization **P** (dipole moment per unit volume) is

$$\mathbf{P} = \varepsilon_0\boldsymbol{\chi}^{(1)}:\mathbf{E} + \boldsymbol{\chi}^{(2)}:\mathbf{EE} + \boldsymbol{\chi}^{(3)}:\mathbf{EEE} + \cdots \qquad [11]$$

where $\boldsymbol{\chi}^{(1)}$, $\boldsymbol{\chi}^{(2)}$, and $\boldsymbol{\chi}^{(3)}$ are the linear, second and third-order nonlinear susceptibilities, respectively. The macroscopic parameters $\boldsymbol{\chi}$ are related to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by the local field correction factors.

A well-known process associated with the second-order nonlinear polarization is second harmonic generation, in which two incident photons of frequency ω are converted into a photon of frequency 2ω. Symmetry rules require that a material should be non-centrosymmetric in order to exhibit second-order nonlinear polarization. In centrosymmetric nematic liquid crystals, an applied dc field or discontinuity at a surface or interface, or director orientational curvature, could provide the required symmetry-breaking mechanism.

In ferroelectric or smectic C* liquid crystals, the molecules possess permanent polarization, i.e., the centrosymmetry is broken, and it is possible to observe second harmonic generation associated with the second-order nonlinear polarization. Typically, the observed nonlinear coefficients β_{2HG} are on the order of $\sim 10^{-14}\ \text{mV}^{-1}$. In contrast to inorganic crystals of much larger dimensions, these thin liquid crystal films are not useful for practical devices. Nonetheless, because of the sensitivity of second harmonic generation to surface and interface conditions, second harmonic generation has been proven to be an effective spectroscopic tool.

In dc field induced second harmonic generation, the nonlinear polarization involved is actually the

Table 2 Electronic optical nonlinearities of nematic liquid crystals

Nonlinear optical phenomena and nonlinearities	*Magnitude*	*Time-scale (pulse duration of laser used)*
Second harmonic generation (dc field induced),		
$\chi_{//}^{(3)}$	2.7×10^{-12} esu	ns
$\chi_{\perp}^{(3)}$	0.55×10^{-12} esu	ns
Second harmonic generation β_{2HG}	10^{-12} mV^{-1}	ps (ferroelectric)
Two-photon absorption, β	5 cm GW^{-1}	80 ps
Effective nonlinear absorption, β_{eff}	200 cm GW^{-1}	10 ns
Degenerate four-wave mixing		
electronic $\chi^{(3)}(\omega; \omega, -\omega, \omega)$	10^{-12} esu	ns

third-order nonlinear polarization given by $P(2\omega) = \boldsymbol{\chi}^{(3)}E(0)E(\omega)E(\omega)$. Typically measured values of $\boldsymbol{\chi}^{(3)}$ are $\sim 10^{-12}$ esu, expressed in the commonly used cgs units. The conversion factor between cgs and SI unit is given by

$$\boldsymbol{\chi}^{(3)}[\text{esu}] = 3^4 \times 10^{17} \boldsymbol{\chi}^{(3)}[\text{SI unit}] \qquad [12]$$

There is relatively little work done on measuring or studying the third-order nonlinear electronic polarizabilities of nematic liquid crystals. Most of the studies are centered on the molecular correlation effects associated with the nematic–isotropic phase transition. The typical magnitude of electronic $\boldsymbol{\chi}^{(3)}$ is 10^{-12} esu.

Nonlinear Absorption and Limiting

Two-photon absorption occurs where the molecule is excited to a real electronic excited state by simultaneous absorption of two photons from the incident light. Since the absorption is proportional to the square of the optical intensity, the incident light will experience an increasing absorption effect, and its transmission at high intensity will be 'quenched' to an almost constant level. Two-photon absorption is usually characterized by the two-photon absorption coefficient β, which is related to the imaginary part of $\boldsymbol{\chi}^{(3)}$. For liquid crystals, they are on the order of 5×10^{-12}–10^{-11} mW^{-1} (equivalent to 0.5–5 cm GW^{-1} in the more commonly used units).

Liquid crystalline materials in their (less scattering) isotropic phase have been extensively studied in the last few years for application in optical limiting devices as a result of their large nonlinear absorption constant. In the nanosecond time regime, the RSA and ESA effects, cf. Figure 3, act to give rise to a very large effective nonlinear absorption coefficient β on the order of several hundreds of cm GW^{-1}. In combination with special device structure such as a nonlinear fiber array, it has become possible to clamp the transmitted nanosecond laser pulse energy to below the retina damage level (below 0.5 μJ).

In Table 2, we give a concise summary of these electronic optical nonlinearities of nematic liquid crystals.

Concluding Remarks

We have presented a concise review of the nonlinear optics of liquid crystals in their nematic and isotropic phases. Nematic liquid crystals in their various pure and doped forms possess many interesting nonlinear optical responses, with nonlinearities ranging over 16 orders of magnitude. Some of the extraordinarily large nonlinearities enable the performance of several all-optical switching and limiting, image modulation and sensing processes at unprecedented low threshold power. In particular, self-defocusing and optical limiting effects with nanowatt-power lasers, and incoherent–coherent image conversion at μW cm^{-2} optical intensity have been demonstrated. Liquid crystal films have been used in selective attenuation of glares and blinding laser light. Stimulated scattering that used to be the domain of high-power lasers can now be performed in thin nematic films with mW-power cw laser. Readers can consult the Further Reading section for some of these advanced photonic devices and applications based on nonlinear optics of liquid crystals. Current active studies by several research groups worldwide are likely to result in even more interesting nonlinear optical effects and useful applications in the near future.

List of Units and Nomenclature

Absorption coefficient	α [cm^{-1}]
Density	ρ_0 [kg m^{-3}]
Eleastic constant	K [dynes]
Intensity (optical)	I [W cm^{-2}]
Nonlinear susceptibility	$\chi^{(3)}$ [esu]
Second harmonic generation coefficient	β_{2HG} [mV^{-1}]
Specific heat	C_V [m^2 s^{-1}]
The nonlinear index coefficients	n_2 [cm^2 W^{-1}]
Time constants	τ [s]

Two-photon absorption constant	β [cm GW^{-1}]
Viscosity	η [kg m^{-1} s^{-1}]
Viscosity coefficient	γ [poise]

See also

Nonlinear Optics, Basics: Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction.

Further Reading

Khoo IC (1995) *Liquid Crystals: Physical Properties and Nonlinear Optical Phenomena*. New York: Wiley Interscience.

Khoo IC (1996) Orientational photorefractive effects in nematic liquid crystal films. *IEEE Journal of Quantum Electronics* 32: 525–534.

Khoo IC and Liang Y (2000) Stimulated orientational and thermal scatterings and self-starting optical phase conjugation with nematic liquid crystals. *Physical Review* E62: 6722–6733.

Khoo IC and Wu ST (1993) *Optics and Nonlinear Optics of Liquid Crystals*. Singapore: World Scientific.

Khoo IC, Chen PH, Wood MV and Shih MY (1999) Molecular photonics of a highly nonlinear organic fiber core liquid for picosecond–nanosecond optical limiting effect. *Chemical Physics* 245: 517–531.

Khoo IC, Wood MV, Shih MY and Chen PH (1999) Extremely nonlinear photosensitive liquid crystals for image sensing and sensor protection. *Optics Express* 4(11): 431–442.

Khoo IC, Shih MY, Wood MV, *et al.* (1999) Dye-doped photorefractive liquid crystals for dynamic and storage holographic grating formation and spatial light modulation. *IEEE Proceedings* 87: 1897–1911.

Marrucci L and Papara D (1997) Photoinduced molecular reorientation of absorbing liquid crystals. *Physical Review* E56: 1765–1772.

Rau H (1990) Photoisomerization of Azobenzenes. In: Rabeck FJ (ed.) *Photochemistry and Photophysics*, vol. 2, chap. 4, pp. 119–141. Boca Raton, FL: CRC Press.

Shen YR (1990) *Principles of Nonlinear Optics*. New York: Wiley Interscience.

Organic Nonlinear Materials

F Kajzar and I Rau[*], Commissariat à l'Energie Atomique, Gif sur Yvette, France

© 2005, Elsevier Ltd. All Rights Reserved.

Nonlinear Medium Polarization

Macroscopic

Photons are neutral, noninteracting bosons. Interaction is possible only through a material medium. Indeed, photons can be absorbed by electrons leading to excited states. A reverse transition may go through radiative channels, with emission of photons, usually with a different energy than the excitation energy. Interaction of photons with matter may also lead to other types of excitation, such as vibronic, rotational, libronic, etc., or just to a change in the medium polarization. This change may be expanded in the power series of the electric forcing field strength with *i*th component given by:

$$\begin{aligned} P_I(\omega_\sigma) = P_{0I}(\omega_\sigma) + \varepsilon_0[K_1\chi^{(1)}_{IJ}(-\omega_\sigma;\omega_\sigma)E_J^{\omega_\sigma} \\ + K_2\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2)E_J^{\omega_1}E_K^{\omega_2} \\ + K_3\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E_J^{\omega_1}E_K^{\omega_2}E_{\mathrm{L}}^{\omega_3} + \cdots] \end{aligned} \quad [1]$$

where the expansion coefficients $\chi^{(n)}$ are $(n+1)$ rank three-dimensional tensors, describing linear $(n=1)$ and nonlinear $(n>1)$ responses of the medium, $P_{0i}(\omega_\sigma)$ is its permanent polarization (at frequency ω_σ) and the coefficients K_i $(i = 1, 2, 3, \ldots)$ take account of the conventions for the Fourier transform of electric field and polarization vectors used, as well as the degeneracy of nonlinear optical processes. For historical reasons two conventions in quantitative description of NLO processes are used:

$$\mathrm{E}(\mathrm{r},t) = \tfrac{1}{2}[\mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.}] \qquad \text{(Convention I)} \quad [2]$$

and

$$\mathrm{E}(\mathrm{r},t) = \mathrm{E}(\mathrm{r})\mathrm{e}^{\mathrm{i}\omega t} + \mathrm{c.c.} \qquad \text{(Convention II)} \quad [3]$$

and similarly for the polarization field. As researchers often use different conventions, it is important to bear this in mind when comparing NLO properties of materials coming from different sources. The tabulated data and the formulas used here are within convention I, with the permutation factors included explicitly in coefficients K_i (eqn [1]). In centrosymmetric media and in dipolar approximation, the static polarization P_0, as well as all even order NLO susceptibilities, are equal to zero, as it follows from time reversal symmetry.

---

[*] On leave from POLITECHNICA University of Bucharest, Romania.

Relations Between Macroscopic and Microscopic Polarizations

Microscopic Polarization

Similar expansions to eqn [1] also obey the conventions on a molecular level. Similarly, as for bulk polarization, the ith component of the molecule dipole moment can also be expanded into the series of the local field strength giving:

$$\begin{aligned}\mu_i(\omega_\sigma) &= \mu_{0i}(\omega_\sigma) + K_1\alpha_{ij}(-\omega_\sigma;\omega_\sigma)E_j^{\omega_\sigma} \\ &\quad + K_2\beta_{ijk}(-\omega_\sigma;\omega_1,\omega_2)E_j^{\omega_1}E_k^{\omega_2} \\ &\quad + K_3\gamma_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)E_j^{\omega_1}E_k^{\omega_2}E_l^{\omega_3} + \cdots \\ &= \mu_{0i}(\omega_\sigma) + \mu_{\text{ind}}(\omega_\sigma,\omega_1,\omega_2,\ldots,\omega_n,\alpha,\beta\ldots,\zeta, \\ &\quad E^{\omega_1},E^{\omega_2},\ldots,E^{\omega_n}) \end{aligned} \qquad [4]$$

where α is the linear molecular polarizability, β is the first order and γ is the second molecular hyperpolarizability. μ_{in} in eqn [4] denotes the induced dipole moment. The expansions [1] and [4] may be also considered as the Tamm–Dancoff perturbation development as functions of an increasing number of interacting photons. In this case, the second term on the right-hand sides of eqns [1] and [4] describes the interaction of two photons, the third term that of three photons, etc. The electric field in eqn [4] is the effective field experienced by the molecule and is given by

$$E = fE \qquad [5]$$

where f is the local field factor, depending on the form of molecule. In eqns [1] and [4] Einstein's notation is used.

For a single crystal and for noninteracting molecules there exists a straightforward relationship between the macroscopic and microscopic quantities given by:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma} f_n^{\omega_1} f_n^{\omega_2} N_n \\ &\quad \times \sum_{ijk} a_{Ii}a_{Jj}a_{Kk}\beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2)\end{aligned} \qquad [6]$$

for the second-order NLO susceptibility and

$$\begin{aligned}\chi^{(3)}_{IJKL}(-\omega_\sigma;\omega_1,\omega_2,\omega_3) &= \sum_n f_n^{\omega_\sigma} f_n^{\omega_1} f_n^{\omega_2} f_n^{\omega_3} N_n \\ &\quad \times \sum_{ijkl} a_{Ii}a_{Jj}a_{Kk}a_{Ll}\gamma^n_{ijkl}(-\omega_\sigma;\omega_1,\omega_2,\omega_3)\end{aligned} \qquad [7]$$

for the third-order NLO susceptibility, respectively. The coefficients a_{Ij} in eqns [6] and [7] are Wigner rotation matrix elements, N_n denoting the number density of species (n). The subscripts ($ijkl$) refer to the molecule, and ($IJKL$) to the laboratory reference frames, respectively.

For disordered (or partly ordered) materials, the macroscopic NLO susceptibility depends on the orientation average of nonzero molecular hyperpolarizability components:

$$\begin{aligned}\chi^{(2)}_{IJK}(-\omega_\sigma;\omega_1,\omega_2) &= \sum_n f_n^{\omega_\sigma} f_n^{\omega_1} f_n^{\omega_2} N_n \\ &\quad \times \sum_{ijk} < \beta^n_{ijk}(-\omega_\sigma;\omega_1,\omega_2) >_{IJK}\end{aligned} \qquad [8]$$

and similarly for $\chi^{(3)}$ susceptibility. The average in eqn [8] is taken over all NLO molecule orientations and all β tensor components. Fortunately, the number of the nonzero β tensor components (21 for β tensor) is limited by the molecular symmetry and in the case of charge transfer (CT) molecules, only the component in the charge transfer direction β_{xxx} may be retained. The others are usually significantly smaller and can be neglected with respect to this one.

Second-Order Nonlinear Optical Materials

The research on nonlinear optical materials is boosted by the potentiality of their applications in photonics, and particularly in optical signal processing, image recognition, frequency conversion, and light amplification. New fields of applications open in biology, medicine (photodynamic therapy), material processing (optical engineering), etc. Depending on the targeted application, the materials showing enhanced second-, third-, and/or higher-order nonlinear optical properties are developed and optimized. The optimization is performed on both molecular and macroscopic levels. Indeed, the bulk nonlinear optical response depends not only on the individual hyperpolarizabilities but also on the way the molecules are assembled into the bulk material. As already mentioned, only the noncentrosymmetric materials exhibit the second-order nonlinear optical response.

Molecules

The most promising molecules for the application in second-order NLO devices are intramolecular CT compounds, composed from an electron donating (D), an electron accepting (A) part, linked by a π conjugated backbone (Figure 1). The CT molecules exhibit a large first-order susceptibility β tensor component β_{xxx}, enhanced in the charge transfer direction x. The value of β_{xxx} depends on the strength of electron donating and electron accepting groups (amount of the charge transfer)

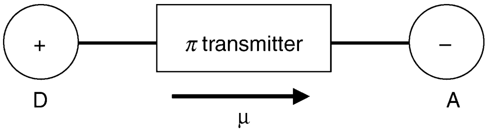

Figure 1 Schematic representation of a charge transfer molecule.

Table 1 Chemical structures of commonly used electron donating and electron accepting groups

Donors	*Acceptors*
$-N(CH_3)_2$	$-NO$
$-NH_2$	$-NO_2$
$-OCH_3$	$-CHO$
$-OH$	$-CN$
$-CH_3$	$-SO_2CH_3$
$-SCH_3$	$-SO_2CF_3$
$-N(CH_3)_2$	$=C(CN)_2$

and on the conjugation length of the transmitter. Table 1 shows the chemical structures of typical electron donating and electron accepting groups. Chemical structures of selected CT molecules are given in Table 2, together with the values of the β_{xxx} tensor component, as determined by the electric field induced second-harmonic generation technique (EFISH).

The dependence of β_{xxx} on the conjugation length L of the conjugated transmitter is given by the scaling law:

$$\beta_{xxx} \propto L^{\alpha_s} \qquad [9]$$

with $\alpha_s > 1$. Usually the conjugated backbone is composed of a number of alternate single/double (or single/triple) bonds and in this case, the scaling law is tested as a function of the number of double (or triple) bonds N:

$$\beta_{xxx} \propto N^{\alpha_\beta} \qquad [10]$$

For a number of CT oligomers with the same electron donating and electron accepting groups, differing only in the length of the conjugated transmitter, the exponent $\alpha_\beta \sim 2.5$. The β_{xxx} value saturates for a larger number of double bonds ($N > 4$, in the case of phenyl oligomers used as transmitter).

A number of studies are in progress on octupolar molecules which exhibit large β hyperpolarizabilities. However, their processing into efficient, noncentrosymmetric bulk materials is still at its initial stage and up to now no practical application has been demonstrated.

Bulk Materials

In order to obtain noncentrosymmetric materials, different approaches are used such as:

(i) single crystals growth;
(ii) epitaxy or heteroepitaxy;
(iii) Langmuir–Blodgett films;
(iv) self-assemblies;
(v) intermolecular charge transfer structures;
(vi) poled polymers; and
(vii) chiral structures.

The growth of good-quality optical single crystals is costly and difficult, especially as the charge transfer molecules usually exhibit a large ground state dipole moment (Table 2). The strong dipole–dipole interaction usually leads to their antiparallel alignment, and as a consequence, a centrosymmetric bulk structure. Only a few molecules crystallize in a noncentrosymmetric structure. One of the approaches used consists of design and synthesis of molecules with a negligible ground state dipole moment such as POM (Table 3). Therefore, a lot of effort has been put into forcing molecules to build into noncentrosymmetric bulk materials (ii)–(vi). Owing to the fact that most of applications are targeted in waveguiding configuration, a significant effort has been made to grow noncentrosymmetric thin films (ii)–(vi). While the chiral molecules form spontaneously noncentrosymmetric films, in other techniques, this is obtained more easily by application of an external electric field (poled polymers), by use of an adequate substrate (epitaxy, pseudoepitaxy, self assembly), intramolecular charge transfer in ordered layered structures (v), and interaction between molecules and molecules–substrate (Langmuir–Blodgett films). In the last case, three types of well-ordered structures, with controllable thickness are obtained: X, Y, Z (Figure 2). Only X and Z type structures give noncentrosymmetric thin films exhibiting second-order NLO properties.

Poled polymers

Up to now, the best results in getting efficient materials for second-order NLO application are obtained with poled polymers. The materials consist of a polymer matrix, exhibiting excellent optical properties, functionalized with active NLO chromophores (CT molecules). Four kinds of structures, shown schematically in Figure 3, are realized:

(i) guest–host systems;
(ii) side chain polymers;
(iii) main chain polymers; and
(iv) thermally or photo-crosslinked polymers.

Table 2 Molecular first hyperpolarizabilities and dipole moments of selected CT molecules

Molecule	*Chemical structure*	λ_{max} *[nm]*	*Solvent*	*Ground state dipole moment,* μ *[D]*	*Product* $\mu\beta$ *[10^{-77} Cm^5 V^{-1}]*	β *[10^{-40} m^4 V^{-1}]*	λ_{meas} *[nm]*	β_0 *[10^{-40} m^4 V^{-1}]*
p-NA (*para*-nitroaniline)	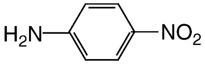	358	1,4 dioxane	5.5	1.30 0.74	71.0 40.2	1064 1907	35
MNA (2-methyl-4-nitroaniline)	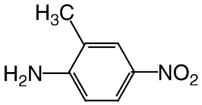	361	1,4 dioxane	7.4	1.72 0.98	69.9 39.8	1064 1907	33
NPP (*N*-(4-nitrophenyl)-(L)-prolinol)	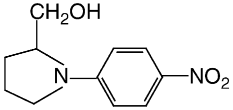	397	acetone	7.3	4.28	176	1064	67
POM (3-methyl-4-nitropyridine-1-oxide)	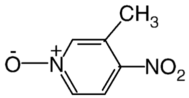	320	from crystal data	0	0	36	1064	21
PNP (1-(*N*-prolinol)-5-nitropyridine)	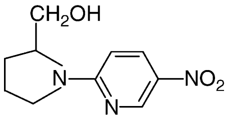	376	1,4 dioxane	7.2	2.28 1.41	95 59	1064 1907	38
MMONS (3-Methyl-4-methoxy-4′ nitrostilbene)	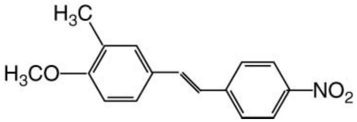	366	1,4 dioxane	5.2	2.14	123.9	1907	90
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)]amino-4′-nitroazobenzene (dispersed red 1))	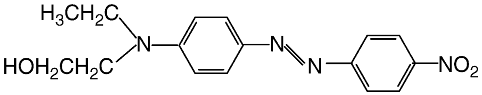	509 455	DMSO 1,4 dioxane			525 234.1	1356 1907	197 150
DADAB (4-(*N*,*N*-diethylamino)-4′-tricyanovinylazo benzene)	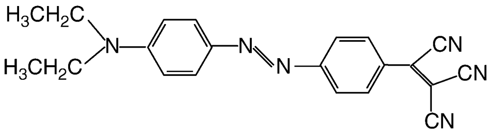	582	DMSO			1640	1580	648

1D = (1/3) × 10^{-39} Cm.

Table 3 Bulk second-order nonlinear optical properties of selected organic materials

Material	*Chemical structure*	*Form*	*Point group*	*λ_{meas} [nm]*	*Refractive index*	*SHG susceptibility $\chi^{(2)}_{ijk}(-2\omega;\omega,\omega)$ [pm V^{-1}]*	*Electrooptical coefficient, r, [pm V^{-1}]*
MNA (2-methyl-4-nitroaniline)	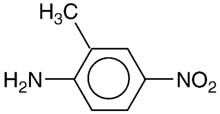	Single crystal	*m*	514.5 1064	2.0	$\chi^{(2)}_{111} = 500$ $\chi^{(2)}_{122} = 76$	70 ± 14
POM (3-methyl-4-nitropyridine-1-oxide)	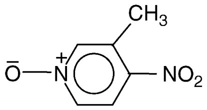	Single crystal	222	514.5 600 10 600 1064	1.712 1.919 1.638	$\chi^{(2)}_{123} = 19.2$ $\chi^{(2)}_{123} = \chi^{(2)}_{251} = \chi^{(2)}_{312}$	4.0 5.1 3.7
COANP (2 cyclooctylamino-5 nitropyridine)	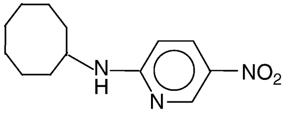	Single crystal	*mm*2	514.5 600 10 600 1064	1.672 1.781 1.647	$\chi^{(2)}_{322} = 108$ $\chi^{(2)}_{333} = 48$	9.3 ± 1.5 15.4 ± 3.2
PNP (2-(N-prolinol)-5-nitropyridine)	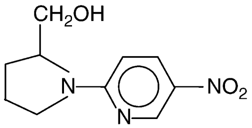	Single crystal	2	514.5 600 10 600 1064	1.990 1.788 1.467	$\chi^{(2)}_{211} = 170$ $\chi^{(2)}_{223} = 55$	8.0 ± 1.4 13.0 ± 2.2
MNBA (4′-nitrobenzylidene-3-acetamino-4-methoxyaniline)	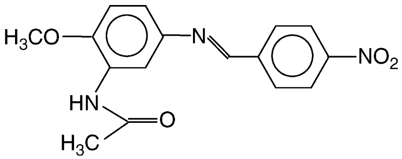	Single crystal	*m*	514.5 600 10 600 1064	2.024 1.585 1.652	$\chi^{(2)}_{111} = 437$	–
MMONS (3-Methyl-4-methoxy-4′-nitrostilbene)	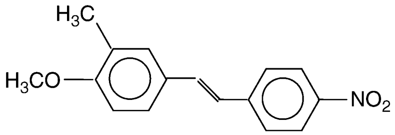	Single crystal	*mm*2	514.5 600 10 600 1064	1.569 1.693 2.129	$\chi^{(2)}_{333} = 368^a$ $\chi^{(2)}_{223} = 142^a$	19.3 ± 4 39.9 ± 7

PMMA + DR 1 (Poly(methyl methacrilate) + 4-[N-ethyl-N-(2-hydroxyethyl)] amino-4′-nitroazobenzene (Dispersed red 1))	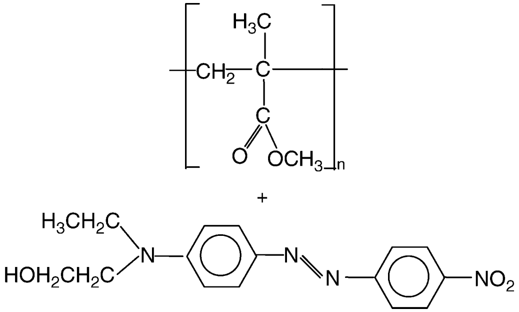	PP/GH	1580	$\chi^{(2)}_{333} = 5$	
			633		$r_{33} = 2.5$
PMMA + DCV (Poly(methyl metacrilate) + 4-dicyanovinyl-4′- [N-ethyl-(2 hydroxyethyl)] azobenzene)	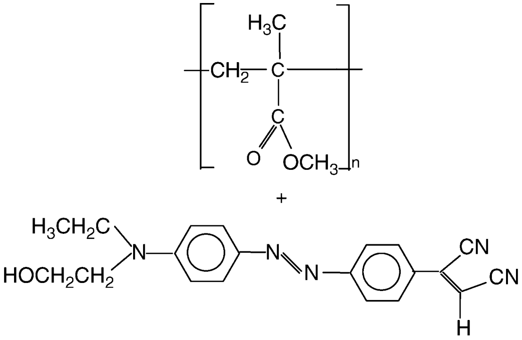	PP/SC	1580	$\chi^{(2)}_{333} = 42$	
			799		$r_{33} = 18$
PMMA + 3RDCVXY (Poly(methyl metacrilate) + 4-dicyanovinyl-4′ (diethylamino) diazobenzene)	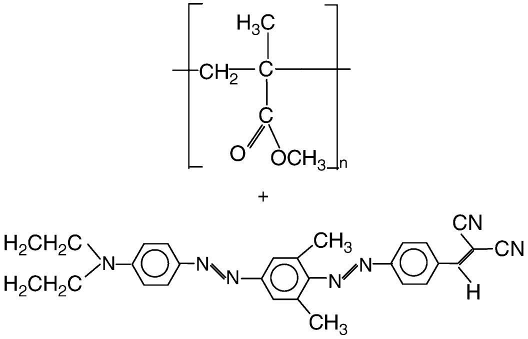	PP/SC	1064	$\chi^{(2)}_{333} = 840$	
			633		$r_{33} = 40$

Table 3 (*Continued*)

Material	*Chemical structure*	*Form*	*Point group*	*λ_{meas} [nm]*	*Refractive index*	*SHG susceptibility $\chi^{(2)}_{ijk}(-2\omega;\omega,\omega)$ [pm V^{-1}]*	*Electrooptical coefficient, r, [pm V^{-1}]*
NNDN + NAN (N,N-diglycidyl-4-nitroaniline + N-(2-aminophenyl)-4-nitroaniline)	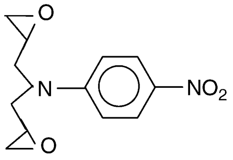 +	PP/CL		1064		$\chi^{(2)}_{311} = 28$	
	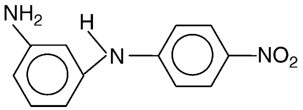			530.9			$r_{13} = 6.5$
DCANP	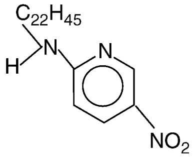	LB film		632.8	1.598 (TE$_0$-mode)		
				1064		$\chi^{(2)}_{333} = 19.5 \pm 2.5$	
FA06	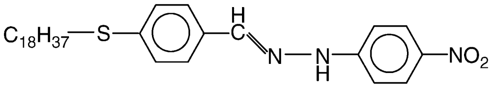	LB film		1064		$\chi^{(2)}_{322} = 20$	

PP, poled polymer; GH, guest host system; SC, side chain polymer; CL, cross linked polymer; LB, Langmuir–Blodgett film.

[a] Calibrated with KTP single crystal ($\chi^{(2)}_{333} = 27.4$ pm/V).

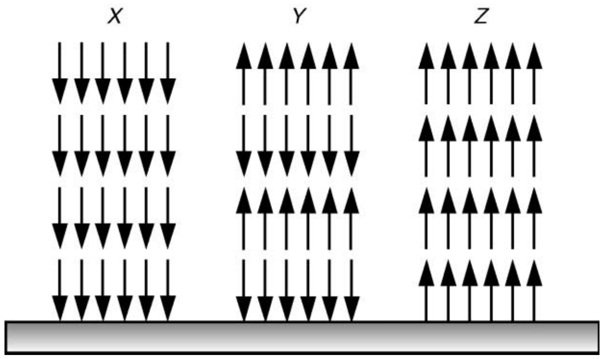

Figure 2 Schematic representation of *X, Y, Z* type Langmuir–Blodgett films. The arrows symbolize the direction of dipole moments, which usually are tilted with respect to the normal to film surface.

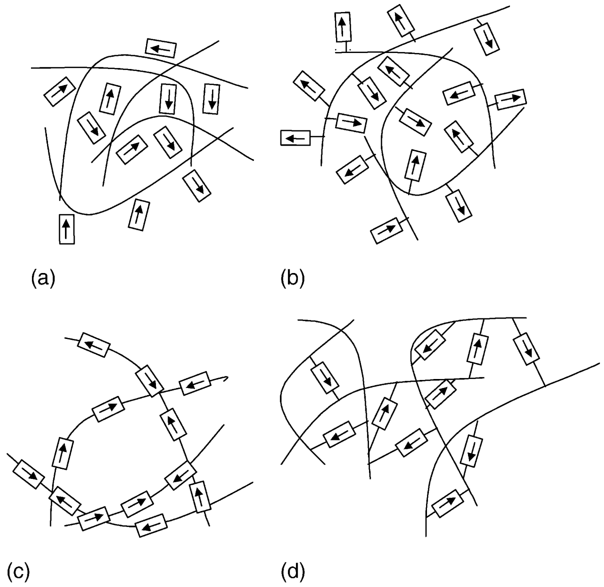

Figure 3 Schematic representation of functionalized polymers for second-order NLO effects: guest–host systems (a), side chain polymers (b), main chain polymers (c), photo and thermally cross-linking polymers (d). Arrows represent the NLO chromophores.

The guest–host systems are just solid solutions of NLO chromophores in an inert matrix. The concentration of NLO chromophores is limited because of already-mentioned strong dipole–dipole interaction leading to antiparallel alignment, resulting in an increase in propagation losses and decrease of $\chi^{(2)}$ susceptibility. The polar order is obtained by applying an external field. In order to increase the rotational mobility of chromophores, the material is heated to the glass transition temperature of polymer. The interaction of dipole moments with the external field forces them to orient in its direction. The orientation is frozen by cooling the material (usually a thin film) to room temperature. However, the stability induced in this way results in a polar order that is poor. Better stability and higher concentration of chromophores is obtained by attaching chemically the NLO active molecules to the polymer backbone (side-chain polymers). However, it requires a significant effort in chemical synthesis. The stability of induced order depends strongly on the difference between the glass transition and the operation temperatures. The larger the difference, the less relaxation is observed. Therefore, a lot of effort is going into developing functionalized high glass transition temperature polymers such as polyimides, polyurethanes, polyvinyl carbazoles, etc.

The main-chain polymers (iii) attract a relatively weak interest. This is due, not only to the difficulties encountered with chemical synthesis, but also with the efficient chromophore orientation. More promising is the last class of functionalized polymers (iv). Here the active chromophores are used to link covalently the polymer chains through a thermal or photo-crosslinking process. However, the poling process is very delicate as the glass transition temperature is increasing with the poling time, so the time process has to be carefully monitored. Also the cross-linking may lead to shrinkage and crackling of thin films and as a consequence, to the increase of propagation losses.

A special class of functionalized materials, similar to poled polymers, represent sol–gels. Here the active chromophores either fill the pores of sol–gel or are chemically bonded to the matrix. The poling process is very similar to that applied to the thermally and photo cross-linking polymers and also has to be carefully monitored. The main advantage of these materials consists of excellent optical properties, and the disadvantage on a relatively low chromophore content, limiting the bulk $\chi^{(2)}$ susceptibility value. The representative values of second-order nonlinear optical susceptibility $\chi^{(2)}_{ijk}$ for different materials for second-order NLO optics, are listed in Table 3.

Poling efficiency

The DC poling of functionalized polymers leads to birefringent noncentrosymmetric materials with two nonzero $\chi^{(2)}$ tensor components: the diagonal:

$$\chi^{(2)}_{ZZZ}(-\omega_0;\omega_1,\omega_2) = NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times\langle\cos^3\Theta\rangle \qquad [11]$$

and the off-diagonal:

$$\chi^{(2)}_{XXZ}(-\omega_\sigma;\omega_1,\omega_2) = \frac{1}{2}NF\beta_{xxx}(-\omega_\sigma;\omega_1,\omega_2) \times\langle\sin^2\Theta\cos\Theta\rangle \qquad [12]$$

where we assumed for simplicity, that only one nonlinear species, as is usually the case. Z, in eqns [11] and [12], is the direction of the applied external field and Θ is the angle between the electric field and the molecule dipole moment. The efficiency of poling is measured by the ratio:

$$a = \frac{\chi^{(2)}_{ZZZ}(-\omega_\sigma; \omega_1, \omega_2)}{\chi^{(2)}_{XXZ}(-\omega_\sigma; \omega_1, \omega_2)} \quad [13]$$

It varies between 1 and ∞ (perfect alignment). For isotropic poled polymers $a = 3 \div 6$ and for nematic liquid crystal side-chain polymers, values of $a \propto 18$ were obtained.

Usually, the second-order NLO susceptibility is measured by the second-harmonic generation technique ($\omega_\sigma = 2\omega$; $\omega_1 = \omega_2 = \omega$), giving the electronic part of $\chi^{(2)}$ susceptibility, with ultrafast response time as the generated harmonic wave has to follow the fundamental one. This susceptibility is important, not only for frequency conversion through the second-harmonic generation, but also for sum and difference frequency generation as well as for optical parametric oscillation (OPO) and amplification (OPA). Another important effect described by $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility is the linear electro-optic effect (Pockels effect). It finds application in electro-optic modulation for signal transmission. Traditionally, this effect is described by the electro-optic coefficient r_{ijk}, which is a third-rank three-dimensional tensor. The variation of the medium refractive index under the applied electric field is given by the following formula:

$$\Delta n_i = -\frac{2}{n_i^3} r_{ij} E_j \quad [14]$$

The electro-optic coefficient is directly linked to the $\chi^{(2)}(-\omega; \omega, 0)$ susceptibility:

$$r_{ijk} = -\frac{2\chi^{(2)}_{ijk}(-\omega; \omega, 0)}{(n_i n_j)^2} \quad [15]$$

The measured values of electro-optic coefficient for some selected organic single crystals and thin films are also given in Table 3.

Third-Order Nonlinear Optical Materials

There is no symmetry restriction on molecules and materials for third- (or higher-odd-) order nonlinear optical effects. These materials are used in significantly larger classes of applications than the materials for second-order NLO effects, particularly in logic systems, all optical switching, frequency conversion (e.g., frequency tripling in third-harmonic generation), light amplification, optical bistability, all optical switching, etc. In the case of organic materials, it was shown that the conjugated quasi 1D π electron systems, such as semiconducting polymers (Figure 4) exhibit an enhanced third-order NLO response.

Molecules

Similarly, as in the case of second-order NLO materials, the NLO response depends strongly on the conjugation length. In the case of conjugated quasi 1D molecules, such as polymers with a chemical structure as shown in Figure 4, the γ tensor component in the conjugation direction (largest hyperpolarizability) is strongly enhanced and described by a scaling law:

$$\gamma_{xxxx} \propto L^{\alpha_t} \quad [16]$$

where L is the conjugation length and the exponent $\alpha_t > 1$. The scaling law, eqn [16], can also be presented as dependent on γ on a number of π conjugated double or triple bonds N:

$$\gamma_{xxxx} \propto N^{\alpha_\gamma} \quad [17]$$

It was extensively studied for a series of conjugated oligomers polyenes, oligothiophenes, etc., and the exponential coefficients are listed in Table 4, together with the calculated values. For a larger value of double (or triple) bonds, a saturation in $\chi^{(3)}$ susceptibility is observed.

The molecular second hyperpolarizability measurements are derived from solution measurements. Some typical values, obtained by the third-harmonic generation technique (THG) are listed in Table 5. Although other NLO property determination techniques were used, the THG method gives the fast, electronic part of molecular hyperpolarizability.

Bulk Materials

Importance of Order

Similarly, as in the case of materials for second-order NLO effects for practical applications the value of $\chi^{(3)}$ susceptibility is important, which on the other hand depends on the molecular second hyperpolarizability γ. Although the noncentrosymmetry is not required in the case of third-order NLO materials, the values of the bulk $\chi^{(3)}$ susceptibility depend on the order, as it follows from eqn [7]. This is particularly clear in the case of conjugated molecules, with enhanced γ_{xxxx} component. In the case of parallel polarizations

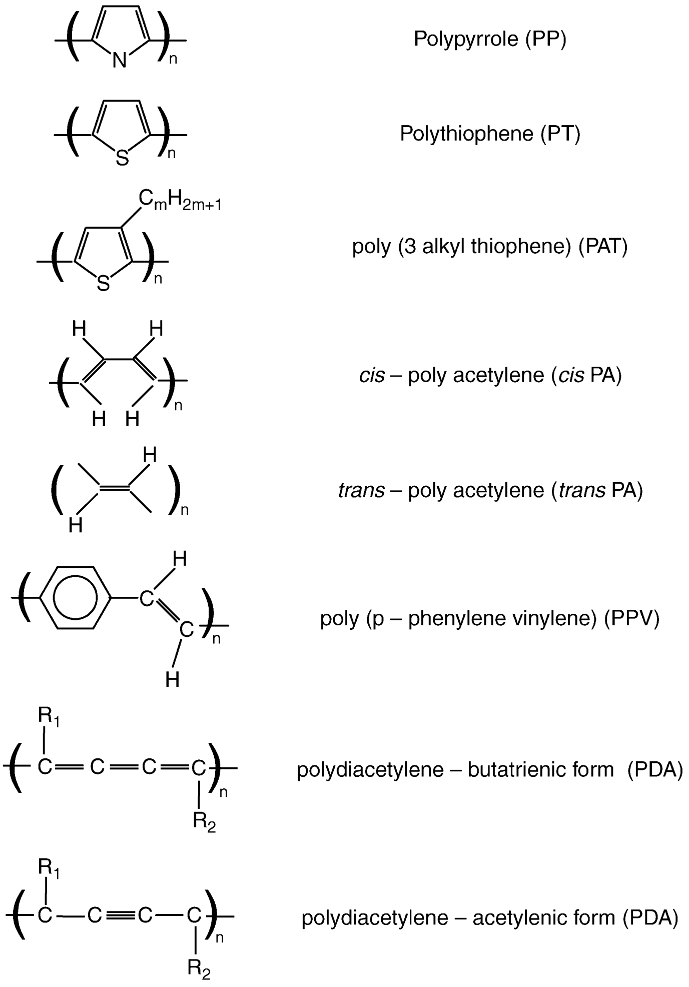
Polypyrrole (PP)
Polythiophene (PT)
poly (3 alkyl thiophene) (PAT)
cis – poly acetylene (*cis* PA)
trans – poly acetylene (*trans* PA)
poly (p – phenylene vinylene) (PPV)
polydiacetylene – butatrienic form (PDA)
polydiacetylene – acetylenic form (PDA)

Figure 4 Chemical structures of selected conjugated quasi 1D organic polymers.

Table 4 Calculated (calc) and measured (meas) values of the scaling low exponent $\gamma \propto N^{\alpha_\gamma}$ for different conjugated oligomers

Oligomers	$\alpha_{\gamma calc}$	*Remark*	$\alpha_{\gamma meas}$	*Remark*
Polyene oligomers	5	free electron model	2.82	CT polyenes ($\omega = 0$)
		INDO ($\omega = 0$)	3.5	$\lambda_\omega = 1.907$ μm
	3	SCF/STO-3G	2.3	Symmetric A-A
	3.27	SCF/3-21G		THG 1.907 μm
	3.14	$N = 2$	3.3	Asymmetric D-A
	3.67	$N = 6$, decrease		THG 1.907 μm
	3.98	For $N > 6$ DVB-PPP	4.4	Symmetric D-D
				EFISH 1.34 μm
				no saturation up to $N = 14$
	3.8	neutral solitons		
	3.85	charged polarons		
	6.32	*trans* form		
	5.4	*cis* form		
	4.7			
Thiophene oligomers	4.5	CNDO $N < 6$	4.05	DFWM in solutions
		Saturation $N > 6$	4.54	EFISH in solid solutions $N = 3, 5, 6$
			3.4	THG 1.907 μm

of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields, the bulk susceptibility is given by:

$$\chi^{(3)}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) = NF\gamma_{xxxx}(-\omega_\sigma; \omega_1, \omega_2, \omega_3) \times \langle \cos^4 \Phi \rangle \qquad [18]$$

where F is the total field factor and Φ is the angle between the polarization direction of $\omega_\sigma, \omega_1, \omega_2, \omega_3$ fields and a chosen direction in the laboratory reference frame. The value of $\langle \cos^4 \Phi \rangle$ varies between 0.2 for an isotropic distribution of polymer chains and 1 for perfectly aligned, respectively. It gives a factor of 5 difference in the macroscopic $\chi^{(3)}$ susceptibility. Some examples of third-order NLO susceptibilities for representative conjugated polymers are given in Table 6. These were obtained by third-harmonic generation technique which measures the fast, electronic contribution to $\chi^{(3)}$.

Table 5 Molecular off resonance second hyperpolarizability γ for selected organic molecules

Molecule	*$\gamma(-3\omega; \omega, \omega, \omega)$ 10^{-50} m^5/V^2*	*λ_{meas} nm*
Carbon tetrachloride	5.1	1064
	4.7	1907
Chloroform	3.8	1064
	3.2	1907
Acetone	2.5	1064
	2.6	1907
Methanol	1.16	1064
	1.21	1907
Ethanol	1.8	1064
	2.1	1907
Propanol	2.4	1064
	2.5	1907
Water	0.46	1064
Nitrobenzene	9.2	1907
Chlorobenzene	5.8	1907

Calibrated with silica value: $\chi^{(3)}(-3\omega; \omega, \omega, \omega) = 3.9 \times 10^{-22}$ m^2/V^2 at 1907 nm.

Two-Photon Absorbing Materials

Two-photon absorbing (TPA) materials are a special class of third-order NLO materials exhibiting a complex Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$. In this case, the nonlinear refractive index, derived from eqn [1] is complex too.

$$n = n_0 + n_2 I = n_0 + (n_2^r + \kappa_2)I \qquad [19]$$

where I is the light intensity, n_2^r is the real part and κ_2 is the imaginary part (nonlinear extinction coefficient) of the nonlinear index of refraction, respectively. The nonlinear index of refraction is directly related to the Kerr susceptibility $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$:

$$n_2 = \frac{3\chi^{(3)}(-\omega; \omega, -\omega, \omega)}{4\varepsilon_0 c n_0^2} \qquad [20]$$

where c is the light velocity. If $\chi^{(3)}(-\omega; \omega, -\omega, \omega)$ is complex (close to a resonance) then n_2 is complex too and the light beam propagating in such a medium in direction z is attenuated through the two-photon absorption process:

$$I(z) = I_0\, e^{-(\alpha_L + \alpha_{TPA} I)z} \qquad [21]$$

Table 6 Resonant and nonresonant values of cubic susceptibility $\chi^{(3)}$ for selected organic materials

Material	*Form*	*λ_{meas} nm*	*$\chi^{(3)}(-3\omega; \omega, \omega, \omega)$ 10^{-22} m^2/V^2*	*Remark*
PDA-PTS	Single crystal	1890	119 (70) × 10^3	‖ chain, 3ph
			<1400	⊥ chain
		2620	22.4 (14) × 10^3	‖ chain
			<280	⊥ chain
PDA-TCDU	Single crystal	1890	9.8 (7) × 10^3	‖ chain
			<70	⊥ chain
		2620	51.8 (20) × 10^2	‖ chain
			<40	⊥ chain
DEANS	Single crystal	1900	1.4 × 10^4	‖ (mol. axis)
			127	⊥ (mol. axis)
PDA-AFA	LB film (blue)	1350	7.8 (1.4) × 10^3	2ph
PDA-AFA	LB film (blue)	1907	11.6 (1.2) × 10^3	3ph
PA	Isotropic oriented	1907	78.4 (22) × 10^4	3ph
			2.4 × 10^6	3ph
PDA-C_4UC_4	Oriented film	1907	2.9 × 10^4	‖ chain, 3ph
			3 × 10^2	⊥ chain
C_{60}	Isotropic film	1064	11.5 (1.1) × 10^3	3ph
C_{70}	Isotropic film	1420	13 (1.4) × 10^3	3ph
PPV	Isotropic film	1450	1.96 × 10^4	3ph

Numbers in brackets are standard deviations. Data calibrated with silica: $\chi^{(3)}(-3\omega;\omega,\omega,\omega) = 3.9 \times 10^{-22}$ m^2/V^2; 2ph, two photon resonance enhancement; 3ph, three photon resonance enhancement; for chemical structures see Figure 4.

Table 7 Two photon absorption cross-sections σ_{TPA}, maximum TPA wavelength (λ_{max}), measurement wavelength (λ_{meas}), pulse duration and measurements technique for a series of organic molecules

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
BDBAS (Bis(di-*n*-butylamino)stilbene)	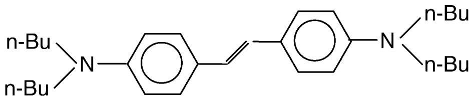	Toluene	600	310	4 ps	600	NLT
				9300	5 ns	600	
				1000	5 ns	700	
			600	210	5 ns	600	TPF
				6	5 ns	698	
				110	100 fs	620	
		Acetone		17 700	5 ns	600	NLT
	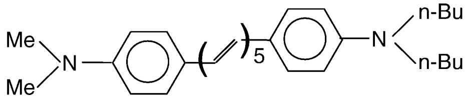	Toluene	755	10 000	5 ns	700	NLT
				17 000	5 ns	755	
				290	4 ps	700	
			730	300	5 ns	698	TPF
				790	5 ns	730	
BDPAS (Bis(diphenylamino)stilbene)	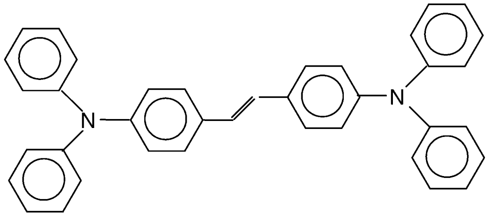	Toluene		13 000	5 ns	700	NLT
				110	4 ps	700	
Trans-stilbene	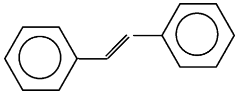	Toluene	514	12	5 ns	514	TPF
	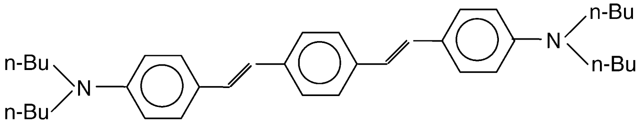	Toluene	730	995	5 ns	730	TPF
			~725	635	100 fs	725	
	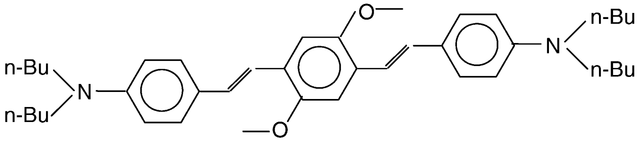	Toluene	730	900	5 ns	730	TPF
			~725	680	100 fs	725	

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
	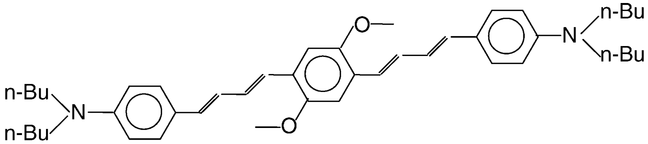	Toluene	775 ~750	1250 1270	5 ns 100 fs	775 750	TPF
	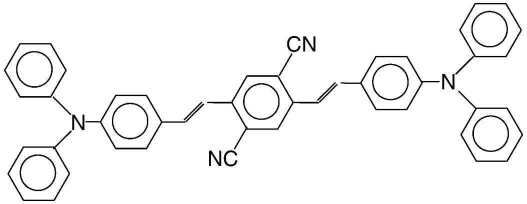	Toluene	835 810	1940 3670	5 ns 100 fs	835 810	TPF
	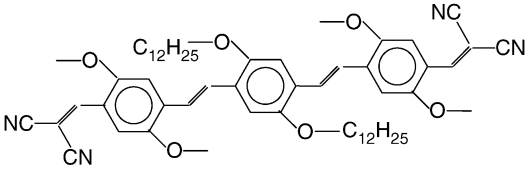	Toluene	825 940 815 910	480 620 650 470	5 ns 100 fs	825 940 815 910	TPF
	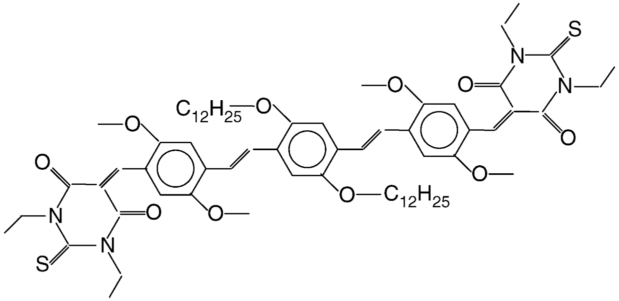	Toluene	970	1750	5 ns	970	TPF
	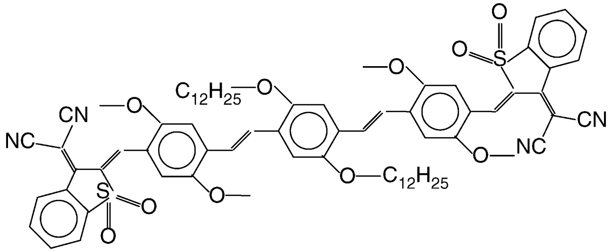	Toluene	975 945	4400 3700	5 ns 100 fs	975 945	TPF

	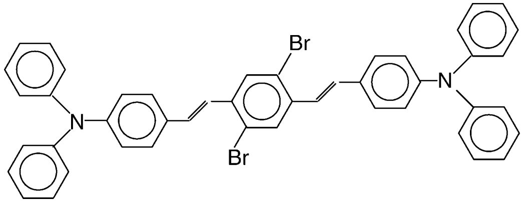	Toluene	~800	450	5 ns	800	TPF
CZ-DTT-CZ (2,6 -Bis [2-(9-ethylcarbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)	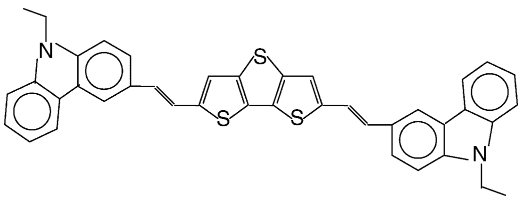	1,1,2,2-Tetra chloro-ethane		105 000	8 ns	810	NLT
CZ-DTT-OX (2-[2-(4-(5-(4-*tert*-Buthyl phenyl-1,3,4-oxadiazol-2-yl) phenyl) vinyl)-6-[2-(9-ethyl carbazol-3-yl) vinyl] thiopheno [3′,2′-2,3] thiopheno [4,5-b] thiophene)	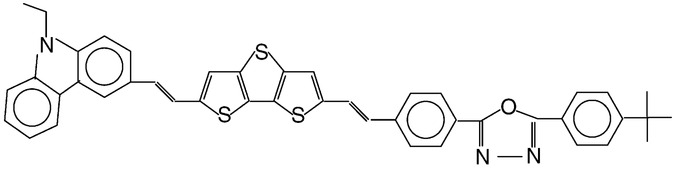	1,1,2,2-Tetra chloro-ethane		33 500	8 ns	810	NLT
TP-DTT-TP ([4-(2-(6-(2-(4-Diphenylamino) phenyl)vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl-vinyl) phenyl] diphenylamine)	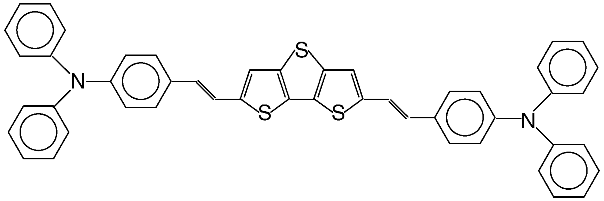	1,1,2,2-Tetra chloro-ethane		199 000	8 ns	810	NLT
TP-DTT-OX ([4-(2-(6-(2-(4-(5-(4-*tert*-Buthylphenyl-1,3,4-oxadiazol-2-yl)) phenyl) vinyl) thiopheno [2′,3′-4,5] thiopheno [3,2-b] thiophen 2-yl)-vinyl) phenyl] diphenylamine)	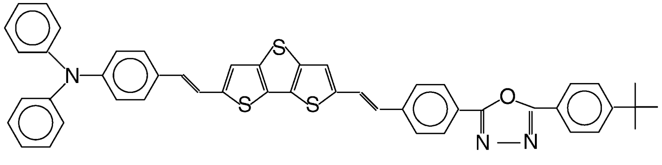	1,1,2,2-Tetra chloro-ethane		119 000	8 ns	810	NLT
Biphenyl	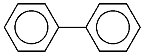			80			Theoretical calculus
Terphenyl	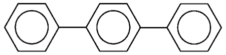			330			Theoretical calculus

Table 7 (*Continued*)

Molecule/material	*Chemical structure*	*Solvent*	λ_{max}^{TPA} *[nm]*	σ_{TPA} *[GM]*	*Pulse duration*	λ_{meas} *[nm]*	*Measurement technique*
Sexiphenyl	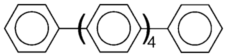			2090			Theoretical calculus
AF 50	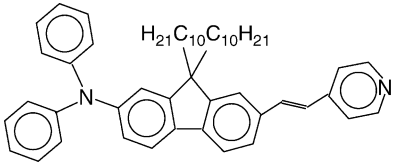	Tetrahydro-furan Benzene		11 560 19 300	8 ns	800	NLT
AF 80	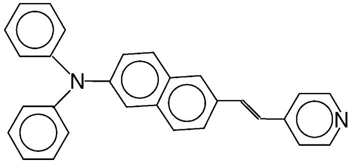	Tetrahydro-furan		6840	8 ns	800	NLT
ASPT	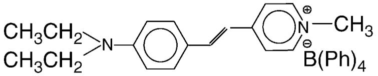	Poly hydroxyethyl metacrylate		12 000	10 ns	1064	NLT
DEANST	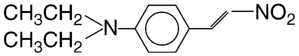	Tetrahydro-furan		200	46 ps	1064	KE
Rhodamine B	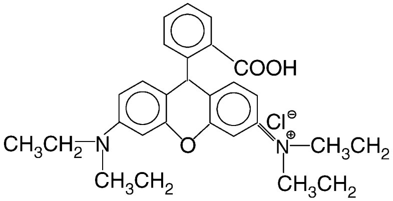	Methanol		50	100 fs	730	TPES
DR1 (4-[*N*-ethyl-*N*-(2-hydroxyethyl)] amino-4′-nitroazobenzene (dispersed red 1))	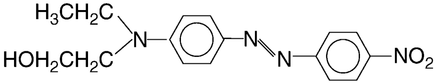	Tetrahydro-furan		102	42 ps	1064	KE

KE, optical Kerr effect; NLT, nonlinear transmission; TPF, two photon fluorescence; TPS, two photon excitation spectroscopy.
1 GM = 10^{-50} cm^4 s/photon-molecule.

Table 8 Two photon absorption coefficients for different molecules and extrapolated bulk values

Molecule	α_{TPA} *cm/GW*	λ_{max} *nm*	λ_{TPA} *nm*	*Concentration g/l*	*Solvent*	*Bulk value cm/GW*
DR 1	0.8	490	835	70	THF	11.4
DBANS	1.1	450	808	70	THF	15.7
DCNP	0.06	460	920	30	THF	2.3
DEANST	0.64	440	920	105	THF	6.1
DCM	0.1	480	950	11	DMSO	9.1
PDA-PTS	80	620	1064 not max of λ_{TPA}	Bulk	Single crystal	80
PDA-PTS	800	620	925	Bulk	Single crystal	800
Poly(3-octyl thiophene)	0.27	435 (142)	720 (94)	25	THF	9.6
Poly(3-octyl thiophene) regioregular	0.85	550 (170)	730 (70)	11.2	THF	76

λ_{max} corresponds to the maximum linear and λ_{TPA} to the maximum nonlinear absorption wavelengths, respectively, DCNP: 3-(1,1-dicyanoethenyl)-1-phenyl-4,5-dihydro-1H-pyrazole, for other molecules see **Figures 4** and **5** and preceding tables.

where α_{L} is the linear and α_{TPA} is the two-photon absorption coefficient, respectively. The last coefficient can be explicitly expressed by the imaginary part of the nonlinear index of refraction:

$$\alpha_{\mathrm{TPA}} = \frac{2\pi\kappa_2}{\lambda_\omega} \qquad [22]$$

where λ_ω is the wavelength of the propagating beam. On the molecular level one also defines the two-photon absorption cross-section given by

$$o_{\mathrm{TPA}} = \frac{\hbar\omega}{N}\alpha_{\mathrm{TPA}} \qquad [23]$$

where N is the number density of molecules. The values of TPA absorption cross-sections and two-photon absorption coefficients for selected organic molecules and solids are listed in Tables 7 and 8, respectively.

Photorefractive Materials

The photorefractive effect consists of the generation of a refractive index grating in an NLO material by its inhomogeneous illumination. This is a necessary condition but not a sufficient one. The sufficient condition consists of the possibility of creating the space charge field, which combined with the applied external field, leads to the refractive index modulation through the Pockels effect. In that case, the refractive index grating is phase mismatch with the intensity grating, leading to the transfer of photons from one beam to another, as occurs in two-beam coupling experiments, and thus applies to light amplification. Although the photorefractive effect was first observed in ferroelectric single crystals, the researchers at IBM demonstrated, at the beginning of 1990s, the photorefractive effect in organic composite materials. However, the main contribution in these materials comes from the rotational mobility of nonlinear optical (NLO) chromophores under the applied external electric field. Thus, to exhibit a photorefractive effect, the organic material has to possess the following properties:

(i) charge generation and separation capacity;
(ii) charge mobility;
(iii) linear electro-optic (Pockels) effect; and
(iv) rotational mobility.

Composite Materials

The different properties satisfying the above requirement to exhibit a photorefractive effect, are obtained by mixing together four types of molecules:

(i) photoconducting polymer;
(ii) highly second order nonlinear CT molecules;
(iii) photoensitizer for charge generation under illumination; and
(iv) plasticizer, lowering the glass transition temperature, thus increasing the rotational mobility.

Some examples of different types of molecules used are shown in Figure 5. Although these structures exhibit large exponential gain coefficients Γ and high diffraction efficiencies η, their main disadvantage consists of slow, seconds-duration response times

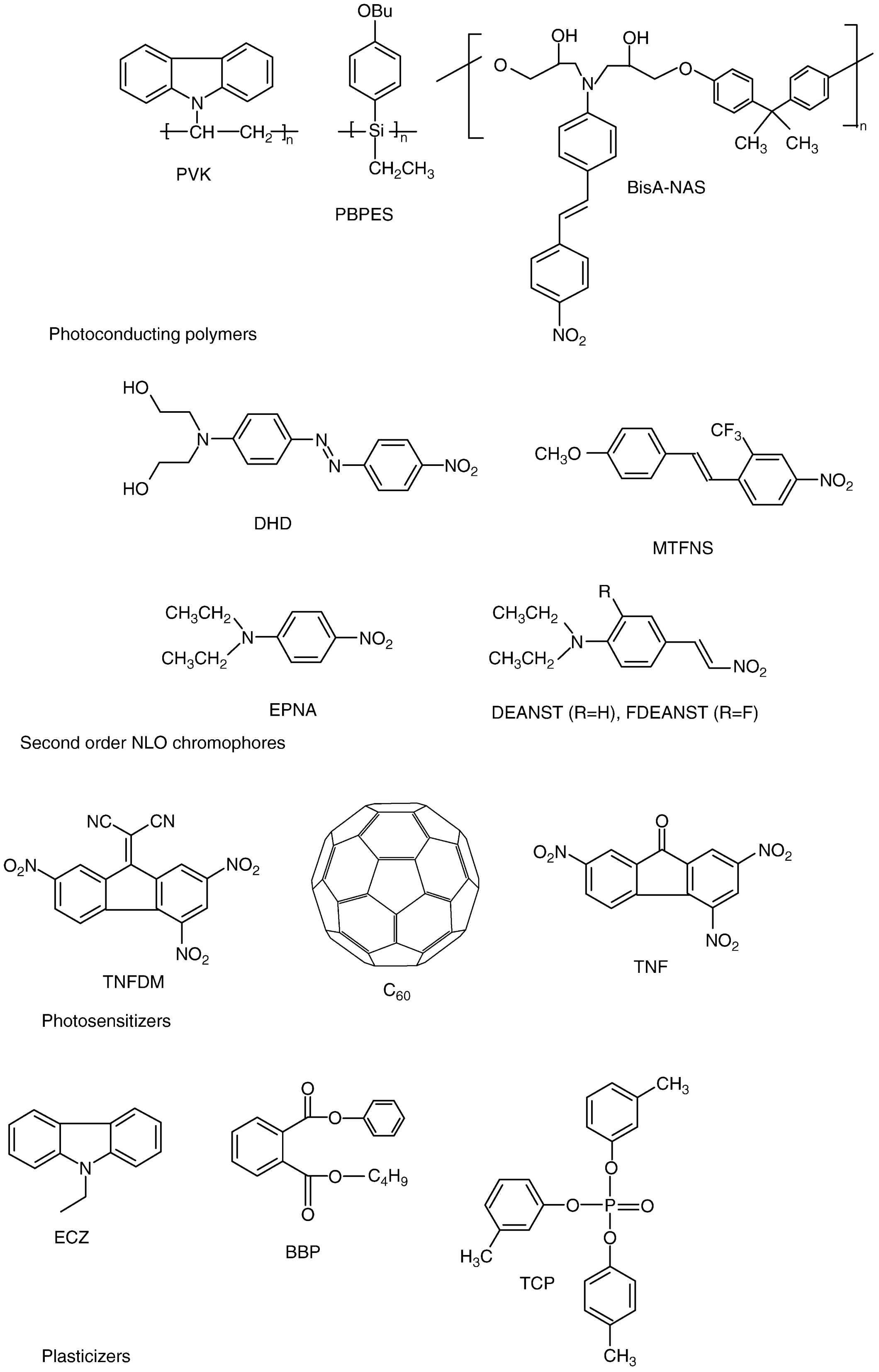
PVK
PBPES
BisA-NAS
Photoconducting polymers
DHD
MTFNS
EPNA
DEANST (R=H), FDEANST (R=F)
Second order NLO chromophores
TNFDM
C_{60}
TNF
Photosensitizers
ECZ
BBP
TCP
Plasticizers

Figure 5 Chemical structures of photoconducting polymers, CT chromophores, photosensitizers and plasticizers commonly used for the fabrication of photorefractive polymers.

Table 9 Photorefractive properties of selected single crystals at zero external field, composite photorefractive polymers and liquid crystals

Structure	*Form*	*Thickness* μm	*Composition w%*	*Laser intensity W/cm*^{2}	*Exponential gain* Γ *cm*$^{-1}$	*Electric field V/*μm	*Response time s*	*Diffraction efficiency %*	λ_{meas} *nm*
COANP: TCNQ	Single crystal			0.08			100–1000 depending on grating period Λ		514.5
MNBA	Single crystal			0.1	2		10–100		514.5
DMNPAA/PVK/TNF/ECZ	Composite material	105	50: 33: 16: 1		220 (p-polarized) -40 (s-polarized)	90	1	86	514.5
Doped: 5 CB/C_{60}	Composite material	25	95: 5		2890	0.07	0.3		514.5
Doped: 8 OCB/5 CB /PER/NI	Composite material	37			640	0.04	0.04		514.5
SCLCP/E7/C_{60}	Composite material	50	49.975: 49.975: 0.05		600	0.7	3.5		632.8
PVA/E44/C_{60}	PDLC	25				40	On: 8 s Off: >3 days	52	
PMMA/E49/ETZ/TNF	PDLC	53	45: 33: 21: 1		41	22		8	675
POT/LC	HPCPLC	10			965	0.6	0.011	26	514.5
POT/DR1/LC	HPCPLC	10			2600	0.6	0.010	31	514.5
PVK/TNF/LC	HPCPLC	10			3700	0.9	0.010	45	514.5

LC, liquid crystals; PDLC, polymer dispersed liquid crystals; HPCPLC, hybrid photoconducting polymer–liquid crystal; COANP-2-cyclooctalamino-5 nitropyridine, TCNQ, 7,7,8,8-tetracyanoquinodimethane; MNBA4′, nitrobenzylidene-3 acetamino-4-methoxyaniline; 5-CB-4′, (*n*-octyloxy)-4-cyanobiphenyl; 8OCB-4′, (*n*-octhyloxy)-4-cyanobiphenyl; PER, perylene; NI, *N*,*N*-di(*n*-octyl)1,4,5,8-naphtalenediimide; SCLCP, side chain liquid crystalline polymer (PMMA-4-cyanophenyl benzoate); E7, E44, E49, commercial LC mixtures (Merck, Germany); other molecules in **Figure 5** and **Tables 2** and **3**.

and require large external operation electric fields (~100 V/μm).

Photorefractive Liquid Crystals

In order to improve the performances of organic photorefractive materials the use of nematic liquid crystals was proposed as active media. These materials exhibit a large rotational mobility, requiring a relatively small electric field with response time of milliseconds. The reorientation of domains requires 2 ÷ 3 orders of magnitude electric fields as compared with the composite materials (reorientation of individual molecules). Several types of structures of liquid crystal based photorefractive structures are realized, such as:

- composite materials in which the liquid crystals mesogens are NLO chromophores;
- polymer dispersed liquid crystals (PDLCs);
- hybrid photoconducting polymers – liquid crystals (HPCPLCs), in which the rotational mobility is separated from the charge mobility. Moreover the use of photoconducting polymer allows the modulation of the electric field on the liquid crystal layer, enhancing the effect.

The performances of different, representative, organic photorefractive structures are listed in Table 9.

See also

Nonlinear Optics, Basics: Photorefraction; $\chi^{(2)}$–Harmonic Generation. **Polarization:** Introduction.

Further Reading

Agranovich VM and La Rocca GC (eds) (2002) Organic Nanostructures: Science and Applications. *Proceed. International School of Physics "Enrico Fermi", Course CXLIX*. Amsterdam: IOS Press.

Bosshard Ch, Sutter K, Pretre Ph, Hulliger J, *et al.* (1995) In: Kajzar F and Garito AF (eds) *Organic Nonlinear Optical Materials*, *Advances in Nonlinear Optics*, vol. 1. Amsterdam: Gordon and Breach Publishers.

Bredas JL and Silbey R (eds) (1991) *Conjugated Polymers*. Dordrecht: Kluwer Academic Publishers.

Charra F, Agranovich V and Kajzar F (2003) *Organic Nanophotonics*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Chemla DS and Zyss J (eds) (1987) *Nonlinear Optical Properties of Organic Molecules and Polymers*. Orlando, FL: Academic Press.

Graja A, Bulka B, Kajzar F (eds) (2002) *Molecular Low Dimensional-and Nano-Structured Materials for Advanced Applications*. NATO Science Series 2. Dordrecht: Kluwer Academic Publishers.

Hornac LA (ed.) (1992) *Polymers for Lightwave and Integrated Optics. Technology and Applications*. New York: Marcel Dekker, Publishers.

Kajzar F and Agranovich VM (eds) (2000) *Multiphoton and Light Driven Multielectron Processes in Organics: New Phenomena, Materials and Applications*. Dordrecht: Kluwer Academic Publishers.

Kajzar F and Reinisch R (eds) (1998) *Beam Shaping and Control with Nonlinear Optics*. NATO ASI Series, Series B: Physics, vol. 369. New York: Plenum Press.

Kajzar F and Swalen J (eds) (1996) *Thin Films for Waveguiding Nonlinear Optics*. Amsterdam: Gordon and Breach Scientific Publishers.

Kajzar F, Agranovich VM and Lee CY-C (1996) *Photoactive Organic Materials: Science and Application*. Dordrecht: Kluwer Academic Publishers.

Kuzyk M and Dirk C (eds) (1998) *Characterization Techniques and Tabulations for Organic Nonlinear Optical Materials*. New York: Marcel Dekker.

Lee KS and Wegner G (eds) (2003) *Polymers for Photonics Applications II: Nonlinear Optical, Photorefractive and Two-Photon Absorption Polymers*. Advances in Polymer Sciences 161. Berlin: Springer-Verlag.

Lessard R and Franke H (eds) (1997) *Materials Characterization and Optical Probe Techniques*. Critical Reviews of Optical Science and Technology, vol. CR69. Bellingham: SPIE.

Messier J, Kajzar F, Prasad P and Ulrich D (1989) *Nonlinear Optical Effects in Organic Polymers*. Dordrecht: Kluwer Academic Publishers.

Messier J, Kajzar F and Prasad P (1991) *Organic Molecules for Nonlinear Optics and Photonics*. Dordrecht: Kluwer Academic Publishers.

Miyata S and Sasabe H (eds) (1997) *Poled polymers and their Application to SHG and EO Devices*. Amsterdam: Gordon and Breach Scientific Publishers.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Richardson T (ed.) (1999) *Functional Polymers and Polymeric Materials*. Chichester, UK: John Wiley & Sons.

MICROSCOPY

Contents

Overview
Confocal Microscopy
Imaging Multiple Photon Fluorescence Microscopy
Interference Microscopy
Nonlinear Microscopy
Phase Contrast Microscopy

Overview

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The basic construction and imaging properties of the optical microscope, one of the most important optical instruments, are described. But the restrictive meaning of a microscope as a device that the user looks down to see a magnified image is far from the present terminology that includes scanning systems (where the image is stored in a computer), nonvisible radiation, such as electrons or X-rays, and a variety of different contrast mechanisms.

According to the *Oxford English Dictionary* a microscope is ‘an instrument magnifying objects by means of lenses so as to reveal details invisible to the naked eye’. Thus, basically a microscope forms a magnified image of an object. The first microscopes used visible light and formed an image showing variations in the intensity of light scattered by the object. This intensity is in general related to the optical properties of the object in a complicated way.

Nowadays there are many different types of microscope. Our understanding of a microscope must therefore be generalized in the following ways.

- The microscope may not be a conventional imaging system using lenses or mirrors. For example, it may be a scanning imaging system. A confocal system is a combination of a conventional and a scanning system. Another important category of microscopes is that of probe microscopes, including near-field optical microscopes. Some probe microscopes, for example the atomic force microscope (AFM) do not even seem at first glance to rely on the direct use of radiation.
- The microscope may not use visible light, but other electromagnetic radiation, or even other forms of radiation.
- The image can be formed using a variety of different contrast mechanisms. Some of these, such as phase contrast, are designed to image particular optical properties of the sample. Others image the generation of a form of radiation when the object is stimulated by another form of radiation. In particular, in principle, virtually any form of spectroscopy can be the basis of building up an image by measuring the spatial variations in the signal.

Different Forms of Microscopes

The Conventional Microscope

In a simple microscope, an objective lens forms a real, magnified, and inverted image of an object. In a compound microscope, an eyepiece is added, forming a virtual image that is viewed by eye to give real image on the retina. The eye can be replaced by CCD camera to record moving or still images. In early microscopes, the objective lens and eyepiece are mounted in the two ends of a brass tube. The objective screws into the tube using an RMS (Royal Microscopical Society) thread (0.8″ Whitworth 36 threads per inch), until its mounting face is flush with the end of the tube. The length of the tube is the mechanical tube length. The distance from the mounting face of the objective to the plane of the real image is the optical tube length, which varied according to different manufacturers in the range 160–210 mm. Now most objectives have infinity tube length, so that they collimate the light from the object. The collimated light is brought to a focus by an additional tube lens. An objective of correct tube length should always be used as otherwise spherical aberration is introduced. As the objective, tube lens, and eyepiece are designed together as a system, great care should be exercised when mixing components from different manufacturers.

Abbe showed that in order to form a perfect lateral image the objective must satisfy a different condition called the Herschel condition. Otherwise the image of an off-axis point will suffer from coma. An aberration-free system satisfying the sine condition is called an aplanatic system. A microscope produces a three-dimensional (3D) image of a 3D object. A point of the object closer to the lens appears further from the lens in the image. If the lateral magnification is M, the longitudinal magnification is approximately equal to M^2. However, it can be shown that it is impossible to devise an instrument that can produce a perfect 3D image of a 3D object, because to have a perfect transverse image the system must satisfy Abbe's sine condition, but to have a perfect axial image the system must satisfy the Herschel condition. These two conditions are not mutually compatible. Actually, we find that the longitudinal magnification is not exactly equal to M^2, and is not even constant in space. Different depths of the object are imaged differently: the microscope objective only produces a good image for one plane of focus, and other planes will exhibit spherical aberration. However, all of these problems can be overcome by bringing the part of the object under observation into focus, without refocusing the eye or other detector. A 3D image can thus be generated by scanning the object stage in the axial direction, and recording a stack of images from different depths in a computer. For modern objectives of infinite tube length, focusing can be alternatively achieved by piezoelectric scanning of the objective lens.

Resolution

For an object consisting of a single bright point object in a dark background, the image produced by a perfect microscope, according to paraxial diffraction theory, is the so-called Airy disk, consisting of a bright central spot, surrounded by a series of rings. The intensity in a meridional cross-section through the Airy disk is the same for any aperture of objective, or wavelength of light, but its width varies. The intensity of the first bright ring is 1.75% of that at the peak. The radius of the first dark ring of the Airy disk is

$$r_0 = \frac{0.61\lambda}{n \sin \alpha} \qquad [1]$$

where λ is the wavelength, n is the refractive index of the immersion medium, and α is the angle subtended at the edge of the objective aperture. The radius depends on $n \sin \alpha$, which is called the numerical aperture and should be as large as possible for good resolution. Thus, high-magnification lenses often use an immersion fluid, usually oil with a refractive index of 1.518 (ISO 8036/1).

The optical coordinate v is defined as

$$v = \frac{2\pi}{\lambda} rn \sin \alpha \qquad [2]$$

so that the dark ring of the Airy disk occurs at a value of v of $2\pi \times 0.61 = 3.83$. Resolution is sometimes expressed in terms of the full width at half maximum (FWHM) of the image of a point object, which is 3.232 in optical coordinates. A full nonparaxial theory does not give a very different figure.

Resolution of a microscope is often specified by two-point resolution, which describes whether the image of two points can be distinguished from that of a single point. According to the Rayleigh criterion of two-point resolution, two points are just resolved if the second point is placed on the first dark ring of the first. The separation is then $0.61\lambda/(n \sin \alpha)$. A cross-section through the image in an incoherent optical system of two bright point objects of equal strength and different separations is shown in Figure 1. The intensity in the image between the points decreases as the separation is increased. The ratio of the intensity midway between the points to that at the points for the Rayleigh criterion to be satisfied is then found to be 0.735.

The concept of resolution should be distinguished from sensitivity or precision. An object much smaller than the resolution limit can still be detected, perhaps weakly, in a microscope: this detection of weak contrast depends on the sensitivity of the microscope rather than its resolution. The size of an object can also be measured with a precision much greater than the resolution of the microscope.

Illumination

If the object is not self-luminous, in order to be imaged in a microscope it must be illuminated. A semitransparent object, such as a biological slice, is illuminated in transmission. For observation of bulk objects, or surfaces, we use illumination in the reflection geometry, called epi-illumination. Either a tungsten halogen lamp or an arc source is usually used. In critical illumination, the source is focused on to the object by a condenser lens. The disadvantage of this approach is that variations in emission of the source are imaged directly into the image.

Cheaper microscopes avoid this problem by using a diffuser. Better microscopes use instead the Köhler illumination system, where the source is placed in the front focal plane of the condenser lens. Figure 2 shows a Köhler illumination system in which the source is imaged into the front focal plane of the condenser lens by another, projector, lens. The system incorporates a separate field stop and aperture stop. In practice, the size of the field stop should be reduced to illuminate as small a region of the object as necessary in order to minimize stray light. An important part of setting up a microscope to operate properly consists in centering the aperture stop of the condenser. This is often done using a so-called Bertrand lens that allows imaging of the aperture stop.

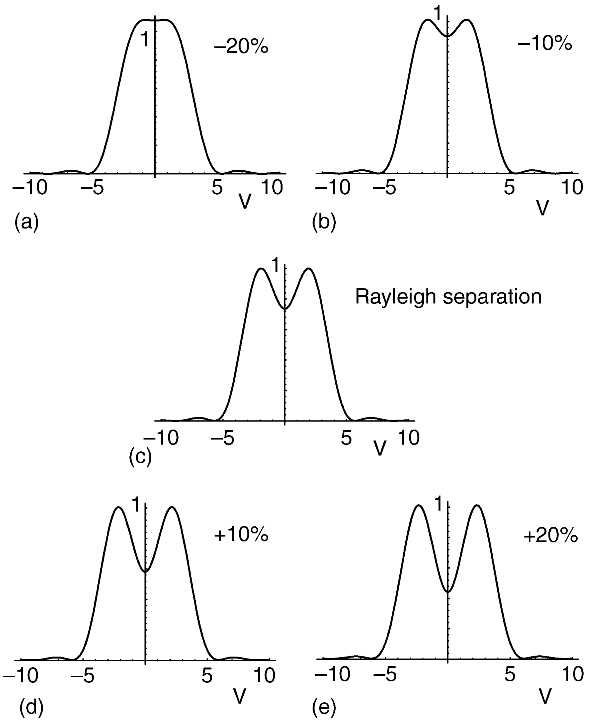

Figure 1 A cross-section through the image of two equal point objects in an incoherent microscope. The separation of the points is that for the Rayleigh criterion to be satisfied, and for changes of ±10, 20% of the Rayleigh separation.

Image formation

The Rayleigh criterion was originally specified for an incoherent optical system. It is thus applicable in fluorescence microscopy, as fluorescence is an incoherent process as two fluorescent point objects fluoresce independently. But for the image of a trans-illuminated object, the resolution depends on the coherence of the illumination. This is controlled by the numerical aperture of the condenser lens as compared with that of the objective. The ratio of these numerical apertures is called the coherence parameter, S, so that

$$S = \frac{n_c \sin \alpha_c}{n \sin \alpha} \qquad [3]$$

For $S = 0$ the illumination is purely coherent, while for $S \rightarrow \infty$ imaging becomes purely incoherent, thus corresponding to the Rayleigh criterion, so S should really be called an incoherence parameter. An important in-between case is when $S = 1$, when the apertures of objective and condenser are equal. This is termed matched, full, or complete illumination. According to the generalized Rayleigh criterion, the points are just resolved when the ratio of the intensity midway between the points to that at the points is 0.735, which is called the generalized Rayleigh criterion. We require that the distance between the points is as small as possible when this occurs. The generalized Rayleigh separation is the same for incoherent illumination and for $S = 1$, and corresponds to a distance of $2v_0 = 3.83$ in optical coordinates. Note, however, that for other separations of the points, the ratio is different for these two cases: it is not true, but often erroneously stated, that $S = 1$ corresponds to incoherent imaging. Resolution improves as the aperture of the condenser is increased ($2v_0 = 5.15$ in optical coordinates for coherent illumination) reaching a maximum when $S = 1.46$, when resolution is 9% better than for incoherent imaging ($2v_0 = 3.58$). This maximum in resolution is achieved when the numerical aperture of the condenser is larger than that of the objective, which is not

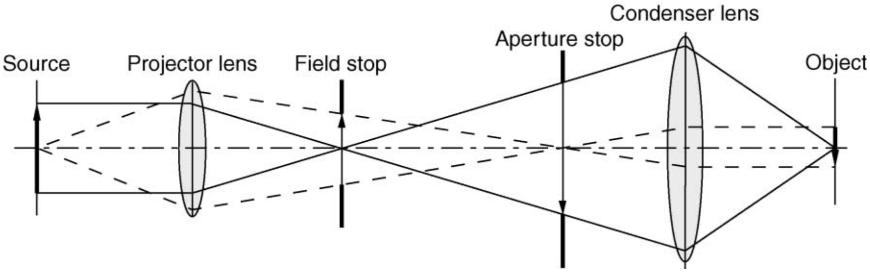

Figure 2 Geometry of the Köhler illumination system. In practice an arrangement with a smaller number of lenses is usually used.

physically achievable if the objective has the highest possible numerical aperture. Similarly, it is impossible to achieve incoherent illumination with very large objective numerical apertures.

The effect of coherence on the resolution of an image depends on the form of the object, and on the particular resolution criterion employed. In practice, contrast also depends on the aperture of the condenser. Usually, opening the condenser improves resolution, but reduces contrast, so there will be an optimum condenser aperture size.

Abbe theory of microscope imaging

Abbe argued that for coherent illumination of a grating, the grating spectra can be observed in the back focal plane of the objective. The tube lens then forms an image from the grating components. Thus, this is an early description of Fourier optics. The strength of the grating components can be altered in the back focal plane to give various optical effects, such as phase contrast. Abbe's theory was not properly appreciated at the time because it was known that in practice, resolution could be improved by using a larger condenser aperture, which does not give coherent illumination. There was much controversy about the merits of Köhler versus critical illumination, and different sizes of condenser aperture.

We now know that Köhler and critical illumination are in principle equivalent. The aberrations of the condenser lens are not important, so that the source and condenser together behave simply as a partially coherent effective source. Imaging can still be described by Fourier optics within the framework of partially coherent imaging theory.

Although Köhler and critical illumination are equivalent in principle, Köhler illumination is in practice better because of its improved illumination uniformity.

Depth of focus

The longitudinal image of a point object is found to be approximately invariant when expressed in terms of an optical coordinate u, defined as

$$u = \frac{8\pi}{\lambda} zn \sin^2 \frac{\alpha}{2} \qquad [4]$$

The axial resolution, and thus also the depth of focus, varies strongly with aperture. The FWHM for longitudinal imaging is 11.13 in optical coordinates.

Microscope objectives

Because microscope objectives often have large apertures, aberrations will be very strong, unless the objective is designed correctly. The basic design principle for a high aperture lens is the aplanatic front element (Figure 3). It is found that all rays from a point A in a sphere of refractive index n, where A is a distance r/n from its center, appear to come from a point B, distance rn, without any aberration. This principle is used by employing a front element that has one surface of the same radius of curvature as the sphere. The front surface is a sphere centered on A (Figure 3b), so that the rays are not deviated on crossing it.

A low-power achromat uses just two components with different optical dispersion to cancel chromatic aberration and correct for other aberrations (an achromatic doublet). For higher powers, two separated doublets are necessary. For the highest apertures, an aplanatic front is used. However, this can never converge the light, but only makes it less divergent. It is thus used together with an achromatic doublet. For oil immersion, the interface between the oil and the front element is not very important, so it is made planar to simplify manufacture. In practice, it is

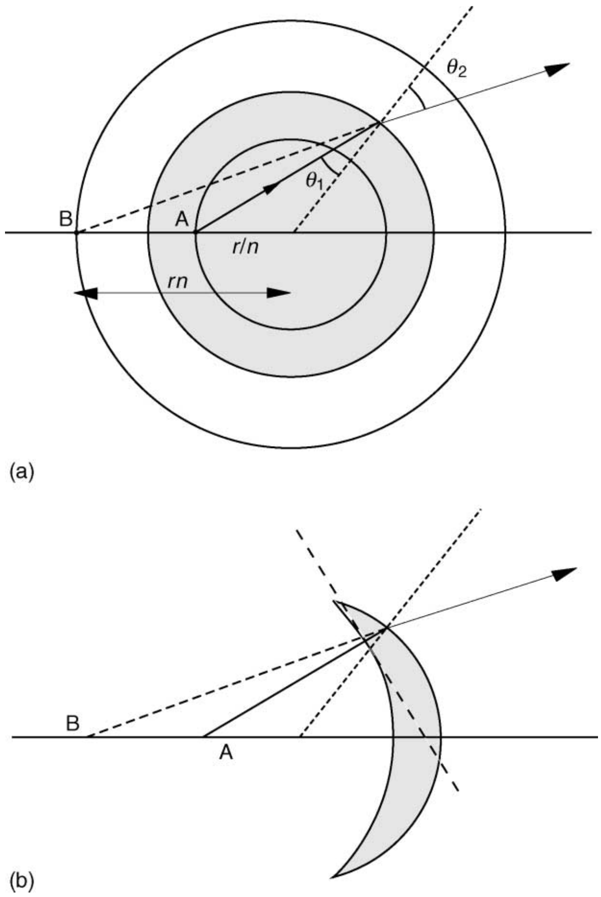

Figure 3 The principle of the aplanatic front: (a) an aplanatic surface; (b) a meniscus lens that acts as an aplanatic system.

usually made planar even for a dry lens, and the resulting aberrations cancelled out elsewhere in the objective. Apochromats, corrected for three colors, use more elements, which may include two or more stages of aplanatic front. The final objective thus consists of a number of elements that must be accurately aligned relative to one another. The objective is adjusted by the manufacturer to give a good star image (image of a point object). First a spacer is selected to optimize spherical aberration. Then a sleeve is adjusted to make the objective parfocal with others. Finally a screw is used to center the assembly to remove coma.

Conventional and Scanning Microscopes

Conventional microscope

Various different types of transmission microscope are illustrated in Figure 4. We consider image formation in a conventional microscope as illustrated schematically in Figure 4a, showing a microscope with critical illumination. A large area incoherent source is focused by the condenser lens on to the specimen, illuminating a comparatively large area of the specimen, corresponding to the whole field of view of the objective. Information from each illuminated point in the specimen is simultaneously transmitted by the objective lens to form the primary image. The objective is responsible for forming the image, with the condenser playing only a secondary role in determining the resolution of the system, through control of the coherence of the illumination. Figure 4b shows a conventional microscope in which the image is measured point by point by a detector. In practice this could be achieved by using a CCD detector.

Scanning microscope

An image can be generated by a scanning system, as illustrated in Figure 4c. A probe of light is formed by demagnification of a source, and is scanned over the object in a raster. The transmitted (or reflected) light is detected by a photodetector and thus builds up an image. The size of the probe of light limits the resolution of the system, that is, the smallest detail in the object that can be seen in the image. It should be noted that the magnification of the image is given simply by the ratio of the distance scanned in the image to the distance scanned by the probe. It is thus unrelated to the demagnification of the source. The image magnification can be altered without changing the lens and is not in any way related to the resolution.

It has been shown that the imaging properties of scanning and conventional microscopes are identical under analogous conditions, this being known as the principle of equivalence. This property is based on the principle of reciprocity, which is a very general physical law that holds even with diffraction,

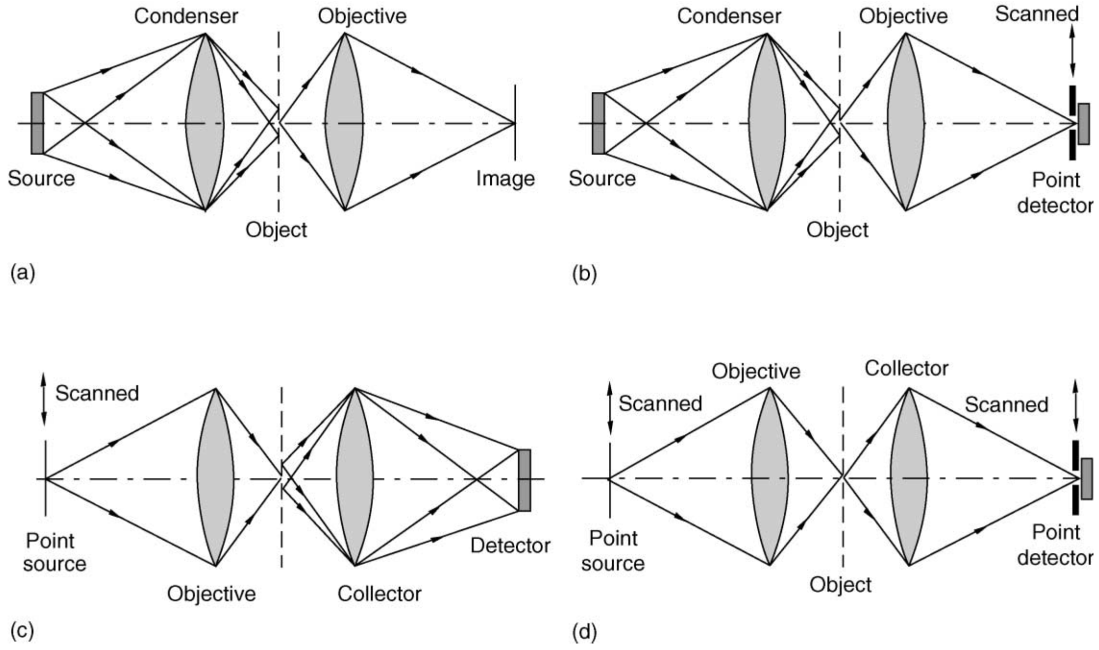

Figure 4 Different geometries of transmission microscope: (a) a conventional microscope; (b) a conventional microscope with a point detector; (c) a scanning microscope; (d) a confocal microscope.

absorption, multiple scattering, and stray light. The principle of equivalence is generally valid except if there are nonreciprocal magnetic or polarization effects, or energy losses involved (such as in fluorescence microscopy, and with inelastic scattering in electron microscopy).

Scanning systems exhibit a number of important advantages over conventional systems. Broadly, these are based on two classes of property. First, in a scanning system the image is in the form of an electronic signal. As a result, this is advantageous for quantitative measurements, as well as for image processing, including image enhancement, image restoration, and image analysis. Second, in a scanning system the object is illuminated by a focused spot, which extends the range of imaging modes available.

Confocal microscope

Finally in Figure 4d, we combine the arrangements of Figures 4b and c to give a confocal scanning optical microscope, in which a point source illuminates just a small region of the object, and a confocal point detector detects light from this illuminated region. If the point source and detector are scanned in unison, a two-dimensional image is generated. However, this system now behaves very differently from the previous ones. The confocal microscope is thus not a special case of the general partially coherent conventional imaging system. From the symmetry of Figure 4d, it is clear that the two lenses play an equal part in the imaging process. This results in an improvement in resolution. In fact, the confocal system behaves as a coherent imaging system, but with a sharper effective point spread function than in a conventional coherent microscope.

Although Figure 4 is drawn for the transmission geometry, in practice most confocal systems operate in the reflection or epi-illumination mode, in which the same objective lens is used both for illumination and detection. For a specimen placed in the focal plane, the properties of confocal transmission and reflection systems are identical. However, once the object is moved from the focal plane, certain differences arise: in particular in confocal reflection a strong optical sectioning effect occurs that allows a single section through a thick object to be imaged. This is the major advantage of the confocal microscope arrangement. Figure 4 also applies equally well to fluorescence imaging. However, in this case, because of the incoherent nature of fluorescence emissions after excitation by coherent light, imaging is then incoherent.

Confocal microscopes can be achieved by using either a point detector, in practice performed by placing a pinhole in front of a photodetector, or by using a coherent detector. Such a coherent detector, sensitive to the amplitude of the incident radiation, is not directly available for light, but is available for acoustic radiation, as in a scanning acoustic microscope. For light, a confocal effect can also be achieved by using an interferometric method to synthesize a coherent detector. Thus, in interference microscopes using the high aperture condenser and objective lenses, an optical sectioning effect arises similar to that in the confocal microscope. Another way of producing a coherent detector is to use a single-mode optical fiber.

Probe Microscopes

Scanning tunneling microscope

The scanning tunneling microscope was the first of the family of probe microscopes, relying on a fundamentally different principle from the usual forms of microscope. A physical tip in the nanometer range is brought close, also in the nanometer range, to a conducting sample. If an electric potential is applied, electrons can tunnel across between the sample and the tip. By scanning the tip mechanically across the sample, an image can be generated with a resolution in the subnanometer range. The surface topography can also be measured with subnanometer sensitivity. While interference microscopy can measure surface topography with subnanometer sensitivity, in this case the profile is averaged over the lateral resolution of the microscope, of the order of the wavelength in dimensions. The scanning tunneling microscope can image individual atoms. The signal is related to the work function of the material. Altering the bias voltage allows the band structure of the material to be investigated.

Atomic force microscope

In the atomic force microscope, the force between the surface and the tip is measured using a cantilever beam and used to build up an image. Atomic force microscopy can be performed in either a contact or noncontact mode. It can be used with insulating specimens. Optical methods, often using a position-sensitive detector, are used to measure displacements of the cantilever.

Near-field scanning optical microscope (NSOM)

In a near-field scanning optical microscope, a very small tip or aperture is scanned relative to the specimen to attain resolution greater than the classical limit set by the wavelength of the radiation. As Figure 5 shows, various different designs of near-field microscope have been proposed. It is evident that

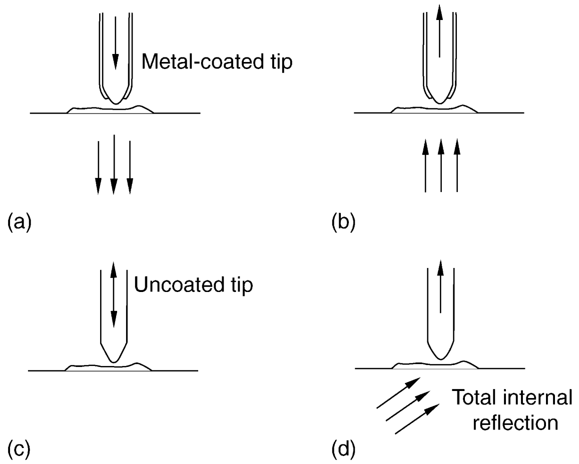

Figure 5 Different forms of near-field microscope: (a) illumination-mode NSOM; (b) collection-mode NSOM; (c) a near-field probe used for both illumination and collection in a confocal arrangement; (d) photon tunneling microscope (PTM). Reproduced with permission from Sheppard CJR (1978) The scanning optical microscope. *The Physics Teacher* 16: 648–651. © 1978 American Association of Physics Teachers.

these again form into the conventional/scanning categories described earlier. The sample can be illuminated using the near-field probe, as in Figure 5a. This is called the illumination-mode NSOM. Or the sample can be uniformly illuminated, and a signal detected using a near-field probe as in Figure 5b, giving a collection-mode NSOM. Or a confocal arrangement can be used, with a near-field probe used for both illumination and collection (c). In this case, an uncoated tip can be used; as otherwise the detected signal will be very small. Finally, Figure 5d shows the photon tunneling microscope (PTM), in which the sample is illuminated with evanescent waves produced by total internal reflection and an uncoated tip used to probe the evanescent field in the presence of the sample.

Different Types of Radiation

Electromagnetic Radiation

UV radiation and X-rays

The first microscopes used visible light, but, in fact, we do not need to use visible light, but can use any electromagnetic radiation, over a broad range of different available wavelengths. In Figure 6 we show the wavelength of the electromagnetic spectrum, illustrating how the wavelength in air varies with frequency. Instead of using visible light, shorter wavelengths allow greater resolution to be achieved. Ultra-violet light is commonly used to excite fluorescence. X-rays have also successfully been used to image biological samples.

IR and microwave radiation

In the longer wavelength region, infra-red radiation is used for observation of semiconductors or for molecular spectroscopic imaging. Imaging of emitted midinfra-red radiation can show up variations in temperature, such as hot spots in semiconductors. Microwaves have also been used, but resolution is limited by the long wavelength. For this reason microwaves (and mid-IR) have also been used for near-field microscopy.

Other Radiation

Electron microscope

The other important class of radiation that can be used for illumination is matter waves. The most common of these, of course, is electrons, the wavelength of which, as a function of acceleration voltage, is shown in Figure 7. The bend in the curve at very high voltages is caused by relativistic effects. The wavelength of the electrons is very small: even for 100 V electrons, the wavelength is only 0.1 nm. But aberrations caused by the electron lenses limit

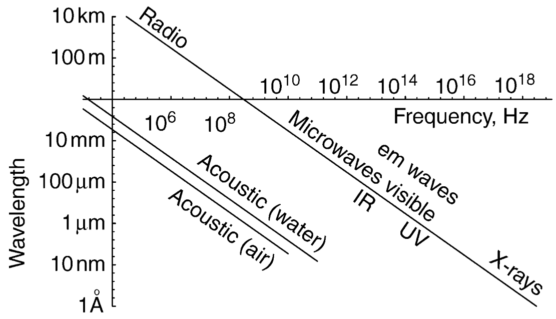

Figure 6 The wavelength of different forms of electromagnetic radiation and acoustic radiation. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

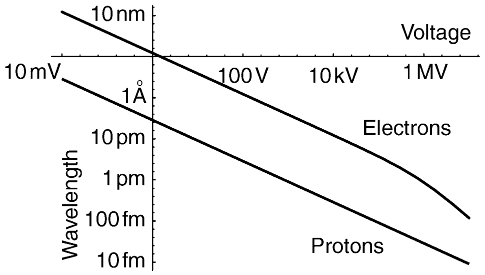

Figure 7 The wavelength of matter waves (electrons and protons) as a function of accelerating voltage. Reproduced with permission from Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications and Advances*, pp. 1–18. New York: Wiley-Liss.

resolution so that at present atomic resolution is only just achievable, even with much higher voltages.

In a transmission electron microscope (TEM), a beam of electrons is produced by an electron gun. The specimen is illuminated via a condenser lens (or lenses). Usually these are magnetic lenses in which a current in a coil produces a magnetic field that focuses the electrons. The electrons that are transmitted through the specimen are then focused by a series of lenses to form a final image that can be viewed directly on a phosphor screen. An X-ray detector can be used to characterize the elements in the specimen. An electron-energy-loss spectrometer, together with a detector, can measure the energy of the electrons in a particular region of the image.

Other matter radiation

Instead of electrons, other particle beams can be used for illumination. Figure 7 shows the wavelength of a beam of protons, as function of the accelerating voltage. Because of the larger mass of protons, relativistic effects are not apparent for protons in the curve in Figure 7. Microscope images have also been formed using ions, and even neutral particles such as neutrons. Neutral atom microscopes are also under development.

Acoustic radiation

Instead of electromagnetic waves, acoustic radiation can be used to form a scanning acoustic microscope (SAM). Ultrasonic waves are focused into a water immersion medium. Mechanical scanning is again used to build up an image. Figure 6 shows the wavelength of the acoustic wave as a function of frequency. By using microwave frequencies, wavelengths, and therefore resolution, in the submicron range, can be generated. The SAM acts as a confocal microscope because the detector is sensitive to the amplitude of the acoustic wave. The acoustic microscope provides information on variations in the elastic properties (viscosity and elasticity) of the sample.

Different Contrast Mechanisms

The bright field microscope detects the intensity of the light transmitted through the object. A microscope with a condenser aperture of appreciable size behaves as a partially coherent system, and the image will, in general, depend on the phase of the light coming from the object as well as its intensity. Many objects are weakly scattering, so the contrast of this phase information is weak. In this case, the microscope effectively images variations in the transmittance or reflectance of the sample. Different designs of microscope can be used to image various properties of the object, thus providing new and complementary information. Phase contrast microscopy provides but one example of a different contrast mechanism. In fact, in general, any physical or chemical interaction of the illuminating radiation with the sample can be used as the basis for a contrast mechanism. These different contrast mechanisms can be performed in either a conventional or scanning arrangement, or in a confocal system.

Phase Contrast Microscopy

Phase contrast is a widely used technique in both biological microscopy (in transmission) and materials microscopy (in reflection). Its importance in biological microscopy stems, firstly, from the fact that contrast may otherwise be too weak to be visible, but secondly because the phase variations are caused by changes in optical thickness, related to physical density. In reflection microscopy, phase is related to surface height variations, but also depends on the optical material properties of the sample.

Polarization Microscopy

The polarization state of light transmitted through or reflected from a sample can also be changed. This change is detected in polarization microscopy, for example by placing the sample between crossed polarizers. The birefringence detected can be either material birefringence, or form birefringence (caused by the shape of the microstructure). Polarization microscopy is widely used in biological microscopy, and in mineralogy.

Fluorescence Microscopy

In fluorescence microscopy, electrons excited by illumination decay to the ground state and emit photons of light. Autofluorescence is natural fluorescence of the sample. Fluorescent dyes can be used as labels for particular biological or chemical constituents. Fluorescence microscopy is usually performed in the epi-illumination geometry to assist rejection of light from the source. Often it is performed in a confocal microscope, allowing 3D localization of the fluorescent labels.

Raman Microscopy

Raman spectroscopy can be performed on a microscopic scale to produce images of presence of particular molecular bonds. It is usually performed in a scanning system, often of the confocal type.

This is one example of the wider class of spectroscopic microscopies: in principle, any form of spectroscopy can be used as the basis for a microscope contrast mode.

Nonlinear Microscopy

Another class of contrast mechanisms is based on nonlinear optical interactions. Again, any nonlinear mechanism can be used as the basis for a contrast mechanism. These include two- (or multiphoton) fluorescence, second- (or higher-order) harmonic generation, sum or difference frequency generation, and coherent anti-Stokes Raman scattering (CARS). Probably other nonlinear optical mechanisms will be exploited in microscopy in the future.

Other Contrast Mechanisms

The contrast mechanisms described above, all rely on illumination and detection of light. But in many other modes, radiation of one form can be converted into another. These include cathodoluminescence microscopy (electrons to light), X-ray microanalysis (electrons to X-rays), photo-emission microscopy (light, usually UV, to electrons), photo-acoustic microscopy (light to ultrasonics), photothermal microscopy (light to thermal waves), optical or electron beam induced current (OBIC/EBIC). Again, in many of these, the illuminating or emitted radiation can be focused, or a combination can be used in a confocal arrangement.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Interference Microscopy; Nonlinear Microscopy; Phase Contrast Microscopy.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Bracegirdle B (1997) *Introduction to the Optical Microscope*. Oxford: Bios Scientific Publishers.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Martin LC (1966) *Theory of the Microscope*. London: Blackie.

Paesler MA and Moyer P (1996) *Near-Field Optics: Theory, Instrumentation, and Applications*. New York: Wiley InterScience.

Pluta M (1988) *Advanced Light Microscopy, Volume 1: Principles and Basic Properties*. Amsterdam: Elsevier.

Sheppard CJR (1987) Scanning optical microscopy. In: Barer R and Cosslett VE (eds) *Advances in Optical and Electron Microscopy*, vol. 10, pp. 1–98. London: Academic Press.

Sheppard CJR (2002) The generalized microscope. In: Diaspro A (ed.) *Confocal and Two-Photon Microscopy: Foundations, Applications, and Advances*, pp. 1–18. New York: Wiley-Liss.

Slayter EM and Slayter HS (1992) *Light and Electron Microscopy*. Cambridge: Cambridge University Press.

Wiesendanger R (1994) *Scanning Probe Microscopy and Spectroscopy*. Cambridge: Cambridge University Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

Confocal Microscopy

T Wilson, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is probably fair to say that the development and wide commercial availability of the confocal microscope has been one of the most significant advances in light microscopy in the recent past. The main reason for the popularity of these instruments derives from their ability to permit the structure of thick specimens of biological tissue to be investigated in three dimensions. It achieves this important goal by resorting to a scanning approach together with a novel (confocal) optical system.

The traditional wide-field conventional microscope is a parallel processing system which images the entire object field simultaneously, which is a severe requirement for the optical components. We can relax this requirement if we no longer try to image the whole object at once. The limit of this relaxation is to require a good image of only one object point at a time. The price that we have to pay is that we must scan in order to build up an image of the entire field.

A typical arrangement of a scanning confocal optical microscope is shown in Figure 1 where the system is built around a host conventional microscope. The essential components are some form of mechanism for scanning the light beam (usually from a laser) relative to the specimen and appropriate photodetectors to collect the reflected or transmitted light.

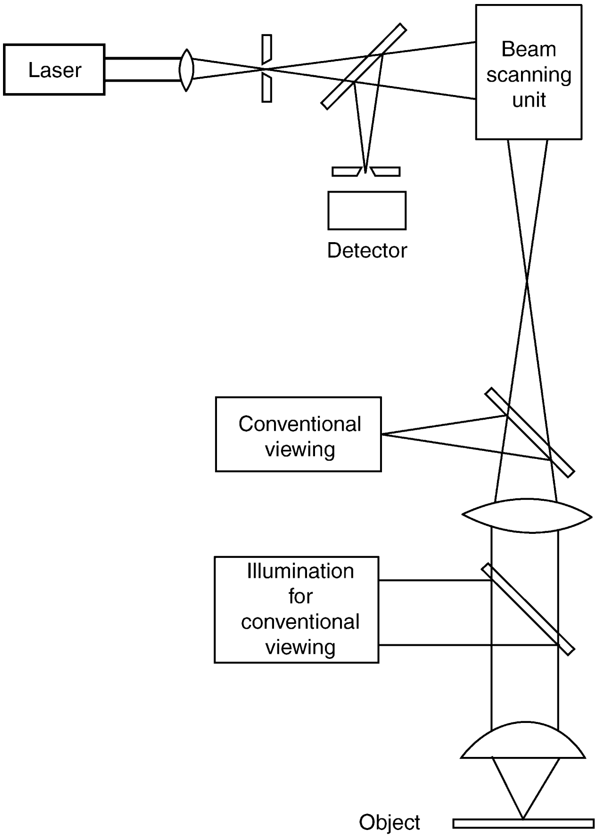

Figure 1 Schematic diagram of a confocal microscope.

Most of the early systems were analog in nature, but it is now universal, due to the serial nature of the image formation, to use a computer both to drive the microscope and to collect, process, and display the image.

In the beam scanning confocal configuration of Figure 1, the scanning is typically achieved by using vibrating galvanometer-type mirrors or acousto-optic beam deflectors. The use of the latter gives the possibility of TV-rate scanning, whereas vibrating mirrors are often relatively slow when imaging an extended region of the specimen, although significantly higher scanning speeds are achievable over smaller scan regions. We should note that other approaches to scanning may be implemented, such as specimen-scanning and lens-scanning. These methods, although not generally available commercially, do have advantages in certain specialized applications. Since this article is necessarily limited in length much additional material may be found in other sources listed in the Further Reading at the end of this article.

Image Formation in Scanning Microscopes

We will not discuss the fine detail of the optical properties of confocal systems. However, the essence is shown in Figure 1 where we see that the confocal optical system consists simply of a point source of light which is then used to probe a single point on the specimen. The strength of the reflected or fluorescence radiation from the single object point is then measured via a point, pinhole detector. The confocal – point source and point detector – optical system therefore merely produces an 'image' of a single object point and hence some form of scanning is necessary to produce an image of an extended region of the specimen. However, the use of single-point illumination and single-point detection results in novel imaging capabilities which offer significant advantages over those possessed by conventional wide-field optical microscopes. In essence, these are enhanced lateral resolution and, perhaps more importantly, result in a unique depth discrimination or optical sectioning property. It is this latter property which leads to the ability to obtain three-dimensional images of volume specimens.

The improvement in lateral resolution may at first seem implausible. However, it can be explained simply by the principle which states that resolution can be increased at the expense of field of view, which can then be increased by scanning. One way of taking advantage of this principle is to place a very small aperture extremely close to the object. The resolution is now determined by the size of the aperture rather than the radiation. In the confocal microscope we do not use a physical aperture in the focal plane but rather use the back-projected image of a point detector in conjunction with the focused point source. Figure 2 indicates the improvement in lateral resolution that may be achieved.

The confocal principle was introduced in an attempt to obtain an image of a slice within a thick specimen which was free from the distracting presence of out-of-focus information from surrounding planes. The confocal optical systems fulfills this requirement and its inherent optical sectioning or depth discrimination property has become the major

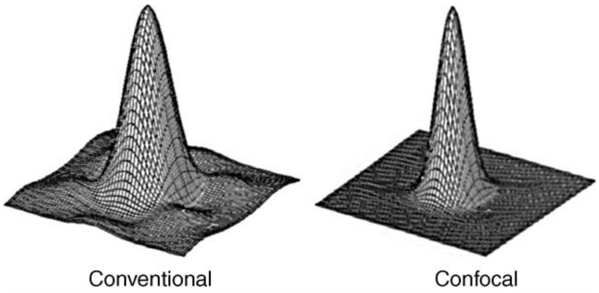

Figure 2 The point spread functions of the conventional and confocal microscopes showing the improvement in lateral resolution which may be obtained in the confocal case.

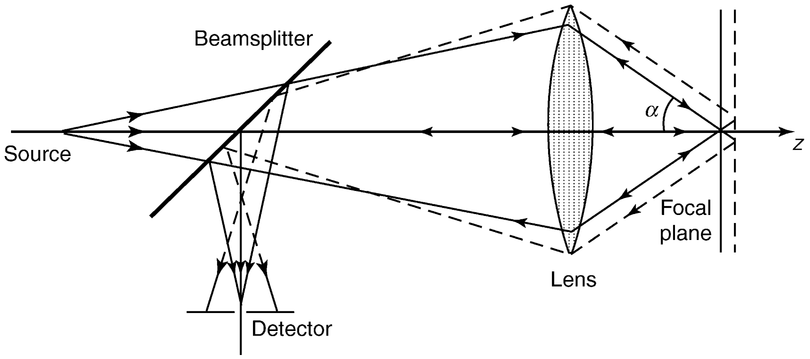

Figure 3 The origin of the optical sectioning or depth discrimination property of the confocal optical system.

motivation for using confocal microscopes, and is the basis of many of the novel imaging modes of these instruments. The origin of the depth discrimination property may be understood easily from Figure 3, where we show a reflection-mode confocal microscope and consider the imaging of a specimen with an undulating surface. The full lines show the optical path when an object feature lies in the focal plane of the lens. At a later scan position, the object surface is supposed to be located in the plane of the vertical dashed line. In this case, simple ray tracing shows that the light reflected back to the detector pinhole arrives as a defocused blur, only the central portion of which is detected and contributes to the image. In this way the system discriminates against features which do not lie within the focal region of the lens. A very simple method of both demonstrating the effect and giving a measure of its strength is to scan a perfect reflector axially through focus and measure the detected signal strength. Figure 4 shows a typical response. These responses are frequently termed the $V(z)$, by analogy with a similar technique in scanning acoustic microscopy. A simple paraxial theory models this response as

$$I(u) = \left[\frac{\sin(u/2)}{u/2} \right]^2 \qquad [1]$$

where u is a normalized axial coordinate which is related to real axial distance, z, via

$$u = \frac{8\pi}{\lambda} nz \sin^2(\alpha/2) \qquad [2]$$

where λ is the wavelength and $n \sin \alpha$ the numerical aperture. As a measure of the strength of the sectioning, we can choose the full width at half intensity of the $I(u)$ curves. Figure 5 shows this value as a function of numerical aperture for the specific case of imaging with red light from a helium neon laser. These curves were obtained using a high aperture theory which is more reliable than eqn [1]

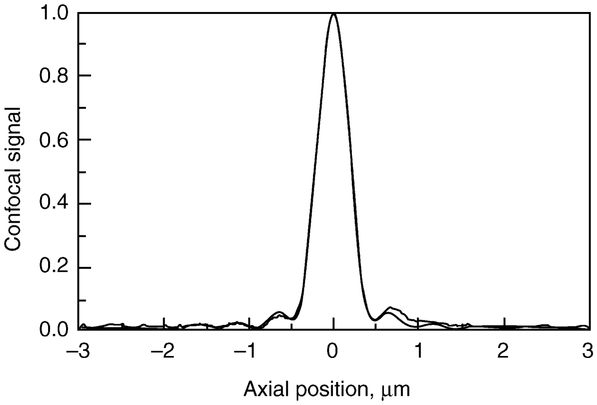

Figure 4 The variation in detected signal as a plane reflector is scanned axially through focus. The measurement was taken with a 1.3 numerical aperture objective and 633 nm radiation.

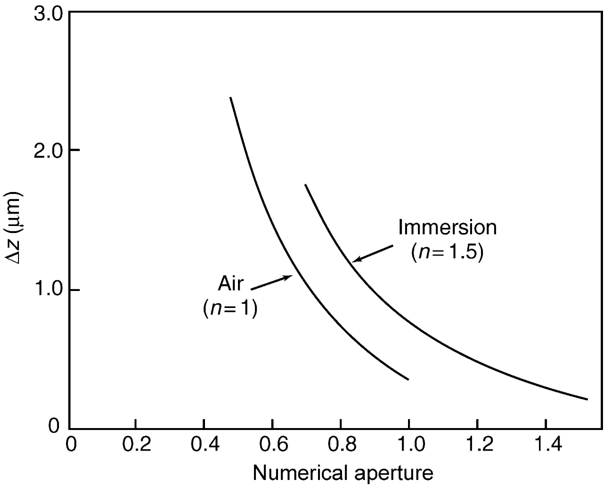

Figure 5 The optical sectioning width as a function of numerical aperture. The curves are for red light (0.6328 µm wavelength). Δz is the full width at the half-intensity points of the curves of $I(u)$ against u.

at the highest values of numerical aperture. We note, of course, that these numerical values refer to nonfluorescence imaging. The qualitative explanation of optical sectioning, of course, carries over to the fluorescence case but the actual value of the optical sectioning strength is different in the fluorescence case.

Applications of Depth Discrimination

Since this property is one of the major reasons for the popularity of confocal microscopes, it is worthwhile, at this point, to review briefly some of the novel imaging techniques which have become available with confocal microscopy.

Figure 6 illustrates the essential effect: Figure 6a shows a conventional image of a planar microcircuit which has deliberately been mounted with its normal at an angle to the optic axis. We see that only one portion of the circuit, running diagonally, is in focus. Figure 6b shows the corresponding confocal image: here the discrimination against detail outside the focal plane is clear. The areas which were out of focus in Figure 6a have been rejected. Furthermore, the confocal image appears to be in focus throughout the visible band, which illustrates that the sectioning property is stronger than the depth of focus.

This suggests that if we try to image a thick translucent specimen, we can arrange, by the choice of our focal position, to image detail exclusively from one specific region. In essence, we can section the specimen optically without having to resort to mechanical means. Figure 7 shows an idealized schematic of the process. The portion of the beehive-shaped object that we see is determined by the focus position. In this way it is possible to take a through-focus series and obtain data about the three-dimensional structure of the specimen. If we represent the volume image by $I(x, y, z)$, then by focusing at a position $z = z_1$, we obtain, ideally, the image $I(x, y, z_1)$. This, of course, is not strictly true in practice because the optical section is not infinitely thin.

It is clear that the confocal microscope allows us to form high-resolution images with a depth of focus sufficiently small that all the detail which is imaged appears in focus. This suggests immediately that we can extend the depth of focus of the microscope by adding together (integrating) the images taken at different focal settings without sacrificing the lateral resolution. Mathematically, this extended-focus image is given by

$$I_{\mathrm{EF}}(x, y) = \int I(x, y, z)\mathrm{d}z \qquad [3]$$

As an alternative to the extended-focus method, we can form an auto-focus image by scanning the object axially and, instead of integrating, selecting the focus at each picture point by recording the maximum in the detected signal. Mathematically, this might be written as

$$I_{\mathrm{AF}}(x, y) = I(x, y, z_{\max}) \qquad [4]$$

where $z_{\max}$ corresponds to the focus setting giving the maximum signal. The images obtained are somewhat similar to the extended focus and, again, substantial increases in depth of focus may be obtained. We can go one step further and turn the microscope into a noncontacting surface profilometer. Here we simply display $z_{\max}$.

It is clear by now that the confocal method gives us a convenient tool for studying three-dimensional structures in general. We essentially record the image as a series of slices and play it back in any desired fashion. Naturally, in practice it is not as simple as this, but we can, for example, display the data as an $x - z$ image rather than an $x - y$ image. This is somewhat similar to viewing the specimen from the side. As another

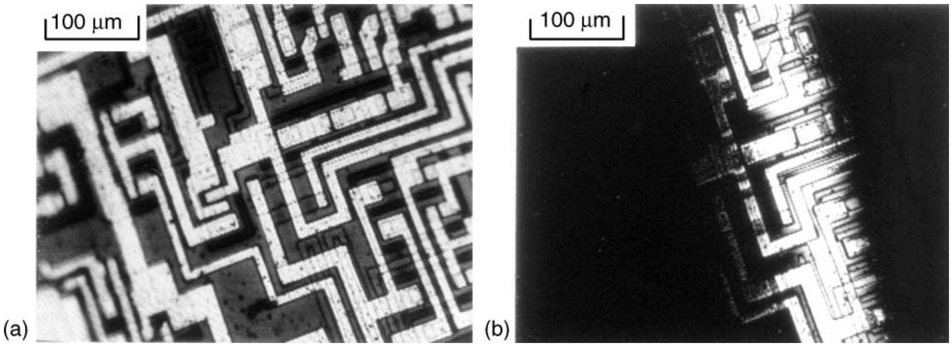

Figure 6 (a) Conventional scanning microscope image of a tilted microcircuit: the parts of the object outside the focal plane appear blurred. (b) Confocal image of the same microcircuit: only the part of the specimen within the focal region is imaged strongly.

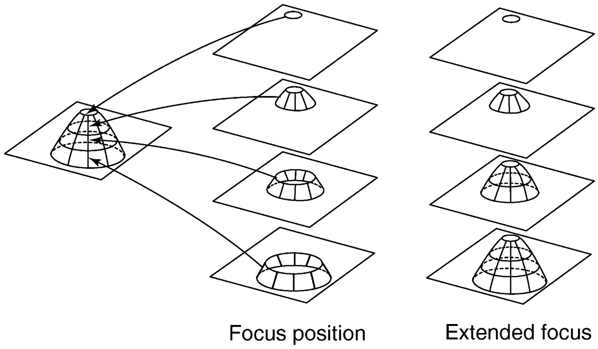

Figure 7 An idealization of the optical sectioning property showing the ability to obtain a through-focus series of images, which may then be used to reconstruct the original volume object at high resolution.

example, we might choose to recombine the data as stereo pairs by introducing a slight lateral offset to each image slice as we add them up. If we do this twice, with an offset to the left in one case and the right in another, we obtain, very simply, stereo pairs. Mathematically, we form images of the form:

$$\int I(x \pm \gamma z, y, z)\mathrm{d}z \qquad [5]$$

where γ is a constant. In practice it may not be necessary to introduce offsets in both directions to obtain an adequate stereo view.

Our discussion so far on these techniques has been by way of a simplified introduction. In particular, we have not presented any fluorescence images. The key point is that, in both brightfield and fluorescence modes, the confocal principle permits the imaging of specimens in three dimensions. The situation is, of course, more involved than we have implied. A thorough knowledge of the image formation process, together with the effects of lens aberrations and absorption, is necessary before accurate data manipulation can take place.

In conclusion to this section it is important to emphasize that the confocal microscope does not produce three-dimensional images. It essentially produces very high-quality two-dimensional images of a (thin) slice within a thick specimen. A three-dimensional rendering of the entire volume specimen may then be generated by suitably combining a number of these two-dimensional image slices from a through focus series of images.

Fluorescence Microscopy

We now turn our attention to confocal fluorescence microscopy, because this is the imaging mode which is usually employed in biological applications. Although we have introduced the confocal microscope in terms of bright-field imaging the comments we have made concerning the origin of the optical sectioning, etc., carry over directly to the fluorescence case although the numerical values describing the strength of the optical sectioning are, of course, different and we will return to this point later.

If we assume that the fluorescence in the object destroys the coherence of the illuminating radiation and produces an incoherent fluorescent field proportional to the intensity of the incident radiation, $I(v,u)$, then we may write the effective intensity point spread function, which describes image formation in the incoherent confocal fluorescence microscope, as

$$I(v,u)I\left(\frac{v}{\beta},\frac{u}{\beta}\right) \qquad [6]$$

where the optical coordinates u and v are defined relative to the primary radiation and $\beta = \lambda_2/\lambda_1$ is the ratio of the fluorescence to the primary wavelength. We note that $v = (2\pi/\lambda)rn \sin\alpha$, where r denotes the actual radial distance.

This suggests that the imaging performance depends on the value of β. In order to illustrate this we show in Figure 8 the variation in detected signal strength as perfect fluorescent sheet scans through focus. This serves to characterize the strength of the optical sectioning in fluorescence microscopy in the same way that the mirror was used in the brightfield case. We note that the half-width of these curves is

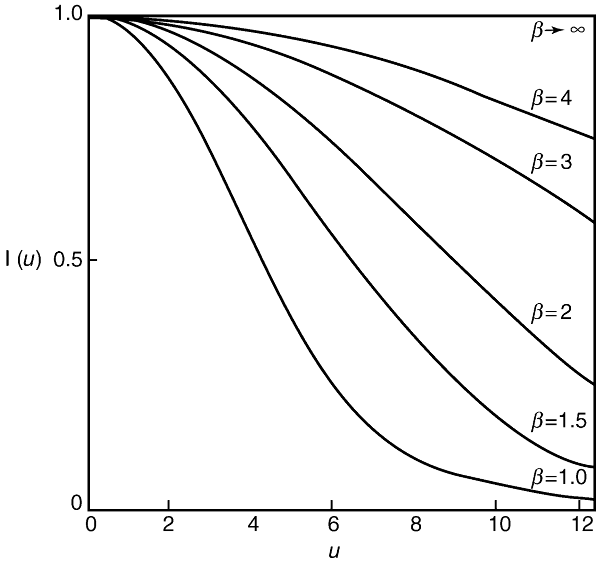

Figure 8 The detected signal as thin fluorescent sheet is scanned axially through focus for a variety of fluorescent wavelengths. We note that if we measure the sectioning by the half width of these curves, the strength of the sectioning is essentially proportional to β.

essentially proportional to β and so for optimum sectioning the wavelength ratio should be as close to unity as possible.

We have just discussed what we might call one-photon fluorescence microscopy in the sense that a fluorphore is excited by a single photon of a particular wavelength. It then returns to the ground state and emits a photon at the (slightly longer) fluorescence wavelength. It is this radiation which is detected via the confocal pinhole. Recently, however, much interest has centered on two-photon excitation fluorescence microscopy. This process relies on the simultaneous absorption of two, longer wavelength photons, following which a single fluorescence photon is emitted. The excitation wavelength is typically twice that used in the one-photon case. The beauty of the two-photon approach lies in the quadratic dependence of the fluorescence intensity on the intensity of the illumination. This leads to fluorescence emission which is always confined to the region of focus. In other words the system possesses an inherent optical sectioning property. Other benefits of two-photon fluorescence over the single photon case include the use of red or infra-red lasers to excite ultra-violet dyes, confinement of photo-bleaching to the focal region (the region of excitation), and the reduced effects of scattering and greater penetration. However, it should also be remembered that, compared with single photon excitation, the fluorescence yield of many fluorescent dyes under two-photon excitation is relatively low.

In order to make a theoretical comparison between the one- and two-photon modalities, we shall assume that the emission wavelength λ_{em} is the same, irrespective of the mode of excitation. Since the wavelength required for single-photon excitation is generally shorter than the emission wavelength, we may write it as $\gamma\lambda_{\mathrm{em}}$ where $\gamma < 1$. Since two-photon excitation requires the simultaneous absorption of two photons of half the energy we will assume that the excitation wavelength may be written as $2\gamma\lambda_{\mathrm{em}}$ which has been shown to be a reasonable approximation for many dyes.

If we now introduce optical coordinates u and v, normalized in terms of λ_{em}, we may now write the effective point spread functions in the one-photon confocal and two-photon case as

$$I_{\mathrm{1p\text{-}conf}} = I\left(\frac{v}{\gamma}, \frac{u}{\gamma}\right) I(v, u) \qquad [7]$$

and

$$I_{\mathrm{2p}} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) \qquad [8]$$

respectively. We note that although a pinhole is not usually employed in two-photon microscopy it is possible to include one if necessary. In this confocal two-photon geometry the effective point spread function becomes

$$I_{\mathrm{2p\text{-}conf}} = I^2\left(\frac{v}{2\gamma}, \frac{u}{2\gamma}\right) I(v, u) \qquad [9]$$

If we now look at eqns [7] and [8] in the $\gamma = 1$ limit, we find that

$$I_{\mathrm{1p\text{-}conf}} = I^2(v, u) \qquad [10]$$

and

$$I_{\mathrm{2p}} = I^2\left(\frac{v}{2}, \frac{u}{2}\right) \qquad [11]$$

We now see that because of the longer excitation wavelength used in two-photon microscopy the effective point spread function is twice as large as that of the one-photon confocal in both the lateral and axial directions. The situation is somewhat improved in the confocal two-photon case but it is worth remembering that the advantages of two-photon excitation microscopy are accompanied by a reduction in optical performance as compared to the single-photon case.

From the practical point of view it is worth noting that the two-photon approach has certain very important advantages over one-photon excitation, in terms of image contrast, when imaging through scattering media apart from the greater depth of penetration afforded by the longer wavelength excitation. In a single-photon confocal case it is quite possible that the desired fluorescence radiation from the focal plane may be scattered after generation in such a way that it is not detected through the confocal pinhole. Since the fluorescence is generated throughout the entire focal volume it is also possible that undesired fluorescence radiation, which was not generated within the focal region, may be scattered so as to be detected through the confocal pinhole. In either case this leads to a reduction in image contrast. The situation is, however, completely different in the two-photon case. Here the fluorescence is generated only in the focal region and not throughout the focal volume. Furthermore since all the fluorescence is detected via a large area detector – no pinhole is involved – it is not so important if further scattering events take place. This leads to high-contrast images which are less sensitive to scattering. This is particularly important for specimens which are much more scattering at the fluorescence ($\lambda/2$) wavelength than the excitation (λ) wavelength.

The Use of Structured Illumination to Achieve Optical Sectioning

An alternative approach to obtain optical sectioning is to use structured illumination in a conventional microscope. This approach is attractive since the conventional light microscope already possesses many desirable properties: real-time image capture, standard illumination, ease of alignment, etc. However, it does not produce optically sectioned images in the sense usually understood in confocal microscopy. In order to see how this deficiency may be corrected via a simple modification of the illumination system let us begin by looking at the theory of image formation in a conventional fluorescence microscope and start by asking in what way the image changes as the microscope is defocused. We know that in a confocal microscope the image signal from all object features attenuates with defocus and that this does not happen in a conventional microscope. However, when we look closely at the image formation process we find that it is only the zero spatial frequency (constant) component which does not change with defocus (Figure 9); all other spatial frequencies actually do attenuate with defocus to a greater or lesser extent. Figure 10 illustrates this by showing the image of a single spatial frequency one-dimensional bar pattern object for increasing degrees of defocus. When the specimen is imaged in focus, a good image of the bar pattern is obtained. However with increasing defocus the image becomes progressively poorer and weaker until it eventually disappears leaving a uniform gray level. This observation is the basis of a simple way to perform optical sectioning in a conventional microscope.

If we simply modify the illumination path of the microscope so as to project onto the object the image of a one-dimensional, single spatial frequency fringe pattern, then the image we see through the microscope will consist of a sharp image of those parts of the object where the fringe pattern is in focus together with an out-of-focus blurred image of the rest of the object. In order to obtain an optically sectioned image it is necessary to remove the blurred out-of-focus portion as well as the fringe pattern from the in-focus optical section. There are many ways to do this – one of the simplest involves simple processing of three images taken at three different spatial positions of the fringe pattern. The out-of-focus regions remain fairly constant between these images and the relative spatial shift of the fringe pattern allows the three images to be combined in such a way as to remove the fringes. This permits us to retrieve both an optically sectioned image as well as a conventional image in real time.

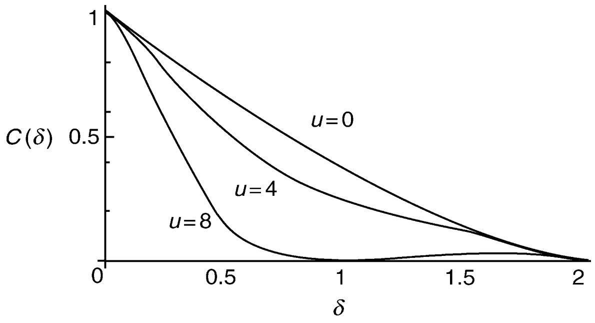

Figure 9 The transfer function for a conventional fluorescence microscope for a number of values of defocus (measured in optical units). We notice that the response for all nonzero spatial frequencies decays with defocus.

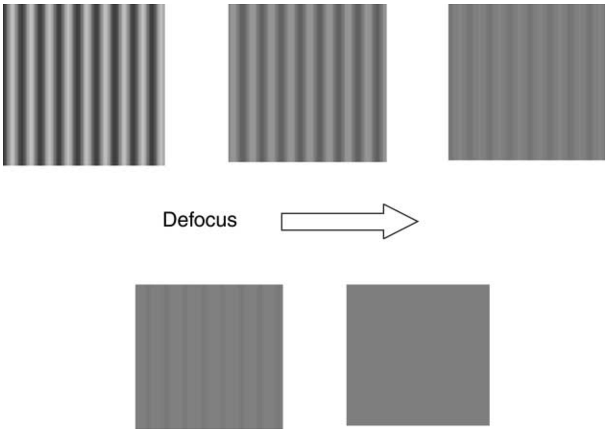

Figure 10 The image of a one-dimensional single spatial frequency bar pattern for varying degrees of defocus. We note that for sufficiently large values of defocus the bar pattern is not imaged at all.

Since the approach involves processing three conventional microscope images, the image formation is fundamentally different from that of the confocal microscope. However, the depth discrimination or optical sectioning strength is very similar and this approach, which requires very minimal modifications to the instrument, has been used to produce high-quality three-dimensional images of volume objects which are directly comparable to those obtained with confocal microscopes. A schematic of the optical arrangement is shown in Figure 11a, together with experimentally obtained axial responses in Figure 11b. It will be noted that these responses are substantially similar to those obtained in the true confocal case. As an example of the kind of images which can be obtained with this kind of microscope, we show in Figure 12 two images

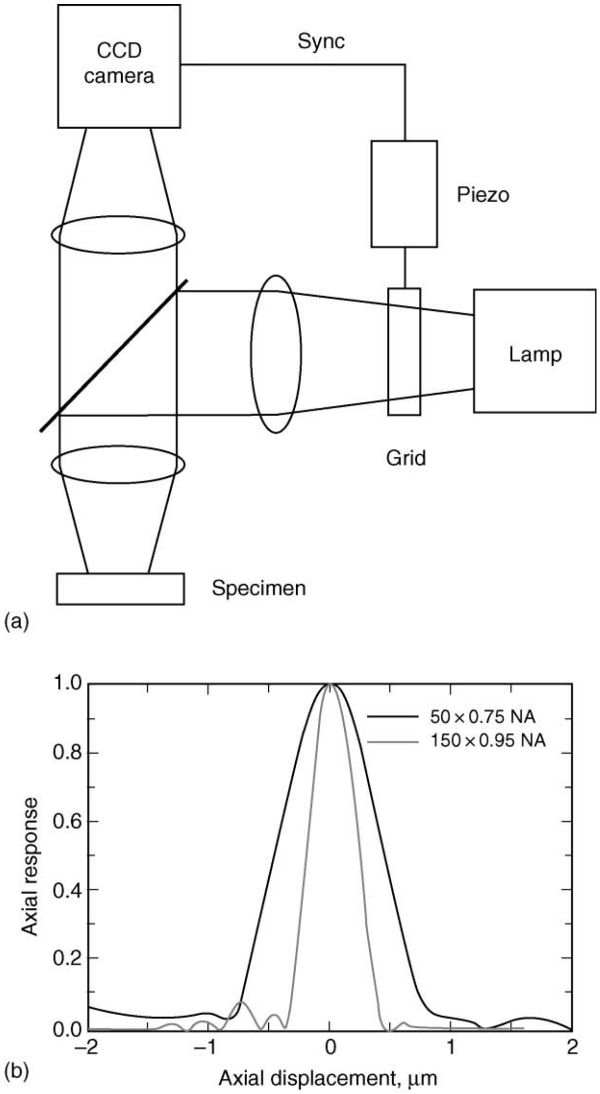

Figure 11 (a) Shows the optical system of the structured illumination microscope together with experimentally obtained axial responses in (b) which confirm the optical sectioning ability of the instrument.

of a spiracle of a head louse. The first is an auto-focus image of greatly extended depth of field constructed from a through-focus series of images. The second is a conventional image taken at mid-focus. The dramatic increase in depth of field is clear when compared with a mid-focus conventional image. We note that these images were taken using a standard microscope illuminator as light source. Indeed, the system is so light efficient that good-quality optical sections have been obtained or transistor specimens using simply a candle as light source!

Imaging using fluorescence light is also possible using this technique. However, an alternative approach that does not require a physical grid is possible if a laser is used as the light source. In this

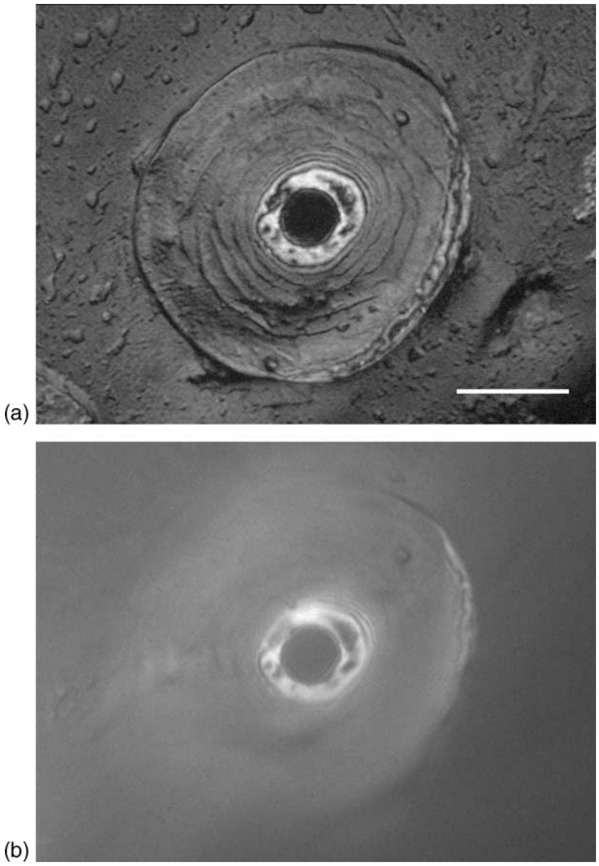

Figure 12 Two images of the region around the spiracle of a head louse. As in the previous figure, (a) is an autofocus image and (b) shows a mid-focus conventional image. The scale bar depicts a length of 10 μm.

system the laser illumination is split into two beams which are allowed to interfere at an angle in the fluorescent specimen. This has the effect of directly 'writing' a one-dimensional fringe pattern in the specimen. Spatial shift of the fringe pattern is achieved by varying the phase of one of the interfering beams. As before, three images are taken from which both the optically sectioned image and conventional image may be obtained. The beauty of this approach is that no imaging optics are required at the illuminating wavelength and system.

Summary

We have discussed the origin of the optical sectioning property in the confocal microscope in order to introduce the range of imaging modes which this unique form of microscope leads to. A range of optical architectures have also been described. By far the most universal is that shown in Figure 1 where a confocal module is integrated around a conventional optical microscope. Other, more recent, real-time implementations have also been described. A number of practical aspects of confocal microscopy have not been discussed. This is because they are readily available elsewhere, such as advice on the correct choice of detector pinhole size or because they are still the focus of active research, such as the development of new contrast mechanisms for achieving enhanced three-dimensional resolution such as the stimulated emission depletion method, STED or 4-Pi which are still under development but offer great promise.

See also

Microscopy: Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Corle TR and Kino GK (1996) *Confocal Scanning Optical Microscopy and Related Imaging Systems*. New York: Academic Press.

Denk W, Strikler JH and Webb WW (1990) Two photon laser scanning fluorescence microscopy. *Science* 248: 73.

Diaspro A (2002) *Confocal and Two-Photon Microscopy.* New York: Wiley-Liss.

Gu M (1996) *Principles of Three-Dimensional Imaging in Confocal Microscopes*. Singapore: World Scientific.

Hell S and Stelzer EHK (1992) Fundamental improvements of resolution with a 4Pi-confocal fluorescence microscope using two-photon excitation. *Optics Communication* 93: 277.

Hell S and Wichmann J (1994) Breaking the diffraction limit by stimulated emission. *Optics Letters* 19: 870.

Lukosz W (1966) Optical systems with resolving power exceeding the classical limit. *Journal of the Optical Society of America* 56: 1463.

Masters BR (1996) *Confocal Microscopy.* Bellingham: SPIE.

Minsky M (1961). Microscopy Apparatus. United States Patent 3,013,467, Dec 19, 1961 (Filed Nov 7, 1957).

Neil MAA, Juskaitis R and Wilson T (1997) Method of obtaining optical sectioning by using structured illumination in a conventional microscope. *Optics Letters* 22: 1905.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy.* New York: Plenum Press.

Wilson T (ed.) (1990) *Confocal Microscopy.* London: Academic Press.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy.* London: Academic Press.

Imaging Multiple Photon Fluorescence Microscopy

M Previte, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The field of light microscopy started with Ernst Abbe's introduction of the diffraction-limited objective lens in the late nineteenth century. Over the last 60 years, new imaging modalities, such as video, fluorescence, confocal, near field, and more recently multi-photon microscopy, have enabled the experimentalist to observe a plethora of new phenomena in the biological and physical sciences. Of these techniques, multi-photon microscopy has emerged not only as a valuable imaging technique but it also forms the technology base for emerging fields of optical memory and microfabrication.

The history of multi-photon microscopy (MPM) began with Göppert-Mayer's theoretical prediction of two-photon excitation in 1931. Although the theoretical groundwork was established, it took another 30 years for this process to be experimentally realized when Kaiser and Garrett observed two-photon absorption fluorescence emission from $CaF_2:Eu^{+2}$. Soon after this observation was made, Singh and Bradley witnessed higher-order multi-photon processes in 1964.

Historically, three-dimensional imaging was first achieved using confocal microscopy. In contrast to MPM, photo-interaction and fluorescence emission occurs throughout the excitation light path in confocal microscopy. Confocal microscopy selects the signal from the focal plane by the use of a conjugated pinhole aperture. In order to construct 3D images, successive planes must be imaged. As a result, the entire sample volume is exposed to the excitation light multiple times, and inevitably, photobleaching and photodamage occurs throughout the imaging volume. Furthermore, the pinhole aperture in the emission path causes additional signal reduction in optically thick specimens where the emission photon is scattered. Consequently, the utility of confocal microscopy is limited for imaging living biological specimens. The introduction of multi-photon microscopy, by Denk and Webb in 1990, addressed many of these limitations. By combining the methodology of multi-photon excitation with scanning microscopy, MPM offered advantages such as 3-D localized, femtoliter excitation volumes, reduced scattering and absorption of the infrared excitation light by the specimen, and a large separation between the excitation and emission wavelengths. MPM can produce excellent three-dimensional images with increased signal to noise ratios.

Theoretical Basis of Multi-Photon Microscopy

Multi-photon Excitation Selection Rules

Unlike direct ground state to excited state one-photon transitions, multi-photon transitions depend on the presence of virtual electronic intermediate states to reach an excited state (see Figure 1).

Using perturbative expansion, the probability of excitation for one-photon transitions is described by the first-order solution, and multi-photon transitions can be determined by including the higher order solutions. The interaction of an electric field vector E_λ with a molecule's dipole, $\vec{r}$, is described by the respective one-photon and two-photon transition probabilities:

$$P_{1h\nu} \approx \langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle^2$$

$$P_{2h\nu} \approx \left| \sum_m \frac{\langle \Psi_f | \vec{E}_\lambda \cdot \vec{r} | \Psi_m \rangle \langle \Psi_m | \vec{E}_\lambda \cdot \vec{r} | \Psi_i \rangle}{\varepsilon_\lambda - \varepsilon_{m,i}} \right|^2 \qquad [1]$$

where Ψ_i, Ψ_m, and Ψ_f are the initial, intermediate, and final transition states, $\vec{E}_\lambda$ is the electric field associated with the excitation light, ε_λ is the energy difference between the ground state and excited states, $\varepsilon_{m,i}$ is the energy difference between the intermediate state m and the ground state, i, and the summation is over all possible intermediate states m. Although electronic transitions, that result from the interaction of a strong electric field with a molecular dipole, are affected by vibrational mode overlap (Franck Condon factors) and anharmonic perturbative effects, the presence of intermediate transition states creates a set of general selection rules for multi-photon absorption processes.

Molecular states have eigenfunctions with a defined parity: 'gerade' describing even or symmetric states (no change in the sign of the wavefunction when the sign changes in the spatial coordinates) and 'ungerade' describing odd or antisymmetric states (the sign of the wavefunction changes when the sign changes in the spatial coordinates). Selection rules arise from the physics inherent in these even and odd eigenfunctions and the odd dipole operator that couples transitions between ground and excited states. In one-photon excitation, transitions are allowed for ground state and excited state transitions of opposite parity. In contrast, two-photon excitation occurs between ground and excited states with the same parity, by coupling through intermediate states with opposite parity. In general, ground and excited states with the same parity have a high excitation probability for an even number of photons; ground and excited states with different parity have a high transition probability for excitations involving an odd number of photons.

In the absence of major vibrational perturbations, inductive effects, or solvent relaxation, the probability of the transition depends on the magnitude of the integral overlap between the ground, intermediate, and excited states. As a result, molecule absorption cross-sections for one-photon, two-photon, three-photon, or multi-photon excitation processes can be estimated. Choosing fluorophores with higher multi-photon cross-sections allows more efficient excitation with lower excitation intensity. Finally, it should be noted that while the absorption cross-sections for one-, two-, and multi-photon excitation of a given fluorophores are different, the molecules rapidly relax to the same vibrational level in the excited electronic state. The excited state residence time (fluorescence lifetime), and the fluorescence decay processes depend only on the molecular structure and its microenvironment. Therefore, the fluorescence

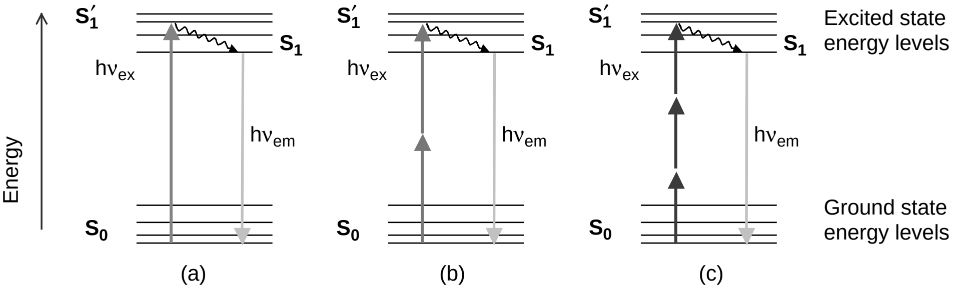

Figure 1 Jablonski diagram of a fluorescent dye molecule in one-, two- and *n*-photon excitation mode.

quantum yield and emission spectra is independent of the initial excitation process.

Enhanced Microscope Resolution Based on Multi-photon Excitation

The number of photons that are absorbed in a multi-photon process is proportional to the fluorophore concentration, $C(\vec{r}, t)$, the molecule's absorption cross-section, σ_n, which depends on n, the number of photons involved in the absorption process, and the intensity of the excitation source, such that

$$N_{\mathrm{abs}}(t) = \int \mathrm{d}\vec{r} I_0^n(\vec{r}, t) C(\vec{r}, t)\sigma_n \qquad [2]$$

When estimating $N_{\mathrm{abs}}(t)$, the spatial and temporal intensity distribution of the excitation intensity are considered independently:

$$I_0(\vec{r}, t) = S(\vec{r}) I_0(t) \qquad [3]$$

such that for an n-photon process, $I_0(t)$ is the temporal dependence of the excitation intensity at the maxima of the spatial distribution and $S(\vec{r})$ is the unitless spatial distribution of the excitation light. Combining eqns [2] and [3] gives

$$N_{\mathrm{abs}}(t) = \int \mathrm{d}\vec{r} S^n(\vec{r}) I_0^n(t) C(\vec{r}, t)\sigma_n \qquad [4]$$

If it is assumed that fluorophore distribution is uniform and the system does not undergo photo-bleaching or ground-state depletion, $C(\vec{r}, t)$ can be considered a constant and the previous equation can be rewritten as

$$N_{\mathrm{abs}}(t) = C\sigma_n I_0^n(t) \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [5]$$

By expanding the incident laser beam and overfilling the back aperture of the objective lens, the paraxial form of the normalized intensity point spread function (PSF, $h^2[u, v]$) determines the properties for a diffraction-limited lens. The dimensionless radial distance, v, from the optic axis and the distance from the in-focus plane, u, are given by

$$v = \frac{2\pi(\mathrm{NA})r}{\lambda} \qquad [6]$$

$$u = \frac{2\pi(\mathrm{NA})^2 z}{n_0\lambda} \qquad [7]$$

where $\mathrm{NA} = n_0 \sin\theta$ (θ being the half-angle of collection for the lens and n_0 is the index of refraction the medium) is the numerical aperture. Expressing the spatial distribution function with the optical units, v, u, and θ gives the following equation:

$$S(\vec{r}) = h^2[u, v] = 2\left|\int J_0(v\theta)\mathrm{e}^{-(1/2)\mathrm{i}u\theta^2}\,\theta\,\mathrm{d}\theta\right|^2 \qquad [8]$$

where J_0 is the zero-order Bessel function. If an oil immersion objective with a $\mathrm{NA} < 1.0$ is used, the paraxial approximation accurately estimates the spatial distribution, and the integral of $S(\vec{r})$ evaluates as

$$\int_{V=\infty} S^n(\vec{r})\mathrm{d}\vec{r} = a_n\left[\frac{n_0\lambda^3}{8\pi^3(\mathrm{NA})^4}\right] \qquad [9]$$

where $a_n = \int_0^\infty 2\pi v\,\mathrm{d}v \int_{-\infty}^{\infty}\mathrm{d}u[h(u, v)]^{2n}$. Since this integral cannot be solved analytically, the values of a_n are determined using numerical integrations of $h^2(u, v)$ over all space. As a result, the volume integral of the spatial distribution is always proportional to NA^{-4} and therefore independent of the number of photons involved in the excitation process. For higher NA objectives, the paraxial form of the normalized intensity PSF may no longer be an adequate approximation. The expression for temporal part of eqn [5] can be written as

$$I_0(t) = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)P(t) \qquad [10]$$

where $P(t)$ is the excitation power and λ is the center wavelength of the pulsed excitation source. The temporal dependence can be described with the time average expression:

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)^n \frac{1}{T}\int_0^T P^n(t)\mathrm{d}t \qquad [11]$$

where T is an arbitrary time interval for continuous wave lasers and $T = 1/f_\mathrm{p}$ for a pulsed laser with a repetition rate, f_p. Since multi-photon microscopy uses pulsed laser systems, the integral in eqn [11] must be evaluated accordingly. For a pulsed laser with a pulse width, τ, repetition rate, f_p, and averaged power, P_0, the pulse profile can be approximated as

$$P(t) = \frac{P_0}{f_\mathrm{p}\tau} \qquad \mathit{for}\ 0 < t < \tau \qquad [12]$$

$$P(t) = 0 \qquad \mathit{for} \qquad \tau < t < \frac{1}{f_\mathrm{p}} \qquad [13]$$

such that eqn [11] can be expressed as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi\mathrm{NA}^2}{\lambda^2}\right)^n \left(\frac{P_0}{f_\mathrm{p}\tau}\right)^n \frac{1}{T}\int_0^T \mathrm{d}t \qquad [14]$$

and evaluated as

$$\langle I_0^n(t)\rangle = \left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \qquad [15]$$

Assuming no stimulated emission and no self-quenching, the quantity of emitted photons collected per unit time, $F(t)$, is

$$F(t) = \frac{1}{n}\phi\eta N_\mathrm{abs}(t) \qquad [16]$$

where η and ϕ are the fluorescence quantum efficiency of the dye and collection efficiency of the system, respectively. Since n photons are needed for each individual excitation event, $1/n$ is a normalization factor for N_abs which describes the total number of photons absorbed. Rewriting this expression in terms of the average number of absorbed photons, $\langle N_\mathrm{abs}(t)\rangle$, gives the average fluorescence intensity:

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta\langle N_\mathrm{abs}(t)\rangle \qquad [17]$$

where $\langle N_\mathrm{abs}(t)\rangle$ is

$$\langle N_\mathrm{abs}(t)\rangle = C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [18]$$

such that the measured fluorescence intensity or time averaged photon flux is expressed as

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n\langle I_0^n(t)\rangle \int_V S^n(\vec{r})\mathrm{d}\vec{r} \qquad [19]$$

Since the time averaged fluorescence is proportional to $\langle I_0^n(t)\rangle$, eqns [9] and [15] can be substituted into eqn [19] to give a practical expression that relates average fluorescence to measurable laboratory quantities, such as P_0, τ_p, f_p, C, σ_n, n, η, and ϕ, NA, and associated numerical constants such that

$$\langle F(t)\rangle = \frac{1}{n}\phi\eta C\sigma_n\left(\frac{\pi \mathrm{NA}^2}{\lambda^2}\right)^n (P_0)^n \left(\frac{1}{f_\mathrm{p}\tau}\right)^{n-1} \times a_n\left[\frac{n_0\lambda^3}{8\pi^3(\mathrm{NA})^4}\right] \qquad [20]$$

As eqn [19] shows, the fluorescence intensity has an *n*th order dependence on the spatial and temporal distribution of the excitation power. Therefore, appreciable multi-photon probability occurs only at a region of high temporal and spatial concentration of photons. Confocal pinholes can be eliminated making this technique better suited for 3D biological imaging.

The higher order (>2) dependence of the fluorescence signal upon the excitation intensity distribution results in axial confinement. Experimentally, it can be shown that over 80% of fluorescence signal emanates from a 1 μm thick region about the focal point for objectives with a numerical aperture of 1.25 under two-photon excitation. This confinement is significantly different from one-photon excitation where a spatially uniform fluorescent sample contributes equal fluorescence intensities from each *z*-section above and below the focal plane (assuming negligible excitation attenuation). For an excitation wavelength of 960 nm, the typical point spread function of two-photon absorption has FWHM of 0.3 μm in the radial direction and 0.9 μm in the axial direction. In comparison with confocal microscopy, the localization of the volume of excitation instead of the volume of observation, as in confocal microscope, allows 3D imaging with minimal photobleaching and damaging.

Basic Imaging Applications of Multi-Photon Microscopy

Multi-photon microscopy finds an increasingly broad range of applications, especially in biological and medical areas. For example, multi-photon microscopy has been utilized to study the physiological and biological states in various highly scattering tissues such as cornea, skin, and pancreas islet. Also, multi-photon microscopy is used in embryology to study the development of sea urchin, *Caenorhabditis elegans*, hamster, and zebra fish. Neurobiological applications include qualifying calcium dynamics dendritic spines, memory and structural plasticity, hemodynamics, and plaque evolution in Alzheimer's disease. MPM has been applied to study pharmacological distribution in tissues and in the quantification of transdermal drug delivery. Due to minimally invasive capability of multi-photon microscopy, intra-vital microscopy imaging into thick tissues of living animals is possible leading to applications in the study of cancer biology.

Although the advantages of MPM are less compelling in studying cell cultures where the specimen is optically thin, a number of uniquely exciting uses have been identified. For instance, the significantly broader two-photon absorption spectra of many fluorophores allow convenient imaging of specimens labeled with multiple color fluorophores. The localized excitation volume reduces photo-interaction with fluorophores outside the excitation volume allowing better 3D images to be obtained from cellular systems labeled with easily photobleachable fluorophores. The use of three and higher photon excitation has also allowed the excitation and imaging of important biomolecules with deep ultraviolet absorption. Another interesting application of multi-photon excitation in cellular systems is based on the ability of this method to produce localized chemical reactions. Through the photoactivation or photolysis of 'caged' molecules, their chemical structures are altered to produce biologically active molecules, fluorescent species, or other observable chemical changes. Multi-photon excitation has been used to locally uncage neurotransmitters for mapping ion channels in neurons and to uncage calcium in studying signal transduction events. In cellular systems, MPM is being used to image normally nonfluorescent biomolecules, such as serotonin in neurons, based on multi-photon induced chemical conversion and multi-photon excitation. Finally, multi-photon microscopy has also found applications in studying the conformation, aggregation, diffusion, and transport of single molecules and proteins. Two-photon fluorescence correlation spectroscopy and photon counting histograms are powerful techniques that have been applied to study protein aggregation, conformation changes, and diffusion *in vitro* with applications ranging from cell biology to high throughput drug discovery. The large separation of excitation and emission spectra of two-photon excitation also allows high sensitivity detection of fluorescence from single chromophore; however, the higher photobleaching rate at the excitation volume for some chromophores may lessen or even reverse this advantage.

Advanced Multi-Photon Microscopy Techniques

As described in the previous sections, excitation processes involving multiple numbers of photons offers considerable advantages for the imaging process in fluorescent microscopy. 4Pi Confocal Microscopy and high-speed Multifocal Multi-photon Microscopy MMM are advanced multi-photon microscopy techniques that capitalize on these advantages to improve image resolution and acquisition times in fluorescence imaging.

4Pi Confocal Microscopy

In contrast to near-field microscopy, far-field light microscopes, i.e. confocal and multi-photon microscopes, can create three-dimensional images of the interior of cells. The lateral and axial resolution of these systems is limited between 180 nm and 300 nm and 500–1000 nm, respectively. This limitation is dictated by the Abbe theory where resolution is inversely proportional to the numerical aperture of the focusing lens. When focusing with a single lens, geometric constraints limit the numerical aperture to about 1. 4Pi microscopy was developed to overcome this limitation. By utilizing two opposing objective lenses whose spherical wavefronts are coherently combined, a greater total effective numerical aperture is achieved. This results in a higher axial resolution, while the lateral resolution of the 4Pi microscope remains unchanged with respect to its confocal counterpart.

In the setup of a 4Pi microscope, the illuminating wavefronts emanate from a common point light source. The path is split by a beamsplitter, guided through two optically identical paths, and through two opposing high NA lenses that share a common focus. The fluorescence light is then collected through one lens and a confocal pinhole. In the axial direction, this arrangement yields a PSF with one main maximum lobe that is 3–4 times sharper along the optic axis than the PSF of a conventional confocal microscope, and two undesired side-lobes.

These side lobes can produce artifacts in the image and their effect needs to be suppressed for unambiguous 3D imaging. Lobe suppression can be achieved in a 4Pi system using two-photon excitation resulting in a quadratic decrease in the ratio of the side-lobe and the main lobe amplitudes. Using an excitation wavelength of 800 nm and 1.4 NA oil-immersion objectives, a FWHM along the axial direction of the main lobe is about 145 nm and side lobes are approximately 25–35% of the maximum lobe intensity. These side lobes can be further mathematically reduced by means of a linear filter after (or during) image acquisition. The filter can be applied at lobe heights below 50%, and an effective 3D PSF consisting of a single spot with a 250 nm lateral and 145 nm axial FWHM is achieved. When combined with nonlinear image restoration, 4Pi confocal microscopy achieves a 3D resolution in the 100 nm range. This resolution is similar to that obtained with near-field optical microscopy in biological applications with an added advantage of 3D imaging of the cell interior.

Since resolution improvement in 4Pi microscopy can be negated by the presence of aberrations, the usage of oil immersion lenses requires the specimen to be fixed and mounted in glycerol to minimize aberrations resulted from index mismatch. The sample thickness is further limited in the range of 20–30 μm. Due to significant index mismatches between oil and water, the imaging of live cells in aqueous medium is extremely difficult, if not impossible. In order to circumvent this difficulty, water immersion objective lenses are introduced into the 4Pi system. These lenses have a lower aperture angle $\alpha = 62°$ (NA = 1.2) than oil immersion lenses

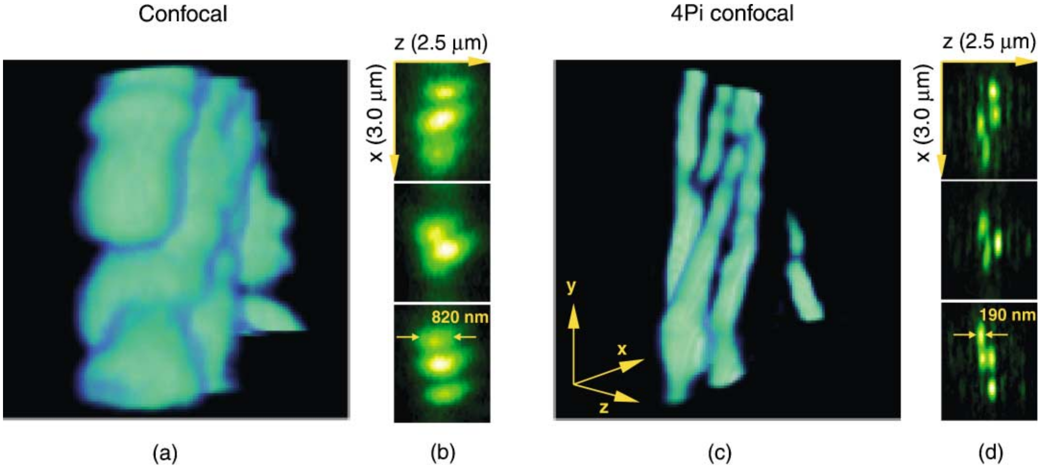

Figure 2 (a) Two-photon confocal and (c) 4Pi confocal point deconvolved images of an immunolabelled microtubule network in fixed NIH3T3 cells. The 3D rendered images in (a) and (c) were generated out of a large XZ-image-stack, partly shown in (b) and (d). As indicated in the XZ-images, the axial resolution is 820 nm and 190 nm for the two-photon confocal and the two-photon 4Pi confocal images, respectively. The images were taken with water immersion lenses (NA = 1.2). Reproduced from Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164, with permission from Elsevier.

$\alpha = 68°$ (NA = 1.4) and therefore are expected to produce higher side lobes, which are reduced by 35% with careful aberration studies such that linear filtering is possible. Nonetheless, the use of water-immersion lenses solves two major problems. First, they extend the possible sample thickness to the working distance of the objective lens (~200 μm); second they allow the imaging of live cells in aqueous medium. As a result, microtubule networks (see Figure 2) and live bacteria can be imaged with a resolution of 280 nm in lateral and 190 nm in axial direction. The water immersion 4Pi microscope, utilizing two-photon excitation, is capable of the highest 3D resolution so far reported in live cell imaging to date.

Multifocal Multi-Photon Microscopy (MMM)

Improving imaging speed is critical for applying 3D microscopy in biological specimens. An MMM scans an object with an array of high-aperture foci whereas a conventional multi-photon microscope scans with a single beam. An MMM combines nonlinear fluorescence excitation with parallelized 3D imaging. The technical design consists of a densely packed microlens array; each microlens transmitting a collimated beam at a slightly different angle, overfilling the back aperture of the objective. Consequently, from the arrangement of the microlenses, an array of foci is reproduced in the image plane. The foci are scanned rapidly across the image plane to ensure full coverage of the focal plane and real-time imaging (a stack of 50 images of 30–70 μm diameter can be acquired within a few seconds). The foci's movements are either generated by rotating a microlens array that is arranged in a Lasserie pattern or by rectilinear scanning of a square microlens matrix.

MMM development was initiated because nonlinear optical damaging processes limit the usable focal peak intensity to about 200 GW/cm^2, which can be reached by a tightly focused 7–15 mW femtosecond laser beam. This power consumption makes use of only a fraction of the laser output of a typical mode-locked titanium-sapphire laser light source. The MMM better exploits the 1–2 W of the beam power by dividing the power among 25–40 foci and scanning each in parallel. With this method, typical acquisition times are cut from ~1000 ms to 10–50 ms. MMM provides an excellent method for real time 3D multi-photon imaging.

A fundamental problem in parallelized 3D microscopy is that increasing density of the foci over a certain level decreases the axial resolution. The reduction of the distance between the foci increases not only the speed or image brightness, but it also increases the interference between neighboring focal fields; especially in planes shortly before and after the focal plane. For neighboring foci in high NA objective lenses, interference effects can be significant. The relatively short laser pulses that are commonly used in multi-photon excitation provide a means to reduce this interference by using time multiplexing. Experimentally, the femtoseconds pulses of the Ti:sapphire laser, going through each lenslet, can be scheduled to arrive at the focal plane at different times by a mask of glass in front of each microlens, varying the optical path lengths of these beams. The different arrival time eliminates interference effects. The resolution of the parallelized system remains unchanged as compared with the conventional one foci system. In conclusion, the MMM enables increased image acquisition speed by a factor of the number of foci in the optical plane without compromising image resolution.

Advanced Multi-Photon Microscopy Applications

As described by the previous section, MPM technology is undergoing rapidly development with improved resolution and increased instrument throughput. Recently, MPM has also found applications in exciting new areas of optical memory and microfabrication.

Multi-Photon 3D Optical Memory

Researchers in the field of optical memory are challenged to develop instrumentation that records three-dimensional, high density, robust bitwise information in low-cost and readily available storage materials. Initially, early optical memory techniques relied on higher energy pulsed recording beams and red shifted reading beams. Aided by the development of the Ti:sapphire laser and high numerical aperture objectives that can confine laser light to a volume of the order, λ^3, the power of the focused femtosecond pulses at the focal point can reach over 10^{18} W/cm^2 with electrical field strengths reaching 2.7×10^{10} V/cm. These fields are substantial enough to locally change the optical properties of selected materials. Thus, multi-photon processes were intertwined with the field of optical memory. Recent experiments showed that with properly chosen materials and instrumentation, high density, three-dimensional data storage was possible.

Optical memory is based on photochromic material whose optical properties are changed upon irradiation such that initial and final chemical states have different absorption spectra or indices of refraction. In 1989, Parthenopoulos and Rentzepis first showed that 3D information could be stored in a polymer film doped with photochromic molecules. Although the recorded information is lost when the photo-excited compound spontaneously reverts back to the original nonfluorescent compound, Dvornikov and Rentzepis later improved the material such that the stored information was stably recorded at room temperature.

For photochromic materials, the recording and reading mechanisms often required multiple laser sources and complex optical setups. Kawata *et al.* addressed these issues with the development of an alternative optical method that used photorefractive crystals as the storage media, a Ti:sapphire as the recording device, and a phase-contrast microscope configuration for data readout of bits separated by 20 μm. Using this method, Kawata avoided the complication of different recording and read-out laser sources, thereby simplifying the instrumentation. However, in photorefractive material, poor refractive index matching, between the storage medium and the objective, results in aberration and degrades storage densities. Recent advances in objective lens design have attempted to address these limitations. Kawata and Kawata describe many of these optical memory devices in a recent review. A recent publication by Olson *et al.* describes a method for multi-photon data storage in molecular glasses and highly cross-linked polymers. These materials prove to be highly efficient and robust for long-term data storage.

Although the field of optical memory has made great progress since 1989, further advances are required before this field becomes a widely adopted technology. For example, simpler recording and reading techniques are required to insure a cost-effective and reliable data storage technology. Consequently, further progress in the production of compact ultrafast lasers and high-speed scanning instrumentation will be needed before broader acceptance of optical memory storage devices.

Multi-Photon 3D Microfabrication

Another promising emerging application of multi-photon excitation is 3D microfabrication. 3D microfabrication processes can be performed with two-photon absorption. Maruo *et al.* and Sun *et al.* provide examples of microfabricated structures generated with femtosecond pulses to induce photopolymerization. Much like the optical memory techniques, this methodology relies on photoinitiators with large two-photon cross-sections to initiate polymerization of surrounding monomer or polymer molecules. Kawata *et al.* and Sun *et al.* showed that two-photon photopolymerization could be used successfully to create intricate three-dimensional microstructures. These systems offer lower polymerization thresholds, higher polymerization rates, and wider dynamic write ranges such that rapid fabrication and low-power writing are possible. Recently, Cumpston *et al.* described a class of π-conjugated compounds with large two-photon cross-sections, which can be effectively used for two-photon polymerization initiators in three-dimensional optical data storage and microfabrication processes.

As with most of these processes, the techniques have proven to be successful, but the mechanisms behind the multi-photon initiated chemical processes are not completely understood. With further work to optimize the reaction parameters, it is possible to design experiments that will allow new materials to be synthesized or patterned based on processes such as dielectric breakdown, microexplosions, optical damage, and pyrolytic effects.

Conclusion

Relying on femtoliter excitation volumes and large separations between excitation and emission wavelengths, multi-photon microscopy has emerged as a powerful technology that allows for increased resolution and enhanced signal-to-noise ratios. It has also enabled scientists to avoid complications in imaging biological systems due to decreased sample photodamage and in some instances, photobleaching of fluorophores. Advancements in multi-photon imaging applications, such as 4Pi Confocal microscopy and multifoci multi-photon microscopy, have further enhanced the resolution of imaging technology and improved the speed of data acquistion respectively. Furthermore, the application of multi-photon technology in the fields of microfabrication and optical memory promises to address some of the difficulties involved in designing three-dimensional high-density storage devices and microfabricated components. Overall, multi-photon microscopy provides the scientific community with an invaluable tool in the study and design of biological and physical systems.

See also

Microscopy: Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Bahlmann K, Jakobs S, *et al.* (2001) 4Pi-confocal microscopy of live cells. *Ultramicroscopy* 87(3): 155–164.

Bewersdorf J, Pick R, *et al.* (1998) Multifocal multiphoton microscopy. *Optical Letters* 23: 655.

Denk W, Strickler JH, *et al.* (1990) Two-photon laser scanning fluorescence microscopy. *Science* 248(4951): 73–76.

Kawata S and Kawata Y (2000) Three-dimensional optical data storage using photochromic materials. *Chem. Rev.* 100(5): 1777–1788.

Olson CE, Previte MJR, *et al.* (2002) Efficient and robust multiphoton data storage in molecular glasses and highly crosslinked polymers. *Nature Materials* 1(4): 225–228.

Pawley JB (ed.) (1995) *Handbook of Biological Confocal Microscopy.* New York: Plenum Press.

So PT, Dong CY, *et al.* (2000) Two-photon excitation fluorescence microscopy. *Annu. Rev. Biomed. Eng.* 2: 399–429.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy.* New York: Plenum Press.

Xu C, Zipfel W, *et al.* (1996) Multiphoton fluorescence excitation: new spectral windows for biological nonlinear microscopy. *Proceedings of the National Academy of Science of the USA* 93(20): 10763–10768.

Interference Microscopy

E Novak, Veeco Instruments, Inc., Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Interference microscopy is a powerful noncontact metrology technique for reflection and transmission characterization of samples. The field can be divided roughly into two main categories, polarization-splitting interferometers and amplitude-splitting interferometers. In both cases, the light incident on a test sample is first divided into two beams and later recombined to form an interference pattern. This interference pattern is analyzed by one of a variety of methods to determine the difference in the optical paths the two beams traversed. This optical path difference may be used to quantify surface shape, refractive index variations, or thickness variations of the sample to a very high accuracy. This article describes the background of the two categories of interference microscopy, specifics of their implementations for different types of measurements, current state of the technologies, and examples of different measurements performed using these techniques.

Background

Interference fringes were first observed in a microscope by Robert Hooke in 1665. In the late 1800s, several practical versions of interference microscopes entered use to examine transparent specimens, though none gained widespread commercial success. Then in the 1950s, differential interference contrast (DIC) microscopy was invented to increase image contrast when examining semi-transparent samples under a microscope, particularly biological specimens. As this technique does not require staining of samples, and avoids the halo effect common in phase contrast microscopy (*see* **Microscopy:** Phase Contrast Microscopy), such systems became widely used in a variety of biological applications. In the early 1960s, Krug, Rienitz and Shulz published a comprehensive book to summarize all of the various interference microscope designs that had been produced thus far, showing the large depth and breath of research to that point.

With the advent of affordable computers and CCD cameras in the 1970s and 1980s, various commercial systems were developed using a variety of interferometric configurations to quantify the measured interference patterns from small samples. Due to their noncontact nature and ability to provide area measurements, as opposed to the two-dimensional traces produced by contact stylus techniques, such systems gained rapid acceptance in a variety of industries. Performance has steadily improved such that vertical resolution is sub-nm and accuracy on the nm-level is readily achievable. Lateral resolution is limited only by the optics of the microscope, with high numerical aperture objectives achieving resolution below 0.5 μm. Today, commercial systems are used in such measurements as index variations of biological samples, roughness of paper, microfluid channel volume, MEMS properties, hard drive head dimensions, drug coating thickness, and machine part characterization.

Polarization-Division Interference Microscopes

While the concepts behind polarization division interference microscopes have been known for well over a century, they did not gain widespread use until after Georges Nomarski invented the DIC microscope, also referred to as a polarization interference contrast microscope, in 1955. A schematic of a transmission-based and reflection-based Nomarski setup is shown in Figure 1, where the displacement of the beams has been greatly exaggerated for clarity.

The illumination in Nomarski systems is generally from an unpolarized broadband source, such as a

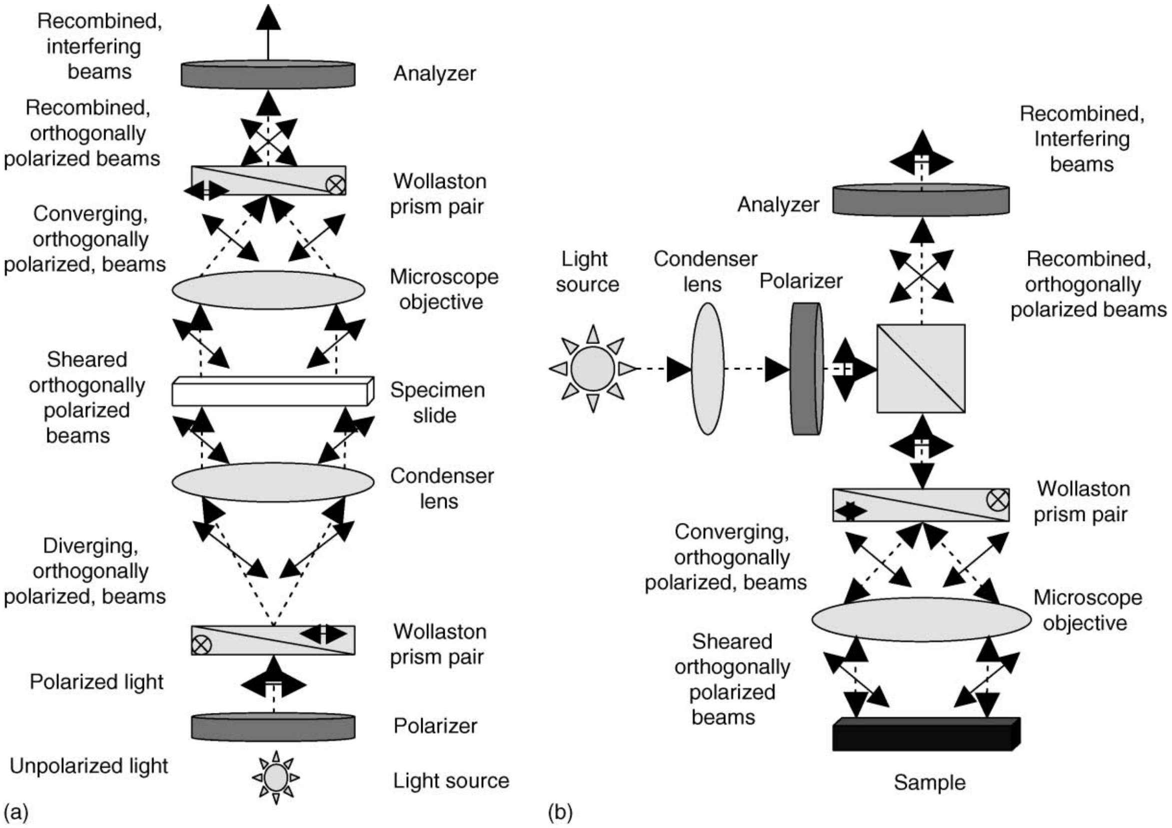
Recombined, interfering beams
Analyzer
Recombined, orthogonally polarized beams
Wollaston prism pair
Converging, orthogonally polarized, beams
Microscope objective
Sheared orthogonally polarized beams
Specimen slide
Condenser lens
Diverging, orthogonally polarized, beams
Wollaston prism pair
Polarized light
Polarizer
Unpolarized light
Light source
(a)
(b)
Recombined, Interfering beams
Analyzer
Light source
Condenser lens
Polarizer
Recombined, orthogonally polarized beams
Wollaston prism pair
Converging, orthogonally polarized, beams
Microscope objective
Sheared orthogonally polarized beams
Sample

Figure 1 (a) Schematic of transmission-based Nomarski interferometer and (b) reflection-based Nomarski interferometer.

tungsten–halogen bulb. A polarizer in the illumination path is then used to set the polarization state of the light incident on a modified Wollaston prism pair. The modified Wollaston prism consists of two birefringent wedge prisms which divide the beam in two parts, each with orthogonal polarization states. The two beams travel in slightly different directions from one another. The condenser lens brings the light to illuminate the sample, with each of the two orthogonally polarized beams slightly sheared from one another. This shear is typically much less than the resolution limit of the microscope objective, typically from tens of nanometers to several hundred nanometers, such that only one image is seen by the observer. After passing through or reflecting from the sample and the imaging microscope objective, a Wollaston prism pair is used to recombine the two polarized beams. Finally, an analyzer is used to bring both beams to a common polarization state, allowing interference between the two beams.

As the system is very sensitive to the position and optical properties of the prisms, different Wollaston prism pairs are used for each objective, usually rotating into the path automatically on a turret as the objectives are changed. In a reflection system, this prism also recombines the displaced beams after they reflect from the test piece; a transmission system will usually have only one Wollaston prism pair after the sample that can handle all objectives, instead of requiring a different recombining prism for each objective.

The observed interference pattern is an image of the gradient in the sample in the direction of shear created by the Wollaston prism pairs. In order to fully characterize the surface, the sample or optics must be rotated to capture shear images in both orthogonal directions. Essentially the technique acts as a high-pass filter, emphasizing edges and lines on the test piece.

With a white-light source, dispersion effects will cause one side of a feature to appear brighter or in a specific color range whereas the other side appears darker or in different colors. This gradient effect confers a pseudo three-dimensional appearance to the specimen. The images are quite colorful, with different colors shearing differently; by adjusting the Nomarski prisms along the optical axis, one can vary the retardation between the two beams, producing more or less contrast or color variation in the observed images.

The Nomarski system has several appealing qualities for surface and index characterization over other interferometric and polarization-based techniques. First, it is a common-path interferometer; with the exception of the slight shear, both beams pass through exactly the same optics. This means that these interferometers are mostly insensitive to vibration, as both interfering beams will be affected in the same manner. In addition, the common path nature of the system reduces the quality requirements on the optics, improving measurement quality and reducing overall system cost. In addition, Nomarski techniques do not suffer the halos associated with phase contrast imaging techniques (*see* **Microscopy:** Phase Contrast Microscopy) and provide better out of focus rejection and higher light levels since they can use the entire aperture of the microscope objective.

For reflection-based systems, the observed fringes relate to the surface slope along the direction of shear; for transmission systems, index of refraction gradients or thickness variations are observed, producing much higher contrast images of biological samples than conventional bright-field techniques. The good out-of-focus rejection mentioned earlier makes possible more accurate optical sectioning of samples, where each vertical layer of the specimen is in focus while all others are out of focus. These features, combined with the fact that high contrast images can be obtained without chemical staining, have made these systems widely used in biology. In addition, the high sensitivity to slope has made DIC systems commonly used in semiconductor inspection applications and other precision surface characterization and defect inspection applications, where automatic image processing is combined with the highly sensitive measurements to quantify both surface and subsurface defects. For instance, Figure 2 shows a bright-field image of a laser diode facet as

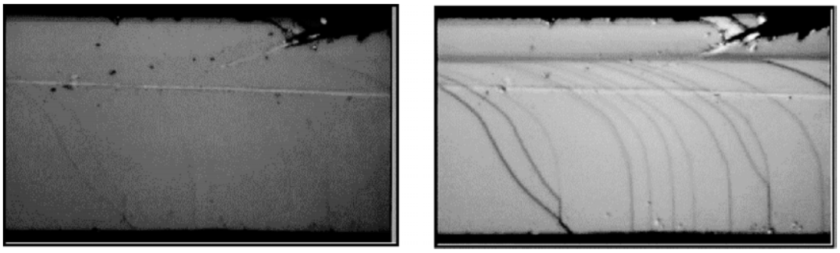

Figure 2 Bright field (left) and Nomarski (right) images of laser diode facet defects. Courtesy of Royce Instruments, Inc., Napa, CA.

well as a Nomarski image, with the facet defects greatly emphasized by the Nomarski optics.

Initial systems using differential interference contrast were strictly qualitative, with the user manually rotating the incoming polarization to produce the best contrast for a given sample and visually evaluating the resulting image. As mentioned above, visual inspection of biological specimens and defect review remain primary applications of these devices, though quantitative. By translating the Wollaston prisms laterally or rotating the analyzer in the system, one can phase shift the interference pattern. Storing multiple phase shifted images on a computer, one can obtain quantitative data about the shear using standard phase shifting interferometry equations. In addition to phase shifting through translation of the prism, quantitative data may be performed through analysis of the colors of the observed fringes or through geometric phase shifting techniques. More details on these methods can be found in the articles on **Interferometry:** Phase Measurement Interferometry; and **Interferometry:** White Light Interferometry.

Once quantitative information about the shear between the beams is obtained, the resulting data may be integrated to get back the original surface or index profile. Performing this operation on data with shear in two orthogonal directions allows the complete surface profile to be obtained. Unfortunately, since Nomarski setups are typically employed as add-on modules to standard microscope platforms rather than custom-designed metrology instrumentation, such quantitative phase shifting capability has not been widely commercialized; only a few custom or laboratory systems are in use when compared with white-light optical profilers described below. However, many research groups report nm-level surface characterization, and have developed a variety of techniques to accurately quantify thickness and index changes. The primary limitation of the quantitative techniques comes from the depth of focus of the microscope objective, which limits the vertical heights, thickness range, and index gradients that can be measured in a single measurement. Phase shifting at multiple vertical locations can be performed, but unless greatly increases the complexity of the measurement and reconstruction algorithms.

Amplitude-Division Interference Microscopes

White Light Optical Profilers

Probably the most common amplitude division interferometric microscopes are white-light interferometers, also often called optical profilers. These systems consist of three main elements: the base microscope, with illuminator and imaging optics; an interferometric element to divide and recombine the wavefront, and a scanning mechanism by which the optical path between the test and reference beams may be varied in a known manner. As the name implies, these systems typically employ spectrally broad sources, generally from a tungsten–halogen bulb. Unlike Nomarski interferometers, white light interferometers are used only in reflection-mode, as their noncommon-path nature and low-coherence source preclude interference from thick samples measured in transmission.

Figure 3 shows a schematic of a typical white light interferometer. Light from the illuminator arm, typically in a Kohler configuration, is coupled into the system with a first beamsplitter and passes through a second beamsplitter which is typically inside the microscope objective. This beamsplitter divides the beam into two parts, one of which travels to the sample and reflects from it, while the other reflects from a very high-quality reference surface. The reflected light is then recombined by the same beamsplitter and the resulting interfering light is imaged onto a camera. The static interferogram thus produced is usually by changing the optical path difference between the test and reference beams through translation of the objective with respect to the sample. The resulting signal is analyzed by one of a variety of methods as described below.

White light interferometers generally have two distinct modes of operation. In the first mode, generally referred to as phase shifting mode, a filter

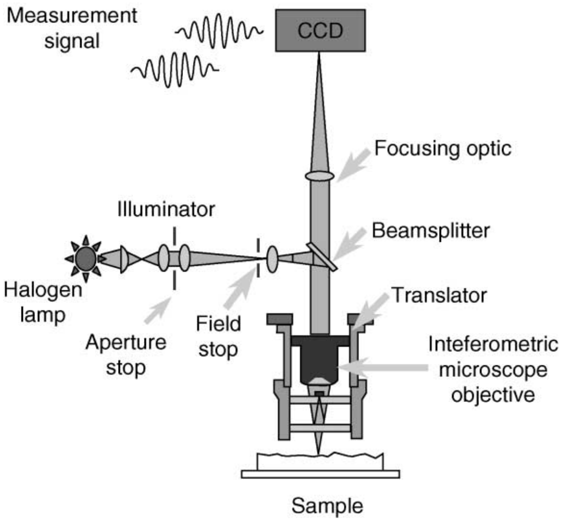

Figure 3 Schematic of white light optical profiler showing the illuminator, and translating interferometric objective.

is used to produce quasi-monochromatic light for the measurement process. This produces a large number of high-contrast interference fringes that obey the classical interference formula:

$$I_i(x,y) = I_t(x,y) + I_r(x,y) + 2\sqrt{I_t I_r}(\cos(\phi(x,y) + \alpha_i) \qquad [1]$$

In the above $I_i(x,y)$ is the irradiance observed at any one point of the image, $I_t(x,y)$ and $I_r(x,y)$ are the irradiances of the test and object beams, $\phi(x,y)$ is the optical path difference between the two beams, and α_i is a phase offset introduced by the interferometric instrument. By varying α_i in a known manner and collecting at least three irradiance patterns, any of the three unknowns in the above equation can be determined; specifically, the optical path difference may be solved for every point. Assuming measurements are in a medium of known and constant refractive index, such that the only contributions to the OPD are height variations rather than index variations, surface height at each point is simply:

$$H(x,y) = \frac{\lambda}{4\pi}\phi(x,y) \qquad [2]$$

Phase shifting techniques produce sub-nm repeatability and very high accuracy, but suffer from a 2π phase ambiguity from the calculation of phase from irradiance data. Thus, this technique cannot handle surfaces with point-to-point steps greater than one quarter of the wavelength of the light used. Calibration methods and further information may be found in the article on **Interferometry:** Phase Measurement Interferometry.

In order to extend interference microscopy to cover large steps, steep slopes, and vertical ranges greater than the depth-of-focus of the microscope objectives, a second technique of fringe analysis may be employed, that of coherence sensing. In this method an incoherent light source is used, such as a tungsten–halogen bulb. Due to the wide spectral range of the source, as interference patterns from different wavelengths add in the observation plane, the observed interference fringes will rapidly wash out for nonzero optical path difference between the two beams. See the article on **Interferometry:** White Light Interferometry, for further details and illustrations of this effect.

With the localization of fringes only at the point of zero optical path, the instrument scans the interference objective over the desired range and then calculates the position at which maximum fringe contrast occurs for each point. Calculating this position across all points within the field creates a height map of the surface; the interference fringes serve as a tool for depth discrimination in much the same way that the pinhole in a confocal microscope does (*see* **Microscopy:** Confocal Microscopy), another technique finding wide use in surface profiling. No phase unwrapping is required, with each pixel height calculated independently of all others, avoiding ambiguities associated with phase shifting methods. The vertical range is limited only by the scan range of the measurement and the working distance of the microscope objectives; current commercial instruments can achieve vertical ranges of over 8 mm. Figure 4 shows the fringes one would observe at a given scan location during measurement of a test piece with three steps. As the OPD between the test and reference arms varies, the fringes will

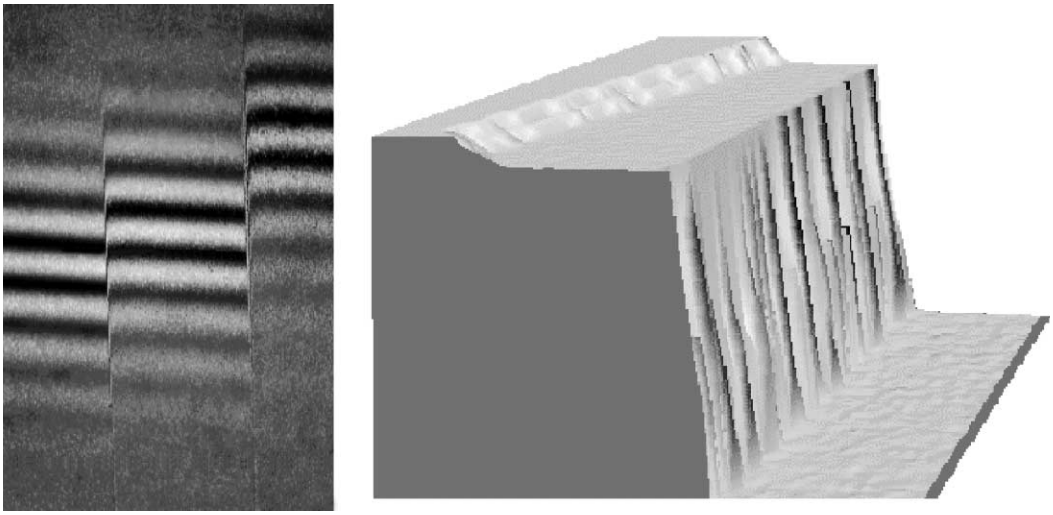

Figure 4 Localized fringes (left) of a 3-tiered surface structure (right). As the OPD is varied, the fringes will scan through the field, with the point of maximum contrast corresponding to the relative surface height of each pixel.

translate across the steps, and the surface plot at the right can be generated by calculating the scanner location of the peak contrast condition for each pixel.

There are a variety of methods for calculating the position of maximum fringe contrast using the coherence-sensing method. These include center-of-mass calculations of the peak fringe position, various frequency-space techniques and correlation of the measured fringe pattern with a theoretical pattern. Calibration of these systems generally involves repeated measurement of a known step height standard and varying the scanner motion until the step height is accurately calculated, though some systems employ separate distance-measuring interferometers to perform on-the-fly calibration for each scan. Generally speaking, coherence sensing techniques have repeatability in the range of a few nm, as opposed to the sub-nm repeatability of phase shifting techniques. This reduced sensitivity is the trade-off for avoiding the height and slope limitations of phase-shifting techniques. More details can again be found in the article on **Interferometry:** White Light Interferometry.

Recent advances in computing power have enabled an additional method of calculating surface height by combining coherence sensing with phase shifting. Coherence sensing methods are used to determine the point of best fringe contrast, and uses this information combined with the phase calculations to overcome ambiguities caused by large steps or steep slopes. Thus, one can obtain the sub-nm repeatability of phase-shifting methods without limitations on the vertical range or slope of the sample to be measured.

New fringe analysis techniques are extending the utility of these systems even further. Strobed sources are now available which effectively freeze the periodic motion for the instrument; this allows dynamic properties of samples to be measured, from surface deformation under motion to resonance frequency and switching time calculations. By analyzing the multiple interference patterns generated by transparent films deposited on a surface, film thickness, homogeneity, and differences between substrate and film surface properties may be determined.

Interferometric Microscope Objectives

One of the most critical elements in white light interferometers is the interferometric objective which splits and recombines the wavefront to form an interference pattern. Three primary interferometric configurations are used to divide and recombine the wavefront: Michelson, Mirau, and Linnik. Each of the three types of objectives is described below.

For low-magnification objectives, typically 5X or below, a Michelson interferometer configuration is typically used. A schematic of such an objective is shown in Figure 4. The illuminating beam passes through the objective and is divided by a beamsplitter placed just underneath the objective. A portion of the light travels to the test sample and the rest to a high-quality reference surface. The light reflected from each beam path is recombined by the beamsplitter and propagates to the observation plane. Note that it is important that the distance between the beamsplitter and reference surface match the distance between the beamsplitter and focus position of the sample. If the reference surface is defocused, fringe contrast will be reduced, aberrations increased, and overall system accuracy and repeatability will suffer.

This objective configuration has several advantages. First, one can add a Michelson housing to almost any standard objective with sufficient working distance for the beamsplitter, making these objectives easy to produce and relatively inexpensive. Second, the reference surface can be easily changed by the user to closely match the reflectivity of the test piece, ensuring maximum fringe contrast in the interfered beam (*see* **Interferometry:** Phase Measurement Interferometry). Third, provided there is distance between the beamsplitter and reference, one can insert a dispersion-compensating plate into the reference arm to allow measurement of parts beneath a cover glass or other dispersive medium; without such compensation, fringe contrast is generally insufficient for accurate measurements for transmissive media beyond a few hundred micrometers thickness.

As one moves towards higher magnifications, however, the working distance of the objective is typically insufficient to accommodate a large beamsplitter and still leave a reasonable distance between the objective and sample. In addition, a large cube beamsplitter will generally introduce unacceptably high aberrations in these higher numerical aperture objectives. For this reason, objectives from 10X magnification to 50X magnification are typically of the Mirau configuration, shown schematically in Figure 5.

Mirau objectives contain two additional plates compared to a standard bright-field objective: one is coated and acts as a beamsplitter, splitting the incident beam, while a second holds a small, high-quality reference surface and acts as a compensating plate so that optical paths in both arms of the interferometer are balanced (Figure 6). Such objectives are more expensive than Michelson designs, but maintain long working distances even at higher magnifications. Thermal stability becomes

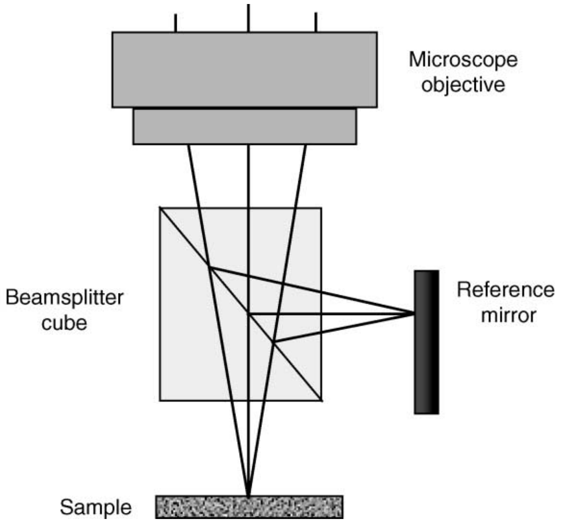

Figure 5 Schematic of a Michelson interference objective used for low-magnification applications.

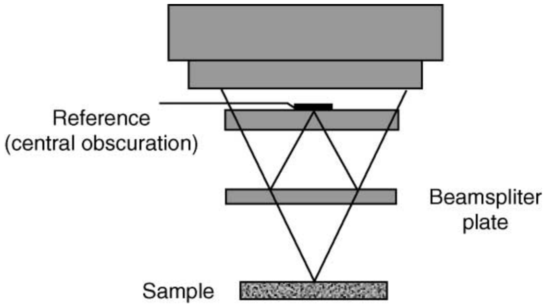

Figure 6 Schematic of a Mirau interference objective used for medium-magnification applications.

very important, as depth of focus is much reduced over lower-magnification objectives and the path from beamsplitter to reference still must be the same as from beamsplitter to in-focus test object, and most Mirau designs are therefore thermally compensated.

As one moves to even higher magnifications and numerical apertures, such as 75X or 100X, the microscope objectives do not have sufficient working distance to allow for any reasonable beamsplitter after the objective unless the numerical aperture is greatly reduced. To maintain high numerical aperture and high lateral resolution, a Linnik configuration must be used. A schematic of this configuration is shown in Figure 5. In this configuration, two separate microscope objectives are used after a beamsplitter, with one beam again reflecting from a high-quality reference and the other from the sample.

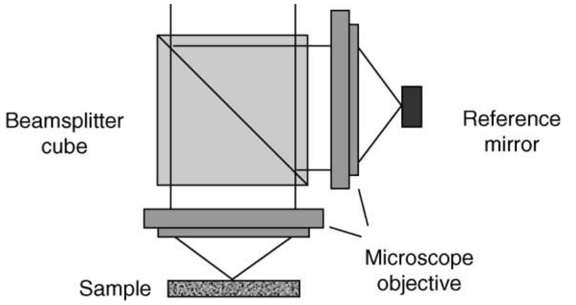

Figure 7 Schematic of a Linnik interference objective used for high-magnification applications.

A Linnik objective is generally the most difficult to use and expensive interferometric configuration (Figure 7). High numerical aperture typically limits working distance to a few hundred micrometers. In addition, depth of focus is very small, and thus these objectives must often actively adjust the reference mirror position to maintain focus over long periods of time and environmental conditions. Also, the dispersion and aberrations in the two objectives must be closely matched to ensure good fringe contrast, as there are many noncommon-path optical elements in this design; typically many objectives must be evaluated to find two that suitably match, another contributor to cost. However, these objectives do have the highest lateral resolution, and with an NA of up to 0.95, can measure the steepest slopes due to a large collection angle.

Applications

White-light interferometers are noncontact, and thus will not damage samples under test. They provide true three-dimensional surface characterization with nm-level accuracy and repeatability, and can be employed in production and laboratory environments. The various interferometric objectives allow for lateral resolutions and ranges to meet most needs; fields of view may even be stitched together to create surface maps whose extent is limited only by the range of the translation stage employed. In addition, each scan is typically quite fast, from under one second in phase-shifting mode to vertical scan rates of up to 100 μm/sec in coherence-sensing mode, with no sample preparation time. While the systems must be properly calibrated, vibration minimized, contrast maximized, and potential error sources such as dispersion and phase change on reflection effects must be managed, they remain one of the most accurate surface characterization tools available.

All of the above factors have enabled the use of white light interferometers in a variety of fields. The semiconductor and data storage industries

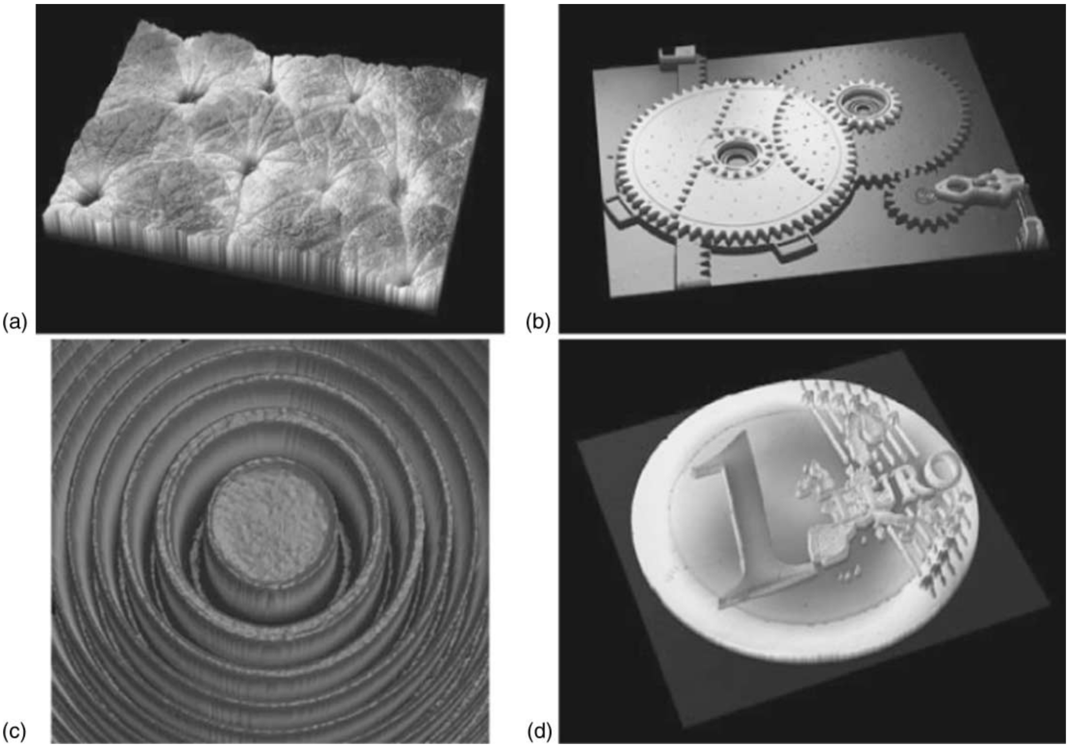

Figure 8 Surface plots from a white light optical profiler. (a) Human skin sample, (b) MEMS microgears, (c) binary optic, and (d) 1 Euro coin. Pictures courtesy of Veeco Instruments.

perform critical dimension and surface characterization of various components using these systems, some of which are configured for full wafer-handling capabilities. Systems are also used to measure precision machine parts, optics, and medical instrumentation and implants. The high sensitivity of the coherence sensing techniques allows surfaces with less than 0.1% reflectivity to be measured, so that characterization of paper, abrasives, and even different foods has been measured with these instruments. With the addition of the advanced technologies mentioned above, these systems are also now doing dynamic measurements of MEMS and thermal response characterization of precision surfaces, and also are measuring thicknesses of adhesives, drug coatings, photoresist, and a variety of other semi-transparent layers in various applications. Figure 8 shows several surface measurements, including human skin, MEMS microgears, a binary optic, and 1 Euro coin.

Laser-based Interference Microscopes

Another, less common type of amplitude-division interference microscope has recently been commercialized, primarily for use in examining micro-optics. These laser-based systems employ a Tywman–Green interferometric configuration combined with a microscope, and are generally called micro-LUPI (laser unequal path interferometer) systems. A schematic of a micro-LUPI interferometer is shown in Figure 9. A beamsplitter divides the incident laser beam, half of which travels to a generally flat reference surface, while the other half typically travels through a

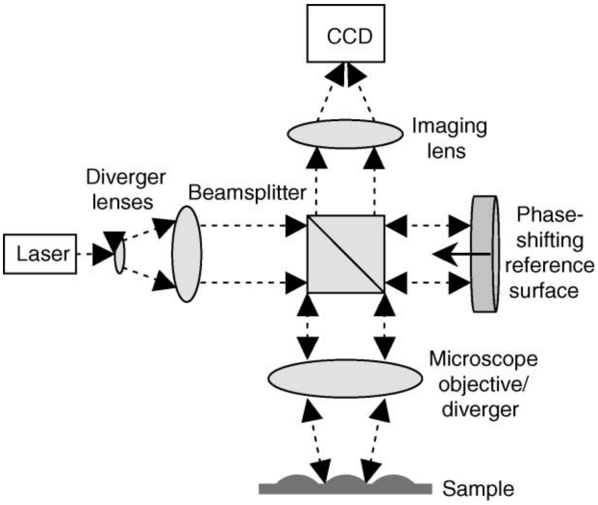
CCD
Imaging lens
Diverger lenses
Beamsplitter
Laser
Phase-shifting reference surface
Microscope objective/ diverger
Sample

Figure 9 Schematic of a micro-LUPI interferometer operating in reflection mode.

microscope objective that acts as a diverger lens and is incident on the test sample.

Due to the long-coherence length of the laser source, the two optical paths do not have to be equal in order to observe interference fringes, and samples may be measured either in reflection or transmission, depending on the desired test result. Collimated, diverging or converging illumination may be brought incident on the sample, depending on the optics in the test beam and their placement. The diverger elements are noncommon-path optics, and reduce the ultimate accuracy of the system to approximately lambda/10, similar to laser-based Fizeau interferometers.

Micro-LUPI-based systems have several advantages for testing small optics over other interference microscopes. First, the ability to measure both transmission and reflection properties of the optics, as well as different incident illumination schemes, means they can test the optics in their use configuration rather than inferring optical properties based solely on surface measurements. They can determine imaging properties, inhomogeneities and thickness variations in addition to characterizing the surfaces under test. Also, the curved test wavefront better matches spherical test surfaces, minimizing the number of observed fringes and making testing of small aspheres easier. Lastly, these systems can have curved reference surfaces as well as flat, and with the proper configuration aberrations in measurements due to the interferometric microscope can be minimized, ensuring high-quality measurements of the optic under test.

These systems have only phase-shifting measurement capability, as the coherence length of the source is too long for white-light interferometric techniques. Transmission tests necessitate a more complex part handling structure, with both the test piece and reflective return optic needing adjustment for tip/tilt and vertical and lateral translation. Therefore, such systems are more costly and not as versatile as a white light optical profiler, but do provide a valuable capability to the micro-optics community.

See also

Coherence: Coherence and Imaging. **Imaging:** Interferometric Imaging. **Interferometry:** Overview; Phase Measurement Interferometry; White Light Interferometry. **Microscopy:** Confocal Microscopy; Overview; Phase Contrast Microscopy.

Further Reading

Advanced Light Microscopy: Measuring Techniques; *Maksymilian Pluta*; North-Holland (1993).

Born M and Wolf E (eds) (1999) *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th edn. Cambridge University Press.

Creath K (1988) Phase-measurement interferometry techniques. In: Wolf E (ed.) *Progress in Optics*, vol. XXVI, pp. 349–393. North-Holland, Amsterdam.

Hariharan P (2003) *Optical Interferometry*, 2nd edn. Academic Press.

Hariharan P and Malacara D (eds) (1995). Selected Papers on Interference, Interferometry, and Interferometric Metrology. *SPIE Milestone*. MS 110.

Krug W, Rienitz J and Shulz G (1964) *Contributions to Interference Microscopy* (translated from the German edition of 1961 by J. Home Dickson): Hilger and Watts Ltd.

Malacara D (ed.) (1992) *Optical Shop Testing*. John Wiley & Sons.

Massie NA (ed.) (1988). Interferometric Metrology. *SPIE* 816.

Polarvarapu PL (ed.) (1997) *Principles and Applications of Polarization-Division Interferometry*. John Wiley & Sons.

Tolansky S (1970) *Multiple-beam Interferometry of Surfaces and Films*. Dover Publications.

Nonlinear Microscopy

S Lévêque-Fort and P Georges, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear microscopy has become an important tool for the investigation of biological applications where high resolution and three-dimensional imaging are essential for understanding the underlying biological function. Two-photon excitation fluorescence (TPEF) microscopy has opened the way to nonlinear techniques and is now widely available. As this new technology enables noninvasive study of biological specimens with high resolution, the number of applications in biology and medicine is still increasing. By exploiting autofluorescence of cells or tissues components, or by coupling with genetically modified fluorescent indicators or small extrinsic probes, high contrast and localization can be obtained. Different detection techniques (fluorescence lifetime imaging (FLIM), fluorescence correlation spectroscope (FCS), intensity, and spectral imaging) permits the analyze of several fluorescence properties and so obtains complementary informations.

The successful development of two-photon techniques has generated substantial interest in the scientific imaging community in further developing and exploiting optical nonlinearities for generating new image contrast. All these new developments have strongly benefited from the major advances in ultrafast lasers which have become more compact and reliable. Second- and third-harmonic generation (SHG, THG) as well as coherent anti-Stokes Raman scattering microscopy (CARS) have thus benefited from the fact that it is now easier to generate ultrashort pulses at almost any desired wavelengths. Association of these new nonlinear techniques in the same instrument (TPEF-SHG/TPEF-CARS, TPEF/THG) is commonly encountered and provides a powerful instrumental tool for probing biological samples.

Basic Principles of Two-Photon Microscopy (TPM)

Two photon microscopy is preferred over single photon method in many biological applications due to several intrinsic advantages which arise from having a nonlinear intensity depend absorption. In general, the nonlinear polarization for a material can be expressed as

$$P = \chi^{(1)}E^1 + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \quad [1]$$

where P is induced polarization, $\chi(n)$ is the nth order nonlinear susceptibility, and E the electric field vector of the incident light. The first term describes normal absorption and reflection of light; the second term describes SHG, hyper-Rayleigh scattering, sum and difference frequency generation, and the third terms covers multiphoton absorption, third-harmonic generation and stimulated Raman processes.

In 1931 Maria Goppert Mayer predicted that photons of lesser energy can together cause an excitation 'usually' produced by the absorption of a single photon of higher energy, in a process call multiphoton excitation. The hypothesis could not be confirmed until the invention of pulsed ruby lasers, in the 1970s, and was first applied to microscopy with the work of Denk *et al.* Two-photon microscopy used the simplest version of her theory: two photons of equal energy (from the same laser) can excite a molecule, and the excitation is similar to that obtained with a single photon possessing twice the energy (Figure 1). Thus, the molecule can emit a fluorescent photon just as if it was excited with a single higher energy photon (Figures 1 and 2). Of course, the absorption of the two photons must occurs nearly simultaneously ($\sim 10^{-16}$ s), resulting in a quadratic dependence on the light intensity rather than the linear dependence of conventional one photon fluorescence.

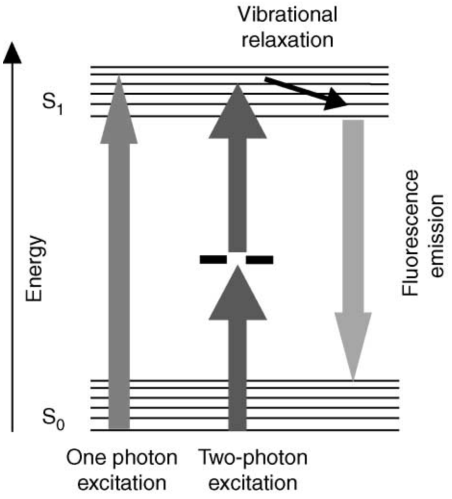

Figure 1 Jablonski diagram for one and two-photon excitation.

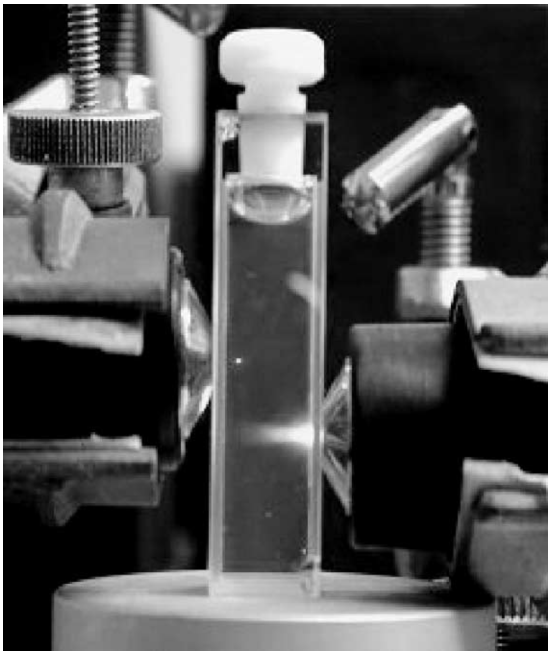

Figure 2 Localization of excitation under two photon excitation (left) compared to one photon excitation (right).

The nonlinear optical absorption property of two-photon excitation, limits the fluorochrome excitation to the point of focus. This localized excitation volume results in greatly improved axial resolution and image contrast compared with conventional microscopy, but also significantly decreases sample photobleaching and photodamage by reducing the region of photointeraction (Figure 2).

Another major advantage of TPEF is its ability to image deeper inside biological samples, owing to the reduced scattering and absorption of near infrared wavelength compared to UV and visible light. Near infrared light (700–1000 nm) is located within the 'therapeutic optical window' where the absorption of the main cells and tissue components is low.

TPEF also permits a high rejection of residual excitation light: in standard one-photon microscopy, the excitation wavelength can be spectrally close to the fluorescence emission band. Filters used to eliminate the leakage of excitation light into the detection channel often cuts off a part of the fluorescence emission. For TPEF excitation and emission band are well separated, and highly efficient filters can remove the residual excitation light without compromising the microscope sensitivity.

Under two-photon excitation, the fluorescence photon flux depends on the square of incident photon flux:

$$F \propto \sigma_{2\mathrm{P}} C I^2 \tag{2}$$

where $\sigma_{2\mathrm{P}}$ is the two-photon cross-section (cm^4 s) with 10^{-50} cm^4 s, called a Goppert–Mayer (GM), C the fluorophore concentration, and I^2 the incident photon flux. The two-photon cross-section is a quantitative measure of the probability of two-photon absorption and as a second-order process is low. As it is difficult to directly measure two-photon cross-sections, we usually use the cross-section of two photon-induced fluorescence (σ_{TPEF}) as the product of two photon absorption cross-section ($\sigma_{2\mathrm{P}}$) with the fluorescence quantum yield $\phi_{\mathrm{F}}\,\sigma_{\mathrm{TPEF}} = \sigma_{2\mathrm{P}}\phi_{\mathrm{F}}$.

Intrinsic fluorophore, such as nicotinamide adenine dinucleotide (NADH), have low cross-sections ($<10^{-4}$ GM), usual fluorescent dye in biology ranging between 1 and 300 GM and quantum dots approaching 50 000 GM. As the probability of two-photon absorption is quiet low, high incident photon flux needs to be delivered to the sample to generate efficient absorption:

$$I \propto E \frac{1}{\Delta\tau\omega^2} \tag{3}$$

I depends on the energy per impulsion E (mean power/pulse frequency), the pulse duration $\Delta\tau$, and the beam waist ω. The expression of I clearly demonstrates that by using ultrashort pulses and by reducing the beam waist, one can maximize the incident flux. Currently the most commonly used laser source for multiphoton microscopy is the femtosecond titanium-sapphire (Ti:Sa) system. These pulsed femtosecond lasers deliver a 100 fs pulse train at a repetition rate of ~80 MHz with a tuning range from 700 to 1000 nm. The beam waist depends on the objective numerical aperture. If we consider a Gaussian beam profile, with radial and axial parameters of σ_{r} and σ_{z} respectively and for a high numerical aperture $\mathrm{NA} > 0.8$:

$$V = \left(\frac{\pi}{2}\right)^{3/2} \sigma_{\mathrm{r}}^2 \sigma_{\mathrm{z}} \tag{4}$$

with $\sigma_{\mathrm{r}} \approx 0.52\lambda/\sin\alpha$ and $\sigma_z \approx 0.76\lambda/(1-\cos\alpha)$. where λ is the excitation wavelength, α is the one-half of the objective angular aperture ($\mathrm{NA} = n \sin\alpha$, n index of the immersion medium). So for a wavelength of 780 nm and an objective working with water immersion with a numerical aperture of 0.9, the excitation volume can be described as an obloidal shape with a radial axis of typically 500 nm and an axial axis of 1.90 μm. It corresponds to an excitation volume of ≈1 fL ou 1 $\mu\mathrm{m}^3$. Higher resolution can be achieved using an immersion oil objective (NA = 1.3) at the cost of a higher cost and a shorter working distance.

Typical powers used to generate two-photon excitation without damaging the sample are between 1 and 10 mW for the mean incident power. For a laser beam of 1 mW, at a rate of 76 MHz with 100 fs pulse duration, and with an objective leading to a surface in the focal plane of 1 $\mu\mathrm{m}^2$, the peak power is 0.13×10^9 $\mathrm{W/cm}^2$ at the focal plane of the sample. Of course each samples has its own limits regarding incident intensity, and the key to use TPEF safely is always to reduce the excitation intensity in order to avoid photodamage.

Fluorescent Probes

Biological systems often possess endogenous fluorescent molecules that can be imaged, revealing important sample characteristics without the need of labeling (Figure 3a,b). Cells or tissues can be imaged through their autofluorescent, with the emission of several components, NADH, flavins, collagen fibers, etc. Pharmacology studies also exploit the natural autofluorescence of aromatic drugs, to

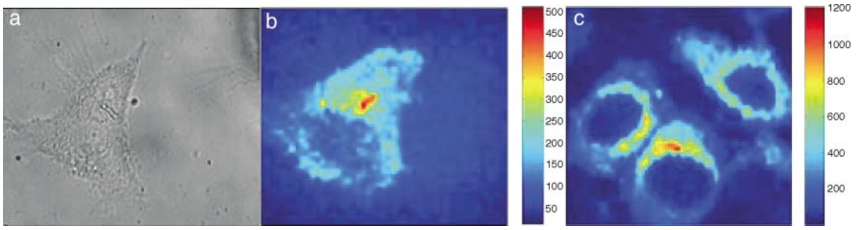

Figure 3 Applications showing various capabilities of TPEF: (a) Hela cells in transmitted image and corresponding autofluorescence image (b) under TPE (780 nm). (c) After adding 40 μM of an anti-HIV drug on similar cells, intrinsic fluorescence of the drug allows us to find its preferable localization inside the cytoplasm.

localize drug inside cells and study their reactivity without any labeling.

However, for some specific applications where the autofluorescence is too low, or if one needs to specifically identify different cells components (Figure 3g), or follow protein mobility, the samples need to be labeled your with extrinsic fluorophores. Extrinsic fluorophores are organic molecules, first developed for one-photon excitation, and are used under two-photon excitation with more or less efficiency as OPE and TPEF have different quantum mechanical selection rules, and the TPE spectrum is not always scaled to the half of OPE spectrum. Current works are focused on the development of specific markers for TPE or SHG to enhance contrast and also reduce incident laser power.

Two new generations of marker have been developed: the development of the green fluorescent protein (GFP) and its mutant introduces a convenient method for monitoring gene expression in cells and tissues; quantum dots are the latest focus in fluorescence labeling. They are promising reporters as their size (1–100 nm) induces a specific narrow emission spectrum in the visible. Absorption can occur over a wide range of short wavelengths because of the existence of multiple electronic states. Time-domain measurement has been performed with CdSe quantum dots which shows multiexponential decay with fluorescence lifetime of 3.4 ns, 16.1 ns, and 35.6 ns, much longer than autofluorescence lifetimes, hence providing greater sensitivity than intensity-based measurement. However, like GFPs, quantum dots display blinking behavior.

Two-photon Applications

Two-photon excitation has become the technique of choice for fluorescence microscopy and is applied to a variety of imaging tasks on samples ranging from single molecules to thick tissues. Owing to the high NIR penetration depth, biopsies and living tissue can be noninvasively investigated by morphological and functional fluorescence imaging of endogenous (Figure 3b) or exogenous fluorophores to depths of more than 300 μm. In genomics, multiphoton multicolour fluorescence *in situ* hybridization (FISH) are used to label various specific regions of DNA, and identified different genes and chromosome regions. In pharmacology, a higher penetration and localization of photodynamic therapy is observed, detection of low drug concentration inside cells facilitates the reactivity comprehension (Figure 3b,c). Neurobiologists investigate calcium dynamics inside brain slices and live animals, or follow neurotransmitter diffusion. Of course, these are only a few examples among the increasing number of applications, and some of them are illustrated in Figure 3.

Two-photon Instrumentation

Two-photon microscopes are now commercially available, however a large number of groups have built their own from separate components or by modifying an existing confocal microscope. Home-made microscopes present a higher configuration flexibility for excitation/detection at a much lower price.

Femtosecond, picosecond, and continuous wave (CW) laser sources have been used for TPM. By using a cw laser sources such as ArKr and NdYag, the average power increases by ~200-fold to achieve the same excitation rate obtained with a femtosecond laser, which is not highly desirable for biological samples. So as previously shown, the TPE efficiency is higher with shorter pulses, so Ti:Sa femtosecond lasers are most commonly used.

One of the drawbacks of TPM is that it is a laser scanning technique. To reconstruct an image you need either to move the sample or the laser beam. Most commonly $x-y$ mirrors driven by a galvanometer scanner are used to scan the laser at the focal plane of the sample. Lenses are used to expand the beam to overfill the objective back aperture and so obtain a diffraction limited focal volume. The excitation light enters the modified inverted epi-fluorescence microscope (Figure 4).

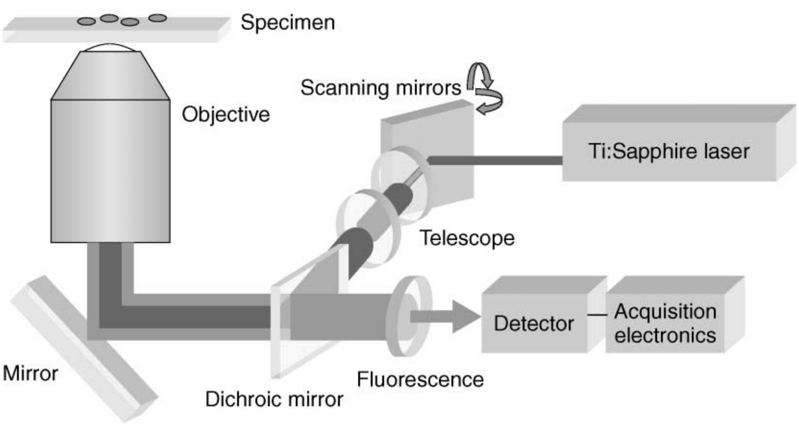

Figure 4 Schematic of a typical two-photon microscope.

A dichroic mirror transmits the excitation light and reflects fluorescence. High numerical aperture objectives are used to provide a high resolution but also maximize fluorescence collection. They are usually mounted on piezo-electric translators to provide a fast *Z*-scan. The fluorescence is collected by the same objective, reflected by the dichroic to the detector. A filter is placed in front of the detector to remove any residual excitation light. Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), avalanche photodiode or charge-coupled device (CCD) cameras. PMTs are the most commonly implemented because they are robust, low-cost, have large, active areas, and good sensitivity.

As TPE is intrinsically confined to the focal volume, two-photon microscopy unlike confocal microscopy, does not need emission pinholes and descanning optics to achieve axial depth discrimination Some authors have, however, added a pinhole to enhance resolution but at the cost of important sign al loss. Concerning the design of the detection pathway, depending on the size of the sample scanned, the objective magnification and the detector area, one may need to use a lens to collect the fluorescence onto the detector.

Fluorescence Detection

Contrast in fluorescence imaging can be provided through several complementary parameters such as steady-state intensity, spectral dependence (at absorption and emission), or fluorescence lifetime.

Intensity Imaging

Intensity images are obtained by summing all the incoming photons during an acquisition time (typically ~1 to 10 ms) which strongly depends on the fluorescence quantum yield of the sample. The whole samples intensity image is obtain by scanning the beam on the focal plane with the $X-Y$ scanner, leading to an acquisition time of 25 s for a 50×50 μm image (1 μm step).

Spectral Imaging

Spectral imaging can be a powerful way to discriminate different molecules, and to identify cellular components or perform co-localization of different proteins. It can be performed by either using an interference filter in front of the detector to extract the contribution in a specific band, or more precisely with a spectrometer. However, autofluorescent spectral bands are usually broad and sometimes it is difficult to separate the different emissions. Extrinsic trackers are thus used to labels specifically, for example, the nucleus, the plasma membrane, and mitochondria, and each tracker usually emits in a specific and well-known spectral band, allowing to easily discriminate and localize the different cellular components. Quantum dots offer narrow emission spectra and allow high spatial resolution, and with the rapid development of their functionalization, they should lead to rapid progress in intracellular imaging.

Fluorescence Lifetime Imaging (FLIM)

Imaging of fluorescence lifetimes is a useful contrast-enhancing method that provides information related to the microenvironment with spatial resolution. The fluorescence lifetime τ corresponds to the average time a fluorophore stays in its excited state is given by $\tau = 1/(k_f + k_{nr})$ with k_f the radiative decay and k_{nr} the nonradiative decay rate.

The radiative decay rate is a characteristic property of a molecule and therefore may be used to contrast different chemical species or fluorophores.

The nonradiative decay rate is a function of the local environment of an excited molecule, and so fluorescence lifetime measurements can provide information about environmental changes (e.g., pH, pO_2, $[Ca^{2+}]$). Furthermore, different cellular probes or tissue components with similar emission spectra but different lifetimes can be distinguished. Another advantage of lifetime measurement is that as a relative measurement of the fluorescence intensity, it does not depend on a non-uniform intensity excitation of the sample, and also does not depend on molecular concentration, allowing one to detect very low quantity of molecules.

Following a delta excitation, the fluorescence intensity de-excitation is given by$I(t) = I_0\, e^{(-t/\tau)}$, with I_0 the initial amplitude. Fluorescence lifetime measurement has been demonstrated either in frequency or time-domains. Time-resolved measurements in the frequency domain are based on the determination of demodulation M and phase shift $\Delta\varphi$ and require modulated light sources (modulation frequency $\omega = 2\pi f$).

The excitation source $I_E(t)$ and induced fluorescence $I_F(t)$ can be written as:

$$\begin{aligned} I_E(t) &= I_{E0} + I_{E\omega}\cos(\omega t + \varphi_E) \\ I_F(t) &= I_{F0} + I_{F\omega}\cos(\omega t + \varphi_F) \end{aligned} \qquad [5]$$

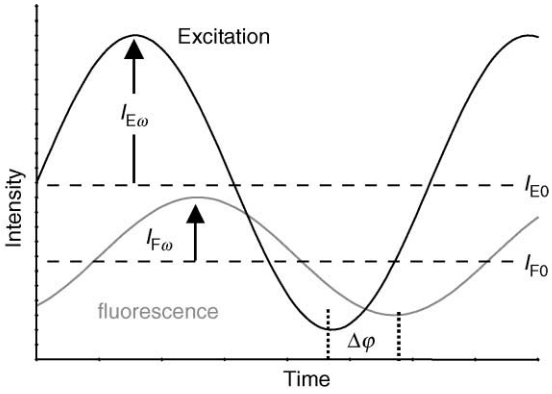

Figure 5 Time course of excitation and emission light in frequency domain fluorescence lifetime measurements techniques.

In the case of fluorescence with mono-exponential decay, the lifetime τ can deduced from the relations (Figure 5):

$$M = \frac{I_{F\omega}/I_{F0}}{I_{E\omega}/I_{E0}}. \qquad \omega\tau = \tan(\Delta\varphi) = \sqrt{\frac{1}{M^2} - 1} \qquad [6]$$

In order to translate the phase and demodulation information from the 80 MHz frequency to a lower kHz frequency, a modulated gain $G(t)$ at the frequency ω' can be applied to the detector (heterodyne detection):

$$G(t) = G_0 + G_{\omega'}\cos(\omega' t + \varphi_G) \qquad [7]$$

The detector response $S(t)$ is thus proportional to the fluorescence intensity:

$$\begin{aligned} S(t) = {} & G_0 I_{F0} + \frac{G\omega' I_{F\omega}}{2}\frac{1}{\sqrt{1 + (\omega\tau)^2}} \\ & \times \cos[(\omega - \omega')t + \varphi_G - \varphi_E + \Delta\varphi] \end{aligned} \qquad [8]$$

In the time-domain, the single point detection system is called time-correlated single-photon counting (TCSPC). The principle of the TSCPC system is shown in Figure 6. For each excitation pulse a trigger signal is sent to a time-to-amplitude converter (TAC) within which a capacitor starts charging linearly with time. Upon arrival of each fluorescence photon on the detector a second trigger pulse stops the capacitor charging. The charge is proportional to the delay between the excitation pulse and the fluorescence photon. By repeating the measurement over many thousands of excitation pulse, a histogram of photon arrival can be build up. As each photon must be record individually, the incident power is reduce to avoid the detection of only photons from the beginning of the decay (pile-up) which would lead to shorter lifetimes. To match detector dead-time and TAC counting rate, the TAC is oftenly used in reverse mode: the fluorescence photon starts the TAC and the following pulse stops it.

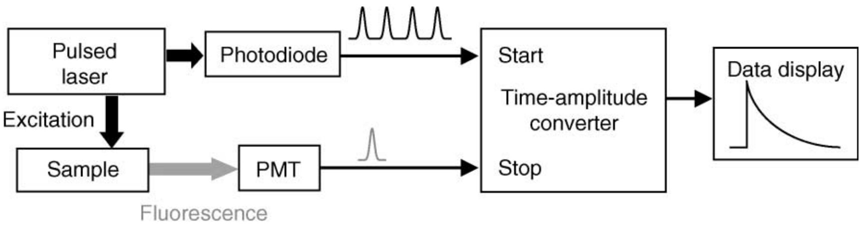

Figure 6 Schematic of time correlated single photon counting system.

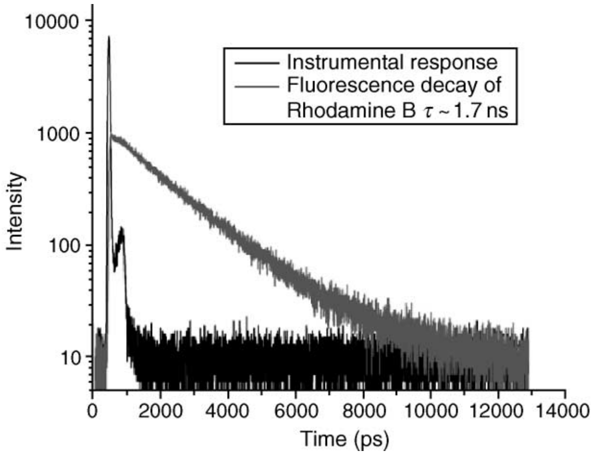

Figure 7 Typical instrumental response of MCP-PMT (black curve), and fluorescence decay curve of rhodamine B (grey curve) with $\tau \sim 1.7$ ns.

Typical detectors used are the photomultiplier tube (PMT), microchannel plate-PMT (MCP-PMT), and avalanche photodiode.

Typical instrumental response of our MCP-PMT is represented in Figure 7 (black curve) leading to a resolution of 30 ps. Depending on the fluorescence quantum yield of the sample, in order to have a good signal to noise ration, acquisition time is ranging between 1 and 30 s per point. Fluorescence decay of rhodamine B with a lifetime $\tau \sim 1.7$ ns is also presented in Figure 7.

For both detection approaches, single point lifetime measurements can be extended to lifetime imaging by synchronizing the sample scanning and the lifetime detection. However, due to the long time acquisition needed, images are only limited to small interesting areas rather than the whole sample. We will see in the recent development section below, new ways to overcome this issue.

Fluorescence Correlation Spectroscopy (FCS)

FCS experiments are performed by recording fluorescence intensity fluctuations (i.e., fluorescence bursts (Figure 8)) from the small focal volume (<1 fl). Applications include diffusion, chemical reactions, molecule concentration, or hydrodynamic flow–FCS under TPE benefits from all the advantages of TPE versus OPE.

Various experimental parameters can be extracted through temporal analysis of fluorescence intensity fluctuations by calculating the autocorrelation or cross-correlation of the fluorescence signal defined as:

$$g(\tau) = \frac{\langle I_F(t) I_F(t+\tau)\rangle}{\langle I_F^2(t)\rangle} = \frac{\int I_F(t) I_F(t+\tau)\mathrm{d}t}{\int I_F^2(t)\mathrm{d}t} \qquad [9]$$

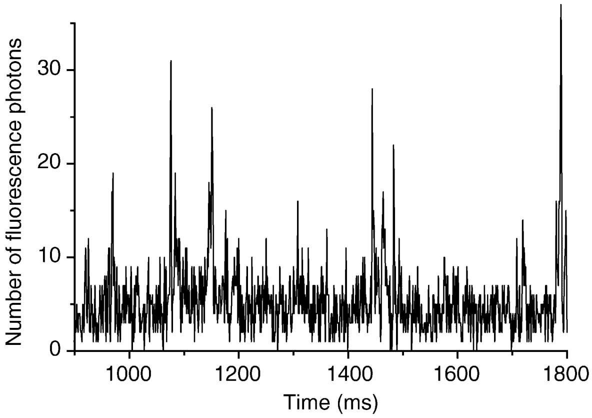

Figure 8 Fluorescence intensity raw data along the time. Bursts are observed when molecules cross the focal volume.

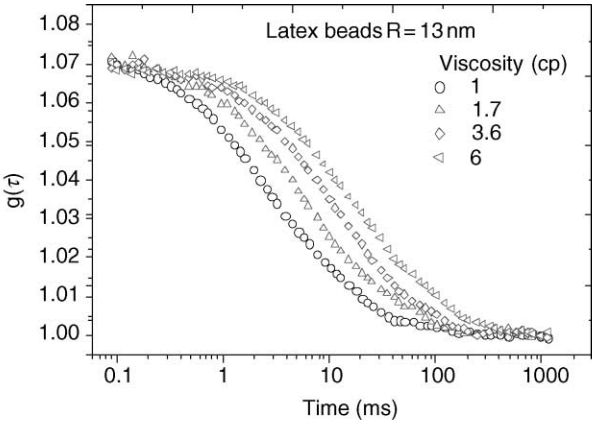

Figure 9 Correlation function observed for microspheres placed in different viscous solution.

where I_F is the time-dependent fluorescence signal. By considering that the fluorescence intensity fluctuates around a mean value:

$$I_F(t) = \langle I_F\rangle + \delta I_F(t) \qquad [10]$$

the autocorrelation function can be written as:

$$g(\tau) = 1 + \frac{\langle \delta I_F(t)\delta I_F(t+\tau)\rangle}{\langle I_F^2\rangle} \qquad [11]$$

FCS have been extensively used to extract diffusion coefficient of small fluorescent latex bead in solution, showing the influence of the bead radius or solvent properties. Figure 9 clearly exhibits the influence of the viscosity of the solvent on the diffusion of microspheres. Single molecule detection has also been reached. FCS as also been extended to biological applications where the low background associated with TPE and the reduced autofluorescence is critical. Diffusion of microbeads inside cells cytoplasm or bacterial biofilms have been successfully demonstrated.

Recent Developments in Two-Photon Microscopy

Towards video microscopy

All the nonlinear microscopies share an important limitation: the image build up pixel-by-pixel by scanning across the sample, leading to long-time acquisition. Depending on the sample's emission and the area studied, scanning rates between 0.5 Hz and 30 Hz can be reached for intensity images. To speed up this acquisition time, a first approach is based on a line excitation, in combination with widefield CCD detection. This provides real-time image acquisition rates at the expense of a reduced axial resolution (5 μm instead of 1 μm). A more ideal way of spreading the available excitation power over the sample and speed up image buildup, is the simultaneous use of several well separated foci, each of them generating a diffraction limited spot. In combination with whole field detector, as with a CCD camera, image acquisition rates can be strongly enhanced without compromising resolution. The spacing between the foci must be carefully chosen in order to avoid out-of-focus interference. By introducing time-delay between the different foci, larger than the pulse duration, each foci of a microlens array can be position very closed without loss of resolution. An alternative way has been proposed; a special beamsplitter has been developed to create multiple beams with a higher efficiency, furthermore this setup inherently induced time-delay between each foci. Time-domain has been demonstrated in both approach, and FLIM maps on cells imaging have been obtained in only few second with 200 ps time resolution.

Sharpen the focal point

Nonlinear microscopy facilitates 3D-imaging but the resolution is still limited to 180 nm in the focal plane and 500–800 nm along the optical axis. According to the diffraction theory, the resolution of a focusing light microscope is related to the size of its focal point. In the mid-1990s, new concepts appeared to break this diffraction theory and sharpen the focal volume. In 4-pi microscopy the sample is placed between two large NA objective lenses in a single axis at their common focus. Coherent interference of beams passing through the two lenses generate a point spread function whose central maximum is 3–4 times narrower than in the normal confocal case, but which also has side lobes. This last issue can be overcome using two-photon excitation. A two-photon 4-pi microscopy can thus result in a resolution of ~100 nm in all directions, and has been demonstrated successfully on filamentous actin and immunofluorescently labeled microtubules. However, a 4-pi microscopy requires the sample to be thin and mounted between two coverslips, unless one of the lenses is a dipping lens. Recent development on sample chambers allows one to overcome this issue, and results in live mammalian cells have been shown. A compact 4-pi commercial version has recently been proposed (Leica), achieving a seven-fold-improved axial resolution (80 nm) compared to the classical two-photon microscope.

Another technique, also developed by Professor Hell's group, is called stimulated emission depletion (STED) microscopy. The fluorophore in the fluorescent state S1 is stimulated to the ground state S0 with a doughnut-shaped beam. The depletion of the S1 saturates with increased I_{STED} intensity, and therefore establishes a nonlinear relationship between the fluorescence and the intensity applied to the STED. The saturation of S1 depletion is the essential element for breaking the diffraction limit as it confines fluorescence to the center. Spot size is not limited, but in practise depends on the quality and the saturation factor limited depletion. Using STED microscopy with wavelengths of 750–800 nm leads to a lateral resolution of up to 28 nm on single molecules experiments.

STED microscope has been successfully combined to 4-pi microscopy, and lateral resolution of 30–40 nm has been demonstrated.

Others Nonlinear Imaging Techniques

The success of two-photon microscopy has opened the way for other nonlinear microscopy. As opposite to multiphoton fluorescence absorption techniques where the induced signal is incoherent fluorescence, the signal generated by harmonic generation and stimulated Raman scattering is coherent with specific directional and polarization properties, as such fluorescence and harmonic images are derived from fundamentally different contrast mechanisms. However, both techniques share many features like intrinsic three-dimensionality, reduce photobleaching or deep sectioning, and require essentially identical equipment, allowing one to combine these two techniques as discussed further below.

Second-Harmonic Generation Microscopy

SHG is a second-order nonlinear optical process that can only arise from a medium lacking a center of symmetry, for example, an anisotropic crystal or at an interface of a membrane. SHG can be used to image tissue, muscles, cells membranes, or protein array, and has been recently reported to enhance membrane potential by using voltage sensitive dye.

SHG does not arise from an absorptive process, as the intense laser field induces a nonlinear, second-order, polarization in the assembly of molecules, resulting in the production of a coherent wave at exactly twice the incident frequency (half the wavelength). As represented in Figure 10, for an excitation at 800 nm, the SHG is observed at 400 nm and depending on the fluorescence properties of the molecule, a broadband fluorescent spectra is emitted in the visible.

Contrary to TPEF, where the emission properties depend only on the molecular properties and are independent of the laser source, the spectral and temporal characteristics of SHG are derived from the laser source. The SHG is temporally synchronous with the excitation pulse and its bandwidth is $1/\sqrt{2}$ of the laser's bandwidth. As SHG is an instantaneous process, the signal is only generated during the duration of the laser pulse (Figure 11), in contrary to TPEF where de-excitation of the molecules can occur over several nanoseconds after the pulse excitation.

A major difference between SHG and TPEF is that SHG requires a noncentrosymmetric environment to produce a signal. By examining the second term of eqn [1] $P = \chi^{(2)}E^2$ where the polarization and electric field are vector quantities and $\chi^{(2)}$ is a tensor, this condition is only satisfied for a birefringent crystal, an order array, or an interface. For example, the induced polarization from molecules in a dye cuvette (centrosymmetric sample) would be equal and opposite in all directions, leading to a vector sum zero. The SHG propagation is essentially in the forward direction, and the emission pattern is confined to twin lobes in the case of surface generation, or a cone in the case of volume generation. The off-axis of the two lobes critically depend on the tight focus of the driving field.

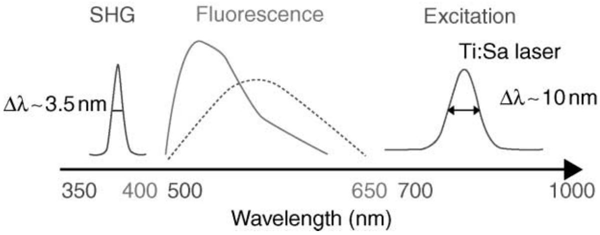

Figure 10 Wavelength domain for two-photon induced fluorescence and second-harmonic generation: excitation meets the same criteria and Ti: Sa are typically used in the range 700–1000 nm. For a typical excitation at 800 nm, fluorescence will be observed over ~450 nm, and SHG, at exactly 400 nm.

The total radiated SHG power can be derived by integrating the radiation profile from all the emission directions: $P_{\mathrm{SHG}} = \frac{1}{2}\theta_2 N^2 \sigma_2 I_0^2$ with $\sigma_2 \propto |\chi_2|^2$, I_0 the driving field intensity at the focal center, N the effective number of radiative molecules, and $\theta_2 \approx (0.1\lambda/w_0)^2$.

A schematic of a typical SHG/TPEF microscope is shown in Figure 12. The excitation source is a mode locked Ti:Sa laser which delivers ~120 fs at 76 MHz. The laser light is focused into the sample with a water immersion microscope objective (Olympus, LUMplanFl 60x) and the resultant SHG is collected in a forward direction, while TPEF is collected in the backward direction.

Figure 13 illustrates SHG and TPEF images of Di-6-ASPBS molecules under 880 nm excitation.

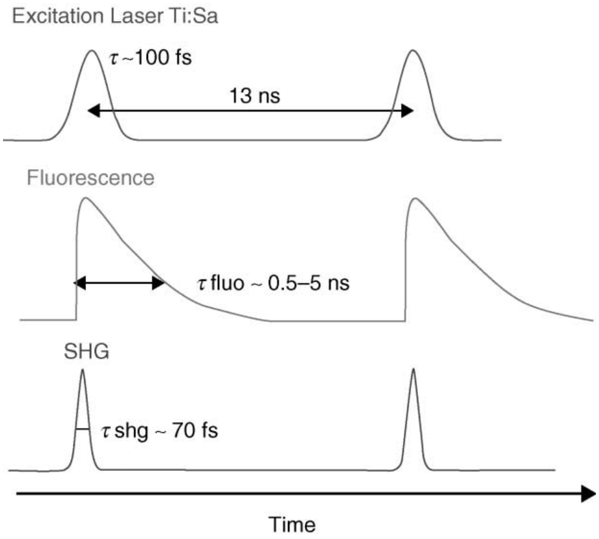

Figure 11 Typical time-response for two-photon induced fluorescence and SHG. SHG signal is an instantaneous process whereas fluorescence emission can occur several nanoseconds after the pulse excitation.

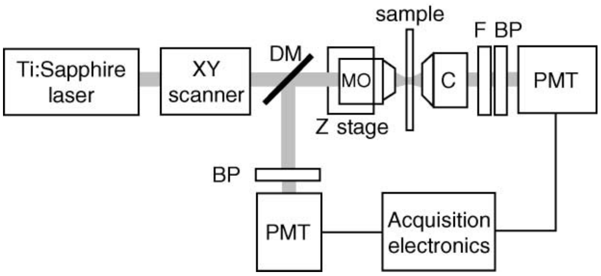

Figure 12 Schematic of SHG setup combined with TPEF extracted from (DM dichroic mirror, MO microscope objective NA 0.9 water immersion, C condenser, F colored glass filter, BP bandpass filter, PMT photomultiplier tube). Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

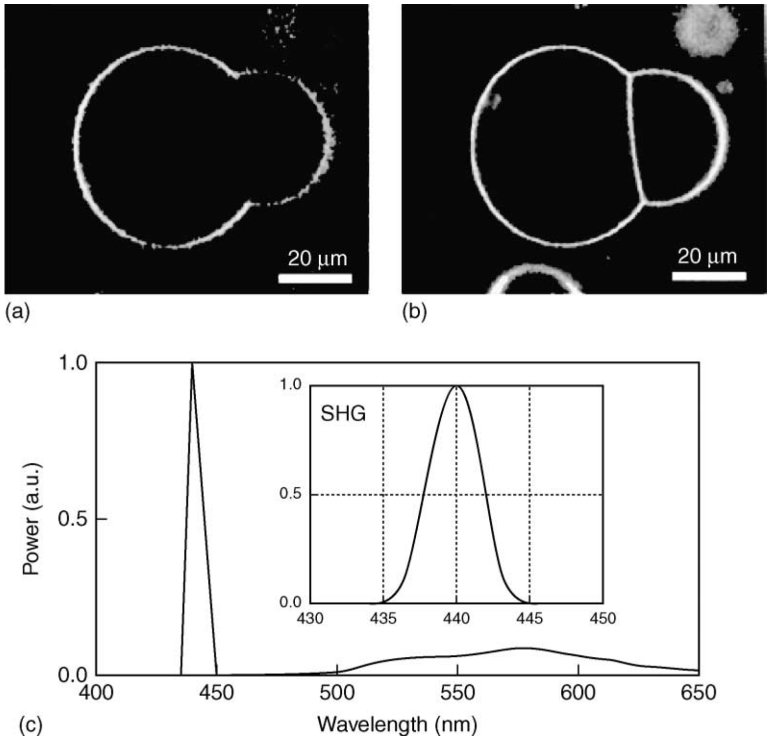

Figure 13 (a) SHG and (b) TPEF of two adhering vesicles labeled with Di-6-ASPBS (equatorial slice) excited at 880 nm. (c) The forward-detected SHG spectra (left peak) and the backward broad band fluorescence spectra (right peak). The Stokes shift in fluorescence leads to an important spectral separation between the two peaks. An expanded view of the SHG peak is presented in the inset. Reproduced with permission from Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322, © 2000 Optical Society of America.

The total acquisition time for the images was 1.5 s for an excitation power less than 1 mW. The adhesion area where the membranes are fused exhibits a centrosymmetric molecular distribution wherein TPEF is allowed, but not SHG. The corresponding forward-detected emission spectra of SHG (left peak) and TPEF (right peak) is shown in Figure 13c.

Third-Harmonic Generation (THG)

Third harmonic generation takes advantage of the discontinuity of third-order nonlinear susceptibility $\chi^{(3)}$ at material interfaces. At the focus point of an input laser beam with wavelength λ, this discontinuity gives rise to a new wave in the forward direction at the third-harmonic, $\lambda/3$. This signal can be analyzed in the transmission mode either with a single pixel detector or with a CCD camera. As with the other nonlinear techniques, THG offers a high 3D sectioning capability. Since all materials have nonvanishing third-order susceptibilities, THG microscopy can utilized as a general-purpose microscopy technique, with no need for fluorescence labeling or staining. Experimental setup for THG are very closed to SHG microscope. The need for efficient THG in the sample, together with low average laser power to avoid heating, dictates a laser source that operates with femtosecond pulses. The lasers that are often used THG microscopy are a synchronously pumped OPO at the wavelength of 1500 nm, a low repetition-rate OPA at 1200 nm and a Cr:Forsterite laser at 1240 nm. All these lasers are in the infrared above 1200 nm, thus generating THG signals in the visible range. These visible signals are compatible with the collection optics. A infrared laser undergoes low scattering, however absorption in water strongly increases.

THG have been demonstrated from subcellular to millimeter-size organisms with imaging of nerve cells, chloroplasts of a leaf, or Ca^{2+} intracellular dynamics.

Coherent Anti-Raman Stokes (CARS)

Chemical imaging by use of inherent molecular vibration signals avoids the photobleaching problem and perturbations to cell function induced by fluorophore labeling. Confocal Raman microscopy can provide 3D spatial resolution but it requires high average power due to small cross-section of Raman scattering, or long-time exposure which is not desirable in studying dynamic living system.

The above difficulties can be circumvented by multiphoton vibrational microscopy based on coherent anti-stokes Raman scattering (CARS). CARS is a four-wave mixing process in which a pump field $E_p(\omega_p)$, a Stokes field E_S (ω_S), and a probe field $E_{p'}$ $(\omega_{p'})$. Interact with a sample and generate an anti-Stokes field E_{AS}, at the frequency $\omega_{AS} = \omega_p - \omega_S + \omega_{p'}$. In practise, the pump and the probe fields derive from the same laser $\omega_p = \omega_{p'}$ and $\omega_{AS} = 2\omega_p - \omega_S$. The energy diagram of CARS is shown on Figure 14.

The CARS intensity is the square modulus of the induced polarization, $P^3 = \chi^{(3)} E_p^2 E_s$, with $\chi^{(3)}$ the third-order susceptibility.

The application of the CARS process to microscopy opens up a new method for chemical imaging. In CARS microscopy, the signal is also dependent on the square of the pump field intensity, so the temporally and spatially overlapped pump and Stokes laser pulses tightly focused into a sample, generate a signal of a small excitation volume ($<1\ \mu m^3$). As with all other nonlinear imaging techniques, the image is obtained by raster scanning the sample or the laser beams. The vibrational contrast in CARS microscopy arises from signal enhancement when $\omega_p - \omega_s$ is tuned to Raman-active vibrational frequency Ω. CARS signal generation needs to fulfill the phase-matching condition, $l < lc = \pi/|\Delta \mathbf{k}| = \pi/|\mathbf{k}_{as} - (2\mathbf{k}_p - \mathbf{k}_s)|$, where $\mathbf{k}_p$, $\mathbf{k}_s$, and $\mathbf{k}_{as}$ are wavevectors of the pump, Stoke and CARS fields, respectively, and $\Delta \mathbf{k}$ is the wavevector mismatch.

CARS presents several advantages similar to TPEF: high 3D sectioning capabilities, a low scattering and absorption of the excitation beams. CARS signal is higher in frequency than one-photon fluorescence, so it can be detected in the presence of a strong fluorescent background. Vibrational contrast is intrinsic to biological samples. Fluorescent probes are not necessary and photobleaching is thus circumvented. The coherent signal accumulation in the CARS process produces a strong and directional signal, making CARS microscopy much more sensitive than the conventional Raman microscopy. Consequently, CARS microscopy requires only a moderate average power tolerable by biological samples and has a high data acquisition speed. However, CARS microscopy is not without shortcomings. First, the intrinsically weak induced nonlinear polarizability requires sophisticated laser excitation sources with high peak power and moderate average power, and with low jitter. Second, unlike fluorescence imaging, CARS measurement is not background free, and the nonresonant signal limits the image contrast.

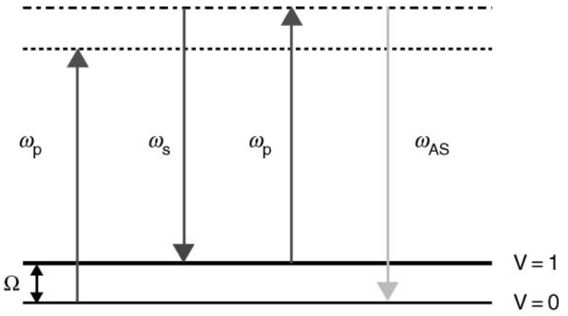

Figure 14 Energy diagram for Coherent anti-Raman spectroscopy (CARS).

The typical setup is very similar equipment compared to TPEF and SHG. The main difference arise from the excitation, two laser beams, the pump, and Stokes laser beams overlapping (in space and time) and are tightly focused into the sample using a high NA objective. The CARS signal can be detected in the forward (F-CARS) and the epi-detection (E-CARS). Unlike fluorescence emission and spontaneous Raman scattering, the radiation pattern of CARS is dependent on both the size and shape of a scatterer. The two detection geometry are thus complementary, the forward detection is suitable for imaging objects of a size comparable to or larger than the excitation wavelength.

For smaller objects, the F-CARS contrast is limited by the large nonresonant background from the solvent. Epi-detected CARS (E-CARS) microscopy provides a sensitive means of imaging objects having an axial length much smaller than the excitation wavelength because it avoids the large background from the solvent.

On the application side, CARS microscopy has been used for mapping the distribution, orientation, and diffusion of specific molecules in dynamic samples, for imaging lipids, proteins, and chromosomes in unstained live cells, and for monitoring cellular processes. CARS is a very promising technique to image small molecules such as lipids, hormones, and drug molecules, for which fluorescence labeling is prone to alter their function in cells and tissues. As for SHG, CARS microscopy is often combined with TPEF in order to obtain simultaneously the complementary contrasts.

Conclusion

Since its introduction more than a decade ago, two-photon fluorescence microscopy has quickly become a standard tool in biophotonic with applications ranging from the study of single molecules to thick tissue imaging. TPEF has opened the way to new coherent nonlinear imaging techniques such as SHG, THG, or CARS which all shared a high spatial resolution. Based on different contrasts these different approaches are often combined for a more complete diagnosis. With developments of new low cost, compact, and reliable lasers with tuning wavelength covering the spectra of a wide range of endogenous or extrinsic markers, nonlinear imaging will continue its successful implementation outside physics labs. Development of endoscopy systems in which ultrafast pulse duration is preserved, will open a new range of *in vivo* studies. Nonlinear imaging still has more to offer and will definitely be a powerful tool for biomedical applications.

See also

Microscopy: Overview. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Cheng J and Xie XS (2004) Coherent anti-stokes Raman scattering microscopy: instrumentation, theory and applications. *Journal of Physical chemistry B* 108: 827–840.

Gratton E, Breusegem S, Sutin J, Ruan Q and Barry N (2003) Fluorescence lifetime imaging for the two-photon microscope: time-domain and frequency-domain methods. *Journal of Biomedical Optics* 8: 381–390.

Hell SW (2003) Towards fluorescence nanoscopy. *Nature Biotechnology* 21: 1347–1355.

Konig K (2000) Multiphoton microscopy in life sciences. *Journal of microscopy* 200: 83–104.

Moreaux L, Sandre O, Blanchard-Desce M and Mertz J (2000) Membrane imaging by simultaneous second-harmonic generation and two-photon microscopy. *Optics Letters* 25: 320–322.

Schwille P, Haupts U, Maiti S and Webb WW (1999) Molecular dynamics in living cells observed by fluorescence correlation spectroscopy with one- and two-photon excitation. *Biophysics Journal* 77: 2251–2265.

So PTC, Dong CY, Masters BR and Berland KM (2000) Two photon micrscopy excitation fluorescence microscopy. *Annual Review of Biomedical Engineering* 399–429.

Squier J and Muller M (2001) High resolution nonlinear microscopy: a review of sources and methods for achieving optimal imaging. *Review of Science Instruments* 72: 7.

Zipfel WR, Williams RM and Webb WW (2003) Nonlinear magic: multiphoton microscopy in the biosciences. *Nature Biotechnology* 21: 1369–1377.

Zipfel WR, Williams RM, Christie R, Nikitin AY, Hyman BT and Webb WW (2003) Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. *Proceedings of the National Academy of Science of USA* 100: 7075–7080.

Phase Contrast Microscopy

C J R Sheppard, National University of Singapore, Singapore

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many biological specimens behave as phase objects, that is, they alter the phase of the light transmitted through them, rather than its intensity. Similarly, reflection from a surface with height variations alters the phase of the reflected light. Observation of phase objects requires special imaging techniques. These include defocusing, dark field, Zernike phase contrast, Schlieren imaging, Hoffmann modulation contrast, differential phase contrast, interference microscopy, shearing interferometry, Nomarski differential interference microscopy, and digital phase retrieval. The imaging performance of these different methods are most easily appreciated and compared by consideration of the transfer function approach. In particular, we can consider weakly scattering objects and slowly varying phase objects as special cases.

When a thin unstained biological specimen is illuminated, in general the transmitted light suffers changes in amplitude, phase, and polarization. The change in polarization is related to the birefringence of the sample, which is neglected in the following. Thus the object (complex) amplitude transmittance can be written:

$$t(x, y) = a(x, y)e^{i\phi(x,y)} \qquad [1]$$

where $a(x, y)$ is real and represents the modulus (also called amplitude) of the light wave on exiting from the sample, and $\phi(x, y)$, also real, represents the phase. In a perfect microscope, we observe the intensity of the transmitted light, so the image is given by

$$I(x,y) = |a(x,y)e^{i\phi(x,y)}|^2 = a^2(x,y) \qquad [2]$$

Thus the phase information is invisible in the image. A perfect microscope does not image phase variations in the object, but results in an image that depends on the amplitude of the transmittance, called amplitude contrast. Usually the change in modulus is also small, so that the image will exhibit weak contrast, which may not be discernable. In order to render the phase variations visible it is necessary to modify the microscope in one of a number of different ways.

This conclusion is also relevant in incident light microscopy of surface structures. Then phase changes result from variations in surface height, and can be made visible by phase contrast microscopy. Quantitative measurement of the phase allows quantitative measurement of the surface profile. But care must be taken to avoid errors caused by phase change on reflection from a conducting material, or from effects of the numerical aperture of the optical system. In transmission microscopy quantitative phase measurement can be useful too, for example giving the concentration of a solute.

Some of these phase contrast methods are easily understood using the simple imaging model described above. Others are more easily appreciated by considering imaging in terms of Fourier optics, equivalent to Abbe's theory of microscope imaging. We introduce the angular spectrum of the object amplitude transmittance:

$$T(m,n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t(x,y)\exp[-i2\pi(mx+ny)]\,dx\,dy \qquad [3]$$

where m, n are spatial frequencies in the x, y directions. We can write for the object amplitude transmittance:

$$t(x,y) = e^{b(x,y)} \qquad [4]$$

where $b(x,y)$ is in general complex. For a weak object we have

$$t(x,y) \approx 1 + b(x,y) \qquad [5]$$

and the spectrum is

$$T(m,n) = \delta(m)\delta(n) + B(m,n) \qquad [6]$$

where δ is a Dirac delta function and $|B| \ll 1$. The problem with imaging of phase arises because a change in the imaginary part of b produces only a small change in the modulus of t.

For a coherent imaging system, with coherent transfer function $c(m,n)$, the image intensity is

$$I(x,y) = \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m,n)T(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]dm\,dn\right|^2 \qquad [7]$$

where M is the magnification. The coherent transfer function $c(m,n)$, is equal to the scaled pupil function $P(\lambda fm, \lambda fn)$, where f is the focal length of the objective. Substituting eqn [6] into eqn [7] and dropping terms of second order we obtain:

$$I(x,y) = |c(0,0)|^2 + 2\Re\left\{c^*(0,0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \times c(m,n)B(m,n)\exp\left[\frac{i2\pi}{M}(mx+ny)\right]dm\,dn\right\} \qquad [8]$$

Normalizing the transfer function to unity at the origin:

$$I(x,y) = 1 + 2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(m,n)B(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]dm\,dn\right\} \qquad [9]$$

For an axially symmetric aberration-free system, the coherent transfer function is real and even, so that a pure phase object is not imaged (only the Hermitian component of B is imaged). Many phase contrast methods in principle rely on appropriate modification of the coherent transfer function, either by making it complex or asymmetric.

However, this theory is an approximation as microscope systems are in practice partially coherent. For a partially coherent system the image intensity is

$$I(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(m,n;p,q)T(m,n) \times T^*(p,q)\exp\left\{\frac{i2\pi}{M}[(m-p)x+(n-q)y]\right\}dm\,dn \qquad [10]$$

where $C(m,n;p,q)$ is the partially coherent transfer function, sometimes called the transmission cross-coefficient, which can be calculated as an integral over the objective and condenser pupils. The partially coherent transfer function is nonzero over a finite region of $m,n;p,q$ space, thus limiting the resolution of the system. Then eqn [8] for a weak object becomes, normalizing so that $C(0,0;0,0)=1$:

$$I(x,y)=1+2\Re\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}C(m,n;0,0)B(m,n)\times\exp\left[\frac{i2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \quad [11]$$

where $C(m,n;0,0)$ is called the weak object transfer function.

For the image of a general object that is not necessarily weak, eqn [10] is often difficult to interpret. We can gain much insight into the imaging process by considering another special form of object, this time a slowly varying object:

$$t(x,y)=a(x,y)\exp\left[i\left(x\frac{\partial\phi}{\partial x}+y\frac{\partial\phi}{\partial y}\right)\right] \quad [12]$$

where a and ϕ are assumed slowly varying. Thus, the object spectrum, assumed now as slowly spatially varying, is

$$T(m,n)\approx a\delta\left(m-\frac{1}{2\pi}\frac{\partial\phi}{\partial x}\right)\delta\left(n-\frac{1}{2\pi}\frac{\partial\phi}{\partial y}\right) \quad [13]$$

Substituting in eqn [10], we then find:

$$I(x,y)=a^2(x,y)C(m,n;m,n) \qquad m=\frac{1}{2\pi}\frac{\partial\phi}{\partial x},\ n=\frac{1}{2\pi}\frac{\partial\phi}{\partial y} \quad [14]$$

in which C is called the phase gradient transfer function, and is real. For an ordinary bright field system, C is an even function. Thus a phase gradient results in a change in image intensity, but for an axially symmetric system the sign of the change is independent of the sign of the phase change.

Defocused Imaging

Historically, the oldest method of phase contrast is simply to defocus the object. The transfer function is then complex, and phase detail is imaged. If the system is aberration-free, the contrast of the phase information reverses on the opposite side of focus. The strength of the contrast increases with defocus, until eventually artifacts are introduced when the imaginary part of the transfer function becomes negative. If the relative condenser aperture is too large, contrast is reduced.

Assume a pure weak phase object:

$$t(x,y)=1+id\cos(2\pi\nu x) \quad [15]$$

so that

$$T(m,n)=\left[\delta(m)+i\frac{d}{2}\delta(m-\nu)+i\frac{d}{2}\delta(m+\nu)\right]\delta(n) \quad [16]$$

Then:

$$I(x,y)=1-dC_i(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \quad [17]$$

where C_i is the imaginary part of C. Thus the phase information is imaged by the imaginary part of the weak object transfer function.

Dark Field Microscopy

Dark field microscopy can be used to visualize weak amplitude and phase information. The constant background in eqn [5] is eliminated, so we obtain an image of b. Note that eqns [8] and [11] are no longer valid because we cannot neglect the second-order term. For a coherent system, the background can be eliminated by use of a central stop to intercept the nondiffracted light. For a partially coherent microscope, an annular condenser aperture (larger in diameter than the objective pupil) is used (Figure 1). In both cases, in practice a range of low spatial frequencies are eliminated, resulting in the appearance of halo artifacts.

For the pure weak phase object of eqn [12], the image in a coherent microscope is

$$I(x,y)=\frac{1}{2}d^2c^2(\nu,0)\left[1+\cos\left(\frac{4\pi\nu x}{M}\right)\right] \quad [18]$$

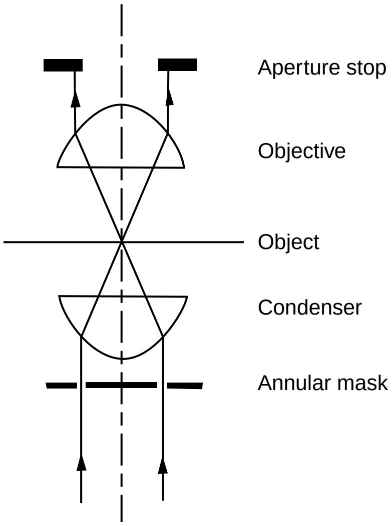

Figure 1 Schematic diagram of dark field microscopy.

so that the spatial frequency is double that in the original object. For the partially coherent microscope, we have

$$I(x,y) = \frac{1}{2}d^2\left[C(\nu,\nu;0,0) + C(\nu,-\nu;0,0)\cos\left(\frac{4\pi\nu x}{M}\right)\right] \qquad [19]$$

in which the two terms represent difference and sum frequency components. However, it can be shown that for the annular dark field system the transfer function for sum frequencies is zero, so the image intensity is a constant. The system images only difference frequencies of pairs of spatial frequencies, and not sum frequencies.

For a slowly varying phase object, zero intensity results from a region of constant phase, but the intensity rises quickly for small values of phase gradient, thus giving strong contrast. For larger phase gradients, the intensity falls off slowly.

Zernike Phase Contrast

The Zernike phase contrast method is similar to dark field except that the direct (nondiffracted) light has its relative phase changed by 90°, rather being eliminated. In practice this is achieved using an annular condenser aperture and an objective with a phase ring (Figure 2). The phase change can be ±90°, giving positive or negative phase contrast. The phase ring is usually only partially transmitting, which has the effect of enhancing the sensitivity. From eqns [8] or [11], imaging of a weak object is linear in phase. The halo artifact is present, as in dark field microscopy. Weak amplitude information is imaged as in dark field microscopy.

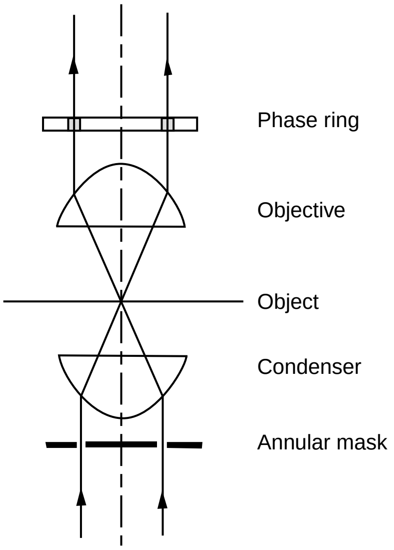

Figure 2 Schematic diagram of Zernike phase contrast microscopy.

For the weak phase object eqn [12], we then have

$$I(x,y) = 1 \pm \frac{2d}{g}C(\nu,0;0,0)\cos\left(\frac{2\pi\nu x}{M}\right) \qquad [20]$$

for positive or negative phase contrast, respectively, where g is the amplitude transmittance of the phase ring. The response of the weak object transfer function is improved as a result of the annular condenser aperture.

Schlieren Imaging, Hoffman Modulation Contrast, and Differential Phase Contrast (DPC)

In Schlieren imaging with coherent illumination, a half plane mask (a Foucault knife edge) is inserted to eliminate say the negative spatial frequencies in x. Taking now $c(0,0) = 1/2$, eqn [8] becomes, after normalization:

$$I(x,y) = 1 + 4\Re\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty} c(m,n)B(m,n) \times \exp\left[\frac{i2\pi}{M}(mx+ny)\right]\mathrm{d}m\,\mathrm{d}n\right\} \qquad [21]$$

For the weak phase object eqn [15], we then have:

$$I(x,y) = 1 - dc(\nu)\sin\left(\frac{2\pi\nu x}{M}\right) \qquad [22]$$

The image exhibits a variation in phase quadrature to the original phase variation, that is, it is related to the x-derivative of the phase. The effect occurs because the coherent transfer function contains an odd (but real) part.

For a partially coherent microscope a similar effect can be achieved simply by offsetting the illumination system, but now the transfer function is no longer single-sided. It can be resolved into odd and even parts that result in differential phase contrast and amplitude contrast, respectively. The amplitude contrast component merely contributes to the background for a phase-only object, and can be removed simply by contrast enhancement of a digital image. The offset can be introduced in practice by inserting a half-plane mask into the condenser aperture stop. Imaging of a slowly varying object can be explained by eqn [14], in which the transfer function is not symmetrical. The image shows a pseudo-three-dimensional bas relief effect. A commercial implementation of an asymmetric transfer function is Hoffman modulation contrast, which uses offset illumination together with an amplitude mask in the objective pupil (Figure 3). Advantages of Hoffman contrast are that it uses standard bright-field objectives and avoids problems from birefringent specimens.

A similar method can be used in scanning laser microscopy. In a scanning microscope the effect of the illumination and detection systems are interchanged. Thus DPC can be achieved in a scanning laser microscope by offsetting the detector. In practice this is conveniently done by using a split detector (Figure 4), a detector divided into two semicircular elements. If the signal of one half is subtracted from that of the other, the amplitude contrast component cancels out, so that pure differential phase contrast results. For weak phase gradients, the intensity varies linearly with phase gradient, again showing the bas relief effect. It is thus bi-directional, and the difference signal can become negative. A constant value is, therefore, usually added to it before display. The sum of the two detector elements results in a conventional bright-field image. From eqn [14], both difference and sum signals are proportional to $a^2(x, y)$, so that if the ratio of these two signals is extracted, an image of the phase gradient in the x direction results. The response to either weak phase gradients or fine detail can be enhanced by modifications using annular split detectors. Use of a quadrant detector allows the phase gradients in the x and y directions to be imaged. DPC microscopy has been demonstrated to be sensitive enough to observe monomolecular films and atomic surface steps.

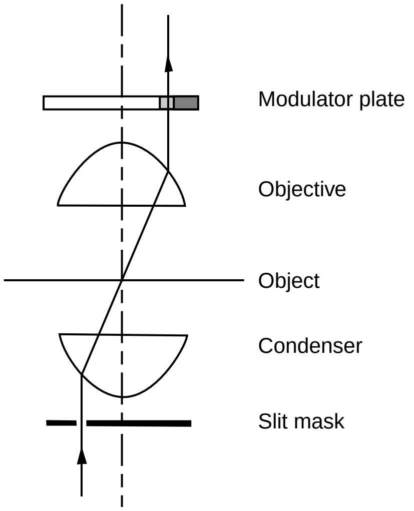

Figure 3 Schematic diagram of Hoffmann modulation contrast microscopy.

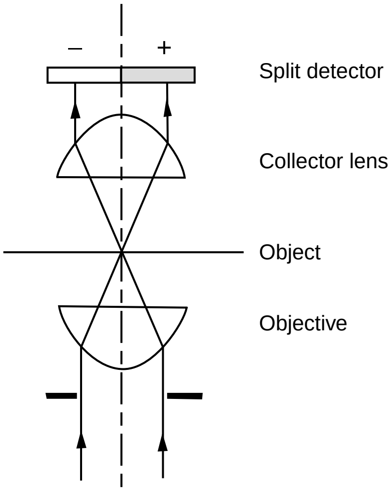

Figure 4 Schematic diagram of differential phase contrast (DPC) microscopy.

Interference Microscopy

Interference microscopy allows quantitative measurement of the object phase by interference of the object beam with a reference beam. In the transmission geometry, a Mach–Zehnder interferometer is used, which is effectively two parallel microscopes, each with matched condenser and objective lenses. In the reflection geometry we can use a Michelson, a Mirau, or a Linnik interferometer. The Michelson interferometer has an inclined beamsplitter and the reference mirror situated between the objective and object, and hence can be used only with low numerical aperture objectives. The Linnik interferometer has matched objectives in the object and reference beam paths. There is no limit to the numerical aperture that can be used, but it is not a common-path interferometer and is hence sensitive to vibrations and air currents. Probably the most practicable system for interference microscopy in the reflection geometry is based on the Mirau interferometer that uses a beamsplitter with its normal parallel to the optic axis. Until recently, interference microscopes used a small condenser aperture giving nearly coherent illumination. The image consists of three components: an object beam term (noninterference image) a reference beam term, and the interference term. The phase (and amplitude) of the object beam can be extracted using phase stepping or heterodyning techniques. The measured phase is wrapped, and requires unwrapping to obtain the absolute phase.

If the condenser aperture is opened up, the behavior is modified. The interference term is given by

$$I_{\text{int}}(x, y) = 2\Re\left\{ r^* \iint c_{\text{int}}(m, n) T(m, n) \times \exp\left[\frac{i2\pi(mx + ny)}{M} \right] \mathrm{d}m \, \mathrm{d}n \right\} \qquad [23]$$

where $c_{\text{int}}(m, n)$ is the coherent transfer function for interference imaging, given by a convolution integral of the object path objective pupil with the product of the reference pupil and the source. The first thing to note is that for a high numerical aperture, the spacing of the interference fringes is increased by 20–40%, and depends on defocus, tilt, and central obscuration. Thus accurate measurement of phase requires careful calibration. The objective pupil exhibits a phase variation when defocused that reduces the modulus of the coherent transfer function, resulting in the fringes being modulated by an envelope the width of which becomes smaller as the aperture becomes larger. The envelope is similar to that of the optical sectioning effect in confocal microscopy. This envelope allows the zero-order fringe to be identified, although simulations have suggested that fringe hopping could occur at phase jumps as a result of diffraction effects.

Multiple Beam Interferometry

The topography of a surface can be investigated by using a Fizeau interferometer formed by bringing a coated optical flat near to the surface. Then if the surface and flat have high reflectivity, multiple beam interference occurs and the width of the bright fringes can be much reduced. The system can be used with a tilted flat, to give fringes that are substantially straight, or with an almost parallel flat, so that the entire field of view is occupied by a single fringe. In the latter case, the sensitivity can be in the subnanometer range. Multiple beam interference also enhances the contrast of weak amplitude information.

Shearing Interferometry

In conventional interferometry, the object beam is arranged to interfere with a constant reference beam. In shearing interferometry, on the other hand, it interferes with a shifted version of itself.

Lateral Shearing Interferometry

An approach that is suitable for quantitative optical path measurements is to use an interferometer that shears the microscope image. The Interphako system uses a Mach–Zehnder interferometer that can measure optical path differences as low as ~3 nm in either the reflection or transmission geometry. The interferometer allows the shear distance to be varied.

Differential Interference Microscopy (DIC)

The most common form of shearing interferometry is differential interference microscopy, usually associated with the name of Nomarski. DIC microscopy can be performed in either the transmission or the reflection geometry. A plane polarized illumination beam is split into two orthogonally polarized beams traveling at slightly differing angles by a Wollaston prism in the front focal plane of the condenser lens (Figure 5). The beams are thus sheared laterally in the object plane. The light from the sample travels through the objective and a second Wollaston prism, which recombines the two beams into one. In the reflection geometry, a single lens and Wollaston prism are used. An analyzer set at 45° allows the two beams to interfere. The shear is arranged to be smaller than the width of the point spread function. Different prisms are supplied for use with different magnification objectives. The system results in

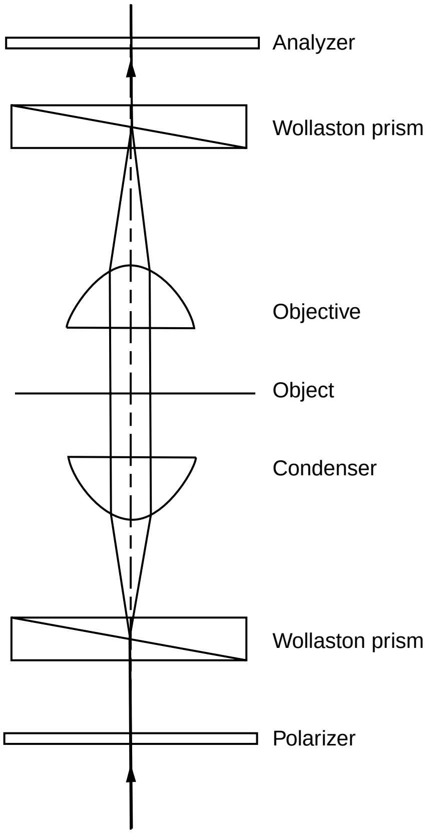

Figure 5 Nomarski differential interference contrast (DIC) microscopy.

differential phase contrast, but there are differences as compared with the DPC microscope described above.

There are two different implementations of DIC in use. In one implementation, the illumination polarizer and analyzer are crossed. The relative phase difference between the two beams (the bias retardation) is adjusted by moving the Wollaston prism laterally. It is claimed that rotation of the analyzer can be useful when observing mixed phase/amplitude objects. In the second arrangement (the Sénarmont compensator) the position of the Wollaston prism is fixed, and the analyzer rotated to change the relative phase.

A recent variation of DIC uses circularly polarized illumination. The Wollaston prism still produces two plane polarized components, but with an additional phase difference of $\pi/2$. The advantage of using circular polarization is that the prism can then be rotated to change the direction of shear.

If the shear is 2Δ, and the phase difference is $\pi - 2\phi_0$ (i.e., if $\phi_0 = 0$ then the two beams are in anti-phase), then the effective object spectrum is:

$$T_{\text{eff}}(m,n) = -2\sin(2\pi m\Delta - \phi_0)T(m,n) \qquad [24]$$

Thus the effective transfer function can be written:

$$\begin{aligned} C_{\text{eff}}(m,n;p,q) &= 4\sin(2\pi m\Delta - \phi_0)\sin(2\pi p\Delta - \phi_0) \\ &\quad \times C(m,n;p,q) \\ &= 2\{\cos[2\pi(m-p)\Delta] \\ &\quad - \cos 2\phi_0 \cos[2\pi(m+p)\Delta] \\ &\quad - \sin 2\phi_0 \sin[2\pi(m+p)\Delta]\} \\ &\quad \times C(m,n;p,q) \qquad [25] \end{aligned}$$

It is the final term of the three in braces that, because it is an odd function, results in differential phase contrast. We can now examine imaging of special forms of object. For a weak object, putting $p = 0$ as in eqn [11]:

$$\begin{aligned} C_{\text{eff}}(m,n;0,0) &= 2\{(1-\cos 2\phi_0)\cos[2\pi m\Delta] \\ &\quad - \sin 2\phi_0 \sin[2\pi m\Delta]\}C(m,n;0,0) \end{aligned} \qquad [26]$$

The first term in braces corresponds to an amplitude contrast background component because it is real and even. The second term results in DPC, because it is real and odd. For zero bias retardation, dark-field imaging results. As the bias retardation is increased from zero, the strength of both the amplitude contrast and the DPC components increase, but that of the DPC component increases more rapidly. For visual observation we require good contrast, and hence we choose a small value of bias retardation, just large enough so that the background is large enough to avoid artifacts in the DPC image. In digital microscopy, contrast is not so important as it can be enhanced digitally, and hence we can increase ϕ_0 to a value of $\pi/4$, which maximizes the strength of the DPC component. Further increase in bias retardation decreases the DPC signal, until for $\pi/2$ pure amplitude contrast results. The actual form of the weak object transfer function depends on the magnitude of the shear. The condition $m_0\Delta = 1/4$, where m_0 is the cut-off in spatial frequency, gives optimum performance for the imaging of weak objects.

The phase gradient transfer function is:

$$C_{\text{eff}}(m,n;m,n) = 4\sin^2(2\pi m\Delta - \phi_0)C(m,n;m,n) \qquad [27]$$

For the bright-field microscope the phase gradient transfer function falls off symmetrically from its value at zero phase gradient. The effect of the premultiplying factor in eqn [27] is to increase the fall-off for positive phase gradients and reduce it for negative ones. A bas-relief effect occurs, with highlighting for negative phase gradients for some values of the parameters. Increasing the value of the shear increases the highlighting effect, but, unlike in DPC, the behavior is not anti-symmetric.

Because interference is dependent on the wavelength of the light, useful and beautiful effects can be seen with DIC microscopy using white light. Usually it is regarded as a qualitative, rather than a quantitative, imaging method. However, as from eqn [14], the intensity can be written:

$$I(x,y) = 2a^2(x,y)[1 - \cos(2\pi m\Delta - \phi_0)]C(m,n;m,n) \qquad [28]$$

By recording images at different bias retardations, conventional phase stepping methods can be used to extract the phase gradient $2\pi m = \partial\phi/\partial x$.

Axial Shearing Interferometry

Instead of shearing in the transverse direction, a shear in the axial direction can be used. In the reflection geometry, the complex amplitude from the point of observation is compared with the average over a defocused spot. This method has proved to be extremely sensitive, and can detect height changes smaller than 0.1 nm.

Phase Retrieval

Digital Phase Retrieval

The phase of a single two-dimensional (2D) measurement of intensity can be retrieved by an iterative procedure, if constraints such as nonnegativity and band-limitedness can be assumed. If the modulus of both an object and its Fourier transform are known, then the phases can also be recovered. In this case, the two 2D real data sets have information content sufficient to reconstruct the 2D complex object. This approach has been extended to the case of measurement of two defocused images, using the method of phase diversity.

Imaging Using the Transport of Intensity Equation

The transport equation of intensity can be used to recover the phase of a partially coherent wavefield directly (i.e., noniteratively) from two 2D intensity measurements. If the wavefield is represented in terms of its intensity I and phase ϕ in a plane, the transport equation can be written in the paraxial approximation and in the absence of phase singularities:

$$-k\frac{\partial I}{\partial z} = \nabla_T \cdot [I\nabla_T \phi] \qquad [29]$$

where ∇_T is a 2D operator acting in the plane. This expression can be inverted to give ϕ quantitatively, if I and $\partial I/\partial z$ are known. However, it gives the phase of the image field, rather than the phase of the object, so that obtaining quantitative values for a strong object may be difficult. The transport of intensity method is valid with partially coherent illumination: if the relative condenser aperture is too great then contrast becomes weak. Further, an effect similar to the variation in fringe spacing with aperture in interference microscopy is introduced because oblique illuminating rays traverse the sample at an angle to the axis. In practice, a series of three images are usually taken at different focus positions. Then, from the quantitative phase data, various different contrast mechanisms such as DIC, Zernike phase contrast, and so on, can be simulated.

Surface Profiling Methods: Triangulation and Coherence Domain Methods

Although strictly they are not phase imaging methods, other surface profiling methods can be used to measure optical path. These could also be adapted to measure optical path in transmission. It is important also to be aware of these methods, as they could affect measurements made using other phase contrast techniques.

Apart from the phase measurement techniques described above, surface profiling methods fall into one of two categories: triangulation and coherence domain methods. An example of a technique that relies on triangulation is confocal microscopy. Confocal microscopy exhibits an optical sectioning effect similar to that described for interference microscopy. The effect is stronger for higher numerical aperture. As a reflector is scanned through focus the intensity goes through a maximum. By locating the position of the maximum, the local surface height can be obtained. In a similar way, in a confocal transmission system, introduction of an object with some optical thickness has the effect of altering the axial position of the confocal pinhole, with a resultant drop in intensity. This effect can be used to image the optical thickness of the sample.

Coherence domain methods rely on the finite coherence length of the light. In an interferometer, this results in an envelope that multiplies that arising from the nonzero relative condenser aperture. Measurement of the peak of the envelope allows local surface height to be measured. Unlike the confocal approach, the width of this coherence envelope is independent of the numerical aperture of the system.

In reflection, both the confocal and the coherence domain methods have the advantage over direct phase imaging techniques that they allow measurement of surface height changes without errors caused by the phase change on reflection from a conducting medium.

See also

Interferometry: Phase Measurement Interferometry. **Microscopy:** Interference Microscopy. **Phase Control:** Phase Conjugation and Image Correction.

Further Reading

Bennett AH, Jupnik H, Osterberg H and Richards OW (1952) *Phase Microscopy*. New York: Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge: Cambridge University Press.

Bradbury S and Evenett PJ (1996) *Contrast Techniques in Light Microscopy*. Oxford: Bios Scientific Publishers.

Françon M (1950) *Le Contraste de Phase en Optique et en Microscopie*. Paris: Revue d'Optique.

Françon M (ed.) (1952) *Le Contraste de Phase et le Contraste par Interférences*. Paris: Revue d'Optique.

Inoue S and Spring KR (1997) *Video Microscopy: The Fundamentals*, 2nd edn. New York: Plenum.

Lessor DL, Hartman JS and Gordon RL (1979) Quantitative surface topography determination by Nomarski reflection microscopy. *Journal of the Optical Society of America* 69: 357–366.

Pluta M (1989) *Advanced Light Microscopy*. Specialized Methods, vol. 2. Amsterdam: Elsevier.

Sheppard CJR and Wilson T (1980) Fourier imaging of phase information in conventional and scanning microscopes. *Philosophical Transactions of Royal Society A* 295: 513–536.

Wilson T and Sheppard CJR (1984) *Theory and Practice of Scanning Optical Microscopy*. London: Academic Press.

MODULATORS

Contents

Acousto-Optics
Electro-Optics
Modulation and Demodulation of Optical Signals

Acousto-Optics

M Gottlieb and D Suhre, Carnegie Mellon University, Pittsburgh, PA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Acousto-optics deals with the interaction of light waves with sound waves and has given rise to a large number of devices related to various laser systems for display, information handling, optical signal processing, and numerous other applications requiring the spatial or temporal modulation of coherent light. The basic principles and phenomena controlling these interactions were largely understood by the mid-1930s, but had little practical significance because very high acoustic power levels were required to attain good optical efficiency and there were few good sources of well-collimated monochromatic light. During the period from 1960 to the present, several key technologies were developing rapidly, at the same time that many applications of the laser were being suggested which require high-speed, high-resolution scanning methods. These new technologies gave rise to high-efficiency, wideband acoustic transducers capable of operation to several gigahertz, high-power wideband solid-state amplifiers to drive such transducers, and the development of a number of new, synthetic acousto-optic crystals with low-drive-power requirements and low acoustic losses at high frequencies. This combination of properties makes acousto-optics feasible for many systems, and for several is the method of choice to satisfy demanding requirements. This chapter will summarize the basic features of acousto-optic interactions and the operating principles of the most common acousto-optic devices.

The Photo-Elastic Effect

The change induced in the refractive index of a transparent material by the pressure change produced by an acoustic disturbance is the underlying mechanism of all acousto-optic interactions. An acoustic wave produces periodic regions of compression and rarefaction in the material, which modulates the density. The Lorentz–Lorenz relation relates the refractive index to the density, for the simplest case of an ideal gas

$$\frac{(n^2-1)}{(n^2+2)} \propto \rho \qquad [1]$$

where n is the refractive index and ρ is the density. This relation is followed to a good approximation for most simple solid materials as well. The elastooptic coefficient is

$$\frac{\rho\,\mathrm{d}n}{\mathrm{d}\rho} = \frac{(n^2-1)/(n^2+2)}{6n} = p \qquad [2]$$

The fundamental quantity given by eqn [2], also known as the photo-elastic constant p, is related to the pressure applied by

$$p = \frac{1}{\beta}\left(\frac{\mathrm{d}n}{\mathrm{d}p}\right) \qquad [3]$$

where P is the applied pressure and β is the compressibility of the material. The photo-elastic constant of an ideal material with refractive index of 1.5 is 0.59. The photo-elastic constants of a wide variety of materials lie in the range from about 0.1 to 0.6.

The relation in eqn [3] follows from the usual definition of the photo-elastic constant:

$$\Delta(1/\varepsilon) = \Delta(1/n^2) = pe \qquad [4]$$

where ε is the dielectric constant, $\varepsilon = n^2$, and e is the acoustic strain amplitude. The induced change in refractive index, Δn, is

$$\Delta n = -n^3 pe \qquad [5]$$

Strain amplitudes typically lie in the range of 10^{-8} to 10^{-5}, with Δn in the range of about 10^{-8} to 10^{-5} (for $n = 1.5$). Devices based upon such a small change in refractive index are capable of generating large effects because these devices are configured in a way that can produce large phase changes at optical wavelengths.

The more complete relation defining the photo-elastic interaction is more complex than the simple scalar eqn [5] in which the photoelastic constant is independent of the directional properties of the material. In fact, even for an isotropic material such as glass, longitudinal acoustic waves and transverse (shear) acoustic waves cause the photoelastic interaction to assume different parameters. A tensor relation between the dielectric properties, the elastic strain, and the photo-elastic coefficient describes the interaction, particularly for anisotropic materials. The tensor equation is

$$\Delta(1/n^2) = \sum_{kl} p_{ijkl}\, e_{kl} = \Delta(1/\varepsilon)_{ij} \qquad [6]$$

where $(1/\varepsilon)_{ij}$ is a component of the optical index ellipsoid, e_{kl} are the Cartesian strain components, and p_{ijkl} are the components of the photo-elastic tensor. The crystal symmetry of the photoelastic material places limits on the possible configurations of interaction geometry.

Diffraction by Acoustic Waves

For typical acousto-optic devices the acoustic wave acts as a diffraction grating, made up of periodic changes in optical phase, moving at sound velocity. These features determine the properties of acousto-optic diffraction. Using a quantum-based model, the light and sound may be thought of as particles, photons and phonons, undergoing collisions in which energy and momentum are conserved. Either of these descriptions may be used to obtain all the important diffraction effects, but some are more easily understood on the basis of one or the other. A description of both is given here. Consider Figure 1 in which the light wave, of frequency ω and wavelength λ, is incident into the material with an acoustic wave of frequency f and wavelength Λ. If the refractive index of the medium is $n + \Delta n$ in the presence of the acoustic wave, the phase of the optical wave will be changed by an amount

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [7]$$

where L is the length of the interaction region. Some typical values of $\Delta\phi$ can be obtained by assuming $L = 2.5$ cm and $\lambda = 0.5$ μm, with Δn reaching a peak value of 10^{-5}. This yields a phase change of π rad, which is, of course, quite large. It is large because L/λ, the number of optical wavelengths, is 50 000, so that a very small Δn can still produce a sizable $\Delta\phi$. If the electric field incident on the delay line is represented by

$$E = E_0\, e^{i\omega t} \qquad [8]$$

then the field of the phase-modulated emerging light will be

$$E = E_0\, e^{(i\omega t + \Delta\phi)} = e^{i\omega t}\, e^{2\pi i (L/\lambda)\sin(ft)} \qquad [9]$$

where f is the acoustic frequency.

A detailed derivation of the resulting temporal and spatial distribution of the light field is mathematically complex, but, analogy with radio-wave modulation can be used to arrive at the resultant fields. The spectrum of a phase-modulated carrier of frequency ω consists of components separated by multiples of the modulation frequency f, as shown in Figure 2. Sidebands are produced about the carrier frequency, such that the frequency of the mth sideband is

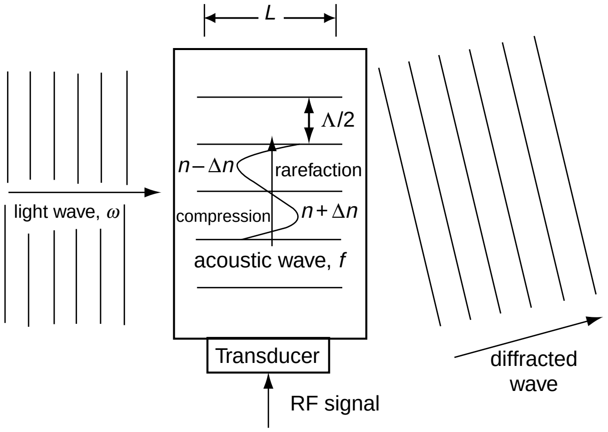

Figure 1 Interaction of light waves with acoustic waves.

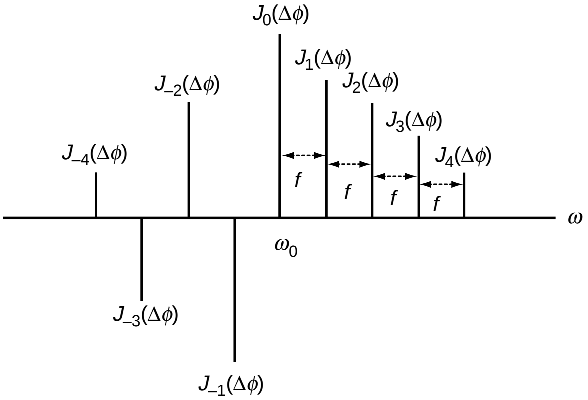

Figure 2 Spectrum of a phase-modulated wave of carrier frequency ω_0, and modulation index $\Delta\phi$.

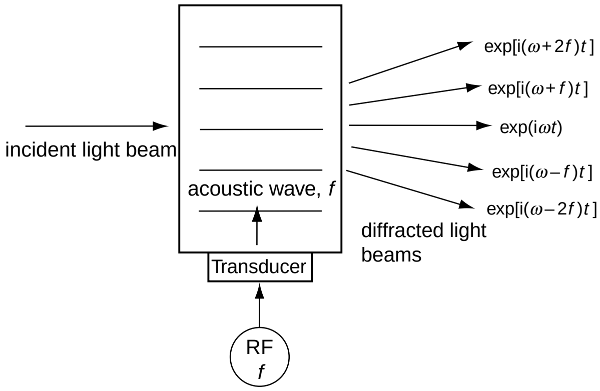

Figure 3 Diffraction of light in the Raman–Nath limit.

$\omega + mf$, where m may be positive or negative. The amplitude of each of the sidebands is proportional to the Bessel function of order equal to the sideband number, and whose argument is the modulation index $\Delta\phi$. The odd-numbered negative orders are 180° out of phase with the even-numbered ones, a feature that may be useful for certain signal processing applications. The light emerging from the delay line is composed of a number of light waves whose frequencies have been shifted by mf from the frequency ω of the incident light. The relative amplitudes will be determined by the peak change in the refractive index. The two extremes of the diffraction process are simplified in the 'thin grating' and the 'thick grating' models. The ratio of the interaction length, L, to the acoustic wavelength, Λ, will determine the character of the diffraction process. The plane wave approximation in which the acoustic energy propagates as a plane wave is valid when this ratio is very large. However, when this ratio is small, the acoustic propagation can be described in terms of a sum of plane waves, the angular spectrum of which increases as the ratio decreases. In this phenomenological model the partial wave which is propagating at an angle λ/Λ to the forward direction may diffract the light second time into an angle $2\theta = 2\lambda/\Lambda$, and the frequency of this light will once again be up-shifted, for a total frequency shift of $2f$. If the spectrum of acoustic waves contains sufficient power of still higher orders, then this rediffraction process can be repeated, so that light will be multiply diffracted m times into higher-order angles, $m\theta = m\lambda/\Lambda$ each with a frequency shift mf. A similar argument holds for the negative orders, so that a complete set of diffracted light beams will appear as shown in Figure 3, where the deflection angle corresponding to the mth order is given by $\sin\theta_m = m\lambda/\Lambda$ and the frequency of the light deflected into the mth order is $\omega_m = \omega + mf$. The intensity of the carrier wave, or zero order, will be zero when the modulation index $\Delta\phi$ is equal to 2.4. The first order will be a maximum for $\Delta\phi = 1.8$, corresponding to the maximum of the first-order Bessel function B_1, and decreasing for larger modulation indices. These phenomena were described by Debye and Sears and so are often referred to as Debye–Sears diffraction. An extensive theoretical analysis of the effect was given by Raman and Nath and is alternatively referred to as Raman–Nath diffraction. As the interaction length is increased the Raman–Nath diffraction gradually weakens. The weakening begins around an interaction length $L \sim \Lambda^2/4\lambda$. This value of L is expressed in the Q parameter

$$Q = 4L\lambda/\Lambda^2 \qquad [10]$$

which is known as the Raman–Nath parameter. A different regime of diffraction takes effect for values of the interaction length $Q \gg 1$.

The phenomena in this regime can be more easily understood in terms of the quantum description of the light and sound waves as collisions between photons and phonons. In this model, the dynamics of the collisions of the light and sound are governed by the laws of conservation of energy and momentum. The momentum vectors magnitudes of the light and sound, $\mathbf{k}$ and $\mathbf{K}$, are given by the well-known expressions

$$|\mathbf{k}| = 2\pi n/\lambda \text{ and } |\mathbf{K}| = 2\pi/\Lambda \qquad [11]$$

where the acoustic wavelength, Λ, is related to the acoustic velocity, V, by $\Lambda = V/f$.

Conservation of momentum is expressed by the vector relation

$$\mathbf{k}_\mathrm{i} + \mathbf{K} = \mathbf{k}_\mathrm{d} \qquad [12]$$

the diagram for which is shown in Figure 3, where $\mathbf{k}_i$, and $\mathbf{k}_d$ represent the momentum of the incident photon and the diffracted photon, respectively. In this notation

$$\mathbf{k}_i = 2\pi n_i/\lambda, \qquad [13a]$$

and

$$\mathbf{k}_d = 2\pi n_d/\lambda \qquad [13b]$$

If the material is anisotropic (birefringent), n_i and n_d may be different.

Conservation of energy requires that

$$h\omega_d = h\omega_i + hf \qquad [14]$$

in which h is Planck's constant. Since ω_i lies in the optical frequency range, $\sim 10^{13}$ Hz, and f lies in the RF or microwave range, 10^6–10^9 Hz, then $\omega_d \sim \omega_i$. This results, for isotropic materials, in the magnitudes of $\mathbf{k}_i$ and $\mathbf{k}_d$ being equal and resulting in the isosceles momentum triangle of Figure 4. The angles of incidence and diffraction (with respect to the planar acoustic wavefronts), called the Bragg angles, are equal for this case, and are

$$\theta_B = (1/2)\lambda/\Lambda \qquad [15]$$

A schematic diagram for this process is shown in Figure 5. The interaction will be large only if the light is well-collimated, and incident at this angle. Unlike the Debye–Sears regime, there is only a single diffracted beam. The energy of the phonon may either be added to that of the incident photon, increasing its frequency to $\omega_d = \omega_i + f$, or the reverse, resulting in $\omega_d = \omega_i - f$. The sense of the momentum vectors determines which of these occurs. The Debye–Sears effect and Bragg diffraction are not different phenomena, but are the limits of the same mechanism. The Raman–Nath parameter Q determines which is the appropriate limit for a given set of values λ, Λ, and L.

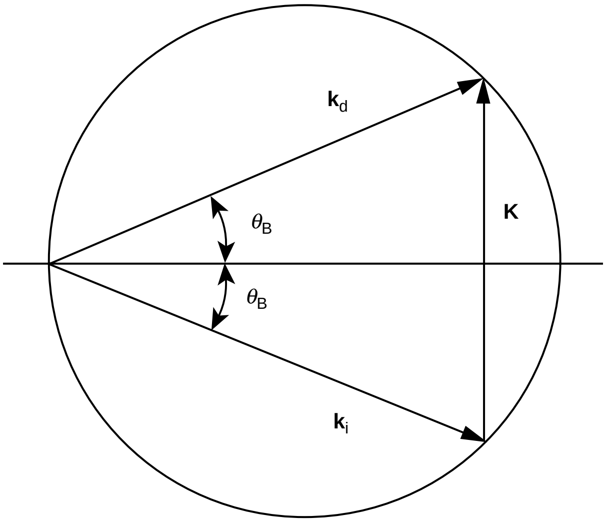

Figure 4 Momentum diagram for diffraction of light.

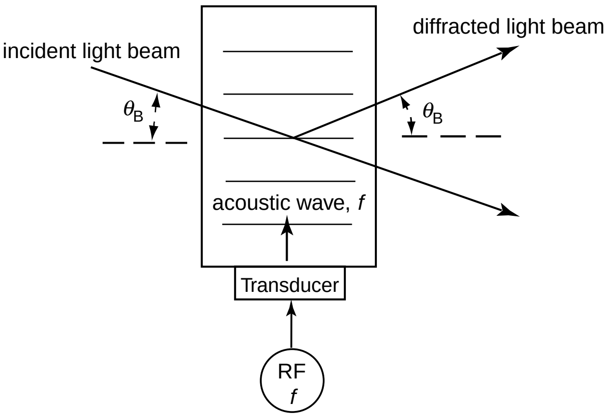

Figure 5 Acousto-optic diffraction cell.

Anisotropic Bragg Diffraction

In optically anisotropic materials, acousto-optic Bragg diffraction can occur between an ordinary and extraordinary optical beam, and vice versa. This is generally referred to as birefringent diffraction, and it offers additional design capabilities compared to the isotropic case, such as enhancing the angular aperture, and extending the aperture–bandwidth product.

The theory of diffraction of light so far described has assumed that the optical medium is isotropic. A number of important acousto-optic devices make use of the properties of birefringent materials. It is different from diffraction in isotropic media in that the magnitude of the momentum vector of the light will in general be different for different light polarization directions. Thus, the vector diagram representing conservation of momentum will no longer be the simple isosceles triangle of Figure 4. The momentum vector for ordinary polarized light will, in general, be different from the momentum vector for light that is extraordinary polarized. To understand the effect of anisotropy on diffraction, it is necessary to mention another phenomenon which occurs when light interacts with shear acoustic waves, i.e., waves in which the displacement of matter is perpendicular to the direction of propagation of the acoustic wave. A shear acoustic wave may cause the direction of polarization of the diffracted light to be rotated by 90°. The underlying reason for this is that the shear disturbance induces a birefringence which causes the plane of polarization to be rotated. The acousto-optic tunable filter (AOTF) is a particularly interesting birefringent device. It was first developed for the case of collinear optical and acoustic beams, and was used for wavelength selection in a dye laser.

Diffraction Efficiency

The acoustic power required to diffract light will be determined by the geometric parameters as well as by the properties of the medium. A simplified calculation leads to results that are useful. Referring to the spectrum of a phase-modulated wave shown in Figure 2, we can see that the ratio of the intensity in the first order to that in the zero order is

$$I_1/I_0 = [B_1(\Delta\phi)/B_0(\Delta\phi)]^2 \qquad [16]$$

By analogy this same result comes about for acousto-optically diffracted light. The acoustic power flow is given by

$$P_a = \frac{1}{2}cVe^2 \qquad [17]$$

where c is the elastic stiffness constant. The elastic stiffness constant is related to the bulk modulus β, the density ρ, and acoustic velocity v through the expression

$$c = \frac{1}{\beta} = \rho V^2 \qquad [18]$$

and the acoustic power density is

$$P_a = \frac{1}{2}\rho V^3 e^2 \qquad [19]$$

The optical phase modulation of the medium resulting from the change in refractive index, Δn, is

$$\Delta\phi = 2\pi(L/\lambda)\Delta n \qquad [20]$$

Using eqn [5] for Δn, the phase modulation is related to the acoustic power density by

$$\Delta\phi = -\pi(L/\lambda)\,n^3 p\,(2P_a/\rho V^3)^{1/2} \qquad [21]$$

The diffraction efficiency can now be calculated by using this value for the optical phase change in eqn [16]. The first- and second-order Bessel functions can be approximated for small modulation index by

$$\begin{aligned} &B_0(\Delta\phi) \sim \cos(\Delta\phi) \sim 1 - (\Delta\phi)^2, \quad \text{and} \\ &B_1(\Delta\phi) \sim \sin(\Delta\phi) \sim \Delta\phi \end{aligned} \qquad [22]$$

The small signal approximation to the diffraction efficiency is then

$$I_1/I_0 \sim (\Delta\phi)^2 = \left(\pi^2 L^2/2\lambda^2\right)\left(n^6 p^2/\rho V^3\right)P_a \qquad [23]$$

The expression $n^6 p^2/\rho V^3$ is known as the figure of merit of the material, and is designated as M_2; it is comprised entirely of intrinsic material properties.

Acousto-optic Materials

Acousto-optic device technology has matured to the point that performance is chiefly limited by material parameters, particularly the figure of merit and acoustic attenuation. Nature has arranged that materials with high figures of merit usually have high attenuation and vice versa. The widely used acousto-optic materials are fused quartz, tellurium dioxide, and lithium niobate. Development work on new infrared materials has been reported recently. A list of commonly used acousto-optic materials is given in Table 1. For materials with a low figure of merit, a higher drive power can be used to obtain the required efficiency.

Experience has indicated that the upper limit for very small devices (active area ~0.1 mm^2) is a drive power density of 100–500 mW/mm^2, provided there is proper heat sinking to transfer the heat energy. For larger devices, sizes greater than ~1 cm^2, the limit is closer to a few W/cm^2. At the high drive power levels, the acoustic attenuation may cause significant optical distortion.

AO Devices

Resolution, bandwidth, and speed are the important characteristics of acousto-optic scanners, shared by all types of scanning devices. Depending upon the application, only one, or all, of these characteristics may have to be optimized. Consider the acousto-optic scanner in Figure 6 with a collimated incident beam of width D, diffracted to an angle θ_B at its RF bandcenter, and whose bandwidth is Δf. If the diffracted beam is focused onto a plane by a lens, or lens combination, at the scanner, the diffraction spread of the optical beam will be

$$\Delta x = F\Delta\phi \sim F - \lambda/D \qquad [24]$$

Table 1 Selected acousto-optic materials

Material	*Transmission range (μm) mission*	*Acoustic mode & propagation direction*	*v (cm/s × 10^5)*	*Acoustic attenuation (dB/cm GHz^2)^c*	*n*	*M_2 (s^3/g × 10^{-18})*
Visible–near-infrared (VIS-NIR)						
$LiNb0_3$	0.04–4.5	L[100]^a	6.57	0.15	2.20	7.0
		S[001]^b	3.59	2.6	2.29	2.92
TiO_2,	0.45–6	L[001]	10.3	0.55	2.58	1.52
α-SiO_2;	0.12–4.5	L[001]	6.32	2.1	1.54	1.48
		L[100]	5.72	3.0	1.55	2.38
TeO_2.	0.35–5	L[001]	4.20	15	2.26	34.5
		S[110]	0.616	90	2.26	793
Far IR						
Ge	2–20	L[111]	5.50	30	4.00	840
$T1_3AsS_4$ (TAS)	0.6–12	L[001]	2.5	29	2.63	510
GaAs	1–11	L[110]	5.15	30	3.37	104
ZnTe	0.55–20	L[110]	3.37	130	2.77	18
GaP	0.6–10	L[110]	6.32	60	3.31	30
$T1_3PSe_4$	0.85–9	L[100]	2.0	150	2.9	2069
Te	5–20	L[100]	2.2	60	4.8	4400
CdS	0.5–11	L[100]	4.17	90	2.44	12
GaP	0.6–10	L[110]	6.32	60	3.31	30
ZnS	0.4–12	L[001]	5.82	27	2.35	3.4
		S[001]	2.63	130	2.35	8.4

[a] Longitudinal acoustic waves propagating along the [100] crystallographic direction.
[b] Shear acoustic waves propagating along the [001] crystallographic direction.
[c] Attenuation constant at 1 GHz; the frequency dependence of the attenuation for most crystals is quadratic.

where F is the focal length of the lens. The light intensity will be distributed in the focal plane with a sinc^2 distribution by aperture diffraction. The number of resolvable spots, N, will be the angular scan range divided by the angular diffraction spread,

$$N = \frac{\Delta\theta}{\Delta\phi} \qquad [25]$$

where $\Delta\theta$ is the range of the angular scan. Differentiating the Bragg angle formula yields

$$\Delta\theta = (\lambda/V \cos\theta_0)\Delta f \qquad [26]$$

and

$$N = \Delta f(D/V\cos\theta_0) = \Delta f\tau \qquad [27]$$

where θ_0 is the Bragg angle at band center, and τ is the time that it takes the acoustic wave to cross the optical aperture. The resulting expression is the time–bandwidth product of the acousto-optic scanner, a concept applied to a variety of electronic devices as a measure of information handling capacity. The time–bandwidth product of an acousto-optic Bragg cell is equivalent to the number of bits of information which may be instantaneously processed by the system. In order to maximize the number of resolution elements, it is desirable to have as large a bandwidth as possible (i.e., large frequency range) and also as large an aperture delay time as possible. There are two factors limiting the bandwidth of an acousto-optic device: the bandwidth of the transducer structure (discussed later), and acoustic absorption in the delay medium. The acoustic absorption increases with increasing frequency; for high-purity single crystals the increase generally goes with the square of the frequency. For glassy materials, on the other hand, the attenuation will increase more slowly with frequency, often approaching a linear function. The maximum frequency is generally taken as that for which the attenuation of the acoustic wave across the optical aperture is equal to

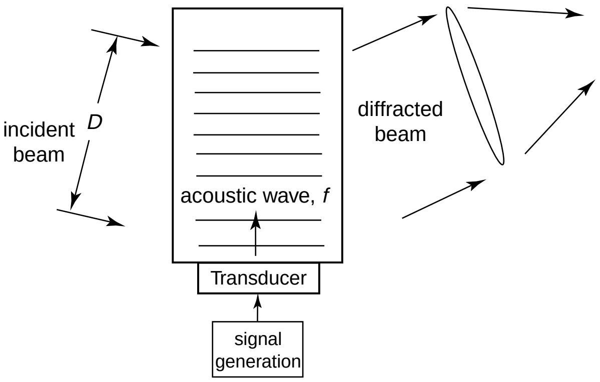

Figure 6 Acousto-optic scanner.

3 dB. A reasonable approximation of the maximum attainable bandwidth is

$$\Delta f = 0.7 f_{\max} \qquad [28]$$

Birefringent Scanners

The birefringent scanner is a significant use for anisotropic Bragg diffraction. There are a number of advantages with the birefringent scanner over the anisotropic scanner, such as a larger angular scan range along with lower frequencies. There are also disadvantages, such as lower speed due to generally lower shear wave velocities, and the particular application will dictate whether an isotropic or birefringent scanner is better. Scanners represent a fairly important aspect of acousto-optic devices due to the widespread commercial use of laser beam deflectors for displays and laser printers. For applications where the required speed is beyond that of mechanical scanners, the acousto-optic scanner is an ideal candidate. However, unlike mirrors, acousto-optic deflectors are wavelength sensitive, and can only be used with single-wavelength laser beams.

The birefringent scanner can be described with the wavevector diagram, as shown in Figure 7 for a positive uniaxial crystal where the extraordinary index of refraction is larger than the ordinary. The acoustic wavevector is tangent to the diffracted surface, which produces the largest scan angle for a given acoustic bandwidth. This is also the degenerate case, where only a single diffracted beam results, whereas two beams at two different acoustic frequencies will generally result from arbitrary input and acoustic propagation directions. The azimuthal acceptance angles (angles normal to the plane of incidence) of the acoustic and optical wavevectors can be different, although propagation and polarization directions must be selected that will produce high efficiency.

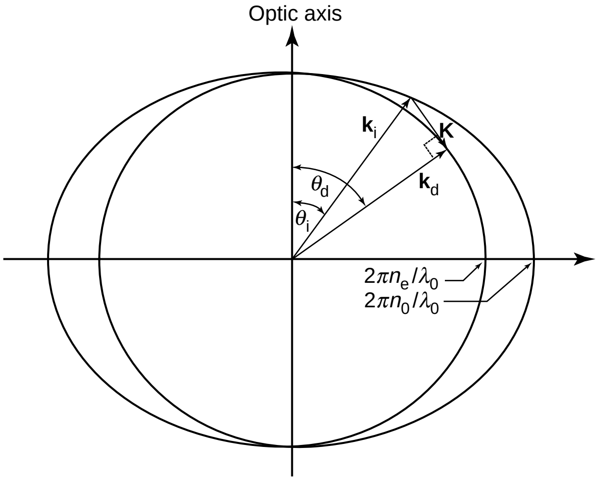

Figure 7 Wavevector diagram for birefringent scanner using a positive uniaxial crystal.

For the positive uniaxial case, the acoustic and diffracted wavevectors will be perpendicular at the design point, which allows the center frequency to be calculated from geometry as

$$f_0 = (V/\lambda)(n_i^2 - n_0^2)^{1/2} \qquad [29]$$

where n_i is the index of refraction at the incidence angle. Due to the typically low velocity of shear waves, this frequency can be much lower than an isotropic scanner, and the incidence angle can also be chosen to adjust the frequency. It is also possible to use the optical activity of certain materials such as TeO_2, along with circularly polarized light to further reduce the acoustic frequency. It is important to maintain as low a frequency as possible due to acoustic attenuation, which is especially high with the soft materials typically used for birefringent applications.

The bandwidth over which the scanner can efficiently operate is fairly large due to having the acoustic and diffracted wavevectors perpendicular, and is approximately $\Delta f = 2f_0$, assuming an octave of bandwidth for the isotropic scanner. This will produce a larger scan angle than an isotropic scanner, and more spots of resolution. The number of resolution spots is determined through diffraction by the aperture size and diffraction, along with the scan angle. Since a larger aperture requires a longer time for the acoustic waves to propagate across the aperture, the response time τ to access the scanner will increase, and the product of τ and f is related to the number of spots by eqn [27]. The advantage of the birefringent scanner is that it can operate at a lower frequency f_0 with better performance. However, since other factors such as acoustic attenuation are important in designing a scanner, for some applications it might be better to operate an isotropic scanner at a higher frequency.

AOTFs

With anisotropic Bragg diffraction, the magnitude of the diffracted wavevector $\mathbf{k}_d$ will differ from the incident wavevector $\mathbf{k}_i$, which cannot occur for the isotropic case. This is illustrated in Figure 8 for an AOTF utilizing a negative uniaxial crystal where an extraordinary input wave propagating at an incident angle θ_i relative to the crystal axis is diffracted into an ordinary output wave at an angle θ_d. This occurs through an acoustic wave propagating at an angle θ_a

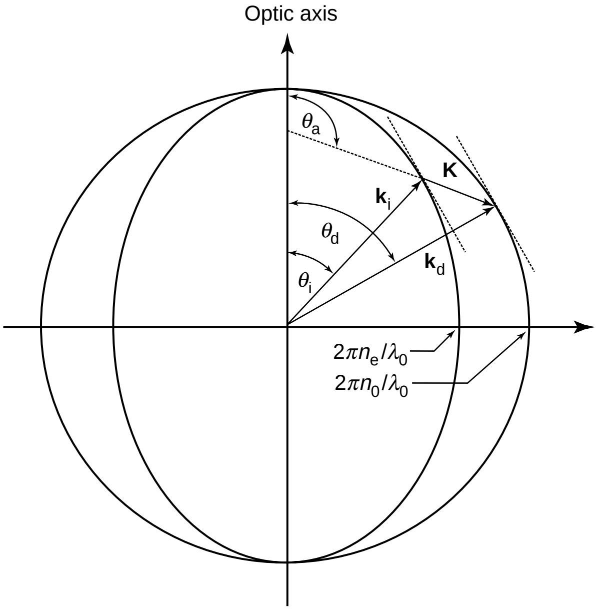

Figure 8 Wavevector diagram for an AOTF using a negative uniaxial crystal.

with wavevector K_a, where all the wavevectors lie in a plane through the optic axis of the crystal. This diagram is identical to the index ellipsoid for the crystal, scaled by $2\pi/\lambda$, where λ is the vacuum wavelength. The acoustic wavevector is shown as adding to the incident wavevector, which increases the frequency of the optical wave by the acoustic frequency. It can also be represented in the reverse direction, which would reduce the optical frequency by the same amount.

An AOTF spectrally filters the optical input, and maintains its spectral purity over a large angular aperture. These conditions are maximized when the power flow or ray directions of the input and output beams are collinear. This produces a parallel-tangents condition, where lines drawn tangent to the wavevector surfaces and connecting the input and diffracted wavevectors are parallel. For beams at 90° to the optical axis, it is referred to as the collinear case, whereas all other parallel-tangents conditions are noncollinear. The diffracted beam is rotated by 90° during anisotropic Bragg interaction, so that crossed polarizers can be used to separate the input and diffracted beams. For the collinear case polarization separation must be used, whereas angular separation can also be used for the noncollinear case.

The AOTF design equations can be derived from the geometrical conditions imposed by the wavevector matching condition

$$\mathbf{k}_d = \mathbf{k}_i + \mathbf{K} \qquad [30]$$

along with decomposing the acoustic wave into its Fourier components which result from the finite interaction length produced by the transducer. These components allow for phase matching over a range of input angles, and they also allow for a spectral spread of the interaction. The full width half maximum (FWHM) resolution is given by

$$\Delta\lambda = \frac{1.8\lambda^2}{bL \sin 2\theta_i} \qquad [31]$$

where λ is the vacuum wavelength, b is the dispersion term (essentially $2\pi\Delta n$, where the birefringence Δn is the difference between the ordinary index n_0 and extraordinary index n_e of refraction), L is the interaction length of the input beam defined geometrically by the acoustic beam, and θ_i refers to the angle of the central or cardinal ray of the input beam, and the AOTF is designed around this angle. It is therefore a sensitive function of the incidence design angle, with the highest resolution occurring for the collinear case. The resolution is also narrower for larger birefringence.

The geometry of a noncollinear AOTF is shown schematically in Figure 9. The acoustic waves propagate in the correct direction and are generated by a transducer of length L_t, which is related geometrically to the interaction length. These waves are generally absorbed at the other side of the AOTF to prevent interfering reflections, and the sides may also be wedged to eliminate reflections into the interaction region. The optical beam enters the AOTF at the correct angle to the crystal optic axis, and either diffracts or passes through as the zero order. The input beam will have an angular spread, producing an acceptance angle that is a function of the interaction length, wavelength, and crystal parameters, and the external solid angular aperture

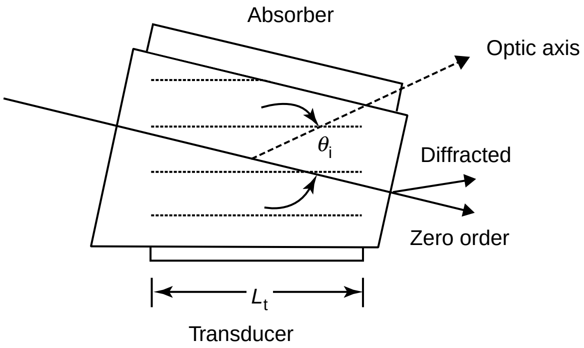

Figure 9 Noncollinear AOTF orientation of optical and acoustic beams.

is roughly given by

$$\Delta\Omega = \frac{\pi n^2 \lambda}{\Delta n L} \quad [32]$$

where n refers to the diffracted beam index of refraction, which can be either n_0 or n_e depending upon the design. The resolution and solid angle are therefore related through the transducer length, and the product of the resolving power times the solid angle is given by

$$(\lambda/\Delta\lambda)(\Delta\Omega) = 2\pi n^2 \sin^2\theta_i / |F_\theta F_\varphi|^{1/2} \quad [33]$$

where

$$F_\theta = 2\cos^2\theta_i - \sin^2\theta_i \quad [34a]$$

$$F_\varphi = 2\cos^2\theta_i + \sin^2\theta_i \quad [34b]$$

are form factors for the polar and azimuthal components of the optical field. Since all the wavectors lie in a plane containing the optic axis, the azimuthal components of the incident, diffracted, and acoustic wavevectors must always be identical under the parallel-tangents condition.

The product $(\lambda/\Delta\lambda)(\Delta\Omega)$ forms a figure of merit for spectrometers, and for the collinear case, the product is identical to a Fabry–Perot etalon having an index of n, and at other angles it is more complicated due to the angular dependence. The angular field is also symmetrical for the collinear AOTF, but nowhere else other than along the optic axis, where the resolving power becomes zero.

The great advantage of the AOTF designed under the parallel-tangents condition is the large angular aperture compared to the general case. Under this condition, the dependence of the angular aperture on resolving power is second order rather than first order, and the angular aperture can be tens of degrees for a typical resolution, which is useful for imaging applications. At the special angle of 54.7°, the second-order dependence is also zero, and the aperture can be extremely large relative to the resolving power, although this requires a high-resolution device with a corresponding narrow field, since the condition would soon be violated at larger field angles.

The efficiency of the AOTF under phase matching is given by

$$\eta = \sin^2[\pi^2 M_2 P_a L^2/(2\lambda^2)]^{1/2} \quad [35]$$

which depends on θ_i and various material properties. Since the interaction is anisotropic, M_2 must be taken as a tensor quantity, in which both the polarization and propagation directions of the acoustic and optical waves must be accounted for. For acoustic waves, the polarization direction is the particle motion, which is perpendicular to the propagation direction for shear waves. In general, M_2 is much larger for a specific configuration of a particular material, and AOTFs are designed around this condition. The range of useful design angles is also usually limited since M_2 is generally a sensitive function of θ_i. The angular dependence on M_2 is important in designing an efficient device since it can be near zero at specific design angles for some materials, such as with TeO_2 for the collinear case.

The AOTF must be designed to optimize performance. The resolution is generally given as a system requirement, and the design angle along with the transducer length must be adjusted to optimize the throughput, or total optical power through the AOTF. This requires maximizing the efficiency, angular aperture, and aperture dimension. The maximum crystal size that can be grown ultimately limits these parameters, both in the interaction length and in the aperture size.

AOTFs have been used for a wide range of spectral filtering applications for spectroscopy and laser applications. Both collinear and noncollinear devices have electronically tuned a variety of lasers, including dye, semiconductor, and Ti:sapphire lasers. The AOTF is placed within the laser cavity, which requires high transmission on the laser line, with enough resolution to separate nearby laser transitions. A related application is the multiplexing of optical communications systems, where the AOTF is used to separate the various laser diode wavelengths. Perhaps the most widely used application is spectroscopy, and single-point detection systems have been used for a variety of biological and chemical applications in the visible and infrared. These techniques have been extended to spectral imaging, again in the visible and the infrared. Due to a variety of AOTF aberrations the image quality will be somewhat degraded. Since the AOTF is electronically tunable, it can be rapidly modulated both in amplitude and wavelength, which allows for modulation spectroscopy to detect small signals in a large background. By applying multiple frequencies, the AOTF has also been used as a rejection filter, in which all wavelengths other than that selected are allowed to pass.

Modulators

Acousto-optic modulators are used to vary and control laser beam intensity. A Bragg configuration gives a single first-order output beam, where intensity is directly linked to the power of RF control signal. The rise time of the modulator is the time required for the acoustic wave to traverse the laser beam. For minimum rise time the laser beam will be focused down, forming a beam waist as it passes through the modulator.

A Bragg configuration gives a single first-order output beam, the intensity of which is controlled by the RF signal power, and diffaction angle controlled by the drive frequency. Two deflectors can be used in series and at right angles to each other to give full two-dimensional scanning.

One requirement in the design of acousto-optic modulators is that the diffracted beam and undiffracted beam must be well separated. For an adequate extinction ratio, the Bragg angle should be at least as large as the divergence of the optical beam. This condition puts a minimum value on the center frequency. Equating the Bragg angle

$$\theta_B = \frac{\lambda f_0}{2nV} \qquad [36]$$

and the diffraction angle of the Gaussian beam

$$\delta\theta_0 = \frac{4\lambda}{\pi(nd)} \qquad [37]$$

where d is the beam diameter at the minimum, it follows that the lower limit of the acoustic frequency is given by

$$f_0 = \frac{8}{\pi\tau} \qquad [38]$$

We may combine these expressions to determine a limit of modulator bandwidth of acousto-optic modulators, with the result

$$\Delta f \sim \frac{1}{4} f_m \qquad [39]$$

i.e., the modulation bandwidth is approximately equal to 25% of the midband acoustic frequency, f_m. In view of the present status of transducer technology, the modulation bandwidth of acousto-optic modulators is limited to several hundred MHz.

A frequency shifter uses the shift inherent in the acousto-optic interaction to up- or down-shift a laser's frequency. Two kinds of shifters can be distinguished: the fixed frequency shifter, and the variable frequency shifter. The frequency shift is equal to the signal frequency applied to the transducer. Various device configurations can be used to shift the laser beam, such as multiple travels inside the shifter to double or triple the frequency shift, or a combination of two frequency shifters in series. Acousto-optic frequency shifters can be used, for example, in optical heterodyning and interferometry, or in laser Doppler velocimetry for particle velocity analysis.

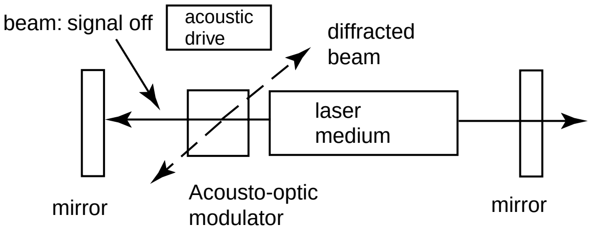

Figure 10 An acousto-optic *Q*-switch in a laser cavity.

Q-Switches

A *Q*-switch is a special modulator which introduces high repetition rate losses inside a laser cavity (typically 1 to 100 KHz). They are designed for minimum insertion loss and to be able to withstand very high laser powers. In normal use an RF signal is applied to diffract a portion of the laser cavity flux out of the cavity. This increases the cavity losses and prevents oscillation. When the RF signal is switched off, the cavity losses decrease rapidly and an intense laser pulse evolves. A diagram of a typical acousto-optic *Q*-switched laser cavity is shown in Figure 10. *Q*-switches are the preferred method of intracavity modulation where high speed, high stability, and modest cost are important design factors.

List of Units and Nomenclature

b	dispersion constant	–
c	elastic stiffness	dynes cm^{-2}
D, d	optical beam diameters	cm
e	elastic strain	–
f	acoustic frequency	hertz
F	optical focal length	cm
F_ϕ, F_θ	angular form factors	–
$\mathbf{k}$	optical momentum wavevector	cm^{-1}
$\mathbf{K}$	acoustic momentum wavevector	cm^{-1}
M_2	acousto-optic figure of merit	sec^3 gram^{-1}
n	refractive index	
p	photo-elastic constant	–
P	pressure	dynes cm^{-2}
P_a	acoustic power	watts cm^{-2}
V	acoustic velocity	cm sec^{-1}
β	compressibility	cm^2 dyne^{-1}
η	diffraction efficiency	–
ε	dielectric constant	–
ω	optical frequency	hertz
Ω	solid angle	steradians
ρ	density	grams cm^{-3}
θ	Bragg angles	radians
τ	acoustic wave travel time	sec

See also

Modulators: Acousto-Optics. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Chang IC (1976) Acousto-optic devices and applications. *IEEE Transactions: Sonics and Ultrasonics* SU-23: 2–22.

Chang IC (1981) Acousto-optic tunable filter. *Optical Engineering* 20: 824–829.

Dixon RW (1967) Photo-elastic properties of selected materials and their relevance for applications to acoustic light modulators and scanners. *Journal of Applied Physics* 38: 5149–5153.

Gordon EI (1966) A review of acousto-optical deflection and modulation devices. *Proceedings of the IEEE* 54: 1391–1401.

Gottlieb M, Ireland CLM and Ley JM (1983) *Electro-optic and Acousto-optic Scanning and Deflection*. New York: Marcel Dekker.

Goutzoulis AP and Pape DR (1994) *Design and Fabrication of Acousto-optic Devices*. New York: Marcel Dekker.

Uchida N and Niizeki N (1973) Acousto-optic deflection materials and techniques. *Proceedings of the IEEE* 61: 1073–1092.

Xu J and Stroud R (1992) *Acousto-optic Devices: Principles, Design, and Applications*. New York: Wiley.

Electro-Optics

L R Dalton, University of Washington, Seattle, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

The electro-optic (EO) effect can be described as any one of a number of phenomena that occur when an electromagnetic wave in the optical spectrum (e.g., characterized by wavelengths in the range 200 to 2000 nm) interacts with an electric field, or with matter under the influence of an electric field. Two of the most important electro-optic phenomena are the Kerr effect (discovered by John Kerr in 1875) and the Pockels effect (discovered by Friedrich Pockels in 1893), in which birefringence is induced or modified in a liquid (the Kerr effect) or a solid (the Pockels effect). Birefringence is the difference in refractive indices for light of orthogonal line polarizations, one of which is parallel to the induced optical axis. The Kerr effect involves creation of birefringence in a liquid that is otherwise not birefringent. The degree of birefringence is quadratically proportional to the applied electric field strength. Hence, the Kerr effect is frequently referred to as the quadratic electro-optic effect while the Pockels effect involves a linear dependence on the applied electric field and is referred to as the linear electro-optic effect. In considering practical applications of the electro-optic effect, a commonly encountered term is that of 'electro-optic modulator', which can be described as a device wherein a signal control element is used to modulate a beam of light. The control element is typically an electric field with frequency components in the zero (DC) to hundreds of gigahertz ($1\ \text{GHz} = 10^9\ \text{Hz}$) range (or even tens of terahertz ($1\ \text{THz} = 10^{12}\ \text{Hz}$). The modulation may be imposed on the phase, frequency, amplitude, or direction of the modulated optical beam.

Electro-optic effects are one class of second-order nonlinear optical phenomena. Other important second-order nonlinear optical phenomena include sum (e.g., second-harmonic) and difference frequency generation. Such phenomena derive from the second-order term in the power series expansion of macroscopic (material) polarization, P, in terms of applied electric fields:

$$P_i = \chi_{ij}^{(1)} E_j + \chi_{ijk}^{(2)} E_j E_k + \chi_{ijkl}^{(3)} E_j E_k E_l + \cdots \qquad [1]$$

where $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$, and $\chi_{ijkl}^{(3)}$ are the linear, second-, and third-order optical susceptibilities, respectively. Second-harmonic generation (SHG) or frequency doubling (2ω) effects, where a beam at twice the frequency of the input beam is generated, can be seen by substituting a sinusoidal field, $E_\omega \cos(\omega t - kz)$; for E_j and E_k. After using a well-known trigonometric identity, the equation for macroscopic polarization becomes (through second-order and dropping the subscript i):

$$\begin{aligned} P = {} & \chi^{(1)} E_\omega \cos(\omega t - kz) + (1/2)\chi^{(2)} E_\omega^2 \\ & \times [1 + \cos(2\omega t - 2kz)] + \cdots \end{aligned} \qquad [2]$$

The electro-optic effect can be appreciated by considering interaction of the medium with an optical field $E_\omega \cos(\omega t - kz)$ and a low-frequency field E_0. Note that although we have used the symbol zero to denote the frequency of the low-frequency electrical field, actual frequencies can extend to hundreds of gigahertz or even tens of terahertz. Again substituting (and now keeping terms to third order), the equation for polarization can be written as:

$$\begin{aligned} P &= \chi^{(1)}E_\omega \cos(\omega t - kz) + 2\chi^{(2)}E_0E_\omega \cos(\omega t - kz) \\ &\quad + 3\chi^{(3)}E_0^2E_\omega \cos(\omega t - kz) + (3/4)\chi^{(3)}E_\omega^3 \cos(\omega t - kz) \\ &= \chi_{\text{eff}}E_\omega \cos(\omega t - kz) \end{aligned} \qquad [3]$$

The equation for the nonlinear index of refraction, n, can be expressed as

$$\begin{aligned} n^2 &= 1 + 4\pi\chi_{\text{eff}} \\ &= 1 + 4\pi[\chi^{(1)} + 2\chi^{(2)}E_0 + 3\chi^{(3)}E_0^2 + (3/4)\chi^{(3)}E_\omega^2] \end{aligned} \qquad [4]$$

Denoting the linear index of refraction as n_0, the above equation can be rewritten as

$$n^2 - n_0^2 = 8\pi\chi^{(2)}E_0 + 12\pi\chi^{(3)}E_0^2 + 3\pi\chi^{(3)}E_\omega^2 \qquad [5]$$

which in turn leads to

$$\begin{aligned} n &= n_0 + (4\pi/n_0)\chi^{(2)}E_0 + (6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad + (3\pi/2n_0)\chi^{(3)}E_\omega^2 \end{aligned} \qquad [6]$$

The definition of light intensity in cgs units is $E_\omega^2 = (8\pi/cn_0)I(\omega)$, which when substituted into the above equation gives

$$\begin{aligned} n &= n_0 + (4\pi/n_0)\chi^{(2)}E_0 + (6\pi/n_0)\chi^{(3)}E_0^2 \\ &\quad + (12\pi/cn_0^2)\chi^{(3)}I(\omega) \end{aligned} \qquad [7]$$

The index of refraction can now be written as

$$n(\omega) = n_0(\omega) + n_1(\omega)E_0 + n_2(0)E_0^2 + n_3(\omega)I(\omega) \qquad [8]$$

where the terms $n_1(\omega)$, $n_2(0)$, $n_3(\omega)$ correspond to the linear electro-optic effect, the quadratic electro-optic effect, and the optical Kerr effect, respectively. The above equation can also be expressed as

$$n = n_0 - (1/2)n^3rE_0 - (1/2)n^3sE_0^2 - \cdots \qquad [9]$$

where r and s are the linear and quadratic electro-optic coefficients, respectively.

Unfortunately, a great deal of confusion exists concerning the use of the terms electro-optic effect and electro-optic modulator. First of all, the terms 'electro-optic' and 'opto-electronic' are frequently confused and used interchangeably. Opto-electronic devices function as electrical-to-optical or optical-to-electrical signal transducers. Examples of commonly used opto-electronic devices include light-emitting diodes (LEDs), photodiodes (PDs), injection laser diodes (ILDs), and integrated optical circuit (IOC) elements. An electro-optic device can also function as an electrical-to-optical signal transducer (e.g., a Mach–Zehnder interferometer or a birefringent modulator employing polarizers at the input and the output); however, the mode of operation is entirely different for electro-optic and opto-electronic devices. For example, with LEDs or modulated lasers, the applied electric field produces modulation of transmission of light by altering the excited state population of the light emitting (lasing) state of the material. With electro-optic materials, no actual excited state population is generated. Electro-optic devices typically operate in regions of the optical spectrum removed from resonant transitions, i.e., regions of relatively high transparency. The applied electric field acts only to perturb the charge distribution of the material (the spatial positions of the electrons and nuclei). The altered charge distribution interacts with the optical beam transmitting the material with the result that the speed of light in the material (i.e., the index of refraction or birefringence) is altered. The response times of opto-electronic and electro-optic phenomena (and hence bandwidths of devices exploiting these phenomena) are quite different. In the former case, response will be limited by the lifetime of the relevant (emitting) excited state while in the latter case, response will be defined by the relaxation time of the material (e.g., reorientation time of a liquid or lattice relaxation time of a solid) following removal of the perturbing electric field.

An even greater confusion can arise due to the jargon used in particular technologies, such as telecommunications. A telecommunications engineer frequently refers to 'electro-optic switching' when describing opto-electronic transduction of an optical signal to an electrical signal, followed by re-routing in the electrical domain, and finally opto-electronic transduction of the electrical signal back to the optical domain. Electro-optic switching (in the sense used in this article) involves quite a different operation and would be described by that same telecommunications engineer as 'all-optical switching' – a term reserved by physicists for the optical Kerr effect (control of one light beam by another light beam).

A second point of confusion involves distinguishing between the terms 'electro-optic' and 'electro-absorptive' modulation. Again, the terms are frequently used interchangeably. However, they involve quite different physical mechanisms. An electro-absorptive modulator is, like a modulated laser or LED, a 'resonant' device. In a device such as a gallium arsenide (GaAs) or indium phosphide (InP) quantum dot electro-absorptive modulator, the position and width of an optical absorption (resonant transition) are defined by the physical dimensions of the quantum dot. Application of an electric field shifts the spectral wavelength of the quantum dot absorption. This phenomenon can be used to dramatically change optical transmission through (or equivalently, absorption by) the material by choosing the wavelength of the propagating beam of light to be near the resonant absorption. With the electric field off, the material is reasonably transparent but becomes strongly absorbing when the electric field is applied, shifting the resonant absorption to the wavelength of the propagating beam. Obviously, the voltage required to achieve a desired change in optical transmission will depend on control of quantum dot dimensions in fabrication, which has been and continues to be a topic of research focus for electro-absorptive materials. Likewise, control of insertion losses of such devices is a concern as the wavelength used for the transmissive state of the device must be sufficiently close to the optical resonance to achieve the desired attenuation with application of a modest control voltage. Another issue to be faced with electro-absorptive devices is that of 'chirp' (optical signal distortion arising from the fact that both absorption and index of refraction – the imaginary and real parts of the optical susceptibility – change with application of the electric control field). On the positive side, electro-absorptive modulators are 'polarization-insensitive' modulators and are thus conveniently used with multimode (as well as single mode) optical transmission.

Further confusion exists because widely different types of materials can be used in electro-optic and electro-absorptive devices. As these materials are frequently competing for the same applications (signal transduction, optical switching, etc.), the advantages and disadvantages of different materials are often compared without maintaining the context that quite different phenomena are involved. Even restricting our discussion to electro-optic materials, we note that materials can range from organic liquid crystalline materials (nematic, smetic, ferroelectric, etc.) to organic electro-optic materials (both crystalline and 'macromolecular') to inorganic crystalline materials (lithium niobate, $LiNbO_3$; barium titanate, $BaTiO_3$; barium borate, BBO; potassium dihydrogen phosphate, KDP; lithium tantalate, $LiTaO_3$; zinc telluride, ZnTe, etc.). The response times of these materials to transient application of an applied electric field will be quite different, reflecting the different types of lattice motion involved. With liquid crystalline materials, particularly nematic materials, considerable molecular reorientation is involved and due to the mass that must be moved, response times are quite slow although the index change will be relatively large. With use of liquid crystalline materials, device bandwidths will typically be limited to tens of megahertz (MHz) or less. In contrast, conjugated π-electron organic materials typically exhibit response times of tens to hundreds of femtoseconds, which translates to potential device bandwidths of tens of terahertz (actual device bandwidths may be less due to factors other than material response). For conjugated π-electron organics, the response time is the phase relaxation time of the π-electron system. Because only a slight change in bond length alternation of the conjugated π-electron system occurs with application of an applied electric field and because strong electron–phonon coupling and resonance stabilization of the π-electron system act to reduce the barrier to lattice relaxation, the response times of π-electron organic materials are among the fastest observed in nature. With crystalline inorganic materials, the ionic constituents move to new locations with application of an electric field with the exact movement determined by the field strength, the charge on the ions, and the restoring force. Unequal restoring forces along the three mutually orthogonal (perpendicular) axes of the crystal lead to anisotropy in the optical properties of the material. The applied electric field will induce a change in the anisotropy (the principal refractive indices). The symmetry of the electro-optic tensor will be closely related to the symmetry of the piezoelectric tensor. The linear electro-optic effect requires that the crystal exhibit noncentrosymmetric (acentric or ferroelectric) symmetry. A centrosymmetric crystal possesses a center of symmetry defined by identical particles existing in the lattice at vectors $\mathbf{r}$ and $-\mathbf{r}$, where $\mathbf{r}$ is a position vector measured from an arbitrary origin. A centrosymmetric crystal, like an isotropic liquid or gas, can exhibit a quadratic electro-optic effect.

For the sake of completeness, it can also be noted that index of refraction changes can be induced by acoustic waves (acousto-optic modulation) and by heating (thermo-optic modulation). Also, elasto-optic and photo-elastic effects can produce index of refraction changes.

To keep this article to a manageable length, discussion is limited to solid-state electro-optic materials. Design principles being used to produce new organic electro-optic materials will be illustrated. Since inorganic materials exist as crystals, limited chemical modification of such materials is possible, although processing techniques for fabricating thin films of such crystalline materials have been developed in a number of cases. For organic macromolecular materials, development of new materials is an on-going activity. Several representative devices being fabricated, using electro-optic materials, will be discussed. Four of the most commonly encountered types of EO devices include: (i) stripline devices; (ii) prism, cascaded prism, and superprism devices; (iii) resonated devices, such as ring microresonator and photonic bandgap devices, and (iv) waveplates. Of course, devices such as stripline devices can be configured in a variety of ways to produce polarization-sensitive and polarization-insensitive electrical-to-optical signal transducers, optical switches either using simple 1 × 2 or 2 × 2 switch architectures or multimode interference (MMI) switches, optical gyroscopes, photonically controlled phased array radar systems, spectrum analyzers, optical digital signal processors, analog-to-digital (A/D) and digital-to-analog (D/A) signal converters, electromagnetic (radiofrequency to millimeter wave) signal generators, voltage sensors, etc. Prism-type devices are typically used for spatial light modulation or laser beam deflection. Ring microresonator devices afford a rich array of applications ranging from active wavelength division multiplexing (WDM) to active optical interconnect reconfiguration, to voltage-controlled wavelength tuning of laser outputs.

Basic Relationships

The effect of the applied field is to deform the index of refraction ellipsoid, which can be represented as

$$\left(\frac{1}{n^2}\right)_1 X^2 + \left(\frac{1}{n^2}\right)_2 Y^2 + \left(\frac{1}{n^2}\right)_3 Z^2 + 2\left(\frac{1}{n^2}\right)_4 YZ + 2\left(\frac{1}{n^2}\right)_5 XZ + 2\left(\frac{1}{n^2}\right)_6 XY = 1 \qquad [10]$$

leading to the corrections

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^{3} r_{ij} E_{0j} \qquad [11]$$

where the electro-optic tensor components are related to the tensor components of the second-order nonlinear optical susceptibility by $r_{ij} = -(8\pi/n_{0i}^2 n_{0j}^2)\chi_{ji}^{(2)}$. The full tensorial equation for change of the 'indicatrix' (index of refraction ellipsoid) can be expressed as

$$\begin{pmatrix} \Delta\left(\frac{1}{n^2}\right)_1 \\ \Delta\left(\frac{1}{n^2}\right)_2 \\ \Delta\left(\frac{1}{n^2}\right)_3 \\ \Delta\left(\frac{1}{n^2}\right)_4 \\ \Delta\left(\frac{1}{n^2}\right)_5 \\ \Delta\left(\frac{1}{n^2}\right)_6 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \\ r_{41} & r_{42} & r_{43} \\ r_{51} & r_{52} & r_{53} \\ r_{61} & r_{62} & r_{63} \end{pmatrix} \begin{pmatrix} E_{0X} \\ E_{0Y} \\ E_{0Z} \end{pmatrix} \qquad [12]$$

The electro-optic tensor reduces considerably for specific materials reflecting the symmetry of the particular material. For inorganic crystalline materials, the electro-optic tensor can be reasonably complex (containing many nonzero elements). It is also helpful to note that for crystalline inorganic materials, the electric-field-induced change in crystal shape leads to strain and the orientation of the indicatrix is altered, leading to an additional contribution to the change in index coefficients:

$$\Delta(1/n^2)_i = \sum_k p_{ik} S_k + \sum_j r_{ij} E_j \qquad [13]$$

where S_k is a component of the strain and p_{ik} is the elasto-optic tensor. At high frequencies, the inertia of the crystal prevents macroscopic straining so that the first term of the above equation vanishes. At low frequencies, elasto-optic effects cannot be ignored. Because the deformation leading to strain is generally caused by the inverse piezoelectric effect, it is linearly related to the applied electric field (the same dependence as the linear electro-optic effect). This can lead to a frequency dependence of the 'effective' electro-optic coefficient for crystalline materials.

A brief comment on the photo-elastic effect (the change in index coefficients produced directly by applied stress) is warranted. This effect has the form:

$$\Delta(1/n^2)_i = \sum_l \pi_{il} \sigma_1 \qquad [14]$$

where σ_l are the components of the stress and π_{il} are the piezo-optical coefficients. Note that this effect is independent of the applied electric field.

For axially symmetric 'charge-transfer type' organic chromophores prepared by deposition or electric field poling, only two unique tensor elements, r_{33} and r_{13}, survive:

$$\tilde{r} = \begin{pmatrix} 0 & 0 & r_{13} \\ 0 & 0 & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{13} & 0 \\ r_{13} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad [15]$$

The electro-optic effect for organic macromolecular materials derives from molecules with conjugated π-electrons confined within molecules (chromophores). The electro-optic tensor elements can be related to molecular first hyperpolarizability, β, and the chromophore number density, N (molecules/cm^3), by

$$\begin{aligned} r_{33} &= -2Nf(0)\beta\langle\cos^3\theta\rangle/n_0e^4 \\ r_{13} &= -Nf(0)\beta\langle\sin^2\theta\cos\theta\rangle/n_{0o}^2 n_{0e}^2 \end{aligned} \qquad [16]$$

The expressions in brackets are 'order parameters' describing the degree of noncentrosymmetric order. $\langle\cos^3\theta\rangle = 1$ corresponds to all 'dipolar' (charge transfer type) molecules pointing in the same direction (perfect noncentrosymmetric or ferroelectric order) while $\langle\cos^3\theta\rangle = 0$ corresponds to complete disorder and no electro-optic effect. The reason that three angles appear in the order parameter is that an angle is required to represent the principal symmetry axis for the chromophore, the principal symmetry axis of the optical field, and the principal symmetry axis of the electric control field. Thus, averaging (denoted by brackets) must be carried out over each of these three angles. The tensor element r_{33} corresponds to the electric control field applied along the principal axis of the ordered chromophores and the principal axis of the optical field vector, transverse magnetic (TM) polarized light. The tensor element r_{13} corresponds to the control electric field applied along the principal axis of the optical field, transverse electric (TE) polarized light, and orthogonal to the principal chromophore axis. The notation r_{33} is contracted notation for r_{333}. The factor $f(0)$ takes into account the dielectric nature of the medium into which the chromophores (molecules with large first hyperpolarizability) are embedded.

For organic electro-optic materials, elasto-optic effects do not appear to make significant contributions. Moreover, electro-optic activity can be systematically improved by improving β and by optimizing the product of chromophore number density and order parameter. Due to the presence of intermolecular electrostatic interactions, W, among chromophores, order parameters and number density are not independent, e.g., for materials that are prepared by electric field poling the order parameter relevant to the principal component of the electro-optic tensor can be expressed approximately as

$$\langle\cos^3\theta\rangle = L_3(\mu f(0)E_\mathrm{p}/kT)[1 - L_1(W/kT)^2] \qquad [17]$$

where L_3, and L_1 are Langevin functions, μ is the chromophore dipole moment, E_p is the strength of the electric poling field, k is the Boltzmann constant, and T is the Kelvin poling temperature. W is a quadratic function of chromophore number density, N. The above equation indicates that a maximum will be observed in the graph of r_{33} versus N.

Materials

With inorganic and organic crystalline materials, very little can be done to optimize electro-optic activity other than discovering new EO crystalline materials or employing isotopic (e.g., deuterium for hydrogen) or ion substitution with existing materials.

As with organic liquid crystalline materials, the electro-optic activity of macromolecular organic second-order nonlinear materials can be systematically improved by molecular modification. As shown in Figure 1, quantum mechanical calculations can provide useful guidance as to which structural modifications will lead to improvements in molecular (first) hyperpolarizability and ultimately electro-optic activity. The molecular hyperpolarizability (long wavelength limit) can be increased by a factor of two by simple variation of the acceptor structure. The calculated values of wavelengths for the interband charge transfer transitions are 390 nm, 403 nm, and 430 nm. Since dipole moments do not change significantly with structure, intermolecular electrostatic interactions should be similar for these three chromophores and the improvement in β should translate to an improvement in electro-optic activity. This theoretical prediction has been experimentally verified. This is just one example of use of quantum mechanics to guide the improvement of electro-optic activity for organic materials. This figure also illustrates the typical structure of an organic electro-optic chromophore, which consists of an electron donor region (an amine donor in the example shown), a connective conjugated π-electron bridge, and an electron deficient electron acceptor region. The theoretically predicted (by density functional theory, DFT, calculations) variation of molecular first hyperpolarizability with molecular structure has been verified for the structures shown. However, the critical point is that electro-optic activity of organic materials (currently exhibiting a maximum value of 182 pm/V at 1.3 microns wavelength) is likely to be improved for some time to come. This value is to be compared to a value of 32 pm/V for lithium niobate. Another factor influencing the value of macroscopic electro-optic activity realized for electrically poled organic materials is the effect of intermolecular electrostatic interactions on optimization of $N\langle\cos^3\theta\rangle$. Statistical mechanical calculations can provide guidance as to how this is achieved (see Figure 2 where the role of chromophore shape is illustrated). Approximating the chromophore shape by a rigid prolate ellipsoid with embedded dipole moment yields nearly quantitative reproduction of the experimental data. If the π-electron structure (dipole moment) is left unchanged but the shape of the chromophore is altered to a spherical shape, a significant improvement in electro-optic activity is predicted. The shape effect arises because there are two components to the electronic dipole–dipole intermolecular electrostatic potential function describing the many body interaction of chromophores.

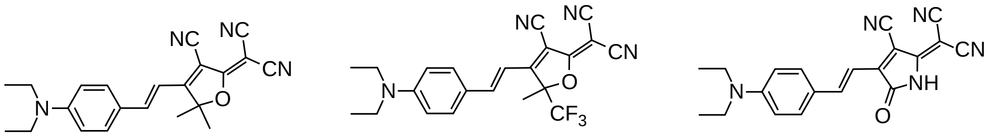

	1	2	10	
Dipole moment:	11.8	11.9	11.7	Debye
Hyperpolarizability	23.8	34.4	43.0	$*10^{-30}$ esu

Figure 1 Density functional calculations (DFT) of molecular first hyperpolarizability (β) and dipole moment (μ) for three chromophore structures.

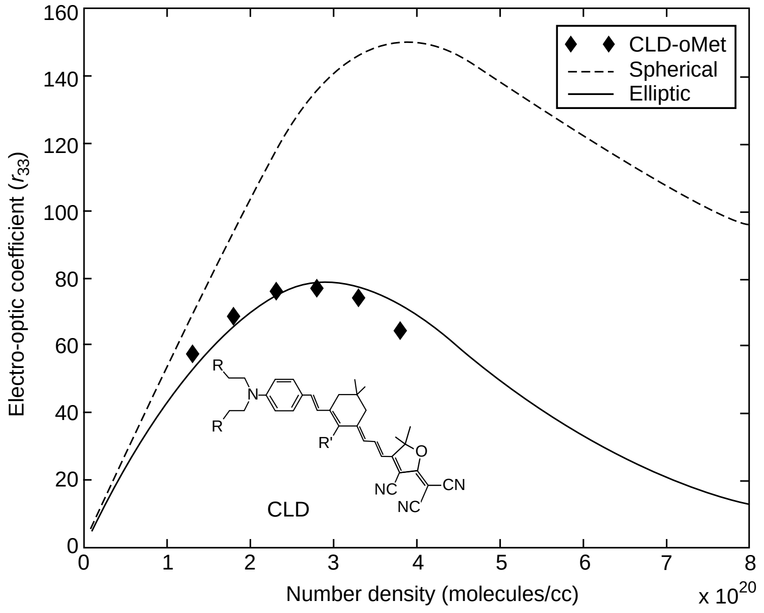

Figure 2 Two theoretical (statistical mechanical) simulations of the variation of electro-optic activity with chromophore number density are shown and compared with experimentally observed results for the CLD chromophore (structure shown) dissolved in an amorphous polycarbonate host polymer.

One component favors centrosymmetric ordering while the second component favors noncentrosymmetric ordering. Their relative weighting in defining macroscopic order depends on chromophore shape. Theoretical calculations have also shown that embedding chromophores in nanoscopic macromolecular objects (such as dendrimers and dendronized polymers) can dramatically enhance electro-optic activity. This type of engineering can also be used to control auxiliary properties such as optical loss and stability. The exceptional processability of organic electro-optic materials has permitted conformal and flexible devices to be fabricated and has permitted electro-optic circuitry to be integrated vertically on top of semiconductor VLSI (very large-scale integration) electronics. Mach–Zehnder modulators, with 3 dB operational bandwidths of 200 GHz, have been fabricated from organic electro-optic materials as have Mach–Zehnder devices with operational drive voltages of less than one volt. The major advantage that crystalline inorganic materials possess, relative to macromolecular organic EO materials, is superior thermal and photostability. A comparison of lithium niobate and the best organic EO materials, is given in Table 1. Cost is a significant factor if electro-optic materials are to compete with modulated lasers and other opto-electronic devices for applications in communications, computing, transportation, and defense. Currently, manufacturing costs for both crystalline and macromolecular materials are very high (several thousand dollars per modulator). Recently, it has been demonstrated that organic macromolecular materials can be used with soft lithography techniques to mass-produce electro-optic circuitry. This could conceivably dramatically reduce manufacturing costs. Also, the fact that organic macromolecular materials can be directly integrated with silicon photonics and VLSI semiconductor electronics, may influence future commercial utilization of electro-optic materials, which is currently dominated by lithium niobate.

Table 1 Comparison of electro-optic performance for inorganic and organic EO materials at 1.3 microns wavelength

Parameter	*$LiNbO_3$*	*NLO polymer*
Effective EO coefficient, r_{eff}, pm/V	32	182
Index of refraction (n)	2.2	1.6–1.7
Dielectric permittivity (ε)	30	3
Material Figure of Merit, $n^3(r_{\text{eff}})/\varepsilon$	10	270
Bandwidth × length product, GHz cm	7	>100
$V_\pi L$, V cm	5	1.5
Optical loss, dB/cm	0.2	0.2–1.0
Processing temperature (°C)	1000	<200
Multiple layers possible	No	Yes

Devices

Representative simple device structures are shown in Figure 3. For a simple stripline phase (birefringence) modulator (Figure 3, upper left), the phase retardation produced by application of an electric field is given by

$$\Delta\phi = 2\pi\Delta n L/\lambda = \pi n^3 rEL/\lambda = \pi n^3 rVL/\lambda d \qquad [18]$$

where L is the electrode length (the distance over which the electrical and optical fields co-propagate),

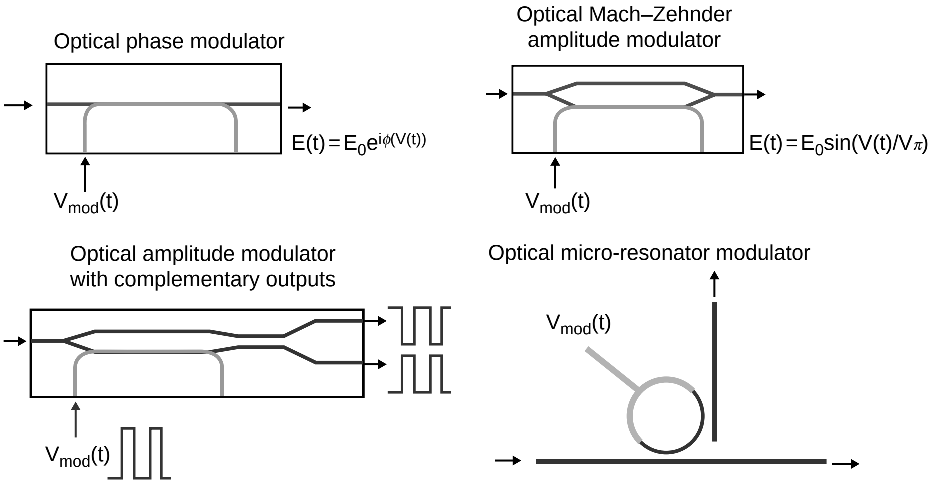
Optical phase modulator
$E(t) = E_0 e^{i\phi(V(t))}$
$V_{mod}(t)$
Optical Mach–Zehnder amplitude modulator
$E(t) = E_0 \sin(V(t)/V_\pi)$
$V_{mod}(t)$
Optical amplitude modulator with complementary outputs
$V_{mod}(t)$
Optical micro-resonator modulator
$V_{mod}(t)$

Figure 3 Schematic representation of simple electro-optic device structures are shown: Upper left (birefringent modulator), upper right (Mach–Zehnder interferometer), lower left (1 × 2 optical switch), lower right (ring microresonator).

λ is the optical wavelength, $V = E/d$ is the electric field felt by the material, and d is the electrode spacing. Electro-optic device performance is frequently characterized by the voltage, V_π, required to produce a phase shift of π radians. For a TM polarized light propagating through a birefringent modulator:

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \qquad [19]$$

where Γ characterizes the overlap of the optical and electrical waves.

A simple Mach–Zehnder interferometer (electro-to-optical signal transducer) is shown in the upper right of Figure 3. When such a device is constructed from electrically poled organic electro-optic chromophores and a voltage is applied to one arm, a retardation of light propagating in that arm will be effected. This, in turn, will result in a change in the constructive/destructive interference of the beams at the output of the device. The consequence of applying an electrical signal to one arm of a Mach–Zehnder interferometer is that the applied voltage is transduced onto the optical output beam as an amplitude modulation of that beam. The voltage required to achieve full wave modulation for an optical beam of TM polarization is

$$V_\pi = \lambda d/(n_{0e}^3 r_{33} L\Gamma) \qquad [20]$$

Amplitude modulation can also be achieved using a birefringent modulator by placing polarizers at the input and output of the device. If an input polarizer is used to launch TM and TE modes, the birefringent modulator functions to produce a rotation of light, which is turned into an amplitude modulation by inserting a polarizer at the output. For such a device, V_π depends upon the difference in r_{33} and r_{13}. Note that for poled organic chromophores, $r_{33} \approx 3r_{13}$, $(V_\pi)_{\mathrm{BRM}} \approx 1.5(V_\pi)_{\mathrm{MZM}}$. For the above devices, it is clear that V_π will depend on the length over which the electrical and optical fields co-propagate in phase. For low electrical field frequencies, L will be the electrode length. The electrode spacing d, that can be used, will depend on the index of refraction difference between the core (electro-optic material) and a cladding material used to prevent the optical field experiencing the metal electrodes (an event that would dramatically increase propagation loss). For poled organic macromolecular materials and standard commercial claddings, d is typically in the order of 10 microns. Electrode lengths usually range from 1 to 3 cm. Two factors limit bandwidth in stripline devices: (i) velocity mismatch between the electrical and optical waves (which relates to the difference between n_0^2 and ε, where ε is the dielectric permittivity of the material – see Table 1); and (ii) microwave and millimeter wave loss in the metal electrodes. For organic electro-optic materials, the latter defines operational 3 dB bandwidth of devices.

Optical routing switches (see Figure 3, lower left) are more complicated to understand and typically require somewhat larger V_π voltages to effect a switching operation (e.g., $(V_\pi)_{\mathrm{simple2\times2crossbar}} \approx 1.7(V_\pi)_{\mathrm{MZM}}$).

Two conditions must be satisfied for light to be coupled into a ring microresonator device (see Figure 3, lower right): (i) the circumference of the ring must be a multiple of the wavelength of light, and (ii) the velocity of light in the ring must equal the velocity of light at the input. If a ring microresonator is fabricated from an electro-optic material, then an applied electrical voltage can be used to control coupling of light into and out of the ring microresonator. The same electrode structure, that is used to achieve voltage-controlled wavelength selective filtering, can also be used to transduce electrical information onto the optical beam while it is in the ring resonator. The quality (Q) factor of a ring microresonator defines both the wavelength selectivity and the bandwidth of the device. Ring microresonator devices also have the advantage of permitting a long interaction length to be achieved in a very compact device structure. The dimensions (radii) of ring microresonator structures are defined by bending loss and thus by the difference between core and cladding materials. Typical radii can range from hundreds of nanometers to hundreds of microns. With ring microresonators, a reduction in drive voltage requirements can be achieved by trading off bandwidth. Another price paid in the use of ring microresonator device structures is that of bending loss, which is not present in the other device structure shown. Ring microresonators can be used for active wavelength division multiplexing (WDM), voltage-controlled optical signal routing, and wavelength tuning of optical sources.

The interaction length of a prism optical beam deflector can be increased by cascading prisms together. For such a device structure, the angle of deflection becomes

$$\theta = n_0^3 r_{33}(V/d)(L/h) \qquad [21]$$

where h is the height of an individual prism and L is the length of the prism array. For currently available electro-optic materials only small deflection angles (a few degrees or less) can be achieved with practical voltage levels. Superprism structures may permit large deflection angles to be achieved but such devices have yet to be demonstrated in a convincing manner.

See also

Interferometry: Overview. **Modulators:** Acousto-Optics. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Dagli N (1999) Wide-bandwidth lasers and modulators for RF photonics. *IEEE Trans. Microwave Theor. Techniques* 47: 1151–1171.

Dalton LR (2003) Rational design of organic electro-optic materials. *Journal of Physical Condensed Materials* 15: R897–R934.

Dalton L, Harper A, Ren A, *et al.* (1999) Polymeric electro-optic modulators: from chromophore design to integration with semiconductor very large scale integration electronics and silica fiber optics. *Ind. Eng. Chem. Res.* 38: 8–33.

Hornak LA (1992) *Polymers for Lightwave and Integrated Optics*. New York: Marcel Dekker.

Khazaei HR, Wang WJ, Berolo E and Ghannouchi F (1998) High speed GaAs/AlGaAs traveling wave electro-optic modulators. *Proceedings of SPIE* 3278: 94–100.

Lawetz C, Cartledge JC, Rolland C and Yu J (1997) Modulation characteristics of semiconductor mach-zehnder optical modulators. *IEEE Journal of Lightwave Technology* 15: 697–702.

Lee KS (2002). *Advances in Polymer Science*, vol. 158. Heidelberg: Springer.

Lee M, Katz HE, Erben C, *et al.* (2002) Broadband modulation of light using an electro-optic polymer. *Science* 298: 1401–1403.

Nalwa HS and Miyata S (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton, FL: CRC Press.

Noguchi K, Mitomi O and Miyazawa H (1998) Millimeter-wave Ti:LiNbO$_3$ optical modulators. *IEEE Journal of Lightwave Technology* 16: 615–619.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Shi Y, Zhang C, Zhang H, *et al.* (2000) Low (sub-1 volt) halfwave voltage polymeric electro-optic modulators achieved by controlling chromophore shape. *Science* 288: 119–122.

Wise DL, Wnek GE, Trantolo DJ, Cooper TM and Gresser JD (1998) *Electrical and Optical Polymer Systems. Fundamentals, Methods and Applications*. New York: Marcel Dekker.

Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.

Modulation and Demodulation of Optical Signals

R A Minasian, University of Sydney, Sydney, NSW, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since earliest times, light has been used for communication purposes. In order to convey information, a physical variable of light, such as the optical intensity or field amplitude, must be modulated to encode a signal onto the optical carrier, and it must be demodulated or decoded after transmission. Laser sources provide the capability of extremely high modulation speeds, which result in high information-carrying capacity in optical communication systems. After transmission, semiconductor photodetectors demodulate the optical signals to recover the original information (*see* **Optical Communication Systems:** Basic Concepts).

The modulation of optical signals can be classified into two types: field modulation and intensity modulation, which is illustrated in Figure 1. In field modulation, the physical variable that is modulated is the frequency, phase, or amplitude of the optical field. These three techniques are sometimes referred to as frequency shift keying (FSK), phase shift keying (PSK) and amplitude shift keying (ASK). Figure 1b shows an example of frequency modulation of light. Field modulation requires keeping track of the phase of the carrier, which imposes strict requirements on the optical coherence (*see* **Coherent Lightwave Systems**). Because the optical carrier is at an extremely high frequency of around 2×10^{14} Hz at fiber-optic communication wavelengths of 1.5 μm, the required optical coherence is very difficult to realize, and moreover the receiver requires complex optical heterodyne methods to detect the phase of the optical field.

In intensity modulation, the physical variable that is modulated is the optical power. Figure 1c shows that it is the envelope of the light power that contains the information. The optical field oscillations at 10^{14}–10^{16} Hz in this case are not related to the

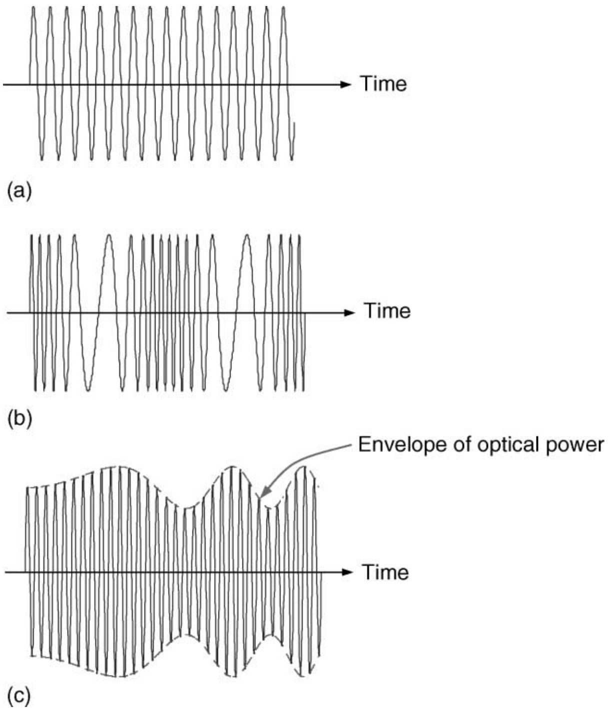

Figure 1 Types of modulation of optical signals. (a) unmodulated optical field; (b) frequency-modulated field; (c) intensity or envelope modulation of the optical power.

modulation and demodulation process, and only the optical power variation is important. This is a key advantage, and as a result intensity modulation systems are prevalent for optical communications. Here, we discuss how varying the injected current in a semiconductor laser or light emitting diode, or by operating the source with constant output and introducing an external modulator, can realize modulation of the optical power of the source. The process by which the optical power variations can be detected at the receiver with semiconductor photodiodes are also discussed, including detection noise characteristics and means of enhancing the responsivity by introducing internal gain mechanisms or wavelength selectivity. The speed capabilities of optical modulation and demodulation in semiconductor lasers and photodetectors are discussed, based on the rate equations for carrier and photon dynamics.

The Modulation of Semiconductor Lasers

Townes, Basov and Prokhorov, who shared the Nobel Prize in 1964, developed the principle of lasers. Semiconductor lasers emit light through a stimulated emission mechanism. The required optical gain for the lasing action is obtained by injecting a forward bias current into a heavily doped p-n junction, thereby injecting carriers into the active junction layer and creating a population inversion. Here we limit our discussion to the optical modulation characteristics of the laser in response to changes in the injection current.

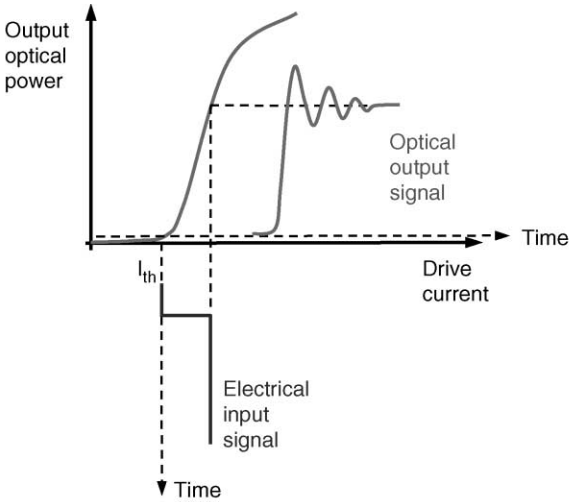

Figure 2 Light-current properties and the dynamic modulation characteristics of semiconductor lasers.

Figure 2 shows the typical light-current characteristics of semiconductor lasers. For drive currents above the threshold value I_{th}, the emitted optical power is predominantly by stimulated emission and it increases linearly with the drive current up to moderate currents. The differential responsivity of the laser, defined by the ratio of optical power increase to electric current increase above threshold, is typically around 0.5 mW/mA. At high current levels, the light output deviates from a proportional dependence on the current because of saturation effects and the differential responsivity decreases.

The dynamic response of the semiconductor laser to modulation of the injection drive current is governed by the rate equations for the carrier density and the photon density in the active junction region. Figure 3 shows the active layer under consideration (*see* **Lasers:** Edge Emitters). The electron concentration in the active layer is denoted N and the photon concentration in the active layer is denoted S. The dynamic behavior of the electron concentration and photon density is governed by the following rate equations:

$$\begin{aligned}\frac{\partial N(x)}{\partial t} =& \frac{I(x)}{q v_a} + D_n \frac{\partial^2 N(x)}{\partial x^2} \\ & - \frac{N(x)}{\tau_{ns}} - G(x)\cdot S(x) \end{aligned} \qquad [1]$$

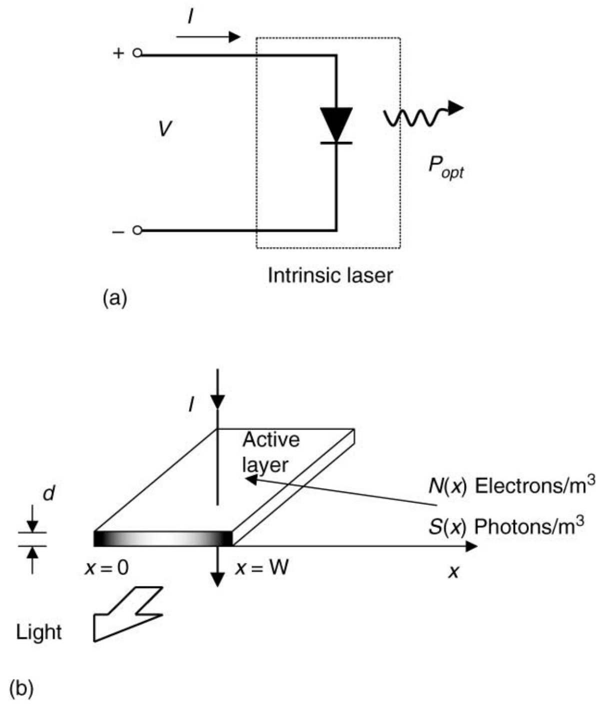

Figure 3 Semiconductor laser. (a) intrinsic device; (b) active region with injected current *I*, electron concentration *N*, and photon concentration *S*.

$$\frac{\partial S(x)}{\partial t} = \Gamma \cdot \beta \frac{N(x)}{\tau_{ns}} + \Gamma \cdot G(x) \cdot S(x) - \frac{S(x)}{\tau_p} \qquad [2]$$

where I is the injected current, q is the electronic charge, ν_a is the volume of the active layer, D_n is the diffusion constant, τ_{ns} is the carrier lifetime, Γ is the confinement factor (defined as the fraction of the optical power in the active region), β is the fraction of the spontaneous emission power coupled into the lasing mode, τ_p is the cavity or photon lifetime, and

$$G(x) = \gamma[N(x) - N_t] \qquad [3]$$

where G is the net rate of stimulated emission, and N_t is the electron concentration at transparency.

Equations [1] and [2] are statements of the fundamental continuity equations for electrons and photons in the active layer. Hence, eqn [1] expresses the fact that the rate of increase of electron concentration in the active layer is given by the rate of increase of electron concentration due to injected current, the rate of increase of electron concentration due to diffusion current, the rate of decrease of electron concentration due to spontaneous recombination, and the rate of decrease in electron concentration due to stimulated photon emission. Equation [2] expresses the fact that the rate of increase of photon density in the active layer is given by the rate of increase in photon density due to spontaneous emission coupled into the lasing mode, the rate of increase of photon density due to stimulated emission coupled into the active layer, and the rate of decrease of photon density due to cavity loss mechanisms.

The actual electron and photon distribution varies across the active layer. However, for simplicity, it is convenient to use average values of N and S, and the simplified rate equations in terms of position independent variables are given by:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{I_i}{q\nu_a} - \frac{N}{\tau_{ns}} - G(1 - \varepsilon S)S \qquad [4]$$

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \Gamma \cdot \beta \frac{N}{\tau_{ns}} + \Gamma \cdot G(1 - \varepsilon S)S - \frac{S}{\tau_p} \qquad [5]$$

where the scaling factor $(1 - \varepsilon S)$ accounts for gain saturation arising from several phenomena such as spatial hole burning, spectral hole burning, carrier heating and two-photon absorption, and ε is a nonlinear gain parameter.

The modulation response of the laser is obtained by solving the rate eqns [4] and [5] for a time dependent input current of the form

$$I(t) = I_b + I_m x(t) \qquad [6]$$

where I_b is the steady-state bias current, I_m is the modulation current amplitude and $x(t)$ represents the normalized shape of the modulation waveform.

The small-signal modulation bandwidth of the laser is obtained by considering the response to small amplitude sinusoidal current modulation at frequency ω, so that $x(t) = \sin(\omega t)$, around a steady-state bias current above lasing threshold. Under the small signal condition, the rate equations can be linearized and combined to result in a second order differential equation. The solution for the photon output exhibits a damped resonant type response. Hence the modulation response displays some peaking at the relaxation oscillation resonance frequency ω_r, and is flat for frequencies much lower than ω_r. The resonance frequency $\omega_r \propto \sqrt{S_o}$, where S_o is the photon density at the laser bias level, so the bandwidth can be increased by biasing the laser further above its threshold current. Semiconductor laser bandwidth to 30 GHz has been observed; however, it is often limited to around 10 GHz due to electrical parasitics in the current drive circuit and interface.

The large-signal modulation response of the laser is obtained by solving the differential eqns [4] and [5]. The response of the carrier-photon coupled system to a step change of drive current has the potential for damped oscillation. This is caused by a resonant type of behavior in which, for example, an increase in photon density causes a depletion in population inversion which leads to a decrease in photon density enabling the carrier density to recover, thus producing a time lag between cause and effect resulting in damped oscillations before settling to the steady state value. These dynamic characteristics are illustrated in Figure 2. For a step modulation of the laser drive current, the optical output response exhibits a delay time, a risetime, an overshoot, and damped oscillations. The risetime of the laser output is typically in the range of 100–300 ps. The turn-on time of the laser and the overshoot are improved by ensuring that the minimum drive current does not fall below threshold. The damped oscillations in the optical output occur at the relaxation oscillation frequency of the laser.

These laser pulse waveform imperfections cause deteriorations in the eye diagram of the laser transmitter that is modulated by very high speed digital bit streams (*see* **Optical Communication Systems:** Lightwave Transmitters). Figure 4 shows the eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream of length $2^{15}-1$. The output waveform imperfections cause a partial closure of the eye, which ultimately limits the modulation speed that can be obtained by directly modulating the laser drive current. Although the semiconductor laser optical output does not exactly replicate the applied electrical drive waveform, the deviations are small enough so that these lasers can be used for data transmission up to around 10 Gbit/s.

Another inherent consequence of direct modulation of semiconductor lasers is the production of laser frequency chirp. This arises from the fact that modulation induced changes in the carrier density in the active region of the laser cavity cause a change in the refractive index of the active layer, which results in a chirping in the output wavelength during modulation. Hence amplitude modulation of semiconductor lasers is fundamentally accompanied by phase modulation from the basic physical mechanism of carrier induced changes in the refractive index. The resulting transient changes in the output wavelength of the laser causes a chirping phenomenon as it is being modulated. The order of chirp magnitude is about 0.1–1 GHz/mA of drive current change. During digital pulse modulation, the laser wavelength shifts to the blue side near the leading edge and towards the red side near the trailing edge of the pulse. Chirp results in a broadening of the laser linewidth. This wavelength change effect in combination with the dispersion characteristics of optical fiber (*see* **Fiber and Guided Wave Optics:** Dispersion; Light Propagation) causes an important limitation to the transmission capacity of fiber optic communication systems. The chirp in semiconductor lasers can be reduced, but not entirely eliminated, by using modulation-doped strained multiquantum-well lasers, which reduce the linewidth enhancement factor.

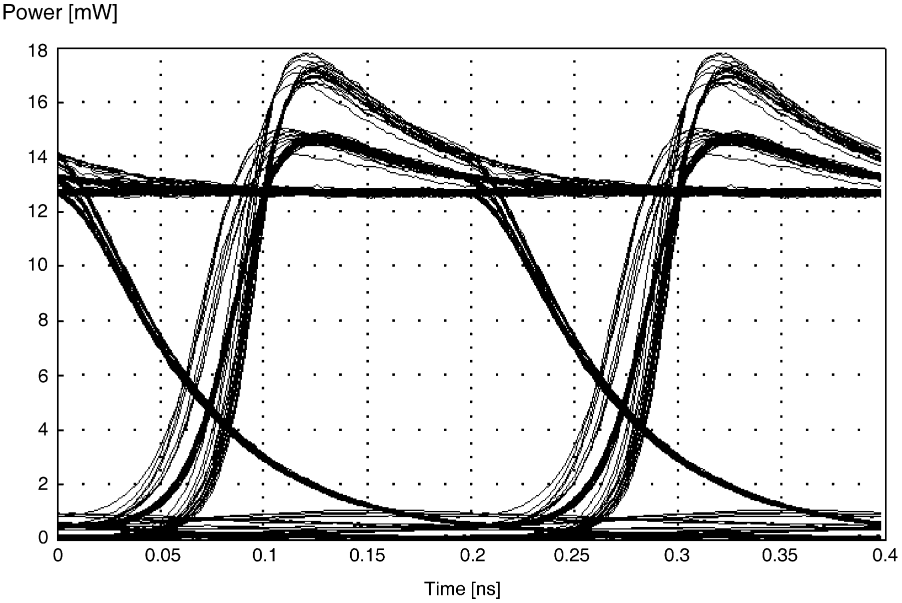

Figure 4 Eye diagram for the optical output of a semiconductor laser directly modulated by a 5 Gbit/s pseudo-random bit stream.

The Modulation of Light Emitting Diodes

The light emitting diode (LED) generates light through spontaneous emission. A semiconductor p-n junction is pumped with a forward current, so that electrons are provided in the upper energy state that can recombine radiatively to produce incoherent light by a spontaneous emission mechanism. The rate of photon generation is directly proportional to the carrier injection rate, and hence the modulating drive current.

Figure 5 shows the typical light-current properties and the modulation characteristics of a LED. The responsivity of the LED, defined as the ratio of the emitted optical power to the injected current, is typically in the range of 10 to 50 μW/mA. The optical output power generated is linearly related to the injection (drive) current up to moderate currents, and is given by

$$P = \eta \cdot h\nu \cdot \frac{I}{q} \qquad [7]$$

where η is the emission quantum efficiency (defined as the ratio of the radiative recombination rate to the total recombination rate), h is Planck's constant, ν is the optical frequency, I is the injected current, and q is the electronic charge. For larger currents, the light output deviates from a proportional dependence on the current because of saturation effects and the responsivity decreases.

The modulation response of the LED is governed by the rate equation for the carrier density N

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{I}{qv_a} - \frac{N}{\tau_c} \qquad [8]$$

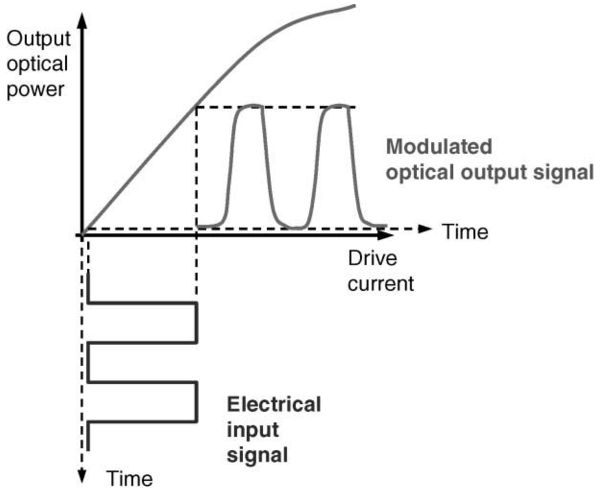

Figure 5 Light-current properties and the modulation characteristics of light emitting diodes.

where I is the injected current, q is the electronic charge, v_a is the volume of the active region, and τ_c is the carrier lifetime. Equation [8] expresses the fact that the rate of increase of carrier concentration is given by the rate of increase of carriers through current injection and the rate of carrier loss through radiative and nonradiative recombination. The small signal response arising from this first order differential equation, exhibits a simple single pole response and the resulting modulation bandwidth of the LED is given by

$$B = 1/(2\pi\tau_c) \qquad [9]$$

The pulse modulation response of the LED is shown in Figure 5. The output modulated optical waveform is not square but exhibits risetimes and falltimes determined by the carrier lifetime τ_c. This sets a limit to the maximum modulation rate that the LED can achieve. The bandwidth can be increased by decreasing the radiative lifetime, which can be controlled to a limited extent by the choice of doping level. Typical lifetimes for InGaAsP LEDs are in the range of 1–5 ns, which corresponds to attainable modulation data rates of several hundred Mbit/s.

The External Modulation of Lasers

Modulation of lasers realized by changing the injection current, as described above, is classed as direct modulation. This is the simplest method of modulation; however, it is limited by several physical phenomena. These include bandwidth limitations due to carrier dynamics in the laser active region, waveform limitations due to relaxation oscillations and overshoot effects, and chirping which causes transmission limitations in optical fibers arising from dispersion effects.

A superior modulation technique, classed as external modulation, operates the laser with continuous wave (CW) output and impresses the modulation by means of a device external to the laser. Although more complex, this approach has many modulation performance advantages. Because the laser is operated CW, this technique removes chirping effects and also enables higher power lasers to be used. Moreover, it is capable of much higher modulation bandwidths, larger on–off extinction ratios, and superior modulation spectral purity.

Electro-optic Modulation of Light

The linear electro-optic effect in materials was first described by Friedrich Pockels in 1893. The Pockels effect is the change in refractive index of the material resulting from the application of an electric field, and is given by

$$n(E) = n - \frac{1}{2} r n^3 E \qquad [10]$$

where n is the refractive index, E is the applied electric field, and r is the electro-optic coefficient. Typical values of r lie in the range 10^{-12} to 10^{-10} m/V, and the most common electro-optic materials used include $LiNbO_3$, GaAs, InP, InGaAsP, and synthetic polymers such as polymethylmethacrylate (*see* **Modulators:** Electro-optics).

If light traverses an optical waveguide of length L, to which an electric field E is applied, it undergoes a phase shift given by

$$\Delta\phi = -\frac{\pi r n^3 E L}{\lambda} \qquad [11]$$

The electric field can be produced by applying a voltage V across the waveguide of width d, hence $E = V/d$. It is therefore possible to modulate the phase of the light by varying the voltage V that is applied to the material through which the light passes. The voltage required to produce a phase shift of π is given by

$$V_\pi = \frac{\lambda d}{n^3 r L} \qquad [12]$$

The parameter V_π, known as the half-wave voltage, is an important characteristic that expresses the applied voltage required to give a phase shift of π.

Intensity modulation of light can be obtained by placing the phase modulator in one arm of an optical interferometer. The structure, known as a Mach-Zehnder modulator, is shown in Figure 6a. The optical output is the phasor sum of two lightwaves that travel through the two arms, resulting in constructive or destructive interference depending on the optical phase difference. The electric field is applied to the optical waveguide in a direction perpendicular to the direction of light propagation using the metal electrodes, see Figure 6a. Hence the refractive index of the waveguide and thus the phase of the light can be modulated by varying the voltage V. When no voltage is applied, the optical fields in the two arms of the interferometer experience identical phase shifts, and hence interfere constructively at the output. When a voltage is applied, the waveguide index changes, and the light in that arm experiences a phase shift with respect to the light in the other arm. This destroys the constructive nature of the interference at the output, and reduces the transmitted light emitted at the output. The extreme case is when the phase difference between the two arms equals π, which causes complete destructive interference resulting in no light transmission.

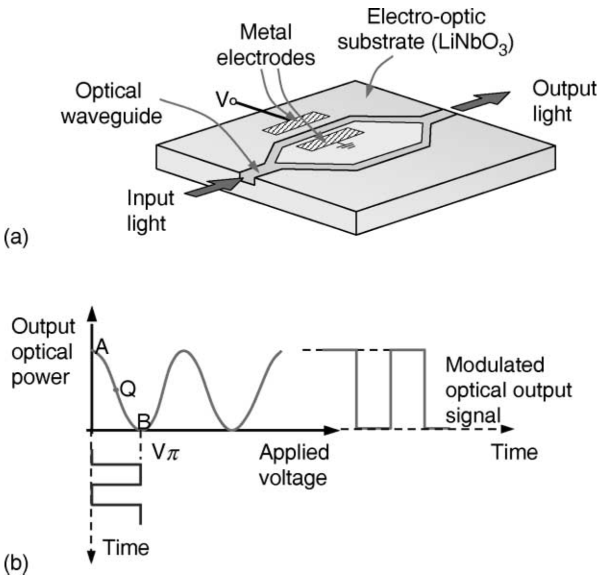

Figure 6 Mach-Zehnder electro-optic intensity modulator. (a) Structure; (b) transfer function and modulation characteristics.

The modulation transfer function is given by

$$P_{\text{out}} = \frac{1}{2} P_{\text{in}}[1 + \cos(\Delta\phi)] \qquad [13]$$

where $\Delta\phi$ is the phase difference encountered by light as it travels through the two arms. In terms of the applied voltage, the optical power transmitted through the modulator is given by

$$P_{\text{out}} = \frac{1}{2} P_{\text{in}}\left[1 + \cos\left(\pi \frac{V}{V_\pi}\right)\right] \qquad [14]$$

This transfer characteristic is shown in Figure 6b. The modulator may be operated as an analog optical modulator by operating in the nearly linear region around point Q. Alternately, it may be operated as a digital modulator by operating between points A and B, thus switching the light on and off as V is switched between 0 and V_π.

Improved efficiency of modulation, and hence a reduction in the value of V_π (typically a few volts) can be obtained by placing a second electrode on the other arm of the interferometer, and using a push-pull configuration to drive the interferometer. Hence the phase modulators in both arms are driven with modulating voltages of opposite polarities so the phase is advanced in one arm and is retarded in the other arm. This produces a net phase difference, which is double that produced in either arm.

The Pockels effect is an extremely fast physical mechanism, hence in principle external modulation is capable of achieving extremely high speeds. In practice, the modulation speed of an electro-optic modulator is limited by the capacitive effects at the electrodes and by the transit time of the light through the material. If the electric field varies significantly within the time it takes for light to transit through the waveguide, then the travelling optical wave will be subjected to different electric fields as it traverses the waveguide, and this results in a transit time limitation to the modulation bandwidth. A technique to avoid this limitation is to apply the voltage at one end of the waveguide and to design the electrodes to form a transmission line. By this means it is possible to make the velocity of the electrical wave match that of the optical wave to essentially eliminate the transit time effects. These external modulators have demonstrated speeds well in excess of 100 GHz.

Electro-absorption Modulation of Light

Electro-absorption external modulators are based on the principle of semiconductor bandgap decrease that occurs upon application of an electric field (the Franz-Keldysh and Stark effect). With no applied voltage, the semiconductor waveguide is transparent to light having a photon energy smaller than the bandgap, and such wavelengths are transmitted with very little attenuation. When a voltage is applied, the bandgap of the semiconductor decreases and the material absorbs the light once the photon energy exceeds the bandgap energy, and this produces optical attenuation. Hence by applying a modulation voltage to the semiconductor waveguide, the absorption property of the waveguide and hence the output light intensity can be modulated.

Improved modulation efficiency, and hence a reduction in the voltage required to decrease the bandgap energy, is obtained by using multiquantum-well semiconductor structures, which exhibit enhanced electro-absorption effects arising from the quantum confined Stark effect. The conventional technique of applying a transverse field across a multiquantum-well is to apply a reverse voltage across it. Modulator drive voltages of a few volts can be obtained. The on/off extinction ratio of modulation depends on the value of the field-dependent absorption coefficient at the operating voltage, and typical on/off extinction ratios are 10–15 dB.

The modulation bandwidth is determined by intrinsic factors related to carrier dynamics, such as exciton formation and dissociation times and the time taken to remove electrons and holes from the wells by thermionic emission and tunelling. It is also determined by extrinsic factors related to the capacitance of the junction diode. These modulators can have bandwidths of around 50 GHz. They also enable low chirp modulation to be obtained. The optical power handling capability of electro-absorption modulators is lower than for electro-optic modulators. However, a principal advantage of the electro-absorption modulator is that the semiconductor waveguide can be fabricated onto one substrate with the semiconductor laser, enabling the realization of monolithic integration on the same chip.

The Demodulation of Optical Signals

The process of demodulation converts the modulated intensity of an optical wave into an electrical signal to recover the information content. Semiconductor photodetectors are based on the principle of absorption of photons to generate mobile electron-hole pairs, and the transport of these mobile carriers via an applied electric field to create a flow of current.

Demodulation using PIN Photodiodes

The PIN photodiode comprises a P-doped/intrinsic/N-doped semiconductor structure. An applied reverse voltage bias creates a depletion region with a high electric field in the intrinsic layer. When the intrinsic layer is illuminated with light, photons are absorbed and electron-hole pairs are created, which are swept by the built-in electric field. The resulting flow of current is proportional to the incident optical power.

The quantum efficiency of the photodetector is defined as the ratio of electron generation rate to the photon incidence rate, and is given by

$$\eta = \frac{h\nu}{q} \cdot \frac{I_p}{P} \qquad [15]$$

where I_p is the photocurrent, P is the incident optical power, ν is the optical frequency, h is Planck's constant and q is the electron charge. If the width of the absorbing intrinsic region is large enough to capture all the light, quantum efficiency values approaching 100% can be achieved.

The current-voltage characteristics of the photodiode under optical illumination are shown in Figure 7a. The dark current I_d, which is the current generated in the photodetector in the absence of light originates from thermally generated electron-hole pairs, and is typically extremely small (<5 nA). Figure 7b shows the transfer characteristic of the

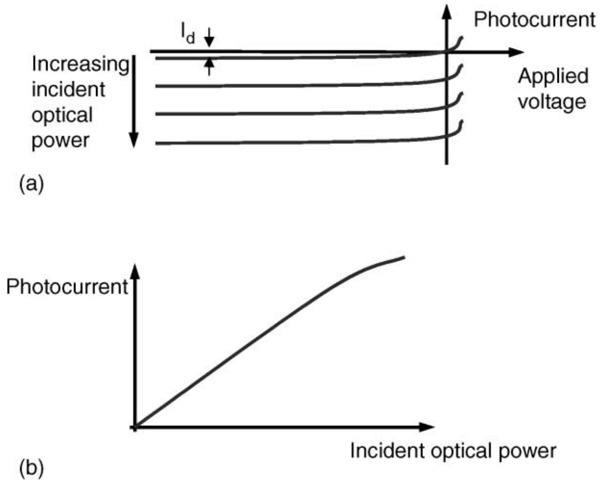

Figure 7 Semiconductor photodiode. (a) Current-light characteristics under optical illumination; (b) transfer function and demodulation photocurrent versus incident optical power.

photodetector. The current is linearly proportional to the optical power up to moderate power levels, when saturation mechanisms occur. The responsivity of the photodiode $\Re$, defined as the ratio of the generated electric current to the incident optical power, is given by

$$\Re = \frac{\eta q}{h\nu} \qquad [16]$$

The responsivity is an important parameter, and provides a measure of the photodetection efficiency of the device for demodulating optical signals. Typical responsivity values for PIN photodiodes are around 1 A/W.

The bandwidth B of demodulation of the photodetector is determined by the speed with which it responds to variations in the incident optical power. This is limited by the transit time τ_{tr} and the RC time constant τ_{RC} and is given by

$$B = \frac{2\pi}{\tau_{tr} + \tau_{\mathrm{RC}}} \qquad [17]$$

The transit time limitation arises because of the time it takes for the generated carriers to traverse the absorption region and be collected. To maximize the bandwidth, the transit time needs to be reduced, which can be achieved by decreasing the width of the absorption region. However, this requires a trade-off with the responsivity, which drops because a smaller fraction of the light can be absorbed under these conditions. The capacitance C and the resistance R of the photodetector and its circuitry form the RC time constant. To maximize the bandwidth, the photodiode capacitance needs to be reduced by decreasing the area of the photodiode. The demodulation bandwidth of PIN photodiodes can exceed 100 GHz.

Photodetection shot noise is a fundamental noise mechanism that introduces noise in the demodulation process of optical signals. This leads to fluctuations in the photocurrent, even for a constant incident optical power, because the photocurrent actually consists of a stream of electrons that are generated at random times. The photodiode current generated by a constant incident optical power is given by

$$I(t) = I_{\mathrm{avg}} + i_{\mathrm{shot}}(t) \qquad [18]$$

where I_{avg} is the average photocurrent generated, and $i_{\mathrm{shot}}(t)$ is the current fluctuation due to shot noise and which is described by a stationary random process with Poisson statistics. The mean-square noise current is given by

$$< i_{\mathrm{shot}}^2(t) >= 2qI_{\mathrm{avg}}\Delta f \qquad [19]$$

where Δf is the effective noise bandwidth of the receiver. The shot noise degrades the signal to noise ratio of the demodulated signal, and causes a fundamental limit to the sensitivity of the PIN photodetection.

Demodulation using Avalanche Photodiodes

Demodulation using avalanche photodiodes detectors (APD) has the advantage of significantly increased responsivity. This is achieved by providing an internal current gain mechanism within the device. The APD operates by generating more than one electron-hole pair per absorbed photon. This is achieved by strongly reverse biasing the diode to create a large junction electric field, so photoelectrons accelerate and acquire enough energy to produce new carriers by the process of impact ionization. These devices incorporate a modification to the PIN structure, by including an additional doped layer known as the multiplication layer, which under reverse bias provides a region of very high electric field that can produce avalanche breakdown. Hence a single primary electron generated through absorption of a photon, generates many secondary electrons and holes in the avalanche multiplication layer, and all these secondary carriers contribute to the photocurrent. As a result of this current gain mechanism, the responsivity of the APD is increased by a multiplication factor M, and is given by

$$\Re_{ADP} = \frac{M\eta q}{h\nu} \qquad [20]$$

where M can range between 5–50 depending on the bias voltage.

The bandwidth of the APD depends on the multiplication factor M. This is because the avalanche process takes additional time. The frequency response can be written

$$M(\omega) = \frac{M_0}{[1 + (\omega\tau_e M_0)^2]^{\frac{1}{2}}} \qquad [21]$$

where M_0 is the low frequency gain and τ_e is the effective transit time that depends on the ionization coefficient. The intrinsic APD bandwidth (assuming no RC limitations) is approximately given by

$$B_{\mathrm{APD}} = \frac{1}{2\pi\tau_e M_0} \qquad [22]$$

Hence the bandwidth decreases as the multiplication factor increases, and there exists a trade-off between the bandwidth and the responsivity. The demodulation characteristics of the avalanche photodiode can be expressed by means of a gain-bandwidth product $B_{\mathrm{APD}} \cdot M_0$, and this can exceed 150 GHz.

Photodetection noise in APDs has an additional contribution because the avalanche gain itself is a random variable. Hence the secondary electron-hole pairs are generated at random times and this adds a noise component to the basic shot noise associated with the generation of primary electro-hole pairs. The total shot noise is given by

$$< i^2_{\mathrm{shot,APD}}(t) >= 2qM^2F(M)I_{\mathrm{avg}}\Delta f \qquad [23]$$

where $I_{\mathrm{avg}} = \Re P$, and F is the excess noise factor expressed by

$$F = \frac{< M^2 >}{< M >^2} \qquad [24]$$

and is approximately given by $F \approx M^x$, where $0.2 < x < 1$. Hence the excess noise increases with the multiplication ratio, so there exists an optimum avalanche photodiode gain that maximizes the signal to noise ratio of the demodulated signal. Despite the fact that additional noise is introduced, because of the higher responsivity, APDs can provide higher sensitivity in demodulating optical signals than PIN photodiodes.

Demodulation using Metal-Semiconductor-Metal (MSM) Photodiodes

Demodulation using MSM photodiodes enables extremely high bandwidths to be achieved. The MSM photodiode comprises two reverse-biased metal-to-semiconductor Schottky contacts formed on a semiconductor layer that prevents flow of electrons from metal to semiconductor. Photons incident between the contacts create electron-hole pairs that are collected by the electric field to generate a photocurrent proportional to the incident optical power. The metal contacts are usually fabricated on the same side of the semiconductor to create a planar structure, which is suitable for monolithic integration. However, a drawback of the MSM device is that the electrodes block some of the light from reaching the absorption region, which results in a reduced responsivity.

The bandwidth of demodulation of the MSM photodiode is determined by the transit time of the photogenerated carriers and the RC time constant. Since the electrode spacing can be around 1 μm, the transit time is extremely small. Moreover, since in the planar structure the parasitic capacitance is inherently low, the RC time constant is also extremely small. This results in demodulation bandwidths of up to 300 GHz.

Wavelength Selective Demodulation

A technique to obtain both wavelength selective demodulation and high-speed, high-responsivity demodulation involves the use of microcavity resonant photodetectors. This is useful for very high-speed detection in wavelength multiplexed systems.

The microcavity photodiode comprises a Fabry-Perot cavity placed around the photodetector absorption region, to enhance the responsivity. The mirrors are typically formed using Bragg reflectors or dielectric mirrors implemented in the semiconductor. A resonance is built up in the cavity at the wavelengths of incoming light for which the round-trip phase shift is 2π, which corresponds to constructive interference. Hence at the resonant wavelength, the incoming light is reflected at the two mirrors and permits the absorbing layer to be reused many times by effectively recirculating the optical signal, thus enhancing the absorption and the responsivity. However, the transit length can remain extremely small, because the photogenerated carriers only have to traverse the thickness of the absorption region to be collected. This solves the problem of the fundamental responsivity-bandwidth trade-off in normal photodetectors, in which increasing the absorption layer width to increase responsivity increases the transit time thus reducing the bandwidth. Microcavity resonant photodiodes enable the simultaneous achievement of large bandwidth and high responsivity possible. Also, the resonant behavior leads to wavelength selectivity, which can be used to advantage in some wavelength division multiplexed WDM applications (*see* **Optical Communication Systems:** Wavelength Division Multiplexing). This approach can provide demodulation of optical signals with quantum efficiencies close to 100% and with bandwidths in excess of 100 GHz.

Optical Receivers – Preamplification and Clock Recovery

The front end of an optical receiver comprises the photodiode and a preamplifier to amplify the electrical signal for further processing. In order to obtain a high demodulation bandwidth, it is important to minimize the photodetector capacitance and to minimize the input resistance of the preamplifier that acts as the load resistance. However, because the thermal noise current is inversely proportional to the load resistance, minimizing the load resistance causes excessive noise and degraded demodulation sensitivity. An alternative approach is to use a high-impedance preamplifier. This reduces the noise at the price of a lower bandwidth. Thus there is a trade-off between the bandwidth and sensitivity. The trans-impedance front-end, in which the load resistor is connected as a feedback resistor around an inverting amplifier, is often used as an optimum compromise that combines high sensitivity together with large bandwidth. This enables the use of a large resistance, which minimizes noise, but at the same time the effect of negative feedback reduces the input impedance by a factor equal to the amplifier gain, thus increasing the bandwidth.

In digital optical transmission systems the data recovery section of the optical receiver comprises a clock recovery circuit and a decision circuit. Accurate timing (clock) information about the bit slot, in order to synchronize the decision process, is obtained from the amplified front-end waveform using a timing extraction circuit (*see* **All-Optical Signal Regeneration**). Typical clock recovery circuits comprise a high-Q resonant circuit which rings at the bit rate frequency or a phase locked loop. This enables precise data time sampling in the following decision circuit within the bit intervals of the original pulse train.

See also

All-Optical Signal Regeneration. **Fiber and Guided Wave Optics:** Dispersion; Light Propagation. **Lasers:** Edge Emitters. **Modulators:** Electro-Optics. **Optical Communication Systems:** Basic Concepts; Lightwave Transmitters; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fiber-Optic Communication Systems*. New York: Wiley.

Coldren L and Corzine S (1995) *Diode Lasers and Photonic Integrated Circuits*. New York: Wiley.

Gowar J (1993) *Optical Communication Systems*. Englewood Cliffs: Prentice Hall.

Kazovsky L, Bendetto S and Willner AE (1996) *Optical Fiber Communication Systems*. Boston: Artech House.

Keiser GE (1991) *Optical Fiber Communications*. New York: McGraw-Hill.

Miller SE and Kaminow I (eds) (1988) *Optical Fiber Telecommunications II*. New York: Academic Press.

Personick SD (1985) *Fiber Optics: Technology and Applications*. New York: Plenum Press.

Petermann K (1991) *Laser Diode Modulation and Noise*. Boston: Kluwer Academic.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Senior J (1992) *Optical Fiber Communications*. NJ: Prentice Hall.

Siegman AE (1986) *Lasers*. Sausalito, CA: University Science Books.

NONCLASSICAL LIGHT

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics,
Garching, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The pioneering experiments by Hanbury-Brown and Twiss, and the subsequent development of the quantum theory of electromagnetic radiation by Glauber and Sudarshan, raised the interest in the question of photon statistics. It followed from their theories that states of the electromagnetic field which can only exist in a quantum treatment. These nonclassical states show up in photon statistics through the phenomenon of antibunching. Later, further nonclassical states of the radiation field were found in the so-called squeezed fields, which are fields with phase-sensitive quantum fluctuations being at certain phase angles smaller than those of a perfect classical coherent field. In order to achieve squeezing of laser radiation, various forms of nonlinear optical interactions can be used.

Classical Versus Quantum Waves

In classical electrodynamics, a monochromatic wave (e.g., an extended light beam) is often described using the complex representation:

$$\begin{aligned} E(r,t) &= E^{-}(r,t) + E^{+}(r,t) \\ &= \frac{1}{2}[\varepsilon(r,t)\mathrm{e}^{-\mathrm{i}\omega t} + \varepsilon^{*}(r,t)\mathrm{e}^{\mathrm{i}\omega t}] \end{aligned} \quad [1]$$

In classical theory, this complex description is merely a tool for the simplification of some calculations. However, there is some physical significance when using the quantum description, where ε and ε^* lead to the lowering and raising operators of the field. The above field can also be represented in the following way:

$$E(r,t) = \frac{\varepsilon + \varepsilon^*}{2} \cos\omega t + \frac{\varepsilon - \varepsilon^*}{2} \sin\omega t \quad [2]$$

with real coefficients $\varepsilon_1 = (\varepsilon + \varepsilon^*)/2$ and $\varepsilon_2 = (\varepsilon - \varepsilon^*)/2\mathrm{i}$. This description gives the state of the field in a two-dimensional phase diagram where the sin-component is plotted versus the cos-component. In this plot, the wave above is represented by one point giving phase and amplitude of the field. These phase diagrams will play an important role in characterizing the nonclassical fields in a quantum phase diagram, as discussed below.

The properties of a field are described by the first- and second-order correlation functions. The first-order correlation function is given by

$$G^{(1)}(r_1, t_1, r_2, t_2) = \langle \varepsilon^*(r_1, t_1)\varepsilon(r_2, t_2)\rangle \quad [3]$$

It is obtained by multiplying the electric field amplitudes at two spacetime points and taking the ensemble average. For $r_1 = r_2$ and $t_1 = t_2$, $G^{(1)}$ is proportional to the intensity at this point and is, therefore, called the first-order in intensity. The first-order correlation function is connected to the spectrum $I(\omega)$ of the field through its Fourier transform.

The second-order correlation function is necessary to characterize the radiation field completely. It describes the intensity fluctuations, being the classical equivalent to photon statistics, and is given in the general form:

$$\begin{aligned} &G^{(2)}(r_1, t_1, r_2, t_2, r_3, t_3, r_4, t_4) \\ &\quad = \langle \varepsilon^*(r_1, t_1)\varepsilon^*(r_2, t_2)\varepsilon(r_3, t_3)\varepsilon(r_4, t_4)\rangle \end{aligned} \quad [4]$$

If $G^{(2)}$ is considered for two spacetime points $r_1 = r_2$, the connection to intensity fluctuations becomes obvious:

$$G_{\mathrm{I}}^{(2)}(\tau) = 4\langle I(t_1)I(t_2)\rangle = \langle \varepsilon^*(t_1)\varepsilon(t_1)\varepsilon^*(t_2)\varepsilon(t_2)\rangle \quad [5]$$

The subscript I indicates that we have to deal with intensity correlations; τ stands for $t_2 - t_1$.

In connection with the Hanbury-Brown, Twiss experiment, in general, the normalized intensity correlation is given by

$$g_{\mathrm{I}}^{(2)}(\tau) = \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 \qquad [6]$$

$I(t)$ stands for the intensity which is given by the mean value $\langle I(t)\rangle$ plus fluctuations $\Delta I(t)$ for which the mean value is zero. The mean value of the squared intensity is then $\langle I^2\rangle = \langle I\rangle^2 + \langle \Delta I^2\rangle$. It follows for the normalized intensity correlation function:

$$\begin{aligned} g_{\mathrm{I}}^{(2)}(\tau) &= \langle I(t)I(t+\tau)\rangle / \langle I(t)\rangle^2 \\ &= \left\{ \begin{array}{ll} 1 + \langle \Delta I^2\rangle / \langle I\rangle^2 & \text{for } \tau = 0 \\ 1 & \text{for } \tau \to \infty \end{array} \right\} \end{aligned} \qquad [7]$$

For $\tau \to \infty$, the intensities are not correlated and $g_{\mathrm{I}}^{(2)}(\tau)$ becomes 1. Otherwise $\langle \Delta I^2\rangle / \langle I\rangle^2 \geq 0$; this means that $g_{\mathrm{I}}^{(2)}(\tau)$ is always equal to or larger than 1, and $g_{\mathrm{I}}^{(2)}(\tau)$ smaller than 1 is not possible classically. The photon correlation is usually measured by the classical Hanbury-Brown, Twiss setup shown in Figure 1.

The maximum of $g_{\mathrm{I}}^{(2)}(\tau)$ at $\tau = 0$ is called intensity bunching and was first observed by Hanbury-Brown and Twiss. The decrease of $g_{\mathrm{I}}^{(2)}(\tau)$ as a function of τ is a measure of the coherence length of the light source.

When the experiment in Figure 1 is considered in the photon picture there is a principal difference compared to the classical intensity. In the classical case, the intensity on the two detectors is the same if the beamsplitter is dividing the light beam in equal portions. Both detectors, therefore, measure the same fluctuations. In the quantum picture, however, a photon can only be detected in one of the two detectors; this leads to the result that values for $g_{\mathrm{I}}^{(2)}(\tau)$ smaller than 1 can be found. The effect of photon bunching was initially thought to be an inherent property of light owing to the boson nature of photons. Therefore, it came as a big surprise when Glauber showed theoretically that for the quantized radiation field $g_{\mathrm{I}}^{(2)}(\tau)$, <1 is possible. This effect, known as photon antibunching, and its experimental observation constitutes an unambiguous proof of the quantum nature of light.

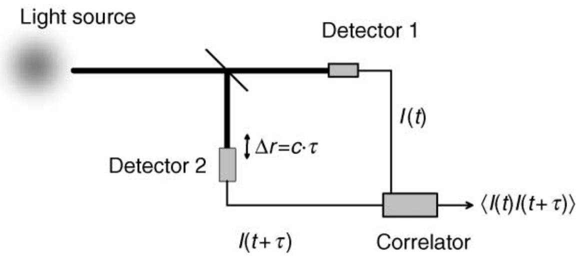

Figure 1 Experimental setup for the Hanbury-Brown, Twiss experiment.

In the quantum description of the electromagnetic field, the electric field amplitudes ε and ε^* are replaced by $\hat{a}$ and $\hat{a}^+$, the photon annihilation and creation operators.

In this case, the electric field operator is written as

$$\hat{E} \propto \frac{\hat{a} + \hat{a}^+}{2} \cos\omega t \frac{\hat{a} - \hat{a}^+}{2\mathrm{i}} \sin\omega t \qquad [8]$$

In analogy, the operators $\hat{a}_1 = (\hat{a} + \hat{a}^+)/2$ and $\hat{a}_2 = (\hat{a} - \hat{a}^+)/2\mathrm{i}$, can be defined describing the amplitude of the cosine and sine parts of the wave. $\hat{a}_1$ and $\hat{a}_2$ do not commute $[\hat{a}_1, \hat{a}_2] = (\mathrm{i}/2)1$, therefore, the expectation values of the two operators $\hat{a}_1$ and $\hat{a}_2$ underlie an uncertainty relation given by

$$\langle \Delta \hat{a}_1^2\rangle = \langle \hat{a}_2^2\rangle - \langle \Delta \hat{a}_i\rangle^2 \geq 1/16 \qquad [9]$$

with $\langle \Delta \hat{a}_i^2\rangle = \langle \hat{a}_i^2\rangle - \langle \Delta \hat{a}_i\rangle^2$. In the corresponding phase diagram, the expectation values for $\hat{a}_1$ and $\hat{a}_2$ are represented by a shaded area defined by the uncertainty (Figure 2). The state of the radiation field, where both phase and amplitude fluctuations are symmetrically minimized, $\langle \Delta \hat{a}_1^2\rangle = \langle \Delta \hat{a}_2^2\rangle = \frac{1}{4}$, best approximates the classical wave with well-defined phase and amplitude. The state with symmetrical minimum uncertainty is the coherent state describing laser radiation. It is an eigenstate of $\hat{a}$ with a photon number fluctuation given by a Poissonian distribution with $\langle \Delta n^2\rangle = \langle n\rangle$.

There are other minimum uncertainty states, however, with unsymmetrical uncertainty regions

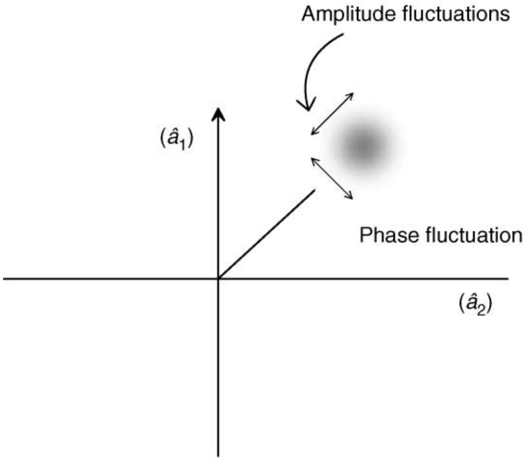

Figure 2 Phase diagram of the quantum mechanical field amplitudes $\langle \hat{a}_1\rangle$ and $\langle \hat{a}_2\rangle$ for coherent radiation.

with the uncertainty squeezed into mainly one component. These states, which are called squeezed states, are defined by $\langle \Delta \hat{a}_1^2 \rangle < 0.25$ or $\langle \Delta \hat{a}_2 \rangle < 0.25$. Squeezed state may also refer to a state which is not a minimum uncertainty state but where, nevertheless, the mean square fluctuation in one of the two components is less than 0.25.

A classical electromagnetic field consists of waves with well-defined amplitude and phase. This is not the case in a quantum treatment: fluctuations are associated with both conjugate variables. The case of a coherent state that most nearly describes a classical electromagnetic field has an equal amount of uncertainty in the two variables (when normalized to the field of a single photon). Equivalently, the field can be described in two conjugate quadrature components and the uncertainties in the two conjugate variables satisfy an uncertainty relation. The coherent state represents a minimum uncertainty state with equal uncertainties in the two quadrature components. This case is usually called the shot-noise limit; it is represented in Figure 2 as a phase diagram.

In a quantum treatment of radiation, it is also possible to generate states that are not present in the classical limit. They can show fluctuations reduced below the classical limit in one of the quadrature components, while the canonically conjugate quadrature component must display enhanced fluctuations in order to fulfill the uncertainty relation. Those states are called 'squeezed states.' An electromagnetic field with fluctuations below the standard quantum limit in one of the quadrature components has, in principle, many attractive applications, for example, in optical communication, in precision and sensitive measurements such as gravitational wave detection, or in noise-free amplification. Therefore, there has been great interest in generating squeezed radiation. Optical measurements have three characteristics that enable them to reach the quantum noise level more readily than in other fields of physics: (i) optical signals are naturally immune to external sources of noise; (ii) thermal noise at room temperature is negligible in the optical domain; and (iii) the outstanding equalities of the optical sources and detectors allow a very-low-level instrumental noise.

Squeezed states with reduced photon number fluctuations (sub-Poissonian photon statistics) correspond to a light beam in which the photons arrive more equally spaced than in a coherent beam for which $g_I^{(2)}(0) = 1$. Therefore, sub-Poissonian number fluctuations correspond to photon antibunching. This shows that photon antibunching and squeezing are related. The state with the largest possible antibunching is the number state, represented by a circle in the phase diagram. It has large fluctuations in both components a_1 and a_2 and, therefore, is not a squeezed state. Thus, it is obvious that antibunching does not imply squeezing. On the other hand, weak squeezing of a coherent state may result in antibunching. If the state is squeezed harder the a_1 component will be better defined. The amplitude fluctuations, however, will start increasing and there will be a transition from antibunching to bunching.

The quantum calculation of the normalized second-order correlation function leads to the following result:

$$g_I^{(2)}(0) = 1 + (\langle \Delta n^2 \rangle - \langle n \rangle / \langle n \rangle^2) \qquad [10]$$

A pure number state of the radiation field with $\langle \Delta n^2 \rangle = 0$, leads to $g_I^{(2)}(0) < 1$, as discussed above. For a coherent state, $\langle \Delta n^2 \rangle = \langle n \rangle$ yielding $g_I^{(2)}(0) = 1$, corresponding to the classical calculations.

It can be concluded that although antibunching and squeezing are related and nonclassical phenomena, one does not imply the other. Thus, the detection schemes necessary to determine the existence of one or the other are quite different. Antibunching or sub-Poissonian statistics can be detected via intensity correlations. Squeezing, on the other hand, is a property of either $\langle a_1 \rangle$ or $\langle a_2 \rangle$, and phase sensitive homo- or heterodyne detection is needed.

Nonclassical states of the radiation field are very sensitive to linear attenuation. Coherent and thermal light fields are the only ones that do not change their photon statistics when attenuated. This is different for nonclassical light. The change depends on the particular case and has to be carefully evaluated.

The first observation of squeezing was achieved in 1985, in an experiment of parametric generation involving four-wave mixing in sodium vapor.

Nonclassical Light from a Single Atom

Resonance fluorescence of atoms is a basic process in radiation–atom interactions, and has therefore always generated considerable interest. The methods of experimental investigation have changed continuously due to the availability of new experimental tools. A considerable step forward occurred when tunable and narrowband dye laser radiation became available. These laser sources are sufficiently intense to easily saturate an atomic transition. In addition, the lasers provide highly monochromatic light with coherence times much longer than typical natural lifetimes of excited atomic states. Excitation spectra with laser light using well-collimated atomic beams lead to a width, being practically the natural width of the resonance transition, therefore it became possible to investigate the frequency spectrum of the fluorescence radiation with high resolution and to also study the photon statistics of the fluorescent light.

Previous experiments to investigate antibunching in resonance fluorescence have been performed by means of laser-excited collimated atomic beams. The initial results obtained in 1977 showed for the second-order correlation function $g^{(2)}(t)$, a positive slope characteristic of photon antibunching, but $g^2(0)$ was larger than $g^2(t)$ for $t \rightarrow \infty$. This was due to number fluctuations in the atomic beam and to the finite interaction time of the atoms. In 1978, the analysis of the experiment was refined and in 1982, another experiment with a longer interaction time was performed. In the latter experiment, the photon correlation was also measured for very low laser intensities. The ideal experiment to study the photon statistics in resonance fluorescence is the investigation of a trapped ion since, in such an experiment, the fluctuations of the atomic number are excluded.

The fluorescence of a single ion should also display the following property. The probability distribution of the photon number recorded in a finite time interval t is narrower than Poissonian, which means, that the variance is smaller than the mean value of the photon number. This is because the single ion can only emit a single photon. Antibunching and sub-Poissonian statistics are often associated. They are, nevertheless, distinct properties and need not necessarily be simultaneously observed, as is the case in the experiment described here. Although there is evidence of antibunching in the atomic-beam experiments, the photon counts are not sub-Poissonian as a result of fluctuations in the number of atoms. In further experiments this effect was excluded by use of a special trigger scheme for the single-atom event. In the setup with a trapped ion these precautions are not necessary since there are no fluctuations in the atomic number. Figure 3 shows the photon correlation of a single trapped ion. The measurements are performed using the $3^2S_{1/2} - 3^2P_{3/2}$ transition of the $^{24}Mg^+$-ion at a wavelength of 280 nm. The laser is tuned slightly below resonance in order to Doppler-cool the secular motion of the ion.

The investigation of the photon correlation employed the ordinary Hanbury-Brown and Twiss setup with two photomultipliers and a beamsplitter (Figure 1). The time delay τ between the photomultiplier signals was converted by a time-to-amplitude converter into a voltage amplitude proportional to the time delay. A delay line of 100 ns in the stop channel allowed for the measurement of $g^{(2)}(\tau)$ for both positive and negative τ in order to check the symmetry of the measured signal. The output of the time-to-amplitude converter was accumulated by a multichannel analyzer in pulse height analyzing mode. Three typical measurements, each at small values of the laser intensity but with different detunings, are shown in Figure 3. The solid curve is a theoretical fit to the measurements. For small time delays (<20 ns), the nonclassical antibunching effect is observed, superimposed with Rabi oscillations being damped out with a time constant corresponding to the lifetime of the excited state.

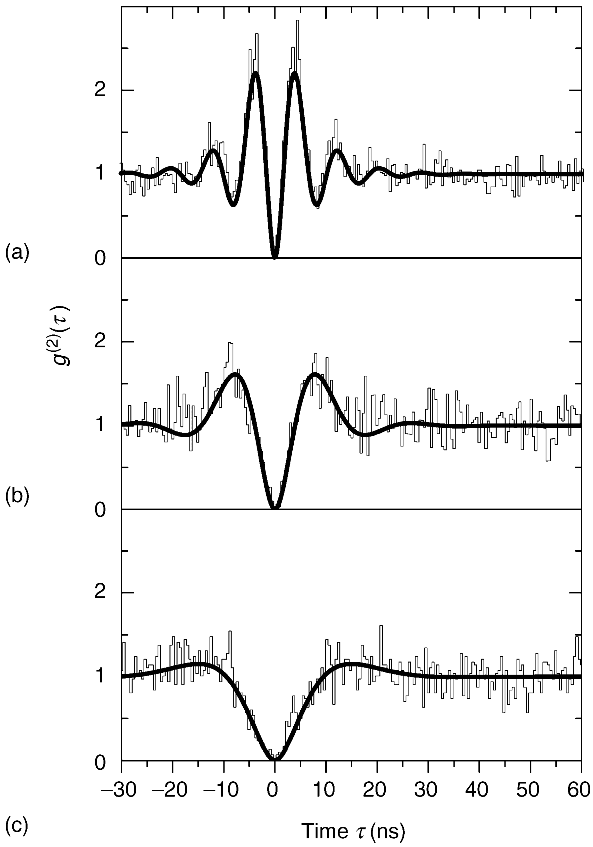

Figure 3 Antibunching signals of a single $^{24}Mg^+$-ion for different laser intensities (decreasing from (a) to (c)). At larger intensities (a) and (b) the Rabi oscillations owing to a strong coherent coupling between ground and excited state are observed.

Recently it also became possible to combine a trapped ion with a cavity. In this way the emission of a photon in the cavity mode can be achieved. This leads to a directed emission of the photon being more efficient than the fluorescence occurring in the entire 4π geometry. In this way, a deterministic single photon source is obtained.

Survey on the Experiments on Squeezing

In order to produce squeezed light from nonlinear optical phenomena, a variety of methods has been used, depending on the type of nonlinearity and the number of interacting modes involved. The choice of optical nonlinearity reduces to either three-wave or four-wave mixing, leading to quadratic or cubic nonlinear response functions, respectively. Each of these methods has advantages and disadvantages. Three-wave mixing requires a pump field at twice the frequency of the field-mode being squeezed. However, it has the advantage as any sidebands induced on the pump, owing to phase noise, have a large frequency offset from the squeezed fields. The other method, the much simpler four-wave mixing, does not require any frequency-doubling step. This was, therefore, the first method used to obtain squeezing with quadrature noise below the vacuum level.

Most of the early experiments on squeezing were performed with one or two cavity modes with the squeezed modes being transferred to propagating external modes via beamsplitters for measurement. This narrowband procedure is followed in more recent experiments commonly used for three-wave (parametric) squeezing. Also squeezing in waveguides was achieved leading to a relatively simple and broadband implementation of squeezing, with, mostly a four-wave mixing with short pulses or solitons being performed.

The main limiting factors that remain include control over the driving laser phase and intensity noise, absorption in the dielectric material and associated devices, and phase fluctuations caused both by low-frequency technical noise in components, and by intrinsic material properties of phonon interactions and refractive index fluctuations. It appears not impossible to reach squeezing levels of 10 dB or more below shot noise, over bandwidths of up to 1 THz. At this level of noise reduction, the main problem is the issue of improving detector technology to allow such low noise levels to be investigated.

Squeezing in Lasers

The quantum noise properties of the laser were studied intensively after its invention in the early 1960s. The conclusion of these studies is that the laser output approaches quite closely the noise properties of a coherent state, provided the laser is operated well above threshold. That is, the spectral variance of the photon number and phase fluctuations at high enough Fourier frequencies will be at the Poissonian or quantum noise limit (also referred to as the shot noise level). On the other hand, the phase of the laser is unconstrained and drifts as a function of time, so-called laser phase diffusion. As a result, the phase fluctuations show excess noise at low frequencies.

In the early 1980s, it was realized that if the pumping mechanism for the laser could be made sub-Poissonian then amplitude squeezing could be obtained at low frequency. A few years later it was shown that sub-Poissonian pumping could be realized in semi-conductor lasers by regularizing the pump current. An experimental demonstration of amplitude squeezing from semi-conductor lasers was soon achieved. Although still requiring rather specialized lasers, many groups now routinely obtain squeezing from semi-conductor lasers.

There are a number of disadvantages to semi-conductor lasers, including their poor beam quality and short coherence length. As a result, interest in other possible squeezing mechanisms continues. It has been found that intrinsic dynamical effects can also lead to amplitude squeezing in the laser output.

Laser squeezing is of a different character to that produced by more ‘traditional squeezers.’ For example, in parametric amplification, the de-amplification of phase fluctuations is accompanied by an equal and opposite amplification of amplitude fluctuations which preserves the uncertainty relationship between them. However, in laser squeezing the reduction in amplitude fluctuations is not linked to any increase in phase fluctuations. Thus, in laser squeezing we take advantage of the fact that the laser output is not in a minimum uncertainty state to suppress amplitude noise. This means that laser squeezing is always amplitude or intensity squeezing.

Application of Nonclassical Light

The applications of nonclassical light are numerous. The single photon sources discussed above have their main applications in quantum communication with single photons where they may be used to establish secure signal transmission. There is also a decisive advance in the field of quantum information expected from the combination of local atomic quantum processing with photonic transmission of quantum states. Conceptually, new possibilities would arise to enhance quantum computational capabilities. For example, separate local quantum processors of moderate size could be linked in a quantum network, with quantum information being exchanged by means of photons.

Although a large number of phenomena have been predicted in which the quantum nature of the squeezed vacuum plays an essential role in modifying atomic properties, only the linear dependence of the two-photon absorption rate of a three-level ladder atom has been confirmed experimentally but classical sources show a quadratic dependence. In 1998 an important series of experiments was reported which investigated the interaction of a beam of Cs atoms with squeezed light in the Fabry–Perot cavity environment. The principal phenomena examined were the modification of the transmission spectrum of a weak, tunable probe beam due to the presence of the squeezed light, and the variation of the transmission modulation with the phase of the squeezed vacuum relative to that of an applied, saturating, coherent field.

There have been many proposed practical applications of squeezed light, and some experiments have already demonstrated its potential in precision measurements and improvement of spectroscopic sensitivity. Squeezed light may be used in optical communications to enhance signal-to-noise ratio, gravitational wave detection, and quantum nondemolition measurements.

The wider applications of squeezed light are numerous. An idea of the range and variety of applications is provided by the following selection: the protective measurement of a single squeezed harmonic oscillator state, coherent nonlinear spectroscopy using number squeezed light, the improvement of quantum nondemolition measurements using a squeezed meter input, sub-shot-noise Doppler anemometry with amplitude squeezed light, noiseless transfer of nonclassical light through bistable systems, generation of few photon states from squeezed atoms, and spin squeezing in a collection of atoms illuminated with squeezed light.

Einstein–Podolsky–Rosen (EPR) Correlations and Entanglement of Photons

In 1986, a quantum nondemolition experiment confirming the quantum correlation of the quadrature phase amplitudes of two spatially separated fields, was reported. In 1987, two-mode squeezing with two spatially separated detectors was experimentally demonstrated using parametric oscillation. With the experimental realization of such two-mode squeezing and quantum nondemolition measurements, the possibility of testing for continuous variable EPR-type correlations became apparent. In these proposals, the conjugate 'position' and 'momentum' observables would be the two orthogonal quadrature phase amplitudes of the field. The high efficiency of detectors used in the quadrature phase amplitude squeezing experiments was an important issue in this connection. Experimental tests of EPR correlations at the time were based on Bohm's version of the EPR paradox, and in fact, focused on Bell's extension of this work. These tests involved measurements performed on individual photon pairs, and used very inefficient detectors.

A specific proposal and a criterion to demonstrate the continuous variable EPR correlations for real experiments by way of a realization of the 1935 EPR paradox, was put forward in 1989. This proposal employed the two-mode squeezed state as the EPR source. The first experimental achievement, using the parametric oscillator, of the 1989 EPR criterion, for efficient measurements with continuous variable outcomes, was reported in 1992.

The EPR fields have proved significant in enabling the experimental realization of continuous variable quantum teleportation, and may have application also to quantum cryptography. Again, because of the efficient nature of the measurement of field quadrature phase amplitudes, and the intrinsic macroscopic nature of the correlated fields, continuous variable teleportation and cryptography may provide large advantages over protocols using a finite number state basis, where signals of individual photons are sent. At the moment, several groups are working in this promising direction.

Conclusions

In this contribution the generation of nonclassical light was reviewed. The generation of nonclassical light became possible after stable and reliable laser sources were available. There are many sources of nonclassical light available now. It is clear that we will see many new applications in the future. Especially in connection with optical communication, rather interesting improvements can be expected.

See also

Coherent Lightwave Systems. **Photonic Crystals:** Electromagnetic Theory. **Quantum Optics:** Atomic Coherence Effects.

Further Reading

Caves CM and Drummond PD (1994) Quantum limits on Bosonic communication rates. *Review of Modern Physics* 44: 481–537.

Cresser JD, Häger J, Leuchs G, Rateike FM and Walther H (1982) Resonance fluorescence of atoms in strong monochromatic laser fields. *Topics in Current Physics* 27: 21–59.

Dagenais M and Mandel L (1978) Investigation of two-time correlations in photon emissions from a single atom. *Physical Review A* 18: 2217–2228.

Diedrich F and Walther H (1987) Nonclassical radiation of a single stored ion. *Physical Review Letters* 58: 203–206.

Drummond PD and Ficek Z (eds) (2004) Springer Series on Atomic, Optical, and Plasma Physics, vol. 27. Berlin, Heidelberg, New York: Springer Verlag.

Gatti A, Lugiato LA and Wiedemann H (1997) In: Giacobino E, Reynaud S, and Zinn-Justin J (eds). *Quantum Fluctuations*, p. 431. Amsterdam: Elsevier.

Georgiades NP, Polzik ES, Edamatsu K, Kimble HJ and Parkins AS (1995) Nonclassical excitation for atoms in a squeezed vacuum. *Physical Review Letters* 75: 3426–3429.

Glauber R (1963) The quantum theory of optical coherence. *Physical Review* 130: 2529–2539.

Glauber R (1963) Coherent and incoherent states of the radiation field. *Physical Review* 130: 2766–2788.

Guthöhrlein GR, Keller M, Hayasaka K and Lange W (2001) A single ion as a nanoscopic probe of an optical field. *Nature* 414: 49–51.

Heidmann A, Horowicz RJ, Reynaud S, Giacobino E, Fabre C and Camy G (1987) Observation of quantum noise reduction on twin laser beams. *Physical Review Letters* 59: 2555–2557.

Höffges JT, Baldauf HW, Eichler T, Helmfrid SR and Walther H (1997) Heterodyne measurement of the fluorescent radiation of a single trapped ion. *Optical Communications* 133: 170–174.

Hood CJ, Kimble HJ, Turchette QA, Georgiades NPh and Parkins AS (1998) Squeezed excitation in cavity QED: experiment and theory. *Physical Review A* 58: 4056–4077.

Jakeman E, Pike ER, Pusey PN and Vaugham JM (1977) Effect of atomic number fluctuations on photon antibunching in resonance fluorescence. *Journal of Physics A: Mathematical and General* 10: L257–L259.

Kimble HJ, Dagenais M and Mandel L (1977) Photon antibunching in resonance fluorescence. *Physical Review Letters* 39: 691–695.

Kimble HJ, Dagenais M and Mandel L (1978) Multiatom and transit-time effects on photon-correlation measurements in resonance fluorescence. *Physical Review A* 18: 201–207.

Kimble HJ, Carnal O, Hu Z, Mabuchi H, Polzik ES, Thompson RJ and Turchette QA (1995) *Annals of the New York Academy of Sciences* 755: 87.

Levenson MD, Shelby RM, Reid MD and Walls DF (1986) Quantum nondemolition detection of optical quadrature amplitudes. *Physical Review Letters* 57: 2473–2476.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, New York: Oxford University Press.

Ou ZY, Pereira SF, Kimble HJ and Peng KC (1992) Realization of the Einstein–Podolsky–Rosen paradox for continuous variables. *Physical Review Letters* 68: 3663–3666.

Reid MD (1989) Demonstration of the Einstein–Podolsky–Rosen paradox using nondegenerate parametric amplification. *Physical Review A* 40: 913–923.

Short R and Mandel L (1983) Observation of sub-Poissonian photon statistics. *Physical Review Letters* 51: 384–387.

Short R and Mandel L (1984) Sub-Poissonian photon statistics in resonance fluorescence. In: Mandel L and Wolf E (eds). *Coherence and Quantum Optics*, vol. V, pp. 671–678. New York: Plenum.

Slusher RE, Hollberg LW, Yurke B, Mertz JC and Vally JF (1985) Observation of squeezed states generated by four-wave mixing in an optical cavity. *Physical Review Letters* 55: 2409–2412.

Walls DF (1979) Evidence for the quantum nature of light. *Nature* 280: 451–453.

Wu LA, Kimble HJ, Hall J and Wu H (1986) Generation of squeezed states by parametric down conversion. *Physical Review Letters* 57: 2520–2523.

Yamamoto Y and Machida S (1987) High-impedance suppression of pump fluctuation and amplitude squeezing in semiconductor lasers. *Physical Review A* 35: 5114–5130.

NONLINEAR OPTICS AT THE CRITICAL FIELD LIMIT

G A Mourou, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is now possible to produce, with compact lasers, intensities so large that the relativistic character of the electrons dominates laser–matter interactions. This is a fundamentally new regime in optics where optics, confined to the study of eV characteristic phenomena, are now addressing phenomena with characteristic energies vastly greater in the MeV–GeV (TeV in the near future) range. With relativistic optics, the field of nonlinear optics found a number of novel applications, including fast ignition of inertially compressed fusion target, bright sources of energetic electrons, protons, ions, positrons, pions, etc., but also nuclear physics, high energy physics, astrophysics, and general relativity.

Relative optics has been one of the most active fields of optics and it is impossible to describe all the applications here. So, this new regime will be illustrated with few examples of the key relativistic mechanisms – reminiscent of conventional nonlinear optics, such as relativistic harmonics, relativistic optical rectification (wake-field acceleration), and relativistic self-focusing in refraction and reflection. We finally show that relativistic optics offers a new and efficient gateway to attosecond physics and nonlinear quantum electrodynamics (QED).

Optics is the study of how electrons respond to light. All material optical properties are the consequence of how light interacts with electrons in materials. Because, in the visible regime the photon energy is of the order of 1 eV, optics deals with phenomena with eV characteristic energy. Light is a wave composed of coupled electric and magnetic fields oscillating in synchrony at high frequencies. The electric and magnetic field oscillate perpendicular to each other, and perpendicular to the direction of propagation, k. Because of the electric field E, electrons under the action of the Lorentz force, $F = eE$, moves a distance $x = a_1 E$ transversally to the propagation direction. In classical optics, the oscillation amplitude x is small compared to the wavelength, and the electrons oscillate and radiate at the laser frequency. The radiated field amplitude and phase depend on how the electron is bound to the atom of the material. The invention of the laser in 1960 gave access to much larger intensities and electromagnetic fields. The laser field became strong enough that the electron displacement, x, in this field could no longer be considered as strictly proportional to the laser field. This departure from linearity is addressed by the addition of new terms in the Taylor series expansion of the displacement:

$$x = a_1 E + a_2 E^2 + a_3 E^3 + \cdots \qquad [1]$$

These additional terms are responsible for all the effects that form the field of nonlinear optics. Note here that the field is still not strong enough to ionize an electron from its parent atom and is therefore below the field that could irreversibly damage the material. Thus, the typical intensity must be kept to a value of less than $\sim 10^{14}$ W/cm^2, where the quiver energy E_Q starts to challenge the ionization energy of a few eV, depending on the material:

$$E_Q = \left\langle \frac{e^2 E^2}{4 m \omega^2} \right\rangle \qquad [2]$$

The bracket denotes the time average over one cycle and ω is the laser angular frequency. Above this limit, the material response is dominated by multiphoton ionization or tunneling, as shown in Figure 1. During the ionization event, high harmonics are produced. Note here again that because the electron is still bound to its nucleus during the emission, we can consider the process of high harmonic generation as bound-electron nonlinear optics. As the intensity increases from 10^{14} to 10^{18} W/cm^2 (Figure 1), we are entering the regime of relativistic intensity or relativistic nonlinear optics. This is a regime that was predicted and discussed in the 1960s but that could be accessed only after the technique of chirped pulse amplification (CPA) was demonstrated, followed by significant progress in the areas of short pulse generation and pulse focusing.

Relativistic Optics

In classical optics one always assumes that during a laser oscillation period the electron mass m_0 stays constant. It is a reasonable assumption since the quiver velocity v of the electron is small compared to the speed of light c. As the intensity is increased a point is reached where the electron quiver velocity approaches the speed of light. This has several implications; first, the mass of the electron changes to become

$$m = \gamma m_0 \qquad [3]$$

where γ is the Lorentz factor given by

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1} \qquad [4]$$

Second, because v is close to c, the magnetic force component needs to be added and the Lorentz force becomes:

$$\vec{F} = e\vec{E} + \frac{\vec{v}}{c} \times \vec{B} \qquad [5]$$

Until recently, in classical optics and laser physics, the $v \times B$ component was totally ignored. However, in the relativistic regime, this component is as large as

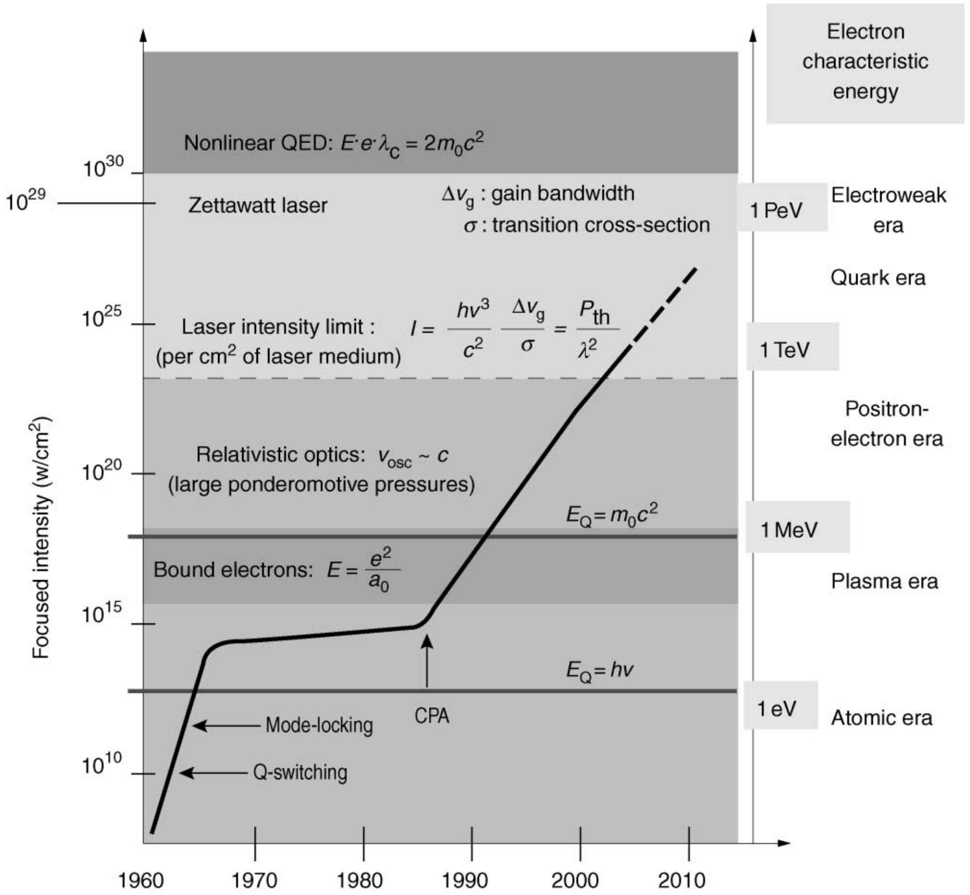

Figure 1 Laser intensity versus years. Note the very steep slope in intensities that occurs during the 1960s. This period corresponds to the discovery of most nonlinear optics effects. Since 1985 after the CPA invention followed by the successive laser short pulse improvements, we are experiencing a similar increase in intensity opening a new regime in optics at 10^{18} W/cm^2 dominated by the relativistic character of the electrons.

the electric one. It pushes the electron forward and is the tenet of the relativistic regime.

At this point, it is convenient to introduce the normalized vector potential:

$$a_0 = \frac{e\lambda E}{m_0c^2} \qquad [6]$$

which represents the energy gain by the electron over one period (λ) normalized to the electron rest mass energy. At $a_0 = 1$, corresponding to 10^{18} W/cm^2 for 1 μm wavelength light, the laser–matter interaction is strongly dominated by the relativistic character of the electrons. At these intensities, the motion of the electrons becomes relativistic and the trajectory of the electron dramatically changes. Figure 2 depicts the difference between the two regimes. It can be shown that independently of the intensity regime, the electron transverse motion is always proportional to a_0, while the longitudinal motion is always proportional to a_0^2. This means that for classical

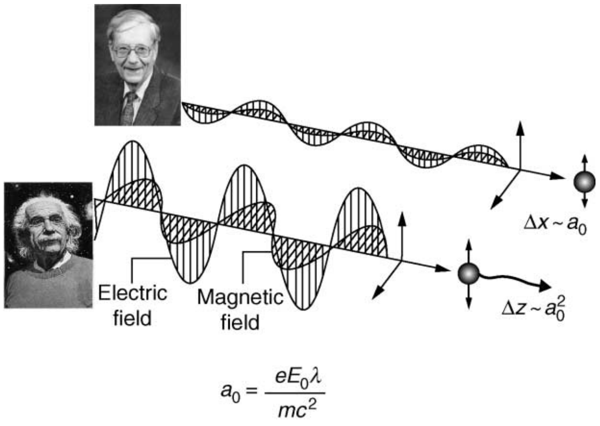

Figure 2 A comparison between classical optics and relativistic optics. The transverse electron displacement scales as a_0 and the longitudinal scales as a_0^2. (a) In classical optics $a_0 < 1$ the transverse electron displacement is proportional to a_0. It is larger than the longitudinal proportional to a_0^2. (b) In relativistic optics it is the opposite, $a_0 > 1$ the transverse displacement becomes smaller than a_0^2.

optics or even bound-electron nonlinear optics with a_0 varying from 10^{-10} to 10^{-4}, the longitudinal motion is $\sim 10^{-10}$ to 10^{-4} smaller than the transverse motion. Of course, the situation reverses completely for $a_0 > 1$, where the longitudinal motion becomes larger than the transverse motion. The departure of the electron motion from the linear regime, due to its relativistic character, gives rise to a new kind of nonlinear optical effects that constitutes the field of relativistic nonlinear optics. This area has become one of the most active in laser physics. In many cases, there are strong similarities between conventional and relativistic optics, and it is the purpose of this article to illustrate the field of relativistic optics by using some of the effects well established in bound electron nonlinear optics.

Considering the rapid progress in laser intensities that the field experienced in the 1960s, as indicated in Figure 1, scientists expected to observe the manifestation of relativistic effects in the 1970s. Formidable intensities on the order of 10^{18} W/cm^2 for 1 μm wavelength light were necessary, however, and could not be readily obtained. Such intensity levels could only be produced in the early 1980s with large-scale sources such as the Helios laser at Los Alamos National Laboratory (LANL) (Figure 3a). Work in relativistic optics virtually could not be conducted in university research laboratories, but rather only in government laboratories. The first relativistic harmonics produced by the critical surface nonlinear motion were observed in 1980 with a kJ-pulse-energy laser at mHz repetition rates (one shot/20 m).

The advent of CPA (Figure 4) revolutionized the field of relativistic optics, placing it within reach of more researchers through enormous reductions in laser size and cost. Further advances in the ultrashort-pulse duration of lasers and in laser focusing made relativistic optics more accessible, especially to universities. Relativistic intensities can be achieved with mJ energy levels at kHz repetition rates by university-size lasers (Figure 3b) instead of with large, kJ, sub-Hz-repetition-rate government-type lasers. Figures 3a and b illustrate the revolution in the relativistic laser evolution representing the Helios (LANL) system. Figure 3b depicts the relativistic λ^3 laser target chamber. The λ^3 laser reaches its relativistic intensity (10^{18} W/cm^2) with pulse energy only on the order of 1 mJ and pulse duration on the order of 10 fs, i.e., a few light periods. Its energy is concentrated on a single wavelength spot size using an f#1 paraboloid corrected by a deformable mirror. Note that all the energy is contained in a focal volume of a few λ^3, hence its name.

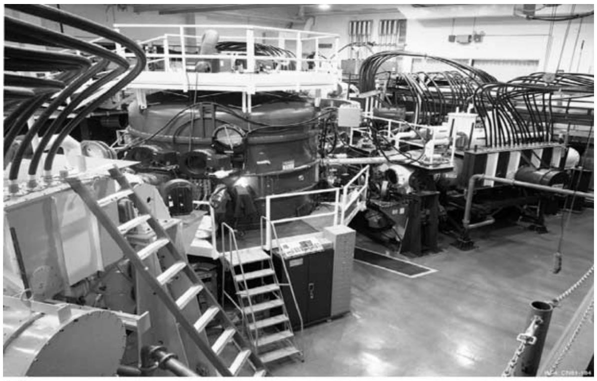
(a)

(b)

Figure 3 Two relativistic lasers. (a) Helios at LANL, was the first relativistic laser with $a_0 \sim 1$ at mHz repetition rate. (b) The λ^3 at the University of Michigan has an $a_0 \sim 1$ at kHz repetition rate.

Some Examples of Relativistic Nonlinear Optics

We will illustrate the field of relativistic optics by comparing relativistic effects with some well-known nonlinear effects.

(a) Relativistic optical rectification (laser wake-field acceleration of electrons and ions)

One well-known bound-electron optical effect coming from the second-order term in eqn [1] is optical rectification. This effect, which takes place in noncentrosymmetric crystals, gives rise to a DC field transverse to the propagation direction. In the relativistic regime, a similar effect known as wake-field acceleration will also contribute to rectification of the optical field. We have seen that in vacuum, electrons are strongly pushed forward by $v \times B$ force. In a plasma, the electromagnetic wave will also push

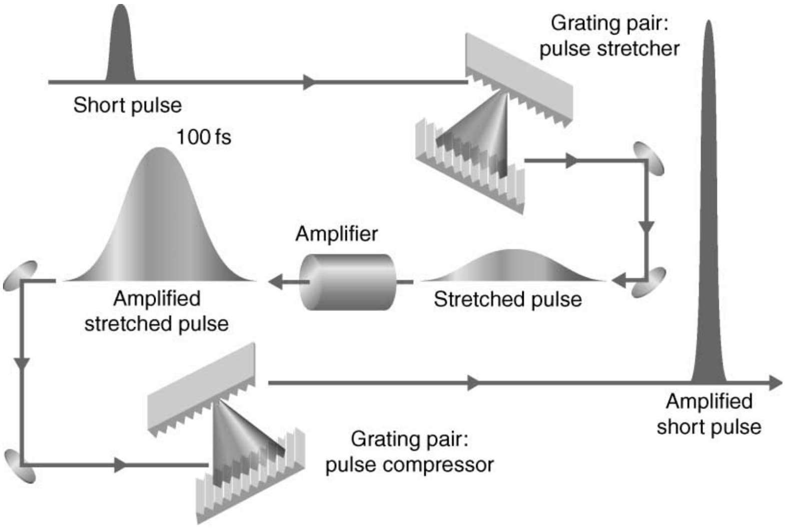

Figure 4 The concept of CPA is the key technique to get relativistic intensity. An initial short pulse is first stretched out (chirped) by a factor $\sim 10^4$, for instance by a pair of diffraction gratings. The stretched pulse intensity becomes lowered by the same factor of $\sim 10^4$. The stretched pulse intensity can then be amplified to a level 10^4 times higher. Once the laser energy has been extracted from the amplifier the stretched pulse is recompressed by the same factor ($\sim 10^4$) to its initial value. This way the laser intensity has been boosted to 10^4 times the peak intensity that the amplifier could have withstood.

forward the electrons, dragging the much heavier positive ions behind, as shown in Figures 5a and b. The charge separation that ensues will produce a strong electrostatic field that can be of the order of the transverse light electric field, which is huge. Unlike in classical optical rectification, in the relativistic regime the effect can be very efficient. Of the order of 10% or more of the transverse field may be rectified. For an intensity of 10^{18} W/cm^2, the electric field can be on the order of 10^{12} V/m (CERN on one meter), or 10^5 times larger than the field that can be obtained by more conventional technology. In addition, the rectified field is flipped along the longitudinal direction. Note also that since the field is created within a plasma, the medium cannot be broken down further and can support these extremely high fields. This effect had to wait for the relativistic-intensity lasers to become a reality before being demonstrated. Today energies up to 200 MeV have been demonstrated by using a 50 TW laser focused on a gas jet. Because the electrons drag the ions, a beam of ions is also observed concomitant with the electron beam. The generation of protons with energy up to 56 MeV has been reported by LLNL.

(b) Relativistic harmonic generation

Harmonic generation in the relativistic regime can be produced in two ways, as a result of the laser plasma interaction with electron density n_e, either below or above the critical density n_c. The first is for $n_e < n_c$ and corresponds to an underdense plasma where light can propagate. Harmonics generation was predicted in the late 1960s. Even-order harmonics are observed with linear and circular polarizations. This is the signature of relativistic effects since even-harmonics cannot occur in conventional nonlinear optics in centrosymmetric media.

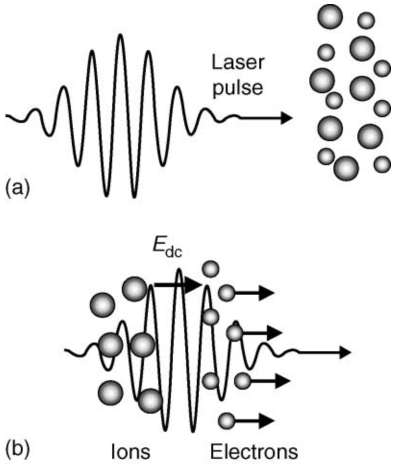

Figure 5 Relativistic rectification in plasma. (a) The high intensity pulse before entering the plasma. (b) Illustrate the $v \times B$ pushing first the plasma electrons. The electrons then drag the heavy ions behind them like a horse pulling a cart. The longitudinal electrostatic field that is generated is almost as large as the transverse laser field.

Harmonic generation can also be produced in over-critical plasmas, where $n > n_c$. In this case, the light cannot propagate and be reflected. The strong $v \times B$ oscillatory force applied to the critical surface will produce a self-modulated reflected beam. This effect was first observed using the Helios laser. The idea of the moving-mirror model was first put forward in the early 1980s. Generation of frequencies up to the 70th harmonic has been demonstrated so this method could provide an efficient way to produce high harmonics. This technique has the advantage over conventional high harmonic generation in that it is capable of accommodating a pulse of large energy, since one can focus the beam over a small area, limited only by the laser wavelength, without fear of disrupting the plasma.

(c) Relativistic self-focusing in refraction

In bound-electron nonlinear optics, the third-order term in eqn [1] is responsible for self-focusing. A similar effect also exists in relativistic optics; it is described in Figure 6a. In the relativistic regime, the electron driven by the laser field changes its mass. In turn, this mass change will modify the index of refraction according to:

$$n^2 = 1 - \left(\frac{\omega_p}{\omega}\right)^2 \qquad [7]$$

where the plasma frequency is given by

$$\omega_p^2 = \frac{e^2}{m} n_e \qquad [8]$$

where n_e is the plasma electron density, and m is the intensity-dependent mass according to eqn [1].

For $\omega > \omega_p$, the index of refraction is positive and the pulse can propagate in the plasma. For the higher-beam-intensity part of the beam, i.e., on the beam axis, the plasma frequency will be less than the frequencies corresponding to the pulse wing. This will cause a modification of the laser wavefront and the beam to self-focus.

(d) Relativistic self-focusing in reflection and isolated attosecond pulse generation

For $\omega < \omega_p$, (eqn [7]) is negative and the pulse cannot propagate and so is reflected at the critical surface. However, the large light pressure will push the plasma critical-density surface at twice the laser frequency, making it cancel out the beam spatial intensity profile. In the relativistic λ^3 regime, the laser is focused at one wavelength. If the light is linearly polarized, the electron density will be collectively moved back and forth along the critical surface. This time-dependent deformation of the critical surface will produce a deflection of each electromagnetic half-cycle in a specific direction (Figure 7), providing a simple means for generating isolated attosecond pulses efficiently. The relativistic motion of the critical surface compresses (relativistic Doppler) the incident femtosecond pulse to the attosecond regime. To understand this compression, recall that the incident pulse travels opposite to the mirror motion formed by the critical surface (for a fraction of the cycle). As we have seen in this regime the critical surface moves also at close to the speed of light, v. Since the reflected light must also travel at c, it has no choice but to be compressed, producing an electromagnetic accumulation along the critical surface. The compression factor will

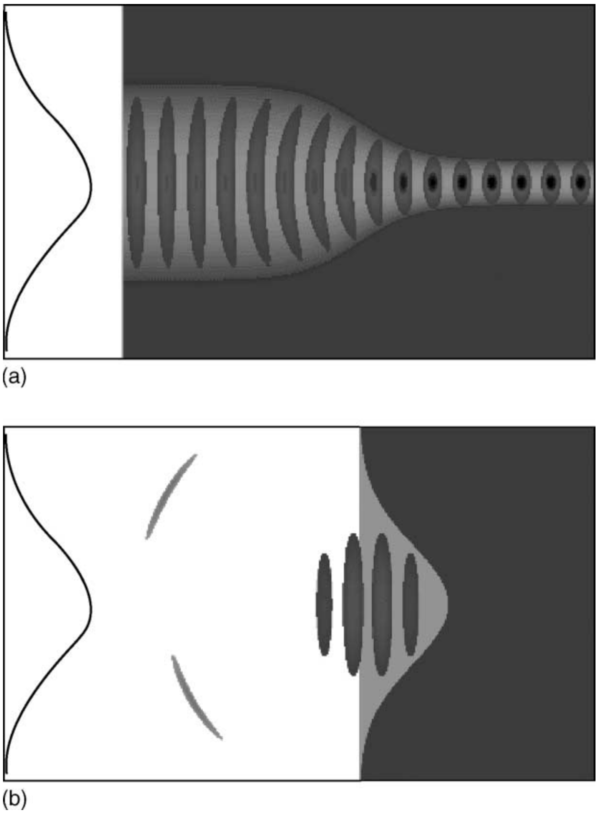

Figure 6 Relativistic self-focusing. (a) Relativistic self-focusing in refraction. Here the index of refraction is modified by the change in the electron mass. The electron mass change will in turn modify the plasma frequency across the beam and consequently the index of refraction. (b) Relativistic self-focusing in reflection showing the deformation of the critical surface by the light pressure from a strongly focused beam. The light is reflected in a series of ultrashort pulses in the attosecond domain.

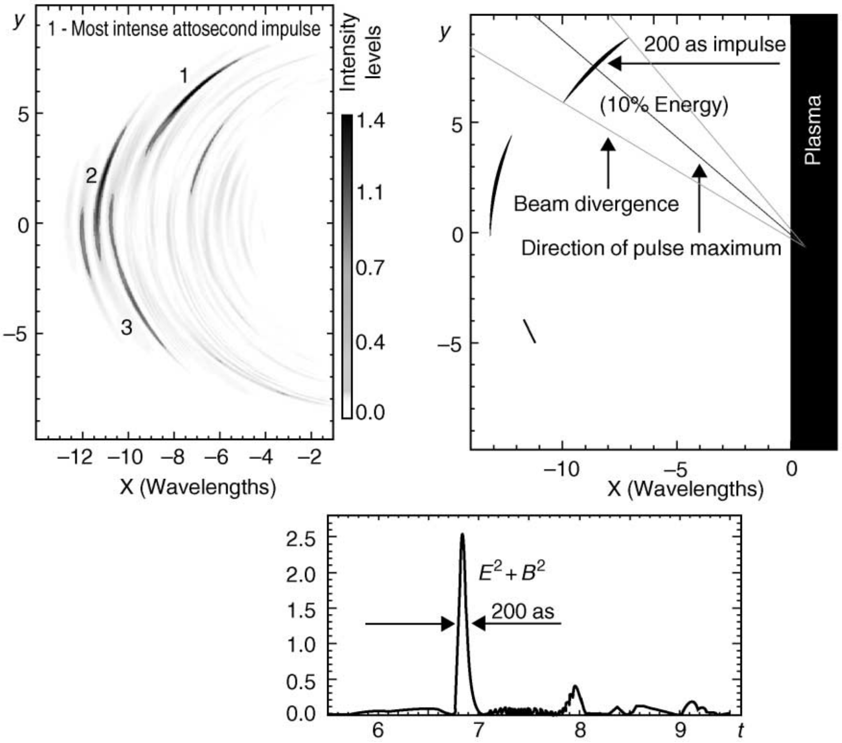

Figure 7 Compressed and deflected attosecond pulses. The combination of relativistic deflection and compression leads to 200 as isolated pulses.

depend on the critical surface velocity and is dependent on the plasma stiffness, i.e., the plasma density. Low-density plasmas are softer than high-density ones; here, n_c is the plasma critical density. The compressed pulse duration has been determined by a particles-in-cell (PIC) simulation to be given by $\tau = 600/a_0$. From the PIC simulation, it can be determined that the compressed pulse optimum duration is obtained for $n_e = a_0 n_c/2$. This technique is an interesting alternative to the one based on high-harmonic generation due to bound electrons, which has been recently demonstrated. The advantage of the relativistic optics approach is that it can be scaled up to large energies, from the mJ to the several-joule level, in contrast with demonstrated attosecond schemes that can only compress mJ pulse energies with an efficiency of 10^{-6}. For a_0 ~100, corresponding to incident intensities of ~10^{22} W/cm^2, a 10-as pulse can be produced by relativistic compression and deflection.

(e) Nonlinear QED on a table top

Looking at Figure 1, we see that there is a regime around 10^{29} W/cm^2 where it becomes possible to produce electron–positron pairs. The condition to fulfill this is given by

$$E_s \lambda_c e = 2m_0c^2 \qquad [9]$$

Here $\lambda_c = h/m_0c$ is the Compton wavelength, E_s the Schwinger field, and h the Planck constant. This expression states that for pair creation the work of the laser field over a Compton wavelength must be equal to the electron–positron rest-mass energy. This field E_s, also known as the critical field, E_s ~10^{16} V/cm, corresponds to a laser intensity of 10^{29} W/cm^2.

An alternative way to depict the process of electron–positron generation from vacuum is by considering the vacuum as a dielectric medium with a gap $E_G = 2m_0c^2$. The field to produce an electron–positron pair is the one necessary to bend the bands so that an electron can tunnel from the Dirac sea into the vacuum states (Figure 8). The probability for such an event is given by

$$W \propto \exp\left(-\frac{\pi E_s}{E}\right) \qquad [10]$$

Today, the intensity necessary to reach pair creation is more than six orders of magnitude higher

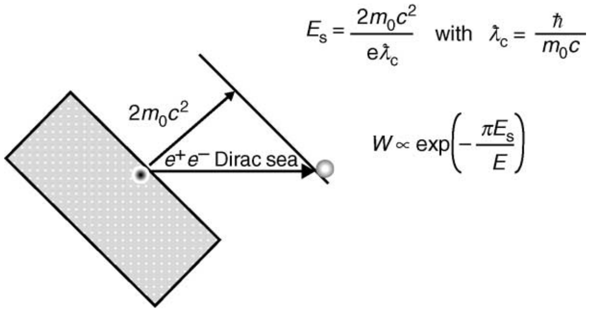
$E_s = \frac{2m_0c^2}{e\bar{\lambda}_c}$ with $\bar{\lambda}_c = \frac{\hbar}{m_0c}$
$2m_0c^2$
e^+e^- Dirac sea
$W \propto \exp\left(-\frac{\pi E_s}{E}\right)$

Figure 8 Pair creation. Electron tunneling from the Dirac sea into the vacuum state.

than what we could produce, $\sim 10^{22}$ W/cm^2. To bridge this enormous gap the electric field has been enhanced by colliding a short and intense pulse with a Stanford Linear Accelerator Center (SLAC) super-relativistic, 50 GeV electron beam. The Lorentz factor is $\gamma = 10^5$, and the laser field of $2 10^{10}$ V/cm (corresponding to an intensity of 10^{18} W/cm^2) can be enhanced by 10^5 to reach the critical field.

The progress in relativistic compression described above could make possible the generation of electron–positron pairs directly from the laser. We are on the verge of having a tabletop laser that could produce an intensity close to 10^{23} W/cm^2, corresponding to $a_0 > 100$. From the scaling law described previously, we should get $\tau < 6$, where the wavelength of such a compressed pulse would also be reduced by ~100. The focal volume would then be reduced by six orders of magnitude. If we assume an efficiency of 10^{-1}, an intensity close to the critical intensity of 10^{29} W/cm could be reached with a tabletop laser and boil vacuum.

Conclusion

Today, laser intensities are such that during light–matter interaction, the electron motion is dominated by relativistic effects responsible for a new type of nonlinear optics called relativistic nonlinear optics. Relativistic nonlinear optics is characterized by extreme field gradients (TV/m) and light pressure in the Gbar regime. Relativistic nonlinear optics offers an attractive and novel way for efficient attosecond pulse generation with intensities near the critical (Schwinger) field, opening optics to vast new horizons.

See also

Relativistic Nonlinear Optics.

Further Reading

Agostini P, Barjot G, Bonnal JF, Mainfray G and Manus C (1968) *IEEE Journal of Quantum Electronics* QE-4: 667.

Albert O, Wang H, Liu D, Chang Z and Mourou G (2000) *Optics Letters* 125: 1125.

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Boyd RW (1992) *Nonlinear Optics*. San Diego: Academic.

Brabec T and Krausz F (2000) *Review of Modern Physics* 72: 545.

Brezin E and Itzykson C (1970) *Physics Review D* 2: 1191.

Bula C, McDonald KT, Prebys EI, *et al.* (1996) *Physics Review Letters* 76: 3116.

Bulanov SV, Naumova NM and Pegoraro F (1994) *Physics Plasmas* 1: 745.

Carman RL, Benjamin RF and Rhodes CK (1981) *Physics Review A* 24: 2649.

Chen S-Y, Maksimchuk A, Esarey E and Umstadter D (2000) *Physics Review Letters* 84: 5528.

Eberly J (1969) *Progress in Optics* 7: 361.

Franken PA, Hill A, Peters C and Weinreich G (1961) *Physics Review Letters* 7: 118.

Hentschel M, Kienberger R, Spielmann Ch, *et al.* (2001) *Nature* 414: 509.

Joshi C and Katsouleas T (2003) Plasmas accelerators at the energy frontier and on tabletops. *Physics Today* 56: 47.

Litvak AG (1969) *Soviet Physics JETP* 30: 344.

Malka V, Fritzler S, Lefebvre E, *et al.* (2002) Electron acceleration by a wakefield forced by an intense ultrashort laser pulse. *Science* 22: 298.

Max C, Arons J and Langdon A (1974) *Physics Review Letters* 33: 209.

Mourou GA, Barty CPJ and Perry MD (1998) *Physics Today* 51(1): 22.

Naumova NM, Nees JA, Sokolov IV, Hou B and Mourou GA (2004) Relativistic generation of isolated attosecond pulses in a λ^3 focal volume. *Physics Review Letters* 92: 063902.

Norreys PA, *et al.* (1996) *Physics Review Letters* 76: 1832.

Sarachik E and Schappert G (1970) *Physics Review D* 1: 2738.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Snavely RA, *et al.* (2000) *Physics Review Letters* 85: 2945.

Strickland D and Mourou G (1985) *Optics Communications* 55: 219 and 447.

Tajima T and Dawson JM (1979) *Physics Review Letters* 43: 262.

Tarasevitch A, Orisch A, von der Linde D, *et al.* (2000) *Physics Review A* 62: 023816.

Umstadter D, Chen S-Y, Maksimchuk A, Mourou G and Wagner R (1996) *Science* 273: 472.

NONLINEAR OPTICS, APPLICATIONS

Contents

Phase Matching
Pulse Compression via Nonlinear Optics
Raman Lasers
Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)
Three-Dimensional Microfabrication

Phase Matching

A V Smith, Sandia National Laboratories,
Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase matching is important in any process that coherently transfers energy from one optical mode to another over a long interaction length. We will start by illustrating the principle, by imagining that the original mode is a single transverse mode of a waveguide, and that microscopic scattering centers distributed uniformly throughout the waveguide weakly scatter light into a second mode with orthogonal polarization. Then suppose that the scattering process preserves phase so the phase of the wavelet scattered at each point into mode two is the same as that of the original wave in mode one. Such a process is shown in Figure 1 where the propagation constant of the initially excited mode is k_1 and the propagation constant of the second mode is k_2. At each z location the light scattered into the second mode has the phase $\phi = k_1 z$, and after propagating to the end of the waveguide ($z = L$) with the phase velocity characteristic of the second mode, it has the phase $\phi_z = [k_1 z + k_2(L - z)] = [k_2 L + z(k_1 - k_2)]$. Coherently summing the fields scattered at all of the z positions, gives the net output field in the second mode. Figure 2 shows how this field can be constructed graphically. The rotation angle of each phasor represents its phase, ϕ_z, while the length of each phasor represents the magnitude of the field contribution from the corresponding z segment. It is apparent that the energy transfer is maximized if the contributions from all z positions have the same phase, in which case the phasors form a straight line, and the power in the second mode is proportional to L^2. This corresponds to $(k_1 - k_2 = 0)$. When this condition is met the waves in the two

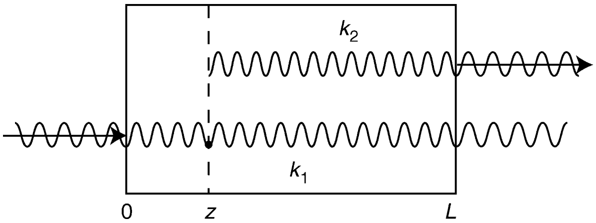

Figure 1 An input wave with propagation constant k_1 in a waveguide mode is weakly scattered into a second waveguide mode with propagation constant k_2. The phase mismatch is $\Delta k \equiv k_2 - k_1$.

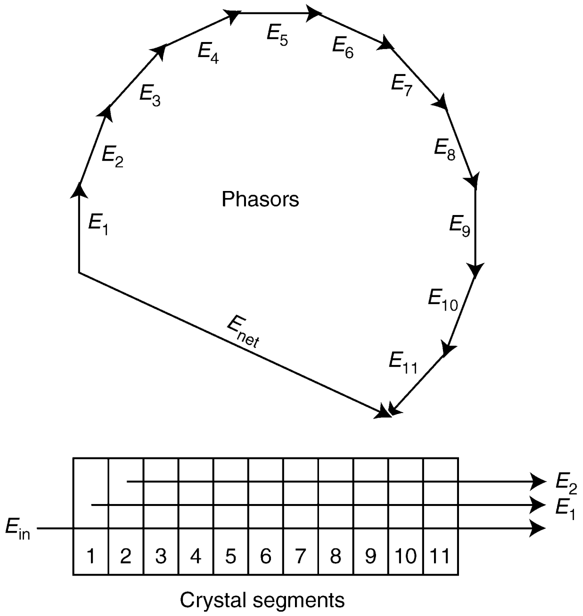

Figure 2 Phasor diagram showing the net output field contributions from segments 1–11. The net field at the output end, E_{net}, is the vector sum of the phasors. The phasors are complex quantities whose angles represent their relative phases according to $E_n = E_o \exp(i\phi_n)$ where $\phi_n = \phi_o - \Delta k(z_n - z_1)$.

modes have equal phase velocities, and the process is phase matched.

The quantity $(k_2 - k_1)$ is called Δk, the phase mismatch. If $\Delta k \neq 0$ the phasors curl to form a section of a circle, reducing the net field in mode two. If all the phasors have the same length, as they would if the wave in mode one is undepleted, the size of the mode two field at position z is

$$E(z) = Kz\frac{\sin(\Delta kz/2)}{\Delta kz/2} = Kz\,\mathrm{sinc}\left(\frac{\Delta kz}{2}\right) \qquad [1]$$

where K is a constant describing the scattering strength. The corresponding power in mode two is plotted in Figure 3 for various values of Δk. When $\Delta k = 0$ the power grows quadratically with z, but for $\Delta k \neq 0$ it is proportional to $\sin^2(\Delta kz/2)/\Delta k^2$. If we plot the normalized output power in mode two $(z = L)$ versus Δk, we have the curve shown in Figure 4 which is $\mathrm{sinc}^2(\Delta kL/2)$.

When $\Delta k = 2\pi/L$ the phasors form a complete circle so the power in mode two first grows with z but then shrinks back to zero at $z = L$. For a given value of Δk the length for which the phasors form a complete circle is known as the coherence length, $L_{\mathrm{coh}} = 2\pi/\Delta k$.

Reducing the z increments to infinitesimals, we can write the output field for mode two in integral form:

$$E_{\mathrm{net}} = \int_0^L S(z)e^{-i\Delta kz}\,\mathrm{d}z \qquad [2]$$

This shows that when Δk is independent of z, the output field is the Fourier transform of a source term, $S(z)$, transformed into Δk space. Where $S(z)$ has a constant value from $z = 0$ to $z = L$ and is zero outside that range, the Fourier transform has the form given by eqn [1], with $z = L$.

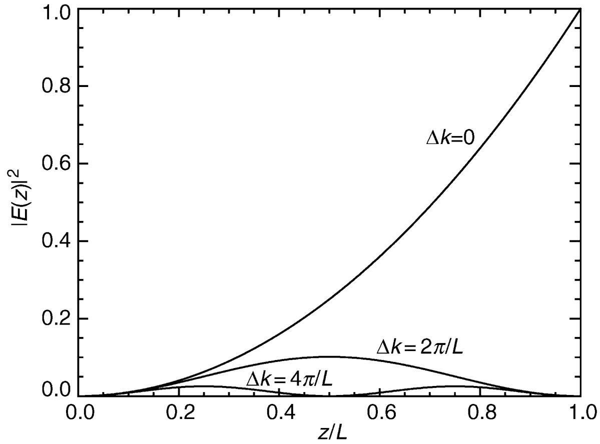

Figure 3 Plot of power in second mode versus z for different values of Δk. For $\Delta k = 0$ the power is proportional to z^2. For $\Delta k = 2\pi/L$ the phasors form a closed circle so the net field is zero at $z = L$. For $\Delta k = 4\pi/L$ the phasors trace two complete overlapped circles so again the field is zero at the $z = L/2$ and at $z = L$.

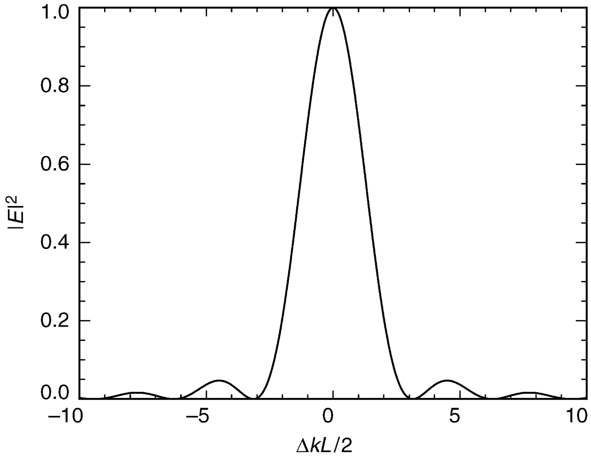

Figure 4 The plotted function indicates the relative irradiance of the output field, $|E|^2$, versus the phase mismatch Δk. The full width at half maximum is $\Delta k = 1.77\pi/L = 5.57/L$.

The $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ form for the output power is ubiquitous in functions that describe the efficiency of nonlinear optical interactions, as well as acousto-optic and electro-optic modulation of light waves. However, it should be kept in mind that $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ holds only when $S(z)$ is independent of z. If the scattering is strong enough to alter the phase or amplitude of the driving wave, or if the coupling strength varies with z for any other reason, $S(z)$ is not independent of z and the $L^2\,\mathrm{sinc}^2(\Delta kL/2)$ dependence is modified.

Our simple example relied on scattering in a waveguide, with a source term that was linear in the field in mode one. In practice, most optical devices that couple light between modes, rely on nonlinear interactions among multiple optical fields, or between an optical field and an acoustic or an rf electric wave. These nonlinear interactions produce a polarization wave in a suitable material at a sum or difference frequency. Each point of this polarization radiates a wavelet in all directions. These wavelets interfere constructively only in directions that are nearly phase matched. Examples include traveling wave acousto-optic filters, in which an acoustic wave traveling parallel to an input light beam of one polarization in a birefringent material, couples light into the orthogonal polarization via the elasto-optic effect. Traveling wave electro-optic modulators are similar but rely on the linear electro-optic effect for the coupling. Optical frequency mixing uses the nonlinearity of various materials to mix multiple optical waves, creating polarization waves at sum and difference frequencies. A wide variety of such nonlinear interactions is discussed in textbooks on nonlinear optics, electro-optics, and acousto-optics. In all of these cases the phase velocity mismatch of interest is that between an induced polarization wave and a radiated mode, and the phase mismatch is $\Delta k = k_{\text{radiation}} - k_{\text{polarization}}$.

To find $S(z)$ and Δk for these nonlinear mixing processes, we must construct the polarization wave. We consider only monochromatic plane waves here, considering beams with spatial or temporal structure in a later section. The electric field of a monochromatic plane optical wave propagating in direction $\mathbf{r}_o$ can be described by

$$\begin{aligned} E &= E_o \cos(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o + \phi_o) \\ &= \tfrac{1}{2}[E\, e^{-i(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o)} + E^*\, e^{i(\omega_o t - \mathbf{k}_o \cdot \mathbf{r}_o)}] \end{aligned} \quad [3]$$

The complex notation is a convenience that simplifies algebraic book-keeping. Similarly, an acoustic field propagating in direction $\mathbf{r}_a$ can be written:

$$\begin{aligned} A &= A_o \cos(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a + \phi_a) \\ &= \tfrac{1}{2}[A\, e^{-i(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a)} + A^*\, e^{i(\omega_a t - \mathbf{k}_a \cdot \mathbf{r}_a)}] \end{aligned} \quad [4]$$

In a suitable material the polarization induced by the product of the optical and acoustic waves has a term of the form:

$$P = sEA \quad [5]$$

P has components at both the sum and difference frequency. The sum frequency component is:

$$P_{\omega_o + \omega_a} = \frac{1}{4} s[EA\, e^{-i(\omega_o + \omega_a)t + i(\mathbf{k}_o + \mathbf{k}_a)\cdot \mathbf{r}} + \text{cc}] \quad [6]$$

where 'cc' indicates the complex conjugate of the first term. The component at the difference frequency is

$$P_{\omega_o - \omega_a} = \frac{1}{4} s[EA^*\, e^{-i(\omega_o - \omega_a)t + i(\mathbf{k}_o - \mathbf{k}_a)\cdot \mathbf{r}} + \text{cc}] \quad [7]$$

An important point is that the k vector of the sum-frequency polarization is $(\mathbf{k}_o + \mathbf{k}_a)$. Phase matching requires that the radiated sum-frequency wave has the same propagation vector. The radiated difference frequency wave must have the propagation vector $(\mathbf{k}_o - \mathbf{k}_a)$ to be phase matched. Phase matching for these two cases is shown in Figure 5. Note the general rule that if the frequency of a wave appears as $+\omega$ in the induced polarization, the associated k vector contribution to the polarization wave is $+\mathbf{k}$, while if it appears as $-\omega$, the k vector contribution is $-\mathbf{k}$. For example, the acoustic wave in $P_{\omega_o+\omega_a}$ is associated with $+\mathbf{k}_a$ contribution to the polarization wave, while in $P_{\omega_o-\omega_a}$ it is associated with a $-\mathbf{k}_a$ contribution.

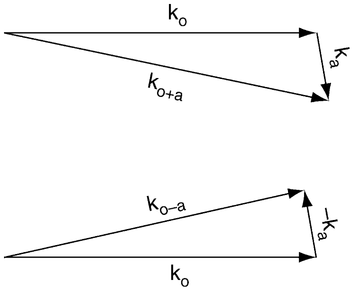

Figure 5 An optical wave with propagation vector k_o and an acoustic wave with propagation vector k_a produce a polarization wave at the sum frequency with the propagation vector k_{o+a} (top diagram), and a polarization wave at the difference frequency with the propagation vector k_{o-a} (bottom diagram).

Using the definition of k:

$$\mathbf{k} = \frac{2\pi}{\lambda}\hat{\mathbf{k}} = \frac{\omega}{v}\hat{\mathbf{k}} \quad [8]$$

where v is the phase velocity, we can write Δk as

$$\begin{aligned} \Delta \mathbf{k} &= \mathbf{k}_{o+a} - \mathbf{k}_o - \mathbf{k}_a \\ &= \left(\frac{\omega_{o+a}}{v_{o+a}}\hat{\mathbf{k}}_{o+a} - \frac{\omega_o}{v_o}\hat{\mathbf{k}}_o - \frac{\omega_a}{v_a}\hat{\mathbf{k}}_a \right) \end{aligned} \quad [9]$$

Clearly, if all of the waves had equal phase velocities, phase matching would be achieved for collinear interactions, where all of the propagation vectors are parallel, because the frequencies inside the parentheses sum to zero. This is seldom the case as for example, the phase velocity of an acoustic wave is much less than that of an optical wave, making an acoustic wave's k-vector disproportionately long compared with that of an optical wave. On the other hand, an rf wave usually has a disproportionately short k-vector compared with an optical wave. Even when all of the waves are optical, dispersion of the refractive index usually makes the k-vector of the bluer waves disproportionately long compared with those of the red waves. A wide variety of methods are used separately or in combinations to compensate for these effects in order to achieve phase matching and these are discussed below.

Phase Matching Techniques

Noncollinear Phase Matching

If one of the low frequency waves in a three wave mixing process, such as the acousto-optic interaction discussed above, has a lower phase velocity than the bluest wave, then noncollinear phase matching may be possible. Acousto-optic deflectors are often phase matched with a noncollinear geometry, such as that shown in Figure 6. An input optical wave of frequency ω_o reflects from a traveling acoustic wave with frequency ω_a in an optically isotropic material to produce an output optical wave of frequency $\omega_o + \omega_a$. The Bragg condition for efficient reflection is

$$\sin\theta = \frac{k_a}{2k_o} \qquad [10]$$

which may be seen from the figure to be the same as the phase matching condition, assuming k_{out} and k_{in} have the same magnitude, which is a good approximation for typical acoustic frequencies less than a few GHz.

Stimulated Brillouin scattering is another example of noncollinear mixing of acoustic and optical waves. One possible geometry has the input optical and acoustic waves traveling in opposite directions to create a reflected optical wave at the sum frequency. The phase matching condition is then

$$k_{\text{acoustic}} \approx 2k_{\text{optical}} \qquad [11]$$

which implies

$$\omega_{\text{acoustic}} = \frac{2n\omega_{\text{optical}}}{c} v_{\text{acoustic}} \qquad [12]$$

The acoustic frequency that phase matches such counter propagating Brillouin scattering is usually a few GHz in liquids and solids.

If noncollinear phase matching in an isotropic material is not possible, as is usually the case for electro-optic modulation and for all-optical sum and difference frequency mixing, it is sometimes possible to use the properties of nonisotropic media to adjust the phase velocity of one or more of the waves. There are several possibilities, such as natural birefringence, form birefringence, photonic band edge tuning, wave guide modal propagation, and various forms of quasi phase matching. Each is briefly described below.

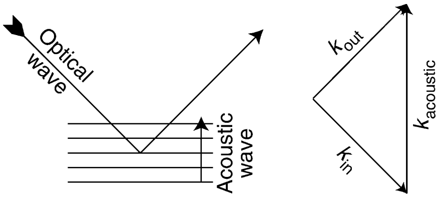

Figure 6 Phase matching in an acousto-optic deflector.

Birefringent Phase Matching

Natural birefringence

Naturally birefringent crystals can be categorized as either uniaxial or biaxial. In either case, light propagating along any direction (except along an optical axis) has two orthogonal eigen polarizations with different refractive indices. In uniaxial crystals, one of these indices is independent of the propagation direction and is called the ordinary index, n_o. The refractive index for the other polarization, known as the extraordinary polarization, is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \qquad [13]$$

where θ specifies the propagation direction measured from the z, or optical, axis, and n_e and n_o are the refractive indices for light polarized parallel and perpendicular to z. The two eigen polarizations are aligned with the lines of constant or latitude (o-wave) or constant longitude (e-wave). The variation of the two refractive indices with propagation angle θ is shown in Figure 7 for a negative uniaxial crystal which is characterized by $n_e < n_o$, and for a positive uniaxial crystal, characterized by $n_e > n_o$.

The acousto-optic tunable filter relies on natural birefringence to achieve a phase match. An optical and an acoustic wave propagate normal to the optical axis of a uniaxial crystal. If the crystal is negative uniaxial, with an input optical polarization of 'e', and an output optical polarization of 'o', the k vector of the acoustic wave must make up the difference

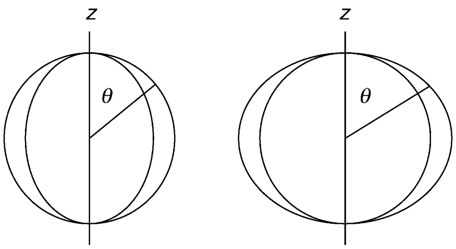

Figure 7 The circle and ellipse represent the two refractive indices for a negative uniaxial crystal (left diagram) and a positive uniaxial crystal (right diagram) for different propagation directions specified by the angles (θ, ϕ) with ϕ fixed. These curves are independent of the azimuthal angle ϕ, so in three dimensions the ordinary refractive index, n_o, forms a sphere in both cases, while the extraordinary index, n_e, forms a prolate ellipsoid for a negative uniaxial crystal, or an oblate ellipsoid for a positive uniaxial crystal.

$k_o - k_e$. This may be expressed as

$$k_a = \frac{\omega_a}{v_a} = \frac{(\omega + \omega_a)n_o}{c} - \frac{\omega n_e}{c} \qquad [14]$$

or, ignoring the difference between ω and $\omega + \omega_a$, the optical frequency that is phase matched is

$$\omega = \frac{\omega_a c}{v_a \Delta n} \qquad [15]$$

where v_a is the phase velocity of the acoustic wave, and $\Delta n = n_o - n_e$. The phase matched optical frequency is seen to be tunable with ω_a. A typical acoustic frequency for a visible filter is a few hundred MHz. The phase matching bandwidth set by the $\mathrm{sinc}^2(\Delta kL/2)$ function can be quite narrow, with a subnanometer optical bandpass in a crystal a few cm long.

Natural birefringence is also widely used to phase match all-optical sum and difference frequency mixing. To counter normal optical dispersion, the bluest wave is polarized in the direction associated with the smaller refractive index, while at least one of the two redder waves has the opposite polarization. By adjusting the propagation angle, θ, the refractive index of the extraordinary waves can be adjusted until exact phase matching is achieved. If the two redder waves both have the high index polarization the process is sometimes called type I, whereas if they have opposite polarizations it is called type II.

Figure 8 illustrates type I phase matched second harmonic generation in a negative uniaxial crystal. The fundamental wave with frequency ω is o-polarized, while the second harmonic wave at frequency 2ω is e-polarized. The circle drawn in dashes represents the line swept out by the polarization wave vector $2k_\omega$ as the propagation direction θ is swept, while the ellipse represents the loci of $k_{2\omega}$ for the same angle. Collinear phase matching is achieved at the point of intersection of the two curves. This sets the angle θ_{pm}. For angles greater than θ_{pm}, dispersion is over-compensated and $\Delta k < 0$ so noncollinear phase matching is possible. All ϕ angles have equivalent propagation properties, so the loci of phase matched propagation directions form a cone at $\theta = \theta_{pm}$. Usually the propagation angle ϕ is chosen to maximize the nonlinear response of the crystal.

Biaxial crystals share with uniaxial crystals the property of two orthogonal eigen polarizations with different refractive indices, but the angular dependence of the indices is more complex. The two indices associated with propagation along direction (θ, ϕ) can be found by solving

$$\frac{\sin^2\theta\cos^2\phi}{1/n^2 - 1/n_x^2} + \frac{\sin^2\theta\sin^2\phi}{1/n^2 - 1/n_y^2} + \frac{\cos^2\phi}{1/n^2 - 1/n_z^2} = 0 \qquad [16]$$

for the two possible values of n. Here n_x, n_y, and n_z are the refractive indices for light polarized along the respective principal axes, where the convention is $n_x < n_y < n_z$. There are two optical axes lying in the x–z plane at an angle determined by the values of n_x, n_y, and n_z. The phase matching loci are no longer lines of constant latitude as they were for uniaxial crystals, but instead are curves that encircle either the optical axes or one of the principal axes. Figure 9 shows an

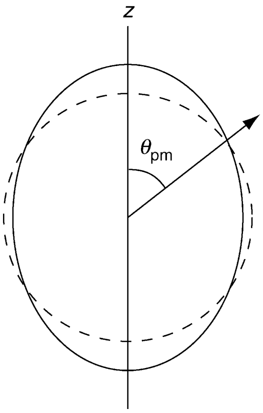

Figure 8 Collinear phase matching of second harmonic generation in a negative uniaxial crystal. The second harmonic light is e-polarized with a refractive index indicated by the solid line while the fundamental light is o-polarized as indicated by the dashed line. The crossing point of the two refractive indices determines the phase matching angle θ_{pm}.

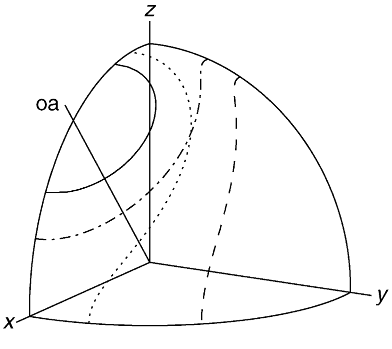

Figure 9 Calculated loci of propagation directions that achieve birefringent phase matching for type I collinear second harmonic generation of different wavelengths in the crystal CsB_3O_5. The fundamental wavelengths are 2,700 nm (dotted curve looping around the x axis), 3,000 nm (dashed curve looping around the y axis), 1,000 nm (solid curve looping around the optical axis 'oa'), and 800 nm (chained curve looping around the z axis). For clarity only a single octant of the sphere of propagation directions is shown. The loci in the other octants of the sphere are formed by reflections in the three principal planes.

example of each type of phase matching curve associated with second harmonic generation of different wavelengths in the crystal CsB_3O_5.

Note that because these curves always intersect at least one of the principal planes, if phase matching in biaxial crystals is possible at all, it is also possible in a principal plane. Biaxial crystals are usually phase matched in a principal plane, but phase matching outside the principal planes can in some cases maximize the nonlinear response, minimize the temperature sensitivity, or adjust the relative group velocities.

Form birefringence

Some materials, such as GaAs, have a strong nonlinear optical response but are not birefringent, so optical frequency mixing cannot be phase matched using natural birefringence. However, form birefringence can be engineered by layered growth, and it can be used to phase match. If many alternating layers of two materials with different refractive indices are stacked together with a layer thickness comparable to the wavelength of light, the material is birefringent for light propagating parallel to the layers. Light polarized with its E field parallel to the layers will have a higher refractive index than light polarized with its field perpendicular to the layers. The strength of birefringence, $(n_{\|} - n_{\perp})$, can be engineered by varying the relative thicknesses and refractive indices of the two layered materials. This has been demonstrated in GaAs/AlAs waveguide structures in which the AlAs layers were oxidized to form Al_2O_3, which has a refractive index of 1.6, compared with an index of 3.4 for GaAs. The resulting large birefringence was used to phase match infrared difference frequency mixing.

Photonic Lattice Phase Matching

The propagation constant of a light wave can be controlled by tuning its frequency near the bandgap of a Bragg structure. The Bragg structure can be a stack of layers with alternating refractive index. If the stack consists of layers of thickness and refractive index l_1, n_1 and l_2, n_2, the Bragg condition at normal incidence is

$$\phi_1 + \phi_2 = \frac{\omega_{\text{Bragg}}}{c}(n_1 l_1 + n_2 l_2) = m\pi \qquad [17]$$

where m is an integer. Frequencies near ω_{Bragg} are strongly reflected. The width of the reflective band of a structure with many layers is proportional to the refractive index contrast, $(n_1 - n_2)/(n_1 + n_2)$. At certain frequencies, on both sides of the bandgap, the reflectivity vanishes. The first transmission peak on the red side of the bandgap has an effective refractive index that is greater than the average $(n_1 + n_2)/2$, while the first peak on the blue side has an effective index less than the average. This effect can be used to adjust the effective value of Δk in the structure. As a bonus, operating on one of the transmission peaks near the bandgap enhances the optical field in the material, increasing the strength of the nonlinear interaction.

It is also possible to use a Bragg structure in reflection mode. Tuning to the red side of the band center increases the effective refractive index of the reflected wave, while tuning to the blue side decreases it. The internal fields are not enhanced in this case and the interaction length may be quite short.

Quasi-Phase Matching (QPM)

Quasi-phase matching uses spatial modulation of the polarization source with a modulation period of $L_{\text{mod}} = 2\pi/\Delta k$ to boost the efficiency of the mixing process even though $\Delta k \neq 0$. We can consider the periodic modulation as another interacting wave that is included in the creation of the polarization wave. The modulation wave has $\omega = 0$ and is stationary with period L_{mod}. Because the modulation is usually a square wave rather than sinusoidal, it is made up of multiple sinusoidal waves with periods L_{mod}/m, all of which can participate in generating a polarization wave. Like traveling waves, the stationary modulation waves can be oriented in any direction, and any Fourier component of the modulation wave with wave vector $\Lambda_m = 2\pi m/L_{\text{mod}}$ can participate in (quasi) phase matching, as shown in Figure 10.

Usually the spatial modulation is achieved by periodically reversing the sign of the nonlinear coupling coefficient. Figure 11 shows the associated phasor diagram for perfect first-order QPM. The net fields from the first two modulation periods are labeled E_1 and E_2. If the sign of the nonlinearity in the first half of the modulation period were not reversed in the second half, the phasors would form a complete circle resulting in a null net field. All following periods would also produce null net fields. However, if the sign is reversed midway through the first period, the phasors from the second half are sign reversed and the net field is E_1 as shown. The magnitude of E_1 is smaller than for a truly phase matched segment by the

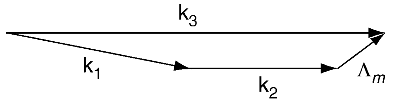

Figure 10 Quasi-phase matching vector diagram for sum frequency mixing. The *m*th Fourier component of the grating vector is indicated by Λm.

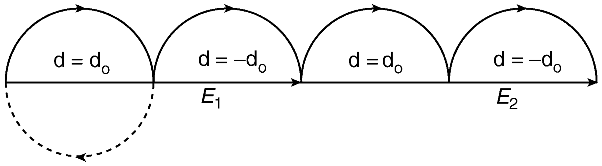

Figure 11 Phasor diagram for a quasi-phase matched interaction in a material with alternating signs of the nonlinear coefficient. The contributions from four segments, each of length $L_{\text{coh}}/2$ are shown. The net field from the first QPM period is labeled E_1, and that from the second period is E_2. The dashed line shows the path the phasors would trace if the nonlinear coefficient were not sign reversed in the even numbered segments.

ratio of the diameter of the circle to its circumference, or by $2/\pi$. Coherence lengths in crystals are typically 5–50 μm so crystals that are several mm long have a large number of periods, in which case the net phasors $E_1, E_2, \ldots$ can themselves be considered elementary phasors that form the net field for the full crystal. For example, if the modulation period L_{mod} is not exactly equal to $2\pi/\Delta k$, QPM is imperfect and phasors $E_1, E_2, \ldots$ curl up to give a net field of $\text{sinc}(\Delta k' L/2)$ where $\Delta k' = \Delta k - 2\pi/L_{\text{mod}}$. For all practical purposes, a QPM crystal containing many modulation periods performs the same as a phase matched crystal with its nonlinear coupling reduced by $2/\pi$.

Referring to eqn [2], we note that periodic modulation of the source term $S(z)$ produces peaks in the phase matching curves at $\Delta k = m2\pi/L_{\text{mod}}$ where m is an integer. If $S(z)$ is a square wave of alternating sign over $0 < z < L$, its Fourier transform is the convolution of $\text{sinc}(\Delta kL/2)$ with a picket fence function with spikes at $\Delta k = m2\pi/L_{\text{mod}}$. If $L_{\text{mod}} \ll L$, the crystal contains many periods of the square wave and the spikes are separated by many times the width of the $\text{sinc}(\Delta kL/2)$ function. This results in well separated sinc functions centered at $\Delta k = 2\pi m/L_{\text{mod}}$, and any process with a coherence length of L_{mod}/m will be quasi-phase matched. That is, one of the Fourier peaks in Δk space will be centered on the corresponding value of Δk.

The magnitudes of the QPM peaks are maximized if the modulation of $S(z)$ is a sign reversal, as described above. If the positive and negative segments have unequal lengths, we can define the duty cycle as $D = l_+/(l_+ + l_-)$. The heights of the pickets in the picket fence are then

$$h_m = \frac{2}{m\pi}\sin(\pi m D) \qquad [18]$$

The largest magnitude peak is for $m = 1$ and $D = 0.5$ in which case $h_1 = 2/\pi$.

Unlike birefringent phase matching, QPM can, in principle, phase match any set of polarizations for the interacting waves. This makes it possible to use a polarization set that maximizes the nonlinear response of the crystal.

Ferroelectric poling

There are various ways to achieve QPM. By far the most popular and versatile is by ferroelectric poling. Ferroelectric crystals have an internal electric field that can be reversed by the application of a sufficiently strong external poling field. Reversing the internal field also reverses the sign of the nonlinearity for some light polarizations but does not alter the refractive indices. Crystals of $LiNbO_3$ and $KTiOPO_4$, with thickness up to 1 mm, can be poled with periods as short as 5 μm. This is sufficient to quasi-phase match frequency mixing of visible or near infrared light. Figure 12 shows an example of the orientation of the internal fields of patterned ferroelectric domains.

It is possible to pole a crystal aperiodically so that multiple mixing processes are quasi-phase matched simultaneously. To design the poling pattern for such a device, the Fourier transform of an initial trial pattern can be calculated and compared with the target spectrum in Δk space. Iterative algorithms can be used to modify the poling pattern, zeroing in on the target spectrum.

Ferroelectric poling can also form two-dimensional poling patterns, corresponding to a two-dimensional set of Fourier components in Δk space. It can also produce chirped gratings which have a broad phase matching peak. Chirped gratings can be used to simultaneously frequency convert and tailor the temporal shapes of the output pulse when the input pulses are frequency chirped. Poling can also yield fanned poling patterns so the poling period and thus the phase matching wavelength can be shifted by translating the grating.

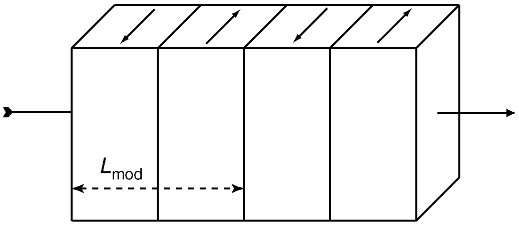

Figure 12 Diagram of a poled ferroelectric crystal. The four crystal segments represent four ferroelectric domains with internal crystal field directions indicated by the arrows. Reversal of the internal field is accompanied by reversal of the nonlinear coefficient, so this is a quasi-phase matched structure with alternating signs of d_{eff}.

Wafer stacking and patterned crystal growth

Another method of quasi-phase matching relies on stacks of thin plates oriented so the sign of the nonlinearity reverses in adjacent plates. Each plate has a thickness of $mL_{\mathrm{coh}}/2$ where m is an odd integer. If all the plates are identically prepared crystals, it is often possible to reverse the sign of their nonlinearity by reorienting them. It is sometimes possible to optically contact the plates or even to diffusion bond them, avoiding reflective losses at the interfaces. This technique is conducive to large aperture QPM crystals. However, the plates are thin so fabrication is a challenge.

Quasi-phase matched crystals of GaAs have also been epitaxially grown on patterned substrates. The patterning serves as a template for growth of zones with alternating crystal orientation and sign of the nonlinear coefficient.

Total internal reflection

Total internal reflection in a thin slab has also been exploited for QPM (Figure 13). Two effects contribute to the source modulation, one is the rephasing of the waves on total internal reflection; the other is the change in nonlinearity that may accompany the change in propagation direction. Usually plates much thicker than one coherence length are used, so the quasi-phase matching is of high order ($m \gg 1$).

Wave Guide Modal Phase Matching

In waveguides the size and shape of the guide influence the transverse field distributions and the propagation constants for the transverse modes. Generally the higher order transverse modes have higher phase velocities than the lower order modes. This can be used to counteract index dispersion in three wave mixing if the bluest wave propagates in a higher order mode than the redder waves. A problem frequently encountered in using this effect is that the transverse modes are orthogonal, or nearly so. Because the nonlinear response is proportional to the overlap integral

$$\mathrm{d}_{\mathrm{eff}} \propto \int\int_A \mathrm{d}(x,y)E_1(x,y)E_2(x,y)E_3(x,y)\mathrm{d}x\,\mathrm{d}y \qquad [19]$$

where the E's are the transverse field distributions of the three waves, the nonlinear response is often quite small. If the local nonlinearity $\mathrm{d}(x,y)$ can be engineered to maximize the integral, this problem can be alleviated. This may be thought of as transverse quasi-phase matching.

Waveguides can also be designed for Cerenkov phase matching where the phase velocity of the radiated wave is less than that of the polarization wave. Figure 14 shows an example. The emission angle is

$$\cos\theta = \frac{2k_\omega}{k_{2\omega}} = \frac{v_{2\omega}}{v_\omega} \qquad [20]$$

where the v's are phase velocities. The transverse momentum Λ must be supplied by the waveguide. In other words, the source harmonic polarization must have a transverse Fourier component with spatial frequency Λ. This is another example of transverse quasi-phase matching.

Another application of waveguide phase matching is in traveling wave electro-optic modulation, where the rf wave travels along an rf waveguide parallel to the light path. For maximum modulation bandwidth, the phase velocity of the rf wave should equal that of the light wave over the desired modulation bandwidth. Because the phase velocity of the rf wave in the bulk medium is usually faster than that of the light wave, the rf guide must be designed for a slow phase velocity.

Compensated Phase Matching (CPM)

Compensated phase matching is somewhat like QPM in that multiple crystal segments are used, but it differs by requiring that sequential crystal segments have ΔkL's of equal magnitude but opposite sign. Ideally the nonlinearity would be the same in all segments. Figure 15 shows the phasor diagram for one such arrangement, in this case with unequal nonlinearity in alternating segments. CPM is realized

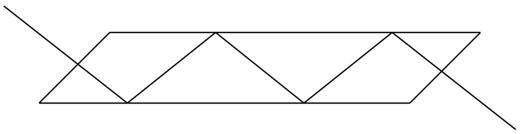

Figure 13 Diagram of quasi-phase matching based on total internal reflection. Relative phases acquired by the waves upon total internal reflection, combined with the path lengths and possible changes in the value of d_{eff} for alternating legs, can be combined to achieve quasi-phase matching.

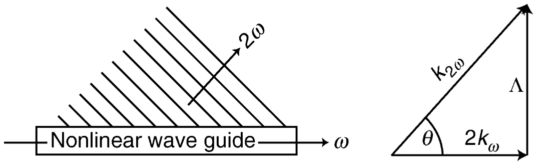

Figure 14 Cerenkov phase matched second harmonic generation. The phase velocity of the fundamental wave in the waveguide is greater than the phase velocity of the harmonic wave in the cladding.

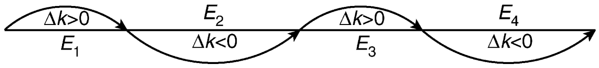

Figure 15 Phasor diagram for four segments of a device with compensated phase matching device. The magnitude of ΔkL is equal in each segment but the sign of Δk alternates. The length of the E_n field from segment n is proportional to $\mathrm{sinc}(\Delta kL/2)$ so the net field ΣE_n must also be proportional to this function.

in walkoff compensated geometries where a sequence of crystals is arranged so $\Delta k = 0$ in each for a particular monochromatic plane wave. In the case of spatial walkoff compensation the walkoff direction in alternating crystals is reversed so the walkoff is exactly compensated in pairs of adjacent crystals. Tilting the input beams by a small amount changes the values of the Δk's in such a way that CPM is achieved. In the case of temporal walkoff compensation, temporal walkoff is reversed in alternating crystals, in which case tuning the input frequency leads to CPM. The advantage of CPM is that the acceptance bands, which will be discussed in the next section, are those of the individual crystal segments which are larger than those of a single crystal of equivalent length without walkoff compensation.

Spatial and Temporal Effects

Acceptance Bands

Our discussion to this point has been based on monochromatic plane waves. Real beams have spatial and temporal structure which means they can be decomposed into sets of monochromatic plane waves with tilted propagation directions relative to the nominal direction, and with detuned frequencies relative to the nominal frequency. As we showed above for monochromatic plane waves, there is a range of angles and frequencies, called the acceptance bands, over which $\Delta kL < 2\pi$. If the interacting beams' angular spectra and frequency spectra lie within these bands, the process is well phase matched for all angular and frequency components, and efficiency is maximized. This statement is equivalent to saying that if the spatial walkoff due to beam tilts and birefringence, and the temporal walkoff due to differing group velocities are small enough that the interacting beams stay well overlapped in space and time, the efficiency of energy transfer to the output channel does not suffer. Consider, for example, second harmonic generation of short pulses. The frequency acceptance bandwidth for the fundamental wave, $\Delta\omega_1$, is found using

$$\frac{\mathrm{d}(\Delta k)}{\mathrm{d}\omega_1}\Delta\omega_1 = \left(\frac{2\mathrm{d}k_2}{\mathrm{d}\omega_2} - \frac{2\mathrm{d}k_1}{\mathrm{d}\omega_1}\right)\Delta\omega_1 = \frac{2\pi}{L} \qquad [21]$$

Using the usual definition of the group velocity:

$$v_\mathrm{g} = \frac{\mathrm{d}\omega}{\mathrm{d}k} \qquad [22]$$

this can be written

$$\Delta\omega_1 = \frac{\pi}{L(1/v_{\mathrm{g}2} - 1/v_{\mathrm{g}1})} = \frac{\pi}{\tau_2 - \tau_1} = \frac{\pi}{\Delta\tau} \qquad [23]$$

where τ_1 and τ_2 are the passage times through a crystal of length L for the fundamental and harmonic pulses, and $\Delta\tau = (\tau_2 - \tau_1)$ is the temporal walkoff between the two. This shows that the acceptance bandwidth is approximately equal to the transform bandwidth of a pulse whose duration is equal to the temporal walkoff $\Delta\tau$. Input pulses shorter than $\Delta\tau$ will not convert efficiently because the fundamental and harmonic pulses do not remain overlapped as they propagate through the crystal. In the frequency domain it is said that pulses shorter than the $\Delta\tau$ have spectral widths larger than the acceptance bandwidth, meaning they are not well phase matched across their full spectra, and hence do not convert efficiently. Parallel arguments apply to spatial walkoff and angular acceptance bandwidths. Beams that do not remain spatially overlapped through the crystal do not convert efficiently. Equivalently, if the angular spectra of the beams exceed the acceptance angle, efficiency suffers.

Sometimes these notions are described in terms of an effective interaction length or an effective crystal length. For short pulses the effective interaction length is that for which temporal walkoff is equal to the pulse duration. Equivalently it is the length for which the acceptance bandwidth is equal to the spectral width of the pulse. Crystals shorter than this will have mixing efficiencies nearly equal to that expected for a phase matched process, while longer crystals have efficiencies that fall short of this expectation.

Focused beams

Focused Gaussian beams are a particularly important class of spatially structured beams. In passing through a focus, a lowest order Gaussian beam acquires a phase shift relative to an unfocused beam. The phase velocity near the focus is higher for the focused beam, leading to a phase shift of π through the focus. For second harmonic generation the phase shift of the second harmonic polarization is twice this, or 2π.

A second harmonic wave also acquires a phase shift of π, but this leaves a π phase shift between the harmonic polarization source wave and a freely propagating harmonic wave. The phasors representing harmonic radiation from symmetrically located points well before and well after the focus will therefore be 180° out of phase. Point pairs closer to the focus will have less severe cancelation, but it is not surprising that the maximum efficiency is achieved, not at $\Delta k = 0$, but for a slightly negative value Δk. The phase matching function sinc($\Delta kL/2$) is shifted and distorted for focused beams. Its exact shape depends on the nature of the mixing process. For example, the phase matching peak is different for third harmonic generation and second harmonic generation, and it is different for mixing with small birefringence and mixing with large birefringence.

Phase Conjugation

Degenerate four-wave mixing

Some wave mixing processes are automatically phase matched. One example is phase conjugation by degenerate four wave mixing. This process creates a polarization wave that is the time reversed image of one of the input waves. Figure 16 compares the reflection from a conventional mirror with that from a phase conjugate mirror.

The mixing process shown in Figure 17 is one that can act as a phase conjugate mirror. Waves *f* and *b* are strong pump waves that are phase conjugates of one another. In their presence a weak signal field at the same frequency produces a polarization given by

$$P = sE_fE_bE_s^* \, e^{-i\omega t + i(\mathbf{k}_f - \mathbf{k}_b - \mathbf{k}_s)\cdot\mathbf{r}} \qquad [24]$$

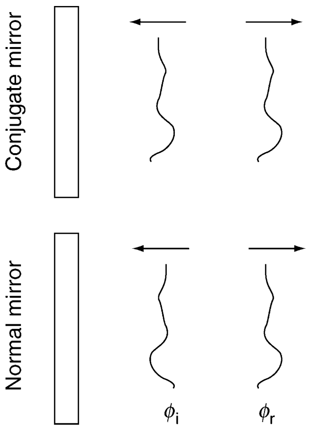

Figure 16 Comparison of a normal and a phase conjugate mirror. The wavy lines represent the phases of the incident and reflected waves.

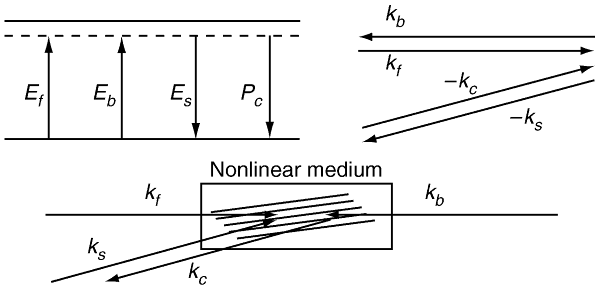

Figure 17 Phase conjugation via degenerate four wave mixing near a one photon resonance. Two counter propagating pump waves with fields E_f and E_b interact with a signal beam, E_s, to produce a polarization wave P_c where the frequencies obey $\omega_f - \omega_s + \omega_b = \omega_c$ (upper left), implying the *k*-vector of the polarization wave is given by $k_c = k_f - k_s + k_b$ (upper right). The lower diagram shows the grating induced by waves *f* and *s* that scatters wave *b* into the direction of *c*.

The two pump propagation vectors cancel one another leaving

$$P = sE_fE_b(E_s^* \, e^{-i\omega t - i\mathbf{k}_s\cdot\mathbf{r}}) \qquad [25]$$

The quantity in parentheses is the phase conjugate of the signal wave, so the polarization wave exactly phase matches a radiated conjugate wave.

Brillouin scattering

Other possibilities for automatic phase matching are processes that involve a stimulated wave in the nonlinear medium such as an acoustic wave (Brillouin scattering) or waves of coherent molecular vibration or rotation (Raman scattering). Phase conjugation, based on Brillouin scattering, is widely used to reduce phase distortions acquired on passing through an amplifier, for example. A single beam of light is focused into a material that has a high polarizability. Example media include high pressure gases Xe and SF_6, liquids acetone and freon, and fused silica. The conjugation process is not as simple to describe as it was in the case of four wave mixing because many acoustic waves can be generated. The key is that the phase conjugate wave has the highest gain under certain conditions and can dominate the Brillouin scattering process, so the reflected wave is nearly a phase conjugate of the incoming wave.

See also

Modulators: Acousto-Optics; Electro-Optics. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing; Nonlinear Optical Phase Conjugation; $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources**: Harmonic Generation in Gases. **Optical Parametric Devices:** Overview. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed).

Phase Control: Phase Conjugation and Image Correction. **Photonic Crystals:** Nonlinear Optics in Photonic Crystal Fibers. **Scattering**: Raman Scattering.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.
Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.
Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1999) *Handbook of Nonlinear Optical Crystals*. Berlin: Springer-Verlag.
Hanna DC, Yuratich MA and Cotter D (1979) *Nonlinear Optics of Free Atoms and Molecules*. Berlin: Springer-Verlag.
Mills DL (1991) *Nonlinear Optics*. Berlin: Springer-Verlag.
Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.
Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.
Yariv A and Yeh P (1984) *Optical Waves in Crystals*. New York: John Wiley & Sons.
Zel'dovich BY, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.
Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Pulse Compression via Nonlinear Optics

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort pulses have a great potential in several areas of research and application, such as measurements of ultrafast physical processes, infrared time-resolved spectroscopy, and sampling systems, but they are also invaluable for communications. With increasing demands being put on the performance of optical fiber communications, the emphasis now is on producing a source of ultrashort optical pulses that can be used in high-bit rate and long-distance optical communication systems.

Several optical pulse compression techniques are described in the literature. These techniques utilize dispersion, possibly from gratings or prisms, simultaneously to compensate for chirp and achieve pulse compression. In fact, when a chirped optical pulse propagates in a linear dispersive medium, it acquires a dispersion-induced chirp. If the initial chirp is in the opposite direction to that imposed by the group-velocity dispersion (GVD), the two tend to cancel each other, resulting in the compression of the optical pulse.

Early work on optical pulse compression did not make use of any nonlinear optical effects. Only during the 1980s, when the use of single-mode silica fibers as nonlinear media became widespread, the self-phase modulation (SPM) effect was used to achieve pulse compression. It has led, in 1987, to the production of optical pulses as short as 6 fs in the 620 nm wavelength region. Such advances were possible, only after having understood the evolution of optical pulses in silica fibers.

We can consider two main categories of pulse compressors which make use of nonlinear effects in optical fibers: grating-fiber and soliton-effect compressors. In a grating-fiber compressor, the input pulse is propagated in the normal-dispersion regime of the fiber, which imposes a nearly linear, positive chirp on the pulse through a combination of SPM and GVD. The output pulse is then sent through a grating pair where it experiences anomalous GVD and gets compressed. In the case of a soliton-effect compressor, the fiber itself acts as a compressor without the need of an external grating pair. The input pulse propagates in the anomalous-GVD regime of the fiber and is compressed through an interplay between SPM and GVD. This compression mechanism is related to a fundamental property of the higher-order solitons. These solitons follow a periodic evolution pattern such that they go through an initial narrowing phase at the beginning of each period. If the fiber length is suitably chosen, the input pulses can be compressed by a factor that depends on the soliton order.

Grating-Fiber Compressors

The grating-fiber compressor is generally used at wavelengths $\lambda < 1.3\ \mu m$. At these wavelengths, optical pulses acquire a positive chirp across its entire width during their passage through conventional silica fibers. A schematic of a grating-fiber compressor in the double pass configuration is shown in Figure 1. After the propagation in the fiber, the optical pulse is sent through a grating pair, which provides the anomalous (or negative) GVD required to get the pulse compression.

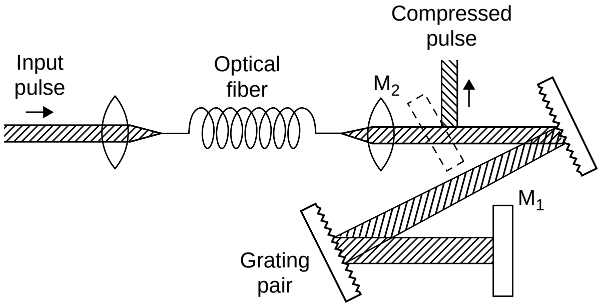

Figure 1 Schematic representation of a grating-fiber compressor in the double-pass configuration. (Reproduced with permission from Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.)

Different frequency components of an optical pulse incident at one grating are diffracted at slightly different angles. As a consequence, they experience different time delays during their passage through the grating pair, the blue-shifted components arriving earlier than the red-shifted ones. In the case of an optical pulse with a positive chirp, the blue-shifted components occur near the trailing edge of the pulse and the passage through the grating pair provides its compression.

A limitation of the grating pair is that the spectral components of the pulse are dispersed not only temporally but also spatially. As a consequence, the optical beam becomes deformed, which is undesirable. This problem can be avoided simply by using a mirror to reflect the pulse back through the grating pair. Mirror M_1, in Figure 1, performs this function. Reversing the direction of propagation not only allows the beam to recover its original cross-section, but also doubles the amount of GVD, thereby reducing the grating separation by a factor of 2. The mirror M_2 in Figure 1 is used to deflect the compressed pulse out of the compressor.

The phase shift acquired by the spectral component of the pulse at the frequency ω passing through the grating pair can be written in the form:

$$\phi_g = \phi_0 + \phi_1(\omega - \omega_0) + \frac{1}{2}\phi_2(\omega - \omega_0)^2 + \frac{1}{6}\phi_3(\omega - \omega_0)^3 + \cdots \quad [1]$$

where ω_0 is the pulse center frequency, ϕ_0 is a constant, ϕ_1 is a constant delay, and ϕ_2 and ϕ_3 are parameters that take into account the GVD effects associated with the grating pairs. These parameters depend on the grating period (line spacing), as well as on the orientation and separation of the two gratings. In most cases of practical interest, the spectral width of the pulse satisfies the condition $\Delta\omega \ll \omega_0$ and the cubic and higher-order terms in eqn [1] can be neglected.

To achieve optimum performance from a grating-fiber compressor, it is necessary to optimize the fiber length, as well as the grating separation. Concerning the first aspect, the effects of both GVD and SPM during the propagation of the pulse inside the fiber must be considered. SPM alone determines a linear chirp only over the central part of an optical pulse. Since the grating pair compresses only this region, while a significant amount of energy remains in the wings, the compressed pulse is not of high quality in this case. The effect of GVD turns out to be important, since it broadens and reshapes the pulse, which develops a nearly linear chirp across its entire width. In these circumstances, the grating pair can compress most of the pulse energy into a narrow pulse.

The balanced effects of both GVD and SPM explain the existence of an optimum fiber length, z_{opt}, for maximum pulse compression. In fact, when the fiber length is less than z_{opt}, the SPM-induced chirp has not yet been linearized. On the other hand, when it is longer than z_{opt}, the SPM effects become negligible due to the GVD-induced pulse broadening.

If the input pulse is unchirped and presents a hyperbolic secant shape, the optimum fiber length for maximum pulse compression is well approximated by

$$z_{\mathrm{opt}} \approx (6L_{\mathrm{D}}L_{\mathrm{NL}})^{1/2} \quad [2]$$

where

$$L_{\mathrm{D}} = \frac{T_0^2}{|\beta_2|} \quad [3]$$

and

$$L_{\mathrm{NL}} = \frac{1}{\gamma P_0} \quad [4]$$

are length scales associated with the GVD and the SMP effects, respectively. In eqns [3] and [4], T_0 is the input pulsewidth, β_2 is the fiber GVD parameter, γ is the nonlinearity parameter of the fiber, and P_0 is the peak power of the input pulse.

If the input pulse presents a negative chirp, this must be compensated for by the positive chirp provided by the fiber. As a result, the optimum fiber length increases. The opposite occurs in the case of an input pulse with a positive chirp.

The optimum compression factor, F_{opt}, in the case of a grating-pair compressor, is given by

$$F_{\mathrm{opt}} = \frac{T_{\mathrm{FWHM}}}{T_{\mathrm{comp}}} \approx 0.625\left(\frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}}\right)^{1/2} \quad [5]$$

where T_{FWHM} is the full width at half maximum (FWHM) of the input pulse ($T_{\mathrm{FWHM}} \approx 1.76T_0$ for a hyperbolic secant pulse) and T_{comp} is the FWHM of the compressed pulse. Equations [2] and [5] provide a good estimate even for pulse shapes other than a hyperbolic secant as long as $\frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}} > 100$.

The previous results were obtained neglecting the higher-order nonlinear and dispersive effects, which is acceptable for pulsewidths $T_0 > 0.1$ ps. However, when shorter pulses are used at the input, their spectral width $\Delta\omega$ is large enough that the cubic term in the expansion [1] must be taken into account. In this case, the compression factor turns out to be smaller than that given by eqn [5].

The third-order dispersion (TOD) of the grating pair, resulting from cubic term in eqn [1], has been compensated for in one experiment using a combination of gratings and prisms. Performing such compensation, a reduction of the pulsewidth to about 6 fs has been achieved. Such a pulse at 620 nm consists of only three optical cycles.

Even though the compression factor $F_{\mathrm{c}} \propto \sqrt{P_0}$, the peak power of the input pulse must be kept below the Raman threshold to avoid the loss of pulse energy through the stimulated Raman scattering (SRS) process. If this condition is not then satisfied, besides the energy loss problem, the Raman pulse can interact with the input pulse through cross-phase modulation and deform the linear nature of the frequency chirp. As a result, the compression factor is reduced from the predictions of eqn [5].

Fiber Bragg Gratings

A uniform fiber Bragg grating (FBG) reflects light whose wavelength is within the stop-band centered at the Bragg wavelength. The GVD is anomalous on the high-frequency side of the stop-band and becomes normal on the low-frequency side. Near the stop-band edges, the grating exhibits large GVD. Since a FBG acts as a dispersive delay line, it can be used in place of the bulk-grating pair in a grating-fiber compressor, providing a compact all-fiber device. In fact, a typical 1 cm long fiber grating may provide as much dispersion as a bulk-grating pair with more than one meter spacing. However, the grating induced TOD can affect significantly the quality of the pulses when the optical frequency falls close to the edges of the stop-band.

The compression factor as well as the pulse quality can be significantly improved by using a FBG in which the grating optical period changes along its length – the so-called chirped fiber grating. In this case, different frequency components of the pulse are reflected from different regions of the grating. Such a device has been used since the mid-1990s to compensate for dispersion-induced broadening of pulses in fiber-optic communication systems, and compression factors above 100 have been achieved. In practice, the only disadvantage of a chirped fiber grating is that the compressed pulse is reflected rather than transmitted.

Soliton-Effect Compressors

Optical pulses at wavelengths exceeding 1.3 μm generally experience both SPM and anomalous GVD during their propagation in silica fibers. Such a fiber can act as a compressor by itself without the need of an external grating pair. In the case of a fundamental or first-order soliton, the effects of SPM cancel the effects of anomalous GVD perfectly, and the soliton propagates whilst preserving its hyperbolic secant shape in a lossless optical fiber. A higher-order soliton, on the other hand, changes its shape periodically as it propagates in a fiber. Such a soliton always experiences an initial pulse narrowing phase before recovering its original hyperbolic secant profile at integral multiples of the soliton period z_0, and this behavior is exploited to achieve pulse compression. Optimum compression is achieved when the higher-order soliton leaves the fiber just as it attains its narrowest width.

The evolution of a soliton of order N inside an optical fiber is governed by the nonlinear Schrödinger equation (NLSE), which can be written in the following form:

$$\mathrm{i}\frac{\partial q}{\partial \xi} + \frac{1}{2}\frac{\partial^2 q}{\partial \tau^2} + N^2|q|^2 q = 0 \qquad [6]$$

where q is the normalized amplitude, $\xi = z/L_{\mathrm{D}}$, $\tau = T/T_0$, and the parameter N is the soliton order, given by

$$N^2 = \frac{L_{\mathrm{D}}}{L_{\mathrm{NL}}} = \frac{\gamma P_0 T_0^2}{|\beta_2|} \qquad [7]$$

The soliton period, z_0, is given by

$$z_0 = \frac{\pi}{2}L_{\mathrm{D}} = \frac{\pi T_0^2}{2|\beta_2|} \qquad [8]$$

In writing eqn [6], the fiber losses were neglected, since fiber lengths employed in practice are relatively small. Even though higher-order solitons follow an exact periodic evolution only for integer values of N, eqn [6] can be used to describe pulse evolution for arbitrary values of N.

In practice, soliton effect compression is carried out by initially amplifying optical pulses up to the power level required for the formation of higher-order solitons. The peak optical power of the initial pulse required for the formation of the *N*th-order soliton is given by

$$P_\mathrm{N} = 3.11\frac{|\beta_2|N^2}{\gamma T_\mathrm{FWHM}^2} \qquad [9]$$

These *N*th-order solitons are then passed through the appropriate length of optical fiber to achieve a highly compressed pulse. The optimum fiber length, z_opt, and the optimum pulse compression factor, F_opt, of a soliton-effect compressor can be estimated from the following empirical relations:

$$z_\mathrm{opt} \approx z_0\left(\frac{0.32}{N} + \frac{1.1}{N^2}\right) \qquad [10]$$

$$F_\mathrm{opt} \approx 4.1N \qquad [11]$$

These relations are accurate to within a few percent for $N > 10$.

Compression factors as large as 500 have been attained using soliton-effect compressors. In fact, considering the same input pulse, the soliton-effect compressor provides a pulse compression that is larger with a fiber that is shorter than a grating-fiber compressor. However, the pulse quality is poorer, since the compressed pulse carries only a fraction of the input energy, while the remaining energy appears in the form of a broad pedestal. From a practical point of view, the pedestal is deleterious since it causes the compressed pulse to be unstable, making it unsuitable for some applications. Despite this, soliton-effect compressed pulses can still be useful because there are some techniques that can eliminate the pedestal.

One difficulty faced, when using the soliton-effect compressor, is that pulses with high peak power are required for the formation of high-order solitons in conventional fibers. As suggested by eqn [9], the use of dispersion-shifted fibers (DSFs) with small values of β_2 at the operating wavelength can reduce the peak power required for soliton generation by an order of magnitude. However, because the soliton period z_0 is inversely proportional to $|\beta_2|$ (see eqn [8]), eqn [10], indicates that longer lengths of DSFs will be required for solitons to achieve optimum compression. As a result, the total fiber loss experienced by those solitons will be larger and the loss-induced pulse broadening will have a significant impact on the compressor global performance.

Another problem introduced by the use of DSFs in soliton-effect compressors is the TOD, which generally degrades the quality of the compressed pulse. In the case of soliton-effect compressors using conventional optical fibers, the effects of TOD become significant only when the widths of the compressed pulse become very short ($\ll$1 ps). In DSFs, however, the relative importance of TOD is increased and those effects become more pronounced because the GVD parameter, β_2, is small. In this case, TOD is detrimental in the compression, even of pulses with widths of a few picoseconds, resulting in serious degradation of the optimum compression factor F_opt.

The result given by eqn [11] for the optimum compression factor holds only for the case of an ideal soliton-effect compressor, when high-order nonlinear and dispersive effects, such as intrapulse Raman scattering (IRS), TOD and self-steepening, are neglected. However, these effects cannot be neglected for the highly compressed pulses because of their subpicosecond widths and high intensities.

For not-too-short optical pulses (widths >50 fs) propagating not too close to the zero-dispersion wavelength of the fiber, the dominant higher-order effect is IRS, which manifests as a shift of the pulse spectrum toward the red side – the so-called soliton self-frequency shift. As a consequence, the group velocity is reduced and there is a delay of the optical pulse. This delay affects substantially the interplay between GVD and SPM that is responsible for pulse compression.

Interestingly, the IRS effect can be used to improve the quality of the compressed pulse, by removing the pedestal mentioned above. In fact, as a consequence of the change in the group velocity induced by the IRS effect, the sharp narrow spike corresponding to the compressed pulse travels more slowly than the pedestal and separates from it. Moreover, the pedestal can be removed by spectral filtering and a red-shifted, pedestal-free, and highly compressed pulse is then produced.

Compression of Fundamental Solitons

As described in the last section, the propagation of higher-order solitons provide a rapid compression method, but it suffers from the existence of residual pedestals. Some techniques can help to reduce or even eliminate those pedestals, but energy is always wasted in this process.

A less rapid technique that provides better pulse quality is adiabatic amplification of fundamental or first-order solitons, which are of primary importance in the domain of optical communications. Solitons have fixed area, so the increased energy from amplification is accommodated by an increase in power and a decrease in width. To avoid distortion, the amplification per soliton period cannot be too great.

The same effect as adiabatic amplification can be achieved using an optical fiber with dispersion that decreases along the length of the fiber. In fact, for optical solitons, a small variation in the dispersion has a similar perturbative effect as an amplification or loss: such a variation perturbs the equilibrium between the dispersion and nonlinearity in such a way that when, for example, the dispersion decreases, the soliton pulse is compressed. It can be seen from eqn [7] that if the value of $|\beta_2|$ decreases along the fiber and to keep the soliton order N ($N = 1$ for the fundamental soliton), the pulsewidth must indeed decrease as $|\beta_2|^{1/2}$. Hence, the use of fibers with variable dispersion is viewed as a passive and effective method to control optical solitons in soliton communication systems. Dispersion-decreasing fibers (DDFs), in particular, have been recognized to be very useful for high-quality, stable, polarization-insensitive, adiabatic soliton pulse compression and soliton train generation. These fibers can be made by tapering the core diameter of a single-mode fiber during the drawing process, and hence changing the waveguide contribution to the second-order dispersion.

In a DDF of certain length, the dispersion is monotonically and smoothly decreased from an initial value to a smaller value at the end of the length according to some specified profile. Provided the dispersion variation in the DDF is sufficiently gradual, soliton compression can be an adiabatic process where an input fundamental soliton pulse can be ideally compressed as it propagates, while retaining its soliton character and conserving the energy.

The evolution of the fundamental soliton in a DDF can be described by the following NLSE:

$$\mathrm{i}\frac{\partial q}{\partial \xi} + \frac{1}{2}p(\xi)\frac{\partial^2 q}{\partial \tau^2} + |q|^2 q = 0 \qquad [12]$$

where the variable coefficient $p(\xi) = |\beta_2(\xi)/\beta_2(0)|$ takes into account the variation of dispersion along the fiber. Using the transformations $s = \int_0^{\xi} p(y)\mathrm{d}y$ and $U = q/\sqrt{p}$, eqn [12] assumes the form:

$$\mathrm{i}\frac{\partial U}{\partial s} + \frac{1}{2}\frac{\partial^2 U}{\partial \tau^2} + |U|^2 U = \mathrm{i}GU \qquad [13]$$

where

$$G = -\frac{1}{2p}\frac{\mathrm{d}p}{\mathrm{d}s} \qquad [14]$$

Equation [13] shows clearly that the effect of decreasing dispersion is mathematically equivalent to the effect of an optical amplifier, adding a gain term to the NLSE. The effective gain coefficient G is related to the rate at which GVD decreases along the fiber.

Since decreasing dispersion is equivalent to an effective gain, a DDF can be used in place of a conventional fiber amplifier to generate a train of ultrashort pulses. To achieve such an objective, a continuous wave (CW) beam with a weak sinusoidal modulation imposed on it is injected into the DDF. The sinusoidal modulation can be imposed, for example, by beating two optical signals. As a result of the combined effect of GVD, SPM, and decreasing GVD, the nearly CW beam is converted into a high-quality train of ultrashort solitons, whose repetition rate is governed by the frequency of the initial sinusoidal modulation.

Different approaches have been developed to determine the optimum GVD profile and its dependence on the width and peak power of input pulses. In the case of picosecond soliton pulse compression, direct numerical simulations of the NLSE show that linear, Gaussian, and exponential dispersion profiles may all be used effectively to provide ideal, adiabatic compression, where the input pulse energy is conserved and remains localized within the pulse. In this case, the final pulse compression factor at the end of a certain length L of DDF is equal to the ratio of the initial to the final second-order dispersions:

$$F_{\mathrm{c}} = \frac{T_{\mathrm{FWHM}}}{T_{\mathrm{comp}}} = \frac{\beta_2(0)}{\beta_2(L)} \qquad [15]$$

Compression factors larger than 50 are possible launching input pulses with peak powers corresponding to the fundamental soliton into a DDF whose length is about one soliton period. This technique takes advantage of soliton-effect compression but requires lower peak powers and produces compressed pulses of better quality.

In the case of subpicosecond soliton pulses, the influence of higher-order nonlinear and dispersive effects must be accounted for. Some studies show that, in this case, the linear and Gaussian dispersion profiles are the most suitable to achieve high-quality, pedestal-free, adiabatic compression of fundamental solitons. However, in the presence of higher-order effects, the final compression factor is generally lower, and after reaching a maximum at a particular length of DDF, it decreases steadily. This can be explained in terms of pulse compression stabilization, which originates from a competition between the rate of dispersion decrease in the DDF and the rate of dispersion increase due to the combined effects of soliton self-frequency shift and third-order dispersion.

List of Units and Nomenclature

F_c	compression factor
F_{opt}	optimum compression factor
G	amplifier gain
L	fiber length (m)
L_D	dispersion length (m)
L_{NL}	nonlinear length (m)
N	soliton order
P_N	peak power of the Nth-order soliton (W)
P_0	pulse peak power (W)
q	normalized pulse amplitude
T_0	input pulse (s)
T_{comp}	compressed pulse FWHM (s)
T_{FWHM}	pulse full width at half maximum (s)
z_0	soliton period (m)
z_{opt}	optimum fiber length (m)
β_2	group-velocity dispersion coefficient ($s^2\ m^{-1}$)
γ	fiber nonlinearity coefficient ($W^{-1}\ m^{-1}$)
ξ	normalized distance
τ	normalized time
ϕ_g	phase shift induced by the grating pair (rad)
ω_0	pulse center frequency (rad s^{-1})
ω	optical frequency (rad s^{-1})
$\Delta\omega$	pulse spectral width (rad s^{-1})

See also

Chirped Pulse Amplification. **Fiber Gratings**. **Scattering:** Raman Scattering. **Solitons:** Soliton Communication Systems. **Ultrafast Laser Techniques:** Pulse Characterization Techniques.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agrawal GP (1997) *Fiber-Optic Communication Systems*, 2nd edn. New York: John Wiley & Sons.

Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Barty C, White W, Sibbett W and Trebino R (eds) (1998) Special issue on ultrafast optics. *IEEE Journal of Selected Topics on Quantum Electronics* 4: 157.

Buck JA (1995) *Fundamentals of Optical Fibers*. New York: John Wiley & Sons.

Kashyap R (1999) *Fiber Bragg Gratings*. San Diego, CA: Academic Press.

Mamyshev PV (1992) Generation and compression of femtosecond solitons in optical fibers. In: Taylor JR (ed.) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.

Raman Lasers

M Santagiustina, Istituto Nazionale di Fisica della Materia, Università degli Studi di Padova, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

The Raman effect bears the name of the Indian scientist Chandrasekhara V. Raman (1888–1970) who received the Nobel Prize in the year 1930 for the discovery, published in 1928, of this scattering phenomenon.

The Raman effect is an inelastic light scattering process, occurring in gaseous, liquid, and solid media. The frequency of the scattered photon can be either down-shifted or up-shifted and the result of the process is the generation of new radiation. Frequency down-shifted and up-shifted waves are respectively called Stokes and anti-Stokes waves. The energy difference between input and output photons is either absorbed by, or released from, the atoms or molecules of the medium, which are left in different energy states; jumps can be between electronic states, but more often they are among vibrational or rotational states. The upper energy state involved in the scattering process can be either a true energy state, and this case is referred to as resonant Raman scattering, or a virtual one. The off-resonance effect has a lower probability to occur and high-power beams are needed to observe it. The emission of anti-Stokes waves requires either an inverted Raman medium or particular four-wave-mixing phase-matching conditions to be satisfied in order to occur. The second condition is referred to as coherent anti-Stokes Raman scattering. Here we will consider mainly Stokes wave generation, which is the most common in Raman lasers.

When pumping is very hard, or a seed wave is provided, stimulated Raman scattering can also occur and a high percentage of the input pump radiation can be converted into Stokes, or anti-Stokes, radiation. The stimulated process efficiency, defined as the ratio between the output Raman wave energy and the input pump wave energy, depends on the imaginary part of the third-order nonlinear susceptibility coefficient and on the pump intensity. Stimulated Raman scattering observation dates back to 1962 during the pioneering studies with lasers sources by Woodbury and Ng; their experiment with a ruby laser was actually the first example of an intracavity Raman laser.

Since its discovery, Raman scattering has become a valuable tool for spectroscopy and a practical technique to extend laser light emission to wavelengths not directly generated by laser active materials. This fundamental property stems from the fact that Raman scattering does not require an upper energy state to occur. Hence, the frequency of the emitted radiation can be any, if a suitable incident radiation wavelength is provided. New frequencies are constrained only by the available powerful pump laser sources and by Raman medium shifts.

The frequency shifts provided by some of the materials that are most often used in Raman lasers are presented in Table 1. Finally, it is important to note that Raman scattering can also occur for the newly generated radiation through an identical process in which a Stokes photon is Raman scattered. This radiation, which is generated if the Stokes radiation intensity grows large enough, is referred to as a second-order Stokes wave. For very large pump intensity, a cascading process to several Stokes orders can take place, thus extending the range of generated frequencies to the far-infrared portion of the spectrum.

We will describe the design of Raman lasers that are most commonly used, as well as their performances, also taking into consideration the Raman medium that sets the main constraint to the laser characteristics. Continuous wave as well as pulse Raman lasers will be described and the tuning capability will be discussed.

Raman Shifters

The simplest conversion scheme is obtained when a Raman medium is placed at the output end of a laser. This single-pass setup, shown in Figure 1, provides a fairly efficient frequency shifter.

Single-pass amplification of the spontaneous scattering or of an injected seed copropagating with an intense pump is usually modeled by a couple of equations whose derivation, under a plane, continuous wave approximation, can be found in the

Table 1 Measured Raman frequency shifts and steady-state gain for some of the most important Raman media

Medium	*Frequency shift [THz]*	*Gain [cm/GW]*
Calcite ($CaCO_3$)	32.58(1)	5.5(1)
Lithium iodate ($LiIO_3$)	23.1–24.66(2)	4.8(3)
Barium nitrate ($Ba(NO_3)_2$)	31.46(4)	11(5)
Potassium gadolinium tungstate ($KGd(WO_4)_2$)	27.03 or 23(6)	6(6)
Silica fiber (SiO_2)	13.2(7)	0.01(7)
Hydrogen gas (H_2) ($p = 10$ atm, $T = 293$ K)	124.65(8)	1.53(9)
Nitrobenzene	40.35(10)	3(10)

(1) Pump wavelength 694 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.

(2) Values refer to a propagation direction respectively perpendicular or orthogonal to the crystal axis; from Guadalberto GM and Arguello CA (1974) Raman spectra from oblique phonons in powdered samples. *Solid State Communications* 14: 911–914.

(3) Pump wavelength 1064 nm; from Falk J and Moshrefzadeh (1985) Oblique Raman and polariton scattering in kithium iodate. *IEEE Journal of Quantum Electronics* QE-21: 110–113.

(4) Pump wavelength 1064 nm; from Basiev TT, Sobol AA, Zverev PG, Ivleva LI, and Osiko VV (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Materials* 11: 307–314.

(5) Pump wavelength 1064 nm; from Basiev TT, Voitsekhovskii VN, Zverev PG, *et al.* (1987) Conversion of tunable radiation from a laser utilizing an LiF crystal containing F_2^- color centers by stimulated Raman scattering in $Ba(NO_3)_2$ and $KGd(WO_4)_2$ crystals. *Soviet Journal of Quantum Electronics* 17: 1560–1561.

(6) Values refer respectively to two linear and orthogonal pump beams at the wavelength 1064 nm; from Zverev PG, Basiev TT and Prokhorov AM (1999) Stimulated Raman scattering of laser radiation in Raman crystals. *Optical Materials* 11: 335–352.

(7) Pump wavelength 1064 nm; from Stolen RH (1978), Fibre Raman lasers. In Ostrowsky D (ed) *Fiber and Integrated Optics*, NATO ASI Series B41. New York, Plenum Press, 157–182.

(8) From Bischel WK and Dyer MJ (1986) Temperature dependence of the Raman linewidth and lineshift for the Q(0)–Q(1) transition in normal and para H_2. *Physical Review* A 33: 3113–3123.

(9) Pump wavelength 1064 nm; from Ottusch JJ and Rockwell DA (1988) Measurements of Raman gain coefficients in hydrogen, deuterium and methane. *IEEE Journal of Quantum Electronics* 24: 2076–2080.

(10) Pump wavelength 694 nm; from Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

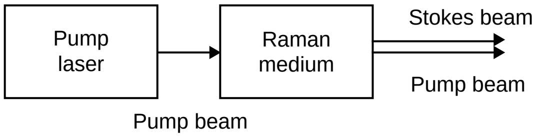

Figure 1 Experimental setup of a single-pass Raman shifter.

literature:

$$\frac{dI_S(z)}{dz} = g_R I_S(z) I_P(z) - \alpha_S I_S(z) \qquad [1]$$

I_S, I_P are the continuous wave intensities of the Stokes and pump waves, z is the propagation distance, g_R the Raman steady-state gain coefficient, and α_S, α_P are the attenuation coefficients at the Stokes and pump frequencies. The Raman gain coefficients, whose general expression as a function of the microscopic material properties can be found in many textbooks, of some of the most commonly used Raman laser media are listed in Table 1.

For short propagation lengths the medium loss can usually be neglected, and if the pump beam is well collimated, the pump intensity can be considered as a constant. For a long interaction length or very absorptive media, the evolution of the pump intensity must also be considered. Though the pump beam is depleted by the scattering process an undepleted pump approximation is often used until high conversion is reached. Under this approximation, the evolution of the pump intensity is mainly governed by diffraction spreading, by losses or by both. In bulk media, single-pass scattered radiation has a coaxial component whose divergence is dictated by the geometry of the pump beam, because Stokes radiation is close to the diffraction limit. Besides this radiation, a nonaxial component can also be observed; this phenomenon stems from the well-known effect of conical emission, which is a phase-matched parametric process. If the Raman medium is within a waveguide, losses are usually the dominating effect. In the undepleted approximation, the evolution of the pump beam intensity as a function of the propagation distance z, either dictated by diffraction or by losses, is known. Then the general solution of eqn [1] is:

$$I_S(z) = I_S(0)\exp\left[g_R \int_0^z I_P(z')dz' - \alpha_S z\right] \qquad [2]$$

where $I_S(0)$ is either the input Stokes signal intensity or an equivalent spontaneous scattering intensity. Then, the amplification gain G of the Stokes beam in the single-pass shifter, for a medium of length L, is given by

$$G = \frac{I_S(L)}{I_S(0)} = \exp\left[g_R \int_0^L I_P(z')dz' - \alpha_S L\right] \qquad [3]$$

Though single-pass Raman shifters are appealing for their simplicity, their application is hampered by the high power necessary for reaching a sufficient efficiency due to the short interaction length that can be obtained, in particular in bulk media, where diffraction dominates. Q-switched pulsed pump lasers are often required, but the propagation in the Raman medium of very powerful pulses requires special precautions to prevent other linear and nonlinear effects occurring. In fact, optical breakdown, thermal and Kerr self-focusing, four-wave mixing, and Brillouin scattering can take place, thus reducing the efficiency of the converter. Precautions may include the optimization of the pump pulse duration, of the focusing, of the interaction length, and the control of the input beam spot size. The multipass configuration has also been exploited to avoid many of the single-pass configuration shortcomings; the typical experimental setup is presented in Figure 2. Note that neither the pump nor the Stokes wave are resonant within the cavity.

The primary advantage of the multipass configuration is the lowering of the pump power required to realize an appreciable conversion. This is due to the increased total gain coefficient, that has been found to be equal to

$$G_n = G\frac{1-R^n}{1-R} + (n-1)\log R \qquad [4]$$

where G is the single-pass gain of eqn [3], R the reflectivity of the mirrors at the Stokes wavelength, and n the number of passes through the Raman medium. There are two main drawbacks of multipass Raman shifters that finally limit the conversion efficiency of this setup. The first is the simultaneous generation of many undesired and uncontrolled Stokes orders, because the first-order Stokes intensity grows rapidly during propagation. The second is the

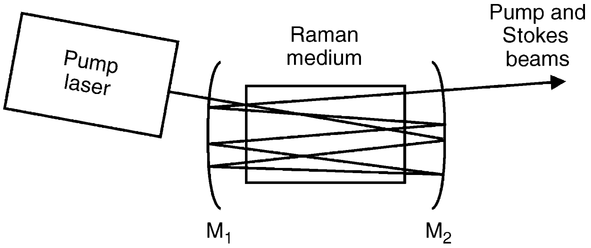

Figure 2 Experimental setup of a multipass Raman shifter. Mirrors M_1 and M_2 reflect both Stokes and pump waves.

intensity decrease associated with the dispersion of pulses, in particular if the Raman medium is a solid. These problems have limited the applications of multipass shifters to hydrogen gas.

Raman Lasers

By placing the Raman medium into an optical cavity, Stokes radiation can be selectively enhanced and Raman oscillation takes place. The two main experimental setups, either with the Raman medium placed in an external cavity or within the pump laser cavity (intracavity), are represented in Figures 3 and 4. In some cases, for the intracavity configuration, the Raman medium may also be the host of the doping laser active material.

For the external resonator Raman laser, the reflectivity of the mirrors is such that they are almost transparent for the pump but highly reflective (resonant) for the shifted radiation. The threshold condition of Raman lasers can be obtained by requiring that the net round trip gain equals the total cavity losses. The net round trip gain, under the plane wave, steady-state, undepleted pump approximation, is given by the single-pass gain G, multiplied by a factor of two to account for the forward and backward amplification. In fact, Raman scattering takes place also between the pump and the counter-propagating Stokes wave reflected by mirror M_2. Losses are given by the medium contribution, which is already accounted for in eqns [2] and [3], and by the energy leakage from the mirrors. Assuming that the mean reflectivities of the mirrors, at the Stokes wavelength, are respectively R_1, R_2, then the threshold condition is $GR_1R_2 = 1$. Through eqn [3] we obtain

$$g_R \int_0^L I_P(z')\mathrm{d}z' - \alpha_S L = -\log(R_1 R_2) \qquad [5]$$

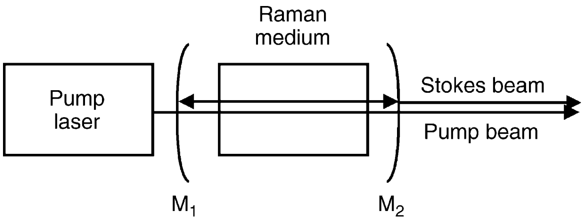

Figure 3 Experimental setup of an externally pumped Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave.

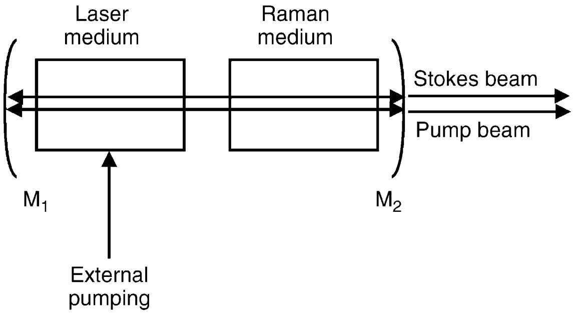

Figure 4 Experimental setup of an intracavity Raman oscillator. Mirrors M_1 and M_2 are highly reflective for both the Stokes and the pump wave.

Equation [5] can be used to evaluate the pump intensity because at threshold low depletion occurs and the approximation taken to obtain [5] is valid. Raman steady-state gain is not very different for gases, liquids and crystals, as shown in Table 1 and oscillation thresholds depend more on the interaction length and the focusing properties of the pump and Stokes beams. Nonetheless, the type of Raman medium imposes many practical constraints for laser design, which are addressed in the next paragraphs.

With gas media it is hard to maintain a high pump intensity over the entire length of the cell, which must be used to contain the gas. Usually gas cells can be from 5 to 30 cm long; this means that Q-switched pump lasers are required to set intensities above 10^4 GW m^{-2}, that is the typical threshold value of Raman oscillation. Continuous wave oscillation of gas Raman lasers is, nonetheless, possible when the pump wavelength is near a medium resonance. In fact, the imaginary part of the third-order non-linear polarization coefficient, to which the Raman gain is proportional, is enhanced near energy transitions because of the resonance in correspondence of that photon energy. The main limitations of this setup are that pump wavelengths are dictated by the Raman medium transitions, and thus also Stokes wavelengths are fixed, and finally conversion efficiency is usually low (about 5% is reported) possibly because of the high absorption. Off-resonance, continuous wave operation can also be attained by using high finesse cavities in a doubly resonant cavity. In fact, by resonating both Stokes and pump wave with high reflectivity mirrors ($R \geq 0.999$), the threshold intensity is lowered and conversion efficiency is high (about 35% has been demonstrated). In spite of the relatively high Raman gain and the large Stokes shifts obtainable, the exploitation of gas cell Raman oscillators is mainly hampered by the need of a high-pressure-resistant vessel and a constant gas flow. The latter is needed because of the low thermal conductivity of gases that might cause local temperature variations and thermal-optical distortions that diminish the output beam uniformity and stability. Moreover the repetition rate of pump pulses must be small (below 50 Hz) so as not to impose a too strong thermal load on the gas.

Liquid media suffer basically the same problems as gases and, moreover, they show higher non-linear refractive indices that result in strong self-focusing. This lens effect results in filamentation, optical breakdown, and high beam divergence. Often Raman liquids are also toxic and very volatile; their use has been slowly but steadily decreasing.

Solid media, in particular crystals, do not show many of the problems listed above and present the advantages of solid-state technology such as compactness, low maintenance, and opportunity of integration. The high density and the restricted atom motion result in high Raman steady-state gains and minimal line broadening. Precise resonance enhancement, as well as low threshold intensities, can be obtained by choosing an adequate mirror reflectivity spectral response. Pump intensities as low as 10^3 GW m^{-2} were reported for the oscillation of an external cavity Raman laser whose medium was a 5-cm-long, barium nitrate (see Table 1) crystal, with mirror reflectivity $R = 0.49$ at the Stokes wavelength. Though the intensity threshold is about an order of magnitude smaller than for gas Raman lasers, Q-switched input pump pulses are still required. Solid-state Raman lasers can be easily forced to oscillate on a TEM$_{0,0}$ mode with divergence close to the diffraction limit even with multimode pumps. Moreover the Stokes beam is subject to a very complex clean-up process; this can be basically explained by the fact that the Stokes beam, growing from the spontaneous emission, does not retain the pump spatial characteristics. Pulse duration is equal to or slightly less than the input one. Conversion efficiency, which depends strongly on mirror reflectivity, can reach 60%, where saturation occurs due to higher-order Stokes wave generation. Solid-state Raman lasers can be easily forced to oscillate also at these higher-order Stokes frequencies if mirror reflectivity is large enough at the relative wavelengths. Second- and third-order Stokes conversion efficiencies of 30% and 20% have been reported.

The theoretical determination of the cavity pump and Stokes fields does not have an analytical solution. For continuous wave or long Q-switched pulsed lasers, eqns [1] must be extended to include all forward and backward Stokes and pump waves and boundary conditions due to the mirrors must be imposed. This problem can only be solved numerically.

The experimental setup of intracavity Raman lasers is shown in Figure 4; both the Raman and the laser medium are placed within the cavity and sometimes the two media coincide. This is, for example, the case of potassium gadolinium tungstate which can be doped with neodymium ions (see Table 1). The cavity is usually resonant for both the pump and the Stokes waves; very high conversion efficiency (up to 80%) is attained when special dichroic coatings are provided such that pump output is suppressed (high-finesse cavity for the pump). Besides the nearly diffraction-limited property of output radiation and clean output beams that are common characteristics of all Raman lasers, enhanced brightness is also observed in this setup. This effect is probably due to the high thermal loading, the consequent linear and nonlinear lens effects, and also to a short pulse generation mechanism that increases output beam intensity. This regime can be reached by reducing the Stokes cavity length with respect to the pump cavity length. A first setup is to use an additional dichroic intracavity mirror that is transparent for the pump but highly reflective for the Stokes; an alternative setup exploits a double ring cavity configuration as depicted in Figure 5.

The principle governing the compression can be explained as follows. The optical paths of the pump and Stokes pulses are made different by an amount ΔL. Then the Stokes pulse generated within the Raman medium, once re-injected into the cavity from mirror M_{S1}, will overlap with an undepleted portion of the recirculating pump pulse. Hence, another Stokes pulse is generated at each round trip, until the pump pulse is completely depleted; the Stokes output finally consists of a train of pulses of duration $\Delta L/c$, at a repetition rate equal to the cavity round-trip. This mechanism leads to pulses as short as the limit imposed by the Raman gain bandwidth. In fact, when the pulse duration becomes comparable to the inverse of the gain bandwidth, transient effects take place because of the delay in the nonlinear medium response. Ring Raman lasers have been obtained with

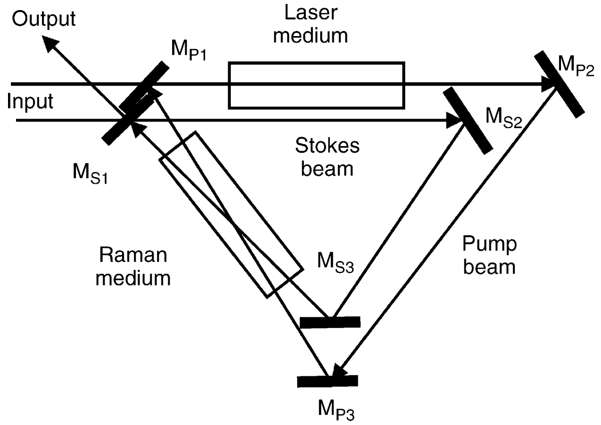

Figure 5 Experimental setup of an intracavity, double ring, Raman oscillator. Mirrors M_{S2}, M_{S3}, M_{P2}, and M_{P3} are highly reflective respectively for the Stokes and the pump wave. M_{S1}, M_{P1} are partially reflective respectively for the Stokes and pump wave.

gases (e.g., hydrogen) and crystal media (barium nitrate and lithium iodate).

Another commonly used experimental setup is that of synchronously pumped ring Raman lasers. In such devices, whose typical setup is sketched in Figure 6, a new input pump pulse is synchronized with the output Stokes pulse thus enhancing the conversion within the Raman medium.

Though electronic trigger detection is needed to lock the pump pulse time repetition rate to a multiple or submultiple of the cavity round-trip, synchronously pumped Raman lasers are relatively simple, stable, high repetition rate sources of radiation. Very compact and reliable synchronously pumped Raman lasers can be realized with optical fibers; in this case the feedback of the ring resonator can be realized by means of a fiber directional coupler. The device, besides the high integration achieved by the all-fiber technology, can also be continuously tuned over a large bandwidth.

Continuous tuning of the same laser source is a very desirable property; the Raman lasers so far described can generate radiation at wavelengths restricted to a few values. In fact, output spectral lines can be changed either by replacing the Raman medium, or the pump laser, or by exploiting cascade Raman scattering or finally resonating a different vibrational line of the same material (if there are many as for example in hydrogen gas); all these cases require substantial cavity changes. Frequency switching without setup changes is attained only in case of potassium gadolinium tungstate. For this medium there exist two independent directions, orthogonal to the crystal optical axis, for which the Raman shifts are different if the pump polarization changes. Then, the wavelength selection is just a matter of controlling the input pump polarization. Though Raman lasers can greatly expand the wavelength range of available laser sources, controlled, continuous tuning of the same Raman laser over a certain frequency band can be obtained only in a few particular cases.

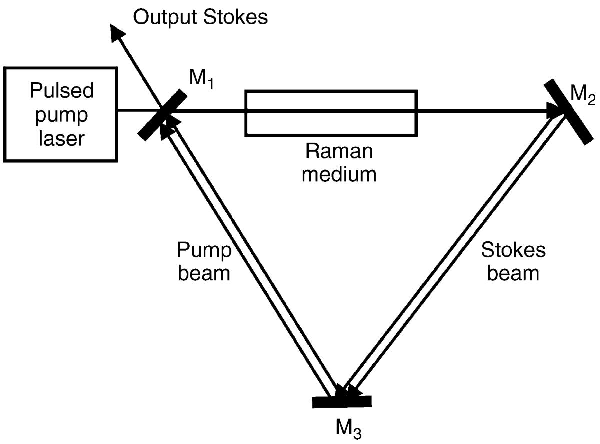

Figure 6 Experimental setup of a synchronously pumped, ring, Raman oscillator. Mirrors M_2 and M_3 are highly reflective for both the Stokes and the pump wave. M_1 is partially reflective for both the Stokes and pump wave.

Wavelength tuning can be attained by selecting the direction of propagation in polar crystals (like lithium niobate and lithium iodate). Different directions allow for wavevector matching of polariton scattering at different frequencies. Also phase matching of parametric interactions, like four-wave mixing, is a possible technique to attain wavelength tuning. However, such Raman lasers cannot operate at maximum gain, and this fact decreases the output energy, the conversion efficiency, and the stability, and increases the pump intensity threshold. Continuous tuning with an external control parameter is attained in the spin-flip Raman semiconductor laser. Semiconductors are Raman media that can be exploited to build far-infrared or even submillimeter Raman lasers, with pump intensity thresholds that are comparable with those of gas and liquid media. Though frequency shifts are fixed, under normal conditions, tuneable operation can be achieved if the semiconductor is placed in a magnetic field. Each energy level of an electron in the solid is, in fact, split into two, depending on the orientation of the electron spin with respect to the magnetic field. When there is a transition involving Raman scattering, the electron spin can change its direction. Then the wavelength of the laser can be continuously varied by changing the magnetic field applied. Tuning over a range of several micrometers in the infrared has been attained and spin-flip Raman lasers have been developed with semiconductors like InSb, $Hg_{0.77}Cd_{0.23}Te$, and $Pb_{0.88}Sn_{0.12}Te$. Use of this technique is severely limited by the need for magnetic field sources.

The simplest method to achieve tuneable radiation is through the use of a material that shows a broad Raman linewidth. Inhomogeneous broadening of the energy level is a well-known effect that takes place in noncrystalline solid media, such as glasses. Though line broadening is accompanied by a decrease in the steady-state Raman gain of about two orders of magnitude with respect to a crystalline medium, this disadvantage can be over-compensated by the increase in the Raman medium length that can be of 3 or 4 orders of magnitude. In fact, glasses are commonly exploited to make optical waveguides, in particular optical fibers, that present very low optical absorption coefficients, in the near-infrared part of the spectrum (1200–1600 nm).

Optical fiber Raman lasers based on silica telecommunication fibers are very versatile, compact, and simple to implement. In principle the setup is not very different from the corresponding external cavity or intracavity Raman lasers so far presented; however, fiber lasers show the great advantage of an integrated, all-fiber technology.

Cavity mirrors can be replaced by fiber Bragg grating reflectors (Figure 7) that can be constructed with a very well-defined spectral reflectivity, thus allowing fine enhancement or suppression of the desired wavelengths. By splicing the Raman fiber, which can be several hundred meters long, with the gratings, the net losses are very low, finally decreasing the pump intensity threshold. The pump threshold for first-order Stokes generation can be easily as low as 200 mW; this means that diode lasers can be used for pumping. Moreover, the need for cavity alignment is eliminated, the laser stability increased and the construction simplified. This simplification allows us to build, much more easily than by using bulk optics, nested multiple resonant cavities, to generate higher-order Stokes beams (Figure 8).

Continuous-wave lasing of up to six Stokes orders has been reported for a fiber Raman laser of this type, with a total injected power of about 7 W. Another key advantage of Bragg gratings is that their reflectivity can be adjusted by means of a piezoelectric stretching of the fiber support; this allows one to control the output intensity of all waves. As previously mentioned, ring cavities can also be constructed by exploiting an all-fiber technology, if feedback is provided by a fiber directional coupler (Figure 9).

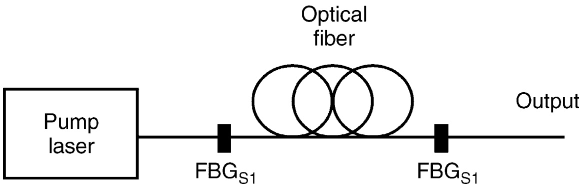

Figure 7 Experimental setup of a continuous-wave, all-fiber, Raman oscillator. FBG$_{S1}$ are fiber Bragg gratings highly reflective for both the first Stokes wave.

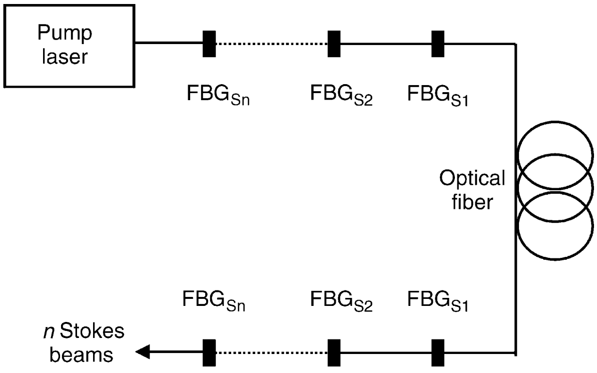

Figure 8 Experimental setup of a continuous-wave, all-fiber, *n*th-order Stokes, Raman oscillator. FBG$_{Si}$ are fiber Bragg gratings highly reflective for the *i*th Stokes wave.

Pulsed operation of fiber Raman lasers has been obtained too; due to the long interaction length, Q-switched pump lasers are no longer needed but rather mode-locked sources can be used. This means that trains of very short laser pulses, down to the hundreds of femto-second regime, can be obtained from fiber Raman lasers, the limit being represented by the Raman bandwidth, which is as large as about 7 THz, at the half maximum, in silica fibers. Propagation of short and intense laser pulses in an optical fiber may induce several other nonlinear effects, like self-, cross-phase modulation, and four-wave mixing. Besides nonlinearity, linear dispersion and velocity mismatch become important too, due to the long propagation. All these linear and non-linear effects limit the Stokes pulse quality. In more detail: self- and cross-phase modulations lead to the spectral broadening of both pump and Stokes waves; broadening is asymmetric, because of the contemporary action of group velocity mismatch. In fact during the propagation the Raman pulse walks off the pump pulse, because of velocity mismatch, and this induces an asymmetry in gain depletion and in nonlinear index modulation. The final outcome is that Stokes waves can be asymmetric both in time and in frequency, and also highly chirped. Four-wave mixing is usually negligible, if phase-matching conditions are not imposed and Brillouin scattering is not an issue for short pulses. Very symmetric and low chirped pulses are attained as soon as the fiber Raman laser is working in the soliton regime, that is when compensation between the dispersion and the nonlinear phase shift occurs. Synchronously pumped, fiber Raman, soliton lasers are able to provide 200-fs pulses, at various wavelengths and with very high conversion efficiency (50%). Intracavity fiber Raman lasers have

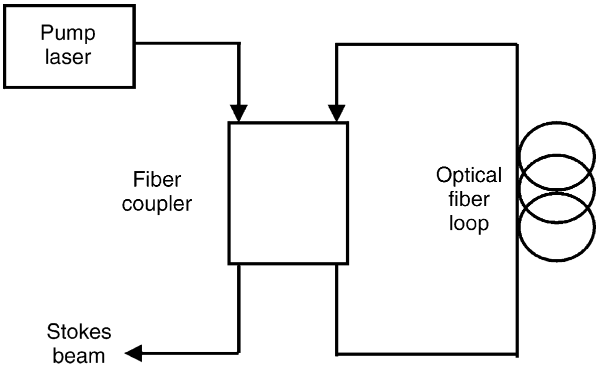

Figure 9 Experimental setup of an all-fiber, ring Raman oscillator.

also been realized by doping the fiber with neodymium, erbium, or ytterbium ions.

The extremely broad Raman spectral linewidth of the glass allows for broadband tuning of the emitted wavelength. For continuous and pulsed wave operation the change in the peak Stokes wavelength can be attained through an intracavity frequency selective element, like for example a prism or a grating. In this case tuning is not obtained in an all-fiber configuration but rather with bulk optical elements and thus additional lenses or objective are needed to collimate the Stokes wave at the fiber ends. The experimental setup is shown in Figure 10.

The laser can be continuously tuned, over about 6 THz, by adjusting the prism or grating rotation. All-fiber Raman laser tuning is possible in a double ring configuration with an inserted optical filter, whose experimental setup is shown in Figure 11. Stokes and pump waves are counterpropagating in the Raman fiber; the Stokes wave frequency is selected by the tuneable optical filter.

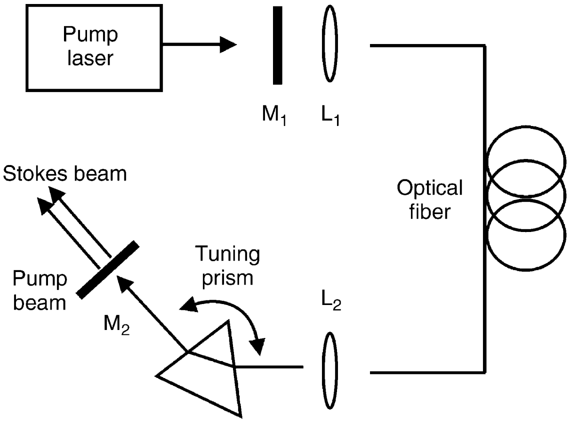

Figure 10 Experimental setup of a tuneable, fiber Raman oscillator. Mirrors M_1 and M_2 are highly reflective for the Stokes wave but not for the pump wave. L_1 and L_2 are collimating lenses or objectives. The tuning prism is rotated to control wavelength selection.

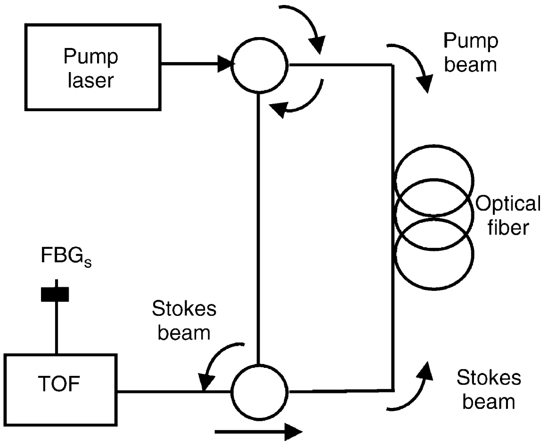

Figure 11 Experimental setup of a double-ring fiber Raman oscillator. OC_1 and OC_2 are fiber optic circulators, and TOF is a tuneable optical filter.

The devices of Figures 10 and 11 stop emitting Stokes radiation for frequencies at which the Raman gain has decreased to a level such that laser threshold is higher than the pump intensity. Due to the fact that Raman gain in fibers is almost constant over about 3 THz, tuning within this frequency range does not lead to any sensible decrease in conversion efficiency. For synchronously pumped ring fiber Raman lasers another technique, named time dispersion tuning, can be used. The experimental setup is shown in Figure 12.

The output fiber end is placed on a moveable mount, that controls the relative delay between the Stokes pulse and the new pump pulse at the fiber input. Since Stokes pulses are dispersed during fiber propagation, to each time slot within a pulse there corresponds a certain portion of the Stokes spectrum. By properly delaying the time of injection of the Stokes pulse back into the fiber the most powerful part of the pump pulse is overlapped to a different part of the Stokes pulse, and this causes a preferred amplification of certain frequencies with respect to others. The shift of the Stokes peak wavelength $\Delta\lambda$ is given by:

$$\Delta\lambda = \frac{\Delta l}{cLD(\lambda)} \tag{6}$$

where Δl is the translation step, c the speed of light, L the fiber length, and $D(\lambda)$ the fiber dispersion at the Stokes wavelength λ. A total tuning range of 10 nm can be obtained with this technique that can also be used as a test measurement of fiber dispersion.

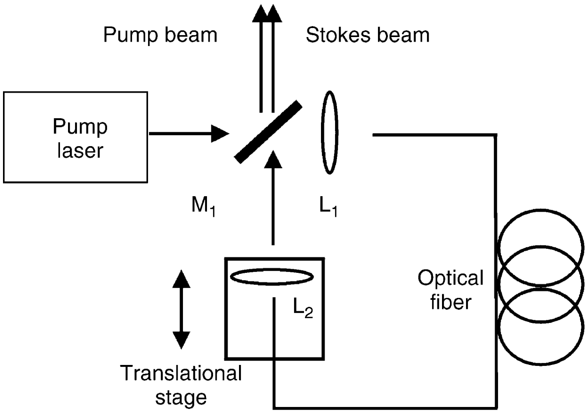

Figure 12 Experimental setup of a time dispersion tuned, ring fiber Raman oscillator. L_1 and L_2 are lenses or objectives. The output collimating lens L_2 and the fiber end are mounted on a translation stage to control wavelength selection by time delaying the re-entrant Stokes pulse.

Applications

Raman lasers have found a large variety of applications. Gas and liquid Raman lasers are fundamental tools for spectroscopic investigation. Solid-state Raman lasers have been shown to be very powerful, eye-safe sources for light detection and ranging (LIDAR) and medical applications. Fiber Raman lasers are exploited for long-haul fiber optics transmission. They can be either sources of short pulses and solitons or high-power continuous-wave sources at multiple wavelengths for subsequent link-distributed Raman broadband amplification.

List of Units and Nomenclature

Amplification gain	[dimensionless]
Efficiency	[dimensionless]
Energy	[J]
Frequency	[Hz]
Intensity	[W m^{-2}]
Length	[m]
Loss coefficient	[m^{-1}]
Power	[W]
Raman bandwidth	[Hz]
Raman gain	[m W^{-1}]
Reflectivity	[dimensionless]
Wavelength	[m]

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Measuring Fiber Characteristics; Nonlinear Effects (Basics); Overview. **Fiber Gratings**. **Optical Amplifiers:** Optical Amplifiers Long-Haul Transmission Systems. **Imaging:** Lidar. **Laser-Induced Damage of Optical Materials**. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Magneto-Optics:** Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics; Phase Matching; Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Optical Materials:** Optical Glasses. **Solitons:** Temporal Solitons. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy.

Further Reading

Band YB, Ackerhalt JR, Krasinski JS and Heller DF (1989) Intra-cavity Raman lasers. *IEEE Journal of Quantum Electronics* QE-25: 208–213.

Basiev TT and Powell RC (eds) (1999) Special issue on solid-state Raman lasers. *Optical Materials* 11: 301–390.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Eckhardt G (1966) Selection of Raman laser materials. *IEEE Journal of Quantum Electronics* QE-2: 1–8.

Grasiuk AZ and Zubarev IG (1978) High power tunable IR Raman lasers. *Applied Physics* 17: 211–232.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-Dubois D (eds) *Laser Handbook*. Amsterdam: North Holland.

Koechner W (1999) *Solid-state Laser Engineering*. Berlin: Springer-Verlag.

Pask HM (2003) The design and operation of solid-state Raman lasers. *Progress in Quantum Electronics* 27: 3–56.

Stolen RH (1978) Fibre Raman lasers. In: Ostrowsky D (ed.) *Fiber and Integrated Optics*, NATO ASI Series B41, pp. 157–182. New York: Plenum Press.

Suto K and Nishizawa J (1994) *Semiconductor Raman Lasers*. Norwood, MA: Artech House.

Weber A (1979) *Raman Spectroscopy of Gases and Liquids*. Topics in Current Physics, vol. 11, Berlin: Springer.

Self-Focusing and Related Effects (Solitons and Multiphoton Absorption)

R L Sutherland, Science Applications International Corporation, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When an intense beam of light with axial symmetry propagates in a medium possessing a nonlinear index of refraction, it will induce its own lens in the medium and be focused. This phenomenon, first discovered just a few years after the invention of the laser, is called self-focusing. It is one of many phenomena under the general classification of self-action, which includes such effects as self-trapping, self-bending, self-phase modulation, and self-steepening. This general class of phenomena involves action on the beam of light induced by the beam itself: light causes a change in an optical property of the medium, which in turn produces an effect back on the light beam. We will discuss the mechanisms responsible for self-focusing in nonlinear optical media and the resultant action on the optical beam. In this we will see how the topics of self-trapping and self-phase modulation are intimately connected to self-focusing. This will lead to a discussion of related effects as well as many potential applications.

It is instructive to first review the properties of a conventional thin lens, illustrated in Figure 1. By 'thin', we mean that just before and just after the lens, the beam size and shape are approximately the same. However, the phase of the optical wave is changed on passage through the lens. We apply the paraxial (aberrationless) approximation where only rays that lie near the lens axis are considered. In this treatment the spherical surfaces of the lens are approximated as parabolic surfaces. For a symmetric double-convex lens (as shown in Figure 1), this phase change is

$$\Delta\varphi(r) = -k(n-1)\frac{r^2}{R} \equiv -\frac{kr^2}{2f}, \quad f^{-1} = (n-1)\frac{2}{R} \qquad [1]$$

where n is the index of refraction of the lens material, and R is the radius of curvature of the lens surfaces. The light beam has a wavevector $k = \omega/c$, where ω is the frequency and c the speed of light in a vacuum. The phase retardation is a maximum on the lens axis (z) and decreases quadratically with the radial coordinate r. In other words, the phase advances quadratically with distance from the axis and forms a spherical, converging wave that comes to focus on the z-axis a distance f (the focal length) from the lens.

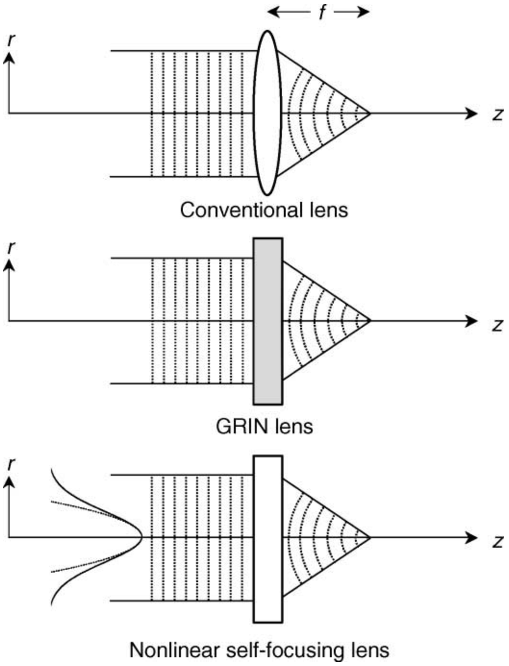

Figure 1 Focusing with a conventional lens, a graded-index (GRIN) lens, and a nonlinear self-focusing lens.

Next consider a graded-index (GRIN) lens, also illustrated in Figure 1. This is a flat slab of material with thickness L and an inhomogeneous index of refraction given by $n(r) = n_0(1 - 1/2\alpha^2 r^2)$. The index has a maximum (n_0) on the lens axis and decreases quadratically with radial distance from the axis. The constant α is a measure of the rate of the index change with r. Upon passage of light through a GRIN lens, the phase change is

$$\Delta\varphi(r) = -kn_0\alpha^2 L\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \quad f^{-1} = n_0\alpha^2 L \qquad [2]$$

Because of the parabolic nature of the index distribution, the phase retardation again decreases quadratically with r and thus forms a converging spherical wave. Notice that for both the conventional and GRIN lenses, the optical thickness of the lens is parabolic. In the conventional lens the index is constant, but the physical thickness varies continuously. In the GRIN lens, the physical thickness remains constant while the index varies continuously. Both produce the same kind of focusing action.

Finally, consider a flat slab of homogeneous material that has a nonlinear index of refraction. Light of intensity I induces a change in the refractive index given by $\Delta n = n_2 I$. The parameter n_2 is called the nonlinear refractive index coefficient and has physical units of $\text{cm}^2\ \text{W}^{-1}$. There are other definitions of n_2 given in the literature where the index change is given in terms of the square of the electric field, which would obviously have different physical units. We will adopt the intensity form of n_2 in this discussion. We now assume the light incident on this slab is a plane wave with a Gaussian intensity distribution given by $I(r) = I_0 \exp(-2r^2/w^2)$ as shown in Figure 1. I_0 is the peak intensity along the z-axis, and w is the radius of the beam at $1/e^2$ of the peak intensity. Consistent with the paraxial approximation, we expand the exponential function, keeping only the first two terms, to yield $I(r) \approx I_0(1 - 2r^2/w^2)$. This parabolic approximation to the intensity is sketched as the dotted curve in Figure 1. The index distribution induced in the medium by the intensity approximately follows this parabolic shape, and the resultant phase change is

$$\Delta\varphi(r) = -k4n_2I_0\frac{L}{w^2}\frac{r^2}{2} \equiv -\frac{kr^2}{2f}, \qquad f^{-1} = \frac{4n_2I_0L}{w^2} \qquad [3]$$

Note the similarity to the GRIN lens, with the association of the peak index $n_0 \to n_2I_0$, and $\alpha^2 \to 4/w^2$. We can say that the nonlinear medium acts like a GRIN lens with an intensity-dependent focal length. However, this correspondence is only accurate within the parabolic approximation, and the nonlinear lens can introduce significant aberrations to the focal properties of the full beam.

In the paraxial approximation, the conventional, GRIN, and nonlinear lenses all have the same effect on an incident beam of light. Although we have considered only plane parallel incident waves, it should be obvious that the GRIN and nonlinear lenses will produce analogous effects on converging and diverging wavefronts as the conventional lens. The effect of a stack of nonlinear lenses in an optical train can also be analyzed in the same way as a stack of conventional lenses, with the exception that the focal length of each lens will depend on the local intensity and hence size of the beam. Finally, we note that if we change the sign of R, α^2, or n_2, we obtain a negative focal length. This is a diverging lens, and in the nonlinear optics realm the effect is called self-defocusing.

Nonlinear Index of Refraction

The concept of an intensity-dependent refractive index has its roots in the power series expression of the nonlinear dielectric polarization P in terms of the electric field E of the optical wave, which in its simplest form is

$$\begin{aligned} P &= \varepsilon_0\left(\chi^{(1)}E + D^{(2)}\chi^{(2)}E^2 + D^{(3)}\chi^{(3)}E^3 + \cdots\right) \\ &= \varepsilon_0\left(\chi^{(1)} + D^{(2)}\chi^{(2)}E + D^{(3)}\chi^{(3)}E^2 + \cdots\right)E \end{aligned} \qquad [4]$$

where $\chi^{(m)}$ $(m = 1, 2, 3\ldots)$ is called the m-th order susceptibility. The quantity $D^{(m)}$ is called the degeneracy factor and is 3 for $m = 3$ in the self-focusing case. This leads directly to the definition of a nonlinear susceptibility $\chi(E)$. All materials have first- and third-order susceptibilities. Centrosymmetric materials (e.g., gases, liquids, cubic crystals) do not possess a second-order susceptibility. Here, we will consider only centrosymmetric media and briefly discuss second-order materials later.

The refractive index is $n = [1 + \chi(E)]^{1/2}$. Consequently, the linear index (small electric field) is $n_0 = (1 + \chi^{(1)})^{1/2}$. The leading nonlinear term in $\chi(E)$ is generally small. Expanding the square root in a binomial series, the first nonlinear term in the refractive index can be approximated by $3\chi^{(3)}E^2/2n_0$. Since $I = 2\varepsilon_0 n_0 cE^2$, we see that

$$n = n_0 + n_2I \qquad [5]$$

where $n_2 = 3\chi^{(3)}/4\varepsilon_0 n_0^2 c$. Third-order (cubic) materials that exhibit an intensity-dependent refractive index described by eqn [5] are called Kerr or Kerr-like.

Several physical mechanisms produce a refractive index of the form given in eqn [5]. A list of the significant ones is presented in Table 1. We emphasize that these mechanisms lead to eqn [5] over a limited range of intensity. For example, in materials where the main mechanism is electronic state population redistribution, when the incident intensity becomes comparable with the saturation intensity, the index begins to saturate (i.e., the change of index with

Table 1 Physical mechanisms responsible for the nonlinear index of refraction

Mechanism	*Typical n_2 ($cm^2\ W^{-1}$)*	*Typical sign*	*Typical response time (seconds)*
Electronic polarization	$\sim 10^{-16}$–10^{-15}	+	$\sim 10^{-15}$
Molecular orientation			
Isotropic liquids	$\sim 10^{-15}$–10^{-14}	+	$\sim 10^{-12}$
Liquid crystals	$\sim 10^{-4}$–10^{-3}	+	$\sim 10^{-3}$–10^{-2}
Electrostriction	$\sim 10^{-13}$–10^{-12}	+	$\sim 10^{-9}$
Population redistribution			
Atomic vapors (ω below resonance)	$\sim 10^{-10}$–10^{-9}	−	$\sim 10^{-8}$
Atomic vapors (ω above resonance)	$\sim 10^{-10}$–10^{-9}	+	$\sim 10^{-8}$
Semiconductors	$\sim 10^{-6}$	+	$\sim 10^{-8}$
Band filling (semiconductors)	$\sim 10^{-6}$–10^{-4}	±	$\sim 10^{-8}$–10^{-7}
Exciton saturation (quantum wells)	$\sim 10^{-4}$	−	$\sim 10^{-8}$
Thermal	$\sim 10^{-8}$	±	$\sim 10^{-4}$

intensity becomes sublinear). The simple expression in eqn [5] no longer applies. These types of effects can be important and will be discussed later. Finally, we mention that other mechanisms exist that produce self-focusing related effects, but they are not Kerr-like and will be briefly discussed later.

We learn from Table 1 that values of n_2 range over several orders of magnitude. High-intensity lasers are required to observe the effects of small optical nonlinearities. Hence pulsed lasers are often used in self-focusing applications. Laser pulse widths can generally range from 100 s of nanoseconds to 10 s of femtoseconds. Consequently, one must keep in mind the response time of the mechanism/material when analyzing experiments. If the pulse width is long compared to the response time, a quasi-steady-state self-focusing regime results. If the pulse is too short, transient effects will appear. Also, as one goes to longer response time mechanisms, lower intensities can be used. In some cases it is even possible to use cw lasers.

Self-Lensing

Thin nonlinear media can be employed in self-lensing applications, where self-focusing is used to modify the far field properties of a beam, just as a conventional lens would do, but with effects dependent on the beam intensity. Gaussian beams are usually employed in the paraxial approximation.

A popular technique employing this effect to measure the magnitude and sign of n_2 is the Z-scan. A typical experiment is illustrated in Figure 2. A thin nonlinear medium is scanned axially between a focusing lens and an aperture in the far field. Consider a material with positive n_2. When the sample is close to the lens the intensity is too low to induce significant self-lensing. As the sample approaches the focal point of the lens, the intensity rises and a positive self-lens is induced that focuses the beam more strongly. This displaces the focus to the left, causing the beam to spread out more in the far field and less power to pass through the aperture. As the sample passes to the right of the focal point, the positive self-lens reduces the divergence of the beam so that more power is passed through the aperture. Near the focal point the self-lens is strongest, but a thin lens at this point has no effect on the beam passing through the far-field aperture. The measured aperture transmittance as a function of scan-position Z is sketched in Figure 2. The peak and valley of this dispersion-shaped curve will be interchanged for a negative n_2. For a small aperture and fixed power of the laser, the difference between the peak and valley transmittance is proportional $|n_2|$.

The same experimental configuration is also popular for measuring small absorption coefficients or thermo-optic coefficients ($\mathrm{d}n/\mathrm{d}T < 0$ for most materials). Under steady-state conditions (cw beam), the thermal n_2 is given by

$$n_2 = \frac{\alpha\tau}{\rho C}\frac{\mathrm{d}n}{\mathrm{d}T} \qquad [6]$$

where α is the absorption coefficient, τ is the thermal diffusion time, ρ is the mass density, and C is the specific heat. This technique is extremely sensitive and can easily be observed, for example, with a ~1 mW laser pointer and a 1 cm cuvette filled with minute quantities of ink or food coloring dissolved in methanol.

This combination of self-lensing and a hard aperture has also been put to use in devices. One such device is the optical power limiter. A typical configuration is similar to the Z-scan experiment in Figure 2, with the sample fixed near the lens focal point. Thick media in the focal region have also been used. At sufficiently low optical power, the medium acts like a flat window. As the input power rises, self-lensing sets in and spreads the beam out at the aperture, limiting the transmitted power. The higher the incident power gets, the more the beam spreads. The result is a transmitted power that is clamped below a preferred level. Thermal lensing has been used for cw beams, while positive Kerr

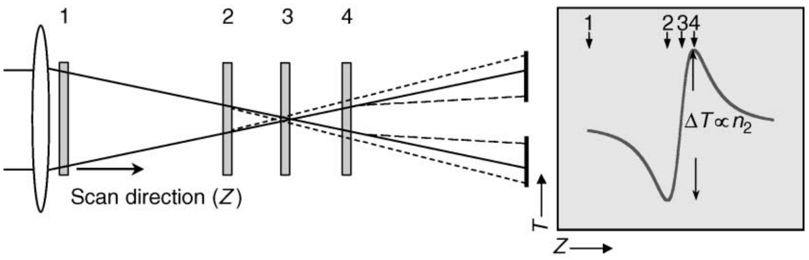

Figure 2 A Z-scan experiment. The transmittance through the aperture as a function of the position (Z) of the thin nonlinear medium is sketched at the right.

self-lensing has been employed for Q-switched lasers.

A less obvious but very important application of this configuration is for mode-locking lasers (see Figure 3). Mode-locking is a coherent phenomenon that locks the relative phase of multiple longitudinal modes in a laser having a broadband gain profile. These stabilized modes interfere to form a train of short pulses with pulsewidth inversely proportional to the gain bandwidth. A common technique for passively mode-locking a laser is to insert a saturable absorber into the laser cavity. A saturable absorber blocks low-power light, keeping the round-trip gain below the oscillation threshold, but transmits high-power light as the absorption coefficient saturates. Thus the mode-locked condition is favored for gain and oscillation since the peak power of the pulses is high. It turns out that transverse spatial modulation, due to self-focusing combined with an aperture, can produce the same type of effect in a laser cavity. The laser rod acts as both the gain medium and nonlinear lens. For example, titanium: sapphire ($Ti:Al_2O_3$) has $n_2 = 3 \times 10^{-16}$ cm^2 W^{-1}. Thus, in a $Ti:Al_2O_3$ laser the rod produces self-lensing that modifies the beam shape propagating through an intracavity aperture. The loss due to the aperture is intensity dependent, as in the Z-scan. Therefore, the net round-trip gain of the cavity is low for low-power but high for high-power pulses, and the mode-locking condition is fulfilled. These types of lasers are called Kerr lens mode-locked (KLM) lasers. KLM lasers are generally not self-starting, but can be self-sustaining. They, therefore, require some active means of starting the mode-locking. Ultrashort pulses (<100 fs) have been obtained with a cw KLM $Ti:Al_2O_3$ laser. For pulsewidths this short, the simple quasi-steady-state self-focusing picture is not entirely adequate to describe the dynamics. Other factors, such as group-velocity dispersion (GVD) and self-phase modulation (SPM), must be considered. Moreover, since the laser rod is typically several millimeters long, it is more appropriate to model it as a thick rather than a thin nonlinear lens.

Beam Trapping

When an intense beam propagates in a thick nonlinear medium, a question arises as to the ultimate fate of the beam. A possibility that has intrigued researchers for a long time is diffractionless propagation of a beam over long distances. In other words, self-focusing exactly balances diffraction to force the beam into a nonlinear waveguiding mode. This phenomenon is called beam trapping or self-trapping. A simplified analysis of a uniform beam captures the essential physics (see Figure 4). The refractive index inside the beam is higher than outside by eqn [5]. Therefore, light incident on the 'interface' at the edge of the beam can experience total internal reflection if its angle of incidence exceeds the critical angle, dependent on the value of the nonlinear index. If the complement of the critical angle matches the diffraction angle of the beam, the beam is trapped in its own self-induced waveguide. The critical power $\mathscr{P}_{cr2}$ at which this occurs is given in Table 2. The condition for beam trapping is independent of beam size since both diffraction and self-focusing depend on beam area, and these size effects cancel.

A detailed solution of the wave equation for a Gaussian beam using the parabolic approximation yields a slightly different critical power $\mathscr{P}_{cr1}$, also given in Table 2. Numerical analyses of the wave

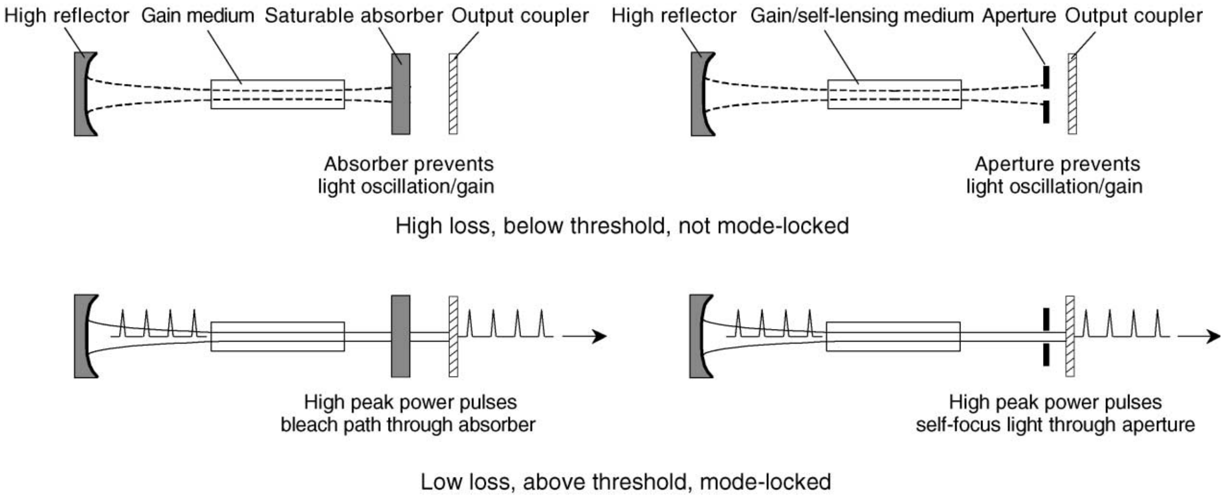
High reflector
Gain medium
Saturable absorber
Output coupler
High reflector
Gain/self-lensing medium
Aperture
Output coupler
Absorber prevents light oscillation/gain
Aperture prevents light oscillation/gain
High loss, below threshold, not mode-locked
High peak power pulses bleach path through absorber
High peak power pulses self-focus light through aperture
Low loss, above threshold, mode-locked

Figure 3 Laser mode-locking using a saturable absorber (left) and self-lensing in combination with an intracavity aperture (right).

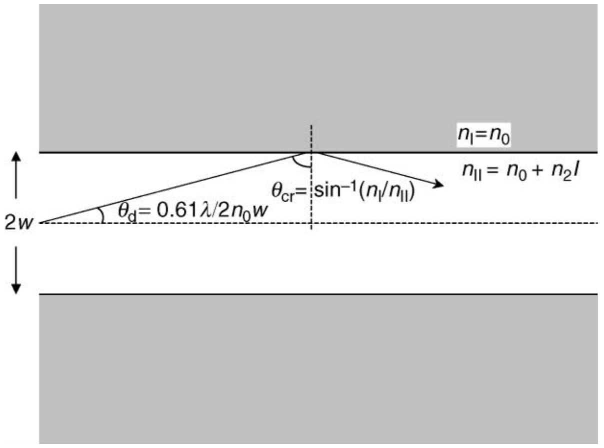
$n_{\mathrm{I}}=n_0$
$n_{\mathrm{II}}=n_0+n_2I$
$\theta_{\mathrm{cr}}=\sin^{-1}(n_{\mathrm{I}}/n_{\mathrm{II}})$
$\theta_{\mathrm{d}}=0.61\lambda/2n_0w$
$2w$

Figure 4 Self-trapping conditions for a uniform beam of intensity I and width $2w$ in a Kerr medium.

Table 2 Critical power for beam trapping (collapse)

Gaussian beam (parabolic approximation)	$\mathscr{P}_{\mathrm{cr1}}=\dfrac{1}{2}\dfrac{\lambda^2}{4\pi n_0n_2}$
Gaussian beam ('whole beam' collapse)	$\mathscr{P}_{\mathrm{cr2}}\approx\dfrac{(1.22\pi)^2}{8}\dfrac{\lambda^2}{4\pi n_0n_2}$
Townes soliton, lower bound for arbitrary beam shape	$\mathscr{P}_{\mathrm{cr}T}\approx1.86225\dfrac{\lambda^2}{4\pi n_0n_2}$
Arbitrary beam (upper bound)[a]	$\mathscr{P}_{\mathrm{cr3}}=G[f]\dfrac{\lambda^2}{4\pi n_0n_2}$

[a] $G[f]$ defined in text.

equation for a whole Gaussian beam have shown for incident power such that $\mathscr{P}_{\mathrm{cr1}}<\mathscr{P}<\mathscr{P}_{\mathrm{cr2}}$ the beam comes to a 'soft' focus and then diffracts (partial beam focusing), but for $\mathscr{P}>\mathscr{P}_{\mathrm{cr2}}$ the whole beam catastrophically collapses (i.e., the beam approaches a singularity). Catastrophic self-focusing has been observed experimentally, but as the beam collapses the intensity becomes so large that additional nonlinear effects, such as multiphoton absorption and stimulated Raman scattering, come into play. These effects rob the beam of energy and terminate the collapse. At this stage the beam will then spread by diffraction.

A more rigorous analysis of this problem is obtained by examining the nonlinear wave equation in the slowly varying amplitude approximation (propagation along z), but including the transverse variation of the wave. If the electric field of the optical wave is of the form $E(\mathbf{r})=A(\mathbf{r})\exp(ikz)$, where now $k=n_0\omega/c$, then the wave equation can be written as

$$2ik\frac{\partial A}{\partial z}+\nabla_\perp^2A+4\varepsilon_0n_2ck^2|A|^2A=0 \qquad [7]$$

where the transverse Laplacian $\nabla_\perp^2$ is $\partial^2/\partial x^2+\partial^2/\partial y^2$ in rectangular Cartesian coordinates, or $\partial^2/\partial r^2+(1/r)\partial/\partial r$ in cylindrical coordinates (assuming azimuthal symmetry). This equation is known as the nonlinear Schrödinger equation (NLSE). In normalized coordinates $\xi=x/w$, $\eta=y/w$, $\rho=r/w$, $\zeta=z/2kw^2$, where w is the input beam radius, and $\psi=2kw(c\varepsilon_0n_2)^{1/2}A$, eqn [7] takes the normalized form of the NLSE:

$$\begin{aligned} &i\frac{\partial\psi}{\partial\zeta}+\nabla_\perp^2\psi+|\psi|^2\psi=0\\ &\nabla_\perp^2=\frac{\partial^2}{\partial\xi^2}+\frac{\partial^2}{\partial\eta^2}=\frac{\partial^2}{\partial\rho^2}+\frac{1}{\rho}\frac{\partial}{\partial\rho}\end{aligned} \qquad [8]$$

A waveguiding (i.e., diffractionless) solution of the form $\psi(\zeta,\rho)=\exp(i\zeta)R(\rho)$ exists for eqn [8], where $R(\rho)$ is a solution of

$$\nabla_\perp^2R-R+R^3=0,\quad R'(0)=0,\quad R(\infty)=0 \qquad [9]$$

$R(\rho)$ is known as the Townes soliton. It has a radial profile similar to a Gaussian but differing in fine detail. Since the Townes soliton exactly balances diffraction and self-focusing, it carries a critical power $\mathscr{P}_{\mathrm{cr}T}$ given by the expression in Table 2. For comparison, $(1.22\pi)^2/8=1.83624$. Therefore, $\mathscr{P}_{\mathrm{cr}T}$ and $\mathscr{P}_{\mathrm{cr2}}$ differ by only about 1%. However, the Townes soliton is unstable. An analysis of the solutions of the NLSE shows that a beam will not catastrophically collapse (i.e., its intensity will not blow up) if its power is less than $\mathscr{P}_{\mathrm{cr}T}$. Thus $\mathscr{P}_{\mathrm{cr}T}$ forms a lower bound for the critical power. An upper bound can also be found from analysis of the NLSE. It depends on the input beam profile $[\psi(\rho,0)\propto f(\rho)]$ and is given as $\mathscr{P}_{\mathrm{cr3}}$ in Table 2. The quantity $G[f]$ characterizes the input beam profile and is given by

$$G[f]=\frac{2\int|f|^2\rho\mathrm{d}\rho\int|\nabla f|^2\rho\mathrm{d}\rho}{\int|f|^4\rho\mathrm{d}\rho} \qquad [10]$$

For example, a Townes soliton, a Gaussian beam, and a hyperbolic secant beam yield $G=1.86225$, $G=2$, and $G=1.86257$, respectively. Thus, if the beam power is such that $\mathscr{P}_{\mathrm{cr}T}\le\mathscr{P}\le\mathscr{P}_{\mathrm{cr3}}$, the beam may collapse, but it will definitely collapse if $\mathscr{P}>\mathscr{P}_{\mathrm{cr3}}$. We note that if the input beam is elliptical, with ellipticity $e=b/a$ (where a and b are the semi-major and semi-minor widths, respectively), then

$$\mathscr{P}_{\mathrm{cr}T}(e)\approx\left(0.6+0.2\frac{e^2+1}{e}\right)\mathscr{P}_{\mathrm{cr}T} \qquad [11]$$

This expression is accurate to within 1.5% of numerical simulations of the NLSE, and also holds well for Gaussian input beams.

The conclusion is that there is no stable self-trapped (2 + 1)-D beam (two diffracting dimensions, one propagating dimension) in a Kerr medium. The beam will diffract, blow up in a finite distance, or go through a single focusing–defocusing cycle. For an incident plane wave, the distance from the input plane to the blowup point is (assuming a Gaussian input)

$$z_{\mathrm{f}} \approx \frac{0.369\pi w^2/\lambda}{(\mathscr{P}/\mathscr{P}_{\mathrm{cr2}})^{1/2} - 0.858} \qquad [12]$$

where the approximation holds well for $\mathscr{P} > 1.2 \times \mathscr{P}_{\mathrm{cr2}}$.

Why, then, does self-trapping appear to occur in Kerr media? Time integrated photographs show what appears to be the formation of a filament, and laser damage tracks in solids follow the shape of a self-trapped filament. This can be explained in terms of the moving focus model. In the quasi-steady-state approximation, each time slice of the pulse comes to focus independently at different times and positions. Figure 5 shows a series of snapshots of a Gaussian spatial and temporal pulse (assume axial symmetry), where the peak power of the pulse is greater than $\mathscr{P}_{\mathrm{cr2}}$. Slices of the pulse with $\mathscr{P} > \mathscr{P}_{\mathrm{cr2}}$ will come to a focus at a finite distance $z_{\mathrm{f}}(\mathscr{P})$, which can be approximated by eqn [12]. The time at which that focus appears is determined by the time it takes for that slice of power $\mathscr{P}$ to propagate to the input plane ($z = 0$) plus propagate to $z_{\mathrm{f}}(\mathscr{P})$. As illustrated in Figure 5, slice 1a takes the minimum time to come to a focus. At a later time slices 2a and 2′ come to focus simultaneously but at different positions, 2a to the left and 2′ to the right of 1a, since $\mathscr{P}(2\mathrm{a}) > \mathscr{P}(1\mathrm{a})$ and $\mathscr{P}(2') < \mathscr{P}(1\mathrm{a})$. The solid curves in the left of Figure 5 are the trajectories of the beam rays showing where each slice will come to focus. Slice 2b, having the same power as 2a, comes to focus at the same position as 2a, but later in time, and so on. The original focus thus appears to split in two, with one branch propagating forward and the other backward. The backward branch comes to a turning point when slice 4 of maximum power in the pulse center comes to focus, then propagates forward as time slices in the back part of the pulse come to focus. The motion of the foci traces out the U-shaped curve shown to the right in Figure 5. The time-integrated picture that appears like a filament is thus just the tracks of moving foci.

The question remains, do conditions exist under which stable self-trapped beams can form? The answer is yes, and these types of beams are called spatial solitons. The (1 + 1)-dimensional NLSE admits the fundamental spatial soliton solution:

$$A(x, z) = A_0 \operatorname{sech}(x/w) \exp(iz/2kw^2)$$
$$|A_0|^2 = \frac{1}{2\varepsilon_0 c n_2 (kw)^2} \qquad [13]$$

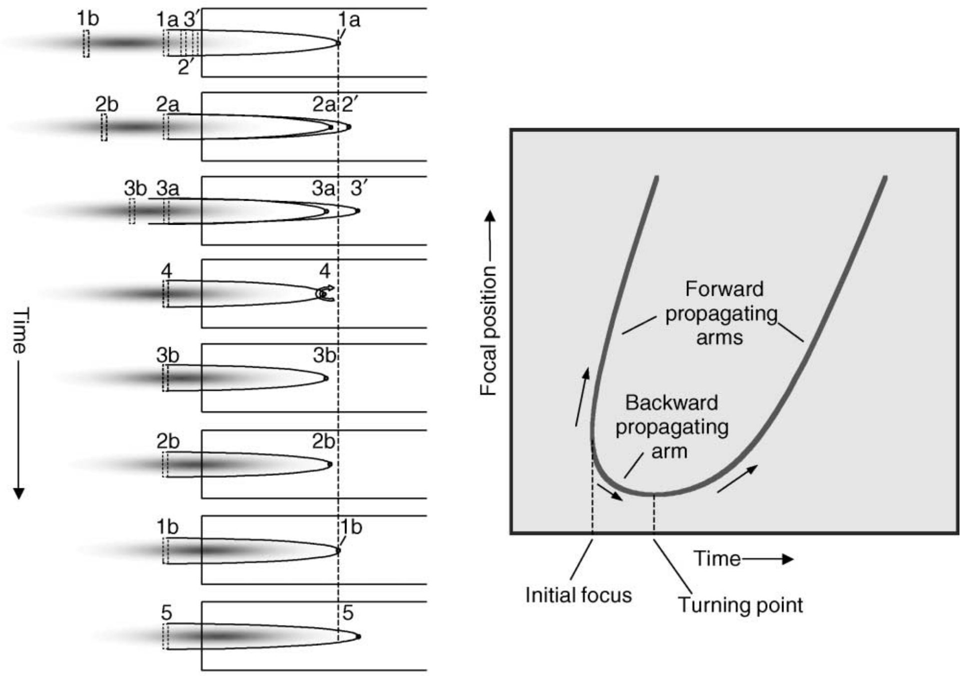

Figure 5 Moving focus model of apparent filament formation by quasi-steady-state self-focusing of a pulsed Gaussian beam. The series of snapshots on the left illustrate when and where different time slices of the pulse come to focus. The tracks of moving foci trace out the curve shown on the right.

The wave given by eqn [13] propagates along z with a constant width w along the x-axis. Such a (1 + 1)-D soliton can be formed in a slab waveguide, as illustrated in Figure 6, consisting of a nonlinear thin film deposited on a linear slab, where the wave is confined along y but can diffract along x. These types of solitons have been confirmed experimentally and shown to be quite robust. They are stable against perturbations and propagate as a self-induced waveguide. Since a soliton changes the refractive index where it propagates, it can also attract and guide other beams. The interactions or 'collisions' of solitons are an active area of current research. Intriguing potential applications include all-optical reconfigurable interconnects for telecommunications and optical computing based on the properties of soliton collisions.

Although (2 + 1)-D spatial solitons are not stable in ideal Kerr media, they can be in Kerr-like media with a saturating nonlinearity. For example, a two-level saturable absorber (e.g., an atomic vapor) at a wavelength tuned above the natural absorption frequency will exhibit a positive Kerr-like response at low intensity ($I \ll I_{\text{sat}}$), but then saturates as $(1 + I/I_{\text{sat}})^{-1}$ for $I \sim I_{\text{sat}}$ or larger, where I_{sat} is the saturation intensity. This can form a stable trapped beam, as was first reported in 1974 for a cw laser in Na vapor, because as the beam self-focuses and the intensity rises, the nonlinear index levels off asymptotically until a balance between diffraction and focusing is reached. This type of situation is stable against perturbations, but may undergo almost periodic focusing–defocusing cycles.

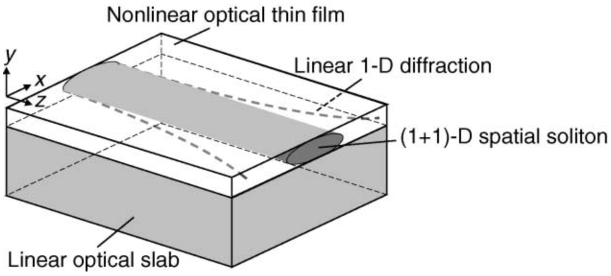

Figure 6 Spatial soliton propagation in a nonlinear slab waveguide.

Spatial solitons employing mechanisms other than the Kerr nonlinearity are also being vigorously studied. One of these is the photorefractive effect. An illustration of the generation of a slab soliton is shown in Figure 7. A Gaussian beam is focused to a line (e.g., by a cylindrical lens) and propagated as an extraordinary wave along the optic (z) axis of a uniaxial photorefractive crystal. A dc electric field is also applied along z. Charge carriers are generated in the slab beam by photo-ionization of dopant atoms and drift in the dc field. The lower resistivity in this region causes the voltage drop to be smaller and hence the field to be lower. The field induces a refractive index change by the Pockels effect ($\Delta n = -n_0^3 | r_{\text{eff}} E|/2$, where r_{eff} is the effective Pockels coefficient and n_0 is the background index). The index change is negative, but is reduced in magnitude where the field is lower. The index change depends on the local intensity, and the

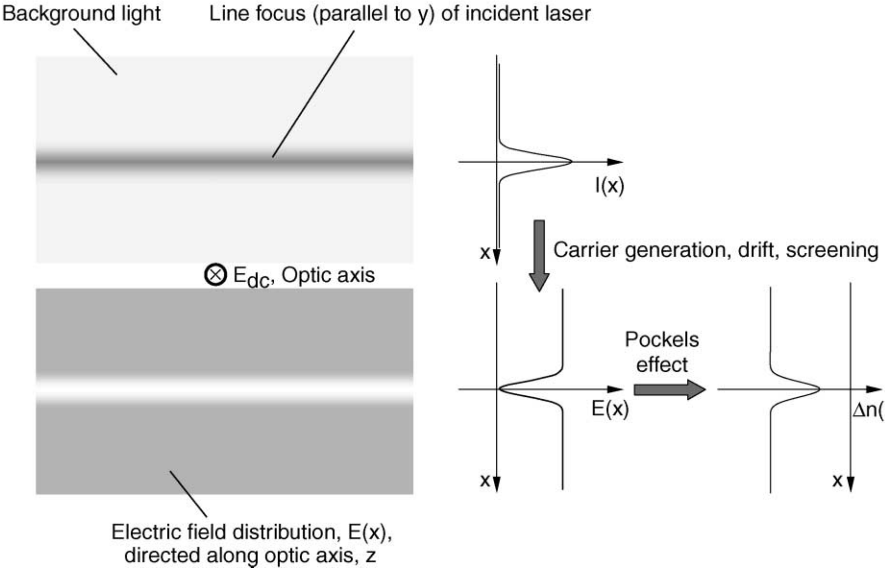
Background light
Line focus (parallel to y) of incident laser
I(x)
x
Carrier generation, drift, screening
⊗ E_dc, Optic axis
Pockels effect
E(x)
Δn(x)
x
x
Electric field distribution, E(x), directed along optic axis, z

Figure 7 Mechanism of slab soliton formation in a photorefractive material.

photorefractive mechanism produces an effective saturable self-focusing nonlinearity. (There are a limited number of charge carriers, and the field cannot be reduced below zero.) This produces a waveguide that traps the slab beam, which then propagates without spreading. Thus the photorefractive soliton is stable. This effect can be observed with milliwatt-level cw lasers, but the space-charge field takes time to build up to a steady state (typically of the order of seconds). Needle (circularly symmetric) solitons have also been observed by launching a TEM_{00} Gaussian beam in a photorefractive crystal, but the theoretical description is not as simple. Other types of spatial solitons and other mechanisms have been explored using the photorefractive and photovoltaic effects. The reader is referred to the literature for a more in-depth discussion.

Another mechanism useful for soliton formation is cascaded second-order (quadratic) effects. This includes such phenomena as second-harmonic generation (SHG) and other frequency conversion processes. Consider SHG with imperfect phase matching. (The reader is referred to other articles discussing harmonic generation and phase matching.) A fundamental (F) wave generates its second harmonic (SH), but because of dispersion $n(\omega)$, the two waves travel at different phase velocities. The SH(2ω) beats with the F(ω) wave to generate, via the second-order nonlinear polarization, another wave at ω that adds to the incident *F* wave. This produces a net phase shift of the *F* wave that accumulates continuously as the waves propagate together and periodically exchange power; this is called cascading. In the nondepleted pump approximation, this phase accumulation can be expressed as

$$\Delta\varphi = -\frac{\Delta k z}{2}\left\{1 - \left[1 + \left(\frac{2\kappa\sqrt{I}}{\Delta k}\right)^2\right]^{1/2}\right\} \qquad [14]$$

$$\kappa = \frac{\omega\chi_{\text{eff}}^{(2)}}{\sqrt{2\varepsilon_0 n_0^3 c^3}}$$

where $\chi_{\text{eff}}^{(2)}$ is an effective second-order susceptibility (dependent on crystal symmetry, polarization vectors, and propagation direction), $\Delta k = 2k(\omega) - k(2\omega)$ is the phase mismatch, I is the intensity of the *F* wave, and n_0 is the approximate refractive index for both F and SH waves. Although not due to a refractive index change, this phase change produces a self-focusing effect for a bell-shaped intensity. For a small enough pump intensity or sufficiently large phase mismatch, an equivalent n_2 can be approximated by

$$n_2 = \frac{\omega(\chi_{\text{eff}}^{(2)})^2}{2\varepsilon_0 n_0^3 c^2 \Delta k} \qquad [15]$$

Note that the sign of n_2 depends on the sign of Δk. For $I \sim (\Delta k/2\kappa)^2$ or larger, eqn [15] is not valid and eqn [14] must be used. Note that the equivalent index change is then sublinear with respect to I, and soliton formation is stable. When pump depletion cannot be ignored, the F and SH coupled-wave equations must be solved numerically to analyze soliton dynamics. Figure 8 illustrates the difference between a beam diffracting normally and a diffractionless quadratic soliton. When an F beam is launched in a nonlinear medium, it does not immediately form a soliton, but evolves into one, shedding some energy in the process. Note that this soliton does not consist of a single frequency beam, but depends on the nonlinear coupling between the ω- and 2ω-waves that continuously exchange energy as the beams propagate. For this reason, these solitons are called simultons.

Self-Focusing in Space–Time

The moving focus model describing the self-focusing of optical pulses becomes inadequate when dispersion effects must be included. This is true in the case of ultrashort pulses (typically ~100 fs or less). To properly describe the propagation of a pulse with group velocity v_g, eqn [7] must be modified by $\partial A/\partial z \rightarrow [\partial/\partial z + (1/v_g)\partial/\partial t]A$. In addition, to include the effects of GVD, a term $-k_2 k(\partial^2 A/\partial t^2)$ is added, where $k_2 = (\partial^2 k/\partial\omega^2)$ is the GVD coefficient. For anomalous GVD, $k_2 < 0$. Taking this to be the case and defining another normalized coordinate $\tau = (t - z/v_g)/(|k_2|k)^{1/2}w$, the normalized NLSE (eqn [8]) becomes

$$i\frac{\partial\psi}{\partial\zeta} + \left(\frac{\partial^2}{\partial\xi^2} + \frac{\partial^2}{\partial\eta^2} + \frac{\partial^2}{\partial\tau^2}\right)\psi + |\psi|^2\psi = 0 \qquad [16]$$

Note the complete symmetry in eqn [16] between the transverse space (ξ, η) and time (τ) variables. There is thus a correlation between diffraction (space) and dispersion (time). The corresponding phenomenon to self-focusing is self-phase modulation (SPM). This effect denotes the additional phase retardation experienced by a pulse in a Kerr medium due to its time-dependent intensity: $\Delta\varphi(t) = -(\omega_0/c)n_2 I(t)z$, where ω_0 is the center (carrier) frequency of the pulse. SPM produces new frequencies ($\mathrm{d}\varphi/\mathrm{d}t$) that are red-shifted on the leading edge of the pulse and blue-shifted on the trailing edge. The frequency shifts in a self-focused filament can be quite sizeable and create a large continuum (a white light pulse) that

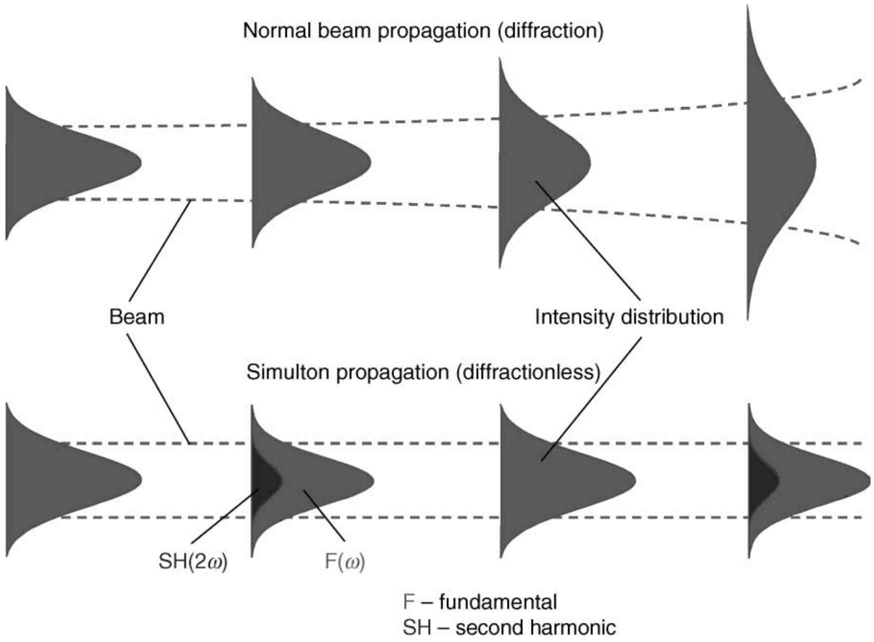
Normal beam propagation (diffraction)
Beam
Intensity distribution
Simulton propagation (diffractionless)
SH(2ω)
F(ω)
F – fundamental
SH – second harmonic

Figure 8 Comparison of normal beam propagation to simulton propagation in a quadratic nonlinear medium. In a simulton, the F and SH waves periodically exchange energy as they co-propagate in the medium.

can be used for ultrafast spectroscopy. In a medium with normal GVD the pulse spreads in time because blue frequencies in the trailing edge travel slower than red frequencies in the leading edge. However, for anomalous GVD the opposite is true, and the pulse can 'focus' or collapse in time.

The simultaneous spatiotemporal collapse of a beam of ultrashort pulses leads to the consideration of $(3+1)$-D soliton generation, where the third 'transverse' dimension is time. While the NLSE leads to unstable solitons for transverse dimensions >1, a saturating nonlinearity could possibly bound the collapse and lead to the formation of stable, diffractionless, dispersionless pulses. These types of pulses are called light bullets. Light bullets have a particle-like nature and display many interesting, exotic properties that could lead to ultrafast switching rates for optical digital logic systems in telecommunications and computing.

The saturable nature of self-focusing in quadratic nonlinear media makes them natural candidates in the search for light bullets, and $(2+1)$-D spatiotemporal simultons have been observed in crystals of lithium iodate. The drawback of ultrashort pulse simulton generation in quadratic media is the added complications of requiring anomalous GVD at both ω and 2ω, and dealing with the effects of group-velocity mismatch (GVM). The phenomenon of GVM involves the different group velocities of the F and SH pulses, which makes them walk away from one another. Without the physical overlap of these two pulses, the simulton will cease to exist. By tilting the amplitude front of the pulses with respect to their phase fronts, it is possible (over a limited distance) to compensate pulse walkoff due to GVM with spatial walkoff due to the birefringence of the quadratic medium. Accomplishing this, however, requires the beam to be in the form of a narrow stripe, and the propagation distance over which pulse overlap is good is greater for a longer stripe. This can be achieved by focusing a beam onto the medium with a cylindrical lens. The simulton then propagates without a change of width temporally or spatially (in 1-D). This is not a light bullet, however, because the confinement is not in all three dimensions.

The elliptic nature of the diffraction–dispersion operator in eqn [16] makes pulse collapse a possibility in nonlinear media with anomalous GVD. With normal GVD, the operator is hyperbolic. For such a medium, it has been predicted and observed that instead of collapsing, a self-focused pulse in a Kerr medium will split in time. Since the red-shifted frequencies generated at the front of the pulse travel faster than the blue-shifted frequencies produced at the back, the front of the pulse separates from the back part. Collapse of the pulse, as well as self-trapping, is not favored under conditions of pulse splitting.

Given this property of the NLSE, it is somewhat surprising that the reportedly first observation of light bullets was in a quadratic medium with normal GVD.

Focusing a beam of Gaussian profile in both space and time near the entrance to a lithium triborate crystal under the condition of a large positive phase mismatch, the pulses propagated without spread temporally or spatially (2D) over a distance ~20 mm (the length of the crystal). However, the pulse did not retain its shape, but evolved into a so-called 'X-wave', or X-shaped soliton. A generic X-shaped soliton (or, strictly speaking, simulton) is shown in Figure 9. The pulse has a conical (clepsydra, or hourglass-shaped) 3D structure with a central hump and slowly decaying conical tails. Interestingly, an X-shaped pulse has been shown to propagate without spreading in a linear optical medium exhibiting normal GVD. But such a pulse must be specially contrived at the input and may be impossible in many cases. In a quadratic nonlinear medium with normal GVD, the X-shaped light bullet appears to form spontaneously. The physics of this evolution lies buried in the coupled wave equations for the fundamental and second harmonic. In the strong phase-mismatched SHG process, weak perturbations in the waves at specific spatial and temporal frequencies can grow exponentially. The hyperbolic nature of the diffraction–dispersion operator with normal GVD leads to amplification of conical wave perturbations with frequencies approaching that of the X-wave. This amplification of the proper frequency components acts as a trigger that drives the evolution of the Gaussian spatiotemporal input pulse into the X-wave. This basic shape has been confirmed in experiments that outline the spatiotemporal intensity profile of the pulse exiting the rear of the crystal.

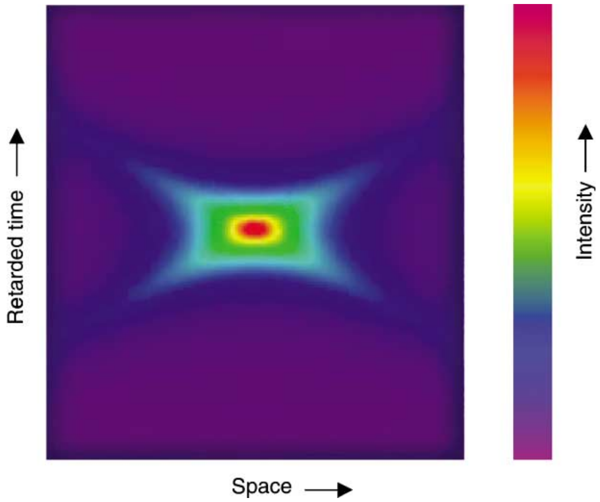

Figure 9 Generic shape of an X-wave light bullet as a function of space and retarded time ($t - z/v_g$) coordinates. The light bullet has radial symmetry.

Other Related Effects (Multiphoton Nonlinearities)

At this point we should note that the susceptibility is frequency dependent. Optical waves couple principally to bound electrons, which can only occupy discrete energy levels. When the photon energy $\hbar\omega$ is nearly resonant with a difference between these energy levels, the dielectric response of the medium is enhanced. This well-known property of linear optics, accounting for absorption of light as well as normal and anomalous dispersion, is also applicable in nonlinear optics. The third-order susceptibility is actually a complex number. A sketch of its real and imaginary parts is given in Figure 10, the real part relating to n_2. The imaginary part is associated with optical loss, if positive, or gain when it is negative. In the present case the frequency dependence of the susceptibility displays a resonance (i.e., peak of the imaginary part) at a frequency $\omega = 2\omega_0$. Referring to Figure 10, when the optical frequency is significantly less than ω_0, the imaginary part of $\chi^{(3)}$ is small compared to the real part and nonlinear refraction dominates. However, when the frequency is comparable to ω_0, the resonance at $2\omega_0$ plays a major role. The real part of $\chi^{(3)}$ is resonantly enhanced, but the imaginary part signifies loss: two photons are simultaneously absorbed. This is called two-photon absorption (2PA). The intensity-dependent absorption loss per unit length is given by $\alpha_2 I$, where α_2 is called the two-photon absorption coefficient (often designated by β) and has physical units of cm W^{-1}.

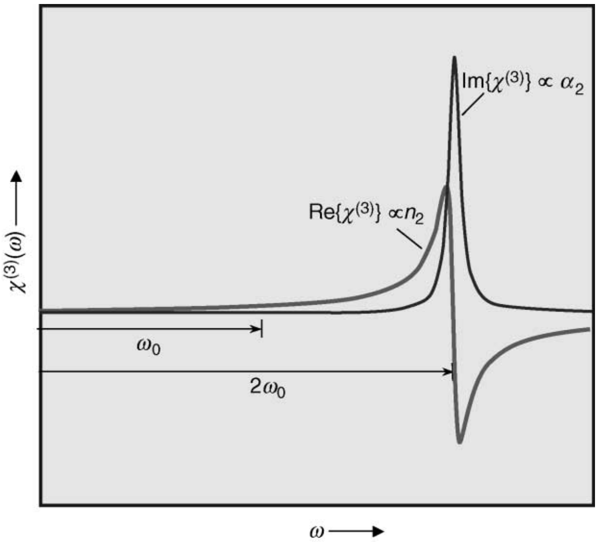

Figure 10 Sketch of the real and imaginary parts of the frequency-dependent third-order susceptibility $\chi^{(3)}$ illustrating a two-photon resonance.

In many materials at a variety of optical frequencies, 2PA will naturally accompany self-focusing. Since these two phenomena are both intensity dependent, while one decreases the intensity and the other increases it, there is an obvious interplay between the two. We illustrate the effect of 2PA in Figure 11 by comparison of the phase retardance of a Gaussian beam (in the parabolic approximation) with and without 2PA in a 'thin' medium, for which the size of an optical beam is nearly constant (diffraction ignored). Obviously, $\Delta\varphi(r)$ does not reproduce the parabolic shape when strong 2PA is present. Actually, it is possible to have multiphoton absorption (MPA) of m photons with corresponding coefficient α_m (related to $\chi^{(m)}$). The effects of 3PA and 4PA are also shown in Figure 11. 2PA, or more generally MPA, distorts the phase and modifies the focusing properties of the beam. This shows up in the Z-scan, for example, as a distortion of the dispersion-like transmittance curve: the valley is deeper and the peak is suppressed.

The reason for this phase distortion is that nonlinear absorption changes the shape of the beam: absorption is stronger where the intensity is higher. For MPA, the intensity at position z in a thin medium is

$$I(r,z) = \frac{I(r,0)}{\left[1+(m-1)\alpha_m I^{m-1}(r,0)z\right]^{1/(m-1)}} \qquad [17]$$

The other obvious effect of MPA is that it reduces the strength of self-focusing since it robs power from the beam. The interplay of the two effects shows up more in thick media where the beam can diffract and change size. When self-focusing is dominant at the front of the medium, the intensity will start to blow up as the beam collapses, but the intensity becomes so large that MPA becomes much stronger. This robs energy from the beam and can stop or reverse the collapse. When MPA is dominant at the start, the beam will still self-focus, but MPA flattens the beam in the center causing the edges of the beam to experience dramatic diffraction effects; for example, a ring structure can develop.

The intensity of a self-focused filament can be quite high. At a level $\sim 10^{11}$ W cm^{-2} or higher, it is possible for MPA to lead to multiphoton ionization (MPI) of the medium. This is illustrated in Figure 12 where the absorption of m photons promotes an electron from the ground state to the continuum in a gas, or from the valence band to the conduction band in a solid. The ionization rate is $\sigma_m I^m$, where m depends on the optical frequency and the ionization potential of the material. For example, at a wavelength of 810 nm, $m = 11$ for nitrogen and $\sigma_{11} = 5.08 \times 10^{-144}$ s^{-1} cm^{22} W^{-11}. We note that for intensities larger than $\sim 10^{14}$ W cm^{-2}, the dominant ionization mechanism is tunneling, where a 'quivering' electron has sufficient energy to tunnel through the coulombic barrier. MPI can create a plasma whose optical properties depend on the density of free electrons. The refractive index of the plasma is nonuniform, and for $\omega > \omega_p = (N_e e^2/\varepsilon_0 m_e)^{1/2}$, the plasma frequency is <1; N_e, e, and m_e are the electron density, charge, and mass, respectively. Hence, plasma formation leads to defocusing. Consequently, self-focusing in air (which has a positive n_2) can lead to plasma formation by MPI, and the resultant defocusing can stabilize the formation of a self-trapped filament of light propagating over a distance of meters. These filaments have been called 'gas-induced solitons'. High peak power femtosecond pulses are typically used to generate these. The pulse width is short compared to the mean

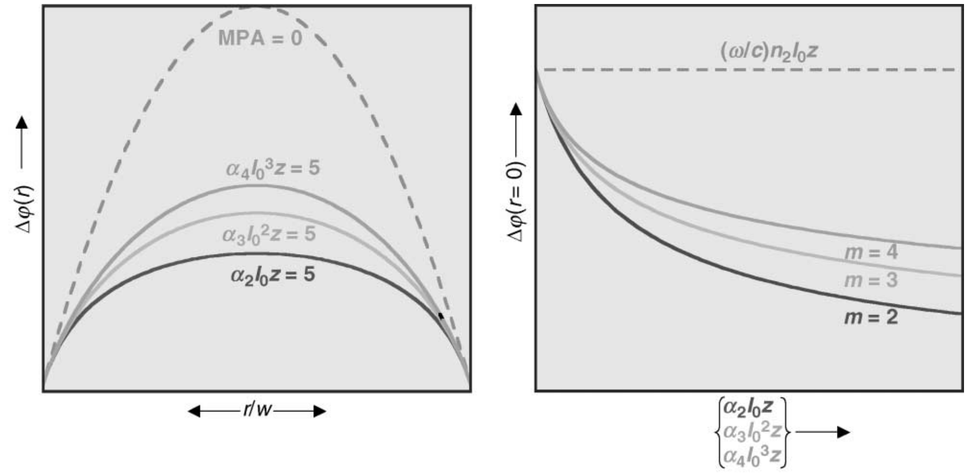

Figure 11 The effects of two-, three-, and four-photon absorption on the intensity-induced phase retardation due to a Kerr self-focusing nonlinearity. The radial dependence of $\Delta\varphi$ on MPA in the parabolic approximation is shown on the left, and the peak of $\Delta\varphi$ as a function of nonlinear absorption loss is plotted on the right.

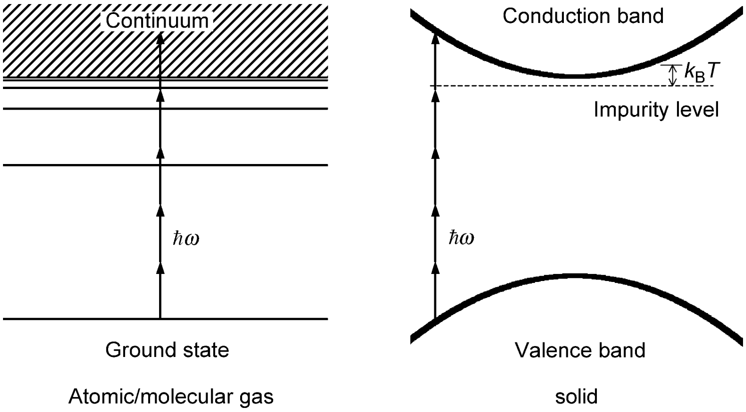

Figure 12 Multiphoton ionization in a gas (left) and a solid (right).

collision time of molecules, and the energy is low so that plasma heating is negligible.

The electron density in plasmas induced by MPI in gases is typically low enough that absorption by free electrons can be ignored. This is generally not the case in solids. MPI can provide enough 'seed' electrons to induce avalanche ionization. As illustrated in Figure 12, impurities may also supply these seed electrons by thermal ionization if they lie within an energy $\sim k_B T$ of the conduction band (k_B is the Boltzmann constant and T is the temperature). Free electrons can absorb electromagnetic energy while colliding with ions or atoms (a process called inverse bremsstrauhlung) and accelerate in a high-intensity field. They strip electrons from other atoms or molecules by impact ionization. The process multiplies free electrons and avalanche ionization ensues. A plasma forms, absorbs more energy and expands, developing a shock wave that can lead to a permanent crack in the solid. This phenomenon is called optical breakdown. Typically observed with nanosecond and picosecond pulses, permanent laser damage of the material also robs energy from the beam and generally terminates the collapse of a self-focused beam. When the electron density is sufficiently high so that $\omega_p > \omega$, the plasma absorbs and scatters light, a process that can be put to use in optical power limiters.

Many additional nonlinear phenomena play important roles when the intensity of a self-focused beam gets this large (e.g., stimulated Raman scattering, stimulated Brillouin scattering, etc.). The reader is referred to the literature on nonlinear optics for more in-depth discussions of these effects.

List of Units and Nomenclature

Critical power	[W]
Electric field	$[\mathrm{V\,m^{-1}}]$
Four-photon absorption coefficient	$[\mathrm{cm^5\,W^{-3}}]$
Intensity	$[\mathrm{W\,cm^{-2}}]$
Multiphoton (m-photon) absorption coefficient	$[\mathrm{cm}^{2m-3}\,\mathrm{W}^{-(m-1)}]$
Multiphoton (m-photon) ionization rate coefficient	$[\mathrm{s}^{-1}\,\mathrm{cm}^{2m}\,\mathrm{W}^{-m}]$
Nonlinear refractive index coefficient	$[\mathrm{cm^2\,W^{-1}}]$
Pockels coefficient	$[\mathrm{m\,V^{-1}}]$
Second-order susceptibility	$[\mathrm{m\,V^{-1}}]$
Third-order susceptibility	$[\mathrm{m^2\,V^{-2}}]$
Three-photon absorption coefficient	$[\mathrm{cm^3\,W^{-2}}]$
Two-photon absorption coefficient	$[\mathrm{cm\,W^{-1}}]$
Vacuum permittivity	$[\mathrm{C\,V^{-1}\,m^{-1}}]$

See also

Laser-induced Damage of Optical Materials. Materials Characterization Techniques: $\chi^{(3)}$. **Nonlinear Optics, Applications:** Pulse Compression via Nonlinear Optics. **Nonlinear Optics, Basics:** Cascading; Kramers–Krönig Relations in Nonlinear Optics; Nomenclature and Units; Photorefraction; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(2)}$–Harmonic Generation. **Photonic Crystals**: Nonlinear Optics in Photonic Crystal Fibers. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. New York: Cambridge University Press.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.
Kivshar YS and Stegeman GI (2002) Spatial optical solitons. *Optics and Photonics News* 13: 59–63.
Marburger JH (1975) Self-focusing: theory. *Progress in Quantum Electronics* 4: 35–110.
Sheik-bahae M, Said AA, Wei TH, Hagan DJ and Van Stryland EW (1989) Sensitive measurement of optical nonlinearities using a single beam. *IEEE Journal of Quantum Electronics* 26: 760–769.
Sheldon SJ, Knight LV and Thorne JM (1982) Laser-induced thermal lens effect: a new theoretical model. *Applied Optics* 21: 1663–1669.
Shen Y (1984) *The Principles of Nonlinear Optics*. New York: Wiley.
Silberberg Y (1990) Collapse of optical pulses. *Optics Letters* 15: 1282–1284.
Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.
Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.
Trillo S and Torruellas WE (eds) (2001) *Spatial Solitons*. Berlin: Springer-Verlag.
Wise F and Di Trapani P (2002) The hunt for light bullets – spatiotemporal solitons. *Optics and Photonics News* 13: 28–32.

Three-Dimensional Microfabrication

S M Kuebler, University of Central Florida, Orlando, FL, USA

M Rumi, Georgia Institute of Technology, Atlanta, GA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear optical three-dimensional microfabrication (3DM) is a photolithographic technique that enables topologically complex 3D microstructures with feature size as small as 1 μm or less to be generated in a single exposure step by nonlinear photopatterning in a material. In the broadest context, 3DM encompasses schemes based on single-beam serial patterning and parallel patterning achieved by multiple-beam interference exposure. In both cases, the photopatterning is initiated by coherent multiphoton excitation (MPE). The material may be a glass, a polymerizable resin, or even a heterogeneous composite, such as a resin containing dispersed nano-particles. The limiting resolution, or the dimensions of the smallest feature that can be generated, is determined by the optical parameters of the excitation geometry and the physical and chemical response of the material to MPE.

3DM offers great promise as a tool for generating complex microdevices, such as micro-electromechanical systems (MEMS), microfluidics, and micro-optical components. New frontiers in the technology of microdevices require that complex 3D structures can be produced in one or more materials that are optimized for a target application in terms of physical, chemical, and mechanical properties. Rapid and versatile fabrication is also essential. Extensive work has been done with conventional photolithography or electron-beam lithography, and more recently using soft-lithography approaches, such as microcontact printing. 3DM provides a new set of tools that overcomes some of the limitations inherent to other methods.

Single-Beam Three-Dimensional Microfabrication

3D Confinement

The simplest form of 3DM involves patterning a microstructure within a material using a single tightly focused laser beam that can activate the medium or one of its constituents by two-photon excitation (TPE) (Figure 1). The TPE initiates a photochemical or photophysical process (e.g., a polymerization reaction or a phase change) within the focal volume and results in an irreversible change in the material. The pattern of the target structure is impressed point by point as the focus is translated within the volume of the material. For some applications the final 3D microstructure is obtained when the unexposed material surrounding the photopatterned regions is removed. In the first report of this type of 3DM in 1990 two-photon-induced polymerization (TPIP) of an acrylate resin was used to fabricate free-standing 3D micron-scale letters on a substrate (Figure 2). The unexposed material does not have to be removed in a post-exposure process when the target structure consists of the phototransformed material supported within the matrix of the unexposed media, as in the case of 3DM of waveguide circuits embedded within a host slab.

3DM is possible because, under tight focusing, TPE and the subsequent material transformation are

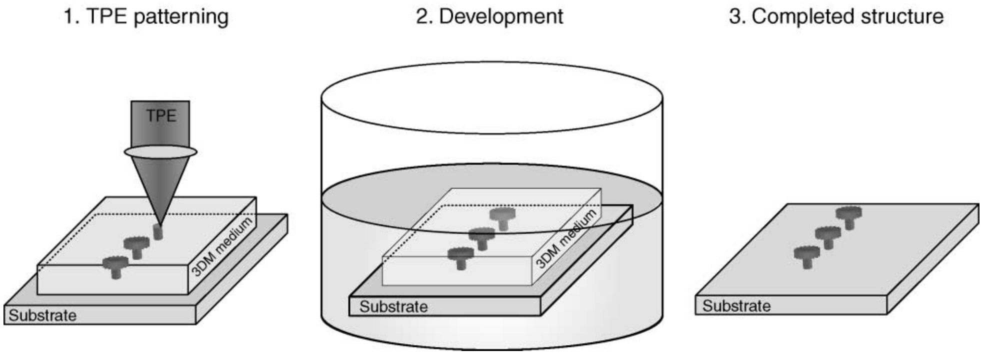

Figure 1 General scheme for single-beam 3DM based on TPE. (1. TPE patterning): The 3DM medium is supported on a substrate and translated in 3D-space relative to a tightly focused laser beam. The laser transforms the material only in the region around the focal point through a photochemical or photophysical process initiated by TPE. (2. Development): In cases where the final structure is needed free from the unexposed material, the sample is treated with a 'developer' that removes the unexposed material, for example by dissolving it in a solvent. (3. Completed structure): Following removal of the unexposed material, the final 3D structure is obtained free-standing on the substrate.

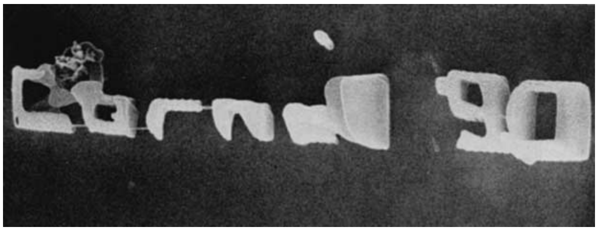

Figure 2 Scanning electron micrograph of a structure obtained by TPIP-3DM in an acrylate-ester resin. The excitation wavelength was 630 nm. The voxel dwell time was 10 ms along the letters. The thin filaments connecting some of the letters are approximately 100 nm wide. Reproduced from Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorscence microscopy. *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 1398: 107–118. Copyright (1990), with permission of SPIE.

confined at the focus within a volume of $\sim(\lambda/n_0)^3$, where λ and n_0 are the vacuum wavelength of the exciting radiation and the refractive index of the material, respectively. Degenerate TPE is a resonant third-order nonlinear optical process in which a species is promoted to an excited electronic state by the simultaneous absorption of two photons of equal energy. Relative to conventional one-photon excitation (OPE), TPE is typically achieved using longer-wavelength radiation (often red or near infrared), for which the combined energy of two photons is sufficient to promote the species into one of its lower lying excited electronic states (Figure 3, left). The rate of TPE, that is the number of species excited per unit volume and time, $R_{2h\nu}$, is proportional to the square of the intensity, I. For a focused Gaussian beam, the on-axis intensity attains its maximum, I_0, at the focus and decreases with distance from the focal plane, z, approximately as $I \propto z^{-2}$. For a focused Gaussian beam, the on-axis intensity is given by $I = I_0/[1 + (z/z_R)^2]$, where z_R is the Rayleigh range of the focused wavefront. Thus, $I \propto z^{-2}$ in the limit of $z \gg z_R$. Consequently, the excitation rate decreases as $R_{2h\nu} \propto z^{-4}$. Similarly, the excitation rate decreases rapidly away from the propagation axis. This results in a region of high excitation, and thus material transformation, that is tightly confined both laterally and longitudinally in space about the focal point (Figure 3, right). This situation may be contrasted with that for OPE, for which the excitation rate, $R_{1h\nu}$, is linearly proportional to I, so $R_{1h\nu} \propto z^{-2}$. The weaker dependence of $R_{1h\nu}$ on z results in material being excited appreciably throughout the irradiation volume, with less longitudinal confinement at the focus than is achieved under TPE. Similar arguments can be used to describe 3DM involving higher-order excitation, for which the excitation is confined to an even smaller volume for a given λ.

The rate for molecular n-photon excitation, $R_{nh\nu}$, represents the average number of molecules that are excited per unit volume per unit time and is given by

$$R_{nh\nu} = \frac{1}{n(h\nu)^n}\sigma^{(n)} N I^n \qquad [1]$$

where n is the order of the excitation process, h is Planck's constant, ν is the frequency of the exciting radiation, $\sigma^{(n)}$ is a molecular n-photon absorption cross-section, which has units of $[\mathrm{cm}^{2n}\,\mathrm{s}^{n-1}\,\mathrm{photon}^{-(n-1)}]$, N is the number density of the photo-excitable species $[\mathrm{cm}^{-3}]$, and I is the intensity $[\mathrm{W\,cm}^{-2}]$. The factor of $1/n$ accounts for the fact that n photons must be absorbed to promote a single molecule to an excited state. Note that $\sigma^{(n)}$ is a

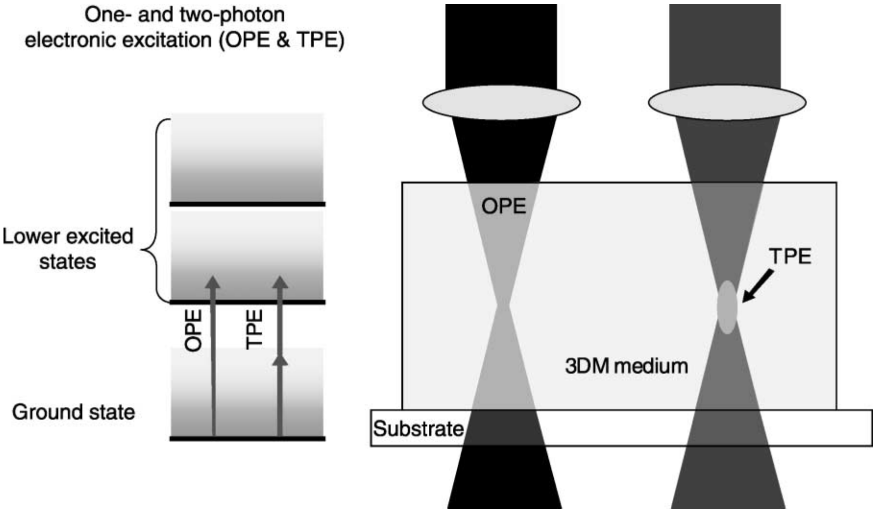

Figure 3 (Left) Jablonski representation of electronic one- and two-photon excitation (OPE and TPE, respectively). (Right) Illustration of a 3DM medium undergoing OPE and TPE using focused short- (black) and long-wavelength (dark gray) radiation, respectively. The regions that are appreciably excited are shown in light gray. Under OPE the medium is excited throughout the length of the interaction volume. Under TPE the excitation is confined in three dimensions to a small region around the focal point.

molecular parameter that is independent of the intensity. The molecular cross-sections for OPE and TPE, $\sigma^{(1)}$ and $\sigma^{(2)}$, are respectively equivalent to the parameters σ and δ that are commonly encountered in the literature.

For MPE to occur, the intensity must be sufficiently large that there is a high probability of more than one photon arriving simultaneously at the species to be excited. Typical magnitudes of the OPE and TPE cross-sections are $\sigma^{(1)} \approx 10^{-17}\,\mathrm{cm}^2$ and $\sigma^{(2)} \leq 10^{-47}\,\mathrm{cm}^4\,\mathrm{s\,photon}^{-1}$, and under conventional excitation conditions, $R_{2h\nu} \ll R_{1h\nu}$. To achieve $R2h\nu \approx R1h\nu$, $I \approx 2h\nu\sigma^{(1)}/\sigma^{(2)}$ is required. For the cross-sections given above, this condition corresponds to $I \approx 500\,\mathrm{GW\,cm}^{-2}$ at $\lambda = 800\,\mathrm{nm}$. Such peak powers can be achieved by focusing ultra-short laser pulses, so mode-locked femtosecond- or picosecond-pulse lasers are generally used for 3DM. Tight focusing in 3DM is desirable then, not only to provide a means for 3D confinement, but also to increase the intensity to a level that $R_{nh\nu}$ becomes large enough to produce a significant material change.

Opto-Mechanical Implementation

One method for implementing single-beam 3DM is illustrated in Figure 4. An excitation laser beam is directed onto a high numerical aperture (NA) objective. A beamsplitter, detector, attenuator, and shutter are placed upstream of the objective to measure and control the average power at the sample, $\langle P \rangle$. A substrate is coated with the photo-active medium and attached to a 3-axis nano/micro-positioner under the objective. The interior of the photo-active material is patterned during the exposure by translating the sample relative to the focus of the laser beam and shuttering the beam as needed. The process is automated as a computer controls the sample translation coordinates and the exposure conditions for each volume element (voxel). This configuration is well suited for patterning solid or semi-solid photo-media. An index-matching fluid must be used with high-NA objectives to achieve the smallest focal spot size. If the photo-medium is a liquid, a coverslip or other thin transparent barrier must be placed between the index matching fluid and the photomedium so that the latter does not become contaminated.

Alternatively, an inverted configuration may be used. Here, the laser is focused through the substrate into the medium. The objective and the index matching fluid are kept out of contact with the medium by the substrate. In this case, the maximum depth that can be patterned is limited by at least two conditions: (i) The working distance of high-NA objectives is small, often no more than 200 μm, so the substrate must be kept thin (~100 μm); and (ii) unless the viscosity of the medium is very high, patterning must begin at the medium/substrate interface and proceed layer by layer away from it, so that the structure remains anchored to the substrate. This could be disadvantageous, as the change in refractive index in the previously patterned layers could alter the intensity distribution

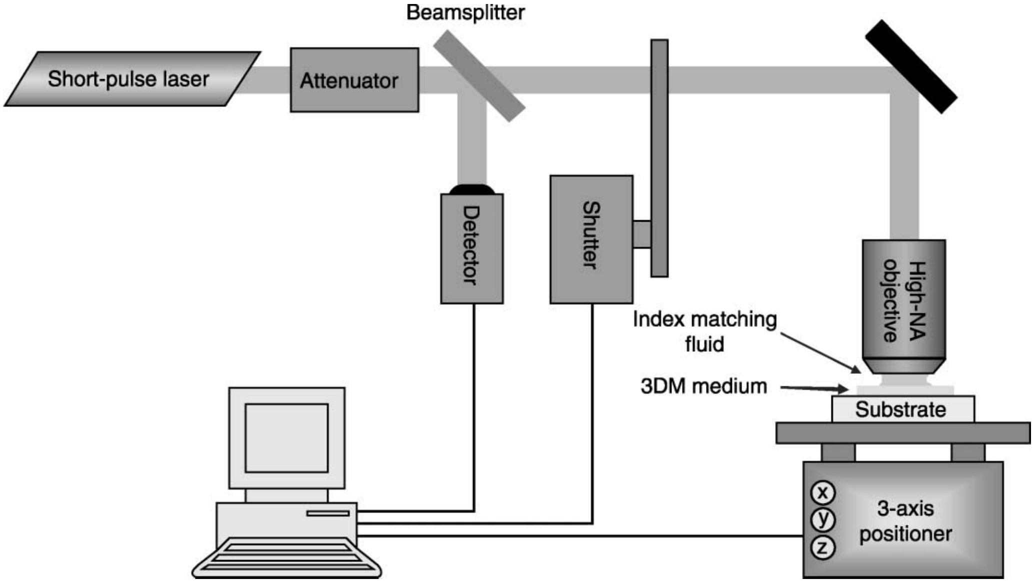

Figure 4 Optomechanical setup for single-beam 3DM.

at the focus. In all configurations, it is best to minimize the number of optical interfaces to maintain the quality of the focus.

A scanning confocal microscope system provides a convenient platform for single-beam 3DM. The nano/micro-positioner can be mounted directly on the sample stage of the microscope, and the optical components of the confocal system can be used directly for 3DM. The confocal system also enables the medium and the structure to be imaged *in situ*, using confocal reflectance or fluorescence imaging. If the material is fluorescent under MPE, a 3D image can be obtained by scanning multiphoton fluorescence microscopy. These imaging modes facilitate 3DM by providing a means for locating the focus at an absolute position within the photo-active medium.

3DM can also be realized by moving the focused beam relative to a fixed sample. One convenient implementation utilizes galvanometer-driven mirrors to scan the beam in planes parallel to the substrate (xy-plane). Motion in the z-direction is accomplished by raising and lowering the objective relative to the substrate. A galvo-based system typically enables faster fabrication, but may not afford motion as precise as that obtained using a multi-axis nanopositioner.

Excitation Sources

The most commonly used and arguably the most convenient and reliable laser for 3DM is the continuous-wave (CW) mode-locked titanium sapphire (Ti:S) laser. Ti:S lasers routinely produce sub-100 fs pulses at high repetition rates (~80 MHz) with single-pulse energies, E_p, of ~1 nJ. The Ti:S laser emission spans the range of 700–1050 nm, making it particularly well suited for two- and three-photon excitation. $\langle P \rangle$ can be as high as several watts at the peak of the gain curve ($\lambda \approx 800$ nm) with high pulse-to-pulse and time-averaged stability (RMS $\approx 1\%$). Such systems may also produce a nearly Gaussian TEM_{00} mode, which is helpful in achieving the tightest focus in the material. Note that focusing a 1 nJ, 100 fs pulse to a 1 μm-radius spot generates a focused average peak intensity of 320 GW cm^{-2}, as is needed to achieve photo-excitation rates of $R_{2h\nu} \approx R_{1h\nu}$.

For higher peak powers, amplified femtosecond (AFS) lasers can be used for 3DM. These systems can produce ~100 fs pulses with $E_p \approx 1$ mJ or higher at repetition rates of 1–100 kHz. Given that a continuous microfabricated structure should be generated from partly overlapping voxels, a low repetition rate reduces the speed with which the laser beam may be scanned within the material. An AFS laser can be used to pump an optical parametric generator to extend the output wavelength range.

Recently, some researchers have turned their attention toward identifying compact solid-state turn-key lasers for 3DM that could be more convenient and cheaper alternatives to the complex systems described above. A commercial Nd:YAG microlaser ($\lambda = 1064$ nm, $E_p = 5.4$ μJ, $\tau_p = 640$ ps, 13.6 kHz) and a mode-locked erbium-doped fiber laser ($\lambda = 780$ nm, $E_p \approx 0.5$ nJ, $\tau_p = 100$ fs, 50 MHz) have both been shown to be satisfactory for 3DM in certain acrylate-based media.

Material Systems

Structural Materials

The most widely used materials for 3DM are based on acrylates that are patterned by multiphoton induced polymerization. The process begins when an initiator chromophore **C** is excited by n-photon absorption to $\mathbf{C}^*$, with an efficiency specified by the n-photon cross-section $\sigma^{(n)}$. $\mathbf{C}^*$ then generates an initiating radical **I·**. Acrylate monomers (**M**) can react with **I·** and undergo a free-radical chain reaction that leads to high molecular weight polymer. With sufficient exposure, the starting material is converted to less soluble high-molecular weight polymer chains or densely cross-linked polymer networks. The photo-patterned structure can then be 'developed' by immersing the sample into a solvent that removes the unexposed material, leaving behind a free-standing structure that is a replica of the photopattern.

Photo-polymerization starts when the MPE exposure generates enough **I·** to react with and consume all of the inhibitor in the local volume at point $\vec{r}$ within a homogeneous medium. In this discussion we assume that the local concentration of inhibitors is not changed by diffusion on the exposure timescale. The total number of **I·** generated during an exposure at $\vec{r}$ per unit volume, $N_{\mathrm{I}\cdot}(\vec{r},t)$, is obtained from the time-integrated MPE rate:

$$N_{\mathrm{I}\cdot}(\vec{r},t) = \frac{\phi_{\mathrm{c}}^{(n)}\sigma^{(n)}N_{\mathrm{C}}}{n(h\nu)^n}\int_{-\infty}^{t} I(\vec{r},t')^n \mathrm{d}t' \qquad [2]$$

Here, N_{C} is the number density of **C**, and $\phi_{\mathrm{c}}^{(n)}$ is the n-photon chemical quantum yield, defined as the number of **I·** generated divided by the number of $\mathbf{C}^*$ present after excitation. If we identify the initial number density of the inhibitors as N_{Inh}, then polymerization occurs when $N_{\mathrm{I}\cdot}(\vec{r},t) > N_{\mathrm{Inh}}$. In multiphoton-induced polymerization, $N_{\mathrm{I}\cdot}(\vec{r},t)$ is a nonlinear function of I, which generally involves a rapidly varying time envelope, as from a mode-locked ultrashort-pulse laser.

Several different commercial acrylate systems have been used successfully for 3DM, such as Nopcocure 800 (San Nopco), NOA 72 (urethane acrylate oligomer, Norland Products), and SCR 500 (blend of urethane acrylate monomers and oligomers, Japan Synthetic Rubber Co.). Researchers have also successfully developed custom formulations using blends of commercially available monomers, oligomers, and polymer additives to create resins with tailored physical, chemical, and optical properties. This aspect of acrylates, and more generally polymer-based systems, makes them one of the most attractive classes of materials for 3DM. Conventional one-photon UV- or blue-sensitive radical generators (e.g., benzil, benzoin methyl ether, 4,4′-bis(*N*,*N*′-dimethylamino)benzophenone, to name just a few) have most often been used, and these were activated by promoting the molecule into one of its low-lying electronic states by MPE over the range of 730–800 nm.

3DM based on epoxide polymerization has also been demonstrated using commercial resins including cross-linkable small molecular weight monomers (e.g., 4-vinyl-1-cyclohexene diepoxide) and epoxide oligomers (e.g., SU-8, first developed by IBM). Epoxide polymerization is most commonly initiated by molecules that, after excitation, generate a Brønsted acid, H^+. The propagating species is a carbocation, and bases present in the medium act as inhibitors. Relative to acrylates, epoxides shrink substantially less upon polymerization, leading to less distortion of the microstructure after development.

Acrylate chemistry has also been exploited for 3DM of functional composite materials. Acrylamide resin mixtures were used to fabricate bi-layer micro-cantilevers that deflect from the surface when illuminated with UV light. This type of microstructure could form the basis of a photo-activatable MEMS. 3D gratings were fabricated from an acrylate pre-polymer containing a dendrimer functionalized with cross-linkable acrylic *L*-phenylalanine groups. The dendrimer acts as a host encapsulation site that enables a laser dye to be loaded in the material at high weight-percent (see below). A variety of complex micro-structures have also been fabricated using 'Ormocer-I', an ORganically MOdified CERamic. Ormocer-I is an inorganic-organic hybrid polymer consisting of a Si–O–Si (siloxane) backbone that is functionalized with organic moieties, including cross-linkable methacryloxypropyl groups. Ormocers are extremely promising for 3DM because of the thermal stability and chemical inertness of the backbone and their attractive physical and optical properties.

3DM could be a powerful tool for fabricating biologically active structures and devices. In this context, hydrogels and Ormocers should be useful material systems as some formulations are known to be bio-compatible. Microstructures can also be fabricated from proteinaceous material. 3D microstructures were produced by patterned MPE of solutions containing the dye rose bengal and the proteins bovine serum albumin or fibrinogen.

The material systems discussed thus far are 'negative-tone' media. This means that only the exposed regions of the material remain after the post-exposure development, so the final structure is a replica of the exposure pattern. In contrast, 'positive-tone' media are solid-state systems for which the exposed regions become soluble in a developer. The final structure is then the inverse of

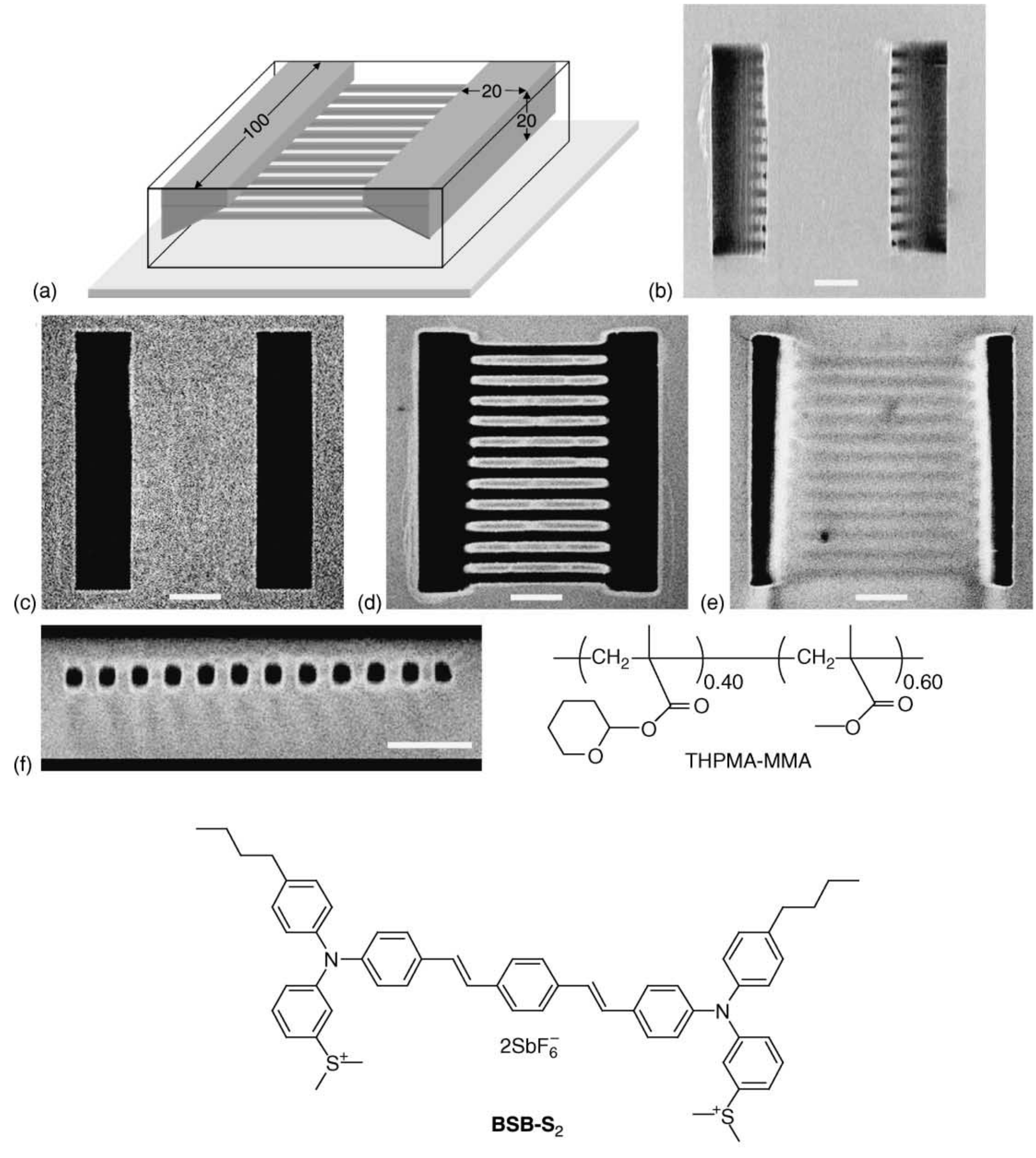

$\sigma^{(2)} = 690 \times 10^{-50}$ cm^4 s photon^{-1} at 705 nm
$\phi_{H+} = 0.50 \pm 0.05$

Figure 5 A 3D micro-channel structure fabricated by TPE-3DM using the positive-tone chemically amplified resist THPMA-MMA containing the two-photon PAG **BSB-S$_2$**. A 50 μm-thick film of the resist was exposed in the pattern of the target structure at 745 nm with tightly focused 80 fs pulses at $\langle P \rangle = 40$ μW and a linear scan speed of 50 μm s^{-1}. (a) Target structure consisting of two rectangular cavities (width: 100 μm, length: 20 μm, depth: 20 μm) with a sloped side-wall, that are connected by 12 channels (length: 50 μm; 4 μm by 4 μm cross-section) lying 10 μm below the surface and spaced apart by 8 μm (center-to-center). (b) Scanning electron micrograph of the final structure, viewed normal to the substrate. (c to e) Two-photon-fluorescence images of the final structure (viewed normal to the substrate): (c) at the surface of the film; (d) 10 μm below the surface; (e) 19 μm below the surface. (f) Two-photon-fluorescence cross-sectional image of the buried channels. The scale bar in b to f corresponds to 20 μm. Reproduced from Zhou W, *et al.* (2001) *Science* 296: 1106–1109. Copyright (2001), with permission of The American Association for the Advancement of Science.

the exposure pattern. Positive-tone 3DM was first demonstrated as early as 1992 with the generation of high aspect ratio trenches in thin layers of commercial photoresists. A positive-tone material system was designed for 3DM (Figure 5), is based on a chemically amplified resist and a high-sensitivity two-photon-activatable photo-acid generator (PAG) (discussed below). Using this medium, it was possible to fabricate subsurface microchannel structures and microgratings, using moderate laser scan speeds (50 μm s^{-1}) and exceptionally low $\langle P \rangle$ (40 μW).

Composites have been formulated that enable 3D-patterning of nanoscale metal features within a host matrix. A two-step process has been described that involves MPE of a Ag^+-containing sol–gel followed by thermal development of the latent image with a $AgClO_4$ solution. Using this system, 3D spirals of silver nano-particles embedded within the sol–gel matrix were generated. A $HAuCl_4$-doped poly(vinyl alcohol) composite was used to 3D-pattern gold nano-particles. It was shown that after TPE a particular sensitizer dye hosted in a poly(vinylcarbazole) matrix containing $AgBF_4$ and ligand-stabilized Ag nanoparticles, electron transfer from the dye to Ag^+ leads to nucleation of Ag^0 at the nanoparticles and the formation of conducting bulk metal features (Figure 6). Material systems like these should find wide application for generating micro- and nano-scale structures, having unusual and useful optical and electronic properties.

3DM in glasses is of great interest for photonic applications. Several groups have demonstrated that tightly focused infrared or visible amplified ultrashort laser pulses can be used to 3D-pattern pure and doped silica and other glasses. The patterned voxels can be generated as a result of photo-induced changes in the refractive index or extreme dielectric breakdown, resulting in pits within the material. The mechanisms for these two cases are not fully understood, but both are thought to be activated through an initial MPE event. The leading edge of the laser pulse generates a free-carrier plasma through MPE. The remainder of the pulse is increasingly absorbed by the plasma leading to avalanche ionization. The optical energy deposited into the local lattice results in cavitation, creating a pit at the focal point. Based on the electronic bandgap of silica, the process likely involves three- and six-photon absorption when 400- and 800-nm light is used, respectively. This approach has been used for the 3D-patterning fused silica and germanium doped glasses.

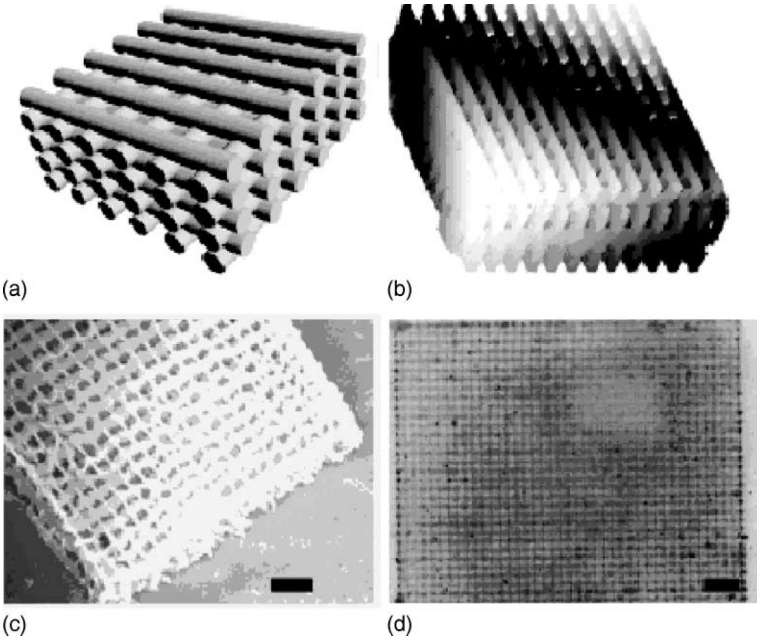

Figure 6 Metallic structures fabricated by TPE-3DM using a silver nanoparticle/polymer composite. (a) Model of the ‘stack-of-logs’ target structure. (b) 3D image constructed from a series of two-photon fluorescence microscopy images of the silver structure immediately following 3DM. (c) Scanning electron micrograph of the free-standing silver structure following removal of the surrounding unexposed composite. The scale bar corresponds to 10 μm. (d) Transmission optical microscopy image of the silver structure immediately following 3DM. The scale bar corresponds to 25 μm. Reproduced from Stellacci F, *et al.* (2001) *Advanced Materials* 14: 194–198. Copyright (2001), with permission of Wiley.

Multiphoton Initiators

For 3DM to develop into a widely used fabrication tool, material systems and patterning methods must be developed that enable rapid microfabrication at low cost. 'Soft materials', such as the polymer systems described above, are promising in this respect, as both the initiating species and the host resin can be engineered to optimize the multiphoton response and the physical and chemical properties of the patterned medium. It may also prove possible to create doped glasses that have physical properties close to that of neat silica and yet offer higher multiphoton sensitivity.

To develop this concept, let us consider 3DM in acrylates further. A high-sensitivity acrylate medium is one for which: (i) **C** strongly absorbs *n*-photons (large $\sigma^{(n)}$); (ii) $\mathbf{C}^*$ efficiently generates the initiating species **I·** (large $\phi_c^{(n)}$); and (iii) **I·** is an efficient initiator (high initiation rate). As criteria (i) and (ii) are predominantly determined by the molecular properties of **C**, it should be possible to engineer soft materials for 3DM by optimizing multiphoton initiators and incorporating these into pre-polymer resins that once patterned have the correct properties for the target application. The sensitivity of materials can be compared on the basis of the minimum average laser power, $\langle P \rangle_{\min}$, that is required to pattern a material using a fixed scan speed (or exposure time). Alternatively, we can use the product $\sigma(n)\phi c(n)$ as a multiphoton initiator figure-of-merit (FOM) that accounts for the strength of the multiphoton absorption and the efficiency with which **I·** is generated.

The chromophores that have been used most frequently for polymerization-based 3DM do efficiently generate highly active initiating species once excited; however, they were developed for one-photon response to UV and blue radiation and generally have low multiphoton FOMs. The use of xanthene dyes was explored as initiators for free-radical polymerization of acrylamides and it was found that the multiphoton sensitivity was low. As a result, 3DM using conventional initiators requires high laser power (typically $\langle P \rangle_{\min} \approx 10-100$ mW) and slow laser scan speeds ($\sim 50\ \mu\text{m s}^{-1}$). Under tight-focusing conditions, the resulting peak powers are often near the damage threshold of the materials, rendering the process unreliable from a manufacturing stand-point.

A series of **D–π–D** molecules was examined (Figure 7), where **D** is an electron-donating group and **π** is a conjugated π-electron bridge. This class of **D–π–D** chromophores was found to have large TPE cross-sections ($\sigma^{(2)} > 200 \times 10^{-50}\ \text{cm}^4\ \text{s photon}^{-1}$) and to initiate the polymerization of acrylates following TPE in the visible and near infrared, with $\langle P \rangle_{\min}$ a factor of 30–50 times lower than for conventional one-photon initiators. The sensitivity achieved with **D–π–D** initiators is so enhanced that TPIP-3DM was performed at laser scan-speeds as high as $9\ \text{cm s}^{-1}$, using $\langle P \rangle = 3$ mW, enabling a volume of $180\ \mu\text{m} \times 180\ \mu\text{m} \times 20\ \mu\text{m}$ to be polymerized in 200 s.

D–π–A molecules have also been examined for TPIP (Figure 7), where **A** is an electron accepting functionality. Such molecules can have large TPE cross-sections, and some have been shown to activate TPIP efficiently in acrylates at $\lambda \approx 800$ nm. A derivative of Michler's ketone was synthesized with an extended π-conjugated system (Figure 7). This molecule exhibited strong two-photon absorption over the broad range of 800–1100 nm and was also shown to be active toward acrylate polymerization at 1064 nm. A strategy was implemented for enhancing two-photon sensitivity based on up-converted fluorescence and used for 3DM of waveguide structures. The strongly two-photon-absorbing dye AF-183 was incorporated into a commercial acrylate photoresist containing a radical-generator that is activated by OPE at 315–450 nm. After AF-183 undergoes TPE at 800 nm, it fluoresces efficiently at ~465 nm. The up-converted emission then excites the one-photon initiator and radicals are generated.

The strategies discussed above are oriented toward enhancing multiphoton sensitivity by increasing $\sigma^{(2)}$ of the initiating chromophore. Little attention has been paid to the need of ensuring that the MPE couples to a mechanism for efficiently generating an active initiating species. This problem was addressed with the design of a two-photon PAG by covalently linking a **D–π–D** chromophore and sulfonium moieties, which can generate H^+ when reduced by photo-induced electron transfer (**BSB-S$_2$** in Figure 5). **BSB-S$_2$** features strong two-photon absorption and its quantum yield for the generation of H^+ per photo-excited molecule (ϕ_{H^+}) exceeds that of some commercial one-photon PAGs. Based on $\langle P \rangle_{\min}$, **BSB-S$_2$** was found to be 10 to 100 times more sensitive toward multiphoton epoxide polymerization than one-photon UV/blue-sensitive PAGs.

It should be understood that this discussion of multiphoton initiators is not comprehensive of all work that has been done in the field. Rather, it is intended as a highlight of the many outstanding examples of progress in this area. The use of inexpensive low-power lasers for 3DM should become more practical as high-sensitivity

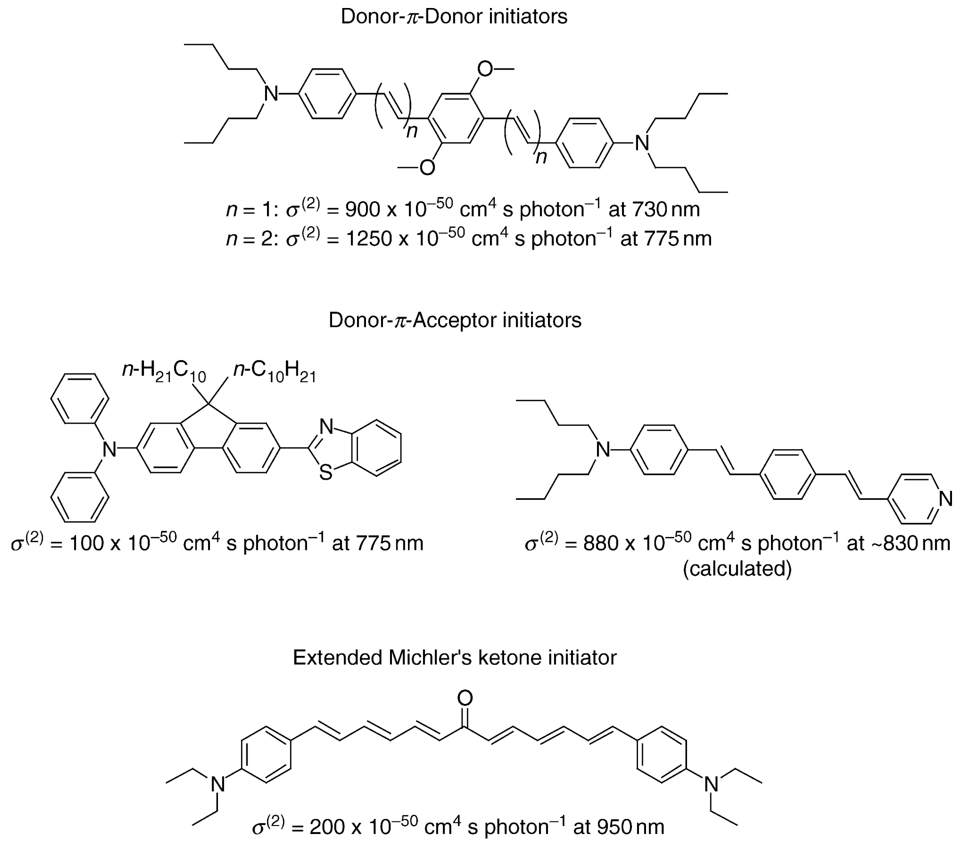
Donor-π-Donor initiators
N
O
O
N
n
n
n = 1: σ(2) = 900 x 10−50 cm4 s photon−1 at 730 nm
n = 2: σ(2) = 1250 x 10−50 cm4 s photon−1 at 775 nm
Donor-π-Acceptor initiators
n-H21C10
n-C10H21
N
N
S
σ(2) = 100 x 10−50 cm4 s photon−1 at 775 nm
N
N
σ(2) = 880 x 10−50 cm4 s photon−1 at ~830 nm (calculated)
Extended Michler's ketone initiator
O
N
N
σ(2) = 200 x 10−50 cm4 s photon−1 at 950 nm

Figure 7 Examples of high-sensitivity initiators for 3DM based on TPIP of acrylates.

3DM media are developed. This should, in turn, facilitate the widespread use of 3DM as a microfabrication tool.

Resolution

The resolution of single-beam 3DM is determined by the size of the smallest voxel that can be reliably produced. Both the optical parameters and the material response under MPE affect the size of the voxel. Given the wide range of physical and chemical mechanisms that can be involved, no single set of guidelines can be outlined regarding the effect of the material response on voxel size. Additionally, this aspect of 3DM remains relatively unexplored. However, as 3DM based on TPIP of acrylates has been examined in some detail, an overview of this material response provides some valuable insight into the overall problem of resolution. The following simplified discussion is based on the local response of the material to photo-excitation.

TPIP initiated in the vicinity of the focus could, in principle, propagate throughout the entire volume of the medium, were it not for termination by inhibitors. Let us consider further the expression for $N_{\mathrm{I}}.(\vec{r},t)$, expanding the intensity as:

$$I(z,r,t') = i(z,r)f(t') \qquad [3]$$

Here, $f(t')$ is a time-dependent amplitude function, $i(z,r)$ is the peak intensity, and the position vector $\vec{r}$ is expressed in cylindrical coordinates (z,r) with the origin at the focus. As before, z denotes the direction in which the beam propagates (longitudinal). The radial coordinate orthogonal to the z-axis (lateral) is denoted by r. The beam is taken to have cylindrical symmetry, so I does not depend on the angular coordinate. For a fixed exposure time $t \rightarrow \tau_{\mathrm{ex}}$, we can identify a polymerization-threshold peak intensity, I_{th}, for which $i(z,r) > I_{\mathrm{th}}$ results in $N_{\mathrm{I}}.(z,r,\tau_{\mathrm{ex}}) > N_{\mathrm{Inh}}$ and polymerization occurs. The polymerized voxel consists then of all points within the volume centered at the focal point and bounded by the I_{th} isophote.

The shape of the I_{th} isophote depends upon the optical parameters of the system. The tightest focusing is achieved when the highest NA objective is used to focus parallel incident rays that overfill the entrance aperture. This situation generates a diffraction limited intensity pattern about the focal point. Within the focal plane the lateral intensity distribution follows the Airy pattern:

$$i(0, r) = [2J_1(\rho)/\rho]^2 I_0 \qquad [4]$$

where J_1 is the first-order Bessel function:

$$\rho = 2\pi r \mathrm{NA}/\lambda \qquad [5]$$

and I_0 is the peak intensity at the origin. Near the focus, the longitudinal intensity distribution is given by

$$i(z, 0) = [\sin(\zeta/4)/(\zeta/4)]^2 I_0 \qquad [6]$$

where

$$\zeta = 2\pi \mathrm{NA}^2 z/(n_0 \lambda) \qquad [7]$$

Let us define the longitudinal and lateral widths of the focal spot, Z and R respectively, to be twice the distance along the corresponding z- and r-axes over which the peak intensity of the central bright fringe decreases to $I_0/2$. If $\langle P \rangle$ is adjusted so that $I_0/2 = I_{\mathrm{th}}$, then to first order we can expect the polymerization voxel to be a solid of rotation centered about the focus having a longitudinal and lateral width of Z and R, respectively. Numerical evaluation shows that the $I_0/2$-points occur at $\rho = \pm 1.6154$ and $\zeta = \pm 5.5661$. For $\lambda = 800$ nm, NA $= 1.4$, and $n_0 \approx 1.5$, we find $R = 0.29$ μm and $Z = 1.1$ μm.

Several groups have demonstrated that voxels generated under conditions similar to those discussed above do indeed have dimensions comparable to those calculated, based on diffraction limited propagation. As shown in Figure 8, voxels with lateral and longitudinal widths below the diffraction limit can be fabricated by steadily reducing τ_{ex} (or I_0) so that an increasingly smaller portion of the central bright fringe exceeds I_{th}. In this way, voxels having $R \approx 100$ nm and $Z < 800$ nm have been generated, which convincingly demonstrates the potential of this technique as a nano-fabrication tool.

Post-exposure material processing can also have a profound impact on the resolution. In the case of TPIP-3DM, material near the boundary of the I_{th} isophote may not have sufficient molecular weight or cross-linking density to resist removal in a development step. In this case, the final voxel will be smaller than the I_{th} isophote defined strictly in terms of a polymerization threshold. Additionally, a solvent-based developer can swell or otherwise distort the voxels. Overall, three important factors determine the final size of the voxel: (i) the exposure and focusing conditions; (ii) the response of the material to photo-excitation; and (iii) the behavior of the photoconverted material during the development process. It should be emphasized that 3D confinement is due to the combined effects of nonlinear photo-excitation and the nonlinearity inherent in the photoconversion and post-exposure development of the material. More advanced treatments have also attempted to address the effects of radical and inhibitor diffusion on the size of the voxel.

Previously, we assumed that the voxel is cylindrically symmetric. The intriguing result of some recent work has shown that the widths of the voxel in the transverse directions can actually differ by as much as 10% when linearly polarized radiation is used because the symmetry of the electric field vectors in the x- and y-directions is broken in high-NA focusing geometries. Cylindrically symmetric voxels can be obtained by using depolarized or circularly polarized radiation. Clearly, study of the combined effects of optical parameters and material response remains an important area for further research in this rapidly advancing field. Although this discussion has centered on TPIP, similar arguments can be developed for 3DM based on other photomaterials and higher-order processes.

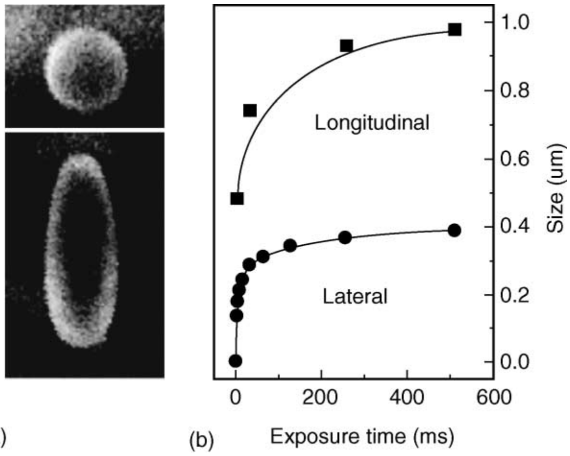

Figure 8 Single voxels produced by TPIP-3DM in an acrylate medium. (a) Scanning electron microscope (SEM) images of a single voxel showing its longitudinal (top) and lateral (bottom) width. (b) Dependence of the longitudinal and lateral widths of voxels on exposure time, as measured from SEM images following fabrication. Reproduced from Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158. Copyright (2003), with permission of The American Institute of Physics.

Multibeam-Interference Three-Dimensional Microfabrication

Multibeam-interference (MBI) 3DM is an alternative approach to 3DM that enables large volumes of material to be simultaneously patterned into a 3D periodic structure. In MBI-3DM, multiple coherent laser beams interfere within a photo-active medium. A periodic structure results because the material undergoes photo-induced changes that occur maximally in the regions of peak intensity across the periodic interference pattern. Depending on the number of beams utilized and their reciprocal orientation, periodic structures in one, two, or three dimensions can be generated, and the periodicity is of the order of λ. This technique is often referred to as holographic lithography or recording.

MBI-3DM has most often been achieved using one-photon-induced polymerization; however, higher-order excitation may be used as well. In this discussion we will focus on multiphoton MBI-3DM, with the understanding that most of the concepts apply equally well to the case of linear photo-excitation. When the power of the incident beams is adjusted so that the intensity at the peaks of the interference pattern just exceeds I_{th}, polymerization only occurs in the local vicinity of these points, whereas the 3DM medium is not appreciably polymerized in regions where the intensity is below I_{th} and may be removed by dissolution in a solvent. The result is a periodic porous structure with highly interconnected parts, that extends over the whole volume where the beams overlapped.

Let us assume that m-beams are directed onto a photosensitive material. The intensity of light at a given point $\vec{r}$ in the region where the beams overlap is proportional to:

$$\begin{aligned} I(\vec{r}) \propto \vec{E}\cdot\vec{E}^{*} &= \left(\sum_{j=1}^{m} \vec{E}_{0j}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j\cdot\vec{r}-\omega t)}\right)\cdot\left(\sum_{l=1}^{m} \vec{E}^{*}_{0l}\, \mathrm{e}^{-\mathrm{i}(\vec{K}_l\cdot\vec{r}-\omega t)}\right) \\ &\propto \sum_{j,l=1}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0l}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j-\vec{K}_l)\cdot\vec{r}} \\ &\propto \sum_{j=1}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0j} + \sum_{j\neq l}^{m} \vec{E}_{0j}\cdot\vec{E}^{*}_{0l}\, \mathrm{e}^{\mathrm{i}(\vec{K}_j-\vec{K}_l)\cdot\vec{r}} \end{aligned} \qquad [8]$$

where $\vec{E}_{0j}$ and $\vec{K}_j$ are the electric field and wavevector of beam j, and ω is its angular frequency ($\omega = 2\pi\nu$). Equation [8] indicates that the intensity distribution depends on the difference between pairs of wavevectors of the interfering beams, $\vec{K}_j - \vec{K}_l$. This distribution is periodic in space and can be described as a crystal lattice with reciprocal lattice vectors equal to $\vec{K}_j - \vec{K}_l$ and with a unit cell structure determined by the pre-factors $\vec{E}_{0j}\cdot\vec{E}^{*}_{0l}$. By choosing the intensity of the incident beams and the exposure time appropriately, it is thus possible to generate a pattern in the medium that has the same periodicity and structure as the I_{th} isophote of the interference pattern.

The simplest implementation of MBI-3DM involves interfering two linearly polarized beams at angles $\pm\theta/2$ with respect to the normal to the film being patterned (θ is the angle between the two beams in vacuum). From eqn [8], it can be shown that the intensity is maximum if $(\vec{K}_1 - \vec{K}_2)\cdot\vec{r} = 2q\pi$, where q is an integer, or $K[\sin(\theta/2)]r' = q\pi$, where K is the magnitude of both $\vec{K}_1$ and $\vec{K}_2$, and r' is the component of $\vec{r}$ along $(\vec{K}_1 - \vec{K}_2)$. Thus, the two beams generate an interference pattern with periodicity (in one dimension) given by $\Lambda = \lambda/[2\sin(\theta/2)]$, if λ is the

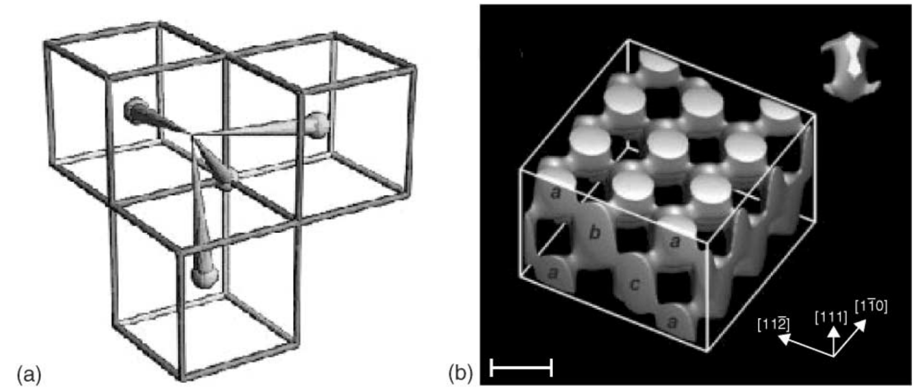

Figure 9 (a) Relative orientation of the beams in a four-beam interference scheme to obtain a face-centered cubic lattice. The beams have wave vectors along the directions [−3/2, −3/2, −3/2], [−5/2, −1/2, −1/2], [−1/2, −5/2, −1/2], and [−1/2, −1/2, −5/2]. (b) One of the isophotes generated by the interference of the four beams in (a). The inset represents the shape of the unit cell. Reproduced with permission from *Nature*, Campbell M, *et al.* (2000) *Nature* 404: 53–56. Copyright (2000), Macmillan Magazines Limited.

vacuum wavelength of the exciting beam ($\lambda = 2\pi/K$). When a film of a polymerizable material is placed at the intersection of the beams, a series of ridges is generated in the resist, spaced by Λ and with a width dependent on the beam intensity. For example, if a laser beam at 800 nm is split in two components forming an angle of 30°, $\Lambda = 1.5$ μm. The periodicity becomes smaller as the angle between the beams is increased.

By interfering three beams it is possible to obtain a two-dimensional periodic structure. For example, the interference of three beams forming an angle θ with respect to the medium normal and symmetrically arranged around it was used to generate an hexagonal lattice of rods that extends parallel to the normal. At least four beams must interfere to obtain a microstructure with 3D periodicity. The beam geometry shown in Figure 9a was used to produce a microstructure with the symmetry of a face-centered cubic lattice (Figure 9b). Beam geometries for generating interference patterns having other symmetries, such as body-centered cubic and body-centered tetragonal, have also been described.

An advantage of MBI-3DM is that it enables rapid patterning of large volumes during a single exposure. In fact, the polymerization takes place, in principle, throughout the region in which the beams overlap. When the process is activated by OPE, the beams can be attenuated significantly by absorption in the material. This can limit the thickness of the fabricated structure or result in spatial variation in the filling ratio. The use of TPIP can overcome this limitation and allow for thicker media to be patterned uniformly. Structures with the same symmetry are generated if either one- or two-photon processes are exploited to induce the material transformation. However, the structures will be different for at least two reasons: (i) if the same initiator is used in both cases, the wavelength used for TPE is approximately twice that for OPE, and

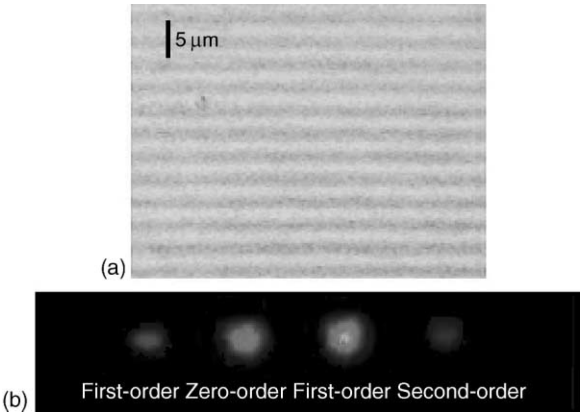

Figure 10 (a) Optical image of a diffraction grating obtained by TPIP using a two-beam interference scheme ($\lambda = 810$ nm). The spacing period is 3.4 μm and film thickness is 40 μm. The grating is made of an acrylate-based resin. (b) Diffraction pattern of the grating in (a) generated using a Helium–Neon laser (632.8 nm). The maximum diffraction efficiency was measured to be 57%. Reproduced from Guo H, *et al.* (2003) *Chemical Physics Letters* 374: 381–384. Copyright (2003), with permission from Elsevier.

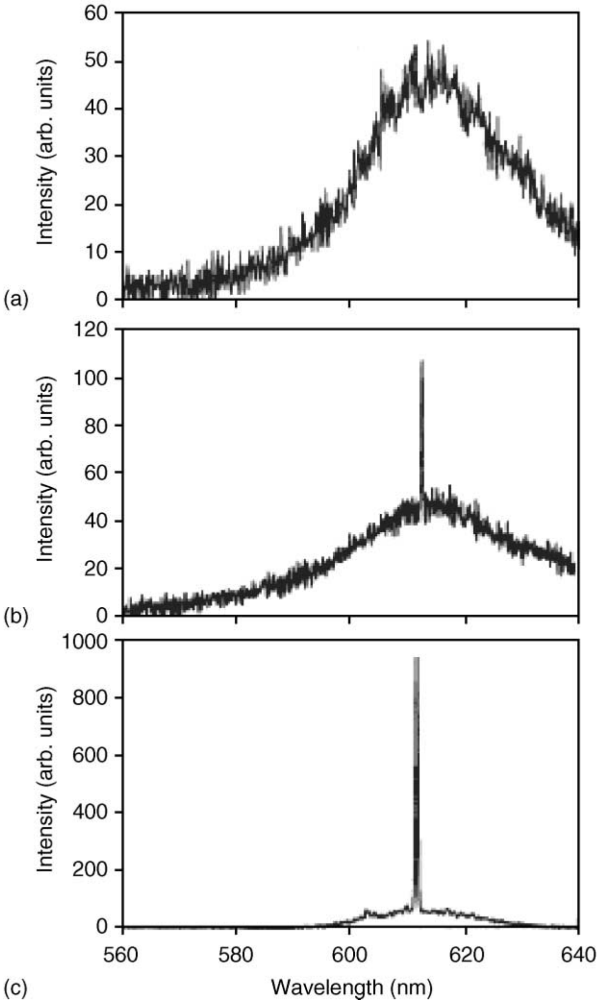

Figure 11 Emission spectra of the laser dye DCM encapsulated in a dendrimer-based resin that was patterned into a grating microstructure by single-beam 3DM. The grating was pumped at 532 nm (8 ns pulses) with excitation energies of (a) 0.03, (b) 0.05, and (c) 0.20 μJ. Starting at excitation energies as in (b), a narrow band at 612 nm appears over the spontaneous emission. The relative intensity of this lasing emission increases with excitation intensity with respect to the spontaneous emission. Reproduced from Yokoyama S, *et al.* (2003) *Thin Solid Films* 438–439: 452–456. Copyright (2003), with permission from Elsevier.

the periodicity of the structure is correspondingly different; and (ii) the I_{th} isophote, and thus the details of the polymerized structure inside each unit cell, is different under OPE and TPE because the excitation rate is proportional to I and I^2, respectively. In principle, better defined features can be obtained using two-photon or higher-order excitation because the transition between regions of low- and high-exposure is sharper than that under OPE, due to the nonlinear dependence on I. Interference fabrication techniques can enable the generation of features considerably smaller than the diffraction limit, when the beam geometry and excitation intensity are chosen appropriately.

The beams used to generate the interference pattern are usually obtained by splitting the output of a single laser source into the desired number of components. The intensity and polarization of each beam are then independently controlled. Temporal overlap of the beams at the sample can be ensured by introducing an optical delay into each of the beam paths. As fluctuations in the beam intensity can cause changes in the lattice unit cell or its filling ratio, the use of a laser with good stability is imperative in cases where exposure time extends over more than a single laser pulse.

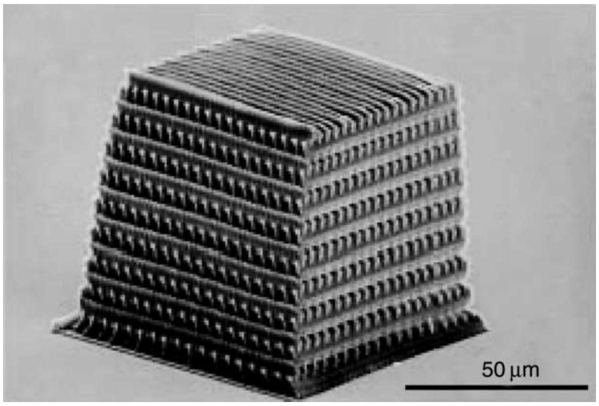

Figure 12 Example of a wood-pile PhC structure obtained by single-beam TPIP-3DM. The periodicity is 5 μm. The change in the cross-section of the structure from top to bottom, is due to shrinking of the material during development, whose effect is largest in the region farthest from the substrate. Reproduced with permission from *Nature*, Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimentional optical data storage and microfabrication. *Nature* 398: 51–54, Copyright (1999), Macmillan Magazines Limited.

Structures and Functional Devices

The versatility of 3DM is easily appreciated by surveying the variety of microstructures that have been fabricated over the years. These may be divided into two categories: (i) structures whose characteristics lay in their form and in the interconnectivity of their parts; and (ii) structures that incorporate an active functionality that can be exploited in a specific application. Some of these microstructures are described below.

Fabricating functional micro-optical devices is a promising application of 3DM. Figure 10 shows a grating fabricated by two-beam-interference TPIP in an acrylate with a period of 3.4 μm, whose diffraction efficiency is 28% as fabricated and 57% after annealing. The same method was used to generate a surface relief grating in a resin containing a polycationic peptide. In the presence of aqueous silicic acid, the peptide catalyzes the formation of ordered arrays of silica nanospheres in

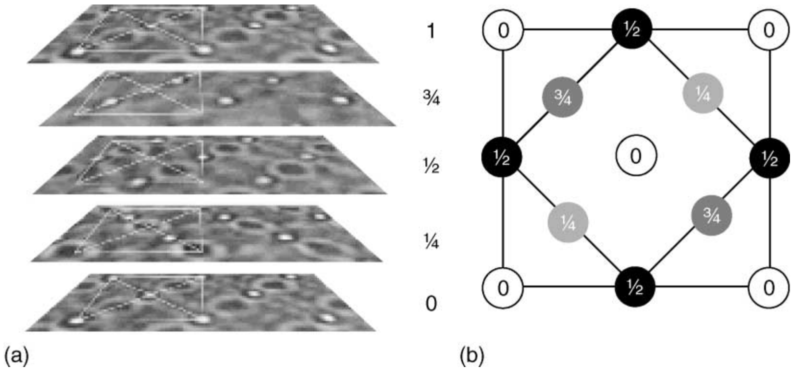

Figure 13 (a) Cross-sectional optical images of various planes of a diamond lattice fabricated in glass. The numbers to the right are the relative positions of the planes, in units of the lattice constant. (b) Diagram of the expected position of lattice points in a diamond structure. Points located in different planes have different shading. Reproduced from Mizeikis V, *et al.* (2001) *Journal of Photochemistry and Photobiology A* 145: 41–47. Copyright (2001), with permission from Elsevier.

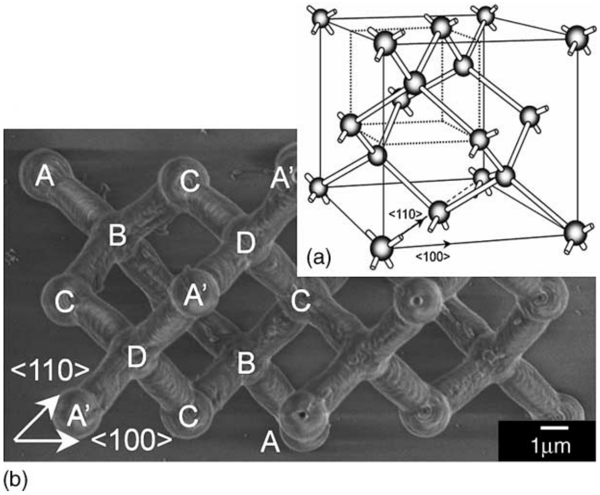

Figure 14 (a) Illustration of a diamond lattice. (b) Scanning electron micrograph of a polymeric microstructure with the symmetry of a diamond lattice, obtained by single-beam TPE-3DM. The letters A, B, C, D, A′ refer to the lattice planes at 0, 1/4, 1/2, 3/4, and 1 lattice units in the vertical direction. Reproduced from Kaneko K, *et al.* (2003) *Applied Physics Letters* 83: 2091–2093. Copyright (2003), with permission of The American Institute of Physics.

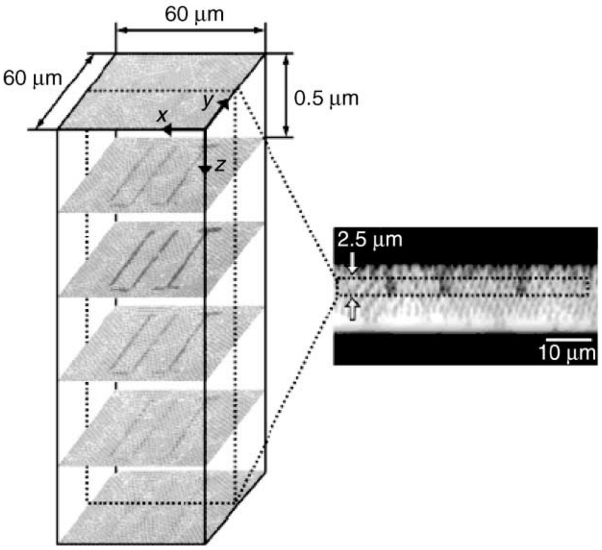

Figure 15 (Left) Images obtained by laser scanning confocal microscopy of linear defects generated by three-photon-induced polymerization of a triacrylate resin backfilled into a silica colloidal crystal. The spacing between the planes of the images is 0.5 µm. (Right) Cross-sectional image in the *xz*-plane of the same structure. Reproduced with permission from Lee W, *et al.* (2002) *Advanced Materials* 14: 271–274. Copyright (2002), with permission of Wiley.

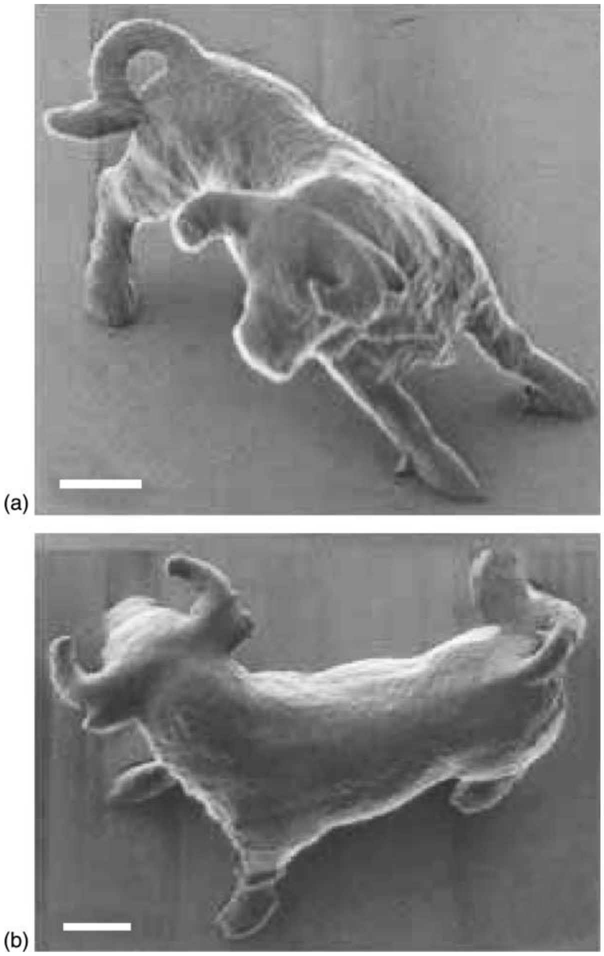

Figure 16 Microstructure of a bull fabricated by single-beam 3DM. The scale bar corresponds to 2 µm. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

the troughs of the grating, which increases the diffraction efficiency 50-fold. 3DM was used to fabricate a grating structure in an acrylate resin containing the laser dye 4-(dicyanomethylene)-2-methyl-6-(4-dimethylaminostyryl)-4H-pyrane (DCM) encapsulated in a host dendrimer (see above) as an active optical medium. When pumped at 532 nm above a certain threshold, lasing emission from DCM was observed at 612 nm, with a bandwidth of 0.15 nm (Figure 11).

Several groups have explored the potential of 3DM as a tool for patterning waveguide-based optical circuitry. For example, a mixture of thermally and photocurable resins was used to generate embedded waveguides by photocross-linking within the thermally set medium. A Y-coupler was created in fused silica through the refractive index change generated by patterned MPE.

3DM provides a powerful route to 3D photonic crystals (PhC). PhCs are structures in which a basic motif, or ‘unit cell’, is repeated within the material with a given periodicity. PhCs have attracted great interest, because the modulation of the refractive

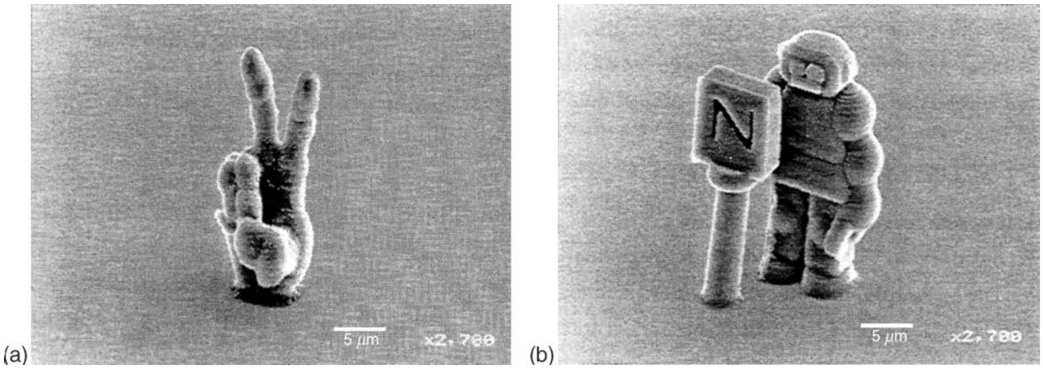

Figure 17 Scanning electron micrographs of two structures in the form of (a) a hand and (b) a robot which were fabricated by single-beam 3DM. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. Copyright (2000), with permission of SPIE.

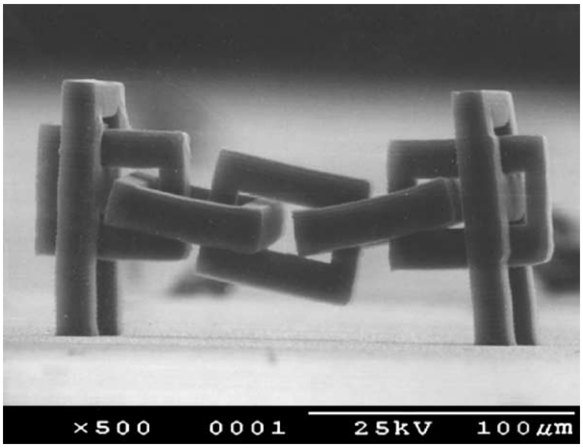

Figure 18 Scanning electron micrograph of a series of connected chain links obtained by single beam TPIP-3DM in an acrylate resin. Notice the tilt angle between the central links, indicating that each link can move with respect to the others. Reproduced from Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimentional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668. Copyright (2001), with permission of the Technical Association of Photopolymers Japan.

index between the 'filled' areas and the surrounding space provides a means for controlling the propagation of electromagnetic radiation within the structure. PhC with periodicities in the submicron and micron range exhibit their peculiar properties in the visible and near infrared spectral region. Optical wavelength PhCs can be fabricated in a straightforward way by 3DM, either by serial patterning using the single beam approach or by parallel patterning using the interference method, for which the number and orientation of the beams is dictated by the symmetry of the desired structure. Typical examples are woodpile structures (Figure 12), which consist of layers of rods of a material fabricated side by side with a fixed spacing, and alternate layers are rotated by 90° with respect to one another. If the rods in one layer are displaced by half the lattice constant with respect to the previous layer with the same orientation, the structure obtained has the symmetry of a face-center cubic lattice. PhCs with the symmetry of the diamond crystal have also been fabricated, either by inducing changes in a glass

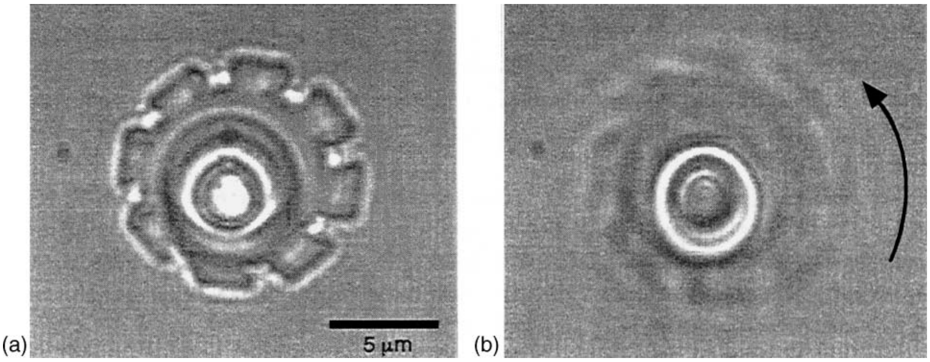

Figure 19 Optical image of a micro-wheel obtained by single-beam 3DM. (a) Structure before development. (b) When ethanol is used to remove the unpolymerized material, the wheel rotates around the axle in the solvent. Reproduced from Maruo S and Ikuta K (2000) *Proceedings of the Society of Photo-Optical Instrumentation Engineers* 3937: 106–112. With permission of the Institute of Electrical Engineers of Japan.

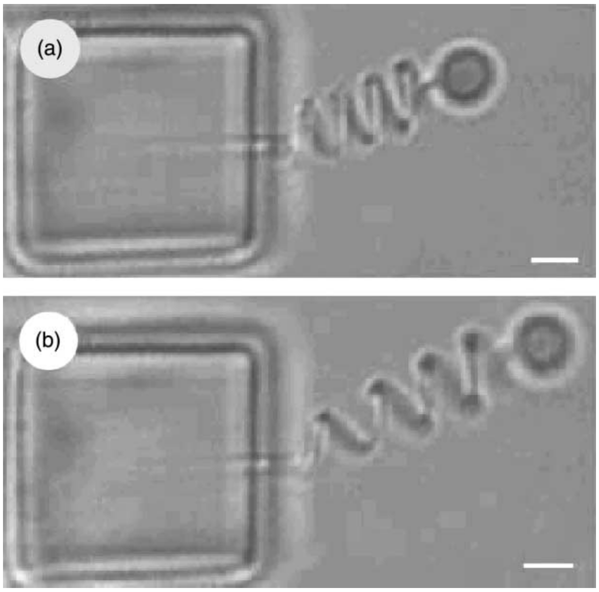

Figure 20 Optical image of a micro-spring tethered to a polymer block. The structure in (a) is at its resting point in ethanol. In (b) the spring was elongated by displacing the polymer bead at the free end using laser trapping. Reproduced with permission from *Nature*, Kawata S, *et al.* (2001) *Nature* 412: 697–698. Copyright (2001), Macmillan Magazines Limited.

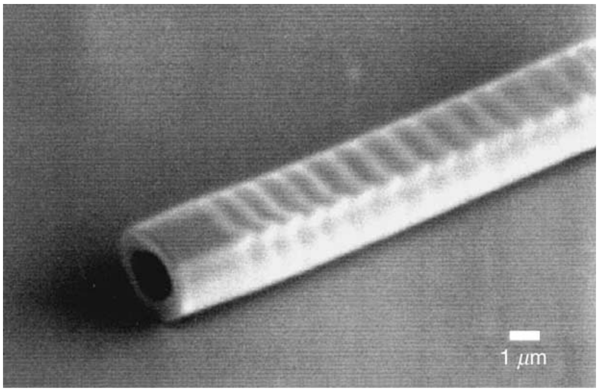

Figure 21 Hollow micro-tubule generated by single-beam TPIP-3DM in a urethane acrylate. The inner diameter is 1.8 μm. Reproduced from S Maruo and S Kawata (1998) *Journal of Microelectromechanical Systems* 7: 411–415. Copyright (1998), with permission of IEEE.

substrate at lattice points (Figure 13), or by constructing a network of polymeric rods connecting nearest neighbor points (Figure 14). It was showed that a PhC created by TPIP-3DM does exhibit a partial bandgap in the infrared spectral region. The position of the bandgap can be tuned by fabricating the structure with the appropriate periodicity.

The ability to introduce defects in PhCs in a controlled fashion is of key importance for the application of these structures in optical circuitry, as they can act as waveguides for channeling light into preferential directions. Point, line, and plane defects can be generated easily, by locally altering the fabrication pattern in the PhC. Point-by-point MPE of a Ge-doped glass was used to pattern a cubic lattice containing systematic defects that defined a Y-shaped waveguide splitter in a [100] plane. 3DM can also be used to modify a structure that is created using another fabrication method, and additional materials can be introduced into the structure in this second fabrication step. A silica colloidal crystal was backfilled with a polymerizable material and waveguides were generated by photocross-linking channels in the medium through three-photon-excitation 3DM (Figure 15).

The full potential of single-beam 3DM can be appreciated by surveying some of the structures that have been fabricated, featuring topologically complex shape, high interconnectivity, and extreme under-cut. Examples include those of Figures 16, 17, and the diamond-lattice PhC of Figure 14. Structures with freely movable parts, like those needed for MEMS, are difficult to obtain by other fabrication techniques, but can often be made by 3DM in a single exposure step. The ability to generate structures with independent but interconnected parts is exemplified by the microchain of Figure 18. A further example is the microgear in Figure 19, which is shown to spin around its axle when the sample is flooded with a liquid. Other functional microstructures have also been reported, such as the microspring in Figure 20, fabricated in a urethane acrylate resin and characterized by a core diameter of only 300 nm. When the sample is immersed in ethanol, the pendant mass can be captured and extended with optical tweezers. When the mass is released, the spring behaves like a damped oscillator.

Promising results have also been obtained in the production of structures to be used in micro-fluidics applications. Of key importance, in this case, is the ability to pattern channels, cavities, and other hollow structures that can be used to store or transfer small amounts of fluids between different parts of a device, as needed to perform, for example, analytical processes on a femtoliter scale. Examples include the hollow microtubule (Figure 21), and the submersed channels structure fabricated in a positive-tone resist (Figure 5).

List of Units and Nomenclature

A	Electron accepting molecular fragment
AFS	Amplified femtosecond
C	Chromophore
$\mathbf{C}^*$	Excited chromophore
CW	Continuous wave
D	Electron donating molecular fragment
3D	Three-dimensional
3DM	Three-dimensional microfabrication
$\vec{E}_0$	Spatial component of the electric field vector [V m^{-1}]
E_p	Laser pulse energy [J]
$f(t)$	Normalized temporal distribution of the beam intensity [dimensionless]
FOM	Figure of merit
h	Planck's constant [J s]
$i(z, r)$	Spatial distribution of the beam intensity [W cm^{-2}]
I	Intensity of a propagating beam [W cm^{-2}]
I_0	Maximum on-axis intensity for a Gaussian beam [W cm^{-2}]
I_{th}	Threshold intensity (minimum peak intensity at which polymerization occurs for a given exposure time) [W cm^{-2}]
I·	Initiating radical
j, l, q	Integers [dimensionless]
J_1	First order Bessel function [dimensionless]
K	Amplitude of a beam wave vector ($K = 2\pi/\lambda$) [m^{-1}]
$\vec{K}$	Wave vector of a beam [m^{-1}]
m	Number of beams in a multi-beam interference configuration [dimensionless]
M	Monomer
MBI	Multi-beam interference
MEMS	Micro-electromechanical systems
MPE	Multi-photon excitation
n	Number of photons involved in MPE [dimensionless]
n_0	Linear refractive index of the medium [dimensionless]
N	Number density (concentration) of a species [cm^{-3}]
NA	Numerical aperture of a lens [dimensionless]
N_C	Number density of chromophores present [cm^{-3}]
Nd:YAG	Neodymium:Yttrium aluminum garnet
$N_{I\cdot}$	Number density of initiating radicals generated [cm^{-3}]
N_{Inh}	Number density of inhibitors present [cm^{-3}]
OPE	One-photon excitation
Ormocer	Organically modified ceramic
$\langle P \rangle$	Average laser power [W]
$\langle P \rangle_{min}$	Minimum average laser power for patterning a material [W]
	Repetition rate (number of laser pulse per unit time) [Hz]
PAG	Photo-acid generator
PhC	Photonic crystal
r	Radial coordinate (distance from optical axis) [m]
r'	Component or $\vec{r}$ along a given wave vector [m]
$\vec{r}$	Position vector in three-dimensional space [m]
R	Twice the radial coordinate of the point at which the intensity is $I_0/2$ [m]
$R_{1h\nu}$	Molecular one-photon excitation rate (number of molecules excited per unit volume and time via OPE) [cm^{-3} s^{-1}]
$R_{2h\nu}$	Molecular two-photon excitation rate (number of molecules excited per unit volume and time via TPE) [cm^{-3} s^{-1}]
$R_{nh\nu}$	Molecular n-photon excitation rate (number of molecules excited per unit volume and time via n-order MPE) [cm^{-3} s^{-1}]
RMS	Root-mean-square
t	Time [s]
TEM_{00}	Lowest order transverse electromagnetic mode of a laser
Ti:S	Titanium:sapphire
TPE	Two-photon excitation
TPIP	Two-photon induced polymerization
z	Coordinate along the beam propagation direction [m]
z_R	Rayleigh range of a Gaussian beam; axial distance between the point of minimum beam waist and the point at which the waist has increased by a factor of $\sqrt{2}$ [m]
Z	Distance between the points of intensity $I_0/2$ along the coordinate z [m]
ζ	Longitudinal optical coordinate ($\zeta = 2\pi Z(\mathrm{NA})^2/n_0\lambda$) [dimensionless]
θ	Angle between a beam wave vector and the normal to a substrate [rad]
λ	Wavelength (in vacuum) [m]
Λ	Period of a one-dimensional grating [m]

ν	Frequency [s^{-1}]
$\boldsymbol{\pi}$	π-conjugated molecular fragment
ρ	Transverse optical coordinate ($\rho = 2\pi r\mathrm{NA}/\lambda$) [dimensionless]
$\sigma^{(1)}$	One-photon absorption cross-section [cm^2]
$\sigma^{(2)}$	Two-photon absorption cross-section [cm^4 s $photon^{-1}$]
$\sigma^{(n)}$	n-photon absorption cross-section [cm^{2n} s^{n-1} $photon^{-(n-1)}$]
τ_{ex}	Exposure time [s]
τ_p	Laser pulse duration [s]
$\phi_c^{(n)}$	n-photon chemical quantum yield [(# of chemical species generated)/(# of n-photon-excited species)] [dimensionless]
ϕ_{H+}	Chemical yield for the generation of H^+ (number of H^+ generated per excited molecule) [dimensionless]
ω	Angular frequency ($\omega = 2\pi\nu$) [Hz]

See also

Chemical Applications of Lasers: Nonlinear Spectroscopies. **Diffractive Systems:** Applications of Diffractive and Micro-Optics in Lithography; Design and Fabrication of Diffractive Optical Elements. **Holography, Techniques:** Overview. **Imaging:** Volume Holographic Imaging. **Laser-Induced Damage of Optical Materials**. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy; Nonlinear Microscopy; Overview. **Photonic Crystals:** Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Self-Assembled and Functionalized Photonic Crystals. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Belfield KD, Schafer KJ, Liu YU, Liu J, Ren XB and van Stryland EW (2000) Multiphoton-absorbing organic materials for microfabrication, emerging optical applications and non-destructive three-dimensional imaging. *Journal of Physical Organic Chemistry* 13: 837–849.

Born M and Wolf E (1980) *Principles of Optics*, 6th edn. Cambridge, UK: Cambridge University Press.

Cumpston BH, Ananthavel SP, Barlow S, *et al.* (1999) Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* 398: 51–54.

Joshi MP, Pudavar HE, Swiatkiewicz J, Prasad PN and Reinhardt BA (1999) Three-dimensional optical circuitry using two-photon-assisted polymerization. *Applied Physics Letters* 74: 170–172.

Kawata S and Sun H-B (2003) Two-photon photopolymerization as a tool for making micro-devices. *Applied Surface Science* 208–209: 153–158.

Kuebler SM, Rumi M, Watanabe T, *et al.* (2001) Optimizing two-photon initiators and exposure conditions for three-dimensional lithographic microfabrication. *Journal of Photopolymer Science and Technology* 14: 657–668.

Kuebler SM, Braun KL, Zhou W, *et al.* (2003) Design and application of high-sensitivity two-photon initiators for three-dimensional microfabrication. *Journal of Photochemistry and Photobiology A: Chemistry* 158: 163–170.

Martineau C, Anémian R, Andraud C, Wang I, Bouriau M and Baldeck PL (2002) Efficient initiators for two-photon induced polymerization in the visible range. *Chemical Physics Letters* 362: 291–295.

Odian G (1981) *Principles of Polymerization*, 2nd edn. New York: Wiley.

Pitts JD, Campagnola PJ, Epling GA and Goodman SL (2000) Submicron multiphoton free-form fabrication of proteins and polymers: studies of reaction efficiencies and applications in sustained release. *Macromolecules* 33: 1514–1523.

Ren Y, Yu X-Q, Zhang D-J, *et al.* (2002) Synthesis, structure and properties of a new two-photon photopolymerization initiator. *Journal of Materials Chemistry* 12: 3431–3437.

Strickler JH and Webb WW (1990) Two-photon excitation in laser scanning fluorescence microscopy. *Proceedings of the Society of Photo-Optical Instrumental Engineers* 1398: 107–118.

Sun H-B and Kawata S (2003) Two-photon laser precision microfabrication and its applications to micro-nano devices and systems. *Journal of Lightwave Technology* 21: 624–633.

Sun H-B, Maeda M, Takada K, Chon JWM, Gu M and Kawata S (2003) Experimental investigation of single voxels for laser nanofabrication via two-photon photopolymerization. *Applied Physics Letters* 83: 819–821.

Wu ES, Strickler JH, Harrell WR and Webb WW (1992) Two-photon lithography for microelectronic application. *Proceedings of the Society of Photo-Optical Instrumental Engineering* 1674: 776–782.

Wu P-W, Cheng W, Martini IB, Dunn B, Schwartz BJ and Yablonovitch E (2000) Two-photon photographic production of three-dimensional metallic structures within a dielectric matrix. *Advanced Materials* 12: 1438–1441.

Xu C and Webb WW (1997) Multiphoton excitation of molecular fluorophores and nonlinear laser microscopy. In: Lakowicz JR (ed.) *Topics in Fluorescence Spectroscopy*, vol. 5, pp. 471–540. New York: Plenum.

NONLINEAR OPTICS, BASICS

Contents

Cascading
$\chi^{(2)}$–Harmonic Generation
$\chi^{(3)}$–Third-Harmonic Generation
Four-Wave Mixing
Kramers–Krönig Relations in Nonlinear Optics
Nomenclature and Units
Nonlinear Optical Phase Conjugation
Photorefraction
Ultrafast and Intense-Field Nonlinear Optics

Cascading

G Assanto, University 'Roma Tre', Rome, Italy

G I Stegeman, University of Central Florida, Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Historically the field of nonlinear optics has evolved along two main axes, depending on the order of the nonlinearity used. The phenomena associated with the second-order nonlinearity $\chi^{(2)}$ have dealt with frequency conversion requiring wavevector matching (i.e., momentum conservation), such as second harmonic and optical parametric generation. On the other hand, effects associated with the third-order susceptibility $\chi^{(3)}$ have focused on inherently phase-matched phenomena at a single frequency, for example an intensity-dependent refractive index or degenerate four-wave mixing.

In cascading, multiple sequential nonlinear events based on $\chi^{(2)}$ can mimic some well-known $\chi^{(3)}$ phenomena. For example, it leads to nonlinear phase shifts $\Delta\phi^{\mathrm{NL}}$ reminiscent of those obtained through self-phase modulation over a distance L, i.e.,

$$\Delta\phi^{\mathrm{NL}}(L, I) = k_0 L n_2 I \qquad [1]$$

where I is the irradiance, k_0 the vacuum wavevector, and n_2 the Kerr coefficient or nonlinear refractive index. These can be used for all-optical switching, the generation of spatio-temporal solitons, etc. Some of these cascading effects have been known from the very earliest days of nonlinear optics, but their effectiveness was not realized until recently, when it proved possible to phase match second-order interactions with usefully large nonlinearities. Note that the medium must be $\chi^{(2)}$-active, i.e., noncentrosymmetric, which usually requires crystals rather than amorphous materials (such as glass).

Here we review the basic concepts, the characteristic interactions, some of the numbers involved and, finally, a few applications.

Physics of Cascading

Nonlinear Phase Shift in Collinear Second Harmonic Generation

The most important cascading effect, the nonlinear phase shift, is rooted in the fundamentals of the parametric interaction when two or three waves of different frequencies ($\omega_3 = \omega_1 \pm \omega_2$) are coupled by the second-order susceptibility $\chi^{(2)}(-\omega_3;\ \omega_1,\ \pm\omega_2)$. The simplest case is 'Type I' second harmonic generation (SHG), in which one eigenwave at the fundamental (FF) and one at the harmonic frequency (SH) interact with $\omega_1 = \omega_2 = \frac{1}{2}\omega_3$. When the waves have different phase velocities, this process involves simultaneously up-conversion ($\omega_1 + \omega_1 \rightarrow \omega_3$) and down-conversion ($\omega_3 - \omega_1 \rightarrow \omega_1$). By the latter the FF is regenerated from the harmonic after a characteristic propagation distance called the coherence length. This phenomenon is cyclic (see Figure 1a).

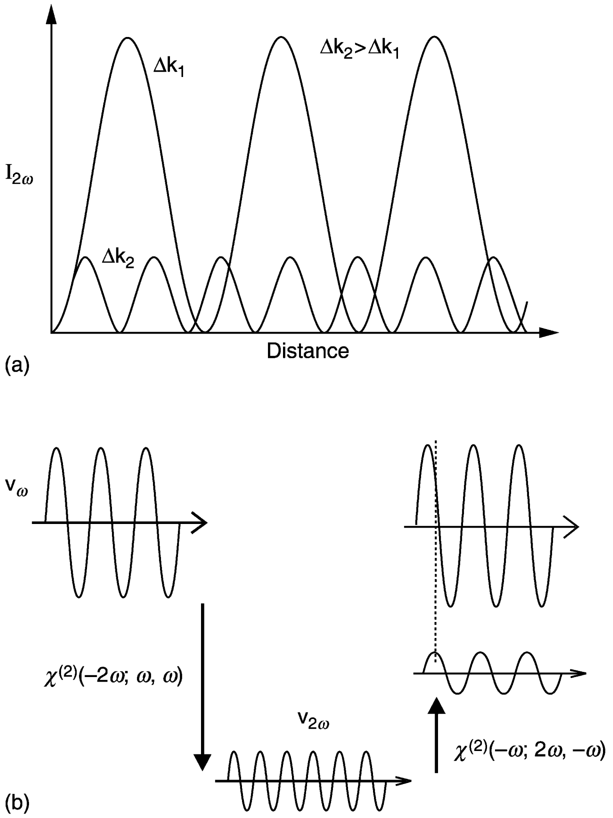

Figure 1 Phase-mismatched SHG. (a) Typical evolution of the generated second harmonic irradiance versus propagation for different mismatches. The period is twice the coherence length. (b) Intuitive picture of phase shift through cascading: the fraction of the initial energy which is first up-converted (SHG) and later down-converted (DFG), having traveled at a different phase velocity due to the wavevector mismatch, recombines with the through-FF causing an overall phase shift.

The field amplitudes in normalized form are

$$E_1(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_1}}\, a_1(z) e^{i(\omega_1 t - k_1 z)} + \text{c.c.}$$
$$E_3(z,t) = \frac{1}{2}\sqrt{\frac{2}{c\varepsilon_0 n_3}}\, a_3(z) e^{i(\omega_3 t - k_3 z)} + \text{c.c.} \qquad [2]$$

where $|a_1|^2$ and $|a_3|^2$ are power densities, c is the speed of light in vacuum, and n_j and k_j are the refractive indices and wavevectors at angular frequencies ω_j $(j = 1, 3)$, respectively. As is common in nonlinear optics, the efficiency of these processes depends on the irradiance of the different waves, and on the wavevector mismatch between them, i.e., $\Delta k = 2k_1 - k_3$ for collinear beams. Since the phase velocities of the two waves, $v_{\omega_1} = \omega_1/k_1$ and $v_{\omega_3} = \omega_3/k_3$, are generally different and $\Delta k \neq 0$, their phase fronts fall out of synchronism during propagation. This leads to a periodic oscillation with distance (coherence length $= \frac{1}{2}$period $= \pi/|\Delta k|$) in the irradiances of the fundamental and harmonic waves, conserving the total energy but giving rise to a cascading phase shift, as pictured in Figure 1b.

The equations that describe the evolution of the fields stem from coupled mode theory, and are given by

$$\frac{d}{dz} a_3(z) = -i\Gamma a_1^2(z) e^{-i\Delta k z}$$
$$\frac{d}{dz} a_1(z) = -i\Gamma a_3(z) a_1^*(z) e^{i\Delta k z} \qquad [3]$$

where

$$\Gamma = \frac{\omega_1 \chi^{(2)}_{\text{eff}}}{\sqrt{2n^3 c^3 \varepsilon_0}} \qquad [4]$$

$\chi^{(2)}_{\text{eff}}$ is the 'effective' nonlinearity (for a specific crystal orientation and field polarization) and we assumed $n = n_1 \cong n_3$. In the limit of a large mismatch, the energy conversion to the harmonic is negligibly small, and it is possible to find an analytical expression for an effective 'n_2', $n_{2,\text{eff}}$. In this limit $|a_1(z)| = |a_1(0)| =$ constant, and, in the absence of loss, solving for the small SH wave and substituting back into the equation for $da_1(z)/dz$ gives

$$\frac{d}{dz} a_1(z) = -i\frac{\Gamma^2}{\Delta k}\Big\{1 - \cos(\Delta k z) - i\sin(\Delta k z)\Big\} |a_1(0)|^2 a_1(z) \qquad [5]$$

By comparing this with the corresponding equation based on n_2 $(\propto \chi^{(3)}[-\omega; \omega, -\omega, \omega])$ and describing self-phase modulation, namely

$$\frac{d}{dz} a_1(z) = -in_2 \frac{\omega_1}{c} |a_1(z)|^2 a_1(z) \qquad [6]$$

we can define an equivalent nonlinear refractive index

$$n_2(z) = \frac{\pi\{\chi^{(2)}_{\text{eff}}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k}[1 - \cos(\Delta k z)] \qquad [7]$$

which turns out to be strictly nonlocal because it relies on propagation. Since the $\cos(\Delta k z)$ term oscillates rapidly in space, the above averages to

$$n_{2,\text{eff}} = \frac{\pi\{\chi^{(2)}_{\text{eff}}\}^2}{2\varepsilon_0 c n^4 \lambda_1 \Delta k} \qquad [8]$$

In this so-called 'Kerr limit' cascading leads to a Kerr-like nonlinear response, with either a self-focusing ($\Delta k > 0$, $n_{2,\mathrm{eff}} > 0$) or a self-defocusing nonlinearity, depending on the sign of Δk.

It is worthwhile to link the above equations to the physics of cascading with a simple argument. The evolution of the waves on propagation depends on their fields (in both amplitude and phase) at the input. When only the fundamental beam at $\omega = \omega_1$ is incident (the classical case of SHG), the harmonic is initially up-converted $\pi/2$ out of phase with the input (first of eqns [3]) via $\chi^{(2)}(-2\omega; \omega, \omega)$. If $\Delta k = 0$, the phase fronts travel at the same velocity and the relative phase is locked in at $\pi/2$. Difference frequency generation (DFG, second of eqns [3]) results in a down-converted fundamental which is an additional $\pi/2$ out of phase, i.e., exactly out of phase with the original input field. This leads to the monotonic depletion of the input and the growth of the SH with distance. If $\Delta k \neq 0$, there is only partial and periodic conversion to the harmonic, the wavefronts travel at different velocities (v_ω and $v_{2\omega}$, respectively), and the relative phase between the two waves changes with distance. As a result, the down-converted FF is no longer π out of phase with the input, thereby changing the overall phase of the total input-frequency wave by a cascaded amount $\Delta\phi^{\mathrm{NL}}$. This is the mechanism sketched in Figure 1b and buried in the better-known Maker-fringe effect represented in Figure 1a. Its magnitude depends on the irradiance and the mismatch, and its sign on that of Δk (see eqn [8]). The higher the input irradiance, the larger the harmonic conversion, and hence the larger $\Delta\phi^{\mathrm{NL}}$.

Alternatively, with a harmonic input at high irradiance and a weak seed at the FF, parametric amplification makes the fundamental grow via down-conversion. If $\Delta k = k_3 - 2k_1 \neq 0$, self-focusing (-defocusing) is obtained on the strong harmonic for $\Delta k > 0$ ($\Delta k < 0$). These phenomena can be generalized to the nondegenerate cases of sum and difference frequency generation, keeping in mind that when using three input waves their mutual phase relationship is essential in determining the details of the interaction.

The solutions to eqns [3] can be obtained numerically or written in terms of Jacobi elliptic integrals subject to boundary conditions ($a_3(0) = 0$, $a_1(0) \neq 0$) at $z = 0$, providing additional insight into the cascading phenomenon. In contrast to the Kerr case given by eqn [1], the variation in n_2 with distance implies a staircase-like increase in $\Delta\phi^{\mathrm{NL}}$ with distance for various detunings ΔkL, as shown in Figure 2a. This is consistent with our intuitive picture above, the increase in $\Delta\phi^{\mathrm{NL}}$ being maximized where the fundamental is the smallest (see Figure 2b for the oscillations in the fundamental amplitude), with the step-period corresponding to twice the coherence length. Conversely, for a fixed input irradiance, the evolution of $\Delta\phi^{\mathrm{NL}}$ exhibits two extrema with detuning ΔkL, as shown in Figure 3a. Note also that, for small ΔkL and large

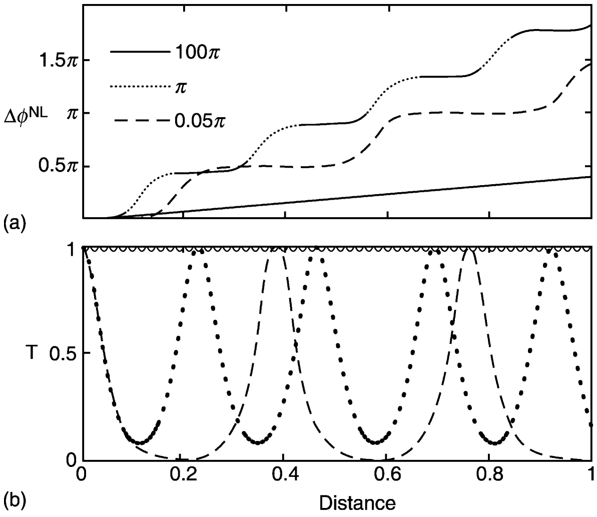

Figure 2 Calculated nonlinear phase (a) and throughput (b) of the fundamental field versus normalized propagation distance, for three values of phase mismatch.

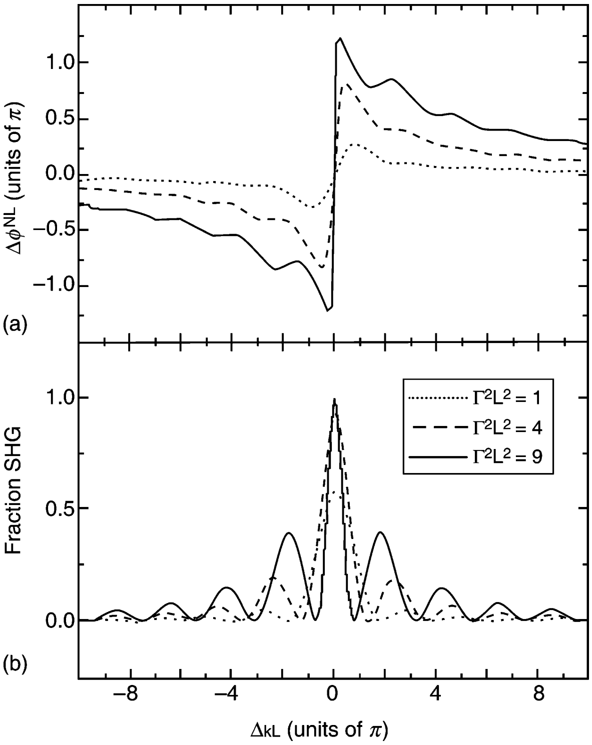

Figure 3 Calculated FF phase shift (a) and generated SH fraction (b) versus phase mismatch, i.e., the detuning curve, for three different nonlinear strengths (unitary FF irradiance).

(unrealistically high) irradiances, the increase in $\Delta\phi^{\mathrm{NL}}$ tends to become proportional to the field amplitude, a clear indication that cascading depends on a different order of nonlinearity than $\chi^{(3)}$, namely $\chi^{(2)}$. On the other hand, for large ΔkL the growth of $\Delta\phi^{\mathrm{NL}}$ is slower but linear with distance and irradiance, as in eqn [1].

There are trade-offs between the fundamental output and the net nonlinear phase shift. This is clearly visible in Figures 3a,b: the maxima in $\Delta\phi^{\mathrm{NL}}$ correspond to a reduced FF throughput. Although it is possible to operate cw at the zeros of the SHG, their location changes with the irradiances required for useful phase shifts ($>\pi$), complicating the trade-offs and the response to temporal pulses (encompassing a continuous irradiance distribution). The problem can be alleviated at the expense of reducing the net phase shift for a given irradiance and propagation length, by introducing regions with a nonuniform matching condition at both ends of the sample. The goal at the input end is to first adiabatically generate the SH required for optimum $\Delta\phi^{\mathrm{NL}}$. At the output, this harmonic is adiabatically converted back to the FF. The corresponding output for wavelengths larger than the phase-matching condition exhibits a $\Delta\phi^{\mathrm{NL}}$ with negligible fundamental depletion. This approach has been used in temperature-tuned and quasi phase matched (QPM, where a periodic modulation of the nonlinearity provides phase matching via the additional grating momentum) lithium niobate waveguides to generate phase shifts of π with less than 10% net loss of the fundamental wave to the harmonic.

While the preceding discussion has focused on Type I SHG, it is also valid for Type II SHG (i.e., with two FF input waves), provided the two fundamental inputs have equal input energy. In fact, an input imbalance between the two fundamental polarizations could give rise to a phase shift on the weaker field even when $\Delta k = 0$, leading to all-optical modulation and transistor action, as demonstrated in potassium titanyl phosphate (KTP) crystals.

Finally, due to the nature of the quadratic nonlinearity, the phase and/or amplitude of a coherent SH seed can have a large effect on the cascaded outcome, both in the phase and amplitude of the fundamental, as shown in Figure 4 for a fixed irradiance seed 1000 times weaker than the FF input. This can be used in seed-controlled optical processors.

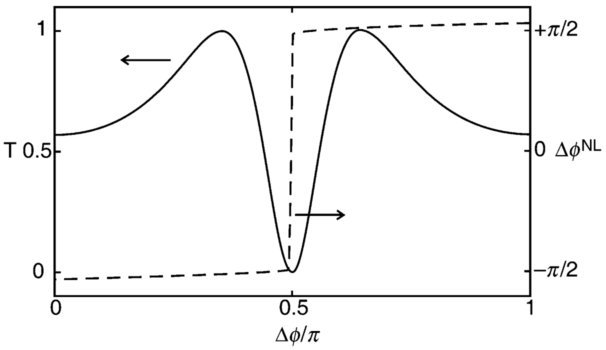

Figure 4 FF transmission (solid line) and phase shift (dashed line) versus the relative phase of a much weaker (three orders of magnitude) SH seed. Here perfect phase matching was assumed.

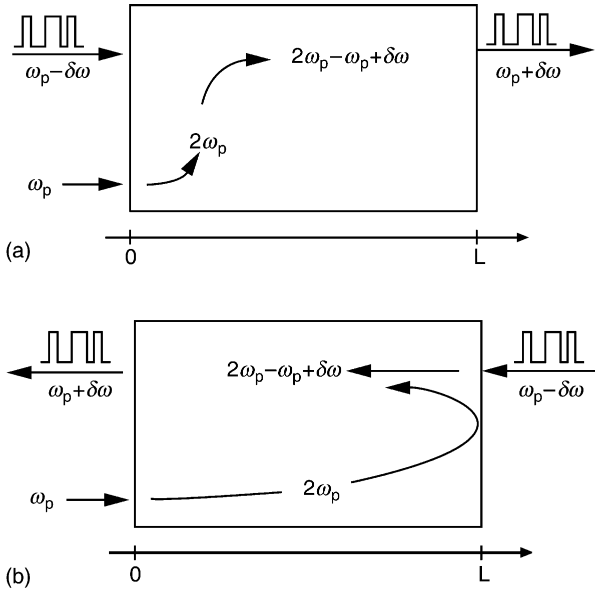
$\omega_p - \delta\omega$
$2\omega_p - \omega_p + \delta\omega$
$\omega_p + \delta\omega$
$2\omega_p$
ω_p
(a)
0
L
$\omega_p + \delta\omega$
$2\omega_p - \omega_p + \delta\omega$
$\omega_p - \delta\omega$
ω_p
$2\omega_p$
(b)
0
L

Figure 5 Wavelength shifter: (a) single-pass and (b) double-pass configurations. In (b) a dichroic mirror reflects the second harmonic of the pump.

Frequency Shifting

In the nondegenerate case of $\chi^{(2)}$ sum and difference frequency generation, two successive operations can perform the useful function of a frequency shifter (Figure 5). The idea is to shift the signal frequency ($\omega_p - \delta$) by first generating from the pump (ω_p) a second harmonic ($2\omega_p$), followed by DFG involving the latter and the signal. This leads to an output at $\omega_p + \delta$ through $\chi^{(2)}(-(\omega_p + \delta); 2\omega_p, -(\omega_p - \delta))$, i.e., a frequency shift of 2δ with an output spectrally mirror-imaged about the pump frequency. This feature is useful in WDM (wavelength division multiplexing) systems, and guarantees a protocol-insensitive wavelength shift with an additional phase conjugation useful in dispersion-compensating schemes. Notice that, since $\omega_p + \delta \cong \omega_p - \delta$ and $\delta << \omega_p$, the bandwidth is large because the interaction is near the degeneracy point.

Figure 5 shows two possible configurations for the same device length. The efficiency of the process can be increased in a double-pass geometry by using a back-propagating signal and a mirror-coated facet to reflect the SH, as sketched in Figure 5b. The best results to date were obtained at communication wavelengths with single-pass waveguides in periodically poled (QPM) $LiNbO_3$, including shift of multiple wavelength-channels with low cross-talk.

Multiple Beam Generation

Two successive $\chi^{(2)}$ processes can also be used to mimic three-wave mixing, a classic noncollinear $\chi^{(3)}$ interaction where two photons are taken out of one beam and one out of the second. In cascading, first a second harmonic is generated by one of the inputs, and then difference frequency mixing (DFM) between that SH and the second input leads to the signal of interest. In the simplest configuration, two equi-frequency (ω) beams of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ are incident in slightly different directions close to phase-matching and overlap inside the sample. When wave 1 is doubled (2ω) and wave 2 mixes with this harmonic via DFM, a new beam (at $2\omega - \omega \rightarrow \omega$) is generated with wavevector $2\mathbf{k}_1 - \mathbf{k}_2$. Of course the complementary output at $2\mathbf{k}_2 - \mathbf{k}_1$ is also generated via the SHG of wave 2. If these two new beams are strong enough, they can generate another pair via cascading, and so on. The generation of more than six extra beams has been demonstrated in β-barium borate with 1064 nm inputs.

This interaction can be applied to the control of the frequency chirp in pulses. For example, if wave 1 has a frequency chirp α_1 and wave 2 a chirp α_2, then the chirp of the beam scattered into the direction $2\mathbf{k}_1 - \mathbf{k}_2$ is $2\alpha_1 - \alpha_2$. This becomes especially interesting if $\alpha_2 = -\alpha_1$ so that the chirp in the new beam is $3\alpha_1$, as observed and subsequently employed in enhancing pulse compression. Note that, since a cascaded phase shift occurs on the down-conversion step back to the original fundamental (wave 1), it is not present in the scattered beam resulting in a clean transfer of the chirp.

Typical Numbers

From eqn [8] it is clear that the figure of merit (FOM) for cascaded processes is the same as for all second-order processes, namely $|\chi^{(2)}|^2/n^4$. The key is to find materials phase-matchable at the wavelengths of interest. Therefore, basically any material that is suited for frequency conversion, either SHG or parametric generation, will work. The most versatile material, which also has the largest easily usable nonlinear coefficient (18 pm/V), is lithium niobate when QPM is employed. As a reference, Table 1 gives the maximum $n_{2,\mathrm{eff}}$ for 1-cm-long samples, obtained by extrapolating eqn [8] to the ΔkL which maximizes the nonlinear phase shift. This is not strictly valid for such small ΔkL, but it is a useful estimate.

Other Implications of Cascading

A straightforward conclusion that could be drawn is that cascading with its $n_{2,\mathrm{eff}}$ and $\Delta\phi^{\mathrm{NL}}$ can generally replace the Kerr n_2. This is true for self-phase modulation and the applications stemming from it, including all-optical guided-wave switches such as directional couplers, Mach–Zehnder interferometers, etc. However, this is not necessarily true for other applications, because all the waves have to be coupled through $\chi^{(2)}$ interactions (at most three at a time) and must be nearly phase-matched for efficient effects. For example, cross-phase modulation between two input FF polarization components is feasible via Type II SHG cascading, but specific wavelengths will require different crystals, or crystal cuts. The nonlinear phase shift can also be used in all-optical devices for analog processing, such as isolators, modulators, and transistors, in laser mode-locking, etc.

In the following subsections we discuss a few additional applications. Specifically, we describe a non-reciprocal device that operates as an isolator, and the cascading of beams of finite extent in space or time. The latter can counterbalance either spatial diffraction or temporal dispersion or both and gives rise to solitons of various dimensionalities.

An Optical Diode

A nonlinear response, when combined with a device structure that is asymmetric about its midpoint, can provide nonreciprocal behavior upon excitation of the structure (crystal, waveguide) from different ends.

Table 1 Figures for cascading in various materials. The effective nonlinear coefficient is estimated under optimum conditions and a 1 cm propagation length

Material L = *1 cm*	d_{ii} *(pm/V)*	d_{ij} *(pm/V)*	$n_{2,eff}$ *(cm^2/W)*
$LiNbO_3$	36		2×10^{-11}
$LiNbO_3$		5.8	5×10^{-13}
MNA(1)	165		7×10^{-10}
NPP(2)		84	2×10^{-10}
DAST(3)	600		6×10^{-9}

(1)MNA: 2-methyl-4-nitroaniline
(2)NPP: N-(4-nitrophenyl)-(L)-prolinol
(3)DAST: dimethyl amino stilbazolium tosylate

In analogy to the standard electronic element, we can refer to such device as a power-dependent optical diode. Dealing with quadratically nonlinear waveguides, simple ways to induce a directional asymmetry are variations in wavevector mismatch, or in QPM periodicity. A mismatch profile which depends on distance from one end, or a localized defect (i.e., an isolated domain in a QPM grating), will give rise to different throughputs when the waveguide is excited from opposite sides, up to complete FF depletion in one direction, and complete transmission in the opposite one. This can be explained by the fact that the output depends on the evolution and interaction of the FF field with its harmonic. The amount and phase of the latter, however, do depend on the details of the propagation path and can substantially change the FF throughput as in the case of seeded SHG.

The sketch of an all-optical diode in a QPM waveguide with its calculated response is shown in Figure 6a,b. Clearly, total depletion is obtained via SHG in backward propagation, whereas unity FF transmission is recovered by the insertion of a phase discontinuity (and seeded interaction) in the forward path. As demonstrated at 1.55 μm in lithium niobate, different degrees of isolation are available at various excitations, with a large extinction ratio at a given power when launching FF waves from opposite ends.

Quadratic Solitons

When beams propagate in a two- (planar waveguides) or three-dimensional (bulk) medium with a quadratic response, cascading induces a nonlinear phase-front distortion which, coupled to the inherent parametric gain, can effectively counteract diffraction and lead to diffractionless propagation or 'spatial solitons' (Figure 7). In contrast to the Kerr case, such solitons are multifrequency solutions or 'simultons' with field components (FF and SH in the degenerate SHG process) bound together in propagation despite possible walk-off or transverse velocity mismatch. Moreover, since the second-order nonlinearity is inherently saturable, the solitons are stable and robust even in bulk media, and provide 'clean' filtered beam profiles at FF and SH frequencies, as has been demonstrated in KTP. Just like other cascading phenomena, quadratic solitons can be controlled by a weak seed, by polarization imbalance at the input (Type II SHG), by direction of propagation (due to birefringence), by phase mismatch, etc. Employing quadratic spatial solitons, phenomena such as collisions, soliton fusion, spatial reshaping, and transverse light localization and instabilities have been studied and constitute an active field of investigation.

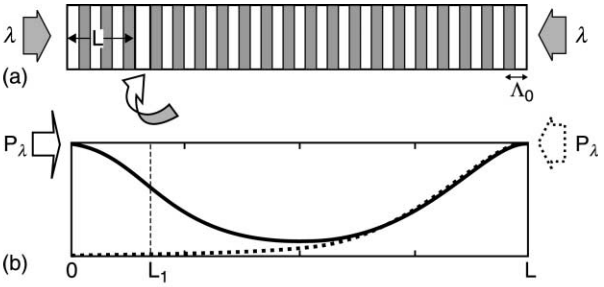

Figure 6 All-optical diode: (a) sketch of implementation with an engineered 'defect' in a QPM grating (*z*-cut periodically poled lithium niobate), and (b) example of FF transmittance *T* versus propagation for forward (solid line) and backward (dashed line) FF excitation of power P_λ.

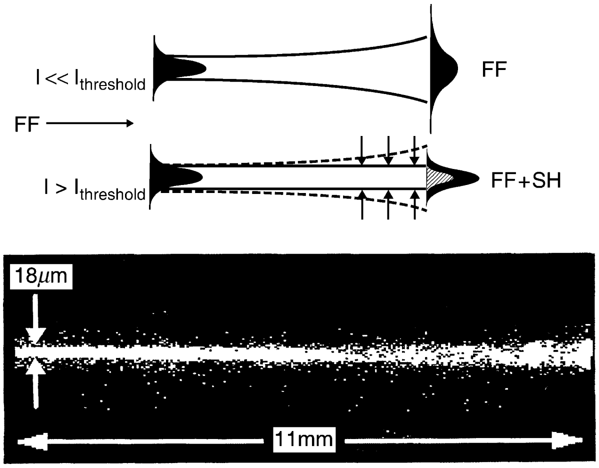

Figure 7 Top: sketch of a quadratic spatial soliton excited by an FF input: when the irradiance is high enough, linear diffraction is overcome by parametric self-trapping and a two-color simulton is obtained. Bottom: sample photograph (top view) of a quadratic soliton experimentally observed in a potassium niobate crystal excited by a 1064-nm laser beam.

When employing pulses, the cascaded phase shift can balance dispersion (material, chromatic, or photonics bandgap) leading to temporal simultons with frequency components locked together despite their individual group-velocity differences and dispersion. This feature can also substantially broaden the bandwidth available for parametric generation. Finally, in analogy to the cubic case, a combination of space and time effects is expected to lead to the formation of 'light bullets' in (3 + 1) dimensions.

List of Units and Nomenclature

Angular frequency	$[\mathrm{s}^{-1}]$	ω
Dielectric constant of vacuum	$[\mathrm{FV}^{-1}\,\mathrm{m}^{-1}]$	ε_0

Electric field	[V m^{-1}]	E
Irradiance, intensity	[W m^{-2}]	I
Length	[m]	L
Refractive index	[ ]	n
Speed of light in vacuum	[m s^{-1}]	c
Second-order electronic susceptibility	[m V^{-1}]	$\chi^{(2)}$
Third-order electronic susceptibility	[m^2 V^{-2}]	$\chi^{(3)}$
Wavevector in medium	[m^{-1}]	$k = n\omega/c$
Wavevector in vacuum	[m^{-1}]	$k_0 = \omega/c$
Wavelength	[m]	$\lambda = 2\pi n/k = 2\pi/k_0$

See also

Solitons: Bright Spatial Solitons; Soliton Communication Systems. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Assanto G (1995) All-optical switching in integrated structures. In: DeGiorgio V and Flytzanis C (eds) *Nonlinear Optical Materials: Principles and Applications*. Amsterdam: IOS Press.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland E (1993) All optical switching devices based on large nonlinear phase shifts from second harmonic generation. *Applied Physics Letters* 62: 1323–1325.

Assanto G, Stegeman GI, Sheik-Bahae M and VanStryland EW (1995) Coherent interactions for all-optical signal processing via quadratic nonlinearities. *IEEE Journal of Quantum Electron* 31: 673.

Assanto G, Stegeman GI and Schiek R (1998) Thin film devices for all-optical switching and processing via quadratic nonlinearities. *Thin Solid Films* 331: 291–297.

Boardman AD (ed.) (1999) *Advanced Photonics with Second-Order Optically Nonlinear Processes*. Dordrecht, The Netherlands: Kluwer.

Boardman AD (ed.) (2001) *Soliton Driven Photonics*. Dordrecht, The Netherlands: Kluwer.

Gallo K and Assanto G (1999) Analysis of lithium niobate all-optical wavelength shifters for the third spectral window. *Journal of the Optical Society of America B* 16: 741–753.

Gallo K, Assanto G, Parameswaran KR and Fejer MM (2001) All-optical diode in a periodically poled lithium niobate waveguide. *Applied Physics Letters* 79: 314–316.

Kajzar F and Reinisch R (eds) (1997) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Stegeman GI, Hagan DJ and Torner L (1996) $\chi^{(2)}$ cascading phenomena and their applications to all-optical signal processing, mode-locking, pulse compression and solitons. *Journal of Optics and Quantum Electronics* 28: 1691.

Stegeman GI and Assanto G (1999) Nonlinear integrated optical devices. In: Murphy EJ (ed.) *Design and Application of Integrated Optical Circuits and Components*. New York: Marcel Dekker.

$\chi^{(2)}$–Harmonic Generation

R C Eckardt, Cleveland Crystals, Inc., Highland Heights, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical harmonic generation is nonlinear conversion of laser output to light, of a frequency that is an integer multiple of that of the fundamental. It is used over a wide range of applications, from battery-operated, pencil-sized pointers emitting less than a milliwatt of light to installations that occupy large buildings and generate thousands of Joules of ultraviolet energy in nanosecond-duration pulses. The applications also range from frequency-stable continuous-wave light to generation of pulses a few femtoseconds in duration. Laser output with specific properties may only be available in narrow spectral regions, and harmonic generation is a method to convert the laser output to spectral regions of shorter wavelength. It is customary to call light of twice the frequency of the fundamental the second harmonic; three times the third harmonic, and so on.

The conversion of light by harmonic generation is a process that is driven by the nonlinear electric polarization. The total polarization $\mathbf{P}(t)$, including the linear term, is a vector quantity related to the electric field $\mathbf{E}(t)$ by the susceptibility tensors

χ of a material:

$$\mathbf{P}(t) = \varepsilon_0\chi^{(1)} : \mathbf{E}(t) + \varepsilon_0\chi^{(2)} : \mathbf{E}(t)\mathbf{E}(t) + \varepsilon_0\chi^{(3)} : \mathbf{E}(t)\mathbf{E}(t)\mathbf{E}(t) + \ldots, \qquad [1]$$

where $\varepsilon_0 \cong 8.854 \times 10^{-12}$ C/N·m^2 is the permittivity of free space. The first-order tensor describes linear optical properties on the material. The second-order tensor $\chi^{(2)}$ describes second-harmonic generation and sum- and difference-frequency generation. This is the term that is of interest here. The second-order nonlinear optical tensor, equal to half the second-order susceptibility, is used widely instead of the susceptibility:

$$d_{ijk} = \chi^{(2)}_{ijk}/2 \qquad [2]$$

In calculations of harmonic generation and sum- and difference-frequency generation, the coefficient is usually reduced to a scalar quantity d_{eff}, the effective nonlinear coefficient that expresses the relationship between only the specific vector components involved in an interaction. The second-order optical nonlinearity, when present, is usually significantly larger than higher-order optical nonlinearities. When optical frequency conversion is possible, using the second-order nonlinearity, it is generally preferred. Any material that has inversion symmetry, however, will have vanishing second-order nonlinearity. Although there is continuing interest in developing other materials, such as poled polymers for future applications, materials for current practical applications are restricted to noncentrosymmetric crystals.

This article will provide a description of the process of optical harmonic generation and discuss some of the concerns in the design of harmonic generation systems. In many cases, harmonic generation can be modeled accurately by analytic or numerical techniques. Some of the theoretical development that is the basis for this modeling will be presented along with examples. There are several essential considerations that must be made in selecting a harmonic generation process. These involve both material properties and conditioning of the fundamental radiation.

General Considerations

The nonlinear optical material must be highly transmitting at the fundamental and harmonic wavelengths and must have a laser-induced-damage threshold high enough to allow optical intensities that will provide adequate conversion efficiency. Phase matching must be possible in the material. Other material characteristics must be considered for specific applications. The harmonic generation process requires that the phase velocities of the fundamental and harmonic radiation be matched. Without phase matching, power converted to the harmonic is back converted to the fundamental in a rapid oscillation, and harmonic conversion remains at low levels. The material also needs to be of good optical quality to provide optical transmission without distortion, which would destroy the phase relationship. Continuous-wave harmonic generation may involve local intensities of the order of 1 megawatt per square cm (1 MW/cm^2). Pulsed applications with Q-switched laser output of a few nanoseconds duration typically employ intensities of several hundred MW/cm^2, and harmonic-generation applications with mode-locked lasers may involve pulse durations of tens of femtoseconds with peak intensities of tens of gigawatts per square cm. Many other properties are desirable, but it is necessary to work within the limitations of available materials.

Phase matching is an essential requirement. When harmonic and fundamental waves of different eigen polarizations are coupled through the nonlinear optical tensor, it may be possible to use crystal birefringence to provide phase matching. The birefringence, if large enough, can compensate dispersion. The fundamental wave is introduced into the crystal with the polarization of the larger refractive index, and the harmonic is generated in the polarization of the smaller index. This is a description of type-I birefringent phase matching. Phase matching is achieved by selecting the direction of propagation in the crystal or by changing the temperature of the crystal. Type-II phase matching with the fundamental wave resolved into the two eigen polarizations is possible with large birefringence. Quasi phase matching involves periodically structuring the nonlinear material in a way that spatial modulation of the optical nonlinearity compensates for dispersion. Quasi phase matching is tuned by selection of the modulation period, and temperature control may provide a fine tuning. Periodic poling, one method of quasi phase matching, involves changing the orientation of ferroelectric domains after each coherence length of the harmonic generation process. These methods of phase matching are sketched in Figure 1.

The physical arrangement of the harmonic generation process depends on properties of the available fundamental light and the properties of the selected nonlinear material. The single-pass techniques, shown in Figure 1, may have the advantage of simplicity, but expanding or focusing the fundamental beam may be required. Crystal length and clear

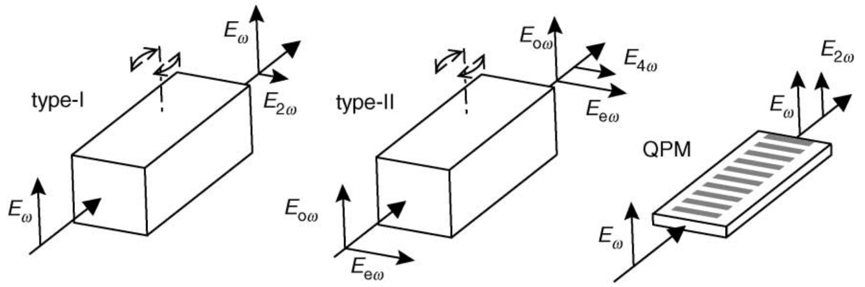

Figure 1 Different methods of phase matching optical harmonic generation are illustrated. The arrows indicate the directions of electric-field polarization and directions of propagation.

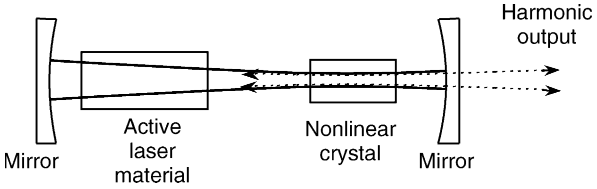

Figure 2 Placing a harmonic-generation crystal inside a laser cavity with mirrors highly reflecting at the laser wavelength is one technique of increasing the intensity of the fundamental beam.

aperture need to be selected. Intracavity harmonic generation (Figure 2), or external-resonant-cavity harmonic generation, may provide advantages in other cases. Analysis of harmonic generation provides the means to select among the options.

Basic Equations

Much of the analysis of second-harmonic generation follows from the differential equations that describe the interaction between the fundamental and harmonic fields, and it is appropriate to describe the development of these coupled equations. The wave equation:

$$\nabla^2\mathbf{E}(\mathbf{r},t) - \mu_0\sigma\frac{\partial \mathbf{E}(\mathbf{r},t)}{\partial t} - \mu_0\varepsilon_0\frac{\partial^2\mathbf{E}(\mathbf{r},t)}{\partial t^2} = \mu_0\frac{\partial^2\mathbf{P}(\mathbf{r},t)}{\partial t^2} \quad [3]$$

is obtained from Maxwell's equations. Here $\mu_0 = 4\pi\times 10^{-7}$ is the permeability of free space, and σ is the conductivity, which later will be incorporated into the optical absorption coefficient $\alpha = \mu_0\sigma c/2$, and $(\varepsilon_0\mu_0)^{-1/2} = c \cong 2.998\times 10^8$ m/s is the speed of light in vacuum. The electric field of a planewave, at angular frequency ω, is expressed in complex notation as the product of a slowly varying complex amplitude and a carrier wave:

$$\mathbf{E}_\omega(\mathbf{r},t) = (1/2)\{\mathbf{E}_\omega \exp \mathrm{i}(\mathbf{k}_\omega\cdot\mathbf{r} - \omega t) + \mathrm{c.c.}\} \quad [4]$$

The intensity of this plane wave is

$$I_\omega = (nc\varepsilon_0/2)|E_\omega|^2 \quad [5]$$

Here n is the index of refraction of the material in which E is specified. The nonlinear polarization resulting from the second-order susceptibility driven by the plane wave $\mathbf{E}_\omega(\mathbf{r},t)$ is expressed as

$$\mathbf{P}^{\mathrm{NL}}_{2\omega}(\mathbf{r},t) = (1/2)\{\mathbf{P}^{\mathrm{NL}}_{2\omega} \exp \mathrm{i}(2\mathbf{k}_\omega\cdot\mathbf{r} - 2\omega t) + \mathrm{c.c.}\} \quad [6]$$

The individual components of the polarization amplitude vector are

$$P^{\mathrm{NL}}_{2\omega,i} = \varepsilon_0\sum_{j,k} d_{ijk}(-2\omega;\omega,\omega)E_{\omega,j}E_{\omega,k} \quad [7]$$

If the nonlinear polarization is produced by two distinguishable waves, such as for sum-frequency generation of the third harmonic by the second harmonic and the fundamental, the nonlinear polarization is

$$\mathbf{P}^{\mathrm{NL}}_3(\mathbf{r},t) = (1/2)\big\{\mathbf{P}^{\mathrm{NL}}_3 \exp \mathrm{i}[(\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} - (\omega_1+\omega_2)t] + \mathrm{c.c.}\big\} \quad [8]$$

where

$$P^{\mathrm{NL}}_{3,i} = 2\varepsilon_0\sum_{j,k} d_{ijk}(-\omega_3;\omega_1,\omega_2)E_{1,j}E_{2,k} \quad [9]$$

and

$$\omega_3 = \omega_1 + \omega_2 \quad [10]$$

A number of simplifications are used to arrive at the coupled equations for harmonic generation with monochromatic plane waves. All waves are assumed to be propagating collinearly in the z direction. Only the vector components involved in the interaction are retained, and the nonlinear optical tensor is replaced by the scalar effective nonlinear coefficient d_{eff}. Finally, the slowly varying envelope approximation is invoked and the higher-order derivatives of the amplitudes are ignored, as are time derivatives of the envelopes. The wave equation simplifies to

$$\frac{\mathrm{d}E(z)}{\mathrm{d}z}+\alpha E(z)=\mathrm{i}\frac{\mu_0 c\omega}{2n}P(z) \qquad [11]$$

where n is the index of refraction.

With three distinguishable waves, as with sum-frequency generation and type-II second-harmonic generation, there are three coupled equations:

$$\frac{\mathrm{d}E_3(z)}{\mathrm{d}z}+\alpha_3 E_3(z)=\mathrm{i}\frac{\omega_3 d_{\text{eff}}}{n_3 c}E_1(z)E_2(z)\exp(-\mathrm{i}\Delta kz) \qquad [12]$$

$$\frac{\mathrm{d}E_1(z)}{\mathrm{d}z}+\alpha_1 E_1(z)=\mathrm{i}\frac{\omega_1 d_{\text{eff}}}{n_1 c}E_3(z)E_2^*(z)\exp(\mathrm{i}\Delta kz) \qquad [13]$$

$$\frac{\mathrm{d}E_2(z)}{\mathrm{d}z}+\alpha_2 E_2(z)=\mathrm{i}\frac{\omega_2 d_{\text{eff}}}{n_2 c}E_3(z)E_1^*(z)\exp(\mathrm{i}\Delta kz) \qquad [14]$$

Here the wave vector mismatch is

$$\Delta k=k_3-k_1-k_2 \qquad [15]$$

and

$$\omega_3=\omega_1+\omega_2 \qquad [16]$$

with the convention of $\omega_1 \le \omega_2 < \omega_3$. The indices of refraction of the three waves are n_1, n_2, and n_3, respectively, and α_1, α_2, and α_3 are the respective amplitude absorption coefficients.

With type-I second-harmonic generation, the fundamental field consists of only a single eigen polarization, and there are two coupled equations:

$$\frac{\mathrm{d}E_{2\omega}(z)}{\mathrm{d}z}+\alpha_{2\omega}E_{2\omega}(z)=\mathrm{i}\frac{\omega d_{\text{eff}}}{n_{2\omega}c}E_\omega(z)E_\omega(z)\exp(-\mathrm{i}\Delta kz) \qquad [17]$$

$$\frac{\mathrm{d}E_\omega(z)}{\mathrm{d}z}+\alpha_\omega E_\omega(z)=\mathrm{i}\frac{\omega d_{\text{eff}}}{n_\omega c}E_\omega(z)E_\omega^*(z)\exp(\mathrm{i}\Delta kz) \qquad [18]$$

Indices of refraction at the harmonic and fundamental are $n_{2\omega}$ and n_ω respectively; $\alpha_{2\omega}$ and α_ω are the respective absorption coefficients; and the wavevector mismatch is

$$\Delta k=k_{2\omega}-2k_\omega \qquad [19]$$

The equations for type-II second-harmonic generation, where $\omega_1=\omega_2$ and $\omega_3=2\omega_1$, become equivalent for those for type-I second-harmonic generation, providing intensities of the two fundamental polarizations are equal, $I_1=I_2$ and $\alpha_\omega=\alpha_{2\omega}$.

Steady-State Solutions

Three monochromatic planewave solutions for type-I second-harmonic generation are discussed. These solutions address type-I harmonic generation; first for low levels of conversion with arbitrary phase mismatch; next for high levels of conversion with perfect phase matching; and last for arbitrary levels of conversion with arbitrary amounts of phase mismatch. Near-field and focused-beam approximations also exist. Beyond that, numerical techniques, such as split-step Fourier transform methods can handle more complicated conditions. The simple approximations of monochromatic planewaves, or steady-state solutions, however, yield useful information.

The coupled equations for type-I second-harmonic generation can be solved by simple integration for low levels of fundamental depletion, where $E_\omega(z)$ is treated as constant and there is insignificant absorption. Initial conditions are $I_\omega(0)$, incident on the crystal and $I_{2\omega}(0)=0$. Integration of the equation for the harmonic electric field yields an expression for electric field after propagating a distance of l in the crystal:

$$E_{2\omega}(l)=\frac{\omega d_{\text{eff}}E_\omega^2(0)}{n_{2\omega}c}\frac{1}{\Delta k}\left\{\exp\mathrm{i}(k_{2\omega}l)-\exp\mathrm{i}(2k_\omega l)\right\} \qquad [20]$$

Expressed as intensity the solution is

$$I_{2\omega}(l)=I_\omega(0)\Gamma^2 l^2\,\mathrm{sinc}^2(\Delta kl/2) \qquad [21]$$

where $\mathrm{sinc}(x)=\sin(x)/x$ and

$$\Gamma^2=2\omega^2 d_{\text{eff}}^2 I_\omega(0)/(n_{2\omega}n_\omega^2 c^3\varepsilon_0) \qquad [22]$$

When $\Delta k=0$ and $\alpha_\omega=\alpha_{2\omega}=0$, conditions of perfect phase matching and no absorption, the coupled equations for type-I second-harmonic generation can be combined into a single equation. This is done using conservation of energy, which requires $E_\omega^2(z)+E_{2\omega}^2(z)=E_\omega^2(0)$, in this case with no initial harmonic intensity. The equation can then be manipulated into a standard integral that yields the inverse hyperbolic tangent. Rewritten and expressed in terms of intensity the result is

$$I_{2\omega}(l) = I_{\omega}(0)\tanh^2(\Gamma l) \qquad [23]$$

The two solutions for $I_{2\omega}(l)$ given above serve well for a number of applications. The $\mathrm{sinc}^2(\Delta k \cdot l/2)$ function is widely used to determine phase-matching acceptances. The $\tanh^2(\Gamma l)$ function is a good approximation for local conversion with large, well-collimated fundamental beams.

An exact solution to the coupled equations, however, gives insight to conditions such as high initial fundamental intensity with small amounts of phase mismatch. After extensive algebraic manipulation, a solution in the form of a Jacobian elliptic function $\mathrm{sn}(u|m)$ is obtained:

$$I_{2\omega}(l) = I_{\omega}(0)v_{\mathrm{b}}^2\,\mathrm{sn}^2\left(\frac{\Gamma l}{v_{\mathrm{b}}}\middle| v_{\mathrm{b}}^4\right) \qquad [24]$$

where

$$1/v_{\mathrm{b}} = \Delta s/4 + \sqrt{1 + (\Delta s/4)^2} \qquad [25]$$

$$\Delta s = \Delta k/\Gamma$$

and Δk and Γ retain the earlier definitions. This solution can be extended to provide the phase of the fundamental and harmonic as the conversion process evolves. Expressing the complex harmonic and fundamental field amplitudes as the product of a real amplitude and a complex exponential:

$$E(z) = A(z)\exp \mathrm{i}\phi(z) \qquad [26]$$

the harmonic phase is

$$\phi_{2\omega}(z) = (\pi - \Delta k z)/2 \qquad [27]$$

The fundamental phase, however, is found to be dependent on the conversion process:

$$\begin{aligned}\phi_{\omega}(z) = {} & (1/2)\arccos[A_{2\omega}(z)\Delta k/(2\kappa A_{\omega}^2(z))] \\ & - (\pi - \Delta k z)/z \end{aligned} \qquad [28]$$

where $\kappa = \omega d_{\mathrm{eff}}/(n_{\omega}c)$. Small amounts of phase mismatch significantly change the phase of the fundamental at high levels of conversion. This phase change is used for switching applications in waveguide interferometers. It also makes phase-matching requirements more stringent at high levels of conversion. The three planewave solutions are illustrated in Figure 3. The narrowing of phase matching at higher conversion is illustrated in Figure 4.

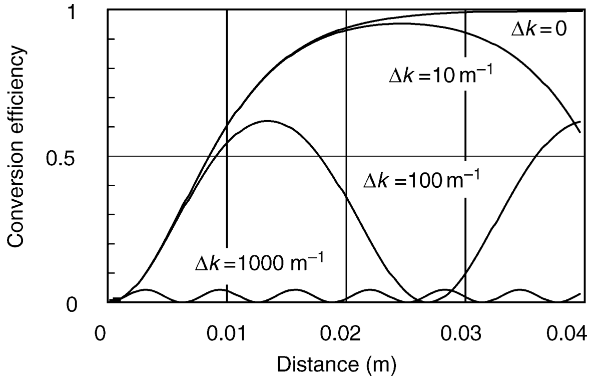

Figure 3 Monochromatic plane-wave solutions of the coupled equations for second-harmonic generation are illustrated for different amounts of wavevector mismatch. The solution takes the form $\tanh^2(\Gamma l)$ when $\Delta k = 0$ and $(\Gamma l)^2\mathrm{sinc}^2(\Delta k l/2)$ when conversion is small.

There is a Jacobian elliptic function solution for intensities with the coupled waves of type-II second-harmonic generation and sum-frequency generation. The phases of the individual waves can be recovered with elliptic integrals of the third kind. Both sets of equations, the two coupled equations for type-I phase matching and the three coupled equations for type-II phase matching, are amenable to numerical solution by Runge–Kutta methods.

Phase Matching

Phase matching is achieved in birefringent materials by balancing the dispersion of the material with the difference between the indices of refraction of the fast and slow waves. With adequate birefringence it may be possible to obtain phase matching by selecting the direction of propagation in the crystal. Equations describing the dispersion and birefringence of nonlinear crystals, necessary for calculating phase matching, along with the nonlinear optical coefficients and expressions for the effective nonlinear optical coefficients, are tabulated in handbooks. Values for a few examples are given here.

The negative uniaxial crystal potassium dihydrogen phosphate (KH_2PO_4 or KDP) is used as an illustrative example. The harmonic wave is chosen to have extraordinary polarization with the smaller index of refraction in this crystal with negative birefringence. With type-I phase matching, the fundamental consists of only an ordinary polarized wave. To have phase matching, $\Delta k = 0$, it is necessary to choose a direction of propagation that makes an angle θ with the optic axis that satisfies:

$$n_{\mathrm{e},2\omega}(\theta) = n_{\mathrm{o},\omega} \qquad [29]$$

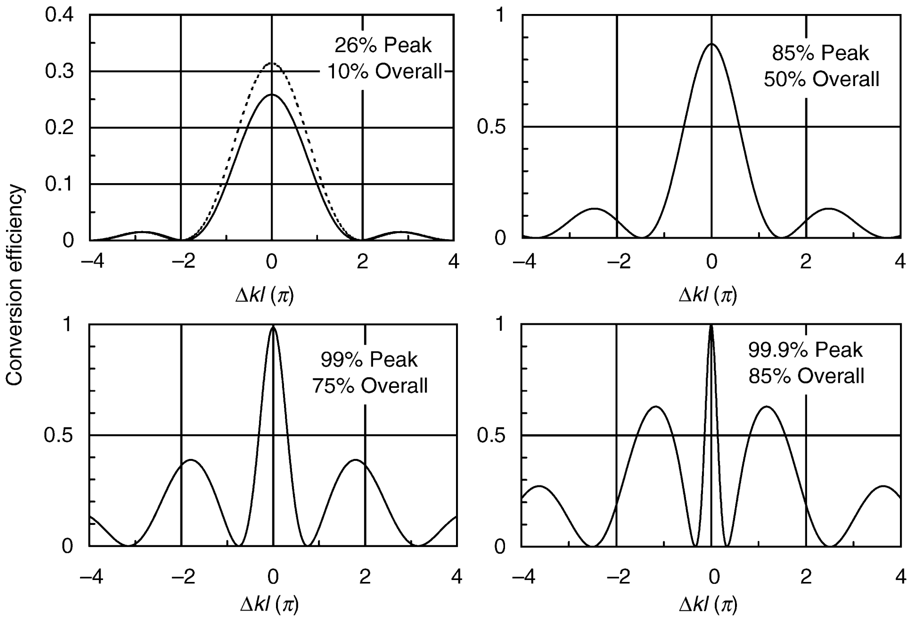

Figure 4 High levels of overall second-harmonic conversion require higher levels of conversion at peak pulse intensity. Phase-matching bandwidths narrow at the higher levels of peak conversion. The broken curve is the $(\Gamma l)^2\text{sinc}^2(\Delta kl/2)$ small conversion approximation.

The extraordinary index for propagation at an angle of θ to the optical axis in a uniaxial crystal $n_e(\theta)$ is given by

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \qquad [30]$$

where n_o is the ordinary index and n_e is the extraordinary index for propagation at 90° to the optic axis. The phase-matching angle is obtained by using the identity $\cos^2\theta = 1 - \sin^2\theta$ and solving for θ:

$$\theta_{\text{PM}} = \sin^{-1}\left\{\frac{n_{e,2\omega}^2(n_{o,2\omega}^2 - n_{o,\omega}^2)}{n_{o,\omega}^2(n_{o,2\omega}^2 - n_{e,\omega}^2)}\right\}^{1/2} \qquad [31]$$

where the subscripts refer to ordinary and extraordinary indices at the fundamental and second-harmonic frequencies.

The effective nonlinear coefficient is determined by crystal symmetry, the type of phase matching, the birefringence of the crystal, and the direction of propagation. For crystals of point group $\bar{4}2m$ with negative birefringence, such as KDP, the effective nonlinear coefficient for type-I phase matching is given by

$$d_{\text{eff}} = -d_{36}\sin\theta\sin 2\phi \qquad [32]$$

The azimuthal angle ϕ is measured from the crystal x-axis. The x-axis and y-axis are specified by X-ray diffraction and piezoelectric coefficients and can be determined from the growth morphology. The azimuthal orientation of a finished crystal, however, is not obvious, and requires X-ray orientation, piezoelectric characterization, or nonlinear optical measurements, if the record of crystal orientation is lost. The nonlinear coefficient d_{36} is an abbreviation for the d_{312} and d_{321} tensor components, which are equal. For harmonic generation with a 1064 nm fundamental in KDP, $d_{36} = 0.39 \times 10^{-12}$ m/V. The unit of 10^{-12} m/V is usually expressed as picometers per volt (pm/V). The effective nonlinear coefficient for type-II phase matching for this class of crystals is

$$d_{\text{eff}} = d_{36}\sin 2\theta\cos 2\phi \qquad [33]$$

Type-I and type-II crystals cut from the same material are not interchangeable. The type-II phase-matching angle for KDP can be found by iterative solution of the equation:

$$n_{e,2\omega}(\theta) = \{n_{o,\omega} + n_{e,\omega}(\theta)\}/2 \qquad [34]$$

The birefringent walkoff angle, that is the angle the Poynting vector of an extraordinary wave makes with the wavevector, is another consideration for birefringent phase matching. There will be no further harmonic generation when the extraordinary and ordinary fundamental beams of type-II phase matching no longer overlap. For type-I phase matching, birefringent walkoff can limit the interaction length, but there can be several interaction lengths in a crystal. The birefringent walkoff of an extraordinary wave propagating at an angle θ to the optic axis in a uniaxial crystal is given by

$$\rho = \tan^{-1}\left\{\frac{n_e^2(\theta)}{2}\left[\frac{1}{n_e^2}-\frac{1}{n_o^2}\right]\sin(2\theta)\right\} \qquad [35]$$

Refractive indices for 1064 nm to 532 nm second-harmonic generation in KDP are $n_{e,2\omega} = 1.4709$, $n_{o,2\omega} = 1.5129$, $n_{e,\omega} = 1.4603$, and $n_{o,\omega} = 1.4944$. The dispersion of KDP and indices at the phase-matching angles are shown in Figure 5. For type-I phase matching $d_{\text{eff}} = 0.26$ pm/V, $\theta_{\text{PM}} = 41.2°$, and $\rho_{2\omega} = 28$ milliradians (mr), and for type-II phase matching $d_{\text{eff}} = 0.34$ pm/V, $\theta_{\text{PM}} = 59.1°$, $\rho_{\omega} = 20$ mr, and $\rho_{2\omega} = 25$ mr. Type-II phase matching in KDP for this application offers an advantage of larger effective nonlinear coefficient. Also the angular acceptance is larger for type-II phase matching.

The angular acceptance is calculated from dispersion equations by first evaluating the derivative $\partial\Delta k/\partial\theta$ and then finding the value of $\Delta\theta$ corresponding to the full width at half maximum (FWHM) of the $\text{sinc}^2(\Delta k\cdot l/2)$ function:

$$\frac{\partial\Delta k}{\partial\theta}\Delta\theta_{\text{FWHM}} \cong 4\times 1.39\ \text{radians}/l \qquad [36]$$

When θ is 0° or 90° for uniaxial crystals, or for propagation along one of the refractive index principal axes in a biaxial crystals, the first derivative $\partial\Delta k/\partial\theta$ becomes zero, and the angular acceptance must be calculated from the second derivative $\partial^2\Delta k/\partial\theta^2$. This case is called noncritical phase matching (NCPM), and the case for θ between 0° and 90° is called critical phase matching.

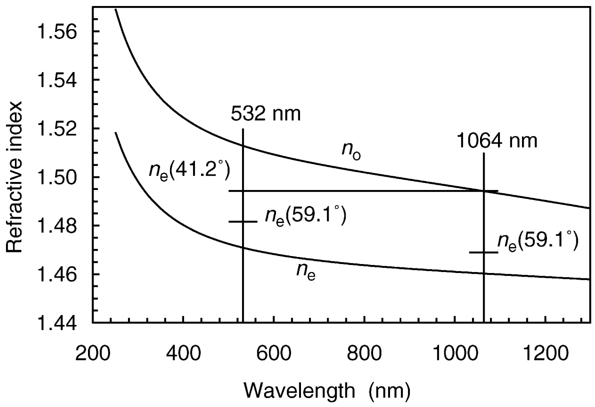

Figure 5 Dispersion of the birefringent nonlinear material KH_2PO_4 with refractive indices indicated for type-I and type-II 1064 nm to 532 nm second-harmonic generation. (Adapted from Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.)

Second-harmonic generation of 1064 nm fundamental radiation in the crystal lithium triborate (LBO or LiB_3O_5) is an example of NCPM. At 149°C, this crystal will phase match for type-I SHG for propagation along the x-axis in the xyz frame determined by the indices of refraction of this biaxial crystal $n_x < n_y < n_z$. The following are calculated from dispersion equations for this case: $\partial\Delta k/\partial\theta = 0$; $\partial^2\Delta k/\partial\theta^2 = 2.04\times 10^4$ (rad/cm)/rad^2; $\partial\Delta k/\partial\phi = 0$; and $\partial^2\Delta k/\partial\phi^2 = 6.45\times 10^3$ (rad/cm)/rad^2. The FWHM angular acceptances for a 1.5 cm-long LBO crystal are $\Delta\theta_{\text{FWHM}} = 1.09°$ and $\Delta\phi_{\text{FWHM}} = 1.94°$. Angle-tuned CPM at 23°C in LBO occurs at $\theta = 90°$ and $\phi = 11°.5$. For this case $\partial\Delta k/\partial\phi = 1.32\times 10^4$ (rad/cm)/rad, and the azimuthal angular acceptance decreases to $\Delta\phi_{\text{FWHM}} = 0.16°$. Also there is a birefringent walkoff angle of $\rho_{2\omega} = 19$ mr, which results in a displacement of 0.29 mm of the extraordinary-like harmonic beam for a 1.5 cm-long crystal. Both the larger angular acceptance and the absence of birefringent walkoff are advantages for harmonic generation. The effective nonlinear coefficient of LBO is reported as 0.85 pm/V and 1.04 pm/V by different sources.

Quasi phase matching offers the possibility, in principle, of noncritical phase matching at any wavelength in the transmission range of a material. However, it may be difficult producing the periodic structure that is required for phase matching in a particular application. The ferroelectric crystals lithium niobate ($LiNbO_3$), lithium tantalate ($LiTaO_3$), and potassium titanyl phosphate ($KTiOPO_4$ or KTP) can be processed by electric-field poling to produce domains of alternating polarity. Lithography techniques are used to apply a patterned electrode structure on the surface of a wafer or thin slab. Domains are reversed by the application of a carefully monitored and controlled electric field. Thicknesses of several mm have been achieved in periodically poled KTP (PPKTP). The higher coercive field of 21 kV/mm has limited the thicknesses that can be achieved in periodically poled lithium niobate (PPLN), particularly at shorter wavelengths for which poling periods become small. Progress is now being made with zincblende materials such as gallium arsenide. These materials have large linearity and good transmission in the infrared, but they are not birefringent. Epitaxial growth on patterned substrates is now being used to produce periodically structured GaAs on a research basis, and bonded stacks of wafers of alternating orientation also have been used. Harmonic conversion with quasi phase matching is compared with conventional phase matching and nonphase-matched harmonic generation in Figure 6.

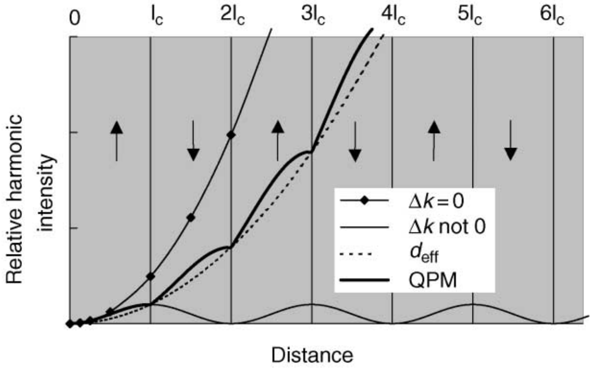

Figure 6 The growth of harmonic intensity is illustrated schematically for a quasi-phase-matched process in which the polarity of the nonlinearity is reversed after each coherence length. Conversion for $\Delta k = 0$, $\Delta k \neq 0$, and a uniform effective nonlinear coefficient are shown for comparison. (Adapted from Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.)

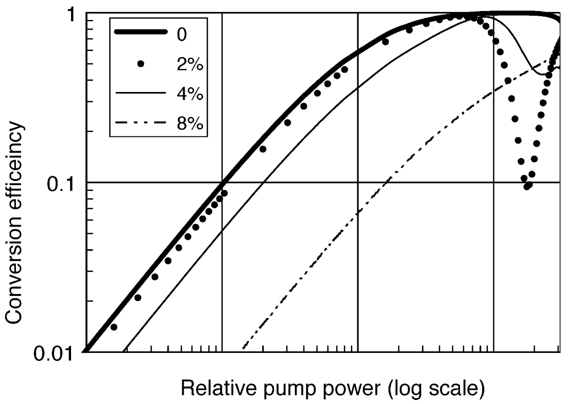

Figure 7 Random variations in domain length will lower harmonic conversion efficiency and have effect on back conversion. RMS variations in domain length are indicated in the legend for this specific example.

Quasi phase matching also provides access to tensor components that are not accessible with birefringent phase matching. This is the case for lithium niobate; the largest tensor component is $d_{333} = d_{33}$, which has a value of 25 pm/V. The effective nonlinear coefficient is the relevant Fourier component of the modulated nonlinearity:

$$d_{\text{eff}} = (2/\pi) d_{33} \approx 16 \text{ pm/V} \qquad [37]$$

This is a larger effective nonlinearity than is available with common birefringently phase matched materials in the visible and near-infrared. A more accurate treatment of second-harmonic generation involves application of the Jacobian-elliptic-function solution to each domain. After nearly complete conversion is reached, the phase perturbations on the individual domains have effect, and back conversion takes over (Figure 7). Conversion is lower and back conversion can set in sooner with random differences in the domain length. There is some experimental evidence that this type of back conversion has been observed in guided-wave harmonic conversion devices. Guided-wave nonlinear optical interaction can have long interaction lengths not limited by diffraction. Quasi phase matching is particularly useful in waveguide second-harmonic generation because the direction of propagation is fixed.

Gaussian Beams and Pulses

A radially symmetric Gaussian light beam has transverse intensity of

$$I(r) = I_0 \exp(-2r^2/w^2) \qquad [38]$$

where r is the radial coordinate and w is called the spot size. Harmonic generation in the low-conversion limit by a Gaussian beam, with a focus of spot size w_0 at the center of a crystal with optimized phase matching, is described by

$$p_{2\omega}(l)/p_\omega(0) = 2\omega^2 d_{\text{eff}}^2 p_\omega(0) l k h(B, \xi)/(\pi n^3 \varepsilon_0 c^3) \qquad [39]$$

The harmonic power after propagating distance l in the crystal is $p_{2\omega}(l)$, and $p_\omega(0)$ is the incident fundamental power. The function $h(B, \xi)$ is the Boyd and Kleinman focusing factor. It is a function of the walkoff parameter:

$$B = \rho\sqrt{lk}/2 \qquad [40]$$

and the focusing parameter:

$$\xi = l/(k w_0^2) \qquad [41]$$

The spot size w_0 of the fundamental beam at its waist is used to described the fundamental intensity distribution at the focus at the center of the crystal:

$$I_\omega(r) = (2p_\omega(0)/(\pi w_0^2))\exp(-2r^2/w_0^2) \qquad [42]$$

Values of the focusing factor for arbitrary focusing and walkoff are usually displayed in the form of a graph (Figure 8). In the case of weak focusing $\xi \ll 1$ and relatively small walkoff $w_0 > \rho l$, the focusing factor is given by

$$h(B, \xi) \approx \xi(1 - t^2/12 + t^4/120 - t^6/1344 + \ldots) \qquad [43]$$

where $t = 2B(2\xi)^{1/2}$.

An example of continuous-wave second harmonic generation is used. This calculation models a conversion experiment published in the literature. External-resonant-cavity second-harmonic generation of 6.5 W of 532 nm radiation was reported. The 1064 nm fundamental radiation with 378 W of circulating power in a resonant cavity external to the laser was focused to a 32 μm spot size in the center of a 6 mm-long LBO crystal (Figure 9). This corresponds to a focusing factor of $\xi = 0.62$ in the index $n = 1.6053$ material. The value of the focusing parameter is $h(0, 0.62) = 0.58$. The nonlinear optical coefficient for this process is reported as 0.85 pm/V and 1.04 pm/V by different sources. When these values are inserted into the earlier equation for $p_{2\omega}(l)/p_{\omega}(0)$, the higher coefficient yields a conversion efficiency of 0.0273 and 10.3 W of 532 nm output, whereas the lower coefficient yields a conversion efficiency of 0.0182 and 6.9 W of 532 nm radiation. Differences in published values of nonlinear optical coefficients are common. The reported precision of the experiment and the measured values tend to support the lower value. Practical applications are useful to check the accuracy of parameter values. With many variables and the added complication of operation at high intensity there is uncertainty that demands caution in accepting the derived parameter values. It is interesting to note that the power and beam size of the example result in a maximum intensity of 17 MW/cm^2 at the surface of this crystal. Many vendors will warranty antireflection coatings for only 1 MW/cm^2 maximum continuous-wave intensity.

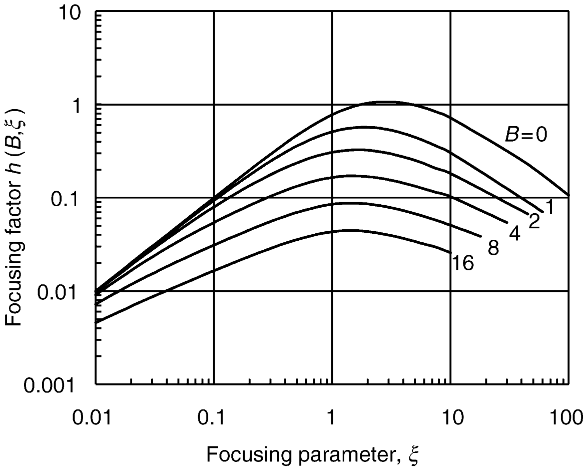

Figure 8 The Boyd and Kleinman focusing factor $h(B,\xi)$ for type-I second-harmonic generation is shown as a function of the focusing parameter ξ for different values of the walkoff parameter B. The curves shown were generated with a split-step Fourier transform calculation. (Adapted from Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.)

For a pulsed fundamental with temporal shape $p_{\omega}(t) = p_{\omega}(0)\exp(-t^2/\tau^2)$, integration over time gives a calculated energy conversion efficiency of:

$$\eta_{\text{calc}} = U_{2\omega}(l)/U_{\omega}(0)$$

$$= \sqrt{2}\omega^2 d_{\text{eff}}^2 U_{\omega}(0) lkh(B,\xi)/(\tau\sqrt{\pi^3}n^3\varepsilon_0 c^3) \quad [44]$$

where $U_{2\omega}(l)$ is the harmonic energy generated after propagation distance l in the crystal and $U_{\omega}(0)$ is the incident fundamental energy. This is still a low conversion approximation. An estimate of conversion efficient at arbitrary levels of depletion η_{est} is obtained with the empirical expression:

$$\eta_{\text{est}} = \eta_{\text{calc}}/(1 + \eta_{\text{calc}}) \quad [45]$$

Returning to the example 1.064 nm to 532 nm, type-II SHG in KDP and using $l = 3$ cm, $U_{\omega}(0) = 0.3$ J, and $\tau = 3.6$ ns for a 6 ns FWHM pulse duration, the above two relationships predict a conversion efficiency of 50%. These parameters give a peak intensity of 330 MW/cm^2, a value typical for efficient harmonic conversion, while avoiding laser-induced damage in a long-term, repetitive-pulse application using KDP. Single-shot laser-induced-damage thresholds for KDP are about 10 times higher.

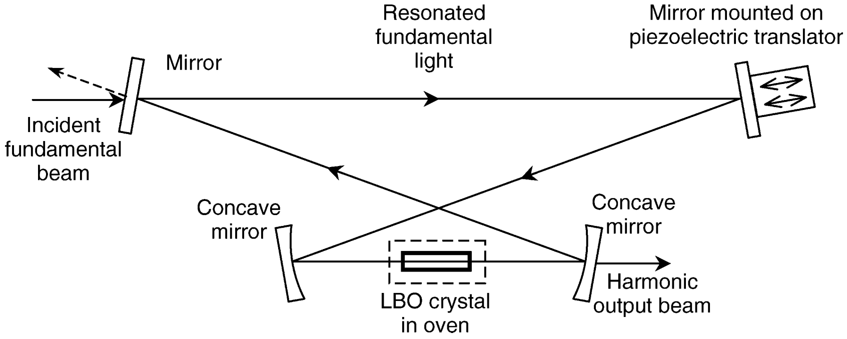

Figure 9 A bow-tie cavity for externally resonant second-harmonic generation is illustrated. The reflected fundamental beam is used to develop an error signal to drive the piezoelectric translator and hold the cavity in resonance.

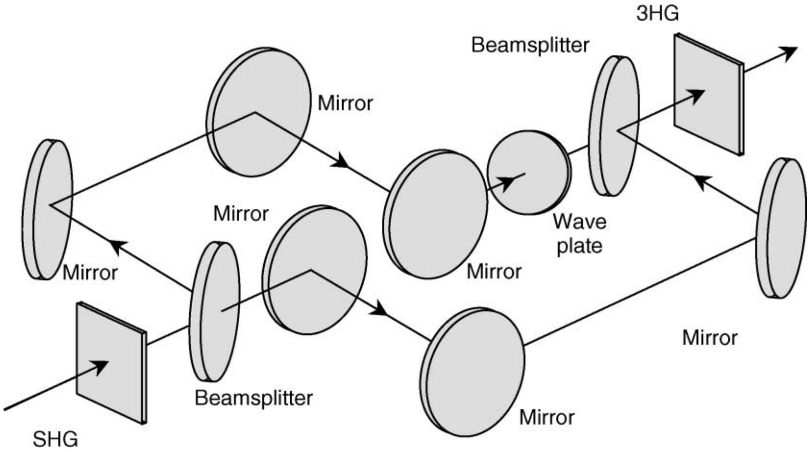

Figure 10 Short crystals less than 1 mm in length are typically used for second-harmonic generation with fundamental pulses that are only ten's of femtoseconds in duration. It is necessary to separate fundamental and second-harmonic pulses and recombine them with appropriate delay to compensate for group-velocity mismatch before sum-frequency generation to the third harmonic.

As a final example, consider second-harmonic generation of a 100 fs ($1\ \text{fs} = 10^{-15}$ s) duration pulse from a mode-locked $Ti:Al_2O_3$ laser operating at 800 nm. Typical laser output might consist of 10 nJ (10×10^{-9} Joule) pulses at 80 MHz repetition rate. Group velocity walkoff is a concern for this short pulse duration (Figure 10). Barium borate (BaB_2O_4 or BBO) is a common crystal for this application. The group velocity walkoff rate between the ordinary-polarization 800 nm fundamental pulse and the extraordinary-polarization 400 nm harmonic pulse is 190 fs/cm. A crystal length no longer than 0.5 mm is needed to avoid lengthening the harmonic pulse in time. Fortunately, the laser-induced-damage-threshold intensity increases approximately as the inverse of the square root of the pulse length between roughly 10 ns and 1 ps, and it is possible to focus the beam to achieve higher intensity without crystal damage and compensate for the short crystal length. However, the 68-milliradian birefringent walkoff of the harmonic will limit the size of the focus. A 20 μm spot size will allow an interaction length that roughly matches the crystal length and provides a peak fundamental intensity around 16 GW/cm^2 and a conversion efficiency of nearly 30%. Now suppose the 100 fs pulse is amplified to 100 mJ in a large chirped-pulse regenerative amplifier. A larger diameter crystal is required, and KDP is an appropriate choice. The group velocity walkoff rate in KDP is 78 fs/mm, and a crystal of 1.5 mm thickness could be used. With the beam spot size increased to 30 mm, the peak intensity will be 71 GW/cm^2, and the conversion efficiency will be about 40%. Material availability, such as the large sizes possible with KDP, often determines a choice.

Conclusion

Basic techniques for describing and modeling second-harmonic generation have been discussed. Birefringent phase matching and quasi phase matching were reviewed. Three examples were used to illustrate considerations for implementing harmonic-generation systems. There are many more parameters to be considered for other applications and for more detailed analysis. The examples presented illustrate the demands nonlinear optical frequency conversion places on materials. A few materials have been established as primary choices in common applications. But for many applications, available materials or laser system performance are limiting and it is necessary to use the materials and systems that are available. The importance of using the analysis methods is stressed. It is possible to sacrifice performance by improperly specifying a crystal. Choosing the best crystal for a harmonic-generation application with an appropriate optical design will optimize harmonic conversion.

See also

Nonlinear Optics, Applications: Phase Matching. **Nonlinear Sources:** Harmonic Generation in Gases. **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Akhmanov SA (1975) Optical harmonic generation and optical frequency multipliers. In: Rabin H and Tang CL (eds) *Quantum Electronics. A Treatise, vol. 1, part B*, pp. 475–583. New York: Academic Press.

Armstrong JA (1962) Interaction between lightwaves in a nonlinear dielectric. *Physics Review* 127: 1918–1962.

Boyd GD (1968) Parametric interactions of focused Gaussian light beams. *Journal of Applied Physics* 39: 3597–3639.

Boyd RW (2003) *Nonlinear Optics*, 2nd edn. New York: Academic Press.

Byer RL (1977) Parametric oscillators and nonlinear materials. In: Harper PG and Wherrett BS (eds) *Nonlinear Optics*, pp. 47–159. San Francisco, CA: Academic Press.

Dmitriev VG (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn. Berlin: Springer-Verlag.

Fejer MM (1992) Quasi-phase-matched second harmonic generation: tuning and tolerances. *IEEE Journal of Quantum Electronics* 28: 2631–2654.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (2003) *Handbook of Nonlinear Optics*, 2nd edn. New York: Marcel Dekker.

Yang SL (1991) 6.5 W, 532 nm radiation by cw resonant external-cavity second-harmonic generation of an 18 W Nd:YAG laser in LiB_3O_5. *Optical Letters* 16: 1493–1495.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

$\chi^{(3)}$–Third-Harmonic Generation

B Y Soon, Thales Electro-Optics PTE Ltd., Singapore

J W Haus, University of Dayton, Dayton, OH, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The invention of the ruby laser, in 1960, provided a revolutionary new light source, which emits an intense, highly collimated polarized beam of light. The new source opened the way for developments that have enabled new technologies, thus profoundly changing our way of life. One of these new frontiers opened by the laser was the field of nonlinear optics. In 1961, Franken *et al.* were the first to report the observation of second-harmonic generation from a quartz crystal, using ruby laser light. Third-harmonic generation (THG) was soon discovered by Terhune and co-workers in 1962.

THG has found applications as a spectroscopic tool to probe optical properties of materials, but recently it has been developed as a microscopy tool and a pulse measurement technique. More efficient THG is possible by engineering the materials; for instance, one way to improve conversion efficiency is to use periodic structured dielectric materials, called photonic bandgap structures.

Polarization and the Third-Order Susceptibility

THG is a nonlinear optical phenomenon, where light of angular frequency ω is converted into light at an angular frequency 3ω that is three times higher. Light is made of particles called photons and the nonlinear medium enables the process whereby three photons of frequency ω are annihilated to create one photon of frequency 3ω. The nonlinear interactions are strongly dependent on the polarization or the dipole moment per unit volume, which is written as a series

$$P(t) = \chi^{(1)} \times E(t) + \chi^{(2)} : E^2(t) + \chi^{(3)} \vdots E^3(t) + \cdots \qquad [1]$$

where χ is the susceptibility of the optical material, the superscript of χ in parentheses denotes the order of nonlinearity, t is the time variable, and E is the electric field. Both the electric field and the polarization are vectors; therefore the susceptibilities are tensors. The first-order or linear susceptibility $\chi_{ij}^{(1)}$ is a second-order tensor. This term constitutes the linear polarization of the medium denoted a P^{L}. It has nine elements, but only up to three independent elements depending on the symmetry of the lattice. The other terms represent an expansion for the nonlinear polarization of the medium, P^{NL}. Note that the superscripts 'L' and 'NL' denote linear and nonlinear polarization contributions, respectively. The third-order susceptibility $\chi_{ijkl}^{(3)}$ is a four-rank tensor, which usually has 81 elements ($3 \times 3 \times 3 \times 3 = 81$). In asymmetric crystalline solids, all 81 elements can be of nonzero value and independent. With a higher degree of symmetry in other optical materials, the 81 elements are no longer all independent, and thus the number of independent elements can be reduced to a lower number. For example, isotropic material such as glass, liquid, or vapor has only three independent elements. For illustration, we consider a general case where the applied frequencies ω are arbitrary, and the third-susceptibility is given as

$$\chi_{ijkl} \equiv \chi_{ijkl}^{(3)}(\omega_4 = \omega_1 + \omega_2 + \omega_3) \qquad [2]$$

Note that the superscript has been removed. In the isotropic material, the coordinate axes must be equivalent. Thus, the susceptibility has symmetry given by

$$\chi_{1111} = \chi_{2222} = \chi_{3333} \qquad [3a]$$

$$\chi_{1122} = \chi_{1133} = \chi_{2211} = \chi_{2233} = \chi_{3311} = \chi_{3322} \qquad [3b]$$

$$\chi_{1212} = \chi_{1313} = \chi_{2323} = \chi_{2121} = \chi_{3131} = \chi_{3232} \qquad [3c]$$

$$\chi_{1221} = \chi_{1331} = \chi_{2112} = \chi_{2332} = \chi_{3113} = \chi_{3223} \qquad [3d]$$

For the isotropic material, there are only 21 nonzero elements, as given in eqns [3a]–[3d], because these elements have a Cartesian index (1, 2, or 3) that appears an even number of times. The reason for nonexistence of the elements with an index appearing an odd number of times is because, for example, χ_{1222} would give a response in the $\hat{x}_1$ direction due to a field applied in the $\hat{x}_2$ direction. This response must vanish in an isotropic material, because there is no reason why the response should be in the $+\hat{x}_1$ direction rather than in the $-\hat{x}_1$ direction. In summary, the elements in the third-order susceptibility of the isotropic material have the relationship given by

$$\chi_{ijkl} = \chi_{1122}\delta_{ij}\delta_{kl} + \chi_{1212}\delta_{ik}\delta_{jl} + \chi_{1221}\delta_{il}\delta_{jk} \qquad [4]$$

where δ is the Kronecker delta function; it has a unit value, $\delta_{ij} = 1$, if and only if $i = j$, otherwise, it is zero, $\delta_{ij} = 0$. Thus, we see from eqn [4] that there are only three independent elements in the isotropic material. Crystalline materials have more independent components and they have been determined for the 32 crystals classes. The relationships between the elements for these crystalline classes are tabulated in books by Boyd and by Butcher and Cotter.

A block diagram of a material illustrates THG and its corresponding energy-level diagram is depicted in Figure 1. Figure 1a shows that fields of frequency ω excite the χ^3 optical material to generate the third-harmonic field with the residual of the fundamental harmonic field. Figure 1b shows that three photons of frequency ω are destroyed to generate a photon of frequency 3ω in the third-harmonic generation.

For third-harmonic generation, the frequency dependence of the susceptibility is expressed as $\chi_{ijkl}(3\omega = \omega + \omega + \omega)$. With the intrinsic permutation symmetry of the nonlinear susceptibility, the elements of the susceptibility tensor has a relationship such that $\chi_{1122} = \chi_{1212} = \chi_{1221}$. So we can further reduce eqn [4] for the third-harmonic generation as

$$\chi_{ijkl}(3\omega = \omega + \omega + \omega) = \chi_{1122}(3\omega = \omega + \omega + \omega) \times (\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \qquad [5]$$

Therefore, we see that the third-harmonic susceptibility tensor has only one independent element.

Wave Equation for χ^3 Third-Harmonic Generation

Theoretically, we have to use the wave equation to observe how an intense laser field interacts with the nonlinear material, so that the polarization of the medium could develop new frequency components that are not present in the incident field. Thus, these new frequency components of the polarization become the sources of the new frequency components of the electromagnetic field.

To formally describe THG, we begin with Maxwell's equations (in Gaussian units), and we have

$$\nabla \times D = 4\pi\rho \qquad [6a]$$

$$\nabla \times B = 0 \qquad [6b]$$

$$\nabla \times E = \frac{-1}{c}\frac{\partial B}{\partial t} \qquad [6c]$$

$$\nabla \times H = \frac{-1}{c}\frac{\partial D}{\partial t} + \frac{4\pi}{c}J \qquad [6d]$$

where H is the magnetic field, B is the magnetic flux density, D is the electric flux density, ρ is the charge density, and J is the current density. The desired solution is in the region of space in the nonlinear material that has no free charge ($\rho = 0$), and therefore, there is no free current ($J = 0$). We assume

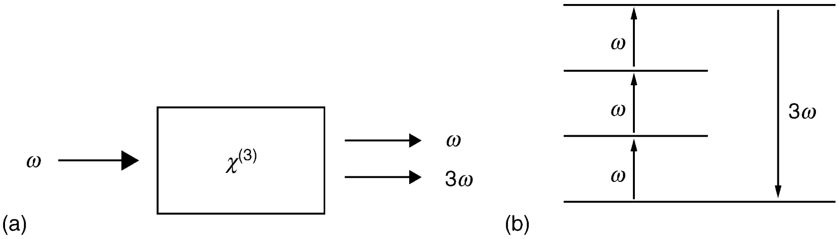

Figure 1 The third-harmonic (a) block diagram; and (b) energy level description.

that the nonlinear material is nonmagnetic, so that $B = \mu H = H$, where $\mu = 1$ in a nonmagnetic environment. In nonlinear material, the electric flux is defined as

$$D = E + 4\pi P \qquad [7]$$

so that the polarization depends nonlinearly on the strength of E. Using eqns [6c] and [6d], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 D}{\partial t^2} = 0 \qquad [8]$$

Substituting eqn [7] into [8], we have

$$\nabla \times \nabla \times E + \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{-4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \qquad [9]$$

Using an identity from the vector calculus, we have

$$\nabla \times \nabla \times E = \nabla(\nabla \times E) - \nabla^2 E \qquad [10]$$

In linear optics, we see from eqn [6a] that when there is no free charge in source-free isotropic material, $\nabla \times D = \nabla \times E = 0$. Thus, the first term on the right-hand side of eqn [10] has no effect. Whereas in nonlinear optics, $\nabla \times E \neq 0$ in isotropic material, but the contribution from the first term on the right-hand side of eqn [10] is negligible, therefore, we can ignore it. Thus, eqn [9] can be simplified as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P}{\partial t^2} \qquad [11]$$

For distinction of the linear and nonlinear contributions, we modified eqn [7] to get

$$D = E + 4\pi P^{\mathrm{L}} + 4\pi P^{\mathrm{NL}}$$

where $P = P^{\mathrm{L}} + P^{\mathrm{NL}}$. Let $D^{\mathrm{L}} = E + 4\pi P^{\mathrm{L}}$, and we rewrite eqn [11] as

$$\nabla^2 E - \frac{1}{c^2}\frac{\partial^2 D^{\mathrm{L}}}{\partial t^2} = \frac{4\pi}{c^2}\frac{\partial^2 P^{\mathrm{NL}}}{\partial t^2} \qquad [12]$$

Equation [12] is the most general form of the wave equation seen in nonlinear optics, and could be used as our starting point for our study of third-harmonic generation.

Methods of Third-Harmonic Generation

Since nonlinear optics is a popular research topic, there are a handful of existing methods that can be used for third-harmonic generation, and possibly new methods may be developed in the future. Here, we will discuss three methods using χ^3 material: (i) photonic crystal; (ii) defect mode using photonic crystal; and (iii) homogeneous nonlinear crystal.

Photonic Crystal

After the invention of electronic bandgap structure, the optical scientists and engineers saw the possibility to extend the operating waveband into the visible, and thus we have the photonic bandgap structure. The geometry and material choices of the photonic bandgap structure create a unique bandgap in frequency spectrum that has potential applications in laser optics, telecommunication, medical imaging, optical limiting, etc. A photonic bandgap structure is also commonly known as the photonic crystal. For a one-dimensional photonic crystal, two materials of certain thicknesses each, repeat themselves a number of times. For example, the first and second materials are of thickness 'a' and 'b', respectively, so that their combined thickness is 'd', and have a periodicity of 10. The schematic of such a one-dimensional photonic crystal, with its corresponding dielectric function as a function of axial distance z, is depicted in Figure 2.

By using χ^3 materials for photonic crystal, and using method of multiple scales (MMS), we have the nonlinear polarization given as

$$P^{\mathrm{NL}} = \lambda \chi^{(3)} E^3 \qquad [13]$$

where λ is the perturbation parameter. With reference to Figure 2, assume that 'a' is equal to 'b' so that $d = 2a$, then the Fourier expansion of the dielectric

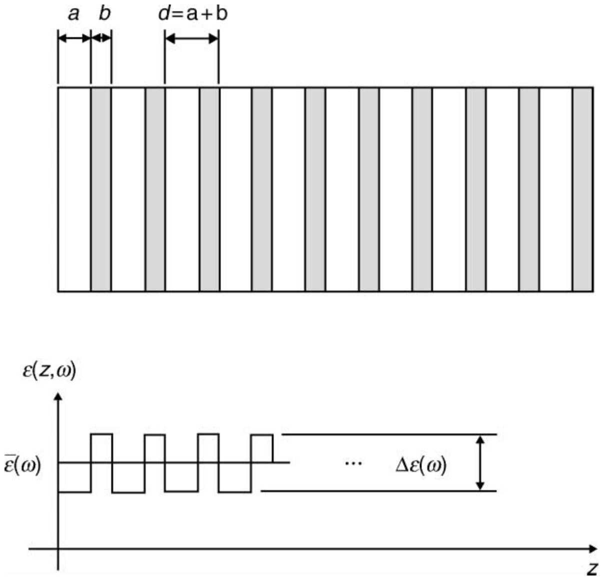

Figure 2 The schematic of the one-dimensional photonic crystal.

function is given as

$$\varepsilon(z,\omega) = \bar{\varepsilon}(\omega) + \lambda\Delta\varepsilon(\omega)\cos(2\pi z/d) - \frac{1}{3}\lambda\Delta\varepsilon(\omega)\cos(6\pi z/d) + \cdots \qquad [14]$$

where $\bar{\varepsilon}(\omega)$ is the zeroth-order coefficient, and $\Delta\varepsilon(\omega)$ is first-order coefficient of the Fourier series expansion. Without going through the mathematics of MMS, we use eqns [12] and [13] to obtain the solution of the wave equation and two sets of coupled mode equations. The solution to the wave equation in eqn [12] is given by

$$E_0 = \left(\frac{1}{\sqrt{k_1}}A_{\mathrm{f}1}(z_1,t_1)\mathrm{e}^{\mathrm{i}k_1 z_0} + \frac{1}{\sqrt{k_1}}A_{\mathrm{b}1}(z_1,t_1)\mathrm{e}^{-\mathrm{i}k_1 z_0}\right)\mathrm{e}^{-\mathrm{i}\omega t_0} + \left(\frac{1}{\sqrt{k_3}}A_{\mathrm{f}3}(z_1,t_1)\mathrm{e}^{\mathrm{i}k_3 z_0} + \frac{1}{\sqrt{k_3}}A_{\mathrm{b}3}(z_1,t_1)\mathrm{e}^{-\mathrm{i}k_3 z_0}\right)\mathrm{e}^{-3\mathrm{i}\omega t_0} + \mathrm{c.c.} \qquad [15]$$

where A is the amplitude of the field, k is the wavenumber such that $k_1^2=\omega^2\varepsilon(\omega)/c^2$, and $k_3^2=(3\omega)^2\varepsilon(3\omega)/c^2$ where c is the speed of light, and 'c.c.' refers to the complex conjugate. Note that the subscript 'f' and 'b' refer to the forward and backward modes, respectively. Also, the subscript '0' refers to the slowest scale, and '3' refers to the fastest scale chosen in the analysis using MMS. Let $a_{\mathrm{f}1}=\mathrm{N}A_{\mathrm{f}1}\mathrm{e}^{\mathrm{i}\pi\delta_1 z_1/d}$, $a_{\mathrm{b}1}=\mathrm{N}A_{\mathrm{b}1}\mathrm{e}^{-\mathrm{i}\pi\delta_1 z_1/d}$, $a_{\mathrm{f}3}=\mathrm{N}A_{\mathrm{f}3}\mathrm{e}^{\mathrm{i}\pi\delta_3 z_1/d}$, and $a_{\mathrm{b}3}=\mathrm{N}A_{\mathrm{b}3}\mathrm{e}^{-\mathrm{i}\pi\delta_3 z_1/d}$, such that

$$N^2 = \frac{2d\sqrt{3k_3}(3\omega)^2\chi^{(3)}}{k_3 c^2 k_1\sqrt{k_1}}$$

The coupled mode equations for the fundamental field are given as

$$\frac{\partial a_{\mathrm{f}1}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}1}}\frac{\partial a_{\mathrm{f}1}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{f}1} + \mathrm{i}\delta_1 a_{\mathrm{f}1} - \mathrm{i}\kappa_1 a_{\mathrm{b}1} - \frac{\alpha_1}{2}a_{\mathrm{f}1} + \mathrm{i}a_{\mathrm{f}1}[2|a_{\mathrm{b}1}|^2 + |a_{\mathrm{f}1}|^2] + \mathrm{i}a_{\mathrm{f}1}^{*2}a_{\mathrm{f}3} \qquad [16a]$$

$$\frac{-\partial a_{\mathrm{b}1}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}1}}\frac{\partial a_{\mathrm{b}1}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{b}1} + \mathrm{i}\delta_1 a_{\mathrm{b}1} - \mathrm{i}\kappa_1 a_{\mathrm{f}1} - \frac{\alpha_1}{2}a_{\mathrm{b}1} + \mathrm{i}a_{\mathrm{b}1}[2|a_{\mathrm{f}1}|^2 + |a_{\mathrm{b}1}|^2] + \mathrm{i}a_{\mathrm{b}1}^{*2}a_{\mathrm{b}3} \qquad [16b]$$

where, at the fundamental harmonic, $v_{\mathrm{g}1}$ is the group velocity of the homogeneous medium, δ_1 is the laser detuning, κ_1 is the coupling coefficient, α_1 is the absorption coefficient, $\nabla_\perp^2$ is the transverse Laplacian and the whole term ($\nabla_\perp^2 a_{\mathrm{f}1}$ and $\nabla_\perp^2 a_{\mathrm{b}1}$) accounts for diffraction effects, and ' $*$ ' denote the complex conjugate of the corresponding term. Similarly, the coupled mode equations for the third-harmonic field are given as

$$\frac{\partial a_{\mathrm{f}3}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}3}}\frac{\partial a_{\mathrm{f}3}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{f}3} + \mathrm{i}a_{\mathrm{f}3}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}(|a_{\mathrm{f}1}|^2 + |a_{\mathrm{b}1}|^2)\right] - \mathrm{i}\kappa_3 a_{\mathrm{b}3} - \frac{\alpha_3}{2}a_{\mathrm{f}3} + \mathrm{i}a_{\mathrm{f}1}^3 \qquad [17a]$$

$$-\frac{\partial a_{\mathrm{b}3}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g}3}}\frac{\partial a_{\mathrm{b}3}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{b}3} + \mathrm{i}a_{\mathrm{b}3}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}(|a_{\mathrm{b}1}|^2 + |a_{\mathrm{f}1}|^2)\right] - \mathrm{i}\kappa_3 a_{\mathrm{f}3} - \frac{\alpha_3}{2}a_{\mathrm{b}3} + \mathrm{i}a_{\mathrm{b}1}^3 \qquad [17b]$$

where at the third harmonic, $v_{\mathrm{g}3}$ is the group velocity of the homogeneous medium, δ_3 is the laser detuning, κ_3 is the coupling coefficient, and α_3 is the absorption coefficient. Note that in the limit when $k_3 \approx 3k_1$, the coupling coefficient between the fundamental and third harmonics are related by

$$\kappa_3 = 3\left(\frac{d}{\pi}\right)\frac{\Delta\varepsilon\omega_0^2}{2k_3c^2} + 3\delta_1 = \kappa_1 \qquad [18]$$

Also, note that the laser detuning between the fundamental and third harmonics are expressed as

$$\delta_3 = \left(\frac{d}{\pi}\right)\Delta k + 3\delta_1 \qquad [19]$$

where $\Delta k = k_3 - 3k_1 \times$ Equation [19] described the phase matching condition, which determines the efficiency of the third-harmonic generation. When the phase is completely matched ($\Delta k=0$), then we can achieved the most efficienct in third-harmonic generation when $\delta_3 = 3\delta_1$. Equations [16] and [17] are useful equations which are commonly used for numerically simulation of beam propagation.

Defect Mode Using Photonic Crystal

A defect mode can be generated in the photonic crystal if we introduce a defect in the geometry of the photonic crystal. The schematic of the photonic crystal with a defect of length d_{d} located at the center of the photonic crystal, where the subscript *'d'* denotes defect, is depicted in Figure 3. For the convenience of the analysis, we divide the entire photonic crystal into three regions: (i) Region I; (ii) Region II; and (iii) Region III. Since we arbitrarily

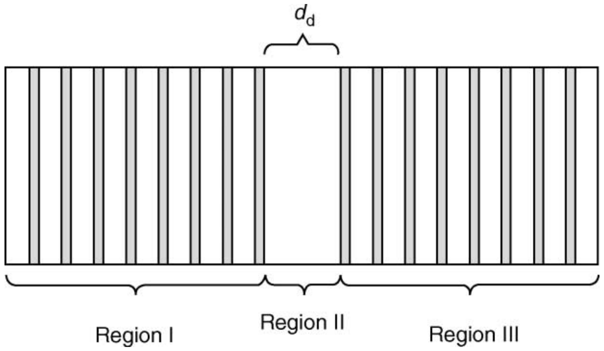

Figure 3 The schematic of the photonic crystal with a defect of length d_d.

put the defect in the center of the photonic crystal, Region I and II will be identical, thus it will be easier for us in our analysis. The strategy of our analysis is to solve the wave equation for Region I of the photonic crystal, and apply the result to the Region II analysis. Similarly, we apply the corresponding result of Region II to Region III. Thus, by breaking up the entire photonic crystal into three regions, we can then easily analyze the propagation of a field through the entire photonic crystal.

Assume that Region I and II are each of length $L/2$. Using the coupled mode equations without considering the effect of absorption, the fundamental fields in the forward and backward modes in Region I is given as

$$a_{\mathrm{f1(REG)}}(z_1) = \left(\cos(\Delta_1 z_1) - \frac{\mathrm{i}\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{\mathrm{f1(REG)}}(0) + \frac{\mathrm{i}\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{\mathrm{b1(REG)}}(0) \qquad [20a]$$

$$a_{\mathrm{b1}(REG)}(z_1) = \left(\cos(\Delta_1 z_1) + \frac{\mathrm{i}\delta_1}{\Delta_1}\sin(\Delta_1 z_1)\right)a_{\mathrm{b1(REG)}}(0) - \frac{\mathrm{i}\kappa}{\Delta_1}\sin(\Delta_1 z_1)a_{\mathrm{f1(REG)}}(0) \qquad [20b]$$

where

$$\Delta_1 = \sqrt{\delta_1^2 - \kappa^2}, \qquad \kappa = \frac{d\omega_0^2\Delta\varepsilon(\omega)}{2\pi k_1 c^2}$$

and the subscript '(REG)' denotes solution in either Region I, II, or III. The boundary conditions for the fundamental harmonic field, at the interface between Region I and II, are given as

$$a_{\mathrm{f1(II)}}(d_\mathrm{d}) = \mathrm{e}^{ikd_\mathrm{d}}a_{\mathrm{f1}(I)}(L/2) \qquad [21a]$$

$$a_{\mathrm{b1(II)}}(d_\mathrm{d}) = \mathrm{e}^{-ikd_\mathrm{d}}a_{\mathrm{b1}(I)}(L/2) \qquad [21b]$$

where $k = n\omega/c$. The field that propagates through Region II can be determined using eqns [20] and [21]. Also, the boundary conditions at the front and end faces of the Region III of the photonic crystal are given as

$$a_{\mathrm{f1(III)}}(0) = a_{\mathrm{f1(II)}}(d_\mathrm{d}) \qquad [22a]$$

$$a_{\mathrm{b1(III)}}(L/2) = 0 \qquad [22b]$$

Using eqns [20] and [22], we can determine the field that propagates through Region III. Similarly, we can use the same approach to analyze the third-harmonic field.

Homogeneous Nonlinear Crystal

If we examine Figure 2, we see that when $\Delta\varepsilon(\omega) = 0$, the photonic crystal becomes a homogeneous nonlinear crystal. In another word, $\kappa_3 = 0$ when we set the laser detuning to zero, thus there is no more coupling between the layers, and we are left with only the forward mode. Without going through the mathematics of four-wave mixing to describe the THG in a homogeneous nonlinear crystal, we can obtain the fundamental and third-harmonic coupled mode equations from eqns [16] and [17]. This will make it easier to make comparisons between the photonic crystal and the homogeneous nonlinear crystal, rather than comparing the coupled mode equation with the four-wave mixing equations among them. Since no backward wave is generated in the homogeneous nonlinear crystal when there is no coupling between the forward and backward modes, eqn [16] can remove the coupling term and the backward mode equation, and the modification is given as

$$\frac{\partial a_{\mathrm{f1}}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g1}}}\frac{\partial a_{\mathrm{f1}}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_1}\nabla_\perp^2 a_{\mathrm{f1}} + \mathrm{i}\delta_1 a_{\mathrm{f1}} - \frac{\alpha_1}{2}a_{\mathrm{f1}} + \mathrm{i}a_{\mathrm{f1}}|a_{\mathrm{f1}}|^2 + \mathrm{i}a_{\mathrm{f1}}^{*2}a_{\mathrm{f3}} \qquad [23]$$

$$\frac{\partial a_{\mathrm{f3}}}{\partial z_1} + \frac{d}{\pi v_{\mathrm{g3}}}\frac{\partial a_{\mathrm{f3}}}{\partial t_1} = \frac{\mathrm{i}d}{2\pi k_3}\nabla_\perp^2 a_{\mathrm{f3}} + \mathrm{i}a_{\mathrm{f3}}\left[\delta_3 + 6\sqrt{\frac{3k_1}{k_3}}|a_{\mathrm{f1}}|^2\right] - \frac{\alpha_3}{2}a_{\mathrm{f3}} + \mathrm{i}a_{\mathrm{f1}}^3 \qquad [24]$$

for the fundamental and third-harmonic fields, respectively.

Applications

Generation of short wavelength radiation is possible using THG. Therefore, it becomes less expensive to generate green, blue, or ultraviolet light with the infrared and the red light. In order to improve THG conversion efficiency, a beam with a tight focus is used, which results in an electric field of highest intensity, and has the highest possible χ^3 polarization effect. Normally, due to the phenomenon of the Gouy phase shift at a tight focal point of the beam, third-harmonic generation is suppressed. However, in the presence of an interface within the focal volume, the phase is disturbed and a third-harmonic signal is generated. This phenomenon has been used to image cells in *in vivo* microscopy on a femtosecond time-scale. The ultrashort pulses have little energy and therefore little heat is deposited in the material. The cells can be continuously imaged without damage.

See also

Materials Characterization Techniques: $\chi^{(3)}$. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Barad Y, Eisenberg H, Horovitz M and Silberberg Y (1997) Nonlinear scanning laser microscopy by third harmonic generation. *Applied Physics Letters* **70**: 922–924.

Boyd RW (1992) *Nonlinear Optics*. Boston, MA: Academic Press.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge: Cambridge University Press.

Franken PA, Hill AE, Peters CW and Weinreich G (1961) Generation of optical harmonics. *Physics Review Letters* 7: 118.

Markowicz PP, Tiryaki H, Pudavar H, Prasad PN, Lepeshkin NN and Boyd RW (2004) Dramatic enhancement of third-harmonic generation in three-dimensional photonic crystals. *Physics Review Letters* 92: 083903.

Petrov GI, Shcheslavskiy V, Yakovlev VV, Ozerov I, Chelnokov E and Marine W (2003) Efficient third-harmonic generation in a thin nanocrystalline film of ZnO. *Applied Physics Letters* **83**: 3993–3995.

Schins J, Schrama T, Brakenhoff G, Muller M and Squier J (2002) Determination of material properties using third harmonic generation microscopy. *Journal of the Optical Society of America B* 19: 1627.

Squier J, Mueller M, Brakenhoff GJ and Wilson KR (1998) Third harmonic generation microscopy. *Optics Express* 3: 315.

Terhune RW, Maker TD and Savage CM (1962) Optical harmonic generation in calcite. *Physics Review Letters* 8: 404.

Viswanathan R (1999). Enhanced Frequency Conversion in Periodic Structures. Ph.D. dissertation, Rensselaer Polytechnic Institute.

Four-Wave Mixing

L Canioni and L Sarger, University of Bordeaux, Talence, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Light waves propagate along straight lines in a vacuum and in materials. However, their amplitude and phase can only be affected by the intrinsic properties of the sample, such as absorption and refraction index. Experimental work from the early 1960s, has deviated from this simple approach, mainly through propagating intense light beams in crystals and glasses. Among many puzzling new effects, frequency conversion and beam coupling have triggered intense studies and opened a whole new field, theoretically and experimentally. These various phenomena have been classified according to the order of their optical nonlinearity. The second order mainly occurs in frequency generation and has already been discussed in previous articles of this encyclopedia. Therefore, this article will be devoted to the higher order, in which self focusing, coupling, and mixing of light beams will be explained. We will illustrate some simple ideas with a convenient model before describing a few applications.

The framework of electromagnetic wave interaction with matter is embedded in the Lorentz force. The oscillating electric field drives the components, ions, and electrons, but due to the large mass differences, only the electron movement is strongly perturbed. In dielectric media, these carriers are connected with static bonds such that the applied electric field modulates the spatial position of equilibrium of the charges. As a result, a dipolar moment is created and oscillates quasi-adiabatically at very high frequencies (between 10^{14} and 10^{17} Hz) for an applied optical wave. These collective oscillations of the induced dipoles compose the so-called matter polarization, proportional to the driving electric field amplitude. Magnetic field interaction is usually much weaker and so can be neglected in nonmagnetic media.

As an electromagnetic wave, lasers are a coherent source of very high intensity, i.e., large electric fields. For example, in standard pulsed lasers focused on a target, the light intensity can easily reach 10 Tw/cm^2, corresponding to an electric field of the order of magnitude of the internal bonding field (10^{10} V/m). At this level of interaction, dipolar moments are no longer strictly proportional to the applied electric field, as depicted in Figure 1. In the weak field regime, the simple perturbation approach displaces the carriers in a power series of the applied electric field. The polarization is thus developed in series, in respect to this electric field.

Formalism

Relation Between Field and Polarization: The Response Function

Although numerous works have been published with a frequency description of the material polarization, we will discuss here the study of nonlinearity in the time domain in order to follow the physical reality. In most processes, both spaces are equal and can be selected at will, as long as the entire requirements are fully understood; the electromagnetic field is classically described.

In the time domain, the polarization can be developed as

$$P(r,t) = P_{\text{Linear}}(r,t) + P_{\text{Nonlinear}}(r,t) \qquad [1]$$

where the linear part of the polarization can be written as

$$P_{\text{Linear}}(r,t) = \varepsilon_0 \int_{R^3} \int_{-\infty}^{\infty} R^{(1)}(r_1,t_1) \times E(r-r_1, t-t_1)\mathrm{d}t_1\,\mathrm{d}r_1 \qquad [2]$$

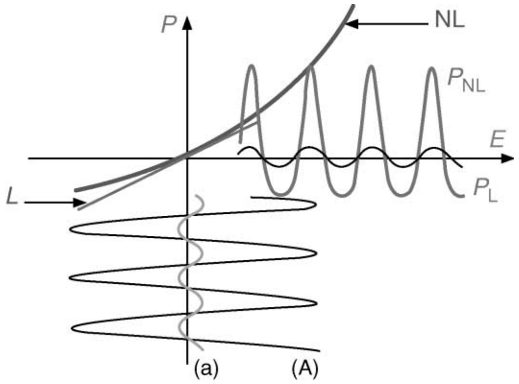

Figure 1 Polarization for various field amplitude: (a) Weak field case, linear response. (A) Strong field, distortion and nonlinearity.

This general expression is rather complex but shows that the polarization, due to collective movement of carriers in a macroscopic volume (thousands of dipoles) centered at position r, always depends on the applied electric field. This expression is also nonlocal as dipoles–dipoles interactions play an important role.

In general, for the nonlinear (NL) part of the polarization, an approximation of the local interaction is described for simplicity. For a second-order interaction, one can write:

$$P_{\mathrm{NL}}^{(2)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(2)}(t_1,t_2) \times E(r,t-t_2)E(r,t-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2 \qquad [3]$$

For the other series expansion – at third order, one obtains

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t_1,t_2,t_3)E(r,t-t_3) \times E(r,t-t_3-t_2)E(r,t-t_3-t_2-t_1)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3 \qquad [4]$$

This third-order polarization already combines three electric fields. This nonlinear polarization also radiates a fourth wave and is therefore the adopted framework for describing four-wave mixing experiments (Figure 2).

As the physical properties of materials depend only on time intervals between the different pulse interactions, eqn (4) can be rearranged as

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t-\tau_1,t-\tau_2,t-\tau_3) \times E(r,\tau_1)E(r,\tau_2)E(r,\tau_3)\mathrm{d}\tau_1\,\mathrm{d}\tau_2\,\mathrm{d}\tau_3 \qquad [5]$$

If the sample response time is much shorter than pulse duration, the so-called adiabatic limit (as for electronic cloud deformation excited distant from

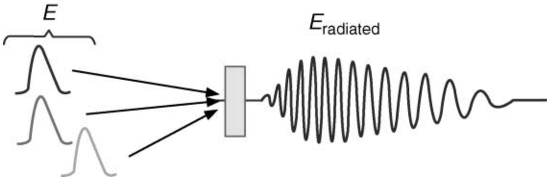

Figure 2 Three incoming beams E_1, E_2, E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4.

resonance) the polarization is written as

$$P_{\mathrm{NL}}^{(3)}(r,t)=\varepsilon_0\sigma^{(3)}E(r,t)E(r,t)E(r,t) \qquad [6]$$

We will see later how this expression is useful for third-harmonic generation and optical Kerr effects, etc. This expression is accurate for materials excited at a distance from resonances and care must be taken to reserve such a model for corresponding processes. In this case, the frequency approach is quite simple too and both spaces, time or Fourier, can be chosen at will.

For very long response time, such as orientational or structural molecular reorganization, one can use the Born–Oppenheimer approximation and present the nonlinear polarization as

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 E(r,t)R'(t) \qquad [7]$$

Similar is linear polarization, where the response function takes into account the changes of the molecular structure under the light intensity stress:

$$R'(t) = \int_{-\infty}^{+\infty} d^{(3)}(t-t_1)E(r,t_1)E(r,t_1)\mathrm{d}t_1 \qquad [8]$$

The material response will then be mostly driven by the field amplitude rather than the frequency.

Properties of Nonlinear Susceptibilities

The third-order nonlinear susceptibility $\chi^{(3)}$ is the Fourier transform of the response function R and has proved to be a most practical tool in optical nonlinear theory. The symmetry properties of the material lead to simplification on the tensor of rank 4 (fourth rank tensor), involved in four-wave mixing:

$$P_i^{(3)}(\omega_m)=\varepsilon_0\sum_{n_1,\ldots,n_p}\sum_{i_1,\ldots,i_p}\chi^{(3)}_{i,i_1,\ldots,i_p}(\omega_m;\omega_{n_1},\ldots,\omega_{n_p}) \times E_{i_1}(\omega_{n_1})\cdots E_{i_p}(\omega_{n_p}) \qquad [9]$$

The nonlinear susceptibility tensor must account for all the spatial and symmetrical order of the sample. This will eventually reduce the number of independent elements from a possible 81.

Wave Equation in Nonlinear Regime

From Maxwell's equations, with a sample without free carriers and current, one can deduce the wave equation in the time domain, as:

$$\Delta E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} - \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{L}}}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \qquad [10]$$

In the frequency domain, this equation reduces to

$$\Delta\tilde{E}(r,\omega) + k(\omega)^2\tilde{E}(r,\omega) = -\mu_0\omega^2\tilde{P}_{\mathrm{NL}}(r,\omega) \qquad [11]$$

In order to solve these (coupled) wave equations in relation to four-wave mixing, we must develop the electric field and the polarization in a slowly varying envelope approximation, using the following conventions:

$$E_i(r,t)=\frac{1}{2}\left(A_i(r,t)\mathrm{e}^{\mathrm{i}(\omega_i t-k(\omega_i)z)}+\bar{A}_i(r,t)\mathrm{e}^{-\mathrm{i}(\omega_i t-k(\omega_i)z)}\right) \qquad [12]$$

$$P_{\mathrm{NL}}(r,t)=\frac{1}{2}\left(p_{\mathrm{nl}}(r,t)\mathrm{e}^{\mathrm{i}(\omega t-k(\omega)z)}+\bar{p}_{\mathrm{nl}}(r,t)\mathrm{e}^{-\mathrm{i}(\omega t-k(\omega)z)}\right) \qquad [13]$$

Within the framework of small perturbations, diffraction can be neglected and the wave equation can be easily reduced to one dimension:

$$\frac{\partial A}{\partial z}=\frac{\mathrm{i}\omega}{2n(\omega)\varepsilon o c}p^{\mathrm{NL}}(z,\omega)\exp(-\mathrm{i}k(\omega)z) \qquad [14]$$

where ω represents the sum or difference permutations over all the frequencies ω_i, thus reflecting the law of energy conservation.

In the case of four-wave mixing, one has to solve four coupled equations for the four interacting waves. Using the polarization eqn [13], one gets:

$$\frac{\partial A_j}{\partial z}=\frac{\mathrm{i}\omega_j}{2n(\omega_j)c}\chi^3 A_1^{(*)}A_2^{(*)}A_3^{(*)} \times\exp\left(\sum_{i=1,2,3}\exp(\pm\mathrm{i}k_i(\omega_i)z)\right)\exp(-\mathrm{i}k(\omega_j)z) \qquad [15]$$

This set of generic equations respectively represent indifferently, either the three incoming beams ($j=1,2,3$), or the radiated field E_4 at ω. The algebra used here associates at each complex conjugate field (*) with the counter propagating wave ($-k$). The amplitude variation of the j field at frequency ω_j is due to the coupling of the three other fields within the sample. The accumulated field is nevertheless always strongly dependent of the phase term and will vanish except when this phase term is zero. This is known as the phase matching condition which has to be fulfilled to ensure an efficient energy conversion. A more general approach of this condition can be found in the corresponding article of this encyclopedia.

Example and Selected Applications

When inserting the full material polarization (linear and nonlinear) into the propagation equation, even restricted at third order, a whole set of processes must be considered. All of them are described by the generic equation already presented, but only a few very important processes – with increasing complexity – will be described here.

Optical Kerr Effect

We consider one single beam incident at frequency ω on a sample. The non-linear polarization is given by

$$P_{\mathrm{NL}}^{(3)}(r,t)=\frac{\varepsilon_0}{8}\chi^{(3)}\left(E(r,t)^3+E^*(r,t)^3+3E(r,t)^2E^*(r,t)+3E(r,t)E^*(r,t)^2\right) \quad [16]$$

The polarization at frequency ω reduces to

$$P_{\mathrm{NL}}^{(3)}(r,t)=\frac{3\varepsilon_0}{4}\chi^{(3)}|E(r,t)|^2E(r,t) \quad [17]$$

This term has the same phase as the incident beam and, therefore, the phase matching is automatic. Using eqns [12] and [13], one can rewrite the wave equation for the optical Kerr effect in the thin sample approximation:

$$\frac{\partial A}{\partial z}=\frac{3\mathrm{i}\omega}{8nc}\chi^{(3)}|A|^2A \quad [18]$$

The solution of this equation is

$$A(z)=A(0)\exp\left(\frac{3\mathrm{i}}{8nc}\chi^{(3)}|A(0)|^2z\right) \quad [19]$$

Along the beam, the amplitude of the input pulse stays constant for this process but its phase is affected by a nonlinear shift proportional to power density I.

In literature, one can find the concept of n_2, the sample nonlinear index, in the following formula:

$$\psi_{\mathrm{NL}}(z) = \frac{\omega}{c}n_2Iz \quad [20]$$

In this way of describing the total phase along the beam, the index of refraction in the nonlinear regime can be expressed more conveniently as

$$n = n_0 + n_2I \quad [21]$$

One of the most significant demonstrations of this effect is the self-focusing phenomenon. The usual Gaussian transverse structure of a laser beam presents high intensity at the center of the beam, therefore creating an index variation in the isotropic sample due to this optical Kerr effect. An induced lens can then be applied to this refractive index gradient, following the field distribution. As a result, the beam focuses itself along the propagation and can result in a complete breakdown.

Third Harmonic Generation

In the third harmonic generation, i.e., the third-order nonlinear process where the frequency of the generated wave is three times the frequency of the input wave at ω, only two waves are in the sample (at ω and 3ω). For the nonlinear wave, the master equation is then:

$$\frac{\partial A_{3\omega}}{\partial z}=\frac{3\mathrm{i}\omega}{8n(3\omega)c}\chi^{(3)}A_{\omega}^3\exp((3\mathrm{i}k(\omega)-\mathrm{i}k(3\omega))z) \quad [22]$$

Here, propagation effects have to be carefully considered as theses waves do travel at different velocities and will eventually be out of phase. Obviously this discrepancy of phase velocities precludes the co-propagation and strongly affects the global efficiency. The only way to ensure a coherent generation of the harmonic wave throughout the material is to fulfill the relation:

$$k(3\omega)=3k(\omega) \quad [23]$$

This is the so-called phase matching condition. Although difficult because the usual dispersion of materials is very large for such a frequency difference, this can be satisfied using either geometrical arrangement or, if possible, using the birefringent properties of the material. The efficiency of this method is poor and a better approach is implemented when two $\chi^{(2)}$ processes are cascaded (doubling and sum frequency mixing).

Degenerate Four-Wave Mixing and Phase Conjugation

One of the most popular and intriguing interactions is depicted in Figure 3. Where the sample is excited by three fields at the same frequency ω in this particular geometry: two-waves, named pump beams E_1 and E_2, are exactly counter-propagating with opposite wave vectors of $+k_1$ and $k_2 = -k_1$, while the third wave (E_3) arrives at the sample with a given wave vector of k_3.

Of possible couplings, one contribution E_4 corresponds to a re-emitted beam along $k_4 = -k_3$, i.e., opposite to the E_3 wave.

The third-order nonlinear polarization is given here by

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{\varepsilon_0}{8}\chi^{(3)}(E_1(r,t) + E_1^*(r,t) + E_2(r,t) + E_2^*(r,t) + E_3(r,t) + E_3^*(r,t))^3 \quad [24]$$

This E_4 wave (along $-k_3$) originates in this product from source terms including:

$$E_1(r,t)E_2(r,t)E_3^*(r,t) \quad [25]$$

and thus present an evolution driven by:

$$\frac{\partial A_4}{\partial z} = \frac{3\mathrm{i}\omega}{4n(\omega)c}\chi^{(3)}A_1A_2A_3^* \quad [26]$$

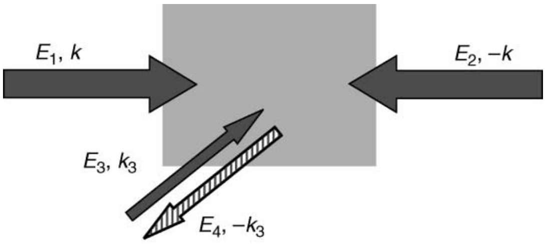

Figure 3 Two counter-propagating fields (E_1, E_2) and the third field E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 opposite to the E_3 wave.

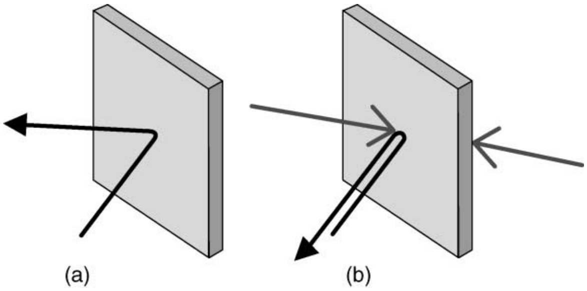

Figure 4 (a) Snell refraction in a conventional mirror; (b) Phase conjugate mirror.

where the phase matching condition is automatically fulfilled. Interesting facts can be described using this expression. First, the radiated amplitude is proportional to the complex conjugate of the E_3 field. This nonlinear process is not only able to reverse the direction of propagation but also the whole phase of an arbitrary incoming beam of light. This 'phase conjugator' can be considered as a kind of mirror with very unusual reflection properties. Unlike a conventional mirror, where a ray is redirected according to the ordinary law of reflection, a phase-conjugate mirror (PCM) retro-reflects all incoming rays back to their origin (Figure 4).

Second, the reflectivity of this mirror depends on the susceptibility and the amplitude of the first two fields, usually called pump beams. This will allow reflectivity that is much higher than the unity, unlike with conventional mirrors.

The remarkable reflection properties of the PCM have found many important applications. The most useful undoubtedly is related to distortion correction. If the image information has been distorted by the transmitting medium on its way to the PCM, then these aberrations will be corrected when the reflected signal retraces its original path through the medium. Through this amazing 'healing' property of optical phase conjugation, a high-quality optical beam can be double passed through a distorting medium, such as an imperfect imaging device or a turbulent atmosphere, without any loss of beam quality. Figure 5 shows the basic two-pass geometry for imaging and aberration correction using a PCM.

The spherical wave from the object point P is aberrated by the distorting medium. The wave-front is reversed by the PCM and a second passage through the same distorting path restores the initial field. The resulting spherical wave converges to the image point P', separated from P by the beamsplitter.

Numerous other applications for PCM and the various underlying nonlinear optical interactions have been proposed. These include effects as diverse

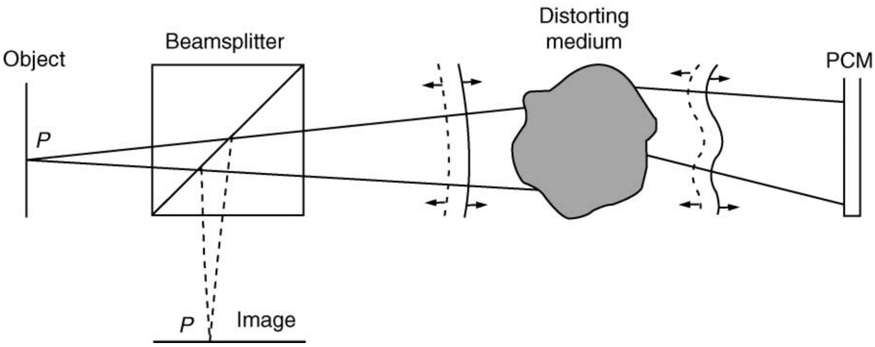

Figure 5 A basic two-pass system for imaging and distortion compensation using optical phase conjugation.

as parametric oscillation, optical tracking and pointing, spatial and temporal image processing, optical computing, optical filtering, etc. The PCM is also successfully implemented for aberration correction in high-power laser resonators.

Nondegenerate Four-Wave Mixing

If the frequencies of the impinging fields build a nonlinear polarization with an oscillating term close to absorption (or emission) frequencies of the sample, the nonlinear susceptibility increases tremendously. The radiated field amplitude changes by orders of magnitude and allows spectroscopic studies of the sample. The most popular process used to perform gas spectroscopy is CARS, an acronym for coherent anti-Stokes Raman scattering (or spectroscopy). In this case, the three incoming beams have different frequencies and the phase matching condition is eventually fulfilled in the 3D space.

Usually, two beams at the same frequency ω_P (pump beams), are mixed with a third beam at a tunable frequency ω_S, according to the energy scheme in Figure 6. As the radiated frequency ω_{CARS} is higher, the Raman spectroscopic notation is used and the subscripts 'S' hold for 'Stokes' and 'P' for 'anti-Stokes'.

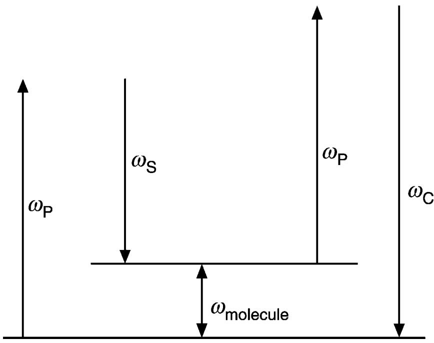

Figure 6 Energy diagram for a CARS experiment. Four-wave mixing efficiency increases when the frequency difference $\omega_P - \omega_S$ is close to a resonant frequency of the sample, $\omega_{molecule}$.

Whenever $\omega_P - \omega_S = \omega_{molecule}$, then an increase of the CARS intensity by many orders of magnitude is observed. Therefore, a Raman spectrum can be obtained by continuously changing, the Stokes laser and simultaneously recording the intensity of the CARS beam. This procedure is called scanning CARS, an interesting technique insofar as it allows taking high-resolution spectra, as well as measuring temperature of the sample while scanning over the Boltzmann distribution of the electronic ground state.

Figure 7 displays a classical geometrical arrangement where the phase matching condition fully determines the directionality of the radiated field at ω_{CARS}. Although the alignment of the setup can become tedious, this technique has been widely used, even in hostile industrial environments. The coherent aspect of this four-wave mixing process, imbedded in the phase matching condition, has a huge benefit over classical spectroscopic techniques and thus isotropic techniques, such as laser-induced fluorescence or Raman spectroscopy.

The CARS process is an interference process comparable to diffraction of a grating. The two fields at ω_P and ω_S form a laser-induced moving grating and the third field, at ω_P, undergoes a Doppler shift at ω_{CARS}. This scattered field undergoes a coherent amplitude addition in the Bragg direction.

Conclusion

In this article, we have discussed a set of optical nonlinear processes occurring at the third order of the development in series of the material polarization. At this order, wave mixing obviously shows most

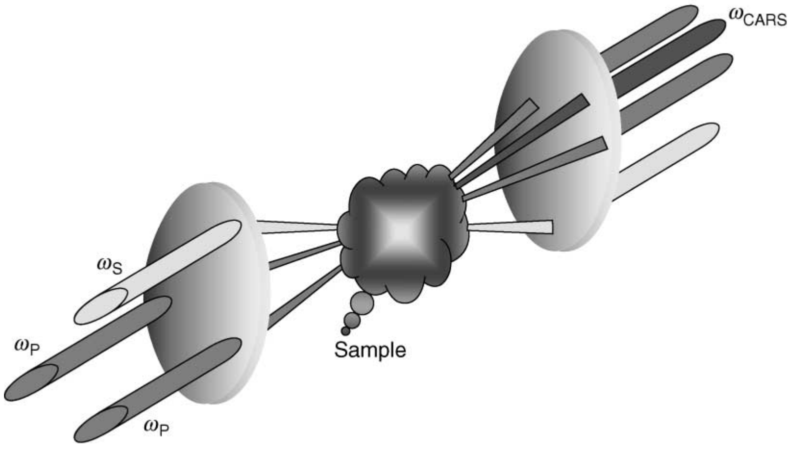

Figure 7 Layout of the CARS beams near a sample (a flame here). This particular arrangement is called folded (or 3D) Boxcars.

significant behavior and has been extensively studied and applied in many scientific areas. The general framework presented here can be further investigated using some books listed in Further Reading below.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(3)}$–Third-Harmonic Generation. **Nonlinear Sources:** Harmonic Generation in Gases. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed).

Further Reading

Bloembergen N (1996) *Nonlinear Optics*. Singapore: World Scientific Pub Co.

Bloom DM and Bjorklund GC (1977) Conjugate wavefront generation and image reconstruction by four-wave mixing. *Applied Physics Letters* 31: 592–594.

Born M and Wolf E (1999) *Principles of Optics*. Cambridge, UK: Cambridge University Press.

Butcher PN and Cotter D (1991) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Eaton DF (1991) Nonlinear optical materials. *Science* 253: 281–287.

Fisher RA (1984) *Optical Phase Conjugation*. San Diego, CA: Academic.

Flytzanis C (1975) Theory of nonlinear optical susceptibilities. In: Rabin H and Tang CL (eds) *Quantum Electronics*, vol. 1, pp. 9–207. New York: Academic Press.

Hellwarth RW (1977) Generation of time-reversed wave fronts by nonlinear refraction. *Journal of the Optical Society of America* 67: 1–3.

Jackson JD (1998) *Classical Electrodynamics*. New York: Wiley.

Marburger JH (1975) Self-focusing: Theory. *Progress in Quantum Electronics* 4: 35–110.

Mittra R and Habashy TM (1984) Theory of wave-front-distortion correction by phase conjugation. *Journal of the Optical Society of America A* 1: 1103–1109.

Rockwell DA (1988) A review of phase-conjugate solid-state lasers. *IEEE Journal of Quantum Electronics* 24: 1124–1140.

Shen YR (2002) *The Principles of Nonlinear Optics*. New York: Wiley.

Yariv A (1977) Compensation for atmospheric degradation of optical beam transmission. *Optical Communication* 21: 49–50.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.

Yariv A and Pepper DM (1977) Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing. *Optical Letters* 1: 16–18.

Zel'dovich BYa, Pilipetsky NF and Shkunov VV (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Kramers–Krönig Relations in Nonlinear Optics

M Sheik-Bahae, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Since their introduction nearly 75 years ago, the Kramers–Krönig (KK) dispersion relations have been widely appreciated and applied in the analysis of linear optical systems. Because they are a consequence of strict causality, the KK relations apply not only to optical systems, but also to any linear, causal system such as electrical networks and particle scattering. In this article, we review the formulation and application of these relations in nonlinear optical systems. Simple logical arguments are used to derive dispersion relations that relate the nonlinear absorption coefficient to the nonlinear refraction coefficient. More general formalisms are then derived that apply to all nonlinear susceptibilities including the harmonic generating cases. Examples of recent successful application of these dispersion relations in analyzing various nonlinear materials will be presented.

The mathematical formalism of the KK dispersion relations in nonlinear optics was studied in the formative days of the field. The great usefulness of these relations was appreciated only recently, however, when they were used to derive the dispersion of the optical Kerr effect in solids from the corresponding nonlinear absorption coefficients, including two-photon absorption.

Before examining the details of KK relations in nonlinear optical systems, it is instructive to revisit the linear dispersion relations and their derivation based on the logic of causality. We will begin this task by introducing the definition of the linear as well as nonlinear susceptibilities $\chi^{(n)}$. In most nonlinear optics texts, the total material polarization (P) that drives the wave equation for the electric field (E) is expressed as

$$P_i(t)=\varepsilon_0\int_{-\infty}^{\infty}R_{ij}^{(1)}(t-t_1)E_j(t_1)\mathrm{d}t_1+\varepsilon_0\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\times R_{ijk}^{(2)}(t-t_1,t-t_2)E_j(t_1)E_k(t_2)\mathrm{d}t_1\,\mathrm{d}t_2+\varepsilon_0\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}R_{ijkl}^{(3)}(t-t_1,t-t_2,t-t_3)\times E_j(t_1)E_k(t_2)E_l(t_3)\mathrm{d}t_1\,\mathrm{d}t_2\,\mathrm{d}t_3+\cdots \quad [1]$$

where $R^{(n)}$ is defined as the nth-order, time-dependent response function or time-dependent susceptibility. The subscripts are polarization indices indicating, in general, the tensor nature of the interactions. The summation over the various indices $j, k, l, \ldots$ is implied for the various tensor elements of $R^{(n)}$. Upon Fourier transformation, we obtain:

$$\vec{P}_i(\omega)=\varepsilon_0\int_{\infty}^{-\infty}\mathrm{d}\omega_1\,\chi_{ij}^{(1)}(\omega_1)\vec{E}_j(\omega_1)\delta(\omega-\omega_1)+\varepsilon_0\int_{\infty}^{-\infty}\mathrm{d}\omega_1\int_{\infty}^{-\infty}\mathrm{d}\omega_2\times\chi_{ijk}^{(2)}(\omega_1,\omega_2)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\delta(\omega-\omega_1-\omega_2)+\varepsilon_0\int_{\infty}^{-\infty}\mathrm{d}\omega_1\int_{\infty}^{-\infty}\mathrm{d}\omega_2\int_{\infty}^{-\infty}\mathrm{d}\omega_3\times\chi_{ijkl}^{(3)}(\omega_1,\omega_2,\omega_3)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\vec{E}_l(\omega_3)\times\delta(\omega-\omega_1-\omega_2-\omega_3),\cdots \quad [2]$$

where δ is the Dirac delta-function. Here the $E(\omega)$ are Fourier transforms of the corresponding electric field. The nth-order susceptibility is defined as the Fourier transform of the nth-order response function:

$$\chi_{ijk\ldots n}^{(n)}(\omega_1,\omega_2,\ldots,\omega_m)=\int_{-\infty}^{+\infty}\mathrm{d}\tau_1\int_{-\infty}^{+\infty}\mathrm{d}\tau_2\cdots\int_{-\infty}^{+\infty}\times\mathrm{d}\tau_n R_{ijk\ldots m}^{(n)}(\tau_1,\tau_2,\ldots,\tau_m)\mathrm{e}^{\mathrm{i}(\omega_1\tau_1+\omega_2\tau_2+\cdots+\omega_m\tau_m)} \quad [3]$$

For simplicity, we drop the polarization indices $i, j, \ldots$, and thus ignore the tensor properties of $\chi^{(n)}$ as well as the vector nature of the electric fields.

Let us for the moment concentrate on the linear polarization alone and derive the linear KK relations for the first-order susceptibility $\chi^{(1)}(\omega)$. For this, we rewrite eqn [3] for $n = 1$:

$$\chi^{(1)}(\omega)=\int_{-\infty}^{\infty}R^{(1)}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \quad [4]$$

(As defined above, $\chi^{(1)}(\omega)$ and $R^{(1)}(\tau)$ are not a strict Fourier transform pair because of a missing factor of 2π.) Causality means that the effect cannot precede the cause. This can be restated mathematically as:

$$R^{(1)}(t)=R^{(1)}(t)\Theta(t) \quad [5]$$

i.e., the response to an impulse at $t = 0$ must be zero for $t < 0$. Here $\Theta(t)$ is the Heaviside step function defined as $\Theta(t) = 1$ for $t > 0$ and $\Theta(t) = 0$ for $t < 0$. Upon Fourier transforming this equation, the product in the time domain becomes a convolution integral in frequency space

$$\chi^{(1)}(\omega)=\chi^{(1)}(\omega)\left[\frac{\delta(\omega)}{2}+\frac{\mathrm{i}}{2\pi\omega}\right]=\frac{\chi^{(1)}(\omega)}{2}+\frac{\mathrm{i}}{2\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega-\omega'}\mathrm{d}\omega'=\frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega'-\omega}\mathrm{d}\omega' \quad [6]$$

which is the KK relation for the linear optical susceptibility. The symbol $\wp$ stands for the Cauchy principal value of the integral. The KK relation is thus a restatement of the causality condition [5] in the frequency domain. Taking the real part we have,

$$\Re e\{\chi^{(1)}(\omega)\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(1)}(\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \quad [7]$$

Taking the imaginary part of eqn [6] leads to a similar relation relating the imaginary part to an integral involving the real part. It is conventional to write the optical dispersion relations in terms of the more familiar quantities of refractive index, $n(\omega)$, and absorption coefficient, $\alpha(\omega)$. For $|\chi^{(1)}| \ll 1$ then $n - 1 = \Re e\{\chi^{(1)}\}/2$ and $\alpha = \omega\Im m\{\chi^{(1)}\}/c$, and eqn [7] is transformed into

$$n(\omega)-1=\frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\alpha(\omega')}{\omega'2-\omega^2}\mathrm{d}\omega' \quad [8]$$

where we additionally used the reality conditions of $n(\omega) = n(-\omega)$, and $\alpha(\omega) = \alpha(-\omega)$ to change the lower integral limit to 0. More rigorous analysis shows that eqn [8] is general and valid for any value of $|\chi^{(1)}|$. Although the KK dispersion relations and the extent of their applications in linear optics are well understood, some confusion sometimes exists about their applications to nonlinear optics. Causality clearly holds for both linear and nonlinear systems. The question is: what form do the resulting dispersion relations take in a nonlinear system? The linear Kramers–Krönig relations were derived from linear system theory, so it would appear to be impossible to apply the same logic to a nonlinear system. The key insight is that one can linearize the system. This is illustrated in Figure 1 where a linear

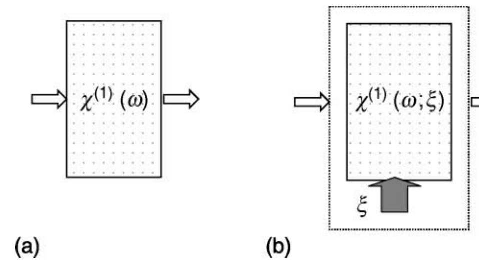

Figure 1 (a) A causal linear system obeying KK relations. (b) The system in (a) when externally perturbed by ξ. The dotted box now represents our new linear causal system whose altered $\chi^{(1)}$ obeys the KK relations.

(and of course, causal) optical material is transformed into a 'new' linear system that now contains the material and an external perturbation denoted by ξ. Although we are interested in perturbations of an optical nature, this formalism is general under any type of perturbation. It is important to appreciate the fact that our new system is causal even in the presence of the perturbation. This allows us to write down a modified form of the Kramers–Krönig relation linking the index of refraction to the absorption:

$$[n(\omega) + \Delta n(\omega;\zeta)] - 1 = \frac{c}{\pi}\wp \int_0^\infty \frac{\alpha(\omega') + \Delta\alpha(\omega';\zeta)}{\omega'2 - \omega^2} \mathrm{d}\omega' \qquad [9]$$

which, after subtracting the linear relation between n and α leaves a relation between the changes in index and absorption:

$$\Delta n(\omega;\zeta) = \frac{c}{\pi}\wp \int_0^\infty \frac{\Delta\alpha(\omega';\zeta)}{\omega'2 - \omega^2} \mathrm{d}\omega' \qquad [10]$$

where ζ denotes a general perturbation. An equivalent relation also exists whereby the change in absorption coefficient can be calculated from the change in the refractive index. It is essential that the perturbation be independent of frequency of observation, ω', in the integral (i.e., the excitation ζ must be held constant as ω' is varied).

Equation [10] has been used to determine refractive changes due to 'real' excitations such as thermal and free-carrier nonlinearities in semiconductors. In those cases, ζ denotes either ΔT (change of temperature) or ΔN (change of free-carrier density), respectively. In the former case, one calculates the refractive index change resulting from a thermally excited electron–hole plasma and the temperature shift of the band edge. For cases where an electron–hole plasma is injected (e.g., optically), the change of absorption gives the plasma contribution to the refractive index. In this case, the ζ parameter in eqn [10] is taken as the change in plasma density regardless of the mechanism of generation or the optical frequency.

Let us now extend this formalism to the case where the perturbation is virtual, occurring at an excitation frequency Ω that is below any material resonance. To the lowest order in the excitation irradiance I_Ω, we write

$$\Delta\alpha(\omega;\zeta) = \Delta\alpha(\omega;\Omega) = 2\alpha_2(\omega;\Omega)I_\Omega \qquad [11]$$

and

$$\Delta n(\omega;\zeta) = \Delta n(\omega;\Omega) = 2n_2(\omega;\Omega)I_\Omega \qquad [12]$$

where n_2 and α_2 are the nonlinear refractive index and absorption coefficients of the material, respectively. By definition, these coefficients are related to the third-order nonlinear susceptibility $\chi^{(3)}(\omega_1,\omega_2,\omega_3)$ via (*see* **Nonlinear Optics, Basics**: Nomenclature and Units)

$$n_2(\omega;\Omega) = \frac{3}{4\varepsilon_0 n_0(\omega) n_0(\Omega) c} \Re e\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [13]$$

and

$$\alpha_2(\omega;\Omega) = \frac{3\omega_\mathrm{a}}{2\varepsilon_0 n_0(\omega) n_0(\Omega) c^2} \Im m\{\chi^{(3)}(\omega,-\Omega,\Omega)\} \qquad [14]$$

We can therefore write the dispersion relations between α_2 and n_2:

$$n_2(\omega;\Omega) = \frac{c}{\pi}\wp \int_0^\infty \frac{\alpha_2(\omega';\Omega)}{\omega'2 - \Omega^2} \mathrm{d}\omega' \qquad [15]$$

Note that even when the degenerate $n_2(\omega) = n_2(\omega;\omega)$ is desired (at a given ω), the dispersion relation requires that we should know the nondegenerate absorption spectrum $\alpha_2(\omega';\omega)$ at all frequencies ω'.

Let us pause here and discuss some physical mechanisms that can be involved for a given system of interest. Consider a material characterized by an optical resonance occurring at, say ω_0 (i.e., a degenerate two-level system). For a solid, this resonance can be regarded as that of the fundamental energy gap; $\omega_0 = \omega_\mathrm{g} = E_\mathrm{g}/\hbar$ in a two-band system. Now, let us examine how the presence of an optical excitation at $\Omega < \omega_0$ can alter the absorption spectrum (at a variable probe ω'). In the quantum mechanical picture, this gives rise to a 'new' material whose perturbed wave functions are 'dressed' by the intensity and frequency of the applied optical field.

The lowest-order correction to the absorption is given by $\alpha_2(\omega';\Omega)$ which involves three major physical processes. Recalling that $\Omega < \omega_0$, these processes include (1) two-photon absorption (2PA) when $\omega' + \Omega \rightarrow \omega_0$ and (2) Raman-induced absorption when $\omega' - \Omega \rightarrow \omega_0$, both implying an absorption of a photon at the probe frequency ω' (i.e., $\alpha_2 > 0$). The third process can be identified as resulting from the blue-shift (for $\Omega < \omega_0$) of the resonance (known as the quadratic optical Stark effect) caused by the excitation field. For our two-level system, the latter results in a decrease followed by an increase in absorption in the vicinity of ω_0. An example of the overall absorption changes due to such processes is shown in Figure 2 where $\alpha_2(\omega';\Omega)$ is qualitatively plotted for a degenerate two-level system. We should note that the relative magnitude of each contribution as well as the width and shape of the resonances are chosen arbitrarily for the purpose of illustration. Using the KK relation in eqn [15], we can now arrive at the nonlinear index coefficient $n_2(\omega;\Omega)$. The result of this transformation is also given in Figure 2. The above simple example elucidates the key concepts involving the relationship between nonlinear absorption and refraction in materials for third-order processes. These concepts, when applied more rigorously to semiconductors, have been successful in predicting the sign, magnitude, and dispersion of n_2 due to the anharmonic motion of bound electrons. This will be briefly discussed later. Returning to the mathematical foundation of KK relations, we use eqns [13] and [14] to write eqn [15] in terms of the nonlinear susceptibility $\chi^{(3)}$:

$$\Re e\{\chi^{(3)}(\omega_1,\omega_2,-\omega_2)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(3)}(\omega',\omega_2,\omega_2)\}}{\omega'-\omega_1}\,d\omega' \qquad [16]$$

The above dispersion relation for $\chi^{(3)}$ was obtained using the physical and intuitive arguments that followed the linearization scheme depicted in Figure 1. General dispersion relations can be formulated following a mathematical procedure that is similar to the derivation of the linear KK relations. In this case we apply the causality condition directly to the nth-order nonlinear response $R^{(n)}$. For example, without loss of generality, we can write

$$R^{(n)}(\tau_1,\tau_2,\ldots,\tau_n) = R^{(n)}(\tau_1,\tau_2,\ldots,\tau_n)\Theta(\tau_j) \qquad [17]$$

and then calculate the Fourier transform of this equation. Here j can apply to any one of the indices $1,2,\ldots,n$. Following the same procedure as for a linear response, we obtain

$$\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n) = \frac{-i}{\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)}{\omega_j-\omega'}\,d\omega' \qquad [18]$$

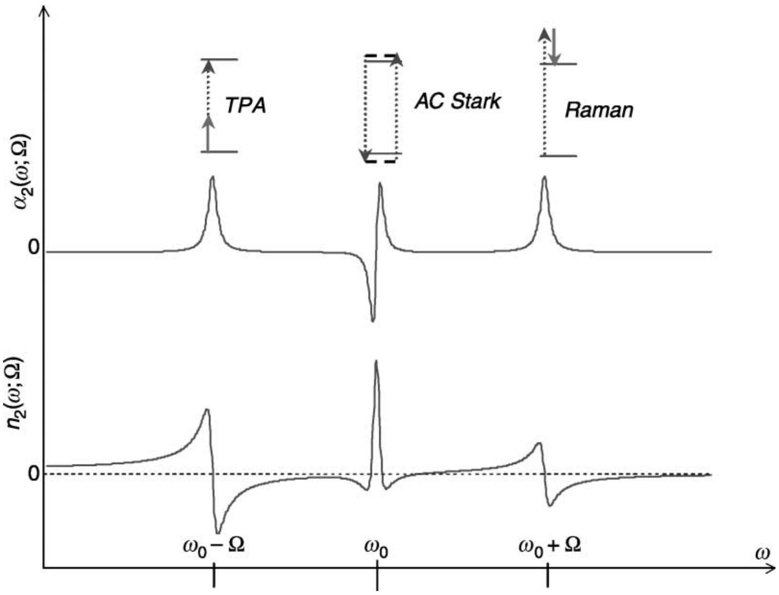

Figure 2 Upper trace: the nonlinear absorption coefficient in a fictitious 'degenerate' two-level system. Lower trace: the resulting nonlinear refractive index obtained using the KK relations. The insets show the three possible physical mechanisms involved.

By separating the real and imaginary parts of this equation, we get the generalized Kramers–Krönig relation pairs for a nondegenerate, *n*th-order nonlinear susceptibility:

$$\Re e\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \qquad [19]$$

and

$$\Im m\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\} = -\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Re e\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \qquad [20]$$

In particular, for $\chi^{(3)}$ processes having $\omega_1=\omega_a,\omega_2\omega_b$, and ${}_3b$, this becomes identical to eqn [16].

Note that in describing the nonlinear susceptibilities, no special attention was given to the harmonic generating susceptibility $\chi^{(N)}(N\omega)\equiv\chi^{(N)}(\omega,\omega,\ldots\omega)$, i.e., the susceptibility generating the *N*th harmonic at $N\omega$. It turns out that in addition to the KK relations given by eqns [19] and [20], the real and imaginary parts of $\chi^{(N)}(N\omega)$ can also be related in different sets of dispersion integrals that involve only the degenerate forms of the susceptibilities. A more general yet simple analysis gives the most general form of KK relations for any type of $\chi^{(n)}$:

$$\chi^{(n)}(\omega_1+p_1\omega,\omega_2+p_2\omega,\ldots,\omega_m+p_m\omega) = \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1+p_1\Omega,\omega_2+p_2\Omega,\ldots,\omega_m+p_m\Omega)}{\Omega-\omega}\mathrm{d}\Omega \qquad [21]$$

for all $p_1,p_2,\ldots,p_m\geq 0$. Setting $\omega_1=\omega_2=\cdots=\omega_m\equiv 0$, and $p_1=p_2=\cdots=p_m=1$ in eqn [21] yields an interesting form of the KK relations for the *N*th-harmonic susceptibilities:

$$\Re e\{\chi^{(N)}(N\omega)\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(N)}(N\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \qquad [22]$$

These dispersion relations have allowed calculations of $\chi^{(2)}(2\omega)$ and $\chi^{(3)}(3\omega)$ in semiconductors using full band structures.

At the beginning of this article, it was noted that all the KK relations for nonlinear optics were known in the early days of the field. Their application in unifying nonlinear absorption (in particular two-photon absorption) and the optical Kerr effect (n_2) in solids came only much later. More recent work demonstrated that the KK relations are a powerful analytical tool in nonlinear optics. Following the picture of a degenerate two-level system shown in Figure 2, a simple two-band model has been used to calculate the nonlinear absorption coefficient, $\alpha_2(\omega_1;\omega_2)$, resulting from three mechanisms: 2PA, the Raman absorption process, and the ac Stark effect. The optical Kerr coefficient $n_2(\omega_1;\omega_2)$ was then calculated using eqn [15]. Of particular practical interest is the degenerate case ($\omega_1=\omega_2=\omega$), from which the 2PA coefficient $\beta(\omega)=\alpha_2(\omega;\omega)$ can be extracted. Figure 3 depicts the calculated dispersion of n_2 and β as a function of $\hbar\omega/E_g$ where E_g is the bandgap energy of the solid. The dispersion of n_2 and its sign reversal shown in Figure 3 has been observed experimentally in many optical solids.

Finally, let us discuss a related implication of causality in nonlinear optics. The KK dispersion relations are traditionally derived in terms of internal material parameters such as susceptibility, absorption coefficient, and refractive index. Similar to the case of electrical circuits, one can obtain dispersion relations that apply to an external transfer function of the system that relates an input signal to an output signal. In this case, the dispersion of the transfer function includes system structure as well as the intrinsic dispersion of the material. As an optical (and linear) example, consider a Fabry–Perot etalon. The optical transmission of this system has well-known spectral

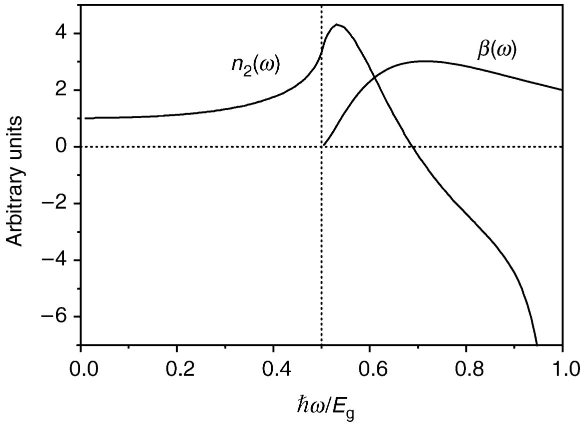

Figure 3 The two-photon absorption coefficient in semiconductors (β) calculated for a two-band model. The resultant nonlinear refractive index (n_2) obtained using a KK transformation of the calculated nondegenerate nonlinear absorption coefficient includes all major mechanisms.

features that are primarily caused by structural dispersion (i.e., interference) in addition to the intrinsic dispersion of the material. Causality still demands that the transmitted signal has a phase variation whose value and dispersion can be determined using a KK relation linking the real and imaginary parts of the transmission coefficient. In other words, the KK relations provide a spectral correlation between the real and imaginary components of the transfer function which in turn may translate to a spectral correlation between the phase and amplitude of the transmitted signal. However, the variations in phase do not necessarily imply the presence of a varying index of refraction, nor does an amplitude variation suggest the existence of material absorption (dissipation). Ultimately, this implies that any mechanism causing a variation in amplitude (including reflection, scattering, or absorption) must be accompanied by a phase variation. (One should note that the reverse of the previous statement is not necessarily true; i.e., a variation in phase does not have to be accompanied by an amplitude modulation.)

In nonlinear optics with the 'black box' approach of Figure 1, the optical perturbation ξ (with frequency Ω) can render an amplitude variation in the probe (at ω) using various frequency mixing schemes in a noncentrosymmetric material (i.e., with nonzero $\chi^{(2)}$). For instance, the probe at ω can be depleted by nonlinear conversion to $\omega_{\text{sum}} = \omega + \Omega$ via sum-frequency generation involving $\chi^{(2)}(\omega, \Omega)$ and/or to $\omega_{\text{diff}} = \omega - \Omega$ via difference-frequency generation involving $\chi^{(2)}(\omega, -\Omega)$. Such a conversion (or depletion) should be accompanied by a phase variation according to the KK dispersion relations. This type of nonlinear phase modulation is known as a $\chi^{(2)}$:$\chi^{(2)}$ cascaded nonlinearity. Such cascaded processes are routinely (and more simply) analyzed with Maxwell's equations governing the propagation of beams in a second-order nonlinear material. The KK relations, however, provide an interesting physical perspective of the process. We find that cascaded second-order nonlinearities are yet another manifestation of causality in nonlinear optics.

List of Units and Nomenclature

α	linear absorption coefficient
α_2	nonlinear absorption coefficient
β	two-photon absorption coefficient
$\chi^{(n)}$	nth-order nonlinear optical susceptibility
n	linear refractive index
n_2	nonlinear refractive index coefficient, coefficient of optical Kerr effect
$\Theta(t)$	step function
$\wp$	principal value
2PA	two-photon absorption
KK relations	Kramers–Krönig relations

See also

Materials for Nonlinear Optics: Liquid Crystals for NLO. **Nonlinear Optics, Basics:** Cascading; Nomenclature and Units.

Further Reading

Bassani F and Scandolo S (1991) Dispersion-relations and sum-rules in nonlinear optics. *Physical Review B–Condensed Matter* 44: 8446–8453.

Caspers PJ (1964) Dispersion relations for nonlinear optics. *Physical Review A* 133: 1249–1251.

Hutchings DC, Sheik-Bahae M, Hagan DJ and Van Stryland EW (1992) Kramers–Krönig relations in nonlinear optics. *Optical and Quantum Electronics* 24: 1–30.

Kogan SM (1963) On the electromagnetics of weakly nonlinear media. *Soviet Physics JETP* 16: 217–219.

Nussenzveig HM (1972) Causality and Dispersion Relations. New York: Academic Press.

Price PJ (1964) Theory of quadratic response functions. *Physical Review* 130: 1792–1797.

Ridener FLJ and Good RHJ (1975) Dispersion relations for nonlinear systems of arbitrary degree. *Physical Review B* 11: 2768–2770.

Sheik-Bahae M and Van Stryland EW (1999) Optical nonlinearities in the transparency region of bulk semiconductors. In: Garmire E and Kost A (eds) *Nonlinear Optics in Semiconductors I*, 58, pp. 257–318. Academic Press.

Sheik-Bahae M, Hutchings DC, Hagan DJ and Van Stryland EW (1991) Dispersion of bound electronic nonlinear refraction in solids. *IEEE Journal of Quantum Electronics* 27: 1296–1309.

Smet F and Vangroenendael A (1979) Dispersion-relations for *N*-order non-linear phenomena. *Physical Review A* 19: 334–337.

Toll JS (1956) Causality and the dispersion relation: logical foundations. *Physical Review* 104: 1760–1770.

Nomenclature and Units

M P Hasselbeck, The University of New Mexico, Albuquerque, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Nomenclature Associated with the Excitation Light

Particular care must be used when characterizing the excitation beam in nonlinear optical experiments compared to linear measurements. By definition, nonlinear optical phenomena depend on the electric field to high order. The higher the order, the more sensitive the observed behavior depends on the input. Extracting a representative nonlinear coefficient from an experiment, for example, becomes progressively more difficult as the order of the optical nonlinearity increases. In other words: the errors associated with optical beam characterization get magnified by the order of the nonlinearity under study.

There is an extensive nomenclature for characterizing the light (almost always laser light) interacting with the nonlinear optical medium. Different descriptions may be used depending on whether the excitation light is pulsed, continuous, or a continuous train of pulses. When a laser beam is constant or continuous, it is often described by its ‘cw power’. ‘Cw’ stands for ‘continuous wave’, an acronym taken from the nomenclature of electronics. The cw power of a laser is determined by placing a power meter in the path of a beam. Repetitively pulsed lasers can be characterized in the same way, provided the response time of the power meter is slower than the pulse separation period. This will almost certainly be the case with a continuously pumped mode-locked laser such as a dye laser, fiber laser, or Ti:sapphire laser, which produce pulses at repetition frequencies of tens of megahertz. Cw power may also be suitable for describing pulsed flashlamp lasers, gas discharge lasers, or any laser that is excited in a periodic fashion.

When the optical output can be distinguished as individual pulses, additional metrics are used. A pulse, by definition, exists during a window of time, i.e., there is a time when light is present and a time when light is absent. How one characterizes the time light is present – the pulse duration – is critically important and not always obvious. There are a myriad of ways the pulse temporal envelope can manifest itself; common examples include square pulses, triangular pulses, Gaussian pulses, and hyperbolic secant pulses. These names refer to the mathematical waveforms that map the pulse envelope as a function of time.

A pulse waveform that exhibits symmetry about the peak in its temporal envelope is commonly characterized by its ‘full-width, half-maximum’, abbreviated FWHM. One obtains the full-width, half-maximum by locating the two points on the pulse profile that are at half the peak value. The temporal separation of these two points is the FWHM. One uses this measure because in a strict mathematical sense, waveforms such as the Gaussian or hyperbolic-secant exist even at times $t = \pm\infty$. A less common term is the ‘half-width, half-maximum’ or HWHM. As the name implies, the HWHM is exactly half the FWHM value. Not all light-pulses are temporally symmetric, however, so greater detail may be needed when giving a mathematical description of asymmetric pulses.

Light pulses are also characterized by their energy and peak power. The temporal envelope recorded by a so-called square-law detector (all laboratory detectors are square-law detectors) shows the power of the pulse as a function of time, provided the detector response time is sufficiently fast. Integrating this waveform gives the pulse energy. It makes no sense to talk about the energy of a cw beam of light, unless that beam is composed of repetitive, distinguishable pulses. Each pulse in the train carries a distinct amount of optical energy.

‘Ultrafast’ or ‘ultrashort’ laser pulses generally refer to light pulses that are of such small duration they cannot be measured directly with detectors and oscilloscopes because of bandwidth limitations. To infer the pulse duration, so-called intensity autocorrelation measurements are often made. The temporal power profile is then deduced from the autocorrelation signal with the appropriate mathematical conversion factor. Equally important is the spectrum of a short pulse. The optical bandwidth determines the lower limit of temporal compression; such ideally compressed pulses are said to be ‘transform limited’. If different components of the spectrum can be distinguished at specific positions on the pulse temporal profile, then the pulse is ‘chirped’. Information on the methods of ultrafast laser science can be found in **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Nonlinear optical effects take place when light is concentrated on a target – there is a cross-section or ‘footprint’ of the beam on the medium.

The interaction of the light and material takes place in a region called the beam area, beam cross-section, focal area, or focal volume. One must then define the appropriate interaction area or volume.

Specification of the beam area allows one to define two important quantities used frequently in nonlinear optics. The power/area is known as irradiance or power density. This quantity is commonly called the intensity, although the strict radiometric definition of intensity is power/(solid angle). The second important quantity is energy/area, which is called the energy density or fluence.

The convention in nonlinear optics is to define the optical electric field as:

$$\mathbf{E}(\mathbf{z},t) = \frac{1}{2}E_0 \exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{z} - \omega t)] + \mathrm{c.c.} \qquad [1]$$

where E_0 is the peak field, $\mathbf{k}$ is the propagation vector, ω is the angular frequency of the light, and c.c. stands for complex conjugate. This represents a forward- and backward-propagating (in the z-direction) infinite, transverse plane wave. Other definitions are also used. In SI/mks units and using the convention of eqn [1], the irradiance (I in units of W/m^2) inside a material of refractive index n is related to the electric field vector of the light ($\mathbf{E}$ in units of V/m) as follows:

$$I = \frac{1}{2}cn\varepsilon_0|\mathrm{E}|^2 \qquad [2]$$

Here c is the speed of light (2.998×10^8 m/s) and ε_0 is the permittivity of free space (8.854×10^{-12} F/m). Nonlinear optics has an unfortunate tradition of mixing mks and cgs units, resulting in a lot of confusion. Irradiance is usually expressed in units of W/cm^2. One can calculate the peak electric field (in units of V/cm) of a laser beam given its irradiance (in units of W/cm^2) by using this simple formula:

$$E_0 = 38.82\sqrt{\frac{I}{n}}\left(\frac{\mathrm{V}}{\mathrm{cm}}\right) \qquad [3]$$

The root-mean-square value of the electric field is obtained by replacing the factor 38.82 by 27.45. In Gaussian/cgs units, irradiance is expressed in units of ergs/(cm^2 sec) and is related to the field as:

$$I = \frac{cn}{8\pi}|\mathbf{E}|^2 \qquad [4]$$

where the field is in units of statvolt/cm and $c = 2.998 \times 10^{10}$ cm/sec. Conversion between mks and cgs for the electric field is 3×10^4 V/m = 1 statvolt/cm; irradiance is converted using 1 erg/(cm^{2}sec) = 1×10^{-3}W/m^2. The following point must be emphasized: depending on how the electric field is defined, there can be factors of 2 or even 4 discrepancies in the irradiance values quoted by different authors. The common definitions are used here, although they are certainly not universal.

A very common realization of the spatial irradiance profile of a laser beam is the radially symmetric Gaussian function:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [5]$$

where I_0 is the peak irradiance at the center of the beam ($r = 0$) and w is the 'spot size'. The spot size changes continuously as the beam propagates and the minimum spot size is known as the waist. One often hears the spot size referred to as the radius, but the radius of a Gaussian beam traditionally denotes the curvature of the phase front. By definition, the phase front radius is infinite at the waist. The factor of 2 appearing in the argument of the exponential in eqn [5] stems from the fact that the spot size w is conventionally defined for the electric field of the Gaussian beam. When the field is squared to obtain the irradiance, the factor of 2 appears. Using this function for the irradiance spatial profile leads to the common parameter '1/e^2 diameter'. The irradiance falls to 1/e^2 (0.1353) of its peak value when $r = w$.

The Gaussian spatial profile or 'Gaussian beam' results when the laser has been designed to operate in the lowest-order transverse mode, usually denoted TEM$_{00}$. The TEM$_{00}$ Gaussian beam is particularly convenient because the same relative power profile is maintained in the near- and far-field, whether or not the beam is collimated or focused.

The cross-sectional area of a TEM$_{00}$ Gaussian beam normally incident on a surface is found as follows. The power in the beam is obtained by integrating the irradiance profile (I) over the surface:

$$P = \iint_{\pm\infty} I \mathrm{d}A \qquad [6]$$

Referring to eqn [5], the integral becomes:

$$P = \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} r\,\mathrm{d}r\, I_0 \exp\left(-\frac{2r^2}{w^2}\right) = I_0\frac{\pi w^2}{2} \qquad [7]$$

Hence the effective area of a TEM$_{00}$ Gaussian beam normally incident on a surface is:

$$A = \frac{\pi w^2}{2} \qquad [8]$$

There are many situations where the Gaussian formulation of eqn [5] is not suitable. Beams produced by unstable laser resonators, for example, are not Gaussian. A multitransverse mode beam profile obtained from a stable resonator does not usually lend itself to a simple mathematical characterization.

Nomenclature Associated with the Nonlinear Optical Medium

In nonlinear optics texts, the subject is often introduced by writing the macroscopic polarization in Maxwell's equations as a power series expansion in the electric field. This approach, proposed by Bloembergen and coworkers in the early 1960s, has been spectacularly successful for interpreting experiments, though it has also led to confusion in the definition of nonlinear optical coefficients. The confusion, which stems from arbitrary definitions and nomenclature used by different authors, cannot be resolved here. The reader should, however, be alert for these discrepancies. Comparison of results published by different laboratories requires that the fundamental equations for extracting the nonlinear coefficients from their data are known. As the discipline has matured, the nonlinear optics community largely recognized the ambiguities and confusion; precise definitions of terms and coefficients are now commonly provided in the research literature.

It should also be noted that the power-series framework may not always be the best formulation for modeling and characterizing nonlinear optical effects. A carefully designed optical limiter, for example, may exhibit an abrupt decrease of transmission at a specific 'threshold irradiance' or 'threshold fluence', perhaps at a given laser pulse duration. The physical phenomenon or phenomena driving this behavior may not readily succumb to characterization by an *n*th-order coefficient in a power series expansion. In this situation it may be appropriate to specify a threshold optical input parameter. Another example is a homogeneously broadened saturable absorber, in which the absorption of an optical medium decreases with increasing input irradiance. The irradiance-dependent absorption coefficient α (refer to eqn [26]) is characterized by a 'saturation irradiance' I_{sat}:

$$\alpha(I) = \frac{\alpha_0}{1 + \dfrac{I}{I_{\text{sat}}}} \qquad [9]$$

where α_0 is the linear absorption coefficient.

Nonlinear Susceptibility

The power-series expansion of the macroscopic polarization is the standard approach for modeling nonlinear optical behavior and categorizing the various phenomena. A common way to write the polarization, nonlinear in the electric field, is (mks units):

$$\mathbf{P} = \varepsilon_0\left(\chi^{(1)}\cdot\mathbf{E} + \chi^{(2)} : \mathbf{EE} + \chi^{(3)} \vdots \mathbf{EEE} + \ldots\right) \qquad [10]$$

where the polarization **P** is a time- and space-dependent vector and the terms $\chi^{(n)}$ are the various orders of the nonlinear susceptibility (the cgs equation is obtained by dropping the free-space permitivitty coefficient ε_0). A second-order effect is associated with $\chi^{(2)}$, a third-order with $\chi^{(3)}$, and so on. In general, there are distinguishable electric field vectors (in eqn [10], the fields have not been individually designated for clarity). The susceptibility terms are generally tensors, which means the medium is sensitive to the orientation of the input fields. The dots in eqn [10] indicate tensor products. The input field vectors can have different frequencies and the presence of complex conjugates in eqn [1] indicates that the frequency components will have both + and − signs, i.e. the *j*th electric field term will have associated frequency factors $\exp(\pm \mathrm{i}\omega_j t)$.

The nomenclature used here can be illustrated by example. In the case of the second-order nonlinear polarization (mks units):

$$\mathbf{P}_j = \varepsilon_0 \chi^{(2)}_{jkl} : \mathbf{E}_k\mathbf{E}_l \qquad [11]$$

Note that subscripts have been introduced. The indices correspond to Cartesian space vectors. This equation says that the polarization in the *j*-direction results from the tensor product of the appropriate $\chi^{(2)}$ with the input fields at $\mathbf{E}_k$ and $\mathbf{E}_l$. The *j*th component of polarization has an oscillation frequency that is determined by mixing of the input fields. In this example, the indices *j*, *k* and *l* can each take the value of *x*, *y*, and *z*. Let each input field be at one of two possible frequencies, call them ω_1 and ω_2. There are an enormous number of permutations (81 to be exact) of eqn [11]. One term is:

$$P_x(\omega_1+\omega_2) = \varepsilon_0\chi^{(2)}_{xzy}(\omega_1+\omega_2)E_z(\omega_1)E_y(\omega_2) \times \exp[-\mathrm{i}(\omega_1+\omega_2)t] \qquad [12]$$

Another possibility is:

$$P_z(\omega_2-\omega_1) = \varepsilon_0\chi^{(2)}_{zyx}(\omega_2-\omega_1)E_y(\omega_2)E_x(-\omega_1) \times \exp[-\mathrm{i}(\omega_2-\omega_1)t] \qquad [13]$$

Equation [12] corresponds to sum frequency generation $(\omega_1 + \omega_2)$ and eqn [13] shows difference frequency generation $(\omega_2 - \omega_1)$.

Not explicitly written on the right-hand side of the above equations is the exponential containing the mixing of the propagation vectors. To maximize the nonlinear polarization, which is the source of nonlinear behavior in Maxwell's equations, the vector sum $\sum \mathbf{k}_j$ for *all* the interacting fields should be close to zero. Arranging the propagation vectors to accomplish this is known as 'phase matching'.

It is important to point out that there are eight more terms describing the nonlinear polarization at $P_x(\omega_1 + \omega_2)$ in addition to eqn [12]; likewise for $P_z(\omega_2 - \omega_1)$. Also note that the indices k and l may be identical, hence a nonlinear polarization driven by $\chi^{(2)}_{xyy}$ for example, is allowed. Higher-order nonlinear polarizations get progressively more complicated. The general third-order nonlinear polarization is

$$\mathbf{P}_j = \varepsilon_0 \chi^{(3)}_{jklm} \vdots \mathbf{E}_k \mathbf{E}_l \mathbf{E}_m \qquad [14]$$

Sometimes the second-order nonlinear susceptibility is written with the coefficient d instead of $\chi^{(2)}$. The relation between the d-coefficient and $\chi^{(2)}$ depends arbitrarily on how it is defined; a common definition is

$$d_{jkl} = \frac{1}{2}\chi^{(2)}_{jkl} \qquad [15]$$

but the reader is cautioned that the factor $\frac{1}{2}$ is sometimes missing. In the preceding discussion it was pointed out how there can, in principle, be a large number of tensor components involved in the expansion of the nonlinear polarization. Simplification occurs when it is realized that there is no discernible physical difference between the frequency terms ω_j and $-\omega_j$ and that ordering of the fields in eqn [14] – which suggests a time order in the arrival of the fields – is irrelevant. These symmetry arguments show that $d_{jkl} = d_{jlk}$, which reduces the number of independent d-coefficients from 81 to 18. A simpler subscript notation is then introduced that uses the integers 1–6 to represent pairs of Cartesian components k and l. This reduces the number of subscripts from three to two. An example of this notation is $d_{xzz} = d_{14}$.

In certain situations, one can take advantage of crystal symmetry to further reduce the complicated summations to a single scalar coefficient for the nonlinear susceptibility, which is referred to as 'd-effective'. For these specialized cases, the nonlinear polarization is

$$P = \varepsilon_0 d_{\text{eff}} E_1 E_2 \qquad [16]$$

In SI/mks units, the polarization is in units of coul/m^2. The second-order susceptibility (i.e. $\chi^{(2)}$, d_{jkl}) must therefore have units of m/volt. The units of $\chi^{(3)}$ are m^2/volt^2, $\chi^{(4)}$ is m^3/volt^3, and so on. In Gaussian/cgs units, $\chi^{(2)}$ has dimensions cm/statvolt, $\chi^{(3)}$ will be cm^2/statvolt^2, etc. although sometimes in the Gaussian system *all* the coefficients $\chi^{(n)}$ are discussed with shorthand 'electrostatic units' or 'esu'.

Complex Quantities

In linear optics, the susceptiblity has real and imaginary parts

$$\chi^{(1)} = \chi^{(1)}_{\text{real}} + \mathrm{i}\chi^{(1)}_{\text{imaginary}} \qquad [17]$$

The complex linear index is, in turn, written as the sum of real and imaginary components, derived from the linear susceptibility as follows (mks units):

$$\sqrt{1 + \chi^{(1)}} = n_0 + \mathrm{i}\kappa \qquad [18]$$

where n_0 is the linear refractive index and κ is the imaginary term leading to absorption of light. In general, the nonlinear susceptibilty $\chi^{(n)}$ is also a complex number:

$$\chi^{(n)} = \chi^{(n)}_{\text{real}} + \mathrm{i}\chi^{(n)}_{\text{imaginary}} \qquad [19]$$

As in linear optics, the complex notation is a bookkeeping method that conveniently accounts for what is known as 'resonant enhancement'. Nonlinear optics research has revealed that the nonlinear susceptibility $\chi^{(n)}$ can be a strong function of frequency. Specifically, this quantity will be strongly enhanced when sums and/or differences of the photon energies in the interacting light beams coincide with quantum mechanical energy resonances in the material. When a resonance condition is achieved, $\chi^{(n)}$ will be dominated by its imaginary component. Far from the resonances, $\chi^{(n)}$ behaves more like a real quantity. These complex terms directly enter the wave equations describing how light propagates in a nonlinear medium and are particularly important for $\chi^{(3)}$; the complex quantity $\chi^{(3)}$ determines whether light will be refracted or absorbed and by how much. The real part of $\chi^{(3)}$ drives nonlinear refraction while the imaginary portion characterizes nonlinear absorption (e.g., two-photon absorption) and the inverse effect – gain. This rich subject – in particular the critical importance of the wavelength dependence of the nonlinear phenomena – is considered in **Nonlinear Optics, Basics:** Kramers–Krönig Relations in Nonlinear Optics.

Nonlinear Refraction

The most common manifestation of nonlinear refraction arises from the third-order nonlinear susceptibility, the so-called optical Kerr effect. In this case, the refractive index is linearly proportional to the irradiance (I) of a monochromatic light beam. The irradiance-dependent refractive index is

$$n(I) = n_0 + n_2 I \qquad [20]$$

where n_0 is the linear, irradiance-independent refractive index and n_2 is the nonlinear refractive index coefficient. Because n is a dimensionless quantity, n_2 must be in units of area/power.

The optical Kerr effect is usually written in mks units with the expression shown in eqn [20]. In cgs units, the common (but by no means universal) convention is to write:

$$n(E) = n_0 + \frac{1}{2}\tilde{n}_2 |\mathbf{E}|^2 \qquad [21]$$

where the field is in units of statvolts/cm. The nonlinear coefficient $\tilde{n}_2$ (cgs) must therefore be in units of cm^2/statvolt^2, which is sometimes abbreviated to 'esu'. The conversion for n_2 between these two equations is:

$$\tilde{n}_2(\text{cgs}) = \frac{n_0 c}{40\pi} n_2(\text{mks}) \qquad [22]$$

where n_2(mks) is in units of m^2/W and c is in m/sec. Also useful is the relation between n_2 and $\chi^{(3)}$, written in mks units as:

$$\Delta n = n_2 I = \frac{\mathrm{Re}\left(\chi^{(3)}\right)}{4\varepsilon_0 n^2 c} I \qquad [23]$$

where 'Re' denotes the real part of $\chi^{(3)}$. In cgs units we have:

$$\Delta n = \frac{1}{2}\tilde{n}_2 |\mathbf{E}|^2 = \frac{4\pi^2 \mathrm{Re}\left(\tilde{\chi}^{(3)}\right)}{n^2 c} I \qquad [24]$$

where I is in units of ergs/(cm^2 sec) defined by eqn [4]. The reader is again advised that different authors will show discrepancies of 2, 4, or even 8 when writing these equations. The susceptibilities $\chi^{(3)}$ are related as follows:

$$\tilde{\chi}^{(3)}(\text{cgs}) = \frac{9 \times 10^8}{4\pi} \chi^{(3)}(\text{mks}) \qquad [25]$$

If the coefficients n_0 and n_2 exist, does the nomenclature imply there are terms n_1, n_3, and others? These coefficients are certainly allowed, but not often seen in discussions of nonlinear refraction. When one refers to nonlinear refraction, the common understanding is that the index depends linearly on irradiance, which is conveniently modeled by eqn [20]. But there are other physically relevant situations to consider. The coefficient n_1, for example, describes the linear electro-optic effect in which the change in index is linearly proportional to the electric field (although it is called the linear electro-optic effect, it is actually derived from the second-order nonlinear susceptibility). The electric field can be the oscillating field of a laser beam, for example, or a dc field applied to an electro-optic crystal (e.g., Pockel's cell). When other terms are added to eqns [20] or [21], the units of the coefficients must be chosen to keep the equation dimensionless.

In the preceding discussion, there is an implication that the nonlinear index is an instantaneous function of the irradiance or field. Although the material system can never respond instantaneously, this is a good approximation in many situations. Sometimes it is not. Consider a pulsed laser beam that heats the material. When the local temperature increases, the linear refractive index can change. A short, Q-switched laser can have a pulse duration far less than a microsecond, while the temperature change it induces can last many orders of magnitude longer. This means optical modification of the refractive index may persist long after the exciting pulse has vanished, i.e., when the irradiance is at zero.

Another example is when a laser beam promotes electrons from their ground state to higher energy states of the material system. These excited electrons may modify the refractive index of the material. In general, the excited electrons remain in high-energy states for some period of time before relaxing to the ground level – if this time is longer than the excitation, the change of the refractive index persists beyond the time the laser beam is present. In these situations, the model of nonlinear refraction suggested by the above equations is inaccurate. One cannot obtain the nonlinear refraction from a simple algebraic analysis. A system of time-dependent equations – describing the dynamics of excitation and relaxation – must be solved using numerical procedures. Such phenomena are sometimes loosely categorized as 'dynamic linear optical effects' or 'effective third-order nonlinearities'.

Nonlinear Absorption

Consider a single-frequency light beam passing through an optically absorbing region of length L.

For simplicity, neglect reflections caused by surfaces that may define the region of interest, i.e., ignore reflections at surfaces that may be located on the optical axis z at points $z = 0$, $z = L$, or any other point in the path. Linear absorption means that the optical power extracted from the light beam as it traverses the absorbing medium is a direct function of the power at a given point. This is described mathematically by an elementary, linear differential equation known as Beer's law:

$$\frac{\mathrm{d}}{\mathrm{d}z} I(z) = -\alpha I(z) \tag{26}$$

The constant of proportionality is α, which is the linear absorption coefficient. Equation [23] is solved by direct integration:

$$\int_{I(0)}^{I(L)} \frac{\mathrm{d}I}{I} = -\alpha \int_0^L \mathrm{d}z \tag{27}$$

This has the solution:

$$\frac{I(L)}{I(0)} = \exp(-\alpha L) \tag{28}$$

which means the irradiance decreases exponentially as a function of propagation distance in a linearly absorbing medium.

In the nonlinear regime, we don't expect a classical Beer's law model to hold. By definition, the absorption will be a nonlinear function of irradiance at a given point. One makes the following power-series expansion to describe nonlinear absorption of monochromatic light:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -\alpha I - \beta I^2 - \gamma I^3 - \cdots \tag{29}$$

The coefficient β corresponds to a two-photon absorption process and γ is the coefficient of three-photon absorption. What about four-photon and even higher-order processes? These are rarely encountered, but certainly possible. To make these distinctions, eqn [29] is sometimes written with coefficients K_i or α_i instead of the sequential Greek alphabet:

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_1 I - K_2 I^2 - K_3 I^3 - K_4 I^4 - \cdots \tag{30}$$

The units of the various absorption coefficients must maintain the dimensional consistency of eqns [29] and [30], which is irradiance/length or power/(length)^3. These are shown in Table 1.

If there is linear absorption, the nonlinear processes at that wavelength are often (but not always) negligibly weak. This means we have the following inequalities:

Table 1 Dimensions of optical absorption coefficients

Process	*Coefficient*	*Units*
Linear absorption	α, K_1	(length)$^{-1}$
Two-photon absorption	β, K_2	length/power
Three-photon absorption	γ, K_3	length^3/power^2
Four-photon absorption	K_4	length^5/power^3

$$K_1 I \gg K_i I^i, \qquad \text{where } i \geq 2 \tag{31}$$

Nonlinear absorption is generally observed in the wavelength region where the medium is transparent to low-irradiance light, i.e., where linear absorption is negligible. It should be emphasized that there are situations where linear absorption is large and in fact a crucial component of the aggregate nonlinear effect. We will return to this point later in the discussion. For the moment, we neglect linear absorption. If the energy of the incident photons and energy levels of the system permit it, the lowest-order nonlinear process is two-photon absorption, described by the equation

$$\frac{\mathrm{d}}{\mathrm{d}z} I = -K_2 I^2 \tag{32}$$

which can be solved by elementary integration:

$$\frac{I(L)}{I(0)} = \frac{1}{1 + I(0) K_2 L} \tag{33}$$

Unfortunately, the situation is rarely this convenient. The above integration has ignored the fact that nonlinear absorption may enable significant linear absorption. This is possible because nonlinear absorption promotes electrons from low- to high-energy states in the medium. The change in excited electron density associated with absorption of light (both linearly and nonlinearly) is called 'photocarrier generation'. These photocarriers (e.g., photo-electrons) may modify the linear absorption as well as the refractive properties of the material. If two-photon absorption causes the linear absorption to increase, for example, the assumption that allowed us to ignore K_1 when writing eqn [32] ceases to be valid. While eqn [32] was appropriate at the very start of the light–matter interaction, the generation of photocarriers after the passage of time may change that condition. The behavior of the system is therefore time-dependent, i.e., it is dynamic. Simple analytic solutions are almost always not possible when dealing with nonlinear absorption.

The time rate of change of the photocarrier population (N) is modeled by the following equation:

$$\frac{\partial}{\partial t}N = \frac{K_1 I}{\hbar\omega} + \frac{K_2 I^2}{2\hbar\omega} + \frac{K_3 I^3}{3\hbar\omega} + \frac{K_4 I^4}{4\hbar\omega} + \cdots - \text{recombination} - \text{diffusion} \qquad [34]$$

where ω is the angular frequency ($\omega = 2\pi f$) of the incident light. Also included are place-holders for loss processes such as recombination and diffusion, which may themselves have complicated mathematical descriptions. The coefficients K_i have exactly the same units, dimensions, and interpretation as in eqn [30] and Table 1; in the spirit of the preceding discussion, these coefficients may be time dependent. Note that the dimensions of eqn [30] describe absorption of light (irradiance/length), while eqn [34] models the photocarrier population density (length^{-3} time^{-1}). The units of these two equations are very different.

List of Units and Nomenclature

CARS	Coherent anti-Stokes Raman scattering
c.c.	Complex conjugate
CSRS	Coherent Stokes Raman scattering
DFG	Difference frequency generation
DFM	Difference frequency mixing
DFWM, FWM	(Degenerate) Four wave mixing
DRO	Doubly resonant OPO
EIT	Electromagnetically induced transparency
ESA	Excited state absorption
GVD	Group velocity dispersion
GVM	Group velocity mismatch
NLA	Nonlinear absorption
NLR	Nonlinear refraction
OFID	Optical free-induction decay
OPA	Optical parametric amplification/amplifier
OPG	Optical parametric generation/generator
OPL	Optical power limiter
OPO	Optical parametric oscillation/oscillator
2PA, TPA	Two-photon absorption
PCM	Phase conjugate mirror
QPM	Quasi phase matching
RIKES	Raman-induced Kerr effect spectroscopy
RSA	Reverse saturable absorption
SBS	Stimulated Brillouin scattering
SFG	Sum frequency generation
SHG	Second harmonic generation
SIT	Self-induced transparency
SRO	Singly resonant OPO
SRS	Stimulated Raman scattering
SRWS	Stimulated Rayleigh wing scattering
SVEA, SVAA	Slowly varying envelope (amplitude) approximation
THG	Third harmonic generation
TPF	Two-photon fluorescence

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Bloembergen N (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Boyd RW (1991) *Nonlinear Optics*. Boston, MA: Academic Press.

Levenson MD and Kano S (1988) *Introduction to Nonlinear Laser Spectroscopy*. Boston, MA: Academic Press.

Miller A, Miller DAB and Smith SD (1981) Dynamic nonlinear optical processes in semiconductors. *Advances in Physics* 30: 697–800.

Mills DL (1998) *Nonlinear Optics: Basic Concepts*. New York: Springer.

Newell C and Moloney JV (1992) *Nonlinear Optics*. Redwood City, CA: Addison-Wesley.

Sauter EG (1996) *Nonlinear Optics*. New York: Wiley.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Yariv A (1991) *Optical Electronics*, 3rd edn. New York: Holt, Rinehart, and Winston.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: Wiley.

Nonlinear Optical Phase Conjugation

B Y Zeldovich, University of Central Florida,
Orlando, FL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Phase conjugation (PC) beams and their applications are best illustrated by Figure 1, which shows the passage of a coherent planewave incident upon a transparent optical element with inhomogeneities of refractive index (Figure 1a). Reversibility of light propagation implies the existence of such an 'anti-distorted' beam, which reverts back to a planewave after reversing through the same inhomogeneities (Figure 1b). A real-valued monochromatic optical wave, $E_{\text{real}}(\mathbf{R}, t)$, may be represented by the complex amplitude $E(\mathbf{R})$ via $E_{\text{real}}(\mathbf{R}, t) = 0.5[E(\mathbf{R})\exp(-i\omega t) + E^*(\mathbf{R})\exp(i\omega t)]$, where $E^*(\mathbf{R})$ represents the complex conjugate of $E(\mathbf{R})$:

$$\begin{aligned} E(\mathbf{R}) &= |E(\mathbf{R})|\exp[i\varphi(\mathbf{R})] \\ E^*(\mathbf{R}) &= |E(\mathbf{R})|\exp[-i\varphi(\mathbf{R})] \end{aligned} \qquad [1]$$

Mathematical expression of time-reversal, $E_{\text{real}}(\mathbf{R}, t) \rightarrow E_{\text{real}}(\mathbf{R}, -t)$, becomes the exchange $E(\mathbf{R}) \leftrightarrow E^*(\mathbf{R})$ for monochromatic waves. The reversibility of propagation means that the complex propagating field $E(\mathbf{R})$ (for example, planewave $\exp(i\mathbf{k}\cdot\mathbf{R})$), is a wave equation solution equivalent to $E^*(\mathbf{R})$ (in this example, $\exp(-i\mathbf{k}\cdot\mathbf{R})$). Conjugation of complex amplitude means reversal of the sign of the phase, and a mixed label PC has been coined and nowadays is firmly established. Terms such as 'wave front reversal' and 'generation of time-reversed replica of the beam' are also used to describe PC.

Figure 1 also illustrates one of the most important applications of PC. If the element in question is a laser-type amplifier, then double-passage allows the extraction of energy from the optically inhomogeneous laser medium in the form of a perfectly collimated beam of diffraction-limited divergence. A PC device may also serve as a mirror of a laser resonator, resulting in the same beam-correction purpose.

One of the most practical and robust methods of PC is based on stimulated Brillouin back-scattering (SBS). The incident beam illuminates the SBS-active transparent medium, for example, liquids (CS_2, CCl_4, acetone, etc.), compressed gases (CH_4, SF_6, etc.), or solids (fused or crystalline quartz, glass, etc.). This 'pump' beam, $E(\mathbf{R}) \equiv E(\mathbf{r}, z)$, must possess well-developed transverse inhomogeneities of intensity $|E(\mathbf{r}, z)|^2$. Here, $\mathbf{r} = \{x, y\}$ is the part of coordinate vector transverse with respect to the central direction z of the beam in question, and $\mathbf{R} = \{\mathbf{r}, z\}$. These inhomogeneities may constitute a narrow focal waist in the case of weakly distorted focused beams, speckle-inhomogeneities in cases of strong distortions, or a combination thereof; see Figure 2, where solid lines symbolize the 'rays' of the incident pump.

Spontaneous scattering of the pump results in a multitude of possible transverse profiles $S(\mathbf{r}, z = 0)$ of the signal, whose 'rays' are depicted via the dotted lines on Figure 2. The signal is amplified exponentially due to SBS processes. This exponent in a simplified form may be represented as $|S(\mathbf{r}, z)|^2 \propto \exp(\int g dz)$, with the gain $g(z)$ being the result of transverse overlapping of the intensity profiles of signal and pump:

$$g(z) \approx \text{const} \cdot \langle |E(\mathbf{r}, z)|^2 \cdot |S(\mathbf{r}, z)|^2 \rangle / \langle |S(\mathbf{r}, z)|^2 \rangle \qquad [2]$$

Thus, the similarity of the intensity profile of the signal to that of the pump is rewarded by the exponential preference in the output signal level. However, this is not enough to guarantee that the output backward-scattered signal $S(\mathbf{r}, z)$ is phase

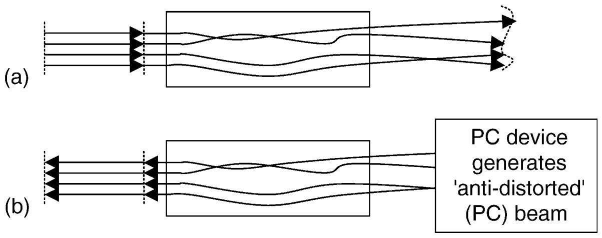

Figure 1 (a) Collimated beam is distorted by propagation through inhomogeneous medium. (b) Conjugate, a.k.a. antidistorted beam becomes collimated after backward passage through the same medium.

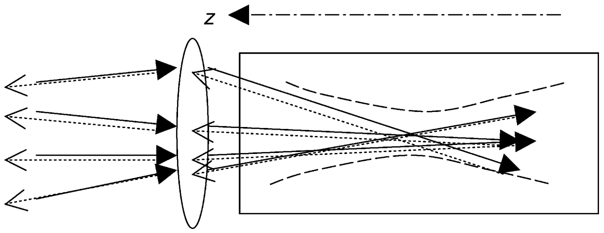

Figure 2 Stimulated Brillouin Scattering (SBS) method of PC. Solid lines symbolize the rays of incident 'beam', while dotted lines depict the rays of the signal amplified by the SBS process.

conjugate with respect to the pump. It is here that the mutual reversibility of propagation of strongly inhomogeneous signal and pump becomes crucial. Namely, good transverse overlapping is sustained along the entire interaction length z, only if $S(\mathbf{r}, z) \propto E^*(\mathbf{r}, z)$. This constitutes the 'discrimination mechanism' of SBS-PC: the 'conjugate mode' of the signal has a z-sustained advantage in exponential gain and, under appropriate conditions, suppresses all nonconjugate signal configurations in competition for pump power. Nonlinear-optical wave theory of this discrimination process has been developed further.

The holographic mechanism of PC has a different nature. It may take the form of static holography or dynamic holography; the latter essentially may be considered as a part of nonlinear optics. Here is a simplified description of one of the variants. Suppose one wants to obtain a conjugate replica of the incident monochromatic signal, whose complex profile is $S(\mathbf{R})$. A plane reference wave $A(\mathbf{R}) = \exp(i\mathbf{k}_A \cdot \mathbf{R})$ is also directed to the registering medium. If $S(\mathbf{R})$ and $A(\mathbf{R})$ are mutually coherent, then an interference pattern of intensity is produced $\{S^*(\mathbf{R})A(\mathbf{R}) + \text{compl. conj.}\}$. The medium records this pattern in the form of the modulation of refractive index and/or of the absorption coefficient, $\delta\varepsilon(\mathbf{R}) \propto (c_1 + ic_2)\{S^*(\mathbf{R})A\ (\mathbf{R}) + \text{c.c.}\}$ (Figure 3a). At the 'reconstruction' stage, the hologram is illuminated by another plane reference wave, $B(\mathbf{R}) = B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$. In the specific case, when $\mathbf{k}_B = -\mathbf{k}_A$, the source δD_{conj} of the reconstructed wave becomes, as shown by Figure 3b:

$$\delta D_{\text{conj}} = (c_1 + ic_2)[S^*(\mathbf{R})A(\mathbf{R})] \cdot B(\mathbf{R}) \propto S^*(\mathbf{R}) \times (c_1 + ic_2)A_0 B_0 \qquad [3]$$

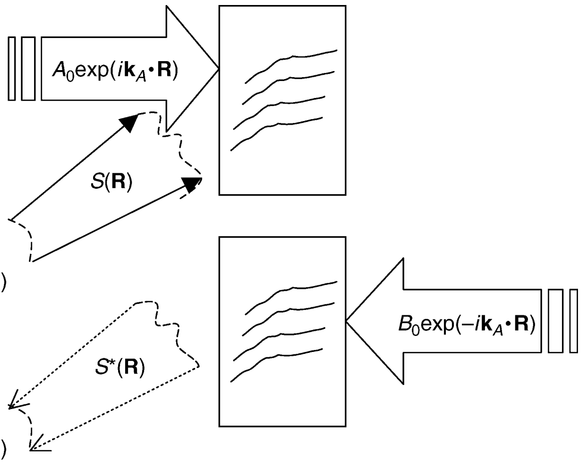
$A_0\exp(i\mathbf{k}_A\cdot\mathbf{R})$
$S(\mathbf{R})$
(a)
$B_0\exp(-i\mathbf{k}_A\cdot\mathbf{R})$
$S^*(\mathbf{R})$
(b)

Figure 3 (a) Recording of a hologram in a photosensitive medium by a reference wave A and signal wave S. (b) Read-out of that hologram by the wave B counter-propagating to A results in generation of the conjugated signal S^*.

Reversibility of propagation laws guarantees that this source will indeed excite a conjugate wave with high efficiency. Both processes, recording of the interference pattern and reconstruction of the conjugate wave, may occur simultaneously, if one deals with the dynamic holography. The same process may also be described as nonlinear-optical four-wave mixing (FWM) via cubic nonlinearity $\chi^{(3)}$:

$$\delta D_{\text{conj}}(\mathbf{R}) = \chi^{(3)} S^*(\mathbf{R}) A(\mathbf{R}) B(\mathbf{R}) = \chi^{(3)} S^*(\mathbf{R}) A_0 B_0 \qquad [4]$$

with the four waves; A, B, S, S^*.

Important characteristics of a PC device are: fidelity of conjugation, efficiency of returning back-reflected power/energy, and reaction time or build-up time. The fidelity parameter must show how close the output of the device $E_{\text{out}}(\mathbf{r})$ is to the perfectly conjugate profile $S^*(\mathbf{R})$ of the incident signal (up to an arbitrary constant complex factor). Among other definitions, the square modulus of the normalized transverse overlapping integral of the fields is often discussed:

$$f = \frac{|\iint E_{\text{out}}(x,y) S(x,y) \mathrm{d}x\mathrm{d}y|^2}{\left(\iint |E_{\text{out}}(x,y)|^2 \mathrm{d}x\mathrm{d}y \cdot \iint |S(x,y)|^2 \mathrm{d}x\mathrm{d}y\right)} \qquad [5]$$

To achieve good (close to 1) fidelity of PC in the holographic or FWM scheme, one has to guarantee that the reference waves, $A_0 \exp(i\mathbf{k}_A \cdot \mathbf{R})$ and $B_0 \exp(i\mathbf{k}_B \cdot \mathbf{R})$, are exactly conjugate to each other, $\mathbf{k}_A = -\mathbf{k}_B$. That, in turn, requires the absence of any distortions in the holographic or nonlinear medium. To the contrary, there are very modest requirements on the phase inhomogeneities of the medium in the SBS scheme of PC, and for that reason it is usually labeled as a scheme of self-phase conjugation.

A hybrid scheme, Brillouin-assisted FWM, has been suggested and implemented, where mutually conjugate reference waves are produced via SBS-PC, while FWM, in a separate medium, exploits SBS nonlinearity. This scheme has produced extremely high, up to about 10^{10}, reflectivity, accompanied by an extremely low-noise detection of incoming signals.

A number of other PC schemes were first realized with the use of an important class of materials for nonlinear optics: photo-refractive crystals (PRC). These are crystals where ionization of the dopants by the incident light creates interference pattern of charge separation. The linear electro-optic (Pockels) effect transforms the resulting patterns of the electrostatic field into patterns of refractive index, thus creating dynamic holograms. A remarkable property of these crystals is that most of them act as very good electrical insulators. Therefore, the only source of conductivity, which tends to erase the hologram, is photoconductivity, the value of the latter being proportional to the intensity of incident light itself. For this reason, the steady-state strength of the hologram turns out to be independent of the intensity. It is the build-up time that must be increased, if intensity is low. One can achieve various processes in PRC, including those similar to the $\chi^{(3)}$-type optical Kerr effect, stimulated scattering, etc.

The simplest use of PRC for generation of PC waves is to devise a FWM scheme (see Figure 3). However, a multitude of other, nontrivial schemes have been devised and implemented. The most impressive use of PRC for PC is based on interactions of the beams in the vicinity of a corner of a rectangular crystal, typically $BaTiO_3$, these interactions being of the stimulated scattering type. To honor J. Feinberg's cat, whose image was reconstructed in the first demonstration, this scheme is universally called 'cat conjugator'. It turned out to be a very reliable and robust scheme of self-phase-conjugation (SPC).

Another important group of the SPC scheme is called 'tail-biting'. One of the variants of the latter is shown in Figure 4. Incident 'pump' wave $A(\mathbf{R})$, whose transverse profile one wants to conjugate, enters the medium, and then is redirected by mirrors back into the same medium, to 'bite itself'. For the purpose of discussion, this redirected pump is labeled as wave $B(\mathbf{R})$ on its second arrival into the medium. Spontaneous scattering results in a seed for the amplification of the wave $C(\mathbf{R})$ via the process of stimulated scattering (SS) $A(\mathbf{R}) \rightarrow C(\mathbf{R})$.

Wave $C(\mathbf{R})$ is also redirected into the medium by the same mirrors, and is labeled as wave $D(\mathbf{R})$ on its second arrival into the medium. Both $C(\mathbf{R})$ and $D(\mathbf{R})$ are depicted in Figure 4 by white arrows. Wave $D(\mathbf{R})$ is also amplified due to the SS process $B(\mathbf{R}) \rightarrow D(\mathbf{R})$.

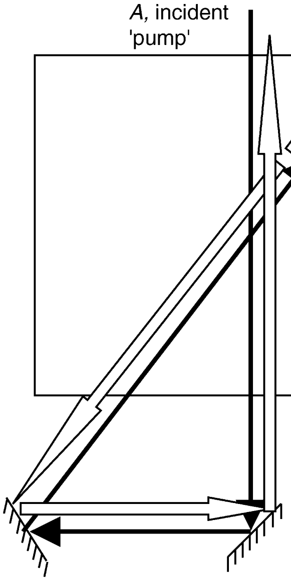

Figure 4 Tail-biting scheme of phase conjugation.

As it follows from the macroscopic description of SS process, volume gratings of refractive index are recorded in the medium, $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R}) + B^*(\mathbf{R})D(\mathbf{R})]$. Among all the transverse profiles for the seed $C(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})$ will be reflected by the mirrors into the medium in the form $D(\mathbf{R}) \propto A^*(\mathbf{R})$. The latter case is again a consequence of the reversibility of propagation laws, where both gratings have the same profile and add coherently in the form $A^*(\mathbf{R})B^*(\mathbf{R})$. What is even more important, the grating $B^*(\mathbf{R})D(\mathbf{R}) \propto B^*(\mathbf{R})A^*(\mathbf{R})$ serves to close the feedback loop for seeding the appropriate $C(\mathbf{R})$. This, and other variants of tail-biting schemes, have also proved to be reliable and robust. Quite often a laser amplifier is inserted in the path $A \rightarrow B$ and $C \rightarrow D$, thus enhancing the feedback.

Another important scheme is traditionally called 'double PC'; albeit it may be better described as 'mutual PC'. Figure 5 should help in understanding that scheme. Consider the beam $A(\mathbf{R})$ incident to a medium active with respect to the stimulated-scattering-type process $A(\mathbf{R}) \rightarrow C(\mathbf{R})$. This process originates from a random seed and results in a fan of different waves $C(\mathbf{R})$. Such a process is characteristic of PRCs and is a consequence of recording dynamic gratings of refractive index $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R})]$. Suppose another beam $B(\mathbf{R})$, typically incoherent with respect to $A(\mathbf{R})$, illuminates the medium from the other side, and also is engaged in the fanning process, this time $B(\mathbf{R}) \rightarrow D(\mathbf{R})$, with the grating of refractive index $\delta n(\mathbf{R}) \propto -i[B^*(\mathbf{R})D(\mathbf{R})]$ involved. Among all the transverse profiles for the seed grating $C(\mathbf{R})A^*(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})A^*(\mathbf{R})$ will be supported by similar profiles in the $B(\mathbf{R}) \rightarrow D(\mathbf{R})$ scattering. In this way, two fanning

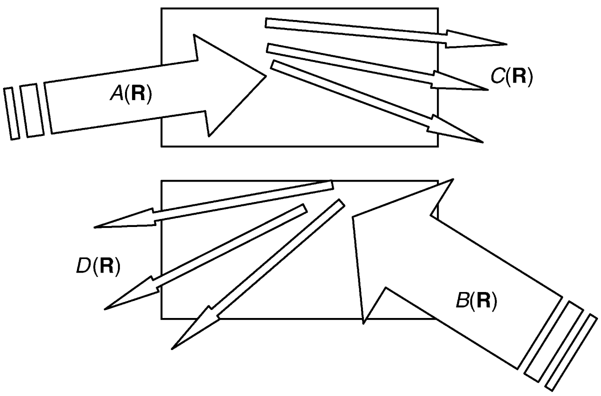

Figure 5 Towards the mechanism of double PC (mutual PC).

processes enhance each other and close the feedback loop. An important consequence is that one obtains the following output waves:

$$\begin{aligned} C(\mathbf{R}) &= \text{const}_1 \times B^*(\mathbf{R}) \\ D(\mathbf{R}) &= \text{const}_2 \times A^*(\mathbf{R}) \end{aligned} \qquad [6]$$

One can say that the output wave $C(\mathbf{R})$ presents a conjugate replica of $B(\mathbf{R})$-beam, while the output wave $D(\mathbf{R})$ presents a conjugate replica of $A(\mathbf{R})$-beam; hence the label 'mutual PC'. In some experiments the incident beams, $A(\mathbf{R})$ and $B(\mathbf{R})$, were of somewhat different colors (i.e., from different lasers), or their pulses did not overlap in time in the medium. It is the grating memory that 'informs' one beam about the presence of the other. The mechanism of cat conjugator is considered to be a combination of tail-biting and double PC schemes.

Pumped laser material with saturable gain often serves as an important nonlinear-optical medium for implementing PC. One of the evident advantages is that possible low efficiency of properly nonlinear process may be compensated by the gain in the same piece of material.

Considerable attention is paid nowadays to the chirp reversal of optical pulses that are used in fiber optical communications. Such chirped pulses, $S(t) = \exp(-i\omega_0 t + i\beta t^2/2 - t^2/\tau_1^2)$, with the value of the 'chirp' $\mathrm{d}\omega/\mathrm{d}t = -\beta$ and an increased pulse duration τ_1, arise as a result of propagation in a fiber which possesses group velocity dispersion (GVD). Pulse stretching and chirp are the consequences of different propagation times for the different frequency constituents of the pulse. If some device changes the sign of that chirp, then subsequent propagation of the pulse, through a piece of fiber with the same GVD, restores the duration of the pulse to original shorter value τ_0. A scheme has been suggested to use nonlinear optical FWM in a $\chi^{(3)}$-medium or three-wave mixing (ThWM) in a $\chi^{(2)}$-medium:

$$\begin{aligned} \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) A(t) B(t) \\ \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) C(t) \end{aligned} \qquad [7]$$

Indeed, complex conjugation of the time dependence of the signal's field is equivalent to the change of the chirp sign. The second (ThWM) expression is written assuming that the reference wave $C(t) \propto \exp(-2i\omega_0 t)$, so that is has the frequency double that of the signal. The same process of ThWM has been shown to yield the conjugation of transverse phase profile of the beams, but for a number of reasons it was not applied.

As for the applications of PC, one of them was discussed above (Figure 1: double-pass scheme of a laser amplifier). Lasers with one or two PC mirrors, or with more complicated PC scheme, promise generation of high-transverse-quality beams with the use of realistic laser media, the latter inevitably possessing thermal and other distorting inhomogeneities. Another important potential application is free-space optical communications: Earth–Earth or Earth–satellite through atmosphere, or satellite–satellite communications. An 'interrogating' beam may be sent in one direction through an imperfect optical system and/or through turbulent atmosphere. Conjugation of transverse profile of the 'interrogating' beam at the other end of the communication link leads to an almost perfect redirection of the reflected beam towards the 'beacon' source, while time modulation may carry the information.

Readout of information data from a holographic storage is almost always performed in the regime of reading the conjugate wave, since it corrects for the most part of aberrations of optical systems in question.

To conclude, one may use a citation from the two consecutive *Scientific American* popular papers on PC: 'at present the number of applications would seem to be limited only by imagination.'

Reviews of various aspects of phase conjugation may be found in monographs and *Scientific American* papers listed below. The author of the present article was introduced to PC by his colleague V. V. Ragulskii, and has greatly benefited from collaboration with V. V. Ragulskii and V. V. Shkunov.

See also

Microscopy: Phase Contrast Microscopy. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Scattering Theory.

Further Reading

Bespalov VI and Pasmanik GA (1994) *Nonlinear Optics and Adaptive Laser Systems*. Commack, NY: Nova Science Publishers.

Feinberg J (1986) Self-pumped, continuous wave phase conjugator using internal reflection. *Optics Letters* 7: 486.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. New York: Academic Press.

Gower M and Proch D (eds) (1994) *Optical Phase Conjugation*. Berlin: Springer Verlag.

Hellwarth RW (1982) Optical beam conjugation by stimulated backscattering. *Optical Engineering* 21: 257.

Nosatch OYu, Popovichev VI, Ragulskii VV and Faizullov FS (1972) Compensation of phase distortions in an amplifying medium by a 'Brillouin mirror'. *JETP Letters* 16: 435.

Pepper DM (1986) Applications of optical phase conjugation. *Scientific American* 254: 74–83.

Shkunov VV and Zeldovich BYa (1985) Optical phase conjugation. *Scientific American* 253: 54–59.

Yariv A (1978) Phase conjugate optics and real-time holography. *IEEE Journal of Quantum Electronics* 14: 650.

Zeldovich BYa, Popovichev VI, Ragulskii VV and Faizullov FS (1972) On relation between wavefronts of reflected and exciting radiation in stimulated Brillouin scattering. *JETP Letters* 16: 109.

Zeldovich BYa, Shkunov VV and Pilipetsky NF (1985) *Principles of Phase Conjugation*. Berlin: Springer-Verlag.

Zeldovich BYa, Mamaev AV and Shkunov VV (1995) *Speckle-Wave Interactions in Application to Holography and Nonlinear Optics*. Boca Raton, FL: CRC Press.

Photorefraction

M Cronin-Golomb, Tufts University, Medford, MA, USA

B Kippelen, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction and Basics of the Photorefractive Effect

In 1966 Ashkin and coworkers were pursuing research on optical devices using the electro-optic material lithium niobate. They noticed that the refractive index of lithium niobate would change when it was exposed to laser light, and upset the expected operation of their devices. They called the effect optical damage. Shortly thereafter, Chen, LaMacchia, and Fraser reported on the use of the optical damage effect for holographic data storage. Thus began the field of photorefractive nonlinear optics, which has been used in various applications such as real-time holography, optical data storage, optical image amplification, nondestructive testing, distortion compensation by phase conjugation, pattern recognition, and radar signal processing. Many inorganic and organic materials have been investigated for their photorefractive effects, including ferroelectrics, semiconductors, and sensitized polymers. The most well-known inorganic materials are lithium niobate, bismuth silicon oxide, barium titanate, and strontium barium niobate. Most organic photorefractive materials are based on polymeric photoconductors such as those used in xerography that are doped with electro-active molecules, some plasticizers, and sensitizers.

While in its broadest interpretation, the photorefractive effect occurs whenever light incident on a material causes a refractive index change, one usually applies the term to electro-optic index changes resulting from optically generated space charge fields. Materials that are photorefractive in this sense share the following properties

- high transmission at the operating wavelengths;
- linear electro-optic coefficients or orientational Kerr effects;
- charge carriers that become mobile when optically excited;
- trapping centers for these charge carriers to enable spatially non-uniform redistribution of charge.

Consider two beams from the same laser crossing inside a photorefractive material such as barium titanate. The interference pattern will have bright and dark fringes. Charge carriers are excited where the light is bright, then drift and diffuse to regions of relative darkness where they preferentially recombine into trapping centers. In this way, a net excess charge develops in the dark regions, and a net deficit of charge develops in the bright regions. The spatially varying charge distribution has an electric field associated with it and this electric field causes a spatially varying refractive index profile. Because the space charge, its field, and resulting refractive index have the same spatial periodicity as the original interference pattern we have a holographic phase grating. The diffraction efficiency of the hologram can easily approach 100% for materials such as barium titanate and strontium barium niobate which have high electro-optic coefficients. Likewise, such high diffraction efficiencies are easily obtained in 100 micrometer-thick photorefractive polymer films.

The Standard Rate Equation Model

The development of photorefractive gratings can be modeled using standard semiconductor rate equations. Figure 1 shows two beams incident

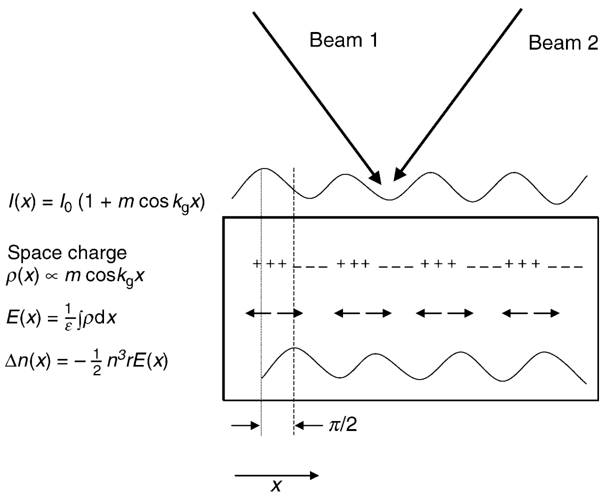
Beam 1
Beam 2
$I(x) = I_0\,(1 + m\cos k_g x)$
Space charge
$\rho(x) \propto m\cos k_g x$
$E(x) = \frac{1}{\varepsilon}\int \rho\,dx$
$\Delta n(x) = -\frac{1}{2}n^3 r E(x)$
+++ --- +++ --- +++ --- +++ ---
$\pi/2$
x

Figure 1 Diagram of two-beam coupling process showing optical interference pattern, space charge, resultant electric field, and $\pi/2$ phase shifted electro-optically induced refractive index grating.

symmetrically on a photorefractive crystal or polymer. They form an interference pattern whose intensity may be written:

$$I(x) = I_0(1 + m\cos(k_g x)) \qquad [1]$$

where x is the direction perpendicular to the interference fringes, I_0 is the average intensity, m is the modulation index, and k_g is the wavenumber of the interference pattern.

The crystal may be considered a wide bandgap semiconductor containing electron donors in the bandgap with density N_D and electron acceptors with density N_A. It is assumed that some electrons ionized from the donors permanently occupy these acceptors so that when charge in the crystal is distributed uniformly in the dark, the number density of ionized donors N_D^i is equal to N_A. Likewise, polymers contain donor and acceptor-like molecules that can be neutral or ionized. The spatially varying light distribution ionizes the donors at the following rate, assuming that $N_D^i \ll N_D$:

$$\frac{\partial N_D^i}{\partial t} = sIN_D - \gamma_R n_e N_D^i \qquad [2]$$

where s is a photoionization coefficient, γ_R is a recombination parameter, and n_e is the density of excited charge carriers, which we assume here to be electrons. The model can easily be generalized to include holes. We also use the equation of charge conservation:

$$\frac{\partial N_D^i}{\partial t} = \frac{\partial n_e}{\partial t} - \frac{1}{e}\nabla\cdot\mathbf{j} \qquad [3]$$

where e is the charge of an electron, and $\mathbf{j}$ is the current in the conduction band.

$$\mathbf{j} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e \qquad [4]$$

where μ is the electron mobility, k_B is Boltzmann's constant, and T is the temperature. The electric field obeys Gauss's law:

$$\nabla\cdot\mathbf{E} = -e(n_e + N_A - N_D^i)/\varepsilon \qquad [5]$$

where N_A is the density of acceptors. These equations may be linearized by approximating the electron density, ionized donor density, and electric field with their first Fourier components:

$$\begin{aligned} E &= E_0 + \tfrac{1}{2}(E_1\exp(ik_g x) + E_1^*\exp(-ik_g x)) \\ N_D^i &= N_{D0}^i + \tfrac{1}{2}(N_{D1}^i\exp(ik_g x) + N_{D1}^{i*}\exp(-ik_g x)) \\ n_e &= n_{e0} + \tfrac{1}{2}(n_{e1}\exp(ik_g x) + n_{e1}^*\exp(-ik_g x)) \end{aligned} \qquad [6]$$

This assumption is strictly valid only when the modulation index m is much less than unity. Otherwise, a generalization to higher orders in the Fourier series is required. However, the linearized theory is sufficient to illustrate the most important features of the photorefractive effect. The solution for the space charge field E_1 for the case when the interference pattern is applied at time $t = 0$ is

$$E_1 = m\frac{-iE_Q(E_0 + iE_D)}{E_0 + i(E_D + E_Q)}(1 - \exp(-t/\tau)) \qquad [7]$$

where E_0 is an externally applied or photovoltaic field (if any), E_D is the diffusion field

$$E_D = \frac{k_B T k_g}{e} \qquad [8]$$

and E_Q is the limiting space charge field

$$E_Q = \frac{eN_A}{\varepsilon k_g} \qquad [9]$$

Some photorefractive crystals, most notably $LiNbO_3$, exhibit the photovoltaic effect in which optical illumination induces a dc field across the crystal.

The sinusoidally varying space charge field E_1 operates through the linear electro-optic effect with effective coefficient r to produce a sinusoidal variation in the refractive index n of the crystal:

$$n(x) = n_0 + \tfrac{1}{2}(n_1\exp(ik_g x) + n_1^*\exp(-ik_g x)) \qquad [10]$$

where

$$n_1 = -\frac{1}{2}rn_0^3 E_1 \qquad [11]$$

The effective electro-optic coefficient r may be found from tensor analysis of the electro-optic tensor and the vector space charge and optical fields. Notice that there is a 90-degree phase shift between the interference pattern and the refractive index grating when E_0 is zero. The time constant τ is given by

$$\tau = \frac{N_A}{sN_D I_0} \frac{E_0 + i(E_D + E_\mu)}{E_0 + i(E_D + E_Q)} \qquad [12]$$

where E_μ is a mobility field

$$E_\mu = \frac{\gamma_R N_A}{\mu k_g} \qquad [13]$$

When E_0 is nonzero, there is an oscillatory component to the time constant.

In contrast to the case of ordinary optical nonlinearities such as the optical Kerr effect, in which the nonlinear coupling strength is proportional to the optical intensity, the steady state strength of the photorefractive effect is independent of optical intensity while the time constant is inversely proportional to intensity in the basic model described above. The time constant varies with the photoconducting performance of a given material. In the fastest polymeric and inorganic materials it is, at the time of writing, of the order of a few milliseconds at 1 W cm^{-2} of incident optical intensity.

At low intensities (below the equivalent dark intensity, 1 W cm^{-2} in barium titanate), the above equations need to be modified to include the effect of dark conductivity. Even in the dark, there will be a few mobile charge carriers in the conduction band that tend to erase the grating. This will result in the effect appearing more Kerr-like, except still with the 90-degree phase shift between the index grating and the interference pattern.

Coupled Wave Equations

The change in refractive index n_1 can be large enough to produce substantial interactions between the writing beams. The writing beams generate a phase grating that diffracts the beams into each other. The grating influences the writing beams, which in turn influence the grating. In the cases where the diffusion field dominates, for example when the externally applied or photovoltaic field E_0 is absent, one beam is amplified by in-phase diffraction of the other beam from the grating. As shown in Figure 2, this amplification results from a 90-degree phase shift due to diffraction from a phase grating coupled with a −90-degree phase shift from the spatial phase shift between the interference pattern and the index grating. The second beam is attenuated by destructive interference with the first beam diffracted by the grating. These interactions can be modeled by standard coupled wave theory. Let the electric field amplitude associated with the jth beam be

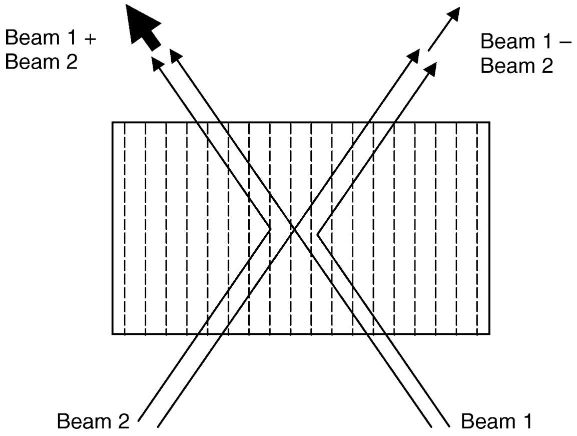

Figure 2 Two-beam coupling amplification. Beam 1 is amplified by constructive interference. Beam 2 is de-amplified by destructive interference.

$$\mathbf{E}_j(\mathbf{r}, t) = \mathbf{e}\Big[A_j(\mathbf{r}) \exp(i(\mathbf{k}_j \cdot \mathbf{r} - \omega t)) + A_j^* \exp(-i(\mathbf{k}_j \cdot \mathbf{r} - \omega t))\Big] \qquad [14]$$

where $\mathbf{e}$ is the polarization unit vector. Using the scalar wave equation

$$\nabla^2 \mathbf{E} + k^2 \mathbf{E} = 0 \qquad [15]$$

and the slowly varying envelope approximation

$$\left|\frac{d^2 A_j}{dz^2}\right| \ll k \left|\frac{dA_j}{dz}\right| \qquad [16]$$

we find

$$\begin{aligned} \frac{dA_1}{dz} &= +\gamma \frac{A_1 |A_2|^2}{I_0} \\ \frac{dA_2^*}{dz} &= -\gamma \frac{A_2^* |A_1|^2}{I_0} \end{aligned} \qquad [17]$$

with the coupling constant γ given by

$$\gamma = -\frac{i\omega \Delta n_m}{c \cos \vartheta} \qquad [18]$$

where ϑ is the half-angle between the writing beams and Δn_m is given by

$$\Delta n_m = -\frac{1}{2} n_0^3 r \frac{E_1}{m} \qquad [19]$$

Equation [17] shows that beam 1 is amplified and beam 2 is attenuated. That beam 1 is amplified instead of beam 2 is a result of the choice of crystal orientation and hence the sign of γ. These nonlinear equations can be solved for normalized intensity $I_j = |A_j|^2$:

$$I_1(z) = \frac{I_0}{1 + (I_2(0)/I_1(0))\exp(+\Gamma z)}$$
$$I_2(z) = \frac{I_0}{1 + (I_1(0)/I_2(0))\exp(-\Gamma z)} \qquad [20]$$

where $\Gamma = \gamma + \gamma^*$ is the intensity coupling constant. In the limit where $I_1 \ll I_2$, the weak beam, beam 1, experiences exponential amplification. This amplification can be used to build optical oscillators with many of the same properties as conventional laser oscillators. The solution can be generalized to the case where γ has an imaginary component due to an applied or photovoltaic field E_0. Linear absorption can also be included.

Materials

Photorefractive materials may be separated into two broad classes: inorganic and organic. The first materials investigated were inorganic oxides and semiconductors. Their success led to efforts to endow more easily produced organic materials with the required photoconductivity, charge trapping centers, and electro-optic coefficients.

Inorganic Materials

The first requirement in a classic photorefractive material is a linear electro-optic coefficient, such as appears in sillenites such as bismuth silicon oxide $Bi_{12}SiO_{20}$ and bismuth germanium oxide $Bi_{12}GeO_{20}$. These were some of the first photorefractive materials used for image processing and phase conjugation applications. However, their electro-optic coefficients (a few pm/V) are not large enough to easily give rise to large diffraction efficiencies, or to photorefractive oscillators. Materials at a temperature near a phase transition generally have higher electro-optic coefficients because their crystal structures are on the verge of changing. They are extremely susceptible to the effect of any external influence such as the application of electric fields including photorefractive space charge fields. That is why ferroelectric materials are good candidates for photorefractive materials. These include barium titanate $BaTiO_3$, potassium niobate $KNbO_3$, and strontium barium niobate $Sr_xBa_{1-x}Nb_2O_6$. The mean free path of charge transport is less than that in sillenites, so they require more photons to reach a steady state. Therefore, ferroelectric materials are typically less sensitive than sillenites.

The second requirement is for photoconductivity. This requires the existence of photoexcitable charge carriers, either from valence band to conduction band or from intraband trapping centers. The latter source of photocarriers is used most often because the absorption length of light whose energy is greater than the bandgap is usually only a few micrometers. This places a severe restriction on the beams' interaction length ℓ. Thus there has been considerable research on suitable dopants, most commonly $Fe^{2+}$and Fe^{3+}. These dopants also act as trapping centers.

Lithium niobate is an example of a material with a large photovoltaic effect. When illuminated, this crystal develops a large dc field, which acts to enhance the grating strength. In those cases where the photorefractive effect is not wanted, such as in the design of lithium niobate waveguide devices, the photovoltaic effect can be greatly diminished by the addition of MgO to the crystal melt during growth.

For photorefractivity in the near infrared, semiconductors such as gallium arsenide and indium phosphide have been used with success. These materials also have the advantage that they can be grown in layered structures to produce, for example, multiple quantum wells that can be used to tailor the characteristics of optical excitation and charge transport.

Organic Materials

First-generation photorefractive polymers were designed to mimic the properties of their inorganic counterparts. Owing to their rich structural flexibility, organic synthetic materials with suitable charge transport, trapping, linear electro-optic effects and low optical absorption, were developed. This approach did build on the know-how developed previously in making photoconducting polymers for xerography and electro-optic polymers for optical modulation. Current polymers are based on an orientational photorefractive effect that leads to higher refractive index changes compared with traditional photorefractive materials. In materials with orientational photorefractivity, the refractive index change is produced by the field induced reorientation of anisotropic conjugated molecules that possess a permanent dipole moment and that have a high polarizability anisotropy. The photorefractive space charge field together with the applied field will periodically reorient these molecules and produce a periodic refractive index modulation through an orientational Kerr effect. The build-up and dynamics of this space-charge field are similar to those of traditional photorefractive crystals and can be described by eqns [2]–[7]. The time constant of the hologram is a convolution of the build-up time of the space-charge and the orientational diffusion time of the dipolar molecules in the total electric field. In contrast to crystals, polymers are nearly amorphous and the transport properties are described by charge hopping processes instead of band-type transport in crystals. Consequently, the photoconducting properties of polymers are strongly field dependent and photorefractivity is mainly observed under strong applied voltages. Numerous polymer composites have been developed using the hole transport polymer poly(N-vinylcarbazole). Several materials with a refractive index modulation amplitude of the order of a percent and two-beam coupling constants $\gamma > 100\ \mathrm{cm}^{-1}$ have been reported. New polymer composites are continuously being tested. Photorefractivity is also studied in other organic materials including organic crystals, liquid crystals, nanocomposites such as polymer-dispersed liquid crystals, or hybrid materials such as sol-gels.

Applications

Holographic Data Storage

Holographic data storage takes advantage of the Bragg selectivity of thick gratings. This allows many holograms to be superimposed in the same small volume, typically of the order of one cubic centimeter. A page of data is displayed on a spatial light modulator and a laser beam passing through the modulator is holographically recorded in the crystal with a reference beam at a specific angle. Many pages of data can be superimposed by recording many holograms with angularly multiplexed reference beams. Other multiplexing schemes are implemented by changing the shape of the wavefront of the reference beam. In principle, the upper limit of recording density is determined by the wavelength of light: one bit per cubic wavelength. If the recording wavelength is 0.5 μm, then one cubic centimeter can contain 1000 gigabytes of data. In practical circumstances, when noise is taken into consideration, the capacity is more realistically of the order of one gigabyte if a 1000 × 1000 spatial light modulator is used.

Distortion Compensation by Phase Conjugation

Photorefractive materials can be used to make high-reflectivity phase conjugate mirrors. The phase conjugate of a laser beam is produced when a hologram of the beam is read by another beam traveling in the opposite direction to the original reference beam. The phase conjugate reconstruction is a time-reversed copy of the original beam. If the original beam has passed through distorting optics, then the phase conjugate beam will retrace the path of the original beam through the distortion and emerge in its undistorted original state. In the standard nomenclature of phase conjugation, the input beam is called the signal, or probe, and the two reference beams are called the pumps. The output beam at $z = 0$ is called the phase conjugate beam and has zero amplitude at its input at $z = \ell$, where ℓ is the interaction length. Applications for phase conjugation exist, for example, in signal transmission through distortions and laser cavity design. If a phase conjugate mirror is used as a cavity mirror, then the effects of intracavity distortions are removed.

Since the photorefractive gratings can be very strong, the diffraction efficiency of the counterpropagating reference beam can be so high that the phase conjugate reflectivity can exceed unity. The simplest generalization of eqn [17] to the four-wave mixing phase conjugation case is when only transmission gratings are important, as occurs in many circumstances depending on the mutual coherence properties of the beams, and the material's properties. The coupled wave equations are

$$\begin{aligned}
\frac{\mathrm{d}A_1}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_4}{I_0} \\
\frac{\mathrm{d}A_2^*}{\mathrm{d}z} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_3^*}{I_0} \\
\frac{\mathrm{d}A_3}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_2}{I_0} \\
\frac{\mathrm{d}A_4^*}{\mathrm{d}z} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_1^*}{I_0}
\end{aligned} \qquad [21]$$

In the undepleted pumps approximation, $(I_1, I_2 \gg I_3, I_4)$, the equations become linear and the solution for phase conjugate reflectivity $R = I_3(0)/I_4(0)$ is

$$R = \left|\frac{\sinh[\gamma\ell/2]}{\cosh[(\gamma\ell + \ln r)/2]}\right|^2 \qquad [22]$$

where $r = I_2/I_1$ is the ratio between the intensities of the pumping beams.

The fact that the reflectivity of the phase conjugate mirror can be greater than unity means that we can build an optical oscillator bounded by a regular mirror and a phase conjugate mirror only. Not only does it not require any additional optical gain, but it also compensates for intracavity distortions. The regular mirror can have any shape, provided that it is sufficiently reflective.

Self-pumped Phase Conjugate Mirrors

If a laser beam passes through a crystal placed inside an optical cavity formed by two facing mirrors, light scattered by imperfections in the crystal can be amplified through the photorefractive effect. The cavity provides feedback and optical oscillation can build up. The oscillation beams pump the crystal as a phase conjugate mirror for the incident beam, thus forming a self-pumped phase conjugate mirror. The feedback can even be provided by total internal reflection in the crystal in which case the crystal by itself can become a phase conjugate mirror.

Pattern Recognition and Image Filtering

Photorefractive wave mixing can be used to perform pattern recognition by matched filtering. One example would be to identify a tank in a battlefield scene, another would be to identify all of the occurrences of a particular word on a page of text. Suppose the input beams contain the corresponding pictorial information, such as might be obtained by passing the beams through an image-bearing transparency or spatial light modulator. Equation [21] shows that the source for beam 3 contains a term proportional to the product of the amplitudes of the three input beams. If lenses are placed in the input beams so that the crystal receives the Fourier transforms of those beams, then the output beam at the crystal, beam 3, will be proportional to the product of the Fourier transforms of the input beams 1, 2, and 4. A lens may then be used to perform the inverse Fourier transform of the product of the Fourier transforms of the input beams, producing an output proportional to beam 1 convolved with beam 2 correlated with beam 4. If beam 1 is a point source before its Fourier transforming lens, then it will be a plane wave at the crystal. Beam 3 after inverse Fourier transformation by its lens will be the correlation of beams 4 and 2. For example, suppose we want to find all the occurrences of a particular word, say ‘optics’ in a given page of text. Then we would make a transparency of the word ‘optics’ and place that in input beam 4. We would then place an image of the page of text in beam 2. Beam 3 would then contain a field with bright spots at the places containing the word ‘optics’ in the original text.

The real-time holographic recording properties of photorefractive materials can also be exploited in medical imaging applications by performing time-gated holography. In this method, a hologram is formed by the temporal overlap in the photorefractive sample of a reference pulse and the first-arriving (ballistic photons) light from a stretched image-bearing pulse that has propagated through a scattering medium. The filtering of the useful photons from the scattered ones is achieved by reconstructing the hologram formed with the ballistic photons in a four-wave mixing geometry.

Optical Limiting, the Novelty Filter, and Laser Ultrasonic Inspection

The attenuation of beam 2 in eqn [17] can be used in several applications including optical limiting and novelty filtering. If one wants to protect a sensor from high-intensity laser radiation, then one could split a small portion of the input beam directed at the sensor and use it as beam 1 in a photorefractive recording setup with the input beam acting as beam 2. If the laser intensity is above the equivalent dark intensity such that the optically excited charge density is greater than the thermally excited charge density, the photorefractive effect will be activated and the input beam will be de-amplified by destructive interference with beam 1, thus protecting the sensor. In materials with high gain–length products ($\gamma\ell > 1$), separate provision of beam 1 is unnecessary because light scattered from the input beam by imperfections in the material and other scattering centers will be greatly amplified, often to such an extent that the input beam is almost completely attenuated. This photorefractive amplification of scattered radiation is known as the fanning effect, because the amplified scattered light often appears as a fan, or brush of light, as can be seen in Figure 3. The effect can also be used to make a novelty filter, which transmits only the moving portion of a scene. The crystal is only fast enough to

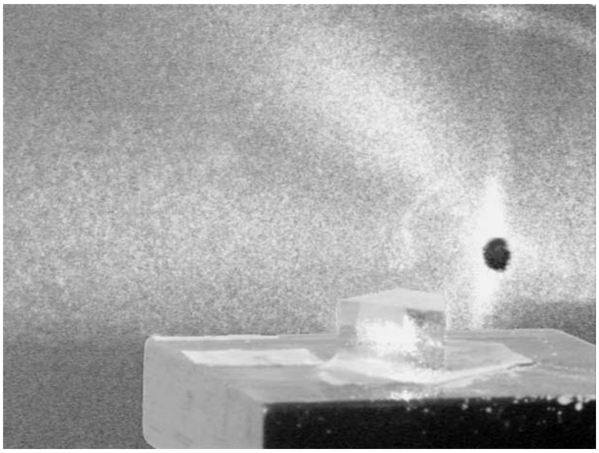

Figure 3 Photorefractive barium titanate exhibiting amplified scattering. A helium neon laser beam is incident from the lower left, passes through the crystal to a screen where it is blacked out to prevent saturation of the camera. The screen shows brushes of amplified scattering, or fanning.

respond to the slowly changing or static components of an image-bearing beam. Since the grating formed is Bragg-matched only to the slow component, the grating will only attenuate that slow component. Any rapidly changing parts of a scene pass through the crystal unattenuated. Such a filter is useful for picking out moving objects in a static cluttered background, for example a tank on a battlefield, or a micro-organism swimming against a stationary background.

A related application is in laser ultrasonic inspection. Defects in industrial material processing such as welding have characteristic ultrasonic signatures. The part under test is pinged by a pulsed laser and a probe laser is reflected from the part. As the part is shaken by the ultrasound, speckle in the reflected beam vibrates. A photorefractive recording is made of the speckle beam. Electrodes are placed on the photorefractive material, so that any optically induced currents can be detected. If the speckle pattern is not moving, or is moving more slowly than the speed of response of the photorefractive material, the photorefractive grating will be essentially at steady state: the drift currents balance the diffusion currents so there is no net current. There is no signal as the part moves through the process line. However, if the speckle beam is moving faster than the response time of the material, it will move photoexcited charge back and forth past the quasistatic grating and generate a net current for detection via the electrodes.

Adaptive Signal Processing

The relatively slow speed of photorefractive devices can be used to advantage in radio-frequency (RF) signal processing, such as signal extraction and coherent combination of signals from antenna arrays. The signal extraction application depends on grating competition in two-beam coupling. Suppose we wish to separate signals on two different RF carrier frequencies ω_1 and ω_2, respectively. The combined signal is applied to an optical carrier beam using a high-speed modulator. The resulting optical field may be represented as

$$S_1(t)\exp(i\omega_1 t) + S_2(t)\exp(i\omega_2 t) \qquad [23]$$

It is used to pump a unidirectional ring resonator so that grating competition allows oscillation only on the strongest component of the combined signal, say S_1. The output of the oscillator is proportional to the extracted component S_1. The effectiveness of the intersignal competition is enhanced by placing another photorefractive material in the cavity. A portion of the intracavity beam is picked off by a beamsplitter and used as a two-beam coupling pump in the second material. The crystal is oriented so that the photorefractive grating diffracts the picked-off beam back into the cavity. The return of the picked-off component is most effective for the stronger component S_1 thus decreasing its loss compared to that of the weaker component S_2. This beamsplitter/crystal combination is known as a reflexive coupler.

Photorefractive Solitons

Under favorable conditions, a single beam incident on a photorefractive crystal will induce a positive refractive index change at the center of the beam. This provides a self-focussing tendency that counteracts the beam's divergence due to diffraction. When these two effects balance each other, the beam can propagate with a constant diameter. Such a beam is known as a spatial soliton in analogy with temporal solitons in optical fibers and can occur when there is a component of the photorefractive response due to drift. The drift component of the photorefractive effect appears when a dc field E_0 is applied to the material. Potential applications are optically written waveguides and couplers.

List of Units and Nomenclature

Amplitude gain	[cm^{-1}]	γ
Applied dc field	[Vm^{-1}]	E_0
Current	[Am^{-2}]	$\mathbf{j}$
Dielectric permittivity	[Farads m^{-1}]	ε
Diffusion field	[Vm^{-1}]	E_D
Electron acceptor density	[m^{-3}]	N_A
Electron charge	[C]	e
Electron donor density	[m^{-3}]	N_D
Grating wavenumber	[m^{-1}]	k_g
Intensity gain	[cm^{-1}]	Γ
Interaction length	[cm]	ℓ
Ionized donor density	[m^{-3}]	N_D^i
Limiting space charge field	[Vm^{-1}]	E_Q
Mobility	[m^2 V^{-1} sec^{-1}]	μ
Mobility field	[Vm^{-1}]	E_μ
Normalized intensity	[V^2 m^{-2}]	I
Photo-ionization coefficient	[kg sec^{-2}]	s
Photorefractive time constant	[sec]	τ
Pump ratio	[1]	r
Radian frequency	[sec^{-1}]	ω
Recombination coefficient	[m^3 sec^{-1}]	γ_R
Slowly varying optical electric field	[Vm^{-1}]	A

See also

Holography, Applications: Holographic Recording Materials and Their Processing. **Holography, Techniques:** Holographic Interferometry. **Imaging:** Volume Holographic Imaging. **Information Processing:** Free-Space Optical Computing; Optical Neural Networks. **Materials for Nonlinear Optics:** Organic Nonlinear Materials. **Modulators:** Electro-Optics. **Nonlinear Optics, Basics:** Four-Wave Mixing; Nonlinear Optical Phase Conjugation. **Phase Control:** Phase Conjugation and Image Correction. **Scattering:** Stimulated Scattering. **Solitons:** Bright Spatial Solitons.

Further Reading

Coufal HJ, Psaltis D and Sincerbox GT (eds) (2000) *Holographic Data Storage*. Berlin: Springer.

Gunter P and Huignard J-P (eds) (1988) *Photorefractive Materials and their Applications I*. Berlin: Springer-Verlag.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Berlin: Springer-Verlag.

Nalwa HS and Miyata S (eds) (1997) *Nonlinear Optics of Organic Molecules and Polymers*. Boca Raton: CRC Press.

Nolte DD (ed.) (1995) *Photorefractive Effects and Materials*. Kluwer.

Pepper DM, Feinberg J and Kukhtarev NV (1990) The photorefractive effect. *Scientific American* 263: 62.

Solymar L, Webb DJ and Grunnet-Jepsen A (1996) The Physics and Application of Photorefractive Materials. Oxford: Clarendon Press.

Yeh P (1993) Introduction To Photorefractive Nonlinear Optics. New York: Wiley.

Yu F and Yin S (eds) (2000) *Photorefractive Optics: Materials, Properties, and Applications*. San Diego: Academic Press.

Ultrafast and Intense-Field Nonlinear Optics

A L Gaeta, Cornell University, Ithaca, NY, USA

R W Boyd, University of Rochester, Rochester, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The tremendous development of high-powered femtosecond laser systems since the 1980s, has opened up new areas of research in nonlinear optics, plasma physics, atomic and molecular dynamics, and intense-field physics. For many of these applications, it is important to understand how ultrashort light pulses propagate through a medium under conditions in one or more of the processes in which nonlinear optics play an important role.

The starting point for the modeling of light propagation under these conditions is Maxwell's wave equation for the electric field $E(\mathbf{r},t)$ which is given in Gaussian units as

$$\nabla^2 E - \frac{\partial^2 E}{\partial t^2} = 4\pi \frac{\partial^2 P}{\partial t^2} \qquad [1]$$

where the $P(\mathbf{r},t)$ is the polarization inside the medium. Typically, the polarization is separated into a term P_l that depends linearly on the field E and into a term P_{nl} that depends nonlinearly on the applied field. The Fourier transform of the linear polarization can be expressed as $\tilde{P}_l(\mathbf{r},\omega) = \chi^{(1)}(\omega)\tilde{E}(\mathbf{r},\omega)$, where $\tilde{E}(\mathbf{r},\omega)$ is the Fourier transform of electric field $E(\mathbf{r},t)$. For the laser–matter interactions that we consider here, we assume that the linear susceptibility $\chi^{(1)}(\omega)$ is real, in which case the wave equation can be expressed as

$$\nabla^2 \tilde{E} + \left[\frac{n_l(\omega)\omega}{c}\right]^2 \tilde{E} = -\frac{4\pi\omega^2}{c^2}\tilde{P}_{\mathrm{nl}} \qquad [2]$$

where $n_l(\omega) = \sqrt{1 + 4\pi\chi^{(1)}(\omega)}$ is the frequency-dependent linear refractive index of the medium.

For light pulses that are longer than an optical period, the envelope description is valid and the electric field can be described by an amplitude envelope $A(\mathbf{r},t)$ and a carrier frequency ω_0 such that

$$E(\mathbf{r},t) = A(\mathbf{r},t)e^{i(k_0 z - \omega_0 t)} + c.c \qquad [3]$$

where $k_0 = k(\omega_0) = n_0\omega_0/c$ is the wavevector amplitude and $n_0 = n_l(\omega_0)$. This envelope description is advantageous for performing analytical and numerical studies of pulse propagation. However, for sufficiently short laser pulses, where the shape of the envelope function does not depend on the carrier phase of the carrier wave, such a description is no longer applicable. Nevertheless the envelope description can be made valid for pulses that are nearly as short as a single cycle or, alternatively, that have spectral bandwidths that are comparable to the central frequency ω_0. To derive an equation for the spatio-temporal evolution of the pulse envelope, the relation for the electric field is substituted into the wave equation (eqn [2]) and the following two approximations are made: i) $k_0 \partial A/\partial z \ll \partial^2 A/\partial z^2$, which signifies that the envelope varies slowly in space over a distance compared to the central wavelength; ii) $v_{\mathrm{ph}} \sim v_{\mathrm{gr}}$ where $v_{\mathrm{ph}} = c/n_0$ is the phase velocity and $v_{\mathrm{gr}} = (\mathrm{d}k/\mathrm{d}\omega)^{-1}$ is the group velocity. This latter approximation is invariably well satisfied when the frequencies contained in the laser field are highly nonresonant with any transition frequencies of the medium.

For an input pulse at $z = 0$ with a peak amplitude A_0, and characteristic widths in space and time given by w_0 and τ_p, respectively, the equation for the normalized amplitude $u(r_\perp, z, t) = A(r_\perp, z, t)/A_0$ can be expressed as

$$\frac{\partial u}{\partial \xi} = \frac{i}{4}\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right)^{-1} \nabla_\perp^2 u - i \sum_{n=2} \frac{L_{\mathrm{ds}}}{n! L_{\mathrm{ds}}^{(n)}} \frac{\partial^n u}{\partial \tau^n} + i\left(1 + \frac{i}{\omega_0 \tau_p}\frac{\partial}{\partial \tau}\right) p^{\mathrm{nl}} \qquad [4]$$

where $\tau = (t - z/v_g)/\tau_p$ is the normalized retarded time for the pulse traveling at the group velocity v_g, $L_{\mathrm{ds}}^{(n)} = \tau_p^n/\beta_n$ is the nth-order dispersion length, $\beta_n (n \geq 2)$ is the nth-order dispersion constant, $L_{\mathrm{ds}} = |L_{\mathrm{ds}}^{(2)}|$ is the dispersion length, $\zeta = z/L_{\mathrm{df}}$ is the normalized distance, $L_{\mathrm{df}} = k w_0^2/2$ is the diffraction length, p_{nl} is the normalized nonlinear polarization, and $\nabla_\perp^2$ is the transverse Laplacian. The presence of the operator $(1 + i\partial/\omega_0 \partial \tau)$ in the diffraction term $(\nabla_\perp^2 u)$ of eqn [4] leads to space–time focusing, while its presence in the nonlinear term results in self-steepening and both these terms arise from not making the slowly varying envelope approximation in time (i.e., $k_0 \partial A/\partial t \ll \partial^2 A/\partial t^2$) in deriving eqn [4]. For a self-consistent analysis of pulse propagation in the nonlinear regime, both the effects of space–time focusing and self-steepening must be included.

Nonlinear Refractive Index

For an isotropic medium in the highly nonresonant limit, the third-order term is the lowest-order contribution to the nonlinear susceptibility. This term gives rise to the nonlinear refractive index, that is, the index of refraction depends on the local intensity of the laser field. For many materials there are two contributions to the nonlinear refractive index: i) an instantaneous part that arises from the electronic response of the medium; and ii) a noninstantaneous contribution due to the nuclear motion of the molecules (i.e, the Raman contribution). For such a medium, the nonlinear polarization may be expressed as

$$p_{\mathrm{nl}}(\zeta, \tau) = \frac{L_{\mathrm{ds}}}{L_{\mathrm{nl}}}\left[(1 - f)|u(\zeta, \tau)|^2 + f\gamma_R \int_{-\infty}^{\tau} \mathrm{d}\tau' R(\tau - \tau')|u(\zeta, \tau')|^2\right] u(\zeta, \tau) \qquad [5]$$

where $L_{\mathrm{nl}} = (c/\omega_0 n_2 I_0)$ is the nonlinear length, $I_0 = n_0 c|A_0|^2/2\pi$ is the peak input intensity, f is the fraction of the Raman contribution to the nonlinear refractive index, and $R(\tau) = \{[1 + (\Omega_R \tau_R)^2]/\Omega_R \tau_R\}$ $\exp(-\tau/\tau_R)\sin(\Omega_R \tau)$ is the Raman response function, τ_R is the characteristic Raman response time, Ω_R is the characterstic Raman frequency, and $\gamma_R = \tau_p/\tau_R$. For example, for fused silica $f = 0.15$, $\tau_R = 50$ fs, and $\Omega_R \tau_R = 4.2$. For a noble gas such as Xe, there is no Raman contribution and $f = 0$.

Self Focusing, Supercontinuum Generation, and Filamentation

The presence of the nonlinear refractive index with $n_2 > 0$ can lead to self-focusing of a laser beam. For sufficiently long pulses such that dispersion, self-steepening, and space–time focusing effects can be neglected, it is found that laser beams with input powers greater than the critical power $P_{\mathrm{cr}} = \alpha\lambda^2/4\pi n_0 n_2$ will undergo catastrophic self-focusing collapse. The dimensionless parameter $\alpha \geq 1.86$ depends on the spatial profile of the input beam and for a Gaussian input beam is given by $\alpha = 1.9$, in which case the ratio of the input power P to the critical power satisfies the relation $P/P_{\mathrm{cr}} = 1.055 L_{\mathrm{df}}/L_{\mathrm{nl}}$. Extensive studies have been made on the dynamics of laser beams undergoing self-focusing. For example, it has been shown that the shape of the collapsing beam evolves to a radially symmetric profile as it approaches the collapse point and that the total power contained in the collapsing portion always corresponds to the minimum value (i.e., $\alpha \sim 1.86$) regardless of the initial power in the beam.

For light pulses shorter than a picosecond, the role of material dispersion plays an important role and completely alters the dynamics of the self-focusing process. These dispersive effects lead to a temporal splitting of the pulse into two pulses and the arrest of its collapse. At even higher powers, other phenomena can occur, such as ‘optical shock’ formation at the rear edge of the pulse, due to self-steepening and space–time focusing. Shock formation leads to the emission of a broad spectrum of radiation extending from the ultraviolet to the mid-infrared, known as supercontinuum generation (SCG). This phenomenon was first observed in 1970, and since then it has been observed in many different solids, liquids, and gases, under a wide variety of experimental conditions.

If the peak intensity of the pulse approaches intensities of $> 10^{13}$ W/cm^2, either through self-focusing or external focusing, multiphoton ionization occurs and a plasma is formed in the medium. (See section below on Plasma Nonlinearities and Relativistic Effects.) The generation of the plasma lowers the refractive index and effectively counteracts the self-focusing process, resulting in the spatial confinement of the pulse for distances far beyond what would be allowed by ordinary linear diffraction and has been observed in gases, liquids, and solids. A striking example of this apparent self-waveguiding is the observation of 'light strings' in air which can extend more than 10 km into the atmosphere. This phenomenon was first observed with 100 fs laser pulses in the near infrared ($\lambda = 800$ nm). Researchers found that pulses with energies greater than 10 mJ undergo self-focusing collapse in air and produce a highly intense ($> 10^{13}$ W/cm^2) 100-micron-diameter light filament tens of meters long.

Multiphoton Absorption

Multiphoton absorption is a process in which an atom or molecule makes a transition from a ground state to an excited state by means of the simultaneous absorption of N photons. In the lowest order of perturbation theory, such a process can be described by means of a susceptibility of order $(2N-1)$, that is, by $\chi^{(2N-1)}$. Alternatively, this process can be described in terms of an N-photon cross-section $\sigma^{(N)}$ defined such that the transition rate per atom is given by

$$R^{(N)} = \sigma^{(N)} I^N \qquad [6]$$

where I is the intensity of the laser field. Quantum mechanical expressions for the N-photon cross-sections are readily obtained. One finds, for instance, that

$$R^{(2)}_{\mathrm{ng}} = \left| \sum_m \frac{\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^2}{\hbar^2(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{ng}} - 2\omega)$$

$$R^{(3)}_{\mathrm{og}} = \left| \sum_{mn} \frac{\mu_{\mathrm{on}}\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^3}{\hbar^3(\omega_{\mathrm{ng}} - 2\omega)(\omega_{\mathrm{mg}} - \omega)} \right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{og}} - 3\omega) \qquad [7]$$

In each of these expressions, the quantity ρ_{f} represents the density of final states, or equivalently the atomic lineshape function, evaluated at the N-photon transition frequency.

Optical Damage

High-intensity laser fields can produce unwanted damage to optical materials. As a point of reference, the threshold for laser damage to fused silica at a wavelength of 1.05 micrometers for a pulse of 30 ps duration corresponds to an intensity of 230 GW/cm^2 or a fluence of 7 J/cm^2. Over a wide range of pulselengths (approximately 1 ps to 1 μs), the threshold intensity for laser damage decreases with pulse length T as $T^{-1/2}$ and correspondingly the threshold fluence for laser damage increases with pulse length $T^{1/2}$. In this range of pulse durations, the dominant mechanism of laser damage is avalanche breakdown. In this process, free electrons are accelerated by the laser field until they acquire sufficient energy to impact-ionize other atoms in the sample. These additional electrons are similarly accelerated and create still more free electrons. The combined action of the breaking of chemical bonds and the deposition of heat energy leads to the fracturing of the optical material. For pulses shorter than 1 ps, processes such as multiphoton absorption and multiphoton dissociation contribute to the mechanism of optical damage. For laser pulses longer than approximately 1 μs (including continuous wave laser beams), the dominant damage mechanism is direct heating of the optical material by linear absorption.

High-Harmonic Generation

Let us consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects are traditionally modeled using a power-series expansion, such as

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \cdots \qquad [8]$$

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{\mathrm{at}} = e/a_0^2 = 2 \times 10^7$ statvolt/cm $= 6 \times 10^9$ V/cm. This field strength corresponds to a laser intensity of $I_{\mathrm{at}} = 4 \times 10^{16}$ W/cm^2, which constitutes the threshold intensity for exciting nonperturbative nonlinear optical response.

One of the consequences of excitation with intensities comparable to the atomic unit of intensity I_{at}, is the occurrence of high-harmonic generation. In a typical experimental arrangement, a gas jet is irradiated by high-intensity laser radiation, and all odd harmonics of the fundamental laser frequency, up to some maximum value N_{max}, are observed. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with $N_{\max}$ as large as 341.

The phenomenon of high-harmonic generation can be understood in terms of a simple physical model. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse, and each time the electron passes near the nucleus it undergoes strong acceleration and emits a short pulse of radiation. This radiation will occur in the form of a train of short pulses; the spectrum of the radiation will be the square of the Fourier transform of this pulse train, which will contain the odd harmonics of the oscillation frequency up to some maximum frequency, that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made quantitative to show that the maximum harmonic number is given by

$$N_{\max}\hbar\omega = 3.17K + U_p \qquad [9]$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ionization energy of the atom.

Plasma Nonlinearities and Relativistic Effects

The process of multiphoton ionization can liberate a sufficient number of free electrons to transform the optical medium into a plasma, that is, a fully or partially ionized gas. The process of plasma formation is described by the equation

$$\frac{\mathrm{d}N_\mathrm{e}}{\mathrm{d}t} = \frac{\mathrm{d}N_\mathrm{i}}{\mathrm{d}t} = (N_\mathrm{T} - N_\mathrm{i})\sigma^{(N)}I^N - rN_\mathrm{e}N_\mathrm{i} \qquad [10]$$

where N_e is the number density of electons, N_i is the number density of ions, N_T is the total number of atoms (both ionized and un-ionized) in the material, and r is the electron–ion recombination coefficient. The optical properties of plasmas are very different from those of typical dielectric materials; the plasma contribution to the dielectric constant is given by

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \qquad [11]$$

where $\omega_p = \sqrt{4\pi Ne^2/m}$ is known as the plasma frequency.

Nonlinear effects can occur in the propagation of light through a plasma. One example is the nonlinear response resulting from the relativistic change in mass of the electron due to the large velocity that it can attain in the field of an intense laser beam. Detailed consideration of this effect shows that the nonlinear change in refractive index can be described as $\Delta n = n_2 I$ where

$$\begin{aligned} n_2 &= \frac{2\pi\omega_p^2 e^2}{n_0^2 m^2 c^3 \omega^4} \\ &= \frac{1}{2\pi n_0^2}\left(\frac{\omega_p}{\omega}\right)^2 1.1\times 10^{-26}\,\frac{\mathrm{cm}^2}{\mathrm{W}} \end{aligned} \qquad [12]$$

Nonlinear Quantum Electrodynamics

One can imagine an electric field so intense that it could lead to the spontaneous creation of electron–positron pairs. Such a field would have a strength of the order of $E_\mathrm{QED} = mc^2/e\bar{\lambda}_c$ where $\bar{\lambda}_c = \hbar/mc$ is the reduced Compton wavelength of the electron. The intensity of a light beam with a peak field amplitude of E_QED is $I_\mathrm{QED} = 4\times 10^{29}$ W/cm^2. Detailed consideration shows that even for fields weaker than I_QED there will be a field-induced change in the dielectric tensor given by

$$\epsilon_{ik} = \delta_{ik} + \frac{e^4\hbar}{45\pi m^4 c^7}\left[2(E^2 - B^2)\delta_{ik} + 7B_iB_k\right] \qquad [13]$$

Because of the unusual tensor properties of this relation, it displays a different polarization dependence than typical optical nonlinearities. Nonetheless, to an order of magnitude one can describe the strength of this response as

$$n_2 = \frac{7}{15c}\,\frac{e^2}{\hbar c}\,\frac{1}{E_\mathrm{QED}^2} = 5.6\times 10^{-34}\ \mathrm{cm}^2/\mathrm{W} \qquad [14]$$

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Agostini P, Fabre F, Mainfray G, Petite G and Rahman NK (1979) Free-free transitions following six photon ionization of xenon. *Physics Review Letters* 42: 1127.

Alfano RR (ed.) (1989) *The Supercontinuum Laser Source*. New York: Springer-Verlag.

Bloembergen N (1997) A brief history of light breakdown. *Journal of Nonlinear Optical Physics and Materials* 6: 377.

Boyd RW (2003) *Nonlinear Optics, Second Edition*, Section 12.5. Amsterdam: Academic Press.

Braun A, Korn G, Liu X, Du D, Squier J and Mourou G (1995) Self-channeling of high-peak-power femtosecond laser pulses in air. *Optical Letters* 20: 73.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: Frontiers of nonlinear optics. *Reviews of Modern Physics* 72: 545.

Chang Z, Rundquist A, Wang H, Murnane MM and Kapteyn HC (1997) Generation of coherent soft X-rays at 2.7 nm using high harmonics. *Physics Review Letters* 79: 2967.

Corkum PB (1993) Plasma perspective on strong field multiphoton ionization. *Physics Review Letters* 71: 1994.

Corkum PB, Rolland C and Srinivasanrao T (1986) Supercontinum generation in gases. *Physics Review Letters* 57: 2268.

Gaeta AL (2003) Collapsing light really shines. *Science* 301: 54.

Kasparian J, Rodriguez M, Mejean G, Yu J, Salmon E, Wille H, Bourayou R, Frey S, Andŕe Y-B, Mysyrowicz A, Sauerbrey R, Wolf J-P and Woste L (2003) White light filaments for atmospheric analysis. *Science* 301: 61.

Lewenstein M, Balcou Ph, Ivanov M Yu, L'Huillier A and Corkum PB (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

Max CE, Arons J and Langdon AB (1974) Self-modulation and self-focusing of electromagnetic waves in plasmas. *Physics Review Letters* 33: 209.

Moll KD, Fibich G and Gaeta AL (2003) Self-similar wave collapse: observation of the Townes profile. *Physics Review Letters* 90: 203902.

Monot P, Auguste T, Gibbon P, Jakober F, Mainfray G, Dulieu A, Louis-Jacquet M, Malka G and Miquel JL (1995) Experimental demonstration of relativistic self-channeling of a multiterawatt laser pulse in an under-dense plasma. *Physics Review Letters* 74: 2953.

Rairoux P, Schillinger H, Niedermeier S, *et al.* (2000) *Applied Physics B—Lasers Optics* 71: 573.

Ranka JK, Schirmer RW and Gaeta AL (1996) Observation of pulse splitting in nonlinear dispersive media. *Physics Review Letters* 77: 3783.

Rothenberg JE (1992) Pulse splitting during self-focusing in normally dispersive media. *Optical Letters* 17: 583.

Sprangle P, Tang C-M and Esarey E (1987) Relativistic self-focusing of short-pulse radiation beams in plasmas. *IEEE Transactions on Plasma Science* 15: 145.

Stuart BC, Feit D, Rubenchik AM, Shore BW and Perry MD (1995) Laser-induced damage in dielectrics with nano-second to subpicosecond pulses. *Physics Review Letters* 74: 2248.

Wagner R, Chen S-Y, Maksemchak A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self guided channel. *Physics Review Letters* 78: 3125.

Wood RM (1986) *Laser Damage in Optical Materials*. Bristol, UK: Adam Hilger.

NONLINEAR SOURCES

Harmonic Generation in Gases

P Villoresi, University of Padova and INFM, Padova, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The generation of laser harmonics in gases provides high-brightness ultraviolet and soft X-ray radiation induced by an intense laser pulse. This phenomenon was discovered experimentally in the late 1980s, in a simple setup as that sketched in Figure 1. The laser pulse is focused in the interaction medium, which may be a gas jet formed by a pulsed electromagnetic valve, a capillary, or a small chamber filled with the gas. The beam leaving the interaction region includes a part of the laser pulse and a number of odd harmonics.

The harmonic orders generated in this way may reach and exceed an order of 300, so that this mechanism, often referred as high-order harmonics (HHs) generation (HHG), directly up-converts low frequency radiation, usually in the near-infrared spectrum, to extreme-ultraviolet (XUV) or soft X-rays. An example of a HH spectrum is shown in Figure 2, where several tens of discrete harmonic peaks are present. The striking difference with ordinary nonlinear processes is the flatness of the photon yield versus nonlinear order, up to a sharp cut-off point.

The HHG is based on the sequence of three processes: the electric field of an intense laser pulse ionizes atoms placed near the laser focus; the liberated electrons are driven by the laser electric field in a trajectory which returns on the parent ion; and the electron recombines releasing a photon whose energy is the ionization potential plus the electron's kinetic energy. The process is periodic in time, with the period of the laser pulse, and the emitted spectrum is then expressed by discrete components, multiples of the laser frequency. The emission is in the form of a traveling wave in the direction of the laser beam and so only odd harmonics are present.

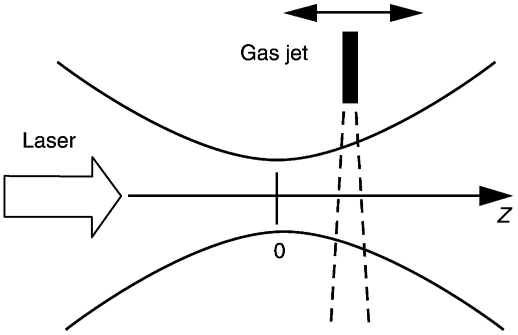

Figure 1 Scheme of the interaction between a focused laser pulse and a gas jet. *z* axis originates at laser focus.

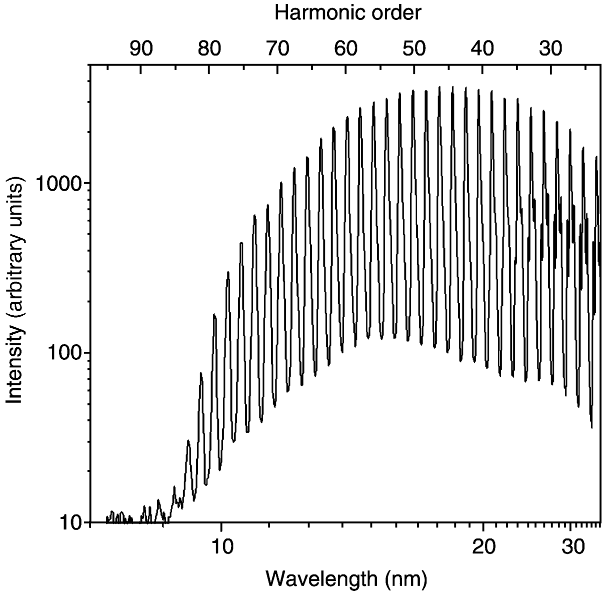

Figure 2 HHs experimental spectrum, obtained from helium gas, and a Ti:sapphire laser pulse at 790 nm, with 25 fs duration and 6×10^{14} W/cm^2 intensity.

The individual processes in HHG have highlighted the coherent response of electrons in the laser field, that show up interferences between quantum trajectories, as well as a nonlinear phase-matching issue – which have provided a simple method to investigate the interaction of single particles with the electromagnetic field, where plasma effects and bonds may be of minor influence.

Background

The interest to produce laser harmonics using gases instead of crystals was expressed by many. For the gas medium; the mass-density is three order of magnitude lower than for solids and the nonlinear susceptibility can be strongly enhanced for suitable resonance frequencies. Moreover, the phase-matching of the phase velocities may be achieved by tuning the gas pressure or by mixing diverse gases, so realizing a highly homogenous conversion region, free from walk-off and with an extension much longer than for crystals. In addition, gases have a high threshold for the intensity-induced damage, which is also non-permanent.

This research was made possible with the introduction of chirped-pulse laser amplifiers in the early 1980s, when picosecond lasers pulses reached intensities capable of directly ionizing solids and matter with their own electric field. This effect, known as optical ionization, was further enhanced when the laser pulses were reduced to subpicosecond duration, up to a few femtoseconds (10^{-15} s), close to their natural limit, which is the optical cycle. The experimental observation of HHs and of the ejection of electrons with unexpected high energy, the above threshold ionization (ATI), dramatically changed the scenario of laser matter interaction and has indicated a new regime in the coupling of laser radiation with the matter. The ordinary nonlinear optics, using the perturbative approach, was not capable of interpreting the findings and so the request was made for new models for the description of the interaction, which has been now named nonperturbative.

Typical Experimental Parameters

In the experimental investigation of the HHG, many laser wavelengths and pulse durations were used, including radiation from excimer, visible, Ti:sapphire, Nd:YAG (first and second harmonics) and CO_2 lasers, and pulses of duration from hundreds of picoseconds to a few femtoseconds. Among these, the more effective strategy in the generation of high harmonic order with high conversion efficiency, has been that of near infrared pulses with shortest duration. In order to induce optical single-ionization, the required laser intensity needs to span in the 10^{14} W/cm^2 range. Different elements are a possible choice for the gas in the interaction medium, though noble gases are used for the generation of the highest orders, since these have the higher ionization potential. Due to the very short duration of the laser pulse, the gas temperature is not relevant, while a pressure range of 10–500 mbar is usually chosen to balance, on one side, the needs of a large number of atoms that take part in the HHG and on the other side, the limit posed by the attenuation of the XUV and soft X-ray radiation after the generation, caused by the gas absorption. For the same reason, a powerful pumping system is needed, in order to reduce the ambient pressure in the apparatus below 10^{-5} mbar level.

The Models for the High Harmonics Generation

The physical processes involved in HHG are many: in the emission of the HH photons, the laser field interacts with a single atom of gas and induces the ionization, the electron's motion and its recollision with its parent ion; besides, the beam of HHs is the result of the superposition of the radiation from all the ionized atoms, as in an ordinary nonlinear medium, where the phase velocity of the laser field and of HHs is a function of the position and of the laser intensity and phase; therefore, the phase-matching issue has to be addressed in order to understand the global yield of HHG. The more relevant causes for the mismatch are the geometrical phase shift at the beam focus, which for Gaussian-profile beams is the Gouy phase, the nonlinear phase from the atomic polarization, and the effects of the plasma induced by the laser pulse itself. In addition, other effects have to be considered, such as the nonuniform density distribution of the target gas or the fact that the propagation of XUV radiation through a gas is usually influenced by a strong absorption, which is frequency dependent.

The study of HHG is first approached by introducing what is known as the simple man model, in which simplifications are introduced in order to treat the problem in a simple classical scheme, but from which the essential characteristics are derived. We then introduce the more complex quantum model based on the path integrals.

First Step: Ionization

The first step in the HHG model is the process with which the laser electric field applies to an electron, belonging to an atom or a molecule, a force that causes its detachment. This takes place only if such a field is strong enough to liberate the electron from its atomic bond. This may happen when the laser, which is taken here as linearly polarized and periodic in time, has modified the potential well so that the electrons may tunnel through it or pass over it. In fact, the sum of the atomic potential with that associated to the laser, which can be thought of as a plane that swings between a positive and negative slope at the laser frequency, gives rise to a finite sized well, through which the electron escapes by tunneling, or eventually to an aperture from which the electron is drawn away from the nucleus. In Figure 3, is shown such a well and the tunnel path for the electron.

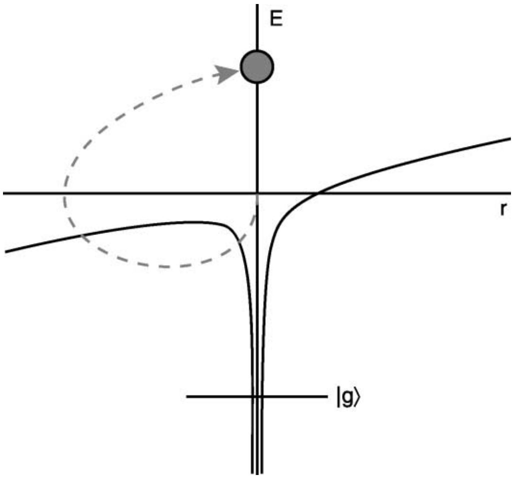

Figure 3 Electron potential when the laser pulse shines on the atom; |g⟩ indicated the level of the ground state; in red, a possible trajectory of the liberated electron after tunneling.

Once freed, the electron responds to the periodic force from the laser which makes it oscillate. The average kinetic energy on such oscillations is known as the ponderomotive potential U_p, and is related to the laser intensity and wavelength, according to the following: $U_p = e^2E_0^2/4m\omega_0^2$, which reads numerically as $U_p(\mathrm{eV}) = 9.3I_0(10^{14}\ \mathrm{W/cm^2})\lambda^2(\mu\mathrm{m})$, where λ is the laser wavelength, E_0 and I_0 the laser peak electric field and intensity.

Second Step: Free-electron Trajectory

On the freed electron are applied both the time-varying force due to the laser field and the ionic potential. Its motion can be approximated according to the simple man model as a nonrelativistic classical-mechanics trajectory; its initial velocity is set to zero, and its initial position is by the ion; finally, the laser exerts a force much stronger than that from the ion as well as from the laser magnetic field, thus these two latter forces are cancelled out.

As mentioned above, the electron trajectories that induce HHs are those where the electron returns by the ion with strong kinetic energy. If the laser is linearly polarized, the induced trajectory takes place along a line, at which also belongs to the parent ion. Moreover, the electron final velocity is related to the actual instant of ionization, or phase of the field, with respect to the laser peak. Therefore, electron final energies span from zero to a maximum value $K_{\mathrm{max.}}$ By considering the first semi-cycle, K_{max} occurs at about 17° of field phase angle, while for the second semi-cycle it is at 197°.

In Figure 4 the electron trajectories for different ionization in the first semi-cycle are shown: the electron displacement in the laser-field direction with

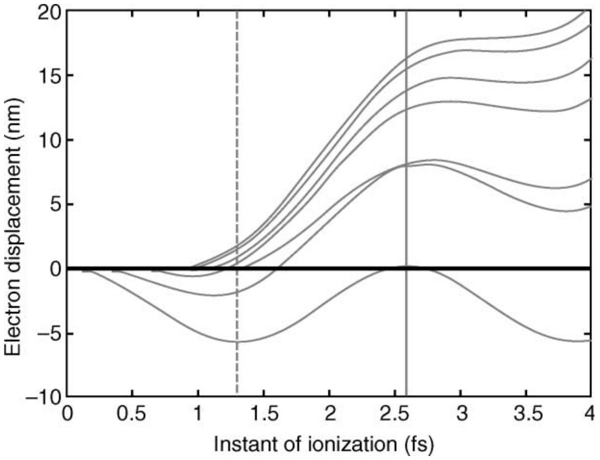

Figure 4 Electron trajectories relative to different ionization instant within the first semi-cycle; in dotted red, half-cycle, in continuous red, the whole optical cycle. Trajectories for later times do not return to the ion, do not contribute to HHG and correspond to ejected photoelectrons.

respect to the ion position, which is the initial as well as the final position, is represented versus time, where the time origin is at the laser peak, where field phase is equal to zero. The laser intensity directly influences the value of $K_{\max}$, according to the following expression: $K_{\max} = 3.17\,U_p$. In the case shown in the figure, the peak laser intensity is $I_0 = 8 \times 10^{14}$ W/cm^2 and the ponderomotive potential is $U_p = 45$ eV and $K_{\max} = 143$ eV.

Moreover, for a given final energy below $K_{\max}$, there are more than one value of the phase of the laser field where the trajectory results with the same final kinetic energy. This indicates that multiple trajectories may produce a given HH, and this is relevant in terms of the interference between them.

Third Step: Recombination

The final process which occurs in the HHG, is the recombination of the ionized electron with the parent ion. The energy conservation means that the kinetic energy and the ionization potential have to be delivered to the emitted photon. In the simple man model, the spectrum of HH then spans from I_p to the maximum values of $E_{\max} = I_p + 3.17\,K_{\max}$. In the case of Figure 2, obtained using helium gas at 6×10^{14} W/cm^2, the spectral position of cutoff can be estimated at about 9 nm, or 140 eV, which is in good agreement with the prediction of the relationship described above.

The electron recombination is not a unique process occurring at this point: the electron may be scattered by the ion and be emitted with a kinetic energy of up to 10 U_p, which is known as the rescattering plateau in ATI, or it may ionize the ion by collision, leaving a double ionized atom via sequential ionization, with a strong increase in the observed cross-section of this process.

Feynman Path-Integral Approach

The ideas in the simple man model have been expanded in order to include the quantum mechanical description of the laser-electron-ion interaction. By using the Feynman path-integral approach, the evolution of the electron driven by the laser force is associated to a probability which depends on the sum of an infinite number of amplitudes $\mathbf{f}_n$. These are complex numbers and correspond to different trajectories, or quantum orbits, $\mathbf{r}_n(t)$, followed by the electron. Their phase is proportional to the classical action of the electron motion, $S[\mathbf{r}_n(t)]$, calculated along $\mathbf{r}_n(t)$, which begins at time t_i and ends at a later time $t_f = t_i + \tau$. Here τ defines the duration of the liberated electron trajectory. In our case, the action is the sum of three terms:

$$S[\mathbf{r}_n(t)] = S_{\text{bound},n} + S_{\text{free},n} + S_{\text{final},n} \qquad [1]$$

that correspond respectively to the electron ionization:

$$S_{\text{bound},n} = -E_0 t_{i,n} \qquad [2]$$

where E_0 is the binding energy of electron in the atom ground state, $E_0 < 0$, to the free electron trajectory:

$$S_{\text{free},n} = \int_{t_i}^{t_f} [\mathbf{k} - e\mathbf{A}(t)^2]\mathrm{d}t \qquad [3]$$

where $\mathbf{k}$ is the electron momentum after ionization, directed parallel the laser electric field, and $\mathbf{A}(t)$ is the vector potential of the laser field, and to the final recombination with the parent ion:

$$S_{\text{final},n} = (\hbar\omega_q + E_0)t_{f,n} \qquad [4]$$

The condition that normally points out the trajectory made by the electron may be obtained from the principle of least action. Within the path integral approach, this principle can be used to deduce some equations which have to be obeyed by the quantum orbits, and that express the energy conservation in both the ionization and recombination process and the fulfilment of boundary conditions for electron trajectory. These are known as saddle-point equations, and reads as:

$$[\mathbf{k} - e\mathbf{A}(t_i)]^2 = 2mE_0 \qquad [5]$$

for the ionization process:

$$\mathbf{k}\tau = e\int_{t_i}^{t_f} \mathbf{A}(t)\mathrm{d}t \qquad [6]$$

which compares the initial momentum with that exchanged with the laser field during the trajectory, and

$$[\mathbf{k} - e\mathbf{A}(t_f)]^2 = 2m(\hbar\omega_q + E_0) \qquad [7]$$

for the recombination process which emits an harmonic of order q.

From eqn [5] we note that the solution for the ionization time t_i is a complex number, since the right hand side is negative. This suggests that the quantum orbits are complex as well. The reason for this is that the tunnel through the potential well is a nonclassical process.

The orbits solutions to eqns [5]–[7] have different durations, and for some of them τ is shorter than an optical cycle, while for the others the electron returns in a later cycle than the one where it was ionized. This latter solution is of minor influence, since the electron wavefunction spreads out along the trajectory, decreasing the recombination probability. Those within the initial optical cycle sum the amplitudes to give the probability of the generation of q-th harmonic, and their interference is evident from the irregular single-atom spectrum.

The time evolution of the laser pulse as well as the interaction geometry are experimental tools that influence the relevance of the different trajectories. In particular conditions it is possible to set the condition for HHG from a single trajectory, with a very simple phase structure and clear spectrum; some cases are described below.

HHs Generated by Few-Optical-Cycle Pulses

The laser electric field considered so far has been taken as stationary, relying on the hypothesis that the pulse envelope has a duration much longer than the laser optical cycle. For the analysis of the electron trajectories, that last a fraction of the laser optical cycle, this hypothesis is valid for pulses down to about 20 fs. For shorter pulses, as those generated with the hollow fiber technique, there is a noticeable variation of the electric field between a cycle and the next. The generation process is now aperiodic, and the characteristics of HH spectrum and conversion is strongly influenced by the pulse actual duration.

In order to show the striking changes in this regime, Figure 5 shows the laser electric field, the envelope and the ionization probability of helium gas in the case of a pulse of 5 fs of full width at half-maximum (FWHM) duration and a laser intensity of 9×10^{14} W/cm^2. The atom can be ionized during only three intervals, of sub-fs duration but where only one is dominant.

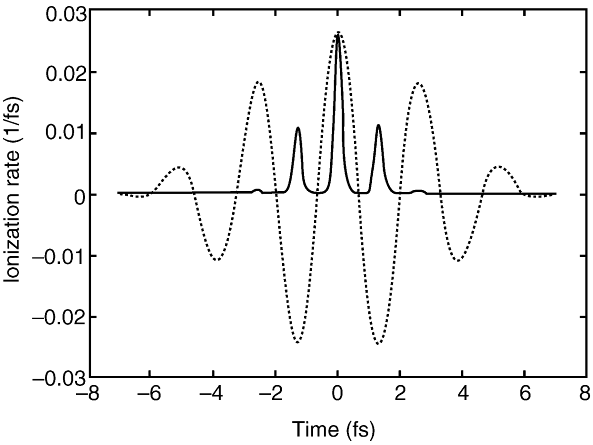

Figure 5 Ionization rate for helium gas shined by a 5 fs FWHM pulse of intensity 9×10^{14} W/cm^2. The dotted line shows the normalized evolution of electric field of the pulse. Only the three central half-cycles induce ionization.

The HHG experiment in this temporal regime have revealed that the influence of the plasma in the interaction region is progressively reduced and eventually vanishes for pulses of duration of about 5 fs. This is because the atoms are exposed directly to the maximum electric field, and so there is no pre-ionization of the gas caused by the initial part of the pulse. This effect, which is present in the case of many-cycle pulses, causes a reduction in the HHG efficiency, due to a depletion of the neutral density available at the time of maximum electric field, where the higher HHs are produced with a large dispersion, which deteriorates the phase-matching.

Moreover, the effect of the carrier-envelope-phase in the HHs spectrum has been demonstrated as an evident spectral shift of the cut-off harmonics generated with few-cycle pulses. The reason can be found in the variation of the phase of the emission from each electron after its trajectory. For the few-cycle pulses, this phase changes for every laser semi-cycle, due to the change in the actual electric field which drives the electron motion.

This portion of the spectrum is also relevant to its phase structure, since here the HHs are generated by a single electron trajectory. This condition sets a constant phase difference between the subsequent harmonic orders. In this way, similarly to the mode locking laser, this effect allows the synthesis of a pulse, from the combination of several harmonic orders, which has a duration much lower than the individual harmonics. This process has been exploited to produce a soft X-ray pulse of duration of 650 as (1 as, attosecond = 1×10^{-18} s).

HHs Beam Characteristics

The HHG is a coherent process, in that the emitted radiation has a direct relationship with the laser pump. The emitted HH beam also reflects this fact in its good directional characteristics and in its spatial and temporal coherence, as different experiments have demonstrated.

Using a 110 fs laser pulse, the interference fringes produced in the far field of two spatially separated HH beams were measured as a function of the mutual delay. The HH coherence time t_c resulted in agreement with the expected duration of the HHs, which is a fraction of that of the laser pulse. Moreover, it has been observed at a different t_c in the central and external regions of the far field of harmonic 23rd, 50 fs and 20 fs respectively. This is because different phase-matching in the HHG contribute to the on-axis and off-axis part of the HHs beam, where the more prominent quantum orbit is also different.

The measurement of the beam divergence for different harmonic orders is relevant for both the estimate of the beam brightness and the irradiance available in the case of the HH applications. Moreover, it is useful for the understanding of the phase-matching which is realized in actual interaction geometry. The experimental measurements requires an astigmatic spectrometer, as is described later. In this kind of experiment, the focalplane images of the HH spectrum are taken for different generation conditions, usually moving the gas-jet position with respect to the laser focus, the z coordinate in Figure 1. The harmonic peaks appear as elongated spots, as shown in Figure 6 for helium gas and an intensity of 4×10^{14} W/cm^2 and duration of 20 fs, whose length in a transverse direction to the dispersion plane can be related, by a optical calculation, to the beam divergence.

A comparison between divergences for HHG driven by few- or many-optical-cycle pulses has revealed another aspect of the phase-matching condition. In this case, the laser pulse was compressed in time by means of the hollow fiber technique, and so the resulting phase front is of the Bessel-beam type, truncated at the first zero. The experiment indicates a divergence for plateau harmonics of about 3 mrad FWHM of the beam, in the case of the 7 fs pulse and larger values by more than 50% using the 20 fs pulse. In both cases, the beam divergence increase moving downstream the gas jet and with increasing harmonic order. The reduced divergence can be ascribed to a more regular wavefront of the harmonic beam, which can be related to the simpler phase structure of the few-cycle case.

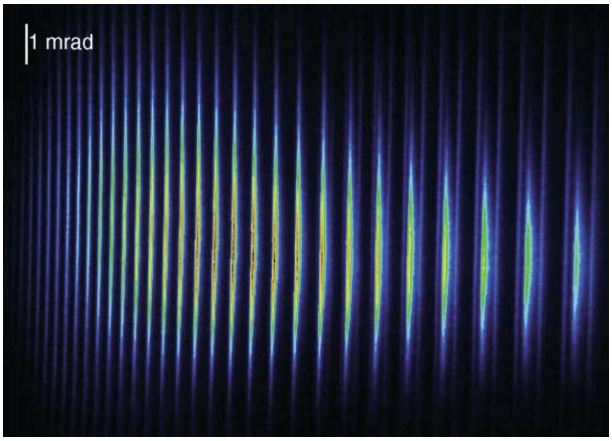

Figure 6 Focal plane image obtained from the spectrograph shown in Figure 9, relative to the HHs spectrum of helium with 20 fs and 4×10^{14} W/cm^2. The bar in upper-left corner indicates the scale-length of vertical axis corresponding to a divergence of the HH beam of 1 mrad.

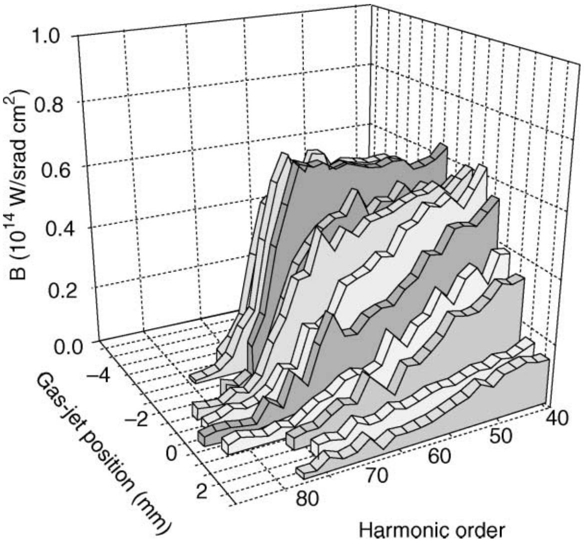

Figure 7 Peak brightness of the HHs beam determined from absolute-intensity and divergence measurements, using neon gas and a 790 nm laser pulse with 7 fs duration and 9×10^{14} W/cm^2 intensity.

By using an intensity-calibrated detector, the number of photons per harmonic order per laser shot can be measured. Combining this measure with that of the divergence, the HH beam brightness can be calculated. The source size and the HH pulse duration can be easily modeled from the experimental parameters. The result for the case of neon with 7 fs pump-pulse with intensity of 9×10^{14} W/cm^2 is shown in Figure 7, as a function of the gas-jet position, with positive values when the jet is downstream with respect to laser focus.

The large divergence reduces the brightness of downstream positions. Maximum values of about 8×10^{13} W/srad cm^2 are found close to the focus, relatively flat with respect to harmonic order.

The conversion efficiency η_{HH} from the fundamental intensity to that of the HHs can also be derived from this kind of measurement. The more convenient parameters for HHG in the range 41–71 were found using neon gas jet from a pulsed valve and a pulse duration of 7 fs, which leads to η_{HH} in the scale of 10^{-7}, while for lower orders, from 17 to 31, a value about two orders of magnitude larger has been obtained using a short hollow fiber filled with argon gas. A remarkable intracycle optimization mechanism was introduced in this case, in order to maximize the given harmonic order of chosen. This was done by tailoring the shape of the pump pulse in spectrum and time, using an adaptive component in the laser compressor. In this case, the optimal pulse shape to maximize the efficiency in the driving of the electron on a trajectory producing the required harmonic order, is sought by means of a genetic algorithm.

A strong efficiency enhancement has also been demonstrated by manipulating the laser wavefront by means of a deformable mirror (DM). In this case, the DM was controlled by a genetic algorithm, aiming to increase the HHs intensity in a given spectral interval. The purpose of this technique is the correction of the optical aberrations on the beam, introduced by both the optical components in the beam focusing line and the phase distortion caused by nonlinear effects leading to an optimum laser front. The effect is shown in Figure 8, where the spectrum obtained with the original beam is compared with that optimized by the genetic algorithm. It can be noticed there is both a remarkable extension of the HHs spectrum, of about 20 orders, as well as a strong enhancement of the plateau intensity.

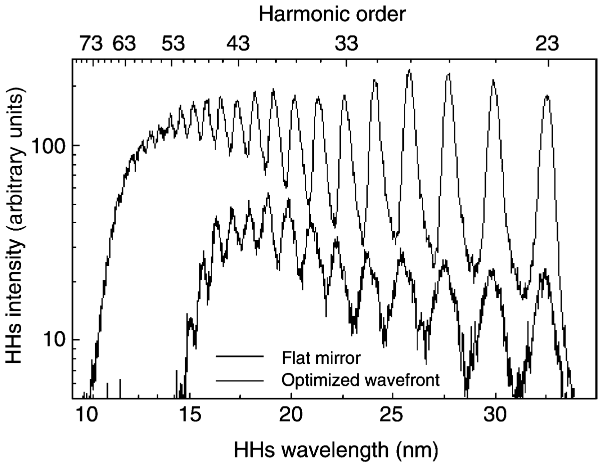

Figure 8 Comparison of the HHs spectrum obtained with and without the wavefront correction using a deformable mirror. Neon gas was used and a 790 nm laser pulse with 6 fs duration and 0.3 mJ energy.

Spectrometers and Monochromators for HHs

The extended spectrum of HHs reaches the optical domains of the extreme-ultraviolet (XUV, 10–100 nm) and soft X-ray (1–10 nm). In these regions, the optical materials have very poor or zero transmission, and so the optical design of instruments for HHs analysis uses reflective optics only. Exceptions sometimes used are the frequency-selective XUV transmission filters, whose thickness does not exceed a few tenths of a micrometer, and the diffracting zone-plates. Moreover, for the radiation of wavelength shorter than 30 nm, even the reflectivity of optics becomes poor, and this forces the use of grazing-incidence mountings or the narrow band XUV and soft X-ray multilayer mirrors.

On the other hand, HHs as optical sources have favorable characteristics, in that they are emitted in a region smaller than the laser waist focus, so usually of a few tens of micrometers, and their divergence is small, as seen above. In addition, the very short duration of the pump pulses broadens the HHs spectrum, due to time-bandwidth inequality, usually smearing details below 0.1 nm of width.

The spectrograph is then realized with the least number of components, usually combining in one or two optics of the spectral dispersion, the gathering of the HH beam and its focusing on the detector. Moreover, it is not feasible to collect with a single detector all the broad HH spectrum in once, so usually a spectral interval of interest is selected and acquired with a linear or bidimensional detector, or scanned sequentially by rotating the grating.

An example of spectrograph with 2D-detector, is shown in Figure 9, the layout of the instrument developed for the simultaneous measure of the divergence and the intensity of HHs generated from few-optical-cycle pulses.

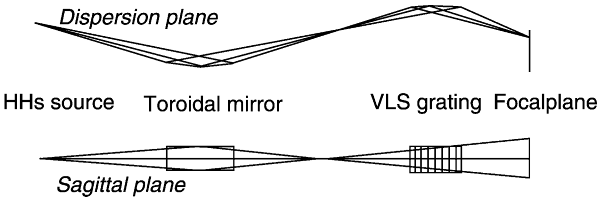

Figure 9 Optical layout of the astigmatic spectrometer for the divergence measurement.

The upper part of Figure 9 shows the optics in the dispersion plane, where the toroidal mirror gathers the HH radiation and relays it to the source point for the spherical grating. The latter has a varied line spacing (VLS) of the grooves, in order to image an erected spectrum in the focalplane. This feature is convenient in order to use a bidimensional detector. In the sagittal plane, the grating has negligible effect, and the toroidal mirror is designed to realize an astigmatic image of the source in the focalplane, from which the divergences can be extracted. The detector needs to be blind at the radiation of the laser and at the lower harmonics, and to have an extended linear response in the XUV and soft X-ray. A good solution can be obtained with an open micro-channel plate (MCP), with a suitable photocathode such as MgF_2, placed in the focalplane, coupled with a high-resolution and a high dynamic CCD camera.

For the selection of a single harmonic order or of a narrow spectral region, an XUV monochromator is used. This device is based on the spectral dispersion of the HHs beam by using a diffraction grating. The desired harmonic order is selected by scanning the wavelength which falls on the slit, usually by rotating the grating. This technique uses a single optical component, a concave grating, to reduce the reflection losses. On the other hand, the single-grating monochromator introduces a remarkable time broadening in the selected radiation. In fact, the grating diffraction is based on the optical path length difference for the rays diffracted by subsequent grating grooves. Therefore the selected beam, of wavelength λ, is composed of rays whose difference in length is $\Delta L = N_g m \lambda$, where m is the diffraction order and N_g is the number of illuminated grooves, given by the beam diameter on grating multiplied by the groove density. The time broadening of the selected pulse is then $\Delta\tau = \Delta L/c$. For a HH XUV pulse, the diffracted pulse results is 7 ps of duration, if beam size is of 3 mm and it is diffracted at 87° by a 1200 gr/mm grating.

The solution to this problem may be found in adopting a twin grating design. In this case, the dispersion induced by the first grating is used to select the desired spectral portion of the HH spectrum, and will be cancelled by the second grating. This latter grating has to be mounted in order to compensate for the different optical path length of different rays within the aperture of the HH beam. The layout of this optical scheme is reported in Figure 10. The ray tracing of this compensated monochromator indicates that for both the grazing incidence and the normal incidence regions, the residual broadening results is of the order of one femtosecond.

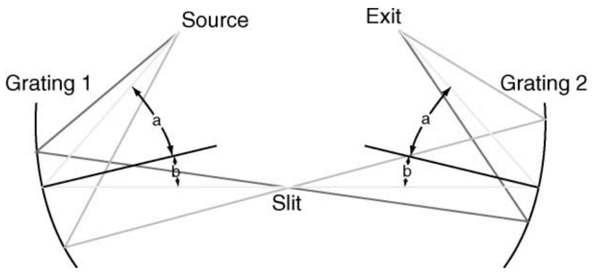

Figure 10 Scheme of the compensated XUV monochromator, for the selection of a spectral portion of the HHs spectrum without introducing temporal broadening. See also Villoresi P (1999) Compensation of optical path lengths in extreme-ultraviolet and soft X-ray monochromators for ultrafast pulses. *Applied Optics* 38: 6040.

HHs at Work: Applications and Perspectives

The advantages of HHs with respect to other XUV and soft X-ray sources, as laser-plasma, synchrotron radiation or X-ray lasers, are mainly the very short duration of their emission, their high brightness, and the source compactness and their intrinsic synchronization with a laser pulse. These aspects have been already exploited in a few experiments, based either on the detection of a photoelectron emitted in the time-dependent interaction of the matter with the HHs plus fundamental laser beams, or the observation of changes in the HHs XUV spectrum.

As examples of direct application, the interferometry of a laser-plasma has been obtained with HHs, to determine spatial distribution of the electron density. In this case, two beams of harmonic radiation were generated by splitting in equal parts the laser pulse, with one of them passing through a laser-generated plasma from a solid aluminum target. By means of a grating monochromator, a single harmonic order is selected, of both beams, and their interference pattern in the far field is acquired. From the shift in the fringes pattern, the map of the plasma electron density was estimated. Here, the broad spectrum of the HHs was exploited to operate the interference at the most convenient wavelength, far from atomic resonances and with fair fringe contrast.

The direct monitoring of a chemical reaction at the surface was achieved by pumping, with a part of the laser pulse, a platinum surface with some adsorbed molecular oxygen. By observing changes in the spectrum of the photoelectrons generated by the delayed HHs beam, the oxygen-platinum bond switch from a superoxo (O_2^-) to a peroxo (O_2^{2-}) state was observed in the time domain.

The electron bond in an atom may be dressed by a laser field, and the changes induced in its orbital are diagnosed by a HHs beam, which photo-ionizes the atom. This phenomenon have been exploited as a diagnostic of the temporal structure of a HHs beam: the sidebands in the photoelectron spectrum have been related to the phase of different harmonics. The findings indicate that HHs are locked in phase and their superposition produces a train of 250 as pulses.

By using the mixing scheme of some suitable orders of a HHs beam generated with few-optical-cycle pulses, mentioned already in section 'HHs beam characteristics,' a pulse with sub-femtosecond duration may be generated. This was used in the detailed analysis of the photoelectron spectrum in a pump-probe experiment, with time resolution of 150 as. Again using the IR and HHs radiation together, this technique has accessed the direct probe of the laser field oscillation, as demonstration by the use of HHs in the new domain of attosecond spectroscopy and metrology.

Finally, the HH radiation can also be considered as the seed for an amplifier of radiation in the soft X-ray region, as in the case of a free electron laser.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Instrumentation:** Spectrometers. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Bartels R, Backus S, Zeek E, *et al.* (2000) Shaped-pulse optimization of coherent emission of high-harmonic soft x-ray. *Nature (London)* 406: 164.

Bauer M, Lei C, Read K, *et al.* (2001) Direct observation of surface chemistry using ultrafast soft-X-ray pulses. *Physics Review Letters* 87: 025501.

Brabec T and Krausz F (2000) Intense few-cycle laser fields: frontiers of nonlinear optics. *Review Modern Physics* 72: 545.

Corkum PB (1993) Plasma perspective on strong-field multiphoton ionization. *Physics Review Letters* 71: 1994.

Descamps D, Lynga L, Norin J, *et al.* (2000) Extreme ultraviolet interferometry measurements with high-order harmonics. *Optic Letters* 25: 135.

Hentschel M, Kienberger R, Spielmann CH, *et al.* (2001) Attosecond metrology. *Nature (London)* 414: 509.

Kulander KC, Schafer KJ and Krause JL (1993) Dynamics of short-pulse excitation, ionization and harmonic conversion. In: Piraux B, L'Huillier A and Rzazewski K (eds) *Super-Intense Laser-Atom Physics*, pp. 95–110. New York: Plenum.

Lewenstein M, Balcou Ph, Ivanov M, *et al.* (1994) Theory of high-harmonic generation by low-frequency laser fields. *Physics Review A* 49: 2117.

L'Huillier A and Balcou P (1993) High-order harmonic generation in rare gases with a 1-ps 1053-nm laser. *Physics Review Letters* 70: 774.

McPherson A, Gibson G, Jara H, *et al.* (1987) Studies of multiphoton production of vacuum-ultraviolet radiation in the rare gasses. *Journal of the Optical Society of America B* 4: 595–601.

Nisoli M, Priori E, Sansone G, *et al.* (2002) High-brightness high-order harmonic generation by truncated bessel beams in the sub-10-fs regime. *Physics Review Letters* 88: 033902.

Nisoli M, Sansone G, Stagira S, *et al.* (2003) Effects of carrier-envelope phase differences of few-optical-cycle light pulses in single-shot high-order-harmonic spectra. *Physics Review Letters* 91: 213905.

Paul PM, Toma ES, Breger P, *et al.* (2001) Observation of a train of attosecond pulses from high harmonic generation. *Science* 292: 1689.

Poletto L, Tondello G and Villoresi P (2001) High-order laser harmonics detection in the EUV and soft x-ray spectral regions. *Review Scientific Instruments* 72: 2868.

Salières P, L'Huillier P, Antoine A, *et al.* (1999) Study of spatial and temporal coherence of high order harmonics. In: Bederson B and Walther H (eds) *Advances in Atomic, Molecular, and Optical Physics*, vol. 41, pp. 83. New York: Academic Press.

Villoresi P, Ceccherini P, Poletto L and Tondello G (2000) Spectral features and modeling of high-order harmonics generated by sub-10-fs pulses. *Physics Review Letters* 85: 2494.

Villoresi P, Bonora S, Pascolini M, *et al.* (2004) Optimization of high-order harmonic generation by adaptive control of sub-10 fs pulse wavefront. *Optic Letters* 29: 207.

O

OPTICAL AMPLIFIERS

Contents

Basic Concepts
Erbium Doped Fiber Amplifiers for Lightwave Systems
Optical Amplifiers in Long-Haul Transmission Systems
Raman, Brillouin and Parametric Amplifiers
Semiconductor Optical Amplifiers

Basic Concepts

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The transmission distance of a fiber-optic communication system is limited by fiber loss and dispersion. For long-haul lightwave systems, the loss limitations have traditionally been overcome by periodic regeneration of the optical signals at repeaters applying conversion to an intermediate electric signal. Because of the complexity and high cost of such regenerators, the need for optical amplifiers became obvious in the mid-1980s. The optical amplifier is ideally a transparent box that provides gain and is also insensitive to the bit rate, modulation format, power and wavelength of the signal passing through it.

Several means of obtaining optical amplification has been suggested since the 1970s, including direct use of the transmission fiber as gain medium through nonlinear effects, semiconductor amplifiers, or doping optical waveguides with an active material (rare-earth ions), that could provide gain. Due to the spectacular results on erbium-doped fiber amplifiers, which are particularly suitable in the third transmission window (around 1.5 μm), an intense worldwide research activity on optical amplifiers has developed. As a consequence, the development of erbium-doped fiber amplifiers has reached an industrial level, and commercial devices are now available.

Semiconductor amplifiers, on the other hand, have the same technical basis as semiconductor lasers. Although the strong nonlinearity of semiconductor amplifiers degrades the performances of transmission systems, the state-of-art semiconductor devices seem to be the most interesting amplifiers for transmission in the second transmission window (around 1.3 μm).

Amplifier Gain and Bandwidth

In a perfect amplifier, the amplification process would be insensitive to the bit rate, modulation format, power, state of polarization, wavelength, and optical bandwidth of the signal passing through it. On the other hand, no interaction would take place if more than one signal were amplified simultaneously. In practice, however, the optical gain depends not only on the wavelength (or frequency) of the incident signal, but also on the electromagnetic field intensity at any point inside the amplifier. Details of wavelength and intensity dependence of the optical signal depend on the amplifying medium.

We consider a case in which the gain medium is modeled as a homogeneously broadened two-level system. In such medium, the gain coefficient (i.e., the gain per unit length) can be written as:

$$g(\nu, P) = \frac{g_0}{1 + \frac{(\nu - \nu_0)^2}{\Delta\nu_0^2} + \frac{P}{P_{\text{sat}}}} \qquad [1]$$

where g_0 is the peak value of the gain coefficient determined by the pumping level of the amplifier, ν the optical frequency, ν_0 the atomic transition frequency, $\Delta\nu_0$ the 3 dB local gain bandwidth, P the optical power of the signal, and P_{sat} the saturation power, which depends on the gain medium parameters. It must be emphasized that $\Delta\nu_0$ and P_{sat} refer to the local gain. However, from the communication system point of view, it is more desirable to use the related concepts of amplifier bandwidth and amplifier saturation power that will be evaluated below.

The amplifier gain G is defined as

$$G = \frac{P_{\text{out}}}{P_{\text{in}}} \qquad [2]$$

where P_{in} is the input power and P_{out} the output power of a continuous wave (CW) signal being amplified. The amplifier gain G may be found by using the relation:

$$\frac{\mathrm{d}P}{\mathrm{d}z} = g(\nu, P)P \qquad [3]$$

where $P(z)$ is the optical power at a distance z from the amplifier input end.

If the signal power obeys the condition $P \ll P_{\text{sat}}$ throughout the amplifier, the gain coefficient given by eqn [1] can be considered independent of the signal power. In such a case, the amplifier is said to be operated in the unsaturated regime and works as a linear device. The gain coefficient presents in this situation a Lorentzian profile that is characteristic of homogeneously broadened two-level systems. However, the gain spectrum of actual amplifiers can deviate significantly from the Lorentzian profile.

The solution of eqn [3] in the unsaturated regime is an exponentially growing signal power, given by

$$P(z) = P_{\text{in}} \exp(gz) \qquad [4]$$

For an amplifier length L, we then find that the linear amplifier gain is

$$G(\nu) = \exp(gL) = \exp\left[\frac{g_0 L}{1 + (\nu - \nu_0)^2/\Delta\nu_0^2}\right] \qquad [5]$$

Both the amplifier gain $G(\nu)$ and the gain coefficient $g(\nu)$ are maximum when $\nu = \nu_0$. However, $G(\nu)$ decreases much faster than $g(\nu)$ with the signal detuning $\nu - \nu_0$, because of the exponential dependence of G on g. As a consequence, the amplifier bandwidth $\Delta\nu_A$, which is defined as the FWHM of $G(\nu)$, is much smaller than the gain bandwidth $\Delta\nu_0$ (Figure 1). This can result in signal distortion in the case where a broadband optical signal is transmitted through the amplifier. From eqn [5] we can obtain the following relation between the amplifier bandwidth and the gain bandwidth:

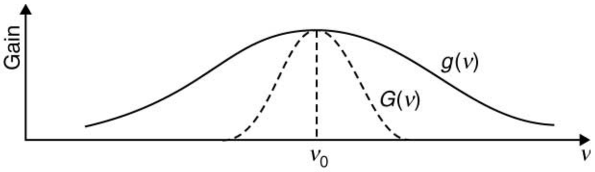

Figure 1 Gain coefficient profile $g(\nu)$ and the corresponding amplifier gain spectrum $G(\nu)$.

$$\Delta\nu_A = \Delta\nu_0 \sqrt{\frac{\ln 2}{g_0 L - \ln 2}} \qquad [6]$$

Gain Saturation

An important limitation of the nonideal amplifier is related with the power dependence of the gain coefficient given by eqn [1]. This property is known as gain saturation and it appears when the signal power ratio P/P_{sat} is non-negligible. Since the gain coefficient is reduced when the signal power P becomes comparable to the saturation power P_{sat}, the amplifier gain G will also decrease.

Assuming that $\nu = \nu_0$ and substituting g from eqn [1] in eqn [3] gives

$$\frac{\mathrm{d}P}{\mathrm{d}z} = \frac{g_0 P}{1 + P/P_{\text{sat}}} \qquad [7]$$

Considering the initial condition $P(0) = P_{\text{in}}$, we obtain from eqns [2] and [7] the following implicit relation for the amplifier gain:

$$(1 - G)\frac{P_{\text{in}}}{P_{\text{sat}}} = \ln\left(\frac{G}{G_0}\right) \qquad [8]$$

where $G_0 = \exp(g_0 L)$. The input saturation power P_{in}^{s} is defined as the input power for which the amplifier gain G is reduced by a factor of 2 from its unsaturated value G_0 (Figure 2). Indeed, it is obtained by using $G = G_0/2$ in eqn [8]:

$$P_{\text{in}}^{\text{s}} = \frac{2\ln(2)P_{\text{sat}}}{(G_0 - 2)} \qquad [9]$$

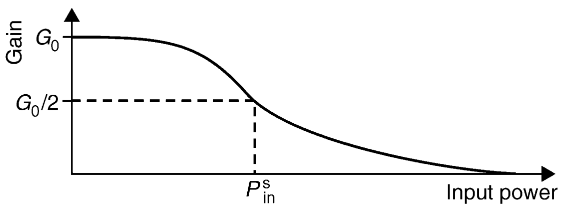

Figure 2 Saturated amplifier gain as a function of the input power.

As observed from eqn [9], the input saturation power P^s_{in} does not coincide with P_{sat}. The output saturation power is given by $P^s_{out} = G_0 P^s_{in}/2$. In practice, $G_0 \gg 2$ and P^s_{out} is found to be smaller than P_{sat}, by about 30%.

Gain saturation can be seen as a serious limitation, particularly for multichannel communication systems. However, the self-regulating effect of gain saturated amplifiers can be useful in long-haul lightwave communication systems, including many concatenated amplifiers. In fact, if the signal level in a chain of amplifiers is unexpectedly increased along the chain, the saturation effect causes a lower gain provided by the following amplifiers and vice versa for a sudden signal power decrease.

Amplifier Noise

Besides the bandwidth and gain saturation limitations, another property must be considered concerning practical optical amplifiers. In fact, optical amplifiers always add spontaneously emitted photons to the signal during the amplification process. Those photons are amplified besides the signal photons so that, at the amplifier output, an amplified spontaneous emission (ASE) noise is presented. Since spontaneous emission always takes place, ASE noise is unavoidable and does not depend on the amplifier temperature. This is one of the most important differences between optical and electrical amplifiers, where amplifier noise is of thermal origin and can be reduced by lowering the amplifier temperature.

The ASE determines a degradation of the signal-to-noise ratio (SNR). The SNR degradation is usually characterized by the amplifier noise figure, which is defined as the SNR ratio between input and output:

$$\mathrm{NF} = \frac{\mathrm{SNR_{in}}}{\mathrm{SNR_{out}}} \qquad [10]$$

The SNR is usually referred to the electrical power generated when the optical signal is converted to electrical current by using a photodetector. Therefore, the noise figure as defined in eqn [10] would usually depend on several detector parameters, which determine the shot noise and thermal noise associated with the practical detector. We will consider the case of an ideal detector, whose performance is limited by shot noise only.

The SNR of the input signal is simply determined by the detection shot noise and can be written as:

$$\mathrm{SNR_{in}} = \frac{P_{in}}{2h\nu\Delta f} \qquad [11]$$

where Δf is the detector bandwidth.

To evaluate the term $\mathrm{SNR_{out}}$, one should add the contribution of spontaneous emission to the receiver noise. The ASE noise spectral density is assumed to be constant and can be written as

$$S_{sp}(\nu) = (G-1)n_{sp}h\nu \qquad [12]$$

where G the amplifier gain and

$$n_{sp} = \frac{N_1}{N_1 - N_0} \qquad [13]$$

is known as the spontaneous emission factor or the population inversion factor. In eqn [13] N_0 and N_1 are the atomic populations for the ground and excited states, respectively.

Considering a low noise amplifier, the signal power impinging the photodetector is larger than the optical noise power and the shot noise power. As a consequence, the electrical noise, due to the signal-ASE beating, is the dominant contribution and the SNR of the amplified signal is given by:

$$\mathrm{SNR_{out}} \approx \frac{GP_{in}}{4S_{sp}\Delta f} \qquad [14]$$

Using eqns [11]–[14], the amplifier noise figure defined by eqn [10] becomes:

$$\mathrm{NF} = 2n_{sp}\frac{G-1}{G} \approx 2n_{sp} \qquad [15]$$

where the approximation holds when the gain is much higher than one. In the case of an ideal amplifier, $n_{sp} = 1$ and eqn [15] show that the SNR is degraded by 3 dB. For most practical amplifiers, NF can be as large as 6–8 dB.

Basic Amplifier Configurations

The relative importance of the different limiting factors discussed above depends on the actual

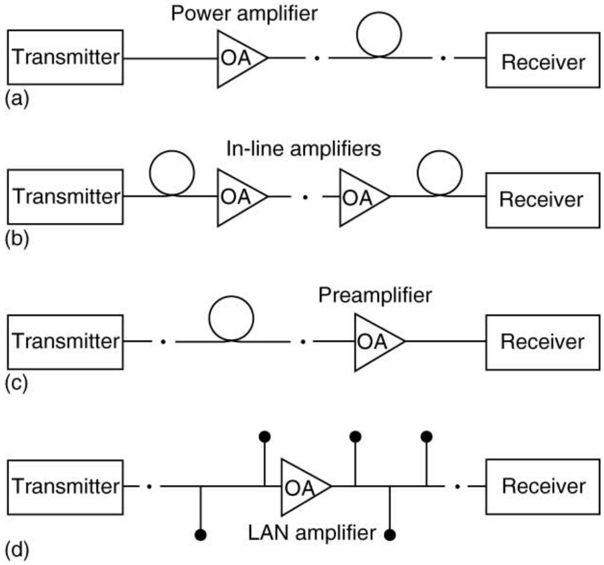

Figure 3 Four generic configurations for incorporating optical amplifiers into transmission systems: (a) as a power amplifier; (b) as in-line amplifiers; (c) as a preamplifier; (d) for compensation of distribution losses in local-area networks.

amplifier application. Figure 3 shows the four basic system configurations envisioned for the incorporation of optical amplifiers. The first configuration is to place the amplifier immediately following the laser transmitter to act as a power amplifier or booster (Figure 3a). The main purpose of such amplifiers is to boost the signal power, which can provide an increase of the transmission distance by 100 km or more. Since the signal input power is typically large (0.1–1.0 mW), the key parameter for the power amplifier will be to maximize the saturation output power and not necessarily the absolute gain.

The second configuration is to place the amplifier in-line and perhaps incorporated at one or more places along the transmission path (Figure 3b), replacing the electronic regenerators. The in-line amplifier corrects for periodic signal attenuation and may exist in a cascade form. The use of in-line optical amplifiers is particularly attractive for multichannel communication systems, since they can amplify all channels simultaneously.

The third configuration consists of using the amplifier immediately before the receiver, so it functions as a preamplifier (Figure 3c). The purpose of such an amplifier is to improve the receiver sensitivity. The main figures of merit are high gain and low amplifier noise, because the entire amplifier output is immediately detected.

In local-area networks (LANs), distribution losses often limit the number of possible nodes. The fourth application of optical amplifiers consists of using them for compensating such distribution losses (Figure 3d).

List of Units and Nomenclature

g_0	peak value of the gain coefficient at peak	m^{-1}
$g(\nu)$	gain coefficient	m^{-1}
G_0	unsaturated amplifier gain	
$G(\nu)$	amplifier gain	
L	amplifier length	m
n_{sp}	spontaneous emission factor	
P	signal power	W
P_{in}	input signal power	W
$P_{\mathrm{in}}^{\mathrm{s}}$	input saturation power	W
P_{out}	output signal power	W
$P_{\mathrm{out}}^{\mathrm{s}}$	output saturation power	W
P_{sat}	saturation power	W
$S_{\mathrm{sp}}(\nu)$	ASE noise spectral density	J
Δf	detector bandwidth	Hz
$\Delta\nu_0$	bandwidth of the gain coefficient	Hz
ν	optical frequency	Hz
ν_0	atomic transition frequency	Hz

See also

Optical Amplifiers: Optical Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*. New York: Wiley.

Bjarklev A (1993) *Optical Fiber Amplifiers: Design and System Applications*. Boston, MA: Artech House.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Green PE Jr (1993) *Fiber Optic Networks*. Englewood Cliffs, NJ: Prentice Hall.

Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Boston, MA: Artech House.

Shimada S and Ishio H (eds) (1994) *Optical Amplifiers and their Applications*. New York: Wiley.

Erbrium Doped Fiber Amplifiers for Lightwave Systems

P Bollond, JDS Uniphase Corporation, Ewing, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The telecommunications industry has undergone a revolution since the 1980s, by using glass optical fibers for the transmission of information encoded as pulses of light. A single telecommunications-grade optical fiber has been shown to support the propagation of more than 1 Tbit per second (1×10^{12} pulses per second) over distances comparable to typical city separations (>100 km). From this technology, optical fiber links have evolved to become a network on a planet-wide scale, and form the physical backbone of the information age.

Erbium-doped fiber amplifiers (EDFAs) are an enabling technology for optical fiber communication networks. They have several important properties that make them the amplification component of choice in long-distance commercial data transport networks. Erbium ions deposited in silica-based glass allow amplification in the lowest loss region of commercial-grade optical fiber (<0.25 dB/km from approximately 1530–1620 nm). Erbium-doped fiber (EDF) is manufactured in a form that allows for low-loss fusion splicing to standard communications fiber. Compact semiconductor laser diodes are available to excite the erbium ions into an amplification state. The long lifetime of the excited state of erbium provides the ability to simultaneously amplify multiple wavelength channels without significant cross-channel interference. Multiple channel systems have been deployed with more than 100 optical channels (or wavelengths) through each EDFA, and this has allowed network capacities to be dramatically increased. The low noise properties of the EDFA also allow networks to be constructed with many amplified spans before the optical signal has to be electronically regenerated. Practically unlimited transmission distance has been demonstrated using a small number of optical soliton channels through periodically amplified EDFA lightwave systems.

The development history of the EDFA can be traced back to the first optical amplifier. In 1962, a neodymium-based fiber amplifier was invented that operated at 1064 nm. During the 1980s, the need for an optical amplifier at telecommunications wavelengths initiated research at many locations throughout the world. In 1987, the University of Southampton (UK) was first to demonstrate an EDFA that had optical gain at 1550 nm, and during the following years the design of erbium-doped fiber was optimized for this application. In 1989, a practical semiconductor laser diode became available to pump EDF, and the first compact optical fiber amplifier modules soon appeared for commercial deployment. The traditional method of signal regeneration, prior to 1990, was to use electronics to detect the optical signal after each transmission span, recover the digital signal, and then retransmit using another laser diode. The EDFA allows practical wavelength division multiplexing (WDM) of multiple optical signals with all optical signal amplification, and provides significant performance and cost advantages over electronic regeneration.

EDFAs have emerged from the laboratory to be widely deployed in communication networks. EDFAs are used to boost transmitted power of the signal lasers (booster amplifier), amplify signals in transit to compensate losses sustained in the fiber (line amplifier), or amplify signals before a receiver (pre-amplifier). Typically, the amplifier module is specifically manufactured for particular systems that are mounted on electronic circuit boards. These circuit packs are then housed in central offices (local telephone exchanges), remote 'repeater huts', or even in undersea 'bottles' as part of a transoceanic cable. The high cost of network failure requires that the manufactured EDFA modules comply with stringent reliability criteria, to provide a useful operating lifetime greater than 25 years when subject to extreme environmental conditions.

Amplifiers are constructed for incorporation into either existing fiber links as part of an upgrade, or for newly planned systems. Because of the high cost of installing new fiber into the ground and securing property rights, it has become economically desirable to upgrade many existing fiber links rather than to build new systems. Transmission cables usually have many pairs of individual optical fibers, some of which will not initially be transporting data, and these 'un-lit' or 'dark fibers' can be activated as consumer demand increases over time. Typically, each fiber of a pair is used to carry either 'east'- or 'west'-bound traffic. Around city areas metropolitan area networks can be expanded in this way, but for long-distance links (long haul networks with distances >1000 km) operation is designed for a larger number of channels (40 to 120 wavelengths) at higher data rates (10 or 40 GHz per channel) over specialized transmission cables containing low numbers of fiber pairs. Typically communication systems are designed to meet certain cost targets, expressed as dollars per Gbit/s per km, for total transmission distances. The total transmission distance is limited by the optical signal to noise ratio (OSNR) degradation after each fiber-amplifier link, with a smaller permissible degradation at higher bit rates. The required gain and OSNR performance for the system is then translated to an EDFA module specification.

This article discusses EDFA design and applications, and shows the elements involved in producing reliable modules for commercial lightwave systems.

Components for EDFAs

EDFAs are comprised of passive optical components, erbium-doped fiber, and pump lasers. Passive components are chosen to meet optical and environmental specifications, while the erbium-doped fiber is selected based on the optical power, gain, and noise figure requirements. Pump lasers are a key influence on the price and performance of optical amplifiers.

The amplifier module is typically connected to the transmission fiber using fiber connectors. These polished fiber connectors have a higher insertion loss and reflectivity than fusion splicing, but allow for easy deployment in the network. Internal optical components are fusion spliced together to provide low loss, low back reflection, high strength, and high reliability joins. Fusion splicing is tailored to particular fiber types, so that optical components with dissimilar fiber types are joined with the lowest loss.

The signal and pump radiation is combined with low loss using optical components called wavelength division multiplexers (see Figure 1). These components are based on fused fiber or interference filter based technology. Fused fiber WDMs offer the lowest insertion loss (<0.1 dB is commercially available) but is restricted to widely spaced wavelengths (e.g., 980 nm pump and 1550 nm signal). Interference filter based WDMs are available for closely spaced wavelengths (e.g., 1480 nm pump and 1530 nm signal) and have a very low wavelength dependent insertion loss (flatness). Interference filters can be designed to produce sharp low-pass, high-pass, or bandpass type filters suitable for combining closely spaced wavelengths, as well as broader peaks suitable for lowering the gain in particular signal wavelength regions to produce gain flattened amplifiers.

Signal reflections can cause an amplifier to act as a laser, and this detrimental effect is eliminated in EDFAs by using optical isolators. In an isolator the signal light is coupled out of the single-mode fiber through a graded index (GRIN) lens and passes through a nonlinear crystal before being coupled back into the optical fiber. The isolator consists of a birefringent rutile (TiO_2) or Yttrium OrthoVanadate (YVO_4) wedge, followed by a Yttrium Iron Garnet (YIG) Faraday rotator, followed by another birefringent wedge. The YIG crystal is surrounded by a permanent magnet that rotates the light's polarization by 45 degrees. The 45-degree polarization rotation, coupled with the two birefringent wedges, ensures that light is efficiently coupled to the output fiber but not in the reverse direction. Commercial isolators are available with low insertion losses across the signal band, with some samples below 0.35 dB. Note that the YIG crystal works well for the 1480 nm pump and 1530–1620 nm signal bands, but currently there is no suitable isolator material that covers 980 to 1550 nm and this puts some limitations on certain EDFA designs.

The isolator design has been extended to make multiport circulators. A three-port circulator has the input into port 1 and output of port 2, light entering port 2 is directed to port 3, and light entering port 3 is blocked with an isolator. The circulator allows for adding and dropping of individual channels when a

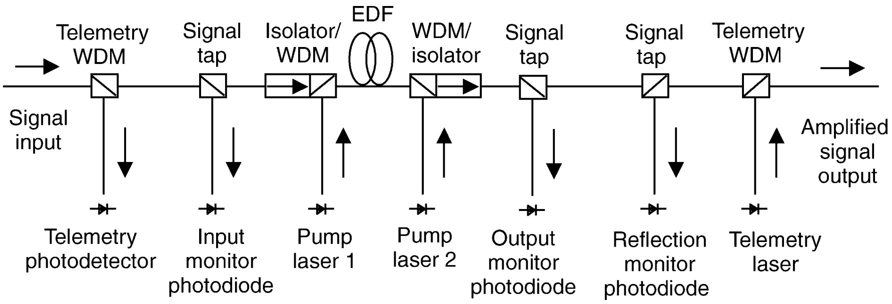

Figure 1 Single-stage EDFA with features.

narrow bandwidth reflective filter is placed on port 2. Circulators also can be used to separate co-propagating and counter-propagating traffic on a single transmission fiber before amplification is done.

Erbium-doped fiber has two strong absorption bands around 980 nm and 1480 nm that are suitable for achieving a population inversion in the erbium ion. The 980 nm wavelength allows low noise amplification, while 1480 nm lasers provide higher EDFA output. 980 nm pump lasers usually incorporate fiber Bragg gratings to stabilize the laser wavelength to match the narrow EDF absorption peak, and also have lower drive current requirements. A 1480 nm pump can provide more amplification, since there are more photons in each mW of laser power at 1480 nm than 980 nm. High-output power pump lasers incorporate thermo-electric coolers (TECs) within the pump laser package and can provide fiber coupled power greater than 500 mW. Both high-power lasers have been qualified to meet the most stringent reliability standards, with typical mean time before failure beyond 25 years. Low-cost 980 nm pump lasers, that do not have TECs, are also available in smaller packages and can supply up to ~200 mW. By using wavelength division or polarization combining techniques, it is possible to further increase the available pump power in the erbium fiber.

The EDFA module is assembled into a package that may contain passive optical components, several meters of coiled EDF, pump lasers, photo-detectors, and electronic circuit boards. In some cases, it is advantageous to thermally stabilize the components so that the amplifier performance can be maintained when exposed to extreme environmental conditions. This may occur when the central office's environmental control is compromised (e.g., air conditioning failure). In particular, EDF can exhibit undesirable spectral gain changes if the ambient temperature changes by more than 5 °C.

Pump lasers with internal TECs can dissipate more than 5 Watts of heat per laser and since cooling fans usually do not have the required reliability, passive cooling is commonly used in the module in the form of a metal heat sink. The amplifier module's size can be compact, limited by the height of optical components or pump laser diodes, or by the size of a built-in heat sink. The module is designed to survive conditions of electrostatic discharge, humidity, temperature, thermal shock, extreme vibration, and other stresses that may be inadvertently present during operation in the field. In addition, all material in the EDFA module is also qualified against problems with out-gassing (e.g., hydrogen release), combustion and chemical or biological exposure.

The Single-Stage Amplifier

A simple single-stage EDFA consists of an erbium-doped fiber spool with signal and pump combining multiplexer. The fiber is optically pumped by 980 nm and/or 1480 nm laser(s), whose light is coupled into the signal fiber by a passive component called an optical multiplexer. The pump wavelengths are readily absorbed by erbium ions embedded in silica raising them to an excited state. Amplification occurs when, stimulated by a nearby signal photon, an excited erbium ion relaxes back to the ground state producing a second, identical signal photon. The erbium ion can be approximated as a three-level atomic system that can be completely inverted by a 980 nm photon, to provide the lowest noise amplification. In contrast, the 1480 nm pump will excite the erbium ion directly into the upper laser level as a two-level system, and because of rapid spontaneous emission from this level, the maximum inversion in this case cannot exceed approximately 75%. Note that as the pump photons are absorbed, the inversion will be nonuniform along the EDF length.

There are two signal wavelength regions commonly amplified by EDFAs, the C-band (conventional band) from approximately 1528 to 1565 nm, and the L-band (long band) from approximately 1570 to 1620 nm. Amplification in the C-band readily occurs when moderate pump power is available, and relies on the erbium ion's spectral absorption and emission wavelength window that is suited to high levels of atomic inversion. L-band amplification is also achieved with moderate pump powers, but because of the lower absorption and emission cross-sections, similar gain is reached using approximately five times more EDF with a lower average inversion. The C-band amplifier is typically less costly because less EDF is used, while high-concentration erbium fibers are available specifically for L-band EDFAs.

An EDFA's most critical performance parameters are its amplified signal output power (typically stated in dBm) and its noise figure (stated in dB). Output power is mainly determined by total pump power and the amplifier internal loss. The noise figure (NF) is defined as the ratio of the signal-to-noise ratio at the input to the signal-to-noise ratio at the output.

A single-stage amplifier typically has 1 or 2 pump lasers but can have more if polarization- or wavelength-pump-combining is implemented for higher power. When the pump radiation propagates in the same direction as the signal, the amplifier is co-pumped, while counter-pumping denotes the case when the pump laser propagates against the signal. For a single pump, a co-pumping 980 nm laser minimizes the NF (suitable for a preamplifier) while counter-pumped 1480 nm architecture optimizes output power at some expense to the NF (suitable for a booster amplifier). Recent semiconductor pump laser improvements have enabled 980 nm pump lasers with output powers >500 mW to be commercially available, allowing most single- or dual-stage EDFAs to be energized by one laser.

Single-stage designs can be enhanced, as shown in Figure 1. To control the network an optical signal may be used as a telemetry or network supervisory channel, and this is removed with a filter. Telemetry wavelengths are usually outside of the useful EDF amplification window, and commonly range from 1500–1520 nm and from 1620–1640 nm. An isolator may be used at the input and/or output to prevent pump laser or amplified spontaneous emission from the erbium-doped fiber 'leaking' into the transmission path. Optical taps may be included to provide information about signal spectra at the input and output sides. Their feedback can be used to control pump laser biases for tuning output power, monitoring amplifier performance, or simply to trigger alarms. In addition, a reflection monitor is sometimes placed at the output to observe backwards propagating optical signals arising from reflections. Electronics in the module will continuously monitor the pump lasers and photodetectors and, for example, can place the module in an 'eye-safe' low-output power (<10 mW) mode within milliseconds if a transmission fiber break is detected through increased back-reflected optical power.

The amplifier has a signal input power of −30 to −10 dBm (1–100 μW) for each optical channel, and has a gain of 25 dB to compensate for a typical span loss of 100 km optical fiber link. This signal level allows high powers at the receiver photodetector for high-quality signal detection, yet is low enough to avoid nonlinear propagation effects in the transmission fiber. The typical total output signal power of an 80 channel (wavelength) C-band EDFA is less than 200 mW (+23 dBm) to avoid stimulated Raman scattering (SRS) in standard transmission fiber. The use of improved low nonlinearity transmission fibers and longer transmission distances can lead to specified total EDFA output powers to be greater than 400 mW (+26 dBm). To provide the best performance, the amplifier usually will allow only a single direction of propagation, with bi-directional communications systems using one transmission fiber and circulators to separate 'east' and 'west' traffic into individual EDFAs.

Figure 2 shows the results of a numerical simulation for a single-stage amplifier as a function of the input power and EDF length. The amplifier was assumed to have EDF with peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 100 mW at 980 nm. The amplifiers signal loss before and after the EDF was taken to be 0.8 and 1.2 dB, respectively. As amplifier input power increases there is a decrease in gain provided by the medium since there is a fixed amount of pump power available. The maximum gain is usually achieved near 1530 nm where the difference between the EDF's emission and absorption cross-sections is greatest. Selecting the length of EDF is critical to achieving the desired performance, and this can be examined using numerical simulation for a wide range of design options.

Undersea systems, with their long distances and large costs of network failure, place stringent design requirements on EDFAs. These systems mostly use single-stage designs with emphasis on low noise operation. This can be achieved by eliminating most of the optical components prior to the EDF, and also by co-propagating a strong 980 nm pump using a low loss fused fiber WDM. Undersea repeaters are spaced by 30 to 80 km, shorter than terrestrial networks, with each channel operated at higher power to achieve multi-thousand kilometer distances. For example, the trans-Pacific TPC-5J cable spans 8600 km from Coos Bay, Oregon (USA) to Ninomiya (Japan), using EDFAs spaced every 33 km. The tight electrical power budgets available to each repeater (powered from land) necessitate using pump lasers without TECs. During installation, the EDFA will experience large mechanical shock, as the cable and metal repeater bottles are unwound aboard ship and dropped into the ocean. Other design considerations are done for operation at the constant ambient temperature of the ocean bottom (~+2 to 4 °C) or for seasonal temperature changes on the continental shelves, where the optics and EDF will operate at 15 to 30 degrees above ambient due to heating from the electronics.

Single-stage optical amplifiers are suited for a wide range of applications such as single-channel amplifiers, simple WDM amplifiers, and low-cost amplifiers. Using high-power pump lasers or combined pump laser schemes allows such amplifiers to deliver output powers >20 dBm. However, single-stage amplifiers cannot meet the requirements of all telecommunications architectures, leading to increasing demand for multiple-stage EDFAs that are discussed below.

Multiple-Stage EDFAs

The most common implementation of a multiple-stage EDFA is to improve the noise and output power

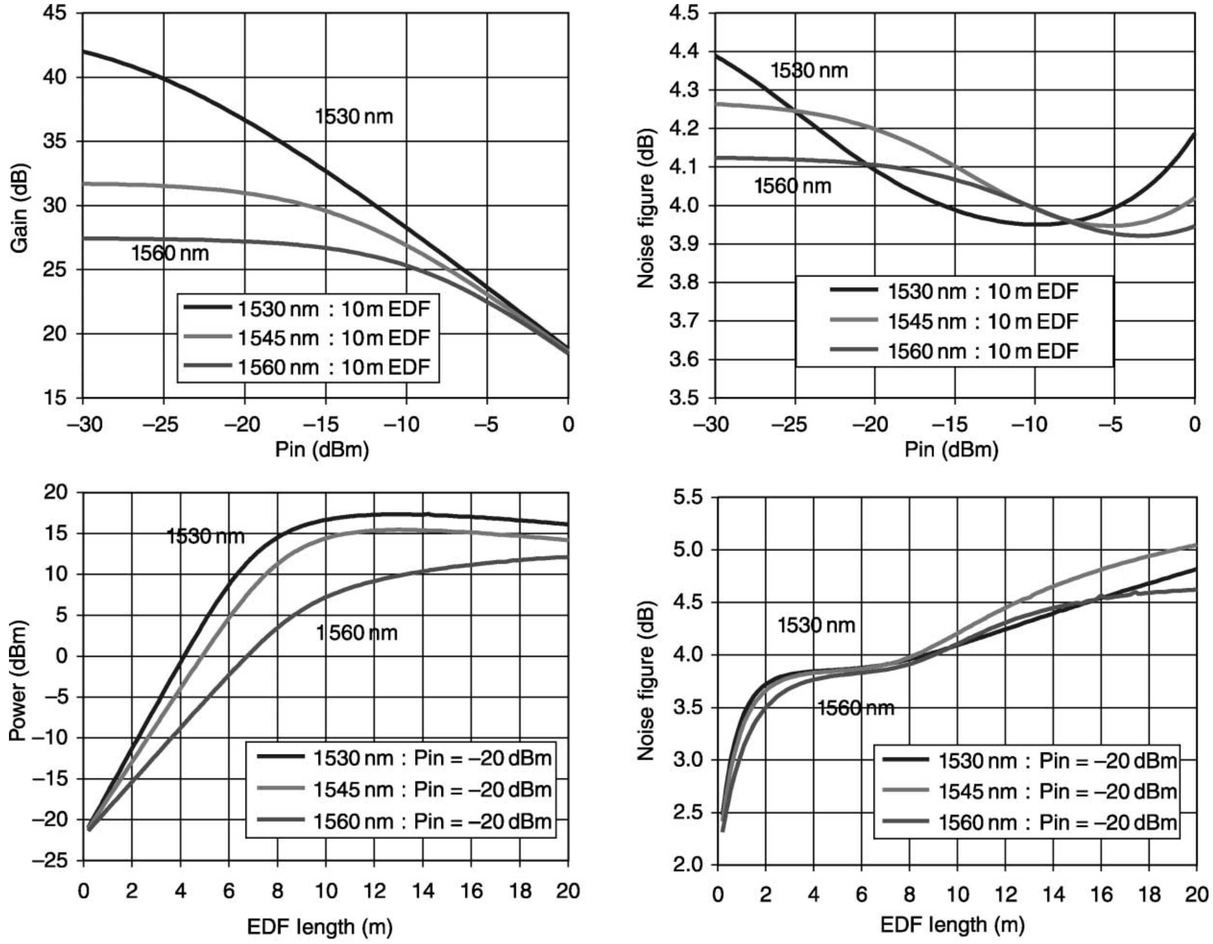

Figure 2 Typical single-channel performance of a single-stage EDFA.

characteristics by imbedding an optical isolator between two sections of EDF. The isolator blocks backward-traveling amplified spontaneous emission to improve the efficiency of the amplifier. This is shown in Figure 3 for two common single-pump designs. The first is the 'pump by-pass', design where the residual pump radiation from the first stage is redirected around the mid-stage isolator. This is often used when pumping with 980 nm since signal band optical isolators will not transmit 980 nm radiation. This design can introduce problems when using low-quality components as small signal levels can propagate around the pump by-pass fiber to create multiple path interference (MPI) effects. Although MPI effects can be eliminated by using a pump splitting coupler to directly pump each EDF section, the pump by-pass design makes the most efficient use of available pump power. The second design is called the 'pump through' design, and is used when pumping at 1480 nm since both the pump and signal band (e.g., 1550 nm) photons will transmit through commercial isolators with only small loss. Note that for cost and space constraints within the module, it is sometimes advantageous to use hybrid optical components, e.g., an isolator and WDM can be combined into one compact package.

An EDFA with multiple input signals will have a very nonuniform output gain profile. This is a consequence of the erbium ion's wavelength-dependent emission and absorption cross-sections in the host glass material. Figure 4 shows gain spectra for both C-band and L-band EDFAs with and without gain-flattening filters (GFFs). The gain spectrum is dependent on the EDF's chemical composition and EDFA design features. For an amplifier with 25 dB gain, a GFF with peak loss less than 10 dB is usually needed to correct for these gain deviations. Although many technologies are available for GFFs (e.g., thin-filter interference filters, Bragg gratings, tapered fiber filters, etc.) the basis function of these technologies is usually not identical to the EDF's gain profile, and this mismatch results in gain

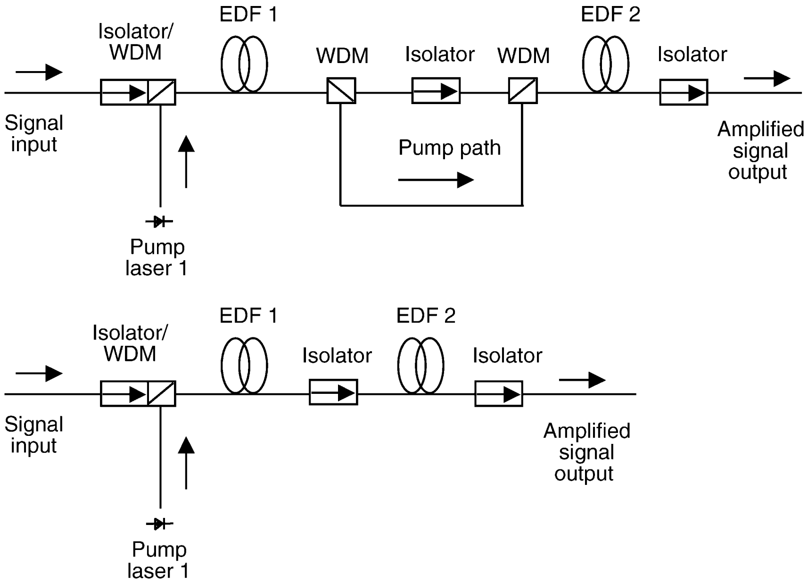
EDF 1
Isolator/ WDM
WDM
Isolator
WDM
EDF 2
Isolator
Signal input
Pump path
Amplified signal output
Pump laser 1
EDF 1
Isolator/ WDM
Isolator
EDF 2
Isolator
Signal input
Amplified signal output
Pump laser 1

Figure 3 Imbedded Isolator EDFA designs.

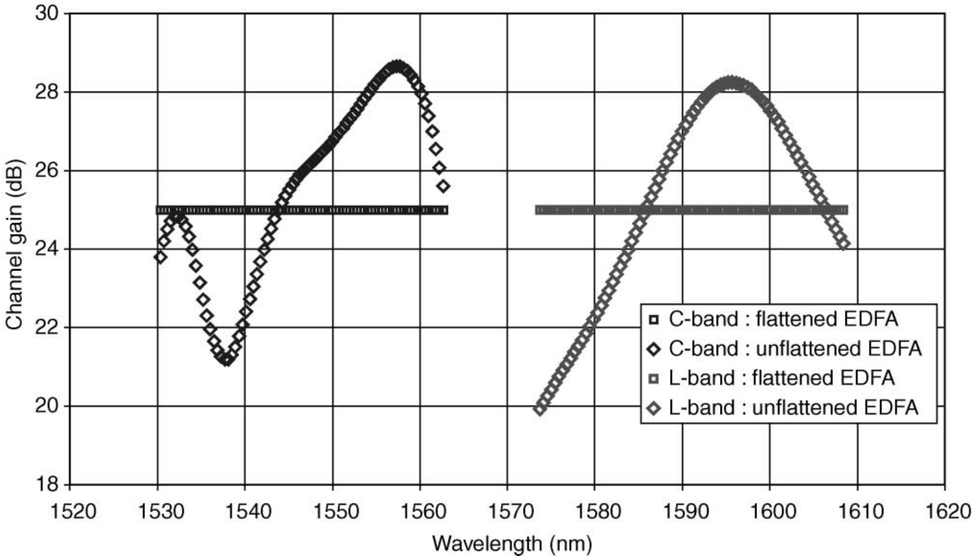

Figure 4 Gain spectrum of a multi-staged optical amplifier with and without gain flattening filters.

flatness error. However, when using these technologies, manufactured EDFAs can have a gain error (flatness) less than 1.0 dB over bandwidths greater than 35 nm.

Early EDFAs for WDM systems used 16 optical channels spaced by 200 GHz in the 1540 to 1560 nm part of the C-band. In this case, the EDF gain profile is relatively smooth and no GFFs are needed to achieve a 1.0 dB gain flatness specification. It is important to note that for a particular level of signal and pump power, a longer length of EDF in the EDFA will produce more gain at the long wavelength end of the spectrum, hence by shortening the EDF length in Figure 4 the result will be relatively flat gain from approximately 1540 to 1560 nm. As EDFA technology has matured, the EDF can be the bandwidth-limiting component in a system, and it has become necessary to use GFFs to realize useable bandwidths of up to 55 nm when using regular silica-based EDF. The most economic EDFAs operate in the C-band where relative to the L-band, pump laser efficiency is highest, EDF lengths are shortest, and the chromatic dispersion of most installed transmission fiber limits nonlinear optical effects.

In single-stage amplifiers, the gain-flattening filter is at the amplifier output, and this results in lost output power. Wideband gain-flattened amplifiers usually have a multiple EDF stage design, with the GFF component between two EDF stages. The EDF sections that follow the attenuating GFF provide additional amplification, to give high power out of the EDFA module that has large internal losses.

In addition to GFF components inside the EDFA, an attenuator can be used to compensate for input signal power changes, which will produce a spectral gain rotation or tilt. The attenuator reduces the power on all optical channels equally and allows the amplifier to operate in a 'fixed gain' mode. In WDM systems, optical amplifiers may also need mid-stage access where the signal is fed into an external device between the amplifier stages. Reasons for doing this include monitoring, adding or dropping of individual channels, and dispersion compensation. Since additional losses up to 10 dB are introduced in the amplification path, the amplifier design has to be optimized for those losses.

Typical two-stage gain-flattened optical amplifier architecture is shown in Figure 5. Input and output couplers and telemetry WDMs can be included if required. A single pump can be split and shared between stages to save cost. When multiple pumps are needed, the most common configuration is a 980 nm pump laser co-pumping the first stage (EDF1) for low noise and a 1480 nm laser counter-pumping the second spool (EDF2) for high gain. A significant loss element, e.g., a gain-flattening filter, add/drop module or dispersion compensation module, is situated between the stages. Note that mid-stage access can be located before or after the gain flattening or between additional EDF stages. The gain spectrum of multistage gain-flattened EDFAs is shown in Figure 6, showing the gain equalization possible using a multistage amplifier design and gain-flattening filters. Using deeper GFFs can increase the usable bandwidth of the amplifiers, but this requires higher-power pump lasers to maintain the same gain and optical signal-to-noise ratio (OSNR) performance.

The two-stage EDFA shown in Figure 5 highlights a common problem for amplifier design. Given high-grade optical components and pump lasers, what

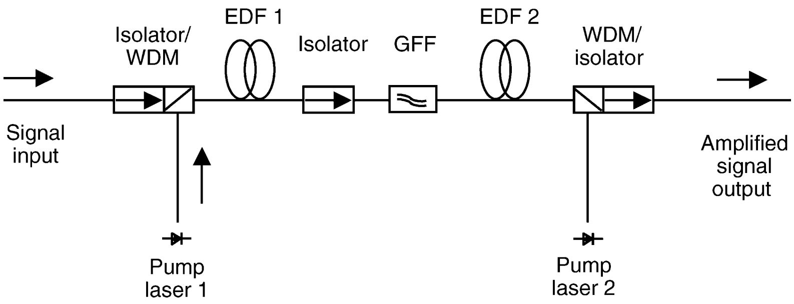

Figure 5 Typical multistage gain-flattened optical amplifier architecture.

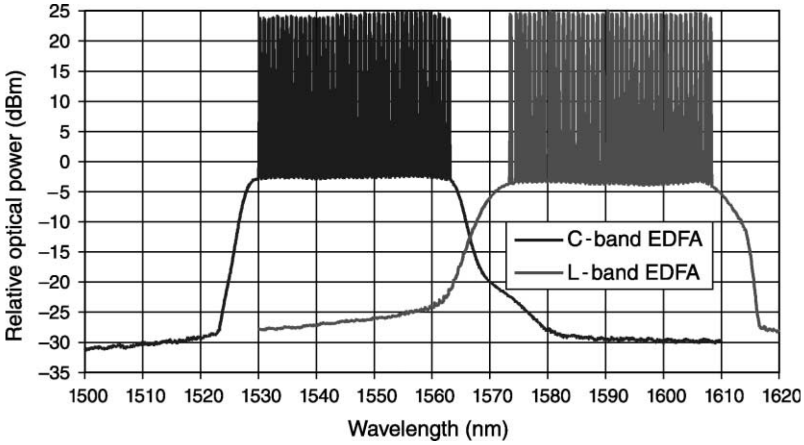

Figure 6 Measured gain and noise spectra of multistaged gain-flattened C-band and L-band EDFAs.

length of erbium-doped fiber should be used to give the best amplifier performance? This question is best answered by using the results of extensive numerical simulations of the optical amplifier.

Figure 7 shows the results of numerical simulation for a two-stage amplifier as a function of the first and second EDF stage lengths. The amplifier was assumed to have EDF with a peak absorption of 5.5 dB/m near 1530 nm, and was pumped with 130 mW at 980 nm in the first stage and 160 mW at 1480 nm in the second stage. The total input power was assumed to be −2.5 dBm for 80 channels distributed from 1530 nm to 1563 nm. Note that each of the 80 channels is separated by 100 GHz, and will generally support a long-distance communications system with each wavelength modulated at 10 GHz.

Figure 7a shows that the optical output power is largest when both EDF stages are approximately 20 meters. For this design, EDF lengths significantly longer than 20 meters will produce less output power, since the pump radiation will have been completely absorbed in the first few meters of EDF. The spectral gain flatness is shown in Figure 7b. For this design, approximately 17 meters of EDF was needed to produce a flat gain spectrum, i.e., each optical channel had the same gain. Other combinations of EDF lengths were not matched to the particular gain-flattening filter and produced large gain variations across the band. In general as the amplifier's total EDF length increases the gain spectrum exhibits a positive tilt (e.g., longer wavelength channels will experience more gain). Figures 7c, d show the average channel and maximum channel NF, respectively. From a system design perspective, a link can be limited by the optical channel with the lowest optical signal-to-noise ratio (OSNR) or highest noise figure, and this is a critical parameter of interest when designing optical amplifiers. In practice the best amplifier design is a compromise between high-gain, low-gain flatness, and low NF. From extensive numerical simulations, for the two-stage gain flattened amplifier example it was for

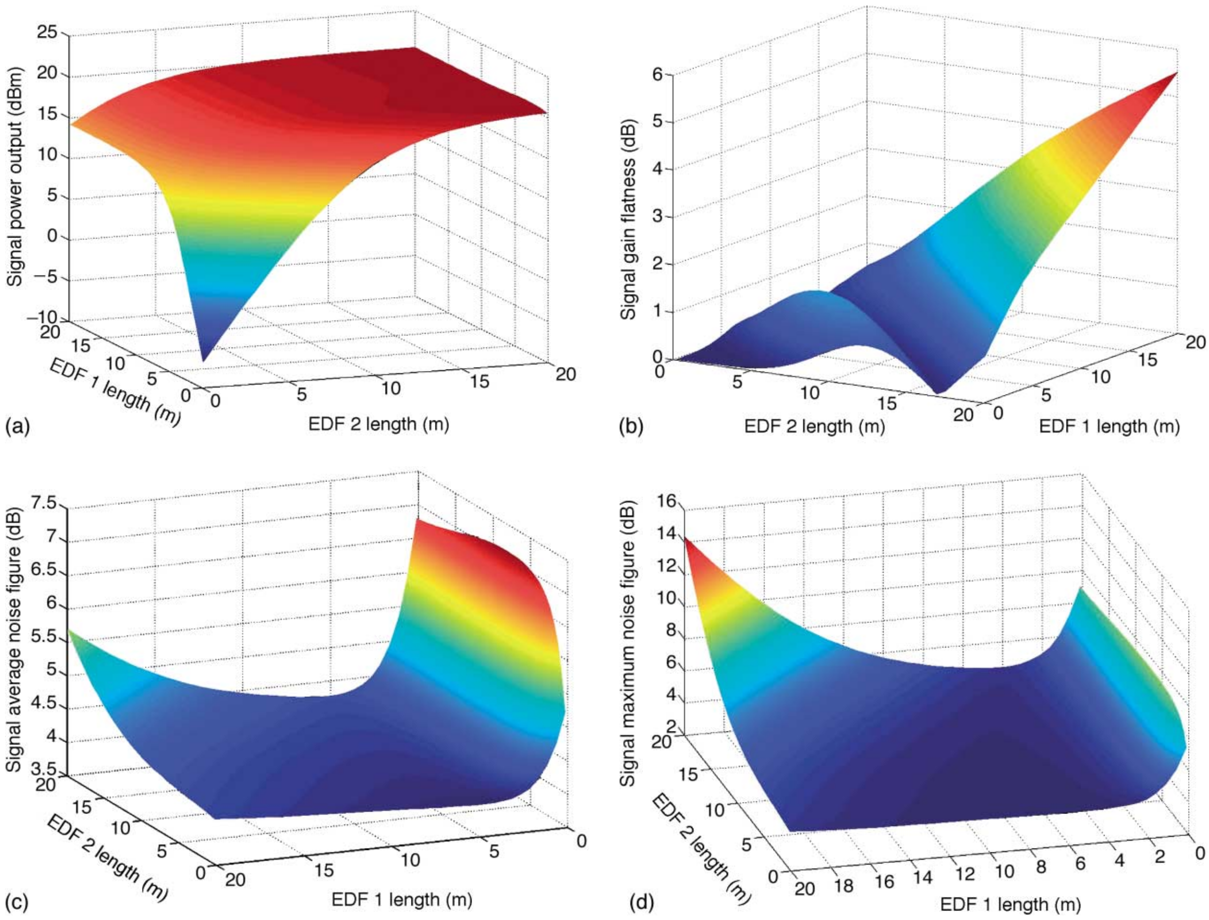

Figure 7 (a) Total output power, (b) spectral flatness, (c) average noise figure, and (d) maximum channel noise figure of a two-stage optical amplifier with a gain-flattening filter.

approximately 4 meters in EDF stage 1, and 13 meters in stage 2.

Successful commercial EDFA products are more than a single amplifier design that can accommodate all applications. In some cases an economical solution is a modular design approach, where smaller capacity, lower-cost amplifiers are initially installed in a network. Additional gain can be added to the basic amplifier for longer-distance spans or as network traffic increases over time. This ‘pay as you grow’ approach can be done using additional gain stages, or pump lasers, and can be upgraded without interruption of revenue-generating network traffic.

Low-Noise Design of EDFAs

A key design feature of commercial EDFAs is low-noise operation since the degradation of the OSNR limits the reach of the system. The NF of an EDFA can be defined as

$$\mathrm{NF} = \mathrm{SNR}_{\mathrm{in}}/\mathrm{SNR}_{\mathrm{out}} \qquad [1]$$

The NF can also be calculated from

$$\mathrm{NF} = \frac{1}{G} \times \left[\frac{P_{\mathrm{ASE}}(\nu_{\mathrm{s}}, L)}{h\nu_{\mathrm{s}}} + 1 \right] \qquad [2]$$

where P_{ASE} is the amplified spontaneous emission (ASE) noise power at the signal frequency ν_{s} for an EDFA of gain G (h = Plancks constant). An alternative equation for the NF for a single section of EDF is

$$\mathrm{NF} = \int_0^L \frac{\gamma_{\mathrm{e}}(\nu_{\mathrm{s}}, z) - \gamma_{\mathrm{a}}(\nu_{\mathrm{s}}, z)}{P_{\mathrm{s}}(z)} \times P_{\mathrm{s}}(0)\mathrm{d}z + 1 \qquad [3]$$

where γ_{e} and γ_{a} are the EDF emission and absorption factors at the signal frequency ν_{s}, at distance z along the amplifying fiber. Equation [3] shows that low NF can result from rapid signal gain in the initial fiber along the EDF. This result indicates that co-propagating pump and signal photons, both into the same end of the EDF, combined with the maximum inversion obtained by using a 980 nm pump, will result in low-noise amplification. The NF is typically expressed in dB units, and using a full quantum mechanical theory, the lowest possible NF for a fully inverted EDF with a large signal gain can be shown to be approximately 3 dB. Actual EDFAs will only approach the ‘3 dB quantum noise limit’ when the signal is significantly lower than the available pump power (e.g., low channel count applications) so that high inversion can be maintained along the EDF length. Optical components in the EDFA prior to the first EDF will attenuate the signal but not the noise, since ASE is only generated in the EDF, and this input stage loss will directly add to the EDFA’s NF.

The use of 1480 nm pump lasers can also result in a further NF penalty of 0.3 to 1.5 dB since complete inversion cannot be achieved when the erbium ions are operating as a two-level atomic system and more ASE is generated. The ASE, like the signal gain, has a wavelength-dependent spectral profile and this results in each channel having a different NF. Typically, the optical channel with the largest NF (lowest OSNR) is used to define the EDFA’s noise performance. EDFAs with low-input signal power (e.g., −15 dBm) and high gain (e.g., 25 dB) may be described as low noise when the NF is <4 dB. For large total input signal power (e.g., +2.5 dBm for high channel count operation) low noise may be NF < 5 dB.

For an amplifier with multiple EDF sections, the noise figure can be calculated from

$$\mathrm{NF}_{\mathrm{System}} = \frac{\mathrm{NF}_1}{L_1} + \frac{\mathrm{NF}_2}{L_1 G_1 L_2} + \ldots + \frac{\mathrm{NF}_N}{L_1 G_1 L_2 G_2 \ldots L_{N-1} G_{N-1}} \qquad [4]$$

where L_i is the signal loss prior to the gain G_i of the EDF stage i. From eqn [4], multistage amplifiers with large inter-stage losses will have numerically larger NFs, but the total module NF is dominated by the first EDF stage. Undersea EDFAs, with a single short EDF section, very small L_i and 980 nm pumping, can have NF <3.5 dB.

Advanced EDFA Functions

An important trend in optical amplifiers, and systems in general, is the move towards dynamic control. Several technologies are being considered and the primary idea is to dynamically control the intensity level of each optical channel or wavelength to correct for any gain/loss inequalities in the network. This is particularly important as channel counts and bit rates rise concurrently with the functionality demanded of optical communication networks. A common requirement in this category is fast amplifier response so that in a network where individual optical channels can be added or dropped there are no power distortions at other surviving wavelengths. In a typical drop event, in less than 1 microsecond, 31 of 32 wavelengths can be completely removed from the input to the EDFA by an add/drop module located elsewhere in the network. This event will cause a total input power decrease of 15 dB and without rapid pump power adjustment the remaining channel will be excessively amplified.

Erbium ions have an excited state lifetime of approximately 10 ms, and this sets the time-scale for the transient signal response. A network transient usually manifests as a sharp optical amplifier (OA) gain fluctuation lasting up to several tens of microseconds that will pulse large signal levels through the entire optical network. Dynamic network loading necessitates controlling temporal gain transients, and this may be achieved by using a microprocessor in the amplifier module. These intelligent EDFA modules can rapidly adjust the pump laser power based on feedback from internal photo detectors to control gain transients.

The need for dynamic control also exists in the frequency domain. Wideband amplifiers have a gain variation among the optical channels of typically up to 1 dB over all operating conditions, and since these EDFAs are concatenated over many spans the electronic receiver circuit at the end of the network will have to detect optical channels with widely varying power levels. This can limit the number of spans that the network can bridge before electronic regeneration becomes necessary. This problem can be corrected using a dynamic gain equalizing filter (DGEF) to actively attenuate individual optical channels levels. A DGEF can be embedded inside an amplifier to create a loss-less component, and can be used to equalize channel powers for an entire network.

The problems of accumulation of gain ripple become more apparent in systems that employ amplifiers based on stimulated Raman scattering. A conventional EDFA amplified system may have 10 spans transmitting 100 optical channels at 10 Gbit/s each over a distance of approximately 800 km before electronic regeneration becomes necessary. To span longer distances with higher capacity, Raman amplifiers are used to amplify the transmission fiber along with EDFAs. This system architecture has the advantage of a higher effective OSNR for the individual spans, and has been demonstrated for distances over 4000 km with 40 Gbit/s channels. However, the variability of the Raman amplifier gain, due to network loading and differing transmission fiber properties, necessitates the use of loss-less DGEFs.

High-capacity transmission uses individual channels that are each modulated up to typical 2.5, 10, or 40 GHz, that requires that the optical amplifiers have low polarization mode dispersion (PMD) and low chromatic dispersion (CD). Polarization mode dispersion distorts the temporal spread of an ultra-short pulse as different polarization states travel at different speeds through the transmission fiber and amplifier components. PMD is particularly a problem in the first generation of installed transmission fiber and through polarization components (e.g., isolators) that are not PMD compensated. Without PMD compensation, 40 Gbit/s transmission, for example, can be limited to very short distances (several km). Fortunately, PMD compensators have been developed for upgrading aged installed networks, and also amplifiers are available with very low PMD.

To achieve long-distance transmission, a chromatic dispersion map is produced for all the optical spans, and by using a dispersion compensating module in each amplifier, the net dispersion across the channel wavelength window is controlled. Although optical amplifiers have approximately 1000 times less optical fiber in them compared to the transmission span that it is amplifying, chromatic dispersion in the amplifier can also be of concern for very long-haul networks. The CD of the amplifier is primarily in the EDF, and is specific to the EDF's geometric and chemical composition. Furthermore, since the erbium ion is a resonant atomic system, there may be a pump and signal power contribution to the dispersion (resonant dispersion) that could degrade network performance. Characterization of these optical effects is typically done on manufactured 'field grade' amplifiers as part of a quality control process.

Conclusion

As optical networks evolve EDFAs will continue to be an enabling technology for higher capacity and more dynamic communications networks. EDFA technology has already advanced to accommodate multiple channels, to span several wavelength windows and to provide features such as dispersion compensation, gain-transient suppression, and dynamic gain flatness. As future network architectures are introduced, to incorporate Raman amplification and additional wavelength windows, the demands placed upon EDFA design will continue to expand.

See also

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing. **Scattering:** Raman Scattering.

Further Reading

Agrawal G (1989) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Agrawal G (1997) *Fiber Optic Communication Systems*. New York: John Wiley and Sons.

Becker PC, Olsson NA and Simpson JR (1999) *Erbium-doped Fiber Amplifiers – Fundamentals and Technology*. San Diego, CA: Academic Press.

Desurvire E (1995) *Erbium Doped Fiber Amplifiers*. New York: John Wiley and Sons.

Pedersen B, Bjarklev A, Povlsen JH, *et al.* (1991) The design of erbium-doped fiber amplifiers. *Journal of Lightwave Technology* 9(9): 1105.

Sudo S (1997) *Optical Fiber Amplifiers: Materials, Devices and Applications*. Boston, MA: Artech House.

Optical Amplifiers in Long-Haul Transmission Systems

B M Desthieux, Xtera Communications, Inc., Allen, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical amplifiers have been a key element for enabling the development of optical transmission systems, which have reached total capacity in the order of a few terabits per second. The first type of optical amplifier to see application in the telecom industry has been the erbium-doped fiber amplifier. Following its invention in the late 1980s, the mid-1990s saw its industrial application in submarine and terrestrial transmission systems.

Since then, optical transmission systems have been the predominant source for transmitting information over distances ranging from a few 100 km to more than 10 000 km, connecting cities, countries, and continents.

This article will describe the applications of optical-fiber amplifiers in long-haul transmission systems, focusing on erbium-doped fiber amplifiers and Raman amplifiers, the most popular type of optical amplifiers used in modern transmission optical systems.

Fundamentals of Long-Haul Transmission Systems

This section explains the key elements and parameters of an optical transmission system.

Key Elements

An optical transmission system consists of a transmitter, a receiver, one or more optical amplifiers, and one or more spans of transmission fiber. Figure 1 shows a simplified schematic of a bidirectional transmission system using a fiber pair.

Optical amplifiers are used to compensate for the loss of the transmission fiber and the other optical elements placed along the signal path. Boosters and pre-amplifiers refer to optical amplifiers which are located at, respectively, the input end and output end of the system. In-line amplifiers refer to optical amplifiers located between the transmission fiber spans.

The transmitter includes laser diodes to generate the signal power and optical modulators to transfer the data stream from the electrical input signals to optical signals. The optical signal is transmitted to the receiver after traveling through the transmission fiber. At the receiver are located the photodiodes, which transfer back the information into electrical output signals. In order to increase the total bandwidth of the transmission system, time-division multiplexing (TDM) and wavelength-division multiplexing (WDM) can be used. TDM techniques combine the electrical domain data stream of several electrical channels, which is carried by a single optical channel. An optical channel refers to an optical signal located at a particular wavelength and modulated with the transmission data. TDM increases, therefore, the bit rate supported by the optical channel. Typical bit rates used in modern optical transmission systems are 2.5 Gbit/s and 10 Gbit/s. With WDM, several optical channels are combined, in the optical domain, into a single optical beam, which is launched into the transmission fiber. For this technique, optical multiplexers (Mux) and de-multiplexers (Demux) are required at, respectively, the transmitter and the receiver.

Long-haul transmission systems typically operate at wavelengths around 1.5 μm, where the attenuation of fibers is the lowest. We label C-band the region located between 1530 nm and 1565 nm. Its lower bound is labeled S-band; the upper bound the L-band.

Apart from attenuation, several other effects existing in fiber can be responsible for degrading the quality of the transmission. Among those is chromatic dispersion, which distorts the optical signal as it travels along the transmission fiber. Several types of fiber have been deployed with different chromatic dispersion characteristics. The most common one is the standard single-mode fiber (SSMF), which has a zero-dispersion wavelength – wavelength for which no chromatic dispersion is observed – near 1.3 μm. The dispersion-shifted fiber (DSF) has a zero-dispersion wavelength near 1.55 μm, and the nonzero

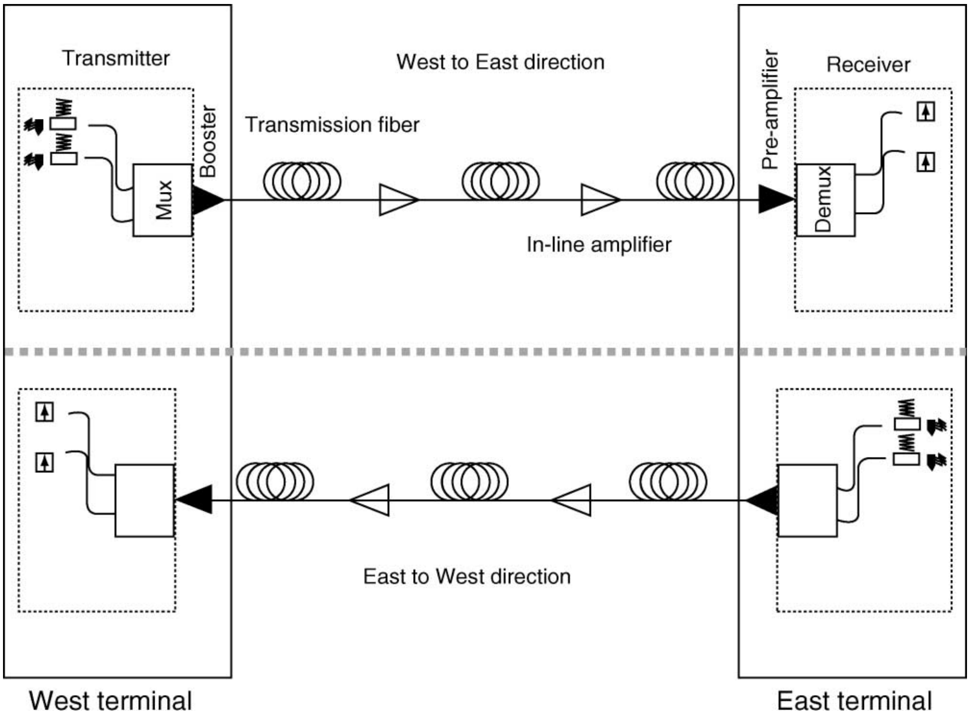

Figure 1 Schematic of a bidirectional long-haul transmission system.

dispersion shifted fiber (NZDSF) has a zero-dispersion wavelength close to 1.5 μm. To remedy this effect, dispersion compensating modules (DCM) are placed along the link. These modules – generally a special type of fiber – have chromatic dispersion with inverted characteristics with respect to those of the transmission fiber.

Key Optical-Amplifier Parameters

Optical amplifiers spontaneously generate optical noise – referred to as amplified spontaneous emission (ASE) – that is added to the modulated signal and consequently degrades the quality of the transmission. It is essential to quantify the level of ASE for a given optical transmission system.

We define the optical signal-to-noise ratio (OSNR) in the bandwidth $\Delta\nu$ as

$$\mathrm{OSNR} = \frac{P_{\mathrm{Signal}}}{P_{\mathrm{ASE}}} \qquad [1]$$

where P_{signal} is the signal power and P_{ASE} is the ASE power in the bandwidth $\Delta\nu$.

The required OSNR at a receiver, to detect the information with high quality, depends on the modulation format and on the other sources of noise or signal degradation existing in the transmission system. If we assume a nonreturn to zero (NRZ) modulation format, and signal detection limited only by ASE noise, an OSNR, in 0.1 nm bandwidth, of about 18 dB, is required at the receiver for a bit-error rate – number of erroneously transmitted bits divided by number of transmitted bits – of 10^{-12}. Introduction of other transmission degradation sources requires a higher OSNR to maintain the same bit-error rate.

The decibel unit (dB) for a given ratio, R, is defined by

$$R_{\mathrm{dB}} = 10\,\log_{10} R_{\mathrm{linear}} \qquad [2]$$

Similarly we defined the power decibel unit (dBm) as

$$P_{\mathrm{dBm}} = 10\,\log_{10} P_{\mathrm{mW}} \qquad [3]$$

where P_{mW} is the power in mW.

In order to quantify the amount of ASE generated by an optical amplifier, we define the equivalent input noise factor (n_{eq}), in a linear unit, by

$$n_{\mathrm{eq}} = \frac{P_{\mathrm{ASE}}}{h\nu\Delta\nu G} \qquad [4]$$

where P_{ASE} is the output ASE power in one polarization state in the bandwidth $\Delta\nu$, h the Planck's constant, ν the frequency of the ASE, and G the gain of the amplifier.

We also define the noise figure of the amplifier (NF) – a universal parameter to characterize amplifier noise – as

$$\mathrm{NF} = 2n_{\mathrm{eq}} + \frac{1}{G} \qquad [5]$$

If an optical component, having a loss T_{in} ($T_{\mathrm{in}} < 1$), is inserted at the input of the amplifier, the total NF of the amplifier becomes $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF}/T_{\mathrm{in}}$. If an optical component, having a loss T_{out} ($T_{\mathrm{out}} < 1$), is inserted at the output of the amplifier, $\mathrm{NF}_{\mathrm{total}} = \mathrm{NF} + 1/G(1/T_{\mathrm{out}} - 1)$. When $G \gg 1$, we have $\mathrm{NF}_{\mathrm{total}} \approx \mathrm{NF}$.

The OSNR in 0.1 nm bandwidth obtained after N fiber spans of loss, $\mathrm{Loss}_{\mathrm{Fiber}}$ (in dB), each having an amplifier, is in decibels:

$$\mathrm{OSNR} = P_{\mathrm{Signal}} - \mathrm{Loss}_{\mathrm{Fiber}} - \mathrm{NF} - 10\log_{10} N + 58 \qquad [6]$$

where P_{Signal} is the signal power at the output of the amplifier.

When the optical signal power is sufficiently high, nonlinear effects also become responsible for the degradation of the system performance. Therefore, choosing the value of the signal power launched into the transmission fiber results in a trade-off between increasing the OSNR, and reducing nonlinear effects.

Optical Fiber Amplifier

An optical fiber amplifier consists of fiber with an active medium and at least one optical pump source, which is launched simultaneously with the signal into the fiber by means of a dichroic coupler.

We define as forward and backward pumping, the case where the pump is launched, respectively, in the same direction as the signal and in the opposite direction to the signal. We define as bidirectional pumping, the case where the fiber is pumped from both ends. The most commonly used pump sources are the semiconductor laser diodes. These lasers may use different technologies to meet the various performance requirements with respect to the amplifier types and applications. The most popular pump diodes are the distributed Bragg reflector (DBR), distributed feedback (DFB), or external-grating stabilized lasers.

Erbium Fiber Amplifier

Fiber doped with the rare-earth Er^{3+} ion can amplify a signal located at 1.5 μm with the appropriate fiber host material and if pumped near 0.98 μm or near 1.48 μm. Silica is the most common material used for the fiber host. Er^{3+} has three energy levels, as described in Figure 2.

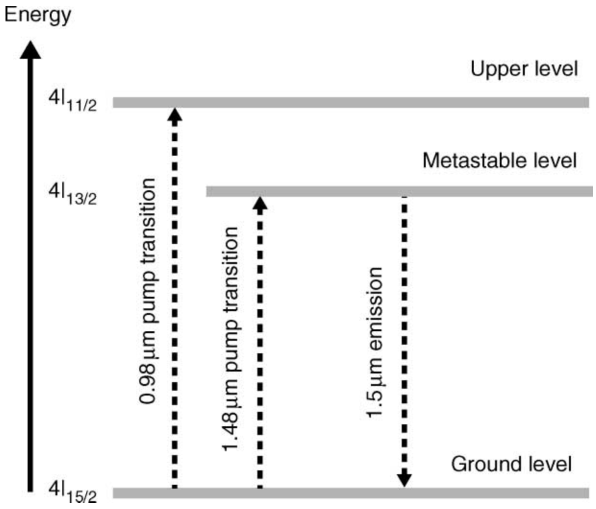

Figure 2 Energy level diagram for Er^{3+} ions.

0.98 μm pumping occurs between the upper level $^4\mathrm{I}_{11/2}$ and the ground level $^4\mathrm{I}_{15/2}$, and 1.48 μm pumping occurs between the metastable level $^4\mathrm{I}_{13/2}$ and the ground level.

The signal amplification takes place between the metastable level and the ground level. With 0.98 μm pumping, a rapid energy transfer occurs between the upper level and the metastable level via a phonon-assisted process. The lifetime of the upper level is much lower than the lifetime of the metastable level, about 10 ms. Therefore, with 0.98 μm pumping we can assume a two-level energy transfer, between the metastable and the ground levels, for the description of the amplification process in the Er^{3+} medium.

We define the normalized medium inversion (D) by

$$D = \frac{N_2 - N_1}{N_t} \qquad [7]$$

where N_2 is the population of the metastable level, N_1 the population of the ground level, and $N_t = N_2 + N_1$ the total population. $D = +1$ when the metastable level is fully populated and $D = -1$ when only the ground level is populated.

A simple approach for calculating the gain (G) with erbium-doped fiber amplifiers (EDFA) is given, in decibel units, by

$$G(\lambda) = \big((\alpha(\lambda) + g(\lambda)\big)\frac{D+1}{2} - \alpha(\lambda)L \qquad [8]$$

where λ is the wavelength of the signal, α the absorption coefficient in dB unit length at the signal wavelength, g the emission coefficient in dB unit length at the signal wavelength, and L the length of the erbium-doped fiber (EDF). Here D is the normalized path-average medium inversion of the EDF.

The gain per unit length, $G(\lambda)/L$, is equal to $g(\lambda)$ for $D = +1$ and to $-\alpha(\lambda)$ for $D = -1$.

The absorption and the emission coefficients depend on the EDF host material and the codoping element used for the realization of the EDF. With silicate EDF the most commonly used codoping elements are germanium, aluminum, and phosphorus.

In Figure 3 we show the dependence of the gain per unit length on the normalized path-average medium inversion, for a signal wavelength ranging from 1500 to 1560 nm.

For a more comprehensive description of the EDFA characteristics we use the standard confined-doping model. This model, which is also based on the two-level system, uses the following differential equations:

$$\frac{dP_S^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right)P_S^+(z) \qquad [9]$$

$$\frac{dP_{ASE}^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_S + g_S) - \alpha_s\right) \times P_{ASE}^+(z) + \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \qquad [10]$$

$$\frac{dP_{ASE}^-(z)}{dz} = \left(\alpha_s - \frac{N_2}{N_t}(\alpha_S + g_S)\right) \times P_{ASE}^-(z) - \frac{N_2}{N_t} g_S 2h\nu_S \Delta\nu \qquad [11]$$

$$\frac{dP_P^+(z)}{dz} = \left(\frac{N_2}{N_t}(\alpha_P + g_P) - \alpha_P\right)P_P^+(z) \qquad [12]$$

$$\frac{dP_P^-(z)}{dz} = \left(\alpha_P - \frac{N_2}{N_t}(\alpha_P + g_P)\right)P_P^-(z) \qquad [13]$$

where $P_S(z)$, $P_{ASE}(z)$, and $P_P(z)$ in W are, respectively, the signal power, the ASE power, and the pump power along the length (z-axis) of the EDF; $^+$ and $^-$ indicate, respectively, the backward-traveling and the forward-traveling directions; α_s and α_P in m^{-1} are the absorption coefficients at, respectively, the signal and pump wavelengths, g_s and g_p in m^{-1} are the emission coefficients at, respectively, the signal and pump wavelengths, ν_s and ν_p are the frequencies of, respectively, the signal and the pump beams; $\Delta\nu$ is the frequency bandwidth of the ASE; and h is the Planck's constant. N_2/N_1 is the medium inversion along the z-axis of the EDF, and is given by

$$\frac{N_2}{N_t} = \frac{\sum_k \frac{P_k(z)\alpha_k \tau}{h\nu_k B_{eff}^2 N_t}}{1 + \sum_k \frac{P_k(z)(\alpha_k + g_k)\tau}{h\nu_k B_{eff}^2 N_t}} \qquad [14]$$

where the index k represents all the beams, i.e., signal, ASE, and pump, traveling backward and forward along the EDF, τ is the lifetime of the metastable level, and B_{eff} is the effective erbium-doping area. For a typical silicate EDF, having a peak absorption coefficient of 1 dBm^{-1}, we have $\tau = 10$ ms, $B_{eff} = 12\ \mu m^2$, and $N_t = 5 \times 10^{23}\ m^{-3}$.

Using the standard confined-doping model we calculate the dependence of the EDFA noise figure on the normalized path-average medium inversion, as shown in Figure 4. We vary the path-average normalized medium inversion by changing the length of the EDFA, while adjusting the pump power so that the signal gain is kept constant.

The noise figure decreases when the medium inversion increases, and approaches the quantum limit value of 3 dB when the path-average normalized medium inversion is close to unity. Near full inversion

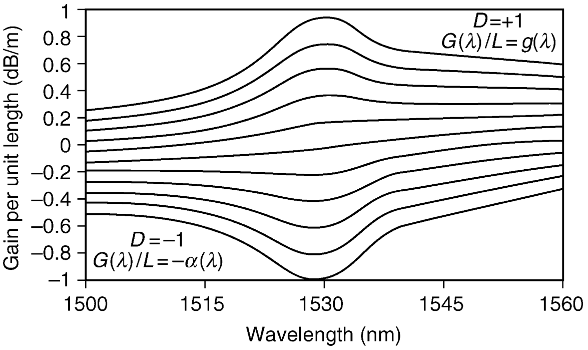

Figure 3 Dependence of the gain per unit length on the normalized path-average medium inversion, D. D ranging from −1 to +1 with a 0.2 step (alumino-silicate EDF; peak absorption of 1 dB m^{-1}; peak emission of 0.93 dB m^{-1}).

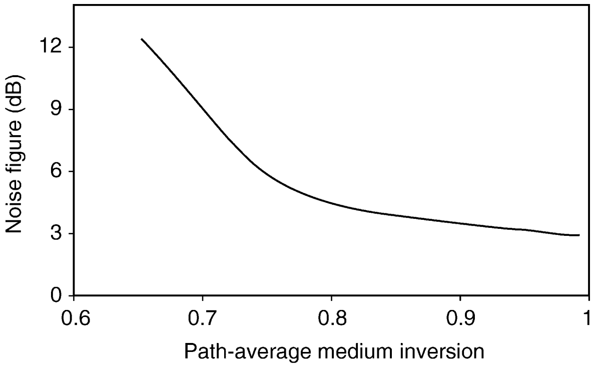

Figure 4 Dependence of the noise figure on the normalized medium inversion (alumino-silicate EDF with peak absorption of 1 dB m^{-1}; EDF lengths ranging from 16 m to 60 m; gain equals to 15 dB; input signal power of −40 dBm at 1530 nm; backward pumping at 980 nm).

may be obtained with a pump located at 0.98 μm because the emission coefficient is null at this pump wavelength. However, this is not the case at 1.48 μm where the ratio g_p/α_p is about 0.3. The maximum medium inversion that can be obtained with 1480 nm is $D = 0.8$. Therefore, higher noise figures are obtained with 1.48 μm pumping than with 0.98 μm pumping. The pumping direction also has an effect on the noise figure, in particular with 0.98 μm. Since at 0.98 μm the emission coefficient is null, when backward pumping configuration is used the 0.98 μm pump is quickly absorbed along the EDF length. This leads to a low-medium inversion at the input end of the EDF, resulting in high noise figures. However, in order to achieve forward pumping, a dichroic coupler must be placed at the input end of the EDF, thus inducing an additional loss at the input of the amplifier. This increases the noise figure and, therefore, to some extent, counterbalances the NF improvement associated with forward pumping.

Raman Fiber Amplifier

Stimulated Raman scattering, an optical-phonon based nonlinear effect present in silica fiber, is used to convert energy from a pump located at a frequency ν_p, to a signal located at a lower frequency ν_s. The efficiency of the so-called Raman amplification depends on the frequency shift $\Delta\nu_r = \nu_p - \nu_s$. Figure 5 shows the dependence of the Raman gain on the frequency shift in standard single-mode fibers. The exact Raman gain spectrum depends on the doping elements used during the fabrication process of the fiber. With standard telecommunications fibers the Raman gain is maximum for $\Delta\nu_r$ close to 13 THz.

Raman amplification can be achieved within the transmission fiber of the system – distributed amplification – or within a specific fiber, such as a dispersion compensation fiber – discrete amplification. Contrary to EDFA, Raman amplification has a very fast response time – in the order of a few hundred femtoseconds – and is strongly dependent on the polarization state of the signal and the pump, the highest gain efficiency being obtained when the polarization states of the signal and the pump are parallel. Because of the nonpolarization-maintaining nature of standard telecommunications fibers, depolarized pump sources are preferable for Raman amplification. To avoid pump fluctuation being transferred to the signal, Raman amplifiers generally use backward pumping, which virtually averages the fluctuation. Forward pumping can also be employed with Raman amplifiers but with careful control of the pump characteristics.

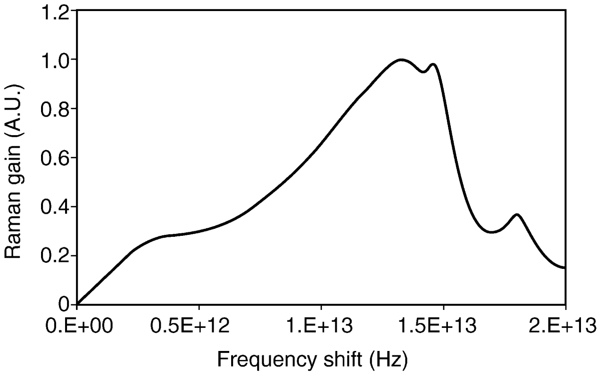

Figure 5 Raman gain spectrum in standard single mode fiber.

If the signal is sufficiently low the pump depletion can be neglected. In that case the gain of the Raman amplifier is given by

$$G = \exp\left(\frac{g_R P_P}{A_{\text{eff}}\alpha_p}\left(1 - \exp^{-\alpha_p L}\right) - \alpha_S L\right) \qquad [15]$$

where P_p is the launched pump power, L is the length of the fiber, g_R is the Raman gain coefficient, and A_{eff} is the effective area of the fiber. We assume for the simulation examples given in this article that the peak value of g_R is equal to 6×10^{-14} m/W.

A more comprehensive model for Raman amplification is given by the following set of differential equations:

$$\frac{dP_S^+(z)}{dz} = -\alpha_s P_S^+(z) + \left(\frac{g_R}{A_{\text{eff}}}\right)\left(P_p^+(z) + P_p^-(z)\right)P_S^+(z) + r_S P_S^-(z) \qquad [16]$$

$$\frac{dP_S^-(z)}{dz} = \alpha_s P_S^-(z) - \left(\frac{g_R}{A_{\text{eff}}}\right)\left(P_p^+(z) + P_p^-(z)\right)P_S^-(z) - r_S P_S^+(z) \qquad [17]$$

$$\frac{dP_{\text{ASE}}^+(z)}{dz} = -\alpha_s P_{\text{ASE}}^+(z) + \left(\frac{g_R}{A_{\text{eff}}}\right)\left(P_p^+(z) + P_p^-(z)\right) \times \left[P_{\text{ASE}}^+(z) + 2h\nu_S\Delta\nu\right] + r_S P_{\text{ASE}}^-(z) \qquad [18]$$

$$\frac{dP_{\text{ASE}}^-(z)}{dz} = \alpha_s P_{\text{ASE}}^-(z) - \left(\frac{g_R}{A_{\text{eff}}}\right)\left(P_p^+(z) + P_p^-(z)\right) \times \left[P_{\text{ASE}}^-(z) + 2h\nu_S\Delta\nu\right] - r_S P_{\text{ASE}}^+(z) \qquad [19]$$

$$\frac{dP_p^+(z)}{dz} = -\alpha_p P_p^+(z) - \left(\frac{\nu_p}{\nu_S}\right)\left(\frac{g_R}{A_{\text{eff}}}\right)P_p^+(z) \times \left(P_S^+(z) + P_S^-(z) + P_{\text{ASE}}^+(z) + P_{\text{ASE}}^-(z)\right) + r_P P_p^-(z) \qquad [20]$$

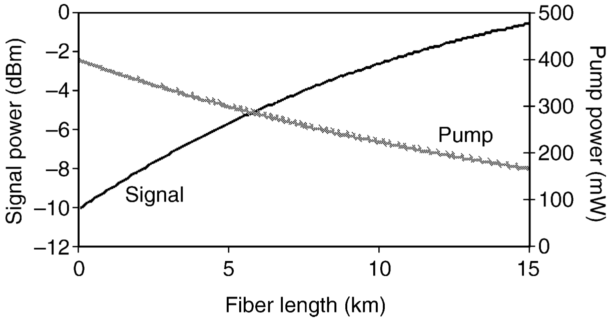

Figure 6 Evolution of the signal power and pump power along fiber length with Raman amplification (standard single-mode fiber; fiber length of 15 km; input signal power of −10 dBm at 1550 nm; forward pumping with 400 mW at 1450 nm; total amplifier gain of 9.5 dB).

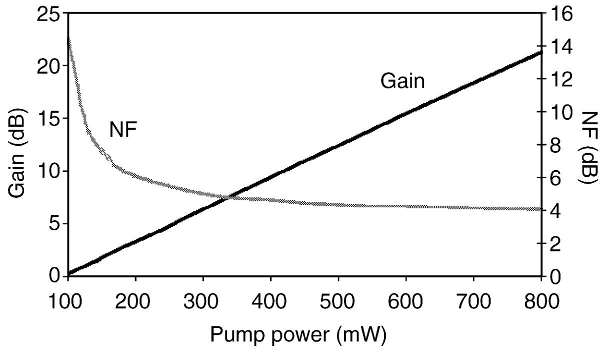

Figure 7 Dependence of the Raman amplifier gain and the noise figure on the pump power (standard single-mode fiber; fiber length of 15 km; input signal power of −10 dBm at 1550 nm; forward pumping at 1450 nm with power ranging from 100 mW to 800 mW).

$$\frac{\mathrm{d}P_{\mathrm{p}}^{-}(z)}{\mathrm{d}z} = \alpha_{\mathrm{p}} P_{\mathrm{p}}^{-}(z) + \left(\frac{\nu_{\mathrm{p}}}{\nu_{\mathrm{S}}}\right)\left(\frac{g_{\mathrm{R}}}{A_{\mathrm{eff}}}\right) P_{\mathrm{p}}^{-}(z) \times \left(P_{\mathrm{S}}^{+}(z) + P_{\mathrm{S}}^{-}(z) + P_{\mathrm{ASE}}^{+}(z) + P_{\mathrm{ASE}}^{-}(z)\right) - r_{\mathrm{P}} P_{\mathrm{p}}^{+}(z) \qquad [21]$$

where r_{S} and r_{p} are the Rayleigh backscattering coefficients at, respectively, the signal wavelength and the pump wavelength.

An example of the evolution of the signal power and pump power along fiber length is given by Figure 6.

Figure 7 shows the dependence of the Raman amplifier gain and noise figure on the pump power.

Distributed and Discrete Amplification

A system uses distributed amplification when the transmission fiber is used as the medium for generating the optical signal amplification. It uses discrete amplification when an optical amplifier is located between the transmission fiber spans.

Discrete Amplification

Both EDFA and Raman amplifiers can be used to provide discrete amplification. Discrete amplifiers can be used as booster, pre-amplifier, or in-line amplifier. They can also be used to compensate for the loss of additional elements such as DCM, which may be located between the fiber spans. In Figure 8 we show the path of the signal when a discrete amplification is used to compensate for the loss of the transmission fiber and for the loss of a DCM. In this example, we assume that the discrete amplifier is made of two amplifiers. The first-stage amplifier compensates mainly for the loss of the transmission fiber, while the second-stage amplifier compensates mainly for the loss of the DCM.

If a 'lossy' element is inserted between two amplifiers, the noise figure of the composite amplifier increases the total noise figure of the dual-stage amplifier.

The total noise figure of a dual-stage amplifier, $\mathrm{NF}_{\mathrm{total}}$, is given by

$$\mathrm{NF}_{\mathrm{total}} = \mathrm{NF}_1 - \frac{1}{G_1} + \frac{\mathrm{NF}_2}{G_1 T_{\mathrm{mid}}} \qquad [22]$$

where NF_1 is the noise figure of the first-stage amplifier, G_1 is the gain of the first-stage amplifier, NF_2 is the noise figure of the second-stage amplifier, and T_{mid} is the loss of the optical element inserted between the two amplifiers ($T_{\mathrm{mid}} < 1$).

Distributed Amplification

When fibers are deployed it is generally assumed, at least for terrestrial networks, that it can be used with different types of transmission systems, which may or

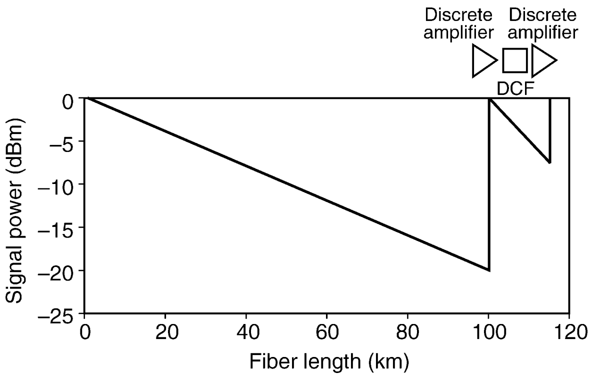

Figure 8 Schematic of a transmission system using discrete amplification to compensate for the loss of the transmission fiber and the dispersion compensating module.

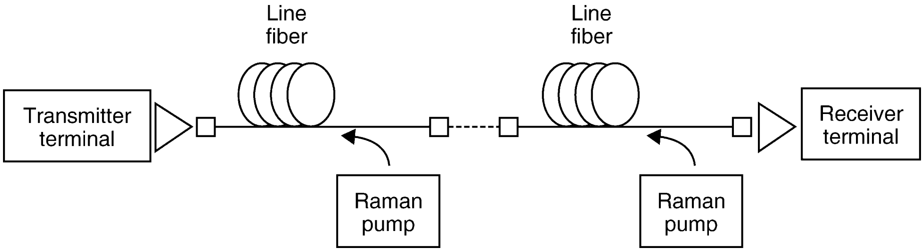

Figure 9 Schematic of a transmission system using distributed Raman amplification.

may not use distributed amplification. Doping the transmission fiber with erbium is not a practical solution since the absence of a pump would result in high fiber loss. On the other hand, Raman amplification can be achieved with conventional fibers, and is consequently the preferred solution to generate distributed amplification. Raman amplification is then obtained by launching a pump signal into the transmission fiber. Figure 9 shows the schematic of a transmission system using distributed Raman amplification with backward pumping.

Even if distributed amplification can be used to compensate fully for the loss of the transmission fiber, discrete amplifiers are generally required to compensate for the loss of the transmitter and receiver terminals. In order to compare distributed amplification with discrete amplification we define the equivalent noise figure as follows:

$$\mathrm{NF}_{\mathrm{Equ}} = \frac{1}{G_{\mathrm{On/Off}}}\left(\frac{\mathrm{Pase}}{h\nu\Delta\nu}+1\right) \qquad [23]$$

where Pase is the total ASE power at the output end of the transmission fiber and where the on/off gain, $G_{\mathrm{On/Off}}$ is given by

$$G_{\mathrm{On/Off}} = \frac{P_{\mathrm{signal}}^{\mathrm{On}}}{P_{\mathrm{signal}}^{\mathrm{Off}}} \qquad [24]$$

where $P_{\mathrm{signal}}^{\mathrm{On}}$ and $P_{\mathrm{signal}}^{\mathrm{Off}}$ are the signal power levels at the output end of the transmission fiber when the distributed amplifier is, respectively, turned on and off.

To illustrate the meaning of the on/off gain, we show in Figure 10 the evolution of the signal power along a fiber with and without Raman pump power. In this case, the on/off gain is 20 dB.

The equivalent noise figure, which has no real physical meaning, is equal to the noise figure of an imaginary discrete amplifier that would be located at the output end of the transmission fiber and would lead to the same OSNR. Figure 11 shows the dependence of the equivalent noise figure on the Raman gain, with no Rayleigh backscattering and with a Rayleigh backscattering level of −74 dB m^{-1}.

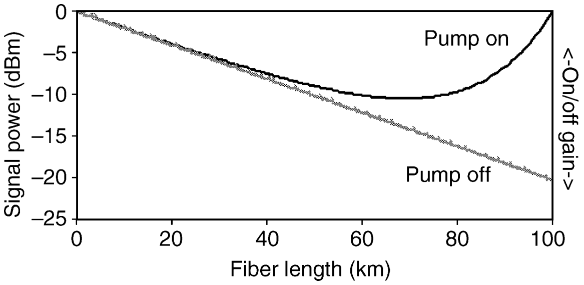

Figure 10 Evolution of signal power along the fiber length with and without pump power (standard single-mode fiber; fiber length of 100 km; fiber attenuation of 0.2 dB/km at signal wavelength; input signal power of 0 dBm at 1550 nm; backward pumping 400 mW at 1450 nm).

Once the on/off gain is higher than ~10 dB the equivalent noise figure becomes negative and it decreases when the on/off gain further increases. However, once the on/off is higher than 20 dB, Rayleigh backscattering limits the minimum equivalent noise figure that can be achieved. The lowest equivalent noise figure is observed for an on/off gain of ~33 dB. In addition, Rayleigh backscattering causes the reflected portion of the signal to be superimposed onto the transmitted signal, leading to multipath interference (MPI), which level increases with the gain. This results in an additional source of noise for the signal, and consequently degrades the system performance.

Hybrid Amplification

Discrete and distributed amplification can be used simultaneously in a same-transmission system. An example of the evolution of the signal power with hybrid amplification is represented in Figure 12, for the case where distributed Raman amplification and a dual-stage discrete amplifier are used to compensate for the loss of the transmission fiber and for the loss of a DCM.

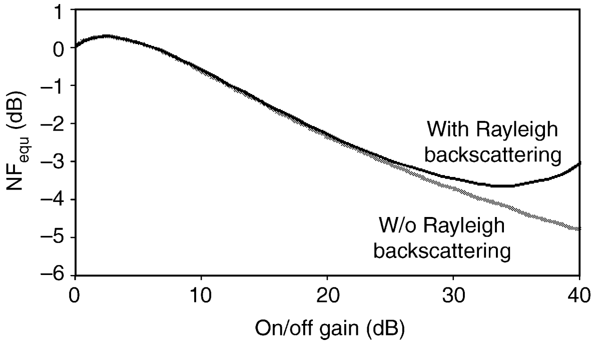

Figure 11 Equivalent noise figure versus distributed on/off gain (standard single-mode fiber; fiber length of 250 km; fiber attenuation of 0.2 dB/km at 1550 nm; input signal power of 0 dBm at 1550 nm; backward pumping ranging from 0 to 1 W at 1450 nm).

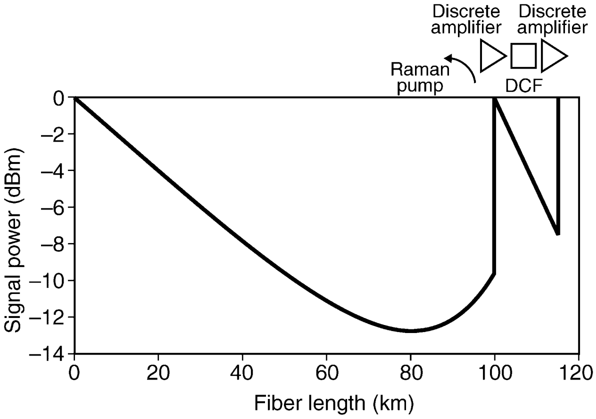

Figure 12 Evolution of the signal power using discrete and distributed amplification (standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; DCM loss of 8 dB; input signal power of 0 dBm at 1550 nm; Raman backward pumping with 200 mW at 1450 nm; Raman on/off gain of 10 dB).

Similar to the case of distributed amplification, the total equivalent noise figure of a hybrid amplifier can be defined by

$$\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Total}} = \mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}} + \frac{\mathrm{NF}^{\mathrm{Discrete}} - 1}{G_{\mathrm{On/Off}}} \qquad [25]$$

where $\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}}$ is the equivalent noise figure of the distributed amplifier and $\mathrm{NF}^{\mathrm{Discrete}}$ is the noise figure of the discrete amplifier.

In Figure 13 we show the dependence of the total equivalent noise figure on the on/off gain for a hybrid amplifier with fixed discrete noise figure. The dependence of the distributed equivalent noise figure on the on/off gain is also plotted on the same graph.

Since the equivalent noise figure of the distributed amplifier is relatively low, the total equivalent noise figure is lower than the noise figure of the discrete amplifier once distributed amplification is used. Moreover, as the on/off gain becomes large, the total equivalent noise figure of the hybrid amplifier converges toward the distributed equivalent noise figure.

Wavelength-Division Multiplexing

The transmission of WDM signals over long distances requires meticulous control of the spectral characteristics of the amplifier gain. The amplifier gain excursion induces a power excursion between the channels of the WDM signal, which increases with the number of spans. Consequently, the available bandwidth for WDM transmission narrows with the total distance of the link. This is illustrated in Figure 14, where the power excursion observed after an amplifier chain is plotted for different numbers of amplifiers, a fiber span being included between each amplifier. With this example, the 1 dB-bandwidth drops from 9 nm after one span to 1.5 nm after 50 spans.

For a comprehensive analysis of the spectral gain characteristic of EDFA and Raman amplifiers we generalize the single-channel model described in the amplifier section of this article to the multichannel case. The generalization of the models is obtained by adding a frequency dimension to each beam propagating in the amplifier, as described in Table 1, where each k element represents the signal, the ASE noise or the pump, λ is the wavelength of the kth beam, $\Delta\nu$ is the frequency separation between the pump beam p, and the signal beam s.

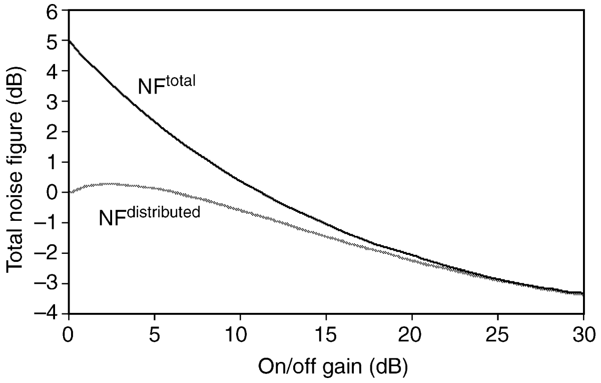

Figure 13 Total noise figure of the hybrid amplifier and noise figure of the distributed amplifier versus on/off gain (noise figure of the discrete amplifier equals to 5 dB; standard single-mode fiber of 100 km; fiber loss of 20 dB at signal wavelength; input signal power of 0 dBm at 1550 nm; Raman backward pumping ranging from 0 mW to 600 mW at 1450 nm).

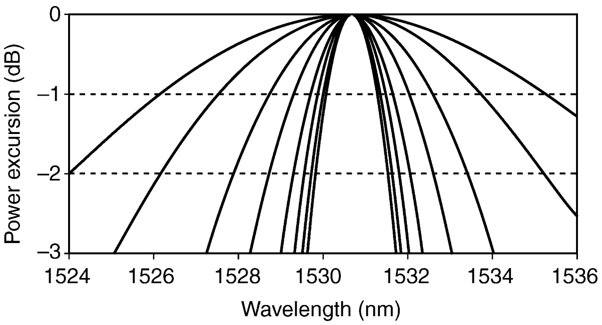

Figure 14 Dependence of the power excursion on the number of fiber spans. Number of spans equals, respectively, to 1, 2, 5, 10, 20, 30, 40, and 50 (amplifier with 15 m long alumino-silicate EDF; peak absorption of 1 dB m^{-1}; $D = 0.6$; fiber loss of 8.5 dB; flat fiber loss spectrum).

Table 1 Correspondence between parameters used for the EDFA and Raman amplifier models with the single-channel and the multichannel approaches (the multichannel approach is also used to describe the power evolution of pump beam when WDM pump scheme is used)

	Single-channel model	*Multichannel model*
Power	$P_k^{+,-}(z)$	$P_k^{+,-}(\lambda, z)$
Sum of power	$P_k^{+}(z)$ $+P_k^{-}(z)$	$\sum_\lambda P_k^{+}(\lambda, z)$ $+P_k^{-}(\lambda, z)$
Absorption coefficient	α_k	$\alpha(\lambda)$
Emission coefficient	g_k	$g(\lambda)$
Raman gain coefficient	g_R	$g_R(\Delta\nu_{p,s})$

With EDFAs, the main factors in the gain flatness are the path-average medium inversion, the EDF codopant composition, and fiber host material. Depending on the bandwidth and the wavelength of the WDM signal, the way the EDF is optimized may vary.

In the following example, we focus on the path-average medium inversion for the case of C-band applications. Figure 15 shows the output spectrum of a WDM signal amplified by an alumino-silicate EDFA having, respectively, a medium inversion of $D = +0.4$ and $D = 0$. The peak-to-peak gain excursion is, respectively, 2 dB and 14 dB for $D = +0.4$ and $D = 0$.

Application with C-band WDM signals requires amplifiers operating with high-medium inversion. As can be extrapolated from Figure 15, if the WDM signal is centered around 1560 nm, a lower medium inversion is more suitable. Indeed, as shown in Figure 3, when the medium inversion is close to zero the peak gain wavelength changes from 1530 nm to 1560 nm and higher; the optical window for L-band applications.

Regarding the doping elements, we show in Figure 16 the gain spectrum for a germano-silicate and an alumino-silicate EDF, respectively. The peak-to-peak gain excursion is 8 dB with the germano-silicate EDF compared to 2 dB with the alumino-silicate EDF.

With Raman amplifiers, the gain flatness can be adjusted by using wavelength-multiplexed pumps. In Figure 17 we show the power spectrum when a Raman amplifier is pumped with three wavelength-multiplexed pumps located, respectively, at 1430 nm, 1455 nm, and 1480 nm. The maximum power excursion between the channels is 2 dB.

For comparison, we also show in Figure 18 the power spectrum of the Raman amplifiers when only

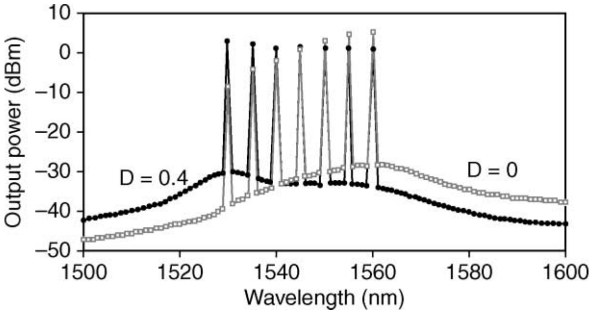

Figure 15 Gain spectrum of an alumino-silicate EDF for $D = 0.4$ and $D = 0$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; EDF with peak absorption of 1 dB m^{-1}; EDF length of 30 m for $D = 0.4$; EDF length of 100 m for $D = 0$).

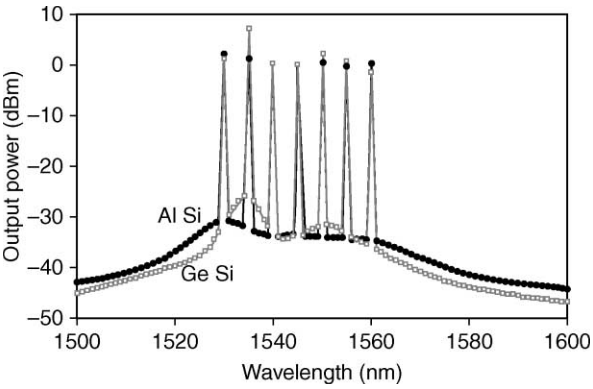

Figure 16 Gain spectrum of an alumino-silicate EDF and of a germano-silicate EDF for $D = 0.4$ (C-band WDM signal made of 7 channels equally spaced ranging from 1530 nm to 1560 nm; flat input power of −10 dBm per channel; gain of channel located at 1545 nm of 10 dB for $D = 0.4$ and $D = 0$; alumino-silicate EDF of 30 m with peak absorption of 1 dB m^{-1}; germano-silicate EDF of 40 m with peak absorption of 1 dB m^{-1}).

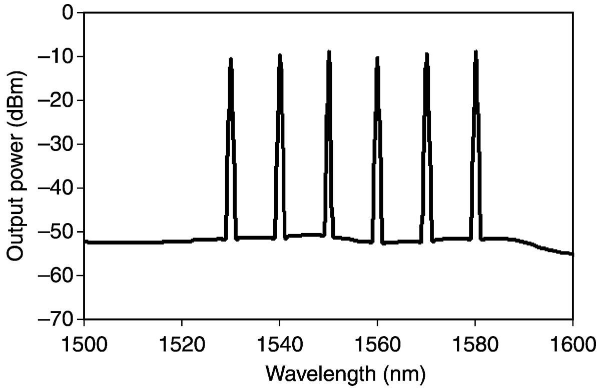

Figure 17 Output power spectrum of a Raman amplifier pumped with three wavelength-multiplexed pumps (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 100 mW per pump located at 1430 nm, 1455 nm, and 1480 nm).

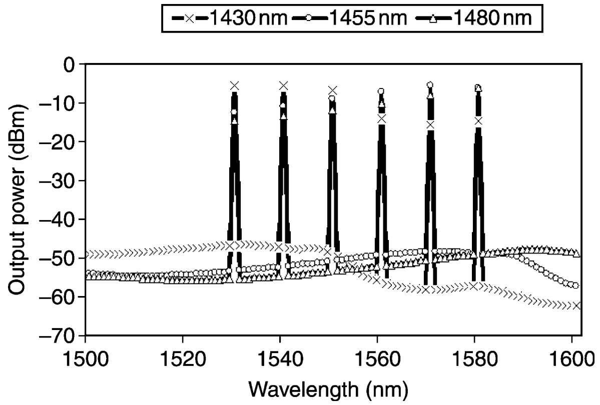

Figure 18 Output power spectrum of Raman amplifiers pumped at three different wavelengths (standard single-mode fiber of 100 km; WDM signal made of 6 channels equally spaced ranging from 1530 nm to 1580 nm; pumping in forward direction with 300 mW at, respectively, 1430 nm, 1455 nm, and 1480 nm).

one of the three multiplexed pumps is used, with the same total pump power. The power excursion becomes 10 dB for the 1430 nm pump, 7 dB for the 1455 nm pump, and 8.5 dB for the 1480 nm pump, which are much higher values than are obtained with the multiplexed pump scheme.

To further reduce the gain excursion of the amplifiers an external gain equalizing filter can be used. The filter, which is located next to the amplifier, is manufactured so that its attenuation spectrum matches the gain spectrum of amplifier, as illustrated in Figure 19. Gain equalizing filters are generally made with thin film filters or fiber Bragg gratings (FBG).

In order to obtain amplification over ultra-wide bandwidths, amplifiers of smaller bandwidths, e.g., C-band and L-band EDFAs, can be operated in parallel by means of Mux and Demux. Another option is to use a Raman amplifier with multiple pumps having their wavelengths distributed over a very large bandwidth.

Long-Haul Transmission System Types

Long-haul optical transmission systems are appropriate for different types of applications. We distinguish here three types of transmission system: the submarine systems; the terrestrial systems; and the unrepeatered systems. Each of these applications requires a careful design of the system, with particular attention to the optical amplifiers. In this section we focus on the key points of their design.

Submarine Systems

Submarine systems are used to transmit information across the oceans, i.e., over distances of up to several thousand kilometers. The transmission fiber and the in-line optical amplifiers, which are packaged in the repeater modules, are deployed undersea at the same time. The drastic constraints imposed on the reliability of the repeaters lead to their high cost, and consequently minimizing their number in a link is essential. The optimum span length, and therefore the number of in-line amplifiers, depends on the total distance of the transmission system. To illustrate this point, we derive from eqn [6] which number of spans of a particular length can be used for a given received OSNR, launched signal power, amplifier noise figure, and span loss per km. Based on the number of spans and span length, we then calculate the maximum distance that can be reached. The result is represented in Figure 20.

With the assumption made here regarding the amplifier characteristics, in order to reach distances of more than 5000 km, span lengths close to 50 km are required.

Terrestrial Systems

Terrestrial systems are used to connect traffic from cities to cities and from cities to submarine-system landing sites. Their total distances vary from a few hundred kilometers to several thousand kilometers. Generally with terrestrial systems, the deployments of the transmission fibers and the transmission systems are done at different times. During its lifetime, the same fiber plant might be used with a different generation of transmission system. Furthermore, because of the geographical constraints that can be imposed on the

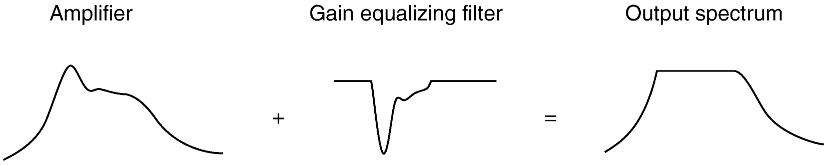

Figure 19 Combination of optical amplifier and gain equalizing filter to provide a flat gain amplifier.

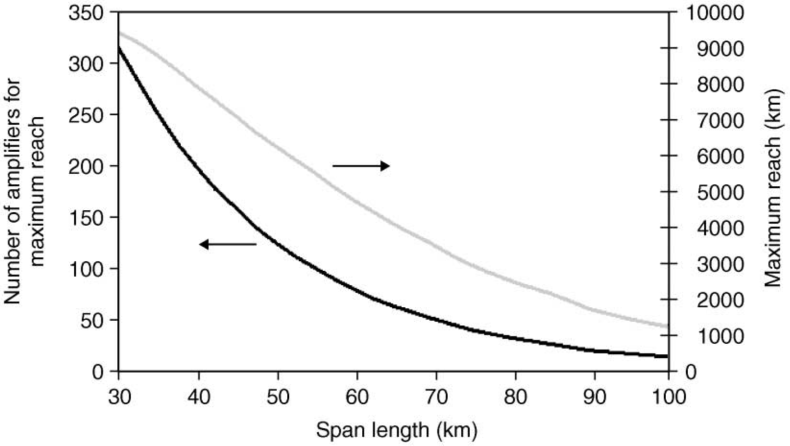

Figure 20 Dependence of the maximum distance and span number of a system on its span length (signal power at the output of the amplifier of −2 dBm; amplifier noise figure of 5 dB; fiber loss of 0.2 dB/km; required OSNR of 20 dB/0.1 nm).

locations of the terminals and the in-line optical amplifiers, the different span lengths may greatly vary in a given link. Consequently, an important characteristic of the terrestrial systems is their capability to operate with different span parameters.

For terrestrial systems using large numbers of WDM channels, maintaining the gain excursion of the optical amplifiers as little as possible under different operating conditions is crucial. Approaches to limit the gain excursion under these variations can be taken. With EDFAs of a given length, the medium inversion depends on its operating gain. The EDFA can be forced to operate at constant gain, independent of the span loss, by adding an adjustable attenuator to the amplifier, e.g., in the middle of a dual-stage amplifier, such that the sum of the attenuator loss and the span loss remains constant. With Raman amplifiers, the control of the gain excursion can be realized by adjusting the power of the pumps when a wavelength-multiplexed pumping scheme is used. Another possibility is to insert dynamically adjustable gain flattening devices along the link. The automatic adjustment of the gain flatness requires the need of a feedback generally provided by an optical channel power monitor.

Unrepeatered Systems

Unrepeatered systems are used to connect mainland sites separated by a water area, such as islands or cities located on the coast. Their main characteristic is that they do not use any electronic component along the transmission fiber; therefore optical amplifiers cannot be placed under water. The length of such systems is typically of a few hundred kilometers, meaning that, even for those systems where very low loss fiber is used, the loss of the transmission fiber is in the order of several tens of dB. To allow sufficiently high OSNR at the receiver, several options can be used. Among them are the use of very high-power boosters, distributed Raman amplification, and remotely pumped discrete amplification. Remote amplification is obtained by inserting, along the submerged fiber cable, a piece of EDF. A 1.48 μm pump is launched into the EDF from the mainland via the transmission fiber or via a dedicated fiber running along the transmission fiber. Figure 21 shows a schematic of a possible configuration for unrepeatered systems when a very high-power booster, distributed Raman amplification, and remote discrete amplification are used simultaneously.

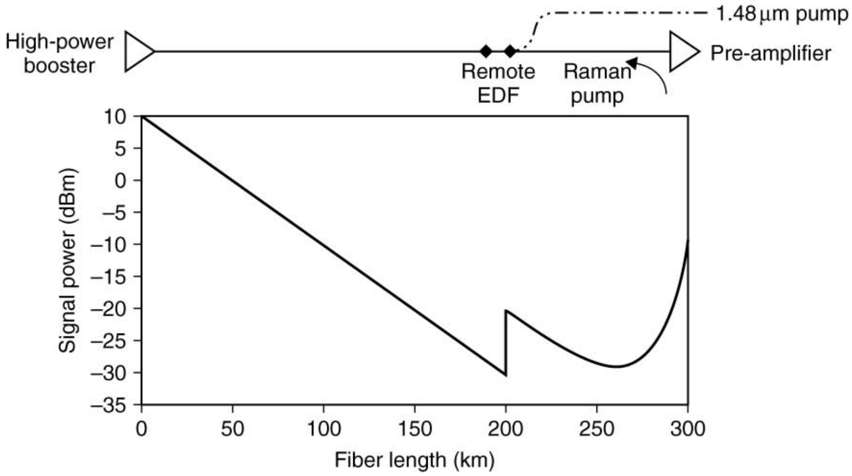

Figure 21 Schematic of unrepeatered systems using high-power booster, remote EDFA, and distributed Raman amplification.

The total equivalent noise figure, $\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Total}}$, of an unrepeatered transmission system, using remote discrete amplification and distributed amplification, is given by

$$\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Total}} = T\,\mathrm{NF}^{\mathrm{Remote}} + \frac{1}{G_{\mathrm{Remote}}} \times \left(\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}} + \frac{\mathrm{NF}^{\mathrm{Discrete}} - 1}{G_{\mathrm{On/Off}}} - T \right) \quad [26]$$

where $\mathrm{NF}^{\mathrm{Remote}}$ and G_{Remote} are, respectively, the noise figure and the gain of the remote amplifier; T is the loss of the transmission fiber between the remote amplifier and the pre-amplifier ($T < 1$); $\mathrm{NF}_{\mathrm{Equ}}^{\mathrm{Distributed}}$ and $G_{\mathrm{On/Off}}$ are, respectively, the equivalent noise figure and on/off gain of the distributed amplifier; and $\mathrm{NF}^{\mathrm{Discrete}}$ is the noise figure of the pre-amplifier.

Assuming that the gain and noise figures of the remote EDFA are, respectively, 10 dB and 7 dB, that the remote EDFA is located at 100 km from the pre-amplifier, that the fiber loss per km is 0.2 dB/km, that the Raman distributed amplifier provides 30 dB of on/off gain with an equivalent noise figure of −3 dB, and that the pre-amplifier noise figure is 5 dB, the equivalent total noise figure of the unrepeatered system is −10 dB.

List of Units and Nomenclature

A_{eff}	effective area
ASE	amplified spontaneous emission
B_{eff}	effective erbium doping area
D	normalized medium inversion
dB	ratio in decibel
dBm	power in decibel
DBR	distributed Bragg reflector
DCM	dispersion compensating modules
Demux	optical de-multiplexers
DFB	distributed feedback
DSF	dispersion-shifted fiber
EDF	erbium-doped fiber
EDFA	erbium-doped fiber amplifiers
FBG	fiber Bragg gratings
g_{R}	Raman gain coefficient
G	gain
$G_{\mathrm{On/Off}}$	on/off gain
h	Planck's constant
Length	meter (m)
MPI	multi-path interference
Mux	optical multiplexers
n_{eq}	equivalent input noise factor
NF	noise figure of the amplifier
NRZ	non-return to zero
NZDSF	non-zero dispersion shifter fiber
N_2/N_1	medium inversion
OSNR	optical signal-to-noise ratio
Power	watt (W)
SSMF	standard single-mode fiber
TDM	time-division multiplexing
Time	second (s)
WDM	wavelength-division multiplexing
τ	lifetime

See also

Coherent Lightwave Systems. **Dispersion Management**. **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics). **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Raman, Brillouin and Parametric Amplifiers.

Optical Communication Systems: Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal G (2001) *Nonlinear Fiber Optics*. London: Academic Press.

Chesnoy J (2002) *Undersea Fiber Communication Systems*. London: Academic Press.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers, Principles and Applications*. New York: John Wiley & Sons.

Desurvire E, Bayart D, Desthieux B and Bigo S (2002) *Erbium-Doped Fiber Amplifiers, Device and System Developments*. New York: John Wiley & Sons.

Emori Y, Kado S and Namiki S (2002) Broadband flat-gain and low-noise Raman amplifiers pumped by wavelength-multiplexed high-power laser diodes. *Optical Fiber Technology* 8: 107.

Hansen PB and Eskildsen L (1997) Remote amplification in repeaterless transmission systems. *Optical Fiber Technology* 3: 221.

Islam MN (2004) *Raman Amplifiers for Telecommunications*. New York: Springer-Verlag.

Kaminow I and Li T (2002) *Optical Fiber Telecommunications IV*. London: Academic Press.

Keiser G (2000) *Optical Fiber Communications*. New York: McGraw-Hill.

Raman, Brillouin and Parametric Amplifiers

M F S Ferreira, University of Aveiro, Aveiro, Portugal

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Glass fibers for optical communications are made of fused silica, an amorphous material, to which dopant materials of various kinds can be added to produce changes in refractive index. A number of third-order nonlinear processes can occur; these can grow to appreciable magnitudes over the long lengths available in fibers, even though the nonlinear coefficients in the materials are relatively small. The effects are particularly important in single-mode fibers, in which the small mode field dimensions result in substantially high light intensities with relatively modest input powers.

Nonlinear optical effects, such as stimulated Raman scattering (SRS), stimulated Brillouin scattering (SBS), and four-wave mixing (FWM), can cause some problems such as nonlinear loss, pulse distortion, and cross-talk between signals in wavelength division multiplexed (WDM) transmission systems. However, these effects are also useful for efficient amplification and frequency conversion of injected signals, as well as for generation of new frequencies. In this chapter the main properties of fiber Raman amplifiers (FRAs), fiber Brillouin amplifiers (FBAs), and fiber parametric amplifiers (FPAs) will be reviewed, regarding their applications in the field of fiber optic communications.

Fiber Raman Amplifiers

Stimulated Raman scattering is likely to occur in fibers, producing appreciable amplification for down-shifted signals when moderate pump power levels are used. This is largely because the process is phase-matched, so that the long interaction lengths available in fibers will enhance the process. Additionally, the small mode field cross-section yields high intensities for relatively modest power levels.

In the SRS process one has simultaneously the absorption of a photon from the pump beam at frequency ω_p and the emission of a photon at the signal (Stokes) frequency, stimulated by the incident signal beam at frequency ω_s (Figure 1a). The difference in energy is taken up by a high energy

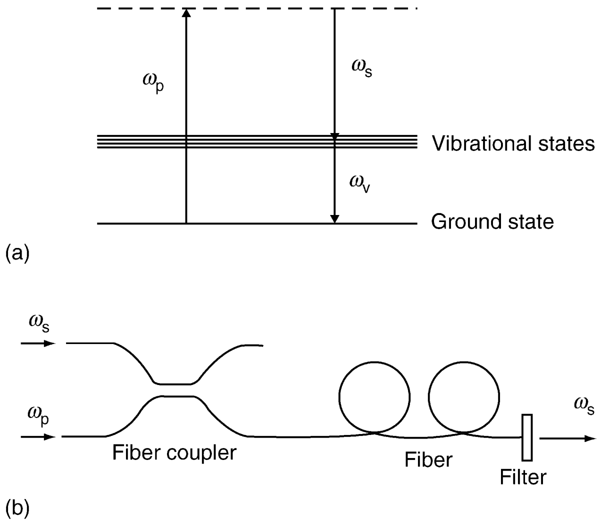

Figure 1 (a) Energy-level illustration of the process of stimulated Raman scattering and (b) schematic diagram of a fiber Raman amplifier.

phonon (molecular vibration) at frequency ω_v. Thus, SRS provides for energy gain at the signal frequency at the expense of the pump.

Figure 1b shows schematically a fiber Raman amplifier. The pump and the signal beams are injected into the fiber through a WDM fiber coupler. The case illustrated in Figure 1a shows the two beams co-propagating inside the fiber, but the counter-propagating configuration is also possible. In fact, the SRS can occur in both directions, forwards and backwards. It was confirmed experimentally that the Raman gain is almost the same in both cases.

Raman Gain and Bandwidth

Figure 2 shows the Raman gain coefficient for fused silica as derived from the spontaneous Raman spectrum. At the peak, the Raman gain coefficient for silica fibers is 9.4×10^{-14} m W^{-1} for a pumping wavelength $\lambda_p = 1.0\,\mu$m and varies as λ_p^{-1}. At this wavelength, a shift of 15 THz corresponds to 50 nm. The most significant feature of the Raman gain in silica fibers is that $\bar{g}_R$ extends over a large frequency range (up to 40 THz) with a broad dominant peak near 13 THz. This behavior is due to the amorphous nature of silica glass, whose molecular vibrational energy levels merge together to form a band.

The use of oxide glasses as dopants for fiber fabrication is suitable to enhance the Raman gain coefficient. For example, in the case of pure GeO_2, the Raman gain coefficient is about 9.2 times that for pure SiO_2 glass. In fact, doping the core of an SiO_2 Raman fiber amplifier with GeO_2 has two beneficial effects. First, the index difference Δ is raised, which increases the effective waveguiding and in turn reduces the effective cross-sectional area A_e. As a consequence, the rate of SRS is increased. Secondly, the Raman gain coefficient increases above that of pure SiO_2 in proportion to the concentration GeO_2, and is given by

$$g_R = (1 + 80\Delta)\bar{g}_R \qquad [1]$$

where $\bar{g}_R$ is the Raman coefficient for pure SiO_2.

In spite of the positive aspects mentioned above, doping with GeO_2 also increases the fiber losses, due to the concomitant rise in Rayleigh scatter and in the number of dopant-dependent loss centers. Thus, greatly increasing the GeO_2 concentration will not be worthwhile if the increased fiber losses outweigh the improvement in amplifier gain.

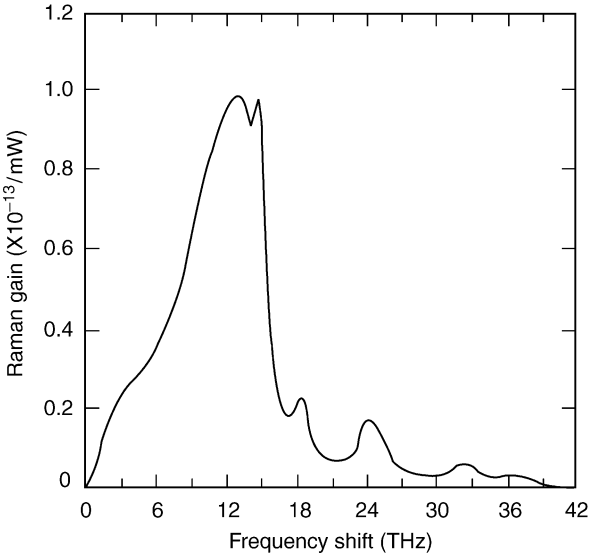

Figure 2 Measured Raman gain spectrum for fused silica at a pump wavelength $\lambda_p = 1.0\,\mu$m. The Raman gain scales inversely with λ_p. Reproduced with permission from Stolen RH and Ippen EP (1973) *Applied Physics Letters* 22: 276. © 1973 American Institute of Physics.

Gain Saturation

The interaction between the pump and Stokes waves is governed by the following coupled equations:

$$\frac{dP_s}{dz} = -\alpha_s P_s + \frac{g_R}{A_e} P_p P_s \qquad [2]$$

$$\frac{dP_p}{dz} = -\alpha_p P_p - \frac{\omega_p}{\omega_s}\frac{g_R}{A_e} P_s P_p \qquad [3]$$

where A_e is the effective cross-sectional area of the fiber mode, α_s and α_p are absorption coefficients which account for the fiber loss at the signal and pump frequencies, respectively, and the signal wave is considered to be co-propagating with the pump wave. Assuming that $\alpha_s = \alpha_p = \alpha$ (which is a reasonable approximation around the 1.5 µm wavelength region in a low loss fiber), the following approximate solutions of eqns [2] and [3] can be written when $P_p(0) \gg P_s(0)$:

$$P_p(z) = \frac{P_p(0)\exp(-\alpha z)}{1 + F(z)} \qquad [4]$$

$$P_s(z) = \frac{P_s(0)\exp(\Gamma(1 - e^{-\alpha z}) - \alpha z)}{1 + F(z)} \qquad [5]$$

where

$$\Gamma = \frac{g_R P_p(0)}{\alpha A_e} \qquad [6]$$

and

$$F(z) = \frac{\omega_p P_s(0)}{\omega_s P_p(0)} \exp(\Gamma(1 - e^{-\alpha z})) \qquad [7]$$

The dimensionless parameter $F(z)$ accounts for the effects of pump depletion and the attendant saturation of the gain seen by the signal.

Figure 3 shows the output signal power $P_s(L)$ versus the input signal power $P_s(0)$ for a Raman amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_R = 6.7 \times 10^{-14}$ m W^{-1}, $\alpha = 0.2$ dB km^{-1}, $A_e = 30 \times 10^{-12}$ m^2, $\lambda_p = 1.46$ μm and $\lambda_s = 1.55$ μm were assumed. The transfer characteristics are linear for $P_p(0) = 0.1$, 0.2, and 0.4 W, but the effects of pump depletion become discernible for input signal levels $P_s(0) > -15$ dB m when $P_p(0) = 0.6$ W. For a pump power $P_p(0) = 0.8$ W the linear behavior of the Raman amplifier cannot be observed, even for signal power levels as low as -25 dB m.

Since, in the absence of Raman amplification, the signal power at the amplifier output would be $P_s(L) = P_s(0) \exp(-\alpha L)$, the amplifier gain is given by

$$G_R = \frac{P_s(L)}{P_s(0) \exp(-\alpha L)} = \frac{\exp(g_R P_p(0) L_e / A_e)}{1 + F(L)} \qquad [8]$$

where

$$L_e = \frac{1}{\alpha}[1 - \exp(-\alpha L)] \qquad [9]$$

is the effective interaction length and L is the actual fiber length. The gain given by eqn [8] is seen to be a function of the input signal power (i.e., a saturation nonlinearity) through the term $F(L)$.

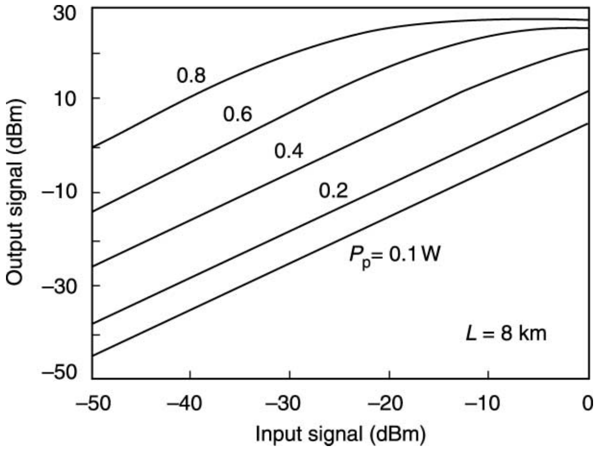

Figure 3 Transfer characteristics for a Raman amplifier with a length $L = 8$ km and several values of the input pump power.

When $F(L) \gg 1$ we have from eqn [5], that

$$P_s(L) = \frac{\omega_s}{\omega_p} P_p(0) \exp(-\alpha L) \qquad [10]$$

In this case, the amplifier output reaches the pump level irrespective of the input signal level. This implies that any spontaneous Raman scatter in the fiber will be amplified up to power levels comparable to that of the pump. This is obviously inadmissible in an amplifier application, thus this operating regime must be avoided.

On the other hand, when $F(L) \ll 1$ we have, from eqn [5], that

$$P_s(L) = P_s(0) \exp(g_R L_e P_p(0)/A_e - \alpha L) \qquad [11]$$

In this case, the amplifier gain is

$$G_R = \frac{P_s(L)}{P_s(0) \exp(-\alpha L)} = \exp(g_R L_e P_0 / A_e) \qquad [12]$$

From eqn [12], the Raman gain in decibels is expected to increase linearly with the pump input power. Raman gains were experimentally confirmed to be almost the same for forwards and backwards pumping.

Figure 4 illustrates the variation of the amplifier gain (full curves, left scale) and the saturation parameter $F(L)$ (dashed curves, right scale) with the pump power $P_p(0)$ for a fixed input signal power $P_s(0) = 0$ dBm and amplifier lengths $L = 2$, 4, and 8 km. The full curve corresponding to $L = 2$ km shows a linear dependence of the gain on the pump power. However, the other two full curves show how the pump depletion limits the achievable gain. The black point on these curves marks the value of the pump power at which the depletion parameter F

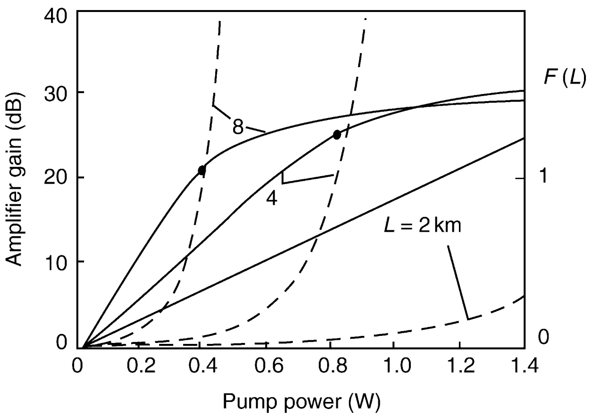

Figure 4 Raman amplifier gain (full curves, left scale) and parameter $F(L)$ (dashed curves, right scale) against input pump power for three values of the amplifier length.

becomes equal to unity. Only operating at pump powers sufficiently below each point ensures linear operation of the amplifier, whilst choosing to operate the amplifier significantly above that point will mean $F \gg 1$ and spontaneous scatter will be amplified to intolerable levels.

Noise Light Power

The dominant noise light in fiber Raman amplification is due to the amplified spontaneously Raman scattered light. In fact, a part of the pump energy is spontaneously converted into Stokes radiation extending over the entire bandwidth of the Raman gain spectrum and is amplified together with the signal. The output thus consists not only of the desired signal but also of background noise extending over a wide frequency range (~10 THz or more).

It has been shown that, for both forwards and backwards pumpings, the noise light power is equivalent to a hypothetical injection of a single-photon per unit frequency at the fiber input end for forwards pumping, and at some distance away from the fiber output for backwards pumping. Assuming that $G_R \gg 1$ and $\alpha_L \gg 1$, the noise light powers for forwards pumping, $P_{\text{fasp}}(L)$, and backwards pumping, $P_{\text{basp}}(L)$, are approximately given by

$$P_{\text{fasp}}(L) \approx h\nu_s \Delta\nu_R (G_R - 1)\exp(-\alpha L) \qquad [13]$$

$$P_{\text{basp}}(L) \approx h\nu_s \Delta\nu_R (G_R - 1)/\ln G_R \qquad [14]$$

where $\Delta\nu_R$ is the Raman gain bandwidth, h is Planck's constant, and ν_s is the signal frequency. We can observe from eqn [13] that the noise light power in the case of forward pumping decreases as the fiber length is increased. However, in the case of backwards pumping, it depends mainly on the Raman gain, being nearly independent of fiber length and loss.

FRA Performance and Applications

One main limitation affecting the operation of fiber Raman amplifiers is the requirement of a relatively high-power laser as a pump source. The experiments near 1.55 μm are usually carried out using color-center lasers for this purpose. However, such lasers are too bulky for communication applications and the use of compact high-power laser diodes (LDs) becomes desirable. This objective has recently been achieved with the advent of high-power multimode LDs. In some cases, the output lights from various LDs are combined to attain the necessary pump power.

In a Raman amplifier, SRS has to compete frequently with other nonlinear effects. For example, since the intrinsic gain coefficient for SBS is two orders of magnitude larger than that for SRS, in some cases the SBS may occur at lower pump powers and may affect significantly the SRS process. In particular, the Raman gain may not only be reduced but also become unstable. However, the SBS can be suppressed taking into account its narrow linewidth, which is typically less than 100 MHz. Therefore, if a pump laser with linewidth broader than the Brillouin bandwidth is employed, the SBS effect would not be significant. When a semiconductor laser is used as pump source, its spectral linewidth can be artificially increased by direct frequency modulation, resulting in effective SBS suppression.

Considering its broad bandwidth (>5 THz), fiber Raman amplifiers can be used to amplify several channels simultaneously in a multichannel communication system. The same characteristic also makes such amplifiers suitable for amplification of short optical pulses.

Considering a repeaterless transmission system, the FRA can be used in three different configurations: forwards pumping, backwards pumping, and bi-directional pumping. In the first case, the signal amplification occurs near the fiber input, where the signal power is relatively high. This configuration for the Raman amplification does not generally affect the transmission quality and can compensate for the transmission loss by approximately the value of the Raman gain. For example, 10 dB Raman gain provides an increase in transmission distance of 50 km when the fiber attenuation is 0.2 dB km^{-1}. In the case of backwards pumping, the signal amplification occurs near the fiber output end before photodetection, where the signal power may be very weak, comparable with the Raman noise power. This scheme for the Raman amplification can considerably reduce the minimum detectable signal power, which results in a corresponding increase of the transmission distance. However, the transmission quality is affected due to backwards noise light. The bi-directional pumping configuration uses both forwards and backwards pumping simultaneously and its performance can be estimated from those for the respective unidirectional schemes considered above.

The FRAs can also be used in nonregenerative all-optical multirepeater systems. In this case, the transmission loss is compensated for by both forwards and backwards Raman gain providing the periodically injected pump light. This scheme can be used to overcome the fiber loss in soliton-based communication systems. In this case, the pulse broadening due to chromatic dispersion of the fiber is also eliminated, which provides the possibility of achieving ultralong distance distortionless transmission.

Fiber Brillouin Amplifiers

The process of SBS can be described as a classical three-wave interaction involving the incident (pump) wave of frequency ω_p, the Stokes wave of frequency ω_s, and an acoustic wave of frequency ω_a. The pump creates a pressure wave in the medium owing to electrostriction, which in turn causes a periodic modulation of the refractive index. This process is illustrated schematically in Figure 5. Since the acoustic wavefronts are moving away from the incident pump wave, the scattered light is shifted downward in frequency to the Stokes frequency:

$$\omega_s = \omega_p - \omega_a \qquad [15]$$

The response of the material to the interference of the pump and Stokes fields tends to increase the amplitude of the acoustic wave. Therefore, the beating of the pump wave with the acoustic wave tends to reinforce the Stokes wave, whereas the beating of the pump wave and the Stokes waves tends to reinforce the acoustic wave. This explains the appearance of the stimulated Brillouin scattering process.

In spite of the apparent similarity between SBS and SRS, these processes differ in three important aspects: i) Brillouin amplification occurs only when the pump and signal beams counter-propagate inside the fiber; ii) the Stokes shift for SBS is smaller by three orders of magnitude compared with that of SRS; and iii) the Brillouin gain spectrum is extremely narrow, with a bandwidth <100 MHz. These characteristics significantly affect the performance and the fiber Brillouin amplifiers.

Brillouin Gain and Bandwidth

Physically, each pump photon in the SBS process gives up its energy to create simultaneously a Stokes photon and an acoustic phonon. Since both the energy and the momentum must be conserved in this process, the frequencies of the three waves must satisfy eqn [15], whereas the wave vectors are related by

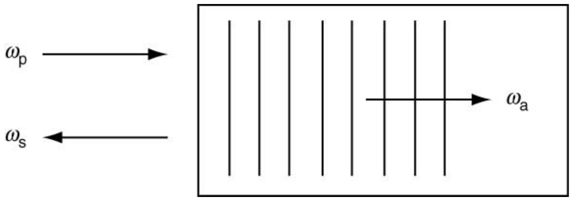

Figure 5 Schematic illustration of the stimulated Brillouin scattering process.

$$\vec{k}_p = \vec{k}_s + \vec{k}_a \qquad [16]$$

where $\vec{k}_p$ and $\vec{k}_s$ are the wave vectors of the pump and signal beams, respectively. The frequency ω_a and the wave vector $\vec{k}_a$ of the acoustic wave satisfy the dispersion relation:

$$\omega_a = |\vec{k}_a| v_a = 2 v_a |\vec{k}_p| \sin(\theta/2) \qquad [17]$$

where v_a is the acoustic velocity, θ is the angle between the pump and signal waves, and $|\vec{k}_p| \approx |\vec{k}_s|$ was used. Equation [17] shows that ω_a vanishes in the forward direction ($\theta = 0$), whereas it is maximum in the backward direction ($\theta = \pi$). The frequency shift ν_B in the backward direction is given by

$$\nu_B = \frac{\omega_a}{2\pi} = \frac{2 n v_a}{\lambda_p} \qquad [18]$$

where eqn [17] was used with $|\vec{k}_p| = 2\pi n/\lambda_p$, λ_p is the pump wavelength, and n is the refractive index. Considering $n = 1.45$ and $v_a = 5.96$ km s^{-1} as typical values for silica fibers, we obtain $\nu_B = 11.1$ GHz at $\lambda_p = 1.55$ μm.

Assuming a monochromatic pump, the spectrum of the scattered light depends on the uniformity of the fiber parameters. For a uniform medium the spectrum is Lorentzian with a width $\Delta\nu_B$ determined by the acoustic attenuation. For 1.5 μm light the width is about 17 MHz. However, both the frequency shift ν_B and the gain bandwidth $\Delta\nu_B$ can vary from fiber to fiber because of the guided nature of light and the presence of dopants in the fiber core. In particular, the inhomogeneities in the core section along the length determine an increase of the amplifier bandwidth, which can exceed 100 MHz, although typical values are 50–60 MHz for $\lambda_p \approx 1.55$ μm.

The peak value of the Brillouin gain coefficient is given by

$$g_B(\nu_B) = \frac{2\pi n^7 p_{12}^2 \gamma}{c \lambda_p^2 \rho_0 v_a \Delta\nu_B} \qquad [19]$$

where p_{12} is the longitudinal elasto-optic coefficient, ρ_0 the average material density, n the refractive index, c the vacuum speed of light, and γ is a numerical factor that depends on the polarization properties of the optical fiber, which assumes the value $\frac{1}{2}$ for long lengths of nonpolarization-preserving optical fiber. By using the parameter values typical of fused silica, g_B is estimated to be about 2.5×10^{-11} m W^{-1}, which is two orders of magnitude larger than the Raman gain coefficient at $\lambda_p = 1.55$ μm.

The expression for the Brillouin gain in eqn [19] assumes that the spectral width of the pump beam ($\Delta\nu_p$) is much narrower that the Brillouin linewidth ($\Delta\nu_B$). When such condition is not verified, the Brillouin gain is reduced according with the relation:

$$\tilde{g}_B(\nu_B) = \frac{\Delta\nu_B}{\Delta\nu_B + \Delta\nu_p} g_B(\nu_B) \qquad [20]$$

For this reason, fiber Brillouin amplifiers need to be pumped by narrow-linewidth semiconductor lasers ($\Delta\nu_p < 10$ MHz).

Gain Saturation

The slowly varying steady-state Fourier amplitudes of the Stokes (E_s) and pump (E_p) electric fields are related by the following coupled-wave equations:

$$\frac{\partial E_s}{\partial z} = \left[j\frac{K_1 K_2 |E_p|^2}{\delta - j\Gamma} + \frac{\alpha}{2} \right] E_s \qquad [21]$$

$$\frac{\partial E_p}{\partial z} = -\left[j\frac{K_1 K_2 |E_s|^2}{\delta + j\Gamma} + \frac{\alpha}{2} \right] E_p \qquad [22]$$

where Γ^{-1} is the acoustic phonon lifetime resulting in a spontaneous Brillouin scattering linewidth $\Delta\nu_B = \Gamma/\pi$, $\alpha = \alpha_s \approx \alpha_p$ is the absorption coefficient, and the coupling constants are

$$K_1 = \frac{K_2 \rho_0 n^2 \varepsilon_0}{2\nu_a}, \qquad K_2 = \frac{n^3 p_{12} \omega_s (\hat{e}_p \times \hat{e}_s)}{2c\rho_0} \qquad [23]$$

ε_0 being the free-space permittivity and $\hat{e}_p$ and $\hat{e}_s$ the unit vectors in the directions of the pump and Stokes optical fields, respectively. In eqns [21] and [22] it is assumed that the pump wave is launched at $z = 0$ and propagates in the $+z$ direction, whereas the signal wave is launched at $z = L$ and propagates in the $-z$ direction. The parameter $\delta \equiv \omega_p - \omega_s - \omega_a$ takes into account a possible detuning of the signal from the SBS gain line center.

Equations [21] and [22] can be written in terms of optical power as follows:

$$\frac{dP_s}{dz} = \alpha P_s - \frac{g_B(\delta)}{A_e} P_p P_s \qquad [24]$$

$$\frac{dP_p}{dz} = -\alpha P_p - \frac{g_B(\delta)}{A_e} P_s P_p \qquad [25]$$

where A_e is the effective core area and $g_B(\delta)$ is the Brillouin gain coefficient, given by

$$g_B(\delta) = \frac{\Gamma^2}{\delta^2 + \Gamma^2} g_B(\nu_B) \qquad [26]$$

where $g_B(\nu_B)$ is given by eqn [19].

Considering that $a \ll 1$, the following approximate solutions of eqns [24] and [25] can be written:

$$P_s(z) = P_s(0) D(z) e^{\alpha z} \qquad [27]$$

$$P_p(z) = P_p(0) D(z) H(z) e^{-\alpha z} \qquad [28]$$

where

$$D(z) = \frac{[P_p(0) - P_s(0)]}{[P_p(0) H(z) - P_s(0)]} \qquad [29]$$

and

$$H(z) = \exp\left\{ g_B [P_p(0) - P_s(0)][1 - \exp(-\alpha z)]/(A_e \alpha) \right\} \qquad [30]$$

The Brillouin amplifier gain is given from eqn [27] as

$$G_B = \frac{P_s(0)}{P_s(L) e^{-\alpha L}} = \frac{1}{D(L)} \qquad [31]$$

Figure 6 shows the output signal power $P_s(0)$ versus the input signal power $P_s(L)$ for a Brillouin amplifier with length $L = 8$ km and several values of the input pump power. Typical values $g_B(0) = 2.5 \times 10^{-11}$ m W^{-1}, $\alpha = 0.2$ dB km^{-1}, $A_e = 30 \times 10^{-12}$ m^2 were assumed. The transfer characteristics are linear for $P_p(0) = 0.2$ mW, but the effects of pump depletion become discernible

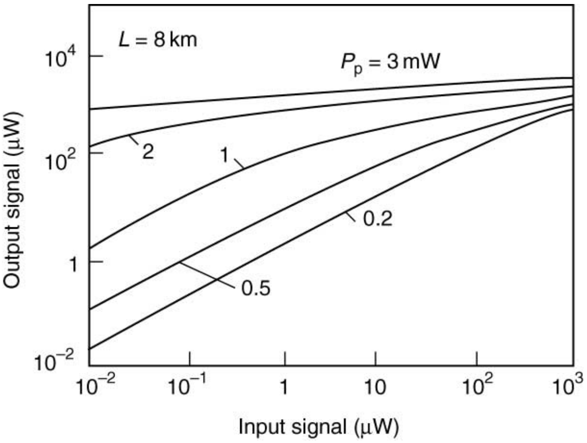

Figure 6 Transfer characteristics for a Brillouin amplifier with a length $L = 8$ km and several values of the input pump power.

for input signal levels $P_s(0) > 2\ \mu W$ when $P_p(0) = 1$ mW. For a pump power $P_p(0) = 2$ mW, linear behavior of the Brillouin amplifier cannot be observed even for signal powers as low as 10 nW.

The dependence of the Brillouin amplifier gain on the input pump power is illustrated in Figure 7 for several amounts of the signal detuning and two values of the signal power. The fiber length is $L = 5$ km and the gain bandwidth is $\Delta\nu_B = 30$ MHz, whereas the other parameter values are equal to those assumed in Figure 6. The saturated gain shows a significant dependence on the input signal power. An input signal of 1 nW, which is comparable to the original value of spontaneous emission, can be amplified by about 60 dB for a pump power $P_p(0) \approx 5$ mW. However, the amplifier gain can be considerably reduced when the signal is detuned from the gain peak. For example, for an input signal power $P_s(L) = 1$ nW and a pump power $P_p(0) \approx 3$ mW, a detuning of 15 MHz (which corresponds to a reduction of the gain coefficient to half of its maximum value) determines a reduction of about 25 dB on the amplifier gain.

Substituting eqn [28] for $P_p(z)$ in eqn [21], the following result is obtained for the Stokes field at the amplifier output:

$$E_s(0) = E_s(L)\sqrt{G_B}\exp\{-\alpha L/2 - j\phi\} \qquad [32]$$

where

$$\phi = \frac{\delta}{2\Gamma}\ln(G_B) \qquad [33]$$

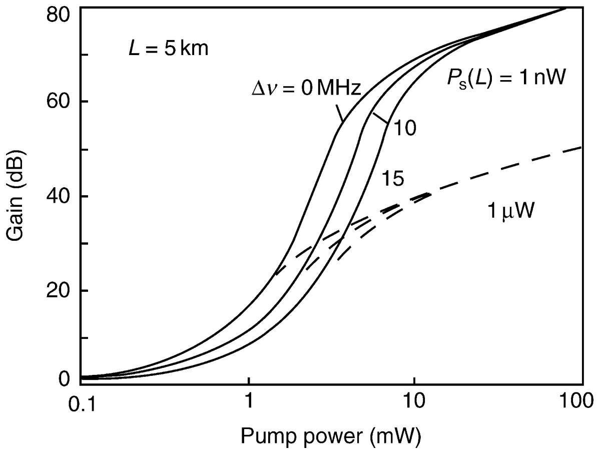

Figure 7 Brillouin amplifier gain against the input pump power for an amplifier length $L = 5$ km and different amounts of detuning from the gain center. The full curves and the dashed curves correspond to input signal powers $P_s(L)$= 1 nW and $P_s(L)$= 1 μW, respectively. Reproduced with permission from Ferreira MFS, *et al.* (1994) *Optical and Quantum Electronics* 26: 35, with permission from Kluwer Academic Publisher.

represents a nonlinear phase change, which depends on the SBS gain, input signal power, and frequency detuning. This nonlinear phase shift can impose a stringent limit on the pump- and signal-frequency stability when FBAs are used in some phase-sensitive detection schemes.

Noise Light Power

To describe the spontaneous emission noise we must add a spontaneous emission term to the equations describing the evolution of the Stokes and pump waves. The amplified spontaneous scattered power per unit frequency is then obtained in the form:

$$P_{asp}(\nu, z) = (h\nu/A_e)g_B(\delta)(N+1)G(\nu, z) \times \int_z^L P_p(z')[G(\nu, z')]^{-1}dz' \qquad [34]$$

where

$$G(\nu, z) = \exp\left\{\int_z^L [P_p(z)g_B(\delta)/A_e - \alpha]dz\right\} \qquad [35]$$

is the gain function, h is Planck's constant, and N is the thermal equilibrium number of acoustic phonons, given by

$$N = \frac{1}{\exp(h\nu_B/kT)^{-1}} \approx kT/h\nu_B \approx 500 \qquad [36]$$

where k is the Boltzmann constant and T is the absolute temperature. The total amplified signal and spontaneous powers are given by integration of $P_s(\nu, z)$ over the gain profile.

Similarly to the behavior of the gain, the noise power depends significantly on the signal magnitude in the saturation regime and becomes particularly high for low signal powers. This high noise level, which is about 20 dB above that of an ideal amplifier, imposes some limitations on the use of FBA as a receiver preamplifier.

FBA Performance and Applications

Fiber Brillouin amplifiers can provide high gains at low pump powers. However, the gain bandwidth is small and the amplified spontaneous emission noise is substantially larger than that observed with other amplifiers, leading to a noise figure that is quite large (>15 dB). This fact limits their usefulness as a preamplifier. Use as in-line amplifiers at high signal levels is possible in some configurations. Some problems in this application are concerned with the narrow bandwidth and the small saturation power of the FBAs. In particular, the Brillouin linewidth strictly limits the bandwidth of data signals that can be amplified in a FBA.

The Brillouin gain bandwidth can be intentionally extended by more than one order of magnitude by applying frequency modulation to the pump laser. Of course, the bandwidth enhancement is accompanied by a reduction of the peak gain. Consequently, a higher pump power will be necessary to achieve the same gain.

The narrow bandwidth of the FBAs can be used to advantage in coherent lightwave systems to amplify selectively the carrier of a phase- or amplitude-modulated wave, thus allowing homodyne detection, using the amplified carrier as local oscillator. The scheme should work well at bit rates >100 Mb s^{-1} because the modulation sidebands then fall outside the amplifier bandwidth, and the optical carrier can be amplified selectively. However, an inevitable accompaniment of this narrow-band amplification is the nonlinear phase shift induced by the pump on the signal, given by eqn [33]. This phase shift imposes the most stringent limit on pump- and signal-frequency stability when FBAs are used for self-homodyne detection. In the case of amplitude-shift keyed (ASK) signals, for example, a phase stability of about 0.1 rad for the amplified signal carrier may be required, which corresponds to a pump-signal frequency offset <100 kHz. The nonlinear phase shift can be used with advantage in some self-homodyne schemes which require some specific adjustment of the carrier phase. In the case of a self-homodyne coherent receiver for phase-shift keyed (PSK) signals, for example, a quadrature phase correction is required. Typically, this requires a detuning of only a few MHz from the SBS gain peak, well within the Brillouin gain bandwidth.

Another application of the narrow linewidth associated with the Brillouin gain, is as a tunable narrow-band filter for channel selection in a densely packed multichannel communication system. A channel can be selectively amplified through Brillouin amplification by launching a pump beam at the receiver end so that it propagates inside the fiber in a direction opposite to that of the multichannel signal. Different channels can be selectively amplified by tuning the pump laser. The adjustable bandwidth and high out-of-band rejection can be used to advantage in this case. In another approach, the fiber Brillouin amplifier can be used to selectively amplify multiple subcarrier-multiplexed (SCM) channels carried by a single optical carrier within a WDM signal, or to amplify a backward-propagating supervisory signal.

Fiber Parametric Amplifiers

In both stimulated Raman scattering and stimulated Brillouin scattering, the optical fiber plays an active role through the participation of molecular vibrations or acoustic phonons. However, in many other nonlinear phenomena, the fiber plays a passive role except for mediating the interaction among several optical waves. Since these phenomena involve the light-induced modulation of some medium parameter, they are known as parametric processes.

The parametric processes can be classified as second-order or third-order processes, depending on whether the second-order susceptibility or the third-order susceptibility is responsible for them. Concerning the second-order parametric processes, they are not significant in silica fibers, due to the small value of the second-order susceptibility. In contrast, some third-order parametric processes, namely third-harmonic generation, four-wave mixing, and parametric amplification, can exhibit relatively high conversion efficiencies. The parametric gain in optical fibers can be used to make fiber parametric amplifiers, which find some interesting applications for squeezing and optical fiber communications.

Phase-Matching in Parametric Processes

The efficiency of the four-wave mixing process depends on the relative phase among the interacting optical waves. A high efficiency requires matching of the frequencies as well as of the wave vectors. The latter requirement is often referred to as phase-matching. In contrast with SRS and SBS, where phase-matching is automatically fulfilled as a result of the active participation of the nonlinear medium, the phase-matching condition in parametric processes requires a specific choice of the frequencies and the refractive indices.

Considering the case in which two photons at energies ω_1 and ω_2 are annihilated with simultaneous creation of two photons at frequencies ω_3 and ω_4, the matching of the frequencies imposes the condition:

$$\omega_1 + \omega_2 = \omega_3 + \omega_4 \qquad [37]$$

On the other hand, the phase-matching is given by the condition $\Delta k = 0$, where

$$\begin{aligned}\Delta k &= k_3 + k_4 - k_1 - k_2 \\ &= (\tilde{n}_3\omega_3 + \tilde{n}_4\omega_4 - \tilde{n}_1\omega_1 - \tilde{n}_2\omega_2)/c \qquad [38]\end{aligned}$$

where the $\tilde{n}_i$ are the effective refractive indices of the fiber modes. It is relatively easy to satisfy the phase-matching condition when $\omega_1 = \omega_2$. In this case, a strong pump at $\omega_1 = \omega_2 \equiv \omega_p$ generates a low-frequency sideband at ω_3 and a high-frequency sideband at ω_4, when we assume $\omega_4 > \omega_3$. These sidebands are referred as the Stokes and anti-Stokes bands, respectively, which are also often called the signal and idler bands. The frequency shift of the two sidebands is given by

$$\Omega_s = \omega_p - \omega_3 = \omega_4 - \omega_p \qquad [39]$$

The effective indices $\tilde{n}_i$ in eqn [38] can be written as

$$\tilde{n}_i = n_i + \Delta n_i \qquad [40]$$

where Δn_i is the change of index refraction due to waveguiding. As a consequence, considering the eqn [40] and the case $\omega_1 = \omega_2$, eqn [38] can be written in the form:

$$\Delta k = \Delta k_m + \Delta k_w \qquad [41]$$

where

$$\Delta k_m = (n_3\omega_3 + n_4\omega_4 - 2n_1\omega_1)/c \qquad [42]$$

results from the material dispersion and

$$\Delta k_w = (\Delta n_3\omega_3 + \Delta n_4\omega_4 - (\Delta n_1 + \Delta n_2)\omega_1)/c \qquad [43]$$

results from the waveguide dispersion. We must note that although the two pump waves have the same frequency, Δn_1 and Δn_2 are, in general, different if these waves propagate in different modes.

The material contribution Δk_m can be expressed in terms of the frequency shift Ω_s given by eqn [39], if we use an expansion of Δk_m about the pump frequency $\omega_p = \omega_1 = \omega_2$. By retaining up to terms quadratic in Ω_s in this expansion, one obtains

$$\Delta k_m = \beta_2\Omega_s^2 \qquad [44]$$

where $\beta_2 = d^2k/d\omega^2$ is the group-velocity dispersion (GVD) coefficient at the pump frequency. In the description of the evolution of the Stokes and anti-Stokes waves, an effective propagation constant appears, given by

$$\kappa = \Delta k + \Delta k_{NL} = \Delta k_m + \Delta k_w + \Delta k_{NL} \qquad [45]$$

In eqn [45], Δk_{NL} represents a nonlinear contribution, given by

$$\Delta k_{NL} = 2\gamma P_p \qquad [46]$$

where $P_p = P_1 = P_2$ is the power in each pump wave and

$$\gamma = \frac{\bar{n}_2\omega}{cA_e} \qquad [47]$$

is the nonlinearity coefficient, $\bar{n}_2$ being the nonlinear-index coefficient.

Parametric Gain

The evolution of the Stokes and anti-Stokes waves are governed by the following coupled equations:

$$\frac{dE_3}{dz} = 2j\gamma P_p \exp(-j\kappa z)E_4^* \qquad [48]$$

$$\frac{dE_4^*}{dz} = -2j\gamma P_p \exp(j\kappa z)E_3 \qquad [49]$$

The general solution of eqns [48] and [49] is given by

$$E_3(z) = (c_1\exp(g_P z) + c_2\exp(-g_P z))\exp(-j\kappa z/2) \qquad [50]$$

$$E_4^*(z) = (c_3\exp(g_P z) + c_4\exp(-g_P z))\exp(j\kappa z/2) \qquad [51]$$

where the coefficients c_i are determined from the boundary conditions and the parametric gain g_P is given by

$$g_P = \sqrt{(2\gamma P_p)^2 - \left(\frac{\kappa}{2}\right)^2} \qquad [52]$$

In the range

$$-6\gamma P_p \le \Delta k \le 2\gamma P_p \qquad [53]$$

the radicand in eqn [52] is positive, so a gain occurs. The maximum gain is achieved when $\kappa=0$ and it is given by

$$g_{P\max} = \Delta k_{NL} = 2\gamma P_p = \bar{g}_P\frac{2P_p}{A_e} \qquad [54]$$

where

$$\bar{g}_P = \frac{\omega_p\bar{n}_2}{c} \qquad [55]$$

Considering that $\bar{n}_2 = 3.2\times10^{-20}\ m^2\ W^{-1}$ and assuming a pump wavelength $\lambda_p = 1\ \mu m$, we find that the gain coefficient $\bar{g}_P$ is larger by a factor of 2 than the peak of the Raman gain coefficient g_R. Consequently, the threshold for the parametric process is lower than with the Raman effect. However, the Raman effect dominates in general for long fibers, since it is difficult to maintain phase matching because of fluctuations of the core diameter. One can define a coherence length, given by

$$L_{coh} = 2\pi/\kappa_{max} \qquad [56]$$

where κ_{max} is the maximum tolerable phase mismatch. Significant four-wave mixing occurs only if $L < L_{coh}$.

In the case of a monomode fiber, Δk_w is very small and can be ignored. For λ_p below the zero of dispersion $\lambda_D \approx 1.3\ \mu m$ ($\beta_2 > 0$), one has to keep both Δk_m and Δk_{NL} small by using small frequency shifts Ω_s and low pump powers. When Δk_m dominates, we have

$$L_{coh} \approx \frac{2\pi}{|\Delta k_m|} = \frac{2\pi}{\beta_2 \Omega_s^2} \quad [57]$$

In the visible range, one has $\beta_2 \approx 50 - 60\ ps^2\ km^{-1}$, and $L_{coh} \approx 1$ km for $f_s = \Omega_s/2\pi \approx 50$ GHz. For a pump wavelength above the zero dispersion ($\beta_2 < 0$), the negative value of Δk_m must be compensated by Δk_{NL}. For phase matching one has

$$\kappa = 0 \approx \Delta k_m + \Delta k_{NL} = \beta_2 \Omega_s^2 + \Delta k_{NL} \quad [58]$$

From this, we obtain

$$\Omega_s = \sqrt{2\gamma P_p / |\beta_2|} \quad [59]$$

which coincides with the frequency shift characteristic of the modulation instability.

Amplifier Gain and Bandwidth

Four-wave mixing can be used, for example, in a fiber-optical parametric amplifier. In fact, the main difference between the four-wave mixing experiments and the parametric-amplification experiments is whether a signal at the phase-matched frequency is co-propagated with the pump or not. In the absence of a signal, the signal and idler waves are generated from amplifications of noise.

Let us consider the situation in which a signal wave with a power P_3 is launched at $z = 0$, and simultaneously an idler wave appears, with $P_4(z = 0) = 0$. In this case, considering the solution given by eqs [50] and [51] and taking into account eqn [52], we can write the following results for the signal and idler powers:

$$P_3(z) = P_3(0)\left[1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z)\right] \quad [60]$$

$$P_4(z) = P_3(0)\left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P z) = P_3(z) - P_3(0) \quad [61]$$

The single-pass gain of the parametric amplifier becomes

$$G_P = \frac{P_3(L)}{P_3(0)} = 1 + \frac{P_4(L)}{P_3(0)} = 1 + \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L)$$
$$= 1 + \frac{(2\gamma P_p)^2}{g_P^2}\sinh^2(g_P L) \approx \left(1 + \frac{\kappa^2}{4g_P^2}\right)\sinh^2(g_P L) \quad [62]$$

The approximation holds for $g_P L \gg 1$. According to eqn [52], amplification ($g_P > 0$) occurs only in the case:

$$|\kappa| < 4\gamma P_p \quad [63]$$

Since $|\kappa|$ must be small, this also means good phase-matching. If one has complete phase-matching ($\kappa = 0$) and still $g_P L \gg 1$,one obtains exponential behavior for the amplifier gain:

$$G_P = \sinh^2(g_P L) \approx \frac{1}{4}\exp(2g_P L) = \frac{1}{4}\exp(4\gamma P_p L) \quad [64]$$

From Eqs [44] and [45], we have that the phase mismatch κ and the frequency shift Ω_s are related by

$$\kappa = \beta_2 \Omega_s^2 + \Delta k_w + \Delta k_{NL} \quad [65]$$

From eqn [65] we have

$$\frac{\Delta\kappa}{\Delta\Omega_s} = 2|\beta_2|\Omega_s \quad [66]$$

In order to estimate the amplification bandwidth $\Delta\Omega_s$, we set

$$\Delta\kappa = \kappa_{max} = 4\gamma P_p \quad [67]$$

From Eqs [66] and [67] we obtain

$$\Delta\Omega_s = \frac{2\gamma P_p}{|\beta_2|\Omega_s} = \frac{2}{|\beta_2|\Omega_s L_{NL}} \quad [68]$$

where $L_{NL} = (\gamma P_p)^{-1}$ is the nonlinear length. Considering typical values $|\beta_2| = 20\text{–}60\ ps^2\ km^{-1}$, $\Omega_s/2\pi = 10\text{–}100$ THz, and $L_{NL} \sim 1$ m, we obtain a bandwidth in the range ~10–100 GHz. This bandwidth lies between that of a fiber Raman amplifier (~5 THz) and a fiber Brillouin amplifier (~100 MHz).

For $|\kappa| > 4\gamma P_p$ the parametric gain becomes imaginary and there is no longer amplification, but rather a periodical power variation of the signal and idler waves. In particular, if $|\kappa| \gg 4\gamma P_p$, then

$g_P \approx j\kappa/2$ and the parametric amplifier gain is

$$G_P = (2\gamma P_p L)^2 \frac{\sin^2(\kappa L/2)}{(\kappa L/2)^2} \qquad [69]$$

A convenient measure for the width in this case corresponds to a wavevector mismatch $\Delta\kappa = 2\pi/L$, where L is the fiber length. Using eqn [66], we then have

$$\Delta\Omega_s = \frac{\pi}{L|\beta_2|\Omega_s} \qquad [70]$$

Amplifier Performance and Applications

Parametric amplification can induce considerable crosstalk in WDM communication systems. Such crosstalk affects the system performance, particularly for channel spacings less than 10 GHz. Considering a channel spacing of 5 GHz and −20 dB allowed crosstalk, calculations show that the input power per channel should be smaller than 1 mW for a 15 km-long monomode fiber, and 0.1 mW for 100 km-long fiber. However, this effect can be avoided by an appropriate frequency separation of the channels.

Besides the limitation mentioned above, parametric amplification may be useful in the field of optical fiber communication systems. In fact, parametric amplifiers have a bandwidth of ~100 GHz, which is smaller that the bandwidth of fiber Raman amplifiers (~5 THz) but still sufficiently large for many applications. On the other hand, the frequency shift in FPAs can be as large as ~100 THz, which must be compared with the pump-signal frequency shift of ~13 THz of FRAs. These characteristics of FPAs allow a considerable flexibility in the choice of the pump source.

Parametric amplifier gains above 40 dB, with a 30-m-long fiber for input signal powers less than 1 mW and a pump power $P_p \approx 70$ W, were already measured. Such large gain values show clearly the potential of fiber parametric amplifiers provided phase-matching can be achieved. In fact, this condition puts stringent limits on the control of frequency shift Ω_s between the pump and signal waves. One method to satisfy the phase-matching requirement consists in using large birefringence of polarization-preserving fibers. Several parametric amplifiers using this kind of fiber have been demonstrated. Moreover, the possibility to realize a parametric laser was also demonstrated, by placing the fiber inside a Fabry–Perot cavity.

Fiber-optic parametric amplifiers have also attracted considerable attention in the context of a phenomenon known as squeezing. In contrast to coherent light containing equal minimum noise fluctuations in cophasal and quadrature components, in squeezed states one of these components acquires a noise fluctuation below the quantum-noise level at the expense of the other component. This phenomenon can occur in a parametric amplifier for certain values of the relative phase between the signal and idler waves. In fact, spontaneous emission at the signal and idler frequencies generates photons with random phases. Four-wave mixing increases or decreases the number of specific signal-idler photon pairs, depending on their relative phases. The noise reduction below the quantum-noise level can be verified using a phase-sensitive detection scheme in which the phase of the local oscillator is conveniently adjusted. The phase-sensitive nature of parametric amplifiers can also be usefully exploited to reduce the timing jitter in soliton communication systems.

List of Units and Nomenclature

A_e	effective fiber core area, m^2
c	vacuum speed of light, m s^{-1}
E_p	amplitude of the pump wave electric field, V m^{-1}
E_s	amplitude of the signal wave electric field, V m^{-1}
g_B	Brillouin gain coefficient, m W^{-1}
G_B	Brillouin amplifier gain
g_P	parametric gain, m^{-1}
G_P	parametric amplifier gain
g_R	Raman gain coefficient, m W^{-1}
G_R	Raman amplifier gain
$\vec{k}_a$	wave vector of the acoustic wave, rad m^{-1}
$\vec{k}_p$	wave vector of the pump wave, rad m^{-1}
$\vec{k}_s$	wave vector of the signal wave, rad m^{-1}
L	amplifier length, m
L_{coh}	coherence length, m
L_e	effective amplifier length, m
L_{NL}	nonlinear length, m
n	refractive index
$\bar{n}_2$	nonlinear-index coefficient, m^2 W^{-1}
p_{12}	longitudinal elasto-optic coefficient
P_{asp}	amplified spontaneous scattered power, W
P_p	pump power, W
P_s	signal (Stokes) power, W
v_a	acoustic velocity, m s^{-1}
ν_B	Brillouin frequency shift, s^{-1}
α_p	absorption coefficient at the pump frequency, m^{-1}

α_s	absorption coefficient at the signal frequency, m^{-1}
β_2	group-velocity dispersion coefficient, $s^2\ m^{-1}$
γ	nonlinearity coefficient, $W^{-1}\ m^{-1}$
Γ	inverse acoustic phonon lifetime, s^{-1}
δ	frequency detuning from Brillouin gain peak, s^{-1}
$\Delta\nu_B$	Brillouin gain bandwidth, s^{-1}
$\Delta\nu_p$	pump linewidth, s^{-1}
$\Delta\nu_R$	Raman gain bandwidth, s^{-1}
ε_0	free-space permittivity, $F\ m^{-1}$
ρ_0	average material density, $kg\ m^{-3}$
ϕ	nonlinear phase shift, rad
ω_a	acoustic wave frequency, rad s^{-1}
ω_p	pump wave frequency, rad s^{-1}
ω_s	signal (Stokes) wave frequency, rad s^{-1}
Ω_s	sideband frequency shift, rad s^{-1}

See also

Optical Communication Systems: Wavelength Division Multiplexing. **Nonlinear Optics, Basics:** Four-Wave Mixing. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (1992) *Fiber-Optic Communication Systems*, Ch. 8. New York: John Wiley & Sons.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn, Ch. 8–10. San Diego, CA: Academic Press.

Boyd RW (1992) *Nonlinear Optics*, Ch. 8–10. San Diego, CA: Academic Press.

Buck JA (1995) *Fundamentals of Optical Fibers*, Ch. 7 and 8. New York: John Wiley & Sons.

Cotter D (1987) Fiber nonlinearities in optical communications. *Optical Quantum Electronics* 19: 1–17.

Cotter D (1993) Nonlinearity in optical fiber communication. In: Eason RW and Miller A (eds) *Nonlinear Optics in Signal Processing*, Ch. 9. London: Chapman & Hall.

Forghieri F, Thach RW and Chraplyvy AR (1997) Fiber nonlinearities and their impact on transmission systems. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications – IIIA*, Ch. 8. San Diego, CA: Academic Press.

Sauter EG (1996) *Nonlinear Optics*, Ch. 5–11. New York: John Wiley & Sons.

Schubert M and Wilhelmi B (1986) *Nonlinear Optics and Quantum Electronics*, Ch. 11 and 12. New York: John Wiley & Sons.

Shen YR (1984) *The Principles of Nonlinear Optics*, Ch. 9–11, 14, 26. New York: John Wiley & Sons.

Semiconductor Optical Amplifiers

M J Connelly, University of Limerick, Limerick, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

There has been rapid growth in the deployment and capacity of optical fiber communication networks over the past 25 years. This growth has been made possible by the development of new optoelectronic technologies that can be utilized to exploit the enormous bandwidth of optical fiber. Today, systems are operational which operate at bit rates in excess of 100 Gb/s. Optical technology is the dominant carrier of global information. It is also central to the realization of future networks that will have the capabilities demanded by society. These capabilities include virtually unlimited bandwidth to carry communication services of almost any kind, and full transparency that allows terminal upgrades in capacity and flexible routing of channels. Many of the advances in optical networks have been made possible by the optical amplifier.

Optical amplifiers can be divided into two classes: the optical fiber amplifier and semiconductor optical amplifier (SOA). The former dominates conventional system applications such as in-line amplification to compensate for optical link losses. However, due to advances in optical semiconductor fabrication techniques and device design, the SOA is showing great promise for use in evolving optical communication networks. It can be utilized as a general gain element but also has many functional applications, including optical switching and wavelength conversion. These functions, where there is no conversion of optical signals into the electrical domain, are required in transparent optical networks.

We will review SOA basics, technology (materials and structures), signal transmission performance (pattern effects, crosstalk, and ultrashort pulse amplification), and some important functional applications (optical switching and wavelength conversion).

Basic Principles

The SOA is based on similar technology as a laser diode. Optical gain is achieved by electrically pumping a suitable semiconductor material, such that a population inversion occurs between the material conduction and valence bands. An incoming photon can then be amplified when the resulting stimulated emission exceeds losses due to stimulated absorption and internal material losses. SOAs can be designed to operate in either the 1300 nm or 1550 nm optical communication windows. The principle of operation of an SOA with low residual reflectivities ($R \approx 0$) is shown in Figure 1. An input optical power experiences a single-pass gain $G = \exp(gL)$ after traveling through the SOA active waveguide of length L. The net gain coefficient $g = \Gamma g_{\mathrm{m}} - \alpha_{\mathrm{int}}$, where Γ is the optical confinement factor (the fraction of the propagating signal power confined to the SOA waveguide), g_{m} the material gain, and α_{int} the optical loss coefficient. g_{m} is a function of the injected carrier (electron) density and wavelength. The single-mode active waveguide can support two orthogonal polarization modes: transverse electric (TE) and transverse magnetic (TM).

The amplification process also adds broadband noise, amplified spontaneous emission (ASE) to the output signal. The amplifier noise figure (NF) is a measure of the signal-to-noise ratio (SNR) degradation of the signal after amplification. It is defined as the ratio of the SNR of the input to that of the output when the input noise is shot noise limited. Expressions for the output SNR are usually calculated assuming that the SOA is followed by a narrowband optical filter and an ideal photodetector, such that the main source of detector noise is the beat noise between the signal and the ASE falling within the filter bandwidth. In this case the SOA noise figure is given by

$$\mathrm{NF} = 2n_{\mathrm{sp}}K \qquad [1]$$

with the excess noise factor K given by

$$K = \frac{(1 + RG)(G - 1)\Gamma g_{\mathrm{m}}}{(1 - R)G(\Gamma g_{\mathrm{m}} - \alpha_{\mathrm{int}})} \qquad [2]$$

where n_{sp} is the effective population inversion parameter. To achieve a low NF, the internal losses must be small and a value of n_{sp} close to one is required. A favorable value of n_{sp} is achieved by operating the SOA at high gain.

A low R is required to prevent the SOA from oscillating at high gains. The residual reflectivity is manifest as ripples in the amplifier gain spectrum and ASE spectrum.

SOA Structures

The key parameters required for a practical SOA are:

- low reflectivity ($<10^{-4}$) to ensure low gain ripple (<0.5 dB);
- low optical loss coefficient to achieve a high gain;
- high material gain to allow low drive current operation;
- low polarization sensitivity (<0.5 dB) because the polarization state of the optical signal coming from a link fiber is usually random;
- high saturation output power (P_{sat}), defined as the output power at which the gain is reduced by 3 dB; and
- low fiber-to-chip coupling losses (<3 dB per facet).

A schematic diagram of a commercial SOA chip is shown in Figure 2. The active waveguide consists of a 0.2 μm thick InGaAsP bulk active layer sandwiched between 0.1 μm thick InGaAsP separate confinement heterostructure (SCH) layers. The central section of the active waveguide is 600 μm long with a constant width of 1.4 μm. The mode-expanding active waveguide tapers are 150 μm long with a width that changes linearly from 1.4 μm to 0.4 μm at the tip. The tapers allow optical coupling to an underlying passive waveguide enabling efficient coupling to the input and output optical fibers. The structure provides a high confinement factor because of the refractive index mismatch between the layers in the gain section. The p-n junction formed by the p-InP

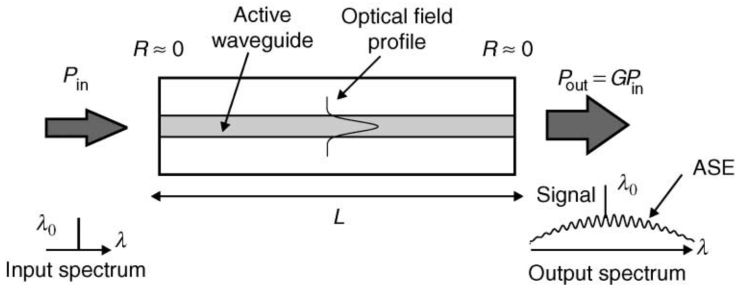

Figure 1 SOA basic structure.

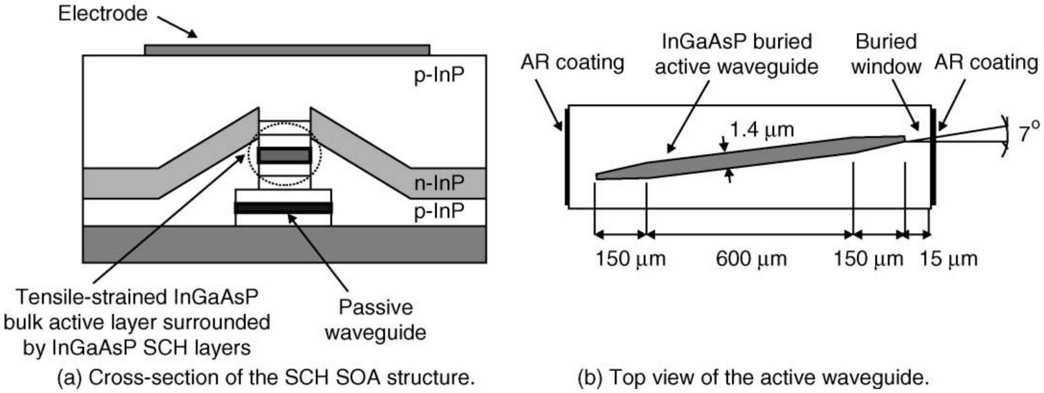
Electrode
p-InP
n-InP
p-InP
Tensile-strained InGaAsP bulk active layer surrounded by InGaAsP SCH layers
Passive waveguide
AR coating
InGaAsP buried active waveguide
Buried window
AR coating
1.4 μm
7°
150 μm
600 μm
150 μm
15 μm

(a) Cross-section of the SCH SOA structure.

(b) Top view of the active waveguide.

Figure 2 SCH SOA with tensile-strained bulk active layer.

Table 1 Typical specifications of a commercial SOA

Wavelength of peak gain	1550 nm
Peak fiber-to-fiber gain	22 dB
Noise figure	6.5 dB
Saturation output power	10 dBm
Polarization sensitivity	0.2 dB
Gain ripple	0.1 dB
3 dB bandwidth	40 nm
Drive current	200 mA

and n-InP layers acts as a current block, thereby providing good confinement of the injected carriers from the drive current to the active layer. Very low-reflectivity ($<10^{-6}$) is obtained by combining buried windows with antireflection-coated tilted facets (7° tilt angle).

Because of the active waveguide asymmetry, the TE confinement factor is larger than the TM confinement factor. g_m is isotropic in bulk material. The introduction of tensile strain, due to the lattice mismatch between the active layer and SCH layers, causes the bulk material band structure to change in such a way that the TM material gain is greater than the TE material gain. The introduction of the correct amount of tensile strain compensates for the difference in the TE and TM confinement factors leading to low polarization dependence.

Other SOA structures that can achieve low polarization dependency are based on active waveguides with near square cross-section having almost the same TE and TM confinement factors or the combination of compressively strained quantum wells (higher TE gain) and tensile strained quantum wells (higher TM gain). Typical specifications and characteristics of a commercial SOA are listed in Table 1 and shown in Figure 3.

As Figure 3 shows, gain saturation in conventional SOAs is manifest at output signal powers well below P_{sat}. Because the gain recovery time (carrier lifetime) in SOAs is fast (typically 0.1–1 ns), this can lead to pattern effects in single-channel systems and severe crosstalk in wavelength division multiplexed (WDM) systems. This problem can be greatly reduced by the use of a gain-clamped SOA (GC-SOA). In a GC-SOA, lasing action is produced, at a wavelength remote from the operating wavelength range, by the introduction of wavelength specific feedback. Once lasing begins, the carrier density in the SOA active layer is clamped at a fixed value. Changes in the input signal power lead to opposing changes in the lasing mode power. This has the effect of keeping the carrier density fixed (i.e., clamped), thereby making the signal gain relatively insensitive to changes in the total input power. A common method of providing this feedback is the use of distributed Bragg reflectors (DBRs) as shown in Figure 4. A typical GC-SOA gain versus output power characteristic is shown in Figure 5.

Basic Applications of SOAs in Optical Communication Systems

The three basic applications of SOAs in optical communication systems are: booster amplifier, in-line amplifier, and preamplifier. The main requirements of SOAs for such applications are listed in Table 2.

A booster amplifier is used to increase the power of a high power signal. Boosting laser power in an optical transmitter can be used to overcome external modulator losses, compensate for splitting and tap losses in optical distribution networks and to increase the power budget of optical links. The most critical requirement of a booster amplifier is a high P_{sat}, to obtain a high output signal power and minimize pattern effects. SOAs with P_{sat} in excess of 10 dBm are now available commercially.

The sensitivity of a conventional optical receiver is limited by thermal noise. The sensitivity is the

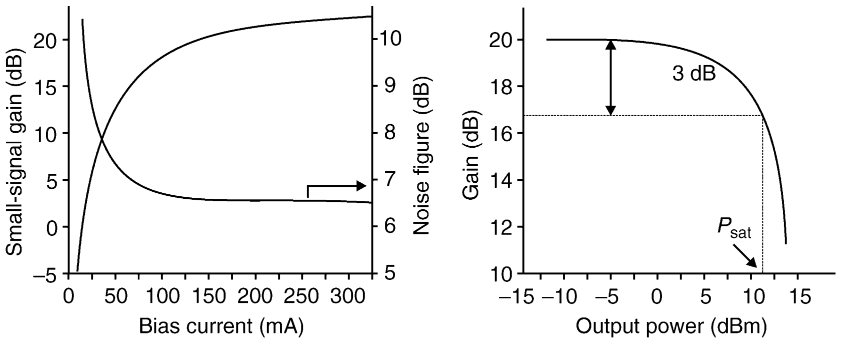

Figure 3 Typical commercial SOA small-signal gain and noise figure versus bias current and gain versus output power characteristics.

minimum signal power required at the receiver input to achieve a desired bit-error-rate (typically 10^{-9}). An optical preamplifier can be used to increase the power level of an optical data signal prior to detection and demodulation, leading to an increase in sensitivity. The performance of a preamplified optical receiver is dependent on the detector signal-to-noise ratio, $\mathrm{SNR} = i_{\mathrm{sig}}^2 / \bar{i}_{\mathrm{noise}}^2$. The signal photocurrent i_{sig} is proportional to the amplified signal power. The mean square noise current $\bar{i}_{\mathrm{noise}}^2$ includes circuit noise (thermal noise and detector dark current noise), signal shot noise, spontaneous emission shot noise, signal-spontaneous beat noise, and beat noise between the spectral components of the spontaneous emission. The best improvement in SNR occurs when the SOA is operated in the signal-spontaneous beat noise limit. In this regime, the signal power is sufficiently large such that the dominant receiver noise is the signal-spontaneous beat noise that falls within the signal bandwidth. This usually requires that the spontaneous emission from the SOA be reduced using a narrowband optical filter. The noise figure is a critical parameter in this application and should be as low as possible.

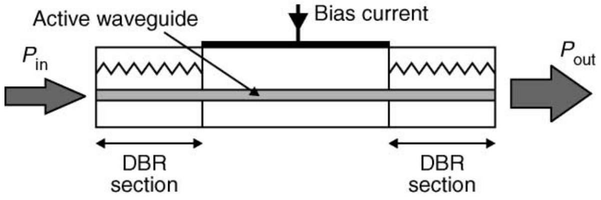

Figure 4 Schematic structure of a GC-SOA.

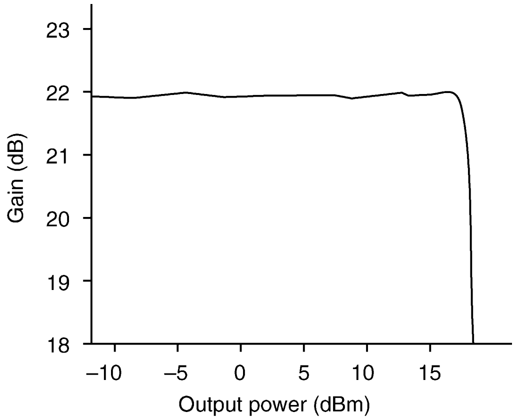

Figure 5 Typical GC-SOA gain versus output power characteristic.

In long-haul optical transmission systems, in-line optical amplifiers can be used to compensate for link losses thereby increasing the spacing between optical regenerators. The main advantages of in-line SOAs are transparency to data rate and modulation format (unsaturated operation), bidirectionality, WDM capability, simple mode of operation, low power consumption, and compactness. The latter two advantages are important for remotely located optical components. To avoid noise accumulation in the link, it may be necessary to follow each SOA by a narrowband optical filter, but this can prevent capacity enhancements using WDM.

An example of a single channel optical transmission system utilizing booster, in-line, and preamplifier SOAs is shown in Figure 6. The transmitter comprised a 1309 nm gain-switched laser diode directly modulated with a 10 GHz sinusoid to produce a train of 40 ps wide pulses at a repetition rate of 10 GHz. At 1309 nm fiber dispersion is small and the maximum transmission distance is mainly limited by the link losses. The laser output was connected to an external modulator, driven by a 2^{31}-1 pseudo-random bit sequence (PRBS) to produce an optical data stream with an extinction ratio of 13 dB. A booster SOA was used to increase

Table 2 Requirements of basic SOA applications

Requirement	*Booster amplifier*	*In-line amplifier*	*Preamplifier*
High gain	Yes	Yes	Yes
High P_{sat}	Yes	Yes	Not critical
Low noise figure	Not critical	Yes	Yes
Low polarization sensitivity	Not critical	Yes	Yes
Optical filter	Not necessary	Not critical	Yes

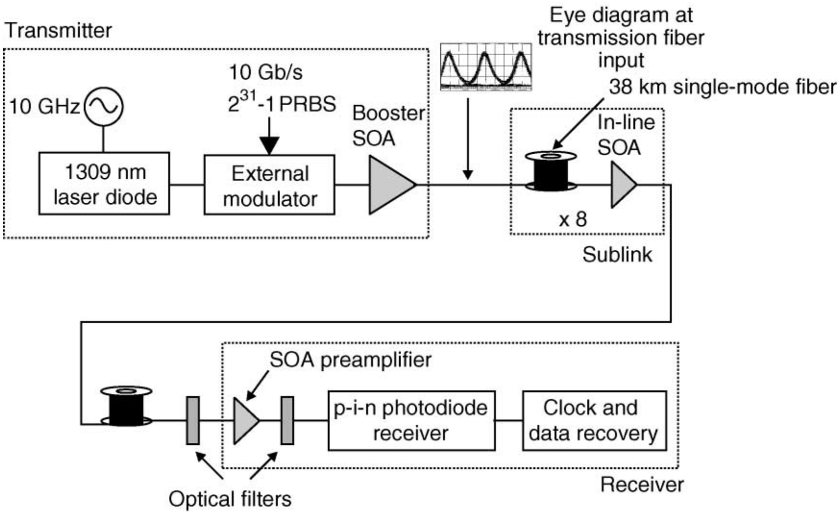

Figure 6 10 Gbit/s transmission experiment utilizing booster, in-line and preamplifier SOAs. (Adapted from Kuindersma *et al.* (1996) 10 Gbit/S RZ transmission at 1309 nm over 420 km using a chain of multiple quantum-well semiconductor optical amplifier modules at 38 km intervals. European Conference on Optical Communications 1996.)

the average transmitted power to between 0 and 2 dBm (7–9 dBm peak power). The transmission fiber length was 420 km with 12 in-line SOAs, used to compensate for fiber loss, spaced at 38 km intervals. At the receiver the signal was passed through a 1 nm bandpass filter, to reduce the accumulated spontaneous emission, amplified by an SOA preamplifier and filtered by a 1 nm bandpass filter. The signal was then detected by a p-i-n photodiode followed by an electronic clock and data recovery circuit. The SOA preamplifier and filter increased the receiver sensitivity from −14 dBm to −31 dBm for a BER of 10^{-10}. The receiver penalty after 420 km was 5 dB. In this experiment, the main limitation on transmission distance was the accumulation of spontaneous emission within the optical filter bandwidth.

Pattern Effects and Crosstalk

When an SOA is operated in the unsaturated regime, the amplifier gain is independent of the number of input signals and signal data rate. Outside this regime, the SOA will cause distortion because at high input powers, the gain saturates and compresses. Dynamic gain saturation occurs on the same time scale as the gain recovery time. This leads to pattern effects where the power of an incoming bit affects the gain experienced by subsequent bits. This is particularly important when the bit rate is of the same order as the inverse of the gain recovery time as shown in Figure 7. In WDM systems cross-gain modulation (XGM) between the channels can lead to severe interchannel crosstalk.

A further complication in WDM systems using SOAs is interchannel crosstalk caused by four-wave mixing (FWM). FWM is a coherent nonlinear process between two optical fields within the SOA, resulting in gain modulation at the beat frequency between the fields, and in the process generating new sidebands. In WDM systems with equally spaced wavelengths, the net effect of FWM is the generation of crosstalk terms interfering with the desired signals. Although the crosstalk can be relatively low in power terms, it can produce significant power penalties due to the coherent beat noise phenomena. The level of FWM crosstalk increases as the channel spacing decreases and the channel output power increases.

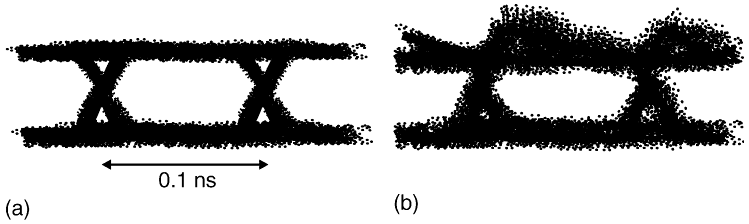

Figure 7 Typical eye diagrams of 10 Gb/s nonreturn to zero data (a) before and (b) after amplification by an SOA. ASE noise is present in the output eye diagram.

FWM crosstalk is of particular importance in dense WDM systems where the inter channel spacing is less than 100 GHz. Figure 8 shows a typical SOA output spectrum for an 8-channel multiplex with FWM-generated crosstalk signals.

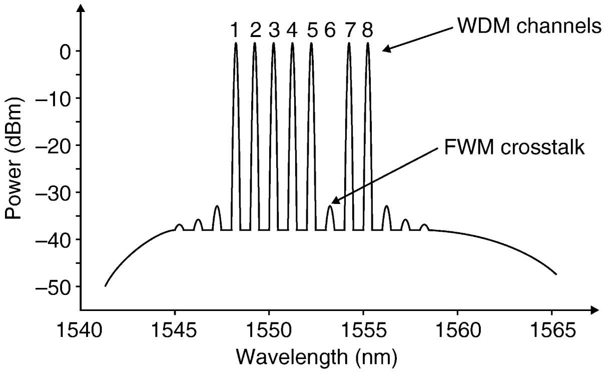

Figure 8 Typical SOA output spectrum for an 8-channel WDM multiplex with 100 GHz spacing. Channel 6 has been dropped to show the FWM crosstalk signal caused by channels 5 and 7.

Ultrashort Pulse Amplification

Optical time division multiplexing is an efficient method for increasing the bit rate of optical transmission systems. This requires the time interleaving of short optical pulses. The transmission distance is limited by the fiber group velocity dispersion, which is proportional to the pulse spectral width. Soliton transmission is also possible where the pulse shape is preserved as it propagates in the fiber.

Because an SOA has a very large bandwidth (typically 5 THz) it is capable of amplifying pulses as short as 100 fs. An input pulse can be amplified without significant distortion if the pulse energy is much less than the saturation energy E_{sat} of the SOA. Typical SOA saturation energies are of the order of a few pico Joules. As the pulse energy approaches E_{sat}, considerable spectral broadening and distortion can result. For an input pulse width τ_{p} (full width at half maximum) of the order of 10–100 ps spectral broadening is primarily due to self phase modulation (SPM). SPM is caused by gain saturation, which leads to intensity-dependent changes in the SOA active layer refractive index in response to carrier density variations. The degree of spectral broadening and distortion also depends on the input pulse shape. If τ_{p} is much less than the carrier lifetime, the output pulse power $P_{\text{out}}(t)$ and phase $\phi_{\text{out}}(t)$ are given approximately by

$$P_{\text{out}}(t) = P_{\text{in}}(t)\exp[h(t)] \qquad [3]$$

$$\phi_{\text{out}}(t) = \phi_{\text{in}}(t) - \frac{1}{2}\alpha h(t) \qquad [4]$$

with

$$h(t) = \left[1 - \left(1 - \frac{1}{G_0}\right)\exp\left(\frac{-U_{\text{in}}(t)}{E_{\text{sat}}}\right)\right]^{-1} \qquad [5]$$

$$U_{\text{in}}(t) = \int_{-\infty}^{t} P_{\text{in}}(\tau)\mathrm{d}\tau \qquad [6]$$

where $P_{\text{in}}(t)$ and $\phi_{\text{in}}(t)$ are the input pulse power and phase, respectively. G_0 and α are the amplifier unsaturated gain and linewidth enhancement factor, respectively. Typical values for α are in the range 2–10 and depend on the amplifier active region material and operating conditions. The output pulse spectrum can be obtained from

$$S(\omega) = \left|\int_{-\infty}^{\infty}[P_{\text{out}}(t)]^{1/2}\exp[i\phi_{\text{out}}(t) + i(\omega - \omega_0)t]\mathrm{d}t\right|^2 \qquad [7]$$

where $\nu_0 = \omega_0/2\pi$ is the pulse optical frequency. The output pulse chirp (frequency variation) is given by

$$\Delta\nu_{\text{out}}(t) = -\frac{1}{2\pi}\frac{\partial\phi_{\text{out}}}{\partial t} = \Delta\nu_{\text{in}}(t) + \frac{\alpha}{4\pi}\frac{\partial h}{\partial t} \qquad [8]$$

where $\Delta\nu_{\text{in}}(t)$ is the input pulse chirp. Using the above theory the shape, chirp and spectrum of an amplified zero-chirp (transform limited) Gaussian pulse of input power

$$P_{\text{in}}(t) = \frac{E_{\text{in}}}{\tau_0\sqrt{\pi}} \exp\left(-\frac{t^2}{\tau_0^2}\right)$$

is shown in Figure 9. E_{in} is the pulse energy and $\tau_p \approx 1.665\tau_0$. The amplified pulse is asymmetric because the leading edge of the pulse experiences a larger gain than the trailing edge. The amplified pulse spectrum is broader than the input pulse and also develops a multipeak structure. This is due to the SPM-induced frequency chirp imposed on the pulse as it propagates through the amplifier. In this case the chirp is linear around the peak of the pulse power spectrum. It can be compensated by transmitting the pulse through an optical fiber with anomalous group-velocity dispersion. In practice optical pulses can be far from Gaussian and can also have an initial chirp. In this case the induced chirp and resulting pulse spectrum can be more complex and more difficult to compensate.

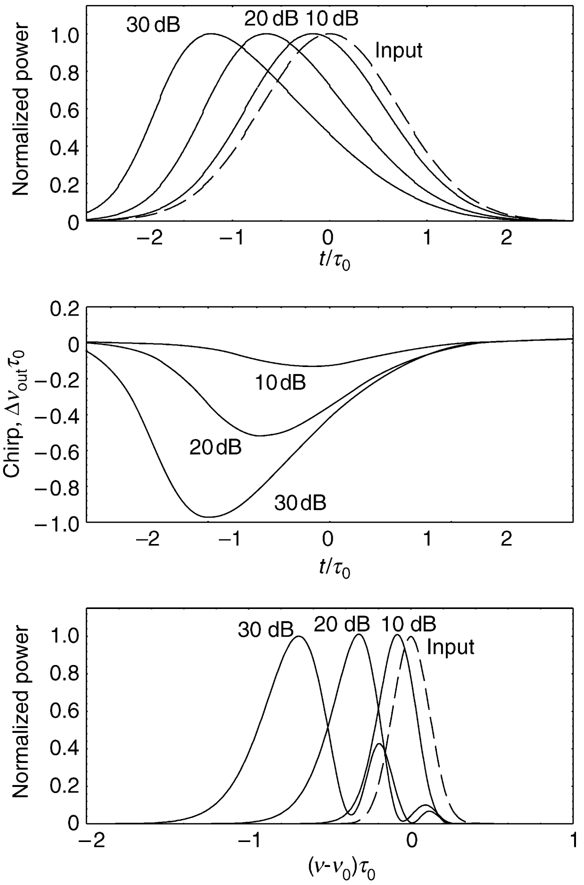

Figure 9 Amplified pulse shape, frequency chirp and spectrum for a transform-limited Gaussian input pulse with $E_{\text{in}}/E_{\text{sat}} = 0.1$ and $\alpha = 5$. ν is optical frequency. The parameter is the amplifier unsaturated gain.

When the input pulse width is less than ~10 ps, the above theory is no longer adequate and other nonlinear effects within the SOA such as carrier heating and spectral hole burning must be taken into account. The resulting pulse power and spectral distortions induced by an SOA on such pulses can be very complex.

Functional Applications

SOAs can be used to perform functions that are particularly useful in optically transparent networks. Recent advances in photonic integrated circuit and optoelectronic device packaging technology have made the deployment of SOA functional elements feasible. Two of the most important applications of SOAs are optical switching and wavelength conversion.

An example of a 2 × 2 SOA switch module used for optical routing is shown in Figure 10. Larger switch matrices can be constructed using this basic element. The module is based on a hybrid structure consisting of four integrated SOAs mounted on a silicon submount. The SOA-array is aligned to input and output polymer waveguides through V-grooves and alignment indentations. An incoming data packet can be routed to any output port by switching on the appropriate SOA. Switching times of the order of 1 ns are possible.

Ultrafast switching (<100 fs) can be achieved by incorporating SOAs in non-linear loop mirror structures, such as the terahertz optical asymmetric demultiplexer (TOAD) shown in Figure 11. Switching is achieved by placing an SOA offset from the center of an optical fiber loop mirror and injecting data into the loop via a 50:50 coupler. The two

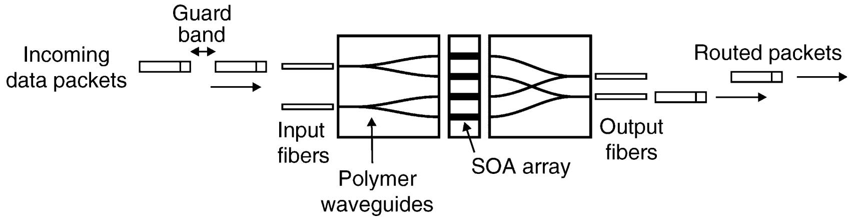

Figure 10 2 × 2 hybrid SOA switch module.

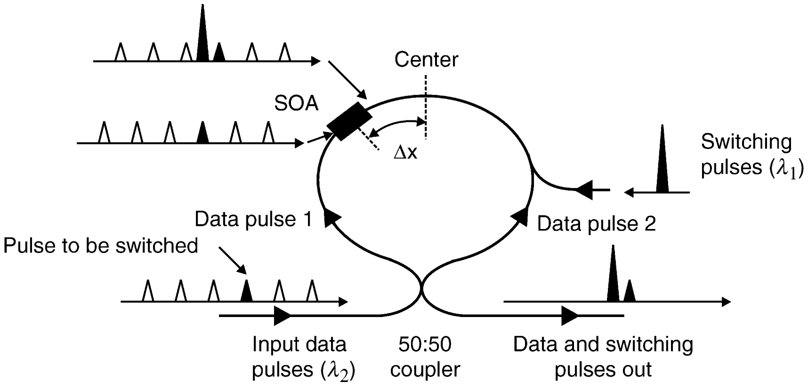
Center
SOA
Δx
Switching pulses (λ_1)
Data pulse 1
Data pulse 2
Pulse to be switched
Input data pulses (λ_2)
50:50 coupler
Data and switching pulses out

Figure 11 Ultrahigh speed optical switching using a TOAD.

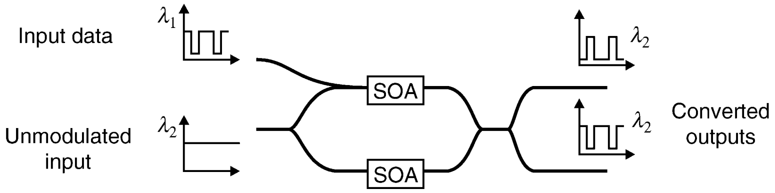

Figure 12 SOA Mach–Zehnder wavelength converter.

counter-propagating data pulse streams arrive asynchronously at the SOA. A switching pulse is timed to arrive after one data pulse but just before its replica. The switching pulse power is adjusted to impart a phase change of π radians onto the replica, so the data pulse is switched out when the two counter-propagating components interfere on their return to the coupler. The TOAD can also be used to demultiplex high-speed time division multiplexed pulse streams.

All-optical wavelength converters will play an important role in broadband optical networks. Their most important function will be to avoid wavelength blocking in optical cross-connects in WDM networks. Wavelength converters increase the flexibility and capacity of a network using a fixed set of wavelengths and can be used to centralize network management. In packet switching networks, tunable wavelength converters can be used to resolve packet contention and reduce optical buffering requirements.

Wavelength conversion in SOAs can be achieved using XGM, cross-phase modulation (XPM), or FWM. The design of an SOA wavelength converter based on XPM requires that one or more SOAs be incorporated into an interferometer, an example of which is the Mach–Zehnder interferometer (MZI) shown in Figure 12. An input data signal at wavelength λ_1 is used to modulate the upper SOA refractive index, which controls the phase shift experienced by a second unmodulated input signal at λ_2 in one of the interferometer arms. When the input data signal is low (logic 0), the arms are in phase and the λ_2 signal appears at the top output. When the input pump signal is high (logic 1) it produces an extra π phase shift between the two arms, causing them to be totally out of phase and the λ_2 signal to appear at the bottom output. Both outputs contain a copy of the original data signal imposed onto the new wavelength. Up and down conversion is possible. The lower SOA is used to equalize the gain experienced by the λ_2 signal in each of the arms. An optical filter is required at the output to remove λ_1. Such wavelength converters can operate at very high bit rates (>10 Gb/s). An important advantage of the MZI structure is that it also gives 2R regeneration (re-amplification and reshaping) of the input data signal with wavelength conversion. This is because the interferometer nonlinear response increases the converted signal extinction ratio compared to the input data. All-optical 3R (2R + retiming) regenerators are of great interest in future optical communication networks to restore degraded transmission signals. SOA-based interferometric structures can be used to achieve 3R (2R + retiming) regeneration at data rates >80 Gb/s.

See also

Coherent Lightwave Systems. **Lasers:** Semiconductor Lasers. **Optical Amplifiers:** Basic Concepts; Erbrium Doped Fiber Amplifiers for Lightwave Systems; Optical

Amplifiers in Long-Haul Transmission Systems. **Optical Communication Systems:** Optical Time Division Multiplexing; Wavelength Division Multiplexing. **Semiconductor Physics:** Band Structure and Optical Properties.

Further Reading

Agrawal GP and Olsson NA (1989) Self phase modulation and spectral broadening of optical pulses in semiconductor laser amplifiers. *IEEE Journal of Quantum Electronics* 25: 2297–2306.

Connelly MJ (2002) *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic.

Desurvire E (1994) *Erbium-Doped Fiber Amplifiers: Principles and Applications*. New York: Wiley.

Durhuus T, Mikkelsen B, Joergensen C, Lykke Danielsen S and Stubkjaer KE (1996) All-optical wavelength conversion by semiconductor optical amplifiers. *Journal of Lightwave Technology* 14: 942–954.

Ghafouri-Shiraz H (1995) *Fundamentals of Laser Diode Amplifiers*. New York: Wiley.

Shimada S and Ishio H (1992) *Optical Amplifiers and their Applications*. New York: Wiley.

OPTICAL COATINGS

Contents

Anti-Counterfeiting and Decorative Coatings
Diamond Optical Devices and Coatings
Laser Damage in Thin Film Coatings
Optical Black Surfaces
Thin Film Optical Coatings
X-Ray Coatings

Anti-Counterfeiting and Decorative Coatings

R W Phillips and R L Bonkowski, Flex Products, Inc., Santa Rosa, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Scientists are well known for the diversity of their interests. One topic that captures the interest of virtually all scientists, and people in general, is money. Not only are people engaged in obtaining money, but they are also fascinated about the manufacture of money and the protection of its value.

The relationship between manufacturing techniques and the protection of value documents (currency, stock shares, bonds, etc.) is not accidental. Technology has always been employed to counter the contemporary threats of counterfeiters. These threats vary, depending on the skill set and resources available to the criminal, as well as the counterfeiting strategy adopted. While we refer to banknotes repeatedly in this chapter, the information applies, as well, to passports, stamps, tax banderoles, identification cards, visas, licenses, checks, and bankcards. Banknotes can be considered the most challenging application for optical anticounterfeiting technologies. Stamps, banderoles, and checks are single-use items, and do not have to be as durable as currency. Identification cards, passports, and the like are used many times but are generally in protective covers and are examined by trained personnel. Currency must endure the rigors of commerce and travel, while still being easily authenticated by the average citizen.

The ever-increasing availability of low-cost computer hardware and software, along with color photocopiers and printers, has stimulated the use of optical and electro-optical technologies that were previously applied in a limited fashion or not at all. These optical techniques include: transmitted light devices, such as watermarks and threads; luminescent materials; reflective, interference, and diffraction schemes; modulation techniques (Moiré, for example); and light polarization effects.

Each of these concepts either has been or is likely to be applied to value documents to prevent counterfeiting. Each has its own set of potential benefits and application challenges. Also, some of these effects have been successfully applied to commercial decorative products. It is important to understand the difference between commercial and security uses. Effects aimed at commercial uses are mainly decorative or aesthetic, whereas security devices are functional and integrated into a security strategy. To ensure economic success, decorative products must be easily manufactured and widely distributed. In contrast, security products are carefully controlled in both their availability and the diffusion of manufacturing expertise.

Historical Perspective

Paper currencies (banknotes) have been printed for hundreds of years. Their history has been marked by a succession of technical changes in an effort to prevent counterfeiting. Sophisticated manufacturing techniques, such as intaglio printing and watermarks, along with special papers, made it difficult for counterfeiters to produce acceptable imitations. Intaglio is the centuries old method of printing, using incised or engraved metal plates. The tremendous pressure of the intaglio press forces the paper into an ink-coated plate engraving, forming patterns of ink on the paper.

In the eighteenth and nineteenth centuries, banknote printers often used unique engravings of guilloche patterns (intricate, overlapping curves), vignettes, and portraits as anticounterfeiting devices. These sophisticated patterns effectively discouraged early, relatively unskilled, counterfeiters. The development of photography and lithographic printing in the late nineteenth century created new opportunities for counterfeiting. Photochemical technology made it possible to produce suitable reproductions using print shop facilities throughout the world.

Photochemical techniques had significant limitations, however. The printed notes were passable but not high-quality imitations of genuine intaglio notes. Making lithographic plates required specific skill, considerable equipment, and left behind meaningful evidence (films, plates, press blankets, etc.) for police to find.

Until the 1980s, counterfeiting detection methods by governmental officials were relatively uncomplicated, as three general perpetrators could be classified in fairly predictable ways (Table 1).

Until the advent of home computers, optical scanners, color printers, and copiers, the challenge for law enforcement was difficult but focused. Amateur counterfeiting was virtually nonexistent. The effort required to make a few notes was daunting, the risk high, and potential profit meager. This meant that anticounterfeiting police efforts could be concentrated on criminal counterfeiters. Government-sponsored counterfeiting was certainly known (both the Germans and the British attempted to de-stabilize their opponents' national economies during World War II), but this was more an issue of foreign policy, diplomacy, and espionage and not the main province of national police.

The situation changed quickly and radically in the late 1980s. Color photocopiers had improved in cost, quality, and size to the point that commercial establishments were buying hundreds of machines. The 1990s introduced numerous new technologies to seduce the casual counterfeiter. These included home computers, graphics editing software, color printers, and low-cost scanners.

The amateur counterfeiter now had an array of sophisticated tools that could copy, edit, and reproduce virtually any kind of document. Banknotes made by the techniques of 'digifeiting' (a term coined by Dr. Sara Church of the US Treasury Department) were not of high quality, but passable in small quantities in low risk situations. It was now possible, economical, and relatively safe to make one or a dozen bills, especially in low denominations.

Police efforts were now severely challenged as the proliferation of amateur counterfeiters made investigation and prosecution costly. Especially in the United States, average citizens were lured by the prospect of augmenting their incomes by printing one or two notes a week. When apprehended, effective prosecution and punishment of such small-stakes criminals was expensive and unpopular.

Large-production criminal counterfeiting was augmented by new digital and electro-optical

Table 1

Class of threat	*Typical volume of notes*	*Resources and skills (1980)*
Amateur	Small (<100)	Retail paper, ink. Household chemistry
Criminal	Medium to large (1000 to 1 000 000)	Litho press, darkroom, plate making, cutting equipment. Skilled printing personnel
State-sponsored	Large (millions)	Intaglio press, currency paper. Skilled engravers and printers, Major distribution networks

breakthroughs. Lithographic plates could now be made without cameras or chemistry. Laser ablated plates made directly on the printing press (such as Heidelberg/Presstek) accurately duplicated features as small as 10 or 15 microns. Further, computer-to-plate technology with erasable plates presented an unusual opportunity for counterfeit production in small press shops.

The use of new electronic and optical tools by amateur and criminal counterfeiters coincided with a rise in state-sponsored counterfeiting. The United States, in particular, was victimized in the early 1990s by high-quality, intaglio-printed notes. One such note, dubbed the 'super $100' was allegedly produced in huge volumes in the Middle East and circulated widely in Europe.

Various means of detecting counterfeits were promoted but remain largely ineffective today. The so-called 'detector pen' applies a small amount of iodine solution onto a banknote. The iodine reacts with starch (often used to stiffen paper to give it a banknote 'feel'), turning the pen mark from light brown to black. Unfortunately, the presence or absence of starch in the paper is not a foolproof criterion. US currency paper contains no starch, while the currencies of other countries, such as Canada, have a starch component.

Concurrent with the availability of new high-tech counterfeiting tools, optical science has been hard at work to create protection schemes, utilizing the latest optical effects and manufacturing technologies. These new technologies have been largely outside the skill set and resources of amateur, criminal, and even state-sponsored counterfeiters.

An analysis of all the types of optical materials used for anticounterfeiting purposes as well as for decorative purposes shows that all are based on the control of electromagnetic radiation in the ultraviolet (UV), visible, and infrared. Exploiting various degrees of light transmission, reflectance, absorption, or scatter controls this radiation. In particular, all anticounterfeit devices fall into one or more of five basic categories: light patterning, light re-emission, light diffraction, polarization, or light interference. Characteristics of an effective anticounterfeiting solution include:

- quick verification by the average person;
- ability to function in a variety of lighting situations;
- deterrence of scanning and electronic photocopying;
- compatibility with existing banknote production schemes and equipment;
- ability to be verified by machines;
- durable and reliable despite strenuous handling conditions;
- tight control of materials and manufacturing processes.

Devices Based on Light Patterning

Watermarks

Images can be created in the paper substrate during the manufacturing process. Watermarks have been in use for at least 500 years and are produced by the arrangement of the paper fibers in the cylinder-mold paper machine or by impression in wet paper in the Fourdrinier paper machine. Because they are not on the surface of a document, watermarks are not readily visible in reflected light, but become apparent when light is transmitted through the paper substrate. Credible counterfeits, however, have been produced using images created by imprinting paper with waxy or oily substances. The fatty 'ink' is not readily detected on the surface of the document and serves to change the light transmission characteristics of the paper.

Watermarks serve chiefly to authenticate the paper substrate and to deter the practice of 'raising' notes, that is, bleaching a genuine low-denomination note and re-printing it as a higher value. The most effective watermarks are produced by the cylinder-mold process and use an image that is a duplicate of that image printed in ink (Figure 1) elsewhere on the document. This gives the examiner a reference point to compare the watermark as, for example, on all high-denomination US currency.

Watermarks are an effective counterfeiting deterrent when the general population is educated about their presence and their function. The limited opportunity to expose banknotes to a source of transmitted light (outdoors or in a dimly lit restaurant, for example) somewhat reduces their effectiveness.

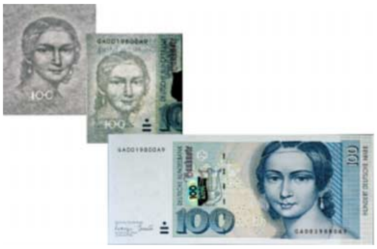

Figure 1 Watermark portrait of Clara Schumann, famous pianist and wife of composer Schumann: far left – portrait in transmitted light; center – portrait in reflected and transmitted light; and right – banknote in reflection. Courtesy of Papierfabrik Louisenthal.

Threads

Thin strips of polymer or metallized polymer can be imbedded in the paper substrate during the paper-making process. Such strips are approximately 12 microns thick and 1–5 millimeters wide. These strips are known as threads because Crane & Company developed the original concept in the 1860s, using colored silk thread. Typically, threads are installed border-to-border in the narrow direction of the banknote. Because they are inside the paper, they are not seen in reflected light, but are readily visible when light is transmitted through the substrate. Such threads are effective deterrents against counterfeiting achieved by scanning or photocopying, as the scanner uses only reflected light and cannot duplicate the buried thread. Threads have evolved into several varieties:

- De-metallized threads: Thin polymer coated with an opaque metal. The metal is selectively removed to create text or images.
- ‘Window’ threads: Metallized polymer that is ‘woven’ in and out of the paper to permit observation in both reflected and transmitted light (Figure 2).
- Fluorescent threads: Polymer impregnated with ultraviolet inks that emit a specific color under UV irradiation (Figure 3). See the section ‘Devices Based on Light Re-emission’ later in this article.
- Diffractive threads: Metallized polymer strips that have been embossed with diffractive or holographic patterns and attached by hot-stamping or adhesive to the surface of the note. See the section ‘Devices Based on Light Diffraction.’

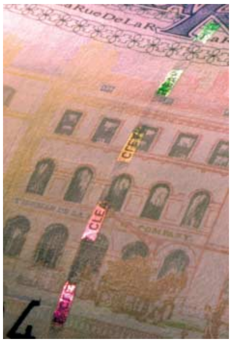

Figure 2 Windowed thread. Reproduced with permission from De La Rue International Limited.

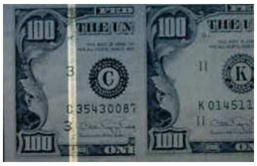

Figure 3 Fluorescent thread embedded into US $100 note. Reproduced with permission from Crane and Co., Inc.

Potential drawbacks to the use of threads are durability and the temptation of the general public to try to extract them from the banknote. Also, some threads can be simulated by printing or by attaching a polymer film to the surface of the banknote.

Laser Perforation

A modern ‘watermark’ concept is laser perforation of the banknote substrate. Precise patterns of microscopic holes (ca. 100 microns diameter) can be drilled using a high-power laser apparatus. The 200-Franc Swiss banknote features denomination numbers that are not readily visible in reflected light. However, when a light source is held behind the note, the numerals fabricated by laser perforation are clearly visible (Figure 4). In addition to being easily authenticated by non-expert observers with no special viewing apparatus, laser perforation has forensic advantages. The penetrations made by laser beams have a particular shape, spacing, and dimension, which are not easily mimicked. By clever adaptation of laser scanning software, holes of different shapes can be made and inter-mixed for additional security complexity. As a result, manual

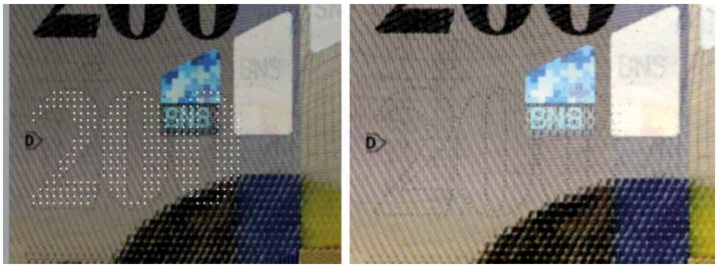

Figure 4 Microperf$^{®}$ laser holes in banknote. Left image in transmitted light; right image in reflected light. Reproduced with permission from Orell Füssli Security Printing Ltd.

pin pricks and mechanically pierced holes can be discriminated as forgeries.

Reflective Foils

One of the earliest defenses against photocopying and digital scanning was the use of reflective metallic foils affixed to the surface of the document. Such foils yield >80% reflection during scanning and the resultant scanned image is usually a black field.

Attachment of foils to currency paper or documents is carried out by applying a thermoset adhesive under heat and pressure and releasing the optical film from its carrier web by means of a separate polymeric release layer. This process is known as hot-stamping. Alternatively, an ultraviolet curing adhesive can be used to affix the foil to the paper substrate using a cold roll-on process.

Foiling may impede the amateur counterfeiter; since decorative foiling of stationery and packaging is well known, it presents little challenge to the professional perpetrator.

Planchettes

Planchettes are small flakes or fragments a few millimeters in diameter and 10–25 microns thick. These are often colorful or reflective materials and can feature a variety of effects including iridescence, diffraction, fluorescence, magnetic properties, or thermochromic properties. These fragments are attached to the banknote substrate by an adhesive (Figure 5). Planchettes discourage scanning and photocopying, since the copier cannot duplicate their special properties. Unfortunately, planchettes have been found to be non-durable and easy to simulate. The generally random placement makes it difficult for an examiner to authenticate a genuine note.

Thermochromic Inks

Inks that react to temperature changes have been employed on checks and identity cards. Thermochromic inks fluctuate from one color to another when the temperature of the printed image is raised. While it is possible for body temperature to activate this change, the effect is more certain if infrared heat is used. Since the process requires several seconds to activate (depending on the heat source), thermochromic materials have not been widely used on banknotes where immediate authentication is necessary.

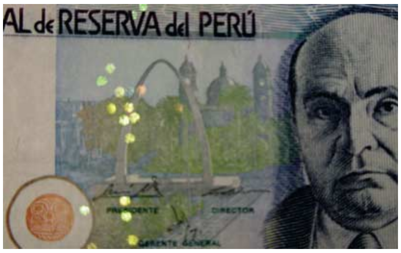

Figure 5 Peruvian 100 Nuevo Soles note showing iridescent planchettes. Reproduced with permission from Flex Products, Inc.

Colored Fibers

Extremely fine fibers (ca. 15 microns diameter) have been incorporated into currency papers for decades. The United States, for example, uses a few hundred grams of red and blue fibers per ton of paper in making currency. While of some forensic interest, the amount and size of the fibers makes them nearly undetectable by the average citizen.

Alternatively, fibers may be impregnated with fluorescent materials, or with absorbers that selectively absorb the wavelengths of light typically used in copy machines, producing obvious forgeries when copied.

Moiré Effects

Images can be created with very fine details that produce aliasing errors when digitized. Aliasing is the creation of new and spurious patterns through the superposition of two patterns whose periodicities are unequal. It represents interaction between the spatial frequency of the genuine document and the sampling frequency of the copying device. Aliasing is commonly known as moiré in the printing industry. Typical moiré generating patterns are constructed as closely spaced concentric circles, often found around the eyes in portraits on banknotes. When scanned digitally, the resulting copy will contain spurious images (scan traps) that are easily identified as fraudulent.

For example, if a genuine document is designed in a 1000 dpi (dots per inch) printing process it can later be scanned on a 1000 dpi scanner and, theoretically, all the data can be reproduced accurately. If, however, the scanner and the document are not in synchrony, the scanner will miss pieces of data, depending on the registration error, resulting in a digitized image that does not resemble the original.

In the scanned section of a US $5 note (Figure 6), one can see patterns of curved lines intersecting in the background, whereas the authentic image consists of fine parallel lines that do not intersect. The curvature and intersections are artifacts of the moiré created by digital scanning.

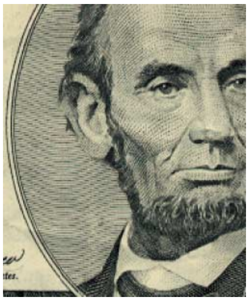

Figure 6 Scanned section of US $5 note showing moiré effect. Authentic image has no intersecting lines. Courtesy of Flex Products, Inc.

Modulation Techniques

Images produced by offset lithography are based on patterns (screens) of fine dots. By varying dot density, an array of shadings and tones can be created. More progressive modulation techniques, such as amplitude modulation (varying dot sizes) and frequency modulation (varying dot spacing) are often used to produce high-resolution images. To deter counterfeiting, banknotes can be printed using pixels, which are lines of equal length but assorted angular orientation. This technique is known as screen angle modulation (SAM).

SAM techniques can be used to defeat digital scanners by preventing the scanning of all available information due to under-sampling, or by concealing hidden patterns which are not demodulated by the human eye, but can be detected with suitable filters (Figure 7). A variant of this technology, called Scrambled Indicia®, has been used to protect US postage stamps (Figure 8).

Devices Based on Light Re-emission

Ultraviolet Fluorescent Inks

Ultraviolet (UV) inks are widely used on banknotes as semicovert deterrents. Images printed in UV inks are challenging to detect in normal lighting situations but are readily visible under UV (black light) illumination. Such printed images require a special light source for authentication and are subject to degradation over time from the handling and chemical exposure of banknotes. Typically, these products are sensitive to either short UV at 254 nanometers or long UV at 366 nanometers.

UV inks provide good protection against photocopying but can be readily simulated by more advanced counterfeiters.

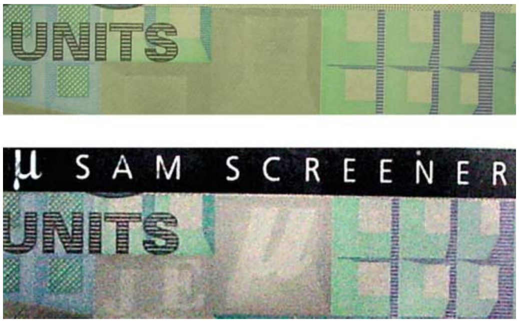

Figure 7 Top image – hidden image as normally seen; Bottom image – Image decoded with visible line screen. Courtesy of Joh Enschedé Security Printing Ltd.

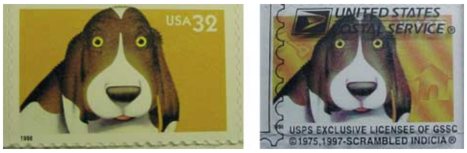
USA32
UNITED STATES POSTAL SERVICE
USPS EXCLUSIVE LICENSEE OF GSSC
©1975,1997-SCRAMBLED INDICIA®

Figure 8 Hidden image in 32¢ US postage stamp decoded by invisible line screen (left image – normal view; right image – with decoder in place). Courtesy of Flex Products, Inc.

UV Fibers

Several countries mix minute quantities of UV luminescent fibers into the paper substrate. In most cases, these are randomly placed and invisible to the naked eye.

Devices Based on Light Diffraction

Diffractive structures range from simple periodic surface relief structures (textured surfaces) to multilevel holographic structures that form a continuum of products ranging from decorative coatings to anticounterfeit devices. Figure 9 shows a generalized structure for all known types of reflective-based relief structures.

Most devices are based on reflection; however, in a few instances where clear polymer is used, devices may also be based on diffractive light transmission. The sizes, shapes, and orientation of the grooves vary according to the exact device. The opaque aluminum layer insures that the design works in reflected light. Most of these structures exhibit a rainbow effect where the color moves from short to long wavelength as the viewing angle increases. This rainbow effect can be found for all types of structures with grooves that have sizes near that of light wavelengths. The color shifts in a direction opposite to that of a thin film interference structure. Further, the rainbow effect can only be seen in a direction perpendicular to the groove direction. In the special case where the groove dimensions are less than the wavelength of light, the

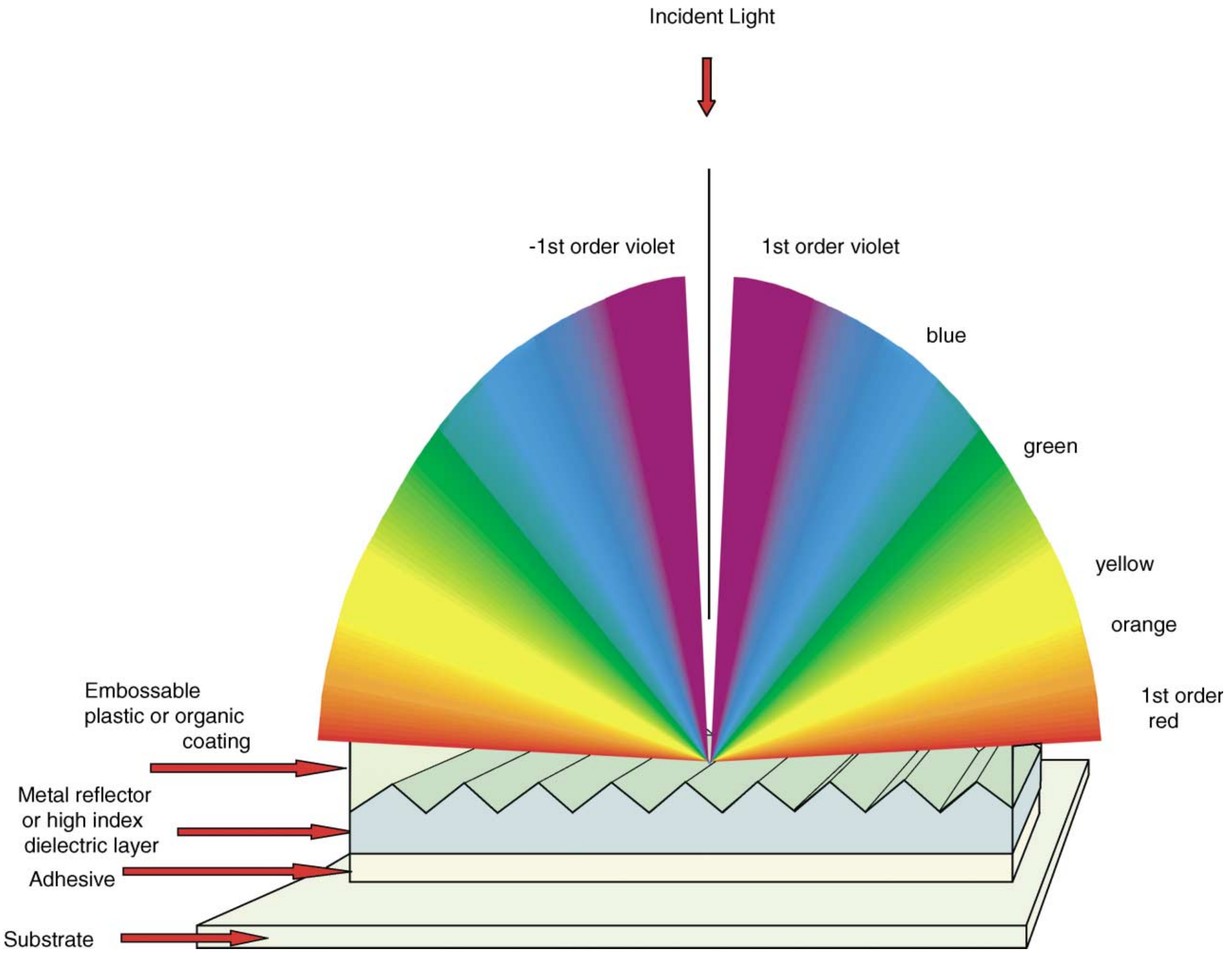

Figure 9 Cross-section of diffraction device. Courtesy of Flex Products, Inc.

rainbow effects disappear and are replaced by another type of diffraction (zero-order diffraction), where only specific colors are seen in reflection. Groove height and width for all diffraction products range from 0.1 to 10 microns.

For a unique set of discrete angles and for a given spacing d, the diffracted light from each facet is in phase with the light diffracted from any other facet, so they combine constructively:

$$Gm\lambda = \sin\alpha + \sin\beta \qquad [1]$$

where $G = 1/d$ is the groove density or pitch, α is the angle between the incident light and the normal to the grating, β is the angle between the diffracted beam and the normal to the grating and m is an integer called the diffraction order.

For $m = 0$, $\beta = -\alpha$ for all λ, the grating acts as a mirror and the wavelengths are no longer separated. This is called specular or zero-order reflection.

In contrast to grating structures (linear grooves) that only modify light intensities, holographic structures are complex surface contours that modify both the phase and intensity of the reflected light. Light reflecting off a metallized holographic structure is capable of producing an image in either two dimensions (2D) or three dimensions (3D). Heat and pressure may emboss the relief structure into a metallized polymeric film by hot-stamping. Alternatively, an embossing metal roller presses the relief image into a thermoplastic film or thin ultraviolet curable coating. The embossed structure is subsequently metallized with opaque aluminum or with a high-index dielectric material such as zinc sulfide or zirconium dioxide. The latter layers allow printed information on the substrate to show, while at the same time provide a holographic image. However, such high-index holograms are not as bright, since some of the reflective light is lost in transmission. Polymethylmethacralate (PMMA), polycarbonate (PC), polyvinyl chloride (PVC), and polyester (type G) are typical materials that can accept such an embossing.

Thick metallized embossed plastic films of ca. 50 microns are used for decorative labels. Thin metallized embossed hardcoats of ca. 7 microns, which can be released from carrier sheets or embossed thin plastic films, are used for the security field.

Holograms used for high-security applications have one characteristic that distinguishes them from ordinary holograms or other diffractive devices used for decorative purposes. This characteristic finds itself in the complication of images, multilevel images at different focal depths, different images that appear at various viewing angles, images of 3D objects superimposed with synthetic 2D images, micro-images, or images that appear to move as the device is tilted back and forth. Specific types of diffractive structures are described below. The list starts with simple holograms used in decorative markets and progresses to high-security devices.

Diffractive grating: A structure comprised of square or sinusoidal linear grooves that breaks light into its component colors. An assemblage of abutting islands, each randomly oriented and with its own parallel grooves, gives the viewer a twinkling color effect (Figure 10).

2D hologram: A relief structure that, upon viewing with a point source of light, shows a 2D image of objects, text, logos, or lines that emerge over a background. Such holograms are often referred to as 'cartoon images' since they are images of original drawings (Figure 11). Because of their simplicity and frequent use as decorative packaging, these holograms are generally not used in the security market.

Multilevel 2D hologram: A relief structure that shows a simulated 3D effect. This is accomplished by creating an image with two or more layers of flat artwork, with each layer at a slightly different focal depth (Figure 12). These images are well suited for labels, cards, and other decorative products. They have also been used on coupons, gift certificates, and tickets.

3D hologram: A relief structure that reconstructs a 3D image from a real or computer-synthesized 3D object (Figure 13). A 3D dove has been used on the VISA credit card as a security device for many years, though many now consider it more of a marketing logo than a security feature.

2D/3D hologram: A relief structure that reconstructs an image showing both a 3D and a 2D object, created from a 3D sculpture and flat art drawings (Figure 14). The artwork is created in such a way that the images appear at different depths of view.

Stereo-hologram: A relief structure that reconstructs a 3D image such that upon tilting from side to side one can see around the object, and additional features can then be viewed. At any given angle, it appears to be a 3D hologram. The image is created through multiple exposures of a 3D object during the holographic mastering process.

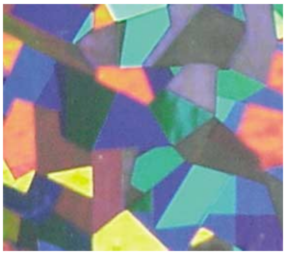

Figure 10 Picture of a typical diffractive assemblage of gratings. Reproduced with permission from Amagic Holographics Inc./K-Laser Technology, Inc.

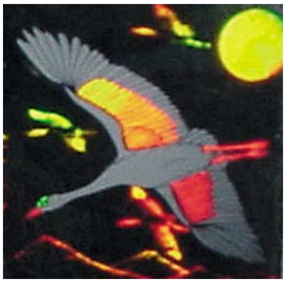

Figure 11 2D hologram. Courtesy of American Bank Note Holographics, Inc.

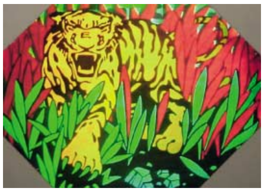

Figure 12 Multilevel, 2D hologram. Courtesy of American Bank Note Holographics, Inc.

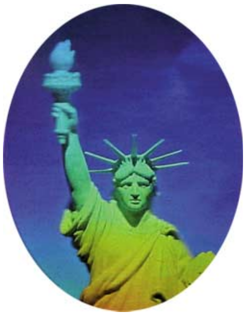

Figure 13 3D hologram. Courtesy of American Bank Note Holographics, Inc.

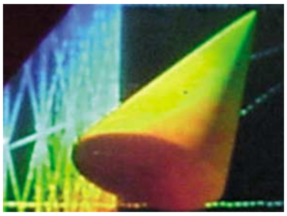

Figure 14 2D/3D hologram. Courtesy of American Bank Note Holographics, Inc.

Figure 15 True color hologram. Reproduced with permission from Light Impressions International, Inc.

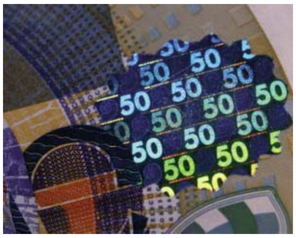

Figure 16 High-definition security hologram. Number 50 appears in various colors in first angle of view. Colored shells appear at another angle of view. Reproduced with permission from DeLaRue International Limited.

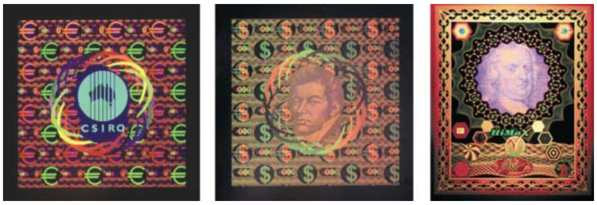

Figure 17 Left and center images – Exelgram™ at two different viewing angles. Reproduced with permission from CSIRO. Right image – HiMax® OVD high security device showing portrait of Sir Isaac Newton. Reproduced with permission from Light Impressions International, Inc.

True color hologram: In contrast to the holograms described above, a true color hologram is created by diffraction angle control of the colors red, green, and blue from a 3D object or color transparency. The relief structure, when viewed in white light, reveals a picture in true color, as opposed to the rainbow effect seen in regular 2D and 3D holograms (Figure 15).

High-definition security hologram: A relief structure that produces an image composed of multiple elements, such as fine line text and complex patterns that simulate movement as the device is tilted back and forth. Different images also appear at different angles (Figure 16). These images occur because different images are recorded in the holographic mastering process at different angles. The complicated imagery and the use of multiple images is the major holographic technique in use today for anticounterfeiting.

High-definition diffraction gratings: The relief structure in Figure 17 shows a multitude of elements, each with colors that change upon tilting. It consists of a very fine line grating structure, 1000 lines or greater per millimeter. These gratings are sometimes combined with holographic images that change depictions upon tilting, thus providing a high level of security.

Kinegrams®: This grating structure, based on computer-generated graphic images, is composed of light diffracting microstructures of less than one micron, with blaze (saw-tooth) contours that when viewed display optokinematic movements. This security device was first used in 1987 on Austrian 5000 Schillings. When the orientation of the blaze grating changes across the device, two distinct images are possible, depending on the view with respect to the blaze angles. The images do, however, require an aluminum reflective layer deposited onto the relief structure in order to generate the highly reflective colored images.

A 180° rotation of the device can lead to a marked change in the diffraction properties and, consequently, the image properties. This stands in contrast to holographic symmetrically based structures that yield identical images. An example of a Kinegram® security device is shown in Figure 18. The letter 'K' on the front facet of the cube is filled with an asymmetric structure, whereas the background is realized by using the same structure rotated by 180°. When viewed from one side, the letter 'K' appears bright on a dark background. Upon rotation in its plane by 180°, the contrast between the letter 'K' and the background is reversed.

Zero-order diffraction grating: This relief structure is quite different from holographic or normal grating structures, as it is based on an abrupt change in the

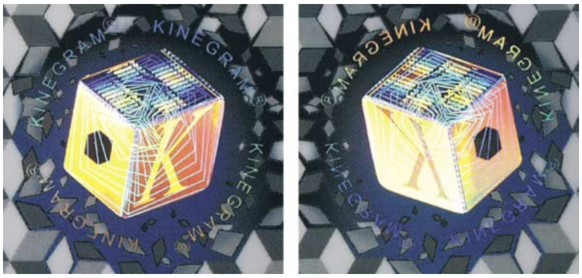

Figure 18 The Kinegram® security device as seen in two views differing from each other by a rotation of 180° in its plane, while illumination and viewing conditions remain unchanged. Reproduced with permission from OVD Kinegram Corporation.

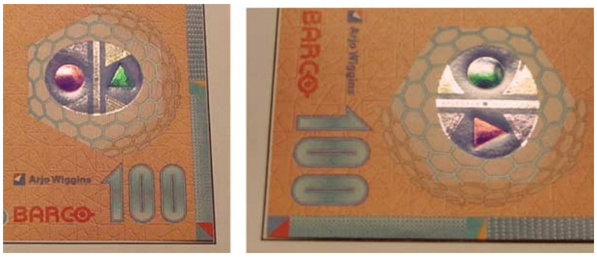

Figure 19 Zero-order diffraction images in a DID® security device: two views 180° apart in planar rotation. Reproduced with permission from Hologram Industries.

refractive index from approximately 2.3 to 1.5. Rather than employing an aluminized layer for reflection, the structure uses a thin, high-index layer ($n = 2.3$, for example) embedded in an organic polymer with a refractive index of about 1.5. The high-index layer appears as a parallel series of lines, approximately 0.2 microns wide, with another parallel series of lines beneath the first set, but oriented under the spacing of the first set of lines at a depth of about 0.2 microns.

This structure not only changes color as the device is tilted back and forth but also as it is rotated. In rotation, the color changes from red to green (Figure 19). Further, the light is polarized as it reflects. When a linear polarizer is inserted between the eye and the device, the color changes to black. The colors seen are based on specular zero-order reflection, unlike standard diffraction gratings where the reflected color is based on first-order reflection. The preeminent effect is when the device is placed over a black background. Over a white background the colors are subtler. Viewing reflection in zero order allows one to see the reflected colors in both point and diffuse lighting. The combination of polarization effects and color in rotation is ideally suited to high-security applications.

Devices Based on Light Interference

Thin-Film Foils

There are two types of thin-film interference devices. The first is based on an all-dielectric filter, and the second is based on a Fabry–Perot resonance cavity structure of the type absorber/dielectric/reflector. The all-dielectric filter uses pairs of high and low refractive index materials at a quarter-wave thickness. A typical filter is the one used initially on Canadian currency that employed the design: ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW/SiO_2 3QW/ZrO_2 1QW, where the QW (quarter-wave optical thickness) is centered at 585 nm. This design exhibits a gold-to-green color shift with viewing angle.

A Fabry–Perot filter also has a gold-to-green color shift and typically uses the design Cr 5 nm/MgF_2 4QW @ 604 nm/Al (opaque). This filter has slightly higher color saturation than the all-dielectric filter. The design is opaque and, therefore, the color observed is independent of the substrate color onto which it is affixed. This stands in contrast to the all-dielectric filter that must be laid over a black background for optical color saturation. If the all-dielectric filter is placed on a white background, the transmitted light passing through the optical filter will reflect off the white background and combine with the filter's reflection to give white light. However, with a varied colored background, the reflected color will vary depending on the combination of reflected color from the substrate and the reflected color from the filter.

The gamut of reflected colors from a metal–dielectric foil covers all four-color quadrants in the CIE Lab color space. Brightness of the foil depends on the materials used and on the design. Figure 20 shows the theoretical gamut of colors as plotted on the a^* b^* plane for the design, using the materials Cr, MgF_2 and Al.

Foils based on thin film interference layers contain a reflectance peak property moving from long wavelength to short wavelength as the viewing angle is increased. This is true across the UV, through the visible and into the near-infrared. To the human eye, the color moves from red to blue. The anti-counterfeit foil shown in Figure 21 exemplifies this effect.

Shifts from long to short wavelength occur because the path light difference (see the bold line in Figure 22) between the rays of light that reflect from the surface and the reflector are a function of the cosine of the incidence angle.

The path length difference is shown by the term $2d \cos \theta$, where θ is the angle of incidence within the

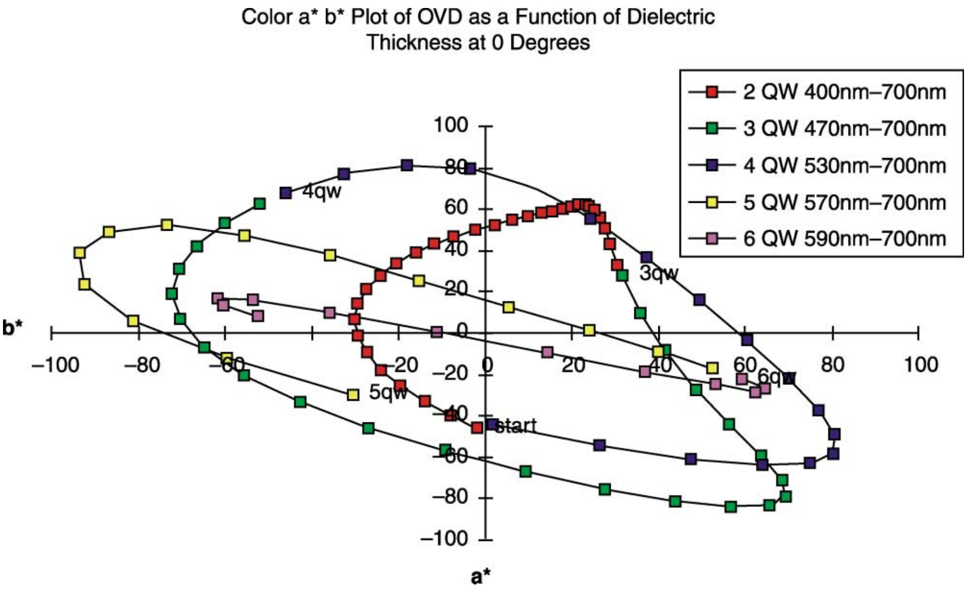

Figure 20 Gamut of colors at 0° incidence viewing under Illuminant C produced from Fabry–Perot designs as a function of quarter-wave optical thickness. Courtesy of Flex Products, Inc.

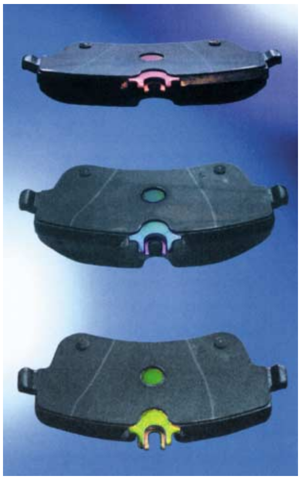

Figure 21 Anticounterfeit interference foil on brake shoes. Reproduced with permission from November AG.

filter and d is the distance between the surface of the filter and the reflector. The angle θ and the angle ϕ are related by Snell's law: $n_0 \sin \phi = n_1 \sin \theta$.

Light interference occurs where the wavefronts superimpose on one another. The fractional number of wavelengths that can fit into this path length difference is given by $\Psi = 2n_1 d \cos(\theta)/\lambda$, where λ is the wavelength of light. Waves reflected from these two surfaces will interfere constructively when Ψ is an integral number of wavelengths and interfere destructively when Ψ is at multiples of half wavelengths. As the viewing angle (ϕ) increases, the angle θ also increases. Thus, constructive interference

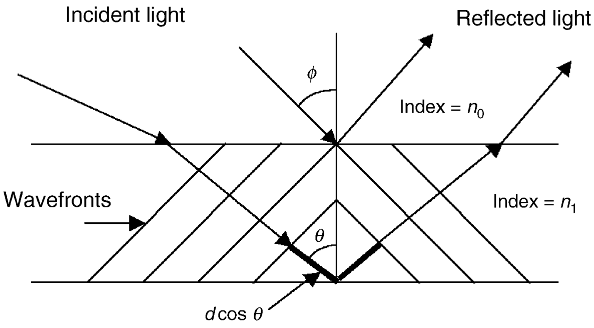

Figure 22 Diagram showing the principle of optical shift with viewing angle. Courtesy of Flex Products, Inc.

translates into a reflectance peak that moves toward shorter wavelengths as the angle of viewing is increased due to its cosine dependence.

The Fabry–Perot filter can use any transparent dielectric (colored or clear), a reflective metal, although optical metals are preferable since they have the highest reflectance (e.g., Ag, Al, Ni, Cu), and an absorber that is partially transmissive and partially reflective. Thin metals, absorbing dielectrics, and metal nitrides are typical materials used as an absorbing layer.

Light Interference Inks

Inks based on interference pigments have been developed over the last thirty years, starting with pigments based on high-index thin film coated micas (typically TiO_2 coated micas) and culminating with

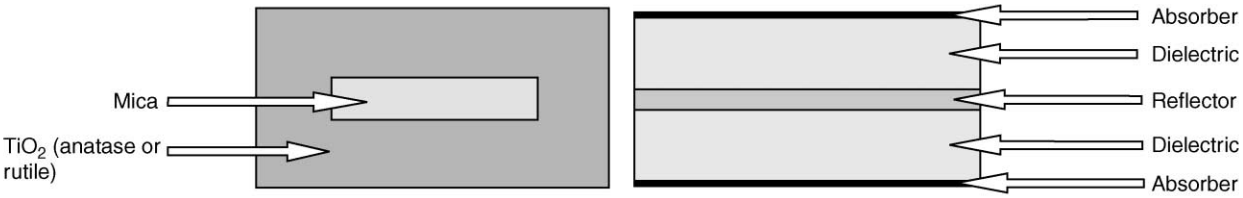

Figure 23 Left image – cross-section of interference mica; right image – cross-section of optically variable pigment (OVP®). Courtesy of Flex Products, Inc.

OVP® pigments based on the Fabry–Perot structure (Figure 23).

Inks created from micas are known as pastel or pearlescent inks, whereas inks formulated from OVP® are known as optically variable inks (OVI®). The latter inks are used as a primary defense against counterfeiting, defeating color copying and color printing from computer images. The difficulty with using interference micas for security features is that they can be readily purchased in the commercial world. In contrast, OVP® is a tightly controlled material and cannot be purchased by the general public.

OVI® security ink has been formulated into intaglio, silkscreen, flexographic and gravure applications. Its use on security documents encompasses over 75 banknotes worldwide, including the newly issued Euro notes, post-1996 \$10, \$20, \$50, and \$100 US banknotes, and banknotes for China. With the use of such inks, postage stamps, tax stamps, identification cards, and other security documents have also been effectively protected against counterfeiting. Color control for OVI® is assured because OVP® can be blended according to additive color theory to produce a desired color and color shift. Two examples of OVI® – the 100 Pula Botswana note and the new 100 Euro note – can be seen respectively in Figures 24 and 25.

OVI® inks have been extremely successful for banknote protection and might even be a *de facto* standard, because they meet the seven characteristics of an effective anticounterfeit solution as described earlier.

Devices Based on Light Polarization

The use of light polarization as an anticounterfeit feature has taken several avenues, one as coatings or pigments, the other as invisible images that are detected with a linear polarizer. With the exception of the zero-order diffraction device described earlier, polarization devices are based on cholesteric liquid crystals, using either nematic (planar oriented) layers or helical twisted layers.

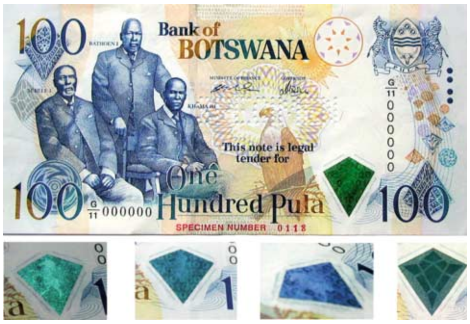

Figure 24 100 Pula Botswana note showing green-to-blue color shift OVI® with latent images. Reproduced with permission from SICPA.

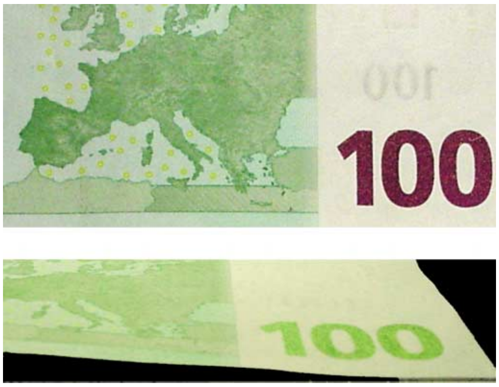

Figure 25 100 Euro note showing magenta-to-green OVI®. Courtesy of Flex Products, Inc.

Liquid crystal pigments are constructed from cholesteric polysilozanes that have been oriented and cross-linked in place with photopolymerization. With this cross-linking, pigments are relatively stable to temperature changes. These materials can form helical structures in either right or left twist orientation. Such materials reflect 50% of a band of wavelengths, centered at wavelength, λ_0, of the incoming light as circularly polarized light and 50% of the band centered at λ_0 as transmitted, circular polarized light, but with the opposite twist. The remaining wavelengths pass through the structure with no interaction. The amount and type of the chiral phase (components that inherently orient into a helical structure) in the polymer determine the pitch of the helical structure. This, in turn, determines the wavelength of reflection, i.e. $\lambda_0 = nP$ where n is the average refractive index and P is the pitch length of the material (pitch length determined by 360° rotation of the helix). At tilted viewing angles, the color changes according to the formula $\lambda = \lambda_0 \cos\theta$ where θ is the viewing angle relative to the vertical, and λ_0 is the reflected wavelength seen perpendicular to the surface. By the judicious combination of left- and right-handed helical based layers, or by inserting a half-wave retarder layer between two layers having the same helical structure, the resulting pigments reflect nearly 100% at λ_0.

When these materials are properly processed and ground (particles 2–50 microns thick, with an aspect ratio of 3:1 in lateral/thickness dimensions), particles with the helical structure perpendicular to the planar surfaces can be realized. Such pigments, or liquid crystal polymeric coatings, when printed, have color-shifting properties with viewing angles that defeat color copiers and the like, since toners have no color-shifting properties (Figure 26). These pigments also have an additional security feature, namely, that when viewed through a circular polarizer having the opposite twist, the color-shifting pigments appear black.

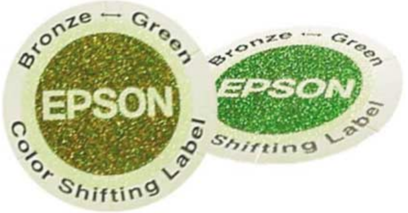

Figure 26 Anticounterfeit label showing liquid crystal coating viewed straight-on and at tilt angles. Courtesy of Flex Products Inc.

A special technology known as linear photo-polymerization (LPP) is a process to produce photo-patterned liquid crystal polymer (LCP) $\lambda/4$ phase retarders, which, when combined with a series of layers involving a reflector, and a linear polarizer, shows positive and negative images. Without the diagnostic linear polarizer, however, the image is invisible (Figure 27).

Adding a nematic liquid crystal layer, which exhibits a helical twist configuration, extends this technology. The latter technology can produce colored images when viewed with a linear polarizer. Such devices can be coated or hot-stamped onto a substrate containing viewable information.

Combination Devices

Beginning in 2000, combinations of holographic structures with liquid crystal or interference thin films were developed, which could make counterfeiting even more difficult. Such structures emerge as a hologram floating on a color shifting background. Magnetic layers, integrated into part of the optical interference stack, allow spatial orientation and 3D effects during printing.

Decorative Coatings

Coatings Based on Diffractive Pigments

When pigments based on diffractive effects are created, individual flakes have surface contours of groove-like structures of 2000–3500 lines mm^{-1}. In spray-out coatings using such pigments, diverse interplay of subtle rainbow colors results (Figure 28). The effect is particularly striking in full sunlight as, for example, on curved surfaces found on cars and motorcycles. Some manufacturers have successfully used finely shredded diffraction gratings

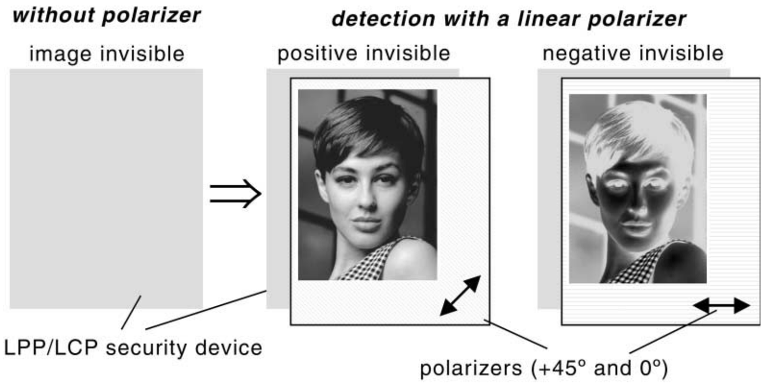

Figure 27 Reflective optical LPP/LCP security device showing a complex LPP/LCP retarder image visualized with one polarizer; rotation of the polarizer changes the image from positive to negative. Reproduced with permission from Rolic Research, Ltd.

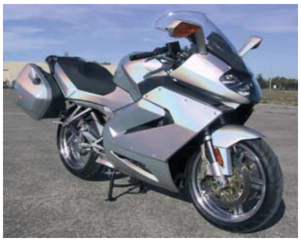

Figure 28 Motorcycle painted with diffraction-based pigments. Courtesy of Flex Products, Inc.

(ca. 25 microns thick) as flakes mixed in paints for decorative purposes. The use of such thick flakes seriously limits the applicability and uniformity of the coating.

Coatings Based on Interference Pigments

Pigment-based interference effects are widely known and have been used for decades in the paint, plastics, and cosmetic industries. These structures are based on TiO_2 coated mica, where the mica thickness is between 300 and 600 nm with planar diameters ranging from 1 to 200 microns, depending on the application.

Depending on the thickness of the TiO_2, different reflective colors of silver, gold, red, blue, and green are achievable. For maximum color reflection, a black substrate or viewing at a particular angle is required so as to prevent dilution of the reflected color from light reflected from the substrate surface. Colored mica pigments are platelets of mica coated with colored oxides. Metallic-like colors are also formed from iron oxide coated micas. In the first instance, two colors may be seen: the platelet color (due to absorption) and the interference color at one angle. Only the interference color may be seen at another viewing angle. If the substrate is colored, the color switches from the interference color to the substrate color as the viewing angle changes. More recently, laminar pigments with an all-dielectric optical stack have been produced with a low-index central core, encapsulated with high-index layers that cause interference.

Interference pigments based on absorber/dielectric/metal/dielectric/absorber have also been introduced into the decorative marketplace. These structures are based on aluminum as the central layer, a dielectric spacer layer, and either a wavelength selective absorber layer such as iron oxide (Fe_2O_3) or a nonwavelength selective absorber such as chromium (Cr). In Fe_2O_3 based pigments, the face-on colors are mainly found in the yellows and reds whereas for the Cr-based pigments, a full color gamut is possible.

Diffractive Foils

These foils are based on diffractive and grating surfaces where the contours are generally less than 2000 lines mm^{-1}. These materials show a glittering effect, i.e., multiple points of color, as the viewing angle is changed.

Interference Polymeric Films

These films are made from multilayer extrusions of pairs of polymers of 50 or more layers. Each pair is composed of two different polymers with a slightly different refractive index. Each layer is approximately one-quarter wave in thickness. Due to the small refractive index differences between layers, the resultant film is approximately 25 microns thick. Uniform color at a given viewing angle is generally not possible due to the lack of thickness control. Hence, they tend to look like oil slicks, with each color changing as the viewing angle changes (Figure 29). Interference polymeric films have found use as decorative gift wrapping material.

Holographic Foils

As indicated earlier, holograms designed for non-security markets may range from simple 2D holograms to more complex ones that function for either decorative or brand protection purposes. Another application of holographic foils is in holographic fabrics, whereby small holograms approximating several millimeters in diameter are affixed in a pattern to the cloth (Figure 30).

Aluminum Flake Based Inks and Coatings

These inks are composed of leafing or nonleafing aluminum powders and a polymeric vehicle. The use of leafing pigments produces images and coatings of high brilliance, whereas nonleafing pigments produce a matte or satin-like finish.

Figure 29 Visual effect from multilayer extruded polymeric film. Reproduced with permission from Englehard Corporation.

Summary and Challenges for the Future

Optical scientists have shown tremendous creativity in discovering and employing clever technical devices in the fight against counterfeiting of valuable documents. Optical solutions have been created that meet the functional as well as the economic needs of the document security industry. A cadre of research institutions and manufacturing companies has accepted the challenge of protecting the instruments of our economic and physical security.

This vigilance will have to continue in the future. Nearly as fast as new defenses are developed, criminals are illicitly using new tools to defeat the protection schemes. Financing for sophisticated counterfeiting efforts is likely to increase, whether sourced by international terrorists or drug dealers. Likewise, the more we move toward an information society, the more we depend on machines to count, sort, and verify documents and money. This machine-driven trend will mandate new features, especially in banknotes and cards.

Figure 30 Fabrics showing holographic effects. Reproduced with permission from Amagic Holographics, Inc./K-Laser Technology, Inc.

It is likely that the future of optical anticounterfeiting technology will be strongly influenced by:

- Active, rather than passive, devices. That is, devices that can be applied to a document and later stimulated or interrogated by a machine to elicit a specific effect or signal.
- Schemes that may be merged with new biometric technologies, especially for individualized documents, such as passports and identification cards. In such applications, it will be necessary to confirm that a document is genuine, and that it is in the possession of its rightful owner.
- The need to clearly differentiate security applications from decorative or aesthetic uses. Effective security products must always be carefully controlled and their availability limited.
- The willingness of the public to accept regular changes in the appearance of value documents. Currency will have to undergo redesign on a three-to-six year cycle, rather than the long cycles typical in the past.

This chapter was intended only as a survey of the issues involved in anticounterfeiting protection. To learn more about the optical methods discussed here, or to learn more about the science of anticounterfeiting, please consult the "Further Reading" section and the Internet sites listed in Table 2.

Table 2 Internet sites: Banknote Design & Protection

Sponsor	*Web Address*
Amagic Holographics, Inc.	http://www.amagic-usa.com/
American Bank Note Holographics, Inc.	http://www.abholographics.com/
Bank of England	http://www.bankofengland.co.uk/banknotes/index.htm
Bundesdruckerei GmbH	http://www.bundesdruckerei.de/en/products/index.html
Crane & Co.	http://www.crane.com/specialty_papers/default.asp
Engelhard Corporation	http://www.engelhard.com/
European Central Bank	http://www.ecb.int/
Flex Products, Inc.	http://www.ocli.com/markets/flex_products.html
Giesecke & Devrient	http://www.gdm.de/eng/main/home/index.php4
Graphic Security Systems Corp.	http://www.graphicsecurity.com
Hologram Industries	http://www.hologram-industries.com/pages/index.htm
Joh Enschedé	http://www.joh-enschede.nl/divisies/jess-engels.html
K-Laser Technology, Inc.	http://www.klaser.com.tw
Light Impressions International, Inc.	http://www.lightimpressions.com/
November AG	http://www.november-ag.de
OVD Kinegram Corporation	http://www.kinegram.com/
Papierfabrik Louisenthal GmbH	http://www.louisenthal.de/
Rolic Research, Ltd.	http://www.rolic.com/
SICPA Security Inks	http://www.sicpa.com/default.asp?lang=e
Thomas De La Rue	http://www.delarue.com/
United States Bureau of Engraving	http://www.bep.treas.gov/

The authors express their gratitude to Ms. Sue Spadini for the technical editing of this work, and to the many individuals and companies who provided exemplars and illustrations. Please see the figure captions and the reference lists for their names.

List of Units and Nomenclature

2D	two-dimensional
3D	three-dimensional
CIE	Commission Internationale de l'Eclairage; International Commission on Illumination
CIELAB color space	(L^* = Lightness; $a^*.b^*$ = chroma coordinates)
CMYK	cyan-magenta-yellow-black
QW	quarter wave

See also

Holography, Techniques: Color Holography; Holographic Interferometry; Holographic Recording Materials and their Processing; Sandwich Holography and Light in Flight. **Microscopy:** Interference Microscopy. **Optical Coatings:** Thin-film Optical Coatings. **Polarization:** Introduction; Matrix Analysis.

Further Reading

Committee on Next-Generation Currency Design, National Materials Advisory Board (1993) *Counterfeit Deterrent Features for the Next Generation Currency Design.* Washington, DC: National Academic Press.

Demus D, Goodby J, Gray GW, Spiess HW and Vill V (eds) (1998) *Handbook of Liquid Crystals*. Weinheim, Germany: Wiley-VCH.

Fagan WF (1990) *Optical Security and Anticounterfeiting Systems*. Volume 1210. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Faris SM (1994) US Patent 5,364,557: *Aligned Cholesteric Liquid Crystal Inks*.

Glausch R, Kieser M, Maisch R, *et al.* (eds) (1998) *Special Effect Pigments*. Hanover, Germany: Vincentz Verlag.

Hutley MC (1982) *Diffraction Gratings*. London: Academic Press.

Phillips RW and Bonkowski RL (2000) *Security Enhancement of Holograms with Interference Coatings*. Volume 3973. SPIE – The International Society for Optical Engineering. Bellingham, WA.

Phillips RW and Nofi M (1999) *Colors by Chemistry or by Physics?* Society of Vacuum Coaters 42nd Annual Technical Conference Proceedings, pp. 494–499, Chicago, IL.

Phillips RW, Nofi M and Slusser R (1994) *Color Effects From Thin Films Designs*. Eighth International Conference on Vacuum Web Coating, pp. 270–284. Las Vegas, NV.

Proceedings of the International Conference on Security Documents for the 21st Century (1987) San Diego, CA. Prepared by Oak Ridge National Laboratories, CONF-87040175, Contract DE-A-C05-840R2140, Oak Ridge, TN 37831.

Rancourt JD (1987) *Optical Thin Films, Users' Handbook*. New York: Macmillan.

US House of Representatives (1984) *The Currency Design Act*. Hearing Before the Sub-Committee on Banking, Finance, and Urban Affairs, Second Session on H.R. 6005, Serial No. 98-98. Washington, DC: US Government Printing Office.

Van den Assem B, Brongers D, Rath J, *et al.* (1994) *Practical Guide for the Security Documents Printing Sector.* The Hague: SDU Uitgevererij Koninginnegracht.

Van Renesse RL (1994) *Optical Document Security.* London: Artech House.

Van Renesse RL (1996) *Optical Security and Counterfeit Deterrence Techniques*. Volume 2659. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL (1998) *Optical Security and Counterfeit Deterrence Techniques II*. Volume 3314. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Van Renesse RL and Vliegenthart WA (2000) *Optical Security and Counterfeit Deterrence Techniques III*. Volume 3973. SPIE Conference – The International Society for Optical Engineering. Bellingham, WA.

Diamond Optical Devices and Coatings

D M Aslam, Michigan State University, East Lansing, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

What do a human, a plant, the photoresist used in microfabrication, and a polymer have in common? Why has carbon become a crucial part of all known living systems? The answer to these questions may be embedded in carbon's hybrid bonding structure (sp, sp^2, and sp^3 C–C bonds). Carbon has 4 electrons in its outermost shell (the valence shell), which is the second shell. At first sight, one would expect that these 4 electrons are distributed among its one 2s and three 2p orbitals. However, it is well known that, in the diamond lattice, each carbon atom bonds with its four neighbors. Are the electrons in 2s and 2p orbitals responsible for the four bonds? One possible answer is that the 2s, $2p_x$, $2p_y$, and $2p_z$ orbitals of the carbon atom form bonds with their neighbors. There are two problems with this answer. First, these bonds will have different strengths (s and p orbitals), which is in contradiction with the experimental fact indicating an equal strength for all the four C–C bonds in the diamond. Second, the computed strength of such bonds will be lower than that found experimentally. What is then the nature of C–C bonds found in organic compounds?

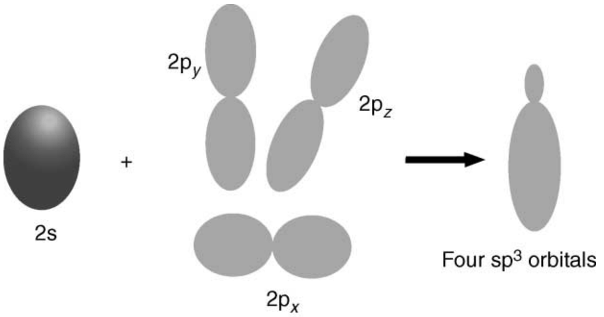

Figure 1 Hybridization of s and p orbitals in carbon to form sp^3 orbitals, which form the so-called sigma bonds responsible for the strong diamond lattice.

The answer to this difficult question was provided in 1931 by Linus Pauling (1901–1994), the winner of two Nobel prizes. With the help of quantum mechanical calculations, Pauling showed how one s orbital and three p orbitals can mix, or hybridize, to form four equivalent atomic orbitals with tetrahedral orientation. The resulting hybrid orbitals are called sp^3 orbitals as shown in Figure 1. The sp^2 or sp^1 orbital results if two or one p orbital combine with the s orbital, respectively, to form the hybrid. While sp^3 C–C bonds lead to a diamond lattice, the sp^2 C–C bonds lead to graphite, and a material which contains both sp^2 and sp^3 C–C bonds is called diamond-like carbon (DLC). The optical properties, such as the bandgap, of DLC depend on the sp^3/sp^2 ratio. If a sheet consisting of a monolayer of graphite, which consists of sp^2 bonds, is wrapped into a pipe, the resulting structure is called a single-wall nanotube (SWNT), as shown in Figure 2a. A multiwall nanotube (MWNT) has many such layers (Figure 2b–c). Other elements such as oxygen (in H_2O) and nitrogen (in NH_3) also form sp^3 orbitals but carbon seems to be only one which is capable of forming sp, sp^2, or sp^3 bonds, depending upon the bonding environment (Figure 3). This may explain why nature has selected carbon

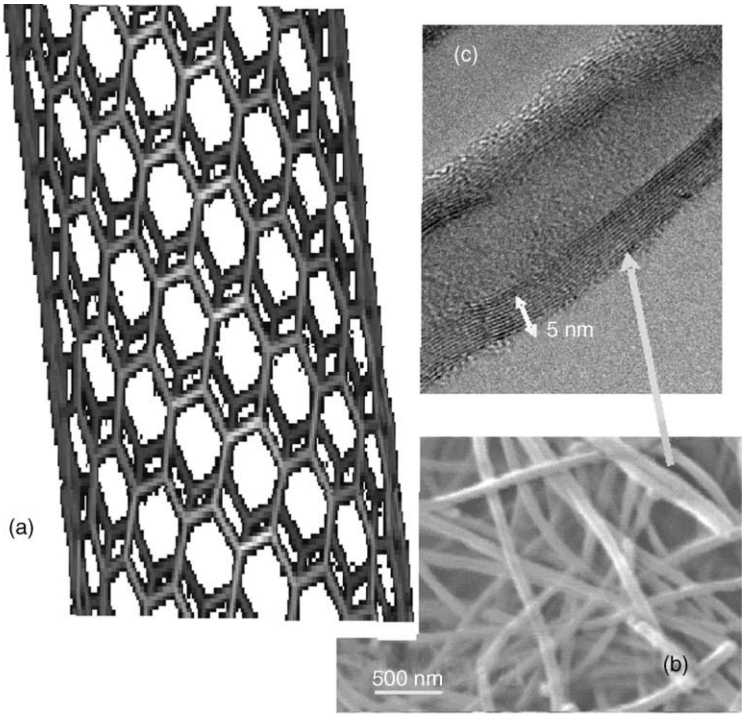

Figure 2 CNT: (a) drawing of a single wall CNT showing sp^2 bonded layer; (b) a mat of multi-wall CNTs; and (c) a TEM of MWNT showing 13 sp^2 bonded layers.

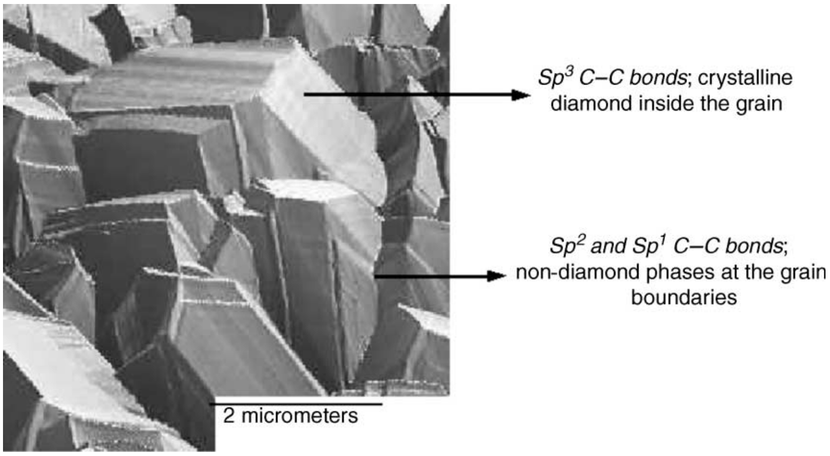
Sp^3 C–C bonds; crystalline diamond inside the grain
Sp^2 and Sp^1 C–C bonds; non-diamond phases at the grain boundaries
2 micrometers

Figure 3 Polycrystalline diamond film; crystalline regions separated by grain boundaries containing non-diamond phases.

Table 1 Comparison of semiconductor properties

Property	*Diamond*		*Si*	*GaA*	*β*-SiC
	Poly-C	*Crystalline*			
Fig. of merit					
Johnson, 10^{23} W Ω s^{-2}		73 856	9	62.5	10 240
Keyes, 10^{2} W cm^{-1} s^{-2} °C		444	13.8	6.3	90.3
Sat. electron vel., 10^{7} cm s^{-1}		2.7	1	1	2.5
Carrier mobilities, (cm^2 V^{-1} s^{-1})					
Holes	1–165	1600	600	400	50
Electrons		2200	1500	8500	400
Resistivity, Ω cm	10^{-2}–10^{8}	10^{-3}–10^{13}	10^{-4}–10^{8}	?–10^{8}	150
Breakdown field, MV cm^{-1}	0.1–1	10	0.3	6	4
Band gap, eV		5.45	1.12	1.42	3
Dielectric constant	6.7	5.5	11.7	12.5	9.7
Thermal expan. coeff., 10^{-6} °C^{-1}	2.6	1.1	2.6	5.9	4.7
Thermal cond., W cm^{-1} K^{-1}	20	4–22	1.5	0.5	5
Lattice constant, Å		3.57	5.43	5.65	4.36
Density, g/cm^3		3.52	2.32	5.31	3.215
Melting point, °C		4000^{a}	1412	1240	2540
Hardness, kg mm^{-2}		10 000	1000	600	3500
Poisson ratio	0.11	0.15	0.23	0.315	
Young's modulus, 10^{12} Pa	0.8–1.2	1.1–1.2	0.155	0.085	0.7

[a] Diamond may convert to graphite well below 4000 °C depending upon the ambient and it oxidizes at 650 °C to form CO_2/CO in O_2.

(perhaps due to its bonding flexibility, to form long C–C chains and rings) as the material of choice for building living systems.

The sp^3 bonding in a diamond lattice makes it a very unique material with a combination of properties not matched by any other known materials (Table 1). In fact, the C–C sp^3 bond in diamond is responsible for its remarkable properties leading to current and potential applications as shown in Figure 4. Although diamond (for centuries) and its bonding structure (for over 70 years) has been known for a long time, the growth of diamond films using chemical vapor deposition (CVD) took a long time to mature. Again, this delay relates to the difficulty with which the diamond lattice with sp^3 C–C bonds is fabricated.

Diamond Growth by CVD

First evidence of diamond growth by CVD, by Eversole in 1952–1953, led to the use of H_2 and CH_4 in the hot filament CVD (HFCVD) to grow diamond on diamond substrates (homoepitaxial growth) by Angus in 1971. The inexpensive CVD polycrystalline diamond (poly-C) was grown on non-diamond substrates by Deryagin in 1976, Spitsyn in 1981, and by Matsumoto *et al.* in 1983. Currently,

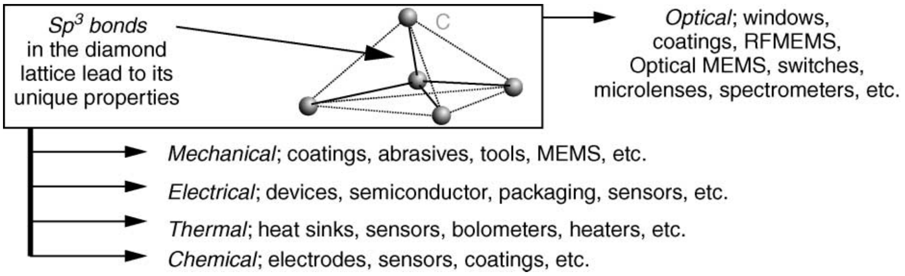
*Sp*³ *bonds*
in the diamond
lattice lead to its
unique properties
C
Optical; windows,
coatings, RFMEMS,
Optical MEMS, switches,
microlenses, spectrometers, etc.
Mechanical; coatings, abrasives, tools, MEMS, etc.
Electrical; devices, semiconductor, packaging, sensors, etc.
Thermal; heat sinks, sensors, bolometers, heaters, etc.
Chemical; electrodes, sensors, coatings, etc.

Figure 4 Sp^3 bonds in diamond make it a unique material.

Table 2 Different diamond deposition methods

Methods	*HFCVD*	*MPCVD*	*DC-arc Jet CVD*	*Combustion synthesis*	*Multiple pulsed laser*
Deposition rate (μm/hr)	0.1–10	0.1–10	30–150	4–40	3600
Substrate temperature (°C)	300–1000	300–1200	800–1100	600–1400	50
Deposition area (cm^2)	5–900	5–100	<2	<3	N/A
Advantages	Simple, large area	Quality, stability	High rate, good quality	Simple, high rate	Ultra-high rate
Disadvantages	Contamination, fragile filament	Rate	Contamination, small area	Small area	Expensive

HFCVD, hot filament CVD.

it is believed that, during the CVD of diamond, the CH_3 is responsible for deposition of C as diamond and non-diamond phases. The atomic hydrogen, present in the growth environment, removes the non-diamond phases leaving behind the diamond phase. Near the end of the 1990s, basic science of CVD diamond was well understood, and today diverse plasma and thermal techniques have been developed to produce poly-C films several mm thick and over 12 inches in diameter. Optically smooth 300-micrometer thick undoped poly-C wafers are available on the market. Although there are some reports of n-type poly-C and crystalline diamond growth, the well-established techniques exist only for *in situ* doping of p-type diamond.

A number of diamond growth techniques are currently available for the growth of diamond and carbon nanotubes. Microwave plasma CVD (MPCVD) can be used for the growth of both diamond and carbon nanotubes. As shown in Table 2, the deposition rate of diamond shows a large variation (0.1–3600 micrometers/hour) depending on the growth technique. For applications of diamond in optical coatings, optoelectronic devices and optical systems, the study of transmission, reflection and absorption of light for diamond is very important. These properties of diamond are responsible for its beauty and glamor. For applications in optoelectronic materials, devices and systems, the micro-fabrication and -machining technologies are crucial, particularly in the case of inexpensive poly-C films.

Absorption, Reflection, and Transmission in Diamond

The optical properties, which have been studied for hundreds of years, of naturally occurring single crystal diamond have led to the gem commerce. The fundamental concept involved in the tremendous historical fame of diamond is its high refractive index, n_d (in the range of 2.38–2.64 depending upon the wavelength of light). The refractive index of a material is a measure of decrease of velocity of light (which is also true for all electromagnetic waves) when the wave enters the material. Mathematically, n_d is defined by

$$n_d = V_{vac}/V_d \tag{1}$$

where V_{vac} and V_d are velocities of light in vacuum and diamond, respectively.

When a beam of light of intensity I_0 in air is incident normally on the surface of diamond, the reflected part of the beam, characterized by the reflection coefficient R, depends on n_d, and R is given by

$$R = \left(\frac{n_a - n_d}{n_a + n_d}\right)^2 \tag{2}$$

where n_a is refractive index of air and n_d is given by Sellmeier dispersion equation:

$$n_d = \sqrt{1 + \frac{4.3\lambda^2}{\lambda^2 - 0.0121} + \frac{0.33\lambda^2}{\lambda^2 - 0.031}} \qquad [3]$$

where λ is the wavelength of light expressed in micrometers. When the beam enters diamond, part of it is absorbed in diamond. The absorption, characterized by the absorption coefficient α, is caused by: (i) impurities (such as nitrogen, boron, hydrogen, etc.); and/or (ii) intrinsic defects (such as dangling bonds, sp^2 bonds, and sp bonds). However, if the energy of light is close to or above the bandgap of diamond (5.5 eV), the absorption can be caused by electronic transitions from valence to conduction bands. The intensity $I(x)$ at a distance x below the diamond surface is given by

$$I(x) = I_0(1 - R)\exp(-\alpha x) \qquad [4]$$

The fundamental expressions for α are different for indirect and direct type semiconductors. However, the modeling of α, for energies close to the bandgap of diamond, can be accomplished empirically using Urbach's rule according to the following equation:

$$\alpha = \alpha_0 \exp\left[\frac{\sigma(T)(E - E_0)}{kT}\right] \qquad [5]$$

where α_0 and σ are Ubach tail parameters, and E_0 (6.5 eV) is direct gap energy of diamond. K is Boltzmann constant and T is the temperature. As at $T = 300$ K, $\alpha_0 = 4.23 \times 10^{11}$ cm^{-1} and $\sigma = 0.585$, eqn [5] can be simplified:

$$\alpha = 4.23 \times 10^{11}\, e^{22.6(E-6.5)} \qquad [6]$$

where the units for α and E are in cm^{-1} and eV, respectively. Using the expression $E = 1.24/\lambda$, where E is in eV and λ is in micrometers, eqn [6] can be written as

$$\alpha = 4.23 \times 10^{11}\, e^{27.9(1/\lambda-5.24)} \qquad [7]$$

where λ is in micrometers. The variations of α and n_d with wavelength and energy of light are shown in Figure 5, which are plotted using eqns [3] and [7]. Impurities and intrinsic defects in diamond strongly affect α, n_d, and its other properties.

Impurities in Diamond

A pure diamond with perfect lattice structure is transparent to ultraviolet (UV), visible, infrared (IR), millimeter and microwave regions of the electromagnetic spectrum. However, even an impurity-free diamond has some absorption in the infrared due to phonon excitations and an ultraviolet cutoff due to band-to-band absorption. The defects, impurity-related and intrinsic, in single crystal diamond have been extensively studied because they strongly affect its properties. In fact, diamonds are classified according to the type of defects that are dominant.

Impurities in diamond give rise to additional energy levels in its bandgap. If the density of these impurities is sufficiently high, they can cause absorption of electromagnetic radiation (including light) on one hand and break the lattice symmetry on the other. The breaking of the lattice symmetry can lead to one-phonon IR absorption, which is forbidden in pure diamond. The fact that impurities affect the optical properties of diamond led to its classification into types Ia, Ib, IIa, and IIb which correspond to pale yellow, golden yellow, no color, and blue, respectively. The transmission properties of these various types of diamond are depicted in Figure 6. The dominant impurities are nitrogen

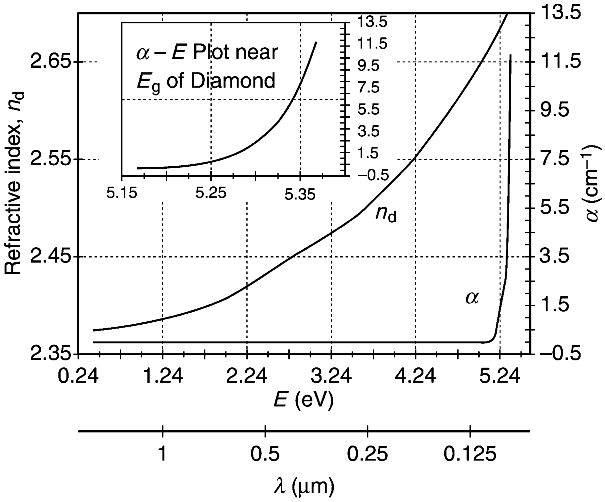

Figure 5 Wavelength and energy dependence of absorption coefficient and refractive index of diamond.

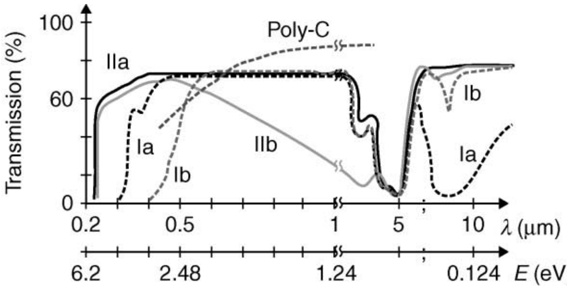

Figure 6 Transmission spectra of natural diamond and CVD polycrystalline diamond (poly-C).

(5–2500 ppm atomic) in type I and boron (10 ppm atomic) in type II. It may be pointed out that boron-doped diamond is a semiconductor with p-type conductivity measured at room temperature. Nitrogen-doped diamond does not show n-type conductivity at room temperature as the donor levels are too deep to be ionized.

Polycrystalline CVD Diamond

Due to the lack of a reliable hetero-epitaxial diamond growth technique, the crystalline diamond films are very expensive (require single crystal diamond substrates for their growth) and, consequently, their widespread commercial use is severely limited. Polycrystalline diamond (poly-C) grown by CVD, which can be as inexpensive as any other material if mass produced, is very attractive for optical and other applications (Figure 2). The properties of high-quality (with high density of sp^3 C–C bonds) poly-C films are comparable to those of crystalline films.

As shown in Figure 6, the transmission properties of poly-C are comparable to or better than those of type I and II diamonds for part of the energy range. Free-standing commercial poly-C films are typically polished and have thicknesses in the range of 50–300 micrometers (the least expensive are 300 μm thick). White poly-C films transparent to visible light are also available. Such films, with optically smooth surface, can be considered free of scattering effects. These films can be considered as a transparent plate in air and, for normal incidence of light with energy less than the bandgap, the transmission coefficient is given by

$$T = \frac{(1 - r_1^2)(1 - r_2^2)}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos(2\delta)} \qquad [8]$$

where the Fresnel coefficients are given by

$$r_1 = \frac{n_a - n_d}{n_a + n_d}, \qquad r_2 = \frac{n_d - n_a}{n_d + n_a} \qquad [9]$$

where n_a and n_d are refractive indexes of air and poly-C, respectively, and δ is phase change for a poly-C film with a thickness of t:

$$\delta = \frac{2\pi}{\lambda} n_d t \qquad [10]$$

Using $n_a = 1.0003$, $n_d = 2.4$, and $t = 100$ μm, eqn [8] can be written as

$$T = \frac{1}{1.49 - 0.493 \cos\left(\frac{3015}{\lambda}\right)} \qquad [11]$$

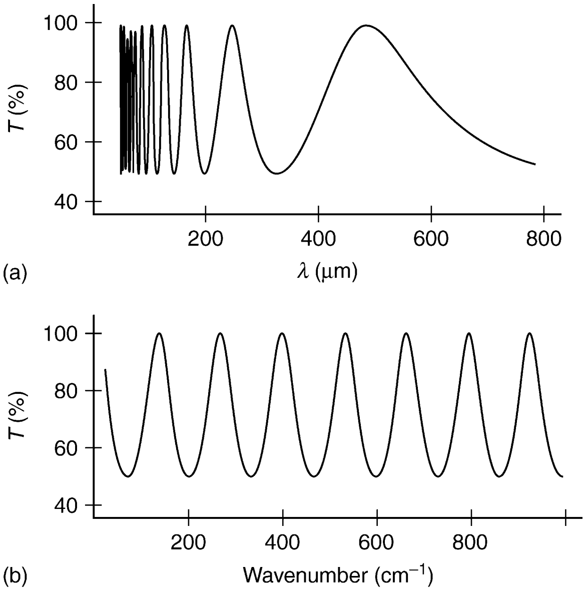

Figure 7 Transmission coefficient of a polished poly-C film with a thickness of 100 μm as a function of (a) wavelength and (b) wavenumber.

or

$$T = \frac{1}{1.49 - 0.493 \cos(0.048\, k)} \qquad [12]$$

where λ is in μm and k, the wavenumber, is in cm^{-1}. Variation of T as a function of λ and k, is shown in Figure 7. A similar behavior is observed for non-polished poly-C films if the wavelength is larger than the surface roughness of the films. For a sample with a very rough surface, scattering effects need to be included in eqn [8].

Electroluminescence in Poly-C

In a luminescence process, the electron must first be excited to higher energy, which can be accomplished by light (photoluminescence), electron beam (cathodoluminescence) or electrical energy (electroluminescence). A subsequent electron transition to lower energies leads to emission of light. A common example of electroluminescence is a forward biased p–n junction made out of direct bandgap type semiconductor such as GaAlAs. The light emitting diodes (LED) and LASERS are based on electroluminescence. Red, green, and blue LEDs are made from GaAlAs, GaP, and SiC/GaN, respectively. Indirect bandgap type semiconductors, such as diamond, typically do not exhibit electroluminescence. The fact that electroluminescence has experimentally been observed in crystalline diamond and poly-C seems to be, at first sight, a contradiction of its indirect bandgap. In fact, defects in diamond are believed to be responsible for their luminescence properties. All types of diamond are found to have a number of defects related to vacancies and impurities. Due to a unique combination of properties of diamond (Table 1), its electroluminescence properties offer a very unique application potential in optical microelectromechanical systems (MEMS), displays and communication devices in IR, visible and UV regions.

Electroluminescence (EL) in optoelectronic devices is produced by excitation of electrons by electronic current. Yellow-green EL has been observed in Schottky diodes made of boron-doped poly-C. Blue-green and blue-violet EL has been observed in free-standing poly-C films, with thicknesses in the range of 15–400 micrometers, when the samples were subjected to applied fields in the range of 2 V/μm. The EL peaks are observed at energies in the range of 3–3.85 eV, with the main peak occurring at 3 eV.

Recently, a very interesting type of EL has been observed in poly-C and carbon nanotubes, which seems to have an application potential in displays and optical MEMS. Under applied electric fields in the range of 20–40 V/μm, electrons are known to tunnel out of a poly-C film into vacuum and the process is called field emission. EL has been observed during the field emission process, and this new effect is called field emission EL (FEEL). A typical FEEL spectra is shown in Figure 8. It is believed that FEEL results from direct type defects in poly-C, which are populated during the field emission process. The electrons in these defects, instead of being field emitted, make transitions to lower energy levels leading to light emission. The main FEEL peak is observed at 2.6 eV which is in the visible range.

Poly-C Applications

The current applications of diamond, which include optical windows, temperature sensors, heat sinks, cutting tools, surgical instruments, and optical coatings, are based on natural, synthetic, and CVD diamond. The cost of producing diamond has always been prohibitive in widespread diamond applications. Fortunately, over the past 20 years, the development of CVD diamond technologies has eradicated the fears of diamond being too expensive for its widespread applications. As the CVD poly-C can be as inexpensive as any other electronic material, widespread applications of diamond are expected for poly-C. In the area of micro- and nanostructures, the poly-C can provide a unique opportunity for both polycrystalline and crystalline microdevices. Using the present fabrication techniques, the grain size in poly-C can be varied in the range from a few tens of nanometers (nanocrystalline diamond) to hundreds of micrometers (commercial free-standing poly-C films). This leads to a whole range of device structures ranging from inter- to intra-grain devices. An intra-grain device made in poly-C will be an inexpensive single crystal device with properties better that those found in natural/synthetic diamonds, because an intra-grain poly-C device will have a higher degree of purity (because it is grown by CVD).

The applications of poly-C in microsystems (current, in-progress, and future) include optical MEMS, biological/protective/hard coatings, sensors,

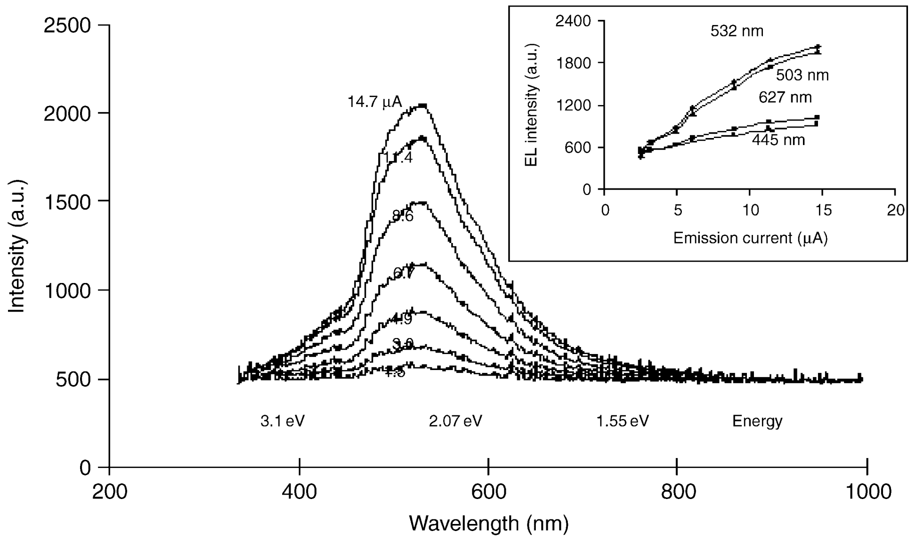

Figure 8 FEEL spectra in poly-C as function of wavelength and emission current (inset).

field emission displays, and WIMS (wireless integrated microsystems). Intra-grain microlenses, RFMEMS, heat pipes, sensors, microabrasives, and spectrometers may lead the way for unique poly-C applications.

Optical MEMS

A basic MEMS structure, with electrostatic actuation mechanism, is shown in Figure 9. Currently, most of the Optical MEMS and RFMEMS (radio frequency MEMS) devices are made from poly-Si (polycrystalline Si) using surface or bulk micro-machining fabrication techniques. The ability to integrate micro-optical elements with movable structures and micro-actuators has opened up many new opportunities for optical and optoelectronics systems, as MEMS allow a more effective manipulation of optical structures than the conventional methods. Optical MEMS have applications in displays, sensing, data storage, and communication systems. Some of the examples are optical switches, optical cross-connects, wavelength division add/drop mixers, tunable filters/lasers/detectors, dispersion components, polarization dispersion components, and spectrometers.

Material-related limitations put some constraints on the current optical MEMS technologies and applications. Diamond is an excellent material for optical MEMS and can extend the application range and reliability. The use of poly-C in optical MEMS can also remove the traditional cost constraints associated with the use of single crystal diamond. Although the poly-C optical MEMS are in their infancy, recent developments are very exciting. Figure 10 shows a poly-C resonator structure fabricated using a poly-C fabrication technology compatible with Si processing techniques. Such structures, if used as filters and switches, can lead to a superior performance as compared to any other currently used material. Poly-C on-chip spectrometers are also expected to operate in a large wavelength range.

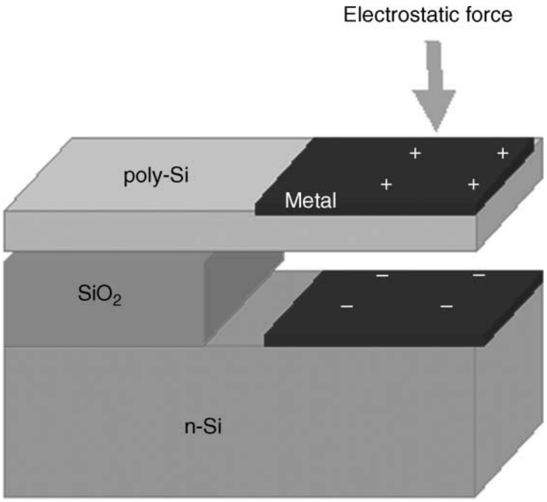

Figure 9 A simple MEMS structure with electrostatic moving of the cantilever beam.

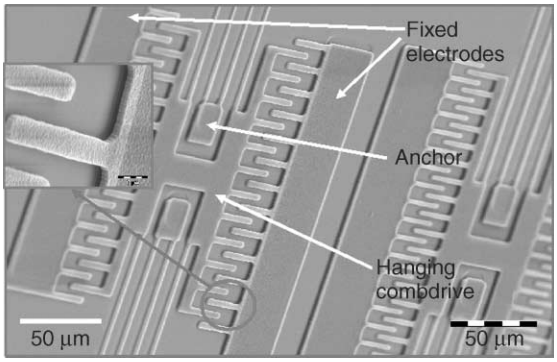

Figure 10 A poly-C MEMS combdrive for high-quality factor resonators, inertial sensors, optical MEMS, etc. Poly-C is superior to poly-Si as its Young's modulus is 4 times higher than that of poly-Si.

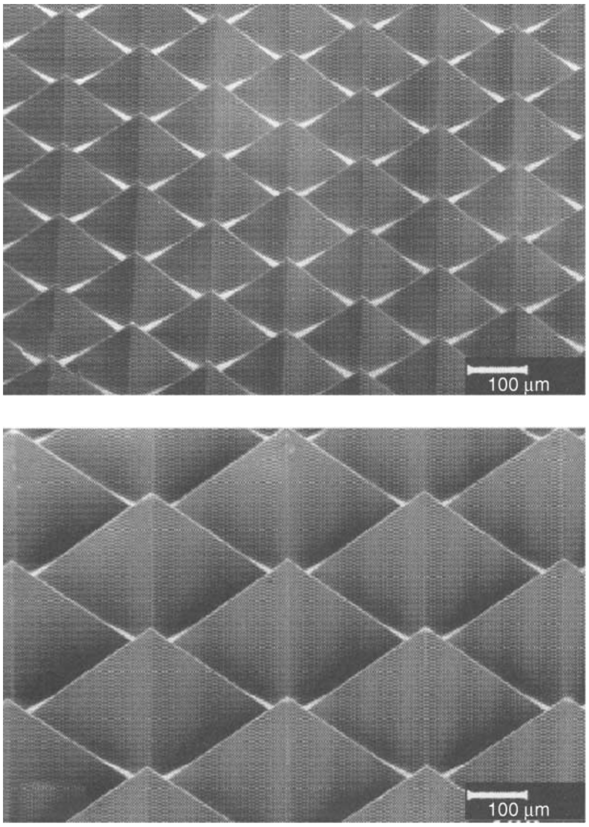

Figure 11 Poly-C microabrasive structures two different densities of pyramids. (Courtesy of Dr. H. Bjorkman of Upsala University.)

Microabrasives

Poly-C is also a very promising material for microabrasives. Micromachined poly-C MEMS structures have recently been fabricated for possible use for microabrasives. Figure 11 shows micromachined MEMS structures which are expected to enhance the life of a tool. Such abrasives by design are inherently cost-effective because they are batch fabricated.

See also

Light Emitting Diodes.

Further Reading

Asmussen J and Reinhard D (2002) *Diamond Handbook*. New York: Marcel Dekker.

Kim U and Aslam DM (2003) Field emission electroluminescence on diamond and carbon nanotube films. *Journal of Vaccum Scientific Technology B* 21(4): 1291–1296.

McMurry J (1996) *Organic Chemistry*. New York: Brooks/Cole.

Pan LS and Kania DR (1996) *Diamond: Electronic Properties and Applications*. Boston, MA: Kluwer.

Spear KE and Dismukes JP (1994) *Synthetic Diamond: Emerging CVD Science and Technology*. New York: Wiley.

Laser Damage in Thin Film Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since the very beginning of laser technology, Laser-induced Damage Thresholds (LIDT) of optical components were always an obstacle for the development of laser systems operating at high power levels. In such systems, the surface or the bulk of the optical components can be damaged by laser radiation of sufficiently high power or energy density resulting in an uneconomical standstill of the laser and its application. In the early days of high power laser development, mainly inclusions in laser rod materials were discussed as a major complication for an augmentation of the output power in solid state laser systems. Nowadays, in the course of the development of optical materials with excellent quality and power handling capability, the problem of laser-induced damage has shifted from the bulk to the surface of the optical component. The optical surface is subjected to various production steps and environmental influences, which modify its structure and composition. Especially, the thin film coating, which is deposited on the optical surface to adapt its reflectance and transmittance to the application, contributes predominantly to the reduction of the LIDT values. As a consequence, the measurement and optimization of the power handling capability of thin films is considered as one of the primary research areas in modern optics technology and is supported by an extensive scientific community.

In the following, a brief review will be given on selected fundamental damage mechanisms in thin films, considering also the scaling of LIDT values for different operation conditions. Also, current standards for the measurement of LIDT will be described, and examples illustrating some practical aspects of high power optical coatings will be presented. Finally, a summary of the present state of the art will be given and recent trends in laser technology will be discussed in respect to research in laser-induced damage.

Fundamental Damage Mechanisms in Thin Films

During thirty years of intense research in laser damage of thin films, a broad spectrum of different damage and degradation mechanisms had been discovered and studied for a variety of laser operation conditions and wavelengths. For prominent high-power laser systems often applied in industrial production fields or fundamental research, specific models for laser-induced breakdown have been developed, which are also of practical relevance. In many experiments, damage sites with the morphology illustrated in Figure 1 are observed. The coating is delaminated from the surface of the optical component in the center of the laser beam area. Obviously, mechanical stress has built up in the coating by heating, which is induced by the absorbed laser power. At a certain stress level, exceeding the adhesion strength of the coating to substrate surface, the thin film cracks and may even delaminate from the component. In other cases of this absorption-induced damage effect, the film reaches its melting point prior to the mechanical threshold and

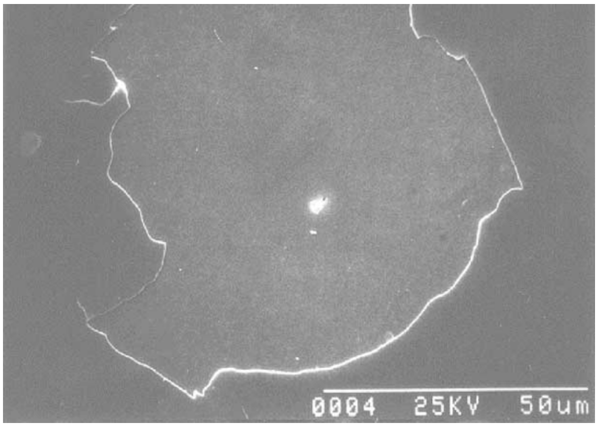

Figure 1 SEM-picture of a damage site on an alternating coating of SiO_2/HfO_2. The system was irradiated by a Nd:YAG-laser with an energy density of 45 J/cm^2. As a dominant damage mechanism, a delamination of the coating by extreme mechanical stress is suggested. Reproduced with permission from Ristau D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.49.

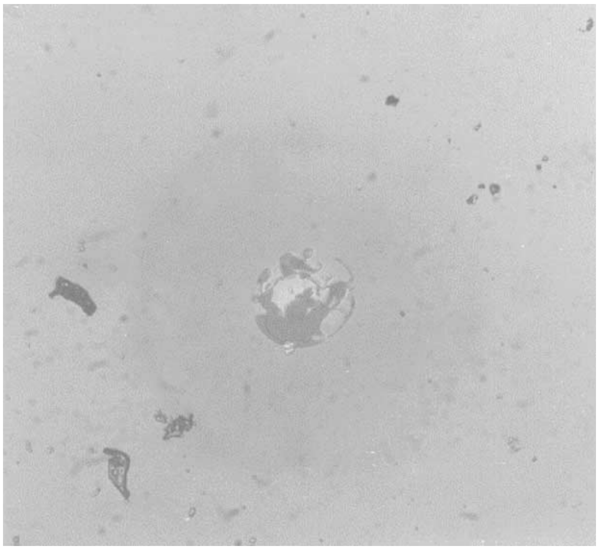

Figure 2 Nomarski micrograph (magnification 250×) of a damage site on an antireflecting coating of SiO_2/Ta_2O_5. The system was irradiated with an energy density of 51 J/cm^2 at a wavelength of 1.064 μm. A delamination of the coating in the center of the beam is surrounded by a recrystallization zone.

evaporates or changes its crystalline structure. The morphology of such damage sites exhibits a discoloring or an increased surface roughness in the center of the laser beam (see Figure 2). Absorption-induced damage is dominated by the instantaneous heating of the coating material in the area of interaction with the laser beam and can be described on the basis of the heat diffusion equation. In this approach, the source term in the heat diffusion equation is given by the heat coupled into the system by the absorbed laser power, and the boundary conditions are determined by the geometry and the structure of the layer system as well as the spatial laser beam profile. Apparently, absorption-induced damage has to be considered only for components with significant absorptance at their operation wavelength. Typical examples for coating materials and wavelength regimes, where absorption-induced damage is frequently observed, are the wavelength 10.6 μm of the CO_2-laser with ZnSe, ZnS, and fluorides as coating materials, or the DUV/VUV-wavelengths dominated by excimer lasers and fluoride coating materials. The effect of absorption-induced damage in optical materials is also of practical relevance for cw- and long-pulse operation conditions of the laser system corresponding to long thermal diffusion lengths. For cw-irradiation, the temperature rise in the component can be calculated by numerical methods involving finite elements or differences. For boundary conditions with circular symmetry and a Gaussian beam profile (beam diameter w, power P), an analytical expression can be derived for the temperature rise ΔT in the center of an irradiated circular component

$$\Delta T = \frac{2\beta_s P}{\pi^{3/2} k w} \tan^{-1}\left(\frac{16\kappa t_I}{w^2}\right)^{1/2} \qquad [1]$$

In this model the temperature rise is dependent on the thermal properties (k: thermal conductivity, κ: thermal diffusivity), on the surface absorption of the component (β_S), and on the beam diameter. For irradiation times t_I, long compared to the typical heat diffusion time w^2/κ, eqn [1] reduces to the asymptotic dependence

$$\Delta T \rightarrow \frac{\beta_s P}{\pi^{1/2} k w} \qquad t_I \rightarrow \infty \qquad [2]$$

Obviously, the temperature rise ΔT scales with the P/w for these long-time irradiation conditions compared to the short pulse regime, where the scaling of the temperature rise is proportional to P/w^2. For the application of optical components in high power cw- and long-pulse lasers, this P/w scaling law has to be considered carefully, because the power handling capability depends critically on the size of the irradiated area. Also, the onset of damage may be delayed in respect to the start of irradiation by time constants in the range of several 10 seconds.

In the course of the development of improved coating processes, optical absorptance could be reduced to very low levels in the near infrared and visible spectral region. For example, high reflecting mirrors for the wavelength of the Nd:YAG-laser with total losses below 1 ppm can be produced with advanced ion beam sputtering processes nowadays. Therefore, another principal damage mechanism, which is based on inclusions or defects in the coatings, is often found in the near infrared spectral range. A detail of a damage site representing inclusion dominated damage induced by a pulsed Nd:YAG-laser in a coating of HfO_2/SiO_2 is shown in Figure 3. In the laser beam area numerous craters are located, which extend from circular voids with different diameters in a range below 1 μm. At the high magnification of the employed electron microscope, the multilayer structure of the coating system can be resolved indicating an origin of damage approximately in the same depth of the layer structure at the substrate surface. The underlying damage mechanism is initiated by inclusions or other defects with an absorption, which is significantly higher than for the surrounding thin film material. As a consequence, the inclusion is rapidly heated during the interaction with the laser beam and explodes removing the covering layer structure. This inclusion breakdown can be modeled by calculating the temperature behavior $T(r)$ at the radial position r of a sphere (radius g, thermal conductivity k_E, diffusivity κ_E) which is embedded in a medium with defined properties (see eqn [3]). The interaction of the sphere, which has a size in the range of the wavelength λ and a refractive index n', with the laser irradiation, is described by the Mie absorption coefficient Q. For the instant of damage, a critical temperature at the outer surface of the sphere ($r = g$) is assumed, which is often related to the melting point of the layer material

$$T(r) = \frac{3QI}{4\pi k_E g}\left[\frac{k_E}{3k} + \frac{1}{6}\left(1 - \frac{r^2}{g^2}\right) - \frac{2gb}{r\pi}\int_0^\infty \mathrm{d}y \exp\left(-\frac{y^2 t_p}{\tau_1}\right) \times \frac{(\sin y - y\cos y)(\sin ry/g)}{y^2\left[(c\sin y - y\cos y)^2 + b^2y^2\sin^2 y\right]}\right] \quad [3]$$

$$Q = \pi g^2\left[1 + \frac{2\exp(-8\pi n' g/\lambda)}{8\pi n' g/\lambda} + \frac{\exp(-8\pi n' g/\lambda) - 1}{(8\pi n' g/\lambda)^2}\right]$$

$$\tau_1 = \frac{g^2}{\kappa_E} \quad b = \frac{k}{k_E}\left(\frac{\sqrt{\kappa_E}}{\sqrt{\kappa}}\right) \quad c = 1 - \frac{k}{k_E}$$

Besides the thermal properties of the host material (thermal conductivity k, diffusivity κ), the properties and the statistic distribution of the inclusions have to be known for a quantitative description of inclusion breakdown.

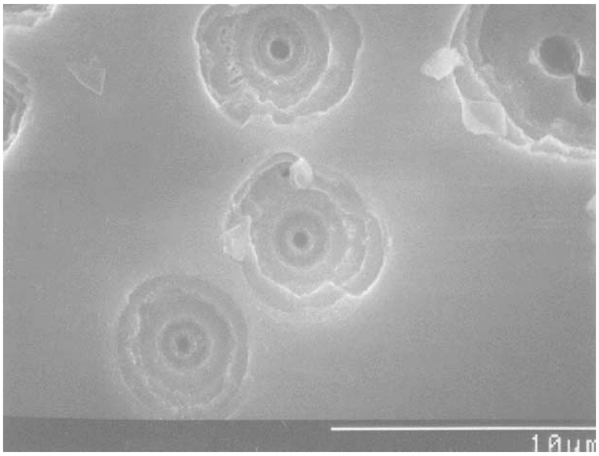

Figure 3 SEM-picture of damage site on a high reflecting coating of SiO_2/HfO_2. This inclusion-dominated breakdown mechanism was initiated at an energy density of approximately 130 J/cm^2 with a Nd:YAG-laser.

In contrast to the nearly deterministic behavior of absorption induced damage, defect dominated damage mechanisms often exhibit a statistical nature. Since the inclusions are not homogeneously distributed in the layer structure, a variation of the threshold value is observed for different damage sites on a single sample. The assessment of the inclusion parameters is extremely difficult in practice. Therefore, the inclusion model is mainly employed for the prediction of tendencies for the LIDT as a function of the beam parameters and the properties of the layer materials.

Different aspects related to the substrate polishing, cleaning, and deposition process have to be considered for an investigation in the various origins of inclusions or defects in a layer structure. Besides contamination of the substrate surface prior to the deposition process, particles may be generated during the deposition process by mechanical abrasion of material from moving parts in the plant or by delamination of material coated on the inner walls. Also, the deposition source may contribute to the formation of defects by sputtering of particles or droplets from the melt or the target. Especially in low energetic thermal deposition processes, defects are often the origin for the formation of nodules. These conical faults in the layer structure reduce the LIDT because they are less resistant against intensive laser radiation than the bulk of the coating system.

Besides models, which are based on the transformation of laser radiation into heat, direct electronic excitation has to be considered as a breakdown mechanism for short-pulse lasers. In these intrinsic breakdown models for dielectric materials, the generation of a high electron density in the conduction band is assumed. This carrier generation can be explained by high field ionization of electrons in the valence band of the dielectric material. Many theoretical concepts introduce the avalanche effect as an important mechanism contributing to the generation of free electrons. Electrons excited by strong field ionization can interact with the electric field and accumulate excitation energy. By collisions with the lattice, this energy can be also transferred to the electrons in the valence band. Thus, when the energy of the free electrons exceed a critical value, other electrons can be excited to the conduction band, and a carrier generation like an avalanche arises. Another theoretical approach is the multiphoton model which is formulated on the basis of multiphoton absorption for the generation of free electrons. In most intrinsic models it is assumed, that laser-induced damage breakdown occurs when the electron density in the conduction band reaches a critical value of approximately 10^{21} 1/cm^3. At this high electron density, a plasma state is reached and the impinging radiation is coupled efficiently into the layer structure resulting in its destruction.

Intrinsic damage is often characterized by a deterministic damage behavior with a well defined threshold value which is also characteristic for damage in bulk materials. As a prominent example, damage induced by ultra-short pulses with a duration below 1 ps can be often attributed to intrinsic mechanisms, because the diffusion length of thermal effects can be neglected in comparison to the intensive interaction of the laser radiation with the electrons of the coating materials. In this case, the generation rate dn/dt of electrons in the conduction band can be described on the basis of the impact ionization rate and multiphoton excitation

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha n(t) q I(t) + \beta_m \left[q I(t) \right]^m \qquad [4]$$

In this model $I(t)$ is the power density of the laser radiation, and q is a correction factor representing interference effects in the coatings, which may result in a local power exceeding the power density of the radiation impinging on to the layer system. The electron excitation mechanisms are described by the impact ionization coefficient α and the m-photon absorption coefficient β_m. The relative contributions of both excitation mechanisms to laser-induced breakdown are depending on the pulse duration and the bandgap energy of the layer material.

In a rough approximation, the damage threshold increases with the bandgap energy of the materials as a consequence of the photon absorption coefficient β_m decreasing with its order m. For short pulse durations below a few 10 fs, multiphoton processes contribute significantly in the range of some 10% to the free-carrier generation. In most materials, avalanche ionization develops to the dominant generation mechanism for pulse durations above 100 fs. These tendencies could be demonstrated for single layers of selected oxide material deposited by ion beam sputtering on quartz substrates, which were tested in a pulse duration regime from 20 fs to 1 ps. According to the theoretical model, the expected increase of the damage also influencing values with the pulse duration was observed in the damage experiments.

In practice, the power handling capability of optical components is also often limited by imperfections like scratches, digs, and areas with high roughness on the optical surface. In most deposition techniques applied in optical technology, the layer system tends to replicate or even enhance surface imperfections of the substrate. In the application, additional weak points may be introduced by improper handling or contamination of the optical surfaces. If the contaminants are not removed by the impinging high-power laser radiation, they increase the surface absorption and act as initiation points for laser-induced damage. Voids, grooves, pores, or scratches reduce the power capability of the optical element, because they act as concentrators for the electric field.

Units and Scaling of Laser-Induced Damage Threshold

According to the theoretical models of laser-induced damage, the appropriate units of measurement for LIDT values are mainly given by the dominant damage mechanism and the irradiation time. For pulsed laser radiation and dielectric breakdown effects, the laser-induced damage threshold is usually reported in W/cm^2. In the case of absorption-induced damage or inclusion dominated breakdown and a pulse duration short compared to the thermal diffusion time in the layer structure, LIDT values are often scaled in J/cm^2. The unit of linear power density (W/cm) is indicated for the LIDT of long-pulse, cw-lasers or other sources, which induce a temperature increase in the optical component with rise times in the regime of seconds.

Laser-induced damage thresholds of optical thin films are critically dependent on the operation conditions of the applied laser system and on the design of the layer structure. For example, even the damage mechanism can change if the wavelength or the pulse duration is varied. Therefore, LIDT values should be scaled with the irradiation parameters only for small intervals, in cases where the fundamental damage mechanism is clarified. In all other situations, the extrapolation of threshold values is extremely difficult and may lead to inaccurate results. Especially, an overestimation of the LIDT value holds severe dangers in practical applications, where the replacement of a damaged component and the repair of other parts impaired by the damage event causes high expenses. If the damaged components consist of toxic materials (e.g., ZnSe, GaAs, CdTe, ThF_4, chalcogenides, Be, Cr, etc.) severe health hazards may occur, and an expensive decontamination of the environment may be necessary. Therefore, the determination of the laser-induced damage threshold using the actual laser parameters under controlled environmental conditions is always recommended for unclear cases.

In the present state of research in laser-induced damage mechanisms, only few tendencies are generally accepted as confident for scaling of LIDT values. As a function of the pulse duration, laser-induced damage thresholds increase for longer pulses. For inclusion and absorption dominated breakdown, a $t_p^{1/2}$-law is often used in the pulse duration regime between 10^{-10} to 10^{-8}s. This dependency can be extended to other pulse regimes up to pulse lengths in the ms-range, if the exponent is replaced by a value between 0.5 to 2. In respect to the laser wavelength, a decrease of the LIDT value with decreasing wavelength is observed for most materials and operational conditions. Investigations in the influence of the beam diameter have been performed by many research groups, indicating a decrease of the LIDT values for increasing beam diameters. Especially for inclusion dominated breakdown, the event of damage for a certain laser irradiated site will be dependent on the distribution of inclusions at that position. If the beam size is small, the probability for interrogating a defect vulnerable to damage is low. By increasing the spot diameter, this probability will asymptotically reach unity, because at a certain beam diameter, at least one defect will always be covered by the beam. Therefore, the onset of laser-induced damage is not dependent on the beam diameter for inclusion-dominated breakdown, which is often encountered in conventional optical coating systems. Another special case is the scaling of the cw-damage threshold with the beam diameter. Since the cw-LIDT is expressed in linear power density, the power handling capability increases more slowly with the beam diameter for the cw-lasers than expected from the normal pulsed operation. For example, extrapolating the LIDT value of 100 W at a beam diameter of 1 mm for a laser mirror to a beam diameter of 10 mm, results in a maximum power load for the component of 1 kW according to the correct P/w-law. The threshold power would be extremely overestimated to approximately 3.2 kW, if the general P/w^2-dependence is applied.

Measurement of Laser-Induced Damage Thresholds

As a consequence of the complicated relation between the laser damage mechanism and a broad spectrum of thin film properties and laser parameters, laser-induced damage threshold measurements have to be performed under well-defined conditions. To investigate the comparability of LIDT measurements, an extended international round-robin experiment has been conducted on coated optics for the wavelength of 1.064 μm at the beginning of the 1980s. This experiment indicated the need for a clearly specified LIDT-measurement procedure, and conceptual work was initiated to develop a corresponding ISO-Standard series. During recent years, an International Standard (ISO 11254) has been adopted covering testing conditions relevant for most typical laser applications. In the first part of ISO 11254 1 on 1-testing of optical surfaces, in respect to laser damage, is described. The fundamental approach of the standard measurement procedure is illustrated in Figure 4. A laser source operating in transversal and longitudinal single mode is employed for the irradiation of the sample surface. The beam

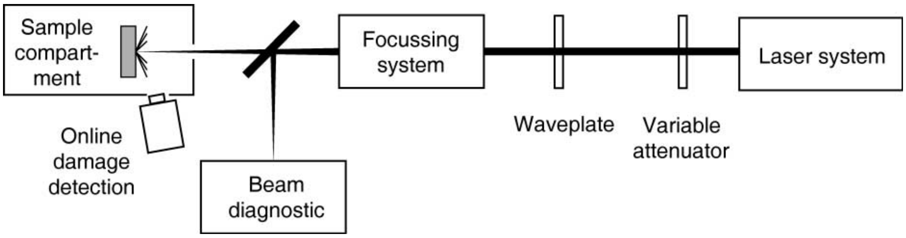

Figure 4 Fundamental setup for the measurement of laser-induced damage thresholds.

parameters of the laser are assessed by a beam diagnostic system which monitors the spatial and temporal profile as well as the energy of the laser radiation in the target plane. In order to achieve the high energy density levels necessary to destroy the surface of the specimen, a well characterized focusing system is installed. For adjustment of the laser energy in the target plane, an attenuating system is employed. In the 1 on 1 measurement protocol, each site on the sample surface is subjected to a single laser pulse once only. In a test sequence, the test surface is examined by pulses of different energies, covering low values without damage and high values causing damage. After the test, the specimen will be inspected with a Nomarski interference contrast or a darkfield microscope at a magnification of 150× or higher to identify the damaged sites. For the evaluation of the damage test, the damage probability method is recommended. In this evaluation scheme, the ratio of the number of damaged sites to the total number of sites objected to a certain energy or power level is interpreted as the damage probability. The plot of these damage probability values as a function of energy or power, which is called the survival curve of the optical component, provides an insight into the damage mechanisms involved. The damage threshold is given by the highest quantity of laser radiation for which the extrapolated probability of damage is zero. A typical example for a 1 on 1 damage test at the wavelength 1.064 μm and the extrapolated LIDT value is depicted in Figure 5.

In most catalogs of optics manufacturers, the 1 on 1 LIDT values are used to illustrate the power handling capability of their products, even though 1 on 1 data are of limited importance for practical applications, where an optical component is always subjected more than one time to a laser beam. This customary situation is covered by Part 2 of the ISO-LIDT-Standard which describes a damage test procedure for a series of pulses (S on 1-tests). For an assessment of the reliability, the concept of the characteristic damage curve is introduced by this standard. This curve is deduced directly from the S on 1 test data by plotting the energy density for a selected damage probability as a function of the number of pulses (see Figure 6). By an extrapolation of the characteristic damage curve to high pulse numbers in the order of 10^9 to 10^{12} shots, the lifetime of the optical component can be roughly estimated. For a certification of optical components in respect to their power handling capability, a third part of ISO 11254 is under development, which is concentrated on different testing protocols of a defined surface fraction at power levels expected in the application. The fundamental approach of these tests is a simulation of conditions at the upper limits of operation parameters encountered in practice.

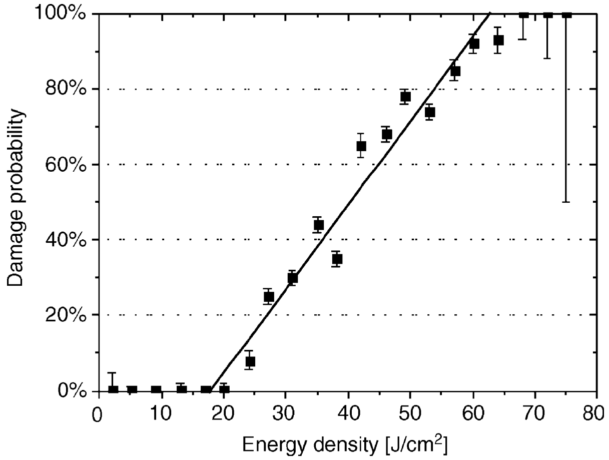

Figure 5 Survival curve of an anti-reflective layer system (two layer V-coating, SiO_2/Ta_2O_5) on fused silica substrate for the Nd:YAG-laser wavelength. The measurement was performed at a wavelength of 1.064 μm, a beam diameter of 420 μm, and a pulse duration of approximately 15 ns, respectively.

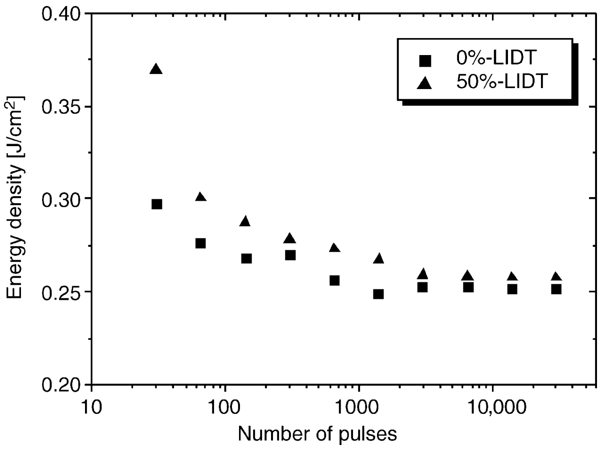

Figure 6 Characteristic damage curve of a high reflecting mirror of SiO_2/TiO_2 for fs-laser systems. The layer structure was tested with a high repetition (1 kHz repetition rate) fs laser system operating at a pulse duration of 150 fs.

Optical Coatings for High Power Lasers

For the development and application of coating systems with high LIDT values, several major aspects have to be considered. In the first approach, appropriate materials and processes have to be selected which deliver coatings with sufficient power resistance. In many studies, a correlation of the LIDT values of the constituent single layers to the power handling capability of a layer system was observed. Therefore, single layers of potential deposition materials were investigated at the wavelengths of prominent laser systems from the VUV- to the FIR-spectral region. For specific deposition processes, often a clear ranking of the coating materials in respect to their power handling capability could be demonstrated. An example for a Nd:YAG-laser with a pulse duration of 15 ns is illustrated by Table 1. With the exception of the ZrO_2-layer, which exhibited severe inhomogeneity and a high inclusion density, a correlation can be assumed between the LIDT value and the melting point of the bulk materials under these operation conditions. This dependency can be attributed to absorption-induced damage as well as to the inclusion model, and therefore, the damage morphology has to be studied in each case to identify the dominating damage mechanism. In the fs-regime or the UV/VUV-spectral range, often a dependence of the power handling capability on the bandgap of the materials is found. According to the relationships of the avalanche and other dielectric breakdown models, materials with highest bandgaps are frequently encountered at the top of an LIDT ranking.

For the influence of the coating process on the laser-induced breakdown of optical coatings, no clear tendencies can be detected. For example, Ion Beam Sputtering (IBS), which is considered as the deposition process for coatings with extremely low losses and contamination, cannot always surpass conventional thermal deposition processes in laser stability. In general, ion or plasma assisted deposition techniques produce coatings with lower LIDT values than conventional coating processes for most operation conditions. Sol-gel processes, which were developed for the deposition of removable coatings on large optics in laser fusion, can reach superior power handling capabilities. In general, as a consequence of the complicated relation between the production parameters and laser-induced breakdown, production processes for optical coatings have to be optimized separately for different wavelength regimes and irradiation conditions.

Table 1 Laser induced damage thresholds of selected single-layer coatings in relation to the melting points of the corresponding bulk materials. In the last column, LIDT values of anti-reflective coatings (AR-Coating) composed of the high index material and SiO_2 are compiled to illustrate the correlation between single LIDT values and the damage threshold of layer systems

Material	*Melting point (°C)*	*LIDT value (J/cm^2) single layer*	*LIDT value (J/cm^2) AR-coating*
TiO_2	1775	13 ± 1	14 ± 1
Ta_2O_5	1918	28 ± 2	32 ± 5
HfO_2	2758	41 ± 3	46 ± 6
ZrO_2	2700	34 ± 4	28 ± 1
Al_2O_3	2072	39 ± 1	
SiO_2	1723	34 ± 7	

Besides the fundamental production parameters, the design of the coating system is of similar importance for the achievement of high LIDT values, because damage is directly driven by the electric field strength in the layer structure. A standing wave field pattern is depicted in Figure 7 for a high reflecting stack of Ta_2O_5 and SiO_2. The design consisting of layers with an optical thickness of 1 QWOT is typical for most standard laser mirrors. The power density reaches extreme values always at the interfaces between the layers and even exceeds the incoming irradiation power density (100%). Interfaces between the layers can be considered as weak points in a coating system, because additional contamination may occur during the switching of the material in the production process. Also, the adhesion between the adjacent layers may be reduced, and mechanical stress may be built up by the different materials. To improve damage thresholds of laser mirrors, the thickness of the first few layer pairs can be adjusted to shift the points of maximum field strength into the bulk of the layer with higher damage resistivity (see Figure 8). Also, a thick layer can be attached to the system in order to stabilize the outer layer pair in respect to thermal or mechanical stress. Another technique, which can be applied to enhance the stability of the interface is the codeposition of materials, resulting in a gradual interface with a mixing zone of materials between the layers. This region of codeposited material exhibits a higher resistance against mechanical stress resulting in an improvement of LIDT values of up to 20%.

The effect of the internal electric field strength distribution on the power handling capability of a layer system should be most apparent for intrinsic damage mechanisms, which are dominant in the ultra short pulse regime below 1 ps. According to the fundamental model (see eqn [4]), a direct relationship of the damage threshold to the maximum field strength value within the layer system is expected. As an example, an investigation in the threshold behavior of ion beam sputtered coating systems with different field strength values is illustrated in Figure 9. In this experiment, the maximum field strength in the last low index layer of SiO_2 has been adjusted to factors between 0.4 and 1.6 of the impinging field strength (see upper diagram in Figure 9) by depositing different designs on a basic 1-QWOT layer stack of TiO_2/SiO_2. In the lower part of Figure 9, the S on 1 LIDT values, measured with a fs-laser, are depicted for selected pulse numbers N. For all layer systems,

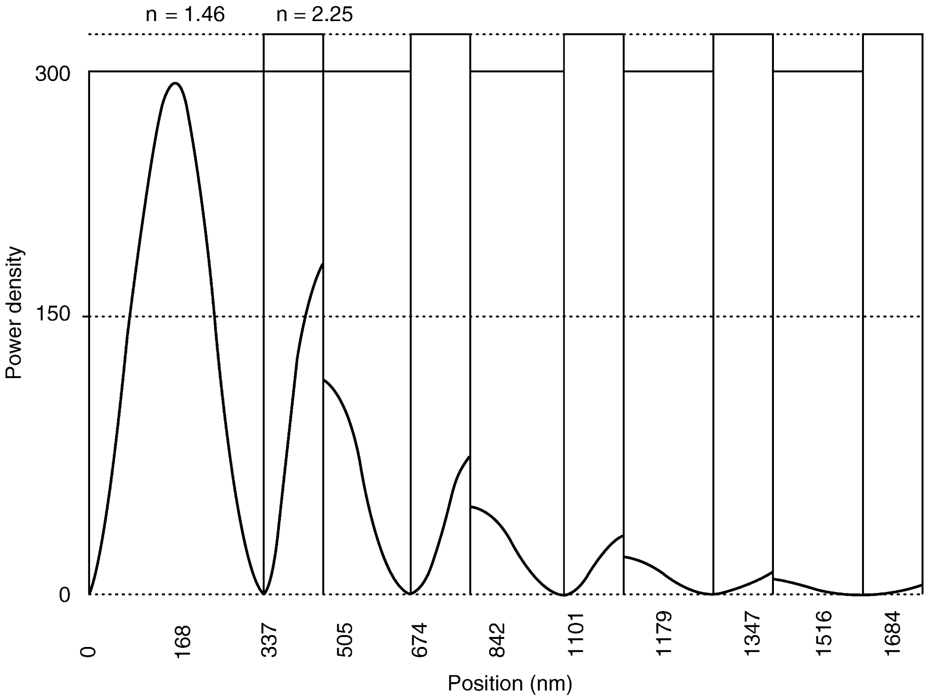

Figure 7 Distribution of the power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system and a final thick stabilizing layer of SiO_2 next to the air interface are depicted. The incoming energy density is calibrated to 100%. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

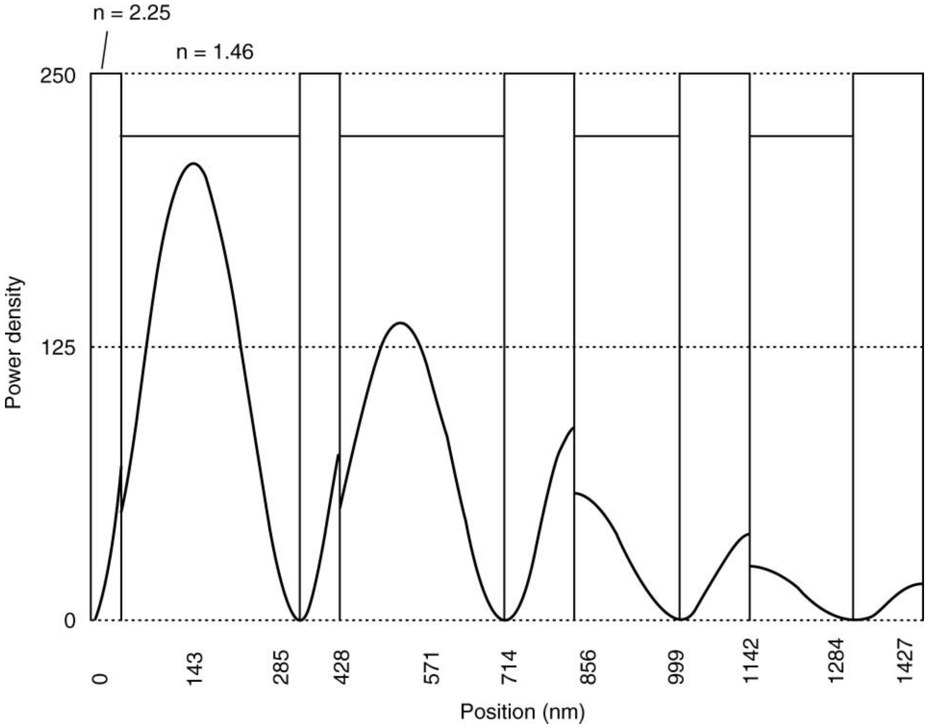

Figure 8 Distribution of power density in a high reflecting layer stack of SiO_2/Ta_2O_5. The first nine layers of the system, which is designed for reduced power density at the interface in the first layer pair are depicted. Reproduced with permission from Ristan D, Ebert J and Welling H (1989) Optische Beschichtungen für Hochleistungslaser. *Laser und Optoelektronik* 21(4): S.53.

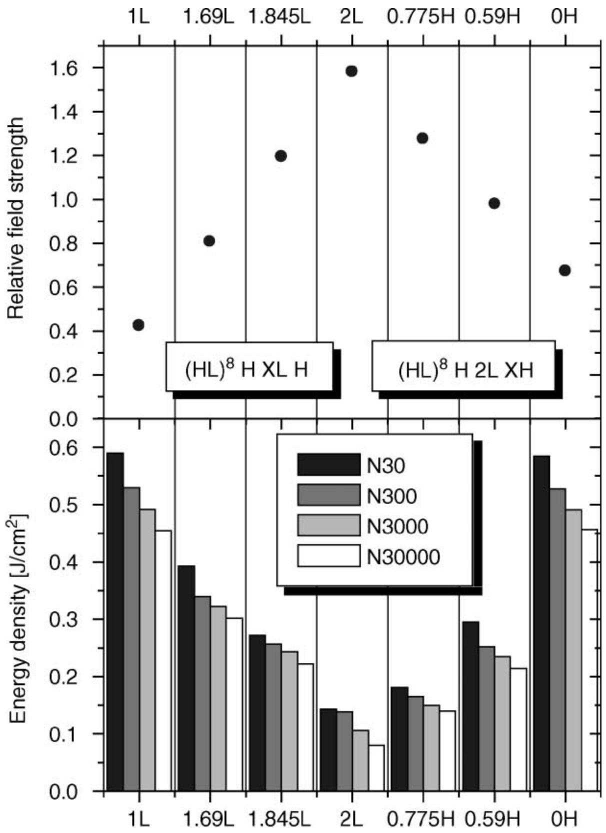

Figure 9 Damage thresholds (lower diagram) and internal field strength relative to the field strength of the impinging wave (upper diagram). The damage thresholds were measured with an ultrashort pulse laser operating at a repetition rate of 1 kHz, a pulse duration of 150 fs, and a beam diameter of 100 μm on the sample surface. Damage thresholds for selected numbers between $N =$ 30 to $N =$ 30 000 of pulses are indicated by columns. The plotted maximum field strength in the last SiO_2-layer is adjusted on the basis of design variation.

which were deposited in separate deposition runs, a strong correlation between the maximum field strength and the laser threshold is observed. This experiment clearly demonstrates the role of the electric field strength distribution in high-power coatings and the potential of advanced thin film design strategies.

Besides the properties of the coatings, the quality of the substrate has to be considered in respect to surface imperfections and to the polishing procedure. For example, substrates for the production of high-quality optics with enhanced power handling capability in the VIS/NIR-spectral range should be polished to a surface roughness of less than 1 nm_{rms} with a surface imperfection value $5/1 \times 0,010$ according to ISO 10110. For an illustration of the effect of the polishing grade on the LIDT values, selected results are compiled in Table 2 for substrates of BK7-glass and high reflecting mirrors.

Table 2 Laser-induced damage threshold values of samples polished using compounds of different grain size. Besides the LIDT values of the uncoated surfaces, also data for anti-reflective coatings of selected materials on these surfaces are summarized

Coating type	*LIDT value (J/cm^2) polishing type I*	*LIDT value (J/cm^2) polishing type II*
Bare surface	76 ± 1	68 ± 12
Ta_2O_5/SiO_2	44 ± 2	38 ± 1
HfO_2/SiO_2	47 ± 16	35 ± 5
Nd_2O_3/MgF_2	34 ± 6	24 ± 8

Substrates with the polishing Type I and Type II were processed with powder of different grain diameter (2 μm and 3 μm) in conventional pitch polishing. For the LIDT values of the bare substrate and the coatings, a clear relation to the surface roughness can be observed. Especially for optical coatings with significant transmittance, subsurface damage in the substrate has to be taken into account. As a consequence of the chemical and mechanical interaction of the surface with the polishing compounds, impurities and dislocation are introduced in the surface structure resulting in a reduction of the damage threshold.

Summary

In the course of the rapid development of laser technology, a large background in high power optical coatings had been built up during the last three decades. Nowadays, the corresponding experience in the production of high power coatings is mainly located at industrial companies and a few research institutes, which are also involved in the characterization of optical coatings. For an illustration of the present state in thin film technology, damage threshold values of advanced optical components are summarized in Table 3 for laser systems and operation conditions often applied in industrial production environments.

Besides the spectacular experiments in laser fusion, isotope separation, and fundamental physics, industrial applications of lasers in material processing, medicine, information technology, and semiconductor lithography are considered as major pacemakers for the progress of high power optics. Therefore, trends in these technology fields will dominantly govern the future development of thin film technology. For example, in semiconductor lithography at a wavelength of 157 nm, which would open the way to feature sizes well below 100 nm, optical coatings with extended lifetimes are still on rank 6 of the list of challenges to achieve an effective lithographic production facility. In next generation lithography, optics

Table 3 Laser-induced damage threshold values of selected optical coating systems for laser applications (Types: HR: high reflecting mirror, AR: antireflective coating, th: thermal evaporation, IBS: ion beam sputtering)

Laser system wavelength	*Type*	*Laser-induced damage threshold ISO 11254*
193 nm, ArF-excimer	AR/th	1–2 J/cm^2 (1 on 1, 20 ns)
	HR/th	2–4 J/cm^2 (1 on 1, 20 ns)
248 nm, KrF-excimer	AR/th	10 J/cm^2 (1 on 1, 30 ns)
	HR/th	$>$20 J/cm^2 (1 on 1, 30 ns)
	HR/IBS	$>$3 J/cm^2 (1 on 1, 30 ns)
1,064 μm, Nd:YAG	AR/th	$>$60 J/cm^2 (12 ns, 0,25 mm)
	HR/th	$>$100 J/cm^2 (12 ns, 0,25 mm)
	HR/IBS	$>$80 J/cm^2 (12 ns, 0,25 mm)
10,6 μm, CO$_2$-Laser	AR/th	$>$20 J/cm^2 (100 ns, 1,4 mm)
		$>$2 kJ/cm^2 (1,2 ms, 250 μm)
		$>$3 kW/mm (cw, 100 μm)
	HR/th	$>$25 J/cm^2 (100 ns, 1,4 mm)
		$>$2 kJ/cm^2 (1,2 ms, 250 μm)

of even smaller wavelengths, around 13 nm, have to be optimized to achieve the 60 nm node. Ultra-short pulse lasers gain importance as innovative tools for material processing, laser medicine, and biology, as well as the analysis and control of chemical reactions. For the development of fs lasers needed in these applications, special high-power broadband coating systems are required. In addition to high LIDT values, these coatings have to fulfil demands with respect to their group delay dispersion for the compensation of dispersion effects in the laser systems. Other challenges have to be expected from applications, where a combination of the high-power handling capability with additional properties has to be achieved. As a typical example, high-power laser coatings with improved mechanical or chemical stability for applications in laser medicine can be considered.

List of Units and Nomenclature

$\mathrm{d}n/\mathrm{d}t$	generation rate of electrons in the conduction band
g	radius of a spherical inclusion; (cm)
$I(t)$	power density of laser radiation,
k	thermal conductivity; (W/(cm °C))
k_E	thermal conductivity of a spherical inclusion; (W/(cm °C))
LIDT	laser induced damage threshold (J/cm^2; W/cm^2; W/cm)
n'	refractive index of a spherical inclusion; (1)
P	output power of a laser; (W)
q	correction factor representing interference effects in coatings
Q	Mie absorption coefficient of a spherical inclusion;
QWOT	quarter wave optical thickness: unit for the thickness of the layer
t_I	irradiation time for a cw-laser; (s)
$T(r)$	temperature in a spherical inclusion at the radial position r; (°C)
w	diameter of a laser beam with Gaussian profile; (cm)
α	impact ionization coefficient; (1)
β_m	m-photon absorption coefficient; (1)
β_S	surface absorption of an optical component; (1)
ΔT	temperature rise in the center of an irradiated circular spot; (°C)
κ	thermal diffusivity; (cm^2/s)
κ_E	thermal diffusivity of a spherical inclusion; (cm^2/s)
λ	wavelength of a laser; (cm)

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Bloembergen N (1973) Role of cracks, pores and absorbing inclusions on laser damage threshold of transparent dielectrics. *Applied Optics* 12(4): 661–664.

Genin FY and Stolz CJ (1996) *Morphologies of laser induced damage in hafnia-silica multilayer mirror and polarizer coatings*, Third International Workshop on Laser Beam and Optics Characterization. In: Morin M and Giesen A (eds). *Proc. SPIE 2870*, 439–448.

Guenther AH, *et al. Proceedings of the Symposium on Laser Induced Damage in Optical Materials, NBS (NIST) Special Publication 1969–1989*, SPIE vols. 1989–2003. CD-Rom version of collected papers from 1969 to 1998 available: Washington: SPIE.

Guenther KH, Ebert J, Kiesel E, *et al.* (1984) 1.06-m laser damage of thin film optical coatings: a round-robin experiment involving various pulse lengths and beam diameters. *Applied Optics* 23: 3743.

ISO 11254 (2001) *Test methods for laser induced damage threshold of optical surfaces*. Part 1: 1 on 1-test, 2000, Part 2: S on 1 test.

ISO 11254-3 (2002) *Laser and laser-related equipment – Determination of laser-induced damage threshold of optical surfaces – Part 3: Assurance of laser power handling capabilities*. Draft Standard. International Organization of Standardisation.

Jasapara J, Nampoothiri AVV, Rudolph W, Ristau D and Welsch E (2001) Femtosecond laser pulse induced breakdown in dielectric thin films. *Physical Review B* 63(4): 045117.

Puttick K, Holm R, Ristau D, *et al.* (1997) *Continuous wave CO$_2$ laser induced damage thresholds in optical* *components*, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. SPIE vol. 3244, 188–197.

Rudolph W, Mero M, Liu J, *et al.* (2003) *Femtosecond pulse damage and pre-damage behavior of dielectric thin films*, Proceedings of the 34th Annual Symposium on Optical Materials for High Power Lasers. SPIE vol. 4932, 202–215.

Walker TW, Vaidyanathan A, Guenther AH and Nielsen P (1979) *Impurity breakdown in thin films*, Proceedings of the Symposium on Laser Induced Damage in Optical Materials. NBS Special Publication, No.568, 479–495.

Wood R (1990) *Laser Damage in Optical Materials.* Bristol and New York: Adam Hilger. Reprint with corrections.

Optical Black Surfaces

S M Pompea, National Optical Astronomy Observatory, Tucson, AZ, USA

S H McCall, Stellar Optics Research International Corporation (SORIC), Thornhill, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Black surfaces play an important role in many, if not most, optical systems. Optical instruments and systems of all types often rely on black surfaces to help minimize the effect that stray or off-axis light can have in degrading optical system performance. The strategic selection and placement of appropriate black surfaces can often limit significant detrimental stray light effects in an optical systems, thereby dramatically improving system performance. The high emissivity of black surfaces also gives them an important role in the design and construction of black bodies, calibration surfaces, and radiator surfaces.

The characterization and selection of black surfaces is an important field of optics and optical system design. However, the selection of black surfaces is a specialized undertaking requiring careful study to ensure the proper selection of particular surfaces for each optical system. A black surface that works well in one application may not be at all appropriate for a different application, where different systems performance goals are desired. The wavelength of operation of the system is an important consideration since a surface which has low reflectance at one wavelength band may have different reflective properties at another wavelength or range of wavelengths. Specialized measuring devices called goniophotometers are often used to characterize the reflectance and scatter of black surfaces for a given wavelength or set of wavelengths. Goniophotometric measurements can be used to create a specialized functional description of reflectance and scatter off a surface, called the Bidirectional Reflectance Distribution Function (BRDF) or Bidirectional Scatter Distribution Function (BSDF). In practice, having the BRDF or BSDF of a surface is extremely valuable for surface selection and is more useful than other reflectance and scatter measurement measurements or descriptions. Comparing the measured BRDF or BSDF of several surfaces helps to distinguish the general optical properties of these surfaces and allows a direct comparison of the value of these different surfaces in an optical system as they contribute to overall system performance.

The BRDF distinguishes, at a particular wavelength, how a surface specularly reflects or how it scatters light in different directions when incident light from a particular direction relative to the surface (at a given angle of incidence) interacts with the surface. By understanding the BRDF of a surface, the optical designer can understand how the surface behaves in a given system location when illuminated from a specific direction. The BRDF of a black surface is also important because it can be used as a mathematical function in stray light analysis programs to predict system performance. The BRDF and other properties of surfaces can be input into stray light codes so that the propagation of light from black surfaces can be modeled through Monte Carlo-based propagation models. Using these models, the stray light can be directed through the placement of black surfaces into directions less likely to degrade system performance. Thus a measure of stray light control can be gained through absorption by black surfaces or by redirection of the stray light into less critical directions.

Black surfaces may be considered a subset of a class of surfaces known as spectrally selective surfaces. The term ‘spectrally selective surface’ indicates that the spectral properties of many surfaces are different when examined in various spectral regions. For applications such as the design of blackbodies, space radiators, and baffles, surfaces with unique spectral emissive and reflection properties are needed. Large databases of spectrally selective surfaces are available to allow the optical designer flexibility in choosing surfaces for applications where specific optical reflectance and optical scatter properties are desired. Spectrally selective surfaces are important for applications that range from solar energy to telescope design. In some applications, it is desirable for a surface to be highly absorbing at one wavelength and more reflective at another, rather than to be highly absorbing of light over a wide range of wavelengths. An analysis of the spectral properties of a large number of surfaces allows the optical designer to choose surfaces optimized for a particular application.

Black surfaces come in many types, but may be generally characterized as being more shiny (specular) or more diffuse (Lambertian) in their reflective and scattering properties, depending on the form of their BRDF curve. Surfaces are often described in computer models as a combination of their respective specular and Lambertian properties for a given wavelength. Many black surfaces have significant reflectivity at some region of the electromagnetic spectrum, even if they are highly absorbing at one wavelength. Some black surfaces are constructed to be remarkably nonreflecting over a range of wavelengths; these surfaces are often used as baffle surfaces for telescopes and instruments that must operate over a wide range of wavelengths. No black surface can work well in all optical systems. Optical designers must understand the specific optical properties of a black surface to determine if the surface can work in their particular system. To characterize a surface, it is helpful to know at least the BRDF of the surface at the wavelength of interest as well as basic materials properties of the surface. The specific BRDF at the most appropriate wavelength for the system can be used with the optical system stray light model to determine if the surface can fulfill the designer's, expectations of it for the system performance. The placement and choice of black surfaces in an optical system are often critical decisions with highly significant effects on overall system performance.

For this reason, optical designers rely greatly on databases of the reflection and scatter properties of black surfaces in order to make the right choice of surface. These databases have been created to provide the optical engineer with the information needed to understand a wide variety of candidate surfaces and select the most appropriate ones. These databases not only describe the optical properties of surfaces but also elucidate material properties so that the selection process can examine the heat resistance, outgassing rates, fade resistance, durability, and a myriad of other surface or material properties that may be relevant.

The reflectance and scatter data on black surfaces are often taken using a variety of reflectance-measuring instruments and the data format often varies greatly. When published and unpublished data sets are compiled, they are often confusing and sometimes even contradictory. This has made the comparison of surfaces a difficult proposition. In recent years, there have been successful attempts to gather existing large amounts of existing data so that comparisons among black surfaces can be made in a data-rich environment so that anomalous data can be ignored. A significant amount of reflectance data on black surfaces has been created by private or governmental organizations such as NASA and NIST – the US National Institute for Standards and Technology. These data now appear in databases of BRDF and reflectance measurements. Designers are relying increasingly on databases of BRDF surfaces rather than the individual measurements that can be found in the optical physics and applied optics published literature.

Uses of Black Surfaces

Table 1 lists some of the major areas in which black surfaces or black coatings may be used in instruments and optical systems. The optical properties of desirable surfaces may be different for the specific categories in the table. Plausible surfaces for these different applications can be found in materials databases such as the SOLEXIS™ database which contains a useful compilation of data from private

Table 1 Possible uses for black surfaces and black coatings in optical systems

Apertures	Baffles	Barrels	Calibration screens	Choppers
Cold shields	Detectors	Dewars	Diffusion screens	Domes for telescopes and instruments
Enclosures and curtains for rooms and instruments	Integrating sphere enclosures	Lens edges	Light traps	Mounts
Radiators	Radiometers	Solar collector absorbers and concentrators	Standards and references of reflectance	Targets

and governmental data compilations from the US, Canada, and the former Soviet Union.

Selection of Optical Black Surfaces

The selection of appropriate black surfaces for a given application becomes progressively harder as the number of optical requirements (e.g., reflectivity, scatter, emissivity, etc.) and materials requirements (e.g., stiffness, coefficient of thermal expansion, etc.) increase. Because of this, the selection process for specialized surfaces for space applications can be quite involved, since there are a large number of criteria, each of which may be regarded as an essential requirement.

The optical system designer must consider at least the following aspects of the system in selecting a surface:

- the wavelength(s) and waveband(s) of interest;
- the position of the surface in optical system;
- the general robustness required of the surface;
- the environment of the optical system;
- the cost including inspection, handling, and assembly; and
- the production and delivery schedule, including spares.

The selection process must carefully address the system needs as well. Poor decisions are often made when only a cursory study of potential black surfaces is made or when databases and the literature on the subject are neglected. Therefore, an important part of the selection process is the management of the timing of the selection, in order to give the designer time to access available information and to influence the system design in a timely manner.

If black surfaces are used for stray light control, then the use of these surfaces must be planned at the beginning of the optical system design process. If the placement of black surfaces and the array of potential surface treatments are examined early in the program, the optimal choices can often be determined in a careful manner that enhances the system performance. If the selection is saved until the very end of the program, the choice of surfaces will be more limited and improvements in system performance may be minimal, even with the most excellent black surfaces. This is because the location of the surfaces in the system often proves to be as critical as the choice of surface. If the surface is in the wrong location, the best optical black may not be able to improve the stray light performance. For best results, the layout of the system and the placement of baffles must be considered first, before the baffle surfaces are selected.

An optimal flowchart of good system design for black surface selection is given in Figure 1. Figure 2 describes a work flowchart where black surface selection is neglected until the end of the program, often with dire consequences for system performance, cost, and schedule.

For space-based systems, the effect of atomic oxygen in a low-earth orbit environment is an important consideration. For space systems, the designer must also examine how black surfaces may be affected by or contribute to contamination and outgassing. Other considerations are the generation of particles during launch, the effect of solar

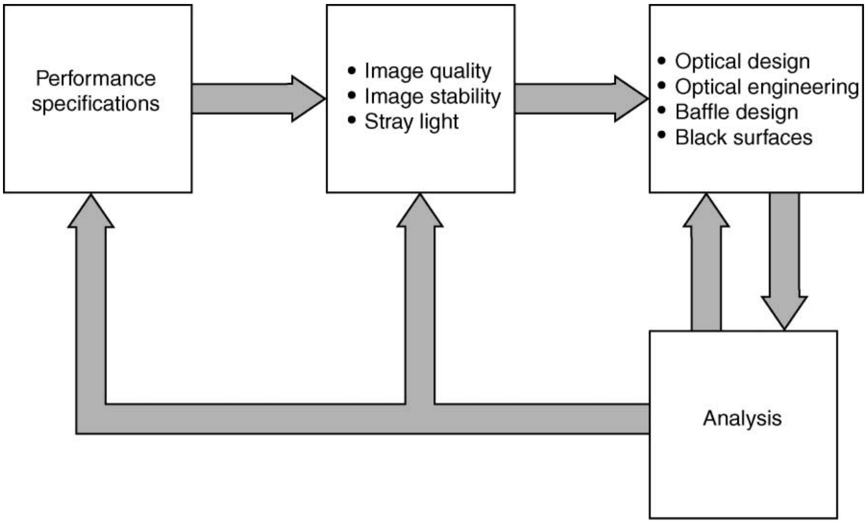

Figure 1 Diagram of the system design process which treats black surface selection as an important part of basic system design, to be done early in the design process. This gives the designer an opportunity to conduct system tradeoffs to optimize system performance.

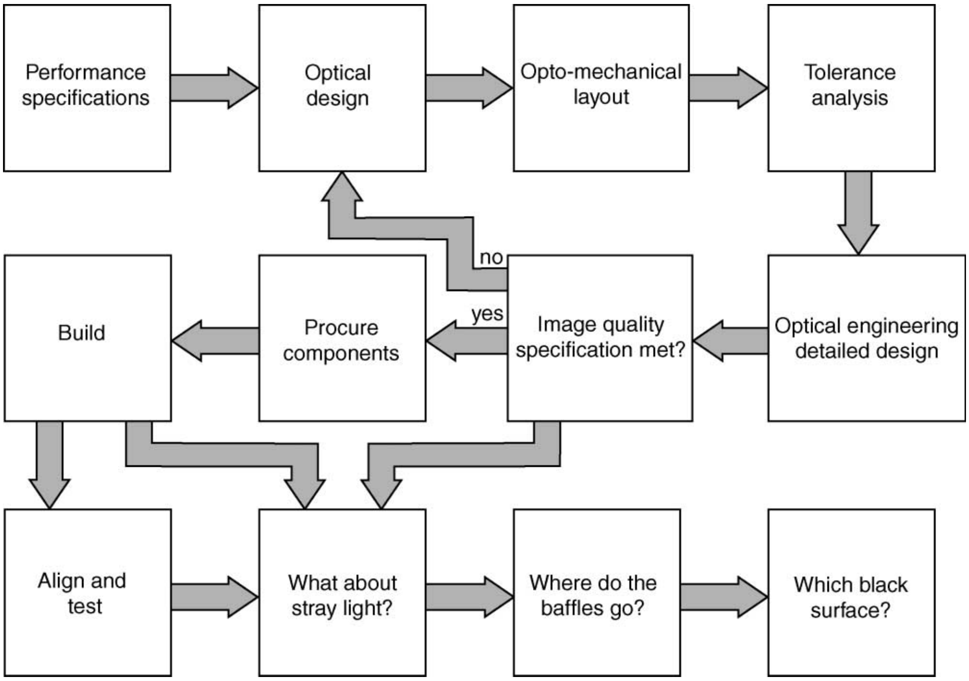

Figure 2 Diagram of a poor system design process. This process addresses the selection of black surfaces late in the system design process, where selection is significantly more difficult, expensive, and risky. Even with a late selection, there may be severe impacts on schedule.

ultraviolet radiation on black surfaces, and how spacecraft charging may affect conductive, dielectric, or insulating black surfaces. Since black surfaces may have a large surface area in an optical system, the designer must ensure that the black surfaces do not contribute significantly to the risk of system failure through some degradation process that may contaminate other optical surfaces. For space-based systems, data on surface outgassing and the past performance of a surface in a specific space environment is essential.

There are a variety of resources available to assist in the selection of black surfaces. The best databases allow the user to create 2-dimensional and 3-dimensional BRDF plots for extensive lists of potential black surfaces. For most surfaces, with no preferred directionality in the structure of the surface, a 2-dimensional BRDF plot is adequate. Most of these databases provide information on a wide variety of spectrally selective materials including black, white, and reflective surfaces for thermo-optical applications for ground and space-based instrumentation. The BRDF plots can be examined by the user or fed into existing, commercially available optical design and analysis software packages. The results of these models indicate whether the chosen surfaces give adequate performance or whether more specialized coatings are necessary to achieve the desired level of performance.

A significant problem for many designers is that the formulations of many paints and other surface treatments often change. Sometimes the name of the coating remains the same even as the optical properties change, due to a change in the compounds or methods used to prepare the coating. Sometimes the name changes with or without a change in the formulation. These often unpublished changes increase the risk of using older plots from the published literature as definitive optical modeling data. Experts in the optical blacks area continuously investigate these changes and their implications and can advise optical designers of the implications of recent changes in the paint and coating arena. These experts are often aware of situations where coating failures have occurred or where mission performance was degraded by unanticipated changes in coatings, even if reports of this incidents were never published.

In summary, the selection process, by its very nature, must be a deliberate and careful effort to examine the best surface options. Some of the better surfaces which were considered laboratory standards (e.g., 3M Black) are no longer easily available and replacements must be found. For the most systematic decision-making process, access is needed to the best optical black surface databases, often commercial databases with hundreds to thousands of BRDF curves on several hundred materials. The use of these optical surface databases can substantially reduce the costs, time, and scope of a measurement program, as well as streamline the selection process. Additionally it allows the designer to consider materials used successfully in other programs which may not be familiar to the designer. Finally, the designer may avoid using materials that have been rendered unreliable through changes in the manufacturing process or in the coating chemistry.

Characterization of Black and Spectrally Selective Surfaces

Optical Characterization

An understanding of black surfaces relies heavily upon an understanding of the meaning and use of the BRDF. While the optical properties of surfaces can be characterized through specular and diffuse reflectance measurements, the use of BRDF measurements has become extremely important. BRDF measurements are equally adept at characterizing highly specular and highly diffuse surfaces and the myriad of surfaces with both characteristics. The BRDF is often used in general to describe the angle-dependent optical scatter from these surfaces. This is important as the angular distribution of scatter can be used to calculate if the scatter from a black surface will be a limiting factor in the resolution or noise level of the system. BRDF measurements can even be used in computer graphics visualization programs that illustrate what a surface may look like under various illumination conditions. For example, the BRDF is used in computer graphics models to render a certain color for an automobile or a countertop in a kitchen. For some black surfaces, the angular distribution of scatter can be used to calculate a variety of surface parameters such as the surface roughness, given certain assumptions about the surface. For opaque black surfaces, BRDF measurements have proved to be one of the best overall descriptors of the optical properties of the surface.

BRDF is defined as the ratio of scattered radiance (watts/(cm^2 sr)) to surface irradiance (watts/cm^2); its units are inverse steradians. Radiance is used in the definition, in order to make the BRDF independent of the parameters of the measuring instrument, such as the detector aperture and distance to the detector. BRDF measurements are made by a variety of instruments for a number of wavelengths. For example, BRDF measurements of surfaces have been made in wavelengths ranging from the vacuum ultraviolet to the far infrared. The most common BRDF measurements made on black surfaces are made at laser wavelengths of 0.6328 and 10.6 micrometers. However, these measurements may be of limited value if the system is operating at substantially different wavelengths, where the surface's BRDF is significantly different.

To make a BRDF measurement requires a light source, a sample mounted and illuminated by that source, a receiver (detector) to measure the scatter from the sample, and the computer/electronics package to accurately record the detection of the scattered light. Additionally the sample and/or light source must be able to be moved so that the sample can be illuminated from nearly every angle, and scattered light from the surface can be detected in a hemisphere around the sample. The geometry of the BRDF measurement is illustrated in Figure 3.

To make a BRDF measurement on black surfaces, appropriate sources, detectors, and low-scatter optics must be used. This becomes more difficult for wavelengths less than about 0.25 micrometers. At longer, infrared wavelengths, diffraction effects become important. Despite the instrumental difficulties, it is often better to measure the amount of scatter at a specific wavelength λ and angle θ_s, rather than try to predict it from theoretical models. This is because the extrapolation to new angles or wavelengths using scaling theories is not always accurate or even possible. Methods for wavelength scaling (extrapolating from BRDF data for one wavelength to data for any other wavelength) or angle scaling (where data taken at particular incident and scatter angles are extrapolated to other angles) will not be discussed as BRDF curves are now widely available for many surfaces at a variety of wavelengths and angles.

The definitions of BRDF are given in Table 2.

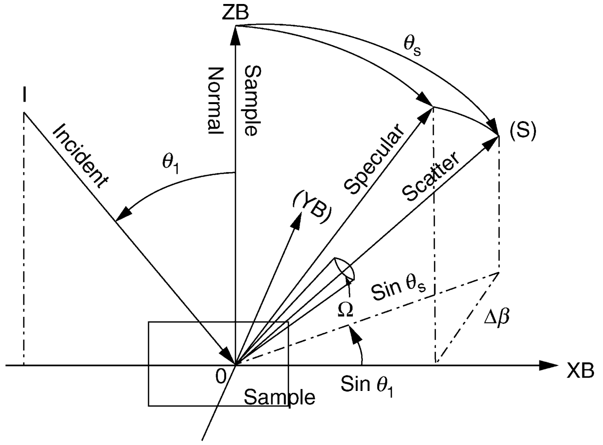

Figure 3 Geometry of BRDF measurement and the angle conventions accepted by the American Society for Testing Materials (ASTM). The plane of incidence is the I-0-ZB plane and the scatter plane is the S-0-ZB plane.

Table 2 BRDF term definitions

Bidirectional reflectance distribution function BRDF	The sample radiance divided by the sample irradiance. The procedures given in this practice are correct only if the field of view (FOV) determined by the receiver field stop is sufficiently large to include the entire illuminated area
Scatter	The radiant flux that has been redirected (by interaction with the sample) over a range of angles
Sample coordinate system	A coordinate system is fixed to the sample and used to a specify position on the sample surface for the measurements. The sample coordinate system may be application and sample specific. A Cartesian coordinate system is recommended for flat samples, with the origin at the geometric center of the sample face with the Z axis normal to the sample
Plane of incidence, PLIN	The plane defined as containing the central ray of the incident flux and the sample normal
Beam coordinate system, XB, YB, ZB	A Cartesian coordinate system defined with the origin on the central ray of the incident flux at the sample surface, the XB axis in the plane of incidence (PLIN) and the ZB axis normal to the surface. The angle of incidence, scatter angle, and incident and scatter azimuth angles are defined with respect to the beam coordinate system
Incident direction	The central ray of the incident flux specified by θ_i and φi in the beam coordinate system
Angle of incidence, θ_i	Angle between the central ray of the incident flux and the ZB axis
Incident azimuth angle, φ_i	The fixed 180° angle from the XB axis to the projection of the incident direction onto the XB–YB plane. It is convenient to use a beam coordinate system in which $\varphi_i = 180°$ since this makes φ_s the correct angle to use directly in the familiar form of the grating equation. Conversion to a sample coordinate system is straight forward, provided the sample location and rotation are known
Scatter direction	The central ray of the scattered flux specified by θ_s and φ_s in the beam coordinate system
Scatter angle, θ_s	Angle between the central ray of the scattered flux and the ZB axis
Scatter azimuth angle, φ_s	Angle from XB axis to the projection of the scatter direction onto the XB–YB plane
Specular direction	The central ray of the reflected flux that lies in the PLIN with $\theta_i = \theta_s$ and $\varphi_s = 0$
Delta theta, $\Delta\theta$	The angle between the specular direction and the scatter direction
Delta beta, $\Delta\beta$	The projection of $\Delta\theta$ onto the XB–YB plane, that is, the $\Delta\theta$ angle measured in direction cosine space. For scatter in the PLIN, $\Delta\beta = \sin\theta_s - \sin\theta_i$. For scatter out of the PLIN, the calculation of $\Delta\beta$ becomes more complicated
Receiver solid angle, Ω	The solid angle subtended by the receiver aperture stop from the sample origin
Sample radiance, L_e	A differential quantity that is the reflected radiant flux per unit projected receiver solid angle Ω per unit sample area. L_e is often an average calculated from scattered power, P_s collected by the projected receiver solid angle $\Omega \cos\theta_s$, from the illuminated area, A. The receiver aperture and distance from the sample determines Ω and the angular resolution of the instrument
Sample irradiance, E_e	The radiant flux incident on the sample surface per unit area. In practice, E_e is an average calculated from the incident power, P_i, divided by the illuminated area, A. The incident flux should arrive from a single direction; however, the acceptable degree of collimation or amount of divergence is application specific and should be reported

Table adapted from ASTM E1392-90 "Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces," American Society for Testing Materials, 1990.

Given these definitions, BRDF, is mathematically described by

$$\mathrm{BRDF} = \frac{L_e}{E_e} = \frac{(P_s/\Omega A \cos\theta_s)}{(P_i/A)} = \frac{P_s}{P_i \Omega \cos_s}\ [\mathrm{sr}^{-1}]$$

The BRDF is a function of many measurable quantities. It depends on the angle of incidence of the incident flux and the wavelength of this radiation. It also depends on the direction in which the scattered light is measured as well as the wavelength and polarization of the incident and scattered flux. The BRDF can be calculated for a given angular configuration of incident source and detector from the average radiance divided by the average irradiance. The BRDF is therefore referenced by the incident power. The cos θ_s term in the formula accounts for the foreshortening of A as the angle θ_s increases, so that true radiance is being used in the calculation. Because of this, the BRDF is constant for a surface with Lambertian (uniformly diffuse) properties. The detector must have a field of view so that all of A can be viewed and such that P_i includes the entire beam. In practice, P_i is measured by removing the sample from the measurement position. Then the detector is moved into the straight-through position. Often a reference detector is used to monitor the incident beam power so that any changes after the initial V_i, measurement can be recorded.

For BRDF measurements to be meaningful, the instrument signature must be examined and understood. This is particularly important in the measurement of low-scatter surfaces. The instrument signature is the BRDF when the sample is removed and there is no sample scatter. In this case the stray light and scatter signature of the device will be apparent. Angular positioning errors and electronic noise can also be important. For example, as θ_s approaches 90° the accuracy of measuring θ_s becomes critical due to the $1/\cos\theta_s$ term in the BRDF equation. Ideally the instrument signature is a small fraction of the BRDF being recorded. An understanding of the instrument signature provides the needed information to decide where the BRDF data loses its meaning.

The BRDF is most often displayed in a plot, where the BRDF value (vertical axis) and the angle of scatter (horizontal axis) are plotted on logarithmic scales. In these 'log–log' plots it is easy to examine the scatter distribution close to the specular direction which, for shiny surfaces, represents the bulk of the scattered light. For surfaces that are diffuse in nature, a log–log plot of the BRDF generally yields a straight line. The slope of this plot is characteristic of the surface of the process which is used to produce the surface. This straight-line plot is useful in comparing samples made with the same process. Deviations from the slope of the line or the value of the points on the line indicate that there may be problems with a sample. Thus the simplest scattering instrument can be used for quality control of black samples. For a fixed angle of incidence and a fixed detection angle, variations in the BRDF value may indicate changes in the materials or the processes which created their surfaces. The quality control process can be performed by looking for variations in the BRDF among the different samples.

The BRDF can be used as a measure of surface roughness – an important attribute for a diffuse black surface. As random surface roughness features increase on a surface, the scatter from the surface also increases. Because of this, the BRDF can be used to characterize surface roughness parameters. If the surface roughness is all in the form of small, micro-rough surface features, the BRDF value then is very much proportional to the surface power spectral density function (PSD), which is a measure of the power plotted against spatial frequency. The characterization of surface roughness by BRDF scatter measurements only works if the scattered energy is due to surface micro-roughness. If the scatter is from larger surface features or from surface contamination, then it will be difficult to accurately characterize the surface roughness. Thus only smooth, mirror-like black surfaces can be used for this simple leap from BRDF values to surface roughness parameters. Some sample BRDFs of black surfaces are found in Figures 4–7.

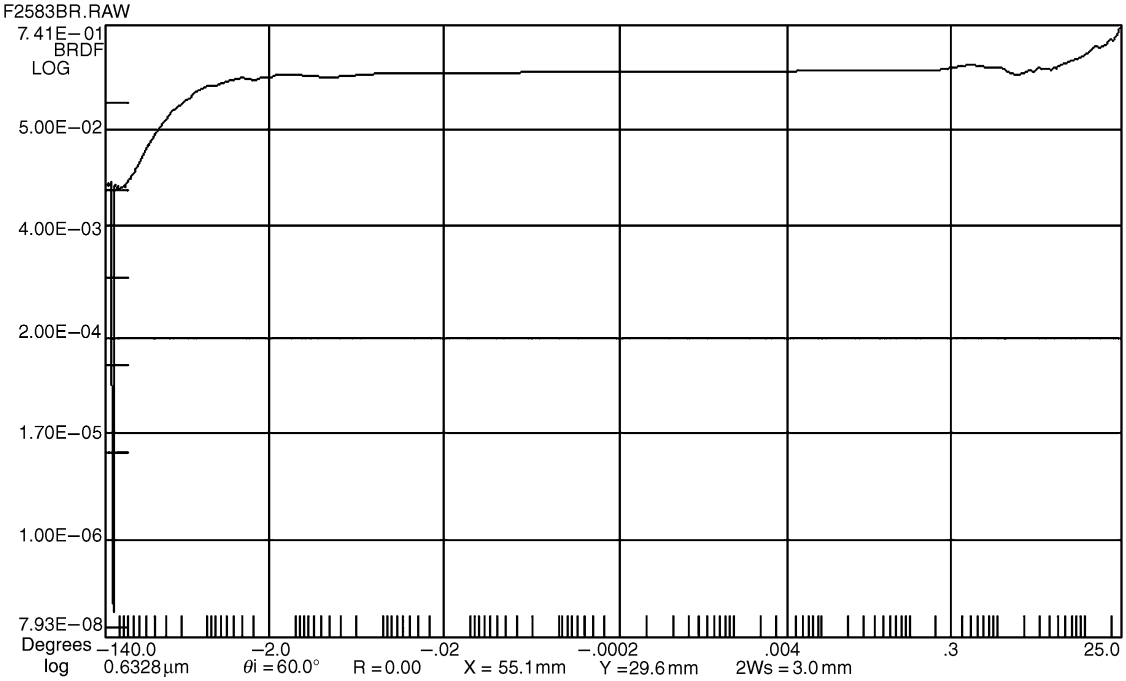

Figure 4 BRDF at wavelength of 0.633 micrometers and 60° incidence angle of black anodize sample. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

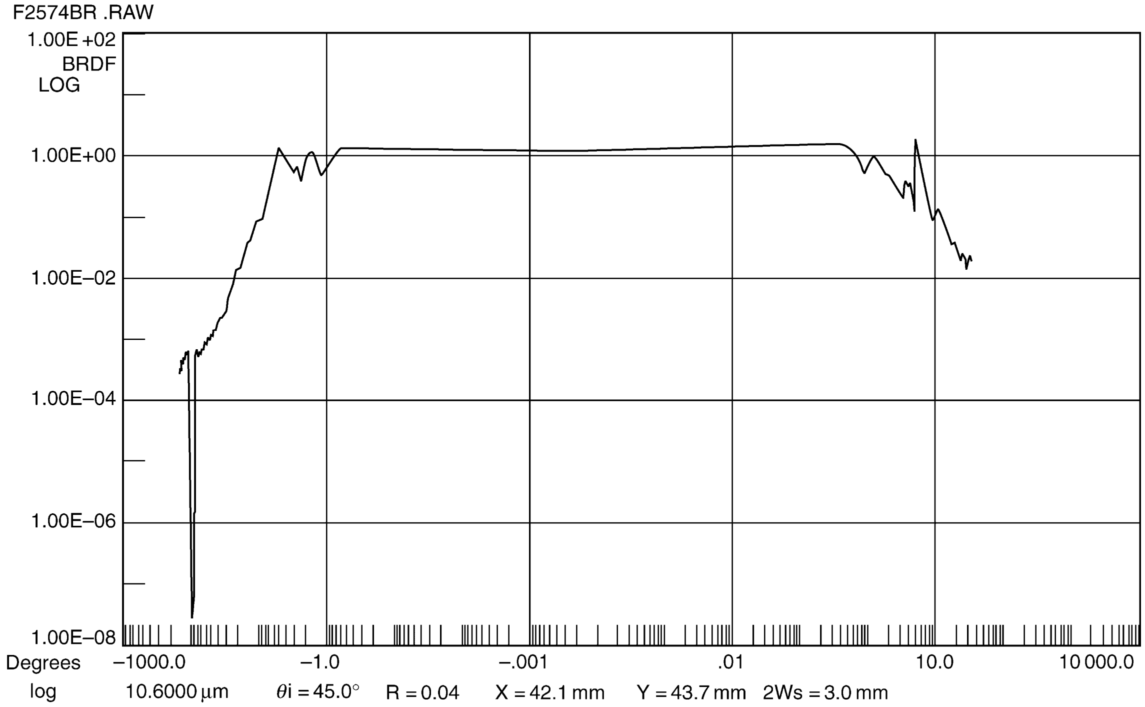

Figure 5 BRDF at wavelength of 10.6 micrometers of black anodize at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

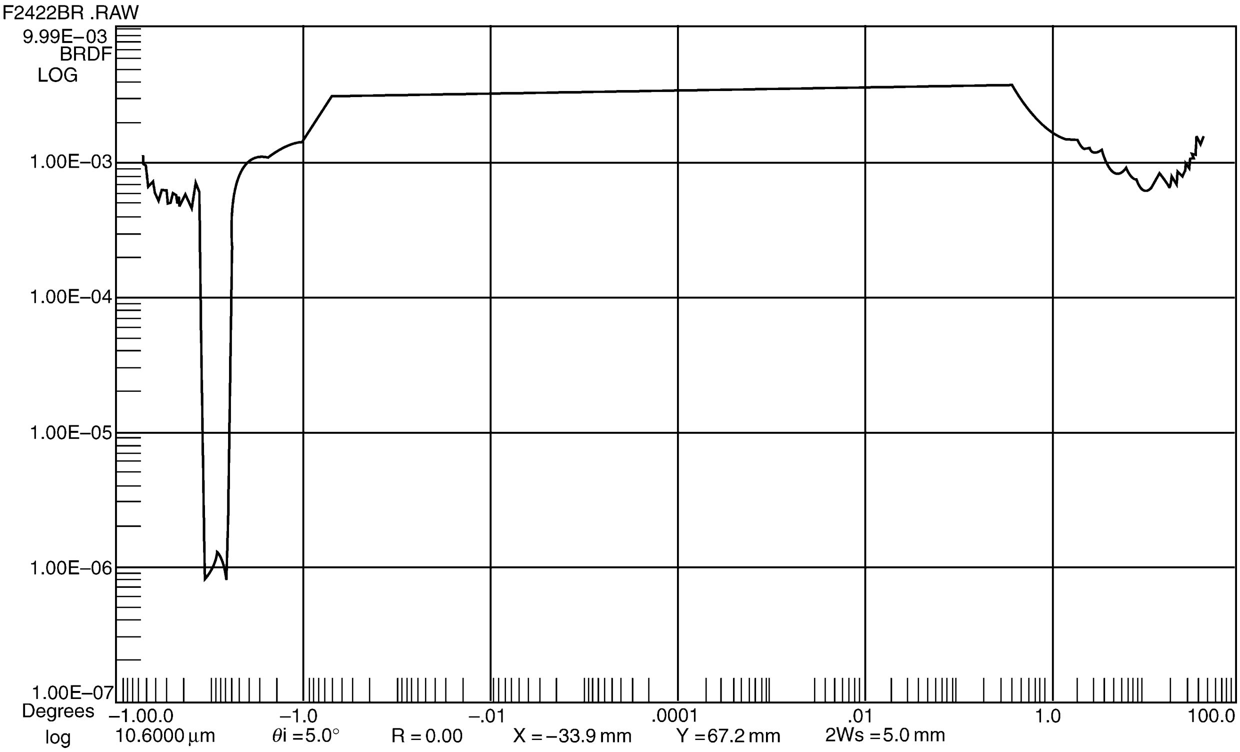

Figure 6 BRDF at a wavelength of 10.6 micrometer of Martin Optical Black at 45° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

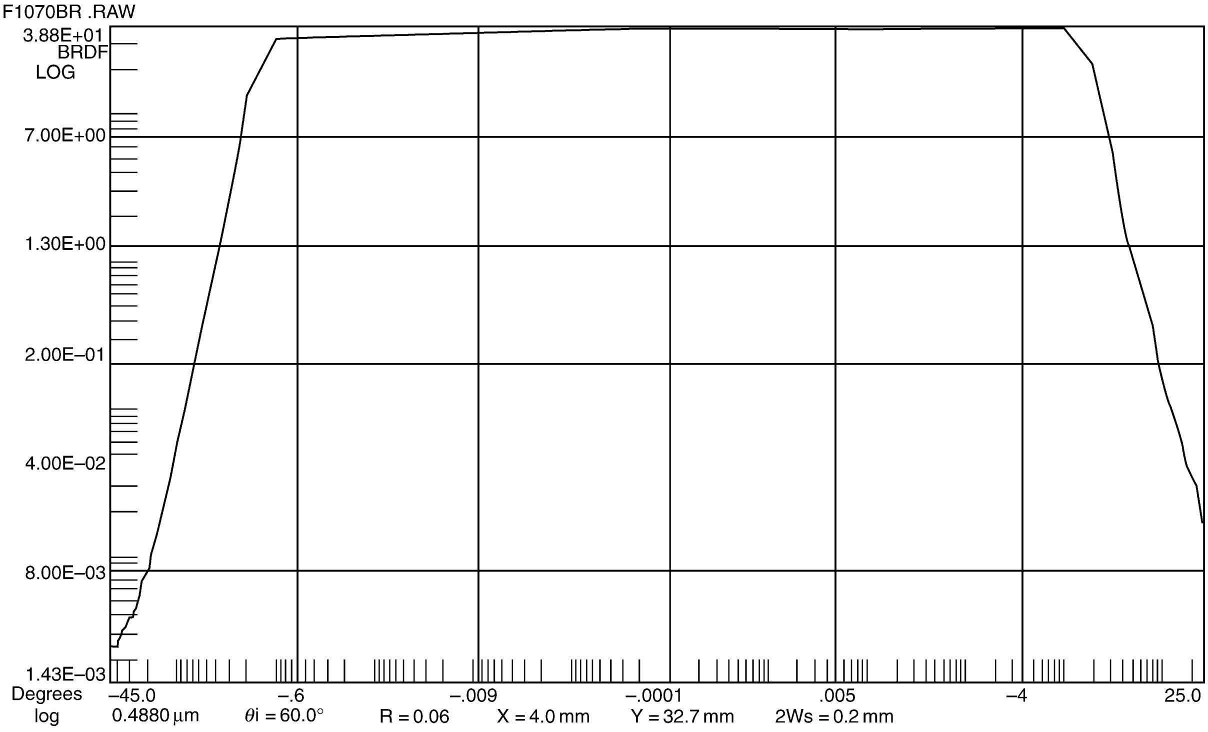

Figure 7 BRDF at a wavelength of 0.488 micrometers of a glossy black surface at 60° incidence. Plot generated by SOLEXIS using CASI software from Schmitt Measurement Systems, Portland, OR, USA.

Examples of Black Surfaces

Black surfaces appear in many forms. They may be either diffuse or specular in nature, or both. Black surfaces may be created by painting a surface, by altering a surface (such as through an anodization process), or by any number of other surface treatment effects. For example, absorbing compounds may be added to paint or to a surface to decrease reflection. Adding large (compared to the wavelength) craters or holes to a surface may result in multiple reflections within the surface leading to decreased reflection. Surfaces may be modified to increase their light-scattering ability, with the hope that reflection in a specific direction will be reduced. Sometimes surfaces may be created in such a way as to enhance thin-film interference effects, leading to reduced reflection. These and other techniques may be combined together in an attempt to reduce reflection and make the material appear blacker in a particular wavelength range of interest or in a particular direction.

Some sample black surfaces are given in Table 3, although databases are much more current on current names and manufacturers.

Many black surfaces use surface texture to enhance their light-absorbing ability. Scanning electron micrographs of surfaces reveal the surface morphology and how multiple reflections can be used for light trapping. Surface chemistry, changes in the index of refraction, and the use of particles of different sizes are also used to create better absorbers. A variety of other proprietary methods are used as well (Figures 8–11).

Future Developments

There are many new kinds of black surfaces being developed for specialized applications. Black surfaces can be created using a variety of new techniques and materials. For example, new surfaces can be created using laser ablation or chemical microetching of surfaces. Surfaces can also be created using new materials such as carbon nanotubes or by using the technology used to create flocked materials. These new kinds of surfaces generally have good absorption at a variety of wavelengths. Even with the creation of these new surfaces, the greatest gain in an optical system will probably be attained through a concerted effort to choose the appropriate black surfaces early in the program and to place them properly in the optical system. An early selection of surfaces and an analysis of their performance will lead to a higher level of system performance than a search for an exotic, ultrablack surface at a late stage

Table 3 Selected optical black surfaces

Name	*Manufacturer*	*Notes*	*Type*
Aeroglaze L300	Lord Corporation Erie, PA	Formerly called Chemglaze L300	Paint
Aeroglaze Z004	Lord Corporation Erie, PA	Formerly called Chemglaze Z004	Paint
Aeroglaze Z302	Lord Corporation Erie, PA	Formerly called Chemglaze Z302	Paint
Aeroglaze Z306	Lord Corporation Erie, PA	Formerly called Chemglaze Z306	Paint
Aeroglaze Z306 with microspheres	Lord Corporation Erie, PA	Formerly called Chemglaze Z306 with microspheres	Paint
Aeroglaze Z307	Lord Corporation Erie, PA	Formerly called Chemglaze Z307	Paint
Aeroglaze Z313	Lord Corporation Erie, PA	Formerly called Chemglaze Z313	Paint
Ames 24E (Ames 24E2)	NASA Ames Research Center Moffett Field, CA		Paint
Cardinal 6450	Cardinal Industrial Finishes South Elmonte, CA	Formerly 'Cardinal 6550'	Paint
Cornell Black	Cornell University		Paint
DeSoto Flat Black	Courtaulds		Paint
Electrically Conductive Black Optical Paint	Jet Propulsion Laboratory Pasadena, CA	Has no trade name	Paint
IITRI Bone Black D-111 (IITRI D111)	Illinois Institute of Technology Chicago, IL		Paint
LMSC Black	Lockheed Martin Palo Alto		Painted multilayer coating
MH21-I	Illinois Institute of Technology Chicago, IL		Paint
MH55	Illinois Institute of Technology Chicago, IL		Paint
MH2200	Illinois Institute of Technology Chicago, IL	Formerly 3M's ECP 2200 paint, but sold to IIT	Paint
Solarchem	Eastern Chem Lac Corporation Malden, MA		Paint
463-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 463-3-8' diffuse black paint	Paint
443-3-8	Akzo Coatings, Inc. Orange, CA	Formerly called 'Cat-a-lac 443-3-8' glossy black	Paint
443-3-17	Akzo Coatings, Inc. Orange, CA	Formerly called 'Sikkens 443-3-17'	Paint
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Martin Black, Enhanced	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Infrablack	Lockheed Martin Denver, CO	For Al substrates only	Anodization process
Black Chrome Type Surfaces	Lockheed Martin Denver, CO	For many kinds of metal substrates	Electrodeposition process
Black Cobalt Type Surfaces (Cobalt Black)	Lockheed Martin Denver, CO and many more companies	For many kinds of metal substrates. References to Black Copper, Black Steel etc. are covered by Black Cobalt	Electrodeposition process, and can be followed by chemical or thermal oxidation
Black Nickel	−NIST	For many kinds of metal substrates	Deposition and etching process
(NBS Black)	−Ball Aerospace		
(Ball Black)	−Lockheed Martin and others	Ball modified the process developed by NBS	
Black Kapton™ Foil			Foil
Black Tedlar™ Foil			Foil
Boron Black	Lockheed Martin Denver, CO	For many kinds of metals	Plasma spray deposition process
Boron Carbide	Lockheed Martin Denver, CO	For Ti substrates only	Proprietary process
Silicon Carbide	Lockheed Martin Denver, CO and other companies	For many kinds of metals	Chemical vapor deposition
Textured Surfaces	−NASA Ames Research Center Moffat Field, CA −SPIRE Corporation Bedford, MA −Optics MODIL Oak Ridge National Laboratory	For many kinds of metal substrates	Either: −sputter coated −ion beam etched −sputter coated then etched

Note that finish names and manufacturers change regularly. Table adapted from SOLEXIS™ database, Stellar Optics Research International Corporation (SORIC), 78 Normark Dr., Thornhill, Ontario, Canada L3T 3R1.

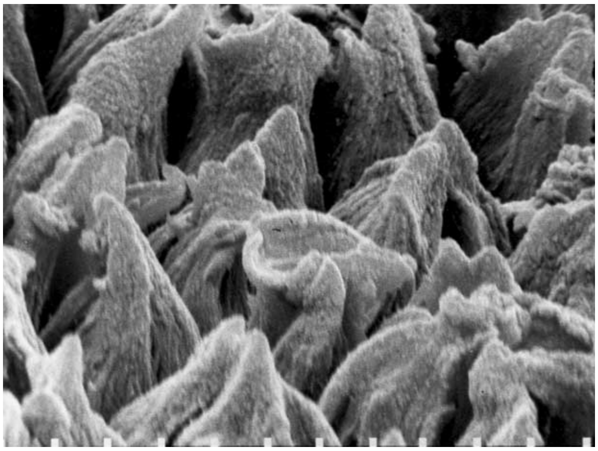

Figure 8 Scanning electron micrograph of Martin Optical Black surface on 6061T6 aluminum. This surface is a good absorber at ultraviolet, visible, and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

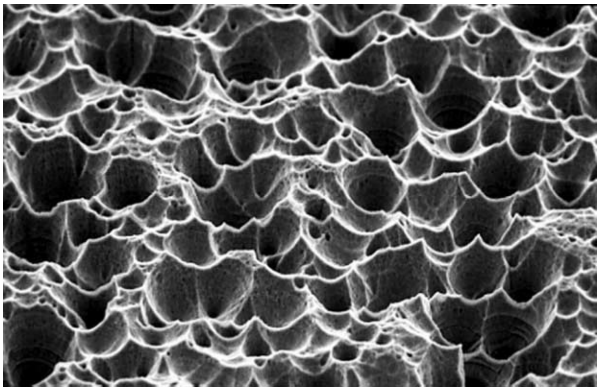

Figure 9 Scanning electron micrograph of phosphoric acid etched electroless nickel surface. This surface is a good absorber in the visible wavelength region. (Courtesy of Stephen Pompea and Donald Shepard.)

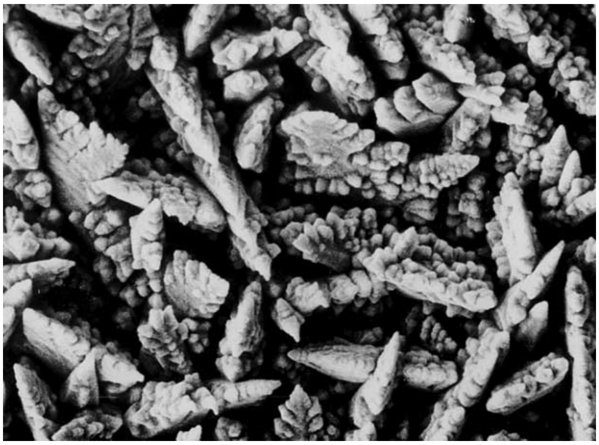

Figure 11 Scanning electron micrograph of tungsten hexafluoride surface. This surface is a good absorber at visible and near infrared wavelengths. (Courtesy of Stephen Pompea and Donald Shepard.)

in the program. Although better black surfaces are constantly being developed, the most important future development in the area of black surfaces may be a wider realization of the importance of an early and comprehensive surface selection process. The value of starting the selection process early using black surface databases has been confirmed repeatedly by optical designers.

See also

Instrumentation: Photometry; Telescopes. **Scattering:** Scattering from Surfaces and Thin Films.

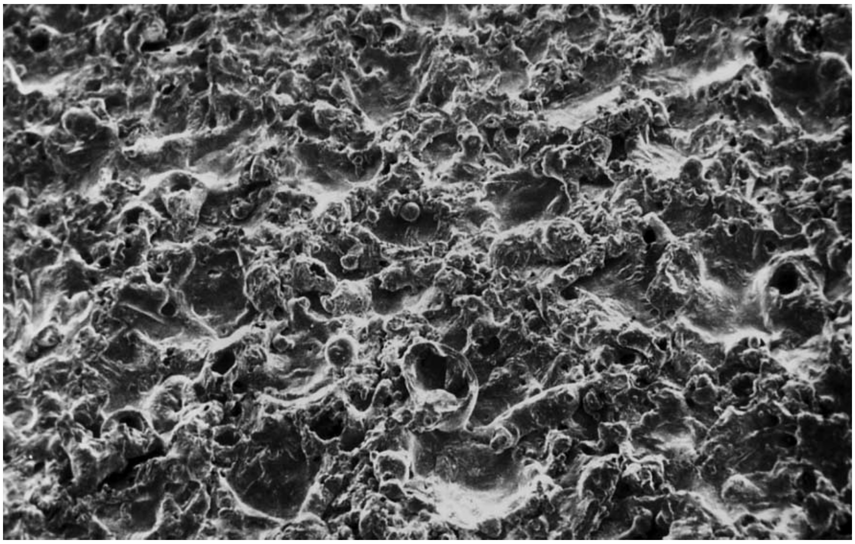

Figure 10 Scanning electron micrograph of Aeroglaze Z-306 painted surface with embedded microspheres. This surface is used primarily at visible wavelength. (Courtesy of Stephen Pompea and Donald Shepard.)

Further Reading

American Society for Testing Materials (ASTM) (1990) ASTM E1392-90. *Standard Practice for Angle Resolved Optical Scatter Measurements on Specular or Diffuse Surfaces*. West Conshohocken, PA: ASTM.

Bennett JM and Mattsson L (1989) *Introduction to Surface Roughness and Scattering*. Washington, DC: Optical Society of America.

Bohren CF and Huffman DF (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley and Sons.

Breault RP (1995) Control of stray light. In: *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Church EL and Takacs PZ (1995) Surface scattering. In: Bass M (chief ed.) *Handbook of Optics*, 2nd edn. New York: Optical Society of America.

Granqvist CG (1989). *Spectrally Selective Surfaces for Heating and Cooling Applications*. Volume TT 1. Bellingham, WA: SPIE Optical Engineering Text.

Gunderson JA (1977) *Goniometric Reflection Scattering Measurements and Techniques at 10.6 Micrometers*. Thesis, University of Arizona.

Henninger JH (1984) Solar absorptance and thermal emittance of some common spacecraft thermal control coatings. *NASA Reference Publication 1121*.

Jelinsky P and Jelinsky S (1987) Low reflectance EUV materials: a comparative study. *Applied Optics* 26(4): 613–615.

Persky MJ (1999) Review of black surfaces for space-borne infrared systems. *Rev. Sci. Instruments* 70(5): 2193–2217.

McCall SHCP, Pompea SM, Breault RP and Regens NL (1992) Reviews of properties of black surfaces for ground and space-based optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation II*, vol. 1753.

Pompea SM and Breault RP (1995) Optical black surfaces. In: *Handbook of Optics*, 2nd edn. New York: Optical society of America.

Pompea SM, McCall SHCP (1992) Outline of selection process for black buffle surfaces in optical systems. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Radiation in Optical Systems* II, vol. 1753.

Pompea SM, Shepard DF and Anderson S (1988) BRDF measurements at 6328 Angstroms and 10.6 micrometers of optical black surfaces for space telescopes. In: Breault RP (ed.) *Proceedings of the SPIE: Stray Light and Contamination in Optical Systems*, vol. 967, pp. 236–248.

Smith SM (1980) Far infrared (FIR) optical black bidirectional reflectance distribution function (BRDF). *Proceedings of the SPIE: Radiation Scattering in Optical Systems* 362: 161–168.

SOLEXIS™ Database of BRDF Measurements (2003) Stellar Optics Research International Corporation (SORIC), Ontario, Canada.

Stover J (1995) *Optical Scattering: Measurement and Analysis*, 2nd edn. Bellingham, WA: SPIE Optical Engineering Press.

Wolfe WL (1978) Optical materials. In: Wolfe WL and Zissis GJ (eds) *The Infrared Handbook*. Ann Arbor, Michigan: ERIM.

Thin Film Optical Coatings

D Ristau, Laser Zentrum Hannover e.V., Hannover, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Thin-film optical coatings are employed to adapt the properties of an optical surface to its specific application. Commonly, optical coatings comprise a sequence of single layers of different transparent materials with a thickness in the range of the operation wavelength. Depending on this coating design, the spectral characteristics of an optical surface can be tuned to nearly any required transfer function including a suppression of reflection for a broad spectral range by an antireflective coating or even the achievement of highest reflectance values up to 99.999% using a high reflecting layer stack. Besides these basic functions, optical coatings can be employed to realize a variety of transfer spectra specified for multiple wavelengths or wavelength ranges to manipulate radiation according to even extremely sophisticated demands in optical systems for laser technology, communication, or fundamental research. Moreover, the desired optical properties can be combined with additional specifications, for example, concerning the mechanical or chemical stability of an optical surface by selecting adapted coating materials. As a consequence of this enormous versatility, thin-film optical coatings can be found in nearly every product and development of modern optical technology today. Aside from the large market of consumer optics including ophthalmology, camera objectives, binoculars, and optical systems for data storage, progress in optical coating technology is predominantly driven by demanding applications in many future technology fields. As one prominent pace-setter, semiconductor lithography imposes new demands on optical coatings extending the operation wavelength of stepper systems in chip production to the extreme ultraviolet (EUV)-spectral range. Other examples can be found in communication technology, where thin-film filters with a bandwidth below

0.5 nm are required, or also in laser material processing, where coatings with high power handling capability and lifetime are indispensable for the development of cost-efficient laser tools.

In this article a concentrated review will be given on the design, production, and on specific properties of optical coatings. After an introduction to the fundamental principles of operation of coating systems, basic expressions for the optical transfer function of a single layer will be derived and extended to layer structures on the basis of the matrix method. Selected conventional and ion processes for the production of coatings will be discussed with respect to the achievable optical and structural properties of the coatings. Also, the thickness control of the coatings during production will be considered introducing the crystal monitoring technique as well as optical measurement strategies at a single wavelength or in a broad spectral range. The article will conclude with a review of the present state of optical coatings and an analysis of future trends.

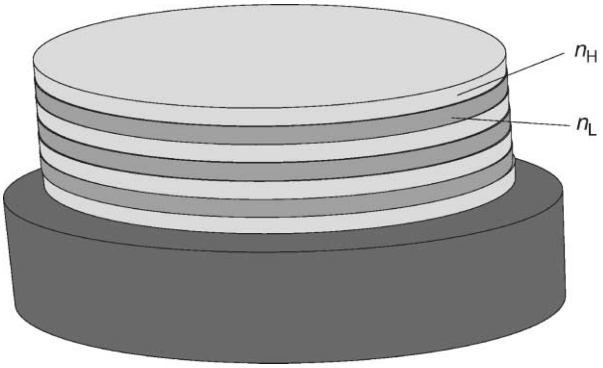

Figure 1 Basic structure of a thin-film system. Transparent layers of at least two different materials (n_H, n_L) are deposited on a substrate with refractive index n_T.

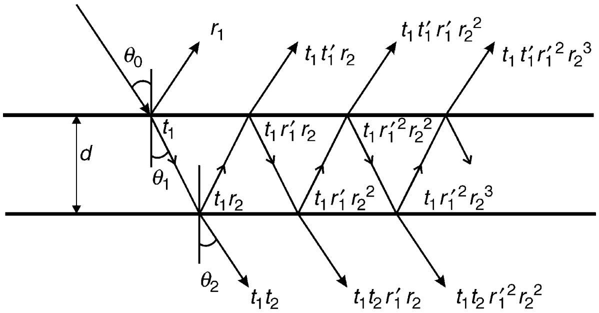

Figure 2 Schematic path of a plane wave incident onto a single layer with refractive index n_1. The reflection and transmission coefficients for the amplitudes at the interfaces $m = 1, 2$ are denoted by t_m and r_m, respectively. A prime added to the coefficients indicates a transfer of the wave opposite to the direction of incidence.

Design of Optical Coatings

The colorful appearance of optical coatings and microstructures, which can be observed on some butterflies and other insects, was present in nature long before mankind invented them. The beginning of the technical evolution of optical surfaces may be dated back to Greek culture, where polished metal surfaces were used as mirrors. Considering the application of transparent coatings, the investigations of Joseph Fraunhofer in the reflectivity of aged glass surfaces during the year 1819 should be mentioned as a first approach towards an antireflective coating. On the basis of controlled etching, this technique was further developed for different glass materials by Dennis Taylor in 1891 to realize the first antireflective coatings on lenses for applications in optical devices. In 1899, the Fabry–Perot theory was developed as a milestone in the theoretical description of multilayer structures. This theory describes the interference of partial waves reflected from two parallel optical surfaces, which can be considered as the fundamental arrangement of a multilayer structure (see Figure 1). In an optical thin-film system, layers of different transparent materials are composed on a stack deposited on the surface of the optical component (substrate refractive index n_T). In general, to adjust the spectral characteristics of the layer system to the demanded specifications, at least two layer materials with a high refractive index n_H and a low index n_L have to be selected. The choice of the thin-film materials is dependent on the application wavelength and the desired stability of the coating system as well as on their optical properties essential for the specific design and the targeted optical quality. Similar to the Fabry–Perot principle, the optical function of the layer structure is based on the interference of partial waves reflected and transmitted at the interfaces between the layers. The trace of the incoming wave and the induced partial waves are depicted in Figure 2 for a single layer, which can be considered as the basic unit for coating systems even with highest complexity. For a theoretical description of the transfer function of this single layer, the contributions of the individual partial waves have to be calculated and accumulated to form the total transmitted and reflected wave. In this classic approach, the incoming wave can be described by the function of a plane wave with amplitude E_0, wavenumber $k = 2\pi/\lambda$ incorporating the wavelength λ, and frequency ω:

$$E(x,t) = E_0 \exp[-\mathrm{i}(kx - \omega t)] \qquad [1]$$

At the first interface, a part of this plane wave is directly reflected with an amplitude of $A_0 = E_0 r_1$

(order zero), which can be calculated using the reflection coefficient r_1 given by the Fresnel formulas. After passing the first interface, a reflection and transmission of the incoming wave at the second interface has to be considered. Also a phase shift has to be taken into account, which is introduced by the path of the wave from the front to the rear interface of the single layer. This phase shift δ_i is dependent on the thickness d_i of the layer and the refractive index n_i:

$$\delta_i = \frac{2\pi n_i d_i \cos\theta_i}{\lambda} \qquad [2]$$

For oblique incidence also the angle of incidence θ_i in the layer, which has to be calculated by Snell's law, has to be taken into account. The index i has been chosen in this consideration for convenience to account for the layer number in coating systems, which will be discussed later. Following the wave reflected the first time inside the layer, a further phase shift for the path from the rear side to the front side has to be taken into account. This partial wave is then transmitted by the front surface resulting in an amplitude A_1 of the first-order wave leaving the layer:

$$A_1 = t_1 t'_1 r_2 E_0 \exp(-i2\delta) \qquad [3]$$

where t'_1 is the transmission coefficient for the amplitude of the wave passing the layer at the interface to the ambient. By following the wave further through the layer, the amplitudes A_k of each partial wave of order k can be determined:

$$A_k = t_1 t'_1 r_2 E_0 \exp(-i2\delta) * (r'_1 r_2 \exp(-i2\delta))^{k-1} \qquad [4]$$

Obviously the contributions of the partial waves follow the rule of a geometric expansion, and therefore the total amplitude of the wave reflected by the layer, which is the sum of the amplitudes of the partial waves, can be expressed in closed form:

$$A_{\text{tot}} = \left[r_1 + \frac{t_1 t'_1 r_2 \exp(-i2\delta)}{1 - r'_1 r_2 \exp(-i2\delta)} \right] E_0 \qquad [5]$$

The reflection coefficient r_S of the single layer is given by the ratio between the total amplitude of the reflected wave and the amplitude of the incoming wave:

$$r_S = \frac{r_1 + r_2 \exp(-i2\delta)}{1 + r_1 r_2 \exp(-i2\delta)} \qquad [6]$$

In this equation, the coefficients r'_1 and t'_1 have been replaced by using the relations

$$t_1 t'_1 = (1 - r'_1)(1 - r_1) = (1 - r_1)(1 + r_1) = 1 - r_1^2 \qquad [7]$$

In principle, this method of compiling the contributions of the partial waves reflected at the layer interfaces can be extended also to the calculation of systems with more than one layer. But, considering the enormously increasing number of partial beams, the resulting equations become extremely complicated for multilayer structures. Therefore, the matrix formalism, which can be deduced from the boundary conditions of the electric and magnetic field at the interfaces, is usually employed for the calculation of thin-film systems. The major advantage of this approach, which was developed in the 1940s, is the representation of each layer by a single matrix M_i containing all specific parameters of the film. The matrix M_i of a thin film at position i within the layer system relates the electric (E_{i-1}) and magnetic (H_{i-1}) field strength at the front face to the field strength values (E_i and H_i) at the rear face of the layer:

$$\begin{pmatrix} E_{i-1} \\ H_{i-1} \end{pmatrix} = \mathbf{M}_i \begin{pmatrix} E_i \\ H_i \end{pmatrix} = \begin{pmatrix} \cos\delta_i & \frac{i}{n_i}\sin\delta_i \\ i n_i \sin\delta_i & \cos\delta_i \end{pmatrix} \begin{pmatrix} E_i \\ H_i \end{pmatrix} \qquad [8]$$

By considering the boundary conditions for the field strengths at the interfaces, the matrix components given in eqn [8] can be determined. On the basis of the matrix formalism, the transfer function of a layer system can be simply accomplished by multiplication of the matrices M_i representing the constituent single layers. Thus, the transfer matrix M_S of a stack, which is formed by a number K of single layers with layer 1 located at the first interface with respect to the incoming wave, is given by

$$\mathbf{M}_S = \mathbf{M}_1 \mathbf{M}_2 \mathbf{M}_3 \ldots \mathbf{M}_i \mathbf{M}_{i+1} \ldots \mathbf{M}_K \qquad [9]$$

For calculating the reflection coefficient r_{SK} of the entire arrangement including the substrate (index of refraction n_T) and the ambient medium (index of refraction n_0), again the ratios of the amplitudes have to considered:

$$r_{SK} = \frac{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} - i\mathbf{M}_{21} - n_T \mathbf{M}_{22}}{n_0 \mathbf{M}_{11} + i n_0 n_T \mathbf{M}_{12} + i\mathbf{M}_{21} + n_T \mathbf{M}_{22}} \qquad [10]$$

As a consequence of its perceptible structure, the matrix method can be readily transferred into computer codes, which cover the calculation of all important properties of even extremely complicated layer systems. In particular, considering the fundamental relations of the matrix method, also the field strength distribution within the layer structure can be determined. In software packages, which are commercially available as standard tools for optical thin-film technology, often the film thickness D_i of the layers is expressed in the units of QWOT (quarter wave optical thickness) at the design wavelength λ_Z:

$$D_i = \frac{4n_i d_i}{\lambda_Z} \qquad [11]$$

In the case of normal incidence, the matrices M_i reduce to simple expressions for layers of integer QWOT values, because the corresponding phase shift terms are multiples of $\pi/2$. Clearly, an integer QWOT value of optical thickness also represents an extreme value in the spectral behavior at the design wavelength. This fact is often exploited for a precise thickness control by optically monitoring the reflectance or transmittance of the growing layer at the design wavelength λ_Z: by recording the transfer behavior as a function of time during production, extreme values and accordingly thickness values of multiple QWOT can be accurately determined. Thus, if a layer structure is composed only of multiple QWOT layers, switching between the layer materials can always be precisely performed at a minimum or maximum read-out of the optical monitor tuned to the design wavelength. In respect of this advantage in production, most coating designs are based on QWOT layers indicating the practical importance of the unit QWOT. For an efficient description of the designs, often a notation in capital letters is used, which indicate a 1 QWOT layer of a certain coating material. For example an alternating QWOT-stack with 11 layers of a high-index material (H, refractive index n_H) and a low-index material (L, refractive index n_L) is described by the sequence HLHLHLHLHLH or even more condensed by $(HL)^5H$. The QWOT stack is the fundamental design for dielectric high reflecting mirrors and represents the basis for a variety of functional coatings in optical thin-film technology. By applying the matrix method, an approximate expression for the reflectance R_S at the design wavelength can be derived for such a stack:

$$R_S \approx 1 - 4n_0 n_T \frac{n_L^{2N}}{n_H^{2(N+1)}} \qquad [12]$$

This expression is only valid for a large number N of QWOT layers and underlines the role of the materials for the practical production of optical coatings. Highest reflectivity with lowest number of layers can be achieved only by selecting materials with a high contrast of their refractive indices. This contrast also influences the spectral broadness of the high-reflecting region of the stack (Figure 3).

Besides highly reflecting mirrors, antireflective coatings are another important class of functional coatings for modern optics and laser technology. Antireflective coatings often involve non-QWOT layer structures of more than two materials in order to accomplish the demands of lowest residual reflectivity for a broad spectral range. Eye glasses with antireflective coatings covering the entire visible spectral range can be considered as a prominent example for consumer optics, which impose requirements on the coatings also in respect to their stability against cleaning and other environmental influences during normal use. In this context, the introduction of optical plastic materials is an additional challenge in consumer optics and ophthalmology, which has to be mastered by adapted designs and production processes. In contrast to the typical demands of consumer optics, antireflective coatings on laser windows are

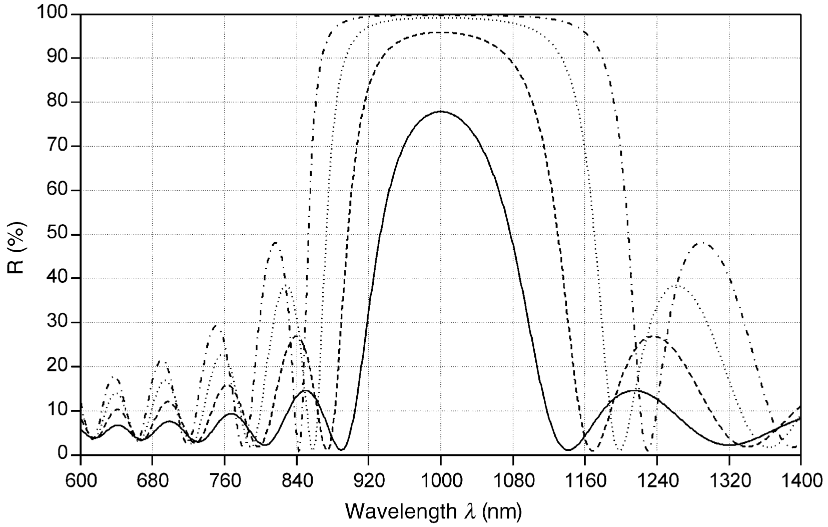

Figure 3 Spectra calculated for high-reflecting QWOT stacks (15 layers, substrate index $n_T = 1.52$) composed of a low-index material ($n_L = 1.46$) and different materials with high indices of refraction varying between $n_H = 1.7$ (solid), 1.9 (dashed), 2.1 (dotted) and 2.3 (dash-dotted). The spectral broadness and the reflectivity increase with the ratio n_H/n_L.

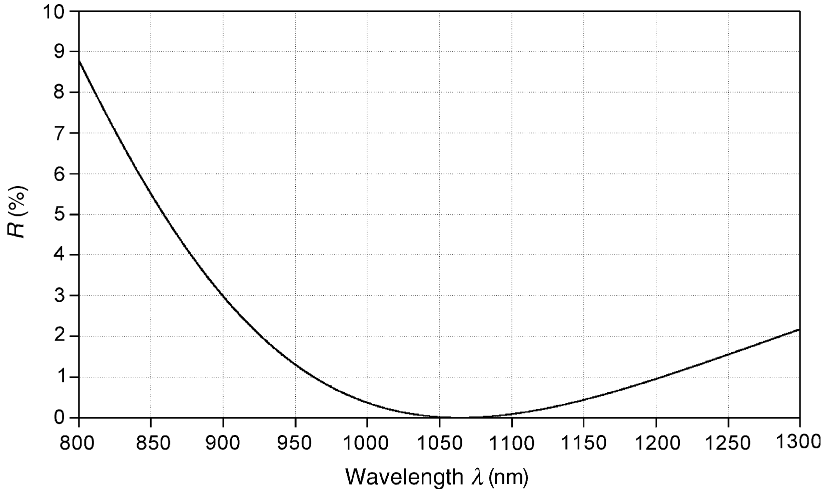

Figure 4 Spectrum of an antireflective coating for the wavelength 1.064 μm, which is designed as a V-coating with two layers of Ta_2O_5 ($n_H = 2.03$; $d_H = 0.383$ QWOT) and SiO_2 ($n_L = 1.46$; $d_L = 1.308$ QWOT) for a quartz substrate (substrate index $n_T = 1.46$).

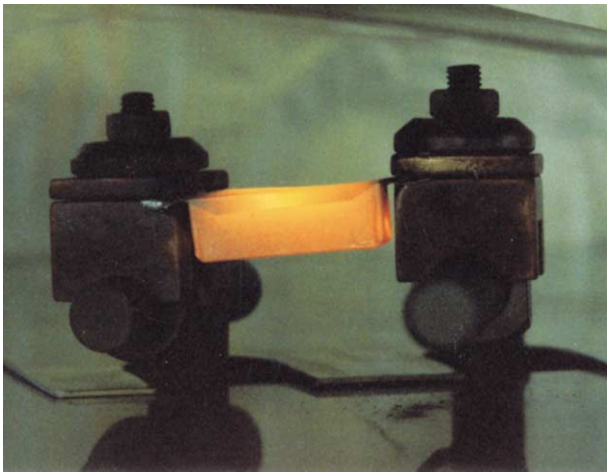

Figure 5 Boat evaporation source in operation. The deposition material is put into the boat which is made of a metal with high melting point. A high electric current is conducted through the boat for heating to temperatures up to 2000 °C.

often restricted to one wavelength, but have to fulfill extreme requirements with respect to optical losses and power handling capability. Therefore, most antireflective coatings in laser technology are composed only of two thin layers to minimize the total thickness, which is often a direct measure of the expected losses of the layer system. With such a coating, which is often called V-coating according to its spectral characteristic (Figure 4), the reflectivity of the window surface can be suppressed ideally for one wavelength.

Production of Optical Coatings

The industrial production of optical coatings started in the 1940s on the basis of glass bell jar systems equipped with diffusion pumps and boat evaporators for the deposition of single layers of metals for mirrors and fluoride materials for antireflective coatings. Metal containers are used as boat evaporation sources. These containers are filled with the deposition material and heated by electric currents up to several 100 A (Figure 5). In the course of the rapid improvement of vacuum systems and especially the development of the electron beam evaporator over the next three decades, new materials with improved optical and mechanical quality could be combined to form more complicated dielectric coating systems. The layout of an e-beam deposition plant of these times, which can still be considered as fundamental for the advanced deposition systems of today, is illustrated in Figure 6. In such a box coater, the

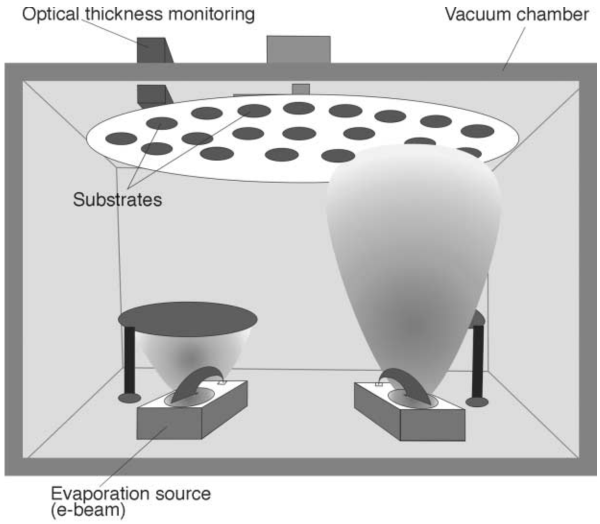

Figure 6 Outline of a modern box coater used for the production of optical coatings with e-beam deposition processes.

substrates are installed in a calotte which is rotated to achieve a uniform distribution of the layer thickness. The deposition materials are put in water-cooled crucibles and are subjected to an electron beam operated at currents in the range between 0.1 A to more than 10 A and voltage levels of several kilovolts. As a consequence of the interaction with the energetic electrons, the deposition material can be heated to high temperatures above 5000 K allowing the evaporation of most of the desired deposition materials. These achievable high temperatures in conjunction with the good controllability are the major advantages of e-beam sources (Figure 7) in comparison with the classical boat evaporation process. In particular, e-beam evaporation opened the way for the deposition of oxide materials and other optical materials with high evaporation temperatures, which exhibit optical properties as well as a high mechanical and chemical stability superior to the material classes accessible with boat evaporation. However, the cost efficient boat evaporation technique is still applied today for the deposition of metals or temperature-sensitive materials for coating systems in the vacuum ultraviolet (VUV)/UV and mid-infrared (MIR) spectral ranges. If the shutter above the deposition source is opened, the evaporated material reaches the substrates and forms the layer by condensation. This condensation process is sensitively dependent on the environmental conditions of the substrates and the deposition sources. Sufficient microstructural properties of the growing layer can generally be achieved only under fine vacuum below 100 Pa and heating of the substrates to temperatures in the range of some 100 °C. Even under well-optimized conditions, the typical microstructure of such thermally evaporated coatings exhibits specific deficiencies (Figure 8) which influence the optical and mechanical properties of the layer systems. Often, microcrystallites with diameters in the range of a few nm are observed, which are arranged in columns of a typical size growing from the substrate interface to the outer surface of the layer. For a simulation of the growing mechanisms for this low energetic deposition process, uncomplicated models with a few input parameters already lead to satisfactory results. Assuming a kinetic energy of 0.1–0.3 eV, which is typical for thermal evaporation according to the Boltzmann distribution, the impinging atoms can move only a small distance or even adhere directly at their position of incidence on the growing layer. As a consequence of this small mobility, the atoms are not able to reach their energetically optimum positions in the layer, and they even obstruct the path of the following atoms by shadowing. On the basis of these assumptions, computer simulations of the growing process result in structures which are comparable to the columnar growth observed in practice. The packing density of such a columnar structure is significantly below the bulk values of the corresponding deposition materials, and it contains also a high number of defects and voids, which reduce the optical quality of the coating. In particular, the coating tends to

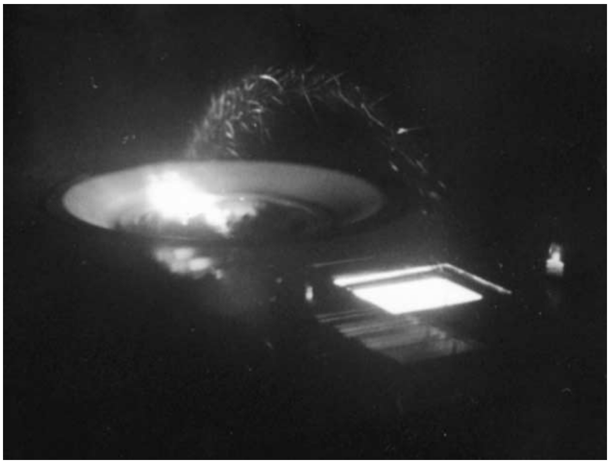

Figure 7 Electron beam source in operation. The deposition material is put in a water-cooled crucible and subjected to an e-beam. Since the zone heated by the e-beam is entirely embedded in the surrounding deposition material, very low contamination levels and temperatures above 5000 °C can be achieved. For the generation of the e-beam an arrangement similar to the system of a cathode ray tube is employed: the electrons emitted by a hot filament are focused by an aperture and accelerated by a high voltage.

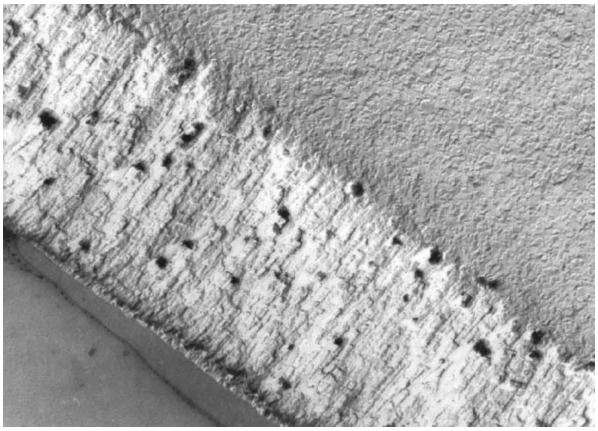

Figure 8 Microstructure of a MgF_2 single layer (geometric thickness approximately 780 nm) deposited by conventional boat evaporation. For the preparation of the TEM picture, the coating has been fractured and replicated.

adsorb water from the normal atmospheric environment into its internal structure leading to a shift of the spectral characteristics as a function of the ambient conditions. Besides this loss of spectral stability, the adsorbed water also induces an increased optical absorption of the coatings in the VUV and the MIR spectral ranges.

The increasing demand in laser technology and other optical applications for coatings far exceeds the quality level of conventionally deposited coatings and this was the driving force for the development of advanced concepts for the production of optical coatings during the last three decades. Considering the fundamental growth mechanism of conventional optical coatings, two basic approaches for an improvement of the microstructure are available. On one hand, the mobility of the atoms on the surface of the growing layer can be increased by the insertion of additional energy from other sources. This concept leads to the development of ion-assisted deposition processes, where the additional energy is transferred to the adatoms by collisions with ions generated directly in the plant by an ion source at energy levels of approximately 30–300 eV (Figure 9). Especially, by the deposition of oxide materials, which can be assisted directly with oxygen ions for balancing their stoichiometry, coating systems with high environmental stability and low optical losses can be achieved for the VIS up to the MIR spectral range. On the other hand, the energy of the adatoms can be directly increased in sputter processes which deliver deposition energies orders of magnitude higher than the thermal energy range. Besides adapted activated reactive magnetron sputter processes, particularly ion beam sputtering (IBS) was developed for a coating process delivering high-end coatings mainly for laser technology with an unsurpassed optical quality (Figure 10). For example, mirrors with total losses below one ppm can be produced for the Nd:YAG laser (wavelength 1.064 μm) with adapted IBS processes.

In addition to stable deposition processes, precise control of the layer thickness with an accuracy below 1 nm is a further important step towards the realization of complex optical coatings. Most deposition plants are equipped with quartz crystal monitors, which measure the optical thickness on the basis of the resonance frequency shift of an oscillating quartz crystal induced by the increase in mass of the deposited coating. The relative accuracy of these monitors ranges from 2% to 10% and is well suited for the regulation of the deposition rate. In order to achieve the required precision for high-quality coatings, optical monitors are often employed, which record the transmittance of the growing layer system at a single wavelength or even in a broad spectral range. Since these systems account directly for the optical properties of the deposited layer system, an efficient and accurate control of the film thickness is possible allowing the production of extremely sensitive filters with bandwidths below 1 nm for telecommunications or mirrors with defined group delay dispersion for fs-laser systems (Figure 11).

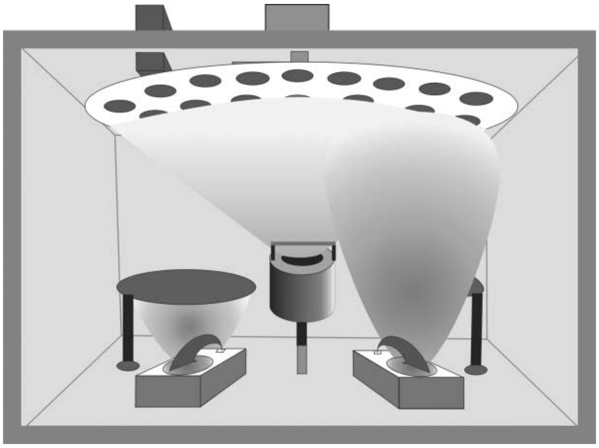

Figure 9 Ion assisted deposition. An additional ion source is installed, which delivers inert or reactive ions impinging on the growing layer. The energy transfer from the ions to the adatoms results in improved microstructure of the produced coatings with higher packing density.

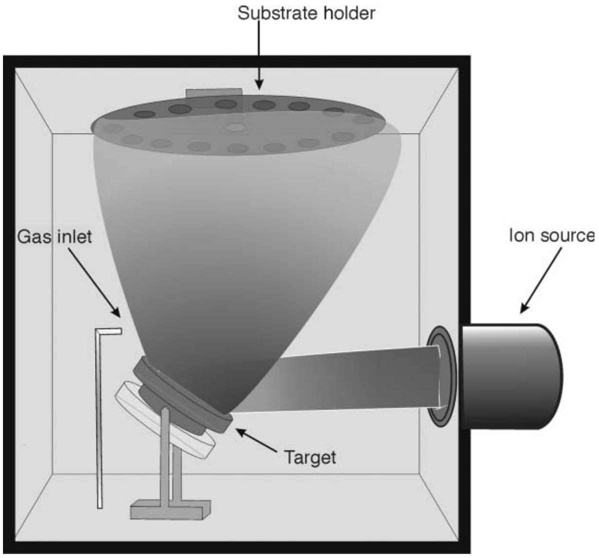

Figure 10 Principle of the ion beam sputtering (IBS) process. Inert ions with an energy in the range 1–2 KeV are generated in an ion source and directed onto the target which consists of the desired deposition material. Material is sputtered from the target and reaches the substrates. As a consequence of the high energy of the adatoms (5–30 eV) the formed layer exhibits a superior microstructure.

Quality Parameters of Optical Coatings

Besides the spectral characteristics, a variety of specific parameters has to be considered for each application of optical coating systems (Table 1). For laser systems the laser-induced damage threshold and the optical losses have to be evaluated in the context of technical and also economic requirements. Absorption is a loss channel, which transforms a fraction of the impinging radiation energy into heat, inducing a temperature increase in the coating. This generated temperature profile influences the imaging properties of the optical component via thermal expansion and temperature-induced alterations of the refractive indices of the coating as well as the substrate materials. Despite numerous technical problems, absorption always implies the loss of laser energy which is produced with high economic expense. Absorption losses are related to defects and impurities, which are introduced during the production process into the layer by a variety of contamination mechanisms. Furthermore, an unbalanced stoichiometry with a higher metal content is often observed as a cause for increased absorption in dielectric optical coatings, which are mainly composed of binary metal compounds. Absorptance values of optical components are determined by laser calorimetric techniques described in ISO 11551. In this measurement scheme, the temperature increase in the sample is monitored during irradiation with a well-characterized laser beam. After this heating phase, the laser beam is blocked and the temperature curve during the cool-down cycle is also recorded. On the basis of the measured temperature development, the absorptance can be calculated taking into account the laser power and the heat capacity of the specimen.

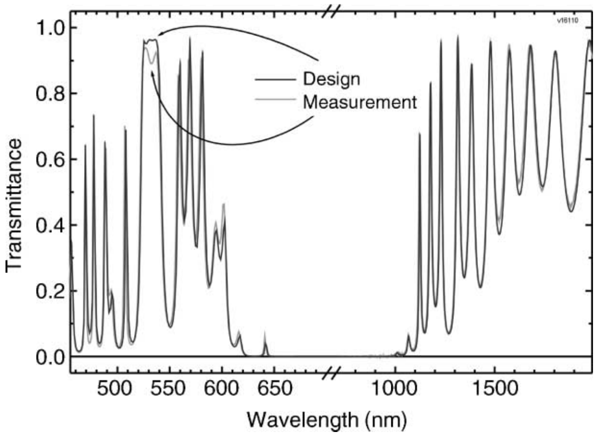

Figure 11 Example of a complicated layer system deposited under the control of an advanced optical broadband monitoring device: double chirped mirror (56 non-QWOT layers of SiO_2/TiO_2). Only small deviations between the calculated spectrum and the spectrum of the produced coating can be observed.

All mechanisms, which deflect radiation from its specular direction, contribute to optical scattering representing the second major loss channel in optical coatings. Since scatter losses are mainly attributed to microstructural imperfections and the surface roughness of the layers, a scaling with wavelength can be deduced from the corresponding theoretical models. Especially in the VUV/UV spectral range, high scatter values are apparent and may dominate the losses of the optical components. For longer wavelengths, the contribution of scattering is decreased to a point in the MIR spectral range where it can be neglected in comparison to the absorption losses. Scatter effects

Table 1 Selected quality parameters of optical coatings and surfaces in conjunction with corresponding ISO standards and measurement principles

Specification	*Parameter/Unit*		*Standard/Measurement principle*
Laser induced damage threshold (LIDT)	cw-LIDT	W/cm	ISO 11254-1: cw-laser irradiation
	1 on 1-LIDT	J/cm^2	ISO 11254-1: irradiation with single pulses
	S on 1-LIDT	J/cm^2	ISO 11254-2: repetitive irradiation with pulses
	Assurance	J/cm^2	DIS 11254-3: irradiation sequence
Optical losses	Absorptance	ppm	ISO 11551: laser calorimetry
	Total scattering	ppm	ISO 13696: integration of scattered radiation
Transfer function	Reflectance	%	CD 13697: precise laser ratiometric method
	Transmittance	%	DIS 15368: spectrophotometry
Surface quality	Scratch/Digs		ISO 10110: 13 parts containing different types of imperfections
	Roughness		
Stability	Abrasion		ISO 9211: different test methods
	Environmental Stability		ISO 9022: 21 parts containing a variety of conditioning methods

may impair the imaging properties or the resolution, and they reduce the efficiency of an optical system. Besides the economic aspect of each loss effect, scattering can also be considered as relevant for safety in high-power laser systems, where some 100 ppm of scattered radiation carry enough power to injure or blind the operator. In the ISO 13696 standard, measurement procedures based on the collection or integration of the total radiation scattered by the sample are described.

Depending on their application, coatings have to withstand also manifold mechanical, chemical, and environmental influences. Severe mechanical and chemical requirements are imposed onto coatings for ophthalmology or consumer optics, which are often cleaned with abrasive and aggressive cleaning solvents and tissues. To qualify coatings for these applications, various testing procedures have been developed and are routinely employed in industrial quality management. Coatings for laser systems and precision optics are predominantly certified according to conditioning methods described in the ISO standard series 9211. Test procedures with respect to environmental influences are extensively outlined in the more than 20 parts of ISO 9022, which is dedicated to a variety of climatic, mechanical, and chemical conditioning cycles.

Summary and Outlook

Optical coatings are the major tool for engineering the properties of functional optical surfaces for an extensive spectrum of applications. On the basis of the matrix algorithm, a comprehensive theoretical model for the spectral transfer characteristics of optical coatings is available, which has been transferred to commercial computer programs. Now, most design problems can be solved with these computer codes, whereas the reproducible and precise production of optical coatings with high quality could only be partially achieved until today (Table 2). Besides conventional thermal processes, new ion-assisted deposition processes and sputtering concepts are gaining importance in optical thin-film technology. For the quality management of optical coatings, a variety of often standardized testing procedures has been developed during the last two decades including the determination of the power handling capability, the measurement of the optical losses and a broad spectrum of conditioning methods for optical coatings.

Demanding future challenges will be imposed on optical coatings by the development of semiconductor lithography in the EUV spectral range to achieve nodes well below 100 nm in chip production. Another important trend is directed by new laser applications in material processing, medicine and communication technology, operating at extremely high repetition rates and power levels, or with ultrashort pulses in the fs-regime. Other requirements will be defined by the rapid development of optical technology towards a broad range of special products with short innovation cycles obligating significant progresses in thin-film technology in the direction towards flexible and reproducible process concepts

Table 2 Selected quality parameters of optical coating systems for laser applications

Laser, wavelength	*Type*	*Absorption ISO 11551*	*Total scattering ISO 13696*	*Laser induced damage threshold, ISO 11254*
157 nm, F_2 excimer	HR/th		1–4%	
193 nm, ArF excimer	AR/th	0.7–2.5%	0.2–0.5%	1–2 J/cm^2 (1on1, 20 ns)
	HR/th	0.4–2.0%	0.2–2.5%	2–4 J/cm^2 (1on1, 20 ns)
248 nm, KrF excimer	AR/th		<0.025%	10 J/cm^2 (1on1, 30 ns)
	HR/th	<500 ppm	<0.2%	>20 J/cm^2 (1on1, 30 ns)
	HR/IB		<0.1%	>3 J/cm^2 (1on1, 30 ns)
633 nm, HeNe laser	HR/th	<30 ppm	<30 ppm	–
	HR/IB	<5 ppm	<5 ppm	–
1.064 μm, Nd:YAG	AR/th	<20 ppm	<100 ppm	>60 J/cm^2 (12 ns, 0.25 mm)
	HR/th	<50 ppm	<100 ppm	>100 J/cm^2 (12 ns, 0.25 mm)
	HR/IB	<1 ppm	<1 ppm	>80 J/cm^2 (12 ns, 0.25 mm)
10.6 μm, CO_2 laser	AR/th	<0.16%	–	>20 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm) >3 kW/mm (cw, 100 μm)
	HR/th	<0.10%	–	>25 J/cm^2 (100 ns, 1.4 mm) >2 kJ/cm^2 (1.2 ms, 250 μm)

Types: HR: highly reflecting mirror; AR: antireflective coating; th: thermal evaporation; IB: ion beam sputtering.

for coatings with increasing complexity. In all fields of modern optics, optical thin films will represent an enabling technology which will play a key role for many future applications and products.

List of Units and Nomenclature

A_0	amplitude of partial wave of order zero
A_1	amplitude of first-order partial wave
A_k	amplitude of the partial wave of order k
d_i	thickness of layer number i
D_i	film thickness expressed in units of QWOT
E_0	amplitude of a plane wave
E_i, H_i	electric field strength and magnetic field strength at the rear front of layer number i
IAD	ion-assisted deposition
IBS	ion beam sputtering
k	wavenumber of a plane wave $k = 2\pi/\lambda$
M_i	matrix of layer number i
M_S	composite matrix of a layer stack of K single layers
n_0	index of refraction of the ambient medium
n_H	refractive index of a high-index material
n_i	refractive index of layer number i
n_L	refractive index of a low-index material
n_T	refractive index of the substrate
QWOT	quarter wave optical thickness: unit for the thickness of the layer
r_1, r'_1	reflection coefficients of the ambient–layer interface
r_2	reflection coefficient at the layer–substrate interface
r_S	reflection coefficient of a single layer
r_SK	reflection coefficient of a layer stack of K single layers
R_S	reflectance of a QWOT stack
t_1, t'_1	transmission coefficients of the ambient–layer interface
δ_i	phase shift of a plane wave in layer number i
λ	wavelength of a plane wave
λ_Z	design wavelength
θ_i	angle of incidence in layer number i
ω	frequency of a plane wave

See also

Optical Coatings: Laser Damage in Thin Film Coatings.

Further Reading

Guenther AH (ed.) (2002) *International Trends in Applied Optics*. Bellingham, Washington: SPIE Press.

ISO 11551 (1997) *Test Method for Absorptance of Optical Laser Components*. International Organization for Standardization.

ISO 13696 (2002) *Test Method for Radiation Scattered by Optical Components*. International Organization for Standardization.

ISO 9022 (1996, 1997) *Optics and Optical Instruments – Environmental Test Methods – Parts 1–21*. International Organization for Standardization.

ISO 9211 (1994–2001) *Optical Coatings, Parts 1–4*. International Organization for Standardization.

ISO/FDIS 15368 (2003) *Measurement of Reflectance of Plane Surfaces and Transmittance of Plane Parallel Elements*. International Organization for Standardization.

Jacobson MR (1989) *Deposition of Optical Coatings*. SPIE Milestone Series, Volume MS 6. Bellingham, Washington: SPIE Press.

Kaiser N and Pulker HK (2003) *Optical Interference Coatings*. Berlin: Springer.

Liddell HM (1980) *Computer-Aided Techniques for the Design of Multilayer Filters*. Bristol, UK: Adam Hilger.

Macleod HA (2001) *Thin Film Optical Filters*. Bristol, UK: Institute of Physics Publishing.

Rancourt JD (1996) *Optical Thin Films*. Bellingham, Washington: SPIE Press.

Thelen A (1988) *Design of Optical Interference Coatings*. New York: McGraw-Hill.

X-Ray Coatings

P Dhez, Université Paris Sud, Orsay, France

© 2005, Elsevier Ltd. All Rights Reserved.

In all spectral ranges, there exist two principal options for coating an optical element. This may be a simple layer, as for example a metallic layer is frequently used in the visible spectral region. The second option is to use a multilayer coating, making use of the optical properties of two or more materials to achieve enhanced reflectivity over a limited spectral range. In both cases, a knowledge of the refractive index of the materials (real or complex) is required, and of course one must choose materials that can be coated in a practical deposition process. A frequent side-advantage of coating is to make it possible to choose the material of the substrate for its mechanical or cooling properties without regard to the optical properties of the coatings.

Originally (at the beginning of the twentieth century) the progress of X-ray optics was hampered by a poor understanding of the nature of these rays, and in particular by the observation that they are very weakly reflected and refracted. After the electromagnetic nature of these rays had been confirmed, the problem of producing polished surfaces that were sufficiently smooth on an atomic scale to reflect X-rays still remained. Optical coatings gave little hope of improvement due to the growth of roughness during the coating process. Thus early attempts to make X-ray mirrors produced poor results and little progress was made until the 1960s. However, in the past 30 years, much technical progress has been made, and now many optical principles used in the visible spectral region have been extended to very short wavelengths. Even refractive and diffractive focusing lenses (not treated here) have been implemented in the X-ray region.

In the first section we outline the history of the discovery of the properties of X-rays, which were achieved mainly in experiments using crystals. This allows us to understand how their electromagnetic nature was discovered, and how they interact with materials. We then describe the phenomenon of the 'total reflection' of X-rays, which occurs only at very high angles of incidence ('grazing incidence') and its main applications in X-ray optics. Finally, we outline the principles of mirrors based on interference, their applications in the X-ray region, their advantages and their limitations.

X-rays Compared to the Other Electromagnetic Spectral Regions

Infrared radiation was discovered by W. Herschel in 1800, and ultraviolet rays by J. Ritter in 1801. These were the only 'invisible rays' known before Maxwell's contribution to the understanding of the electromagnetic spectra through the development of his equations in 1869. The discovery of radio waves by H. Hertz in 1887 confirmed that our eyes are blind to most of the electromagnetic spectrum.

In 1895, W. Roentgen discovered some previously unknown 'rays' that had the surprising property of being able to penetrate low-density material opaque to visible light. This discovery was a side-effect of his studies of low-pressure electrical discharges, fashionable at the time. The nature of these unknown or 'X'-rays as Roentgen named them, remained unknown for 12 years or so. Were they electromagnetic waves or particles? Roentgen used lenses, prisms and mirrors to try to detect a reflective, refractive or focusing effect, with inconclusive results. Although the electromagnetic nature of the X-rays was favored, the magnitude of the frequency or the refractive index was unknown. We note that natural radioactivity and gamma rays were discovered a year after X-rays, and the same questions concerning their nature were raised. However, in neither case did the ignorance of the nature of the radiation impede their use in medical radiography or in cancer therapy.

In 1906, the experiments of C. Barkla provided the first evidence of the electromagnetic nature of X-rays. In scattering X-rays from blocks of a light element (in this case, carbon), he was able to show that the scattered radiation was polarized, and therefore had the nature of a transverse wave. This kind of experiment had been performed almost a century previously to demonstrate the transverse wave nature of visible light. In 1916, following a suggestion by M. von Laue that X-rays might be diffracted from a grating of period comparable to their wavelength, Friedrich and Knipping succeeded in diffracting X-rays from the atomic planes of a natural crystal. This experiment established definitively the electromagnetic nature of X-rays and also the scale of their wavelength.

In 1915 W. L. Bragg summarized the diffraction rules for crystals through his equation [1], now known as the simplified Bragg law:

$$2d_i \sin\theta_i = N\lambda_i \qquad [1]$$

This equation relates the diffracted X-rays wavelength λ_i to a set of parallel atomic planes separated by a distance d_i and to the incidence angle θ_i. N is an integer indicating the order of diffraction, as in the grating equation. In using this equation, care is needed because of two peculiarities compared to the usual mirror or grating equations in the visible range. In the Bragg equation, the angle of incidence θ is measured from a given atomic plane to the incident rays. However, inside a crystal several different atomic planes exists, not only the planes parallel to the surface as indicated in Figure 1. In addition, θ is not measured from the normal as is usual in optics; hence the angle θ is referred to as the glancing or grazing angle of incidence.

We note that the Bragg equation is not restricted to X-rays, but is a coherent scattering phenomenon which is common to all cases where waves, such as

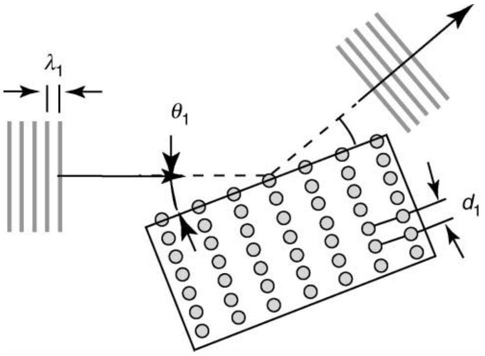

Figure 1 The geometry for the conditions leading to selective reflection on a periodic medium.

sound waves or ripples on water, are diffracted by regular arrays of scattering centers.

In order to visualize the phenomena, we refer to Figure 1 which is a schematic representation of a plane wave incident upon a two-dimensional array of scattering centers (electrons around nucleus, in the case of X-rays), shown here as dots. As the sine function has a maximum value of 1, the maximum wavelength λ that can be reflected from a spacing d is given by $\lambda = 2d_{\max}$. As the largest value of $2d$ that can be found in natural crystals is about 2 nm (in natural muscovite mica, for example), X-rays of 2 nm, with an equivalent energy of about 620 eV, are the 'softest' (least energetic) that can be reflected in this manner.

The model shown in Figure 1, similar to that proposed by M. von Laue, does not take into account the detailed interaction of the atoms with the X-rays, and therefore ignores effects such as absorption and refraction, and discrete atomic transitions. A.H. Compton attempted the first treatment of the interaction of matter with X-rays from an atomic viewpoint. In 1917, he developed a classical theory in which the electrons bound to the nucleus are treated as harmonic oscillators. In this way, he was able to derive an expression for the atomic scattering coefficients and to describe resonance effects at frequencies corresponding to specific atomic transitions. A few years later, he was able to establish a link between these 'atomic scattering factors' and the macroscopic concept of refractive index. This new formulation allowed the use of the Fresnel equations, widely used in other spectral regions, to calculate the intensity of X-rays reflected from smooth, polished surfaces. These surfaces may be amorphous as well as crystalline.

The refractive index can be written in several ways; since it is close to unity over the X-ray range, the most common form is complex:

$$n = 1 - \delta - i\beta \qquad [2]$$

where δ is the refractive index decrement, and the imaginary part β the absorption index. For photon energies corresponding to X-rays, both have small negative values, typically 10^{-3}–10^{-6}, and are correcting factors to the vacuum index 1. As in other photon energy ranges, the absorption and the refraction coefficient variations are linked through the Kramers–Krönig relations. In the visible range the index values are usually larger than 1, being below 1 only around specific optical thresholds. In the UV region, the index values become smaller than unity at energies above about 30 eV, which corresponds to the energy range for which atoms can be ionized.

At an early stage, Roentgen demonstrated X-ray absorption by sheets of materials. The demonstration of refraction took much longer and was more difficult. The first direct X-ray refraction index measurement was achieved in 1924, using a prism and an X-ray tube. Figure 2 is the scheme of the experiment, which is easy to perform in the visible range to measure the refraction of a transparent material. The footprint of the direct beam on the photographic plate is made before inserting the prism. In the visible range for which $n > 1$, the beam is refracted at larger angle than the glancing incidence angle, and so below the direct beam direction as shown in Figure 2a. At the contrary, as indicated in Figure 2b, the X-ray beam is refracted toward the prism surface and such an experiment effectively demonstrated that the index was slightly inferior to unity, as suggested by A.H. Compton. We will return

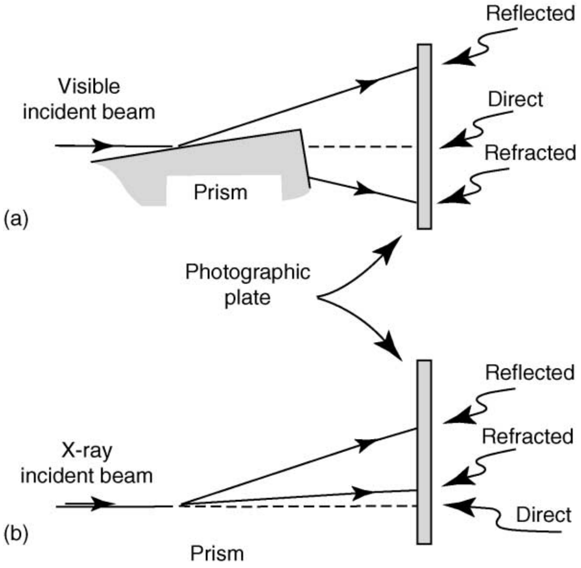

Figure 2 The geometry of the experiment used to evaluate the magnitude and the sign of the refractive index in the X-ray and the visible ranges.

to this peculiarity in the following section, when the total reflection phenomenon is described.

More details about the successive steps in our understanding and the pioneering works on X-rays can be found in standard textbooks.

Total Reflection X-Ray Mirrors

Using Fresnel's equations and tabulated values of the refractive index one can verify that, on any material, grazing incidence is necessary to reflect X-rays. Reflectivity calculations over the range from EUV to X-rays (30 eV to 10 keV), with approximations to allow calculations for rough surfaces, can be carried out using the web site of the Center for X-ray Optics of Lawrence Berkeley Laboratory http://www-cxro.lbl.gov, and curves similar to those presented below can be reproduced.

In the case of normal incidence, Fresnel's equation indicates that for any material the X-ray reflectivity $R^2 = |(n-1)/(n+1)|^2$ is very low, typically below 10^{-4}. This is a direct consequence of the X-ray index values, which are close to unity, as mentioned above.

In the visible range, normal incidence reflectivity from the surface of a transparent medium is low but can be increased just by going to glancing incidence. This is also the case in the X-ray region but in addition as the real part of the X-ray refractive index is lower than that of the vacuum, we have an 'external' total reflection phenomenon. From Figure 2 one can see how, for a sufficiently small incidence angle, the refracted wave becomes parallel to the interface. This is well known in the visible range (in which case it is the 'internal' reflection) and the phenomenon is used, for example, in fiber optics.

Figure 3 shows examples of calculated reflectivity variations versus grazing incidence angles, for perfectly smooth glass (SiO_2), nickel, and platinum surfaces. Curves are drawn for the 8 keV energy photons, corresponding to the Cu Kα line produced by conventional X-ray tubes. One observes that for this energy the reflectivity is low beyond a grazing angle around 1°, then rises quickly at low angles. This is typical of the total reflection phenomenon. The grazing angle corresponding to a 50% reflectivity is called the critical angle Θ_c, its value being roughly proportional to the refractive index decrement at the relevant photon energy. Notice that weaker absorption materials, lower Z, lead to more abrupt reflectivity variation but with smaller critical angles.

Such reflectivity variations at grazing incidence were systematically studied with X-ray tubes as early as the 1930s. The penetration depth of X-rays at grazing incidence was also studied. The evanescent wave inside the material, under total reflection conditions, is confined to a surface layer of a few tens of wavelengths. Hence, in the total reflection of X-rays, a coating having a thickness of a few tenths of a micron, on any substrate, is all that is required.

Figure 4 shows examples of the variation in nickel reflectivity calculated over the energy range 0.3–10 keV, for several grazing incidence angles. Nickel is one of the coatings commonly used for X-ray mirrors. The abrupt reflectivity variations and the dips correspond to the energies of the nickel atomic thresholds L (0.85 keV) and K (7.5 keV) for which the index tables predict a rapid change in absorption and refraction coefficients. Choosing an incidence angle provides a practical way of limiting the reflected spectral range. The energy cut-off of the

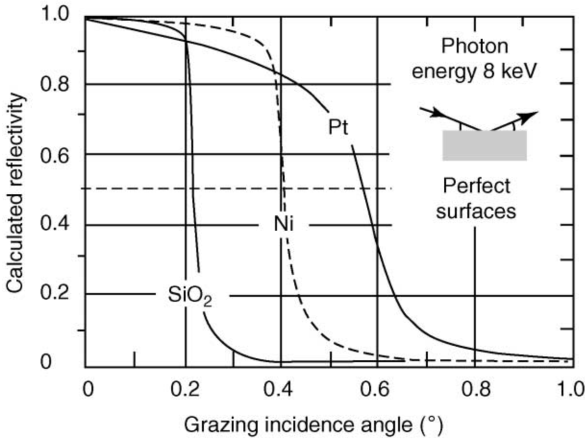

Figure 3 Examples of calculated reflectivity variations versus the incidence angle in the case of 8 keV X-rays on perfect platinum, nickel and glass surfaces.

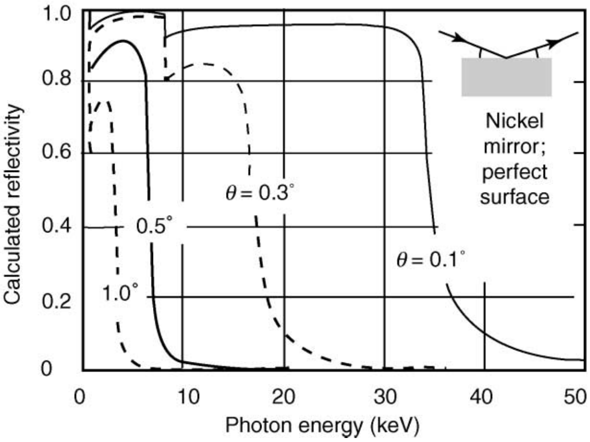

Figure 4 Calculated reflectivity variations versus photon energy, on a perfect nickel surface, for different incidence angles.

X-ray mirror acts as a filter for the higher-energy photons. That is effectively the method of suppressing the short wavelengths from an incident spectrum if they are not desired in an experiment.

When using highly divergent sources such as X-ray tubes, the restriction to operation at grazing incidence leads to a strong limitation of the collected solid angle, and thus of the beam size and angular aperture. These constraints are less stringent on highly collimated beams emitted by synchrotron sources. Thus, most synchrotron beam lines use total reflection optics in the form of curved mirrors, to refocus the beams and remove the undesired high X-ray energies. Recently, X-ray imaging optics, consisting of a series of parabolic and hyperbolic nested grazing incidence mirrors to increase the collected aperture, have been launched on satellites.

The reflectivity curves in Figures 3 and 4 are calculated for perfect surfaces. Real surfaces, even polished surfaces, have roughness created by defects over a large and continuous range of spatial frequencies, ranging from several centimeters down to the atomic scale. Coating a surface replicates the large period defects. Eventually new defects are added at near-atomic scale by nucleation of the coating material. Interferometric methods have been developed to study the range of defects from centimeter to micron spatial periods, whereas the atomic force microscope (AFM) is able to measure roughness from the micron down to the 0.1 nm scale. As waves are scattered most efficiently by asperities of a size comparable to their wavelength, scattering measurements are used in the X-ray region to probe defects of atomic scale. The prevalence of scattering explains why mirrors that reflect well in the visible range were disappointing when used in early X-ray experiments. The recent development of adaptive polishing methods has almost solved this problem.

The main effect of scattering by an optical system is not to decrease the number of photons coming out, but to modify their angular distribution. Scattering measurements are experimentally studied to test and evaluate the spatial defect distribution. In X-rays, one first characterizes the incident beam angular distribution in measuring its angular variation intensity after a scanning slit. Performing the same measurement after the optic, across the reflected beam in the case of mirrors, provides the modified angular distribution. From such a measurement, a few theoretical models adapted to the X-ray range can be used to deduce the mirror spatial defect distribution and so to check progress in polishing methods. The simplest model, coming from early crystal studies, is describing defects by a single parameter named roughness and denoted σ. Figure 5 shows an example of calculated reflectivity; on nickel surfaces versus photon energy and for increasing roughness; it shows how many photons have deviated from the symmetrical direction. Corresponding angular distributions can also be calculated.

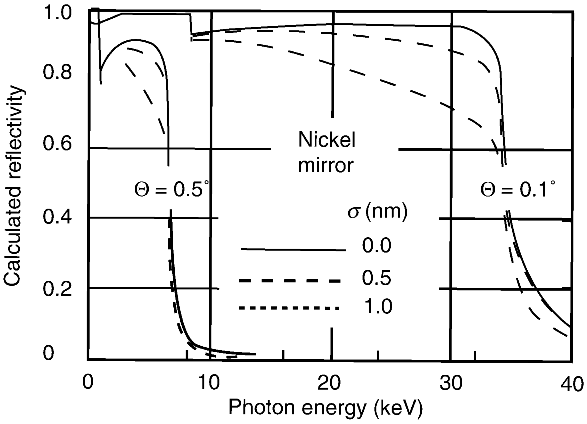

Figure 5 Calculated reflectivity variations versus photon energy, on nickel surfaces having different roughness, for different incidence angles.

The coating systems used for X-ray mirrors are material dependent. Deposition methods are similar to those used for visible optics coatings, but specific effort has been devoted to avoid nucleation growth to achieve uniform smooth layers. In general, one tries to give the evaporated particles, ions or clusters a kinetic energy lying within a specific range, which promotes mobility on the surface but is below the energy needed for implantation. In the case of electron beam deposition the energy distribution is thermal, but an auxiliary ionic bombardment can be used during deposition to provide the necessary energy to the previously deposited atomic layer (ion assisted deposition). For each material, the specific sputtering deposition and ionic bombardment conditions need to be carefully studied.

Until the 1970s, Bragg reflection by crystals and the total reflection phenomenon was the only way to reflect X-rays.

Multilayer X-Ray Mirrors

In the visible range, optics are frequently coated with stacks of dielectric multilayers obtained by periodic deposition of layers of two amorphous transparent materials. Such stacks are highly selective reflectors, or bandpass filters. They operate on the principle of constructive interference between all the waves partially reflected by the successive interfaces. They can be regarded as analogous to two-dimensional Bragg reflectors, in which the interfaces play the same role as atomic planes in crystals. Hence the Bragg equation [1] can be applied, the relevant period *d* being the thickness of an individual bilayer, if the stack is truly periodic.

The concept of using artificially deposited stacks of thin films, of alternating high and low electron density, to reflect and diffract X-rays dates back at least to the 1920s. It was not until 1944, however, that DuMond and Youtz were able to demonstrate significant X-ray reflection from such an artificially layered structure. Their work used alternating layers of copper and gold, laid down by evaporation. Unfortunately these materials diffuse into one another, reducing the electron density contrast and hence the X-ray reflectivity. In 1967 Dinklage and Frerichs succeeded in making a structure of alternating layers of iron and magnesium, which do not interdiffuse quickly, and thus produced the first stable X-ray reflecting multilayer.

Since then, enormous progress has been made in the fabrication and application of such multilayer reflectors, principally because of the efforts of two pioneers. Troy Barbee, while at Stanford University, developed open loop sputtering methods for controlling the thickness of the many thin (~nm) layers required to make up such a stack, while Eberhard Spiller at IBM developed an evaporation method using *in situ* X-ray monitoring to achieve similar thickness control. Almost all of the many multilayer deposition systems in use around the world are based on one of these two methods.

As for the total reflection mirror, only recent progress in methods for deposition and controls has made this new kind of X-ray optics practical. Present applications range from the extreme ultra violet (30 eV) region to hard X-rays having energies of a few tens of keV.

Prediction of the reflectivity variations can be simulated by calculations on the same web site mentioned in the preceding section. Multilayer mirrors are based on interference, so the reflectivity coefficient and the width of the reflected bandpass are determined by the amount of absorption inside the stack and the partial reflectivity at each interface. The more efficient pairs of materials have the lowest X-ray absorption and the highest refractive index difference at the required energy. The selection of real materials is constrained by the coating conditions and the need to achieve smooth interfaces. In particular, a low interdiffusion coefficient between the successive material pairs is a requirement for temporal stability. Epitaxial deposition methods have been considered but no useful pair of materials has yet been found.

An example of a flourishing application is multilayer-coated normal incidence X-ray optics. Optical schemes used in the visible region for telescopes and microscopes have excellent optical performance and high numerical aperture but a very low reflectivity in the X-ray region if coated with a single layer. However, if coated with multilayers they can be used, providing that extremely low-roughness surfaces can be produced. The interface qualities of the deposited thin layers presently limit normal incidence optics to the soft X-ray energy range, typically below 300 eV so around the carbon edge. This corresponds to periods of about 2 nm, or a few atomic layers, and the resulting roughness seriously decreases the reflectivity.

Figure 6 is a reflectivity curve versus energy for a quasinormal incidence Mo/Si multilayer mirror optimized to get the first-order peak at 93 nm. On Mo/Si mirrors, precise control of the layer period and the quality of the interfaces allow the manufacture of mirrors with reflectivity only a few percent below the theoretical limit shown here. Considerable effort has been invested in such mirrors, with a view to applications to EUV projection lithography.

It is possible to build polarizer mirrors over the soft X-ray range. An incidence angle of 45° corresponds to the Brewster angle, because the X-ray index is close to unity. X-ray multilayer and semitransparent selective beam splitters now permit the development of Michelson interferometers for newly developed soft X-ray lasers.

Figure 7 is an example of a tungsten/carbon (W/C) reflectivity curve for a multilayer adapted to the 8 keV energy of Cu Kα, mentioned before. The curve

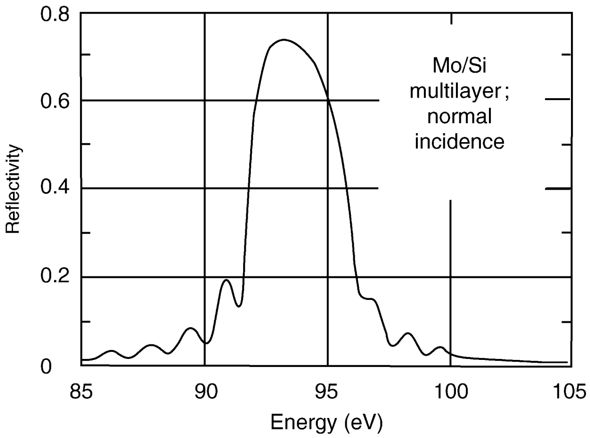

Figure 6 Calculated reflectivity versus incidence angle for a typical Mo/Si multilayer, optimized around 93 eV, under quasi-normal incidence.

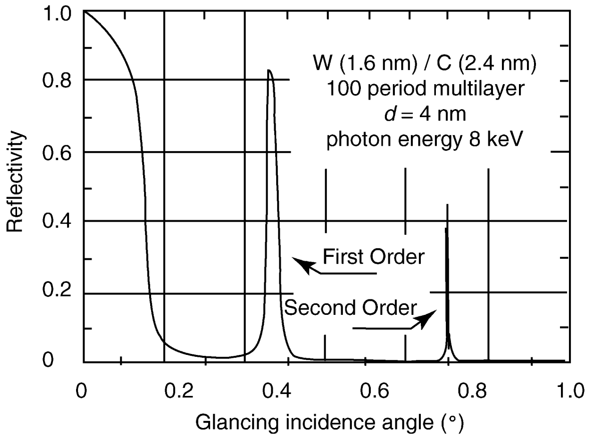

Figure 7 Calculated reflectivity versus incidence angle for a typical W/C multilayer adapted to reflect 8 keV photons.

of reflectivity versus glancing angle has been calculated from the total reflection range to the second-order peak. This kind of multilayer is being used in X-ray tubes to premonochromatize the beam and make it parallel. In such an application, the period of the coating is continuously graded laterally to satisfy locally the Bragg law at each point of the curved mirror.

As a final example of the new possibilities offered by X-ray multilayers, we mention the possibility of widening the reflected bandpass by varying continuously the period from the bottom to the top of the deposited stack. This is the equivalent of multilayer neutron ‘supermirrors’. Such interference mirrors are usable at quite high X-ray energies, and give high reflectivity and a larger incidence angle than total reflection mirrors working over the same energy range.

The continuous fast development of coating technologies and X-ray multilayer applications can be followed through the proceedings of dedicated meetings at, for example, on a web site like http://PXRM.

Summary and Perspectives

After the discovery of X-rays in 1895, almost 25 years were necessary to understand their characteristics compared to the visible range. Only in 1923 did F. Holweck and R.A. Millikan demonstrate the continuity of electromagnetic radiation from X-rays to ultraviolet. The selective Bragg reflection by natural crystals is limited to energy commonly larger than about 2 keV photons. Below this energy the total reflection mirror was the only optical device to reflect and focus wavelengths shorter than the far UV until the development of adapted interferential mirrors.

During the 1960s, the EUV and X-UV ranges gained interest due to the spread of fusion research and in the astrophysics community studying the Sun and stars. For both communities, the continuum and the lines emitted at short wavelengths by ionized matter were a convenient tool to evaluate high plasma temperatures. At about the same time, the development of synchrotron radiation sources increased the need for X-ray optics. In addition, this last source offered the possibility to attempt new optical solutions with convenient intense and parallel beams from UV to X-rays. Finally, the concomitant progress in various microtechnologies afforded the necessary tools to produce new adapted optical devices for short wavelengths. At the present time, adapted optics are under development to take advantage of various applications from microscopy to interferometry over the very large electromagnetic energy X-ray range.

As a matter of fact, it is now possible to tailor a choice of mirror coatings adapted to each X-ray photon energy and application. Such coatings are specific because they must be optimized to take into account the inevitable absorption energy dependence.

Progress achieved in blank polishing and evaporation methods has led to efficient and high-quality X-ray mirrors even able to keep the spatial coherence of synchrotron sources. Such technological advances are also steps to master surfaces and interfaces at the atomic scale. They are now becoming applicable to the other spectral ranges and to various micro- and nanotechnologies.

List of Units and Nomenclature

Crystal period	d	[nm]	1 nm = 10^{-9} m
Incidence Angle	Θ	[degree]	
Photon Energy	E	[eV]	$(h/2\pi)\omega.\lambda = hc =$ 1239.842 eV.nm
Wavelength	λ	[nm]	1 nm = 10^{-9} m

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Incoherent Sources:** Synchrotrons. **Optical Coatings:** Thin-film Optical Coatings.

Further Reading

Attwood D (1999) *Soft X-Rays and Extreme Ultraviolet Radiation*. Cambridge: Cambridge University Press.

Ditchburn RW (1963) *Light*. Appendix XIX, 2nd edn. London: Blackie.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Henke BL, Gullikson EM and Davies JC (1993) *Atomic Data and Nuclear Data Tables* 54: 181.

James RW (1982) *The Optical Principles of the Diffraction of X-Rays*. Woodbridge, CT: Ox Bow Press.

Michette A and Pfauntsch S (1997) *X-Rays: The First Hundred Years*. New York: John Wiley.

Mould RF (1993) *A Century of X-Rays and Radioactivity in Medicine*. Bristol, UK: Institute of Physics Publishing.

Paratt LG (1954) Surfaces studies of solids by total reflection of x-rays. *Physical Review* 95: 359.

Spiller E (1972) Low-loss reflection coatings using absorbing materials. *Applied Physics Letters* 20(9): 365–367.

Underwood JH and Attwood DT (1984) The renaissance of x-rays optics. *Physics Today, April*: 44–51.

OPTICAL COMMUNICATION SYSTEMS

Contents

Basic Concepts
Historical Development
Architectures of Optical Fiber Communication Systems
Free Space Optical Communications
Lightwave Transmitters
Local Area Networks
Optical Time Division Multiplexing
Wavelength Division Multiplexing

Basic Concepts

S Lee and A E Willner, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Since modern lightwave communication started in the 1970s, extensive efforts have been made to increase data rate and transmission distance. Recently, the bit rates for single-channel and multichannel systems have exceeded 40 and 2 Tb/s on single fibers respectively, for transoceanic distances. One revolutionary development in lightwave systems has been the deployment of erbium doped fiber amplifiers (EDFAs). These amplifiers facilitate transmission of an optical signal over long distances, by providing periodic analog-like amplification rather than digital-like regeneration. The wide bandwidth provided by EDFAs has made possible the increased use of wavelength division multiplexing (WDM) in which multiple lightwave signals, each having a different wavelength, are co-propagated on a single fiber. EDFAs and WDM techniques can enhance the lightwave system capacity, both in terms of obtainable transmission distance, and total number of data rates (see Figure 1).

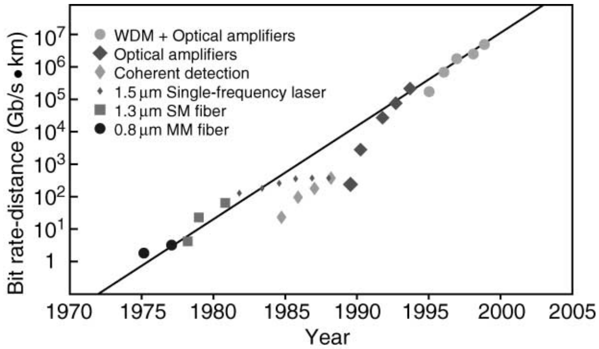

Figure 1 Bit rate product. Data from T. Li (private communication).

Optical Fiber

Most telecommunication fibers are made of silica, therefore following the attenuation of silica material. Figure 2 shows the silica fiber attenuation as a function of wavelength. The downward slope towards 1.6 μm and upward slope away from 1.6 μm, are the theoretical limits due to Rayleigh scattering and absorption of silica material, respectively. The absorption peak near 1.4 μm is due to water molecule absorption. There are two low attenuation regions: ~1.3 μm and 1.55 μm, both of which are used for communications in general, and some other wavelengths are also used for shorter distance communications. Figure 2 also shows the comparison between gain bandwidth of EDFA (~3 THz) and the low attenuation window around 1.55 μm (~25 THz). The typical values of attenuation at 1.3 μm and 1.55 μm for single-mode fiber (SMF) are ~0.35 dB/km, and 0.2 dB/km, respectively.

When an electromagnetic wave interacts with the bound electrons of a silica fiber, the medium response depends upon optical frequency (ω). This property, referred to as material dispersion, manifests itself through the frequency dependence of the refractive index $n(\omega)$. Fiber material dispersion plays a critical role in propagation of short optical pulses since different spectral components associated with the pulse travel at different speeds, given by $c/n(\omega)$. Consequently, the optical pulse at the output of the fiber is broadened in time. The commonly used system parameter is called the dispersion parameter D (the negative D value is called normal dispersion, and the positive D value is called anomalous dispersion), and is measured in ps/(nm·km); D for SMF is ~+ 17 ps/(nm · km) at 1.5 μm. However, because of dielectric waveguiding, the effective mode index is slightly lower than the material index $n(\omega)$, with the reduction itself being ω-dependent (waveguide dispersion). This results in a waveguide contribution that must be added to the material contribution to obtain the total dispersion (see Figure 3).

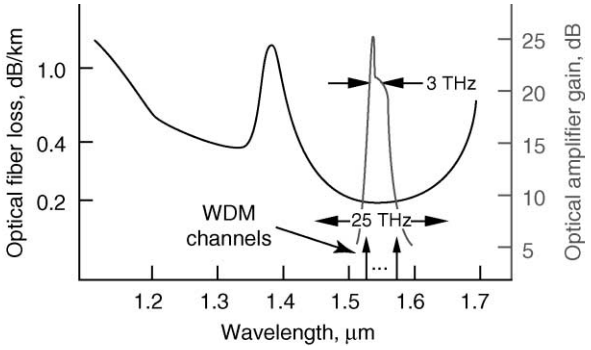

Figure 2 Silica fiber attenuation versus wavelength.

Generally, the waveguide contribution to dispersion D is to shift zero dispersion wavelength (λ_0) to longer wavelengths. Furthermore, the waveguiding can be tailored to shift the λ_0 from 1.3 μm to 1.5 μm, the wavelength at which the fiber has the minimum attenuation loss. The fiber having λ_0 in the neighborhood of 1.5 μm, is called dispersion-shifted fiber (DSF); D for DSF is usually between −2.5 and +2.5 ps/nm·km, at a wavelength of 1.5 μm. The typical dispersion parameter, D, as a function of a wavelength for both SMF and DSF, is shown in Figure 4.

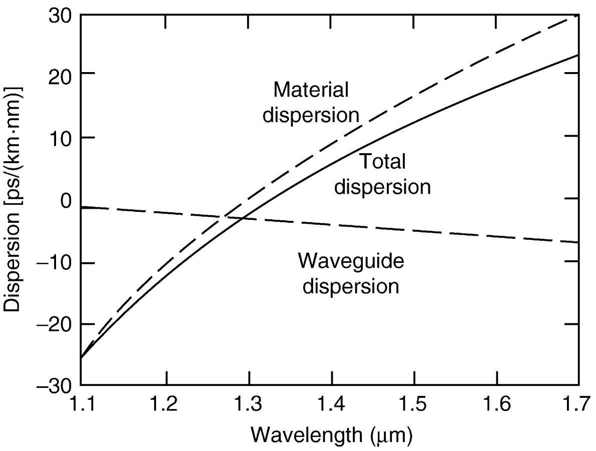

Figure 3 Total dispersion (D) and contribution of material dispersion and waveguide dispersion in conventional SMF.

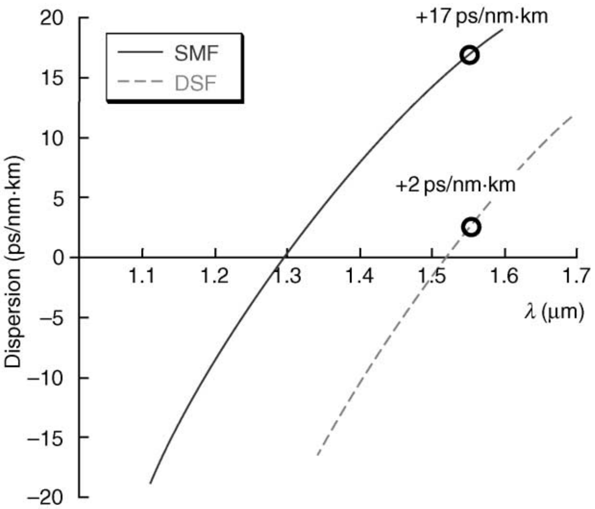

Figure 4 Dispersion versus wavelength for conventional SMF and DSF.

In addition to chromatic dispersion, optical fiber has another dispersion properties related to polarization. An optical wave of arbitrary polarization can be represented as the superposition of two orthogonally polarized modes. In an ideal fiber, these two modes are indistinguishable, and have the same propagation constants owing to the cylindrical symmetry of the waveguide. However, in real fibers there is some residual anisotropy due to unintentional circular asymmetry, usually caused by noncircular waveguide geometry or asymmetrical stress around the core, as shown in Figure 5.

In either case, the loss of circular symmetry gives rise to two distinct orthogonally-polarized modes with different propagation constants (differential phase velocity) responsible for polarization mode dispersion (PMD) in the fiber, and can be related to the difference in refractive indices (birefringence) between the two orthogonal polarization axes. The differential phase velocity indicated is accompanied by a difference in the group velocities for the two polarization modes, therefore the pulse at the end of the transmission fiber is broadened by this differential group delay (DGD). DGD is usually expressed in units of ps/km for a short length (1 m to 1 km) of birefringent fiber. However, DGD does not accumulate along a long fiber link in a linear fashion. Instead, because of random variations in the perturbations along a fiber span, the DGD in one section may either add to or subtract from another section of the fiber. As a result, average DGD in long fiber spans accumulates in a random-walk-like process that leads to a square root of transmission-length dependence. Therefore, average DGD is expressed in ps/km$^{1/2}$ in long fiber spans, referred to as the PMD of the fiber, and the typical PMD parameter has a value of 0.01 to 10 ps/km$^{1/2}$. Because of the many perturbations that act on a real-world fiber (e.g., temperature, vibration, etc.), transmission properties typically vary with time. Therefore, the PMD of the fiber link fluctuates randomly, thus causing random fluctuations in system performance.

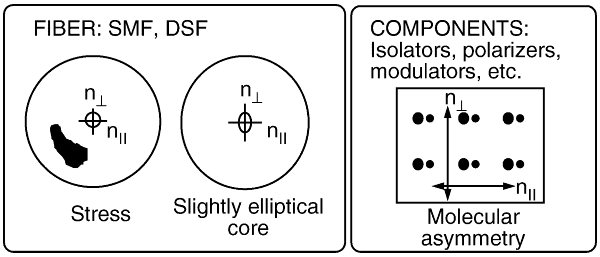

Figure 5 Origin of PMD.

It is well known that optical fibers show nonlinear behavior under conditions of high power and long interaction length. The power-times-distance products for amplified transmission systems can be large enough to make fiber nonlinear effects the dominant factor in the design of long-distance systems.

There are two categories of fundamental optical nonlinear effects: stimulated scattering effects and refractive index effects. Stimulated scattering effects arise from parametric interactions between light and acoustic or optical phonons in the fiber. Two nonlinear effects fall into this category: stimulated Brillouin scattering (SBS) and stimulated Raman scattering (SRS). The main difference between the two is that optical phonons participate in SRS, while acoustic phonons participate in SBS. In a simple quantum-mechanical picture, a photon of the incident field is annihilated to create a photon at a longer wavelength. The new photon is co-propagated and counter-propagated along the original signal in SRS, while counter-propagated in SBS. Refractive index effects are caused by modulation due to changes in light intensity. There are three types of refractive index effects: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). SPM introduces the change of optical phase by its own intensity, and, this leads to frequency chirp of the pulse, depending on the relative position from the peak. When this nonlinear frequency chirp interacts with the fiber dispersion, the pulse either broadens or compresses, depending on the sign and the amount of dispersion along the fiber. In XPM, the phase of the signal in one wavelength channel is modulated by the intensity fluctuation of the other wavelength channels. In a WDM system, XPM imposes far more damaging effects than SPM because XPM is stronger by a factor of two than SPM when the channel power is the same, and large number of WDM channels can contribute to XPM. FWM is the generation of modulation sideband at new frequencies, due to the phase modulation of channels between lights at different frequencies in multichannel system. FWM causes penalties in a WDM system if the newly generated frequency is either equal to or close to the frequency of existing WDM channels.

Fiber nonlinearities impose significant degradation in optical transmission system, and limit the system on allowable number of channels, channel power, and channel spacing (see Figure 6). But, the effects can be well suppressed by carefully designing the fiber dispersion profile (i.e., dispersion map) in the transmission system, and controlling the optical launch power for each fiber span.

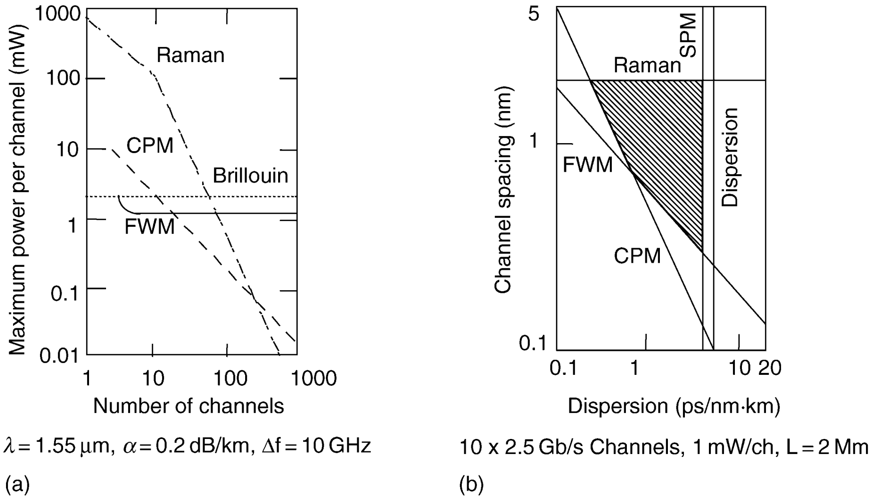

Figure 6 Limitation due to fiber nonlinearities in (a) maximum channel power and number of channels, and (b) channel spacing and dispersion. Reproduced with permission from Chraplyvy AR (1990) Limitation on lightwave communications imposed by optical-fiber nonlinearities. *IEEE Journal of Lightwave Technology* 8:1548–1557. Copyright © 1990 IEEE.

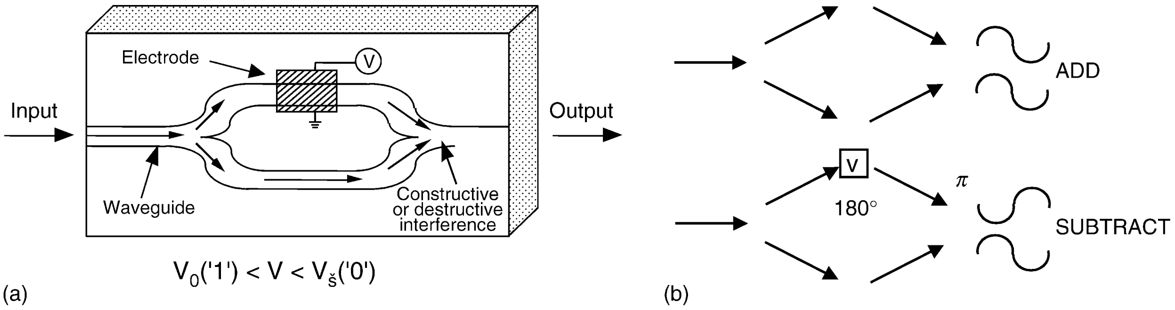
Electrode
V
Input
Output
Waveguide
Constructive or destructive interference
$V_0('1') < V < V_s('0')$
(a)
V
180°
π
ADD
SUBTRACT
(b)

Figure 7 (a) EOM: MZ interferometer on LiNbO2, (b) operation of EOM; constructive and destructive interference resulting in intensity modulation.

Communication Components

Distributed feedback laser (DFB) is widely used as a light source for metro, long-haul, and undersea applications, due to its narrow spectral width, and wavelength stability. Fabry–Perot (FP) lasers, and vertical cavity surface emitting lasers (VCSEL) are used for local area networks (LAN), and for access applications, because these are cheaper than DFB lasers. For an application, where the cost is the most important parameter with low data modulation speed, light-emitting diode (LED) is used as a light source.

In optical data modulation, the simplest and easiest technique is turning ON or OFF the laser, depending on the binary logic '1' or '0' (direct modulation). This type of modulation is simple and cheap compared to external modulation technique. But, its application is limited by the bandwidth of modulation, and the induced frequency chirp (i.e., difference in optical frequency at the turn ON state and just before the turn OFF state of laser) that limits the transmission distance. Two different types of external modulators are widely used for digital optical data transmission. Electro absorption modulator (EAM) uses the Franz–Keldysh effect, where it is observed that the wavelength of the optical absorption edge in a semiconductor is lengthened by applying an electric field. Electro-optic modulator (EOM) utilizes the linear electro-optic effect, called the Pokels effect, which changes the refractive index of the material caused by and proportional to an applied electric field, so therefore changes the phase of the optical signal. This phase change is used to modulate the intensity of a lightwave through a Mach–Zehnder (MZ) interferometer (see Figure 7). Because the Pockels effect exists only in crystal, lithium niobate crystal (LiNbO$_2$) or electro-optic polymers are used for EOM.

EDFA is a wideband optical amplifier that has merits in that: (i) erbium ions (Er^{3+}) emit light in the 1.55 μm loss-minimum band of optical fiber, (ii) a circular fiber-based amplifier is inherently compatible with a fiber optics system; (iii) it provides

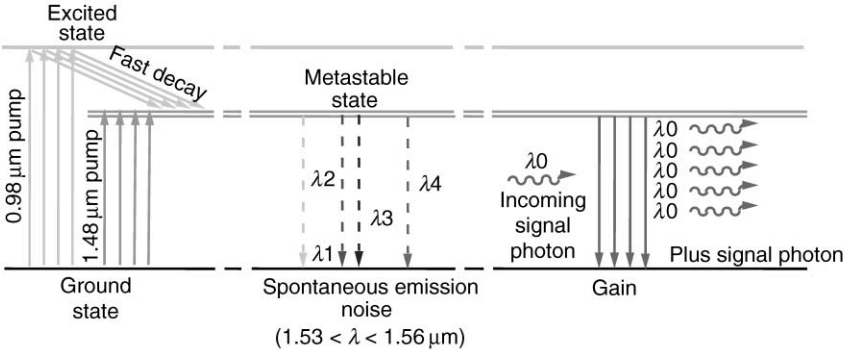

Figure 8 Energy level and transition diagram of EDFA. Reproduced with permission from Willner AE (1997) Mining the optical bandwidth for a terabit per second. *IEEE Spectrum* Apr.: 32–41. Copyright © 1997 IEEE.

amplification independent of bit-rate, modulation format, power, and wavelength; and (iv) it has low distortion and low noise during amplification. EDFA contains a gain medium (i.e., erbium-doped fiber) that must be inverted by a pump source. A signal initiates stimulated emission, resulting in gain, and spontaneous emission occurs naturally, which results in noise.

Figure 8 shows the energy level and transition diagram of EDFA. A 0.98 μm pump photon is absorbed and excites a carrier (Er^{3+} ions) into higher excited states, and the excited carrier decays rapidly to the first excited state that has a very long lifetime of ~10 ms (metastable state). In contrast, a 1.48 μm pump photon is absorbed and excites a carrier into a metastable state. This long metastable state is one of the key advantages of EDFA, and intersymbol distortion and interchannel crosstalk are negligible as a result of these slow dynamics. Depending on the external optical excitation signal, this carrier will decay in a stimulated or spontaneous fashion to the ground state and emit a photon. Both absorption and emission spectra have an associated bandwidth depending on the spread in wavelengths that can be absorbed or emitted from a given energy level. This is highly desirable because (i) a pump laser does not need to be an exact wavelength; and (ii) the signal may be at one of several wavelengths, especially in a WDM system. Figure 9 shows the bandwidth in the 1.48 μm absorption and 1.55 μm fluorescence spectrum of a typical erbium-doped amplifier.

When the noise characteristics of EDFA are considered, noise figure (NF) is an important parameter, which is defined by $\mathrm{NF} \equiv \mathrm{SNR}_{\mathrm{in}}/\mathrm{SNR}_{\mathrm{out}}$, where $\mathrm{SNR}_{\mathrm{in}}$ ($\mathrm{SNR}_{\mathrm{out}}$) is the electrical equivalent signal-to-noise ratio (SNR) of the optical wave going into (coming out) of the amplifier, if it were be detected. The typical value of NF in commercially-available EDFA is 4.5–6 dB. The importance of NF in optical communication systems is presented later in this chapter, when the required optical SNR (OSNR) for error-free optical data transmission is considered. Figure 10 shows an example of EDFA diagram.

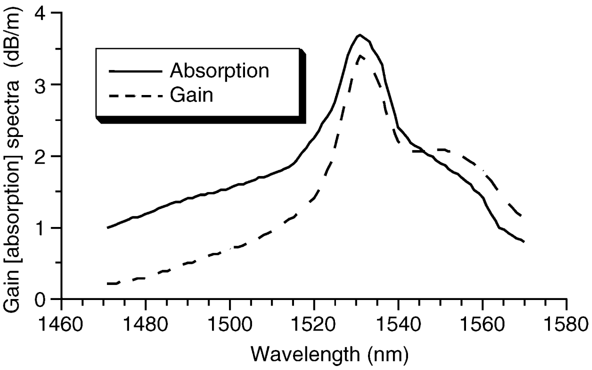

Figure 9 The absorption and fluorescence spectra for erbium near 1.5 μm. Reproduced with permission from Miniscalco WJ (1991) Erbium-doped glasses for fiber amplifiers at 1550 nm. *IEEE Journal of Lightwave Technology* 9: 234–250. Copyright © 1991 IEEE.

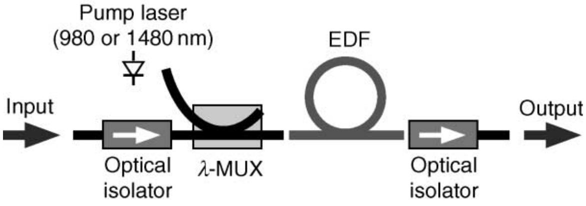

Figure 10 Typical EDFA diagram.

Recent development of EDFA in longer wavelength regimes or L-band (1570–1620 nm) has doubled the useful transmission bandwidth over the conventional band or C-band (1525–1265 nm). L-band gain of EDFA is generally achieved by proper arrangement of pump lasers through longer lengths of erbium-doped fiber than conventional C-band EDFA. Figure 11a shows the concept of wideband EDFA. The incoming

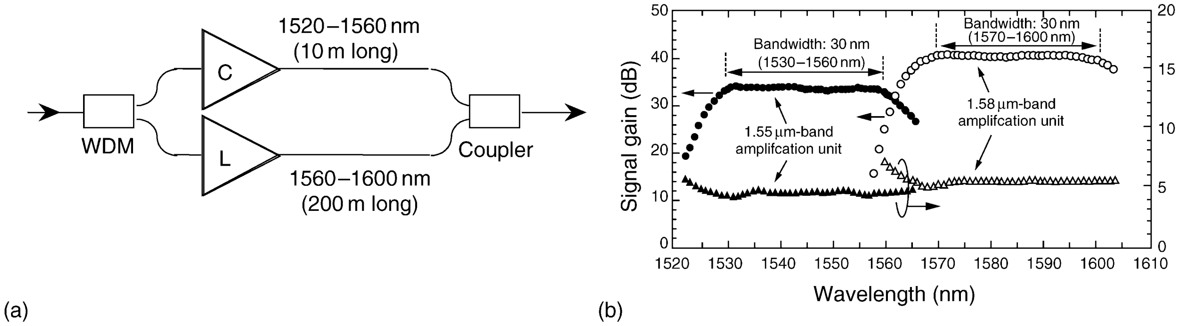

Figure 11 Wide band EDFA. Reproduced with permission from Yanada M (1997) *Electronics Letters*. Copyright © 1997 IEEE.

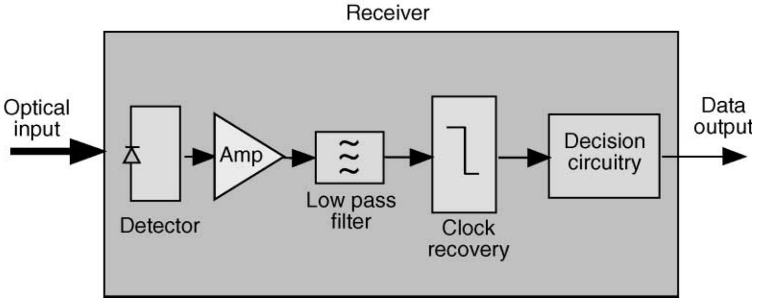

Figure 12 Block diagram of functions performed in advanced receiver packages.

WDM signal is split into two bands (C- and L-bands), amplified separately, and combined at the output. Gain and NF characteristics of this wideband EDFA is shown at Figure 11b.

Because SRS transfers the energy of a lower wavelength channel into the higher wavelength channel in the optical fiber, the transmission fiber itself can be used as a gain medium on specific channel (or channels) when the proper pump signal (or signals) are provided in the fiber. Even though Raman gain is much lower than the gain of EDFA (i.e., <0.1 dB gain/mW pump in Raman amplifier (RA) compared to a few dB gain/mW pump in EDFA), RA has several advantages over EDFA (i) RA has the capability to provide gain at any signal wavelengths, (ii) the amplification window is expandable by combining multipump wavelengths; and (iii) RA offers improved noise performance because the signal is amplified over transmission fiber (i.e., distributed amplifier). Especially improved noise performance increases system OSNR, which can be used to extend system reach or create more wavelength channels on the system. The Raman pumps are typically located ~100 nm shorter than the wavelength of the signal to be amplified, and a few 100s mW pump lasers are used to get a 10–15 dB gain.

In general, telecom receiver (RCVR) consists of photo diode (PD) with transimpedence amplifier (TIA), followed by limiting amplifier (LA) or automatic gain control amplifier (AGC), low pass filter, and clock data recovery circuit (CDR) (see Figure 12).

Vertically illuminated positive-intrinsic-negative (PIN) photodiode and avalanche photo diode (APD) are the most common PD that is used in RCVR. In PIN, incoming light is absorbed in the n^- (low doped or intrinsic) region and generates an electron-hole pair, which drifts to a depletion region and collected n^+ and p^+ regions, therefore generating electric current. APD is used for high-sensitivity detection up to 10 Gb/s. It is structured as absorption, grading, and multiplication layers. The multiplication layer of APD provides additional transit time delay compared to PIN, therefore the bandwidth of APD is smaller than PIN in general. TIA is a broadband low noise amplifier that is used to convert and amplify the weak current from PD to the desired output voltage. A low pass filter is used to suppress the noise and data distortion from high frequency components with a typical cut-off frequency ~0.7 bitrate. LA and AGC are amplifiers that limit the output voltage as a constant. LA is a nonlinear amplifier that makes a decision either to make logic '1' or '0' based on input voltage level, and limits output voltage as fixed values. AGC is a linear amplifier that limits the output voltage by adjusting the gain of the amplifier based on input voltage level. CDR consists of a phase lock loop (PLL) with a voltage controlled oscillator (VCO)

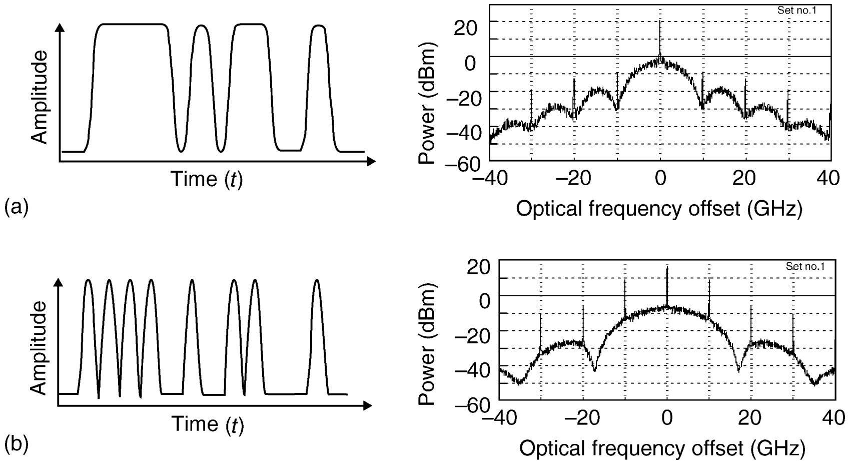

Figure 13 Amplitude modulation (a) nonreturn-to-zero (NRZ), and (b) return-to-zero (RZ).

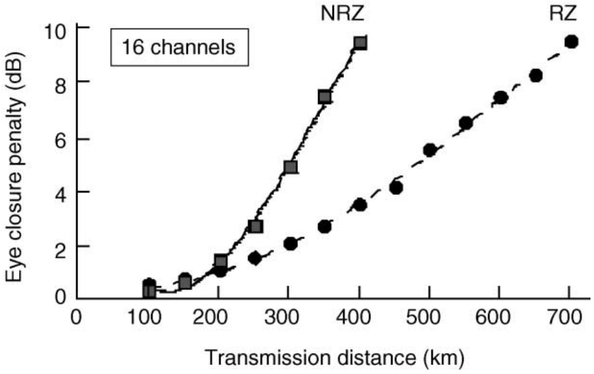

Figure 14 System performance of different modulation format (NRZ vs. RZ).

and data flip-flop circuit. The PLL is used to synchronize the VCO to the incoming high-speed bit stream, therefore recovering the system clock. The recovered clock is used in D-FF to retime and regenerate the data.

Data Modulation Formats

In digital optical communications, numerous modulation formats have been developed and utilized, in order to achieve higher spectral efficiency (i.e., ratio of individual channel bit rate to dense WDM (DWDM) separation), and to improve system performance over fiber impairments. A simple amplitude modulation of lightwave is still the most widely used method for optical communication. The easiest amplitude modulation technique is to using a nonreturn-to-zero (NRZ) format. Figure 13a shows an example of NRZ binary sequence in time domain, and the corresponding frequency domain spectrum. In NRZ, data '1' and '0' correspond to the ON and OFF state of a lightwave transmitter, respectively, occupying entire bit time (i.e., binary symbol time). When the ON or OFF state of a lightwave transmitter does not occupy the entire bit time (T_B) (i.e., when the data pulse duration (T_d) is less then (T_B), it is called the return-to-zero (RZ). Figure 13b shows a binary sequence of RZ, and its frequency spectrum when $T_d \approx 0.5\ T_B$. The Fourier transform of the square wave is the Sinc function. Therefore, the frequency spectrum of an ideal NRZ and RZ follows the pattern of a Sinc square function. Note that the first null of frequency spectrum of RZ is about twice as high as the null of NRZ.

Even though NRZ is better in spectral efficiency than RZ, RZ has an advantage over NRZ in tolerance of fiber impairments. Figure 14 shows an example on eye closure penalty comparison between NRZ and RZ data formats at a 16-channel WDM system at 40 Gb/s data rate. It is clearly seen that the penalty of RZ is much less severe than the penalty of NRZ, as the transmission distance increases. (Note, nonlinearities and dispersion increases as the transmission distance increases.)

There has been much effort to increase the spectral efficiency, to improve the system performance over fiber impairments, or to achieve both simultaneously. Figure 15 shows some of examples of these. Chirped RZ (CRZ) is the modulation format inducing frequency chirp on conventional RZ. CRZ is robust on fiber impairment at the cost of losing spectral efficiency. In carrier-suppressed RZ (CS-RZ), a strong un-modulated optical carrier is removed from conventional RZ, therefore suppressing the nonlinear effects. Duo-binary is a three-level signal modulation format generated by a series of delay and added

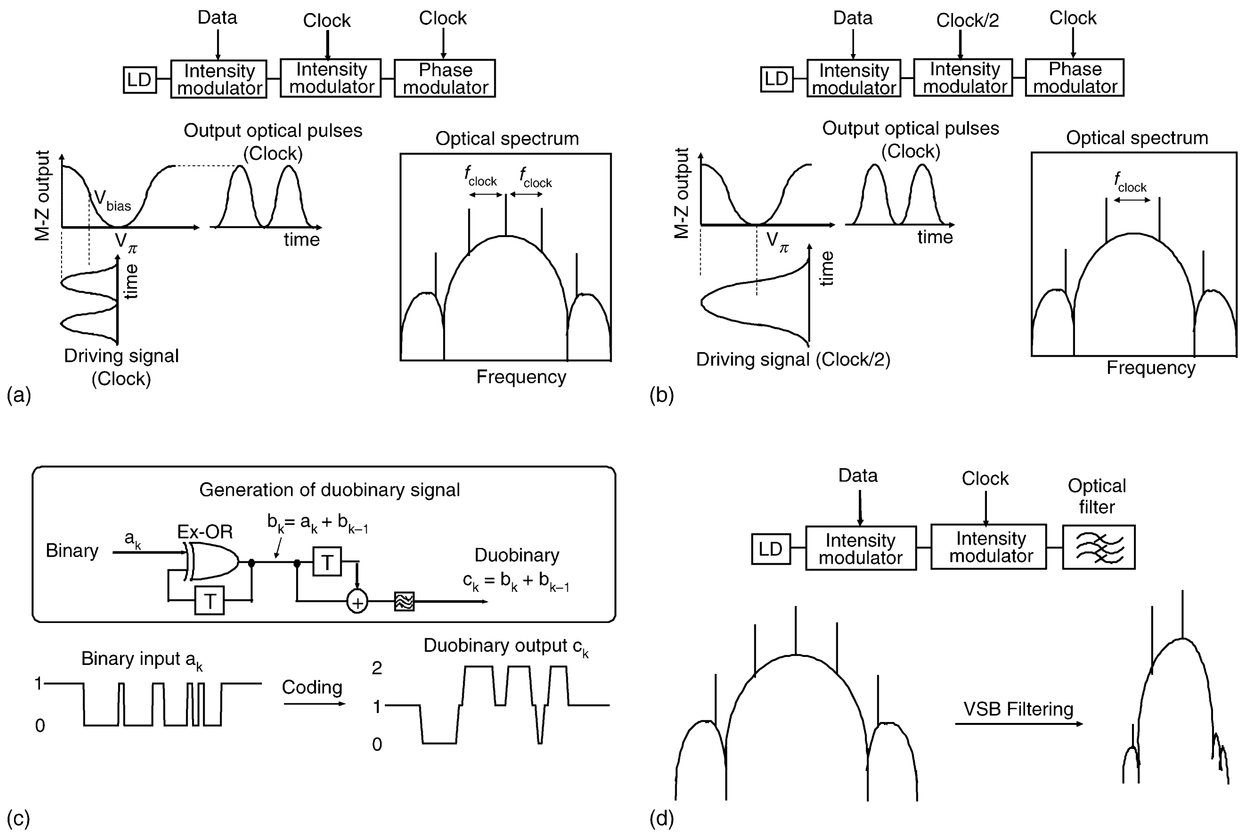
Data
Clock
Clock
LD
Intensity modulator
Intensity modulator
Phase modulator
Output optical pulses (Clock)
Optical spectrum
M-Z output
V_{bias}
V_π
time
time
f_{clock}
f_{clock}
Driving signal (Clock)
Frequency
(a)
Data
Clock/2
Clock
LD
Intensity modulator
Intensity modulator
Phase modulator
Output optical pulses (Clock)
Optical spectrum
M-Z output
V_π
time
time
f_{clock}
Driving signal (Clock/2)
Frequency
(b)
Generation of duobinary signal
Binary
a_k
Ex-OR
$b_k = a_k + b_{k-1}$
T
T
Duobinary
$c_k = b_k + b_{k-1}$
Binary input a_k
1
0
Coding
Duobinary output c_k
2
1
0
(c)
Data
Clock
Optical filter
LD
Intensity modulator
Intensity modulator
VSB Filtering
(d)

Figure 15 Examples of bandwidth efficient modulation formats. (a) Chirped RZ (CRZ) generation, (b) carrier suppressed-RZ (CS-RZ) generation, (c) duo-binary generation, and (d) vestigial single side band (VSB) RZ generation.

circuits in the transmitter. The advantage of duo-binary is that its bandwidth is reduced to one half of conventional NRZ, therefore having enhanced spectral efficiency, and has a high tolerance in chromatic dispersion, and suppression of fiber nonlinearites.

Single-sideband (SSB) or vestigial singleside band (VSB) can be generated into either NRZ or RZ, by either optically utilizing optical filtering with a sharp cut-off filter, or electrically applying tapped-delay-line filter approximation on both amplitude and phase modulators. SSB (or VSB) has about a two times higher spectral efficiency and higher tolerance on chromatic dispersion than conventional modulation formats.

Data Multiplexing

At their most basic, optical networks can imitate electrical networks in which time division multiplexing (TDM) is overwhelmingly used for digital data transmission. A fiber can carry many time-multiplexed channels, in which each channel can transmit its data in an assigned time slot. A typical TDM link is shown in Figure 16, in which N transmitters are sequentially polled by a fast multiplexer to transmit their data. The time-multiplexed data are sequentially and rapidly demultiplexed at the receiving node.

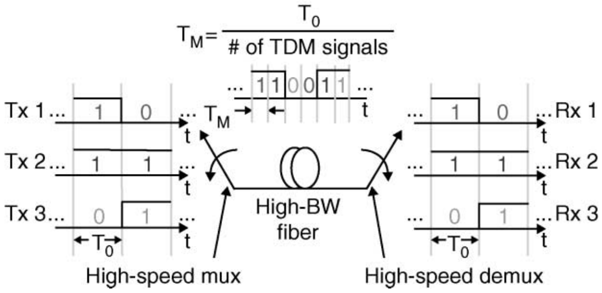
$T_M = \frac{T_0}{\text{\# of TDM signals}}$
Tx 1
Tx 2
Tx 3
1 0
1 1
0 1
T_0
t
T_M
1 1 0 0 1 1
High-BW fiber
High-speed mux
High-speed demux
Rx 1
Rx 2
Rx 3
T_0

Figure 16 Concept of bit-interleaving TDM.

Time multiplexing and demultiplexing functions can be performed either electrically with an electrical time (DE)MUX switch and an optical transmitter/receiver, or optically with multiple optical transmitters/receivers and an optical time (DE)MUX switch. The major advantage of TDM is that there is no output–port contention problem (each data bit occupies its own time slot and there is only a single high-speed signal present at any given instant). The disadvantage of TDM is that the scheme requires a ultra-high-speed switching component if the individual signals are themselves high-speed and if there are many users.

In WDM, multiple wavelength channels are transmitted through a single fiber, therefore enabling the fiber to carry more throughput. By using wavelength-selective devices for the ON and OFF ramps, independent signal routing also can be accomplished. Figure 17a shows a simple point-to-point WDM system in which several channels are multiplexed at one node, the combined signals transmitted across a distance of fiber, and the channels demultiplexed at a destination node. As shown in Figure 17b, the wavelength becomes the signature address for either the transmitter or the receivers, and the wavelength will determine the routing path through an optical network.

Many interesting challenges face the eventual implementation of WDM systems and networks. Several of these challenges involve the control and management of the data through this novel high-speed network. Figure 18 shows a small subset of critical component technologies typically required for a WDM system, including multiple-wavelength transmitters, multiport star couplers, passive and active wavelength routers, EDFAs, and tunable optical filters.

Some generic goals that a WDM-device technologist aims to achieve include; large wavelength tuning range; multi-user capability; wavelength selectivity and repeatability; low cross-talk; high extinction ratio; minimum excess losses; fast wavelength tunability; high-speed modulation bandwidth; low residual chirp; high finesse; low noise; robustness; high yield; potential low cost. Depending on system requirements, device availability and cost, WDM technologies divide into two; course-WDM (CWDM) and dense-WDM DWDM. Figure 19 shows the wavelength allocation for CWDM and

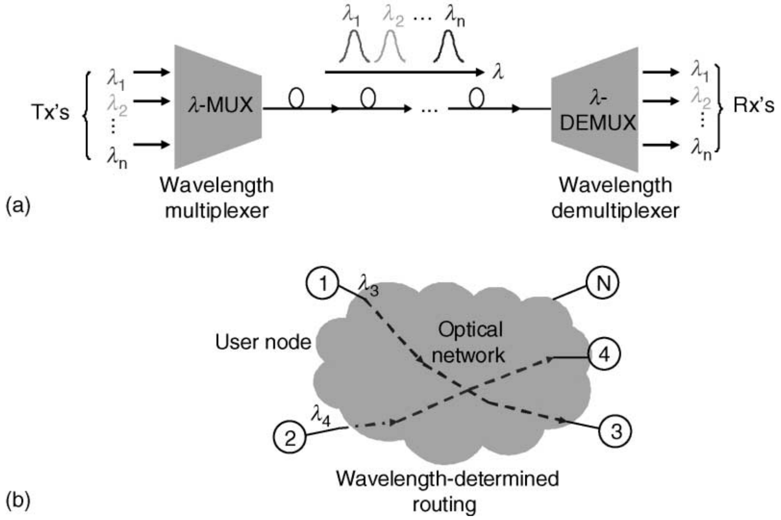

λ_1 λ_2 ... λ_n
λ
Tx's
λ_1
λ_2
λ_n
λ-MUX
Wavelength multiplexer
λ-DEMUX
Wavelength demultiplexer
λ_1
λ_2
λ_n
Rx's
(a)
1
λ_3
N
User node
Optical network
4
λ_4
2
3
Wavelength-determined routing
(b)

Figure 17 (a) A simple point-to-point WDM transmission system; (b) a generic multiuser network in which the communication links and routing paths are determined by the wavelengths used within the optical switching fabric.

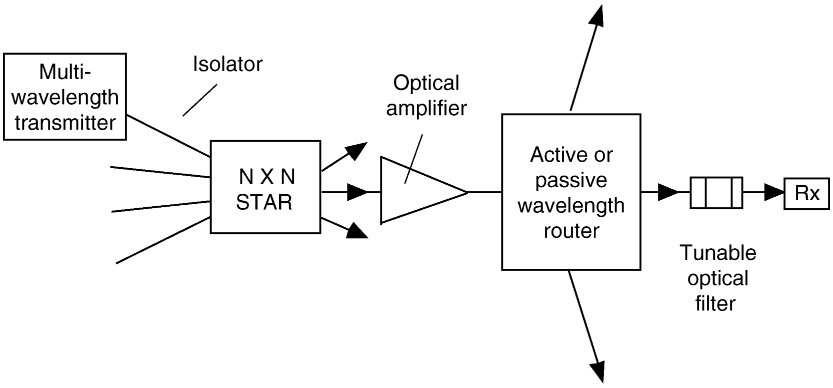

Multi-wavelength transmitter
Isolator
Optical amplifier
N X N STAR
Active or passive wavelength router
Rx
Tunable optical filter

Figure 18 Schematic of a small subset of enabling device technologies for a WDM system.

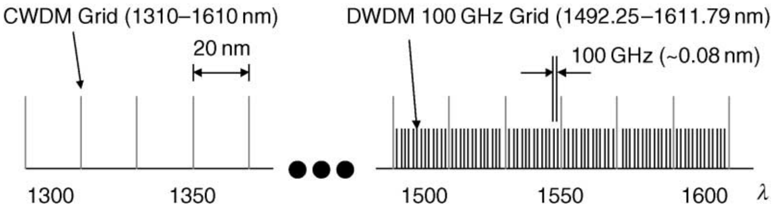
CWDM Grid (1310–1610 nm)
20 nm
DWDM 100 GHz Grid (1492.25–1611.79 nm)
100 GHz (~0.08 nm)
1300
1350
1500
1550
1600
λ

Figure 19 WDM wavelength allocation.

DWDM. CWDM technology uses an International Telecommunication Union (ITU) standard 20 nm spacing between the wavelengths from 1310 nm to 1610 nm. Also DWDM technology uses an ITU standard 100 GHz or 200 GHz between wavelengths, arranged in several bands at around 1500–1600 nm.

Dispersion Management and Compensation

As mentioned previously, chromatic dispersion is one of the most basic characteristics of fiber, although it is possible to manufacture fiber that induces zero chromatic dispersion. But such fiber is incompatible with the deployment of a WDM system since harmful nonlinear effects are generated, therefore chromatic dispersion must exist, and must be compensated for. The effect of chromatic dispersion is cumulative and increases quadratically with the data rate (see Figure 20).

The quadratic dependence of dispersion with the data rate is a result of two effects. First, a doubling of the data rate will double the Fourier-transformed frequency spectrum of the signal, thereby doubling the effect of dispersion. Second, the same doubling of the data rate makes the data pulse only half as long in time and therefore twice as sensitive to temporal spreading due to dispersion. The conventional wisdom for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit time period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by $L_D = 1/(D \cdot B \cdot \Delta\lambda)$ (see Figure 21).

In theory, compensation of chromatic dispersion for high-speed or long-distance systems can be fixed in value if each link's dispersion value is known. A simple yet elegant solution is to create a dispersion 'map', in which the designer of a transmission link alternates elements that produce positive and then negative dispersion (see Figure 22). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, effectively

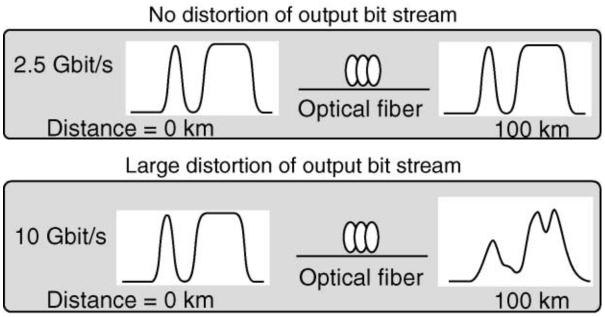
No distortion of output bit stream
2.5 Gbit/s
Distance = 0 km
Optical fiber
100 km
Large distortion of output bit stream
10 Gbit/s
Distance = 0 km
Optical fiber
100 km

Figure 20 Pulse broadening at two different data rates (2.5 Gbit/s, and 10 Gbit/s) as a result of quadratic nature of chromatic dispersion.

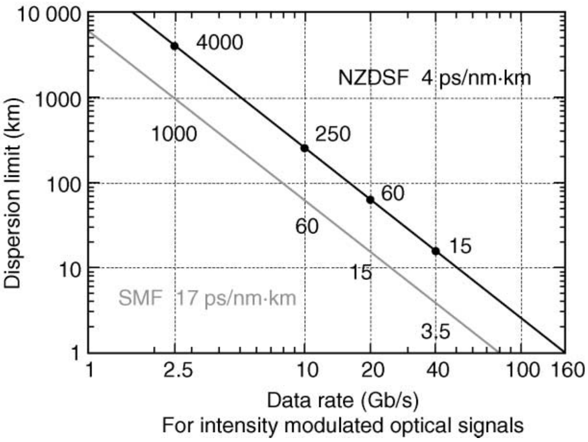

Figure 21 Dispersion-limited (uncompensated) transmission distance in single-mode fiber (SMF) as a function of data rate for intensity-modulated optical signals. Courtesy of L.D. Garrett.

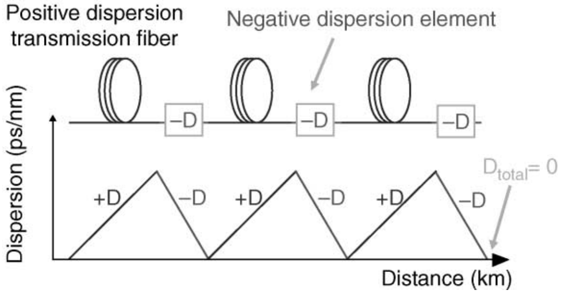
Positive dispersion transmission fiber
Negative dispersion element
–D
–D
–D
Dispersion (ps/nm)
+D
–D
+D
–D
+D
–D
D_{total}= 0
Distance (km)

Figure 22 Typical static management of chromatic dispersion.

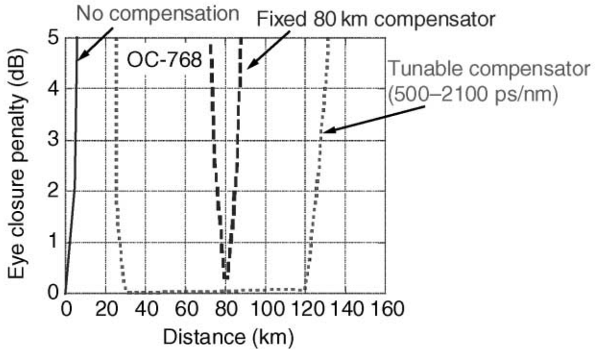

Figure 23 Effect of tunable compensation at 40 Gb/s (OC-768); it allows a wide range of transmission distance. On the contrary, the 80 km fixed compensator allows only small range of transmission around 80 km.

eliminating FWM and XPM, but the total accumulated dispersion at the end of the fiber link is zero, so that minimal pulse broadening is induced. The specific system design, as to the periodicity of management, depends on several variables, but a typical number for SMF as the embedded base is a compensation at every 80 km in a 10 Gb/s system. Dispersion-compensating fiber (DCF) has been generally used as a dispersion compensating element.

However, there are several important aspects of optical systems and networks that make tunable dispersion compensation solutions attractive, including: (i) it significantly reduces the inventory of different required types of compensation modules; (ii) it tunes to adapt to routing path changes in a reconfigurable network, (iii) it tracks dynamic changes in dispersion due to environment, and (iv) it achieves a high degree of accuracy necessary for 40 Gb/s channels (see Figure 23).

One brute-force method to achieve tunable dispersion compensation is to build a module with optical switches used to add or remove sections of fixed DCF to achieve a discrete set of dispersion compensation. Many other elegant, yet viable, approaches have been developed for tunable dispersion compensation, and these approaches include linearly chirped FBG with nonuniform heating, nonlinearly chirped Fiber Bragg grating (FBG) with a simple mechanical stretcher, virtually imaged phase array, electronic tap delay filter with weights, multiple stage all pass filter, etc.

System Performance Parameters

An eye diagram provides a simple and useful tool to visualize intersymbol interference between data bits. Figure 24a shows a perfect eye diagram. A square bit stream (i.e., series of symbol ‘1’s and ‘0’s) is sliced into sub-bit stream with predetermined eye intervals (i.e., several bit periods), and displayed through bit analyzing equipment (e.g., digital channel analyzer), overlapping the sliced sub-bit stream in order to obtain the eye diagram. When a perfect transmitter and receiver (i.e., infinite receiver bandwidth with zero noise characteristics and jitter), and a perfect transmission media (i.e., no dispersion and nonlinearites) are used, the received eye diagram is shaped as a perfect rectangular. In reality, the transmitter and receiver have a limited bandwidth with noise and jitter, and the transmission media (i.e., optical fiber) has dispersion and nonlinearites. Therefore, the eye diagram deviates from the perfect rectangular shape. Figure 24b shows the eye diagram close to a real situation. The shape of the eye is generally broadened and distorted (i.e., eye is closed) due to limited bandwidth and fiber impairments, and noise and timing jitter are added onto this broadened and distorted eye shape.

The Q-factor (q) is also an important system parameter widely used in long-distance optical transmission system design. It is defined as the electrical signal-to-noise ratio before the decision circuit at receiver. The Q-factor is directly related to bit-error-rate (BER: the percentage of bits that has an error relative to total number of bits received in a specific time) by: $\mathrm{BER} = 0.5\cdot\mathrm{erfc}(q/2^{0.5})$, where $\mathrm{erfc}(x)$ is the complementary error function. The Q-factor is an unitless linear ratio, and is express in dB by $20\cdot\log(q)$. As a conventional relationship, Q-factor of 15.6 dB (i.e., linear ratio 6) is required to achieve $\mathrm{BER} = 10^{-9}$. Figure 25 shows the relationship of eye diagram, Q-factor, and BER.

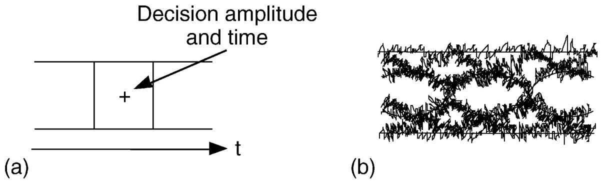

Figure 24 (a) A perfect square eye diagram, (b) a closed eye diagram due to bandwidth, fiber impairments, noise, and timing jitter.

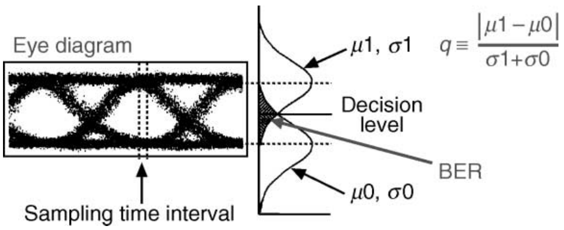

Figure 25 The relationship of eye diagram, Q-factor, and BER ($\mu1$, $\mu2$: mean levels of logic ‘1’ and ‘0’, $\sigma1$, $\sigma2$: standard deviation of logic ‘1’ and ‘0’).

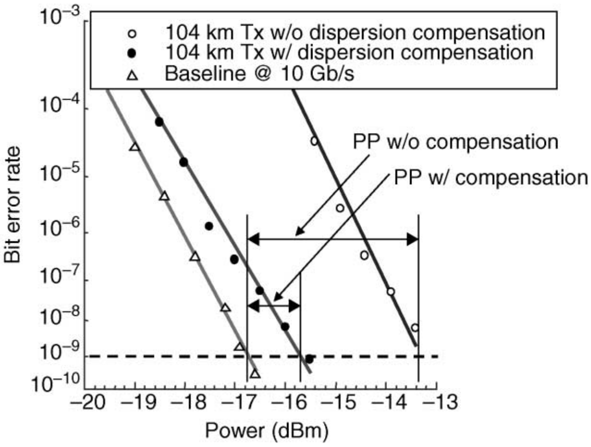

Figure 26 BER curve and power penalty.

Power penalty (PP) is one of the most important system parameters, and is defined as the received optical power difference in dB with and without signal impairments at a specified BER (conventionally 10^{-9}), from measured BER versus optical power curve. Therefore, 1 dB PP means that a system with signal impairments requires 1 dB more optical power at the receiver in order to achieve the same BER performance compared to the system without signal impairments. Figure 26 shows the typical BER curve as a function of received optical power and PP. Note that PP reduces from ~3.5 dB to ~1 dB with dispersion compensation.

In long-haul and undersea transmission system with many EDFA chains, OSNR is an important system parameter to design and characterize the optical transport system. OSNR is defined as the ratio of the power of signal channel to the power of ASE in a specified optical bandwidth (conventionally 0.1 nm). The OSNR (in dB) of a signal channel at the end of the system is approximated by: $\text{OSNR (dB)} = 58 + P_{\text{out}} - L_{\text{span}} - \text{NF} - 10\log(N_{\text{amp}})$, when the system has N_{amp} fiber spans, span loss L_{span} (in dB) followed by an optical amplifier with output power P_{out} (in dBm) per channel launched into the span and noise figure (NF) (in dB). As an example, a typical OSNR requirement for $\text{BER} = 10^{-9}$ is about 17 dB at 10 Gb/s data rate. The required OSNR should be increased by 6 dB when the data rate is increased by a factor of 4. Improving the amplifier's NF can increase OSNR, and the improved OSNR can be used to increase the system reach, reducing the channel power in order to suppress the nonlinearities, etc.

See also

Fiber and Guided Wave Optics: Light Propagation. **Imaging:** Multiplex Imaging. **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Leonid K, Sergio B and Alan W (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House Inc.

Govind PA (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Emmanuel D (1994) *Erbium-Doped Fiber Amplifiers*. New York: Wiley.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV A Components*. San Diego, CA: Academic Press.

Ivan K and Tingye L (2002) *Optical Fiber Telecommunications IV B Systems and Impairments*. San Diego, CA: Academic Press.

Historical Development

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

This article describes the development of optical fiber communication systems. After describing some of the motivations for using optical fiber communications and the advantages of this technology, the key milestones and the principal people involved in developing optical fibers and compatible light sources are presented. Following this, the article looks at the evolution of fielded systems and the use of optical fiber links in undersea applications.

One of the principal needs of people since antiquity has been to communicate. This need created interests in devising communication systems for sending messages from one distant place to another. Many forms of such systems have appeared over the years. The basic motivations behind each new one were either to improve the transmission fidelity, to increase the data rate so that more information could be sent, or to increase the transmission distance between relay stations. Prior to the nineteenth century, all communication systems operated at a very low information rate and involved only optical or acoustical means, such as signal lamps or horns. One of the earliest known optical transmission links, for example, was the use of a fire signal by the Greeks in the eighth century BC for sending alarms, calls for help, or announcements of special events.

The invention of the telegraph by Samuel FB Morse in 1838 ushered in the era of electrical communications. In the ensuing years an increasingly larger portion of the electromagnetic spectrum, shown in Figure 1, was utilized for the conveying of information from one place to another. The reason for this trend is that, in electrical systems, the data usually are transferred over the communication channel by superimposing the information onto a sinusoidally varying electromagnetic wave, which is known as the carrier. When reaching its destination the information is removed from the carrier wave and processed as desired. Since the amount of information that can be transmitted is related directly to the frequency range over which the carrier operates, increasing the carrier frequency theoretically increases the available transmission bandwidth and, consequently, provides a larger information capacity. Thus the trend in electrical communication system developments was to employ progressively higher frequencies (shorter wavelengths), which offer corresponding increases in bandwidth or information capacity. This activity led to the birth of communication mechanisms such as radio, television, microwave, and satellite links.

Another important portion of the electromagnetic spectrum encompasses the optical region shown in Figure 1. In contrast to electrical communications, transmission of information in an optical format is carried out, not by frequency modulation of the carrier but by varying the intensity of the optical power. Similar to the radio-frequency spectrum, two classes of transmission medium can be used: an atmospheric channel or a guided-wave channel. Whereas transmission of optical signals through the atmosphere was done thousands of years ago, the use of a guided-wave optical channel such as an optical fiber is a fairly recent application.

Fiber optic communication systems have a number of inherent advantages over their copper-based and radio-transmission counterparts. Fiber optic cable can transmit at a higher capacity, thereby reducing the number of physical lines and the amount of equipment needed for a given transmission span. The lower weight and smaller size of optical fibers offer a distinct advantage over heavy, bulky copper cables in crowded underground cable ducts, in ceiling-mounted cable trays, and in mobile platforms such as aircraft and ships. Their dielectric composition is an especially important feature of optical fibers, since this ensures freedom from electromagnetic interference between adjacent fibers, eliminates ground loops, and results in extremely low fiber-to-fiber crosstalk. In addition, an optical fiber affords a high degree of data security since the optical signal is well confined within the optical waveguide.

Optical Fiber Development

Since an atmospheric channel requires a line-of-sight link and because it can be adversely affected by weather conditions, a guided-wave channel is the preferred approach in most cases for communication system applications. A challenge in using an optical fiber channel is to have a flexible, low-loss medium that transfers the optical signal over long distances

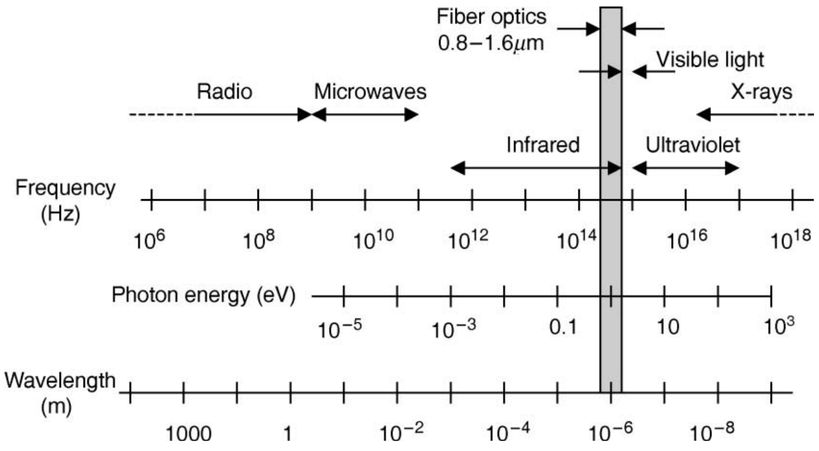

Figure 1 Electromagnetic spectrum for electrical and optical communications.

without significant attenuation and distortion. Glass is an obvious material for such applications. The earliest known glass was made around 2500 BC and glass already was drawn into fibers during the time of the Roman Empire. However, such glasses have very high attenuations and thus are not suitable for communication applications. One of the first known attempts of using optical fibers for communication purposes was a demonstration in 1930 by Heinrich Lamm of image transmission through a short bundle of optical fibers for potential medical imaging. However, no further work was done beyond the demonstration phase since the technology for producing reasonably low-loss fibers with good light confinement was not yet mature.

Further work and experiments continued on using optical fibers for image transmission and by 1960 glass-clad fibers had attenuations of about one decibel per meter. This attenuation allowed fibers to be used for medical imaging, but it was still much too high for communications. Optical fibers had attracted the attention of researchers at that time, because they were analogous in theory to plastic dielectric waveguides used in certain microwave applications. In 1961, Elias Snitzer demonstrated this similarity by drawing fibers with cores so small they carried light in only one waveguide mode. He published a classic theoretical description of single-mode fibers in May 1961. However, to be useful for communication systems, optical fibers would need to have a loss of no more than 10 or 20 decibels per kilometer.

As a reminder, decibels measure the ratio of the output power to the input power on a logarithmic scale. The abbreviation for a decibel is ‘dB.’ The power ratio in decibels is given by the expression ‘10 log (power out/power in).’ As an example, a power loss of 20 dB over a 1-km distance in an optical fiber means that only 1% of the light injected into the fiber comes out of the other end. Table 1 gives some typical examples of various power ratios and their decibel equivalents.

In the early 1960s Charles Kao pursued the idea of using a clad glass fiber for an optical waveguide, building on optical waveguide research being done at the Standard Telecommunication Laboratories in England. After he and George Hockman painstakingly examined the transparency properties of various types of glass, Kao made a prediction, in 1966, that losses of no more than 20 dB/km were possible in optical fibers. In July 1966, Kao and Hockman presented a detailed analysis for achieving such a loss level. Kao then went on to actively advocate and promote the prospects of fiber communications, which generated interest in laboratories around the world to reduce fiber loss. It took four years to reach Kao’s predicted goal of 20 dB/km, and the final solution was different from what many had expected.

Table 1 Examples of some optical power ratios and their decibel equivalents

Power ratio (P_{out}/P_{in})	*Decibel equivalent (dB)*
0.01	−20
0.10	−10
0.50	−3
1	0
2	3
10	10
100	20

To understand the process of making a fiber, consider the schematic of a typical fiber structure shown in Figure 2. A fiber consists of a solid dielectric (glass or plastic) cylinder of refractive index n_1 called the core. This is surrounded by a dielectric cladding which has a refractive index n_2, that is less than n_1, in order to achieve light guiding in the fiber. This structure is then encapsulated by a buffer coating to protect the fiber from abrasion and outside contaminants. The first step in making a fiber is to form the clear glass rod or tube called a preform. This normally is done by a vapor-phase oxidation process. The preform has two distinct regions that correspond to the core and cladding of the eventual fiber. Fibers are made from the preform by precision feeding it into a circular furnace that softens the end of the preform to the point where it can be drawn into a very thin filament which becomes the optical fiber.

Most researchers had tried to purify the compound glasses used for standard optics, which are easy to melt and draw into fibers. A different approach was taken at the Corning Glass Works where Robert Maurer, Donald Keck, and Peter Schultz started with fused silica. This material can be made extremely pure, but has a high melting point and a low refractive index. Silica has the approximate attenuation versus wavelength characteristic shown in Figure 3. Note that for early silica fibers there are regions of low attenuation around 850, 1310, and 1550 nm, which the literature refers to as the first, second, and third windows, respectively. The large attenuation spikes in the 1000- and 1400-nm spectral

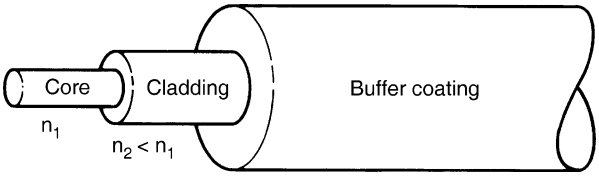

Figure 2 Typical optical fiber structure.

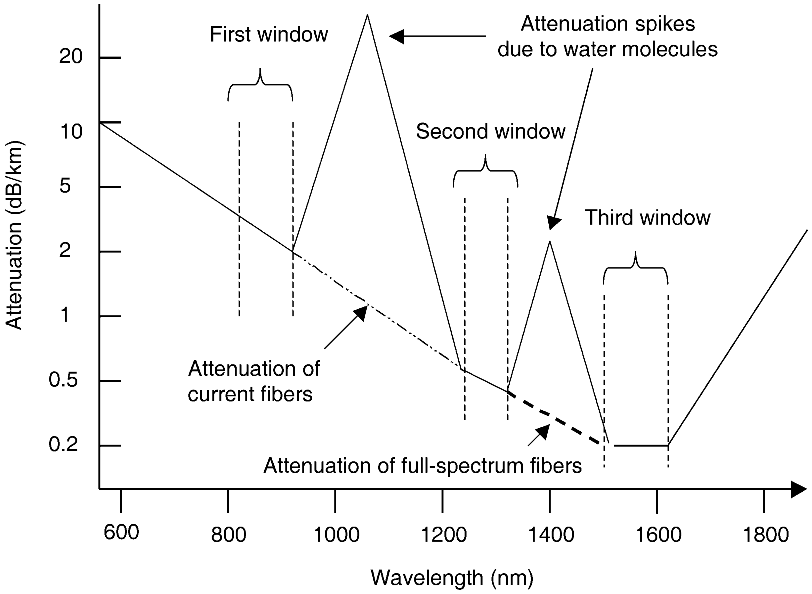

Figure 3 Attenuation versus wavelength approximation for silica fibers.

regions are due to absorption by residual water molecules in the glass. Although ultra-pure material processing techniques can eliminate these spikes to produce what are known as full-spectrum or low-water-peak fibers, most installed fibers still have a relatively large attenuation between 1350 and 1500 nm.

Note that since the attenuation spikes no longer are present in various types of fibers, the idea of operating windows has been replaced by the concept of spectral bands. The International Telecommunications Union (ITU) has designated six spectral bands for use in intermediate-range and long-distance optical fiber communications within the 1260 to 1675-nm region. The regions, which are known by the letters O, E, S, C, L, and U, are defined as follows:

- Original Band (O-Band): 1260 to 1360 nm
- Extended Band (E-Band): 1360 to 1460 nm
- Short Band (S-Band): 1460 to 1530 nm
- Conventional Band (C-Band): 1530 to 1565 nm
- Long Band (L-Band): 1565 to 1625 nm
- Ultra-long Band (U-Band): 1625 to 1675 nm

The Corning team made cylindrical preforms by depositing purified materials from the vapor phase. To produce a fiber that has light guiding properties they carefully added controlled trace levels of titanium to the core to make its refractive index slightly higher than that of the cladding without raising the attenuation significantly. In September 1970, they announced the fabrication of single-mode fibers with an attenuation of 17dB/km at the 633-nm helium-neon line. The fibers were fragile, but independent tests at the British Post Office Research Laboratories facility in Martlesham Heath, England confirmed the low loss.

This dramatic breakthrough was the first among the many developments that opened the door to fiber optic communications. The ensuing years saw further reductions in optical fiber attenuations. By mid-1972, Maurer, Keck, and Schultz at Corning had made multimode germania-doped fibers with a 4-dB/km loss and much greater strength than the earlier brittle titania-doped fibers. In order to couple a sufficient amount of optical power into a fiber, early applications used multimode fibers with a refractive-index gradient between core and cladding of around 2% and core diameters of 50 or 62.5 micrometers.

Single-mode fibers have much smaller core diameters on the order of 9 micrometers in order to allow only one propagation mode. This type of fiber has a much higher transmission capacity since the effect of modal dispersion is eliminated. The first single-mode fibers were optimized to have a zero dispersion value at 1310 nm, since the silica material used at that time exhibited a low loss within a spectral band around this wavelength. Figure 4 shows the dispersion characteristic of this type of fiber as a function of wavelength in the S-, C-, and L-bands. As a result of its widespread use in early single-mode transmission systems, this fiber design has been standardized by the International Telecommunication Union (ITU-T) under the designation Recommendation G.652.

Standard G.652 silica fibers provide the lowest attenuation at 1550 nm, but have a much larger signal

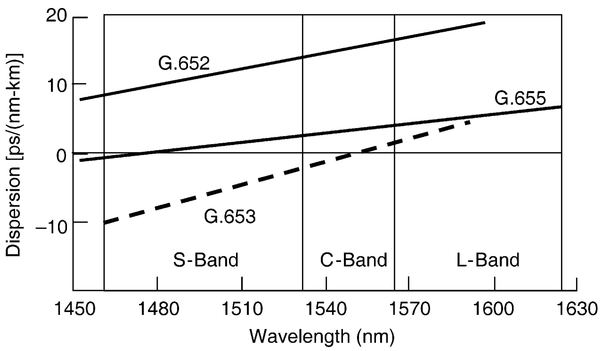

Figure 4 Dispersion characteristics of three types of standard optical fibers.

dispersion at this wavelength than at 1310 nm. Since system designers wanted to use fibers at a 1550-nm wavelength, in order to transmit high-speed data over longer distances, fiber manufacturers overcame the larger signal dispersion limitation by creating the so-called dispersion-shifted fibers. This was done through a clever manipulation of the core and cladding designs that allowed the zero-dispersion point to shift to longer wavelengths. In particular, the ITU-T created a specification for operation at 1550 nm which is designated as recommendation G.653. Figure 4 also shows the dispersion characteristic of this type of fiber as a function of wavelength.

Although the G.652 and G.653 fibers work well for single-wavelength operation, a different type of fiber, having non-zero dispersion within a broad spectral range, is needed when implementing systems that use many independent light sources simultaneously within a particular wavelength band. This led to the specification of the non-zero dispersion shifted fiber (NZDSF) in the ITU-T Recommendation G.655 that are designated to operate around 1550 nm. The main purpose of having a positive dispersion value over the entire operating spectrum is to mitigate a nonlinear optical effect called four-wave mixing (FWM), which is analogous to intermodulation distortion in electrical systems. Figure 4 shows the dispersion characteristic of the G.655 fiber in the S-, C-, and L-bands.

Light Sources

The fiber by itself is not practical unless there is a compatible optical source for launching light signals into it. The most suitable device for this is a semiconductor light-emitting diode (LED) or a laser diode. In the 1960s, a great deal of effort took place to achieve laser action in pn-junction diodes. The early devices were GaAs and GaAsP lasers that operated at a temperature of 77 K, emitted at a wavelength around 850 nm, and had high lasing threshold current densities. To make devices that were more application-friendly by operating at room temperature, structures consisting of sandwiched layers of AlGaAs and GaAs were investigated. Finally, in 1970, researchers at Bell Laboratories and a team at the Ioffe Physical Institute in Leningrad made the first semiconductor diode lasers based on a layered AlGaAs/GaAs/AlGaAs structure that were able to emit continuous-wave light in the 850-nm region at room temperature.

Major improvements in laser diode performance and reliability followed this achievement during the next decade. In addition, around 1976, room-temperature laser diodes operating at longer wavelengths, in the 1100 to 1600-nm region, started to appear. Of particular interest were GaInAsP/InP-based laser diodes emitting in the 1310-nm and 1550-nm low-loss windows of optical fibers. The progressive development of ever-improving devices during the 1980s and 1990s, included single-frequency emission with narrow linewidths under continuous operation, low levels of chirp under direct modulation, high output power, and the ability to tune specially constructed laser diodes over a wavelength range of up to 30 nm.

Fielded Systems

The bit rate-distance product B × L, where B is the transmission bit rate and L is the repeater spacing, measures the transmission capacity of optical fiber links. Since the inception of optical fiber communications in the mid-1970s, the link transmission capacity has experienced a tenfold increase every four years. Several major technology advances spurred this growth. Among the technology developments are laser diodes emitting over an extremely narrow spectral band, optical amplifiers, fibers with low losses and low dispersions, and the concepts of wavelength division multiplexing.

Some of the initial telephone-system field trials in the USA were carried out in 1977 by GTE in Los Angeles and by AT&T in Chicago. These transmission links were largely for the trunking of telephone lines, which are digital links consisting of time-division-multiplexed 64-kb/s voice channels. Similar demonstrations were carried out in Europe and Japan. Applications ranged from 45 to 140 Mb/s with repeater spacings around 10 km.

With the advent of high-capacity fiber optic transmission lines in the 1980s, service providers established a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. These standards define a synchronous frame structure for sending multiplexed digital traffic over optical fiber trunk lines. The basic building block and first level of the SONET signal hierarchy is called the synchronous transport signal–level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an optical carrier–level N (OC-N) signal. Thus the OC-N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems, the fundamental building block is the 155.52-Mb/s synchronous transport module–level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM-N signal. Table 2 shows commonly used SDH and SONET signal levels.

The development of optical sources and photodetectors capable of operating at 1310 nm allowed a shift in the transmission wavelength from 850 nm to 1310 nm. This resulted in a substantial increase in the repeaterless transmission distance for long-haul telephone trunks, since optical fibers exhibit lower power loss and less signal dispersion at 1310 nm. Intercity applications first used multimode fibers, but in 1984 switched exclusively to single-mode fibers, which have a significantly larger bandwidth. Bit rates for long-haul links typically range between 155 and 622 Mb/s (OC-3 and OC-12), and in some cases up to 2.5 Gb/s (OC-48), over repeater spacings of 40 km. Both multimode and single-mode 1310-nm fibers are used in local area networks, where bit rates range from 10 Mb/s to 100 Mb/s over distances ranging from 500 m to tens of kilometers.

In the next step of system evolution, links operating in the low-loss window around 1550 nm attracted much attention for high-capacity, long-span terrestrial and undersea transmission links. These links routinely carry traffic at 2.5 Gb/s over 90-km repeaterless distances. By 1996, advances in high-quality lasers and receivers allowed single-wavelength transmission rates of 10 Gb/s (OC-192).

In 1989, the introduction of optical amplifiers gave a major boost to fiber transmission capacity. Although there are GaAlAs-based solid-state optical amplifiers for the first window and InGaAsP amplifiers for the second window, the most successful and widely used devices are erbium-doped fiber amplifiers (commonly called EDFAs) operating in the 1530 to 1560-nm range and Raman fiber amplifiers that are used for operation in the 1560 to 1600-nm region.

During the same time period, impressive demonstrations of long-distance high-capacity systems were made using optical soliton signals. A soliton is a nondispersive pulse that makes use of nonlinear dispersion properties in a fiber to cancel out chromatic dispersion effects. As an example, solitons at rates of 10 Gb/s have been sent over a 12 200-km experimental link using optical amplifiers and special modulation techniques.

Entrance of Wavelength Division Multiplexing

The use of wavelength division multiplexing (WDM) offers a further boost in fiber transmission capacity. The basis of WDM is to use multiple sources operating at slightly different wavelengths to transmit several independent information streams over the same fiber. Although researchers started looking at WDM in the 1970s, during the ensuing years it generally turned out to be easier to implement higher-speed electronic and optical devices than to invoke the greater system complexity called for in WDM. However, a dramatic surge in WDM popularity started in the early 1990s, as electronic devices neared their modulation limit and high-speed equipment became increasingly complex.

Figure 5 shows the concept of implementing many closely spaced wavelengths within a spectral band centered around 1552.524 nm. This scheme is referred to as dense WDM or DWDM. Conceptually, the DWDM scheme is the same as frequency

Table 2 Commonly used SONET and SDH transmission rates

SONET level	*Electrical level*	*Line rate (Mb/s)*	*SDH equivalent*	*Common rate name*
OC-1	STS-1	51.84	–	
OC-3	STS-3	155.52	STM-1	155 Mb/s
OC-12	STS-12	622.08	STM-4	622 Mb/s
OC-24	STS-24	1244.16	STM-8	
OC-48	STS-48	2488.32	STM-16	2.5 Gb/s
OC-96	STS-96	4976.64	STM-32	
OC-192	STS-192	9953.28	STM-64	10 Gb/s
OC-768	STS-768	39,813.12	STM-256	40 Gb/s

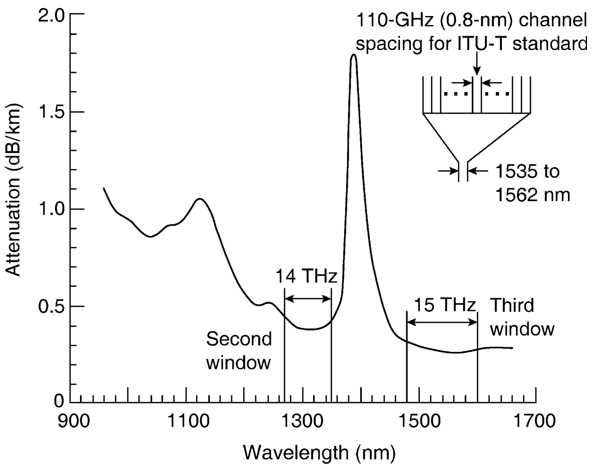

Figure 5 Wavelength division multiplexing (WDM) concept.

division multiplexing (FDM) used in microwave radio and satellite systems. Just as in FDM, the wavelengths (or optical frequencies) in a DWDM link must be properly spaced to avoid interference between channels. In an optical system this interference may arise from the fact that the center wavelength of laser diode sources and the spectral operating characteristics of other optical components in the link may drift with temperature and time, thereby giving rise to the need for a guard band between wavelength channels.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the ITU developed DWDM standards that specify channel spacings in terms of frequency. The ITU-T Recommendation G.694.1, which is entitled 'Dense Wavelength Division Multiplexing (DWDM),' specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm). This implementation requires the use of stable, high-quality, temperature-controlled and wavelength-controlled (frequency-locked) laser diode light sources.

With the production of full-spectrum (low-water-content) fibers, the development of relatively inexpensive optical sources, and the desire to have low-cost optical links operating in metro and local area networks, came the concept of coarse WDM (CWDM). In 2002, the ITU-T released a standard aimed specifically at CWDM. This is Recommendation G.694.2, which is entitled 'Coarse Wavelength Division Multiplexing (CWDM).' The CWDM grid is made up of 18 wavelengths defined within the range 1270 nm to 1610 nm (O-through L-bands) spaced by 20 nm with wavelength-drift tolerances of ± 2 nm. This can be achieved with inexpensive light sources that are not temperature-controlled. The targeted transmission distance for CWDM is 50 km on single-mode fibers, such as those specified in ITU-T Recommendations G.652, G.653, and G.655.

Wavelength tunability of a source is an important property of WDM systems. Obviously it is not desirable or practical to maintain an inventory of dozens of lasers that emit at different wavelengths for WDM applications. The ideal tunable laser should be adjustable to emit at a specific wavelength across a broad spectral range. One such device is a distributed Bragg reflector (DBR) laser diode that can be tuned over a 10 to 20 nm spectral range. Work on perfecting such devices are still underway.

Starting in the mid-1990s, a combination of EDFAs and WDM was used to boost fiber information capacity to even higher levels and to increase the transmission distance. A major system consideration in these super-high capacity links is to ensure that there is appropriate link and equipment redundancy, so that alternate paths are available in case of disruptions in communications resulting from cable ruptures (for example, caused by errant digging from a backhoe) or equipment failures at an intermediate node. Such disruptions otherwise could have a devastating effect on a large group of users.

Undersea Optical Cable Systems

The first transoceanic fiber optic cable systems were installed in the Atlantic and Pacific Oceans in 1988 and 1989. Initially these systems operated at 280 Mb/s per fiber pair using 1310-nm lasers and single-mode fibers. The links consisted of a series of point-to-point optical fiber segments between electronic-based undersea regeneration points that were located nominally 60 km apart. Later the transmission capacity of these links was upgraded to 2.5 Gb/s and the regenerator spacing was increased to 100 km by converting the 1310-nm multiple-frequency light sources to 1550-nm single-frequency laser diodes. Later, the regenerator spacing was increased to 140 km.

Although these cable systems significantly improved the quality of the international telephone service, the optical-to-electrical conversion process at each regeneration point remained a capacity bottleneck. The introduction of erbium-doped optical fiber amplifiers (EDFA) eliminated this bottleneck from undersea lightwave systems by amplifying signals directly in the optical domain. Since EDFAs operate over a 30-nm wavelength band, they are well suited for use with undersea WDM links, which have provided a further capacity increase. These undersea optical amplifiers are typically spaced about 45 km apart.

One example of the many worldwide installations of optically amplified WDM networks is the SEA-ME-WE-3 Cable System. This undersea network runs from Germany to Singapore, connecting more than a dozen countries in between. Hence the name SEA-ME-WE, which refers to Southeast Asia (SEA), the Middle East (ME), and Western Europe (WE). The network has two pairs of undersea fibers with a capacity of eight STM-16 wavelengths per fiber.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Wavelength Division Multiplexing.

Further Reading

Bergano NS (1997) Undersea amplified lightwave system design. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 302–335. New York: Academic.

Goralski WJ (2002) SONET/SDH, 3rd edn. New York: McGraw-Hill.

Hecht J (1999) *City of Light*. New York: Oxford University Press.

Joyner CH (1997) Semiconductor laser growth and fabrication technology. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIB, pp. 163–199. New York: Academic.

Kaiser P and Keck DB (1988) Fiber types and their status. In: Miller SE and Kaminow IP (eds) *Optical Fiber Telecommunications II*, pp. 29–54. New York: Academic.

Kao KC and Hockman GA (July 1966) Dielectric-fibre surface waveguides for optical frequencies. *Proceedings IEE*, vol. 113, 1151–1158.

Keiser G (2003) *Optical Communications Essentials*. New York: McGraw-Hill.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit rate long-distance transmission. In: Kaminow IP and Koch TL (eds) *Optical Fiber Telecommunications*, vol. IIIA, pp. 373–460. New York: Academic.

Nyman B, Farries M and Si C (Oct 2001) Technology trends in dense WDM demultiplexers. *Optical Fiber Technology* 7(4): 255–274.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco, CA: Morgan Kaufmann.

Snitzer E (May 1961) Cylindrical dielectric waveguide modes. *Journal Optical Society of America* 51: 491–498.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communication Magazine* 36(2): 62–66.

Architectures of Optical Fiber Communication Systems

G Keiser, PhotonicsComm Solutions, Inc.,
Newton Center, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Simple Optical Fiber Links

The most basic optical fiber communication system architecture is a point-to-point link consisting of the elements shown in Figure 1. The key sections are a transmitter consisting of a light source and its associated drive circuitry, a cable offering mechanical and environmental protection to the optical fibers contained inside, and a receiver consisting of a photodetector plus amplification and signal-restoring circuitry. Very long links may include periodic optical amplifiers for boosting the level of the optical signal. In networks that use only simple point-to-point links, the optical fiber is used principally as a transmission medium and all the switching and processing of information bits is done by electronics.

At the transmitting end of an optical fiber link, information bits consisting of electronic pulses are used to modulate the light output of the source. The resulting light pulses are coupled into an optical fiber. As they travel along the fiber these signal pulses become progressively weakened and distorted due to attenuation and dispersion mechanism that are characteristic of an optical fiber. At the end of the fiber link a photodetector converts the optical pulses back to an electrical format for switching to another link or for processing at the end station.

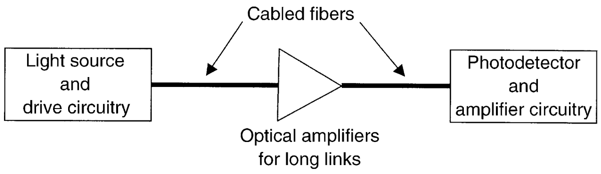
Cabled fibers
Light source and drive circuitry
Photodetector and amplifier circuitry
Optical amplifiers for long links

Figure 1 Basic elements of a point-to-point fiber optic link.

The cabled fiber is one of the most important elements in an optical fiber link, since it determines operational parameters such as transmission distance and achievable data rates. In addition to protecting the glass fibers during installation and service, the cable may contain copper wires for powering optical amplifiers or signal regenerators, which are needed periodically in long-distance links for amplifying and reshaping the signal.

Analogous to copper cables, the installation of optical fiber cables can be either aerial, in ducts, undersea, or buried directly in the ground. As a result of installation and/or manufacturing limitations, individual cable lengths will range from several hundred meters to several kilometers. Practical considerations such as reel size and cable weight determine the actual length of a single cable section. The shorter segments tend to be used when the cables are pulled through ducts. Longer lengths are used in aerial, direct-burial, or undersea applications. Splicing together individual cable sections forms continuous transmission lines for these long-distance links. For undersea installations, the splicing and repeater-installation functions are carried out on board a specially designed cable-laying ship.

One of the principal characteristics of an optical fiber is its attenuation as a function of wavelength, as shown in Figure 2. Early technology used the 800 to 900 nm wavelength band, since in this region the fibers made at that time exhibited a local minimum in the attenuation curve, and optical sources and photodetectors operating at these wavelengths were available. This region originally was referred to as the ‘first window’. By reducing the concentration of hydroxyl ions and metallic impurities in the fiber material, in the 1980s manufacturers were able to fabricate optical fibers with very low loss in the 1100 to 1600 nm region. Two windows are defined here, these being the ‘second window’ centered around 1310 nm and the ‘third window’ centered around 1550 nm.

In 1998 a new ultra-high purifying process patented by Lucent Technologies eliminated virtually all water molecules from the glass material. The resulting fiber is designated as either a low-water-peak or a full spectrum fiber. By dramatically reducing the water-attenuation peak around 1400 nm, this process opens the transmission region between the second and third windows to provide around 100 nm more bandwidth than in conventional single-mode fibers. This particular fiber allows wider spectral tolerances on optical components, which reduces costs for moderate-distance metropolitan network applications.

Since the attenuation of low-water-peak fibers makes the designation of transmission windows obsolete, the concept of operational spectral bands arose for the 1260–1675 nm region. As shown in

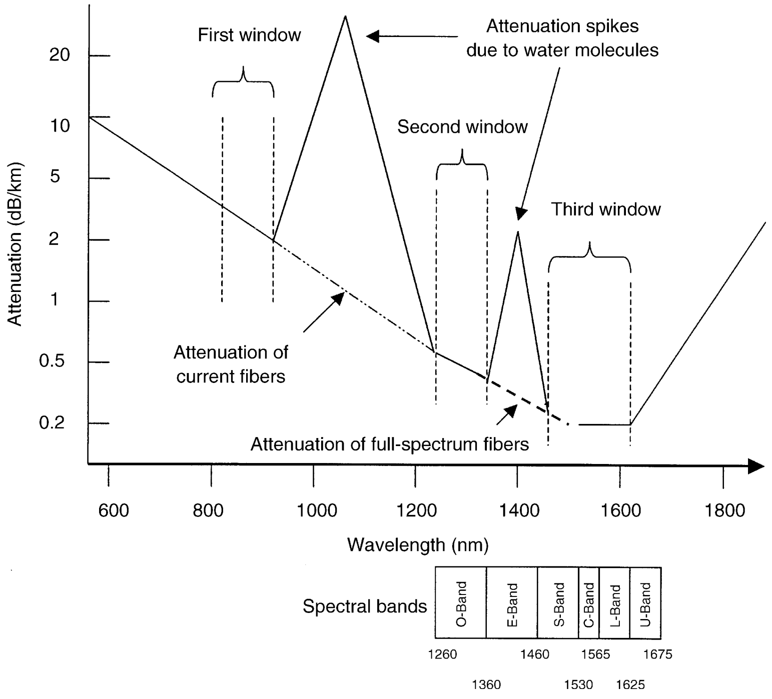

Figure 2 Attenuation versus wavelength for silica fibers and the concept of spectral bands.

Figure 2, the regions are known by the letters O, E, S, C, L, and U, which are defined as follows:

- Original band (O-Band): 1260–1360 nm;
- Extended band (E-Band): 1360–1460 nm;
- Short band (S-Band): 1460–1530 nm;
- Conventional band (C-Band): 1530–1565 nm;
- Long band (L-Band) 1565–1625 nm;
- Ultra-long band (U-Band): 1625–1675 nm.

Once the cable is installed, a light source that is dimensionally compatible with the fiber core is used to launch optical power into the fiber. Semiconductor light-emitting diodes (LEDs) and laser diodes are suitable for this purpose, since their light output can be modulated rapidly by simply varying the bias current at the desired transmission rate, thereby producing an optical signal. The electric input signals to the transmitter circuitry for the optical source can be either of an analog or digital form. For high-rate systems (nominally greater than 1 Gb/s), direct modulation of the source can lead to unacceptable signal distortion. In this case, an external modulator is used to vary the amplitude of a continuous light output from a laser diode source.

After an optical signal is launched into a fiber, it will become progressively attenuated and distorted with increasing distance because of scattering, absorption, and dispersion mechanisms in the glass material. At the receiver a photodiode will detect the weakened optical signal emerging from the fiber end and convert it to an electric current. The design of an optical receiver is inherently more complex than that of the transmitter, since it has to interpret the content of the weakened and degraded signal received by the photodetector. The principal figure of merit for a receiver is the maximum optical power necessary at the desired data rate to attain either a given error probability for digital systems or a specified signal-to-noise ratio for an analog system. The ability of a receiver to achieve a certain performance level depends on the photodetector type, the effects of noise in the system, and the characteristics of the successive amplification stages in the receiver.

Networks of Links

With the advent of fiber optic transmission lines, the next step in the evolution of the digital multiplexing scheme used by telecommunication providers was a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. In the mid-1980s, efforts started on developing a standard that would allow network engineers to interconnect fiber optic transmission equipment from various vendors through multiple-owner trunk networks. This resulted in a series of standards for SONET by the American National Standards Institute (ANSI) and a series of recommendations for SDH from the International Telecommunications Union (ITU). Examples of particular interest are the ANSI T1.105.06 standard and the ITU-T G.957 recommendation. Although there are some implementation differences between SONET and SDH, all SONET specifications conform to the SDH recommendations. A key characteristic of SONET and SDH is that they are usually configured as a ring architecture. This is done to create loop diversity for uninterrupted service protection purposes in case of link or equipment failures. The SONET/SDH rings commonly are called self-healing rings, since the traffic flowing along a certain path can be switched automatically to an alternate or standby path following failure or degradation of the link segment.

The basic building block and first level of the SONET signal hierarchy is called the Synchronous Transport Signal – Level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving *N* STS-1 frames, which are then scrambled and converted to an Optical Carrier – Level *N* (OC-N) signal. Thus the OC-N signal will have a line rate exactly *N* times that of an OC-1 signal. For SDH systems the fundamental building block is the 155.52 Mb/s Synchronous Transport Module – Level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing *N* different STM-1 signals to form the STM-N signal.

Three main features, each with two alternatives, classify all SONET/SDH rings, thus yielding eight possible combinations of ring types. First, there can be either two or four fibers running between the nodes on a ring. Second, the operating signals can travel either clockwise only (which is termed a unidirectional ring) or in both directions around the ring (which is called a bidirectional ring) Third, protection switching can be performed either via a line-switching or a path-switching scheme. Upon link failure or degradation, line switching moves all signal channels of an entire OC-N channel to a protection fiber. Conversely, path switching can move individual payload channels within an OC-N channel (e.g., an STS-1 channel within an OC-12 channel) to another path.

Of the eight possible combinations of ring types, the following two architectures have become popular for SONET and SDH networks:

- two-fiber unidirectional path-switched ring (two-fiber UPSR);
- two-fiber or four-fiber bidirectional line-switched ring (two-fiber or four-fiber BLSR).

The common abbreviations of these configurations are given in parentheses. They also are referred to as unidirectional or bidirectional self-healing rings (USHR or BSHR).

Figure 3 shows a two-fiber unidirectional path-switched ring network. By convention, in a unidirectional ring the normal working traffic travels clockwise around the ring, as indicated by the heavy arrows. For example, the connection from node 1 to node 3 uses links 1 and 2, whereas the traffic from node 3 to node 1 traverses links 3 and 4. Thus, two communicating nodes use a specific bandwidth capacity around the entire perimeter of the ring. If nodes 1 and 3 exchange information at an OC-3 rate in an OC-12 ring, then they use one-fourth of the capacity around the ring on all the primary links. In a unidirectional ring the counter-clockwise path is used as an alternate path for protection against link or node failures. To achieve this, the signal from a transmitting node is dual-fed into both the primary and protection fibers. This establishes a designated protection path on which traffic flows counter-clockwise, i.e., from node 1 to node 3 via links 5 and 6 (in that order), as shown in Figure 3b. A heavy line and a dashed line indicate the primary and protection paths, respectively.

Consequently, two identical signals from a particular node arrive at their destination from opposite directions, usually with different delays, as denoted in Figure 3b. The receiver normally selects the signal from the primary path. However, it continuously compares the fidelity of each signal and chooses the alternate signal in case of severe degradation or loss of the primary signal. Thus, each path is individually switched based on the quality of the received signal. For example, if path 2 breaks or equipment in node 2 fails, then node 3 will switch to the protection channel to receive signals from node 1.

Figure 4 illustrates the architecture of a four-fiber bidirectional line-switched ring. Here two primary fiber loops (with fiber segments labeled 1p through 8p) are used for normal bidirectional communication, and the other two secondary fiber loops are standby links for protection purposes (with fiber segments labeled 1s through 8s). In contrast to the two-fiber UPSR, the four-fiber BLSR has a capacity advantage since it uses twice as much fiber cabling and because traffic between two nodes is only sent partially around the ring.

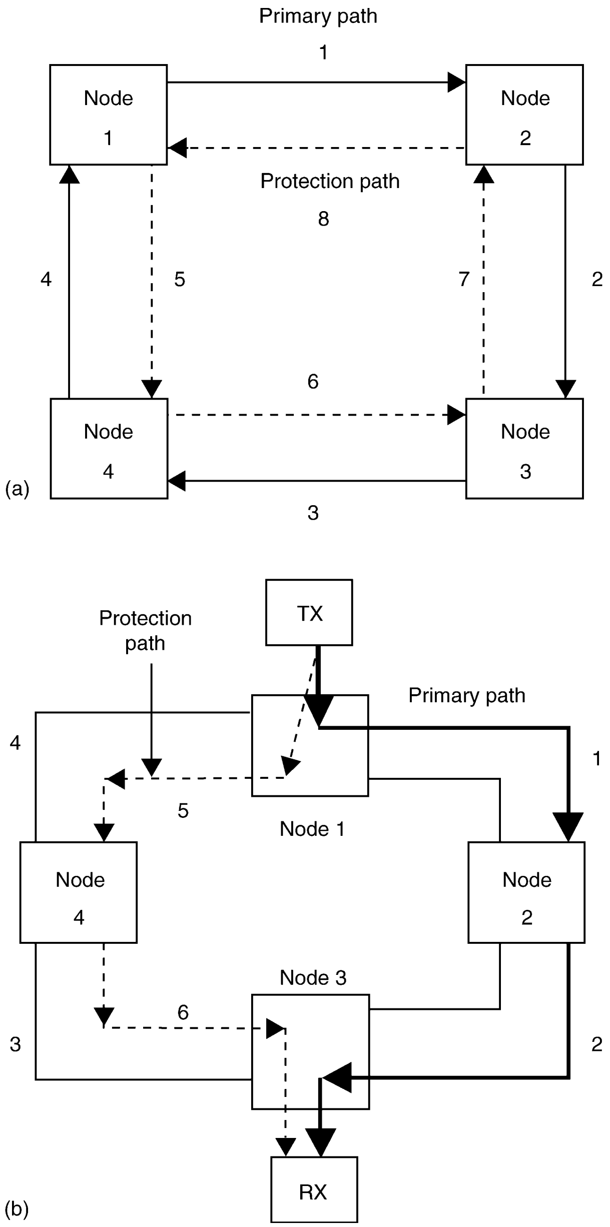

Figure 3 Schematic of a two-fiber unidirectional path-switched ring network.

To see the function and versatility of the standby links in the four-fiber BLSR, consider first the case where a transmitter or receiver circuit card used on the primary ring fails in either node 3 or 4. In this situation the affected nodes detect a loss-of-signal condition and switch both primary fibers connecting them to the secondary protection pair, as shown in Figure 5. The protection segment between nodes 3 and 4 now becomes part of the primary bidirectional loop. The exact same reconfiguration scenario will

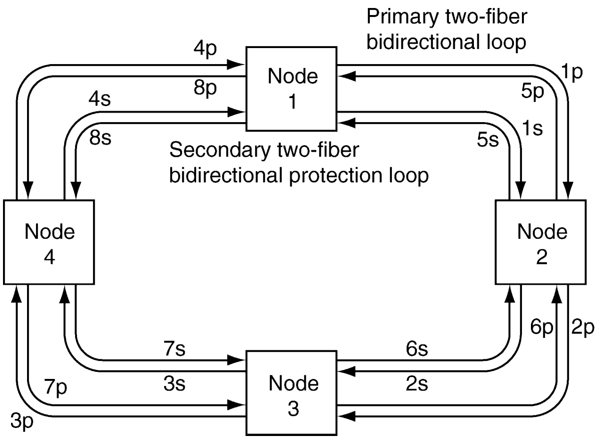
Primary two-fiber
bidirectional loop
4p
8p
4s
8s
Node
1
1p
5p
1s
5s
Secondary two-fiber
bidirectional protection loop
Node
4
Node
2
6p
2p
7s
6s
7p
3s
2s
3p
Node
3

Figure 4 The architecture of a four-fiber bidirectional, line-switched ring.

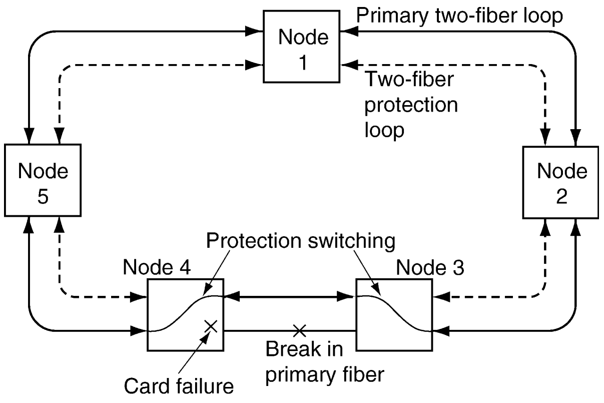
Node
1
Primary two-fiber loop
Two-fiber
protection
loop
Node
5
Node
2
Protection switching
Node 4
Node 3
Break in
primary fiber
Card failure

Figure 5 Reconfiguration of a four-fiber BLSR under transceiver failure.

occur when the primary fiber connecting nodes 3 and 4 breaks. Note that in either case the other links remain unaffected.

Now suppose an entire node fails, or both the primary and the protection fibers in a given span are severed, which could happen if they are in the same cable duct between two nodes. In this case the nodes on either side of the failed span internally switch the primary-path connections from their receivers and transmitters to the protection fibers, in order to loop traffic back to the previous node. This process again forms a closed ring, but now with all of the primary and protection fibers in use around the entire ring, as shown in Figure 6.

Single-Span WDM Links

An interesting and powerful aspect of an optical communication link is that many different wavelengths can be sent along a fiber simultaneously in the 1300–1600 nm spectrum. The technology of combining a number of wavelengths onto the same fiber is known as wavelength division multiplexing or WDM. Figure 7 shows the basic WDM concept. Here N independent optically formatted information streams, each transmitted at a different wavelength, are combined with an optical multiplexer and sent over the same fiber. Note that each of these streams could be at a different data rate. Each information stream maintains its individual data rate after being multiplexed with the other streams, and still operates at its unique wavelength. Conceptually, the WDM scheme is the same as frequency division multiplexing (FDM) used in microwave radio and satellite systems.

To see the potential of WDM, consider the characteristics of a high-quality optical source such as a distributed feedback (DFB) laser, which has a very narrow frequency spectrum on the order of 1 MHz, which is equivalent to a spectral linewidth of 10^{-5} nm. When using such a source, a guard band of 0.4–1.6 nm is typically employed. This is done to take into account possible drifts of the peak wavelength due to aging or temperature effects, and to give both the manufacturer and the user some leeway in specifying and choosing the precise peak emission wavelength. With such spectral band widths, simplex systems make use of only a very

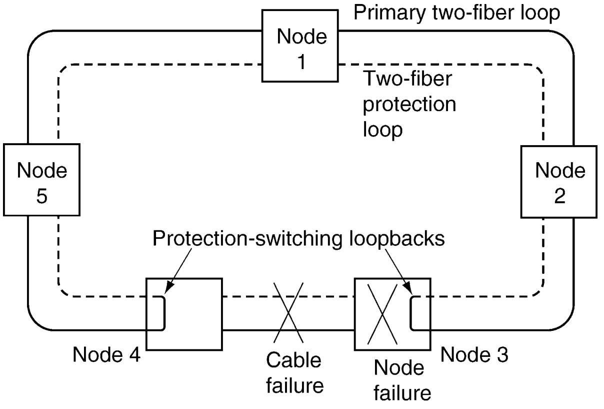
Node
1
Primary two-fiber loop
Two-fiber
protection
loop
Node
5
Node
2
Protection-switching loopbacks
Node 4
Cable
failure
Node
failure
Node 3

Figure 6 Reconfiguration of a four-fiber BLSR under node or fiber-cable failure.

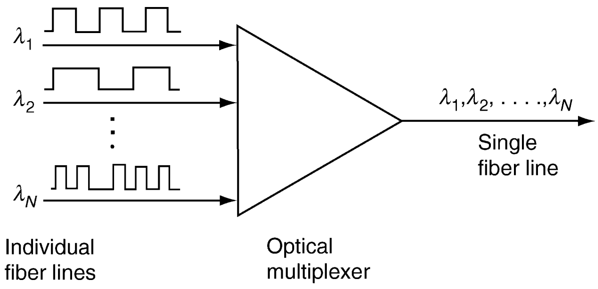
λ_1
λ_2
λ_N
$\lambda_1, \lambda_2, \ldots, \lambda_N$
Single
fiber line
Individual
fiber lines
Optical
multiplexer

Figure 7 The basic WDM concept.

small portion of the transmission bandwidth capability of a fiber. This can be seen from Figure 2, which depicts the attenuation of light in a silica fiber as a function of wavelength. The curve shows that the two low-loss regions of a single-mode fiber extend over the wavelengths ranging from about 1260 to 1360 nm (the O-band) and from 1460 to 1625 nm (the S-, C-, and L-bands).

These regions can be viewed either in terms of spectral width (the wavelength band occupied by the light signal and its guard band) or by means of optical bandwidth (the frequency band occupied by the light signal). To find the optical bandwidth corresponding to a particular spectral width in these regions, consider the fundamental relationship $c = \lambda\nu$, which relates the wavelength λ to the carrier frequency ν, where c is the speed of light. Differentiating this yields $\Delta\nu = (c/\lambda^2)\Delta\lambda$, where the deviation in frequency $\Delta\nu$ corresponds to the wavelength deviation $\Delta\lambda$ around λ. Thus the optical bandwidth is $\Delta\nu = 14$ THz for a usable spectral band $\Delta\lambda = 80$ nm in the O-band. Similarly, $\Delta\nu = 15$ THz for a usable spectral band $\Delta\lambda = 120$ nm in the region spanning the S-, C-, and L-bands.

Since the spectral width of a high-quality source occupies only a narrow optical bandwidth, the two low-loss windows provide many additional operating regions. By using a number of light sources, each emitting at a different peak wavelength that is sufficiently spaced from its neighbor so as not to create interference, the fidelities of the independent messages from each source are maintained for subsequent conversion to electrical signals at the receiving end.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the WDM standards developed by the ITU specify channel spacings in terms of frequency. A key reason for selecting a fixed frequency spacing, rather than a constant wavelength spacing, is that when locking a laser to a particular operating mode it is the frequency of the laser that is fixed. The ITU-T Recommendation G.692 specifies selecting the channels from a grid of frequencies referenced to 193.100 THz (1552.524 nm) and spacing them 100 GHz (0.8 nm at 1552 nm) apart. Suggested alternative spacings include 50 GHz and 200 GHz.

Historically the term dense WDM (DWDM) was used somewhat loosely and generally referred to the spacings denoted by Recommendation G.692. In 2002 the ITU-T released an updated standard aimed specifically at DWDM. This is Recommendation G.694.1, which specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm).

Alternatively, by using the larger spectral window available in low-water-content fibers, one can relax the spectral tolerances on the optical components and space the individual sources much further apart. This is the basis of coarse WDM (CWDM), which is useful for metropolitan area networks where variations in attenuation between different wavelength channels is not critical over short transmission distances. In 2002 the ITU-T released Recommendation G.694.2, which is aimed specifically at CWDM. The CWDM grid is made up of 18 wavelengths defined within the range 1270–1610 nm (O- through L-bands) spaced by 20 nm with wavelength-drift tolerances of ±2 nm.

A key feature of WDM is that the discrete wavelengths form an orthogonal set of carriers which can be separated, routed, and switched without interfering with each other. This holds as long as the total optical power intensity is kept sufficiently low to prevent nonlinear effects such as stimulated Brillouin scattering and four-wave mixing processes from degrading the link performance.

The implementation of WDM networks requires a variety of passive and/or active devices to combine, distribute, isolate, and amplify optical power at different wavelengths. Passive devices require no external control for their operation, so they are somewhat limited in their application in WDM networks. These components are mainly used to split and combine or tap off optical signals. The performance of active devices can be controlled electronically, thereby providing a large degree of network flexibility. Active WDM components include tunable optical filters, tunable sources, and optical amplifiers.

Passive Optical Networks

Another step towards realizing the full potential of optical fiber transmission capacity is the concept of all-optical WDM networks to extend the versatility of communication networks beyond architectures such as those provided by SONET. These networks can be classified as either broadcast-and-select or wavelength-routing networks. In general, broadcast-and-select techniques employing passive optical stars, buses, or wavelength routers are used for local network applications, whereas active optical components form the basis for constructing wide-area wavelength-routing networks. Broadcast-and-select networks can be categorized as single-hop or multihop networks. Single-hop refers to networks where information transmitted in the form of light reaches its destination without being converted to an electrical form at any intermediate point. On the other hand, intermediate conversion can occur in a multihop network.

Figure 8 shows two alternate physical architectures for a WDM-based local network. Here *N* sets of transmitters and receivers are attached either to a star coupler or a passive bus. Each transmitter sends its information at a different fixed wavelength. All the transmissions from the various nodes are combined in a passive star coupler or coupled onto a bus and the result is sent out to all receivers.

Each receiver sees all wavelengths and uses a tunable filter to select the one wavelength addressed to it. In addition to point-to-point links, this configuration can also support multicast or broadcast services, where one transmitter sends the same information to several nodes. An interesting point is that these passive WDM networks are protocol transparent. This means that different sets of communicating nodes can use different information-exchange rules (protocols) without affecting the other nodes in the network. This is analogous to current time-division-multiplexed telephone lines in which voice, data, or facsimile services are sent in different time slots without interfering with each other.

Although the architectures of single-hop broadcast-and-select networks are fairly simple, there needs to be careful dynamic coordination between the nodes. For example, if the network is set up so that a transmitter sends its information at a unique fixed wavelength, then the destined receiver needs to be informed when a message is being sent to it, so that it can tune its selective filter to that wavelength. Also conflicts need to be resolved in cases where two stations transmitting at different wavelengths want to send information to the same recipient simultaneously.

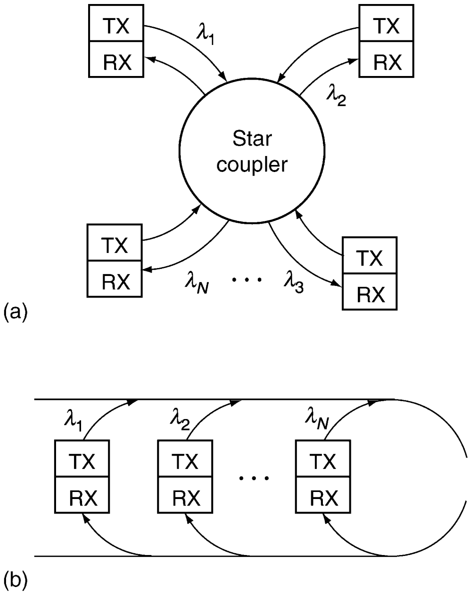

Figure 8 Two alternate physical architectures for a WDM-based local network.

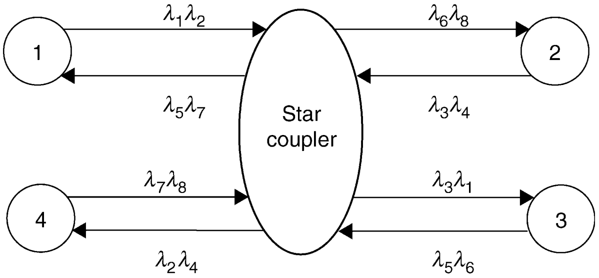

Figure 9 Example of a four-node broadcast-and-select multihop network.

A drawback of single-hop networks is the need for rapidly tunable lasers or receiver optical filters. The designs of multihop networks avoid this need. Multihop networks generally do not have direct paths between each node pair. Each node has a small number of fixed-tuned optical transmitters and receivers. Figure 9 shows an example of a four-node broadcast-and-select multihop network where each node transmits on one set of two fixed wavelengths and receives on another set of two wavelengths. Stations can send information directly only to those nodes that have a receiver tuned to one of the two transmit wavelengths. Information destined for other nodes will have to be routed through intermediate stations.

The flow of traffic can be seen from Figure 9. If node 1 wants to send a message to node 2, it first transmits the message to node 3 using λ_1. Then node 3 forwards the message to node 2 using λ_6. In contrast to single-hop networks, with this scheme there are no destination conflicts or packet collisions in the network, since each wavelength channel is dedicated to a particular source–destination link. However, for *H* hops between nodes, there is a network throughput penalty of at least $1/H$.

Wavelength-Routed Optical Networks

Wavelength-routed networks overcome the limitations of passive optical networks through wavelength reuse, wavelength conversion, and optical switching. The physical topology of a wavelength-routed network consists of optical wavelength routers interconnected by pairs of point-to-point fiber links in an arbitrary mesh configuration, as illustrated in Figure 10. Each link can carry a certain number of wavelengths, which can be directed independently to different output paths at a node. Each node may have logical connections with several other nodes in the network, where each connection uses a particular wavelength. Provided the paths taken by any two connections do not overlap, they can use the same wavelength. Thereby the number of wavelengths is greatly reduced. For example, the connection from node 1 to node 2 and from node 2 to node 3 can both be on λ_1, whereas the connection between nodes 4 and 5 requires a different wavelength (λ_2).

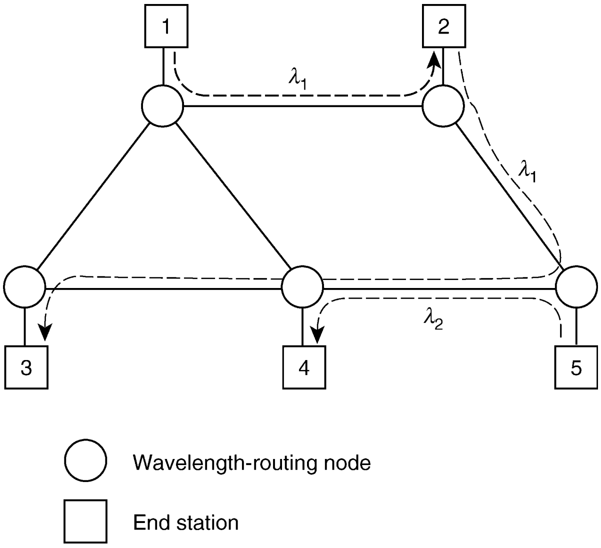

Figure 10 The physical topology of a wavelength-routed network.

A high degree of path modularity, capacity scaling, and flexibility in adding or dropping channels at a user site can be achieved by introducing the concept of an optical cross-connect architecture in the physical path structure (the so-called path layer) of an optical network. These optical cross-connects (OXCs) operate directly in the optical domain and can route very high-capacity WDM data streams over a network of interconnected optical paths.

Contentions could arise when channels having the same wavelength but traveling on different input fibers enter the OXC and need to be switched simultaneously to the same output fiber. This could be resolved by assigning a fixed wavelength to each optical path throughout the network, or by dropping one of the channels and retransmitting it at another wavelength. However, in the first case, wavelength reuse and network scalability (expandability) are reduced, and in the second case the add/drop flexibility of the OXC is lost. These blocking characteristics can be eliminated by using wavelength conversion at any output of the OXC.

As an example consider the 4 × 4 OXC shown in Figure 11. Here two input fibers are each carrying two wavelengths. Either wavelength can be switched to any of the four output ports. The OXC consists of three 2 × 2 switch elements. Suppose that λ_2 on input fiber 1 needs to be switched to output fiber 2 and that λ_1 on input fiber 2 needs to be switched to output fiber 1. This is achieved by having the first two switch elements set in the bar state (the straight-through configuration) and the third element set in the cross state, as indicated in Figure 11. Obviously without wavelength conversion there would be wavelength contention at both output ports.

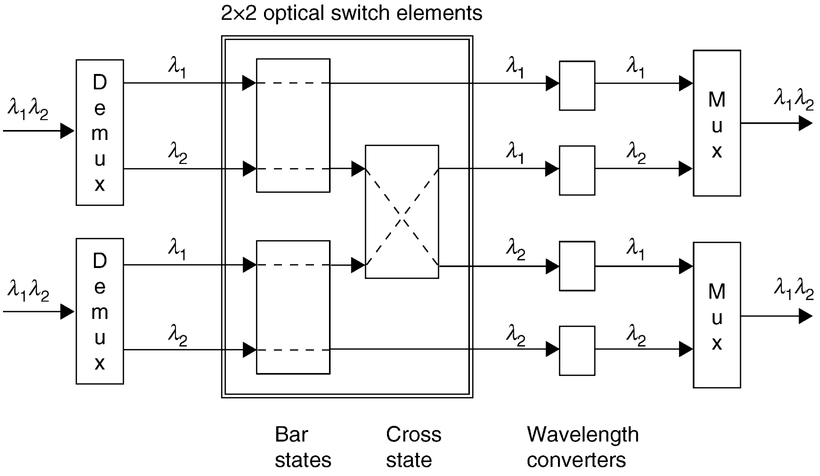

Figure 11 Example of a 4 × 4 optical cross-connect (OXC).

By using wavelength converters, the cross-connected channels can be converted to noncontending wavelengths.

See also

Lasers: Optical Fiber Lasers. **Optical Communication Systems:** Wavelength Division Multiplexing.

Further Reading

Blumenthal DJ, Prucnal PR and Sauer JR (1994) Photonic packet switches: architectures and experimental implementations. *Proceedings of the Institute of Electrical and Electronics Engineering* 82(11): 1650–1667.

Forouzan BA (2001) *Introduction to Data Communications and Networking*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Goralski WJ (2002) *SONET/SDH*, 3rd edn. New York: McGraw-Hill.

Gumaste A and Antony T (2003) *DWDM Network Designs and Engineering Solutions*. Upper Saddle River, NJ: Prentice Hall.

Hecht J (2002) *Understanding Fiber Optics*, 4th edn. Upper Saddle River, NJ: Prentice Hall.

Iannone E and Sabella R (1996) Optical path technologies: a comparison among different cross-connect architectures. *Journal of Lightwave Technology* 14(10): 2184–2196.

Keiser G (1999) A review of WDM technology and applications. *Optical Fiber Technology* 5(1): 3–39.

Keiser G (2000) *Optical Fiber Communications*, 3rd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2002) *Local Area Networks*, 2nd edn. Burr Ridge, IL: McGraw-Hill.

Keiser G (2003) *Optical Communications Essentials*, New York: McGraw-Hill.

Ramaswami R and Sivarajan KN (2002) *Optical Networks*, 2nd edn. San Francisco: Morgan Kaufmann.

Special Issue (1995) *Undersea Communications Technology. AT&T Technical Journal* 74(1): 4–92.

Trischitta PR and Marra WC (1998) Applying WDM technology to undersea cable networks. *IEEE Communications Magazine* 36(2): 62–66.

Zang H, Jue JP, Sahasrabuddhe L, *et al.* (2001) Dynamic lightpath establishment in wavelength-routed WDM networks. *IEEE Communications Magazine* 39: 100–108.

Free Space Optical Communications

R Martini, Stevens Institute of Technology,
Hoboken, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Free space optical (FSO) communication is the wireless transmission of data via a modulated optical beam directed through free space, without fiber optics or other optical systems guiding the light. The fundamental idea goes back to ancient times, as light (or smoke) signals were used to transmit information. From a more modern point of view, Graham Bell's patent on the photophone may mark the onset of modern FSO techniques, as it transmitted audio signals (i.e., voice) via the modulation of sunlight. A renaissance of FSO systems started with the availability of lasers, light sources with high output power and high coherence, which allowed the accurate direction of the light beams over long distances. During the 1970s and 1980s the main proposed application of FSO systems was for secure and long distance (50–1000 km) communication, mainly targeted for ground–satellite or satellite–satellite communication. This focus changed drastically over the last decade as a new market for FSO grew in the establishment of high bandwidth data link and their integration over a locally restricted area.

The main competitors in this market are the fiber-based optical network, the RF communication system, as well as the low bandwidth copper cable-based system. In comparison to the closely related wireless radio frequency transmission, the higher frequency of the optical carrier ($\sim 10^{14}$–10^{15} Hz) thereby allows for much higher transmission rates, comparable to those of typical fiber optic networks. On the other hand, the use of an optical carrier also results in much more directed beam propagation, which requires an undisturbed line-of-sight between emitter and detector. This restricts the application of most FSO systems to a range between a few hundred meters up to several kilometers, which are still favorable for distribution of high bandwidth networks over a locally restricted area. This makes FSO a highly attractive candidate for the 'last-mile' distribution of high-bandwidth Ethernet to the individual homes.

The simplicity of setup of FSO links, as well as their modularity, is thereby their biggest advantage, compared to fiber-based networks. It makes them not only highly cost efficient, but it eases the maintenance or allows for fast and easy upgrades, as it does not require any extensive and time-consuming installation – in contrast to the installation of fiber optic cables. This advantage makes FSO highly attractive for temporary installations (emergency or short-time high broadcast situations, such as the Olympic Games), as well as to overcome geometrical restraints (river, seas, etc.). As most detector and emitter systems are typically based on the same electro-optical components used for fiber optic networks, they seamlessly integrate and expand an existing network without complications in data handling.

Standard FSO System

Figure 1 shows schematically the setup for a typical FSO system connecting two separate local networks. It consists of: an electronic data input; a small but powerful light source which can be modulated; emitter optics which shape the emitted beam to a highly directed beam; the atmosphere as transmitting device; detector optics which receive the transmitted light and focus it onto a photodetector; and an electronic amplifier which serves as data output.

Compared to a fiber optical communication system, the main difference is the use of the atmosphere as the transmission medium, in contrast to guided optical propagation. The transmission properties are thereby determined by the atmospheric conditions, which can lead to degradation and/or redirection of the beam. The main influences which have to be taken into account are atmospheric scattering, molecular absorption, turbulence effects (including beam wander and fading), as well as the impact of low-visibility weather situations like rain, snow, or fog. These depend strongly on wavelength and linewidth of the used emitter system, distance between emitter and receiver, as well as the environmental conditions. As a consequence, the characteristics of the atmospheric channel determine the layout of the other FSO components – and have to be considered first.

Atmospheric Losses

The overall losses of the atmospheric transmission are determined by the Beer–Lambert absorption law:

$$I = I_0 \exp[-\alpha L] \qquad [1]$$

where I is the intensity at the receiver, I_0 the intensity of the transmission beam after the emitter optics, and L is the range (i.e., distance between emitter and detector optics). The total extinction coefficient $\alpha = \alpha_{\mathrm{Rayleigh}} + \alpha_{\mathrm{Mie}} + \alpha_{\mathrm{Absorption}}$ is based on three different contributions: Rayleigh scattering, Mie scattering, and absorption.

In contrast to absorption, the losses due to scattering originate from redirection of light and will be discussed first. Depending on the size of the scattering partner, two separate regimes have to be taken into account. For scatter smaller than the wavelength, the effect can be described based on Rayleigh's classical formula. The scattering cross-section of an element with mass m, charge q, natural oscillation wavelength λ_0, is then:

$$\sigma_{\mathrm{s}} = \frac{fq^4\lambda_0^4}{6\pi\varepsilon_0^2 m^2 c^4}\frac{1}{\lambda^4} \qquad [2]$$

where ε_0 is the dielectric constant, c the speed of light, and λ the wavelength of the incident light, leading to an extinction coefficient of $\alpha_{\mathrm{Rayleigh}} = \sigma_{\mathrm{s}} N_{\mathrm{s}}$, where N_{s} is the concentration of the scattering elements. The λ^{-4} dependence in the scattering predicts stronger scattering for shorter wavelengths and Rayleigh scattering is therefore dominant for short and visible wavelengths (which gives the sky its blue color). For typical FSO

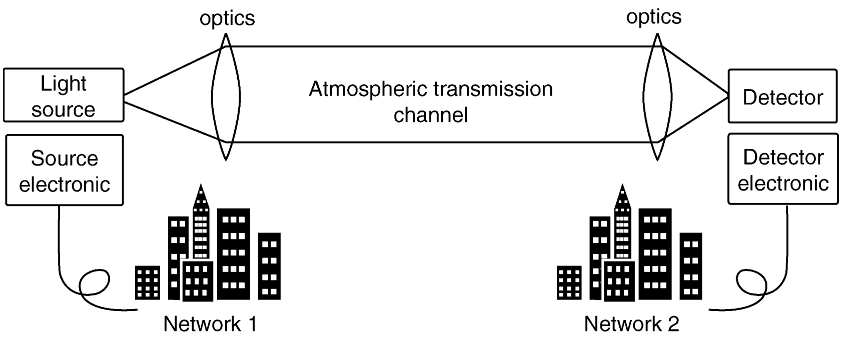

Figure 1 Typical components of a FSO system.

wavelengths in the region of near-infrared radiation (NIR, 800–2000 nm) or mid-infrared radiation (MIR, 2000–20 000 nm), the contribution of Rayleigh scattering is typically relatively low and can be neglected.

On the other hand, Mie scattering is, in most cases, the strongest contribution for FSO systems – but also the hardest to predict. It is due to resonant scattering effects, where the particle size is of the wavelength of the scattered light. For the typical FSO wavelengths these particles include water and oil droplets, dust, smog, fly ash, and aerosols (which can originate from pollution). A correct calculation of the Mie scattering coefficient is possible, but requires knowledge of the exact particle concentration as well as their size distribution. As these numbers are not only difficult to measure, but also fluctuate continuously depending on weather situation, time of day, season, temperature, climate, pollution, and other factors, a numerically correct evaluation is not possible.

Historically, most systems rely on an empirical formula, which estimates the attenuation coefficient to

$$\sigma_{\mathrm{Mie}} = \frac{3.91}{V}\left(\frac{\lambda}{550}\right)^{-p} \qquad [3]$$

where V denotes the visibility and the value of p ranges from 1.6 for high visibility ($V > 50$ km) and 1.3 for average visibility (50 km $> V >$ 6 km) down to 0.585 $V_{1/3}$, respectively for low visibility. The formula is based on data collected in the 1950s and 1960s. More accurate measurements show that the actual scattering cross-sections differ quite strongly, especially for low-visibility cases. Especially, the predicted dependence on wavelength is not correct. The failure of such a general model is not surprising, as the actual scattering properties are strongly dependent on the weather situation and the environmental conditions. It is the focus of ongoing research to find more accurate models describing the scattering condition in multiple environments and conditions to provide the background for more accurate empirical models.

Apart from the light redirection due to scattering, absorption either by molecules or aerosols is nearly as important for the transmission calculation of a FSO system. The strongest molecular absorptions are typically due to water, carbon dioxide, as well as ozone. A typical absorption spectrum can be easily calculated using standard software together with the publicly available HITRAN database, both available from ONTAR. The upper part of Figure 2 shows a low-resolution transmission spectrum for the interesting wavelength region for a path length of 1 km.

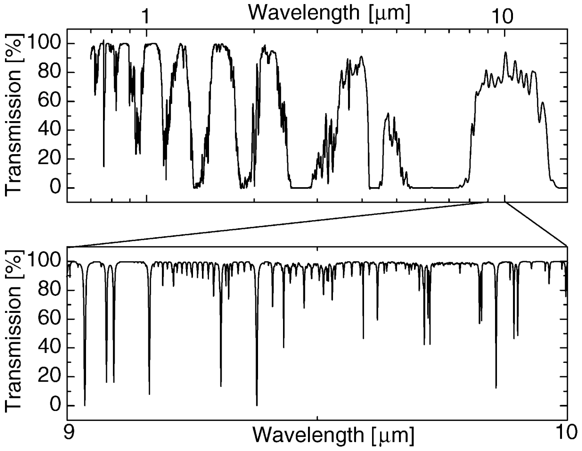

Figure 2 Upper: Low-resolution transmission spectrum of the atmosphere showing molecular absorption for a path length of 1 km. Lower: high-resolution spectrum of the atmospheric windows around 9.5 μm.

Within this spectrum, several well-known atmospheric windows exist, which are naturally preferred for FSO applications. These include the region around 830 nm, 1550 nm, and from 8 μm to 12 μm. For the exact selection of a wavelength, a closer look is needed. As an example, a high-resolution analysis of the atmospheric windows around 9.5 μm is shown in the lower part of Figure 2, which identifies additional absorption lines as well as highly transparent wavelength ranges within the window. Such substructure has to be taken carefully into account for the FSO wavelength selection, especially if lasers are employed, which have a linewidth comparable to the width of the absorption line.

Besides the molecules, aerosol particles also contribute to the absorption spectrum. Calculation of their contribution can also be performed using available software package, such as Air Force's MODTRAN and LOWTRAN program as well as OPAC from the LMU University in Munich. Figure 3 shows a typical low-resolution transmission spectrum for 50% humidity as well as 99% humidity, where by scattering as well as absorption effects have both been included in the calculation. The occurrence of sharp absorption features, as well as the increasing losses for wavelengths above 10 μm, is typically due to the aerosol absorption, whereas the scattering dominates the function for wavelengths below 3 μm. Losses due to the aforementioned Rayleigh scattering are also shown in the same diagram, demonstrating its low influence at longer wavelengths.

MODTRAN and LOWTRAN also allow the calculation of transmission spectra under rain and snow conditions, where rain and snow consist of

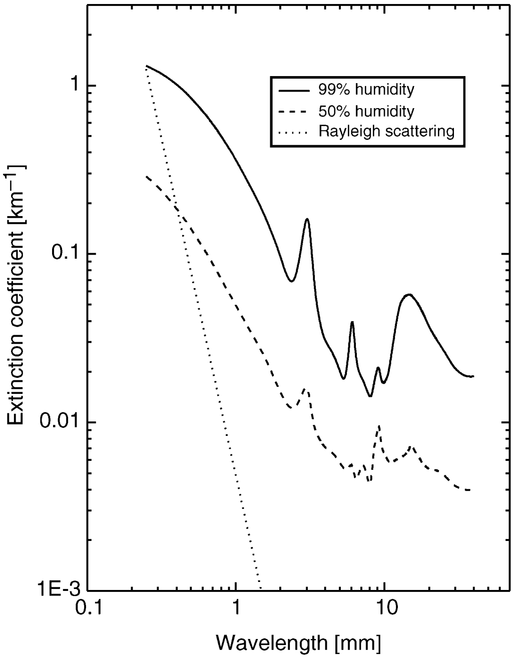

Figure 3 Typical low-resolution transmission spectrum calculated based on absorption and scattering losses from aerosols for 50% humidity (dashed line) and 99% humidity (solid line). The dotted line shows additionally the losses due to Rayleigh scattering.

clusters of water and aerosol with diameters from 200 μm to 2 mm and above. Depending on the severity of the precipitation, the losses can range from ~0.5 dB/km for light rain up to 13 dB/km for cloudbursts, requiring an additional power margin in the preparation of the FSO link. Significantly stronger is the impact of fog, as typically particles with much lower sizes are present. The losses inflicted by a moderate fog (visibility 500 m) range above 20 dB/km, surpassing the effect of a cloudburst; for thick fog the attenuation can even grow beyond 200 dB/km. However, depending on the size distribution within the fog itself, the losses can be highly wavelength dependent. Recent experiments show that long wavelength FSO systems operating in the MIR wavelength region have a much stronger stability under the influence of thick fog than one operating at NIR wavelengths.

Turbulence and Scintillation

In contrast to a static medium, the atmosphere is a dynamic mixture which undergoes continuous changes due to wind, temperature, and other influences. This creates a constant change in the parts of the atmosphere which are contributing to the FSO channel and therefore in the optical properties of the beam path. The index of refraction can differ, not only in its total value, but also within subsections of it, creating a random fluctuation in the beam paths. Such localized changes in the refractive index of the air are schematically shown in Figure 4, where turbulences of different sizes modify the beam path and disturb the beam front. The turbulence in the atmosphere can be quantified by the refractive-turbulence structure coefficient C_n^2, which can range from $10^{-13}\,\mathrm{m}^{-2/3}$ in the middle of the afternoon (highest value) to $10^{-17}\,\mathrm{m}^{-2/3}$ one hour after sunrise or sunset. As the turbulence effect is typically due to strong temperature difference, C_n^2 is strongest near hot surfaces (e.g., tar roofs or asphalt in the summer) and decreases with altitude. The effects of such turbulences on a directed light beam (such as used in FSO systems) include mainly beam wander and scintillation effects.

Beam wander is thereby based on large size turbulences which act as a lens on the propagating beam. Consequently, a narrow focused beam could be randomly steered away from the detector area, creating a fluctuation in the detected intensity.

Stronger impact for FSO links has scintillation, which is better known as the twinkling of a star or the heat shimmer of the horizon on a hot day. It is based on the continuous variation of the beam front which can lead to intensity fluctuations (like the twinkling) and loss in the image resolution (heat shimmer). Strong intensity fluctuations can thereby cause signal loss as well as receiver saturation, both resulting in downtime in the link. Such local intensity disturbances can occur up to sizes of $\sqrt{\lambda L}$, reaching, for example, 4 cm for a 1.55 μm link after 1 km. For small fluctuations, the intensity follows a classical stochastic behavior and can be described by the irradiance variation (normalized to its mean value) as

$$\sigma^2 = \exp\left[0.5\left(\frac{2\pi}{\lambda}\right)^{6/7} L^{11/6} C_n^2\right] - 1 \qquad [4]$$

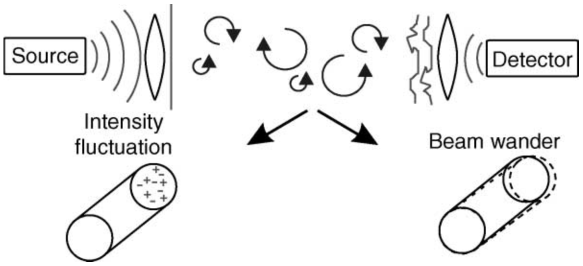

Figure 4 Schematic depiction of atmospheric turbulences with different sizes in FSO systems. Shown below are illustrations of its main effects on a directed beam causing intensity fluctuations and beam wander.

with usual values in the range from 5 to 70%. Nonetheless, it should be pointed out that actual turbulence effects can differ quite strongly from these predictions, due to local inhomogeneities in temperature distribution. Consequently, test measurements are necessary for the correct evaluation of a specific installation site.

What makes scintillation effects difficult to handle for communications systems is the relatively slow frequency of fluctuations in the order of a few hundred Hz or less. The corresponding downtimes for a link can range from 1 ms to 10 ms, or even up to seconds. This is magnitudes higher than typical downtimes in standard communications links and cannot be compensated by standard forward error correction systems. One possibility to reduce the influence of scintillation is the method of aperture averaging, whereby multiple beams originating from slightly different positions are deployed and targeted at the same receiver. Assuming that the variation in each of the separate beam paths is random and uncorrelated, the signal-to-noise ratio will decrease by the square root of the number of beams. Recent experiments have impressively demonstrated this effect and allow reduction of intensity fluctuation to an acceptable level.

Another alternative to decreasing the scintillation effects is in the application of longer wavelength FSO systems. Whereas beam wander is relatively independent from the wavelength, scintillations effects scale with $\lambda^{7/6}$, as turbulences smaller than the wavelength cannot obscure its beam front. Using a 2 times longer wavelength, for example, should have the same impact as signal averaging with 4 different beams.

Beam Spreading and Pointing

Based on optical diffraction theory, the beam leaving the emitter optics will be divergent and its beam diameter will consistently increase during transmission through the optical channel. Assuming an optimal alignment and using standard geometrical optics, the losses due to beam spreading can estimate the ratio of emitted power (P_E) to received power (P_R):

$$R_\mathrm{BS} = \frac{P_\mathrm{R}}{P_\mathrm{E}} = \frac{D_\mathrm{R}}{(D_\mathrm{E} + L\theta)} \qquad [5]$$

where D_R is the diameter of the receiving optics, D_E the diameter of the emitting optics, and θ the divergence of the beam. The minimal divergence of the beam is typically determined by the light source and is given by $\theta_\mathrm{L} = \lambda / D_\mathrm{L} M^2$, where D_L is the initial size of the laser beam and M^2 the laser-mode structure parameter, for a typical semiconductor laser θ_L below 0.1 mrad. In many cases, the divergence of the emitted beam is commonly enlarged by slight defocusing of the emitting optics – if the power budget allows for the additional losses. The beam size at the receiver is typically several times larger than the receiving aperture, which not only eases the alignment of the system, but also reduces pointing losses. This includes the effect of beam wander due to turbulence as well as swaying of the buildings or the mount. If those effects are solely compensated for by an increased beam spread, divergence can easily reach values of 2–20 mrad for short-distance links. Obviously this is not an option for longer-distance FSO links, as the losses will increase dramatically. In these cases, active tracking systems are incorporated in the emitter optics allowing for continuously adjustment to achieve and keep an optimal alignment. Typical tracking devices are thereby based on servo-driven Gimbal mounts, steering mirrors, or recently, MEMS devices.

Laser Eye Safety

As the atmospheric transmission channel of most FSO systems is not completely isolated, there exists the possibility of a human person intercepting the beam. To allow for this, most FSO systems are designed to be eye-safe following the voluntary ANSI standard, which limits the output irradiance to a certain value depending on the deployed wavelength. As radiation below 1400 nm is still being focused on the retina, it causes a higher irradiance, which results in a lower damage threshold. For 800 nm laser systems, the minimum permissible exposure (MPE) level for a 10 s long exposure is therefore recommended not to exceed 1 mW cm^{-2}, whereas for 1.55 μm or longer wavelength systems, an MPE of 100 mW cm^{-2} is suggested. The recommendations for incoherent light sources such as LEDs are thereby less stringent as they cannot be focused quite as strongly as laser light.

The Light Source

Following the discussion above, the optimal light source for FSO system should be emitting at a wavelength away from any absorption line (or should be much broader then the absorption line), have a relatively long wavelength as well as a relative low divergence, be eye-safe, and allow for high modulation speed. In reality, most FSO systems use either GaAsAl laser diodes or LEDs emitting at 0.8 μm, or the typically more expensive 1.5 μm laser diodes based on the InGaAsP semiconductor system. The output power of both laser systems is typically in the range of 1 mW to 100 mW, which for the longer wavelength case can be boosted to 1–10 W using erbium-doped fiber amplifiers (or even higher for specific applications). The longer wavelength system has the additional advantage of higher ANSI standards as discussed above, allowing stronger eye-safe emission power, but is typically far more expensive.

Both laser types were developed and optimized for standard fiber-based communication systems allowing direct modulation up to 10 Gbit/s. To achieve higher bandwidth in FSO systems, standard techniques known from fiber-based networks can easily be deployed, like wavelength division multiplexing which increases the bandwidth up to state-of-the-art levels of fiber-based networks.

Contrary to semiconductor lasers, LEDs provide a much less expensive alternative, but allow only for modulation up to 10–100 MHz. Their broad emission spectrum is thereby an advantage, as it is less sensitive to sharp absorption lines which – as discussed earlier – can have a drastic impact on the FSO link performance. Consequently, LEDs have been successfully implemented in several commercial FSO systems targeted at the short-range market.

Recently developed sources, which have also been applied for FSO systems, include vertical-cavity surface-emitting lasers as well as quantum cascade lasers (QCL) which emit in the mid-infrared regime. The latter devices have been predicted to have a strong impact on the FSO market, as they combine high-output power (up to 1 W) together with fundamentally high modulation bandwidth (2.5 Gbit/s recently demonstrated and higher rates predicted) at wavelengths ranging from 3.5 μm to 12 μm and beyond. They can access the long-wavelength atmospheric window atmospheric for the FSO systems (typically reserved for CO_2 lasers) with the convenience of direct semiconductor lasers like direct electrical modulation, small size, and low power consumption. Recently, first FSO experiments, employing experimental QCLs, demonstrated the predicted higher stability of the link in low-visibility situations, including strong fog as well as high dust concentration. Nonetheless, to date the high cooling requirement for QCLs restricts more general use of them, but their very fast development, recent reports of cw-operation of room-temperature QCLs as well as FSO links with TE-cooled QCLs in pulse operation, give hope for a solution to this problem.

The Detector System

Once the light reaches the receiver, it is typically collimated using a short focal length telescope or a single lens system. To optimize efficiency without sacrificing ease of alignment a compromise has to be made between a wide aperture for greater light collection and a short focal length. The concentrated light then hits the detector, which, depending on the wavelength, is either Si or InGaAs-photodiodes. Si-based devices can detect up to wavelengths of 1100 nm and InGaAs systems are typically used for the 1.5 μm region, respectively. To enhance the sensitivity of the detectors, avalanche photodiodes are preferred which internally enhance the detected signal by a factor ranging from 20 to 100. To ensure high bandwidth operation (10 MHz–10 GHz), the size of the detecting device has to be kept relatively small (20 to 100 μm) avoiding parasitical capacitance effects – but raising the requirement on the detecting optics.

The sensitivity of these NIR detectors is typically limited by the shot-noise, which is due to statistical fluctuations of the received phonons and scales with the square root of its number. For practical purposes the internal noise of the detector is quantified using the corresponding amount of incident optical power leading to a comparable signal. The values for this noise equivalent power (NEP) can range from microwatts to tens of nanowatts. Other sources for detector noise can include a background current from the detector–amplifier combination, typically referred to as Johnson noise.

Mid-infrared detectors are additionally objected to strong background radiation due to thermal emission, which peaks around the 10 μm at temperature of 300 K. The use of additional cold filters and high-resolution spectral filters can help to overcome this restriction, but still the sensitivity of those detectors is lacking compared to NIR systems.

For communication system the noise is more commonly quantified as bit error rate (BER), quantifying the average percentage of wrongly received bits (submitted '1', detected '0' or vice versa). A BER greater than 10^{-9} is usually referred to as an error-free system, as the residual errors can be corrected employing forward error correcting algorithms. In contrast, the performance of FSO systems can more easily be measured in its signal-to-noise ratio (SNR). Following standard communication theory, a peak SNR of 12 and an average SNR of 6 is required to achieve a BER of 10^{-9}, guaranteeing error-free communication. Consequently range limits for FSO links are typically based on $\text{SNR} = 6$.

Link Budget and FSO Range Equation

Classical link budget estimation is a commonly used tool to access the performance of FSO systems. To take into account the losses due to atmospheric absorption, scattering, turbulence, and beam spreading, as well as optical losses in the used receiver and emitter optics, they are deducted from the starting power emitted from the FSO light source. The estimated intensity at the detector is then compared with the intensity needed to establish error-free communication (for example $\text{SNR} = 6$), the calculated difference is the link margin of the specific system, i.e., the amount of additional losses under which the FSO link will still be working.

Comparing this number to losses from nonpredictable fluctuations (like heavy rain or fog) gives an insight under which implications the link will fail. Using weather station data, the frequency of such extremely low-visibility occurrences can be obtained and a number for the reliability of the FSO link be estimated. Typical well-installed FSO systems estimate very high reliabilities, between 99.9% and 99.999%, which have been obtained in installation and verified in long-term measurements.

Another way to estimate the performance for a given system is based on the FSO range equation, where the received power on the detector (P_R) is calculated to

$$P_R = P_E \frac{D_R}{(D_E + L\theta)^2} TK \exp[-\alpha L] \qquad [6]$$

T denotes the optical losses of the transmission and receiver system, and K denotes additional losses for incoherent lights sources and is given by the ratio of LED emitting area to detector area, and is $K = 1$ for laser light sources. This equation can also be used to estimate losses for different weather situations and obtain the maximum range L of a specific FSO system.

Recent Research and Future Considerations

Existing FSO systems have impressively demonstrated their versatility and stability, as well as their seamless integration into existing networks, on multiple occasions. A prime example is the re-establishment of communication links between Merrill–Lynch Brokerage and Wall Street in New York, in the aftermath of the terrorist attack on September 11th, which destroyed one of the central fiber optic hubs in downtown Manhattan. The installation of a commercial FSO system by Terabeam ensured a fast replacement and a stable continuation of their work.

The demonstrated performance of such links compares to standard fiber-based links. The inclusion of wavelength division multiplexing as well as other standard or nonstandard techniques in FSO systems proves the strengths of a system, which is based on identical hardware for emitter and detectors as fiber-based networks. Nonetheless, FSO allows a multitude of applications not possible in fibers, which current research projects are targeted at. This includes, for example, long wavelength FSO communication systems, continuously ground-satellite links and free-space quantum communication for extreme secure data transmission – to list only a few of the possibilities. One very obvious example is extreme wavelength division multiplexing, which uses one link operating in all atmospheric windows at the same time. Such a FSO system would not only allow very high data rates and extreme stability against weather effects, but could also employ new kinds of wavelength-based modulation and encoding schemes increasing the security of the data link. Concluding the actual state of the FSO system as well as ongoing research, free space optical communication has all the ingredients needed to be one of the mayor key ingredients of future networks – and based on the increasing demand of bandwidths they will contribute strongly.

List of Units and Nomenclature

c	speed of light
C_n	refractive-turbulence structure coefficient
D_E	diameter of the emitting optics
D_R	diameter of the receiving optics
I	intensity
L	range (i.e., distance between emitter and detector optics)
m	mass
M^2	laser-mode structure parameter
N_s	concentration
P_E	emitted power
P_R	received power
q	charge
T	optical losses of the transmission and receiver system
V	visibility
α	extinction coefficient
ε_0	dielectric constant
θ	divergence of the light beam
λ	wavelength
λ_0	natural oscillation wavelength
σ_s	scattering cross-section

See also

Environmental Measurements: Doppler Lidar; Optical Transmission and Scatter of the Atmosphere. **Imaging:** Infrared Imaging; Lidar.

Further Reading

Andrews LC and Phillips RL (1998) *Laser Beam Propagation through Random Media*. Bellingham, CA: SPIE Optical Engineering Press.

Begley DL (ed.) (1991) *Selected Papers on Free-Space Laser Communications*. SPIE Milestone Series. Bellingham, CA: SPIE Optical Engineering Press.

IEE Proceedings Optoelectronics, *Special Issue on Free Space Optical Communications*. December 1996.

Jha AR (2000) *Infrared Technologies*. New York: John Wiley & Sons.

Killinger D (2002) *Free Space Optics for Laser Communications Through the Air.* Washington: Optics and Photonics News, Optical Society of America, p. 36.

Measures R (1984) *Laser Remote Sensing*. New York: John Wiley & Sons.

Mecherle GS (ed.) (2003) *Free-Space Laser Communication Technologies XV.* Bellingham, CA: SPIE Optical Engineering Press.

Weichl H (1990) *Laser Beam Propagation in the Atmosphere*. Bellingham, CA: SPIE Optical Engineering Press.

Willebrand H and Ghuman BS (2001) *Free-Space Optics: Enabling Optical Connectivity in Today's Networks.* Indianapolis, IN: SAMS.

Wolfe WL and Zissis GJ (1978) *The Infrared Handbook*. Ann Arbor: Enviromental Research Institute of Michigan.

Zuev VE (1988) *Laser Beams in the Atmosphere.* New York: Consultant Bureau.

Lightwave Transmitters

J G McInerney, National University of Ireland-Cork, Cork, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical fiber telecommunications has changed human society forever, providing the capacity for affordable and ubiquitous communication. It provides data rates and transport economies far in excess of those available using purely electronic means. It is doubtful whether the Internet and worldwide web, or the vast infrastructure of wireless cellular telephony, would be viable without optical fiber technology. Its rapid development and acceptance has been driven largely by contemporaneous successes in manufacturable low-loss optical fiber cables, sensitive pin-FET photoreceivers, and reliable high-performance laser diode-based transmitters. The earliest developments in silica glass fibers at Standard Telecommunication Laboratories, UK, (then part of the US-based ITT Corporation) and Corning Glass Works (USA) in the 1960s and 1970s, were pivotal in determining the directions of modern fiber technology, as were later key inventions of rare-earth doped fiber amplifiers at the University of Southampton (UK) and AT&T Bell Laboratories (USA) in the 1980s and early 1990s. However, perhaps the greatest contributors to the success of fiber optics have been semiconductor injection lasers, or laser diodes, which in various forms have made up the vast majority of all fiber optic transmitters, and the entirety of long-haul transmitters, as well as pump sources for the doped-fiber amplifiers which enable modern networks. The fascinating story of the development of fiber optics is to a large extent driven by the dramatic progress in the development of high-performance, reliable semiconductor lasers. These emitted at wavelengths near 1300 and 1550 nm, which respectively correspond to the dispersion and attenuation minima of single-mode silica optical fibers.

Semiconductor Laser Principles

Semiconductor lasers were first demonstrated in research laboratories at General Electric, IBM Corporation and the MIT Lincoln Laboratory (all USA) as early as 1962, although it took almost two decades for the basic science and engineering of materials, fabrication, reliability, and performance design issues to be developed sufficiently for their use in practical communication systems. Early successes were obtained using the GaAs/AlGaAs lattice-matched material system, in which a lightly doped GaAs or ternary compound AlGaAs active layer is sandwiched between two lattice-matched heterojunctions with n- and p-doped AlGaAs layers with higher Al fractions than the active layer, so that their refractive indices are lower and their bandgaps higher, providing optical and charge-carrier confinement. In all cases the material is epitaxially grown, meaning that the entire laser chip forms a single crystal. The gain, and hence the laser emission, occurs at the direct bandgap of the active layer, at wavelengths in the range 750–870 nm depending on the Al fraction.

Later evolutions used the quaternary compound GaInAsP clad by InP guiding layers, where the In and P fractions are chosen to match the lattice constant to the InP substrate.

In general, optical gain in a laser diode is generated by creating an electron-hole plasma in the vicinity of the forward-biased junction, a situation corresponding to population inversion in a conventional laser, and with sufficient forward current the material becomes transparent, in that the gain exactly equals the absorption and scattering losses, at a particular wavelength. When this arrangement is enclosed within a suitable optical cavity, and the forward current is increased further, the system begins to lase, that is to oscillate continuously to produce coherent optical radiation at a wavelength which satisfies two basic conditions: it must be within the set of wavelengths at which the material produces sufficient gain, and it must be close to one of the electromagnetic modes at which the optical cavity is resonant. These conditions, which define lasing threshold, are characterized by the equality of the optical gain and the total losses, that is the material losses (absorption, scattering, and free-carrier plasma) and the cavity losses which include the light output.

Early laser diodes in both the GaAs/AlGaAs and GaInAsP/InP systems were of the edge-emitting type, in which the light is emitted from the edge of the laser chip, perpendicular to the growth direction in the plane of the wafer. Although this geometry complicates laser production flow, particularly the testing function which requires that the wafer be scribed into bars to define the laser output facets, edge-emitting lasers (EELs) make up the vast majority of optical fiber transmission sources. An alternative laser diode structure is the vertical cavity surface-emitting laser (VCSEL), in which emission occurs in the growth direction perpendicular to the plane of the wafer. VCSELs offer major advantages in spectral stability, beam quality, manufacturability and cost, but their output powers are low (a few mW) and they are not yet available at the key telecommunication windows around 1300 nm and 1550 nm, due to difficulty in forming the necessary high-reflectivity Bragg mirrors in the GaInAsP/InP material system. Major initiatives are currently underway to produce long-wave VCSELs in the GaAs/AlGaAs system using InAs quantum dots, or dilute nitrides such as InGaAsN, or hybrid approaches such as InP-based gain media fused to high-reflectivity mirrors using GaAs/AlGaAs, dielectric coatings, air gaps, and others. At present, however, all long-haul optical fiber transmission is based on edge-emitting laser diodes fabricated in GaInAsP/InP.

General Structure and Requirements of Optical Fiber Communication Systems

The first optical fiber communication links were simple point-to-point affairs, consisting essentially of series-connected transmitters, fibers, and receivers. The transmitters were simple Fabry–Perot laser diodes, onto which data were encoded by direct digital modulation of the injection current at rates ~100 Mbit/s. The first commercial optical fiber link was built in 1976, a single fiber cable linking two switching centers of the Illinois Bell Telephone Co just 2.5 km apart in the Chicago metropolitan area, using 850 nm GaAs laser technology. In 1988, the first transatlantic fiber link, AT&T's TAT-8, was completed linking endpoints in New Jersey, England, and France with three fiber pairs carrying signals generated by 1300 nm GaInAsp single mode lasers. In these early systems, data were regenerated frequently by repeaters each consisting of a receiver, signal conditioner and transmitter. Each fiber in TAT-8 carried a single optical frequency modulated at 280 Mbit/s and the construction cost of the system was ~$50 k/km, resulting in an economic figure of merit of over $600 k/Mbit/s. Twenty-five years later, current systems cost the same per kilometer to build in absolute dollars despite inflation, a major effective cost reduction enabled largely by replacement of the expensive repeaters by doped-fiber optical amplifiers. Moreover, each fiber is now highly multiplexed, with the potential for hundreds of wavelengths each carrying 10 Gbit/s signals generated by 1550 nm laser transmitters, so that the unit cost has plummeted to ~$200/Mbit/s, a reduction of three and a half orders of magnitude.

In the remainder of this article, we will describe transmitters for both short- and long-haul systems where the channel spacing is only tens of GHz in the optical carrier frequency. Modern short-haul systems either feed client networks as part of bi-directional transceivers, or drive metro and local access networks connected directly by photonic switches to terminals of the long-haul systems.

DWDM Transceivers

Modern practice uses modular units containing transmitters and receivers allowing bi-directional data flow. In the send mode, data are applied to a long-haul transmitter at the local (client) source and coupled through the external fiber network to the remote long-haul receiver. In the receive mode, data arrive from a remote long-haul transmitter and are coupled (via a regenerator including signal conditioning, clock recovery, and error correcting steps) to a local short-haul transmitter leading to the client's internal network. Each transceiver card therefore contains a long-haul transmitter and receiver set (known collectively as dense wavelength division multiplexing (DWDM) transport devices) and a short-haul transmitter and receiver set (known as client interfaces or local transport devices).

Transmitter Requirements

Short-haul transmitters may be simple Fabry–Perot edge-emitting laser diodes, or vertical cavity laser diodes, or even superluminescent LEDs for sub-Gbit/s data rates. They are generally not required to be at one of the minimum-loss telecom windows near 1550 nm, as the transmission distances are short. However, as client networks grow into large metropolitan networks, it is likely that 1300 nm sources will be used to minimize dispersion. Power requirements are modest, a few mW peak, at data rates ~1–10 Gbit/s. Single mode laser emission and polarization control are not required by present day systems, although polarization-mode dispersion may limit transmission at higher data rates. Both return-to-zero (RZ) and nonreturn-to-zero (NRZ) coding may be used, and direct modulation of the laser amplitude is the norm.

For long-haul transmitters the situation is dramatically different. Operation in one of the standard erbium-doped fiber amplifier (EDFA) bands near 1550 nm is mandatory: the C (conventional) band extends from 1530–1565 nm, the S (short-wave) band from 1460–1530 nm) and the L (long-wave) band from 1565–1625 nm. In the C-band, output powers ~10 mW are generally sufficient, and powers above ~30 mW are limited by nonlinear effects such as self-phase modulation and four-wave mixing in the fiber. In the S- and L-bands, where the available gain from EDFAs is less, lasers may need to operate close to the nonlinear limit, several tens of mW, although various schemes (differential pumping, gain-flattening filters, etc.) are being evaluated to flatten the global EDFA gain spectrum.

In current practice, a single fiber utilizing DWDM, using only the C-band, can carry up to 1 Tbit/s comprising 100 channels at 10 Gbit/s on each channel, the current standard. For such performance a channel spacing of 50 GHz in optical frequency is required. Using all three bands with similar density would enable ~3 Tbit/s, which could be doubled by using 25 GHz channel spacings. Ultimately, even higher channel densities and hence overall system data rates are possible, with a practical limit ~10 Tbit/s. The channel spacing is a key parameter in that it determines the laser design, specifically its spectral stability and linewidth, and the overall transmitter design in that laser chirp (modulation induced dynamic spectral shift) must be less than half the channel spacing. In practice, only single-wavelength and coarse wavelength division multiplexing (CWDM) systems may be modulated directly. All modern DWDM systems require external modulation, for example using electro-absorption or Mach–Zehnder interferometric devices outside the laser cavity. For the latter cases, stabilization of the laser wavelength is required for example by on-chip gratings to form distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, or by external fiber gratings.

Directly Modulated Lasers

Direct modulation of semiconductor lasers is convenient and effective: by modulating the laser injection current, one modulates the carrier density and hence the optical gain, resulting in modulation of the laser output up to ~10 Gbit/s. The actual modulation bandwidth is determined by interactions between photons and carriers, whose respective decay lifetimes, τ_s and τ_p, are determined by cavity losses and total recombination rates. The simplest form of such interactions is described by the photon and carrier rate equations for the injected carrier density N and photon density S in a single lasing mode:

$$\mathrm{d}N/\mathrm{d}t = J/ed - gNS - N/\tau_s \qquad [1]$$

$$\mathrm{d}P/\mathrm{d}t = gNS + \beta N/\tau_s - S/\tau_p \qquad [2]$$

where J is the injected current density, e the electronic charge, d is the thickness of the active region (hence J/ed is the rate of injection of carriers per unit volume), g is the gain coefficient per unit length and β is the fraction of the spontaneous emission coupled into the lasing mode, typically $\sim 10^{-2}$–10^{-5} for edge-emitting lasers (product of geometric and spectral overlap factors). These equations simply state that the rate of change of the carriers or photons is given by the rate of generation less the loss rate. Additional terms are required in the presence of optical feedback or coupling (in which case the photon density is replaced by the complex amplitude and phase of the optical electric field) or for extremely rapid modulation where the traveling wave nature of the disturbances in the photon fields is significant. These so-called traveling-wave rate equations for the photons are of the form:

$$\mathrm{d}S_+/\mathrm{d}t + c\mathrm{d}S_+/\mathrm{d}z = gNS_+ + \beta N/\tau_s \qquad [3]$$

$$dS_-/dt - c dS_-/dz = gNS_- + \beta N/\tau_s \qquad [4]$$

with the same carrier rate equation as [1] using the total photon number S made up of the forward-traveling component S_+ and the backward-traveling component $S_- : S = S_+ + S_-$.

Both sets of rate equations are based on approximations such as homogeneous gain broadening, neglect of transverse field effects, and diffusion. These are valid in most situations but care is required when analyzing lasers with large transverse apertures, or when ultrashort (picosecond) pulses are involved.

Small-Signal Modulation

Small-signal modulation may be analyzed by a linear perturbation analysis of the rate equations, leading to a conjugate pair of poles in the response function, with damped relaxation oscillations at frequency:

$$f_R = (1/2\pi)\sqrt{(gS_0/\tau_p)} \qquad [5]$$

or, in terms of the injected current density J (assuming $S \sim G(J - J_{th})/ed$ with Γ the electromagnetic confinement factor of the mode):

$$f_R = (1/2\pi)\sqrt{[\Gamma g(J - J_{th})/ed]}$$

f_R can be regarded as the resonant frequency for the interaction between the carriers and photons. As a practical matter, lasers can be modulated up to $\sim 2f_R$ but the modulation response rolls off rapidly with increasing frequency above f_R. Using explicit expressions for the threshold gain, we can write f_R as:

$$f_R = (1/2\pi)\sqrt{[(GN_T g\tau_p + 1)(J/J_{th} - 1)/\tau_s\tau_p]} \qquad [6]$$

where N_T is the transparency carrier density $\sim 10^{18}\,\text{cm}^{-3}$. f_R can take values in excess of 10 GHz for a well-designed laser. It should be noted, however, that high modulation bandwidth almost always requires high differential gain dg/dN and short τ_p. Increasing dg/dN requires p-doping (which increases optical loss) and/or some special quantum-confining structure such as wells, wires, or dots. Decreasing τ_p inevitably leads to greater optical loss and thus a higher threshold.

The damping rate of the relaxation oscillations is also important in limiting modulation bandwidth, and also (when increased) in reducing the sensitivity of the laser to optical feedback. For bulk or quantum well lasers, the condition for critical damping is $A^2 = 4B$ with $A = (S/g\tau_s) - \sigma[N - N_T]/g\tau_p$ and $B = \sigma[1 + (S_0/g\tau_s) - (1 - \beta)(N - N_T)]$. For quantum dot lasers there are additional damping terms due to carrier transport and thermalization which may restrict small-signal modulation bandwidths to a few GHz.

Experimentally, small-signal modulation properties may be determined using a simple sampling oscilloscope, low-noise tunable signal generator and spectrum analyzer. A combination of these elements with automated frequency sweeping may be found in a scalar network analyzer acting as an s-parameter test set. The modulated laser is connected to port 1 and a high speed photodetector connected to port 2, then a swept-frequency measurement of the transfer characteristic s_{21} is performed. In addition to the limitations on modulation response due to the carrier-photon resonance, practical lasers are limited by RC parasitics in the laser chip (junction impedance) and its package. For the highest modulation speeds, packages need to be designed as microwave transmission line components with effective impedance matching and low back reflections, as characterized by the voltage standing wave ratio (VSWR).

Large-Signal Modulation

Large-signal direct modulation can be analyzed by numerical integration of the rate equations. In general, its results are beyond the scope of this review, but for effective high-speed response, digital modulation of laser diodes should be carried out with a pre-bias close to threshold. Modulation bandwidth also increases with laser power, which is limited for systems considerations by fiber nonlinearities, and for reasons of laser reliability. In pulse code modulation, care must be taken that the laser is near critical damping, to minimize thermal patterning effects due to long strings of ones (high power) or zeros (low power). Experimentally, large signal modulation is analyzed by constructing eye diagrams, in which pulse traces on a fast oscilloscope are continuously overlaid when the laser is modulated using pseudorandom binary pulses. When the signal quality is high, the high and low digital levels are easily distinguishable, resulting in an open 'eye' in the accumulated traces.

Requirements for Externally Modulated Lasers

The requirement for external modulation occurs when the laser chirp exceeds half the desired channel spacing in the communication system. Chirp occurs due to changes in the refractive index, and hence the optical phase, during modulation. This dynamic phase shift then results in an instantaneous frequency shift.

For direct modulation we have instantaneous wavelength

$$\lambda(t) = \lambda(0)(\mu(t)/\mu(0) \qquad [7]$$

Where $\mu(t)$ is the refractive index at time t. μ shifts with carrier density N due to combinations of several factors (free carrier plasma dispersion, thermal bandgap shrinkage, bandgap renormalization, dynamic Burstein–Moss effects) which give $\mathrm{d}\mu/\mathrm{d}N \sim -3 \times 10^{-20}\ \mathrm{cm}^3$, which for large signal digital modulation can result in tens of GHz. The actual degree of chirp depends on the so-called antiguiding or linewidth enhancement factor a given by

$$\alpha = -2k_0(\mathrm{d}\mu/\mathrm{d}N)/(\mathrm{d}g/\mathrm{d}N) \qquad [8]$$

where $k_0 = 2\pi/\lambda_0$ is the free-space wavenumber. α as defined is positive, typically in the range 2–7 depending on the material, operating carrier density, and wavelength relative to the gain peak. Distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, which use integrated gratings to control the lasing wavelength, can have lower chirp by detuning from the gain peak but therefore are likely to require careful temperature stabilization. Quantum dot lasers, based on 3D nanoclustered active regions, have the potential for very low chirp, narrow linewidths, and reduced wavelength shifts, so that directly modulated DWDM transmitters may be possible.

Integration of laser diodes with electro-absorption or Mach–Zehnder type modulators is achieved in the transmitter optical subassembly and may in some cases be accomplished by monolithic integration on the same chip. Using separate modulators allows each device to be optimized independently but requires optical assembly, alignment, and retention. Mach–Zehnder modulators in X-cut lithium niobate, for example, have essentially zero chirp and can operate to 40 Gbit/s and above, with dynamic extinction ratios (ratio between fully on and fully off) of ~20 dB. Integrated semiconductor M-Z modulators have chirp an order of magnitude less than those of directly modulated lasers, so that channel spacings ~25 GHz are possible in DWDM systems with suitable temperature and wavelength controls. Although the details of such modulators are beyond the scope of this article, the laser requirements are to generate ~ 10 mW of continuous power with stable center wavelength and narrow linewidth. Such lasers are usually mounted in hermetic packages (e.g., the current 14-pin butterfly standard) with integrated thermo-electric cooler, power monitor photodiode, an optical isolator to suppress optical backreflections which cause instabilities and self-pulsing, and optional wavelength locking optics.

Reliability of Lasers in Fiber Optic Systems

From the early days, when laser diode lifetimes were measured in seconds even at cryogenic temperature, enormous progress has been made in achieving materials purity and reducing crystalline defects. Today's fiber optic laser transmitters have projected lifetimes of decades. In common with electronic devices, diode lasers fail at a rate given by a 'bathtub' curve, that is the failure rate $r(t)$ defined as the probability of failure per unit time at time t, has relatively high values at low t (early failures) and high t (wearout failures) and very low values in between. In terms of the population of lasers $n(t)$ we have:

$$r(t) = (-1/n(t))\,\mathrm{d}n(t)/\mathrm{d}t \qquad [9]$$

If Δn is the number of samples which fail in time Δt, then assuming Δt begins at $t = 0$, the effective or average failure rate r_{eff} over the interval is

$$r_{\mathrm{eff}}(t) = \Delta t \Delta n/n(0) \qquad [10]$$

In reliability science it is customary to define r_{eff} in FITs (failures integrated in time) with the time-span chosen to yield statistically significant numbers. For electronic and photonic devices, it is customary to select the time interval $\Delta t = 10^9$ h (1 billion operating hours), so that if 1% of the devices fail in 10 years (~10^5 hours) we have $r_{\mathrm{eff}} \sim 100$ FITs, which is the order of magnitude required by modern fiber optic laser sources.

In terms of actual statistical models, failure rates can be estimated using failure probability density functions $f(t)$. This is related to $r(t)$ by $f(t) = r(t)S(t)$ where $S(t)$ is the cumulative probability of surviving until t. When $r(t)$ is nearly constant, as in early failures due to material defects or process errors, the statistics are approximately exponential:

$$f(t) + (1/\tau)\exp(-t/\tau) \qquad [11]$$

or Weibull-distributed (an exponential is a Weibull distribution with $\beta = 1$):

$$f(t) = (\beta/t^{\beta})t^{b-1}\exp[-(t/\tau)^b] \qquad [12]$$

but this description is not suited to wearout failures for which $f(t)$ shows a significant increase as the population ages. For such cases, the lognormal distribution is often applicable:

$$f(t) = (1/\sigma t\sqrt{}(2\pi))\exp[-(\ln t - \ln \tau)^2/2\sigma^2] \qquad [13]$$

where in each case τ is approximately the mean time to failure (MTTF).

Laser diodes undergo standardized qualification tests to demonstrate sufficient reliability prior to being used in commercial systems. Operationally, failure statistics must be tabulated and the best fit obtained. Burn-in procedures are used to screen early failures. Chip and module failures should be distinguished clearly, as the latter includes many additional factors such as thermal and power management, light coupling, mechanical or chemical integrity.

Given that effective lifetimes of decades are required, it is clearly necessary to accelerate aging to produce statistically meaningful failure rates in reasonable times $\sim 10^3$ h. For the most common modes of laser chip degradation, due to recombination-induced aggregation of defects in the crystalline epitaxial material, these are thermally activated and current driven, so that it is customary to use a modified Arrhenius law at temperature T:

$$\tau \sim (1/J^n)\exp(E_a/k_B T) \qquad [14]$$

where J is the operating current density, n is an index ~ 2, E_a is the activation energy, and k_B is Boltzmann's constant. In terms of thermal acceleration, we have a factor $F = \tau(T_A)/\tau(T_B)$ with:

$$F = \exp[(E_a/k_B)(1/T_A - 1/T_B)] \qquad [15]$$

Typical values of E_a for laser diodes are ~ 1 eV so that acceleration factors $\sim 10^3$ are possible.

Finally, it should be noted that laser diodes are generally subject to catastrophic failure in the event of overdriving or static discharge, the failure mode being thermal facet damage at $\sim 1-10$ MW/cm^2 for continuous-wave operation, the exact value depending on the material, surface preparation and specifically surface state absorption and its temperature coefficient.

Conclusions

Laser diodes are ideal sources for optical fiber communication systems and have propelled the development of fiber optics from its origins in the 1960s to the present day. They are compact, rugged, efficient, and reliable sources of light at key wavelength ranges such as 1300–1310 nm (short haul high speed links) and 1500–1600 nm (long haul amplified systems). They are capable of direct modulation at gigabit rates for simple systems but require external modulation and careful wavelength control for dense wavelength division multiplexed terabit systems. Transmitter lasers for real systems must satisfy stringent reliability conditions.

See also

Lasers: Edge Emitters; Semiconductor Lasers. **Semiconductor Physics:** Quantum Wells and GaAs-based Structures.

Further Reading

Telcordia Technologies (formerly Bellcore): document family FR-796 on Reliability and Quality Generic Requirements for telecommunications equipment. Latest issue is 003 dated Jul 2003. The specific document related to laser reliability is GR-468 – Generic Reliability Assurance Requirements for Optoelectronic Devices used in Telecommunications Equipment. Standards for other network photonic, electronic and mechanical components may be found in the document family FR-2063 Network Equipment-Building System Family of Requirements, issue 003 dated Feb 2003 and specifically FR-357 – Network Equipment-Building System Family of Requirements: Component Design for Reliability.

Adams MJ and Thomas B (1977) Detailed calculations of transient effects in semiconductor injection lasers. *IEEE Journal of Quantum Electronics* QE-13: 580.

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: Wiley.

Alferov ZhI, Andreev VM, Portnoi EL and Trukan MK (1970) AlAs-GaAs heterojunction injection lasers with a low room-temperature threshold. *Soviet Physics – Semiconductors* 3: 1107.

Basov NG, Krokhin ON and Popov YuM (1961) Production of negative-temperature states in p-n junctions of degenerate semiconductors. *Soviet Physics – JETP* 13: 1320.

Bernard MGA and Duraffourg G (1961) Laser conditions in semiconductors. *Physica Status Solidi* 1: 699.

Bhattacharya P (1996) *Semiconductor Optoelectronic Devices*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

Cheng J and Dutta NK (eds) (2000) *Vertical Cavity Surface-Emitting Semiconductor Lasers: Technology and Applications*. London: Taylor & Francis.

Chuang S-L (1995) *Physics of Optoelectronic Devices*. New York: Wiley.

Coldren LA and Corzine SW (1995) *Laser Diodes and Photonic Integrated Circuits*. New York: Wiley.

Dupuis RD (1987) An introduction to the development of the semiconductor laser. *IEEE Journal of Quantum Electronics* QE-23: 651.

Fukuda M (1991) *Reliability and Degradation of Semiconductor Lasers and LEDs*. Norwood, MA: Artech House.

Fukuda M (1998) *Optical Semiconductor Devices*. Wiley.

Guekos G (ed.) (1998) *Photonic Devices for Telecommunications*. New York: Springer.

Hall RN, Fenner GE, Kingsley JD, Soltys TJ and Carlson RO (1962) Coherent light emission from GaAs junctions. *Physical Review Letters* 9: 366.

Hayashi I, Panish MB, Foy PW and Sumski S (1970) Junction lasers which operate continuously at room temperature. *Applied Physics Letters* 17: 109.

Holonyak N Jr and Bevacqua SF (1962) Coherent (visible) light emission from Ga(As1-xPx) junctions. *Applied Physics Letters* 1: 82.

Ito R, Nakashima H, Kishino S and Nakada O (1975) Degradation sources in GaAs-AlGaAs double-heterostructure lasers. *IEEE Journal of Quantum Electronics* QE-11: 551.

Kaminow IP and Li T (eds) (2002) *Optical Fiber Telecommunications IV (two parts)*. Amsterdam: Elsevier.

Kaminow IP and Koch TL (eds) (1997) *Optical Fiber Telecommunications III (two parts)*. London: Academic Press.

Kapon E (1998) *Semiconductor Lasers, (two parts)*. London: Academic.

Kartalopoulos SV (1999) *Introduction to DWDM Technology*. New York: Wiley.

Kartalopoulos SV (1999) *Understanding SONET/SDH and ATM*. IEEE Press.

Kasap SO (2001) *Optoelectronics and Photonics: Principles and Practices*. Englewood Cliffs, NJ: Prentice-Hall.

Kazovsky L, Benedetto S and Willner A (1996) *Optical Fiber Communication Systems*. Norwood, MA: Artech House.

Keiser G (1991) *Optical Fiber Communications*. New York: McGraw-Hill.

Kressel H and Lockwood HF (1974) A review of gradual degradation phenomena in electroluminescent diodes. *Journal de Physique* 35: C3–223.

Lasher G and Stern F (1964) Spontaneous and stimulated recombination radiation in semiconductors. *Physical Review* 133: A553.

Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley.

Morthier G and Vankwikelberge P (1997) *Handbook of Distributed Feedback Laser Diodes*. Norwood, MA: Artech House.

Nathan MI, Dumke WP, Burns G, Dill FH Jr and Lasher G (1962) Stimulated emission of radiation from GaAs p-n junctions. *Applied Physics Letters* 1: 62.

Numai T (2004) *Fundamentals of Semiconductor Lasers*. New York: Springer.

Paoli TL (1977) Changes in the optical properties of cw (AlGa) as junction lasers during accelerated aging. *IEEE Journal of Quantum Electronics* QE-13: 351.

Paoli TL and Ripper TE (1970) Direct modulation of semiconductor lasers. *Proceedings of IEEE* 58: 1457.

Pearsall TP (2003) *Photonics Essentials: An Introduction with Experiments*. New York: McGraw-Hill.

Piprek J (2003) *Semiconductor Optoelectronic Devices: Introduction to Physics and Simulation*. London: Academic.

Quist TM, Rediker RH, Keyes RJ, Krag WE, Lax B, McWhorter AL and Zeiger HJ (1962) Semiconductor maser of GaAs. *Applied Physics Letters* 1: 91.

Reinhart FK, Hayashi I and Panish MB (1971) Mode reflectivity and waveguide properties of double-heterostructure injection. *Lasers* 42: 4466.

Saleh BEA and Teich MC (1991) *Fundamentals of Photonics*. New York: Wiley.

Senior JM (1992) *Optical Fiber Communications*, 2nd edn. Engelwood cliffs, NJ: Prentice-Hall.

Siegman AE (1986) *Lasers*. University Science Books.

Verdeyen JT (1994) *Laser Electronics*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: Wiley.

Zachos TH and Ripper JE (1969) Resonant modes of GaAs junction lasers. *IEEE Journal of Quantum Electronics* QE-5: 29.

Local Area Networks

E Wong, University of Melbourne, Melbourne, VIC, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Local area networks (LANs) are privately owned networks that are used to carry internal traffic loads within an organization. A LAN may exist on many levels categorized by its application and speed. This is illustrated in Figure 1 whereby, at the simplest requirement, a low-speed LAN may be used to interconnect a cluster of personal computers and workstations. Such a scenario may exist in a department within a university campus, or a building within a technology park. If need be, these individual departmental or building LANs may be interconnected using a moderately high-speed backbone LAN. The backbone LAN may also provide interconnection to servers which support mass storage devices that are shared by the individual lower speed LANs. In larger organizations, a much higher speed LAN is linked to the backbone LAN to support bulk data transfer between mainframes, supercomputers, and other office equipment.

The common transmission mediums for traditional low-speed LANs are coaxial cable and twisted pairs. High-speed backbone LANs advocate the use of the optical fiber. The continuing progress in fiber-optic technology and devices, and the availability of appropriate standards have given rise to LANs that can span distances of hundreds of kilometers or more, and that operate at data rates in the order of hundreds to thousands of megabits per second (Mb/s).

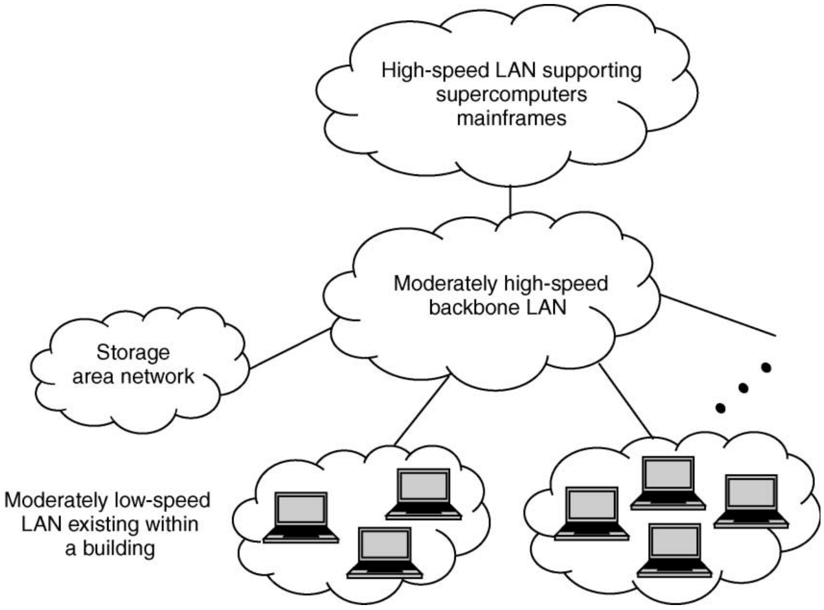
High-speed LAN supporting supercomputers mainframes
Moderately high-speed backbone LAN
Storage area network
Moderately low-speed LAN existing within a building

Figure 1 Interconnection of local area networks (LANs).

The increase in demand, to extend LAN capabilities towards higher data rates and across larger geographical areas, is fueled primarily by the need to provide new bandwidth intensive services such as multimedia teleconferencing, image processing, and streaming video, to a significantly higher number of end-users.

This article focuses on three existing fiber-optic LAN standards, namely Ethernet, Fiber Distributed Data Interface (FDDI), and Distributed Queue Dual Bus (DQDB). While the current uses of FDDI and DQDB are not as widespread as Ethernet, both standards have significantly contributed to the initial deployment of fiber-optics for LAN applications, with standard-based products still commercially available. In general, standards are required to ensure interoperability between components and equipment from different vendors and suppliers. Two key features are addressed by the standards: the physical topology, and the medium access control (MAC) protocol. A network's physical topology defines the way in which the users, or nodes, are interconnected to each other and to the transmission medium. Common optical LAN topologies are the star (e.g., Ethernet), bus (e.g., DQDB and Ethernet), and ring (e.g., FDDI). In comparison, the MAC protocol controls network access such that its capacity is efficiently and fairly shared amongst all nodes.

Access control can be either centralized or decentralized. In the former, a node wishing to transmit must wait until it receives permission from a controller, whereas in the latter, nodes collectively determine the transmission order. Further, capacity allocation to each node can be either synchronous or asynchronous. The synchronous approach guarantees a specific capacity to each node, e.g., using time division multiple access (TDMA), whereby nodes are allowed to transmit only during a designated time slot. In contrast, the asynchronous approach allows capacity to be dynamically allocated to each node. The asynchronous approach can be further categorized into round robin, reservation, and contention schemes. With round robin, nodes take it in turn to transmit in a cyclic manner and a node may transmit up to a maximum amount of data during its turn. With the reservation scheme, time on the medium is divided into slots, and a node wishing to transmit must reserve future slots by transmitting additional control information into the network. With the contention scheme, nodes randomly access the network. The Ethernet standard, discussed in the following section, is based on the contention scheme.

Ethernet

Overview

The IEEE (Institute of Electrical and Electronics Engineers) 802.3 standards committee first standardized 10 Mb/s Ethernet in 1985. It is now the most widely used technology for high-speed LANs, with standards extended to cater for operating speeds of 100 Mb/s, 1 Gb/s and 10 Gb/s, and encompassing the use of the optical-fiber transmission medium. Low-speed 10 Mb/s and 100 Mb/s Ethernet LANs use the carrier-sense multiple access with collision detection (CSMA/CD) MAC protocol, whereas a switched technique is used for the 1 Gb/s and 10 Gb/s options.

In CSMA/CD, a node wishing to transmit an Ethernet frame first listens to the transmission medium by detecting the presence of signals on the channel (i.e., carrier-sense). This process allows the node to determine if another transmission is in progress. If the medium is in use, the node continues to listen until the medium is idle, and then transmits immediately. While transmitting, the node continues to listen to the medium. It may happen that two or more nodes may attempt to transmit at the same time. If this occurs, there will be a collision, and data from both transmissions will become garbled.

A collision can be indicated by a combined signal that exceeds a predefined CD threshold voltage at the carrier-sense receiver, taking into account the collision between frames from nodes that are located furthest apart. If a collision is detected during transmission (i.e., collision detection), the transmitting node ceases transmission and proceeds to transmit a brief jamming signal to notify all nodes of the collision. After transmitting the jamming signal, the node must back off before attempting to transmit the same data frame again. The back-off period is random, and is determined by a binary exponential back-off algorithm. Every time a node attempts to retransmit, the mean value of the random back-off period is doubled until the tenth attempt. After that the node is allowed to retransmit a further six times with the same mean random delay before discarding the frame and reporting an error to higher control layers.

The transmission medium remains unusable when frames collide but only until the collision is detected. As such, the minimum transmission time of a frame must be longer than the maximum propagation delay of the network so that collision is detected prior to the end of transmission. If shorter frames are used, then collision detection does not occur, and CSMA/CD thus exhibits the same performance as its less efficient predecessor protocol, CSMA. Accordingly, the CSMA/CD standards specify a minimum frame length and maximum network size, which place fundamental limitations on the practical application of a pure CSMA/CD LAN.

IEEE 802.3 10 Mb/s Ethernet

In 1993, the IEEE 802.3 committee extended the 10 Mb/s Ethernet standard to include the optical-fiber transmission medium. The standard notation of this option is 10BASE-F. Each transmission link comprises a pair of graded-index 62.5 μm core multimode optical fibers (MMF), one for each direction of transmission. The transmission wavelength is 850 nm. 10BASE-F can be further categorized into three specifications. The first, denoted 10BASE-FP (fiber passive star), specifies a broadcast star topology that can interconnect up to 33 nodes and repeaters. Each of these devices is attached to a passive star coupler via two optical-fiber links with distances up to 500 m. The passive star coupler splits the optical power of any incoming frame equally amongst its output links, allowing all nodes or repeaters in the network to receive the frame, albeit at a lower optical power level.

The second and third specifications, denoted 10BASE-FL (fiber link) and 10BASE-FB (fiber backbone) respectively, define point-to-point links over distances of up to 2 km. 10BASE-FL supports both nodes and repeaters, whereas 10BASE-FB supports only repeaters. A repeater is a device that detects the optical signals on its incoming link, performs electrical signal regeneration, and then optically retransmits the signals on its output link. Repeaters used in 10BASE-FB have an additional function in which the detected signals are retimed with a local clock before being retransmitted. This prevents the propagation of timing distortions, allowing distances of up to 15 cascaded repeater links to be implemented. The characteristics of the three specifications for 10BASE-F, along with the 100 Mb/s and 1 Gb/s specifications to be discussed in the following sections, are summarized in Table 1.

IEEE 802.3 100 Mb/s Ethernet (Fast Ethernet)

Fast Ethernet refers to the standard developed by the IEEE 802.3 committee to provide low-cost, Ethernet-compatible LANs operating at 100 Mb/s. The optical-fiber transmission medium option is denoted 100BASE-FX. Originally defined for FDDI (covered in the section on FDDI below), 100BASE-FX specifies the use of two 62.5 μm MMF links, each operating at a unidirectional rate of 100 Mb/s. While 10 Mb/s Ethernet can only operate in half-duplex mode, whereby a station can either transmit or receive a frame but cannot do both simultaneously, Fast Ethernet can operate in either half- or full-duplex modes. A node operating in full-duplex mode can simultaneously transmit and receive data frames.

In order to deploy full-duplex Fast Ethernet, all repeaters must be replaced with switching hubs. Figure 2a illustrates a repeater located in the central point in an active star topology. Each node is attached to the repeater via two optical links. The repeater retransmits frames received from any one of its input

Table 1 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Mb/s, 100 Mb/s and 1 Gb/s Ethernet

	10BASE-FP	*10BASE-FL*	*10BASE-FB*	*100BASE-FX*	*1000BASE-SX*	*1000BASE-LX*
Data rate	10 Mb/s	10 Mb/s	10 Mb/s	100 Mb/s	1 Gb/s	1 Gb/s
Transmission wavelength	850 nm	850 nm	850 nm	850 nm	850 nm	1310 nm
Transmission medium	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	62.5 μm MMF pair	50 μm or 62.5 μm MMF pair	10 μm SMF, 50 μm or 62.5 μm MMF pair
MAC protocol	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)	CSMA/CD (half duplex) Switched (full duplex)

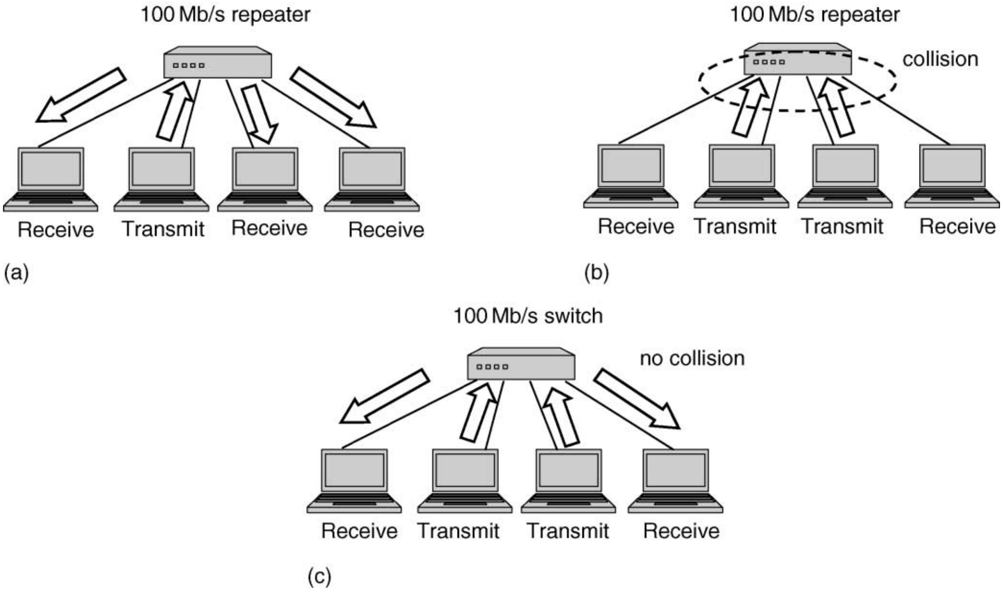

Figure 2 (a) Shared LAN repeater. (b) Collision at a shared LAN repeater due to simultaneous transmission of frames. (c) LAN switch (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

links to all of its output links. If more than one node transmits at the same time, then collisions will occur at the repeater, as illustrated in Figure 2b. A switching hub, on the other hand, switches an incoming frame from a particular input to an appropriate output line. The same switching function can be performed simultaneously on other input–output links without interference to the frames involved, as illustrated in Figure 2c. As such, each node has a dedicated capacity equal to that of the transmission rate of the LAN, and the switching hub can support mixed capacity links, e.g., 10 Mb/s and 100 Mb/s. With dedicated access to the medium and appropriate buffering at the switch to avoid contention of the same output link, there is no need to implement the CSMA/CD access algorithm. Nonetheless, the conventional Ethernet frame format is used to ensure inter-operability between existing 10 Mb/s and 100 Mb/s LANs.

IEEE 802.3 1 Gb/s Ethernet (Gigabit Ethernet)

The IEEE 802.3 1 Gb/s Ethernet standard, commonly referred to as Gigabit Ethernet, was finalized in 1998. An example of a Gigabit Ethernet backbone LAN is shown in Figure 3, whereby a Gigabit Ethernet backbone switch connects central servers to high-speed workgroup switches. In turn, each workgroup switch supports gigabit high-performance workgroup servers, as well as 100 Mb/s workstations and hubs. As in the 100 Mb/s standard, Gigabit Ethernet uses the conventional Ethernet frame format and allows both half-duplex and full-duplex operation. For half-duplex operation, the CSMA/CD access protocol is

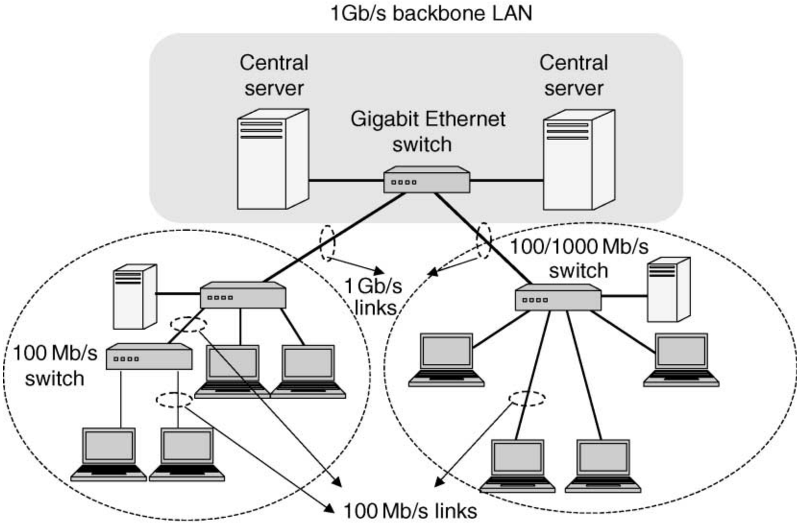

Figure 3 Gigabit Ethernet backbone network (adapted from William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education).

equipped with two additional features. The first feature, known as carrier extension, extends the propagation time of short frames by adding nondata symbols to the end of frames. The minimum length of a frame for Gigabit Ethernet is 512 bytes as opposed to the specified minimum of 64 bytes in the preceding two standards. The second feature, known as frame bursting, allows a node to transmit multiple consecutive frames, up to a specified maximum, without having to undergo the contention process for each transmission. Both features improve the efficiency of the protocol.

There are two optical-fiber transmission medium specifications for Gigabit Ethernet. The short wavelength option, denoted 1000BASE-SX, is for 850 nm transmission. Specified maximum link distances are 275 m using 62.5 μm MMF, and 550 μm using 50 μm MMF. The long wavelength option, denoted 1000BASE-LX, is for 1310 nm transmission. Achievable duplex link distances are 550 m using 62.5 μm or 50 μm MMF, and 5 km using 10 μm SMF.

IEEE 802.3 10 Gb/s Ethernet (10GbE)

The Ethernet standard was extended to incorporate 10 Gb/s (10GbE) in 2002, fueled by ever-increasing Internet traffic and bandwidth intensive applications. 10GbE fulfills not only LAN specifications, but also metropolitan and wide-area network specifications. It is envisioned that with the emergence of 10GbE metropolitan and wide-area networks, the deployment of 10GbE LANs will enable a seamless transport of Ethernet frames across different network boundaries, thus simplifying network management and optimizing operational costs.

Unlike its predecessor standards, the physical layer options for 10GbE include only the optical-fiber transmission medium and allow only for full-duplex operation. Four specifications are defined for 10GbE. 10GBASE-S is designed for 850 nm transmission on 63.5 μm MMF, while 10BASE-L and 10BASE-LE use 10 μm single mode fiber (SMF) for 1310 nm transmission and 1550 nm transmission, respectively. The corresponding maximum achievable distances are 300 m, 10 km, and 40 km, respectively. An alternative specification, 10GBASE-LX4, is designed for wavelength division multiplexed (WDM) transmission links. The specification caters for four 2.5 Gb/s transmission channels within the 1270–1355 wavelength region on either 10 μm SMF, 50 μm MMF, or 62.5 μm MMF. Single-mode transmissions can reach distances of up to 10 km, while multimode transmissions can reach distances up to 300 m. The characteristics of these physical layer specifications are summarized in Table 2.

Fiber Distributed Data Interface (FDDI)

Overview

The Fiber Distributed Data Interface (FDDI) standard was developed by the Accredited Standards Committee X3T9.5, and approved by both ANSI (American National Standards Institute) and ISO (International Standards Organization) in 1989. While not as

Table 2 Summary of IEEE 802.3 physical and MAC layer specifications for 10 Gb/s Ethernet

	10GBASE-S	*10GBASE-L*	*10GBASE-E*	*10GBASE-LX4*
Data rate	10 Gb/s	10 Gb/s	10 Gb/s	4 × 10 Gb/s
Transmission wavelength	850 nm	1310 nm	1550 nm	1270 nm to 1355 nm
Transmission medium	62.5 μm MMF pair	10 μm SMF pair	10 μm SMF pair	10 μm SMF or 50 μm, 62.5 μm MMF pair
MAC protocol	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)	Switched (full duplex)

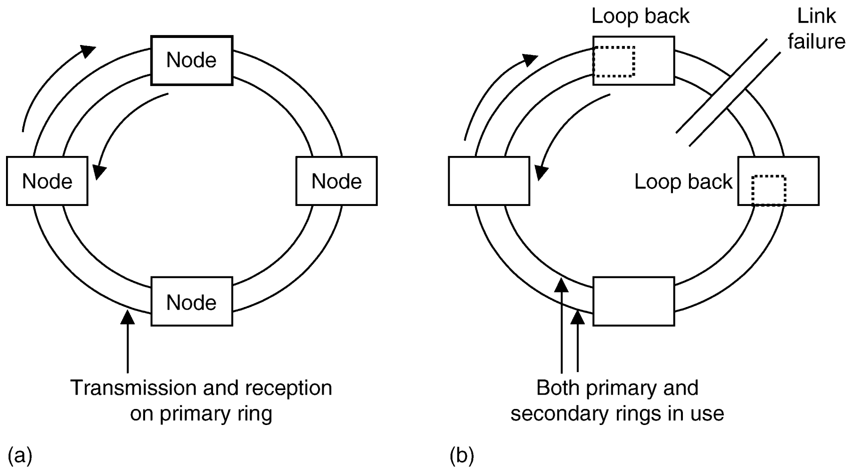
Node
Node
Node
Node
Transmission and reception on primary ring
(a)
Loop back
Link failure
Loop back
Both primary and secondary rings in use
(b)

Figure 4 (a) FDDI ring network. (b) Use of secondary ring for protection against link failures.

widely deployed as the Ethernet family of LANs, due to the high cost of equipment and devices, it remains as the first to be standardized for high-speed fiber-optic LAN applications and is now mainly used to interconnect existing FDDI-compatible LANs. Extensions have been made to the FDDI standard to incorporate single-mode fiber operation for increased signal reach.

FDDI specifies a dual counter-propagating ring network as shown in Figure 4a. Each ring operates at the speed of 100 Mb/s, and can reach a maximum length of 100 km. The dual ring topology is self-healing against fiber-cuts and node failures. Under normal operation, frame transmission and reception are performed on a primary ring. The secondary ring is utilized to form a closed loop when a node or link failure occurs, as illustrated in Figure 4b. In FDDI, the interface between each node and the optical-fiber transmission medium is a repeater. Data frames, each of which carries its source and a destination address in a defined field in the frame header, are transmitted sequentially around the ring from one repeater to another. Each frame can vary in length by up to 4550 bytes. Each repeater copies the destination address field of all incoming frames and, upon recognizing the destination address as its own, additionally copies the remainder of the frame. The repeater attached to the source node also facilitates the removal of frames it has transmitted after a round-trip propagation around the ring.

FDDI employs a distributed timer-controlled, token-passing mechanism, referred to as timed token rotation protocol, to control access to the shared optical fiber ring. While Ethernet-based LANs are highly suited to support throughput sensitive services which can sustain reasonable delay in return for high throughput, FDDI provides an additional capability to support delay sensitive services such as packetised voice and video. For such services, frames must be received within a certain time interval for quality of service requirements. A node connected to an FDDI network can therefore transmit both classes of services: the former referred to as burst traffic and the latter as stream traffic. In addition, the timed token rotation protocol can limit the amount of frames that can be transmitted by a particular node for each class of service, ensuring efficient and fair network access amongst all nodes and services. FDDI is thus well-suited to carry the mixture of burst and stream traffic expected in a backbone network which supports a number of smaller LANs.

MAC Protocol

Token-passing mechanism

In FDDI, access to the shared transmission medium is governed by a timed token rotation protocol. A token is a special frame created during network (re)-initialization by a pre-assigned node. The token circulates the ring network and gives the node that holds it the right to transmit. Figure 5 gives an example of a single frame transmission from Node A to Node C, and from Node B to Node A. Node A has a frame to transmit to Node C and thus waits for a token (Figure 5a). Upon recognizing that a token frame is passing through, it aborts the token transmission without regenerating the token bits. After Node A has completely received the token, the node begins transmitting frame FA addressed to Node C, and appending a new token upon completion (Figure 5b). Meanwhile, Node B regenerates and retransmits FA as it circulates past. When FA reaches Node C, the node copies FA in addition to regenerating and retransmitting FA. Node B captures the token issued by A and transmits FB to Node A (Figure 5c). Eventually FA recirculates back to its originating node, Node A, whereby it is removed from the ring (Figure 5d). At the same time, Node B finishes transmitting FB and releases the token. Nodes C and D, with no frames to transmit or receive, regenerate and retransmit both FB and the token. When FB reaches Node A, it is copied by the node. Node A also regenerates and retransmits both FB and the token. FB is finally removed from the ring by Node B (Figure 5e), and the token remains in the ring to be received by the next node with a frame to transmit (Figure 5f).

Aside from source and destination address fields, each dataframe also contains a frame status field which comprises error detected (E), address recognized (A), and frame copied (C) indicator bits. Each node can check passing frames for errors and can set the E indicator if an error is detected using the standard cyclic redundancy check procedure. If a node detects its own address, it sets the A indicator, and upon copying the frame, the station may also set the C indicator. When a dataframe recirculates back to its source node, the indicator bits are examined. If both A and C are not set, then the destination node does not exist or is inactive. On the other hand, if the frame has not been copied by an active destination node or an error has been detected, higher control layers are notified for future retransmission.

Time-controlled mechanism

FDDI uses a time-controlled mechanism to prevent nodes from dominating the network capacity. During initialization, a target token rotation time (TTRT) is established based on the shortest allowable time between token arrivals at each node. The same TTRT value is then stored at all nodes. In addition, each node has a maximum capacity guaranteed for the transmission of its stream traffic. The total transmission time of this stream traffic and the latency of the ring (i.e., total round-trip propagation time, the transmission time of a maximum length frame and the token frame) must never exceed the TTRT. During network operation, each node measures the

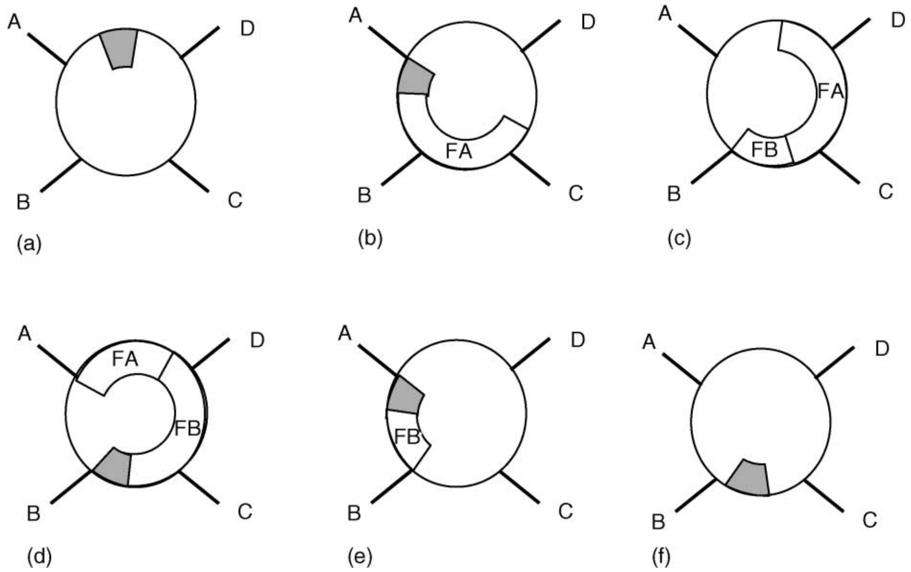

Figure 5 Fiber Distributed Data Interface (FDDI) token passing mechanism.

time interval between successive token arrivals, using a token-rotation timer (TRT). The TRT is enabled only upon a token arrival, and expires if the counter reaches the TTRT value. In turn, a late counter (LC) is incremented and TRT is reset to zero. The number of TRT expirations since a token was last received is, therefore, counted by LC. Any token that arrives at a node with a nonzero LC value is referred to as a late token. Likewise, an early token is one which arrives at a node with a zero LC value.

The action of a node when it receives a token depends on whether the token is early or late. When a late token is received, LC is reset to zero but TRT continues to run. The node then transmits for the period of time which corresponds to the predetermined bandwidth for steam traffic transmission. The node is not allowed to transmit any bursty traffic frames. In the event of an early token reception, the node transfers the current value of TRT to a second timer, the token-holding timer (THT) but does not enable it. TRT is reset to zero and is enabled to track the next rotation of the counter. The node then transmits stream traffic frames for the allocated period of time. After that, THT is enabled and bursty traffic frames are transmitted until THT reaches the TTRT value. The THT is then reset and the node releases the token into the ring. The transmission of bursty frames may be further controlled by two additional mechanisms. The restricted-token mechanism allows only two nodes in the network to use all the available bursty bandwidth, and restricts all other nodes to their stream bandwidth. This is facilitated by the use of a restricted token which is issued by one of the two nodes. The priority mechanism provides each station with the capability to transmit bursty frames according to eight priority levels. Each priority level corresponds to a different time threshold in the THT.

Note that the time-controlled mechanism of the MAC protocol guarantees a maximum response time for the ring because, in the worst case, the time between the arrival of two successive tokens will never exceed twice the value of TTRT. Most multimedia applications, such as video conferencing, real-time video, and entertainment video can be supported on FDDI with stream service and a small TTRT. However, since the TTRT cannot be less than the ring latency, applications requiring time bounds less than twice the ring latency cannot be supported by FDDI. Long and continuous isochronous services which must adhere to strict periodic access require FDDI-II.

FDDI-II

FDDI-II provides the additional circuit-switching capability to support constant bit-rate telecommunication applications. Like FDDI, the transmission bit-rate of FDDI-II is 100 Mb/s. FDDI-II can operate in either complete isochronous mode, complete packet-switching mode, or in hybrid mode in which both isochronous as well as bursty and stream services are supported. Strict periodic access required by isochronous services are accomplished with transmission opportunities repeated every 125 μs. At 100 Mb/s, time on the ring is divided into several frames, each 1562.5 bytes long. Out of the 1562.5 bytes, 1560 bytes are used for data transmission while the remaining 2.5 bytes are used as interframe gap. In turn, the 1560 bytes are divided into 16 wideband channels of 96 bytes each. Each wideband channel thus provides a bandwidth of 6.144 Mb/s, sufficiently adequate for television broadcast and multiple voice channels. For operation in the hybrid mode, some of the wideband channels are allocated for packet mode transmission. Likewise, if complete packet-switching operation is required, all wideband channels are allocated for packet switching transmission.

IEEE 802.6 Distributed Queue Dual Bus (DQDB)

Overview

The IEEE 802.6 Distributed Queue Dual Bus (DQDB) standard, finalized in 1990, specifies a dual bus network with an operating speed of 150 Mb/s. DQDB is now mainly used for MAN applications, although it can be readily deployed for LAN backbone interconnection. Figure 6 illustrates the basic architecture of a DQDB network with access nodes, Node 1 to Node *N*, distributed along two counter propagating buses, BUS A and BUS B. Each access node is considered to be connected to a lower speed LAN. At the head end of each bus, a slot generator continuously generates fixed-size time slots which propagate along its respective bus. The contents of each time slot are regenerated at each access node it passes, and upon reaching the end of the bus, a sink terminal receives and removes the time slots from the network. Data are transported across the network from one access node to another in the form of fixed-size segments. A node transmits a segment by changing the contents of a passing time slot. Each segment needs only be transmitted on one of the two buses, the choice depending on the locations of the transmitting and receiving nodes. For example, referring to Figure 6, a downstream transmission from Node i to Node $i+1$ requires the use of BUS A. On the contrary, an upstream transmission from Node i to Node $i-1$ uses BUS B.

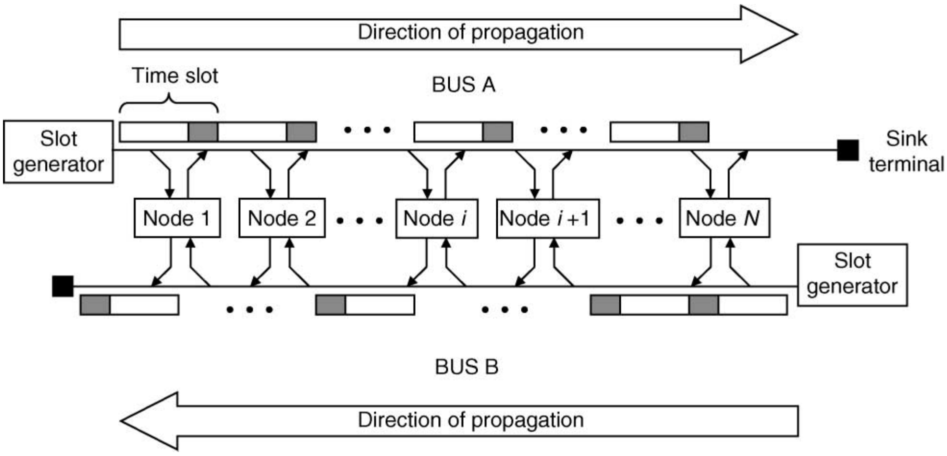
Direction of propagation
Time slot
BUS A
Slot generator
Sink terminal
Node 1
Node 2
Node *i*
Node *i* +1
Node *N*
Slot generator
BUS B
Direction of propagation

Figure 6 Distributed Queue Dual Bus (DQDB) architecture.

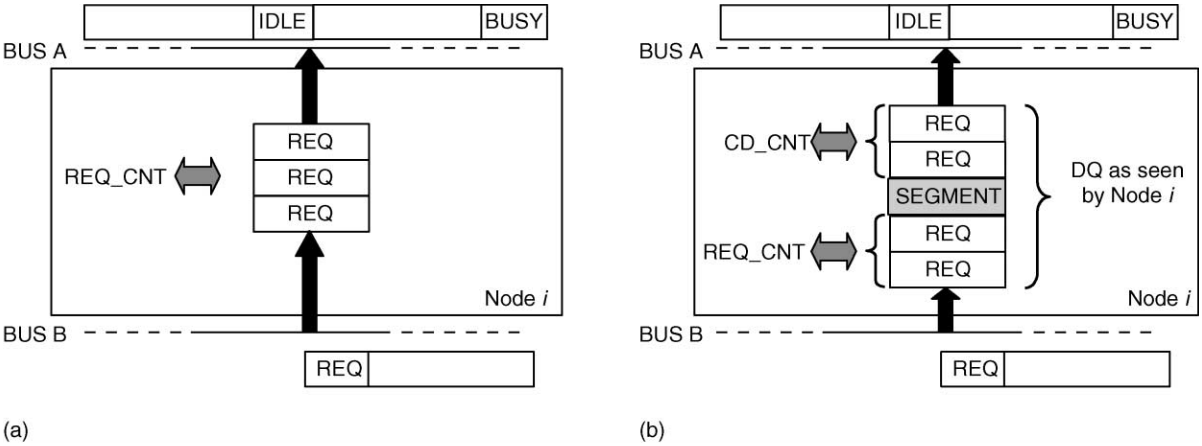
IDLE
BUSY
BUS A
REQ_CNT
REQ
REQ
REQ
Node *i*
BUS B
REQ
(a)
IDLE
BUSY
BUS A
CD_CNT
REQ
REQ
SEGMENT
REQ_CNT
REQ
REQ
DQ as seen by Node *i*
Node *i*
BUS B
REQ
(b)

Figure 7 Behavior of Node *i* during (a) idle state; and during (b) countdown state.

MAC Protocol

In DQDB, permission to transmit segments into the network is determined by a global distributed queueing algorithm. The algorithm allows each node to continuously monitor both buses, counting the requests and available time slots that pass by. In addition to maintaining a global first-in-first-out queue (DQ) for each bus, the algorithm relies on information in the header of time slots, namely the BUSY and REQ fields, and counters at the access nodes to coordinate transmissions. A ‘BUSY = 0’ field represents the availability of a time slot for the transport of a segment. For example, time slots generated at the head end of each bus have default ‘BUSY = 0’ and ‘REQ = 0’ fields. The BUSY field is set to 1 when a segment is transmitted into an available time slot.

The REQ field is used in conjunction with the counters to inform nodes of an impending transmission from downstream. To simplify the explanation, consider only packets propagating on BUS A. A node wishing to transmit a segment downstream needs to queue the segment in the appropriate DQ. The node with an impending transmission must then notify all upstream nodes that it has inserted a segment in the DQ so that an available time slot can be reserved for the transmission. This is accomplished by setting the first time slot propagating in the opposite bus, BUS B, with REQ = 0 to REQ = 1. As such, the node maintains the outstanding segments in the DQ belonging to downstream nodes via the value in the REQ field.

To illustrate, Figure 7a shows an arbitrary node, Node *i*, in an idle state, whereby the node has no segments to transmit downstream on BUS A. With every ‘REQ = 1’ observed on BUS B, a request counter, REQ_CNT is incremented by one. However, with every empty slot, i.e., ‘BUSY = 0’, observed on

Table 3 Comparison of standardized local area networks

	Ethernet	*Fast Ethernet*	*Gigabit Ethernet*	*10GbE*	*FDDI*	*DQDB*
Data rate	10 Mb/s	100 Mb/s	1 Gb/s	10 Gb/s	100 Mb/s	150 Mb/s
MAC protocol	CSMA/CD	CSMA/CD or Switched technique	CSMA/CD or Switched technique	Switched technique	Timed token rotation	Distributed queue
Standard	IEEE 802.3	IEEE 802.3	IEEE 802.3	IEEE 802.3	ANSI-FDDI	IEEE 802.6

the BUS A, REQ_CNT is decremented by one. When Node *i* has a segment to transmit, it enters the countdown state as illustrated in Figure 7b. The value in REQ_CNT is transferred into an independent countdown counter, CD_CNT, and REQ_CNT is reset to zero. Node *i* also sets 'REQ = 1' in the opposite bus, BUS B. As in the idle state, each 'REQ = 1' observed on BUS B increments REQ_CNT while each 'BUSY = 1' observed on BUS A decrements it. When CD_CNT reaches zero, Node *i* transmits its segment in the first time slot with 'BUSY = 0' on BUS A and sets 'BUSY = 1'. The node returns to the idle state if there are no other segments to transmit. The algorithm allows the node to determine the number of time slots reserved by downstream nodes and the time to transmit a segment into an available time slot. Note that both the idle and countdown states are replicated for transmissions on BUS B.

The DQDB standard also specifies a limited support of isochronous services. Nodes which provide and receive such services are reserved a certain amount of time slots which are indicated with additional header information. The remaining time slots are used by all other access nodes for the transport of packet-switched traffic, and access to these time slots is managed by the global distributed queueing algorithm. A priority mechanism based on using multiple DQs, each with a different priority level, may also be implemented to guarantee access to certain segments in the network.

Summary

Due to the explosive demand for high-bandwidth applications, fiber-optics has become an essential technology in LANs. Table 3 compares the various fiber-optic LAN standards discussed in this article. The Ethernet family remains the most prevalent of LAN technologies. Its widespread use is contributed by the availability of low-cost chipsets, mature and familiar management and analysis tools. The future of fiber-optic LANs operating beyond 10 Gb/s will depend upon WDM solutions. The current usage of one transmission channel per network is impeded primarily by the limitation of the speed of electro-optic conversion, and thus does not fully exploit the massive transmission bandwidth of the optical fiber. With WDM, multiple transmission channels are concurrently transmitted on a single optical fiber, thereby greatly increasing the capacity of an optical LAN. Ongoing research efforts have been directly towards advancing WDM component technology and addressing prevailing issues such as polarization control, chromatic dispersion, and four-wave mixing among others.

List of Units and Nomenclature

CD_CNT	countdown counter
CSMA/CD	carrier-sense multiple access with collision detection
DQ	global first-in-first-out queue
DQDB	distributed queue dual bus
FDDI	fiber distributed data interface
LAN	local area network
LC	late counter
MAC	medium access control
MAN	metropolitan area network
MMF	multimode fiber
REQ_CNT	request counter
SMF	single mode fiber
TDMA	time division multiple access
THT	token-holding timer
TRT	token rotation timer
TTRT	target token rotation timer
WDM	wavelength division multiplexing

See also

Information Processing: Optical Neural Networks.

Further Reading

IEEE Std 802.6 (1990) IEEE standards for local and metropolitan area networks: distributed queue dual bus (DQDB) subnetwork of a metropolitan area network (MAN).

IEEE Std 802.3 (2002, Revision of IEEE Std 802.3, 2000 ed.) IEEE Standard for Information Technology – Telecommunications and Information Exchange Between Systems – Local and Metropolitan Area

Networks – Specific Requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications.

ISO 9314-1,2 (1989) Information processing systems – Fibre Distributed Data Interface (FDDI) – Part 1: Token Ring Physical Layer Protocol (PHY) and Part 2: Token Ring Media Access Control (MAC).

Marco C, Enrico G and Luciano L (1997) *Metropolitan Area Networks*. New York: Springer-Verlag.

Mills A (1995) *Understanding FDDI*. Englewood Cliffs, NJ: Prentice Hall.

William S (2004) *Data and Computer Communications*, 7th edn, Ch. 15–16. Upper Saddle River, NJ: Pearson Education.

Optical Time Division Multiplexing

L P Barry, Dublin City University, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

As the demand for bandwidth continues to grow, driven by the massive increase in Internet usage, so will the necessity to have communications networks that can handle very high data rates. The use of optical fiber networks is the clear choice for such systems given the huge available bandwidth of the fiber transmission medium. However, in a basic optical communication system comprising a laser transmitter, an optical fiber transmission medium, and a receiver, the capacity is essentially limited by the speed at which light can be modulated at the transmitter. To overcome this limitation, which is basically due to the speed of available electronics, it is necessary to use optical multiplexing techniques, and one such technique, known as optical time division multiplexing (OTDM), will be the subject of this article.

Principles of Time Division Multiplexing

Time division multiplexing (TDM) has long been the traditional method for electrically combining information channels. If we take the most fundamental data rate to be that of a simple voice call at 64 kbit/s, then the transmission of data at higher bit rates is achieved by electrically multiplexing a large number of 64 kbit/s channels in the time domain. With the evolution of standards, a number of different data rates have been specified as standard transmission rates. In Europe the synchronous digital hierarchy (SDH) has a basic data rate of 155.52 Mbit/s, which is known as the synchronous transport module – Level 1 (STM-1). This particular data rate is essentially obtained by electrically multiplexing over 2000 voice calls in the time domain (with some of the capacity required for overhead information). By subsequently multiplexing a number of STM-1 channels together we can obtain transmission at the higher standard data rates of STM-4 (622 Mbit/s), STM-16 (2.48 Gbit/s), and STM-64 (9.88 Gbit/s), as outlined in Table 1.

Figure 1 illustrates how basic electrical TDM is used in standard optical communication systems. Clearly as we approach the higher data rates of STM-64, a serious level of electrical multiplexing is required, and as the data rates increase so does the cost and complexity of the electrical equipment at the transmitter and receiver. Indeed the present state of the art in electronics seems to suggest 40 Gbit/s as the limit for electrically multiplexed communication systems. However, as we stated earlier, communications traffic has been growing explosively over the last decade and will continue to do so. In order to meet this demand for capacity, and better exploit the massive available bandwidth of optical fiber, it is necessary to use optical multiplexing techniques for communications systems. The two main optical multiplexing techniques available are wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM). WDM essentially involves transmitting data at a number of different wavelengths on the same fiber. Although electrical multiplexing may be limited to data transmission rates of around 40 Gbit/s using one laser, by multiplexing together *N* different wavelength channels each carrying 40 Gbit/s, we can achieve overall data rates up to and beyond a terabit/s.

The second of these optical multiplexing techniques, OTDM, is the subject of this article. Whereas WDM multiplexes optical data channels in the

Table 1 Standard data rates for SDH transmission systems

SDH standard	*Data rate (Mbit/s)*
STM-1	155.52
STM-4	622.08
STM-16	2488
STM-64	9953
STM-256	39 813

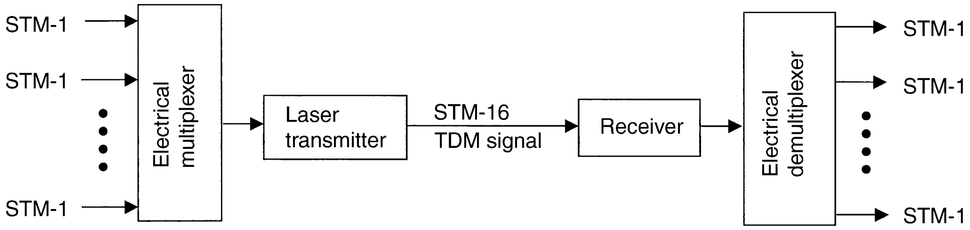
STM-1
STM-1
STM-1
Electrical multiplexer
Laser transmitter
STM-16 TDM signal
Receiver
Electrical demultiplexer
STM-1
STM-1
STM-1

Figure 1 Electrical time division multiplexing (ETDM) in a basic optical communication system.

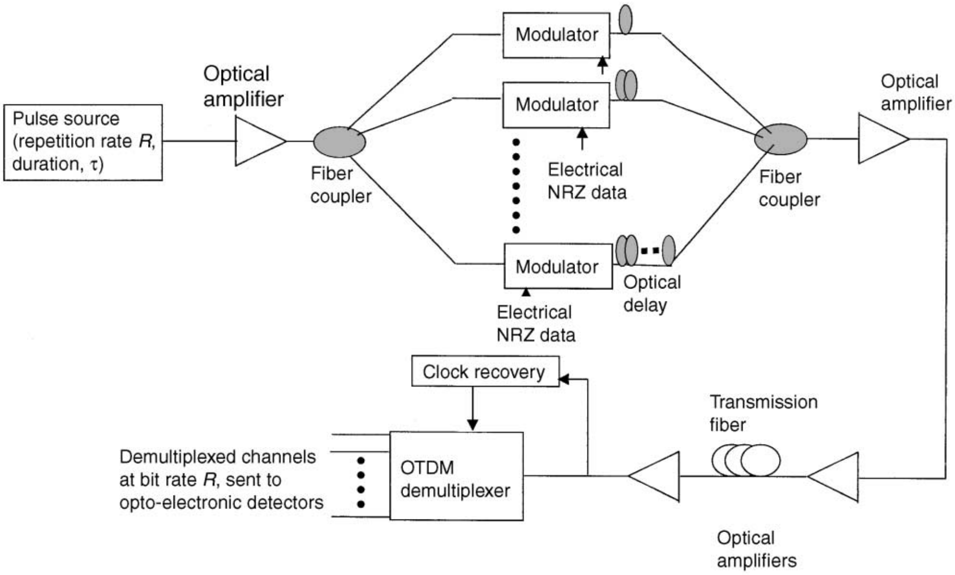
Pulse source (repetition rate *R*, duration, τ)
Optical amplifier
Fiber coupler
Modulator
Modulator
Modulator
Electrical NRZ data
Electrical NRZ data
Optical delay
Fiber coupler
Optical amplifier
Clock recovery
Transmission fiber
Demultiplexed channels at bit rate *R*, sent to opto-electronic detectors
OTDM demultiplexer
Optical amplifiers

Figure 2 Basic configuration for an OTDM transmission system comprising transmitter, transmission fiber, and receiver.

wavelength domain, the basic principle of OTDM communications is to increase the system capacity by multiplexing optical data channels in the time domain. This multiplexing is usually achieved by a process known as bit-interleaving. This process can be explained by considering Figure 2, which shows a basic schematic of an OTDM based transmission system. The main component of the overall system is a source of ultrashort optical pulses (pulse duration τ) at a certain repetition rate, R. The optical pulse source is initially split into N channels using a passive fiber coupler, and each pulse train is subsequently modulated by electrical data which is at a data rate of R. The resulting output from each modulator is essentially an optical data channel where the data are represented using ultrashort optical pulses (return-to-zero data format). The data from each modulator then passes through a fixed fiber delay line which delays each channel by a time equal to $1/RN$ relative to its adjacent channel (as shown in Figure 2). The N modulated and delayed optical data channels are then recombined in another passive fiber coupler to form the OTDM data signal. We can consider that the delay lines essentially assign each data channel to a specific bit slot (of width $1/RN$) in the overall multiplexed signal. The multiplexed data signal may then be transmitted over optical fiber before arriving at the receiver which is responsible for demultiplexing the optical signal into its discrete channels.

The duration of the optical pulse source is extremely important in determining the maximum overall data rate which can be achieved. The overall data rate is basically defined by the temporal separation between channels in the multiplexor, but in order to avoid cross-channel interference, this separation must be significantly greater than the pulse duration. Thus to increase the overall data rate we must use shorter optical pulses. However, as we reduce the optical pulse width to raise the data rate we need to take into account the problems that may be encountered as this high-speed data signal propagates over optical fiber, and also the difficulty in demultiplexing a high-speed OTDM data signal. The following section will look in greater detail at these key elements of the OTDM communication system, namely the optical pulse source, the transmission of the ultrahigh capacity signal, and the demultiplexing at the receiver.

Key Elements of an OTDM Communication System

Ultra-Short Optical Pulse Sources

As stated earlier, the optical pulse source is a key element in any OTDM-based communication system. The important characteristics of the pulse source that will affect its usefulness in an OTDM system are the pulsewidth, the spectral width, and the temporal jitter. The pulse duration clearly has to be short enough to support the desired overall transmission rate. For example, if we wish to design an OTDM system with an aggregate data rate of 100 Gbit/s (which will have a delay of 10 ps between each channel in the multiplexed signal), then the pulsewidth should normally be less the 30% of the channel spacing to avoid cross-channel interference (i.e. around 3 ps). For terabit/s OTDM systems we would require pulsewidths less than 0.5 ps. The spectral width of the pulse source is also important, as it will have a major impact on how the pulse will evolve during propagation in the fiber. A standard figure of merit which is employed is the time–bandwidth product, and ideally we require the pulse source to be transform limited, which implies that the spectral width is as small as possible for the associated pulsewidth. The impact of temporal jitter on the pulses can be easily understood by considering the above example of a 100 Gbit/s system, employing 3 ps pulses, with the multiplexed channels spaced by 10 ps. Obviously if the jitter on the pulses becomes significant in relation to the channel spacing, then this can also lead to interference between adjacent channels in the overall OTDM system. A large number of techniques have been employed to develop ultrashort pulse sources suitable for use in OTDM systems, but the three main methods which will be described below are active mode-locking, gain-switching, and external modulation of a CW light signal.

Active mode-locking of laser diodes normally involves modulating the amplitude of the optical field inside the laser cavity at a frequency which is equal to the mode spacing of the laser. This can be achieved by applying an electrical sinusoidal signal at the correct frequency, and results in the generation of optical pulses at the repetition rate of the applied signal. This technique has been successfully employed in the generation of subpicosecond pulses at repetition rates up to and beyond 40 GHz, with excellent spectral and temporal jitter characteristics. However, an inherent problem in all mode-locked pulse sources is the difficulty in synchronizing the mode-locked frequency to a specific SDH standard data rate.

Gain switching of semiconductor laser diodes is probably the simplest and most reliable technique to generate optical pulses. The technique, which is presented in Figure 3, involves applying a high-power electrical pulse (or electrical sinusoidal signal) to the laser in conjunction with a certain bias current. By ensuring that the electrical pulse signal and the bias signal have the correct level, the relaxation oscillation phenomenon of the laser results in the production of optical pulses with durations between 10 and 30 ps, at the repetition rate of the applied electrical signal. The frequency of the electrical signal applied to the laser is essentially arbitrary (provided it is not larger than the modulation bandwidth of the diode), thus making it straightforward to synchronize the optical pulse train to a SDH line-rate. The main problem with this technique is that the spectral width of the pulses generated is such that the pulses are far from transform-limited, which would affect their subsequent propagation in the fiber. In addition, temporal jitter on gain-switched pulses can also be a problem. However, by using novel arrangements such as external injection into the gain-switched laser, this difficulty can be overcome.

The third pulse generation technique mentioned above involves external modulation of a cw light signal with an electro-absorption modulator. The experimental configuration for this pulse generation technique is shown in Figure 4. By biasing the modulator around its null point, and applying an electrical sinusoidal signal to it, the cw light passing through the modulator becomes shaped into optical pulses. The optical pulse train which is generated due

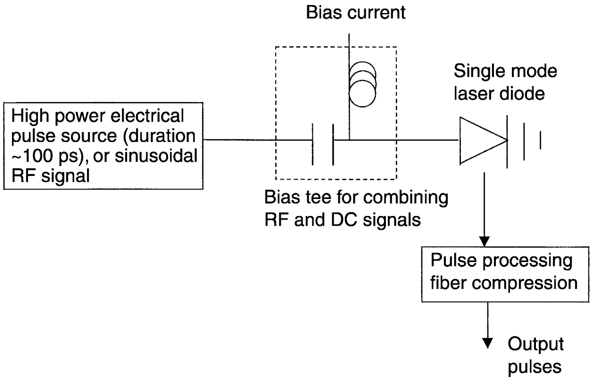
Bias current
Single mode laser diode
High power electrical pulse source (duration ~100 ps), or sinusoidal RF signal
Bias tee for combining RF and DC signals
Pulse processing fiber compression
Output pulses

Figure 3 Pulse generation using the gain-switching technique followed by pulse compression.

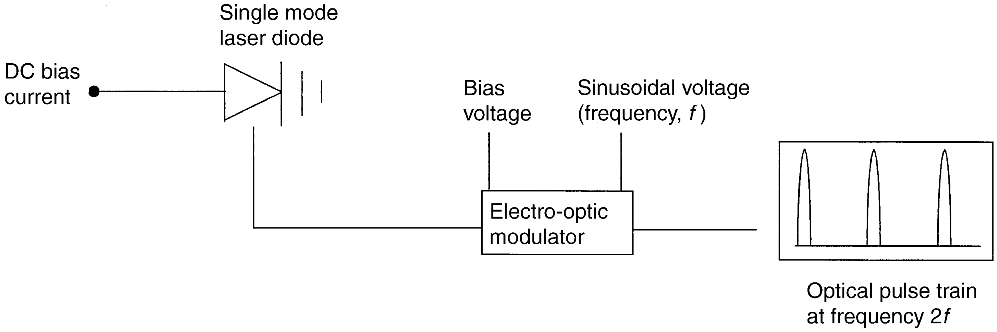

Figure 4 Pulse generation based on shaping of cw light using the nonlinear response of an external modulator biased about its null point.

the nonlinear response of the electro-absorption modulator is at a repetition rate of twice the applied electrical signal. The pulses generated are normally transform limited with extremely low temporal jitter, and the repetition rate is arbitrary as in the gain-switching technique (with the limit being ultimately determined by the modulation bandwidth of the modulator). This method may be readily used to generate pulses at repetition rates up to 40 GHz with pulsewidth around 5 ps, and it has the advantage that the optical source and the modulator can be integrated in a single device to form a compact pulse source suitable for OTDM communication systems.

It should also be noted that in addition to the various techniques that have been outlined, it is possible to use pulse compression in order to reduce the pulse width. The main issues concerned with pulse compression are the shape and spectral width of the optical pulses after compression. One of the most attractive methods of achieving pulse compression involves using nonlinear compression in dispersion-decreasing fiber, with the main advantage of this technique being the ability to maintain a transform limited pulse after compression. By employing this compression scheme, optical pulse sources at 10 GHz with pulsewidths below 200 fs have been developed, and such pulsewidths would be suitable for use in Tbit/s OTDM systems.

Transmission of an OTDM Signal over Fiber

The transmission performance of an OTDM data signal is vital in determining the distance over which the data can be transmitted successfully. The main fiber parameters, which will affect the signal transmission, are attenuation and dispersion. If we consider OTDM systems operating at a wavelength of around 1550 nm (minimum loss wavelength), to overcome the fiber attenuation and maintain a suitable optical power budget for the system, optical amplifiers are normally employed. If we thus assume that the amplifiers overcome the fiber loss problem, then the maximum transmission distance will be limited by the fiber dispersion. In a very basic OTDM system operating at 1550 nm, with transmitter and receiver linked using standard fiber (dispersion parameter of about 16 ps km^{-1} nm^{-1}), the maximum transmission distance will be limited by the overall data rate, and the pulsewidth and spectral width of the optical pulse source. For example, consider a 40 Gbit/s OTDM data signal which is formed using 8 ps optical pulses with a spectral width of 40 GHz (0.32 nm). From the spectral width we can calculate the signal broadening due to dispersion to be around 5 ps km^{-1}, and as the pulses broaden and spread into adjacent channels of the OTDM signal then this interference will make it increasingly difficult to correctly detect the signal at the receiver. In this case after propagation through 5 km of fiber, the signal pulses will have already dispersed to around 25 ps duration, the same value as the temporal bit slot into which each channel is placed. This will clearly result in serious loss of signal integrity. A straightforward possibility to increase the transmission distance of OTDM systems is to employ dispersion shifted fiber; the dispersion parameter is now around 1–2 ps km^{-1} nm^{-1}, at an operating wavelength of 1550 nm. This reduction in dispersion will obviously increase the allowed transmission distance by about an order of magnitude for the example described above. However, to develop ultrahigh-speed, long-haul OTDM communications we require more complex transmission schemes. Two possibilities for this include dispersion compensation, and soliton transmission techniques.

Dispersion compensation basically involves compensating for the dispersion that has been encountered during transmission by using some fiber with a total dispersion of opposite sign but equal magnitude to the transmission fiber. Dispersion compensation may also be achieved using a suitably designed fiber grating. For a long-haul OTDM system, a dispersion compensator may be used every 50 or 60 km, in conjunction with the optical amplifier, thus allowing us to compensate both fiber loss and attenuation periodically along the link. The main limitation, however, with the dispersion compensation technique is caused by the dispersion slope of the fiber, as for ideal compensation it is necessary to compensate completely for the dispersion slope in addition to the overall fiber dispersion. The second technique that may be employed to greatly extend the transmission distance of high-speed OTDM communication systems is the use of soliton transmission. The basic principle of soliton transmission is to use optical data pulses with a particular shape, pulsewidth, and peak power, such that as the pulse propagates, the effects of fiber dispersion and nonlinearity counterbalance to allow the signal to propagate undistorted. By using optical amplifiers to ensure that the optical pulse peak power does not vary too much along the transmission link, OTDM transmission at data rates greater than 100 Gbit/s, over distances approaching 500 km have been demonstrated.

Demultiplexing of OTDM Signal at Receiver

For OTDM communication systems with data rates above 40 Gbit/s, it is not feasible to use electrical switching. Indeed for ultrahigh-speed systems operating at 100 Gbit/s and beyond, the only solution for demultiplexing is to use all-optical switching in which optical control signals are used to switch the OTDM data signal. All-optical switches normally use nonlinear optical effects either in optical fiber or semiconductors. A typical scenario (as presented in Figure 5) involves the injection of both the OTDM data signal and an optical control signal into the nonlinear device. The control signal consists of high-power ultrashort pulses at the repetition rate of the individual channels within the temporally multiplexed signal, and by synchronizing it with one of the OTDM channels it is possible to demultiplex this channel from the high-speed OTDM signal.

Two of the most popular all-optical demultiplexers are the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD). The NOLM is based on the nonlinear refractive index of optical fiber. It essentially consists of a 2 × 2 fiber coupler with its two outputs joined using a certain length of fiber. When an OTDM signal is injected into an input port of the coupler, it splits into two counterpropagating components in the fiber loop, and when these components recombine and interfere at the coupler, the overall signal is output through its initial input port. If we now inject a high-power control signal directly into the loop, such that it propagates unidirectionally, and is

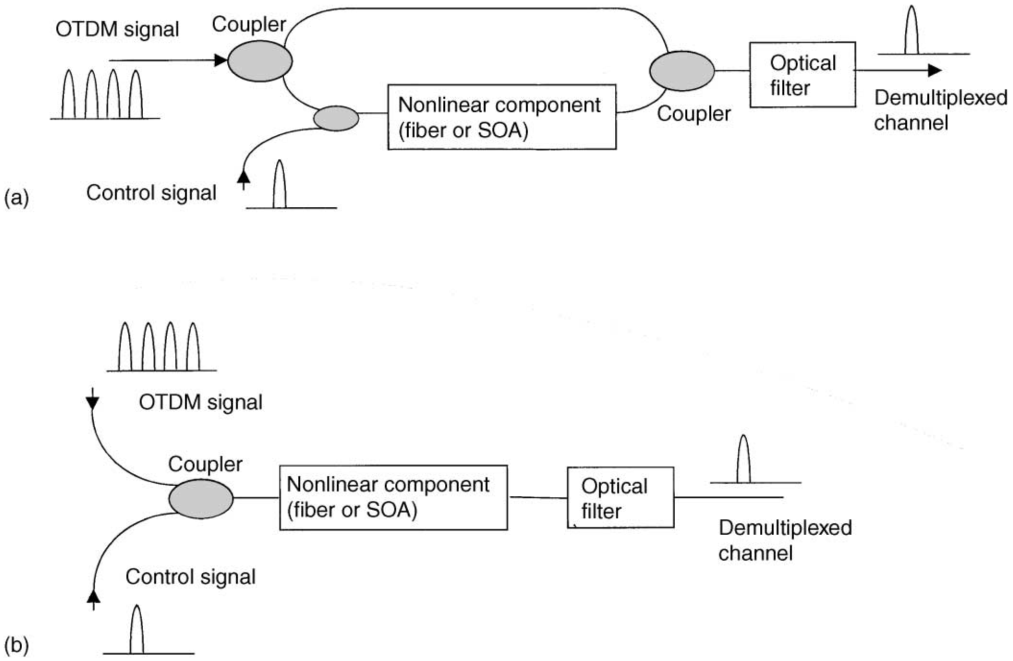

Figure 5 Configuration of OTDM demultiplexers using (a) an interferometer based on nonlinear phase shift in fiber or semiconductor optical amplifiers, and (b) four-wave mixing in fiber or semiconductor optical amplifiers.

synchronized with one of the OTDM data channels, then the phase shift induced by the control on the copropagating signal channel (via the nonlinear refractive index of the fiber) results in that particular channel being switched out to the second input port of the coupler. The remaining data channels of the OTDM signal are once again output through the same port that they entered the coupler. As this particular nonlinear effect in fiber (known as the Kerr effect) has a femtosecond response time, it should be possible to realize extremely high-speed switching with the NOLM. Demultiplexing at data rates in excess of 100 Gbit/s has already been achieved using this technique. One disadvantage of this method is that the nonlinear index coefficient of standard fiber is very small, so in order to achieve the required phase shift from the control pulse, it is necessary to use a fiber loop of around 1 km in length (depending on the power of the control pulse used). This length may affect the overall stability of the demultiplexer and its ability to be readily integrated into high-speed OTDM systems. However, it should be noted that recent developments in the design of high-nonlinearity fibers may reduce this problem.

The second all-optical switch mentioned above is known as a TOAD. Whereas the NOLM is based on the nonlinear refractive index of the fiber, the TOAD is based on a nonlinear optical effect in semiconductor optical amplifiers (SOA). The overall setup for a TOAD is very similar to that for a NOLM, except that a semiconductor amplifier, as shown in Figure 5, replaces the length of fiber in the loop. As for the NOLM, the OTDM data signal is injected into the TOAD using one input of the coupler, such that it splits into two counterpropagating signals, while the control signal propagates unidirectionally in the loop containing the SOA. By synchronizing the control with one channel in the OTDM signal, this particular channel is switched out. Although the response time of a TOAD-based switch may not be as fast as the NOLM, it does have the major advantage that it can be developed into a very compact demultiplexer, suitable for deployment in OTDM-based communication systems.

In addition to the NOLM and the TOAD structures, which are both interferometric-based switches, it is also possible to use four-wavemixing (FWM) in optical fiber or semiconductor optical amplifiers to carry out the demultiplexing of OTDM signals. In this case, the high-power control pulse is synchronized with one of the OTDM channels such that the wavelength of this channel is shifted as it passes through the nonlinear device. Optical filtering may then be used to select out the demultiplexed channel.

As we discussed in the section on demultiplexing, whatever specific all-optical switching technique is employed, we need to generate an optical clock signal for use as the control pulses. This implies that it is necessary to extract a clock at the base rate of the individual data channels from the high-speed OTDM signal. Depending on the base data rate, which may be anywhere from 2.5 to 40 Gbit/s for typical OTDM systems, different clock recovery techniques have been demonstrated. These range from basic electronic clock recovery schemes to more advanced techniques based on self-pulsating laser diodes and optical phase-locked loops.

OTDM Networking Issues

The previous section has predominantly dealt with the basic elements required to implement a high-speed OTDM transmission link. However, if we are to use OTDM technology for high-speed networking, then additional elements will be required. One of the most important of these is an add/drop multiplexer (ADM). The ADM allows us to switch out one channel from the OTDM signal, and then insert another data channel for transmission in the vacant bit slot. If we consider employing OTDM in a ring configuration network, then a typical architecture for the network will be as shown in Figure 6. The high-speed OTDM data signal propagates around the closed fiber loop, and each node in the network is responsible for switching out the data it requires, in addition to inserting the data it wishes to transmit onto the fiber ring. Add/drop multiplexing is thus imperative in developing such an OTDM ring network.

The configuration for each node in an OTDM ring network is shown in Figure 7. At each node it is initially necessary to recover the base rate clock signal. The recovered clock is then used in conjunction with the OTDM signal to switch out the required data channel from the multiplexed signal. When an add/drop multiplexer is used for this purpose we obtain two outputs, one being the demultiplexed data, and the other being the OTDM signal less the switched-out data channel. A simple fiber coupler can subsequently be used to allow this particular node to insert the data it wishes to send into the free bit slot of the overall OTDM signal. This obviously requires control of the relative delay between the OTDM signal and the local data channel to be inserted onto the fiber ring, to ensure that the local data are inserted into the vacant bit slot. A number of techniques have been used for implementing add/drop multiplexers in

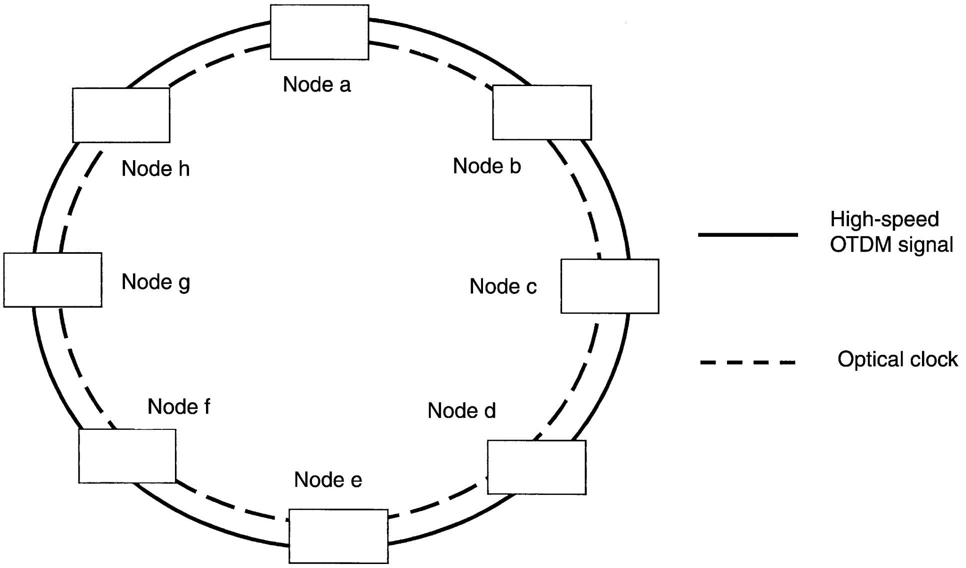
Node a
Node b
Node c
Node d
Node e
Node f
Node g
Node h
High-speed OTDM signal
Optical clock

Figure 6 Schematic representation of a ring network employing OTDM with a separate fiber for clock distribution.

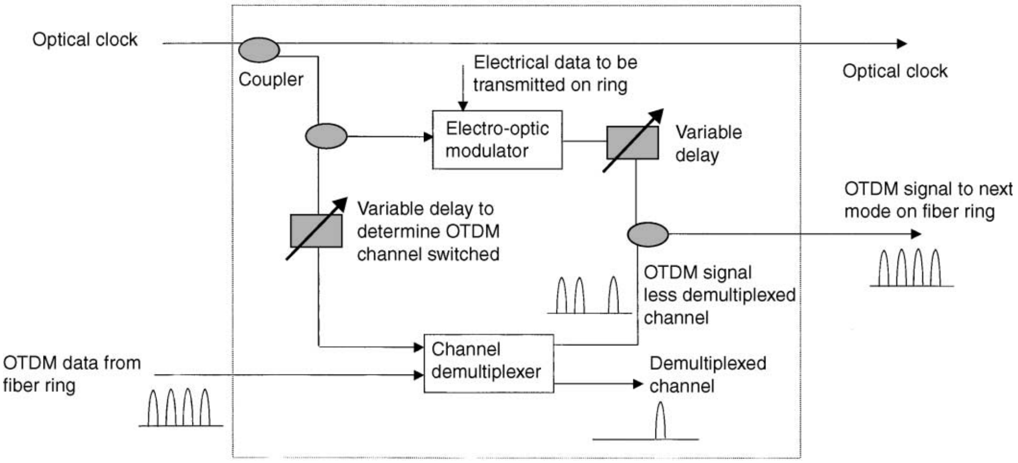
Optical clock
Coupler
Electrical data to be transmitted on ring
Electro-optic modulator
Variable delay
Optical clock
OTDM signal to next mode on fiber ring
Variable delay to determine OTDM channel switched
OTDM signal less demultiplexed channel
Channel demultiplexer
OTDM data from fiber ring
Demultiplexed channel

Figure 7 Configuration of an add/drop multiplexing node for use in OTDM-based communication systems.

OTDM-based networks. These have used either fiber interferometers, or interferometers using semiconductor optical amplifiers, in which we obtain both the demultiplexed data signal and the OTDM signal less the switched-out channel (to which the local data channel is then added). The continuing development and enhancement of ADM devices for OTDM systems will be vital for the future implementation of OTDM communication systems.

Conclusion

In this article we have explained the main ideas involved in building OTDM communication systems, and also introduced the principal technologies which are required to do so. OTDM is an extremely wavelength-efficient technique for delivering high-capacity data signals, and it may be used both for long-haul transmission and networking around metropolitan areas. In addition, unlike WDM, it does not require a different wavelength source for each channel in the multiplexed signal. Despite these and other advantages, OTDM is still way behind WDM when it comes to commercial maturity, with no OTDM systems currently available in the telecommunications market. However, the technologies required to implement high-speed OTDM systems (outlined in Table 2), including ultrashort pulse sources, clock extraction circuitry, and demultiplexing devices, are developing rapidly. It therefore seems

Table 2 Main elements of an OTDM communications system

Device	*Principal function*	*Technical implementation*
Optical pulse source	Generate ultrashort optical pulses suitable for transmission in high-speed OTDM systems	Gain-switching, mode-locking, or external modulation techniques
Multiplexer	Multiplex individual pulse data channels in the temporal domain using appropriate delays to achieve an OTDM signal	Passive couplers with fixed fiber delay lines, or planar waveguide circuits (for accurate control of delays)
Optical amplifier	Amplify the OTDM signal during fiber transmission to overcome fiber loss	Fiber doped with rare earth materials (e.g. erbium) in conjunction with pump laser
Transmission fiber	Transmit the OTDM data signal from transmission site to required receiver	Standard single-mode fiber, dispersion shifted fiber, or other specialty fiber may be used
Clock recovery	Extract clock (at base rate of individual channels) from OTDM data signal at receiver, for use in demultiplexing	Electronic or optical phase-locked loops, self-pulsating lasers, or mode-locked ring laser techniques
Demultiplexer	Demultiplex one of the channels from the overall OTDM signal	Interferometric or four-wave mixing techniques based on fiber or semiconductor nonlinearity
Add/drop demultiplexer	Extract one channel from OTDM data signal, and insert a local channel into the empty time slot for transmission	Similar technical implementations to demultiplexer described above

likely that in the future, OTDM technology may be used for enhancing the overall capacity of optical communication systems.

One possibility for employing OTDM systems would be for upgrading installed WDM networks. The current method for improving WDM networks involves using more wavelength channels, but it may also be feasible to increase the overall capacity by developing hybrid WDM/OTDM communications systems. In such a system, each wavelength channel may carry aggregate data rates in excess of 100 Gbit/s, achieved using OTDM. The design of these hybrid networks would permit switching of the optical data signals to be carried out at two different levels in the overall network. The WDM signals could be switched coarsely using passive filtering devices, and the OTDM data channels could be switched with fine granularity using one of the all-optical switching techniques available.

List of Units and Nomenclature

Dispersion	[ps km^{-1} nm^{-1}]
Transmission data rate	[bit s^{-1}]
ADM	Add drop multiplexer
CW	Continuous wave
FWM	Four-wave mixing
NOLM	Nonlinear optical loop mirror
OTDM	Optical time division multiplexing
SDH	Synchronous digital hierarchy
SOA	Semiconductor optical amplifier
STM	Synchronous transport module
TOAD	Terahertz optical asymmetric demultiplexer
WDM	Wavelength division multiplexing

See also

All-Optical Signal Regeneration. **Optical Communication Systems:** Basic Concepts; Wavelength Division Multiplexing.

Further Reading

Agrawal GP (1997) *Fibre Optic Communications*, 2nd edn., pp. 329–335. New York: Wiley.

Cole M (2000) *Introduction to Telecommunications: Voice, Data, and the Internet*, chapter 5. Englewood Cliffs, NJ: Prentice Hall.

Islam M (1992) *Ultrafast Fiber Switching Devices and Systems*. Cambridge, UK: Cambridge University Press.

Kawanishi S (1998) Ultrahigh-speed optical time-division-multiplexed transmission technology based on optical signal processing. *IEEE Journal of Quantum Electronics* 34: 2064–2079.

Midwinter JE and Guo YL (1992) *Optoelectronics and Lightwave Technology*, pp. 148–178. Chichester, UK: Wiley.

Mukherjee B (1997) *Optical Communications Networks*, pp. 493–506. New York: McGraw-Hill.

Nakazawa M, Kubota H, Suzuki K, Yamada E and Sahara A (2000) Ultrahigh-speed long distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363–396.

Prucnal PR, Santoro MA and Sehgal SK (1986) Ultrafast all-optical synchronous multiple access fiber networks.

IEEE Journal of Selected Areas of Communication 4: 1484–1493.

Spirit DM, Ellis AD and Barnsley PE (1994) Optical time division multiplexing: systems and networks. *IEEE Communications Magazine* 32(12): 56–62.

Spirit DM, Blank LC and Ellis AD (1995) Optical time division multiplexing for future high capacity network applications. In: Smith DW (ed.) *Optical Networking Technology*, pp. 54–77. London: Chapman and Hall.

Tucker RS, Einstein G and Korotky SK (1988) Optical time-division multiplexing for very high bit rate transmission. *IEEE Journal of Lightwave Technology* 6: 1737–1749.

Wavelength Division Multiplexing

J Bowers and H F Chou, University of California, Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Wavelength division multiplexing (WDM) has enabled a revolution in communications technology. This article describes the technology, critical components of WDM systems, and transmission impairment and limits to transmission distance and capacity.

Multiplexing

There are three techniques used in fiber optics to multiplex the information that is transmitted from one point to another. The first is space division multiplexing (SDM), which simply consists of having multiple fibers in a cable or conduit. In a typical installation, there may be between 8 and 1000 fibers in parallel. The second approach is time division multiplexing (TDM) (*see* **Optical Communication Systems:** Optical Time Division Multiplexing), which consist of interleaving multiple data streams into one higher bit rate stream. For example, 672 voice channels (DS0) at 64 kbit/s are multiplexed into one T1 data stream at 1.544 Mbit/s. Eighty-four of these T1 channels are then multiplexed into a SONET (Synchronous Optical NETwork) OC-3 signal at 155 Mbit/s. The third multiplexing approach is wavelength division multiplexing (WDM), which consists of sending multiple signals at different wavelengths. These signals can be used to fill up the entire low loss transmission band of an optical fiber (Figure 1). The low loss transmission band extends from 1200 nm to 1600 nm, which corresponds to over 40 THz. One approach is to pack the signals densely together, as in dense wavelength division multiplexing (DWDM). This is a relatively expensive, but very high capacity approach. The other approach is to space the signals widely apart, (coarse WDM or CWDM) resulting in very low cost components, but with fairly limited transmission capacity.

Dense Wavelength Division Multiplexing (DWDM)

Most DWDM systems use multiple beams spaced at 100 GHz spacing centered at 193.1 GHz as defined by an International Telecommunications Union (ITU) standard (Figure 2a). Some higher capacity systems use finer spacing, 50, 25, or even 12.5 GHz spacing to pack more information into the fiber. The 100 GHz spacing used in DWDM systems corresponds to 0.8 nm spacing at 1.55 mm. In principle, 40 THz of information can be sent down an optical fiber. If high spectral efficiency (~1 bit/Hz) is maintained across the band, then this would correspond to 40 Tbit/s. Practical DWDM systems are limited by the bandwidth of erbium-doped fiber amplifiers (EDFAs) and by signal impairments discussed below. Consequently, most systems utilize a single band, typically the C-band extending from 1530 nm to 1565 nm. Additionally some transmission systems use multiplexers to combine other bands, typically C and L bands for additional transmission capability. The wavelength definitions are shown in Figure 1. Subbands are often used within a band to facilitate adding and dropping a group of wavelengths within a band. Typically 40 10 Gbit/s signals spaced at 100 GHz are used to fill the C-band. Higher capacity

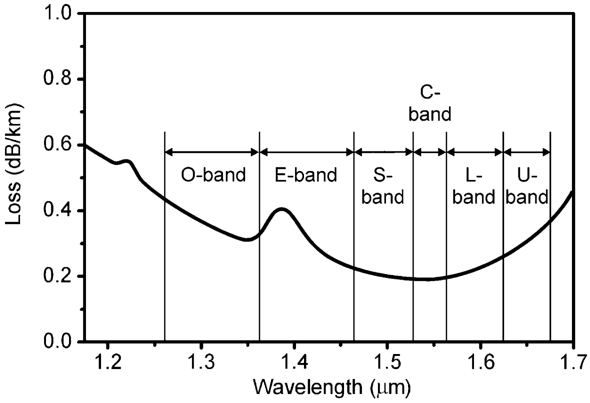

Figure 1 Transmission spectrum of optical fiber. Also shown are the standard transmission bands.

systems with 80 to 200 wavelengths have also been deployed commercially.

The historical evolution of optical transmission systems can be seen in Figure 3. The lower line shows the increasing bit rate demonstrated in experimental TDM systems, which began in 1978 with 45 Mbit/s systems and has now reached 1.28 Tbit/s. The upper line shows what has been demonstrated with DWDM with over 10 Tbit/s transmission demonstrated.

Coarse Wavelength Division Multiplexing (CWDM)

The fine frequency requirements of DWDM systems result in significant cost for transmitters, multiplexers and other components. A low-cost approach is to space the colors widely apart. The ITU standard is 20 nm spacing from 1271 to 1611 nm (Figure 2b). Note that the spacing is uniform in wavelength, but varies in frequency because the frequency spacing is related to the wavelength spacing by

$$\Delta f = -\frac{c}{\lambda^2}\Delta\lambda \qquad [1]$$

The wide spacing of signal wavelengths means that temperature stabilization of elements is not needed and the yield of components is high, resulting in a lower-cost system.

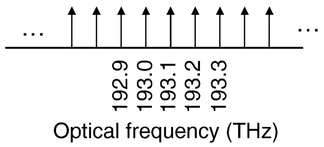

(a) Dense wavelength division multiplexing spectrum

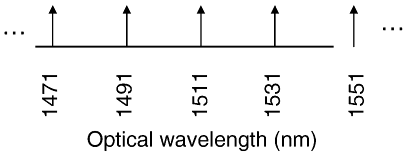

(b) Coarse wavelength division multiplexing spectrum

Figure 2 (a) DWDM spectrum (b) CWDM spectrum.

Point-to-Point DWDM Links

Point-to-point DWDM links were widely deployed during the 1990s. The essential elements are shown in Figure 4, namely single frequency lasers, optical multiplexers and demultiplexers, optical amplifiers, optical add drop multiplexers (OADM), and optical receivers. These will be described in more detail in subsequent sections. Essential parts of the design of these links are managing the signal-to-noise ratio and managing the pulse broadening. With periodic amplification in EDFAs and dispersion compensation, links of 1000 km are common. The widespread use of forward error correction (FEC) circuits allows error-free transmission over a link with limited signal-to-noise ratio. With careful design, ultra-long-haul

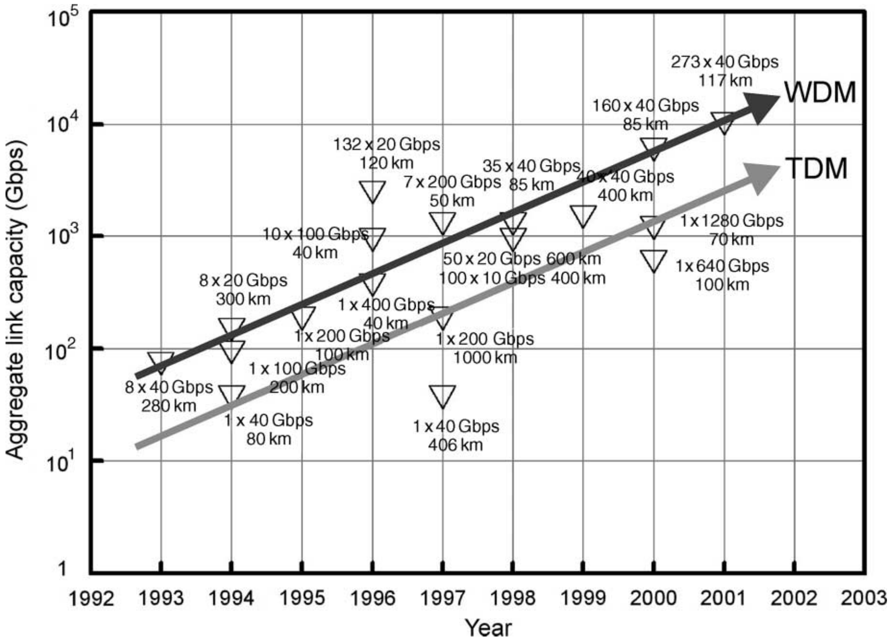

Figure 3 Historical evolution of transmission systems.

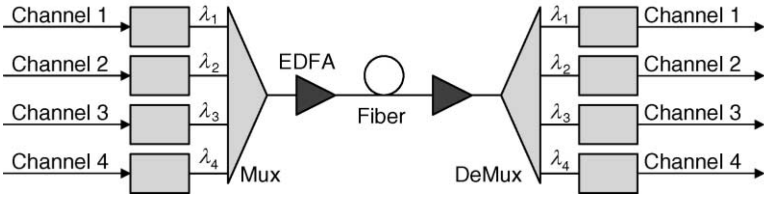
Channel 1
Channel 2
Channel 3
Channel 4
λ_1
λ_2
λ_3
λ_4
Mux
EDFA
Fiber
DeMux
λ_1
λ_2
λ_3
λ_4
Channel 1
Channel 2
Channel 3
Channel 4

Figure 4 Point-to-point WDM link.

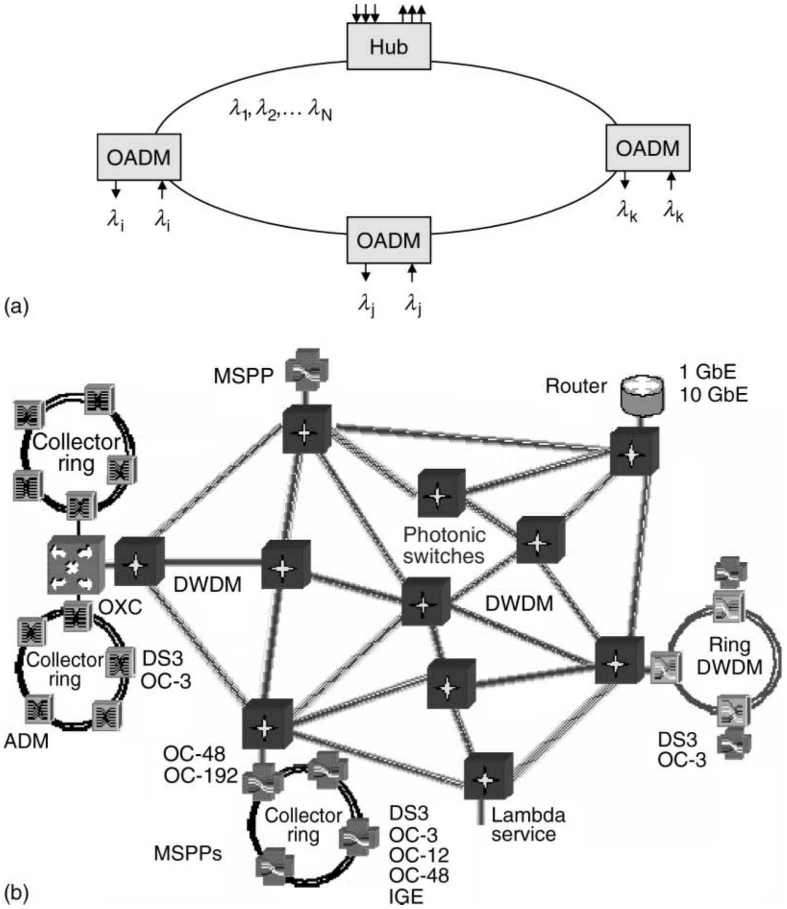
Hub
$\lambda_1, \lambda_2, \dots \lambda_N$
OADM
λ_i
λ_i
OADM
λ_k
λ_k
OADM
λ_j
λ_j
(a)
MSPP
Router
1 GbE
10 GbE
Collector ring
Photonic switches
DWDM
OXC
DWDM
Collector ring
DS3
OC-3
Ring DWDM
ADM
OC-48
OC-192
DS3
OC-3
Collector ring
MSPPs
DS3
OC-3
OC-12
OC-48
IGE
Lambda service
(b)

Figure 5 (a) WDM ring network. (b) DWDM mesh networks (after Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings).

transmission over 5000 km is possible, even with many 10 Gbit/s channels.

DWDM Networks

Point-to-point DWDM links are widely deployed in opaque networks with electronic regeneration and switching at nodes. The next stage in optical networks is the deployment of DWDM transparent ring and mesh networks. An example of a simple ring network is shown in Figure 5a where most of the light signals transit nodes transparently. The OADM elements drop particular wavelengths at each node. As more wavelengths are deployed in these ring networks, it becomes necessary to dynamically reallocate wavelengths and resources. Reconfigurable optical add/drop multiplexers (ROADMs) and optical switches have been developed to allow for dynamic transparent mesh networks (Figure 5b). This allows service providers to provision their networks remotely and quickly adapt to changing demand and restore vital connections when service interruptions occur.

Critical WDM Components

The previous sections described WDM networks. This section describes in more detail the critical components used in these networks.

Multiplexers/Demultiplexers

Multiplexers (MUXs) are used to combine multiple wavelength channels into a single fiber. They are usually wavelength selective and utilize wave phenomena such as diffraction and interference. A simple cascade of 3 dB couplers is not a practical multiplexer since the output power decreases exponentially with channel count. Due to the reciprocity of electromagnetic waves, most MUXs can also serve as a DEMUX. For high channel count systems, a multistage design may be necessary to overcome the limitations of a single MUX/DEMUX. This approach also provides a more flexible upgrading plan. In this case, wavelength channels can be grouped into bands or interleaved. Diffraction gratings and arrayed waveguide gratings (AWGs) are among the available technologies for MUX/DEMUX.

In a diffraction grating, narrow slits or micro structures are spaced evenly on the grating plane and serve as a secondary light source for the input light. The transmitted or reflected light (depending on the design) forms a diffraction pattern on the image plane. The locations of peak intensity on the image plane depend on the wavelength and the grating pitch. Therefore, wavelength channels can be separated in space and fibers can be positioned properly to couple individual channels. An AWG is a planar integrated and digitized form of diffraction grating as shown in Figure 6. It is usually fabricated on a silicon substrate with silica waveguides (silica-on-silicon). The input light is coupled into an array of waveguides with incremental difference in length. These tributaries are then combined in a free-propagating output coupler to several output waveguides. Due to the interference of these tributaries from the arrayed waveguides, different wavelength channels are focused to different output waveguides.

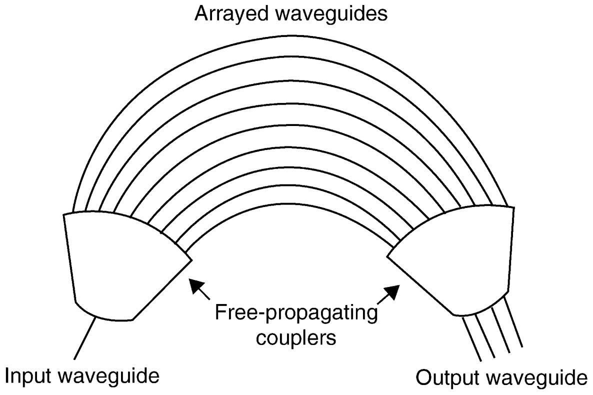

Figure 6 Schematic diagram of an AWG.

Optical Add/Drop Multiplexer

An optical add/drop multiplexer (OADM) provides critical networking capability for WDM systems. In an OADM, some of the wavelength channels are dropped from the input and are either terminated or routed. New channels at these wavelengths are then added with the through channels. This function can be done with the combination of a MUX and a DEMUX of any kind as shown in Figure 7a. However, if only one wavelength channel needs to be processed while others pass through, a simpler configuration can be used where a circulator drops the reflected wavelength from the fiber Bragg grating and a new channel is added with a coupler (Figure 7b).

The two approaches described above are termed static since the wavelength processed is fixed. To enable a more flexible network, reconfigurable OADM (ROADM) are required, where the wavelengths to be added or dropped can be dynamically specified. Figure 7c shows a configuration of ROADM based on a high port-count optical switch and a set of MUX/DEMUX. A tunable laser is highly desirable in this scenario since the wavelength of the added channel can be reconfigured to resolve wavelength contention in the network.

Optical Amplifiers

Erbium-doped fiber amplifiers (EDFA) (*see* **Optical Amplifiers:** Erbrium Doped Fiber Amplifiers for Lightwave Systems) provide simultaneous amplification of wavelength channels in the entire C-band (1530 to 1565 nm) and part of the L-band (1565 to 1605 nm). EDFAs were a key enabling technology for the deployment of WDM systems, especially for long-haul transmission where costly regenerators were eliminated. Both 980 nm and 1480 nm pumps can be utilized to achieve gain in erbium-doped fibers. The 980 nm pumps provide low-noise and high-gain operation while the 1480 nm pumps are more favorable for high-output power. One of the most significant advantages of an EDFA over a semiconductor optical amplifier (SOA) is that crosstalk among wavelength channels is much smaller in EDFA due to a millisecond-scale spontaneous emission lifetime. One critical issue for EDFAs is that the gain is not flat over the operation wavelength range and peaks around 1532 nm. However, this can be equalized by specially designed filters.

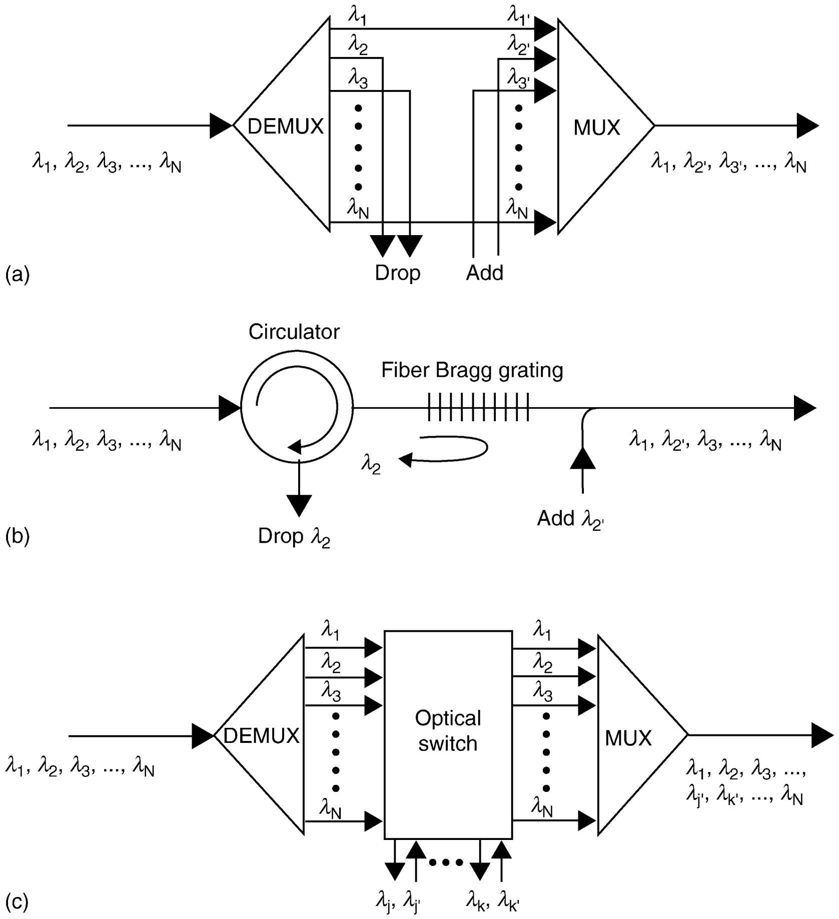
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
DEMUX
λ_1
λ_2
λ_3
λ_N
$\lambda_{1'}$
$\lambda_{2'}$
$\lambda_{3'}$
λ_N
MUX
$\lambda_{1'}, \lambda_{2'}, \lambda_{3'}, ..., \lambda_N$
Drop
Add
(a)
Circulator
Fiber Bragg grating
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
λ_2
$\lambda_1, \lambda_{2'}, \lambda_3, ..., \lambda_N$
Drop λ_2
Add $\lambda_{2'}$
(b)
$\lambda_1, \lambda_2, \lambda_3, ..., \lambda_N$
DEMUX
λ_1
λ_2
λ_3
λ_N
Optical switch
λ_1
λ_2
λ_3
λ_N
MUX
$\lambda_1, \lambda_2, \lambda_3, ...,$
$\lambda_{j'}, \lambda_{k'}, ..., \lambda_N$
$\lambda_j, \lambda_{j'}$
$\lambda_k, \lambda_{k'}$
(c)

Figure 7 (a) A parallel OADM architecture which allows multiple simultaneous add's and drop's; (b) a linear OADM architecture for processing a single channel; (c) an ROADM which enables reconfigurable processing of multiple wavelength channels.

Outside the gain spectrum of EDFAs, Raman amplifiers can be used, where the gain peak is about 13 THz (100 nm) below the frequency of the pump signal. Also, Raman amplification occurs in the same fiber used for transmission so that it can provide a distributed gain over the span of the fiber, which results in a better signal-to-noise ratio. This is particularly important for ultra-long-haul transmission. A counter-propagating pump scheme is preferable to reduce the influence of pump noise and depletion. The biggest challenge for Raman amplifiers is the availability of high-power pump lasers at the required wavelengths.

Optical Filters

Filters can find many applications in WDM systems, such as in multiplexers, demultiplexers, OADM, receivers and optical amplifiers. Fiber Bragg grating, Fabry–Perot filter, and multilayer thin-film filters are among the most widely used filters and utilize the interference property of optical waves. There are some general requirements for filters in WDM systems. First, the top of the pass-band must be flat enough so that cascading of filters in the network does not lead to a significant decrease of bandwidth. In addition, the roll-off of the pass-band must be sharp so that cross-talk from other channels is minimized.

A fiber Bragg grating has a periodic perturbation of refractive index made by exposing a photosensitive fiber to interfering ultraviolet lights. These perturbations reflect the input light as it propagates. At a specific wavelength, these reflected beams add in phase, which results in a strong overall reflection. The reflections at other wavelengths do not add up and these wavelengths will eventually pass through the grating. In short, the fiber Bragg grating acts as a wavelength-selective reflector and is usually used with a circulator or isolator to manage the reflected light. A Fabry–Perot filter is composed of two partially transmitting interfaces and a cavity in between. The underlying mechanism is similar to the fiber Bragg grating but the transmission is periodic in optical frequency. The spacing is called the free spectral range, which depends on the optical length of the cavity. The bandwidth of the passband is a strong function of the reflectivity of the partially transmitting interfaces. A multilayer thin-film filter is an extended form of Fabry–Perot filter where several layers of dielectric reflectors are formed with dielectric cavities in between. This allows for more tuning dimensions in the design of the passband shape in order to meet the requirements of WDM systems and the temperature tolerance is better than other kinds of filter.

Laser Sources

Semiconductor lasers (*see* **Lasers:** Semiconductor Lasers) are widely used in optical communication because they are compact, low cost and efficient. As a light source for transmission, single frequency operation is necessary for high-capacity WDM systems. Multilongitudinal mode lasers such as Fabry–Perot lasers have excess bandwidth which causes excessive penalty due to dispersion unless the transmission distance is short. They also result in crosstalk in DWDM systems. The most widely used single-frequency laser in WDM systems is the distributed-feedback laser (DFB), where a corrugated grating structure is fabricated on top of the gain region and only the wavelength which satisfies the Bragg condition can oscillate in the laser cavity and all other modes are suppressed.

As WDM networks advance, tunable lasers are strongly desired to reduce the excessive inventory of fixed wavelength lasers for DWDM and also to provide reconfigurability in the network. One possible structure is the distributed Bragg reflector (DBR) laser, in which the Bragg grating is decoupled from the gain region and the Bragg wavelength can be tuned independently by isolated current injection. More complicated variations of this concept such as the three-section DBR laser and the sampled grating DBR laser can provide more continuous and wider tuning in wavelength.

Photonic Switches

Photonic switches are necessary in transparent networks to allow reconfigurability in networks, and allow rapid provisioning and fault recovery. They provide connectivity between rings and between point-to-point links without requiring the DWDM signal to be broken into individual wavelengths and to be electronically regenerated. Photonic switches can be used to switch fibers, wavebands, or individual wavelengths. They are inherently transparent to bit-rate and data format. A number of different technologies have been developed.

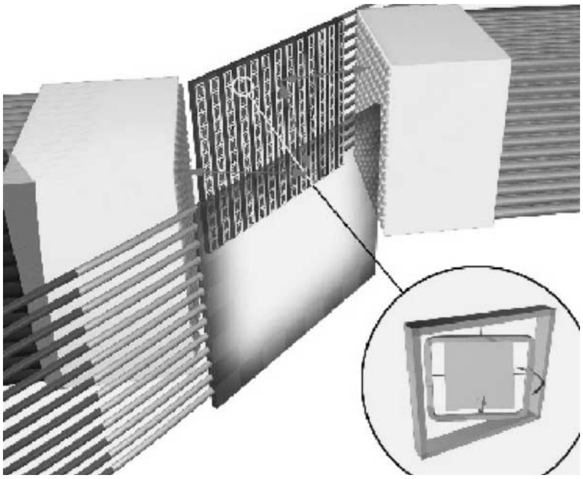

Figure 8 3D MEMS optical switch.

Microelectromechanical system (MEMS) switches are widely used for 2 × 2 protection switches, wavelength selective switches and for nonblocking crossbar switches with sizes from 8 × 8 to 1024 × 1024. The technology usually used for large switches is called 3D MEMS switches (Figure 8), in contrast to 2D technologies where the optical beams are confined to a two-dimensional plane.

Transmission Issues

Chromatic Dispersion

Chromatic dispersion (*see* **Fiber and Guided Wave Optics:** Dispersion) means that the group velocity, which is the propagation speed of an optical signal, changes with wavelength. Since every real pulse has a finite bandwidth, each slice of wavelength within this bandwidth propagates with a different speed and results in the distortion of the pulse as it propagates along the fiber. Depending on the sign of the dispersion and the initial phase distribution, the pulse can either be broadened or compressed. However, if the fiber is long and nonlinear effects are small, the pulses will eventually be broadened and result in intersymbol interference. Dispersion compensating fibers (DCFs), which have an opposite sign of dispersion, can be used to reverse this process and restore the original pulse shape. In general, it is advantageous to minimize the length of DCF by keeping the dispersion low. Dispersion-shifted fibers (DSF) have a zero-crossing of dispersion around 1.5 μm and hence a reduced dispersion for the C-band.

For a WDM system, a very low dispersion will result in more severe nonlinear impairments since the distance pulses can overlap with each other along the fiber is long. Nonzero-dispersion-shifted fibers

(NZ-DSF), which introduce a suitable amount of dispersion, can be used instead to solve this issue. Chirped fiber Bragg grating can also be used, where different wavelengths are reflected at different positions in the grating so as to compensate the dispersion caused by transmission. It is capable of compensating a given amount of dispersion with a much shorter length when compared to DCF. For high channel count WDM systems, the dispersion experienced by each channel may be different and dispersion compensation needs to be implemented on a per-channel basis after a demultiplexer.

Polarization Mode Dispersion

Polarization mode dispersion (PMD) means that the group velocity dispersion of the two orthogonal polarizations in the fiber are different and cannot be simultaneously compensated by a single DCF. It is usually manifested as the splitting of a pulse at the receiver even after chromatic dispersion compensation and causes intersymbol interference. PMD can originate from the imperfections in the fiber or environmental factors like temperature and mechanical pressure. Therefore, the amount of PMD can vary with time for a given link, which means that dynamic compensation is necessary. Traditional electrical equalization after the optical signal is detected can be used for compensation but the highest bit-rate achievable is limited by the speed of electronics. Optical compensation is another choice. The optical signal is split into fast and slow components and the fast component is delayed properly before being recombined with the slow component.

Nonlinear Impairments

The most important nonlinear impairments for WDM systems are the cross-phase modulation (CPM) and the four-wave mixing (FWM). They are both third-order nonlinear effects, where the nonlinear dielectric polarization is proportional to the third power of the electric field of the signal. In CPM, the phase of a wavelength channel can be modified by other channels, which increases the chirp and results in more dispersion penalty. On the other hand, FWM can generate new waves at other wavelengths, causing unwanted crosstalk if the generated wave happens to coincide with another channel. The existence of dispersion in fiber can help to reduce the effects of both impairments since the pulses at different wavelengths propagate at different speeds and walk away from each other so that nonlinear interactions are reduced. This is why NZ-DSF is preferred over DSF. Self-phase modulation (SPM) is another third-order nonlinearity, where the phase distribution of a pulse is altered by its intensity profile. It will lead to the distortion of the pulse. Other nonlinear effects include stimulated Brillouin scattering and stimulated Raman scattering. Both originate from the interactions of phonons and photons in the fiber and cause power fluctuations at different wavelengths.

Forward Error Correction

To reduce the errors of a transmission link, extra information relating to the data can be sent by the transmitter and used by the receiver to detect and correct errors in the received data. This technique is called forward error correction (FEC) and imposes redundancy to the transmission which means that the bit-rate transmitted will be higher than the actual data rate. A measure of performance for FEC is the coding gain, which is defined as the decrease in receiver sensitivity at a given bit-error-rate. Reed–Solomon codes are one of these codes. With a 7% redundancy, up to 6 dB of coding gain can be achieved. This could correspond to more than 9 orders of magnitude improvement in bit-error-rate. Therefore, the use of FEC can increase the spacing between regenerators in ultra-long-haul transmission and reduces the cost. It is also useful to overcome error floor problems due to crosstalk in a WDM system, where a further increase in received power cannot reduce the bit-error-rate. Crosstalk among wavelength channels can happen due to nonlinear impairments during fiber transmission or due to imperfect channel isolation in components like multiplexers, demultiplexers, filters, and switches. The use of FEC can greatly increase the system margin.

Conclusions

WDM technology has enabled a huge increase in the transmission capacity of optical fibers and a huge reduction in the cost of fiber optic transmission since the cost of the installed fiber is shared between the signals traveling on different wavelengths. As new components have been developed, WDM technology has expanded simple point-to-point links to transparent networks. Great progress has been made over the past 20 years to solve transmission impairments and extend the reach and bandwidth of these systems. This is likely to continue to meet the increasing communication needs of our society.

List of Units and Nomenclature

Distance (km, m, μm)
Bit rate (bit/s)
Wavelength (nm)

See also

Color and the World. **Information Processing:** All-optical Multiplexing/Demultiplexing. **Fiber and Guided Wave Optics:** Optical Fiber Cables; Passive Optical Components. **Lasers:** Edge Emitters; Semiconductor Lasers. **Modulators:** Modulation and Demodulation of Optical Signals. **Optical Amplifiers:** Basic Concepts; Optical Amplifiers in Long-Haul Transmission Systems; Semiconductor Optical Amplifiers; **Optical Communication Systems:** Architectures of Optical Fiber Communication Systems; Basic Concepts; Historical Development; Lightwave Transmitters; Optical Time Division Multiplexing. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2002) *Fiber Optic Communication Systems*, 3rd edn. New York: John Wiley & Sons Inc.

Basch B, Gringeri S, Goudreault R and Ravinkumar S (2002) Evolution of the photonic core within metropolitan and regional networks. *NFOEC*. Technical Proceedings.

Green PE (1993) *Fiber Optic Networks*. Upper Saddle River, NJ: Prentice Hall.

Ramaswami R and Sivarajan KN (2002) *Optical Networks: A Practical Perspective*, 2nd edn. San Francisco, CA: Morgan Kaufmann Publishers.

OPTICAL MATERIALS

Contents

Color Filter and Absorption Glasses
Heterogeneous Materials
Lightweight Mirrors
Measurement of Optical Properties of Solids
Optical Glasses
Plastics
Sculptured Thin Films
Smart Optical Materials
Sol-Gel Materials

Color Filter and Absorption Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical filter and absorbing glasses can be categorized by either the optical behavior of the filter, for example, bandpass, shortpass or longpass, neutral density, or interference filters – or by the method used to filter the light, i.e., glasses which absorb light themselves and those which serve as high quality, transparent substrates for coatings which either absorb or reflect light. The term ‘filter’ usually refers to glasses used in optical systems, to either transmit or prevent transmission of specific wavelengths. The term ‘absorber’ is usually reserved for glasses which are used for controlling light to provide protection from damage to humans, for example, for eye protection, as in sunglasses, or welder and laser goggles, or to materials exposed to light, for example, plastics and other components in the passenger compartment of automobiles. Manufacturing technology is normally based on the method used to produce the device, such as control of absorbing species in the glass or method of application of optical coatings. Terminology for optical applications is almost always based on optical behavior and not on the method used to produce that behavior.

Bandpass Filters

The first glasses used to filter light were colored by inclusion of ions from the first row of transition metals. These colored glasses still provide most of

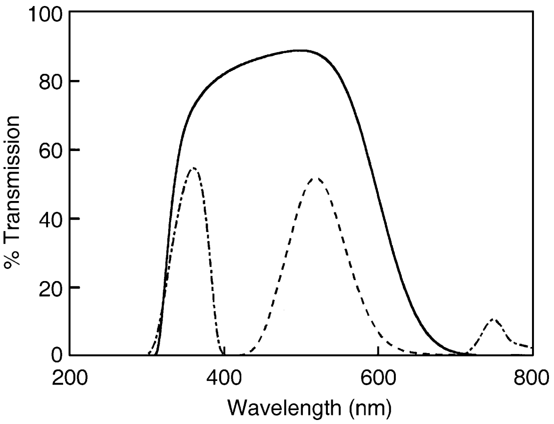

Figure 1 Transmission spectra for bandpass filters with a large bandwidth (solid line), medium bandwidth (dashed line), and narrow bandwidth (dot dash line).

the glasses used to control the wavelength of light. They are widely used as absorbers in windows in buildings and automobiles, to control the entrance of ultraviolet and infrared light into these environments. These filters allow the passage of light over a significant range of wavelengths, while absorbing wavelengths outside that range. They are designated by the peak transmission wavelength, or central wavelength, and the bandwidth, or wavelength range, over which light is transmitted with at least some designated transmission value. Filters based on absorption by transition metal or rare earth ions usually have larger bandwidths and somewhat diffuse cutoffs at the edges of the transmitted spectral region (Figure 1).

The most common bandpass filters absorb radiation in a desired wavelength region. Historically, the first bandpass filters, which are still the most common, relied on absorption of light as a result of ligand field (also known as crystal field) effects and/or charge transfer processes within a glass, to create absorption bands in the spectrum.

Ligand field effects will produce optical absorption bands in glasses which contain transition metal (usually row 3d) or rare earth (lanthanide, or 4f) ions. The differences in the electronic structures of the transition metal and rare earth ions result in very different absorption spectra. Absorption spectra, due to transition metal ions, are characterized by broad bands and are frequently accompanied by charge transfer bands, due to the same ions, while the bands due to absorption of light by rare earth ions are very sharp and are not accompanied by charge transfer absorption bands.

Iron oxide was probably the first, and is still the most widely used additive for production of glasses which primarily transmit light in the visible portion of the spectrum. A large, broad absorption band due to ferrous ions is centered at ≈1050 nm in common soda-lime-silica glasses. The tails of this band extend into the visible portion of the spectrum, allowing passage of the blue-green portion of the spectrum, but absorbing much of the red light, and to wavelengths in excess of 2000 nm, absorbing much of the near infrared spectrum.

Ferric iron produces relatively small ligand-field absorption bands in the region from 380 to 440 nm. These bands remove most of the blue part of the spectrum, allowing passage of light with longer wavelengths. Both ferrous and ferric ions also display charge transfer bands in the ultraviolet, effectively eliminating transmission of light with a wavelength <300 nm. Glasses containing enough iron oxide, to effectively eliminate both ultraviolet and near infrared transmission, exhibit various shades of green, depending on the ratio of ferrous to ferric ions present. Melting under strongly reducing conditions yields blue-green glasses, whereas melting under strongly oxidizing conditions yields yellow-green glasses. Glasses containing ≧0.5 wt% iron oxide are used in windows of automobiles to minimize the amount of ultraviolet and infrared radiation entering the passenger compartment.

Copper oxide is also used to remove a portion of the near infrared spectrum from sunlight and to produce glasses which transmit in the green and blue region. The primary ligand-field absorption band of copper is due to cupric ions and is located near 800 nm. This band in similar in shape to that due to ferrous ions, with tails extending into the visible and near infrared regions of the spectrum. The shift in band position toward the visible, however, results in less absorption in the infrared and more in the visible than for ferrous iron. As a result, glasses containing copper are more strongly colored than those containing ferrous iron and are less effective filters for infrared light. Glasses colored with cupric ions can be used to produce a transmission window from about 350 to 600 nm in soda-lime-silica glasses. Cuprous ions do not create any significant absorption bands.

Glasses containing hexavalent chromium, which must be produced in a strongly oxidizing environment, effectively remove ultraviolet light between 300 and 400 nm, but allow transmission of ultraviolet light in a narrow range centered around 300 nm. A small tail of the charge transfer band due to Cr^{6+}extends into the visible, producing a yellow-green glass. The presence of a set of much weaker ligand-field bands, centered around 650 nm due to trivalent chromium, also contributes to the green color of these glasses. Combining chromium with copper ions in a glass yields a filter with a small transmission window centered at about 520 nm.

Combination of chromium with iron results in suppression of both hexavalent chromium and ferrous iron, yielding a green glass without either the strong ultraviolet absorption of hexavalent chromium or the strong near infrared absorption of ferrous iron.

Manganese oxide can be used to produce bandpass filters which block most of the visible, but allow transmission of ultraviolet and infrared light. Addition of manganese ions to a glass results in a set of absorption bands with a maximum intensity at about 500 nm and a tail extending to around 900 nm. Since these ions do not absorb in the ultraviolet, the transmission spectra of these glasses has a window between about 300 and 400 nm, which overlaps the transmission window due hexavalent chromium. The absorption bands due to manganese ions are relatively weak, so that concentrations of manganese oxide in excess of 1 wt% are needed to effectively block visible transmission.

Cobalt oxide is the strongest colorant of the oxides of 3d transition metals. A small addition of cobalt oxide to silicate glasses creates strong absorption bands between about 450 and 700 nm. The optical window of these glasses in the ultraviolet is centered near 380 nm and is somewhat broader than that of glasses containing manganese oxide, which have similar spectra. Since cobalt oxide is such a strong absorber, black glasses are easily produced by use of cobalt oxide in combination with other oxides. If cobalt oxide is combined with nickel oxide, the resulting glass retains a bandpass in the ultraviolet, while combining cobalt oxide with iron oxide, which has a complementary absorption spectrum, yields a black glass which also effectively eliminates transmission in the ultraviolet and near infrared.

The optical spectra of glasses containing most rare earth ions are characterized by sharp absorption bands. Some rare earth ions create spectra with many absorption bands, while others display only a few bands. The band structures which result from rare earth ions are not particularly useful for creation of bandpass filters, although Yb^{3+} ions do effectively remove light around 1100 nm, without absorbing light in the rest of the spectrum.

While not commonly used for filters, glasses containing rare earth ions have been used for wavelength calibration standards for uv-vis spectrophotometers. A common wavelength standard called a ‘didymium’ glass contains a combination of neodymium and prasedymium oxides. Since the optical spectrum from each of these ions contains many sharp bands, these glasses provide known wavelengths for absorption bands over a wide region of the spectrum. Neodymium ions, which have absorption bands near the D line of sodium, are used in goggles for eye protection for glassblowers, to remove the intense yellow light resulting from the sodium in the glass.

Interference Filters

Interference filters are a special case of bandpass filters which are created by application of dielectric films to glass substrates. The dielectric films are deposited on a glass substrate in alternating layers of high and low refractive index materials. The combination of refractive indices and thickness of the dielectric films determines the wavelength regions where constructive and destructive interference occur, i.e., those wavelength regions which will transmit or reflect the incident light. Since the choice of materials will control the indices, production of filters for different wavelengths usually relies on deposition of layers of differing thicknesses.

These filters can be produced with very narrow bandwidths (±10 nm) or with broader bandwidths (commonly ±50 or ±80 nm). The peak transmission frequently does not occur at the midpoint of the bandwidth. Since the wavelengths for constructive and destructive interference depend upon the path length through the dielectric layers, these filters are extremely dependent upon the incident light angle, with a shift toward lower wavelengths for incident angles other than normal to the filter surface. Interference filters typically reduce throughput of light as compared to absorption-based bandpass filters, but provide for much narrower transmission windows. If the dielectric layers are deposited in such a manner that their thickness varies along the length of the substrate, interference filters can be produced with a continually varying central transmission wavelength along the length of the substrate. While convenient for applications, which might otherwise require several filters of different central wavelengths, these linearly variable interference filters are much more expensive than those designed to block specific wavelengths.

Dispersion Filters

Dispersion filters are based on the passage of light through a medium consisting of a matrix material and dispersed, very small particles of a material which has a very different dispersion curve. The original version of these filters, called a Christiansen filter, consists of small glass particles with a controlled particle size suspended in a liquid. The dispersion curve for the glass is much less wavelength-dependent than that of the liquid. Transmission of light occurs at the wavelength where the refractive index of the glass and the liquid are equal, while light of other wavelengths is scattered. Since the refractive index of a typical liquid is highly temperature sensitive, the transmitted, or peak, wavelength can be shifted over a wide region by controlling the temperature of the mixture. The bandwidth of dispersion filters can be very narrow (<± 10 nm).

The large temperature dependence of the traditional dispersion filter can be useful in that it provides an optical filter with an adjustable peak wavelength. On the other hand, this sensitivity to temperature also results in a serious control problem, since a temperature change of only a few degrees can shift the transmission window by up to 100 nm. This problem can be decreased or even eliminated by making a solid dispersion filter using two glasses of different dispersions, with one glass distributed in the other, or a glass dispersed in some other transparent solid such as a polymer with a similar thermal expansion coefficient. Since the refractive index of solids tends to vary with thermal expansion coefficient, these solid state dispersion filters can be almost temperature independent.

Neutral Density Filters

Neutral density filters are used to reduce the throughput of light to some desired value over a wide spectral region. Gray glasses are used to reduce the intensity of light from a bright source such as the sun. Black glasses are used to completely block light in the visible region. Neutral density filters can be produced using a combination of transition metal ions or by use of optical coatings. They are characterized by their percent transmission, range of wavelength where the transmission is constant, and the 'ripple', or spectral flatness, within the designated spectral region (Figure 2).

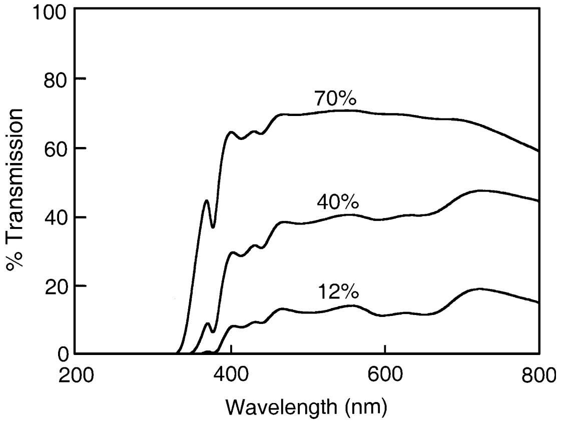

Figure 2 Transmission spectra for neutral density filters with different % transmission.

An approximation to a neutral density filter for visible light is often created by combining the absorptions in the ultraviolet and near infrared from ferrous and ferric ions, with the multiple absorption bands centered in the region of 500 to 600 nm due to divalent cobalt ions. While the resulting absorption spectrum displays small ripples throughout the region from 350 nm to beyond 1200 nm, the transmission of the glass can be controlled to within a few percent by the proper balance of ferric and ferrous ions combined with the desired concentration of CoO. These glasses are generally gray, or, if enough of iron and cobalt oxides are present, black. Combination of CoO with NiO can produce similar filters covering a smaller wavelength range which maintain transmission in the ultraviolet. Filters with different transmission values can be produced by either varying the composition of the glass or by use of different glass thicknesses. A continuously graded neutral filter can be produced from a single piece of glass by etching or polishing to form a wedge. With more effort, a step-graded filter can be produced by etching or polishing to form sharply delineated thickness steps along the length of a glass plate.

Neutral density filters can also be produced by coating a glass substrate with a metal film. Coated neutral density filters have less ripple than absorptive filters based on iron and cobalt oxides dissolved in the glass. In order to produce a constant transmission throughout the desired wavelength region, the metal used (commonly inconel) should have a constant reflectance across the desired wavelength region. Use of a metal with a varying wavelength dependence in its reflectivity, such as silver, will result in a wavelength dependent transmission. Continuously graded neutral filters can be produced by varying the thickness of the film along the length of a strip of glass.

Longpass/Shortpass Filters

Longpass filters are designed to pass light with wavelengths longer than the specified cutoff value. They are frequently used to block ultraviolet light while transmitting visible light or to block visible light while transmitting infrared light. Shortpass filters provide the opposite effect, i.e. they pass light with wavelengths shorter than the specified cutoff value, while blocking light with longer wavelengths. Shortpass filters are often used to block passage of heat (infrared light), while allowing transmission of visible light (Figure 3).

Ultraviolet blocking longpass filters can be produced by varying the basic composition of a glass or

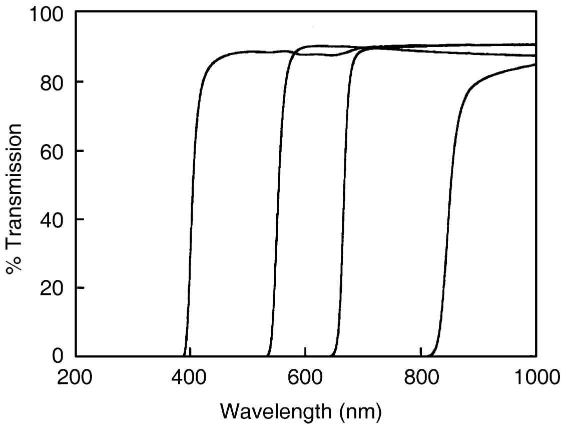

Figure 3 Transmission spectra for longpass filters with different cutoff wavelengths.

by addition of specific ions which absorb strongly in the ultraviolet. Varying the primary glassforming oxide can alter the ultraviolet cutoff frequency of a glass. Vitreous silica, for example, has an ultraviolet cutoff beyond 200 nm, so that it is a very effective transmitter of ultraviolet light. Vitreous germania, on the other hand, has an ultraviolet cutoff near 400 nm, so that it acts as an effective longpass filter for most ultraviolet light. While vitreous silica is widely used as an optical material, the poor chemical durability of vitreous germania prevents its use in most optical applications.

The ultraviolet cutoff wavelength of silicate glasses is shifted toward the visible by addition of alkali and alkaline earth oxides to silica. The addition of these oxides creates nonbridging oxygens in the vitreous network, which shift the ultraviolet cutoff toward the visible. These glasses are effectively opaque beyond their cutoff wavelength, creating longpass filter glasses with cutoff wavelengths which vary with the nonbridging oxygen concentration. Impurities such as iron or copper can also reduce the ultraviolet transmission of silicate glasses due to the creation of charge transfer bands, producing longpass filters with cutoffs in the range of 250 to 300 nm.

Longpass filters with tightly controlled ultraviolet cutoff wavelengths are often made by adding either cerium or a combination of cerium and titanium oxides to silicate glasses. Additions of cerium oxide result in a strong absorption band near 320 nm, which provides protection from ultraviolet light. Cerium also is routinely used to reduce radiation-induced coloration of glasses exposed to high energy radiation. An absorption band due to trivalent titanium also serves to cut off ultraviolet light, while producing a yellow glass. The use of a combination of cerium and titanium produces a very effective ultraviolet filter which eliminates transmission of almost all ultraviolet light.

Long bandpass absorption filters can also be based on precipitation of colloids of metals or semi-conductors in an otherwise colorless glass. The glasses are initially prepared in a colorless state, with the materials which will form the colloids dissolved as ions. The glasses are then reheated into the glass transformation region, where the desired ions are reduced to the atomic state and precipitated as colloids. Since this process is known as ‘striking’, these filters are known as striking filters. Colloids can be produced by striking from several elements, including silver, gold, copper, arsenic, antimony, lead, and bismuth, and from a number of compounds, including the cadmium chalcogenides and ZnS.

Formation of silver colloids in common soda-lime-silicate glasses results in a very strong optical absorption band centered near 400 nm. Glasses containing silver colloids are yellow, with a bandpass window in the ultraviolet. Since these glasses absorb so strongly at a wavelength commonly used to irradiate photosensitive polymers in the production of electrical circuit boards, they are frequently used for the production of beam masks. A pattern of silver colloids is easily produced in the surface of common float glasses by depositing metallic silver through a mask onto the surface of the glass and then heating the glass to 500 °C for a short time.

Copper can also be used to form longpass filters with a cutoff in the red portion of the spectrum by a similar process to that used to form the silver colloid filters. In this case, it is necessary to use a copper salt in place of the metallic silver, but the remainder of the process is essentially the same (Figure 4).

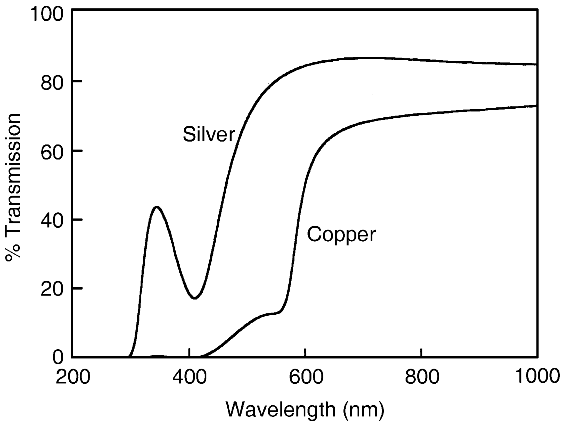

Figure 4 Transmission spectra for longpass filters containing silver or copper colloids.

Precipitation of semi-conducting colloids containing CdS, CdSe, and CdTe in silicate glasses results in sharp cutoffs in the range from the near ultraviolet to the near infrared. The cutoff is the direct result of the bandgap of these crystallites. Since these materials form a broad series of solid solutions, the desired cutoff wavelength can be obtained by use of an appropriate mixture of two of these compounds in the glass batch. Glasses containing only CdS, for example, have cutoffs in the range of 400 to 500 nm. Combining CdS with CdSe will shift the cutoff toward the infrared, with cutoffs in the range of 500 to 700 nm. Combining CdS with CdTe produces glasses with cutoffs near the edge of the visible region. Use of CdTe alone results in filters with cutoffs around 800 nm.

While the most common explanation for the shift in the cutoff in glasses containing the cadmium chalcogenides is based on solid solution arguments, it has been shown that the size of the particles also strongly affects the position of the cutoff for glasses containing CdSe. In this case, the energy gap, which determines the cutoff wavelength, varies inversely with the square of the particle size.

Glasses containing CdS and ZnS can be heat-treated to form filters with cutoffs in the range from 340 to 518 nm. Heat treatment at lower temperatures yields CdS colloids, with a cutoff around 500 nm. At higher heat treatment temperatures, the absorption edge shifts toward the ultraviolet, approaching the value of 340 nm representative of the lattice absorption of ZnS. This behavior has been interpreted as indicating that CdS crystals precipitate at lower temperatures, while at higher temperatures, solid solution crystals of ZnS and CdS are formed.

Colloids of arsenic, antimony, bismuth, and lead form less interesting filter glasses. The optical properties of these metals are such that there is no sharp cutoff wavelength. Growth of colloids produces wavelength-dependent scattering, with a long tail extending into the visible or infrared. These glasses can be produced in colors ranging from light tan to black by progressive heat treatments of the originally colorless glasses in hydrogen. The hydrogen serves to reduce the ions to the atomic state. The atoms then diffuse to form colloids. The growth process creates a sharp interface between the colored and colorless region of the glass, which can be used to produce windows for visible light of highly transmitting glass surrounded by essentially opaque glass.

Absorber Glasses

Glass filters are used to remove unwanted wavelengths from light. When the application is less demanding than those of complex optical devices or systems, the glasses are commonly called ‘absorbers’ rather than filters. These applications usually involve protection of humans or materials from damage by light, for example, sunglasses and goggles for welders and glassworkers, or to enhance the environment by controlling the entrance or exit of ultraviolet and infrared light. Most of these applications are variations on the filters discussed above.

The most common use of glass as an absorber for both protection from damage and enhancement of the environment is found in the soda-lime-silicate glasses used as windows in buildings and automobiles. The presence of a relatively large concentration of iron oxide provides absorption in both the ultraviolet and infrared regions of the spectrum. Since glasses containing iron do absorb some light in the visible, a compromise between total absorption of unwanted, nonvisible light and the desired visible light must be made. As a result, these glasses usually contain between 0.5 and 2 wt% iron oxide. Changes in melting conditions which alter the ratio of ferrous to ferric ions in the glass are used to tune the glass for specific applications. While these glasses are technically broad bandpass filters, their common applications are based on the favorable location and bandwidth of their transmission window, which roughly corresponds to the visible region.

Iron is also used for protection from ultraviolet and infrared radiation in welder goggles. These glasses contain much more iron (up to 8 wt% or more) than those used for windows, since the demand for protection from the intense light created in welding is much greater. The high absorption coefficient for infrared light results in considerable heating of the glass during lengthy use. A thin film of gold is often applied to the glass to reflect infrared radiation and reduce heating of the glass.

Other absorber applications are more demanding in terms of absorption wavelength. As mentioned earlier, glasses containing neodymium have a strong absorption band at the wavelength of the yellow sodium emission line. These glasses transmit visible light at other wavelengths, allowing their use by glassblowers, who need to be able to see the unheated part of the glass and their immediate surrounding area, while simultaneously reducing the intense yellow light given off by the sodium in the glasses normally used for glassworking.

Protection of the eye from laser radiation is even more demanding due to the very high intensity over very narrow wavelength ranges. Common ultraviolet longpass filters are usually satisfactory for absorption of light from ultraviolet lasers. Specific ions must be added to glasses to provide protection from many visible and near infrared lasers, since the glass must simultaneously permit transmission of visible light of other wavelengths and total absorption at the laser wavelengths. Light from lasers which operate in the far infrared is readily absorbed by most glasses, since the infrared cutoff of oxide glasses lies around 5 micrometers. Great caution should always be used in selecting glasses for eye protection from lasers. No assumptions regarding the degree of protection should ever be made without insuring that the glass is designed for the conditions under which it will be used.

See also

Optical Coatings: Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Bamford CR (1977) *Colour Generation and Control in Glass*. Amsterdam: Elsevier Scientific Publishing Company.

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds). *Commercial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Edmund Industrial Optics Catalog N017 (2001) pp. 68–87. Barrington, NJ: Edmund Industrial Optics.

Hillmann E and Neuroth N (1998) Glasses for eye protection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 370–379. Heidelberg, Germany: Springer-Verlag.

Kolberg U (1998) Coloured glasses. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 350–370. Heidelberg, Germany: Springer-Verlag.

Neuroth N (1998) Transmission and reflection. In: Bach H and Neuroth N (eds). *The Properties of Optical Glass*, pp. 82–95. Heidelberg, Germany: Springer-Verlag.

Paul A (1982) *Chemistry of Glasses*, pp. 204–276. London: Chapman and Hall.

Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds). *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Weyl WA (1951) *Coloured Glasses*. Society of Glass Technology: Sheffield, England.

Heterogeneous Materials

U Kreibig and M Quinten, Aachen University of Technology, Aachen, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The term heterogeneous materials – being a quite diffuse classification of matter with discrete building units of different kinds, irrespective of their sizes – is defined here by scaling these inhomogeneities in relation to optical light wavelengths. Matter consisting of atoms in a regular structural order (or, alternatively, in fully statistical disorder) is classified as homogeneous since the inhomogeneities are too small to effect optical fields. This applies to crystalline, liquid, and glassy or amorphous matter likewise.

All geometrical/topological and chemical heterogeneities on a larger length-scale than the atomic scale, i.e., single crystallites, organic macromolecules, domains, extended lattice defects, clusters, and nanoparticles of chemical compositions differing from their host material, etc., result in a macroscopic material that is heterogeneous, in optical properties and optical experiments. Examples are multilayer-structures, heterogeneous in one dimension, three-dimensional heterogeneous composites, i.e., isometric building units distributed in some homogeneous host (photographic systems, drugs, cosmetics, color pigments in binders, etc.), and three-dimensional heterogeneous densely packed systems (poly- or nanocrystalline materials, ceramics, or nanoceramics, etc.).

The materials most intensively investigated, in the field of nanoscience, represent a subclass of heterogeneous materials.

Most of the common materials in our everyday environment have heterogeneous structures, often on larger characteristic size-scales. More generally, heterogeneous materials are so widespread that they are much more frequent than homogeneous materials. Some more examples, relevant for their optical properties, may illustrate this: pigment colors, photovoltaic composites, aerosol-systems, hydrosol-systems, many minerals, photonic crystals, island film devices, nanosensors, special biosystems, and polymer-systems. Most interesting optical features arise when these inhomogeneities are below or of the order of light wavelengths. Recent research has concentrated on a narrow size region on the lower nanometer scale where many-atoms-systems change from molecular to solid-state behavior with increasing sizes. This ‘transition’ is accompanied by drastic changes of almost all physical and chemical properties, hence these properties develop strong dependences on particle size in that size region, and materials with uncommon properties can be developed.

In contrast to the widespread applications and research perspectives, the basic theoretical understanding of heterogeneous material is generally still poor. One reason is that, since the beginning of the development of optics as a science, the homogeneous materials like glass, water, etc. were preferentially treated since they were easier to describe theoretically. Another reason is the extremely wide variability of existing heterogeneous structures usually based upon low symmetries. This variability is too large, and the resulting optical properties too complex and multifarious to enable a general and uniform description of heterogeneous matter on a quantitative level, like the one successfully developed for the much simpler topology of homogeneous crystals.

Qualitative general features, however, can be obtained from investigations of model systems, with especially simplified topological structures. In the following we describe the model system of nanoparticle/host-matter (or, synonymously, cluster-matter), which is based upon nanoparticles/clusters packed closely in a host material, thus forming macroscopic or microscopic heterogeneous many-particle systems. Even this model system exhibits broad variability of topologies and of chemical compositions, as shown in Figure 1.

The restriction in the present contribution to this topologically simple model system appears to be justified by the fact that the optical properties discussed below for this system are also qualitatively characteristic of other arbitrary heterogeneous materials. We also admit that the nanos may consist of a nucleus and a surrounding shell of different structure and/or different material.

Heterogeneous matter can be defined by its heterogeneous building units and by the borders of these units and by the host material and the borders between building units and host, i.e., the internal interfaces.

It is a general feature that structures are visible because of property changes at their terminating borders: borders create structures, and structures require borders. The more refined the structures are, the more increases the topological extent of the borders. This means that when sizes of building units, for example, in cluster matter decrease, the relative amount and the effects of surface/interface regions increase. Such interfaces in heterogeneous matter are not two-dimensional (as suggested by the term 'interface') but consist of three-dimensional

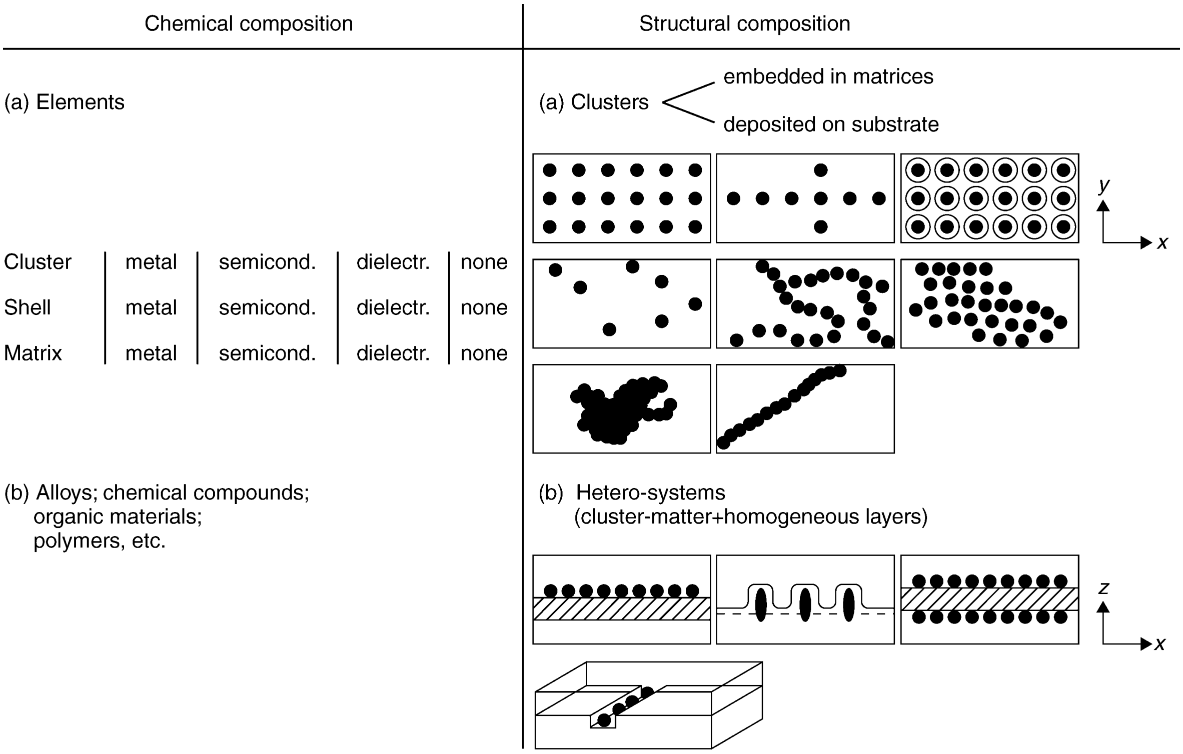
Chemical composition
Structural composition
(a) Elements
(a) Clusters
embedded in matrices
deposited on substrate
y
x
Cluster | metal | semicond. | dielectr. | none
Shell | metal | semicond. | dielectr. | none
Matrix | metal | semicond. | dielectr. | none
(b) Alloys; chemical compounds; organic materials; polymers, etc.
(b) Hetero-systems (cluster-matter+homogeneous layers)
z
x

Figure 1 Examples of heterogeneous nanoparticle/host materials. Left: possible material combinations of the three building units: cluster, shell, matrix. Right: some examples of sample topologies.

interlayers, the atomic and electronic structure and the chemical composition of which are usually in complex disorder. In the case of very small particles the interface volume can easily exceed the volume of the building units themselves. There are, however, also examples where the atomic arrangement forms regular interfaces, such as at coherent grain boundaries.

Hence, heterogeneous materials are characterized by various building units, homogeneous matrices and the interface regions. To describe them, we will focus on our introduced model system of nanoparticle/host matter, where we further specialize to spherical shapes of the particles.

The investigation and description of optical properties of heterogeneous materials have to deal with the interior of the building units, e.g., characteristic optical size effects, and with the interfaces, e.g., optical interface effects. The importance of both increase with decreasing characterizing structural lengths. The role of the host materials, when present, can be different, depending on their concentration: in dilute particle systems they may act as bulk-like homogeneous matter (e.g., the electrolyte in colloidal liquid systems); at low amounts, the host may also develop size effects (e.g., the water layers in the opal).

General Optical Properties

Among the experimental characterization methods most frequently applied to heterogeneous matter, the optical ones are essential, ranging from spectroscopy in the frequency domain (from FIR to UV), to femtosecond analysis in the time domain.

Optics can mean physical optics, i.e., the science of optical material properties, including light emission (the 'active' properties) and light absorption (the 'passive' properties), and instrumental optics, i.e., special experimental techniques based upon propagation of light to analyze such samples.

The most usual experiments for the investigation of linear optical response are refraction, absorption, reflection, and elastic scattering. Also, there are the emission responses and the broad field of nonlinear optical responses.

Refraction, reflection, and absorption are typical of homogeneous matter. Elastic light scattering, on the other hand, can be used to define optical heterogeneous matter, since elastic scattering is a direct consequence of the existence of interior interfaces forming deviations from the homogeneous structure. While the grainy structure on the atomic level in homogeneous matter does not give rise to disturbances of the shape of a plane interacting optical wave, this does occur with heterogeneous matter. Each inhomogeneity gives rise to a wavefront deformation and, hence, to light scattering. Scattering spectra and intensities provide important information about the heterogeneous material, so such optical methods are frequently used for its identification.

Besides scattering, the electrical charging of the interior interfaces induced by the light field, can lead to spectacular multipolar optical absorption features, unseen in related macroscopic homogeneous materials, such as Mie resonance, discussed below.

The optical properties of our model system are composed of various contributions which all act together in a complex way, which can be described as:

1. the optical properties of the single building units. In the case of solid state units, they are characterized by the optical material properties, frequency dependent and, generally, complex dielectric function $\varepsilon(\omega)$ or, alternatively, the (generally complex) refraction index $n(\omega)$, which, in the nanometer-size region, both develop strong variations with size.
2. the optical interactions among all building units in the sample. There are different effects:
    - interference effects among the scattering radiation of the units;
    - electromagnetic coupling effects;
    - the contributions of host materials, if present; and
    - the interactions between heterogeneous building units and host.

The Optical Nanoparticle Resonances (Mie Resonances)

Metallic and semiconductor nanoparticles can exhibit special, spectrally selective absorption and scattering bands which do not have counterparts in the molecular state, in thin film structures, or in the extended bulk. They represent the most spectacular optical properties of matter with nanosized heterogeneities. Therefore they will now be treated in more detail in the frame of our model material of spherical inclusions in a homogeneous embedding medium. We begin with a comparison to thin film structures.

The interband transitions (in Ag below 325 nm) are almost equal in both spectra shown in Figure 2, of a thin Ag film and spherical Ag nanoparticles, respectively, while above 325 nm the flat relaxator spectrum of the conduction electrons in the metal film differs strongly from the oscillator spectrum of the same electrons in the clusters. The inset demonstrates the excitation of the dipolar surface oscillator mode by

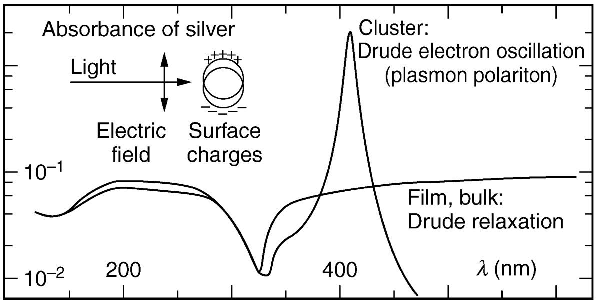

Figure 2 Comparison of absorbances of a thin Ag film and spherical small Ag nanoparticles.

incident polarized light. This figure exemplifies the above-mentioned effect that charging of interfaces by the incident electromagnetic lightwave can give rise to electromagnetic multipole resonances, the Mie-resonances.

The monotonic conduction electron absorption spectrum of the thin film, ranging from UV to zero frequency (the dc current), is thereby compressed in the nanoparticles into the narrow Mie-absorption band with peak position in the blue, containing almost the whole oscillator strength of the conduction electrons. In particular, the Ag-spheres are therefore transparent in the low frequency region, and if we assume the conductivity according to the Drude model as being definitive for a metal, we must assume that these particles are no longer metallic. The conduction electrons no longer follow the Drude relaxator behavior; instead the narrow peak indicates the properties of an oscillator.

The origin of this oscillator and its excitation can be described, being a quantum phenomenon; due to the incident electrical field, the electron plasma, the 'electronic particle', is shifted against the fixed ionic background, the 'ionic particle', causing surface charges modulated with the light frequency. The surface charges produce restoring forces to the plasma, following the periodic field strength, which is the condition required for oscillator behavior. As every free oscillator the particle develops resonances at special eigenfrequencies, and in the case of our metal clusters these are close to the peak frequency of the Mie-absorption, the oscillations of which not being free but forced by the incident field. Analogous to the mechanical case, the resulting resonance frequencies of free and forced oscillations differ slightly. To distinguish them, Mie resonances in metallic particles, coupled to the external electromagnetic field, are called plasmon polaritons.

In the frame of classical electrodynamics each single particle in the heterogeneous material acts as a spherical nano-antenna. This acts as receiving antenna by absorption and, as with every other kinds of antenna, simultaneously becomes a transmitting antenna by re-radiating elastically scattered light.

In the case of Cu, Ag, and Au, and many other metals, these absorption frequencies are in the visible, hence beautiful colors appear. However, there are a lot of metals where this oscillator is overdamped due to low electron mobility. Under particular conditions, such resonances also occur in semiconductor nanoparticles, but due to the lower charge densities, in the IR. At weak fields, the optical response of the oscillator is linear. At higher fields it becomes nonlinear, and thus second and higher harmonics can be excited.

Such unique collective excitations are not restricted to electrons but also spherical surface phonon polaritons can be excited if the particles are nonmetallic (in this case, free electron excitations do not cover the spectral region of phonon excitations). Only in the most simple case of particles which are extremely small compared to wavelength ($R/\lambda \leq 1/20$), the dipolar mode of excitation described so far (i.e., homogeneous polarization of the whole particle) comprises the whole response. For larger particle sizes, a complicated, nonhomogeneous polarization of the whole particle occurs, which can be described by an orthogonal multipole expansion, i.e., multipolar oscillations. The according size (R-) dependent multipolar contributions, which are excited simultaneously and sum up to the total optical response. Further complications arise in larger particles by retardation-induced phase shifts of the lightwaves along the particles.

A complete description of this complex multipole excitation was given by Gustav Mie on the basis of classical electrodynamics. Mie's theory is the exact solution within the frame of Maxwell theory for spherical symmetry, as are the Fresnel-formulae for the planar geometry. However Mie's theory does not describe exactly the conditions which we observe in real heterogeneous systems, hence several extensions of Mie's theory will be presented in the section entitled *Beyond Mie's Theory,* below.

Mie's theory also gives some qualitative insight into the optical behaviour of other kinds of heterogeneous matter. A highly abridged summary of this theory will be presented below.

The Theory of Mie

In the following, an outline of the Mie theory for one spherical particle is given by listing the mathematical steps:

1. Introduction of spherical coordinates r, θ, ϕ. Particle radius R;

2. Incidence of a plane, monochromatic, harmonic electromagnetic wave (circular frequency ω, wavelength λ);
3. Solution of Maxwell's equations for the incident and scattered fields in the surrounding dielectric host medium with (real) dielectric constant $\varepsilon_{\text{matrix}}$, and in the interior of an arbitrary spherical particle of (complex) dielectric function $\varepsilon(\omega)$.

   There are two solutions for transverse fields: one with the radial component of the electric field $E_{\text{r}} = 0$, called 'transverse electric (TE) solution', and one with the radial component of the magnetic field $H_{\text{r}} = 0$, called 'transverse magnetic (TM) solution'. These solutions can be obtained in terms of vector spherical harmonics which can be derived from corresponding scalar potentials Π_{E} and Π_{H}. These are solutions of the Helmholtz-equation:

   $$\begin{aligned} &\nabla^2 \Pi_{\text{E,H}} + K^2_{\text{transverse}} \times \Pi_{\text{E,H}} = 0, \\ &K^2_{\text{transverse}} = \varepsilon_{\text{transverse}} \omega^2/c^2 \end{aligned} \qquad [1a]$$

   In spherical coordinates, the Helmholtz equation is solved by a separation of variables ansatz and a multipole expansion, yielding an infinite number of independent partial solutions with multipolar order L, the TM partial waves or 'electric partial waves', and the TE partial waves or 'magnetic partial waves'.

   A third solution, the potential Γ, that only affects the electric field, can be found as the solution of

   $$\nabla^2 \Gamma = 0, \qquad \varepsilon_{\text{longitudinal}} = 0 \qquad [1b]$$

   and represents longitudinal excitations in the particle.
4. Boundary conditions at $r = R$:
   for the transverse fields:

   $$\begin{aligned} E_\theta^{\text{incident}} + E_\theta^{\text{scattered}} &= E_\theta^{\text{interior}} \\ E_\phi^{\text{incident}} + E_\phi^{\text{scattered}} &= E_\phi^{\text{interior}} \end{aligned} \qquad [2a]$$

   $$\begin{aligned} H_\theta^{\text{incident}} + H_\theta^{\text{scattered}} &= H_\theta^{\text{interior}} \\ H_\phi^{\text{incident}} + H_\phi^{\text{scattered}} &= H_\phi^{\text{interior}} \end{aligned} \qquad [2b]$$

   for the longitudinal fields:

   $$E_{\text{r}}^{\text{incident}} + E_{\text{r}}^{\text{scattered}} = E_{\text{r}}^{\text{interior}} \qquad [3]$$

   (Sauter–Forstmann condition)

   The linear response function of the particle material used here, is ε, where $\varepsilon = \varepsilon_1 + i\varepsilon_2$ is averaged over the particle volume, including the surface–interface region.

   The magnetic permeability μ is set 1 for the investigated high frequency regions.
5. Computation of cross-sections for extinction, scattering, and absorption from Poynting's law:
   - Extinction, i.e., the sum of absorption and scattering losses:

     $$\sigma_{\text{ext}} = \frac{2\pi}{k^2} \sum_{L=1}^{\infty} (2L+1) \text{Re}\{a_L + b_L\} \qquad [4]$$

   - Elastic scattering:

     $$\sigma_{\text{sca}} = \frac{2\pi}{k^2} \sum_{L=1}^{\infty} (2L+1) \left(|a_L|^2 + |b_L|^2 \right) \qquad [5]$$

   - Absorption, i.e., the energy dissipation (production of heat and thermal radiation):

     $$\sigma_{\text{abs}} = \sigma_{\text{ext}} - \sigma_{\text{sca}} \qquad [6]$$

   L indicates the multipolar order of the Mie resonances.

The coefficients a_L and b_L, following from the Maxwell boundary conditions for transverse fields, are

$$a_L = \frac{m\psi_L(mx)\psi'_L(x) - \psi'_L(mx)\psi_L(x)}{m\psi_L(mx)\eta'_L(x) - \psi'_L(mx)\eta_L(x)} \qquad [7a]$$

$$b_L = \frac{\psi_L(mx)\psi'_L(x) - m\psi'_L(mx)\psi_L(x)}{\psi_L(mx)\eta'_L(x) - m\psi'_L(mx)\eta_L(x)} \qquad [7b]$$

with $m = 2x/n_{\text{Matrix}}$, where m denotes the complex-valued index of refraction of the particle $\varepsilon = n'^2$ and n_{matrix} the real-valued index of refraction of surrounding medium. k is the wavevector and $x = k \cdot R$, the size parameter. $\Psi_L(z)$ and $\eta_L(z)$ are Riccati–Bessel functions of the first and third kind. The prime indicates differentiation with respect to the argument in parentheses. Including longitudinal fields, eqn [7] will be slightly modified.

Applying Lambert–Beer's law, these cross-sections are related to the absorption-, scattering-, and extinction-constant $A(\omega)$, $S(\omega)$, and $E(\omega)$, respectively, and the resulting intensity loss ΔI of a many-particle system of number-density Z of the (noninteracting) particles, is

$$\begin{aligned} &\Delta I = I_0 \left(1 - \exp(-Z \cdot \sigma_{\text{ext}} \cdot d)\right), \\ &d = \text{sample thickness} \end{aligned} \qquad [8]$$

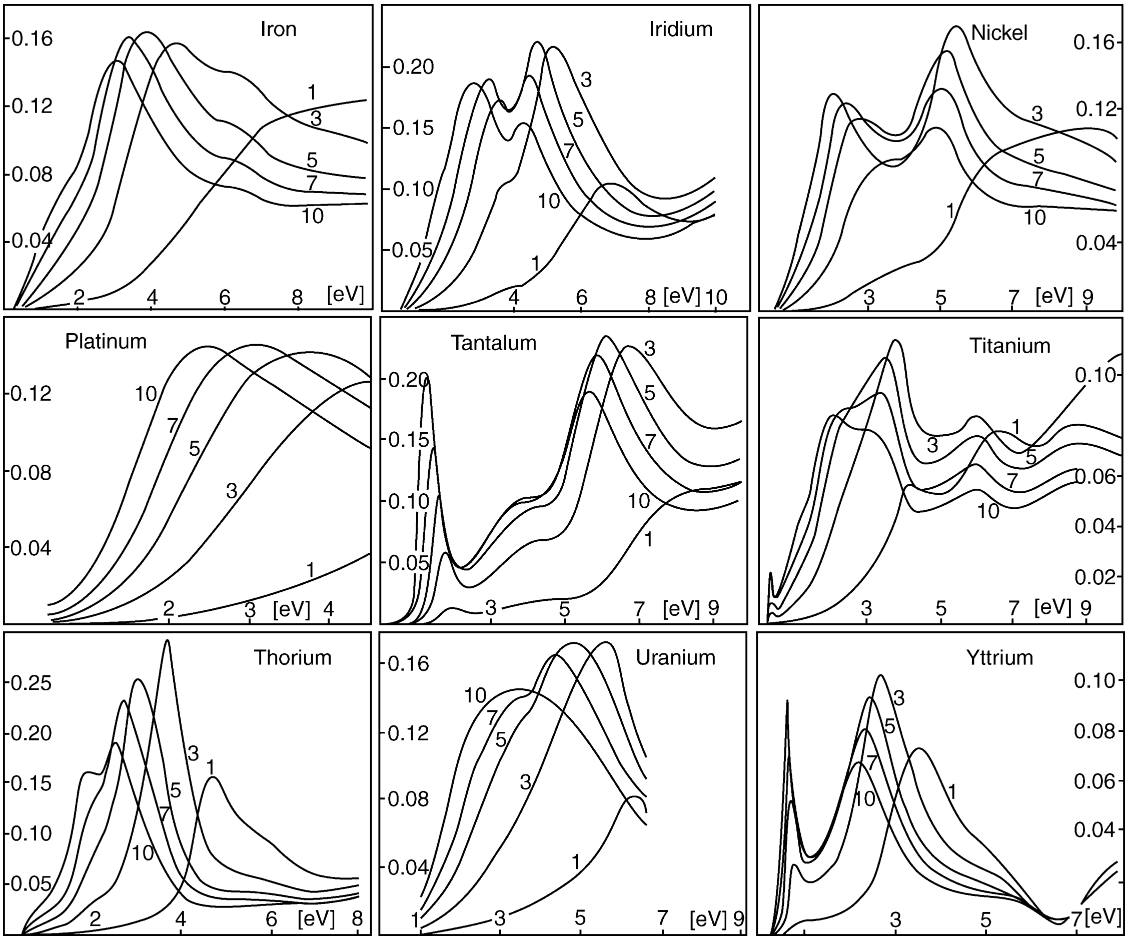

Figure 3 Mie extinction spectra of various particle materials. $2R = 10\,\mathrm{nm}$. The clusters are embedded in matrix materials with dielectric constants between 1 and 10, which are indicated for each spectrum. Obviously, the peak structure of the spectra strongly depends on the matrix.

Figure 3 gives a compilation of some spectra of general interest determined by evaluating Mie's theory.

Beyond Mie's Theory

Application of Mie's theory to experimental results is disappointing because, in most cases, quantitative coincidence is not obtained. One reason is that the simple assumptions incorporated either in Maxwell's theory or in Mie's derivation are not met by reality.

Mie's theory does not apply, if:

1. particle sizes in a diluted cluster matter sample are not uniform;
2. particle shapes differ from the sphere;
3. $\varepsilon_{\mathrm{matrix}}$ varies with frequency, $\varepsilon_{\mathrm{matrix}}(\omega)$
4. cluster-dependent dielectric functions differ from the according functions of the bulk;
5. the (step-function like) Maxwellian boundary conditions at the particle-matrix interface are not applicable;
6. dielectric functions of particles and/or matrix are nonlinear;
7. surrounding/embedding matrix material is absorbing;
8. particle structures are heterogeneous (core-shell structures, multi-grain clusters, etc); and
9. particles form close-packed aggregates instead of being well-separated, and hence, electromagnetic coupling among particles is important.

Of course, this list may be continued. Regarding realistic samples, it is obvious already from these deviations that quantitative correspondence between experiment and Mie theory usually cannot be expected. But, over time, each of these extra effects has been treated and included into extensions of Mie's theory, which should therefore be applied, instead of the original theory. In the following several of these effects will be treated in some detail.

Particle Shapes Deviating From the Sphere

Even metals of high structure-symmetry, like the fcc-Au, can yield nanoparticles with low symmetry habits. Figure 4 shows Au nanoparticles with exotic shapes ranging from triangle and hexagonal platelets over fivefold symmetry multitwin-particles to rodlike shapes. The majority of particles are multigrain structures of irregular isometric shapes.

Multishell Particles

In experiments, particle surfaces often react chemically with the surrounding media, thereby forming compound shells. Core/shell structures can be clearly visible in the optical absorption spectra. Figure 5 shows, as an example, the absorption spectra of nanoparticles of Ag core/Au shell and of Au core/Ag shell, respectively with varied shell thicknesses.

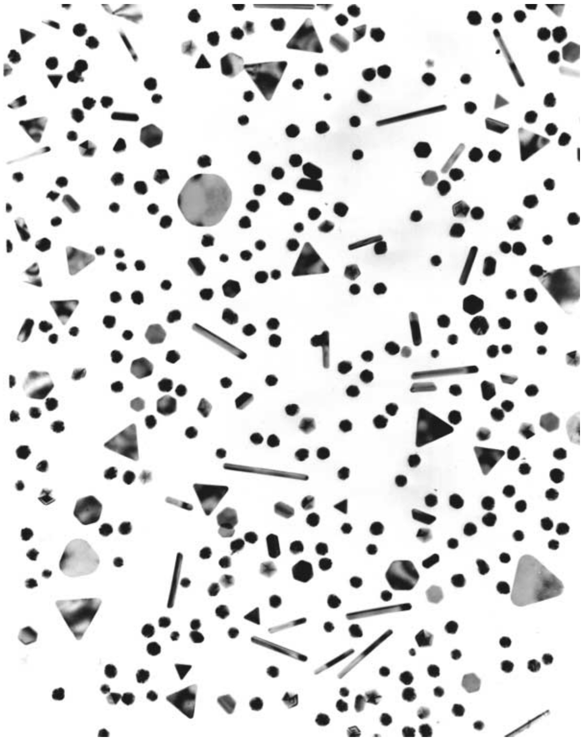

Figure 4 Chemically (Esiguondy method) produced Au clusters with extraordinary variety of different shapes. (Of special interest are the small particles of five-fold symmetry.) Mean sizes of the irregular, almost spherical particles: 36 nm.

Resonances will also occur, e.g., in metal-coated dielectric particles. An example is shown in Figure 6, with extinction spectra of silver-coated titanium-dioxide particles, computed by an extension of Mie's theory.

Aggregates of Nanoparticles with Electrodynamic Particle–Particle Coupling

The optical extinction and scattering spectra of many-particle systems with dense packing forming aggregates (see Figure 4) are strongly influenced by electromagnetic interactions: each particle feels the

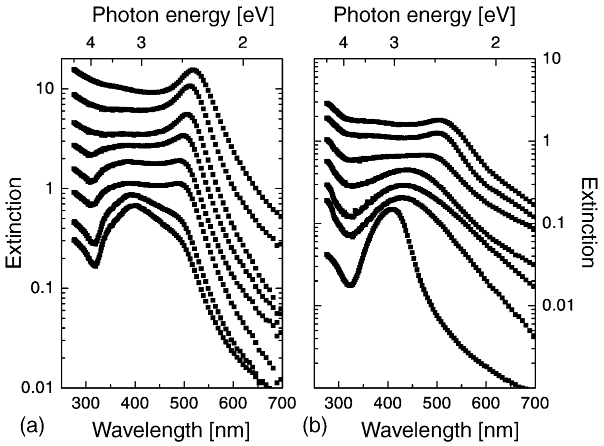

Figure 5 Measured extinction spectra of (a) silver-coated gold particles (gold core $2R = 17$ nm, shell thickness varying from $d = 0$ to $d = 3.6$ nm from top to bottom), and (b) gold-coated silver particles (silver core $2R = 18$ nm, shell thickness varying from $d = 0$ to $d = 5$ nm from bottom to top).

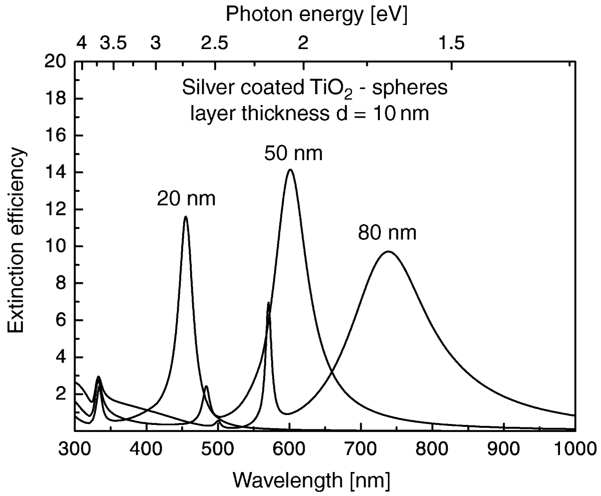

Figure 6 Calculated extinction spectra of spherical silver-coated titanium dioxide particles. The shell thickness is kept constant at $d = 10$ nm, the core sizes amount to $2R = 30$ nm, 50 nm, and 80 nm, respectively, as indicated in the graph.

scattering near fields of all its neighbors, which give rise to splitting, broadening, and shifts of resonances. This is clearly demonstrated with colloidal Ag and Au particles in Figure 7.

The splitting, broadening, and shifts can be assigned to the shape and size of the resulting many-particle aggregates. An example for the influence of the topology of the many-particle system is given in Figure 8 with extinction and scattering spectra of silver particle aggregates of $N = 5$ identical silver particles with $2R = 40$ nm, which were computed by the Generalized Mie Theory (GMT). As a rough rule-of-thumb, such electrodynamic coupling effects of the oscillators only can be ignored if the smallest next neighbor center-to-center distance d_{cc} exceeds $d_{cc} \approx 5 \cdot R$.

Size Effects

The extraordinary practical success of Mie's theory is based upon the fact that this example of classical electrodynamics enables us to compute numerically the optical response for arbitrary realistic particle materials. In contrast quantum theoretical calculations are restricted, beyond the molecule-solid state transition, to jellium and jellium-related hypothetical material.

Classical electrodynamics, being a phenomenological theory to describe light propagation, does not yield information about optical material properties. These enter via the dielectric functions inserted into the Maxwell boundary conditions. They must be taken from elsewhere, such as from experiments or from quantum solid state theory model calculations. This twofold basis of Mie's theory is illustrated in Figure 9.

There are two different sources for particle size dependences of the optical extinction spectra:

1. The electrodynamics: the larger the particles, the larger are the number of contributing multipoles and the influences of retardation effects, both influencing strongly the optical spectra;
2. The optical material properties describing the particle material which, in the nanoscale size region, exhibit size dependences. These size dependences can be treated by corrections entering the bulk material function.

In general, the dielectric function $\varepsilon(\omega)$ of a metal or semiconductor is the sum of susceptibility contributions of the lattice, of the conduction electrons, and of the interband (electron–hole) excitations:

$$\begin{aligned}\varepsilon(\omega) &= \varepsilon_1(\omega) + i\varepsilon_2(\omega)\\ &= 1 + \chi^{\text{lattice}}(\omega) + \chi^{\text{free electrons}}(\omega)\\ &\quad + \chi^{\text{interband}}(\omega)\end{aligned} \qquad [9]$$

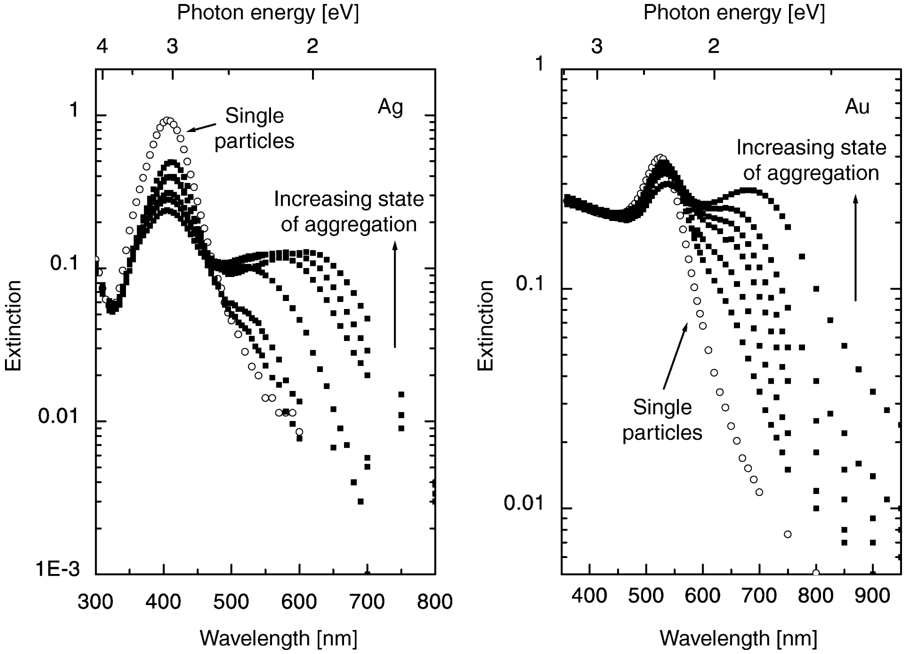

Figure 7 Measured extinction spectra of aggregated samples of (a) silver particles with $2R = 28$ nm, and (b) gold particles with $2R = 38$ nm. The samples contain random distributions of aggregates of various topologies, but with increasing state of aggregation from bottom to top.

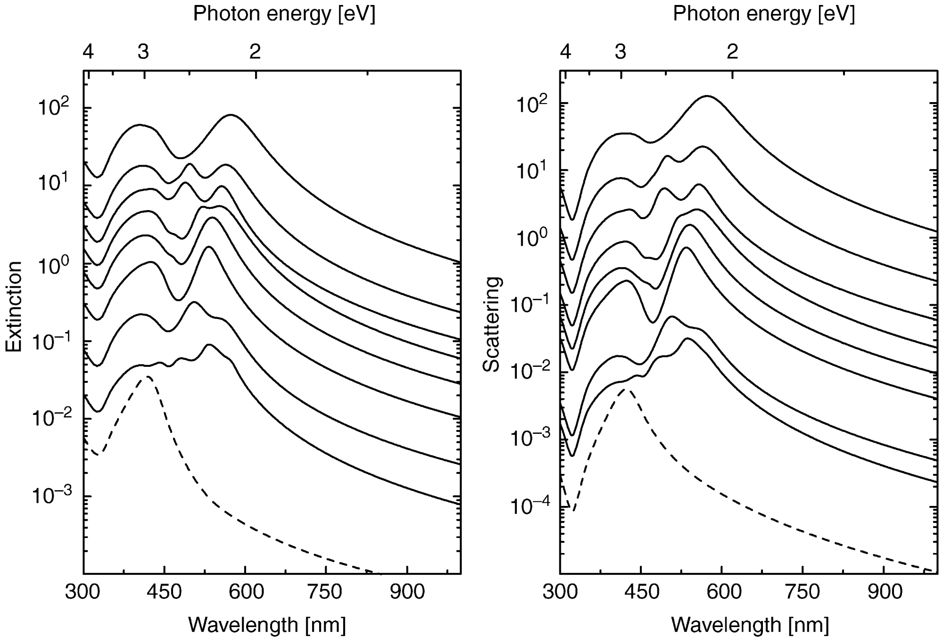

Figure 8 Calculated extinction and scattering spectra of aggregates with $N = 5$ almost touching identical silver spheres of $2R = 40$ nm. From top to bottom, the spectra belong to the following particle arrangements in the aggregate: . The dashed line represents the single particle spectra.

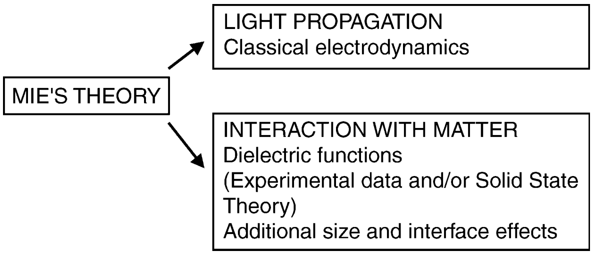
MIE'S THEORY
LIGHT PROPAGATION
Classical electrodynamics
INTERACTION WITH MATTER
Dielectric functions
(Experimental data and/or Solid State Theory)
Additional size and interface effects

Figure 9 Scheme of the basics of Mie's theory.

A detailed expression was given by Bassani (here, the lattice contributions, mainly important in the IR, were disregarded):

$$\varepsilon(\omega) = 1 - \underbrace{\frac{(ne_0^2/\varepsilon_0 m_{\text{eff}})}{\omega^2+i\gamma\omega}}_{\text{conduction electrons}} + \underbrace{\chi^{\text{interband}}(\omega)}_{\text{interband transitions}} \qquad [10]$$

with

$$\chi^{\text{interband}} = \frac{2\hbar^2 e^2}{\varepsilon_0 m_{\text{eff}}^2} \sum_{i,f} \int_{\text{BZ}} \frac{2d^3k}{(2\pi)^3} |\mathbf{e}\times M_{\text{if}}(\mathbf{k})|^2 \times \left\{ \frac{1}{[E_f(\mathbf{k})-E_i(\mathbf{k})][(E_f(\mathbf{k})-E_i(\mathbf{k}))^2-\hbar^2\omega^2]} + i\frac{\pi}{2\hbar^2\omega^2}\delta(E_f(\mathbf{k})-E_i(\mathbf{k})-\hbar\omega) \right\} \qquad [11]$$

In this equation, the quantity n is the conduction electron density, $\mathbf{k}$ the wavevector of the electrons with norm k, and BZ denotes the Brillouin Zone. The vector $\mathbf{e}$ is the polarization vector of the incident light. M_{if} is the transition matrix element with i = initial and f = final states.

To which degree of approximation a single dielectric function can describe the optical response of a nanoparticle consisting of inner core and surface with differing local optical properties (i.e., polarizabilities), and how this function has to be modelled as an average, is still a problem. An easy, but only approximate answer is to identify terms in the eqns [10] and [11], which may change if the cluster size is reduced down to the nanometer scale, and to introduce correction terms.

In the case of the conduction electrons of eqn [10], these are the electron density n, the effective electron mass m_{eff}, and the relaxation frequency γ. In the terms of the interband transitions (index inter), these are the transition matrix elements M_{if} and the band structure energies of the initial and final states E_i and E_f. All of these quantities change, when the particle size is reduced, in different size regions and to different amounts, and detailed quantum theoretical investigations are required to derive proper correction terms.

The application of the continuum-based dielectric functions and of Mie's theory fail when the molecule-solid state transition region is reached.

The effects of size-dependent dielectric functions upon cluster absorption spectra are demonstrated in the series of absorption spectra of Figure 10, measured on highly diluted heterogeneous systems containing spherical Ag, Au, and Cu particles which were embedded in a glass matrix. If $\varepsilon(\omega)$ were independent of size, then the spectra should be identical, in the investigated particle size region. So, their differences reflect clearly that $\varepsilon = \varepsilon(R)$.

To describe the size effects of the conduction electrons, the relaxation frequency γ of eqn [10] has been modified introducing a parameter that is called A-parameter which also contains size effects and also particle–host interface effects. To explain the electronic size effects in detail, quantum effects have to be considered taking the discretization of the bandstructure into account, but they can also be treated by a simple classical approach: this is the mean free path effect (FPE), assuming the mean flight time of the effective conduction electrons (i.e., those with Fermi energy E_{F}) between subsequent surface collisions amount to

$$\tau\gamma_{\mathrm{FPE}} = A_{\mathrm{FPE}}\frac{v_{\mathrm{Fermi}}}{R} \qquad [12]$$

The A-parameter is a measure of the effectiveness of the single collision in view of relaxation. Complete 'memory loss' of the drift momentum in each collision would lead to $A_{\mathrm{FPE}} = 1$. Taking electron correlation effects and energy relaxation into account, A-values larger than 1 may also be realistic. Values larger than unity also occur in multigrain particles, due to grain boundary 'collisions'.

Various quantum mechanical models of electronic particle size effects result in relations, formally coinciding with the $1/R$-dependence in eqn [12]. Most relevant is the value of A for particles *in vacuo* with clean surface. This value has been determined theoretically and experimentally to $A \approx 0.25$ for Ag particles.

Properties of Internal Interfaces

Structures of Surfaces/Interfaces

Only mathematical interfaces in geometry can, in fact, be two-dimensional planes separating two different volume elements. This is assumed in the frame of Maxwell boundary conditions entering Mie's theory. In real nanostructured matter, however, the interface regions are three-dimensional layers formed by disturbances of atomic and/or of electronic structures, and/or of chemical compositions. In addition, due to the finite wavelengths of the conduction electrons, the according electron density 'spills out' and, as a consequence, we find electron density beyond the geometric surface. It should be kept in mind that these effects hold, with proper changes, for interfaces in nanostructured materials in general. Hence, an understanding of the interface regions is of key importance for all nano-material.

Electronic Interface States

Up to now, this discussion has been restricted to particles with uncovered, 'clean' surfaces. In most practical examples of heterogeneous materials however, the building units are embedded in a host matrix, deposited upon the surface of a substrate, or are in close contact to neighboring crystallites, etc., to stabilize the samples. Then, special electronic

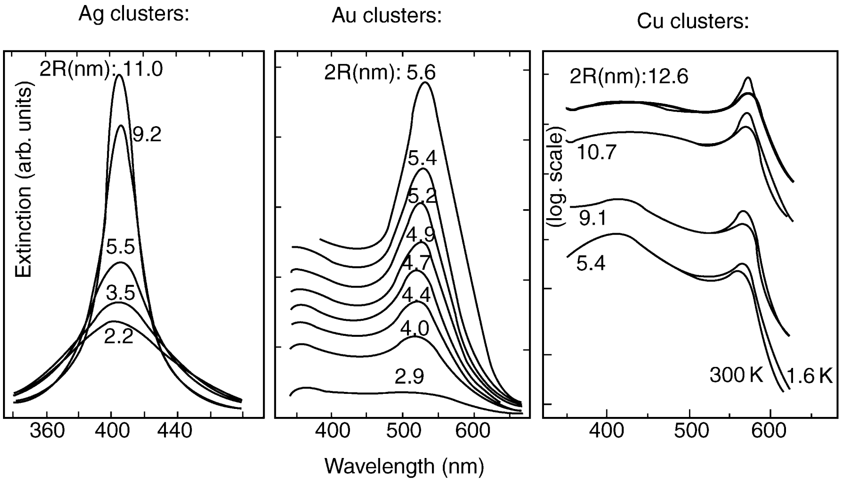

Figure 10 Measured extinction spectra of Ag, Au, and Cu clusters of various mean sizes, grown in glass matrix. In the case of Cu-particles, measurements at 1.5 K are added.

interface states are created by contributions of particle electrons and host electrons, which are (disregarding their density on the surface) similar to adsorbate states.

One might assume that the broad knowledge about these states in surface science, obtained from samples with extended crystallographic planes, should be very useful here. However, the surface of a cluster, in principle, differs from the planar surface of a single crystal by containing a multitude of different local surface structures, atomic sites, and varying atomic coordination numbers. Even if the surface structure is not irregular (as is usual in nanoparticles), the surface consists of several facets of different orientation separated by edges and corners. The coordination numbers of surface atoms, thus, vary strongly (e.g., between 5 at a corner and 9 on a plane), indicating surface sites to have different energies.

We have shown that, if the particles are metallic, the 'particle/host'-interface states give rise to two different electronic charge transfer processes of particle conduction electrons, the static and the dynamic charge transfer. This will be demonstrated qualitatively for metal particles by the example presented in Figure 11.

The Static Interface Charge Transfer and its Effects Upon Mie Resonance Positions

At the left side of Figure 11, the conduction band of the metal cluster close to its surface is shown, which is occupied with electrons up to the Fermi energy E_F. Chemisorption of a free atom (one energy level E_a of which is plotted at the right-hand side) is assumed to occur by approaching the atom to the cluster surface. The resulting energy states are also shown. Let us assume the marked electronic level E_a in the free atom is nonoccupied. During the chemisorption process, this level is shifted and broadened, developing density of nonoccupied states below the particle Fermi energy, which is localized at the three-dimensional interface region. After complete the process is, such a state will be occupied permanently by electrons of the cluster (static charge transfer), and the nanoparticle will have lost electrons, thus its electron density being reduced. The driving impact of this process is the equilibration of the chemical potential in the whole cluster-matrix system.

Now assume we embed a cluster into some surrounding material. Then all those atoms of the latter, which are situated directly at the interface, become chemisorbed at the cluster. Hence remarkable amounts of charges can be transferred into the interlayer region and an electrical double layer is created. As an example, a very strong amount of static charge transfer was observed experimentally when applying solid Fullerite (C_{60} molecules) as embedding matrix for Ag-nanoparticles: here, each C_{60} in direct contact with the Ag particle attracts one electron, forming (C_{60}^-). Hence, in this experiment with 500 atoms large particles, in total about 20% of the conduction electrons of the Ag-particle were transferred into the interlayer. Strictly speaking, such a nanoparticle of the heterogeneous material is no longer a silver particle because of the remarkable changes of electronic properties by static charge transfer.

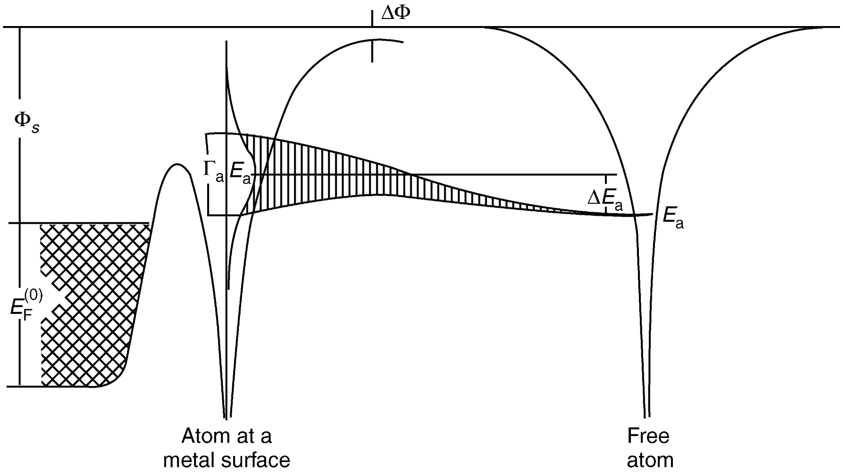

Figure 11 Electron energy scheme for atomic adsorbation at a metal particle surface. The free atom (energy spectrum: right) approaches the particle surface (left; conduction band with Fermi-level E_F). Its energy level E_a is thereby shifted and broadened (including ensemble broadening of many atoms deposited at many clusters). The atoms are assumed to be separated from the metal by a tunnel barrier (see text).

Some consequences are listed in the following:

In a first step, let us assume that the electronic bandstructure is not altered by the charge transfer. Then we expect: (i) changes of level occupation; (ii) changes of Fermi energy and Fermi velocity; (iii) changes of effective masses close to E_F; (iv) changes of surface 'spill out'; (v) additional adsorbate levels and special surface states; and (vi) changes of the optical interband transition edge.

In a second step, we may consider possible changes of electronic bandstructure (e.g., energies) and, finally, also changes of atomic structures (i.e., lattice structures and/or surface structures).

Since plasmon frequencies depend on the conduction electron density, the resulting Mie resonance can monitor and give quantitative information of this effect. We find shifted by the static charge transfer:

$$\Delta\omega_{\text{resonance}} \approx (\sqrt{n_2} - \sqrt{n_1}) \times \sqrt{\frac{e_0^2}{\varepsilon_0 m_{\text{eff}}(2\varepsilon_{\text{matrix}} + 1 + \chi_1^{\text{interband}})}} \qquad [13]$$

with n_1 being the conduction electron density in the cluster *in vacuo*, and n_2 being the electron density after contact with foreign material (adsorbate, substrate, or host-matrix). This peak shift of the Mie resonance, however, is superimposed in the experiment by a much larger shift, following simply from the change of Maxwell boundary conditions by the altered dielectric constant of the particle surrounding $\varepsilon_{\text{matrix}}$, when going from vacuum to a matrix material. Hence, the experimental determination of static charge transfer processes is rendered more difficult.

The Dynamic Interface Charge Transfer and its Effects Upon Mie Resonance Widths

After the static charge transfer in Figure 11 has taken place, electrons close to E_F can fluctuate between the cluster and empty adsorbate levels, if present at E_F. The residence times in these levels, which are usually separated from the Fermi sea of the metal cluster by some tunnel barrier, vary statistically. If a plasmon is excited, i.e., mainly a collective, phase ordered motion of all conduction electrons superimposed to their statistical motion, we may assume that, after such electron transfer back and forth, the reminiscence to the primary drift momentum, and the phase of this excitation, has been lost. Thus, the phase relaxation of the plasmon is increased.

The consequences of this dynamic charge transfer effect ('chemical interface damping') have been investigated in detail. Here, we only state that the directly observable main consequence is the broadening of Mie resonances. If the particles are sufficiently small to neglect radiation damping (this is the case for Ag clusters smaller ≈15 nm), band broadening is only due to internal energy dissipation and the band width can be modeled quantitatively with the modified relaxation frequency γ in eqn [10] which yields an additive correction term:

$$\Gamma(A, R) \approx \Gamma_0 + \frac{2ne_0^2}{\varepsilon_0 m_{\text{eff}}\omega^3} \times \frac{A}{\sqrt{(\partial\varepsilon_1/\partial\omega)^2 + (\partial\varepsilon_2/\partial\omega)^2}} \frac{v_{\text{Fermi}}}{R} \qquad [14]$$

with the A-parameter extended to the additional electronic excitations in the particle/host interlayer:

$$A = A_{\text{size effects}} + A_{\text{interface effects}} \qquad [15]$$

The A-parameter of eqn [15] now expresses both the strengths of the particle size effects and of the chemical interface damping, due to statistically varying occupation of interface states near E_F.

An important feature of this model is that the dependence on the kinds of materials is described by specific material parameters ($A_{\text{size effects}}$ and $A_{\text{interface effects}}$), and the direct size dependence effects are given by an $1/R$-dependence, for both the direct size effects and the interface effects. For one fixed cluster size R, the magnitude of $A_{\text{interface effects}}$ directly characterizes the probability of the single temporal charge transfer, including the number and energies of contributing interface states; hence, important properties of the adsorption process can be investigated by determining the A-parameter.

Experimental Results

In the experiments presented in the following, we deal with the interfaces between clusters and a more or less homogeneous solid embedding medium. We selected very small silver nanoparticles of about 2 to 4 nm size with accordingly large ratio of surface- to volume-atoms, in order to optimize the sensitivity of the measurements. Silver is unique in having the most pronounced Mie resonance of all known materials.

In Figure 12 two classes of matrix materials are presented, fluorides and oxides. The absorption spectrum of the free particles with uncovered 'clean' surface (measured in a cluster beam in UHV) is added in both diagrams which is only influenced by size effect. General features are, that embedding in a host medium causes:

- drastic Mie resonance peak shifts toward IR, depending on the dielectric constant of the matrix

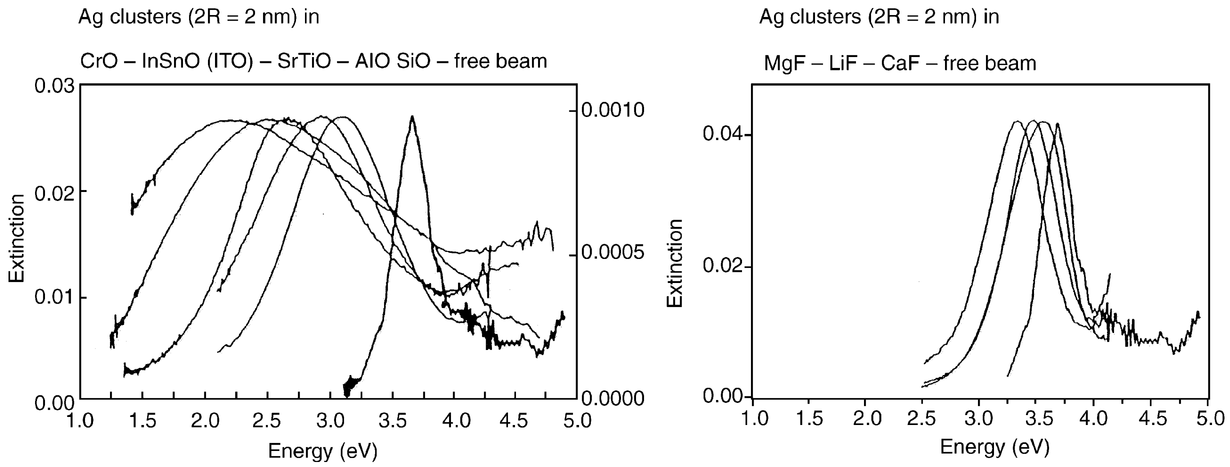

Figure 12 Measured optical extinction spectra of 2 nm Ag nanoparticles, embedded in various different embedding materials. For comparison, the spectrum of the free nanoparticle beam is added. Left side: oxidic materials, right side: fluoridic materials.

$\varepsilon_{\mathrm{matrix}}$, which classically follow from Mie's theory and by comparingly small additional shift due to static charge transfer;

- drastic broadening (exceeding, in some cases, all known size effects), which according to eqn [14], can be described by characterizing values of the *A*-parameter. Recently, these parameters have been experimentally determined for Ag nanoparticles embedded in a broad variety of host materials. Their interpretation is rendered difficult, however, due to inhomogeneous broadening of varying magnitude.

We can immediately judge from Figure 12 that both shift and broadening are larger in the case of oxidic matrices than of fluoridic ones, i.e., oxides possess more interface states close to E_F of the material. In Figure 13, observed peak positions $\hbar\omega_{\max}$ and broadenings expressed in the resulting *A*-parameters, are summarized for a broad variety of different embedding materials.

Comparison of Plasmon Polariton Lifetimes with Femtosecond Experiments

In the last section, experimental examples were presented to show how fundamental physical properties are changed by forming heterogeneous materials. Because the Mie resonance widths are mainly a consequence of finite plasmon polariton life times τ, (in addition, there are interband contributions, in larger particles radiation damping, and dispersion influence), the latter can be evaluated from measured optical absorption spectra. This supposes the bandwidth to be essentially homogeneous. From a width analysis fitting Mie's theory numerically we

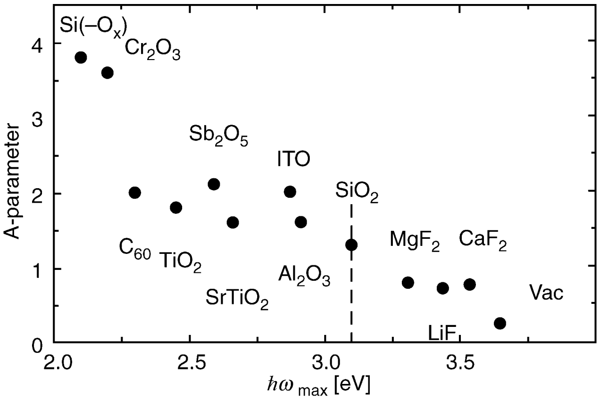

Figure 13 Silver nanoparticles ($2R = 2$ nm) embedded in various solid host matrices: measured Mie peak positions and resonance halfwidths expressed by the *A*-parameters.

obtained for Ag particles, free flying in UHV with uncontaminated free surfaces and 2 nm in size, a plasmon lifetime of $\tau = 6$ fs. Under these experimental conditions, the measured bandwidth is the homogeneous one.

Concerning the size dependences, we can extrapolate to larger particles. The longest lifetimes are expected for 15 nm particles, for which $\tau = 15$ fs was obtained by extrapolating the experimental data for 2 nm particles to this size. The same width analysis was performed for the nanoparticles embedded in matrix materials, and as a result, the lifetimes drastically decrease down to 1–2 fsec. The observed increase of the *A*-parameter, as shown in the previous section, is thus directly correlated to the measured decrease of the plasmon polariton lifetime. Hence, we conclude that the decrease of the lifetime is mainly due to increased phase relaxation caused by the dynamic charge transfer effect.

These results can be compared with recent femtosecond spectroscopy experiments which have recently given much impact on the experimental investigation of the dynamics of Mie resonances in silver and gold nanoparticles. In these experiments data between 6 and 10 fsec were obtained for deposited Ag and Au particles of different sizes and different substrates. Regarding the complexity of the different experiments the correspondence of the data appears convincing.

Obviously the lifetimes evaluated in the frequency domain, from our spectra of particles *in vacuo*, coincide with the lifetimes determined in the time domain. The latter, though, were measured with larger particles deposited on substrates (i.e., not from free particles but including interactions with substrates).

As a concluding remark, it may be pointed out again, that the optical properties and the heterogeneity effects presented here for the model system of nanoparticle/host-matter, are also to be found in other examples of the large field of heterogeneous materials.

List of Units and Nomenclature

Absorption bandwidth	Γ
circular frequency [$\mathrm{s^{-1}}$]	ω
dielectric constant/dielectric function [–]	ε
effective electron mass [kg]	m_{eff}
electric field [$\mathrm{V\,m^{-1}}$]	E
(electron) relaxation frequency [$\mathrm{s^{-1}}$]	γ
electron state energies for initial and final state [J]	E_i, E_f
elementary charge [C]	e_0
extinction, scattering, absorption cross section [$\mathrm{m^2}$]	σ_{ext}, σ_{sca}, σ_{abs}
Fermi energy [J]	E_{F}
Fermi velocity of electrons [$\mathrm{m\,s^{-1}}$]	v_{Fermi}
intensity [$\mathrm{W\,m^{-2}}$]	I
Kronecker energy distribution [$\mathrm{J^{-1}}$]	$\delta(E)$
magnetic field [$\mathrm{A\,m^{-1}}$]	H
norm of wave vector of electrons [$\mathrm{m^{-1}}$]	k
number density of conduction electrons [$\mathrm{m^{-3}}$]	n
Riccati-Bessel functions of first kind [–]	$\psi_n(z)$
particle radius [m]	R
(phase-) life time [s]	τ
Planck constant [$\mathrm{J_s}$]	$\hbar$
Riccati-Bessel functions of third kind [–]	$\eta_n(z)$
scalar electromagnetic potentials [V]	Π, Γ
susceptibility [–]	χ
transition matrix element [$\mathrm{kg\,m\,s^{-1}}$]	M_{if}
vacuum permittivity [$\mathrm{CV^{-1}\,m^{-1}}$]	ε_0
vacuum velocity of light [$\mathrm{m\,s^{-1}}$]	c
vacuum wavelength of light [m]	λ
wave vector of electrons [$\mathrm{m^{-1}}$]	$\mathbf{k}$

See also

Scattering: Scattering Theory. **Semiconductor Physics:** Polarons.

Further Reading

Bassani F and Parravicini G (1975) *Electronic States and Optical Transitions in Solids*. Oxford: Pergamon.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.

Born M (1960) *Optik*. Berlin: Springer-Verlag.

Duncan M (ed.) (1998) Advances in Metal and Semiconductor Clusters, vol. IV. London: JAI Press Inc.

Haberland H (ed.) (1994) *Clusters of Atoms and Molecules*. Berlin: Springer Series in Chemical Physics 56.

Nalwa HS (ed.) (2001) *Handbook of Surfaces and Interfaces of Materials*, vol. 3. San Diego: Academic Press.

Hölzl J, Schulte F and Wagner H (1979) *Solid Surface Physics*. Springer Tracts in Modern Physics 85. Heidelberg: Springer.

Hummel R and Wissmann P (eds) (1997) *Handbook of Optical Properties*, vol. II: Optics of Small Particles, Interfaces and Surfaces. Boca Raton, FL: CRS Press.

Jena P, Khanna SN, Rao BK (eds) (2000) *Cluster and Nanostructure Interfaces*. New Jersey: World Scientific.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kreibig U and Vollmer M (1995) *Optical Properties of Metal Clusters*. Berlin: Springer Series in Material Science 25.

Nanometer Scale Science and Technology, Internatl. School of Physics 'Enrico Fermi' (2001) CXLIV. Amsterdam: IOS Press.

Quinten M (1998) *Physical and Chemical Properties of Surfaces in Nano-Particle Systems*. Aachener Beiträge zur Physik kondensierter Materie, vol. 24. Mainz, Aachen: Wissenschaftsverlag.

Ruhbahn HG (2002) *Nanophysik und Nanotechnologie*. Stuttgart: Teubner.

Schmid G (1994) *Clusters and Colloids*. Weinheim: VCH Verlag.

van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: Wiley.

von Plessen G (1998) *Ultrafast Dynamics of the Surface Plasmon in Gold Particles*. Advances in Solid State Physics 253. Berlin: Vieweg.

Lightweight Mirrors

J W Bilbro, George C Marshall Space Flight Center, AL, USA

Published by Elsevier Ltd.

The desire to probe the heavens is as old as humanity; the realization that the human eye could be aided in these investigations is somewhat more recent. In May 1609, Galileo Galilei, while visiting in Venice, heard about the new 'telescope' invented by the Dutch lens-maker Hans Lipershay and decided that it would be useful in his astronomical investigations. He returned to his home in Padua and made his own telescope comprised of 4.2 cm diameter plano-convex and plano-concave lenses. With this telescope, Galileo discovered the satellites of Jupiter; got tenure; had his salary doubled – and the race was on! It did not take too long to realize that 'bigger' was 'better', and from this point forward, there has been a continuous striving to increase collecting area and resolution, thereby enabling the astronomer to 'see' farther and farther into the heavens.

The limitations of refractive optics soon prompted investigations into reflecting optics – mirrors. In 1668, Isaac Newton produced the first reflecting telescope having a 2.5 cm diameter mirror made of speculum (copper (6), tin (2), and arsenic (1)). A little over 100 years later, in 1789, William Herschel constructed his 'Great Telescope', having a 1.22 m diameter mirror made of bronze. Of course as telescopes grew, so did mirror weight. In 1845, William Parsons' 'Leviathan' was built. It had a 1.84 m diameter mirror made of speculum that weighed 3600 kg – a whopping 1354 kg m^{-2}.

Problems with corrosion in speculum and bronze mirrors led to a renewed interest in glass, and the development of highly reflective coatings soon made metal mirrors obsolete. In 1917, the mirror for the Hooker telescope arrived on Mount Wilson. Made by the Saint Gobain glassworks in France from the same material used for wine bottles, the mirror was 2.54 m in diameter, 33 cm thick and weighed over 4000 kg (789 kg m^{-2}).

The increasing weight of larger and larger mirrors produced practical problems for telescope makers and astronomers. Herschel's telescope was so unwieldy that he found it almost impossible to use and returned to smaller telescopes for most of his observations, and the 2.54 m Hooker telescope required a 100 ton mount and drive system. Obviously, if telescope apertures were to increase, steps had to be taken to make mirrors lighter. The first major success in this arena was accomplished for what was to become the Hale Telescope commissioned at the Palomar Observatory in 1948. In 1936, Corning cast a mirror made of a new borosilicate glass blend called Pyrex. The blank was lightweighted by placing hexagonal shaped obstructions in the mold which created corresponding voids in the glass. The finished mirror (Figure 1) was 5.08 m in diameter, 61 cm thick at the edge and weighed approximately 13 000 kg (641 kg m^{-2}). Even with this degree of 'lightweighting' the mounting and drive mechanisms still weighed around 182 000 kg.

Although the desire for larger ground-based telescopes provided the initial impetus for developing lightweight mirrors, the advent of the space program and the desire to put telescopes in space became the main driver in the pursuit of 'serious' lightweighting techniques. The cost of launching

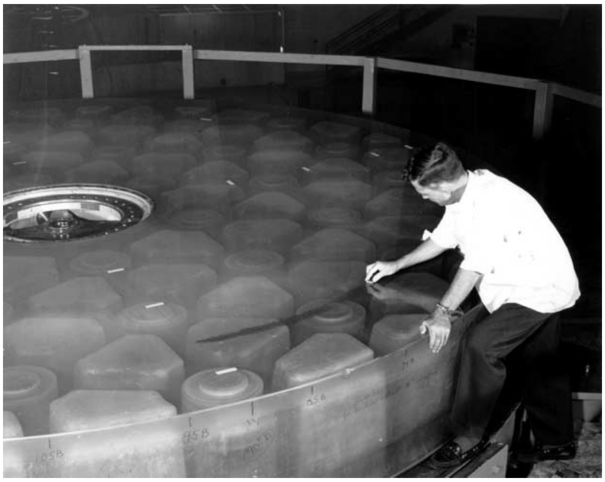

Figure 1 5.08 m Pyrex™ mirror made by Corning/Caltech for the Hale Telescope. Reproduced with permission from the Palomar Observatory/California Institute of Technology.

large telescopes into space quickly becomes prohibitive as aperture size increases. Consequently, the current emphasis is on developing ultra-lightweight, deployable telescopes that permit ever-increasing apertures to be placed in orbit. Fortunately, the zero gravity environment of space permits satisfactory operation of such ultra-lightweight telescopes and therefore the extent to which a mirror can be lightweighted is primarily dependent upon the mirror material and the manufacturing techniques utilized.

Manufacturing

Mirror manufacturing can, in some sense, be divided into two general classes, one that removes material (direct generation) and one that deposits material (replication). The two classes have significantly different limitations relative to producing lightweight mirrors. Direct generation encompasses the more traditional methods of mirror manufacturing, blank fabrication, and optical surface generation, i.e., grinding and polishing. Many innovative techniques have been developed over the past decade that have not only made significant strides in improving fabrication times, optical figure, and surface finish, but have also made possible the reduction of weight to unprecedented levels.

Although extremely lightweight mirrors are being fabricated using direct generation, in the end, there are limitations in the fabrication processes that are not present in replication. This process offers the fascinating potential to create mirrors that are only a few molecules thick, thereby bringing a new meaning to the term ‘lightweight optics’.

Direct Generation

Mirror blank fabrication

Casting lightweight glass mirror blanks has evolved substantially since 1936. Four 8.2 m diameter Zerodur™ mirror blanks were spun cast for the European Southern Observatory’s Very Large Telescope Interferometer. The largest mirror cast to date was the 8.4 m borosilicate mirror cast in a spinning oven for the Large Binocular Telescope on Mount Graham. Although not exactly lightweight by space standards – the blank weighs 15 455 kg ($266\ \mathrm{kg\ m^{-2}}$) – it is substantially less than a conventionally cast solid blank.

It has been postulated that an areal density of around $64\ \mathrm{kg\ m^{-2}}$ probably represents the limit of lightweight glass castings. Therefore, as the demand for lighter and lighter mirrors increased, methods other than casting had to be developed. Investigations in the 1960s indicated that mirrors could be lightweighted to about $180\ \mathrm{kg\ m^{-2}}$ by first fabricating a ‘honeycomb’ core and then fusing plates on the front and back of the core using a high-temperature fusion process. This process was used to produce the Hubble Space Telescope mirror (Figure 2). Hubble had a 2.5 cm thick face-sheet and back-sheet and a honeycomb interior constructed of individual pieces 0.5 cm thick. These pieces were assembled and placed in a large furnace where the temperature was increased until they fused together. The resulting 2.4 m diameter mirror weighed 828 kg ($183\ \mathrm{kg\ m^{-2}}$).

During the 1980s, Frit bonding technology was developed. Unlike the fusion process, Frit allowed the bonding of glass elements at a lower temperature. This allowed the cross-sections of the glass elements to be significantly reduced. The thicknesses of face-plates were typically reduced to around 7 mm and areal densities in the $40-50\ \mathrm{kg\ m^{-2}}$ range became routine. New advances were also made in core manufacturing. Diamond tooling was developed along with computer controlled systems to machine cores from solid blanks. High-pressure water jets were developed that allowed cores to be made much faster than with diamond tooling. In addition, the water jet cutting provided the ability to create very flexible geometries optimized to minimize weight while retaining stiffness. Water jets could machine core walls that could be held to a thickness of around 1.5 mm over a core depth of 20 cm. Computer controlled diamond machining has created even thinner cores (less than 1 mm in thickness) but the process requires longer generation times.

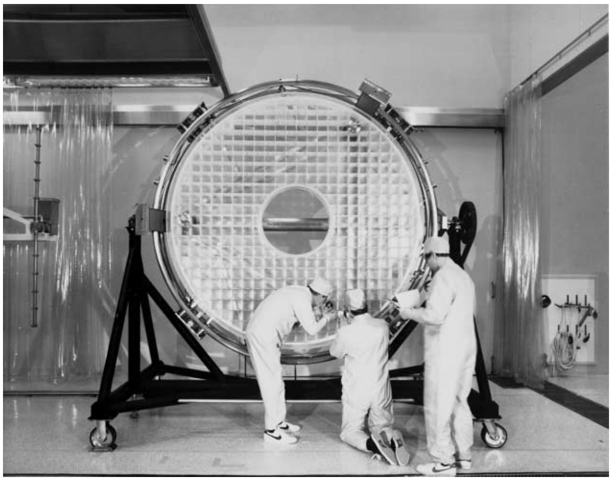

Figure 2 2.4 m ULE™ mirror made by Corning/Perkin Elmer for the Hubble Space Telescope. Photography courtesy of NASA.

Material removal

The evolution in blank manufacturing required a corresponding evolution in the process of material removal. Many new grinding and polishing techniques were developed over the years to more accurately and rapidly generate precision optical surfaces. These included adaptive force grinding, electrolytic in-process dressing (ELID), water jet figuring, computer controlled polishing, stress lap polishing, magneto-rhelogical figuring (MRF), plasma-assisted chemical etching (PACE), ion figuring, and vacuum lap figuring. Most of these processes induce localized forces on the face-sheet in the process of material removal, and for a 'honeycomb' mirror, the face-plate deflects away from the grinding/polishing tool as the tool reaches the center of the cell. This means that less material will be removed in the center of the cell than at the edge of the cell where the core wall holds the face-sheet against the tool. This creates a 'quilted' pattern in the surface that must be removed in subsequent steps. As face-sheets become thinner, quilting becomes more pronounced and the process of removing 'quilting' becomes exacerbated. For a number of years, this was the limiting factor in the development of high-precision lightweight mirrors.

Ion figuring is a noncontact process that has eliminated the problems associated with mechanical material removal. It bombards the surface of the optic with neutralized ions causing atoms to be dislodged. The process creates no deflection of the surface and therefore can remove quilting and other irregularities without generating further distortion in the surface. This has permitted 2–3 mm thick face-plates to be manufactured without concern for quilting effects. Primary limitations now being encountered are in face-plate generation and the various associated handling processes. Great care must be taken to avoid fracture. Spin casting ultra-thin face-plates is being investigated as a means of further reducing problems in the generation process. The 0.75 m mirror made for IKONOS, the high-resolution imaging satellite, represents the state-of-the-art in all-glass lightweight mirrors (28 kg m^{-2}).

Replication

Replication techniques are quite common in low-cost production mirrors and lenses for commercial purposes. However, there are also a number of techniques that hold great potential for precision mirrors as well. Chemical vapor deposition (CVD), plasma vapor deposition (PVD), plasma spray deposition, and electro-forming are all being investigated for producing lightweight precision mirrors. This class of mirror generation does not start with a mirror blank *per se*, rather it replicates a 'mandrel' that has the appropriate surface finish and the inverse of the shape desired in the final mirror. The mandrel can be made very rigid and therefore unperturbed by either gravitational forces or the localized forces generated by grinding and polishing. Extremely thin mirrors with areal densities on the order of 0.1 kg m^{-2} are currently being manufactured using replication techniques. These 'membrane' mirrors are currently being used for solar energy collection; however, recent advances in membrane quality and wavefront correction techniques hold out the promise of producing high-resolution, membrane imaging systems.

The Future?

Theoretically, single-molecule thick mirrors could be manufactured in space using self-replicating nanotechnology systems that 'grow' mirrors according to a mathematical description of the desired optical surface. These 'smart' mirrors could self-correct in the presence of physical disturbances and self-heal when damaged. They could be 'grown' to hitherto unimagined size, being limited only by the availability of raw material. Such mirrors, however, are definitely of the future and should merit re-examination 20 years hence.

Materials

A variety of materials have been developed over the years that have led to lighter and lighter mirrors. Recently, metals are coming back into play as means of fabricating stiff, lightweight mirrors. These metals include beryllium, nickel, nickel alloys, aluminum, and aluminum alloys. Composite materials are also playing a large role in lightweight mirror fabrication, both as a support structure and also as a reflecting surface (with appropriate coatings). Hybrid mirrors comprised of a combination of glass and composites have also been developed, as well as combinations of glass and beryllium. Various forms of silicon carbide have been extensively investigated and lightweight mirrors of pure silicon are being manufactured. Glass still plays a significant role in large lightweight mirrors, and various plastics are under investigation for the production of ultra-thin membrane mirrors.

Glass, Glass Ceramics, and Hybrids

The most common glass/glass ceramic materials used in large telescope mirrors are fused silica, ULE™ (Ultra-Low Expansion) glass, and Zerodur™, a glass ceramic. As was mentioned earlier, there are a variety

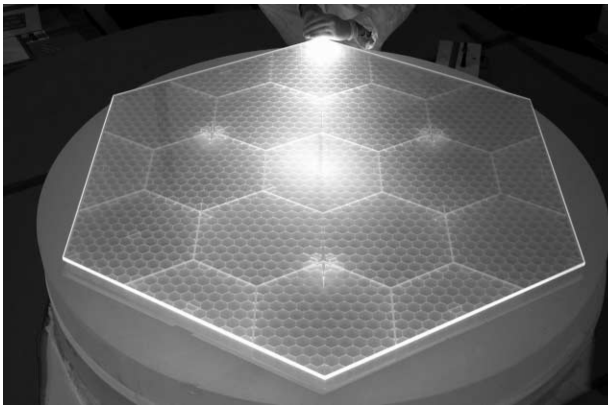

Figure 3 1.5 m ULE™ mirror made by Corning/Kodak for NASA's Next Generation Space Telescope Advanced Mirror System Demonstrator technology program. Photograph courtesy of NASA.

of methods available for the manufacture of glass mirrors depending upon the extent of lightweighting desired. Most of the current effort in lightweight mirrors is centered around hybrid mirrors. One method bonds very thin (2–3 mm) glass face-sheets to composite cores to form 'sandwich' mirrors. Another method attaches face-sheets to composite back-plane structures via actuators. In the latter case, the actuators can be used to actively control mirrors to correct for distortions. These mirrors were developed as part of the James Webb Space Telescope technology program, and typically have areal densities on the order of 15 kg m^{-2} (including back-plane structures). An all glass face-sheet and core is shown in Figure 3.

Silicon Carbide

Silicon carbide has long been a promising material for lightweight mirrors. It has many of the desirable properties of beryllium and is nontoxic. Until recently difficulties were encountered in producing mirrors larger than ~0.5 m in diameter. Of late, several manufacturers have produced large (<1 m) lightweight mirrors using different processes; sintering, liquid silicon infiltration and chemical vapor deposition.

The sintering process consists of preparing a finely ground silicon carbide powder mixed with organic binders and additives and then compacting the powder by isostatically pressing it at high pressure. The resulting compact is then machined to the desired shape and weight and the blank heated to cause agglomeration of the particles. The resulting material is known as reaction-bonded silicon carbide; it can be polished to around 30 angstroms and can be ion figured. If smoother surfaces are required, then a layer of CVD or PVD silicon carbide can be added. A 1.35 m diameter mirror has been fabricated as a demonstration for the European Space Agency's Herschel Space Observatory. The mirror was fabricated in nine segments that were then brazed together to form a unified 26 kg m^{-2} blank. This mirror was not designed to test the limits of lightweight and, consequently, there is no reason to believe that this process could not achieve an areal density comparable to the 7.8 kg m^{-2} of Cesic™ described in the following paragraph. The size of mirrors produced using this technique is currently limited to around 4 m due to the availability of the furnaces required for segment brazing.

The liquid infiltration process is used to create a biphase carbon, silicon carbide material known as Cesic™. This material is manufactured by creating a block of chopped carbon fibers embedded in a phenolic resin. The block is machined to the desired shape and then put into a furnace where the temperature is elevated until the resin is turned to carbon. The temperature is further elevated and liquid silicon infiltrated to produce silicon carbide. Bare Cesic™ can be polished to approximately 100 angstroms. In a subsequent step, a silicon carbide slurry can be applied to the blank that can be polished to around 30 angstroms. It is also possible to add a CVD or PVD silicon carbide layer to the blank that can be polished to obtain angstrom-level surfaces. A 1.04 m diameter Cesic™ mirror has been constructed for the NASA program, Solar Lite. The mirror weighs approximately 60 kg for an areal density of 56 kg m^{-2}. Again, this mirror was not designed to test the limits of lightweighting, but in a separate demonstration, an areal density of 7.8 kg m^{-2} was achieved. The size of mirrors produced using this technique is also currently limited to around 4 m due to the availability of existing furnaces required for infiltration.

The CVD process produces a SiC material that is highly polishable, yielding surface finishes on the order of angstroms. It has been used quite successfully as a coating layer, and has the potential to provide a means of replicating very lightweight mirrors by depositing a thin layer of SiC onto a mandrel. In one approach, the mandrel has the appropriate shape, but not the surface finish. In this case, once the SiC has been deposited, the mandrel is burned away and a SiC mirror blank is left. The blank is then ground and polished to produce the desired figure and finish. If the blank is too thin, this can present a problem in achieving a high-performance mirror. A second approach involves the deposition of SiC onto a figured and polished mandrel. In principle, both the figure and surface finish can be replicated. Success in this approach, however, has been somewhat limited, and although 20-angstrom surface finishes have been replicated, they have not been repeatable. A 1 m diameter blank was manufactured using the first approach, but it too has not been repeated and the lack of success appears to be due to high stress developed in the deposition process. The residual stress in the blank causes large-scale deformation and frequently fracture.

Beryllium

The extremely high stiffness of beryllium makes it a highly desirable candidate for lightweight mirrors. Unfortunately, it is a highly toxic material and there are only a very few companies involved in the manufacture of beryllium mirrors and even fewer that produce the raw material. Blanks are manufactured by isostatically compressing beryllium powder at high temperature and pressure. The blanks are then machined and subsequently ground and polished. Early beryllium mirrors often exhibited interesting, mysterious changes in figure after seemingly benign handling events. Unpredictable changes in figure also occurred as they were cycled between ambient and cryogenic temperatures. These problems appear to have been resolved with the production of spherical powder. The older powder consisted of irregularly shaped particles and when the powder was compressed, small, irregularly shaped voids were created throughout the material. It is hypothesized that these voids changed in size and shape as the blank was stressed, thereby causing unpredictable dimensional changes in the blank. The spherical powder permits uniform, high-density packing yielding homogeneous blanks that do not appear to exhibit unpredictable dimensional changes.

The lightest-weight beryllium mirror to date ($11.8\ \mathrm{kg\ m^{-2}}$) was made for NASA's James Webb Space Telescope technology program. This technology was selected for the flight mirror. The weight reduction was achieved by machining the blank to create an open back honeycomb structure with an integral face-sheet. The ability to fabricate large beryllium mirrors is limited by the availability of facilities capable of isostatically pressing the material at the requisite pressures and temperatures. Currently, blank sizes are limited by available isostatic presses to around 1.2 m in width and 1.6 m in length. The expense of construction of larger facilities combined with limited need make it highly unlikely that any larger mirrors will be manufactured in the near term. A brazing process, however, has been demonstrated for joining segments to form a large blank. This process could, in principle, permit the fabrication of mirrors several meters in diameter. Bare beryllium can be polished to around 30 angstroms surface finishes. If smoother surfaces are required, beryllium can be vacuum deposited on the blank. Nickel may also be deposited on the blank; this has the added benefit that it can be polished without the special facilities required for the hazardous beryllium polishing operations. The drawback of using nickel is that of the bimetallic effect if the mirror is to be operated over a wide temperature range. The Vavilov Institute in Russia has developed an alternative approach to obtaining good surfaces by creating a glass that matches the coefficient of thermal expansion (CTE) of beryllium over a 50° temperature range. The glass is then fused to the beryllium blank. There have been numerous attempts over the years to produce replicated mirrors using a CVD process to deposit beryllium on a mold. So far these attempts have been unsuccessful, the process producing a material structure that no longer has high stiffness.

Silicon

Silicon as a mirror material has the advantage of being a very homogeneous material that can be polished to very smooth surfaces (angstrom level) very rapidly. A lightweight mirror has been fabricated using a silicon face-sheet brazed to a central core of silicon foam. This mirror 'sandwich' has demonstrated an areal density of $<15\ \mathrm{kg\ m^{-2}}$. Technology has been developed to braze segments together in a manner that has been shown to have no adverse effects on either polishability or cryogenic performance (i.e., no distortion across the braze). Consequently, it should be possible to manufacture lightweight silicon mirrors of 1 m diameter or larger.

Composites

There have been many attempts to manufacture mirrors from composite materials, beginning initially with graphite/epoxy and more recently with graphite/cyanate-ester. These mirror are produced by laying up graphite sheets impregnated with epoxy, or cyanate-ester on a near net shape mold and then curing the material at elevated temperatures. The optical surface is typically created by coating another mold (this time with the desired optical surface) with a passivation layer and then with epoxy and placing the near net shape composite substrate on the epoxy. When the epoxy cures, the completed mirror is separated from the mold and the epoxy surface ion milled to remove residual errors such as fiber print-through. These mirrors can be very lightweight ($\sim5\ \mathrm{kg\ m^{-2}}$), but typically have problems with large-scale distortion due to inhomogeneities due to the lay-up process. The epoxy surface replication can also be a problem if the mirror is subjected to temperature extremes. A 3.5 m diameter composite mirror was made as a demonstration for the FIRST program. It has suitable performance for submillimeter and infrared wavelengths, but not for shorter wavelengths. At this time, the accuracy required for large (>1 m) visible imaging systems has not yet been achieved.

Gossamer Mirrors

The term ‘gossamer’ refers to that category of mirrors that in and of themselves have no structural rigidity beyond spatial scales of fractions of millimeters. Several materials are currently being used for these types of materials including Kapton™ and various other polyimide films. The initial shape of the membrane is formed by a variety of techniques including vacuum forming, pressure forming, spin casting, hot forming, and others. Early mirrors were manufactured by joining together triangular-shaped segments (gores) to form the desired shape. When deployed, these mirrors require some means of applying pressure to achieve the desired shape. Some methods have involved applying force at the edge to put the mirror in tension; some involve using a clear membrane as a face-sheet and then pressurizing the combination to create a ‘lens’ with a reflecting inner surface. Most recently, techniques have been demonstrated involving various forms of electrostatic deflection to achieve the desired shape after deployment. At present the primary applications are in the area of solar collection, power beaming, and thermal propulsion. The mirrors are approaching the quality required for imaging telescopes, particularly for small sizes. Techniques for correcting these mirrors with ‘downstream’ active optics have reached the point where it is not unrealistic to imagine that membrane optics could soon reach the accuracy required for imaging. There have been two space-based demonstrations, one by the Russians – a 20 m reflector – and the other by the USA – a 14 m inflatable antenna (Figure 4).

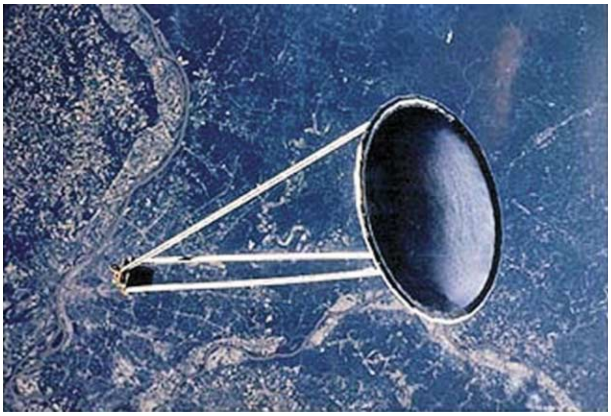

Figure 4 15 m antenna made by L’Garde for NASA’s Inflatable Antenna Experiment. Photograph courtesy of NASA.

Grazing Incidence Mirrors

Broadband X-ray telescopes make use of the reflection property of materials at angles of grazing incidence. The most common design is based on the X-ray microscope objective designs of Wolter in the early part of the twentieth century. These mirrors are hyperboloids and paraboloids of revolution. This configuration provides a structural rigidity not present in normal incidence mirrors; consequently, X-ray mirrors can be made much lighter. The largest mirror of the Chandra X-ray Observatory is made from Zerodur™ (Figure 5). It is a 1.2 m diameter, 0.9 m long cylindrical mirror with a wall thickness of only 24 mm. Such a thickness in a normal incidence mirror of this size could be easily deformed.

Lightweight approaches in grazing incidence optics are almost all concentrated in various forms of replication. Recent developments in electroformed nickel alloy mirror technology have resulted in a 0.5 m diameter 0.6 m long mirror that weighs only 1.2 kg (Figure 6). A similar-sized mirror in the Chandra telescope weighs over 100 kg. Extremely lightweight cylindrical mirrors have been made from graphite/epoxy and from CVD SiC. Various investigations are underway in producing segments of cylinders that are then mounted to form a complete

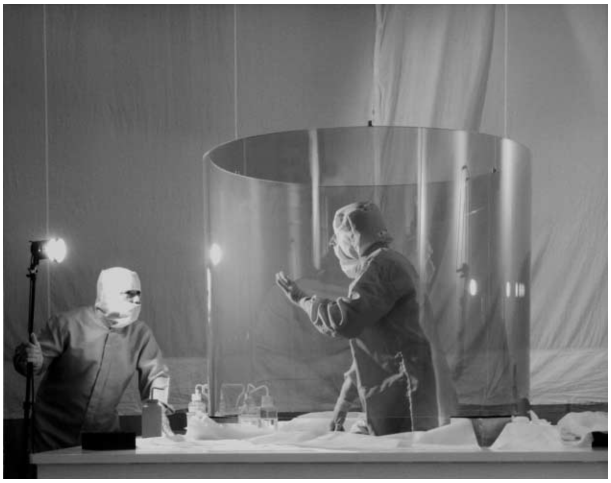

Figure 5 1.2 m Zerodur™ made by Schott Glass Werken/Hughes Danbury Optical Systems for NASA’s Chandra X-ray Observatory. Photograph courtesy of NASA.

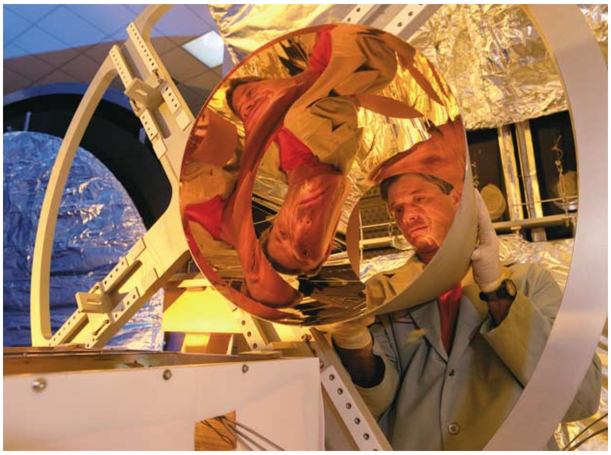

Figure 6 0.5 m electro-formed nickel–cobalt technology mirror made by NASA's George C. Marshall Space Flight Center for NASA's Constellation X-ray Mission technology program. Photography courtesy of NASA.

cylindrical mirror. These techniques involve a variety of materials including aluminum foil, and beryllium plates. In segment replication, the mandrel is passivated and then coated with the desired reflecting material of the final optic (the passivation process prevents the coating from adhering too strongly to the mandrel.) After the mandrel has been coated with the reflecting material, an epoxy coating is applied and a preformed shell segment is placed on the epoxy. After the epoxy cures, the segment can be removed from the mandrel along with the desired reflective surface (typically gold). In this case, the preformed shell does not have to have as precise a figure, or surface as the mandrel, since the epoxy will accommodate minor differences.

See also

Geometrical Optics: Lenses and Mirrors. **Imaging:** Adaptive Optics; Wavefront Sensors and Control (Imaging through Turbulence). **Instrumentation:** Astronomical Instrumentation; Telescopes. **Optical Materials**: Smart Optical Materials.

Further Reading

Bilbro J (2001) Optics in orbit. *SPIE OE Magazine* (Aug. issue).

Breckinridge JB and Jakobsen P (eds) (2000) UV, Optical, and IR Space Telescopes and Instruments. *Proceedings of SPIE* 4013.

Chen PC, Bowers CW, Content DA, *et al.* (2000) Advances in very lightweight composite mirror technology. *Optical Engineering* 39(9).

Ealey MA (ed.) (1995) Silicon carbide materials for optics and precision structures. *Proceedings of SPIE* 2543.

Engelhaupt DE, Ramsey BD, O'Dell SL, *et al.* (2000) New alloys for electroformed replicated X-ray optics. *Proceedings of SPIE* 4138: 22.

Goodman WA and Jacoby MT (2000) Dimensionally stable ultra-lightweight silicon optics for both cryogenic and high-energy laser applications. *Proceedings of SPIE* 4198: 260–270.

Jenkins C (vol. ed.) (2001) Gossamer spacecraft: membrane and inflatable structures technology for space applications. In: Zarchan P (series ed.) *Progress in Astronautics and Aeronautics* 191. Reston, VA: American Institute of Aetronautics and Astronautics.

Kendrick SE, Reed T and Streetman S (2001) In-process status of the 1.4-m beryllium semi-rigid Advanced Mirror System Demonstrator (AMSD). In: Stahl HP (ed.) *Optical Manufacturing and Testing,* vol. IV. *Proceeding of SPIE* 4451: 58–66.

Mathews G (2001) Semi-rigid active mirror technology advancements. *Proceedings of SPIE* 4451.

Müller C, Papenburg U, Goodman WA and Jacoby M (2000) C/SiC high precision lightweight components for optomechanical applications. *Proceedings of SPIE* 4198: 51–57.

Truemper JE and Aschenbach B (eds) (2000) X-Ray optics, instruments, and missions III. *Proceedings of SPIE* 4012.

Measurement of Optical Properties of Solids

P Lucas, University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The interaction of light with solids takes place through different mechanisms, depending on the type of material and the range of wavelength investigated. Insulators or dielectrics are typically transparent to visible light while most semiconductors are opaque to visible light yet transparent to infrared radiations; in contrast metallic solids appear shiny because they reflect all wavelength up to the ultraviolet region. The optical properties of a solid depends on its chemical composition and its structural properties and vary for every material, though, one parameter, the complex refractive index **n**, is sufficient to characterize entirely the optical properties of a specific material.

In this article we will briefly review the physics underlining the nature of the refractive index **n**, in order to understand its relationship to the observable quantities such as absorption, reflection, and transmission routinely measured in optics experiments. A description of the experimental techniques, as well as a physical interpretation of the data, will also be presented.

Origin of Light Interaction with Matter

Light is described as a transverse electromagnetic wave consisting of an electric and magnetic field oscillating perpendicular to the direction of propagation at optical frequencies within $10^{13} \rightarrow 10^{17}$ Hz. This range of frequencies comprises the ultraviolet, visible, and infrared domain. On the other hand, solid mediums are composed of charged particles, negative electrons, and positive ions producing electric dipoles, which can be polarized under the action of an electric field. Hence, when electromagnetic radiation impinges upon a material it interacts by polarizing the molecular units, producing oscillating dipole moments. This interaction results in several observable optical phenomena such as reflection, transmission, absorption, or scattering, which will be described in more detail in a following section. The classical model of light propagation assumes that the oscillating electric field can interact with several different types of dipole oscillators within the material. Different dipoles are usually accessed by light wave from different frequency range depending on their mass.

Electronic Oscillators

Lorentz originally proposed that the electrons bound to the nucleus of an atom behave as oscillators. In this representation, electrons behave as if they were held by springs and subjected to a linear restoring force, therefore oscillating up and down (Figure 1a). This representation of an atom might seem to contradict the conventional picture of electrons orbiting around the nucleus; however, it is valid as far as light–matter interactions are concerned. The resonant frequency of such an oscillator is proportional to the inverse of its reduced mass μ defined as

$$\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_N} \qquad [1]$$

where m_N and m_e are the mass of the nucleus and electrons, respectively. The resonant frequency ω_0 is then related to μ according to

$$\omega_0 \propto \sqrt{\frac{1}{\mu}} \qquad [2]$$

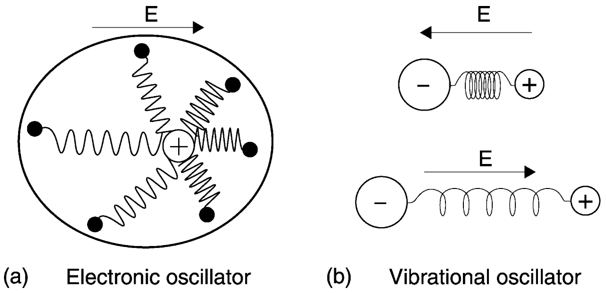

Figure 1 (a) Classical oscillator model of a polyelectronic atom. The electrons depicted as black dots are bound to the positive nucleus by springs, which represent the electrostatic restoring forces between the charged particles. The electronic oscillators are distorted under the effect of an electric field. (b) Classical model of a diatomic vibrational oscillator. The chemical bond is represented as a spring and the charged ions vibrate around their equilibrium position under the effect of an oscillating electric field.

Considering that $m_N \ll m_e$ we can assume that $\mu = m_e$ and show that the smaller mass of the electron determines the resonant frequency of the oscillator. Hence, resonant electronic oscillations are accessed at higher frequencies in the ultraviolet and visible region of the spectrum. In solids, the quantum theory shows that these resonances correspond to excitation of electrons across the gap between a full valence band and an empty conduction band.

Vibrational Oscillators

The second type of dipoles contained in the medium are based on pairs of charged atoms vibrating around their equilibrium position in the solid. Every atom has a different electronegativity, hence the electronic density on atomic pairs is distorted and result in a charge imbalance, which can couple to the electric field of light. This generates oscillating dipoles, which involves the motion of atomic nucleus (Figures 1b). The nuclear mass being several orders of magnitude larger than the electronic mass, we can show by analogy with eqns [1] and [2] that the resonant vibration will occur at lower frequencies in the infrared region. In solids, the quantum theory shows that the lattice atoms vibrate cooperatively and generate quantized lattice waves called phonons.

Free Electron Oscillators

In contrast with the bound electrons from the electronic oscillator model, free electrons can move without being subjected to a restoring force. These correspond to the conduction electrons present in metals and to a smaller extent in doped semiconductors. The free electron oscillator model is therefore

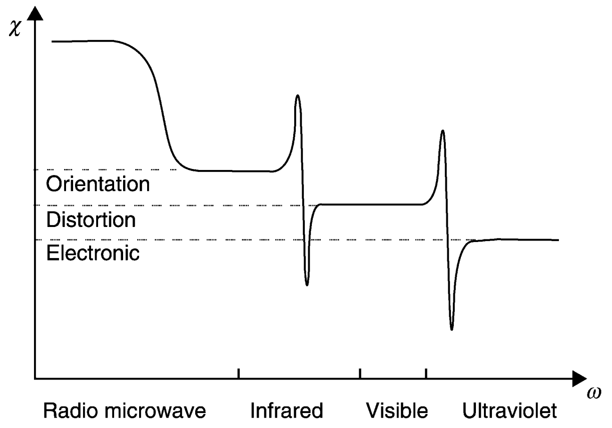

Figure 2 Contributions to the polarizability of a medium as a function of light frequency. At optical frequencies, the molecules cannot re-orient fast enough to follow the reversing electric field and the contribution of dipole orientation is lost.

principally used to describe the optical properties of metals such as reflectivity. The free electrons are well described as conventional oscillator with a natural resonant frequency $\omega_0 = 0$.

The relationship between the applied electric field E and the resulting polarization of the medium P is expressed through the dielectric susceptibility χ as

$$P = \varepsilon_0 \chi E \qquad [3]$$

where ε_0 is the permittivity of vacuum. We can generally recognize three contributions to the dielectric susceptibility or the polarizability of an optical material (Figure 2). Applied fields in the radio or microwave region oscillate slowly enough that polar molecular units can reorient and align in the direction of the field. The characteristic time for orientational motion is fairly long and this contribution is quickly lost when the molecules cannot follow the field oscillating at higher frequency. When the field reaches optical frequency, the two remaining contributions are the vibrational and electronic dipoles mentioned above. The vibrational dipoles involve a distortion of the chemical bond and are the slowest process occurring only in the infrared region. This contribution is lost at visible frequencies and leaves only the fast electronic polarization processes.

If we treat it rigorously, the polarization P is actually expressed as a function of higher order terms according to

$$P = \varepsilon_0\left(\chi E + \chi^{(2)} EE + \chi^{(3)} EEE + \ldots\right) \qquad [4]$$

This gives rise to nonlinear effects such as self-focusing and second or third harmonic generation. However, the higher order susceptibilities $\chi^{(n)}$ are very small and these effects are only significant for very high intensity light source such as lasers. In the following text we will only consider linear optical processes.

The Dipole Oscillator Model (Lorentz Oscillator)

In the Lorentz model, the bound electrons behave as harmonic oscillators and are constrained to vibrate at their natural resonant frequency ω_0. These oscillators are also submitted to a damping force corresponding to a resistance to the motion due to collisional processes. The equation of motion is then

$$m\left(\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} + \omega_0^2 x\right) = qE \qquad [5]$$

where x is the displacement along E, γ is the damping factor, m is the mass of the electron and q its charge. The passage of an electromagnetic wave through a medium exerts an oscillating force on the electrons causing them to vibrate up and down. In other words these electronic dipoles are submitted to a forced oscillation induced by an electric field varying periodically with time. The electric field felt by the atomic dipoles is written as

$$E = E_0 e^{i\omega t} \qquad [6]$$

where ω is the frequency of the light.

The electric field of the light wave then drives the dipole oscillations at its own frequency ω and the dipoles displacement oscillates according to

$$x = x_0 e^{i\omega t} \qquad [7]$$

Solving eqn [5] for x in terms of E gives the expression for the displacement of the electrons

$$x = \frac{q/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [8]$$

The atomic dipole moment p induced by the electrons displacement is therefore

$$p = qx = \frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [9]$$

And for a medium with N atoms per unit volume the total polarization P is

$$P = N\frac{q^2/m}{\omega_0^2 - \omega^2 + i\gamma\omega} E \qquad [10]$$

Our result in eqn [10] only contains one natural frequency ω_0 and therefore only accounts for one single type of oscillator in the medium. However, normal optical mediums are composed of different kinds of atoms, each with several characteristic resonant frequencies ω_j. In our model, we can simply treat these multiple resonances by considering that each oscillator acts separately and add the contributions of all oscillators.

$$P = \left(\frac{q^2}{m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \right) E \qquad [11]$$

Comparing eqns [3] and [11] then gives an expression for the polarizability χ of the medium

$$\chi = \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \qquad [12]$$

Additionally, the polarization P is associated with Maxwell's displacement vector, D, through the relationship

$$D = \varepsilon_0 E + P \quad \text{and} \quad D = \varepsilon E \qquad [13]$$

where ε is the permittivity of the medium. The permittivity ε is normally expressed relative to the permittivity of vacuum ε_0 in terms of the dimensionless quantity $\varepsilon_r = \varepsilon/\varepsilon_0$. Where ε_r is the familiar dielectric constant of a material, which is related to the refracted index n as $n = \sqrt{\varepsilon_r}$.

Combining eqns [3] and [13] we obtain an important relationship linking χ to ε_r, and in turn χ to n.

$$\varepsilon_r = 1 + \chi \quad \text{or} \quad n^2 = 1 + \chi \qquad [14]$$

Finally, combining eqns [12] and [14] we obtain the expression for n in terms of the oscillator model.

$$n^2 = 1 + \frac{q^2}{\varepsilon_0 m} \sum_j \frac{N_j}{\omega_j^2 - \omega^2 + i\gamma_j \omega} \qquad [15]$$

The Refractive Index

Two important points should be made concerning the result of eqn [15]. The refractive index is a wavelength-dependent quantity and is a complex quantity. The complex refractive index **n** is indeed usually expressed as

$$\mathbf{n} = n + i\kappa \qquad [16]$$

where n is the real refractive index also defined as the ratio of the wave velocity in vacuum to the velocity in the medium $n = c/v$. And κ is the extinction coefficient, which is directly related to the absorption coefficient α as we shall see in the next section.

As we mentioned in the introductory section, the complex refractive index **n** is sufficient to characterize the optical properties of the solid. The real part of the index describes the change in velocity or wavelength of a wave propagating from a vacuum into a medium while the imaginary part is a measure of the dissipation rate of the wave in the medium.

By inspection of eqn [15] we can predict the form of the variation of n with wavelength. At frequencies ω below the resonant frequency ω_j, the terms ω^2 and $i\gamma_j\omega$ are much smaller than ω_j^2 and to a first approximation n is constant. When ω reaches a resonant frequency ω_j, the term $(\omega_j^2 - \omega^2)$ goes to zero and since the term $i\gamma_j\omega$ is very small, the value of n increases very sharply and exhibits a resonance line as depicted in Figure 3.

On closer inspection of eqn [15] however, we can see that n is not exactly constant in between resonance lines. As ω rises, the denominator slightly decreases and therefore n rises slowly with frequency. The refractive index is then slightly larger for wave of higher frequency. This is why a glass prism bends blue light more strongly than red light. This phenomenon is called dispersion as waves of different frequencies are 'dispersed' by the prism.

Another important point raised by eqn [15] is that because $(\omega_j^2 - \omega^2)$ goes to zero at a resonant frequency, the damping term $i\gamma_j\omega$ dominates and n becomes almost entirely imaginary. **n** is then mostly governed by the extinction coefficient κ. The damping associated with the coefficient γ_j corresponds to a friction force causing a loss of energy, which becomes the dominant effect during resonance. The resonance region is then associated with strong attenuation or

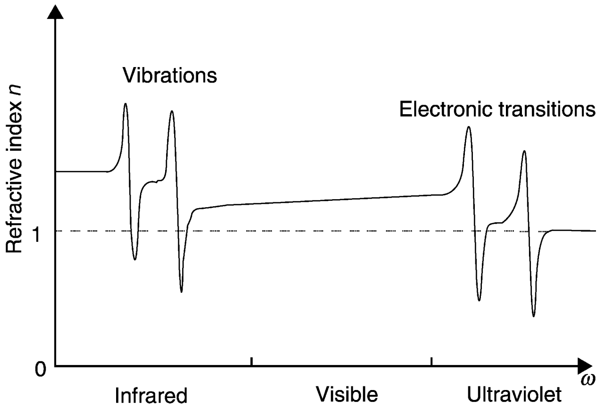

Figure 3 Variation of the refractive index of a hypothetical dielectric solid in the optical frequency range. At high frequency, ω is larger than all resonant frequencies and the refractive index n reaches unity.

absorption of the wave. In contrast, between resonant frequencies, the refractive index is almost entirely real as the imaginary part $i\gamma_j\omega$ is negligible in comparison with $(\omega_j^2 - \omega^2)$. Absorption is then very small and that region corresponds to the transparency domain of the material. In fact, we can assume that $\mathbf{n} = n$ in the transparent region, that is why tables of optical constants of transparent optical materials only list the real part of the refractive index n. However, at resonant frequencies on either side of the transparency domain, absorption becomes significant and the imaginary part of the refractive index must be taken into account so that $\mathbf{n} = n + i\kappa$.

The refractive index $\mathbf{n}$ being a complex quantity it ensues that ε_r must also be complex, since $n = \sqrt{\varepsilon_r}$. Hence we also define the complex dielectric constant as

$$\varepsilon_r = \varepsilon_1 + i\varepsilon_2 \qquad [17]$$

Both parameters describe equivalently the optical properties of a solid so that $\mathbf{n}$ or ε_r are called the optical constants of the material. Note that it is an unfortunate misnomer since they vary distinctly with wavelength (Figure 3).

Using equation [17] and $n = \sqrt{\varepsilon_r}$, we can establish the relationships between the real and imaginary parts of $\mathbf{n}$ and ε_r.

$$\varepsilon_1 = n^2 - \kappa^2 \qquad [18]$$

$$\varepsilon_2 = 2n\kappa \qquad [19]$$

We can see that $\mathbf{n}$ and ε_r are not independent variables but that the real and imaginary part of ε_r can be calculated, knowing the real and imaginary part of $\mathbf{n}$ and vice versa.

Moreover, it can be shown that the real and imaginary parts of each individual parameter are not independent either but can be computed from one another with use of the following Kramers–Kronig relations.

$$\varepsilon_1(\omega) = 1 + \frac{2}{\pi}\int_0^\infty \frac{\omega'\varepsilon_2(\omega')}{\omega'2 + \omega^2}\mathrm{d}\omega' \qquad [20]$$

$$\varepsilon_2(\omega) = -\frac{2\omega}{\pi}\int_0^\infty \frac{\varepsilon_1(\omega')}{\omega'2 + \omega^2}\mathrm{d}\omega' \qquad [21]$$

In these equations, only the principal value of the integral is calculated so that ε_1 can be computed if ε_2 is known over all frequencies and vice versa. The same equation connect n and κ together.

This analysis shows that it is only necessary to know one real or one imaginary part over a wide frequency range to determine all others. The interaction of light with matter is in fact characterized by a single independent real quantity.

Absorption

All electromagnetic phenomena are governed by Maxwell's equations. The form of an electromagnetic wave propagating through a medium is obtained by solving Maxwell's equation. For a wave propagating along z through a homogeneous, optically isotropic medium the form of the oscillating electric field is given by:

$$E = E_0 e^{i(kz-\omega t)} \qquad [22]$$

where k is the wave vector and ω the angular frequency of the light. For a wave traveling into a transparent medium of refractive index n, the relationship between k and ω is

$$k = \frac{n\omega}{c} \qquad [23]$$

However, for an ordinary optical material there are regions of absorption and the refractive index is therefore complex. We should then rewrite eqn [23] as

$$k = (n + i\kappa)\frac{\omega}{c} \qquad [24]$$

By substituting this value of k in eqn [22] we obtain an expression with two exponential terms

$$E = E_0 e^{-\frac{\omega\kappa z}{c}} e^{i\left(\frac{\omega n z}{c} - \omega t\right)} \qquad [25]$$

The term $e^{i\left(\frac{\omega n z}{c} - \omega t\right)}$ represents a wave traveling at a speed $v = c/n$, while the term $E_0 e^{-(\omega\kappa z/c)}$ represent the amplitude of this wave which decays exponentially with distance z. The intensity I of the wave is proportional to the square of the amplitude so that

$$I \propto e^{-\frac{2\omega\kappa z}{c}} \quad \text{or} \quad I \propto e^{-\alpha z} \qquad [26]$$

where α is the familiar absorption coefficient routinely measured by absorption spectroscopy. Hence we obtain an important relationship relating the observable quantity α to the imaginary part of the refractive index κ according to

$$\alpha = \frac{2\kappa\omega}{c} \qquad [27]$$

Local Field Correction

It should be pointed out that the derivation of our refractive index expression eqn [15] assumes that only the electric field of the light affects the polarization of the dipole oscillators. However, this assumption is not correct in dense materials because of the contribution from the electric field of neighboring atomic dipoles. The field generated by each dipole will affect all other dipoles in the vicinity. It is possible to approximate this contribution by accounting for the effect of dipoles within a local sphere. The result of this approximation is known as the Clausius–Mossotti equation, which is accurate for isotropic medium such as glasses and cubic crystals

$$\frac{n^2 - 1}{n^2 + 2} = \frac{\chi}{3} \qquad [28]$$

Measurable Optical Parameters

Four things can happen to a beam of light propagating through a thick slab of optical material (Figure 4). Some of the light can be reflected at the surface of the solid, some can be absorbed by the sample, some can be scattered in different directions, and some of the light can be transmitted through the sample. The effect on the light beam, resulting from these phenomena, can be quantified by a number of optical coefficients, which characterize the macroscopic properties of the material.

Reflection of light at the surface of a solid is described by the reflectance R defined as the ratio of the reflected intensity I to the incident intensity I_0. For a beam falling perpendicularly on a flat surface, the reflection is called specular and is governed by the complex refractive index according to the Fresnel equation

$$R = \frac{I}{I_0} = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2} \qquad [29]$$

This provides us with a second important formula relating a measurable quantity (R) to the optical constant of the material. For measurements performed within the transparency region of a solid, the value of κ is much less than n and eqn [29] reduces to the more familiar form

$$R = \frac{(n-1)^2}{(n+1)^2} \qquad [30]$$

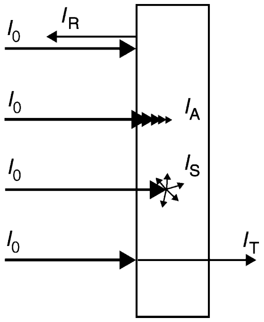

Figure 4 Representation of the four main optical processes happening to a beam as it propagates through a slab of transparent material.

Absorption occurs when the light frequency reaches the natural resonance of some dipole oscillators in the medium. The energy of the electromagnetic wave is partly transferred to the material and generally dissipated in the form of heat. The light intensity is therefore attenuated as it propagates through the material and the attenuation efficiency is quantified by the absorption coefficient α. The intensity decreases exponentially with pathlength z according to Beer's law

$$I(z) = I_0 e^{-\alpha z} \qquad [31]$$

where I_0 is the incident intensity. The imaginary part κ of the optical constant can then be directly obtained from measurements of α using eqn [27].

Scattering is the process whereby light is redirected in different directions due to microscopic density fluctuation in the sample produced by defects, impurities, or structural inhomogeneities. The intensity of light propagating in the forward direction is attenuated by the scattering event and it can be quantified in a way equivalent to absorption. The intensity has an exponential dependence on pathlength z analogous to Beer's law

$$I(z) = I_0 e^{-Sz} \qquad [32]$$

where S is the scattering coefficient. When the scattering center is smaller than the wavelength of light this phenomenon is called Rayleigh scattering and the scattering coefficient S vary with the inverse fourth power of the wavelength

$$S(\lambda) \propto \frac{1}{\lambda^4} \qquad [33]$$

By measuring light attenuation we cannot tell the difference between absorption and scattering and the total attenuation is $\alpha_{\text{Tot}} = \alpha + S$. However, the scattering contribution is generally much weaker than the absorption and can be neglected so that $\alpha_{\text{Tot}} = \alpha$.

Transmission occurs when the light is neither reflected, absorbed, or scattered. The beam is then transmitted through the sample and the fraction of light exiting the back surface is quantified as the transmittance T. If we disregard scattering in comparison to absorption then conservation of energy require that

$$R + T + A = 1 \qquad [34]$$

where A is the fraction of light absorbed. The transmittance can then be compiled from R and α for an incident beam I_0 traveling across a sample of thickness l. In this case we must consider the reflection of the light on the front surface as well as on the back surface when it exits the transparent medium. The expression for the transmittance accounting for dual reflection is

$$T = (1 - R)^2 e^{-\alpha l} \qquad [35]$$

It should be noted that a rigorous treatment of eqn [35] should account for multiple reflections between the front and back surface in the interior of the sample. In this case, interference effects complicate the determination of **n** and a modified version of eqn [35] can be computed by summing up the intensities due to the contribution of the multiple reflections.

Transmission and reflection measurement are readily obtained using conventional spectrometers. Two types of spectrometers must be used to cover the entire optical spectral region. FTIR (Fourier Transform Infra Red) spectrometers use a glow bar as the light source, which enable it to cover the infrared and near infrared domain. The UV-VIS (ultraviolet visible) double beam spectrometers use a tungsten and deuterium lamp and cover the spectrum from the ultraviolet down to the near infrared. Using eqns [27], [29] and [35], the reflection and transmission measurements thus obtained are used to calculate the real and imaginary component of the optical constant over the entire optical frequency range.

An example of transmission measurement is shown in Figure 5, for a dielectric and a semiconductor. The two spectra have the same principal features. At short wavelength the edge of the transparency window is due to absorption of energy by the valence electrons. In classical terms this corresponds to the resonance of electronic oscillators, and in quantum mechanical terms to the excitation of electrons from the valence to the conduction band. This edge is due to a sharp increase of κ, which extends over a significant range of frequency corresponding to the domain of opacity or high absorption. The wavelength at the edge determines the minimum energy of the photon necessary to promote an electron across the bandgap E_g. Insulators have a large bandgap and appear transparent to the human eye because visible light is not energetic enough to promote an electron across E_g. In contrast, semiconductors have a smaller bandgap and appear black (opaque) because all the visible light is absorbed to induce electronic transitions. On the low frequency side the transparency window is limited by atomic vibrations in the solid structure. The wavelength of the vibrations increases with the mass of the atomic oscillators. Semiconductor compounds have higher atomic masses and therefore transmit at longer wavelength in the infrared while insulator have lower atomic mass and exhibit their vibrational absorption edge at shorter wavelength.

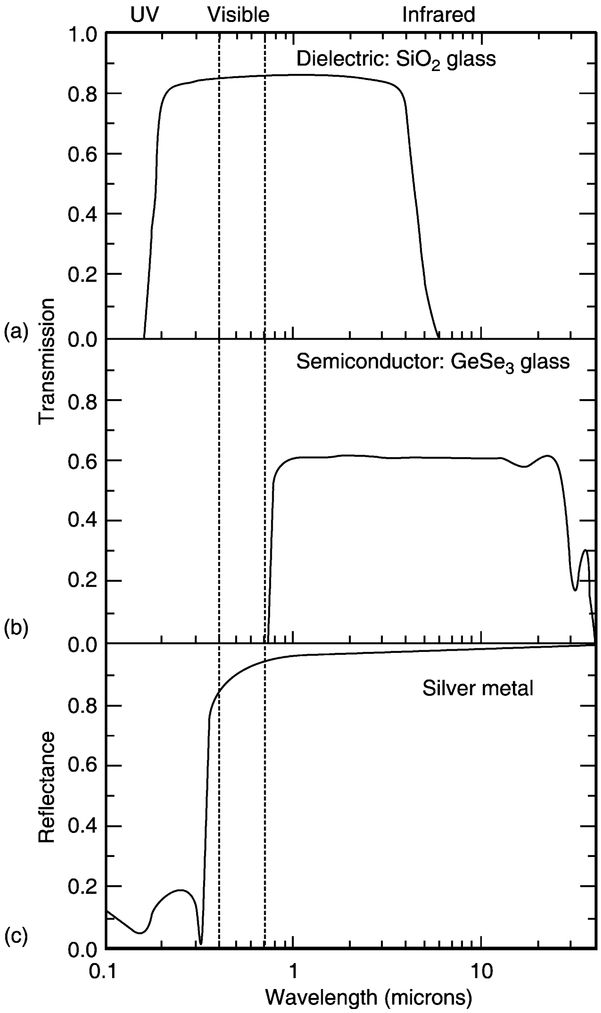

Figure 5 (a) Transmission spectrum of SiO_2 glass, a standard dielectric solid. (b) Transmission spectrum of $GeSe_3$ glass, a semiconductor ($E_g = 1.6$ eV). (c) Reflectance spectrum of silver metal.

In between the two absorption regions, κ is very small and the material is transparent. The refractive index is almost entirely governed by the real part n. Eqn [15] shows that n is proportional to the number N of electronic oscillators. The refractive index will then be higher for heavy atoms with a high electronic

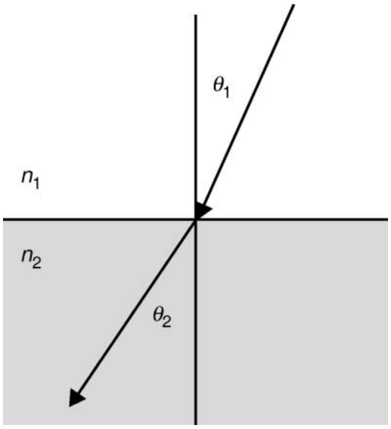

Figure 6 Refraction of a beam propagating across the interface between two mediums of different refractive index. The incident angle and the refraction angle are related to the index of refraction through Snell's Law.

density. Consequently, semiconductors exhibit a maximum transmission of only 60% because of significant surface reflectivity (eqn [35]) resulting from their higher refractive index n (eqn [30]).

Most applications of materials in optics are only concerned with the transparency domain so that the useful optical constant is reduced to the real part n. Hence a number of experimental techniques have been developed to obtain n, often by measuring angles of refraction with the use of Snell's law

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad [36]$$

where θ_1 is the angle of the beam incident in the medium of index n_1, and θ_2 the angle of the beam refracted at the interface with a medium of index n_2 (Figure 6). The basic principle is to compare the sample with a standard glass of known refractive index and measure the refraction angle at their interface. The sample's index can then be obtained following eqn [36]. Since n varies with wavelength, these techniques are normally performed with monochromatic light and provide n at a single wavelength. Among these methods are: the *Abbe refractometer*, using a glass hemisphere as a standard, the *V-block refractometer*, using a V shaped prism as a standard, and the *prism goniometer* based on the relative deviation between a known glass prism and a prism shaped sample.

Ellipsometry is another technique widely used to measure the index n. This technique is normally used on thin films deposited on substrate, as it allows to simultaneously measure the thickness and the refractive index. However, the method can also be applied to bulk samples. The principle of ellipsometry is based on measuring the change in polarization of a beam reflected off the sample surface as a function of incidence angle. The reflected light is elliptically polarized to an extent depending on n. This technique is especially useful to measure n at frequency range above the absorption edge where the sample is highly absorbing.

Metals

The optical properties of metals are mainly characterized by their very high reflectivity, which causes their shiny appearance. All metals reflect light in the infrared and visible region up to a cutoff frequency in the ultraviolet. This critical frequency is called the plasma frequency ω_p. The physical significance of the plasma frequency can be understood using the Drude–Lorentz model of the free electron oscillator. As mentioned previously, the free electron can be described accurately as dipole oscillators with a resonant frequency $\omega_0 = 0$. If we consider a system with negligible damping, we can introduce $\gamma = 0$ and by assuming all free electrons to be equivalent we can consider only one type of oscillators so that eqn [15] reduces to

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2} \qquad [37]$$

where

$$\omega_p^2 = \frac{Nq^2}{\varepsilon_0 m_e} \qquad [38]$$

Equation [37] means that in the low frequency domain $\omega < \omega_p$, the refractive index n must be imaginary ($n^2 < 0$) and in the high frequency domain $\omega > \omega_p$, n is a real positive number. The low frequency region is therefore dominated by the extinction coefficient κ and the reflectance R, given by eqn [29], is essentially unity. In the high frequency limit, R decreases and n is real. These features are shown in Figure 5c. Metals are opaque and highly reflective below the plasma frequency, while they become transparent in the ultraviolet.

List of Units and Nomenclature

A	absorbance
c	light velocity in vacuum
D	electric displacement
E	electric field
I	intensity of light (power per unit surface)
k	wave vector
m_e	mass of electron
m_N	mass of nucleus
$\mathbf{n}$	complex refractive index
n	real refractive index

p	dipole moment
P	polarization
q	charge of electron
R	reflectance
S	scattering coefficient
T	transmittance
v	phase velocity
α	absorption coefficient
ε_0	permittivity of vacuum
ε_r	relative dielectric constant
κ	extinction coefficient
λ	wavelength
μ	reduced mass
χ	dielectric susceptibility
ω	angular frequency
ω_0	resonant frequency
ω_p	plasma frequency

See also

Instrumentation: Ellipsometry; Spectrometers. **Scattering:** Scattering from Surfaces and Thin Films.

Further Reading

Abeles F (1972) *Optical Properties of Solids*. North Holland Publishing Company.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge University Press.

Dragoman D and Dragoman M (2002) *Optical Characterization of Solids*. Springer.

Feynman RP (1970). *Lectures on Physics*, vol. I–III. Addison–Wesley.

Fox M (2001) *Optical Properties of Solids*. Oxford University Press.

Hecht E (1998) Optics, 3rd edn. Addison–Wesley.

Klingshirn CF (1997) *Semiconductor Optics*. Springer.

Nazarewick W, Rolland P, Dasilva E and Balkanski M (1962) ABAC chart for fast calculation of the absorption and reflection coefficients. *Applied Optics* 1(3): 369–370.

Palik ED (1985) *Handbook of Optical Constants of Solids*. Academic Press.

Savage JAR (1990) Signals radar. *Optical Materials* 1:323–419.

Simmons CJ and El-Bayoumi OH (1993) *Experimental Techniques of Glass Science*. The American Ceramic Society.

Simmons JH and Potter KS (2000) *Optical Materials*. Academic Press.

Optical Glasses

J E Shelby, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

While advances continue in this field, some optical glasses have histories which can be traced to the seventeenth century. These glasses are produced by direct melting of raw materials of relatively high purity. Modern, or nontraditional, optical glasses are more commonly used to transmit light outside the visible region and are often made by nontraditional methods, including many variations on chemical vapor deposition, reactive melt processing, and sol–gel techniques. These glasses are used for ultraviolet and infrared lenses and other optical components, for optical fibers, for glass lasers, and for photonic devices such as optical isolators. Purity requirements and production demands for these glasses are significantly greater than those for traditional optical glasses.

Traditional Optical Glasses

Traditional optical glasses are primarily used as components such as lenses, prisms, or windows for optical systems operating in the visible region. These applications require glasses of high homogeneity, with precisely known refractive indices and dispersions. Glasses, as defined in this article, are nonmetallic, inorganic materials which have no long-range atomic structure and which display a temperature region of time-dependent properties.

Crown and Flint Glasses

Traditional optical glasses are designated as either crown or flint glasses. Crown glasses were originally based on the soda-lime-silica ternary system and were developed for use as window glass. The crown process involved spinning a spherical gob of molten glass to form a flattened disk which was much thicker in the center than at the edges. The edges of these disks were used for window panes, while the ‘crown’ in the center was often used as a crude lens. Crown glasses have low refractive indices and low to moderate dispersions.

The first flint glasses were potash-lead-silicate glasses prepared using English ‘flint’ as the source of silica. The presence of lead in these glasses results in higher refractive indices and dispersions than those of crown glasses.

The designations of crown and flint have since been expanded to include glasses from many other compositional systems. Glasses with lower refractive indices and dispersions are designated as crowns, while those with higher refractive indices and dispersions are designated as flints. Modern crown glasses are based on silicate, phosphate, and borosilicate systems. Additions of fluorine are particularly useful for production of low refractive index, low dispersion glasses, and for adjusting these parameters to meet specific needs. The lowest combinations of refractive index and dispersion are found for glasses based on BeF_2, while the highest combinations of refractive index and dispersion are found in rare earth borate systems. Other common optical glasses contain barium or lanthanum, which can be used to produce either crown or flint glasses, depending on the other components of the glass.

The need for glasses with different combinations of refractive index and dispersion can be traced back to the development of the telescope. Use of lenses made of a single glass made it difficult to obtain sharp images. Newton recognized that this problem was due to divergence of light of different wavelengths while passing through a lens. He felt that this problem was inherent to lenses and could not be eliminated. Others eventually recognized that a compound lens containing a positive crown glass and a negative flint glass would result in cancellation of the divergences and produce a sharper image. While this development did not totally eliminate the colored fringes in telescope images due to the presence of the secondary spectrum, it did lead to major improvements in optical imaging. Binary compound lenses are still fashioned from a combination of crown and flint glasses, with much greater flexibility today due to development of glasses with much more varied combinations of refractive index and dispersion.

Notation for Optical Glasses

Since the refractive index and the dispersion are the two most important characteristics of an optical glass for use as a lens, a notation system has been developed to designate any glass by a six digit number. This notation system requires that the dispersion be expressed by the Abbe number, ν_d, which is defined as

$$\nu_d = (n_d - 1)/(n_F - n_c) \qquad [1]$$

where d, F, and c indicate the value of the refractive index measured at the yellow helium line at 587.6 nm, the blue hydrogen line at 486.1 nm, and the red hydrogen line at 656.3 nm, respectively. Since the Abbe number is a reciprocal dispersion, a large Abbe number indicates a low dispersion, while a small Abbe number indicates a larger dispersion. A glass is then designated by the expression

$$1000(n_d - 1)(10\nu_d) \qquad [2]$$

where each term is rounded to 3 significant digits. A common crown glass known as BK7, for example, which has a n_d of 1.516 800 and ν_d of 64.17 is designated as 517 642. A very high index flint glass, with $n_d = 1.807\,410$ and $\nu_d = 31.61$ would be designated as 807 316. This system can be used for all but the exceptional glasses with $n_d \geq 2.00$. Some tabulations are based on the green mercury line, or e line, at 546.1 nm instead of the yellow helium (d) line. The choice of the d or e line results in a small difference in the reported refractive index and Abbe number for any given glass (Figure 1).

Dispersion Formulae for Optical Glasses

The combination of a refractive index near the middle of the visible region of the spectrum and the Abbe number describe the contribution of an optical glass to the refractive power and chromatic aberration of a lens. This information was sufficient for many early applications of lenses and is still sufficient for the design of optical devices of modest expectations. More sophisticated optical devices require more information for proper design. The simple Abbe number is often insufficient for expressing the dispersion, since it is necessary to provide a

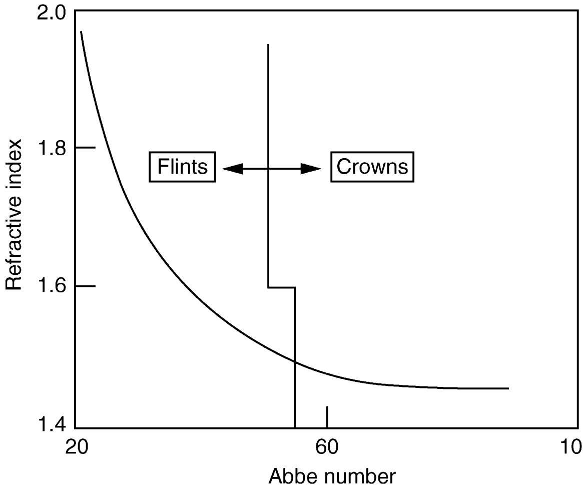

Figure 1 Relation between the refractive index and Abbe number of optical glasses. Commerical optical glasses lie in the regions labeled flint and crown above the curve.

detailed set of values of the refractive index as a function of wavelength across the entire visible region. The refractive index can be listed at a number of wavelengths, or described by an equation covering the entire wavelength range of interest. Although there are a number of expressions for the wavelength dependence of the refractive index, the most commonly used are the Cauchy dispersion formula, or

$$n = a + b\lambda^{-2} + c\lambda^{-4} \qquad [3]$$

where a, b, and c are constants derived from a best fit of the data to this expression, and variations of the Sellmeier dispersion formula, which is often simplified to the expression

$$n^2 = A_0 + A_1\lambda^2 + A_2\lambda^{-2} + A_3\lambda^{-4} + A_4\lambda^{-6} + A_5\lambda^{-8} \qquad [4]$$

where the six A_n constants are tabulated. The Cauchy dispersion formula is accurate to $\approx 10^{-4}$, while the modified Sellmeier dispersion formula is accurate to $\approx 3 \times 10^{-6}$. The validity of the latter expression can be extended further into the ultraviolet by adding a λ^{-10} term and into the infrared by adding a λ^4 term.

Annealing of Optical Glasses

Much of our knowledge of the behavior of glasses in the glass transformation region is based on studies of the annealing of optical glasses. Glass-forming melts have much higher thermal expansion coefficients than those of solid glasses of the same composition. During cooling of a melt, the density increases rapidly so long as the melt is relatively fluid. As the viscosity increases, the rate of structural rearrangement of the melt slows and eventually becomes fixed at some effective, or fictive, temperature. If the melt is cooled rapidly, the structure of the glass will be representative of a high temperature melt. If the melt is cooled more slowly, the structure will have time to equilibrate at lower temperatures and the glass formed will have the structure of a lower temperature melt and hence will be more dense. Since the refractive index of a solid is related to its density, the refractive index will be less for the glass produced by rapid cooling than for one produced by slower cooling.

There are two ramifications of the cooling effect on refractive index. First, the refractive index will be different for different cooling rates, even though the composition is identical. Statements regarding the refractive index of a glass are only applicable for a known thermal history. Second, but less obvious, the outside of a piece of glass will cool more rapidly than the inside due to the kinetics involved with removing heat from the glass. As a result, a refractive index gradient will exist in the glass, with a lower refractive index at the surface. If the glass is ground and polished to form a lens, the lens will not perform as expected if the shape is based on the assumption that the refractive index is constant through the entire lens. While cooling-induced differences in refractive index between the surface and center of a large block of glass are usually of the order of 0.000 1, they can be ≥0.001 for a rapidly cooled block.

Control of the cooling rate is essential for the production of a block of glass of uniform refractive index. The cooling rates used to reduce stresses to acceptable levels in the production of nonoptical glasses are much too fast for the production of optical glasses. Use of a constant cooling rate throughout the glass transformation region will not yield a block of uniform refractive index. Ideally, the cooling rate should be decreased gradually as the temperature decreases throughout the glass transformation region. The production of a fine annealed glass requires a very slow cooling process, with a gradually decreasing cooling rate with decreasing temperature within the temperature region where permanent changes in the structure of the glass can occur.

Athermal Lenses

While the thermal expansion coefficient of many glasses is relatively low, it is not zero. As a result, the dimensions and refractive index of a glass are functions of temperature. Temperature changes in an optical system therefore result in changes in optical path length. While changes in optical path length of this magnitude may be unimportant for optical systems used within a narrow temperature range, those used over ranges exceeding 10 to 20 K must be designed to minimize thermal effects. Production of athermal optical systems, i.e., systems of constant power over a range of temperatures, requires use of glasses where the thermal changes in refractive index are exactly offset by the changes in dimensions, i.e., the optical path length is independent of temperature. This condition is met if the temperature coefficients of the index and the thickness are equal and of opposite sign.

The overall effect of changes in temperature on the optical path length are tabulated in terms of the thermo-optical constant, G, given by

$$G = \alpha(n - 1) + \mathrm{d}n/\mathrm{d}T \qquad [5]$$

where α is the linear thermal expansion coefficient, n is the refractive index, and $\mathrm{d}n/\mathrm{d}T$ is the thermal coefficient of the refractive index. Perfect compensation between changes in index and dimensions are obtained when $G = 0$. Since α is positive for most optical glasses, with values in the range of 3 to 15 ppm/K, this condition requires a negative dn/dT. Negative values of dn/dT are found primarily for glasses containing some fluorine, such as fluoride glasses and those in which oxygen is partially replaced by fluorine, such as fluorosilicates, fluoroborates, etc., or for some phosphate glasses. Silicate glasses rarely, if ever, have a negative value of dn/dT. For the same glassformer, e.g., P_2O_5, use of modifier ions with large diameters and low field strengths favor increasingly negative values of dn/dT.

The value and sign of dn/dT is determined by the competition between the decrease in refractive index induced by an increase in molar volume and the increase in the refractive index induced by an increase in polarizability of the ions in the glass with an increase in temperature. Equation [5] indicates that the absolute value of the refractive index is also important in determining the value of G. If the refractive index is high, then dn/dT must be more negative to yield $G = 0$. As a result, most athermal glasses are crown glasses. The value of dn/dT is also wavelength-dependent, generally decreasing with increasing wavelength. In particular, dn/dT becomes more positive at wavelengths near the UV edge of the glass. An athermal lens is only perfectly athermal at the design wavelength.

A few special glasses and glass-ceramics used for optical applications, e.g., vitreous silica, titanium-doped silica, and some aluminosilicate glass-ceramics, have thermal expansion coefficients which approach zero or are slightly negative. Vitreous silica has a very low refractive index and thermal expansion coefficient and can be used to produce lenses which, while not truly athermal, have relatively low temperature sensitivity. These low thermal expansion materials are often used as substrates for coated reflective optics (mirrors), where dimensional stability is far more important than any change in the refractive index of the substrate.

Production of Traditional Optical Glasses

Quality requirements for optical glasses are much greater than those for other common glasses. A constant refractive index throughout the glass requires an extremely high degree of homogeneity. The glass must be free from defects such as striae (regions of refractive index variation due to density/composition variations) and stones (undissolved material) and the bubble content must be as low as possible. The thermal history of the glass must be such that the refractive index variation, due to differing fictive temperatures, is minimized. Internal stresses must be minimized to yield a low birefringence. The glass must also be free of any species which contribute to absorption in the desired spectral region.

The quality demands of optical glasses require use of much higher quality, and thus more expensive, raw materials. Concentrations of third-row transmission metal ions must be held to rigid specifications, with maximum values usually in the ppm, or even in the ppb range, for the highest quality glasses. Defects are reduced by melting in platinum containers instead of the oxide containers used in the past. Stirring the melt, using platinum wherever possible, is essential for obtaining acceptable homogeneity. Control of the melting atmosphere may be essential for controlling the redox state of the melt. Rigorous annealing schedules must be used to produce glasses with the desired thermal history so that the problems of stress-induced birefringence and refractive index gradients are reduced to acceptable levels.

Non-Traditional Optical Glasses

While most traditional applications of optical glasses involve transmission in the visible, many modern applications require high transmission in the ultraviolet, near-infrared, or infrared regions of the spectrum. Transmission in each of these regions is determined by both the bulk composition of the glass and by the concentration of impurities present. Most nontraditional applications place even more stringent requirements on the purity of the materials used in manufacturing the glass. An extreme example of the demand for purity is found in the glasses used for telecommunication optical fibers, where concentrations of some impurities must be held to the low ppb level.

Ultraviolet-Transmitting Glasses

The ultraviolet transmission of a glass is determined by the intrinsic electronic transitions of the material, by charge transfer absorption bands due to impurities, and by surface and bulk scattering. The first of these is determined by the electronic bandgap of the material, while the other two are controlled by the care taken during production of the glass. These glasses are used in spectroscopy for windows, lenses, UV-lamps, and sample cuvets, for microlithography, and for lenses in excimer laser optical systems.

Vitreous silica is widely used as an ultraviolet transmitting glass for transmission optics at wavelengths well below 200 nm. The lowest energy absorption band for this glass occurs at 10.2 eV, or 122 nm. Many UV–VIS spectrometers, which normally are capable of measurements extending to 185 to 190 nm, use vitreous silica as optical elements. Additions of alkali or alkaline earth oxides to silica produce nonbridging oxygens in the network. Nonbridging oxygens do not bind the excitable electrons as tightly as bridging oxygens, which shifts the absorption edge to lower energies, i.e., longer wavelenths. These glasses are incapable of high transmission beyond 5 eV, or about 250 nm, and are not used for applications requiring good ultraviolet transmission.

Addition of elements which eliminate nonbridging oxygens favors good ultraviolet transmission. Alumina and/or boric oxide are frequently used in conjunction with alkali or alkaline earth oxides to counter the formation of nonbridging oxygens and preserve the ultraviolet transparency of silicate glasses. This approach can be taken to its logical conclusion by eliminating silica from the composition and producing alkali/alkaline-earth-aluminoborate glasses with good ultraviolet transparency.

In theory, fluoride glasses should provide better ultraviolet transmission than oxide glasses due to the larger bandgaps of fluorides. This assumption is true for single crystals, where fluorides such as CaF_2 are under development for transmission to higher energies than those obtainable with oxide crystals or glasses. In practice, the superiority of fluoride glasses for ultraviolet transmitting glasses has proven difficult to achieve. Vitreous BeF_2 and related glasses have superior ultraviolet transmission when compared to silicate glasses. Unfortunately, BeF_2 is very volatile and quite toxic, which restricts production of glasses based on BeF_2. Some fluoroaluminate and fluorophosphate glasses also have excellent ultraviolet transparency, but difficulties in producing these glasses in large sizes and of optical quality has limited their application.

Impurities are particularly important in production of ultraviolet transmitting glasses. The third-row transition metals, platinum, and cerium are all known to absorb strongly in the region from 200 to 300 nm when dissolved in glasses. Ferric iron has such a strong absorption band at $\approx$230 nm that concentrations of ferric ions must be kept in the low ppm range in glasses used for ultraviolet transmission. Since iron is a very common impurity in the sands used to produce silicate glasses, special care must be taken to minimize the Fe^{3+} content of ultraviolet-transmitting silicate glasses. Since ferrous iron has lower absorption in the ultraviolet, the effect of iron can be reduced by melting under reducing conditions. Platinum can enter melts from the melting unit and can be present as either ions or metallic particles. The ionic form of platinum absorbs ultraviolet light, while the metallic form scatters the same light.

Near-Infrared-Transmitting Glasses

Most glasses have good intrinsic optical transmission in the near-infrared. Absorption in this region is primarily due to impurities. Ferrous iron is especially detrimental to transmission in this region as a result of an extremely broad ligand field absorption band centered at about 1050 nm and extending into the visible and out to over 2000 nm. Reduction of absorption due to ferrous ions is obtained by careful control of raw materials to minimize the amount of iron oxide present and by melting under oxidizing conditions, which converts the ferrous ions to ferric ions. Conversion from ferrous to ferric ions increases the ultraviolet absorption of the glass. If good transmission in both the ultraviolet and near infrared regions is required, the iron content must be held to very low levels.

Although many rare earth ions also have absorption bands in this region, these elements are rarely found as impurities in sufficient levels to impair transmission in the near-infrared region for lenses and other common optical components. A more serious problem occurs for glasses used as optical fibers, where the greater optical path length magnifies the effect of impurities which may be undetectable in the spectra of samples only a few mm thick. The same can be said of absorptions of other ions which have very weak absorption bands in this region. Even higher order overtones/combination bands of infrared vibration absorption bands due to hydroxyl become important in optical fibers. While the hydroxyl absorption in infrared-transmitting glasses used in mm thicknesses is insignificant for concentrations less than 1 ppm, the hydroxyl concentration of optical fibers must be reduced to levels of a few ppb to reduce the effect of the overtone/combination bands to acceptable levels.

Infrared-Transmitting Glasses

Infrared transparency is limited by the intrinsic vibration, or multiphonon, edge of the glass and by the presence of impurities, primarily hydroxyl and carbon dioxide, which absorb in the infrared. The position of the infrared edge is determined by the force constant, which is proportional to the bond strength, and the masses of the ions. The best infrared transmission is found for materials with weak bonds and heavy ions. Infrared transmitting glasses are used in infrared spectrometers, lenses for pyrometers, infrared transparent domes for aircraft and missiles, and optical components for many systems using infrared lasers.

The fundamental absorptions of B_2O_3, SiO_2, P_2O_5, and GeO_2 lie in the region from 7.5 to 12 μm. These bands have such strong absorptions that the first overtone, at half the wavelength of the fundamental vibration, effectively determines the infrared edge for samples of more than a few hundred micrometers in thickness. The best infrared-transmitting oxide glasses do not contain significant amounts of B_2O_3, SiO_2, or P_2O_5. The lower fundamental vibrational frequency of GeO_2 allows production of some glasses of limited use for infrared transmission to about 5 μm. Comparable transmission is obtained for calcium aluminate glasses. Better transmission has been obtained for glasses combining the oxides of lead, bismuth, or cadmium with gallium oxide. These gallate glasses are termed heavy metal oxide, or HMO, glasses and transmit to ≈7 μm in mm thicknesses. The HMO glasses transmit into the visible, with a cutoff at 400 to 500 nm. Since these glasses are transparent to visible light, they offer advantages over many other infrared-transmitting glasses which are opaque in the visible.

Recently, heavy metal halide (HMH) glasses have been found to exist in a wide range of compositional systems. HMH glasses transmit much further into the infrared than the best oxide glasses, with cutoff wavelengths extending beyond 20 μm. These glasses are much more difficult to prepare than oxide glasses and are highly prone to crystallization, less durable in water, and mechanically weaker than the HMO glasses. Improvements in infrared transmission correlate with degradation in the other properties needed for most applications. Oxide impurities severely degrade the infrared transmission of these glasses. While much greater infrared cutoffs can be obtained in the laboratory, glasses currently produced commercially transmit to ≈8 μm.

Chalcogenide glasses contain either S, Se, or Te as the glassforming anion and are free of oxygen. Commercial compositions transmit to 16 μm, while laboratory samples transmitting beyond 20 μm can be made. These glasses have refractive indices in the infrared in the range of 2 to 3, very low glass transformation temperatures, which limit their temperature range of application, and are, with few exceptions, opaque in the visible. Toxicity is also a problem for Se and Te. Oxygen impurities must be kept below the 1 ppm level to avoid unacceptable absorption bands.

All infrared-transmitting glasses suffer from transmission interference in the region from 2.5 to 6 μm, due to hydroxyl in the glass. Since hydroxyl readily forms in any melt exposed to an atmosphere containing water vapor, great care must be used to reduce the hydroxyl content to acceptable levels. Glasses can be melted in sealed containers, under reactive atmospheres containing fluorine or chlorine, or under vacuum. Some glasses are prepared under dry conditions and then remelted under a vacuum to reduce the hydroxyl content. Similar care must be used to exclude CO_2 from HMH glasses, where a band at 4.25 μm has been assigned to the asymmetric stretching vibration of the dissolved molecules (Figure 2).

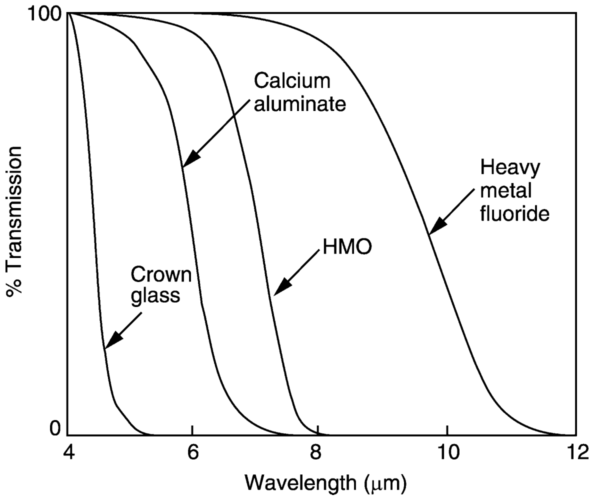

Figure 2 Infrared transmission spectra of several optical glasses showing the effect of glass composition on the infrared cutoff frequency.

Optical Isolators

Optical isolators based on the Faraday effect contain high concentrations of rare earths for paramagnetic devices and lead for diamagnetic devices. The highest concentrations of rare earth ions are obtained in aluminosilicate, aluminoborate, and aluminogermanate systems, where excellent quality glasses can be produced, while the best diamagnetic isolators are obtained using the lead bismuth gallate glasses.

See also

Optical Materials: Color Filter and Absorption Glasses. **Magneto-Optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping.

Further Reading

Deeg EW (1986) Optical glasses. In: Boyd DC and MacDowell JF (eds) *Commericial Glasses*, pp. 1–34. Columbus, OH: The American Ceramic Society, Inc.

Fanderlik I (1983) *Optical Properties of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Hoffmann HJ (1998) Differential changes in the refractive index. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 96–123. Heidelberg, Germany: Springer-Verlag.
Lentes FT (1998) Refractive index and dispersion. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 19–58. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Ultraviolet-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 288–297. Heidelberg, Germany: Springer-Verlag.
Liepmann MJ and Neuroth N (1998) Infrared-transmitting glasses. In: Bach H and Neuroth N (eds) *The Properties of Optical Glass*, pp. 297–308. Heidelberg, Germany: Springer-Verlag.
Pfander HG (1996) *Schott Guide to Glass*. London: Chapman & Hall.
Rawson H (1980) *Properties and Applications of Glass*. Amsterdam: Elsevier Scientific Publishing Company.
Shelby J (1997) *Introduction to Glass Science and Technology*. Cambridge, UK: Royal Society of Chemistry.
Vogel W (1991) Optical properties of oxide glasses. In: Ullmann DR and Kreidl NJ (eds) *Optical Properties of Glass*, pp. 1–36. Westerville, OH: The American Ceramic Society, Inc.

Plastics

T Bauer, Wahl Optoparts GmbH, Triptis, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Injection molded precision plastic optics in high volumes were first produced during the 1960s. After the development of sophisticated measuring and manufacturing methods in the late 1960s, precise aspheric surfaces were as easy to make as spheric contours in plastics. Today plastic is a widely used low-cost option compared to glass with even more degrees of freedom for optical and component design. Polymer optical systems are used in sensor applications, visual systems, cameras, mobile phones, video-conferencing cameras, scanners, security systems, and so on.

The physical and chemical properties of plastic materials are very different from those of glass. Generally speaking, glass materials are harder, more durable, and more stable against temperature and humidity than plastics. The variety of optical glasses comprises hundreds of different materials. Compared with this the choice for plastic materials is limited only to about 10 different materials (and even less optical parameter variations). However, plastic optics offers other design freedoms that are not achievable with glass optics.

The manufacturing technologies for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process whereas precision plastic lenses are usually made by injection molding, compression molding, or casting. Because of the material's characteristics and the manufacturing process, plastic optics have some unique advantages.

High production numbers at low costs

Injection molding is ideal for high-volume production with low-unit costs. Moderate raw material costs and multicavity mold designs (up to 32 cavities) allow large production volumes at a reasonable unit price. In spite of considerable tooling costs, these costs compared to glass design versions, can be relatively low.

Lightweight and hardiness

For a given volume glass is much heavier than plastic (by a factor of 2.3 to 4.9). However, plastic materials are relatively shatter and impact resistant. These features are important for head-mounted systems and other weight-sensitive applications (mobile phones, cameras, etc.).

Design potentials

Injection molding makes it economical to produce sophisticated optical shapes such as aspheres, diffractive optical elements, or even freeform surface structures. From the design point of view, the more sophisticated surface shapes reduce costs or obtain better performance (Fresnel structures, lens arrays, diffractive optical elements, etc.).

Optical systems and component assembly

For typical optical system designs optical components (mirrors, lenses, prisms, etc.) must be fixed in a mounting. With plastic optics it is possible to mold mounting elements, posts, or alignment notches integrally with the optical component. This can reduce part and assembly costs considerably.

Technologies adapted to plastic materials, such as ultrasonic and laser welding, gluing and integrated snap-in structures, allow fast and cost-efficient automated and manually operated assembling solutions.

History of Transparent Plastic Materials

A Material With No Use

In 1840, the Austrian chemist Redtenbacher (1810–1870), a student of Justus Liebig, discovered during a chemical experiment a malodorous transparent liquidity, the acrylic acid. It was an unwanted byproduct which was not further analyzed at that time. Nevertheless, that discovery can be considered to be the origin of today's plastic optics.

The development of plastics began in the US in 1860 as people were searching for an alternative for ivory in the production of billiard balls.

In 1869 John Wesley Hyatt developed from cellulose nitrate and camphor the first plastic material: celluloid. At first it was used for dental plates and afterwards for shirt collars and as a film base material. Despite its flammability and its lack of light resistance, it was a first economical success.

In the following decades the first synthetical plastics were developed, for example, the material with the trade name Bakelite in 1906.

Otto Röhm's Doctorate

The basis of acrylates was rediscovered in the records of Redtenbacher in 1900. From 1901 onwards, Prof. Pechmann (1850–1902) asked his postgraduates to do more research into the acrylic acid. The acrylic acid is an acetylene compound which can be extracted from carbon, petroleum, or lime.

In 1901 Otto Röhm, a student of Pechmann, wrote his dissertation 'About polymerisation products of the acrylic acid' – the basis of the methacrylate chemistry.

Hermann Staudinger (1881–1965), head of the Institute of Chemistry in Freiburg started theoretical investigations into the structure and natural characteristics of natural and synthetical polymers in 1920. He assumed that the molecules of plastic materials consist of numerous small molecule units. Experiments as proof of the theory led to a rapid development of scientific research and to significant breakthroughs in plastic chemistry. Staudinger won the Nobel Prize for Chemistry for his work in the field of macromolecules in 1953.

In 1928 methyl methacrylate was syntheticized for the first time and patented. Acrylic resin was used as a binder in multi-layer glass.

In the 1920s and 1930s a number of plastic materials were developed, including cellulose acetate CA (photo film, synthetic fiber); polyvinyl chloride PVC (pipes, coatings, isolations); urea formaldehyde resin (dishes, electrical equipment).

Polystyrene resins were commercially produced for the first time in 1937. Polytetrafluoroethene (PTFE) was produced for the first time in 1938, and distributed as 'Teflon' since 1943. Another key development was the synthesis of Nylon, the first technical high-performance plastic material.

Plexiglas Captures the World

The industrial production of methacrylic acid methyl ester began in 1934. With its polymerization to a hard and transparent plastic material, polymethylmethacrylic (PMMA or acrylic) was developed. It was patented in 1936 and caused, under the registered trade name Plexiglas®, a sensation at the world exhibition in Paris in 1936, almost 100 years after the discovery of acrylic acid.

Plates of Plexiglas were manufactured by means of effusion and polymerization between glass plates. At first it was used for watch glass, cockpits, and the construction of street lighting and luminous advertising. People already experimented with Plexiglas injection moulding in 1935. Arthur Kingston, the founder of Combined Optical Industries Limited (COIL) had invented and patented the first plastic lens in 1933.

The First Plastic Lenses in Cameras

The first camera, incorporating plastic optical parts, to be mass produced was the 'Purma Special' introduced by RF Hunter Ltd., in 1937, using 'Perspex' (another brand name for Acrylic), for the direct-vision viewfinder.

Due to the scarcity of natural raw materials at the beginning of World War II, the plastic industry became a source of outstanding substitute materials. This development had a lasting effect far into the post-war period. Results were mainly achieved with technical plastics, such as polycarbonates, acetates, and polyamides. Other plastics were developed as a substitute for metals as well as for hardhats, for high-temperature constructions, and for products which are resistant against acids and bases.

The German chemist Karl Ziegler discovered polyethylene (Polyethen, PE) in 1953 and the Italian chemist Giulio Natta Polypropylene (PP) in 1954.

The Eastman Kodak company is credited with giving birth to the modern-day plastic Fresnel lens. In 1946 Kodak developed a tooling and manufacturing process to mass produce plastic Fresnel lenses.

The circular Fresnel lenses made at that time had spiral grooves like music records.

It took another 10 years of improvement in injection molding techniques to the first plastic lens camera – the Kodak Brownie 44A from 1959.

Between 1953 and 1959 the Polycarbonate polymer was developed by Bayer and General Electrics. Currently acrylics and polycarbonates are still the most important materials for plastic optics.

But not only thermoplastic materials were developed over this time. In May 1940 research work resulted in the monomer allyl diglycol carbonate (ADC). In the following years more than 180 different compounds of this monomer were investigated and examined.

During World War II Columbia Southern Chemical Company (a subsidiary of Pittsburgh Plate Glass) began research into ways of obtaining nonthermoplastic materials. The project was given the name 'Columbia Resins'. The 39th tested (CR 39) turned out to be the most significant due to its unique properties and was used since 1947.

At first, solely the low-refractive CR 39 was available for organic eyeglasses. Only in the mid-1980s did the development begin in the field of high-refractive nonthermoplastic materials.

Properties of Plastics for Optics

As already mentioned, the choice for plastic optical material is limited only to about 10 different types of material. Optical properties (abbe value, refractive index, transparency) as well as mechanical, thermal, and humidity boundary conditions are decisive for the material choice. Different to glass during plastic processing (injection molding, casting, etc.) the process affects not only the geometry but also inner properties like refractive index, transparency, and birefringence.

Optical Properties

Fundamental optical properties are defined by optical transmission, refractive index, and dispersion. For plastic optics birefringence is an important parameter too.

Although the total number of plastic materials has increased in recent years, the range of refractive index and dispersion characteristic is limited to almost two major groups – crown-like materials such as Acrylic (PMMA), Polyolefin (COC, COP) and flint-like materials such as Polystyrene (PS), Polycarbonate (PC) and Styrol-Acrylnitril-Copolymer (SAN).

This strongly limited variety of plastic optical materials significantly restricts the optical design freedom (Figures 1 and 2).

Physical Properties

Important physical properties are weight, impact, and abrasion resistance – and thermal properties like temperature resistance and thermal expansion. Already, during system design, both mechanical and thermal properties have to be taken into consideration. The thermal expansion of optical plastics is approximately ten times higher than that of glass materials. In an optical system this effect has to be compensated for by optical design or mounting.

Typically most optical plastics can withstand temperatures up to 90°C. The maximum service temperature of Polycarbonate and Poly-Olefin materials reaches 120°C.

The specific gravity of plastic optical materials ranges from 1 to 1.3.

Polycarbonate has the highest impact resistance of all optical plastics and so is used for windshields and crash helmets. Acrylic has the best abrasion resistance (Table 1).

Common Materials

Thermoplastics

Thermoplastics are polymers with a linear molecule structure – with or without side chains – which can be deformed reversibly without any modification of their thermoplastic characterization by means of the impact of heat. Optical elements can be manufactured by injection molding or hot embossing.

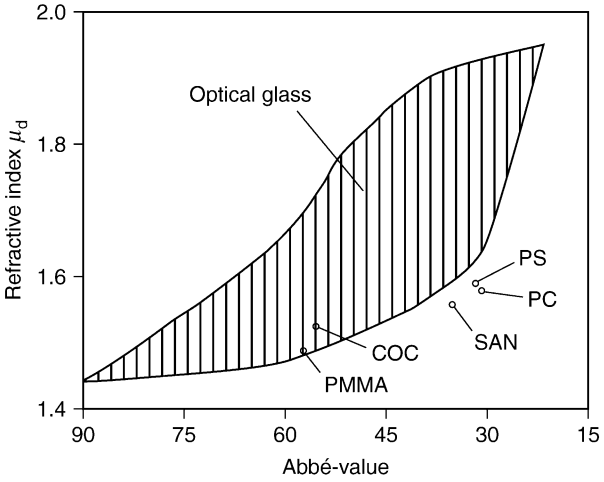

Figure 1 Abbe diagram.

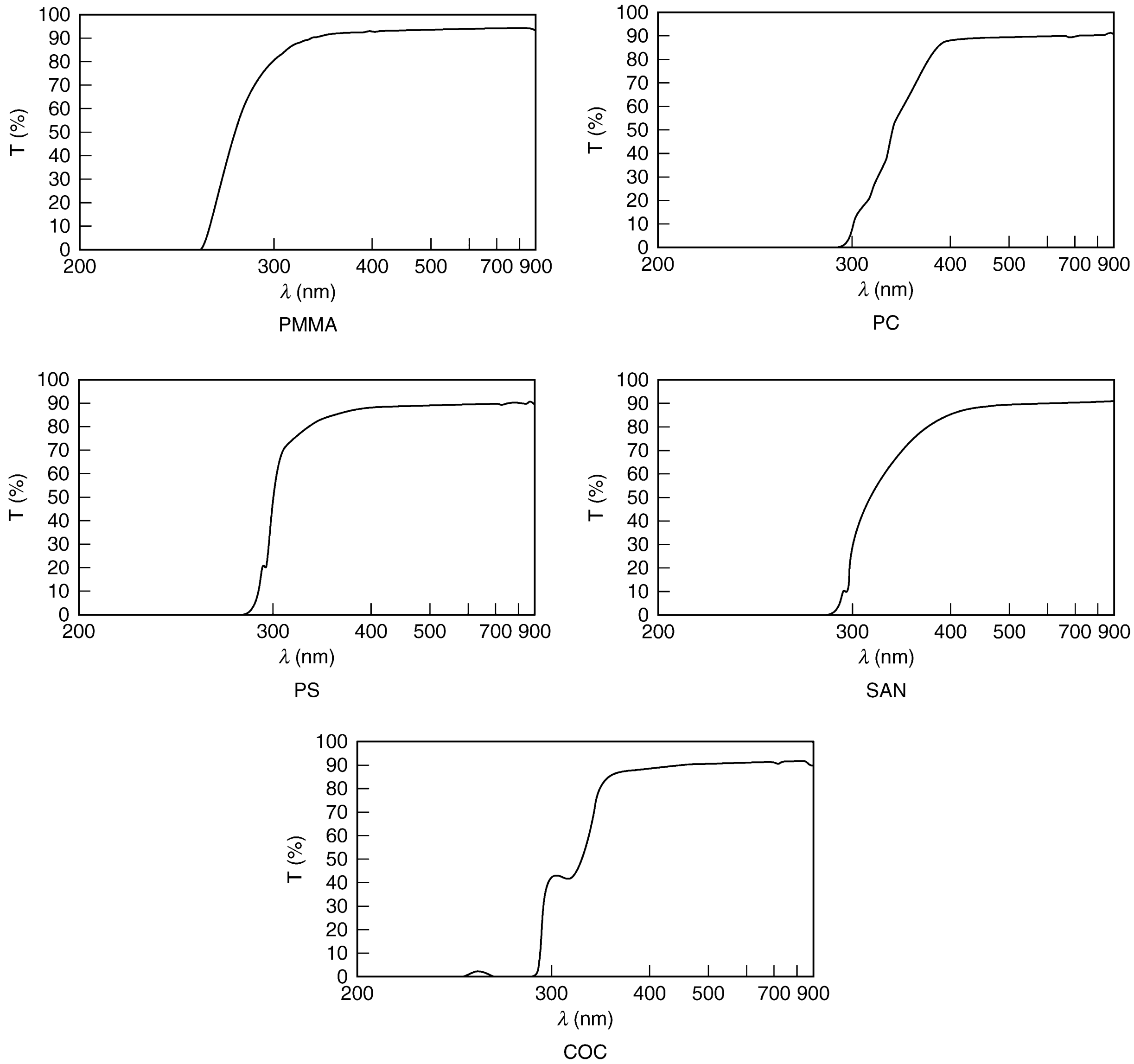

Figure 2 Transmission characteristics.

Typical optical thermoplastic polymers are listed below.

Acrylic or PMMA (Polymethyl methacrylate)

Acrylic is the most commonly used optical plastic. It is moderately priced, easy to mold, scratch resistant, and not very water absorptive. Its transparency is greater than that of most optical glasses throughout the visible range of the spectrum (organic glass). PMMA can replace glass in every application as long as the service temperature is less than 90°C and a low chemical resistance is required. Additives to acrylic (as well as to several other plastics) considerably improve its ultraviolet transmittance and stability.

PC (Polycarbonate)

Polycarbonates are linear polyesters of the carbonic acid which combine many characteristics of metal, glass, and plastics.

Polycarbonate is very similar to styrene in terms of optical properties such as transmission, refractive index, and dispersion. Polycarbonate, however, has a much broader operating temperature band of up to more than 120°C. It is used as the flint material for systems that have to withstand severe thermal conditions.

Another advantage is the high-impact resistance of polycarbonate. Safety glasses and systems requiring durability often consist of polycarbonate. Because of its high ductility polycarbonate is not easily machined.

Table 1 Material properties

Material	*Characteristics*	*Acrylic (PMMA)*	*Polystyrene (PS)*	*Polycarbonate (PC)*	*Styrol-Acrylnitril (SAN)*	*Cyclic Olefin Polymer (COP)*	*Cyclic Olefin Copolymer (COC)*	*Polyether-sulfone (PES)*	*Acrylnitril-Butadien-Styrolco-polymere (ABS)*	*CR39*	*Optical glass (B7)*
Optical	Spectral Passing Band (nm)	390–1600	400–1600	360–1600	395–1600	300–1600					
	Refractive index at 587 nm and 20 °C	1.4918	1.5905	1.5855	1.5674	1.5261		1.6600	1.5380	1.501	1.517
	Abbe value $(nD - 1)/(nF - nC)$	57.2 (53)	30.8	34 (29.9)	34.8	55.8	58	19.4		58	64.4
	Transmittance (%) thickness 3.2 mm	92	87–92	85–91	88	92	92	80	85		
	Haze% thickness 3.2 mm	1.3	2–3	1.7	1.5	1.5	1.5	1.5			
Physical	Specific Gravity (g/cm)	1.18	1.06	1.25	1.07	1.01	1.03	1.37	1.05	1.32	2.53
	Max. Service Temperature (°C)	92	82	124	95	123	130	200	90	130	400
	Linear Expansion Coefficient $(1/K)$	6.8×10^{-5}	8.0×10^{-5}	6.6×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	7.0×10^{-5}	5.5×10^{-5}	8.5×10^{-5}		1.1×10^{-4}
	Abrasion Resistance (1–10)	10	4	2		6	6				
Environmental	dn/dT ($\times 10^{-5}$)	−8.5	−12	−11.8…−14.3		−8	−10.1				
	Sensitivity to Humidity	high	low	low	medium	low		high	medium		
	Water Absorbtion (weight %) 23 °C, ISO 61	0.30	0.20	0.15	0.30	<0.01	<0.01	0.70	0.45		0
Manufacturability	Processability	excellent	good	poor	excellent	good					
	Birefrigence	low	high	high		low					
Chemical	Resitance to Alcohol	limited	good	limited				good			
Costs	approx. Material costs (EUR/kg)	3.3	2.5	4.4	4.4	27.1		21.0	3.5		25.0

PS (Polystyrene)

Polystyrene or Styrene is a low-cost material with excellent molding properties. Styrene has a higher index and a lower numerical dispersion value than other plastics. It is often used as the flint element in color-corrected plastic optical systems.

Compared with acrylic, styrene has lower transmission in the UV range of the spectrum. It has a lower resistance to UV than acrylic but is more easily scratched than acrylic.

Because its surface is less durable, styrene is more typically used in nonexposed areas of a lens system. A styrene lens, when paired with an appropriate acrylic lens, offers an effective achromatic solution.

SAN (Styrol-Acrylnitril-Copolymer)

SAN is a copolymer of certain fractions of Polystyrene and Acrylic (typically 70/30), which allows adjusting refractive indexes. It is a glassy polymer with a low tendency to stress cracks. SAN has slightly more chemical durability than PS and is a low-cost material with excellent molding properties.

ABS/MABS (Acryl-Butadien-Styrol-Copolymer)

ABS is developed from polymerization of styrene and acrylnitrit on polybutadien during the emulsion process. Process variations lead to different material characteristics. Compared to PS, the chemical resistance and temperature stability of ABS is considerably better. ABS has a natural color so it cannot be used for every optical application.

PES (Polyethersulfon)

PES shows a remarkable temperature stability, a low transmission against UV-light, a limited transmission with a natural yellow color, and cannot be used for every optical application. Because of its excellent temperature stability, PES is suitable for assembling, even on electronic motherboards by reflow soldering. This process causes temperature stress up to 220°C (Figure 3).

COP and COC (Cyclic olefin polymer and copolymer)

Cyclic olefin (co-)polymer provides a high-temperature alternative to acrylic. Its refractive index and transmittance are similar but the heat distortion temperature is about 30°C higher than for acrylic. It has a remarkably low water absorption value of less than 0.01% (compared with Polycarbonate 0.2% and PMMA 0.3%).

Cyclic Olefin Polymer (COP) is a family of amorphous plastic resins with low native stress birefringence properties. Zeonex is a resin developed by the Nippon Zeon Company – one grade, E48R, has a heat distortion temperature of about 122°C. Zeon Chemicals has another class of polymers called Zeonor, that can be provided as a lower-cost alternative to Zeonex.

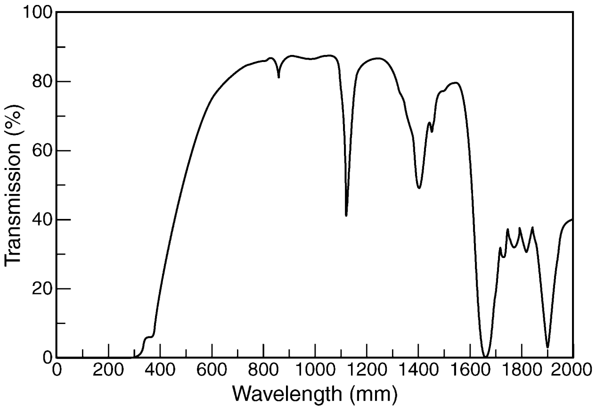

Figure 3 PES.

Cyclic Olefin Copolymer (COC), with the trade name of Topas, is another high-temperature alternative to acrylic. Generally, it is a lower-cost alternative to COP. Particularly favorable is the high flowability (good injection moldability) of these thermoplastics.

PMP (Polymethylpenten – TPX®)

PMP has a structure similar to PE, except that only the methyl groups are replaced with isobutyl groups. The chemical resistance can be compared to PP, although it tends to stress crack under the impact of ketones or chlorinated solvents. The main advantages of PMP are its excellent transparency and its good mechanical characteristics, even at high service temperatures of up to 150°C.

PE-HD (high density polyethylene)

If polymerization is controlled catalytically, molecules with a low number of side chains will be received. Compared to PE-LD (low density), the molecules are quite compact with a higher stability and chemical resistance and service temperature can be up to 105°C. PE shows a bad transparency (milky appearance) in the visible spectral range. However, this material is permeable for infrared beams and so is frequently used for motion detectors.

Elastomere

Elastomeres are plastics with loose and networked molecules which are rubber elastic at a normal temperature. The most popular elastomeres are natural rubber and silicone rubber. These materials are optically used as flexible light guides.

Duromeres

Organic glass is a fully synthetic plastic material available in a vitreous state. It consists of macromolecular organic compounds which do not follow any principle of periodic arrangement and are hence amorphous.

Duromeres are plastics with spacial closely networked molecules which are very hard and refractory at a normal temperature. Impact of heat causes an irreversible induration. In most cases, duromers are used to produce plastic lenses made of organic glass. Once they have been thermally treated after production, their shape can no longer be changed.

The well-known plastic CR 39 is one of the organic materials used for plastic lenses.

CR39 plastic is made by polymerization of the allyl diglycol carbonate (ADC). It is a transparent, thermosetting resin which combines an exceptional range of qualities which are not available in other plastic transparent materials. The CR39 plastic is colorless and completely transparent to the visible light and almost completely opaque in the infrared and ultraviolet region of the spectrum. For these reasons, it is largely used for the production of sunglass lenses. They can be colored by surface dyeing or bulk tinting, have high abrasion resistance, and high-quality optical properties. Their weight is about half as much as glass, they keep their excellent optical properties despite long-term exposure to chemicals and resist heat distortion up to 100°C. These qualities, combined with the exceptional optical characteristics, make CR39 the best choice for applications where severe conditions of use exclude all other optical materials.

Additives and Colors

The characteristics of many plastics, particularly of thermoplastic materials, can be changed by adding so-called additives.

UV-filter materials are frequently added to improve the stability of materials in the sunlight. Sometimes nonstabilized plastics tend to yellow (transmittance decreases). Color filters can also be generated by coloring. Infrared coloring permits the production of optical elements which appear black in the visible spectral range. These materials are particularly used for sensors, optical scanners, or remote control front windows. Even characteristics, such as temperature stability, can be influenced by additives.

Manufacturing Methods

The manufacturing processes for glass and plastic optics are totally different. Glass lenses are made by a grinding and polishing process. By contrast, typical manufacturing methods for precise technical plastic optical parts are diamond turning of plastic blocks, injection, and compression molding of granulates. Other methods, such as casting, have their application in special fields like ophthalmics (eyeglasses). In any case, processing transparent plastic materials for optical purposes should take place in a clean environment.

Diamond Point Turning

Diamond point turning is an ultra-precision machining process carried out on special high-performance, ultra-precision diamond CNC-machining systems.

In combination with vibration isolation systems, digital signal processor-based machine control and integrated measuring systems with nanometer programming resolution manufacturing of optical surfaces is possible. This method can be expanded to full 3D-milling systems. With such an arrangement the generation of free-form surface profiles is possible.

Because of the long production time and the high machine costs, this technology is used for prototyping in plastics and nonferrous metals, mold inserts, and series production plastic parts – which cannot be manufactured by injection molding (because of size or precision demands).

Injection Molding

Injection molding produces several parts per shot (molding cycle) in single or multicavity tools. This production method is used for most of the plastic optic parts. Other fabrication techniques are used only when molding is inappropriate.

A plastic injection-molding machine consists of a fixed platen, a moving platen, a clamping unit, and an injection unit. Molding of optical parts requires special machine configurations and auxiliary equipment.

Compression Molding

Compression molding is used in the making of Fresnel lenses or other micro-optical structures. The material is pressed between heated platens with accurately defined temperature cycling during pressing. Mold inserts are formed by electroplated copies, replicated from master structures. These masters can be, for instance, diamond turned structures or diffraction gratings made by holographic methods.

The compression molding process allows realizing small structures with high aspect ratios and tight angular and positioning tolerances.

Manufacturing Process

Optical and System Design

On the one hand, optic design for plastic optics in principle uses the same mathematical algorithms as optic design for glass. On the other hand, designing plastic optics requires a profound understanding of the material properties and the manufacturing processes. Knowledge of production technologies, material characteristics, and assembling methods, together with design expertise, are needed to fully exhaust what precision plastic optics can be. Simply substituting the indices of refraction and re-optimizing, the design will not succeed. Expert design assistance is essential at this stage.

Designing plastic optics with modern design tools allows greater design freedom. The advantage of combining integral mounting structures with the optical surfaces to create mounting flanges, alignment, and snap features, provides the ability to automated assembly. Aspheric, cylindric, or toroidal surfaces are as easy to realize as spherical ones. Microstructures such as diffractive optical elements can be integrated too.

Prototyping

After designing a plastic optical system, lens prototypes can be made by diamond point turning in various plastic materials. The best surface finishes can be obtained with PMMA. Materials such as polycarbonate, pleximid, or Zeonex do not yield smooth surface finishes.

A major problem is the availability of semi-finished plastic blocks in various materials. For PMMA, a wide variety of bars or plates are available. For other materials or colored options, semi-finished items have to be made by injection molding. Because of the high manufacturing costs, diamond point turning is only recommended for making a limited number of prototypes to verify functionality of the optical design and to perform first tests.

In this stage optical systems are normally assembled from single elements. Housing parts often are made from aluminum. The resulting surface quality and system performance cannot be a validation of the manufacturability by injection molding. To get reliable knowledge about this, making a molded prototype from a single-cavity prototype mold is recommended.

The Injection Mold

A high-quality injection mold is obviously essential for precise plastic optic parts. The parts can never be better than the tool – but good tooling, however, does not guarantee good parts. A strong understanding of the whole manufacturing process is the key to producing precision plastic components.

Any injection mold consists of three main parts. The upper half, which is affixed to the injection-side platen, the lower mold half, which is affixed to the ejector-side platen, and the mold-ejection mechanism. Guide pins and taper locks ensure proper alignment of the mold halves.

Production volume and precision required from a tool influences the selection of mold materials and built-in maintenance features.

Single-cavity and multi-cavity molds from two to eight, 16, or even 32 cavitys are used for plastic optic parts.

Thermoplastic materials shrink during cooling in the mold. This geometric effect has to be compensated for in the injection mold. Exact shrink rates can be calculated and the tool can be modified.

For manufacturing optical plastic parts by injection molding, the optical surfaces (plano or aspherical shapes, diffractive, conical, lenticular, and cylindrical surfaces) are generated as separate inserts in the tool.

Aspheric inserts are manufactured by two steps. First a best-fit curve is generated on a stainless steel substrate. The substrate is then subjected to a nickel-plating process (electroless nickel) that deposits a thin layer of nickel (up to 500 μm). In the second step, single-point diamond turning produces the final aspheric or diffractive curve in the nickel. Because the hardness of a nickel-plated insert is less than that of a steel spherical insert, it will be more susceptible to scratch defects (Figure 4).

Pre-Production

A pre-production stage is recommended to check the manufacturing process. Typically this is done by using a single-cavity prototype mold. This mold can be used to find optimal molding conditions.

Optical and mechanical design can be verified with real molded components and design revision is possible to affordable conditions. Often the prototype molds is used to start limited production since production tooling may take much more time.

Series Production

For series production of high volumes, multicavity production molds are required. Depending on quality,

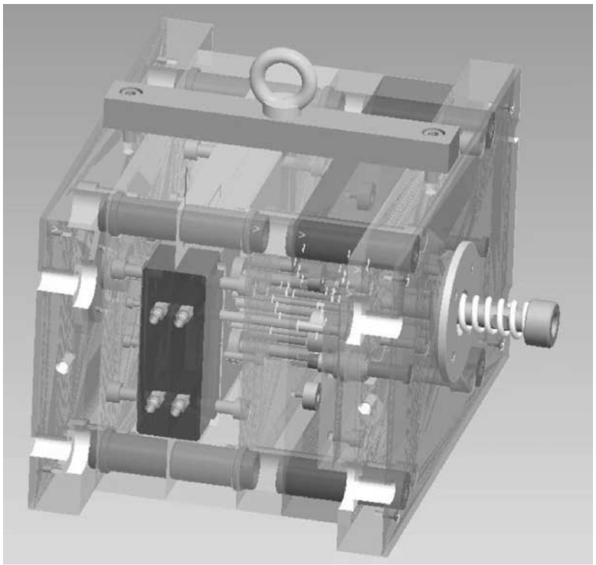

Figure 4 Injection mold.

volume, throughput, and cost, the production tool may have 2, 4, 8 or even up to 32 cavities. The production molds function for at least several hundred of thousands of injection cycles.

Coating and Component Assembly

Coating

Because of their limited temperature and UV resistance, plastic lenses must not be coated in an elevated temperature and radiation environment. During the deposition of thin films onto plastic, the coating chamber temperature is significantly lower than that for glass optics. This requires deposition techniques such as ion-assisted deposition to apply antireflective, conductive, mirror, and beamsplitter coatings.

Today, multi-layer dielectric coatings are routinely deposited on plastic components. Typical broadband antireflection coatings reduce reflection to about 0.5% per surface across the entire visible spectrum. Narrowband, multilayer antireflection coatings can achieve surface reflectances less than 0.2%. Multilayer dielectric coatings can be modified easily to scratch-resistant designs for front lenses and windows.

Several front and back surface reflector coatings are available for plastic substrates. Standard coating metals include aluminum, silver, and gold. Aluminum coatings provide surface reflectances greater than 88% across the visible spectrum and gold coatings greater than 95% for the near-infrared region.

Component Assembly

Many advantages of using plastic optics will come out in manufacturing components. The integration of optical, mechanical, and electronic elements allows building low-cost polymer optical components such as small camera lenses or scanner heads.

Plastic optical assemblies are usually carried out in a clean room environment by manual, semi-, or fully automated processes. The components have to be designed to ensure ease of assembly. Snap-on features, UV-cementing, heat-staking, and ultrasonic and laser welding, can also be employed with plastic materials.

Since most optical tolerances are additive, it is essential to establish quality check points in the manufacturing process to sort out nonconforming sub-assemblies.

Automatic in-line optical performance, monitoring such as MTF testing, can be implemented. SPC techniques should be used here to ensure the process is not drifting out of the controllable range.

Summary

For centuries, glass proved to be the material of choice for all optical applications. During the twentieth century, applications for glass elements and complex optical systems greatly expanded. Early uses of plastic molded elements included toy objectives, gauge windows, and watch crystals. Throughout the 1970s and up to the present day, high-grade polymers were developed specifically for optical applications. These advancements in materials, coupled with improved mold design have enabled plastic optics to replace glass optics in a wide and growing number of applications.

Plastic optic technology permanently expands their traditional limits and fields of application. On the one hand, plastic optic is still limited by material properties but, on the other hand, design and assembling freedoms allow new approaches.

See also

Diffraction: Fresnel Diffraction. **Geometrical Optics:** Lenses and Mirrors. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Color Filter and Absorption Glasses.

Further Reading

Corning Precision Lens Inc. (2000) *The Handbook of Plastic Optics*, 2nd edn. Cincinatti, OH: Corning Precision Lens Inc.

OPTICAL MATERIALS

Sculptured Thin Films

K Robbie, Queen's University, Kingston, ON, Canada

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical materials, like all materials, ultimately derive their physical properties from their chemical composition and physical form. Aluminum behaves differently from copper because of a chemical difference, while graphite behaves differently from diamond because of a crystallographic difference. For materials that are not infinite perfect crystals, an important contribution to their electromagnetic (optical) response comes from the morphological form of the material. Morphological form describes the physical structure of the material: a perfect crystal is one type of morphological structure, a porous fibrous bundle is another. An example of a morphological difference is that between snow and ice. Snow and ice are chemically identical, composed of solid water, but they exhibit different optical properties because of a difference in morphology. The polished glasses and large perfect crystals that comprise most optical components have morphologies more like ice than snow; their morphologies are dense and crystalline, with few and small disordered 'defects'. Unlike ice, snow is rich with morphological structure (see Figure 1) arising from the complex growth of each flake in the clouds, and the aggregate growth of snow on the ground. The morphology of snow scatters all colors of visible light, giving it an opaque white appearance, while ice is transparent and nearly colorless.

Sculptured thin films (STFs) are materials with controlled and designed morphology that exhibit a rich behavior unlike that of the constituent materials in bulk; they are the snow to most optical materials' ice. Unlike snow, sculptured thin film morphology is designed to scatter light in a controlled manner to produce a desired optical response. A sculptured thin film could be a coating applied to an optical element where the coating converts polarizations of light, circular or linear, but is composed entirely of an optically isotropic material (which doesn't usually affect the polarization of light). The film morphology is designed to create the desired optical properties. Sculptured thin films are also useful for magnetics, where an example of a sculptured thin film is the metal film coated onto plastic tape for video, audio, and data recording. The film is deposited under conditions that impose an elongated shape onto the magnetic domains, altering the magnetic properties. The morphological form of a sculptured thin film defines the way it responds in electromagnetic, chemical, and biological interaction, and there is hope that these unique materials will someday form the core of technologies that make the world a better place.

Structure in Thin Films

As chemical composition and morphology determine the electromagnetic response of materials, it would be wonderful to have a technique where we could build materials atom by atom, placing each exactly where we wished, and varying the element at will. With this level of control it would be possible to design and fabricate materials for almost any conceivable optical application. In some sense this is what is done now to fabricate lasers and optical switches. Microelectronic and micromachining processes have produced fantastic optical devices by controllably mixing materials (mostly semiconductors) and patterning designed structures. These techniques typically employ a repeated series of thin-film depositions and photolithographic patterning. Examples of thin-film coatings include antireflection coatings on eyeglasses, hard coatings on grocery store barcode scanners,

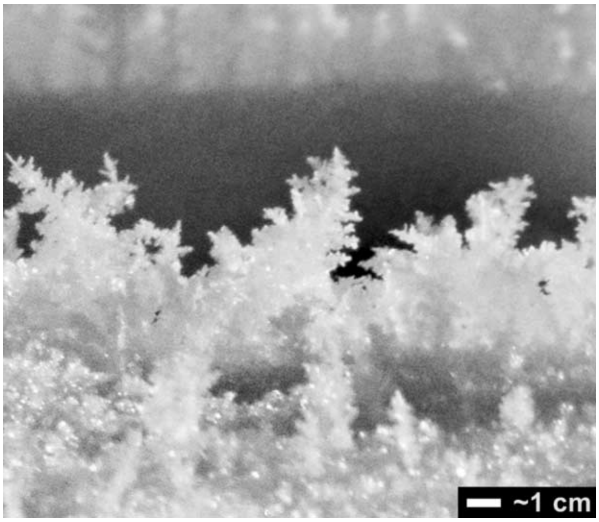

Figure 1 Photograph of snow accumulating at $-20\,^{\circ}\mathrm{C}$. Photo courtesy of Chelsea Elliott.

ultraviolet blocking aluminum inside potato chip bags, and all integrated electronic circuits. Thin films are usually produced by vapor deposition in vacuum, and can be made with most elements and inorganic compounds (and a growing list of organics). Most thin films grow as uniform, dense, two-dimensional (2D) layers, and three-dimensional structure (3D) is produced by removing portions of the film with photolithography and etching. Photolithography is a process where ultraviolet light is projected through a mask and microscope system onto a photosensitive polymer layer atop the thin film. The light hardens areas of the polymer in the pattern of the mask, preventing them from dissolving in a chemical developer solution. Exposed portions of the thin film can then be etched away with chemical or other means. Repeating as desired, many steps can create complex 3D designed structures. Computer chips and other integrated circuits are semiconductor crystals covered with a layered structure of conductors (aluminum, copper, platinum) and insulators (metal and semiconductor oxides) patterned with thin film deposition and lithography. Semiconductor lasers and many other optical devices are made this way as well, usually employing compound semiconductors (GaAs, AlGaAs, GaAsP, InP, etc.). The micrometer-scale patterning made possible with thin-film deposition and etching techniques has produced a wealth of semiconductor optical devices.

Optical lithography, however, is limited in size scale to dimensions comparable to the wavelength of light used to expose the photomask. Modern UV lithography can pattern 100–200 nm structures. Visible light has wavelengths from 400 nm (violet) to 700 nm (red). Shrinking the dimensions of physical structures is a great benefit for optical devices as it allows light to be controlled coherently without strong random scattering. Structures with dimensions comparable to the wavelengths of light strongly scatter light waves; scratches on optical components or imperfections in materials with dimensions comparable to the wavelength are considered defects, and are usually deleterious. However, materials with morphology much smaller than the wavelength (nanometer structure, or nanostructure) are seen by propagating light as homogeneous, with polarization and dispersion effects characteristic of the underlying structure. Sculptured thin films are materials where complex nanostructure has been created by controlling growth parameters dynamically during deposition of a thin film. The simplest sculptured thin film might be one where the temperature is periodically varied during thin-film deposition (although this is a very simple example, lacking the anisotropy central to the sculptured thin-film concept). Variation in the growth process with temperature affects film density, producing a porosity variation (periodic inhomogeneity) as a function of film thickness. Varying porosity produces varying refractive index, and this film would be an optical interference filter. The largest limitation of sculptured thin films is that all attainable structures must be characteristic of the growth process. As all morphological control is accomplished by varying film growth parameters, structure that is not characteristic of thin film growth is not possible. Fortunately, thin film deposition and growth is sufficiently complex to allow the fabrication of a large range of useful structures. As with lithographic patterning, the structural control is two dimensional, but now the third dimension is constructed by continuing the deposition, not by repeating the deposition and lithography cycle. Complex structures can be produced on the scale of tens or hundreds of atoms (approximately tens of nanometers) with controlled 3D morphology. This reduction in size scale compared to lithographic patterning is a huge benefit. Materials can now be designed and created with subwavelength structure to coherently scatter light. Because the morphological structure (<100 nm) is much smaller than the wavelength of visible light (~500 nm), light traveling in a sculptured thin film sees a medium with averaged properties. The material can be designed to control refractive index through a wide range, and birefringence and circular polarization effects are controllable.

The morphological structure of thin films is a vast topic. Thin films are produced with a nonequilibrium growth process, somewhat like the growth of clouds, and the mechanics of the growth determine the structure, which is often very different from that seen in a bulk sample of the same substance. Crystal and morphological structure vary with deposition process conditions, and vary during deposition to record a history of growth much like the rings of a tree, or stratified rock. If chemical composition is varied during the deposition, that change will be reflected in layers in the film. If vapor energy or temperature is varied, variations in density and crystal structure will result. Thin film growth is an aggregation process (the film accumulates over time), and exhibits many of the features of similar processes, sometimes producing materials with structures similar to broccoli or clouds. This self-affine (self-similar) structure is observed over a large range of sizes under some film deposition conditions. Under other deposition conditions, crystallographic effects dominate growth and films become layers of faceted crystals. In sculptured thin films, these growth characteristics are controlled and exploited in a functional manner to design material properties. For example, magnetic video and audio tapes that have the symbol ME, abbreviating ‘metal evaporated’, employ a structured magnetic thin film that has superior magnetic response for recording. Similar effects are observed in optical materials, and sculptured thin films appear promising for improving existing optical devices or enabling new ones to be developed.

The most studied form of sculptured thin films is that produced by an asymmetric deposition process. In deposition by vacuum evaporation, a thin film is fabricated by evaporating some source material in vacuum, and condensing a vapor onto a substrate. Eyeglasses, mirrors, television phosphor screens, etc. are coated this way. In most thin films the vapor arrives straight down (perpendicular) to the plane of a flat substrate. The film structure reflects the symmetry of the vapor, and is typically composed of a dense layer with some defects or voids oriented perpendicular to the substrate surface. A thin film exhibiting this morphology is shown in the scanning electron microscope image of Figure 2. When vapor arrives asymmetrically onto a substrate, however, that asymmetry is reflected in the growth, producing structural asymmetry in the resulting film, and altered physical properties. A second scanning electron microscope image is shown in Figure 3, of a sculptured thin film where the orientation of the substrate relative to the vapor is changed twice, producing three layers with different morphologies. The thin uniform gray at the absolute bottom of the image is the edge of the silicon crystal substrate which has been cleaved to reveal an interior section of the film. The bottom and first film layer is inclined to the right, toward the vapor source for the duration of this first part of the deposition. The middle and second film layer has a twisted structure created by slowly rotating the substrate relative to the vapor, while keeping the tilt fixed. The top layer has a vertical structure created by rapidly rotating the substrate. The entire film was produced with the vapor arriving at an angle of 45°, measured from either the plane of the substrate or its perpendicular. Varying the tilt angle has a profound effect on the morphology of thin films. The void structure in these films records the geometry of the deposition. When we cleave a film we are able to observe a 2D slice of this rich morphology. At large tilt angles, glancing angle deposition (GLAD) produces films with highly porous and complex structure. At these angles, the film structure becomes isolated columns of material where growth variations are now reflected in variations in the shape of individual columns. At less oblique vapor angles, growth variations are reflected in the shape of networks of atomic-scale voids within a largely dense film. Indeed, for the best optical devices fabricated to date (with the serial bideposition technique described later) it is difficult to distinguish any variation in structure even using a high-resolution scanning electron microscope. The film deposition process can be designed to control the morphology of the growing film, and to vary the structure as a function of time and thickness. The morphology of these films determines their optical properties; optical birefringence, second harmonic generation, filter bandwidth and attenuation, etc. can all be controlled in sculptured thin films.

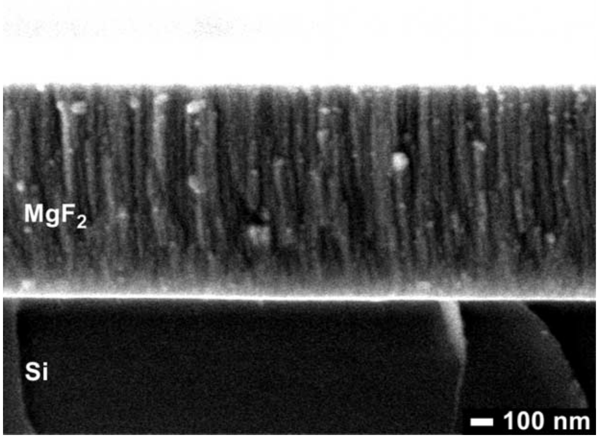

Figure 2 Micrograph of MgF_2 film deposited onto a 75 mm silicon wafer substrate, then fractured to reveal the edge shown.

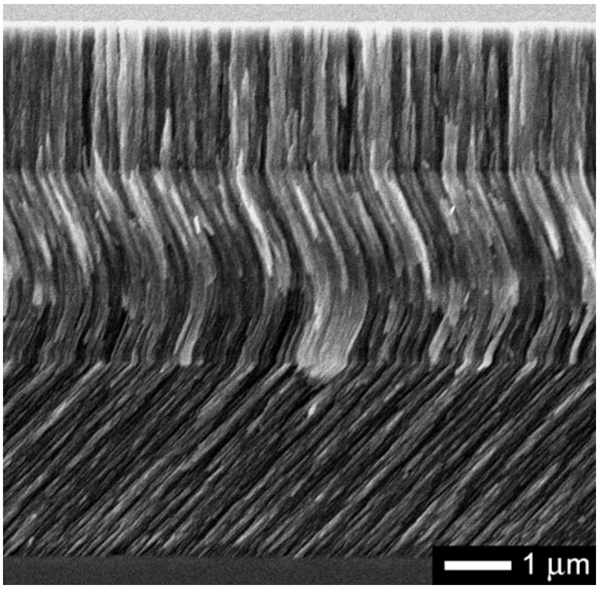

Figure 3 Micrograph of SiO film deposited with a tilt angle of 45° with one layer of no rotation, one layer with one rotation, then one layer with rapid rotation.

Optical Thin Films

Light can be described as a traveling wave with electric and magnetic fields oscillating perpendicular to each other, and to the direction of propagation. If the orientation of the electric field remains constant as the wave propagates, the light is said to be linearly polarized. If the orientation of the electric field rotates through a circular path, the light is circularly polarized, and can be left or right handed. In the most general description, the electric field traces an ellipse during propagation and the light is said to be elliptically polarized. The experimental study of light polarization is called ellipsometry. When light propagates through vacuum or air, the state of polarization remains constant. When light propagates through some types of materials, however, crystallographic or morphological asymmetry (or more correctly, anisotropy) results in different propagation conditions for different polarizations. Upon passage through these materials, the orientation and form of the state of polarization is changed, for example from left circularly polarized (LCP) to right circularly polarized (RCP). When multiple layers of anisotropic materials are combined into an optical circuit, as in the sculptured thin film concept, highly controlled optical response characteristics can be created.

The discussion of the optical properties of thin films begins with definitions of isotropic, uniaxial, and biaxial materials. Isotropic materials have optical properties that are independent of the direction that the light propagates through them. Vacuum and air are isotropic, as is glass. Rotation of the optical material relative to the light propagation direction has no effect on how the light propagates, and the state of polarization remains constant. Uniaxial materials have properties that vary with direction, but with one preferred direction (or axis) where response is independent of rotation. Biaxial materials have properties that are even more dependent on propagation direction, and require two axes to describe their optical properties. For biaxial materials there is no propagation direction where the optical response is independent of rotation. The mineral calcite is an example of a uniaxial material where the refractive index, and therefore the speed of light propagation, varies with direction relative to the crystal structure. This is a direct consequence of the trigonal crystal structure of calcite, and is an example of linear birefringence. Birefringence is the effect caused by anisotropic materials whereby light of different polarizations, even within one ray of light, travel with different speeds. When light propagates in a birefringent material, the orientation of the light's electric field relative to the material's optical axis determines the speed of propagation. Light with different polarization (orientation) propagates at different speeds, and a beam's polarization state is altered on passage through the material. Thin films can be isotropic, uniaxial, or biaxial, depending on the conditions under which the film is deposited. Sculptured thin films are films fabricated in a manner that controllably enhances the anisotropic optical response to produce materials with designed optical properties based on control of their morphological structure. As in calcite where the anisotropic crystal structure yields a uniaxial optical response, anisotropic morphological structure in sculptured thin films yields biaxial optical response. A plan view scanning electron microscope image of an anisotropic thin film is shown in Figure 4. The film was deposited with two vapor sources with the substrate placed between and the vapor arriving from both sides at an angle of 70° from the substrate perpendicular. This film is similar to those produced with the serial bideposition technique, and is optically birefringent. Polarization filters and other optical components can be constructed by designing the morphological structure of the thin film.

As discussed above, when a thin film is deposited from a vapor arriving perpendicular to a flat substrate surface, a structure composed of small (tens of nanometers diameter) columns of film material grows, with the columns oriented perpendicular to the substrate (Figure 2). This film is optically uniaxial, with the preferred axis perpendicular to the substrate, and parallel to the vapor flux. Light propagating along this optical axis (perpendicular to the substrate) travels at one speed regardless of polarization, and independent of rotations of the substrate around this axis. Light propagating along other paths will travel at a speed that depends on the polarization of the light

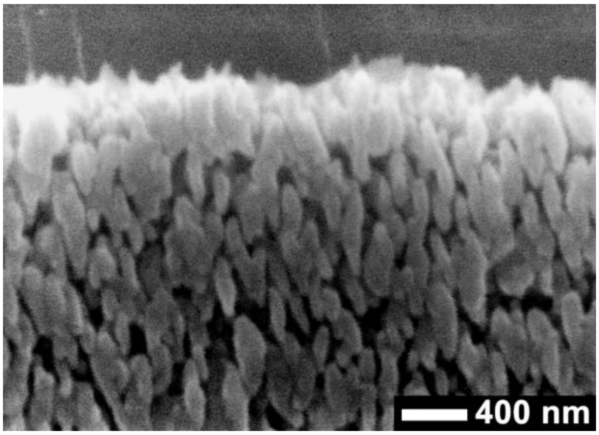

Figure 4 Micrograph of MgF_2 film, viewed from above, deposited with simultaneous growth from two vapor sources.

(orientation of the electric field relative to the propagation direction). The effect is small, and arises from the morphological structure of the film. More pronounced optical response is produced when the vapor arrives onto the substrate at an angle from the substrate perpendicular, producing optically biaxial materials, and enabling exciting polarization and filtering effects. If a thin film of magnesium fluoride is deposited onto a polished piece of glass with the vapor arriving at 45° (relative to an axis perpendicular to the substrate), a slanted columnar film is produced. The morphology of the film is tightly packed columns of polycrystalline magnesium fluoride with diameters between 50 and 200 nanometers (nm), inclined toward the vapor source, and penetrated by a network of atomic-scale voids. Because of the anisotropic nature of the deposition, the voids are anisotropic, and the film is optically biaxial. If the substrate is then held at that tilt angle and is rotated during deposition, the anisotropy in the structure rotates with the deposition, and an important type of sculptured thin film is produced. The electromagnetic description of this type of film is a 'helicoidal bianisotropic medium', describing both the helix shape of the rotation of the structure and the local bianisotropic (biaxial) physical, and optical, structure. Films with helicoidal symmetry are ideally suited for studies involving circular polarizations of light, and many demonstrations of sculptured thin films have involved these films and circularly polarized light. Like many organic molecules and some mineral crystals, these film are chiral, existing in two mirror image forms, called enantiomers. A right-handed structure rotates clockwise traveling away from an observer; a left-handed structure rotates counterclockwise. Also like some organic molecules and mineral crystals, helicoidal films are optically active – they affect the polarization of light. The observed effect depends on the wavelength of light relative to the physical helicoidal structure. If these two lengths match, a strong resonance effect is observed where the light wave coherently scatters from the anisotropic structure. Filters to produce and convert circular polarized light have been constructed with sculptured thin films of this type.

The range of structural control over the morphology of thin optical films limits the ability to tailor optical response for desired applications. While optical systems must always be treated as a whole to determine response properties, the description of some basic elements is useful. A film deposited at normal incidence will be uniaxial with the axis perpendicular to the substrate. For light propagation along this axis (common for optical systems), the material exhibits no birefringence. This structure is often called an isotropic element even though the columnar structure is slightly optically anisotropic. A truly isotropic layer could be produced by ion bombardment or elevated temperature during deposition, either of which would have the effect of eliminating the columnar structure and reducing anisotropy in the film. The second elemental structure is a slanted columnar film that is commonly described as a matchstick morphology. It is produced by changing the vapor incidence away from the substrate normal, leading to an inclination of the column structure toward the vapor source, as described above. Further structures can include: helicoidal structure where the anisotropy rotates through the thickness of the film, and periodically bent nematic structure where the anisotropy tilts in an s-shaped pattern through the film (fabricated with a complex pattern of rotations and depositions). These four elementary microstructures are shown in Figure 5. Electrical, optical, and magnetic anisotropy can also be increased by depositing obliquely from two sides instead of one, and this technique (called serial bideposition) can be used to increase the anisotropy in sculptured thin films. At extremely oblique deposition angles, a regime called glancing angle

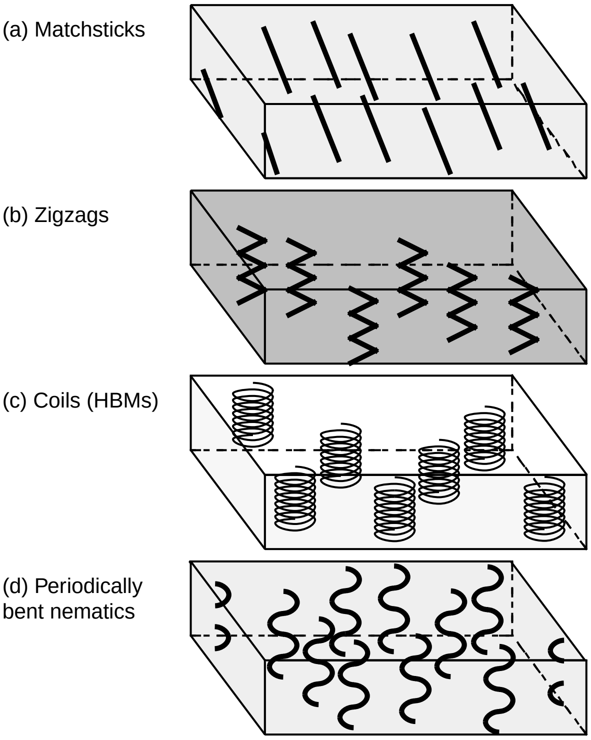

Figure 5 Elementary microstructures of sculptured thin films. From Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 1(2): 165–176, reproduced with permission.

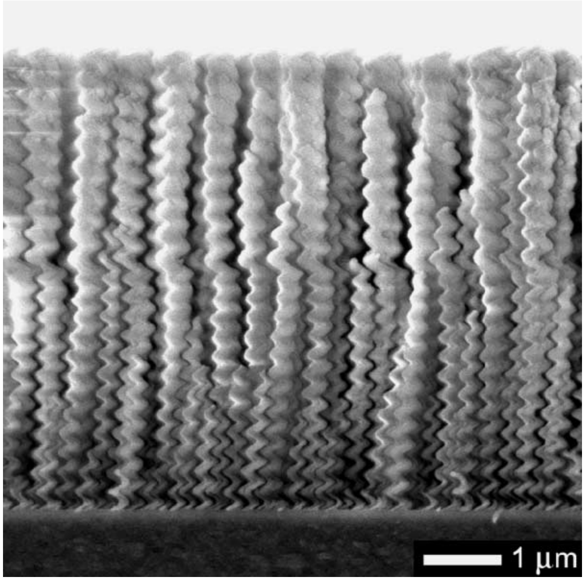

Figure 6 Micrograph of MgF_2 film deposited with a tilt angle of 85° and substrate rotation clockwise for the first half and counterclockwise for the second.

deposition, a rapid increase in porosity with approximately constant column tilt results from a competition between deposition processes. Films deposited here are highly porous and strongly scatter light, limiting their usefulness for conventional optical devices, though some application has been seen in these highly porous films when impregnated with liquid crystals, and switched similarly to liquid crystal displays. An example of a highly porous optical filter with circular polarization effects is shown in Figure 6, where two layers of helicoidal porous films are deposited with different handedness and pitch. Sculptured thin films for most optical devices are deposited at less oblique angles to reduce porosity and therefore diffuse light scatter (haze) as well as sensitivity to humidity and other environmental changes.

Combining the concepts of film growth and optical response to morphology leads to the hoped-for vision of sculptured thin films: optical circuits in a chip. Combining the elementary microstructures described above, and employing powerful computer growth modeling and optical design programs, researchers are currently defining and producing sculptured thin films with increasingly complex optical response characteristics. There are substantial steps required to move from the fairly simple optical filters fabricated today to large-scale devices with complex functionality, but continuing success and increasing interest suggest that sculptured thin films could become a dominant optical technology in years to come.

Current Research

While thin films deposited at off-perpendicular incidence angles have been studied for many years, the last decade has produced significant new insights into these materials, leading to the concept of sculptured thin films. Understanding that morphologically rich materials can be used to produce new optical effects has yielded some impressive demonstrations. Research is also underway to explore these materials for chemical, magnetic, biological, and other applications.

Significant research has been conducted in the past few years to develop and extend the concepts of sculptured thin films. Optical studies, both theoretical and experimental, have revealed a range of remarkable properties including circular birefringence and Bragg reflection, liquid crystal alignment and switching, optical interconnection, and even suggestions of photonic bandgaps. Studies of optical sensing of gases and other materials is progressing, and simulation and modeling of the growth processes is ongoing. The following are two examples of optical materials that have been fabricated with sculptured thin-film techniques; one is a polarization discriminatory light-handedness inverter and the second is a liquid crystal switching cell that can be used to electrically modulate the transmission of polarized light. The handedness inverter converts left circularly polarized light to right circularly polarized within a spectral band, and blocks right circular polarized light. It is constructed by combining a layer with vertical columnar, or matchstick, morphology together with a layer with helicoidal morphology. In keeping with the sculptured thin film concept, the films are deposited sequentially during the same deposition, and optical properties are monitored during growth. The transmission spectrum for the devices is shown in Figure 7, where T_{RL} and T_{LR} are the light transmittances of right circularly polarized light when the incident is LCP, and of LCP when the incident is RCP, respectively. LCP light is almost entirely converted to RCP upon transmission, and RCP light is blocked. The bandwidth of the devices is about 50 nm, determined by matching of the circularly polarized light with the layer with helicoidal morphology.

When films are deposited with helicoidal morphology at glancing angles, i.e., glancing angle deposition, highly porous structures are produced where the columns stand alone and are individually shaped by the deposition process. Optical devices can be constructed by filling these porous films with gases or liquids, and there have been demonstrations of display-type switching with liquid crystals, and discussions of chemical sensing possibilities.

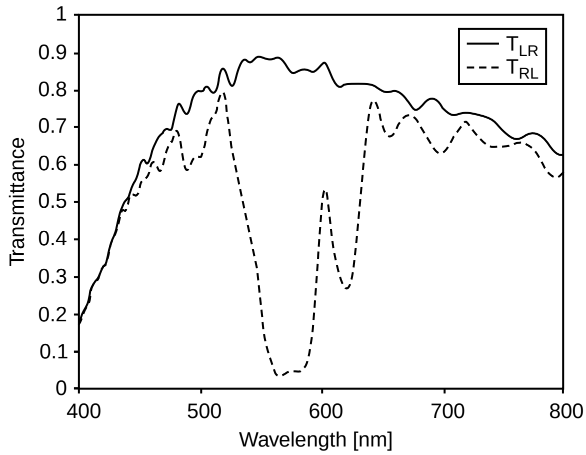

Figure 7 Transmission spectrum for visible light through a polarization-discriminatory light-handedness inverter. Adapted with permission from Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters, *Optical Engineering*, **39**: 2831–2834.

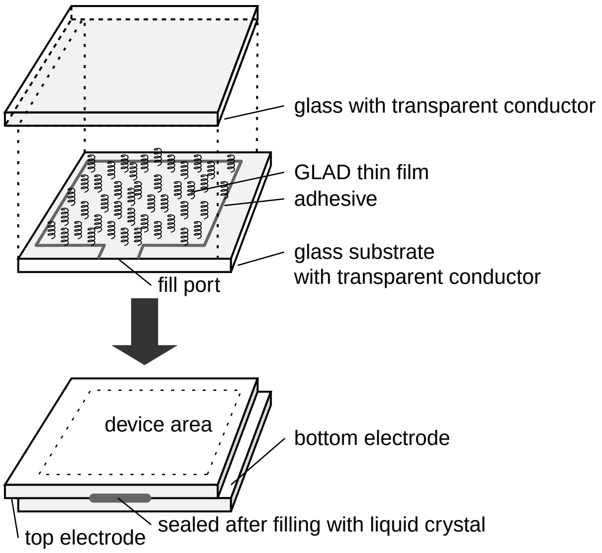
glass with transparent conductor
GLAD thin film
adhesive
glass substrate
with transparent conductor
fill port
device area
bottom electrode
top electrode
sealed after filling with liquid crystal

Figure 8 Liquid crystal switching cell schematic with highly porous glancing angle deposited (GLAD) film between glass plates coated with transparent electrodes. Reproduced with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

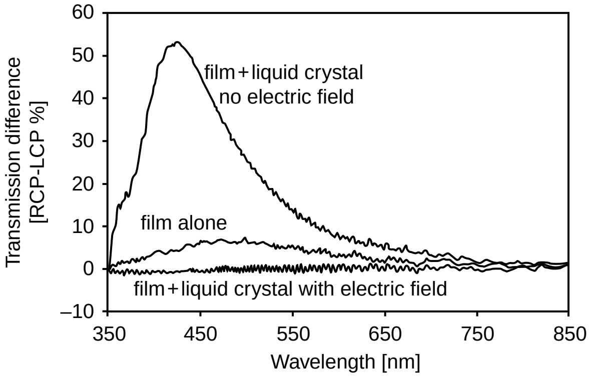

Figure 9 Transmission spectrum for visible light through a liquid crystal cell where the difference between right- and left-circularly polarized light transmission is eliminated with applied electric field. Adapted with permission from Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391, www.tandf.co.uk/journals.

The device shown in Figure 8 is a liquid crystal switching cell where the space between the electrodes, usually composed solely of liquid crystalline material, is now filled with a highly porous helical structured film. The film structure is somewhat similar to helicoidal film produced for the handedness inverter discussed above, but glancing angle deposition has produced a film of individual formed columns instead of a more dense film with a helicoidal void morphology. When the porous portions of the film are filled with simple nematic liquid crystalline material, the orientation of the rod-like molecules is strongly influenced by the helical columns. An orientational phase occurs in the liquid crystals mirroring the helical structure of the columns in the film, and producing optical properties similar to chiral nematic liquid crystals. The discrimination of left and right circularly polarized light increases relative to the unfilled film. When an electric field is applied across the liquid crystal cell, dipole moments in the liquid crystal molecules force them to align with the field, and chiral optical properties vanish (see Figure 9). Used as an electro-optic filter for circularly polarized light, this cell can switch light transmission on and off. Initial work has not shown clear benefits of this system over conventional liquid crystal display technology, but control of liquid crystal orientation throughout the cell instead of just at the glass plates appears promising. Possibly more complex optical circuits, incorporating layers of porous films filled with liquid crystals, will be useful.

Theoretical and experimental study of sculptured thin films is in its infancy, and many exciting uses for these materials are expected in coming years. Continuing study of thin film growth dynamics will improve and expand the range of structures attainable. Theoretical investigations will aid in the design and understanding of the optical response.

List of Units and Nomenclature

Aluminum (Al)	white metal used in airplanes, mirrors, and microelectronic circuits

Anisotropic	having properties that vary with orientation (rotation)
Biaxial	having anisotropic optical response characteristics that require two axes to describe
Chiral	having a shape that exists in two mirror image forms, called enantiomers
Diamond	carbon crystallographically ordered in a face-centered-cubic lattice, and used in cutting tools and jewelry
Electromagnetic waves	traveling waves of energy that compose light, x-rays, radio-waves, etc.
Epitaxy	growth of a thin film that preserves the crystal structure of the underlying film or substrate
Etching	removing material with chemical or physical atomic scale process
Graphite	carbon crystallographically ordered in hexagonal sheets, and used in pencils and lubricants
Inhomogeneity	varying with translation (position)
Isotropic	having properties that are identical in all directions
Lithography	process where patterns are produced in a sensitive layer by irradiating the layer through a mask with the desired pattern
Magnesium fluoride	ceramic (MgF_2) that is used in thin-film or single-crystal form in many optical systems
Morphology	microstructural shape, or physical arrangement of atoms in a material
Nanometers (nm)	one billionth of a meter, 10^{-9} m
Nematic liquid crystal	optical material composed of rod-shaped organic molecules that usually exhibits uniaxial optical properties
Pitch	physical period for one full rotation of a helical structure
Polarization	orientation and phase of the electric and magnetic fields in a traveling electromagnetic wave; can be linear, circular, or generally, elliptical
Refractive index	the ratio of the speed of light in vacuum to the speed of light in a material
Substrate	object that a thin film is deposited onto, often polished glass or semiconductors
Ultraviolet (UV)	electromagnetic radiation (light) with wavelength shorter than visible light and longer than X-rays
Uniaxial	having optical response characteristics that are anisotropic, but are rotationally symmetric about one axis
Vacuum	absence of matter, typically produced inside sealed glass or steel chambers by removing air with pumps
Vapor flux	moving gas-like collection of material, emitted from a source such as an evaporator or effusion cell, that will condense to form a thin film

See also

Coherence: Overview. **Coherent Lightwave Systems**. **Fiber and Guided Wave Optics:** Overview. **Imaging:** Imaging Through Scattering Media. **Information processing:** All-optical Multiplexing/Demultiplexing; Coherent Analogue Optical Processors; Optical Neural Networks. **Instrumentation:** Ellipsometry. **Optical Processing Systems**. **Optical Coatings:** Anti-Counterfeiting and Decorative Coatings; Optical Black Surfaces; Thin-Film Optical Coatings. **Optical Materials:** Measurement of Optical Properties of Solids. **Polarization:** Introduction; Matrix Analysis. **Scattering:** Scattering from Surfaces and Thin Films; Scattering Theory. **Semiconductor Materials:** Band Structure Engineering. **Terahertz Technology:** Coherent Terahertz Sources.

Further Reading

Ball P (1997) *Made to Measure*. Princeton, NJ: Princeton University Press.

Barabasi A-L and Stanley HE (1982) *Fractal Concepts in Surface Growth*. Cambridge, UK: Cambridge University Press.

Collins PJ (1990) *Liquid Crystals*. Princeton, NJ: Princeton University Press.

de Gennes PG and Prost J (1993) *The Physics of Liquid Crystals*. Oxford, UK: Clarendon Press.

Hodgkinson IJ and Wu Q-H (1997) *Birefringent Thin Films and Polarizing Elements*. Singapore: World Scientific Publishing.

Hodgkinson IJ, Lakhtakia A and Wu Q-H (2000) Experimental realization of sculptured-thin-film polarization-discriminatory light-handedness inverters. *Optical Engineering* 39: 2831–2834.

Hodgkinson IJ, Wu Q-H, Thorn KE, *et al.* (2000) Spacerless circular-polarization spectral-hole fitters using chiral sculptured thin films: theory and experiment. *Optics Communications* 184: 57–66.

Lakhtakia A, Messier R, Brett MJ and Robbie K (1996) Sculptured thin films (STFs) for optical, chemical and biological applications. *Innovations in Materials Research* 2: 165–176.

Liu F, Umlor MT, Shen L, *et al.* (1999) The growth of nanoscale structured iron films by glancing angle deposition. *Journal of Applied Physics* 85: 5486–5488.

Messier R, Venugopal VC and Sunal PD (2000) Origin and evolution of sculptured thin films. *Journal of Vacuum Science and Technology* 18: 1538–1545.

Ohring M (1992) *The Materials Science of Thin Films*. New York: Academic Press.

Robbie K, Brett MJ and Lakhtakia A (1996) Chiral sculptured thin films. *Nature* 384: 616.

Singh ON and Lakhtakia A (eds) (2000) *Electromagnetic Fields in Unconventional Materials and Structures*. New York: John Wiley.

Sit JC, Broer DJ and Brett MJ (2000) Alignment and switching of nematic liquid crystals embedded in porous chiral thin films. *Liquid Crystals* 27: 387–391.

Toader O and John S (2001) Proposed square spiral microfabrication architecture for large three-dimensional photonic band gap crystals. *Science* 292: 1133–1135.

Wang L and Clancy P (2001) Kinetic Monte Carlo simulation of the growth of polycrystalline Cu films. *Surface Science* 473: 25–38.

Smart Optical Materials

P M Martin, Pacific Northwest National Laboratory, Richland, WA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Smart optical materials are those materials whose optical properties (transmittance, reflectance, absorptance, emittance, fluorescence) can be controlled by external stimuli; either by an applied electric field or voltage, mechanical stress, incident light or electromagnetic field intensity, temperature variation, or time duration. This includes electrochromic, photochromic, thermochromic, nonlinear optical, light emitting, fluorescing, emissive, and piezochromic materials. These materials can be in various states, such as solids, gels, composites, and mesophases. Electrochromic and photochromic materials, however, are used most extensively.

The mechanism for adaptive control of light must be reversible, have a well-defined response time, and have an activation threshold, otherwise, the material is considered passive. The primary mechanisms that can cause variations in optical properties are crystal phase transitions, nonlinear effects in the polarizability, optical energy band transitions and band bending, compositional changes, defect and color center formation, and lattice strain. Crystal phase transitions change the bonding morphology of a material, which affects the energy bandgap. Most optical materials have several possible crystalline phases, depending on temperature. The optical properties of the anatase and rutile phases of TiO_2, for example, are very different, and result from different energy bandstructures (recall that the energy bandstructure depends on lattice spacing and crystalline plane orientation). The crystalline phase of several thermochromic materials changes from the martensite structure to the austenite structure with change in temperature, and an associated change in optical properties. Compositional changes are caused by chemical or physical reactions in the material, such as a metal ion bonding to an atom in the lattice, or loss of an atom in the lattice. The reaction can be caused by an energetic photon or ion driven by an electric field; electrochromic materials fit into this category. If their energy is high enough, incident photons can create defect states in the bandgap of materials, which in turn form color centers that absorb light. Modulated electric and electromagnetic fields can cause changes in the linear and nonlinear polarizability of the atoms in a lattice. Nonlinear optical and piezochromic effects are based on this mechanism.

Photochromic Glass

Photochromic glass is one of the most widely used smart optical materials, and definitely the most widely marketed. It has been around for over 30 years with hundreds of associated patents. The transmission of photochromic materials adjusts to the light level of their environment, but these materials give no output information for data processing or communications. They do, however, require no other external control mechanism other than the intensity of light (characteristics of a true smart material). Photochromic materials darken reversibly when exposed to sunlight or a specific optical (UV–LWIR) wavelength band. Absorption in specific narrow or broad wavelength bands is due to the formation of color centers, or point defect states in the lattice. The kinetics of darkening has been described by a two-color center model with two rate constants. Energy (usually in the ultraviolet wavelength range) of the incident photons, ionizes atoms in the glass, thus forming ‘F’ centers. F centers are highly absorptive and the absorption generally occurs in a broad wavelength band. The earliest photochromic glass developed utilized a dispersed and sensitized colloidal silver halide as the darkening agent. Copper oxide was a typical sensitizer and AgCl is also a fairly common impurity used in PC glass. This material darkens by the equation

$$\mathrm{Ag^I} + \mathrm{Cu^I} \rightleftharpoons \mathrm{Ag^0} + \mathrm{Cu^{II}} \qquad [1]$$

This type of reaction is fairly typical of these materials. The defect state is created in the silver and copper, creating optical absorption. Photochromic glass can darken down to 22% transmittance in sunlight and become nearly clear indoors, transmitting 85% of visible light. Figure 1 shows the transmittance of the darkened and normal states. Temperature, however, does have an effect on the lenses. The glass will not darken completely at temperatures above 32 °C, and at cold temperatures the lens will get extremely dark in a shorter period of time. Recently, $BaO–R_2O_3–SiO_2$ (R = Al, B, La) nanoparticles have been introduced to enhance photochromicity.

Thermochromic Materials

Thermochromic materials darken and lose transmittance when heated above a critical temperature. They can be as sophisticated as a multilayer thin film optical interference coating or as mundane as a cereal bowl that changes color when heated.

Thermochromic products are used to darken windows of automobiles, for example, when the internal temperature gets above a specific value. The conductivity of these materials also decreases with darkening in many cases because additional electrons are excited into the conduction band due to the increased number of defect states in the bandgap. The most common inorganic thermochromic thin film materials are transition metal oxides, and among these, the vanadium oxide family (V_2O_5, VO_2) shows the best performance. Figure 2 shows the transmittance of a VO_2 thin film before and after heating. Here T_t is the transition temperature. When heated above T_t, the film darkens. Stoichiometric VO_2 darkens at a temperature near $T_t = 68\,°C$ due to a semiconductor to metal transition. The resistivity of VO_2 decreases by three orders of magnitude and the emittance can be increased up to 90% at the transition temperature. This material has promising applications for thermal control of space structures, where surfaces are exposed to intense solar radiation and periods of darkness at very low temperatures. The transition temperature can be engineered (increased or decreased) by doping with metals such as W and Ta.

Loss of transmission can be achieved by absorption or increased reflectance. Reflectance increases with heating are observed in substoichiometric WO_3 films and absorptance increases in stoichiometric films. The reflectance increase is attributed to a decrease in W/O ratio and an associated increase in the refractive index (n) and extinction coefficient (k).

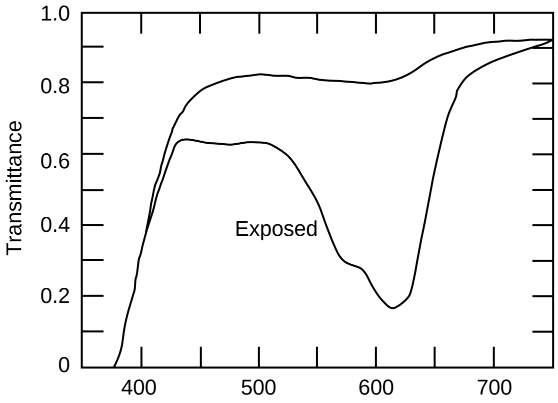

Figure 1 Transmittance of photochromic glass in darkened and normal states. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

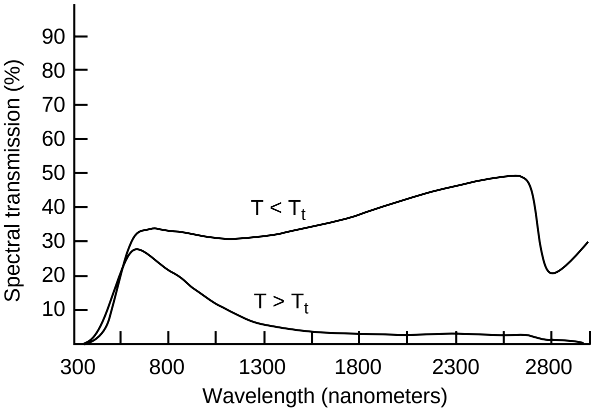

Figure 2 Change in transmittance of a thermochromic VO_2 film with heating. $T < T_t$ for normal transmission and $T > T_t$ when heated. (Reproduced with permission from Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.)

Electrochromic Materials

Electrochromic materials, also called switchable materials, are usually used in the form of thin film devices. Electrochromic windows, or smart windows, have finally reached maturity and promise to be the next major advance in energy-efficient window technology. These devices are used in architectural glazings, automotive mirrors, rear view mirrors, sunroofs, sunglasses, and other high-end applications. Electrochromic devices are well suited for both solar and thermal control. A window can be controllably darkened during periods of bright sunlight and switched back to transparency during the evening or during cloudy periods. Windows with multilayer thin film electrochromic coatings change color and darken with an applied voltage, and bleach with a reverse bias. The magnitude of the applied voltage determines the degree of coloring and bleaching. In addition to a change in visible transmission, the NIR transmission and reflectance, and the associated thermal properties, can also be controlled. The major advantages of electrochromic devices are: (1) power is used only during switching; (2) switching voltages are small (1–5 V); (3) reflectance and transmittance is always specular; (4) gray scale variations are possible; (5) polarization is minimal (reduces birefringence and distortion); and (6) the memory is adjustable in many cases.

The electrochromic effect occurs in inorganic compounds by dual injection (cathodic) or ejection (anodic) of ions (M) and electrons (e^-). A typical reaction for a cathodic coloring material is

$$WO_3(\text{colorless}) + yM^+ + ye^- \rightleftharpoons MyWO_3(\text{blue}) \quad [2]$$

where M is H^+, Li^+, Na^+, Ag^+, etc.

A typical anodic reaction is

$$Ir(OH)_x(\text{colorless}) \rightleftharpoons IrO_x(\text{black}) + XH^+ + xe^- \quad [3]$$

Movement to the right creates color or darkening, and movement to the left results in bleaching. Figure 3 shows the spectral data for the switching of a tungsten oxide film over the solar spectrum; the spectral response of the coating in the bleached and colored (or 'switched') states, and Figure 4 shows the bleached and darkened states of an actual electrochromic window.

Figure 5 shows the design of a simple electrochromic coating. The basic electrochromic coating consists of a transparent top electrode (usually indium tin oxide – ITO), which also injects electrons into the electrochromic layer beneath. By far the most studied and most promising electrochromic material is tungsten oxide (WO_3), which can have either an amorphous or crystalline microstructure. WO_3 has a high coloration efficiency and capacity. Other inorganic materials that are presently being developed for solar control applications are various forms of WO_3, NiO, $WMoO_3$, MoO_3, and IrO_x. Many of these are doped with a conduction ion such as Li^+ or H^+. The ion conductor layer is used to transport protons (H^+), which are supplied by the ion storage or counter-electrode layer, into the electrochromic material. The most promising ion conductors are certain immobile solvent polymer systems, ionic glasses, and open-channel metal oxide structures such as perovskites. Oxides of Ta and Nb also show promise. The bottom layer is another transparent conducting coating. It is possible to inject both electrons and protons into the electrochromic material, thereby inducing a strong absorption band in a given region, e.g., the visible spectrum, that results in a color change. In an

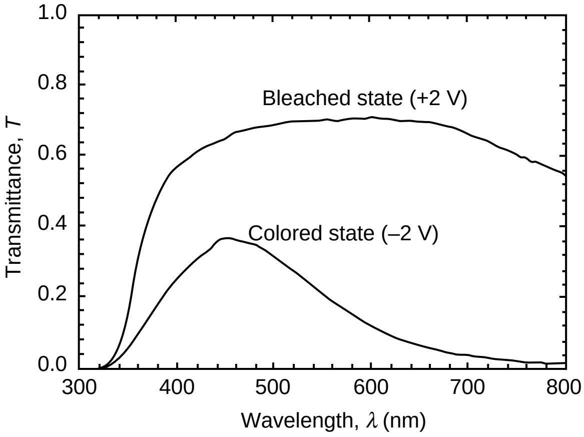

Figure 3 Spectral switching of an electrochromic window for darkened (−2V bias) and bleached (2V bias) states. (Reproduced from Kre J, Topic M, Smole F, *et al.* (2002) *Solar Energy Materials and Solar Cells* 71: 387–395, with permission from Elsevier.)

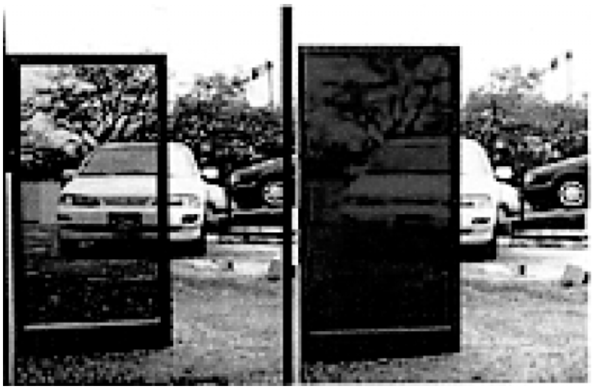

Figure 4 Bleached (left) and darkened (right) states of an electrochromic window.

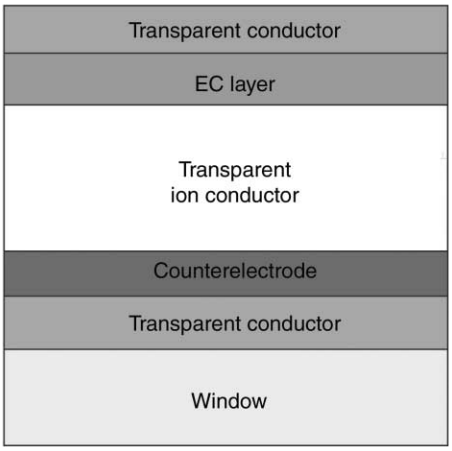

Figure 5 Basic layer design of electrochromic coating for use on a window.

Table 1 Electrochromic device structures

Electrochromic layer	*Ion conductor or electrolyte*	*Ion storage*
a-Li_xWO_3	$LiClO_4$ + PC	Redox Cple/NiO
a-WO_3	Ta_2O_5	$Ir_xSn_yO_2$:F
a-Li_xWO_3	$LiClO_4$ + PC	Prussian blue
a-H_xWO_3	Polymer	Polyaniline
a-H_xWO_3	SiO_2/metal	WO_3
Viologen	PMMA + organic	none
a-WO_3	Ta_2O_5	a-IrO_2
c-WO_3	Li–B–SiO Glass	IC/$Li_xV_2O_5$
a-WO_3	Li–PEO	CeO_x
a-Li_xWO_3	PPG–$LiClO_4$–MMA	$Li_yV_2O_5$
a-H_xWO_3	Ta_2O_5	NiO
a-WO_3	Li–PEO	NiO
NiO	a-PEO copolymer	Nb_2O_5, WO_3:Mo
a-WO_3	a-PEO copolymer	Polymer
Polyaniline	HCl	–

a, amorphous; c, crystalline; IC, ion conductor; MMA, methyl methacrylate; PC, polycarbonate; PEO, poly (ethylene oxide); PMMA, poly (methyl methacrylate); PPG, poly (propylene glycol).

electrochromic device, an externally applied electric field is used to control the injection process (voltage is applied across the two transparent electrodes). The coloration remains for some time, even after the external field is removed. The system returns to the initial state upon reversing the polarity of the external field. It is interesting to note that electrochromic windows being developed by NREL now employ solar cells to facilitate optical switching.

There are two major categories of electrochromic materials: transition metal oxides (including intercalated compounds) and organic compounds (including polymers). Table 1 shows a few of the most common electrochromic materials. Organic electrochromic materials are based on the viologens, anthraquinones, diphthalocyanines, and tetrathiafulvalenues. With organic compounds, coloration is achieved by an oxidation–reduction reaction, which may be coupled to a chemical reaction. The viologens are the most studied of the organic electrochromics. Originally, organic electrochromics tended to suffer from problems with secondary reactions during switching, but recently more stable organic systems have been developed. However, their reliability in harsh environments is questionable, and protective encapsulation may be required.

Electrochromic nickel oxide (NiO) is also being developed. NiO has been combined with manganese oxide, cobalt oxide, and niobium oxide in all alkaline devices.

Nonlinear Optical Materials

Nonlinear optical materials have evolved slowly over the past 20 years. At relatively low light intensities, the properties of most optical materials are independent of intensity of illumination. There is no interaction between the light waves in the medium. However, if the intensity of illumination gets high enough, nonlinear optical effects can occur and the optical properties become intensity dependent. When an electromagnetic field oscillating at optical frequencies ($\sim 10^{13}$–$\sim 10^{17}$ Hz) is applied to the lattice of a dielectric material, the electrical charges in the lattice become polarized and form electric dipoles. The positive ion cores do not move significantly, but the electrons are displaced in a harmonic motion. The polarization $\mathbf{P}$ is defined as $-Ne\mathbf{X}$, where N is the number of electric dipoles, $\mathbf{X}$ is the displacement of the electrons, and e is the charge on an electron. Here $\mathbf{X}$ is a complex number. Now, $\mathbf{P}$ is related to the electric field E by $\mathbf{P} = \varepsilon_0 \chi E$, where χ is the susceptibility and ε_0 is the permitivity of free space.

The linear dielectric constant is $\mathrm{Re}(1 + \mathbf{X})^{1/2}$, which is just the index of refraction. The imaginary part causes optical absorption.

When the intensity of illumination is high, the elastic limit of the electron's displacement is exceeded and the harmonic restoring force becomes nonlinear in $\mathbf{X}$, which adds quadratic and higher order terms to the expression for $\mathbf{X}$ and $\mathbf{P}$. In addition to the fundamental resonance, the dipoles now also oscillate at twice and thrice their fundamental resonant frequency (second and third harmonics). The polarization then is also nonlinear in displacement and can be expressed by the well-known relation

$$\mathrm{P} = \varepsilon_0(\chi^{(1)}\mathrm{E} + \chi^{(2)}\mathrm{E}^2 + \chi^{(3)}\mathrm{E}^3 + \ldots) \qquad [4]$$

where $\chi^{(2)}$ and $\chi^{(3)}$... are called the nonlinear susceptibilities. The second-order polarization $P^{(2)}$ causes optical mixing phenomena, involving sums and difference frequencies. When mixed with a DC field, the refractive index of the $\chi^{(2)}$ material can be controllably varied by varying the DC field. This is the well-known Pockels effect and is used in optical modulators. Figure 6 shows one method that is used to modulate the refractive index by the photorefractive effect. Here two laser beams (A_1 and A_2) are mixed in the $\chi^{(2)}$ material with the resulting intensity $I = I_o(1 + m\cos(kx))$. The material is polarized in phase with the mixed beams $\rho = \rho_o(1 + m\cos(kx))$, which results in a modulated electric field E_{sc}. The refractive index (Δn in the figure) is modulated in phase with the mixing signal. The $\chi^{(2)}$ materials are also used for frequency doubling.

Barium titanate ($BaTiO_3$), lithium niobate ($LiNbO_3$), KH_2PO_3 (KDP), lithium tantalite ($LiTaO_3$) and potassium niobate ($KNbO_3$) are $\chi^{(2)}$ optical switching and frequency doubling materials that have been studied for decades. Note that because they can be easily polarized, these are also excellent piezoelectric/ferroelectric smart materials.

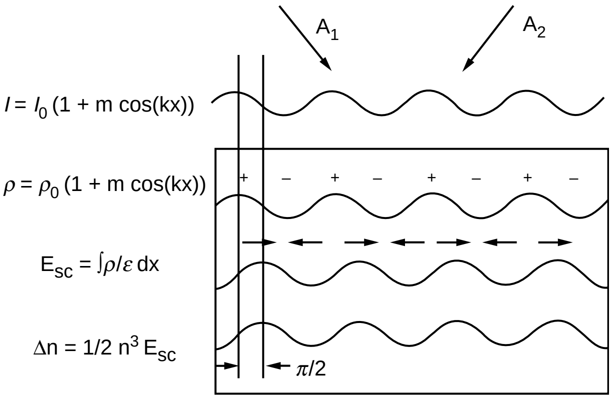

Figure 6 Modulation of the refractive index of a nonlinear optical material with incident laser modulation. From top to bottom: modulated laser intensity I, density ρ, electric field E_{sc}, refractive index Δn, and phase angle $\pi/2$. (Reproduced with permission from Bass M (ed.) (1995) *Handbook of Optics II*. New York: McGraw-Hill.)

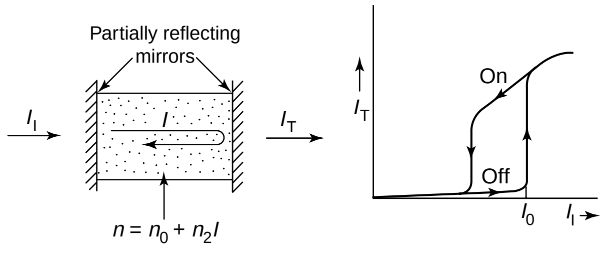

Figure 7 Optical switching behavior of a Fabry–Perot filter. I_I and I_T are the incident and transmitted intensities respectively. (Reproduced with permission from Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.)

The $\chi^{(3)}$ effects are sought after for applications such as laser eye protection, optical sensor protection, optical switching, optical limiting, and signal processing. In this case the refractive index (real and imaginary) depends on the intensity of the incident radiation. Figure 7 shows how this nonlinear effect is used in an optical switching device. It is all based on the change in the optical path length of a Fabry–Perot filter. At normal intensities, the refractive index of the spacer layer is n_0 and the filter is in the 'off' state. The refractive index becomes $n_0 + n_2 I$ and the optical thickness (nt) of the spacer layer changes at high illumination intensities, which in turn shifts the resonant wavelength (frequency) of the Fabry–Perot filter and creates the optical switch ('on' state).

There are a large number of $\chi^{(3)}$ and $\chi^{(2)}$ organic materials, including dyes, dimethylamino nitrostilbene, methyl nitroanaline, poly-BCMU, polydiacetylenes, and urea. Because they are essentially chains, many organic molecules can be easily polarized and possess higher order susceptibilities. The polarizability of organic materials is often enhanced by the mobility of delocalized π electrons in C–C bonds in aromatic rings. Hybrid nonlinear materials have even been formed by combining organic and inorganic components such as metallophthalocyanines which display very strong excited state absorptions. Most of these materials are deposited by spin casting.

Emissive Materials

Two of the newest applications of smart optical materials are emissive displays and electronic ink. This encompasses light-emitting polymers and small molecules, light-emitting diodes, and quantum cascade lasers. Organic light-emitting device (OLED) technology is leading the revolution in the flat panel display industry. This device was first invented by Tang and Van Slike and first reported in *Applied Physics Letters* in 1987. Figure 8a shows the layer structure of the OLED and Figure 8b shows a blue OLED pixel. The device is deposited onto an indium tin oxide (ITO)-coated glass or plastic substrate, with the ITO acting as the anode. The active layers can be either light-emitting polymers or small molecules. About 500 Å of a hole transport material (hole transport layer, HTL), here a tertiary amine, is deposited over the ITO. An aluminum trisquinalate (Alq), which transports electrons (electron transport layer, ETL) is deposited next. The device is topped off with

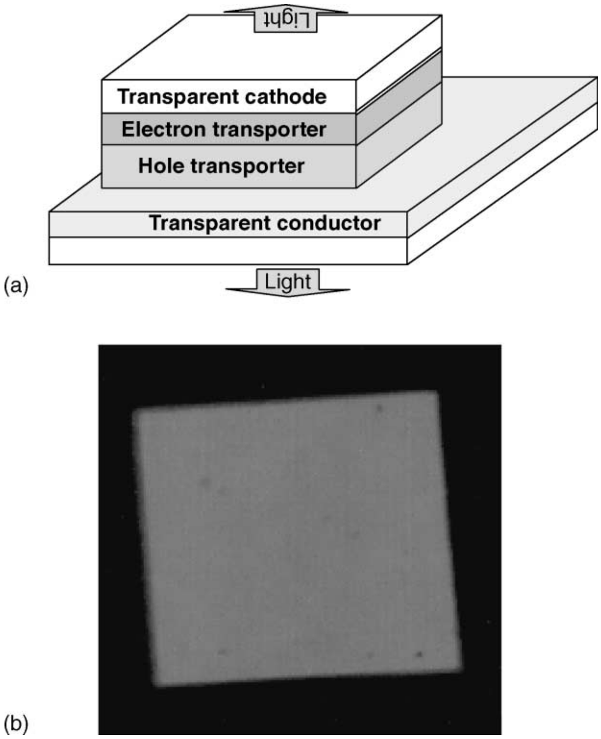

Figure 8 (a) Layer diagram of an OLED. (b) Blue OLED pixel.

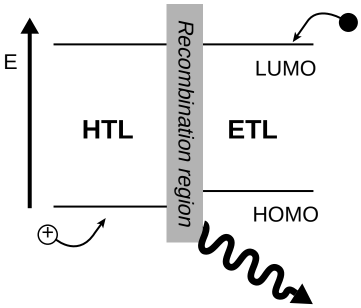

Figure 9 Simple energy level diagram of an OLED showing the relationship between the ETL and LUMO, and HTL and HOMO.

a low work function cathode, usually Ag:Mg or Ca. The entire device thickness is about 1000 Å. Even a dust particle can ruin the operation of an OLED by creating inactive regions or shorting out the electrodes.

The OLED emits light in a blue, green, or red wavelength band when activated by an electric current, i.e., electron and hole injection. The device functions by injecting holes from the ITO into the HTL. As shown in the simplified energy band diagram (Figure 9) for an OLED, the hole moves into the HOMO (highest occupied molecular orbital), which loosely corresponds to the valence band of a semiconductor. Electrons are injected from the cathode into the ETL, where it is transported into the LUMO (lowest unoccupied molecular orbital), which loosely corresponds to the conduction band of a semiconductor. When a voltage is applied between the anode and cathode, an electron flows and the holes and electrons combine in the ALQ layer and form excitons. Light is emitted when the excitons decay back a lower energy state. The color of the emitted light (green, red, or blue) can be changed by changing the HTL and ETL material composition. Light is emitted by fluorescence processes, from both singlet and triplet states if the material chemistry is correct, which results in 100% internal quantum efficiency. Twenty percent external quantum efficiency results from using all available exciton states. Photons can be lost through absorption in the layers, scattering, reflection between layers, and waveguide modes. These factors must be minimized to maximize the external quantum efficiency.

OLED video displays now have outputs near 100 cd/m (20 lm/W), which is about twice the normal video brightness. Red, green, and blue OLEDs now emit with very high efficiencies. For comparison, a 100 W light bulb emits at 17 lm/W and the best fluorescent light emits at 80 lm/W. Today, OLEDs are already sitting at intensities of 50–60 lm/W, and could easily push 150 lm/W in the near future. Another advantage of OLED displays is their wide viewing angle, up to 160°.

There are several types of OLED display configurations. Figure 10 shows a passive matrix display geometry and placement of the OLED layers. Here the OLED is sandwiched between crossed cathode and anode contacts. By injecting current between sets of crossed anode and cathode rows, the OLED will light up, each contact point forming a pixel. Displays can typically have as many as 3 million pixels. OLED displays seem brighter and livelier to the eye because the colors are additive, rather than subtractive as with LCD displays. For a full-color passive matrix display green, red, and blue pixels must be located very close together. Sony has developed an active matrix 13-inch diagonal display with 800 × 600 pixels (~150 000 functioning OLEDs).

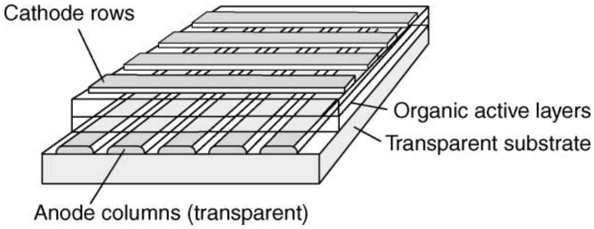

Figure 10 Geometry of a passive matrix display.

Among the more promising pixel geometries is the SOLED, the stacked OLED, where, red, green, and blue pixels are stacked on top of each other. It is mandatory that the pixels and the electrical contacts be transparent for this design, or light will not be transmitted through each layer. Very thin Mg contacts are used with a top sputtered ITO cathode. The emitted color is simply changed by applying current across the pixels, with white being obtained when all pixels are activated.

The future of OLED displays is plastics, or the flexible OLED (FOLED). Plastic substrates are lightweight, low cost, and rugged. However, they have an inherent problem of being very permeable to oxygen and water vapor, both of which severely degrade performance of the OLED by reacting with the anode layer. For displays to be practical, they must have lifetimes in excess of 10 000 hr and preferably 20 000 hr. Barrier coatings are now being developed for plastics to prevent oxygen and water ingress and extend the life of the OLED to this goal.

Electronic ink is a display medium that can be formatted using a passive matrix. It consists of millions of tiny negatively charged microcapsules, about 100 microns in diameter, floating inside a liquid medium. Each microcapsule is filled with hundreds of pigmented chips. One half of each microcapsule is white and the other half is dyed blue. When an electric charge is applied to the microcapsules, the chips are drawn either to the bottom or top of the microsphere. The viewer sees dark areas when they are drawn to the top. In another version being developed by Xerox Corp., half the chip is white and half the chip is dark. When an electric charge is placed on the chip, the dark side is oriented up and the white side down, creating dark regions where charge is placed. This can be easily erased and reformatted into text, symbols, signs, and pictures.

Future smart optical materials will simulate biological systems such as an insect's eye and will employ quantum structures to achieve optical switching, emission, and control of optical properties.

List of Units and Nomenclature

Absorptance (A)	The amount of light absorbed by an optical element divided by the incident intensity. $A = 1 - T - R$.
Active matrix display	A display geometry in which the voltage to the pixel is controlled individually by a transistor.
Candela (cd)	Formerly known as a candle. The luminous intensity equal to 1/60 of the luminous intensity per square meter of a black body radiator operating at the temperature of freezing platinum.
Exciton	An electron–hole pair, usually generated by a photon.
FOLED	Organic light emitting device on a flexible plastic substrate.
LCD	Liquid crystal diode.
Lumen (lm)	A unit of luminous flux equal to the luminous flux emitted in a solid angle of one steradian for a point source having an intensity of one candela.
NIR	The near-infrared spectral region, consisting of wavelengths between 700 and 2500 nm.
Passive matrix display	A display geometry consisting of a grid of horizontal and vertical wires. At the intersection of each grid is an active element (LCD or OLED) which constitutes a single pixel, either letting light through or blocking it.
Piezoelectric	Capable of generating a voltage when a force or stress is applied.
Polarization (coulomb/m^3)	A vector quantity equal to the electric dipole moment per unit volume.
Quantum efficiency	The number of electrons, holes, or electron–hole pairs generated per incident photon at a given optical wavelength.
Reflectance (R)	The intensity of light reflected from an optical element divided by the incident intensity. $R = 1 - T - A$.
Susceptibility (no units)	A scalar quantity equal to the ratio of the electric polarization of a material to the electric field strength.
Transmission (T)	The intensity of light transmitted through an optical element divided by the incident intensity. Note that transmission + reflection + absorption = 1 ($T + R + A = 1$).
UV	The ultraviolet spectral region consisting of wavelengths between 200 and 400 nm.
Visible	The spectral region consisting of wavelengths between 400 and 700 nm.
Wavelength (μm or nm)	The length of one oscillation of a light wave.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Optical Coatings:** Thin-Film Optical Coatings. **Optical Materials:** Optical Glasses.

Further Reading

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge University Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 1. Orlando, FL: Academic Press.

Chemla DS and Zyss J (1987) *Nonlinear Optical Properties of Organic Molecules and Crystals*, vol. 2. Orlando, FL: Academic Press.

Clark DT (1994) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Crowson A (1996) Smart Materials Technology and Biomimetics, *SPIE Proceedings*, vol. 2716. Bellingham, CA: SPIE.

Gobin P-F and Friend CM (2000) Smart Structures and Materials, *SPIE Proceedings*, vol. 4073. Bellingham, CA: SPIE.

Janocha H (1999) *Adaptronics and Smart Structures, Basics, Materials, Designs, and Applications*. Berlin: Springer.

Jorgenson GV and Lee JC (1985) Optical Materials for Energy Efficiency and Solar Energy Conversion IV, *SPIE Proceedings*, vol. 562. Bellingham, CA: SPIE.

Goff AH-L, Granqvist C-G and Lampert CM (1992) Optical Materials Technology for Energy Efficiency and Solar Energy Conversion XI: Chromogenics for Smart Windows, *SPIE Proceedings*, vol. 1728. Bellingham, CA: SPIE.

McDonuch A, Gardiner APT and McEwen KS (1994) In: Culshaw B (ed.) Second European Conference on Smart Structures and Materials, *SPIE Proceedings*, vol. 2361. Bellingham, CA: SPIE.

Zerbi G (ed.) (1993) *Organic Materials for Photonics*, European Materials Research Society Monographs, V6. Amsterdam: North-Holland.

Sol-Gel Materials

Lisa C Klein, Rutgers the State University of New Jersey, Piscataway, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Sol-gel optics represents a methodology for fabricating optical materials in fiber, thin film, and bulk geometries. Defined broadly, sol-gel optics encompasses a variety of optical devices that contain sol-gel processed components. The optical device may be based wholly on sol-gel processing or incorporate a sol-gel material, along with conventional optical materials. Principally, the optical devices prepared by sol-gel processing or assembled with sol-gel materials are traditional optics or, in some cases, opto-electronics. Among the current sol-gel optical materials, by far the most common are oxides. Further, the majority of sol-gel optics is for transparent components, which typically means glasses, and, consequently, silicates.

In fact, most sol-gel materials have been developed to duplicate conventional silica optical materials. While glass melting technology is highly advanced for making quality lenses, an equivalent material can be made using a sol-gel route. Similarly, ultralow-loss optical fiber can be made by chemical vapor deposition (CVD). Now sol-gel processing can be used to prepare comparable fibers for telecommunications. In cases where sol-gel processing produces an exact duplicate of optical materials produced other ways, the perceived advantage is that sol-gel precursors have a higher purity and better level of mixing than powder raw materials. Perhaps, a more significant advantage is that sol-gel formulations can be designed to promote a particular geometry in the final material, essentially making near-net shape materials.

While sol-gel processing is not designed to produce new optical phenomena, it does offer new ways to assemble materials to produce new optical designs, such as filters, Bragg gratings, amplifiers, and beam splitters. One noted case of sol-gel materials in nonlinear optics is the case where semiconductor particles are trapped in nanometer-sized pores. This leads to so-called quantum dots. The microporosity of the gel, no doubt, assists in this effect, but it is not the gel itself that has produced the optical effect. The optical effect is derived from the nanometer size of the particles. The gel serves as a transparent host that restricts the size of the particles and provides for their uniform distribution.

Sol-Gel Chemistry of Oxide Optical Materials

The sol-gel process for oxides is the name given to any one of a number of processes involving a solution or sol that undergoes a sol-gel transition. Solution refers to a single-phase liquid, while a sol is a kinetically stabilized suspension of colloidal particles. At the transition, the solution or sol becomes a rigid, porous mass by destabilization, precipitation, or supersaturation. Once the solution or sol goes through a sol-gel transition, it is a rigid two-phase system of solid and solvent filled pores. The sol-gel transition is not reversible.

In using a sol-gel process, it is important to be familiar with the reagents, the intermediates and the products. The most important reagent in this case is a hydrolyzable organometallic which is called a metal alkoxide, $M(OR)_x$ where (OR) is an alkoxy group. Alkoxides are the precursors for SiO_2, GeO_2 or B_2O_3, among others. Hydrolysis occurs when the metal alkoxide and water are mixed in a mutual solvent.

The first step is choosing the right reagents. Since most of the glasses are silicates, we start with the silica precursor. Of the available silicon alkoxides, tetraethylorthosilicate (TEOS) is used most often, because it reacts slowly with water and comes to equilibrium as a complex silanol. The colorless liquid $Si(OC_2H_5)_4$ has a density of about 0.9 g/cm^3, is easy to handle safely and is extremely pure when distilled. Most metal alkoxides are prepared by the reaction of halogenated species or pure metals with alcohol.

The two other ingredients in a typical formulation are alcohol and distilled water. In the case of TEOS, ethanol serves as the mutual solvent for TEOS and water. Once TEOS is dissolved in ethanol to make it soluble in water, the chemical reactions hydrolyzation and polymerization begin. The chemical reactions are approximately:

Hydrolyzation

$$-\underset{|}{\overset{|}{\mathrm{Si}}}-\mathrm{O}-\mathrm{C_2H_5}+\mathrm{H_2O}\rightarrow -\underset{|}{\overset{|}{\mathrm{Si}}}-\mathrm{OH}+\mathrm{C_2H_5OH} \qquad [1]$$

Polymerization

$$-\underset{|}{\overset{|}{\mathrm{Si}}}-\mathrm{O}-\mathrm{C_2H_5}+\mathrm{Si}-\mathrm{OH}\rightarrow -\underset{|}{\overset{|}{\mathrm{Si}}}-\mathrm{O}-\underset{|}{\overset{|}{\mathrm{Si}}}+\mathrm{C_2H_5OH} \qquad [2]$$

An acid is usually added to control the rates of these reactions. The temperature is often ambient, and mixing is continued for between 1 and 3 hours. Complete hydrolysis does not occur. Instead, condensation occurs between silanol and ethoxy groups to give bridging oxygens or siloxane groups. The intermediates that remain soluble in the alcohol–water medium are silanols, ethoxysilanols, and polysiloxanes.

The equipment needed for laboratory-scale production consists of a narrow-necked glass flask with magnetic stir bar on a stirring hot plate. The reactants are added volumetrically to the flask in the order: solvent, TEOS, water, and acid. When everything is dissolved and the solution is clear, the hot plate may be turned on to accelerate reactions or a reflux condenser may be attached.

The parameters that influence the chemical reactions are temperature, pH, amount of water, solvent, and precursor. More complex compositions are easier to mix in longer chain alcohols, and longer chain organic groups give slower reaction rates. Low pH favors a polymerization scheme that gives linear molecules and the solution remains clear. High pH favors cluster growth and the solution may become cloudy. At high pH or high water, solutions produce cross-linked polymers or branched clusters. A small volume fraction of branched clusters restricts flow, when the same volume fraction of linear polymers does not. At a higher volume fraction, the linear polymers become tangled, and then low water solutions will gel. The result of these reactions under all conditions is an increase in the molecular weight of the oxide polymer. Eventually, the solution reacts to a point where the molecular structure is no longer reversible, and the sol-gel transition has been reached.

Shapes, Compositions and Precursor Concentrations

A sampling of compositions and precursors is given in Table 1. Some compositions such as SiO_2 and $TiO_2{-}SiO_2$ have been prepared in bulk, to duplicate materials made by melting for lenses and mirror blanks. Other compositions such as $B_2O_3{-}SiO_2$ have been prepared for thin-film applications. The distinguishing aspect between the solutions formulated for bulk objects and those formulated for thin films is the oxide concentration in the solution. An indication of the solution concentration is the molar ratio of water to alkoxide. The higher ratios are used when bulk objects are desired.

Table 2 gives a sampling of the geometries for sol-gel silica. For each geometry, an example of a shape is given, with an application for each shape and the approximate dimensions of the sol-gel material.

In terms of near-net shape processing, when the material goes through its sol-gel transition, it proceeds to dry and shrink. This phase is characterized either by isotropic shrinkage or anisotropic shrinkage. Monoliths fall into the first category with isotropic shrinkage following molding. Thin films on substrates fall into the second category with anisotropic shrinkage following dipping. Fibers also show

Table 1 Typical compositions of sol-gel optical materials

Oxides	*Precursors*	*Compositions in mol %*	*R*[*]
Single oxide			
SiO_2	$Si(OC_2H_5)_4$	100 SiO_2	16
Two oxides			
$B_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $B(OCH_3)_3$	80 SiO_2–20 B_2O_3	1.5
GeO_2-SiO_2	$Si(OCH_3)_4$ $Ge(OC_2H_5)_4$	90 SiO_2–10 GeO_2	4
TiO_2-SiO_2	$Si(OC_2H_5)_4$ $Ti(OC_3H_7)_4$	94 SiO_2–6 TiO_2	50
$P_2O_5-SiO_2$	$Si(OC_2H_5)_4$ H_3PO_4	90 SiO_2–10 P_2O_5	16
Three oxides			
$Li_2O-Al_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $Al(OC_4H_9)_3$ $LiNO_3$	82 SiO_2–3 Al_2O_3–15 Li_2O	8
$Na_2O-B_2O_3-SiO_2$	$Si(OC_2H_5)_4$ $B(OCH_3)_3$ $NaNO_3$	82 SiO_2–12.4 B_2O_3–5.6 Na_2O	10

[*] Approximate ratio of water to alkoxide, e.g., $R = 16$ means 16 moles of H_2O per mole of $Si(OC_2H_5)_4$.

Table 2 Geometries and approximate dimensions for selected applications of sol-gel silica

Geometry	*Shape*	*Dimensions*	*Application*
Monolith	Flat disk	10 cm diameter 1 cm thick	Lens
	Cylinder	0.5 cm diameter 2 cm long	Rigid host for chromophores
Thin film	Film on Si wafer	500 nm thick	Antireflection
	Film on Hollow tube	~10 000 nm thick	Solar collector panel
Fiber	Rod	0.8 cm diameter 25 cm long	Preform for drawing
	Filament	150 micron diameter/100 m length	Lightguide

anisotropic shrinkage, though it depends on whether or not the fiber is drawn from a preform many times its diameter or produced from a filament that is closer to its diameter.

Sol-Gel Monolith Processing

Monoliths formed from an alkoxide solution are achieved only in dilute solutions. Therefore, alkoxide gels are somewhat difficult to dry because of their small pores (<10 nm). To avoid cracking in alkoxide gels, different treatments have been tried. Aerogels are dried in an autoclave by hypercritical techniques. That is, the solvent is removed above its critical point. The resulting gel is about 10% dense and shows no shrinkage. Xerogels are dried by natural evaporation. Xerogels are 60% dense and have reductions 40–60% in volume.

Alternatively, monoliths can be formed from colloidal solutions. These particles coalesce at gelling. Colloidal gels are easily dried, and they are less likely to crack than alkoxide xerogels. While alkoxide gels have small pores, colloidal gels have larger pores or voids between particles. For the same reason, colloidal gels densify at higher temperatures than alkoxide xerogels. The higher temperature for colloidal gels may lead to undesirable crystallization.

By either route, a serious problem for sol-gel processing is the presence of water. Especially when monoliths are used for fiber drawing, residual OH ions have a deleterious effect on optical properties. That is why non-aqueous solvents, as well as fluorination or chlorination treatments, have been investigated to reduce the OH content.

Sol-Gel Fiber Processing

While monolithic pieces need high water content, fibers can be drawn out of low water content solutions. The sol-gel process allows one to bait and draw a string of gel about the same diameter as the desired fiber directly from the solution. This has been shown where the water to alkoxide ratio was controlled to give the property of spinnability. The time and manner in which the system loses fluidity depend on composition and catalyst. Low water solutions show a gradual increase in viscosity before losing fluidity, while high water solutions lose fluidity more abruptly from a lower viscosity. A problem with this method is the shape of the fiber, which tends not to be cylindrical.

Another approach to making near-net shape fibers from a sol-gel process involves coating a shell or concentric shells of sol-gel glass on a filament of textile fiber. This approach is referred to as the volatile host method, where a plastic filament is used as a template. Multiple dipping or adjustment of the viscosity can lead to different thicknesses. Upon exposure to atmospheric moisture, the films gel. After gelling, a rigid, continuous shell runs the length of the filament.

Once the shell has hardened, solvent and water can escape through interconnected pores. The shell shrinks in the radial dimension, but remains continuous along its axis. During heat treatment and desiccation, little if any densification occurs at temperatures well below the glass transition. The interconnected pore structure remains open until the volatile host is eliminated and undesirable gases are removed. This method is capable of producing optical fibers, but is more suitable for hollow fibers or ion exchange media.

Sol-Gel Thin Film Processing

Like sol-gel fibers, sol-gel thin films are prepared from low water content solutions. The process is exceedingly simple. A solution containing the desired oxide precursors is applied to a substrate by dipping or spinning.

At this time, the majority of sol-gel coatings are applied by dipping. This is an approach that allows the properties of the solution to control the deposition. A substrate is lowered into a vessel containing the solution. A meniscus develops at the contact of the liquid and the substrate. As the substrate is withdrawn, the meniscus generates a continuous film on the substrate. The process is able to apply a coating to the inside and outside of complex shapes simultaneously.

The equipment is inexpensive, especially in comparison to any deposition techniques that involve vacuum. Coatings can be applied to metals, plastics, and ceramics. Typically, the coatings are applied at room temperature, although most need to be calcined and densified with heating. Dip films are typically one micron, uniform over large areas and adherent. The time-to-gel is especially important when it comes to coatings because film formation, drying, and creation of pores must be rapid. Optimum film formulations correspond to those solutions that lose tackiness quickly.

In most cases, the effect of the viscosity is far greater than the surface tension. Two simple relationships are that the film thickness increases with increasing withdrawal rate and, for a given withdrawal speed, the film thickness increases with an increase in oxide content. The process is frequently a batch process, certainly on the laboratory scale. However, the batch process can be scaled up. Repeated dipping builds up a thicker film.

For one-side coating, a technique that is often used is spin coating. In this case, the substrate, usually a silicon wafer, is placed on a spinner, rotated at perhaps 200 rpm while solution is dripped on the center of the substrate. As with dip coatings, a film between 50 and 500 nm results.

After the Sol-Gel Transition

Having selected a shape and designed the formulations accordingly, there are several further steps common to monoliths, fibers, and films. First of all, the gels must be dried. Using the example of silica, TEOS is less than 30% by weight silica and the solution is even less. From the time the solution is applied to the time it gels, there is about 50% weight loss. Then, as the gel dries there is another reduction in weight by one-half. To go along with the weight loss there is about 70% volume reduction.

Films can be dried quickly in air because of the one thin dimension. It has been shown repeatedly that all shrinkage is taken up in the thin dimension and not in the plane of the substrate. Yet, the films remain adherent and continuous and maintain complete surface coverage. To go from a tacky film to a hard film may take only a few minutes.

Following drying and shrinkage comes the heat treatment. It would be inaccurate to say that glass has been formed at room temperature by simply reacting, gelling, and drying the solutions. In this state, the materials have many of the characteristics of the corresponding glass, but they are more or less porous. Water and solvent escape through interconnected pores that remain open at the surface until the gels are fired to temperatures well above 600°C. When needed to fulfill the requirements of an application, the hard gel is heated to various degrees of collapse.

The microporosity in silica xerogels is not removed entirely until 1000°C.

Applications of Sol-Gel Optical Coatings

Sol-gel processing has a special relevance to optical coatings. In most cases, the applications are substitutions of a sol-gel coating for a coating obtainable by other means. Most optical coatings achieve their effect by selective absorption and interference. By depositing 1–14 or more layers of dielectric thin films, using two or more different compositions, a variety of antireflection (AR), half-mirror, hot-mirror and cold-mirror coatings can be designed. An early design was for rear-view mirrors for cars, where the coatings consisted of titania-silica-titania interference filters that gave the effect of total reflectance and replaced conventional metallizing. Another early application was solar reflecting films for windows. These coatings consisted of Pd-containing titania films which show selective absorption.

A sampling of optical coatings is given in Table 3. The application, the mechanism, and the oxide composition are listed. Generally, these coatings can be prepared by other deposition techniques. Those applications which are specific to sol-gel coatings relate to the microporous nature of the sol-gel processed films. In Table 3, there are antireflective coatings of several types. Both borosilicates and titania silicates have been developed for this application. Changes in the composition, changes in the microstructure, and changes in the porosity can be used to grade the index of refraction. Antireflection in the ultraviolet range has been accomplished with thoria and hafnia films. Broadband antireflection has been used for laser optics. The sol-gel coating technique provides control over composition and deposition thickness. The layers, being quarter wavelength in thickness, require precise control.

Table 3 Thin-film applications

Optical application	*Mechanism*	*Typical composition*
Color filters	Selective Absorption	Cr_2O_3–SiO_2
Mirror	Interference	14-layer TiO_2: SiO_2
Antiglare	Absorption	CoO–SiO_2
Antireflection		
Narrow	Graded porosity	SiO_2
Broadband	Interference	Three-layer
	1/4 wavelength	SiO_2–TiO_2:ZrO_2: SiO_2
	Graded porosity	SiO_2
IR reflector	Interference	Three-layer TiO_2:Ag:TiO_2

Sol-Gel Rigid Hosts

An exciting way of looking at sol-gel materials is to use the gel as a host. The important point to note is that the sol-gel materials are transparent. While most of the materials are porous, the porosity is on a scale that does not scatter visible light. Therefore, most light is transmitted. The appeal of the bulk shape host is that it can be molded to near-net shape and can be cut and polished.

Hybrids

Several approaches have been taken to prepare so-called hybrid gels incorporating organic and inorganic components. Those methods that mix the components on the nanometer scale include the infiltration of previously formed oxide gels with monomers, reacting of alkoxide and organic monomers, or organic polymers with alkoxides in mutual solvents. The products of these processes are variously called ormosils (organically modified silicates), ceramers (ceramic polymers), interpenetrating networks (IPNs) and nanocomposites.

The original route to hybrids was one where a gel was formed first as a rigid matrix. Monomer was infiltrated into the open porosity of the shaped gel and polymerized inside the gel using UV curing, for example. Since silica gel is the easiest gel to prepare in monolithic form, one combination that has been investigated widely is silica gel/polymethyl methacrylate (PMMA).

Silica and PMMA, which both have good visible light transmission and similar index of refraction, have been used in hybrids, although it turns out that the match in index of refraction is less critical on the nanoscale. They are also glassy materials with similar strength. PMMA has much higher fracture energy than silica and is often used as glazing in applications where impact resistance is a priority. However, PMMA has a much lower hardness than silica and is less resistant to abrasion. Ideally, silica–PMMA hybrid nanocomposites have the strength and impact resistance of PMMA, and the hardness of silica in one. One obvious application is transparent, lightweight windows.

Encapsulation

Sol-gel hosts have been used to encapsulate a variety of organic and inorganic molecules, chromophores, and biological agents. Whatever is incorporated into the gel can act in two ways, to give information about the gel host or to be influenced by the gel in its interaction with light. For example, spiropyrans have been incorporate for photochromics, CdS for nonlinear optics, Nd for tunable lasers, and various enzymes for sensors. Early on, silica gel hosts were used to encapsulate the electro-optically active organic compounds 2-methyl-4-nitroaniline (MNA). In comparison to conventional optical cells, MNA doped silica hosts retained optical quality at higher temperature and were more abrasion resistant. Since this discovery, silica hosts, in particular, and sol-gel processing, in general, has been used extensively for encapsulating molecules and species that cannot be heated, but need to be protected in a rigid matrix.

Future Directions

As more and more is learned about the sol-gel process, the greater the number of possibilities for useful optical devices. In the selected examples given above, a key feature is that the sol-gel material is transparent. In addition, the sol-gel material is suitable for application to a variety of substrates. Also, the process is a low-temperature process that is compatible with polymers and organic chromophores. In summary, sol-gel processing has had a number of successes in its application to optical materials, and its use is likely to expand.

See also

Optical Coatings: Thin-Film Optical Coatings.

Further Reading

Andrews MP and Najafi SI (eds) (1997) *Sol-Gel and Polymer Photonic Devices*. Bellingham, CA: SPIE Critical Reviews, CR68-03.

Brinker CJ and Scherer GW (1990) *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*. Boston, CA: Academic Press.

Dunn B and Zink JI (1991) Optical properties of sol-gel glasses doped with organic molecules. *Journal of Material Chemistry* 1: 903–913.

Klein LC (1993) Sol-gel optical materials. *Annual Reviews of Material Science* 23: 437–452.

Klein LC (ed.) (1994) *Sol-Gel Optics: Processing and Applications*. Dordrecht: Kluwer Academic.

Najafi SI (ed.) (1998) *Selected Papers on Sol-Gel for Photonics*. Bellingham, CA: SPIE Milestones Series, MS148.

Riesfeld R and Jorgenson CK (1992) Optical properties of colorants or luminescent species in sol-gel glasses. *Structure and Bonding* 77: 208–256.

Sol-Gel Optics. SPIE vols. 1328, 1758, 2288 and 3943, 1990–2000. Bellingham, CA: SPIE.

OPTICAL MICROLENSES

H Ottevaere and H Thienpont, Vrije Universiteit Brussel, Brussel, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In today's world of information processing the role of optics and opto-electronics is expected to become increasingly important as the performance of communication, processing, sensing, and display technologies is continuously evolving. Making these photonics technologies faster and smaller requires, at the same time, the introduction of massive parallelism and miniaturization. As a result, high-quality, high-precision, and low-cost microlens arrays are becoming indispensable components. Moreover, the introduction of micro-optics provides new degrees of freedom for the system design and, therefore, also for the functionality. It allows solutions which are not possible with purely conventional optics such as the integration of different micro-optical components and systems, hence reducing the additional mechanical assembly cost as required in conventional optics.

Since the beginning of the 1990s, several research groups and industrial research labs have been focusing their attention on the development of fabrication techniques for refractive and diffractive microlenses. Microlenses and microlens arrays are used for beam shaping purposes such as collimation and focusing (e.g., in combination with laserdiode arrays, detector arrays, and fibers), for illumination (e.g., in display systems and projection systems), and for imaging purposes (e.g., in photocopiers).

In this article we will mainly focus on refractive microlenses and we intend to provide an introductory guide to those basic geometrical and optical parameters that are used to characterize refractive microlenses. Indeed, some elementary principles of geometrical optics (the geometrical laws for optical propagation without the inclusion of diffraction) and wave optics (in wave optics the imaging is studied as a result of the propagation of a wavefront to the image plane; diffraction effects can be explained on the basis of this wave theory) are required to understand the characterization and performance evaluation of these lenses. We, therefore, bring forward the basic nomenclature of refractive micro-optics and will highlight how the performance of an optical lens is affected when reducing its physical dimensions.

We start this article by introducing the imaging properties of both thick and thin lenses in the paraxial approximation. In the section on lens aberrations, we discuss the different primary aberrations that influence the imaging properties of a lens. Combined effects of aberrations and diffraction give rise to some frequently used lens performance criteria which we discuss in the section on diffraction. Different aspects have to be considered when discussing the quality of a lens, because each type of microlens application requires specific performances. It is, therefore, not possible to come up with a single figure of merit which satisfactorily evaluates the overall quality of a microlens. Moreover, one needs specific information about the characteristics of the light sources and detectors that are used in the system and also details about the imaging system geometry and the required imaging contrast. We review them in the section on quality criteria.

In the section on interference, we describe these phenomena and focus on two-beam interference that forms the basis of the photonic instrumentation that can be used to characterize discrete microlenses and microlens arrays. For a complete evaluation of a microlens it is essential to measure both its geometrical properties (diameter, sag, surface profile, etc.) and its optical performance (focal length, wave aberrations, etc.). In the final section of this article we introduce three interference-based instruments which are crucial to the full quantitative characterization of refractive microlenses: a noncontact optical profiler based on a Mirau interferometer, a transmission Mach–Zehnder interferometer, and a Twyman–Green interferometer.

Geometrical Optics

Introduction

Within the paradigm of geometrical optics or ray-optics, light travels from the source along straight lines or rays. When a light ray traverses an optical system consisting of several homogeneous media in sequence, the optical path is a chain of straight-line segments. Discontinuities in the line segments occur each time the light is refracted or reflected. The laws of geometrical optics that describe the direction of the rays are described below.

The law of refraction, also called Snell's law

The refraction of a ray at a 'well-polished' surface between two isotropic media, with indices of refraction n_1 and n_2, is given by

$$n_1 \sin \varphi_1 = n_2 \sin \varphi_2 \qquad [1]$$

where φ_1 and φ_2 are the angles between, respectively, the incident and the refracted ray and the normal to the surface.

The law of reflection

When a ray is reflected on a flat 'polished' interface dividing two uniform media, the reflected ray remains within the plane of incidence and the angle of reflection equals the angle of incidence.

Geometrical wavefronts can then be constructed from the optical rays propagating through the system. Locally they are orthogonal to the rays and the optical path lengths along all rays from one wavefront to another are equal. In regions where the wavelength is considered to be negligible, as compared to the dimensions of the relevant components of the optical system, geometrical wavefronts are close approximations of the physical wavefronts.

The Paraxial Approximation

The basic optics of image location and magnification is demonstrated by the application of paraxial or first-order optics, also called Gaussian optics. In the paraxial approximation, only those points and rays will be considered which lie in the immediate vicinity of the optical axis, so that all angles φ are small ($\varphi \leq 15°$). This assumption is true in lens systems with small object and image field sizes and small apertures of pupils. The terms involving squares and higher powers of the off-axis distances or of the angles, which the rays make with the optical axis, will not be discussed here. This leads to:

$$\sin \varphi = \tan \varphi = \varphi + O_{\geq 2}(\varphi) \approx \varphi \qquad [2]$$

With this assumption we can write the refraction law by the linear relation:

$$n_1 \varphi_1 = n_2 \varphi_2 \qquad [3]$$

Refraction at a spherical surface

In Figure 1 a spherical surface with a radius of curvature R forms an interface between two materials with indices of refraction n_1 and n_2. Two rays are shown emanating from an axial object point S. One is an axial ray, normal to the surface of the lens at its vertex V. The vertex of a lens is defined as the point on the refracting surface at the center of its free aperture. This assumes that the aperture is circular

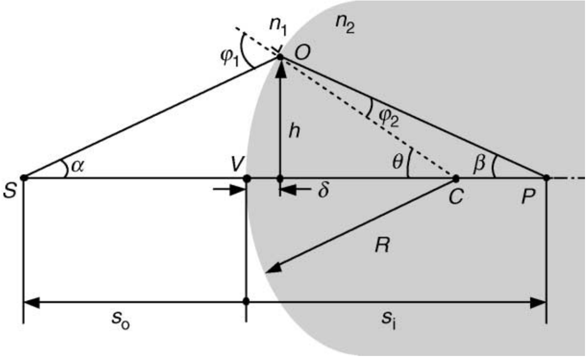

Figure 1 Refraction at a spherical surface.

and centered and the surface is spherical. More generally, if the surface is not spherical but has rotational symmetry, the vertex is the point where the axis of symmetry intersects the optical surface. It is refracted without any change in direction. The other ray is an arbitrary ray incident at O and refracting there according to Snell's Law.

We can rewrite the paraxial refraction law (eqn [3]) and derive a relation between the object and image distances, s_o and s_i, respectively.

From the triangles SOC and POC, we have:

$$\varphi_1 = \alpha + |\theta| \qquad |\theta| = |\beta| + \varphi_2 \tag{4}$$

where $|\theta|$ and $|\beta|$ represent the absolute values of θ and β.

We also have:

$$\tan\alpha = \frac{h}{|s_o| + \delta} \qquad \tan|\beta| = \frac{h}{s_i - \delta}$$
$$\tan|\theta| = \frac{h}{R - \delta} \tag{5}$$

As discussed above, for paraxial rays we may approximate the sine and tangent of an angle by the angle itself and one can show that the small distance δ may be ignored. In accordance with the sign convention the object distance, being to the left of the refracting surface, is to be taken as negative. Inserting eqns [4] and [5] into [3] results in:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \frac{n_2 - n_1}{R} \tag{6}$$

This equation does not contain the angle α, so the image distance s_i is the same for all paraxial rays starting from S.

The spherical surface power $\wp$ is now defined as:

$$\wp \equiv \frac{n_2 - n_1}{R} \tag{7}$$

We can relate this power of the surface $\wp$ to the object and image distance, by combining eqns [6] and [7]:

$$\frac{n_2}{s_i} - \frac{n_1}{s_o} = \wp \tag{8}$$

The distance from the object point to the vertex of the refractive surface is called the front focal length f_{FFL}, if the image is at infinity ($s_i = \infty$). Equations [6] and [8] then give:

$$f_{\mathrm{FFL}} = -\frac{n_1}{n_2 - n_1}R = -\frac{n_1}{\wp} \tag{9}$$

Similarly, if the image is placed at the back focal length f_{BFL}, such that the object distance equals infinity ($s_o = \infty$), then:

$$f_{\mathrm{BFL}} = \frac{n_2}{n_2 - n_1}R = \frac{n_2}{\wp} \tag{10}$$

Thick lenses

A thick lens can be treated as a component consisting of two spherical refracting surfaces separated by a distance d between their vertices V_1 and V_2, as shown in Figure 2. The determination of the focal length as a function of the object and image distance and of the radii of curvature can also be solved using the matrix formalism.

The first and second focal points or the object and image foci, F_o and F_i, can be measured from the two vertices or poles and are called the front and the back focal length denoted by f_{FFL} and f_{BFL}, respectively (Figure 2). The incident and the emerging rays will virtually meet at points that belong to two curved surfaces. These surfaces, approximating planes in the paraxial region, are termed the primary and secondary principal planes, PP_1 and PP_2. Points where the primary and secondary principal planes intersect the optical axis are known as the first and second principal points, H_1 and H_2, respectively.

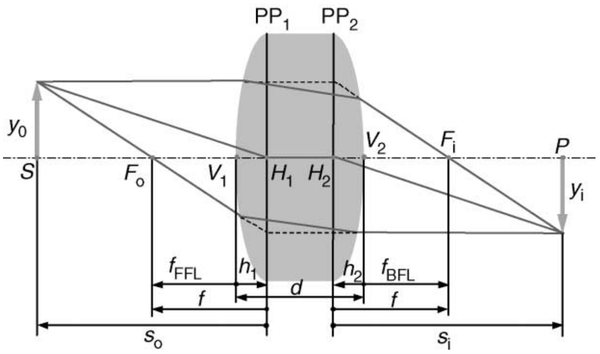

Figure 2 The thick lens geometry.

They provide a set of very useful references from which to measure several of the system parameters. To fully describe a system by its cardinal points, two more points are needed in addition to the two focal points, F_o and F_i and the two principal points, H_1 and H_2. These points are called the nodal points. However, they coincide with the principal points when the lens is surrounded on both sides by the same medium, for example, air. Note that h_1 and h_2 determine the positions of the principal points relative to the vertices V_1 and V_2, while f determines the focal point position relative to the principal points. One can show that these distances are given by:

$$\begin{aligned} h_1 &= \frac{-R_1 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d} \\ h_2 &= \frac{R_2 n_m d}{n_l(R_2 - R_1) + (n_l - n_m)d} \\ f &= \frac{n_m}{(n_l - n_m)\left[\dfrac{1}{R_1} - \dfrac{1}{R_2} + \dfrac{d(n_l - n_m)}{n_l R_1 R_2}\right]} \end{aligned} \qquad [11]$$

We now derive the Gaussian formulae that describe imaging through two surfaces which are rotationally symmetric about the same axis. We assume here that the lens, with refractive index n_l, faces the same medium of refractive index n_m on both sides (Figure 3). We can correctly use this hypothesis since the microlenses which we discuss throughout this work are mainly surrounded by air. Let R_1 and R_2 be the radii of curvature of the surfaces at their axial points, measured as positive when the surface is convex towards the incident light.

The image of a given object, formed by refraction at the first surface with radius of curvature R_1, becomes the object for refraction at the second surface. The image formed by the second surface with R_2 as radius is then the final image due to the action of the thick lens. We know from eqn [6] that the paraxial rays issuing from S at a distance s_{o1} from V_1 will meet at P', at a distance which we call s_{i1}. This gives:

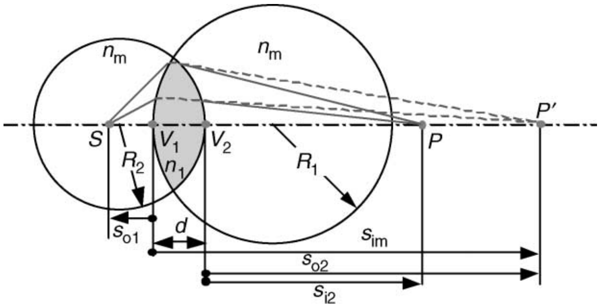

Figure 3 Refraction at both spherical surfaces of a thick lens.

$$\frac{n_l}{s_{i1}} - \frac{n_m}{s_{o1}} = \frac{n_l - n_m}{R_1} \qquad [12]$$

Thus, at the second spherical surface, eqn [6] becomes:

$$\frac{n_m}{s_{i2}} + \frac{n_l}{(s_{i1} - d)} = -\frac{n_m - n_l}{R_2} \qquad [13]$$

As discussed above, we have:

$$\begin{aligned} -\frac{1}{f_{FFL1}} &= \frac{n_l - n_m}{n_m}\frac{1}{R_1} \\ \frac{1}{f_{BFL2}} &= -\frac{n_m - n_l}{n_m}\frac{1}{R_2} \end{aligned} \qquad [14]$$

where f_{FFL1} is the front focal length of the first refracting surface and f_{BFL2} the back focal length of the second refracting surface.

Eliminating s_{i1} in eqns [12] and [13] and inserting [14] yields:

$$d = \frac{n_l}{n_m\left(\dfrac{1}{s_{o1}} - \dfrac{1}{f_{FFL1}}\right)} - \frac{n_l}{n_m\left(\dfrac{1}{f_{BFL2}} - \dfrac{1}{s_{i2}}\right)} \qquad [15]$$

where s_{o1} is measured with respect to V_1 and s_{i2} is measured with respect to V_2. After some calculations this equation may also be put in the form:

$$\begin{aligned} &s_{o1}s_{i2} - \frac{f_{BFL2}(n_l f_{FFL1} + dn_m)}{n_l(f_{FFL1} + f_{BFL2}) + dn_m} \\ &s_{o1} - \frac{f_{FFL1}(n_l f_{BFL2} + dn_m)}{n_l(f_{FFL1} + f_{BFL2}) + dn_m} \\ &s_{i2} + \frac{dn_m f_{FFL1} f_{BFL2}}{n_l(f_{FFL1} + f_{BFL2}) + dn_m} = 0 \end{aligned} \qquad [16]$$

Obviously eqn [16] is much too cumbersome to find image locations. However, we can transform and reduce the latter equation to the form:

$$\frac{1}{s_{i2} - h_2} - \frac{1}{s_{o1} - h_1} = \frac{1}{f} \qquad [17]$$

where h_1, h_2 and f are the unknowns.

Indeed, rewriting eqn [17] gives:

$$s_{o1}s_{i2} - (f + h_2)s_{o1} + (f - h_1)s_{i2} + f(h_1 - h_2) + h_1 h_2 = 0 \qquad [18]$$

Equations [16] and [18] will be identical if the coefficients of their corresponding terms are equal;

that is when:

$$
\begin{aligned}
f+h_2 &= \frac{f_{\mathrm{BFL2}}(n_l f_{\mathrm{FFL1}}+dn_{\mathrm{m}})}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}}\\
f-h_1 &= -\frac{f_{\mathrm{FFL1}}(n_l f_{\mathrm{BFL2}}+dn_{\mathrm{m}})}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}}\\
f(h_1-h_2)+h_1h_2 &= \frac{dn_{\mathrm{m}}f_{\mathrm{FFL1}}f_{\mathrm{BFL2}}}{n_l(f_{\mathrm{FFL1}}+f_{\mathrm{BFL2}})+dn_{\mathrm{m}}}
\end{aligned}
\qquad [19]
$$

These three equations may be solved for the three unknowns h_1, h_2, and f:

$$
\begin{aligned}
f &= \frac{n_{\mathrm{m}}}{(n_l-n_{\mathrm{m}})\left[\dfrac{1}{R_1}-\dfrac{1}{R_2}+\dfrac{d(n_l-n_{\mathrm{m}})}{n_lR_1R_2}\right]}\\
h_1 &= \frac{-R_1n_{\mathrm{m}}d}{n_l(R_2-R_1)+(n_l-n_{\mathrm{m}})d}\\
h_2 &= \frac{R_2n_{\mathrm{m}}d}{n_l(R_2-R_1)+(n_l-n_{\mathrm{m}})d}
\end{aligned}
\qquad [20]
$$

Thus we see that the relation between the object position and the image position (eqn [16]) is expressed in the more convenient form of eqn [17] with the values of h_1, h_2, and f given by eqn [20]. Here h_1, h_2 and f depend on the constants of the system where, as shown in eqn [11], h_{i} represents the distance between the vertex V_{i} and the principal point H_{i}, and f is the focal length measured from a principal point on the axis and not from the vertex of the lens. If we now assume that:

$$s_{\mathrm{o}} = s_{\mathrm{o}1} - h_1, \qquad s_{\mathrm{i}} = s_{\mathrm{i}2} - h_2 \qquad [21]$$

we can write the thick lens equation as:

$$\frac{1}{s_{\mathrm{i}}} - \frac{1}{s_0} = \frac{1}{f} \qquad [22]$$

where the object and image distance, s_{o} and s_{i}, respectively, are measured from the principal points of the lens. We can conclude that the formula that describes imaging through two surfaces may be simplified to the thick lens formulation given in eqn [22] by choosing the principal points H_{i} instead of the vertices V_{i}, from which S and P are to be measured.

Thin lenses

The preceding eqn [20] takes a particularly simple form when the lens is so thin that the axial thickness of the lens d may be neglected. But obviously there exists no real lens with a thickness equal to zero. A thin lens is only a concept which is an extremely useful tool in optical system design. When a lens or an optical system has a zero thickness, the object and image calculations can be greatly simplified. In this case, the two principal points are coincident and are located where the thin lens is situated.

As a further simplification, we assume that the surrounding medium of the lens is air which means that $n_{\mathrm{m}} \approx 1$. When $n_l = n$, we find:

$$\frac{1}{f} = \frac{1}{s_{\mathrm{i}}} - \frac{1}{s_{\mathrm{o}}} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \qquad [23]$$

These equations are often called the thin-lens equation and the lensmaker's formula, respectively, because the latter predicts the focal length of a lens, fabricated with a given refractive index and radii of curvature. This formula is very useful in the optical design process as a first-order approach to calculate the features of the thin lenses and results in a set of lens powers (or focal lengths) for the different lenses and in the spacings between them. Once the thin lens optical system is solved and the primary aberrations are corrected for, one can design the thick lens system. The corresponding thick lens system must have the same component powers and the lenses must be spaced apart by the same distances as for the thin lens design, but the thick lens spacings must be measured from the principal points of the lenses instead of from the vertices of the lenses.

Other lens characteristics

Another basic parameter of a refractive lens is its focal number $f\#$. The f-number, also called the speed of the lens, is defined as the ratio of its effective focal length and the lens aperture diameter D (Figure 4):

$$f\# = \frac{f}{D} \qquad [24]$$

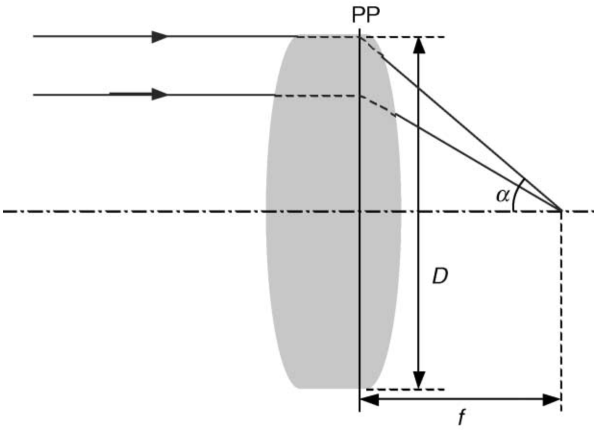

Figure 4 A lens with an aperture diameter D, a focal length f, and a semi-angle α.

The effective focal length f of a lens or often called the focal length is defined from a principal point on the axis of the lens and not from its vertex. Typical focal numbers for refractive microlenses range between 0.8 and 30.

The numerical aperture (NA), is another way of defining the same characteristic. It is described as the sine of the angular semi-aperture α in the object space multiplied by the refractive index n of the object space (Figure 4):

$$\mathrm{NA} = n \sin \alpha \qquad [25]$$

Both the focal number ($f\#$) and the NA are related in the paraxial approximation by

$$f\# = \frac{n}{2\mathrm{NA}} \qquad [26]$$

Lens Aberrations

Introduction to Third-Order Monochromatic Aberrations

An exact ray trace or a measurement on a real lens will certainly disclose inconsistencies with the corresponding paraxial description. Such deviations from the idealized Gaussian optics are known as aberrations. For monochromatic light there are five third-order or Seidel aberrations: three of them, namely spherical aberration, coma, and astigmatism, deteriorate the image by blurring it, while the remaining two, curvature of field and distortion, deform the image. An additional aberration, chromatic aberration, results from the wavelength dependence of the imaging properties of an optical system. We will not discuss the latter as this aberration can be minimized by selecting a lens material with a refractive index that only varies by a negligible amount over the wavelength range of interest.

For a brief discussion of aberrations we consider the case of a perfect lens focusing a collimated light beam. In this case, the lens transforms the incoming planewave into a spherical wave, also called the reference sphere. For a lens with aberrations, however, the real wavefront behind the lens shows deviations from the ideal spherical shape (Figure 5). The optical path length [AB] may be called the wavefront aberrations $\psi(x, y)$ and can be calculated as:

$$\psi(x, y) = W(x, y) - R(x, y) \qquad [27]$$

where $W(x, y)$ represents the spherical wavefront in the Gaussian approximation and $R(x, y)$ depicts the real wavefront. The wave aberration is given by the optical path difference along a ray between the reference sphere and the real wavefront behind the lens. Alternatively, the aberrations can be described in terms of the ray aberrations. In this case, not all the rays intersect the paraxial image plane at the same point and the distribution of the ray intercepts around the paraxial image point can be used to measure the amount of light blur. These two ways of quantitatively describing the aberrations are apparently not independent of each other, the ray aberrations being the derivate of the wavefront aberrations.

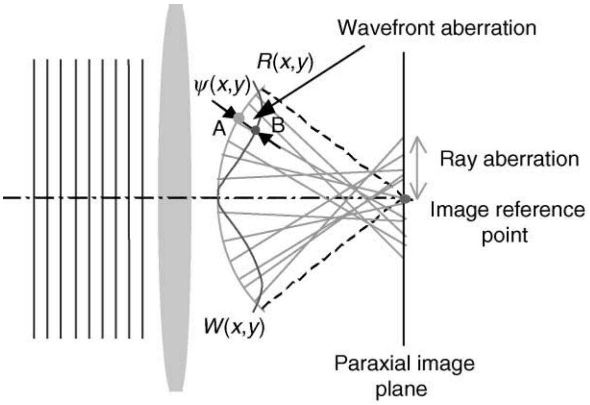

Figure 5 Relation between wavefront and ray aberration.

There are several sets of suitable polynomials over which the wavefront aberrations can be expanded. The Zernike polynomials will be further used in this work (see, for example, the Mach–Zehnder and Twyman–Green interferometers in the section on Microlens testing).

For rotational symmetric optical systems, the most common approach is to expand the aberration function ψ to the fourth order as a power series in polar pupil coordinates. This corresponds to the third-order ray aberrations or primary aberrations. The different terms in the aberration polynomial ψ are to the fourth power of the object coordinate h' and the pupil coordinate r. In that case, the wavefront aberration polynomial ψ, as a function of the object coordinate h' and the polar pupil coordinates r and θ (Figure 6), is given by

$$\psi(h', r, \theta) = {}_0C_{40}r^4 + {}_1C_{31}h'r^3 \cos \theta + {}_2C_{22}h'^2 r^2 \cos^2 \theta + {}_2C_{20}h'^2 r^2 + {}_3C_{11}h'^3 r \cos \theta \qquad [28]$$

The ${}_iC_{jk}$ coefficients are subscripted by the numbers that specify the powers of the term dependence on h', r and $\cos \theta$, respectively. These terms comprise the five monochromatic or Seidel aberrations as shown in Table 1.

Each aberration is characterized by its dependence on the deviation from the optical axis h', on the aperture of the refracting surface r, and on the

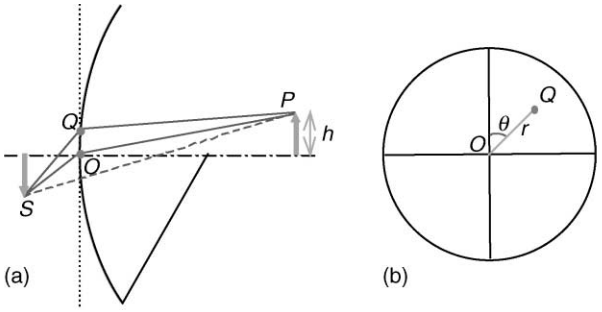

Figure 6 (a) Imaging of an off-axis point *P*; (b) frontal view of a portion of the wavefront.

Table 1 The five monochromatic or Seidel aberrations

r^4	Spherical aberration
$h' r^3 \cos\theta$	Coma
$h'^2 r^2 \cos^2\theta$	Astigmatism
$h'^2 r^2$	Curvature of field
$h'^3 r \cos\theta$	Distortion

symmetry around the optical axis θ. Equation [28] can be used to estimate how the aberration blurs change if the aperture or field of view of a lens is changed. In the following section we briefly describe each of these aberrations in terms of their visual effects.

Spherical Aberration

The first term in eqn [28], ${}_0C_{40}r^4$, is the only term in the third-order wave aberration that does not depend on the deviation h' from the optical axis. Thus, spherical aberration (SA) is the only monochromatic aberration that occurs on the optical axis and that exists even for axial object and image points. As shown in Figure 7a, SA results from the rays that enter via different zones of the aperture and are focused at different distances from the lens. It is evident from Figure 7a that the best focus is not at the paraxial plane. The best focus, however, is found at the location where the circular image blur is the same over the entire field of view and is called the circle of least confusion.

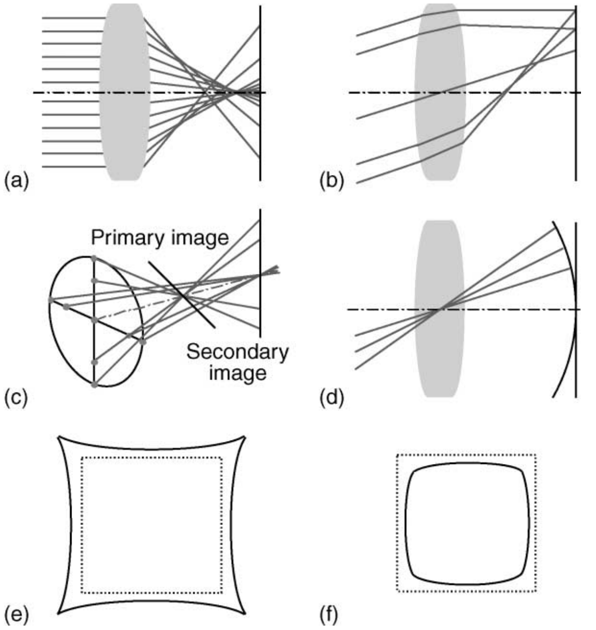

Figure 7 The primary aberrations: (a) spherical aberration – negative SA: the rays through the outer zones of the lens focus closer to the lens than the rays through the central zones; (b) coma – positive coma: the rays through the outer zones of the lens form a larger image than those going through the center; (c) astigmatism – negative astigmatism: the tangential or vertical fan of rays is focused to the left of the horizontal or sagittal fan; (d) curvature of field – negative curvature of field: images farther from the axis focus nearer to the lens than the on-axis images; (e) positive distortion – the magnification increases as the field angle increases; (f) negative distortion.

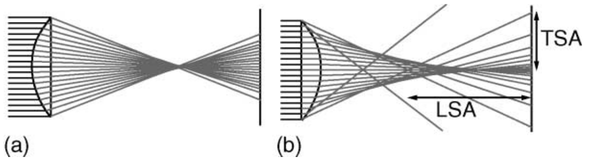

Figure 8 Ray diagram of a planoconvex lens ($D = 200\ \mu$m; $f = 249\ \mu$m) with planewave illumination incident on (a) the curved surface and (b) on the plane surface. LSA (longitudinal SA) represents the difference between the paraxial and the marginal focus while TSA (transversal SA) shows the radius of the spot in the paraxial focal plane.

Because most applications consist of on-axis illuminated planoconvex microlenses, we will firstly focus our attention on how microlenses have to be positioned to minimize SA. When we consider a planoconvex lens with the curved surface towards the infinite conjugate for an incident parallel beam, both surfaces are contributing approximately equal amounts of ray bending. Figure 8a shows the ray diagram for a planoconvex lens as obtained with the ray-tracing program Solstis. Solstis is a commercial photonics design software, consisting of different modules, each of which is dedicated to a certain optical modeling approach. If, however, the flat surface is turned toward the incident rays, then the first surface does not bend the rays and all the ray-bending work is done by the second surface (Figure 8b). We can conclude here that the best orientation is that which distributes the ray bending equally between the two surfaces of the lens.

Because the ray blur varies with the third power of the lens aperture, lowering the aperture diameter will cause a rapid decrease in the extent of the spherical

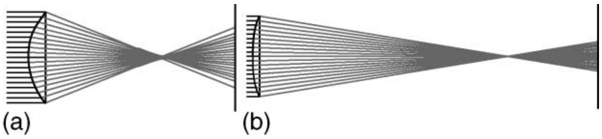

Figure 9 Ray diagram of a planoconvex lens (a) with a focal length *f* and (b) with a focal length *2f*.

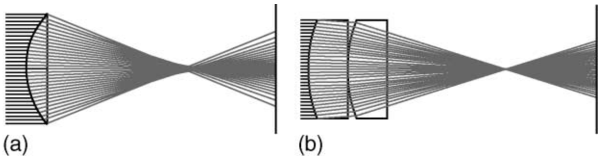

Figure 10 (a) Ray diagram for a planoconvex lens ($D = 200$ μm; $f = 249$ μm). (b) The spherical aberration is reduced with a factor 4 by splitting the optical power of the original lens and using a doublet ($D = 200$ μm; $f = 621$ μm for both lenses).

aberration blur. For a given aperture of the lens, the spherical aberration is a function of the focal length of the lens and increases with a decreasing focal length or an increasing NA (Figure 9).

A change in SA, contributing to the wave aberration, for a variation of the lens diameter D or the focal length f can be estimated with the following rules:

$$\text{Spherical aberration} \div D^4 \quad \text{for a constant } f$$

$$\text{Spherical aberration} \div 1/f^3 \quad \text{for a constant } D$$

However, by preserving the focal length and hence the optical power of the lens, the SA can be changed by varying the radii of the surfaces. This would be the same as physically bending a lens made of flexible plastic. In Figure 10 we have reduced the spherical aberration with a factor of four by splitting a lens into two elements, each with a longer focal length, keeping the total optical power of the elements the same as the original lens. This means that, in practice, high NA microlenses can be replaced by a number of microlenses reducing the SA significantly.

However, in optical systems, where only a small number of surfaces are allowed, aspherical surfaces can be used to correct the SA. In Figure 11 we show that making the lens shape elliptical while maintaining the other characteristics of the microlens ($D = 200$ μm; $R = 150$ μm) results in a reduction of the SA. The most common form of an aspheric surface is a rotationally symmetric surface with the sag, defined as:

$$z_{\text{conic}} = \frac{R^{-1}(x^2 + y^2)}{1 + \sqrt{1 - (1 + K)R^{-2}(x^2 + y^2)}} \qquad [29]$$

where R is the radius of curvature, K the conic parameter, and x and y the coordinates in the substrate plane. The meaning of K is given in Table 2.

In eqn [28] we have shown that the wavefront aberrations can be written as the sum of the Seidel or third-order aberrations. Furthermore, we have seen that only the SA contributes to the wavefront aberration when we illuminate the microlens on-axis ($h' = 0$). When we limit our experiments to on-axis illumination during the optical characterization of the microlenses, the wavefront aberrations are a measure for the SA. To prove this, we have calculated and plotted in Figure 12, the SA and the wavefront aberrations as a function of the focal length for a 200 μm lens ($n = 1.48$).

Both the SA and wavefront aberration values are obtained with the ray-tracing software Solstis. The wavefront aberration is calculated here as the optical path difference [BA] between a real wavefront surface and a spherical reference surface. From these results we can observe that for microlenses with an $f > 330$ μm or an NA lower than 0.3 only the spherical aberration contributes to the root-mean-square (RMS) wavefront aberration. The RMS or standard deviation σ of a number N of data x_i ($i = 1, \ldots, N$) with a mean value X can be written as $\sigma = \sqrt{\sum_{i=1}^{N} (X - x_i)^2 / N}$. For smaller focal length microlenses, however, we notice that higher-order aberrations appear and compensate the SA, resulting in a lower wavefront aberration compared to the SA. We can conclude that the wavefront aberration values represent the optical performance of the microlenses with a main contribution from the SA for on-axis illumination.

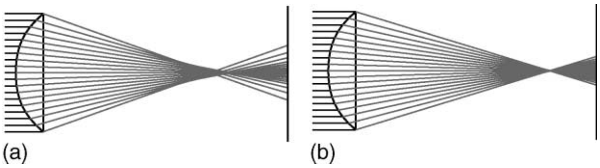

Figure 11 (a) Ray diagram of a spherical planoconvex lens ($D = 200$ μm; $R = 150$ μm; $K = 0$) with planewave illumination. (b) The spherical aberration is reduced by making the lens shape elliptical ($K = -0.57$).

Table 2 The surface type for different conic parameters K

Conic constant K	*Surface type*
$K = 0$	Sphere
$K < -1$	Hyperboloid
$K = -1$	Paraboloid
$-1 < K < 0$	Ellipsoid
$K > 0$	Oblate ellipsoid

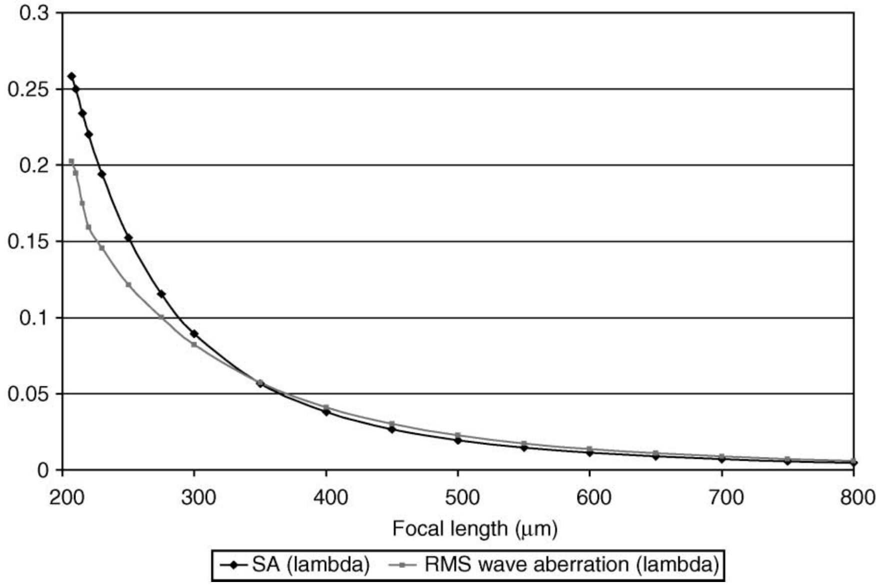

Figure 12 The spherical aberration SA and the RMS wave aberration as a function of the focal length for a 200 μm lens ($n = 1.48$).

Other Seidel Aberrations

Coma is an image-degrading aberration that increases rapidly with the lens aperture r. It indicates an off-axis aberration ($h' \neq 0$) that is nonsymmetrical around the optical axis ($\cos\theta \neq$ constant). Figure 7b illustrates the aberration due to a vertical fan of parallel rays refracted by a spherical lens. It results from a lens having different magnifications for rays passing different zones of its aperture. Each circular zone of the lens forms a circular image called the comatic circle. Every other fan of rays forms images that complete the comatic circle. The combination of all these circles causes a comet-shaped flare.

Astigmatism is the imaging of an off-axis point as two perpendicular lines at different distances. In this aberration the rays from a point object converge at a certain distance from the lens to a line which is called the primary image. This line lies in the sagittal plane and is perpendicular to the plane defined by the optical distance and the object point. At a somewhat different distance from the lens they converge to a second line, called the secondary image, which is parallel to the tangential plane. This effect is shown in Figure 7c. The elimination of astigmatism requires that the tangential and sagittal surfaces be made to coincide. The resulting surface is then called the Petzval surface. Although in that case astigmatism has been eliminated, the associated aberration, called curvature of field, remains. So off-axis images are focused on a curved saucer-shaped surface as drawn in Figure 7d, instead of on an ideal flat image surface.

The last of the five primary aberrations is distortion and causes straight lines which do not intersect the axis to be imaged as curved lines. This effect is not related to a lack of sharpness of the image but results from the fact that the magnification varies across the field and induces the image of a square object to be bowed outward (Figure 7e) or sagged inward (Figure 7f).

We have shown that three of the five Seidel aberrations (namely SA, coma, and astigmatism) are responsible for a lack of sharpness of the image. The other two (i.e., curvature of field and distortion) are related to the position and the form of the image. In general, it is impossible to design a system which is free from all the primary as well as the higher-order aberrations, and therefore, a suitable compromise as to their relative magnitude has to be made. In some cases, the effects of the Seidel aberrations are reduced by balancing them against aberrations of higher orders. In other cases, one has to eliminate certain aberrations completely, even at the price of introducing aberrations of other types. So far we have studied the aberration effects on the basis of geometrical optics only. However, if the aberrations are very small (wave aberrations of the order of a wavelength or less), diffraction starts to play an important role. That is why, in the next section, we will discuss the role of diffraction on the imaging quality of a lens.

Diffraction

Introduction

So far we have used simple geometrical optics to evaluate the effects of aberrations – deviations of the wavefront behind the lens from the ideal shape – that influence the quality of an image generated by a lens. However, we know that even for an aberration-free lens, blur of the spot in the image plane occurs. This means that diffraction effects at the lens aperture can also influence the spot size and the point spread function (PSF), certainly because diffraction gains importance when miniaturizing optical components. Therefore, we have to introduce diffraction before we can bring forward the most important figures of merit for microlens performance and study its effects on imaging quality.

Fraunhofer and Fresnel Diffraction

In certain regions of space occupied by our optical lens system, the simple geometrical model of energy propagation is inadequate. In particular, deviations from this model can be expected in the immediate vicinity of the boundaries of shadows and in regions where a large number of rays meet. These regions are of great practical interest, as they include that part of the image space in which the optical image is situated. These deviations are manifested by the appearance of diffraction fringes. Also, the phenomena involving the superposition of coherent wavefronts cannot be treated with a geometrical model of energy propagation. Thus a more accurate treatment is necessary where the wave nature of light and diffraction are taken into account. Wave optics is such an approach with which the actual light intensity, that will be observed as a result of the propagation of a wavefront to the image plane, can be constructed. According to Huygens' construction, every point of a wavefront may be considered as a center of secondary disturbance, which gives rise to spherical wavelets. The wavefront at any later instance may be regarded as the envelope of these wavelets. Fresnel was able to account for diffraction by supplementing Huygens' construction with the postulate that the secondary wavelets mutually interfere. We discuss diffraction here, on the basis of wave theory, by the application of Huygens' construction together with the principle of interference. The latter principle, also called the Huygens–Fresnel principle, sufficiently describes the propagation of light in free space.

This Huygens–Fresnel principle may be regarded as an approximate form of the integral theorem of Helmholtz and Kirchhoff. Kirchhoff's theory makes an approximation to the diffraction of scalar waves. This scalar theory is adequate enough for the treatment of problems in instrumental optics where polarization effects are of minor importance and can be neglected. If we consider a strictly monochromatic electric field $\varepsilon(x,y,z,t) = E(x,y,z)\cdot \mathrm{e}^{-\mathrm{i}\omega}$ in any ideal isotropic dielectric, the space-dependent part will satisfy the homogeneous time-independent wave equation: $(\nabla^2 + k^2)E = 0$, also known as the Helmholtz equation where $k = \omega/c$. This theorem expresses the solution E of the Helmholtz equation at an arbitrary field point P, in terms of its solution E and its first derivatives $\partial E/\partial n$ at all points on an arbitrary closed surface surrounding P, provided the medium is a homogeneous dielectric:

$$E(P) = -\frac{1}{4\pi}\oiint_S \left[E\bar{\nabla}\left(\frac{\mathrm{e}^{-\mathrm{i}ks}}{s}\right) - \frac{\mathrm{e}^{-\mathrm{i}ks}}{s}\bar{\nabla}E\right]\bar{n}\,\mathrm{d}S \qquad [30]$$

Here, s is the distance from point P to the point on surface S, n is the outward normal to S, and $k = \omega/c$ is the wave number in vacuum.

When we consider a monochromatic wave, from a point source P_o propagating through an aperture Ω in a plane opaque screen, and P is the point at which the light disturbance has to be calculated, the theorem can be reduced to an approximate but much simpler form (Figure 13a) on condition that we can assume that the linear dimensions of the aperture Ω, although large compared to the wavelength, are small compared to the distances of both P_o and P from the screen.

As a result of Kirchhoff's theorem we obtain:

$$E(P) = \frac{\mathrm{i}A}{2\lambda}\int\int_\Omega \frac{\mathrm{e}^{-\mathrm{i}k(r+s)}}{rs}[\cos(n,r) - \cos(n,s)]\mathrm{d}S \qquad [31]$$

where A is a constant, λ the wavelength, and r and s are the distances from element $\mathrm{d}S$ on Ω to P_o and P, respectively. Equation [31] is known as the Fresnel–Kirchhoff diffraction formula. In case a more general field E_o is incident at the aperture Ω, replacing the point source $A\,\mathrm{e}^{-\mathrm{i}kr}/r$, an equivalent formulation, known as the Huygens integral, can be found:

$$E(P) = \frac{\mathrm{i}}{\lambda}\int\int_\Omega E_\mathrm{o}(x_\mathrm{o}, y_\mathrm{o}, z_\mathrm{o})\frac{\mathrm{e}^{-\mathrm{i}ks}}{s}\cos(n,s)\mathrm{d}S \qquad [32]$$

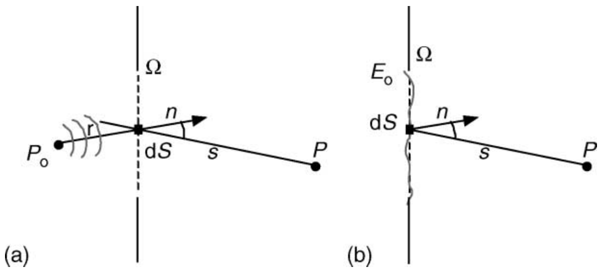

Figure 13 Geometry for evaluating (a) the Fresnel–Kirchhoff diffraction formula and (b) the Huygens' integral.

In the case of the Huygens integral an alternative Green's function is chosen, namely $G(P) = \exp(-\mathrm{j}ks)/s - \exp(-\mathrm{j}kr)/r$ instead of $G(P) = \exp(-\mathrm{j}ks)/s$ used in the Fresnel–Kirchhoff diffraction formula. This then results in an obliquity factor $\cos(n,s)$ instead of $\cos(n,s) - \cos(n,r)$. If the distances of the points P_o and P are large compared to the dimensions of the aperture, the factor $[\cos(n,r) - \cos(n,s)]$ in eqn [31] will not vary appreciably over the aperture and may be replaced by 1. In this case, the optical beams are close to the optical axis and this is considered as the paraxial diffraction theory. In the phase factors, we can expand $r(x,y,z)$ and $s(x,y,z)$ as power series. If the quadratic and higher-order terms may be neglected in the series expansion $(r,s \gg \lambda)$, one speaks of Fraunhofer diffraction, when the previous assumption is not fulfilled $(r,s \geq \lambda)$, the quadratic terms cannot be neglected and one speaks of Fresnel diffraction. The regions where the Fresnel and Fraunhofer diffraction integral may be applied are called the near-field (NF) and far-field (FF) diffraction regions respectively. The Rayleigh range z_R marks the boundary between the NF and FF regions for the beam emerging from the aperture (Figure 14).

As an example, we discuss the diffraction of a uniform planewave by a circular aperture of diameter $2a$. The Rayleigh range for this circular aperture is given by

$$z_R = \frac{\pi a^2}{\lambda} \qquad [33]$$

The FF diffraction pattern $(z \gg z_R)$ for a uniformly illuminated circular aperture is the well-known Airy disk pattern and is given by

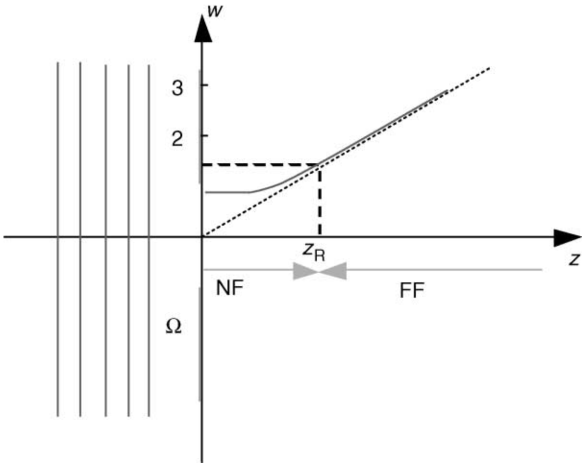

Figure 14 The Rayleigh range z_R marks the boundary between the NF and FF regions for a beam emerging from the aperture Ω.

$$E(r,z) \sim \frac{2J_1(2\pi a\theta/\lambda)}{(2\pi a\theta/\lambda)} \qquad [34]$$

where J_1 is the first-order Bessel function. This pattern has a single dominant central lobe surrounded by a series of increasingly weaker circular rings. The first zero of this pattern occurs at a half angle θ_1 or a radius r_1 in the FF, given by

$$\theta_1 = \frac{r_1}{z} \approx \frac{1.22\lambda}{2a} \qquad [35]$$

The NF or Fresnel diffraction patterns for a uniformly illuminated circular aperture $(z < z_R)$ consist of a series of circular rings modulating a constant-amplitude background. When moving even closer to the aperture, the beam profile becomes more and more square and the frequency of the Fresnel ripples increases.

We now know that light is diffracted at the apertures of optical elements (lenses) and that this diffraction can affect its performance. Ideally, the focus of a planewave should be infinitely small, being the image of a point source located at infinity. But even in the ideal case, when no geometrical aberrations are introduced, diffraction will limit the lens performance and the focus of such diffraction-limited lenses will have a finite extension.

Quality Criteria for Lens Performance

In this section we discuss some of the most important figures of merit for diffraction-limited lens performance. This survey raises the problems related to this issue. A more detailed overview can be found in the literature.

Diffraction in a Lens System: The Point Spread Function (PSF)

Diffraction at the lens aperture D causes a blur of the focus. The light distribution in the focal plane of an aberration free lens essentially arises from Fraunhofer diffraction at the aperture of the lens. From eqn [34] we know that the diffraction image for an aberration-free lens with a circular aperture D is the scaled Airy disk, given by

$$E(r,z) \sim \frac{\pi J_1(rD/\lambda f)}{(rD/\lambda f)} \qquad [36]$$

where λ denotes the wavelength of the illuminating light beam and f the focal length.

The normalized intensity of this pattern is called the diffraction-limited (PSF) of the lens:

$$\mathrm{PSF}(r, z) = |E(r, z)|^2 \qquad [37]$$

Although irradiance replaces nowadays the word intensity in optics, we will still use intensity to nominate the flow of energy per unit area per unit time. We can define the irradiance as $I \equiv \frac{1}{2}\varepsilon_0 c\langle \mathbf{E}^2 \rangle$, with ε_0 and c the electric permittivity in vacuum and the speed of light, respectively. The radius of this pattern is known as the Rayleigh diffraction limit or Rayleigh resolution and is compared to the measured spot size in the image plane.

To include the effects of aberrations in the diffraction computation, we have to include both the phase profile and the amplitude of the real wavefront in the pupil of the lens. The pupil function $T(x, y)$ may be written as:

$$T(x, y) = E(x, y)\mathrm{e}^{\mathrm{i}k\psi(x,y)} \qquad [38]$$

where $E(x, y)$ represents the amplitude distribution over the exit pupil and $\psi(x, y)$ is the wavefront deformation on this pupil. Starting from the Huygens–Fresnel approach, the diffraction integral can be rewritten in the form of a Fourier transform:

$$E(r, z) \sim FT(\mathrm{e}^{\mathrm{i}k\psi(x,y)}) \qquad [39]$$

The PSF at the focal position can then be found by calculating the Fourier transform of the wavefront error ψ or the pupil function $T(x, y)$ of the lens and taking the complex square of the latter. Figure 15 shows the shape of the PSF of an ideal or diffraction-limited lens. When the lens aperture D decreases, the extension of the PSF increases proportionally, which means that it becomes easier to achieve diffraction-limited performance. Since the extension of the diffraction-limited PSF becomes larger, the constraints on the shape of the phase profile become less stringent. This has important consequences on the scaling behavior of lenses. In the extreme case, the microlens diameter becomes so small, that diffraction-limited imaging is possible without any specific phase profile in the lens pupil. This is also one of the reasons why small water droplets often form relatively good lenslets. For systems including micro-optic lenses, we therefore will have to find better criteria than the PSF of a diffraction-limited lens to evaluate or compare these optical components.

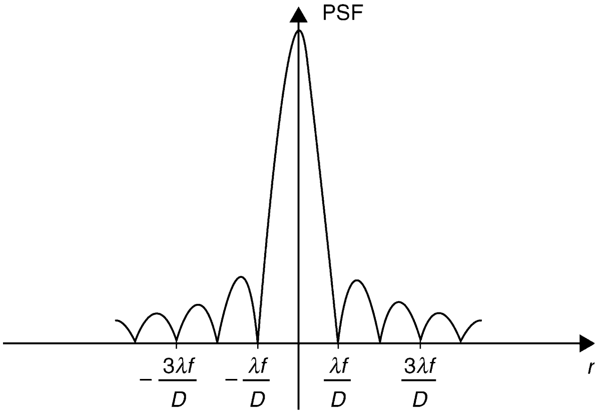

Figure 15 The 1D point spread function of an ideal lens.

The Rayleigh and Maréchal Criterion

The widely used Rayleigh criterion states that a system is diffraction-limited, as long as the maximum wavefront deviation $|\psi_{\max}(x)|$ is less than one quarter of a wavelength (Figure 16). It was shown by Rayleigh that for a system which suffers primarily from SA such that the wavefront deviation is less than a quarter wavelength, the intensity at the Gaussian focus is diminished by less than 20%; a loss of light that can usually be tolerated. Also in the presence of other commonly occurring Seidel aberrations is the quality of the image not seriously affected when the deviation is less than λ/4. This criterion, which determines the amount of aberration that can be tolerated in an image-formation system is, of course, only a rough guideline since the light distribution in the image does not only depend on the maximum deformation but also on the shape of the wavefront, and even more on the type of aberration. Moreover, the loss of light that can be tolerated depends on the particular use to which the optical component is put and more stringent tolerances have to be imposed for certain applications.

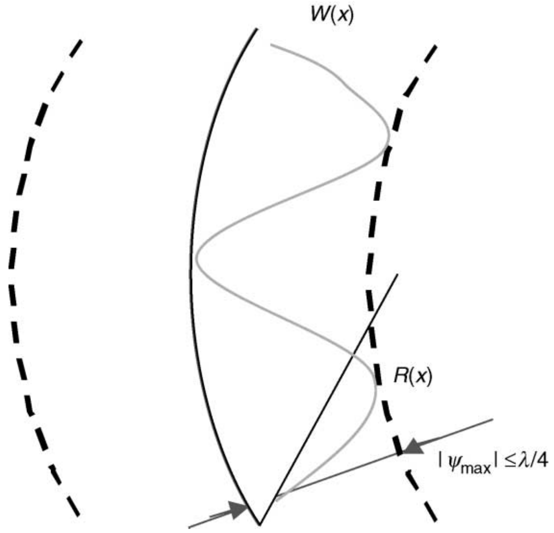

Figure 16 Illustration of the Rayleigh criterion.

Low-order aberrations merely cause a shift of light from the main lobe of the PSF to the side lobes. Higher-order aberrations additionally cause a broadening of the PSF and the correlation between the image quality and the maximum value of the wavefront aberration is usually lost. In the latter case, it seems more appropriate to formulate tolerance criteria restricting the average amount of aberration. Criteria of this type were considered by Maréchal, who used the relation that exists between the intensity at the center of the reference sphere and the RMS deviation of the wavefront from a spherical form.

This Maréchal criterion uses the following RMS wavefront aberration ψ_{RMS} to define the lens quality:

$$\psi_{\mathrm{RMS}} = \sqrt{\int |\psi(x)|^2 \,\mathrm{d}x - \left[\int |\psi(x)| \mathrm{d}x\right]^2} \leq \frac{\lambda}{14} \qquad [40]$$

This criterion is sensitive to statistical noise on the measured phase profile, while a criterion such as the Rayleigh criterion is little affected by it.

Strehl Ratio

For some applications not the imaging quality but the maximum deposited energy is the most essential feature. In such a situation, the Strehl ratio S is a suitable figure of merit. The Strehl ratio is defined as the normalized peak intensity of the PSF of the lens:

$$S = \frac{I_{\mathrm{real}}(0,0)}{I_{\mathrm{ideal}}(0,0)} = \left| \int\int \mathrm{e}^{ik\psi(x,y)} \mathrm{d}x \,\mathrm{d}y \right|^2 \qquad [41]$$

$I_{\mathrm{real}}(0,0)$ and $I_{\mathrm{ideal}}(0,0)$ denote the intensities at the center of the real point image and the ideal PSF without aberrations, respectively. Even for small aberrations, which do not affect the extension of the PSF, the light intensity in the center peak can drop significantly (Figure 17). A Strehl ratio $S \geq 0.8$ is generally considered to correspond to diffraction-limited performance.

By a Taylor expansion of eqn [40], the relation between the Strehl ratio S and the Maréchal criterion can be written as:

$$S = 1 - \left(\frac{2\pi}{\lambda}\psi_{\mathrm{RMS}}\right)^2 \qquad [42]$$

If only low-order aberrations are taken into account, Rayleigh's $\lambda/4$ criterion also results in a Strehl ratio of $S = 0.8$. For large aberrations ($S \leq 0.3$), however, the behavior of the Strehl ratio becomes uncorrelated to the image quality due to interference effects.

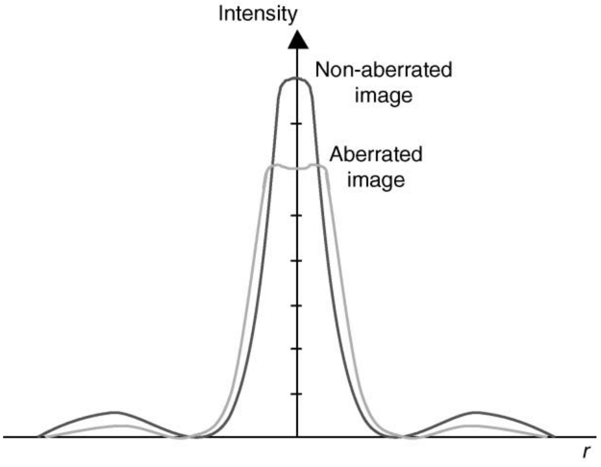

Figure 17 Strehl ratio criterion.

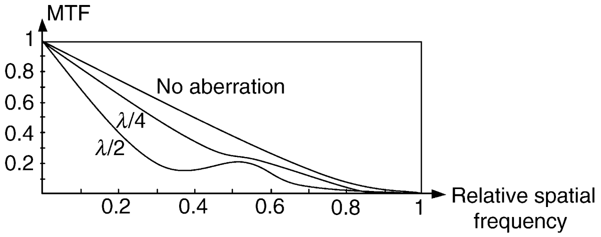

Figure 18 Modulation transfer function for a perfect lens and for small amounts of spherical aberration (λ/2 and λ/4).

Transfer Functions

Transfer functions are most commonly used for evaluating the image quality in lens design. As discussed above, the PSF is the Fourier transform of the wavefront error ψ. For simplicity we can assume that the modulus of the pupil function is 1. The PSF then follows from the Fourier transform:

$$PSF(\mu, \nu) = |\mathrm{FT}(\mathrm{e}^{ik\psi(x,y)})|^2 \qquad [43]$$

The optical transfer function (OTF) is calculated from the PSF data by means of an inverse Fourier transformation:

$$\mathrm{OTF}(\vartheta, \rho) = \mathrm{FT}^{-1}[\mathrm{PSF}(\mu, \nu)] \qquad [44]$$

The OTF is a spatial frequency-dependent complex quantity whose modulus is the modulation transfer function (MTF) and whose phase is, of course, the phase transfer function (PTF). The former is a measure for the reduction in contrast from the object to the image over the whole spectrum. Figure 18 shows the normalized MTF for both a perfect lens and a lens subjected to some small amounts of spherical aberration. The PTF is, in most cases, not so conclusive concerning the optical behavior.

Therefore the wave aberrations, the PSF, the MTF, and the Strehl ratio are the characteristics that are usually provided as an output of lens design programs.

Interference

Many of the processes for the manufacturing of microlenses are not fully understood nor perfectly controlled. It occurs that one generates a microlens array and then measures its performance, hoping that by accurately repeating the processing conditions, a consistent result can be achieved. To achieve optimum conditions, it is therefore crucial to measure the geometrical and optical parameters of the microlenses. Most of the microlens measurement instruments are based on two-beam interference, the phenomena where light from a source is divided into two beams which are later on superposed again. The intensity in the region of superposition is found to vary from point to point between maxima which exceed the sum of the beam intensities and minima which may be zero.

Two-Beam Interference

We consider the interference of two waves, a reference and an object wave, represented by $\mathbf{E}_{\text{ref}}$ and $\mathbf{E}_{\text{obj}}$ respectively, where we take into account the vector properties of the electric fields. In the case of interference, both waves typically originate from a single source and recombine after traveling along different paths. For the moment we will consider only monochromatic linearly polarized waves of the form:

$$\mathbf{E}_{\text{ref}} = \mathbf{E}_{0\text{ref}}\, e^{i\Phi_{\text{ref}}(r) - i\omega t}, \quad \mathbf{E}_{\text{obj}} = \mathbf{E}_{0\text{obj}}\, e^{i\Phi_{\text{obj}}(r) - i\omega t} \qquad [45]$$

The total electric field intensity $\mathbf{E}$ at the output of the interferometer, according to the principle of superposition, is given by:

$$\mathbf{E} = \mathbf{E}_{\text{ref}} + \mathbf{E}_{\text{obj}} \qquad [46]$$

The total intensity at the output, being the time average of the square of the wave amplitude, can be written as

$$I = I_{\text{ref}} + I_{\text{obj}} + I_{\text{ref/obj}} \qquad [47]$$

with $I_{\text{ref}} \sim \langle \mathbf{E}_{\text{ref}}^2 \rangle$; $I_{\text{obj}} \sim \langle \mathbf{E}_{\text{obj}}^2 \rangle$ and $I_{\text{ref/obj}} \sim 2\langle \mathbf{E}_{\text{ref}} \cdot \mathbf{E}_{\text{obj}} \rangle$. Notice that if both $\mathbf{E}$-fields are perpendicular to one another, the interference term $I_{\text{ref/obj}}$ vanishes, while if we assume that both fields are parallel, we can restrict our discussion to scalar theory. Under the latter condition we can write the total intensity as:

$$I = I_{\text{ref}} + I_{\text{obj}} + 2\sqrt{I_{\text{ref}} I_{\text{obj}}} \cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}) \qquad [48]$$

where the presence of the third term is indicative of the wave nature of light which can produce enhancement or diminution of the intensity through interference.

We can rewrite eqn [48] in Michelson's manner of writing:

$$I = I_0(1 + V\cos(\Phi_{\text{obj}} - \Phi_{\text{ref}}))$$
$$\text{where} \quad I_0 = I_{\text{ref}} + I_{\text{obj}}, \quad V = \frac{2\sqrt{I_{\text{ref}} I_{\text{obj}}}}{I_{\text{ref}} + I_{\text{obj}}} \qquad [49]$$

with V the visibility or fringe contrast indicating the degree of modulation.

We can conclude that the measured intensity of a two-beam interferogram depends on the mean intensity I_0, the visibility V, and the phase difference $\Phi_{\text{obj}} - \Phi_{\text{ref}}$. There are a number of techniques to measure the three unknowns: the mean intensity, the visibility, and the interesting spatial phase variation. In the optical instrumentation, further discussed in this article, phase shifting interferometry (PSI) is used which requires a measurement of the intensity in each point of the aperture for at least three different reference phase values, in order to obtain three equations for the three unknowns (I_0, V, and $\Phi_{\text{obj}} - \Phi_{\text{ref}}$). In practice, the reference phase values are obtained by a piezodriven transducer (PZT) mirror and should be equally spaced over one phase cycle of 2π. It has been shown that the five-phase algorithm is the optimal choice as the calculation still remains simple and the second-order nonlinear response of the detector is eliminated. A charge coupled device (CCD) sensor then records the intensity of the resulting interference pattern for different relative phase shifts and converts the intensity data to wavefront phase data through integration.

Microlens Testing

In microlens (lens with a diameter lower than several millimeters) testing, it is essential to measure in a first step the shape or the surface profile of the microlenses or in cases where gradient-index lenses are concerned, their refractive index distribution. In addition, the paraxial parameters, such as the focal length, and the more complex properties like the wave aberrations and surface quality have to be characterized. As discussed in the introduction, massive parallelism and miniaturization ask for the fabrication of microlenses and microlens arrays. The most critical features of these lenslet arrays are the uniformity of the microlenses and their mutual positioning accuracy. As a variation of these properties deteriorates the optical performance in most applications, dedicated test instrumentation is needed.

In the following, we discuss different measurement instruments designed to characterize all the geometrical and optical parameters of microlenses. First, we introduce an optical noncontact profiler which can be used to determine the diameter and the sag (the maximum height of the spherical lenscap) of the surface profile microlenses. In addition, it allows us to quantify the surface roughness on the top of the microlenses. With a transmission Mach–Zehnder (MZ) interferometer the wave aberrations and the focal length of the microlenses can be measured, while a Twyman–Green (TG) interferometer can be used for the measurement of the surface deviations of a microlens from an ideal spherical shape and for the measurement of the radius of curvature of the surface.

Surface Profile Measurements

Optical noncontact profiler

Principle. The surface profile of a microlens can be measured with a Wyko NT2000 noncontact optical profiler (Figure 19a). This surface profiler system uses two approaches to measure a wide range of surface heights. The PSI mode allows measuring fairly smooth and continuous surfaces ($0.1\ \text{nm} < \text{heights} < 160\ \text{nm}$) while the vertical scanning interferometry (VSI) mode can measure rough surfaces and heights ranging between 160 nm and 2 mm. We only discuss here the vertical scanning mode because the principle on which PSI is based has been explained above. As shown in Figure 19b, this noncontact profiler is based on a Mirau interference microscope where a white light beam passes through a beamsplitter which reflects half of the incident beam to a reference surface and transmits the other half onto the sample. The light reflected from the sample and from the reference surface then recombine at the beamsplitter to form interference fringes. The system measures the degree of the fringe modulation or the fringe contrast. Because white light has a short coherence length, interference fringes are present only over a very shallow depth for each focus position. The coherence length can be described here as the optical path distance (OPD) between two wavefronts of an interferometric system for which the phase remains correlated. Or in other words it is a measure of the range of heights over which the instrument will be able to obtain measurable interference fringes. Fringe contrast at a single sample point reaches a peak as the sample is translated through focus. During a measurement, the reference arm containing the interferometric objective moves vertically to scan the surface at varying heights. A linearized PZT precisely controls this motion. As the system scans downward, an interference signal for each point on the surface is recorded. Finally, a series of advanced computer algorithms is used to demodulate the envelope of the fringe signal and to extract the surface information.

The vertical resolution is approximately 3 nm RMS for a single measurement. The lateral resolution is a function of the total magnification, of the cursor position accuracy, and of the detector array size.

Measuring the geometrical lens dimensions. This profiler provides us with information on geometrical dimensions of microlenses. For each microlens we can derive from the contour plot (Figure 20) the sag and the diameter of the microlenses along the x and the y direction. As an example, we show in Figure 21, the

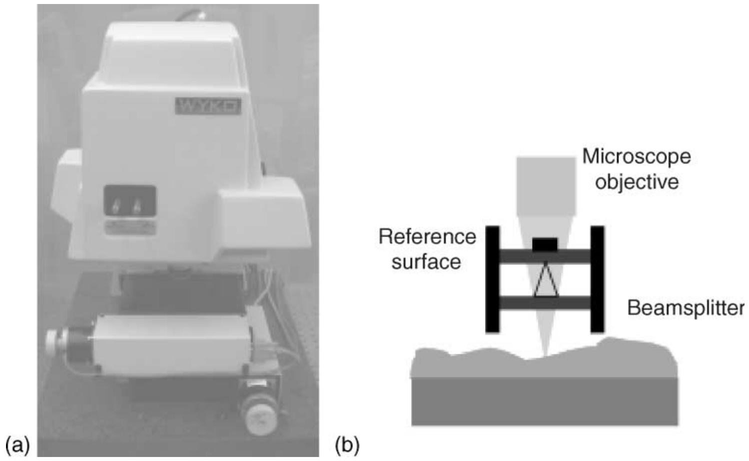

Figure 19 (a) Vertical scanning noncontact profiler (WYKO NT2000); (b) Working principle.

profiles of the cross-section of a 200 μm fused silica microlens ($NA = 0.05$; $D = 200$ μm) along the x and y directions.

The surface roughness on top of the microlens can also be derived from measurements with the noncontact profiler. As an example, we show the fringes on the vertex area (65 × 61 μm^2) of a 200 μm diameter lens as well as the average and RMS roughness on the vertex of the lens (Figure 22).

Measurement of the Optical Lens Performance

To investigate the optical properties of discrete microlenses one of the most accurate and reliable techniques is interferometry. It is well known that microlenses can be tested in transmitted light using instruments with a single pass geometry (e.g., MZ interferometer) as well as in reflected light interferometers using a double pass configuration (e.g., TG interferometer). The interferometers proposed in this section have been especially designed and constructed for microlens testing.

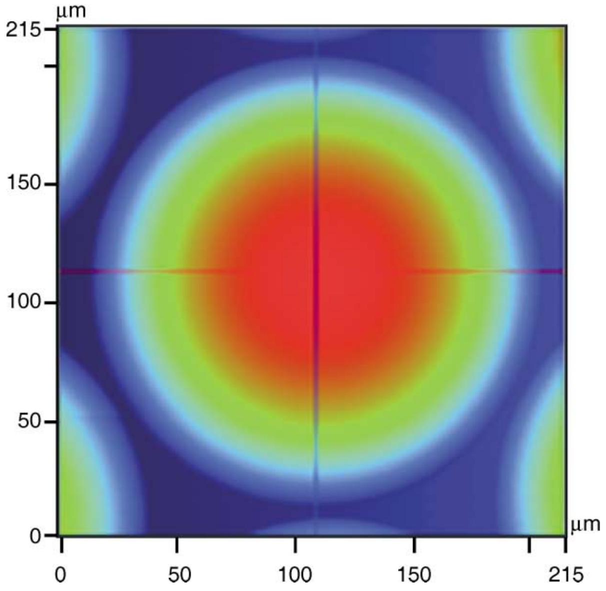

Figure 20 Contour plot of a 200 μm fused silica microlens as obtained with the noncontact profiler.

Mach–Zehnder (MZ) interferometer for transmitted light measurements

Principle. A MZ interferometer, constructed at the Erlangen–Nürnberg University, is used in this work for the measurement of microlenses in transmitted light. The interferometer is installed in a commercial Carl Zeiss Jena microscope, while fiber-optics is used for beam delivery. The scheme is shown in Figure 23. A polarizing beamsplitter (PBS) cube splits the beam of a HeNe laser while a halfwave plate at the entrance enables the balancing of the two interfering beams. The fibers are polarization preserving in order to keep the polarization in the two beams of the interferometer. The planewave of the object arm is either focused by an illumination microscope objective so that a spherical wave impinges onto the microlenses (this is the so-called spherical wave illumination) or the first microscope objective is removed so that a planewave illuminates the object (this is the so-called planewave illumination). By using a telescope consisting of the imaging microscope objective and an achromatic doublet, the expanded planewave then enters a second beamsplitter and is superimposed on a reference planewave generated by the second fiber and an expansion lens. This second beamsplitter is a nonpolarizing one. To force the two beams to interfere, the plane of polarization of the reference beam must be rotated by 90° which is achieved by twisting the fiber end accordingly. This solution has been chosen to avoid a polarizer in front of the CCD camera as it has been shown that the latter would decrease the optical quality. In addition one of the mirrors in the object

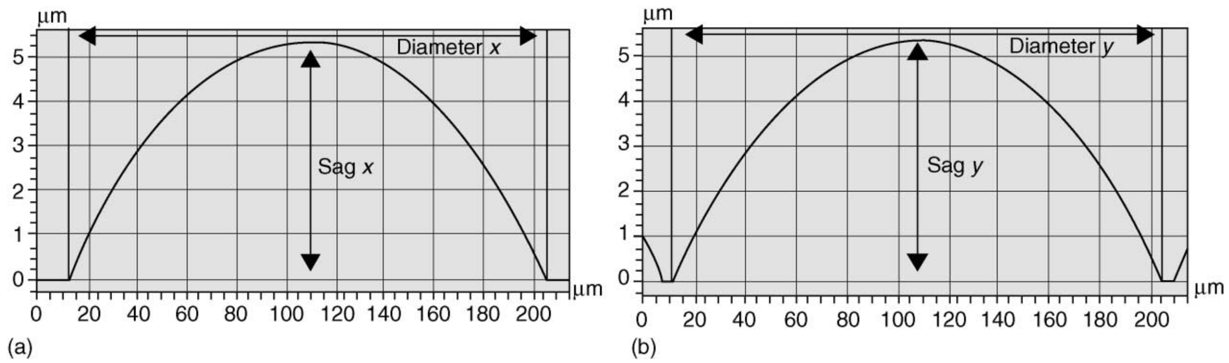

Figure 21 (a) 2D lens profile of the microlens shown in Figure 20 along the x direction and (b) along the y direction.

Roughness values	(nm)
Average roughness R_1	1.55
RMS roughness R_q	1.94

(a)

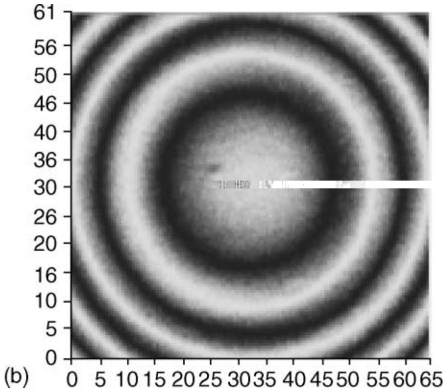

Figure 22 (a) Different roughness values on the vertex area of a 200 μm lens; (b) Fringes in this vertex area (65 × 61 μm^2).

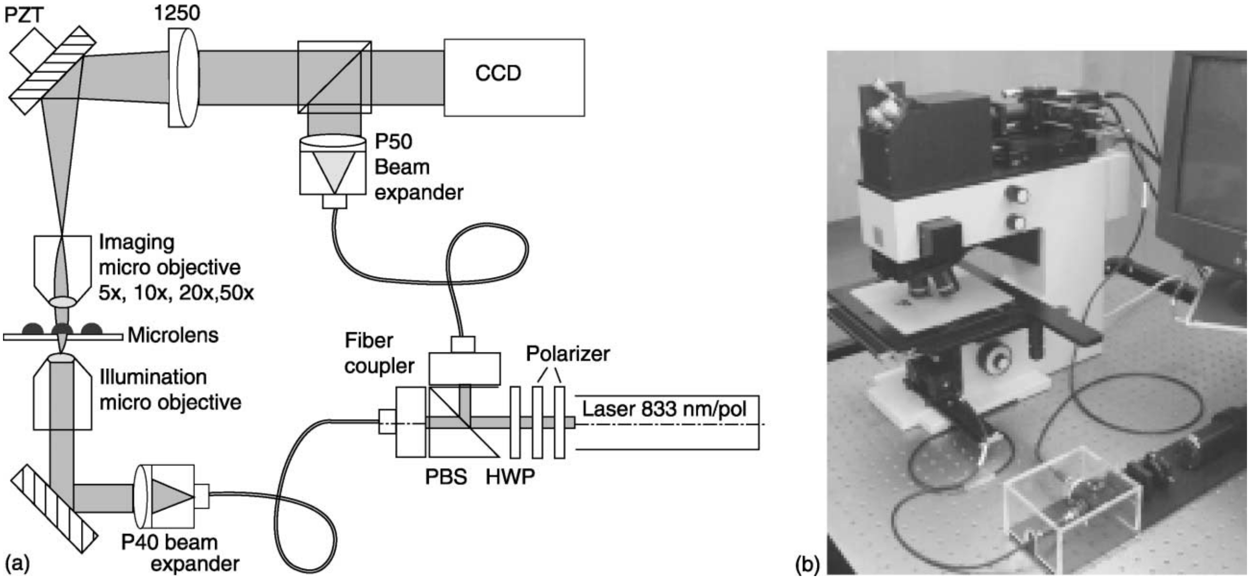

Figure 23 (a) Scheme and (b) picture of the Mach–Zehnder interferometer for measurements in transmitted light.

arm of the MZ is mounted on a PZT driver to enable phase shifting interferometry.

Focal length and array uniformity through plane-wave illumination. The MZ interferometer with planar wavefront illumination offers for slow lenses (NA < 0.1) the opportunity to measure also the total phase lag introduced by the lenses, as shown in Figure 24. This makes it possible to determine the focal length and hence the NA of the microlenses.

For certain applications of arrays of microlenses it is also important to know the uniformity of the lens array or at least of a part of the lens array. There are several different test methods (e.g., Smartt test, shearing test) which can provide a measure of the uniformity. One can even obtain the total phase retardation for a subsection of the microlens array (see Figure 25b).

Measuring the aberrations with spherical wave illumination. Wave aberrations of refractive microlenses can be determined with a spherical wavefront illumination (Figure 26). The measurement of the wave aberrations is carried out by using a PSI algorithm. This requires a measurement of the intensity in each point of the aperture for five different reference phase values. The result of a PSI evaluation is the phase distribution $\Phi(x, y)$ in the exit pupil:

$$\Phi(x, y) = \frac{2\pi}{\lambda}\psi(x, y) \qquad [50]$$

where $\psi(x, y)$ are the wave aberrations of the lens under test.

From these data, the software calculates the unwrapped phase distribution and by subtracting the tilt and defocus, the deviation from an ideal wavefront (phase/lambda) can be found. Fitting this deviation to a fourth degree Zernike polynomial then gives the aberrations (phase/lambda). Besides the aberrations we also derive from the measured phase distribution other optical lens characteristics such as the modulation transfer function and the PSF.

Finally, we emphasise that the measurements should preferably be done in a cleanroom environment as the smallest dust particle on the lens surface decreases the optical performance.

Twyman–Green interferometer

Principle. For most applications microlenses have to feature a spherical or nearly spherical shape. The measurement of the deviations of the microlens surface from an ideal sphere is therefore very important. The basic principle comes down to a planewave which is focused by a high NA microscope objective in such a way that the focus coincides with the center of curvature of the microsphere under test. The light rays are reflected, and behind the microscope objective a nearly planewave carrying the surface deviations and misalignment errors, results. The vertex of the test surface is then imaged onto a CCD camera and the nearly plane object wave interferes with the plane reference wave. The evaluation here is also done with phase shifting interferometry by axially shifting the reference mirror with a piezo transducer.

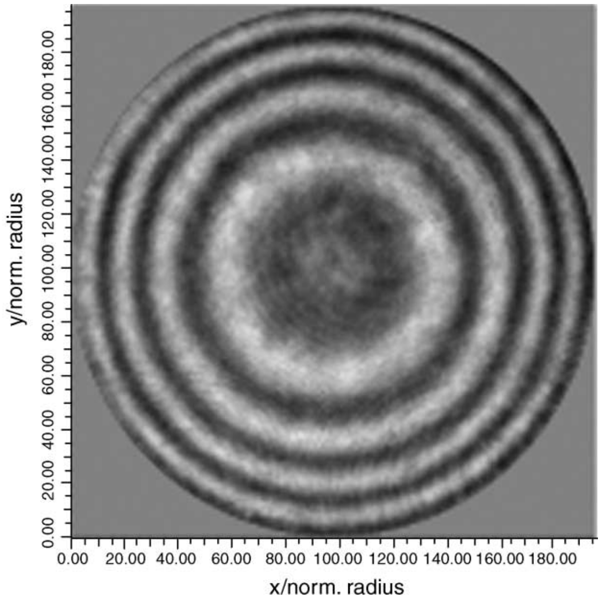

Figure 24 Interferogram of a 200 μm lens with an NA of 0.05 under planewave illumination.

The use of laser illumination allows us to obtain high contrast fringes. However, the high spatial and temporal coherence of the laser light also has important drawbacks, such as speckle noise and spurious fringes across the field. It is convenient to divide coherence into two classifications, temporal and spatial. The former relates directly to the finite bandwidth of the source, the latter to its finite extent in space. Therefore, polarization beamsplitter optics is commonly used to reduce the contribution of spurious reflections. The polarizing beamsplitter group additionally allows balancing the intensity between the object and the reference arm of the interferometer. Even so, the remaining scatter leads to poor and rough interference fringes. To further increase the fringe quality, the spatial coherence has to be reduced while the temporal coherence of the laser is preserved. This can be done by introducing a rotating scatterer in the light beam which gives a smoothing effect of the interference fringes.

Measurement of the deviations from sphericity. To determine the surface deviation of a microlens from a perfect sphere we make use of a TG

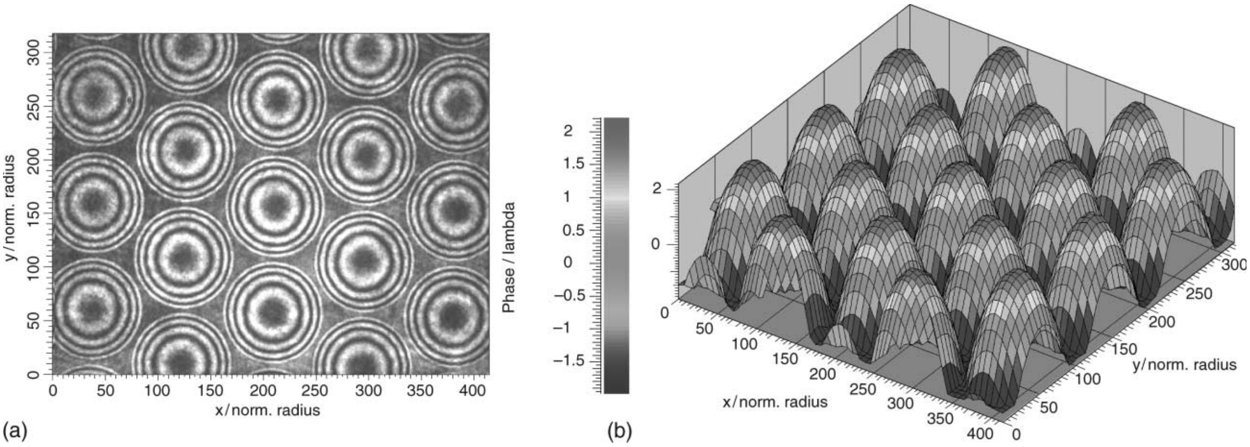

Figure 25 (a) Interferogram of an array of 200 μm diameter microlenses with an average NA of 0.05 and illuminated with a plane wavefront; (b) 3D representation of the phase lag (RMS 1.26λ, *PV* 4.37λ) for the same microlens array.

interferometer, constructed by the Erlangen-Nürnberg University (Figure 27). In this interferometer, the light of a HeNe laser is imaged onto a rotating scatterer by a microscope objective and the scattered light is collimated by an achromat. The collimated laser beam then passes a field stop which is sharply imaged by a 4f system onto the lens under test. The image of the field stop can be focused onto the surface under test by moving the field stop along the optical axis. Once the image of the field stop is sharply imaged onto the CCD array, one can translate the reference mirror until the reference image of this stop in its turn becomes sharp onto the CCD camera. Thus, two conditions have to be met to achieve high contrast fringes: the surface under test and the image of the field stop of both arms of the interferometer have to be sharp on the detector. The half wave plate in front of the PBS adjusts the splitting ratio at the PBS while the polarizer after the beamsplitter is necessary to force the two orthogonally polarized beams to interfere. Finally, the optics at the output of the interferometer gives a sharp image of the intermediate image of the lens surface onto the CCD camera. Analogous to the MZ interferometer, five phase images are grabbed and subjected to different algorithms, leading to the desired information.

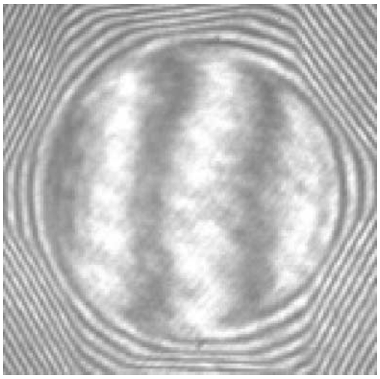

Figure 26 Interferogram of a microlens under spherical wavefront illumination: the resulting planewave is a measure for the wave aberrations.

A condensor objective transforms the radius of curvature of the spherical wavefront such that it matches the radius of curvature of the microlens. Hence, the wavefront impinges everywhere perpendicularly on the lens surface. This means that the sensitivity to surface deviations is independent on the surface position leading to high accuracy measurements. The TG interferometer allows us to test the sphericity of microlenses with diameters ranging

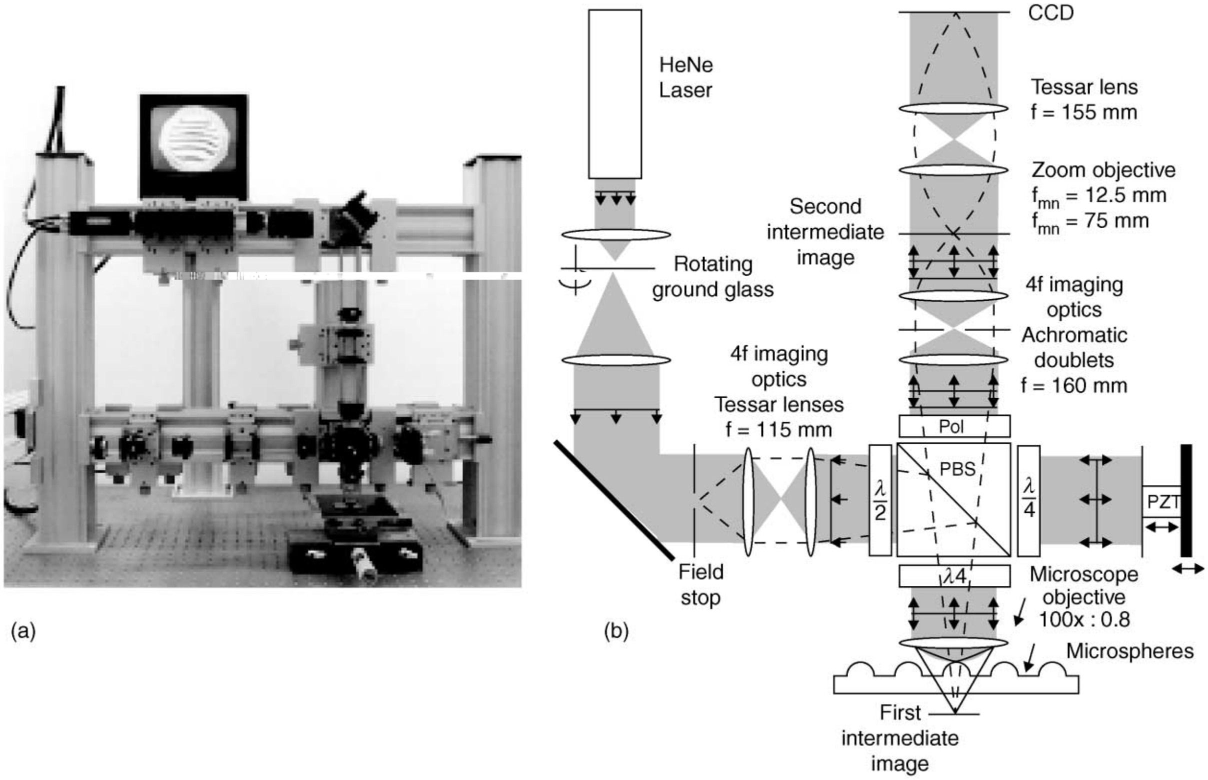
(a)
(b)
HeNe Laser
Rotating ground glass
4f imaging optics
Tessar lenses
f = 115 mm
Field stop
Second intermediate image
CCD
Tessar lens
f = 155 mm
Zoom objective
f_{mn} = 12.5 mm
f_{mn} = 75 mm
4f imaging optics
Achromatic doublets
f = 160 mm
Pol
PBS
λ/2
λ/4
PZT
λ4
Microscope objective
100x : 0.8
Microspheres
First intermediate image

Figure 27 (a) Twyman–Green interferometer setup. (b) Detailed schematic overview of this interferometer.

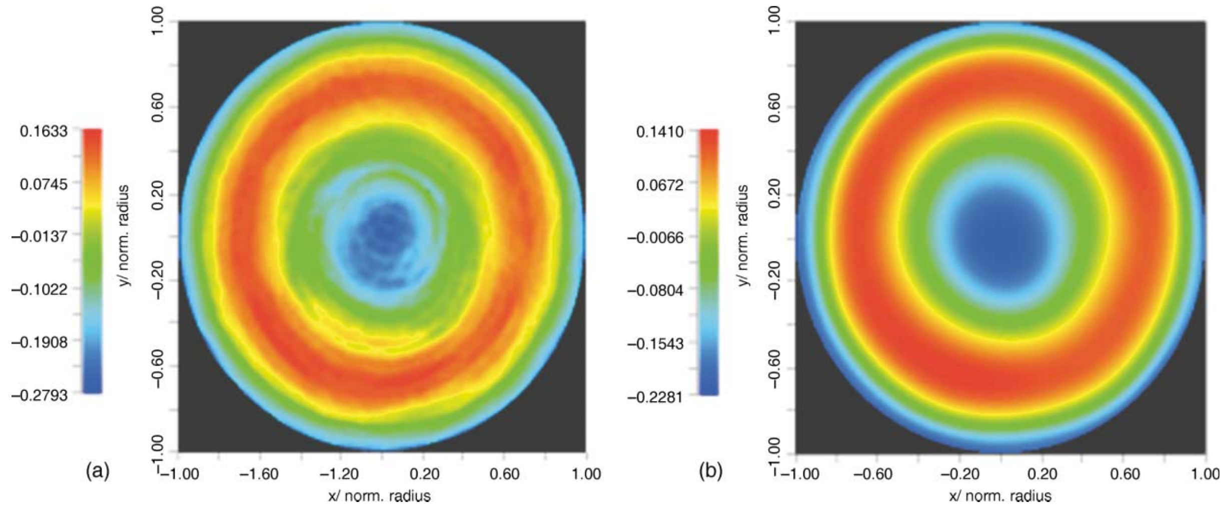

Figure 28 (a) Measured deviation from an ideal sphere (phase/lambda) for a 200 μm lens with an NA of 0.05 as obtained with a Twyman–Green interferometer. (b) Polynomial fitted surface deviation (phase/lambda) of this microlens.

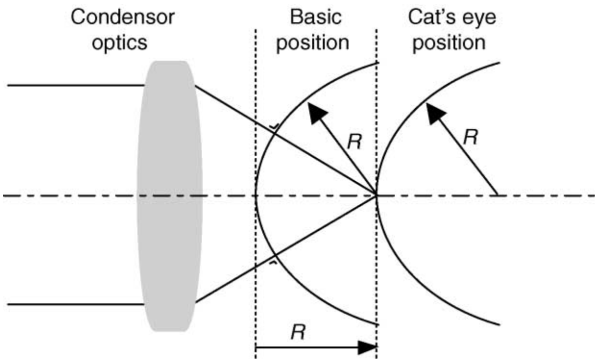

Figure 29 Necessary adjustments of the spherical microlens for radius measurements.

from 50 to 1000 μm and with a maximum radius of curvature of 1960 μm.

An example of a 200 μm fused silica microlens, with an NA of 0.05, is shown in Figure 28. The deviation from an ideal sphere can be derived from the unwrapped phase after subtracting the tilt and defocus. Fitting the deviation to a fourth degree Zernike polynomial then gives the surface deviation.

Measuring the radius of curvature. Using the TG interferometer it is also possible to determine the radius of curvature R of a microlens. To do this, the test object has to be shifted axially from the basic position, where the light rays impinge perpendicularly onto the surface and the focal length of the condensor objective coincides with the center of the sphere under test, to the so-called cat's eye position, where the focus of the impinging spherical wave is at the vertex of the sphere (Figure 29). For both positions straight, parallel, and equidistant interference fringes are obtained, only perturbed by aberrations due to the lens surface and the condensor objective. This forms the criterion for the adjustment of the two positions. The difference between these two positions then gives the radius of the microsphere. It should be mentioned that the cat's eye position causes an inversion of the object wavefront with respect to the reference wavefront, which results in a vanishing contrast with the partially coherent illumination. The latter can only be avoided by providing the necessary coherence. This implies that the rotating scatterer will have to be removed.

Summary of the Interferometric Instrumentation

Figure 30 summarizes the various instruments we discussed in this section: a noncontact optical profiler, a MZ and a TG interferometer. For each of these instruments, we highlight the measurands, the assumptions that are made, and the resulting optical lens characteristics. From this figure, it is clear that one can obtain a measure for the focal length and hence the numerical aperture and the focal number of a refractive microlens, with all three instruments. However, only the focal length obtained with the MZ interferometer is a direct measurement. The focal length obtained with the optical profiler is derived from the sag of the lens under the assumption of an ideal sphere while the radius of curvature measured with the TG interferometer allows to calculate the focal length as well. Nevertheless, the three values make an interesting comparison possible.

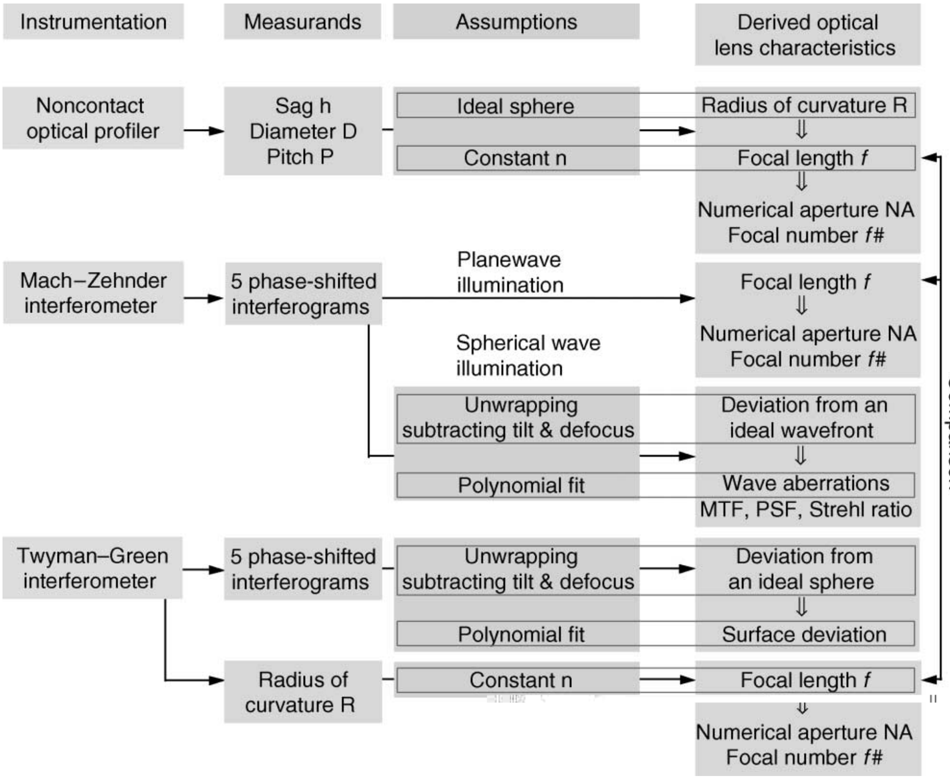
Instrumentation
Measurands
Assumptions
Derived optical lens characteristics
Noncontact optical profiler
Sag h
Diameter D
Pitch P
Ideal sphere
Constant n
Radius of curvature R
⇓
Focal length *f*
⇓
Numerical aperture NA
Focal number *f*#
Mach–Zehnder interferometer
5 phase-shifted interferograms
Planewave illumination
Focal length *f*
⇓
Numerical aperture NA
Focal number *f*#
Spherical wave illumination
Unwrapping subtracting tilt & defocus
Polynomial fit
Deviation from an ideal wavefront
⇓
Wave aberrations
MTF, PSF, Strehl ratio
Comparison
Twyman–Green interferometer
5 phase-shifted interferograms
Unwrapping subtracting tilt & defocus
Polynomial fit
Deviation from an ideal sphere
⇓
Surface deviation
Radius of curvature R
Constant n
Focal length *f*
⇓
Numerical aperture NA
Focal number *f*#

Figure 30 Overview of the instruments that we will use to quantitatively characterize the refractive microlenses: the measurands and the resulting optical lens characteristics.

Conclusions

In this article we introduced geometrical optics to discuss the imaging properties and the aberrations of refractive lenses, while wave optics was used to explain the effects of diffraction on the imaging properties of lenses. Next we put forward some commonly used quality criteria to assess lens performance and we pointed out the unique properties of microlenses. Finally, interference was briefly introduced to clarify the working principle of various instruments to quantitatively characterize refractive microlenses and microlens arrays: a noncontact profiler, a transmission MZ interferometer and a TG interferometer.

We can conclude that micro-optics is a key technology with many applications in various areas of tomorrow's information society. It enables new functionalities and applications previously unachievable with conventional optics and improves product performance, stability and robustness while reducing cost, volume, and weight. While in conventional optics discrete lenses are mounted to form optical systems using mechanical alignment equipment, for microlenses new integration approaches are available. For most of the applications, the integration of micro-optical components into fully functional systems is very important.

Thus micro-optics is a generic technology that allows the manipulation of light with micron and submicron scale structures and components. Micro-optics is therefore the technology to help provide the necessary tools to interface the macroscopic world we live in with the microscopic world of opto- and nano-electronic data processing. Micro-optic processing and fabrication technologies can be adapted to a wide variety of specialty optical materials such as plastics, polymers, semiconductors, sol-gels, and doped glasses. The application of these modified materials permits the widening of the field of photonic functionalities to virtually any region of the optical spectrum and beyond, such that a wealth of novel optical and photonic applications can be realized. Micro-optic processing technologies can be made compatible with scalable industrial mass-manufacturing and replication techniques.

Acknowledgment

This research has been made possible through the financial support of the Fund for Scientific Research-Flanders (FWO-Vlaanderen), the IAP Photon

Network, the Bilateral Scientific and Technological Cooperation Flanders-Poland, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the IWT GBOU project ‘Generic Technologies for Plastic Photonics’, the Research Council (OZR) of the Vrije Universiteit Brussel and the EC Network of Excellence NEMO.

List of Units and Nomenclature

[AB]	optical path length
c	speed of light
d	lens thickness
D	lens aperture diameter
$\mathbf{E}$	electricfield
f	focal length
f_{BFL}	back focal length
f_{FFL}	front focal length
$f\#$	focal number
F_{i}	image focal point
F_o	object focal point
h_1	position of the first principal point relative to the vertex V1
h_2	position of the second principal point relative to the vertex V2
h'	deviation from the optical axis
H_1	first principal point
H_2	second principal point
I	intensity
k	wave number in vacuum
n	index of refraction
n_{i}	index of refraction of medium i
O	object point
P	arbitrary field point
P	image point
P	lens pitch
$\wp$	spherical surface power
r	aperture of the refracting surface
R	radius of curvature
$R(x, y)$	real wavefront
S	axial object point
S	Strehl ratio
s_0	object distance
s_i	image distance
V	lens vertex
V	visibility or fringe contrast
$W(x, y)$	spherical wavefront in the Gaussian approximation
y_{o}	object height
y_{i}	image height
z_{R}	Rayleigh range
α	angular semi-aperture
θ	symmetry around the optical axis
λ	wavelength of light
σ	standard deviation
φ_1	angle between the incident ray and the normal to the flat surface
φ_2	angle between the refracted ray and the normal to the flat surface
$\Phi_{\mathrm{obj}} - \Phi_{\mathrm{ref}}$	phase difference
$\Phi(x, y)$	phase distribution
$\psi(x, y)$	wavefront aberration
ω	angular frequency
Ω	aperture in a plane opaque screen
BFL	Back focal length
CCD	Charge coupled device
FF	Far-field
FFL	Front focal length
MTF	Modulation transfer function
MZ	Mach–Zehnder
NA	Numerical aperture
NF	Near-field
OPD	Optical path difference
OTF	Optical transfer function
PBS	Polarizing beamsplitter
PP	Principal plane
PSF	Point spread function
PSI	Phase shifting interferometry
PTF	Phase transfer function
PZT	Piezo-electric transducer
RMS	Root mean square
SA	Spherical aberration
VSI	Vertical scanning interferometry

See also

Diffraction: Fraunhofer Diffraction. **Diffractive Systems:** Aberration Correction with Diffractive Elements. **Geometrical Optics:** Aberrations; Lenses and Mirrors. **Interferometry:** Overview; Phase Measurement Interferometry.

Further Reading

Born M and Wolf E (1999) *Principles of Optics*, 7th edn. Cambridge, UK: Cambridge University Press.

Hecht E (1987) *Optics*. San Francisco, CA: Addison-Wesley Publishing Company.

Herzig HP (1997) *Micro-Optics: Elements, Systems and Applications*. London: Taylor & Francis.

Jenkins FA and White HE (1981) *Fundamentals of Optics*, 4th edn. New York: McGraw-Hill.

Malacara D and Malacara Z (1994) *Handbook of Lens Design*. New York: Marcel Dekker.

Malacara D, Servin M and Malacara Z (1998) *Interferogram Analysis for Optical Testing*. New York: Marcel Dekker.

Mouroulis P and Macdonald J (1997) *Geometrical Optics and Optical Design*. New York: Oxford University Press.
Pedrotti FL and Pedrotti LS (1993) *Introduction to Optics*, 2nd edn. Englewood Cliffs: Prentice-Hall International.
Shannon RR (1997) *The Art and Science of Optical Design*. Cambridge, UK: Cambridge University Press.
Siegman AE (1986) *Lasers*. Sausalito: University Science Books.
Sinzinger S and Jahns J (1999) *Microoptics*. Weinheim, Germany: Wiley-VCH GmbH.
Smith WJ (1997) *Practical Optical System Layout and Use of Stock Lenses*. New York: McGraw-Hill.
Smith W (2002) Engineering an optical system. *SPIE OE Magazine* 7: 49–54.
Young HD (1992) *Physics*, 8th edn. San Francisco, CA: Addison-Wesley Publishing Company.

OPTICAL PARAMETRIC DEVICES

Contents

Overview
Optical Parametric Oscillators (Continuous Wave)
Optical Parametric Oscillators (Pulsed)

Overview

B J Orr, Macquarie University, Sydney, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Many forms of optical parametric device (OPD) play significant roles as three-wave nonlinear-optical sources of tunable coherent light for laser-based science and technology. Any such device will fall into one of three categories, in increasing order of sophistication: optical parametric generator (OPG); optical parametric amplifier (OPA); and optical parametric oscillator (OPO). This article concentrates largely on nanosecond-pulsed OPOs, with particular emphasis on designs that facilitate their wavelength control and their optical power- or energy-efficiency. The two other important temporal manifestations of OPOs (ultra-fast pulsed and continuous-wave) are briefly outlined. OPDs serve as versatile sources of tunable, intense, coherent radiation covering the infrared, visible, or ultraviolet spectral regions. Their most obvious uses are associated with laser wavelength-conversion and/or tunability, resulting in a variety of spectroscopic applications. Four-wave nonlinear-optical parametric contributions, such as those involved in stimulated Raman scattering and highly nonlinear optical fibers, are also briefly considered.

What is an Optical Parametric Device?

An optical parametric device (OPD) is a source of coherent, laser-like radiation that is typically intense and tunable over a wide range of wavelengths. Its mechanism invariably depends on nonlinear optics, usually through a three-wave mixing process mediated by the nonlinear-optical susceptibility $\chi^{(2)}$ in a noncentrosymmetric crystalline medium. Three forms of OPD are illustrated in Figure 1, and also illustrated is a closely related (but distinct) nonlinear-optical device. Coherent light waves are represented by arrows, with their associated optical frequency ω_j and wave vector $\mathbf{k}_j$ (which are defined below). In Figure 1, input and output waves are shown as arrows on the left and right, respectively, with their breadth intended to indicate typical relative intensities.

An optical parametric generator (OPG) is the simplest form of OPD, with a single input wave (pump, P) from a laser generating two coherent output waves (signal, S; idler, I; $\omega_S \geq \omega_I$), as depicted in Figure 1a. The nonlinear-optical process itself is initiated by low-intensity parametric spontaneous emission/noise/fluorescence processes that occur naturally, effectively ‘splitting’ a pump photon into two new photons. Once a signal and/or idler wave has been generated, it can be coherently amplified by passing it through an optical parametric amplifier (OPA) together with input pump radiation, as depicted in Figure 1b. A further order of sophistication is reached in an optical parametric oscillator (OPO), as depicted in Figure 1c, where the

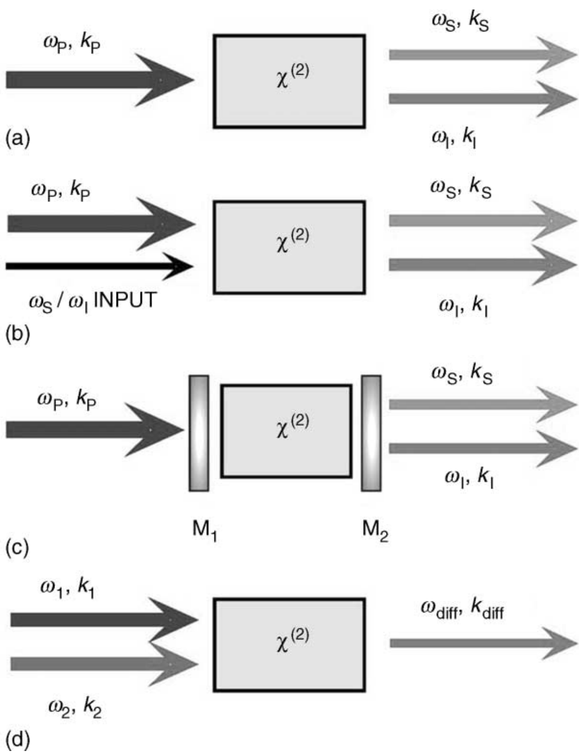

Figure 1 Schematic diagrams of three forms of optical parametric device (OPD): (a) optical parametric generator (OPG); (b) optical parametric amplifier (OPA); (c) optical parametric oscillator (OPO). Note that, by convention, optical frequencies of the signal (S) and idler (I) output waves are defined such that $\omega_S \geq \omega_I$. Also shown is a fourth closely related device: (d) difference-frequency generator (DFG). Nonlinear-optical media are denoted by their susceptibility $\chi^{(2)}$. Arrows are used to represent input and output waves, together with their optical frequencies ω_j and wave vectors $\mathbf{k}_j$. An OPO requires an optical resonator, comprising at least two aligned reflectors (M_1, M_2).

functions of an OPG and an OPA are combined by multipassing the optical waves involved inside a resonant optical cavity, formed by two or more appropriately aligned reflectors (M_1, M_2).

A difference frequency generator (DFG), as depicted in Figure 1d, is not an OPD, although the DFG source term is central to the nonlinear-optical mechanism of an OPD. In a DFG, two input waves (labeled 1 and 2) are combined coherently to generate a third output wave at the difference frequency of the two input waves: there are now two relatively high-power driving waves, rather than one, and photons are effectively 'joined', rather than 'split' as in the OPD case. Nevertheless, the outcome and utility of a DFG can be similar to that of an OPD. For instance, if coherent radiation is required at a particular infrared (IR) wavelength, it can be generated either as the idler wave of an OPD, with frequency $\omega_I = (\omega_P - \omega_S)$ *or* as the output wave of a DFG, with frequency $\omega_3 = |\omega_1 - \omega_2|$.

Many of the desirable attributes of OPDs in general, and OPOs in particular, arise from the fact that any such instrument is a nonlinear-optical device, not a laser. This allows versatile, flexible design features, such as methods of temporal and wavelength control that are not possible with lasers. The latter generally depend on population inversion of an optical gain medium, with associated optical lifetime and saturation limitations. On the other hand, optical parametric gain, oscillation and amplification are more amenable to modular system design because (as is explained below) they depend on nonlinear-optical coefficients and phase-matching conditions.

Nonlinear-Optical Origins of OPDs

Nonlinear optics involves interaction of a number ($\sigma > 2$) of optical waves in a medium with a nonlinear-optical susceptibility tensor $\chi^{(\sigma-1)}$. For inelastic optical processes, these waves (each with an angular frequency $\omega_1, \omega_2, \ldots, \omega_\sigma$) obey two conservation conditions, one for energy (or frequency):

$$\omega_1 + \omega_2 + \ldots + \omega_\sigma = 0 \qquad [1]$$

The other conservation condition (expressed in terms of wave vectors $\mathbf{k}_i$, with $i = 1, 2, \ldots, \sigma$, that have magnitudes $k_i = n_i\omega_i/c = 2\pi n_i/\lambda_i$, where n_i is the refractive index at vacuum wavelength λ_i and c is the speed of light) is for momentum:

$$\mathbf{k}_1 + \mathbf{k}_2 + \ldots + \mathbf{k}_\sigma + \Delta\mathbf{k} = 0 \qquad [2]$$

where $\Delta\mathbf{k}$ is the phase-mismatch between the σ interacting waves. Each frequency component and wavevector is ascribed a positive or negative sign, according to their phase relationships. Equation [2] defines phase-matching conditions in which $\Delta\mathbf{k}$ must be minimized to optimize the efficiency of the nonlinear-optical process of interest.

We shall consider below the two specific three-wave nonlinear-optical processes that are particularly relevant to this article, namely, those for an OPD (i.e., OPG, OPA, or OPO) or a DFG. Each of these is mediated by the nonlinear-optical susceptibility tensor $\chi^{(2)}$, which is nonzero in a crystalline medium only if it is noncentrosymmetric. Many such crystals are available. For example, lithium niobate ($LiNbO_3$) has been popular since the early days of pulsed tunable OPOs. Subsequent interest and activity in OPD technology have been stimulated by the availability of new nonlinear-optical materials such as BBO (β-barium borate, BaB_2O_4) and KTP (potassium titanyl phosphate, $KTiOPO_4$).

Additional recent impetus has come from quasi-phase-matched (QPM) nonlinear-optical media, such as periodically poled lithium niobate (PPLN) and PPKTP, tailored for specific wavelengths by periodic optical structuring of domains. QPM media offer compact, efficient, low-threshold alternatives to conventional birefringently phase-matched (BPM) media. Characteristics of BPM and QPM nonlinear-optical crystalline media are accessible *via* Arlee Smith's SNLO software package, accessible by internet at www.sandia.gov/imrl/XWEB1128/xxtal.htm.

For a three-wave OPD, which is of principal interest in this article, the energy and momentum conservation conditions of eqns [1] and [2] become:

$$\omega_P - \omega_S - \omega_I = 0, \qquad \mathbf{k}_P - \mathbf{k}_S - \mathbf{k}_I - \Delta\mathbf{k} = 0 \qquad [3]$$

where a laser input wave ('pump', frequency ω_P, wave vector $\mathbf{k}_P$) yields two coherent output waves ('signal', frequency ω_S, wave vector $\mathbf{k}_S$; 'idler', frequency ω_I, wave vector $\mathbf{k}_I$; $\omega_S \geq \omega_I$). Note that the idler frequency ω_I equals the difference $|\omega_P - \omega_S|$ between pump and signal frequencies. Equation [3] should be viewed in the context of sections (a–c) of Figure 1.

Equation [2] and the second half of eqn [3] apply strictly only to the conventional case of BPM media. In such media, the phase-matching condition $\Delta\mathbf{k} \approx 0$ is achieved by adjusting the angle and/or temperature of a birefringent nonlinear-optical crystal *via* its ordinary- and extraordinary-ray refractive indices. Such adjustments are used to optimize OPD conversion efficiency and thereby control the output signal and idler wavelengths, λ_S and λ_I. If it is assumed that the three waves are collinear and that $\Delta\mathbf{k}$ is exactly zero, then the signal frequency/wavelength is given simply in terms of the pump frequency/wavelength and the refractive indices n_i $(i = P, S, I)$ as:

$$\begin{aligned} \omega_S &= \omega_P(n_P - n_I)/(n_S - n_I); \\ \lambda_S &= \lambda_P(n_S - n_I)/(n_P - n_I) \end{aligned} \qquad [4]$$

An alternative to birefringent phase matching is the quasi-phase-matching approach first suggested by Nobel Laureate Nicolaas Bloembergen and co-workers in 1962. This approach has been realized experimentally only in the last few years, *via* QPM nonlinear-optical media such as PPLN or PPKTP. Equation [2] and the second half of eqn [3] need to be slightly modified to include QPM grating contributions.

In the corresponding case of a DFG (which, we repeat, is not an OPD), the energy and momentum conservation conditions of eqns [1] and [2] become:

$$\begin{aligned} |\omega_1 - \omega_2| - \omega_{\text{diff}} &= 0, \\ \mathbf{k}_1 - \mathbf{k}_2 - \mathbf{k}_{\text{diff}} - \Delta\mathbf{k} &= 0 \end{aligned} \qquad [5]$$

where two laser input waves (frequencies ω_1, ω_2; wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$) yield one coherent output wave at the difference frequency $\omega_{\text{diff}} = |\omega_1 - \omega_2|$ (with wave vector $\mathbf{k}_{\text{diff}}$), as depicted in Figure 1d. As above, phase matching is defined by $\Delta\mathbf{k} \approx 0$ for BPM media and, in the case of QPM media, there is an additional grating contribution in the second half of eqn [5].

Finally, we mention briefly two other important forms of nonlinear-optical device arising from four-wave mixing processes that are mediated by the nonlinear-optical susceptibility $\chi^{(3)}$ which can be nonzero, even in isotropic or centrosymmetric media such as gases, liquids, optical fibers, and all classes of crystal. Optical parametric processes of this type contribute to stimulated Raman scattering (SRS), involving an optical medium with Raman-active resonance frequencies ω_R that coincide with the difference between two optical frequencies. This is particularly useful in providing a relatively straightforward source of coherent radiation, Raman-shifted at discrete intervals from the frequency ω_L of an input pump laser (either tunable or fixed-wavelength). Four-wave parametric processes are involved, such as that in eqn [1] with $|\omega_1| = |\omega_3| = \omega_L, |\omega_2| = \omega_L - \omega_R$, and $|\omega_4| = |\omega_1| + |\omega_3| - |\omega_2| = \omega_L + \omega_R$, where ω_R is the characteristic Raman frequency of the nonlinear-optical medium used in the Raman shifter (e.g., high-pressure hydrogen, deuterium or methane gas). These Raman-shifted intervals, both added to (anti-Stokes) and subtracted from (Stokes) the laser frequency ω_L, are integer multiples of ω_R. Other nonlinear-optical Raman parametric processes give rise to various forms of nonlinear Raman spectroscopy, such as CARS (coherent anti-Stokes Raman spectroscopy), and to Raman fiber-optical amplifiers, used in optical telecommunications.

Another promising area of developing OPD technology entailing $\chi^{(3)}$ nonlinearities concerns OPGs, OPAs, and OPOs based on highly nonlinear optical fibers, with either pulsed or cw pump lasers. Such processes use two pump waves (P) to generate tunable signal (S) and idler (I) output waves, such that $\omega_I = 2\omega_P - \omega_S$.

Nanosecond-Pulsed OPOs – Design and Wavelength Control

Here we consider pulsed OPOs that operate on nanosecond (ns, 10^{-9} s) time-scales, and design

Table 1 Some elements of OPO applications to spectroscopy

Key properties	*Options*	*Comments on instrument/technique*
Wavelength range: what forms of spectra need to be measured?	UV/visible (0.2–0.7 μm) Near IR (0.7–4.0 μm) Longwave IR (>4.0 μm)	Many nonlinear-optical OPO materials are available, but less well developed for longwave IR applications
OPO phase matching: BPM or QPM?	Birefringent phase matching (BPM) Quasi-phase matching (QPM)	BPM is well established and preferred for high-power operation. Various low-threshold QPM media are now available for both pulsed and cw OPOs
Temporal: continuous-wave (cw) or pulsed?	Pulsed for power and timing Cw for narrowest bandwidth	Ultrafast (ps, fs) OPO output pulses are relatively broadband, owing to the Fourier transform limit
Optical bandwidth: broad or narrow?	Broadband (free-running) Single longitudinal mode	The Fourier transform limit is 44 MHz (0.0015 cm^{-1}) for an ideal 10 ns pulse Cw OPOs offer even lower $\Delta\nu$. Ultrafast-pulsed OPOs are inevitably broadband: e.g., $\Delta\nu \geq 0.44$ THz ($\geq$15 cm^{-1}) for a 1 ps pulse
Mode of recording of spectra: continuous tuning or multiplex?	Scan narrowband signal/idler OPO output wavelength Free-running OPOs operate broadband in multiplex case	Wavelength control yields continuously tuned narrowband spectra. Multiplex spectra use dispersed detection or multi-wavelength spectroscopic tailoring
Wavelength control: intracavity elements or injection seeding?	Intracavity gratings or étalons Injection seeding of signal or idler by a tunable low-power coherent light source	Intracavity-element designs yield broad tunability but can be complicated. Injection seeding facilitates narrowband, mode-hop-free spectra and tailored multi-wavelength experiments

The above options determine ways to use OPOs for spectroscopic applications, such as: (1) linear absorption (e.g., with a multipass cell); (2) cavity ringdown (CRD) absorption spectra; (3) high-resolution spectra ($\Delta\nu \approx$ MHz or kHz); (4) nonlinear-optical (e.g., coherent Raman); (5) atmospheric remote sensing (e.g., DIAL); (6) fast (μs, ns) and ultrafast (ps, fs) processes.

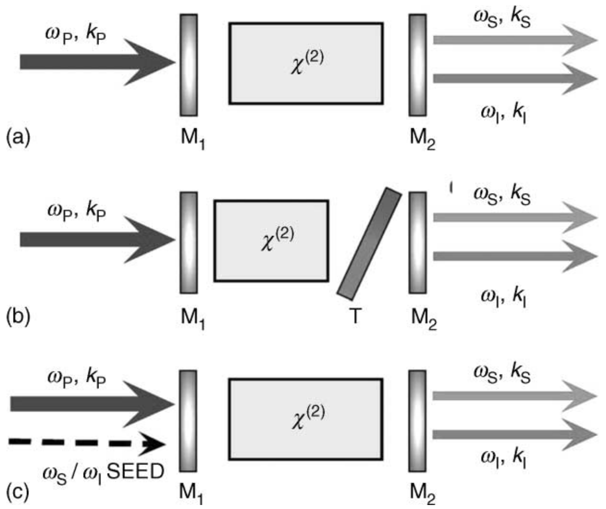

Figure 2 Schematic diagrams of three forms of optical parametric oscillator: (a) free-running OPO (with no active wavelength control), similar to **Figure 1c**; (b) OPO with an intracavity tuning element (T); (c) injection-seeded OPO.

features that make them fit for one of the principal applications of OPDs, namely, spectroscopy. Options that determine ways to use OPOs for spectroscopic applications are summarized in Table 1. Some design and wavelength-control features used in ns-pulsed OPOs are depicted schematically in Figure 2. Some of these applications and features are shared with other forms of OPO, notably the ultrafast-pulsed OPOs and cw OPOs that are considered much more briefly in later sections.

The first OPO, demonstrated by Giordmaine and Miller in 1965, was ns-pulsed; it was based on $LiNbO_3$, tunable over the range ~0.96–1.16 μm, spanning a signal- and idler-wavelength range of ±0.1 μm on either side of the degeneracy point defined by the 529 nm pump radiation (from a frequency-doubled, Q-switched Nd:$CaWO_4$ laser). This advance occurred soon after lasers were discovered and the potential of nonlinear optics had been realized through harmonic-generation processes such as frequency doubling. Pioneering work by Stephen Harris, Robert Byer and their co-workers at Stanford University soon established ns-pulsed OPOs as practical sources of tunable coherent light for significant applications such as spectroscopic sensing of chemical processes, in industrial or environmental diagnostics and in basic science. Such ns-pulsed OPOs had an impressive impact in the 1970s, followed by sporadic progress in much of the 1980s, and a dramatic resurgence of interest and activity since the late 1980s.

This irregular rate of progress is attributable to the availability of suitable high-quality nonlinear-optical crystals (as mentioned above), as well as high-performance pump lasers and advanced tunable OPO system designs.

In a typical ns-pulsed OPO, the pump laser delivers sufficient optical energy to enable parametric gain to build up from noise during the pump–pulse duration, exceeding the threshold for oscillation. To maximize gain, the parametric (signal and/or idler) waves are amplified by multipassing them during the pump pulse in an optical resonator, as depicted in Figures 1c and 2. Light travels ~3 m during a 10 ns pump pulse, so that the parametric waves can make ~15 round trips of a 10 cm linear OPO cavity.

A simple free-running OPO, comprising an optical cavity with input and output mirrors M_1 and M_2 but with no wavelength-selective elements, is depicted in Figure 2a. The output radiation from such an OPO generally has a relatively broad optical bandwidth – typically ~5 cm^{-1} (~150 GHz) or more – depending on several factors: refractivity, dispersion, and absorption of the OPO medium; wavelengths λ_P, λ_S, and λ_I; type of phase matching (BPM or QPM, whether collinear or not); crystal dimensions and orientation; cavity reflectivity and effective number of passes of the resonated wave; optical bandwidth, divergence, pulse duration and pulse energy of the pump radiation. Free-running pulsed OPOs represent one extreme of operational simplicity, yielding relatively broadband output radiation suitable for low-resolution or multiplex spectroscopy. Additional OPO wavelength-control measures are usually necessary for higher-resolution spectroscopic applications.

At the other extreme of operational complexity, intracavity wavelength-selective elements, such as gratings and/or étalons, provide a traditional way to control OPO output wavelength and to achieve narrow optical bandwidths. One such approach is depicted in Figure 2b, where a tuning element T (in this case, a tilted étalon or filter; in other designs, an intracavity diffraction grating replacing the output cavity mirror M_2) is inserted in the cavity. Such approaches were used in early pulsed $LiNbO_3$ OPO designs that were continuously tunable in the near IR with an optical bandwidth of ~0.1 cm^{-1} (~3 GHz), but these were often difficult to operate and were damage-prone (with intracavity losses from gratings and étalons causing the operating threshold to approach the damage threshold of OPO nonlinear-optical materials such as $LiNbO_3$).

Continuous narrowband tunability, preferably on a single-longitudinal-mode (SLM) basis, is a desirable performance characteristic of pulsed OPOs amenable to high-resolution spectroscopy. A commercially viable approach to this ideal was achieved by Walter Bosenberg and Dean Guyer in their advanced KTP OPO/OPA system, which includes a master OPO stage with a grazing-incidence grating. This is continuously tunable under computer control in the near IR (1.3–4 μm) with narrow optical bandwidth – ~0.02 cm^{-1} (~600 MHz) or better. Such a high-performance tunable coherent IR source has been used for various forms of laser spectroscopy (cavity ringdown, degenerate four-wave mixing, long-path absorption, etc.), including investigations of chemically reactive media, combustion diagnostics, and studies of processes in molecular beams.

A popular alternative approach to OPO wavelength control is injection seeding by a low-power tunable coherent source such as a tunable diode laser (TDL). This approach is depicted in Figure 3; in practice, a ring cavity is often used to avoid

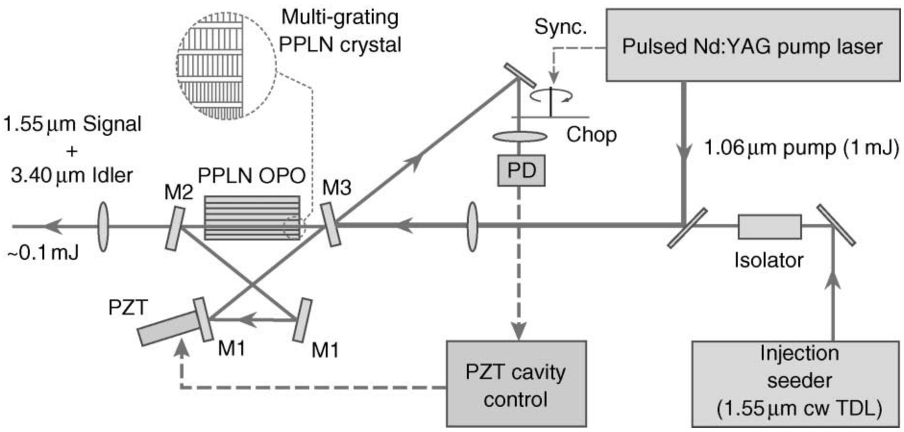

Figure 3 Schematic diagram of an injection-seeded tunable PPLN OPO system with active cavity control, developed in the author's laboratory. Legend: PD = photodetector, PZT = piezoelectric translator, TDL = tunable diode laser, M_{1-3} = cavity reflectors; the inset shows the QPM multi-grating structure of the PPLN nonlinear-optical crystal.

feedback from the OPO to the seed laser. A significant advantage of injection seeding is that OPO construction is simplified by putting the wavelength-control function into a module that is effectively separate from the optical generation and amplification functions. For example, a high-performance, TDL-seeded PPLN OPO system developed in the author's laboratory is illustrated in Figure 3. It features a multigrating QPM gain medium that is temperature-adjustable up to 250°C, and is pumped at a wavelength of 1.064 μm by a ns-pulsed, SLM Nd:YAG laser. An active cavity control scheme devised by Yabai He enables the SLM near-IR signal and idler outputs from this OPO (at ~1.5 μm and ~3.5 μm, respectively) to be tuned continuously without mode hops over optical frequency ranges of ~250 cm^{-1} (~7.5 THz) with a SLM optical bandwidth of <0.004 cm^{-1} (<120 MHz), as has been confirmed in assorted applications to molecular spectroscopy. The PPLN OPO can by itself generate only moderate output pulse energy (~0.1 mJ, approximately 70% of which is signal radiation), since the 1.064 μm Nd:YAG pump laser pulse energy is limited to ~1 mJ to avoid the optical damage threshold of the PPLN crystal. For higher-power applications, a further portion of the Nd:YAG laser energy (~200 mJ/pulse) can be used to pump an OPA stage based on BPM $\mathrm{LiNbO_3}$, thereby generating SLM signal and idler output pulse energies of ~1.3 mJ and ~0.7 mJ, respectively.

An alternative approach to injection-seeded tuning of ns-pulsed OPOs (used extensively by the author's group in the context of BPM media such as BBO and $\mathrm{LiNbO_3}$) entails passive control of the OPO cavity by slightly misaligning one of its reflectors. This facilitates continuous tuning of the injection-seeded OPO signal and idler outputs by decreasing the effective finesse of the OPO cavity, so that it is then not necessary to lock the OPO cavity length to the seed wavelength. This mode of operation is simpler optically and electronically than active control of the OPO cavity. The method depends on the OPO cavity having a series of high-order transverse modes to smooth out the sharp, widely separated resonances that occur when the OPO cavity is well aligned; a resulting disadvantage is that the multiple transverse modes tend to cause some degradation of output beam quality. Nevertheless, this approach has proved useful for many applications of tunable OPOs, with seeding by either pulsed dye lasers or SLM TDLs.

This passive, misaligned-cavity approach to injection seeding of ns-pulsed OPOs is particularly well suited for spectroscopic applications requiring a coherent source that simultaneously generates two or more adjustable output wavelengths. Such a TDL-seeded, dual-wavelength BBO OPO has been used in the author's laboratory for coherent-Raman spectroscopic measurements of nitrogen ($\mathrm{N_2}$) in furnace air; the relative intensity of the two spectroscopic features provides an instantaneous, turbulence-immune estimate of the temperature of gas in the furnace, within a single shot of the OPO's pump laser (i.e., within ~10 ns). This is an example of so-called spectroscopic tailoring, in which an OPO generates a structured set of two or more discrete output wavelengths, each of which is set to be on- or off-resonance with characteristic features in spectra of molecular target species of interest.

Dual-wavelength injection seeding of ns-pulsed OPOs is relevant to atmospheric remote sensing by DIAL (differential absorption lidar), with TDL-seeded OPO output switched between on- and off-resonance IR wavelengths on alternate shots of the pump laser; range-resolved measurements have been made of key atmospheric species such as methane ($\mathrm{CH_4}$) and water vapor ($\mathrm{H_2O}$). Figure 4 illustrates a proposed extension of the spectroscopic tailoring concept, applicable to dual- and multi-wavelength remote sensing applications by long-path absorption or DIAL. This multiplex system comprises a multi-wavelength passive-cavity pulsed OPO, injection-seeded by a set of single-mode TDLs and a computer-controlled fiber-optic switch to code and decode the resulting multiplex spectroscopic signals.

A significant frontier for narrowband pulsed OPOs entails extending the spectral output range beyond the present readily attainable near-IR limit of ~4 μm (as provided by materials such as BBO, KTP, $\mathrm{LiNbO_3}$, PPLN, and PPKTP). This extension yields high-performance tunable coherent spectroscopic sources at longwave-IR wavelengths where many of the strongest molecular fundamental IR absorption bands are located. OPOs that target the longwave IR depend on availability of suitable nonlinear-optical materials and coherent pump sources at sufficiently long wavelengths (typically >2 μm). One of the most promising longwave-IR OPO materials is zinc germanium phosphide ($\mathrm{ZnGeP_2}$), pumped at 2–3 μm by solid-state crystal lasers such as those based on holmium or erbium. An étalon-tuned narrowband (~0.1 cm^{-1}) $\mathrm{ZnGeP_2}$ OPO, itself pumped at 2.55 μm by a broadband (~15 cm^{-1}) Nd:YAG-pumped $\mathrm{LiNbO_3}$ OPO, is able to cover the wavelength range 3.7–8 μm and has been used to record spectra of $\mathrm{H_2O}$ vapor in the 6.2 μm region. Pulsed OPOs based on gallium arsenide (GaAs) also have great potential; they can be pumped at wavelengths as short as 1 μm and

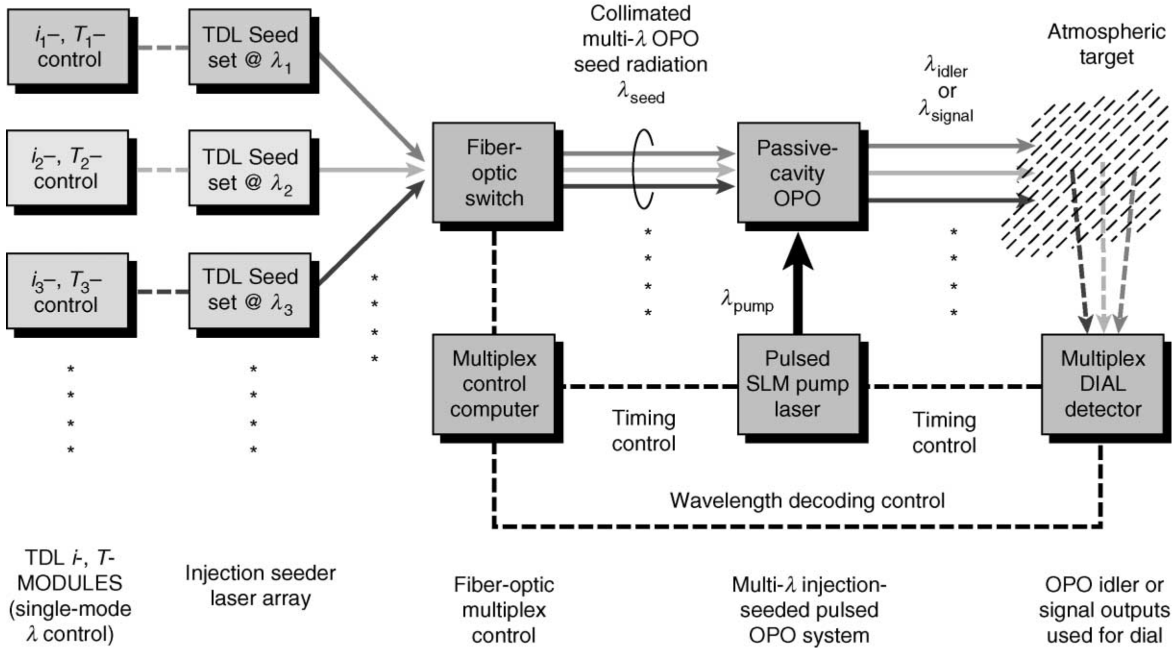
i_1–, T_1– control
TDL Seed set @ λ_1
i_2–, T_2– control
TDL Seed set @ λ_2
i_3–, T_3– control
TDL Seed set @ λ_3
Collimated multi-λ OPO seed radiation λ_{seed}
Fiber-optic switch
Passive-cavity OPO
λ_{idler} or λ_{signal}
Atmospheric target
λ_{pump}
Multiplex control computer
Pulsed SLM pump laser
Multiplex DIAL detector
Timing control
Timing control
Wavelength decoding control
TDL i-, T-MODULES (single-mode λ control)
Injection seeder laser array
Fiber-optic multiplex control
Multi-λ injection-seeded pulsed OPO system
OPO idler or signal outputs used for dial

Figure 4 Schematic diagram of a multi-wavelength pulsed OPO system employing spectroscopically tailored injection seeding. This design has been proposed by the author for atmospheric remote sensing by DIAL (differential absorption lidar) or long-path absorption.

offer (so far unrealized) tunability out to ~16 μm. Despite its cubic structure and lack of birefringence, GaAs serves as a QPM nonlinear-optical medium in the form of orientation-patterned films.

Another approach to operating continuously tunable OPOs is to vary the wavelength of the pump laser, while fixing other phase-matching parameters such as temperature of the gain medium and crystal angle or QPM grating pitch. Suitable tunable pump laser sources that have been used to pump ns-pulsed OPOs include Ti:sapphire and tunable fiber lasers.

The various wavelength-control strategies outlined above illustrate some of the distinctive attributes of OPOs. It should be recognized that the signal and idler wavelengths of an OPO are distinctly different from that of the pump laser itself, thereby opening up fresh regions of the electromagnetic spectrum. Even if an OPO is injection-seeded (e.g., by a cw TDL) at its signal or idler wavelength, it still generates complementary idler or signal radiation at a new wavelength and (in the case of a pulsed OPO) it also pulse-amplifies the original seed light. Moreover, the nonlinear-optical parametric relationships between the pump, signal, and idler waves in an OPO allow various wavelength-control approaches (e.g., spectroscopic tailoring by multi-wavelength injection seeding) that are simply not attainable in tunable lasers.

Ultrafast-Pulsed OPOs

Given the relatively high peak power of very short (sub-ns) pulses of coherent laser light, it might seem quite easy to use such light to pump the gain medium of an OPD and to exceed its OPO threshold. However, such a task is not at all straightforward, because an ultrafast light pulse, on the time-scale of picoseconds (ps, 10^{-12} s) or femtoseconds (fs, 10^{-15} s), does not stay in one place long enough to enable a coherent parametric wave to build up from noise in the nonlinear-optical medium. Light travels a distance of only ~0.3 mm in 1 ps, which does not allow it to make multiple traversals of the nonlinear-optical medium and build up coherent signal and idler waves. This problem is overcome by synchronous pumping (also used in some ultrafast lasers), in which a train of many consecutive ultrafast pulses from a pump laser interact sequentially with a single signal or idler pulse circulating within the OPO cavity. The mode-locking interval of the pump laser (essentially the round-trip transit time in the laser cavity) must therefore equal the round-trip time of the down-converted (signal or idler) pulse in the OPO cavity. The situation is aggravated when sub-ps (e.g., ~100 fs) pump pulses are used, because the pump and signal (or idler) waves have sufficiently different group velocities that they undergo 'temporal walk-off', becoming separated in time after traversing a relatively short length (typically 1–10 mm for 100-fs pulses) of the nonlinear-optical medium.

Practical ultrafast OPOs were introduced around 1990 and their utility grew as improved pump lasers, nonlinear-optical materials (e.g., KTP, PPLN), and OPO cavity designs became available. Research on ps- and fs-OPOs is now a hot topic, and numerous broadly tunable commercial systems are on the market. Ultrafast OPO-based spectroscopy is becoming extremely important in applications to biology and medicine, where many key processes occur on ps- and fs-time-scales.

Continuous-Wave OPOs

The first cw OPO, reported in 1968, operated in the near-IR and was based on barium sodium niobate ($Ba_2NaNb_2O_5$ – known popularly as 'banana'); it was soon followed by visible cw OPOs based on $LiNbO_3$. However, there was not much more substantial progress on practical cw OPOs for another 20 years, until several technological advances had been made: low-threshold QPM nonlinear-optical gain media, such as MgO:$LiNbO_3$ and KTP; efficient, wavelength-selective cavity designs; all-solid-state cw pump lasers (e.g., diode-pumped crystal lasers or MOPA diode systems). More recently, QPM media (such as PPLN and PPKTP) have provided a further stimulus to efficient, compact cw OPOs.

Most ns-pulsed OPOs use singly resonant (SROPO) designs in which the cavity resonates at either the signal or idler wavelength (but not both), doubly resonant (DROPO) schemes in which both signal and idler resonate are often needed to meet the more demanding pump threshold requirements of cw OPOs. Continuous tuning of a DROPO without longitudinal mode hops needs multiparameter approaches, either dual-cavity (with separate signal- and idler-resonant arms) or concerted tuning of pump and signal/idler wavelengths. Efficient cavity designs for cw OPOs include those that are pump-enhanced (i.e., the cavity is resonant at the pump wavelength as well as the signal and/or idler) and intracavity arrangements in which the laser and OPO gain media share a common cavity. Progress is also being made on cw OPOs based on highly nonlinear optical fibers. With such developments, cw OPOs now have great potential as tunable spectroscopic sources, through their narrow optical bandwidth and their compactness.

OPGs and DFGs – Dispensing with the Optical Cavity

Most of this article has focused on OPOs, in which parametric processes are generated by pumping a nonlinear-optical medium in an optically resonant cavity. As already explained, such a cavity is used to enhance the optical parametric gain by multipassing the signal and/or idler radiation. In some cases, the cavity also contributes to wavelength control, temporal characteristics, beam quality, etc. However, not all optical parametric and related devices need a cavity, as is evident from the OPG, OPA, and DFG schematics in Figures 1a, b, and d.

The use of one or more OPA stages, to supplement the intracavity amplification already present in an OPO itself, is widespread in tunable OPO/OPA systems. Moreover, in some applications (e.g., spectroscopic measurements where optical power is not at a premium) of pulsed OPDs, it is found that a single pass of a (relatively long) nonlinear-optical crystal results in a useful OPG (without an OPO cavity). For instance, Scott Bisson, Thomas Kulp, Peter Powers and co-workers have developed and used several OPD systems of this type. A cavity-free OPG/OPA system that is pumped by a 1 kHz Nd:YAG laser, based on 'fan' PPLN (tuned by continuous transverse variation of the QPM grating pitch), and spectrally filtered by a high-finesse étalon, generates ~10 μJ/pulse of idler radiation. This OPG/OPA radiation has been used to measure narrowband ($\Delta\nu < 0.1\ \text{cm}^{-1}$) cavity-ring-down spectra of CH_4 gas.

Likewise, various tunable DFG devices (closely related, but not identical, to OPDs) have been used in many spectroscopic applications. For instance, a DFG system developed by Frank Tittel and colleagues at Rice University took the near-IR output of two room-temperature TDLs (at 1.55 μm and 1.32 μm, respectively), mixed them in a BPM nonlinear-optical medium such as $AgGaSe_2$, and generated 8.7 μm mid-IR radiation for long-path absorption studies of sulfur dioxide (SO_2), an important atmospheric pollutant. The same group has also used QPM media (e.g., PPLN) for DFG mixing of TDL outputs in the 0.8 μm–1.6 μm range, yielding output at ~3.5 μm to monitor gases such as H_2O, CH_4, and formaldehyde (H_2CO); such DFG devices have been used in NASAs Lunar-Mars Life Support Test program.

Relative to OPGs and DFGs, OPOs and OPAs offer higher output power and coherence, both of which are advantageous in some applications (e.g., remote spectroscopic sensing of the atmosphere). Nevertheless, OPGs and DFGs still have a useful role to play, as indicated above.

Concluding Remarks – When to Use an OPD?

Coherent light sources based on optical parametric generation, amplification, and oscillation represent a significant application of nonlinear optics. Such devices can in turn address as many applications as lasers themselves, at least in principle. However, the fact that OPDs need to be pumped by a laser tends to limit their use to situations that take advantage of their wavelength-conversion and tunability capabilities. The most obvious areas of application of an OPO or any other OPD are therefore those in which the wavelength of the output radiation is critical (e.g., spectroscopy, as indicated in various parts of the above text).

From time to time, there have been suggestions that OPOs would largely displace established tunable lasers such as dye and Ti:sapphire lasers. In the early 1990s, one major laser manufacturer ventured to advertise its new OPO-based instrument with the slogan, "Goodbye to Ti: and Dye". However, such a projected monopoly for OPOs has not eventuated and most laser spectroscopy facilities now depend upon a balanced combination of OPOs together with various forms of tunable laser (including semiconductor lasers, which are almost ubiquitous).

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$-harmonic Generation; $\chi^{(3)}$ – Third-harmonic Generation; Four-wave Mixing. **Optical Parametric Devices:** Optical Parametric Oscillators (Continuous Wave); Optical Parametric Oscillators (Pulsed). **Scattering:** Stimulated Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Barnes NP (1995) Optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Lasers Handbook*, pp. 293–348. San Diego, CA: Academic Press.

Byer RL (1975) Optical parametric oscillators. In: Rabin H and Tang CL (eds) *Quantum Electronics: A Treatise*. Nonlinear Optics, part B, chap. 9, vol. 1, pp. 587–702. New York: Academic.

Byer RL and Herbst RL (1977) Parametric oscillation and mixing. In: Shen Y-R (ed.) *Nonlinear Infrared Generation*, pp. 81–137. New York: Springer-Verlag.

Dmitriev VG, Gurzayan GG and Nikogosyan DN (1997) *Handbook of Nonlinear Optical Crystals*, 2nd edn, pp. 339–362.

Dunn MH and Ebrahimzadeh M (1998) Optical parametric oscillators. In: Finlayson DM and Sinclair BD (eds) *Advances in Lasers and Applications*, pp. 61–82. Bristol and Philadelphia: Institute of Physics Publishing.

Ebrahimzadeh M and Dunn MH (2001) Optical parametric oscillators. In: Bass M, Enoch JM, Van Stryland EW and Wolfe WL (eds) *Handbook of Optics, Fiber Optics and Nonlinear Optics*, vol. 4, pp. 22.1–22.73. New York: McGraw-Hill.

Harris SE (1969) Tunable optical parametric oscillators. *Proceedings of the IEEE* 57: 2096–2113.

Orr BJ (2000) IR Lidar applications in air monitoring. In: Myers RA (ed.) *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation, Applications of Instrumental Methods*, vol. 3, pp. 2007–2032. Chichester: Wiley.

Orr BJ, Johnson MJ and Haub JG (1995) Spectroscopic applications of pulsed tunable optical parametric oscillators. In: Duarte FJ (ed.) *Tunable Laser Applications,* chap 2, pp. 11–82. New York: Marcel Dekker.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Sources*. New York: Springer-Verlag.

Sutherland RL (1996) Optical parametric generation, amplification, and oscillation. *Handbook of Nonlinear Optics,* chap. 3, pp. 111–206. New York: Marcel Dekker.

Optical Parametric Oscillators (Continuous Wave)

S Schiller, Heinrich-Heine Universität Düsseldorf, Düsseldorf, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A continuous-wave optical parametric oscillator (cwOPO) is a frequency conversion device in which a continuous-wave (cw) laser field (pump) of frequency ω_p is partially converted into two cw waves of lower frequencies ω_s, ω_i (historically denoted by signal and idler) inside an optical cavity. The three frequencies are related by $\omega_i + \omega_s = \omega_p$. The cavity contains a nonlinear medium with a second-order susceptibility and the mirror reflectivities are chosen such that at least one of signal and idler resonates in the cavity. The conversion efficiency from pump power to signal or idler power or both can be larger than 10%. The frequencies of signal and idler waves can be tuned by means similar to conventional lasers, i.e., by changing the spectral dependence of the cavity loss and by tuning the cavity length. Additional tuning parameters specific to OPOs are the pump frequency and the refractive indices of the nonlinear medium for pump, signal and idler. CwOPOs provide a unique combination of features that makes them suitable sources for a variety of spectroscopic applications. Today's cwOPOs can exhibit the following features:

- spectral coverage from the near-infrared (0.7 μm) to the mid-infrared (5.2 μm); for a single device two optical octaves can be covered without change of optical components;
- high output power (up to several W);
- high pump power conversion efficiency (>30%) and good wall-plug efficiency;
- nearly diffraction-limited output beams (modes are close to lowest-order transverse electromagnetic (TEM_{00}) modes)
- single-frequency output, absence of frequency jumps (mode hops);
- narrow free-running linewidth (<6 kHz);
- high absolute frequency stability (<100 MHz drift per hour);
- continuous frequency tuning exceeding 10 GHz;
- rapid, stepwise tuning over several THz; and
- transportability.

cwOPOs complement their pulsed relatives (*see* **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed)) nanosecond and synchronously pumped pulsed parametric oscillators (*see* **Optical Parametric Devices:** Overview).

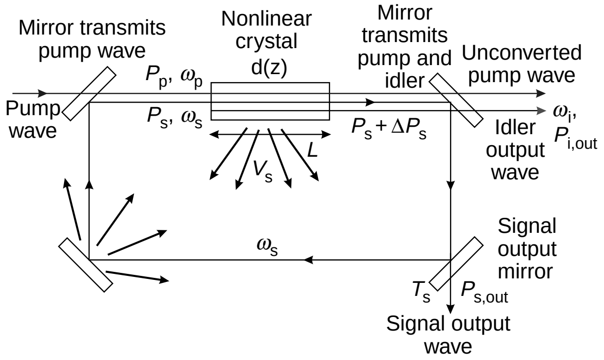

Figure 1 Schematic of a singly resonant cwOPO (SRO) in ring configuration. The signal wave (ω_s) is resonated, the idler wave (ω_i) is not. In the nonlinear crystal, the pump wave overlaps with the signal and idler waves; here they are shown displaced with respect to each other for clarity. The signal wave's attenuation coefficient per roundtrip, S_s, is the sum of loss inside the nonlinear crystal, V_s (indicated by arrows), and transmission through mirrors, here represented by a single mirror of transmission T_s. Scatter and absorption losses at mirrors (indicated by arrows) are included in V_s.

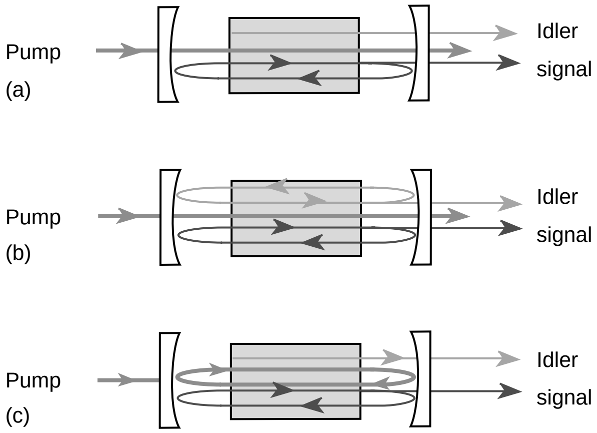

Figure 2 Common types of cwOPOs. (a) basic SRO. (b) Doubly resonant OPO (DRO). Standing-wave cavities are shown. (c) SRO with enhancement of the pump wave (PR-SRO); a common-cavity configuration is shown.

Basic Principle

The operation of a cwOPO is most simply described for a singly resonant OPO (SRO) (see Figure 1 and Figure 2a).

Without loss of generality we take the signal wave to be the resonant wave. When a higher-frequency pump wave (power P_p, angular frequency ω_p) and a lower-frequency (signal) wave (power P_s, angular frequency ω_s) traverse a $\chi^{(2)}$ medium (*see* **Fiber and Guided Wave Optics:** Nonlinear Optics), a transfer of power from the pump wave to the signal wave occurs, described by a power gain:

$$\Delta P_s = G P_s = E P_p P_s \qquad [1]$$

This parametric amplification expression is valid as long as $\Delta P_s \ll P_s, P_p$, i.e., in the limit of small gain and small pump depletion. The (unsaturated) gain coefficient E is a function of: the medium's indices of refraction, orientation and length L, the frequencies ω_s, ω_p, and the spatial modes and overlap of the two beams.

If the wave ω_s is made to circulate inside a cavity and the power attenuation per roundtrip, $S_s P_s$ is smaller than the power gain per roundtrip (eqn [1]), a sustained oscillation is possible. This is satisfied when the input pump power P_p exceeds the threshold:

$$P^{\text{th}}_{p,\text{in}} = S_s / E \qquad [2]$$

As the pump power P_p is increased above $P^{\text{th}}_{p,\text{in}}$, conversion of pump power to idler power occurs, implying that the pump power decreases within the medium. Qualitatively, the average pump power $\langle P_p \rangle$

along the medium remains at the level $P_{\mathrm{p,in}}^{\mathrm{th}}$, so that the roundtrip fractional power gain remains constant, $G \simeq E\langle P_{\mathrm{p}}\rangle \simeq EP_{\mathrm{p}}^{\mathrm{th}} \simeq S_{\mathrm{s}}$. This ensures that the device reaches a stationary state. The circulating signal power P_{s} increases steadily with increasing pump power, and so does the signal power emitted from the OPO, $T_{\mathrm{s}}P_{\mathrm{s}}$, which is equal to the 'useable' fraction of the roundtrip power gain, $(T_{\mathrm{s}}/S_{\mathrm{s}})GP_{\mathrm{s}}$.

The parametric process that allows the signal wave to experience gain (eqn [1]), also generates an idler wave (difference frequency generation). Its power at the end of the medium is given by the Manley–Rowe relation:

$$P_{\mathrm{i}} = \frac{\omega_{\mathrm{i}}}{\omega_{\mathrm{s}}}\Delta P_{\mathrm{s}} \qquad [3]$$

An OPO also requires a process that starts the oscillation. Spontaneous parametric fluorescence provides this trigger. Due to this quantum effect, a pump wave traversing a $\chi^{(2)}$ medium will generate a flux of signal photons even if there is no signal wave present at the crystal input face. In the photon picture, a tiny fraction (on the order 10^{-9} for a 1 W pump) of the pump photons spontaneously annihilate, creating pairs of signal and idler photons (photon splitting). The idler photons have a frequency $\omega_{\mathrm{i}} = \omega_{\mathrm{p}} - \omega_{\mathrm{s}}$, so that photon energy conservation holds. The process may be thought of as being stimulated by the vacuum fluctuations of the electromagnetic field at the signal and idler frequencies. Some of the signal photons are emitted into the cavity mode, so that after a roundtrip they can serve as a seed for further (stimulated) photon splitting. The process repeats, and the signal wave builds up exponentially in time to a macroscopic level if the pump power is above threshold. The spontaneous (as well as the parametric amplification) process is strongest when momentum conservation is essentially fulfilled, $k_{\mathrm{p}} \simeq k_{\mathrm{s}} + k_{\mathrm{i}}$, where $k = \omega n(\omega)/c$.

In the photon picture, the relation (eqn [3]) is a statement that signal and idler photons are always created in pairs.

Types of OPOs

OPO types may be classified according to the number of waves that resonate within cavities. Typically, a realizable roundtrip loss coefficient is on the order of $S_{\mathrm{s}} = 1\%$ (to which the linear absorption of the nonlinear crystal and mirror transmissivities contribute) and the gain coefficient $E \sim 0.1\%/\mathrm{W}–1\%/\mathrm{W}$ for Gaussian waves, depending on the pump wavelength, crystal type, etc. This leads to SRO threshold powers on the order of 1 to 10 W. Until the early 1990s, solid-state pump lasers of this power level and single-frequency output were unavailable. This led to the interest in and development of cwOPOs in which more than one of the three waves is resonated (Figure 2). The corresponding external threshold powers are

$$\begin{aligned} P_{\mathrm{p,in}}^{\mathrm{th}} &\simeq S_{\mathrm{s}}S_{\mathrm{i}}/4E \text{ for a doubly-resonant OPO} \\ &\qquad \text{(DRO, Figure 2b)} \\ P_{\mathrm{p,in}}^{\mathrm{th}} &\simeq S_{\mathrm{s}}S_{\mathrm{p}}/E \text{ for a pump-resonant SRO} \\ &\qquad \text{(PR-SRO, Figure 2c, see eqn [35])} \\ P_{\mathrm{p,in}}^{\mathrm{th}} &\simeq S_{\mathrm{s}}S_{\mathrm{i}}S_{\mathrm{p}}/E \text{ for a triply-resonant OPO (TRO)} \end{aligned} \qquad [4]$$

Here $S = T + V$ is the fractional power attenuation per roundtrip, arising from the mirror transmission T and absorption/scatter loss V. While the mirror transmission maintains power in a given spatial mode, the loss V removes optical power completely and is the fundamental cause for the nonunity efficiency of the frequency conversion process. Typically, each additional resonant enhancement lowers the threshold power by one to two orders.

Examples for threshold values reported in the literature are: 1.9 W for a 920 nm-pumped SRO, 4 mW for a 0.5 μm-pumped DRO, and 140 mW for a 1 μm-pumped common-cavity PR-SRO. DROs can therefore be pumped even by low-power diode lasers. TROs (where pump, signal, and idler are resonated) are not of importance due to their higher complexity.

Output Power

As an example, the signal and idler output powers for a PR-SRO that resonates the signal are (for a derivation, see Appendix B):

$$P_{\mathrm{s,out}} \simeq T_{\mathrm{s}}P_{\mathrm{s}}(0), \qquad P_{\mathrm{i,out}} \simeq (\omega_{\mathrm{i}}/\omega_{\mathrm{s}})S_{\mathrm{s}}P_{\mathrm{s}}(0) \qquad [5]$$

where the circulating signal power is

$$P_{\mathrm{s}}(0) \simeq \frac{\omega_{\mathrm{s}}}{\omega_{\mathrm{p}}}\frac{4T_{\mathrm{p}}}{S_{\mathrm{s}}S_{\mathrm{p}}}P_{\mathrm{p,in}}^{\mathrm{th}}\left(\sqrt{\frac{P_{\mathrm{p,in}}}{P_{\mathrm{p,in}}^{\mathrm{th}}}} - 1\right) \qquad [6]$$

T_{p} is the input mirror transmission for the pump wave, and T_{s} is the output mirror transmission of the signal wave. Signal and idler output powers have the same dependence on pump input power and increase monotonically with $P_{\mathrm{p,in}}$. The conversion efficiencies reach a maximum four times above threshold, i.e., when $P_{\mathrm{p,in}} = 4P_{\mathrm{p,in}}^{\mathrm{th}}$. There, the signal and idler photon (quantum) conversion efficiencies $\omega_{\mathrm{p}}P_{\mathrm{s,out}}/P_{\mathrm{p,in}}\omega_{\mathrm{s}}$ and $\omega_{\mathrm{p}}P_{\mathrm{i,out}}/P_{\mathrm{p,in}}\omega_{\mathrm{i}}$ exceed 25% and 50%, respectively, if the mirror transmissions T_{p}, T_{s} for pump and signal are chosen larger than the respective loss V_p, V_s. This can be achieved in practice, of course, at the expense of a higher threshold.

Nonlinear Materials

While the fundamental developments in cwOPOs were performed using birefringently phase-matched nonlinear crystals ($MgO:LiNbO_3$, LBO, KTP, $KNbO_3$) (*see* **Materials for Nonlinear Optics:** Liquid Crystals for NLO), and substantial spectral coverage is possible with these crystals, most cwOPOs are now operated with quasi-phase-matched (QPM) nonlinear crystals (*see* **Nonlinear Optics, Application:** Phase Matching). The most commonly used material is periodically poled (PP) $LiNbO_3$ (PPLN), other crystals used are PPKTP and $PPLiTaO_3$, crystals. Their use has led to a substantial widening of the spectral coverage, opening in particular the wavelength range $\lambda > 1.6\ \mu m$, a reduction of the required pump powers thanks to their much higher nonlinearities compared to birefringently phase-matched crystals, and a simplification of (gross) tuning through the use of multigrating structures (containing several sections of differing Λ periods side by side within the same crystal). The above PP crystals exhibit a loss of 0.1%/cm at wavelength around 1 μm, and can be coated with broadband dielectric antireflection coatings, in order to maximize the overall transmission. Another important aspect is that the crystals withstand long irradiation times with high-power focused cw light. For PPLN, operation temperature above 100°C is chosen in order to prevent photorefractive effects.

Spectral Coverage

An example of particularly wide spectral coverage reported in the literature is a cwOPO pumped by a frequency-doubled $Nd:YVO_4$ laser ($\lambda_p = 532$ nm) using a multigrating PPLN crystal. With grating periods $\Lambda = 6.5-9.6\ \mu m$, the range 660–1030 nm (signal wave) and 1100–2800 nm (idler wave) was covered. A cwOPO with a Nd:YAG pump source emitting at the fundamental wavelength ($\lambda_p = 1064$ nm) and a multigrating PPLN crystal ($\Lambda = 25.5-31.2\ \mu m$) achieved a range of 1.45–1.99 μm (signal) and 2.3–4.0 μm (idler).

For idler wavelengths beyond 4 μm, idler absorption by the crystals types above becomes relevant, leading to a reduction in the gain G (this effect is not considered in the treatment given in the Appendix) and to an increase of the threshold. To a certain extent, this can be compensated with powerful cw pump lasers. For example, a 1064 nm pumped PPLN-SRO achieved idler emission in the range 3.6–4.7 μm at 11 W pump power. Also, an 850 nm pumped PR-SRO covered an idler range 4.3–5.3 μm when pumped with 750 mW.

In the future, cwOPOs will certainly profit from the development of novel QPM materials with wider IR transparency range.

Frequency Control

The small-signal gain $G(\omega_s)$ of a parametric interaction in a cwOPO is analogous to the homogenously broadened gain in a laser medium. In contrast to it, however, spatial hole-burning cannot occur. Therefore, mode competition is generally strong in a cwOPO, which will oscillate on a single longitudinal mode: the one for which the difference (or ratio) between 'unsaturated' parametric gain G and total loss S is largest (Figure 3). This is equivalent to the condition of minimum threshold.

The functions $G(\omega_s)$ and $S(\omega_s)$ are therefore important characteristics of a cwOPO. Oscillation can only occur for a frequency ω_s that is very close to a cavity mode frequency $\omega_q = 2\pi qc/OPL(\omega_q)$, where q is the integer mode number, and $OPL(\omega_s)$ is the round-trip optical path length at the frequency ω_s.

To tune a cwOPO to a particular signal/idler frequency pair, a coarse tuning is initially performed

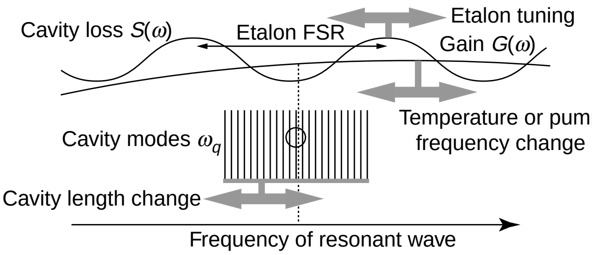

Figure 3 Frequency selection and tuning in a SRO (or PR-SRO) containing an etalon. Oscillation occurs on the signal cavity mode (circled) closest to the frequency where the difference between the small-signal gain G and the loss S is largest (dashed line). The frequency of this mode or the mode number q can be changed by changing the parameters indicated near the thick double arrows. The parameter changes shift the line(s) 'attached' to the corresponding double arrow. Changing the temperature leads to a rough setting of the oscillation frequency. Tilting the etalon allows mode-hop tuning of signal and idler frequencies. Fine tuning by a small amount can be done by changing the cavity length. The frequency of the idler is not shown here; it is always given by the difference between pump frequency and the frequency of the oscillating mode. Thus, tuning of the pump frequency generally leads to a tuning of the idler frequency. For modest pump tuning the amount of signal tuning however depends on SRO type: no signal tuning in a basic SRO and in a dual-cavity PR-SRO, while in a common-cavity PR-SRO pump and signal frequencies are correlated and thus pump tuning leads to signal tuning. FSR: free spectral range.

by tuning the corresponding phase mismatch ΔkL (L is the crystal length) approximately to zero. Taking into account a modulation of the nonlinear susceptibility of the crystal with period Λ, the quantity $\Delta k = k_\mathrm{p} - k_\mathrm{s} - k_\mathrm{i} \pm 2\pi/\Lambda$ describes the deviation from momentum conservation in the photon splitting process. When using a multigrating QPM crystal, the first tuning consists in the selection of a grating of appropriate period Λ, which is moved into the pump wavepath. This is followed by tuning the photon momenta via the refractive indices of the nonlinear medium, which is usually accomplished by temperature tuning, or alternatively by angle-tuning of the crystal. These steps place the maximum of $G(\omega_\mathrm{s})$ to the vicinity of the desired frequency. Options for tuning the emission exactly to the desired frequency are described below for the different OPO types.

Because external perturbations change parameters that enter in $G(\omega_\mathrm{s})$, $S(\omega_\mathrm{s})$, and ω_q, the value of the oscillating frequency will also change in time. The change will be continuous over a certain range of the parameters, but if their perturbation is large enough, a mode hop can occur: the oscillation on the initial mode q will eventually cease and another mode q', for which $G(\omega_{q'}) - S(\omega_{q'})$ is larger, will oscillate instead. A variety of parameters is affected by perturbations: pump power, pump beam direction, cavity length, crystal temperature, and pump frequency.

When such perturbations are applied on intentionally, the OPO output frequencies can be tuned, scanned, or adjusted. The tolerable change in the various parameters before a mode-hop occurs, and the sensitivity of the output frequencies to changes in the parameters depends on the particular cwOPO type. One important approach to suppress mode-hops consists in adding elements into the cavity, such as an etalon, that modify the function $S(\omega)$, typically by modulating it spectrally.

SRO

In the following, we assume that the signal wave is resonated. The SRO is the simplest cwOPO type. The oscillation nominally occurs at the cavity frequency that is closest to the gain maximum, as described above. In practice, environmental disturbances will cause mode-hops on a time-scale of minutes. For this reason, an etalon is added to the cavity. Mode-hop-free oscillation, for several hours, is then achieved.

Typical nonlinear crystals exhibit parametric gain with spectral widths of hundreds of GHz. In contrast, the cavity mode spacing (free spectral range, FSR) c/OPL is on the order of 1 GHz. The tuning range achievable by a cavity length change (using, e.g., a piezo translator) is therefore, at most, one free spectral range (equal but opposite for signal and idler); then a mode-hop to the neighboring cavity mode occurs, with a signal/idler frequency jump by one FSR in the opposite direction to the previous tuning. This process repeats itself as the cavity length is tuned further. It is, therefore, not possible to easily and reliably access a large frequency range by using the cavity length as tuning parameter.

As shown in Figure 3, if an etalon of linewidth smaller than the gain bandwidth is added to the cavity, it becomes the dominant influence (etalon FSRs on the order of one hundred to a few hundred GHz have been used). This means that shifting the etalon's transmission maximum (i.e., the cavity loss S) in frequency space (by tilting a solid etalon or changing the spacing of an air-spaced etalon) as indicated by the double-headed arrows in Figure 3, forces the signal to mode-hop to an adjacent mode, with a corresponding idler frequency hop in the opposite direction. As the etalon is tuned further this occurs repeatedly. This discontinuous 'mode-hop tuning' is useful for tuning the SRO output frequencies over a significant range (e.g., 100 GHz) and allows to reach a desired frequency within one cavity free spectral range.

Continuous tuning of both signal and idler waves is obtainable, if the pump laser is not frequency tunable, by changing the cavity length and synchronously tuning the etalon. This can be achieved by a feedback or feedforward system that controls the etalon angle or spacing. Up to 38 GHz tuning for signal and idler were achieved in this way.

When a tunable pump is available, one can tune the pump frequency while keeping the cavity length and the etalon angle fixed. This keeps the signal cavity mode fixed, and thus the nonresonant wave's frequency (usually the idler) is tuned by an amount equal to the pump frequency change. Continuous tuning ranges exceeding 50 GHz have been obtained in this way. For such pump tuning ranges, the shift of the gain curve, indicated by a double-arrow in Figure 3, is usually sufficiently small that it can be neglected. The use of a widely tuneable pump laser is also possible. A (discontinuous) idler tuning of 9 THz around 3.3 μm has been demonstrated, covered in just 0.3 ms thanks to a large pump tuning rate. In such a case of large pump tuning range, the gain shift is important, causing a significant tuning also of the signal.

The advantages of the basic SRO as compared to other cwOPOs are mainly the simpler tuning behavior, the possibility of continuous tuning without necessitating a tunable pump laser, and the possibility to achieve a very wide tuning range without optics change, since only the spectral range covered by one of the two parametric waves must be provided as a high reflectivity coating on the cavity mirrors. Moreover, in the simplest implementation, it is possible to operate without any electronic servo control of cavity elements, at the expense of a reduced power and frequency stability. In terms of spectral properties (linewidth, frequency stability, time between undesired mode-hops) the SRO does not offer particular advantages compared to other cwOPO types, even if actively stabilized (see below). The disadvantage of the basic SROs based on Gaussian modes is the high threshold, necessitating expensive high-power (several W) cw lasers. This disadvantage is likely to disappear with progress in laser development. The alternative of waveguide SROs is mentioned below.

PR-SRO

In a PR-SRO, the pump wave is also resonated, leading to a substantial reduction of threshold. For example, it is possible to generate mid-IR radiation (3 μm range), even with much more affordable 1 W-level Nd:YAG pump lasers.

One implementation is the common-cavity PR-SRO (Figure 4). The cavity length is either locked to the pump laser frequency or vice-versa. If a frequency-stable pump laser is used, the first solution is favorable, since then both signal and idler waves have a stability comparable to that of the pump laser. With an etalon (coated so as to be transparent for the pump wave) inside the cavity, mode-hop-free operation for several hours has been demonstrated.

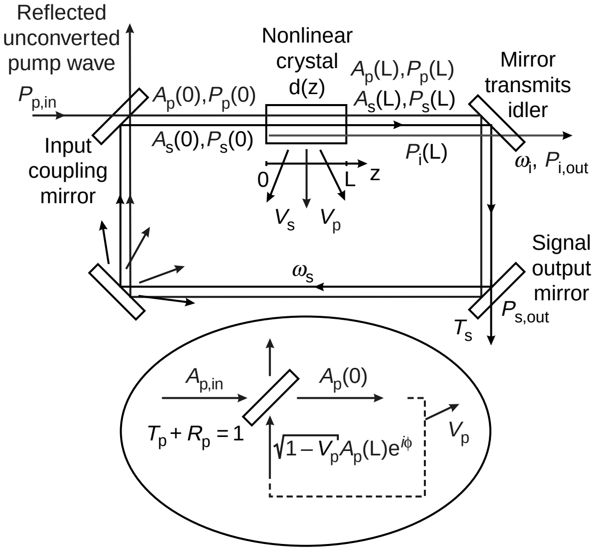

Figure 4 Schematic of a pump-resonant SRO with a common cavity for pump and signal waves. The inset shows the pump wave fields incident on and leaving the input coupling mirror. The arrow symbolizes pump wave loss during the round trip.

Mode-hop tuning in a PR-SRO with cavity-to-laser lock can be achieved:

(i) by tuning the etalon at constant pump frequency: the signal mode-hops from mode to mode, the idler tunes equally but in opposite direction;
(ii) by tuning the pump frequency at constant etalon position: the cavity length must follow and therefore the signal tunes by $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, while the idler frequency tunes by $\Delta\omega_i = (\omega_i/\omega_p)\Delta\omega_p$. This continues until the signal must mode-hop back, causing an equal but opposite frequency jump in the idler. The pump tuning is thus completely transferred to the idler.

Continuous signal and idler tuning is performed by tuning the pump frequency and synchronously adjusting the etalon angle. Thus, a limitation of the common-cavity PR-SRO concept is that continuous tunability necessitates a tunable pump laser.

In a dual-cavity, PR-SRO pump and signal waves resonate in separate cavities. The pump cavity is usually stabilized to the pump frequency and an etalon is placed into the signal cavity for better stability against mode-hops. This configuration combines the tuning simplicity of a basic SRO with a low-threshold power. For example, a threshold for a 1 μm-pumped PPLN dual-cavity PR-SRO at the 0.4 W level was obtained.

Mode-hop tuning (in opposite directions) of the frequencies of both parametric waves in steps of the cavity's FSR can be performed by changing the etalon angle (Figure 3). A range exceeding 50 GHz (at constant output power) has been achieved. Continuous signal and idler wave tuning (by equal but opposite amounts) is possible by changing the signal cavity length (Figure 3). However, the range is limited to the FSR of the cavity, typically on the order of 500 MHz, then the frequencies mode-hop back. It is expected that wider continuous tuning should be possible by synchronous etalon and cavity length scan. If a tuneable pump laser is employed, continuous tuning of only the idler frequency is possible by keeping the etalon and signal wave cavity length fixed; the pump tuning range is fully transferred to the idler.

DRO

In contrast to an SRO, this device type requires both parametrically generated waves to be resonant in the cavity. Here we consider only DROs with a common cavity for both signal and idler waves (dual-cavity DROs have also been demonstrated).

Because of the added resonance condition, mode-hops of the oscillating mode pair occur already upon a very small change (~1 nm) in the cavity length. Under normal environmental conditions such changes are unavoidable and therefore a 'free-running' DRO does not produce frequency-stable output. However, by active control of one additional parameter of the OPO system the mode-hops can be effectively suppressed. A practical solution is to actively control the length of the cavity via the position *a* of one of the cavity mirrors so that a particular mode pair, q_s, q_i, continues to oscillate; this is achieved by maximizing the OPO output power or by minimizing the detuning of one of the two resonant waves from its respective cavity mode frequency. This control system ensures that energy conservation and the two resonance conditions are simultaneously satisfied:

$$\begin{aligned} \omega_s + \omega_i &= \omega_p \\ \omega_s &= 2\pi c q_s / \mathrm{OPL}(d, \omega_s, T) \qquad [7] \\ \omega_i &= 2\pi c q_i / \mathrm{OPL}(d, \omega_i, T) \end{aligned}$$

Experimentally, mode-hop-free oscillation exceeding 10 h was achieved by such active stabilization.

Another system parameter may then be used to tune both signal and idler frequencies. The tuning coefficient for the crystal temperature T is small, but tuning the pump frequency is an efficient way. The tuning coefficient is approximately $\Delta\omega_s = (\omega_s/\omega_p)\Delta\omega_p$, and accordingly for ω_i.

Continuous tuning of signal and idler frequencies over 10 GHz by pump tuning was achieved. This range was limited by the available continuous tuning range of the pump laser. Mode-hop tuning can be induced by mechanical perturbations: by tapping on the DRO cavity, mode-hop tuning over 220 GHz has been demonstrated.

Linewidth and Frequency Stability

For high-resolution spectroscopy a narrow linewidth for signal or idler is desirable. One contribution to the linewidth of a cwOPO arises from the unavoidable spontaneous parametric fluorescence, similar to the Schawlow–Townes limit in lasers due to spontaneous emission. This contribution is negligible if the OPO operates at power levels of practical interest. Thus, other sources of classical noise dominate the linewidth of cwOPOs. As discussed above, the frequency of cwOPOs depends (on a fine scale) on pump frequency and cavity optical path length. Both parameters fluctuate: the pump laser has a finite linewidth as well as frequency jitter, and the cavity optical path length fluctuates due to acoustic noise affecting the mirror positions, to air pressure fluctuations, temperature fluctuations of the crystal and of the cavity structure.

With careful construction and shielding, many of these noise sources can be minimized. Furthermore, it is highly favorable to use pump lasers of narrow linewidth and high intrinsic frequency stability (e.g., diode-pumped monolithic solid-state lasers), or actively frequency-stabilized lasers. Note that this special requirement on the pump source does not exist for lasers.

In an SRO, the linewidth and frequency stability of the signal are mostly determined by the stability properties of the cavity. The idler spectral properties then follow entirely from the signal and pump spectral properties through $\omega_i = \omega_p - \omega_s$. Since the idler is usually the wave of interest, this implies that the SRO cavity and the pump laser must both have a stability appropriate to the spectroscopic application.

If the SRO cavity is not sufficiently stable (e.g., because of insufficient temperature stability of the oven containing the nonlinear crystal), the frequency fluctuation/drift of the signal and idler can be substantial, tens to hundreds of MHz within minutes being typical.

To avoid mode-hops also on long time-scales, the cavity length (or the etalon angle) can be actively stabilized so as to maximize the idler's output power. This ensures that the cavity mode position and the frequency of maximum difference between gain and loss are kept equal.

If a frequency-stable pump wave is used, it is possible to transfer these characteristics by stabilizing the SRO cavity length to the pump wave frequency. In an SRO without resonant pump one can make use of the residual reflection of the pump wave by the cavity mirrors that leads to interference with the pump wave reflected from the pump input mirror. This was implemented in a SRO pumped by a high-power diode-pumped internally doubled Nd:YVO$_4$ laser (532 nm) of high intrinsic frequency stability. Signal and idler drifts of less than 50 MHz/h were achieved. At the same time, a very small 20 kHz linewidth of the (resonant) signal radiation was obtained. Similar levels were obtained for a PR-SRO with a common cavity for pump and signal wave.

In a PR-SRO with dual-cavity configuration, the same requirements of minimization of disturbances to the parametric wave's cavity arise as in a basic SRO. Using a highly frequency-stable kHz linewidth monolithic pump laser, together with the above-mentioned idler power maximization control, frequency drift values for both parametric waves at the level of 30 MHz/h were obtained and a short-term linewidth below 6 kHz.

In a single-cavity DRO, the linewidth and frequency stability of the parametric waves are essentially determined by the pump wave properties (see discussion in the section on DRO above). Here, too, the use of narrow-linewidth, high-frequency-stability pump lasers has led to excellent signal/idler characteristics. For example, less than 40 kHz linewidth and <40 MHz per hour drift were obtained with a DRO pumped by an externally doubled monolithic Nd:YAG laser of 10 kHz linewidth.

If further narrowing of linewidth or reduction of frequency drift is desired, the OPO output frequencies can be stabilized to a stable cavity, atomic, or molecular transition. Locking to a cavity has been demonstrated for a DRO.

Applications

CwOPOs are most interesting as sources of radiation in the IR range beyond 2 μm, especially in the 3–5 μm range, where vibrational transitions of many molecules occur. In comparison to other sources of coherent radiation (color center laser, quantum cascade laser, lead-salt diode laser, CO overtone laser, difference frequency generation), cwOPOs provide a unique combination of desirable features.

A simple application of cwOPOs would be as a source providing a power level/output wavelength/mode quality combination not available from other sources. Possible applications are vibrational spectroscopy in the condensed phase, where a narrow linewidth is not required. The excellent beam quality provides a high spatial resolution.

Concerning high-resolution spectroscopy (mostly in the gas phase), methods demonstrated in conjunction with cwOPOs include:

- Absorption spectroscopy;
- Doppler-free saturation spectroscopy;
- Photo-acoustic spectroscopy (PAS);
- Hole-burning spectroscopy;
- Coherent atomic spectroscopy;
- Cw cavity ring down spectroscopy (cw-CRDS).

Trace gas detection using PAS or cw-CRDS is a particular application where cwOPOs have reached excellent sensitivity (minimum detectable concentrations as low as 1 part in 10^{11}). The high output power and/or the narrow linewidth are specific features that make such sensitivity possible.

Quantum-Optical Properties

The pairwise production of signal and idler photons is at the heart of the parametric process. This perfect quantum correlation is to a certain extent also maintained in the output waves of a cwOPO. In particular, the intensity fluctuations of signal and idler waves emitted by a DRO are strongly (albeit never perfectly) quantum correlated if their output coupler transmissions significantly exceed their respective roundtrip losses. For these 'twin beams' the spectral density of the fluctuations of their power difference can be less than the spectral density of the fluctuations of the power of an individual wave (squeezing). Making use of this correlation permits to increase the obtainable sensitivity in spectroscopic measurements for a given power level of the wave interacting with the sample.

Device Development Trends

CwOPOs using solid-state pump lasers have become commercially available. In the near future, some likely developments will be miniaturization and optimization in order to reduce the influence of perturbations, further extension of continuous and mode-hop tuning ranges and the wider use of diode lasers as pump lasers, leading to cost reduction. To access a larger range of molecules in spectroscopy, extension of the emission range well beyond 5 μm is also of significant interest. Novel nonlinear materials and pump lasers of wavelength longer than the often employed Nd:YAG lasers will become of importance in this development.

Two specific lines of development complementary to the bulk cwOPOs described above are waveguide and intracavity cwOPOs (ICSRO). If pump, signal and idler waves are confined in a waveguide (*see* **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides) the parametric gain coefficient is a few orders higher than for free-space Gaussian waves, due to the small mode cross-section and the scaling with the square of the crystal length, compared with the linear scaling in case of Gaussian beams (see eqn [36]). Values are on the order of $E \sim 1/\mathrm{W}$ at $\lambda_p = 1.5\ \mu\mathrm{m}$ leading to low thresholds even for SROs, with pump lasers of comparatively long wavelength. For example, in a SRO consisting of a 9 cm long titanium:PPLN waveguide, a threshold below 300 mW for a 1.5 μm pump was achieved. The idler emission range was 3.1–3.4 μm. The potential for mass production and, therefore, cost minimization is clear.

An ICSRO consists of an SRO internal to the cavity of a cw laser, where the high circulating laser power is taken advantage of for pumping. An ICSRO is thus similar to a PR-SRO but there is no need to implement a control system for the frequency lock between pump laser and pump cavity nor to mode-match the pump wave to the SRO cavity mode.

Appendix A

Three-Wave Mixing in a $\chi^{(2)}$ Medium with Focused Waves

The wave equation in a nonlinear medium reads

$$\Delta\vec{E} - \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{D}_\mathrm{L} = \mu_0\mu_r\frac{\partial^2}{\partial t^2}\vec{P}_\mathrm{NL} \qquad [8]$$

The displacement field $\vec{D}_\mathrm{L}$ is related to the electric field $\vec{E}$ by linear response. $\vec{P}_\mathrm{NL}$ is the nonlinear polarization. We deal with nonmagnetic media ($\mu_r = 1$) and choose the principal axis system as coordinate system. The Fourier transform of, for example, the x-component of the wave equation then reads:

$$\begin{aligned}&\Delta E_x(\omega,\vec{r}) + \mu_0\omega^2\varepsilon_0\varepsilon_x(\omega)E_x(\omega,\vec{r})\\ &\quad = -\mu_0\omega^2 P_{\mathrm{NL},x}(\omega,\vec{r})\end{aligned} \qquad [9]$$

In the following, we take into account only second-order nonlinear effects, so the nonlinear polarization reduces to $P_\mathrm{NL} = P^{(2)}$. The fields to be considered are monochromatic, propagate in the z-direction and are linearly polarized along x. The notation can then be simplified to:

$$E_x(\omega,\vec{r}) = E(\vec{r})\mathrm{e}^{\mathrm{i}kz}, \quad P_{\mathrm{NL},x}(\omega,\vec{r}) = P_\omega^{(2)}(\vec{r}) \qquad [10]$$

with the wavevector $k = \omega n(\omega)/c$ and the refractive index for the polarization under consideration $n_\omega^2 = n(\omega)^2 = E_x(\omega)$. The electric field is obtained as $\vec{E}(t,\vec{r}) = \frac{1}{2}\mathrm{Re}(E(\vec{r})\mathrm{e}^{\mathrm{i}(kz-\omega t)})\hat{x}$.

Since the parametric gains are small in cw conversion, the derivative of the envelope $E(\vec{r})$ changes only little over a propagation distance on the order of a wavelength $2\pi/k$. Inserting eqn [9] into eqns [6] and [8] and neglecting the term $\sim \mathrm{d}^2E(\vec{r})/\mathrm{d}z^2$, yields the paraxial or slowly varying envelope wave equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + 2\mathrm{i}k\frac{\partial}{\partial z}\right)E(\vec{r}) = -\mu_0\omega^2 P_\omega^{(2)}(\vec{r})\mathrm{e}^{-\mathrm{i}kz} \qquad [11]$$

The electric field envelope can be expanded in a complete set of Gaussian TEM$_{mn}$ – mode functions ψ_{mn}:

$$E(\vec{r}) = \sqrt{\frac{\omega}{n_\omega}}\sum_{m,n=0}^{\infty} A_{mn}(z)\psi_{mn}(\vec{r}) \qquad [12]$$

These mode functions are defined to satisfy the paraxial wave equation in absence of nonlinearity (vanishing right-hand side in eqn [11]) and the orthonormality properties

$$\int\int_{-\infty}^{+\infty}\psi_{mn}^*(\vec{r})\psi_{m'n'}(\vec{r})\mathrm{d}x\,\mathrm{d}y = \delta_{mm'}\delta_{nn'} \qquad [13]$$

A particular mode function set is characterized by position (along z) and size of the waist, which can be chosen arbitrarily. The functions $A_{mn}(z)$ are slowly varying amplitude functions. The power in a particular mode can be calculated from:

$$P_{mn}(z) = \frac{\omega|A_{mn}(z)|^2}{2\mu_0 c} \qquad [14]$$

where c is the speed of light *in vacuo*. Inserting eqn [12] into eqn [11], multiplication with $\psi_{m'n'}^*(\vec{r})$, and integration over x, y yields:

$$\frac{\mathrm{d}A_{m'n'}(z)}{\mathrm{d}z} = \mathrm{i}\frac{\mu_0 c}{2}\sqrt{\frac{\omega}{n_\omega}}\mathrm{e}^{-\mathrm{i}kz}\int\int_{-\infty}^{\infty}P_\omega^{(2)}(\vec{r})\psi_{m'n'}^*(\vec{r})\mathrm{d}x\,\mathrm{d}y \qquad [15]$$

If all electric fields are linearly polarized, the Fourier amplitudes of the nonlinear polarization can be written as:

$$\begin{aligned}P_{\omega_\mathrm{s}}^{(2)}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_\mathrm{p}}(\vec{r})E_{\omega_\mathrm{i}}^*(\vec{r})\\ P_{\omega_\mathrm{i}}^{(2)}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_\mathrm{p}}(\vec{r})E_{\omega_\mathrm{s}}^*(\vec{r})\\ P_{\omega_\mathrm{p}}^{(2)}(\vec{r}) &= 2\varepsilon_0 d(z)E_{\omega_\mathrm{s}}(\vec{r})E_{\omega_\mathrm{i}}(\vec{r})\end{aligned} \qquad [16]$$

where d is the tensor element appropriate to the particular combination of linear polarizations of the three waves. (In eqn [16], $\omega_\mathrm{i} + \omega_\mathrm{s} = \omega_\mathrm{p}$.)

The mode amplitude equations for the three interacting waves follow from eqns [10], [15], and [16] as:

$$\frac{\mathrm{d}A_{mn}^{\left\{\begin{smallmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{smallmatrix}\right\}}(z)}{\mathrm{d}z} = \mathrm{i}\frac{d(z)}{c}\left\{\begin{matrix}\sqrt{\omega_\mathrm{p}/n_\mathrm{p}}\mathrm{e}^{-\mathrm{i}\Delta kz}\\ \sqrt{\omega_\mathrm{s}/n_\mathrm{s}}\mathrm{e}^{+\mathrm{i}\Delta kz}\\ \sqrt{\omega_\mathrm{i}/n_\mathrm{i}}\mathrm{e}^{+\mathrm{i}\Delta kz}\end{matrix}\right\}\int\int\mathrm{d}x\,\mathrm{d}y$$

$$\times\left\{\begin{matrix}\mathrm{E_s}(\vec{r})\mathrm{E_i}(\vec{r})\\ \mathrm{E_p}(\vec{r})\mathrm{E_i}(\vec{r})^*\\ \mathrm{E_p}(\vec{r})\mathrm{E_s}(\vec{r})^*\end{matrix}\right\}\psi_{mn}^{\left\{\begin{smallmatrix}\mathrm{p}\\ \mathrm{s}\\ \mathrm{i}\end{smallmatrix}\right\}}(\vec{r})^{*+\infty} \qquad [17]$$

The quantity $\Delta k = k_\mathrm{p} - k_\mathrm{s} - k_\mathrm{i}$ is the wavevector mismatch. Here, the envelope of each wave has an expansion of the form in eqn [12] with its respective (independent) set of mode functions.

Appendix B

Steady-State Description of the cwOPO

The general eqns [17] are applicable to any type of cwOPO. Below and at threshold at least two of the three waves have given modes, i.e., the corresponding function $E(\vec{r})$ is known up to a slowly z-dependent amplitude: the resonant parametric wave has a mode defined by the cavity geometry and mirrors, while the pump wave mode is defined by the used focusing optics. This knowledge allows to obtain an analytic expression for the threshold of any OPO type and for an arbitrary cavity.

Above threshold two cases can be distinguished:

(i) in a basic SRO the pump mode itself becomes a function of pump power, due to its depletion by signal and idler wave, and only a numerical solution can be given; and
(ii) if two or more waves resonate in one or more cavities (DRO, PR-SRO, TRO), an analytic solution is possible even above threshold.

In the following, the PR-SRO is treated. Figure 4 shows a schematic of a common cavity PR-SRO and the notation; the treatment below also holds for the dual cavity PR-SRO. Without loss of generality, the signal wave is taken to be the signal. For the sake of simplicity, below a further assumption about the spatial mode of the idler will be made.

The pump and signal waves are given by TEM_{00} modes of the cavity (or cavities):

$$\begin{aligned} E_\mathrm{p}(\vec{r}) &= \sqrt{\frac{\omega_\mathrm{p}}{n_\mathrm{p}}} A_{00\mathrm{p}}(z)\psi_{00\mathrm{p}}(\vec{r}) \\ E_\mathrm{s}(\vec{r}) &= \sqrt{\frac{\omega_\mathrm{s}}{n_\mathrm{l}}} A_{00\mathrm{s}}(z)\psi_{00\mathrm{s}}(\vec{r}) \end{aligned} \qquad [18]$$

In the case of a common-cavity PR-SRO, pump and signal resonate in the same cavity and therefore their modes have the same waist position and Rayleigh ranges $z_\mathrm{R} = kw^2/2$, (w is the waist). In the case of a dual-cavity PR-SRO the modes are in principle independent.

Although the idler wave is in general not purely TEM_{00}, it may be approximated by a TEM_{00} mode, whose mode parameters are determined by an optimization criterion. We can then introduce the overlap function:

$$O(z) = \int\!\!\int \mathrm{d}x\,\mathrm{d}y (\psi_{00\mathrm{p}}(\vec{r}))^* \psi_{00\mathrm{s}}(\vec{r})\psi_{00\mathrm{i}}(\vec{r})_{-\infty}^{+\infty} \qquad [19]$$

Under these assumptions, the infinite set of mode eqns [17] reduces to just three coupled equations ($A \equiv A_{00}$):

$$\frac{\mathrm{d}A_\mathrm{p}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{s}(z)A_\mathrm{i}(z)\mathrm{e}^{-\mathrm{i}\Delta kz}O(z) \qquad [20]$$

$$\frac{\mathrm{d}A_\mathrm{s}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{p}(z)A_\mathrm{i}^*(z)\mathrm{e}^{\mathrm{i}\Delta kz}O(z)^* \qquad [21]$$

$$\frac{\mathrm{d}A_\mathrm{i}(z)}{\mathrm{d}z} = -\mathrm{i}\xi(z)A_\mathrm{p}(z)A_\mathrm{s}^*(z)\mathrm{e}^{\mathrm{i}\Delta kz}O(z)^* \qquad [22]$$

where

$$\xi(z) = \frac{d(z)}{c}\sqrt{\frac{\omega_\mathrm{p}\omega_\mathrm{s}\omega_\mathrm{i}}{n_\mathrm{p}n_\mathrm{s}n_\mathrm{i}}} \qquad [23]$$

Two phasematching cases have to be considered: for birefringent phasematching, $d = \mathrm{const}$ and ξ is independent of z. In QPM, the d-coefficient is modulated periodically along z with period Λ. It suffices to consider only the particular term of its Fourier series expansion that gives the smallest wavevector quasi-mismatch. When only the first term is relevant, this is referred to as first-order QPM and $d(z) \simeq \frac{1}{2}\hat{d}(g_1\mathrm{e}^{\mathrm{i}2\pi z/\Lambda} + \mathrm{c.c.})$. This case can be taken into account by replacing $d(z)$ in eqn [23] by $d_\mathrm{eff} = \frac{1}{2}\hat{d}g_1$, and in eqns [20]–[22] the relevant wavevector mismatch becomes $\Delta k = k_\mathrm{p} - k_\mathrm{s} - k_\mathrm{i} \pm 2\pi/\Lambda$, where, for usual materials, the minus sign applies.

We solve eqns [20]–[22] by expanding in powers of the nonlinearity ξ. This is appropriate due to the small gain and implies that the resonant pump and signal wave amplitudes do not change substantially (relative to the respective average amplitude) along the crystal length and around the cavity.

Integrating eqn [22] after approximating the amplitudes on the right-hand side by their values at $z = 0$ yields:

$$\begin{aligned} A_\mathrm{i}(z) &= -\mathrm{i}\xi z A_\mathrm{p}(0)A_\mathrm{s}^*(0)I^*(z) \\ I(z) &= \frac{1}{z}\int_0^z \mathrm{d}z' O(z')\mathrm{e}^{-\mathrm{i}\Delta kz'} \end{aligned} \qquad [24]$$

We insert this result into eqns [20] and [21] and approximate $A_\mathrm{s}(z)$ and $A_\mathrm{p}(z)$ by their values at $z = 0$. The result is

$$A_\mathrm{p}(z) = A_\mathrm{p}(0) - \frac{1}{2}(\xi z)^2 D^*(z)A_\mathrm{p}(0)|A_\mathrm{s}(0)|^2 \qquad [25]$$

$$A_\mathrm{s}(z) = A_\mathrm{s}(0) + \frac{1}{2}(\xi z)^2 D(z)A_\mathrm{s}(0)|A_\mathrm{p}(0)|^2 \qquad [26]$$

$$D(z) = \frac{2}{z^2}\int_0^z \mathrm{d}z'z'I(z')O(z')^* \, \mathrm{e}^{\mathrm{i}\Delta kz'} \qquad [27]$$

This completes the calculation of the propagation through the nonlinear medium. Now we proceed to calculate the propagation around the cavity.

Since the signal wave is resonant with the cavity, we can immediately write the self-consistency equation for the roundtrip:

$$|A_\mathrm{s}(0)| = \sqrt{R_\mathrm{s}(1 - V_\mathrm{s})}|A_\mathrm{s}(L)| \qquad [28]$$

The roundtrip cavity reflectivity and roundtrip loss for the signal wave are denoted by R_s and V_s. At the input mirror (see lower part of Figure 4), characterized by a transmission $T_\mathrm{p} = 1 - R_\mathrm{p}$, the interference of the injected pump wave $A_{\mathrm{p,in}}$ and the internally circulating pump wave must be considered:

$$A_\mathrm{p}(0) = \sqrt{T_\mathrm{p}}A_{\mathrm{p,in}} + \sqrt{R_\mathrm{p}(1 - V_\mathrm{p})}A_\mathrm{p}(L)\mathrm{e}^{\mathrm{i}\phi} \qquad [29]$$

V_p is the pump roundtrip loss. Evaluating eqn [26] at $z = L$ and comparing with eqn [28] yields:

$$|A_\mathrm{p}(0)|^2 = \left[\frac{1}{\sqrt{R_\mathrm{s}(1 - V_\mathrm{s})}} - 1\right]\frac{1}{\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D(L)} \qquad [30]$$

Here the imaginary part of $D(z)$ was neglected since $|\xi L A_\mathrm{p}(0)|^2 \ll 1$. Expression [30] shows that the intracavity pump power is clamped at a fixed value, equal to the value at threshold, independent of the input power.

In eqn [29] we can assume $A_{\mathrm{p,in}}$ to be real without loss of generality, yielding the relation:

$$\sqrt{T_\mathrm{p}}\frac{A_{\mathrm{p,in}}}{A_\mathrm{p}(0)} = 1 - \sqrt{R_\mathrm{p}(1 - V_\mathrm{p})}\mathrm{e}^{\mathrm{i}\phi}\frac{A_\mathrm{p}(L)}{A_\mathrm{p}(0)} \qquad [31]$$

where $\phi = \omega_\mathrm{p}\mathrm{OPL}(\omega_\mathrm{p}, T)/c$. The pump wave is assumed to be exactly resonant; this means that $A_\mathrm{p}(0)$ is maximum. The phase ϕ must therefore be such that:

$$\mathrm{e}^{\mathrm{i}\phi}\frac{A_\mathrm{p}(L)}{|A_\mathrm{p}(0)|} \in \mathbb{R} \qquad [32]$$

Together, with eqn [25], we obtain:

$$\left|\frac{A_\mathrm{p}(L)}{A_\mathrm{p}(0)}\right|^2 = \left|1 - \frac{1}{2}(\xi L)^2 D^*(L)|A_\mathrm{s}(0)|^2\right|^2 \qquad [33]$$

The right-hand side is evaluated neglecting the contribution proportional to ξ^4; inserting the result in eqns [31] and [32] gives:

$$\sqrt{T_\mathrm{p}}\frac{A_{\mathrm{p,in}}}{|A_\mathrm{p}(0)|} = 1 - \sqrt{R_\mathrm{p}(1 - V_\mathrm{p})} + \sqrt{R_\mathrm{p}(1 - V_\mathrm{p})}\frac{1}{2}(\xi L)^2 \mathrm{Re}\, D|A_\mathrm{s}(0)|^2 \qquad [34]$$

Equations [30] and [34] represent the steady-state solution of the PR-SRO: they give the circulating pump and signal powers, from which all desired quantities can be computed.

The external threshold is found by setting $A_\mathrm{s}(0) = 0$ and assuming small losses, T_p, $V_\mathrm{p} \ll 1$

$$P_{\mathrm{p,in}}^{\mathrm{th}} \simeq \frac{T_\mathrm{p} + V_\mathrm{p}}{T_\mathrm{p}} \cdot (T_\mathrm{p} + V_\mathrm{p}) \cdot \frac{(T_\mathrm{s} + V_\mathrm{s})}{E} \qquad [35]$$

with the nonlinearity

$$E = \frac{2\mu_0 d^2}{\pi c^2}\frac{\omega_\mathrm{s}\omega_\mathrm{i}}{n_\mathrm{p}n_\mathrm{s}n_\mathrm{i}}L^2 \,\mathrm{Re}\, D(L) \qquad [36]$$

Note that $\mathrm{Re}\, D(L) = |I(L)|^2$. The second factor in eqn [35] is responsible for the reduction of threshold due to pump resonance. The last factor is the internal pump threshold and constant circulating pump power level:

$$P_\mathrm{p}(0) \simeq \frac{(T_\mathrm{s} + V_\mathrm{s})}{E} \qquad [37]$$

which is also equal to the threshold of a basic SRO. The nonlinearity may be written as

$$E \simeq \frac{4\mu_0 d^2}{\pi c^2}\frac{\omega_\mathrm{s}^2\omega_\mathrm{i}^2}{n_\mathrm{p}^2\omega_\mathrm{p}}Lh_{\mathrm{PA}} \qquad [38]$$

with the dimensionless focusing function h_{PA}. A detailed calculation shows that the value of h_{PA} cannot exceed 1.1 and is close to its maximum when the phase mismatch ΔkL is approximately zero and the pump and signal focusing by the cavity is such that their Rayleigh range $z_\mathrm{R} \simeq L/2$.

The OPO will select the signal cavity mode frequency ω_s and idler spatial mode that minimizes the threshold eqn [35]. Note that this is not necessarily equivalent to a maximization of E, since T_s and V_s may also be dependent on signal frequency, for example, because of the presence of an etalon inside the cavity.

The circulating signal power as derived from eqns [30] and [34], is given in eqn [6].

The signal power extracted from the cavity is $P_{\mathrm{s,out}} = T_\mathrm{s}P_\mathrm{s}(L) \simeq T_\mathrm{s}P_\mathrm{s}(0)$, while the (fully extracted) idler power is given by:

$$P_{\mathrm{i,out}} = P_\mathrm{i}(L) = \frac{\omega_\mathrm{i}}{\omega_\mathrm{s}}EP_\mathrm{s}(0)P_\mathrm{p}(0) \qquad [39]$$

See also

Nonlinear Optics, Applications: Phase Matching. **Optical Parametric Devices:** Overview.

Further Reading

Breitenbach G, Schiller S and Mlynek J (1995) 81% conversion efficiency in frequency-stable continuous-wave parametric oscillation. *Journal of the Optical Society of America B* 12: 2095–2101.

Ebrahimzadeh M and Dunn MH (2000) Optical parametric oscillators. In: Bass M, Enoch JM, van Stryland EW and Wolfe WL (eds) *Handbook of Optics IV*, pp. 2201–2272. New York: McGraw-Hill.

Ebrahimzadeh M, Turnbull GA, Edwards TJ, Stothard DJM, Lindsay ID and Dunn MH (1999) *Journal of the Optical Society of America B* 16: 1499–1511.

Gibson GM, Dunn MH and Padgett MJ (1998) Application of a continuously tunable, cw optical parametric oscillator for high-resolution spectroscopy. *Optics Letters* 22: 40–43.

Klein ME, Laue CK, Lee D-H, Boller K-J and Wallenstein R (2000) Diode-pumped singly resonant continuous-wave optical parametric oscillator with wide continuous tuning of the near-infrared idler wave. *Optics Letters* 25: 490–493.

Klein ME, Gross P, Boller K-J, *et al.* (2003) Rapidly tunable continuous-wave optical parametric oscillator pumped by a fiber laser. *Optics Letters* 28: 920–923.

Müller F, Popp A, Kühnemann F and Schiller S (2003) Transportable, highly sensitive photoacoustic spectrometer based on a continuous-wave dual-cavity optical parametric oscillator. *Optics Express* 11: 2820–2825.

Schiller S, Schneider K and Mlynek J (1999) *Journal of the Optical Society of America B* 16: 1512–1524.

Schreiber G, Hofmann D and Grundkötter W, *et al.* (2001) Nonlinear integrated optical frequency converters with periodically poled Ti:LiNbO$_3$ waveguides. *Proceeding SPIE* 4277: 144.

Special issue (1998) Continuous-wave optical parametric oscillators: materials, devices, and applications. *Applied Physics B* 66(6).

Strössner U, Meyn J-P, Wallenstein R, *et al.* (2002) Single-frequency continuous-wave optical parametric oscillator system with an ultrawide tuning range of 550 to 2830 nm. *Journal of the Optical Society of America B* 19: 1419–1424.

Yariv A and Yeh P (2002) *Optical Waves in Crystals*. Hoboken, NJ: Wiley-Interscience.

Optical Parametric Oscillators (Pulsed)

H Giessen and X P Zhang, University of Bonn, Bonn, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Advantages and Applications of Ultrashort Pulsed OPOs

The advantages of ultrashort pulsed optical parametric oscillators (OPOs) are twofold: on the one hand, it is possible to generate tunable laser radiation with a pulse duration in the femtosecond and picosecond time domain; on the other hand, the short pulses give rise to a high peak intensity of the electric field of the pump wave inside the nonlinear medium. Due to the fact that the OPO single-path gain coefficient reads as:

$$\eta = \frac{P_{\mathrm{Signal}}}{P_{\mathrm{Pump}}} = \frac{8\pi^2 d_{\mathrm{eff}}^2 L^2 I_{\mathrm{P}}}{\varepsilon_0 n_{\mathrm{P}} n_{\mathrm{S}} n_{\mathrm{I}} c \lambda_{\mathrm{S}}^2} \frac{\sin^2(\Delta k L/2)}{(\Delta k L/2)^2} \qquad [1]$$

where I_{P} is the pump intensity, Δk is the phase mismatch between pump, signal, and idler, n_{P}, n_{S}, n_{I}, are the refractive indices at the pump, signal, and idler wavelengths, respectively, and L is the length of the nonlinear medium with an effective nonlinear coefficient d_{eff}, the efficiency grows proportional to the intensity of the pump.

Ultrashort pulsed OPOs can provide widely tunable femtosecond or picosecond pulses, covering the spectral range from the visible to the mid-infrared, which have many applications, such as time-resolved spectroscopy, pump-probe measurements, semiconductor analysis, photochemistry, optical ranging, data transfer, and data processing with high carrier frequency, white-light generation, etc.

Synchronous Pumping

The key issue in operating ultrashort-pulse OPOs is synchronization between the pump pulses and the signal pulses. Consider the following sequence: a pump pulse enters the OPO cavity and generates spontaneously downconverted light at the signal wavelength (see Figure 1, inset). These signal photons travel inside the OPO cavity, entering the nonlinear medium again after exactly one cavity round trip. This light will act as seed light for another downconversion process if a second pump pulse enters the nonlinear medium at exactly the same instance (see Figure 1). The presence of the signal pulse will lead to a much more efficient stimulated downconversion process.

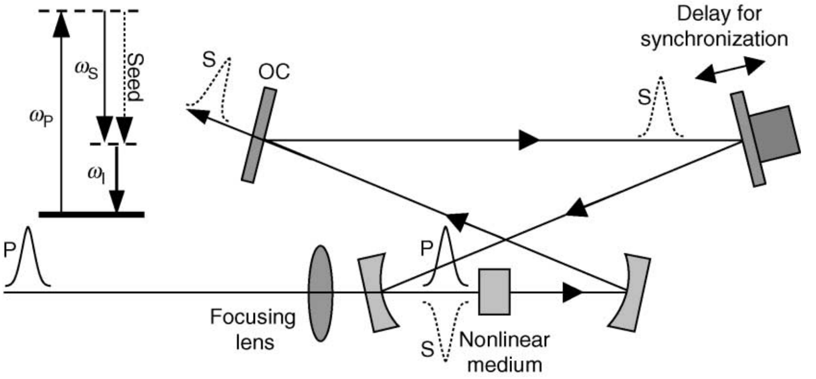

Figure 1 Synchronous pumping of a typical ring-cavity OPO: P, pump pulses; S, signal pulses; OC, output coupler; ω_P, ω_S, ω_I, frequencies of the pump, the signal, and the idler, respectively. Inset: energy level diagram for optical parametric oscillation.

This requires the repetition rate of the pump oscillator and the OPO to be synchronized precisely. Therefore, this pump scheme is called *synchronized pumping*. If the temporal separation between the pump pulses does not exactly equal the roundtrip time of the signal pulses inside the OPO cavity, the generation efficiency will decrease dramatically as the temporal pulse overlap is reduced.

Dispersion in Nonlinear Media and Optical Glasses

Maintaining synchronicity between the pump pulses and the signal pulses is a necessary, but not sufficient prerequisite for successful operation of an ultrashort-pulse OPO. The signal pulses will be temporally broadened during their repeated trip through the nonlinear medium. This will lead to problems, as the newly generated signal pulses are still rather short (as they have been generated by a short pump pulse), but the previously generated signal pulses, which should act as seed pulses and ideally have a similar pulse duration as the pump pulses, have already been broadened temporally, due to dispersion in the nonlinear medium and other intracavity optical elements. Meanwhile, temporal walk-off between the signal (idler) and the pump pulses will take place, resulting in reduced conversion efficiency and broadened signal pulses. Group velocity dispersion (GVD) and group velocity mismatch (GVM) are normally used to characterize the dispersion in nonlinear medium.

The group velocity of the laser pulses propagating in a nonlinear optical medium is defined as:

$$v_G = \frac{c}{n - \lambda \dfrac{dn}{d\lambda}} \qquad [2]$$

where c is the velocity of light in vacuum, n is the refractive index of the nonlinear medium at the center wavelength λ of the laser pulses. Then the GVD can be written as:

$$\frac{dv_G}{d\lambda} = \frac{\lambda v_G^2}{c} \frac{d^2 n}{d\lambda^2} \qquad [3]$$

In most cases, we evaluate the GVD of the optical materials using the group velocity dispersion parameter (k''), which is defined as:

$$k'' = \frac{d^2 k}{d\omega^2} = \frac{\lambda^3}{2\pi c^2} \frac{d^2 n}{d\lambda^2} \qquad [4]$$

where $k = n\omega/c$ is the wave vector, ω is the frequency of the incident wave, and consequently:

$$\frac{dv_G}{d\lambda} = \frac{2\pi c v_G^2}{\lambda^2} \frac{d^2 k}{d\omega^2} \qquad [5]$$

Due to GVD in the nonlinear medium, the spectral components in the broadband spectrum of an ultrashort pulse will travel at different group velocities and will consequently be delayed differently in the crystal, therefore the pulses will become temporally broadened or become chirped. This linear chirp induced pulse broadening effect can be characterized quantitatively by the following formula:

$$\tau_P = \tau_{P0} \sqrt{1 + \left(\frac{z}{z_d}\right)^2} \qquad [6]$$

where τ_{P0} is the incident pulse duration, τ_P is the pulse duration after the pulse propagates in the nonlinear medium over a distance of z, and z_d is called the characteristic length and defined as:

$$z_d = \frac{\tau_{P0}^2}{4 \ln 2 |k''|} \qquad [7]$$

In frequency conversion processes, such as second-harmonic generation (SHG), sum-frequency generation (SFG), difference frequency generation (DFG), optical parametric oscillation (OPO), or optical parametric amplification (OPA), three waves interact with each other while propagating through the nonlinear medium. Due to dispersion of the material, different waves propagate at different group velocities. This effect is generally called group velocity mismatch. In cases of interaction on the femtosecond scale, the interacting pulses may get separated after propagating some distance in the medium, which means that there is a reduced effective interaction length. The GVM is a quantitative evaluation of this effect, which is defined as:

$$\Delta v_G = \left(\frac{1}{v_{G,i}} - \frac{1}{v_{G,j}} \right)^{-1} \qquad [8]$$

where $v_{G,i}$ and $v_{G,j}$ are the group velocities of the two interacting waves i and j. The effective interaction length L_{eff} for wave i and wave j can thus be calculated by the following formula:

$$L_{eff} = \tau_P |\Delta v_G| = \tau_P \left| \left(\frac{1}{v_{G,i}} - \frac{1}{v_{G,j}} \right)^{-1} \right| \qquad [9]$$

where τ_P is the incident pulse duration at full width at half maximum (FWHM). It is assumed that wave i and wave j have a similar pulse duration. This temporal walk-off will lead to two effects. First, the interacting pulses will get separated and the conversion efficiency will be reduced. The second effect is the broadening of the signal pulses. In the case of sub-100 fs operation, to restrict the broadening in pulse length to less than ~10%, the crystal length should be smaller than $L_{eff}/2$.

Figure 2 shows the GVD of some typical OPO crystals, including KTP, KTA, RTA, CTA, PPLN, PPKTP, and LBO. Their Sellmeier equations are as summarized in Appendix A. In addition to the wavelength of the incident beam, the GVD is also dependent on the polarization of the incident beams and the type of phase matching.

We give the GVM values between the pump and the signal (P–S), the signal and the idler (S–I), and between the pump and the idler (P–I) for a KTP crystal in Figure 3.

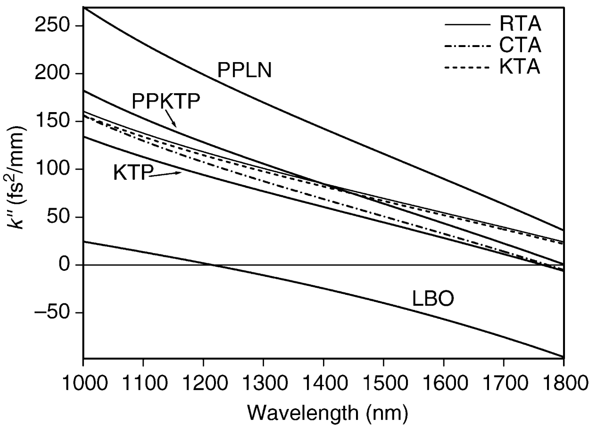

Figure 2 Group velocity dispersion characteristics of some typical OPO crystals.

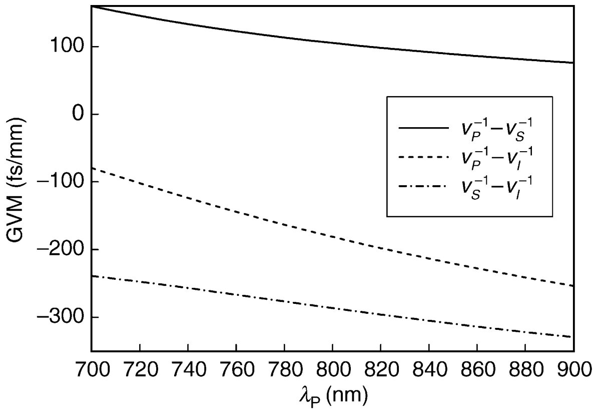

Figure 3 GVM between the interaction waves of a KTP OPO versus pump wavelength (λ_P).

Intracavity Dispersion Compensation and Extra Cavity Pulse Compression

In order to tackle the problem of intracavity dispersion, corrective elements have to be placed inside the cavity in order to compensate the chirp. There have been several solutions to this problem, namely the combination of intracavity prism pairs, chirped mirrors, chirped grating periods (for quasi-phase matching) with oppositely chirped pump pulses, and glass plate dispersion compensators. Certainly, a shorter crystal length will yield a smaller GVM and GVD, but will also give a smaller nonlinear conversion efficiency due to the shorter crystal length (see eqn [1]).

Pump Sources

As pump pulses, mode-locked femtosecond or picosecond lasers are used. The requirements on pulse durations are determined by the desired pulse durations of the OPO signal and by the phase-matching bandwidth of the correspondingly used crystal. Shorter pump pulses normally lead to higher pump power thresholds due to shorter temporal overlap between the interacting pulses. It can also lead to higher instability of the OPO operation due to its higher sensitivity to the cavity length fluctuations. Generally, larger pump bandwidth means larger bandwidth of the signal. Consequently, a potentially shorter signal pulse duration can be achieved if the intracavity dispersion is perfectly compensated. Sometimes it is advantageous for stability to operate the OPO in the slightly negative total GVD regime. In a quasi-phase-matched (QPM) OPO, the phase-matching bandwidth is so large that only a narrow pump bandwidth is required to phase-match a large bandwidth of the signal. Therefore, a pump depletion hole can be commonly observed with the oscillation of the OPO.

Design of Ultrashort Pulsed OPOs

Design of an OPO device may involve the design of the cavity length, cavity mirrors, focal length for focusing the pump beam, intracavity dispersion compensation scheme, frequency stabilization module, crystal length, and some other design depending on the special configuration of the OPO, for example, a heating module (an oven) is required for a PPLN OPO and temperature control module is required for a temperature phase-matched LBO OPO. For an ultrashort pulsed OPO, the most important designs include the design of the crystal length, the design of intracavity dispersion compensation components, and the frequency stabilization control.

Crystal Length

Several factors have to be taken into account when designing a crystal length. The design of a crystal length should be a compromise of all of these considerations. What is special for a femtosecond or picosecond OPO is the dispersion and temporal walk-off between the interacting pulses in the crystal:

1. Effective interaction length: Pump pulse duration, estimated signal pulse duration, temporal walk-off between the pump and the signal give a limit of the crystal length that is characterized by eqn [9].
2. Pulse broadening due to GVM and GVD from the OPO crystal and other intracavity optical glass elements should be estimated. It is reasonable that shorter crystal length is always favorable for ultrashort pulse generation. The crystal length should be as short as possible while maintaining an acceptable conversion efficiency. When designing the crystal length, the available dispersion compensation device should be considered simultaneously.

Dispersion Compensation

For a crystal with a given material and length, the intracavity dispersion can be calculated. The compensation device should introduce an amount of negative GVD such that the total intracavity GVD is equal to or slightly smaller than zero at the desired center wavelength of the signal pulses. Meanwhile, the compensation should be relatively uniform (for instance having a GVD fluctuation of $<50\ \mathrm{fs}^2/\mathrm{mm}$) over the bandwidth of the expected signal spectrum. Then a compensation method (as described above) should be chosen and corresponding device should be designed.

Piezo-Driven Mirror and Frequency Stabilization

To achieve fine tuning of the cavity length, one of the cavity mirrors should be mounted onto a piezoelectric translation stage. When feeding back the signal output of the OPO to this piezo-driven mirror, forming a closed loop to compensate for the drift of the cavity length, a stabilization of the OPO repetition rate is achieved. If the signal is dispersed by a grating, a change in cavity length will also lead to a change in the output spectrum. Using a split photodiode, the difference signal of the dispersed spectrum can be utilized as the measured signal in the feedback loop.

Tuning Characteristics and Methods

Tunability is one of the most important advantages of the OPO over other ultrashort pulsed lasers. A large variety of tuning methods make the OPO more practical and versatile in many applications. For birefringently phase-matched OPOs, the tuning of the signal wavelength can be achieved through changing the pump wavelength (pump wavelength tuning), using an intracavity wavelength selection element (such as a birefringent filter), changing the temperature of the crystal when its refractive index is sensitive to the temperature change (temperature tuning), or changing the orientation of the OPO crystal (angle tuning). Some more methods can be utilized for a QPM OPO, for example through changing the poling period of the crystal (grating period tuning). A special method for an ultrashort pulsed OPO is the so-called cavity length tuning.

As has been discussed above, in synchronous pumping, the OPO must have the same cavity length as the pump laser, so that the signal pulse arrives simultaneously with the pump pulse at the crystal to get the most efficient interaction. Because of dispersion of the nonlinear crystal, the signal will be delayed differently at different wavelengths. In other words, signal pulses have different round-trip times in the cavity at different center wavelengths. When the cavity length of the OPO is changed slightly, the signal has to change its wavelength to keep constant optical path length or constant round-trip time. Cavity length tuning can be achieved easily by tuning the voltage on a piezoelectric transducer (PZT), onto which one of the cavity mirrors is mounted. Therefore, this method is the most common and frequently used to tune the signal wavelength of a synchronously pumped OPO. This is also the basis for stabilizing the operation of a synchronously pumped OPO.

High-Repetition-Rate (HRR) Operation

Synchronously pumping an OPO at high repetition rate produces a high-repetition-rate signal pulse train which extends the application fields of the OPO. Using high-repetition-rate laser pulses (e.g., >1 GHz) as the pump is a straightforward method to run the OPO at a high repetition rate. But a high repetition rate means an extremely short cavity length. For instance, the cavity length is about 30 cm at 1 GHz repetition rate. Such a short cavity length leads to a large mode of the signal beam waist in the crystal. Mode matching between the signal and the pump beams requires that the transverse mode of the pump beam should be slightly larger than that of the signal beam within the interaction zone to maintain stable and efficient oscillation of the OPO. Large modes of the signal and the pump beams mean low interacting intensities, increasing the pump threshold significantly. Using smaller curvature radius of the curved mirrors and shorter focal length of the focusing lens might reduce the mode sizes at the focus in the nonlinear medium, but this will introduce extremely serious astigmatism when the signal beam is reflected from a strongly tilted curved mirror with a quite small curvature radius. In addition to exact synchronous pumping, there are at least two methods to realize high-repetition-rate operation of an OPO. First, pumping at low repetition rate while using a shorter cavity length of the OPO (L_{OPO}) than the pump laser (L_{P}) with $L_{\mathrm{OPO}} = L_{\mathrm{P}}m/N$, where m and N are both integers and $m < N$, as demonstrated in Figure 4a, which is called lower-order synchronous pumping. However, applying this technique, the signal pulses experience more round-trip losses, also resulting in a much higher pump threshold and periodically varying intensities of the output pulses. Second, pumping the OPO at high repetition rate while using a much longer cavity length of the OPO than the pump laser with $L_{\mathrm{OPO}} = L_{\mathrm{P}}N/m$, where m and N are both integers and $m < N$, is called higher-order pumping, as demonstrated in Figure 4b. In higher-order synchronous pumping, the signal pulse train has the same repetition rate as the pump, while the mode-matching problem is removed through a long OPO cavity.

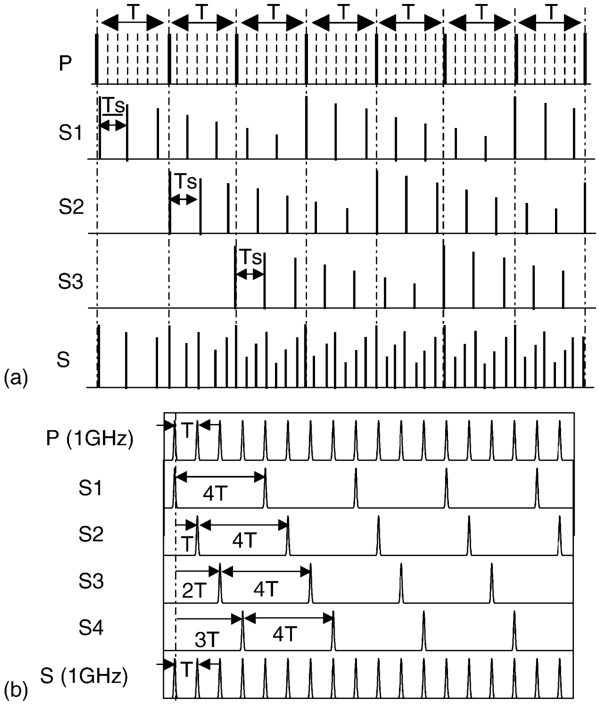

Figure 4 (a) 3/7-order synchronous pumping. Three signal pulse trains (S1–S3) are produced, each having a repetition period (Ts) 3/7 that of the pump pulse train (P): Ts = 3/7*T*. The pump repetition rate is multiplied by a factor of 7 in the signal pulse train (S). (b) Fourth-order synchronous pumping at 1 GHz. The signal pulse train (S) is a combination of four pulse trains (S1–S4), each having a repetition period (4*T*) 4 times that of the pump pulse train (P).

See also

Optical Parametric Devices: Optical Parametric Oscillators (Continuous Wave); Overview.

Appendix A

Sellmeier Equations for some Typical Crystals and Optical Glasses

In the equations, n denotes the refractive index; λ is the wavelength of incident light in μm; $A, B, C, D, A_1, A_2, A_3, A_4, A_5, A_6, B_1, B_2, B_3, B_4, \ldots$, are the Sellmeier coefficients; the subscripts X, Y, Z, stand for the directions of the principal axes of the biaxial crystals.

KTP

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.1146	0.89188	0.20861	0.01320
n_Y	2.1518	0.87862	0.21801	0.01327
n_Z	2.3136	1.00012	0.23831	0.01679

KTA

$$n^2 = A + \frac{B}{\lambda^2 - C} - D\lambda^2$$

	A	B	C	D
n_X	3.1413	0.04683	0.04055	0.01023
n_Y	3.1593	0.04828	0.04710	0.01049
n_Z	3.4435	0.06571	0.05435	0.01460

RTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.04207	1.17785	0.20157	0.01035
n_Y	2.14941	1.09267	0.21462	0.01067
n_Z	2.18962	1.30103	0.22809	0.01390

CTA

$$n^2 = A + \frac{B}{1 - \left(\frac{C}{\lambda}\right)^2} - D\lambda^2$$

	A	B	C	D
n_X	2.34498	1.04863	0.22044	0.01483
n_Y	2.74440	0.70733	0.26033	0.01526
n_Z	2.53666	1.10600	0.24988	0.01711

LiNbO$_3$

$$n_e^2 = A_1 + B_1 f + \frac{A_2 + B_2 f}{\lambda^2 - (A_3 + B_3 f)^2} + \frac{A_4 + B_4 f}{\lambda^2 - A_5^2} - A_6\lambda^2$$

$$f = (T - T_0)(T + T_0 + 546) \text{ and } T_0 = 24.5\,°\text{C}$$

Coefficients	*Values*
A_1	5.35583
A_2	0.100473
A_3	0.20692
A_4	100
A_5	11.34927
A_6	1.5334×10^{-2}
B_1	4.629×10^{-7}
B_2	3.862×10^{-8}
B_3	-0.89×10^{-8}
B_4	2.657×10^{-5}

LBO

$$n^2 = A + \frac{B}{\lambda^2 + C} + D\lambda^2$$

	A	B	C	D
n_X	2.4543	1.1413×10^{-2}	-9.4981×10^{-3}	-1.3900×10^{-2}
n_Y	2.5382	1.2830×10^{-2}	-1.1387×10^{-2}	-1.7034×10^{-2}
n_Z	2.5854	1.3065×10^{-2}	-1.1617×10^{-2}	-1.8146×10^{-2}

$$\frac{dn_X}{dT} = 2.0342 \times 10^{-7} - 1.9697 \times 10^{-8} T - 1.4415 \times 10^{-11} T^2$$

$$\frac{dn_Y}{dT} = -1.0748 \times 10^{-5} - 7.1034 \times 10^{-8} T - 5.7387 \times 10^{-11} T^2$$

$$\frac{dn_Z}{dT} = -8.5998 \times 10^{-7} - 1.5476 \times 10^{-7} T + 9.4675 \times 10^{-10} T^2 - 2.2375 \times 10^{-12} T^3$$

BBO

$$n_o^2 = 1.7018379 + \frac{1.0357554\lambda^2}{\lambda^2 - 0.01800344} + \frac{1.2479989\lambda^2}{\lambda^2 - 91}$$

$$n_e^2 = 1.5920433 + \frac{0.7816893\lambda^2}{\lambda^2 - 0.016067891} + \frac{0.8403893\lambda^2}{\lambda^2 - 91}$$

Sapphire

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

A_1	A_2	A_3	B_1	B_2	B_3
1.023798	1.058264	5.280792	0.00377588	0.0122544	321.3616

Quartz

$$n_{o,e}^2 = A + \frac{B\lambda^2}{\lambda^2 - C} + \frac{D\lambda^2}{\lambda^2 - E}$$

	A	B	C	D	E
n_o	1.28604141	1.07044083	0.0100585997	1.10202242	100
n_e	1.28851804	1.09509924	0.0102101864	1.15662475	100

Fused Silica, BK7, SF2, SF5, SF10, SF11, SF18

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - B_1} + \frac{A_2\lambda^2}{\lambda^2 - B_2} + \frac{A_3\lambda^2}{\lambda^2 - B_3}$$

	A_1	A_2	A_3	B_1	B_2	B_3
Fused silica	0.6961663	0.4079426	0.8974794	0.0046791	0.01351206	97.9340025
BK7	1.03961212	0.23179234	1.01046945	0.0060007	0.0200179	103.560653
SF2	1.40301821	0.23176750	0.93905659	0.0105796	0.04932270	112.405955
SF5	1.46141885	0.24771302	0.94999583	0.0118261	0.05085946	112.041888
SF10	1.61625977	0.25922933	1.07762317	0.0127535	0.05819840	116.607680
SF11	1.73848403	0.31116897	1.17490871	0.0136069	0.06159605	121.922711
SF18	1.56441436	0.29141358	0.96030788	0.0121864	0.05355680	111.451201

Further Reading

Agrawal GP (1989) *Nonlinear Fiber Optics*. Boston, San Diego: Academic Press.

Albrecht TF, Sandmann JHH, Feldmann J, Stolz W, Göbel EO, Hillmer H, Lösch R and Schlapp W (1995) Design and application of a femtosecond optical parametric oscillator for time-resolved spectroscopy of semiconductor heterostructures. *Applied Physics* B 60: 459.

Arbore MA, Marco O and Fejer MM (1997) Pulse compression during second-harmonic generation in aperiodic quasi-phase-matching gratings. *Optics Letters* 22: 865–867.

Arigas D and Reid DT (2002) Efficient femtosecond optical parameteric oscillators based on aperiodically poled nonlinear crystals. *Optics Letters* 27: 851–853.

Bass M, Enoch JM, Van Stryland EW and Wolfe WL (2001) *Handbook of Optics: Volume IV, Fiber Optics and Nonlinear Optics*. New York: McGraw-Hill.

Beddard T, Ebrahimzadeh M, Reid TD and Sibbett W (2000) Synchronously pumped optical parametric oscillator driven by a femtosecond mode-locked fiber laser. *Optics Letters* 25: 1052.

Bierlein JD and Vanherzeele H (1997) Potassium titanyl phosphate: properties and new applications. *Journal of the Optical Society of America B* 6: 622.

Burr KC, Tang CL, Arbore MA and Fejer MM (1997) High-repetition-rate femtosecond optical parametric oscillator based on periodically poled lithium niobate. *Applied Physics Letters* 70: 3341.

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego, CA: Academic Press.

Dmitriev VG, Gurzadyan GG and Nikogosyan DN (1991) *Handbook of Nonlinear Optical Crystals*. New York: Springer-Verlag.

Dunn MH and Ebrahimzadeh M (1999) Parametric generation of tunable light from continuous-wave to femtosecond pulses. *Science* 286: 1513.

Fenimore DL, Schepler KL, Zelmon D, Kuck S, Ramabadran, Von Richter P, and Small D (1996) Rubidium titanyl arsenate difference-frequency generation and validation of new Sellmeier coefficients. *Journal of the Optical Society of America B* 13: 1935.

Ghost G (1995) Temperature dispersion of refractive indices in β-BaB_2O_4 and LiB_3O_5 crystals for nonlinear optical devices. *Journal of Applied Physics* 78: 6752.

Ghost G (1999) Dispersion-equation coefficients for the refractive index and birefringence of calcite and quartz crystals. *Optics Communications* 163: 95.

Hebling J, Giessen H, Linden S and Kuhl J (1997) Mirror-dispersion-compensated femtosecond optical parametric oscillator. *Optical Communications* 141: 229.

Jundt DH (1997) Temperature-dependent Sellmeier equation for the index of refraction, ne, in congruent lithium niobate. *Optics Letters* 22: 1553.

Kato K, Umemura N and Tanaka E (1997) 90° phase-matched mid-infrared parametric oscillation in undoped $KTiOAsO_4$. *Japanese Journal of Applied Physics 2* 36(4A): L403.

Loza-Alvarez P, Reid DT, Faller P, Ebrahimzadeh M and Sibbett W (1999) Simultaneous femtosecond-pulse compression and second-harmonic generation in aperiodically poled KTiOPO4. *Optics Letters* 24: 1071.

Matuschek N, Gallmann L, Sutter DH, Steinmeyer G and Keller U (2000) Back-side-coated chirped mirrors with ultra-smooth broadband dispersion characteristics. *Applied Physics B* 71: 509.

Mokhtari A, Fini L and Chesony J (1987) Efficient frequency mixing of a cw femtosecond laser synchronously pumped by a frequency doubled Nd:YAG laser. *Optics Communications* 61: 421.

Nebel A, Fallnich C, Beigang R and Wallenstein R (1997) Noncritically phase-matched continuous-wave mode-locked singly resonant optical parametric oscillator synchronously pumped by a Ti:sapphire laser. *Journal of the Optical Society of America B* 10: 2195.

Schubert M and Wilhelmi B (1978) *Nonlinear Optics and Quantum Electronics*. New York: John Wiley & Sons.

Seres J and Hebling J (2000) Nonstationary theory of synchronously pumped femtosecond optical parametric oscillators. *Journal of the Optical Society of America B* 17: 741.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: John Wiley & Sons.

Sutherland RL (1996) *Handbook of Nonlinear Optics*. New York: Marcel Dekker Inc.

Szipöcs R and Kohazi-Kis A (1997) Theory and design of chirped dielectric laser mirrors. *Applied Physics B* 65: 115.

Tang Y, Cui Y and Dunn MH (1995) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 12: 638.

Tomaru T and Petek H (2000) Femtosecond Cr4+ YAG laser with an L-fold cavity operating at a 1.2-GHz repetition rate. *Optics Letters* 25: 584.

Tomaru T and Petek H (2000) Thermal dependence of the principal refractive indices of lithium triborate. *Journal of the Optical Society of America B* 18: 388.

Velsko SP, Webb M, Davis L and Huang C (1991) Phase-matched generation in lithium triborate (LBO). *IEEE Journal of Quantum Electronics* 27: 2182.

Yariv A (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Yariv A (1989) *Quantum Electronics*. New York: John Wiley & Sons.

Zernike F and Midwinter JE (1973) *Applied Nonlinear Optics*. New York: John Wiley & Sons.

Zhang XP, Hebling J, Kuhl J, Rühle WW, Palfalvi L, Giessen H (2002) Femtosecond near-IR optical parametric oscillator with efficient intracavity generation of visible light. *Journal of the Optical Society of America B* 19: 2479.

Zhang XP, Kuhl J, Hebling J, Bartels A, Nau D, Rühle XP, Giessen H (2002) 1-GHz repetition-rate femtosecond optical parametric oscillator. *Applied Physics Letters* 80: 1873.

OPTICAL PROCESSING SYSTEMS

D Arbel and N S Kopeika, Ben-Gurion University of the Negev, Beer Sheva, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modern optical systems find many applications in sensing, recording, storage, and transmission of data. These applications stimulate the development of many new optical devices, components, materials, and applied technologies. The basic element of an optical system is the light. The physical implications of the light in optical systems are very analogous to those of electricity, hence, light signals are quite similar to electric signals. High-speed spatial light components are being developed for more stringent applications in coherent optical processing systems, where they may be used in the input plane or in the Fourier transform plane. Research in this area is directed towards the design of robust and compact correlators.

Novel transforms for information processing (also known in electronics) include Fourier and wavelet transforms. Such optical transformations serve as the basic platform for linear optical system processing. Optical devices and systems for information processing include spatial light modulators, which are used to implement optical information processing systems. There are holographic applications in information processing, display systems, memory systems, optical computing systems, and imaging systems. Such devices and systems are based on applications of novel materials for information processing. Optical and digital techniques are used for optical system design, optical computing, image processing, and even for encryption and security systems.

The imaging system is an optical system and integrates hardware as well as software. Figure 1 presents a general model of an extensive imaging system. The starting point of the system is the object (or the light source). An imaging device, such as camera or optical sensor detects the energy radiated in electromagnetic waves from the light source. The image (or series of images – video) formed in the imaging device depends on the received irradiance of the image propagated through the imaging path and is correspondingly diminished from that in the object plane. The properties of the optical medium (transmission path) such as atmospheric properties, background light, motion, and vibration in both the object and the image planes, and the quality of the imaging device, all affect the quality of the images.

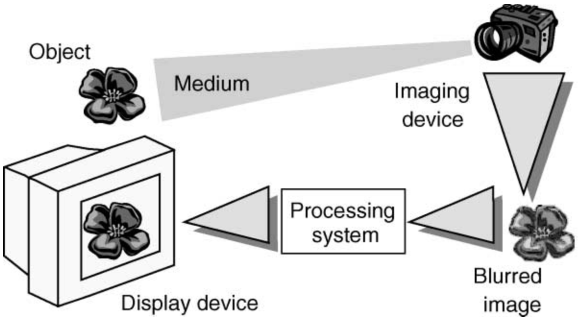

Figure 1 General description of an imaging system.

There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image, as well as atmospheric phenomena that cause blurring of detail. Both types of phenomena degrade image quality.

Image blurring caused by vibration is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. In order to obtain a well processed imaged, as similar as possible to the ideal image (the source), image processing techniques such as image enhancement, restoration, and filtering are implemented in an optical processing system which is a combination of software (algorithms) and hardware (electronic and optical components). The processing procedure involves mathematical and especially optical transformations. The output of the processing system is presented by using a display device or hardcopy. The development of imaging systems (imaging devices and displays) which include integrated optical components and processing units, assist in improving image quality at early stages, and as a result reduce the need of unique processing systems and may serve as the standalone device.

The key issues which must be addressed in the design, evaluation, and implementation of practical systems for optical processing, are provided here in a detailed review. Thus, the review is focused on imaging where considerations such as optical transformations, optical computing, effects of diffraction, scatter, absorption, and design on the performance of optical systems, imaging and the optical transfer function (OTF) for image restoration, and filtering, are all discussed in detail.

Digital Processing and Control

The Fourier transform properties of Fraunhofer diffraction permit data processing operations carried out optically that can, in many situations, be more advantageous than digital signal processing. The primary advantages of optical techniques over electronic analog and digital signal processing, are best realized when the information to be processed has two degrees of freedom. Optical systems correspondingly can have two degrees of freedom. Thus, for handling pictorial data, they can be intrinsically superior to electric or electronic systems, which have only time as an independent variable. This two-dimensional (2D) processing property can be particularly valuable in radar applications where previously velocity and position data had to be processed separately in order for sufficient accuracy in the estimation of each set of data to be obtained.

Many of the basic operations that can be performed by optical computers can also be done with electronic systems. However, if large quantities of wideband data are to be processed, the electronic techniques can often be prohibitively expensive and time consuming. The optical computer is faster than a digital computer because it operates on signals that are functions of positions, not time. The input to an optical data processor is a pattern of light irradiance that varies over some area, not light irradiance that varies as a function of time.

Thus, optical data processing can be very appropriate for such lengthy operations as the calculation of complicated Fourier transforms and the processing of side-looking radar signals. In addition to mathematical operations such as differentiation, integration, and producing Fourier transforms, the two classes of applications for which optical signal processing is most often used are filtering and correlation.

Optical Computing

Optical computing (OC) means the use of light as a primary means for carrying out numerical calculations, reasoning, artificial intelligence, etc. The history of OC is linked to that of radar systems. OC systems received a great push from the invention of laser in 1960. In the 1960s, the first schemes for all-optical digital computers were proposed. Since the 1990s, researchers have been pursuing the development of computers based on optical technology, and emphasis has shifted to optical interconnection of arrays of semiconductor smart pixels. OC provides the first in-depth review of the possibilities and limitations of optical data processing. The partial listing of those points that make OC appealing are direct image processing, massive parallelism and connectivity, immunity to electromagnetic interference (EMI), speed, size, and cost. There are three distinct trends in OC: special purpose analog optical systems, general purpose digital optical systems, and hybrid optical/electronic systems.

Innovations such as the electro-optic spatial filter, the 2D spatial light modulator, and the variable electro-optic mirror provide the fundamental building blocks for both analog optical computing (AOC) and digital optical computing (DOC) technologies. Optical techniques can provide solutions and a number of ways of extending the information processing capability of electronics. The optical processors have to be compatible with existing electronic systems. Free space digital optics is one direction that provides some valuable solution. DOC requires the use of nonlinear optics.

All-optical processing refers to optically based processors employed nonlinear optical resources either of liquid crystal spatial light modulators (SLMs) or nonlinear interference filters (NLIFs). Electro-absorption devices are a bridge between all-optical demonstrators and optically interconnected smart pixels (logic and local on-chip interconnection is electrical whereas chip to chip interconnection is optical). Optically interconnected smart pixels have several advantages like faster data acquisition and low power consumption. A major breakthrough in nonlinear optics is essential to increase the use of optics within computing and the manufacture of optical digital computers.

An optically bistable device used for photonic switching is the self electro-optic effect device (SEED). It is the analogous component for the transistor in optical processing. A SEED consists of multiple elements, biased by an external voltage, which creates an external field that shifts the wavelength of the onset of absorption, thus varying the intensity of the transmitted light. SEEDs rely on changes in the optical transmission of a semiconductor induced by an applied electric field. SEEDs can be configured in pairs (connected electrically) so that a beam of light switching one device can cause a complementary switch in the transmission of the other. Hence, a small change in the intensity of one beam can cause a large change in intensity of the other. This kind of configuration is known as symmetric SEEDs, and acts like an electronic flip-flop and permits logic operations NAND and NOR to be carried out on pairs of optical input signals. With symmetric SEEDs, the higher the input optical power, the faster the switching speed. Cellular logic image processor (CLIP) is implemented using the symmetric SEEDs. CLIP computer architecture is designed to permit parallel information processing, in which logic operations are performed on each element of the array simultaneously.

Holography

Holography is a photographic method of recording information (two mutually coherent beams) about an object, in which one beam is the object beam and the other a reference beam, which enables the construction of objects in three dimensions. Holography relies on the encoding of the object formation in a set of complex interference fringes formed by the interaction of a plane coherent reference wave with the wavelets diffusely scattered by the object.

The microscopic interference fringes recorded on a high resolution photographic plate after development and processing become a hologram. Hologram applications assist the study of small deformations of objects, to replace conventional optical elements such as lenses and beamsplitters, for data/information storage, and use of photo-refractive materials in analog computing, object recognition, and correlation, etc.

The purpose of recording a hologram is not optical data processing but rather a lensless method of forming an image. The uniqueness of a hologram as compared to an ordinary photographic image, for example, is that the latter contains amplitude information only. A hologram, because it involves interference of the complex field amplitude (CFA) of a given scene with a reference beam, records both amplitude and phase of the CFA of the given scene as seen within its field of view (FOV) by a square-low photosensitive device such as a film.

The film is exposed within its region of linearity whereby transmission is proportional to irradiance for a given time exposure. The transparency resulting from the film is the hologram. To view the holographic image, the hologram is illuminated with a reference wave, which causes the CFA of the given scene (both amplitude and phase) to be reconstructed. Since hologram reconstruction involves the CFA of the given scene in its entirety, depth information is included as well. Furthermore, it is even possible to view the holographic scene from different physical points of view and to see around objects in the scene. This ability, however, is limited by the solid angle subtended by the object when the hologram is recorded. Hologram reconstruction cannot supply more information than that recorded in the hologram itself.

Image Processing

Figure 1 demonstrates a general behavior of an imaging process. The ‘Processing System’ block in Figure 1 is expanded and presented in more detail by Figure 2. Figure 2 demonstrates in a flow chart, a general process occurring within the processing system. The input to the processing system is a degraded image (or set of images or video signal) such as blurred image, noisy image, image with clutters, multi-spectral image, etc. The input depends on exterior (environmental conditions) and interior (optical, electronic, and mechanical) effects of the imaging system. It may be generated as an analog signal or digital signal. As a result, the output of the imaging system is expected to be a processed (filtered, enhanced, estimated, restored, cleared, etc.) image.

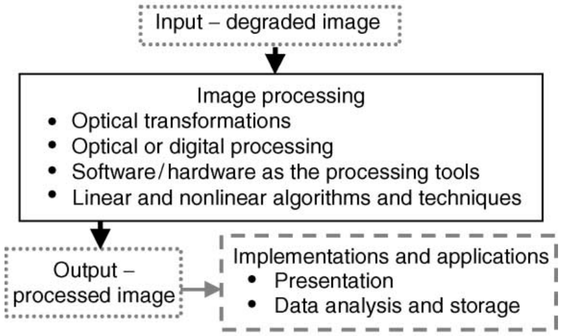
Input – degraded image
Image processing
• Optical transformations
• Optical or digital processing
• Software/hardware as the processing tools
• Linear and nonlinear algorithms and techniques
Output – processed image
Implementations and applications
• Presentation
• Data analysis and storage

Figure 2 Image processing.

Implementations and applications depend on the output obtained by the imaging system. The output can be presented by a display device and/or hardcopy. The output can be stored for future assignments and may also serve for data analysis such as detection, recognition, and identification of patterns and other details of interest (mapping, target acquisition, defects and imperfections in the human body, meteorology, security, intelligence, etc.).

Spatial units are reciprocal to time, and are cycles/seconds or Hz. The spatial units are also reciprocal to the domain of the information being sought. Spatial signals are 2D or even 3D on occasions. 2D spatial transforms may actually exist physically in the spatial plane. Novel transforms for information processing include Fourier and wavelets transforms. Such optical transformations serve as the basic platform for linear optical system processing. Image processing techniques are based on the image spatial frequencies. In most techniques the degraded image is transformed using an optical transformation, as a primary stage before any other process is being conducted. Optical (inverse) transformations also serve as the last processing tool before the output. Practically, optical transformations are involved and integrated in most processes related to image processing.

Image processing refers to both digital and optical image processing, and also considers algorithms and techniques, which are implemented by an integration of software and hardware. Image processing is an extremely broad field. It considers digital image processing, 2D and 3D image processing and pattern recognition, real-time processing, neural networks, image enhancement/improvement, image restoration, estimation, filtering for noise removal, linear and nonlinear filtering techniques for variety of implementations, optimum filtering, etc. Image processing also include varieties of algorithms such as linear and nonlinear algorithms, algorithms for pattern recognition, linear and nonlinear techniques for pattern recognition, neural networks based algorithms, etc. Both optical and electronic implementations (hardware) are still under investigation. Nevertheless, image processing based on software presents reliable results and hence, can be transformed into hardware components, which may become an integrated part of the processing hardware.

Optical Transformations

Mathematical transformations are applied to signals to obtain further information from the signal that is not readily available in the raw signal. Most of the signals, in practice, are time-domain signals in their raw format. That is, whatever that signal is measuring, is a function of time. In many cases, the most important information is hidden in the frequency content of the signal. The frequency spectrum of a signal is basically the frequency components (spectral components) of that signal. The frequency spectrum of a signal shows what frequencies exist in the signal.

There are many transforms that are widely used, often by engineers and mathematicians, in solving problems in science and engineering. Fourier, Laplace, Hilbert, Wigner distributions, Radon, and wavelet transforms, constitute only a small portion of a huge list of transforms that are available. Every transformation technique has its own area of application, with advantages and disadvantages.

Fourier Transform

The Fourier transform (FT) is probably the most popular linear transform being used. The FT, a pervasive and versatile tool, is used in many fields of science as a mathematical or physical tool to alter a problem into one that can be more easily solved. It is used in linear systems analysis, antenna studies, optics, random process modeling, probability theory, quantum physics, and boundary-value problems and has been successfully applied to restoration of astronomical data. Some scientists understand Fourier theory as a physical phenomenon, not simply as a mathematical tool. In some branches of science, the FT of one function may yield another physical function. The FT, in essence, decomposes or separates a waveform or function into sinusoids of different frequency which sum to the original waveform. It identifies or distinguishes the different frequency sinusoids and their respective amplitudes. There are functions for which the Fourier transform does not exist; however, most physical functions have a FT, especially if the transform represents a physical quantity. Other functions can be treated with Fourier theory as limiting cases. Many of the common theoretical functions are actually limiting cases in Fourier theory.

Wavelets

Wavelets were developed independently in the fields of mathematics, quantum physics, electrical engineering, and seismic geology. Interchanges between these fields have led to many new wavelet applications including image processing. Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. Wavelet transforms comprise an infinite set. The different wavelet families make different trade-offs between how compactly the basis functions are localized in space and how smooth they are. Within each family of wavelets are wavelet subclasses distinguished by the number of coefficients and by the level of iteration.

The wavelet transform (WT), which is also a linear operation, provides the time-frequency representation. Often a particular spectral component occurring at any instant can be of particular interest. In these cases it may be beneficial to know the time intervals at which these particular spectral components occur. WT is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. The WT was developed as an alternative to the short-time FT (STFT). The WT was developed to overcome some resolution related problems of the STFT. The frequency and time information of a signal at some certain point in the time-frequency plane cannot be known. What spectral component exists at any given time instant is a mystery. Therefore, it is best to investigate what spectral components exist at any given interval of time. This is a problem of resolution, and it is the main reason why researchers have switched to WT from STFT. STFT gives a fixed resolution at all times, whereas WT gives a variable resolution as follows: higher frequencies are better resolved in time, and lower frequencies are better resolved in frequency. This means that a certain high frequency component can be located better in time (with less relative error) than a low frequency component. On the contrary, a low frequency component can be located better in frequency, compared to high frequency component.

Wavelet versus Fourier Transforms

Both transforms are linear and the mathematical properties of the matrices involved in the transforms are similar. The inverse transform matrix for both transforms is the transpose of the original. Wavelets have advantages over traditional Fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes. The most interesting dissimilarity is that individual wavelet functions are localized in space, Fourier sine and cosine functions are not. This localization feature, along with wavelets localization of frequency, makes many functions and operators using wavelets ‘sparse’ when transformed into the wavelet domain. This sparseness, in turn, results in a number of useful applications such as data compression, detecting features in images, and removing noise from time series. One way to see the time-frequency resolution differences between the FT and the WT is to look at the basis function coverage of the time-frequency plane. In order to isolate signal discontinuities, some very short basis functions are needed. At the same time, in order to obtain detailed frequency analysis, some very long basis functions are needed. A way to achieve this is to have short high-frequency basis functions and long low-frequency ones. This happy medium is obtained with WT. FT utilizes just the sine and cosine functions (single set of basis functions). Therefore, its new domain contains basis functions that are sines and cosines. On the other hand, WT includes an infinite set of possible basis functions. Hence, its new domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets. The basis functions are localized in frequency, making mathematical tools such as power spectra and scalegrams useful at picking out frequencies and calculating power distributions. Thus wavelet analysis provides immediate access to information that can be obscured by other time-frequency methods such as Fourier analysis.

Physical Properties and Analog/Digital Mitigation Processing in Nonideal Systems

The weakest link in most imaging systems is time varying distortions, such as the Earth’s atmosphere, motion, or vibrations. To obtain clear imaging through time varying distortions, it is useful to characterize the physical degradation phenomena using an OTF. It is often important to use the appropriate filter based on the OTF magnitude, modulation transfer function (MTF), to correct (restore) the degraded image (using software or hardware). On the other hand, it may be useful to incorporate real-time phase correction into the imaging system such as adaptive optics.

Optical Transfer Function for Incoherent Imaging

In electronics, the transfer function is redefined as the Fourier transform of the impulse response. The OTF is defined similarly, but is normalized to its own maximum value which normally occurs at zero spatial frequency, that is:

$$\begin{aligned}\mathrm{OTF} &= \tau(\omega) = \tau(\omega_x, \omega_y)\\ &= \frac{\iint_{-\infty}^{\infty} s(x', y') \exp[-j(\omega_x x' + \omega_y y') \mathrm{d}x' \, \mathrm{d}y']}{\iint_{-\infty}^{\infty} s(x', y') \mathrm{d}x' \, \mathrm{d}y'}\\ &= \frac{S(\omega_x, \omega_y)}{S(0,0)} = \mathrm{MTF} \exp(j\mathrm{PTF}) \qquad [1]\end{aligned}$$

where $S(\omega_x, \omega_y)$ is the FT of the spread function or impulse response $s(x', y')$. Ordinarily, for artificial imaging systems, the OTF is maximum at zero frequency. The magnitude of the OTF is called the MTF. The name refers to spatial modulation of light irradiance. If there is no such modulation, irradiance is uniform and there is no image. Thus, spatial modulation is related to image quality and MTF is a measure of the ability of an imaging component or system to transfer such spatial modulation from input (object plane) to output (image plane).

There is also a phase transfer function (PTF) whose importance in resolution is not nearly as widespread as that of MTF but which, nevertheless, cannot be neglected. Spatial phase determines image position and orientation, rather than amount of detail. If a target is displaced bodily in the image plane such that each part of the target image is displaced by the same amount, the target image is not distorted. However, if portions of the target image are displaced more or less than other portions, then the target image is distorted. This information is contained in the PTF.

The physical implications of the OTF are very analogous to those of electronics transfer functions. Both permit determination of output for any given input. In both cases, since the FT of the delta function is a constant (unity) amplitude, a delta function input permits indication of the frequency response of the system, that is, which frequencies pass through the system unattenuated and which are attenuated and by how much. Perfect fidelity or resemblance between input and output requires infinite bandwidth so that an impulse function is obtained at output for impulse function input. In imaging, an impulse formation at output is a point image. This means that, to obtain a point image for a point object, infinite spatial frequency bandwidth is required of the imaging system. Physically, this means that because of diffraction effects optical elements such as lenses, field stops, and mirrors must be of infinite diameter so as to eliminate diffraction at the edges. This, of course, is not attainable.

Pixel Modulation Transfer Function

To measure the point spread function (PSF), it is important to determine which size object can be considered a point object. This is affected primarily by instrumentation pixel size. The pixel is the smallest element recordable in image space. The brightness value represents average irradiance over that small portion of the image scene. Pixel size is often related to detector size. If only a portion of the detector is illuminated, the output current is equivalent to that obtained for the same total radiant power absorbed by the detector but averaged over the entire detector area. No detail smaller than a pixel can be resolved in the image. In practice, dead space between pixels violates isoplanaticism (spatial stationarity), which is a requirement for linear systems. Nevertheless, dead space is small and such effects are often neglected in image system characterization. However, it is possible to overcome distortions in MTF measurements arising from this lack of isoplanaticism by using a 'white' spatial noise random object. This gives rise to a white noise random image. The ratio of the image power spectral density to object power spectral density is equal to the square of the system MTF. In this way the lack of isoplanaticism is overcome by the spatial randomness of the object and image.

Pixel size strongly affects system MTF. If a pixel represents a point image, then pixel size and shape are minimum spread function. A best case OTF for such an imaging system is then a normalized FT of the pixel shape. For example, a square pixel yields a two-dimensional sinc MTF whose width increases as pixel size decreases.

Optical Transfer Functions for Image Motion and Vibration

Image blurring caused by vibrations is a factor whose influence on resolution is often significant in imaging systems that involve mechanical motion. A stabilized system, which obtains image motion corrections by moving optical elements so as to counteract sensor motions, has a residual error typically on the order of 2500 μrad/s involving pitch, roll, yaw, and forward motion compensation errors. The greatest errors usually result from rotational vibration. A stabilized mount-type system has a combined error typically on the order of 100 μrad/s. Even where no motion is involved, as in large fixed-position astronomical telescopes, vibrations deriving from thermal gradients in the walls and other parts of the telescope can be the limiting factor in image resolution.

As a consequence of such image motion, in many high-resolution systems, despite the use of high-quality sensors, resolution is limited by image motion and, as a result, the high-resolution capability of the sensor may be wasted. It can be a significant waste of means to use an expensive high-resolution sensor in such a situation unless the blur can be corrected with image restoration. In imaging system design, a convenient engineering tool is OTF. The overall system OTF is limited by the OTF of the weakest link. The formulation of such image motion blur into an OTF format is thus very convenient for system design and system analysis purposes, as well as for image restoration.

OTFs can describe image quality for many forms of image motion such as linear (uniform) motion, sinusoidal vibrations at high vibration frequencies, sinusoidal vibrations at low vibration frequencies, acceleration, etc. The OTF for any type of image motion can be calculated numerically in real time provided the motion function is known. This is usually determined easily via a gyroscope or microelectronic accelerometer.

In general, the motion of interest is not the actual relative motion between object and imager, but rather the relative motion between imager and image plane. The latter motion is that observed in the image plane. Both of these motions are related to each other by the image system magnification. Hence, image plane relative motion is usually much less than the actual relative motion. However, any motion during an exposure gives rise to blurring of image plane detail. One question to be considered is how much blur can take place as a result of motion and still not be noticeable because of inherent blur deriving from the PSF of the rest of the imaging system. In other words, assuming that image motion and vibration blur cannot be removed completely, how much is tolerable before resolution and target acquisition capability are affected? This depends on focal length and system zoom.

A general method of OTF calculation is obtained by the line spread function (LSF) derived from image motion transverse to the optical axis. The MTF is derived as the modulus of the MTF or FT of the LSF, and the PTF is derived as the phase of the OTF. The LSF is the probability density function (PDF) of the histogram of the relative transverse displacement between object and sensor beginning at time t_s and ending at time $t_s + t_e$ where t_s is measured from the instant the sensor is first exposed and t_e is exposure time.

Recently a method of calculating motion OTF analytically has been developed for any type of motion. Also, methods to calculate motion OTF without a motion sensor have been derived for a sequence of consecutive images and even from a single motion-blurred image. All three methods have been used successfully in image restoration.

Optical Transfer Functions for the Atmosphere

Many properties of the atmosphere affect the quality of images propagating through it. There are atmospheric phenomena that give rise to attenuation of the irradiance of the propagating image, thus reducing the contrast of the final image. There are also atmospheric phenomena that cause blurring of detail. Both types of phenomena prevent small detail from being resolved in the final image, thus degrading image quality.

Electromagnetic (EM) wave absorption and scattering by the constituent gases and particulates of the atmosphere and airborne particulates give rise to attenuation. Scattering of photons by airborne particulates is manifested as deflections of the photons to directions other than that of the original propagation. If such scattering causes the deflected photons to miss the imaging receiver, then the scattering is manifested as attenuation. The received irradiance of the image propagated through the atmosphere is correspondingly diminished from that in the object plane. However, if the light scattering is at very small angles with respect to the original direction of the propagation, and several such small-angle scattering events take place, then forward-scattered radiation can take round-about paths and still be received by the imaging system together with the unscattered radiation. The net effect is image blurring caused by a multitude of angles of arrival at the imaging receiver of radiation emanating from the same point in the object plane. Many multiple-scattering paths give rise to a relatively large blurred-point image rather than a fine sharp-point image. Adjacent image plane points can then appear for a single object plane point, thus degrading image resolution. Such aerosol scatter blur is known as the adjacency effect.

Another effect of light scattering, particularly at large angles, is increased path radiance. The atmospheric background irradiance at wavelengths less than 2–4 μm, consists through the day of scattered background light such as sunlight. This atmospheric background radiation is imaged over the same image space as the received irradiance from the target plane. The result is decreased contrast of the target plane scene. At wavelengths larger than 4 μm, most of the path radiance is not scattered sunlight, but rather thermal emission of the atmospheric constituents, gases in particular. This thermal path radiance decreases contrast in thermal imaging systems, particularly in the longwave infrared (8–13 μm).

Turbulence results from variations in the atmospheric refractive index, which are random in both time and space. These refractive index fluctuations are caused by local fluctuations in atmospheric temperatures, pressure, humidity, etc., all of which are affected by wind speed, which in itself is also random in time and space. Generally, the larger the temperature and humidity gradients, the more serious the image resolution degradation caused by turbulence. The lower the wind speed, the stronger the impact of turbulence. Scattering and turbulence are basically two very different physical mechanisms, each exhibiting quite different properties. Time dynamic properties such as scintillations and image dancing derive from turbulence, while aerosol scatter light changes only slowly with time.

A point that must be strongly emphasized is that there is no such thing as a purely turbulent, purely scattering, or purely absorption atmosphere. All three properties occur simultaneously in the real-world atmosphere. Modeling and interpretation of experimental results must relate to all three processes. Any attempt to pretend otherwise is not realistic. In general, blur is caused by turbulence and small-angle scatter of light by aerosols and particles. Attenuation is caused by large-angle scatter and by absorption.

Generally turbulence blur is dominant for images near the ground, where ground heating gives rise to the turbulence. Aerosol blur is usually dominant at higher elevations, even as little as several meters above ground, depending on weather conditions and wavelength.

The product of both aerosol and turbulence MTFs (in the spatial frequency domain, after FT) yields approximately atmospheric MTF. The atmospheric MTF can be predicted according to atmospheric (meteorological) parameters. Aerosol MTF calculation is based on optical depth, which includes parameters such as scattering and absorption coefficients through the atmosphere over the imaging path. Turbulence MTF calculation is based on the refractive index structure coefficient, which includes parameters such as temperature, humidity, wind speed, etc.

By referring to the edge response in a given scene the atmospheric MTF may be calculated. The edge separating the wide white and black bars is imaged from the scene. The image of the edge is not a step function because of atmospheric blur. The gradient of the edge image is the LSF, from which the overall atmospheric MTF can be obtained after the FT of the LSF is divided by the hardware MTF.

Spatial Filtering

Spatial filtering is a process by which we can alter properties of an optical image by selectively removing certain spatial frequencies that make up an object, for example, filtering video data received from satellite and space probes, or removal of raster from a television picture or scanned image. In digital image processing, the term image refers to a two-dimensional light-intensity function, where amplitude at spatial coordinates (x, y) gives the intensity (brightness) of the image at that point.

Digital image processing techniques are oriented to transform, enhance, restore, or encode an image. Digital image restoration is on the basis of the OTF or even MTF alone describing the image degradation. Resolution is limited usually not by electronics or optics, but by image motion and vibration. For long-range imaging, the image quality may be limited by atmospheric effects.

By image restoration filters we refer primarily to spatial frequency filters placed in the spatial frequency plane in the optical data processor. Although data processing is done optically, the same mathematical concept is usually carried out with a digital computer. It is also possible to do restoration in the spatial domain. The simplest image restoration filter is the inverse filter. When noise becomes a problem, there are often used three classes of minimum mean square error estimations (MMSEE), which are the noncausal Wiener filtering, causal Wiener filtering, and Kalman filtering. Kalman and Wiener filters are proposed to obtain the optimum solution for the MMSEE problem in linear systems. Wiener filters are applied in linear time invariant systems, and Kalman filters in linear systems which may be either time invariant or time variant. A novel area of application for Wiener filtering is image restoration. On the other hand, if the estimation processes are partial and an ideal image is not obtainable, a Kalman filter may present better restoration results than a Wiener filter.

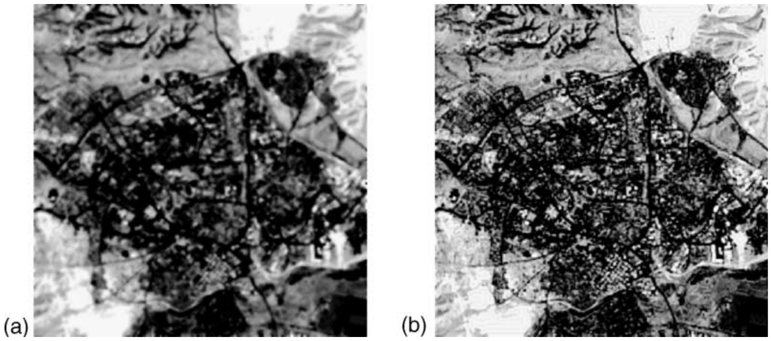

Figure 3 An example for satellite image restoration. (a) Original image of Beer-Sheva, Israel, from Landsat 7 satellite; (b) restored image after Kalman filter using aerosol and turbulence MTFs. Reproduced with permission of the American Society for Photogrammetry and Remote Sensing from Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing* (PE&RS) 70(1): 91–100.

Figure 3 demonstrates the restoration result for a satellite image. Figure 3a presents an original Landsat 7 enhanced thematic mapper plus (ETM+) image of the city of Beer-Sheva, Israel. The image was recorded from 750 km satellite elevation. The image was recorded using an ETM+ sensor with 15 m ground resolution, and its wavelength range is 520–900 nm (Panchromatic band). Figure 3a is of 256 × 256 pixel size, which is a segment defined as a region of interest (ROI) in the original image. The scene was recorded on August 7, 1999 and was degraded (blurred) as a result of the meteorological conditions, which were measured and recorded (as data) at the same time. Figure 3b shows the restored image using the Kalman filter, which was based on the atmospheric MTF, which is obtained by using aerosol MTF with turbulence MTF. The restored image (Figure 3b) is more detailed. For example, roads and buildings were deblurred and became more accurate than in the original image (Figure 3a).

See also

Environmental Measurements: Optical Transmission and Scatter of the Atmosphere. **Fourier Optics**. **Holography, Techniques:** Computer-Generated Holograms. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Arbel D, Greenberg S, Hadar O and Kopeika NS (2003) Criteria for satellite image restoration success. *Optical Engineering* 42(9): 2607–2613.

Arbel D, Cohen E, Citroen M, Blumberg DG and Kopeika NS (2004) Landsat TM satellite image restoration using Kalman filters. *Photogrammetric Engineering and Remote Sensing (PE&RS)* 70(1): 91–100.

Daniels A, Boreman GD and Ducharme AD (1995) Random transparency targets for modulation transfer function measurement in the visible and infrared. *Optical Engineering* 34(3): 860–868.

Gabor D (1948) A new microscope principle. *Nature* 161: 777.

Gabor D (1949) Microscopy by reconstructed wavefronts. *Proceeding of the Royal Society A* 197: 454.

Gabor D (1951) Microscopy by reconstructed wavefronts: II. *Proceeding of the Royal Society B* 64: 449.

Hadar O, Dror I and Kopeika NS (1994) Image restoration limits resulting from mechanical vibrations, part IV: real-time numerical calculation of optical transfer functions and experimental verification. *Optical Engineering* 33: 566–578.

Hadar O, Robbins M, Novogrozky Y and Kaplan D (1996) Image motion restoration from a sequence of images. *Optical Engineering* 35: 2898–2904.

Kopeika NS (1998) *A System Engineering Approach to Imaging*. Bellingham, WA: SPIE Press.

O'Donnell E (1987) Image steadiness: a critical factor in system performance. In: Wright R (ed.) *Electron Optical Imaging System Integration. Proceedings SPIE* 762: 51–56.

Stern A and Kopeika NS (1995) Analytical method to calculate optical transfer functions for image motion and vibrations using moments. *Journal of the Optical Society of America A* 14: 388–396.

Yitzhaky Y and Kopeika NS (1997) Identification of blur parameters from motion blurred images. *CVGIP: Graphical Models and Image Processing* 59(5): 321–332.

OPTICAL TWEEZERS

A Gajraj and J C Meiners, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The concept of radiation pressure or pressure from the propagation of light is not new. In the early seventeenth century, Johannes Kepler proposed that radiation pressure from the sun accounted for the fact that tails of comets always point away from the sun. In 1873, James Maxwell, on the basis of electromagnetic theory, showed that pressure should result from the absorption or reflection of a beam of light. However, early studies in the beginning of the twentieth century that attempted to detect and quantify the effects of radiation pressure were unsuccessful, mainly frustrated by thermal forces from temperature gradients that tended to obscure the weak effects of radiation pressure. It was only in the 1960s, that the advent of lasers provided a light source with the power to generate sufficient radiation pressure to overcome forces due to thermal effects.

As lasers became generally available in the early 1970s, Arthur Ashkin from the Bell Labs pioneered the development of optical tweezers. He showed that the forces of radiation pressure from focused laser beams could be used to affect the dynamics of small transparent micrometer-sized particles. According to Ashkin, his early experiments were motivated by simple back of the envelope calculations of the magnitude of the radiation pressure force of light on a totally reflecting mirror. Let us consider such a calculation: a photon of frequency ω has a momentum of $\hbar\omega/c$. For a total power of I, we have a flux of $I/\hbar\omega$ photons per second incident on the mirror. When the photons are reflected, they transfer a momentum of $2I/c$ to the mirror. The force exerted on the mirror for $I = 1$ W, is approximately 10^{-9} N. If the mirror is a 1 μm size particle weighing 10^{-15} kg, this force will induce an acceleration of 10^6 g!

In the following decade, Ashkin demonstrated that by focusing laser light using a high numerical aperture microscope objective, strong intensity gradients could be created to optically trap small dielectric particles such as microspheres. Ashkin named this trapping scheme a 'single-beam gradient force trap', now more commonly called optical tweezers.

Today optical tweezers are well-established and extremely versatile tools for the physical and life sciences. Using optical tweezers, forces up to 200 pN can be applied with a resolution of 100 aN on particles with dimensions on the order of the wavelength of light. Even particles as small as 5 nm can be trapped, rotated, positioned, and arrayed. Optical tweezers have been used in single molecule studies of biopolymers (DNA, RNA), in the manipulation of various types of cells and in cell sorting, in microsurgical procedures, as well as in viscoelastic studies of membranes, and protein. Optical tweezing has also been the key tool in the understanding of the workings of molecular motors such as myosin, kinesin, RNA polymerases, etc.

Theory

Optical tweezers trap dielectric particles using radiation pressure generated by a tightly focused beam of light from a laser. Depending on the size of the particle in relation to the wavelength of the trapping beam, there are two different mechanisms responsible for trapping small particles in a single beam gradient trap. For particles in the Mie regime, where the particle size is large compared to the laser wavelength, a simple geometric-optics model serves best. Figures 1a–c, illustrate possible cases of a Mie particle in an optical trap. Here, the particle is assumed to be a sphere that is transparent with a refractive index higher than that of the surrounding medium. The particle behaves as a convergent lens refracting and deflecting the beam when its center is displaced from the laser focus. In Figure 1a, the sphere is axially displaced from the focus, away from the microscope objective. Ray A is then refracted toward the surface normal as it enters and exits the sphere, causing a change in the momentum of the ray. Due to conservation of momentum, the sphere reacts with a force F_a in the opposite direction. The situation is similar for ray B. The gradient force, F, obtained from the vectorial addition of F_a and F_b, is directed toward the light source. At the same time, the scattering force, arising from back and scattered reflections, is directed away from the light source. It can be seen that above the focal point, the scattering and gradient forces act in the same direction, pushing the particle away from the light source (Figure 1b). For a particle in the focal plane, but transversely displaced from the trap center, the gradient force acts in a restoring direction pushing the bead into the trap focus (Figure 1c). At a location just below the focus, the scattering force balances the gradient force and stable optical trapping is achieved.

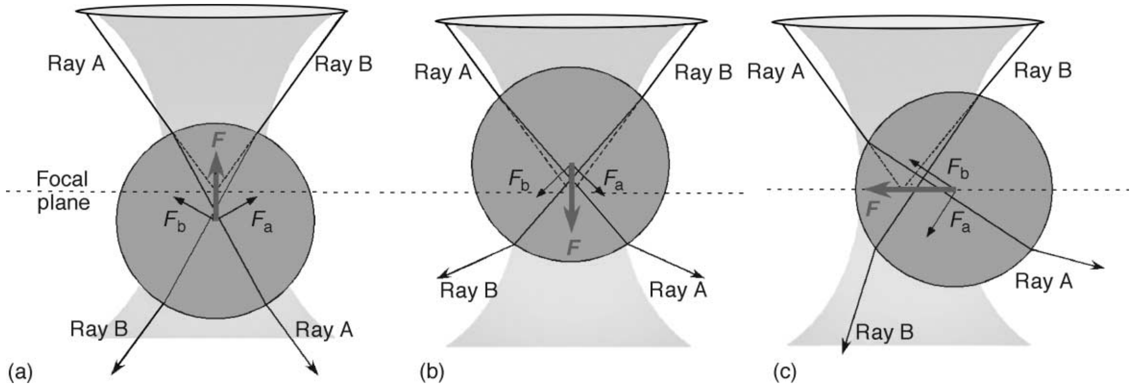

Figure 1 Ray diagram sketches showing three cases of particle trapping in the Mie regime. Rays A and B border the cone of light focused by a high numerical aperture microscope objective. The refraction of these rays gives rise to forces F_a and F_b, the resultant force F tends to restore axial and transverse displacements of the particle from the trap focus (at the intersection of dotted rays A and B).

This simple ray optics model can also be used to estimate the trapping force, an example of such a calculation is shown in Box 1.

For dielectric particles in the Rayleigh regime, where the size of the particle is small compared to the wavelength of the incident light, the particle acts as a simple dipole. The electromagnetic light wave polarizes the particle, inducing a set of oscillating dipoles. The dipoles respond to the electric field of the incident light in the direction of the gradient, resulting in the movement of the particle towards the focal point. Additionally, there is also a scattering force, acting in the direction of beam propagation, from light that is scattered or absorbed. While the scattering force is proportional to the electromagnetic field, the gradient force is proportional to the gradient of the electromagnetic field. Therefore, stable optical trapping can be achieved for a field with a gradient large enough to overcome the scattering force. This can be obtained by using a microscope objective with a high numerical aperture.

Stable trapping is achieved when the energy associated with the trapping significantly exceeds the thermal energy of the trapped particle, $\frac{1}{2}k_B T$ per degree of freedom. Since the optical gradient is weakest in the axial direction of the trap, the trap stiffness in this direction limits its stability. Furthermore, the trap is weakened in the axial direction by the scattering force, which pushes the microsphere away from the objective. To overcome this, two counterpropagating laser beams can be used, not only to increase the axial gradient of the trap, but also to provide a compensating scattering force in the opposite direction.

Force Measurement

Optical tweezers can be used as picotensiometers or highly sensitive force sensors. Measuring the displacement of a particle in a trap with known stiffness k, yields the force F given by

$$\mathbf{F} = -k \cdot x \qquad [1]$$

The displacement can be measured in different ways, ranging from video microscopy to light-scattering techniques. Some of these methods are discussed in the instrument design section below.

The trap stiffness is generally not known *a priori* as it depends on the shape and intensity of the trap, as well as the relative refractive index, shape, and position of the object in the trap. Since theoretical approximations are usually not accurate enough, the trap stiffness must be empirically determined and calibrated.

Various force calibration strategies have evolved over the years, however, they can be broadly divided into two types: (i) schemes that rely on the interaction of the optical trap with viscous forces; and (ii) those that exploit the thermal fluctuations of the trapped object.

Force Calibration by Viscous Drag

The conceptual basis for calibration by viscous drag is simple. A known force is exerted on a trapped object and the deviation in its position is measured as a function of the applied force. The trap stiffness can then be calculated from eqn [1]. Force can be exerted by either flowing liquid past the trapped object, translating the microscope stage while keeping the trap fixed, or translating the optical trap while keeping the sample fixed.

In a situation where the Reynolds number is small, inertial forces can be neglected and Stokes' Law gives the viscous drag force on a sphere of known radius r:

$$F_d = (6\pi\eta r)\nu = \beta\nu \qquad [2]$$

Box 1

Optical Tweezers in the Ray Optics Model

In a simple ray optics model, the trapped microsphere acts like a small lens on the incident laser beam. If the microsphere is not centered on the beam axis, the beam will be refracted. The corresponding change in the momentum of the laser beam gives rise to a restoring force on the microsphere.

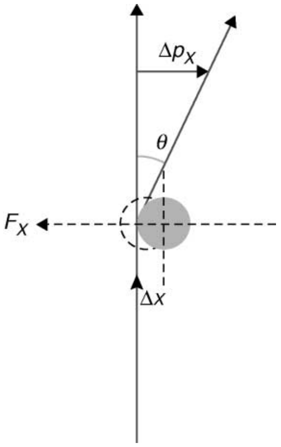

Figure B1

A photon of frequency ω in a medium of refractive index n_1 that is deflected by an angle θ, as shown in Figure B1, acquires a lateral momentum of

$$\Delta p_x = \frac{\hbar \omega n_1}{c} \sin\theta \qquad [B1]$$

Momentum conservation requires a counteracting force on the microsphere, which is for a laser beam of intensity I

$$F_x = \frac{I n_1}{c} \sin\theta \qquad [B2]$$

Assuming that the deflection angle θ is given by the deflection of a beam that is incident on biconvex lens with curvature R and refractive index n_2 at an off-axis displacement Δx, we find for the restoring force on the microsphere:

$$F_x = \frac{2I}{Rc}(n_2 - n_1)\Delta x \qquad [B3]$$

which describes a linear spring with a stiffness of the optical trap of

$$k = \frac{2I}{Rc}(n_2 - n_1) \qquad [B4]$$

Here, η, the viscosity of the sample fluid, is known and immediately allows the determination of the friction coefficient β of the sphere. The fluid velocity, v, can be measured in the case of external fluid flow by either tracking the trapped object when the trap is switched off or measuring the speed of some other marker in the sample. If the stage is moved to generate the fluid flow, the velocity of the fluid surrounding the trapped object can be estimated from the motion of the stage. In either case, the drag coefficient, from eqn [2] may need to be corrected if the trapped object is close to the walls of the sample cell. This correction is well known as Faxen's Law, which accounts for hydrodynamic interactions between the trapped particle and the wall. These effects are particularly significant when the center of the trapped particle is less than a diameter away from the wall. The advantage of these fluid-flow calibration methods is that the trap stiffness at regions far from the trap center can be measured. The drawback is that, in general, only transverse trapping forces can be measured and the trapped objects have to be approximated as spherical or ellipsoidal.

Thermal Motion Methods

These methods measure the Brownian motion of a trapped particle to deduce the trap stiffness. A trapped particle experiences three forces: (i) the thermal forces due to collisions with surrounding solvent molecules; (ii) forces from viscous drag; and (iii) the restoring force of the trap. For low Reynolds numbers, where the inertial effects can be neglected for the microsphere, the equation of motion is

$$\beta\dot{x} + kx = F(t) \qquad [3]$$

Where β is the viscous drag, k is the trap stiffness, and $F(t)$ is a randomly fluctuating Langevin force due to thermal fluctuations. This motion of the microsphere in the trap at a temperature T is characterized by a Lorentzian power spectrum of the form:

$$x_f^2 = \frac{k_\mathrm{b}T}{2\pi^3\beta(f_0^2 + f^2)} \qquad [4]$$

where $f_0 = k/(2\pi\beta)$ is the corner frequency, above which the motion of the particle is Brownian.

To calibrate the optical trap by the corner frequency method, the motion of a trapped particle is observed and its power spectrum, which is the Fourier transform of the autocorrelation function, is calculated. Fitting the power spectrum to a Lorentzian function will yield the corner frequency f_0, which can be directly related to the trap stiffness k, provided that the drag coefficient β can be estimated.

Alternatively, the trap stiffness can be obtained directly from the time constant of the exponential decay of the autocorrelation function of the particle motion. The spring constant is the ratio of the drag coefficient to the time constant of the trap.

Another approach known as the mean square displacement method does not require the estimation of the drag coefficient but does require a well-calibrated position detector. In this method, the mean square displacement $\langle x^2\rangle$ of the fluctuations of a trapped particle is computed from position data and then by the Equipartition theorem:

$$\langle x^2\rangle = k_\mathrm{b}T/k \qquad [5]$$

This enables a simple, straightforward calculation of k, the trap stiffness.

Design

Optical tweezing stations broadly fall into three categories: (i) stand-alone systems; (ii) systems incorporating an optical microscope; and (iii) completely configured commercially available systems. Stand-alone systems feature optical tweezers built around a microscope objective and suitable optical components. These systems provide the maximum flexibility and can be easily reconfigured to adapt to different experimental needs. Furthermore, they also tend to provide the best mechanical stability, if configured properly. Figure 2a shows a picture of a stand-alone optical tweezing setup from our laboratory. This system is capable of the basic tasks of trapping, holding, and positioning microscopic objects. The most common optical tweezing stations are integrated into commercial light microscopes as sketched in Figure 3. The advantage of these hybrid systems is that the commercial microscope provides a flexible, optimized optical imaging platform, in which the optical trapping system can be integrated with only moderate difficulty. Figure 2b shows a picture of a typical optical tweezing station built around a commercial inverted microscope and mounted on a vibration–isolation table. This system includes an acousto-optic modulator that enables beam scanning for line-tweezing experiments. At the time of this writing, optical tweezing systems range from less than $10 000 for basic demonstration stations to between $50 000 to $100 000 for more sophisticated hybrid systems.

Commercial fully integrated systems are available from P.A.L.M. Microlaser Technologies AG (Bernried, Germany), Cell Robotics, Inc (Albuquerque, NM, USA), and Arryx (Chicago, IL, USA). Figure 2c shows a commercial optical tweezers set-up available from P.A.L.M. Microlaser Technologies.

When designing an optical tweezing station, the first item to consider is the choice of laser system. A simple single beam tweezing system can be built around a laser with less than 100 mW of power. Even though the minimum power required for trapping is a few milliwatts in the laser focus, there is considerable power loss in the tweezing and microscope optics. Also, the trap stiffness and thus the maximum trapping force decreases linearly with intensity, therefore most optical tweezers typically use lasers with an output power of 1 W. In some special applications, when beam scanning or multiple traps are required, the power requirement can be significantly larger. The wavelength of the laser also affects the capability of the trap in that the optimum wavelength depends on the size of the objects that are to be trapped. In general, the strongest traps occur

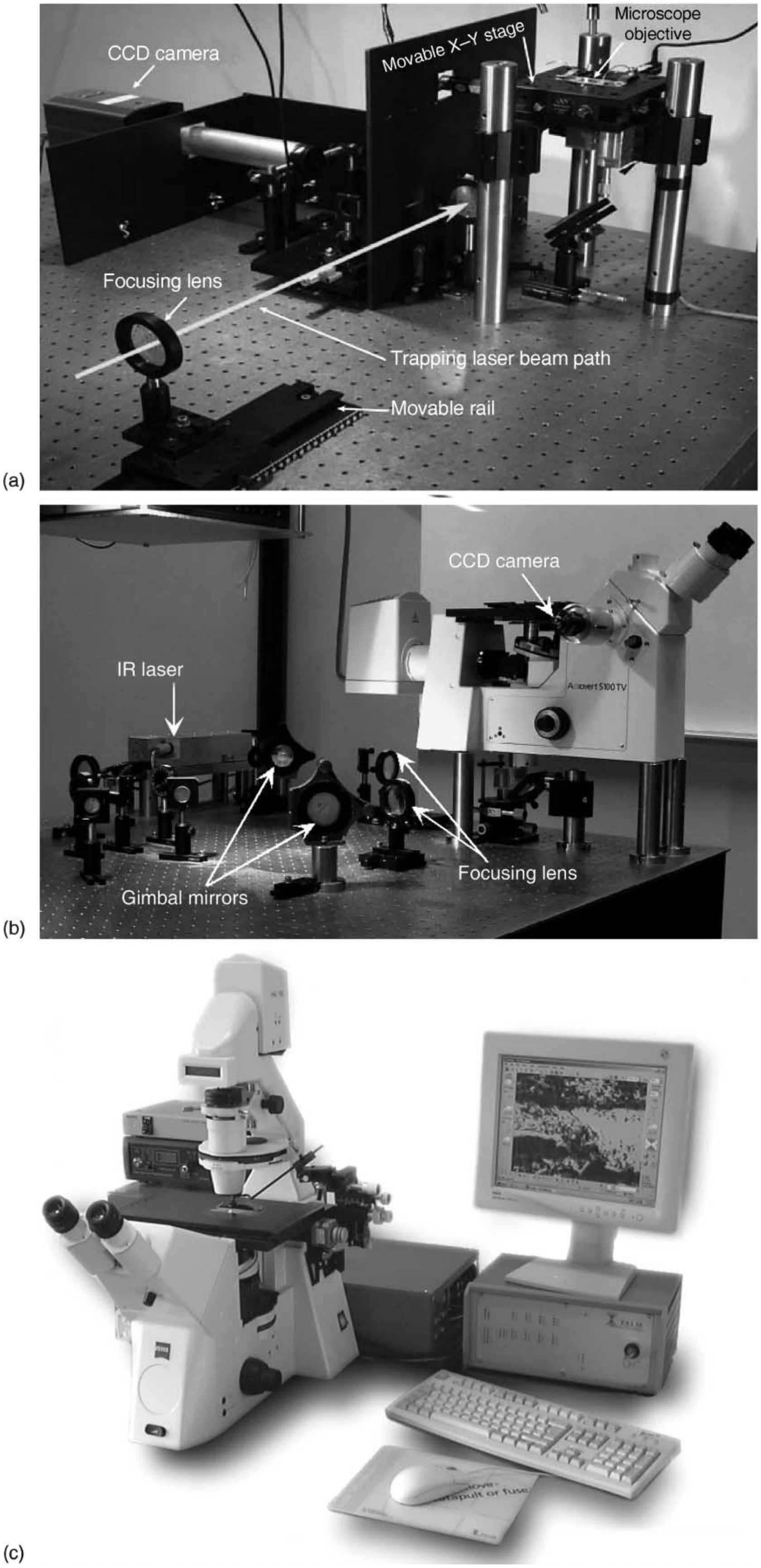
CCD camera
Movable X–Y stage
Microscope objective
Focusing lens
Trapping laser beam path
Movable rail
(a)
CCD camera
IR laser
Axiovert S100 TV
Gimbal mirrors
Focusing lens
(b)
(c)

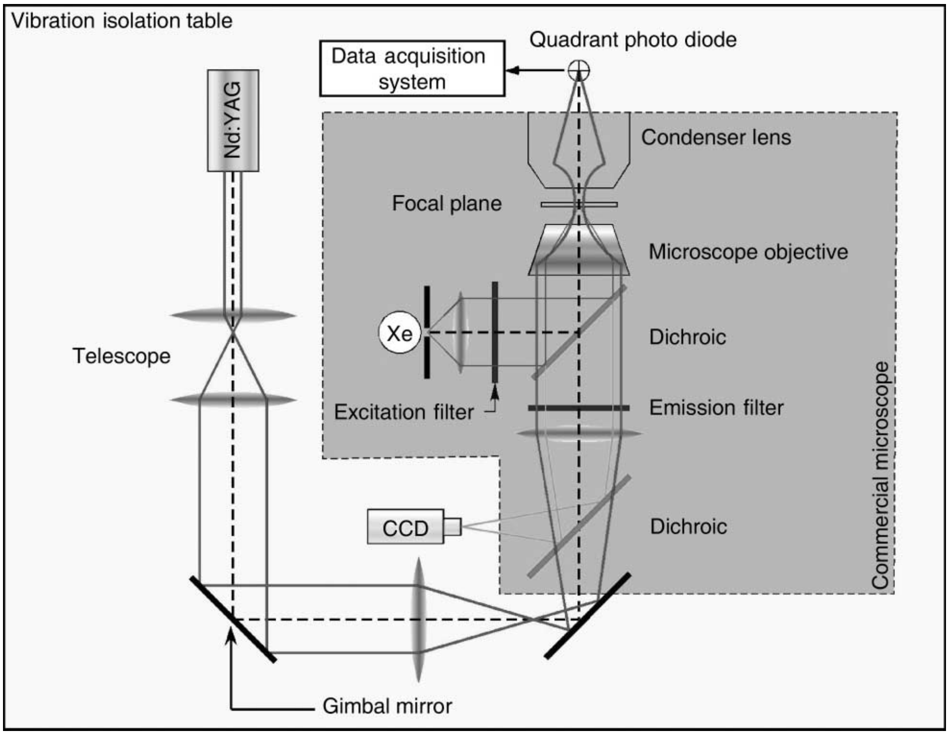

Figure 3 A sketch showing the instrumentation of a typical optical tweezing system incorporated into a commercial microscope.

for particles of size comparable to the wavelength of the laser. Another important concern in the choice of wavelength is possible damage to the sample from absorption and heating. In general, biological material is more transparent to the longer wavelengths of the IR spectrum. Studies on living cells have shown that generally, a range of wavelengths of 700 nm to 1100 nm, termed 'the optical window' of cells and tissue, is most suitable. Importantly, at higher wavelengths there is a reduced possibility of damage from multiphoton effects. DNA (260 nm) and proteins (280 nm) are strong absorbers in the ultraviolet range. Due to multiphoton effects, wavelengths between 700–800 nm have been seen to affect cell metabolism and lead to cell death within seconds. However, at wavelengths greater than 1000 nm, the simultaneous absorption of four photons required is unlikely at typical laser intensities used in trapping experiments.

The beam quality of the laser also restricts the choice of laser system. Diode lasers, in particular, have poor beam quality since the beam is often elliptical in shape and astigmatic, leading to a larger inefficient trap. Over the years, Nd-YAG lasers, at 1064 nm, have emerged as the lasers of choice, providing the best balance between price, power, and quality.

To achieve stable trapping it is essential to create an optical potential in which the gradient forces dominate the scattering force as discussed in the Theory section above. This suggests that the higher the numerical aperture (NA) of the microscope objective the stronger and more stable the trap will be. Oil or water-immersion microscope objectives with NA > 1.2 are commonly used. Further, this requirement also implies that the back aperture (rear entrance) of the microscope objective needs to be completely filled or, in the case of a Gaussian beam, overfilled to a certain degree to achieve optimal trapping efficiency. Since most commercial objectives are corrected for achromatic and spherical aberrations in the visual region of the spectrum, only

Figure 2 (a) A stand-alone tweezing station built around a microscope objective mounted on a vibration isolation table. (b) A typical tweezing station incorporated into a standard commercial microscope. This setup features dual-steerable optical traps and includes a CCD camera. Position detecting instrumentation can be mounted on the existing microscope stage. (c) The commercially available PALM$^{®}$ Combisystem utilizes an IR laser (wavelength 1064 nm) and a UV laser (wavelength 377 nm) coupled via the epi-fluorescence path to a research microscope. The system advertises automated and noncontact laser microdissection as well as trapping, sorting, and positioning. (**Figure 2:** (c), courtesy of P.A.L.M. Microlaser Technologies, Bernried, Germany.)

a small fraction of the working distance is usable for viable trapping. Normally this is limited to within 40 μm of the cover slip.

In addition to providing a beam large enough to overfill the back aperture, there are two other criteria that the external optics have to satisfy: first, the optical trap should be parfocal with the specimen plane. This enables the visualization of trapped objects. Second, the laser beam should be able to pivot around the back aperture without moving laterally to maintain the same degree of overfilling. This allows for a stable two-dimensionally movable trap. In practice, these requirements are fulfilled by using a telescopic system of lenses and Gimbal mounted mirrors. The latter are special mounts that enable a mirror to be tilted around its center position without lateral displacement. The telescope expands the incoming beam to fill the back aperture of the microscope objective and also images the back aperture onto the center of the Gimbal mounted mirror. As seen in Figure 4, since the back aperture of the microscope objective and the Gimbal mounted mirror are in conjugate planes, a small tilting of the mirror will only result in a change in the angle of the laser beam entering the microscope objective. The position of the beam will be unchanged, as will the degree of overfilling.

More sophisticated optical tweezer designs use acousto-optic modulators and galvanometer scanning mirrors to create scanning line tweezers and/or multiple traps. The basic approach is to deflect the laser beam to create multiple traps or to scan a single trap to create a one-dimensional potential well. Additionally, a new generation of holographic optical tweezers, capable of creating three-dimensional arrays of trapped particles, has emerged. Computer generated holograms that modify the phase and amplitude of the trapping laser are reverse engineered from the desired two- or three-dimensional optical trapping patterns. Furthermore, constructing such holograms from liquid crystal light modulators allows for dynamic control of the trapping pattern, yielding a fully computerized reconfigurable trapping system. An interesting example of the application of holographic optical tweezers to colloidal transport is shown in Figure 5, the first demonstration of 'kinetically locked-in states' where particles traveling across a potential mesh display a preference for certain paths.

It is to be noted that the optical tweezer is fundamentally a constant-extension technique, that is, it attempts to maintain the extension of a stretched molecule whereas the force is varied. However, it is of interest to study the action of molecular motors under constant force conditions. A basic optical tweezers setup can be modified to include an electronic feedback loop to provide constant force conditions. More recently, scanning-line optical tweezers have been adapted to implement all-optical constant force schemes that dispense with the need for electronic feedback.

An important aspect of any force measurement is the ability to accurately image and measure the position of a trapped particle. Standard digital video microscopy techniques yield a resolution of ~5 nm but have limited bandwidth and thus cannot be used for systems that employ feedback control of laser intensity or trap position. In these systems, the forward- or back-scattered trapping laser light from the trapped particle can be imaged onto a calibrated quadrant photodiode to provide sensitive position detection at high bandwidths (~1 kHz). If the photodiode is fixed in a plane conjugate to the back focal plane of the microscope objective, the trap can

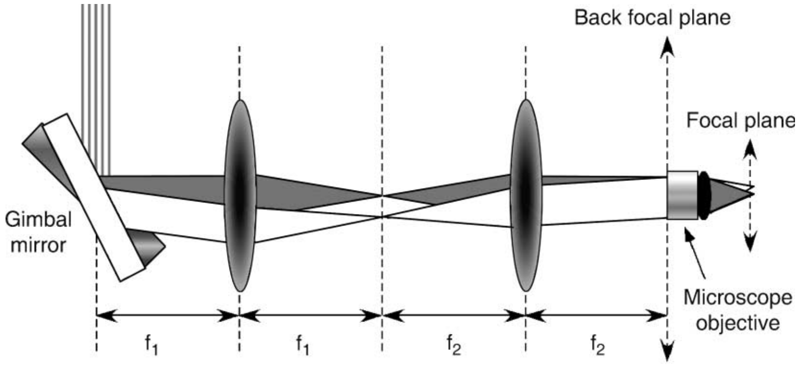

Figure 4 A sketch showing the working of a gimbal mounted mirror in two positions (shown as shaded and as unshaded). For small tilts of the mirror, the beam is seen to 'pivot' about the back aperture of the microscope objective resulting in the displacement of the optical trap in the focal plane. Note that for both beams the degree of overfilling of the back aperture of the microscope objective is unchanged, a crucial requirement for movable traps.

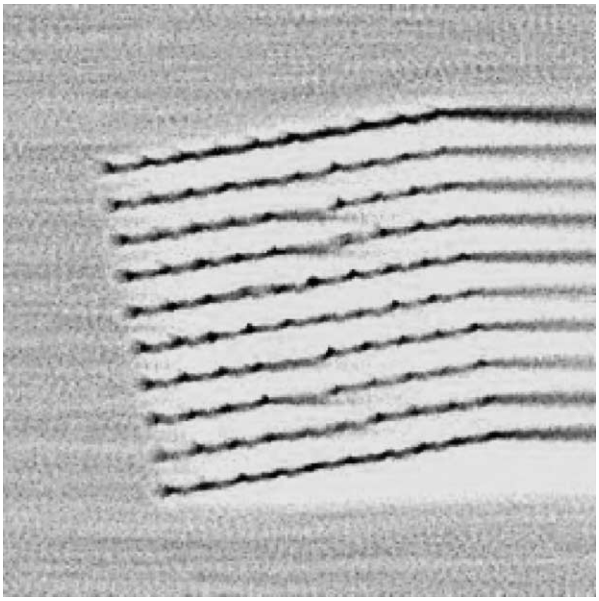

Figure 5 Data showing the superimposed paths of approximately 18 000 micron-sized particles flowing across a 10 × 10 grid of optical traps created by holographic optical tweezers. The particles are deflected laterally assuming a path dictated by the potential energy landscape and display kinetically locked-in states. This technique can be used to sort particles based on size, refractive index or charge. (Picture courtesy of D. Grier.)

be moved without having to move or recalibrate the photodiode detector. Using the quadrant photodiode method 2- and 3-dimensional position sensing is possible. A further improvement is found in schemes that use an additional laser dedicated to position sensing. The most sensitive method, however, for position sensing is interferometry. In this scheme, the trapping laser beam is split into two orthogonally polarized beams that are recombined after passing through the microscope objective and the sample. For a particle perfectly centered in the trap, the resulting combined beam has a circular polarization. In other cases, the bead displacement is sensitively reflected in the ellipticity of the polarization.

Applications

Colloids

Optical tweezers have been widely used to probe the interaction in colloids and other complex fluids. In fact, an optical tweezer apparatus was included in the instrumentation on the International Space Station for use with colloidal crystallization experiments in space. Various aspects of colloidal interactions have been studied using optical tweezers including mutual- and self-diffusion coefficients, the effects of hydrodynamic and electrostatic interactions, and depletion layer effects. More sophisticated scanning-line tweezing setups have been used to probe entropically driven self-assembly and interactions of colloids in polymer suspensions.

Blinking optical tweezers provide an elegant method for positioning colloidal particles in configurations of interest in a solution. When the tweezers are switched off the colloids respond to their environment, reacting to thermal forces and interacting either with other colloids or interfaces. The trajectories of the released colloids can be tracked and analyzed in 2D using digital video microscopy or 3D if combined with an evanescent illumination scheme. Switching on the tweezers again results in a return to the initial configuration, yielding a reproducible and tightly controlled experiment to examine colloidal behavior.

Biopolymers

The study of biopolymers has benefited from the development of optical tweezers. By chemically attaching a trappable latex microsphere to a biopolymer of interest, the molecule can be stretched, positioned, or moved. Measuring the response of the biopolymer to these external forces yields useful information such as elasticity, rigidity, and domain unfolding characteristics. This information obtained at a single molecule level has provided invaluable insight into the relationships between the structure and function of biopolymers. For example, in an early tweezing experiment, characterizing the intrinsic elasticity of λ-phage DNA, it was discovered that the double-stranded DNA undergoes a reversible structural transition when stretched at forces greater than 65 pN, extending to about 1.7 times its crystallographic contour length. These studies complement well-established structural techniques, such as X-ray crystallography and NMR, which provide information in a force-free environment. Furthermore, by anchoring one strand of DNA to a surface and pulling on the other strand, unzipping forces for DNA have been measured. Optical tweezers have also been combined with other mechanical tools, such as a micropipette, and used as sensitive tensiometers. Protein folding and unfolding experiments are generally harder, as their contour length is shorter than that of typical DNA fragments. Nevertheless, the folding and unfolding behavior of single giant muscle protein, titin has been probed by attaching a different latex bead to either end of the molecule. The first bead is held by a movable micropipette and the second bead is optically trapped. As the micropipette is moved, the position of the optically trapped bead provides information about the forces required to unfold the membrane domains of the protein.

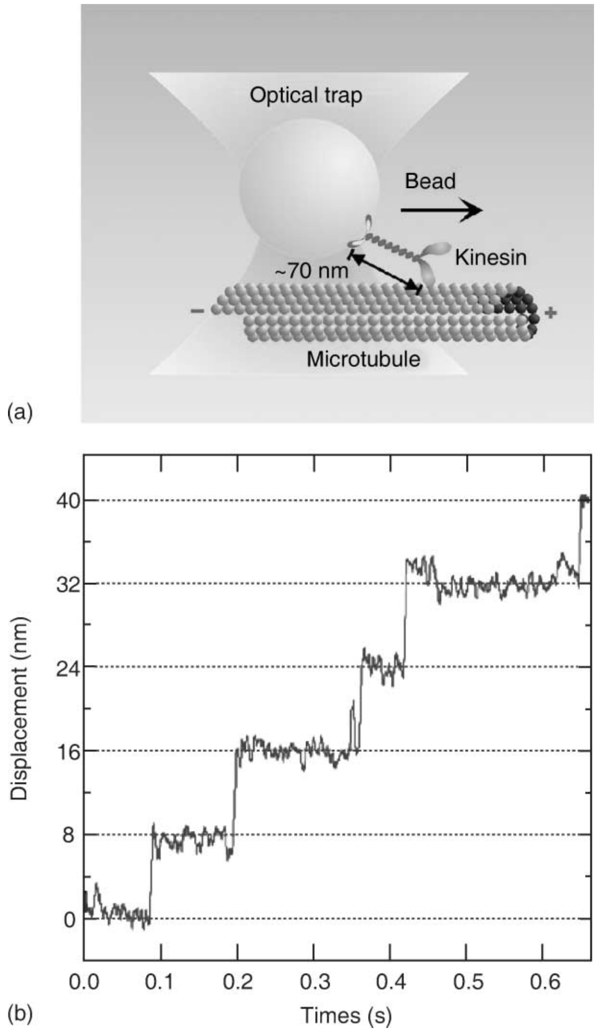

Figure 6 Observation of the stepping motion of the molecular motor kinesin. (a) A microsphere attached to a kinesin molecule is optically trapped; as the kinesin steps along a microtubule track which is immobilized onto a glass substrate, it pulls on the microsphere, whose motion is followed by an electronic feedback mechanism. (b) Motion of a single kinesin motor stepping against a 5 pN force. The data shows the discrete 8 nm steps of kinesin. (Figures courtesy of S. Block.)

Molecular Motors

Molecular motors (proteins that transduce chemical potential energy in to mechanical motion) have been the subject of intense attention since the early 1990s. Optical tweezers have played a crucial role in furthering the understanding of the mechanisms of molecular motors. The characteristics of molecular motor systems that have been studied with optical tweezers include speed, step-size, processivity, ATPs consumed per stroke, stall force, and static force. In fact, many of the advances in optical tweezers technology have been motivated by the study of molecular motors and have lead to spectacular results. An impressive example is the observance of the stepping motion of kinesin, as shown in Figure 6. Other experiments on molecular motors with optical tweezers have measured the stall force of motors as they move along their polymer track, or observed the transcription of DNA and RNA polymers, yielding interesting observations of the pauses during the transcription process.

Cells

The application of optical tweezers to cell biology was first demonstrated by Ashkin and co-workers in the late 1980s, by successfully trapping living objects: large viruses and bacteria. Since then, optical tweezers have been widely used to study various aspects of cellular biology. These include cell reproduction, growth, adhesion, mobility, and membrane elasticity. Additionally, the noninvasive nature of optical tweezers enables manipulations in the interior of an unopened object, allowing for the study of intracellular mechanisms by subcellular manipulations.

Optical tweezers have also been combined with various techniques to yield application-oriented tools for cell biologists. For instance, the combination of optical tweezers and state-of-the-art image recognition algorithms has been used to develop microfluidic cell sorters. In reproductive medicine, two of the major problems are penetration of the egg and sperm motility. A promising new solution combines laser zona drilling to pierce the egg envelope and optical tweezing to transport the sperm cells to the egg.

See also

Modulators: Acousto-Optics.

Further Reading

Ashkin A (2000) History of optical trapping and manipulation of small-neutral particle, atoms and molecules. *IEEE Journal of Selected Topics in Quantum Electronics* 6(6): 841–856.

Greulich KO (1999) *Micromanipulation by Light in Biology and Medicine: The Laser Microbeam and Optical Tweezers*. Basel; Boston; Berlin: Birkhauser.

Grier DG (2003) A revolution in optical manipulation. *Nature* 424: 810–816.

Lang MJ and Block SM (2003) Resource letter: LBOT-1: Laser-based optical tweezers. *American Journal of Physics* 71(3).

Sheetz MP (1998) *Methods in Cell Biology. Volume 55: Laser Tweezers in Cell Biology*. San Diego, CA: Academic Press.

Svoboda K and Block SM (1994) Biological applications of optical forces. *Annual Review of Biophysics Biomolecular Structure* 23: 247–285.

P

PHASE CONTROL

Contents

Phase Conjugation and Image Correction
Wavefront Coding

Phase Conjugation and Image Correction

E N Leith, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Phase conjugation of a wave is a process that has long been recognized. However, the technique for producing high-quality phase conjugation is relatively new. Also, the principal use of phase conjugation, the correction of phase errors in waves that have propagated through phase-distorting media, is relatively new, having originated in holography in the 1960s. Today, phase conjugation is an important area of modern optics. This article describes the conjugation of a wavefield by the process of holography and the holographic phase conjugation method of removing phase distortions from a wavefield, thereby permitting high-quality images to be produced from phase-distorted waves.

The basic meaning of a conjugated wave is explained with the aid of Figure 1. Figure 1a shows a wavefield propagating to the right. For simplicity, we show only one wavefront, i.e., one contour of constant phase. It is understood that a wavefield consists of a sequence of such wavefronts, with each wavefront differing in phase by 2π from the adjacent one, and separated by the wavelength λ. Of the two points, A and B, on the wavefront, we say that the wavefront at A lags that at B, since when the wavefront crosses a plane normal to the direction of travel, point A of the wavefront crosses that plane behind point B. The wavefield produces, at the plane P, a field $ae^{i\phi}$, where a is the amplitude and ϕ is the phase of the wavefront at the plane P.

We place a phase conjugator in the path of the wave. The wave passes through the conjugator, emerging as a phase conjugated wave, in which the point B, which formerly led point A in phase, now lags behind it by a corresponding amount. This illustrates the basic meaning of the term ‘phase conjugation’. This device could be considered as a forward phase conjugator, since the phase conjugated wave travels in the same direction as the original wave.

Next (Figure 1b) we describe a backward phase conjugator. The conjugate wave travels in a direction opposite to that of the incident wave. The conjugate wave emitted by the phase conjugator seems to be

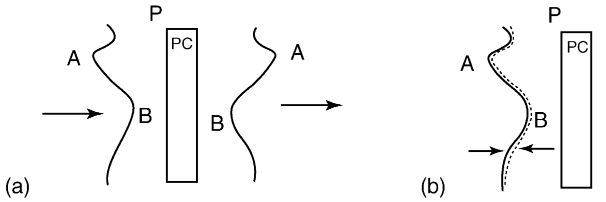

Figure 1 A phase conjugator. (a) shows a forward phase conjugator; (b) shows a backward phase conjugator. For clarity, in (b) the conjugated wave is shown with dotted lines. PC is phase conjugator, and P is the entrance plane of the conjugator.

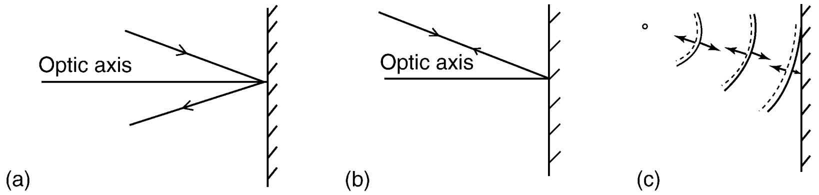

Figure 2 Light reflected from mirrors. (a) Light falling on a conventional mirror and being reflected; (b) light reflected from a conjugating mirror; (c) as in (b), except that illustration is with wavefronts instead of rays.

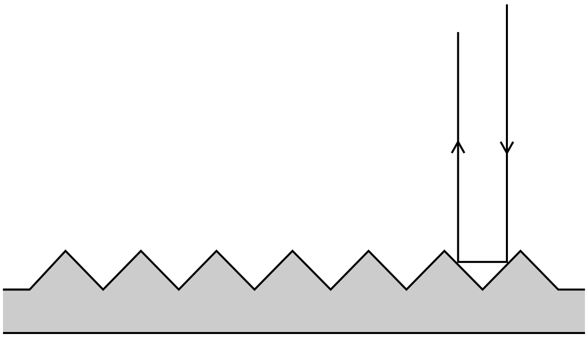

Figure 3 A ray of light reflected from a retroreflector (or corner cube).

identical to the incident wave; we show the two waves to be in coincidence, or more precisely, for purposes of clarity, almost in coincidence. The points A on the two wavefronts coincide, as do the points B. As before, on the incident wave, point A lags behind point B. On the exiting wave, since the wave is traveling in the opposite direction, point B is now lagging point A, i.e., is lagging in phase. In either case, forward or backward conjugation, the wave emerging from the conjugator has the complex amplitude $a(x, y)e^{-i\phi(x,y)}$.

A phase conjugator (backward variety) is like a mirror, in that it obeys the basic mirror law, that the angle of reflection equals the angle of incidence, but in a rather different way, as Figure 2 shows. In fact, the reflected ray travels back on itself. A simple way to generate a conjugate wave, and one that has been known for a very long time, is with the use of an array of retroreflectors, or corner cubes. As shown in Figure 3, each ray undergoes two reflections and emerges traveling in exactly the opposite direction, although with some displacement. If we considered the third dimension, there could be three reflections. The amount of displacement depends on where the ray strikes; if it strikes near the apex of the corner cube, the ray displacement will be slight; if at the edge, the displacement is greatest. Thus, the image formed in reflection is degraded. This geometrical degradation is greater for larger corner cubes. Thus, why not make the cubes very small? Then, however, the image is degraded by diffraction effects, since each corner cube constitutes a small aperture. Thus, there is an optimum size for the cubes; if too large, geometrical factors degrade the image; if too small, diffraction effects degrade the image.

Phase Conjugation by Holography

With holography, however, came a method of conjugating a wavefront that yields high-resolution imagery. The holographic process inherently produces a phase conjugate image, as was noted by Gabor in his early papers on holography. The holographic process is illustrated in Figure 4. For simplicity, we assume the object to be a planar transparency $s(x, y)$. A coherent light wave, generally a plane or spherical wave, passes through the transparency and a diffracted field u is produced at the recording plane P, as in Figure 4a. The field u is typically the Fresnel diffraction pattern of the object s, but if the propagation distance is sufficiently great, u could be the Fraunhofer pattern of s. Of course, lenses can be placed between the object and recording planes, thus modifying the diffraction pattern.

The field u is combined with a reference wave u_o, which is generally an off-axis planar or spherical wave. The two beams produce an interference pattern with intensity

$$I = |u_o + u|^2 = |u_o|^2 + |u|^2 + 2Re\, u_o^* u \qquad [1]$$

where the final term can be written as $u_o^* u + u_o u^*$; the * denotes complex conjugate. The process of forming the interference of the two beams, i.e., forming the intensity of their sum, has generated the conjugate term u^*. When the intensity is recorded, as on photographic film, for example, both terms are stored on the film. Suppose for simplicity that u_o is a plane wave $a_o e^{i2\pi f_o x}$, where a_o is a constant, and $f_o = (\sin\theta)/\lambda$, where θ is the angle that the reference wave makes with the optic axis and λ is the wavelength of the light. Also, u can be

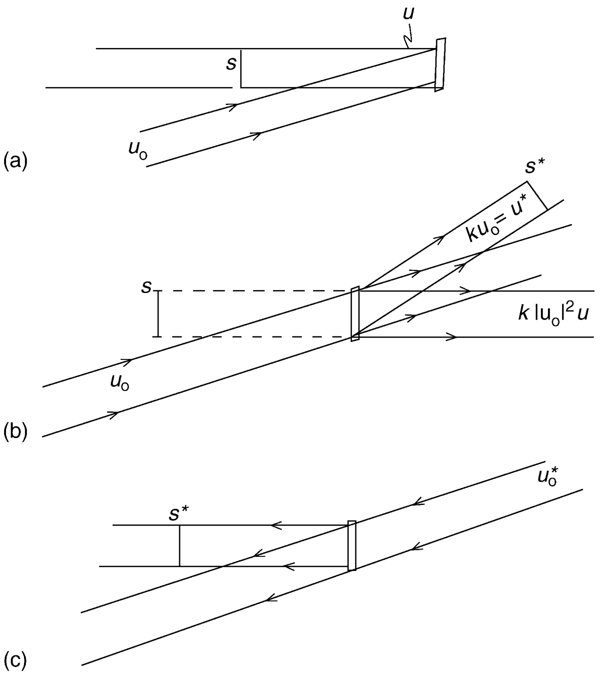

Figure 4 The holographic process. (a) Making the hologram; (b) reading out the hologram, using a readout beam that duplicates reference beam; (c) reading out with reversal of reference beam.

written as a product of an amplitude term a and a phase term $e^{i\phi}$, or $u = ae^{i\phi}$. Let the photographic plate, after exposure and development, have a transmittance t proportional to the intensity I, or $t = kI$, with k a constant; this is an oversimplification of the recording process, but preserves the essence of the holographic process. Now let the photographic plate, which after development becomes a hologram, be illuminated with a readout wave that, for simplicity, we let be u_o, a duplicate of the reference wave. Then, the field emerging from the hologram is

$$u_e = u_o k\left[|u_o|^2 + |u|^2 + u_o^* u + u_o u^*\right] \qquad [2]$$

Only the last two terms are of interest. These can be written as $k|u_o|^2u$ and $ku_o^2u^*$, respectively, or alternatively, as $ka_o^2ae^{i\phi}$ and $ka_o^2ae^{i(4\pi f_o x-\phi)}$. The term $k|u_o|^2u$ represents a wave that is, to within a constant, identical to the object wave u; the other term represents a wave that is, to within a constant and a linear phase term, a conjugate of the original wave. The factor $e^{i4\pi f_o x}$ has the physical meaning that the wave is propagating in a direction 2θ given by $\sin 2\theta/\lambda = 2f_o$. The readout waves are shown in Figure 4b.

The so-called true wave u and the conjugate wave u^* thus propagate angularly away from each other and from the direct transmitted beam, which carries the uninteresting components $|u_o|^2$ and $|u|^2$.

However, a more suitable way to form the conjugate wave is to illuminate the hologram with a readout wave that is, again, a duplicate of the original reference wave, but propagating in the opposite direction, as in Figure 4c. The conjugate wave is now propagating along the z axis, along the path of the original object wave. It is readily shown that this counterpropagating wave is identical in shape to the original object wave recorded by the hologram, except that it travels in the opposite direction, exactly the situation portrayed in Figure 2. This wave propagates so as to form an image that is, to within a constant, identical to the original object, except for the counterpropagation, which produces a conjugate image $s^*(x, y)$. The image intensity, however, is $|s|^2$, which is identical to the original image intensity.

This conjugate wave was, in the early days of holography, a detriment, since without the use of the off-axis method (reference wave introduced at an angle to the object wave) the two images were inseparable, and the true image could be observed only against the background of the defocused conjugate image. Thus, an out-of-focus image overlaid the true image, making it noisy.

In 1962, the same year that the off-axis reference wave for separating the two images was introduced, a possible use for the conjugate image was suggested. If the original propagation path of the object wave were aberrated, as would occur if the wave propagated through an irregular medium, then the conjugate wave propagating through the medium would undergo a correction and would arrive at the original object position perfectly corrected, resulting in a sharp image, a process sometimes called wavefront healing. The idea lay dormant for 3 years, until in 1965, experimental demonstration of this idea was reported. In one case, the object was a transparency containing opaque lettering against a transparent background and the inhomogeneity was a piece of frosted glass, which resulted in extremely severe aberrations, so that the object wave, having traversed the diffuser, was no longer capable of forming an image. However, when the distorted wave was recorded holographically, conjugated, and sent back through the same diffuser, the result was a highly legible image.

The principle of wavefront healing is illustrated in Figure 5, when the inhomogeneity is an extended, or three-dimensional, structure. A point object O sends a spherical wave into the inhomogeneity. As the wavefront propagates, it becomes distorted, and loses its sphericity. When it emerges from the inhomogeneity, the wavefront is highly distorted.

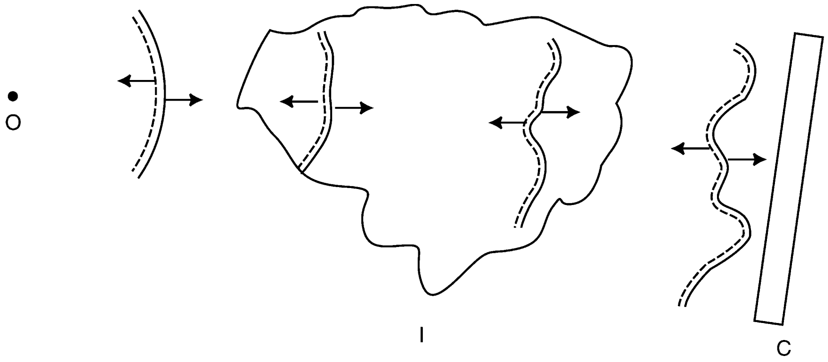

Figure 5 The wavefront healing process. C is the conjugator, I is the inhomogeneity, and O is a point object.

The wave is conjugated at the conjugator C and emerges as a back-propagating wave that retraces the original wave path through the inhomogeneity, gradually losing its aberrations and looking more like a spherical wave. When it emerges from the inhomogeneity, it is a perfect spherical wave that converges to a point image that coincides with the original object. At each plane throughout the inhomogeneity, the back-propagating conjugate wave (dashed lines) has the same shape as did the original wave when it crossed that plane (solid line).

This interesting wavefront healing property of holographic phase conjugation was the beginning of a new branch of optics, called phase conjugation, giving rise to a number of applications. However, the holographic method had the problem that it did not occur instantaneously, but had to await the processing of the photographic film on which the hologram was recorded. This delay limited the range of potential applications. Later, in the 1970s, nonlinear crystals became available that did the equivalent of holography, but in real time, or near real time – a process known as four wave mixing. There are yet other types of real-time phase conjugators with nonlinear devices.

Pseudoscopic Image

As long as the object is only two-dimensional, a function of x and y only, with no extent along the axial dimension, z, the intensity distribution of the true and conjugate images is identical. The conjugateness, being a pure phase phenomenon, disappears when the observable is the image intensity $|s|^2$, and not the complex amplitude s. Since the intensity is always the observable, whether the image is viewed by eye, or with a camera, the conjugateness generally is of no significance.

However, when the object for holography is three-dimensional, the situation is different, and the conjugate image exhibits some strange properties. These arise from conflicting cues pertaining to the three-dimensionality of the image. To explain, we review some basics about how a three-dimensional world is perceived by a viewer. First, there is the stereo effect, resulting from the viewer having two eyes, with each eye seeing an object from slightly different positions, a phenomenon called binocular disparity. The mind uses this stereo effect to perceive three-dimensionality. But there are other cues that also contribute to perception of three-dimensionality. Another is parallax; as the viewer moves his head, objects farther from the observer change their angular position more slowly than objects close by. The observer can sense the three-dimensionality of an object, even with only one eye, by moving the head, thereby inducing parallax. Also, there is interposition, whereby near object portions block the viewing of object portions lying behind them, again giving a sense of depth. There are still other cues, more subjective, such as perspective, parallel lines converging toward infinity, shadowing, etc. With these observations in mind, we consider the hologram under illumination and the two images that it forms (Figure 6). As depicted in Figure 6, these two images lie in mirror symmetrical positions relative to the hologram. Both images have the same side facing the hologram. It must be this way, because when the hologram was recorded, it was side **a** that faced the recording plane. And both images must therefore have side **a** closer to the hologram.

When a viewer observes the true image (a virtual image), he views it looking through the hologram, and he sees the image at the position where the original object was when the hologram was recorded, and it is, in its visual properties, indistinguishable from the actual object; it has all of the three-dimensional and the parallax properties that the original object would exhibit. However, when the viewer observes the conjugate image, he must view it

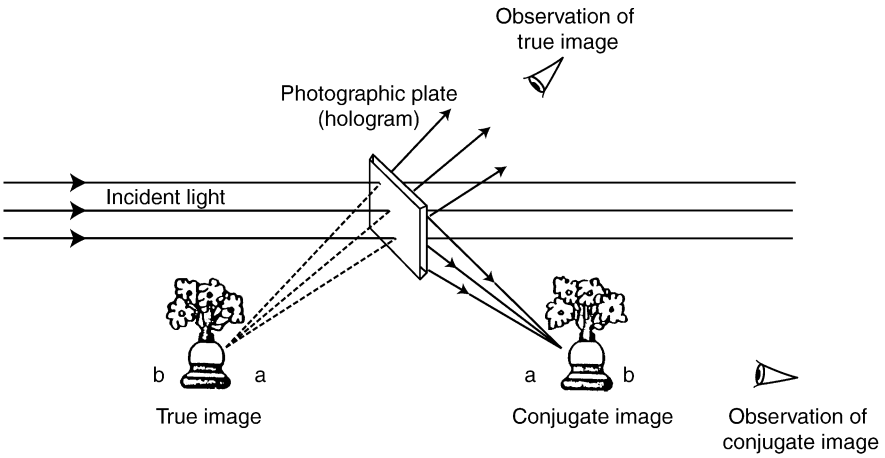
Observation of true image
Photographic plate (hologram)
Incident light
b
a
True image
a
b
Conjugate image
Observation of conjugate image

Figure 6 Formation of the true and conjugate images from a hologram.

from the side of the image opposite from the hologram, i.e., he views the image from the back side. When a viewer observes the conjugate image, he sees it from the back side (side **b**). All of the above cited cues are in agreement about the orientation of the image, viz., that side **b** is closer to the observer than is side **a**. That is, all cues but one. The conjugate image is formed as the hologram recording plate saw it, with side **a** being closer. Thus, all interpositions are formed with the same relation that the object had with respect to the recording plate. The side **a** is the part of the object that was closer to the hologram, so that when interpositions occur, side **a** blocks side **b**. Since the observer thus sees side **a** partially obscuring side **b**, he must conclude that side **a** is closer than is side **b**. However, this observation is in direct conflict with the other cues, especially the stereo and the parallax cues, which are themselves strong cues, as is the interposition cue. The conflict of these cues, with two of them interpreting the image as being formed with one depth orientation, the other with the opposite, gives a number of strange effects. For example, as the viewer moves his head so as to induce parallax, the object appears to rotate. If the image is that of a human head, for example, the observed effect is that as the viewer changes his position, the head appears to rotate in synchronism with the observer, as if the image had its gaze fixed on the viewer, an effect that is often perceived as being eerie.

Similarly, if the object were two flat plates, (Figure 7) in the situation depicted, the planes have no overlap, i.e., no interposition, as seen by the observer. And the observer unequivocally perceives plate A as being closer than plate B. However, if the viewer moves his head to a position where interposition occurs, the conflict of cues arises, A becomes partially blocked by B, which is quite correct as observed at the position of the hologram, but is quite incorrect from the viewer's vantage point. The result of this conflict is that the viewer, who previously perceived the three-dimensionality, now sees the three-dimensionality perception suddenly collapse, and the sense of three-dimensionality is lost. Other strange perception anomalies can also be produced. Such images, with conflicting depth cues, are called pseudoscopic, as opposed to orthoscopic images, where the cues do not conflict.

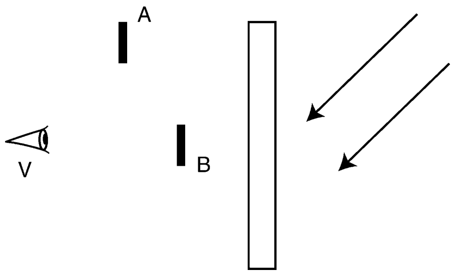

Figure 7 Pseudoscopic conjugate image formed from an object consisting of two plates. The viewer V perceives plate A as being closer than B and there are no conflicting cues, since there is no interposition. However, if the observer moves his head upward, there is partial interposition, thus conflicting cues, and the sense of three-dimensionality is lost.

One Way Phase Conjugation

A significant limitation of phase conjugation aberration correction is that the image and object are on the same side of the inhomogeneity. One would usually want them to be on opposite sides. In a modified form

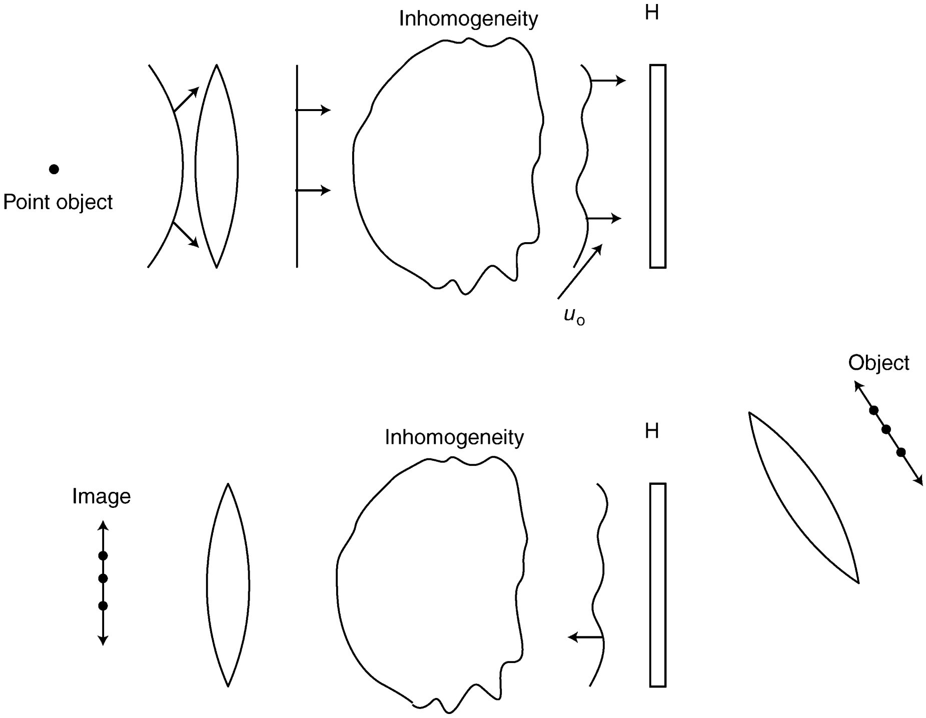

Figure 8 One-way phase conjugation. Top: Making the hologram. Bottom: The viewing process. We show three points on the object. The center point forms a well-corrected image, but the other two points produce imperfectly corrected images.

of the phase conjugation process, shown in Figure 8, this result is achieved. The first step of the process is carried out by passing a plane wave, instead of the object wave, through the inhomogeneity. This plane wave originates from a point source located at the focal plane of a collimating lens. The plane wave thus carries information about the inhomogeneity to the recording plate, where it is combined with a reference wave and a hologram of the distorted plane wave is recorded. Thus, information about the phase defect is stored on the hologram.

Next, the hologram is read out, not with a counterpropagating replica of the reference wave as before, but instead with an object-bearing wave that passes through the hologram. This wave emerges from the hologram and now has on it the recorded aberration information. As it counterpropagates along the path of the original object wave, it passes through the inhomogeneity, whereupon the aberration is cancelled and a corrected image forms at the focal plane of the collimating lens, where previously the point source had originated. The goal is achieved; the object and the corrected image are now on opposite sides of the inhomogeneity.

This correction is only approximate. To explain, we think of the object as being a collection of many points. The radiation from each point is converted into a plane wave by the lens, with different object points producing plane waves propagating at different angles. These plane waves then impinge on the hologram and emerge with each one bearing the aberration stored on the hologram. They continue on their paths and enter the inhomogeneity. The plane wave component, whose path direction matches that of the reference wave used in recording the hologram, will emerge from the inhomogeneity as a perfectly corrected wave, thus forming a well-corrected point image at the position of the original point radiator. However, another plane wave component, traveling in an incorrect direction, will be laterally displaced from the proper position for entering the inhomogeneity, and the aberration correction will be imperfect, the degree of imperfection being in direct relation to this displacement error. The displacement error is the product of the angular error and the propagation distance between hologram and inhomogeneity. If the inhomogeneity is slowly varying as a function of x and y, and if the object does not have a large extent, so that the range of angles is small, the image degradation will be small.

If the last two conditions do not apply, two options are available. First, the distance between inhomogeneity and hologram can be reduced, either by bringing the hologram closer to the inhomogeneity, or by imaging the hologram onto the inhomogeneity. The correction is now valid over a wider extent of the image, especially if the inhomogeneity can be treated as planar rather than extending over a volume. Alternatively, a scanning process can be carried out on the object beam falling on the hologram, so that in sequence, each plane wave component assumes the proper angle to counter-propagate along the path for perfect correction, and will form a well-corrected image point. This output spot then falls on a pinhole positioned to separate it from other, imperfectly corrected points. Then, over time, all points of the object will form well-corrected image points, and the entire object will thus be well-imaged. The one way phase conjugation method was first used to correct the spherical aberration of a lens, and the result was quite successful.

Phase conjugation has become an important branch of modern optics. The process is often carried out using photographic materials. More often nowadays, it is carried out using nonlinear crystals. Some of the crystal methods are essentially holography in real time, but others, such as phase conjugation by Brillouin scattering, go beyond holography.

See also

Nonlinear Optics, Basics: Nonlinear Optical Phase Conjugation. **Phase Control:** Wavefront Coding.

Further Reading

Caulfield HJ (ed.) (1979) *Handbook of Optical Holography*. New York: Academic Press.

Goodman JW (1996) *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill, chap. 9. This chapter is devoted to holography, with section 9.12.3 dealing with phase conjugation aberration correction.

Kogelnik H (1965) Holographic image projection through inhomogeneous media. *Bell Systems Technology Journal* 48: 2451–2452.

Leith EN and Upatnieks J (1965) Holograms: their properties and uses. *SPIE Journal* 4: 3–6. This is the earliest publication demonstrating the principle of wavefront healing by holographic phase conjugation.

Smith HM (1975) *Principles of Holography*, 2nd edn. New York: Wiley.

Wavefront Coding

W T Cathey, University of Colorado, Boulder, CO, USA

E R Dowski, CDM Optics, Inc. Boulder, CO, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

We describe a new paradigm for designing imaging systems that have performance that is otherwise unobtainable. These are hybrid systems that use the optics of the system to acquire an 'intermediate' image, which is processed to produce the final image. The optical components of the system are designed jointly with signal processing to optimize overall performance. It is possible to extend the depth of field of an imaging system, for example, by a factor of eight to ten. On the other hand, by concentrating on the increase of the depth of focus in the detector plane, it is possible to use that increase in the depth of focus to make the system invariant to focus-related aberrations. By doing this, it becomes possible to design a single-lens imaging system with small f-number and large field of view. Another example of the sort of new imaging system that can be designed is a two-lens zoom lens that images with a large field and a very short overall length.

First, the term 'Wavefront Coding' is defined as employed with imaging systems, and a description is given of how such coding is achieved. Examples of means of applying Wavefront Coding are given along with the analysis and design tools that are useful in designing hybrid optical/digital imaging systems. Finally, some potential future applications of Wavefront Coding in imaging systems are suggested.

Background and Basic Concepts

Some early suggested modifications to imaging systems can be viewed as wavefront modification, even though they were not described in that manner at the time. These include the work of Ojeda-Castañeda that described masks to block portions of the lens to extend the depth of field. In addition, it has long been known that the depth of field is slightly greater in lenses that have spherical aberration. In both of these cases, the wavefront is modified, in one case by changing the wave amplitude, and in the other by a deviation from the traditional ideal phase front. However, in both cases, the increase in the depth of field was only a few percent. Häusler extended the depth of field by requiring that the focus be continuously changed during the exposure time of the camera. This resulted in a modulation transfer function (MTF) for the imaging system that was the incoherent superposition of a spread of MTFs for the different foci during the exposure. To form an acceptable image, signal processing was used.

The use of Wavefront Coding here refers to new optical systems arrived at by designing the modulation of the wave and signal processing to achieve previously unobtainable imaging modalities. These new imaging systems can have more than an order of magnitude increase in depth of field, for example. The coding of signals so that particular information is optimally conveyed is not new. For example, in radar, coding of the pulses is done to optimally provide information concerning a target's range. At the radar receiver, signal processing is required to extract the range information. In a similar manner, an imaging system can have a coded wavefront in order to provide range or misfocus information in a manner such that signal processing can easily extract the desired information in the scene. Alternatively, the wavefront can be modified so that the imaging system is invariant to object depth or to misfocus of the system, so that, after processing, good resolution can be obtained over a large image depth.

A primary feature of Wavefront Coding in imaging is that the optical image-gathering portion of the system can be designed in combination with the design of the signal processing. This is normally not done in the design of imaging systems. It is done, however, in tomography, coded aperture imaging, and sometimes, interferometric imaging. In 1984, a group pointed out the potential of increasing the performance of imaging systems by jointly designing the optics and the signal processing. Since that time, major advances have been made in doing this. Now, several analysis and design tools are available, and are described in the following section.

Theory and Applications of Wavefront Coding

Wavefront Coding refers to a modification of the imaging system by placing specialized phase plates (or phase and amplitude masks) in the aperture stop, or in an image of the aperture stop, of the imaging system. This could be a phase plate as a separate optical element, or, if a surface of a lens is close to the aperture stop, the Wavefront Coding surface can be superimposed onto the normal lens surface. Processing to decode the image that is formed by the detector array (the intermediate image) is related to the coding that is impressed on the wavefront. The term 'signal processing' also includes the spatial integration provided by the width and height of the pixels. The final image is hence a function of the optics that form the intermediate image and the signal processing that produces the final image.

In the following section, we discuss the Woodward function as used in analysis and design of hybrid imaging systems, and show visualizations of the effect of Wavefront Coding on the image performance of two dramatically different types of systems. Then, other design techniques are discussed for imaging systems using Wavefront Coding and several examples of systems are given where the depth of field of imaging systems is extended by at least an order of magnitude. By concentrating on the increase in the depth of focus, rather than the depth of field, very simple imaging systems can be designed where Wavefront Coding makes the imaging system invariant to focus-related aberrations, and the shape of the lens is used to minimize the other aberrations.

Wavefront Coding Theory

The Woodward function

P. M. Woodward showed that the capability of a radar system to estimate both the range and the velocity of a target can be described by a mathematical function that became known as the Ambiguity Function. The Ambiguity Function shows how well systems that use different types of radar pulses can measure the transit time of the pulse (hence the range to the target) and the Doppler shift in the frequency of the returned pulse (hence the velocity of the target). Because the parameters, range, and Doppler, are measured with different accuracy, there is ambiguity in the measurement, and the Ambiguity Function is a display of the tradeoffs that are possible, Brenner, Lohmann, and Ojeda-Castañeda showed that the same mathematical function, with differently defined variables, can represent the Optical Transfer Function (OTF) of a one-dimensional imaging system.

The Ambiguity Function for a radar system is given by

$$A(u,v) = \int P(x + u/2)P^*(x - u/2)\exp(\mathrm{i}2\pi v x)\mathrm{d}x \quad [1]$$

where $\mathrm{i} = \mathrm{sqrt}(-1)$, * denotes complex conjugate, u is time, v is frequency, x is a dummy variable, and $P(x)$ is a complex function representing the transmitted waveform of the radar system. The OTF of a one-dimensional imaging system for different values of misfocus can be written as

$$H(u,\psi) = \int \left\{ P(x+u/2)\exp(j(x+u/2)^2\psi) \right\} \times \left\{ P^*(x-u/2)\exp(-j(x-u/2)^2\psi) \right\} \mathrm{d}x \qquad [2]$$

where $P(x)$ represents the complex pupil function of the imaging system. The misfocus parameter ψ can be shown to be equal to the physical parameters:

$$\psi = (\pi L^2/(4\lambda))(1/f - 1/d_o - 1/d_i) = 2\pi W_{20}/\lambda = kW_{20} \qquad [3]$$

where L is the one-dimensional length of the aperture, and λ is the center wavelength of the narrowband illumination. The distance d_o is measured between the object and the first principal plane of the lens, and d_i is the distance between the second principal plane of the lens and the image plane. The focal length of the lens is given by f. The wavenumber is given by k and the traditional misfocus aberration constant is given by W_{20}. A comparison of the form of the ambiguity function and the OTF shows that:

$$H(u,\psi) = A(u, u\psi/\pi) \qquad [4]$$

The mathematics of the Woodward function as used for the Ambiguity Function has been studied extensively. Hence one can apply the theory that was worked out for the Ambiguity Function to the study of imaging systems. This gives a tool that is, in many ways, more powerful than other forms of three-dimensional optical transfer functions. For example, it can be shown that the squared volume under the Woodward function is a constant. This means that whatever one does to modify the Woodward function, there is a constraint on what can be done. Another constant is the power under a vertical cut of the Woodward function at a particular value of spatial frequency. This means that there is constant power for all misfocus values at any given spatial frequency. It can be bunched up about the normal in-focus position, or it can be spread out as shown below, but the squared area does not change.

Examples of the Woodward function for different imaging systems

A display of the Woodward function for a high-quality, one-dimensional lens is shown in Figure 1a. The horizontal line and the slanted line represent in-focus and out-of-focus slices through the Woodward function. A projection onto the horizontal axis gives

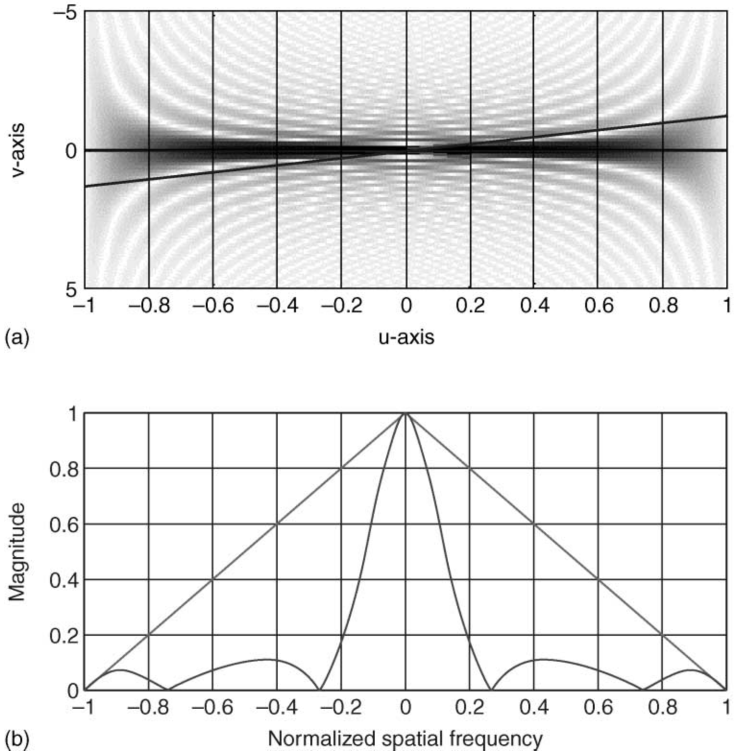

Figure 1 (a) The Woodward function for a high-quality one-dimensional lens with slices that represent in-focus and out-of-focus imaging. (b) The corresponding values of the modulation transfer function for those two cases.

the OTFs for those two focal conditions. The magnitudes of those two OTFs, the MTF, is shown in Figure 1b. Note that the out-of-focus MTF goes through a zero, where there is a phase shift, which gives a contrast reversal in the image.

Extended depth of field. Figure 2a shows the corresponding display of the Woodward function for a one-dimensional lens with a cubic phase function, $\phi = \alpha x^3$ added in the aperture stop. The appearance of the Woodward function immediately shows that this imaging system will have an increased depth of field, because the Woodward function shows that the energy is spread over a larger misfocus region. The slices in the Woodward function are taken for the same misfocus values as were used for the images in Figure 1. Notice that the MTFs of Figure 2b are very similar, and that the MTF for the misfocused case does not go through a zero. This implies that the imaging system will have an extended depth of focus.

The analogy to the Ambiguity Function of radar was key in selection of the cubic function for early demonstrations of extending the depth of field of an imaging system. Work in the radar field had shown that a radar pulse, where the carrier frequency increases quadratically with time, has an Ambiguity Function that is identical to the distribution in Figure 2a. Because a quadratic in time is a cubic in phase, the cubic phase function:

$$\phi = \alpha x^3 \qquad [5]$$

was selected as one that will provide the same Woodward function, and hence an increase in the depth of field. Many families of phase distributions have now been found that extend the depth of focus of an imaging system.

Antialiasing. When the resolution of the optics of an imaging system exceeds the resolution of the image detector array, the high spatial frequencies cannot be detected. They are not lost, however. When a spatial frequency is sampled below the Nyquist sampling rate, the high spatial frequency shows up, or aliases, as a lower spatial frequency. This aliased frequency is equal to the sampling frequency (determined by the pixel spacing) less the difference between the high frequency that has been undersampled and the sampling frequency. This results in an image with broad bands (low spatial frequencies) across the image as is often seen in television images of patterns with high spatial frequency content, such as striped shirts. Because the signal power at a given spatial

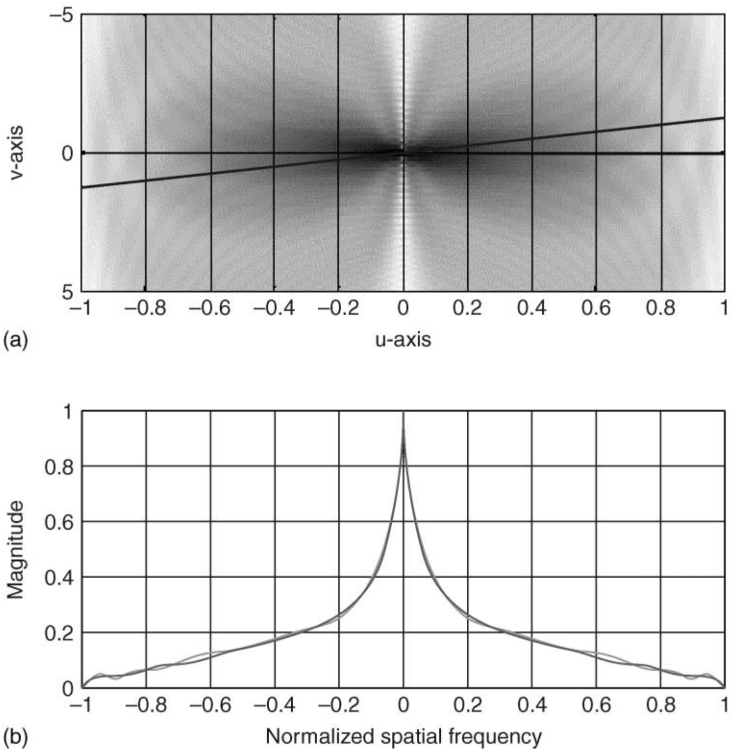

Figure 2 The Woodward function and the modulation transfer functions for a high-quality one-dimensional lens that has a phase plate with a cubic profile in the aperture stop. (a) The Woodward function showing a broadened pattern that is indicative of an increased depth of field. (b) The MTFs that are associated with the slices that represent in focus and out of focus.

frequency over all misfocus (the squared area under a vertical slice of the Woodward function at that spatial frequency) is constant, these frequency components cannot be removed. They can, however, be reduced at a given focus value by spreading the Woodward function vertically. Figure 3 shows a Woodward function that would be ideal. Below the cutoff spatial frequency, all of the signal power is in the focal plane, and beyond the cutoff frequency, the power is spread as much as possible over a large misfocus region. There is no assurance that a phase distribution can be found to give a desired Woodward function, however. Figure 4 shows the effect of one phase function, given by

$$S(r, \theta) = r^3 \cos(3\theta) \qquad [6]$$

upon the MTF of the imaging system. The dashed vertical line shows the cutoff spatial frequency that is due to the sampling rate of the pixels. Note that in Figure 4, the magnitude of the Wavefront Coded spatial frequencies that are aliased have lower values than do those that are passed by a traditional full aperture F/2.5 system. Stopping the traditional system down to F/25 reduces the aliased MTF to an amount close to that obtained with the wide aperture and Wavefront Coding, but it also reduces the amount of light by a factor of 100.

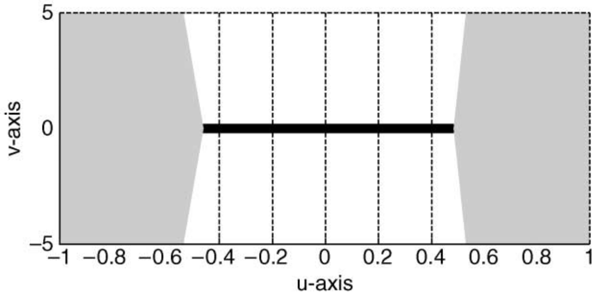

Figure 3 The Woodward function for an ideal antialiasing filter.

Wavefront Coding in Imaging Systems with Extended Depth of Field or Focus

Figure 2 shows the Woodward function for one phase distribution (cubic in x and y for a two-dimensional imaging system) that increases the depth of field of an imaging system. The cubic function was selected by using analogies to the Ambiguity Function of radar. Even though it is doubtful that this phase distribution could have been found by using conventional lens design techniques, it is instructive to show the ray traces and point spread functions that are obtained when this phase distribution is used in the aperture stop.

Ray trace analysis

Figure 5 shows the ray trace from a point object through a conventional (one-dimensional) lens and through a lens that has been modified with the cubic profile of eqn [5]. Figure 5a shows the rays through a conventional lens coming to a focus, and Figure 5b shows the corresponding rays as they go through a lens with a cubic profile added. In the latter case, the

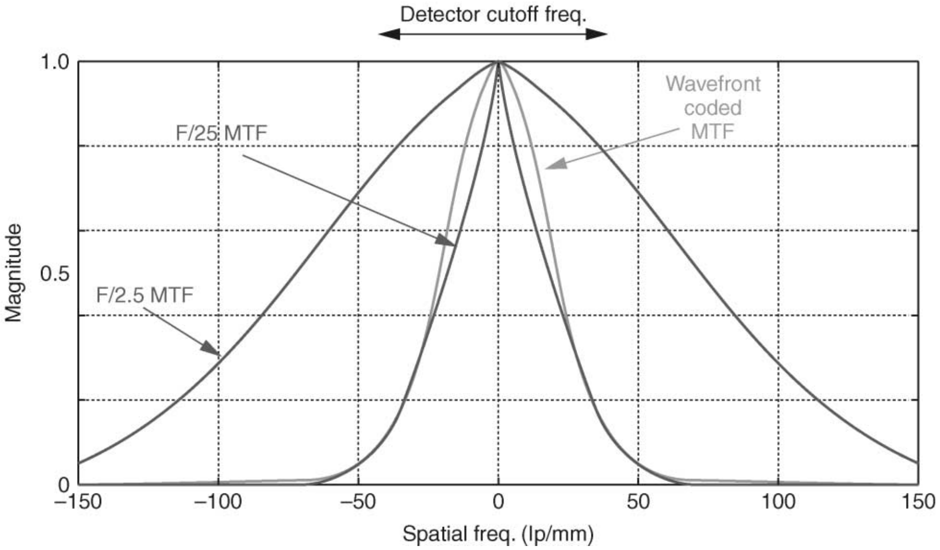

Figure 4 Normalized MTFs for traditional optics and Wavefront Coding. The full aperture F/2.5 system has a large amount of optical power beyond the detector cutoff frequency of 50 lp/mm. With Wavefront Coding the aliased optical power can be greatly reduced. The traditional system aperture needs to be reduced to F/25 to compare to the antialiasing performance of the Wavefront Coded system. This reduction reduces the light level by a factor of 100 in comparison with the F/2.5 system.

rays cross, but have a relatively constant ray density in the region of what was the focal region. Figure 6 shows experimentally acquired point spread functions (PSF) for both cases. Note that the PSF changes very little over the region of misfocus. After signal processing, the PSF can be brought back to that of the conventional imaging system when in focus. Figure 7 illustrates the effect of the signal processing upon the intermediate image that is acquired by the optics. The object is a tilted bar pattern, so that a portion is in focus, and the remainder of the bar pattern is so far out of focus that there is a contrast reversal in the image. With the system that has a coded wavefront to give an extended depth of field, the entire image appears to have the same blur, no matter which part of the tilted bar pattern is observed. This is because the same PSF is convolved with the object distribution for all regions of misfocus. Signal processing removes the effect of the modified PSF, resulting in the image of Figure 7b. Figure 7c shows a trace through the image of Figure 7b. The amount of the misfocus that can be tolerated is a design parameter, and is discussed in the next section.

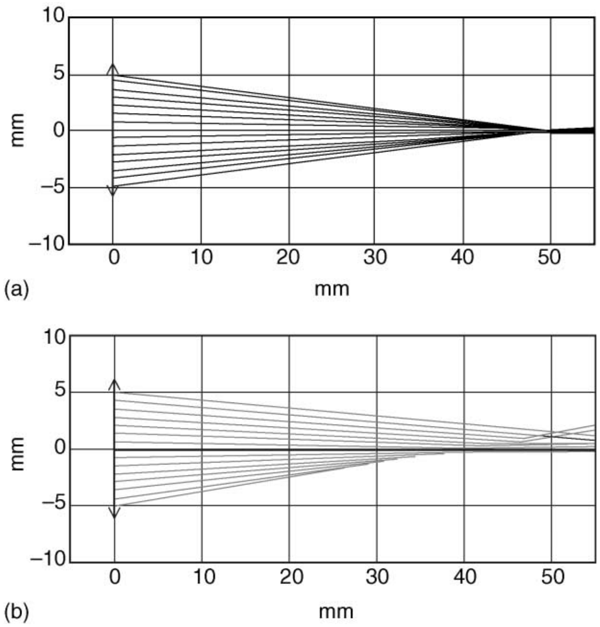

Figure 5 Ray traces from a point object through (a) a conventional one-dimensional lens, and (b) a lens that has been modified to have a cubic profile on the surface.

Signal processing requirements

It is necessary to do signal processing when using an imaging system that employs Wavefront Coding; the signal processing is needed to ‘decode’ the image that is formed by the detector array. This intermediate image is not a clear image. To obtain an intermediate image, the object distribution is convolved with a PSF, or impulse response, that has a shape similar to that

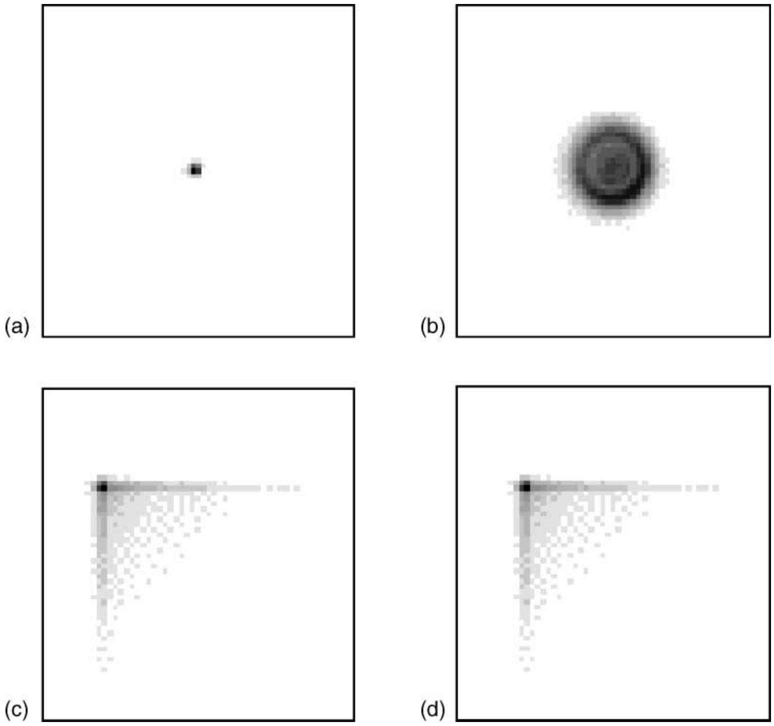

Figure 6 Point spread functions for a conventional lens in focus (a) and with large misfocus (b), and a lens that has been modified to have a cubic profile on the surface when in focus (c) and with a large misfocus (d).

of Figure 6c and d. To obtain the final image, the intermediate image must go through a deconvolution with the PSF. Because of the square law detection that occurs in the detector array, this deconvolution does not simply undo the effects of the convolution with the PSF that describes image formation. A straightforward digital convolution is required in most cases, and has been shown to run at video frame rates in software.

Parameter trade offs in imaging systems with extended depth of field

Depth of field and signal-to-noise ratio. The greater the phase modulation by the phase plate (with a cubic profile in this example), the greater the extension of the depth of field. The greater the phase modulation for extending the depth of field, the larger the PSF, and the greater the drop in the MTF, as seen in Figure 2b, which show the MTF of the imaging system. The use of the phase function to extend the depth of field causes portions of the MTF to sag as compared with the MTF of the in-focus conventional system. This reduces the signal-to-noise ratio. If there is a greater phase alteration, the MTF sags more.

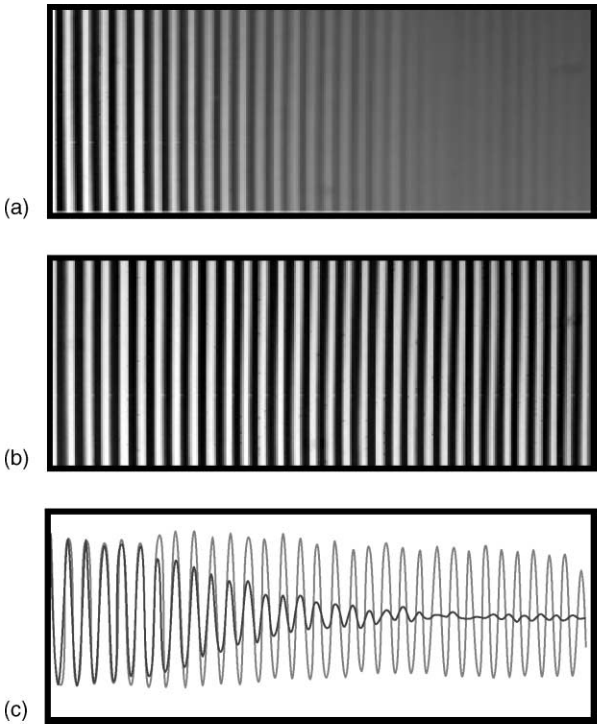

Figure 7 Image of a tilted bar pattern with an extended depth of field imaging system. (a) Image with a conventional imaging system. (b) Final image from the imaging system with Wavefront Coding. (c) Traces through the images to give quantitative comparisons of the image quality.

Focus budget and focus-related aberrations. Another set of trade offs can be seen by considering the imaging systems from the detector plane. Here, we see an increase in the depth of focus of the system. All of this increase in the focal depth can be used to increase the depth of field if the optics of the system are of high quality. Otherwise, the increase in the depth of focus can be used to compensate for focus-related aberrations. For example, chromatic aberration occurs when different colors focus to different planes. Figure 8 shows that if the depth of focus is extended, then there is a region where red, blue, and green, for example, are in focus. Figure 9 shows the red, blue, and green images of a US Air Force test pattern, as formed with a two-element imaging system with severe chromatic aberration, and with the same system when a Wavefront Coding element is used to extend the depth of focus.

Compensation can also be seen for the case of curvature of field, where the image of a plane object falls onto a curved surface. If an extension of the depth of focus causes the focal region to resemble a curved bar instead of a curved line, then it is possible to place a flat detector array in the curved region, and therefore obtain a good image after signal processing. Similar arguments can be made for other focus-related aberrations. This leads to the concept of a focus budget, where a portion of the increased depth of focus can be allocated to extend the depth of field by some amount, and other portions of the depth of focus allocated to provide invariance to focus-related aberrations.

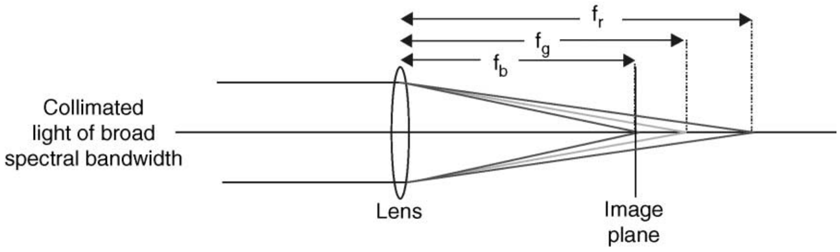

Figure 8 Ray traces showing different focal planes for red, blue, and green, and the effect of extending the depth of focus for each color.

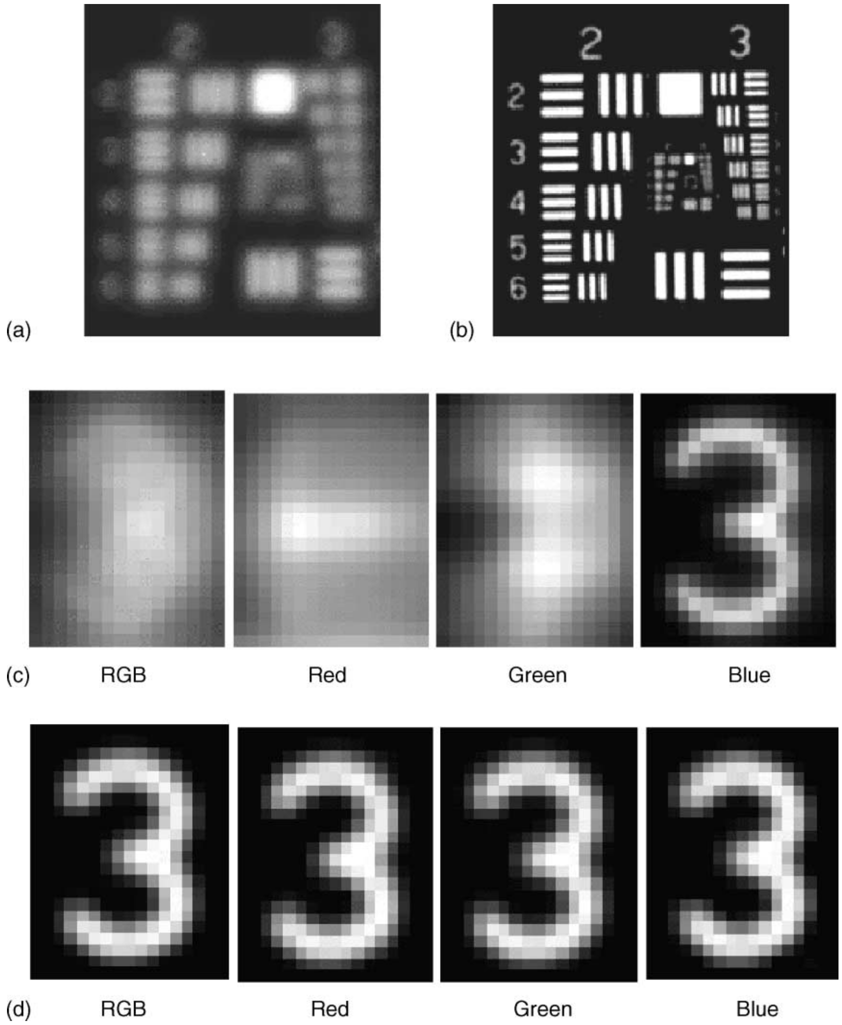
(a)
(b)
(c)
RGB
Red
Green
Blue
(d)
RGB
Red
Green
Blue

Figure 9 Color images of a US Air Force test pattern. (a) Images from a two-element lens of the center portion of the test pattern. (b) Images from an extended-depth-of-focus imaging system of the center portion of the test pattern. (c) Images from a two-element lens of a small number ‘three’ on the test pattern. (d) Images from an extended-depth-of-focus imaging system of a small number ‘three’ on the test pattern. Some residual effects remain due to different magnifications at different wavelengths.

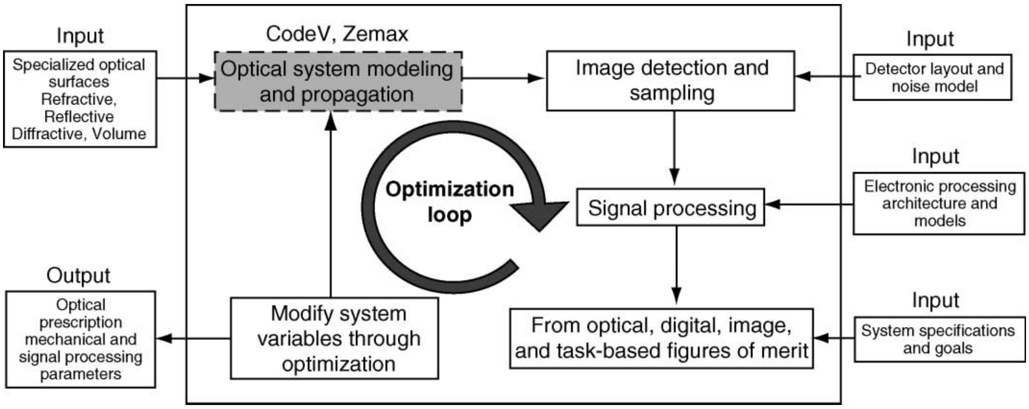
Input
Specialized optical surfaces Refractive, Reflective Diffractive, Volume
CodeV, Zemax
Optical system modeling and propagation
Image detection and sampling
Input
Detector layout and noise model
Optimization loop
Signal processing
Input
Electronic processing architecture and models
Output
Optical prescription mechanical and signal processing parameters
Modify system variables through optimization
From optical, digital, image, and task-based figures of merit
Input
System specifications and goals

Figure 10 Schematic of a tool for designing imaging systems with Wavefront Coding. A ray trace determines the distribution in the exit aperture, and software determines the effect of the detector array and the signal processing. Then the ray trace and signal processing can be modified to optimize the entire system.

New design tools for imaging systems with wavefront coding

The Woodward function is a very good tool for analyzing imaging systems with a coded wavefront and is a good guide for someone who is expert in designing such systems. However, another approach is better for those designers who are more familiar with conventional lens design tools. If one combines a conventional ray-tracing tool with custom software, it is possible to jointly design the optical and the signal processing portions of the system. A schematic of such a system is shown in Figure 10. The ray-tracing tool is used to trace the rays and determine the distribution in the exit aperture. Various software tools then are used to take into account the effect of pixel size and spacing, and the signal processing. The criteria for optimizing the design can include the size of the digital filter, acceptable noise gain, and other parameters related to signal processing as well as the parameters of the optical elements.

Example applications of hybrid imaging systems with an extended depth of field or focus

Barcode and label readers. The images of Figure 7 show that image-based bar code reading is a good application for imaging systems that are based on Wavefront Coding. A bar code or label reader need not be positioned nearly as accurately as without the

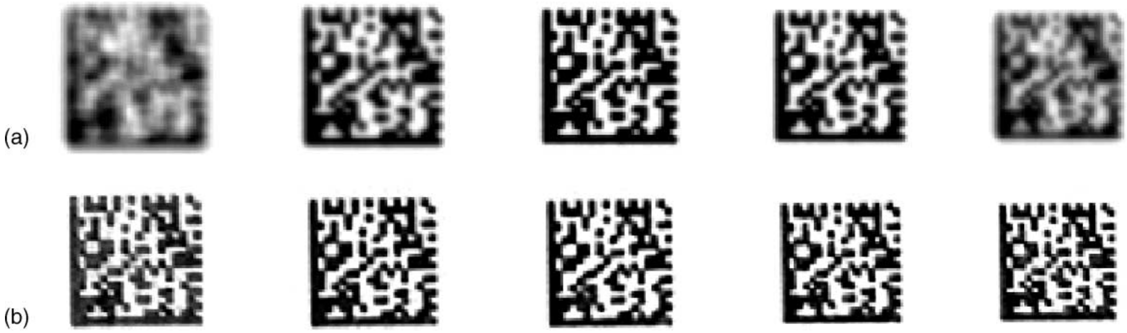

Figure 11 Images of a 2D bar code as the object position is varied for (a) conventional images and (b) Wavefront Coded images after signal processing for the same object positions as in (a).

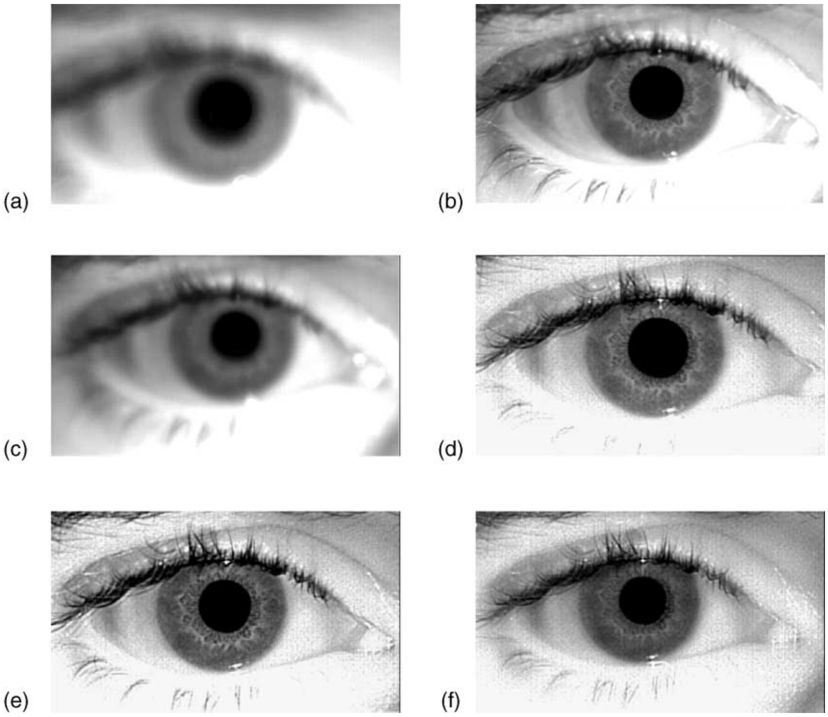

Figure 12 Images of an iris taken with infrared illumination. The top row shows images with a traditional CMOS camera. The lower row is taken with a camera using wavefront coding to extend the depth of field. The misfocus distance in the right and left columns is the same.

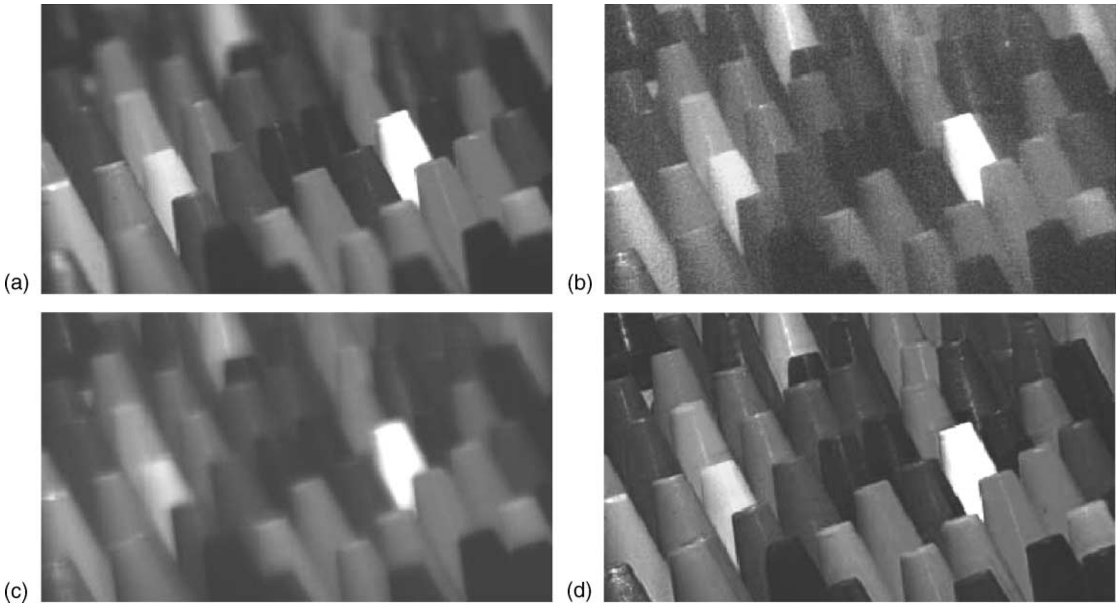

Figure 13 Images from a digital camera (a) Conventional camera with the lens wide open and focused on the white crayon. (b) Conventional camera with the lens stopped down to increase the depth of field, with the resulting increase in noise due to a lower light level. (c) Intermediate image of a camera that was modified with Wavefront Coding to have an extended depth of field. (d) Final image with the modified camera.

extended depth of field. Images of 2D bar codes imaged with and without Wavefront Coding are shown in Figure 11. Images in Figure 11a show the image of a 2D bar code formed by a conventional imaging system as the position of the object is varied. As with most imaging systems with fast optics and moderate magnification, the depth of field of the system is smaller than the application would prefer. Images in Figure 11b show the Wavefront Coded images of a 2D bar code after signal processing for the same object positions as (Figure 11a). These images are sharp and clear over a very large change in the distance to the bar code.

Biometric imaging systems. Biometric imaging applications, such as fingerprint or iris recognition, provide the means to eliminate passwords, secure E-commerce, and securely identify humans or animals. In traditional imaging technology, quite often the trade-off between light throughput and depth of field makes conventional biometric systems difficult to use or not sufficiently reliable. By using Wavefront Coding, the depth of information gathered can greatly increasing the ease of use and possibly reducing the cost of the systems.

Figures 12a,b, and c shows example images of an iris taken with near IR illumination and a traditional CMOS imaging system at three different planes of focus. At best focus (Figure 12b) the detail of the iris is sharp and clear. This iris detail is used by iris recognition algorithms to accept or reject the person.

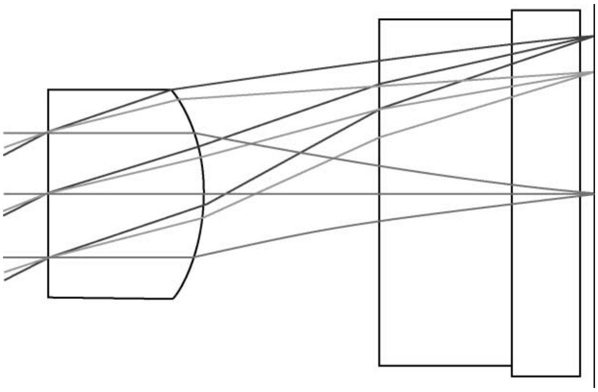

Figure 14 A schematic of a fast single-lens imaging system that has good performance over a wide field of view and the full visible range.

With slight movement towards (Figure 12a) or away (Figure 12c) from the imaging system the detail of the iris is lost. Figures 12d,e, and f show example iris images taken with the same near-IR illumination and a CMOS Wavefront Coded imaging system. At best focus (Figure 12e), the iris detail is as high as that from the traditional system. Movement towards (Figure 12d) or away (Figure 12f) from the imaging system results in images with essentially a constant level of iris detail. The size of the iris changes through magnification change. Increasing the range at which clear images can be made extends the useful image capture volume, enabling a more flexible system that is easier to use.

Color images with digital cameras. A conventional digital camera has a small depth of field as shown in Figure 13a, where the camera was focused on the white crayon. In a conventional imaging system, the only way to increase the depth of field is to stop down the lens. As seen in Figure 13b, this reduces the light that is available, and results in a very noisy image. The intermediate image of an imaging system that has an extended depth of field is shown in Figure 13c, where it is seen that the result of convolving the object distribution with the modified PSF is an intermediate image that appears to have the same blur over the entire image. After signal processing of the intermediate image to obtain the final image, a clear image is seen with good signal-to-noise ratio.

Inexpensive imaging systems with high-quality imaging characteristics. By exploiting the extension of the depth of focus that is made possible with Wavefront Coding techniques, one can design imaging systems that have fewer optical elements than traditional designs that do not use Wavefront Coding. In high-volume applications, reduction in the number of optical elements can lead to greatly reduced costs. Figure 14 shows a schematic of a fast, single-lens Wavefront Coded imaging system that is very compact and has the performance of a longer lens that traditionally requires two or more optical elements. In this example, the entire focus budget was used to compensate for focus-related aberrations, and the shape of the lens was used to compensate for distortion and coma. Even zoom systems can be greatly simplified with the new design techniques. Figure 15 shows a fast two-element Wavefront Coded zoom lens that images over a wide field with performance that is traditionally obtained only with three or more optical elements. Both of these imaging systems use non-rotationally symmetric forms of phase functions and are different from the ones of eqns [5] and [6].

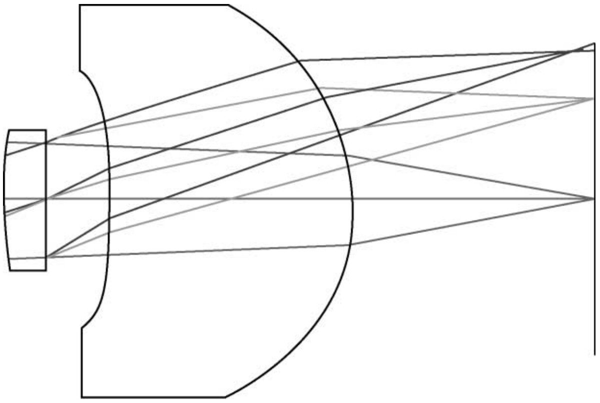

Figure 15 A schematic of a fast two-element zoom system with good performance over a 2.5 × zoom and a wide viewing angle.

Large depth of focus microscopes. A microscope can be modified to employ Wavefront Coding by either modifying the surface of one of the lenses in the objective, or by inserting a phase plate into the system at some appropriate position. One easy means of doing the modification is to place a Wavefront Coding phase plate in the slider position of a microscope, as shown in Figure 16. In this case, the optics is of high quality, and the entire focus budget was allocated to increasing the depth of field. A modified Zeiss microscope was used to obtain the images shown in Figure 17. An image of a diatom in Figure 17 shows the comparison with a conventional microscope in (a) and as with a Wavefront Coded microscope in (b). Notice that the greater depth of field gives more of an appearance of depth in the image.

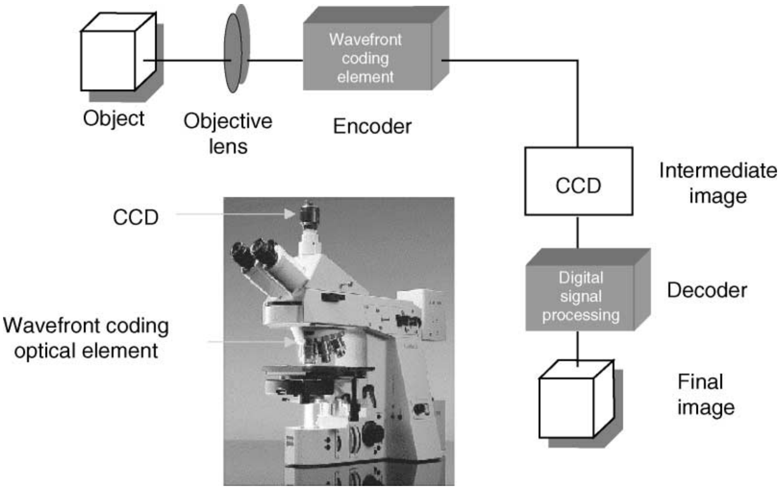

Figure 16 Schematic of a modified Zeiss microscope to employ Wavefront Coding to extend the depth of field.

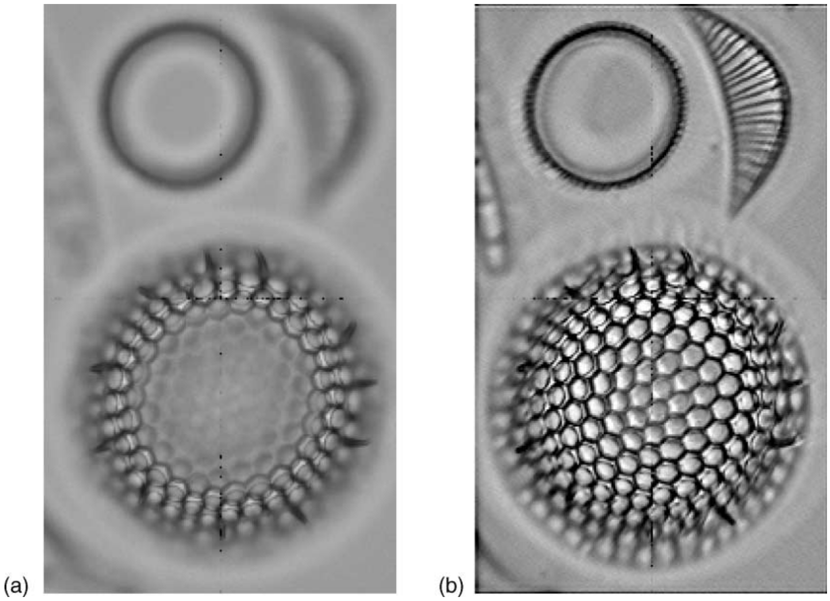

Figure 17 A 100 × image of diatoms made with (a) conventional and (b) modified Zeiss microscope to give an extended depth of field.

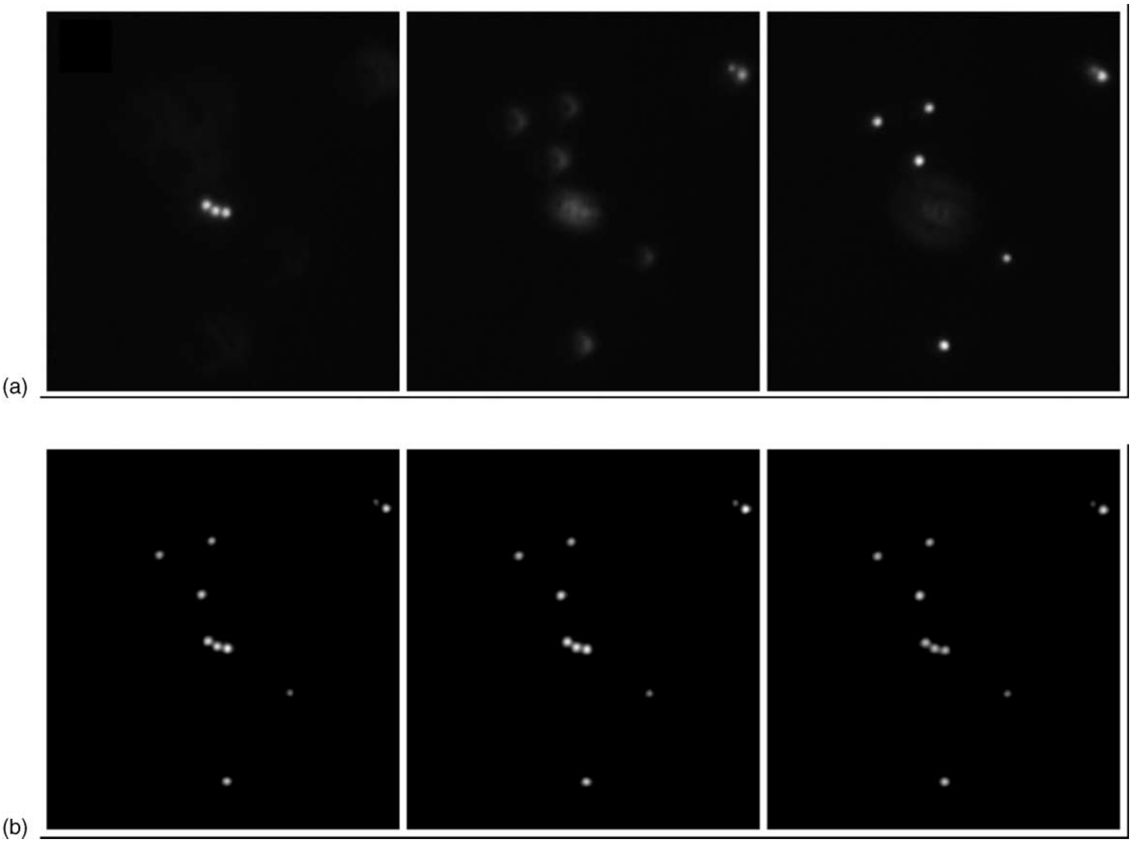

Figure 18 Imaging 0.5 micron diameter fluorescent beads at different sample distances. The images were taken with a 63×, NA = 1.4 imaging system in one micron steps. (a) Three image planes taken using a traditional system. They show that individual beads are sharp only in a single plane. The images in (b), which were taken with a Wavefront Coding system at the same image planes, show all beads are imaged clearly at each plane.

Other modes of microscopy are also possible. In fluorescence microscopy, ultraviolet light illuminates objects that have fluorescing compounds in them. The images are formed using the fluorescence that is excited by the ultraviolet illumination. Figure 18 shows images of fluorescing beads having a 0.5 micron diameter. The beads are at various depths within the sample. The traditional system, operating at 63 × with an immersion objective having a NA of 1.4, has a depth of field that is too small to image all of them. The images in Figure 18a were taken with the same microscope but with a wavefront coding phase element placed in the slider and signal processing of the captured images. All beads are clearly seen at all three of the focus positions.

Potential Future Applications

Wavefront Coding should be useful in many imaging system where signal processing is to be done anyway; in that case, optimizing the optics and the signal processing to the purpose of the imaging system. Examples where an extended depth of field may be desired are endoscopes, video cameras, and lithography. In designing an imaging system to take advantage of a large depth of focus to produce inexpensive digital cameras, high-quality images can be formed with fewer optical elements and much lighter cameras. In lithography, a larger process window is possible.

Because an increase in the depth of focus can permit wide spectral imaging, wide band infrared images can be formed. In systems that are used to provide images for machine interpretation, it is possible to design the optics and signal processing so that the desired process can be optimized globally, with the optics and the signal processing. For example, if the object is to classify cells as cancerous or noncancerous, it may be that the optics should be modified so that the information of primary importance is acquired and transmitted to the detector array. This modification would be done in conjunction with the optimization of the signal processing.

Conclusion

The mathematics and concepts that underlie a new approach to lens design were summarized. The technique has thus far been useful in enhancing the performance of some imaging systems such as microscopes, and in making simpler systems for inexpensive cameras. This new approach requires that some signal processing be done, but in digital cameras, some signal processing is being done already.

List of Units and Nomenclature

Modulation transfer function	The magnitude of the optical transfer function of an imaging system.
Optical transfer function	The response of an imaging system as a function of the spatial frequency of a sinusoidal object distribution.
Spatial frequency	A measure of spatial variation in the form of sinusoids in cycles per millimeter.

See also

Imaging: Wavefront Sensors and Control (Imaging Through Turbulence). **Microscopy:** Imaging Multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Barrett HH and Swindell W (1981) *Radiological Imaging: The Theory of Image Formation, Detection, and Processing*. New York: Academic Press.

Brenner K, Lohmann A and Ojeda-Castañeda J (1983) The ambiguity function as a polar display of the OTF. *Optics Communications* 44: 323–326.

Cathey WT and Dowski ER (2002) New paradigm for imaging systems. *Applied Optics* 41: 6080–6092.

Cathey WT, Frieden BR, Rhodes WT and Rushforth CK (1984) Image gathering and processing for enhanced resolution. *Journal of the Optical Society of America A* 1: 241–249.

Cushman SC (2001) *Finite-Dimensional Matrix Treatment of Partially Coherent Optical Systems*. University of Colorado thesis, Boulder, Colorado.

Dowski ER and Cathey WT (1995) Extended depth of field through wavefront coding. *Applied Optics* 34: 1859–1866.

FitzGerrell AR (1993) *Applications of the Ambiguity Function to Optics*. Thesis, University of Colorado, Boulder, Colorado.

Häusler G (1972) A method to increase the depth of focus by two step image processing. *Optics Communications* 6: 38–42.

Ojeda-Castañeda J, Tepichin E and Diaz A (1989) Arbitrary high focal depth with quasioptimum real and positive transmittance apodizer. *Applied Optics* 28: 2666–2670.

Papoulis A (1974) Ambiguity function in Fourier optics. *Journal of the Optical Society of America* 64: 779–788.

Rihaczek AW (1969) *Principles of High Resolution Radar.* New York: McGraw-Hill.

Van der Gracht J, Dowski ER, Taylor MG and Deaver DM (1996) Broadband behavior of an optical-digital focus-invariant system. *Optics Letters* 21: 919–921.

Wach HB, Dowski ER and Cathey WT (1998) Control of chromatic focal shift through wave-front coding. *Applied Optics* 37: 5359–5367.

Woodward PM (1953) *Probability and Information Theory with Applications to Radar.* New York: Pergamon Press.

PHOTON PICTURE OF LIGHT

S J Bentley, Adelphi University, Garden City, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The photon nature of light is the central topic in the field of quantum optics, which has developed into a leading area of research. The possible technological impacts of this field to many areas, such as imaging, computing, and lithography, have greatly expanded interest into the nature of the photon. A photon description of light goes well beyond the particle nature of light, to include any optical phenomena that cannot be adequately described by classical physical optics. This article will introduce the nature of the photon through a discussion of many such phenomena, covering topics from the beginning of quantum mechanics, including black-body radiation and Compton scattering, to areas of modern interest, such as squeezed light and entangled photons.

The Origins of Quantum Optics

One of the earliest theories in the development of quantum physics was the description of black-body radiation by Max Planck (1858–1947). Using the model of a radiating body as a collection of oscillators, he postulated that the amount of energy emitted or absorbed was proportional to its frequency, ν, and that these oscillators could only have energies corresponding to integer multiples of such a packet of energy, where one unit of energy was given by

$$E = h\nu \qquad [1]$$

where $h = 6.626 \times 10^{-34}$ Js is the universal constant, now known as Planck's constant. He then assumed a Maxwell–Boltzmann distribution to describe the probability of such oscillator states being occupied:

$$P(n) = P_0 e^{-nh\nu/kT} \qquad [2]$$

where $k = 1.381 \times 10^{-23}\ \mathrm{JK}^{-1}$ is Boltzmann's constant and T is the temperature of the blackbody. The energy density in the radiation field, at a given frequency, is given by the oscillator density (at that frequency) times the average energy of the oscillators. Previous models had not only failed to quantize the oscillator energies, but in doing so also assumed equal occupation probabilities that led to a prediction of infinite energy densities at short wavelengths, the so-called ultraviolet catastrophe. The oscillator density is given by

$$n(\nu) = 8\pi\nu^2/c^3 \qquad [3]$$

where $c = 3 \times 10^8\ \mathrm{m\ s}^{-1}$ is the speed of light. This value can be found by a mode density calculation within a box, as is often done in an introductory electromagnetics text. The average energy of the oscillators is found by calculating the sum of the occupation probability times the state energy and dividing by the sum of the occupation probabilities:

$$E_{\mathrm{av}} = \sum_{n=0}^{\infty} [P(n)h\nu n] \Big/ \sum_{n=0}^{\infty} P(n) \qquad [4]$$

Evaluating the above expression for E_{av} and multiplying by $n(\nu)$, one finds the Planck radiation law:

$$I(\nu) = \frac{8h\pi\nu^3}{c^3(e^{h\nu/kT} - 1)} \qquad [5]$$

The black-body curve described by Planck's law is shown in Figure 1. By introducing quantization to solve a problem that classical physics had failed to explain, Planck paved the way for quantum physics.

Albert Einstein (1879–1955) made use of Planck's quantization for two important results that will be described here. The first is the photoelectric effect, a process by which electrons are released from a metal surface when exposed to light. The process has several traits that could not be explained by

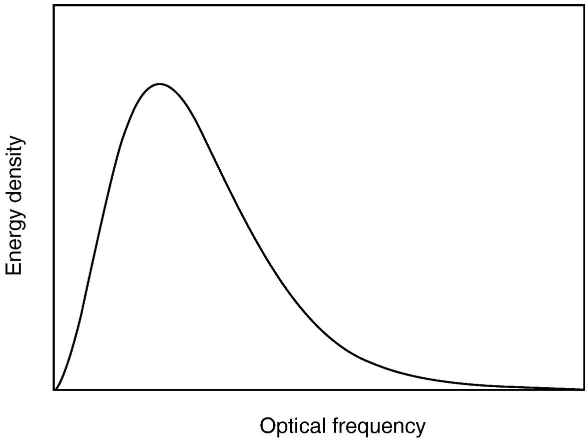

Figure 1 The spectral density emitted by a blackbody at a given temperature as given by Planck's law.

classical physics. A given metal will only release electrons if the incident light is above a certain frequency, regardless of the intensity. However, once sufficient frequency components are present, electrons are immediately released, with their number dependent on the intensity of the light and independent of its frequency. Contrarily, the maximum potential energy of the released electrons depends only on the maximum frequency of light present, independent of intensity. Einstein was able to satisfy all of these conditions by saying that the light energy comes in quantized amounts as described by Planck, and that the electrons in the metal require a minimum energy, or work function W, to be ejected. Then for each quantum of light, which we now call a photon, with energy greater than W that strikes an electron, the electron will be released with kinetic energy given by

$$K = h\nu - W \qquad [6]$$

It is historically important to note that, although Planck introduced the idea of a quantized release of radiation from oscillators, he still believed the light field itself to be continuous and classical. Planck's approach of quantizing energy levels of matter while treating light classically has since become known as the semiclassical method, and is adequate for solving many problems. Thus, it was also in quantizing light itself that Einstein took the next step, and thus can be viewed in many respects as the father of the modern photon, although the term photon was not used until two decades later.

The second area in which Einstein used the idea of energy quantization is in predicting radiative transition probabilities. A radiative transition is a process by which a photon is either absorbed or emitted to satisfy conservation of energy when an atom moves to a higher or lower energy state. Einstein assumed three such processes existed, as shown in Figure 2: absorption of a photon raising the atom to a higher energy level; spontaneous emission of a photon, where the atom emits a photon in going to a lower energy level; and stimulated emission, a process in which a photon interacts with the atom and causes it to emit a second photon (in phase with the first) and thereby go to a lower energy level. For each of these processes he defined a corresponding coefficient, now known as the Einstein A and B coefficients. They are labeled B_{12} for absorption, A_{21} for spontaneous emission, and B_{21} for stimulated emission. Note that there is no A_{12} as absorption by its very nature is a stimulated process. Einstein further stated that the rate of spontaneous emission was only dependent on his material dependent coefficient and the number of atoms in the upper energy state, with the rate given by

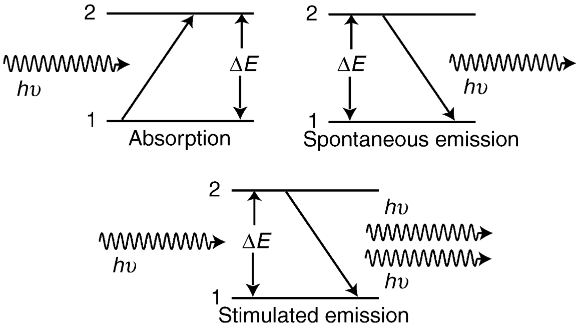

Figure 2 Energy level diagrams representing the processes of absorption, spontaneous emission, and stimulated emission.

$$R_{\text{sp}} = A_{21}N_2$$

However, he said that the rates of the stimulated processes were also dependent upon the strength of the optical field seen by the atoms, $I(\nu)$, and were given by

$$R_{\text{ab}} = I(\nu)B_{12}N_1 \qquad [7]$$

for absorption and

$$R_{\text{st}} = I(\nu)B_{21}N_2 \qquad [8]$$

for stimulated emission. The full absorption rate is given by R_{ab} as defined above, whereas the full emission rate is the sum of the spontaneous and stimulated rates, given by

$$R_{\text{em}} = [A_{21} + B_{21}I(\nu)]N_2 \qquad [9]$$

Einstein was primarily interested in a system in equilibrium, for which $R_{\text{ab}} = R_{\text{em}}$. If one then looks at the limit where the optical field is very large such that spontaneous emission is negligibly small and the population is equally divided between the upper and lower state, the relationship $B_{12} = B_{21}$ is found. Applying the same Maxwell–Boltzmann occupation statistics that Planck used to find his black-body expression, one can find that in thermal equilibrium the ratio of probability of finding an atom in state 1 to finding an atom in state 2 is

$$N_1/N_2 = e^{(E_2-E_1)/kT} = e^{h\nu/kT} \qquad [10]$$

Note that for a two-level system in thermal equilibrium, the population in the upper level can at most equal, never exceed, the population in the lower level, regardless of the strength of the applied field. This is important in showing that the B coefficients are equal, and will also be a crucial point when discussing the theory of the laser later in this article.

Appropriately combining the above expressions leads to

$$I(\nu) = A_{21}/[B_{21}(e^{h\nu/kT} - 1)] \quad [11]$$

Equating this with Planck's radiation law yields the relation:

$$A_{21}/B_{21} = 8\pi h\nu^3/c^3 \quad [12]$$

Thus, to fully specify the transition probabilities for a given system, one needs only to find *A* or *B*. It is rather straightforward using only atomic quantization to calculate *B* (such a derivation can be found in many introductory quantum mechanics textbooks). Direct calculation of *A* is more involved, requiring quantization of both the optical field as well as the atomic system.

Arthur Compton (1892–1962) used the idea of quantized radiation to explain the presence of longer wavelength components present when X-rays are scattered from matter, an effect which now bears his name. The amount of scattering is independent of the scattering material and the wavelength of the source, depending only on the angle of scattering, θ. If the scattering is treated as a collision between a particle-like photon and a massive particle (taken to be initially at rest) as shown in Figure 3, the result can then be found simply by applying conservation of energy and momentum, where the photon energy is as given above and the photon momentum is given by

$$p = E/c = h/\lambda \quad [13]$$

Recalling that the momentum is a vector quantity, the conservation laws provide a set of three coupled equations. Solving them directly one finds that

$$\Delta\lambda = (h/c)(1 - \cos\theta)(2K)/[p^2 - (K/c)^2] \quad [14]$$

where *K* and *p* correspond to the final kinetic energy and momentum of the massive particle. It is important to allow for the possibility of very high velocities for the massive particle, thus implying that relativistic expressions for *K* and *p* are needed. In terms of the kinetic energy, the relativistic expression for momentum is

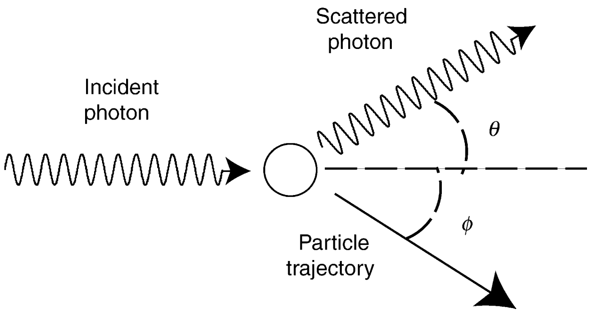

Figure 3 Schematic representation of Compton scattering.

$$p = \sqrt{2Km + (K/c)^2} \quad [15]$$

where *m* is the mass of the scattering particle. Using this relation, the above expression can be simplified, giving the wavelength shift to be

$$\Delta\lambda = (1 - \cos\theta)h/mc \quad [16]$$

The scaling factor is known as the Compton wavelength:

$$\lambda_c = h/mc \quad [17]$$

For the case of an electron, the Compton wavelength has the value 0.02426 angstroms. For a massive particle, the Compton wavelength is often used when wanting to discuss and analyze wavelike properties. As a final note on this process, if the calculation would have been done nonrelativistically, there would have been an extra factor of $[1 - 0.25(v/c)^2]^{-1}$, where v is the velocity of the particle. The appearance of the factor (v/c) in any result implies the need for a relativistic treatment, though often (as in this case) taking the limit of $v/c = 0$ yields the proper result.

Another area in which the photon nature of light is revealed is a form of inelastic scattering of light now known as the spontaneous Raman effect. Discovered by CV Raman (1888–1970), this is a process in which a material illuminated by a light source scatters light into frequency components not found in the source. The frequency shift is due to the energy difference, ΔE, between two energy states in the material. There are two possible cases, each of which is represented in Figure 4. In the first, known as Stokes scattering, an atom in the lower state absorbs a source photon, raising it to the upper state via a virtual level, during which it emits a photon at the Stokes frequency, ν_s, as represented by the relation

$$h\nu_i = h\nu_s + \Delta E \quad [18]$$

where ν_i is the frequency of the incident photon from the source. The second case, anti-Stokes scattering, has an atom in the upper state absorb the source photon and simultaneously fall to the lower state, again through the virtual level during which it emits a photon at the higher anti-Stokes frequency, ν_a,

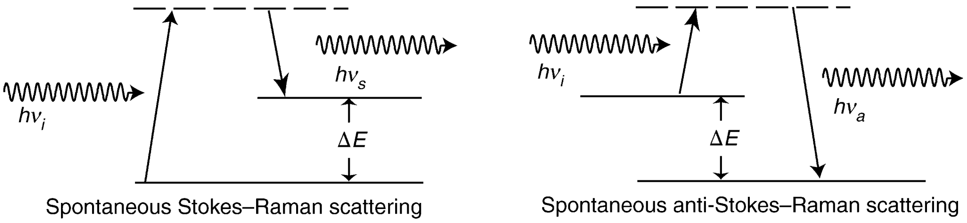
$h\nu_i$
$h\nu_s$
ΔE
Spontaneous Stokes–Raman scattering
$h\nu_i$
$h\nu_a$
ΔE
Spontaneous anti-Stokes–Raman scattering

Figure 4 Energy level diagrams for Stokes and anti-Stokes Raman scattering.

given by

$$h\nu_i = h\nu_a - \Delta E \quad [19]$$

Since the anti-Stokes scattering requires the atom to be initially in the upper state, the probability of such scattering is generally much less than that for Stokes scattering. This can be quantified by the same occupation statistics used above, giving the ratio of probabilities as

$$P_{\text{as}}/P_s = e^{-h\Delta E/kT} \quad [20]$$

It is important to note that the energy difference can either be based on the electronic structure of an atomic system as has been discussed thus far, or other energies such as the vibrational energies in a molecular system. The ability to probe a variety of energy structures without having to tune the source to a particular wavelength has led to wide use of Raman scattering for spectroscopy. A related effect, known as stimulated Raman scattering, will be discussed later.

The Photon in Modern Physics

As an indication of the widespread importance of the results discussed thus far, they led to four Nobel prizes in physics: Planck in 1918; Einstein in 1921; Compton in 1927; and Raman in 1930. Our initial understanding of the photon comes from these results, but it was with the development of modern quantum mechanics by Erwin Schrödinger (1887–1961), Werner Heisenberg (1901–1976), and Paul Dirac (1902–1984); quantum electodynamics by Richard Feynman (1918–1988) and others; and the invention of the laser by Charles Townes (1915–) and others, that the field of quantum optics and our modern view of the photon began to fully take shape. In this section, the basics of modern quantum optics will first be introduced. A brief discussion of the theory of the laser as related to photon transitions will then be given, followed finally by examples of important areas of experimental quantum optics that have been made accessible by the laser.

It is now common to treat the photon with a very similar formalism as that originally created for massive particles in the development of quantum mechanics. That is, one describes the photon by a wave function, $|\Psi\rangle$, which contains all the information about the photon and must satisfy the Schrödinger equation

$$i\hbar\frac{\partial}{\partial t}|\Psi\rangle = H|\Psi\rangle \quad [21]$$

where $\hbar = h/2\pi = 1.054\times10^{-34}$ Js and H is the Hamiltonian for the photon. Just as in classical electromagnetics, the Hamiltonian is related to an electric and magnetic field. Clearly, for this treatment to be in agreement with the well-established classical electromagnetic theory, these fields must satisfy Maxwell's equations. There are several possible representations of the Hamiltonian, but for the purposes of this discussion a convenient choice is

$$H = h\nu(\hat{n} + 0.5) \quad [22]$$

where $\hat{n}$ is called the number operator as it returns the average number of photons in the state, $\langle\hat{n}\rangle$, represented by the operation

$$\langle\hat{n}\rangle \equiv \langle\Psi|\hat{n}|\Psi\rangle \quad [23]$$

Explanations of the notation used can be found in any elementary quantum mechanics text. Note that in defining the Hamiltonian, a single frequency ν was used. This implies a single mode field. For a more general case, one would simply sum over the modes corresponding to various frequencies.

In addition to obeying the Schrödinger equation, the photon is also restricted by the Heisenberg uncertainties as for massive matter, given in a general form as

$$\Delta q\Delta p \geq \hbar \quad [24]$$

where q and p are any set of canonically conjugate variables, such as position and momentum or energy and time. Note that, depending on the exact definitions used for q and p, there can be extra factors such as 0.5, and sometimes normalized variables are used such that the inequality is with respect to 1. However, it is interesting to note that, unlike for massive particles that were considered well localized in classical theory, the idea of uncertainty for light had been accepted in some form long before the development of quantum mechanics. Neglecting the scaling factor $\hbar$, which is purely quantum in origin, an uncertainty principle can be derived for any wavelike phenomena simply using the relationship between Fourier transform pairs.

Using the basic ideas of a photon state and the uncertainty principle, some important properties of light will now be explored. There are several possible forms the wave function can take. It is often convenient to write the wave function in terms of states containing a certain number of photons. Such states are known as Fock or number states and are represented as $|n\rangle$, with n being the number of photons in the state and are said to be eigenstates of the number operator. The energy of a Fock state is given by

$$E_n = h\nu(n + 0.5) \qquad [25]$$

Notice that this predicts a background, or vacuum, energy for the electromagnetic field, even when the average number of photons in the field is zero. This vacuum energy is given by 0.5 $h\nu$. Correspondingly, due to the uncertainty principle, there must be some uncertainty in this energy as well as with any energy. The fluctuations in the vacuum energy are known as vacuum fluctuations. Many processes, such as spontaneous emission that cannot be adequately described by a simple semiclassical approach, where matter is treated quantum mechanically and light is treated classically, can be treated by the addition of vacuum fluctuations, without requiring a full quantum mechanical treatment of the electromagnetic field. Such an approach is used in many texts, but there are several important physical phenomena that do require the fully quantized approach discussed here.

The state most resembling a classical field is known as the coherent state, represented here as a sum of Fock states:

$$|\Psi\rangle = |\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \qquad [26]$$

where α is a parameter describing the field, with the average number of photons in the field given by

$$\langle \hat{n} \rangle = |\alpha|^2 \qquad [27]$$

One key property of coherent states is that they exhibit minimal uncertainty. That is, the uncertainty relations become equalities. Also, the uncertainty is equally divided between quadrature components.

Before applying the above principles to explore some of the dramatic results of quantum optics, a brief discussion of the laser will be given due to its importance in the field. The theory of operation of the laser is based directly on Einstein's transition probabilities. For the transition of interest, one wants to generate a large number of coherent photons (that is, photons in phase with one another). Recall that in the process of stimulated emission, one photon leads to two coherent photons, whereas spontaneous emission leads to photons which are not phase referenced to anything else in the system. Also, if a significant number of the photons are to escape the region where they are generated, it is important to minimize the absorption. Thus, a laser requires that the stimulated emission rate exceed both the spontaneous emission rate and the absorption rate. The first condition places a lower limit on the optical field density present at the lasing frequency. This density can be reasonably large, but presents no fundamental difficulties. In practice, mirrors are placed around the emission region to confine the photons and counteract losses, thereby building up the energy density until the above condition is satisfied. The second condition, however, can be written as

$$N_2 > N_1 \qquad [28]$$

and is the more restrictive requirement. Such a condition is referred to as inversion and, as was previously mentioned, cannot be sustained in a two-level system once equilibrium is reached. Therefore, to be able to maintain lasing, additional levels are needed. Two typical laser energy diagrams are shown in Figure 5. In such multilevel systems, it is possible to create an inversion in the levels of interest and thus allow lasing. Further details on lasers (both in theory and practice) are widely available, though it should be noted that the output of a laser is a good approximation of a coherent state described above.

The invention of the laser revolutionized optics, as well as the whole of modern physics, by making readily available electromagnetic fields of a vastly different nature than that from any previous source of light. This quickly led to the birth of a new field known as nonlinear optics. Nonlinear optics describes interactions between light and matter which have properties dependent on the field strength. Although some nonlinear processes can be described adequately without the use of a photon picture, most nonlinear process can be best described as some combination of annihilation (absorption) and creation (emission) of photons. Two such

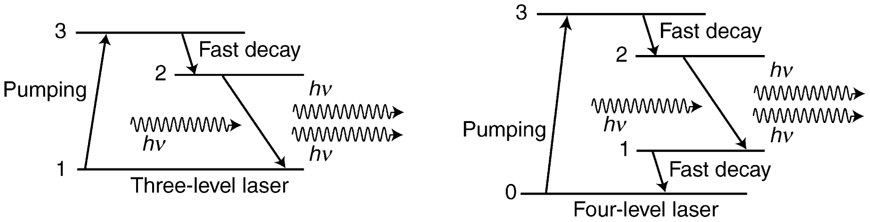

Figure 5 Energy level diagrams for three-level and four-level lasers.

processes, that are of great interest in quantum optics, are stimulated scattering and parametric down conversion.

Earlier, the process of spontaneous Raman scattering was discussed. For such a process, it is assumed that the incident field driving the process is relatively weak such that the scattering process generates very low photon densities in the various modes of the electromagnetic field. However, just as there are both stimulated and spontaneous emission processes in the radiative transitions discussed previously, in the presence of strong fields there can also be stimulated scattering. The total Raman scattering rate can be expressed as

$$R = \eta I_i(I_{\mathrm{R}} + 1) \qquad [29]$$

where η is a scaling factor, including the gain of the process for a given material, I_i is the strength of the incident optical field (generally a laser field to be sufficiently strong to cause stimulated scattering), and I_{R} is the Raman field strength. Note that for I_{R}, the rate is just linear with the applied field, representing the spontaneous process. However, with a strong pump, the generated field will quickly become large enough such that the stimulated process will dominate. The spontaneous scattering is not directional dependent, but since the stimulated process grows along the length of the scattering material in the path of the incident beam, it will tend to be generated along the path of the incident laser. Due to linear and nonlinear dispersion effects, however, the light will form in a narrow cone about the pumping laser rather than exactly collinear. Unlike the spontaneous scattering, a significant percentage of the incident photons can be converted to the Raman frequencies, making stimulated Raman scattering an extremely powerful spectroscopic tool.

Possibly the most important nonlinear optical process for quantum optics is parametric down conversion. This process is interesting both in the nature of the process itself, as well in the properties of the generated light. In parametric down conversion, shown schematically in Figure 6, one photon (known as a pump photon) of frequency ν_p is annihilated and two new photons (generally referred to as signal and idler) are created with frequencies ν_s and ν_i respectively, given by the conservation of energy relationship:

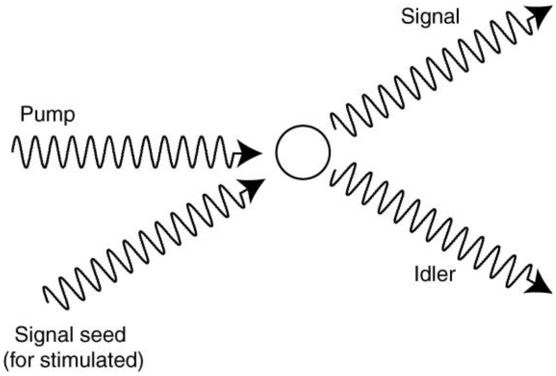

Figure 6 Schematic representation of parametric down conversion.

$$h\nu_p = h\nu_s + h\nu_i \qquad [30]$$

In the case where $\nu_s = \nu_i$, the process is referred to as degenerate. Down conversion can either be spontaneous, in which only pump photons are externally applied, or stimulated, in which both pump photons and signal photons are applied. In the stimulated case, the signal is amplified and idler photons are generated as a byproduct. Parametric down conversion is one of the leading sources of nonclassical states of light. The term nonclassical is used rather than quantum, because all states of light can be accurately described by a quantum picture, but while most can also be described very well with a classical picture the states referred to as nonclassical cannot. Two of the most common nonclassical states of light are entangled states and squeezed states.

Entangled states are multiphoton states in which the state cannot be written as a product of the states of the individual photons. For simplicity, only two-photon entangled states (commonly referred to as EPR states for Einstein, Podolsky, and Rosen) will be discussed. More extensive and general material, including discussions of higher-order entanglement,

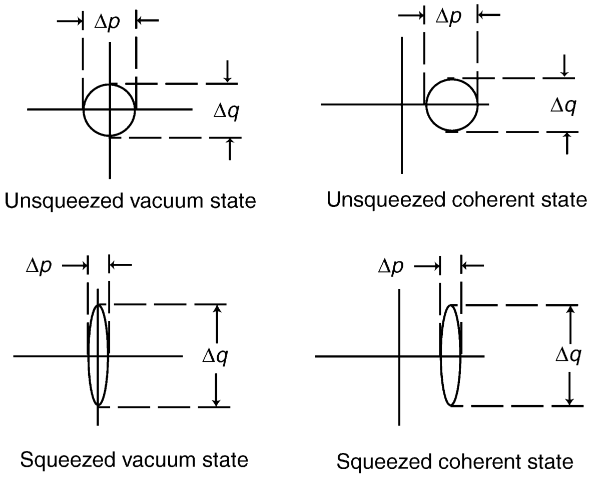

Figure 7 Uncertainty diagrams for squeezed and unsqueezed vacuum and coherent states.

such as the three-photon entangled states (commonly referred to as GHZ states for Greenberger, Horne, and Zeilinger), can be found in a variety of textbooks on modern quantum optics. One of the simplest ways to generate a two-photon entangled state is by the process of spontaneous parametric down conversion described above. If the product of the pumping intensity, generation length, and generation gain is of an appropriate level, individual pairs of photons are emitted. The states of these photon pairs are entangled, with the position, momentum, wavelength, polarization, and all other properties of the photons completely correlated and inseparable. Such states have been a powerful tool for the experimental verification of many of the predictions of quantum mechanics and quantum optics. Entangled states can also act as a source for many of the newer technologies in quantum optics, including cryptography and teleportation.

Another important form of nonclassical light is known as squeezed light. The principle behind squeezed light is that, unlike for coherent states where the uncertainty is equally distributed, the uncertainty in one quadrature can be reduced at the expense of increased uncertainty in the other quadrature. This is shown schematically in Figure 7. As opposed to entangled states which tend to be very weak, containing only a few photons, squeezed states of light can be quite intense. This makes squeezed light more desirable for many applications.

In conclusion, it should be noted that while the photon description of light can explain a wide variety of physical processes and is now leading to many exciting technological developments, much is still unknown about the photon at the fundamental level. Any accurate description of the photon must be in agreement with experimental evidence, including that discussed in this article, but one should be aware that the accepted picture of the photon nature of light is still constantly evolving and growing.

See also

Nonclassical Light. **Nonlinear Optics, Applications:** Raman Lasers. **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Quantum Optics:** Entanglement and Quantum Information. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Boyd RW (2002) *Nonlinear Optics*, 2nd edn. San Diego, CA: Academic Press.

Cohen-Tannoudji C, Diu B and Laloë F (1977) *Quantum Mechanics*. New York: Wiley-Interscience Publications.

Hecht E (2001) *Optics*, 4th edn. Reading, MA: Addison-Wesley Publishing.

Heisenberg W (1949) *The Physical Principles of the Quantum Theory*. New York: Dover Publications.

Lipson SG, Lipson H and Tannhauser DS (1995) *Optical Physics*, 3rd edn. Cambridge, UK: Cambridge University Press.

Loudon R (2000) *The Quantum Theory of Light*, 3rd edn. Oxford, UK: Oxford University Press.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Milonni PW and Eberly JH (1988) *Lasers*. New York: Wiley-Interscience Publications.

Roychoudhuri C and Roy R (eds) (2003) *OPN Trends – The Nature of Light: What is a Photon?* Washington, DC: Optical Society of America.

Schwinger J (ed.) (1958) *Selected Papers on Quantum Electrodynamics*. New York: Dover Publications.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Yariv A (1989) *Quantum Electronics*, 3rd edn. New York: John Wiley & Sons.

PHOTONIC CRYSTALS

Contents

Atomic Physics
Electromagnetic Theory
Microwave Photonic Crystals
Nonlinear Optics in Photonic Crystal Fibers
Photonic Crystal Lasers, Cavities and Waveguides
Self-Assembled and Functionalized Photonic Crystals

Atomic Physics

G Kurizki and A G Kofman, Weizmann Institute of Science, Rehovot, Israel

D Petrosyan, Institute of Electronic Structure and Laser, Heraklion, Greece

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fabrication and study of dielectric structures whose refractive index is periodically modulated on a micron or submicron scale, known as photonic crystals (PCs), are attracting considerable interest at present. One-, two-, or three-dimensional (1D-, 2D-, or 3D-) periodic PCs exhibit photonic band gaps (PBGs), analogously to electronic band gaps in ordinary crystals, and can incorporate defects, designed to form localized narrow-linewidth (high-*Q*) modes at PBG frequencies. The advent of such structures opens up new perspectives in atomic physics and quantum optics, since they are expected to allow an unprecedented control over the spectral density of modes (DOM) and the spatial modulation of narrow-linewidth (high-*Q*) modes, in the microwave, infrared, and optical domains. An interesting situation arises when foreign atoms or ions – dopants – with transition frequencies within the PBG, are implanted in the PC. Then light near one of these frequencies resonantly interacts with the dopants and is concurrently affected by the PBG dispersion. Consequently, highly nonlinear processes with a rich variety of unusual PC-related features are anticipated. Such nonlinear optical processes, involving near-resonant transitions in PCs, undergo basic modifications as compared to the corresponding processes in free space, which are attributed to the strong suppression of the DOM within PBGs, to sharp bandedges and to intra-gap narrow lines associated with high-*Q* defect modes.

Results are detailed below for spontaneous emission of radiation in PCs and for the resonant interaction of atoms with the field of a single high-*Q* defect mode. These results stem from the failure of perturbation theory and the onset of strong field-atom coupling near sharp bandedges or narrow defect-mode lines in 2D and 3D PCs. 3D-PBGs are needed, in order to extinguish spontaneous emission in all possible directions of propagation and dipole orientations. If only one polarized atomic transition is involved in the spontaneous emission, 2D-PCs suffice for its suppression. For controlling strictly unidirectional field propagation, it is sufficient to resort to PBGs in 1D-periodic structures (Bragg reflectors or dielectric multi-layer mirrors).

As a further example of field–atom interactions in doped PCs, we will show that two ultraweak electromagnetic fields (or photons), coupled to appropriate transitions of the dopants in a PC, can mutually induce large phase shifts or drastic changes in absorption. Appreciable photon–photon correlations can then be established, which can be utilized in both classical and quantum optical communications.

Radiative Decay and Photon–Atom Binding in a PC

The interaction of a two-level atom with an arbitrary field-mode continuum can be described by the following second-quantized Hamiltonian in the rotating-wave approximation (RWA)

$$H = \hbar\omega_{\mathrm{a}}|e\rangle\langle e| + \hbar\sum_{\vec{k}} \omega_{\vec{k}} a^{\dagger}_{\vec{k}} a_{\vec{k}} + \hbar\sum_{\vec{k}} [\kappa(\omega_{\vec{k}}) a^{\dagger}_{\vec{k}} |g\rangle\langle e| + \text{hermitian constant}] \qquad [1]$$

Here $|e\rangle$ and $|g\rangle$ are the excited and ground atomic states, respectively, ω_{a} is the atomic transition frequency, $a^{\dagger}_{\vec{k}}$ and $a_{\vec{k}}$ are the creation and annihilation operators of a field mode with wavevector $\vec{k}$, and frequency $\omega_{\vec{k}}$, and $\hbar\kappa(\omega_{\vec{k}})$ is the resonant coupling energy of this mode with the atomic dipole. The $\vec{k}$-summation can be replaced by an integral over the continuum DOM in the frequency domain, $\sum_{\vec{k}} \to \int_0^{\infty} \mathrm{d}\omega\rho(\omega)$, avoiding, for the sake of simplicity, the study of direction-(angle-)dependent DOM effects.

We consider a two-level atom that is excited at time $t = 0$ in an empty, periodic dielectric structure. The wavefunction of the combined system field + atom can be cast in the general form:

$$|\Psi(t)\rangle = \alpha(t)|e, \{0_{\omega}\}\rangle + \int \beta_{\omega}(t)|g, 1_{\omega}\rangle\rho(\omega)\mathrm{d}\omega \qquad [2]$$

where $\{0_{\omega}\}$ signifies the completeness of field modes in the vacuum state and $|1_{\omega}\rangle$ denotes single-photon occupation of the ω-mode. The corresponding Schrödinger equation can be solved with the initial condition $|\Psi(0)\rangle = |e, \{0_{\omega}\}\rangle$ by means of the continuum spectral response:

$$G(\omega) = |\kappa(\omega)|^2\rho(\omega) \qquad [3]$$

The analysis of eqn [2], with $G(\omega)$ appropriate for photonic bandstructures, is aimed at revealing the prominent features of the atomic excitation decay $\alpha(t)$ and the corresponding emission spectrum.

In what follows, we consider a photonic bandstructure with several PBGs, separated by allowed bands. Each PBG is labeled by index i and has lower and upper cutoff frequencies $\omega_{\mathrm{L}i}$ and $\omega_{\mathrm{U}i}$, respectively. Then $G(\omega) = 0$ for $\omega_{\mathrm{L}i} < \omega < \omega_{\mathrm{U}i}$. Naively, one might expect that an excited atom would either be stable against single-photon decay, if ω_{a} is within a PBG, or decay completely at $t \to \infty$, if ω_{a} is anywhere in an allowed band. However, both statements turn out to be inaccurate.

Incomplete decay of $\alpha(t \to \infty)$ occurs if there is a stable eigenvalue (energy level) $\hbar\omega_i$ of the total (field-atom) Hamiltonian eqn [1]. Such an energy level is possible only if $G(\omega_i) = 0$, i.e., for ω_i in a PBG. It must satisfy:

$$\omega_i = \omega_{\mathrm{a}} + \Delta(\omega_i) \qquad [4]$$

$$\Delta(\omega_i) = \int_0^{\omega_{\mathrm{L}i}} \frac{G(\omega')}{\omega_i - \omega'}\mathrm{d}\omega' + \int_{\omega_{\mathrm{L}i}}^{\infty} \frac{G(\omega')}{\omega_i - \omega'}\mathrm{d}\omega' \qquad [5]$$

Here the integrals are the frequency shifts of the atomic resonance $\hbar\omega_{\mathrm{a}}$ induced by the spectral parts of the reservoir situated, respectively, below and above the ith PBG (Figure 1). If eqn [4] holds, then the corresponding term in $\alpha(t)$ is proportional to $\exp(-i\omega_i t)$, and does not decay. Physically, such a stable level can be interpreted as representing the binding of the photon to the atom (photon-dressed atom), without the ability to leave the atomic vicinity, due to Bragg reflection in the PBG.

Assuming that ω_i occurs in the ith PBG, we observe that the first shift is positive whereas the second shift is negative, i.e., each part of the continuum repels the atomic level from its PBG edge. Equation [4] has a solution if ω_{a} falls between the minimum and maximum values assumed by its left-hand side in the ith PBG, i.e., if

$$\omega_{\mathrm{L}i} - \Delta(\omega_{\mathrm{L}i}) < \omega_{\mathrm{a}} < \omega_{\mathrm{U}i} - \Delta(\omega_{\mathrm{U}i}) \qquad [6]$$

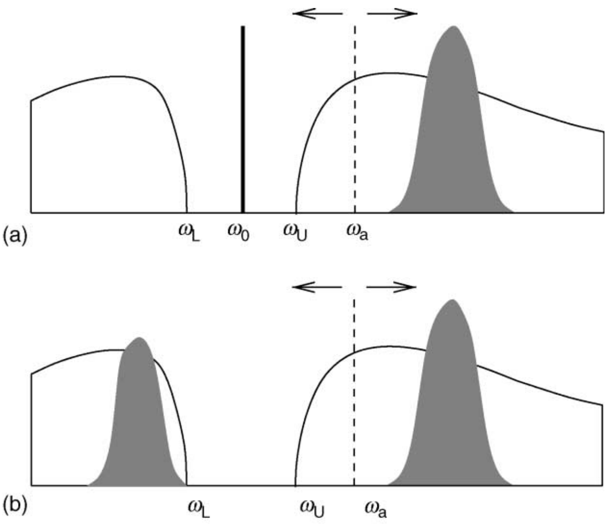

Figure 1 Frequency shifts of the atomic resonance ω_{a} (dashed line) due to 'repulsion' (arrows) by asymmetric spectral parts of the continuum below and above the PBG: (a) Incomplete decay for ω_{a} in allowed band. The interaction with the continuum splits the state into a superposition of a stable part in the PBG (solid line) and a decaying part above ω_{a} (shaded peak). (b) Complete decay for same ω_{a}, due to larger shifts that split the state into decaying parts in two allowed bands.

Incomplete decay occurs if this condition holds at least for one PBG. There may be, at most, one stable state in a single PBG (Figure 1). However, a two-level atom in a periodic structure can have several stable dressed states simultaneously, when conditions hold for several PBGs. It is quite extraordinary that the conditions (eqn [6]) for incomplete decay can be fulfilled with ω_a, not only in a PBG, but even in an allowed band, e.g., when $\Delta(\omega_{\mathrm{U}i})$ is negative or when $\Delta(\omega_{\mathrm{L}i})$ is positive (Figure 1a,b).

Equally counterintuitive is the converse possibility, of complete decay for ω_a within a PBG. When there is a single PBG and one of the inequalities in eqn [6] is violated, complete decay will occur for $\omega_{\mathrm{L}i} < \omega_\mathrm{a} < \omega_{\mathrm{L}i} - \Delta(\omega_{\mathrm{L}i})$, if $\Delta(\omega_{\mathrm{L}i}) < 0$, or $\omega_{\mathrm{U}i} - \Delta(\omega_{\mathrm{U}i}) < \omega_\mathrm{a} < \omega_{\mathrm{U}i}$, if $\Delta(\omega_{\mathrm{U}i}) > 0$.

The possibility that one or several discrete, stable, states exist yields the corresponding wavefunction in the form:

$$|\Psi(t)\rangle = \sum_i \sqrt{c_i}|\psi_i\rangle \mathrm{e}^{-i\omega_i t} + |\Psi_\mathrm{c}(t)\rangle \qquad [7]$$

Here the summation is over all discrete atomic photon-dressed states with energies $\hbar\omega_i$, $|\psi_i\rangle$ is a discrete-(dressed-)state eigenfunction of the Hamiltonian (eqn [1]) normalized to unity, and weighted by the amplitude $\sqrt{c_i} = [1 + \int_0^\infty \mathrm{d}\omega G(\omega)/(\omega - \omega_i)^2]^{-1/2}$. The dressed state $|\psi_i\rangle$ consists of an excited-state component and a photon-bound ground-state component.

The population of the excited state for long times:

$$|\alpha(t)|^2 = \sum_{i,i'} c_i c_{i'} \cos(\omega_i - \omega_{i'})t \ \text{ for } t \to \infty \qquad [8]$$

is a nonzero time-constant if there is one discrete stable state, whereas in the case of several such states, it undergoes beats at the frequencies corresponding to their energy differences. The splitting of an excited state $|e\rangle$ into superposed stable states, oscillating at bandgap frequencies ω_i, whose amplitudes c_i, and eigenfrequencies ω_i are controllable by the atomic transition detuning from cutoff, constitutes spontaneous coherence control.

If, however, ω_a is far from the cutoff, then eqn [2] results in an exponentially decaying amplitude:

$$\alpha(t) \approx e^{-(\gamma_\mathrm{a} + i\tilde{\omega}_\mathrm{a})t} \qquad [9]$$

where $\tilde{\omega}_\mathrm{a} = \omega_\mathrm{a} + \Delta_\mathrm{a}$, Δ_a, and γ_a being the effective spectral shift and width of the decaying atom. This regime holds for a locally smooth $G(\omega \approx \omega_\mathrm{a})$ such that

$$|\gamma'_\mathrm{a}|, |\Delta'_\mathrm{a}| \ll 1;\ |\gamma'_\mathrm{a}\Delta_\mathrm{a}| \ll \gamma_\mathrm{a};\ \gamma_\mathrm{a}, |\Delta_\mathrm{a}| \ll |\omega_\mathrm{a} - \omega_\mathrm{g}| \qquad [10]$$

where ω_g is the bandedge frequency nearest to ω_a and the primes denote differentiation with respect to ω_a.

We now apply the foregoing general results to a model DOM distribution. This distribution is derived on keeping the lowest term in the Taylor expansion of the dispersion relation $\omega(\vec{k})$ near a photonic bandedge ω_U (the effects of the further PBG edge are neglected). This yields the effective mass approximation:

$$\omega \approx \omega_\mathrm{U} + \sum_{i=x,y,z} (k - k_\mathrm{U})_i^2/m_i \qquad [11]$$

with $1/m_i = (1/2)(\partial^2\omega/\partial k_i^2)|_{\omega=\omega_\mathrm{U}}$. In a structure with period L, k_U satisfies the Bragg condition $k_\mathrm{U} = \pi/L$. The corresponding DOM in a 3D-periodic structure with an allowed symmetry may be approximated as

$$\rho(\omega) \sim (\omega - \omega_\mathrm{U})^{(1-D)/2}\theta(\omega - \omega_\mathrm{U}) \qquad [12]$$

where θ is the step function and D is the dimension of the Brillouin-zone surface spanned by bandedge modes with vanishing group velocity. In realistic photonic structures $D \leq 2$, $D = 2$ corresponding to completely isotropic dispersion (spherical Brillouin-zone surface) and $D = 0$ corresponding to an anisotropic three-dimensional Brillouin zone. Both cases can be represented, respectively, as the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ of the function:

$$G(\omega) = \frac{C}{\pi}\frac{\sqrt{\omega - \omega_\mathrm{U}}}{\omega - \omega_\mathrm{U} + \varepsilon}\theta(\omega - \omega_\mathrm{U}) \qquad [13]$$

where ε is the 'cutoff-smoothing' parameter and C is the continuum coupling constant. Depending on whether $C^{2/3}/\varepsilon$ is greater or less than one, the continuum is close to the case $\varepsilon = 0$ ($D = 2$) or $\varepsilon \to \infty$ ($D = 0$), respectively.

Using the properties of the model DOM (eqn [13]), we can infer the criteria for the two regimes discussed above:

(i) The conditions for incomplete decay, eqn [6], are now (Figure 2) $\Delta_\mathrm{c} \equiv \omega_\mathrm{a} - \omega_\mathrm{U} < C\varepsilon^{-1/2}$. Hence, the abrupt, singular cutoff of the DOM with $D = 2$ ($\varepsilon \to 0$) implies the existence of a discrete state for any ω_a, either inside or outside the PBG. The energy of the discrete state, $\hbar\omega_0$, which must lie in the PBG, is found to be a real and positive root of the equation $\omega_\mathrm{a} - \omega_0 = C/(\sqrt{\omega_\mathrm{U} - \omega_0} + \sqrt{\varepsilon})$.

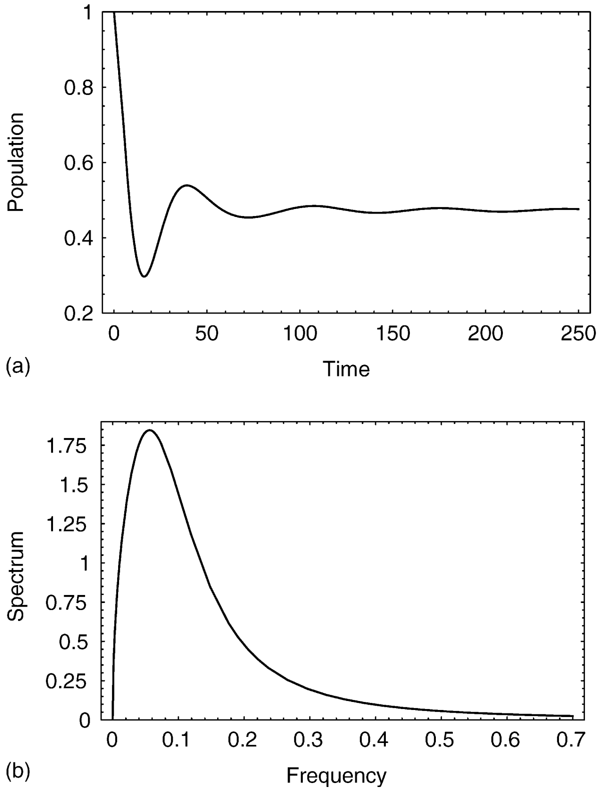

Figure 2 (a) Incomplete decay of the excited state population as a function of $\gamma_c t$, where $\gamma_c = C\varepsilon^{-1/2}$, for $\varepsilon = 10^{-3}$, at cutoff $\Delta_c = 0$. The beat period is $2\pi/(\omega_U - \omega_0)$ and the nondecaying probability is $c_0^2 = 4/9$. (b) The corresponding spectrum. The frequency is normalized to γ_c.

(ii) The conditions for nearly exponential decay, eqn [10], can be shown to reduce now to the requirement that ω_a be in the allowed zone sufficiently far from cutoff, $\Delta_c \gg \min\{C^{2/3}, C\varepsilon^{-1/2}\}$. Under this condition, the intermediate-time exponential decay of the excited state (Figure 3) is given to first approximation by eqn [9], with $\gamma_a = C\sqrt{\Delta_c}/(\varepsilon + \Delta_c)$ and $\Delta_a = -C\sqrt{\varepsilon}/(\varepsilon + \Delta_c)$.

The resulting atomic frequency shift Δ_a is negative in the present model, vanishing for $\varepsilon \to 0$ $(D = 2)$. The long-time behavior of $\alpha(t)$ can be shown to exhibit a tail decaying as $t^{-3/2}$ (or $t^{-1/2}$ at $\Delta_c = C\varepsilon^{-1/2}$) and oscillating at the cutoff frequency ω_U (Figure 3).

With the increase of ε, the smoothing inhibits the decay more strongly for ω_a at cutoff, $\Delta_c = 0$, because $G(\omega_a \simeq \omega_U)$ is now weaker and the stable-state probability c_0 (eqn [7]) is correspondingly larger. By contrast, at a large detuning from cutoff Δ_c, the large-ε smoothing allows the decay to become complete, and the corresponding spectrum is entirely Lorentzian.

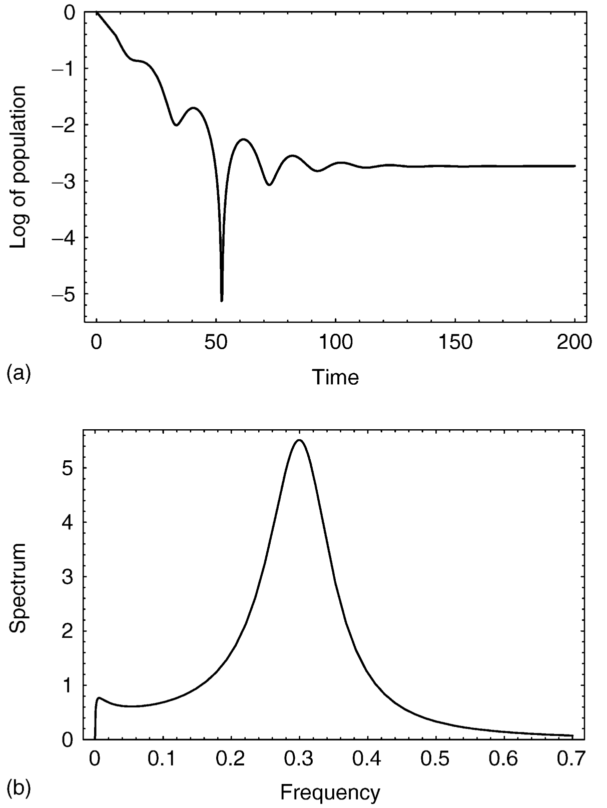

Figure 3 As in Figure 2, for a large detuning from the sharp cutoff, $\varepsilon = 10^{-3}$, $\Delta_c = 0.3$. (a) $\log_{10}|\alpha|t|^2$ is plotted as a function of *i*. The nearly complete decay ($c_0^2 \sim 10^{-3}$) is modulated by beats with frequency Δ_c. A power tail obtains for $t \gg \gamma_a^{-1} \approx \Delta_c^{1/2}/C$, decreasing as $t^{-1/2}$ for $t \ll \Delta_c^2/C^2$ and as $t^{-3/2}$ for $t \gg \Delta_c^2/C^2$. (b) The nearly Lorentzian spectrum, w(ω), has a small peak near ω_U, due to the sharply peaked DOM near the cutoff.

A defect in the periodic structure can produce a narrow-linewidth local mode in the PBG, whose spectral response is describable by a Lorentzian:

$$G_d(\omega) = \frac{\gamma_d}{\pi} \frac{\Gamma_d^2}{\Gamma_d^2 + (\omega - \omega_d)^2} \qquad [14]$$

where γ_d characterizes the coupling strength of the atomic dipole with the defect field, whereas ω_d and Γ_d stand for the line center and width, respectively. The presence of a nonvanishing DOM in the PBG, due to a defect, causes spontaneous emission in the PBG spectral range. This broadens the discrete state ω_0, which becomes metastable (see Figure 4). The oscillator strength of the line at ω_0 is then proportional to c_0.

EIT and Cross-Coupling of Photons in Doped PCs

Nonlinear effects, whereby one light beam influences another, require strong fields or else light confinement in a high-Q cavity. The analysis above has shown that by choosing an appropriate detuning of an atomic transition from the PBG cutoff, the spontaneous coherence is established between the states of an initially excited atom. The ability to control this coherence, by varying the detuning, opens interesting perspectives for optical processes in PCs containing multilevel atoms with a resonant transition near a bandedge. From among such processes, we discuss here electromagnetically induced transparency (EIT) and its applicability to nonlinear photon switching or giant cross-phase modulation.

Let us examine the nonlinear coupling of two weak (single-photon) optical fields E_a and E_b with the frequencies ω_a and ω_b, respectively, propagating along the z-axis in a PC dilutely doped with identical four-level atoms (**Figure 5a**). These fields interact with the atoms via the transitions $|1\rangle \to |2\rangle$ and $|3\rangle \to |4\rangle$, respectively, while the transition $|2\rangle \to |3\rangle$ is coupled to the structured mode-continuum $\rho(\omega)$ in the PC (**Figure 5b**). Initially the atoms are in the ground state $|1\rangle$ and the continuum is in the vacuum state $\{0_\omega\}$. Then the wavefunction of the system at the position z_l of the lth atom reads:

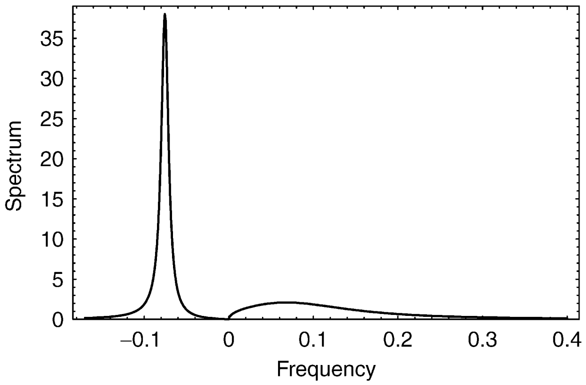

Figure 4 Spectrum w(ω)(same units as in **Figures 2** and **3**) for ω_a at cutoff, $\Delta_c = 0$, in the presence of a defect in the PBG, $\omega_d = -10$, $\gamma = 0.1$, $\Gamma_d = 3$, $\varepsilon = 10^{-3}$. The oscillator strength is now roughly equally shared between the distribution above cutoff (same as in **Figure 2b**) and the defect peak at ω_0.

$$
\begin{aligned}
|\Psi(z_l, t)\rangle = {} & \alpha_1 |1, \{0_\omega\}\rangle + \alpha_2 |2, \{0_\omega\}\rangle \\
& + \int \beta_{3,\omega} |3, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega \\
& + \int \beta_{4,\omega} |4, 1_\omega\rangle \rho(\omega) \mathrm{d}\omega
\end{aligned} \qquad [15]
$$

The consecutive terms in eqn [15] denote the atom being in states $|1\rangle$, $|2\rangle$, $|3\rangle$, or $|4\rangle$, with zero $\{0_\omega\}$ or one $|1_\omega\rangle$ photon in mode ω whose DOM is $\rho(\omega)$, and α_1, α_2, $\beta_{3,\omega}$, or $\beta_{4,\omega}$ are the corresponding probability amplitudes. Upon making the weak-field linear-response approximation, we can set $\alpha_1 \simeq 1$ and solve the set of equations for the slowly varying (compared to an optical cycle) probability amplitudes α and β, using the perturbation theory. Under these conditions, we effectively obtain a free-space propagation of the E_b field. By contrast, the evolution of the E_a field in the slowly varying envelope approximation, is given by $E_a(z, t) = E_a(0, t - z/v_g)\exp(ipz)$. We thus see that the real part of the macroscopic complex polarizability p is responsible for the phase shift ϕ_a of the E_a field, $\phi_a = \mathrm{Re}(p)z$, while the probability of the absorption $\mathscr{A}$ of the field depends on the imaginary part of p, $\mathscr{A} = 1 - \exp[-2\mathrm{Im}(p)z]$. The polarizability is expressed by

$$
p = a_0 \frac{i\gamma_2/2}{\gamma_2/2 - i\Delta_a + I(\Delta_a)} \qquad [16]
$$

where $a_0 \equiv \sigma_0 N$ is the linear resonant absorption coefficient on the atomic transition $|1\rangle \to |2\rangle$, with σ_0 the resonant absorption cross-section and N the density of doping atoms, γ_2 is the radiative width of

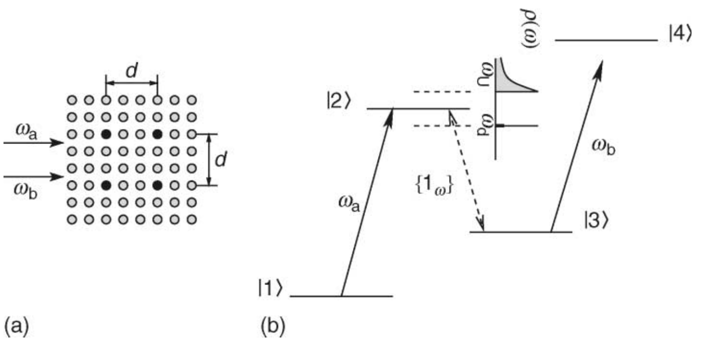

Figure 5 (a) Photonic crystal dilutely doped with atoms located at black dots. (b) Four-level atom coupled to a structured continuum $\rho(\omega)$ near the band-edge or defect mode frequencies (DOM plotted) via the intermediate transition $|2\rangle \to |3\rangle$ and interacting with two weak fields E_a and E_b at the sideband transitions $|1\rangle \to |2\rangle$ and $|3\rangle \to |4\rangle$, respectively.

state $|2\rangle$, $\Delta_a = \omega_a - \omega_{21}$ is the detuning from the atomic resonance ω_{21}, and $I(\Delta_a)$ is the integral of the saturation factor over the structured DOM. The group velocity v_g is expressed as $v_g = \partial_{k_a}\omega_a = [n_a/c + \partial_{\omega_a}\mathrm{Re}(p)]^{-1}$, where n_a is the (averaged) refraction index at the frequency ω_a.

To calculate $I(\Delta_a)$, we assume the isotropic PBG model, eqn [13] with $\varepsilon \to 0$, with the atoms doped at the positions of the local defects in the PC separated by a distance d from each other. These defect modes in the PBG are localized around each atomic site in a volume $V_d \simeq (rL)^3$ of several $(r)^3$ lattice cells L^3, with $L \simeq \pi c/\omega_{23}$, and serve as effective high-Q cavities, eqn [14] with $\Gamma_d \ll 1$. Assuming $|\Delta_b| \gg |\Delta|, |\Delta_a|, \gamma_4$, where $\Delta_b = \omega_b - \omega_{43}$ and $\Delta = \omega - \omega_{23}$, the integration leads to

$$I(\Delta_a) = \frac{B_d^2}{\gamma_{31} - i(\Delta_a - \Delta_d - s_3)} - \frac{B_U^{3/2}}{\sqrt{i\gamma_{31} + (\Delta_a - \Delta_U - s_3)}} \qquad [17]$$

where $\Delta_{d,U} = \omega_{d,U} - \omega_{23} \ll \omega_{23}$ are the detunings of the defect-mode and PBG-edge frequencies from the atomic resonance ω_{23}, γ_{31} the $|1\rangle \leftrightarrow |3\rangle$ decoherence rate, $s_3 = (\mu_{34}^2/\hbar^2\Delta_b)|E_b|^2$ is the E_b field-induced ac Stark shift of level $|3\rangle$ (μ_{ij} is the atomic dipole matrix element on the transition $|i\rangle \to |j\rangle$), and B_d and B_U are the coupling constants of the atom with the structured reservoir, whose main contributions are near ω_d and ω_U.

To illustrate the results of the foregoing analysis, we plot in Figure 6 the imaginary and real parts of the polarizability (eqn [16]). Consider first the case of one incident field $E_a (E_b = 0)$. Clearly, two frequency regions, $\Delta_a \sim \Delta_d$ and $\Delta_a \sim \Delta_U$, where the absorption vanishes and, at the same time, the dispersion slope is steep, are of particular interest. One can see that there is, however, a substantial difference between the spectra in the foregoing frequency regions, for the following physical reasons. First, in the vicinity of Δ_d, the atom interacts with the defect mode as in a high-Q cavity. This strong interaction 'dresses' the atomic states $|2\rangle$ and $|3\rangle$, thereby splitting the spectrum around $\Delta_a \simeq \Delta_d$ by the amount equal roughly to $2B_d$. Near the two-photon Raman resonance $\Delta_a = \Delta_d$, the two alternative transition paths $|1\rangle \to |2\rangle$ (the direct transition) and $|1\rangle \to |2\rangle \to |3\rangle \to |2\rangle$ (transition via state $|3\rangle$) interfere distractively with each other, cancelling thus the absorption of the E_a field and the medium becomes transparent to the radiation. This effect has been widely studied in atomic vapors, where the transition $|2\rangle \to |3\rangle$ is strongly driven by a coherent laser field, and is called electromagnetically induced transparency (EIT). The transparency window is rather broad and is given by the inverse Lorentzian (see the first term on the right-hand side of eqn [17]). Due to the steepness of the dispersion curve, the corresponding group velocity is much smaller than the speed of light, $v_g \simeq 2B_d^2/(\gamma_2 a_0) \ll c$, which leads to a large delay time $T_{\mathrm{del}} = \zeta/v_g$ at the exit $z = \zeta$ from the medium. One has to keep in mind, however, that the absorption-free propagation time is limited by the EIT decoherence time $T_{\mathrm{del}} < \gamma_{31}^{-1}$, which imposes a limitation on the length ζ of the active PC medium. Second, in the vicinity of Δ_U, the strong interaction of the atom with the continuum near the bandedge ω_U causes the Autler–Townes splitting of level $|2\rangle$ into a doublet with a separation equal roughly to B_U. One component of this doublet is shifted out of the PBG, while the other one remains

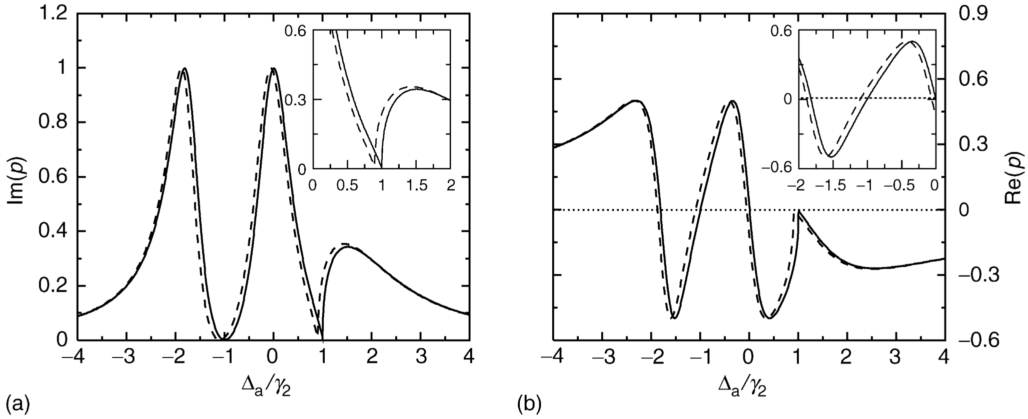

Figure 6 (a) Imaginary and (b) real part of the complex polarizability p as a function of the detuning Δ_a for the case $E_b = 0$ (solid lines) and $E_b \neq 0$ (dashed lines). The parameters (normalized by γ_2) are: $\Delta_d = -1$, $\Delta_U = 1$, $\gamma_{31} = 0.001$, $B_d = B_U = 1$, $s_3 = -0.1$, and $a_0 = 1\ \mathrm{cm}^{-1}$. The insets magnify the important frequency regions.

within the gap and forms the photon–atom bound state. Consequently, there is vanishing absorption and rapid variation of the dispersion at $\Delta_a \simeq \Delta_U$. Since the transparency region is very narrow with a width $\delta\omega \sim \gamma_{31}(\ll \gamma_2, B_U)$, for an absorption-free propagation of the E_a pulse, its temporal width τ_a should satisfy the condition $\tau_a > \pi/\delta\omega$. Simultaneously, a small deviation from the condition $\Delta_a = \Delta_U$ will lead to a strong increase in the absorption of the E_a field.

Let us now switch on the E_b field. As seen from eqn [17], its effect is merely to shift the spectrum by the amount equal to s_3 (Figure 6). This shift, however, will have different implications in the two frequency regions distinguished above: if $s_3 \ll B_d$, i.e., the Stark shift is smaller than the width of the EIT window at Δ_d, the medium will still remain transparent for an E_a field with the detuning $\Delta_a = \Delta_d$, but its phase will experience an appreciable nonlinear shift ϕ_a, given by $\phi_a = Re(p)z \simeq -\gamma_2 a_0 s_3 z/(2B_d^2)$. On the other hand, for an E_a field with the detuning $\Delta_a = \Delta_U$, the medium, which is transparent for $E_b = 0$, $\mathrm{Im}(p) \ll a_0$, will become highly absorptive (opaque) even for such a small frequency shift as s_3 (provided $s_3 < 0$ and $|s_3| > \Delta\omega$), $\mathrm{Im}(p) \simeq \gamma_2 a_0 \sqrt{|s_3|}/(2B_U^{3/2})$ and thus acting as an ultrasensitive, effective switch.

The remaining question is how to maximize the interaction between the E_a pulse, which propagates with a small group velocity, and the E_b pulse, which propagates with a velocity close to the speed of light. The interaction between the fields is maximized if: they enter the medium simultaneously; the transverse shapes of their envelopes overlap completely; and the pulse length l_b of the E_b field satisfies the condition $(l_b + \zeta)/c \leq \zeta/v_g$. Then the E_b pulse leaves the medium not later than the E_a pulse. The effective interaction length between the two pulses is, therefore, $z_{\mathrm{eff}} \sim l_b v_g/c \leq \zeta$, after which the two pulses slip apart. Thus, the presence of the E_b (control) field induces either strong absorption or a large phase shift of the E_a (signal) field, depending on the frequency region employed.

The effects surveyed above reveal unusual features of spontaneous emission and photon–atom binding in PCs, atomic interaction with the field of a high-Q defect mode, and nonlinear coupling of two fields via four-level dopant atoms in PCs. These effects are of fundamental interest. In addition, they can serve as the basis for highly efficient optical communications and data processing, in either the classical or the quantum domain, by providing two key elements: ultrasensitive nonlinear phase-shifters and photon switches.

List of Units and Nomenclature

Atomic or photonic eigenstates [dimensionless]	$\|\ldots\rangle$
Complex polarizability [cm^{-1}]	p
Coupling constant [s^{-1}]	κ, $C^{2/3}$, B
Decay rate [s^{-1}]	γ
Density of modes [s]	ρ
Frequency [s^{-1}]	ω, Δ, s
Phase [dimensionless]	ϕ
Probability amplitudes [dimensionless]	α, β, c
Reservoir spectral response [s^{-1}]	G
Resonant absorption coefficient [cm^{-1}]	a_0
Spontaneous emission spectrum [s]	W
Wave vector [m^{-1}]	k

See also

Electromagnetically Induced Transparency. **Photonic Crystals:** Photonic Crystal Lasers, Cavities and Waveguides.

Further Reading

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

John S and Wang J (1991) Quantum optics of localized light in a photonic band gap. *Physics Review B* 43: 12772–12789.

Kofman AG, Kurizki G and Sherman B (1994) Spontaneous and induced atomic decay in photonic band structures. *Journal of Modern Optics* 41: 353–384.

Lambropoulos P, Nikolopoulos GM, Nielsen TR and Bay S (2000) Fundamental quantum optics in structured reservoirs. *Reports on Progress in Physics* 63: 455–503.

Loudon R (1983) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Meystre P and Sargent M III (1991) *Elements of Quantum Optics*. Berlin: Springer Verlag.

Petrosyan D and Kurizki G (2001) Photon–photon correlations and entanglement in doped photonic crystals. *Physics Review A* 64: 023810-1-6.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Springer Series in Optical Sciences, 80. Berlin: Springer Verlag.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Electromagnetic Theory

S G Johnson and J D Joannopoulos,
Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals are periodically structured electromagnetic media, generally possessing photonic bandgaps: ranges of frequency in which light cannot propagate through the structure. This periodicity, whose lengthscale is proportional to the wavelength of light in the bandgap, is the electromagnetic analog of a crystalline atomic lattice, where the latter acts on the electron wavefunction to produce the familiar band gaps, semiconductors, etc., of solid-state physics. The study of photonic crystals is likewise governed by the Bloch–Floquet theorem, and intentionally introduced defects in the crystal (analogous to electronic dopants) give rise to localized electromagnetic states: linear waveguides and point-like cavities. The crystal can thus form a kind of perfect optical 'insulator', which can confine light around sharp bends, in lower-index media, and within wavelength-scale cavities, among other novel possibilities for control of electromagnetic phenomena. Below is introduced the basic theoretical background of photonic crystals in one, two, and three dimensions (schematically depicted in Figure 1), as well as hybrid structures that combine photonic-crystal effects in some directions with more-conventional index guiding in other directions. (Line and point defects in photonic crystals are discussed in another article.)

Electromagnetic wave propagation in periodic media was first studied by Lord Rayleigh in 1887, in connection with the peculiar reflective properties of a crystalline mineral with periodic 'twinning' planes (across which the dielectric tensor undergoes a mirror flip). These correspond to one-dimensional photonic crystals, and he identified the fact that they have a narrow bandgap prohibiting light propagation through the planes. This bandgap is angle-dependent, due to the differing periodicities experienced by light propagating at non-normal incidences, producing a reflected color that varies sharply with angle. (A similar effect is responsible for many other iridescent colors in nature, such as those of butterfly wings and abalone shells.) Although multilayer films received intensive study over the following century, it was not until 100 years later, when Yablonovitch and John, in 1987, joined the tools of classical electromagnetism and solid-state physics, that the concepts of omnidirectional photonic bandgaps in two and three dimensions was introduced. This generalization, which inspired the name 'photonic crystal', led to many subsequent developments in their fabrication, theory, and application, from integrated optics to negative refraction to optical fibers that guide light in air.

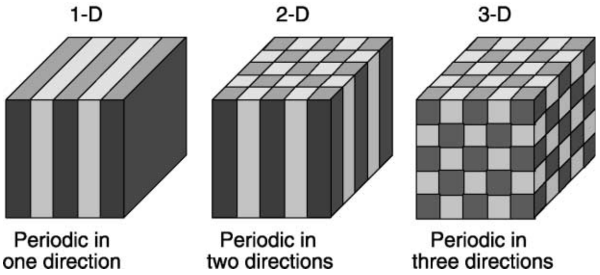
1-D
2-D
3-D
Periodic in one direction
Periodic in two directions
Periodic in three directions

Figure 1 Schematic depiction of photonic crystals periodic in one, two, and three directions, where the periodicity is in the material (typically dielectric) structure of the crystal. Only a 3d periodicity, with a more complex topology than is shown here, can support an omnidirectional photonic bandgap.

Maxwell's Equations in Periodic Media

The study of wave propagation in three-dimensionally periodic media was pioneered by Felix Bloch in 1928, unknowingly extending an 1883 theorem in one dimension by Gaston Floquet. Bloch proved that waves in such a medium can propagate without scattering, their behavior described by a periodic envelope function multiplied by a planewave. Although Bloch studied quantum mechanics, leading to the surprising result that electrons in a conductor scatter only from imperfections and not from the periodic ions, the same techniques can be applied to electromagnetism by casting Maxwell's equations as an eigenproblem in analog with Schrödinger's equation. By combining the source-free Faraday's and Ampere's laws at a fixed (angular) frequency ω, i.e., time dependence $e^{-i\omega t}$, one can obtain an equation in only the magnetic field $\vec{H}$:

$$\vec{\nabla} \times \frac{1}{\varepsilon} \vec{\nabla} \times \vec{H} = \left(\frac{\omega}{c}\right)^2 \vec{H} \qquad [1]$$

where ε is the dielectric function $\varepsilon(x, y, z)$ and c is the speed of light. This is an eigenvalue equation, with eigenvalue $(\omega/c)^2$ and an eigen-operator $\vec{\nabla} \times (1/\varepsilon) \vec{\nabla} \times$ that is Hermitian (acts the same to the left and right) under the inner product $\int \vec{H}^* \cdot \vec{H}'$ between two fields $\vec{H}$ and $\vec{H}'$. (The two curls correspond roughly to the 'kinetic energy' and $1/\varepsilon$ to the 'potential' compared to the Schrödinger Hamiltonian $\nabla^2 + V$.) It is sometimes more convenient to write a generalized Hermitian eigenproblem in the electric field $\vec{E}$, $\vec{\nabla} \times \vec{\nabla} \times \vec{E} = (\omega/c)^2 \varepsilon \vec{E}$, which separates the kinetic and potential terms. Electric fields that lie in higher ε, i.e., lower potential, will have lower ω; this is discussed further in the context of the variational theorem of eqn [3].

Thus, the same linear-algebraic theorems as those in quantum mechanics can be applied to the electromagnetic wave solutions. The fact that the eigen-operator is Hermitian and positive-definite (for real $\varepsilon > 0$) implies that the eigenfrequencies ω are real, for example, and also leads to orthogonality, variational formulations, and perturbation–theory relations that are discussed further below. An important difference compared to quantum mechanics is that there is a transversality constraint: one typically excludes $\vec{\nabla} \cdot \vec{H} \neq 0$ (or $\vec{\nabla} \cdot \varepsilon \vec{E} \neq 0$) eigensolutions, which lie at $\omega = 0$; i.e., static-field solutions with free magnetic (or electric) charge are forbidden.

Bloch Waves and Brillouin Zones

A photonic crystal corresponds to a periodic dielectric function $\varepsilon(\vec{x}) = \varepsilon(\vec{x} + \vec{R}_i)$ for some primitive lattice vectors $\vec{R}_i$ ($i = 1, 2, 3$ for a crystal periodic in all three dimensions). In this case, the Bloch–Floquet theorem for periodic eigenproblems states that the solutions to eqn [1] can be chosen of the form $\vec{H}(\vec{x}) = e^{i\vec{k}\cdot\vec{x}} \vec{H}_{n,\vec{k}}(\vec{x})$ with eigenvalues $\omega_n(\vec{k})$, where $\vec{H}_{n,\vec{k}}$ is a periodic envelope function satisfying

$$\left(\vec{\nabla} + i\vec{k}\right) \times \frac{1}{\varepsilon} \left(\vec{\nabla} + i\vec{k}\right) \times \vec{H}_{n,\vec{k}} = \frac{\omega_n(\vec{k})}{c}^2 \vec{H}_{n,\vec{k}} \qquad [2]$$

yielding a different Hermitian eigenproblem over the primitive cell of the lattice at each Bloch wavevector $\vec{k}$. This primitive cell is a finite domain if the structure is periodic in all directions, leading to discrete eigenvalues labeled by $n = 1, 2, \ldots$. These eigenvalues $\omega_n(\vec{k})$ are continuous functions of $\vec{k}$, forming discrete 'bands' when plotted versus the latter, in a 'band structure' or dispersion diagram – both ω and $\vec{k}$ are conserved quantities, meaning that a band diagram maps out all possible interactions in the system. (Note also that $\vec{k}$ is not required to be real; complex $\vec{k}$ gives evanescent modes that can exponentially decay from the boundaries of a finite crystal, but which cannot exist in the bulk.)

Moreover, the eigensolutions are periodic functions of $\vec{k}$ as well: the solution at $\vec{k}$ is the same as the solution at $\vec{k} + \vec{G}_j$, where $\vec{G}_j$ is a primitive reciprocal lattice vector defined by $\vec{R}_i \cdot \vec{G}_j = 2\pi\delta_{i,j}$. Thanks to this periodicity, one need only compute the eigensolutions for $\vec{k}$ within the primitive cell of this reciprocal lattice – or, more conventionally, one considers the set of inequivalent wavevectors closest to the $\vec{k} = 0$ origin, a region called the first Brillouin zone. For example, in a one-dimensional system, where $R_1 = a$ for some periodicity a and $G_1 = 2\pi/a$, the first Brillouin zone is the region $k = -\pi/a \ldots \pi/a$; all other wavevectors are equivalent to some point in this zone under translation by a multiple of G_1. Furthermore, the first Brillouin zone may itself be redundant if the crystal possesses additional symmetries such as mirror planes; by eliminating these redundant regions, one obtains the irreducible Brillouin zone, a convex polyhedron that can be found tabulated for most crystalline structures. In the preceding one-dimensional example, since most systems will have time-reversal symmetry ($k \rightarrow -k$), the irreducible Brillouin zone would be $k = 0 \ldots \pi/a$.

The familar dispersion relations of uniform waveguides arise as a special case of the Bloch formalism: such translational symmetry corresponds to a period $a \rightarrow 0$. In this case, the Brillouin zone of the wavevector k (also called β) is unbounded, and the envelope function $\vec{H}_{n,k}$ is a function only of the transverse coordinates.

The Origin of the Photonic Bandgap

A complete photonic bandgap is a range of ω in which there are no propagating (real $\vec{k}$) solutions of Maxwell's eqn [2] for any $\vec{k}$, surrounded by propagating states above and below the gap. There are also incomplete gaps, which only exist over a subset of all possible wavevectors, polarizations, and/or symmetries. Both sorts of gaps are discussed in the subsequent sections, but in either case, their origins are the same and can be understood by examining the consequences of periodicity for a simple one-dimensional system.

Consider a one-dimensional system with uniform $\varepsilon = \bar{\varepsilon}$, which has planewave eigensolutions $\omega(k) = ck$, as depicted in Figure 2(left). This ε has trivial periodicity a for any $a \geq 0$, with $a = 0$ giving the usual unbounded dispersion relation. One is free, however, to label the states in terms of Bloch envelope functions and wavevectors for some $a \neq 0$, in which case the bands for $|k| > \pi/a$ are translated (folded) into the first Brillouin zone, as shown by the dashed lines in Figure 2(left). In particular, the $k = -\pi/a$ mode in this description now lies at an equivalent wavevector to the $k = \pi/a$ mode, and at the same frequency; this accidental degeneracy is an artifact of

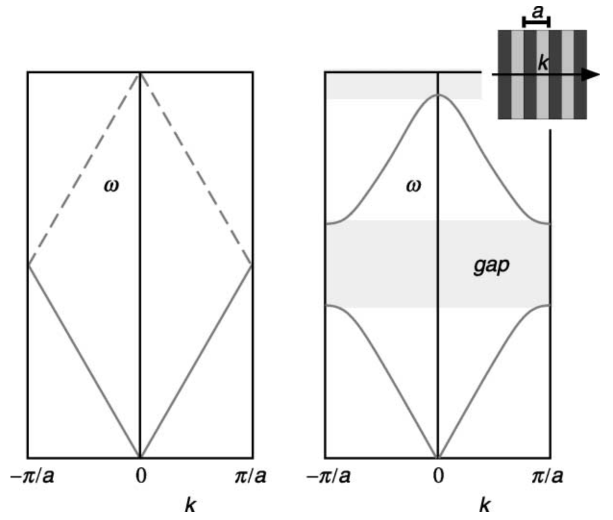

Figure 2 Left: Dispersion relation (band diagram), frequency ω versus wavenumber k, of a uniform one-dimensional medium, where the dashed lines show the 'folding' effect of applying Bloch's theorem with an artificial periodicity a. Right: Schematic effect on the bands of a physical periodic dielectric variation (inset), where a gap has been opened by splitting the degeneracy at the $k = \pm\pi/a$ Brillouin-zone boundaries (as well as a higher-order gap at $k = 0$).

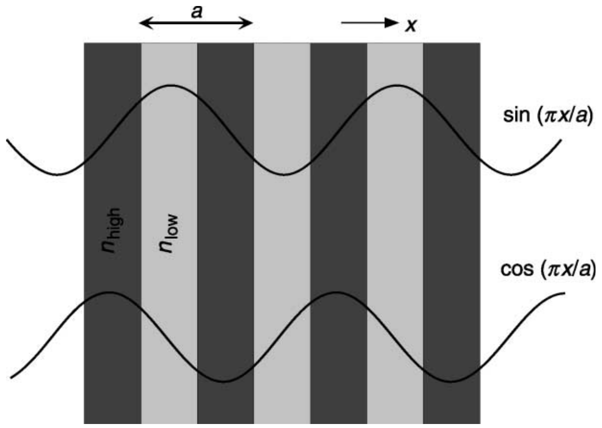

Figure 3 Schematic origin of the band gap in one dimension. The degenerate $ka = \pm\pi/a$ planewaves of a uniform medium are split into $\cos(\pi x/a)$ and $\sin(\pi x/a)$ standing waves by a dielectric periodicity, forming the lower and upper edges of the bandgap, respectively – the former has electric-field peaks in the high dielectric (n_{high}) and so will lie at a lower frequency than the latter (which peaks in the low dielectric).

the 'artificial' period that has been chosen. Instead of writing these wave solutions with electric fields $\vec{E}(x) \sim e^{\pm i\pi x/a}$, one can equivalently write linear combinations $e(x) = \cos(\pi x/a)$ and $o(x) = \sin(\pi x/a)$ as shown in Figure 3, both at $\omega = c\pi/a\sqrt{\varepsilon}$. Now, however, suppose that one perturbs ε so that it is nontrivially periodic with period a; for example, a sinusoid $\varepsilon(x) = \bar{\varepsilon}\cdot[1 + \Delta\cdot\cos(2\pi x/a)]$, or a square wave as in the inset of Figure 2(right). In the presence of such an oscillating 'potential', the accidental degeneracy between $e(x)$ and $o(x)$ is broken: supposing $\Delta > 0$, then the field $e(x)$ is more concentrated in the higher-ε regions than $o(x)$, and so lies at a lower frequency. This opposite shifting of the bands away from the mid-gap frequency $\omega \cong c\pi/a\sqrt{\bar{\varepsilon}}$ creates a bandgap, as depicted in Figure 2(right). (In fact, from the perturbation theory described subsequently, one can show that for $\Delta \ll 1$ the bandgap, as a fraction of mid-gap frequency, is $\Delta\omega/\omega \cong \Delta/2$.) By the same arguments, it follows that any periodic dielectric variation in one dimension will lead to a bandgap, albeit a small gap for a small variation; a similar result was identified by Lord Rayleigh in 1887.

More generally, it follows immediately from the properties of Hermitian eigensystems that the eigenvalues minimize a variational problem:

$$\omega^2_{n,\vec{k}} = \min_{\vec{E}_{n,\vec{k}}} \frac{\int \left|\left(\vec{\nabla} + i\vec{k}\right) \times \vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2} c^2 \qquad [3]$$

in terms of the periodic electric field envelope $\vec{E}_{n,\vec{k}}$, where the numerator minimizes the 'kinetic energy' and the denominator minimizes the 'potential energy' Here, the $n > 1$ bands are additionally constrained to be orthogonal to the lower bands:

$$\int \vec{H}^*_{m,\vec{k}} \cdot \vec{H}_{n,\vec{k}} = \int \varepsilon \vec{E}^*_{m,\vec{k}} \cdot \vec{E}_{n,\vec{k}} = 0 \qquad [4]$$

for $m < n$. Thus, at each $\vec{k}$, there will be a gap between the lower 'dielectric' bands concentrated in the high dielectric (low potential) and the upper 'air' bands that are less concentrated in the high dielectric: the air bands are forced out by the orthogonality condition, or otherwise must have fast oscillations that increase their kinetic energy. (The dielectric/air bands are analogous to the valence/conduction bands in a semiconductor.)

In order for a complete bandgap to arise in two or three dimensions, two additional hurdles must be overcome. First, although in each symmetry direction of the crystal (and each $\vec{k}$ point) there will be a bandgap by the one-dimensional argument, these bandgaps will not necessarily overlap in frequency (or even lie between the same bands). In order that they overlap, the gaps must be sufficiently large, which implies a minimum ε contrast (typically at least 4/1 in 3d). Since the 1d mid-gap frequency $\sim c\pi/a\sqrt{\bar{\varepsilon}}$ varies inversely with the period a, it is also helpful if the periodicity is nearly the same in different directions – thus, the largest gaps typically arise for hexagonal lattices in 2d and fcc lattices in 3d, which have the most nearly circular/spherical Brillouin zones. Second, one must take into account the vectorial boundary conditions on the electric field: moving across a dielectric boundary from ε to some $\varepsilon' < \varepsilon$, the inverse 'potential' $\varepsilon|\vec{E}|^2$ will decrease discontinuously if $\vec{E}$ is parallel to the interface ($\vec{E}_{\parallel}$ is continuous) and will increase discontinuously if $\vec{E}$ is perpendicular to the interface ($\varepsilon\vec{E}_{\perp}$ is continuous). This means that, whenever the electric field lines cross a dielectric boundary, it is much harder to strongly contain the field energy within the high dielectric, and the converse is true when the field lines are parallel to a boundary. Thus, in order to obtain a large bandgap, a dielectric structure should consist of thin, continuous veins/membranes along which the electric field lines can run – this way, the lowest band(s) can be strongly confined, while the upper bands are forced to a much higher frequency because the thin veins cannot support multiple modes (except for two orthogonal polarizations). The veins must also run in all directions, so that this confinement can occur for all $\vec{k}$ and polarizations, necessitating a complex topology in the crystal.

Ultimately, however, in two or three dimensions there are only rules of thumb for the existence of a bandgap in a periodic structure, since no rigorous criteria have yet been determined. This made the design of 3d photonic crystals a trial and error process, with the first example by Ho *et al.* of a complete 3d gap coming three years after the initial 1987 concept. As is discussed by the final section below, a small number of families of 3d photonic crystals have since been identified, with many variations thereof explored for fabrication.

Computational Techniques

Because photonic crystals are generally complex, high index-contrast, two- and three-dimensional vectorial systems, numerical computations are a crucial part of most theoretical analyses. Such computations typically fall into three categories: time-domain 'numerical experiments' that model the time-evolution of the fields with arbitrary starting conditions in a discretized system (e.g., finite-difference); definite-frequency transfer matrices wherein the scattering matrices are computed in some basis to extract transmission/reflection through the structure; and frequency-domain methods to directly extract the Bloch fields and frequencies by diagonalizing the eigenoperator. The first two categories intuitively correspond to directly measurable quantities such as transmission (although they can also be used to compute e.g., eigenvalues), whereas the third is more abstract, yielding the band diagrams that provide a guide to interpretation of measurements as well as a starting-point for device design and semi-analytical methods. Moreover, several band diagrams are included in the following sections, and so the frequency-domain method used to compute them is briefly outlined here.

Any frequency–domain method begins by expanding the fields in some complete basis, $\vec{H}_{\vec{k}}(\vec{x}) = \sum_n h_n \vec{b}_n(\vec{x})$, transforming the partial differential eqn [2] into a discrete matrix eigenvalue problem for the coefficients h_n. Truncating the basis to N elements leads to $N \times N$ matrices, which could be diagonalized in $O(N^3)$ time by standard methods. This is impractical for large 3d systems, however, and is also unnecessary – typically, one only wants the few lowest eigenfrequencies, in which case one can use iterative eigensolver methods requiring only $\sim O(N)$ time. Perhaps the simplest such method is based directly on the variational theorem eqn [3]: given some starting coefficients h_n, one iteratively minimizes the variational 'Rayleigh' quotient using e.g., preconditioned conjugate–gradient descent. This yields the lowest band's eigenvalue and field, and upper bands are found by the same minimization while orthogonalizing against the lower bands ('deflation'). There is one additional difficulty, however, and that is that one must at the same time enforce the $(\vec{\nabla} + i\vec{k}) \cdot \vec{H}_{\vec{k}} = 0$ transversality constraint, which is nontrivial in three dimensions. The simplest way to maintain this constraint is to employ a basis that is already transverse, for example planewaves $\vec{h}_{\vec{G}} e^{i\vec{G}\cdot\vec{x}}$ with transverse amplitudes $\vec{h}_{\vec{G}} \cdot (\vec{G} + \vec{k}) = 0$. (In such a planewave basis, the action of the eigen-operator can be computed via a fast Fourier transform in $O(N \log N)$ time.)

Semi-analytical Methods: Perturbation Theory

As in quantum mechanics, the eigenstates can be the starting point for many analytical and semi-analytical studies. One common technique is perturbation theory, applied to small deviations from an ideal system – closely related to the variational eqn [3], perturbation theory can be exploited to consider effects such as nonlinearities, material absorption, fabrication disorder, and external tunability. Not only is perturbation theory useful in its own right, but it also illustrates both old and peculiarly new features that arise in such analyses of electromagnetism compared to scalar problems such as quantum mechanics.

Given an unperturbed eigenfield $\vec{E}_{n,\vec{k}}$ for a structure ε, the lowest-order correction $\Delta\omega_n^{(1)}$ to the eigenfrequency from a small perturbation $\Delta\varepsilon$ is given by

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int \Delta\varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2} \qquad [5]$$

where the integral is over the primitive cell of the lattice. A Kerr nonlinearity would give $\Delta\varepsilon \sim |\vec{E}|^2$, material absorption would produce an imaginary frequency correction (decay coefficient) from a small imaginary $\Delta\varepsilon$, and so on. Similarly, one can compute the shift in frequency from a small $\Delta\vec{k}$ in order to determine the group velocity $\mathrm{d}\omega/\mathrm{d}k$; this variation of perturbation theory is also called $k \cdot p$ theory. All such first-order perturbation corrections are well known from quantum mechanics, and in the limit of infinitesimal perturbations give the exact Hellman–Feynman expression for the derivative of the eigenvalue. However, in the limit where $\Delta\varepsilon$ is a small shift Δh of a dielectric boundary between some ε_1 and ε_2, an important class of geometric perturbation, eqn [5] gives a surface integral of $|\vec{E}|^2$ on the interface, but this is ill-defined because the field there is discontinous. The proper derivation of perturbation theory in the face of such discontinuity requires a more careful limiting process from an anisotropically smoothed system, yielding the surface integral:

$$\Delta\omega_n^{(1)} = -\frac{\omega_n(\vec{k})}{2} \frac{\int\int \Delta h \left(\Delta\varepsilon_{12} \left|\vec{E}_{\|}\right|^2 - \Delta\varepsilon_{12}^{-1} \left|D_{\perp}\right|^2\right)}{\int \varepsilon \left|\vec{E}_{n,\vec{k}}\right|^2}$$

where $\Delta\varepsilon_{12} = \varepsilon_1 - \varepsilon_2, \Delta\varepsilon_{12}^{-1} = \varepsilon_1^{-1} - \varepsilon_2^{-1}$, and $\vec{E}_{\|}/D_{\perp}$ denotes the (continuous) interface parallel/perpendicular components of the unperturbed electric/displacement eigenfield. A similar expression is required in high index-contrast systems to employ, e.g., coupled-mode theory for slowly-varying waveguides or Green's functions for interface roughness.

Standard perturbation–theory techniques also provide expressions for higher-order corrections to the eigenvalue and eigenfield, based on an expansion in the basis of the unperturbed eigenfields. This approach, however, runs into immediate difficulty because the eigenfields are also subject to the transversality constraint, $(\vec{\nabla} + i\vec{k}) \cdot \varepsilon\vec{E}_{\vec{k}} = 0$, and this constraint varies with ε and $\vec{k}$ – the eigenfields $\vec{E}_{\vec{k}}$ are not a complete basis for the eigenfields constrained at a different ε or $\vec{k}$. For ε perturbations, this problem can be eliminated by using the $\vec{H}$ or $\vec{D}$ eigenproblems, whose constraints are independent of ε. For $\vec{k}$ perturbations, one can employ a transformation by Sipe to derive a corrected higher-order perturbation theory (for e.g., the group-velocity dispersion), based on the fact that all of the non-transverse fields lie at $\omega = 0$. Such completeness issues also arise applying the variational eqn [3], as was noted in the previous section: in order for a useful variational bound to apply, one must operate in the constrained (transverse) subspace.

Two-Dimensional Photonic Crystals

After the identification of one-dimensional bandgaps, it took a full century to add a second dimension, and three years to add the third. It should therefore come as no surprise that 2d systems exhibit most of the important characteristics of photonic crystals, from nontrivial Brillouin zones to topological sensitivity to a minimum index contrast, and can also be used to demonstrate most proposed photonic–crystal devices. The key to understanding photonic crystals in two dimensions is to realize that the fields in 2d can be divided into two polarizations by symmetry: *TM* (transverse magnetic), in which the magnetic field is in the (xy) plane and the electric field is perpendicular (z); and *TE* (transverse electric), in which the electric field is in the plane and the magnetic field is perpendicular.

Corresponding to the polarizations, there are two basic topologies for 2d photonic crystals, as depicted in Figure 4(top): high index rods surrounded by low index (top) and low-index holes in high index (bottom). Here, a hexagonal lattice is used because, as noted earlier, it gives the largest gaps. Recall that a photonic band gap requires that the electric field lines run along thin veins: thus, the rods are best suited to TM light ($\vec{E}$ parallel to the rods), and the holes are best suited to TE light ($\vec{E}$ running around the holes). This preference is reflected in the band diagrams, shown in Figure 4, in which the rods/holes (top/bottom) have a strong TM/TE band gap. For these diagrams, the rod/hole radius is chosen to be $0.2a/0.3a$, where a is the lattice constant (the nearest-neighbor periodicity) and the high/low ε is 12/1. The TM/TE bandgaps are then 47%/28% as a fraction of mid-gap frequency, but these bandgaps require a minimum ε contrast of 1.7/1 and 1.9/1, respectively. Moreover, it is conventional to give the frequencies ω in units of $2\pi c/a$, which is equivalent to a/λ (λ being the vacuum wavelength) – Maxwell's equations are scale-invariant, and the same solutions can be applied to any wavelength simply by choosing the appropriate a. For example, the TM mid-gap ω in these units is 0.36, so if one wanted this to correspond to $\lambda = 1.55\ \mu\mathrm{m}$ one would use $a = 0.36 \cdot 1.55\ \mu\mathrm{m} = 0.56\ \mu\mathrm{m}$.

The Brillouin zone (a hexagon) is shown at left-center, with the irreducible Brillouin zone shaded (following the sixfold symmetry of the crystal); the corners (high symmetry points) of this zone are given canonical names, where Γ always denotes the origin $\vec{k} = 0$, K is the nearest-neighbor direction,

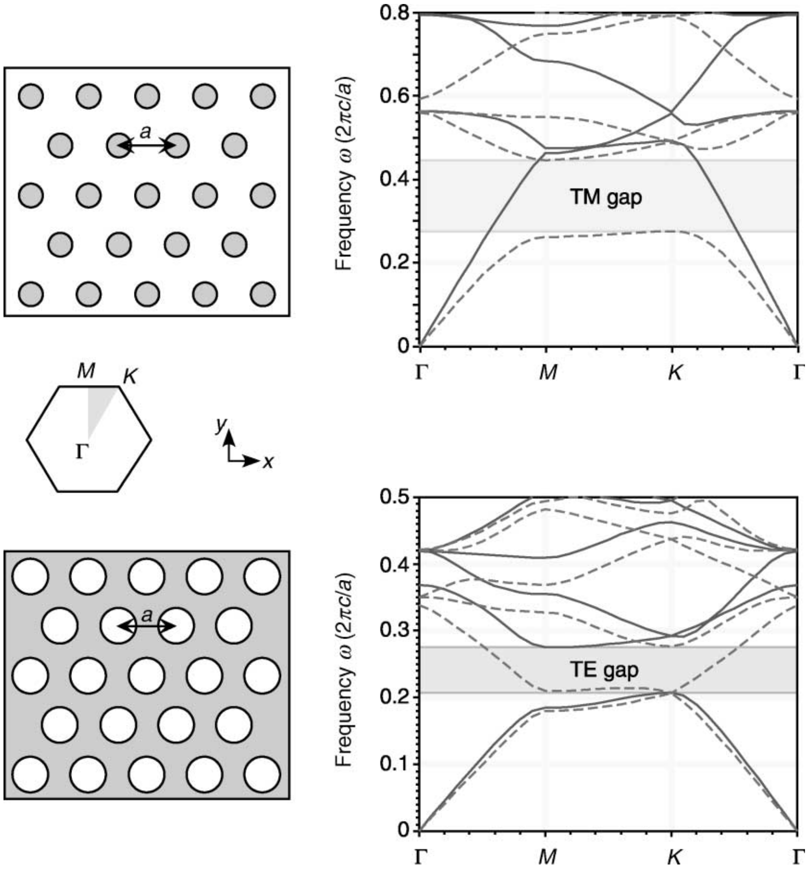

Figure 4 Band diagrams and photonic band gaps for hexagonal lattices of high dielectric rods ($\varepsilon = 12$, $r = 0.2a$) in air (top), and air holes ($r = 0.3a$) in dielectric (bottom), where a is the center–center periodicity. The frequencies are plotted around the boundary of the irreducible Brillouin zone (shaded triangle, left center), with solid/dashed lines denoting TE/TM polarization (electric field parallel/perpendicular to plane of periodicity). The rods/holes have a gap in the TM/TE bands.

and M is the next-nearest-neighbor direction. The Brillouin zone is a two-dimensional region of wavevectors, so the bands $\omega_n(\vec{k})$ are actually surfaces, but in practice the band extrema almost always occur along the boundaries of the irreducible zone (i.e., the high-symmetry directions). So, it is conventional to plot the bands only along these zone boundaries in order to identify the bandgap, as is done in Figure 4.

Actually, the hole lattice can display not only a TE gap, but a complete photonic bandgap (for both polarizations) if the holes are sufficiently large (nearly touching). In this case, the thin veins between nearest-neighbor holes induce a TE gap, while the interstices between triplets of holes form 'rod-like' regions that support a TM gap overlapping the TE gap.

Photonic-Crystal Slabs

In order to realize 2d photonic-crystal phenomena in three dimensions, the most straightforward design is to simply fabricate a 2d-periodic crystal with a finite height: a photonic-crystal slab, as depicted in Figure 5. Such a structure can confine light vertically within the slab via index guiding, a generalization of total internal reflection – this mechanism is the source of several new tradeoffs and behaviors of slab systems compared to their 2d analogs.

The key to index guiding is the fact that the 2d periodicity implies that the 2d Bloch wavevector $\vec{k}_\parallel$ is a conserved quantity, so the projected band structure – all states in the bulk substrate/superstrate (the uniform regions far below/above the slab) versus their in-plane wavevector component (projected wavevector) – creates a map of which states can radiate vertically. If the slab is suspended in air, for example, then the eigensolutions of the bulk air are $\omega = c\sqrt{|\vec{k}_\parallel|^2 + k_\perp^2}$, which when plotted versus $\vec{k}_\parallel$ forms the continuous light cone $\omega \geq c|\vec{k}_\parallel|$, shown as a shaded region in Figure 5. Because the slab has a higher ε (12) than the air (1), and frequency goes as $1/\sqrt{\varepsilon}$, discrete guided bands are 'pulled down' in frequency from this continuum – these bands, lying beneath the light cone, cannot couple to any vertically

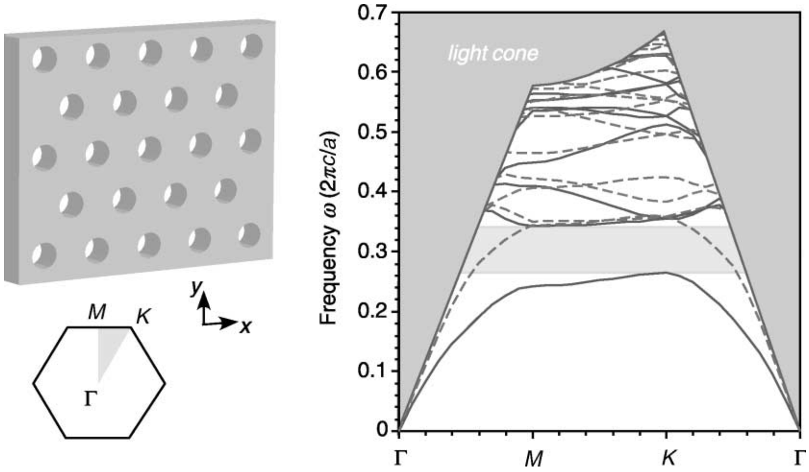

Figure 5 Projected band diagram for a finite-thickness (0.5*a*) slab of air holes in dielectric (cross section as in **Figure 4** bottom), with the irreducible Brillouin zone at lower left. The shaded region is the light cone: the projection of all states that can radiate in the air. Solid/dashed lines denote guided modes (confined to the slab) that are even/odd with respect to the horizontal mirror plane of the slab, whose polarization is TE-like/TM-like, respectively. There is a 'bandgap' (region without guided modes) in the TE-like guided modes only.

radiating mode by the conservation law and so are confined to the slab (exponentially decaying away from it). If the horizontal mid-plane of the slab is a mirror symmetry plane, then just as there were TM and TE states in 2d, here there are two categories of modes: even (TE-like) and odd (TM-like) modes under reflections through the mirror plane (which are purely TE/TM in the mirror plane itself). Because the slab here is based on the 2d hole crystal, which had a TE gap, here there is a 26% 'bandgap' in the even modes: a range of frequencies in which there are no guided modes. It is not a complete photonic bandgap, not only because of the odd modes, but also because there are radiating (light cone) modes at every ω. The presence of these radiating modes means that if all in-plane translational symmetry is broken by a localized change in the structure, say a waveguide bend or a resonant cavity, then vertical radiation losses are inevitable; there are various strategies to minimize the losses to tolerable levels, however. On the other hand, if only one direction of translational symmetry is broken, as in a linear-defect waveguide, ideally lossless guiding can be maintained.

Photonic-crystal slabs have two new critical parameters that influence the existence of a gap. First, it must have vertical mirror symmetry in order that the gaps in the even and odd modes be treatable separately – such symmetry is broken by the presence of a substrate that is not the mirror image of the superstrate, but in practice the symmetry breaking can be weak if the index contrast is sufficiently high (so that the modes are strongly confined in the slab). Second, the height of the slab must not be too small (or the modes will be weakly confined) or too large (or higher-order modes will fill the gap); the optimum height is around half a wavelength λ/n_{eff} (relative to an average/effective index n_{eff} that depends on the polarization). In Figure 5, a height of 0.5a is used, which is near the optimum (with holes of radius 0.3a and $\varepsilon = 12$ as in the previous section).

Three-Dimensional Photonic Crystals

Photonic-crystal slabs are one way of realizing 2d photonic-crystal effects in three dimensions; an example of another way, lifting the sacrifices imposed by the light cone, is depicted in Figure 6. This is a 3d-periodic crystal, formed by an alternating hole-slab/rod-slab sequence in an ABCABC stacking of bilayers – equivalently, it is an fcc (face-centered cubic) lattice of air cylinders in dielectric, stacked and oriented in the 111 direction, where each overlapping layer of cylinders forms a rod/hole bilayer simultaneously. Its band diagram is shown in Figure 6 along the boundaries of its irreducible Brillouin zone (from a truncated octahedron, inset), and this structure has a >21% complete gap ($\Delta\omega$ as a fraction of mid-gap frequency) for $\varepsilon = 12/1$, forbidding light propagation for all wavevectors (directions) and all polarizations. Not only can this crystal confine light perfectly in 3d, but because its layers resemble 2d rod/hole crystals, it turns out that the confined modes created by defects in these layers strongly resemble the TM/TE states created by corresponding defects in two dimensions. One can therefore use this crystal to directly transfer designs from two to three dimensions while retaining omnidirectional confinement. Its fabrication, of

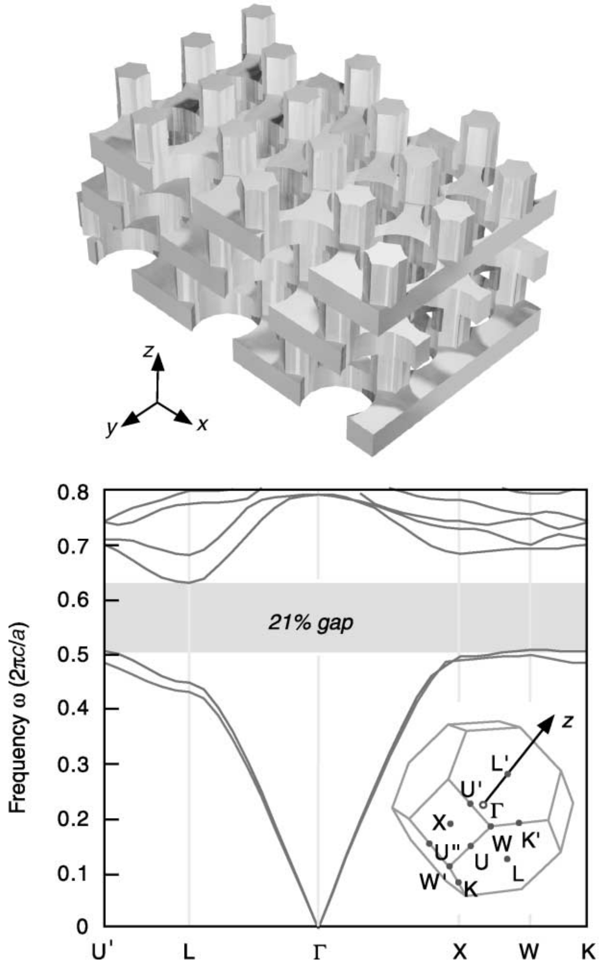

Figure 6 Band diagram (bottom) for 3d-periodic photonic crystal (top) consisting of an alternating stack of rod and hole 2d-periodic slabs (similar to **Figure 4**), with the corners of the irreducible Brillouin zone labeled in the inset. This structure exhibits a $\Delta\omega/\omega_{\mathrm{midgap}} = 21\%$ omnidirectional bandgap.

course, is more complex than that of photonic-crystal slabs (with a minimum ε contrast of 4/1), but this and other 3d photonic crystal structures have been constructed even at micron (infrared) lengthscales, as described below.

There are three general dielectric topologies that have been identified to support complete 3d gaps for $\varepsilon = 12/1$ (e.g., Si:air) contrast: diamond-like arrangements of high dielectric 'atoms' surrounded by low dielectric, which can lead to $>20\%$ gaps between the 2nd and 3rd bands; fcc 'inverse opal' lattices of nearly-touching low dielectric spheres (or similar) surrounded by high dielectric, giving gaps around 10% between the 9th and 10th bands; and cubic 'scaffold' lattices of rods along the cube edges, giving $\sim 7\%$ gaps between the 2nd and 3rd bands. It is notable that the first two topologies correspond to fcc lattices, which have the most nearly spherical Brillouin zones in accordance with the rules of thumb given above. Many variations on these topologies continue to be proposed – for example, the structure of Figure 6 is diamond/graphite-like – mainly in conjunction with different fabrication strategies, such as the following three successful approaches. First, layer-by-layer fabrication, in which individual crystal layers (typically of constant cross-section) are deposited one-by-one and etched with a 2d pattern via standard lithographic methods (giving fine control over placement of defects, etc.); Figure 6 can be constructed in this fashion (as well as other diamond-like structures with large gaps). Second, colloidal self-assembly, in which small dielectric spheres in a fluid automatically arrange themselves into close-packed (fcc) crystals by surface forces – these crystals can be back-filled with a high-index material, out of which the original spheres are dissolved in order to form inverse–opal crystals with a complete gap. Third, holographic lithography, in which a variety of 3d crystals can be formed by an interference pattern of four laser beams to harden a light-sensitive resin (which is then back-filled and dissolved, as with colloids, to achieve the requisite index contrast). The second and third techniques are notable for their ability to construct large-scale 3d crystals (thousands of periods) in a short time.

See also

Photonic Crystals: Atomic Physics; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Philadelphia: Holt Saunders.

Cohen-Tannoudji C, Din B and Laloë F (1977) *Quantum Mechanics*. Paris: Hermann.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Boston: Kluwer.

Johnson SG, Ibanescu M, Skorobogatiy M, Weisberg O, Joannopoulos JD and Fink Y (2002) Perturbation theory for Maxwell's equations with shifting material boundaries. *Physical Review E* 65: 066611.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. Berlin: Springer.

Sipe JE (2000) Vector $k \cdot p$ approach for photonic band structures. *Physical Review E* 62: 5672–5677.

Microwave Photonic Crystals

D F Sievenpiper, HRL Laboratories LLC, Malibu, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Photonic Crystals

Photonic crystals are materials that have a periodically varying dielectric constant that results in a bandgap, or range of frequencies where electromagnetic waves are unable to propagate. A simple one-dimensional example is the Bragg mirror, or quarter-wave stack, in which periodic layers of alternating dielectrics reflect photons within a frequency band determined by the thickness and refractive index of the layers. Photonic crystals are the extension of this concept into two or three dimensions. They are constructed as a lattice of two different dielectrics, and examples include a triangular lattice of circular rods, or a diamond lattice of spheres. Certain geometries will produce a bandgap, within which propagation is impossible in any direction. These materials are the photonic analog of semiconductors, in that their periodic dielectric function provides a bandgap for photons in much the same way that the periodic atomic potential of a crystal lattice can provide a bandgap for electrons.

While most photonic crystal research has focused on optical wavelength structures, many of the first experimental studies were performed on scale models at microwave frequencies, where fabrication and measurements were simpler. Operating at microwave frequencies also allowed a new class of materials to emerge, based on metals and lumped circuit elements. The nearly infinite dielectric contrast provided by the use of metals, and the ability to make complex subwavelength scale structures at radio frequencies, allows photonic crystals with ultra-wide bandgaps to be produced. Furthermore, the use of lumped circuit elements allows the electromagnetic properties to be easily tuned, resulting in unique microwave structures.

Wire Mesh Photonic Crystals

At optical frequencies, conductive losses preclude the use of metals for many applications. However, at microwave frequencies, metals are nearly ideal materials, since they have low loss and are easily formed into complicated geometries. The simplest metallic photonic crystal is a three-dimensional wire mesh, shown in Figure 1. This example resembles a diamond crystal lattice, in which metal wires have been substituted for atomic bonds. It can be constructed using zig-zag-shaped metal strips that fit together to form a continuous conductive network.

The electromagnetic properties of this wire mesh can be understood by considering a traditional microwave analog: an array of metal waveguides. Open tunnels permeate the mesh in various directions, and it behaves as an interpenetrating fabric of waveguides, whose walls are defined by the wires. Shorter wavelengths, that fit between the wires, can propagate through the crystal, while longer waves see the mesh as a continuous metal surface, and are reflected. As for a conventional waveguide, the cutoff frequency occurs approximately where one half wavelength fits between the rows of wires. Similarly, for a diamond lattice of wires, the cutoff frequency is

$$\omega_{\text{cutoff}} = \frac{\pi c}{na} \qquad [1]$$

where n is the refractive index of the material between the wires, and a is the length of the unit cube of the diamond lattice. Because the cutoff frequency is primarily determined by the diameter of the open tunnels, and it is largely independent of the detailed arrangement of the wires, other mesh geometries can

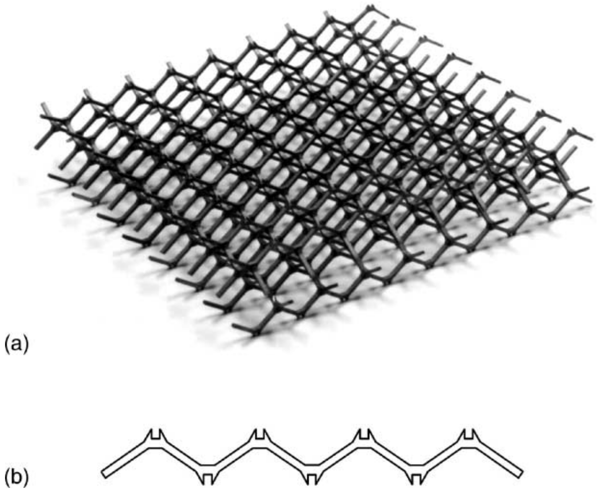

Figure 1 A wire mesh photonic crystal with diamond geometry. (a) The mesh forms continuous conductive structure that is permeated with open tunnels. (b) A diagram of one of the interlocking metal parts that fit together to form the crystal. (Reproduced with permission from Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480. Copyright (1996) by the American Physical Society.)

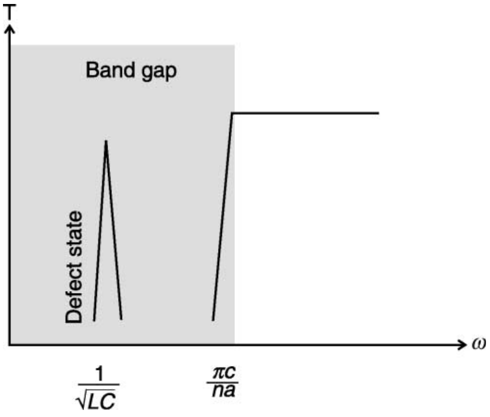

Figure 2 The transmission magnitude through a conductive wire mesh photonic crystal as a function of frequency, ω. Shorter wavelengths can propagate through the open tunnels, while longer wavelengths see the mesh as a continuous block of metal. The cutoff frequency is determined largely by the lattice constant, *a*, and the refractive index, *n*, of the material between the wires. A cut wire forms a defect state within the band gap.

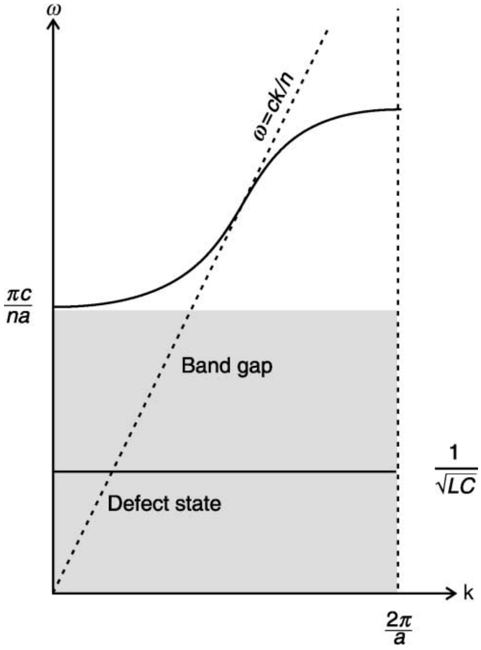

Figure 3 The dispersion diagram for a wire mesh photonic crystal, plotted as frequency, ω, versus wave vector k. The bandgap extends to zero frequency. The lower bandedge of the allowed band is determined largely by the lattice constant. Like other photonic crystals, the wire mesh also supports higher frequency bands, which are not shown. A point defect such as a cut wire creates a mode at a single frequency, which contains all wave vectors.

be expected to give similar results. However, the cutoff frequency can vary by a constant of order unity, depending on the lattice and the unit cell chosen.

The transmission magnitude, as a function of frequency for such a mesh, is depicted in Figure 2. Experimentally, it is found that a typical wire mesh structure attenuates roughly 6 dB per layer for frequencies below cutoff. However, because the exact transmission magnitude above cutoff depends on details of the measurement, such as impedance matching at the crystal boundary, the vertical scale shown here is arbitrary, and the intent is simply to give the qualitative behavior of these and similar crystal structures. Waves just below the cutoff frequency can penetrate a crystal with finite thickness, as depicted by the finite slope of the transmission edge.

We can better understand the behavior of waves within the crystal by examining the dispersion diagram, shown in Figure 3, in which we plot the allowed frequencies as a function of wave vector $k = 2\pi/\lambda$, the spatial frequency of waves inside the material. A band of allowed frequencies is bounded by upper and lower band edges, corresponding to high and low cutoff frequencies, beyond which propagation is forbidden. The reflective nature of the crystal below cutoff is shown as a bandgap that extends down to zero frequency. The slope of the dispersion curve corresponds to the group velocity of waves within the allowed band, and states at the bandedges represent standing waves. The upper edge of this band occurs where the wavelength matches the periodicity of the crystal, and for a diamond mesh, it is approximately

$$\omega_{\text{lattice}} = \frac{2\pi c}{na} \qquad [2]$$

Like traditional dielectric photonic crystals, these wire mesh structures also have higher order bands and gaps that occur at higher frequencies. The nature of the higher band structure depends on the detailed geometry of the mesh to a much greater extent than the lowest order gap, which is our primary interest for radio frequency materials.

Defect States

As in semiconductor crystals, we can alter the behavior of photonic crystals by adding defects to create localized states within the bandgap. For the wire mesh photonic crystal, a defect may be a cut wire, or a lumped circuit element inserted into the mesh. For example, replacing one of the wires with a capacitor creates a parallel resonant circuit, which acts as a small electromagnetic cavity within the mesh. Its resonance frequency is given by

$$\omega_{\text{defect}} = \frac{1}{\sqrt{LC}} \qquad [3]$$

where C is the capacitance of the defect, and L is the inductance of the surrounding wires. The inductance depends logarithmically on the wire diameter, so it is largely fixed for a given mesh lattice. However, the capacitance can be easily changed, since it is a simple lumped circuit element. Thus, by adding capacitive defects we can create tunable local states within the bandgap. The presence of a defect is seen as a narrow transmission band within the bandgap, shown in Figure 2. Because a point defect contains all wave vectors, it is shown as a horizontal line in the dispersion diagram in Figure 3.

If two defects are located in close proximity, their modes overlap, and they interact with each other. For example, two neighboring defects with identical capacitance will exhibit a split pair of resonances, corresponding to a lower frequency symmetric mode, and a higher frequency anti-symmetric mode. The modes are highly localized at the defects, and this interaction is limited to defects that are within one or two lattice constants of each other. If one member of a pair of nearby defects is tuned with respect to the other, the frequencies the two modes exhibit are an avoided crossing behavior, as is typical in many systems of coupled oscillators. The resonance frequencies for more complex arrangements of defects can be calculated by assigning an interaction energy to each neighboring pair, and solving the resulting Hamiltonian matrix to find the energy levels of the complete system. In this way, the interaction among defects in small clusters is analogous to atoms in simple molecules, and the resulting energy level structure is similar.

As more defects are added periodically throughout the mesh, they form a superlattice, and their states merge into an additional band within the original bandgap. The upper frequency of this band depends on the LC resonance frequency of the defects. The lower frequency depends on the density of the remaining wires, since waves at the lower bandedge must be able to fit through the open regions in the mesh. In the extreme case, where every wire is cut and replaced by a capacitor, the mesh is no longer a continuous conductive network. These structures have an allowed band that extends to zero frequency, and they will be explored in greater detail in the next section. Not every wire must be cut to reach this limit, and the conditions for this to occur are the subject of percolation theory.

Capacitive Mesh Photonic Crystals

In a capacitive mesh photonic crystal, small isolated metal islands are distributed on a lattice, and they are capacitively coupled to their neighbors. This represents the extreme case of the wire mesh photonic crystal, where every wire has been cut, and replaced by a capacitor. An example of such a structure, built as a stack of printed circuit boards, is shown in Figure 4. A diamond lattice can be approximated using planar circuit boards, with printed tetrapodal structures consisting of two orthogonal metal strips connected by a metal plated via. Each metallic island couples to its four nearest neighbors through their overlapping capacitor plates.

The capacitive mesh photonic crystal is transmissive at low frequencies because it is not traversed by conducting paths, like the continuous wire mesh version. The crystal supports a lower band that extends from zero frequency, up to an LC resonance frequency:

$$\omega_{\text{lower}} = \frac{1}{\sqrt{LC}} \tag{4}$$

which marks the lower edge of the bandgap. This is determined by the capacitance between the metal islands, and the inductance associated with each island. At higher frequencies, the capacitors appear as shorts, and the crystal behaves similarly to the conductive wire mesh structure described previously, with an upper edge given by

$$\omega_{\text{upper}} = \frac{\pi c}{na} \tag{5}$$

At this frequency, the waves can again fit through the open areas between the metallic islands. The transmission magnitude, as a function of frequency for the capacitive mesh photonic crystal, is shown in Figure 5, and the additional band is illustrated in the dispersion diagram in Figure 6. Like the conductive wire mesh, the capacitive mesh also has a rich upper band structure, that is outside the scope of this article.

With these capacitive mesh structures, it is possible to create materials with finite, but arbitrarily large bandgaps, because the upper and lower edges of the bandgap are controlled independently by the lattice constant and the capacitance, respectively. In conventional dielectric photonic crystals, the width of the bandgap depends on the dielectric contrast between the constituent materials. Waves at the upper edge primarily reside in the lower dielectric material, and waves at the lower edge inhabit the higher dielectric material. This reasoning can also be applied to capacitive mesh structures. The lower dielectric is the substrate material that fills the open spaces between the metal islands, and modes at the upper edge of the gap primarily fill these open spaces. However, for modes at the lower edge of the gap, the electric field is primarily concentrated within the capacitors. The metal regions and the capacitors can

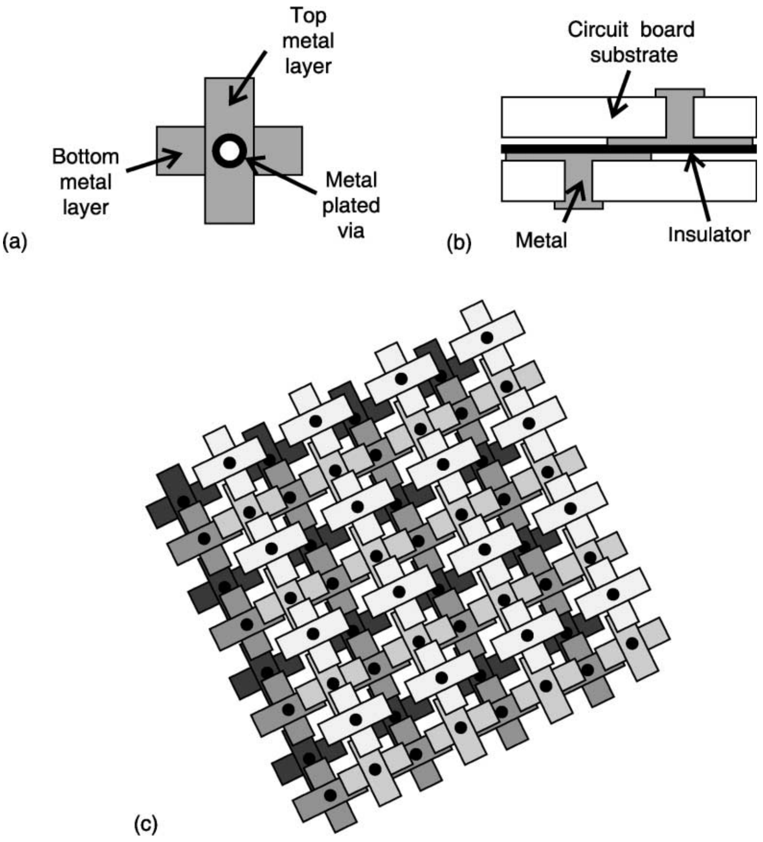

Figure 4 A capacitive mesh photonic crystal, built as a diamond lattice of capacitively coupled metallic islands. It can be built by stacking printed circuit boards. (a) One 'atom' of the crystal, which is made of two small orthogonal metal strips connected by a metal-plated via. (b) A side view, showing how the metal strips overlap to form a capacitor. (c) Several periods of the completed crystal structure, shown without the substrate. (Reproduced with permission from Sievenpiper D, Yablonovitch E, Winn J, et al. (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. Physics Review Lett. 80: 2829. Copyright (1998) by the American Physical Society.)

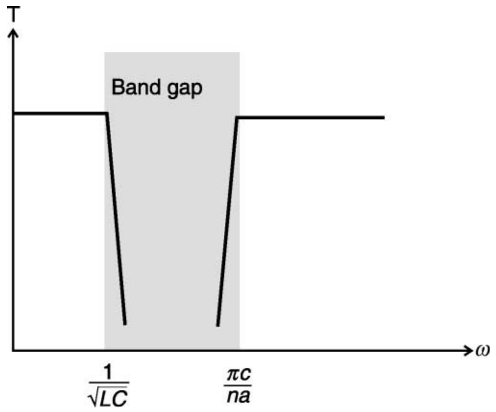

Figure 5 The transmission magnitude through a capacitive mesh photonic crystal, as a function of frequency. The crystal is transparent at low frequencies, because the conducting paths are broken by capacitors. At higher frequencies, the capacitors appear as shorts, and crystal behaves like a solid wire mesh.

be considered as an effective medium, or artificial dielectric whose properties depend on the density, arrangement, and values of the capacitors. By using large capacitors, the dielectric contrast can be made arbitrarily large, and the bandgap arbitrarily wide.

Within the bandgap, these materials can support surface waves, which are modes that are bound to the interface between the crystal and free space. Surface waves propagate along the crystal surface, and they have their own band structure. In the following sections, we examine these waves, and we will introduce periodically textured surfaces that suppress them, much like three-dimensional photonic crystals suppress internal plane waves.

Surface Waves

Electromagnetic waves that are bound to the interface between two materials are known as surface waves. They exist on ordinary metals, and at optical frequencies they are often called surface plasmons. They also exist on photonic crystals within the

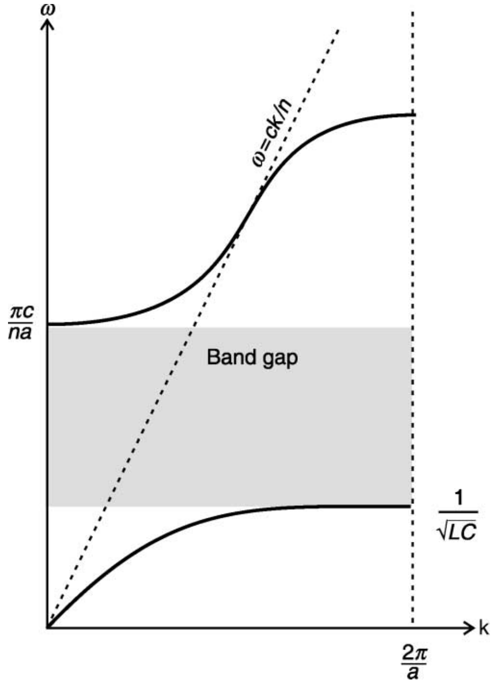

Figure 6 The dispersion diagram of the capacitive mesh photonic crystal. The two edges of the bandgap can be adjusted independently by varying the lattice constant and the capacitance.

bandgap region, as well as on other impenetrable materials. They cannot propagate into the crystal because of Bragg reflection, yet they are prevented from escaping into free space by total internal reflection. They typically interact strongly with the top layer of the crystal, and they can often be manipulated or suppressed by adjusting the geometry of that layer. At radio and microwave frequencies, these waves are usually called surface currents, because the associated fields are only loosely bound to the surface, and extend a great distance into free space. Their presence can be seen in antenna radiation patterns and other phenomena, and by suppressing them we can make a variety of unique radio frequency devices. However, we must first understand the nature of these waves.

To understand the nature of these surface currents, we begin by examining surface waves on metals, where we assume a wave confined to a dielectric interface, and then insert the complex dielectric constant of a metal. Consider a surface in the YZ plane, where $X < 0$ is filled with dielectric ε, and $X > 0$ is filled with vacuum. A wave propagates in the Z direction with propagation constant k. For transverse magnetic (TM) polarized waves, which are illustrated in Figure 7, the transverse electric field $E_y = 0$. The electric field has the form

$$E_1 = (\hat{x}E_{1x} + \hat{z}E_{1z})e^{i\omega x - ikz - \alpha x} \qquad [6]$$

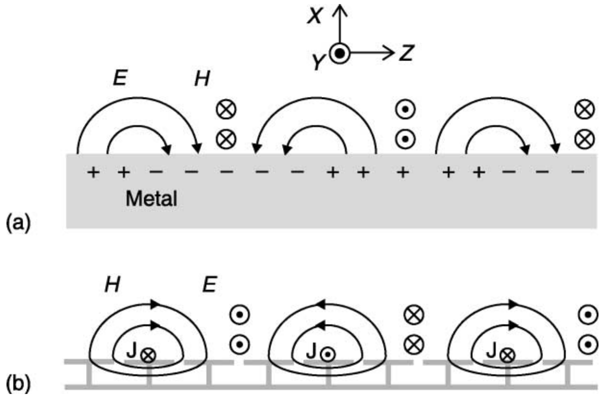

Figure 7 Surface waves can be classified into two polarizations. (a) A TM wave, shown on a metal surface; (b) A TE wave, shown on a textured surface. TM waves consist of electric charges on the metal surface that oscillate along the direction of propagation. They have nonzero electric field components both perpendicular to the surface, and parallel to the direction of propagation. TE waves are associated with transverse currents in the surface. They are similar to TM waves, but with the electric and magnetic fields exchanged.

for $X > 0$ and

$$E_2 = (\hat{x}E_{2x} + \hat{z}E_{2z})e^{i\omega x - ikz + \gamma x} \qquad [7]$$

for $X < 0$, where α and γ are the attenuation constants into free space and dielectric, respectively. The fields of such a wave are shown in Figure 7. We can solve the wave equation for the electric field:

$$\nabla \times \nabla \times \vec{E} = \varepsilon_r \frac{\omega^2}{c^2}\frac{\partial^2 \vec{E}}{\partial t^2} \qquad [8]$$

by matching E_x and $\varepsilon_r E_z$, at the interface to yield

$$k = \sqrt{\frac{\varepsilon}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [9]$$

$$\alpha = \sqrt{\frac{-1}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [10]$$

$$\gamma = \sqrt{\frac{-\varepsilon^2}{1+\varepsilon}}\,\frac{\omega}{c} \qquad [11]$$

A solution for transverse electric (TE) waves, where $H_y = 0$, can be found by applying the principle of duality, by exchanging the electric and magnetic fields, and substituting μ for ε.

For surface waves on a metal surface, we insert the dielectric constant of a metal, which is

$$\varepsilon = 1 - \frac{i\sigma}{\omega\varepsilon_0} \qquad [12]$$

The conductivity, λ, is

$$\sigma = \frac{nq^2\tau/m}{1 + i\omega\tau} \qquad [13]$$

where τ is the mean electron collision time, q is the electron charge, m and n are the effective mass and the density, respectively, of the conduction electrons. For frequencies much lower than $1/\tau$, which includes microwave and radio frequencies, we can substitute eqns [12] and [13] into eqn [9] to find

$$k \approx \frac{\omega}{c} \qquad [14]$$

Thus, surface waves on metals propagate at nearly the speed of light in vacuum, and they travel for many wavelengths with little attenuation. We can also solve for the decay constants into free space using eqn [10]:

$$\alpha = (1 - i)\frac{\omega}{c}\sqrt{\frac{\omega\varepsilon_0}{2\sigma}} \qquad [15]$$

and into the metal using eqn [11]:

$$\gamma \approx (1 + i)\sqrt{\frac{\omega\mu_0\sigma}{2}} = \frac{(1 + i)}{\delta} \qquad [16]$$

where δ is recognized as the skin depth. Thus, we have derived the skin depth of a metal from the penetration depth of surface waves.

From the skin depth, we can derive the impedance of a metal sheet. Assuming that E_0 is the electric field at the surface, the surface current is given by

$$J_z(x) = \sigma E_z(x) = \sigma E_0 e^{-x(1+i)/\delta} \qquad [17]$$

The magnetic field at the surface is found by integrating along a path that surrounds the surface current:

$$H_0 = \overset{-\infty}{\underset{0}{\delta}} J_z(x)\mathrm{d}x = \frac{\sigma\delta}{1 + i}E_0 \qquad [18]$$

The surface impedance of the metal is therefore

$$Z_s = \frac{E_z}{H_y} = \frac{1 + i}{\sigma\delta} \qquad [19]$$

We can see that the resistance of a metal surface is accompanied by an equal amount of inductance, because the real and imaginary parts of the surface impedance are equal, and have the same sign.

The part of the wave that extends into vacuum is very loosely bound to the metal surface, but the currents in the metal occur in a well-defined skin depth. For example, copper has a surface impedance of $0.03(1 + i)$ ohms/square at 10 GHz. Using eqns [15] and [16], a surface wave at that frequency extends about 70 meters, or 2300 wavelengths into free space, but less than 1 micron into the metal. For this reason, surface waves are often described simply as surface currents at microwave and radio frequencies, although they are the same phenomenon as the more tightly bound surface plasmons that occur at optical frequencies. Nonetheless, their presence can be seen in a variety of phenomena at RF frequencies, such as antenna radiation patterns.

The above discussion can be generalized to a simple impedance surface, and it can be shown that TM waves are supported on inductive surfaces having an impedance of

$$Z_s(\mathrm{TM}) = \frac{i\alpha}{\omega\varepsilon} \qquad [20]$$

while TE waves are supported on capacitive surfaces, having an impedance of

$$Z_s(\mathrm{TE}) = \frac{-i\omega\mu}{\alpha} \qquad [21]$$

Thus, while ordinary metals, which are inductive, support TM waves, they do not support TE waves, which require a capacitive surface.

Surface waves on metals can be suppressed over a finite frequency band by applying various kinds of periotic textures. For example, at optical frequencies, a lattice of small bumps on a metal surface can create a surface wave bandgap. Surface waves are scattered by the bumps, and form standing waves when one-half wavelength fits between the rows of bumps. Two modes occur, in which the wave nodes lie either on the bumps or between them. These two modes have slightly different frequencies, separated by a narrow bandgap.

At microwave frequencies, more elaborate structures can be built to control the surface properties. One example is a corrugated metal surface, shown in Figure 8. The corrugations can be considered as transmission lines running into the slab, and shorted at the bottom. They are typically one-quarter wavelength deep, so that the short circuit at the bottom is transformed into an open circuit at the top. Because of these resonant structures built into the surface, it has high impedance for waves that are polarized with the electric field transverse to the corrugations, but it has low impedance for waves that are polarized in the orthogonal direction, with the electric field parallel to the metal ridges. For waves propagating transverse to or parallel to the corrugations, the surface can be described as hard or soft, respectively, by analogy to the acoustic surfaces of the same names. Both of these boundary conditions are used in antennas to control scattering properties or

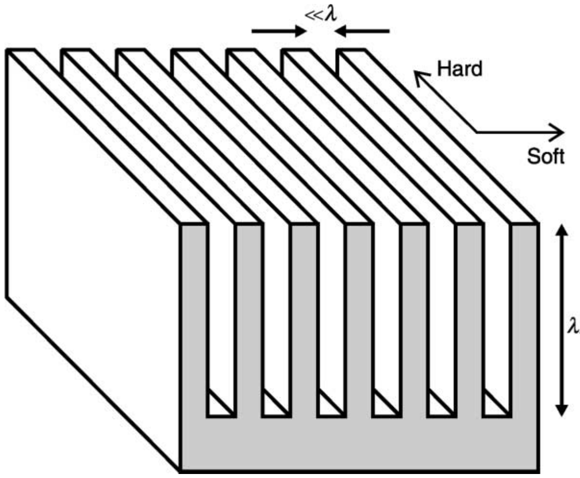

Figure 8 A corrugated metal surface containing quarter-wave resonant grooves presents a soft boundary condition to waves propagating transverse to the grooves, and a hard boundary condition to waves propagating along the grooves.

Table 1 Four basic boundary conditions in electromagnetics, and the surface impedance for each polarization

Boundary condition	*Z_s(TM)*	*Z_s(TE)*
Electric conductor	0	0
Soft surface	∞	0
Hard surface	0	∞
Magnetic conductor	∞	∞

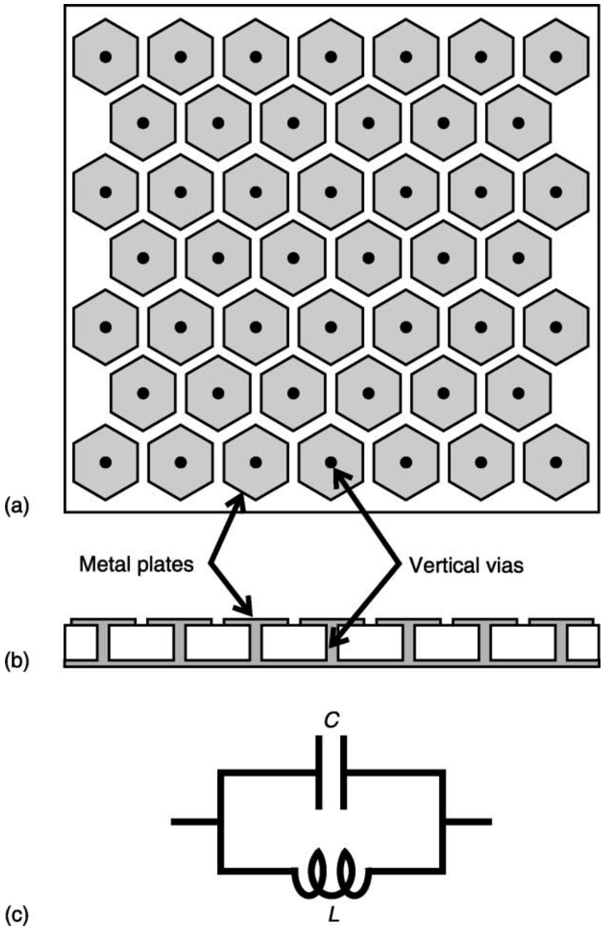

Figure 9 A high-impedance surface, constructed as a printed circuit board. (a) The top view shows an array of hexagonal metal patches; (b) the side view shows metal-plated vias that connect the patches to the ground plane; (c) the effective surface impedance can be modeled as a parallel resonant LC circuit.

field profiles. These two surfaces, along with the electric and magnetic conductors, describe four basic boundary conditions for electromagnetics, which are summarized in Table 1. Like the soft and hard surfaces, the magnetic conductor does not exist in nature, but it can be produced artificially over a finite frequency band, using a kind of two-dimensional photonic crystal known as a high-impedance surface.

High-Impedance Surfaces

Just as three-dimensional photonic crystals forbid internal plane waves, two-dimensional photonic crystals can be designed to suppress surface waves. Like the capacitive mesh described earlier, they can be constructed with lumped circuit elements. They can be understood as corrugated surfaces, in which the resonant corrugations have been folded up into resonant LC circuits, and distributed on a two-dimensional lattice. An example of such a structure is shown in Figure 9. It can be built using printed circuit boards, in which flat metal patches are connected to a common ground plane by metal plated vias.

When the plates are small compared to the wavelength, we may treat the surface as an effective medium, with a surface impedance defined by a parallel resonant LC circuit. The proximity of the neighboring metal plates provides capacitance, while the current paths, through the vias and the ground plane, provide inductance. The surface impedance is given by

$$Z_s = \frac{i\omega L}{1 - \omega^2 LC} \qquad [22]$$

and the resonance frequency is

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad [23]$$

The sheet capacitance, C, is equal to the value of the individual capacitors, multiplied by a geometrical factor that is related to their arrangement. It can be shown that for nonmagnetic materials, the sheet inductance L is equal to

$$L = \mu_0 t \qquad [24]$$

where t is the total thickness of the high-impedance surface. The surface is inductive for frequencies below resonance, and capacitive for frequencies above resonance. Near the resonance frequency, the magnitude of the surface impedance is very high ($Z_s >> \eta$, where $\eta = 377\Omega$ is the impedance of free space).

For a normally incident wave, the reflection coefficient of the surface is

$$R = \frac{Z_s - \eta}{Z_s + \eta} \qquad [25]$$

The phase $\phi = \mathrm{Im}(Ln(R))$ of the reflection coefficient is plotted as a function of frequency in Figure 10. While a low impedance surface such as a conventional electric conductor reflects with a phase shift of π, a high-impedance surface or artificial magnetic conductor reflects with a phase shift of 0. This can be used to build low-profile antennas, because radiating elements can lie very near to a high-impedance ground plane without being shorted out. For example, a thin wire antenna can be placed within a small fraction of a wavelength of the high-impedance surface, yet retain high radiation efficiency. While image currents in a metal ground plane would normally oppose the antenna current, the currents in a high-impedance ground plane reinforce the antenna current because of the opposite reflection phase. This effect occurs when the reflection phase falls between $\pi/2$ and $-\pi/2$, and the corresponding frequency range defines the practical bandwidth of a low-profile antenna on such a surface. Experimentally this bandwidth is found to be roughly equal to the surface wave bandgap, to be described later.

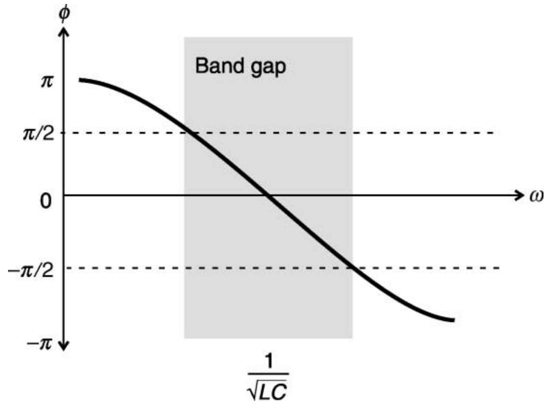

Figure 10 The reflection phase of a high-impedance surface crosses through 0 at the resonance frequency, where the surface appears as an artificial magnetic conductor. The reflection phase falls between $\pi/2$ and $-\pi/2$ for frequencies corresponding roughly to the edges of the surface wave bandgap.

The reflection phase properties of these materials can also be used to build unique microwave reflectors. Since the phase depends on the resonance frequency, one can vary the local resonance frequency in order to tune the reflection phase as a function of position on the surface. For example, a linear phase gradient can be used for beam steering, or a parabolic phase function can be used for focusing. The resonance frequency can be adjusted by tuning the capacitance of the individual unit cells, using varactors or mechanical motion, to build actively controlled microwave reflectors.

The other important property of these textured ground planes is their surface wave bandgap, which can be used to control the radiation pattern of antennas near the surface. We examine the surface wave properties by considering a wave that decays into free space, as in eqn [6]. From Maxwell's equations, we can obtain

$$k^2 = \mu_0 \varepsilon_0 \omega^2 + \alpha^2 \qquad [26]$$

which can be combined with eqs [20] and [21] to obtain the dispersion relations for TM and TE waves on a general impedance surface, respectively. For TM waves, which occur where $\omega < \omega_0$, we have

$$k(\mathrm{TM}) = \frac{\omega}{c}\sqrt{1 - \frac{Z_s^2}{\eta^2}} \qquad [27]$$

and for TE waves, which occur where $\omega > \omega_0$ we have

$$k(\mathrm{TE}) = \frac{\omega}{c}\sqrt{1 - \frac{\eta^2}{Z_s^2}} \qquad [28]$$

These functions describe, to first order, the dispersion of waves on a high-impedance surface. However, they do not predict the bandgap, which can only be obtained by including the effects of radiation damping.

Consider an oscillating sheet of current that sees the impedance of the textured surface on one side, and the impedance of free space on the other side. At very low, or very high frequencies, the current is shunted through the surface. Near the resonance frequency, where the surface impedance is high, power is delivered to free space through radiation damping. The half power bandwidth of this radiation is defined by the frequencies where the magnitude of the surface impedance is equal to the impedance of free space:

$$\left|\frac{i\omega L}{1-\omega^2 LC}\right| = \eta \qquad [29]$$

We solve for λ to yield the frequencies of the two bandedges:

$$\omega^2 = \frac{1}{LC} + \frac{1}{2\eta^2 C^2} \pm \frac{1}{\eta C}\sqrt{\frac{1}{LC} + \frac{1}{4\eta^2 C^2}} \qquad [30]$$

The terms in $1/\eta^2 C^2$ are typically small compared to the terms in $1/LC$, so we neglect them to obtain

$$\omega \approx \omega_0 \sqrt{1 \pm \frac{Z_0}{\eta}} \qquad [31]$$

where

$$Z_0 = \sqrt{\frac{L}{C}} \qquad [32]$$

is the characteristic impedance of the surface. For structures that are thin compared to the wavelength, the characteristic impedance is small compared to η. We can expand the square root in eqn [31], to approximate the fractional bandwidth of the radiation as

$$\frac{\Delta\omega}{\omega_0} \approx \frac{Z_0}{\eta} = \frac{\sqrt{L/C}}{\sqrt{\mu_0/\varepsilon_0}} \qquad [33]$$

Using eqn [23], this can be rearranged to obtain

$$\frac{\Delta\omega}{\omega_0} \approx t\frac{2\pi}{\lambda_0} \qquad [34]$$

where λ_0 is the free space wavelength at the resonance frequency. This defines a surface wave bandgap that is centered at the resonance frequency, between a lower-frequency TM band and a higher-frequency TE band. From eqn [34], we see that the bandwidth of a high-impedance surface, or of an antenna on that surface, is limited by its thickness with respect to its free space wavelength at resonance. The term $\lambda_0/2\pi$ is recognized as the radian length, described by Wheeler, who showed that the fractional bandwidth limit of a small resonant antenna is given by its volume with respect to a notional sphere of this radius. Equation [34] represents the equivalent bandwidth limit for a two-dimensional resonant surface.

The surface wave bandgap can be measured experimentally with a pair of small probes placed near the surface. A surface wave probe can be constructed from the end of a coaxial cable with a short section of the inner conductor exposed. Fields from the probe excite surface waves when they are placed near the surface, and the transmission magnitude between the probes can be used to analyze the surface wave band structure. For a high impedance surface, the bandgap can be seen as a range of frequencies with a significant reduction in surface wave transmission magnitude, shown in Figure 11. Within the bandgap, waves are not bound to the surface, and currents on the surface readily radiate energy into free space. The upper edge of the bandgap

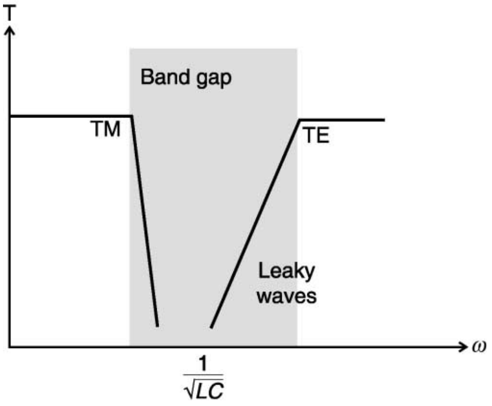

Figure 11 The surface wave transmission magnitude as a function of frequency for a high-impedance surface. TM waves are supported at low frequencies, and TE waves are supported at high frequencies. Leaky TE waves, which are damped by radiation, can occur within the bandgap.

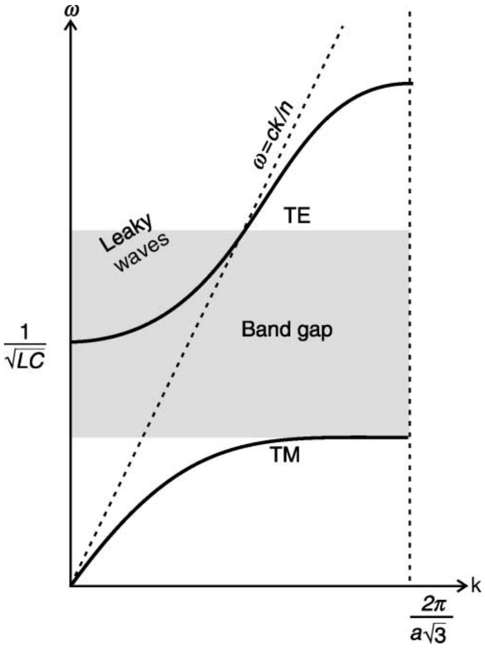

Figure 12 The dispersion diagram for surface waves on a high-impedance surface. The bandgap spans from the upper edge of the TM band, to the point where the TE band crosses the light line. Within the bandgap, electromagnetic waves are not bound to the surface. Modes to the left of the light line represent leaky waves.

is typically less abrupt than the lower edge, due to the presence of leaky TE waves.

The dispersion diagram for surface waves on a high-impedance surface is illustrated in Figure 12. The TM band begins at zero frequency, and bends over flat below the resonance frequency. The TE band begins at the resonance frequency, and slopes upward through the light line. The bandgap, which is the region where bound surface waves are not supported, extends from the upper edge of the TM band to the point where the TE band crosses the light line. This boundary, where $\omega = ck/n$, represents waves in free space propagating parallel to the surface. TE modes, that lie to the right of the light line, are bound to the surface while those that lie left of the light line lose energy to radiation damping. They are not bound to the surface, and are described as leaky waves. The approximate dispersion functions of eqns [27] and [28] are similar to the curves shown in Figure 12, but because they neglect radiation damping, they do not describe these leaky modes. Furthermore, like other photonic crystals, high-impedance surfaces also have higher-order bands and gaps that are not explained by the effective medium model.

The suppression of propagating surface modes can be used for a variety of applications, such as antennas and other microwave structures. One example is to change the radiation pattern of an antenna by covering its ground plane with a textured surface. Figure 13 shows a simple monopole antenna, along with the radiation patterns produced by the monopole when it is placed on a flat or textured ground plane. These plots show the relative magnitude of radiation from the antenna as a function of angle. TM surface currents on the finite metal ground plane form standing waves, which scatter from the edges and corners of a finite ground plane. This causes ripples in the radiation pattern as radiation from the edges interferes with the direct radiation from the antenna. Scattering of the surface currents from the ground plane edges also allows the antenna to radiate power into the backward hemisphere, which would otherwise be shadowed by the ground plane. On the high-impedance ground plane, the suppression of surface

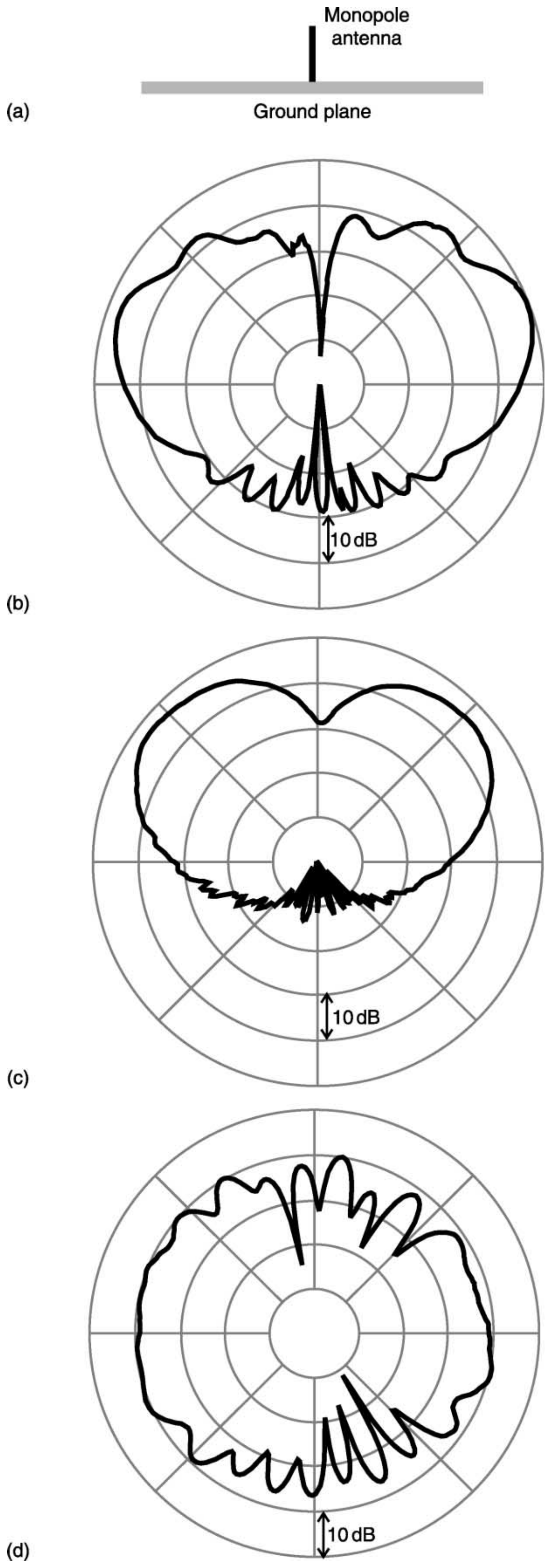

Figure 13 (a) A monopole antenna on a ground plane, used for testing the effect of a textured surface on the radiation pattern. (b) The E-plane radiation pattern of the monopole on a flat metal ground plane shows ripples due to standing surface waves. (c) On a high-impedance ground plane, within the bandgap, surface waves are forbidden, and the radiation pattern is smooth. (d) Outside the bandgap, in the TM band, there is a high density of surface modes.

currents within the bandgap results in a smooth radiation pattern in the forward hemisphere, and very little power in the backward hemisphere. For frequencies outside the bandgap, such as within the TM band, the high density of TM modes can be seen as fine ripples in the pattern, and significant power in the backward hemisphere.

The monopole is useful for illustrating the suppression of TM surface modes, but many other antenna designs are possible. For example, the suppression of both TM and TE surface waves can be used to build patch antennas or aperture antennas with a high degree of symmetry between the *E* and *H* planes. Because these materials behave as artificial magnetic conductors, they can also be used to design a variety of thin antennas that would be impractical on conventional metal ground planes. Finally, by incorporating active elements into the surface, tunable or steerable antennas are possible.

List of Units and Nomenclature

a	lattice constant	meters
c	speed of light in vacuum	meters/second
C	capacitance	Farads
E	electric field	Volts/meter
H	magnetic field	Amps/meter
i	unit imaginary number	–
J	current density	Amps/meter^2
k	wave vector	1/meters
L	inductance	Henrys
m	effective electron mass	kilograms
n	refractive index	–
n	electron density	number/meter^3
q	electron charge	Coulombs
r	radius	meters
R	reflection coefficient	–
t	thickness	meters
T	transmission coefficient	–
Z	surface impedance	Ohms/square
α	decay constant into free space	1/meters
γ	decay constant into material	1/meters
δ	skin depth	meters
ε	electric permittivity	Farads/meter
η	impedance of free space	Ohms
λ	wavelength	meters
μ	magnetic permeability	Henrys/meter
π	diameter/circumference ratio	–
σ	electrical conductivity	1/Ohms
τ	mean electron collision time	seconds
ω	angular frequency	radians/second
∇	vector differential operator	–
∂	scalar differential operator	–
∞	infinity	–

See also

Babinet's Principle. **Diffractive Systems**: Diffractives in Animals; Microstructure Fibers; Omnidirectional Surfaces and Fibers. **Photonic Crystals:** Atomic Physics; Electromagnetic Theory; Nonlinear Optics in Photonic Crystal Fibers; Photonic Crystal Lasers, Cavities and Waveguide; Self-Assembled and Functionalized Photonic Crystals.

Further Reading

Ashcroft N and Mermin N (1976) *Solid State Physics*. Orlando, FL: Saunders College Publishing.

Brillouin L (1953) *Wave Propagation in Periodic Structures: Electric Filters and Crystal Lattices*, 2nd edn. New York: Dover Publications.

Brown E, Parker C and Yablonovitch E (1993) Radiation properties of a planar antenna on a photonic-crystal substrate. *Journal of the Optical Society of America B* 10: 404.

Collin R (1991) *Field Theory of Guided Waves*, 2nd edn. New York: IEEE Press.

Fan S, Villeneuve P and Joannopoulos J (1996) Large omnidirectional band gaps in metallodielectric photonic crystals. *Physics Review B* 54: 11245.

Feynman R, Leighton R and Sands M (1964) *The Feynman Lectures on Physics*, vol. 3. Reading, MA: Addison-Wesley.

Joannopoulos J, Meade R and Winn J (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton University Press.

Kildal P-S (1990) Artificially soft and hard surfaces in electromagnetics. *IEEE Transactions on Antennas and Propagation* 38: 1537.

Kurizki G and Haus J (eds) (1994) Photonic band structures. *Journal of Modern Optics* 41, special issue.

Lee S, Zarrillo G and Law C (1982) Simple formulas for transmission through periodic metal grids or plates. *IEEE Transactions on Antennas and Propagation* 30: 904.

Pendry J, Holden A, Stewart W and Youngs I (1996) Extremely low frequency plasmons in metallic mesostructures. *Physics Review Letters* 76: 4773.

Ramo S, Whinnery J and Van Duzer T (1984) *Fields and Waves in Communication Electronics*, 2nd edn. New York: John Wiley and Sons.

Sievenpiper D (1999) *High-Impedance Electromagnetic Surfaces*. Ph.D. Thesis, Los Angeles: University of California.

Sievenpiper D, Sickmiller M and Yablonovitch E (1996) 3D wire mesh photonic crystals. *Physics Review Letters* 76: 2480.

Sievenpiper D, Yablonovitch E, Winn J, *et al.* (1998) 3D metallo-dielectric photonic crystals with strong capacitive coupling between metallic islands. *Physics Review Letters* 80: 2829.

Sigalas M, Chan C, Ho K-M and Soukoulis C (1995) Metallic photonic band-gap materials. *Physics Review B* 52: 11744.

Smith D, Schultz S, Kroll N, et al. (1994) Experimental and theoretical results for a two-dimensional metal photonic band-gap cavity. *Applied Physics Letters* 65: 645.

Soukoulis C (ed.) (1996) Photonic band gap materials. *Proceedings of the NATO ASI on Photonic Band Gap Materials, Elounda, Crete, Greece, June 18–30, 1995*. The Netherlands: Kluwer Academic Publishers.

Wheeler H (1947) Fundamental limitations of small antennas. *Proceedings of the Institute of Radio Engineers* 35: 1479.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58: 2059.

Nonlinear Optics in Photonic Crystal Fibers

J E Sharping, Cornell University, Ithaca, NY, USA

P Kumar, Northwestern University, Evanston, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystal fibers (PCFs) are very similar to normal optical fibers in that they consist of a core surrounded by cladding, such that light is guided within the core of the fiber. The primary difference between PCF and standard optical fibers is that PCFs feature an air–silica cross-section, whereas standard optical fibers have an all-glass cross-section. An electron micrograph of a typical PCF is shown in Figure 1. The air holes extend along the axis of the fiber for its entire length and the core of the fiber is formed by a defect, or missing hole, in the periodic structure. The core is formed of solid glass, whose refractive index is that of pure silica (or whatever other glass is chosen), and the cladding is formed by the air–glass mixture, whose effective refractive index depends on the ratio of air-to-glass, also known as the air-fill fraction, that comprises the structure. The resulting effective-index of the cladding will be lower compared with that of the core and, as such, will provide the refractive index variation necessary to support total internal reflection at the core-cladding boundary, and guide light in a manner similar to that of standard optical fibers. The fiber design (i.e., size, shape, and the air-fill fraction) dictates solutions to Maxwell's equations for light propagating within the fiber. Valid solutions are referred to as 'modes' which propagate along the fiber in a known manner, and have a well-defined shape in the transverse direction (i.e., they have a well-defined transverse mode structure).

Nonlinear-optical effects in fibers result from the interaction of optical fields with the glass via the $\chi^{(3)}$, or Kerr nonlinearity. The phenomenon of nonlinear refractive index is a manifestation of a light–material interaction mediated by $\chi^{(3)}$. The magnitudes of the components of the third-order susceptibility tensor in glass, $\chi^{(3)}$, are generally quite small compared with the analogous second-order ($\chi^{(2)}$) terms for materials exhibiting such nonlinearities (e.g., lithium niobate, beta-barium borate (BBO), etc.). The relatively small $\chi^{(3)}$ nonlinearity in optical fibers makes them ideal for wavelength-division multiplexed optical communication where light propagation subject to a minimum of nonlinear effects is critical. Nonlinearity does, however, eventually become an issue in wavelength-division multiplexed systems as the launched optical power increases and as the channel spacing decreases. On the other hand, one can utilize nonlinear-optical effects in soliton communication systems and to build useful photonic devices. Despite the weak $\chi^{(3)}$

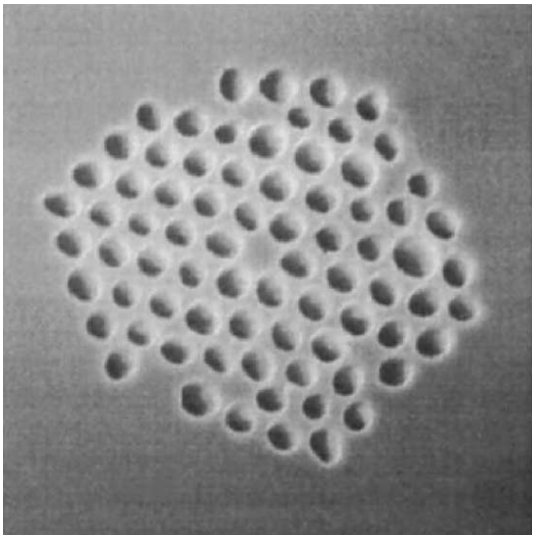

Figure 1 An electron micrograph showing the periodic microstructure of a typical PCF. The core is formed by the 'missing hole' in the center of the microstructure. (Reproduced with permission from Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America, courtesy of OFS.)

nonlinearity, the net nonlinear-optical effect in fibers can be large due to the ability to tightly confine intense fields within the core of an optical fiber and maintain the interaction over a long distance as the guided fields propagate through the fiber.

The study of nonlinear-fiber optics has benefited from dramatic improvements in optical fiber and fiber-optic device fabrication. The importance of understanding nonlinear-fiber optics is driven by the need to develop fiber-integrated devices, and also by the need to understand and mitigate the problems that these nonlinearities cause in optical communication systems.

This section introduces the unique linear- and nonlinear-optical properties of PCFs in order to understand the reasons why nonlinear-optical effects are often enhanced in such fibers. These discussions pertain to PCFs which are 'highly nonlinear'. It is essential to clarify that 'highly nonlinear' in this context does not mean that the $\chi^{(3)}$ is any larger than that of standard telecommunication fibers, rather that the effect of this nonlinearity is enhanced due to the fiber's very small core.

PCF Properties

Photonic crystal fibers feature a variety of interesting properties. From the standpoint of nonlinear-fiber optics there are four very useful fundamental properties of PCFs:

- a mechanically robust optical fiber can be fabricated with an extremely small core (a few μm^2);
- a fiber can be made to guide in a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- there are new degrees of freedom that allow one to manipulate the fiber's group-velocity dispersion (GVD) properties; and
- many, but not all, PCFs are polarization maintaining as a result of form birefringence present in the core.

The fact that small-core PCFs can be fabricated is clear from Figure 1 by taking note of the fact that the center defect region which comprises the core is about 1.7 μm in diameter. Photonic crystal fibers with even smaller cores have been fabricated.

Transverse Mode Structure

A widely accepted model used to describe the transverse modal behavior of PCFs is called the effective-index model. The effective-index model can be used to understand why some PCFs are 'endlessly single mode', meaning that the fiber guides in a single transverse mode over an exceptionally wide wavelength range (370 nm–1600 nm). In the effective-index model, the refractive index of the core, $n_{co}(\lambda)$, is that of glass, and the refractive index of the cladding, $n_{eff}(\lambda)$, assumes a value in between that of glass and air. In the context of PCFs, one makes a modification to the standard expression describing single-mode behavior in step-index fibers:

$$V_{eff} = \frac{2\pi\Lambda}{\lambda}\sqrt{n_{co}(\lambda)^2 - n_{eff}(\lambda)^2} < V_{cutoff} \qquad [1]$$

where Λ is the spacing between air holes, λ is the wavelength of light, and V_{cutoff} is the cutoff condition for the PCF. A similar expression for the V parameter is commonly used to understand the modal behavior of standard fibers where the larger the V is, the more transverse modes are supported within the fiber. In standard fibers the cutoff condition below which only a single mode can propagate within the core of a fiber is given by $V_{cutoff} < 2.405$. In the case of PCFs, a numerical method should be used to determine V_{cutoff}. Mechanically robust PCFs can be fabricated where the dispersion in $n_{eff}(\lambda)$ (i.e., the variation of n_{eff} with λ) offsets dispersion in $n_{co}(\lambda)$ and compensates for the $2\pi\Lambda/\lambda$ coefficient in eqn [1]. Therefore, the light within the fiber propagates in a single, Gaussian-like mode because for all wavelengths $V_{eff} < V_{cutoff}$. A graph of the variation of V_{eff} with Λ/λ is shown in Figure 2, where d represents the size of an air hole.

Conceptually, the effective index model can be understood by noting that at short wavelengths the mode field is confined well within the all-silica core, but as λ increases the mode field extends further into the air–glass cladding and V_{eff} and $n_{eff}(\lambda)$ both decrease.

Dispersion in PCF

Other critical differences between PCFs and standard optical fibers lie in the dispersion properties. When light propagates through a fiber its behavior depends on the light's optical frequency:

$$E(t,z) = A(t,z)e^{i[\omega t - \beta(\omega)z]} \qquad [2]$$

Equation [2] describes the mode as it propagates through the fiber. It is decomposed into a slowly varying envelope, $A(t,z)$, and a rapidly varying exponential component where ω is the frequency of the mode, t is time, z is the position along the length of the fiber, and $\beta(\omega)$ is called the mode-propagation constant. The general term used in describing the

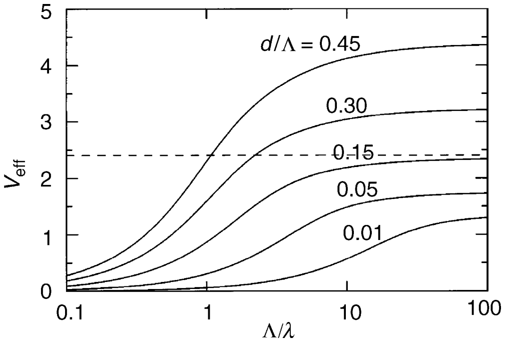

Figure 2 Variation of V_{eff} for different relative hole diameters d/Λ. The calculation assumes a fiber with an air–glass cross-section where the refractive index of air and glass was taken to be 1 and 1.45, respectively. The dashed line marks $V_{\text{eff}} = 2.405$, the cutoff value for a step-index fiber. (Reproduced with permission from Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963. ©1997 Optical Society of America.)

frequency dependence of β is chromatic dispersion, which includes contributions from the material as well as the waveguide. Other types of dispersion present in optical fibers include multi-modal (arising from multiple guided transverse modes) and polarization-mode dispersion.

One way to understand the chromatic dispersion of a mode propagating through an optical fiber is to study the Taylor series expansion of the mode-propagation constant, β, about the center frequency of the field, ω_0:

$$\beta(\omega) = \beta_0 + \beta_1(\omega - \omega_0) + \frac{1}{2}\beta_2(\omega - \omega_0)^2 + \cdots \qquad [3]$$

where $\beta_i = \mathrm{d}^i\beta/\mathrm{d}\omega^i$. The physical significance of the various β_i in eqn [3] are as follows: the phase-fronts of the electric field move at a speed given by $\omega/\beta_0 = v_p$, and the envelope, $A(t,z)$, moves at a group velocity given by $1/\beta_1$. A GVD term, which governs temporal spreading of the envelope, is given by β_2. Higher-order β terms are usually negligible for propagation of pulses of ≥1 ps duration in optical fibers and are lumped into the category of 'higher-order chromatic dispersion'.

The notation 'β_2', as defined above, is often used in the literature with dimensions of ps^2/km. However, another expression is frequently used because of its direct relationship to measured quantities. It is straightforward to measure the relative delay, T, between two pulses having different center wavelengths. Choosing a particular wavelength as a reference, one can then measure relative delay as a function of an injected pulse's center wavelength. The first derivative with respect to λ of the relative delay curve gives the GVD according to

$$D = \frac{\partial\left(\dfrac{1}{v_g}\right)}{\partial\lambda} = \frac{1}{L}\frac{\mathrm{d}T}{\mathrm{d}\lambda} = -\frac{2\pi c}{\lambda^2}\beta_2 \qquad [4]$$

where v_g is the group velocity, and L is the length of the fiber under test. The dimension commonly used for D is ps/(nm km).

Chromatic dispersion in single-mode optical fibers results from two different wavelength-dependent fiber parameters. The medium itself, glass in this case, has a wavelength-dependent refractive index. This 'material' contribution has the same magnitude regardless of the various parameters associated with the waveguide. A second contribution has to do with the design of the optical fiber. This 'waveguide' contribution to dispersion arises from the fact that the wavelength-dependent mode depends on the properties of the waveguide (i.e., the core size and refractive index contrast between the core and cladding). Empirical models can be used to describe the material, waveguide, and total GVD for standard communication fibers. Such a set of curves is given in Figure 3 where it can be seen that it is possible to have positive, negative, or zero values for D. For historical reasons, the regions where D is negative (β_2 is positive) exhibit 'normal GVD', while those where D is positive (β_2 is negative), exhibit 'anomalous GVD'. The wavelength corresponding to $D = 0$ is referred to as the zero-dispersion wavelength (λ_0), which for most silica glass fibers is about 1,300 nm.

In contrast with standard optical fibers, where the waveguide contribution to D is always less than zero, small-core PCFs can be fabricated where the waveguide contribution to GVD is positive and quite large. As such, in PCFs, λ_0 can be shifted to wavelengths shorter than the intrinsic dispersion zero of glass. Control over the GVD is essential for phase matching certain nonlinear-optical interactions involving light

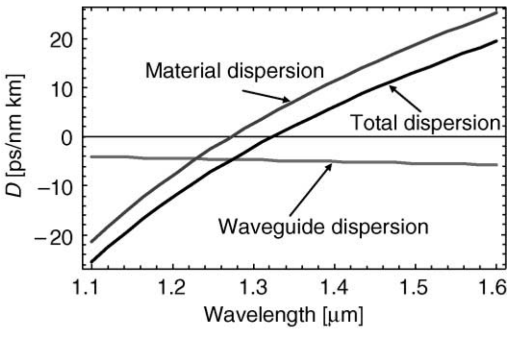

Figure 3 Plots of the theoretical dispersion coefficient, D, as a function of wavelength for a standard optical communication fiber.

of different colors co-propagating within a fiber. Indeed, several exciting applications of nonlinear optics in PCF require a fiber with a $\lambda_0 \sim 800$ nm. The GVD is also of great importance when working with pulsed light in PCF, because GVD results in temporal pulse broadening. It also governs pulse temporal walkoff effects, limiting the effective interaction length between pulses of different colors. This new flexibility to manipulate the GVD curve, by varying waveguide design parameters, is a key advantage associated with using PCFs for nonlinear optics.

Birefringence in PCF

The core of an optical fiber often exhibits some amount of anisotropy. The core may be elliptical in form (shape) which leads to a phenomenon referred to as form birefringence. Since mode propagation depends on the fiber structure, a fiber with an elliptical core will exhibit mode propagation that depends on the electric field's polarization with respect to the axes of the elliptical core. As a result of birefringence, the polarization of the mode varies as it propagates through the fiber (unless care is taken to align the polarization of the injected light with respect to a principal axis of the birefringence).

Polarization-maintaining (PM) fibers are designed to include birefringence in a particular axis of the fiber. By including a well-defined birefringence throughout the length of a fiber that is larger than that induced by external perturbations, fast and slow axes of the optical fiber are created for all guided wavelengths, giving two orthogonal 'polarization modes'. If light is injected into one of the polarization modes (i.e., with its linear polarization along one of the axes) it remains linearly polarized along that axis as it propagates along the fiber. The two polarization modes generally have different group velocities, so pulsed light in each mode will take a different amount of time to propagate through a given segment of fiber. Most PCFs exhibit strong birefringence due to a slightly elliptical core combined with a large core-cladding index difference, and so they behave similarly to PM fibers. Special care must be taken when working with PCF to be sure that the polarization of the light launched into the fiber is aligned with one of the birefringent axes.

In practice, there are a few other features of PCF that are of importance when discussing nonlinear-optic interactions:

- propagation losses are generally larger in PCFs than in standard optical fibers; and
- free-space coupling and splicing are difficult and can result in large coupling losses.

Nonlinear Phenomena

The basic principles determining nonlinear effects in PCFs are the same as those for standard optical fibers (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics)). It is the new flexibility in PCFs to obtain transverse-modal and GVD behavior different from that of standard optical fibers that makes PCFs truly interesting for nonlinear optics. The relevant nonlinear-optical effects are: self-phase modulation (SPM); cross-phase modulation (CPM); third-harmonic generation (3HG); four-wave mixing (FWM); Raman scattering; and Brillouin scattering.

Self-phase modulation (also known as the optical Kerr effect) refers to the self-induced phase shift experienced by an optical field as it propagates through a fiber. It becomes particularly important for the case of pulses of light propagating through optical fibers. In small core PCFs, SPM is enhanced due to the high-intensity light propagating within the core. Self-phase modulation can lead to substantial spectral broadening of pulsed light propagating along an optical fiber.

When a pulse of light experiences normal GVD (i.e., $D < 0$) as it propagates, the longer-wavelength components travel faster than the shorter-wavelength components. Anomalous GVD (i.e., $D > 0$) leads to the opposite, short-wavelength components traveling faster than the long-wavelength components. Group-velocity dispersion generally leads to temporal broadening of pulses as they propagate along a fiber. Under ideal conditions, however, SPM in combination with anomalous GVD, leads to pulses which propagate without any temporal or spectral broadening. These self-sustaining pulses are called 'optical solitons'.

When waves of light having different wavelengths co-propagate along a fiber, CPM can occur. It can be understood as a phase shift induced on one wave, due to the presence of the other wave. Cross-phase modulation also leads to spectral broadening and solitonic pulse propagation.

In 3HG and FWM, one or more photons are destroyed and others are created. In 3HG, three 'fundamental' photons are destroyed to create one with three times the energy of the fundamental photons. In FWM, two fundamental photons are destroyed while two others are created. While it is straightforward to conserve energy in 3HG and FWM, these interactions must be 'phase-matched', meaning that the interacting waves must be made to propagate in-phase over a meaningful length. Such phase-matching conditions need to be carefully considered when studying 3HG and FWM. Nevertheless, 3HG and FWM can be used to obtain frequency shifts and all-optical amplification. In comparison with other types of optical fibers, PCFs are particularly useful for 3HG and FWM applications. The small core of the PCF allows interactions to occur at much lower input powers, and the new flexibility associated with the GVD properties permits phase matching in cases which are not possible using standard optical fibers. Finally, the fiber's endlessly single-mode behavior permits very good transverse mode overlap between interacting waves having widely different center wavelengths.

Self-phase modulation, CPM, 3HG, and FWM are photon–photon interactions wherein no energy is exchanged with the medium itself. In contrast, Raman scattering and Brillouin scattering result from photon–phonon interactions. The differences between Raman and Brillouin scatterings lie in the energy of the phonons involved and the direction in which the interactions occur. Raman scattering is an interaction between a photon and an optical–phonon mode of the molecules making up the material. In the case of Raman scattering in glass, the energy shift, associated with molecular vibrational (Raman) modes, corresponds to frequencies of 1–12 THz. Raman scattering occurs in the forward and backward directions. With pulsed light, stimulated Raman scattering can occur when the lower-frequency spectrum of the pulse overlaps with the spectrum of the Raman resonances excited by the higher-frequency spectrum. When this happens, energy can be efficiently shifted in spectrum towards the peak of the Raman resonance. In the forward direction, this 'Raman self-frequency shift' builds up. Additionally, if the Raman self-frequency shift occurs in the presence of anomalous dispersion, a 'Raman soliton self-frequency shift' can result. As the injected power is increased, the spectral shift between the injected pulse and the resulting Raman soliton increases. The principal advantage associated with PCF is the ability to generate Raman solitons for a broader range of wavelengths than was previously possible.

For Brillouin scattering, interaction with the acoustical phonons results in frequency shifts of about 10 GHz and the interaction only occurs in the backward direction. Brillouin scattering is generally a nuisance in fiber-based devices, leading to intensity noise and other problems. The interesting feature of Brillouin scattering in PCFs is that the threshold intensity where problems begin to occur is higher for PCFs than for standard optical fibers. The higher threshold permits further optimization of fiber-based devices wherein Brillouin scattering limits the performance.

Experiment Examples

In the following subsections, a selection of experiments, demonstrating a few of the relevant nonlinear-optical phenomena, are briefly described.

Supercontinuum Generation

One of the most exciting demonstrations of nonlinear optics in PCF is that of supercontinuum generation. In a typical experiment, 100 femtosecond pulses from a mode-locked Ti:Sapphire laser operating at a wavelength of 800 nm were injected into the PCF. As the injected power was increased, a broad continuum of spectrum was generated from wavelengths of 400 nm up to 1,600 nm. Typical data showing the input and output optical spectra are given in Figure 4, where it can be seen that well over an octave of frequency spectrum is generated from an input pulse whose spectral width is only about 10 nm.

It is widely accepted that supercontinuum generation results from a combination of linear and nonlinear optical effects conspiring to generate the broad spectrum. As the pulse propagates through the PCF, SPM, FWM, and Raman scattering are all likely to occur with relative efficiencies depending heavily on the input pulse's spectral and temporal characteristics, as well as the fiber's properties. Indeed, efficient supercontinuum generation can occur in PCFs for pump wavelengths lying near the zero-dispersion point in the normal or anomalous dispersion regime of the PCF, and for fibers as short as a few millimeters in length.

Spectral broadening, due to SPM, is common in standard optical fibers, but what makes this

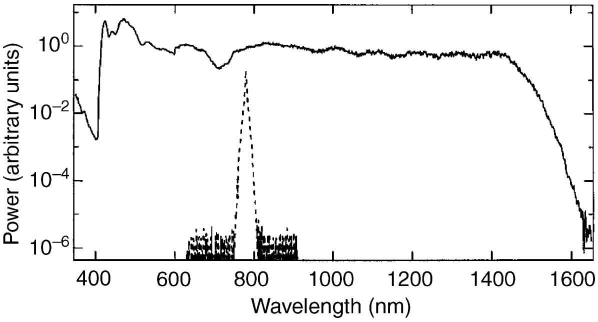

Figure 4 In supercontinuum generation one observes a broad continuum generated after short pulses of light from a Ti:Sapphire laser propagate through a 75 cm section of PCF. The spectrum of the input pulse is shown as a dashed curve, while the output is a solid curve. (Reproduced with permission from Ranka JK, Windeler RS and Strentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27. ©2000 Optical Society of America.)

particular experiment interesting is the remarkable width of spectrum generated with a short piece of fiber (~1 m) and with comparatively small optical powers.

Optical Switching in PCF

Cross-phase modulation can be used to build an all-optical switch. Such a switch can be implemented as a three-port device where the output port for a given optical bit (the signal pulse), is determined by the presence of an optical control (the pump pulse). Switching can then be achieved by dividing the signal pulse equally on two arms of an interferometer and injecting the strong pump pulse only on one arm. Because the pump pulse co-propagates with only one of the two signal pulses, there exists a CPM-induced phase difference $\phi_{\mathrm{NL}} = 2\gamma P_{\mathrm{p}} L$ at the output of the interferometer, where P_{p} is the peak power of the pump pulse and L is the interaction length. By varying the intensity of the pump pulses one can vary the magnitude of this phase difference. If a π-phase shift is achieved, one can switch the interference from destructive to constructive, or vice-versa, thus realizing an all-optical switch.

Figure 5 shows an experimental setup used to observe switching near 1,550 nm (a similar apparatus using bulk-optic rather than fiber-optic components can be used to conduct experiments near 780 nm). The pump and signal are synchronous few-ps-duration pulses with a tunable wavelength separation

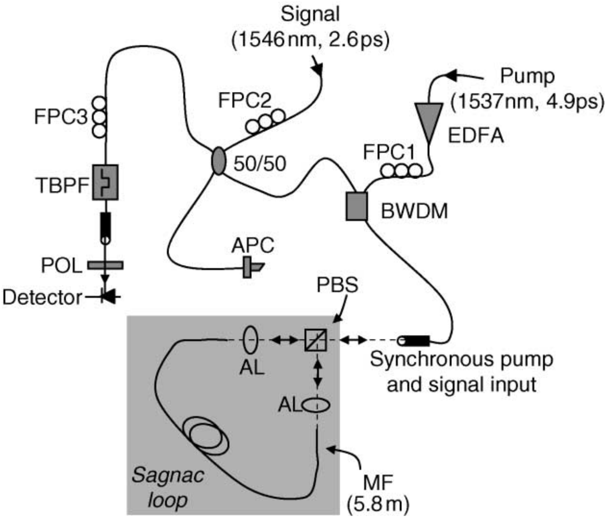
Signal
(1546 nm, 2.6 ps)
Pump
(1537 nm, 4.9 ps)
EDFA
FPC3
FPC2
FPC1
50/50
TBPF
BWDM
APC
POL
Detector
PBS
AL
AL
Synchronous pump
and signal input
Sagnac
loop
MF
(5.8 m)

Figure 5 Experimental setup used to demonstrate all-optical switching near 1,550 nm. (EDFA, erbium-doped fiber amplifier; BWDM, bandpass wavelength-division multiplexor; PBS, polarization beamsplitter; FPC, fiber polarization controller) (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE).

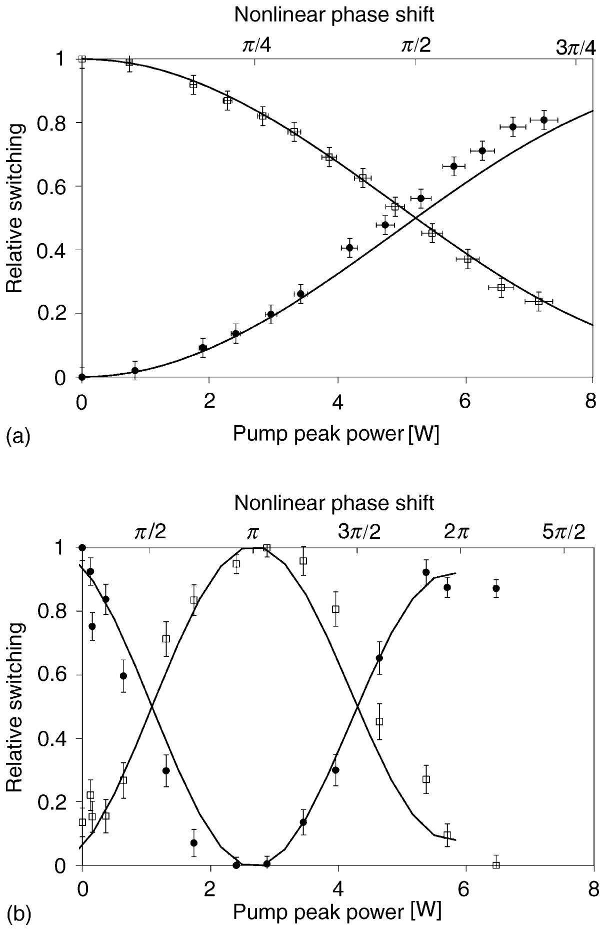

Figure 6 Switching curves (open boxes and filled circles), showing the relative power measured in each port of the switch vs. the pump peak power, for experiments conducted near (a) 1,550 nm and (b) 780 nm. The curves accompanying the data are generated from numerical solutions of coupled wave equations for CPM. (Reproduced with permission from Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79. ©2002 IEEE.)

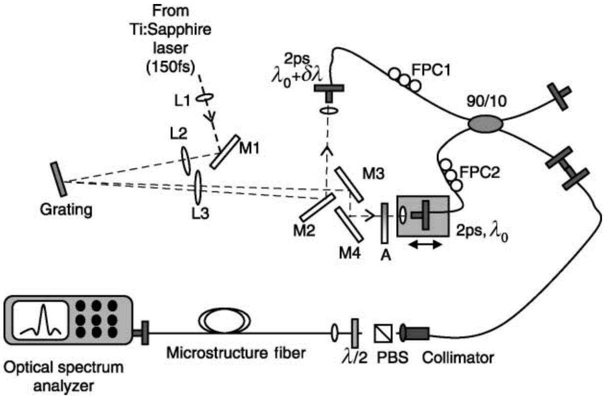
From
Ti:Sapphire
laser
(150fs)
L1
L2
M1
L3
Grating
2ps
λ0+δλ
FPC1
90/10
FPC2
M3
M2
M4
A
2ps, λ0
Optical spectrum
analyzer
Microstructure fiber
λ/2
PBS
Collimator

Figure 7 A schematic of the experimental setup used to investigate FWM in PCFs. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

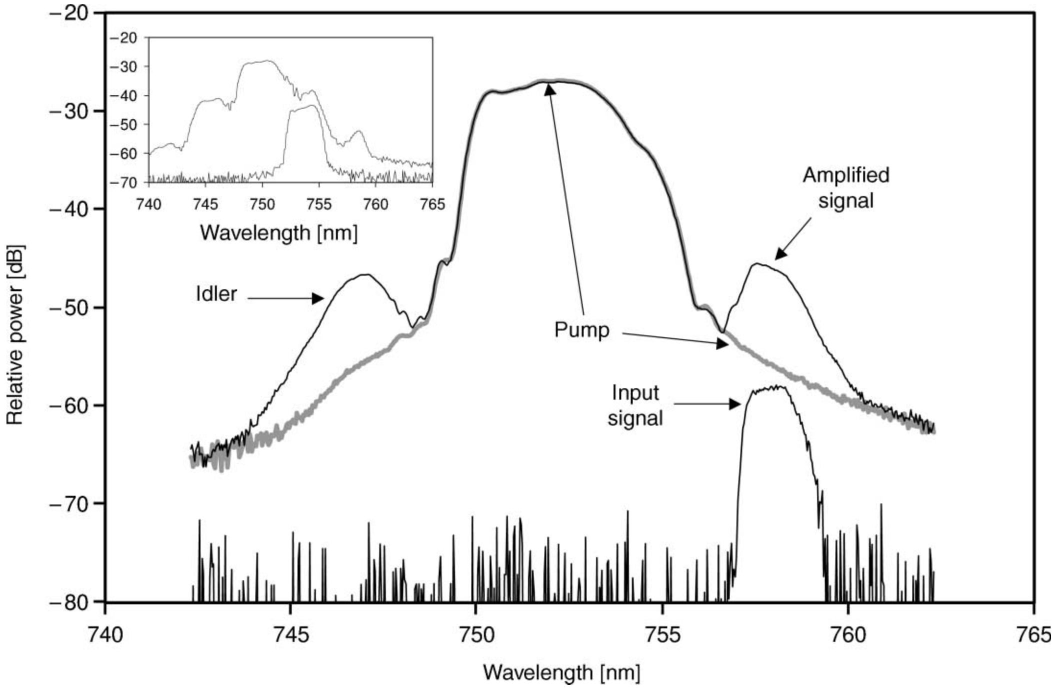

Figure 8 A typical FWM spectrum observed at the output of the microstructure fiber. The inset shows a spectrum where higher-order cascaded mixing is evident. (Reproduced with permission from Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050. ©2001 Optical Society of America.)

of 5–15 nm. The switching characteristics for this implementation are shown in Figure 6. The apparatus has the advantage of requiring short fiber lengths, low switching powers, and allows switching of weak pulses. It demonstrates the feasibility of using nonlinear optics in PCF to perform essential functions in high-speed all-optical processing.

Parametric (Mixing) Processes

The first set of experiments with controlled FWM in a PCF achieved nondegenerate parametric gains over a 30 nm range of pump wavelengths near the λ_0 of the PCF. The experiments also confirmed the wavelength dependence of the GVD coefficient of the PCF near λ_0. Since the dispersion characteristics of these fibers can be adjusted during the fabrication process, the experiments demonstrate the potential for the use of PCFs in broadband parametric amplifiers, wavelength shifters, and other optical communication devices.

The experimental setup used to demonstrate phase-matched FWM in PCF is shown in Figure 7. The pump and the input signal are two synchronous pulsed beams having 3–5 nm wavelength separation with the center wavelength tunable over a 720–850 nm range. The maximum peak power of the pump pulses is ≃12 W. The two synchronous beams are then combined and injected into the PCF. The pump and signal's optical paths are adjusted to obtain temporal overlap in the PCF and their polarizations are aligned by fiber polarization controllers.

Figure 8 shows a typical FWM optical spectrum at the output of a 6.1 m long PCF. Here the strong pump beam and the weak signal beam have wavelengths of 753 nm and 758 nm, respectively. The spectrum shows the undepleted pump, the amplified signal, and the generated idler at 747 nm. The spectra in Figure 8 show that large gain is achievable for a pump-to-signal spacing of 5 nm. Gain values of more than 20 (13 dB) were obtained.

Conclusion

In summary, the advantages of using photonic crystal fibers for demonstrating nonlinear-fiber optical effects arise from four novel properties:

- the nonlinear coefficient is enhanced in small-core PCFs (core area of a few μm^2);
- PCFs can support a single transverse mode over an extremely broad wavelength range (370 nm–1600 nm);
- PCF design parameters allow one to manipulate the fiber's GVD properties; and
- nonlinear interactions are enhanced due to the polarization maintaining properties of PCFs which result from form birefringence present in the core.

These four properties combine to allow efficient interactions to occur in PCFs which are either inefficient or not possible at all in standard optical fibers.

See also

Optical Amplifiers: Raman, Brillouin and Parametric Amplifiers. **Photonic Crystals:** Electromagnetic Theory. **Scattering:** Raman Scattering.

Further Reading

Agrawal GP (2000) *Nonlinear Fiber Optics*, 3rd edn. San Diego, CA: Academic Press.

Birks TA, Knight JC and Russell PStJ (1997) Endlessly single-mode photonic crystal fiber. *Optical Letters* 22: 961–963.

Hansryd J, Andrekson PK, Westlund M, Li J and Hedekvist P (2002) Fiber-based optical parametric amplifiers and their applications. *IEEE Journal of Selected Topics in Quantum Electronics* 8: 506–520.

Monro TM, Richardson DJ, Broderick NGR and Bennett PJ (2000) Modeling large air fraction holey optical fibers. *Journal of Lightwave Technology* 18: 50–56.

Ranka JK, Windeler RS and Stentz AJ (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Optical Letters* 25: 25–27.

Russell PStJ (2003) Photonic crystal fibers. *Science* 299: 358–362.

Sharping JE, Fiorentino M, Coker A, Kumar P and Windeler RS (2001) Four-wave mixing in microstructure fiber. *Optical Letters* 26: 1048–1050.

Sharping JE, Fiorentino M, Kumar P and Windeler RS (2002) All-optical switching based on cross-phase modulation in microstructure fiber. *IEEE Photonics Technology Letters* 14: 77–79.

Photonic Crystal Lasers, Cavities and Waveguides

J O'Brien and W Kuang, University of Southern California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic bandstructure engineering, in which the electromagnetic dispersion relations are intentionally modified, can be accomplished by creating one-, two-, and three-dimensionally periodic dielectric materials. These periodic materials, called photonic crystals, can be used to confine and guide light. This article describes the formation of optical resonant cavities and waveguides in these materials. Most of the discussion will focus on two-dimensional photonic crystals. One-dimensionally periodic systems are treated elsewhere in this volume. This article will begin with a short review of the relevant electromagnetic properties of photonic crystals. The formation and properties of resonant cavities will follow this, and the article will conclude with waveguides formed by linear defects in two-dimensional photonic crystals.

Electromagnetics of Photonic Crystals

A two-dimensional photonic crystal consists of a dielectric material in which the dielectric constant is periodic in two dimensions. The period of these materials is on the order of a half-wavelength of the operating optical wavelength. In this section, we briefly review the electromagnetic properties of two-dimensional photonic crystals. For more details, see the article entitled **Spectroscopy:** Raman Spectroscopy in this volume.

There are five Bravais lattices in two dimensions. Most of the research on photonic crystal resonant cavities and waveguides has focused on square and triangular lattices, and in this article we will consider examples based on the triangular lattice. Since these structures are usually defined lithographically however, there is, in principle, no need to confine the investigation to naturally occurring Bravais lattices. Due to the limitations of current nanofabrication technology, much of the research has focused on photonic crystals that have a finite thickness. The important feature of photonic crystals is that the bandgaps in the electromagnetic spectrum opened up by all of the Bragg planes overlap spatially and spectrally, so that a frequency region exists in which no electromagnetic propagation is allowed. These electromagnetic bandgaps can be used to confine optical modes in very small-volume resonant cavities and in waveguides. In high-contrast dielectric systems, such as a semiconductor/air periodic system, only a few lattice periods may be necessary to confine an electromagnetic mode. In the case of finite thickness photonic crystals, this periodicity has the effect of limiting the bandgap formed to the in-plane directions and in the case of a two-dimensional photonic crystal in a high-index dielectric slab to the formation of a bandgap in the guided modes of the slab. In this case there are still radiation modes of the slab.

Consider a two-dimensional photonic crystal that is periodic in the $x-y$ plane. To start, we can consider materials that are uniform and infinitely extended in z. Electromagnetic fields propagating in the $x-y$ plane are Bloch states. These can be written as:

$$\vec{H}_{k_z,k_\parallel}(\vec{\rho}, z, t) = \mathrm{e}^{\mathrm{i}(\omega t - (\vec{k}_\parallel \cdot \vec{\rho} + k_z z))} \vec{u}_{k_z,k_\parallel}(\vec{\rho}) \qquad [1]$$

where $\vec{\rho}$ labels the in-plane coordinates and $\vec{k}_\parallel$ labels the in-plane wavevectors. $\vec{u}_{k_z,k_\parallel}$ is a periodic function in the $x-y$ plane. Modes propagating in the $x-y$ plane in such a system can be classified as transverse electric (TE) waves or transverse magnetic (TM) waves. TE waves have nonzero E_x, E_y, and H_z field components, and TM waves have nonzero E_z, H_x, and H_y field components. A triangular lattice of air holes patterned into a high refractive index dielectric can be used to create a bandgap for both the guided TE and TM modes of the membrane for a range of hole radii. The TE and TM bandgaps that are formed can be over different spectral ranges, however. For laser applications this is unimportant since in practice the emission occurs from electron-hole recombination in a semiconductor quantum well, and for unstrained or compressively strained quantum well materials this emission is TE polarized. In most cases, waveguides have also been designed to work for a single optical polarization. It is generally true that a photonic lattice that consists of a connected high dielectric region is likely to exhibit a TE bandgap, while a lattice formed by disconnected high dielectric regions is more likely to exhibit a TM bandgap. The bandgap in the TE modes is formed between the first and second bands. The first band is called the dielectric band because the field at the Brillouin zone boundary is a standing wave with its intensity concentrated in the high dielectric regions. The second band is called the air band because at the Brillouin zone boundary this field is a standing wave with its intensity localized in the low dielectric regions. Over a reasonable range of lattice parameters, the bandwidth of the TE bandgap can be changed by changing *r*/*a* where *r* is the radius of the holes and *a* is the lattice constant of the triangular lattice. As *r*/*a* gets larger the dielectric band moves up in frequency. This can be thought of as being due to the fact that this mode has a decreasing effective index as *r*/*a* increases. The air band frequency increases as *r*/*a* increases. Since more of the electric field of the air band is located in the low dielectric regions than the field of the dielectric band, the air band increases in frequency with increasing *r*/*a* faster than the dielectric band and the bandgap therefore increases with increasing *r*/*a*. This trend holds in photonic crystals of finite thickness as well.

For photonic crystals of a finite thickness, the ability to simply classify modes as either TE or TM is lost. In the most simple finite thickness photonic crystals, that of a high index dielectric slab in which a two-dimensional photonic crystal has been patterned surrounded above and below by the same low index material which may be air of sapphire, or silicon dioxide, the modes can be classified as being either even or odd modes with respect to the mid-plane of the high index slab. In this mid-plane, however, modes can be classified as being either TE or TM, with each containing the same nonzero vector field components as in the infinite case. Away from this mid-plane, however, modes have in general six nonzero electromagnetic field components. The photonic bandgap that is created in this case is in the guided mode spectrum of the high index dielectric slab. There is no gap formed in the radiation modes of the slab. The radiation modes consist of modes with a ray vector that is not totally internally reflected at the high index slab/low index cladding interfaces. Figure 1 shows the dispersion relation, plotted over the irreducible Brillouin zone, of the even modes for dielectric slab with a refractive index of 3.4 in which a two-dimensional triangular lattice of holes has been patterned. The surrounding material is air. The figure shows an electromagnetic bandgap

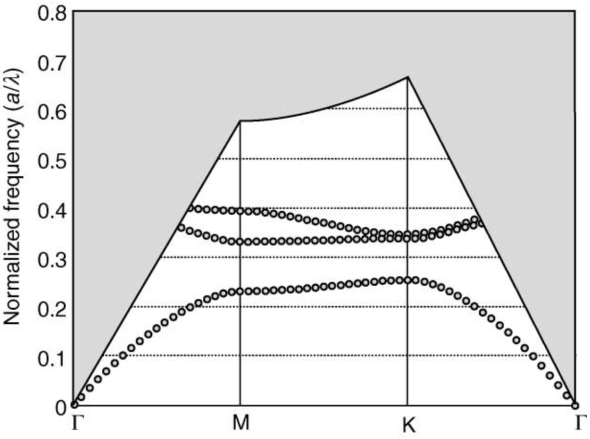

Figure 1 The photonic band diagram for a photonic crystal slab with a triangular lattice of air holes perforating a high-index dielectric slab. The slab, which assumes a refractive index $n = 3.4$, is suspended in the air with a thickness of $d/a = 0.6$. The air holes have radii of $r/a = 0.3$, where a is the lattice constant. Only the lowest three eigenmodes are shown in the plot.

formed between the first and the second guided modes of the slab. Radiation modes of the slab have real valued propagation vectors in the z-direction. These modes lie above the in-plane dispersion relation for air and exist inside the shaded region in the figure. This dividing line between the guided modes and the radiation modes is referred to as the light line.

If the upper and lower cladding layers of the high index slab are different and therefore have different indices of refraction, then the ability to classify modes as being either even or odd about the mid-plane of the slab disappears and it is possible to lose the bandgap in the guided modes altogether. In cases where the refractive index of the bottom cladding is not significantly different from that of the top cladding, such as a case where air serves as the top cladding layer and a material such as sapphire or silicon dioxide forms the bottom cladding, it is found that the even and odd modes describe the field reasonably accurately. To emphasize the approximation involved in the model, these modes are most often referred to in the literature as even-like or odd-like. Overall, there are three primary effects of having asymmetric cladding layers such as an air top cladding and sapphire bottom cladding case. These are a reduction in the effective bandgap width, an increase in the area of the radiation modes on the dispersion relation, and the loss of a rigorous bandgap in the guided modes of the slab. The first two of these effects are the most serious for device designers.

Photonic Crystal Resonant Cavities and Lasers

Photonic crystal resonant cavities can be formed at frequencies inside the bandgap or at the bandedges. Modes in the bandgap are formed as a result of a defect in the lattice. Modes at the bandedge are also sometimes used because the fields with wavevectors at Brillouin zone boundary are standing waves. These standing waves are analogous to the resonant modes that are formed in distributed feedback (DFB) laser structures. Here we will focus on modes confined in the bandgap. Allowed modes in electromagnetic bandgaps in photonic crystal slabs can be introduced by introducing one or more defects into the lattice. By perturbing a single lattice site, we can permit a localized mode or modes that have a frequency in the gap. This phenomenon is similar to the formation of deep levels in an electronic bandgap of a solid due to impurities and identical to the formation of resonant modes in vertical cavity surface emitting lasers (VCSELs) and phase-shifted DFB lasers. In each of these last cases, a defect in the periodicity of a precise thickness occurs between two one-dimensional distributed Bragg reflectors. Defect modes in photonic crystal resonant cavities can be engineered to radiate in a particular direction by trading-off the in-plane versus out-of-plane losses, and the resonant frequency is determined by the parameters of the lattice. Very small mode optical mode volumes can also be obtained in these resonant cavities leading to the possibility of modifying spontaneous emission.

Simple models exist which illustrate the formation of localized modes inside the bandgap as a result of defects in the otherwise periodic lattice. One of these was applied to the formation of deep levels in solids. This model can also be applied to the electromagnetics of defects in periodic structure. For a simple one-dimensional case, assuming that the defect exists at a single unit cell and that the dispersion of the band is dominated by nearest neighbor interactions, this can be solved exactly. Figure 2 shows the allowed frequencies of a ten unit cell chain with a single defect located at the center unit cell. Notice that one mode, called a donor mode, drops out of the band for positive values of $\Delta\varepsilon$ and one mode, called an acceptor mode, rises out of the band for negative values of $\Delta\varepsilon$. This is true in general. A perturbation of the lattice in which the defect has a higher index than the unperturbed lattice will cause a mode to drop out of the air band while a perturbation that has a lower dielectric constant than the background lattice will cause a mode to rise into the bandgap out of the dielectric band. In practice, resonant modes and frequencies in photonic crystals are modeled numerically.

A great deal of variety exists in resonant modes formed by defects in two-dimensional photonic crystal slabs. The simplest examples consist of a single missing hole in a two-dimensional square or

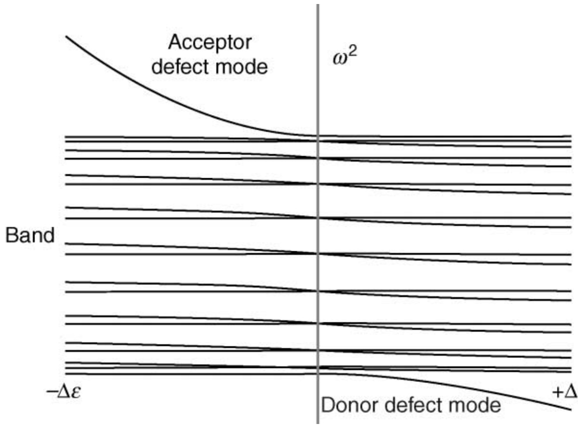

Figure 2 A one-dimensional Slater–Koster model for a single defect in a 10-unit-cell chain, showing the allowed states as a function of the defect perturbation.

triangular lattice patterned into a high-index dielectric slab. However, the resonant mode size, shape, frequency, and polarization can be engineered by tailoring the local dielectric in the region of the resonant mode. Losses in these cavities can be conceptually separated into losses in-plane through the photonic crystal and radiation losses out-of-plane. The in-plane losses are a result of having only a finite number of lattice periods surrounding the localized mode. A finite number of periods results in a finite loss due to tunneling of the fields through the lattice. This loss can in principle be made arbitrarily small by increasing the number of lattice periods surrounding the resonant mode. For a fixed number of lattice periods, the in-plane radiation loss will be reduced as the bandwidth of the photonic bandgap is increased because the magnitude of the in-plane decay constant increases with increasing bandgap width. A resonant mode formed by a defect in a truly three-dimensional photonic crystal will have losses limited by the number of achievable lattice periods and the bandwidth of the photonic crystal bandgap.

Photonic crystal cavities formed by two-dimensional photonic crystals of finite thickness suffer from an out-of-plane radiation loss. Optical confinement in the direction perpendicular to the two-dimensional photonic crystal is due to total internal reflection of the mode that occurs in the high index photonic crystal slab surrounded by lower index materials. Out-of-plane losses are due to the presence of wavevector components in the resonant mode that lie above the light line inside the radiation cone of the cladding. Figure 3 shows the Fourier transform of the magnetic field at the mid-plane of photonic crystal slab in which the resonant mode is formed by a single missing hole in a two-dimensional triangular lattice. In this cavity geometry, a doubly degenerate pair of modes is introduced into the photonic bandgap. The figure shows the Fourier transform of the field for one mode of this double degenerate pair. The boundary of the radiation cone in the figure is marked by the solid circle. Components of the mode inside this circle are not confined to the slab and contribute to out-of-plane radiation losses. These losses can be reduced by engineering the mode so that it has a smaller overlap with the radiation cone. This can be accomplished by allowing the mode to expand in real space in directions in which the wavevectors of the mode extend into the radiation cone. This reduces the spread of the Fourier transform of the mode and reduces the overlap of the mode with the radiation cone.

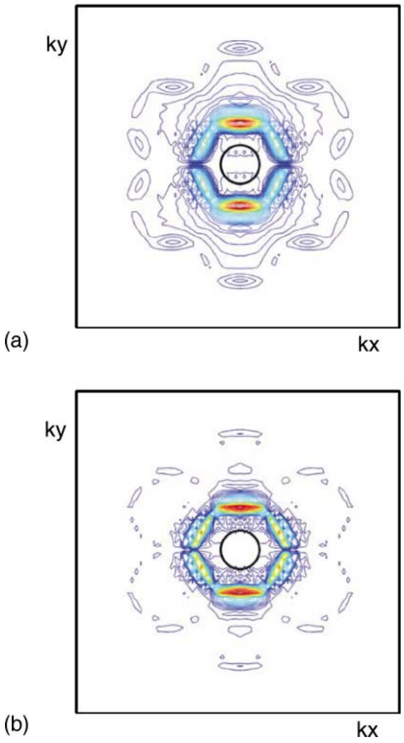

Figure 3 (a) The spatial Fourier transform of the magnetic field component, Hz, of a defect cavity mode at the mid-plane of a triangular-lattice suspended membrane single-defect photonic crystal microcavity. The inner circle indicates the light cone inside which mode components radiate vertically. (b) The spatial Fourier transform of a modified defect cavity mode. The overlap of the mode with light cone has been reduced by modifying the defect cavity.

Symmetry can also be used to reduce the coupling to the radiation fields of the slab. Reducing the resonant frequency of the mode can also reduce out-of-plane radiation losses. This has the effect of reducing the size of the radiation cone in *k*-space. One way to accomplish this is to reduce *r*/*a* in the photonic crystal lattice. This will generally reduce frequencies of all of the photonic crystal modes. It will also, however, generally be accompanied by a reduction in the bandwidth of the photonic bandgap which will increase the in-plane losses. As a result, care must be taken in designing photonic crystal resonant cavities. Nevertheless, in the smallest mode volume photonic crystal resonant cavities, the optical losses will generally be dominated by out-of-plane radiation loss.

Quality factors have been predicted to exceed 100 000 in carefully designed photonic crystal resonant cavities in which the optical mode volume is on the order of a few cubic half-wavelengths in the material. Larger resonant cavities formed by introducing multiple defects in the lattice are often less well confined in the real space lattice leading to a reduction in the components overlapping the radiation cone in *k*-space and a reduction in out-of-plane radiation loss. These cavities also are much more likely to support multiple resonant modes for a given index perturbation.

In analyzing resonance modes of photonic crystal resonant cavities, one of the most important parameters is the quality factor Q of the cold cavity. Practically, a theoretical prediction of the quality factor requires the use of numerical methods and finite-difference time-domain and finite element techniques are commonly used. There are basically three numerical approaches that can be used to calculate the quality factor of the cavity modes. Two of these methods calculate the quality factor from the time domain fields while the third is a frequency domain approach. The first method is to calculate the slope of the exponential decay of the energy of a given cavity mode with time. This decay of energy from the cavity is described by $\exp(-t/t_{\mathrm{ph}})$ where t_{ph} is the photon lifetime which is related to the quality factor Q, by $Q = \omega t_{\mathrm{ph}}$. This method is most useful for relatively low Q modes where the slope of energy decay is visibly greater than zero.

Another method is to calculate the ratio of the full width at half magnitude (FWHM) of the cavity resonance in the frequency domain, $\Delta\omega$, to the center frequency, ω_0. However, distortion to the spectrum is often introduced because the numerical simulation terminates before the impulse response is fully evolved. This has the effect of viewing the true time-domain response through a rectangular window, which translates mathematically into the convolution of the true spectrum with a sinc function. This must be accounted for in the determination of the quality factor.

A third method calculates the ratio of cycle-averaged power absorbed in the boundary to the total energy in the cavity mode. This last method has the advantage of by being able to separate the radiation losses into different directions:

$$\frac{1}{Q} = \frac{\omega P}{U} = \frac{\omega(P_{\perp} + P_{\|})}{U} = \frac{1}{Q_{\perp}} + \frac{1}{Q_{\|}} \qquad [2]$$

in which, the effective vertical quality factor $Q_{\perp}$ is given by the ratio of power lost to the absorber at the top and bottom $P_{\perp}$ to the total cavity energy $U(t)$, and the effective in-plane quality factor $Q_{\|}$ is similarly given by the ratio of in-plane power loss $P_{\|}$ to the total cavity energy.

Photonic crystal lasers and resonant cavities have been demonstrated experimentally. The first demonstration of lasing in a two-dimensional photonic crystal laser occurred in a cavity formed by a single defect in a triangular lattice that was patterned into an InGaAsP membrane. The laser operated under optically pumped conditions at 143 K. Room temperature pulsed operation of photonic crystal defect lasers was reported shortly thereafter. Figure 4 shows an electron micrograph of such a resonant cavity. In the figure, the resonant cavity is formed in a suspended membrane in which a triangular lattice photonic crystal with a defect has been patterned. The fabrication of these devices involves pattern definition by electron beam lithography and pattern transfer by a series of wet and dry etching steps. The observed quality factors are, of course, sensitive to fabrication imperfections. Many other cavity designs and demonstrations have also been reported.

To operate a photonic crystal laser continuously at room temperature, it is necessary to design a high-Q resonant cavity that dissipates heat well. Poor heat dissipation is a major drawback of suspended membrane resonant cavities. One strategy for improving the heat dissipation in photonic crystal laser cavities is to form the two-dimensional photonic crystal membrane cavity on top of a low-index high-thermal-conductivity substrate. Figure 5 shows an electron micrograph of a photonic crystal cavity formed by leaving out 37 holes from a triangular lattice patterned into an InGaAsP membrane. This membrane is bonded to a sapphire substrate, which facilitates heat dissipation. This cavity is capable of room temperature continuous wave operation.

It is believed that these lasers are capable of electrically-pumped room temperature operation. This is based on the fact that large quality factors have been demonstrated in optically-pumped photonic crystal lasers and some quality factor reduction due to free carrier absorption and absorption at metal contacts can be tolerated and still lead to lasing. A reasonable assumption would be that the

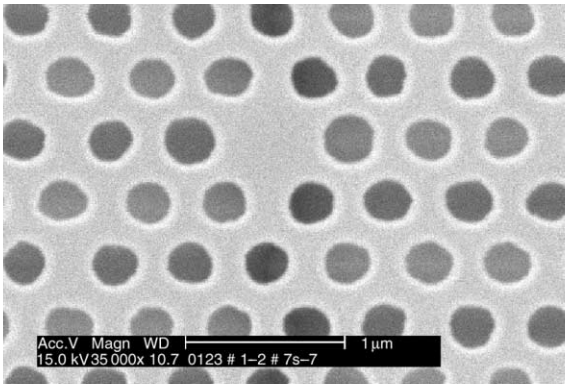

Figure 4 The scanning electron micrograph of a triangular-lattice suspended membrane photonic crystal single defect cavity.

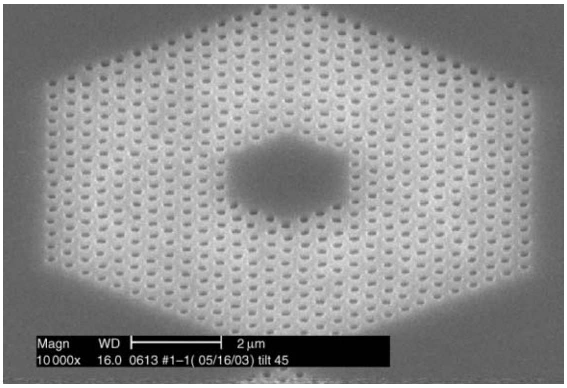

Figure 5 The scanning electron micrograph of a triangular-lattice, 37-missing-hole, photonic crystal defect cavity, bonded on sapphire which serves as lower cladding to improve heat dissipation. CW operation of the laser cavity was demonstrated.

performance of these electrically-pumped photonic crystal lasers would have much in common with the performance of VCSELs. Both VCSELs and photonic crystal cavities have small mode volumes with the cavity formed in some directions by distributed Bragg reflection. The emission direction of photonic crystal lasers can be engineered to be vertical or in-plane, but the basic resonant cavities do have important similarities. The free carrier absorption loss in photonic crystal laser cavities may be expected to be larger than in VCSELs since a photonic crystal laser mode lacks the standing wave behavior in the vertical direction that allows doping at the nodes of the standing wave to reduce the absorption loss. If free carrier absorption loss leads to larger internal losses for photonic crystal lasers than for VCSELs, then some reduction in the achievable slope efficiencies will also occur for photonic crystal lasers.

Photonic Crystal Waveguides

Another important optical element for integrated photonic circuitry is a linear waveguide. Conventional dielectric waveguides confine propagating beams by an index of refraction difference between the waveguide core and the waveguide cladding. Photonic crystal waveguides are most often formed by a linear defect, which consists of a row of modified unit cells of the lattice, patterned into a high index dielectric membrane. The guiding in this case is due to a mixture of total internal reflection at the high index membrane/low index cladding interfaces and distributed reflections from the photonic crystal in-plane. Because this confinement does not exclusively dependent on total internal reflection, the transmission loss through small bending radius waveguide branches and bends can be made very small. Predictions and experimental demonstrations exist which show that photonic crystal waveguides are capable of supporting small bending radius waveguide bends with almost total transmission. This makes photonic crystal waveguides a candidate for waveguides in densely integrated photonic circuits.

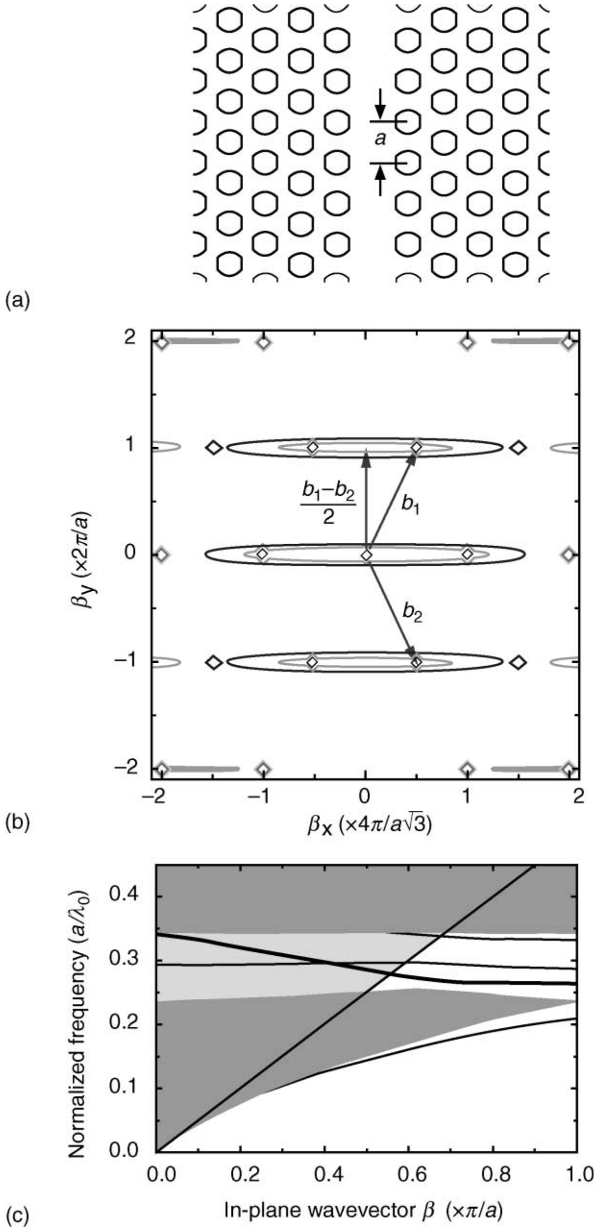

Figure 6 (a) The top view of a triangular lattice photonic crystal single-line defect waveguide with air hole radius $r/a = 0.3$. (b) The 2D spatial Fourier transform of the waveguide dielectric distribution. (c) The photonic band diagram for a suspended membrane single-line photonic crystal waveguide. The membrane has a refractive index of $n = 3.4$ and its thickness is $d/a = 0.6$. The surrounding material is air. The light gray and dark gray area indicate the regimes where field radiates through the air and photonic crystals, respectively.

Figure 6a shows an illustration of the top view of a two-dimensional photonic crystal waveguide.

This is a triangular lattice in which one row of unit cells has been modified. Shown in Figure 6b is the Fourier transform of index of refraction for the situation in which the photonic crystal consists of low index holes in a high index semiconductor. The presence of the linear defects modifies and broadens the reciprocal lattice from a simple reciprocal triangular lattice, before the creation of linear defect, to the distribution shown in the figure. This Fourier transform has components along the direction of propagation at $(\vec{b}_1 - \vec{b}_2)/2$ as shown in the figure where $\vec{b}_1$ and $\vec{b}_2$ are reciprocal lattice unit vectors. These reciprocal lattice components at $(\vec{b}_1 - \vec{b}_2)/2$ couple a planewave with wavevector β to other planewaves with wavevectors $\beta \pm (\vec{b}_1 - \vec{b}_2)/2$. The result of this coupling is that the electric field of the waveguide mode is a Bloch wave. In a photonic crystal waveguide in which the waveguide axis is the z-direction, the Bloch wavefield in the waveguide can be written in the form:

$$\vec{E}(x,y,z,t) = \sum_n c_n \vec{f}_\beta(x,z) \times \exp\left[i\left(\omega t - \left[\beta + n\left(\frac{\vec{b}_1 - \vec{b}_2}{2} \right) \right] y \right) \right] \quad [3]$$

In this expression the periodic function of the Bloch wave has been explicitly expanded as a Fourier series and each term in the Fourier series is called a spatial harmonic. Note that these fields differ from that of a photonic crystal fiber because the cladding of a photonic crystal waveguide has a periodicity along the direction of a waveguide. A photonic crystal fiber does not.

Most of the realizations of photonic crystal waveguides occur in two-dimensional photonic crystals formed in high index dielectric slabs. The membrane is usually located a few microns above a high index substrate. Because the fields of the guided modes decay exponentially outside the membrane, the effect of a substrate several microns away is very nearly negligible. The photonic band structure for $r/a = 0.3$, normalized membrane thickness $d/a = 0.6$ and refractive index of $n = 3.4$ is shown in Figure 6c. Only the modes with even symmetry along the mid-plane of the membrane are shown. A bandgap opens for modes with odd symmetry only for large low index filling factors in the triangular lattice. Filling factors large enough to create a bandgap for the odd modes are often impractical so that the odd modes are rarely considered. The even and odd modes are orthogonal, so that plotting the dispersion relation for the even modes only is justified. In reality, even and odd modes may be weakly coupled as a result of fabrication imperfections. The horizontal axis of the dispersion relation covers the in-plane wavevector β along the propagation direction of the irreducible Brillouin zone. The vertical radiation light cone and transverse radiating region of the photonic crystal are mapped as light gray and dark gray areas, respectively. They correspond to the regimes in which light radiates through the air cladding and photonic crystal cladding, respectively. Waveguide dispersion relations can be calculated using two-dimensional approaches in which case the guiding due to the high index slab is accounted for by using an effective index of refraction for the mode. The effective index is the ratio of the propagation coefficient to the free space wavevector of the guided mode of the slab. The two-dimensional

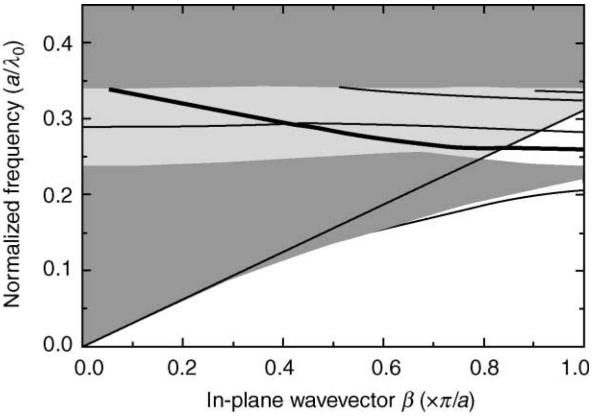

Figure 7 The photonic band diagram for a triangular-lattice sapphire-bonded single-line photonic crystal defect waveguide. The dielectric membrane has a refractive index of $n = 3.4$ and thickness of $d/a = 1.0$. The index of the sapphire bottom cladding is assumed to be 1.6.

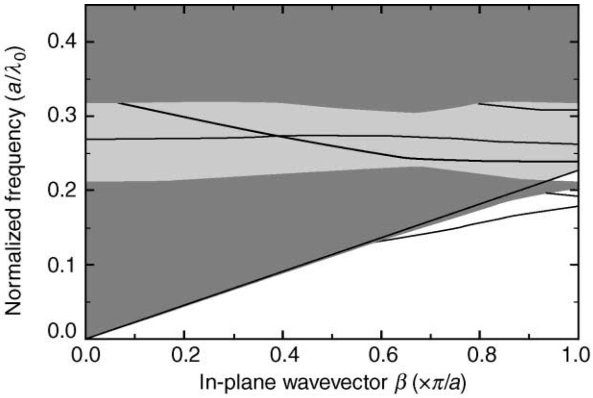

Figure 8 The photonic band diagram for a triangular-lattice deeply etched single-line photonic crystal defect waveguide. The top cladding, guiding membrane, and bottom cladding have an index of refraction of 3.0, 3.4, and 3.0, respectively. Their normalized thicknesses d/a are 1.0, 1.0, and 6.0, respectively.

calculation then uses the effective index of the mode as the high index of material in the photonic crystal. To obtain accurate results, the effective index for each mode calculated separately and the two-dimensional calculation of the photonic crystal waveguide dispersion must be repeated for each waveguide mode.

Examples of photonic crystal waveguide dispersion relations are shown in Figures 7 and 8 for waveguides formed by photonic crystals in dielectric slabs on high index lower cladding layers. Figure 7 is the dispersion relation for an oxide lower cladding layer. The area of the light gray region, which is the projection of the oxide cladding light cone onto waveguide propagation direction, increases as a result of higher index of refraction. Figure 8 shows the dispersion relation for a waveguide with an even larger lower cladding index. In this case the air holes of the two-dimensional photonic crystal are patterned all the

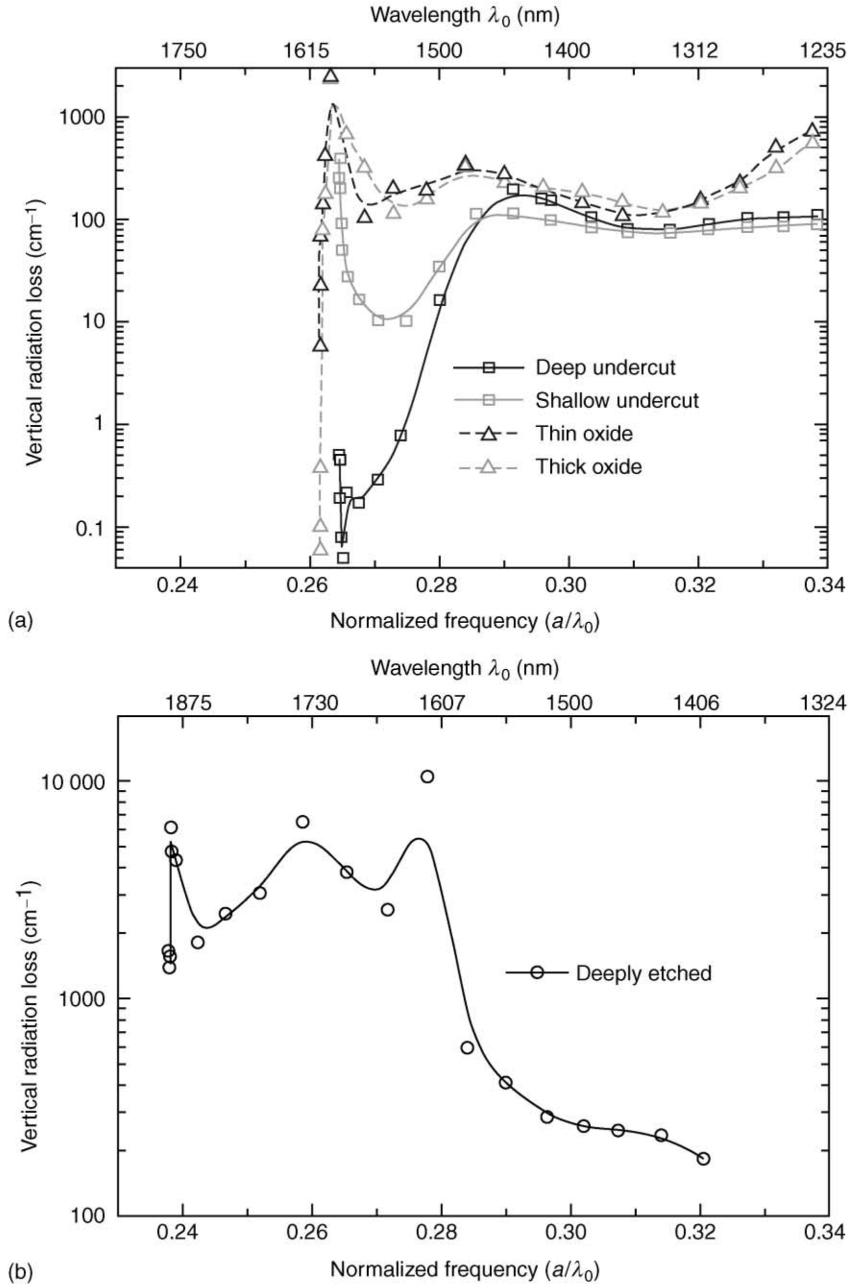

Figure 9 The out-of-plane radiation loss as a function of normalized frequency for the photonic crystal defect slab waveguides illustrated in Figures 6–8. The lattice constant $a = 450$ nm for deeply etched waveguide and 420 nm for the rest of the cases. Points are calculated values and lines are B-spline curve fits.

Table 1 Photonic crystal defect slab waveguides considered in the calculations

Waveguide structures	*Suspended membrane*		*Oxidized lower cladding*		*Deeply etched*
	Deep undercut	*Shallow undercut*	*Thin oxide*	*Thick oxide*	
Layer description: Material (normalized thickness d/a)			Air (>3.0)		
	–	–	–	–	AlGaAs (1.0)
	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (0.6)	GaAs (1.0)
	Air (3.0)	Air (1.0)	Al_xO_y (2.0)	Al_xO_y (5.0)	AlGaAs (6.0)
			GaAs substrate (>3.0)		

Refractive index of 3.4 is assumed for GaAs, 3.0 for AlGaAs, 1.6 for Al_xO_y, and 1.0 for air.

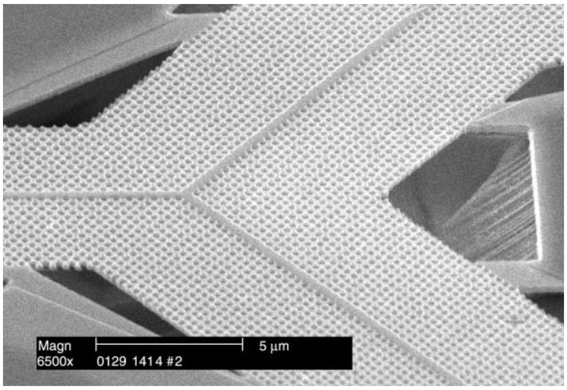

Figure 10 The scanning electron micrograph of a photonic crystal Y-branch fabricated on InP material system.

way through the lower cladding layer, with an index of 3.0, on a high index substrate. The defect modes lying within the in-plane photonic bandgap are all above the cladding light line in this case. Propagation losses are predicted to increase by over two orders of magnitude from the structure with very low index cladding layers shown in Figure 6 to the high index lower cladding structure of Figure 8.

Figure 9a and b shows the results of a fully three-dimensional calculation for the propagation loss of the three structure shown in Figures 6–8. The details of these structures are shown in Table 1. This calculated loss is due to coupling between the guided mode and the radiation modes of the high index substrate. A wavelength scale is also included using $a = 420$ nm in Figure 9a and 450 nm in Figure 9b. The suspended membrane suffers the least vertical radiation loss among all of the waveguides considered and has a minimum loss around 0.2 cm^{-1}. Low-loss waveguides formed on high index cladding layers may be possible, but care should be taken to eliminate coupling between the guided modes and the radiation modes of the cladding. In this case, the design strategy for reducing the optical loss is to reduce the magnitude of the Fourier component of the electromagnetic field inside the light cone. It should also be noted that deeply-etched waveguides formed by removing multiple rows of holes have been demonstrated with significantly lower radiation loss than is predicted for the single line defect waveguides. However, these waveguides are multi-moded at all frequencies. It is also important to remember that the nanofabricated waveguides will suffer additional losses due to fabrication imperfections that were not included in the numerical model. Figure 10 shows a nanofabricated photonic crystal waveguide. The image is an electron micrograph of a y-branch component of a Mach–Zehnder interferometer formed in a two-dimensional photonic crystal waveguide patterned into a suspended InGaAsP membrane.

Finally, it is worth noting that the group velocity of photonic crystal waveguide modes, where the group velocity is given by the gradient of the dispersion surface:

$$v_g = \nabla_{\vec{\beta}} \omega \qquad [4]$$

is very small near the zone boundary. This can be seen from the dispersion relations shown in Figures 6–8. In fact, the slope of the dispersion relations is likely to be a parameter which can be engineered by careful waveguide design. This property may allow photonic crystal waveguides to find applications in optical processing functions such as delay lines.

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory.

Further Reading

Baba T (1997) Photonic crystals and microdisk cavities based on GaInAsP–InP system. *IEEE Journal of Selected Topics in Quantum Electronics* 3: 808–830.

Joannopoulos JD, Meade RD and Winn JD (1995) *Photonic Crystals: Molding the Flow of Light*. Princeton: Princeton University Press.

Mekis A, Chen JC, Kurland I, Fan S, Villeneuve PR and Joannopalos JD (1996) High transmission through sharp bends in photonic crystal waveguides. *Physics Review Letters* 77: 3787–3790.

Painter O, Lee RK, Yariv A, *et al.* (1999) Two-dimensional photonic crystal defect laser. *Science* 284: 1819–1821.

Sakoda K (2001) *Optical Properties of Photonic Crystals*. New York: Springer.

Slater JC (1967). *Insulators, Semiconductors, and Metals, Vol. 3*. New York: McGraw-Hill Book Company.

Srinivasan K and Painter O (2002) Momentum space design of high-Q photonic crystal optical cavities. *Optics Express* 10: 670–684.

Villeneuve PR, Fan S and Joannopoulos JD (1996) Microcavities in photonic crystals: mode symmetry, tenability, and coupling efficiency. *Physics Review B* 54: 7837–7842.

Yariv A and Yeh P (1984) *Optical Waves in Crystals: Propagation and Control of Laser Radiation*. New York: John Wiley & Sons.

Self-Assembled and Functionalized Photonic Crystals

S Bhandarkar, Alfred University, Alfred, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Photonic crystals mimic the role of the atomic structure of a semiconductor in manipulating electrons, in this case for light. Hence, called semiconductors of light, these crystals are not generally found in nature and have to be designed and fabricated. Research and development in this field is taking place at a feverish pace to both understand the various nuances of this concept, as well as to evolve new methods to fabricate such crystals in the laboratory and ultimately on a commercial scale. The driving force behind this is the tremendous potential of these systems. They promise to revolutionize the way light can be used for future devices: from forming perfect, omnidirectional mirrors to the smallest ever lasers, and ultimately, complex optical chips with a wide range of functionality arranged in three-dimensional (3D) circuits.

This article seeks to elaborate on a specific technique to make photonic crystals called colloidal self assembly. Here, we will review the fundamental concepts underlying photonic crystals and see how the above process can lead to a versatile platform in making such systems. We will discuss the basic aspects of this method and look at the various innovative manifestations that have been used to fabricate many types of photonic bandgap materials.

Photonic Bandgap

A key attribute of a semiconductor is the formation of a bandgap, a band of energy states that the electrons cannot acquire. They may exist in the so-called valence band where their energy levels are less than that of the bandgap. Or they may be energized to states above the bandgap, i.e., to the conduction band, but the bandgap is the forbidden region. The continuous band of states that constitutes this bandgap is a result of the long-range order imposed by the crystal structure of the semiconductor, where the atomic lattice presents a periodic potential to a propagating electron. The tremendous usefulness of semiconductors comes from their ability to perform switching and logic functions, which in turn is a result of controlling the availability of electrons above and below the bandgap.

The propagation of light as an electromagnetic wave is influenced by the dielectric properties of the medium. For a waveguide aligned in the z direction, the field is the sum of the two propagating transverse modes, each with the form $E(x) = \Phi(x, y)e^{ik_z z}$, where $\Phi(x, y)$ is the field distribution and k_z is the propagation constant. The wave equation for the transverse mode reduces to

$$\frac{\lambda^2}{4\pi^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Phi(x, y) + n(x, y)^2\Phi(x, y) = (\bar{n})^2\Phi(x, y) \qquad [1]$$

Here, $\bar{n}$ is referred to as the effective refractive index and is given by $\bar{n} = k_z\lambda/2\pi$.

Equation [1] has the same form as the Schrödinger equation, which describes the wave functions of quantum particles such as electrons. This equivalence helps us to understand the similarity between the attractive potential well structures that bind electrons to the wells created by a local increase in the refractive index that confines light into optical modes.

The creation of a structure with a 3D periodic change in the refractive index (Figure 1) opens a gap, due to Bragg scattering off the dielectric interfaces whereby a set of frequencies of light becomes forbidden to propagate in any direction – a photonic bandgap. To understand the formation of the

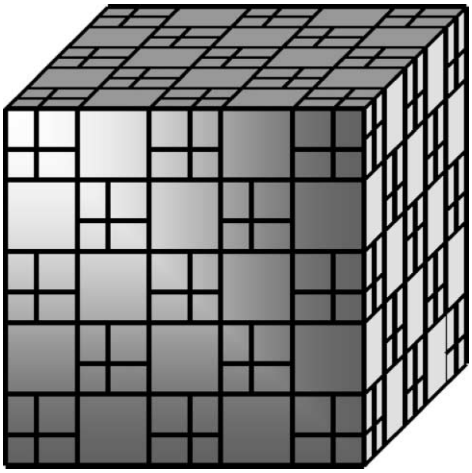

Figure 1 Schematic representation of a three-dimensional photonic crystal. The structure formed is a periodic variation in the dielectric properties of the matrix. Here, the two different dielectric regions are represented by □ and ⊞.

bandgap, one needs to consider the complex pattern of overlapping beams that results from the refractions and reflections from the numerous interfaces. This pattern is determined by the characteristics of lightwaves (wavelength and direction of travel) and those of the periodic dielectric variation (pitch and magnitude of the contrast). Thus, for a select band of wavelengths, complete cancellation through destructive interference occurs in all directions.

One way to breach this restriction is by creating a defect or a disturbance in the periodicity, which would allow localized photonic states in the gap. This defect can be a point defect to form a microcavity, or have a 1D trace that behaves as a waveguide or a planar defect. Thus, a photonic crystal with a specific defect structure can be used to manipulate light in unique ways that can be used for a variety of applications, including lasers, filters, switches, waveguides, and for 3D integrated optics circuits. These structures are also interesting because they have been predicted to exhibit enhanced optical nonlinearities.

Photonic Crystals

The lattice constant for the photonic crystal must be comparable to the wavelength of light, typically in the range of 0.5 to 2 μm. This range is three orders of magnitude greater than the lattice constants found in solids and hence photonic crystals need to be synthesized. A common example of such a structure is a diffraction grating in a waveguide, which serves as a 1D photonic crystal. There are also examples of photonics crystal structures found in nature, such as the iridescent colors seen in certain butterflies and opal gemstones. The structure of opal is a 3D photonic crystal consisting of small spherical particles of hydrated silica (SiO_2). These spheres tend to be irregular in size but in rare precious opals, they are monosized. Diffraction from these spheres leads to the remarkable colors seen in opals.

Fabrication of 3D photonic crystal structures, with two materials with sufficiently different dielectric properties, has been extensively investigated over the past several years. Given the requirement of a tightly controlled periodicity, one obvious means of doing this is to use the microlithographic patterning processes that are so well developed for very large-scale integration (VLSI). Since these are planar and not volumetric techniques, several iterations of layer-by-layer deposition, lithography, and etching are necessary. A popular second method is that of colloidal self assembly.

The Self-Assembly Technique

The term ‘self assembly’ is usually used to describe the act of spontaneous and controlled agglomeration of some basic moieties. These basic units could be molecules, supramolecules, nanoclusters, or particles. It is similar to what is seen in living systems, where organic and inorganic materials are assembled into useful configurations mediated by various proteins. Since there is no intrinsic restriction on the number of such building blocks that can spontaneously organize into periodic structures, this process is readily scalable to large dimensions, and is the main attraction of this process.

The self-assembly technique is well suited for making periodic structures such as photonic crystals. Starting with particles of the appropriately uniform size (in the range 0.5 to 2 μm), one can conceive of organizing them into a periodic 3D array. Particles, rather than molecules or nanoclusters, are the preferred starting materials for photonic crystals because their larger size provides periodicity on the length-scale of optical wavelengths. Large block copolymer molecules have also been used instead, but they are limited in their scope.

The following sections will discuss the salient details of using self assembly to form photonic crystals. The sequence of the discussion follows the path increasing process complexity. To begin, we will examine the simplest of photonic crystals made by neatly packing uniform particles called a colloidal lattice. Given the limited range of monosized particles available, such lattices are inadequate as useful photonic crystals but can be used as templates into which other materials can be incorporated. We shall discuss one particularly attractive incorporation technique, namely electrochemical deposition, in more detail. Finally, since the process is dominated by surface interactions, diverse classes of particles can be generated by simply functionalizing the particle surface. Looking ahead to the future, this is an important aspect that can vastly increase the capability of a colloidal assembly.

Formation of Colloidal Lattices

Colloidal crystallization has been studied for several decades but has received renewed attention due to both the availability of more uniform powders as well as new applications such as photonic crystals. Fine particles are unwieldy in their dry state due to high solid–gas interfacial energy and are typically dispersed in a liquid. Particle concentration is kept low to avoid agglomeration. Further stabilization of the particles may be achieved by either charging the particles' surface (electrostatic stabilization) or by adsorbing a neutral shielding molecule (steric stabilization). A preferred approach is to induce monosized spherical particles to settle into a tightly packed lattice. The settling rate, v, of particles of diameter d_p is given by the well-known Stokes equation:

$$v = \frac{d_\mathrm{p}^2 a \Delta\rho}{18\mu} \qquad [2]$$

where a is the acceleration due to gravity (or a centrifugal field), $\Delta\rho$ is the density difference between the particle and the liquid, and μ is the viscosity of the liquid. Whereas settling is a popular means of self assembling, other methods such as centrifugation, filtering, pressing, or dip coating can be used (Figure 2). For charged particles, electrophoresis, where an electric field induces flow, is applicable.

The typical lattice types in these domains are hexagonal close-packed planes (hcp) as shown in Figure 3 or face-centered cubic (fcc) structures (Figure 4). In fcc crystals, the layer stacking follows the order ABCABC; in hcp it is ABAB. In both stacking sequences, hard spheres occupy a volume fraction of 0.7404 and can be shown to have little entropic differences. However, it is experimentally observed that the preferred ordering of spheres is the fcc close-packed arrangement, which is also the case in high-quality opals.

This method clearly has the advantage of being a simple, cost-effective platform to make 3D photonic crystals that can be readily scaled to various levels. There are no repetitive steps as would be the case in layer-wise lithographic pattering. However, there are still some outstanding challenges. The primary one is the difficulty in ensuring that the organization of the particles does not incur the formation of unwanted defects and domains. Also, introduction of intentional defects such as waveguides is not trivial, although this is being demonstrated in different laboratories. Another shortcoming is the

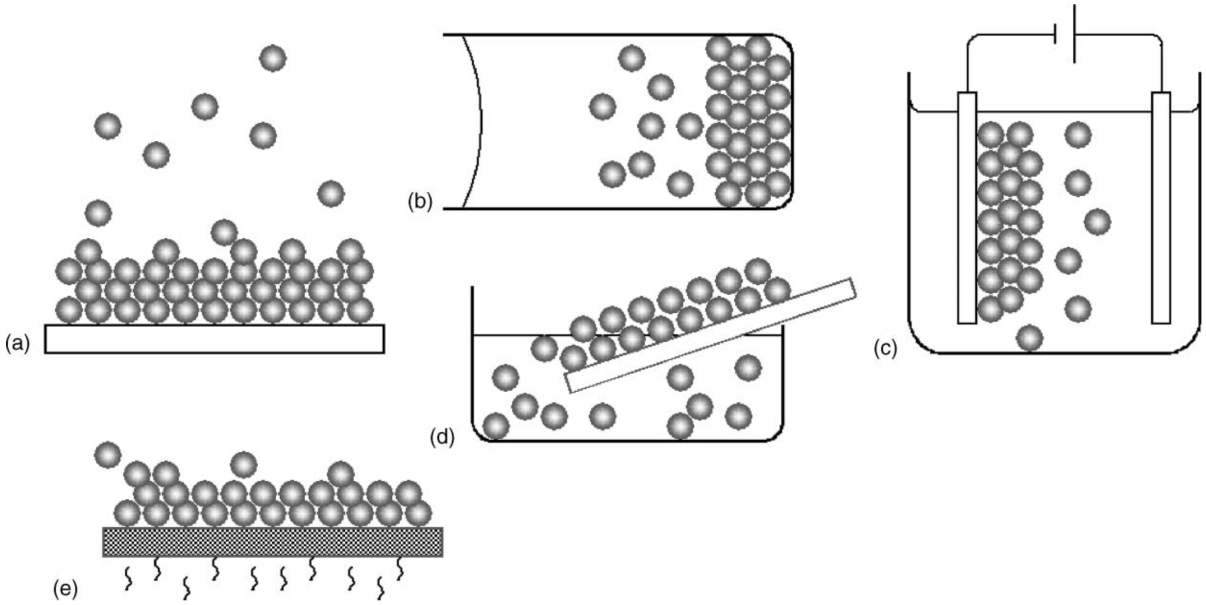

Figure 2 Diagram showing some of the techniques to use self assembly of spherical particles to make colloidal lattices. These include settling (a); centrifuging (b); electrophoresis (c); dip-coating (d); and filter-pressing (e).

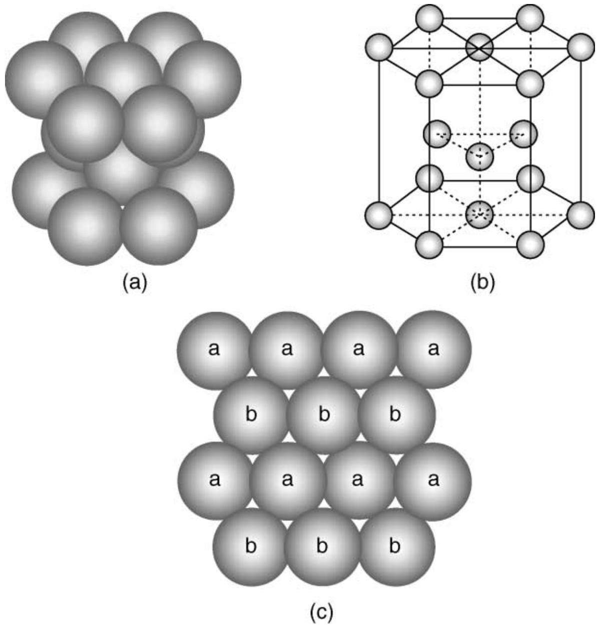

Figure 3 The arrangement of spheres in a hexagonal closed packed (hcp) arrangement (a). The unit cell is shown in (b) whereas the stacking *abab* sequence can be seen in (c).

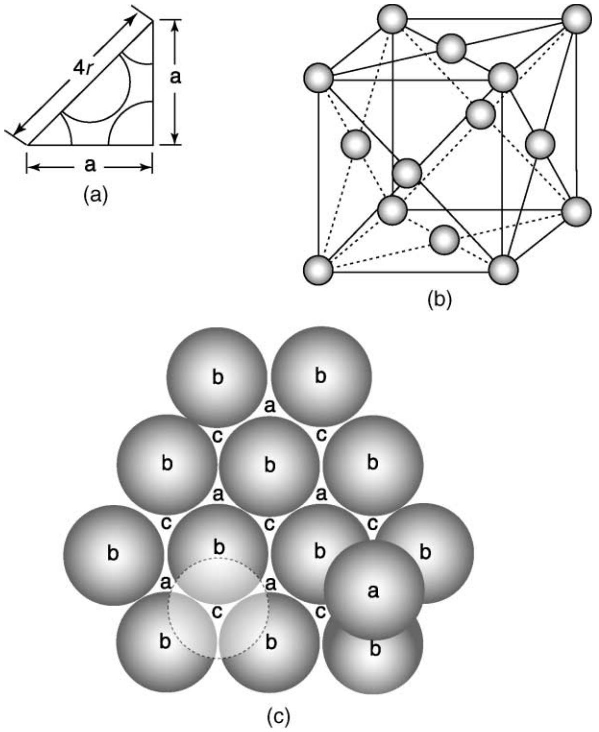

Figure 4 The arrangement of spheres in a face centered cubic (fcc) arrangement with spheres of radius *r* (a). The unit cell is shown in (b). The stacking follows an *abcabc* order. The *a* and *c* sites relative to spheres populating the *b* sites are seen in (c).

unavailability of uniformly sized particles of various different chemical compositions.

The formation of monosized spherical particles is a complex problem in its own right. Usually, such particles can be made from specific polymeric materials such as polystyrene or latex, whereas SiO_2 is the best-known inorganic system. These particles are not perfectly uniform but the typical deviations from the mean size of a few percent are acceptable for most applications. Given the attractiveness of photonic crystals for a wide array of applications in optics, research in this field is burgeoning and more breakthroughs are bound to come. Already first-generation photonic bandgap devices, such as fiber gratings or photonic crystal optical fiber, are commercially available.

Colloidal Lattices as Photonic Crystals

Due to the limited range of composition of the particles available, the particle-to-air refractive index contrast provided by straightforward colloidal lattices is small. This results in low diffraction efficiencies and hence colloidal lattices do not have photonic bandgap in all directions. However, these structures can then be used as templates for various additional modifications, as seen later.

On the other hand, a simple colloidal lattice can be more easily modeled. An analytical solution can be achieved for this system using dynamical diffraction theory for a low diffraction efficiency approximation. For a close-packed stacking of spheres in a fcc array, there exists a minimum in the transmission band at a wavelength λ_c which is given by the Bragg–Snell equation:

$$\lambda_c = 2d_s\sqrt{n_{\text{eff}}^2 - \sin^2\theta} \qquad [3]$$

where d_s is the d-spacing between the planes in the $\langle 111 \rangle$ direction, n_{eff} is the effective refractive index of the assembly and θ is the angle between the incident beam and normal to the (111) plane. This minimum in the transmission is referred to as the primary stop-band.

Photonic crystals are characterized using a Michelson or Mach Zehnder interferometer to obtain the phase and amplitude of the transmission (or reflection) as a function of the wavelength of light (Figures 5 and 6). Ideally, the whole band structure can be calculated from these data. In addition to λ_c, the close-packed structures also exhibit a transmission threshold at a wavelength λ_t, as seen in Figure 5. Both λ_c and λ_t depend on the particle diameter and increase with particle size. In a close-packed structure,

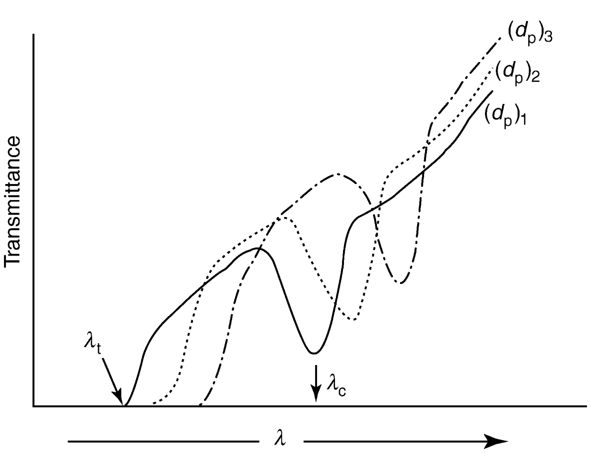

Figure 5 The typical profiles observed from transmission measurements on a photonic crystal. The stopband wavelength (λ_c) increases with d_p, where $(d_p)_1 > (d_p)_2 > (d_p)_3$. The polarization of light is fixed at *s* or *p* polarized.

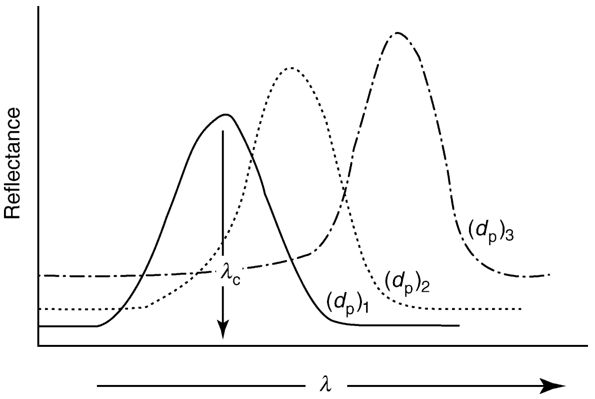

Figure 6 The typical profiles of the reflectance peaks corresponding to the transmission plot seen in **Figure 5**.

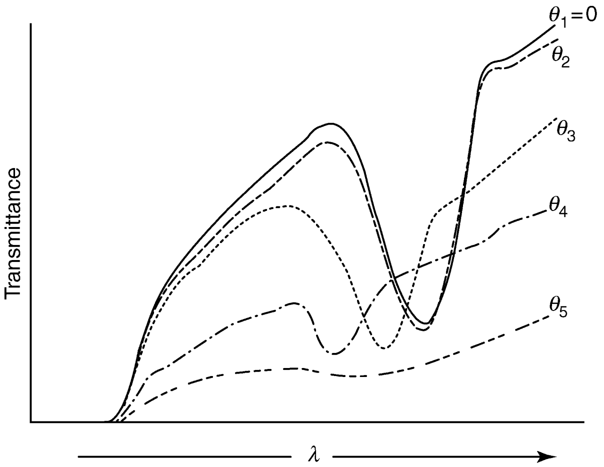

Figure 7 The transmittance profiles as a function of the angle of incidence of light of a fixed polarization. Typical values of θ would be $\theta_1 = 0$, $\theta_2 = 10°$, $\theta_3 = 30°$, $\theta_4 = 50°$, $\theta_5 = 60°$.

the lattice *d*-spacing d_s is related to the particle diameter d_p as $d_s = d_p\sqrt{2/3}$. So when the incident light is orthogonal to the structure, eqn [3] can be simplified to:

$$\lambda_c = 2\sqrt{2/3}(d_p n_{\text{eff}}) \qquad [4]$$

Thus, n_{eff} can be obtained from the slope of a plot of the stop-band wavelength λ_c as a function of the particle diameter d_p. The angular dependence of the reflectance spectra is illustrated in Figure 7. These results depend upon the polarization state of the incident light and hence need to be obtained separately for *s* and *p* polarized light.

More Complex Colloidal Lattices

While the use of SiO_2 and polymer microspheres is ubiquitous, they are optically passive materials. New methods are being developed for making uniform spherical particles from other more optically active systems. The zinc oxide (ZnO) system is one such example. The higher refractive index (~2.2) and the light emission ability of this system make it particularly interesting. Monosized particles in the size range of 100 to 600 nm can be made using a two-stage reaction method, which allows extra control over the particle formation. Other optically active systems include rare-earth doped SiO_2, dye doped PMMA spheres, or composite spheres comprised of core–shell configuration with SiO_2 or zinc sulfide (ZnS) cores and gold (Au) or SiO_2 shells, respectively.

Whereas the above discussion focuses on monosized particles, binary crystals can be made using spheres of two different sizes. The strategy here is to vary the particle size during the deposition of every alternate layer. For instance, a first layer of closely packed spheres is followed by the deposition of layers of smaller particles. The resultant crystal structure depends primarily on the concentration of the smaller particles in the dispersion. As seen in Figure 8, the second layer can vary from having three to six particles around every underlying larger particle. Thus the large-small (LS) stoichiometry of these binary crystals can be varied from LS to LS_2 to LS_3 by increasing the concentration of the smaller particles in the suspension phase. Such crystals open up interesting new avenues for complex photonic crystals.

Colloidal Lattices as Templates

By using the colloidal lattice as porous host for a second material, a greater range of photonic crystals can be synthesized. Here, the role of the colloidal array

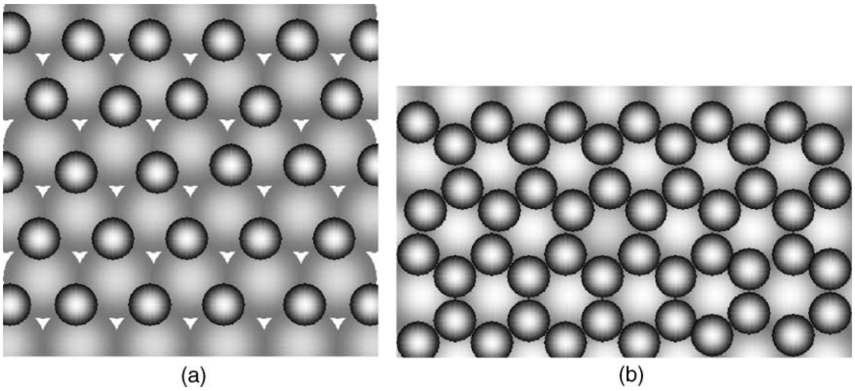

Figure 8 Binary crystals formed using spheres of two different sizes. The large-small (LS) ratio in (a) is 1:1 whereas in (b) it is 1:2.

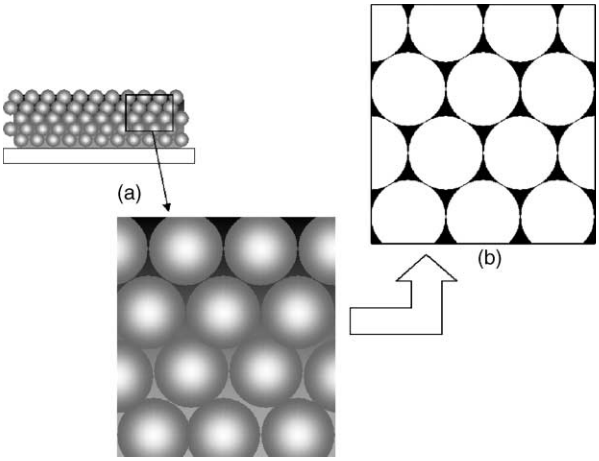

Figure 9 Schematic representation of the templating process using infiltration (a). Upon removal of the particles, a macroporous photonic crystal is formed (b). This is the inverted opal configuration.

is simply as a scaffolding structure. The second material to be infiltrated may be either a liquid (in the form of a dispersion, a solution, or a molten liquid) or a vapor. Both eventually deposit a solid material inside the template. Subsequent selective removal of the template yields a 3D photonic crystal in the form of an inverted opal (Figure 9). This is one of the best and most versatile means of achieving high index contrast, which is required for true photonic bandgap behavior.

It must be noted that to use self-assembled particle arrays as a template, the whole structure has to be often heat treated to confer greater mechanical strength. As assembled, the particles are normally attached by short-range Van der Waals forces and hydrogen bonding, which in the case of SiO_2 is silanol (Si–OH) mediated. Both of these are relatively weak bonds. By heating the assembly, a greater number of siloxane (Si–O) bonds can be formed which additionally strengthens and stabilizes the particulate network. This is important to retain the periodic structure of the assembly during the exposure to high stresses during the infiltration step, for instance. At temperatures around 750 °C, there is significant generation of siloxane bonds but minimal sintering so that the lattice does not change dimensions and retains its original open porosity. With latex particles, the annealing is done just above the glass transition temperature T_g of the polymer for no more than a critical period of time. Longer treatments can destroy the ordering of the spheres.

A template thus strengthened by heat treatment is now ready for infiltration. For example, a template comprised of polystyrene spheres can be infiltrated with a hydrolyzed sol of titanium alkoxide to ultimately form titania in the interstitial regions. Upon standing, the alkoxide sol gels in the interstitial regions leading to a polymeric network called a xerogel. This entire composite can then be further heated to high temperatures ($>300\,°C$). All organics are removed by decomposition and combustion so as to leave behind a titania matrix with periodic voids (in place of the polystyrene spheres). The templating spheres may also be dissolved, rather than combusted, in solvents such as toluene. The refractive index of titania depends on the crystalline phase that results after heat treatment – anatase is ~ 2.5 while rutile is ~ 2.9. High indices can also be achieved with semiconducting materials such as selenium (Se) with an index of ~ 2.5 or cadmium selenide (CdSe), ~ 2.9. These chalcogen-derived materials have low absorption in the visible and near infrared, so they are suitable candidates.

When SiO_2 particles are used for templating, they are eventually removed by etching, typically with dilute, buffered hydrofluoric acid at low temperatures. Given the thermal and chemical stability of the SiO_2 particles, molten Se (or other low melting metals) can be used to rapidly fill the interstitial space. Vapor phase deposition has also been used to fill the porous templates. Examples include the backfilling with silicon (Si) using chemical vapor deposition (CVD), or indium phosphide (InP) using metal-organic chemical vapor deposition (MOCVD).

A variety of structured, macroporous carbon matrices has been made with remarkable optical properties. The templates are artificial opals assembled using SiO_2 microspheres, which are filled with carbon by either: (a) infiltrating with phenolic resin followed by anaerobic pyrolysis of the resin to glassy carbon; (b) CVD of graphitic carbon; and (c) growing diamond-seeded CVD graphite from a plasma. The resultant carbonaceous 'inverse opals' are highly conductive, show intense opalescence due to the ordered arrays of holes, and have a photonic bandgap in the infrared region.

Electrochemical Deposition

One particularly effective backfilling technique is electrodeposition, which has been successfully used to deposit most of the different classes of materials – semiconductors, metals, and polymers. Here, a conducting substrate serves as the appropriate electrode where deposition occurs. By placing a colloidal lattice template on this electrode, the electrochemical deposit is used to fill the interstitial voids in the template (Figure 10). This method has been successfully applied for a host of different metals such as gold (Au), nickel (Ni), platinum (Pt), and alloys such as tin–cobalt (Sn–Co). Likewise, conducting polymers can be efficiently deposited by this method. Although these polymers suffer from inherently low indices, they have favorable electro-optic properties that can be tuned by controlling the composition. In the case of II–VI semiconductors, galvanostatic or potentiostatic deposition can be used: for instance, interstitial CdSe has been deposited (Figure 10) using an acidic solution of electrolytes such as cadmium sulfate ($CdSO_4$) and selenium dioxide (SeO_2).

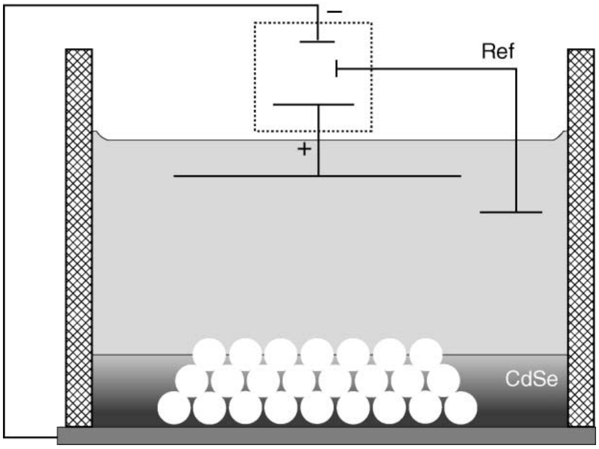

Figure 10 An electrochemical cell for deposition of CdSe from an electrolytic solution on a conducting electrode. The solution contains sulfuric acid (H_2SO_4), $CdSO_4$, and SeO_2. A potentiostat is used to generate the electric fields.

Electrochemically deposited structures have minimal porosity leading to densities close to bulk density. Hence, they have a higher refractive index and are generally more robust than those made by simple infiltration. Also, they exhibit little change in dimensions and a lower tendency toward cracking during the various physicochemical processes such as drying and sintering that occur en route to forming the photonic crystal. Hence, it is a preferred technique in many cases.

Functionalization

The self-assembly process, whether used directly or as a template to synthesize a photonic crystal, is dominated by surface interactions between particles starting from the organization process through to the thermal stabilization step. These interactions can be tailored by specially treating the surface of the particles with functional groups. Functionalized colloidal templates can enhance the infiltration of the high index material and its properties such as smoothness, density, or optical nonlinearity. Latex spheres can be surface functionalized with sulfate and carboxylic acid groups to increase their surface charging capability, which can lead to more robust self-organization. These spheres can also be coated with a layer of SiOH, which gives them a SiO_2-like appearance on the surface and allows them to be used in conjunction with SiO_2 microspheres. The infiltration of metal (such as Au) nanoparticles into a template crystal of SiO_2 microspheres can be significantly enhanced by functionalizing the SiO_2 surface. This can be achieved using long-chain organic molecules with polar groups on one end that bond with silanols on the SiO_2 surface and thiol groups on the other that strongly latch onto the gold nanoparticles. Simulations of light interaction with ordered three-dimensional arrays of these metallodielectric spheres, show that they are promising starting points for the construction of full photonic bandgap materials in the visible part of the optical spectrum. Also, in these systems, the plasmon resonance of the Au shell particles can be controlled over the visible and infrared region by varying the thickness of the Au shell.

The site-selective attachment ability of biological moieties can also be used effectively here by attaching them onto particle surfaces. DNA oligomers are well known to form conjugate specific bonds – adenine to thymine and guanine to cytosine. Thus particles functionalized with these oligomers can be made to organize in specific sequences that are differentiated, depending on the surface group. These particles can have variable sizes or compositions but their surface properties are controlled largely by the attached biomolecules. Likewise, the antigen–antibody attachment ability can be exploited to improve surface bonding.

Positioning Defects in Photonic Crystals

Much of the work on photonic crystals involves the making of a long-range periodic microstructure with high dielectric contrast because the underlying materials engineering aspects are very challenging. However, the usefulness of photonic crystals depends on the successful placement of defects in precise locations within the crystal. The placement of a point defect leads to a photonic microcavity and that of a line defect, a waveguide for propagation of a lightwave (Figure 11). However, there are certain unique advantages to this style of propagation, such as the possibility of bending light around sharp corners. Line defects in the form of air cores have been introduced in a photonic crystal using a combination of lithography and self assembly. First, ridges are defined in a photoresist on top of the substrate by optical lithography (Figure 12). Upon forming a colloidal lattice on top, the ridges are converted to linear voids by removing the photoresist using combustion or etching. Another technique is to infiltrate a photopolymerizable monomer into a colloidal crystal and preferentially polymerize the liquid by focusing a light beam inside the crystal. Thus the locus of the focused beam defines the waveguiding region that is formed upon removal of the unpolymerized liquid.

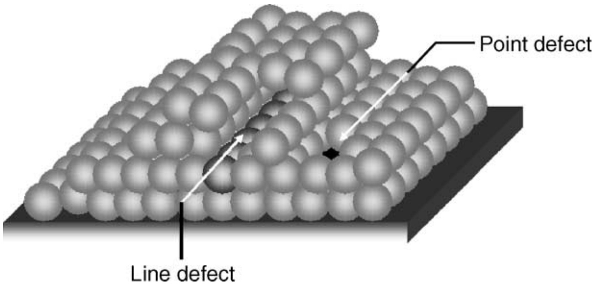

Figure 11 Two common defects in colloidal lattices.

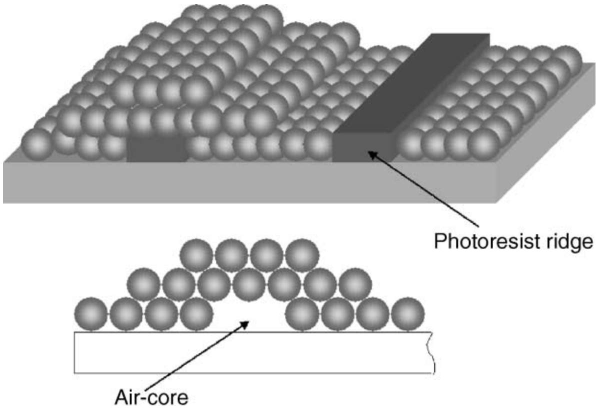

Figure 12 Formation of intentional defect using a lithographic process to define an air-core. This structure can also be used as a template for backfilling a high dielectric material.

Tunable Photonic Crystals

The holy grail in many photonic devices is tunability, and photonic crystals are no exception. Given the wide spectral range that constitutes the domain of light, the ability of devices to switch or translate their performance to another wavelength through an external stimulus is indeed very attractive. The two parameters that can be externally tuned are the dielectric contrast, which would result in a change in the n_{eff} of the crystal and the periodicity of the grating. One technique to tune the former is to form an interference pattern that organizes liquid crystals dispersed in a polymer matrix into droplets to form the photonic crystal. Applying an electric field that changes the refractive index of the droplets can therefore alter the transmission spectrum of the photonic crystal.

To control the lattice parameter, the size of or the spacing between the spheres need to be varied. This concept has also been demonstrated using self-assembled hydrogel-based nanoparticles, which are made of a thermo-responsive material such as poly-*N*-isopropylacrylamide. These particles are soft and conformable and their sizes can be tuned depending on the water content of the system. This ordered matrix is in the form of a gel that can be made to swell and compress upon changing the temperature as water is exchanged between the gel and the ambient. Wavelength tunability of one nanometer has been achieved this way. Tunability has also been sought using methods such as straining the lattice. Calculations show that a 2 or 3% shear strain ought to create a 52–73% distortion of the bandgap. In one novel approach, SiO_2 or polystyrene microspheres were strongly surface charged and ordered with a less than dense packing within a poly-acrylamide

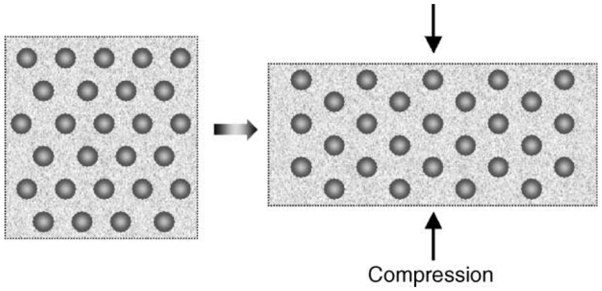

Figure 13 Tunable photonic crystal. Spherical particles are assembled in a compressible media like a hydrogel, which deforms upon applying a stress. The deformation can be used to tune the lattice parameter and hence the optical properties of the photonic crystal.

hydrogel housing (Figure 13). Due to the strong charging the arrangement tends to be body centered cubic (bcc) with the lattice spacing that depends on the concentration of the particles. When the hydrogel is deformed by applying mechanical stress, the diffraction peaks were shown to be tunable over almost the whole visible light wavelength spectrum.

Photonic crystals made from self-organized hydrogels have been used as sensors. Here, they are functionalized with chemical recognition agents that cause the gels to swell upon binding to the analyte. The resultant change in the lattice parameters causes a measurable change in the color. Specific examples include the detection of the toxic organophosphate, parathion, using acetylcholinesterase or of glucose using boronic acid and polyethylene glycol. Utilizing the same principle, even metal ions such as copper can be detected using a hydroxyquinoline-based hydrogel. These serve as a good illustration of the multifaceted potential of photonic crystals.

Summary

Photonic crystals have aroused tremendous interest with their promise to lift the field of photonics to a new level of capability. These crystals are conceptually analogous to electronic semiconductors in their function and hence are called photonic bandgap materials. The big challenge lies in synthesizing them artificially without defects. One of the few methods to synthesize these crystals is by self assembly of micron-sized building units. Self assembly leads to formation of a uniform array of monodispersed particles, typically silica or latex. This process can be further embellished by permeating a second, high-index material using the particulate assembly as a template. Useful applications of photonic crystal require them to have specifically positioned defects and tunability. These aspects, which currently constitute the state-of-the-art research in this field, promise to deliver unique devices and applications.

List of Units and Nomenclature

a	Acceleration (typically due to gravitational or applied centrifugal force)
d_p	Diameter of the particle
d_s	d-spacing or distance between lattice planes
E	Electric field
k_z	Propagation constant for travel in z direction
n	Refractive index of the medium
$\bar{n}$	Modal effective refractive index
n_{eff}	Effective refractive index of the colloidal assembly
v	Settling velocity
$\Delta\rho$	Density difference between the particle and the fluid in which it is suspended
θ	Angle between the incident beam and normal to the lattice plane in question
λ	Wavelength of the lightwave
λ_c	Stopband wavelength
λ_t	Threshold wavelength for the photonic crystal
μ	Viscosity of the suspending fluid
Φ	Electric field distribution

See also

Photonic Crystals: Atomic Physics; Electromagnetic Theory; Photonic Crystal Lasers, Cavities and Waveguide.

Further Reading

Born M and Wolf E (1981) *Principles of Optics: Electromagnetic Theory of Propagation Interference and Diffraction of Light*. Oxford, UK: Pergamon Press.

Braun PV and Wiltzius P (2002) Macroporous materials – electrochemically grown photonic crystals. *Current Opinion in Colloid and Interface Science* 7: 116–123.

Graf C and Van Blaaderen A (2002) Metallodielectric colloidal core-shell particles for photonic applications. *Langmuir* 18(2): 524–534.

Iwayama Y, Yamanaka J, Takiguchi Y, *et al.* (2003) Optically tunable gelled photonic crystal covering almost the entire visible light wavelength region. *Langmuir* 19(4): 977–980.

Joannopoulos JD, Meade RD and Winn JN (1995) *Photonic Crystals*. Princeton, NJ: Princeton University Press.

Johnson SG and Joannopoulos JD (2002) *Photonic Crystals: The Road from Theory to Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kazuaki S (2001) *Optical Properties of Photonic Crystals*. New York: Springer-Verlag.

Koenderink AF, Johnson PM, Galisteo López JF and Vos WL (2002) Three-dimensional photonic crystals as a cage for light. *Comptes Rendus Physique* 3: 67–77.

Soukoulis CM (ed.) (2001) *Photonic Crystals and Light Localization in the 21st Century*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stein A (2001) Sphere templating methods for periodic porous solids. *Macroporous and Mesoporous Materials* 44–45: 227–239.

Velev OD and Lenhoff AM (2000) Colloidal crystals as templates for porous materials. *Current Opinion in Colloid and Interface Science* 5: 56–63.

Xia Y, Gates B, Yin Y and Lu Y (2000) Monodispersed colloidal spheres: old materials with new applications. *Advanced Materials* 12(10): 693–713.

Yablonovitch E (1987) Inhibited spontaneous emission in solid-state physics and electronics. *Physics Review Letters* 58(20): 2059–2062.

PHYSICAL APPLICATIONS OF LASERS

Contents

Free-Electron Lasers in Physics
Industrial Applications
Sum-Frequency Generation at Surfaces

Free-Electron Lasers in Physics

T Dekorsy, Forschungszentrum Rossendorf, Dresden, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The topic of this article is a survey on the main applications of free-electron lasers (FELs) in physical sciences. The merits of FELs are discussed with respect to other available laser sources. For the wavelength ranges, where FELs contributed significantly to advances in physical sciences, specific examples of the research with FELs are given.

The basic physics of FELs, their historic development, and the state of the art of these laser sources are described in detail in this Encyclopedia (*see* **Lasers:** Free Electron Lasers). The present article covers the applications of FELs in physical sciences, but excludes all phenomena involved in the physics of FELs themselves. From the application point of view, one should regard an FEL as any other laser, as a source of coherent electromagnetic radiation with high power, narrow bandwidth, and a diffraction limited beam. Thus, the radiation produced by an FEL significantly differs from that generated by synchrotrons (*see* **Incoherent Sources:** Synchrotrons). The specifications of the FEL, such as wavelength, temporal structure of the output intensity, and peak power, determine the possible applications of the laser. One might now ask, why it is worthwhile to write – and read – an article on its own devoted to the applications of FELs. The answer to this question is the uniqueness of the radiation produced by FELs, which enables certain classes of experiments and applications in specific wavelength ranges not covered by other laser sources, such as gas, solid state, or semiconductor lasers.

Since the construction and operation costs of an FEL facility are usually orders of magnitude higher than those for the equipment of a conventional laser spectroscopy laboratory, FEL facilities have been built, and are planned for the future, in different places all over the world as dedicated-user facilities. These user facilities allow researchers from universities and industry to exploit the unique properties of FEL radiation for their research and provide the basic infrastructure and support for setting up the experiments. Usually potential users apply using a standard procedure for beam time at an FEL by submitting a short proposal to the facility. The proposals are judged by a panel which distributes the available beam time to the different users. So researchers active in the field of optics should keep in mind that these facilities exist and are available to anyone who has specific needs which can be fulfilled by an FEL facility. A list of all major FEL facilities is given in Table 1.

Table 1 FEL user facilities. Facilities marked with an asterisk are under construction at the time of publication

Facility name	*Wavelength ranges*
iFEL (Osaka, Japan)	230 nm–1.2 μm 1–6 μm, 5–22 μm 20–60 μm, 50–100 μm
Duke University Free Electron Laser Laboratory (US)	193–400 nm 2–9 μm
Vanderbilt University Free Electron Laser Center (US)	2.1–9.8 μm
FELIX–FOM (Rijnhuizen, Netherlands)	4.5–250 μm
Stanford Picosecond FEL Center (US)	3–15 μm 15–65 μm
CLIO–LURE (Orsay, France)	3–90 μm
Jefferson Lab (US)	3–6.2 μm
FEL–SUT–Science University of Tokyo (Japan)	5–16 μm
UCSB Center for Terahertz Science and Technology (US)	63–338 μm 338 μm–2.5 mm
ENEA Compact FEL (Frascati, Italy)	2–3.5 mm
FELBE (Dresden, Germany)*	5–150 μm
SCSS SPring-8 Compact SASE Source (Japan)*	X-ray
Institute for Plasma Research (India)*	sub-mm
BESSY FEL (Germany)*	deep UV–X-ray
Brookhaven–NSLS SDL DUV–FEL (US)*	deep UV
University of Maryland–MIRFEL (US)*	far-infrared
University of Hawaii (US)*	UV–far-infrared
SLAC SSRL X-ray FEL–LCLS (US)*	X-ray
TESLA X-ray FEL, DESY (Germany)*	X-ray
Daresbury 4GLS (UK)*	IR, X-ray
MIT Bates Lab X-ray FEL (US)*	X-ray
INFN SPARX X-FEL (Italy)*	X-ray
Tel Aviv University (Israel)*	far-infrared
Center for Advanced Technology (India)*	far-infrared

The basic principles of an FEL are briefly summarized here, while for an in-depth description the reader is referred to the article on free electron lasers (*see* **Lasers:** Free Electron Lasers). As with any other laser, the FEL consists of a laser cavity and a gain medium. The gain media are free electrons, which travel with a velocity close to the speed of light in vacuum through a periodic array of magnets called an undulator or wiggler. The magnets are arranged in a way that the electrons feel an alternating magnetic field. Due to the Lorentz force, the electrons perform an undulating or wiggling motion perpendicular to the magnetic field, which gives rise to spontaneous electromagnetic radiation into the forward direction. The radiation, emitted in due course of the passage of the electrons through the undulator, superposes constructively. The FEL may operate in such a way that the gain from one passage of an electron bunch is high enough to produce a coherent light pulse. This regime is known as self-amplified spontaneous emission (SASE) regime which is especially desirable for short wavelength operation in the UV and X-ray region, where the manufacturing of adequate resonator optics is difficult. At longer wavelengths, a laser cavity provides the feedback for the build-up of the coherent light field. The laser starts lasing when the amplification exceeds the losses of the cavity, including the light intensity coupled out of the cavity.

One of the distinct properties of FELs, in comparison to other laser sources, is their continuous tunability. The emission wavelength of an FEL is given by the following relation (*see* **Lasers:** Free Electron Lasers):

$$\lambda = \frac{d}{2\gamma^2}\left(1 + \frac{K^2}{2}\right) \qquad [1]$$

with d the period of the magnet array, the γ-parameter is the relativistic factor related to the electron velocity u via $\gamma = 1/(1 - u^2/c^2)^{1/2}$, and the K-parameter is given by d, and the magnetic field B as $K = 0.934 \times B[\mathrm{T}] \times d[\mathrm{cm}]$. From eqn [1] it becomes obvious that the emission wavelength of the FEL can be continuously tuned via the electron energy and the magnetic field. This is in distinct contrast to other laser sources, where the emission wavelength is fixed by the energy separation of electronic levels in atoms or by the bandgap energy of semiconductors.

The main advantage of FELs concerns the high power, both average and peak, which can be obtained. In addition, an FEL provides short optical pulses due to the fact that the electron beam passing through the undulator consists of short electron bunches. The temporal width of these electron bunches can be below one picosecond (10^{-12} s), so that optical pulses in the sub-picosecond to picosecond range are obtained. These pulses are extremely adjuvant for performing nonlinear optical experiments (*see* **Spectroscopy:** Nonlinear Laser Spectroscopy) and pump-probe experiments with high temporal resolution (*see* **Chemical Applications of Lasers**: Pump and Probe Studies of Femtosecond Kinetics). Typically the optical pulses are Fourier transform limited, i.e., the product of the spectral bandwidth $\Delta\nu$ times the pulse duration $\Delta\tau$ equals a constant, which depends on the specific shape of the pulse. For a pulse of Gaussian shape in frequency and time, this so called time-bandwidth product is given as $\Delta\nu\,\Delta\tau \geq 0.441$. From this equation the minimum spectral width (temporal duration) can be calculated for a given temporal duration (spectral width) of the laser pulse. When the electron beam driving the FEL is produced by a superconducting radio frequency linear electron accelerator, the repetition rate of the pulses is in the range of some tens of MHz. The energy per optical pulse is obtained by dividing the average power of the FEL through the repetition rate. The peak power within an optical pulse is obtained by dividing the energy per pulse through the pulse duration, which can be in the picosecond to sub-picosecond range. Typical values for FELs are energies per pulse in the μJ to mJ range, pulse lengths from some 100 femtoseconds to a few picoseconds, and accordingly peak powers in the range of kW to GW. Depending on the type of electron accelerator employed, the pulse structure of an FEL may have further peculiarities. Superconducting radio-frequency accelerators provide a quasi continuous-wave (CW) train of electron pulses, thus enabling the FEL to emit a continuous optical pulse train. Electrostatic accelerators provide macropulses of electrons at some Hz repetition rates with a duration of several μs. These macropulses consist of micropulses with MHz repetition rates and picosecond or sub-picosecond duration. Hence the FEL output has the same temporal characteristics with the FEL turning on at the beginning of each macropulse.

Another important development is the synchronization of table-top short-pulse laser systems to FELs. This is accomplished by stabilizing the repetition rate of the table-top laser system via a movable intracavity mirror accurately to the repetition rate (or a higher harmonic of the FEL) of the FEL. Using a high-frequency electronic phase-locked loop, a timing jitter, down to 500 femtoseconds between the FEL pulses and those from a mode-locked titanium-sapphire laser, have been obtained. Nonlinear frequency conversion of the near-infrared pulses from the synchronized titanium-sapphire allows one to perform multi-color pump-probe measurements in a broad spectral range in combination with the FEL.

In the following, FELs will be discussed with respect to the wavelength range which they cover. For each wavelength range alternative laser sources are given and compared to FELs. It should be noted that these alternative sources evolve very rapidly and thus pose challenges on the justification of FELs in certain wavelength and power ranges. For each wavelength region selected scientific achievements obtained from experiments performed with FELs are given.

FELs in the Far Infrared Region

The far infrared region is roughly defined as the wavelength (energy) range from 20 μm to 300 μm (62 meV to 4 meV). At the long wavelength end (wavelengths larger than 220 μm) radiation can be obtained from pure electronic devices, so-called backward wave oscillators (BWO). Furthermore, optically pumped gas lasers operate at certain discrete wavelengths in this range where there exists one fundamental problem for realizing a semiconductor laser based on transitions between the conduction and the valence bands. Auger recombination, where an electron and hole combine by transferring another electron high into the conduction band, becomes increasingly important for smaller bandgaps. Recent advancements in this field are based on quantum cascade lasers, which use inter-sub-band transitions of quantum structures for achieving lasing in the wavelength range from 105 μm to 3.4 μm. Another possibility to generate coherent radiation in this wavelength range is optical rectification or difference frequency mixing of femtosecond laser pulses at near-infrared wavelengths. However, none of these sources outperform FELs in terms of peak or average power, making FELs the most important tunable laser source in this wavelength range.

In the far-infrared, several scientific opportunities exist. In condensed matter many important elementary excitations exist in this frequency range, e.g., phonon-polaritons, plasmons, polarons, and magnons. The quantum confinement in semiconductor heterostructures leads to the formation of discrete electronic levels with energy separation typically in the far-infrared. The associated radiative transitions, so-called inter-sub-band transitions, can be arbitrarily tuned by varying the materials and the dimensions of the confining potential. In surface science, intramolecular vibration of molecules adsorbed to surfaces are of large interest. They can be studied by resonant nonlinear optical experiments such as sum-difference frequency generation or second harmonic generation, giving information about the energy transfer between the molecules and the surface. In biophysics, low-frequency modes of large bio-molecules lie in the far-infrared, whose dynamics can also be studied.

Important scientific results have been obtained with FELs in the following fields, which exclude the important fields of molecular science, chemistry, biology, and medical research.

Scanning Near-Field Optical Microscopy (SNOM)

Increasing the spatial resolution is an important task in spectroscopy. Free-space focusing of a laser beam is diffraction-limited to a spot with the dimensions of the wavelength. Especially in the far-infrared region this would allow to resolve objects larger than 10 μm only. However, the imaging of elementary excitations in solid state or of collective excitation in biomolecules on a sub-μm scale, would be highly desirable. SNOM allows to circumvent the diffraction limit. Several different SNOM techniques have been introduced in recent years, most of them working with apertures smaller than the wavelength. Imaging is performed by scanning the aperture over the sample or vice versa. An evanescent electromagnetic field is present at the aperture which can couple to the specimen brought into the proximity of the aperture. Since the aperture is smaller than the wavelength, the light intensity drops by order of magnitudes. It is here that the high power available from FELs becomes advantageous. Usually, the end of a fiber is tapered and metalized, leaving a small aperture at the tip of the fiber. In the far-infrared region the problem exists in the absorption of the fiber material. Chalcogenide fibers are capable of transmitting in the 2–12 μm range, while hollow glass capillaries coated with metals inside work also at longer wavelengths. Recent approaches try to work without apertures, e.g., by coupling the light locally via a nanometer metal tip to the sample. Another interesting approach is to use the focus of visible laser light to create a small highly reflective spot on a material. By creating free-carriers with the visible laser light in the material, the reflectivity of the far-infrared light of the FEL can be strongly increased, thus a reflecting spot is created which is diffraction-limited by the wavelength of the visible light rather than that of the infrared light. The techniques described here are, of course, not limited to the far-infrared region but may also be applied successfully, for example, in the mid-infrared region.

Rabi Oscillations

Rabi oscillations are a fundamental phenomenon in a quantum mechanical two-level system driven by a light field. The light field promotes electrons from the lower to the upper state and back, depending on the duration and intensity of the light field, i.e., the population oscillates between the lower and upper state of the two-level system with the oscillation period depending on the intensity of the light field. This phenomenon has been studied extensively in atoms. In semiconductors, scattering processes strongly aggravate the observation of Rabi oscillations. By using the Santa Barbara FEL Rabi oscillations between the ground and excited state of a hydrogenic, impurities in GaAs could be observed. From the excited state of the impurity, electrons are transferred to the conduction band, thus changing the photoconductivity of the crystal, which is measured as a function of time delay. Overdamped oscillations with frequencies changing with the laser intensity could be clearly resolved.

Optical Nonlinearities in Semiconductors

An important concept of nonlinear optics is based on the nonlinear susceptibility $\chi^{(n)}$ of a material, where n denotes the order of the nonlinearity. The nonlinear susceptibility has tensorial character and determines, for example, the efficiency of second harmonic generation and difference frequency mixing ($\chi^{(2)}$), third harmonic generation, and the Kerr nonlinearity ($\chi^{(3)}$), etc. Hence the knowledge about the dispersion and the absolute value of $\chi^{(n)}$s is both of fundamental and practical importance. Due to the lack of tunable laser sources, these values are not well known in the far-infrared. Especially close to fundamental lattice resonances in the far-infrared, a strong dispersion is theoretically expected. Here FELs play an important role in the determination of these values. For example, in GaAs, the dispersion of the $\chi^{(2)}$ has been recently determined from second harmonic generation experiments below the fundamental optical phonon resonance at 8 THz, i.e., wavelengths larger than 37 μm, where a resonance enhancement at half the phonon frequency could be observed.

Relaxation Kinetics in Semiconductor Quantum Structures

The investigations of inter-sub-band transitions in semiconductor heterostructures have opened the pathway for the realization of the quantum cascade laser. In these devices, the radiative and nonradiative transition rates between different electronic quantum levels are of prime importance for a working device. The relevant time-scales have been obtained from pump-probe experiments with FELs. In semiconductor quantum dots, the strong coupling of phonons to inter-sublevel transitions could be observed and the polaron lifetime could be determined.

Bloch Oscillations in Semiconductor Superlattices

Bloch oscillations are an intriguing concept of solid state physics. When an electron experiences the force of an electric field in a periodic potential, the resulting dynamics is oscillatory in the absence of scattering events. However, since in real solid state the scattering rates are much higher than the expected oscillation period, such oscillations cannot be observed and the application of an electric field leads to a net electric current. However, in artificial semiconductor superlattices, the Bloch oscillation period can be tailored to be smaller than the scattering rate, leading to the observation of Bloch oscillations and the associated THz emission. The inverse effect of a Bloch oscillator, i.e., the generation of a photocurrent by tuning an FEL to the Bloch oscillation frequency, could be observed.

Cluster Physics

FELs can be an important tool for studying the dynamics of clusters and molecules in the gas phase by infrared resonance enhanced multiphoton ionization (IR-REMPI). In IR-REMPI, the species of interest are irradiated by pulsed, tunable IR radiation from a free electron laser. When the radiation is resonant with a vibrational mode of the molecule or cluster, the absorption of many – up to several hundred – photons can take place. When the internal energy is high enough, the thermal emission of an electron can take place. The resulting ions can be detected mass selectively in a time-of-flight mass spectrometer. Monitoring the mass specific ion yield as a function of IR wavelength then yields the IR spectrum of that species. This technique has given important information on the IR absorption spectra of interstellar dust.

FELs in the Mid-Infrared Region

This wavelength (energy) range extends from 20 μm to 2 μm (62 meV to 0.62 eV). Alternative sources in this wavelength range are nonlinear optical parametric generators and amplifiers (OPGs/OPAs) which compete against FELs in terms of peak power. However, there still exists a large number of interesting phenomena where FELs are better suited, e.g., due to their higher repetition rate in the MHz range as opposed to the kHz repetition rate of OPGs/OPAs and their higher average power exceeding several W up to kW as compared to mW for OPGs/OPAs. These high-power levels find applications in material processing and recently in surgery. Since these applications cover such a wide variety of different materials, we refer the reader to publications from users of the Jefferson Laboratories FEL and the Vanderbilt University FEL (see Table 1). The advantage of FELs in this context is that their wavelength can be tuned exactly to a resonance of the material under preparation. The following significant spectroscopic application was performed with FELs.

Vibrational Relaxation of Defects in Semiconductors

The heat dissipation in submicron electronic devices is of prime importance for the failure behavior of such devices. As the device size shrinks to the nanometer scale, atomic defects play an important role. In silicon, hydrogen passivation of defects is a standard process step in the production of metal-oxide-semiconductor electronic devices. Hence detailed knowledge about the stability and heat dissipation of hydrogen defects is very important. By employing nonlinear optical methods such as four-wave mixing or a transient bleaching technique, the vibrational lifetime of hydrogen vibrational modes in silicon (around a wavelength of 5 μm to 6 μm) could be determined and important information on the dependence of the local atomic arrangement and the vibrational lifetime could be obtained.

FELs in the Near-Infrared to the UV Region

The wavelength (energy) range from 2 μm to 200 nm (0.62 eV to 6.2 eV) is the realm of solid state lasers. Especially diode pumped solid state lasers have made tremendous improvements in terms of power and stability in the past years. Nonlinear frequency conversion of ultrashort optical pulses generated from titanium-sapphire lasers and amplifiers cover almost all of this wavelength range. Hence there is no specific need for operating FELs in this wavelength range for applications in physical sciences.

FELs in the Extended UV to X-ray Region

In the wavelength (energy) range shorter than 200 nm (larger than 6.2 eV) the availability of laser sources is very scarce. In this wavelength range, solid state lasers have the problem of the reabsorption of the laser light by the crystals' hosts. Also nonlinear frequency conversion processes are hampered by the absorption of the nonlinear crystals in this wavelength range. The generation of ultrashort pulses at the wavelength of the Cu K_α line has been obtained on the basis of amplified femtosecond titanium-sapphire pulses, thus enabling the study of the lattice vibrations in the time domain via a modulation of the X-ray diffraction. Since these sources are limited in their wavelength and output power, the development and operation of FELs in this wavelength range is truly important.

Only recently the first experimental verification of the feasibility of a SASE FEL at short wavelengths was obtained at the DESY facility. The lasing wavelength was 110 nm. Since several of these sources are presently under construction worldwide, there are as yet no experimental results published. However, the future development of short-wavelength FELs will enable new types of time-resolved experiments in atoms, solid state physics, chemistry, and biology. One prominent example is protein crystallography with X-ray pulses. The high intensity of FELs in this wavelength range should allow one to obtain a diffraction pattern of a single protein or other important bio-molecules in a single shot on a subpicosecond time-scale, i.e., before the protein is fragmented. The short-pulse duration opens the pathway to study the dynamical behavior of such structures in pump-probe experiments with solid-state lasers synchronized to the FEL. In solid state physics, such experiments could yield important insights into the dynamics of phase transitions.

Other Developments

Beyond the use of an FEL as a light source in the above described wavelength ranges, there exists the possibility of using FEL facilities in a broader way. One example is the development of a high-intensity γ-ray source which can be accomplished by Compton backscattering of FEL photons from the relativistic electrons driving the FEL. This work has been performed at a storage ring FEL (OK-4/Duke University), where 12.2 MeV γ rays were obtained by backscattering 379.4 nm free-electron laser photons from 500 MeV electrons circulating in the storage ring. The advantage of these γ rays is that they are highly monochromatic and linearly polarized, thus making them an ideal source for nuclear spectroscopy and investigations in cancer therapy.

See also

Incoherent Sources: Synchrotrons. **Lasers:** Free Electron Lasers. **Spectroscopy:** Nonlinear Laser Spectroscopy.

Further Reading

Andruszkow J, Aune B, Ayvazyan V, *et al.* (2000) First observation of self-amplified spontaneous emission in a free-electron laser at 109 nm wavelength. *Physical Review Letters* 85: 3825–3828.

Benson SV (2003) Free electron lasers push into new frontiers. *Optics and Photonics News* 14: 20–25.

Budde M, Lupke G, Chen E, *et al.* (2001) Lifetimes of hydrogen and deuterium related vibrational modes in silicon. *Physical Review Letters* 87: 145501–145503.

Cole BE, Williams JB, King BT, Sherwin MS and Stanley CR (2001) Coherent manipulation of semiconductor quantum bits with terahertz radiation. *Nature* 410: 60–62.

Colson WB, Johnson ED, Kelley MJ and Schwettman HA (2002) Putting free-electron lasers to work. *Physics Today* 55: 35–41.

Dekorsy T, Yokovlev VA, Seidel W, Helm M and Keilmann F (2003) Infrared-phonon-polariton resonance of the nonlinear susceptibility in GaAs. *Physical Review Letters* 90: 055508.

Helm M (ed.) (2000) *Long Wavelength Infrared Emitters Based on Quantum Wells and Superlattices*. Amsterdam: Gordon and Breach.

Knippels GMH, van de Pol MJ, Pellemans HPM, *et al.* (1998) Two-color facility based on a broadly tunable infrared free-electron laser and a subpicosecond-synchronized 10-fs-Ti sapphire laser. *Optical Letters* 23: 1754–1756.

Litvinenko VN, Burnham B, Emamian M, *et al.* (1997) Gamma-ray production in a storage ring free-electron laser. *Physical Review Letters* 78: 4569–4572.

Murdin BN, Knippels GMH, Vandermeer AFG, *et al.* (1994) Excite-probe determination of the intersubband lifetime in wide GaAs/AlGaAs quantum wells using a far-infrared free-electron laser. *Sem. Sci. Techn.* 9: 1554–1557.

Palanker DV, Knippels GMH, Smith TI and Schwettman HA (1998) IR microscopy with a transient photo-induced near-field probe. *Opt. Comm.* 148: 215–217.

Sauvage S, Boucaud P, Lobo RPSM, *et al.* (2002) Long polaron lifetime in InAs/GaAs self-assembled quantum dots. *Physical Review Letters* 88: 177402.

Sorokina IT and Vodopyanov KL (eds) (2003) *Solid-State Mid-Infrared Laser Sources*. Berlin: Springer.

Unterrainer K, Keay BJ, Wanke MC, *et al.* (1996) Inverse Bloch oscillator: strong terahertz-photocurrent resonances at the Bloch frequency. *Physical Review Letters* 76: 2973–2976.

von Helden G, Tielens ACGM, van Heijnsbergen D, *et al.* (2000) Titanium carbide nanocrystals in circumstellar environments. *Science* 288: 313–316.

World Wide Web Virtual Library (2004) *Free Electron Laser Research and Applications*. Available online at http://sbfel3.edu/www/UL_fel.html

Industrial Applications

I P Mercer, Imperial College of Science, Technology and Medicine, London, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Lasers are seeing increasing application in industry, driven by emerging market demand and continuing enabling advances in laser technology. Core laser technology advances are based on robustness, performance, and cost. A combination of gas and solid-state lasers currently see application across the spectrum, from the extreme ultraviolet (EUV) to the infrared (IR). Industrial applications continue to emerge across the spectrum and are being extended to the extreme ends of the spectrum at terahertz (THz) and X-ray frequencies (see Figure 1).

Where a step in level of performance is offered by a laser based tool, that is robust enough and cost-effective for industrial application, it can be anticipated that new applications and markets will emerge. For this reason, many established laser manufacturers and new entries in the 1980s and 1990s were driven in part by the exploitation of laser technology advances, as opposed to market application demand. As markets mature for lasers applied to materials processing and diagnostics, so the driver across industry is beginning to shift from being predominantly set by the capability of emerging laser technology, to a combination of laser technology drive and application demand.

Figure 2 shows the current and emerging picture of laser technologies across a core range of wavelengths and pulse durations. Current technologies are shown along with emerging technologies expected to see industrial application in this decade, opening up new applications and markets. Currently established technologies include: CO_2 gas lasers at 10 μm; excimer and KrF gas lasers emitting in the ultraviolet at 248 and 193 nm, respectively; direct diode lasers emitting in the near infrared and visible; and solid state lasers emitting in the near infrared and visible. Current technology developments are centered on extending the range of performance of both gas and solid-state lasers: gas lasers are being extended to the ultraviolet with gases such as F_2; diode lasers are being extended to the ultraviolet with materials such as nitride-based semiconductors and to the infrared with the incorporation of quantum cascade technology; and solid-state diode pumped lasers are being extended to the infrared and ultraviolet via nonlinear conversion. In essence, achievable laser parameter space is seeing continued expansion, not only in terms of pulse duration and wavelength, but also in terms of average power, pulse energy, beam quality, robustness, and cost. Laser technology has now arrived as a major driver for industrial and social change through this decade and into the next.

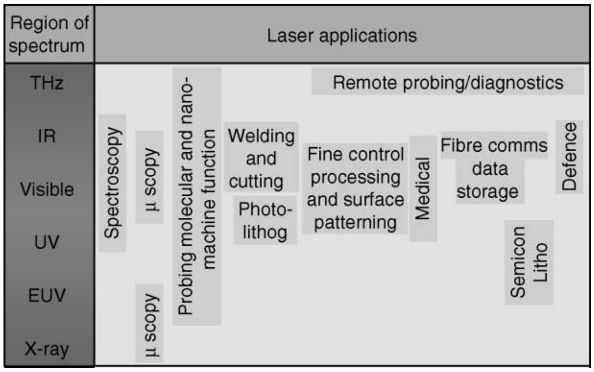

Figure 1 Industrial laser applications across the spectrum.

Infrared wavelengths directly stimulate nuclear motion, coupling to heat without necessary recourse to electronic excitation. Visible wavelengths couple to electronic transitions and at high laser intensities can ionize via multiphoton absorption. Ultraviolet wavelengths couple to electronic transitions and for wavelengths in the region of or below 200 nm, ultraviolet radiation can directly result in bond breaking. For materials processing applications, ultraviolet radiation can have advantages in reducing the laser threshold required for ablation, translating to reduced thermal load imparted to the material.

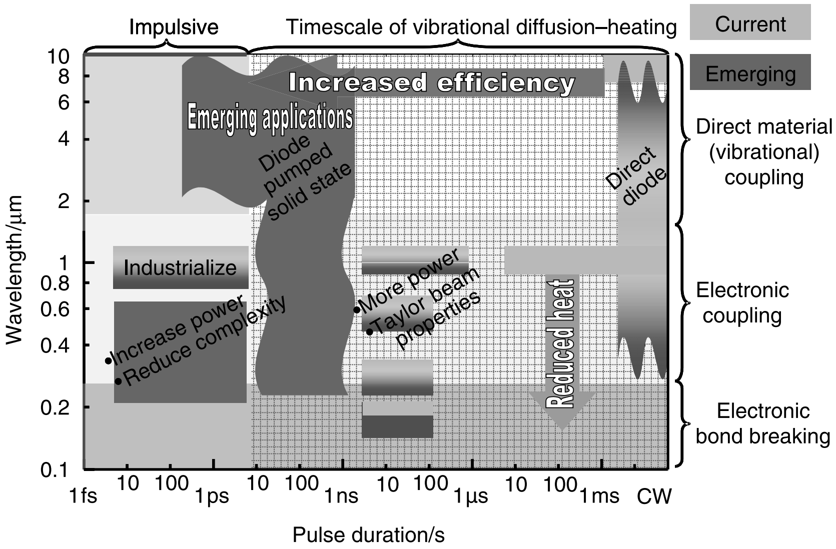

Figure 2 Current and emerging laser technologies across the spectrum as a function of laser pulse duration. Currently established technologies (light gray) are shown with technologies expected to emerge during this decade (dark gray).

Nuclear motions in materials cover a range of timescales of dynamics from 100 ps ($0.3\ \mathrm{cm}^{-1}$) to 10 fs ($3000\ \mathrm{cm}^{-1}$). As the laser pulse duration is reduced below 1 μs, the effects of thermal diffusion are mitigated in solids and as the pulse duration is further reduced below 100 ps, the interaction between radiation and matter is seen to adopt impulsive behavior. Reduction in pulse duration below the microsecond regime can lead to increased efficiency for material ablation combined with improved quality.

Materials Processing – Photochemical

Lithography in Semiconductor Manufacture

Laser technology's role in semiconductor and microelectronics fabrication is growing as manufacturers move towards producing smaller, more powerful devices. The overriding commercial driver for a generation has been to increase the number of integrated components per area of silicon. This has been achieved by reducing the feature size of a transistor on silicon and this trend is anticipated to remain in place through to 2010.

Features on silicon are chemically etched following lithographic exposure. To meet the increasing demand in resolution and throughput, lasers have been applied with a trend in development towards shorter (ultraviolet) wavelengths, with increasing power and reliability.

The semiconductor industry is currently in transition from 248 nm (XeCl gas) to 193 nm (ArF gas) ultraviolet (UV) pulsed laser technology to expose silicon wafers of diameter 30 cm. A typical laser power available at this time at 193 nm, is 40 W at a repetition rate of 4 kHz. To continue the reduction of feature sizes by a factor of two every four to five years, the industry is considering moving to 157 nm (F_2 gas) pulsed laser technology. As wavelengths reach the deep ultraviolet (DUV), transmission through gas is significantly reduced, requiring a move to vacuum-based systems. Associated technologies relating to translation (stepper), mask, and image relay optics also require significant advancement. A technology also being considered at this time is to image the mask to silicon through a medium with refractive index greater than unity, thereby reducing the minimum feature size afforded by diffraction theory, whilst maintaining the same laser wavelength and numerical aperture.

It is estimated that at a feature size of approximately 80 nm, the above technologies will be reaching their resolution limit. This will necessitate a switch to electron beam projection lithography or extreme ultraviolet (EUV, 13 nm) technology. Electron beam projection carries a disadvantage in terms of projected rates of wafer throughput and as such, emphasis has been placed by the industry on the current development of EUV technology. EUV radiation can be produced by laser interaction with a gaseous volume.

Computer to Plate Printing

A common advantage where lasers are replacing current technologies is to reduce the number of associated processes and thereby reduce the cost of process ownership. Computer to plate (CTP) printing involves the laser exposure of a printing plate without the use of a film intermediate. This technology is finding fast traction within the printing industry and is applicable to drum or flat bed platforms. CTP first gained serious attention in the 1990s and has since substantially advanced, with a wide range of technology suppliers (Figure 3).

Traditional methods have involved offset printing plates, imaged used traditional film, and high-power

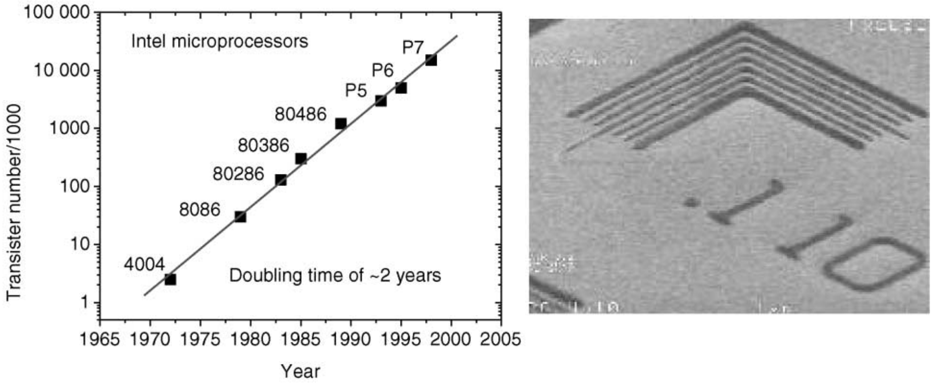

Figure 3 Moore's Law: continuing growth in computing power, showing a doubling in the number of transistors per unit area of silicon every 2 to 2.5 years. Also shown are 110 nm dense lines on silicon (state of the art in 2002). Source: International Sematech.

UV lamps. In contrast, CTP exposes the plate directly with laser radiation. The major saving reported by users of CTP is a faster turnaround time. For color printers, CTP gives almost instant color register on the press and there is a reduced requirement to stop the press to delete dust marks and scratches, as was the case with film. As the industry sees transition to reduced press run lengths, so the importance of CTP is increasing.

Digital plates generally use a high-quality, grained, and anodized aluminum substrate, as for analog plates (see Figure 4). High tolerance electrochemical graining and anodizing ensure stable ink/water balance and good-quality printed results and translates to predictable press behavior. Multiple coatings are applied for CTP, in contrast to a single coating for offset printing, and these take the form of a photopolymer or silver halide and a thermally responsive layer.

On the press, ink and water rollers are continually rolled over the plate. The grained and anodized aluminum substrate serves as a hydrophyllic surface volume to attract water. Ink is oil based and is repelled from areas of water absorbed into the substrate, but is attracted to the hydrophobic dot material.

Silver halide is rated for runs up to 350 000, proving a popular choice for commercial and newspaper applications. It is straightforward to sensitize for a range of writing wavelengths and is the fastest to write.

Thermal plates offer high resolution, equivalent to silver halide and can be baked to enable the longest application lifetimes. Typically the plates are exposed to 830 nm radiation on the external surface of a drum. This technology has served the industry well, but has struggled to find full application in the newspaper business, due to its disadvantage in writing speed.

The most common CTP technology used for newspapers uses a photopolymer coating, which is fast to write and durable. Resolution has traditionally been limited, with Agfa, for example, rating their N91 photopolymer plate to a maximum of 175 lines per inch (LPI), whilst thermal and silver plates are able to image at 200 LPI and above. However, recent advances in shorter wavelength violet diode lasers and violet sensitive polymers have improved resolution. Launched in 2000, violet technology is seeing widespread takeup. This technology is a spinoff from the DVD industry and with laser lifetimes estimated at 10 to 20 years, cost of ownership is one of the drivers. Both silver halide and photopolymers can be sensitized to violet. Current typical laser parameters are for: silver halide, 2–3 μJ/cm^2 at 405 nm; photopolymer, 0.2 mJ/cm^2 at 488 or 532 nm; and thermal, 0.1 J/cm^2 at 830 nm.

The chemical composition of the CTP surface layers is balanced to determine the response to different laser wavelengths, as well as performance characteristics in different print applications. The plate coatings undergo a photochemical change during laser exposure. This is a critical stage since it is important that the plate surface is not under- or over-exposed.

Chemical post processing after exposure is generally required for CTP to date, with the exception of ablative CTP, to remove unexposed material. After imaging, thermal plates go through a pre-bake step, during which the system heats the plate to 130 °C (266 °F) for about 30 seconds. The energy from this

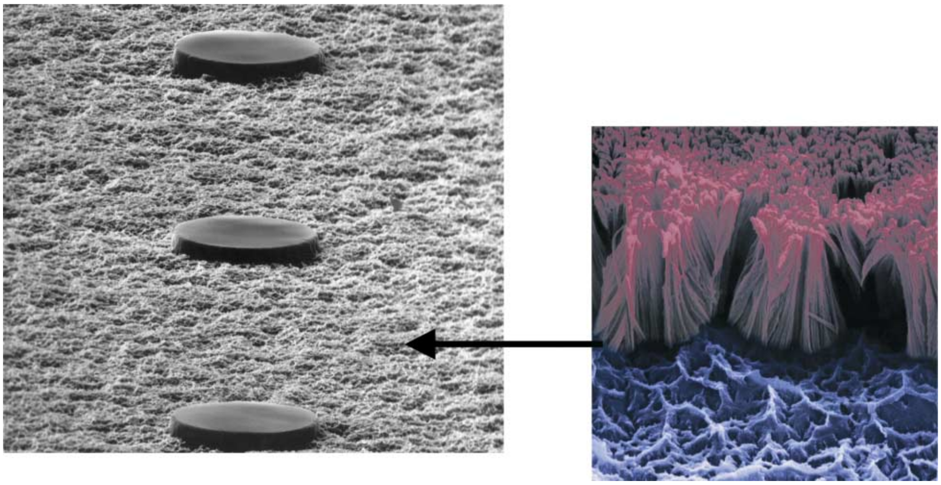

Figure 4 Three typical polymer dots of a few microns depth on an anodized aluminum substrate. Also shown is a further magnified electron microscope image of the substrate material, with the anodized layer cut away for presentation. The porous structure increases water retention. Oil-based ink is repelled from areas other than the dots. Figure courtesy of Agfa (Tony King, tony.king1@agfa.com).

heating stage completes cross-linking of polymer chains in the exposed portions of the emulsion. Chemical processing consists of processing the plate through an alkaline developer solution where the polymer that is not cross-linked washes away. Following the wash is a rinse and application of a coating. After chemical processing, some plates require baking to 200 °C (392 °F) for 35 minutes. Afterwards, the plate continues on to rinsing.

The machine tools required for these processing steps measure 30 feet long and take up to 7 minutes to process a plate after imaging. For both polymer and thermal plate processing, the temperatures are carefully controlled to prevent the aluminum substrate from deforming during the baking stages.

The effluent from thermal and other photopolymer plate processors is fairly benign and is typically directly disposed without significant environmental concern. Spent photopolymer chemistry can, however, pose problems with waste flow, it being a plastic sludge that can solidify and over time, clog plumbing systems.

Silver diffusion plate processors are relatively compact and tend to process plates in less than a minute. Environmental precautions are required with the disposal of accumulated silver particulates. Relatively simple systems are available, which neutralize the pH of the effluent and filter out silver until concentration levels are below 3 parts per million. Under these circumstances, most jurisdictions allow drain disposal.

Ablative plates do not require post processing, other than a rinse to remove residues. This is an emerging technology that has not taken hold at this time in the printing industry and requires more powerful lasers to ablate a layer from the plate. These plates use layers of ink-repellent silicon on top of an aluminum substrate to provide the non-inking portion of the plate. Variations of this technology use a laser to ablate the silicon from the substrate in the image areas of the plate, or to transfer material from a donor layer to the aluminum substrate. A weakness of this technology has been the limited durability of the plates in press. There is continuing development to produce ablative plates with other surface layers that are more durable on press.

Materials Processing – Heating

One way in which lasers can be applied to materials processing is as a directed source of heating. With material quickly raised in temperature to the point of melting and vaporization, applications lie in drilling, welding, cutting, and scintering.

The exploited characteristics of a laser source are its directionality and discrete color. Directionality enables the formation of a restricted area for heating as well as an extended working distance, whilst a discrete color enables absorption to be optimized for the material being worked. Laser technologies typical in application to materials processing through heating, range from diodes at 808 and 940 nm to flash-lamp or diode pumped solid state lasers at 1 μm, and CO_2 gas lasers at 10 μm. A recent entry technology is in diode pumped fiber lasers, reaching output powers in the kilowatt regime at a wavelength of 1 μm and typically being continuous in emission or pulsed with a pulse duration in the millisecond regime.

Direct diode lasers (see Figure 5) have advantages in simplicity and robustness and with the minimum of additional assembled optical components, costs can be expected to continue to reduce, fed by Moore's law for the cost of semiconductor manufacture. Typical powers range from 1 to 10 kW and light can be fiber delivered to the work piece. A disadvantage of direct diode emission is the reduced directionality in comparison with alternative laser sources. For example, a current typical numerical aperture of emission from a diode bar array emitting a power of 60 W, is 10 and 60 degrees in directions parallel and perpendicular, respectively, to the array. These divergences can be reduced to a numerical aperture of approximately one degree with the addition of micro-optics, at a device cost increment of approximately 20%. For an array of diode bars with emitting powers in the kilowatt regime and with emitting dimensions sided 2 cm, a typical optical quality is one thousand time diffraction limited, translating to working distances are on the order of centimeters for a laser spot size at a work piece that is on the order of 1 mm.

CO_2 gas lasers and solid state lasers offer the advantage of reduced divergence and corresponding greater working distance, combined with a smaller laser spot size at the work piece for finer processing.

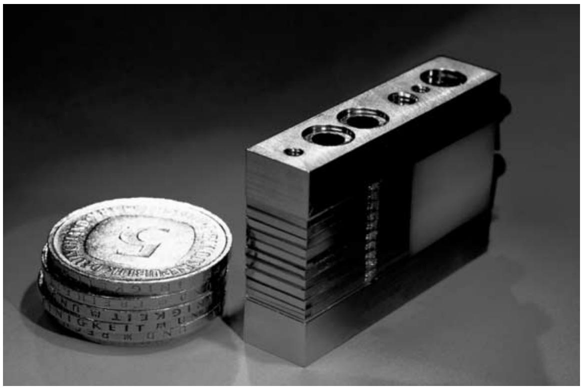

Figure 5 Diode bar array, with an output power exceeding 500 W. Courtesy of Rofin.

Both laser technologies can operate in the kilowatt regime, for application to sheet metal cutting and welding.

With the interaction process involving high temperatures and the ejection of material, a high-pressure gas bathes the area of illumination to remove smelt and either reduce exothermic chemical reactions, or where process acceleration is required, to utilize an assist gas to increase heating through exothermic reaction. Oxygen is typically used as an assist gas for the processing of low-alloy steels, with the process being typically known as laser flame cutting. The gas is, in general, injected coaxially with the laser beam. Laser flame cutting is used in industry for material up to 4 cm thick, albeit the width of the cut becomes significant at this thickness. Both CO_2 and Nd:YAG lasers are applied for this application, the choice of which is dependant on the cut geometry, the cycle time, and the material. Typical cutting speeds are, for example, in the region of 10 meters per minute for 1 mm, 3 meters per minute for 6 mm, and 1 meter per minute for 15 mm thick material.

In the case of high-alloy steels and aluminum, in particular, the material is bathed with an inert gas such as nitrogen or argon. The material is heated solely by the energy in the laser beam, and is typically referred to as laser fusion cutting (see Figure 6). The laser power required is higher than for laser flame cutting. Laser fusion cutting reduces oxidization to the cut edges, which is particularly important where welding is the next process step after cutting.

Laser fusion cutting is used in industry for material up to 25 mm thick, albeit the width of the cut becomes significant at this thickness.

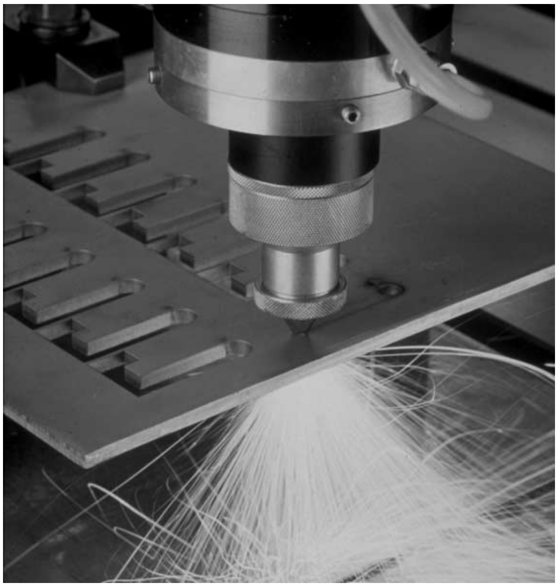

Figure 6 Laser fusion cutting. Courtesy of Rofin.

As with laser flame cutting, both CO_2 and Nd:YAG lasers are suitable for this application, the choice of which is influenced by the geometry of the cut, the cycle time, and the material. Typical cutting speeds in steel are, for example, in the region of ~8 meters per minute for 1 mm, 4.5 meters per minute for 3 mm and 1.5 meters per minute for 8 mm thick material.

Materials Processing – Localized Heating

As a material is heated in a localized volume to the point of vaporization, so also is the surrounding material, which can melt and experience cracking or changes in phase. Such heating is mainly considered deleterious, reducing process tolerances and reducing the tensile strength of the surrounding material.

As the pulse duration is reduced for a localized laser heating source, so are the dimensions that experience heating around the interaction volume (the heat affected zone), translating to improved part quality, with the length for thermal diffusion, l_{th} given by:

$$l_{th} = \sqrt{Dt} \qquad [1]$$

where D is the thermal diffusivity and t, the time of diffusion. As the pulse duration is reduced, the pulse energy required to ablate material is also reduced, with the laser pulse energy required to raise the material temperature to the point of vaporization being proportional to the volume heated. For an example such as silver, a pulse duration of 200 fs is associated with a diffusion length of 6 nm. Extrapolating, a pulse duration of 100 ps, 100 ns, and 100 μs are associated with diffusion lengths of 0.1, 4, and 130 μm, respectively. For typical metals and dielectrics, the process quality and pulse energy threshold for ablation become decoupled from variation in pulse duration for pulse durations less than a few picoseconds. In this regime, the volume for heating is dictated more by the depth for optical absorption and further, the physical mechanism for absorption changes to that of material breakdown and multiphoton ionization (Figure 7).

Additional advantages of operating in the short-pulse regime are a sharpening of threshold to ablation, enabling the processing of features of dimension less than that of the laser spot size and an increase in the number of materials that can be processed by the same color laser source.

Given the advantages of the short-pulse regime, it should be noted that this is currently juxtaposed with the complexity and cost of the required laser sources. Solid-state laser technology is, however, advancing, significantly reducing the number of optical components required to deliver short pulses at multi-kHz

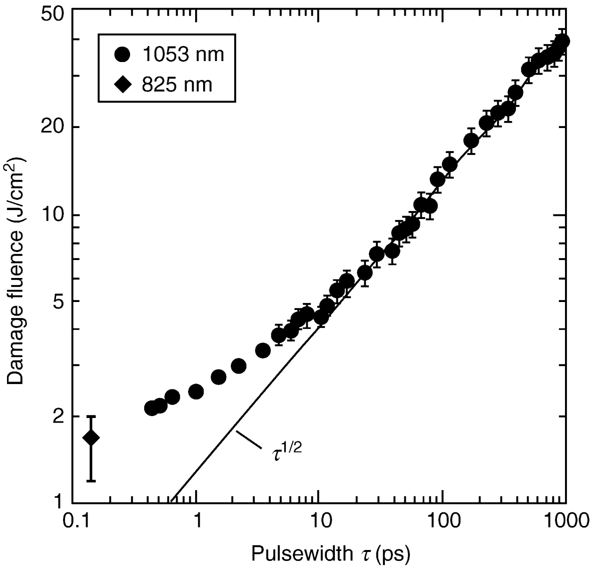

Figure 7 Pulse duration dependence of threshold damage fluence for fused silica. Reproduced with permission from Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53: 1749. Copyright (1996) by The American Physical Society.

repetition rates and multi-mJ pulse energies. At this time of evolution in laser technology, solid-state, diode-pumped high repetition rate (>10 kHz), high pulse energy (>40 mJ) nanosecond (1–200 ns) lasers at a wavelength of 1 μm are seeing increasing application. Such laser sources offer a combination of robustness and cost-effectiveness, combined with a process quality improvement over sources with longer pulse duration.

In the nanosecond regime, the nature of light-matter interaction remains dependent on the wavelength of the laser source and the material properties. In particular, wavelengths in the UV exhibit reduced pulse energy threshold and can offer improved process quality. UV wavelengths less than 300 nm are currently produced by excimer and ArF gas lasers, emitting at 248 and 193 nm, respectively, or solid-state sources that are wavelength converted with nonlinear optical parametric materials to emit at 354 or 256 nm. Significant laser development effort is underway to improve optical parametric materials for nonlinear conversion, to enable greater robustness for solid-state UV sources that then offer reduced overall laser tool cost and complexity.

Semiconductor Lithography

The advantage of nanosecond and shorter pulsed lasers for materials processing extends from localization of heating. With heating being localized, a greater temperature can be achieved. This behavior is fundamental to the production of EUV radiation for application to next-generation lithographic processing.

EUV lithography is considered as an attractive candidate to succeed deep DUV-based optical lithography within this decade. EUV wavelengths around 13 nm will allow the reduction of structure sizes to less than 80 nm in semiconductor devices.

For such a high-power light source that emits at a wavelength of 13 nm, solutions are currently being pursued by multiple industry players in electrical plasma discharge and laser produced plasmas (LPP) (Figure 8). Electron beam and synchrotron sources could, in principle, satisfy the application requirements, but suffer from restriction relating to cost, size, and throughput. The choices of technologies remain to be finalized. It is, however, accepted across the industry that Moore's Law should remain intact and that a solution is required that is scalable to high throughput rates for wafer exposure.

For a laser-produced plasma, the peak wavelength of the light emitted from a plasma is determined by its temperature, T, where according to the blackbody formula:

$$T \sim E_{\mathrm{ph}}/3 \qquad [2]$$

where E_{ph} is the photon energy associated with the wavelength of maximum emission. Liner photopolymerization (LPP) schemes use pulsed laser radiation to heat a target to the optimum temperature required for efficient EUV production. For plasma-emitted light which peaks at a wavelength of 13.5 nm, the optimum temperature is approximately 3.5×10^5 Kelvin (30 eV).

A laser is focused onto a target, transferring its energy into the target in such a way that a plasma is formed and heated. A range of targets have been investigated for use in LPP schemes ranging from gaseous and cluster targets, to liquid targets and macroscopic solid targets. The interaction begins with the production of free electrons via multiphoton ionization, which is strongly dependent on the peak intensity of the laser. As heating progresses, there is further ionization from inelastic collisions between heated electrons and ions to reach a pseudo-equilibrium state, where cooling competes with heating on nanosecond timescales. Given this time-scale for cooling and requirements on source size and temperature, the optimum laser pulse duration lies in the regime of nanoseconds and the optimum laser pulse energy is hundreds of millijoules. Laser technology

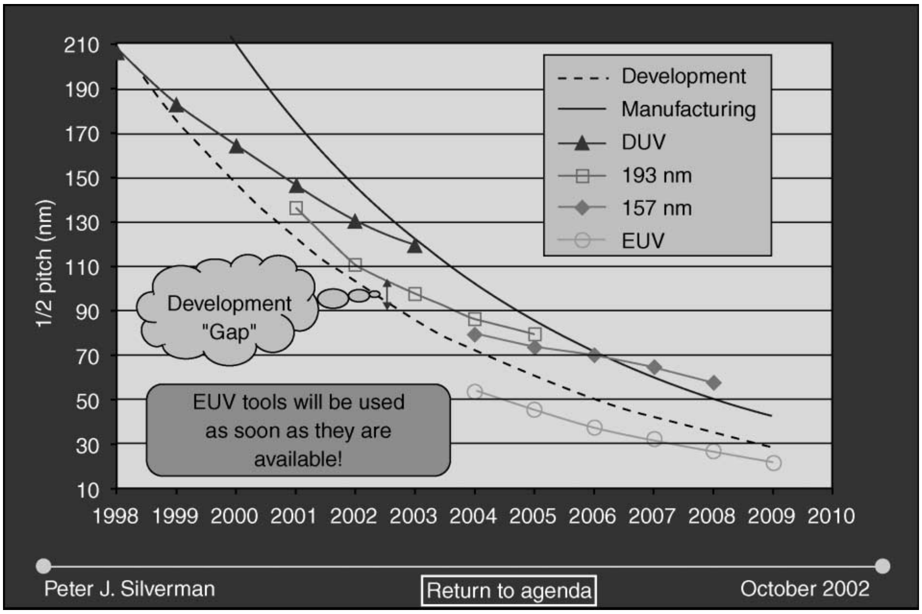

Figure 8 Light source roadmap to 2010. Courtesy of Intel Corporation.

suited to this application is diode-pumped Nd:YAG, emitting at a wavelength of 1 μm and pulsed at multikilohertz repetition rates.

Fine Control Material Removal

Localized heating afforded with the application of lasers with pulse durations in the nanosecond regime and shorter, is seeing increasing takeup to replace processes ranging from wet etching and electro-discharge machining (EDM) in applications such as scribing, drilling, and differential layer removal. Laser methods are seeing advantage in direct ablation of material, for feature sizes on the order of microns to hundreds of microns. Large areas on the order of 2 m^2 are now becoming cost-effective to process with direct laser ablation, with the advent of high average power laser sources of hundreds of watts of average power, combined with nanosecond regime or shorter pulse durations. Typical sources are gas excimer or diode-pumped solid-state.

The finest processing is advantaged by the application of pulse durations in the picosecond regime. A restriction in application of this technology has been robustness and cost of the laser sources to date; however, such technology is continuing to see rapid improvement and may be expected to see wider takeup in industry by 2010.

Also seeing application, are wavelengths in the UV regime combined with nanosecond pulse durations. UV radiation in general offers a reduced threshold of ablation, translating to reduced heating and improved process quality. High absorption at UV wavelengths in most materials, combined with photon energies that can exceed bond energies of the material, contribute to the removal of material with the minimum of heating. Excimer, KrF, and ArF gas lasers are applied here (Figure 9), but it is expected that solid-state lasers will emerge by 2010 to capture a large part of the UV market. CO_2 gas lasers are used for fine processing where the material being ablated is of low thermal diffusivity, which is typical for dielectrics or ceramics. Materials of low thermal diffusivity are associated with a low heat affected zone (see eqn [1]).

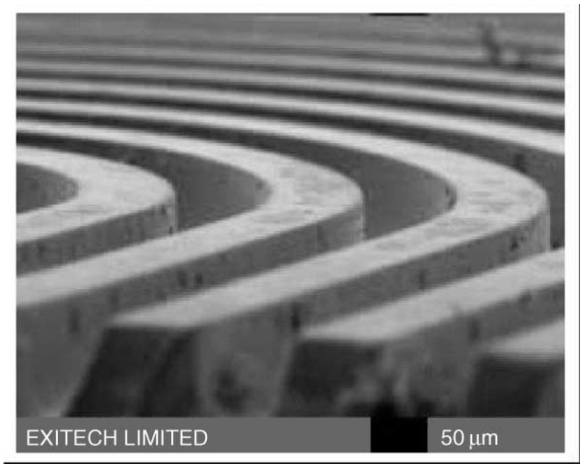

Figure 9 KrF laser micromachined microfluidic channels in polyester for application to lab-on-a-chip technologies or sensing devices. Courtesy of Exitech Ltd.

Scribing is required for dicing of wafers such as silicon and other low-*k* materials, both for computing device manufacture and for solar cell manufacture. Nanosecond and shorter pulsed lasers have seen recent application in this area, enabling higher aspect ratio scribing and a reduced material affected zone. This application has seen laser wavelengths applied from the UV to near-infrared. Short pulse laser scribing can see advantages of high throughput yet reduced or eliminated chipping, delamination, and film peeling.

Fine drilling is required for a wide variety of applications and markets ranging from constructing vias for through-layer interconnects in printed circuit board (PCB) fabrication, to ink jet nozzles (Figure 10) and high aspect ratio holes in hard materials, such as steel or alloys for application to combustion engines.

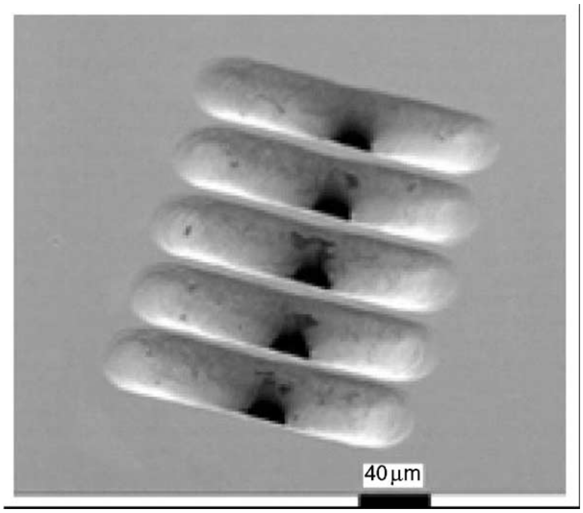

Figure 10 A laser drilled hole in steel and an array of shaped holes in polyamide for inject nozzle application. Courtesy of Exitech Ltd.

Excimer and CO_2 laser drilling has seen industrial application in the drilling of via holes for electronics interconnection packages, since early investigations in the 1980s. More recently, high repetition rate diode-pumped solid-state lasers, nonlinearly converted to emit in the UV, are seeing application. In parallel with the increasing density of transistors in silicon processors, the density of packaging of multichip modules (MCM) and PCB components has also been increasing, requiring the drilling of blind vias on the order of 100 microns in diameter to enable interconnects between multiple conducting layers (Figure 11). Also, as processor power increases, so does the number of connections required to devices. Microvias have enabled the application of two-dimensional grid array connections to the underside of devices, reducing overall package size, including processor electrical connections, by a factor of five.

Laser drilling of microvias in PCBs is expected to increase as packing density increases. PCBs remain cheaper than MCMs and traditionally have been mechanically drilled. As via sizes are reduced below 100 µm, however, it is expected that pulsed CO_2 and Nd:YAG lasers will be applied.

Thin layer ablation is required for applications ranging from surface cleaning to display manufacture and for polymers, dielectrics, and metals. For example, manufacture of plasma displays currently incorporates lithography combined with wet etching, to pattern layers of thin conducting oxide onto a substrate. Multiple process steps can be replaced by application of direct laser ablation and increased laser average power combined with pulse durations in the nanosecond regime are allowing millions of cells to be processed per minute, with the accommodation of process areas currently greater than 2.5 m^2 (see Figure 12).

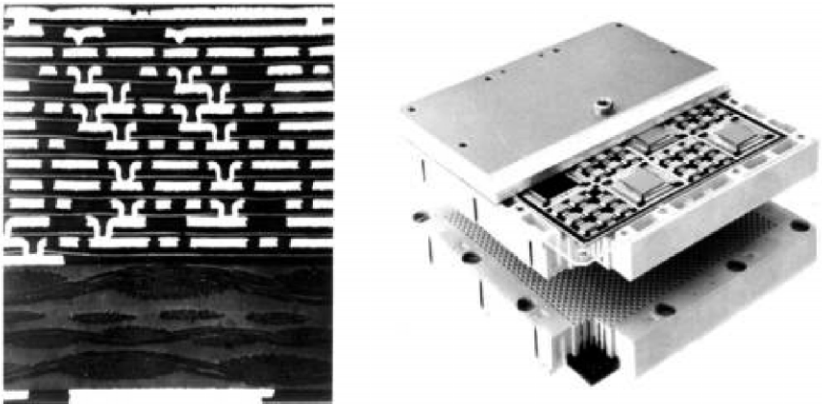

Figure 11 Cross-section of 14-layer copper plated MCM board showing 80 µm diameter microvias drilled by a KrF laser in acrylic resin-polyimide and, fully assembled MCM package. Courtesy of Siemens AG.

Laser-Based Diagnostics

With a continuing increase in reliability, performance, and range of specifications, lasers are seeing growing industrial application for process monitoring and diagnostics, ranging from gas phase detection of molecular specific concentrations to *in vitro* microscopy in the life sciences. Markets are emerging, driven in a large part by advances in laser technology. Chemical sensing with application of lasers is a vast field, which is difficult to cover here completely. Following, are some applications seeing current industrial application and growth.

Microscopy

Confocal microscopy has seen large uptake, in particular, for the application to imaging of cell and membrane function, with more than fifteen manufacturers now supplying confocal microscope tools including Zeiss, Biorad, Leica, Olympus, and Nikon. Confocal microscopy enables increases in resolution and contrast compared to standard microscopy. Techniques such as fluorescence resonant energy transfer (FRET), fluorescence recovery after photobleaching (FRAP), fluorescence loss in photobleaching (FLIP), fluorescence lifetime imaging (FLIM), or fluorescence correlation spectroscopy (FCS), where married with confocal microscopy and, in particular, where relevant molecular structures are known in advance, can offer the additional advantages of specific molecular and structural specificity through the incorporation of tracer fluorophores. This is seen as color coding of the resultant image, with sophisticated microscopes currently offering differential imaging stimulated by three laser colors and with six selectable colors for detection.

The confocal microscope operates with advantage over a standard microscope, by illuminating only a small area in the sample with dimensions of the resolution limit of the objective and re-imaging emission from that area through a pinhole spatial filter (Figures 13 and 14). Structures outside of the illuminated area or at a depth outside that of the focus, are suppressed from detection. Fluorophores (tracers) can be inserted into protein sequences at predetermined locations and are stimulated by either single or two-photon absorption of the stimulating laser light.

Two photon absorption has the additional advantages of increased resolution and detected contrast over single photon absorption. Laser sources applied are generally modelocked oscillators operating near 800 nm, with a pulse duration on the order of 100 fs. This technology is continuing to improve in robustness, with fiber and direct diode pumped solid-state sources available from multiple vendors.

FRAP and FLIP monitor diffusion or transport of fluorescent labeled molecules into or out of the laser focal volume, respectively. FCS monitors the fluctuations in detected signal from emissions within the laser focal volume, to determine the rate of diffusion with sensitivity achievable to the single molecule level (Figure 15).

FRET is seeing application for the probing of protein interactions in cells or at cell membranes. FRET involves the nonradiative transfer of excitation energy from one fluorophore to another, which subsequently emits at a different wavelength. The fluorophores are inserted into molecules, the

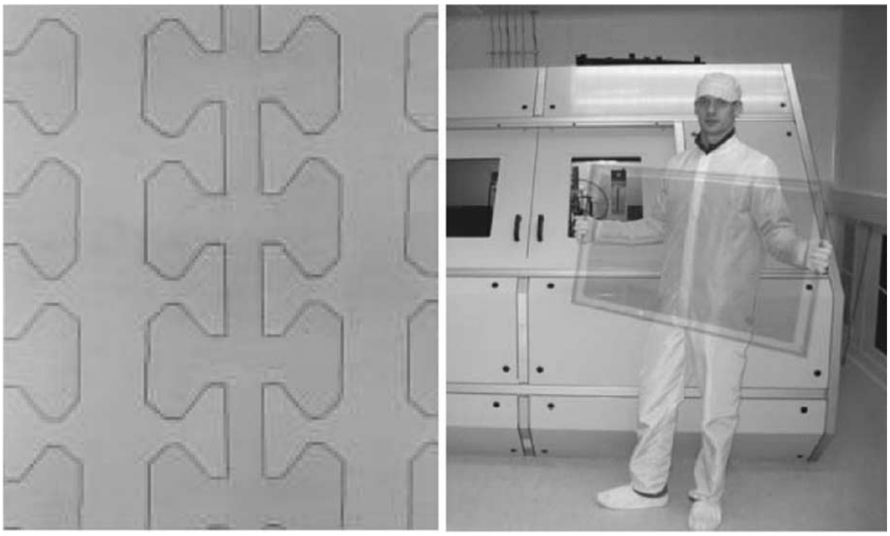

Figure 12 Plasma display manufactured by direct laser ablation of thin-film indium tin oxide (ITO). Courtesy of Exitech Ltd.

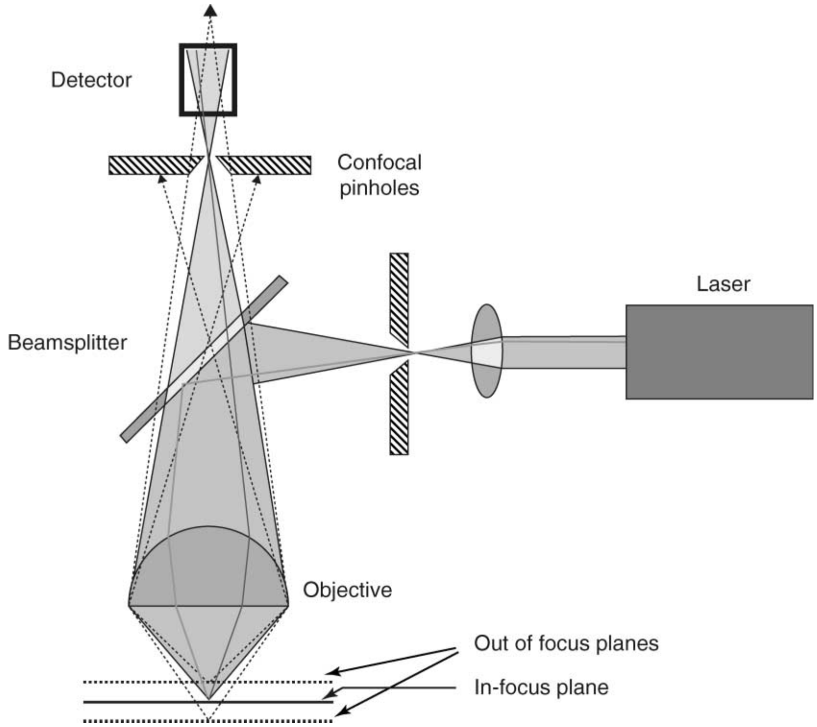

Figure 13 Confocal microscope principle of operation. Courtesy of Leica Microsystems Heidelberg GmbH.

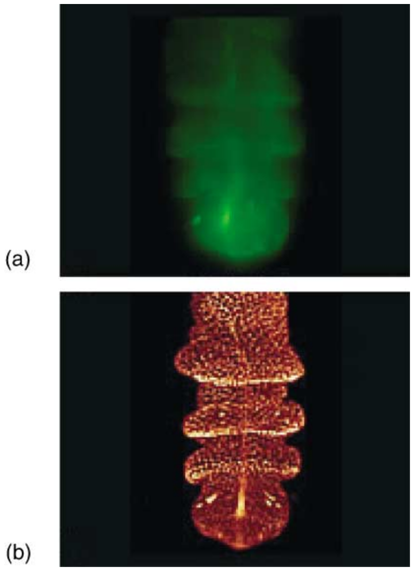

Figure 14 Image (a): *Drosophila melanogaster*, fluorescence nonconfocal configuration and Image; (b): *Drosophila melanogaster*, fluorescence confocal 3D projection. Courtesy of Leica Microsystems Heidelberg GmbH.

interaction of which is being investigated. FRET effectively enables sensitivity to relative position at fluorophore separations less than 10 nm; a significant advance over the 200 nm resolution provided by a standard optical microscope. Concentrations of binding events can be monitored over time (Figure 16).

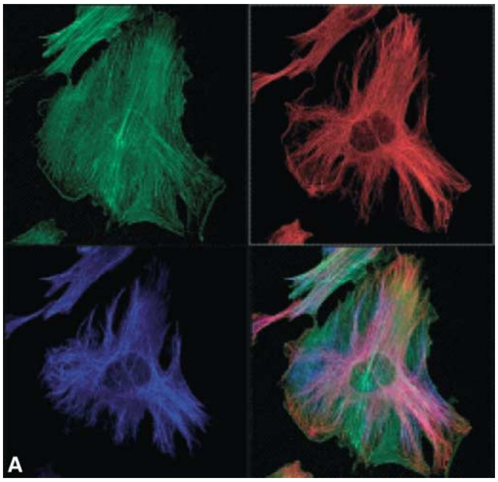

Figure 15 Spectral selective imaging with three fluorescence channels showing triple labeled mouse fibroblast: green: actin fibers, red: microtubules, blue: vimentin. Courtesy of Leica Microsystems Heidelberg GmbH.

Raman Methods and Microscopy

Raman spectroscopy sees broad application in chemistry in the analysis of chemical composition and molecular structure, being used, for example, in pharmaceuticals, polymer materials science, bioclinical science, and forensic science.

The history of Raman scattering stems from theoretical predictions in the early 1920s,

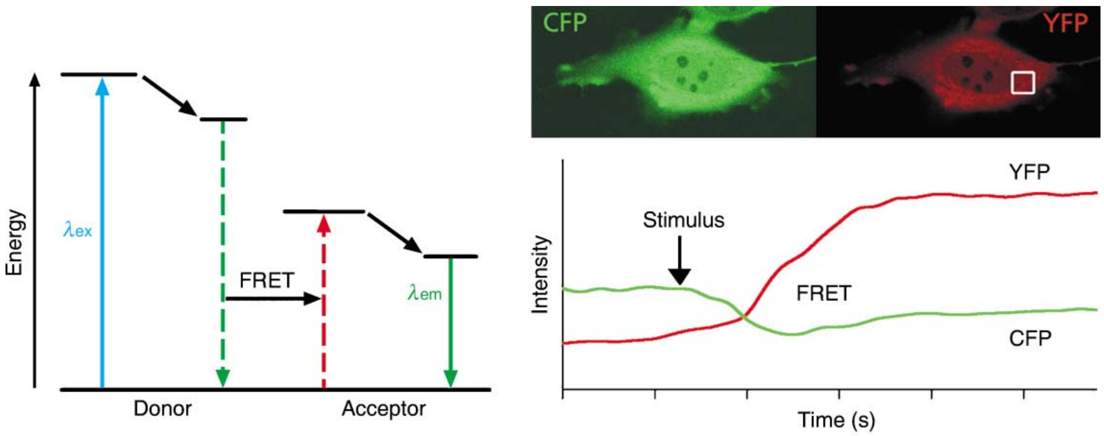
CFP
YFP
Energy
λex
FRET
λem
Donor
Acceptor
YFP
Stimulus
Intensity
FRET
CFP
Time (s)

Figure 16 Fluorescence resonant energy transfer mechanism and example relating to cytoplasmic Ca^{2+} concentration. Cyan Fluorescent Protein (CFP) and Yellow Fluorescent Protein (YFP) are acceptor and donor, respectively, for the FRET. Courtesy of Carl Zeiss, Germany.

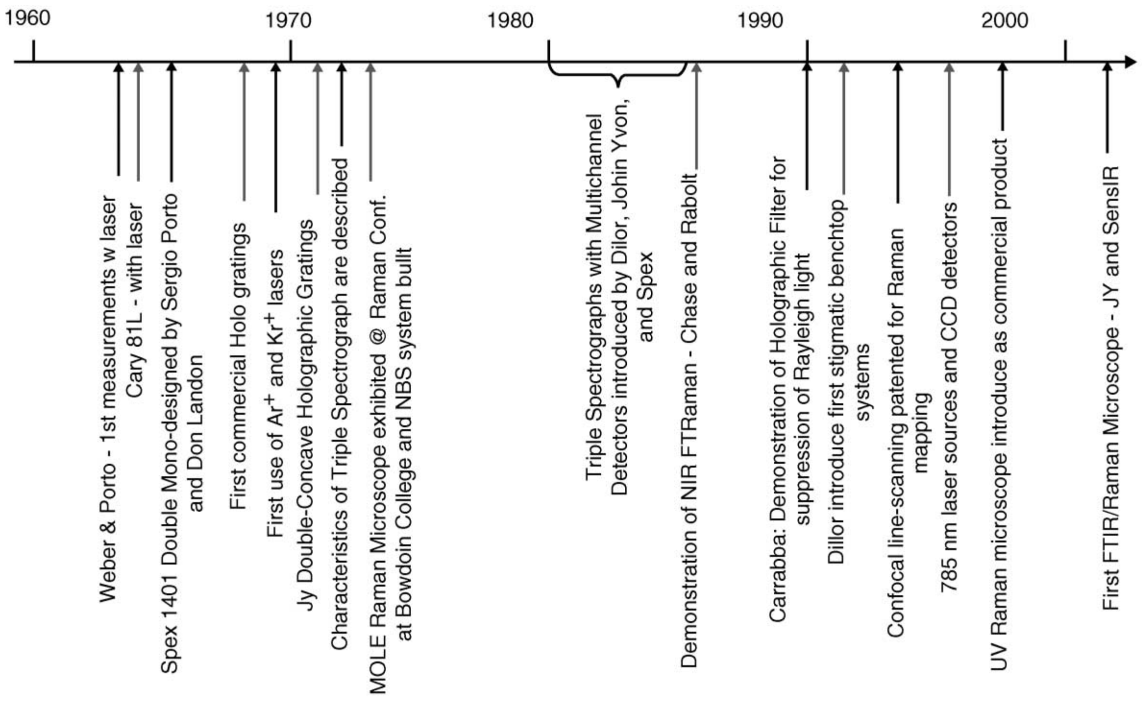
1960
1970
1980
1990
2000
Weber & Porto - 1st measurements w laser
Cary 81L - with laser
Spex 1401 Double Mono-designed by Sergio Porto and Don Landon
First commercial Holo gratings
First use of Ar^+ and Kr^+ lasers
Jy Double-Concave Holographic Gratings
Characteristics of Triple Spectrograph are described
MOLE Raman Microscope exhibited @ Raman Conf. at Bowdoin College and NBS system built
Triple Spectrographs with Multichannel Detectors introduced by Dilor, Johin Yvon, and Spex
Demonstration of NIR FTRaman - Chase and Rabolt
Carrabba: Demonstration of Holographic Filter for suppression of Rayleigh light
Dillor introduce first stigmatic benchtop systems
Confocal line-scanning patented for Raman mapping
785 nm laser sources and CCD detectors
UV Raman microscope introduce as commercial product
First FTIR/Raman Microscope - JY and SensIR

Figure 17 Timeline of Raman spectroscopy from the first incorporation of a laser source. Courtesy of Jobin Yvon.

accompanied also in the 1920s by early measurements, notable being those of C.V. Raman for which he was awarded the Nobel prize, soon followed in the 1930s by hundreds of chemistry papers describing this technique. Raman used for his early experiments filtered sunlight, a prism spectrometer, and visual observation. This was followed by application of filtered Hg emission and photographic plates. Lasers first saw application to Raman scattering in the 1960s, shortly after their innovation (see Figure 17).

In the simplest form of application of the Raman effect, UV or visible radiation illuminates a sample, coupling to excite electronic transitions near or on resonance. Spontaneous emission is shifted to a longer wavelength than that of the excitation (Stokes emission) and this is filtered and detected. In general, a broad spectrum of fluorescence is superimposed by weaker, narrow features that originate from vibration-electronic coupling (Raman emission). Inclusion of laser excitation increases the rate of data collection and also enables nonresonant interaction of the stimulating radiation with the sample whilst maintaining data collection rates. Nonresonant interaction preferentially reduces electronic emission of fluorescence relative to that of the narrow featured Raman emission. Lasers applied to spontaneous Raman spectrometers in general emit in the visible or near-infrared running nonpulsed, such as Ar-ion, Nd:YAG or He:Ne.

Further enhancements are made to deliver spatial as well as vibration-specific information, through tight focusing of the stimulating radiation on to the sample in a confocal microscope arrangement, delivering spatial resolution down to 200 nm (Figure 18).

A limitation of standard Raman spectrometers and microscopes is the rate of signal collection, requiring concentrated samples for interrogation. The restricted signal level derives from a typical cross-section for the spontaneous emission of Raman radiation being $10^{-30}\,\mathrm{cm}^2$; 10^{10} to 10^{14} times less than that typical for spontaneous electronic fluorescence. Enhancements in the level of signal and the ratio of signal to background noise may be made through coherent laser stimulation of anti-Stokes Raman emission, known as coherent anti-stokes Raman scattering (CARS) (Figures 19 and 20).

It can be anticipated that bond-specific imaging will see wide industrial application in future, employing Raman as well as resonant infrared techniques. Application of CARS to microscopy involves two laser colors. A vibration transition is exited at the beat frequency of the two laser frequencies and the shorter laser wavelength is subsequently scattered from the sample to a yet shorter wavelength of emission. Such a coherent process has the advantage of signal scaling proportional to the square of emitters at the sample, thereby increasing the signal over that of spontaneous Raman scattering. Typical laser pump and Stokes powers range from less than 1 mW to 10s of mW, with laser repetition rates ranging from multi-kHz to greater than 100 MHz and typical pulse durations from 10s of picoseconds to 100 fs.

Infrared Methods

Infrared absorption spectroscopy provides a wealth of information on bond-specific sample concentration

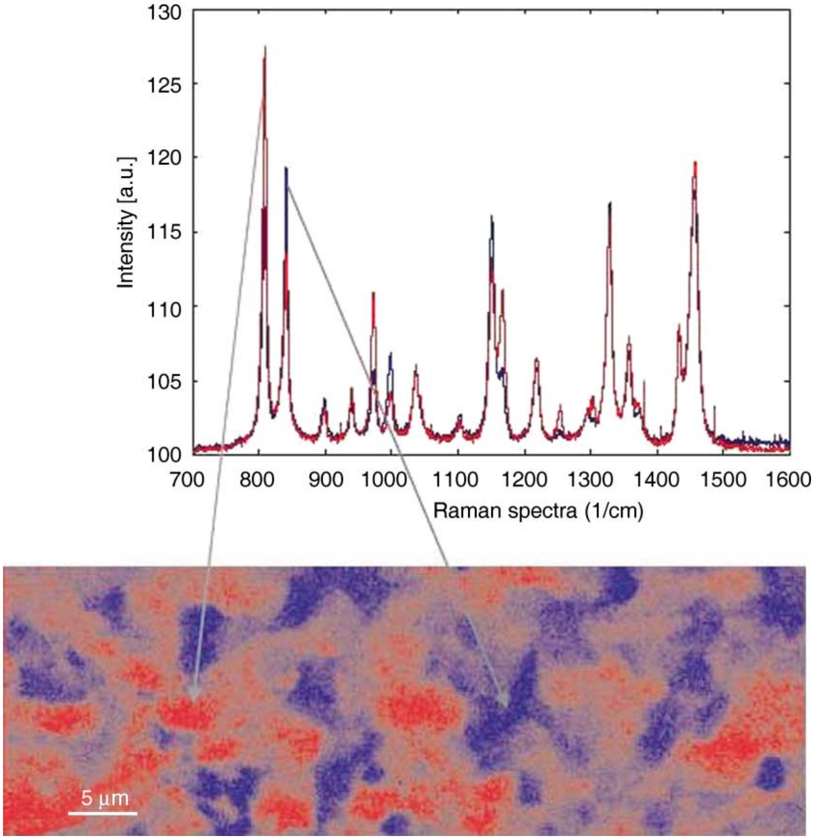

Figure 18 Confocal Raman image of semicrystalline polypropylene film. Two typical Raman spectra are shown. The ratio of the Raman lines at 809 and 841 cm increases with the degree of crystalline order, delivering a two-dimensional image (Hendra *et al.* 1995). Courtesy of Witec GmbH.

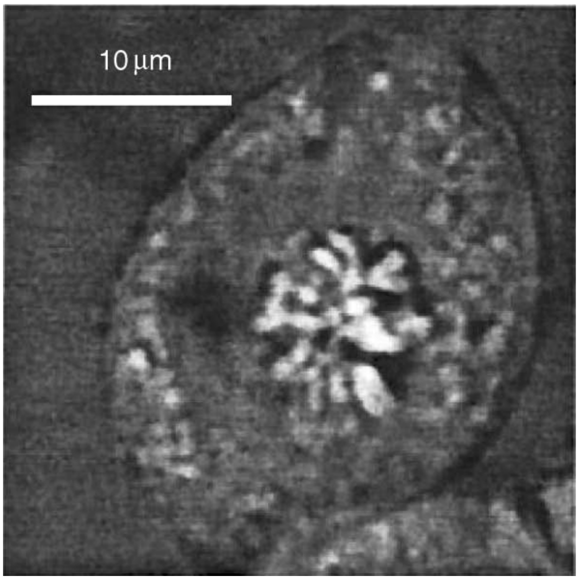

Figure 19 CARS image of a NIH 3T3 cell in metaphase, where Raman vibration stimulation is tuned to the PO_2^- symmetric stretching vibrational frequency at 1090 cm^{-1}. Reproduced from Cheng J-X, Kevin JYK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509, with permission from The Biophysical Society.

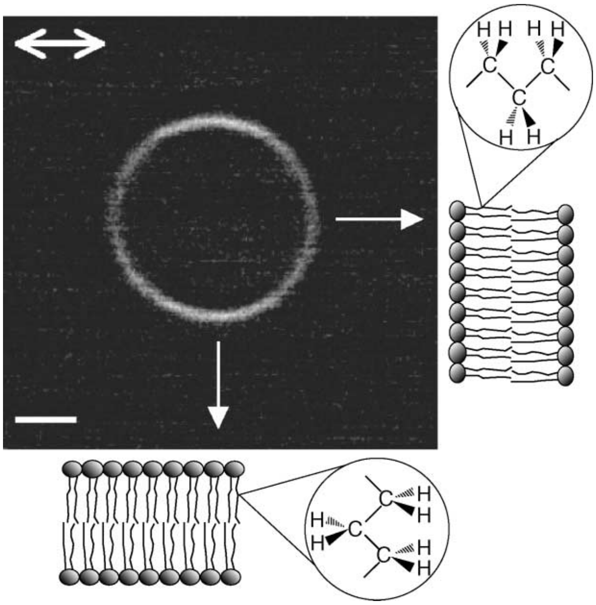

Figure 20 CARS microscope image of an erythrocyte ghost, consisting of a single lipid bilayer, with a Raman shift of 2845 cm^{-1}. The scale bar is of length 2 μm. Reproduced from Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 36: 642–650, with permission from Wiley InterScience.

and has become a standard tool for the chemistry laboratory. Interaction between the radiation and sample is via resonant vibrational absorption and is otherwise analogous to the operation of a standard optical absorption spectrometer. Applied to a microscope arrangement, resolution on the order of 10 μm is standard. One restriction associated with infrared absorption spectroscopy applied to solution, is that absorption lineshapes are broadened and the spectra lose detail at room temperature, hindering identification of the contributing species.

A technique showing early promise for application in bond-specific molecular detection in solution, is coherent resonant infrared wavemixing. Here, cross coupling between vibrational modes are detected, promising information more detailed than that of infrared absorption. Figure 21 shows an example of triply vibrationally enhanced four-wave mixing, where two infrared laser beams mix to deliver cross peaks representing vibration mode coupling. Here, the 2D spectrum contains two sets of cross-peaks corresponding to the couplings of the two strong IR modes at 1943 and 2002 cm^{-1} of $Ni(CO)_2(PPh_3)_2$ and 1996 and 2068 cm^{-1} of $Ni(CO)_3PPh_3$. Concentrations in solution are on the order of 10 mM. Such methods provide a means sensitive to chemical structure without a requirement for fluorophores.

Current application requires complex laser equipment delivering microjoule pulse energies, tunable between 3 and 10 microns, with pulse durations on the order of 1 ps. However, following the example of other laser diagnostic techniques in application

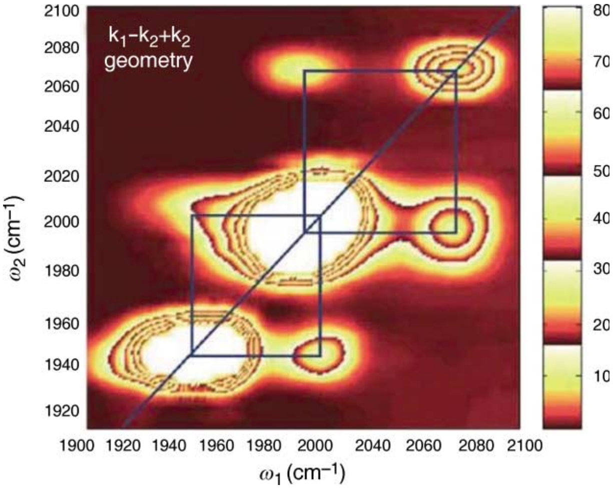

Figure 21 Triply vibrationally enhanced four-wave mixing of $Ni(CO)_2(PPh_3)_2$ and $Ni(CO)_3PPh_3$. Reproduced from Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649, with permission of Elsevier.

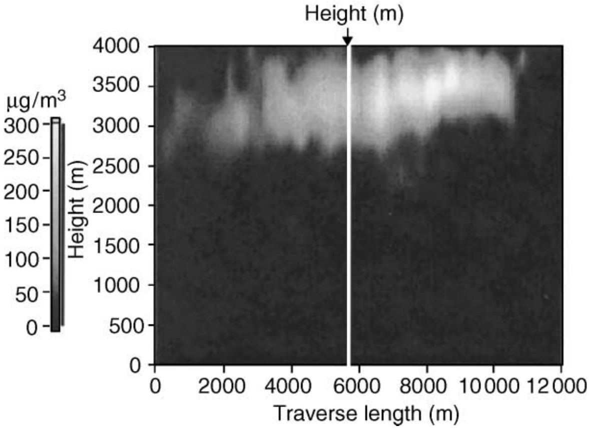

Figure 22 DIAL trace of sulfur dioxide distribution from Mount Etna. Reproduced with permission from Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42: 3583–3594. © 2003 Optical Society of America.

across industry today, as laser technology advances over this decade, we can expect such methods of probing chemical systems to see significant uptake, in particular in the bio-industries.

Laser Diagnostics and Treatment of Tumors

For clinical medicine, fluorescent detection is applied in the visualization of malignant tumors. Important natural chromophores emitting fluorescence in the blue-green region are elastin, collagen, NADH, and NAD^+, whilst administered agents such as porphyrins and phthalocyanines yield sharp peaks in the dark red.

Excitation is via UV or violet light. Where emission is monitored in both the blue-green and red regions of the spectrum and the ratio taken, an enhanced tumor differentiation is possible from that of normal tissue whilst simultaneously reducing sensitivity to geometry, illumination, and detection efficiency. Fiber optic probes are generally applied for delivery and collection of radiation.

Such detection can be combined with techniques of laser photodynamic therapy (PDT) to preferentially remove cancerous tissues. Typical laser wavelengths for photodynamic therapy are in the visible region of 600 to 700 nm with a pulse fluence of tens of joules/cm^2, a pulse duration in the nanosecond regime and a repetition rate from single shot to 10 Hz. Where combined with diagnostic detection, the PDT can be carried out by exciting the same administered sensitizer, with subsequent transfer of ground-state triplet oxygen to a toxic singlet state.

Remote Chemical Sensing

The atmosphere can be monitored by absorption and laser-induced fluorescence (LIF). Three-dimensional mapping of chemical distribution can be achieved with light detection and ranging (LIDAR), using a pulsed laser as a transmitter and an optical telescope as a receiver. Information collated includes pollution concentration, temperature, humidity, and wind velocity. Strong quenching prohibits LIF at atmospheric pressures; however, it does become practical to monitor for example, Li, Na, K, and Ca layers in the mesosphere. Pollution monitoring in the troposphere is most frequently performed with differential absorption LIDAR (DIAL). Range resolved optical transients are taken as a ratio of two laser wavelengths; on and just off resonance for the species of interest. Typical ranges for detection of SO_2, O_3, NO_2, NO, and Hg are 500 m to 5 km, making this technique suitable for urban and industrial monitoring. Sulfur dioxide is the most common pollutant emitted when fossil fuels are burned or sulfide ores are roasted and this technique sees application for the monitoring of industrial emissions. Figure 22 shows a DIAL trace recorded from Mount Etna, from which emissions of SO_2 are estimated at 75 tonnes/h. Laser wavelengths used are 300.02 and 299.30 nm for on and off resonance for absorption, respectively.

Important to the widespread application of LIDAR and LIF for atmospheric chemical monitoring, has been the advance in robustness and flexibility of the associated laser technologies. The laser technology applied for the above measurement is based on solid-sate Nd:YAG, nonlinearly converted via optical parametric amplification to wavelengths tunable between 220 nm and 4 μm, delivering up to 20 mJ in the mid-infrared at a repetition rate of 20 Hz and with bandwidths of less than 0.2 cm^{-1}, in a pulse duration of 4 ns.

See also

Imaging: Infrared Imaging. **Lasers:** Carbon Dioxide Lasers. **Microscopy:** Confocal Microscopy. **Scattering:** Raman Scattering. **Spectroscopy:** Raman Spectroscopy.

Further Reading

Allen SD, Bass M and Teisniger ML, Comparison of pulsed Nd:YAG and pulsed CO2 lasers for hole drilling in printed circuit board materials. CLEO 1982 Conference Summary.

Cheng J-X, Kevin Jia YK, Zheng G and Xie X-S (2002) Laser-scanning coherent anti-Stokes Raman scattering microscopy and applications to cell biology. *Biophysical Journal* 83: 502–509.

Cymer Inc., http://www.cymer.com

Cymer Inc., ASML, TRW, Xtreme Technologies GmbH, Fraunhoffer Institut fur Lasertechnik, Philips Extreme UV GmbH, Royal Institute of Technology Sweden, PLEX LLC.

Disco Corporation, http://www.disco.co.jp/

Exitech Ltd., http://www.exitech.co.uk/

Lamda Physik, http://www.lambdaphysik.de/

Meyer KA, Besemann DM and Wright JC (2003) Coherent two dimensional spectroscopy with triply vibrationally enhanced infrared four-wave mixing. *Chemical Physics Letters* 381: 642–649.

Perkin Elmer, www.perkinelmer.com

Phillips D (1997) Chemical mechanisms in photodynamic therapy with phthalocyanines. *Progress in Reaction Kinetics* 22(3–4): 175–300.

Potma EO and Xie X-S (2003) Detection of single lipid bilayers with coherent anti-Stokes Raman scattering (CARS) microscopy. *Journal of Raman Spectroscopy* 34: 642–650.

Powerlase Ltd., http://www.powerlase.com

Presstek Inc., Hudson, NH, http://www.presstek.com

Pronko PP, Dutta DK, Squier J, Rudd JV, Du D and Mourou G (1995) Machining of sub-micron holes using a femtosecond laser at 800 nm. *Optics Communications* 114: 106–110.

Reiser A (1989) *Photoreactive Polymers – The Science and Technology of Resists*. John Wiley & Sons.

Silverman PJ (2002) Intel Corporation. Presented at 1st Intl EUV Lithography Symposium, 15–17 October 2002, Dallas, USA.

Stuart BC, Feit MD, Herman S, Rubenchik AM, Shore BW and Perry MD (1996) Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. *Physics Review B* 53(4): 1749.

Svanberg S (1996) Chemical sensing with laser spectroscopy. *Sensors and Actuators B* 33: 1–4.

Wang I, Bendsoe N, Klinteberg CAF, *et al.* (2001) Photodynamic therapy vs. cryosurgery of basal cell carcinomas: results of a phase III clinical trial. *British Journal of Dermatology* 144(4): 832–840.

Watson MN (1984) Laser drilling of printed circuit boards. *Circuit World* 11: 13.

Waynant RW (ed.) (2002) *Lasers in Medicine*. CRC Press.

Weibring P, Edner H and Svanberg S (2003) Versatile mobile lidar system for environmental monitoring. *Applied Optics* 42(18): 3583–3594.

WITEC, WITec Wissenschaftliche, Instrumente und Technologie GmbH, Hörvelsinger Weg 6, D-89081 Ulm, Germany, www.WITec.de

Yvon J, www.jobinyvon.com

Sum-Frequency Generation at Surfaces

M B Raschke and Y R Shen, University of California, Berkeley, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Under sufficiently intense illumination the optical properties of a material system depend nonlinearly on the strength of the electromagnetic field. Related to this nonlinear optical response is a large number of phenomena and fundamental processes which are discussed in various articles of this encyclopedia. This article will focus on second-harmonic generation (SHG) and sum-frequency generation (SFG) for the investigation of surfaces and interfaces. The intrinsic surface sensitivity of these techniques allows for investigations of surface properties not readily accessible by other spectroscopies. Here, the basic principles of these optical processes as well as their experimental implementation are discussed, and a summary of the applications to different material systems is given.

Among the various techniques employed for the characterization of surfaces and interfaces, those using light are particularly attractive. They are applicable *in situ* to all interfaces accessible by light, are nondestructive, and offer unprecedented time resolution. However, the penetration depth of optical radiation in matter is generally of the order of a wavelength, which makes isolation of the surface or interface contribution to the optical response from the bulk contribution difficult. In contrast, for reasons of symmetry, SHG and SFG are intrinsically surface sensitive in media with inversion symmetry, and hence the signal generated mainly originates from the topmost surface layer where the inversion symmetry is broken. By means of electronic or vibrational SHG or SFG spectroscopy, information on surface structure, chemical composition and bond or molecular orientation at solid and liquid interfaces can be deduced. To date, SFG and SHG have been well established as important tools for the investigation of surfaces and interfaces of solids ranging from metals and semiconductors to insulators and magnetic materials, liquids, polymers, biological membranes, and other systems. The studies are motivated by fundamental interests as well as applications in many areas such as heterogeneous catalysis, electrochemistry, device fabrication, epitaxial growth, and environmental science.

Theory

The nonlinear optical response of a material, in the electric-dipole approximation, can often be described by an induced polarization in the form of a power series,

$$\begin{aligned}\boldsymbol{P}(\omega) = \epsilon_0[&\chi^{(1)}(\omega)\boldsymbol{E}(\omega)\\ &+\chi^{(2)}(\omega=\omega_1+\omega_2)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)\\ &+\chi^{(3)}(\omega=\omega_1+\omega_2+\omega_3)\boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)\boldsymbol{E}(\omega_3)+\ldots]\end{aligned} \qquad [1]$$

where $E(\omega_i)$ is the optical field with frequency ω_i and $\chi^{(n)}$ is the nth-order linear ($n=1$) or nonlinear ($n>1$) susceptibility tensor. The second term, being the lowest-order nonlinear optical response, is responsible for sum-frequency generation and second-harmonic generation (with $\omega_1=\omega_2$). In the electric-dipole approximation, all even-order terms ($\chi^2, \chi^{(4)}$, etc.) are forbidden in media with inversion symmetry. At the surface or interface, however, the inversion symmetry is necessarily broken and hence $\chi^{(2)} \neq 0$. This makes SHG and SFG surface-sensitive and specific. While electric-quadrupole and magnetic-dipole contributions from the bulk may not be totally negligible, in many cases it has been shown that the surface contribution clearly dominates.

Surface SFG and SHG is best described by radiation from a surface polarization $\boldsymbol{P}_s^{(2)}(\omega)$ induced in a thin sheet with dielectric constant ε' sandwiched between two linear media (1) and (2) as shown schematically in Figure 1. This surface polarization has an in-plane wavevector component equal to the sum of the in-plane wavevector components of the incident fields: $\boldsymbol{k}_{\|}(\omega) = \boldsymbol{k}_{\|}(\omega_1) + \boldsymbol{k}_{\|}(\omega_2)$. The ratio of $\boldsymbol{k}_{\|}(\omega)$ to $\boldsymbol{k}(\omega)$ then determines the direction of the sum-frequency output that appears in both transmission and reflection. The reflection geometry is often used in order to minimize the bulk contribution.

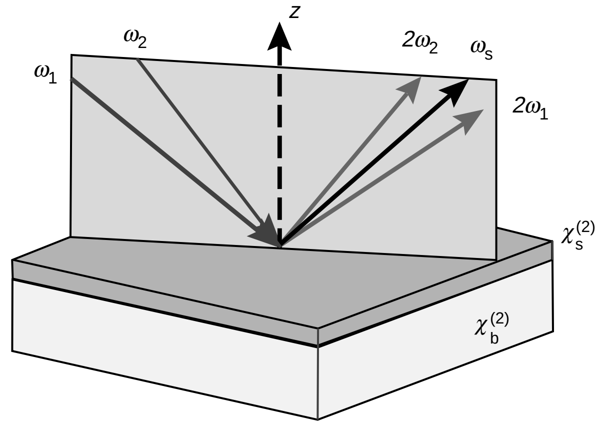

Figure 1 Schematic of SHG and SFG geometry in reflection and transmission from an interface.

With $\boldsymbol{P}_s^{(2)}$ as the source term, the sum-frequency output can be obtained from the solution of the wave equation. With the proper boundary conditions the sum frequency (SF) intensity is given by:

$$I(\omega) = \frac{8\pi^3\omega_s^2\sec^2\theta_\omega}{\hbar c^3\sqrt{\epsilon_1(\omega)\epsilon_1(\omega_1)\epsilon_1(\omega_2)}} \times |\boldsymbol{e}^{\dagger}(\omega)\cdot\boldsymbol{\chi}_s^{(2)} : \boldsymbol{e}(\omega_1)\boldsymbol{e}(\omega_2)|^2 I_1(\omega_1)I_2(\omega_2) \qquad [2]$$

In this expression, $\chi_s^{(2)}$ is the surface nonlinear susceptibility defined by $\boldsymbol{P}_s^{(2)} = \varepsilon_0\boldsymbol{\chi}_s^{(2)} : \boldsymbol{E}(\omega_1)\boldsymbol{E}(\omega_2)$; the quantities $\boldsymbol{e}(\omega_i)$ correspond to the unit polarization vectors after the appropriate Fresnel correction, $\boldsymbol{e}(\omega_i) \equiv \boldsymbol{F}(\omega_i)\hat{\boldsymbol{e}}(\omega_i)$, with $\boldsymbol{F}(\omega_i)$ being the transmission Fresnel factor and $\hat{\boldsymbol{e}}(\omega_i)$ the unit polarization vector of $\boldsymbol{E}(\omega_i)$; θ_ω denotes the SF output angle with respect to the surface normal, $\varepsilon(\omega_i)$ is the dielectric constant at frequency ω_i, and I_i is the input pump intensity at ω_i.

The surface nonlinear susceptibility $\chi_s^{(2)}$ is a third-rank tensor. In Cartesian coordinates it has 27 tensor elements ($\chi_{ijk}^{(2)}$), many of which could vanish or depend on others due to the surface structural symmetry. As an example, $\chi_{ijk}^{(2)}(\omega=\omega_1+\omega_2)$ for an isotropic surface, with the z-direction defined by the surface normal, has only four independent nonvanishing elements: $\chi_{xxz}^{(2)} = \chi_{yyz}^{(2)}, \chi_{xzx}^{(2)} = \chi_{yzy}^{(2)}, \chi_{zxx}^{(2)} = \chi_{zyy}^{(2)}$, and $\chi_{zzz}^{(2)}$. Different combinations of input and output beam polarizations in SFG measurements are often used to deduce values for the non-vanishing elements. Such measurements then allow the determination of surface symmetry or surface molecular orientation. Being a third-rank tensor, $\chi_s^{(2)}$ can reflect up to three-fold rotational symmetry. $\chi_{s,ijk}^{(2)}$ is related to the molecular nonlinear polarizability or hyperpolarizability $\alpha_{\xi\eta\zeta}^{(2)}$, where $\hat{\xi}$, $\hat{\eta}$, and $\hat{\zeta}$ define the molecular coordinates, through the coordinate transformation

$$\chi_{s,ijk}^{(2)} = N_s\left\langle(\hat{\boldsymbol{i}}\cdot\hat{\boldsymbol{\xi}})(\hat{\boldsymbol{j}}\cdot\hat{\boldsymbol{\eta}})(\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{\xi}})\right\rangle\alpha_{\xi\eta\zeta}^{(2)} \qquad [3]$$

The angular brackets denote an average over the molecular orientational distribution, and N_s is the surface molecular density. For simplicity, microscopic local-field correction is neglected in eqn [3]. Knowing $\chi_{s,ijk}^{(2)}$ and $\alpha_{\xi\eta\zeta}^{(2)}$ thus permits deduction of information on the orientational distribution.

An explicit expression for the nonlinear optical polarizability is obtained from a second-order quantum mechanical perturbation calculation:

$$\alpha^{(2)}_{s,ijk}(\omega_s) = -\frac{e^3}{\hbar^2}\sum_{g,n,n'}\left[\frac{\langle g|r_i|n\rangle}{\omega_s-\omega_{ng}+i\Gamma_{ng}} \times\left(\frac{\langle n|r_j|n'\rangle\langle n'|r_k|g\rangle}{\omega_2-\omega_{n'g}+i\Gamma_{n'g}} + \frac{\langle n|r_k|n'\rangle\langle n'|r_j|g\rangle}{\omega_1-\omega_{n'g}+i\Gamma_{n'g}}\right)+\cdots\right]\rho_g^{(0)} \qquad [4]$$

This expression shows how the nonlinear polarizability or susceptibility depends on material parameters such as the dipole transition moments $\langle n|r_i|g\rangle$ and energy levels. It contains a sum over eight resonant terms. The quantities ω_{ng} and Γ_{ng} are the frequencies and half-widths for the transitions between quantum states $|n\rangle$ and $|g\rangle$, and $\rho_g^{(0)}$ denotes the population in $|g\rangle$. It can be seen that $\alpha^{(2)}$, and hence the SF output, should be resonantly enhanced when the pump frequency ω_1 or ω_2 and/or the sum frequency $\omega_1+\omega_2$ approach resonance. The resonant enhancement provides spectral information and makes SHG and SFG effective spectroscopic tools. The resonances could be electronic or vibrational, or more generally, any surface characteristic transition.

Experimental

Here we deal with basic considerations concerning the optical setup for SHG or SFG experiments in reflection from a surface as is shown schematically in Figure 2. In general, the pump radiation from pulsed laser sources is directed onto the surface. High-power tunable light pulses can be obtained from optical parametric generation and amplification (OPG/OPA) together with harmonic, sum- or difference-frequency generation, preferably pumped by picosecond or femtosecond lasers with high repetition rates. By this means radiation tunable from near UV at ~200 nm to mid IR at ~18 μm can be generated. For SFG the input pulses with frequencies ω_1 and ω_2 are directed to overlap spatially and temporally on the sample. The sample could be the surface of a liquid or solid in air, a single crystalline surface in ultrahigh vacuum, buried interfaces, etc. Polarizers and half-wave plates allow for tuning of incident power and polarizations at the sample. Together with the polarizer in the output path, different input/output polarization combinations can be chosen. In the detection system the signal has to be discriminated against reflected and scattered pump light. This is achieved by spatial and spectral filtering with apertures, dielectric mirrors, and interference and color glass filters. An optional monochromator can also be used for additional stray-light suppression. The signal is then detected by a photomultiplier tube and gated electronics are used for signal integration or photon counting.

From eqn [2] the expected signal strength for SHG/SFG can be estimated. With a typical value of $\chi^{(2)}_s$ of 10^{-21} m^2 V^{-1}, a single 1-μm pump pulse, incident at $\theta = 45°$ having pulse energy $E = I A\tau = 100$ μJ, beam cross-section $A = 1$ mm^2, and pulse duration $\tau = 10$ ps, can generate about 10^3 photons per pulse of SHG. By means of photon counting a minimal count rate of 10^{-3} photons/pulse

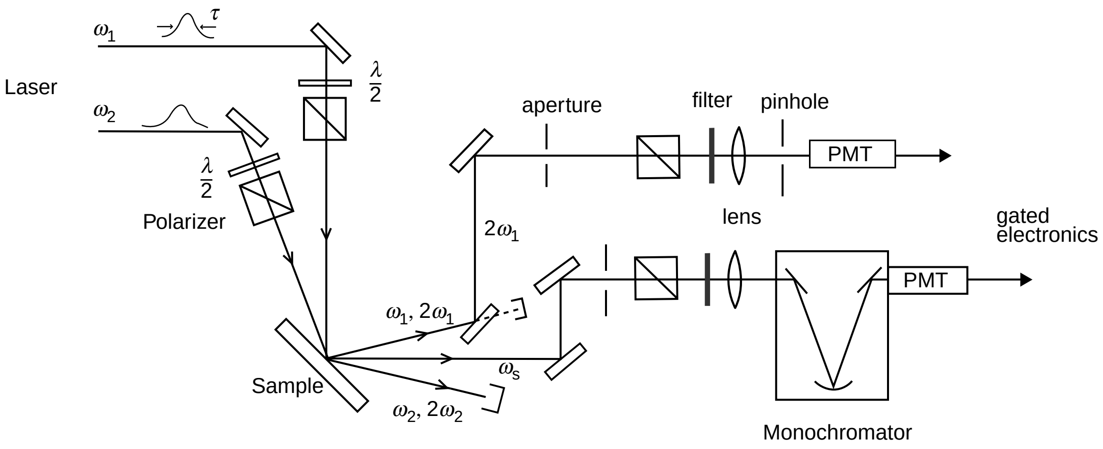

Figure 2 Schematic of experimental setup for second-harmonic and sum-frequency generation.

can be achieved which allows the detection of $\chi_s^{(2)}$ as small as 10^{-24} m^2 V^{-1}. The spectral resolution is limited by the spectral width of the pump pulses. In order to obtain accurate spectral data over a wide tuning range or to obtain absolute values for $\chi_{s,ijk}^{(2)}$ of the material investigated, a reference material such as *z*-cut quartz with known values of χ is to be measured for comparison.

The simplest kind of experiments are SHG at a fixed pump frequency. From the signal variation in response to surface modification one can probe kinetics and dynamics of adsorption, desorption, diffusion, surface melting, phase transitions, etc. By tuning the pump wavelength, second-harmonic spectroscopy can give information on surface electronic states. For surface vibrational spectroscopy, however, IR-vis SFG must be employed. In the latter case, tunable IR input is mixed with visible input to yield a SF signal in the visible region.

Surface Specificity

In many cases, the surface contribution to SFG or SHG from a centrosymmetric medium clearly dominates over the bulk contribution. This generally occurs when the surface or interface is composed of a polar oriented molecular layer. An example is shown in Figure 3, where the vibrational spectra in the C–H stretch region of three buried solid–liquid interfaces are presented. The spectrum for hexadecane ($C_{16}H_{34}$)/fused silica shows no C–H peaks. This is in spite of the fact that liquid hexadecane has strong infrared and Raman activity in this spectral range. The result indicates that the bulk signal contribution is negligible. In contrast, strong resonances are observed from the hexadecane/OTS/silica system, where OTS is a monolayer of octadecyltrichlorosilane ($CH_3(CH_2)_{17}SiCl_3$) chemisorbed on the fused silica plate. A similar spectrum is obtained for the CCl_4/OTS/silica interface. This clearly indicates that it is the contribution from the OTS monolayer that dominates the spectra.

Another example is described in Figure 4 where the SFG spectrum in reflection from a free liquid water surface is shown. The resonant features between 3000 and 3600 cm^{-1} are due to the O–H stretches of hydrogen-bonded OH groups. The sharp peak at 3700 cm^{-1} can be identified with the free OH bonds. Since no dangling OH bonds can exist below the surface its presence indicates that the spectrum must originate from the topmost layer of water molecules. Furthermore it can be shown that the surface water molecules are oriented with one of the OH bonds directed out of the liquid. In addition to the above examples, the surface sensitivity and specificity has been demonstrated and successfully applied to a large number of systems as will be shown below.

A note on the definition of surface or interface is in order. In SHG and SFG, the surface or interface layer refers to a thin layer between two adjacent bulk media that has a different structure from the bulk media and lacks inversion symmetry. If molecules in the surface layer are polar-ordered, then SHG and

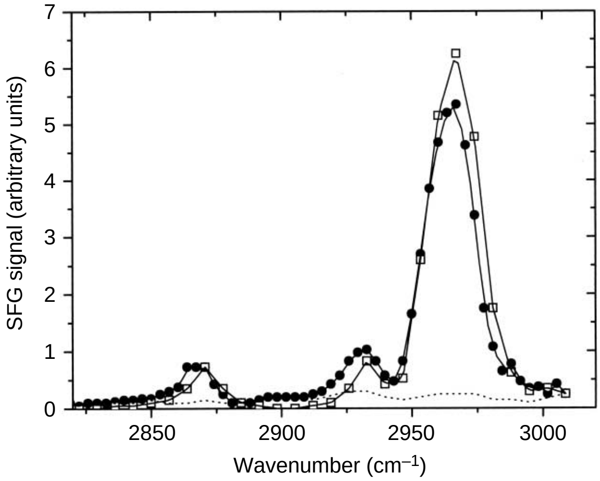

Figure 3 SFG spectra of different interfaces obtained with *ssp* polarization combination: hexadecane/silica (dashed), CCl_4/ OTS/silica (solid circles), and hexadecane/OTS/silica (open squares). OTS refers to a monolayer of octadecyltrichlorosilane.

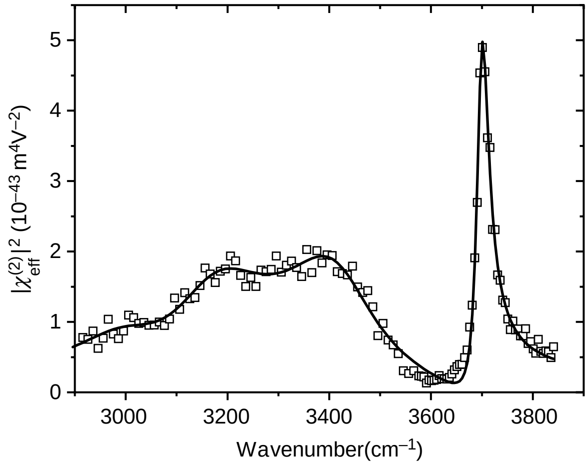

Figure 4 SFG spectrum of the liquid–vapor interface of pure water (*ssp* polarization combination). The mode at 3700 cm^{-1} is due to the dangling OH bonds at the surface.

SFG are often dominated by contributions from the surface layer. However, the bulk contribution is not strictly zero even if the media have inversion symmetry, as $\chi^{(2)}$ can arise from electric-quadrupole and magnetic-dipole contributions. Separation of surface and bulk contributions in SHG and SFG is a subtle problem in general. Experimentally, how surface modification affects the signal can be used to evaluate whether or not the surface contribution dominates.

A special case concerns materials that lack bulk inversion symmetry, for example, III–V and II–VI semiconductors such as GaAs and ZnSe. In this case, the bulk contribution to SHG and SFG is electric-dipole allowed and may be very significant. With an appropriate choice of polarization combinations and surface orientation, it can nevertheless be well suppressed, leaving the surface contribution dominant.

Applications

Both SFG and SHG have been applied to a wide variety of surface and interfacial systems. To illustrate their capabilities and versatility as surface spectroscopic tools, the focus of this summary will be on examples where the information deduced cannot readily be obtained by other techniques. (For details, readers are referred to the review articles cited in Further Reading and the original papers cited therein.)

SFG and SHG have been used for the study of adsorbates at surfaces even under ambient conditions. The adsorbate alters electronic or structural properties of the substrate and the signal is thus related to the surface density of the adsorbate and its orientation.

Elementary surface reactions, in particular adsorption, dissociation, or desorption of simple molecules such as CO, hydrogen, oxygen, or water on metal and semiconductor single-crystal surfaces under ultra-high-vacuum conditions, have been investigated. Sensitivity into the sub-percent monolayer regime is achieved and allows the identifictation of competing reaction channels and adsorbate dynamics on surfaces. Such studies can help to understand epitaxial growth or heterogeneous catalysis because the techniques are applicable at any gas pressure and surface temperature. Related to this is the study of surface diffusion, which can be probed following the temporal decay of a diffracted SHG signal off a submonolayer adsorbate grating formed by laser-induced desorption.

Making use of the spectroscopic capabilities of SHG, the surface electronic structure of selected metals and semiconductors, in particular silicon, were studied with an emphasis on the effects of temperature, crystallographic orientation and different adsorbates. Among buried interfaces, the Si/SiO$_2$ interface has been the subject of intense investigations and the sensitivity of SHG to static electric fields at the interface and to inhomogeneous strain was also demonstrated.

Different in-plane surface symmetries associated with different crystallographic orientations or surface reconstructions yield different SHG and SFG responses. This allowed for the investigation of surface reconstructions (e.g., Si(111)$2\times1 \rightarrow 7\times7$), order–disorder transitions (e.g., Si(111)$7\times7 \rightarrow 1\times1$, Au/Si(111)) and surface melting (e.g., single-crystal ice, Si(111), Ge(111)), providing information on latent heat, superheating, and its dynamics using pump-probe techniques. Magnetization-induced SHG was also used for probing surface and interface magnetism of ferromagnetic metals and bimetallic systems.

Many unique applications of spectroscopic SFG have been developed based on its capability to study large molecules, in particular under ambient conditions. Surface vibrational spectra with different input/output polarization combinations and sample geometries provide information on orientation, conformation and alignment of surface molecules and the composition and structure of the surface layer. The study of self-assembled monolayers (e.g., OTS on quartz) to deduce the orientation and conformation of the alkyl chains in the monolayer is an example. The effect of molecular density and temperature on the spectrum gives information on intermolecular interactions and phase transitions. Another example concerns the adsorption of liquid crystal molecules on nanostructured polymer substrates. Both the chain orientation at the polymer surface and its effect on the alignment of adsorbed liquid crystal molecules can be addressed.

Surface structures of liquids, especially those of pure liquids, are of great interest in science and technology, and SFG is a unique spectroscopic tool for probing liquid surfaces and interfaces. For example, the SFG vibrational spectrum for the water–vapor interface in Figure 4 gives information on the density and orientation of the surface water molecules at the water–vapor interface. The surface layer was found to be more ordered compared to the bulk as was also observed for various other organic liquids studied with SFG. In other cases soluble and insoluble organic molecules at a liquid–vapor, liquid–liquid, or liquid–solid interface have also been studied. They appear as adsorbed monolayer films and their structure and phase behavior are relevant to many applications. Identification of the adsorbed molecular species at solid–liquid interfaces is also important for understanding electrochemical processes. Various studies with SFG have addressed bonding of, e.g., CO, CN^-, and SCN^- at electrodes and its variation with electrode potential.

SFG spectroscopy also finds unique applications in studies of polymer surfaces and interfaces most relevant to modern science and technology. The observed surface vibrational spectra of polymers provide information on surface composition, molecular orientation, and conformation of neat and blended polymers. The effect of environment and surface treatment on the surface structure can also be studied. Investigation of the biological functions of complex molecular systems has attracted increasing interest, and SHG and SFG have been applied successfully to selected systems ranging from the isomerization of retinal – a molecule involved in the vision process – to functional aspects of model membranes.

Both SHG and SFG can be combined with microscopy techniques for surface microscopic studies. Microscopic imaging can be achieved by rastering the probe beam or, preferably, the sample position. Near-field SFG/SHG spectroscopy has recently been developed.

Performing surface SHG and SFG experiments in the time domain can provide information on surface dynamics. Pump-probe experiments allow for measurements of energy relaxation and phase coherence of excitations such as surface electronic states and surface phonons and vibrations.

Outlook

Although SHG and SFG were established as surface analytical tools more than 15 years ago, it is still an active field of research with much potential that has not yet been explored. As an example, it has only recently been demonstrated that doubly resonant infrared–visible SFG, as a two-dimensional spectroscopy, can give access to couplings between vibrational modes and surface electronic transitions. With the commercial availability of suitable laser sources, the techniques can be extended to a much wider range of applications. This includes possible investigations of ultrafast surface dynamics, nanostructures, and biological systems.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon Press.

Bloembergen N (1977) *Nonlinear Optics*. New York: Benjamin.

Boyd RW (1993) *Nonlinear Optics,* 2nd edn. San Diego, CA: Academic Press.

Dumas P, Weldon MK, Chabal YJ and Williams GP (1999) Molecules at surfaces and interfaces studied using vibrational spectroscopies and related techniques. *Surface Review & Letters* 6(2): 225–255.

Guyot-Sionnest P and Harris AL (1995) Surface vibrational dynamics probed by sum frequency generation. In: Ho W and Dai HL (eds) *Laser Spectroscopy and Photochemistry of Metal Surfaces*, pp. 405–458. Singapore: World Scientific.

Guyot-Sionnest P, Superfine R, Hunt JH and Shen YR (1988) Vibrational spectroscopy of a silane monolayer at air/solid and liquid/solid interfaces using sum-frequency generation. *Chemical Physics Letter* 144: 1–5.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Stegeman GI and Ponath H-E (eds), *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam, The Netherlands: North-Holland.

Matthias E and Träger F (eds) (1999) *Applied Physics B, Special Issue: Nonlinear Optics at Interfaces*, Volume 210.

McGilp JF (1999) Optical second-harmonic generation as a semiconductor surface and interface probe. *Physica Status Solidi (a)* 175: 153–167.

Miranda PB and Shen YR (1999) Liquid interfaces: A study by sum-frequency vibrational spectroscopy. *Journal of Physical Chemistry B* 103: 3292–3307.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Interfaces*, 413. Amsterdam, The Netherlands: Elsevier.

Shen YR (1984) The Principles of Nonlinear Optics. New York: Wiley.

Shen YR (1994) Nonlinear optical studies of polymer interfaces. *International Journal of Nonlinear Optical Physics* 3: 459–468.

Shen YR (1994) Surface spectroscopy by nonlinear optics. In: Ḧansch TW and Inguscio M (eds) *Frontiers in Laser Spectroscopy*, pp. 139–165. Amsterdam, The Netherlands: North-Holland.

Shen YR (2001) Vibrational spectroscopy of water at interfaces. In: Morra M (ed.) *Water in Biomaterials Surface Science*, Chapter 8, pp. 215–244. Chichester, UK: Wiley.

Wei X and Shen YR (2001) Motional effect in surface sum-frequency vibrational spectroscopy. *Physical Review Letter* 86: 4799–4802.

POLARIZATION

Contents

Introduction
Matrix Analysis

Introduction

J M Bennett, Michelson Laboratory, China Lake, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Polarization is one property of light waves and will be defined after a brief overview of the properties of light. Light is part of the electromagnetic spectrum of waves that have both particle-like and wave-like properties. Light waves carry energy in the form of photons which act like particles; the photon energy increases in proportion to the frequency of the wave. The particle-like properties of light and other electromagnetic waves are described by quantum mechanics. Light also acts like transverse waves that travel in straight lines in air and vacuum and can be described by classical electromagnetic theory. Polarization deals with the wave-like properties of light, and is described mathematically by Maxwell's equations, the key relations in electromagnetic theory. Polarization effects occur when light interacts with matter. In order to understand polarization, there will be a brief introduction to the subject of electromagnetic waves.

For a more in-depth treatment of the material in this article, the reader is directed to the Further Reading list at the end of this article and in particular to the two chapters on 'Polarization' and 'Polarizers' written by the author in Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chapter 5 and vol. II, chapter 3. New York: McGraw-Hill Inc.

Electromagnetic waves have both electric and magnetic fields associated with them. These are vibrations in directions perpendicular to the direction the wave is traveling, i.e., the direction of propagation. $\vec{E}$ represents the vector of the electric field and $\vec{H}$, perpendicular to $\vec{E}$, represents the magnetic field vector. Both these vectors are complex and have real and imaginary parts. Polarization effects are always associated with the $\vec{E}$ vector. Specifically, the plane of polarization is defined as the plane in which the $\vec{E}$ vector is vibrating. Waves having different amplitudes, phases, or angular orientations (azimuths) of their electric or magnetic vectors can be combined by conventional vector addition methods. Also the $\vec{E}$ vector of a particular vibration can be resolved into two components in mutually perpendicular directions that are vibrating in phase.

If a light source such as the sun, a candle flame, or an electric light bulb is considered on a microscopic scale, each vibrating atom or molecule emits linearly polarized light (see the definition below). But the individual atoms or molecules do not act together, so their vibrations have no fixed phase relationships to each other and they cannot be added into a single linearly polarized beam. Thus, we call light from these sources unpolarized. In an unpolarized light beam, the $\vec{E}$ vector vibrates in all directions perpendicular to the direction of propagation. If a snapshot is taken at a particular instant of time, different parts of the beam will have $\vec{E}$ vectors vibrating with different amplitudes and phases at different angles to each other, but all in a plane perpendicular to the direction of propagation. In the most common convention used in optics, the wave travels in the $+z$ direction in a right hand coordinate system and the $\vec{E}$ vectors are all vibrating at various angles in the x–y plane. The angle of vibration is measured from the positive x axis in a counterclockwise direction when the observer is looking against the direction of propagation of the light beam.

For linearly polarized or plane polarized light, if a snapshot of a light beam is taken at a particular instant, the $\vec{E}$ vector will be vibrating at a certain angle in the x–y plane. As time (or position on the traveling wave) varies, the amplitude of the $\vec{E}$ vector will vary in a sinusoidal manner, but the vibration will remain at the same angle in the x–y plane. As an example, the real part of the electric vector of a linearly polarized beam that is vibrating in the $+x$ and $-x$ directions and traveling in the $+z$ direction is given by the relation:

$$\vec{E}_x(z,t) = \hat{\imath} E_0 \cos(kz - \omega t) \qquad [1]$$

where t is the time, $\hat{\imath}$ is a unit vector in the $+x$ direction, E_0 is the amplitude of the vibration, $\vec{k}$ is the propagation vector in the direction the wave is traveling, i.e., the z direction (its magnitude is $2\pi/\lambda$), and $\omega = 2\pi f$ where f is the vibration frequency of the wave. In free space, $\omega = 2\pi c/\lambda$ where c is the velocity of light and λ is the wavelength of the vibration.

Light can also be circularly or elliptically polarized. Circularly polarized light is produced by adding the electric vectors of two waves, each having the same amplitude but which are 90° out of phase; the resulting vibration sweeps out a circle in the $x-y$ plane. When the first wave is vibrating in the $+x$ and $-x$ directions (eqn [1]), the second wave is vibrating in the $+y$ and $-y$ directions:

$$\vec{E}_y(z,t) = \hat{\jmath} E_0 \sin(kz - \omega t) \qquad [2]$$

and is added to the first wave. The resultant wave is the sum of eqns [1] and [2]:

$$\vec{E} = E_0[\hat{\imath} \cos(kz - \omega t) + \hat{\jmath} \sin(kz - \omega t)] \qquad [3]$$

The direction of vibration of this wave will rotate in a circle as either the time increases or as the distance along the wave increases, but the amplitude of the vibration will not change. As time increases, the vibration will make a circle in the clockwise direction (to negative angles). This light is defined to be right circularly polarized. If the + sign between the two parts of eqn [3] is changed to a minus sign, as time increases, the vibration will make a circle in the counterclockwise direction (to positive angles) and the light is now defined to be left circularly polarized. These definitions are for conventional (traditional) optics. However, the opposite definitions with right (left) circularly polarized light defining circles in the counterclockwise (clockwise) direction are also in use. In modern physics, there is still another convention that defines right (left) circularly polarized light as having negative (positive) helicity.

Right and left elliptically polarized light beams have the same angle conventions as for circularly polarized beams but are produced by adding two electric vectors that have different amplitudes. In eqns [1] and [2], the amplitude terms will be, for example, E_1 and E_2 instead of E_0 and eqn [3] can no longer be used. There is now a major axis and a minor axis of the ellipse. If $E_1 > E_2$, the major axis will be along the x axis; for $E_1 < E_2$, the major axis will coincide with the y axis.

The preceding discussion has dealt entirely with the electric vector of the electromagnetic field. However, one cannot observe the $\vec{E}$ vector. The irradiance (energy per unit area per unit time), $\vec{E}\vec{E}^*$, is what can be observed visually and measured by electronic detectors. Thus, measurements of the polarization are irradiance measurements. Light transmitted by a polarizer is called the transmittance of the polarizer; similarly, light reflected from a polarizer is called the reflectance of the polarizer.

One of the most important parts of the subject of polarization is how to produce linearly polarized light starting with unpolarized light. This is done by using polarizers, as discussed in the section on polarizers below. Certain materials have special properties that make them able to polarize light. Depending on the application, different kinds of polarizers are preferred. Optics textbooks by Hecht and Guenther discuss the most important polarizers and two articles by Bennett describe many kinds of polarizers including special ones (see the Further Reading at the end of this article for full references). Only the basic principles will be described here.

Sunlight scattered by air molecules in the atmosphere (Rayleigh scattering) is also partially linearly polarized. The air molecules act like tiny dipoles and vibrate when they absorb sunlight. They emit radiation that is polarized in certain directions relative to the vibration direction. When the viewer is at a 90° angle with respect to the sun, the polarization of the skylight is a maximum. Rayleigh scattering is the subject of several books and will not be further discussed here.

Retarders are used to change linearly polarized light into circularly or elliptically polarized light and compensators, which are a form of retarders, can analyze an unknown type of polarized light and determine its composition. They are discussed below.

Polarimetry and ellipsometry are closely related techniques that are used to determine the optical properties of a material by shining a known type of polarized light on the material at non-normal incidence and analyzing the polarization properties of the light after it has been reflected from the material. These subjects are treated in other articles in this encyclopedia and in several references at the end of this article.

Changes can be produced in the optical properties of some materials by applying an electric field, a magnetic field, an acoustic field, or another form of variable pressure. The materials changed in these ways are said to be electro-optic, magneto-optic, or piezo-optic. The changes in the optical properties modify some parameter of a light wave passing through a material or reflecting from it.

The amplitude, phase, frequency, polarization, or direction of the light wave can be modified. In this article, we are only concerned with modifications that produce changes in the polarization; these are discussed towards the end of this article.

Matrix methods have been developed to handle problems involving polarization and there is also a visual representation of the matrix algebra that is based on the Poincaré sphere. Both topics are covered at the end of this article.

Polarizers

Basic Relations for Polarizers

A linear polarizer is anything which, when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one direction with only a small component vibrating in the direction perpendicular to it. The transmittance T of the linear polarizer is

$$T = \tfrac{1}{2}(T_1 + T_2) \qquad [4]$$

where T_1, the principal transmittance of the polarizer, is $\gg T_2$, the transmittance of the polarizer at 90° to the principal transmittance. Thus, a perfect polarizer would transmit only 50% of an incident unpolarized beam.

If a linear polarizer is placed in a linearly polarized beam and is rotated about an axis parallel to the beam direction, the transmittance will vary between a maximum value T_1 and a minimum value T_2 according to the law:

$$T = (T_1 - T_2)\cos^2\theta + T_2 \qquad [5]$$

where θ is the angle between the plane of vibration of the principal transmittance and the plane of vibration of the electric vector in the incident beam.

The extinction ratio ρ_P of a polarizer is defined as

$$\rho_P = \frac{T_2}{T_1} \qquad [6]$$

and the degree of polarization of a polarization P is

$$P = \frac{T_1 - T_2}{T_1 + T_2} \qquad [7]$$

When two identical polarizers are placed in an unpolarized beam, and the directions of their principal transmittances, T_A and T_B, are inclined at an angle θ to each other, the transmittance of the pair will be

$$T_\theta = \tfrac{1}{2}(T_1^2 + T_2^2)\cos^2\theta + T_1 T_2 \sin^2\theta \qquad [8]$$

Thus, when the directions of the principal transmittances are aligned, $T_\parallel = \frac{1}{2}(T_1^2 + T_2^2)$, and when they are perpendicular, $T_\perp = T_1 T_2$.

Birefringent Materials (Calcite)

The majority of high-quality polarizers are made from calcite. This is a birefringent (doubly refracting) crystalline material that is uniaxial (i.e., there is one preferred direction in the crystal). A birefringent material acts differently for light going in different directions through the crystal. For example, if an unpolarized light beam passes through the crystal in a certain direction, it will be split into two spatially separated beams that are parallel but are linearly polarized at right angles to each other. A uniaxial crystal has an optic axis (i.e., a certain direction in the crystal). When light rays travel parallel to the optic axis, they travel at the same velocity and there is no difference between them. When light passes through the crystal in other directions, the ray whose vibration direction is perpendicular to the optic axis is governed by the ordinary laws of geometrical optics (the same as for isotropic materials) and is called the ordinary ray. The ray whose vibration direction is parallel to the optic axis does not follow the normal geometrical optics laws and is called the extraordinary ray. One ray travels faster than the other, so there is a phase retardation for one ray relative to the other. This is the principle of a retarder or retardation plate (see the next section).

Calcite is a negative uniaxial crystal which means that the refractive index for the ordinary ray is larger than the refractive index for the extraordinary ray. When the ordinary ray enters a block of calcite from air at non-normal incidence, it is bent more than the extraordinary ray.

Calcite can be easily cleaved along three distinct planes, making it possible to produce rhombs of the form shown in Figure 1. The optic axis, going in the HI direction, makes equal angles with all three faces at point H. Any plane, such as DBFH, which contains the optic axis and is perpendicular to the two opposite faces of the rhomb ABCD and EFGH is called a principal section. If the plane of incidence of light on the rhomb coincides with a principal section, the light entering the crystal will be split into two components polarized at right angles to each other which travel in slightly different directions and leave the crystal as two beams slightly displaced but parallel to each other.

The large birefringence of calcite and its excellent transmission through the visible spectral region and into the ultraviolet and infrared regions has made it possible to make excellent high extinction ratio polarizers with different designs. Some of these are shown in Figure 2. There are two main types: Glan types with rectangular shapes and Nicol types with rhombohedral shapes. The polarizers are made of two pieces of calcite cemented together. Glan types have their optic axes in the plane of the entrance face. In the Nicol types, the principal section is perpendicular to the entrance face, but the optic axis is neither parallel nor perpendicular to the face. The two halves of conventional polarizing prisms are cemented together with cement that has a refractive index intermediate between the ordinary and extraordinary refractive indices. This enables one ray (generally the extraordinary, or *e* ray) to be transmitted and the other to be reflected at the cut, so that only one ray exits from the prism in the direction of the incoming ray. The Glan types are also used without cement with only an air space between the two halves. They can be used at shorter wavelengths in the ultraviolet where the cement absorbs. The extinction ratio can be very high for air-spaced prisms. For example, for a Glan–Foucault prism (an air-spaced Glan–Thompson prism, Figure 2), the extinction ratio can be better than 1×10^{-5} and the prism is usable from about 0.214 μm in the ultraviolet to 2.3 μm in the infrared. However, air-spaced polarizers have very small field angles, so they are mainly used with laser sources where the beam is parallel.

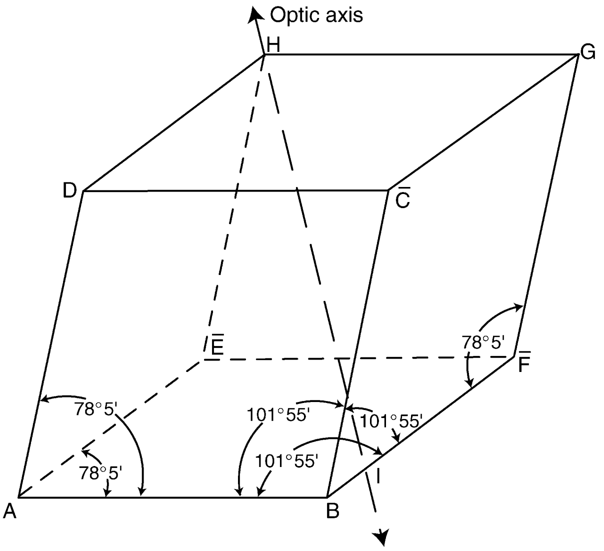

Figure 1 Schematic representation of a rhombohedral calcite crystal showing the angles between faces. The optic axis passes through corner H and point I on side BF. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

The extreme paths of light through a cemented Glan–Thompson prism are shown in Figure 3. This prism has a length that is three times the width of the entrance aperture (i.e., an *L*/*A* ratio of 3). The optic axis is perpendicular to the plane of incidence which is in the plane of the paper. In the first half of the polarizer, the paths of the ordinary and extraordinary rays, both of which follow the conventional law of refraction (Snell's Law, eqn [8] above) nearly coincide. Ray A is incident on the entrance face of the polarizer at an angle such that the angle of incidence on the cut is the smallest angle for which the *o* ray is totally internally reflected. Ray B is incident at an angle such that the angle of refraction in the first half of the prism is essentially equal to the cut angle, S, so that the *e* ray just passes through the cut. The field angle of the polarizer is twice the smaller of angles θ_A and θ_B. Thus, all rays having angles of incidence between rays A and B will be polarized

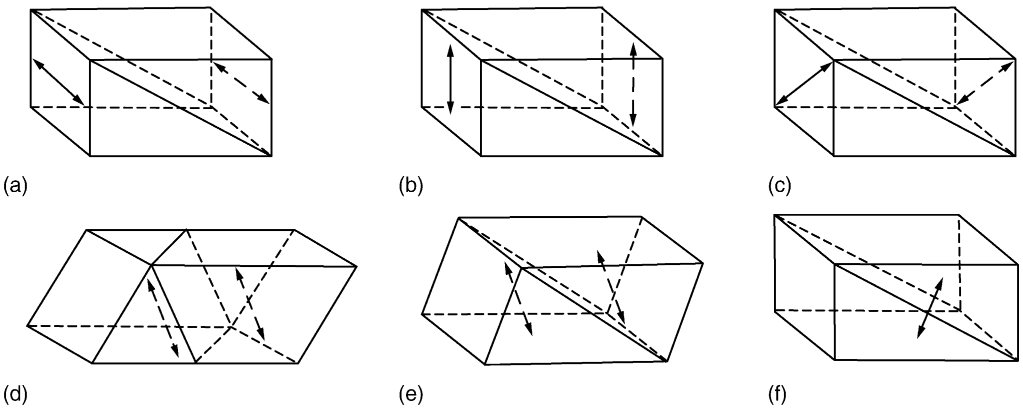

Figure 2 Types of polarizing prisms. Glan types: (a) Glan-Thompson, (b) Lippich, and (c) Frank-Ritter; Nicol types: (d) conventional Nicol, (e) Nicol, Halle form, and (f) Hartnack-Prazmowsky. The optic axes are indicated by the double-pointed arrows. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

when exiting the prism. The paths of similar rays can be traced through the other prism designs.

In addition to conventional polarizing prisms, there are also polarizing beamsplitter prisms and Feussner prisms. In polarizing beamsplitter prisms, the two beams that are polarized at right angles to each other both emerge from the prism but are separated spatially. Figure 4 shows a diagram of several types of polarizing beamsplitter prisms and Figure 5 shows the paths of rays through side views of these prisms.

A Feussner prism is made of isotropic material but the film separating the two halves is birefringent. These prisms have the advantage that much less birefringent material is required but they have a more limited wavelength range when calcite or sodium nitrite (another birefringent material) is used because the transmitted ordinary ray has a shorter transmission range than the extraordinary ray which is transmitted by the conventional and air-spaced polarizing prisms.

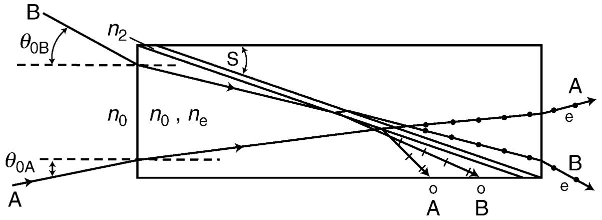

Figure 3 Extreme rays (A and B) that can pass through a cemented Glan-Thompson prism. Both rays change to ordinary and extraordinary rays in the calcite. Both ordinary rays are reflected at the boundary between the two halves of the prism and the extraordinary rays exit in the directions indicated. Rays entering the prism between rays A and B would be transmitted by the prism. The field angle of the prism is twice the smaller of the two angles of incidence (ray A).

Dichroic Absorbers

A dichroic material is one that absorbs light polarized in one direction more strongly than light polarized at right angles to that direction. Dichroic materials are different from birefringent materials because the latter usually have negligible absorption coefficients for both the ordinary and extraordinary rays. Stretched polyvinyl alcohol sheets treated with absorbing dyes or polymeric iodine are the most common type of dichroic absorbers and are primarily sold under the tradename Polaroid. These polarizers do not have as good an extinction ratio as the calcite prism polarizers (see above), but they are inexpensive, come in large sizes, are easily rotated, and produce negligible beam deviation. Also, they are thin, lightweight, and can be made in any desired shape. One of their main advantages is that they are insensitive to the degree of collimation of the beam and can be used in strongly convergent or divergent light.

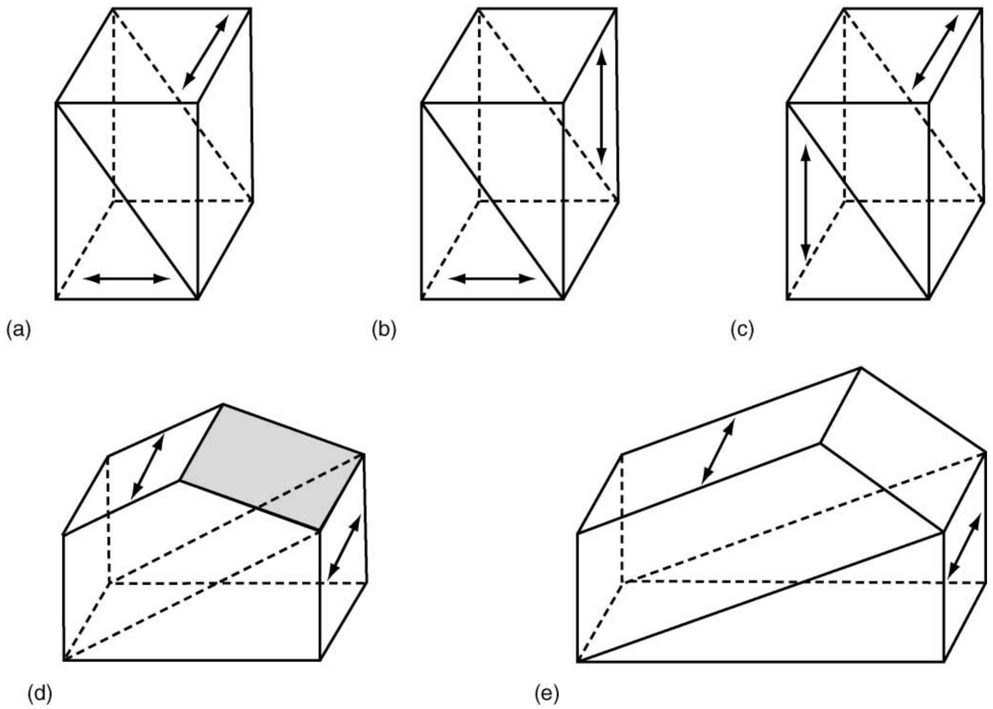

Figure 4 Three-dimensional views of various types of polarizing beamsplitter prisms: (a) Rochon; (b) Sénarmont; (c) Wollaston; (d) Foster (shaded face is silvered); and (e) beamsplitting Glan-Thompson. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

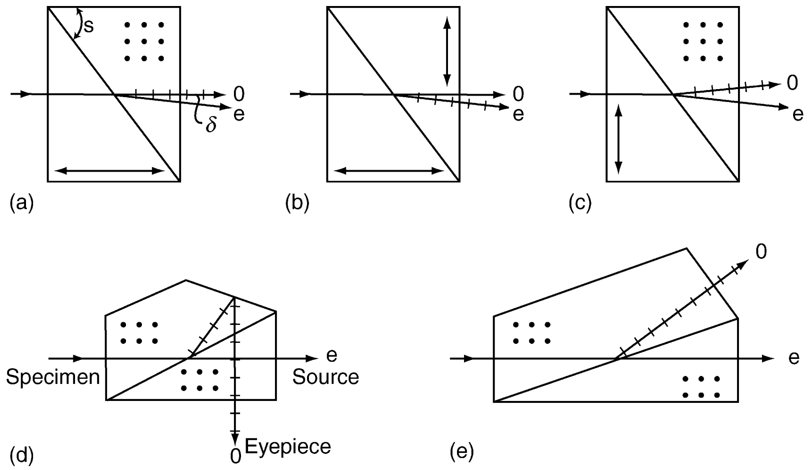

Figure 5 Side views of the polarizing beamsplitter prisms in **Figure 4**. The directions of the optic axes are indicated by the dots and the heavy double-pointed arrows. When the Foster prism is used as a microscope illuminator, the source, specimen, and eyepiece are in the positions indicated. (Reproduced with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

Depending on the density and type of absorbing dye used to make the polarizer, the transmission can be maximized for the visible or near infrared spectrum. The extinction ratio of the Polaroid HN-22 sheet compares favorably with that of the Glan–Thompson prisms throughout the visible spectral region, but the transmission of the Glan–Thompson prism is superior. As the dichroic polarizer transmission increases, the extinction ratio becomes worse. Transmission and extinction ratio curves for various types of dichroic polarizers are shown in Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.

A wire grid polarizer is another kind of dichroic absorber. It is made of a series of equally spaced conducting bars or wires that are either free standing or deposited onto a transparent substrate (backing plate). Energy that is polarized parallel to the length of the bars is absorbed out of the incoming wave by inducing oscillations in the electrons in the metal. Thus, only light polarized perpendicular to the bars will be transmitted. In order to have an appreciable degree of polarization, the wavelength must be at least twice the spacing between grids in the polarizer. Because of the difficulty of making wire grids with small enough spacings, these polarizers are limited to the mid- and long-wavelength infrared spectral region, beyond about 5 μm. The polarizer will have the best extinction ratio if the substrate has a low refractive index; if a high refractive index substrate such as silicon or germanium is used, it must be covered with an antireflection coating before the grid is deposited.

Most of the wire grid polarizers have been used with microwaves, so the theory is in the form for transmission lines. Similar transmission line theory has been applied to infrared polarizers in the form of bars and wires.

Reflection and Transmission

Light can be polarized by reflecting it from the flat surface of a material inclined at non-normal incidence to the light beam or by transmitting it through a transparent plate at non-normal incidence. These polarizers work because light has different reflectances when the electric vector is linearly polarized parallel and perpendicular to the plane of incidence. The plane of incidence contains both the incoming light beam (incident beam) and the reflected light beam. The angles for both the incident and reflected light beams are measured relative to an axis perpendicular to the surface (the surface normal). The reflection coefficients are given by the Fresnel equations and, for nonabsorbing materials, are:

$$R_s = r_s^2 = \frac{\vec{E}_s^2(\text{reflected})}{\vec{E}_s^2(\text{incident})} = \frac{\sin^2(\theta_0 - \theta_1)}{\sin^2(\theta_0 + \theta_1)} \qquad [9]$$

$$R_p = r_p^2 = \frac{\vec{E}_p^2(\text{reflected})}{\vec{E}_p^2(\text{incident})} = \frac{\tan^2(\theta_0 - \theta_1)}{\tan^2(\theta_0 + \theta_1)} \qquad [10]$$

where R_s and R_p are called the reflectances (previously called the intensity reflection coefficients), and r_s and r_p are the amplitude reflection coefficients. $\vec{E}_s$ ($\vec{E}_p$) is the incident or reflected electromagnetic wave vibrating perpendicular (parallel) to the plane of incidence and θ_0 and θ_1 are the angles of incidence and refraction, respectively. The angle of refraction is the angle of the light beam in the material, measured from the surface normal. For a nonabsorbing material, it can be obtained in terms of the refractive index n_1 of the material from Snell's Law:

$$\frac{\sin \theta_0}{\sin \theta_1} = \frac{n_1}{n_0} \qquad [11]$$

where n_0 and n_1 are the refractive indexes of the incident medium (usually air with $n_0 = 1$) and the material. Sometimes the refractive index of the material is expressed as a ratio measured relative to the refractive index of the incident medium: $n = n_1/n_0$. The reflectances of absorbing materials are similar to eqns [9]–[11] but involve complex refractive indexes and angles of refraction.

Figure 6 shows curves for R_s and R_p as a function of angle of incidence for four nonabsorbing materials that have different refractive indexes. In all cases the reflectance is higher for the *s* component than for the *p* component except at 0° and 90° angles of incidence where they are the same. At the so-called Brewster angle, θ_B, $R_p = 0$, $\tan \theta_B = n_1/n_0$, and incident unpolarized light is now completely linearly polarized perpendicular to the plane of incidence (s-polarized light). Note that high refractive index materials produce more intense polarized beams (i.e., have higher reflectances) than low refractive index materials.

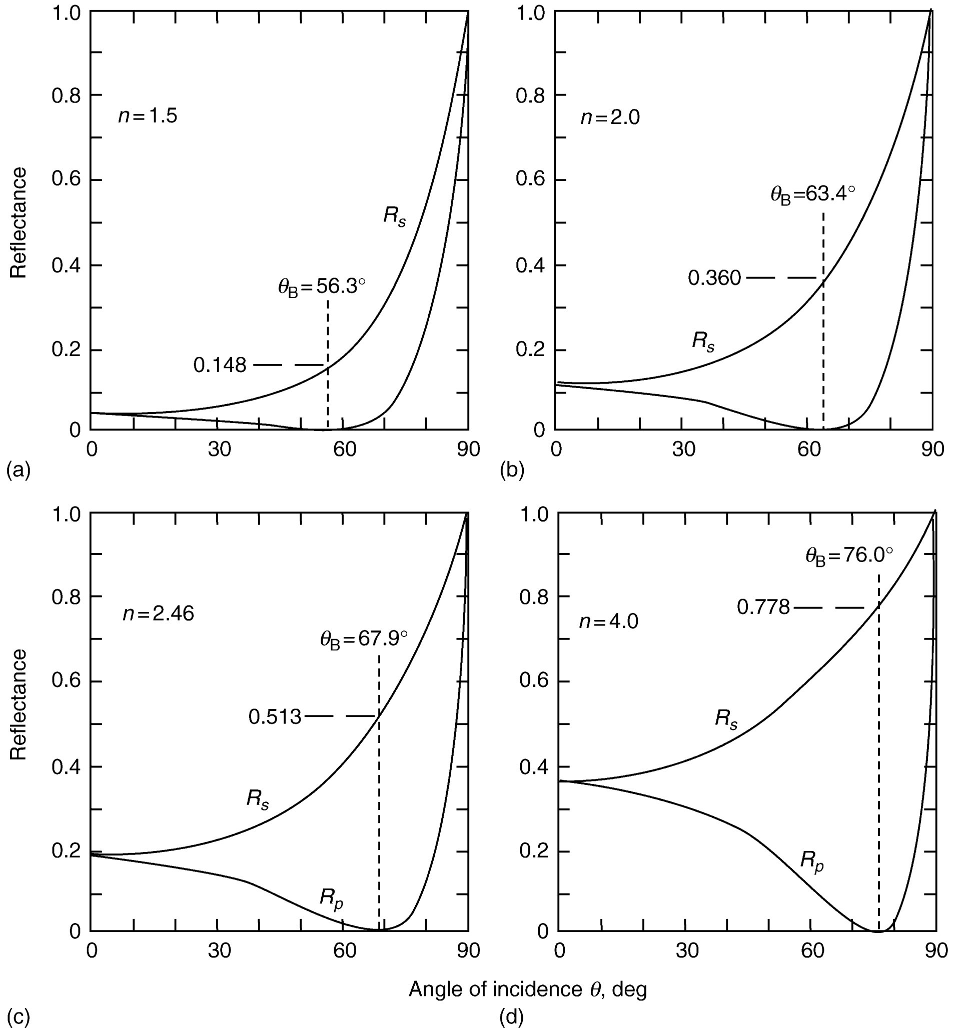

Figure 6 Reflectance of light polarized parallel R_p and perpendicular R_s to the plane of incidence from materials of different refractive index *n* as a function of angle of incidence: (a) $n = 1.5$ (alkali halides in the ultraviolet, glass (approximately) in the visible, and sheet plastics in the infrared), (b) $n = 2.0$ (AgCl in the infrared), (c) $n = 2.46$ (Se in the infrared), and (d) $n = 4.0$ (Ge in the infrared). The Brewster angle θ_B (at which R_p goes to 0) and the magnitude of R_s at θ_B are also indicated. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

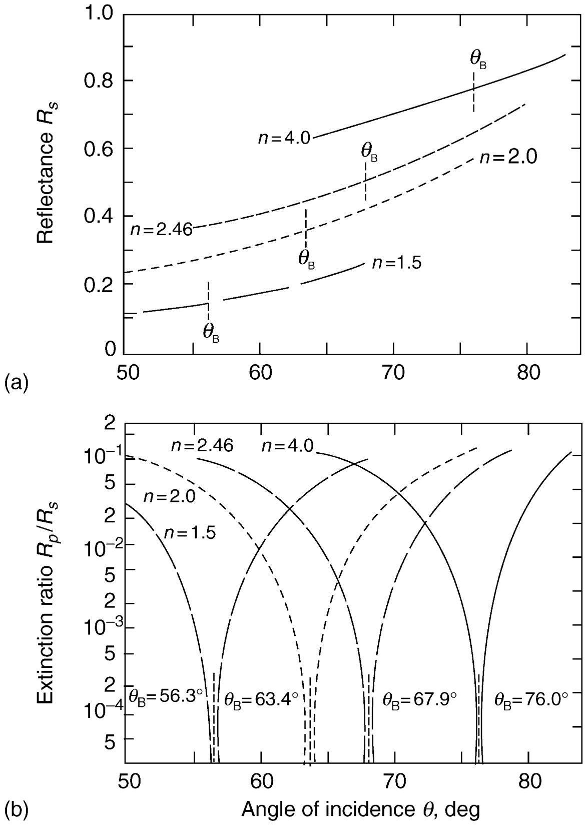

Figure 7 (a) Reflectance R_s and (b) extinction ratio R_p/R_s for materials of different refractive index at angles near the Brewster angle θ_B. A single surface of the material is assumed. (Reproduced with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 7 shows the reflectances and extinction ratios for materials having different refractive indexes. Each curve is for a single reflection. Since plates have two surfaces, there will be two reflections from an actual reflection polarizer so the reflectance and extinction ratios will increase. If the plates are thick, the reflection from the back surface of the plate will be displaced from the reflection from the front surface of the plate. A reflection polarizer is normally used close to the Brewster angle because a high degree of polarization (i.e., a very small extinction ratio) is desired. Because the extinction ratio changes rapidly for angles of incidence close to the Brewster angle, the incident beam must be well collimated to obtain a high degree of polarization. A main disadvantage of reflection polarizers is that the reflected beam is no longer parallel to the incident beam and two additional reflections are required to align the beam.

Light transmitted through a plate will only be partially linearly polarized at any angle of incidence including the Brewster angle because both s- and p-components are partially transmitted. Transmission polarizers are thus made of several plates to increase the degree of polarization. Figure 8 shows the transmittance and extinction ratio for a stack of four plane parallel plates assuming multiple incoherent reflections within each plate and no reflections between plates. At the Brewster angle the p-transmittance is theoretically 1 (assuming that there is no absorption within the material) but the extinction ratio greatly depends on the refractive index of the material. A stack of low refractive index plates ($n = 1.5$) has a poor extinction ratio, while a pile of high refractive index plates ($n = 4.0$) has a much better extinction ratio. The plates are often inclined at small angles to each other so the light multiply reflected between plates (which decreases the polarization) is removed from the transmitted beam. The sides of each plate are plane parallel to increase the polarization, but the plates are too thick for the amplitudes of the multiple internally reflected beams to add or subtract. If the amplitudes of the beams could be added, there would be interference effects and the transmittance would vary with the thickness of each plate and with the wavelength. These are so-called interference polarizers and are mentioned below.

Transmission polarizers do not have the angle deviation problem of the reflection polarizers, but the transmitted beam may be slightly displaced parallel to the incident beam if the plates are thick. The main problem with transmission polarizers is that the light is not completely polarized even with several plates in a stack. If the plates are slightly absorbing, the transmission of the polarizer is reduced.

Miscellaneous Types

There are numerous other types of polarizers that mostly use thin films. Interference effects in thin films can increase the polarization efficiency at certain wavelengths depending on the thicknesses of the films and the design of the multilayer coating. Many of the applications involve non-normal incidence designs. For example, a single high refractive index film or a multilayer coating evaporated onto a low refractive index substrate increases the degree of polarization of reflection and transmission polarizers. Polarizing beamsplitters also use multilayer dielectric coatings. Free-standing films of different thicknesses were formerly used for infrared polarizers before wire grid polarizers became available.

Dichroic absorbers have been made from two-phase lamelar eutectics of thin needles of a conducting material embedded in a transparent matrix. Thin sheets of pyrolytic graphite material also act

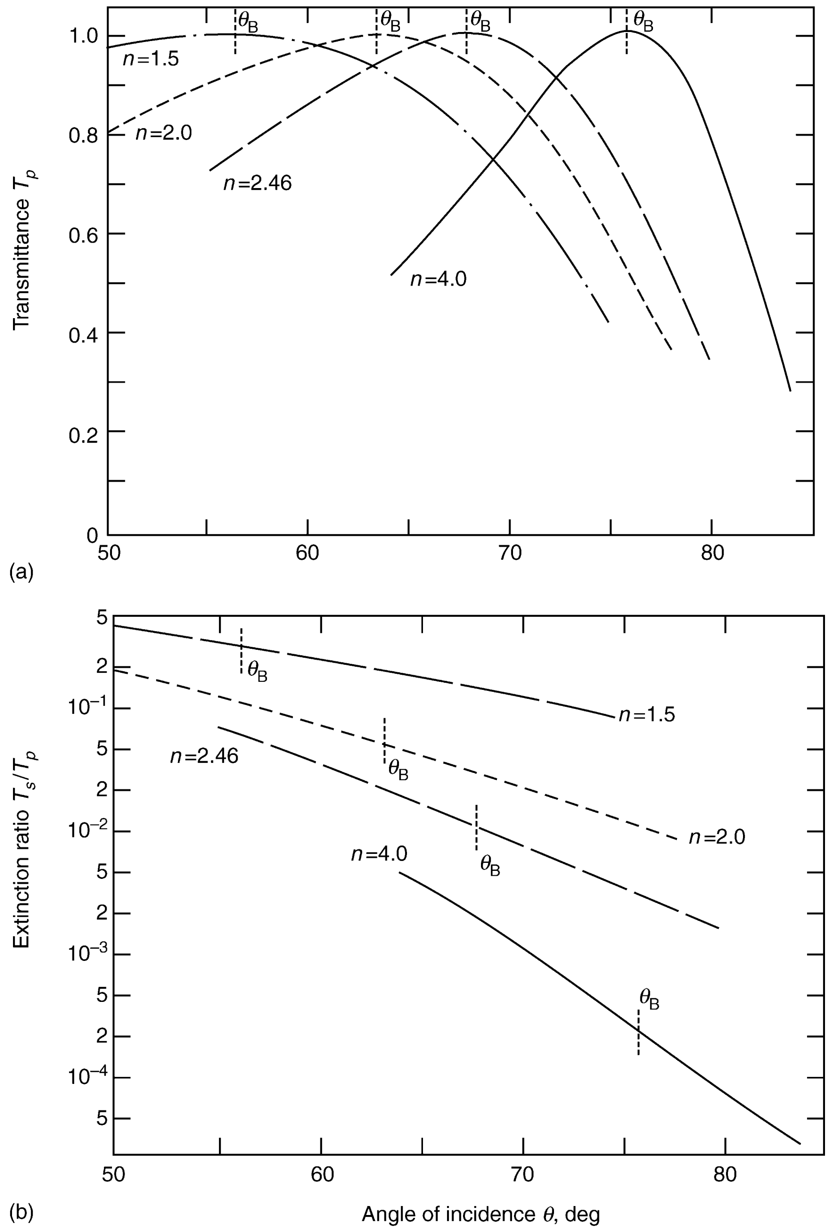

Figure 8 (a) Transmittance and (b) extinction ratio of four plane-parallel plates of refractive index n as a function of angle of incidence, for angles near the Brewster angle. Assumptions are multiple reflections but no interference within each plate and no reflections between plates. (Reproduced with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

as dichroic absorbers. There are also numerous miniature polarizer designs for fiber optics and nanotechnology applications.

Retarders or Retardation Plates

Retarders, or retardation plates, are devices that can change linearly polarized light into circularly or elliptically polarized light. They can also rotate the plane of polarization of linearly polarized light. A retardation plate is made from a uniaxial crystal that has an optic axis (as discussed above); the light travels in a direction perpendicular to the optic axis, as described below and shown in Figure 9. The ordinary and extraordinary rays travel at different velocities through the crystal depending on their refractive indexes: $n = c/v$, where c is the velocity of light in a vacuum and v is the velocity of light in the material. The larger the refractive index, the slower is the velocity of light in the material. The ordinary ray with a refractive index n_o and the extraordinary ray with a refractive index n_e have mutually perpendicular vibration directions. If the ray has a vibration direction at another azimuth in the crystal, the refractive index is intermediate between n_o and n_e. For a positive uniaxial crystal $n_e > n_o$ so the extraordinary ray travels slower than the ordinary ray through the crystal. Thus, the designation 'fast axis' is often used for the ordinary ray and 'slow axis' is used for the extraordinary ray, as shown in Figure 9. In a negative uniaxial crystal, $n_e < n_o$, i.e., the velocities along the two axes are reversed.

Because there is a velocity difference between the ordinary and extraordinary rays traveling through a uniaxial crystal, they get out of phase. Depending on their velocities and the optical thickness of the material (the refractive index times the physical thickness), the addition of their amplitudes results in a wave whose vibration direction is either: (a) perpendicular to the original vibration direction; (b) rotates in a circle (right or left circularly polarized light); or (c) rotates in an ellipse (right or left elliptically polarized light). The path difference N between two rays in the crystal, measured in terms of the wavelength of the light, is $N\lambda = \pm d(n_e - n_o)$, where d is the physical thickness of the material. The corresponding phase difference δ between the two rays is $\delta = 2\pi N = \pm(2\pi d/\lambda)(n_e - n_o)$. The quantity N is important because if beams of light vibrating along the fast and slow axes of a crystal get out of step by one quarter of the wavelength of the light, the device is called a quarter wave (or $\lambda/4$) retardation plate, or simply a quarter-wave plate. There are also half-wave ($\lambda/2$) plates that rotate the plane of polarization, and full-wave (λ) plates that rotate the plane of polarization through 180°. If light passes through a full-wave plate, theoretically it is indistinguishable from the original beam. However, materials are normally temperature sensitive so that as the temperature changes, the retardation is no longer exactly one wave, but may be slightly less than or greater than one wavelength. Other thicknesses of retardation plates produce elliptically polarized light.

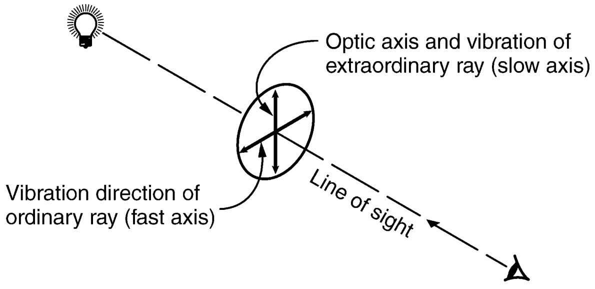
Optic axis and vibration of extraordinary ray (slow axis)
Vibration direction of ordinary ray (fast axis)
Line of sight

Figure 9 Light incident normally on the front surface of a retardation plate showing the vibration directions of the ordinary and extraordinary rays. In a positive uniaxial crystal, the fast and slow axes are as indicated in parentheses; in a negative uniaxial crystal, the two axes are interchanged. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

Figure 10 shows what happens to a beam of linearly polarized light when the axis of vibration is at 45° to the fast and slow axes of a positive

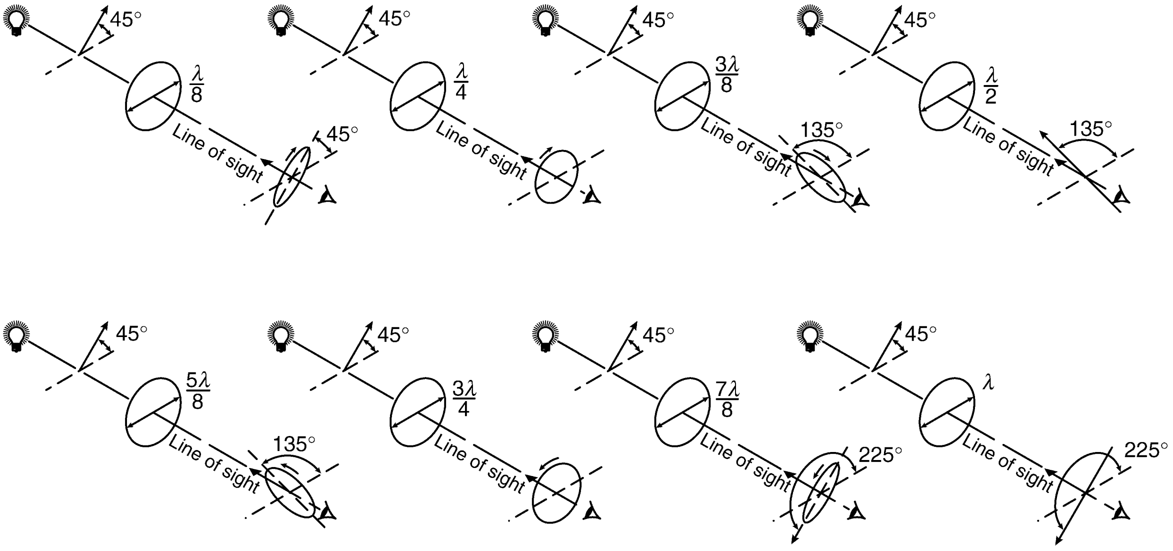

Figure 10 State of polarization of a light wave after passing through a crystal plate whose retardation is indicated in fractions of a wavelength (phase retardation $2\pi/\lambda$ times these values) and whose fast axis is indicated by the double arrow. In all cases the incident light is linearly polarized at an azimuth of 45° to the direction of the fast axis. (Adapted with permission from Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

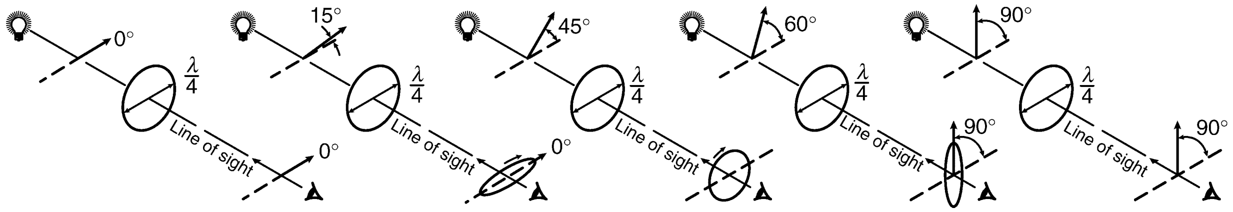

Figure 11 State of polarization of a light wave after passing through a $\lambda/4$ plate (whose fast axis is indicated by the double arrow) for different azimuths of the incident linearly polarized beam. (Adapted with permission from Bennett JM (1995) Polarization. In Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill, Inc.)

uniaxial crystal. The fast axis is in the horizontal direction, the same as in Figure 9.

The main purpose of a quarter-wave plate is to change linearly polarized light into circularly polarized light. However, as Figure 11 shows, the state of polarization of the light will be quite different depending on the orientation of the plane of polarization of the incident beam relative to the fast axis of the quarter-wave plate.

A half-wave plate is used to rotate the plane of polarization of a linearly polarized beam. The plane of polarization is always rotated through an angle that is twice the angle the initial plane of polarization makes with the fast axis of the uniaxial crystal. A linearly polarized beam always remains linearly polarized.

Retardation plates are often made of mica, stretched polyvinyl alcohol, or crystal quartz, although they can also be made of other stretched plastics, sapphire, magnesium fluoride, and other materials.

Variable Retardation Plates and Compensators

Variable retardation plates are devices whose retardation can be varied in a variety of ways. They can be used to modulate or vary the phase of a beam of linearly polarized light or to analyze a beam of unknown polarization (often elliptically polarized light) such as might be produced by transmission through a birefringent material or by reflection from a metal or film-covered surface. The term compensator is often applied to a variable retardation plate since it can be used to compensate for the phase retardation produced by a material. Common types of compensators include the Babinet and Soleil compensators, in which the total thickness of a birefringent material in the light path is changed, the Sénarmont compensator which consists of a fixed quarter-wave plate and rotatable analyzer to compensate for varying amounts of ellipticity in a light beam, and tilting-plate compensators, which change the thickness of birefringent material in the light beam by changing the angle of incidence. Electro-optic and piezo-optic modulators can also be used as high-frequency variable retardation plates since their birefringence can be changed by varying the electric field or pressure (see next section).

The Babinet compensator, shown schematically in Figure 12 consists of two crystal quartz wedges, each with its axis in the plane of the face but with the axes at right angles to each other. One wedge is stationary, and the other can be moved in the direction indicated by the arrow, so that the total amount of quartz through which the light passes can be varied. The total retardation is proportional to the difference in thickness between the two wedges. This type of compensator was used extensively when the light source was a vertical slit and visual measurements were made of the state of polarization of a beam. However the bands representing different phase retardations were too narrow to be used effectively with a photoelectric detector, so the Babinet compensator was replaced by a Soleil compensator (Figure 13). This device, sometimes called a Babinet–Soleil compensator, is similar to the Babinet compensator except that the field of view has a uniform tint if the compensator is constructed correctly. This is because the ratio of the thicknesses of the two crystal quartz blocks (a movable wedge and a fixed wedge attached to a plate with the two axes perpendicular to each other) is the same over the entire field. The Soleil compensator will produce light of varying ellipticity depending on the position of the movable wedge. It is used in the same way as a Babinet compensator with the uniformly dark field of the Soleil compensator corresponding to the black zero-retardation band in the Babinet compensator.

It is sometimes necessary to accurately measure the azimuth of a beam of linearly polarized light using a photoelectric detector with a rotatable analyzer (i.e., a polarizer) directly in front it. The obvious method is to rotate the analyzer until the detector signal is a minimum and then read the analyzer angle, which equals the extinction angle (perpendicular to the azimuth of the linearly polarized beam). However, the

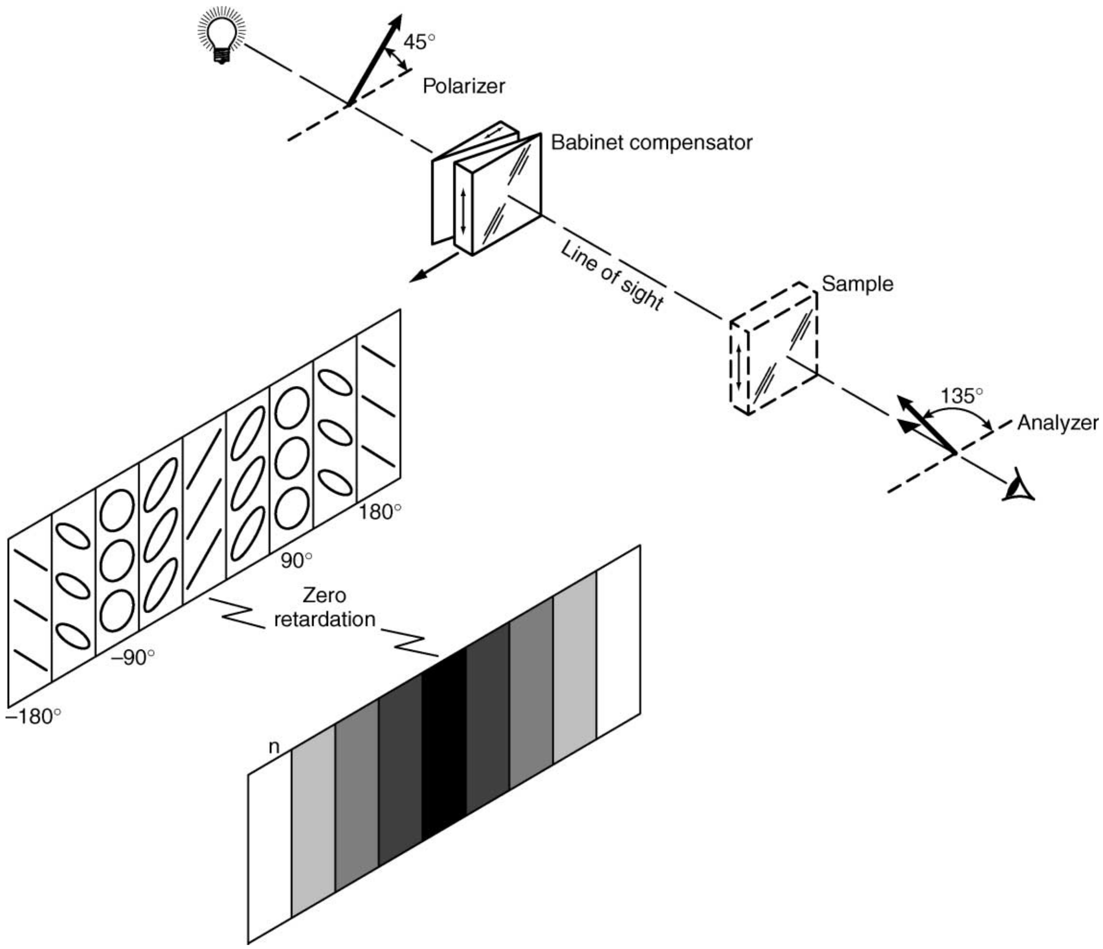

Figure 12 Arrangement of a Babinet compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. Retardations are indicated for alternate regions. After the beam passes through the analyzer, the field is crossed by a series of dark bands, one of which is shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

angle can be determined more precisely if the analyzer is offset by a small angle from the extinction angle and the transmittance noted. Then the analyzer is offset by a small angle on the other side of the extinction angle at the angle where the transmittance is the same. One half the difference between these two azimuthal angles gives a more accurate value of the extinction angle than can be obtained by measuring it directly.

Before the days of photoelectric detectors, half-shade devices were extensively used to determine the azimuth of a linearly polarized beam. The device consisted of two polarizers with their axes inclined at a small angle to each other. As the device was rotated, one part of the field became darker and the other part became lighter. At the match position, both parts of the field appeared equally bright. There were a variety of these devices as well as ellipticity half-shade devices that could detect very small amounts of ellipticity in a nominally linearly polarized beam and hence could verify when a compensator had completely converted elliptically polarized light into linearly polarized light.

Electro-Optic, Magneto-Optic, and Piezo-Optic Devices

The state of polarization of light can be rapidly altered by passing it through a material that has electro-optic, magneto-optic or piezo-optic properties. The voltage, magnetic field, or pressure are varied to make the material birefringent (see above). Some materials are isotropic without the applied field, i.e., the refractive index is the same in all directions. Other materials have an optic axis

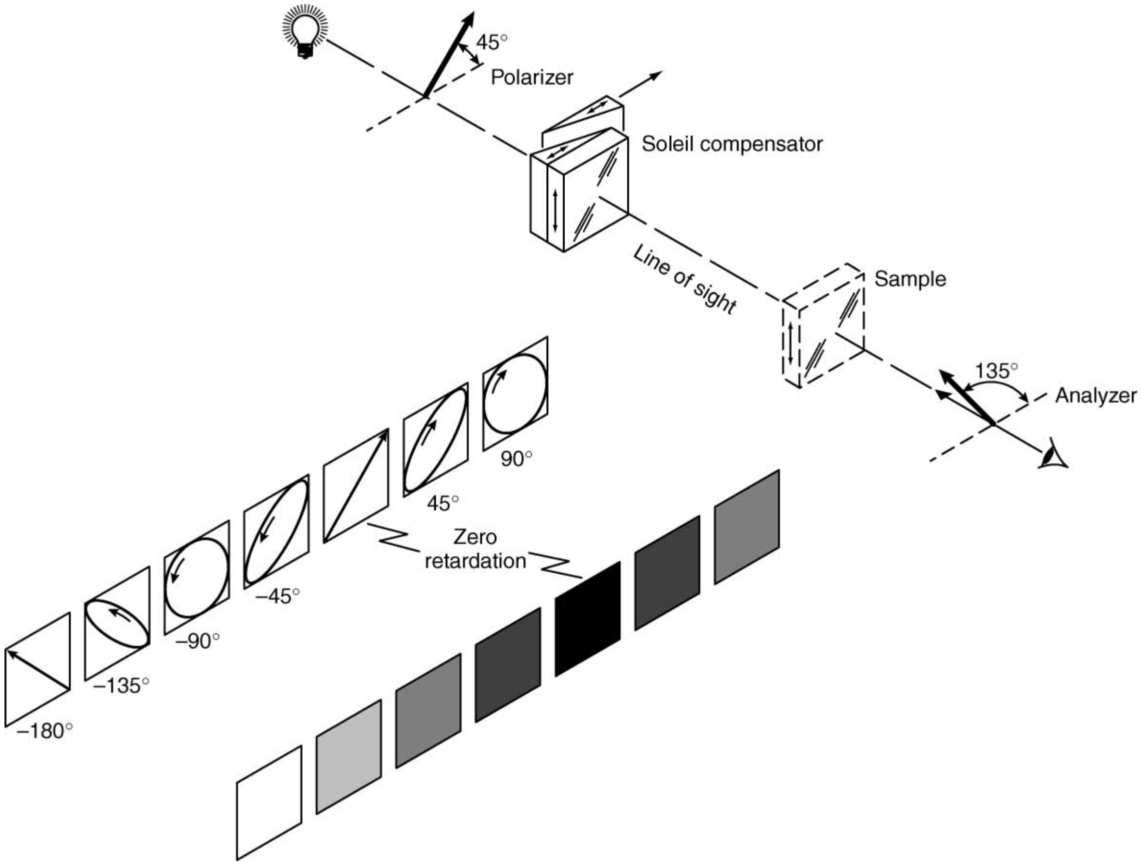

Figure 13 Arrangement of a Soleil compensator, polarizer, and analyzer for measuring the retardation of a sample. The appearance of the field after the light has passed through the compensator is shown to the left of the sample position. After the beam passes through the analyzer, the field appears as one of the shades of gray shown to the left of the analyzer. (Adapted with permission from Bennett JM (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill, Inc.)

(uniaxial materials) or two optic axes (biaxial materials). In these cases, the applied field creates further asymmetries in the material.

The most common electro-optic, magneto-optic and piezo-optic effects are the Pockel's effect (depending on the electric field), the Kerr effect (depending on the square of the electric field), the Faraday effect (depending on the magnetic field), the Cotton–Mouton and Voigt effects (depending on the square of the magnetic field), and stress birefringence or the photoelastic effect (depending on pressure changes). These effects are shown in Table 1 along with order of magnitude strengths needed to produce changes in the refractive index, and materials that most strongly exhibit the effects. The mathematical descriptions of the effects involve tensors and are too involved to present here. However, simple physical descriptions will be given.

If a varying electric field acts on an electro-optic material, electrons, ions, or permanent dipoles in the material are made to reorient, inducing an electric polarization. The induced polarization creates birefringence that modifies the optical polarization of a light beam passing through the material. The electric field strength determines how the polarization is changed (see Figure 10).

Linearly polarized light passing through a magneto-optic material will be rotated in a direction parallel to the direction of the applied magnetic field. This phenomenon, called the Faraday effect, or Faraday rotation, is similar to what happens in optically active materials (see the next paragraph) except that the Faraday effect depends on the direction of the magnetic field, but not on the direction the light is traveling.

There are naturally optically active materials such as camphor, nicotine, sugar solutions, and crystal quartz that can rotate the plane of linearly polarized light passing through the material. For example, the Si–O bonds in crystal quartz form a helical path around the optic axis. This crystal structure interacts with an incoming linearly polarized beam traveling parallel to the optic axis and rotates it in a clockwise direction. A 1 mm-thick piece of quartz will rotate a

Table 1 Electro-optic, magneto-optic, and piezo-optic effects

Effect	*Device*	*Effect proportional to*	*Strength* (Δn)	*Materials*
Pockels (electro-optic)	Pockels cell	E	$10^{-12}E$	$BaTiO_2$, $LiNBO_3$
Kerr	Kerr cell	E^2	$10^{-19}E^2$	Nitrobenzene, Benzonitrile
Faraday (magneto-optic)	Faraday rotation	H	$10^{-14}H$	Zns, GaAs, CS_2
Cotton–Mounton	[a]	H^2	$10^{-25}H^2$	Chloroform, acetone
Photoelastic effect; stress birefringence	[b]	Pressure	–	Fused silica, polystyrene, Ge, KDP, ruby

[a] Too small to be of technological importance.
[b] Depends on the mounting or the processing of the material.
Adapted with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

linearly polarized beam by 21.7°. Other materials have asymmetric structures that have mirror images. A mirror image structure cannot be obtained by simply rotating the group of atoms in space. These materials are also optically active. The amount of rotation produced by an optically active material is proportional to the thickness (optical density) of the material that the light passes through. Dextrose (a form of glucose) and levulose (also known as fructose) rotate the plane of linearly polarized light in clockwise and counterclockwise directions, respectively. Many other organic molecules have D- (right handed, dextro-rotary) and L- (left handed, levo-rotary) mirror image forms.

An isotropic material can become anisotropic when stress or an induced strain is applied because of an elasto-optic interaction known as stress birefringence, or the photoelastic effect. This effect is usually bad because it reduces the performance of optical components and devices by introducing phase distortions caused by improper mounting or unequal thermal expansion between the mounts and components. Another source of distortion is strain that has been frozen into optical components during manufacture. French curves are excellent examples that show colored strain birefringence patterns when viewed between crossed polarizers. In one application, strain birefringence has been used constructively to produce a variable phase retarder made from crystal quartz and fused silica. A block of electro-optic crystal quartz, is cemented to a block of isotropic fused silica. When a variable electric field is applied to the crystal quartz, its length changes at the resonant frequency of the block. This produces strain in the fused silica block which in turn produces a variable retardation of light passing through the strained fused silica. Depending on the magnitude of the varying electric field, the device can act like a variable quarter-wave plate, a variable half-wave plate, or have other applications.

Matrix Methods for Computing Polarization

An optical system containing various polarizing and retarding devices can be modeled using matrices. In general, there is an incident beam (in matrix form), that has some state of polarization. It interacts with a device called an instrument (also in matrix form) that alters the state of polarization, so that the exiting beam (a third matrix, the product of the first and second matrices) has another state of polarization. There can be more than one instrument matrix acting on the same incident matrix to produce the final matrix. The two most common matrix methods are the Mueller calculus and the Jones calculus.

There is also a visual representation of this process, a Poincaré sphere, on which vectors represent different states of polarization. The polarization instrument moves the vector representing the incident polarization around the sphere in a prescribed manner. The vectors are called Stokes vectors and are used in the Mueller calculus. The various states of polarization are represented on the Poincaré sphere as follows: The equator represents various forms of linear polarization, the poles represent right- and left-circular polarization, and other points on the sphere represent elliptically polarized light. Every point on the sphere corresponds to a different polarization form. The radius of the sphere represents the incident irradiance of the light beam (which is usually assumed to be unity). The effects of various polarization instruments are determined by displacements on the sphere. Partially polarized light or absorption may be dealt with approximately by ignoring the irradiance factor, since the state of polarization is generally the quantity desired. The Poincaré sphere is most useful for visualizing problems involving nonabsorbing materials, various polarization instruments including polarizers, retarders, compensators, half-shade devices, and depolarizers, and has also been applied to ellipsometric problems and stress-optical measurements.

The Poincaré sphere is used with Stokes vectors, which are often designated S_0, S_1, S_2, and S_3. S_0 is the incident irradiance of the light beam, corresponding to the radius of the Poincaré sphere. S_1 is the difference in irradiances between the horizontal and vertical polarization components of the beam; for example, when S_1 is positive, the preference is for horizontal polarization. S_2 indicates preference for $+45°$ or $-45°$ polarization depending on whether it is positive or negative, and S_3 gives the preference for right or left circular polarization. Thus, the Stokes vectors S_1, S_2, and S_3 are simply the three Cartesian coordinates of a point on the Poincaré sphere. S_1 and S_2 are perpendicular to each other in the equatorial plane, and S_3 points toward the north pole of the sphere. Any state of polarization of a light beam can be specified by these three vectors. The irradiance vector S_0 is related to the other three by the relation $S_0^2 = S_1^2 + S_2^2 + S_3^2$ when the beam is completely polarized. If the beam is partially polarized, $S_0^2 > S_1^2 + S_2^2 + S_3^2$.

In the Mueller calculus, the incident beam is represented by the four-component Stokes vector, written as a column vector. This vector has all real elements and gives information about irradiance properties of the beam. Thus it is not able to handle problems involving phase changes or combinations of two beams that are coherent. The instrument matrix is a 4×4 matrix with all real elements.

In the Jones calculus, the Jones vector representing the incident beam is a two-component column vector that generally has complex elements. It contains information about the *amplitude* properties of the beam and hence is well suited for handling problems involving the phases of light waves. However, it cannot handle problems involving depolarization, as the Mueller calculus can. The Jones instrument matrix is a 2×2 matrix whose elements are generally complex. The Jones calculus is well suited to problems involving a large number of similar devices arranged in series in a regular manner and makes it possible to obtain a result expressed explicitly in terms of the number of such devices. The Jones instrument matrix of a train of transparent or absorbing nondepolarizing polarizers and retarders contains no redundant information. The matrix contains four elements, each of which has two parts, so that there are a total of eight constants, none of which is a function of any other. The Mueller instrument matrix of such a train contains much redundancy; there are 16 constants but only 7 of them are independent. More information about the Poincaré sphere and the Mueller and Jones calculus is available in other references.

See also

Electromagnetically Induced Transparency. **Magneto-optics:** Faraday Rotation, CARS, ODMR, ODSR, Optical Pumping; Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering.

Further Reading

Auton JP (1967) Infrared transmission polarizers by photolithography. *Applied Optics* 6: 1023–1027.

Azzam, Rasheed MA and Bashara NM (1977) *Ellipsometry and Polarized Light*, chaps. 1 and 2. Amsterdam: North-Holland Publishing Co.

Azzam, Rasheed MA (1995) Ellipsometry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 27. New York: McGraw-Hill.

Bennett JM (1995) Polarization. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. I, chap. 5. New York: McGraw-Hill.

Bennett M (1995) Polarizers. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 3. New York: McGraw-Hill.

Bohren CF and Huffman DR (1983) *Absorption and Scattering of Light by Small Particles*. New York: John Wiley.

Born M and Wolf E (1999) *Principles of Optics*, 7th edn, chap. 1. Cambridge: Cambridge University Press.

Chipman RA (1995) Polarimetry. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 22. New York: McGraw-Hill.

Clark JR (1956) New calculus for the treatment of optical systems. VIII. Electromagnetic theory. *Journal of the Optical Society of America* 46: 126–131.

Guenther RD (1990) *Modern Optics*. chaps. 2, 13, and 14. New York: John Wiley and Sons.

Hass M and O'Hara M (1965) Sheet infrared transmission polarizers. *Applied Optics* 4: 1027–1031.

Hecht E and Zajac A (1997) *Optics*, 3rd edn, chap. 8. Reading, MA: Addison-Wesley Publishing Co.

Jackson JD (1999) *Classical Electrodynamics*, 3rd edn, chap. 7. New York: John Wiley and Sons, Inc.

Jasperson SN and Schnatterly SE (1969) An improved method for high reflectivity ellipsometry based on a new polarization modulation technique. *Reviews of Scientific Instruments* 40: 761–767.

Kerker M (1969) *The Scattering of Light and Other Electromagnetic Radiation*. New York: Academic Press.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*, chaps. 1–5. San Diego: Academic Press.

Maldonado TA (1995) Electro-optic modulators. In: Bass M (ed.) *Handbook of Optics*, 2nd edn, vol. II, chap. 13. New York: McGraw-Hill, Inc.

Markuvitz N (ed.) (1951) *Waveguide Handbook*, pp. 218–229, 285–289. New York: McGraw-Hill Book Co.

Shurcliff WA (1962) *Polarized Light, Production and Use*. Cambridge, MA: Harvard University Press.

van de Hulst HC (1957) *Light Scattering by Small Particles*. New York: John Wiley.

Matrix Analysis

B D Guenther, Duke University, Durham, NC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Hooke postulated, in the seventeenth century, that light waves must be transverse but his idea was forgotten. Young and Fresnel put forward the same idea in the nineteenth century and accompanied their postulation with a theoretical description of light based on transverse waves. Forty years later, Maxwell proved that light must be a transverse wave and that **E** and **H**, for a plane wave in an isotropic medium with no free charge and no currents, are mutually perpendicular and lie in a plane normal to the direction of propagation, **k**.

Convention requires that we use the electric vector to label the direction of the electromagnetic wave's polarization. The direction of the displacement vector is called the direction of polarization and the plane containing the direction of polarization and the propagation vector is called the plane of polarization. The selection of the electric field is not completely arbitrary, except in relativistic situations, when $v \approx c$, the interaction of the electromagnetic wave with matter will be dominated by the electric field.

Assume that a plane wave is propagating in the z-direction. In complex notation, the plane wave is given by

$$\tilde{\mathbf{E}} = \mathbf{E}_0\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi)} = \mathbf{E}_0\, e^{i(\omega t - kz + \phi)} \qquad [1]$$

The procedure used to decompose an arbitrary polarization into components parallel to two axes of a Cartesian coordinate system, is a technique used extensively in vector algebra to simplify mathematical calculations. According to the mathematical formalism associated with this technique, the polarization is described in terms of a set of basis vectors, $\mathbf{e}_i$. An arbitrary polarization would be expressed as

$$\mathbf{E} = \sum_{i=1}^{2} a_i \mathbf{e}_i \qquad [2]$$

The set of basis vectors, $\mathbf{e}_i$, are orthonormal, i.e.:

$$\mathbf{e}_i \mathbf{e}_j^* = \delta_{ij} = \left\{ \begin{matrix} 1 & i = j \\ 0 & i \neq j \end{matrix} \right\} \qquad [3]$$

where we have assumed that the basis vectors could be complex. We mention this mathematical formalism because an identical formalism is encountered in a description of spin.

In a Cartesian coordinate system, the $\mathbf{e}_i$'s are the unit vectors **i**, **j**, **k**. The summation in eqn [2] extends over only two terms because the electromagnetic wave is transverse, confining **E** to a plane normal to the direction of propagation (according to the coordinate convention we have selected, the **E** field is in the x, y plane).

Polarization Ellipse

Following the formalism of eqn [2], a polarized wave can be written in terms of the x and y components of $\mathbf{E}_0$

$$\tilde{\mathbf{E}} = \mathbf{E}_{0x}\, e^{i(\omega t - kz + \phi_1)}\hat{\mathbf{i}} + \mathbf{E}_{0y}\, e^{i(\omega t - kz + \phi_2)}\hat{\mathbf{j}} \qquad [4]$$

We will use only the real part of **E** for manipulation, to prevent errors. We divide each component of the electric field by its maximum value, so that the problem is reduced to one of the following two sinusoidal varying unit vectors:

$$\begin{aligned} \frac{E_x}{E_{0x}} &= \cos(\omega t - kz + \phi_1) \\ &= \cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1 \\ \frac{E_y}{E_{0y}} &= \cos(\omega t - kz)\cos\phi_2 - \sin(\omega t - kz)\sin\phi_2 \end{aligned} \qquad [5]$$

When these unit vectors are added together, the resulting equation will describe the path taken by the tip of the resultant vector. The path will create a Lissajous figure.

To obtain the equation for the Lissajous figure, we eliminate the dependence of the unit vectors on $(\omega t - kz)$. First, multiply the equations by $\sin\phi_2$ and $\sin\phi_1$, respectively and then subtract the resulting equations. Second, multiply the two equations by $\cos\phi_2$ and $\cos\phi_1$, respectively and then subtract the new equations. These two operations yield the following pair of equations:

$$\begin{aligned} &\frac{E_x}{E_{0x}}\sin\phi_2 - \frac{E_y}{E_{0y}}\sin\phi_1 \\ &\quad = \cos(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2] \end{aligned} \qquad [6]$$

$$\begin{aligned}\frac{E_x}{E_{0x}}\cos\phi_2 - \frac{E_y}{E_{0y}}\cos\phi_1 \\ = \sin(\omega t - kz)[\cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2]\end{aligned} \qquad [7]$$

The term in brackets can be simplified using the trig identity

$$\begin{aligned}\sin\delta &= \sin(\phi_2 - \phi_1) \\ &= \cos\phi_1 \sin\phi_2 - \sin\phi_1 \cos\phi_2\end{aligned} \qquad [8]$$

After replacing the term in brackets by sin δ, the two equations are squared and added yielding the equation for the Lissajous figure:

$$\left(\frac{E_x}{E_{0x}}\right)^2 + \left(\frac{E_y}{E_{0y}}\right)^2 - \left(\frac{2E_xE_y}{E_{0x}E_{0y}}\right)\cos\delta = \sin^2\delta \qquad [9]$$

The trig identity

$$\cos\delta = \cos(\phi_2 - \phi_1) = \cos\phi_1 \cos\phi_2 + \sin\phi_1 \sin\phi_2$$

was also used to further simplify eqn [9].

Equation [9] has the same form as the equation of a conic

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

Geometry defines the conic as an ellipse because, from eqn [9]

$$B^2 - 4AC = \frac{4}{E_{0x}^2 E_{0y}^2}(\cos^2\delta - 1) < 0$$

This ellipse is called the polarization ellipse. The orientation of the ellipse, with respect to the x-axis, is

$$\tan 2\theta = \frac{B}{A - C} = \frac{2E_{0x}E_{0y}\cos\delta}{E_{0x}^2 - E_{0y}^2} \qquad [10]$$

The tip of the resultant electric field vector obtained from eqn [4] traces out the polarization ellipse in the plane normal to **k**, as predicted by eqn [9]. The ratio of the length of the minor to the major axis of the ellipse is equal to the ellipticity, φ, i.e., the amount of deviation of the ellipse from a circle

$$\begin{aligned}\tan\varphi &= \pm\left(\frac{E_m}{E_M}\right) \\ &= \frac{E_{0x}\sin\phi_1 \sin\theta - E_{0y}\sin\phi_2 \cos\theta}{E_{0x}\cos\phi_1 \cos\theta + E_{0y}\cos\phi_2 \sin\theta}\end{aligned} \qquad [11]$$

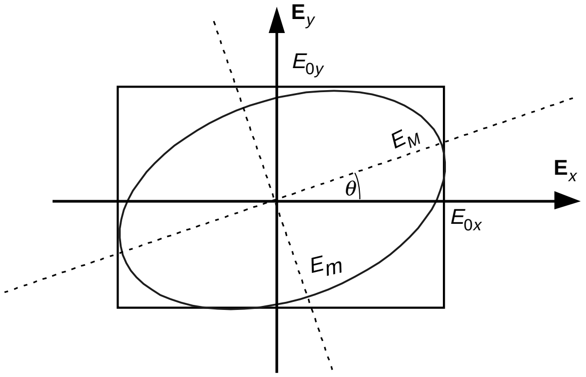

Figure 1 General form of the ellipse described by eqn [9]. Reproduced with permission from Guenther RD (1990) *Modern Optics*. New York: John Wiley and Sons.

In particle physics, the light would be said to have a negative helicity if it rotated in a clockwise direction. If we look at the source, the electric vector seems to follow the threads of a left-handed screw, agreeing with the nomenclature that left-handed quantities are negative. However, in optics the light that rotates clockwise, as we view it traveling toward us from the source, is said to be right-circularly polarized. The counterclockwise rotating light is left-circularly polarized. The association of right-circularly polarized light, with 'right-handedness' in optics, came about by looking at the path of the electric vector in space at a fixed time, then $\tan\psi = \tan(\phi - kz)$ (Figure 1).

Stokes Parameters

In the formalism associated with eqn [2], the expansion coefficients, a_i, can be used to form a 2–2 matrix which, in statistical mechanics, is called the density matrix and in optics, the coherency matrix (Table 1). The elements of the matrix are formed by the rule

$$\rho_{ij} = \mathbf{a}_i \mathbf{a}_j^* \qquad [12]$$

The matrix is Hermitian, so that $\rho_{ij} = \rho_{ji}^*$. We will not develop the theory of polarization using the coherency matrix, but simply use the coherency matrix to justify the need for four independent measurements to characterize polarization. There is no unique set of measurements required by the theory, but normally measurements made are of the Stoke's parameters which are directly related to the parameters of the polarization ellipse of eqn [9].

The Stokes parameters of a light wave are measurable quantities, defined as:

$s_0 \Rightarrow$ Total flux density
$s_1 \Rightarrow$ Difference between flux density transmitted by a

Table 1 Typical Stokes vectors

Horizontal polarization	$\begin{bmatrix}1\\1\\0\\0\end{bmatrix}$	Vertical polarization	$\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}$
+45° polarization	$\begin{bmatrix}1\\0\\1\\0\end{bmatrix}$	−45° polarization	$\begin{bmatrix}1\\0\\-1\\0\end{bmatrix}$
Right circular polarization	$\begin{bmatrix}1\\0\\0\\1\end{bmatrix}$	Left circular polarization	$\begin{bmatrix}1\\0\\0\\-1\end{bmatrix}$

linear polarizer oriented parallel to the *x*-axis and one oriented parallel to the *y*-axis. The *x* and *y*-axes are usually selected to be parallel to the horizontal and vertical directions in the laboratory.

s_2 ⇒ Difference between flux density transmitted by a linear polarizer oriented at 45° to the *x*-axis and one oriented at 135°.

s_3 ⇒ Difference between flux density transmitted by a right-circular polarizer and a left-circular polarizer.

If the Stokes parameters are to characterize the polarization of a wave, they must be related to the parameters of the polarization ellipse. It is therefore important to establish that the Stokes parameters are variables of the polarization ellipse (eqn [9]).

In its current form, eqn [9] contains no measurable quantities and thus must be modified if it is to be associated with the Stokes parameters. The time average of the Poynting vector is the quantity observed when measurements are made of light waves. We must, therefore, find the time average of eqn [9] if we wish to relate its parameters to observable quantities.

The time average eqn [9] can now be written as

$$\frac{\langle E_x^2\rangle}{E_{0x}^2} + \frac{\langle E_y^2\rangle}{E_{0y}^2} - 2\frac{\langle E_x E_y\rangle}{E_{0x}E_{0y}}\cos\delta = \sin^2\delta \qquad [13]$$

where time average is denoted by ⟨ ⟩

$$\langle E_x^2\rangle = \frac{1}{T}\int_{t_0}^{t_0+T} E_{0x}^2[\cos(\omega t - kz)\cos\phi_1 - \sin(\omega t - kz)\sin\phi_1]^2\,\mathrm{d}t \qquad [14]$$

With the time averages, eqn [13] can be written as

$$(E_{0x}^2 + E_{0y}^2)^2 - (E_{0x}^2 - E_{0y}^2)^2 - (2E_{0x}E_{0y}\cos\delta)^2 = (2E_{0x}E_{0y}\sin\delta)^2 \qquad [15]$$

Each term in this equation can be identified with a Stokes parameter:

$$s_0 = \langle E_{0x}^2\rangle + \langle E_{0y}^2\rangle \qquad s_1 = \langle E_{0x}^2\rangle - \langle E_{0y}^2\rangle$$
$$s_2 = \langle 2E_{0x}E_{0y}\cos\delta\rangle \qquad s_3 = \langle 2E_{0x}E_{0y}\sin\delta\rangle \qquad [16]$$

Equation [15] can now be written as

$$s_0^2 - s_1^2 - s_2^2 = s_3^2 \qquad [17]$$

For a polarized wave, only three of the Stokes parameters are independent. This agrees with the requirement placed upon elements of the Hermitian, coherency matrix, introduced above.

With this demonstration of the connection between the Stokes parameters and the polarization ellipse, the Stokes parameters can be written in terms of the parameters of the polarization ellipse.

$$\begin{aligned} s_1 &= s_0\cos 2\varphi\cos 2\theta \\ s_2 &= s_0\cos 2\varphi\sin 2\theta \\ s_3 &= s_0\sin 2\varphi \end{aligned} \qquad [18]$$

Mueller Calculus

Mueller pointed out that the Stokes parameters can be thought of as elements of a column matrix or a 4-vector (Table 2)

$$\mathcal{S} = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} \qquad [19]$$

Each element of the Stokes vector represents a measurable intensity. The vector can represent not only polarized light, but unpolarized or partially polarized light.

In 1929, Soleillet discovered that an optical device performed a linear transformation on the input wave and, in 1942, Perrin put this fact into a matrix formalism involving the Stokes vectors:

$$\mathcal{S}_1 = \mathcal{M}\cdot\mathcal{S}_0 \qquad [20]$$

Mueller used experimentally derived 4 × 4 matrices, the $\mathcal{M}$ in eqn [20], to describe the effect of an optical device on a light wave's polarization. The $\mathcal{S}$'s in eqn [20] are column matrices whose elements

Table 2 Mueller matrices for polarizers

Polarizer	*Transmission axis*	*Mueller matrix*
Linear	Horizontal	$\frac{1}{2}\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
	Vertical	$\frac{1}{2}\begin{bmatrix}1 & -1 & 0 & 0\\ -1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix}1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\\ 1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
	−45°	$\frac{1}{2}\begin{bmatrix}1 & 0 & -1 & 0\\ 0 & 0 & 0 & 0\\ -1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix}1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 1\end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix}1 & 0 & 0 & -1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ -1 & 0 & 0 & 1\end{bmatrix}$

Table 3 Mueller matrices for retarders

Retarder	*Transmission axis*	*Mueller matrix*
Quarter-wave plate	Horizontal	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & -1 & 0\end{bmatrix}$
	Vertical	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & 1 & 0\end{bmatrix}$
	+45°	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & -1\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}$
	−45°	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & -1 & 0 & 0\end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix}$
	±45°	$\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & -1\end{bmatrix}$

are the Stokes parameters (eqn [16]). The matrices are based upon an assumed linear relationship between the incident and transmitted beams (Table 3).

The analysis of the effect of a number of polarizers and retarders is made easier by the use of the Mueller–Stokes matrix calculus, coupled with the use of the Stokes vectors. To determine the Mueller matrices, one must measure the effect a device has on unpolarized, horizontally polarized, linearly polarized at + 45°, and right-circularly polarized light. It is then an algebraic exercise to generate the elements of the matrix. A few optical devices and their Mueller matrix are listed in Tables 2 and 3.

The Mueller matrix contains 16 parameters but there are only seven independent parameters. The matrix contains no information about absolute phase but it handles partially polarized and unpolarized light without modification (Table 4).

Jones Vector

There is one other representation of polarized light, complementary to the Stokes vector, developed by Clark Jones in 1941 and called the Jones vector. It is superior to the Stokes vector in that it handles light of a known phase and amplitude with a reduced number of parameters. However, it is inferior to the Stokes vector in that, unlike the Stokes representation, which is experimentally determined, the Jones representation cannot handle unpolarized or partially polarized light. The Jones vector is a theoretical construct that can only describe light with a well-defined phase and frequency. The density matrix formalism can be used to correct the shortcomings of the Jones vector, but then the simplicity of the Jones representation is lost.

Assuming the coordinate system is such that the electromagnetic wave is propagating along the z-axis, it was shown earlier that any polarization could be decomposed into two orthogonal **E** vectors, i.e., parallel to the x and y directions. The Jones vector is defined as a two-row, column matrix consisting of the complex components in the x and y direction:

$$\mathbf{E} = \begin{bmatrix} E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \\ E_{0y}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_2)} \end{bmatrix} \qquad [21]$$

Table 4 Jones vectors

Horizontal polarization	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	Vertical polarization	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
+45° polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	−45° polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$
Right circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$	Left circular polarization	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$

If absolute phase is not an issue, then we may normalize the vector by dividing by that number (real or complex) that simplifies the components but keeps the sum of the square of the components equal to one. For example, assume that $E_{0x} = E_{0y}$, then

$$\mathbf{E} = E_{0x}\, e^{i(\omega t - \mathbf{k}\cdot\mathbf{r} + \phi_1)} \begin{bmatrix} 1 \\ e^{i\delta} \end{bmatrix} \qquad [22]$$

The normalized vector would be the terms contained within the bracket, each divided by $\frac{1}{\sqrt{2}}$.

The general form of the Jones vector is

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \qquad \mathbf{E}^* = [\mathbf{A}^*\mathbf{B}^*] \qquad [23]$$

Some examples of Jones vectors are shown in Table 4.

Jones Calculus

The Jones calculus is complementary to the Mueller calculus and operates on the Jones vector, (eqn [23]), similar to the way the Mueller matrix operates on the Stokes vector (Table 5).

The Jones matrix contains eight independent parameters with no redundancy, making it simpler than the Mueller calculus. However, the Jones calculus only applies to polarized light. The Jones calculus can be extended, using the density matrix formalism to allow manipulation of unpolarized

Table 5 Jones matrices for retarders

Retarder	*Transmission axis*	*Jones matrix*
	Horizontal	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
	Vertical	$\begin{bmatrix} e^{-i\pi/4} & 0 \\ 0 & e^{i\pi/4} \end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$
Quarter-wave plate	−45°	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}$
Half-wave plate	0° or 90°	$\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$
	±45°	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Table 6 Jones matrices for polarizers

Polarizer	*Transmission axis*	*Jones matrix*
Linear	Horizontal	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	Vertical	$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$
	+45°	$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
	−45°	$\frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$
Circular	Right	$\frac{1}{2}\begin{bmatrix} 1 & -i \\ i & 1 \end{bmatrix}$
	Left	$\frac{1}{2}\begin{bmatrix} 1 & i \\ -i & 1 \end{bmatrix}$

light, but with a loss of simplicity. The matrix equation for Jones calculus is

$$\mathbf{E}_{\text{out}} = \mathcal{W}\mathbf{E}_{\text{in}} \qquad [24]$$

A few optical devices and their Jones matrices, $\mathcal{W}$, are listed in Tables 5 and 6.

For every matrix in Jones calculus, there is a matrix in Mueller calculus, but the converse is not true. For example, it is possible to construct a depolarizer, by using a thick piece of opal glass in the visible or by using gold covered sandpaper in the infrared. Such a device can be described in Mueller calculus, but there is no matrix for such a device in Jones calculus (Tables 5 and 6).

See also

Polarization: Introduction.

Further Reading

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Clark Jones R (1941) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 32: 488–503.

Clark Jones R (1942) A new calculus for the treatment of optical systems. *Journal of the Optical Society of America* 31: 486–493.

Guenther RD (1990) *Modern Optics*. New York: Wiley.

Hecht E (1998) *Optics*, 3rd edn. Reading, MA: Addison-Wesley.

Klein MV (1970) *Optics*. New York: Wiley.

Kliger DS, Lewis JW and Randall CE (1990) *Polarized Light in Optics and Spectroscopy*. Boston, MA: Academic Press.

McMaster WH (1954) Polarization and the Stokes parameters. *American Journal of Physics* 22: 351–362.

Mueller H (1948) The foundation of optics. *Journal of the Optical Society of America* 38: 661.

Q

QUANTUM ELECTRODYNAMICS

Contents

Quantum Theory of the Electromagnetic Field
Cavity QED
Cavity QED in Semiconductors

Quantum Theory of the Electromagnetic Field

I Bialynicki-Birula and Z Bialynicka-Birula,
Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

The quantization of the free electromagnetic field is explained by analogy with the quantization of a collection of harmonic oscillators.

Brief History

The first step in the formulation of the quantum theory of the electromagnetic field (EMF) was the 'quantum hypothesis' put forward in 1900 by Planck. According to this hypothesis the exchange of energy between the EMF and the cavity walls occurs in portions $h\nu$, where h – the quantum of action – is a new fundamental constant (the Planck constant) and ν is the frequency of the material oscillator in the wall. With the help of this hypothesis Planck explained the spectrum of the electromagnetic radiation in thermodynamic equilibrium with the cavity walls. The second step was made in 1905 by Einstein, who postulated that the quantum hypothesis of Planck should be applied directly to the EMF. The assumption that the energy of the electromagnetic radiation is carried by light quanta (named photons by Lewis in 1926), each of energy $h\nu$, enabled Einstein to explain the puzzling properties of the photoelectric effect, namely, the gross disparity between the tiny classical energy flux delivered to an electron and the amount of energy needed to eject the photo-electron from the metal. After the formulation of quantum mechanics, it became possible to complete the quantum theory of the EMF (Dirac 1927) by applying the general rules of quantization to the classical Maxwell theory. However, this theory ran into difficulties when it was applied to the interaction of the EMF with charged matter.

The lightest carriers of the electric charge are the electrons and their dynamics determines to a very large extent the electromagnetic properties of ordinary matter. The main principles of quantum electrodynamics (QED) – the theory unifying the quantum theory of the EMF and the relativistic theory of electrons (Dirac 1928) – have been set forth by Heisenberg and Pauli already in 1929 but the proper tool (renormalization theory) to perform highly accurate calculations was developed by Feynman and Schwinger in 1949. In 1967 QED was unified by Weinberg and Salam with the theory of weak interactions. Later, after the unification with the theory of strong interactions, QED became part of the so called 'standard model' – the present day theory of elementary ingredients of matter: quarks, leptons, gluons, intermediate bosons, and photons. Therefore, the modern quantum theory of the EMF is not a self-contained theory but part of a much more encompassing theory. Often it requires some input from a more general theory or from phenomenology. However, the quantum theory of the EMF with only a very crude description of its coupling to the rest of the world (for example, in the form of boundary conditions or constitutive relations) is still not only a good starting point to a more elaborate theory but also a source of important information about the behavior of photons under realistic conditions. The main reason why treating photons as free particles gives a reasonably good approximation is the smallness of the Sommerfeld fine structure constant $\alpha = e^2/4\pi\varepsilon_0\hbar c \approx 1/137$ that measures the strength of the electromagnetic interactions between photons and charged particles.

Electromagnetic Field as a Collection of Harmonic Oscillators

In order to apply the same quantization procedure as in quantum mechanics, it is useful to formulate classical electrodynamics in a way that exhibits its close relationship with mechanical systems. Many properties of the electromagnetic field may be described and understood in the simplest case of the free field. Therefore, in this article we restrict ourselves to the quantum description of the free EMF.

It was discovered by Rayleigh and Jeans that the electromagnetic field may be viewed as a collection of harmonic oscillators. This analogy enables one to apply well-established methods of simple quantum mechanics to describe the much more complicated problem of the quantum field. The dynamics of a single, one-dimensional harmonic oscillator is governed by the Hamiltonian:

$$H(p,q) = \frac{p^2}{2m} + \frac{m\omega^2 q^2}{2}$$

The equations of motion for the position and momentum of a harmonic oscillator have the same form in classical and in quantum theory:

$$\frac{\mathrm{d}q(t)}{\mathrm{d}t} = \frac{p(t)}{m}, \qquad \frac{\mathrm{d}p(t)}{\mathrm{d}t} = -m\omega^2 q(t)$$

The solutions of these equations describe harmonic oscillations with the frequency ω

$$q(t) = q_0 \cos(\omega t) + (p_0/m\omega)\sin(\omega t)$$

$$p(t) = -m\omega q_0 \sin(\omega t) + p_0 \cos(\omega t)$$

A complete description of a multidimensional harmonic oscillator may be obtained in terms of normal modes. The transformation to normal modes leads to a decomposition of the Hamiltonian of the system into the sum, one term for each mode,

$$H = \sum_k \left(\frac{p_k^2}{2m} + \frac{m\omega_k^2 q_k^2}{2} \right)$$

Thus, each mode may be viewed as a separate system – a one-dimensional harmonic oscillator with characteristic frequency ω_k – noninteracting with the remaining oscillators.

The Maxwell equations for the free EMF read

$$\partial_t \mathbf{B}(\mathbf{r},t) = -\nabla \times \mathbf{D}(\mathbf{r},t)/\varepsilon, \ \nabla\cdot\mathbf{B}(\mathbf{r},t) = 0$$

$$\partial_t \mathbf{D}(\mathbf{r},t) = \nabla \times \mathbf{B}(\mathbf{r},t)/\mu, \ \nabla\cdot\mathbf{D}(\mathbf{r},t) = 0$$

In the vacuum, the constants ε and μ must be replaced by ε_0 and μ_0. The Maxwell equations lead to the wave equation for the vector $\mathbf{D}(\mathbf{r},t)$

$$\left(\frac{1}{c^2}\partial_t^2 - \Delta\right)\mathbf{D}(\mathbf{r},t) = 0$$

and to the same equation for the vector $\mathbf{B}(\mathbf{r},t)$. The solution of this equation obtained by the separation of variables has the form

$$\mathbf{D}(\mathbf{r},t) = a(t)\mathbf{u}(\mathbf{r})$$

where $a(t)$ satisfies the harmonic equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}a(t) + \omega^2 a(t) = 0$$

and $\mathbf{u}(\mathbf{r})$ satisfies the Helmholtz equation

$$(\Delta + (\omega/c)^2)\mathbf{u}(\mathbf{r}) = 0$$

Without going into mathematical details we shall assume that there exist a complete set of vector functions $\mathbf{u}_k(\mathbf{r})$, called the mode functions, satisfying the Helmholtz equations

$$(\Delta + (\omega_k/c)^2)\mathbf{u}_k(\mathbf{r}) = 0$$

the divergence condition $\nabla \cdot \mathbf{u}_k(\mathbf{r}) = 0$, the orthonormality conditions

$$\int \mathrm{d}^3 r \mathbf{u}_k(\mathbf{r}) \cdot \mathbf{u}_l(\mathbf{r}) = \delta_{kl}$$

and also the appropriate boundary conditions. In most cases of physical interest such a set does exist.

Given the set of mode functions $\mathbf{u}_k(\mathbf{r})$ one may expand the vector field $\mathbf{D}(\mathbf{r},t)$ at each t into the series

$$\mathbf{D}(\mathbf{r},t) = -\sum_k p_k(t)\mathbf{u}_k(\mathbf{r})$$

The minus sign is purely conventional; it makes the correspondence with the harmonic oscillator a closer one. The same procedure may be repeated for the vector **B** with a different complete set of functions $\mathbf{v}_k(\mathbf{r})$, resulting in the decomposition

$$\mathbf{B}(\mathbf{r}, t) = \sum_k q_k(t)\mathbf{v}_k(\mathbf{r})$$

The Maxwell equations are satisfied if the coefficient functions $p_k(t)$ and $q_k(t)$ obey the equations of motion for harmonic oscillators

$$\frac{\mathrm{d}q_k(t)}{\mathrm{d}t} = \frac{p_k(t)}{\varepsilon}, \quad \frac{\mathrm{d}p_k(t)}{\mathrm{d}t} = -\varepsilon\omega^2 q_k(t) \qquad [1]$$

and the functions $\mathbf{v}_k(\mathbf{r})$ are related to $\mathbf{u}_k(\mathbf{r})$ by the formulas

$$\mathbf{v}_k(\mathbf{r}) = \nabla \times \mathbf{u}_k(\mathbf{r})$$

Note that the electric constant ε plays the same role as the mass for a mechanical oscillator. The calculation of the Hamiltonian, the total field energy,

$$H = \int \mathrm{d}^3r \left(\frac{\mathbf{D}^2}{2\varepsilon} + \frac{\mathbf{B}^2}{2\mu} \right)$$

completes the reduction of the electromagnetic field to a collection of harmonic oscillators. Upon inserting the representation of the field vectors in terms of the complete set of functions $\mathbf{u}_k(\mathbf{r})$ into the integral, one obtains the following formula

$$H = \sum_k \left(\frac{p_k^2}{2\varepsilon} + \frac{\varepsilon\omega_k^2 q_k^2}{2} \right)$$

which differs from its mechanical counterpart only by the replacement $m \to \varepsilon$. The time dependence has been omitted here because the Hamiltonian is a constant of motion.

Mode Functions

The simplest example of a complete set of mode functions is that obtained for a rectangular metallic cavity. In this case the functions $\mathbf{u}_k(\mathbf{r})$ may be expressed in terms of trigonometric functions. In order to write them down, it is convenient to replace one label k by a collection of three numbers $\boldsymbol{\kappa} = \{\kappa_x, \kappa_y, \kappa_z\}$ and an additional index λ taking on the values 1 and 2. The mode functions $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ satisfying the proper boundary conditions have the following components

$$\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r}) = N \begin{pmatrix} a_x^{(\lambda)} \cos(\kappa_x x)\sin(\kappa_y y)\sin(\kappa_z z) \\ a_y^{(\lambda)} \sin(\kappa_x x)\cos(\kappa_y y)\sin(\kappa_z z) \\ a_z^{(\lambda)} \sin(\kappa_x x)\sin(\kappa_y y)\cos(\kappa_z z) \end{pmatrix}$$

where $N = \sqrt{8/(L_x L_y L_z)}$, the coefficients $\{\kappa_x, \kappa_y, \kappa_z\}$ have the form $\{\kappa_x, \kappa_y, \kappa_z\} = \{\pi n_x/L_x, \pi n_y/L_y, \pi n_z/L_z\}$ and all the n's are natural numbers. The components of both vectors $\mathbf{a}^{(\lambda)}$ are subjected to the conditions $\boldsymbol{\kappa}\cdot\mathbf{a}^{(\lambda)} = 0$ (vanishing divergence) and orthonormalization $\mathbf{a}^{(\lambda)}\cdot\mathbf{a}^{(\lambda')} = \delta_{\lambda\lambda'}$. For each set $\{\kappa_x, \kappa_y, \kappa_z\}$ one must choose two independent solutions of these conditions $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$. These two vectors define two states of linear polarization. The mode function $\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$ is a solution of the Helmholtz equation for the value of ω/c equal to the length of the vector $\boldsymbol{\kappa}$. The expansion of the EMF vectors into the mode functions in a rectangular cavity has the form

$$\mathbf{D}(\mathbf{r}, t) = -\sum_{\boldsymbol{\kappa}\lambda} p_{\boldsymbol{\kappa}\lambda}(t)\mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

$$\mathbf{B}(\mathbf{r}, t) = \sum_{\boldsymbol{\kappa}\lambda} q_{\boldsymbol{\kappa}\lambda}(t)\nabla \times \mathbf{u}_{\boldsymbol{\kappa}\lambda}(\mathbf{r})$$

where the summation extends over all allowed values of $\boldsymbol{\kappa}$ and two values of λ. One may explicitly write down mode functions also for a cylindrical and for a spherical cavity. They involve Bessel functions and are obtained by solving the Helmholtz equation in cylindrical and spherical coordinates, respectively.

One often uses another set of mode functions that does not correspond to any physical boundary conditions but is very helpful in the theoretical analysis. It is obtained by imposing the so-called periodic boundary conditions that were first introduced by Born in the study of crystal lattices. The periodic boundary conditions require that the electromagnetic field is periodic in space; it does not change when the coordinates are displaced by the 'period of the lattice' in all three directions. Assuming, for simplicity, that this period is the same, equal to L, in all directions, one obtains the following conditions for the mode functions

$$\mathbf{u}_k(\mathbf{r}) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_x) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_y) = \mathbf{u}_k(\mathbf{r} + L\mathbf{i}_z)$$

where $\mathbf{i}_i$ are the unit vectors in the directions of the coordinate axes. The complete set of solutions of the Helmholtz equation satisfying the periodic boundary conditions is labeled by the wavevector $\mathbf{k} = \{2\pi n_x/L, 2\pi n_y/L, 2\pi n_z/L\}$, where all the n's are arbitrary integers, and by the polarization index λ. The value of ω/c in the Helmholtz equation is equal in this case to the length of the vector k. Each member of the set $\mathbf{u}_{\mathbf{k}\lambda}$ has the form

$$\mathbf{u}_{\mathbf{k}\lambda}(\mathbf{r}) = \frac{1}{\sqrt{V}} \mathbf{e}_{\mathbf{k}\lambda} \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \qquad [2]$$

where $V = L^3$. The polarization vectors $\mathbf{e}_{\mathbf{k}\lambda}$ must be orthogonal to the wave vector $\mathbf{k}$. It is convenient to impose the unitary orthonormality conditions $\mathbf{e}_{\mathbf{k}\lambda} \cdot \mathbf{e}^*_{\mathbf{k}\lambda} = \delta_{\lambda\lambda'}$ and also to assume that $\mathbf{e}^*_{\mathbf{k}\lambda} = \mathbf{e}_{-\mathbf{k}\lambda}$. The expansion of the EMF vectors into the mode functions obeying the periodic boundary conditions has the form

$$\mathbf{D}(\mathbf{r},t) = -\frac{1}{\sqrt{V}} \sum_{\mathbf{k}\lambda} P_{\mathbf{k}\lambda}(t) \mathbf{e}_{\mathbf{k}\lambda} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$$

$$\mathbf{B}(\mathbf{r},t) = \frac{1}{\sqrt{V}} \sum_{\mathbf{k}\lambda} Q_{\mathbf{k}\lambda}(t) \mathrm{i}\mathbf{k} \times \mathbf{e}_{\mathbf{k}\lambda} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \qquad [3]$$

where the sum extends again over all allowed values of the vectors $\mathbf{k}$ and over two values of λ. The expansion coefficients $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are now complex. In order to guarantee that the field vectors are real, they must obey the conditions

$$P^*_{\mathbf{k}\lambda} = P_{-\mathbf{k}\lambda}, \; Q^*_{\mathbf{k}\lambda} = Q_{-\mathbf{k}\lambda}$$

These conditions are satisfied if $P_{\mathbf{k}\lambda}$ and $Q_{\mathbf{k}\lambda}$ are parametrized by the real (but otherwise arbitrary) variables $p_{\mathbf{k}\lambda}$ and $q_{\mathbf{k}\lambda}$

$$P_{\mathbf{k}\lambda} = \frac{p_{\mathbf{k}\lambda} + p_{-\mathbf{k}\lambda}}{2} + \frac{\varepsilon\omega_{\mathbf{k}}(q_{\mathbf{k}\lambda} - q_{-\mathbf{k}\lambda})}{2i}$$

$$Q_{\mathbf{k}\lambda} = \frac{q_{\mathbf{k}\lambda} + q_{-\mathbf{k}\lambda}}{2} + \frac{\mathrm{i}(p_{\mathbf{k}\lambda} - p_{-\mathbf{k}\lambda})}{2\varepsilon\omega_{\mathbf{k}}}$$

The energy of the EMF splits again into a sum of independent contributions from all the modes

$$H = \sum_{\mathbf{k}\lambda} \omega_{\mathbf{k}} \left(\frac{p_{\mathbf{k}\lambda} p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}} q_{\mathbf{k}\lambda} q_{\mathbf{k}\lambda}}{2} \right)$$

An important property of the periodic mode functions [2] is that not only the Hamiltonian but also the total momentum of the field splits into a sum of independent contributions. The total momentum $\mathbf{P}$ of the EMF is

$$\mathbf{P} = \int \mathrm{d}^3 r \, \mathbf{D} \times \mathbf{B}$$

Upon substituting into this expression the expansions [3] of the field vectors, one obtains, with the use of the orthonormality conditions,

$$\mathbf{P} = \sum_{\mathbf{k}\lambda} \mathbf{k} \left(\frac{p_{\mathbf{k}\lambda} p_{\mathbf{k}\lambda}}{2\varepsilon\omega_{\mathbf{k}}} + \frac{\varepsilon\omega_{\mathbf{k}} q_{\mathbf{k}\lambda} q_{\mathbf{k}\lambda}}{2} \right)$$

Canonical Quantization

The canonical quantization in quantum mechanics consists in replacing the c-numbers representing the classical canonical variables q_k and p_k by the operators $\hat{q}_k$ and $\hat{p}_k$. These operators are assumed to obey the canonical commutation relations

$$[\hat{q}_k, \hat{p}_l] = \mathrm{i}\hbar\delta_{kl} \qquad [4]$$

The canonical quantization of the EMF may proceed along the same lines as in quantum mechanics. All the variables q_k and p_k appearing in the decomposition into normal modes are replaced by the operators. When this is done, the field vectors $\mathbf{B}$ and $\mathbf{D}$ become the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ and the expression for the energy of the field becomes the Hamiltonian operator $\hat{H}$

$$\hat{H} = \sum_k \frac{\hat{p}_k^2}{2\varepsilon} + \frac{\varepsilon\omega_k^2 \hat{q}_k^2}{2}$$

The only difference between quantum mechanics and quantum field theory is that in the second case the number of canonical variables is infinite but this may cause complications only when one wants to give the quantum theory of the EMF a firm mathematical foundation. In physical applications one may think of the number of field oscillators as being very large but finite.

Time evolution of a quantum system is described by associating the time dependence either with the operators (the Heisenberg picture) or with the state vectors (the Schrödinger picture). In both pictures the time dependence is governed by the system Hamiltonian. In the Heisenberg picture the time dependence of the operators is given by the commutators with $\hat{H}$

$$\frac{\mathrm{d}\hat{q}_k(t)}{\mathrm{d}t} = \frac{1}{\mathrm{i}\hbar}[\hat{q}_k(t), \hat{H}] = \frac{\hat{p}_k(t)}{\varepsilon}$$

$$\frac{\mathrm{d}\hat{p}_k(t)}{\mathrm{d}t} = \frac{1}{\mathrm{i}\hbar}[\hat{p}_k(t), \hat{H}] = -\varepsilon\omega_k^2 \hat{q}_k(t)$$

These equations have the same form as in the classical theory (cf. eqn [1]). Therefore, in the Heisenberg picture the field operators will also obey the same equations as in the classical theory – the Maxwell equations:

$$\partial_t \hat{\mathbf{B}}(\mathbf{r},t) = -\nabla \times \hat{\mathbf{D}}(\mathbf{r},t)/\varepsilon, \;\; \nabla\cdot\hat{\mathbf{B}}(\mathbf{r},t) = 0$$

$$\partial_t \hat{\mathbf{D}}(\mathbf{r},t) = \nabla \times \hat{\mathbf{B}}(\mathbf{r},t)/\mu, \;\; \nabla\cdot\hat{\mathbf{D}}(\mathbf{r},t) = 0$$

Of course, the average values of the field operators evaluated in both pictures are the same and for every quantum state they obey the classical Maxwell equations.

The commutation relations [4] for the canonical operators and the completeness of the mode functions imply the following equal-time commutation relations for the field operators

$$[\hat{B}_i(\mathbf{r},t),\hat{D}_j(\mathbf{r}',t)] = \mathrm{i}\hbar \sum_k \varepsilon_{ijk}\partial_k \delta^{(3)}(\mathbf{r}-\mathbf{r}')$$

These relations do not depend on the choice of the mode functions. They summarize in a universal manner the canonical properties of the electromagnetic field operators.

Annihilation and Creation Operators

The canonical variables q_k and p_k are convenient for bringing home the analogy between mechanical oscillators and wave fields. However, there exists also a complementary, corpuscular aspect of the quantum theory of the EMF and that is most easily expressed in terms of photons. The best way to describe photons in quantum field theory is in terms of annihilation and creation operators. These operators were introduced by Schrödinger in the context of the quantum-mechanical harmonic oscillator but they have become really indispensable in quantum field theory. Formally, the annihilation $\hat{a}_k$ and the creation $\hat{a}_k^\dagger$ operators are defined as the following complex combinations of canonical operators

$$\hat{a}_k = \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k + \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right)$$

$$\hat{a}_k^\dagger = \sqrt{\frac{\varepsilon\omega}{2\hbar}}\left(\hat{q}_k - \mathrm{i}\frac{\hat{p}_k}{\varepsilon\omega}\right) \qquad [5]$$

Annihilation and creation operators are dimensionless, one is the Hermitian adjoint of the other, and they obey very simple commutation relations:

$$[\hat{a}_k, \hat{a}_l^\dagger] = \delta_{kl} \qquad [6]$$

The Hamiltonian expressed in terms of annihilation and creation operators reads

$$\hat{H} = \sum_k \hbar\omega_k \hat{a}_k^\dagger \hat{a}_k + \frac{1}{2}\sum_k \hbar\omega_k \qquad [7]$$

This Hamiltonian determines the following time dependence of the annihilation and creation operators

$$\hat{a}_k(t) = \hat{a}_k \exp(-\mathrm{i}\omega_k t),\ \hat{a}_k^\dagger(t) = \hat{a}_k^\dagger \exp(\mathrm{i}\omega_k t)$$

The field operators expressed in terms of annihilation and creation operators corresponding to the periodic boundary conditions are

$$\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar\varepsilon\omega_{\mathbf{k}}}{2V}}\hat{a}_{\mathbf{k}\lambda}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

$$\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_{\mathbf{k}\lambda}\sqrt{\frac{\hbar}{2V\varepsilon\omega_{\mathbf{k}}}}\hat{a}_{\mathbf{k}\lambda}\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \mathrm{H.c.}$$

where H.c. stands for the Hermitian conjugate of the first term. In the limiting case for infinite empty space, when $V \to \infty$, the Fourier sum becomes the Fourier integral:

$$\hat{\mathbf{D}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda \int \mathrm{d}^3k \sqrt{\frac{\hbar\varepsilon\omega_{\mathbf{k}}}{2(2\pi)^3}}\left(\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} - \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\right)$$

$$\hat{\mathbf{B}}(\mathbf{r},t) = \mathrm{i}\sum_\lambda \int \mathrm{d}^3k \sqrt{\frac{\hbar}{2\varepsilon\omega_{\mathbf{k}}(2\pi)^3}}\left(\hat{a}_\lambda(\mathbf{k})\mathbf{k}\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} - \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{k}\times\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\right)$$

In this case, the commutation relations involve the (three-dimensional) Dirac delta function

$$[\hat{a}_\lambda(\mathbf{k}), \hat{a}_{\lambda'}^\dagger(\mathbf{k}')] = \delta_{\lambda\lambda'}\delta^{(3)}(\mathbf{k}-\mathbf{k}') \qquad [8]$$

The field operators may also be expressed in terms of the vector potential operator $\hat{\mathbf{A}}(\mathbf{r},t)$

$$\hat{\mathbf{D}}(\mathbf{r},t) = -\varepsilon\partial_t\hat{\mathbf{A}}(\mathbf{r},t),\ \hat{\mathbf{B}}(\mathbf{r},t) = \nabla\times\hat{\mathbf{A}}(\mathbf{r},t)$$

$$\hat{\mathbf{A}}(\mathbf{r},t) = \sum_\lambda \int \mathrm{d}^3k \sqrt{\frac{\hbar}{2\varepsilon\omega_{\mathbf{k}}(2\pi)^3}}\left(\hat{a}_\lambda(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}\mathrm{e}^{-\mathrm{i}\omega_{\mathbf{k}}t+\mathrm{i}\mathbf{k}\cdot\mathbf{r}} + \hat{a}_\lambda^\dagger(\mathbf{k})\mathbf{e}_{\mathbf{k}\lambda}^*\mathrm{e}^{\mathrm{i}\omega_{\mathbf{k}}t-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\right)$$

The vector potential is needed for the description of the interaction with charged particles.

Space of Quantum States

The standard basis in the space of quantum states consists of the state vectors of stationary states – the eigenvectors of the Hamiltonian.

Vacuum State

The lowest energy state of the EMF – the ground state of the system – is often called the vacuum state. The state vector describing the vacuum state is denoted by $|0\rangle$. The vacuum state is the common ground state of all harmonic oscillators making up the EMF. Since the ground state of a single harmonic oscillator has the energy $\hbar\omega/2$, the energy of the ground state for an infinite number of oscillators is infinite (last term in eqn [7]). This infinity has no physical consequences because only the energy differences are observable. In the presence of obstacles (mirrors, resonators, beamsplitters, etc.), owing to the change in the mode structure, there will be a change in the energy of the new ground state relative to the energy in empty space. The energy differences between various ground states do have physical significance.

Fock Basis

It follows from the commutation relations [6] and from the form [7] of the Hamiltonian that the operator $\hat{a}_k$ decreases the energy of the EMF by the Planck quantum of energy $\hbar\omega_k$. Therefore, all operators $\hat{a}_k$ must annihilate the ground state vector, $\hat{a}_k|0\rangle = 0$, because there is no state with a lower energy. Each creation operator $\hat{a}_k^\dagger$ acting on any state vector adds one excitation with the energy $\hbar\omega_k$. These excitations of the electromagnetic field with characteristic energies $\hbar\omega_k$ are interpreted as photons. The state vectors of the form $\hat{a}_k^\dagger|0\rangle$ describe one-photon states, the state vectors of the form $\hat{a}_k^\dagger\hat{a}_l^\dagger|0\rangle$ describe two-photon states, etc. All the vectors generated from the vacuum state vector by acting with the creation operators form a basis, called the Fock basis. The general state vector of this basis is characterized by a sequence of natural numbers $\{n_1, n_2, \ldots\}$ and is generated from the vacuum state according to the formula

$$|n_1, n_2, \ldots\rangle = (n_1!n_2!\cdots)^{-1/2}(\hat{a}_1^\dagger)^{n_1}(\hat{a}_2^\dagger)^{n_2}\ldots|0\rangle \qquad [9]$$

where the prefactor is needed for normalization. The number n_k tells us how many photons of type k are present in this state. The spatial structure of a field excitation created by $\hat{a}_k^\dagger$ is described by the mode function $\mathbf{u}_k(\mathbf{r})$. Thus, this mode function plays the same role as the wavefunction $\psi(\mathbf{r})$ of a massive particle and the set of indices labeling the mode functions is the analog of quantum numbers for $\psi(\mathbf{r})$. In particular, for the mode function [2], the photons have a definite momentum $\hbar\mathbf{k}$. In a cylindrical waveguide the photons may have a definite momentum only along the axis of the cylinder. In a spherical cavity the photons may have the definite value of square of the total angular momentum and its projection on a given axis. The Fock states are highly nonclassical; all average values of the EMF in these states vanish.

All states of the EMF are either pure or mixed. Pure states are described by all linear superpositions of the basis states. In particular, by superposing photon states with a range of energies, one may form nonmonochromatic photon wavepackets. All superpositions are allowed since photons do not have any absolutely conserved quantum numbers (like, e.g., charge). Mixed states are described by the density operators. Among the mixed states the prominent role is played by the thermal state describing the EMF in equilibrium with a reservoir at temperature T.

Coherent States

The one-mode coherent states $|\alpha\rangle$ are obtained from the vacuum state by the action of the unitary displacement operator $\hat{D}(\alpha, \alpha^*)$

$$|\alpha\rangle = D(\alpha, \alpha^*)|0\rangle = \mathrm{e}^{\alpha\hat{a}^\dagger - \alpha^*\hat{a}}|0\rangle = \mathrm{e}^{-|\alpha|^2/2}\mathrm{e}^{\alpha\hat{a}^\dagger}|0\rangle$$

where $\hat{a}$ and $\hat{a}^\dagger$ are the annihilation and creation operators corresponding to a given mode. Coherent states are superpositions of an infinite number of Fock states. The multimode coherent state is generated by the product of displacement operators

$$|\alpha_1, \alpha_2, \ldots\rangle = \mathrm{e}^{\sum_k (\alpha_k\hat{a}_k^\dagger - \alpha_k^*\hat{a}_k)}|0\rangle$$

In reality, the coherent states are generated by the action of an external (strong) time-dependent electric current. In that case the coefficients α_k are the projections of the current on the mode functions. The average values of the field operators $\langle\hat{B}_i(\mathbf{r}, t)\rangle$ and $\langle\hat{D}_j(\mathbf{r}, t)\rangle$ in coherent states are the solutions of the Maxwell equations in the presence of an external electric current.

Entangled States

The states of the EMF describing photons in different modes may exhibit correlations. The simplest such state is a two-photon state of the form $|\phi\rangle = 2^{-1/2}(\hat{a}_1^\dagger\hat{a}_2^\dagger + \hat{a}_3^\dagger\hat{a}_4^\dagger)|0\rangle$. In this case, when one photon is found, say, in the state 1, the other will be found with the probability 1 in the state 2. The states that exhibit correlations between photons belonging to different modes are called the entangled states.

Correlations occur also in the classical theory, but only for statistical mixtures when there is a loss of information. In quantum theory the correlations may arise even without a loss of information, i.e., for pure states. A striking feature of entangled states is that the decomposition of such a state into a sum of Fock states is not unique. For example, the state vector $|\phi\rangle$ may also be written in the form $|\phi\rangle = 2^{-1/2}(\hat{a}^\dagger_{1'}\hat{a}^\dagger_{2'} + \hat{a}^\dagger_{3'}\hat{a}^\dagger_{4'})|0\rangle$, where $\hat{a}^\dagger_{1'} = 2^{-1/2}(\hat{a}^\dagger_1 - \hat{a}^\dagger_3)$, $\hat{a}^\dagger_{2'} = 2^{-1/2}(\hat{a}^\dagger_2 - \hat{a}^\dagger_4)$, $\hat{a}^\dagger_{3'} = 2^{-1/2}(\hat{a}^\dagger_1 + \hat{a}^\dagger_3)$, $\hat{a}^\dagger_{4'} = 2^{-1/2}(\hat{a}^\dagger_2 + \hat{a}^\dagger_4)$. Entangled states play a crucial role in quantum information theory, especially in quantum cryptography.

Quantum Fluctuations and the Uncertainty Relations

It follows from the basic principles of quantum theory that there are no states of the system for which both canonical variables q and p have sharply defined values. The limits on the quantum fluctuations (variances) of these values are determined by the uncertainty relation. The most stringent form of this relation obtained from the commutation relations [4] (for a single mode) reads

$$(\Delta q)^2(\Delta p)^2 \geq \frac{\hbar^2}{4} + \frac{\langle \delta\hat{q}\delta\hat{p} + \delta\hat{p}\delta\hat{q}\rangle^2}{4} \qquad [10]$$

where $\delta\hat{q} = \hat{q} - \langle\hat{q}\rangle$, $\delta\hat{p} = \hat{p} - \langle\hat{p}\rangle$, $(\Delta q)^2 = \langle(\delta\hat{q})^2\rangle$, and $(\Delta p)^2 = \langle(\delta\hat{p})^2\rangle$. The states for which this inequality is saturated play a distinguished role in quantum theory. Since the uncertainty relation expresses the limitations on the quantum fluctuations of q and p, the best one can do in approaching the classical regime is to achieve equality in the uncertainty relation. That is why the quantum states that saturate [10] are viewed as the counterparts of the classical states. In the quantum theory of the EMF, the uncertainty relations express the complementarity between the magnetic and electric fields. The better one knows the magnetic field, the more uncertain is the electric field, and vice versa. However, the projections of the field operators $\hat{\mathbf{B}}$ and $\hat{\mathbf{D}}$ on the same direction commute (cf. eqn [5]). Therefore, there are no limitations on the magnetic and electric field values in a given direction and only the values in different directions are subjected to the quantum uncertainty relation.

In the vacuum state the variances of q and p for each mode are

$$(\Delta q)^2 = \frac{\hbar}{2\varepsilon\omega}, (\Delta p)^2 = \frac{\hbar\varepsilon\omega}{2} \qquad [11]$$

and they saturate the uncertainty relation. These fluctuations are called the shot noise. Very often one uses dimensionless quadrature operators $\hat{X}_1 = \hat{q}\sqrt{2\varepsilon\omega/\hbar}$ and $\hat{X}_2 = \hat{p}\sqrt{2/\varepsilon\omega\hbar}$, whose fluctuations in the vacuum state are equal to 1. All coherent states have the same fluctuations as the vacuum state. They may be viewed in the phase space as the vacuum state displaced from the origin (corresponding to the vacuum state) by the mean values of q and p.

Coherent states are not the only states that saturate the uncertainty relations. The other states with this property are called the squeezed states. The fluctuations of X_1 and X_2 for squeezed states are different from 1 but their product is still equal to 1, $(\Delta X_1)^2(\Delta X_2)^2 = 1$.

Fock states have well-defined energies and numbers of photons since they are the eigenstates of the Hamiltonian. However, the phases of the EMF are maximally uncertain; they are evenly spread over the whole range from 0 to 2π. That is why the average values of the field vectors vanish in all Fock states.

See also

Photon Picture of Light. **Quantum Electrodynamics:** Cavity QED.

Further Reading

Bialynicki-Birula I (1996) Photon wave function. In: Wolf E (ed.) *Progress in Optics*, vol. 36. Amsterdam, The Netherlands: Elsevier.

Bialynicki-Birula I and Bialynicka-Birula Z (1975) *Quantum Electrodynamics*. London: Pergamon.

Loudon R (1973) *The Quantum Theory of Light*. Oxford, UK: Clarendon Press.

Louisell W (1990) *Quantum Statistical Properties of Radiation*. New York: Wiley.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Power EA (1964) *Introductory Quantum Electrodynamics*. London: Longman.

Scully MO and Zubairy MS (1997) *Quantum Optics*. Cambridge, UK: Cambridge University Press.

Cavity QED

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics, Garching,
Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Modification of the Spontaneous Transition Rate in Confined Space

In order to understand the modification of the spontaneous emission rate in an external cavity-like structure, it must be remembered that in quantum electrodynamics this rate is proportional to the density of modes of the electromagnetic field, i.e., the vacuum field fluctuations above the atomic transition frequency ω_0. As a consequence the spontaneous emission rate is increased if the atom is surrounded by a cavity tuned to the transition frequency. Conversely, the decay rate decreases when the cavity is mistuned. This fact was already recognized in the pioneering days of nuclear magnetic resonance by Purcell.

The spontaneous decay rate of the atom in the cavity, γ_c, will then be enhanced in relation to that in free space, γ_f, by a factor given by the ratio of the corresponding mode densities

$$\frac{\gamma_c}{\gamma_f} = \frac{\rho_c(\omega_0)}{\rho_f(\omega_0)} = \frac{2\pi Q}{V_c \omega_0^3} = \frac{Q\lambda_0^3}{4\pi^2 V_c}$$

where V_c is the volume of the cavity and Q is the quality factor of the cavity which expresses the 'sharpness' of the mode. For low-order cavities $V_c \approx \lambda_0^3$ this means that the spontaneous emission rate is increased by roughly a factor of Q. As mentioned already, when the cavity is detuned, the decay rate will decrease. In this case the atom cannot emit a photon, since the cavity is not able to accept it, and the energy will stay with the atom, i.e., its decay is inhibited.

To change the decay rate of an atom, in principle no resonator has to be present; any conducting surface near the radiator affects the mode density and, therefore, the spontaneous radiation rate. Parallel conducting planes can alter the emission rate somewhat but can only reduce the rate by a factor of 2 owing to the existence of modes, independent of the plate separation with an electric field perpendicular to the planes.

In order to demonstrate experimentally the modification of the spontaneous decay rate, it is not necessary to go to single-atom densities. The experiments where the spontaneous emission is inhibited can also be performed with large atom numbers. However, in the opposite case, when the increase of the spontaneous rate is observed, a large number of excited atoms may disturb the experiment by induced transitions. The first experimental work on inhibited spontaneous emission was done by Drexhage, Kuhn and Schäfer in 1974. The fluorescence of a thin dye film near a mirror was investigated. A reduction of the fluorescence decay by up to 25% was observed. Later experiments with microcavities filled with dye solutions were performed by de Martini *et al.* These experiments demonstrated that thresholdless lasing is possible in such systems.

Inhibited spontaneous emission was observed by Gabrielse and Dehmelt. In these neat experiments with a single electron stored in a Penning trap they observed that cyclotron orbits show lifetimes which are up to 10 times longer than that calculated for free space. The electrodes of the trap form a cavity which decouples the cyclotron motion from the vacuum field leading to the longer lifetime.

A new stage in experiments on cavity QED was reached when it was recognized that highly excited alkaline atoms are very suitable for these experiments. The main quantum numbers are in the range $n = 20-60$. Those so-called Rydberg atoms couple very strongly to the radiation field as will be discussed later. The transitions to neighboring states are in the microwave region. The cavities can therefore be built as low-order cavities with dimensions on the order of the wavelength of the transitions being in the mm or cm regions. The spontaneous rate of these transitions in free space is small owing to the low value of their transition frequency, therefore an enhancement is possible. Experiments with Rydberg atoms on the inhibition of spontaneous emission have been performed by Kleppner and coworkers and by Haroche and coworkers. In the latter experiment a 3.4 μm transition of the Cs atom was suppressed.

The first observation of enhanced atomic spontaneous emission in a resonant cavity was published by Haroche *et al.* This experiment was performed with Rydberg atoms of Na excited in the 23s state in a niobium superconducting cavity resonant at 340 GHz. Cavity tuning-dependent shortening of the lifetime was observed. The cooling of the cavity had the advantage of totally suppressing the black-body field. The latter effect is completely absent if optical transitions are observed, however, in this case it is more difficult to obtain low-order cavities. The first experiments on optical transitions were performed by Feld and collaborators. They succeeded in demonstrating the enhancement of spontaneous transitions even in higher-order optical cavities.

In modern semiconductor devices both electronic and optical properties can be tailored with a high degree of precision. Therefore, electron–hole systems producing recombination radiation analogous to radiating atoms can be localized in cavity-like structures, e.g., in quantum wells. Thus optical microcavities of half or full wavelength size are obtained. Both suppression and enhancement of spontaneous emission in semiconductor microcavities were demonstrated in experiments by Yamamoto and collaborators. Yablonovitch *et al.* proposed the use of photonic band gap structures in semiconductors. In those systems spontaneous emission is forbidden in certain frequency regions owing to the special material properties.

Energy Shift in Confined Space

In the previous section we focused on changes of radiation rates of atoms near conducting walls or in cavity-like structures. Next we will discuss the more subtle phenomenon of energy shifts. While radiation rates are modified by the influence of the vacuum field on the real emission of photons the energy shifts are caused by the modification of a virtual photon interaction.

Resonant and nonresonant phenomena have to be distinguished. The resonant self-energy shift of a decaying atomic dipole in the vicinity of a conducting wall can be determined from the average polarization energy produced by the image dipole field. For distances comparable to the wavelength the near-field condition is satisfied, resulting in the z^{-3} dependence of the static dipole–dipole interaction characteristic for the van der Waals energy; under far-field conditions the distance dependence is given by z^{-1}. The polarization of the atom by the nonresonant parts of the broadband electromagnetic field causes energy shifts, of which the Lamb shift is the most prominent. In the sense of the nonrelativistic treatment by Bethe the major contribution of that shift can be described as being a result of the emission and reabsorption of virtual photons. It is plausible that just as the real emission of a photon is modified in confined space, so also is the virtual process. The latter ‘real’ radiation energy shift is thus a consequence of vacuum fluctuations only. It is identical with the energy shift predicted by Casimir and Polder and is analogous to the better-known result of Casimir on the force between two plane neutral conducting plates.

The question of modification of atomic energies in confined space has recently found considerable interest and many calculations of the phenomenon have been performed. Direct application to the energy shift of Rydberg atoms, which are of special interest for experimental studies, was performed by Barton in 1987. He calculated the direct electrostatic interaction with a conducting wall and the radiation induced (retarded) effects. The result is that in the case of two parallel plates the electrostatic effect is dominant when the distance L between the conducting plates is small, $L < n^3 a_0/\alpha$ (n is the principal quantum number of the Rydberg atom, a_0 the Bohr radius, and α the fine-structure constant), and the radiative effect plays the major role when large distances are used, $L > n^3 a_0 \alpha$.

Experiments to measure the van der Waals force of conducting planes have been performed by Hinds and coworkers. They could clearly demonstrate the z^{-3} dependence of the van der Waals force. The retarded effects were not detectable at large distances but at small distances from the wall a deviation from the z^{-3} dependence was found which was attributed to the retarded QED potential.

The energy shift of rubidium Rydberg atoms in confined space has been measured by Marrocco *et al.* in an experiment with extreme high-resolution using the Ramsey experiment two-field method. The atoms are excited in s-Rydberg states ($n \approx 30$) by two-photon transitions using the light of an ultrastable dye laser, the linewidth of which was less than 10 Hz. The laser intensity was enhanced in a folded cavity locked to the laser frequency. Both interaction zones necessary for the Ramsey method were enclosed in this cavity. Between the two interaction regions the atoms pass through a pair of conducting plates, the distance between which can be changed. The shift of the Ramsey interference was measured as a function of the plate distance. Using the Ramsey method has the advantage that the shift can be determined without a direct probing of the atoms in the space between the plates. A level shift on the order of about 150 Hz could be measured, in very good agreement with theory.

Maser Operation

If the rate of atoms crossing a cavity exceeds the cavity damping rate ω/Q, the photon released by each atom is stored long enough to interact with the next atom. Here ω stands for the cavity frequency and Q for the quality factor of the cavity. The atom–field coupling becomes stronger and stronger as the field builds up and evolves to a steady state. Using Rydberg atoms with a large field–atom coupling constant leads to a new kind of maser which operates with exceedingly small numbers of atoms and photons.

The photons corresponding to transitions between neighboring Rydberg atoms are in the microwave region at about 20–100 GHz. Atomic fluxes as small as a few atoms per second have generated maser action, as could be demonstrated by Walther *et al.* of the University of Munich in 1985 using a superconducting cavity. For such a low flux there is never more than a single atom in the resonator – in fact, most of the time the cavity is empty of atoms. It should also be mentioned that in the case of such a setup a single resonant photon is sufficient to saturate the maser transition.

The scheme for this one-atom maser or micromaser is shown in Figure 1. The setup represents the simplest system in radiation–atom interaction: a single atom is interacting with a single mode of the radiation field. This device was used to verify the complex dynamics of a single atom in a quantized field predicted by the Jaynes–Cummings model. All of the features are explicitly a consequence of the quantum nature of the electromagnetic field: the statistical and discrete nature of the photon field leads to new characteristic dynamics such as collapse and revivals in the exchange of a photon between the atom and the cavity mode. The frequency of this exchange is usually called the Rabi flopping frequency. The field in the cavity is measured through the number of atoms in the lower state of the maser transition. Due to the strong coupling between the atom and maser field both are entangled. The strong coupling can also be used to entangle subsequent atoms.

The steady-state field of the micromaser shows sub-Poisson statistics. This is in contrast to regular masers and lasers where coherent fields (Poisson statistics) are observed. The reason for nonclassical radiation being produced is that a fixed interaction time of the atoms is chosen, leading to careful control of the atom–field interaction dynamics.

Under steady-state conditions, the photon statistics $P(n)$ of the field of the micromaser are essentially determined by the pump parameter $\Theta = N_{\mathrm{ex}}^{1/2}\Omega t_{\mathrm{int}}$, denoting the angular rotation of the pseudospin vector of the interacting atom. Here, N_{ex} is the average number of atoms that enter the cavity during the decay time of the cavity τ_{cav}, Ω the vacuum Rabi flopping frequency, and t_{int} is the atom–cavity interaction time. The normalized photon number of the maser $\langle\nu\rangle = \langle n\rangle/N_{\mathrm{ex}}$ shows the following generic behavior

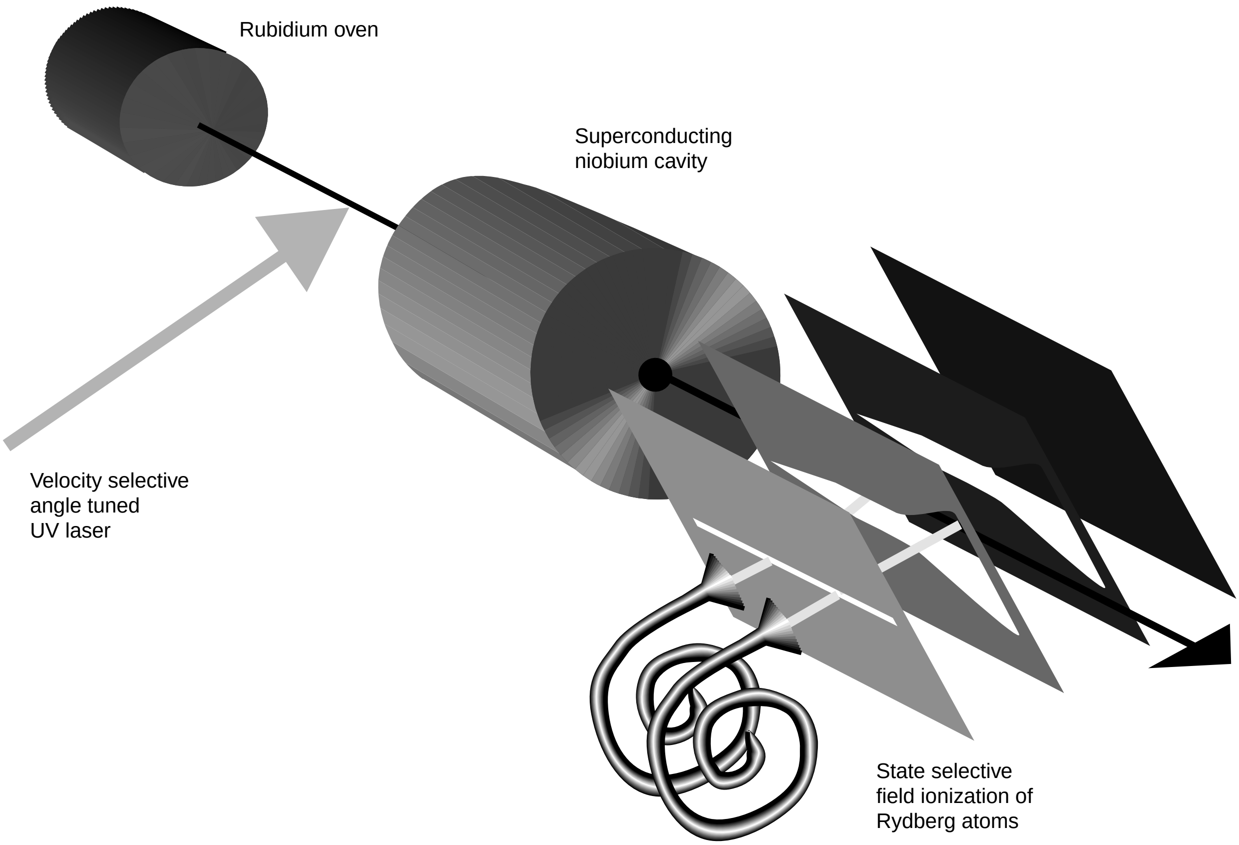
Rubidium oven
Superconducting niobium cavity
Velocity selective angle tuned UV laser
State selective field ionization of Rydberg atoms

Figure 1 Micromaser setup of rubidium Rydberg atoms in the 63p state. The velocity of the atoms is controlled by exciting a velocity subgroup of atoms in the atomic beam. The atoms in the upper and lower maser levels are selectively detected by field ionization. The number of photons deposited in the cavity is determined through the number of atoms in the lower state (61d).

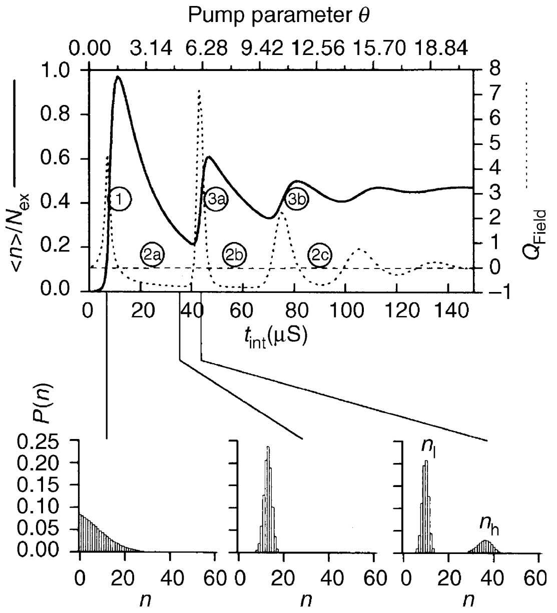

Figure 2 One-atom maser or micromaser characteristics. The upper part of the figure shows the average photon number versus interaction time (solid curve) and the photon number fluctuations represented by the Q-factor (dotted line). Both curves are determined by the photon exchange dynamics between atom and field. The pump parameter gives the angular rotation of the pseudospin vector of the interacting atom (e.g., 2π corresponds to a full rotation, i.e., the atom is again in the upper level). In the lower part of the figure the steady-state photon number distribution $P(n)$ is shown for three values of t_{int}. The distribution on the left side corresponds to the maser threshold, that on the right gives an example of the double-peaked distribution associated with the quantum jump behavior. In this situation, the atom is back in the excited state and can again emit, leading to a higher steady-state photon number n_h. With increasing t_{int} this part will grow and n_l will decrease and disappear. A new jump occurs in the region 3b.

(see Figure 2). It suddenly increases at the maser threshold value $\Theta = 1$ and reaches a maximum for $\Theta \approx 2$ (denoted by 1 in Figure 2). The behavior at threshold corresponds to the characteristics of a continuous phase transition. As Θ increases further, the normalized averaged photon number $\langle \nu \rangle / N_{\text{ex}}$ decreases to a minimum slightly below $\Theta \approx 2\pi$, and then abruptly increases to a second maximum (3a in Figure 2). This general type of behavior recurs roughly at integer multiples of 2π, but becomes less pronounced with increasing Θ. The reason for the periodic maxima of the average photon number is that for integer multiples of $\Theta = 2\pi$ the pump atoms perform an integer number of full Rabi flopping cycles, and start to flip over at a slightly larger value of Θ, thus leading to enhanced photon emission. The periodic maxima for $\Theta = 2\pi, 4\pi$, and so on may be interpreted as first-order phase transitions.

The photon statistics of the maser radiation is usually characterized by the Q-parameter introduced by Mandel:

$$Q_{\text{field}} = \left\{ \left(\langle n^2 \rangle - \langle n \rangle^2 \right) / \langle n \rangle \right\} - 1$$

It can be seen that $Q_{\text{field}} = 0$ corresponds to a Poissonian photon distribution. Q_{field} for the micromaser is plotted as a dotted line in Figure 2. The value drops below zero in the region 2a, 2b, etc. This shows the highly sub-Poissonian character of the one-atom-maser field being present over a large range of parameters.

The reason for the sub-Poissonian atomic statistics is the following. A changing flux of atoms changes the Rabi frequency via the stored photon number in the cavity. Adjusting the interaction time allows the phase of the Rabi nutation cycle to be chosen such that the probability of the atoms leaving the cavity in the upper maser level increases when the flux, and hence the photon number in the cavity, is raised. This leads to sub-Poissonian atomic statistics since the number of atoms in the lower state decreases with increasing flux and photon number in the cavity. This feedback mechanism can be neatly demonstrated when the anticorrelation of atoms leaving the cavity in the lower state is investigated. Measurements of this anticorrelation phenomenon could be made.

The fact that anticorrelation is observed shows that the atoms in the lower state are more equally spaced than expected for a Poissonian distribution of the atoms in the beam. This means, for example, that when two atoms enter the cavity close to each other, the second one performs a transition to the lower state with reduced probability.

The interaction with the cavity field thus leads to an atomic beam with atoms in the lower maser level showing number fluctuations which are up to 40% below those of a Poissonian distribution found in usual atomic beams. This is interesting because atoms in the lower level have emitted a photon to compensate for cavity losses inevitably present. Although this process is induced by dissipation giving rise to fluctuations, the atoms still obey sub-Poissonian statistics.

Generation of Number States (Fock States) of the Radiation Field

When the micromaser is operated at low temperatures, a very interesting feature is observed in the inversion of the population of the two maser levels called trapping states. They are a steady-state feature of the maser field; they occur in the micromaser as a direct consequence of field quantization in a cavity. At low cavity temperatures the number of black-body photons, n_{th}, in the cavity mode is reduced and trapping states begin to appear; at higher temperatures they are washed out by the thermal photons. The trapping states show up when the atom–field coupling Ω and the interaction time t_{int} are chosen such that in a cavity field with n_{q} photons the atoms undergo an integer number k of Rabi cycles. This is summarized by the condition

$$\Omega t_{\mathrm{int}}\sqrt{n_{\mathrm{q}}+1} = k\pi \qquad [1]$$

for the trapping state, denoted by (n_{q}, k). When eqn [1] is satisfied, the cavity photon number is left unchanged after interaction of an atom and hence the photon number is ‘trapped’. This will occur regardless of the atomic pump rate N_{ex}. The trapping state is therefore characterized by the upper-bound photon number n_{q} and the number of integer multiples of full Rabi cycles k. In this situation the field is stabilized. Whenever a photon disappears, e.g., due to dissipation, the next incoming atom experiences a changed Rabi nutation frequency and emits a photon with high probability. At the trapping condition a quantum nondemolition situation is present. Through the dynamics of the Rabi nutation the field is measured, without any net change of the field.

The build-up of the cavity field can be seen in Figure 3, where the emerging atom inversion is plotted against the interaction time and pump rate. At low atomic pump rates (low N_{ex}) the maser field cannot build up and the maser exhibits Rabi oscillations due to the interaction with the vacuum field. At the positions of the trapping states, the field increases until it reaches the trapping-state condition. This is manifested as a reduced emission probability and hence as a dip in the atomic inversion. Once in a trapping state, the maser will remain there regardless of the pump rate.

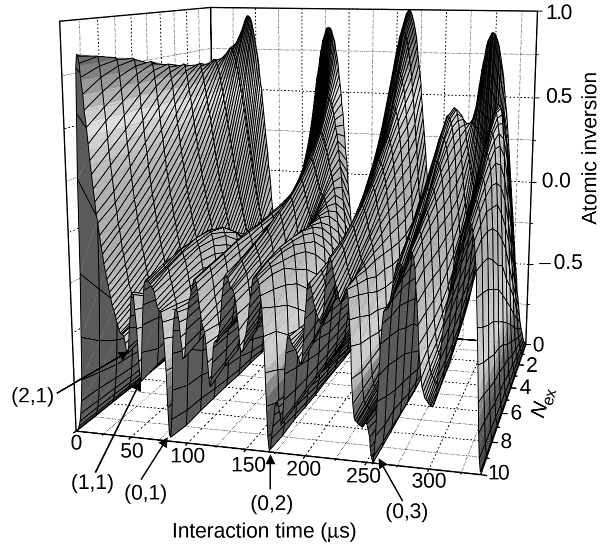

Figure 3 Behavior of the micromaser at low temperature showing the trapping states, manifested as valleys along the N_{ex} axis. For the designation of the trapping states see text.

Under ideal conditions the micromaser field in a trapping state is a Fock state, but when the micromaser is operated in a continuous wave mode, the field state is very fragile and highly sensitive to external influences and experimental parameters. However, in contrast to continuous-wave operation, in pulsed operation trapping states are more stable and more practical and can be used over a much broader parameter range than in continuous-wave operation.

To demonstrate the principle, Figure 4 shows a simulation of a sequence of 20 pulses of the pumping atoms in which an average of seven excited atoms per pulse are present. Under the trapping condition only a single emission event occurs, producing a single lower state atom which leaves a single photon in the cavity. Since the atom–cavity system is in the trapping condition, the emission probability is reduced to zero and the photon number is stabilized. Consequently, excited state atoms following the emitting atom stay in the upper maser level. The variation of the time when an emission event occurs during the atom pulses in Figure 4 is due to the Poissonian spacing of upper-state atoms entering the cavity and the stochasticity of the quantum process.

The lower part of Figure 4 shows the photon number distribution resulting from this process. It was demonstrated experimentally by Brattke *et al.* that a single photon number state could be generated with a success rate of 85%. By improving the experimental parameters one can expect to prepare single-photon Fock states in 98% of the pulses by this method.

Other Cavity Experiments

Besides the micromaser, there are also other experiments using Rydberg atoms in cavities and the strong interaction of these atoms with cavities which lead to interesting applications. One example is an experiment by S. Haroche, J. M. Raimond *et al.*, who succeeded in realizing a Schrödinger cat state in a cavity. Studies on the decoherence of this state could be conducted. The experiments are quite important in exploring the boundaries between the quantum and classical worlds. Another example is the quantum nondemolition detection of a single photon in a cavity by S. Haroche employing the dispersion level shift of probing atoms.

There is an interesting equivalence between an atom interacting with a single-mode field and a quantum

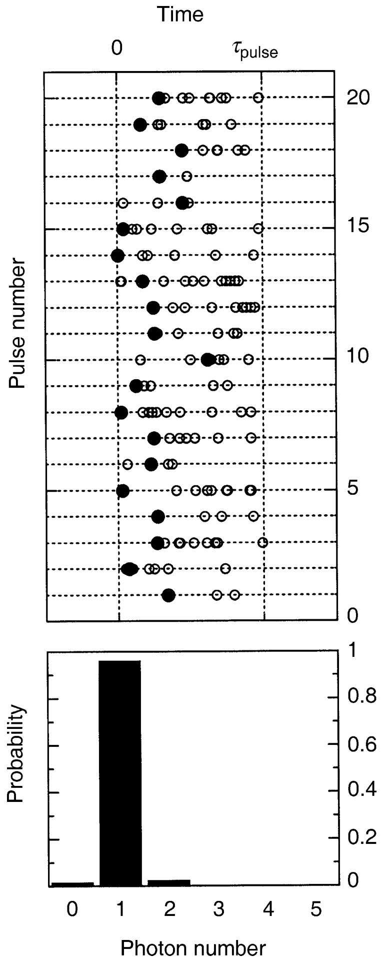

Figure 4 A simulation of a subset of 20 subsequent atom bunches generated by pulsed laser excitation with the associated probability distribution for photons in the cavity or lower-state atom production (filled circles represent lower-state atoms and open circles represent excited-state atoms). The start and end of each pulse is indicated by the vertical dotted lines marked 0 and τ_{pulse}, respectively. For details, see Battke S, Varcoe BTH and Walther H (2001) Generation of photon number states on demand via cavity, quantum electrodynamics. *Physical Review Letters* 86: 3534–3537.

particle in a harmonic potential, as was first pointed out by D. F. Walls and H. Risken; this connection results from the fact that the radiation field is quantized on the basis of the harmonic oscillator. The Jaynes–Cummings dynamics can therefore also be observed with trapped ions, as recently demonstrated in a series of beautiful experiments by D. Wineland *et al.* They also produced a Schrödinger cat state by preparing a single trapped ion in a superposition of two spatially separated wavepackets which are formed by coupling different vibrational quantum states in the excitation process.

Besides the experiments in the microwave region, a single-atom laser emitting in the visible range has also been realized by M. Feld *et al.* Furthermore, cavity quantum electrodynamic effects have been studied in the optical spectral region by J. L. Kimble *et al.*

Today's technology in microfabrication of semiconductor diode structures allows the realization of low-order cavity structures for diode lasers. In these systems the spontaneous emission is controlled in the same way as in the micromaser. Since spontaneous decay is a source of strong losses, control of this phenomenon leads to highly efficient laser systems. Quantum control of spontaneous decay thus has important consequences for technical applications. This topic will be briefly described in the next section.

Microlasers

The simplest approach to fabricating an optical microcavity is to shrink the spacing between the mirrors of a Fabry–Perot resonator to λ/n (where n stands for the refractive index) while reducing the lateral dimensions to a range of the same order of magnitude. This structure provides a single dominant longitudinal field mode that radiates into a narrow range of angles around the cavity axis.

The first optical microcavity experiments used dye molecules between high-reflectivity dielectric mirrors in the Fabry–Perot configuration. Because spontaneous emission is a major source of energy loss, speed limitations, and noise in lasers, the capability to control spontaneous emission is expected to improve laser performance. If the fraction of spontaneous emission coupled into the lasing mode is made close to one, the 'thresholdless' laser is obtained, in which the light output increases almost linearly with the pump power instead of exhibiting a sharp turn-on at the pump threshold.

Semiconductor microcavities provide high-Q Fabry–Perot cavities for both basic studies and potential applications. Molecular beam epitaxy or organometallic chemical vapor deposition techniques are used to deposit high-reflectivity mirrors consisting of alternating quarter-wavelength layers of lattice-matched semiconductors. For example, 15–20 pairs of quarter-wave layers of $Al_{0.2}Ga_{0.8}As$ and AlAs result in a reflectivity greater than 99% and Q values greater than 500. The optically active layer in such a microcavity is typically a GaAs quantum well located at the midplane of the cavity, where the field strength is maximum.

Since the microcavities have an extremely low threshold, their efficiency will also be high. Low-cost, high-density light source arrays and photonic circuits are made possible by the small size and low power consumption of such resonators. It will be possible to produce entire wafers containing millions of microlasers with a multipole arrangement instead of having to cleave each individual semiconductor laser as at present. This improvement will lead to higher yields and lower cost per element. Another advantage of surface-emitting microcavity sources is the efficient optical coupling of their stable symmetric mode patterns into optical fibers or waveguides. These lasers are thus certain to spark a revolution in optical communication in the future.

Conclusion

This article reviewed experiments in cavity quantum electrodynamics. The experiments shed new light on our understanding of the radiation interaction of atoms. The achievable strong coupling between atoms and radiation leads to the possibility to generate entanglement between the generated radiation field and the atoms, presenting a basis for interesting applications in connection with, e.g., quantum computing and quantum information processing. Furthermore Fock states of the radiation field can be generated. The control of spontaneous decay finally leads to interesting new laser systems offering high efficiency.

See also

Quantum Electrodynamics: Cavity QED in Semiconductors; Quantum Theory of the Electromagnetic Field. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman PR (1994) Cavity quantum electrodynamics. In: Bederson B and Bates D (eds). *Advances in Atomic, Molecular, and Optical Physics*, Supplement 2. Boston, MA: Academic Press.

Englert B-G, Löffler M, Benson O, *et al.* (1998) Entangled atoms in micromaser physics. *Fortschritte der Physik* 46: 897–926.

Haroche S (1992) Cavity quantum electrodynamics. In: Dalibard J, Raimond J-M and Zinn-Justin J (eds) *Fundamental Systems in Quantum Optics*. pp. 767–935. Amsterdam: North-Holland.

Haroche S and Kleppner D (1989) Cavity quantum electrodynamics. *Physics Today* 42: 24–30.

Kimble HJ, Carnal O, Georgiades N, *et al.* (1995) Quantum optics with strong coupling. In: Wineland DJ, Wieman CE and Smith SJ (eds) *Atomic Physics*, vol. 14, pp. 314–335. New York: American Institute of Physics.

Marrocco M, Weidinger M, Sang RT and Walther H (1998) Quantum electrodynamic shifts of Rydberg energy levels between parallel metal plates. *Physical Review Letters* 81: 5784–5787.

Meschede D (1992) Radiating atoms in confined space: from spontaneous emission to micromasers. *Physics Reports* 211: 201–250.

Meystre P (1992) Cavity quantum optics and the quantum measurement process. In: Wolf E (ed.) *Progress in Optics*, vol. 30. Amsterdam: North-Holland.

Walther H (1992) Experiments on cavity quantum electrodynamics. *Physics Reports* 219: 263–281.

Cavity QED in Semiconductors

M Kira, W Hoyer and S W Koch,
Philipps-University, Marburg, Germany

G Khitrova and H M Gibbs,
University of Arizona, Tucson, AZ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

According to classical electromagnetism, propagating light fields obey Maxwell's equations under which electric and magnetic fields exchange their strength in an oscillatory manner. If a propagating field is followed at a single reference point, it is convenient to represent the electric field as a complex number $|E(t)|\exp(-i\omega t)$ where $|E(t)|$ is the magnitude of the field and the phase of the field $\exp(-i\omega t)$ oscillates in time t with the optical frequency ω. Figure 1 illustrates how the classical field is determined by one single point in the complex plane; consequently, the phase and amplitude of the field are known exactly. According to quantum mechanics, this simple picture has to be altered because the Heisenberg uncertainty relation states that complementary quantities like phase and amplitude cannot be determined precisely at the same time. As a result, the real and imaginary parts of the field can be determined only with accuracy $\Delta\text{Re}[E]$ and $\Delta\text{Im}[E]$, respectively, and the Heisenberg uncertainty principle determines the best possible accuracy $\Delta\text{Re}[E]\,\Delta\text{Im}[E] \geq 1$. To incorporate the fundamental inaccuracy, the electric field has to be defined by using complex-valued distributions or equivalently wavefunctions as shown in Figure 1. This quantization procedure leads to the Schrödinger equation for light analogous to that for particles. The field called quantum optics investigates the quantum electrodynamics (QED) features of light.

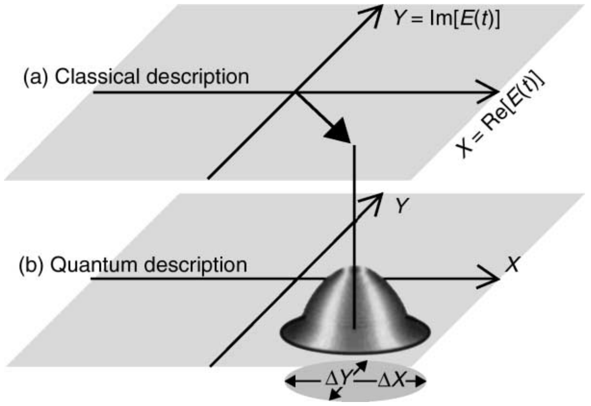

Figure 1 (a) The classically described light field $E(t)$ is a single point in the complex XY-plane; the corresponding (X, Y) vector has a definite phase and angle. (b) Quantum mechanically described light is a distribution with fluctuations ΔX and ΔY described by the shaded circle.

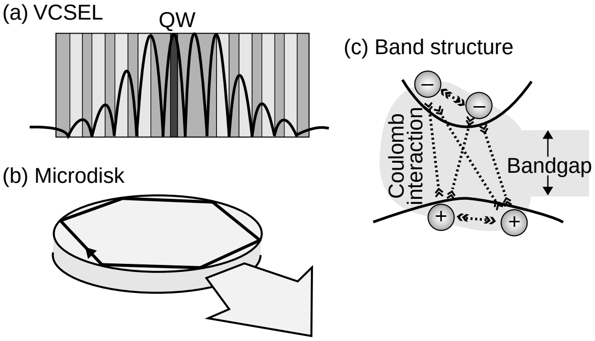

Figure 2 (a) A typical vertical-cavity surface-emitting laser (VCSEL) structure; a stack of quarter-wavelength layers with different refractive index provides a mirror such that the light intensity (solid line) is strongly confined inside the cavity. A quantum well (QW) is positioned at the field maximum. (b) In a microdisk laser, light encounters ideal total reflection when the edge of the disk is reached. The so-called whispering gallery mode emits (leaks) light from the edge. (c) A schematic illustration of semiconductor band structure, bandgap, and Coulomb interaction for an electro–hole system.

In general, light does not propagate freely in space, but it interacts strongly with the surrounding matter. For example, light can be absorbed, and its energy can be converted into excitation of the matter. As a result, the light may be slowed down, reflected, scattered, or diffracted. In order to explain the implications of the light–matter interaction in detail, one obviously has to apply a quantum mechanical description also for the matter. This approach determines the so-called eigenstates of the matter, so that the matter may occupy only certain discrete states with discrete energies. The Rydberg series of a hydrogen atom is a typical example. Similar discrete energy levels are also found for the quantized light; these levels are often referred to as photons which heuristically describe the particle properties of the light. The lowest-order light–matter interaction consists of processes where one photon is absorbed (emitted) while the matter is simultaneously excited (de-excited) from one eigenstate to another. This generic type of interaction leads to a microscopic description of the optical effects mentioned above. The magnitude of these effects depends on the strength of the interaction. By implementing different cavity configurations, the reflected light can be forced to travel across the same matter many times. As a result, the light–matter interaction is enhanced by a factor proportional to the multiple passes of the light. Thus, a cavity can be efficiently used to amplify optical phenomena. Figure 2 shows cavity configurations commonly used for semiconductors. Semiconductor cavity QED investigates quantum optical effects in semiconductor systems by enhancing them with cavity mirrors. In general, quantum optical features produce classically unexpected effects which typically stem from: (a) the discrete nature of eigenstates of light and matter; (b) the superposition principle of quantum mechanics; and (c) the Heisenberg uncertainty principle. In this article, we review both theoretical and experimental advances made so far to predict, observe, and control quantum optical effects in semiconductors with respect to effects (a)–(c).

During the last few decades, atomic quantum optics has already developed toward QED investigations, and semiconductor optics is developing rapidly into the same regime. Advanced atomic QED theories have successfully explained and guided intriguing experimental developments like laser cooling, atom condensation, and photon teleportation. However, the atomic approach can mostly be applied to describe dilute and only weakly interacting atomic gases since relatively simple models of few-level systems are used to describe the material. The elementary electronic excitations in semiconductors consist of electrons and holes (missing electrons) lifted into the conduction and valence bands, respectively. The corresponding eigenstates form continuous energy bands with the band-gap energy separation indicated in Figure 2. Since electrons and holes have opposite charges, they experience Coulomb attraction whereas bare electrons or holes repel each other. Under favorable conditions, the attractive Coulomb interaction prevails and atom-like bound electron–hole pairs (excitons) may be formed. However, since the Coulomb force has an infinite range and an electron–hole system is typically dense, excitons cannot be treated as weakly interacting quasiparticles. As a result, the atomic QED approaches cannot be used to describe quantum optics of semiconductors in general. Thus, this article concentrates mostly on semiconductor QED theory which fully includes the interaction of charge carriers, i.e., electrons and holes, as well as quantum features of light.

The development work on quantum devices and processes is vital for advancements in several key technologies such as computers and telecommunications. The continuous decrease in component dimensions leads to a microscopic structure size (less than 1 μm); at the same time, increase in the device performance is accompanied by ultrafast operation time (faster than 1 ps). Due to these development trends, the properties of components and processes are becoming more and more quantum mechanical. Although the description of the quantum processes is complicated, the microscopic behavior offers entirely new operational functionality such as massively parallel quantum computers. These possibilities are based on the controlled manipulation of the quantum mechanical state of the light and matter. Due to rapid recent developments, optically coupled semiconductor devices have a great potential to become technologically successful and commercially viable quantum components.

Quantized States: Observation via Strong Light–Matter Interaction

In order to enhance the light–matter interaction, we choose a system where the semiconductor is placed inside a cavity in the position where the field intensity is maximum. For an empty cavity, the field intensity has a strong dependence upon frequency with a strong resonance at ω_R when the cavity length is an integer number of light wavelengths in the material. The width of the resonance is directly related to the quality of the cavity as determined by the number of back-and-forth reflections of light inside the cavity. If one chooses a vertical-cavity surface-emitting laser (VCSEL) structure, the most efficient coupling is obtained by placing a thin planar semiconductor structure at the field maximum as shown in Figure 2. If the semiconductor is planar and thin, the structure is called a quantum well because electrons and holes are confined in one direction in analogy to the standard particle-in-a-box problem of fundamental quantum mechanics. As a result, the energy levels of a quantum well are discrete in the confinement direction. If the structure is narrow enough, the system dynamics is confined to the lowest quantum-well level such that the carriers are quasi-two-dimensional due to the free in-plane motion. The optical response of such a system to a weak classical probe beam has been successfully analyzed with the so-called semiconductor Bloch equations. For the quantum-well system alone, the absorption spectrum may have a sharp resonance; this is often referred to as an excitonic resonance since it is located below the fundamental bandgap energy at a position corresponding to the exciton binding energy. Obviously, the light–matter coupling becomes large when the exciton and cavity resonances coincide. However, since these resonances are coupled, the optical response is altered; we observe a splitting into two absorption peaks instead of the original degenerate resonances. This phenomenon is commonly referred to as normal-mode coupling; in general, it is a quite common feature in quantum mechanics that degenerate states split due to an additional interaction. Figure 3 shows the first experimental observation of normal-mode coupling in semiconductors.

Since a semiconductor is a strongly interacting many-body system, several effects alter the specific character of an excitonic resonance when electrons and holes are excited. The attractive interaction between electrons and holes becomes weaker for

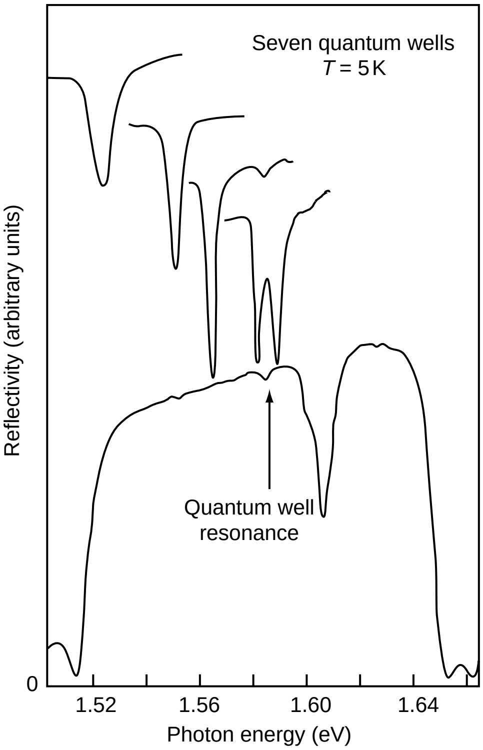

Figure 3 Probe reflectivity of a seven-quantum-well microcavity structure. The various curves correspond to the energy difference between exciton and cavity resonances. When these resonances become degenerate, the reflectivity shows two resonances corresponding to the normal-mode splitting. From Weisbuch C, Nishioka M, Ishikawa A and Arakawa Y (1992) Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Physical Review Letters* 69: 3314. Copyright (2004) by the American Physical Society.

increased carrier density since the surrounding charges screen the bare Coulomb interaction. A particular exciton may also experience scattering from nearby electrons, holes, or other excitons via the Coulomb interaction. Furthermore, electrons and holes are Fermions, which requires that any given carrier state can only be occupied once. This poses fundamental limits on how many carriers can coexist in a specific volume; above a certain limit, additional occupation is prevented, and so-called Pauli blocking is observed. Since an exciton consists fundamentally of Fermionic constituents, eventually Pauli blocking effects become important even for excitons. Due to a combination of these many-body effects, the excitonic resonance is weakened by an increasing carrier density. Figure 4 shows a comparison between theory and experiment of the excitonic absorption spectrum for different carrier densities. For elevated densities, the exciton resonance is broadened and its height reduced. When the same investigations are repeated in a microcavity, we observe that the normal-mode peaks decrease in height but their separation is roughly unchanged for moderate densities. In general, the normal-mode spitting is proportional to the oscillator strength of the excitonic absorption, i.e., the area under the resonance. Thus, the observed constant splitting suggests an unchanged oscillator strength for moderate densities. The invariant oscillator strength was unexpected and can only be explained by using a theory which includes the Coulomb interaction of carriers microscopically. When the carrier density is increased even further, the exciton resonance is completely bleached, and the microcavity transmission has only a single peak, at the cavity mode energy; this is commonly referred to as the weak-coupling regime, in contrast to the nonperturbative strong coupling with two transmission peaks. The strong coupling regime provides several intriguing phenomena; for example, parametric amplification of emission has been demonstrated by applying simultaneously multiple light beams to the sample.

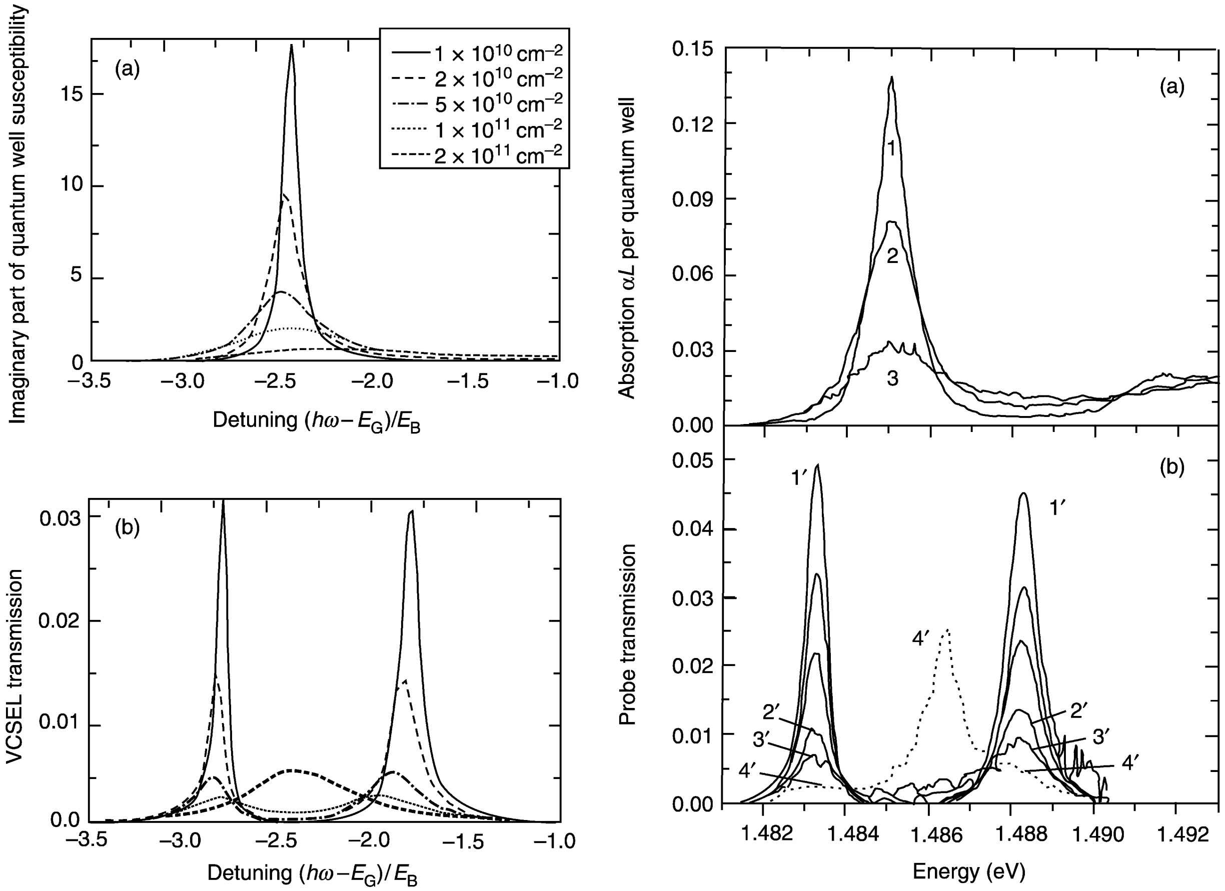

Figure 4 Left column: (a) microscopically computed bare quantum-well absorption spectrum, i.e., the imaginary part of the susceptibility, as a function of carrier density. (b) Calculated transmission of the quantum-well microcavity for corresponding densities. The right column shows experimentally determined (a) absorption and (b) microcavity transmission for the same conditions as in the calculations. From Jahnke F, Kira M, Koch SW, *et al.* (1996) Excitonic nonlinearities of semiconductor microcavities in the nonperturbative regime. *Physical Review Letters* 77: 5257. Copyright (2004) by the American Physical Society.

The ultimate cavity QED limit of normal-mode coupling follows when the light–matter interaction is so strong that even a single photon leads to splitting (genuine strong coupling). We briefly outline some aspects of the quantum statistical limit known from atomic physics. This situation can be analyzed with the so-called Jaynes–Cummings model where only two states of the atom are included together with a single light mode. In this case, the interaction between a single photon and an atom leads to a normal-mode splitting g_0. If two photons interact with one atom, the splitting increases to $2^{1/2}g_0$; more generally coupling between n photons and an atom leads to a splitting of $n^{1/2}g_0$. In the QED limit, the matter response is very nonlinear because one can detect the quantized nature of light directly from the energy splitting. This limit has already been reached in atomic cavity QED, whereas this regime is hard to approach with semiconductors because they behave more like multi-atom systems. If one has N atoms in a cavity, the first photon excites one atom, but there are N different ways of doing this, so the splitting is $N^{1/2}g_0$. If a second photon arrives, it interacts with the remaining unexcited atoms rather than the excited one, so that the splitting remains $N^{1/2}g_0$ as long as the number of photons is much smaller than N. Thus, in an N-atom system, or equivalently in semiconductors, the quantized light effects are much harder to observe than for a single atom. Currently, the normal-mode splitting in a semiconductor microcavity is explainable by the classical features of light, so that only the quantized nature of the matter is observed. In order to approach the ultimate cavity QED limit, one obviously has to reduce N and increase the light–matter coupling. In the future, this objective might be possible in quantum-dot/nanocavity systems where the semiconductor is confined in all spatial directions.

When an excited semiconductor is not under any influence of external classical light fields, it can still emit light via spontaneous emission resulting from the recombination of electron–hole pairs. The resulting light emission is called photoluminescence. Since the emission process takes place in a strongly interacting many-body system, one has to systematically include the Coulomb interaction and Fermionic features. The emission properties can be consistently described by the so-called semiconductor luminescence equations. When the microcavity photoluminescence spectrum is investigated, one observes a similar normal-mode splitting as for the transmission studies. However, this normal-mode coupling is not in the ultimate cavity QED limit even though the quantum fluctuations of the light field trigger the spontaneous emission. Nevertheless, the luminescence shows an interesting new transition as a function of the excitation level. Figure 5 contains a comparison between theory and experiment of normal-mode peak heights and positions as a function of carrier density. For low densities, the high-energy peak is lower in height but it overtakes the low-energy peak for moderate carrier densities that are still below lasing threshold. This threshold-like overtaking was attributed to Boser action, involving exciton formation, final-state stimulation, and Bose condensation. The inset shows an estimate of just how Bosonic the excitons are, i.e., how much they actually behave as independent atoms; the value unity corresponds to the fully Bosonic situation. The nonlinear luminescence transition takes place at a density regime where the underlying Fermionic electron and hole contributions become important (commutator is roughly 0.5). A more detailed analysis with the semiconductor luminescence equations shows that Fermionic carrier nonlinearities in a detuned cavity are responsible for the experimental observations. This example of a misinterpretation based on a Bosonic analysis stresses how important it is to include the Coulomb interaction and Fermion character of carriers when analyzing the properties of semiconductors.

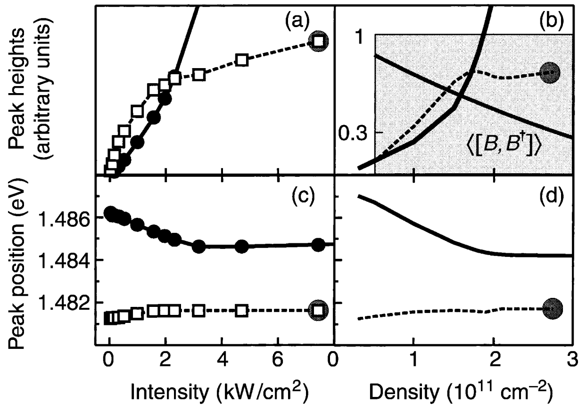

Figure 5 Measured microcavity luminescence intensities (a) and peak energies (c) versus excitation intensity for the high-energy peak (solid line) and the low-energy peak (dashed line). The results of the microscopic theory are shown in (b) and (d). From Kira M, Jahnke F, Koch SW, *et al.* (1997) Quantum theory of nonlinear semiconductor microcavity luminescence explaining "boser" experiments. *Physical Review Letters* 79: 5170. Copyright (2004) by the American Physical Society.

Superposition Principle: Observation of Quantum Optical Correlations

The above normal-mode-coupling investigations have shown interesting nonlinear effects, but they also revealed that the ultimate QED limit – demonstrating discrete states of light directly – has not yet been reached. Alternatively, a QED investigation may concentrate on other fundamental features of quantum mechanics. One possibility is to study the implications of the superposition principle which states that the joint properties of a light–matter system can always be expressed as a superposition that combines light and matter wavefunctions. The most dramatic consequence can be observed in so-called entangled wavefunctions which cannot be factorized into light and matter parts. In a factorized wavefunction, the light and matter parts are independent whereas entangled wavefunctions conditionally connect light and matter wavefunctions. To elaborate the subtle details of entanglement, we first analyze a simple example. Assume that light can be in two different polarization states, $|\sigma^+\rangle$ or $|\sigma^-\rangle$, and the matter part is either excited $|\mathrm{up}\rangle$ or de-excited $|\mathrm{down}\rangle$. A wavefunction, $[|\mathrm{up}\rangle + |\mathrm{down}\rangle][|\sigma^+\rangle + |\sigma^-\rangle]$, is clearly a superposition of the fundamental states, and at the same time the light and matter parts are completely factorized. However, the wavefunction $[|\mathrm{up}\rangle\ |\sigma^+\rangle + |\mathrm{down}\rangle|\sigma^-\rangle]$ is entangled, since the light and matter parts cannot be separated. In the entangled state, measurement on the light state will conditionally determine a definite state of the matter, whereas the factorized state has no such conditionality. The entanglement has far-reaching consequences which cannot be understood with a classical analysis. For example, the principles of teleportation and quantum computing follow directly from the controlled manipulation of different parts of the entangled wavefunction. For atomic systems, wavefunction entanglement has been demonstrated and utilized in several experiments. For semiconductors, the direct manipulation and detection of the wavefunction seems difficult due to the overwhelming complexity of the many-body wavefunctions. Once again, the entanglement and the wavefunction are simplest for low-dimensional structures; direct entanglement effects have been demonstrated recently in a single quantum dot and between a pair of dots. Also for more complicated semiconductor systems, such as a quantum well in a microcavity, the entanglement can be observed as correlations between light and matter. In this case, the existence of QED correlations basically means that light and matter properties depend conditionally on each other. In general, the direct entanglement and QED correlation investigations have a large development potential for semiconductors.

When a semiconductor is excited with an external light pulse, entanglement-related correlations couple the classical and luminescence emission dynamics. The resulting dynamic interplay between the semiconductor Bloch and luminescence equations is mainly mediated by the correlation between photons and electron–hole densities. In the following, we analyze a microcavity configuration where a strong pump pulse generates large QED correlations. The effect of the correlations is then measured by the response of a weak probe beam. Figure 6 shows a comparison of theory and experiment of probe reflection in a configuration where the pump and probe do not have any spectral overlap. The measured probe reflection displays long-living oscillations as a function of time delay, i.e., phase difference, between the pump and the probe. Only by including the QED correlation in the theory can the oscillatory probe reflection be explained. When the QED contributions are omitted from the theory, the phase difference does not have any effect on the probe reflection. Thus, this experiment–theory example represents a direct observation of the cavity QED effects in semiconductors. Figure 6 also shows more pronounced oscillations when two phase-locked pump pulses provide the excitation; again the full QED theory explains the enhanced oscillation features.

The QED features can also alter the normal-mode coupling characteristic of a weak probe beam. Figure 7 shows a comparison of theory and experiment for a situation where the pump spectrum is

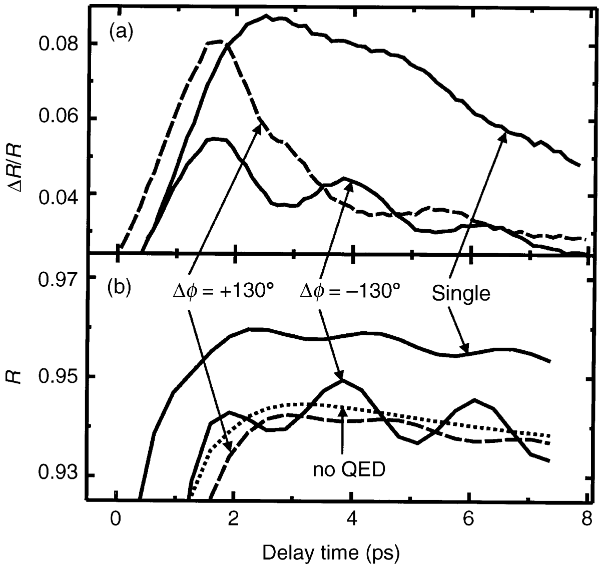

Figure 6 (a) Measured differential reflectivity and (b) computed reflection probability of the probe pulse as function of probe delay with respect to the excitation pulse. For two-pulse excitations, ±130° relative phase shifts are used. The dotted line is computed without the QED correlations. From Lee Y-S, Norris TB, Kira M, *et al.* (1999) Quantum correlations and intraband coherences in semiconductor cavity QED. *Physical Review Letters* 83: 5338. Copyright (2004) by the American Physical Society.

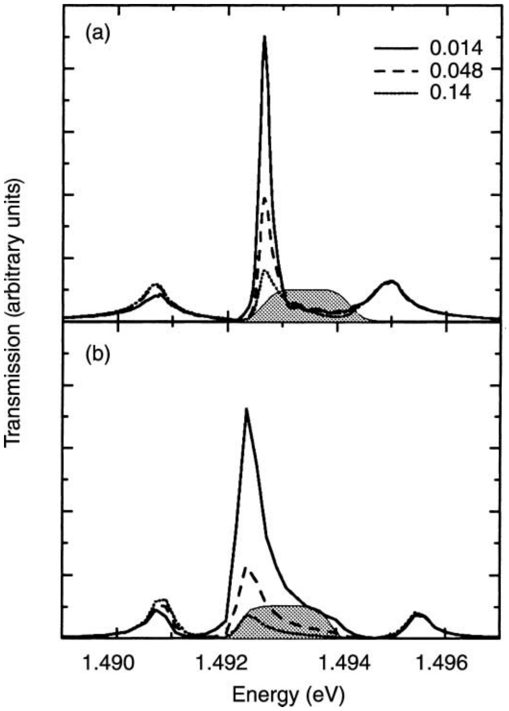

Figure 7 Dependence of probe transmission on probe intensity (nJ cm^{-2}) at a constant pump intensity of 330 nJ cm^{-2} for the microcavity. (a) Experiment, and (b) theory. The shaded regions indicate the pump spectrum. From Ell C, Brick P, Hübner M, *et al.* (2000) Quantum correlations in the nonperturbative regime of semiconductor microcavities. *Physical Review Letters* 85: 5392. Copyright (2004) by the American Physical Society.

located between the two normal-mode coupling resonances. Both the theory and experiment show a new third resonance which follows the energetic position of the pump. The third peak is a direct consequence of the QED correlations. Since the third peak is generated from the QED field–carrier correlations generated by the pump pulse, its magnitude follows the intensity of the pump. As the probe intensity is decreased the relative effect of the QED correlations is increased so that the third peak grows in the probe transmission.

In these QED investigations, the cavity plays an important role since the mirror feedback leads to a significant amplification of the QED effects through the enhanced light–matter coupling. From a practical point of view, normal-mode coupling also provides well-separated spectral features which can be used as classical-emission reference points.

Heisenberg Uncertainty Principle: Squeezing of Light Emission

Quantized light effects can also be investigated by measuring the light fluctuations of the field directly. To maximize the quantum effects, we analyze the field fluctuations in the emission directions where the classical field vanishes. This situation can be realized in planar quantum-well structures which are nearly free of disorder. In such systems, a light pulse propagating perpendicular to the structure leads to transmission and reflection of classical light only along the excitation axis. If the detection is performed at an angle, the so-called secondary emission is purely quantum mechanical. However, the classical and quantum emissions are coupled in the same way as for the QED correlation study. The resulting fluctuations of secondary emission obey the Heisenberg uncertainty principle; in the following, the special quantum features of these fluctuations are investigated.

In the full analysis, we determine the variance ΔX and ΔY of the emission as shown in Figure 8 (see also Figure 1). The Heisenberg uncertainty principle requires that the quadrature fluctuations obey $\Delta X \Delta Y \geq 1$. For a specific quadrature, the minimum uncertainty limit is usually defined to be $\Delta X = 1$ or $\Delta Y = 1$. If ΔX and ΔY are different, the observed light field is squeezed; and if the variance in one quadrature is less than one, the field is squeezed below the minimum uncertainty limit. In both cases, the field has a strong quantum nature; in particular, squeezing below unity suggests that measurements can be more accurate than the standard quantum limit for that quadrature.

To illustrate the behavior of the quantum fluctuations, we excite the quantum well resonantly with a relatively strong pulse. The emission is detected at an angle of 45° away from the excitation axis. Since the excitation pulse is relatively strong, the carrier density starts to oscillate during the pulse because Fermionic carriers can be excited only once; thus, further excitation actually leads to de-excitation. In other words, when these states become almost fully excited, the pulse can no longer excite the system, and we observe periodic de-excitation and excitation analogous to the Rabi-flopping of a strongly excited two-level system. Figure 8 shows the exciting pulse and corresponding quadrature fluctuations during the excitation process. As long as the pulse is present, the squeezing is at the 4% level and oscillates with the Rabi-flopping frequency. The quadrature fluctuations clearly show squeezing below the minimum uncertainty limit. Hence, the field has obvious quantum properties.

Similar squeezing and quantum characteristics statistics have been predicted and observed in the photon statistics of resonance fluorescence from two-level atoms subjected to an intense coherent light beam. The quantum properties of the scattered light

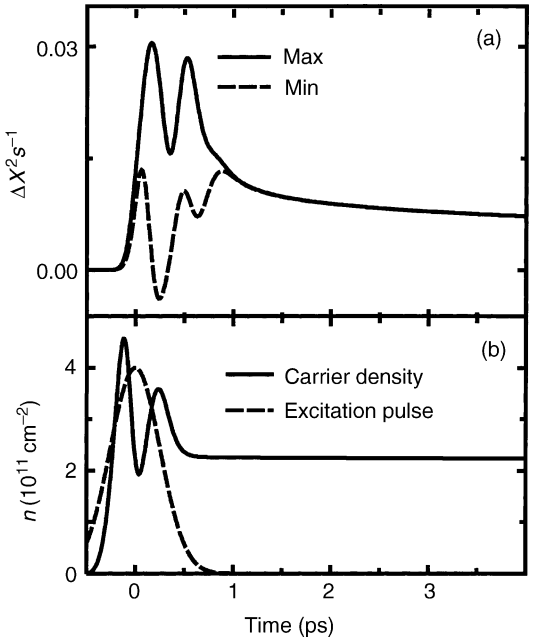

Figure 8 (a) Time-resolved maximum (solid line) and minimum (dashed line) deviations of the mode quadratures from the vacuum value of unity. (b) Corresponding time-resolved carrier density (solid line) and excitation pulse (dashed line). From Kira M, Jahnke F and Koch SW (1999) Quantum theory of secondary emission in optically excited semiconductor quantum wells. *Physical Review Letters* 82: 3544. Copyright (2004) by the American Physical Society.

in both quantum-well and atomic emission are enhanced when the driving field forces the system to oscillate between the excited and de-excited state. Just as described above, when the system becomes de-excited, it can no longer emit an additional photon; this inhibition manifests itself as sub-Poissonian photon statistics and squeezing in the mode quadratures. When the excitation pulse re-excites the system, such restrictions are no longer imposed. Thus, the field properties show quantum features oscillating with the Rabi-flopping frequency. Squeezing effects can also be observed with excitation by an electric current; for example, amplitude squeezing of diode laser emission has been demonstrated by controlling the electron statistics of the current flow.

The squeezing investigations demonstrate that some QED effects are expected to be seen without a cavity. Figure 9 shows the quadrature fluctuations when the cavity is added to the squeezing analysis. We observe qualitatively similar squeezing, but now the level of squeezing is enhanced up to 30% compared to the minimum uncertainty. This is once again a demonstration how QED effects can be amplified by using an increased light–matter coupling provided by the cavity.

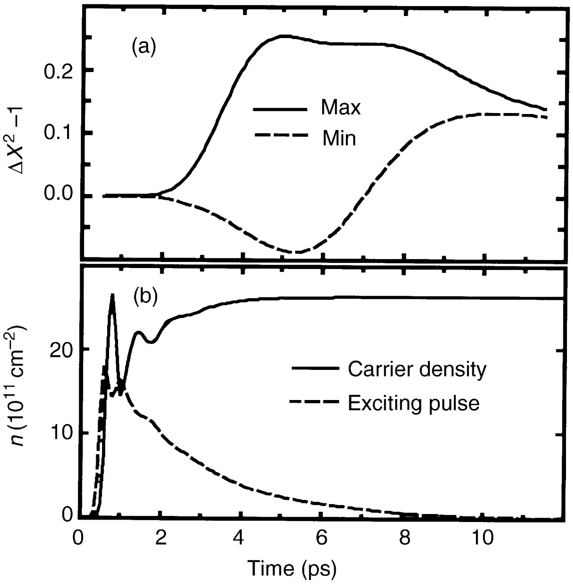

Figure 9 Squeezed light from a quantum well inside a microcavity: (a) the time evolution of maximum and minimum of the quadrature fluctuations; (b) the corresponding density and light field intensity at the quantum well.

Summary

This article overviews some of the most intriguing features of cavity quantum electrodynamics in semiconductors. Even though the related research has started only recently, several important quantum phenomena have already been predicted and measured: (a) the discrete quantum mechanical states of light and/or matter can be measured with strong coupling; (b) the consequences of the superposition principle have been detected as light–matter entanglement; and (c) the Heisenberg uncertainty principle has been tested via the squeezing of light. In all cases, the cavity enhances light–matter coupling leading to more pronounced quantum effects. Compared to atomic systems, the semiconductor has strong Coulomb correlation effects due to the relatively high density, which makes the analysis challenging. However, the semiconductor can also provide new mechanisms like the photon/carrier-density correlations which trigger new unexpected quantum phenomena.

The field of semiconductor quantum optics is very active and is developing rapidly. It benefits both from advances in computer capabilities and in semiconductor crystal growth. The increasing computer capacity allows more profound, accurate, and realistic modeling of semiconductor structures. At the same time, advances in crystal growth technology provide us with improved samples which are almost disorderless. In particular, the growth of quantum wells with narrower exciton linewidths and quantum dots with larger dipole moments and reduced dephasing rates may be achieved, which is certain to make quantum features increasingly apparent and unavoidable.

All these research efforts eventually focus on producing devices utilizing quantum mechanical principles. One of the main objectives endeavors to develop quantum logic components for building blocks of quantum computers. Similar expanding possibilities can be expected, e.g., for accuracy of detection, device efficiency, and component design in general. Considering all of this, we are almost guaranteed to see new astonishing advancements; in this bright future, semiconductor cavity QED research will most likely be a pre-eminent element.

See also

Quantum Electrodynamics: Cavity QED.

Further Reading

Allen L and Eberly JH (1987) Optical Resonance and Two-Level Atoms, 1st edn. New York: Dover.

Banyai L and Koch SW (1993) Semiconductor Quantum Dots, 1st edn. Singapore: World Scientific.

Berman PR (ed.) (1994) Several articles in *Cavity Quantum Electrodynamics*, 1st edn. Boston, MA: Academic Press.

Cohen-Tannoudji C, Dupont-Roc J and Grynberg G (1989) Photons and Atoms, 3rd edn. New York: Wiley.

Haug H and Koch SW (2004) Quantum Theory of the Optical and Electronic Properties of Semiconductors, 4th edn. Singapore: World Scientific.

Khitrova G, Gibbs HM, Jahnke F, Kira M and Koch SW (1999) Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Reviews of Modern Physics* 71: 1591–1639.

Kira M, Jahnke F, Hoyer W and Koch SW (1999) Quantum theory of spontaneous emission and coherent effects in semiconductor microstructures. *Progress in Quantum Electronics* 23: 189–279.

Scully MO and Zubairy MS (1999) Quantum Optics, 1st edn. Cambridge, UK: Cambridge University Press.

Walls DF and Milburn GJ (1994) Quantum Optics, 1st edn. New York: Springer-Verlag.

QUANTUM OPTICS

Contents

Atom Optics
Atomic Coherence Effects
Entanglement and Quantum Information
Laser Cooling of Ions
Quantum Computing with Atoms
Squeezed Phonons in Solids

Atom Optics

A D Cronin and D E Pritchard, Massachussetts Institute of Technology, Cambridge, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

In 1923 Louis de Broglie suggested that particles of matter propagate as waves with a wavelength

$$\lambda_{dB} = h/p \qquad [1]$$

where h is Planck's constant and p is the particle's momentum. This breakthrough idea motivated Schrödinger's equation, which is the equation of motion of quantum mechanics.

Schrödinger's equation is a wave propagation equation, and therefore implies the existence of a whole new type of optics – matter wave optics – in which electron waves, neutron waves, atom waves, and entire molecule waves can be manipulated coherently. Such optics enable new measurement technologies and devices that depend on matter wave interference.

Guided by a knowledge of light waves, one can understand that the de Broglie wavelength of a particle sets the scale for focusing, diffraction, and interference. However, there are differences between matter and light. For example, the vacuum is tremendously dispersive for matter waves. In other words, the group velocity for a particle depends on its de Broglie wavelength, while for light waves the group velocity is independent of wavelength.

To illuminate the mathematical connection with light optics, we note the wave equation for light is

$$\nabla^2 \vec{E} = \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} \qquad [2]$$

where $\vec{E} = \vec{E}(\vec{x}, t)$ is the optical electric field, and the parameters μ and ε may vary throughout space, describing an index of refraction. The wave equation for matter waves is Schrödinger's equation:

$$\left(-\frac{\hbar^2\nabla^2}{2m} + V\right)\psi = i\hbar\frac{\partial}{\partial t}\psi \qquad [3]$$

where $\psi = \psi(\vec{x}, t)$ is the probability amplitude, or wavefunction, for a particle of mass m. Here, the potential energy V can vary throughout space, and thus control the phase and amplitude of propagating atoms. Just as density for photons is proportional to $\vec{E}(\vec{x})^2$, the probability density for particles is given by $|\psi(\vec{x})|^2$.

In a vacuum $V = 0$ and $(\mu_0\varepsilon_0)^{-1/2} = c$, and if we assume that both ψ and E have a time dependence described by $e^{-ik\omega}$, the wave equations become

$$\left[\nabla^2 + \frac{\omega^2}{c^2}\right]\vec{E} = 0 \qquad [4]$$

and

$$\left[\nabla^2 + \frac{2m\omega}{\hbar}\right]\psi = 0 \qquad [5]$$

which look similar, but have a critical difference stemming from the single time derivative in Schrödinger's equation. To see how this makes the vacuum dispersive for matter waves, but not for light, consider a propagating plane wave of either ψ or $\vec{E}$ described by $e^{i(kx-\omega t)}$ (for either kind of wave, $k \equiv 2\pi(\lambda)^{-1}$). Using the wave equation for light waves, the dispersion relation is

$$\omega = ck \qquad [6]$$

where c is the phase (and group) velocity of light, which is independent of k. The same plane wave substituted into Schrödinger's equation for matter waves yields the dispersion relation

$$\omega = \frac{\hbar k^2}{2m} \qquad [7]$$

This is quadratic for matter waves, which means the group velocity $v_g = d\omega/dk$ is twice the phase velocity $v_p = \omega/k$, and both still depend on k.

From an engineering perspective, matter wave optics consists of lenses, mirrors, beam-splitters, and other components of an optics toolkit. The fact that electrons and neutrons can be transmitted through solid material, and that electrons can be reflected or refracted by static electromagnetic fields, has enabled optical elements to be created for these kinds of matter waves. For atoms, however, finding suitable materials for lenses or mirrors is a daunting task, since with rare exceptions (e.g., cold H on liquid He) atoms thermalize on surfaces rather than elastically bouncing off or passing through them. This was the major obstacle which delayed the blossoming of atom optics until the mid-1980s.

This obstacle has now been circumvented, but only with difficulty. Electromagnetic fields (whose spatial configurations are severely constrained by Maxwell's equations) can change the potential energy of neutral atoms, but since an atom's net charge is zero, provide only a weak and highly dispersive index of refraction based on the atom's polarizability or magnetic moment. However, these effects are tremendously enhanced if the electromagnetic fields oscillate near a transition frequency of the atom. Hence, propagating waves and standing waves of single-frequency laser light have become standard tools in atom optics. Standing light waves are often referred to as a light crystal, because the periodic potential can diffract atom beams.

An alternative source of atom optical components is based on nanofabrication. Thin sheets of material with patterned holes serve as absorptive atom optics by transmitting atom waves only through the nanometer scale openings. A variety of techniques have been used to make physical diffraction gratings, zone plates, and holograms for atom waves. Methods such as electron beam lithography and ultraviolet photolithography both rely on sophistocated etching procedures to make the final free-standing structures. Thus, in either approach – near-resonant light, or miniature physical structures – current technologies such as the tunable laser or nanofabrication techniques have been required for the development of useful atom optics.

While coherent manipulation of atom waves is often the goal, incoherent manipulations are possible also, e.g., if photons are spontaneously scattered. Dissipative processes offer an additional class of possibilities such as increasing the brightness of an atom source, and are therefore an important part of atom optics. This article excludes atom slowing and trapping, and focuses instead on atom diffraction – a coherent process which can be used to make beam splitters for atom interferometers.

Nanofabricated Atom Optics

Because atoms stick to surfaces, mechanical structures are absorptive atom optics. Yet the de Broglie waves transmitted through slots of a mask are nearly unaffected, provided the mask is thin enough. This is because the van der Waals potential energy of interaction with nearby surfaces falls off rapidly for neutral atoms.

Consider, for example, an atom wave packet being diffracted by a mechanical grating made with physical bars and slots as shown in Figure 1. Initially, let the atom wave packet have a slowly varying spatial envelope, $f = f(x, y, z)$, and momentum only in the $\hat{x}$-direction, $\vec{p} = \hbar\vec{k}_0 = h(\lambda_{dB})^{-1}\hat{x}$, so the wavefunction can be written as

$$\psi = fe^{i(k_0x-wt)} \qquad [8]$$

Immediately after the grating, only the portions of the wavefunction passing through slots remain nonzero, as depicted in Figure 2. Here, the wavefunction is now periodic both due to the grating lattice vector $k_g\hat{y}$ and the initial wavevector $k_0\hat{x}$. According to the Schrödinger equation, a periodicity in any direction implies a momentum in that direction. Thus, variation of ψ in the $\hat{y}$ variation gives rise to a new momentum distribution.

The square wave modulating the envelope of the wave packet can be described by a Fourier sum times the incident wavefunction

$$\psi = fe^{i(k_0x-\omega t)}\sum_{n=-\infty}^{\infty} a_n \cos(nk_g y) \qquad [9]$$

where $k_g = 2\pi(d_g)^{-1}$ is the grating lattice vector. The only other physical parameter of the grating is the open fraction, γ, and this determines the Fourier amplitudes a_n

$$a_n = \frac{\sin(n\pi\gamma)}{n\pi} \qquad [10]$$

From the principle of superposition (which is applicable here because the Schrödinger equation is linear) the wavefunction can be written as a sum of traveling waves with wavevectors $\vec{k}_n$:

$$\psi = f\sum_{n=-\infty}^{\infty} a_n e^{(i\vec{k}_n\cdot\vec{r}-wt)} \qquad [11]$$

$$\vec{k}_n = \frac{k_0\hat{x} + nk_g\hat{y}}{N} \qquad [12]$$

where, just as in standard optics, the factor N serves to keep $|\vec{k}_n|$ on a spherical shell, i.e., the grating does not add energy to the atom.

Inserting this expression for the wavefunction into Schrödinger's equation we get the dispersion relation

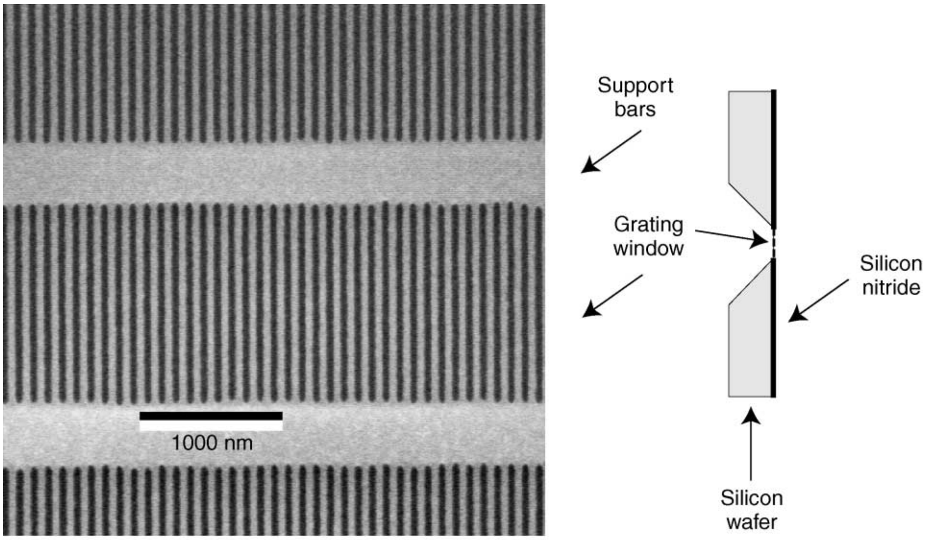

Front view

Side view

Figure 1 Scanning electron microscope image of a silicon nitride diffraction grating for atom waves. The bars (light) absorb, and the slots (dark) transmit atom waves. The grating period is 100 nm, the open fraction is 0.67, and the free-standing bars are supported every 5 μm with a thicker strip of silicon nitride material. The grating was fabricated by Tim Savas and Henry I Smith at MIT NanoStructures laboratory.

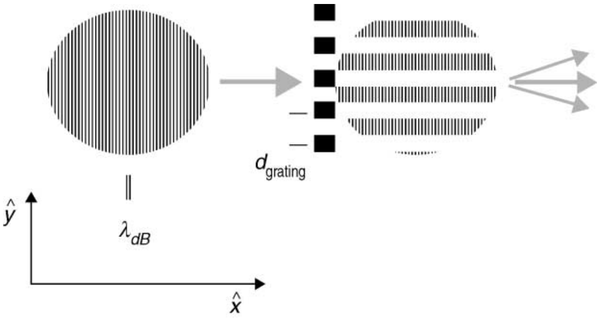

Figure 2 Schematic diagram of an atom wavepacket passing through an absorbing grating, as seen from above. The atom wavepacket is depicted before and after passing through the grating from left to right.

for matter waves and a different group velocity, v_n, for each k_n

$$\hbar\omega = \frac{\hbar^2 k_n^2}{2m} \qquad [13]$$

$$v_n = \frac{d\omega}{dk_n} = \frac{\hbar k_n}{m} = p_n/m \qquad [14]$$

Thus, diffraction through a grating puts an atom into a coherent superposition of different momenta, each separated by $\sim\hbar k_g \hat{y}$. Some time later, the atom will be in a coherent superposition of multiple locations, and the far-field probability density will be peaked at integer multiples of the diffraction angle

$$\Theta_D = \frac{k_g}{k_0} = \frac{\lambda_{dB}}{d_{\text{grating}}} \qquad [15]$$

where d_{grating} is the lattice spacing of the grating and the small-angle approximation for $\sin(\theta)$ has been used. Exactly as in light optics, these diffraction gratings are good momentum spectrometers, i.e., the diffraction angle is a measure of longitudinal momentum. (Contrary to light optics, group velocity can also be used to measure matter wave momentum.) Coherence between the diffraction orders remains until a perturbation occurs which would (even in principle) be sufficient to determine which path the atom took. This assertion has been tested with atoms in an interferometer (discussed below).

Atom flux in the nth diffraction order is given by

$$I_n = \left(\frac{\sin(n\pi\gamma)}{n\pi}\right)^2 \qquad [16]$$

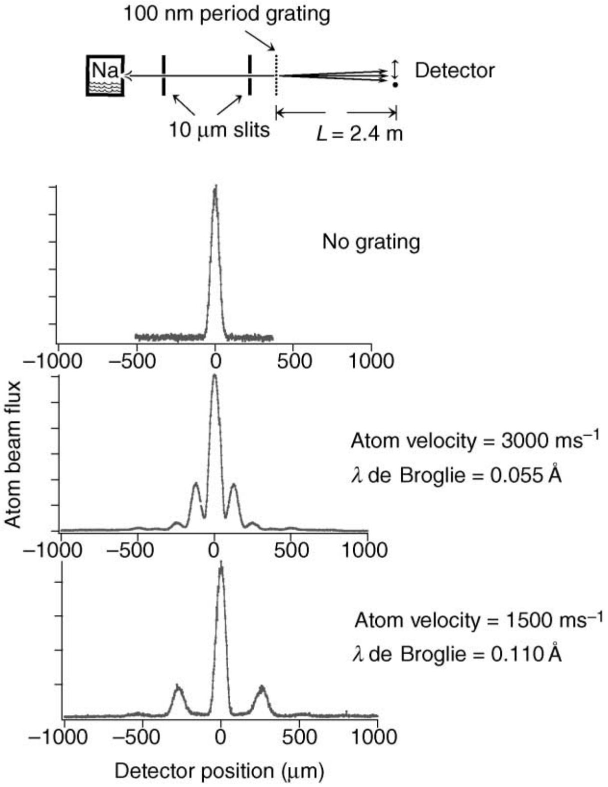

Figure 3 Diffraction data from a beam of sodium formed by passing atoms through two collimating slits followed by a nanofabricated diffraction grating. The momentum of the atoms determines their de Broglie wavelength, and thus their diffraction angle $\sin(\Theta_D) = \lambda_{dB}/d_{\text{grating}}$. These data were obtained at MIT by the authors.

To resolve the diffracted orders the initial transverse momentum distribution of the atoms must be smaller than k_g, and the time until the far-field observation must be sufficiently large given the beam width. At intermediate times, the atom waves are in the Fresnel diffraction region where the Talbot effect can be observed. This effect causes the atom waves to form a replica of the grating in free space at integer multiples of the Talbot length, $Z_T = 2d^2_{\text{grating}}\,\lambda^{-1}_{dB}$.

Figure 3 shows the far-field diffraction pattern from a beam of sodium (^{23}Na) atoms using a material grating with a period of 100 nm. The atom beam velocity peaked at 1500 ms^{-1} corresponds to a de Broglie wavelength of 0.11 Å; thus, $\Theta_D \approx 10^{-4}$ rad. (The beam of 3000 ms^{-1} atoms diffracts at half this angle.) The spread in longitudinal velocity of $\sigma_v/v \approx 6\%$ can be deduced from the data because it slightly increases the width for higher order diffraction peaks. The open fraction of the grating, $\gamma \approx 60\%$, can also be determined from the diffraction data, because it determines the relative intensity of the orders. Effects of the atoms' finite size and interaction with the grating bars make small corrections to this simple theory.

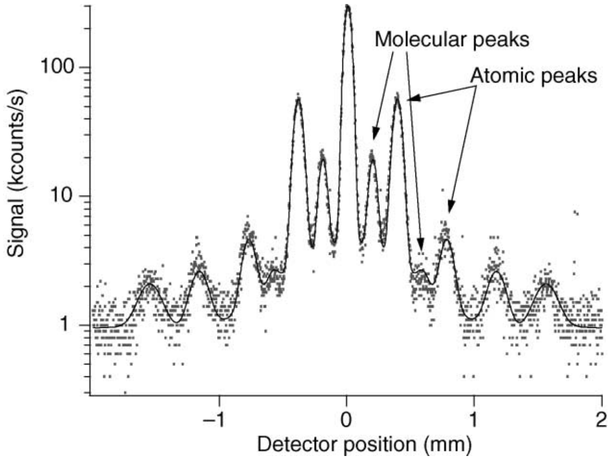

Figure 4 Diffraction data using a mixed beam of sodium atoms (Na) and sodium dimer molecules (Na_2). From this source molecules have the same speed as atoms, thus twice the momentum. Thus, molecule waves are diffracted at half the diffraction angle compared to atom waves. These data were obtained at MIT by the authors.

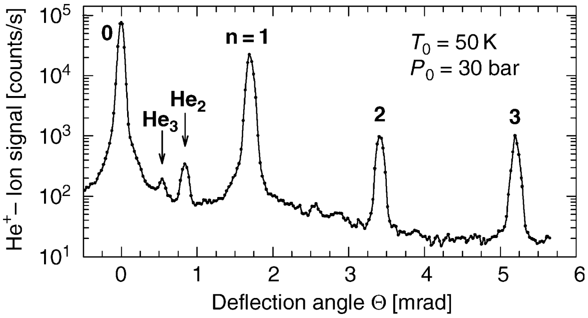

Figure 5 Diffraction of helium atoms and helium molecules through a nanofabricated grating. These data, reproduced with permission from Wieland Schoellkopf, were obtained at the Max-Planck-Institute in Göttingen.

Nanofabricated atom optics differ from those based on light forces in a number of ways: they are amplitude structures (with corresponding loss of transmission intensity), they are species-independent, their scale size can be several times smaller than attainable with light, and they can be arbitrarily patterned since they are fabricated by electron beam lithography. Insensitivity to species is demonstrated in Figure 4, where a beam of Na_2 molecules and Na atoms is diffracted and separated by a nanofabricated diffraction grating. The flexibility of electron beam lithography has allowed fabrication not only of diffraction gratings, but also of spherical and cylindrical zone plates, as well as a combination of lens and a hologram that generates a focused atom image. One stunning application laid to rest a long-standing argument concerning whether a stable bound state of the $^4He_2$ dimer exists. For this a diffraction grating was used to separate and resolve $^4He_2$ and more massive clusters (Figure 5). Subsequently a nano-sieve was used to estimate the size of the $^4He_2$ dimers to be 62 ± 10 Å.

Standing Waves of Light

In contrast to nanofabricated structures, standing waves of near-resonant light are phase gratings for atoms. The optical electric field dynamically polarizes the atoms and causes a shift to the ground state energy (the ac-Stark shift). For optical electric fields oscillating above (below) the atomic resonance frequency, the ac-Stark shift increases (decreases) potential energy for atoms and in either case perturbs the de Broglie wave phase.

In many respects, the interaction of an atom with a standing light wave is richer than the more familiar topic of light–atom interactions in a traveling light wave. Part of this richness reflects two ways in which a standing wave can be considered, either as two counter-propagating traveling waves or as a single, stationary standing wave. The standing wave–atom interaction is capable of transferring momentum in well-determined quantities, coherently splitting atomic wave packets, and generating forces much larger than possible with spontaneously scattered light. Not surprisingly, this interaction has many distinct facets that only appear in different regimes of the interaction parameters (intensity, detuning, and pulse duration) and atomic parameters (mass, initial momentum with respect to the standing wave, and excited state natural lifetime).

We begin with the Bragg scattering of atoms from a standing wave light grating. Although it can be difficult to realize the physical conditions that assure its occurrence, pure Bragg scattering is simpler than intermediate cases involving spontaneous decay (from light too close to resonance) or Kapitza–Dirac diffraction (from shorter interaction times).

Consider a standing wave light grating formed by two counter-propagating plane waves (traveling parallel to the z-axis) of equal amplitude, E_0, wavevector, k, frequency, ω, polarization vector, $\hat{e}$, and temporal envelope function, $f(t)$:

$$\vec{E}(z,t) = E_0 f(t)\sin(kz-\omega t)\hat{e} + E_0 f(t)\sin(kz+\omega t)\hat{e} \qquad [17]$$

$$= 2E_0 f(t)\sin(kz)\cos(\omega t)\hat{e} \qquad [18]$$

We would like to work with momentum states as our atomic basis, thus it is easiest to consider the description of the electric field driving the transitions in terms of two counter-propagating traveling waves of definite momentum (eqn [17]), as opposed to the single standing wave they jointly form (eqn [18]).

Momentum is transferred by paired stimulated absorption and emission processes, resulting in a transfer of photons between the traveling waves. An N_B^{th} order diffraction process transfers N_B photons from one traveling wave to the counter-propagating traveling wave and changes the atomic momentum by $2N_B\hbar k\hat{z}$. Furthermore, atomic population is transferred only between $|g, -N_B\hbar k\rangle$ and $|g, +N_B\hbar k\rangle$, where $|g(e), \pm N\hbar k\rangle$ denotes a two-level atom in its ground (excited) state with momentum $\pm N\hbar k$ parallel to the standing wave axis. The excited state remains nominally unpopulated so long as the temporal envelope function, $f(t)$, does not have strong frequency components near the laser detuning

$$\delta = \omega - \omega_0 \qquad [19]$$

where ω_0 is the unperturbed frequency of the atomic transition. Furthermore, for a given initial state ($|g, -N_B\hbar k\rangle$) the uniqueness of the final state ($|g, +N_B\hbar k\rangle$) occurs because of the fundamental assumption that makes Bragg scattering so simple to describe; the uncertainty in the photon energy driving the transitions is small compared to the energy separation between neighboring momentum states. A quantitative discussion of the validity of this assumption will be given later.

We now calculate the probability, $P_1^B(\tau)$, of the first order ($N_B = 1$) Bragg process. Scattering transfers an atom from $|g, -\hbar k\rangle$ to $|g, +\hbar k\rangle$ when the atoms interact with a constant light intensity for a time τ (i.e., in eqns [17] and [18] $f(t)$ is a square wave of unit amplitude and duration τ). The geometry for Bragg diffraction from a standing wave of light is shown in Figure 6, and the energy/momentum states involved in the transition are depicted in Figure 7.

Considering the electronic degree of freedom, we rewrite the Schrödinger equation as

$$(\mathcal{H}_0 + \mathcal{H}_{\text{int}})|\psi\rangle = i\hbar\frac{\partial}{\partial t}|\psi\rangle \qquad [20]$$

where the total Hamiltonian $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\text{int}}$ is a matrix. The wavefunction is a vector of coefficients c_i in the basis $\{|g, -\hbar k\rangle, |e, 0\rangle, |g, +\hbar k\rangle\}$, with normalized population amplitudes for each component:

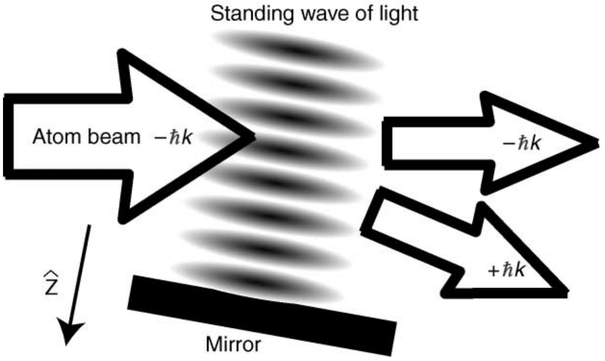

Figure 6 Two counter-propagating running waves of light superimpose to form a standing wave of light. This 'light crystal' acts as a diffraction grating for atoms, shown here in the configuration for Bragg diffraction.

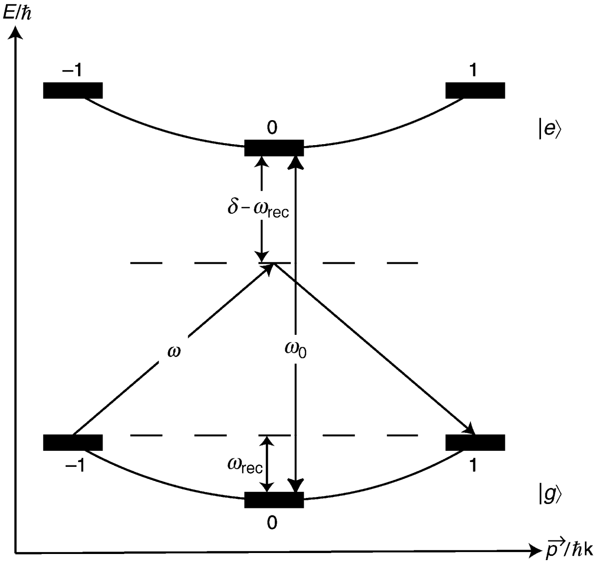

Figure 7 Energy vs. momentum diagram describing atomic states involved in Bragg diffraction. The total energy is due to the electronic state ($|g\rangle$, $|e\rangle$) plus the kinetic energy associated with the recoil from photon emission (or absorption). In Bragg diffraction, atoms make a coherent transition from $|g, -\hbar k\rangle \leftrightarrow |g, +\hbar k\rangle$. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t) \\ c_0(t) \\ c_{+1}(t) \end{pmatrix} \qquad [21]$$

In the electric dipole approximation, the interaction Hamiltonian is $\mathcal{H}_{\text{int}}(t) = -\vec{\mu} \cdot \vec{E}(t)$, where $\mu = \langle e|q\vec{r}|g\rangle \cdot \hat{e}$ is the electric dipole matrix element connecting the ground ($|g\rangle$) and excited ($|e\rangle$) states of the atom. $\hat{e}$ is the polarization vector of the light and q is the charge of the electron. By momentum conservation, only the plane wave traveling in the $\pm\hat{z}$ direction couples the $|e,0\rangle \leftrightarrow |g, \mp\hbar k\rangle$ transitions. By using this argument, we are effectively viewing the electric field as a quantum mechanical operator. Expanding the sinusoidal variation of the electric field in terms of complex exponentials and treating the spatially dependent complex exponential terms as quantum mechanical momentum translation operators yields the interaction Hamiltonian:

$$\mathcal{H}_{\text{int}} = \frac{\hbar\omega_R}{2}\begin{pmatrix} 0 & ie^{i\omega t} & 0 \\ -ie^{-i\omega t} & 0 & ie^{-i\omega t} \\ 0 & -ie^{i\omega t} & 0 \end{pmatrix} \qquad [22]$$

For example, $e^{\pm ik\hat{z}}|g, n\hbar k\rangle = |e, (n \pm 1)\hbar k\rangle$ describes absorption, and for stimulated emission, $e^{\mp ik\hat{z}}|e, n\hbar k\rangle = |g, (n \pm 1)\hbar k\rangle$. This procedure includes the rotating wave approximation. In addition, we have neglected any frequency components associated with the sudden switch on of the fields and the finite duration of the light–atom interaction. In formulating $\mathcal{H}_{\text{int}}$, the strength of the light–atom interaction is parameterized by the single-photon Rabi frequency:

$$\omega_R = \frac{\mu E_0}{\hbar} \qquad [23]$$

Without loss of generality, we will take μ and hence ω_R to be positive, real-valued quantities.

The total Hamiltonian follows simply by including the electronic and kinetic energy terms:

$$\mathcal{H}_0 = \hbar\begin{pmatrix} \omega_{\text{rec}} & 0 & 0 \\ 0 & \omega_0 & 0 \\ 0 & 0 & \omega_{\text{rec}} \end{pmatrix} \qquad [24]$$

where the single-photon recoil energy, E_{rec}, of an atom of mass m is given by

$$E_{\text{rec}} = \hbar\omega_{\text{rec}} = \frac{\hbar^2 k^2}{2m} \qquad [25]$$

Making the *ansatz* for the solution wavefunction as

$$|\psi\rangle = \begin{pmatrix} c_{-1}(t)e^{-i\omega_{\text{rec}}t} \\ c_0(t)e^{-i\omega_0 t} \\ c_{+1}(t)e^{-i\omega_{\text{rec}}t} \end{pmatrix} \qquad [26]$$

and substituting into the Schrödinger equation yields three coupled first-order differential equations:

$$\dot{c}_{\pm 1}(t) = \mp\frac{\omega_R}{2}e^{i\Delta t}c_0(t), \qquad [27]$$

$$\dot{c}_0(t) = \frac{\omega_R}{2}e^{-i\Delta t}\big(c_{+1}(t) - c_{-1}(t)\big), \qquad [28]$$

where $\Delta = \delta + \omega_{\text{rec}} = (\omega - \omega_0) + \omega_{\text{rec}}$. Differentiating eqn [27] and substituting eqn [28] into the result yields two coupled second-order differential equations:

$$\ddot{c}_{\pm 1}(t) - i\Delta\dot{c}_{\pm 1}(t) + \frac{\omega_R^2}{4}\big(c_{\pm 1}(t) - c_{\mp 1}(t)\big) = 0 \qquad [29]$$

With the initial conditions:

$$c_{-1}(0) = 1 \qquad [30]$$

$$c_0(0) = 0 \Rightarrow \dot{c}_{\pm 1}(0) = 0 \qquad [31]$$

$$c_{+1}(0) = 0 \qquad [32]$$

and the assumption $\Delta^2 \gg \omega_R^2$, the solutions to eqn [29] are

$$c_{-1}(t) = e^{-i\frac{\omega_R^{(2)}}{2}t}\cos\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [33]$$

$$c_{+1}(t) = ie^{-i\frac{\omega_R^{(2)}}{2}t}\sin\left(\frac{\omega_R^{(2)}}{2}t\right) \qquad [34]$$

where the two-photon Rabi frequency is

$$\omega_R^{(2)} = \frac{\omega_R^2}{2\Delta} \to \frac{\omega_R^2}{2\delta}, \quad |\delta| \gg \omega_{\text{rec}} \qquad [35]$$

with both transitions driven at equal single-photon Rabi frequencies, ω_R.

Substituting the solution of eqn [33] or [34] into eqn [27] yields an expression for the excited state amplitude:

$$c_0(t) = -i\frac{\omega_R}{2\Delta}e^{-i\Delta t}e^{-i\omega_R^{(2)}t} \qquad [36]$$

This will be important in calculating the rate of spontaneous emission events later.

The solutions for $c_{-1}(t)$ and $c_{+1}(t)$ oscillate with the interaction duration, τ, yielding the result for the $|g, -\hbar k\rangle \to |g, +\hbar k\rangle$ transition probability:

$$P_1^B(\tau) = |c_{+1}(\tau)|^2 = \sin^2\left(\frac{\omega_R^{(2)}}{2}\tau\right) \qquad [37]$$

Thus, the system oscillates between the two momentum states $|g, -\hbar k\rangle$ and $|g, +\hbar k\rangle$ in a manner analogous to the Rabi oscillation of atomic population between two resonantly coupled states. This solution with oscillatory probabilities for the two Bragg coupled states is known as the Pendellösung and has been observed for atoms, neutrons, and X-rays (Figure 8). The nice feature here is that the strength of the grating can be actually controlled by the intensity of the light.

Viewing Bragg scattering as a two-photon transition from the initial ground state to the final ground state with opposite momentum, illuminates the close connection with a Raman transition between two internal sub-states of the ground state manifold, each with its own external momentum state. The formalism describing the Raman transition is basically the same as that presented here, except the two transitions can be driven at different single-photon Rabi frequencies, ω_{R1} and ω_{R2}, so that the generic two-photon Rabi frequency is given by $\omega_R^{(2)} = \omega_{R1}\omega_{R2}/2\Delta$, where Δ is the detuning from the intermediate state.

An N_B^{th} order Bragg process (similar to a $2N_B$-photon Raman process) is a coherent succession of N_B two-photon transitions from $|g, -N_B\hbar k\rangle$ to $|g, +N_B\hbar k\rangle$ with $2N_B - 1$ intermediate states of the form:

$$|e, (-N_B + 1)\hbar k\rangle, |g, (-N_B + 2)\hbar k\rangle \ldots$$

$$\ldots |g, (N_B - 2)\hbar k\rangle, |e, (N_B - 1)\hbar k\rangle$$

Such a process is characterized by a $2N_B$-photon Rabi frequency given by

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{2N_B-1}\Delta_1\Delta_2\cdots\Delta_{2N_B-1}} \qquad [38]$$

where Δ_n is the detuning from the n^{th} intermediate state. Figure 9 shows what this process would look like for an N_B^{th} order Bragg transition where the intermediate state detunings are given by

$$\Delta_n = \left\{ \begin{array}{ll} \delta + (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ odd} \\ (2N_B n - n^2)\omega_{\text{rec}} : & n \text{ even} \end{array} \right\} \qquad [39]$$

Substituting these detunings into eqn [38] yields the N_B^{th} order Bragg transition $2N_B$-photon Rabi frequency, $\omega_R^{(2N_B)}$:

$$\omega_R^{(2N_B)} = \frac{[\omega_R]^{2N_B}}{2^{4N_B-3}[(N_B-1)!]^2\delta^{N_B}\omega_{\text{rec}}^{N_B-1}} \qquad [40]$$

where we have assumed $|\delta| \gg N_B^2\omega_{\text{rec}}$.

To ensure that the system truly undergoes Bragg scattering, and validate the assumption that only states of equal kinetic energy and opposite

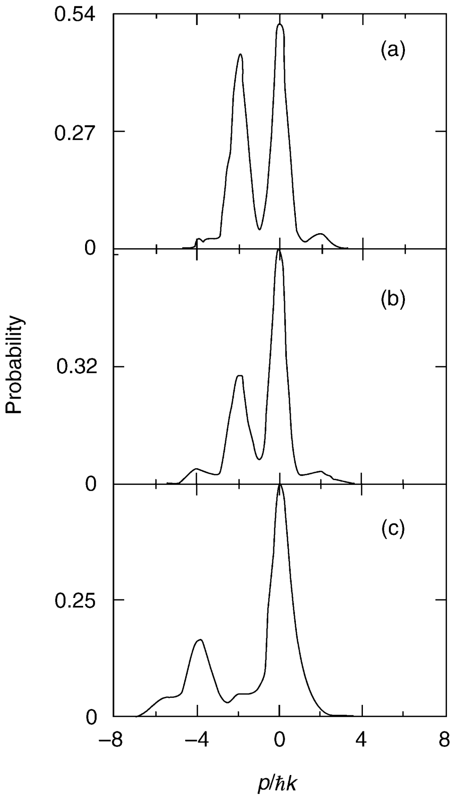

Figure 8 Bragg scattering of an atom beam from an optical standing wave. First-order Bragg scattering at (a) lower power, (b) higher power of the laser beams (Pendellösung has set in, increasing the amplitude of the undeflected peak), (c) second-order Bragg scattering. Reprinted with permission from *Physical Review Letters* 60(6): 515–518. Copyright (1998) by the American Physical Society.

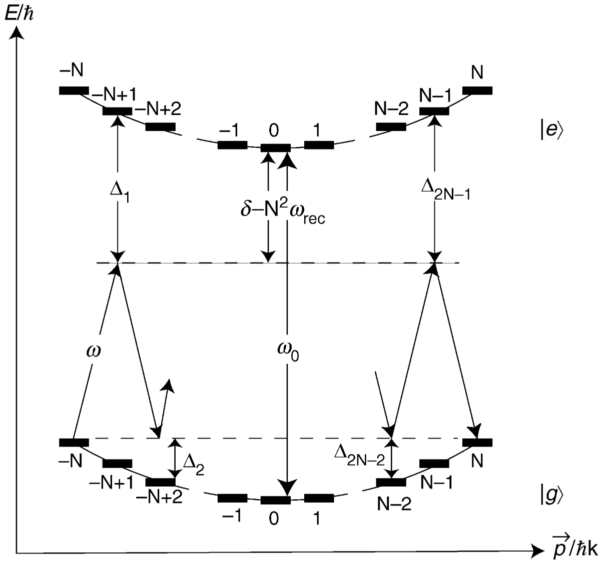

Figure 9 Energy vs. momentum diagram describing atomic states involved in higher-order Bragg diffraction. Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

momentum are coupled, the overall exposure time, τ, of the atoms to the fields must be limited both from below and above.

The lower bound on τ is necessary to resolve the final momentum state ($|g, N_B\hbar k\rangle$) from neighboring momentum states ($|g, (N_B \pm 2)\hbar k\rangle$) two photon recoil momenta away. Avoiding population transfer into these states requires that the energy separation be resolvable, or

$$\tau(\delta E) \gg h \qquad [41]$$

For first-order Bragg scattering processes ($N_B = 1$) the nearest lying momentum states that may be mistakenly populated are $|g, \pm 3\hbar k\rangle$, for which

$$\delta E = \frac{p_{n=3}^2}{2m} - \frac{p_{n=1}^2}{2m} = 8\hbar\omega_{\text{rec}} \qquad [42]$$

Therefore

$$\tau \gg \frac{h}{8\hbar\omega_{\text{rec}}} = \frac{\pi}{4\omega_{\text{rec}}} \qquad [43]$$

For sodium atoms (^{23}Na) in 590 nm light, $\tau \gg 6\ \mu$s is the lower bound required for first-order Bragg scattering.

For all higher-order ($N_B > 1$) Bragg scattering processes, the nearest momentum states are $|g, \pm(N_B - 2)\hbar k\rangle$, which limits the interaction time to

$$\tau \gg \frac{\pi}{2(N_B - 1)\omega_{\text{rec}}} \qquad [44]$$

which for $N_B \gg 1$ reduces to

$$\tau \gg \frac{\pi}{2N_B\omega_{\text{rec}}} \qquad [45]$$

The upper bound on the interaction duration is necessary to avoid spontaneous emission. The interaction duration must be short enough so that the expected number, N_s, of spontaneous emission events per atom during the time τ is negligible. N_s is simply given by the product of the excited state fraction (eqn [36]) and the probability of spontaneous decay given that the atom is in the excited state:

$$N_s = |c_0(t)|^2\Gamma\tau = \frac{\omega_R^2}{4\Delta^2}\Gamma\tau \qquad [46]$$

where Γ is the natural decay rate of the excited state. Avoiding spontaneous emission ($N_s \ll 1$) while still having a significant probability for transitions ($\omega_R^{(2N_B)}\tau \simeq \pi$) forces one to work in the regime where $\Delta \gg \Gamma$, which is a practical requirement to avoid spontaneous emission and maintain coherence. For sodium atoms on the 590 nm transition, $\Gamma = 6.3 \times 10^7\ \text{s}^{-1}$, and a typical choice in Bragg diffraction experiments is $\Delta = 50\Gamma$, or $\Delta/2\pi = 500$ MHz.

Briefly, to relate Rabi frequency to optical intensity, the transition rate W on resonance with a Lorentzian spectral line is given by Fermi's golden rule:

$$W = \frac{\omega_R^2}{\Gamma} = \frac{\lambda_{\text{photon}}^2}{2\pi}\frac{I}{\hbar\omega} \qquad [47]$$

where the last term can be regarded as the cross section for absorption multiplied by intensity in units of photon flux. Hence

$$\omega_R^2 = \Gamma\frac{\lambda_{\text{photon}}^2}{2\pi}\frac{I}{\hbar\omega} \qquad [48]$$

For Na atoms in 12 mW(cm)$^{-2}$ resonant light, $\omega_R = \Gamma = 6.3 \times 10^7\ \text{s}^{-1}$. Both the probability of Bragg scattering and the spontaneous emission increase linearly with $\omega_R^2\tau\Gamma^{-1}$. However, spontaneous emission decreases as Δ^2 while the Bragg scattering probability decreases as Δ. Thus, as shown in Figure 10, at sufficient detuning and laser power, Bragg scattering occurs with no spontaneous emission. For reference, 12 mW(cm)$^{-2}$ × 10 μs corresponds to $\omega_R^2\tau\Gamma^{-1} = 630$ for sodium atoms.

Bragg scattering of atoms from a standing light wave was first observed at MIT in 1988. A supersonic atomic beam was diffracted from a standing wave of near-resonant laser light as depicted in Figure 6. The angle between the atomic beam (of thermal wavelength λ_{dB}) and the light grating (of periodicity $\lambda_L/2$) was tuned to the appropriate Bragg angle, θ_B, where:

$$\lambda_{dB} = \lambda_L \sin(\theta_B) \qquad [49]$$

Population transfer corresponding to both first- and second-order Bragg scattering was observed (Figure 7). The experiment required a sub-recoil transverse momentum spread of the atomic beam in order to resolve the different momentum states in the far field and limit the final state to only one

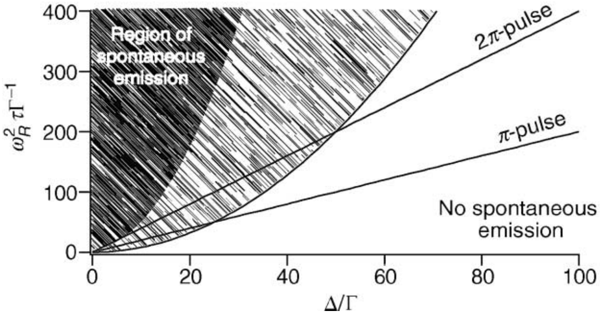

Figure 10 The conditions for which N_s >0.02(0.10) are hashed in light(dark) gray. ω_R^2/Γ is proportional to optical intensity. The line labeled π-pulse indicates the parameters which cause complete first-order Bragg diffraction. The line labeled 2π-pulse indicates conditions where population is completely transferred back to the initial state (Pendellösung oscillation).

diffracted order. The Pendellösung was observed as an oscillation in population transfer as a function of standing wave intensity, $I \propto \omega_R^{(2)}$, for a fixed interaction time, τ.

Atomic beam diffraction from an optical standing wave is a continuous-wave (CW) experiment in which the selectivity needed for the Bragg process is imposed by good angular resolution of the particle beam and a high degree of parallelism between the light crystal planes. This ensures that of the various final Bragg orders allowed by momentum conservation, only one conserves energy (energy conservation is exact in a CW experiment). For atoms scattering from a light crystal, parallelism of the crystal planes requires highly parallel photon momentum that implies a minimum width of the standing wave (the diffraction limit for the collimated photons). The transit time, τ, of the atoms across this width then exceeds the lower bound given in eqn [44].

The excellent collimation required of the atomic beam to ensure resolution of the Bragg scattered atoms reduces the intensity of the source by many orders of magnitude. A Bose–Einstein Condensate (BEC) is an attractive alternative source of atoms because its momentum spread is typically an order of magnitude below a single-photon recoil momentum. To Bragg diffract atoms initially in a stationary BEC, it is easier to move the light crystal than to accelerate the condensate. This is done by simply frequency shifting one of the traveling waves so that the resultant standing wave formed by its interference with the unshifted traveling wave moves with the proper velocity ($N_B \hbar k/m$) relative to the stationary atoms to impart the necessary momentum. The Bragg scattered atoms will then have momentum $2N_B \hbar k$ in the laboratory frame. The resonance condition thus becomes a condition on relative detuning, δ_{N_B}, between the two laser beams forming the diffraction grating. For N_B^{th}-order Bragg diffraction, the relative detuning is given by

$$\delta_{N_B} = \frac{2N_B \hbar k^2}{m} = 4N_B \omega_{\text{rec}} \qquad [50]$$

The first demonstration of Bragg scattering in a BEC was at NIST in 1999. They used Bragg scattering mainly as a tool to manipulate the momentum of the BEC, observing up to sixth-order processes. At MIT the interaction time was lengthened (to ≈100 times the lower bound of eqn [44] for first-order Bragg scattering), creating a new type of spectroscopy called Bragg Spectroscopy. It is a spectroscopic measurement of the shift of the Bragg scattering frequency due to the Doppler shift ($k \cdot v$) from the atoms' motion together with any mean field interaction. (These effects can be separated by going to higher-order Bragg scattering to enhance the Doppler shift.)

Bragg Spectroscopy was used to observe the momentum distribution of a BEC in a magnetic trap. The width of the Bragg resonance curve was primarily due to a 2 kHz Doppler-broadening that yielded the momentum distribution of the condensate. The spread in the corresponding velocity distribution was very small ($\approx 0.5\ \text{mm s}^{-1}$), even smaller than allowed by the Heisenberg uncertainty limit for a particle confined in the ground state of a harmonic trap. This reflects the increase in size of the condensate due to the mean field repulsion (^{23}Na, used in the MIT experiment, has a positive scattering length). The distribution was Heisenberg uncertainty limited at the observed size of the BEC, establishing for the first time that the coherence length of the condensate was equal to its size. In addition, the narrow Bragg resonance was shifted by the repulsive interactions within the condensate, resulting in a spectroscopic measurement of the mean-field energy.

Kapitza–Dirac Scattering

Diffraction of neutral atoms from a standing wave of near-resonant light with a short interaction time ($\tau_{KD} \ll 1/\omega_{\text{rec}}$) has come to be called Kapitza–Dirac scattering, in honor of their pioneering suggestion. In 1933, Kapitza and Dirac predicted that an electron beam propagating in a standing light wave would undergo stimulated Compton scattering and be reflected. This process has a tiny cross-section, equal to the classical electron radius squared: $\sigma_{\text{Compton}} = 8\pi/3(e^2/mc^2)^2 \approx 6 \times 10^{-25}\ \text{cm}^2$, and has only recently been observed using extremely high laser intensities. If the electrons are replaced by atoms, however, the scattering cross-section for resonant light $\sigma_{\text{atom}} = (1/2\pi)\lambda_{\text{photon}}^2 \approx 4 \times 10^{-10}\ \text{cm}^2$ is 15 orders of magnitude larger. This large cross-section, together with the ready availability of tunable lasers, has allowed stimulated scattering, both in the Kapitza–Dirac and Bragg regimes, to become the primary tool for the coherent manipulation of atoms.

In Kapitza–Dirac scattering, atomic motion during the interaction time is small compared to the characteristic dimensions of the interaction potential. This is equivalent to the eikonal approximation for scattering or the thin-lens approximation in optics. The idea is that the phase of the incident particle changes along each classical trajectory, but not the amplitude. Thus, there may be momentum transfer perpendicular to the trajectory, but the trajectory is not significantly displaced (until after the interaction is over). Mathematically, this regime can be treated by neglecting the atomic kinetic energy term in the Hamiltonian (the Raman–Nath approximation).

The standing wave interaction may be treated by considering the standing wave (ac-Stark shift) potential resulting from the applied fields given in eqn [18]:

$$V(z,t) = \frac{\hbar\omega_R^2}{\delta} f^2(t)\sin^2(kz) \qquad [51]$$

where we have assumed $\delta \gg \Gamma$, because the constraint $N_s \ll 1$ (eqn [46]) holds in this regime as well, where N_s is the expected number of spontaneous emission events per atom during the time τ. Although we neglect the kinetic energy term in the Hamiltonian, we continue to will treat z in eqn [51] as an operator.

Given the initial atomic wavefunction in momentum space, $\psi(p_0)$, the atomic wavefunction immediately after the interaction is given by

$$\psi(p) = e^{-\frac{i}{\hbar}\int^{t} dt' U(z,t')}\psi(p_0) \qquad [52]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau} e^{i\frac{\omega_R^2}{2\delta}\tau\cos(2kz)}\psi(p_0) \qquad [53]$$

where $\tau = \int dt' f^2(t')$ and the integral is over the interaction duration. With the use of the identity for Bessel functions of the first kind, $e^{i\alpha\cos(\beta)} = \sum_{n=-\infty}^{\infty} i^n J_n(\alpha)e^{in\beta}$, the atomic wavefunction can be written as

$$\psi(p) = e^{-i\frac{\omega_R^2}{2\delta}\tau}\sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right)e^{i2nkz}\psi(p_0) \qquad [54]$$

$$= e^{-i\frac{\omega_R^2}{2\delta}\tau}\sum_{n=-\infty}^{\infty} i^n J_n\left(\frac{\omega_R^2}{2\delta}\tau\right)\psi(p_0 + 2n\hbar k) \qquad [55]$$

States with $2N\hbar k$ of momentum are populated with the probability:

$$P_N = J_N^2(\theta), \quad N = 0, \pm 1, \pm 2, \ldots \qquad [56]$$

where

$$\theta = \frac{\omega_R^2}{2\delta}\tau = \omega_R^{(2)}\tau \qquad [57]$$

is the pulse area. This leads to a transverse rms momentum of the diffracted atoms that is linearly proportional to the pulse area:

$$p_{\text{rms}} = \sum_{n=-\infty}^{\infty} (n\hbar k)^2 P_n = 2^{1/2}\theta\hbar k \qquad [58]$$

The maximum and minimum bounds on the interaction time, τ, and the amount of momentum transfer to the atoms are discussed in the next section.

Kapitza–Dirac diffraction of atoms was first observed at MIT in 1986. Diffraction of a well-collimated (subrecoil) supersonic atomic beam was observed after passage through the tightly focused waist of a near-resonant standing wave (Figure 11). Significant diffraction into momentum states $|g, \pm 10\hbar k\rangle$ was observed (Figure 12).

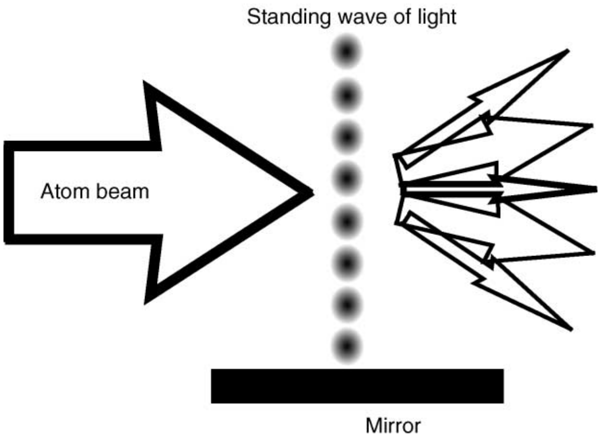

Figure 11 Two counter-propagating running waves of light superimpose to form a standing wave of light. The more narrow beam shown here represents the configuration for Kapitza–Dirac diffraction.

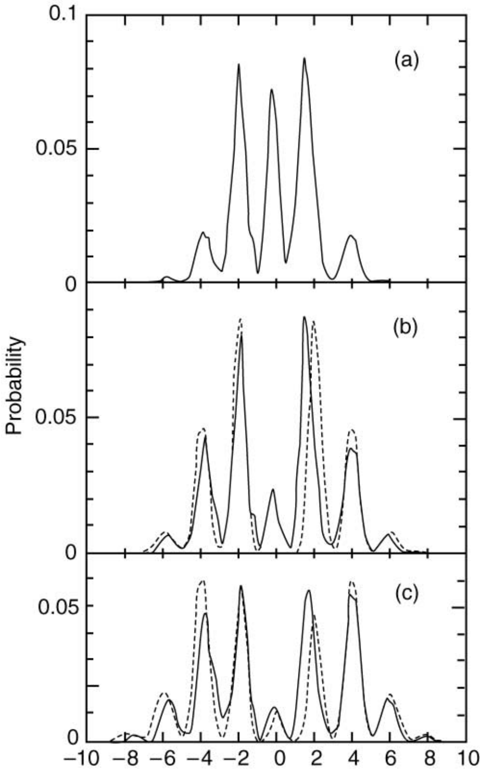

Figure 12 Kapitza–Dirac diffraction of an atomic beam from a standing wavelight grating. The solid lines are experimental data, the dashed lines are theoretical curves. (a) $\theta = 1.69$, (b) $\theta = 2.33$, (c) $\theta = 2.84$. Reprinted with permission from *Physical Review Letters* 56(8): 827–830. Copyright (1986) by the American Physical Society.

Even higher diffracted orders should be observable in the future using laser beams directed at small BECs for somewhat longer times.

Comparison of Atom-Standing Wave Interactions

The atomic motion induced by interaction with a standing light wave varies qualitatively with the interaction parameters. Figure 13 summarizes the various interaction regimes of an atom with a standing wave in terms of the two most basic parameters: the two-photon Rabi frequency, $\omega_R^{(2)} = \omega_R^2/2\delta$, and the duration of the interaction, τ. To render the plot independent of atomic species, $\omega_R^{(2)}$ is given in units of the single-photon recoil frequency, ω_{rec}, and τ is given in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. Thus, for ^{23}Na the point (1,1) on Figure 13 corresponds to (6.4 μs, $2\pi \times 25$ kHz). The product of the two-photon Rabi frequency and the duration of the interaction is the pulse area, $\theta = \omega_R^{(2)}\tau$ (eqn [57]), which is likewise the product of the pair of coordinates forming a point on the plot.

The lines KD1(KD10) on Figure 13 show where the first maxima of $P_{1,10} = J_{1,10}^2(\theta)$ (eqn [56]) occur, corresponding to the maximum possible population transfer into the first and tenth Kapitza–Dirac diffracted order respectively. Since the momentum distribution of diffracted atoms depends only on the pulse area (θ), all Kapitza–Dirac orders are parallel to each other and further offset from the origin in increasing order number.

The Kapitza–Dirac regime ends at large interaction times where the Raman–Nath approximation fails due to motion of the atoms down the slope of the standing wave potential. We show these lines dashed as the interaction time approaches the beginning of the classical oscillation regime, and terminate these lines where the oscillation produces its first focus, at the first focus line. This line corresponds to the atoms completing a quarter period of oscillation in the standing wave potential, $\tau = \tau_{\text{osc}}/4$, where the oscillation time is derived from the approximation that the potential $\langle V(z,t)\rangle_t$ (eqn [51]) is harmonic

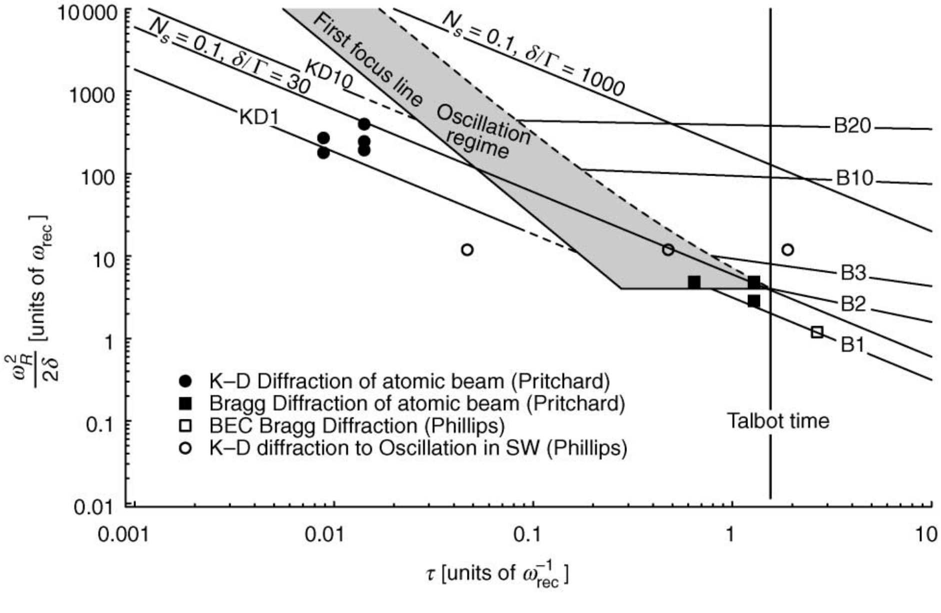

Figure 13 Atomic diffraction from a standing light wave. The vertical axis is the two-photon Rabi frequency, $\omega_R^2/2\delta$, in units of the single-photon recoil frequency, ω_{rec}, and the horizontal axis is the pulse duration, τ, in units of the inverse single-photon recoil frequency, ω_{rec}^{-1}. The scaling is chosen to eliminate the atomic species dependence of the plot. All coherent momentum transfer processes are destroyed by spontaneous decay, which occurs with probability $N_s = 0.1$ along the lines labeled accordingly and parameterized by the given ratio δ/Γ. Lines KD1 and KD10 show conditions for the maximum transfer into the first and tenth Kapitza–Dirac diffracted order respectively. As the interaction time is increased the Raman–Nath approximation is violated (termination of Kapitza–Dirac curves) and the atoms enter the oscillatory regime, executing at least a quarter period of oscillation above the first focus line (shaded area). Cures B1, B2, B3, B10, and B20 correspond to conditions that generate complete Bragg reflection in the first-, second-, third-, tenth-, and twentieth-order respectively. Experimental conditions are shown as points. Filled circles: Kapitza–Dirac diffraction of an atomic beam (Pritchard 1986). Filled squares: Bragg diffraction of an atomic beam (Pritchard 1988). Open squares: Bragg diffraction of a BEC (Phillips 1999). Open circles: transition from Kapitza-Dirac diffraction to oscillation of a BEC in a standing wave light pulse (Phillips 1999). Reprinted from Gupta S, *et al.* (2001) Coherent manipulation of atoms with standing light waves. *Comptes Rendus de L'Academie des Sciences – Series IV – Physics-Astrophysics* 2:479–495, copyright (2001) with permission from Elsevier.

about the minimum:

$$\tau_{\mathrm{osc}} = \frac{\pi}{\omega_R}\left(\frac{|\delta|}{\omega_{\mathrm{rec}}}\right)^{1/2} \qquad [59]$$

In the shaded region of Figure 13, the atoms will oscillate classically about the potential minimum, causing a periodic focusing of the atoms alternately in position and momentum space. However, anharmonicities of the potential away from its minimum will degrade the quality of the focusing effects.

The maximum diffracted order that can be significantly populated by the light–atom interaction is limited by energy conservation. Classically, the maximum momentum transfer to the atoms due to the sudden switch on of the standing wave is delivered to atoms that convert the full height of the standing wave potential ($\hbar\omega_R^2/\delta$), into kinetic energy. This gives

$$\frac{p_{\max}^2}{2m} \simeq \frac{\hbar\omega_R^2}{\delta} \Rightarrow p_{\max} \simeq \omega_R\left(\frac{2\hbar m}{\delta}\right)^{1/2} \qquad [60]$$

Equating the maximum absorbed momentum, $p_{\max}$, to an integer number of two-photon recoil momenta yields the maximum expected diffracted order, $N_{\max}$:

$$p_{\max} = 2N_{\max}\hbar k \Rightarrow N_{\max} \simeq \frac{\omega_R}{2}(\omega_{\mathrm{rec}}\delta)^{-1/2} \qquad [61]$$

This is just half of the square-root of the standing wave potential height measured in units of the single-photon recoil energy:

$$N_{\max} = \frac{1}{2}\left(\frac{\frac{\omega_R^2}{\delta}}{\omega_{\mathrm{rec}}}\right)^{1/2} = \frac{1}{2}\left(\frac{\omega_R^{(2)}}{\omega_{\mathrm{rec}}}\right)^{1/2} \qquad [62]$$

As the two-photon Rabi frequency is reduced, $N_{\max}$ falls below unity. As a result, there is no longer time for the higher-order multi-photon processes to generate significant amplitudes in diffracted orders with $N > 1$ before dephasing due to the kinetic energy term (eqn [24]) becomes significant and the higher-order processes become negligible. Only a small population is ever transferred to the $N = \pm 1$ orders and it does not oscillate at ω_{rec}. Therefore the classical oscillation regime does not extend below $N_{\max} \simeq 1$ where it is terminated on Figure 13.

Classical oscillation would result in atoms with momentum from zero up to the maximum allowed by energy conservation. Therefore, the classical focusing of atoms must be manifest as an atomic population distribution over many of the quantized momentum states (separated by $2\hbar k$) allowed by energy conservation (eqn [61]). However, as the interaction time lengthens and extends into the Bragg regime, the populated momentum states become restricted by energy conservation until only the oscillatory Pendellösung into and back from only one final state remains. Therefore, we have ended the classical oscillation regime where the Bragg condition (eqn [44]) is satisfied. The Bragg regime presupposes a smooth light pulse shape. For a pulse with sharp edges, classically oscillatory behavior can still be observed at longer times than included in the shaded region of Figure 13, which is why we show the large τ boundary of the classical oscillation regime as dashed.

In the Bragg regime, transfer of population is restricted to (and back from) only one final momentum state. The allowable final states are restricted by limiting the frequency bandwidth (i.e., energy uncertainty) of the light fields in the atomic rest frame. This is accomplished by lengthening the interaction time and smoothing the rise and fall of the electromagnetic fields. The parameters for a first-order Bragg transition (Table 1) are typically $\delta/\Gamma \simeq 50$, $\Gamma\tau \simeq 1$, and $\omega_R/\Gamma \simeq 10$ giving $N_s \simeq 10^{-2}$ and $\omega_R^{(2)}\tau \simeq 1$. Obtaining significant population transfer with higher-order Bragg processes requires larger intensities. Various orders (1, 2, 3, 10, and 20) of Bragg diffraction are shown as lines on Figure 13 corresponding to $\omega_R^{(2N_B)}\tau = \pi$, where $\omega_R^{(2N_B)}$ is given in eqn [40]. The lines are terminated at the appropriate interaction time determined from the final momentum state resolution condition (eqn [44]). The Bragg regime extends indefinitely to larger interaction times, which might be termed the region of Bragg spectroscopy. In experiments in this regime with ^{23}Na, atomic velocity resolution below 1 mm s^{-1} was obtained at $\tau = 80\omega_{\mathrm{rec}}^{-1}$. The study of adiabatically expanded BECs would require larger interaction times to resolve their smaller velocity spread and weaker mean field shifts.

The Kaptiza–Dirac and Bragg regimes assume a different initial atomic momentum (in the rest frame of the standing wave) parallel to the standing

Table 1 Natural parameters and typical experimental parameters involved in standing wave diffraction. The relevant frequencies together with the corresponding times (1/ω) are tabulated. The parameters for ^{23}Na have been used for system dependent quantities

ω	$\omega/2\pi$	$1/\omega$
ω_0	500 THz	0.3 fs
δ (Bragg)	500 MHz	0.3 ns
ω_R (Bragg)	100 MHz	1.6 ns
ω_R (K–D)	100 MHz	1.6 ns
Γ	10 MHz	16 ns
ω_{rec}	25 kHz	6.4 μs
ω_{osc}	1 MHz	160 ns

wave axis. Kapitza–Dirac scattering assumes no component of the initial atomic momentum along this axis. The efficiency of Kapitza–Dirac diffraction falls rapidly as the initial momentum in units of the photon momentum approaches $1/\omega_{\text{rec}}\tau$. To observe N_B^{th}-order Bragg scattering, the initial atomic momentum along the standing wave axis must have (nonzero) magnitude $N_B\hbar k$. Without adhering to this constraint, no final momentum state will be energetically degenerate with the initial state and the atomic sample will not respond to the presence of the standing wave, even if the interaction parameters are appropriate ($\omega_R^{(2N_B)}\tau \simeq 1$) for N_B^{th}-order Bragg scattering.

The condition $N_s = 0.1$ (eqn [46]), where N_s is the expected number of spontaneous emission events per atom during the light–atom interaction, is drawn on Figure 13 for two ratios (30 and 1000) of standing wave detuning, δ, to the atomic excited state natural lifetime, Γ. For fixed δ/Γ, we see that $N_s \propto \theta$, thus conditions with an increased pulse area (above and to the right of the $N_s = 0.1$ lines) result in a proportionally increased N_s. These $N_s = 0.1$ lines are extended across all regimes of Figure 13 because restricting spontaneous emission is required for both Kapitza–Dirac and Bragg scattering. As N_s approaches unity, the correct optical potential describing the light–atom interaction no longer is eqn [51], and scattering of atoms becomes an incoherent process which can lead to new possibilities such as a complex valued optical potential. Thus, while the standing wave fields in pure Bragg or Kapitza–Dirac scattering behaved as phase gratings for atoms, light gratings on resonance can effectively become amplitude diffraction gratings.

Atom Interferometry

In the 19th century, Fizeau (1853), Michelson (1881), Rayleigh (1881), and Fabry and Perot (Fabry 1899) exploited the interference properties of light waves to create the light interferometer which has since resulted in many beautiful experiments and precise measurements. Using technologies invented since the Second World War, the initial ideas of de Broglie and Schrödinger have evolved into construction of interferometers for neutrons, electrons, and atoms. These new interferometers are proving to be valuable tools for probing fundamental physics, for studying quantum mechanical phenomena, and for making inertial measurements.

The scientific value of interferometry with atoms and molecules has long been recognized. In fact, the concept of an atom interferometer was patented in 1973 and it has been extensively discussed since. Atom interferometry offers great richness stemming from the varied internal structure of atoms, the wide range of properties possessed by different atoms (e.g., mass, magnetic moment, absorption frequencies, and polarizability), and the great variety of interactions between atoms and their environment (e.g., static E–M fields, radiation, and other atoms).

Light interferometers are generally based either on achromatic beamsplitters such as half-silvered mirrors or on other semi-transparent membranes whose structure is small compared to the wavelength of the wave they are splitting. Lacking material structures that are either transparent to atoms or smaller than their de Broglie wavelength, diffraction gratings have been pressed into service both as beamsplitters and mirrors for atom waves. This means that atom interferometers are constrained to designs which somehow compensate for the dependence of diffraction angle on the wavelength of the individual atoms. In spite of this challenge, a surprising variety of atom and molecule interferometers have been built since 1991. A majority have used the three-grating configuration in which the first grating splits the incident beam, the second reverses the differential momenta given by the first, and the third recombines the two beams at the location where they overlap. Both material gratings and standing waves of light have been used in the Raman–Nath, Bragg, and adiabatic regimes to obtain interference fringes. Some designs render the interference fringes in position space, others in internal state space.

Figure 14 shows the MIT setup for a three-grating atom interferometer which features sufficient separation between interfering paths to accommodate an interposed metal foil. This allows different fields or media to act on atom waves only in one of the two

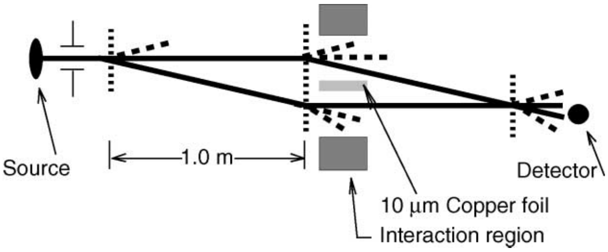

Figure 14 An interferometer for atoms built with three nanofabricated gratings. Each atom propagates in a superposition of two paths, shown in bold. The beam of atoms is well enough collimated, and the diffraction angle is large enough, that a metal barrier can be inserted between the two paths inside the interferometer. The interference fringes are observed by translating one grating transverse to the atom beam. See for example: Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

arms of the interferometer. Such an interaction region has been used to measure phase shifts due to an electric field applied to only one of the interfering paths. The phase difference between the two paths is then given by the Feynman path integral:

$$\phi = \frac{-i}{\hbar}\int V(x)dt \qquad [63]$$

where

$$V = -\frac{1}{2}\alpha E^2 \qquad [64]$$

is the potential energy due to the electric filed E, and α is the electric polarizability of the atoms. The resulting phase shift of the interference fringes (Figure 15) has been used to measure the polarizability of sodium atoms to unprecedented accuracy.

Phase shifts have also been measured due a dilute target gas in one arm of the interferometer. In this case there is a complex index of refraction for the forward scattered atoms (some attenuation, and some phase shift) which depends on matter wave wavelength. Recent measurements reveal glory oscillations in the matter wave index of refraction due to passing through a dilute gas, and are a sensitive probe of interatomic potentials.

In the domain of inertial sensing, atom interferometers have been used to the local gravitational acceleration, $\vec{g}$, Newton's constant, G, gravity gradients, $\nabla\vec{g}$, and rotations Ω, each with sensitivity rivaling if not exceeding any other method. The Sagnac phase shift, due to rotations, for a two-path interferometer is

$$\phi_S = \frac{4\pi}{\lambda_{dB}v}\vec{\Omega}\cdot\vec{A} \qquad [65]$$

where v is the velocity of atoms, $\vec{A}$ is the enclosed area of the interferometer, and $\vec{\Omega}$ is the rotation rate. The two factors in the denominator, λ_{dB} and v, are both $\sim 10^5$ times larger for light than for atoms, which is why atom interferometers have the potential to be 10^{11} times more accurate than optical interferometers. In practice, atom interferometers only outperform laser gyros by a small amount because they have dramatically smaller enclosed areas and particle flux than optical interferometers.

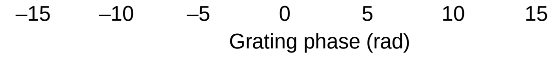

Figure 15 Interference fringes from atoms. 2π of grating phase corresponds to translating the grating one grating period, or 100 nm. These data were obtained at MIT.

Since the first atom interferometers for Na and He* began working in 1991, others have been made for Li, Ar*, Ca, Cs, K, Mg, Ne*, Rb, and Li_2, Na_2, I_2, C_{60} and C_{70} molecules, interferometers starting with trapped atoms have been made for Cs, Ca, He*, Mg, and Rb. Interferometers using Bose–Einstein condensates have been demonstrated with Na and Rb. These lists are still growing. Diffraction from a single grating has been observed for molecules as large as C_{70} buckyballs, which may someday be used in a three-grating interferometer.

Conclusion

The manipulation of atoms using light forces from standing light waves is a rich subject. The seminal suggestion of Kapitza and Dirac laid dormant for 50 years due to lack of experimental technology. However, in the 1980s this suggestion was realized with atomic sources and considerably extended both experimentally and theoretically. In the early 1990s, coherent standing wave manipulation and nonofabricated atom optics became the two major routes to making atom interferometers. As the new century begins, these techniques are being refined further with improved sources, new detection schemes, and better atom optics. Replacing thermal atomic beams by Bose–Einstein condensates (BECs), as sources of atoms, will revolutionize the field of atom optics just as lasers did in light optics.

Conversely, stimulated light scattering and nanostructures provide new ways to study atomic and molecular properties, such as Bragg spectroscopy for BEC characterization. As progress continues, exciting developments should be forthcoming, such as definitive measurements of the fine structure constant, α, atom gyroscopes that are far superior to the best laser gyroscope technology, and improved studies of BECs. Another area that is on the brink of spectacular development is the confinement of coherent matter waves in atom waveguides and the development of the scientific and technological opportunities that these represent, in analogy to fiber optic waveguides for light.

While we have described the application of nanofabrication techniques to atom optics, it is possible to imagine technology transfer in the other direction. The fundamental problem of fabricating ever smaller structures might be tremendously advanced by atom optics because the de Broglie wavelength of atoms is so much smaller than that of light, permitting them to be focused to directly deposit much smaller features than possible with photolithography.

See also

Interferometry: Overview. **Scattering:** Scattering Theory.

Further Reading

Cohen-Tannoudji C, Diu B and Laloe F (1977) *Quantum Mechanics*. New York: Wiley and Sons.

Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.

Berman PR (ed.) (1997) *Atom Interferometry*. San Diego, CA: Academic Press.

Mandel L and Wolf EM (1995) *Optical Coherence and Quantum Optics*. London: Cambridge University Press.

Balykin VI and Letokhov VS (1995) *Atom Optics with Laser Light*. Chur, Switzerland: Harwood Academic Publishers.

Meystre P and Sargent M III (1999) *Elements of Quantum Optics*. Heidelberg, Germany: Springer Verlag.

Rauch H (2000) *Neutron Interferometry*. Oxford, UK: Oxford University Press.

Meystre P (2001) *Atom Optics*. New York: Springer-Verlag.

Campargue R (ed.) (2001) *Atomic and Molecular Beams, the State of the Art 2000*. Berlin: Springer.

Atomic Coherence Effects

A Belyanin, G R Welch and M O Scully,
Texas A&M University, College Station, TX, USA

© 2005, Elsevier Ltd. All Rights Reserved.

History and Perspective

Quantum optics and laser physics have entered a new era. This started with the invention of the laser and the associated theoretical developments such as the theory of photon correlations and statistics, and the quantum theory of the laser. Later, new and novel lasers such as the free electron laser and tunable dye laser fueled many new research directions. With the development of the micromaser, quantum optics became a beautiful new test bed for the study of nonequilibrium statistical physics of photons interacting with matter in a controlled environment. The realization that a deeper understanding was needed to understand the measurement process led to the utilization of the laser as a probe of gravity, and to insight into nonclassical processes such as quantum nondemolition measurements and the production of squeezed states of light.

Effects of atomic coherence in quantum electronics have a long history. One of the first examples was the Hanle effect: excitation by polarized light creates atomic coherence that can evolve in a magnetic field, causing the change of polarization of the resulting fluorescence. Another important example is the interference of decay processes, first suggested by Fano. Modern schemes utilizing quantum coherence in atomic vapors or solid media usually involve strong coupling laser fields (laser drives). In each case, the effect of quantum coherence can be understood as interference of multiple (usually two) transition paths between quantum states of a system.

In the past decade, the concepts of atomic coherence and interference have led to many new and surprising innovations in laser physics such as electromagnetically induced transparency (EIT), refractive index enhancement, lasers without population inversion, ultra-slow light group velocity, storage of the quantum states of light, teleportation of quantum states, and nonlinear optics at an ultra-low intensity level.

Electromagnetically Induced Transparency

One of the most vivid manifestations of quantum coherence effects is the reduction of absorption in a three-level system that is driven by a coherent laser field. This effect is known as electromagnetically induced transparency, or EIT. We consider in more detail the case when the probe and drive fields are in a so-called Λ-configuration (Figure 1a) named after the appearance of the level scheme. In the conventional usage, the two lower levels are coupled to the upper level with two lasers. One of the lasers may be strong and is called the driving or coupling laser, and the other is weak and is called the probe laser.

Consider a quasi-monochromatic laser field E_d (called the 'driving' field, or 'coupling' field) and probe field E_p interacting with a three-level medium of Figure 1a. This interaction is described by a density matrix ρ_{ij}, where i, j take the values a, b, c. Within the

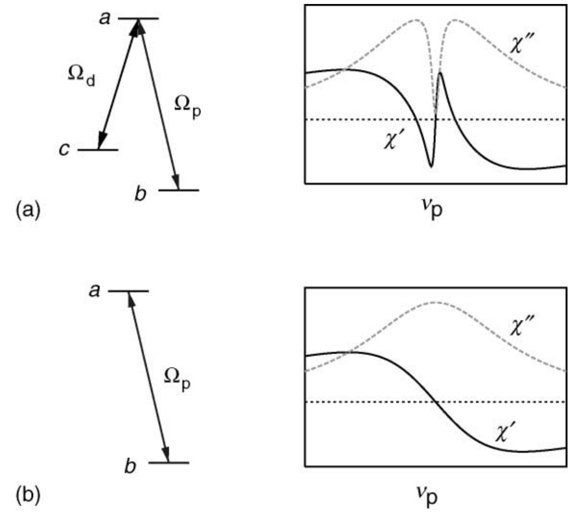

Figure 1 Real and imaginary parts of the susceptibility as functions of the frequency of the probe laser for (a) three-level Λ–type atoms, and (b) two-level atoms.

density matrix approach, quantum coherence in the system has the precise meaning of an off-diagonal element of the density matrix at the two-photon transition c–b. To have strong quantum coherence in a Λ-system, one must excite a large value of ρ_{cb}.

It is convenient to simplify the equations from the very beginning by using the rotating-wave approximation. We introduce slowly varying complex amplitudes of fields and polarizations and neglect their second derivatives. For example, for the probe field applied at the $a \rightarrow b$ transition we write

$$E_{\rm p}(z,t) = \frac{1}{2}\varepsilon_{\rm p}(z,t)\exp(-{\rm i}\omega_{\rm p}t + {\rm i}k_{\rm p}z) + {\rm c.c.} \qquad [1]$$

After introducing the complex Rabi frequency $\Omega_{\rm p}(z,t) = d_{\rm p}\varepsilon_{\rm p}(z,t)/2\hbar$, the wave equation for waves propagating in the z-direction can be written as

$$\frac{\partial\Omega_{\rm p}}{\partial z} + \frac{n_{\rm p}}{c}\frac{\partial\Omega_{\rm p}}{\partial t} + \kappa_{\rm p}\Omega_{\rm p} = \frac{2\pi{\rm i}\nu_{\rm p}d_{\rm p}^2N}{\hbar c n_{\rm p}^2}\sigma_{ab} \qquad [2]$$

Here $d_{\rm p}$ is the dipole moment of the probe transition a–b, N the total volume density of atoms in the active region, $\kappa_{\rm p}$ represents cavity losses, and $n_{\rm p} = k_{\rm p}c/\omega_{\rm p}$ is the refraction index. The variable σ_{ab} is the slowly varying amplitude of the corresponding off-diagonal element $\rho_{ab} = \sigma_{ab}\exp(-{\rm i}\nu_{\rm p}t)$ of the density matrix. The same representation is assumed for the drive field $E_{\rm d}$, with parameters n, d, ν, κ having index 'd'.

The functions σ_{ab}, σ_{ac}, and σ_{cb}, which determine the polarization response of the medium at the frequency of the external field, should be found from the density matrix equations

$$\frac{{\rm d}\sigma_{cb}}{{\rm d}t} + \Gamma_{cb}\sigma_{cb} = -{\rm i}\Omega_{\rm p}\sigma_{ac}^* + {\rm i}\Omega_{\rm d}^*\sigma_{ab} \qquad [3]$$

$$\frac{{\rm d}\sigma_{ab}}{{\rm d}t} + \Gamma_{ab}\sigma_{ab} = {\rm i}\Omega_{\rm p}n_{ba} + {\rm i}\Omega_{\rm d}\sigma_{cb} \qquad [4]$$

$$\frac{{\rm d}\sigma_{ac}}{{\rm d}t} + \Gamma_{ac}\sigma_{ac} = {\rm i}\Omega_{\rm d}n_{ca} + {\rm i}\Omega_{\rm p}\sigma_{cb}^* \qquad [5]$$

where

$$\begin{aligned}\Gamma_{cb} &= \gamma_{cb} + {\rm i}(\nu_{cb} - \nu_{\rm p} + \nu_{\rm d}),\\ \Gamma_{ab} &= \gamma_{ab} + {\rm i}(\nu_{ab} - \nu_{\rm p}),\\ \Gamma_{ac} &= \gamma_{ac} + {\rm i}(\nu_{ac} - \nu_{\rm d})\end{aligned} \qquad [6]$$

Here the γ's are relaxation rates of the polarizations at the corresponding transitions (the so-called homogeneous broadening). The population differences $n_{ik} = \rho_{ii} - \rho_{kk}$ are defined from eqns [2]–[6] together with the three equations for $\rho_{ii}, i = a, b, c$, with phenomenological rates of population relaxation and pumping:

$$\begin{aligned}\frac{{\rm d}\rho_{bb}}{{\rm d}t} &= -2{\rm Im}[\Omega_{\rm p}^*\sigma_{ab}] + r_{cb}\rho_{cc} + r_{ab}\rho_{aa} - r_{ba}\rho_{bb}\\ \frac{{\rm d}\rho_{cc}}{{\rm d}t} &= -2{\rm Im}[\Omega_{\rm d}^*\sigma_{ac}] + r_{ac}\rho_{aa} - r_{cb}\rho_{cc}\\ \frac{{\rm d}\rho_{aa}}{{\rm d}t} &= 2{\rm Im}[\Omega_{\rm p}^*\sigma_{ab}] + 2{\rm Im}[\Omega_{\rm d}^*\sigma_{ac}] + r_{ba}\rho_{bb}\\ &\quad - (r_{ab} + r_{ac})\rho_{aa}\end{aligned} \qquad [7]$$

Here we have included only the pumping term r_{ba} between levels b and a for simplicity. Later, we will consider the effect of incoherent pumping on gain, but in this section we put $r_{ba} = 0$. The sum of populations is evidently conserved in this model. Analysis of open systems with, e.g., flow of atoms through a cavity, leads to qualitatively similar conclusions.

Suppose that the probe field is weak, $|\Omega_{\rm p}|^2 \ll \gamma_{cb}\gamma_{ac}$, and the drive-field intensity is constant in a sample. The complex susceptibility at the probe-field frequency, which is proportional to $\sigma_{ab}/\varepsilon_{\rm p}$, is given by

$$\chi = \frac{{\rm i}d_{\rm p}^2N}{\hbar c}\left\{\frac{1}{\Gamma_{ab} + |\Omega_{\rm d}|^2/\Gamma_{cb}}\left(n_{ba} - \frac{|\Omega_{\rm d}|^2 n_{ca}}{\Gamma_{cb}\Gamma_{ac}^*}\right)\right\} \qquad [8]$$

This expression can be further simplified if there is no pumping to the states $|c\rangle$ and $|a\rangle$, so that we can assume that $n_{ca} = 0$ and drop the second term in the brackets. The behavior of χ as a function of detuning from resonance $\nu_p = \nu_{ab}$ is shown in Figure 1a. In the absence of the drive field Ω_d, we retrieve the usual two-level susceptibility, shown in Figure 1b.

There are two remarkable features in the susceptibility of Figure 1a. The first one is the transparency window at the point of two-photon resonance. The absorption coefficient at resonance is

$$K'' = \frac{2\pi\nu_p d_p^2 N}{\hbar c(\gamma_{ab} + |\Omega_d|^2/\gamma_{cb})} \qquad [9]$$

Clearly, absorption can be arbitrarily small when $\gamma_{cb} \ll \gamma_{ab}$ and $|\Omega_d|^2 \gg \gamma_{cb}\gamma_{ab}$, since it is proportional to $\gamma_{cb}/|\Omega_d|^2$. For any value of $|\Omega_d|^2$, the absorption is much smaller than that resulting from simple Stark splitting of two Lorentzians. Moreover, in the most interesting case

$$\gamma_{cb}\gamma_{ab} \ll |\Omega_d|^2 \ll \gamma_{ab}^2$$

the Stark splitting is negligible while the deep and narrow EIT feature remains, with width $\gamma_{cb} + |\Omega_d|^2/\gamma_{ab}$.

Note that the intensity of the drive field required for the coherent bleaching of the medium is a factor $r_{ab}/\gamma_{cb} \gg 1$ smaller than the intensity needed for usual saturation bleaching.

EIT is closely related to the phenomenon of coherent population trapping (CPT). This can be illustrated by transforming from states $|a\rangle$, $|b\rangle$, $|c\rangle$ to a new set of states $|a\rangle$, $|B\rangle$, $|D\rangle$, where

$$|D\rangle = (\Omega_d|b\rangle - \Omega_p|c\rangle)/\sqrt{|\Omega_d|^2 + |\Omega_p|^2}$$

$$|B\rangle = (\Omega_d|b\rangle + \Omega_p|c\rangle)/\sqrt{|\Omega_d|^2 + |\Omega_p|^2}$$

It can be easily found that under the condition of two-photon resonance $\nu_p - \nu_d = \nu_{cb}$ the dipole matrix element from $|D\rangle$ to $|a\rangle$ goes to zero. This means that the population in state $|D\rangle$ does not interact with the coherent radiation field and does not contribute to absorption. That is why this state is called a non-absorbing or dark state. Under the condition $|\Omega_d|^2/\gamma_{cb}\gamma_{ab} \gg 1$ practically all the population of states $|b\rangle$ and $|c\rangle$ is optically pumped into the dark state, and only a small fraction $\propto\gamma_{cb}$ remains in the absorbing state $|A\rangle$ and contributes to absorption.

The second remarkable feature of a driven Λ-system is sharp variation of the real part of χ near the resonance, which can be used for manipulation of the group velocity of light, as discussed in the next section.

Slow, Ultra-Slow, Stored, and Frozen Light

Introduction

There are five useful definitions for the velocity of a traveling wave:

- the phase velocity, which is the speed at which the zero crossings of the carrier wave move;
- the group velocity, at which the peak of a wave packet moves;
- the energy velocity, at which energy is transported by the wave;
- the signal velocity, at which the half-maximum wave amplitude moves;
- the front velocity, at which the first appearance of a discontinuity moves.

Although these can all be different, in most cases (linear, passive, dispersive media) the last four coincide, and are usually less than the phase velocity. By using the very steep frequency dispersion in the vicinity of the narrow resonance associated with electromagnetically induced transparency (EIT), it has been shown that the group velocity of light can be reduced by approximately 100 million compared with its velocity in vacuum.

In order to understand the concept of group velocity, it is useful to consider the superposition of two waves of the same amplitude E_1 and E_2 where $E_i = E_0\cos(k_i z - \nu_i t)$ with $i = 1, 2$. Addition of these waves gives rise to the modulation shown in Figure 2:

$$E = E_0(\cos(k_1 z - \nu_1 t) + \cos(k_2 z - \nu_2 t))$$
$$= 2E_0\cos(\Delta k z - \Delta\nu t)\cos(kz - \nu t)$$

where $\Delta k = (k_1 - k_2)/2$, $\Delta\nu = (\nu_1 - \nu_2)/2$, $k = (k_1 + k_2)/2$, and $\nu = (\nu_1 + \nu_2)/2$. This superposition

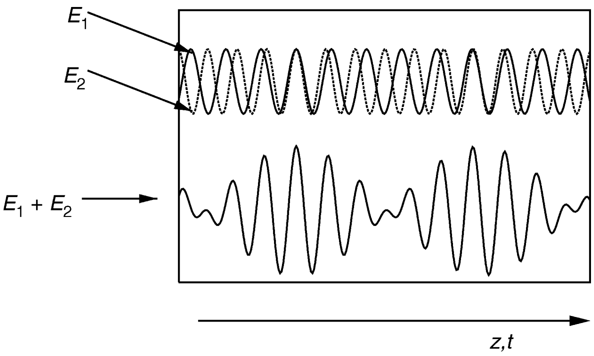

Figure 2 Interference of two monochromatic waves with different frequencies results in a wave modulated in time and space.

displays an interference pattern consisting of a rapid oscillation propagating with the so-called phase velocity,

$$v_{\text{phase}} = \frac{\nu}{k} \qquad [10]$$

and a slowly varying envelope propagating at the group velocity,

$$v_{\text{g}} = \frac{\Delta\nu}{\Delta k} \qquad [11]$$

Usually, the group velocity of wave packets consisting of more than two harmonics is given by transforming the ratio $\Delta\nu/\Delta k$ into $\mathrm{d}\nu/\mathrm{d}k$.

Slow and Ultra-Slow Light

In ordinary absorption and refraction of light, it is sufficient to consider only two atomic energy levels interacting with the light. However, in multilevel systems, coherent excitation of more than one transition can dramatically alter the dispersion and hence the group velocity. The first experiments of this effect demonstrated a group velocity of *c*/13. However, this was progressively lowered in later experiments as the steepness of the dispersion was increased, culminating in a group velocity as low as *c*/300 000 000 (1 m/s) in a Bose–Einstein condensate of ultra-cold sodium atoms. Slow light has been observed in a wide range of systems, from the Bose–Einstein condensates just mentioned, to hot optically dense gases of atoms and molecules, and in solids ranging from rare-earth doped crystals to solid hydrogen.

Most of these experiments occurred in a Λ system. The susceptibility of this medium has very steep dispersion without absorption, as shown in Figure 1.

The minimum possible group velocity slowed by the steep dispersion in EIT is given by

$$v_{\text{g}} = \frac{8\pi\Omega^2}{3\lambda^2 N\gamma} > 0 \qquad [12]$$

where $\gamma = \gamma_{ab} = \gamma_{ac}$. The driving field Ω should be strong enough to meet the condition $\Omega^2 \gg \gamma\gamma_{bc}$ for coherent effects. It can be shown that the group velocity approaches its minimum $v_{\text{g min}} = 8\pi\gamma_{bc}/(3\lambda^2 N)$ when $\Omega^2 \approx \gamma\gamma_{bc}$.

Bringing Light to a Halt: Frozen Light

Consider a laser pulse propagating through a medium with ultra-slow group velocity while the medium itself is moving with some velocity v in the opposite direction to the direction of laser pulse propagation. Let us denote the group velocity of the light in the frame co-moving with the atoms as $\tilde{v}_{\text{g}}$. In this frame, atoms are at rest, and hence, there is no spatial dispersion. The Galilean transformation to the laboratory frame, $k = \tilde{k}$, $\nu = \tilde{\nu} - \tilde{k}v$, where v is the atomic velocity, yields the group velocity $v_{\text{g}} = Re(\mathrm{d}\nu/\mathrm{d}k) = \tilde{v}_{\text{g}} - v$. This simple transformation shows that the light pulse is 'dragged' by the moving atoms. If the velocity of the atoms with respect to the laboratory frame is exactly equal and opposite to the group velocity of light in the frame where atoms are at rest, then we find that the group velocity in the laboratory frame is zero. We refer to this as 'freezing' of light. We may even make the group velocity negative so that the propagation direction is opposite to the wavevector (that is, that $v_{\text{g}} < 0$). The simplest example of spatial dispersion is the so-called drift dispersion corresponding to a mono-velocity atomic beam or moving sample with velocity v.

Slow light experiments in Doppler broadened systems such as warm gases allow group velocities that are much lower than the mean thermal speed (ν_{T}) of the atoms in the thermal sample (Figure 3). This means that freezing of light can occur without actually moving the atomic sample. When the experimentally obtained group velocity is less than or about equal to the mean thermal speed of the atoms, a pulse of light can be stopped in a stationary cell. In this case the delay time for a pulse passing through the medium would tend to infinity (contrary to the case of a finite atomic beam) in the sense that this pulse would never leave the cell.

This is accomplished by using a single velocity group from the Maxwellian thermal distribution of atomic velocities, and adjusting the frequency of the driving field to be resonant with that velocity group (see Figure 4). In this case it is mainly the atoms in this single velocity group that support the ultra-slow group velocity as a slow EIT polariton, and the atoms

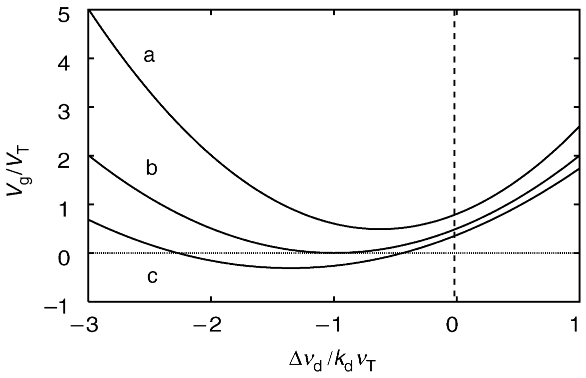

Figure 3 Ultra-slow and negative group velocity of EIT polariton versus detuning of drive laser; $\Omega = 0.25\gamma$, $k_{\text{d}}\nu_{\text{T}} = 100\gamma$, $\gamma_{cb} = 0.001\gamma$, (a) $N = 0.6\,N_{\text{cr}}$; (b) $N = N_{\text{cr}}$; (c) $N = 1.5\,N_{\text{cr}}$.

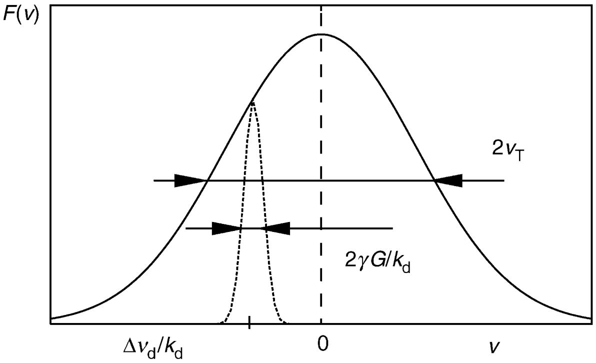

Figure 4 The velocity distribution of atoms in a cell (solid line). Effective drifting beam (dotted) selected by drive laser.

then act as an effective atomic beam. If the driving field is on resonance with the velocity group of atoms moving into the light with speed equal to $\tilde{v}$, we will have $v_g = \tilde{v}'_g - v_d$, where $\tilde{v}'_g$ is the group velocity of the light defined by the effective atom beam if those atoms were at rest. This requires that the intensity of the drive is strong enough to provide EIT for the resonant velocity group of atoms, i.e., $\Omega^2 > \gamma_{bc}\gamma$ (see Figure 4) but at the same time weak enough to avoid interaction with off-resonant atoms moving with the 'wrong' velocities (i.e. $\Omega < kv_T\sqrt{\gamma_{bc}/\gamma}$).

Storing and Retrieving Quantum Information

Photons are potentially the best particles for use in quantum computing and quantum information processing. They are the fastest, and perhaps the simplest and most robust carriers of quantum information, but their storage and manipulation is difficult without destroying their quantum state. However, for application to quantum cryptography, quantum teleportation, and quantum computation it is necessary to achieve a strong coupling between electromagnetic waves and a large nonlinear susceptibility with small losses. For this task, it may be resonant coherent media that offer the promise of success.

Recently, new coherent and reversible methods based on EIT have been found for the transfer and manipulation of photons and their quantum information. Quantum states of photons are trapped in coherently driven atomic media in which the group velocity is reduced adiabatically toward zero. The methods allow for an ideal transfer of quantum correlations between light fields and metastable states of matter. Currently, interesting applications such as quantum state memories and quantum information processing including transporting and time reversing the state of light are being investigated.

Although experiments have not yet been conducted with single quanta, proof-of-principle experiments using classical laser fields have demonstrated the technique. In these experiments control and signal light pulses propagate in a gas of three-level Λ-type atoms and excite a spatial profile of a long-lived coherence of ground spin states. This spin coherence profile contains quantum information about the 'writing' pulses. When the control field is switched back on, the signal pulse is restored via Raman scattering on the atomic coherence. Ideally, the scattered pulse is in the same quantum state as the input signal pulse, preserving entanglement, quantum statistics, etc.

Similar experiments have also shown that different reading pulses (rather than switching back on the same control pulse) can be used to obtain new tools for quantum information storage and processing. Time reversing, teleportation, color switch and multiplexing of the signal pulse of light have all been demonstrated. These new operational possibilities arise as a result of the reading of the stored information with a light pulse that is spatially separated from, has different frequency from, and propagates in the opposite direction to the signal beam.

To demonstrate these effects, one pair of control and signal pulses (the 'writing' fields) prepares a spin coherence of the lower level (Figure 5). A second control pulse (which can be thought of as a 'reading' field) scatters from this coherence resulting in the generation of a new, or 'recovered', field. This recovered field acquires some properties of the reading field because it is a Raman scattered component of the reading pulse. If the reading field pulse is centered about a frequency other than that of the writing fields, and propagates in the opposite direction to the writing fields, then the recovered field propagates in the same direction as the reading field and has a different frequency from the incident writing fields. Even if the writing and reading pulses are separated in space as well as time the scattering effect will persist. This allows the transportation of the coherence grating to another point in the atomic cell via the atomic center-of-mass motion.

In the ideal case the signal pulse profile and quantum statistics are exactly stored by the atomic coherence grating. Thus, reading by a backward propagating pulse results in a generated pulse that is also backward propagating. That is it represents an exactly time-reversed copy of the signal pulse. Similarly, atomic motion in the direction transverse to the light propagation moves the coherence grating to another spatial position. Reading the light by a spatially shifted laser results in the equivalent of

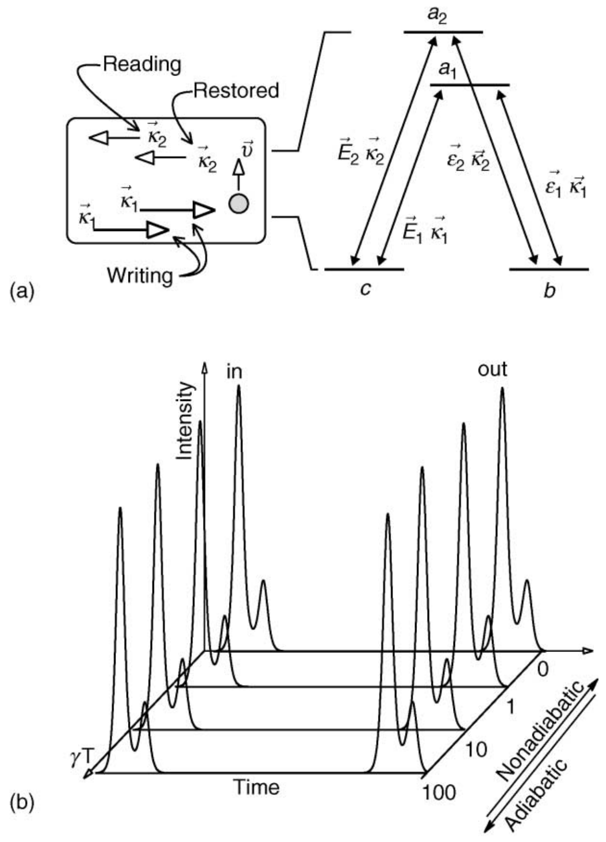

Figure 5 (a) Energy level scheme used in experiments for multiplexing, transporting, and time-reversing stored light. (b) 'Storage of light': intensity of the signal field versus time for different switching times (T) of the reading and writing fields.

'teleportation' of light by atoms in the same spirit as the teleportation of the quantum state of light by atoms that has been discussed recently. It is also possible for the frequency of the 'restored' pulse to be easily shifted by switching the frequency of the reading pulse. This means that the use of a frequency comb of reading pulses propagating in different spatial modes should result in multiplexing of the signal light.

Novel Laser Sources Based on Quantum Coherence

Introduction

In the previous sections we have seen how quantum coherence can be used to overcome resonant absorption on a transition of interest, leading to EIT. A natural question arises whether it is possible to achieve gain without population inversion at this transition, i.e. lasing without inversion (LWI). This subject has been intensively investigated since the early 1990s, and the first proof-of-principle experiments have been successful.

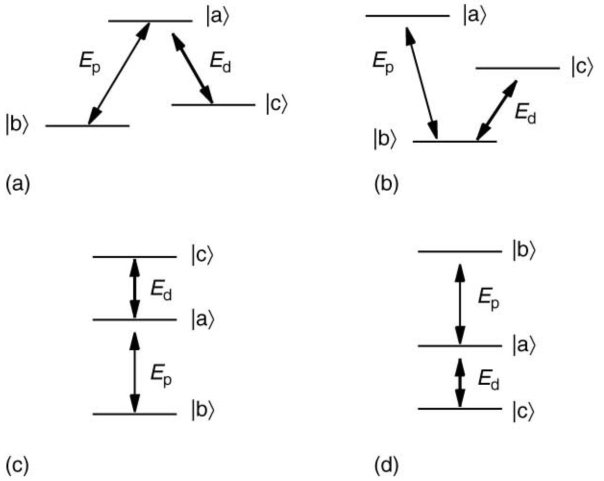

Figure 6 LWI schemes in a three-level system coherently driven by a laser field.

Besides the fundamental physical beauty of this phenomenon, the research effort has been motivated by potentially important applications. It is clear that LWI could be particularly useful in situations where a population inversion is difficult to maintain due to, e.g., very fast relaxation or inefficient pumping. The obvious examples are high-frequency UV, X-ray and gamma-ray lasers, or semiconductor mid/far-infrared lasers based on intersub-band transitions. Another promising application of LWI is efficient frequency up- or down-conversion. Furthermore, if resonant absorption can be eliminated, resonant nonlinear optical interactions become a reality. Since all nonlinear coefficients are maximized at resonance, highly efficient nonlinear optical generators of light can be created.

In this section we discuss the simplest possible schemes of inversionless lasing which allow one to clarify the physical origin of LWI and to see the shortcomings which have so far prevented LWI-based schemes from reaching the level of practical laser devices.

Consider a three-level system, coherently driven by a strong quasi-monochromatic field nearly resonant with one transition, and probed by a field that is nearly resonant with an adjacent transition. Some possible configurations are shown in Figure 6. We consider in detail the case when the two fields are in a Λ-configuration (Figure 6a). This is probably the most popular and well-studied scheme since it can provide probe-field gain for a very low pumping rate, when most of the population is at the lowest energy level. However, we will see that the gain condition for the Λ-scheme is actually rather restrictive and does not allow one to realize large frequency up-conversion. For the latter purpose, ladder-type schemes (Figure 6c,d) turn out to be more favorable.

Weak-field gain in the Λ-scheme

From the general formalism presented earlier, it is straightforward to get the amplification condition

$$K'' = \mathrm{Re}\left\{ \eta_{\mathrm{p}} \frac{1}{\Gamma_{ab} + |\Omega_{\mathrm{d}}|^2/\Gamma_{cb}} \left(n_{ba} - \frac{|\Omega_{\mathrm{d}}|^2 n_{ca}}{\Gamma_{cb}\Gamma^*_{ac}} \right) \right\} + \kappa_{\mathrm{p}} < 0 \qquad [13]$$

where

$$\eta_{\mathrm{p}} = \frac{2\pi\omega_{\mathrm{p}} d_{\mathrm{p}}^2 N}{\hbar c n_{\mathrm{p}}^2}$$

The physical meaning of the two terms in brackets in eqn [13], which originate from the two terms on the right-hand side of eqn [4] for σ_{ab}, can be easily understood. The first term is just the resonant one-photon absorption proportional to the population difference n_{ba}. It can be strongly reduced in the presence of a drive field Ω_{d}, as explained earlier. However, it can change sign and give rise to the field amplification only if $n_{ba} < 0$, meaning there is a population inversion!

The second term, originating from the product $\Omega_{\mathrm{d}}\sigma_{cb}$ in eqn [4], is due to the mixing of the drive field Ω_{d} and polarization σ_{cb} where the latter is parametrically excited by a two-photon process: $\sigma_{cb} \propto e_{\mathrm{p}}\sigma^*_{ac} \propto e_{\mathrm{p}}\Omega^*_{\mathrm{d}} n_{ca}$. It is this term that can overcome one-photon absorption to provide amplification without inversion in the Λ-scheme. As we see, the possibility of LWI depends crucially on our ability to excite a sufficiently large polarization σ_{cb} at the transition resonant with the beat frequency of the probe and drive fields. This conclusion is universal for all LWI schemes with a coherent drive. We see that the sometimes loosely defined term 'quantum coherence' has a precise meaning here as the polarization σ_{cb} at a two-photon transition c–b. It is clear that the effects of quantum coherence are maximized when γ_{cb} is small, i.e., the transition c–b is long-lived.

We are of course interested in the case of weak incoherent pumping. The typical dependence of gain on detuning from two-photon resonance is shown in Figure 7. At resonance, where the gain is maximized, the amplification condition $K'' < 0$ yields (in the absence of nonresonant losses):

$$r_{ac} > 2\gamma_{cb}(r_{ab}/r_{ba} - 1) \geq r_{ab} - r_{ba} \simeq r_{ab} \qquad [14]$$

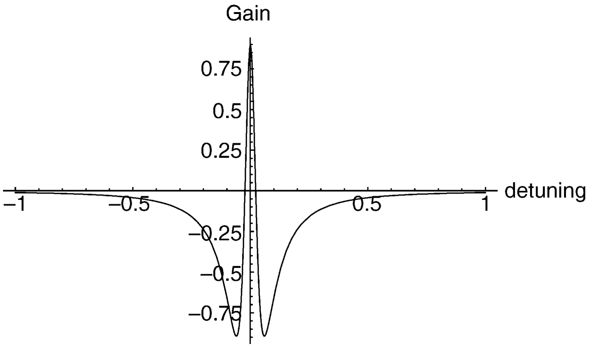

Figure 7 Gain without inversion in Λ-scheme.

because $\gamma_{cb} \geq r_{ba}/2$. Here we assumed that the incoherent pumping is weak, $r_{ba} \ll r_{ab}$, so that it does not produce population inversion at the probe transition.

Equation [14] is a simple and very important result, which is valid for all three-level laser schemes in the absence of inversion at both probe and drive transitions: the relaxation rate at the drive transition should exceed that at the probe transition. This result is an immediate consequence of the fact that the long-lived coherence σ_{cb} is inevitably destroyed by incoherent pumping, and γ_{cb} is now bounded from below. One can follow how the EIT regime is transformed into lasing as follows. With increasing incoherent pumping rate r_{ba}, EIT is actually destroyed due to the growth of $\gamma_{cb} \simeq r_{ba}/2$. In terms of CPT states, this means increase of population of the absorbing state $|B\rangle$. At the same time, the resonant Raman term in [13] also grows since n_{ac} is proportional to r_{ba}. With increase in the incoherent pumping rate, the contribution to gain overcomes the contribution to absorption if $r_{ac} > r_{ab}$.

Propagation effects

Inequality [14] has important implications for the problem of co-propagation of probe and drive fields. Indeed, to provide a significant amplification of the probe field, the absorption length $L_{\mathrm{d}} = \gamma_{ac}/(\eta_{\mathrm{d}} n_{ca})$ of the drive field should be larger than the amplification length $L_{\mathrm{p}} \simeq \gamma_{ac}/(\eta_{\mathrm{p}} n_{ca})$ of the probe. This requirement gives simply

$$\eta_{\mathrm{p}}/\eta_{\mathrm{d}} > 1 \qquad [15]$$

which in the absence of nonradiative relaxation can be rewritten as

$$\frac{r_{ab}}{r_{ac}} \frac{\nu_d^2}{\nu_p^2} > 1$$

In view of the gain condition given in eqn [14], the above requirement can be satisfied only if $\nu_d > \nu_p$, i.e., up-conversion is impossible.

An interesting solution to this problem is to realize population inversion at the driving transition. Then up-conversion can be realized with a lower relaxation rate at the drive transition as compared with probe transition. In this case in the Λ and V-schemes the two-photon term contributes to absorption at resonance, and gain is only possible far off-resonance. At the same time, in the ladder schemes shown in Figure 6c and d the gain is achieved at resonance, as discussed below in more detail.

It should also be noted that to excite laser oscillations in a high-Q cavity, large one-pass gain is not required at all, so that a Λ-scheme could still be employed for up-conversion.

Now, consider a probe field of arbitrary intensity. An important question is where the energy for amplification of the probe field comes from. It is known that for parametric processes in a transparent medium the Manley–Rowe relations are satisfied. Therefore, it might seem that the drive field is also the ultimate energy source for LWI. However, this is the case only for very large detunings of the drive and probe fields from the corresponding transition frequencies. In that case, we can neglect one-photon absorption, and the wave equations take the form typical for stimulated Raman scattering. They possess the first integral

$$\frac{|\varepsilon_d|^2}{\nu_d} + \frac{|\varepsilon_p|^2}{\nu_p} = \text{const} \qquad [16]$$

which expresses the conservation of the total number of quanta in both waves (Manley–Rowe conditions).

However, in the case of exact resonance, one can show that one-photon absorption of the drive field is always important, and the Manley–Rowe-type conservation law does not exist at all. The optimal condition for the probe field amplification is $\eta_p/\eta_d \gg 1$. In this case the stage of exponential amplification of the probe is maintained over a long distance, and the probe field intensity can grow to much larger values than the sum of initial intensities of the drive and probe fields. Apparently, the energy for the probe field amplification comes mainly from the medium, namely from the upper-level population ρ_{aa} supported by incoherent pumping.

V-Scheme

The V-scheme, shown in Figure 6b, is qualitatively similar to the Λ-scheme. An important difference is that there is now no long-lived coherence at the mixing transition c–a. Indeed, even if the transition $a \rightarrow c$ is strongly forbidden, the coherence γ_{ac} is destroyed by fast relaxations at the c–b and a–b transitions. We can immediately suggest that for the V-scheme inversionless lasing takes place at large driving field intensities exceeding the saturation value. However, one must be careful: although it appears that the reduction of absorption in the absence of pumping cannot be greater than a factor of 2, this turns out to be incorrect. Rather, almost complete transparency can be achieved, as confirmed by calculations and recent experiments. Analysis of the laser gain condition shows that the V-scheme shares with Λ-scheme the same negative feature: relaxation at the driving transition should be much faster than at the probe transition.

Schemes with Population Inversion at the Driving Transition

The above analysis illustrates the common problem of LWI schemes: the requirement of rapid relaxation at the driving transition and the resulting strong absorption of the drive field. This difficulty can be overcome in schemes with population inversion at the driving transition. Then, of course, all this makes practical sense only if the relaxation at the drive transition is much slower than that on the probe transition, so that it is easy to provide population inversion for the drive transition as compared with the probe transition. Luckily, such schemes do exist in principle, and seem to have many potential applications. The simplest possibility is the so-called ladder (or cascade) schemes shown in Figure 6c,d. LWI in ladder schemes is usually studied in the regime when there is no inversion at both probe and drive transitions. In this case gain is achieved at large detunings from two-photon resonance and under the same unfortunate condition of fast relaxation at the drive transition as compared with the probe transition. In the opposite case, population inversion at the driving transition is required for LWI at the probe transition. In ladder schemes the maximum gain of the probe field is achieved when both fields are at resonance with atomic transitions, while in V and Λ-schemes the resonance point corresponds to maximum absorption, and the gain is achieved in sidebands.

Note that population inversion at the driving transition means that the drive field can be amplified. This immediately puts forward the possibility to use a drive field which is self-generated in the same active medium. In other words, the system supports lasing with inversion at the drive frequency and LWI at the probe frequency. This situation is especially beneficial for dense gases or solids where the problems of external drive absorption, reflection, and inhomogeneity may become otherwise insurmountable. One specific example considered in the literature is LWI on rapidly decaying UV and soft X-ray transitions in gas lasers in the presence of lasing with inversion on long-lived adjacent transitions in the visible range (as in Figure 6c). Another example is LWI on rapidly decaying mid/far-infrared intersub-band transitions in semiconductor quantum wells in the presence of lasing with inversion on the interband transition, as in Figure 6d. These interband transitions have 10^4 times longer lifetime than the intersub-band transitions.

Experiments

Experimentally, amplification without inversion has been observed in rarefied atomic vapors in the early and mid-1990s. In particular, amplification of a probe beam in the Λ-configuration was realized in atomic samarium, sodium, and cadmium, while the V-scheme was successfully implemented in atomic rubidium. Inversionless laser oscillation was observed in a V-scheme in rubidium vapor and a Λ-scheme in a sodium atomic beam. In all of these proof-of-principle experiments, the frequencies of the probe and drive fields were very close to each other, and the coherence was excited between two hyperfine states or Zeeman sublevels. One-pass gain of a few per cent was observed.

Nonlinear Optical Sources Based on Quantum Coherence

Excitation of a strong coherence has a profound impact on nonlinear optical interactions. Strong scattering on the polarization wave excited at the beat frequency of two or more laser fields allows one to achieve huge nonlinear coefficients and conversion efficiencies. A particularly impressive example is ultra-broadband collinear Raman generation in a resonant Λ-scheme, that has been observed experimentally. A pair of strong laser pulses propagating in molecular deuterium excites a strong polarization wave at a beat frequency close to a vibronic transition in deuterium. As a result, a comb of Raman sidebands from the infrared to the deep-UV region of the spectrum has been observed.

Another example is an extremely efficient simultaneous blue light and mid-infrared generation in a four-wave mixing experiment in rubidium vapor. Here, all four fields are resonant with corresponding transitions.

A final example illustrates the use of self-generated drive fields for nonlinear generation. It has been suggested that one can use two intracavity generated laser modes in a semiconductor quantum-well laser for the difference frequency, sum frequency, and second-harmonic generation in fully resonant three-wave mixing processes. This idea has already been successfully realized in two-wavelength quantum-cascade lasers, and experiments with other types of semiconductor lasers are under way.

The whole field of resonant nonlinear optics, stemming from original ideas of LWI in systems with quantum coherence, is now in explosive development.

Summary

Atomic coherence effects have already found many important applications in metrology and low-intensity nonlinear optics. Novel applications in quantum nonlinear optics, laser physics, and quantum information processing seem to be feasible. The coherence allows the storage of information about the probe light, its transportion in space, time reversing, and multiplexing the light. Moreover, the coherence allows an increase of the coupling between light fields such that it becomes possible to study the interaction between single photons. New coherent radiation sources based on the ideas of LWI and coherent nonlinear optics are at the stage of active research. There are surely many other applications ahead.

See also

Electromagnetically Induced Transparency. **Non-Classical Light**. **Nonlinear Optics, Applications:** Raman Lasers. **Nonlinear Optics, Basics:** Cascading; Four-Wave Mixing. **Quantum Optics:** Entanglement and Quantum Information.

Further Reading

Belyanin AA, Bentley C, Capasso F, *et al.* (2001) Inversionless lasing with self-generated driving field. *Physical Review A* 64: 013814.

Harris SE (1997) Electromagnetically induced transparency. *Physics Today* 50(9): 36–42.

Kocharovskaya O (1992) Amplification and lasing without inversion. *Physics Reports* 219: 175–190.

Mompart J and Corbalán R (2000) Lasing without inversion. *Journal of Optics B: Quantum and Semiclassical Optics* 2: R7–R24.

Rostovtsev Y, Kocharovskaya O, Welch GR and Scully MO (2002) Slow, ultra-slow, stored and frozen light. *Optics Photonics News*, June.

Scully M and Zubairy S (1997) Quantum Optics. Cambridge, UK: Cambridge University Press.

Entanglement and Quantum Information

P G Kwiat, University of Illinois at Urbana-Champaign, Urbana, IL, USA

D F V James, Los Alamos National Laboratory, Los Alamos, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Quantum information is an emerging field of technology that encompasses the application of fundamental quantum mechanical phenomena, such as entanglement, to tasks in information processing and communications. The importance of optical technology in quantum information is not surprising, given the success of quantum optics in the investigation of these fundamental phenomena and the pre-eminent role of optics in modern communications. This article describes some of the important advances in photonic quantum information and discusses the prospects for the future.

Basic Notions: Qubits and Entanglement

Qubits

Quantum information is usually confined to two-level quantum systems, which are referred to as quantum bits, or qubits. The qubits form the register of a quantum computer or carry the information in a quantum communications channel. Physical examples of qubits under active study include the spin degrees of freedom of an electron or spin-1/2 nucleus, two well-isolated energy levels of an atom or trapped ion, and the polarization degrees of freedom of a photon; here we concentrate on the latter.[1] The two states of each qubit are generically denoted by $|0\rangle$ and $|1\rangle$ (although, because of our concentration on optical polarization, we shall use the horizontal and vertical states $|H\rangle$ and $|V\rangle$ interchangeably with $|0\rangle$ and $|1\rangle$ when describing experiments). An arbitrary state of a single qubit, $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, is specified by the complex probability amplitudes α and β associated with the two possible states, constrained by $|\alpha|^2 + |\beta|^2 = 1$. States for which α and β are deterministic quantities are referred to as pure states; when they are stochastic, the state is mixed, and must be described by a density matrix, e.g., $\rho = \sum_i p_i|\psi_i\rangle\langle\psi_i|$, where p_i is the probability of the pure state $|\psi_i\rangle$. Pure states are a useful but unattainable ideal, just like a perfectly coherent laser beam; in reality even the best-prepared quantum states have some level of mixedness, which can be quantified by the entropy: $S \equiv -Tr(\rho \ln\rho) = -\sum_i \lambda_i \ln\lambda_i$, where λ_i are the eigenvalues of the matrix ρ.

Multiqubit States and Entanglement

For two qubits, the possible pure states of the system are specified by four probability amplitudes:[2] $|\psi_2\rangle = \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$, with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$, and, e.g., $|00\rangle \equiv |0\rangle \otimes |0\rangle$ denoting the state in which the first and the second qubits are both in state $|0\rangle$. In general, 2^n amplitudes are needed to characterize a pure state of n qubits. Immediately we can see one advantage of quantum systems for information storage: if each distinct probability amplitude is regarded as a data register, the size of the memory grows exponentially with the number of qubits. Further, if one were to flip just one qubit, for example the second, the state would be transformed into $|\psi'_2\rangle = \beta|00\rangle + \alpha|01\rangle + \delta|10\rangle + \gamma|11\rangle$, where all of the amplitudes have been affected. This simple example demonstrates the notion of quantum parallelism, one of the most powerful properties of quantum information processors.

If the two-particle state can be written as a product of two single-particle states, i.e., if $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle = (A|0\rangle + B|1\rangle) \otimes (C|0\rangle + D|1\rangle)$, then these two particles can be considered as separate, unconnected entities: the state is said to be separable.[3] When the state cannot be written in this form, it is called an entangled state. Entanglement is one of the most fascinating fundamental properties of quantum systems: Erwin Schrödinger described it as '*the* characteristic trait of quantum mechanics … that enforces its entire departure from classical lines of thought.' Consider performing a measurement on the second of the two qubits, to establish which of its two states the particle was in, leaving the first qubit untouched. The result would collapse the state of the first qubit into one of two possibilities: either $|\phi_0\rangle = (\alpha|0\rangle + \gamma|1\rangle)/\sqrt{P_0}$ if the outcome of the measurement were 0 (which occurs with probability $P_0 = |\alpha|^2 + |\gamma|^2$) or $|\phi_1\rangle = (\beta|0\rangle + \delta|1\rangle)/\sqrt{P_1}$ if it were 1 (with probability $P_1 = |\beta|^2 + |\delta|^2$). Thus the state of the first qubit is instantaneously projected into a specific state by performing a measurement on the second qubit, which can be in a completely different location. This example demonstrates the nonlocality of quantum mechanics, which has been confirmed experimentally in some depth by a number of elegant quantum optics experiments over the last 30 years.

---

[1] Recently some attention has been devoted to quantum information based on continuous variables, such as those arising in connection with 'squeezed light'. However, here we will focus on discrete systems, i.e., qubits.

[2] Qubits are assumed to be distinguishable particles, and so the wavefunctions describing their states need not be symmetrized under exchange of particles.

[3] When mixture is also present, the state is separable if it can be written as $\rho = \sum_i p_i \rho_i^A \otimes \rho_i^B$, i.e., as a sum of product states for systems A and B.

The fidelity $F = |\langle\Phi|\Psi\rangle|^2$ characterizes the overlap of two states.[4] The fidelity of $|\phi_0\rangle$ and $|\phi_1\rangle$ is a way to quantify the amount of entanglement in the initial two-qubit state. If the final two states are identical, the fidelity will be unity, and the initial state is separable; conversely if the two final states are orthogonal and occur with equal probability, then in a sense the initial state is maximally entangled, since the largest possible nonlocal influence would occur due to the measurement. Mathematically, the fidelity of the two final states is $|\langle\phi_0|\phi_1\rangle|^2 = 1 - C^2/4P_0P_1$, where $C = 2|\alpha\delta - \beta\gamma|$ is called the concurrence of the state. If $C = 0$, the state is separable; if $C = 1$, the state is maximally entangled.[5]

As discussed below, entanglement forms the heart of a number of quantum information protocols, such as dense coding, teleportation, and one type of quantum key distribution (also known as quantum cryptography), and large-scale entangled states of many qubits seem to be a requirement for the more ambitious goal of practical quantum computing.

Creating Entangled States Experimentally

Entangled states can currently be created in a controlled manner using technologies such as ion traps, cavity quantum electrodynamics, and optical spontaneous parametric down-conversion (SPDC). Down-conversion is a nonlinear optical process by which an incident 'pump' photon can be split (or down-converted) into a pair of longer-wavelength daughter photons (historically called 'signal' and 'idler') in a crystal possessing a $\chi^{(2)}$ nonlinearity, such as beta-barium borate (BBO). Mathematically, the process is described using the creation and annihilation operators of the field modes $(\hat{a}^\dagger_\lambda, \hat{a}_\lambda)$, thus:

$$|\Phi_{\text{out}}\rangle = \exp\left[-i \sum_{p,s,i} c_{p,s,i}(\hat{a}^\dagger_p \hat{a}_s \hat{a}_i + \hat{a}^\dagger_i \hat{a}^\dagger_s \hat{a}_p) \right] |\Phi_{in}\rangle$$

where the triple sum is over all field modes (subscripts p, s, and i refer to pump, signal, and idler modes, respectively) and $c_{p,s,i}$ is a coefficient linearly dependent on the second-order nonlinear susceptibility $\chi^{(2)}_{ijk}$ and also on the birefringent properties of the crystal. Further, $c_{p,s,i}$ will be negligibly small unless both total photon energy and momentum are conserved (i.e., $\omega_p = \omega_s + \omega_i$ and $\vec{\kappa}_p = \vec{\kappa}_s + \vec{\kappa}_i$, where $\vec{\kappa}$ is the intracrystal momentum).

One particularly efficient method for using this phenomenon to create entangled states is as follows. For a specific geometry (type-I phase matching), the daughter photons emerge from the crystal with identical polarizations (perpendicular to the parent polarization and the crystal optic axis) on opposite sides of a cone that is centered about the pump beam (Figure 1a). Because each photon is in a definite state of polarization, the two photons are not in an entangled state. Two crystals, aligned with their axes of symmetry oriented at 90° to each other, as shown in Figure 1b, can then be used to create an entangled state.[6] With crossed crystals, two processes are possible: the parent photon can down-convert in the first crystal to yield two vertically polarized photons, or it can down-convert in the second crystal to yield two horizontally polarized photons. Because it is impossible to distinguish which of these processes has occurred, the state of the daughter photons is a coherent quantum-mechanical superposition of the states that would arise from each crystal alone; the output of the crossed crystals is photons in the maximally entangled state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle)$.[7]

Figure 2 shows how this basic source can be adapted to produce any pure quantum state of two photons by placing rotatable half- and quarter-wave plates (which can be used to transform the

---

[4] The generalized fidelity between mixed states ρ_A and ρ_B is $F = |Tr\sqrt{\sqrt{\rho_A}\rho_B\sqrt{\rho_A}}|^2$.

[5] The concurrence can also be generalized to mixed states, although it has a much more complicated expression. The quantification of entanglement for mixed states with more than two subsystems is presently an active area of research.

[6] Another widely used method employs type-II phase matching in a single crystal. The down-conversion photons are emitted with perpendicular polarizations along a pair of cones. Pairs emitted along particular directions are in the maximally polarization-entangled state $|\psi_-\rangle = \frac{1}{\sqrt{2}}(|HV\rangle - |HV\rangle)$. Such a source was used in the first demonstrations of superdense coding and quantum teleportation.

[7] Note that in an arbitrary basis $|\theta\rangle \equiv \cos\theta|H\rangle + \sin\theta|V\rangle$ and $|\theta^\perp\rangle \equiv -\sin\theta|H\rangle + \cos\theta|V\rangle$, this maximally entangled state has the form $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|\theta\theta\rangle + |\theta^\perp\theta^\perp\rangle)$, demonstrating that the nonlocal correlations are present regardless of the bases used to represent the state.

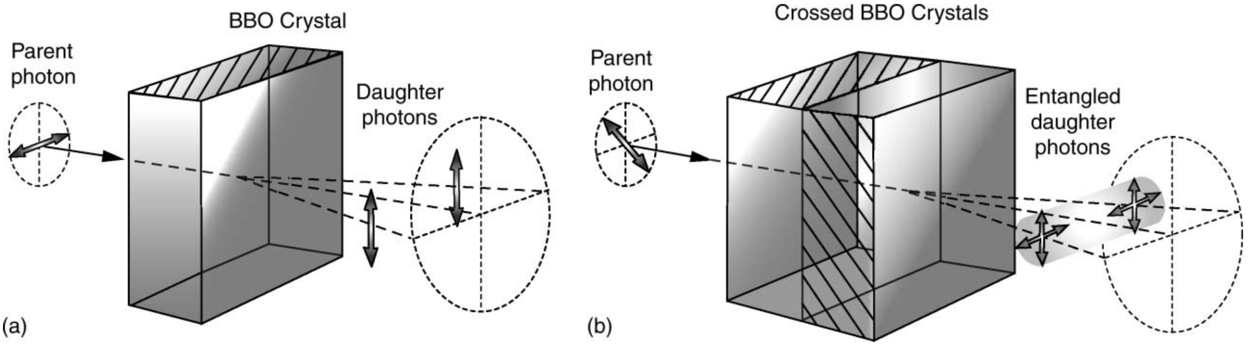

Figure 1 An entangled-photon source. (a) For a given orientation of the nonlinear crystal, a horizontally polarized parent photon produces a pair of vertically polarized daughters. The daughters emerge on opposite sides of an imaginary cone. The cone's axis is parallel to the original direction taken by the parent photon. The two daughter photons are not in an entangled state. Reorienting the BBO crystal by 90° will produce a pair of horizontally polarized daughters if a vertically polarized pump beam is used. (b) Passing a photon polarized at 45° through two crossed BBO crystals can produce two photons in an entangled state. Because of the Heisenberg uncertainty principle, there is no way to tell in which crystal the parent photon 'gave birth,' and so a coherent superposition of two possible outcomes results: a pair of vertically polarized photons or a pair of horizontally polarized photons. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

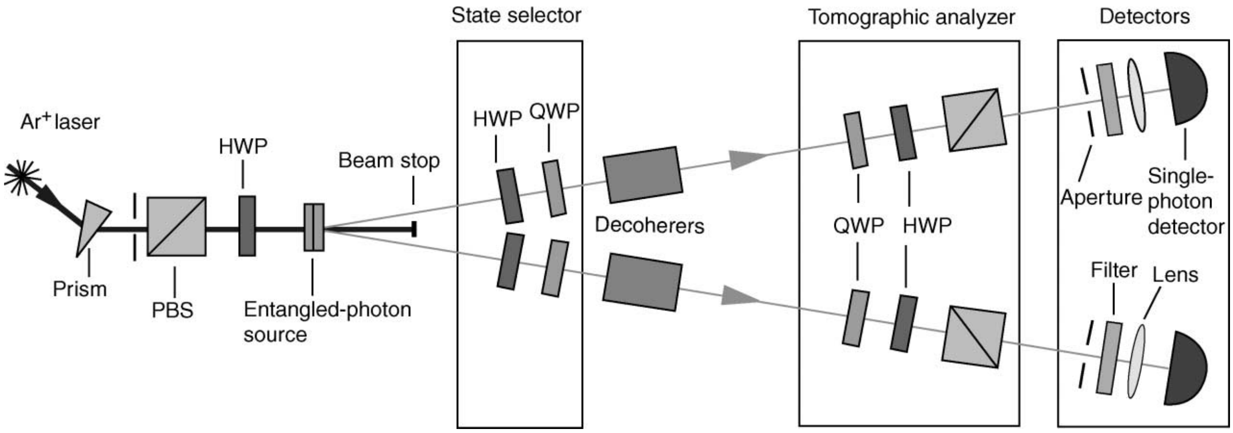

Figure 2 Creating and measuring two-photon entangled states. The 'pump' photons are created, for example, in an argon ion laser and are linearly polarized with a polarizing beamsplitter (PBS). The half-wave plate (HWP) rotates the polarization state before the photon enters the pair of nonlinear crystals that constitute the entangled-photon source; the initial angle of pump polarization controls the entanglement of the pair produced. Each photon's polarization state can be altered at will by the subsequent HWP and quarter-wave plate (QWP). The decoherers following the state selection allow the production of mixed photon states. The optical elements (QWP, HWP, and PBS) in the tomographic analyzer allow the measurement of each photon in an arbitrary basis. Coincidence measurements of photons allow the quantum state to be determined. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

polarization state of a single photon) before the crystal pair and in the paths of the two daughter photons. To create mixed states, a long birefringent crystal can be used to delay one polarization component with respect to the other. If the delay is longer than the coherence time of the photons, the horizontal and vertical components are effectively decohered; that is, the phase relationship between the different states is destroyed.

The figure also shows schematically the apparatus required for measuring the quantum state. In classical optics, four Stokes parameters are required to specify the polarization of a single beam (i.e., an ensemble of uncorrelated photons); for a pair of photons, 16 projective measurements, each with different wave plate settings, is required. From these 16 measurements, all of the elements of the 4×4 density matrix describing the (in general, mixed) state of the photon pairs can be deduced. This is an example of quantum state tomography, a technique that has found application to a number of quantum optical systems.

Quantum Key Distribution

Two parties, historically known as Alice and Bob, want to have a secret conversation. A generic, classical encryption protocol would begin when Alice and Bob convert their messages to separate binary streams of 0's and 1's, which are then encrypted and decrypted with a set of secret 'keys' known only to the two. Each key is a random string of 0's and 1's that is as long as the binary string comprising each message. To encrypt, Alice (the sender) sequentially adds each bit of the key to each bit of her message, using addition modulo 2. She then sends the encrypted message over a public channel to Bob, who decrypts it by simply repeating the addition modulo 2 of the key to the message. This type of encryption, known as a one-time pad, is currently the only provably secure encryption protocol. But the one-time pad is effective only if Alice and Bob never reuse the key, and more obviously, if the key remains secret. A potential eavesdropper, Eve, cannot be allowed to glean any part of the bit stream that makes up the key. Therein lies a central problem of cryptography – how can secret keys be created and then securely distributed?

Quantum key distribution (QKD) exploits the fundamentally indivisible nature of photons to perform this task. There are a variety of QKD protocols; here we describe one that employs entangled photon pairs (see Figure 3). Alice and Bob use a source such as the one described above to produce maximally entangled photons in the state $|\Phi_+\rangle \equiv \frac{1}{\sqrt{2}}(|HH\rangle + |VV\rangle) = \frac{1}{\sqrt{2}}(|45°,45°\rangle + |-45°,-45°\rangle)$. One photon goes to Alice and the other to Bob. For each pair, Alice and Bob randomly and independently analyze their respective photons (using a polarizing beamsplitter) in the H/V or the 45°/−45° basis. They record a bit value of 0 for all H or 45° results, and a 1 for all V or −45° results. After a sufficient number of measurements (dictated by the length of the key), Alice and Bob have a public discussion, e.g., over the internet. For each detected photon, they announce which basis they used for the measurement, but not the actual measurement result. Whenever they made the same basis choice (50% of the time), the correlations of the entangled state ensure their measured bit values agree. By contrast, they discard the results when they used different bases, because their measurements are completely uncorrelated (see Table 1).

An eavesdropper (Eve) cannot tap the line, as she might with conventional communications, because of the indivisibility of individual photons and the fact

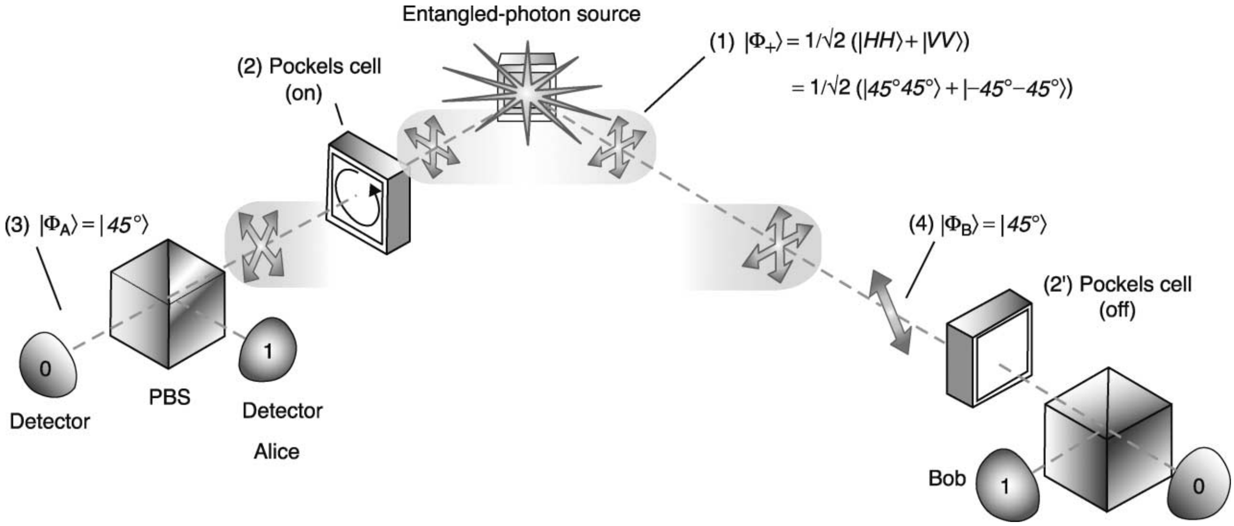
Entangled-photon source
(1) $|\Phi_+\rangle = 1/\sqrt{2}\,(|HH\rangle + |VV\rangle)$
$= 1/\sqrt{2}\,(|45°45°\rangle + |-45°-45°\rangle)$
(2) Pockels cell (on)
(3) $|\Phi_A\rangle = |45°\rangle$
0
Detector
PBS
1
Detector
Alice
(4) $|\Phi_B\rangle = |45°\rangle$
(2') Pockels cell (off)
Bob
1
0

Figure 3 Quantum cryptography using entangled photons. Entangled photons can be used to create a pair of identical cryptographic keys. One member of an entangled pair (1) is sent to Alice, and the other to Bob. Alice and Bob each randomly and independently analyze their respective photon in one of two linear polarization bases with a polarizing beamsplitter (PBS); the basis can be actively chosen using a Pockels cell before the PBS, as shown, to rotate the polarization (alternatively, the 'choice' could be made by directing the photon onto a nonpolarizing 50/50 beamsplitter; in the transmitted path, one analysis basis is used, in the reflected path, the other is used). In the example shown, Alice used the 45°/−45° basis (2), and measured 45° polarization (3), thus projecting Bob's photon into the identical state (4). Since he chose the H/V analysis basis (2′), he is equally likely to detect a 0 or a 1. By subsequent public discussions Alice and Bob determine the events for which they used the same analysis basis (and discard the other events). For these events Alice and Bob will have obtained identical measurement results, which they may interpret as raw key material. After classical error correction and privacy amplification techniques are applied, the remaining string is the sought-after shared secret key. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

Table 1 Polarization entanglement-based quantum cryptography protocol

Alice's analysis basis	H/V	H/V	45°/−45°	H/V	45°/−45°	45°/−45°	H/V	45°/−45°
Alice's measurement result[a]	*H*	–	*−45°*	*V*	*45°*	*45°*	*V*	*−45°*
Alice's bit value	0	–	1	1	0	0	1	1
Bob's analysis basis	H/V	45°/−45°	45°/−45°	45°/−45°	H/V	45°/−45°	H/V	HV
Bob's measurement result	*H*	*−45°*	*−45°*	*45°*	*H*	–	*V*	*H*
Bob's bit value	0	1	1	0	0	–	1	0
Public discussion: Both photons detected and same basis used?	yes	no	yes	no	no	no	yes	no
Remaining secret key	0		1				1	

[a] In some cases Alice or Bob may not detect a photon due to loss or detector inefficiency. These events simply do not contribute to the key data.

that arbitrary quantum systems cannot be accurately cloned.[8] If Eve steals Bob's photon (a 'denial-of-service' attack), the photon's information never becomes part of the key. Thus, although a wiretap would reduce the rate of the transmission, it would not jeopardize the security of the key. Eve can try to intercept the photon, measure it, and send another one to Bob. But any measurement Eve would make to determine the photon's polarization state would perturb the photon and collapse the entangled state. The photon she sends to Bob would therefore only be 'classically' correlated with Alice's photon. Consequently, Eve's intervention necessarily induces additional errors into Bob's key, which Alice and Bob can detect by publicly revealing a small subset of their actual key. Unfortunately, even with no eavesdropper, the encryption keys created by any real-world quantum cryptography system typically possess a few-percent errors. To make sure their key is secure, Alice and Bob ascribe all errors to Eve and then estimate the maximum amount of information available to the eavesdropper. They then use a classical privacy amplification protocol to reduce Eve's knowledge of the secret key to less than one bit by reducing the length of the key. It has been proven that if the initial error probability per bit is greater than ~15%, no secret bits will remain after error detection and privacy amplification.

Researchers are working to make entanglement-based quantum cryptography more practical. Also, a number of longer distance quantum cryptography demonstrations using weak pulses have been carried out, over tens of kilometers in fibers and in free space. In fact, the first commercially available systems have recently been announced (in Europe and the United States).

Table 2 Method to convert the initial state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ into any of the four Bell states

Alice's operation	*Polarization transformation*	*Resultant two-qubit state*
$\hat{I}$, the identity	H → H; V → V	$|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$
$\hat{\sigma}_x$	H → V; V → H	$|\Psi_+\rangle = \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle)$
$i\hat{\sigma}_y$	H → V; V → −H	$|\Psi_-\rangle = \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)$
$\hat{\sigma}_z$	H → H; V → −V	$|\Phi_-\rangle = \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle)$

Quantum Superdense Coding

It is possible for Alice to send Bob two bits of classical information using a single qubit in the quantum superdense coding protocol. Suppose that Alice and Bob share one qubit each of an entangled pair in the maximally entangled state $|\Phi_+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, where we have returned to the generic labeling scheme for the qubit states. Alice has two classical bits of information, which is equivalent to one of four choices. She can encode this information by applying one of four possible transformations on her qubit; a suitable set of transformations are the three Pauli matrices and the identity (i.e., do nothing). This set of operations can be performed experimentally on photon qubits quite easily, e.g., using waveplates, as indicated in Table 2. The four resultant two-qubit states are all maximally entangled and are orthonormal. They form a special basis for the two-qubit states, the Bell basis, which is of particular use in analyzing entanglement.

Alice now sends her qubit to Bob, e.g., through an optical fiber. Bob can then perform a Bell state analysis[9] on his qubit pair, i.e., a projective measurement of the two-qubit state in the Bell basis. The result immediately reveals the choice of operation Alice made, and in effect, two bits of classical information have been encoded on a single qubit.

---

[8] The no-cloning proof is very simple. If we have a copying operation such that $|0\rangle|c\rangle \to |0\rangle|0\rangle$ and $|1\rangle|c\rangle \to |1\rangle|1\rangle$ ($|c\rangle$ is the initial state of the copier), then by linearity, a superposition becomes an entangled state, instead of two copies: $\frac{(|0\rangle + |1\rangle)}{\sqrt{2}}|c\rangle \to \frac{(|0\rangle|0\rangle + |1\rangle|1\rangle)}{\sqrt{2}} \neq \frac{(|0\rangle + |1\rangle)}{\sqrt{2}}\frac{(|0\rangle + |1\rangle)}{\sqrt{2}}$.

[9] Complete discrimination of all four Bell states is presently an unsolved technical problem: for photons one needs either a nonlinear interaction (which is typically very weak) or to exploit so-called 'hyper-entanglement' involving other entangled degrees of freedom of the photons.

Quantum Teleportation

Another application of entanglement is quantum state teleportation (Figure 4), in which the infinite amount of information contained in an arbitrary qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ is transferred by communication of two bits of classical information between Alice and Bob, if they also share two qubits in the maximally entangled state $|\Psi_-\rangle$. The three-photon initial state (i.e., the input photon plus the two entangled photons) is

$$|\psi_0\rangle = (\alpha|H\rangle + \beta|V\rangle) \otimes \tfrac{1}{\sqrt{2}}(|HV\rangle - |VH\rangle),$$

which can be rewritten with the first two photons (the input plus the first half of the entangled pair) represented by the Bell state basis:

$$\begin{aligned} |\psi_0\rangle = \ & \tfrac{1}{2}\{|\Phi_+\rangle \otimes (-\beta|H\rangle + \alpha|V\rangle) \\ & + |\Phi_-\rangle \otimes (\beta|H\rangle + \alpha|V\rangle) \\ & + |\Psi_+\rangle \otimes (-\alpha|H\rangle + \beta|V\rangle) \\ & + |\Psi_-\rangle \otimes (-\alpha|H\rangle - \beta|V\rangle)\}. \end{aligned}$$

Now suppose that a Bell state measurement is performed on the first two photons. The third photon

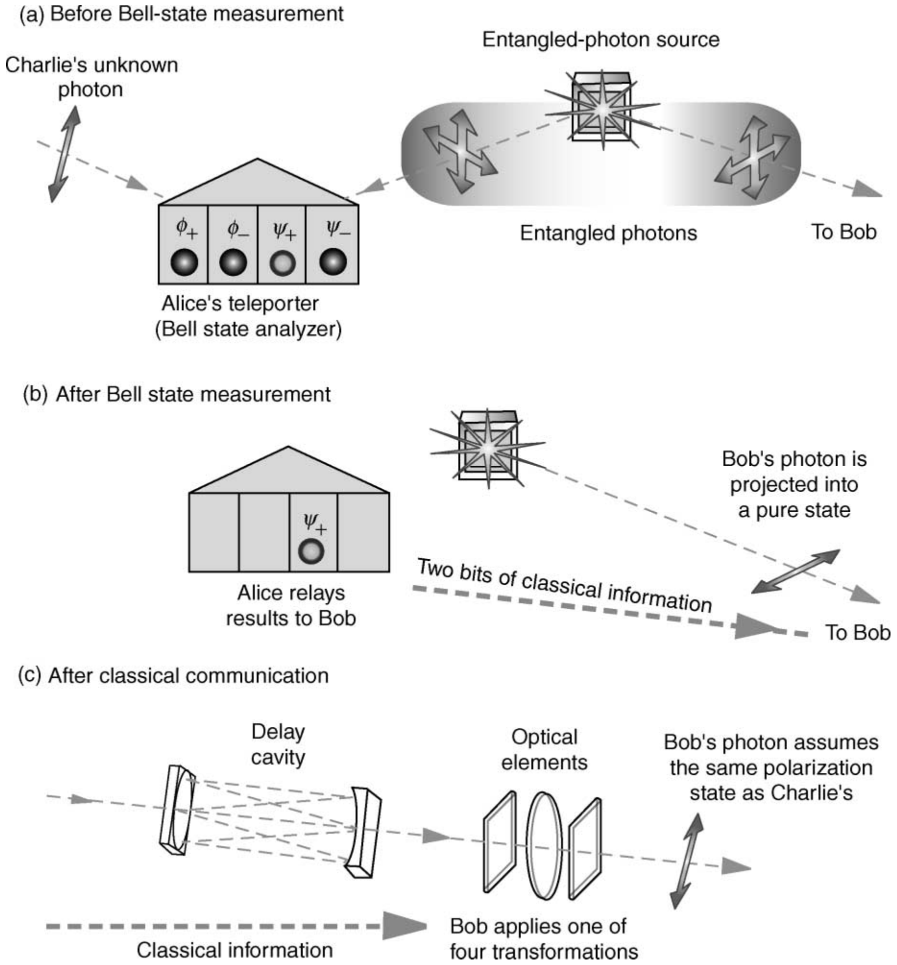
(a) Before Bell-state measurement
Charlie's unknown photon
Entangled-photon source
ϕ_+
ϕ_-
ψ_+
ψ_-
Entangled photons
To Bob
Alice's teleporter (Bell state analyzer)
(b) After Bell state measurement
ψ_+
Bob's photon is projected into a pure state
Two bits of classical information
Alice relays results to Bob
To Bob
(c) After classical communication
Delay cavity
Optical elements
Bob's photon assumes the same polarization state as Charlie's
Classical information
Bob applies one of four transformations

Figure 4 Quantum state teleportation. (a) Teleportation requires an entangled photon source, a Bell state analyzer and a classical communications channel. One entangled photon goes to Bob and the other to Alice, who also receives a photon of unknown polarization state. (b) Alice performs a joint polarization measurement of the two photons and relays the result to Bob using two classical bits of information. The photon going to Bob is projected into a pure state as a result of Alice's measurement. (c) Upon receiving Alice's classical information, Bob performs a simple transformation on his photon (which he has been storing), such as a rotation of the polarization vector. The resulting state of his photon is then identical to the unknown state Alice wished to teleport. (Adapted with permission from James DFV and Kwiat PG (2002) Quantum state entanglement: creation, characterization, and application. *Los Alamos Science* 27: 52–67.)

is immediately projected into one of four possible states, which can be transformed back into the state of the original input photon by a simple operation, e.g., with a waveplate. For example, if the Bell state measurement produced the result $|\Psi_+\rangle$, the third photon is immediately 'collapsed' into the pure state $|\psi_3\rangle = -\alpha|H\rangle + \beta|V\rangle$. By applying a π-phase shift to the horizontal polarization (relative to the vertical), $|\psi_3\rangle$ can be transformed into the original input state: $|\psi'_3\rangle = \alpha|H\rangle + \beta|V\rangle$. Alice (who does the Bell state measurement) need only communicate to Bob (who wants to receive the teleported photon) which of the four Bell states she measured. Note that during the entire teleportation procedure, neither Alice nor Bob can obtain any idea of the values of the parameters α and β, which specify the state. Also, because Alice's measurement collapses the unknown state, there is only a single copy at the end of the protocol.

Quantum state teleportation was first demonstrated experimentally by Zeilinger and coworkers (University of Innsbruck, Austria). The group was able to determine two of the four Bell states unambiguously (the other two states gave the same experimental signature) and proved for those cases that the state of the input photon could indeed be transferred to Bob. Other experiments have realized modified forms of quantum teleportation in different systems. For example, Kimble's group (California Institute of Technology) teleported the coherent state of an optical mode using squeezed (rather than polarization-entangled) light. Researchers have recently suggested how teleportation might form the basis of a distributed network of quantum communication channels, and how it might enable quantum computing in all-optical systems.

Other Application of Optical Entanglement

Quantum Computing

The most challenging and powerful application of quantum information is large-scale quantum computing. Two features that make quantum information processing potentially powerful are the exponentially large Hilbert space, which gives quantum registers very large capacities, and quantum parallelism, which means that data processing tasks can be performed very efficiently. One fundamental drawback, which severely constrains useful applications of quantum computers, is that the final measurement can only produce a number of classical bits equal to the number of qubits. Thus, quantum computers are limited to performing tasks in which a small amount of information is meant to be gleaned from processing a large amount of data; examples include searching an unstructured database for a specific entry (Grover's algorithm) or finding the periodicity of a function (the quantum Fourier transform). This second task is central to Shor's factor-finding algorithm, the most famous quantum computing algorithm to date.

A practicable quantum computer technology must have at least the following features, first identified by DiVincenzo:

- A set of well-characterized, distinguishable qubits to form a quantum data storage register.
- The ability to initialize the qubits of the register in a simple fiducial state.
- Decoherence times that are much longer than the time needed to perform logical operations.
- The ability to perform any single qubit operation on any qubit in the register, and the ability to perform two-qubit conditional logic gates (such as the CNOT gate: $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \rightarrow \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle$). Together, these operations constitute a universal set of quantum gates, from which all other gates can be synthesized.
- The ability to measure each qubit.

All-optical schemes have been used to implement small quantum algorithms. However, most of the approaches are not scalable, due to the limitations of non-linear optics to perform a CNOT gate at the single-photon level. Recent proposals have suggested that very high efficiency single-photon detectors, along with sources of single photons 'on-demand,' allow scalable quantum computing with only linear optics; preliminary two-qubit gates have been experimentally demonstrated. A number of other promising candidate technologies that meet DiVincenzo's five requirements are being pursued vigorously. For example, the ability to create multi-qubit entanglement and perform reliable measurements on trapped ions cooled and manipulated by lasers has recently been demonstrated. Solid state systems offer the possibility of scalability, and a number of schemes, such as quantum dots, isolated impurities with nuclear spins, and superconducting quantum interference devices (SQUIDs), are being investigated. It is possible that the final quantum computing technology may take the form of a hybrid between various present approaches.

Lithography

Lithography, in which a pattern is optically imaged onto some photoresistive material, is the primary method of manufacturing microscale or nanoscale electronic devices. An inherent limitation of this process is that details smaller than a wavelength of light cannot be written reliably. However, quantum state entanglement might circumvent this limitation. Under the right circumstances, the interference pattern formed by beams of entangled photons can have half the classical fringe spacing. Quantum lithography requires two beams of photons, a coherent superposition consisting of the state in which two photons are in beam A while none are in B, and the state in which no photon is in beam A while two photons are in B. Such number-entangled states can be made in the laboratory, and the predictions about fringe spacings have been verified. However, other obstacles must be overcome to surpass current classical-lithography techniques.

Two-Photon Imaging and Microscopy

At present, two-photon microscopy is widely used to produce high-resolution images, often of biological systems. However, the classical light sources (lasers) used for the imaging have random spreads in the temporal and spatial distributions of the photons, and the light intensity must be very high if two photons are to intersect within a small enough volume to cause a detectable excitation. Such high intensity can damage the system under investigation. Because the temporal and spatial correlations may be much stronger between members of an entangled photon pair, much weaker light sources could be used, which would be much less damaging to the systems being observed. The development of such systems is currently an active area of research.

List of Units and Nomenclature

Entanglement: a property of quantum systems consisting of two or more distinct subsystems, often separate particles. When transformations on one system do not affect the other, the state is said to be separable; when this is not the case, and the quantum state of the overall system cannot be resolved into separate states of the individual pieces, the system is entangled.

Pure and mixed states: when the probability amplitudes specifying a quantum state are deterministic, the state is said to be pure; when the amplitudes are random quantities, the state is mixed. The distinction is similar to that of coherent and partially coherent optical fields.

Quantum computer: an information processing device that exploits quantum mechanical phenomena to greatly enhance computational power for certain problems. Some mathematical problems thought to be intractable on conventional computers can, in theory, be performed efficiently on a quantum computer.

Quantum crytography (also known as quantum key distribution): a quantum information protocol by which two parties acquire a shared series of random numbers (a cryptographic key) by exploiting the quantum nature of light. Absolute security can be proven by virtue of the indivisibility and uncopy-ability of individual quanta.

Quantum state tomography: a means by which quantum states can be determined experimentally by performing a series of appropriate measurements on multiple identically prepared systems. From such measurements, the elements of the density matrix, which fully specifies the state, may be inferred.

Quantum superdense coding: A quantum information protocol by which two bits of classical information may be communicated by a single quantum bit, initially part of an appropriate entangled quantum state.

Quantum teleportation: a quantum information protocol by which the unknown quantum state of one particle can be transferred to another distant particle, using a pair of entangled particles, a projective measurement, and exchange of two bits of classical information.

Qubit (quantum bit): a two-level quantum mechanical system, which constitutes the building blocks of quantum information processing devices. Examples include a spin-1/2 particle, an atom with two well-isolated levels, or the polarization of a photon.

See also

Nonlinear Optics, Applications: Three-Dimensional Microfabrication. **Quantum Optics:** Quantum Computing with Atoms.

Further Reading

Berman GP, Doolen GD, Mainieri R and Tsifrinovitch V (1998) *Introduction to Quantum Computers*. Singapore: World Scientific.

Bouwmeester D, Ekert AK and Zeilinger A (eds) (2000) *The Physics of Quantum Information*. Berlin: Springer-Verlag.

Braunstein SL and Lo HK (eds) (2001) *Scalable Quantum Computers: Paving the Way to Realization*. Berlin: Wiley-VCH.

DiVincenzo DP (2000) The physical implementation of quantum computation. *Fortschrcitte der Physik* 48: 771–783.

Gisin N, Ribordy G, Tittel W and Zbinden H (2002) Quantum cryptography. *Review of Modern Physics* 74: 145–195.

Kok P, Braunstein SL and Dowling JP (2002) Quantum lithography. *Optics and Photonics News* 13(9): 24–27.

Lo HK, Popescu S and Spiller T (eds) (1998) *Introduction to Quantum Computation and Information*. Singapore: World Scientific.

Macchiavello C, Palma GM and Zeilinger A (eds) (2000) *Quantum Computation and Quantum Information Theory*. Singapore: World Scientific.

Mandel L and Wolf E (1995) *Optical Coherence and Quantum Optics*. Cambridge, UK: Cambridge University Press.

Nielsen MA and Chuang IL (2000) *Quantum Computation and Quantum Information*. Cambridge, UK: Cambridge University Press.

Williams CP and Clearwater SH (1998) *Explorations in Quantum Computing*. Santa Clara: Telos.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Laser Cooling of Ions

H Walther, University of Munich and
Max-Planck-Institute for Quantum Optics, Garching,
Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that light can exert a mechanical force on atoms by the transfer of linear momentum was recognized as long ago as the 1930s, when Frisch demonstrated the deflection of an atomic beam using resonant light. With the advent of tunable laser light in the early 1970s, a much larger excitation rate for atoms was achieved, leading to a much stronger deflection. A little earlier it was also observed that in addition to the force resulting from photon recoil, a dipole force is also produced in a focused light beam so that even small solid particles can be trapped and manipulated. The original proposal for laser cooling of atomic gases was put forward in 1975 and later laser cooling of trapped ions was proposed. The principle of laser cooling is very simple. The ion is excited by laser light below a resonance frequency; in order that a transition can be made, the residual energy has to be taken from the kinetic energy of the particle, leading to cooling of the particle. In this article, we will discuss this cooling process in more detail.

Laser Cooling of Ions in a Radiofrequency Trap

Of the various proposed and realized methods for laser cooling of atoms or trapped ions, sideband cooling is conceptually the simplest and most efficient. It allows a trapped particle to be cooled to the quantum-mechanical ground state of a harmonic trap potential. The process can be regarded as anti-Stokes Raman scattering of the pseudomolecule formed by the ion and the trap. If the linewidth of an optical transition in this atom is smaller than the vibrational frequency of the trapped ion, both electronic levels of the transition split into a ladder of resolved vibrational levels. The absorption spectrum consists of a carrier at the resonance of the free atom and sidebands at multiples of the vibrational frequency. With a laser tuned to the m-th lower frequency sideband, each absorption of a photon will reduce the vibrational energy of the atom by m quanta. Since the subsequent spontaneous re-emission will, on average, not change the vibrational excitation, repeated photon scattering processes lead to efficient cooling of the atom. This cooling will continue until the absorption probability for the low-frequency sidebands decreases when approaching the vibrational ground state of the ion in the trap. The final temperature is limited by the heating process caused by photon recoil. However, after the ion is cooled to the vibrational ground state this heating process disappears, when the vibrational energy of the ion is larger than the recoil energy of a photon. In this case, the recoil is taken up by the entire trap setup, as it is the case in the Mössbauer effect.

Despite these obvious advantages of sideband cooling, relatively few experiments have been made using this method. This is because the regime, where the oscillation frequency ω is larger than the natural linewidth Γ of an optical transition (the so-called 'strong binding regime'), is not easily accessible. Vibrational frequencies of ions in standard electromagnetic traps are of the order of 1 MHz, whereas linewidths of electric dipole transitions in positive ions are in the range 20–50 MHz. In this 'weak binding regime' with $\Gamma > \omega$, laser cooling of trapped ions is essentially analogous to Doppler cooling of free atoms. The minimum temperature is given by the Doppler limit $k_{\mathrm{B}}T = \hbar\Gamma/2$, and the mean vibrational quantum number is $\langle n\rangle = \Gamma/2\omega > 1$. The first experiments with cooling on resolved sidebands, therefore, used a strongly forbidden optical transition or stimulated Raman transitions. In these cases, however, sideband cooling is inevitably slow and it was used only to extract a few vibrational quanta from Doppler-precooled ions. In both experiments the vibrational ground state was reached to a very good approximation (mean quantum number $\langle n\rangle \approx 0.05$). Raman sideband cooling has also recently been applied to neutral atoms in optical traps.

Laser cooling of trapped ions was motivated by their possible use for optical frequency standards where the relativistic Doppler shift, as well as the Stark shift due to the trap potential, are eliminated at low temperatures. Cooling to the ground state also opens up the possibility of studying a well-controlled mechanical quantum oscillator experimentally. This led to the realization of nonclassical states of motion, investigation of quantum decoherence phenomena, and the demonstration of a quantum logic gate for quantum computing.

Sideband Cooling

Most theoretical treatments of sideband cooling consider a static harmonic trap, whereby the time dependence of the potential of the radio-frequency (rf) ion trap is neglected. In such a model a running laser wave $\propto\exp(\mathrm{i}kr)$ of wave number k induces transitions between the eigenstates $|n\rangle$ and $|n'\rangle$ of a harmonic oscillator. The relevant transition matrix elements are

$$U_{n'n} = \langle n'|\exp(\mathrm{i}kr)|n\rangle \qquad [1]$$

$$U_{n'n} = \exp(-\eta^2/2)\sqrt{\frac{n_<!}{(n_< + \Delta n)!}}(\mathrm{i}\eta)^{\Delta n} L_{n_<}^{\Delta n}(\eta^2) \qquad [2]$$

Here $\eta = kx_0 = k\sqrt{\hbar/2m\omega}$ is the Lamb–Dicke parameter for the spatial extension x_0 of the ground state, $n_<$ the smaller of the two numbers n and n', $\Delta n = |n - n'|$, and L_n^m denotes the generalized Laguerre polynomial. In the case of high vibrational excitation $n, n' \gg 1$, $U_{n',n}$ can be approximated as

$$U_{n+\Delta n,n} \approx \mathrm{i}^{\Delta n} J_{\Delta n}(2\eta\sqrt{n}) \qquad [3]$$

with the Bessel function of n-th order J_n. The argument of the Bessel function is the product of the wave number and the classical vibration amplitude of the highly excited state. This is the result one would expect for the frequency-modulated spectrum of a classical oscillator. For arguments β among the Bessel functions $J_n(\beta)$, those of order $n \approx \beta$ have the largest modulus. Consequently, during the initial stages of cooling it is most efficient to tune the laser so that it induces transitions with $\Delta n \approx 2\eta\sqrt{n} \gg 1$.

For the final stages of cooling one can usually assume that the ion is in the Lamb–Dicke regime, i.e., that it is localized within dimensions of the optical wavelength. A detailed study of the limits of sideband cooling in a static trap was carried out in 1984. The final distribution of population over the vibrational states is thermal and can be characterized by a mean occupation number $\langle n\rangle$ or a temperature $T = \hbar\omega/k_{\mathrm{B}} \ln(1 + 1/\langle n\rangle)$. The number $\langle n\rangle$ is a rational function of the four frequencies involved, i.e., the natural linewidth Γ, the oscillation frequency in the trap ω, the laser detuning δ, and the Rabi frequency ω_{R}. For certain limiting cases, simple expressions for the mean vibrational excitation $\langle n\rangle$ can be derived from the general result. The lowest vibrational excitation is achieved in the case of well-resolved sidebands ($\omega \gg \Gamma$) and under weak laser excitation ($\omega_{\mathrm{R}} \ll \Gamma$), tuned to the first sideband ($\delta = -\omega$):

$$\langle n\rangle = \left(\alpha + \frac{1}{4}\right)\frac{\Gamma^2}{4\omega^2} \qquad [4]$$

The constant α depends on the geometry of excitation and photon re-emission and is of the order of 1. In the case of strong saturation ($\omega_{\mathrm{R}} \gg \Gamma$) of the first sideband ($\delta = -\omega$), the mean vibrational quantum number is proportional to the laser intensity:

$$\langle n\rangle = \frac{\omega_{\mathrm{R}}^2}{8\omega^2} \qquad [5]$$

This case is of practical relevance if one wants to detect a fluorescence signal on the cooling transition, for example, to measure the internal state of the ion in a double-resonance experiment, while keeping the ion cooled. For the choice of parameters $\omega_{\mathrm{R}} = \omega = -\delta$, the ion is still predominantly in the vibrational ground state ($\langle n\rangle \approx 1/8$), but photons are scattered from the wing of the saturation-broadened carrier at a rate $P = \Gamma\omega_{\mathrm{R}}^2/(\Gamma^2 + 2\omega_{\mathrm{R}}^2 + 4\delta^2) \approx \Gamma/6$, which is already one-third of the maximum scattering rate. Finally, in the case of large laser detuning ($\delta \gg \Gamma$, $\omega, \omega_{\mathrm{R}}$), the energy is proportional to the detuning because the lower vibrational levels are no longer depleted by the laser:

$$\langle n\rangle \approx \frac{\alpha + 1}{4}\frac{\delta}{\omega} \qquad [6]$$

If the Lamb–Dicke parameter satisfies $\eta \ll 1$, the strength of the sidebands can be calculated from a power series expansion of the matrix elements of eqn [2] and the thermal distribution of population over the oscillator levels. The $U_{n',n}$ are evaluated to first-order in η only. In this approximation, the height of the first lower-frequency sideband in the absorption spectrum relative to the carrier is $\eta^2\langle n\rangle$ and that of the first higher-frequency sideband $\eta^2(\langle n\rangle + 1)$. This information can be used to determine experimentally the vibrational quantum number from the absorption spectrum. The strengths of the higher-order sidebands contain higher powers of η^2 and therefore decrease rapidly.

Let us now briefly review some facts about the motional spectrum of an ion in a Paul trap, which is not just a static system but possesses an explicitly time-dependent potential. The equation of motion for a classical particle of charge e and mass m in the quadrupole potential of the trap $\phi(r,z) = (V_0 - U_0 \cos \Omega t)(r^2 - 2z^2)/2r_0^2$ is given by the Mathieu differential equation:

$$\frac{\delta^2 r}{\delta \tau^2} + (a_r - 2q_r \cos 2\tau)r = 0 \qquad [7]$$

with dimensionless parameters:

$$\tau = \frac{\Omega}{2}t, \qquad a_r = \frac{4eV_0}{m\Omega^2 r_0^2}, \qquad q_r = \frac{2eU_0}{m\Omega^2 r_0^2} \qquad [8]$$

Here a static voltage V_0 and an alternating voltage with amplitude U_0 and frequency Ω are applied to a Paul trap with the radius of the ring electrode r_0. The equation of motion for the z coordinate is obtained by multiplying the parameters a and q by the factor -2. The general solution of the Mathieu equation in the first stability region of the (a,q) parameter space can be written by using the Floquet ansatz:

$$r(\tau) = A\sum_{j=-\infty}^{\infty} c_j \cos(2j+\beta_r)\tau + B\sum_{j=-\infty}^{\infty} c_j \sin(2j+\beta_r)\tau \qquad [9]$$

where A and B are constants determined by the initial conditions, and the c_j and β_r depend on a and q. In the first stability region, where most of the experiments are carried out, the value of β_r is between 0 and 1. According to eqn [9] the motional spectrum of the ion has resonances at

$$\omega_{r,j} = \left(j \pm \frac{\beta_r}{2}\right)\Omega, \qquad j = 0, 1, 2\ldots \qquad [10]$$

In the adiabatic approximation, which is valid for $a, q^2 \ll 1$, the motion of the ion can be separated into an oscillation at the driving frequency Ω, the so-called micromotion, and a slower oscillation, called the secular motion, describing the oscillation in the time-averaged pseudopotential. This pseudopotential describes the effect of the ponderomotive force of the oscillating trapping field that drives the ion to the field minimum at the center of the trap. Since, in the quadrupole trap the pseudopotential is harmonic, the solution to the equation of motion can be approximated as

$$r(\tau) = r_0 \cos\beta_r\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [11]$$

The approximate expression for β_r is $\beta_r = \sqrt{q_r^2/2 + a_r}$ and the secular frequencies are given by

$$\omega_r = \frac{\Omega}{\sqrt{8}}\sqrt{q_r^2 + 2a_r}, \qquad \omega_z = \frac{\Omega}{\sqrt{2}}\sqrt{q_r^2 - 2a_r} \qquad [12]$$

which are the lowest-order resonances $\omega_{r,z} = \beta_{r,z}\Omega/2$ in the Floquet ansatz, eqn [9]. From eqn [11] it is seen that the amplitude of motion at the resonances $\Omega \pm \omega$ is a factor $q_r/4$ smaller than those at the resonance ω. It is noteworthy that the frequency Ω of the driving field does not appear directly in the motional spectrum, but only in combination with the secular frequency ω. This changes, however, if there is an additional static force acting on the ion, displacing it from the center of the quadrupole. The approximate solution to the equation of motion then becomes:

$$r(\tau) = r_1 + r_0 \cos\beta\tau\left(1 + \frac{q_r}{2}\cos 2\tau\right) \qquad [13]$$

and a resonance at Ω appears. The force that displaces the ion by r_1 may be due to static electric stray fields or the presence of a second ion in the trap. These effects can have a strong influence on sideband laser cooling of the trapped ions. A number of theoretical studies have been published on the quantum-mechanical treatment of the Paul trap. The explicitly time-dependent potential does not allow stationary states with time-independent energy eigenvalues. One can, however, as in the case of the pseudopotential model, distinguish between two time-scales and construct wave functions of the harmonic secular oscillator that show some additional breathing motion at the trap frequency. A quantum-mechanical theory of sideband cooling in the time-dependent potential of the Paul trap thus far exists only for special cases; in 1994 there was a case where the ion was localized in the Lamb–Dicke regime at the node of a standing laser wave and time averaged kinetic energies were calculated. Minima of the energy that agree with the results of the theory for the static harmonic trap to within a factor of 2 were found. The strongest modifications to the predictions of the cooling theory for the static trap arise in the dependence of the energy on the laser detuning. The influence of the micromotion gives rise to heating for certain regions of negative laser detuning (on the high-frequency side of a low-frequency micromotion sideband) and to cooling for some positive detunings (on the low-frequency side of a high-frequency micromotion sideband).

Trapped LASER-cooled ions have a variety of applications in spectroscopy and other fields.

In the following, a phenomenon will be discussed where observations during the cooling process of ions in traps is interpreted as phase transitions between chaotic and ordered ion structures during the cooling process.

Generation of Chaos During the Cooling Process and Phase Transitions to Ordered Configurations

Using a Paul trap, it is possible to study the physics of a few-body phase transition of ions. The ordered state of the trapped ions can be observed directly by monitoring their resonantly scattered light. The transition to an ordered or 'crystalline' configuration is expected when the Coulomb coupling constant which corresponds to the ratio of Coulomb energy to the kinetic energy of a particle, is larger than 1. The ordered structures represent configurations of maximum density for the ions, they are determined by the trap potential pushing the ions towards the trap center, and by the Coulomb repulsion between the ions acting in the opposite direction, so that minimum energy configurations are finally adopted. The system corresponds to an ordered one-component plasma or a strongly coupled plasma.

In the following, computer simulations of the phase transition in the Paul trap will be reviewed. In the trap the ions are subjected to five different forces. These are the force $\mathbf{F}^{(\text{trap})}$, arising from the dynamical rf trapping field, the Coulomb interaction between the ions $\mathbf{F}^{(\text{Coul})}$, the laser cooling force $\mathbf{F}^{(\text{laser})}$, a random force $\mathbf{F}^{(\text{rand})}$ arising from the recoil of the spontaneously emitted photons, and a parasitary force $\mathbf{F}^{(\text{cont})}$, arising from a contact potential, which may be generated unintentionally by locally coating the electrodes with the trapped atoms.

The potential generated by the rf voltage applied to the hyperbolic electrodes of the trap is given by

$$\phi(\mathbf{r},t) = \varphi(t)[x^2 + y^2 - 2z^2]$$
$$\varphi(t) = \frac{U_0 + V_0\cos(\Omega t)}{r_0^2 + 2z_0^2} \qquad [14]$$

where U_0 and V_0 are the dc and ac components of the voltages applied to the trap. From eqn [1] we obtain the trapping force and the resulting equation of motion of a single ion in the trap:

$$\mathbf{F}^{(\text{trap})} = m\ddot{\mathbf{r}} = \mathrm{e}\nabla\phi(\mathbf{r},t) = -2\mathrm{e}\varphi(t)[\mathbf{r} - 3z\mathbf{e}_z] \qquad [15]$$

where m, $\mathbf{r}$, and $z\mathbf{e}_z$ denote the mass of the ion, the position, and the $z\mathbf{e}_z$-component of the position vector, respectively.

The Coulomb interaction between the particles results in the force:

$$\mathbf{F}_i^{(\text{Coul})} = \frac{e^2}{4\pi\varepsilon_0}, \sum_{m\neq 1} \frac{\mathbf{r}_i - \mathbf{r}_m}{[\mathbf{r}_i - \mathbf{r}_m]^3} \qquad [16]$$

In the process of laser cooling, every scattered photon changes the momentum of an ion on the average by an amount $\hbar k$.

The ensemble average of the resulting laser cooling force acting on the i-th ion is thus:

$$\mathbf{F}_i^{(\text{laser})} = \hbar k R[\mathbf{r}_i(t);\, \dot{\mathbf{r}}_i(t)] \qquad [17]$$

The number $N_i(l)$ of spontaneously emitted photons from the i-th ion during the l-th cycle of period $T = 2\pi/\Omega$ is

$$N_i(l) \int_{(l-1)T}^{lT} R[\mathbf{r}_i(t);\, \dot{\mathbf{r}}_i(t)]\mathrm{d}t \qquad [18]$$

In the simulations, it is assumed that these N_i spontaneous photons are emitted one by one, at the end of each rf cycle, rather than emitting them one by one, according to the appropriate photon-statistics during the cycle. The orientation of the unit vector $\hat{\mathbf{u}}_j$, which points in the direction of the j-th emitted photon, is chosen at random, but weighted with a $\cos^2$ distribution (dipole characteristic) with respect to the laser polarization axis. This way, only the linearly polarized spontaneously emitted photons have been simulated in the calculations. The circularly polarized ones, which occur with a $\sin^2$ distribution, have been neglected. The random force $\mathbf{F}_i^{(\text{rand})}$, is thus:

$$\mathbf{F}_i^{(\text{rand})}(t) = \hbar k \sum_{l=-\infty}^{\infty} \delta(t - lT) \sum_{j=1}^{N_i} \hat{\mathbf{u}}_j \qquad [19]$$

In some of the calculations, a contact potential has been taken into account which is generated experimentally by a coating of the quadrupole electrodes by the trapped particles. Close to the center of the trap, the contact potential gives rise to an approximately homogeneous electric field. Altogether, the motion of the *i*-th particle is governed by the sum of the above forces, that is:

$$m\ddot{\mathbf{r}}_i = \mathbf{F}_i^{(\text{trap})} + \mathbf{F}_i^{(\text{Coul})} + \mathbf{F}_i^{(\text{laser})} + \mathbf{F}_i^{(\text{rand})} + \mathbf{F}_i^{(\text{cont})} \qquad [20]$$

Given appropriate initial conditions, the eqns of motion [20] are integrated forward in time to obtain the trajectories of the *n*-ion system. These trajectories provide all the information necessary to find the quantities of interest, such as the occupation probability density of the *i*-th ion in phase space (i.e., the probability density of finding the *i*-th ion at a given position, and with a given velocity), the (total) kinetic energy, and the fluorescence intensity.

Some of the results of the described simulations will be summarized below. Figure 1 shows the excitation spectrum of the ions as a function of laser detuning. The experiment confirms the results of the simulations.

Owing to the Coulomb interaction of the ions at particular phases of the cooling process, chaos in the ion cloud is observed leading to a heating process of the ions by the applied field of the trap. This heating mechanism was investigated in the simulations in detail and the summary of the results is given in Figure 2.

For zero laser power and large *r*, there is no net heating of the ions. This is confirmed by the experiments, in which, even in the absence of a cooling laser, large clouds of ions can be stored in a Paul trap over several hours without being heated out of the trap. The ions are far apart, the Coulomb force is small, and on short time-scales the ions behave essentially like independent singly stored ions. For this reason, this part of the heating diagram is called the 'Mathieu regime' as it is well described by the Mathieu equation [7]. Turning on a small laser power, the rms radius *r* reduces drastically due to laser

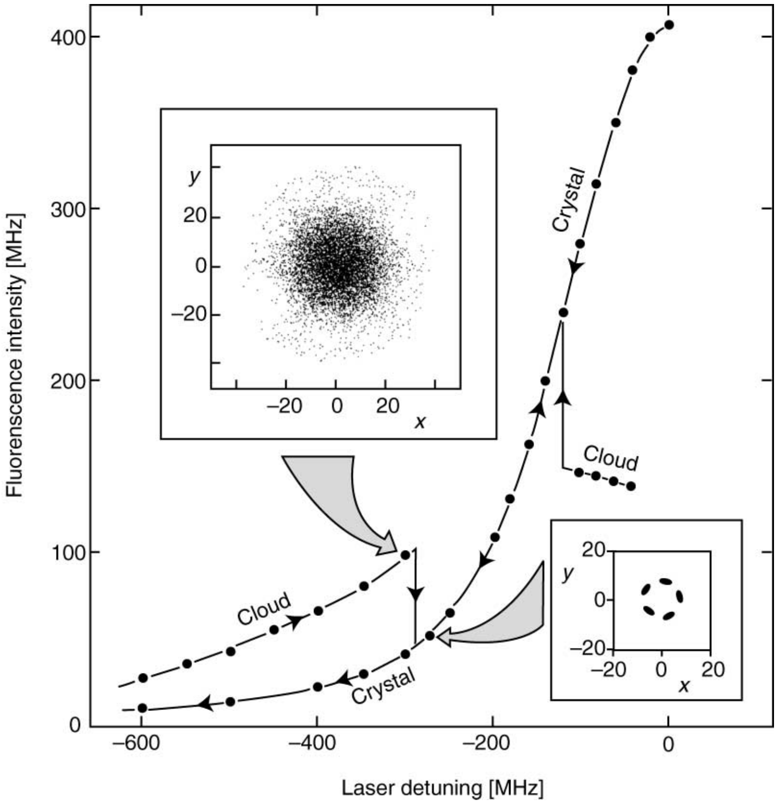

Figure 1 Excitation spectrum of five ions as a function of laser detuning. When the laser detuning is changed the cooling rate of the ion is varying. Two phase transition points (and bistability) are apparent. Insets show the results of the molecular dynamics calculations shortly before and after the jump, which demonstrates directly the existence of a transition from a cloud phase to a crystalline state. The units given at the insets are µm.

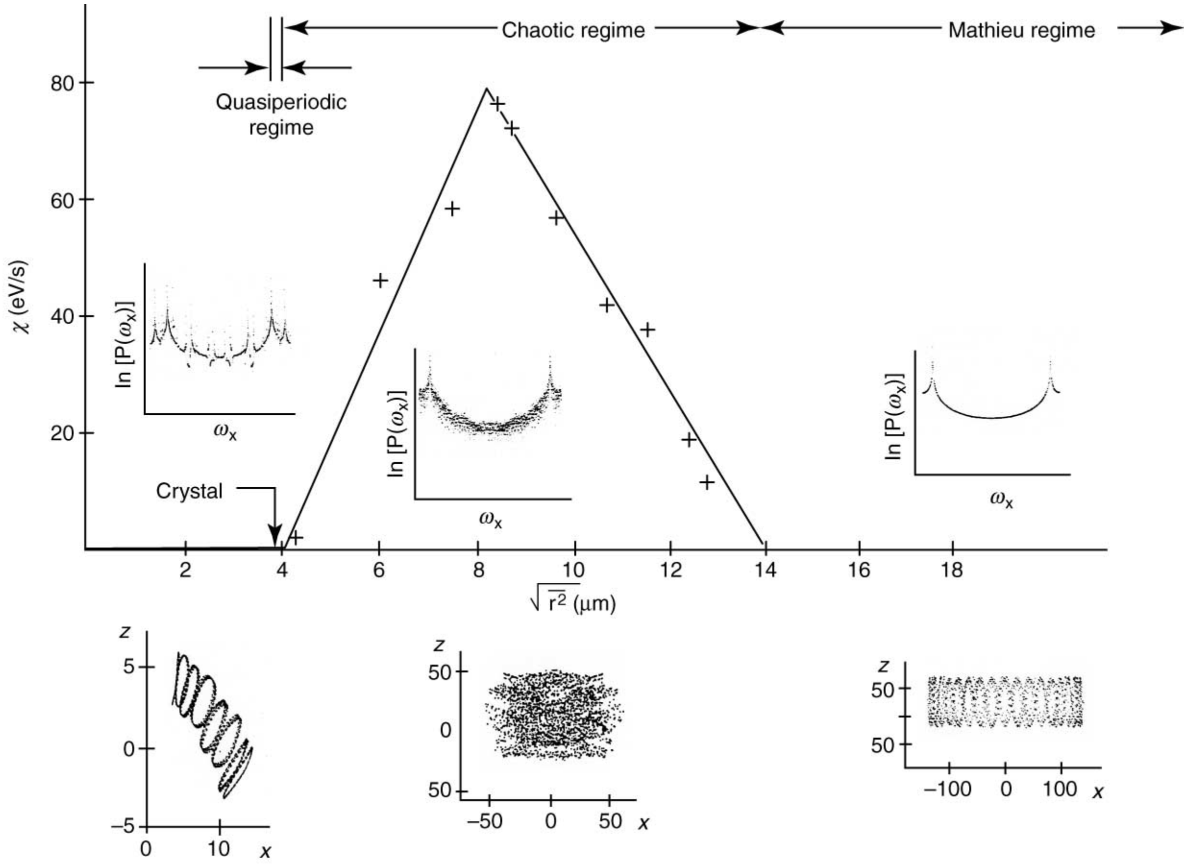

Figure 2 Average heating rate of five ions in a Paul trap versus the rms radius of the ion configuration. The insets show the power spectrum and the corresponding stroboscopic Poincaré sections in the x–z plane of relative separation for two ions in three characteristic domains: the quasiperiodic regime, the chaotic regime and the Mathieu regime. All length scales are in units of micrometers. Units gives at the x–z axis of the insets are in μm.

cooling, but comes to a halt at about 14 μm where chaotic rf heating sets in and balances the cooling power of the laser. Increasing the laser power results in an even smaller cloud. The smaller cloud produces more chaotic rf heating, as seen clearly by the negative slope of the heating curve (Figure 2) in the range $8 < r < 14$ μm.

In the range 4 μm $< r <$ 8 μm, there is still chaotic heating, but the slope of the heating curve is positive. As a consequence of the resulting triangular shape of the heating curve, at about 8 μm, the chaotic heating power can no longer balance the cooling power of the laser, and the cloud collapses into the 'crystalline' state located at $r \approx 3.8$ μm; a 'phase transition' has occurred. Due to this collapse, the behavior of the heating rate in the range 3.8 μm $< r <$ 8 μm cannot be studied by equilibrating laser cooling and rf heating. In this case, we start out from the crystal state and slightly displace the ions to explore the 'vicinity' of the crystal. We observe no heating for 3.8 μm $< r <$ 4 μm, but quasiperiodic motion, and thus name this regime the 'quasi-periodic' regime. We call the upper edge of the quasiperiodic regime ($r \approx 4$ μm) the 'chaos threshold'. An initial condition beyond the chaos threshold, i.e., satisfying $r > 4$ μm, leads to heating and expansion of the ion configuration and numerical data relevant for the shape of the heating curve can be taken during this 'explosion phase'. The laser power P is set to zero for this type of experiment. We conjecture that, apart from the trivial case of a single stored ion, the heating curve is 'universal', i.e., its qualitative shape, including the existence of the chaotic regime, does not depend on the number of simultaneously trapped ions.

For the quasiperiodic, the chaotic and the Mathieu-regime, respectively, the corresponding type of power spectrum are shown in the insets above the abscissa of Figure 2. The data were actually taken from the case of two ions, but would look similar in the five-ion case. A discrete spectrum in the quasi-periodic regime and a complicated

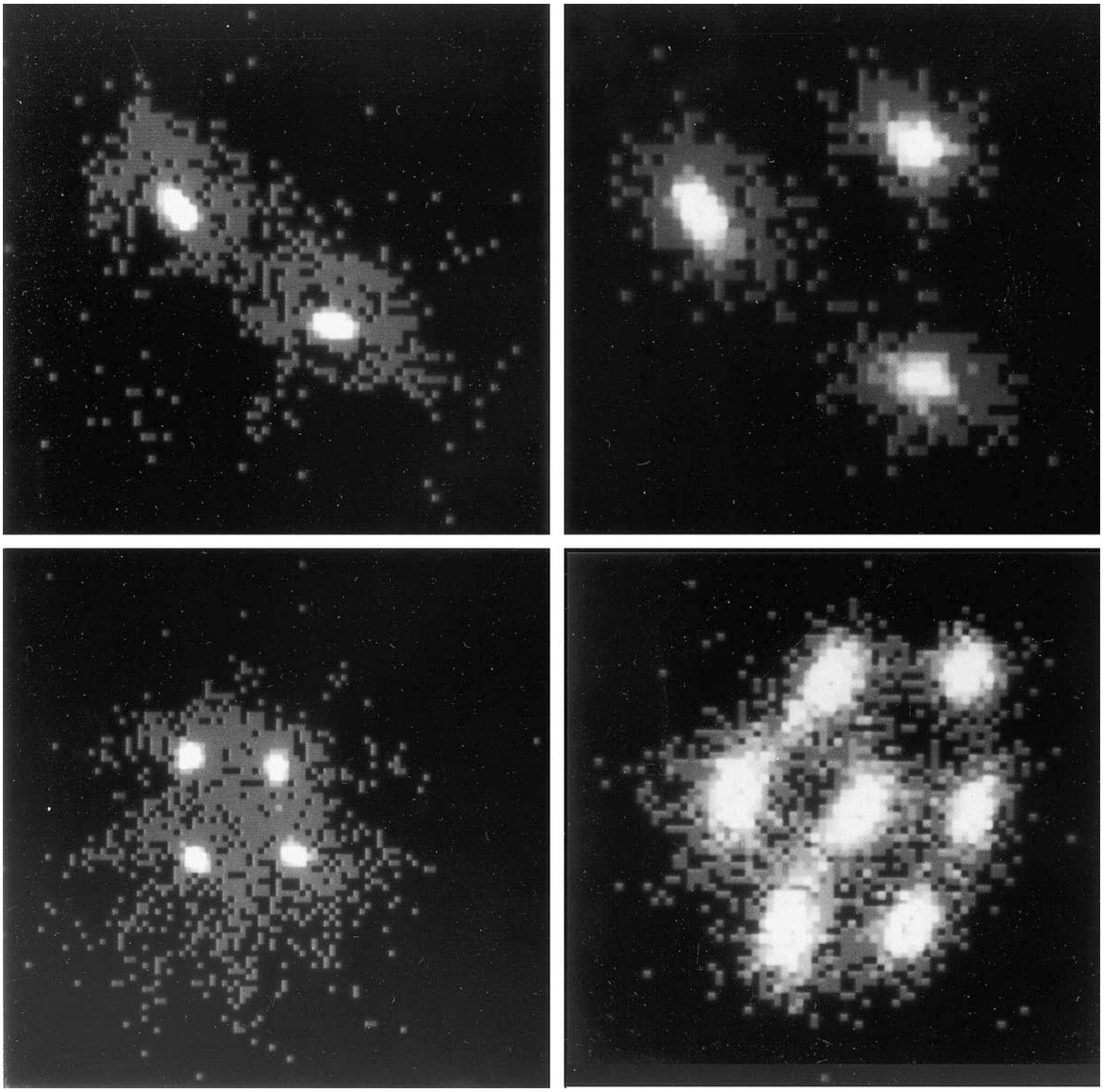

Figure 3 Ordered ion configuration of two, three, four, and seven laser-cooled Mg^+ ions confined in a Paul trap and observed using resonantly scattered light. The average ion separation is 20 μm. The structures of four and seven ions are slightly deformed owing to contact potentials in the trap.

noisy spectrum in the chaotic regime is obtained. The spectrum in the Mathieu regime is again quite simple and dominated by the secular motion frequency. Stroboscopic pictures of the locations of two ions in the $x-y$ plane of the trap characterizing the three regions, are shown on the insets below the abscissa of Figure 2.

Some of the ordered ion structures observed experimentally are shown on Figures 3 and 4. The structures in Figure 3 were obtained in a Paul trap; the others are measured in a storage ring leading to quasi-linear structures of the ion configurations, depending on the symmetry of the trap. Ordered ion structures have also been observed in Penning traps. Due to the symmetry of this trap rotationally symmetric structures are observed.

Conclusion

Laser cooling of ions in traps provides an excellent tool for numerous experiments, varying from tests of basic physics to important and new applications in quantum information processing. The linear ion structure shown on Figure 4 is presently investigated in many laboratories in order to realize gates for

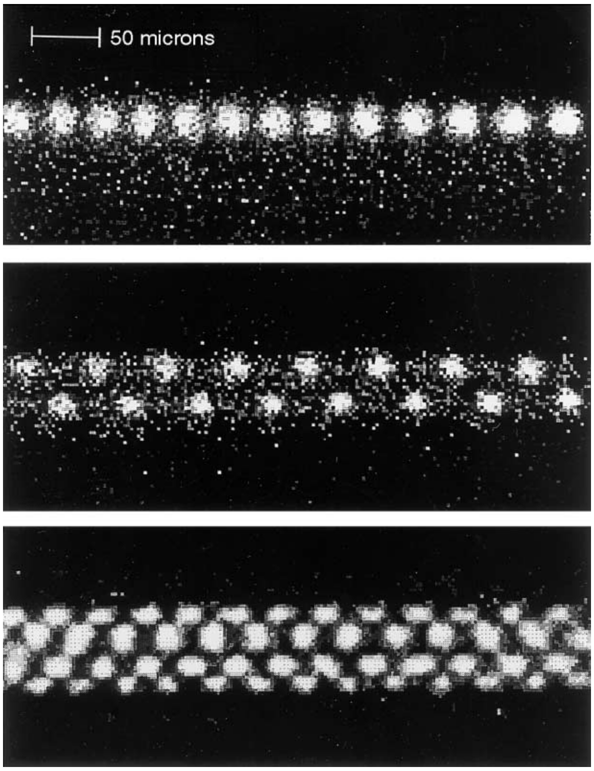

Figure 4 Crystalline structures of laser-cooled $^{24}Mg^{+}$ ions in the quadrupole storage ring. At a low ion density the ions form a string along the field axis (upper). Increasing the ion density transforms the configuration to a zigzag (middle). At still higher ion densities, the ions form ordered helical structures on the surface of a cylinder, e.g., three interwoven helices (lower). As the fluorescent light is projected onto the plane of observation in this case the inner spots are each created by two ions seated on opposite sides of the cylindrical surface, resulting in a single, bright area. (Reproduced by permission of *Nature* from Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^{+}$ ions in a quadrupole storage ring. *Nature* 357: 310–313.)

quantum computing. Single trapped ions in cavities may be used as deterministic light sources for secure single photon communication links. Furthermore, in many laboratories single laser-cooled ions are investigated as the basis for new time and frequency standards, being several orders of magnitude more accurate than our present atomic clock.

See also

Chaos in Nonlinear Optics. **Quantum Optics:** Entanglement and Quantum Information; Quantum Computing with Atoms.

Further Reading

Arimondo E, Philips WD and Strumia F (eds) (1992) Laser manipulation of atoms and ions. *Proceedings of the International School of Physics Enrico Fermi, Course CXVIII*. Amsterdam, Oxford, New York, Tokyo: North Holland.

Ashkin A (1970) Atomic-beam deflection by resonance-radiation pressure. *Physical Review Letters* 25: 1321–1324.

Birkl G, Kassner S and Walther H (1992) Multiple shell structures of laser cooled $^{24}Mg^{+}$ ions in a quadrupole storage ring. *Nature* 357: 310–313.

Blümel R, Chen JM, Peik E, Quint W, Schleich W, Shen YR and Walther H (1988) Phase transitions of stored laser-cooled ions. *Nature* 334: 309–313.

Blümel R, Kappler C, Quint W and Walther H (1989) Chaos and order of laser-cooled ions in a Paul trap. *Physical Review A* 40: 808–823.

Bollinger JJ, Kriesel JM, Mitchell TB, King LB, Jensen MJ, Itano WM and Dubin DHE (2003) Laser-cooled ion plasmas in Penning traps. *Journal of Physics B: Atomic Molecular Optical Physical* 36: 499–510.

Cirac JI, Garay LJ, Blatt R, Parkins AS and Zoller P (1994) Laser cooling of trapped ions: the influence of micro-motion. *Physical Review A* 49: 421–432.

Dehmelt H (1967 and 1969) In: Bates DR and Estermann I (eds) *Advances in Atomic and Molecular Physics*, vols. 3 and 5. New York: Academic Press.

Dehmelt HG (1983) Stored ion spectroscopy – three lectures. In: Arecchi FT, Strumia F and Walther H (eds) *Advances in Laser Spectroscopy*. New York: Plenum.

Diedrich F, Peik E, Chen JM, Quint W and Walther H (1987) Observation of a phase transition of stored laser-cooled ions. *Physical Review Letters* 59: 2931–2934.

Glauber R (1991) In: Arimondo E and Phillips W (eds) *Proceedings of the International School of Physics 'Enrico Fermi,'* vol. 118. Amsterdam: North Holland.

Hänsch TW and Schawlow AL (1975) Cooling of gases by laser radiation. *Optics Communications* 13: 68–69.

Javanainen J, Lindberg M and Stenholm S (1984) Laser cooling of trapped ions: dynamics of the final stages. *Journal of Optical Society of America B* 1: 111–115.

Leibfried D, Blatt R, Monroe C and Wineland D (2004) Quantum dynamics of single trapped ions. *Review of Modern Physics* 75: 281–324.

Peik E, Abel A, Becker Th, von Zanthier J and Walther H (1999) Sideband cooling of ions in radio-frequency traps. *Physical Review A* 60: 439–449.

Schieder R, Walther H and Wöste L (1972) Atomic beam detection by the light of a tunable dye laser. *Optics Communications* 5: 337–340.

Wineland D and Dehmelt H (1975) Proposed 10–14 laser fluorescence spectroscopy on Tl^{+} mono-ion oscillator. *Bulletin of the American Physical Society* 20: 637.

Quantum Computing with Atoms

S F Huelga, University of Hertfordshire, Hatfield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

It is a well-known fact that the speed of commercial microprocessors is doubling, roughly, every 18 months (Moore's law). This speedup is achieved mainly at the expense of reducing the size of the microchip. If this tendency is to continue, it seems unavoidable that quantum mechanical effects will eventually become important.

Remarkably, far from being a technological nuisance, the fact that information carriers are truly quantum mechanical systems offers an entirely new paradigm for computation. As opposed to classical bits, with logical values 0 and 1, quantum bits (qubits) can be prepared in arbitrary, normalized superpositions of the logic values. In addition, quantum states of multiqubit systems can be entangled, i.e., cannot be written as tensor products of states of the individual components. The theoretical description of entanglement requires a huge number of parameters which, in fact, grows exponentially with the number of particles that are involved. This exponential number of new degrees of freedom is responsible for the computational potential that quantum systems offer. Recently it has been demonstrated theoretically how we can harness this potential in the form of quantum algorithms, including Shor's factorization algorithm and Grover's protocol for data base search.

From a theoretical point of view, the implementation of quantum computation is greatly simplified by the fact that any unitary transformation involving an arbitrarily large number of qubits can be decomposed into two building blocks: single qubit operations and two-qubit controlled-NOT gates (CNOT-gates). We only need to identify two-level systems whose interaction can be mediated in a way that allows for conditional dynamics, that is, the second system undergoes certain operation (a NOT operation in our case), conditional on the state of the first qubit. However, it is easy to convince ourselves that from the experimental point of view, the challenge is going to be formidable. The following (DiVincenzo)-checklist summarizes the necessary requirements for the potential physical realization of a quantum computer:

- The system needs to provide a well-defined qubit, with two robust levels to represent the logical values 0 and 1. Moreover, the system should be scalable.
- Both the initial preparation and the final readout of the system has to be done with sufficient accuracy.
- The system has to allow the implementation of a universal set of gates (single-qubit rotations and CNOTs gates) with sufficient high fidelity. Implicit in this requirement is the ability to address individual qubits.

From the above list it is clear that the practical realization of a quantum computer relies on the ability to achieve a delicate equilibrium: the system needs to interact strongly with selected external agents, typically laser light, both during the gate operations and the preparation and readout steps, but should be shielded from any unwanted interaction if the desired dynamics is to be preserved. Perfect, selective, shielding from the environment is not possible in practice and unwanted interactions with the surroundings will unavoidably result in some form of decoherence. In this article we will review the fundamentals of the implementation of a universal set of quantum gates using (i) atomic system coupled to a resonant field sustained in a microwave cavity and (ii) trapped ions confined in an electromagnetic trap. Atomic and ionic systems provide a good means for storing quantum information in long-lived electronic states. In addition, both neutral atoms and ions can be selectively detected with high efficiency. While none of the presented schemes is expected to be scalable to a very large number of qubits, they constitute the first prototypes for quantum registers and are expected to play an important role in applications where only a moderate number of qubits are required. These applications range from the development of new frequency standards to quantum communication protocols. We will discuss in the following sections how a set of universal gates can build within these systems and will present the state of the art of the experiments performed to date.

Quantum Computation: Basic Definitions

As stated before, the basic unit of quantum information is a qubit, an arbitrary quantum mechanical two-level system. We will associate logical values 0 and 1 with quantum states $|0\rangle$ and $|1\rangle$, the eigenvectors of an observable such as the energy, and refer to such states as the computational basis. Any unitary operation on a qubit will be called a single qubit gate. Joint operations on more than one qubit will be denoted as multi-qubit gates. Using the fact that any two-level system is isomorphic to a spin-1/2 particle (e.g., an electron), we can always think of a qubit in terms of a certain pseudo-spin whose dynamics can be represented as rotation of the associated Bloch vector within the Bloch sphere.

For instance, as depicted in Figure 1, a qubit initially prepared in state $|1\rangle$, after experiencing a single qubit gate will be represented by certain Bloch vector $\vec{S}$, whose location on the Bloch sphere is a function of the gate parameters. Note that a classical system can only assume two positions (upwards and downwards), while a quantum system can explore the whole surface of the sphere.

It is also quite standard in quantum computation to use a diagramatic representation in terms of quantum circuits. Quantum circuits are built up of wires to denote the quantum bits and boxes to represent the gates, or quantum operations. Figure 2 shows two basics elements of a quantum circuit. The element on the left is the so-called Hadamard transformation, a particular example of single-qubit gate which transform each element of the computational basis as follows:

$$
\begin{aligned}
|0\rangle &\rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) \\
|1\rangle &\rightarrow \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right)
\end{aligned}
\qquad [1]
$$

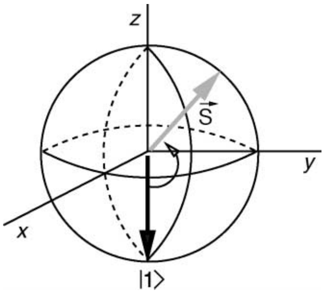

Figure 1 Geometrical representation of qubit systems using the Bloch sphere formalism. See the text for details.

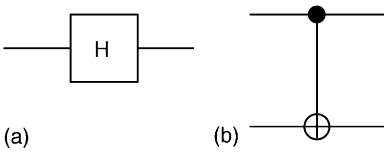

(a) (b)

Figure 2 Examples of single qubit gates (a) Hadamard transformation and two qubit gates (b) CNOT gate as customarily represented in quantum circuits diagrams.

Part (b) of Figure 2 depicts a CNOT-gate. This is now an operation involving two qubits. The first one, the control qubit, remains unaltered after the gate operation while the second qubit, referred to as the target, undergoes a spin flip (logical NOT operation) only when the control is in the state $|1\rangle$. Specifically, the action of a CNOT-gate is defined by the following transformation rule:

$$
\begin{gathered}
\mathrm{CNOT} \\
|00\rangle \rightarrow |00\rangle \\
|01\rangle \rightarrow |01\rangle \\
|10\rangle \rightarrow |11\rangle \\
|11\rangle \rightarrow |10\rangle
\end{gathered}
\qquad [2]
$$

We have already mentioned that single qubit rotations and CNOT gates provide a universal set of gates. Our aim now is to show how such operations can be built when the so far generic qubit are realized in terms of real atomic systems. In the two experimental scenarios we are going to analyze here, the CNOT operation will be achieved by means of a different type of gate, a controlled-phase gate supplemented by two one-qubit Hadamard gates on the target qubit. The arrangement is showed in Figure 3. The controlled-phase gate is again a two-qubit operation defined by the following transformation rule:

$$
\begin{gathered}
\mathrm{CNOT} \\
|00\rangle \rightarrow |00\rangle \\
|01\rangle \rightarrow |01\rangle \\
|10\rangle \rightarrow |10\rangle \\
|11\rangle \rightarrow -|11\rangle,
\end{gathered}
\qquad [3]
$$

i.e., a global phase equal to π is acquired if, and only if, the two basis states are in the logical state 1. It is easy to check that when the target bit is rotated by a Hadamard gate prior and subsequently to the action

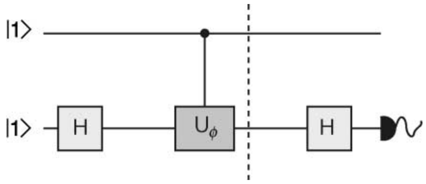

Figure 3 Controlled-phase gate with both inputs, control, and target qubits, in state $|1\rangle$. The first Hadamard gate rotates the target to state $(|0\rangle - |1\rangle)/\sqrt{2}$. The action of the controlled-phase gate with $\phi = \pi$ yields the joint state $|1\rangle + (|0\rangle - |1\rangle)/\sqrt{2}$ and a final Hadamard rotation on the target leaves the system in the product state $|1\rangle|0\rangle$.

of the joint phase gate, the resulting state is tantamount to the action of a single CNOT gate. Therefore, to prove that our atomic system allows for the implementation of a universal set of gates, it suffices to show that both single qubit rotations and controlled phase gates can be implemented with sufficient accuracy.

The next section analyzes a typical cavity QED setup and presents the basic physical ideas to understand how single and two-qubit gates can be implemented within atomic systems interacting with radiation fields. Subsequent sections will rely heavily in this analysis.

Quantum Computing with Atoms I: Cavity QED

A simplified cavity-QED arrangement for performing both single- and two-qubit gate operations is depicted in Figure 4. This configuration is well-known to quantum opticians and it is often referred to as a Ramsey interferometric setup. For concreteness, we will focus here in the experimental set up currently in operation at the ENS in Paris. There, the atomic system is a highly excited Rydberg atom initially prepared in a long-lived circular state. The relevant atomic states are represented in Figure 5, with the qubit being stored in levels $|i\rangle$ and $|g\rangle$ which will play the role of the general basis states $|0\rangle$ and $|1\rangle$.

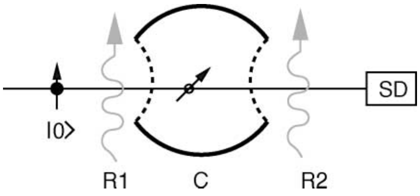

Figure 4 Cavity QED setup for single-qubit and two-qubit operations. Rotations of the atomic pseudo spin are performed via interation with classical radiation (Zone R), while the interaction with a resonant cavity field sustained in C allows the realization of a phase gate where the cavity photon acts as the control qubit.

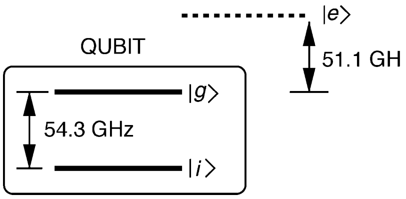

Figure 5 Circular Rydberg levels involved in the ENS experimental setup. The qubit is stored in levels $|i\rangle$ (logical value 0) and $|g\rangle$ (logical value 1). The radiative lifetime of these levels lies in the millisecond range. Dissipation can then be neglected along the 20 cm path inside the apparatus.

Regions denoted by R represent the interaction of the atomic system with classical microwave radiation. As detailed below, this interaction will allow arbitrary rotations of the atomic pseudo spin. Region C represents the interaction of the atomic system with a single mode cavity field resonant the atomic transition $g - e$. The highest Q value reported to date is of the order of 10^8, which corresponds to a photon storage time of a few milliseconds. Therefore, the relaxation time of the cavity field is much longer than the atom-cavity interaction time, which defines the two-qubit gate operation time and is of the order of a few tens of μs. The cavity field can be prepared in either of the Fock states $|0\rangle_\gamma$ or $|1\rangle_\gamma$ and will play the role of the control qubit for the realization of a phase gate, as detailed below. This set up is completed with a detection scheme (SD in the figure) that allows the selective detection of the atomic electronic state.

We will start our discussion showing how to perform an arbitrary single-qubit gate by means of irradiating the atomic system with classical radiation. We will keep the analysis in very general terms with the aim of using it later in the description of the implementation of quantum logic with trapped ions.

Single-Qubit Rotations

Let our qubit be a two-level atomic system with frequency separation ω_0. We will show that an arbitrary rotation can be achieved by means of irradiating the system with classical light, e.g., microwave radiation or a laser field, with tunable frequency ω_L. The radiation field will be specified by two controllable parameters, its Rabi frequency Ω and its phase ϕ. The atom-field Hamiltonian can be written as

$$\begin{aligned} H = {} & \frac{\hbar\omega_0}{2}\left(|1\rangle\langle 1| - |0\rangle\langle 0|\right) \\ & + \hbar\left(g e^{i\omega_L t}|0\rangle\langle 1| + g^* e^{-i\omega_L t}|1\rangle\langle 0|\right) \end{aligned} \qquad [4]$$

where the first term corresponds to the atomic internal energy and the second describes the coupling with the external field. The coupling constant g can be written in terms of the field parameters as

$$g = \frac{i\Omega}{2} e^{-i\phi} \qquad [5]$$

In order to write down a time-independent Hamiltonian, we will move to an interaction picture with respect to $H_0 = \hbar\omega_L/2(|1\rangle\langle 1| - |0\rangle\langle 0|)$. In this frame,

$$H_{\text{int}} = \frac{\hbar\delta}{2}\Big(|1\rangle\langle 1| - |0\rangle\langle 0|\Big) + \hbar\Big(g|0\rangle\langle 1| + g^*|1\rangle\langle 0|\Big) \quad [6]$$

The interaction Hamiltonian can be readily expressed in terms of the Pauli spin operators

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad [7]$$

as

$$H_{\text{int}} = \frac{\hbar\delta}{2}\sigma_z + \frac{\hbar\Omega}{2}(\sin\phi\,\sigma_x - \cos\phi\,\sigma_y) \quad [8]$$

where $\delta = \omega_L - \omega_0$ denotes the detuning from the atomic transition and we have taken into account the explicit form of g as given by eqn [5]. We can now easily derive the time evolution operator by noting that H_{int} can be written in a compact form as $H_{\text{int}} = \vec{\sigma}\cdot\vec{n}$, where the vector $\vec{n}$ has cartesian coordinates $\vec{n} = (\Omega/2\sin\phi, -\Omega/2\cos\phi, \delta/2)$. Note that $|\vec{n}| \equiv \Omega_g = 1/2\sqrt{\Omega^2 + \delta^2}$. The time evolution operator can then be written as follows:

$$U = e^{-iH_{\text{int}}t} = \cos(|\vec{n}|t)\mathbb{1} - i\sin(|\vec{n}|t)\vec{\sigma}\cdot\vec{n}/|\vec{n}| \quad [9]$$

In matrix form:

$$U = \begin{pmatrix} \cos\left(\frac{\Omega_g t}{2}\right) - i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) & \frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi} \\ -\frac{\Omega}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi} & \cos\left(\frac{\Omega_g t}{2}\right) + i\frac{\delta}{\Omega_g}\sin\left(\frac{\Omega_g t}{2}\right) \end{pmatrix} \quad [10]$$

In the limit of large Rabi frequency, where $\Omega \gg \delta$, we can write the following approximate expressions for the time evolution of the computational basis states

$$\begin{aligned} U|0\rangle &= \cos\left(\frac{\Omega_g t}{2}\right)|0\rangle - \sin\left(\frac{\Omega_g t}{2}\right)e^{i\phi}|1\rangle \\ U|1\rangle &= \sin\left(\frac{\Omega_g t}{2}\right)e^{-i\phi}|0\rangle + \cos\left(\frac{\Omega_g t}{2}\right)|1\rangle \end{aligned} \quad [11]$$

The transformed states by U are eigenstates of the spin operator along the direction $\vec{n}$ which is a function of the external field parameters. Therefore, by adjusting the Rabi frequency and the field's phase, it is possible to transform the basis states into spin eigenstates along any arbitrary direction, and to perform an arbitrary rotation of our qubit, we simply have to tailor appropriately the parameters of the driving field. For example, when the Rabi frequency and the interaction time are chosen such that $\Omega t/2 = \pi/4$, and the field's phase is taken equal to π, we get

$$\begin{aligned} U|0\rangle &= \frac{1}{2}\Big(|0\rangle + |1\rangle\Big) \\ U|1\rangle &= \frac{1}{2}\Big(|0\rangle - |1\rangle\Big) \end{aligned} \quad [12]$$

This relationship defines the Hadamard transformation H introduced before. Alternatively, using quantum optics jargon, we have performed a $\pi/2$-pulse. Note that using the Bloch sphere representation, we can visualize the action of a Hadamard gate on the computational basis states as a rotations that brings them to the equatorial plane of the Bloch sphere, the transformed states by H now lying in opposite directions along the x-axis.

The next step is to show that our system also allows the implementation of conditional dynamics and, in particular, allows to build a controlled-phase gate.

Conditional Dynamics: Controlled-Phase Gate

The interaction of the atomic system with the cavity field sustained by the microwave cavity is described by the Jaynes–Cummings (JC) Hamiltonian. We can see this Hamilton operator as a generalization of the operator defined by eqn [4], where the radiation field is now quantized and the terms $e^{\pm i\omega_L t}$ are replaced by the operators a and $a^\dagger$ corresponding to the annihilation and creation of a cavity photon of frequency ω. The interaction with the quantized cavity field couples levels $|e\rangle$ and $|g\rangle$, while the far detuned level $|i\rangle$ will now play the role of a passive spectator. Remember that our atomic qubit is stored in levels $|i\rangle$ and $|g\rangle$. To avoid confusion with the notation, we will write the Hamilton operator in terms of the atomic states e and g. Explicitly:

$$\begin{aligned} H_{\text{JC}} = {} & \frac{\hbar\omega_0}{2}\Big(|e\rangle\langle e| - |g\rangle\langle g|\Big) + \hbar\omega a^\dagger a \\ & - i\frac{\hbar\Omega}{2}\Big(a|g\rangle\langle e| - a^\dagger|g\rangle\langle e|\Big) \end{aligned} \quad [13]$$

where the coupling constant g has been taken to be $g = -i(\hbar\Omega/2)$, proportional to the Rabi frequency averaged over the mode structure, Ω. Using the same techniques employed in the previous section, we can derive the corresponding JC time evolution operator, which yields, for exact resonance, the following transformation rules:

$$\begin{aligned} &|g\rangle|0\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\ &|g\rangle|1\rangle_\gamma \rightarrow \cos\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma - \sin\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma \\ &|e\rangle|0\rangle_\gamma \rightarrow \cos\frac{\Omega t}{2}|e\rangle|0\rangle_\gamma + \sin\frac{\Omega t}{2}|g\rangle|1\rangle_\gamma \end{aligned} \qquad [14]$$

where states $|\rangle_\gamma$ refer to the cavity field. Note that:

- The cavity field will play the role of our control qubit in the performance of a phase gate, the target qubit being the atomic system with computational states defined as before. The computational basis for the photonic qubit is provided by the Fock states $|0\rangle_\gamma$ and $|1\rangle_\gamma$.
- Strictly speaking, the time t entering in the equations above corresponds to an effective interaction time which accounts for the spatial variation of the coupling along the field mode.
- When the atom enters the cavity in state $|i\rangle$, it will acquire a global phase proportional to the ratio Ω^2/δ_i, where δ_i denotes the detuning between the atomic transition $e \leftrightarrow i$ and the mode frequency ω. In conditions of resonance, or quasi-resonance, this factor can be safely neglected, which justifies the term *passive spectator* used before.

We now have all the ingredients to understand how we can perform a controlled phase gate. Consider the situation where $\Omega t = 2\pi$, so the system experiences a full Rabi oscillation. According to the previous discussion, the basis states of two-qubit system atom-cavity field will transform as follows:

$$\begin{aligned} &|i\rangle|0\rangle_\gamma \rightarrow |i\rangle|0\rangle_\gamma \\ &|i\rangle|1\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\ &|g\rangle|0\rangle_\gamma \rightarrow |g\rangle|0\rangle_\gamma \\ &|g\rangle|1\rangle_\gamma \rightarrow -|g\rangle|1\rangle_\gamma \end{aligned} \qquad [15]$$

That is, whenever the atomic system enters the cavity in state $|i\rangle$, it will be left unaffected, irrespective of the photon state. However, if the atom enters the cavity in state $|g\rangle$, the system will undergo a phase shift equal to π conditional to the photon number. And that is precisely the dynamics required for a controlled phase gate as defined earlier.

A detailed account of the experiments in cavity QED, showing coherent gate operation, can be found in Further Reading. These experiments prove the working principle of a universal set of quantum logic gates with neutral atoms and provide the basic elements for the development of a cavity QED quantum register in the near future.

Quantum Computing with Atoms II: Ion Traps

Charged atomic particles, ions, provided the first atomic system where conditional dynamics was demonstrated experimentally. The theoretical proposal in 1995, by Cirac and Zoller, for performing a two-qubit controlled NOT gate between trapped ions cooled to their motional ground state, was followed a few months later by a pioneering experiment at NIST, in Colorado. While the physics involved in the description of laser cooled and trapped ions is conceptually more involved than the dynamics ruled by a purely JC Hamiltonian, we can review the rudiments of the Cirac–Zoller scheme on the basis of the techniques developed in the previous sections. Nevertheless, as discussed briefly at the end of this section, we should point out that the exact procedure followed in the actual experiments departs from the analysis presented here.

Figure 6 depicts a linear ion trap where a string of ions is confined in an essentially 1-D structure. Quantum bits are stored in two long-lived hyperfine states which can be driven with laser light of controllable Rabi frequency Ω and phase ϕ. Ideally, the separation d between ions exceeds the wavelength of the external field so that ions can be addressed individually. Detection of the atomic state relies on the well mastered quantum jump technique which provides an efficiency close to 100%. From our previous discussion, it is clear that we already have all the ingredients for performing arbitrary single qubits gates, but we still need to provide a procedure to achieve two-qubit gate operations. As detailed in the next section, this will be achieved using the motional degrees of freedom of the ions as a bus for information transfer between an arbitrary pair of ions.

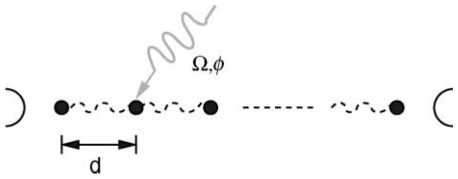

Figure 6 Schematic representation of a linear ion trap. Ions are individually addressed by external laser fields to perform single qubit operations, while two-qubit gates are mediated by the coupling to the motional degrees of freedom, depicted here as mechanical strings linking the chain of ions in the trap. The exact duration of a two-qubit gate operation and the characteristic decoherence time depend very much on the trap design, but orientative values are in the range of 50 μs and 100 − 1000 μs respectively.

The Cirac and Zoller Quantum Computer

At sufficiently low temperatures, the process of absorption and emission of photons produces a non-negligible recoil of the atomic system and the ion's motion can no longer be treated classically. In this regime, we can think of the string of ions as a system of coupled quantum harmonic oscillators whose normal mode configurations correspond to the collective motional states of the ions. The lowest energy solution corresponds to an oscillation where all ions are in phase (center of mass (CM) mode) and will be the one considered in our description. Figure 7 depicts the level scheme for ions initially cooled down to the motional ground state, illustrating both internal levels, $|e\rangle$ and $|g\rangle$, and motional (vibrational) states, denoted by $0, 1, \ldots$ Laser radiation can be tuned to be resonant with the $|g\rangle|0\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$ transition (carrier transition), but may also be tuned to drive transitions with different vibrational states (sideband transition), as represented by the external arrows in the figure. The realization of single qubit gates simply involves tuning the laser towards the carrier transition. The interaction is ruled by essentially the same Hamiltonian discussed above. Therefore, choosing appropriately the laser parameters, it is possible to induce an arbitrary rotation of the ion pseudo-spin.

To prove that the system allows the implementation of a universal set of gates, we will now show that it is possible to operate a controlled phase gate between two arbitrary ions. For that we need to couple internal and motional degrees of freedom which will be done by means of tuning the laser light to a sideband transition. The corresponding Hamiltonian will not be derived here in detail but it can be understood as describing a Jaynes–Cummings form of dynamics where the quantized electromagnetic field within a cavity is replaced by the quantum mechanical oscillator describing the vibrational states of the ion. Explicitly:

$$H = \frac{\eta\Omega}{2}\left(c|e\rangle\langle g| + c^{\dagger}|g\rangle\langle e|\right) \qquad [16]$$

where c and $c^{\dagger}$ denote the annihilation and creation operators of vibrational quanta and η is the so-called Lamb–Dicke parameter, a measure of the recoil energy. The gate operation involves the following sequence:

- Mapping the internal state of the control ion into the collective motional state of the chain,
- performing an operation on the target ion conditional to the excitation number of the vibrational state and
- reverse the initial operation restoring the vibrational state to its ground state.

Imagine ion i, which will play the role of the control, and j, the target, initially prepared in arbitrary superpositions of the corresponding internal states, with the whole system cooled down to their motional ground state $|0\rangle_{cm}$. The joint state of the two particles can be written as a product state of the form:

$$\left(\alpha_i|g\rangle_i + \beta_i|e\rangle_i\right) \otimes |0\rangle_{cm} \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [17]$$

When ion i is irradiated by a laser tuned to the first motional side band $|g\rangle|1\rangle_{cm} \leftrightarrow |e\rangle|0\rangle_{cm}$, and the timing is such that the system undergoes a π-Rabi rotation, the internal state of the ion i is mapped onto the vibrational state (Note the transformation rules given in eqn [14]. As a result, the joint state of the system is still factorizable as

$$\left(|g\rangle_i \otimes \alpha_i|0\rangle_{cm} + \beta_i|1\rangle_{cm}\right) \otimes \left(\alpha_j|g\rangle_j + \beta_j|e\rangle_j\right) \qquad [18]$$

The second step of the protocol consists of applying an operation on ion j conditional on the vibrational state of the joint system. For that we apply a laser pulse resonant with the transition $|e\rangle_j|1\rangle_{cm} \leftrightarrow |a\rangle_j|0\rangle_{cm}$ with an auxiliary level $|a\rangle_j$ so that ion j undergoes a full Rabi oscillation. This results in acquiring a phase shift equal to π only if the system is state $|e\rangle_j|1\rangle_{cm}$ (see eqn [15]). This step entangles the internal and vibrational degrees of freedom of ion j, leading to a joint state of

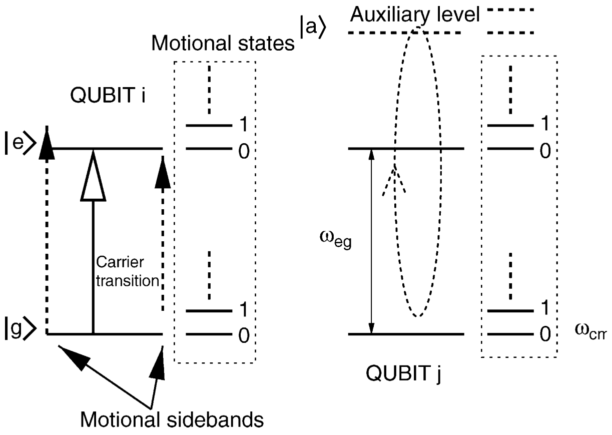

Figure 7 Relevant internal and motional atomic levels for the realization of a controlled NOT gate between two arbitrary ions within a chain. See the text for details.

the form:

$$|g\rangle_i \otimes \Big(\alpha_i|0\rangle_{cm}(\alpha_j|g\rangle_j + \beta_j|e\rangle_j) + \beta_i|1\rangle_{cm}(\alpha_j|g\rangle_j - \beta_j|e\rangle_j)\Big) \qquad [19]$$

Applying a final laser pulse on ion i identical to the initial one, the state of the vibrational mode is transferred back to ion i and the atomic systems are left in the, generally, entangled state

$$\alpha_i\alpha_j|g\rangle_i|g\rangle_j + \alpha_i\beta_j|g\rangle_i|e\rangle_i + \beta_i\alpha_j|e\rangle_i|g\rangle_j - \beta_i\beta_j|e\rangle_i|e\rangle_j, \qquad [20]$$

which completes the controlled-phase gate operation.

A simplified version of this scheme shortly followed the theoretical proposal by Cirac and Zoller, proving the working principle of the ions trap quantum processor. Since then, both theoretical and experimental progress have allowed the development of schemes for implementing faster gates as well as protocols which do not require cooling to the motional ground state. We will briefly revise the current state of the art and future prospects of this technology in the last section of this article.

Frequency Estimations as a Quantum Computation: Building Frequency Standards with Entangled States

Recently, it has been proposed to make use of entangled states of trapped ions for performing high-resolution spectroscopy beyond the shot noise limit. This boundary is the ultimate limit achievable when performing a frequency measurement using uncorrelated particles. Consider n atomic two-level systems with frequency separation ω_0 initially prepared in their ground state $|0\rangle$. The aim is to lock a certain oscillator of frequency ω_L to the atomic frequency. To do that, the currently employed protocols use the so-called Ramsey method of separate oscillating fields, where each atomic system undergoes two $\pi/2$-pulses (or Hadamard transformations) separated by a free time evolution of duration t, much larger than the duration of the pulse. Looking at the Hamiltonian describing the interaction between the ion and the field, eqn [1], one can easily see that the free time evolution can be understood as a controlled phase gate in the following sense. If we set $g = 0$ in eqn [1] and for simplicity re-scale the ion internal energies so that $H = \hbar\delta|1\rangle\langle 1|$, the free evolution operator can be written as $U = -\hbar\delta|1\rangle\langle 1|$, so that the atomic basis states evolve as:

$$\begin{aligned} U|0\rangle &= |0\rangle \\ U|1\rangle &= e^{-i\delta t}|1\rangle \end{aligned} \qquad [21]$$

In other words, the ground state remains invariant, whereas the upper levels pick up a phase which is proportional to the detuning from the atomic transition. The state of each particle after the time evolution can then be written as

$$\frac{1}{\sqrt{2}}\Big(|0\rangle + e^{-i\delta t}|1\rangle\Big) \qquad [22]$$

After the second pulse, which again mixes populations of levels $|0\rangle$ and $|1\rangle$, one measures the population of, say, level $|1\rangle$. The signal is therefore a sinusoidal function of the form

$$P_1 = \frac{1}{2}\Big(1 + \cos(\delta t)\Big) \qquad [23]$$

which is symmetrical with respect to $\delta = 0$. This procedure is repeated a number of times to accumulate a sufficient data. Let us call T the total duration of the procedure, so that we accumulate $N = nT/t$ experimental data. Finally, whenever statistics in the populations are different from 50:50 at half maximum are detected, the oscillator frequency is steered to achieve exact resonance. The ultimate resolution $\Delta\omega_0$ achievable with this technique is limited simply by projection noise and can be written as

$$\Delta\omega_0 = \frac{1}{\sqrt{nTt}} \qquad [24]$$

After time T, the oscillator's frequency is steered towards the resonance and new sequence starts. We will now illustrate how the use of entanglement can yield an uncertainty in the measurement beyond the shot noise limit eqn [24]. Consider the quantum network depicted in Figure 8. The n particles are now initially prepared in a maximally entangled state. This is achieved by means of subjecting the first ion to

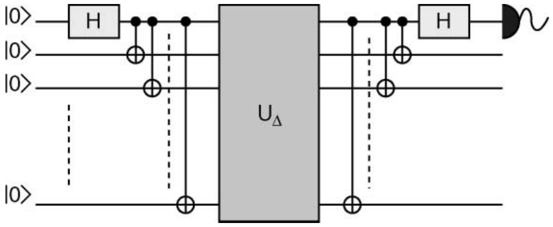

Figure 8 Quantum circuit representation for an experimental setup for a frequency estimation using maximally entangled states.

a Hadamard transformation followed by a cascade of CNOTs between the first ion, which acts as a control, and each of the remaining $(n-1)$ particles.

The resulting state is of the form

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + |1\ldots1\rangle\Big) \qquad [25]$$

The systems now evolve freely for a time t and, in light of the discussion above, we can represent this evolution in terms of a quantum gate which will leave state $|0\rangle$ invariant while introducing a phase factor proportional to the detuning if the system is in state $|1\rangle$. As a result, the joint state of the n particle after the free time evolution is given by

$$\frac{1}{\sqrt{2}}\Big(|0\ldots0\rangle + e^{-in\delta t}|1\ldots1\rangle\Big) \qquad [26]$$

with a phase factor amplified by a factor n as compared with the single particle state given by eqn [22]. The second part of the network disentangles the n particles, leaving the first ion in the state $(1/\sqrt{2})(|0\rangle + e^{-in\delta t}|1\rangle)$. If we now subject it to an additional Hadamard gate, the probability for the ion to be excited is

$$P_{1,ent} = \frac{1}{2}\Big(1 + \cos(n\delta t)\Big) \qquad [27]$$

which oscillates n times faster than the signal eqn [23]. The corresponding uncertainty for the estimation of the atomic frequency is now

$$\Delta\omega_0, ent = \frac{1}{n\sqrt{Tt}} \qquad [28]$$

This corresponds to an increase in the resolution of $\sqrt{n}$ beyond the shot noise bound. In practice, the achievable resolution using entangles states is limited by the decoherence time of the system t_{dec}, which sets the condition $T < t_{\mathrm{dec}}/2$ for the procedure to be valid. Remarkably, entanglement is still advantageous in a situation where T is chosen such that it exceeds the decoherence time. However, maximally entangled states should now be replaced by certain partially entangled ones to overcome the precision achievable with uncorrelated particles.

The application of entangled states in precision spectroscopy is a nice illustration of the potential of the ideas developed in quantum information beyond the ultimate task of building a large-scale quantum computer. The first experimental results showing the enhanced performance of entangled states have been reported by the NIST group in Boulder.

Conclusion and Perspectives

We have revised the fundamental concepts for the practical realization of quantum computation using atomic systems. We have showed how single qubit and two-qubit controlled-phase gates can be implemented using atoms interacting with the cavity field sustained by a microwave resonator and ion traps. We have illustrated how a very simple quantum computation can be used for the development of more precise frequency standards, emphasizing the potential of the ideas developed in quantum information beyond the implementation of large scale quantum computation.

Cavity QED technology has so far allowed the generation of entangled states of two and three particles and has demonstrated coherent gate operation. Ion trap technology has pushed controllable multi-particle entanglement to four particles and has showed the possibility of implementing forms of error correction. The combination of these two technologies offers interesting prospects for the near future. Surrounding traps by optical cavities and networking these devices via optical fibers is expected to allow the construction of the first prototypes for a quantum communication network.

See also

Quantum Optics: Entanglement and Quantum Information.

Further Reading

Cirac JI and Zoller P (1995) Quantum computations with cold trapped ions. *Physical Review Letters* 74: 4091.

Deutsch D, Barenco A and Ekert AK (1995) Universality in quantum computation. *Proc. Roy. Soc. London A MAT* 449: 669–677.

Grover L (1999) Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters* 79: 325–328.

Huelga SF, *et al.* (1997) Improvement of frequency standards with quantum entanglement. *Physical Review Letters* 79: 3865.

Jaynes ET and Cummings FW (1963) Comparison of quantum and semiclassical radiation theories with application to the beam maser. *Proc. IEEE* 51: 89.

Kielpinski D, Meyer V, Rowe MA, *et al.* (2001) A decoherence-free quantum memory using trapped ions. *Science* 291: 1013–1015.

Monroe C, Meekhof DM, King BE, *et al.* (1995) Demonstration of a fundamental quantum logic gate. *Physical Review Letters* 75: 4714–4717.

Raimond JM, Brune M and Haroche S (2001) *Colloquium*: manipulating quantum entanglement with atoms and photons in a cavity. *Review Modern Physics* 73: 565 and references therein.

Sackett CA, Kielpinski D, King BE, *et al.* (2000) Experimental entanglement of four particles. *Nature* 404: 256–259.

Shor P (1994) *Proc. of the 35th Annual Symposium on the Foundations of Computer Science, IEEE Computer Society, Los Alamitos*. Santa Fe, NM: IEEE Computer Society Press.

Sorensen A and Molmer K (1999) Quantum computation with ions in thermal motion. *Physical Review Letters* 82: 1971.

Wineland DJ, Monroe C, Itano WM, Leibfried D, King BE and Meekhof DM (1998) Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl. Inst. Stand. Tech.* 103: 259 and references therein.

Squeezed Phonons in Solids

A V Bragas, Universidad de Buenos Aires, Buenos Aires, Argentina

R Merlin, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The reduction of noise is crucial for measurements requiring high accuracy and, often, for the observation of new phenomena. Among noise sources, noise inherently related to the quantum nature of an object persists even after all the classical sources of noise are removed. Quantum noise is a fundamental limitation, loosely related to the Heisenberg uncertainty principle, which imposes a lower limit to the fluctuations of any pair of conjugated variables, like position and momentum. For conjugated variables A_1 and A_2 we have

$$\Delta A_1 \Delta A_2 \geq \frac{\hbar}{2} \qquad [1]$$

where

$$\Delta A_i = \sqrt{\langle A_i^2 \rangle - \langle A_i \rangle^2} \qquad i = 1, 2 \qquad [2]$$

While quantum mechanics does not forbid reducing the noise in one of the conjugated observables down to zero, the noise for the other variable will increase according to eqn [1]. Squeezing refers to the reduction of noise of a quantum variable at the expenses of its conjugated variable. The term originates from quantum optics. Nonlinear interactions of the electromagnetic field with passive and active atomic media have been successfully used to generate squeezed photons, opening new possibilities for essentially noiseless optical communications and precision measurements. The electromagnetic field is not the only quantum object that can be squeezed. By performing appropriate manipulations of the internal quantum state of a system, it is, in principle, possible to squeeze an arbitrary variable. In the past few years, vibrational states have also been experimentally squeezed in molecules and solids, by exciting the system with ultrafast optical pulses. The concept of squeezing can also be extended to an ensemble of classical objects for which the fluctuations are of thermal origin.

Squeezing applies in particular to situations where the relevant variables are the position Q and momentum P of a harmonic oscillator. Consider the Hamiltonian of a particle of mass m in a harmonic potential of frequency Ω

$$H_0 = \frac{P^2}{2m} + \frac{m}{2}\Omega^2 Q^2 \qquad [3]$$

The ground state wavefunction is

$$\varphi_0 = \frac{1}{\left[2\langle Q_0^2 \rangle \pi\right]^{1/4}} \exp\left(-\frac{Q^2}{4\langle Q_0^2 \rangle}\right) \qquad [4]$$

where $\langle Q_0^2 \rangle = \hbar/2m\Omega$. The stationary probability, $|\varphi_0|^2$, is represented in Figure 1a. This state is also a minimal uncertainty state in that $\Delta Q \Delta P \equiv \hbar/2$. $\langle Q_0^2 \rangle$ gives a measure of the quantum noise and it is known as the vacuum noise level.

Let us now discuss the behavior of a harmonic oscillator that interacts with an external source. There are two types of quantum states for which the Gaussian shape of the wavefunction is preserved after interacting with an external driving force. These are the coherent state and the squeezed state. If the interaction places the system in a state of the same form as eqn [4] but displaced with respect to the center, the wavepacket will oscillate periodically, with period $\tau = 2\pi/\Omega$, without changing its shape as shown in Figure 1b. This quantum state is known as the coherent state. It has the same uncertainty as the ground state but the expectation value $\langle Q(t) \rangle$ depends on the time following the classical trajectory in the harmonic potential. If, instead, we introduce a sudden change in the spring constant at $t = 0$, the wavepacket will begin to 'breathe' for $t > 0$, so that its width will vary periodically with a period given by $\pi/2$. This is the squeezed state. Here the noise of one of the variables is periodically reduced below the level of the vacuum noise (Figure 1c). A state that is a combination of a coherent and a squeezed state is shown in Figure 1d. The wavepacket follows the classical trajectory and breathes at the same time.

The results in Figure 1 apply only to harmonic oscillators. While the concept of squeezing can be extended to anharmonic potentials, there are additional features to consider such as, for example, wavepacket revivals.

Dynamics of a Parametrically Excited Crystal

Harmonic oscillators describe a wide range of physical systems. In particular, the dynamics of a crystal lattice can be described in terms of a set of independent harmonic oscillators, the quantum nature of which is described in terms of phonons. In this section we discuss interactions of a harmonic oscillator with external sources that produce squeezing.

The normal modes in a crystal obey equations of the form:

$$\ddot{Q}_q + \Omega_q^2 Q_q = 0 \qquad [5]$$

where Q_q and Ω_q are the phonon coordinate and frequency, and q is the wavevector.

As briefly mentioned in the previous section, a sudden change in the spring constant produces a squeeze state. To do this, we need what is known as time-dependent parametric coupling. As in the experiments on phonon squeezing, using ultrashort optical pulses, we consider an impulsive driving force proportional to Q. The relevant equation is:

$$\ddot{Q}_q + \Omega_q^2 Q_q = \lambda_q Q_q \delta(t) \qquad [6]$$

where $\delta(t)$ is the Dirac delta function, and λ accounts for the strength of the external impulsive force. This impulsive excitation induces a sudden change in momentum, proportional to the normal coordinate Q at $t = 0$, as can be shown by integrating eqn [6]. Explicitly, $\Delta\dot{Q}_q = \Delta P_q = \lambda_q Q_q(0)$. $\Delta P_q = P_q(0^+) - P_q(0^-)$, where $0^-(0^+)$ is the time immediately before (after) the excitation. This result is illustrated in Figure 2. The main effect of the interaction is to change the width of the Gaussian packet or variance of the distribution. This leads to a time-dependence of the noise defined in eqn [2], $\Delta Q^2(t) = \langle Q^2(t)\rangle$ reflecting the 'breathing' of the wavepacket. Since P_q is the conjugated variable of Q_q, $\langle P_q^2(t)\rangle$ shows the same time-dependent behavior as $\langle Q_q^2(t)\rangle$ but in quadrature with it, so as to fulfill eqn [2] at all time. Assuming an oscillator initially in its ground state, the solution of eqn [6] is

$$\langle Q_q^2(t)\rangle = \frac{\hbar}{2m\Omega_q}\left[1 + \left(\frac{\lambda_q}{\Omega_q}\right)\sin(2\Omega_q t) + \left(\frac{\lambda_q}{\Omega_q}\right)^2 \sin^2(\Omega_q t)\right] \qquad [7]$$

It is then clear that the fluctuations of Q_q periodically fall below the vacuum quantum noise level, determined by $\langle Q_0^2\rangle$, and that the oscillations have a frequency that is twice the frequency of the oscillator.

The considerations above apply to a quantum harmonic oscillator at temperature $T = 0$. In the classical case, we need to consider an ensemble of oscillators in equilibrium with a thermal bath at temperature T. It can be shown that eqn [7] also applies to the classical case, provided $\hbar/2m\Omega$ is replaced by $k_B T/\Omega^2 m$, where k_B is the Boltzman constant. In classical squeezing, the external perturbation is applied simultaneously to the ensemble, leading to a sudden change of the momentum of all the oscillators. The jump in momentum is proportional to the position of the particle at $t = 0$. After the interaction, the noise of the ensemble will fall periodically below the thermal noise at temperature T given by $k_B T/\Omega^2 m$. A pictorial representation of classical squeezing is shown in Figure 3.

To describe a more realistic situation, it is necessary to take into account effects due to dissipation. Theoretically, this is usually accomplished by considering the coupling with a continuum of external modes simulating a thermal bath. The solution to this type of problem is nontrivial even for the simplest case of linear coupling to the bath.

The impulsive excitation of eqn [6] is a parametric excitation containing all the frequencies. A more common way to parametrically excite an oscillator is to use a monochromatic source, as described by the equation

$$\ddot{Q} + [\Omega^2 + \varepsilon\cos(\Omega_0 t)]Q = 0 \qquad [8]$$

where the parametric excitation is characterized by the amplitude ε and the modulation frequency Ω_0. Notice that the excitation induces a time-dependent frequency $\omega^2(t) = \Omega^2 + \varepsilon\cos(\Omega_0 t)$, and, hence, it

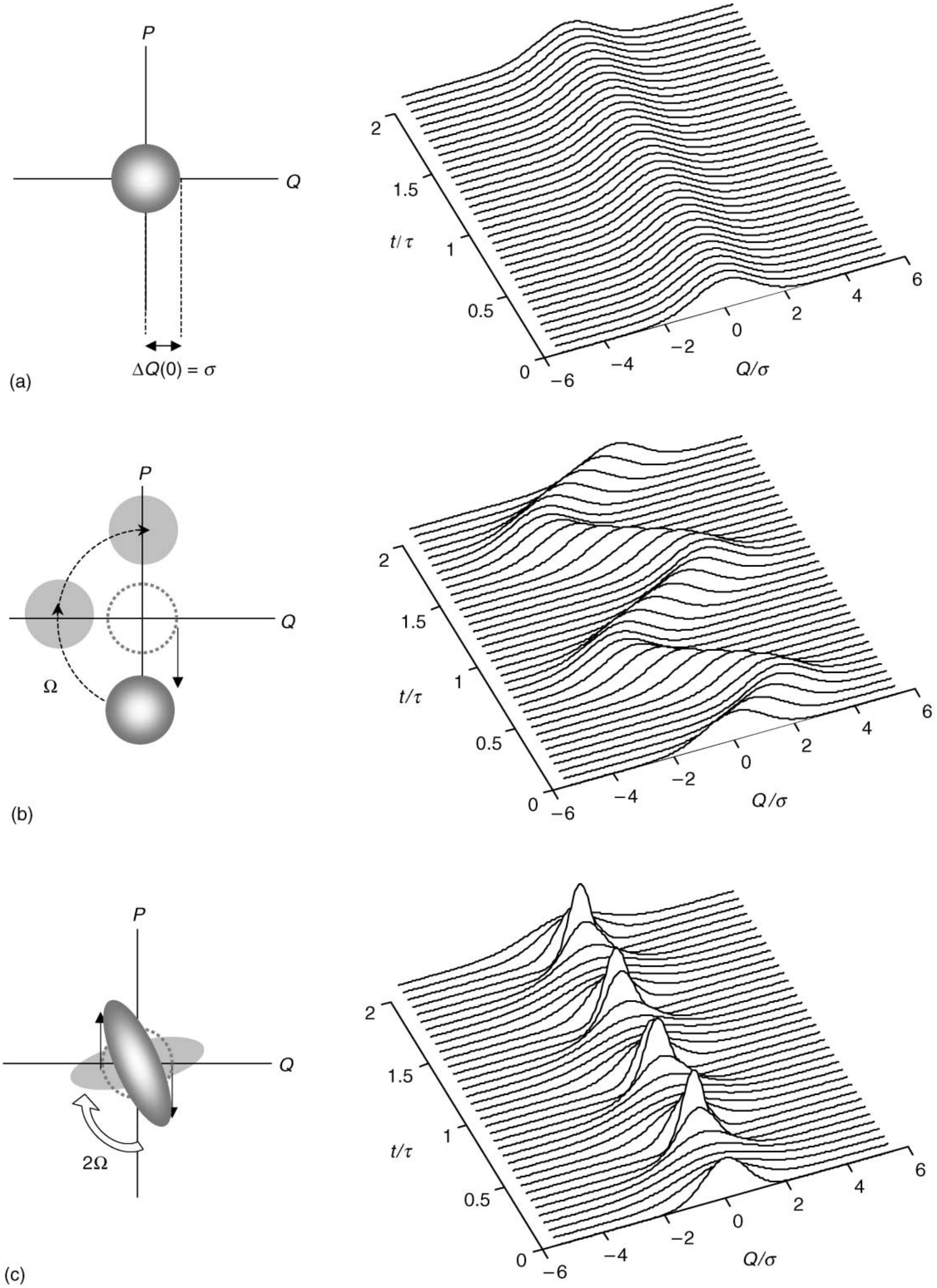

Figure 1 (a) Ground state probability density $|\varphi_0|^2$ for a quantum harmonic oscillator of frequency Ω. Q is the position, P is the momentum and σ is the standard deviation of the Gaussian wavepacket, or the square root of $\langle Q_0^2 \rangle$. The phase-space diagram on the left shows the uncertainty region for the ground state. (b) Coherent state. The uncertainty $\langle Q^2 \rangle$ does not change in time and it is always equal to $\langle Q_0^2 \rangle$, but the wavepacket oscillates with the period τ of the harmonic oscillator. The phase diagram shows the evolution of the uncertainty region, starting from the ground state (dashed line). A sudden constant change in momentum at time $t = 0$ produces a change in the equilibrium position and the wavepacket begins to oscillate. (c) Squeezed state. The wavepacket evolves by changing its width periodically with period $\tau/2$. The uncertainty, $\langle Q^2(t) \rangle$, takes values smaller than $\langle Q_0^2 \rangle$ for a fraction of the cycle. Due to the sudden change on momentum, proportional to Q, at $t = 0$, the distribution becomes ellipsoidal and rotates with frequency 2Ω. (d) Coherent-squeezed state. It is a combination of both (b) and (c). The wavepacket oscillates and breathes at the same time.

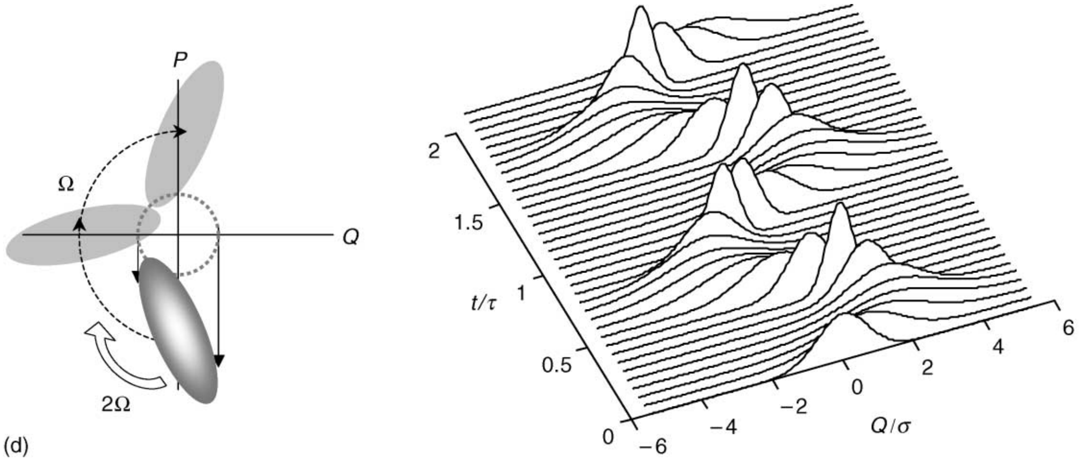

Figure 1 *Continued.*

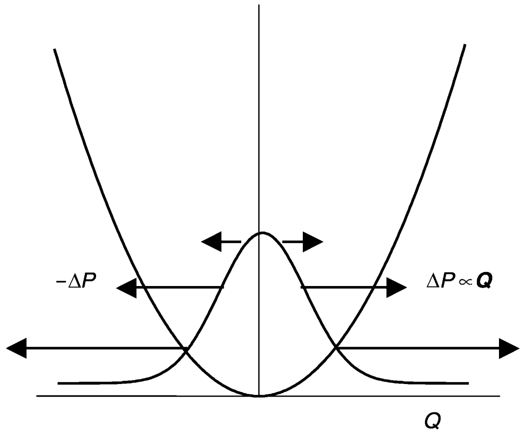

Figure 2 Schematic representation of the effect of an impulsive excitation on the ground state of the harmonic oscillator. At $t = 0$, the pulse induces a change in momentum P, proportional to the position Q. As a result, the wavepacket width in momentum space changes suddenly and starts to oscillate.

shows squeezing effects. The complete set of solutions with and without dissipation has been obtained by Zerbe and Hanggi for the quantum and classical cases.

Physical Interactions Giving Phonon Squeezing

The considerations above are rather general and specific to any physical system that behaves like a set of harmonic oscillators. Here, we consider proposals for phonon squeezing.

Equations [6] and [8] show that we need a driving force proportional to the phonon coordinate Q to achieve squeezing. One way to do this is for the external interaction to produce time-dependent correlations between two modes. Prior to the experimental achievement of phonon squeezing in 1997, using stimulated second-order Raman scattering, models based on (a) three-phonon parametric amplification and (b) polaritons had been proposed.

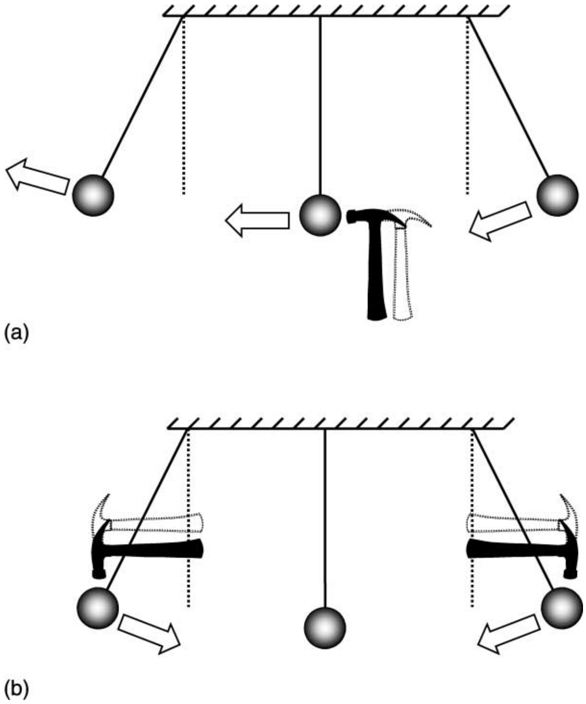

Figure 3 (a) Classical representation of a coherent phonon. An impulsive horizontal force acts on all members of a thermal distribution of pendula, displacing the group in phase space along the momentum axis. The ensemble precesses with an average position $\langle x \rangle \sim \sin(\Omega t)$. (b) Classical representation of a squeezed phonon. An impulsive vertical force produces an impulse proportional to the pendulum's displacement from equilibrium. Those pendula farthest from equilibrium receive the largest kick. The phase space distribution becomes elliptical and precesses with an ensemble averaged variance $\langle x^2 \rangle \sim \sin(2\Omega t)$.

These proposals have not yet been implemented experimentally.

The scheme proposed in (a) is based on three-phonon parametric amplification, similar to photon squeezed states that are generated by optical parametric amplification. Specifically, the proposal relies on a two-step process. First the generation by external excitation with two lasers of a 'pump' longitudinal optical phonon (LO) and, second, the decay of the optical phonon into two longitudinal acoustic phonons (LA), which will be squeezed. This is the kind of parametric oscillator described by eqn [8], for which the parametric control is driven by the coherent optical phonon field.

In case (b) it has been shown that mixed polarization modes known as phonon polaritons are intrinsically squeezed. These hybrid objects, sometimes referred to as dressed photons, are a mixture between phonons and photons that cannot be separated once they are formed.

Squeezing of vibrational states in molecules was confirmed experimentally earlier than in solids. Here, a molecule initially in the ground electronic and vibrational state is excited by an ultrafast optical pulse tuned with an electronic resonance. After the pulse arrives, the molecule is placed in a excited electronic state and the nuclei in a new (excited) potential with a new set of discrete vibrational modes. In Figure 4 we show, as an example of a diatomic molecule, the potential of a sodium dimer. For the sake of simplicity, we consider a single vibrational mode in the excited state and we disregard effects due to anharmonicity. If the optical pulsewidth is much shorter than any vibrational frequency, and the frequencies are different in the ground and excited state, the sudden change in frequency produces a vibrational squeezed state. On the other hand, even if the frequencies in the ground and excited state are equal, but the laser pulse is finite or chirped (frequency sweep on time), the final vibrational state may be squeezed. The formal demonstration of this is not trivial. Analytic solutions for the vibrational wavefunction for an arbitrary external field excitation are known for a very limited number of cases.

Squeezing of a wavepacket in an excited sodium diatomic molecule was achieved by Dunn and coworkers, measuring the time and frequency-resolved spectrum of the spontaneous emission. In this experiment, the molecule can return at any time to its electronic ground state by emitting light, the wavelength of which depends on the internuclear separation at that time. As shown in Figure 4, this is because the equilibrium position for the potential surfaces in the ground and excited states are different.

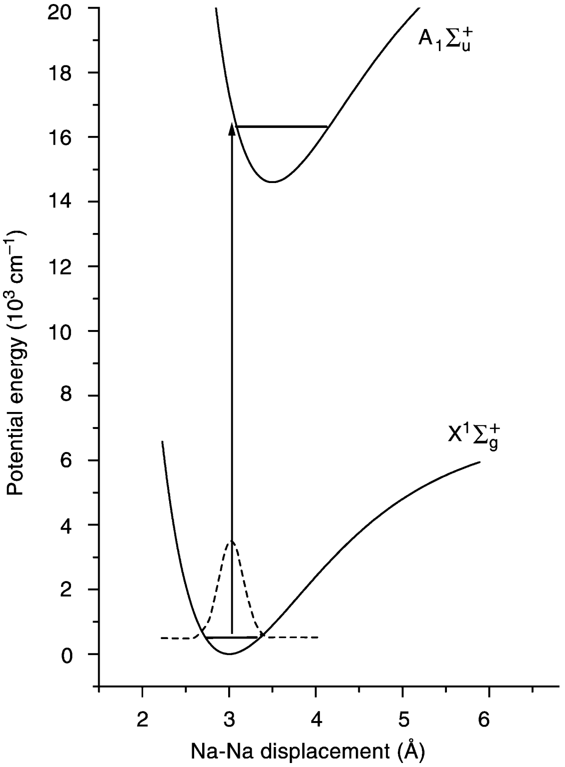

Figure 4 Potential curves for a Na_2 molecule, showing the ground and first excited states. The dashed line represents the ground state wave function. The equilibrium position of both potentials are shifted and the vibrational frequencies are different.

Thus, the time-dependent average wavelength of the emitted light oscillates as the atom vibrates, giving a signature of the dynamics of the wavepacket. Dunn observed oscillations in the signal strength that are consistent with a wavepacket expanding and contracting.

Experimental Generation and Detection of Squeezed Phonons by Second Order Raman Scattering

The first experimental demonstration of phonon squeezing was performed in 1997 by exciting a $KTaO_3$ (potassium tantalate) crystal with an ultrafast pulsed laser and coupling the light to the phonons through second order Raman Scattering. The squeezed phonon field produced by an intense optical pulse, referred to as the 'pump', is detected by a second low intensity pulse, known as the 'probe'. Phonon squeezing was also reported for $SrTiO_3$ (strontium titanate). In addition, correlations between pairs of phonons of different frequencies was reported in a closely related experiment on $KTaO_3$.

The simplest Hamiltonian describing this experiment contains the harmonic contribution from the lattice and a Raman-type interaction term. Raman scattering (RS) refers to light-crystal interaction in which the initial quantum of radiation energy, $\hbar\omega_i$, is divided into a vibrational excitation of the crystal with energy $\hbar\Omega$ and a scattered photon with energy $\hbar\omega_f$. Here, the total energy and momentum are conserved. The interaction energy is proportional to the second-order electronic susceptibility which can be expanded in terms of the atomic displacement Q. The nth term of the Taylor expansion is associated with RS of order n. Two phonons participate in the second order RS process. Although the absence of first-order RS is not required for achieving squeezing, $KTaO_3$ and SrTiO3 (above ~100 K) show no first-order Raman active modes. As a result, the leading contribution is

$$H_{\text{int}} = -\frac{1}{4}|E(t)|^2 \sum_q \chi_2(q) \qquad [9]$$

with $\chi_2(q) = (\partial^2\chi/\partial Q_q^2)Q_q^2$, where $\partial^2\chi/\partial Q_q^2$ is the second order Raman tensor, $E(t)$ is the magnitude of the light electric field, and χ is the linear electronic susceptibility. The Hamiltonian eqn [9], reflecting contributions from pairs of modes at $\pm q$, represents a change in the phonon frequency $\Delta\Omega_q(t) \approx -(1/4\Omega_q)(\partial^2\chi/\partial Q_0^2)|E(t)|^2$ that results in a time-dependent variance $\langle Q_q^2(t)\rangle$ and thus, squeezing. We note that, while individual scattering events produce a two-phonon state, the coherent ensemble of these correlated two-phonon states modulates the variance of Q. Notice that second-order RS excites phonons throughout the whole Brillouin zone instead of just the zone center at $q \approx 0$, as in first-order RS. This means that the light couples to a continuum of modes. The strength of the scattering at any particular frequency is proportional to the number of modes, i.e., the density of states, which is inversely proportional to the gradient $\nabla_q\Omega_q$ and thus, exhibits singularities at points where the gradient vanishes. These so-called critical points lead to associated singularities in the density of states referred to as van Hove singularities. $KTaO_3$ and $SrTiO_3$ show a sharp peak in the density of states at frequencies associated with certain van Hove singularities and, therefore, the continuum of modes becomes quasimonochromatic.

In the experiments, the period of the relevant phonons, $2\pi/\Omega_q$, is large compared with both the pulsewidth and the time the pulse takes for traversing the sample. Then, the time-dependent electric field can be approximated by the delta function $|E(t)|^2 \cong (4\pi I_0/nc)\delta(t)$, where I_0 is the pulse energy density, n the refractive index and c the speed of the light. Within this approximation, the dynamics of the system at $T = 0$ is described by eqn [6]. To lowest order in I_0, we have

$$\begin{aligned}\Delta Q_q &= \langle Q_q^2(t)\rangle^{\frac{1}{2}} \\ &\approx \langle Q_q^2(0)\rangle^{\frac{1}{2}}\left[1 + \left(\frac{\lambda}{\Omega_q}\right)\sin(2\Omega_q t)\right] \qquad [10]\end{aligned}$$

with $\lambda = 2\pi(I_0/nc)(\partial^2\chi/\partial Q_q^2)$.

As before, the variance drops below the quantum limit $\langle Q_q^2(0)\rangle$ for some fraction of the cycle. This zero-temperature description is only accurate for temperatures such that $k_B T$ is very small compared with the phonon energies. For $T \neq 0$, the conditions overcoming the thermal contribution and produce quantum squeezing at low intensities are

$$n_q \ll 1 \quad \text{and} \quad \left(\frac{\lambda}{\Omega_q}\right) > 2n_q \qquad [11]$$

where $n_q = 1/[\exp(\hbar\Omega_q/k_B T) - 1]$ is the Bose factor.

Next we discuss the detection of the squeezed field. As mentioned earlier, the pump pulse creates the squeezed phonon state and the probe is affected by the changes in the material produced by the pump. The scattering of the probe by the squeezed phonon is also described by the nonlinear polarization of the material $P^{NL} = (2V)^{-1}E'(t)\sum_q \chi_2(q)$, where $E'(t)$ is the probe field, and V is the scattering volume. If the probe pulse is Gaussian-shaped of width τ_0 and centered at $\omega = \omega_0$, the total normalized change in the probe transmission ΔT is

$$\begin{aligned}\frac{\Delta T}{T}(\omega) \approx &-\frac{l\pi\tau_0^2\omega(\omega-\omega_0)}{ncV}\sum_q \frac{\partial^2\chi}{\partial Q_q^2}|E'(t)|^2 \\ &\times \lambda\langle Q_q^2(0)\rangle\cos(2\Omega_q\Delta t)\exp(-2\Omega_q^2\tau_0^2) \qquad [12]\end{aligned}$$

where l is the length of the sample. Equation [12] can be obtained from well-known results in coherent phonon detection. Comparing eqn [12] with eqn [10] it follows that the change in probe transmission is proportional to the derivative of the total time-dependent variance $\langle Q^2(t)\rangle = \sum_q \langle Q_q^2(t)\rangle$. Accordingly, the integral of $\Delta T/T$ probes $\langle Q^2(t)\rangle$.

$KTaO_3$

The perovskite $KTaO_3$ was the first material for which phonon squeezing was achieved. This compound has no first-order Raman active modes. The time-domain data in Figure 5a shows behavior

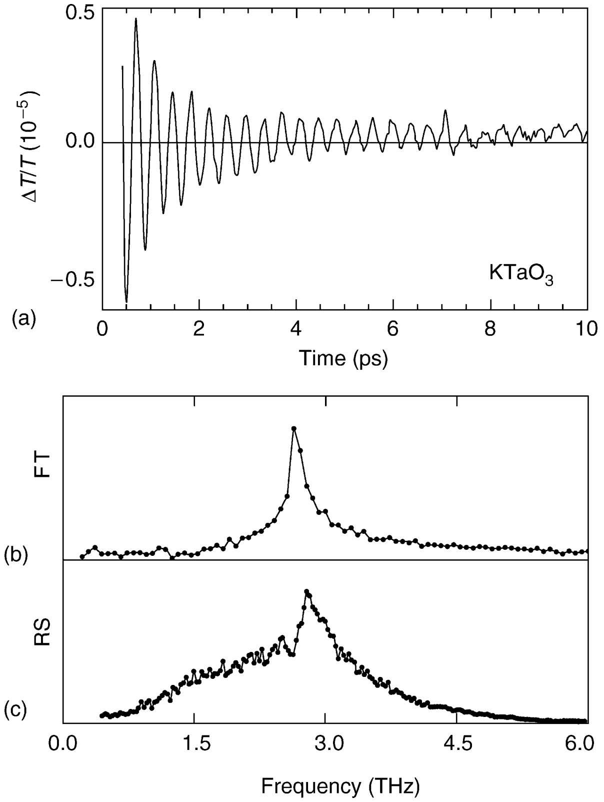

Figure 5 (a) The normalized transmitted probe intensity as a function of the probe delay for KTaO$_3$ at $T = 10$ K. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science*, 275(5306): 1638–1640. Copyright 1997 American Association for Advancement of Science. (b) Fourier transform of the time domain data in (a). The narrow peak appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. (c) Second order Raman cross section.

consistent with eqn [12]. The Fourier transform $\mathcal{F}_{\mathrm{FT}}$ in Figure 5b, is dominated by a narrow peak, strongly dependent on temperature, that appears very close to twice the frequency of the TA (transverse acoustic) mode at the X point of the Brillouin zone. The comparison with the spontaneous Raman scattering data, Figure 5c, indicates that the $\mathcal{F}_{\mathrm{FT}}$ peak is due to the 2TA overtone. The squeezing factor defined as $S = 1 - [\langle Q^2(t)\rangle / \langle Q^2(0)\rangle]^{1/2}$, is depicted in Figure 6. Notice that, whereas the squeezing factor in a quantum optics experiment can reach values as large as 70%, phonons in a solid can be squeezed by no more than, say, 0.01%. This is because larger values may lead to collapse of the solid since, according to the Lindemann criterion, solids melt when the root-mean-square displacement of an atom is a fraction of 0.2–0.25 of the lattice parameter.

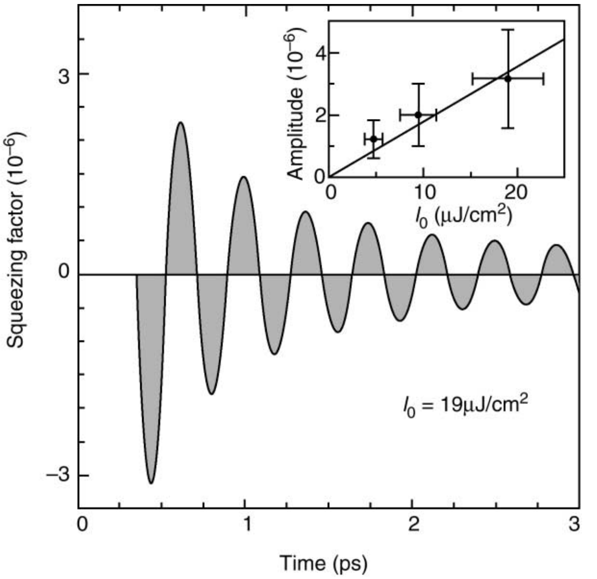

Figure 6 Experimental time dependence of the squeezing factor $S = 1 - [\langle Q^2(t)\rangle / \langle Q^2(0)\rangle]^{1/2}$. I_0 is the pulse intensity. Inset: the amplitude of S as a function of the integrated intensity of the pulse, showing the linear behavior predicted by eqn [12]. Reprinted with permission from Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640. Copyright 1997 American Association for the Advancement of Science.

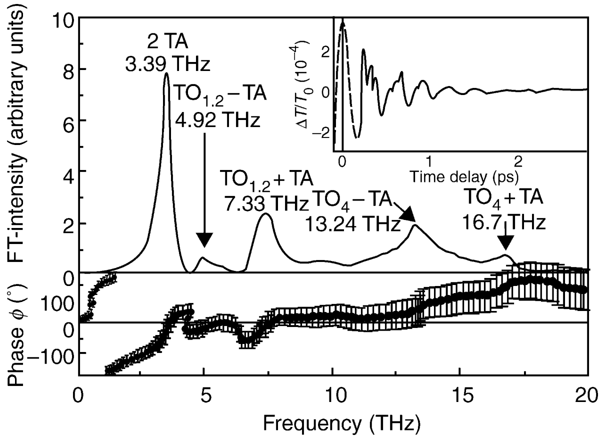

Figure 7 Fourier transform of the time domain data (inset) of the probe differential transmitivity of KTaO$_3$ at room temperature. The excited phonon combinations are indicated in the figure. Reproduced with permission from Bartels A, Dekorsy T and Kurz H (2000) Impulsive excitation of phonon-pair combination states by second-order Raman scattering. *Physical Review Letters* 84(13): 2981–2984. Copyright 2000 by the American Physical Society.

In KTaO$_3$ it was also possible to induce a time-dependent covariance between two different independent modes by using second order RS to couple the two oscillators. In Figure 7 we show

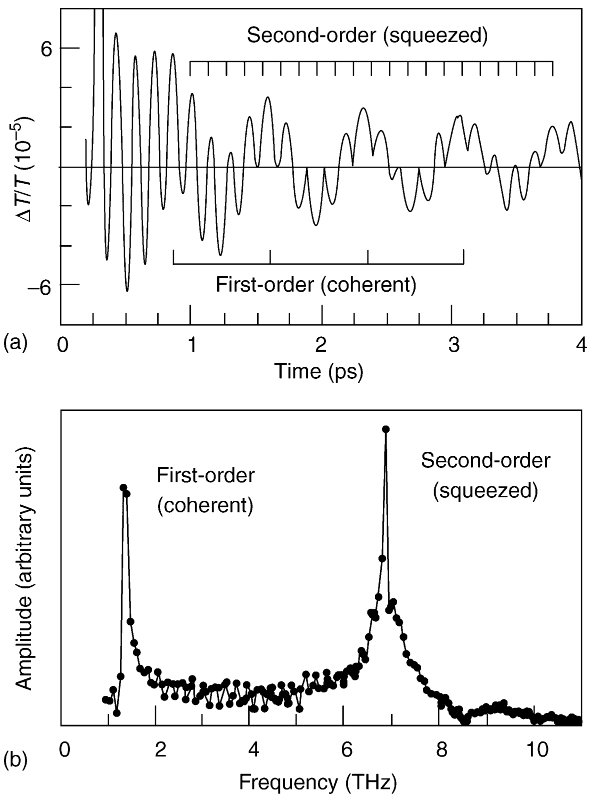

Figure 8 (a) The normalized transmitted probe intensity as a function of the probe delay for $SrTiO_3$ at $T = 7$ K. (b) Fourier transform of the time domain data. The 1.3 THz peak is the soft A_{1g}-phonon (coherent), and the 6.9 THz is the 2 TA overtone (squeezed). The second order feature is dominated by a sharp peak very close to twice the frequency of TA modes at X and M points of the Brillouin zone.

data revealing correlations between transverse optical (TO) and transverse acoustical (TA) modes.

$SrTiO_3$

This material undergoes a structural phase transition, called antiferro-distortive, at $T_C \approx 110$ K. Second-order RS is allowed at all temperatures, but first-order RS is only allowed for temperatures below T_C. Figure 8 shows data at low temperatures, revealing both first and second order features. From the comparison with spontaneous RS, the peak appearing in the $\mathcal{F}_{FT}$ at ~1.3 THz is due to first-order RS, while the peak at ~6.9 THz is due to second order RS (2TA overtone). The quantum state created by the ultrafast pulse is then a mixture of a coherent and a squeezed state. Hence the phonon wavefunction will evolve qualitatively as in the picture shown in Figure 1d.

List of Units and Nomenclature

Gaussian units

I_0	energy density [Jcm^{-2}]
E	electric field [statvolt cm^{-1}]

See also

Scattering: Stimulated Scattering; Raman Scattering. **Semiconductor Physics:** Light Scattering. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses.

Further Reading

Dunn TJ, Sweetser JN, Walmsley IA and Radzewicz C (1995) Experimental determination of the dynamics of a molecular nuclear wave-packet via the spectra of spontaneous emission. *Physical Review Letters* 74(6): 884–887.

Garraway BM and Suominen KA (1995) Wave-packet dynamics: new physics and chemistry in femto-time. *Reports on Progress in Physics* 58(4): 365–419.

Garrett GA, Rojo AG, Sood AK, Whitaker JF and Merlin R (1997) Vacuum squeezing of solids: Macroscopic quantum states driven by light pulses. *Science* 275(5306): 1638–1640.

Hayes W and Loudon R (1978) *Scattering of Light by Crystals*. New York: John Wiley & Sons, Inc.

Lindemann F (1910) Über die Berechnung molekularer Eigenfrequenzen. *Zeitschrift Für Physik* 11: 609–612.

Walls DF and Milburn GJ (1994) Generation and Applications of Squeezed Light. *Quantum Optics*, pp. 137–174. Berlin Heidelberg: Springer Verlag.

Zerbe C and Hanggi P (1995) Brownian parametric quantum oscillator with dissipation. *Physical Review E* 52(2): 1533–1543.

RELATIVISTIC NONLINEAR OPTICS

D P Umstadter, University of Michigan, Ann Arbor, MI, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Immediately after the invention of the laser in 1960, light could be focused to sufficient intensity to cause nonlinear optical effects in atomic media. In conventional nonlinear optics, where the electric field of a light wave E is much smaller than an atomic field, $E_{at} = 3 \times 10^9$ V/cm, various nonlinear phenomena, such as self-focusing, harmonic generation, and Raman scattering, arise due to the anharmonic motion of electrons in the combined fields of atom and laser. Approximate analytical solutions can be obtained by means of perturbation expansion methods, using E/E_{at} as the expansion parameter. At higher light fields, when E approaches E_{at}, this method breaks down and the medium becomes photo-ionized, creating a plasma, as illustrated by Figure 1. Further increases in light intensity enabled nonlinear optical effects of even these free plasma electrons (see Figure 2). The nonlinearity arises, in this case, because the electrons oscillate at relativistic velocities in laser fields that exceed 10^{11} V/cm, resulting in relativistic mass changes exceeding the electron rest mass and the light's magnetic field becomes important. The work done by the electromagnetic field (E) on an electron ($eE\lambda$) over the distance of a laser wavelength (λ), then approaches the electron rest mass energy ($m_e c^2$), where e is the elementary charge of an electron, m_e is the electron rest mass, and c is the speed of light. Effects analogous to those studied with conventional nonlinear optics – self-focusing, self-modulation, harmonic generation, and so on – are all found, but based on this entirely different physical mechanism. Thus, a new field of nonlinear optics, that of relativistic electrons, has been launched, as illustrated by Figure 1.

One outcome of accessing this new optical regime is the generation of frequency-shifted light in a spectral region where there are no other compact sources. Another is the acceleration of other types of particles, such as positrons, ions, and neutrons. These novel radiation sources have properties (femtosecond duration, micron source size, MeV energy) that make them suitable for numerous applications in imaging and spectroscopy in basic research, as well as medical diagnostics, cancer therapy, energy production, and space propulsion. Rapid advancement is underway and new research tools, subfields, and commercial products are on the horizon, e.g., compact and ultrashort pulse duration laser-based electron accelerators and X-ray sources.

Another physical regime will be encountered at even higher intensities ($I\lambda^2 \simeq 10^{24}$ W/cm^2), when even protons will quiver relativistically, i.e., the work done on a proton over the distance of a laser wavelength approaches its rest mass energy. This might be called the nuclear regime of laser–plasma interactions, because of the fusion and fission reactions and the generation of pions, muons, and neutrinos that should occur as nuclei collide in such energetic plasmas.

The recent dramatic increase in light intensity was partly made possible by the development, in the last decade, of compact lasers that have the ability to amplify shorter light pulses. For instance, solid-state lasers use the technique of chirped-pulse amplification to generate femtosecond duration pulses. To accomplish this, an ultrashort, low-energy laser pulse is first stretched in time before it is amplified and then recompressed. Gas or dye lasers, using solid-state switches, have produced picosecond duration pulses. Present day advanced laser systems now have multi-terawatt peak powers and, when focused to micron spotsizes with adaptive optics, can produce electromagnetic intensities $I\lambda^2 \simeq 10^{21}$ W/cm^2, as illustrated by Figure 3. An example of a modern ultra-high-power (0.9 PW) solid-state (Ti:Sapphire) laser

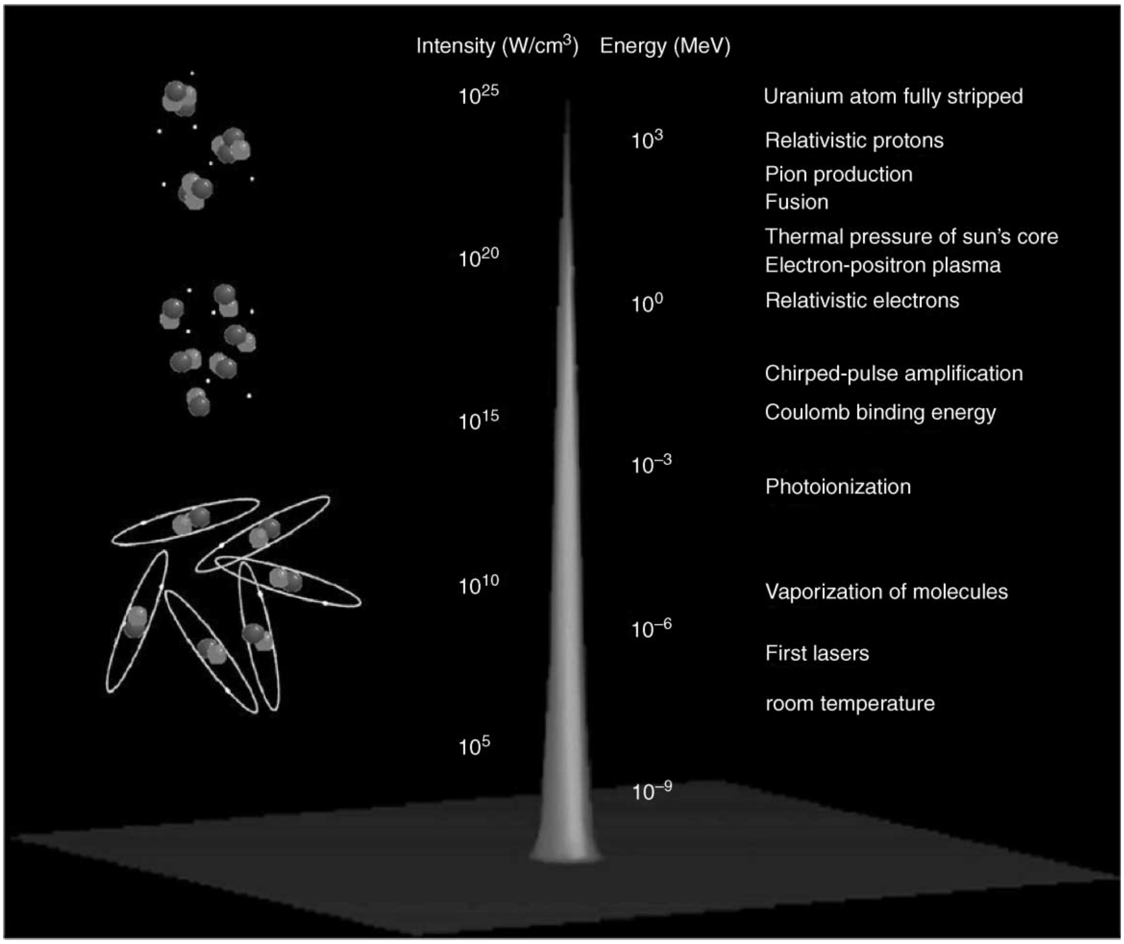
Intensity (W/cm^3)
Energy (MeV)
10^{25}
10^{20}
10^{15}
10^{10}
10^{5}
10^{3}
10^{0}
10^{-3}
10^{-6}
10^{-9}
Uranium atom fully stripped
Relativistic protons
Pion production
Fusion
Thermal pressure of sun's core
Electron-positron plasma
Relativistic electrons
Chirped-pulse amplification
Coulomb binding energy
Photoionization
Vaporization of molecules
First lasers
room temperature

Figure 1 The various regimes of laser–matter interactions, represented by the ideal laser pulse. As the intensity of laser light increases, so does the energy of electrons accelerated in the light field and the regime of conventional nonlinear optics with electrons bound to atoms is replaced by the regime of relativistic nonlinear optics with free electrons in relativistic plasmas. At the highest intensities, even protons become relativistic, giving rise to what might be called the regime of nuclear optics, in which various nuclear processes, such as fusion, can take place. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

system, located at the Advanced Photon Research Center, Kansai Research Establishment, Japan Atomic Energy Research Institute (JAERI), can be seen in the photograph shown in Figure 4.

This paper will discuss the fundamental concepts underlying what might be referred to more broadly as the field of high-field science. For more detailed descriptions of recent progress in experiment or theory, several review papers have been published on related topics: (i) relativistic nonlinear optics; (ii) high-intensity laser development; (iii) laser accelerators; (iv) intense laser–plasma interactions; and (v) relativistic scattering.

Definitions

For an infinite plane wave for the light wave propagating in the $\hat{z}$ direction (shown in Figure 5), the fields may be described by

$$\mathbf{E} = \hat{e}_x E_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_x E_0\, \mathrm{e}^{\mathrm{i}\psi} \qquad [1]$$

and

$$\mathbf{B} = \hat{e}_y B_0 \exp[\mathrm{i}(kz - \omega t)] = \hat{e}_y B_0\, \mathrm{e}^{\mathrm{i}\psi} \qquad [2]$$

where ω and k are the frequency and wavenumber of the laser light, respectively. For the ultrashort pulse durations of the lasers typically used to produce high fields, it is assumed that the ions do not have time to move because of their greater inertia. In this case, $m(\mathrm{d}v/\mathrm{d}t) \approx -e\mathbf{E}$ and the term v/c can be used as an expansion parameter. At even higher fields, electrons can be accelerated to nearly the speed of light as they quiver in the oscillating fields of the light wave. In this case, the force due to the magnetic field can no longer be ignored and the mass of the electron

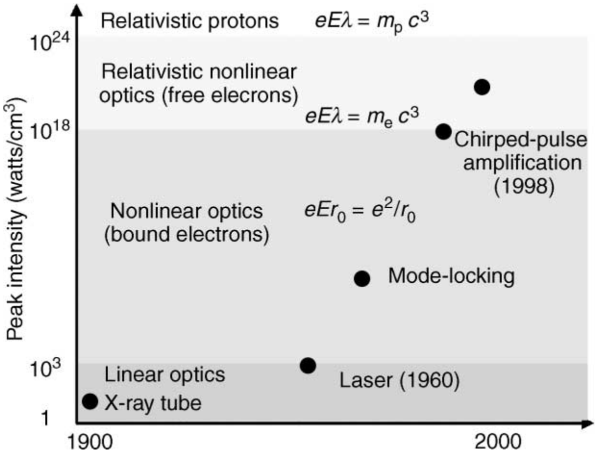

Figure 2 History of light sources over the last century. Each advance in laser power enables a new regime of optics. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

can increase as $m = \gamma m_0$, where m_0 is the electron rest mass and the relativistic gamma factor is given by

$$\gamma = \frac{1}{\sqrt{1-\left(\frac{v}{c}\right)^2}} \qquad [3]$$

The threshold for the relativistic regime of optics is crossed when the electron momentum p exceeds the product of the electron rest mass m_0 and c.

For problems involving charged particle motion in intense electromagnetic fields, it is often convenient to introduce the vector potential: $\mathbf{A}(\mathbf{r},t) = \mathrm{Re}\{\mathbf{A}_0\, e^{i\psi}\}$. From Maxwell's equation $\nabla \cdot B = 0$, it can be seen that without loss of generality, the magnetic field can be written as the curl of a potential $B = \nabla \times A$. Further, from Faraday's law in vacuum:

$$\nabla \times E = -\frac{1}{c}\frac{\partial B}{\partial t} = -\frac{1}{c}\frac{\partial(\nabla \times A)}{\partial t} \qquad [4]$$

it is clear that $\mathbf{E} = -(1/c)\partial\mathbf{A}/\partial t$. To include the space–charge field of the plasma, as we will do in the section on collective plasma response below, we introduce the potential Φ, and the definition:

$$\mathbf{E} = -\nabla\Phi - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{A} \qquad [5]$$

Assuming the Coulomb gauge:

$$\nabla \cdot \mathbf{A} = 0 \qquad [6]$$

or $A_z = 0$ in 1D, then $\mathbf{A}_\perp$ represents the laser field.

If we neglect Φ for now, then, by Fourier analysis:

$$\mathbf{E} = \mathrm{Re}\left\{\frac{i\omega}{c}\mathbf{A}_0\, e^{i\psi}\right\} \qquad [7]$$

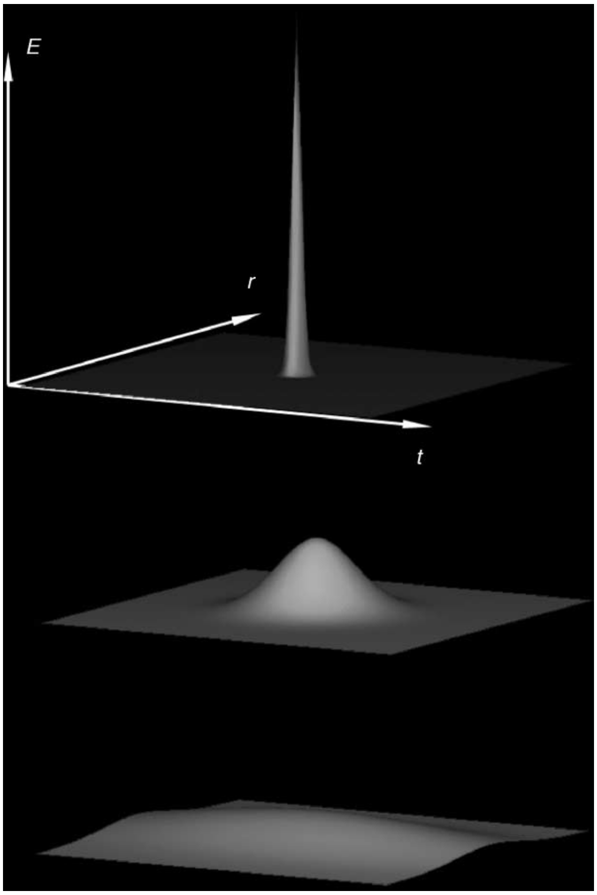

Figure 3 Laser light has become concentrated to ever-smaller regions of space (r) and time (t), dramatically increasing the peak electric field (E) at the laser focus. Prior to the development of chirped-pulse amplification (CPA), the energy of light was produced in long-duration pulses, as shown in the pulse of the bottom figure. After CPA, the pulse duration decreased dramatically as shown in the pulse in the middle. The latest improvement in laser technology has been the use of deformable mirrors, which has allowed lasers to be focused to a spatial dimension that is as small as the temporal dimension, a few laser wavelengths, as shown in the pulse on top. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

$$\mathbf{B} = \mathrm{Re}\{ik\hat{z} \times \mathbf{A}_0\, e^{i\psi}\} \qquad [8]$$

The light intensity is given by the time-averaged Poynting's vector, which can be written for linear (LP), $\mathbf{A}_0 = A_0\hat{e}_y$ and circular polarizations (CP), $\mathbf{A}_0 = A_0(\hat{e}_x \pm \hat{e}_y)$ as:

$$I = \langle|\mathbf{S}|\rangle = \left\langle\left|\frac{c\mathbf{E}\times\mathbf{B}}{8\pi}\right|\right\rangle = \frac{\omega k}{16\pi}A_0^2 \qquad [9]$$

It is often the combination $I\lambda^3$ that matters, which has units of power and for linear polarization can be

Figure 4 Photograph of the 33-fs duration 0.9-PW-peak power laser system at JAERI. (Reproduced with permission from Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.)

written as

$$I\lambda^2 = \frac{\omega k \lambda^2}{16\pi} A_0^2 = \frac{\pi}{2} c A_0^2 = \left[1.37 \times 10^{18} \frac{\mathrm{W}}{\mathrm{cm}^2}\,\mu\mathrm{m}^2\right] a_0^2 \qquad [10]$$

where the normalized vector potential is defined as

$$a_0 = \frac{eA_0}{mc^2} \qquad [11]$$

Relativistic effects become significant when a_0 approaches unity, which, as can be seen from eqn [10] occurs for 1 μm wavelength light at a laser intensity of $\sim 1.37 \times 10^{18}$ W/cm^2, corresponding to a electric field strength of 6×10^9 V/cm, which is four orders of magnitude greater than E_{at}.

Single-Particle Motion

The electron orbit, subject to a linearly polarized electromagnetic wave propagating in the $+z$ direction, is governed by the Lorentz equation:

$$m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = -e\left(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B}\right) \qquad [12]$$

where v is the electron velocity, and E and B are the light's electric and magnetic field.

A zeroth order solution to eqn [12] is found by setting $v/c \ll 1$, which allows the term $(v/c) \times B$ to be neglected and γ from eqn [8] is set to unity. Integrating eqn [12] in this limit once over time (t) yields for the velocity:

$$\mathbf{v} = \mathrm{Re}\left\{\frac{e\mathbf{E}}{im\omega}\right\} = \frac{a_0}{c}\begin{cases} \hat{e}_y \cos\psi, & \text{LP} \\ (\hat{e}_x \cos\varphi \mp \hat{e}_y \sin\psi), & \text{CP} \end{cases} \qquad [13]$$

An electron in low-intensity light oscillates with this velocity in a straight line along the polarization vector ($\hat{e}_y$), which when normalized to c, is the normalized vector potential a_0. Integrating again yields for the

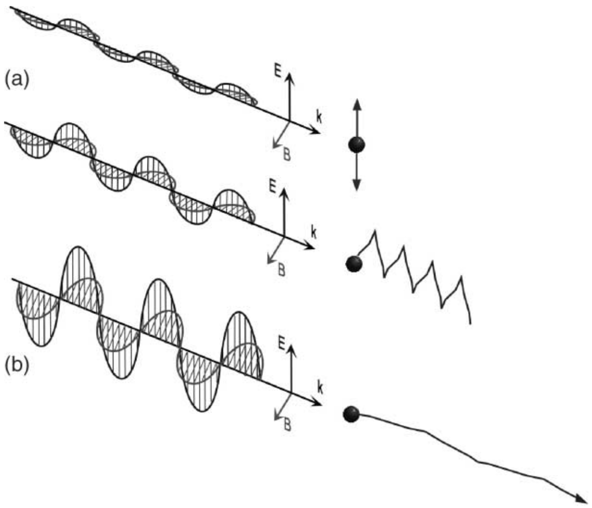

Figure 5 Classical optics versus relativistic optics. (a) In classical optics, the amplitude of the light wave is small, electrons oscillate in the direction of the electric field at the light's frequency, and there is no displacement along the light's propagation direction. Note that only the *E* field acts on the electron and the electron-oscillation velocity is very small compared with the speed of light. (b) In relativistic optics, the amplitude of the light wave is very large, the light's magnetic field becomes important and the combined action of the electric and magnetic fields push the electron forward. In this case, the electron velocity becomes close to the speed of light and its motion is a figure-eight superimposed upon a steady drift. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

transverse displacement:

$$\mathbf{x} = \mathrm{Re}\left\{\frac{e\mathrm{E}}{m\omega^2}\right\} = \frac{a_0}{\omega c}\begin{cases} \hat{e}_y \sin\psi, & \text{LP} \\ (\hat{e}_y \sin\varphi \mp \hat{e}_z \cos\psi), & \text{CP} \end{cases} \qquad [14]$$

Thus, for the $a_0 \sim 1$, the electron excursion during its oscillation is approximately λ.

A first-order approximation for the electron motion can be found by substituting the zeroth order velocity eqn [13] into the $v/c \times B$ term of eqn [12]. The latter then becomes proportional to

$$E \times B \propto \frac{a_0^2}{2}[1 + \cos(2\omega_0)]\hat{e}_z \qquad [15]$$

In the frame in which the electron is on the average at rest, the relativistic motion of an electron is thus described by a figure eight, oscillating twice in the $\hat{e}_z$ or $\hat{k}$ direction for every once in the polarization direction ($\hat{e}_x$); i.e., a figure of eight motion. This originates from the fact that $v \times B \propto E \times B \propto E^2\hat{k}$, which is a product of two functions that vary sinusoidally at frequency ω and thus varies itself at frequency 2ω. In the lab frame, this transverse motion is superimposed upon a steady drift in the ($\hat{e}_z$) direction (shown in Figure 5), originating from the DC term in eqn [15]. The next order approximation would include the mass shift $m = \gamma m_0$. As the field strength increases ($a_0^2 \gg 1$), the longitudinal motion ($\propto a_0^2$) begins to dominate the transverse motion ($\propto a_0$), as shown in Figure 5. This is described more formally below.

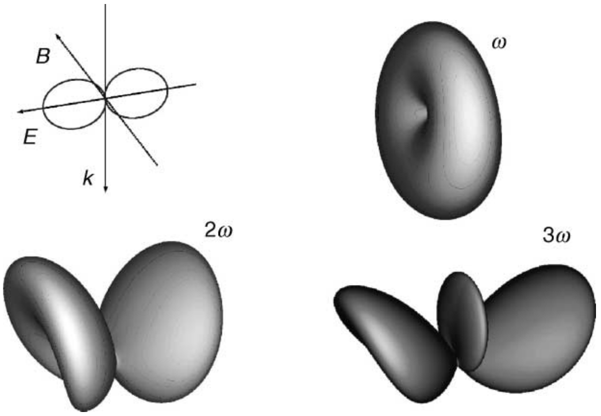

Figure 6 Harmonics driven by relativistic Thomson scattering as the electrons in high-intensity laser fields (a_0^2) undergo figure-eight motion display unique angular distributions. (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

In the regime $a_0 \lesssim 1$, electrons radiate photons at harmonics of a modified laser frequency ω_0, with each harmonic order having its own unique angular distribution, as shown in Figure 6. The radiation at the fundamental is the usual donut pattern, with a maximum in the direction perpendicular, and a minimum along, the polarization vector ($\hat{e}_x$). The second harmonic has two emission lobes with maxima pointing at an angle between $\hat{e}_z$ and $\hat{e}_x$. An additional lobe is added for each additional harmonic order. This is referred to as nonlinear or relativistic Thomson scattering. The unique angular distributions of the second and third harmonics emitted from nonlinear relativistic Thomson scattering were observed experimentally and are shown in Figure 7. As will be shown in detail in the section on radiation from relativistic electrons, below, the situation is more complex for $a_0 \gtrsim 1$, and the scattered light is no longer simply harmonic.

Constants of the Motion

Several constants of the motion can be found from an exact treatment for the motion. The starting point is the relativistically correct Lagrangian, which is written as

$$L(\mathbf{r}, \mathbf{v}, t) = -mc^2\sqrt{1 - \frac{v^2}{c^2}} + \frac{q}{c}\mathbf{v}\cdot\mathbf{A} + q\phi \qquad [16]$$

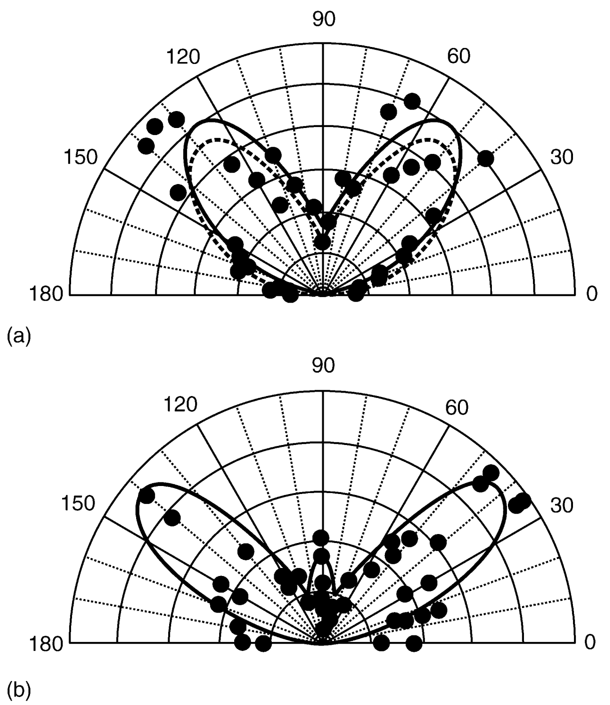

Figure 7 Angular pattern of higher-order harmonic light. Shown are polar plots of the intensity of the second-harmonic light (top) and third harmonic (bottom) as a function of azimuthal angle. Filled circles, experimental data; solid and dashed lines, theoretical results. (Reproduced with permission from Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.)

where ϕ is the scalar potential associated with electrostatic fields. The Lorentz equation, eqn [12], is derived from the Euler–Lagrange equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \mathbf{v}} - \frac{\partial L}{\partial \mathbf{r}} = 0 \qquad [17]$$

For an infinite plane wave, in which the Lagrangian is independent of space in the transverse direction, eqn [17] yields the conservation of transverse canonical momentum:

$$\frac{\partial L}{\partial \mathbf{v}_\perp} = \mathbf{p}_\perp + \frac{q}{c}\mathbf{A}_\perp = \text{constant} \qquad [18]$$

The sum of the transverse momentum and field strength remain constant. The next constant follows from $\mathrm{d}H/\mathrm{d}t = -\partial L/\partial t$, yielding the relation:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\frac{\partial L}{\partial t} = c\frac{\partial L}{\partial z} = c\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial v_z} = c\frac{\mathrm{d}p_z}{\mathrm{d}t} \qquad [19]$$

where E is the time-dependent energy of the particle, which yields our second constant of the motion:

$$E - cp_z = \text{constant} \qquad [20]$$

For a particle initially at rest, the kinetic energy E_{kin} becomes

$$E_{\mathrm{kin}} = E - mc^2 = p_z c \qquad [21]$$

which yields

$$E_{\mathrm{kin}} = \frac{p_\perp^2}{2m} = p_x c = mc^2(\gamma - 1) \qquad [22]$$

Thus, the electron scattering angle θ is related to the transverse and longitudinal momenta or the kinetic energy by

$$\tan^2\theta = \left(\frac{p_\perp}{p_z}\right)^2 = \frac{2mE_{\mathrm{kin}}}{(E_{\mathrm{kin}}/c)^2} = \frac{2}{\gamma - 1} \qquad [23]$$

The angle of electrons produced by photo-ionization with intense lasers has been shown experimentally to obey the conservation of canonical momentum, as demonstrated in experiments that studied the angular distribution of relativistic electrons emitted from barrier-suppression ionization of atoms in intense laser fields (shown in Figure 8). The transverse normalized vector potential, $\mathbf{a} = e\mathbf{A}_\perp/mc^2$, is thus equal to the transverse normalized momentum:

$$\hat{\mathbf{p}}_\perp = \frac{\mathbf{p}_\perp}{mc} = \mathbf{a} = (0, a_y, a_x) \qquad [24]$$

The normalized kinetic energy can also be written in terms of a:

$$\hat{E}_{\mathrm{kin}} = \frac{E_{\mathrm{kin}}}{mc^2} = \gamma - 1 = \hat{p}_z = \frac{\hat{p}_\perp^2}{2} = \frac{a^2}{2} \qquad [25]$$

This can be rewritten as

$$\gamma = 1 + \frac{a^2}{2} \qquad [26]$$

From eqn [25], the normalized momenta in each direction can be found:

$$\hat{p}_z = \gamma\beta_z = \frac{\gamma}{c}\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{a^2}{2} \qquad [27]$$

$$\hat{p}_y = \gamma\beta_y = \frac{\gamma}{c}\frac{\mathrm{d}y}{\mathrm{d}t} = a_y \qquad [28]$$

$$\hat{p}_x = \gamma\beta_x = \frac{\gamma}{c}\frac{\mathrm{d}x}{\mathrm{d}t} = a_x \qquad [29]$$

Note that the momentum in the longitudinal ($\hat{e}_z$) direction is proportional to a^2 while in the two

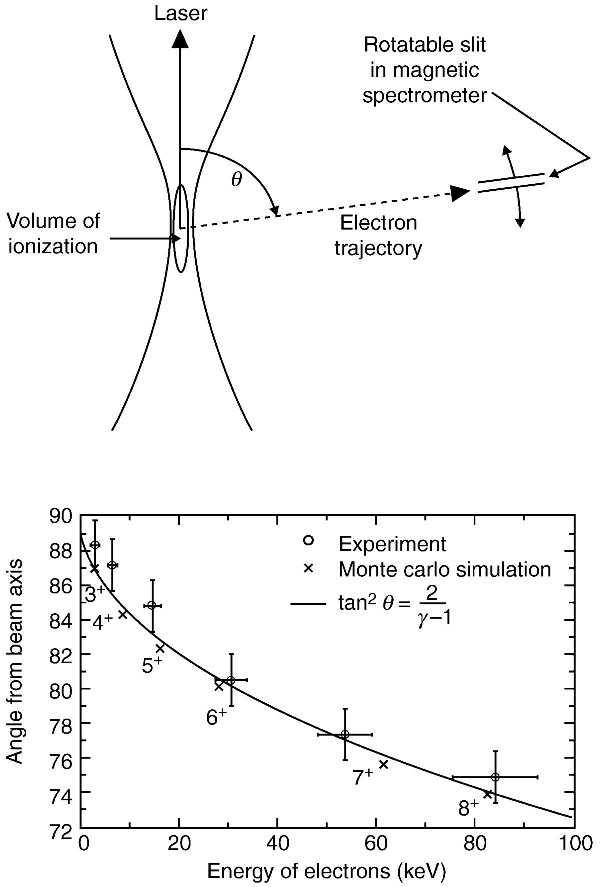

Figure 8 Top: Experiment to measure the angle of electron ejection during photo-ionization. Bottom: Kinetic energy versus angle during the production of 3^+ through 8^+ of neon. (Reproduced with permission from Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439. Copyright (1995) by the American Physical Society.)

transverse directions it is proportional to a. Equations [26]–[29] indicate that in the limit of large a, the normalized velocities are given by

$$\beta_z = \frac{a^2/2}{1+a^2/2} \to 1, \qquad \beta_y = \frac{a_y}{1+a^2/2} \to 0, \qquad \beta_x = \frac{a_x}{1+a^2/2} \to 0 \qquad [30]$$

This indicates that in the deeply relativistic regime, the electron is accelerated to near c in the longitudinal direction, while in the transverse directions its velocity goes to zero. It follows that, as a increases and the electron moves in phase with the light wave, relativistic Thomson scattering will be significantly reduced. Integrating eqn [29] in the variable $\tau = t - x(t)/c$ gives

$$\gamma\frac{\mathrm{d}}{\mathrm{d}t} = \gamma\frac{\mathrm{d}\tau}{\mathrm{d}t}\frac{\mathrm{d}}{\mathrm{d}\tau} = \gamma\left(1 - \frac{1}{c}\frac{\mathrm{d}z}{\mathrm{d}\tau}\right)\frac{\mathrm{d}}{\mathrm{d}\tau}$$
$$= \left(1 + \frac{a^2}{2} - \frac{a^2}{2}\right)\frac{\mathrm{d}}{\mathrm{d}\tau} = \frac{\mathrm{d}}{\mathrm{d}\tau} \qquad [31]$$

which allows the equations for the displacement to be written simply as

$$\frac{\mathrm{d}z}{\mathrm{d}\tau} = c\frac{a^2}{2} \qquad [32]$$

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = ca_y \qquad [33]$$

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = ca_x \qquad [34]$$

For a box-shaped pulse, $a_x = a_0\cos(\omega\tau)$ for $0 < \tau < N(2\pi/\omega)$, $a_z = 0$, $a^2 = a_x^2\cos^2(\omega\tau)$, the trajectory for an electron initially at rest at $x = y = z = 0$ at $\tau = 0$ is then

$$x(\tau) = \frac{ca_0^2}{2}\int_0^\tau \cos^2(\omega\hat{\tau})\mathrm{d}\hat{\tau}$$
$$= \frac{ca_0^2}{4}\left[\tau + \frac{1}{2\omega}\sin(2\omega\tau)\right] \qquad [35]$$

$$y(\tau) = ca_0\int_0^\tau \cos(\omega\hat{\tau})\mathrm{d}\hat{\tau} = \frac{ca_0}{4}\sin(\omega\tau) \qquad [36]$$

Thus, as we saw from the first-order appoximation above, eqn [15], the motion consists of a drift along $\hat{x}$:

$$x_{\mathrm{d}}(t) = \frac{a_0^2}{a_0^2+4}ct \qquad [37]$$

and a figure-eight motion in the drift frame of reference:

$$ky = a_0\sin(\omega\tau), \qquad k(z - z_{\mathrm{d}}) = \frac{a_0}{8}\sin(2\omega\tau) \qquad [38]$$

In the strongly relativistic regime, because the electron is strongly accelerated in the direction of the light wave, the harmonic emission from relativistic Thomson scattering loses its dependence on the laser frequency, ω, but instead depends only on the direction of observation, the amplitude a_0 and the initial phase of the electron, relative to the light wave. Electrons that do not begin their trajectories at zero or π phase with respect to the light wave can drift in the transverse direction. For pulses with finite spatial width, such as a Gaussian pulse with field strength peaked on axis, the ponderomotive force of the light can also act to expel the electrons transversely from the axis.

For circular polarization:

$$\mathbf{a}(\mathbf{r},t) = \mathrm{Re}\{a_0(\hat{e}_y \pm \mathrm{i}\hat{e}_x)\mathrm{e}^{-\mathrm{i}\omega\tau}\} \qquad [39]$$

the electron drifts as it did in the case of linear polarization and described in eqn [37] but orbits in a helical trajectory:

$$y(\tau) = \frac{ca_0}{\sqrt{2}\omega}\sin(\omega t/\gamma) \qquad [40]$$

$$x(\tau) = \mp\frac{ca_0}{\sqrt{2}\omega}\cos(\omega t/\gamma) \qquad [41]$$

Rather than oscillate in amplitude, the relativistic factor remains constant:

$$\gamma = 1 + \frac{a_0^2}{4} \qquad [42]$$

In either polarization case, the light pulse eventually overtakes the electron, leaving the latter at rest. However, net acceleration in vacuum can be obtained only under special circumstances, described by the Lawson–Woodward criterion, such as in the near field of metallic surfaces; with high-order focusing modes instead of infinite plane waves; or with mixed frequencies, or with optics to terminate the light but not the electrons. Regions of a tightly focused Gaussian pulse, where the accelerating field propagates at sub-luminal velocities also exist, but they are effective for only extremely large values of a and thus only very short acceleration lengths.

Role of Initial Phase

If we wish to understand the role of the electron's initial phase, the relativistic Lorentz equation (eqn [12]) may also be written, for an $\hat{x}$-polarized wave traveling in the $\hat{z}$ direction, as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\gamma\beta) = (\hat{x} - \hat{z}\times\beta)a\cos(t-z) \qquad [43]$$

where we normalize time by $1/\omega_0$, velocity by c, and distance by c/ω_0. In eqn [43], $a = eE_0/m\omega_0 c$ is the dimensionless parameter measuring the electric field strength, $\gamma = (1-\beta_x^2-\beta_y^2-\beta_z^2)^{-1/2}$ is the relativistic mass factor, and $\beta = (\beta_x, \beta_y, \beta_z)$ is the electron velocity (in units of c). The electron orbit, subject to the following general initial conditions at time $t = 0$:

$$x = 0, \qquad y = 0, \qquad z = z_{\mathrm{in}} \qquad [44]$$

$$\beta_x = \beta_{x0}, \qquad \beta_y = \beta_{y0}, \qquad \beta_x = \beta_{z0} \qquad [45]$$

$$\beta_{x0} = \beta_{y0} = 0 \qquad [46]$$

has a closed form solution when it is expressed parametrically: $t = t(\theta)$, $r = r(\theta)$, $\beta = \beta(\theta)$, where θ is the phase of the wave, defined by

$$\theta = t - z \qquad [47]$$

Note that $\beta_0 = (\beta_{x0}, \beta_{y0}, \beta_{z0})$ is the unperturbed velocity of the electron ($a = 0$ limit) and that the initial phase that the electron sees is $\theta_{\mathrm{in}} = -z_{\mathrm{in}}$ according to eqns [44] and [47]. This phase can be important in the ionization of the gas by an intense laser.

For the special case $\beta_{x0} = 0$, $\beta_{y0} = 0$, one finds $\beta_y = 0$, and the orbital equation yields the following closed form solution:

$$\gamma = \gamma_0 + \frac{a^2(\sin\theta - \sin\theta_{\mathrm{in}})^2}{2\gamma_0(1-\beta_{z0})} \qquad [48]$$

$$\gamma\beta_z = \gamma - \gamma_0(1-\beta_{z0}) \qquad [49]$$

$$\gamma\beta_x = \frac{1}{\gamma_0(1-\beta_{z0})}a(\sin\theta - \sin\theta_{\mathrm{in}}) \qquad [50]$$

$$x = \frac{a[(\cos\theta_{\mathrm{in}} - \cos\theta) - (\theta - \theta_{\mathrm{in}})\sin\theta_{\mathrm{in}}]}{\gamma_0(1-\beta_{z0})} \qquad [51]$$

$$\begin{aligned} t = {} & \frac{(\theta-\theta_{\mathrm{in}})}{1-\beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] \\ & + \frac{a^2(1+\beta_{z0})}{2(1-\beta_{z0})}\left[-\frac{\sin 2\theta}{4} + 2\cos\theta\sin\theta_{\mathrm{in}}\right. \\ & \left. - \frac{3\sin 2\theta_{\mathrm{in}}}{4}\right] \end{aligned} \qquad [52]$$

In eqns [48]–[52], $\gamma_0 = (1-\beta_{z0}^2)^{-1/2}$ and β_{z0} may either be negative (counter-propagating against the laser), or zero (Thomson scattering), or positive (co-propagating with the laser). Note that the velocity components β_x and β_z are given as explicit functions of θ according to eqns [49] and [50], upon using eqn [48]. They are periodic functions of θ of period 2π. The period, T, of this periodic (Figure 8) motion is thus equal to the increase in t as θ increases by 2π. Thus, we obtain from eqn [53]:

$$T = \frac{2\pi}{1-\beta_{z0}}\left[1 + \frac{a^2}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] \qquad [53]$$

The parametric solution for the z-coordinate of the electron orbit is given by $z = t - \theta$, in which t is given by eqn [52]. Over one orbital period, T, the electron undergoes a net displacement $r_0 = (x_0, 0, z_0)$ where x_0 is given by the increase in eqn [51] as θ increases by 2π, and z_0 is simply $T - 2\pi$:

$$\begin{aligned} z_0 &= T - 2\pi \\ &= \frac{2\pi}{1-\beta_{z0}}\left[\beta_{z0} + \frac{a^3}{2}\left(\frac{1}{2} + \sin^2\theta_{\mathrm{in}}\right)\right] \end{aligned} \qquad [54]$$

$$x_0 = \frac{-2\pi a \sin\theta_{\mathrm{in}}}{\gamma_0(1-\beta_{z0})} \qquad [55]$$

Note that the electron trajectory depends on a, β_{z0}, and θ_{in} in a rather complicated manner. Accordingly, the fundamental frequency ω_1 of the radiation spectrum depends on these three quantities. For backscattered radiation ($\hat{n} = -\hat{z}$), the dependence on electron beam and on the laser becomes decoupled.

Following is a useful formula relating the change of t with respect to θ:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = 1 - \beta_z = \frac{\gamma_0(1-\beta_{z0})}{\gamma} \qquad [56]$$

which may be verified from eqns [47] and [49]. Equation [56] is also valid for arbitrary values of β_{x0}, β_{y0}, β_{z0}, and θ_{in}, in which case γ_0 is the electron's relativistic mass factor in the absence of the laser.

Radiation from Relativistic Electrons

An electron with displacement $r(t)$ and velocity $v(t)$ carries a current density $J(r,t) = ev(t)\delta r - r(t)]$, whose Fourier transform $J_{\mathrm{F}}(k,\omega)$ may easily be obtained. The total work done, W, in ergs, performed by the current $\mathbf{J}(\mathbf{r},t)$ on the electric field $\mathbf{E}(\mathbf{r},t)$ is given by

$$\begin{aligned} W &\equiv \int \mathrm{d}\omega\, \mathrm{d}\Omega \frac{\mathrm{d}^2 W}{\mathrm{d}\omega\, \mathrm{d}\Omega} \\ &= \int \mathrm{d}\omega\, \mathrm{d}\Omega \frac{\omega^2}{c^2}\left[|\mathbf{J}_{\mathrm{F}}(\mathbf{k},\omega)|^2 - |\mathbf{n}\cdot\mathbf{J}_{\mathrm{F}}(\mathbf{k},\omega)|^2\right] \end{aligned} \qquad [57]$$

where the $\mathbf{k}$-space differential volume $\mathrm{d}^3\mathbf{k} = k^2\, \mathrm{d}k\, \mathrm{d}\Omega$ is expressed in terms of the solid angle (Ω) in the direction of the unit vector $\mathbf{n} = c\mathbf{k}/\omega$. We immediately obtain from eqn [56], under the far-field approximation:

$$\frac{\mathrm{d}^2 W}{\mathrm{d}\Omega\, \mathrm{d}\omega} = \frac{e^2\omega^2}{4\pi^2 c}|\mathbf{n}\times[\mathbf{n}\times\mathbf{F}(\omega)]|^2 \qquad [58]$$

$$\mathbf{F}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\, \beta(t) \mathrm{e}^{\mathrm{i}\omega[t - \mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [59]$$

Equation [58] gives the energy radiated by the electron in the direction of the unit vector, per unit solid angle, per unit frequency ω. Radiation damping is ignored throughout.

Let us consider the simplest case where the electron orbit is strictly a periodic function of time with period T and a net displacement $\mathbf{r}_0$ per period. Thus, we have for all integers m (positive, negative, or zero):

$$\beta(t + mT) = \beta(t), \quad \mathbf{r}(t + mT) = m\mathbf{r}_0 + \mathbf{r}(t) \qquad [60]$$

Equation [59] may then be written as, because of Thomson scattering the energy of the scattered photon is much less than the electron rest mass in the average rest frame of the electron, i.e.:

$$\frac{\hbar\omega_{\mathrm{L}}}{mc^3} \ll 1 \qquad [61]$$

$$\begin{aligned} \mathbf{F}(\omega) &= \sum_{m=-\infty}^{\infty} \int_{mT}^{(m+1)T} \mathrm{d}t\, \beta(t)\mathrm{e}^{\mathrm{i}\omega[t - \mathbf{n}\cdot\mathbf{r}(t)/c]} \\ &= \sum_{m=-\infty}^{\infty} f(\omega)\mathrm{e}^{\mathrm{i}m\omega[T - \mathbf{n}\cdot\mathbf{r}_0/c]} \end{aligned} \qquad [62]$$

where

$$f(\omega) = \int_0^T \mathrm{d}t\, \beta(t)\mathrm{e}^{\mathrm{i}\omega[t - \mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [63]$$

and we have used eqn [60]. Upon using $\sum_m \mathrm{e}^{\mathrm{i}mx} = \sum_m 2\pi\delta(x - 2m\pi)$ in the last infinite sum in eqn [64] and the property of the Dirac delta function, $\delta(ax) = (1/a)\delta(x)$, we obtain from eqns [62] and [63], the following expression for the spectrum:

$$\mathbf{F}(\omega) = \sum_{m=-\infty}^{\infty} \mathbf{F}_m \delta(\omega - m\omega_1) \qquad [64]$$

$$\omega_1 = \frac{2\pi}{T - \mathbf{n}\cdot\mathbf{r}_0/c}, \qquad [65]$$

$$\mathbf{F}_m = \frac{\omega_1}{2\pi}\int_0^T \mathrm{d}t\, \bar{\beta}(t)\mathrm{e}^{\mathrm{i}m\omega_1[t - \mathbf{n}\cdot\mathbf{r}(t)/c]} \qquad [66]$$

Note from eqn [64] that the radiation spectrum is discrete, for strictly periodic motion of the electron. The base frequency of this spectrum, ω_1, depends on the periodicity (T) of the electron, on the electron's net displacement ($\mathbf{r}_0$) in one such period, and on the direction ($\hat{n}$) in which the radiation is observed. Thus, the radiation spectrum is, in general, not at the harmonic frequency of the laser (nor at the harmonics of the electron orbital frequency, $2\pi/T$). It would thus be wrong to simply insert $\omega = n\omega_0$ in eqn [63] and to replace the electron's orbital period T there by the laser's optical period $2\pi/\omega_0$ and consider the resultant value of that integral to give the spectral amplitude of the radiation at the nth harmonic of the laser frequency. Erroneous conclusions regarding generation of high laser harmonics have appeared in the literature based on such an intuitive (but incorrect) substitution.

The power, p_m (in erg/s), radiated at the harmonic frequency $\omega = m\omega_1$ per unit solid angle in the direction of the unit vector $\hat{n}$ is then given by (from eqns [58] and [64]):

$$p_m = \frac{e^2 m^2 \omega_1^2}{4\pi^2 c} |\mathbf{n} \times \mathbf{F}_m|^2 \qquad [67]$$

where the dimensionless spectral amplitude $\mathbf{F}_m$ is given in eqn [66]. It is easy to show from eqns [60] and [65] that the integrand in eqn [66] is a periodic function of t of period T. Integrals of this type are readily evaluated by the Romberg method. The radiation spectrum observed *exactly* in the forward direction of the laser ($\hat{n} = \hat{z}$) always has only one discrete frequency, $\omega = \omega_1 = \omega_0$, which is easily shown from eqn [65] upon using the first equality of eqn [53]. This statement is true regardless of the velocity of the electron or the laser intensity, and may easily be deduced from eqn [66] for this case. However, for an energetic electron beam that is almost co-propagating with the laser, such as that produced by the laser itself, high harmonics at the laser frequency may be observed in the direction just slightly off the laser direction.

In the backscattering direction of the laser ($n = -z$), if we set $\theta_{\mathrm{in}} = 0$, one obtains the following expressions for ω_1 and p_m, the backscatter power at $\omega = m\omega_1$ (from eqn [67]):

$$\begin{aligned} \frac{\omega_1}{\omega_0} &= \left(\frac{2}{2+a^2}\right)\left(\frac{1-\beta_{z0}}{1+\beta_{z0}}\right) \\ &= \left(\frac{2}{2+a^2}\right)\gamma_0^2(1-\beta_{z0})^2 \qquad [68] \end{aligned}$$

$$p_m = \frac{A}{\gamma_0^2(1-\beta_{z0})^2}\left(\frac{\omega_1}{\omega_0}\right)^4 s_m \qquad [69]$$

where $A = e^2\omega_0^2/4\pi^2 c = 0.69[\lambda(1\ \mu\mathrm{m})]^{-2}$ erg/s, $s_m = 0$ for $m = 0, \pm 2, \pm 4, \ldots$, and for $m = \pm 1, \pm 3, \pm 5, \ldots$,

$$s_m = (a\pi)^2 m^2 \left[J_{\frac{(m-1)}{2}}(m\kappa) - J_{\frac{(m+1)}{2}}(m\kappa)\right]^2 \qquad [70]$$

$$\kappa = \frac{a^2}{2(a^2+2)} \qquad [71]$$

The $J_\nu(x)$ above are Bessel functions of the first kind of order ν. They also appear in the quantity '$[JJ]$' or '$F_m(K)$' in the FEL/synchrotron literature, where K is the undulator/wiggler parameter K. Setting $K = a$, one finds $s_m = \pi^2(1 + K^2/2)^2 F_m(K)$. It is easy to show that $s_m = s_{-m}$ for all odd integers m. Note that the relative spectral shape of s_m depends only on a, and is independent of the electron beam energy. The discrete spectrum for small a approaches a continuum for $a^2 \gg 1$.

The maximum values of s_m, occurring at $m = M$ with a value s_M, are shown in Figure 9. Note that the frequency component $\omega = M\omega_1$ contains the highest backscattered power. In terms of the laser frequency, the frequency component $\omega = N\omega_0$ would contain the highest backscatter power, where $N = M\omega_1/\omega_0$. The total backscatter power, P_{T} (in ergs/s), per unit solid angle in the $\mathbf{n} = -\mathbf{z}$ direction is then given by $P_{\mathrm{T}} = \Sigma p_m$, where the sum is taken over all odd values of m:

$$P_{\mathrm{T}} \approx \frac{\gamma_0^6(1-\beta_{z0})^6}{(\lambda/1\ \mu\mathrm{m})^3} \times \begin{cases} 13.7a^3, & (a < 0.3) \\ 11.1/a, & (a \gg 1) \end{cases} \qquad [72]$$

To summarize, due to the rapid acceleration of electrons in the direction of the light wave and decrease of the oscillation frequency in the deeply relativistic regime, the photon energy from relativistic nonlinear scattering scales only linearly with laser field strength, $\sim a$. However, the relativistic motion of the electron results in a reduction in the angle of the scattered light, such that the harmonics are generated in a low-divergence-angle forward-propagating beam. There is also a relativistic Doppler shift ($\sim\gamma^2$). For instance, this mechanism will be used in high-energy physics experiments to cleanly make constituent particles in the gamma–gamma ($\gamma\gamma$) collider, in which gamma rays with energy 200 GeV will be generated by Compton scattering 1 eV photons from 250 GeV energy conventionally accelerated electron beams ($\gamma = 5 \times 10^{11}$).

Using a much smaller, laser-based accelerator (discussed in the following section), a 1 eV photon can be upshifted by this Doppler shift to an energy of 50 keV, corresponding to subatomic spatial resolution and of interest in medical diagnostics, by an electron beam with only $\gamma = 200$ (100 MeV). In this case, the maximum efficiency is obtained for laser fields a of order unity. When compared with conventional light sources based on cm-wavelength magnetostatic wigglers, the electromagnetic wigglers of such all-optical-laser-based EUV sources have ten-thousand times shorter wavelength (micron-scale). Thus, the total length of the wiggler region is correspondingly smaller (only mm in length). Another consequence of this is that the frequency upshift required to reach a given output wavelength is also ten thousand times smaller. Also, given that the required electron energy scales as the square root of the upshift, the required electron energy can be one-hundred-times lower (10–100 MeV). It follows from this, and the fact that the field gradients of these accelerators can be ten thousand times higher (1 GeV/cm) than convention RF-based accelerators, that the size of the accelerating region can, in principle, be a million times smaller (only mm in length). Besides its small size, this EUV source can produce femtosecond duration pulse and be synchronized with a relatively low jitter with another femtosecond light pulse having a different wavelength (by virtue of the possibility of deriving the two pulses from the same laser pulse); this is advantageous for the study of ultrafast pump-and-probe photo-initiated processes. The exceptionally low transverse emittance of laser-accelerated electron beams may even make it possible to generate coherent XUV radiation by means of the self-amplified spontaneous emission (SASE) free-electron lasing (FEL) mechanism.

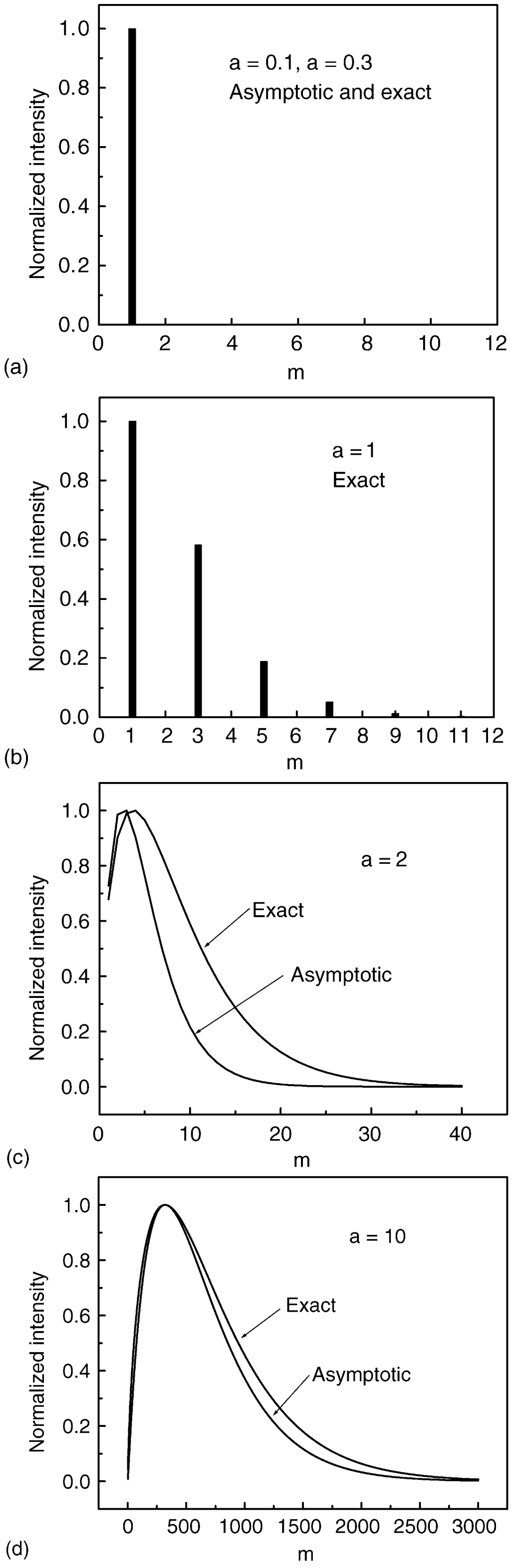

Figure 9 Normalized spectral distribution of s_m, at frequency $\omega = m\omega_1$. Here, s_m is normalized with respect to the maximum value s_M, occurring at $m = M$. (Reproduced from Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155, with permission from the American Institute of Physics.)

Collective Plasma Response

The above equations apply to electron motion in vacuum. In the case of electrons in plasmas, the collective plasma electron response needs to be considered. A theory for the 1D nonlinear interaction of intense laser fields with electrons has been developed.

The variables are normalized such that

$$\mathbf{a} = e\mathbf{A}/mc^2, \qquad \phi = e\Phi/mc^2 \qquad [73]$$

$$\mathbf{u} = \mathbf{p}/mc, \qquad \beta = \mathbf{v}/c, \qquad \mathbf{u} = \gamma\beta \qquad [74]$$

$$\gamma = (1 + u^2)^{1/2} = (1 - \beta^2)^{-1/2} \qquad [75]$$

The relativistic force equation (Lorentz equation) is given by

$$\frac{d}{dt}\mathbf{u} = \nabla\phi + \frac{\partial}{\partial t}\mathbf{a} - \bar{\beta}\times(\nabla\times a) \qquad [76]$$

The energy equation ($\mathbf{u}\cdot$force equation) is given by

$$\frac{d}{dt}\gamma = \beta\cdot\left(\frac{\partial}{\partial z}\phi + \frac{\partial}{\partial t}\mathbf{a}\right) \qquad [77]$$

The transverse force equation (1D) is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [78]$$

Since $\mathbf{a} = \mathbf{a}(z, t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{a}_\perp = \left(\frac{\partial}{\partial t} + \beta_z \frac{\partial}{\partial z}\right)\mathbf{a}_\perp \qquad [79]$$

hence:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{u}_\perp - \mathbf{a}_\perp) = 0 \qquad [80]$$

This is just conservation of transverse canonical momentum. Assuming $\mathbf{u}_\perp = 0$ prior to the laser interaction gives

$$\mathbf{u}_\perp = \mathbf{a}_\perp \qquad [81]$$

which is the same as eqn [24].

The electron response to the normalized scaler and vector potentials of the form $\phi = \phi(z - ct)$ and $\mathbf{a} = \mathbf{a}(z - ct)$, which are a function of only $\xi = z - ct$, is completely described by the following constants of the motion:

$$\beta_\perp - a_\perp/\gamma = 0, \qquad [82]$$

$$\gamma(1 - \beta_z) - \phi = 1, \qquad [83]$$

$$n(1 - \beta_z) = n_0, \qquad [84]$$

where $\gamma = (1 - \beta^2)^{-1/2}$ is the relativistic factor and n is the electron density. This allows the various electron quantities to be specified solely in terms of the potentials:

$$\beta_\perp = \frac{2(1+\phi)a_\perp}{(1 + a_\perp^2) + (1+\phi)^2}, \qquad [85]$$

$$\beta_z = \frac{(1 + a_\perp^2) - (1+\phi)^2}{(1 + a_\perp^2) + (1+\phi)^2}, \qquad [86]$$

$$\gamma = \frac{(1 + a_\perp^2) + (1+\phi)^2}{2(1+\phi)}, \qquad [87]$$

$$\frac{n}{n_0} = \frac{(1 + a_\perp^2) + (1+\phi)^2}{2(1+\phi)^2} \qquad [88]$$

In the single particle limit, $\phi \equiv 0$, eqns [26] and [30] are retrieved. For a long pulse interacting with a dense plasma, $\tau_\mathrm{L} \gg \omega_\mathrm{p}^{-1}$ (i.e., neglecting wakefield effects) it can be shown that $(1 + \phi) \simeq (1 + a_0^2)^{1/2}$ for circular polarization and $(1 + \phi) \simeq (1 + a_0^2/2)^{1/2}$ for linear polarization.

Propagation

To understand the propagation of high-intensity light in plasma, we need to understand how the dielectric properties of a plasma medium are affected by the relativistic electron mass change. A wave equation is found by taking the curl of Faraday's law:

$$\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \qquad [89]$$

and using Ampere's law:

$$\nabla \times \mathbf{B} = 4\pi \mathbf{J}/c + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} \qquad [90]$$

gives

$$\begin{aligned}\nabla \times \nabla \times \mathbf{E} &= -\nabla^2 \mathbf{E} + \nabla(\nabla \cdot \mathbf{E}) \\ &= \frac{1}{c^2}\left[4\pi \frac{\partial \mathbf{J}}{\partial t} - \frac{\partial^2 \mathbf{E}}{\partial t^2}\right]\end{aligned} \qquad [91]$$

where $\mathbf{J} = \sum n_i q_i v_i$ and the summation over the index i is done over all charged species of electrons and ions. Using Coulomb's law:

$$\nabla \cdot \mathbf{E} = 4\pi\rho \qquad [92]$$

where $\rho = \sum n_i q_i$, and assuming a uniform plasma, the second term on the left vanishes. Substituting eqn [13] for the velocity, we find

$$\left(-\frac{\partial^2}{\partial t^2} + c^2\nabla^2 - \omega_\mathrm{p}^2\right)\mathbf{E} = 0 \qquad [93]$$

where

$$\omega_\mathrm{p} = (4\pi n_i q_i^2/m_i)^{1/2} \qquad [94]$$

is the plasma frequency. Assuming again plane waves and Fourier analyzing, yields the well-known dispersion relation for electromagnetic waves in plasma:

$$\omega^2 = \omega_\mathrm{p}^2 + c^2k^2 \qquad [95]$$

The index of refraction can be written fully relativistically as

$$\eta^2 = \frac{c^2k^2}{\omega^2} = \frac{c^2}{v_\phi^2} = 1 - \frac{\omega_\mathrm{p}^2}{\omega^2} \qquad [96]$$

where v_ϕ is the phase velocity of the light wave. Assuming infinitely massive ions, the plasma frequency can be written as

$$\omega_\mathrm{p} = \frac{\omega_{\mathrm{p}0}}{\gamma^{1/2}} = \left(\frac{4\pi n_\mathrm{e} e^2}{\gamma m_0}\right)^{1/2} \qquad [97]$$

where $\omega_{\rm p0}$ is the plasma frequency in a quiescent plasma, e is the electron charge, m_0 is the electron rest mass, and $n_{\rm e}$ is the plasma electron density. A change in mass changes the plasma frequency, which in turn modifies the index of refraction and the velocity of the light wave. The light's phase velocity then depends on the laser intensity. This can be seen clearly if we expand the phase velocity for small field strength ($a_0 \ll 1$):

$$v_\phi = \frac{c}{\eta} \sim c\left[1 + \frac{\omega_{\rm p0}^2}{2\omega^2}\left(1 - \frac{\langle a_0^2\rangle}{2}\right)\right] \qquad [98]$$

An on-axis minimum of the phase velocity [i.e., $v_\phi(r) > v_\phi(0)$] can be created by a laser beam with an intensity profile peaked on axis, such as with a Gaussian beam, causing the wavefronts to curve inward and the laser beam to focus, as shown in Figure 10. When this focusing effect just balances the defocusing caused by diffraction, the laser pulse can propagate over a longer distance than it could in vacuum, while maintaining a small cross-section. This mechanism is referred to as relativistic self-guiding.

Relativistic Self-Focusing

The threshold for relativistic self-focusing can be obtained using the fully relativistic formalism with the vector potential, as in our previous discussion of electron motion in vacuum. In this case, we can rewrite eqn [91] as

$$\nabla^2\mathbf{a} - \frac{1}{c^2}\frac{\partial^2\mathbf{a}}{\partial t^2} = \frac{\omega_{\rm p}^2}{c^2}\frac{n_{\rm e}\mathbf{a}}{n_0\gamma} \qquad [99]$$

where $n_0 = n_i$ is the uniform background density. For a Gaussian beam focused in vacuum, where

$$\mathbf{a}(\mathbf{r}, t) = \mathrm{Re}\left\{\mathbf{a}_0(\mathbf{r}, t)\exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{r} - \omega t)]\right\} \qquad [100]$$

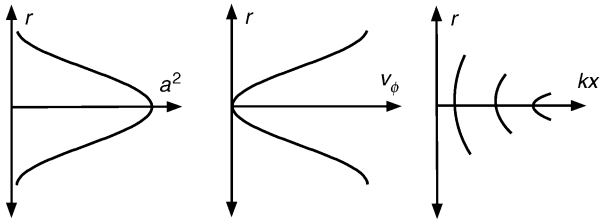

Figure 10 Relativistic self-focusing occurs when an on-axis peak in light intensity (left) produces an on-axis dip in the phase velocity (middle), which acts like a positive lens to cause the phase fronts to curve inward (right). (Reproduced from Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165, with permission from IOP Publishing Ltd.)

and the amplitude $\mathbf{a}(\mathbf{r}, t)$ varies much less less in r and t than does the phase:

$$|\partial a_0/\partial t| \ll |\omega a_0|, \qquad |\partial a_0/\partial z| \ll |k a_0| \qquad [101]$$

then

$$a_0(r, z) = \frac{e^{-r^2/(r_0^2(1+z^2/L_{\rm r}^2))}}{\sqrt{1 + z^2/L_{\rm R}^2}}\exp\left\{-\,\mathrm{i}\arctan\left(\frac{z}{L_{\rm R}}\right) + \mathrm{i}\left(\frac{r}{r_0}\right)^2\frac{z/L_{\rm R}}{1 + z^2/L_{\rm R}^2}\right\} \qquad [102]$$

satisfies the envelope equation (assuming second deriviatives are small compared to first derivatives), where $L_{\rm R}$ is the Rayleigh range and r_0 is the focal radius. In plasma, eqn [99] can then be written in this envelope approximation as

$$\left[\nabla_\perp^2 + 2ik\frac{\partial}{\partial z}\right]a = -\frac{\omega_{\rm p}^2}{c^2}\frac{|a|^2}{4}a \qquad [103]$$

The power at which the first and last terms of eqn [103] are in balance is

$$P = \frac{\pi R^2 I_0}{2} \qquad [104]$$

An equation for the beam radius R can thus be found:

$$\frac{\mathrm{d}^2 R(z)}{\mathrm{d}z^2} = \frac{4}{k^2 r_0^2}\left[1 - \frac{1}{22}\frac{\omega_{\rm p}^2}{c^2}|a_0|^2 r_0^2\right] \qquad [105]$$

Self-focusing occurs when the two terms in the brackets in eqn [105] are equal, which gives for the critical power:

$$P_{\rm crit} \simeq 17.4\cdot(\omega/\omega_{\rm p})^2\ \mathrm{GW} \qquad [106]$$

This corresponds to 1 TW for 1 μm light focused into a gas with electron density of $10^{19}/\mathrm{cm}^3$. Note that this is a power threshold, not an intensity threshold, because the tighter the focusing, the greater the diffraction (the first term in eqn [103]). Numerous recent experiments have confirmed this focusing mechanism when $P > P_{\rm crit}$.

Any spatial variation of the laser intensity will act to push an electron to regions of lower intensity through the so-called ponderomotive force:

$$\mathrm{F} = \nabla\gamma = \nabla\sqrt{1 + a^2} = (2\gamma)^{-1}\nabla a^2 \qquad [107]$$

$$\mathbf{F}_{\text{pond}} = -\frac{m_0c^2}{8\pi}\frac{\nabla a_0^2}{\sqrt{1+\frac{a_0^2}{2}}}\cdot\left[\left(\frac{2}{a_0^2}+1\right)E(2\pi,\kappa) - \frac{2}{a_0^2}F(2\pi,\kappa)\right] \qquad [108]$$

where

$$\kappa = \frac{a_0^2/2}{1+a_0^2/2} \qquad [109]$$

and F and E are the elliptic integrals of the first and second kind, respectively. In the low intensity limit, i.e., when $a_0^2 \ll 1$;

$$\mathbf{F}_{\text{pond}} = -\frac{m_0c^2}{4}\nabla a_0^2 \qquad [110]$$

That is, the laser ponderomotive force is roughly proportional to the gradient of laser intensity.

A Gaussian-shaped laser's ponderomotive force will tend to expel electrons radially from the region of the axis, so called 'electron cavitation.' If the ponderomotive force is high enough for long enough, the charge displacement due to expelled electrons (in either the lateral or longitudinal direction) will eventually cause the ions to move as well through the Coulomb electrostatic force, forming a density channel. Because $n_e(0) < n_e(r)$, and thus $v_\phi(0) < v_\phi(r)$, this enhances the previously-discussed relativistic self-guiding and can itself guide a laser pulse.

For plasmas created by photo-ionization of a gas by a Gaussian laser pulse, the density will be higher on the axis than off the axis. If we instead expand the phase velocity in terms of changes in density:

$$\frac{v_\phi}{c} \sim 1 + \frac{\omega_{p0}^2}{\omega^2}\frac{\Delta n_e}{n_e} \qquad [111]$$

The phase velocity will thus be higher on axis, which will tend to defocus the light and increase the self-guiding threshold. In order to avoid this, gases with low atomic number and thus fewer available electrons, such as H_3, are commonly used as targets.

Raman Scattering, Plasma Wave Excitation and Electron Acceleration

The local phase velocity, described in eqns [98] and [111], can also vary longitudinally if the intensity and/or electron density does. Local variation in the index of refraction can 'accelerate' photons, i.e., shift their frequency, resulting in photon bunching, which in turn bunches the electron density through the ponderomotive force (F), and so on. When the laser pulse duration is longer than an electron plasma period, $\tau \gg \tau_p = 2\pi/\omega_p$, this photon and electron bunching grows exponentially, leading to the stimulated Raman scattering instability. Energy and momentum must be conserved when the electromagnetic wave $(\omega_0, \mathbf{k}_0)$ decays into a plasma wave $(\omega_p, \mathbf{k}_p)$ and another light wave $(\omega_0 - \omega_p, \mathbf{k}_0 - \mathbf{k}_p)$.

From an equivalent viewpoint, the process begins with a small density perturbation, Δn_e, which, when coupled with the quiver motion, eqn [13], drives a current $J = \Delta n_e e v_e$. This current then becomes the source term for the wave equation (eqn [91]), driving the scattered light wave. The ponderomotive force, due to the beating of the incident and scattered light wave, enhances the density perturbation, creating a plasma wave and the process begins anew. In three dimensions, a plasma wave can be driven when transverse self-focusing and stimulated Raman scattering occur together, a process called the self-modulated wakefield instability.

Two conditions must be satisfied for self-modulation to occur in the plasma. First, the laser pulse must be long compared to the plasma wave, $L \gg \lambda_p$. This allows the Raman instability time to grow, and it allows for feedback from the plasma to the laser pulse to occur. Second, the laser must be intense enough for relativistic self-focusing to occur, $P > P_c$, so that the laser can be locally modified by the plasma. Under these conditions, the laser can form a large plasma wave useful for accelerating electrons.

As the long laser pulse enters the plasma, it will begin to drive a small plasma wave due to either forward Raman scattering or the laser wakefield effect from the front of the laser pulse. This small plasma wave will have regions of higher and lower density with both longitudinal and radial dependence. That is, the plasma wave will be three-dimensional in nature with a modulation along the propagation direction of the laser and a decay in the radial direction to the ambient density (see Figure 11). The importance of this lies with how it affects the index of refraction in the plasma. In the regions of the plasma wave where the plasma density is lower, the radial change in the index of refraction is negative, $\partial n(r)/\partial r < 0$. This means that this part of the plasma acts like a positive lens and focuses the laser. Whereas regions of the plasma wave where the density is higher, $\partial n(r)/\partial r > 0$, the opposite occurs and the laser defocuses. This has the effect of breaking up the laser pulse into a series of shorter pulses of length $\lambda_p/2$ which will be separated by the plasma period. The instability occurs because of how the plasma responds to this. Where the laser is more tightly focused, the ponderomotive force will be

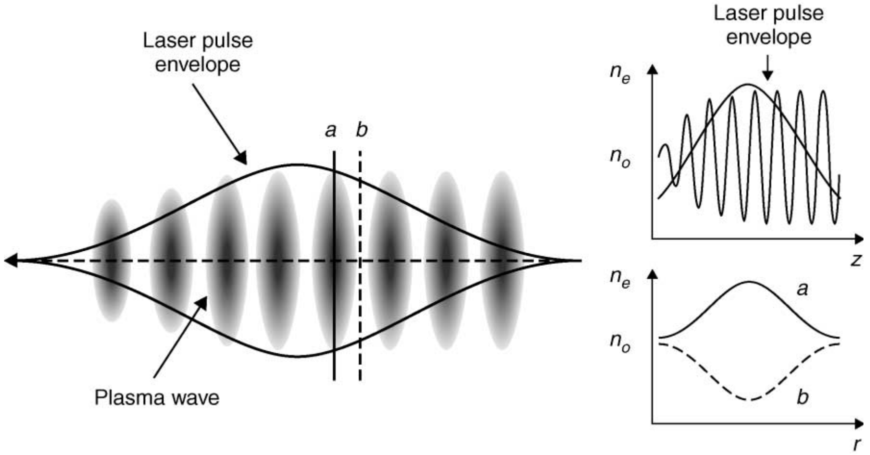

Figure 11 The plasma wave generated by a SMLWFA is three-dimensional in nature. Note that the darker regions correspond to areas of higher plasma density. The graphs to the right represent lineouts of the plasma density longitudinally and radially at the indicated points. (Reproduced with permission from Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.)

greater and will tend to expel more electrons. This decreases the density in these regions even further, resulting in more focusing of the laser. This feedback rapidly grows, hence the instability.

The phase velocity of the plasma wave in the case of forward scattering is equal to the group velocity of the beat wave, which for low-density plasma is close to the speed of light, as can be seen from the relation:

$$v_\phi = \omega_p/k_p = \Delta\omega/\Delta k = v_g = c\eta \sim c \qquad [112]$$

where eqn [94] and $\omega_p^2 \ll \omega^2$ were used to show that η is close to unity. Such relativistic plasma waves can also be driven by short pulses ($\tau \sim \tau_p$). In this case, the process is referred to as laser–wakefield generation, referring to the analogy with the wake driven by the bow of a boat moving through water, but the mechanism is similar (except it has the advantage that the plasma wave is driven linearly instead of as an instability).

In either case, the resulting electrostatic plasma wave can continuously accelerate relativistic electrons with enormous acceleration gradients. The gradient can be estimated from eqn [92] and the fact that because

$$\nabla \cdot \mathbf{E} \sim E\omega_p/c \propto E\sqrt{n_e} \qquad [113]$$

then

$$E \sim \sqrt{n_e}\ \text{eV/cm} \qquad [114]$$

corresponding to 1 GeV/cm for $n_e = 10^{18}\ \text{cm}^{-3}$. Because this gradient is four orders of magnitude greater than achieved by conventional accelerators (based on fields driven by radio-frequency waves pumped into metal cavities), laser-driven plasma accelerators have received considerable recent attention. They have been shown to accelerate an amount of electron charge (100 nC) comparable to that from conventional accelerators and to have superior transverse geometrical emittance (product of divergence angle and spotsize, similar to the $f/\#$ in light optics). However, their longitudinal emittance is currently much inferior, energy spreads of 100%. They have been shown to be useful for much of the same applications: radio-isotope production, radiation chemistry, as well as X-ray, proton, and neutron generation. Once the longitudinal emittance can be reduced, they may be advantageous for, among other applications, injectors (especially of short-lived unstable particles) into larger conventional accelerators for high-energy physics research and light sources, and, as discussed in the section on radiation from relativistic electrons about, as stand-alone all-optically driven ultrashort-pulse duration X-ray sources.

The SIMLAC code has been used to study wakefield generation and laser propagation in the limit $a^2 \ll 1$. It draws from nonlinear optics models and treats propagation in the group velocity frame. In this idealized model (which assumes perfect Gaussian beams), the pulse and wake are maintained over long enough propagation distances to accelerate an electron to GeV energy, as shown in Figure 12. A three-dimensional envelope equation for the laser field was derived that includes nonparaxial effects, wakefields, and relativistic nonlinearities.

The resonant wakefield has been characterized experimentally by temporal interferometry, as shown in Figure 13. However, this was done only for the tight-focusing case in which the laser spotsize is much

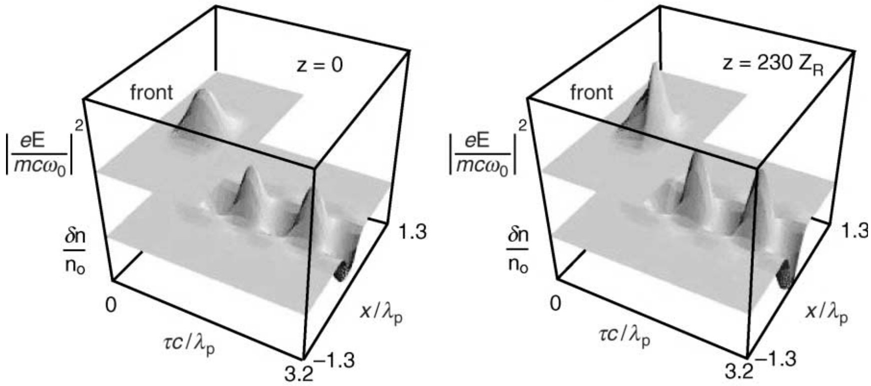

Figure 12 The 'standard' resonant wakefield simulated with SIMLAC, a code that moves at the light pulse's group velocity. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.)

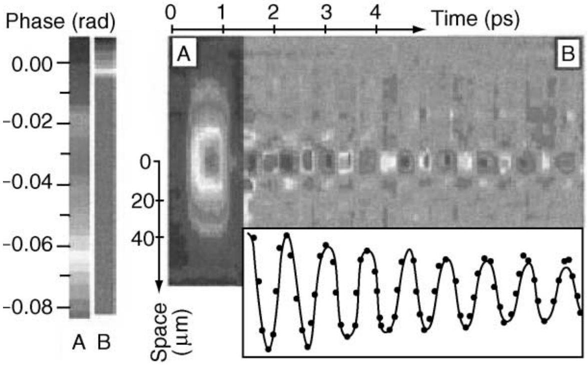

Figure 13 Typical result of a phase shift measurement to study a resonantly excited laser wakefield plasma wave by means of time-domain interferometry. Parts (A) and (B) have different color scales. The bottom graph is a line out of part (B) along the laser axis. (Reproduced with permission from Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463. Copyright (1995) by the American Physical Society.)

smaller than the plasma wave wavelength ($r_1 \ll \lambda_p$) and thus the transverse wakefield was much greater than the longitudinal wakefield.

A typical experimental setup, used to study electron acceleration, is shown in Figure 14.

Dramatic reduction of the angular divergence of a laser accelerated electron beam was observed with increasing laser power above the relativistic self-focusing threshold, as shown in Figure 15.

Interactions with Solid-Density Targets

The generation of electrons by high-intensity laser light interacting with solid targets can generate energetic X-rays, accelerate other types of particles and induce nuclear reactions, as illustrated in Figure 16. For instance, high-order harmonics have been generated by the oscillation of the critical density surface, in the so-called moving mirror model. Bright X-rays, originating from Bremstrahlung caused by electron collisions with high-Z atoms in solid targets, have created isotopes by means of photofission. Laser-accelerated electron energies and

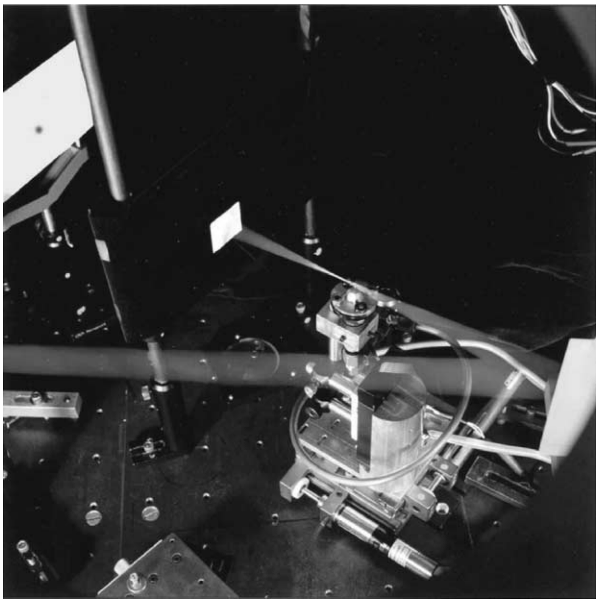

Figure 14 Artistically enhanced photograph of the acceleration of an electron beam by a laser interacting with a gas jet inside a vacuum chamber. The laser crosses the picture from left to right and is focused by a parabolic mirror (right side of the picture). The supersonic nozzle (shown in the middle of the picture) is positioned with micron accuracy with a 3-axis micropositioner. The e-beam makes a small spot on a white flourescent (LANEX) screen, shown in the upper left-hand corner of the picture. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

angular distributions have been inferred from analyzing (γ, n) and $(\gamma, 2n)$ reactions in composite Pb/Cu targets and in Ta/Cu targets. Positrons were created by colliding laser-accelerated electrons with a tungsten target.

When electrons are heated to high temperatures or accelerated to high energies, they can separate from plasma ions. Such charge displacement creates an electrostatic sheath, which eventually accelerates the ions. The ions are pulled by the charge of the electrons and pushed by the other ions' unshielded charges (similar to the 'Coulomb explosion' that can occur during the ionization of atoms). When the charge displacement is driven by thermal expansion, as in long-pulse (low-power) laser–plasma experiments, the maximum ion energies are limited to less than 100 keV. However, when the charge displacement is driven by direct laser heating, as in short-pulse high-power laser–plasma experiments, multi-MeV ion energies are possible. This was first shown with gas jet targets, in which case the ions were accelerated radially into 2π, and then later with thin solid-density-films, in which case the ions were accelerated into collimated beams. In the latter case, hydrocarbons and water on the surface of the film can become ionized and provide a source of protons to be accelerated.

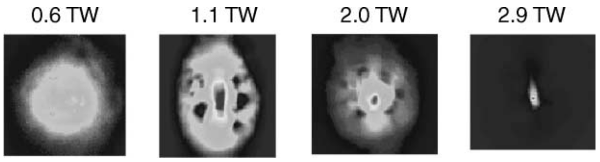

Figure 15 Images of the spatial profiles of the electron beam measured by a ccd camera imaging a LANEX screen at a distance of 15 cm from the gas jet for various laser powers. The divergence angle of the beam decreases to a value of $\Delta\theta = 1°$ at a power of 2.9 TW, corresponding to a transverse geomtrical emittance of just $\varepsilon_\perp \lesssim 0.06\pi$ mm-mrad. (Reproduced with permission from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics.)

An intense laser can ponderomotively heat electrons. If the laser contrast is high, vacuum heating can occur in the following manner. When light encounters a sharp interface between vacuum and solid density, the electromagnetic field becomes evanescent in the region above the critical density. The instantaneous '$v \times B$' force can push electrons in the direction of the light's propagation vector; it also has a frequency twice that of the pump and a magnitude proportional to the square of the normalized vector potential, a_0^3. Thus, electrons can only complete half of their figure-eight orbits, on the vacuum side, gaining relativistic energies; they move through the overdense region without the electromagnetic field to pull them back. An electrostatic sheath can thus form, which will accelerate the ions left behind. Another important heating mechanism is stochastic heating, which occurs when the light that is reflected from the critical surface beats with the incoming wave to create a standing wave. The motion of electrons in such a wave can become chaotic, resulting in a large increase in electron temperature (>100 keV).

As the heated electrons propagate through a solid, they can instantaneously field-ionize the neutral atoms of the solid. This will both modify the solid's

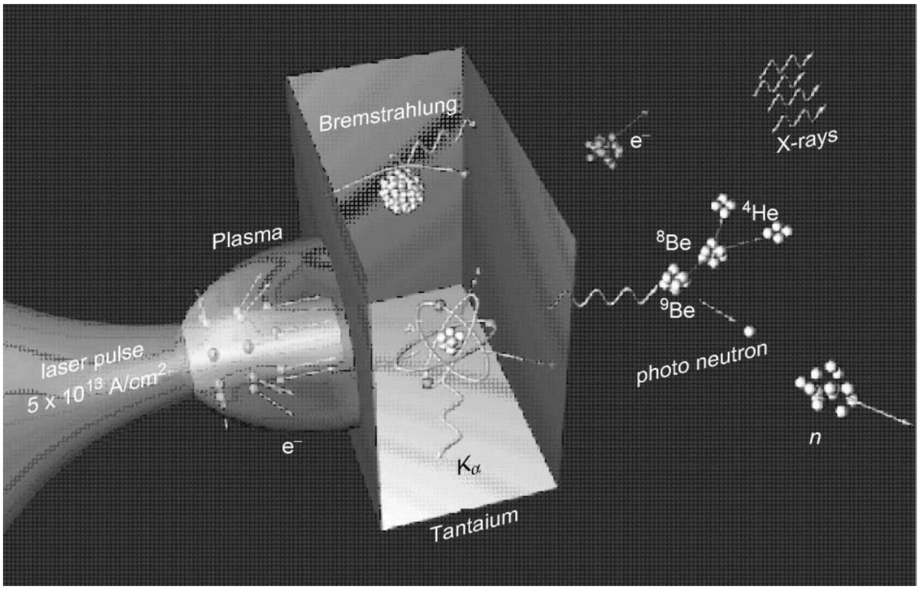

Figure 16 Illustration of the various mechanisms of particle acceleration and X-ray generation in high-intensity laser interactions with solid-density targets. (Courtesy of R. Sauerbrey.)

conductivity and provide a source of protons on the rear-side of the target. If the film is thin enough, the electrons can pass through, and create a sheath on the rear-side of the target. This latter mechanism has been dubbed the target normal sheath acceleration mechanism. The ions from thin foils have been claimed to originate from both the front and rear-side of the foil.

Several groups have reported the observation of ions originating from thin-film solid-density targets. Unlike previous long-pulse experiments, the ions were accelerated along the direction normal to the side of the target, that is opposite to that upon which the laser was incident. The ions generally originate from water or hydrocarbons on the surface of the material. The acceleration results from several different mechanisms, which may be occurring simultaneously. Charge-displacement is again common to all, with the electrons being heated ponderomotively, such as by Brunel, $J \times B$ or stochastic heating. In one case, the electrostatic sheath is formed at the backside of the ionization layer formed on the side of the target upon which the laser is incident (front side). In another case, the electrostatic sheath is formed by field ionization of the ion layer on the opposite side of the thin film target (back-side), the target normal sheath acceleration (TNSA) hypothesis. Numerical simulations show evidence for both front- and back-side acceleration.

The results of these experiments indicate that a large number of protons ($10^{13}\, p$) can be accelerated, corresponding to source current densities (10^8 A/cm^2) that are nine orders of magnitude higher than produced by cyclotrons, but with comparable, or even lower, transverse emittances ($\varepsilon_\perp \leq 1.0\pi$ mm-mrad). Proton energies up to 60 MeV have been observed in experiments at intensities exceeding 10^{20} W/cm^2 (using a petawatt power laser). The high end of the proton spectrum typically has a sharp cut-off, but is a continuum. In one experiment, protons were observed to be emitted in ring patterns, the radii of which depend on the proton energy, which was explained by self-generated magnetic fields.

A typical experimental setup used to observe the acceleration of ions is shown in Figure 17. The production of radionuclides have been used as an ion energy diagnostic. In another example of a nuclear reaction initiated by an intense laser, neutrons

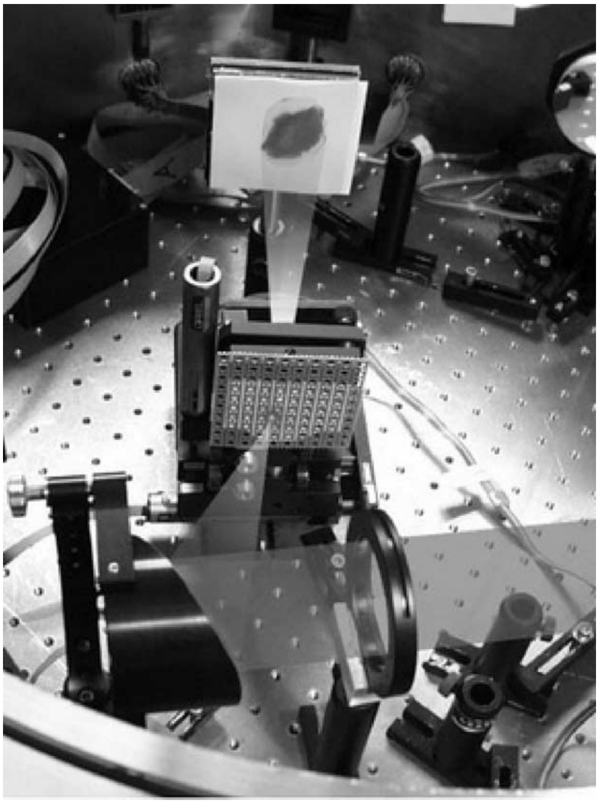

Figure 17 Artistically enhanced photograph of the typical setup used to observe the acceleration of ions. The laser (shown in the foreground) is focused with an off-axis parabola onto a thin-foil, held by a mesh that is positioned by a 3-axis micropositioner. CR-39, a nuclear track detector (shown in the background) is used to detect the ions. An actual proton-produced pattern is shown in red. (Reproduced from Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774, with permission from the American Institute of Physics).

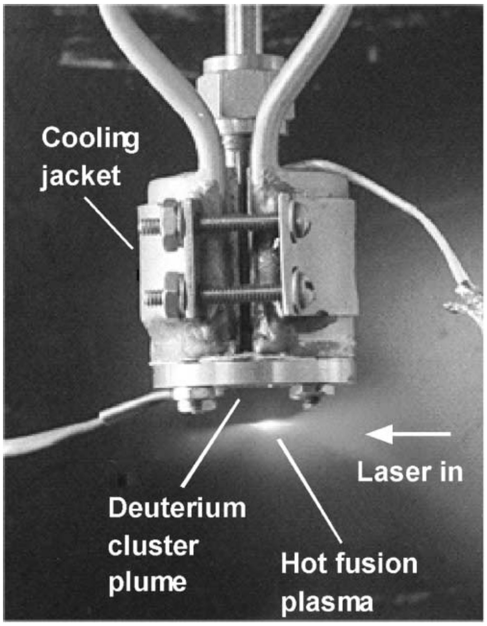

Figure 18 Photograph of a laser interacting with a deuterium cluster jet emanating from cooled gas nozzle. Fusion neutrons are produced with relatively low laser intensities. (Reproduced with permission from Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28, with permission from IEEE.)

have been produced by the He fusion reaction in the focus of 200 mJ, 160 fs Ti:sapphire laser pulses on a deuterated polyethylene target. Optimizing the fast electron and ion generation by applying a well-defined prepulse led to an average rate of 140 neutrons per shot. Neutrons have also been generated from cluster plasmas, which were produced by the cooled-nozzle depicted in Figure 18, but with significantly lower laser intensities than required with planar solid targets.

Concluding Remarks

The field of nonlinear optics, with electrons bound to atoms, has over the last forty years given rise to many scientific discoveries and technologies that are now commonplace. The relatively young field of relativistic nonlinear optics has already begun to do the same. Some of the exciting phenomena and applications that have already been identified have been discussed in this article. Others await discovery as ever higher laser intensities are reached, with no foreseeable limit.

See also

Fiber and Guided Wave Optics: Nonlinear Optics. **Scattering:** Raman Scattering.

Further Reading

Chen S-Y, Maksimchuk A and Umstadter D (1998) Experimental observation of relativistic nonlinear Thomson scattering. *Nature* 396: 653.

Cowan TE, Hunt AW, Phillips TW, Wilks SC, Perry MD, Brown C, Fountain W, Hatchett S, Johnson J, Key MH, Parnell T, Pennington DM, Snavely RA and Takahashi Y (2000) Photonuclear fission from high energy electrons from ultraintense laser–solid interactions. *Physics Review Letters* 84: 903.

Ditmire T (2002) Laser fusion on a tabletop. *Optics and Photonics News* 13: 28.

Eberly JH (1986) Fundamentals of laser interactions, vol. 229 of lecture notes in physics. *Journal Optical Society of America B* 3: 1324.

Esarey E, Sprangle P, Krall J and Ting A (1996) Overview of plasma-based accelerator concepts. *IEEE Transactions on Plasma Science* 24: 252.

Gibbon P and Förster E (1996) Short-pulse laser–plasma interactions. *Plasma Phys. Controlled Fusion* 38: 769.

Hartemann F (2001) *High-Field Electrodynamics*. Boca Raton: CRC Press.

Joshi CJ and Corkum PB (1995) Interactions of ultra-intense laser light with matter. *Physics Today* 48: 36.

Lau Y, He F and Umstadter D (2003) Nonlinear Thomson scattering: A tutorial. *Physics Plasmas* 10: 2155.

Ledingham KWD, Spencer I, McCanny T, Singhal RP, Santala MI, Clark E, Watts I, Beg FN, Zepf M, Krushelnick K, Tatarakis M, Dangor AE, Norreys PA, Allott R, Neely D, Clark RJ, Machacock AC, Wark JS, Cresswell AJ, Sanderson DCW and Magill J (2000) Photonuclear physics when a multiterawatt laser pulse interacts with solid targets. *Physics Review Letters* 84: 899.

Luther-Davies B, Gamaly EG, Wang Y, Rode AV and Tikhonchuk V (1992) Matter in ultrastrong laser fields. *Soviet Journal Quantum Electronics* 22: 289.

Marquès JR, Dorchies F, Audebert P, Ceindre JP, Amiranoff F, Gauthier JC, Hammoniaux G, Antonetti A, Chessa P, Mora PTM and Antonson J (1997) Frequency increase and damping of nonlinear electron plasma oscillations in cylindrical symmetry. *Physics Review Letters* 78: 3463.

Meyer-ter-Vehn J, Pukhov A and Sheng Z-M (2001) Relativistic laser plasma interaction. In: Batani D (ed.) *Atoms, Solids and Plasmas in Super-Intense Laser Fields*, pp. 167. New York: Kluwer Academic–Plenum.

Moore CI, Knauer JP and Meyerhofer DD (1995) Observation of the transition from Thomson to Compton scattering in multiphoton interactions with low-energy electrons. *Physics Review Letters* 74: 2439.

Mourou G and Umstadter D (1992) Development and applications of compact high-intensity lasers. *Physics Fluids B* 4: 2315.

Mourou GA and Umstadter D (2002) Extreme light. *Scientific America* 81.

Mourou GA, Barty CPJ and Perry MD (1998) Ultrahigh-intensity lasers: Physics of the extreme on a tabletop. *Physics Today* 22.

Siders CW, Blanc SPL, Fisher D, Tajima T, Downer MC, Babine A, Stepanov A and Sergeev A (1996) Laser wakefield excitation and measurement by femtosecond longitudinal interferometry. *Physics Review Letters* 76: 3570.

Sprangle P, Peñano JR, Hafizi B, Hubbard RF, Ting A, Gordon DF, Zigler A and Antonsen TM (2002) GeV acceleration in tapered plasma channels. *Physics Plasmas* 9: 2364.

Tajima T and Mourou G (2002) Zettawatt-exawatt lasers and their applications in ultrastrong-field physics. *Physics Review STAB* 5: 031301.

Umstadter D (2001) Review of physics and applications of relativistic plasmas driver by ultra-intense lasers. *Physics Plasmas* 8: 1774.

Umstadter D (2003) Relativistic laser–plasma interactions. *Journal Physics D: Applied Physics* 36: R151–R165.

Umstadter D and Norris TB (eds) (1997) Nonlinear optics with relativistic electrons. *IEEE Journal Quantum Electronics* 33: 1878.

Wagner R (1998) Laser–plasma electron accelerators and nonlinear relativistic optics. PhD thesis, University of Michigan.

Wagner R, Chen S-Y, Maksimchuk A and Umstadter D (1997) Electron acceleration by a laser wakefield in a relativistically self-guided channel. *Physics Review Letters* 78: 3125.

Yamakawa K (2004) Table-top lasers create ultrahigh peak powers. *OYOBUTURI* 73: 186–193.

S

SCATTERING

Contents

Raman Scattering
Scattering from Surfaces and Thin Films
Scattering Phenomena in Optical Fibers
Scattering Theory
Stimulated Scattering

Raman Scattering

F Kannari, Keio University, Yokohama, Japan

© 2005, Elsevier Ltd. All Rights Reserved.

When an electromagnetic wave passes over a small elastically bound charged particle, the particle will be set into motion by the electric field. Scattering takes place for frequencies not corresponding to the natural frequencies of the particles. If the particle is bound by a force obeying Hooke's law, this vibration will have the same frequency and spatial direction as that of the electric force in the electromagnetic wave. The phase, on the other hand, of the forced vibration will differ from that of the incident wave, which causes the basis of dispersion. If the charged oscillation is bound by a force which does not obey Hooke's law, but some more complicated law, it will reradiate not only the impressed frequency, but also various combinations of this frequency with the fundamental and overtone frequencies of the oscillator.

The Raman effect was discovered by C.V. Raman (1888–1971). Ordinary Raman spectroscopy is used as a tool for studying the vibrational energy levels of molecules and of lattice optical branch vibration in crystals. A spectroscopic measurement of the scattered light reveals the existence of frequencies shifted down by increments corresponding to vibrational frequencies of the material. This down-shifted scattering is referred to as Stokes scattering. Frequencies shifted up by the vibrational frequencies are also present in the scattered radiation, which are the so-called anti-Stokes scattering.

Stimulated amplification of scattered Stokes radiation occurs through a third-order nonlinear process. To introduce stimulated Raman scattering (SRS), first we describe the steady-state stimulated scattering theory.

In general, nonlinear optical processes are described by the classical Maxwell equation with the nonlinear polarization of a material $\mathbf{P}^{\mathrm{NL}}$:

$$\nabla \times \nabla \times \mathbf{E} + \varepsilon\mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\mu_0 \frac{\partial^2 \mathbf{P}^{\mathrm{NL}}}{\partial t^2} \qquad [1]$$

where ε and μ_0 are the linear dielectric constant of the material and the magnetic permeability of the vacuum. We consider a relatively intense excitation radiation $\mathbf{E}_{\mathrm{L}}$ (frequency ω_{L}) and a weak signal radiation $\mathbf{E}_{\mathrm{S}}$, having the carrier frequency ω_{s} corresponding to the scattered light frequency ($\omega_{\mathrm{s}} = \omega_{\mathrm{L}} - n\omega_0$);

$$\mathbf{E}_{\mathrm{L}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{L}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{L}} \times \mathbf{r} - \omega_{\mathrm{L}} t) + \mathrm{c.c.}\} \qquad [2]$$

$$\mathbf{E}_{\mathrm{S}} = \frac{1}{2}\{\bar{\mathbf{E}}_{\mathrm{S}}(\mathbf{r}, t) \exp \mathrm{i}(\mathbf{k}_{\mathrm{S}} \times \mathbf{r} - \omega_{\mathrm{S}} t) + \mathrm{c.c.}\} \qquad [3]$$

where $\bar{\mathbf{E}}_L(\mathbf{r}, t), \bar{\mathbf{E}}_S(\mathbf{r}, t)$ are a slowly varying envelope of the complex electric fields. Assuming that the variation of the electric field during the time-scale shorter than the decay time of the induced oscillation in the material is negligible, nonlinear polarization induced by the excitation radiation is described as follows:

$$\mathbf{P}_S^{NL} = \frac{1}{2}\{\bar{\mathbf{P}}_S^{NL}(\mathbf{r}, t)\exp \mathrm{i}(\mathbf{k}_S \times \mathbf{r} - \omega_S t) + \mathrm{c.c.}\} \qquad [4]$$

$$\bar{\mathbf{P}}_S^{NL} = \varepsilon_0 \chi^{NL}(-\omega_S, -\omega_L, \omega_L, \omega_S)\bar{\mathbf{E}}_L^* \bar{\mathbf{E}}_L \bar{\mathbf{E}}_S \qquad [5]$$

where ε_0 is the vacuum dielectric constant and $\chi^{NL}(-\omega_S, -\omega_L, \omega_L, \omega_S)$ is the third-order nonlinear susceptibility, of which the real part corresponds to the variation of the dielectric constant depending on the optical radiation, whereas the imaginary part describes the stimulated scattering. In steady-state approximation, substituting eqns [3]–[5] into [1], we obtain

$$\frac{\mathrm{d}\bar{\mathbf{E}}_S}{\mathrm{d}z} = \frac{1}{2}\left(- a + \mathrm{i}\frac{\omega_S}{cn_S}\chi^{NL}|\bar{\mathbf{E}}_L|^2 \right)\bar{\mathbf{E}}_S \qquad [6]$$

where we ignore the higher-order differential of amplitude. n_S is the refractive index at ω_S, and a is the linear absorption coefficient of the signal radiation. When ignoring the decay of the excitation radiation along the z direction in the material, for making the right-hand term of eqn [6] positive, the next condition must be satisfied:

$$I_L = \frac{1}{2}c\varepsilon_0 n_L |\bar{\mathbf{E}}_L|^2 \geq I_L^{th} = \frac{c^2\varepsilon_0 n_S n_L}{2\omega_S}\frac{a}{-\mathrm{Im}\chi^{NL}} \qquad [7]$$

The signal radiation is then amplified along the z direction at an amplification constant of g,

$$G = g(I_L - I_L^{th}) \qquad [8]$$

$$g = \frac{\omega_S}{c^2\varepsilon_0 n_S n_L}(-\mathrm{Im}\,\chi^{NL}) \qquad [9]$$

Stimulated Raman Scattering

In SRS, the frequency difference $\omega_L - \omega_S$ corresponds to the resonance frequency of vibrational oscillation ω_V. Therefore, the vibrational oscillation mode is excited in the material as well as the generation of scattered radiation. In the following section, we consider the SRS by optical phonon. In general, nonlinear polarization under the optical excitation is described as follows:

$$\mathbf{P}^{NL} = N_0 \alpha' q \mathbf{E}, \qquad [10]$$

$$\ddot{q} = \Gamma_V \dot{q} + \omega_V^2 q = \frac{1}{2}\alpha' \mathbf{E}^2 \qquad [11]$$

where q is the basic coordinate of molecular vibration, Γ_V is the decay rate of the vibrational oscillation, $\alpha' = (\partial\alpha/\partial q)_0$ is the Raman polarizability of a molecule, and N_0 is the molecule density. We can obtain the steady-state nonlinear polarization from eqns [4], [5], [10], and [11] as

$$\chi^{NL} = (-\omega_S, -\omega_L, \omega_L, \omega_S) = \frac{N_0(\alpha')^2}{4\varepsilon_0}\frac{1}{\omega_V^2 - \omega^2 + \mathrm{i}\omega\Gamma_V} \qquad [12]$$

where $\omega = \omega_L - \omega_S$. Thus, the gain coefficient can be obtained from eqns [9] and [12] under the assumption of $\omega_V \gg \Gamma_V$, as follows:

$$g_R(\omega_S) = \pm\frac{N_0(\alpha')^2\omega_S}{2c^2\varepsilon_0 n_S n_L \omega_V \Gamma_V} \times \frac{(\Gamma_V/2)^2}{(\omega_L - \omega_S \mp \omega_V)^2 + (\Gamma_V/2)^2} \qquad [13]$$

In general, Raman scattering is characterized by the scattering cross-section $\mathrm{d}\sigma/\mathrm{d}\Omega$, which is defined for one of the optical polarization directions in terms of the optical power density:

$$g_R(\omega_L - \omega_V) = \frac{16\pi^2 c^2 N_0}{\hbar n_S^2 \omega_S^3 \Gamma_V}\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}\left\{1 - \exp\left(- \frac{\hbar\omega_V}{k_B T}\right)\right\} \qquad [14]$$

The value of $(\mathrm{d}\sigma/\mathrm{d}\Omega)\Gamma_V^{-1}$ can be characterized as a peak intensity in the measured spontaneous Raman scattering spectrum.

When the conversion to the first Stokes wave is large, we must consider both pump depletion and conversion to higher-order Stokes waves (Figure 1a). Pump depletion in first Stokes generation can be accounted for by solving the following coupling

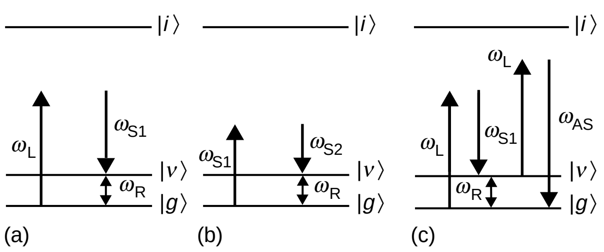

Figure 1 Energy diagrams of stimulated Raman scattering: (a) first Stokes scattering; (b) cascaded second Stokes scattering; and (c) first anti-Stokes scattering.

equations between pump power and Stokes power:

$$\frac{\partial I_L}{\partial z} + \frac{n_L}{c}\frac{\partial I_L}{\partial t} = -g_L I_L I_S \qquad [15]$$

$$\frac{\partial I_S}{\partial z} + \frac{n_S}{c}\frac{\partial I_S}{\partial t} = g_R I_L I_S \qquad [16]$$

where

$$g_L = \frac{\omega_L}{\omega_S} g_R$$

In SRS the power generated in the first Stokes serves as a pump for a second Stokes wave (Figure 1b). Cascade shifts of this type, in which the Stokes wave of order n serves as a pump for the generation of the Stokes wave order $n+1$, can result in the generation of radiation at several frequencies, each one shifted from the original pump frequency by a multiple of the vibrational mode frequency. In addition to the cascade higher-order Stokes generation, multiple Stokes orders can also be generated by four-wave mixing interactions (Figure 2), of the form:

$$\omega_{S,n+1} = 2\omega_{S,n} - \omega_{S,n-1} \qquad [17]$$

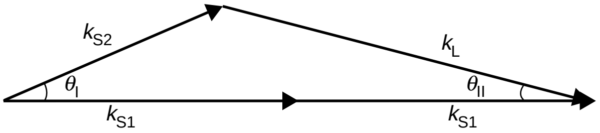

Figure 2 Vector diagram of multiple Stokes generation through four-wave mixing in positively dispersive media.

These interactions are important when the wavevector mismatch among the various waves is small. In most media exhibiting positive dispersion, the phase matching of the various interactions involves off-axis components of the generated waves. Thus, these four-wave mixing interactions are stronger for tightly focused beams than for collimated pump beams.

SRS can be realized in a variety of configurations, including single pass Raman generation, or using a variety of extra-cavity or intra-cavity Raman resonator. Variation of the Raman cross-section for the wavelength is shown in Figure 3 for various gaseous Raman media. The Raman gain coefficient simply increases with decreasing wavelength, as indicted by eqn [13]. Therefore, higher conversion efficiencies are obtainable with UV lasers such as rare-gas halide excimer lasers in a single pass configuration. Note that in some gases, such as H_2, the Raman gain becomes independent of pressure at higher pressures since the increase in linewidth, due to pressure broadening, cancels the increase in gain originated from number density. Typical SRS gains for some gas, liquid, and solid media are listed in Table 1.

Although SRS in high-pressure gases has been widely employed for frequency conversion, SRS in crystals is currently a growing area of research activity. Table 2 lists SRS frequency shifts of various crystals. Since a very high pump laser power ~1 GW/cm^2, which is already close to the laser damage threshold of many crystals, solid-state Raman lasers are

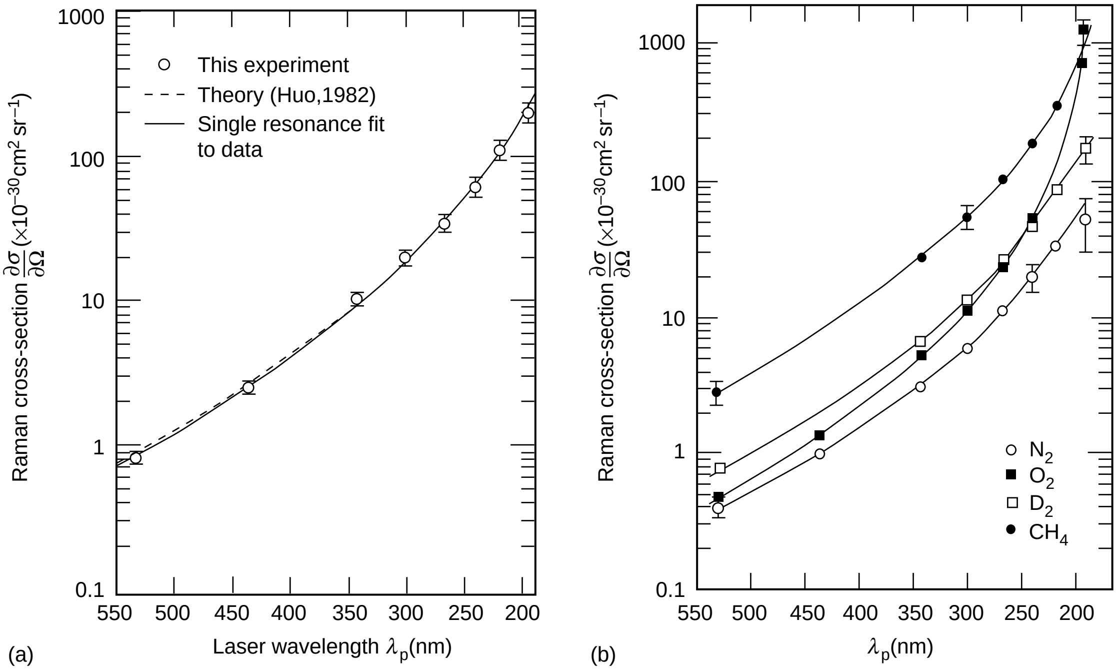

Figure 3 Variation of the Raman cross-section with pump wavelength for various gases. Source: AIP Conference Proceedings No. 100 (Excimer Laser – 1983), 181.

Table 1 Typical Raman media and their parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	$N\frac{d\sigma}{d\Omega}$ *(10^{-6} m^{-1} str^{-1})*	*Raman gain (10^{-5} m/MW)*
H_2 (100 atm)	4155	0.20	3.0×10^3	1.5
H_2O	3420	176	0.51	0.07
Acetone CH_3COCH_3	2921	18	5.4	1.2
Cyclohexane C_6H_{12}	2852	11	3.7	1.2
Liquid N_2	2326.5	0.067	0.29 ± 0.09	17.5 ± 5
Liquid O_2	1552	0.117	0.48 ± 0.14	14.5 ± 4
Nitrobenzene $C_6H_5NO_2$	1345	6.6	6.4	2.1
Toluene C_7H_8	1003	1.94	1.1	1.2
C_6H_5Cl	1002	1.6	1.5	1.9
C_6H5Br	1000	1.9	1.5	1.5
C_6H_6	992	2.15	3.06	2.8
CS_2	655.6	0.5	7.55	23.8
CC_{l4}	459	5	2.3	0.85
Diamond	1332	2.04	17	6.9
Calcite	1086	1.1	2.9	4.4
Si	521	0.8	305	190
Quartz	467	6.7	3.1	0.8
Li^7TaO_3	201	22	238	4.4
	215	12	167	10
Li^7NbO_3	256	23	381	8.9
	258	7	262	28.7
	637	20	231	9.4
	643	16	231	12.6
InSb ($n_e \sim 2 \times 10^{22}$ m^{-3})	100	2	30	16.7

designed with an intra-cavity, coupled cavities, or an extra-cavity. Efficient Raman lasers have been demonstrated using external resonator configuration when the pumping pulse is intense enough. For example, ~40% conversion efficiency to S1 was reported with a Q-switched frequency doubled Nd:YAG laser in a $Ba(NO_3)_2$ crystal using a pump fluence of 1.4 J cm^{-1} and a pulse repetition rate of 30 Hz. The resonator mirror characteristics can be designed to enhance the desired Stokes-order radiation by completely confining the lower-order Stokes radiation inside the cavity. For a lower power pump source, such as cw-pumped Q-switch lasers, a Raman crystal is placed inside a pump laser resonator to reduce the effective threshold for SRS. The resonator configuration has a significant influence on the Stokes output performance. The optimum spot size in the resonator and power density in the Raman crystal is a compromise between maximizing the conversion efficiency and avoiding optical damage. It is noteworthy that the thermal lens in the Raman crystal is caused by the energy dissipation of the phonon and thus depends on the power density of the Stokes field.

A fiber Raman laser is an optical fiber with pump light focused into one end and SRS light appearing at the output. By injecting a weak frequency-shifted signal light together with the pump light, the signal light experienced substantial amplification along the length of the fiber. This fiber Raman amplifier has been actively developed to extend the wavelength band for optical fiber communication. The fiber Raman lasers have utilized fused silica (SiO_2), germanosilicate, or phosphosilicate optical fibers for which the Raman frequency shifts are ~440–490 cm^{-1}. In contrast to Erbium-doped fiber amplifier (EDFA), since the single mode fiber itself can act as an amplifier, the Raman amplifier can be distributed over optical transmission lines in lengths of >20 km, which can keep the transmission power always nearly constant and consequently reduce the signal noise. Moreover, by preparing multipumping wavelengths, the number of signal wavelengths can increase without the limitation of an amplification gain bandwidth. High-power light sources around 1480 nm developed for EDFA can be utilized by the fiber Raman amplifier. By tailoring the power levels among multiwavelength pump lights, a flat amplification gain spectral width of ~100 nm is obtainable.

Transient Effects

In the steady-state SRS analysis, we assume that the pump pulse duration is much longer than the vibrational mode dephasing time ($\Delta\nu_p \ll \Gamma_V$). In a transient regime, when the pump pulse duration is shorter than the dephasing time ($\Delta\nu_p \gg \Gamma_V$), the effect of the transient response of the medium reduces the gain and delays the Stokes pulse relative to the pump pulse. With no pump depletion, the analytical expression for the Stokes intensity for large

Table 2 Typical solid-state Raman materials and parameters

Materials	*Stokes shift (cm^{-1})*	*Raman linewidth (cm^{-1})*	*$\frac{d\sigma}{d\Omega}$ (arb.u.)*
Diamond	1332.9	2.7	100
SO_2	464.5	7.0	2.2
Nitrates and calcite			
$Ba(NO_3)_2$	1048.6	0.4	21
$NaNO_3$	1069.2	1.0	23
$CaCO_3$	1086.4	1.2	6.0
Tungstates			
$CaWO_4$	910.7	4.8	47
$SrWO_4$	921.5	3	–
$BaWO_4$	926.5	2.2	–
$NaY(WO_4)_2$	918	15	–
$KGd(WO_4)_2$	901	5.4	54
$KGd(WO_4)_2$	901	5.4	43
$KGd(WO_4)_2$	768	6.4	19
$KGd(WO_4)_2$	768	6.4	65
$KY(WO_4)_2$	905.6	7	50
$KY(WO_4)_2$	905.6	7	45
$KY(WO_4)_2$	767.4	8.4	20
$KY(WO_4)_2$	767.4	8.4	64
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	908	7.4	48
$KYb(WO_4)_2$	757	15	25
$KYb(WO_4)_2$	757	15	70
Molybdates			
$CaMoO_4$	879.3	5.0	64
$SrMoO_4$	887.7	2.8	55
$BaMoO_4$	892.4	2.1	52
Iodate and niobates			
$LiIO_3$	821.6	5.0	54
$LiNO_3$	872	21.4	44
$LiNbO_3$	632	27	166
$LiNbO_3$	250	28	–
$LaNbO_4$	805	9	22
Phosphates			
$Ca_2(PO_4)_3F$	964.7	2.8	3.4
$Sr_5(PO_4)_3F$	950.3	2.8	3.4

amplification gain is given as

$$I_S(\ell) \propto I_S(0)\exp\left(-\frac{\tau_p}{T_V}\right) \times \exp\left\{2\left[I_L\tau_p\ell\frac{16\pi^2c^2N_0}{\hbar n_S^2\omega_S^3}\left(\frac{d\sigma}{d\Omega}\right)\right]^{1/2}\right\} \qquad [18]$$

Here, T_V is the dephasing time ($T_V = \pi\Gamma_V^{-1}$), and τ_p is the pump pulse width. One can see that the Raman gain is proportional to the square root of the multiplication of the pump pulse energy $I_L \times \tau_p$, the crystal length ℓ, and the total integral Raman scattering cross-section ($d\sigma/d\Omega$), but does not depend on the Raman linewidth Γ_V.

Anti-Stokes Raman Scattering

From eqn [13] the gain coefficient is negative at $\omega = \omega_L + \omega_S$. However, anti-Stokes scattering (Figure 1c) is stimulated through a four-wave mixing interaction of the form:

$$\omega_A = 2\omega_L - \omega_S \qquad [19]$$

The equations describing anti-Stokes Raman scattering in the absence of pump depletion are:

$$\frac{d\bar{E}_S}{dz} + \frac{a}{2}\bar{E}_S = i\frac{\omega_S^2}{2c^2\varepsilon_0 k_{Sz}}\bar{P}_S^{NL} \qquad [20]$$

$$\frac{dE_A^*}{dz} + \frac{a}{2}\bar{E}_A^* = -i\frac{\omega_A^2}{2c^2\varepsilon_0 k_{Az}}\bar{P}_A^{NL*} \qquad [21]$$

$$\bar{P}_S^{NL} = \varepsilon_0\chi^{NL}\left\{|\bar{E}_L|^2\bar{E}_S + \bar{E}_L^2\bar{E}_A^*\exp(i\Delta kz)\right\} \qquad [22]$$

$$\bar{P}_A^{NL*} = \varepsilon_0\chi^{NL}\left\{\bar{E}_L^{*2}\bar{E}_S\exp(-i\Delta kz) + |\bar{E}_L|^2\bar{E}_A^*\right\} \qquad [23]$$

Here, Δk is the wavevector mismatch, given by

$$\Delta k = k_S + k_A - 2k_L \qquad [24]$$

Then, power gain per unit length is obtained to give

$$G \cong -a \pm 2\,\mathrm{Im}\left\{\left(\frac{\Delta k}{2}\right)^2 - \frac{\omega_S}{2cn_S}\chi^{NL}|\bar{E}_L|^2\Delta k\right\}^{1/2} \qquad [25]$$

The interaction of the anti-Stokes wave with the other waves is stronger when the wavevector mismatch is small. The exact value of the phase-matching angle depends on the dispersion of the medium. At the tightly focused pumping condition, strong anti-Stokes radiation is generated so that the wavevector mismatch is minimum. Just as in Stokes scattering, multiple anti-Stokes orders can be generated (Figure 4).

So far, Stokes and anti-Stokes Raman scattering have been used to generate radiation over much of the UV and VUV, ranging from 138 to 400 nm. The shortest wavelength achieved in this manner is at 138 nm, generated by a seventh-order anti-Stokes shift in H_2 using 139 nm radiation of an ArF laser. In general, the pump intensity used in experiments with rare-gas halide excimer lasers was in the range of 100 MW/cm^2 to 5 GW/cm^2. The conversion efficiency and the distribution of power among the various Stokes and anti-Stokes

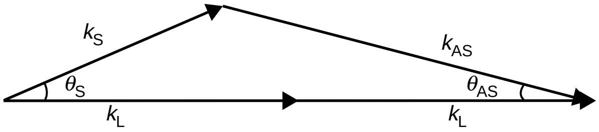

Figure 4 Vector diagram of anti-Stokes Raman scattering in positively dispersive media.

components depend on the focal geometry, and the extent of the spectrum could be varied by changing the gas pressure.

See also

Magneto-Optics: Cyclotron Resonance, Interband Magnetoabsorption, Spin Flip Raman Scattering. **Nonlinear Optics, Applications:** Raman Lasers.

Further Reading

Basiev TT, Sobol AA, Zverey PG, Ivleva LI, Osiko VV and Powell RC (1999) Raman spectroscopy of crystals for stimulated Raman scattering. *Optical Matters* 11: 307.

Bischel WK and Blach G (1983) New lines in the UV: SRS of excimer laser wavelengths. *AIP Conference Proceedings No. 100 (Excimer Lasers-1983)*, 181.

Loree TR, Sze RC, Barker DL and Scott PB (1979) *IEEE Journal of Quantum Electronics* QE-15: 337.

Reinties JF (1985) Coherent ultraviolet and vacuum ultraviolet sources. In: Laser Bass M and Stitch ML (eds) *Laser Handbook*. Amsterdam: North-Holland Physics Publishing.

Weber MJ (ed.) (2001) Raman, Brillouin and Soliton lasers. In: *Handbook of Laser*, p. 291. Boca Raton, FL: CRC Press.

Scattering from Surfaces and Thin Films

A Duparré, Fraunhofer Institut IOF, Jena, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The quality of optical components can be critically limited by light scattering. The majority of these components, such as lenses, mirrors, beamsplitters, and polarizers are covered with thin film optical coatings. To reduce scattering losses, their main sources and mechanisms must be known and, whenever possible, be controlled. Moreover, light scattering techniques have been widely recognized as a powerful tool for roughness and defect analysis as well as for general quality assessment.

Scattering can be divided into surface/interface scattering (SIS) and volume scattering (VS), attributed to the following scatter sources:

- roughness and surface defects of the substrate: SIS,
- surface/interface roughness and defects of the thin film coating: SIS,
- imperfections in the bulk substrate material: VS,
- intrinsic thin film morphology and defect structures: VS,
- surface contaminations: SIS.

Even though in particular cases and applications, VS can constitute the main scattering source, SIS largely dominates both substrate and thin film scattering losses. Studies into light scattering have widely focused on SIS effects from surface and interface micro-roughness. The theoretical outline given in the next section will also follow this concept in describing scattering from surface and interface roughness. It is emphasized, however, that volume scattering theories are accessible in the literature.

Light Scattering Models

Scattering from Rough Surfaces

A randomly rough surface can be considered as a Fourier series of sinusoidal waves with different amplitudes, periods, and phases. Following the grating equation, a single grating with spacing D causes scatter into the angle Θ according to

$$\sin\Theta = \lambda/D \qquad [1]$$

where λ is the wavelength of light. D represents one spatial wavelength in the Fourier series. Accordingly, $f = 1/D$ represents one single spatial frequency. A randomly rough surface contains many different spatial frequencies. This is quantitatively expressed by the power spectral density (PSD), giving the relative strength of each roughness component of a surface microstructure as a function of spatial frequency:

$$\mathrm{PSD}(\mathbf{f}) = \lim_{A\to\infty} \frac{1}{A}\left|\int_{\mathbf{A}} \zeta(\mathbf{r})\exp(-2\pi i\mathbf{f}\cdot\mathbf{r})\mathrm{d}\mathbf{r}\right|^2 \qquad [2]$$

where $\zeta(\mathbf{r})$ represents the height of the surface roughness profile, $\mathbf{r}$ is the position vector, and $\mathbf{f}$ is the spatial frequency vector in the $x-y$ plane. A is the area of the measured region $\mathbf{A}$. We confine our discussion to isotropic surfaces, which represent the majority of cases in optical surface and thin film scattering studies. Thus, a $\mathrm{PSD}(f)$ independent of the surface direction Φ of vector $\mathbf{f}$ is obtained by averaging the two-dimensional function $\mathrm{PSD}(\mathbf{f})$ over all surface directions after transformation into polar coordinates:

$$\mathrm{PSD}(f) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{PSD}(f,\Phi)\mathrm{d}\Phi \qquad [3]$$

Well-established vector scattering theories developed, for instance, by Bousquet *et al.* or Elson provide the link between the PSD and the scattering intensity per solid angle of a surface:

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = F(\lambda, n, \Theta, \Phi)\mathrm{PSD}(f) \qquad [4]$$

This theory is valid for surfaces whose rms roughness is small compared to the wavelength. $\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$ denotes the differential power scattered into the direction (Θ, Φ) per unit solid angle $\mathrm{d}\Omega = \sin\Theta\,\mathrm{d}\Theta\,\mathrm{d}\Phi$ divided by the incident power P_0. Θ and Φ are the polar and azimuthal angles of scattering, respectively. The optical factor F contains all information on the corresponding perfect surface (without the roughness properties), i.e., the refractive index n, wavelength, and the conditions of illumination and observation. Both backscattering and forward scattering can be expressed by eqn [4], according to the illumination and observation conditions chosen. Without loss of generality, all formulas have been written here for normal incidence. The formalism, however, allows consideration of all possible cases, including oblique angles of incidence and arbitrary polarization properties. $\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$ is called angle resolved scattering (ARS) which is related to the well-established term BRDF/BTDF (bidirectional reflectance/transmittance distribution function) by multiplication with $\cos\Theta$:

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = \mathrm{ARS} = \mathrm{BRDF}\ (\text{or BTDF})\cdot\cos\Theta \qquad [5]$$

Total scattering (TS), which is defined as the power P scattered into the backward or forward hemisphere divided by the incident power P_0 (see also section on scattering measurement below), is obtained by integrating eqn [4] over the forward or backward hemisphere:

$$\mathrm{TS} = 2\pi\int_0^{\frac{\pi}{2}}\left(\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega}\right)\sin\Theta\,\mathrm{d}\Theta \qquad [6]$$

If the correlation length of surface roughness is much larger than the wavelength, scalar scattering theories like the one from Carniglia can also be employed. Moreover, in this case, the well-known simple approximate formula for total backscattering can be derived from both vector and scalar theories:

$$\mathrm{TS}_{\mathrm{back}} = R_0\left(\frac{4\pi\sigma}{\lambda}\right)^2 \qquad [7]$$

where σ is the rms roughness and R_0 the specular reflectance.

It must be emphasized that this formula is only valid if the above-mentioned condition is met and as long as only single surfaces without coatings are considered.

Scattering from Dielectric Thin Films

For surfaces coated with a dielectric single layer or multilayer, ARS for a system of N layers is given by

$$\frac{\mathrm{d}P}{P_0\,\mathrm{d}\Omega} = \sum_{i=0}^{N}\sum_{j=0}^{N} F_i F_j^{*}\mathrm{PSD}_{ij}(f) \qquad [8]$$

where F_i is the optical factor at the i-th interface and F_j^* is the conjugate complex number of the optical factor at the j-th interface. The optical factors include both the conditions of illumination and observation and the characteristics of the ideal multilayer (refractive indices, film thickness). PSD_{ij} are the power spectral densities of the corresponding interfaces. For $i \neq j$, they describe the roughness cross-correlation between two interfaces.

From eqn [8] it becomes obvious that the scattering mechanism of coated surfaces is considerably more complex as compared to single surfaces.

The amplitudes of the scattered fields from all individual interfaces add up to the total scatter. In particular, the cross-correlation properties between the interface roughness profiles significantly determine the overall total scatter. So even for coatings having the same statistical roughness, the scatter losses can differ drastically when the cross-correlation between the interfaces is different. An example is demonstrated in a model calculation of total backscattering for $\lambda = 632.8$ nm in Figure 1.

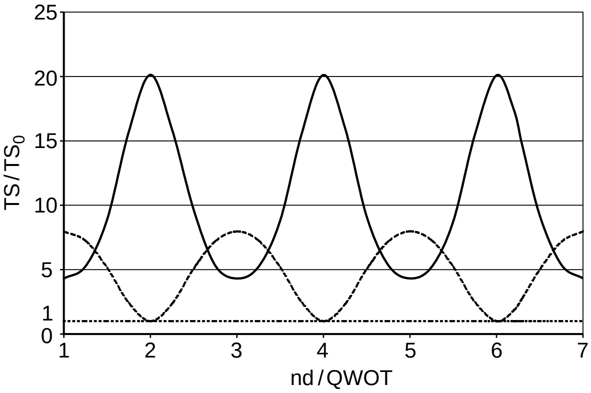

Figure 1 Calculated ratio of the total backscattering of a single high index layer ($n = 2.4$) on BK7 substrate ($n = 1.52$) to the scattering of the bare substrate as a function of optical thickness nd. (QWOT: quarter wave optical thickness). rms roughness was 1 nm, correlation length 1 μm for both the substrate–film and film–air interfaces. Full line: full interface cross-correlation. Dashed line: zero interface cross-correlation. Lower dashed line: bare substrate.

A thin film with refractive index $n = 2.4$ (typical for Titania films), and continuously increasing thickness, was modeled onto a BK 7 substrate ($n = 1.52$). Gaussian autocovariance functions with an rms roughness $\sigma = 1$ nm and a correlation length $\tau = 1$ μm were chosen for both the substrate–film and film–air interfaces. Only the cross-correlation function was varied. The two extreme cases of fully uncorrelated and fully correlated interfaces are depicted in the figure and related to the scattering of the bare substrate (TS_0). As a result of scattering interference effects, in both cases the scattering varies periodically with optical film thickness. The appearance of maxima and minima as well as the total amount of scattering crucially depends on the type of cross-correlation.

These effects have to be taken into careful deliberation whenever scattering in dielectric thin film coatings is considered. Otherwise, misinterpretation of the scatter loss origin is possible and may result in unsuitable technological attempts to minimize scatter losses.

Instrumentation for Light Scattering Measurement

Types of Measurements

Light scattering experiments on optical surfaces and thin films are, in most cases, either performed as ARS measurements using goniophotometers, or the scattering is collected over the backward and forward hemispheres yielding total integrated scattering (TIS) or TS. The latter can be measured either by using a Coblentz sphere or an integrating sphere (see schematic picture in Figure 2).

As Coblentz spheres image the scattered radiation directly onto the detector, they are much less sensitive to scattering by air particles than integrating spheres. So if super-smooth samples with scattering levels in the visible range of 1 ppm and lower shall be measured, integrating spheres require operation in vacuum or He-gas atmosphere whereas Coblentz spheres can still be used under usual clean room conditions.

Angle Resolved Scattering (ARS)

Sophisticated instruments for ARS measurements have been established in a number of laboratories. Most frequently the systems are operated at the He–Ne laser wavelength 632.8 nm, but also at 10.6 μm (CO_2 laser), 325 nm (He–Cd laser), and yet other wavelengths. These techniques are versatile and powerful, but rather unsuitable for routine measurements. For single surfaces, PSD can be calculated from such measurements. A standard procedure for ARS measurements is defined in ASTM standard E 1392 and was successfully verified in various round-robin experiments at different wavelengths. Figure 3 shows the photograph of a typical instrumentation.

The main parts of this setup are the illumination system and a double goniometer. The sample holder and the detector are each mounted on a precision goniometer. The detector head is located on the outer goniometer and can be revolved 360 degrees around the sample with an angular resolution of 0.01°. The performance of a scatterometer depends also considerably on the quality of the illumination system. Laser radiation passes several optical elements, such as a spatial filter, diaphragms, polarizers, and mirrors providing the high beam quality necessary for precision measurement.

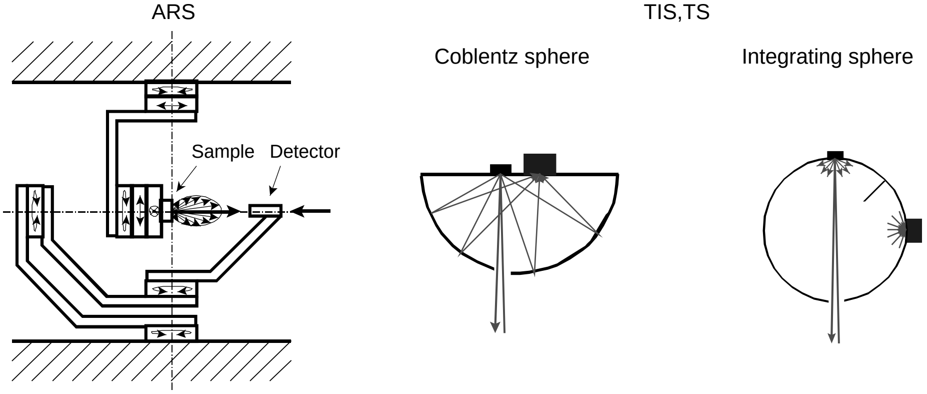

Figure 2 Schematic picture of types of scattering measurement. ARS: angle resolved scattering, TIS, TS: total integrated scattering, total scattering.

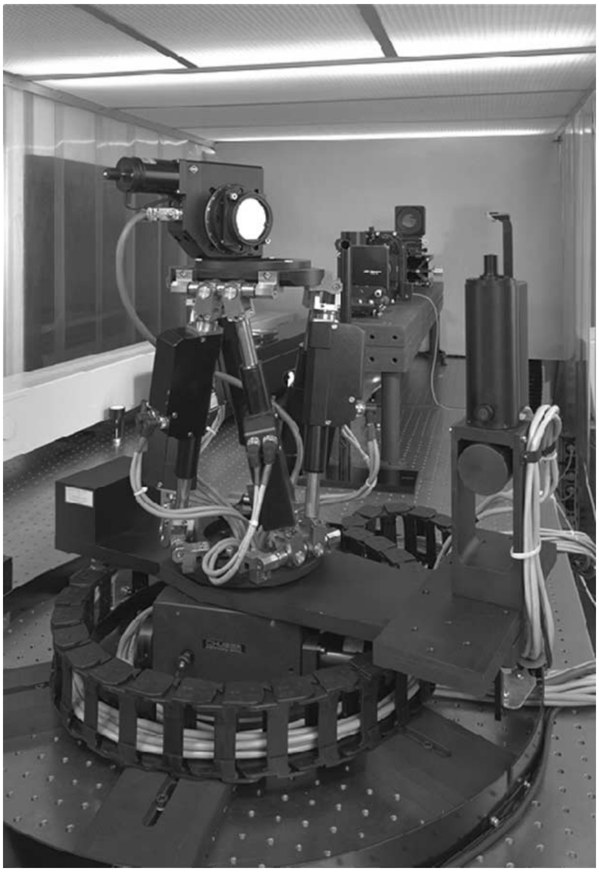

Figure 3 Instrumentation for angle resolved scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

Total Scattering (TS)

TIS measurements collect the light scattered into the backward hemisphere and are defined as the backscattered radiation divided by the total reflectance. The TIS measurement procedure is prescribed in ASTM standard F 1048 which applies to opaque reflective surfaces. Transparent or semi-transparent samples such as substrates, AR coatings, and beamsplitters cannot be measured without ambiguous additional assumption. The new international standard ISO 13696 defines TS as the backscattered and forward scattered radiation P divided by the incident radiation P_0. TS is hence equivalent to the scattering loss of the component. Opaque as well as transparent surfaces and coatings can be measured both in the backward and forward directions. TS and TIS can, however, be converted into one another, if the reflectance of the sample is known. This standard procedure was proved in an international round-robin experiment at 632.8 nm.

An example of a facility for TS measurements, in a wide range of wavelengths, is given in Figure 4. The apparatus for total backscattering and forward scattering measurements is operated from 193 nm to 10.6 μm by using several lasers. The setup is based on a Coblentz sphere imaging the light scattered into the backward or forward hemisphere within an angular range from 2° to 85° onto the detector unit. This angular range complies with the specification given in ISO 13696.

A special arrangement allows easy change from the backscatter to forward scatter operation modus. The detector unit consists of the detector (photomultiplier in the UV, VIS, NIR, and HgCdTe-element in the IR) and a small integrating sphere. The latter is used for homogeneously illuminating the detector. At all wavelengths, the light beam is modulated by a chopper and passes a beam-cleaning element. The incident radiation hits the sample surface at nearly zero degrees and the specular beam is guided back through the entrance/exit aperture of the Coblentz sphere. While performing the measurement, the sample is scanned across its surface by a positioning system, yielding one- or two-dimensional scattering diagrams. Calibration is performed with a commercial diffusing (Lambertian) standard. Filters are used for attenuating the beam during measurement of high-scatter samples and diffuse reflectance standards. Background levels smaller than 0.1 ppm at 632.8 nm were achieved without the necessity of He-gas flow or operation in vacuum.

Whereas in the past the majority of such facilities were designed for the visible, infrared, and near-ultraviolet spectral regions, the increasing demands for low-scatter optical components in the vacuum ultraviolet (VUV) region for application in photolithography has driven the development of scattering facilities for wavelengths as short as 157 nm. Because of the enhanced technical effort needed for the development of VUV scattering measurement instrumentation, only few setups exist that can be operated at 157 nm.

Applications

Fields of Applications

There exists a diversity of purposes for light-scattering investigations of surfaces and thin films, which requires proper selection of the particular method. For a number of purposes, routine measurements without detailed understanding of the scatter sources can be sufficient, while other problems need deeper insight into the scatter mechanisms and hence, require the combination of measurement with theoretical modeling. It is up to the engineer or scientist to decide in each particular case, to which level the

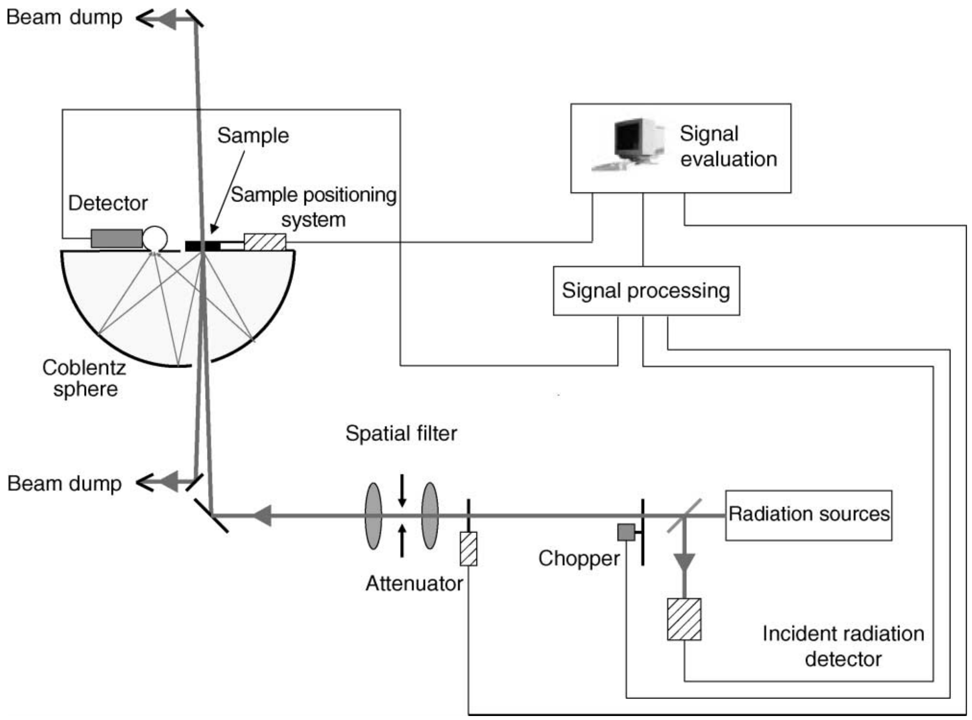
Beam dump
Sample
Sample positioning system
Detector
Coblentz sphere
Signal evaluation
Signal processing
Spatial filter
Beam dump
Radiation sources
Chopper
Attenuator
Incident radiation detector

Figure 4 Schematic picture of an instrument for total scattering measurement (courtesy of Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany).

light-scattering study has to be extended. Examples for typical application tasks arising in industry and institutes are:

- determination of total scattering as an optical loss limiting the performance of a thin film component;
- surface finish assessment (roughness, defects) of optical substrates;
- study into the relation between scattering and the morphology and roughness of optical thin films as a precondition for optimizing deposition processes;
- assessment of cleanliness of coated components;
- detection of degradation effects in coatings and substrate materials;
- separation of interface scattering from the film and substrate;
- distinction of volume scattering from interface scattering in substrates and films.

Scattering measurements are noncontact, do not require sample preparation, enable rapid mapping of large sample areas, and are highly sensitive. This makes them usable for a large variety of applications extending from super-smooth substrates with surface roughness below 0.01 nm, thin film structures in the nanometer range, and small and large defect features up to rough engineering surfaces with micron roughness.

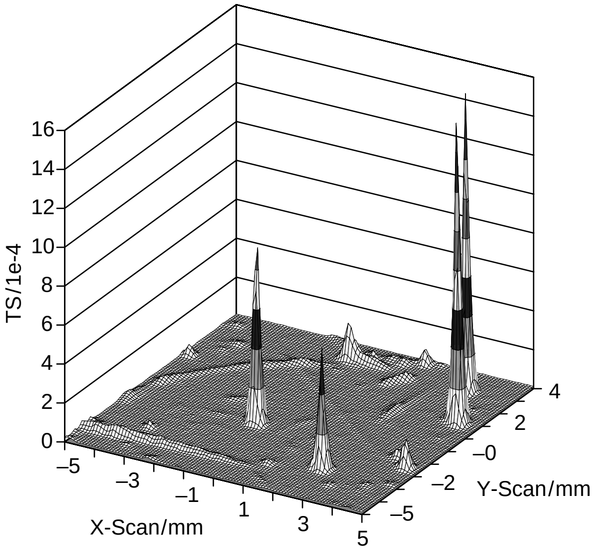

Figure 5 TS measurement (backscattering) on conventionally polished fused silica at 632.8 nm (2D-mapping) for surface quality assessment. Measurement performed with the TS instrument shown in Figure 4.

Examples of Measurements

Surface finish assessment

The result of a scattering measurement (TS, backscattering) at the He–Ne laser wavelength (632.8 nm), on conventionally polished fused silica, is shown in Figure 5. Symmetric defects, as well as scratch-like features, can be recognized in this diagram. The minimum and averaged rms-roughness determined from the minimum (defect-free) and averaged (including defects) scattering levels are 0.51 and 0.72 nm, respectively.

Scattering from thin film components

Figure 6 shows a TS measurement (backscattering) performed on a polished CaF_2 sample. Half of the area of one surface of this sample was coated with a fluoride multilayer mirror system designed for 248 nm. The area mapping of the TS measurement at 248 nm delivers information of both the bare substrate quality and the increase in scattering after coating. According to the explanations in the first section, this increase has to be interpreted as a result of two influences: the increased reflectance (i.e., changed optical factors) and the roughness growth during deposition of the multilayer (i.e., changed PSD).

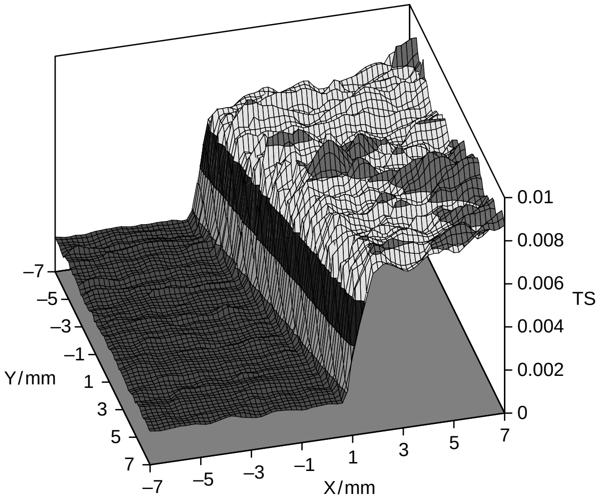

Figure 6 TS measurement (backscattering) at 248 nm on a CaF_2 substrate half of which was coated with a fluoride multilayer mirror system designed for 248 nm. Measurement performed with the TS instrument shown in **Figure 4**.

On the basis of such measurements, detailed interpretation can be accomplished by additional roughness analysis using scanning force microscopy, subsequent scattering modeling employing the formulas given above, and comparison of the modeled and measured results.

ARS techniques can, for example, be employed to compare the angular distribution of the scattering of a coating to that of the bare substrate. Figure 7 displays ARS measurements in the backward hemisphere at 325 nm (He–Cd laser) on a fused silica substrate before and after deposition (magnetron sputtering) of a Ta_2O_5/SiO_2 quarterwave multilayer mirror (325 nm design wavelength). As with the example discussed above, the higher scatter of the multilayer is caused by both the optical factors and increased roughness. The slight bump at large scatter angles in particular reveals high spatial frequency roughness attributed to the film morphology.

Figure 8 displays TS measurements at 157 nm. The diagram contains one-dimensional scans of forward scattering from an uncoated CaF_2 substrate and an antireflective (AR) coating on CaF_2. The background signal level is also included.

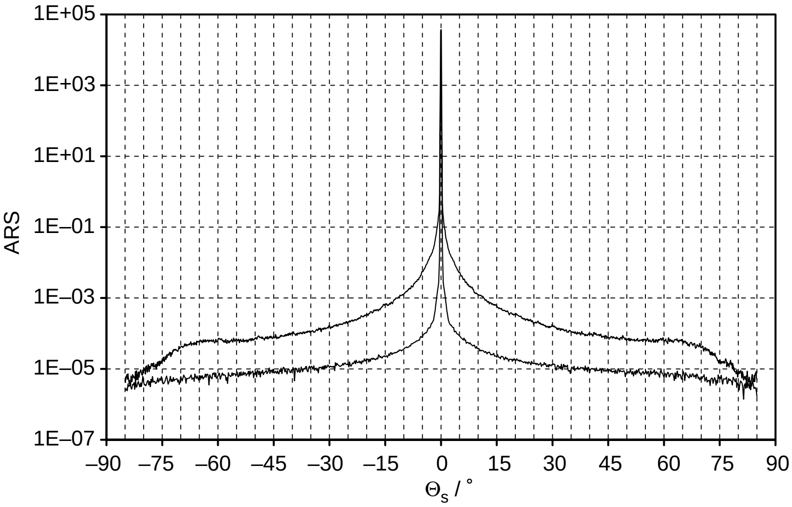

Figure 7 ARS measurement at 325 nm (backward hemisphere) on a fused silica substrate before (lower curve) and after (upper curve) deposition of a Ta_2O_5/SiO_2 quarterwave multilayer mirror with 325 nm design wavelength. Measurements performed with the ARS instrument shown in **Figure 3**.

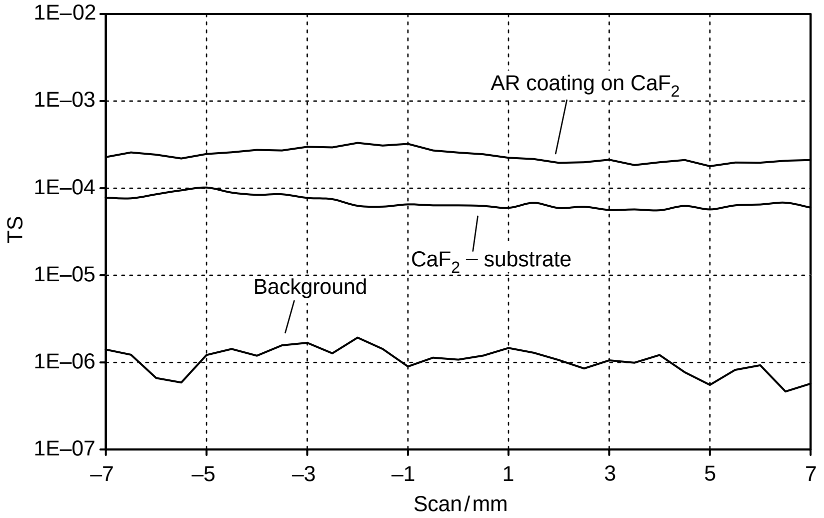

Figure 8 TS measurements at 157 nm (forward scattering): bare CaF$_2$ substrate, AR coating on CaF$_2$ and background scattering level. Measurements performed with the VUV total scattering instrumentation at the Fraunhofer Institute for Applied Optics and Precision Engineering, Jena, Germany.

List of Units and Nomenclature

Angle resolved scattering	[nondimensional]	ARS
Azimuthal scattering angle	[degree]	Φ
Correlation length	[nm]	τ
Differential scattered power	[nondimensional]	$\mathrm{d}P/(P_0\,\mathrm{d}\Omega)$
Grating spacing	[μm]	D
Height of surface roughness profile	[nm]	ζ
Incident power	[W]	P_0
Optical factor	[nm^{-4}]	F
Polar scattering angle	[degree]	Θ
Power spectral density	[nm^4]	PSD
Refractive index	[nondimensional]	n
Rms roughness	[nm]	σ
Spatial frequency	[μm^{-1}]	f
Specular reflectance	[nondimensional]	R_0
Total integrated scattering	[nondimensional]	TIS
Total scattering	[nondimensional]	TS
Wavelength	[nm]	λ

See also

Optical Coatings: Thin-Film Optical Coatings. **Semiconductor Physics:** Light Scattering.

Further Reading

Baeumer S, Duparré A, Herrmann T, *et al.* (1999) SLIOS – a contribution to standard procedures in stray light measurement. *Proceedings SPIE* 3739: 414–421.

Bennett JM and Mattsson L (1999) *Introduction to Surface Roughness and Scattering*, 2nd edn., pp. 28–35, 62–70. Washington, DC: Optical Society of America.

Bousquet P, Flory F and Roche P (1981) Scattering from multilayer thin films: theory and experiment. *Journal of Optical Society of America* 71: 1115–1123.

Carniglia CK (1979) Scalar scattering theory for multilayer optical coatings. *Optical Engineering* 18: 104–115.

Church L, Jenkinson HA and Zavada JM (1979) Relationship between surface scattering and microtopographic features. *Optical Engineering* 18: 125–136.

Duparré A (1995) Light scattering of thin dielectric films. In: Hummel RE and Günther KH (eds) *Thin Films for Optical Coatings*, pp. 273–303. Boca Raton, FL: CRC Press.

Elson JM (1984) Theory of light scattering from a rough surface with an inhomogeneous dielectric permittivity. *Physical Review B* 30: 5460–5480.

Elson JM and Bennett JM (1979) Vector scattering theory. *Optical Engineering* 18: 116–124.

Herman IP (1996) *Optical Diagnostics for Thin Film Processing*, pp. 522–529. London: Academic Press.

Kadkhoda P, Müller A, Ristau D, *et al.* (2000) International round-robin experiment to test the international organization for standardization total-scattering draft standard. *Applied Optics* 39: 3321–3332.

Kassam S, Duparré A, Hehl K, Bussemer P and Neubert J (1992) Light scattering from the volume of optical thin films: theory and experiment. *Applied Optics* 31: 1304–1313.

Leonhard TA and Rudolph P (1995) BRDF round robin test of ASTME 1392. *Proceedings SPIE* 1995: 285–293.

Macleod HA (2001) *Thin-Film Optical Filters*, 3rd edn., pp. 12–52. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Rönnow D and Veszelei E (1994) Design review of an instrument for spectroscopic total integrated light scattering measurements in the visible wavelength region. *Review of Scientific Instruments* 65: 327–334.

Stover JC (1995) *Optical Scattering, Measurement and Analysis*, 2nd edn., pp. 133–176. Bellingham, CA: SPIE Press.

Van de Hulst HC (1957) *Light Scattering by Small Particles*, pp. 28–31, 63–84. London: Chapman and Hall Ltd.

Whitehouse DJ (1994) *Handbook of Surface Metrology*, pp. 49–60, 512–517. Bristol, UK and Philadelphia: Institute of Physics Publishing.

Scattering Phenomena in Optical Fibers

P Tchofo-Dinda and G Millot, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Nonlinear phenomena that result from interaction of intense light beams with dielectric media can be broadly divided into two main categories: namely, parametric effects and scattering phenomena. Parametric effects arise whenever the state of the matter is left unchanged by the interaction, whereas scattering phenomena imply transitions between vibrational or rotational energy levels in the medium. The two dominant scattering phenomena in optical fibers are stimulated Raman scattering (SRS) and stimulated Brillouin scattering (SBS), which are nonlinear processes in which a part of the energy propagating at a given frequency is converted into one or several down-shifted Stokes beams, through interaction with the vibrational modes of the materials. Although SBS and SRS phenomena are beneficial for developing all-optical fiber lasers and amplifiers for long-haul fiber transmissions, they are detrimental to transmission of ultrashort light pulses in silica fibers. In this article we describe the fundamental aspects of SBS and SRS.

General Features

Three well-known kinds of scattering phenomena in optical fibers are Rayleigh, Raman and Brillouin scattering. Although these three phenomena manifest themselves in qualitatively different ways, all of them originate fundamentally from the effects of a light wave on the atomic and molecular charges of the dielectric medium. The Raman and Brillouin scattering phenomena involve the vibrational modes of the material that are associated with the optical and acoustic branches of the dispersion curve of the material, respectively. Such vibrational modes, called phonons, are schematically represented in Figure 1 with the help of typical dispersion curves of a one-dimensional atomic lattice.

In standard silica fibers (SiO_2), the phonons associated with the optical branch correspond to intramolecular vibrations, which give rise to the Raman scattering phenomenon, whereas those associated with the acoustic branch correspond to intermolecular vibrations, which are responsible for Brillouin scattering. In either of these two scattering processes, part of the energy of the incoming light, which propagates at a given frequency ω, is converted into downshifted (Stokes) or upshifted (anti-Stokes) light waves, at frequencies $\omega - \Omega$ and $\omega + \Omega$, respectively. The frequency shift Ω is determined by the vibrational modes involved in the scattering process. Figure 2 represents schematically

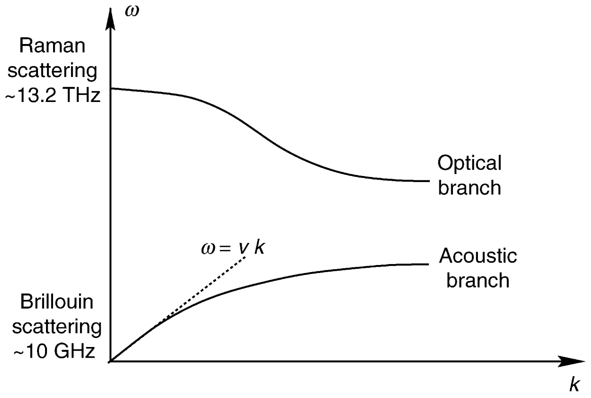

Figure 1 Schematic representation of the dispersion curves for material waves.

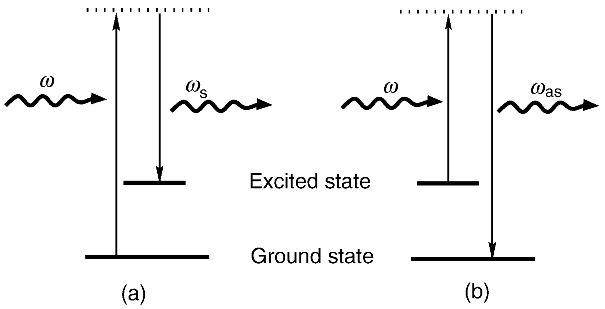

Figure 2 Schematic representation of the diagrams of energy transfer between incoming photons with frequency ω and the material, leading to (a) generation of the Stokes radiation ω_s, and (b) generation of the anti-Stokes radiation ω_{as}.

these energy conversion processes with the help of only two vibrational states, to simplify matters.

As Figure 2a shows, the material, which is initially in the ground state, absorbs a fraction $\hbar\Omega$ of the energy $\hbar\omega$ of incoming photons, to move up to the excited state. As a result, the frequency of the photons involved in this process is downshifted to $\omega_s = \omega - \Omega$. Those photons that are generated at frequency ω_s are commonly called 'Stokes radiation'. On the other hand, when a sufficiently large number of molecules is excited, they can interact with the incoming photons and return back to the ground state. As Figure 2b shows, in this process the energy is transferred from the material to the incoming photons, which thereby generates an upshifted anti-Stokes radiation at frequency $\omega_{as} = \omega + \Omega$. The Raman (Brillouin) frequency shift Ω is of the order of $\Omega_R \approx 13\,\text{THz}$ ($\Omega_B \approx 10\,\text{GHz}$).

In fact, the Raman and Brillouin effects may lead to so-called spontaneous processes, which occur in the limit of low-amplitude for all waves other than the incoming wave. For example, when a laser beam is injected in the fiber with a sufficiently low intensity, one can observe a spontaneous scattering process as schematically represented in Figure 3a.

One can clearly identify in Figure 3a the Raman and Brillouin radiation, as well as radiation scattered at the same frequency as the incoming wave, that is, Rayleigh radiation. Now, when the incoming wave propagates together with light waves whose intensity is no longer negligible with respect to that of the incoming wave, the Raman and Brillouin effects can lead to stimulated processes, in which energy is continually converted from higher to lower frequency photons. As a result, the number of Stokes photons is continually amplified at the expense of anti-Stokes photons, as schematically represented in Figure 3b. The stimulated process can be obtained by either launching the incoming wave ω together with a probe wave at frequency $\omega - \Omega$ (which corresponds to the situation represented in Figure 3b), or by using a sufficiently intense incoming light beam. Indeed, above a certain power threshold (which depends strongly on the fiber length and the amount of dopants in the fiber), for which the gain corresponding to the Stokes amplification compensates for the linear fiber losses, the energy conversion from the incoming wave to the scattered waves can become sufficiently important for the scattering process to become stimulated. This process is commonly referred to as a self-stimulated process, as it does not require the use of a probe wave. Although the SRS and SBS exhibit some common general features as those mentioned above, there exist however some major differences between these two phenomena, from a fundamental point of view as well as for their practical applications to optical communication systems. We discuss separately these specific features for SRS and SBS.

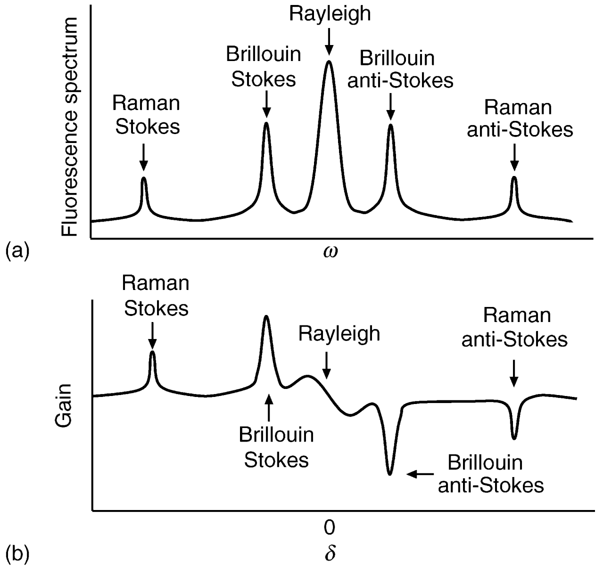

Figure 3 (a) Schematic representation of the frequency spectra for spontaneous and stimulated Raman and Brillouin scattering processes. (a) Fluorescence spectrum; (b) gain spectrum for stimulated Raman and Brillouin scattering processes. A pump and a probe waves are injected together in the fiber. δ is the frequency spacing between the pump and the probe wave.

Stimulated Raman Scattering

Under adiabatic following by the electrons of the optical-field-induced nuclear motions, the third-order nonlinear polarization in optical fibers may be

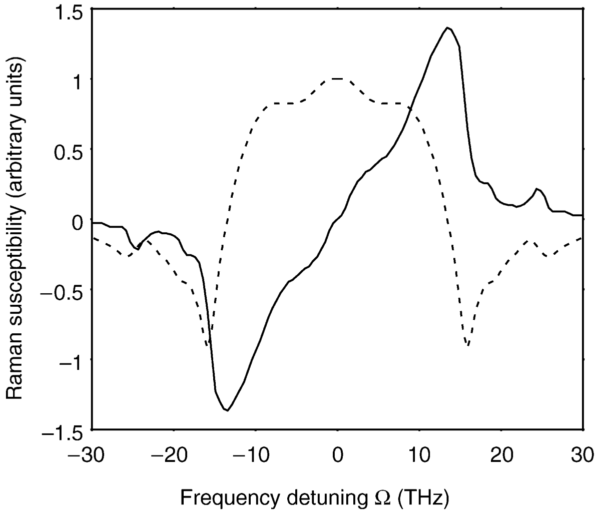

Figure 4 Schematic representation of the Raman susceptibility in an optical fiber. The solid curve represents the imaginary part of the susceptibility, that is, the Raman gain. The dashed curve represents the real part of the susceptibility.

written as

$$P(t) = \sigma|E(t)|^2E(t) + E(t)\int_{-\infty}^{t} ds\sigma_R(t-s)|E(s)|^2 \quad [1]$$

where σ and σ_R represent the instantaneous (Kerr) and noninstantaneous (Raman) parts of the fiber response, respectively. Here, we consider linearly polarized waves, for simplicity. Figure 4 shows schematically the Raman susceptibility of the fused silica, $\chi_R(\Omega) = \sigma_R(\Omega)/2$, which is quite close to that of standard silica fibers.

In Figure 4, the real and imaginary parts of χ_R are represented in solid and dashed curves, respectively. The imaginary part of the Raman susceptibility, which corresponds to the Raman gain $G_R = \mathrm{Im}(\chi_R)$, is an experimentally measurable quantity. For example, at a pump wavelength $\lambda = 1\,\mu\mathrm{m}$, the maximum value of the Raman gain in standard silica fibers is $\approx 10^{-13}\,\mathrm{mW}^{-1}$. In fact the Raman gain $G_R(\Omega)$ depends on the type and amount of dopants used in the fabrication of the fiber. In contrast to crystalline silica where the Raman gain occurs at a well-defined frequency of 13.2 THz, the Raman gain of silica fibers extends over a large frequency range, as can be seen in Figure 4, with a broad prominent peak around 13.2 THz. In the absence of an input signal the broadband noise is amplified within the Raman Stokes components, which are centered about the frequency detuning of 13.2 THz.

If the power of the Stokes wave becomes large enough, the Stokes wave serves as a pump to generate a second-order Stokes wave. This process of cascade SRS can generate multiple Stokes waves whose number depends on the input pump power and fiber length. Figure 5 displays a typical experimental Raman spectrum generated from a Q-switched, frequency-doubled Nd:YAG laser, emitting 5 ns duration (30 GHz spectral width) pulses at 563.63 THz (532.26 nm). The laser output was injected along one birefringence axis of a 14 m length of polarization-maintaining single-mode fiber. At an estimated injected peak power of 200 W, several higher-order Raman Stokes orders S_j ($j = 1, 4$) (separated by 13.2 THz) are generated between 510 and 550 THz (588 and 545 nm). The spectral profile of each resulting Stokes line is near-Lorentzian with a full width at half maximum ranging from about 2 THz for the first Stokes to about 3.5 THz for the fourth Stokes. The broadening of the Raman Stokes wave, as the Stokes order increases, is due to several competing nonlinear processes.

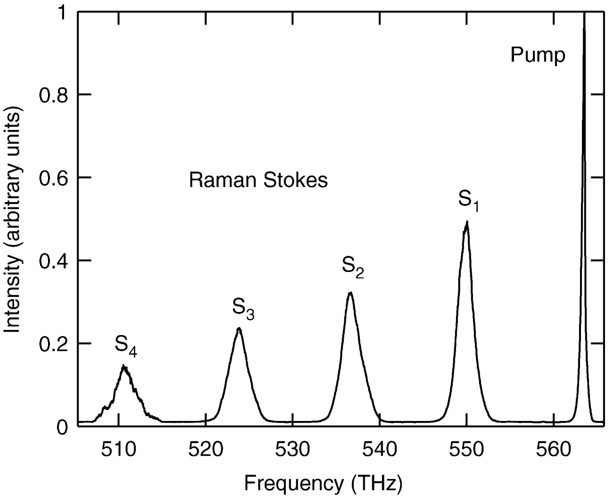

Figure 5 Self-stimulated Raman spectrum (i.e., amplified spontaneous noise), as recorded for a peak pump power $P_0 =$ 150 W and a fiber length $L = 10$ m.

This broadband gain is due to the amorphous nature of fused silica, in which the molecular vibrational frequencies spread out into bands which overlap and create a continuum. This outstanding feature has been fully exploited to develop wide-band amplifiers for multichannel optical communications. Recent work demonstrated the achievement of a Raman amplifier with a bandwidth of 100 nm, by use of a set of height pumps and an appropriate choice of the frequency separations between the pumps. Another important application of the Raman gain in optical fibers lies in the development of fiber-Raman lasers, which can be tuned over a wide frequency range (≈10THz). Such lasers can be obtained by inserting a piece of single-mode fiber inside a Fabry–Perot cavity formed by two partially reflecting mirrors. A rotating output mirror and an intracavity prism allow one to select the Stokes light that provides the laser wavelength. Another example is that the Raman gain can be used to assist non-phase-matched parametric interactions.

A large power-gain enhancement was recently demonstrated for nonphase-matched waves in a three-wave mixing interaction. The Raman-assisted three-wave mixing can be used to achieve wide-bandwidth frequency conversion processes without having to satisfy the strict conditions imposed by phase-matching conditions. The above discussion clearly illustrates some useful effects of the SRS in optical fibers. On the other hand, the SRS may represent a drawback for some optical communication systems, as we discuss below.

Let us consider a pump wave injected in the fiber at frequency ω_0. The total electric field in the fiber may be written as follows:

$$E = \frac{1}{2}\mathcal{A}(z,t)\exp[\mathrm{i}(k_0 z - \omega_0 t)] \qquad [2]$$

where $\mathcal{A}$ is a slowly varying field envelope. By expanding the wavevector k_0 of the electric field in a Taylor series around ω_0, the amplitude $\mathcal{A}$ of the electric field is found to satisfy the generalized nonlinear Schrödinger equation (NLSE):

$$\begin{aligned}\frac{\partial\mathcal{A}}{\partial z}+\frac{\mathrm{i}}{2}\beta\frac{\partial^2\mathcal{A}}{\partial t^2} &= \mathrm{i}\gamma(1-\rho)|\mathcal{A}|^2\mathcal{A}+\mathrm{i}\gamma\rho\mathcal{A}\int_0^\infty \chi_R(s)|\mathcal{A}|^2(t-s)\mathrm{d}s \\ &= \mathrm{i}\gamma|\mathcal{A}|^2\mathcal{A}+\mathrm{i}\gamma\rho C_1\mathcal{A}\frac{\partial|\mathcal{A}|^2}{\partial t}+\frac{\mathrm{i}}{2}\gamma\rho C_2\mathcal{A}\frac{\partial^2|\mathcal{A}|^2}{\partial t^2} \\ &\quad +\frac{\mathrm{i}}{6}\gamma\rho C_3\mathcal{A}\frac{\partial^3|\mathcal{A}|^2}{\partial t^3}+\cdots \qquad [3]\end{aligned}$$

In eqn [3], $C_1 \approx -7.9\times10^{-3}$ps, $C_2 \approx -1.28\times10^{-4}$ps^2, $C_3 \approx 9\times10^{-6}$ps^3, and the parameter ρ, which measures the fractional contribution of Raman effects to the total nonlinearity, is ≈ 0.18 for standard silica fibers.

When the wave propagation involves only a relatively small frequency range around the pump frequency (less that 1 THz) then only the term proportional to C_1 in eqn [3], i.e., the leading term of the Raman contribution, plays a significant role. Most high-capacity transmission systems using a single channel fall into this situation, as long as picosecond pulses are used. The term proportional to C_1 induces a continuous frequency shift of the pulse frequency, as a result of a small transfer of energy from higher- to lower-frequency waves. This phenomenon, known as the soliton self-frequency shift, constitutes one of the undesirable effects of SRS in optical communications.

On the other hand, when the wave propagation involves a relatively large frequency region (of the order of a few THz), then higher-order terms that are proportional to C_2 and C_3 begin to play a significant role. In particular, the term proportional to C_3 indicates a strong transfer of energy from higher- to lower-frequency waves. In optical communications, this effect causes higher-wavelength channels of wavelength-division multiplexing systems to be amplified while depleting the lower-wavelength channels. This effect can be reduced by using filters and more amplification to compensate for the depletion at lower wavelengths (but this is achieved at the expense of a degradation of the signal-to-noise ratio).

Stimulated Brillouin Scattering

In optical fibers, the SBS occurs at input power levels that are much lower that those needed for SRS, and manifests itself through the generation of a backward-propagating Stokes wave. Although SBS is not (strictly speaking) a parametric process, the generation of the Brillouin Stokes radiation is commonly described as a parametric interaction between the pump wave, the Stokes wave and an acoustic wave (material wave). Indeed, one assumes that the annihilation of a pump photon creates simultaneously a Stokes photon and an acoustic phonon, with the following conservation laws for the energy and the momentum:

$$\Omega_\mathrm{B} = \omega_0 - \omega_\mathrm{s} \qquad [4]$$

$$K_\mathrm{B} = k_0 - k_\mathrm{s} \qquad [5]$$

where ω_0 and ω_s are the frequencies, and k_0 and k_s are the wavevectors of the incoming and Stokes waves, respectively. Then, assuming $|k_\mathrm{s}| \approx |k_0|$, the acoustic wave is found to satisfy the following dispersion relation

$$\Omega_\mathrm{B} = K_\mathrm{B}v = 2v|k_0|\sin(\theta/2) \qquad [6]$$

where v is the phase velocity of the acoustic wave, and θ is the angle between the pump and Stokes wavevectors. As eqn [6] shows, the frequency shift

Ω_B depends on the scattering angle. Ω_B is a maximum for $\theta = \pi$ (i.e., in the backward direction) and vanishes for $\theta = 0$ (forward direction). Thus, although eqn [6] predicts that Brillouin scattering should not occur in the forward direction, spontaneous Brillouin scattering can occur in the forward direction in optical fibers. But this phenomenon is quite negligible. In optical fibers, SBS occurs mainly in the backward direction with a frequency shift given by

$$\nu_B = \Omega_B/(2\pi) = 2v|k_0| = 2n_0 v/\lambda_0 \qquad [7]$$

where n_0 is the refractive index and λ_0 is the pump wavelength (incoming wave). In silica fibers, $v \approx 6\,\text{km s}^{-1}$ and $n_0 \approx 1.45$; which leads to $\nu_B \approx 10\,\text{GHz}$ at $\lambda_0 = 1.55\,\mu\text{m}$. Thus, the Brillouin frequency shift is smaller by three orders of magnitude compared with the Raman frequency shift (≈13 THz). The amorphous nature of fused silica leads to a Brillouin-gain coefficient $G_B(\nu)$ which extends over a frequency range of ≈ 10 MHz, with a peak value at $\nu = \nu_B$. The Brillouin gain can be approximated by a Lorentzian profile given by

$$G_B(\nu) = \frac{(\Delta\nu_B/2)^2}{(\nu - \nu_B)^2 + (\Delta\nu_B/2)^2} G_B(\nu_B),$$
$$G_B(\nu_B) = \frac{2\pi n_0^7 p_{12}^2}{c d_0 \lambda_0^2 v \Delta\nu_B} \qquad [8]$$

where $\Delta\nu_B$ is the full width at half maximum, p_{12} is the elasto-optic coefficient and d_0 is the material density.

The Brillouin gain has been exploited to develop fiber-Brillouin lasers. Such lasers are obtained by inserting a piece of fiber inside a (ring or Fabry–Perot) cavity. Another useful application of the Brillouin gain of an optical fiber lies in the amplification of a weak signal. Fiber-Brillouin amplifiers allow one to amplify signals with a frequency shift that corresponds to the Brillouin frequency shift ν_B. Fiber-Brillouin amplifiers can provide up to 40 dB gain at pump powers of a few milliwatts. But the bandwidth of such amplifiers is relatively narrow.

Conclusion

In fact scattering phenomena can be beneficial as well as detrimental for optical communication systems. If signals are transmitted in a communication channel with a power above the threshold of a given scattering phenomenon, this phenomenon will induce a transfer of energy from the signal to the Stokes radiation. The resulting signal depletion would then require more amplification. In multichannel communications, the SBS (SRS) can induce cross-talk between two bidirectional channels if their frequency separation happens to be close to the Brillouin (Raman) shift, ≈10 GHz (≈13 THz). However, those undesirable effects can be avoided by in-line control techniques for pulse transmission in fiber-optics communication systems. On the other hand, the Raman and Brillouin effects are useful for generating light at new frequencies as well as for amplifying weak light signals. In particular Raman amplifiers are one of the key devices for high-speed telecommunication systems.

List of Units and Nomenclature

Group velocity dispersion coefficient β_2 [$\text{ps}^2\,\text{km}^{-1} = 10^{-27}\text{s}^2\,\text{m}^{-1}$]
Nonlinear parameter γ [$\text{W}^{-1}\,\text{m}^{-1}$]
NLS (E): nonlinear Schrödinger (equation)
Optical intensity I [$\text{GW cm}^{-2} = 10^9\,\text{W cm}^{-2}$]
SBS: stimulated Brillouin scattering
SRS: stimulated Raman scattering
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Light Propagation; Nonlinear Effects (Basics). **Optical Amplifiers:** Raman, Brillouin and Parametric Amplifiers. **Solitons:** Soliton Communication Systems.

Further Reading

Agrawal GP (2001) Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Agrawal GP (2001) Applications of Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Bloembergen N (1977) Nonlinear Optics. Reading, MA: Benjamin.
Dianov EM, Mamyshev PV, Prokhorov AM and Serkin VN (1989) Nonlinear Effects in Optical Fibres. Switzerland: Harwood Academic Publishers.
Hellwarth R (1977) Third-order optical susceptibilities of liquids and solids. *Progress in Quantum Electronics* 5: 1–68.
Pocholle JP, Papuchon M, Raffy J and Desurvire E (1990) Non linearities and optical amplification in single mode fibres. *Revue Technique Thomson-CSF* 22: 187–268.
Shen YR (1984) Principles of Nonlinear Optics. New York: Wiley.

Scattering Theory

Y A Eremin, Moscow State University, Moscow, Russia

© 2005 Published by Elsevier Ltd.

Lord Rayleigh (John William Strutt) (1842–1916)

In the nineteenth century, Lord Rayleigh offered the first explanation for the sky's blue color (on a clear day!). From Rayleigh's initial theory, representing a simplification of Maxwell's electromagnetic theory for the case of particles very small compared to the wavelength of the incident light, the theory was extended to describe the scattering of light by larger macromolecules in solution.

Scattering theory has played a central role in twentieth-century mathematical physics. From Rayleigh's explanation of why the sky is blue to modern medical application of computerized tomography, has attracted both scientists and mathematicians for over a hundred years. Scattering theory is concerned with the effect an inhomogeneous medium has on an incident wave. In particular, if the total field is viewed as sum of an incident field and a scattered field, then the direct scattering problem is to determine the scattered field from a knowledge of the incident one, obstacle properties, and the equations governing the wave motion. In contrast, the inverse scattering problem consists of reconstruction of the inhomogeneity from the knowledge of asymptotic behavior of the scattered field. An everyday example is human vision: from the measurements of scattered light that reaches our retinas, our brain reconstructs a detailed three-dimensional map of the world around us. In summary, the task of direct scattering theory is to determine the relation between the incident and scattered waves. The task of inverse scattering theory is to determine properties of the obstacle.

Light propagating can be described in terms of an electromagnetic field, which is composed of two related vector fields, the electric and the magnetic fields. This means that the vectors of electric (**E**) and magnetic (**H**) field have values defined at each point in space and time. The vector fields **E** and **H** interact through the material parameters: electric permittivity ε and magnetic permeability μ. The behavior of electromagnetic fields can be described by Maxwell's equations.

Maxwell Equations. James Clerk Maxwell (1831–1879)

Consider the electromagnetic wave (light) propagation in a homogeneous, isotropic medium with electric permittivity $\varepsilon > 0$, electric conductivity $\sigma \geq 0$, and magnetic permeability $\mu > 0$, containing a local obstacle. We assume that $\sigma = 0$ outside an obstacle. If J is a current density, the electric field $\vec{E}(x,t)$ and magnetic field $\vec{H}(x,t)$ satisfy the Maxwell equations:

$$\begin{aligned} &\nabla \times \vec{E}(x,t) + \mu \frac{\partial \vec{H}}{\partial t}(x,t) = 0 \\ &\nabla \times \vec{H}(x,t) - \varepsilon \frac{\partial \vec{E}}{\partial t}(x,t) = \vec{J}(x,t) \end{aligned} \qquad [1]$$

Also, in an isotropic conductor, the electric field satisfies Ohm's law:

$$\sigma \vec{E}(x,t) = \vec{J}(x,t) \qquad [2]$$

The speed of wave propagation for the electric and magnetic fields is $c = 1/\sqrt{\varepsilon\mu}$ and in free space $c = 1/\sqrt{\varepsilon_0\mu_0} = 2.99792458 \times 10^8 \text{ m s}^{-1}$. Assume that the electromagnetic field is time-harmonic, i.e., of the form:

$$\begin{aligned} \vec{E}(x,t) &= \frac{1}{\sqrt{\varepsilon_0}} \mathbf{E}(x) \mathrm{e}^{-\mathrm{j}\omega t} \\ \vec{H}(x,t) &= \frac{1}{\sqrt{\mu_0}} \mathbf{H}(x) \mathrm{e}^{-\mathrm{j}\omega t} \end{aligned} \qquad [3]$$

where $\omega > 0$ is the frequency. Many phenomenon of light scattering can be described based on monochromatic wave propagation. Let us consider a local obstacle occupying a bounded domain D. Inside the obstacle we introduce relative complex permittivity:

$$\varepsilon = \frac{1}{\varepsilon_0}\left(\varepsilon + \mathrm{j}\frac{\sigma}{\omega}\right) \qquad [4]$$

Then $\mathbf{E}(x)$ and $\mathbf{H}(x)$ obey the time-harmonic Maxwell equations:

$$\nabla \times \mathbf{E}(x) = \mathrm{j}k\mathbf{H}(x), \quad \nabla \times \mathbf{H}(x) = -\mathrm{j}k\varepsilon\mathbf{E}(x) \qquad [5]$$

where $k^2 = \varepsilon_0\mu_0\omega^2$.

Helmholtz Equation. Herman Ludwig von Helmholtz (1821–1894)

A wave equation for vector **E** can be readily derived from the Maxwell equations. We obtain:

$$(\nabla^2 + k^2)\mathbf{E}(x) = 0 \qquad [6]$$

This is known as the homogeneous Helmholtz equation. A general planewave solution of the equation can be written as

$$\mathbf{E}(x) = \mathbf{E}_0 \, \mathrm{e}^{\mathrm{j}\vec{k}\cdot\vec{x}} \qquad [7]$$

here $\mathbf{E}_0$ is a constant vector independent of spatial coordinates and $\vec{k}\cdot\vec{x}$ is a scalar product of the vectors $\vec{x} = \{x_1, x_2, x_3\}$ and $\vec{k} = \{k_1, k_2, k_3\}$, the so called wavevector. Equations [6] and [7] yield the dispersion relation:

$$k^2 = \varepsilon_0 \mu_0 \omega^2 = k_1^2 + k_2^2 + k_3^2 \qquad [8]$$

For a given $\vec{k}$ vector, a constant phase front is determined by $\vec{k}\cdot\vec{x} = \text{const}$, which indicates that the front is perpendicular to the wavevector. So, the phase front is a plane and the amplitude of the electric field on the plane is a constant.

The Maxwell equations for the plane wave solution become:

$$\begin{aligned} &\vec{k} \times \mathbf{E}(x) = k\mathbf{H}(x), \quad \vec{k} \times \mathbf{H}(x) = -k\mathbf{E}(x) \\ &\Rightarrow \vec{k}\cdot\mathbf{E} = 0, \quad \vec{k}\cdot\mathbf{H} = 0 \end{aligned} \qquad [9]$$

Therefore, vectors $\mathbf{E}$ and $\mathbf{H}$ belong to the constant phase plane perpendicular to vector $\vec{k}$ and are orthogonal each other.

Poynting's Vector. John Henry Poynting (1852–1914)

For time-dependent fields, the Poynting vector is introduced as

$$\vec{S}(x,t) = \vec{E}(x,t) \times \vec{H}(x,t) \qquad [10]$$

It is interpreted as the power flow density with the dimension of watts/m^2. The time-average Poynting's vector power density is given by

$$\langle S \rangle = \frac{1}{T} \int_0^T \vec{S}(x,t)\mathrm{d}t \qquad [11]$$

For the time-dependent planewave fields:

$$\begin{aligned} \vec{E}(x,t) &= \mathbf{E}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \\ \vec{H}(x,t) &= \mathbf{H}_0 \cos(\vec{k}\cdot\vec{x} - \omega t) \end{aligned} \qquad [12]$$

The vector Poynting's power density shown in eqn [10], accepts the form:

$$\vec{S}(x,t) = |\mathbf{E}_0|^2 \cos^2(\vec{k}\cdot\vec{x} - \omega t) \qquad [13]$$

The time-average vector power density gives

$$\langle \vec{S}(x,t) \rangle = \frac{1}{2} |\mathbf{E}_0|^2 \qquad [14]$$

which shows that Poynting's power density is also propagating in the direction of the wave vector $\vec{k}$.

Boundary Conditions

Let us consider a bound obstacle D_i with a smooth closed surface S. The Maxwell eqn [5] has been written in differential form. It must be supplemented with boundary conditions enforced at the obstacle surface. The boundary conditions can be derived from the integral form of Maxwell equations. The field vectors $\mathbf{E}$, $\mathbf{H}$ are assumed to be finite but may be discontinuous across the obstacle surface S. In cases where there is no surface current density (which is so for a perfect conductor) the boundary conditions accept the following form:

$$\vec{n} \times (\mathbf{E}_\mathrm{i} - \mathbf{E}_\mathrm{e}) = 0, \quad \vec{n} \times (\mathbf{H}_\mathrm{i} - \mathbf{H}_\mathrm{e}) = 0 \qquad [15]$$

where subscripts i and e correspond to the total fields inside and outside the obstacle, respectively, and $\vec{n}$ is unit outward normal to S. Essentially the boundary conditions state that tangential components for electromagnetic fields are continuous across the boundary.

Silver–Müller Radiation Conditions

Analysis of a direct scattering problem on a local obstacle involves Maxwell eqn [10] and boundary conditions as in eqn [15] and some infinity conditions, which provide the uniqueness of the scattering problem solutions. Let us assume that a local obstacle is excited by an incident electromagnetic field $\{\mathbf{E}_0, \mathbf{H}_0\}$, given as entire solution to the Maxwell equations. Total field outside the obstacle can be represented as

$$\mathbf{E}_\mathrm{e} = \mathbf{E}_0 + \mathbf{E}_\mathrm{s}, \quad \mathbf{H}_\mathrm{e} = \mathbf{H}_0 + \mathbf{H}_\mathrm{s} \qquad [16]$$

where scattered field $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ in the exterior domain $D_\mathrm{e} = \mathbb{R}^3/\bar{D}_\mathrm{i}$ satisfies the Silver–Müller radiation conditions:

$$\left.\begin{aligned} \frac{x}{|x|} \times \mathbf{H}_\mathrm{s} + \mathbf{E}_\mathrm{s} &= \mathrm{o}\left(\frac{1}{|x|}\right) \\ \frac{x}{|x|} \times \mathbf{E}_\mathrm{s} - \mathbf{H}_\mathrm{s} &= \mathrm{o}\left(\frac{1}{|x|}\right) \end{aligned}\right. \quad |x| \to \infty \qquad [17]$$

uniformly for all directions $x/|x|$.

Scattering Amplitude

The following result is valid: every radiating solution $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ to the Maxwell equations has the following asymptotic form:

$$\mathbf{E}_\mathrm{s}(x) = \frac{\mathrm{e}^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\varpi) + \mathrm{O}\left(\frac{1}{|x|^2}\right)\right\} \qquad |x| \rightarrow \infty \qquad [18]$$

$$\mathbf{H}_\mathrm{s}(x) = \frac{\mathrm{e}^{ik|x|}}{|x|}\left\{\mathbf{H}_\infty(\varpi) + \mathrm{O}\left(\frac{1}{|x|^2}\right)\right\}$$

where the vector functions $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ defined on the unit sphere $\varpi \in \Omega = \{0 \leq \vartheta \leq \pi; 0 \leq \varphi \leq 2\pi\}$ are known as the electric far field pattern (scattering amplitude) and magnetic far field pattern, respectively. They satisfy:

$$\mathbf{H}_\infty = \vec{\nu} \times \mathbf{E}_\infty, \quad \vec{\nu} \cdot \mathbf{E}_\infty = \vec{\nu} \cdot \mathbf{H}_\infty = 0 \qquad [19]$$

with the unit outward normal $\vec{\nu}$ on Ω. The latter means that $\mathbf{E}_\infty$ and $\mathbf{H}_\infty$ have only tangential components at the unit sphere.

Far field pattern represents real value analytic functions on a unit sphere. Reillich's lemma establishes one-to-one correspondence between radiating electromagnetic field and their far field pattern. It means that there exists only one scattered field $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ corresponding to a given far field pattern $\mathbf{E}_\infty$.

Reciprocity Relations

Examine the set of far field pattern corresponding to the scattering problem including: Maxwell eqn [5], boundary conditions in eqn [15], and radiation conditions in eqn [17] at the infinity for the scattered fields. Let the incident electromagnetic field is given by the planewave:

$$\mathbf{E}_0(x) = \vec{e}_0\, \mathrm{e}^{\mathrm{j}k\vec{x}\cdot\vec{\kappa}}, \quad \mathbf{H}_0(x) = \vec{\kappa} \times \vec{e}_0\, \mathrm{e}^{\mathrm{j}k\vec{x}\cdot\vec{\kappa}} \qquad [20]$$

where $\vec{\kappa} \in \Omega$ is a unit vector giving the propagation direction and $\vec{e}_0 \in \mathbb{R}^3$ is a constant vector responsible for the polarization. From the previous consideration one can see that:

$$\mathbf{E}_\mathrm{s}(x; \vec{\kappa}, \vec{e}_0) = \frac{\mathrm{e}^{ik|x|}}{|x|}\left\{\mathbf{E}_\infty(\varpi\vec{\kappa}, \vec{e}_0) + \mathrm{O}\left(\frac{1}{|x|^2}\right)\right\}, \quad |x| \rightarrow \infty \qquad [21]$$

Furthermore, the following reciprocity relation for the scattering amplitude holds:

$$\vec{h}_0 \cdot \mathbf{E}_\infty(\varpi; \vec{\kappa}, \vec{e}_0) = \vec{e}_0 \cdot \mathbf{E}_\infty(-\vec{\kappa}; -\varpi, \vec{h}_0)$$

for all vectors $\varpi, \vec{\kappa} \in \Omega$ and $\forall \vec{e}_0, \vec{h}_0 \in \mathbb{R}^3$.

Optical Theorem

Far field pattern enables us to evaluate scattering cross section, which is defined as

$$\sigma_\mathrm{s} = \int_\Omega |\mathbf{E}_\infty(\varpi)|^2 \mathrm{d}\varpi \qquad [22]$$

For the planewave given by eqn [20] and excitation of nonabsorbing obstacle ($\sigma = 0$), the following relation holds:

$$\sigma_\mathrm{s} = \frac{4\pi}{k} \mathrm{Im}\{\mathbf{E}_\infty(\vec{\kappa}) \cdot \vec{e}_0\} \qquad [23]$$

This is so-called optical theorem.

Huygen's Principle (Green's Theorem). Christian Huygens (1629–1695)

The Huygens' principle shows that a wavefield on the surface determines the wavefield off the surface. Let us introduce the scalar Green's function which satisfies the Helmholtz equation:

$$(\nabla^2 + k^2) g(x, y) = -\delta(x - y) \qquad [24]$$

where $\delta(x - y)$ is the Dirac delta function. For three-dimensional problems, the scalar Green's function $g(x, y)$ for isotropic media accepts the form:

$$g(x, y) = \frac{\mathrm{e}^{\mathrm{j}k|x-y|}}{4\pi|x - y|} \qquad [25]$$

Let electromagnetic fields $\{\mathbf{E}_\mathrm{s}, \mathbf{H}_\mathrm{s}\}$ be a radiating solution to the Maxwell eqn [5] in D_e. Then for any $x \in D_\mathrm{e}$ we have the Stratton–Chu formulas:

$$\begin{aligned} \mathbf{E}_\mathrm{s}(x) &= \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{E}_\mathrm{s}(y) \mathrm{d}s_y \\ &\quad + \frac{\mathrm{j}}{k} \nabla \times \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{H}_\mathrm{s}(y) \mathrm{d}s_y \\ \mathbf{H}_\mathrm{s}(x) &= \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{H}_\mathrm{s}(y) \mathrm{d}s_y \\ &\quad - \frac{\mathrm{j}}{k} \nabla \times \nabla \times \int_S g(x, y) \vec{\nu}_y \times \mathbf{E}_\mathrm{s}(y) \mathrm{d}s_y \end{aligned} \qquad [26]$$

here S is a smooth closed surface, and $\vec{\nu}$ is the unit normal vector to S directed into D_e. Similar relations are valid for an electromagnetic field inside a bounded domain D_i. The Stratton–Chu formulas are extensively employed to construct boundary integral equations for scattering problems investigating.

Homogeneous Obstacle

Consider a mathematical statement of the scattering problem by a bound homogeneous obstacle. Let external excitation be a planewave $\{\mathbf{E}_0, \mathbf{H}_0\}$: (eqn [20]). In this case the boundary-value scattering problem can be formulated in the form:

$$\begin{gathered}\nabla \times \mathbf{E}_\tau(x) = \mathrm{j}k\mathbf{H}_\tau(x), \quad \nabla \times \mathbf{H}_\tau(x) = -\mathrm{j}k\varepsilon_\tau \mathbf{E}_\tau(x), \\ x \in D_\tau, \tau = \mathrm{i}, \mathrm{e} \\ \vec{\nu} \times (\mathbf{E}_\mathrm{i} + \mathbf{E}_\mathrm{e}) = \vec{\nu} \times \mathbf{E}_0, \\ \vec{\nu} \times (\mathbf{H}_\mathrm{i} + \mathbf{H}_\mathrm{e}) = \vec{\nu} \times \mathbf{H}_0 \ \text{ on } S \\ \frac{x}{|x|} \times \mathbf{H}_\mathrm{e} + \mathbf{E}_\mathrm{e} = o\left(\frac{1}{|x|}\right), \quad |x| \to \infty \end{gathered} \qquad [27]$$

where $\varepsilon_\mathrm{e} = 1$, $\varepsilon_\mathrm{i} = \varepsilon$ and $\{\mathbf{E}_\mathrm{e}, \mathbf{H}_\mathrm{e}\}$ corresponds to scattered field. For a smooth surface and the case $\mathrm{Im}\, \varepsilon \geq 0$, the problem has a unique solution.

Lippman–Swinger Integral Equation

Assume we deal with the local inhomogeneous obstacle that can be described by dielectric permittivity $\varepsilon(x)$, which is a continuous complex-valued $(\mathrm{Im}\, \varepsilon(x) \geq 0)$ function in D_i. The correspondent scattering problem can be formulated as below in eqn [29]. In this case, the boundary-value scattering problem can be reduced to the solution of the Lippman–Swinger integral equation

$$\mathbf{E}_\mathrm{i}(x) = \mathbf{E}_0(x) + \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x,y) k^2(\varepsilon(y) - 1) \times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{i} \qquad [28]$$

Volume integral eqn [28] is equivalent to the corresponding scattering problem. Once the solution is found, the associated scattered field outside the inhomogeneous obstacle can be represented as

$$\mathbf{E}_\mathrm{s}(x) = \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x,y) k^2(\varepsilon(y) - 1) \times \mathbf{E}_\mathrm{i}(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \qquad [29]$$

The Lippman–Swinger integral equation plays an important role under analysis of a wave scattering by local inhomogeneous obstacle.

Born Approximation

In the case where $k^2|\varepsilon(x) - 1| mes D_\mathrm{i} \ll 1$, or where the second term on the right-hand part of the Lippman–Swinger integral eqn [28] is small compared to the first term, one can write $\mathbf{E}_\mathrm{i}(x) \simeq \mathbf{E}_0(x)$. Then the scattered field can be approximately evaluated as

$$\mathbf{E}_\mathrm{s}(x) = \nabla \times \nabla \times \int_{D_\mathrm{i}} g(x,y) k^2(\varepsilon(y) - 1) \times \mathbf{E}_0(y) \mathrm{d}\nu_y, \quad x \in D_\mathrm{e} \qquad [30]$$

The above representation for the scattered field is known as the first-order Born approximation. It is also the first-order approximation in the Neumann series for the integral eqn [28]. The Born approximation plays an important role in diffraction tomography.

Rayleigh Scattering

Rayleigh scattering characterizes the scattering of electromagnetic waves by particles much smaller then an exciting wavelength. Consider a spherical particle with constant permittivity ε and radius a deposited at the origin of a coordinate system. Let linear polarized planewave be incident upon the particle $\mathbf{E}_0(x) = \vec{e}_x\, e^{\mathrm{j}kz}$, and $\vec{e}_x$ be unit vector Cartesian coordinate system. It is polarized in x-direction and propagates along z-axis. In this case, for the scattered intensity, we get:

$$I_\mathrm{s} = \left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 \frac{k^4 a^6}{r^2} \sin^2 \vartheta \qquad [31]$$

where r is distance from the origin and $\vartheta \in [0, \pi]$.

Corresponding scattering cross-section is calculated as

$$\sigma_\mathrm{s} = \frac{8\pi}{3} \left(\frac{\varepsilon - 1}{\varepsilon + 2}\right)^2 k^4 a^6 \qquad [32]$$

Thus the total scattered power is proportional to the fourth power of the wave number (minus fourth power of the wavelength – Rayleigh's Law). The scattered power is also proportional to the sixth power of the radius.

Mie Scattering

The Mie theory is a complete mathematical–physical theory of the scattering of electromagnetic wave by homogeneous spherical particles, developed by Gustav Mie in 1908. In contrast to Rayleigh scattering, the Mie theory embraces all possible ratios of the particle radius to wavelength. It relays upon the exact solution to Maxwell equation for spherical homogeneous obstacle with planewave excitation. In the frame of Mie theory, the scattered field is represented as an infinite series of fields of spherical multipoles, which are orthogonal at a sphere surface. Mie solution plays an important role in light scattering by clouds of particles.

See also

Scattering: Raman Scattering; Scattering from Surfaces and Thin Films; Scattering Phenomena in Optical Fibers; Stimulated Scattering.

Further Reading

Boheren C and Hoffman D (1983) *Absorption and Scattering of Light by Small Particles*. New York: Wiley.
Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.
Chew WC (1995) *Waves and Fields in Inhomogeneous Media*. New York: IEEE Press.
Colton D and Kress R (1992) *Inverse Acoustic and Electromagnetic Scattering Problems*. New York: Springer-Verlag.
Guenther RD (1990) *Modern Optics*. New York: Wiley.
Kong JA (2000) *Electromagnetic Wave Theory*. Cambridge: EMW Publishing.
Kong JA (2002) *Maxwell Equations*. Cambridge: EMW Publishing.

Stimulated Scattering

M Bashkansky and J Reintjes, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Scattering of light involves conversion of an incoming light wave in one mode to an outgoing wave in a different mode. Scattering can involve changes of wavelength, direction, and polarization, and can occur in all types of materials. Without scattering we would not be able to see objects in the world around us. Scattering processes involving particular physical interactions are often given distinctive names. Familiar examples include Rayleigh scattering from the random distribution of air molecules, which makes the sky blue, and Tyndall scattering from small suspended particles, which makes the sunset red. Other types of scattering, e.g., Raman, Brillouin and Rayleigh-wing scattering, are important spectroscopic tools that provide insight into atomic and molecular properties. In their most commonly encountered form, where the intensity of the scattered light is proportional to the intensity of the incident light, the scattering processes are termed spontaneous. With the intense radiation provided by lasers, however, the intensity of the scattered light can grow exponentially, and the processes are termed stimulated scattering.

Stimulated scattering is a subset of the larger discipline of nonlinear optics. The term was first used to describe nonlinear optical interactions (e.g., Raman, Brillouin, Rayleigh) in which the intensity of the scattered light grows exponentially, in a manner similar to the exponential growth of light intensity associated with lasers. These interactions involve exchange of energy with a second mode that is an internal mode of the scattering medium, and the scattered light wave is at a different wavelength from the incoming light. The use of the term stimulated has subsequently been expanded to include a wider class of nonlinear optical interactions, such as parametric fluorescence or certain aspects of self-focusing, in which the second mode is itself a light wave, with either a different frequency or direction of propagation.

In the wider sense, scattering of light occurs through interaction of the incoming light wave with an 'idler' mode in a material. Stimulated scattering occurs when the idler mode is itself driven by interference of the incoming and scattered light waves. Under these circumstances, exponential growth of the scattered wave results. Depending on the interaction, the idler mode may be an internal mode of the material or another light wave. Most linear scattering processes have stimulated counterparts. However, there are stimulated scattering interactions that have no spontaneous analogs.

The elements of stimulated light scattering can be summarized mathematically as follows. The scattering process is modeled as an incoming laser or pump, an outgoing scattered wave and an idler mode. The interaction of the laser with the idler mode to produce the scattered light wave is described by the equation

$$\frac{\partial}{\partial z} A_S(z,t) + \frac{n}{c}\frac{\partial}{\partial t} A_S(z,t) = \kappa A_L b^* \qquad [1]$$

where A_S and A_L are the slowly varying amplitudes of the scattered and laser light waves whose electric fields are defined as

$$E_{S(L)} = \frac{1}{2}\left\{A_{S(L)} e^{i[\omega_{S(L)}t - k_{S(L)}z]} + cc\right\} \qquad [2]$$

$\omega_{S(L)}$ is the frequency of the scattered (laser) light, $k_{S(L)}$ is the corresponding wavevector, κ is a coupling coefficient, cc denotes complex conjugate, and b is the amplitude of the idler mode given by

$$B = \frac{1}{2}\left\{b e^{i[\omega_b t - k_b z]} + cc\right\} \qquad [3]$$

The interaction of the laser and scattered wave to drive the idler mode is described by the equation

$$\mathcal{L}(b^*) = \kappa' A_L^* A_S \qquad [4]$$

where $\mathcal{L}$ is a differential operator that describes the dynamics of the idler mode, e.g., a damped oscillator or a propagating wave, and κ' is a second coupling coefficient. For the class of stimulated processes that involve excitation of an internal mode of the medium, $\mathcal{L}$ has the form

$$\mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{T_2} - v\frac{\partial}{\partial z} \qquad [5]$$

where T_2 is the damping time of the coherence of the internal mode. When interaction times are long compared to internal response and decay times,

$$b^* \sim \kappa' A_L^* A_S$$

and the scattered wave obeys an equation of the form

$$\frac{\partial}{\partial z} A_S(z,t) = \frac{g}{2}|A_L|^2 A_S(z,t)$$

which has an exponential-gain solution for the scattered wave intensity:

$$I_S(z,t) = I_S(0,t)e^{I_L g z}$$

When the idler mode is a light wave, as in stimulated parametric fluorescence, [4] has the same form as [1] and they can be solved as a coupled pair, leading to transcendental or hyperbolic solutions. Within the framework of nonlinear optics, interactions of these types can be expressed as arising from a nonlinear optical polarization, which forms the right-hand side of [1]. The dynamics of the nonlinear polarization are then described by [4].

Stimulated processes are often associated with pumping thresholds. When these processes are configured with cavity resonators, they possess a threshold in pumping intensity just as with lasers or other oscillators. When they are operated in single or multipass scattering cells, they have no true threshold. However, they still have a practical useful threshold when the growth from quantum noise raises the scattered signal to a level comparable to the pump. By convention, this is taken to be the incident light intensity at which the scattered intensity reaches approximately 1% of the pump intensity, typically involving gains of the order of e^{23}–e^{30}. At this level, a small increase in pump intensity raises the scattered intensity to levels at which the pump starts to be depleted.

Attenuation of the scattered light can be accounted for by replacing $I_L g$ in the above expression by $I_L g - \alpha$, where α is the attenuation coefficient. Exponential gain is now experienced when the pump intensity exceeds the level given by

$$I_L > \frac{\alpha}{g}$$

This value can serve as an estimate of threshold pumping in cavity configurations, although in free propagation situations, net gain must still approach e^{23} to exhibit threshold behavior.

A list of stimulated scattering interactions is given in Table 1.

Stimulated Raman Scattering

Stimulated Raman scattering (SRS) involves conversion of an incoming light wave at one frequency, termed the laser or pump wave, into a scattered wave of different frequency, along with excitation or de-excitation of an internal mode of the medium. The frequency difference between the pump and scattered waves is equal to the frequency of the material mode and is characteristic of the material. When the frequency of the scattered wave is lower than that of the incident wave, the scattered wave is termed a Stokes wave, and the internal mode of the medium is excited in the scattering process. When the frequency of the scattered wave is higher than that of the incident wave, the scattered wave is termed an anti-Stokes wave, and the internal mode of the medium is de-excited in the scattering process.

The internal mode of the medium that is involved in Raman scattering is a nonpropagating collective mode. Many different types of internal modes in all types of materials can be involved in Raman scattering. Examples include molecular vibrations and rotations in gases, liquids, or solids, electronic states that are of the same parity as the ground state in gases or solids, optical phonons or polaritons in solids, and Langmuir waves in plasmas. Raman frequency shifts can range from tens of wavenumbers for lattice vibrations to tens of thousands of wavenumbers for electronic transitions.

Spontaneous Raman scattering was first described theoretically by Smekal in 1923 from a quantum mechanical study of light scattering, and was

Table 1 List of stimulated scattering interactions

Stimulated scattering process	*Internal material mode*	*Light mode change*
Stimulated Raman scattering	Nonpropagating internal energy mode: molecular vibrations and rotations, same parity electronic states, optical phonons and Langmuir waves	Change in wavelength, polarization, direction
Stimulated Brillouin scattering	Propagating sound waves	Change in wavelength, polarization, and direction
Stimulated Rayleigh scattering	Temperature and density fluctuations due to light absorption	Change in direction
Stimulated Rayleigh wing scattering	Molecular distribution and orientation fluctuations	Change in wavelength and direction
Stimulated concentration scattering	Fluctuations of concentration in gas mixtures	Change in direction
Photorefractive stimulated scattering	Diffusion of photo-excited carriers	Change in direction
Stimulated Thomson/Compton scattering	Momentum changes in bunched free electrons	Change in wavelength and direction
Stimulated parametric fluorescence	Nonresonant nonlinear polarization	Change in wavelength and direction
Self-focusing	Nonresonant nonlinear polarization	Change in direction

demonstrated experimentally by Raman, and shortly after by Landsberg and Mandelstamm in 1928. Stimulated Raman scattering was first observed in 1962 by Woodbury and Ng, and Eckhardt *et al.* soon after the development of pulsed lasers.

Stimulated Raman scattering can be observed in many different configurations including cavity resonators, single or multiple-pass generator cells or amplifiers. It can be observed as forward scattering, with the scattered light propagating in the same direction as the pump light, or backward scattering, with the scattered light propagating in the opposite direction to the pump light. Its most widespread use is for generation of coherent light at frequencies different from the pump light. Other applications include phase conjugation, optical gating, beam cleanup, and pulse compression.

The equations describing SRS when depletion of the pump light and diffraction of the optical beams are neglected are:

$$2ik_S\left(\frac{\partial A_S}{\partial z} \pm \frac{1}{v_{gS}}\frac{\partial A_S}{\partial t}\right) = -\frac{N}{2}\frac{\omega_S^2}{c^2}\left(\frac{\partial \alpha}{\partial Q}\right)_S Q^* A_L \qquad [6]$$

$$\frac{\partial Q^*}{\partial t} + \Gamma Q^* = -\frac{i}{4\omega_o m}\left(\frac{\partial \alpha}{\partial Q}\right)_S A_L^* A_S \qquad [7]$$

Here A_S and A_L are the slowly varying optical field amplitudes of the Stokes and laser waves as given in [2], $\omega_{S(L)}$ are the optical frequencies, $k_{S(L)}$ are the corresponding k vectors, $v_{S(L)}$ are the group velocities, and Q is the amplitude of the idler mode, in this case the normal mode coordinate of the material excitation, defined by:

$$\mathrm{Q}(x,y,z,t) = \frac{1}{2}\left(Q(x,y,z,t)e^{-i(\omega_o t - k_o z)} + \mathrm{cc}\right) \qquad [8]$$

$(\partial\alpha/\partial Q)_S$ is the Stokes hyperpolarizability, $\omega_0 = \omega_L - \omega_S$ is the Raman transition frequency, and k_0 is the k vector of the material excitation. The phase matching relation $k_S = k_L - k_0$ is automatically satisfied for SRS because the material mode that is involved is nonpropagating. Γ is the half width at half maximum of the Raman linewidth given by $\Gamma = 1/\mathrm{T}_2$, where T_2 is the dephasing time. N is the number of molecules in the ground state and m is the effective reduced mass of the material oscillation.

When the fields are constant or vary slowly compared to the response time of the material, the Stokes intensity is given by

$$I_S(z) = I_S(0)e^{I_L g_{ss} Z} \qquad [9]$$

where the intensity I is given by

$$I_{S(L)} = \frac{1}{2}(cn_{S(L)}\varepsilon_0)|A_{S(L)}|^2$$

and

$$g_{ss} = \frac{N\omega_s}{4\Gamma n_S n_L m \omega_0 c^2 \varepsilon_0}\left(\frac{\partial \alpha}{\partial Q}\right)^2 \qquad [10]$$

where $n_{S(L)}$ is the index of refraction of the Stokes (laser) in the material and ε_0 is the dielectric permittivity of the vacuum.

The Stokes wave grows exponentially with z at a rate that depends on the pump intensity. If the Stokes wave becomes intense enough, its power can become comparable to that of the original pump wave. In that case depletion of the pump wave must be taken into account with an equation for the pump wave amplitude similar to [6]. The Stokes intensity is now given by

$$I_S(z) = \frac{I_S(0)e^{I_L(0)gz}}{1 + \dfrac{\omega_L}{\omega_S}\dfrac{I_S(0)}{I_L(0)}e^{I_L(0)gz}} \qquad [11]$$

which is valid when $I_S(0) \ll I_L(0)$.

In the limit of strong interactions all of the laser energy can be depleted. The maximum Stokes energy is given by $I_S = I_L(\omega_S/\omega_L)$, where ω_S/ω_L is termed the Manly–Rowe ratio. The energy that is removed from the laser but not converted to Stokes radiation is transferred to the material.

For each photon produced in the scattered wave, a corresponding quantum of energy is given to the material mode and a photon is removed from the laser. In principle, 100% of the laser photons can be converted to Stokes photons, but complete conversion is not achieved in practice. Photon conversion efficiencies greater than 90% have been reported in some pulsed laser experiments.

SRS frequency shifts and gains of some materials are shown in Table 2.

Table 2 List of frequency shifts and gains of selected materials for stimulated Raman scattering

Material, λ_L (nm)	*Shift (cm^{-1})*	*Gain (cm/GW)*
H_2 gas (20 atm), 532	4155	2.5
D_2 gas (60 atm), 532	2987	0.45
CH_4 (115 atm), 532	2913	1.26
H_2O, 694	3290	0.14
CS_2, 694	655.6	24
$LiNbO_3$, 694	637	9.4
$Ba(NO_3)_2$, 532	1047	47

Table compiled in parts from: Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2. pp. 1077–1150. Amsterdam, The Netherlands: North Holland and Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin Scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV, New York: McGraw-Hill.

When pulses are used with durations comparable to or shorter than the response time of the material, ranging from picoseconds to femtoseconds depending on the material, the interaction is transient and the effects of the time derivatives in [6] and [7] must be accounted for. Under these conditions the Stokes intensity is given as a temporal integral with Bessel function dependence on the laser intensity. When the pulse duration is much shorter than the material dephasing time and the incident Stokes functional form is the same as the pump, the amplified Stokes intensity is given by the expression

$$I_S(z,\tau) = I_S(0,\tau)I_0^2\left(\sqrt{2g_{ss}\Gamma z\int_{-\infty}^{\tau} I_L(\tau')d\tau'}\right) \qquad [12]$$

where I_0 is a modified Bessel function. In this regime, the Stokes intensity depends only on the integrated pump energy. The Stokes gain is reduced from its steady state value for pulses of constant intensity, but for pulses of constant energy the highest Raman gain is obtained in the limit of transient interactions.

One of the most important applications of Raman scattering is the generation of new coherent waves at frequencies shifted from the pump. In this case, an incident Stokes wave is not supplied as the starting term in [9] or [12]. Rather, the process is initiated by the vacuum fluctuations in the Stokes and material oscillator fields. Quantum field creation and annihilation operators must replace the classical field amplitudes of [6] and [7]. The intensity of the generated Stokes wave is given by the expectation value of the normally ordered number operator and in the extreme transient regime is given by:

$$\begin{aligned} I_S(z, d\tau) = {} & \frac{1}{2A}\hbar\omega_S\Gamma g_{ss}I_L(d\tau)z \\ & \times\left\{I_0^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right)\right. \\ & \left. - I_1^2\left(\sqrt{2\Gamma g_{ss}z\int_{-\infty}^{\tau} I_L(\tau')\tau'}\right)\right\} \end{aligned} \qquad [13]$$

where A is the cross-sectional area of the interaction geometry. The Stokes light generated in this fashion has all the characteristics of amplified spontaneous emission of lasers operated below threshold until the Stokes intensity becomes high enough to deplete the laser.

The stochastic nature of the SRS quantum mechanical source is manifest in the macroscopic Stokes light when the Stokes generator is operated well below pump depletion levels. For a single spatial mode the probability density distribution of

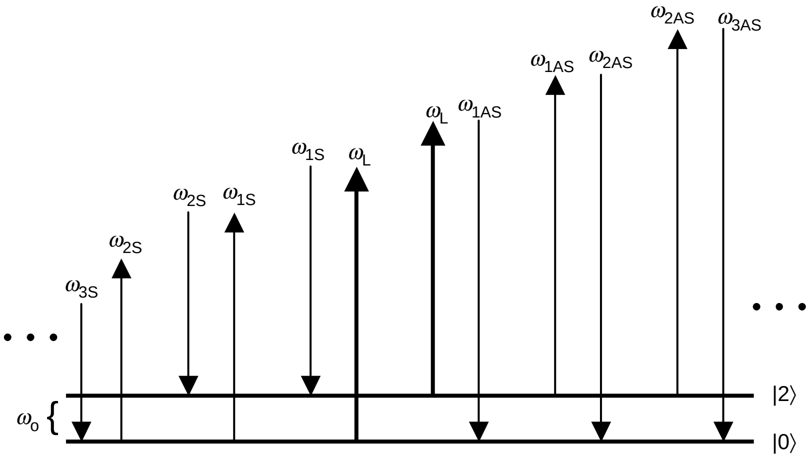

Figure 1 SRS level diagram showing generation of multiple Stokes and anti-Stokes frequencies.

the Stokes energy has the form of a negative exponential:

$$P(W) = \frac{1}{\langle W \rangle} \exp\left[- \frac{W}{\langle W \rangle} \right]$$

Statistical fluctuations in pulse energy, spatial profile, pointing, and in the spectral and temporal structure of the Stokes signal have all been observed.

If the intensity of the generated Stokes light becomes comparable to the pump intensity, it can serve as a pump wave for its own Raman interaction, producing a wave termed the second Stokes wave, shifted from the pump by frequency $2\omega_o$. If the process is driven hard enough third- and higher-order Stokes waves can be produced.

In addition to the interaction between the pump and Stokes beams described above, stimulated Raman scattering can also involve an anti-Stokes wave whose frequency is given by $\omega_{AS} = \omega_L + \omega_o$. The relationship of various Stokes and anti-Stokes waves is depicted in Figure 1. If a population inversion exists between the energy levels of the Raman process, then the anti-Stokes wave experiences exponential gain in a manner similar to that described above for the Stokes gain.

When all the population is in the ground state, which is usually the case, the anti-Stokes wave has exponential loss. The anti-Stokes wave can still grow without population inversion if the phase-matching condition

$$\Delta k = k_{AS} + k_S - 2k_L = 0$$

is approximately satisfied. Because of normal dispersion in the material this condition is usually not satisfied for collinear beams. However, if the Stokes and anti-Stokes beams propagate at small angles as shown in Figure 2, then the phase-matching condition can be met. For angles near but not exactly at phase matching the anti-Stokes wave grows exponentially as part of a mixed Stokes/anti-Stokes mode with a gain given by:

$$g = \mathrm{Re}\left\{ \tfrac{1}{2}(K_3 - K_1)|A_L|^2 - \tfrac{i}{2}\sqrt{\Delta k^2 + 2i\Delta k(K_3 + K_1)|A_L|^2 - (K_1 - K_3)^2|A_L|^4} \right\} \qquad [14]$$

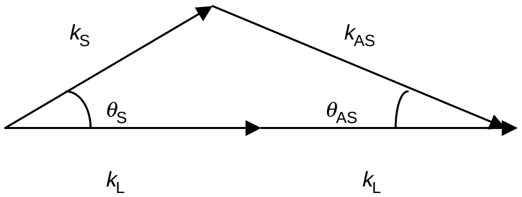

Figure 2 Phase matching diagram for Stokes–anti-Stokes coupling in a medium with positive dispersion.

At exact phase matching the interaction of Stokes and anti-Stokes waves serves to suppress the exponential gain of either wave.

As with Stokes waves, multiple-order anti-Stokes waves can be generated. Stokes and anti-Stokes orders up to at least 13 have been reported.

Stimulated Brillouin and Rayleigh Scattering

Light scattering can also occur through interaction with material density variations. Spontaneous scattering through interaction with this type of internal mode produces a triplet of spectral lines. The up- and down-shifted lines are known as Brillouin anti-Stokes and Stokes lines, respectively, and are due to interaction with acoustic waves. The central peak is termed Rayleigh, or thermal Rayleigh scattering and is due to stationary density fluctuations arising from light absorption. A broad skirt around the central peak, which can be observed in anisotropic media, is due to the fluctuation of molecular distribution and orientation and is termed Rayleigh wing scattering. Each of these processes can give rise to stimulated scattering in the presence of intense radiation.

Density fluctuations are induced directly by light intensity through electrostriction, and indirectly by temperature fluctuations when optical absorption is important. A general classical model of the interaction, including both electrostriction and absorptive heating effects, is described by three equations:

The Navier-Stokes equation:

$$\rho_0\frac{\partial \mathbf{v}}{\partial t} + \frac{v^2}{\delta}\nabla(\Delta\rho) + \frac{v^2\beta\rho_0}{\delta}\nabla(\Delta T) - \eta\nabla^2\mathbf{v} = 2\gamma\varepsilon_0\nabla(\mathbf{E}_{\mathrm{L}}\cdot\mathbf{E}^*_{\mathrm{R/B}}) - 2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho(\mathbf{E}_{\mathrm{L}}\cdot\mathbf{E}^*_{\mathrm{R/B}})\nabla(\Delta T) \qquad [15]$$

The continuity equation:

$$\frac{\partial}{\partial t}\Delta\rho + \rho_0\nabla\cdot\mathbf{v} = 0 \qquad [16]$$

The energy transport equation:

$$\left(\rho_0 C_v\frac{\partial}{\partial t} - \lambda_{\mathrm{T}}\nabla^2\right)\Delta T - \frac{C_v(\delta-1)}{\beta}\frac{\partial}{\partial t}\Delta\rho = 4nc\alpha\varepsilon_0(\mathbf{E}_{\mathrm{L}}\cdot\mathbf{E}^*_{\mathrm{R/B}}) + 4\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\left(T_0\mathbf{E}_{\mathrm{L}}\cdot\frac{\partial}{\partial t}\mathbf{E}^*_{\mathrm{R/B}}\right) \qquad [17]$$

Here $\rho = \rho_0 + \Delta\rho$, $T = T_0 + \Delta T$, $\mathbf{v}$ is the volume element, v is the acoustic wave velocity, $\delta = C_p/C_v$ is the ratio of heat capacities at constant pressure and at constant volume, $\eta = (2\eta_{\mathrm{s}} + \eta_{\mathrm{d}})$ where η_{s} is the shear viscosity and η_{d} is the dilational viscosity, $\gamma = \rho_0(\partial\varepsilon/\partial\rho)_T$, λ_T and β are the thermal conductivity and thermal expansion coefficients, ε is the dielectric constant, and α is the linear absorption coefficient. $\mathbf{E}_{\mathrm{L}}$ is the pump field and $\mathbf{E}_{\mathrm{R/B}}$ is either the Rayleigh or Brillouin field.

Equations [15] and [16] can be combined to eliminate the $\mathbf{v}$ dependence:

$$\left[-\frac{\partial^2}{\partial t^2} + \left(\frac{v^2}{\delta} + \frac{\eta}{\rho_0}\frac{\partial}{\partial t}\right)\nabla^2\right]\Delta\rho + \frac{v^2\beta\rho_0}{\delta}\nabla^2(\Delta T) = 2\gamma\varepsilon_0\nabla^2(\mathbf{E}_{\mathrm{L}}\cdot\mathbf{E}^*_{\mathrm{R/B}}) - 2\varepsilon_0\left(\frac{\partial\varepsilon}{\partial T}\right)_\rho\nabla\cdot[(\mathbf{E}_{\mathrm{L}}\cdot\mathbf{E}^*_{\mathrm{R/B}})\nabla(\Delta T)] \qquad [18]$$

Together with the equations describing the effect of the medium on the optical fields [19a,b], this coupled set of equations describes simultaneously stimulated Brillouin and Rayleigh scattering:

$$\left(\frac{\partial}{\partial z} + \frac{1}{v_{\mathrm{g}}}\frac{\partial}{\partial t} + \frac{1}{2}\alpha\right)A_{\mathrm{L}} = \mathrm{i}\frac{\omega_{\mathrm{L}}}{4cn_{\mathrm{L}}}\left[\frac{\gamma}{\rho_0}\Delta\rho A_{\mathrm{R/B}} + \left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_{\mathrm{R/B}}\Delta T\right] \qquad [19a]$$

$$\left(\frac{\partial}{\partial z} - \frac{1}{v_{\mathrm{g}}}\frac{\partial}{\partial t} - \frac{1}{2}\alpha\right)A_{\mathrm{R/B}} = -\mathrm{i}\frac{\omega_{\mathrm{R/B}}}{4cn_{\mathrm{R/B}}}\left[\frac{\gamma}{\rho_0}\Delta\rho^* A_{\mathrm{L}} + \left(\frac{\partial\varepsilon}{\partial T}\right)_\rho A_{\mathrm{L}}\Delta T^*\right] \qquad [19b]$$

where the slowly varying envelope approximation [2] is used for the waves described by A_{L}, $A_{\mathrm{R/B}}$, $\Delta\rho$, and ΔT. The nonlinear polarization term is related to density and temperature wave amplitudes in the following way:

$$P^{\mathrm{NL}} = \varepsilon_0 E\left[\left(\frac{\partial\varepsilon}{\partial\rho}\right)_T\rho + \left(\frac{\partial\varepsilon}{\partial T}\right)_\rho T\right] \qquad [20]$$

Without the driving terms involving the optical fields, [17] and [18] describe broadband damped fluctuations in the variables $\Delta\rho$ and ΔT. When the optical driving terms are included, resonant behavior emerges. Steady state solutions can be obtained assuming strongly damped material excitations, allowing neglect of $\Delta\rho$ and ΔT derivatives. Separate contributions of the Rayleigh and Brillouin interactions to $\Delta\rho$ can be identified:

Rayleigh contribution:

$$\Delta\rho_{\mathrm{R}} = \left[\frac{\gamma(\Omega + \frac{1}{2}\mathrm{i}\delta\Gamma_{\mathrm{R}}) - \mathrm{i}\gamma_{\mathrm{a}}\Omega_{\mathrm{B}}}{\Omega + \frac{1}{2}\mathrm{i}\Gamma_{\mathrm{R}}}\right]\frac{\varepsilon_0}{4v^2}A_{\mathrm{L}}A^*_{\mathrm{R}} \qquad [21]$$

where γ_{a} is the absorptive coupling constant given by

$$\gamma_{\mathrm{a}} = \frac{2\alpha n v^2\beta}{C_p\Omega_{\mathrm{B}}}$$

The Brillouin contribution is:

$$\Delta\rho_{\mathrm{B}} = \left[\frac{-(\gamma - \mathrm{i}\gamma_a vq/\Omega)q^2\varepsilon_0}{4(\Omega^2 + \mathrm{i}\Omega\Gamma_{\mathrm{B}} - v^2q^2)}\right] A_{\mathrm{L}}A_{\mathrm{B}}^* \qquad [22]$$

In [21] and [22] the frequencies and wavevectors satisfy the following relations:

$$\omega_{\mathrm{L}} = \omega_{\mathrm{R/B}} + \Omega \qquad [23a]$$

$$k_{\mathrm{L}} = k_{\mathrm{R/B}} + q \qquad [23b]$$

where ω_{L} is the incident laser frequency, $\omega_{\mathrm{R/B}}$ is the frequency of the scattered Rayleigh or Brillouin wave, and k_{L}, $k_{\mathrm{R/B}}$ are the corresponding wavevectors. Ω and q are the frequency and wavevector of the density fluctuation. Equations [23a,b] can also be described as energy and momentum conservation relations among incident and scattered photons and a phonon in the quantum mechanical picture.

The Rayleigh and Brillouin contributions exhibit resonant behavior at different frequency shifts. The Rayleigh contribution has its resonant frequency at $\Omega_{\mathrm{R}} = 0$, resulting in typical diffusion behavior. The Rayleigh linewidth $\Gamma_{\mathrm{R}} = 2\lambda_T q^2/\rho_0 C_p$ is the inverse of the decay time of the isobaric density fluctuations. The Brillouin contribution has its resonance at a frequency shift given by

$$\Omega_{\mathrm{B}} = qv$$

where v is the sound velocity. Thus at resonance the density fluctuations associated with the Brillouin effect correspond to sound waves. The Brillouin linewidth

$$\Gamma_{\mathrm{B}} = \frac{q^2}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_v} - 1\right)\right]$$

is the inverse of the phonon lifetime τ_{B}.

In general Brillouin and Rayleigh interactions involve noncollinear scattering with the outgoing wave propagating at an arbitrary angle to the incident wave. In this case [23b] is a vector relation and the interaction geometry can be represented as shown in Figure 3.

The resulting relations among the k-vectors and frequencies of the various waves can place restrictions on the values of frequency shifts that are observed under various circumstances.

The coupled equations for the optical intensities in the steady state take the following form:

$$\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}z} = -g_{\mathrm{R/B}}I_{\mathrm{L}}I_{\mathrm{R/B}} - \alpha I_{\mathrm{L}} \qquad [24a]$$

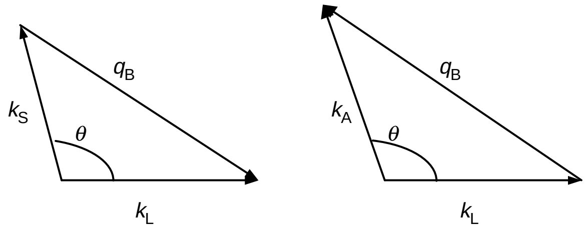

Figure 3 SBS phase matching diagram depicting relations between laser, k_{L}, sound wave, q_{B}, and Stokes and anti-Stokes, k_{S}, k_{A} k-vectors.

$$\frac{\mathrm{d}I_{\mathrm{R/B}}}{\mathrm{d}\zeta} = g_{\mathrm{R/B}}I_{\mathrm{L}}I_{\mathrm{R/B}} - \alpha I_{\mathrm{R/B}} \qquad [24b]$$

where z is the direction of propagation of the incoming wave, ζ is the direction of propagation of the scattered wave, and $g_{\mathrm{R/B}}$ is the gain coefficient of the Rayleigh or Brillouin effect. Equations [24a,b] describe interactions that have exponential gain. Stimulated Brillouin and Rayleigh scattering can be used in both a self-generator and an amplifier mode, but are used most often in a self-generator mode. Because the frequency shifts are small in these interactions, the starting signals for self-generators are thermal excitations of the internal modes rather than vacuum fluctuations.

Stimulated Brillouin Scattering

Stimulated Brillouin scattering (SBS) occurs through interaction with acoustic waves in solids, liquids, and gases and with ion-acoustic waves in plasmas. The Stokes or anti-Stokes frequency shifts for SBS are much smaller than for SRS, with typical values on the order of 0.1–100 GHz, depending on the excitation wavelength and interaction geometry as well as on material properties. Because the internal mode in SBS is a propagating wave, the phase matching condition shown in Figure 3 places restrictions on the frequency of the sound wave that is involved at various scattering angles. SBS is used most often for pulse compression and for correction of phase distortion in propagating light beams through phase conjugation. Although SBS is normally associated with high-power lasers in laboratory experiments, it is also a major factor in limiting the intensity of low-power lasers that can be used in optical fibers because of the long interaction lengths. SBS is also a prominent source of damage in solids and crystals caused by high-power lasers. It can also be used for generation and study of coherent hypersonic waves.

The SBS gain coefficient has both electrostrictive and absorptive contributions, $g_B = g_B^e + g_B^a$ where

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2 n_B n_L c^2 v^2 \rho_0 \Gamma_B} \frac{1}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \qquad [25]$$

and

$$g_B^a = \frac{\omega_B \gamma \gamma_a \Omega_B}{4 n_B n_L \rho_0 v^2 c^2 \Gamma_B} \frac{4(\Omega - \Omega_B)/\Gamma_B}{1 + (2(\Omega - \Omega_B)/\Gamma_B)^2} \qquad [26]$$

The SBS gain from electrostriction has a maximum value of

$$g_B^e = \frac{\omega_B \gamma^2 \Omega_B}{2 n_B n_L c^2 v^2 \rho_0 \Gamma_B} \qquad [27]$$

at the resonant frequency shift

$$\Omega = \Omega_B = qv \qquad [28]$$

For this condition the interaction is with a sound wave in the material. The frequency of the sound wave, and the gain, depend on the angle between the incident and scattered light. Using the vector relation

$$\mathbf{k}_B = \mathbf{k}_L - \mathbf{q}$$

depicted in Figure 3 and the approximations $\Omega_B << \omega_B, \omega_L, |k_B| \approx |k_L|$ we obtain

$$q = 2|k_L| \sin(\theta/2) \qquad [29]$$

The Brillouin frequency shift is then given by

$$\Omega_B = 2 n_L \omega_L (v/c) \sin(\theta/2) \qquad [30]$$

and the Brillouin linewidth is given by

$$\Gamma_B = \frac{4}{\rho_0}\left[\eta + \frac{\kappa}{C_p}\left(\frac{C_p}{C_v} - 1\right)\right] \frac{\omega_L^2 n_L^2}{c^2} \sin^2(\theta/2)$$
$$= \Gamma'' \sin^2(\theta/2) \qquad [31]$$

The peak Brillouin electrostrictive gain is then given by

$$g_B^e = \frac{\omega_B \gamma^2 \omega_L}{n_B c^3 v \rho_0 \Gamma_B} \sin(\theta/2) \qquad [32a]$$

$$g_B^e = \frac{\omega^2 \gamma^2}{n_B c^3 v \rho_0 \Gamma'' \sin(\theta/2)} \qquad [32b]$$

where we have used the approximation $\omega_B \approx \omega_L = \omega$.

The maximum electrostrictive SBS gain is independent of the laser frequency because the Brillouin linewidth is proportional to the square of the laser frequency.

As a function of angle the maximum Brillouin gain is usually observed in the backward direction. This is consistent with the form of the gain coefficient given in [32a], which indicates that the gain is proportional to $\sin(\theta/2)$. However, when the angular variation of the Brillouin linewidth is also accounted for, the gain coefficient is actually minimum in the backward direction [32b]. Typical laboratory experiments involve pencil-like geometries in which the interaction length is significant only in the backward or forward directions. In these situations, the total gain is actually maximum in the backward direction, where the interaction length is longest. Although the formulas appear to indicate that the gain goes to infinity in the forward direction, the damping time, which is proportional to Γ_B^{-1}, also goes to infinity and the interaction becomes transient. Under these conditions the gain parameter is proportional to $\sin(\theta/2)$ and the Brillouin gain goes to zero in the forward direction. Under some geometries Brillouin scattering at 90° to the laser can also be significant. It can be a source of damage to large glass components in high-power lasers.

For thermal Brillouin scattering the gain has a dispersive shape centered on the frequency shift $\Omega = \Omega_B$, with loss occurring on the low-frequency side and gain occurring on the high-frequency side, peaking at a frequency $\Omega = \Omega_B + \Gamma_B/2$. The maximum absorptive SBS gain depends on the laser frequency as ω^{-1}.

Table 3 shows representative values of the Brillouin linewidth and gain coefficients for a selection of substances.

Brillouin-enhanced four-wave mixing is a related effect that involves both a downshifted wave and an upshifted wave in which the coherent interaction is enhanced by the resonance associated with the sound wave. Gains of the order of 10^5 can be realized. Systems based on this effect are often used for phase conjugation.

Stimulated Rayleigh Scattering

Stimulated thermal Rayleigh scattering (R) can be observed in liquids while Stimulated Rayleigh wing scattering (SRWS) is observed in liquids with anisotropic molecules. Additionally, stimulated Rayleigh scattering due to fluctuations of concentration in gas mixtures can also occur. For R the internal mode obeys a diffusion equation. The R gain is given by

$$g_R = g_R^e + g_R^a$$

Table 3 Properties of stimulated Brillouin scattering for selected materials

Substance	*Laser wavelength (nm)*	*Frequency shift (Ghz)*	*$\Delta\nu$ (MHz)*	τ_B (ns)	*g_B^e (cm/GW)*	g_B^a/α (cm/GW)*
Liquid						
Acetone	532	5.93	361	0.44	12.9	22
Benzene	532	8.33	515	0.31	12.3	24
CS_2	532	7.7	120	1.9	130	20
CCl_4	532	5.72	890	0.18	8.77	13
Chloroform	532	5.75	635	0.25	11.7	
Ethanol	532	5.91	546	0.29	12*	10
Methanol	532	5.47	325	0.49	10.6	13
Water	532	7.4	607	0.26	2.94	0.8
Gas						
Xenon (7599 torr)	532	0.654 ± 0.024	98.1 ± 8.9	$0.65\lambda_P^2P$	1.38 ± 0.19	
SF_6(20 bar)	1320	0.2		35		
N_2(100 bar)	1320	0.5		30		
Solid						
BK 7	532	34.65 ± 0.039	165.0 ± 8.6		2.15 ± 0.21	
CaF_2	532	37.164 ± 1.185	45.6 ± 8.8		4.11 ± 0.65	
Plexiglas	532	15.687 ± 0.036	253.7 ± 12.6			
SiO_2	488	35.6	156		4.482	

*Absorptive maximum gain SBS values at 694 nm, with permission from Boyd RW (1992) *Nonlinear Optics*. Academic Press: New York.
Reproduced from Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

where g_R^e and g_R^a are the electrostrictive and absorptive contributions given by

$$g_R^e = \frac{\omega\gamma^2(\delta - 1)}{4\rho_0 n_L n_R v^2 c^2}\,\frac{4\Omega/\Gamma_R}{1 + (2\Omega/\Gamma_R)^2} \qquad [33]$$

$$g_R^a = \frac{-\omega\gamma\gamma_a\Omega_B}{2\rho_0 n^2 v^2 c^2 \Gamma_R}\,\frac{4\Omega/\Gamma_R}{1 + (2\Omega/\Gamma_R)^2} \qquad [34]$$

Both contributions to the STRS gain have dispersive forms. The electrostrictive contribution is maximum for a down-shifted wave with $\Omega = \frac{1}{2}\Gamma_R$. Its maximum value scales linearly with laser frequency. For the absorptive process the gain depends inversely on the laser frequency, and is maximum for an upshifted wave with $\Omega = -\frac{1}{2}\Gamma_R$. Table 4 shows representative values of the Rayleigh linewidth and gain coefficients for a selection of substances.

Table 4 Properties of stimulated Rayleigh scattering for some materials at a wavelength of 694 nm

Substance	*Calculated maximum steady state gain factor*		*Linewidth $\delta\nu_R$ (MHz)*
	g_R^e(max) (cm/MW)	*g_R(max)/α (cm^2/MW)*	
CCl_4	2.6×10^{-4}	0.82	17
Methanol	8.4×10^{-4}	0.32	20
CS_2	6.0×10^{-4}	0.62	36
Benzene	2.2×10^{-4}	0.57	24
Aceton	2.0×10^{-4}	0.47	21
H_2O	0.02×10^{-4}	0.019	27.5
Ethanol		0.38	18

Reproduced from Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman Spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Parametric Fluorescence

Some types of stimulated interactions do not involve exchange of energy with the material. These include parametric down-conversion and self-focusing. In parametric down-conversion an incident wave at frequency ω_L scatters into two longer-frequency light waves at signal and idler frequencies ω_s and ω_i. They satisfy the frequency and wavevector relations

$$\omega_L = \omega_s + \omega_i$$
$$k_L = k_s + k_i$$

The equations for the waves are:

$$\frac{\partial}{\partial z}A_s = -i\left(\frac{\omega_s}{2n_s c}\right)\chi^{(2)}A_L A_i^*$$

$$\frac{\partial}{\partial z}A_i = -i\frac{\omega_i}{2n_i c}\chi^{(2)}A_L A_s^*$$

where A_s and A_i are called the signal and idler fields. The solution for the signal field takes the following form:

$$A_s(z) = A_s(0)\cosh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) - i\sqrt{\frac{\omega_s n_i}{\omega_i n_s}}A_i^*(0)\sinh\left(\sqrt{|\kappa|^2|A_L|^2}z\right) \qquad [35]$$

where

$$|\kappa|^2 = (\omega_s\omega_i/4n_s n_i c^2)|\chi^{(2)}|^2$$

The stimulated parametric interaction provides gain for input signal and idler waves. For large values of $\sqrt{|\kappa|^2 A_L|^2}\ z$ the gain becomes exponential in character, similar to the gain in SRS and SBS. When no incident signals are supplied, the source terms in [35] are the vacuum fluctuations at the signal and idler frequencies. Stimulated parametric fluorescence is the basis for parametric oscillators that provide tunable coherent radiation over a wide wavelength range in the visible and infrared.

Stimulated Photorefractive Scattering

Stimulated light scattering also takes place in photorefractive materials. These materials include crystals such as $LiNbO_3$, $BaTiO_3$, $Bi_{12}SiO_{20}$, and others. They possess a unique property of photo-induced index change with very low light intensities. Some of the potential applications include real-time holography, information recording, phase conjugation, parallel signal processing, and ultrafast optical gating. The nonlinearities leading to stimulated scattering are very large and can produce a significant gain in millimeter thick crystals. In order for amplification of scattered light to take place the nonlinearities have to have a nonlocal character due to diffusion or drift fields. Figure 4 shows a two-beam interaction effect in a photorefractive medium leading to stimulated scattering. Solid lines represent intensity maxima of interfering light beams while dashed lines represent diffused index change maxima. The fact that the two do not overlap allows amplification of the signal beam, with the maximum gain occurring at π/2 phase shift.

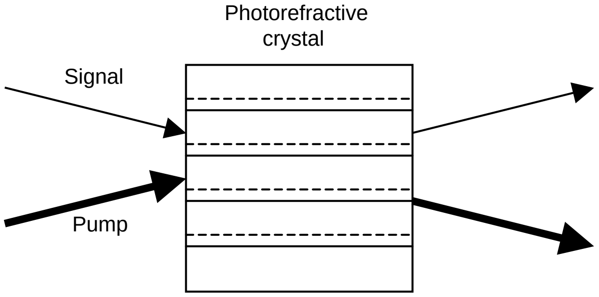

Figure 4 Stimulated photorefractive scattering diagram showing incident pump and signal beams, amplified signal beam, and an index grating formed in the photorefractive crystal.

Stimulated Compton and Thomson Scattering

Light scattering from free electrons is described by Thomson scattering in the low-energy limit, with a cross-section of $(8\pi)/(3)((e^2)/(4\pi\varepsilon_0 mc^2)^2)$, and by Compton scattering in the relativistic limit. Stimulated Thomson or Compton scattering is the source of light emitted from free electron lasers. In these interactions, the electron scatters incoming radiation in its rest frame to a backscattered wave. The ponderomotive force resulting from the interaction with the scattered waves increases the periodic electron bunching, which in turn increases the efficiency of scattering. In some cases, the incoming electromagnetic wave is provided by a microwave source, while in others it arises from a static magnetic wiggler, which appears as an incident electromagnetic wave in the electron rest frame because of a relativistic Lorentz transformation. Free electron lasers are a source of tunable coherent radiation from the microwave region to the ultraviolet.

Self-focusing

In the presence of a strong laser field with intensity I, the refractive index of many materials changes according to the relation

$$n = n_0 + n_2 I$$

where n_2 is termed the nonlinear index. This change in refractive index gives rise to several effects, depending on the temporal and spatial properties of the light. In one type of interaction, a strong forward beam couples to two weaker beams that propagate at small forward angles, providing gain for the weaker beams. One of the weak beams interferes with the strong forward beam to form a phase grating through the nonlinear index, providing gain for the second weak beam. The interference of the second weak beam with the forward beam forms a second phase grating, providing gain for the first weak beam, thereby completing the coupling for the stimulated process. This effect is similar to Stoke/anti-Stokes Raman scattering, except that the coupling is between waves propagating in different directions rather than between waves with different frequencies.

For interactions with an instantaneous relaxation time, the gain for waves propagating at a small angle to the forward direction is given by

$$g = \frac{K_\perp}{2k}\sqrt{K_{\max}^2 - K_\perp^2}$$

where $K_\perp = \sqrt{k_x^2 + k_x^2}$ is the transverse wavevector, corresponding to the angle

$$\theta = \frac{K_\perp}{k}$$

and

$$K_{\max} = 4k^2 n^{\mathrm{L}} n_2 I$$

is the largest transverse k-vector for which gain is observed. The maximum gain is given by

$$g_{\max} = \frac{K_{\max}^2}{4k}$$

and occurs for an angle corresponding to

$$K_\perp = \frac{K_{\max}}{\sqrt{2}}$$

When the physical interaction that causes the refractive index to change with light intensity has a nonzero relaxation time, gain can be observed at angles larger than that given by $K_{\max}$.

This stimulated four-photon effect can be interpreted as the initiating stage of self-focusing. In the self-focusing effect, a laser beam with a normal mode profile (more intense in the center than at the edges) creates a positive lens in the material, focusing the beam. As the beam intensifies, the power of the lens increases, causing the beam to focus more tightly and eventually coming to a catastrophic focus if other nonlinear effects do not limit the intensity. The four-photon stimulated effect described above results in the spread of k-vectors of the forward beam, corresponding to a reduction in its diameter.

Self-focusing is responsible for breakdown and damage in many solids. Self-focusing in high-power lasers is commonly suppressed by restricting transverse k-vectors to values well below those of maximum gain.

See also

Nonlinear Optics, Applications: Self-Focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Photorefraction. **Scattering:** Scattering Theory.

Further Reading

Bloembergen N (1965) *Nonlinear Optics*. New York: Benjamin.

Bloembergen N (1967) The stimulated Raman effect. *American Journal of Physics* 35: 989–1023.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic Press.

Fabellinskii IL (1968) *Molecular Scattering of Light*. New York: Plenum.

Gunter P and Huignard J-P (eds) (1989) *Photorefractive Materials and their Applications II*. Springer-Verlag Topics in Applied Physics, vol. 62. Berlin: Springer-Verlag.

Kaiser W and Maier M (1972) Stimulated Rayleigh, Brillouin and Raman spectroscopy. In: Arecchi FT and Schulz-DuBois EO (eds) *Laser Handbook*, vol. 2, pp. 1077–1150. Amsterdam, The Netherlands: North Holland.

Milanovich F (1986) Stimulated Raman scattering. In: Weber MJ (ed.) *Handbook of Laser Science and Technology. Optical Materials*, vol. III. Boca Raton, FL: CRC Press.

Penzkofer A, Laubereau A and Kaiser W (1979) High intensity Raman interactions. *Progress in Quantum Electronics* 6: 55–140.

Rabin H and Tang CL (eds) (1975) *Quantum Electronics: A Treatise*. Part A, vol. 1. New York: Academic Press.

Raymer MG and Walmsley IA (1990) The quantum coherence properties of stimulated Raman scattering. In: Wolf E (ed.) *Progress in Optics*, vol. XXVIII, p. 183. Amsterdam, The Netherlands: Elsevier.

Reintjes JF (1984) *Nonlinear Optical Parametric Processes in Liquids and Gases*. New York: Academic Press.

Reintjes JF (1995) Stimulated Raman and Brillouin scattering. Handbook of laser science and technology, supplement 2. In: Weber MJ (ed.) *Optical Materials*, pp. 334–364. Boca Raton, FL: CRC Press.

Reintjes J and Bashkansky M (2001) Stimulated Raman and Brillouin scattering. In: Bass M (ed.) *Optical Society of America Handbook of Optics*, Ch. 18, 2nd ed., vol. IV. New York: McGraw-Hill.

Shen YR (1984) *Principles of Nonlinear Optics*. New York: Wiley.

Shen YR and Bloembergen N (1965) Theory of stimulated Brillouin and Raman scattering. *Physical Review* 137: A1787.

SEMICONDUCTOR MATERIALS

Contents

Amorphous Semiconductors
Band Structure Engineering
Dilute Magnetic Semiconductors
GaAs Based Compounds
Group IV Semiconductors, Si/SiGe
III-Nitrides
Large Gap II–VI Semiconductors
Lead Salts
Mercury Cadmium Telluride
Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures
Quantum Dots
Type-II Quantum Wells and Superlattices

Amorphous Semiconductors

J I B Wilson, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Amorphous semiconductors comprise several classes of materials having a range of useful properties that have made many of them commercially important. Their special characteristics are derived from the particular techniques used in their synthesis, and include the ability of some to change their structure under the influence of heat or illumination. Many are photoconductive, making them valuable for photodetectors when large area devices are required (e.g., xerography), or photovoltaic, hydrogenated amorphous silicon being widely used in thin-film solar cells.

'Amorphous' means noncrystalline, thus encompassing glassy materials and disordered covalent-bonded materials. Glasses, such as the amorphous chalcogenides (glassy sulfides, selenides, and tellurides), melt to viscous liquids when heated, whereas disordered semiconductors like amorphous silicon are metastable thin films that vaporize or crystallize when heated. Two other groups of glassy semiconductors, the covalent-bonded II–IV–V_2 compounds and ionic-bonded oxides (e.g., $V_2O_5P_2O_5$) will not be discussed here. Amorphous organic semiconductors are receiving much interest now, because of their potential use in low-cost display devices, but are more logically treated as polymer materials.

The following sections explain the main preparation methods and the structures that result, the importance of defects and the influence these have on optical and electrical behavior, and some of the most important applications.

Amorphous Structure

The common property that distinguishes all amorphous materials from crystalline ones is the absence of long-range order in their atomic structure. These disordered solids have no periodic or translational symmetry. The term is sometimes loosely applied, and it should be noted that elements and compounds that are able to form several structures might contain different phases. An example is 'hard carbon' which in thin-film form may contain tetrahedral amorphous carbon and hydrogenated amorphous carbon, as well as microcrystallites of graphite and diamond. Interestingly, the prediction in 1989 by A.Y. Liu and M.L. Cohen, that crystalline cubic-C_3N_4 would be even harder than diamond, has not been confirmed by experiment. Only a hard amorphous C_xN_y has been produced, even by nitrogen ion beam assisted sputtering of graphite, which it was thought would have provided sufficient energy to form the close C–N bonds.

When the common crystalline semiconductors, Si and Ge, melt they become liquid metallic conductors. This is because the solid is formed by directional covalent bonds that produce tetrahedral packing of the atoms. In the liquid the bonds are weakened and a closer packed structure found in metals results. However a less ordered atomic arrangement results from condensing vaporized Si or Ge on to a cold substrate. In this case the atoms cannot move into their preferred positions for either tetrahedral or close packing arrays and instead are frozen almost wherever they land. Nonetheless, the arrangement of atoms is closer to that of the tetrahedral crystalline semiconductor than that of the metallic liquid. The radial distribution function of an amorphous semiconductor provides the numbers of atoms versus distance from an arbitrary atom. It shows that the nearest-neighbor distance is the same as in the equivalent crystalline material, although with some spread of values, but that the separation of atoms further away has a less well-defined value. Thus pure amorphous Si and Ge retain the directional covalent bonds between nearest neighbors but have distorted bond angles that make next nearest neighbors lie in different positions from the equivalent crystalline solid. Consequently there is only short-range order and the atoms form a continuous random network, a concept introduced by W.H. Zachariasen in 1932 for glasses. The deposition conditions generally lead to further disruptive effects that will incorporate small voids where the distortion is too large to allow a bond between atoms. These voids together with broken bonds give a defective network that is far from the ideal random tetrahedral 'lattice'.

Amorphous Semiconductor Energy Bands

A frequently asked question is 'Why can amorphous materials show semiconducting behavior?' since they do not have the periodic structure that leads to the allowed energy band scheme in crystalline semiconductors. However, note that metals do not become insulators when they melt, demonstrating that electrical conductivity does not require a periodic atomic structure. In fact, electron energy levels in solids may also be derived by considering chemical bonds when discrete atoms are brought closer together. The valence band in a semiconductor arises from the electron bonding orbitals and the conduction band from the antibonding electron orbitals. The forbidden energy gap (bandgap) is the separation between these two bands, which are only partially filled in semiconductors at room temperature. Thus amorphous semiconductors will have bonding and antibonding energy bands, and a more or less well-defined bandgap according to the defect content.

In fact, pure amorphous Si and Ge are extremely poor semiconductors due to an ill-defined bandgap and the large concentration of defects. Distorted bonds produce allowed electron energy levels that lie adjacent to the well-defined bands in crystals, producing tails of energy levels extending into the bandgap. Electrons that take up these 'tail state' energies are not entirely free to move through the solid but can only hop from one defective site to another. Additional defects arise from unsatisfied covalent bonds that 'dangle' in the spaces between the atoms. These produce electron energy levels within the forbidden gap (Figure 1). Indeed, their concentration is so high in amorphous silicon produced from a condensed silicon vapor, that they fix the position of the Fermi level at approximately the middle of the bandgap and the semiconductor has a high resistivity. The usual option of reducing a semiconductor's resistivity by adding small amounts of dopants from adjacent columns of the periodic table has little effect in this case. A donor atom from group V (e.g., phosphorus) may have all five valence electrons satisfied by a locally distorted arrangement of silicon atoms, instead of having only four bonded electrons and a loosely bound electron as in crystalline silicon. A similar picture exists for an acceptor atom from group III (e.g., boron).

Semiconductors containing group VI elements, S, Se, or Te, have two-fold coordination instead of the tetrahedral arrangement of the group IV elements, Si and Ge. The bonding between atoms in the solid is formed from four of the six p-electrons, which give

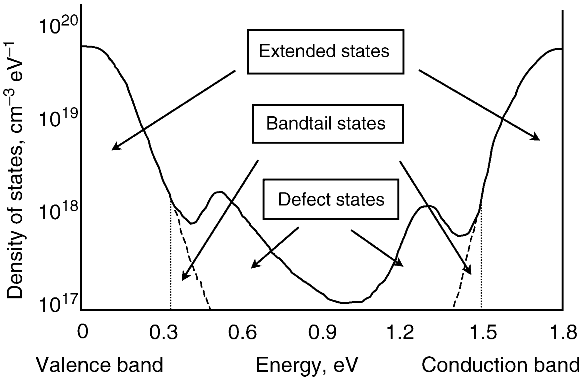

Figure 1 Schematic density of states for hydrogenated amorphous silicon.

rise to the bonding and antibonding energy bands of the solid, and the two remaining p-electrons from each atom form a non-bonding 'lone pair'. The lone pair energy band lies between the bonding and antibonding bands, and is the highest filled band in the solid, thus being the valence band.

Just as in amorphous Si or Ge, dangling bond defects will have energies between the valence and conduction bands, but in lone-pair semiconductors such disturbed energy states will now produce energy levels that fall within the lone pair band instead of within the bandgap. This makes it difficult to dope chalcogenide semiconductors, especially by donors, since the additional energy levels often lie within an existing energy band. Similar considerations apply to ternary chalcogenide alloys: these will have lone-pair bands arising from the group VI element, with additional energy levels from the other components (e.g., from group IV or V). However, amorphous chalcogenides will also contain three-fold coordinated chalcogenide atoms and these defects can associate in pairs or may interact with an adjacent two-fold coordinated atom. Defect centers with two oppositely charged chalcogens may form a 'valence alternation pair' (VAP).

So it may be seen that the semiconducting behavior of solids arises from the nearest-neighbor interactions of the atoms giving rise to allowed electron energy bands, and that long-range order is not necessary to produce the semiconductor's forbidden energy gap. Defects then produce localized energy levels that usually lie within the forbidden gap.

Preparation of Amorphous Silicon

Despite the early use of selenium as a photoconductive semiconductor in xerography, most amorphous semiconductors might have stayed as curious but technologically impractical materials, if an alternative preparation method to physical vapor deposition had not been discovered.

In the 1950s the preparation of high-purity single-crystal silicon was not yet established, and work at Standard Telecommunication Laboratories, UK, was attempting to use radio-frequency heating to produce silicon from low-pressure silane gas (SiH_4). It was noticed that a side effect was a deposit of resistive, noncrystalline material on cool parts of the container, close to unwanted plasma 'glow discharges'. In following up this observation, the technique was developed to produce a variety of other useful compounds from electrical discharges in appropriate gas mixtures, particularly silicon nitrides and oxides. With new scientific interest in amorphous semiconductors, the silicon coatings from silane were examined in more detail, and were shown to be significantly different from amorphous silicon produced by evaporating and condensing silicon. This new form was very photoconductive and could be doped n-type by adding phosphine to the silane. A more detailed study of doping led to the seminal papers by W.E. Spear and P.G. LeComber of Dundee University, UK, and to an explosive growth in the whole subject, once the potential commercial applications became apparent.

Despite many alternative synthesis methods, plasma chemical vapor deposition (CVD) remains the most widely used, and produces material that has not been surpassed in quality. The key features are a vacuum chamber into which a controlled flow of each of the gaseous sources is fed, and a heated plate on which the substrate is placed for coating. The decomposition of the gases is driven by an electrical discharge between two electrodes, which enables the gas temperature to be much lower than in conventional pyrolysis (thermal CVD). An important aspect of the process is the incorporation of hydrogen into the deposited material, which is thus more accurately known as hydrogenated amorphous silicon, a-Si:H. In the best material, hydrogen is mainly bonded singly as Si–H, but if the preparative conditions are not optimum then there may also be large amounts of $Si{-}H_2$ (as may be seen from the infrared absorption spectra). Alternative source gases usually contain hydrogen (e.g., Si_2H_6), and hydrogen may be added as a diluting agent for the doping gases (e.g., B_2H_6 or PH_3). Other precursors include fluorinated gases, but these have failed to give the predicted improvements in performance of the semiconductor in such applications as solar cells, and have greater chemical reactivity.

Since silane is spontaneously flammable in air, and the dopant gases are not only flammable but also very toxic, efforts have been made to produce a-Si:H by other means. Sputtering from undoped or doped silicon targets has been a popular choice, but it is still necessary to incorporate hydrogen into the deposited material, for instance by adding it to the usual argon sputter gas. This is because hydrogen serves to passivate unsatisfied 'dangling' Si bonds and to reduce the stresses that would occur in a pure silicon network. Photo-CVD, using ultraviolet light to drive the reaction without raising the temperature, can produce high-quality material but has the problem of unwanted silicon deposits on the illuminated reaction chamber window. Although conventional silane pyrolysis is not suitable for producing a-Si:H, because the temperatures needed for useful deposition rates also remove much of the hydrogen if the substrate is too hot, it is possible to heat the gas separately from the substrate by means of a hot filament and so produce good-quality films.

Other hydrogenated silicon alloy semiconductors are synthesized by additions to silane of ammonia (giving amorphous silicon nitride, a-Si:N:H), or a hydrocarbon gas such as methane (giving a-Si:C:H), or germane (giving a-Si:Ge:H).

Preparation of Amorphous Chalcogenides

Unlike the tetrahedral group IV elements, useful members of this class of amorphous semiconductors may be produced by vacuum evaporation and condensation on to a relatively cool substrate. Compounds, such as As_2Se_3, do not evaporate congruently and so the deposit will change composition as the evaporation proceeds. More uniform compositions may be obtained by sputtering with argon, and glassy oxides may also be obtained in this way with oxygen added to the inert gas. Chalcogenide glasses may also be prepared by rapid cooling from a melt, or by chemical vapor deposition from heated vapor precursors. Since glasses are thermodynamically unstable, they will tend to transform to a more stable, sometimes crystalline structure. The rate for this process depends on the temperature and may be effectively too slow to observe.

Optical Properties of Amorphous Semiconductors

Optical absorption spectra of amorphous semiconductors clearly show a forbidden energy gap. By plotting a suitable function of the absorption coefficient, α, versus photon energy, $h\nu$, usually that given by J. Tauc, $(\alpha h\nu)^{1/2}$, the 'optical gap' may be determined from the intercept of the linear portion of the plot with the x-axis. Unlike crystalline semiconductors, amorphous ones have no k-selection rules to give indirect and direct transitions across the bandgap, thus a-Si:H behaves as a direct gap material. This 'optical gap' is greater than the bandgap determined from electrical conductivity measurements, since the latter will have contributions from both localized bandtail states (low electrical mobility) and extended states (high electrical mobility) (Figure 2).

The absorption spectra also have a portion at lower photon energies having an exponential dependence on photon energy, known as the Urbach tail. This is related to the structural disorder in the material and the tail states. Transitions involving defect bandgap states lie at even lower photon energy, and the density of defects can be determined from the weak absorption in this region of the spectrum, perhaps by using photothermal deflection spectroscopy to detect the weak optical absorption.

Photoluminescence also provides a probe of the defect content of many materials. Interpretation and identification of the actual defect (e.g., chemical impurity or structural imperfection) is aided by

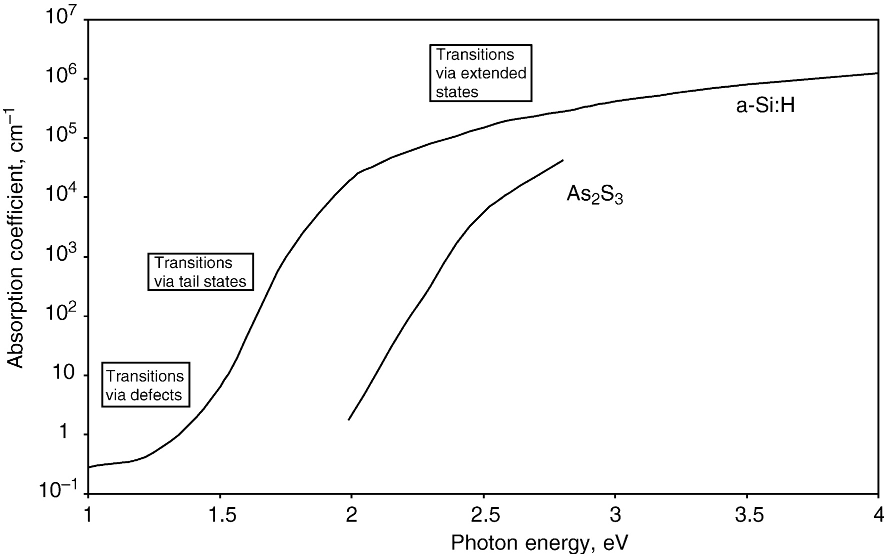

Figure 2 Typical optical absorption spectra of hydrogenated amorphous silicon and arsenic sulfide.

reference to the crystalline form. Both crystalline and amorphous forms of chalcogenide have a wealth of special effects, including reversible photoinduced bleaching and darkening, and laser-induced crystallization and amorphization. In addition the creation and dispersal of defects by appropriate illumination of some amorphous chalcogenides is accompanied by changes in the electron spin resonance (ESR) signal. Some of these effects have been the basis for applications in optical memories. More prosaically, chalcogenide glasses are important infrared transparent materials and amorphous chalcogenide thin films have had historical importance as the photoconductive layer in xerography, although here they are now being supplanted by less toxic amorphous silicon alloys.

Electrical Properties of Amorphous Semiconductors

The electrical conductivity of amorphous semiconductors is often controlled by the defects that are introduced by impurities or by structural flaws. For a-Si:H this provides alternative conduction mechanisms, via extended state electron and hole conduction (as in crystalline silicon), and via hopping of electrons or holes through localized energy levels within the bandtails or within the bandgap. The relative importance of each route depends on the temperature and on the density of states (DOS) in each energy range, as well as on the mobility of the charge carriers.

These different conduction mechanisms are clearly seen in the temperature dependence of the conductivity of a-Si:H and of many amorphous chalcogenides, since a semilog plot of conductivity versus $1000/T$ K will have portions with different slopes as each mechanism dominates a particular temperature range. Near room temperature, good-quality a-Si:H films in the dark exhibit a well-defined Arrhenius relationship with an activation energy that is the energy separation of the Fermi level from the conduction band extended states, electrons dominating the conductivity. The conduction band extended states do not have the high electron mobility found in crystalline silicon, due to the disruption of long-range order and the consequent increased electron scattering. Typical electron mobilities in a-Si:H are $\sim 1\ \mathrm{cm^2\ v^{-1}\ s^{-1}}$, and hole mobilities are even lower.

The band tail states have even worse carrier mobilities, since movement is by hopping to vacant sites that might lie at some distance, and hopping is unlikely between sites that have a large energy difference. (The process has an inverse exponential dependence on the energy difference between the initial and final states.) Hopping of carriers between states that lie within the bandgap is a significant contribution to the conductivity only at low temperatures, unless the density of such defects is high. In materials having weakly localized states, nearest-neighbor hopping is supplanted by variable range hopping, as carriers prefer to hop to a more distant site if it has a smaller energy difference from the starting site than do the neighboring sites. As the temperature is lowered, this hopping distance will increase and the conductivity has N.F. Mott's characteristic $\exp(-B/T^{1/4})$ dependence on temperature, T. Lower-quality films deposited by plasma CVD on to unheated substrates tend to be p-type due to hopping conductivity through the high density of defect states. The link between substrate deposition temperature and material quality is caused by the way in which hydrogen is incorporated into the growing material: high temperatures allow only singly bonded hydrogen, which passivates dangling bonds thus removing bandgap defect energy levels. Low temperatures also allow hydrogen to be incorporated as $Si{-}H_2$ and this disrupts the Si–Si lattice bonding, producing more defect levels. Confirmation of this effect is obtained from sputtered amorphous silicon, with and without hydrogen added to the argon.

When a-Si:H is illuminated it may show strong photoconductivity, depending on the preparative conditions. Free carriers are generated by the illumination but may fall into defect states (energies within the band gap) where they can be released or can recombine with the oppositely charged carriers. The detailed response to light (wavelength, intensity, duration) and temperature thus depends on the distribution and concentration of defects. Luminescence effects are similarly influenced. Most amorphous chalcogenides are only weakly photoconductive due to their high density of gap states, those near the middle of the gap (e.g., VAPs) being highly effective recombination centers.

The electrical properties of other hydrogenated amorphous tetrahedral semiconductors based on silicon are determined by their defect populations, and in general they are less well behaved. a-Si:Ge:H and a-Si:Sn:H have reduced bandgaps and a-Si:C:H and a-Si:N:H have increased bandgaps, which offer many possibilities for multilayer device design, provided that their normally high defect contents can be reduced or controlled.

The study of glassy semiconductors was stimulated in the 1960s by the discovery of electrical switching in sputtered multicomponent thin films.

These simple devices changed from a nonconducting to a conducting state when the applied electric field exceeded a certain threshold, and returned to their initial state when the current fell below a certain value. A variety of other switching types was discovered, especially under pulsed operation, including those having a memory effect based on the reversible crystallization and vitrification of a narrow filament between the contacts. A theoretical model of these devices comprises both electronic and thermal effects, which are responsible for initiating the switching event and for memory effects. A practical problem has been the variable jitter in the switching time.

Applications

Amorphous silicon alloy films are valuable as the active layers in thin-film photovoltaic cells, two-dimensional optical position detectors, linear image sensors (optical scanners), and thin-film transistors used in liquid crystal display panels. They also have uses as antireflection coatings and planar optical waveguides. Amorphous chalcogenide films have been used as electrical switches and memory devices, for optical data storage, and as dry photoresists and electron beam resists. Both types of amorphous material have been used in xerography and medical X-ray imaging detectors.

Amorphous silicon solar cells have power conversion efficiencies of ~12% for the most complicated structures. These are tandem cells that use different alloys (including a-Si:C:H) for the various layers, in order to enhance effective absorption of the solar spectrum. A serious drawback of using amorphous semiconductors in electronic devices is their short carrier diffusion length. In solar cells this is circumvented by ensuring that there is a built-in field to separate photogenerated electron–hole pairs before they can recombine. Because doped amorphous silicon alloys have high defect densities, it is difficult to make the effective pn junctions that provide such built-in fields in conventional crystalline silicon cells. Instead, amorphous silicon cells use pin structures, where the i-layer is effectively undoped and provides an extended electric field between the p-i and i-n junctions. Long periods of illumination increase the dark current in these devices, as additional defects are generated. This 'Staebler–Wronski effect' is believed to be associated with hydrogen movement at a passivated dangling bond (Si–H) and a nearby weak Si–Si bond. It is a cause of slow degradation in unstabilized solar cells although the change may be reversed by thermal annealing.

In contrast to 'minority carrier' photovoltaic devices, thin-film transistors necessitate the control of majority carriers alone. To address each pixel in a liquid crystal display, a switching element is needed and simple thin-film transistors fulfill this role. Although low carrier mobilities in amorphous semiconductors restrict the current passed and switching speeds, the requirement for coating large area glass panels at low temperatures make a-Si:H with a-Si:N:H gate dielectric almost the only options. Improved performance is possible by using pulsed laser irradiation to change the amorphous semiconductor to polycrystalline material.

Further Reading

Brodsky MH (ed.) (1985) *Amorphous Semiconductors*, 2nd edn. Berlin: Springer-Verlag.

Elliot SR (1984) *Physics of Amorphous Materials*. London: Longman.

EMIS Datareviews Series No.1 (1989) *Properties of Amorphous Silicon*, 2nd edn. London: INSPEC.

Joannopoulos JD and Lucovsky G (eds) (1984) *The Physics of Hydrogenated Amorphous Silicon*. Topics in Applied Physics, vols 55 and 56. Berlin: Springer-Verlag.

Journal of Non-Crystalline Solids (2000) *Proceedings of the 18th International Conference on Amorphous and Microcrystalline Semiconductors (1999)*, vol. 266–269. This journal has published the proceedings of most of these well-established annual meetings dealing with all aspects of amorphous semiconductors.

Kramer B and MacKinnon A (1993) Localization – theory and experiment. *Reports on Progress in Physics* 56: 1469–1564.

Luft W and Tsuo YS (1993) *Hydrogenated Amorphous Silicon Alloy Deposition Processes*. New York: Marcel Dekker.

Madan A and Shaw MP (1988) *The Physics and Applications of Amorphous Semiconductors*. San Diego, CA: Academic Press.

Mott NF (1985) Electrons in non-crystalline materials – the last twenty five years. *Contemporary Physics* 26: 203–215.

Pankove JI (volume ed.) (1984) Hydrogenated amorphous silicon. In: Willardson RK and Beer AC (series eds) *Semiconductors and Semimetals*, vol. 21, Parts A–D. Florida: Academic Press.

Zallen R (1983) *The Physics of Amorphous Solids*. New York: Wiley.

Band Structure Engineering

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Direct gap semiconductors have many unique properties which are advantageous for optical and electronic applications, and in particular for detectors and emitters of infrared radiation associated with interband optical transitions. Gallium arsenide, GaAs, based structures have principally been used in the near-infrared, and so-called narrow gap semiconductors in the mid-infrared (MIR) spectral regions. More recently systems based on intersub-band transitions have shown great promise for MIR devices, with the obvious advantage for some applications that they can be constructed from large gap, better controlled materials (mainly InP- and GaAs-based to date) and have a peak operating frequency which is determined simply by the design of the particular quantum structure. Both approaches have been substantially advanced by band structure engineering techniques, reducing Auger losses and phonon losses, respectively, which is the subject of the present work.

The classic examples of narrow gap materials are the III–V semiconductors (InSb, InAs, and their related alloys), II–VI semiconductors (mainly $Hg_{1-x}Cd_xTe$), and the lead salt alloys. As a direct result of the small energy gap, these materials have a very low electron effective mass, very high electron mobility, even at room temperature, and a high saturation velocity. In addition they have a high ratio of stimulated to spontaneous transition probabilities and a large nonlinear susceptibility. However, they have the disadvantage of a high intrinsic carrier density at room temperature and thermal generation rate, so that a fundamental limitation to the performance of devices at elevated temperatures (preferably room temperature) is nonradiative losses resulting from Auger recombination processes. One solution is to cool these systems to 80 K or below which involves expensive and bulky cooling methods such as Stirling-cycle coolers, negating their advantages over other, otherwise inferior, devices. In the past 10 years there has been an intensive effort to circumvent this limitation by quantum structure and so-called 'band structure engineering' techniques made possible by epitaxial growth techniques, whereby the Auger processes are suppressed. Since the emphasis of the band structure engineering work has been mainly towards the optimization of III–V semiconductor lasers rather than detectors, we reflect that emphasis here. Although developed more recently than the bipolar diode lasers, the unipolar quantum cascade (QC) lasers provide perhaps the best example of systems that can be conveniently optimized by band structure engineering techniques. In this case, phonon processes dominate so that the design is to do with phonon suppression. Thus, we are concerned principally with suppression of Auger processes for interband devices, and phonon processes for intraband devices.

Nonradiative Processes and Band Structure Engineering

Suppression of Auger Recombination

Electron recombination processes are dealt with in detail elsewhere in the Encyclopedia (*see* **Semiconductor Physics:** Recombination Processes) but we briefly summarize the main points here. It is well known that there are three principal interband recombination mechanisms by which thermodynamic equilibrium is established in semiconductors: Shockley–Read recombination, radiative recombination, and Auger recombination. These are shown schematically in Figure 1. The Shockley–Read recombination occurs via extrinsic energy levels produced by lattice defects and impurities, Figure 1a, and can in principle be controlled by the procedure used to grow the material; hence it does not provide a fundamental limit to the

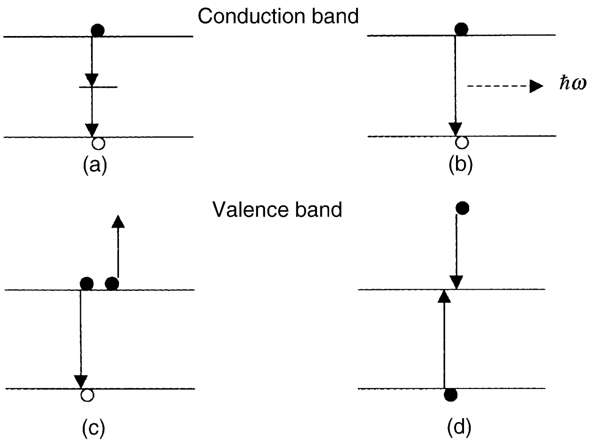

Figure 1 Electronic processes in semiconductors: (a) Shockley–Read recombination; (b) radiative recombination; (c) Auger recombination; (d) impact ionization.

carrier lifetime. The radiative and Auger mechanisms (Figure 1b,c, respectively) are, however, fundamental and are determined by the electronic structure of the material. The former is central to the operation of light-emitting diodes and is approximately inversely proportional to the electron and hole concentration. Optimum detector operation is achieved in materials that are limited by the radiative lifetime mechanism. Thus, in both cases it is desirable to minimize Auger processes.

By contrast with the two-particle radiative recombination process, Auger recombination is a three-particle interaction. For example in one so-called CHCC (or Auger-1) process an electron collides with a second electron; the first electron recombines with a hole in the valence band and the energy produced by the electron–electron collision excites the second electron higher up into the conduction band. This second electron eventually de-excites by phonon emission, and can cause observable heating of carriers by an Auger type of up-conversion. However, for the present article we are only interested in the first part of the process, so we do not consider the heating further here. The Auger process is the inverse of impact ionization where an energetic conduction electron collides with a valence electron, thereby creating an electron–hole pair as shown in Figure 1d. The final state consists of two low-energy electrons and one hole. Analogously, in Auger recombination the initial state has three particles (e.g., two electrons and one hole) and the final state one (e.g., one electron). Due to this three-body nature, the collision probability (or transition rate) per unit volume, R, varies with the cube of the carrier concentration and (if $n = p$) may be written $R = Cn^3$, where n is the electron density and C is referred to as the Auger coefficient. Note that the Auger rate is sometimes defined per carrier as $R' = Cn^2 = 1/\tau_A$, where τ_A is the Auger lifetime.

Narrow gap III–V semiconductors have the familiar three-band (so-called Kane energy band structure – (*see* **Semiconductor Physics:** Band Structure and Optical Properties) consisting of an s-like conduction band separated by a small energy gap from a triply degenerate p-like valence band, part of whose degeneracy is raised by spin–orbit splitting as shown in Figure 2. Thus, a large number of permutations of Auger processes become possible for n- and p-type semiconductors. Those which are initiated by a collision between two electrons are referred to as conduction, or n-n, Auger processes (e.g., the CHCC process), and those by a collision between two holes as valence, or p-p, Auger processes. It can be shown that the n-n process (so-called CHCC or Auger-1) illustrated in Figure 3a is dominant in bulk n-type and the p-p process (so-called CHLH or Auger-7) shown in Figure 3c is dominant in bulk p-type semiconductors, and so these have received the most attention. For larger gap (shorter wavelength) semiconductors there is sometimes a significant resonance between the bandgap and the spin–orbit splitting and the p-p process (CHSH) shown in Figure 3b becomes important. Clearly these simple descriptions require extension when the more complex band structure of strained layer superlattice quantum systems is considered.

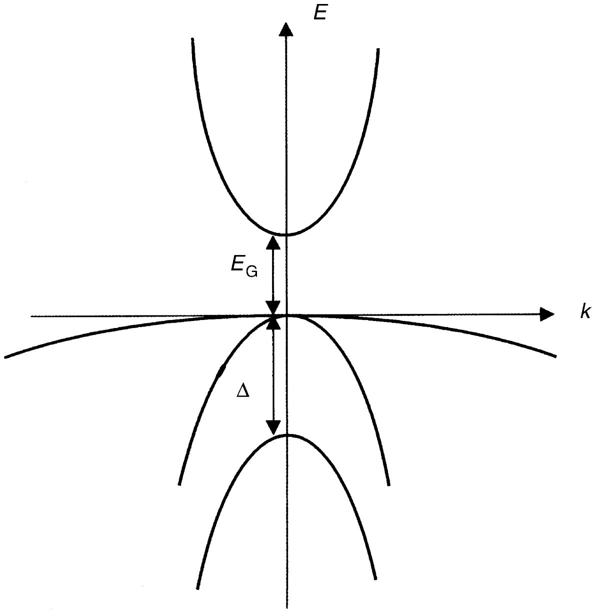

Figure 2 Schematic diagram of the conduction and valence band structure of InSb showing the energy bands around the Γ-point ($k = 0$) that are involved in Auger transitions.

With certain approximations, some analytic solutions for the Auger recombination rates are possible. This model is, while not appropriate for higher levels of band structure engineering, very important for giving an insight into how Auger effects should be studied, and how they themselves might be engineered. The approximations generally used are parabolic bands and Boltzmann statistics, but some further analytic results may also be obtained when these restrictions are relaxed. Bloch functions and Fermi's golden rule are utilized (*see* **Semiconductor Physics:** Recombination Processes) to describe the particle states, so obtaining the general expression for the Auger rate per unit volume. Energy and momentum conservation provide strong constraints on the resulting integral, and give rise to a considerable reduction in the Auger transition rate. However if, for example, the momentum conservation constraint is relaxed due to occupied levels existing with a broad range of momenta, such as at some band extremum with a

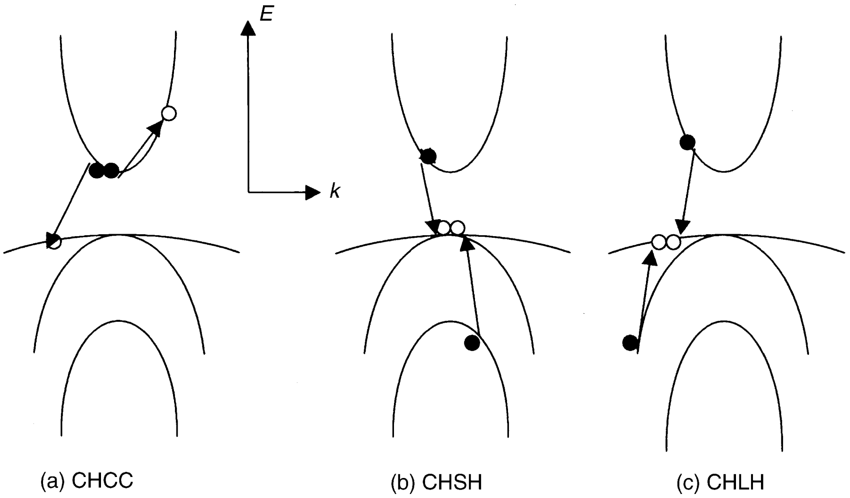

Figure 3 Important Auger recombination mechanisms for mid-infrared (MIR) semiconductor detectors and lasers: (a) CHCC or Auger-1; (b) CHSH; (c) CHLH or Auger-7.

large effective mass, then the Auger rate will be strongly enhanced. Or to put it the other way round, the Auger rate can be reduced by engineering the band structures such that the heavy hole effective mass is reduced (e.g., by stress – see below) to approach that of the conduction band. Furthermore, since we are concerned with nonradiative recombination across the bandgap with energy E_g, any energy levels (again band extrema in particular) separated from the conduction or valence band edges by E_g are particularly susceptible to the excitation part of the Auger process. The Auger rate may therefore be reduced by designing the structure so that there are no band extrema at energies E_g from the conduction and valence band edges.

Interband (Bipolar) Lasers

The technology of band structure engineering, which includes superlattices, quantum confinement, and intentional incorporation of strain, has now become sufficiently advanced that it can be employed in a routine manner. As originally proposed, the modified band structure of strained III–V quantum-well structures can lead to significant benefits for diode laser performance, including reduced threshold current density, improved efficiency, and enhanced dynamic response and speed. Predicted advantages have been demonstrated, although for near-infrared devices these have not been quite as substantial as originally hoped because of the presence of phonon-assisted Auger processes which are not so susceptible to band structure engineering techniques. Strained layer lasers are now commercially available at a variety of MIR wavelengths, where the dominant nonradiative loss is via the direct interband Auger scattering routes.

From the point of view of the present article the main contribution of 'designed' strain splittings and quantum confinement is in the reduction of the Auger coefficient itself, by removing states that easily satisfy the momentum and energy conservation conditions required for the process to take place. However, in addition to this there can be a substantial reduction in the threshold carrier injection level required for population inversion in a laser. In real III–V semiconductors the heavy hole valence band mass, m_h, is so much greater than the electron mass, m_e, that the usual laser picture of degenerate electron and hole distributions does not apply. At normal carrier injection levels the quasi-Fermi level of the holes is above the top of the valence band resulting in a classical heavy hole distribution. Thus, the lower laser levels are almost completely filled with electrons, by contrast with the ideal situation of two mirror bands. This leads to an injection level for the laser threshold of greater than 2:1 required for the actual case compared to the ideal case. Since the Auger recombination rate per unit volume is approximately proportional to the cube of n, this factor of ~ 2 reduction in injection level can produce a factor of almost an order of magnitude reduction in this loss mechanism. Large biaxial strains, either compressive or tensile, can lead to a considerable reduction in the hole mass and hence in the density of states at the top of the valence band, so that the laser characteristics may then approach those of an ideal semiconductor, with $m_e = m_h$.

A further important loss mechanism is free hole absorption of the heavy holes resulting from intervalence band transitions. This can be effectively eliminated, and other improvements made to laser performance, by the introduction of a strained layer structure which raises the degeneracy of the light and heavy hole at $k = 0$ and reduces the mass of the heavy hole. One example of this comes just from the symmetry of the problem, where clearly the cubic symmetry of the p-like valence band is not matched to the uniaxial symmetry of the laser beam. Application of biaxial or uniaxial strain removes the cubic symmetry, changing the relative energies of the p_x, p_y, and p_z valence states, and making it possible for almost all of the injected holes to be in the states of the correct symmetry to take part in the lasing action. The one parasitic phenomenon in semiconductor lasers which is hardly affected by band structure engineering is spontaneous emission, which will always make some contribution to the current density required at threshold.

The band structure of a bulk unstrained direct gap tetrahedral semiconductor is shown again for comparison in Figure 2. The effect of biaxial tension and compression is shown respectively in Figure 4a,b. In each case the lowest conduction band is approximately parabolic near the zone center, but becomes anisotropic with the electron dispersion at small k. The axial strain breaks the cubic symmetry of the semiconductor, introducing a tetragonal distortion which splits the degeneracy of the light and heavy hole states at the valence maximum, Γ, typically by about 60–80 meV for a 1% lattice mismatch. The resulting valence band structure is highly anisotropic (Figure 4a,b), with the band which is heavy along the strain axis, wavevector $k_\perp$, being comparatively light in the plane, along $k_\parallel$, and vice versa. The in-plane mass, $m_\parallel$, determines the density of states, and thus, under compression (Figure 4b), can be reduced according to the principles described above to improve conditions towards those for ideal laser operation: viz, (1) that there is only one band at the valence band maximum; (2) that it has an effective mass as close to that of the electron as possible; and (3) that above the Bernard–Duraffourg transparency condition, the polarization of the gain is anisotropic, with spontaneous emission and gain being suppressed along all directions except along the axis contributing to the laser beam.

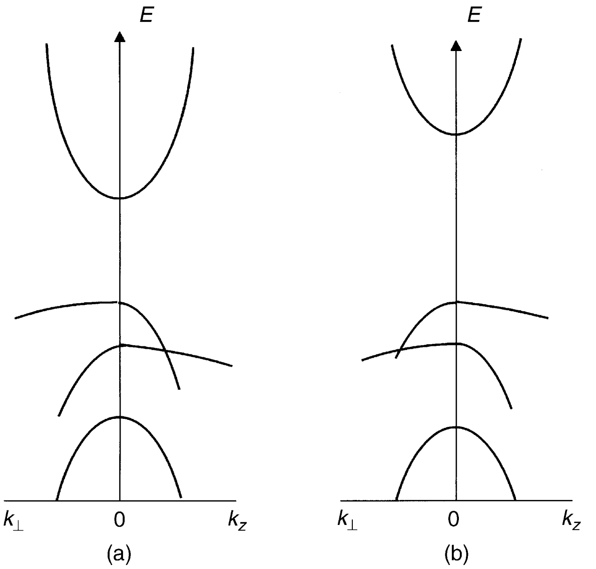

(a) (b)

Figure 4 Schematic representation of the band structure of a direct gap tetrahedral semiconductor: (a) under biaxial tension the hydrostatic component of the tension reduces the mean bandgap, while the axial component splits the degeneracy of the valence band maximum and introduces an anisotropic valence band structure (the highest band being light along the strain axis, k_z and comparatively heavy perpendicular to that axis, $k_\perp$); (b) under biaxial compression the mean bandgap increases and the valence splitting is reversed.

Since the Auger recombination involves three carriers, the Auger current J_{NR} in a device with an undoped active region varies approximately as $J_{\mathrm{NR}}(T) \propto C(T)n_{\mathrm{th}}^3$, where $C(T)$ is the temperature-dependent Auger coefficient per unit volume and nth is the threshold carrier density. The influence of strain on Auger recombination has two aspects, as discussed above. First, J_{NR} is very sensitive to any reduction in n_{th} brought about by strain either through a decrease in h (heavy hole) mass or an increase in the optical transition strength. Secondly, strain may change the magnitude of $C(T)$ for the interband Auger processes described earlier (Figure 3). A simple calculation of Auger recombination in a quantum well heterostructure, assuming parabolic bands and Boltzmann statistics, shows that the coefficient increases exponentially with temperature as $C(T) = C_o\exp(-E_a/kT)$, where the activation energy, E_a, is approximately dependent on $(m_h)^{-1}$. Thus, both of these should be reduced by several orders of magnitude due to a reduction in the heavy hole mass, m_h, resulting from the strain effects discussed above.

Phonon Scattering: QWIPs and QC Lasers

Sub-band transitions are useful for MIR detectors and emitters because they make it possible to cover a broad range of MIR wavelengths using the same combination of large gap, InP-based and GaAs-based state of the art semiconductor alloys that are used for high-speed electronics and optical communications. They are the most advanced of the low-dimensional solids from the point of view of demonstrated detector performance. This superior control of materials results in high uniformity and has thus allowed the demonstration of large 128 × 128 arrays with excellent imaging performance. Detectivities $D^* > 10^{10}\ \mathrm{cm\,Hz^{1/2}\,W^{-1}}$, low noise equivalent temperature differences, and excellent infrared video imaging can be achieved at 10 μm for temperatures $T = 10-80$ K.

Although the QWIP (quantum well infrared photodetectors) detectors showed promise for many years as alternatives to HgCdTe systems, the first intersub-band laser was not realized until the demonstration of the quantum cascade (QC) laser. Intersub-band (QC) lasers differ fundamentally from bipolar diode lasers in that they rely on only one type of carrier and on electronic transitions between conduction sub-band states arising from size quantization in semiconductor quantum heterostructures. These in-plane sub-bands have the same curvature (by contrast with the interband devices which have conduction and valence bands of opposite curvature), so that the joint density of states and corresponding gain spectrum (collision limited) is much narrower and approximately symmetric. Thus, the arguments which show that the Kane-type band structure is particularly disadvantageous for the overall Auger loss mechanism in interband diode lasers do not apply; the problems resulting from having $m_\mathrm{e} \ll m_\mathrm{h}$ are removed and the inversion is much less sensitive to thermal broadening of the electron distribution. In addition, there is no free hole absorption, and, since the component semiconductor materials are of comparatively wide bandgap, there is negligible intrinsic carrier concentration. However, the threshold current density tends to be rather large owing to the rapid nonradiative phonon relaxation of the population inversion.

In the original QC laser structure the electrons streamed down a potential staircase, sequentially emitting photons at the steps, consisting of coupled quantum wells in which the population inversion between discrete conduction band excited states was achieved by control of tunneling. The laser structure was grown with the $Al_{0.48}In_{0.52}As/Ga_{0.47}In_{0.53}As$ heterojunction material system lattice matched to InP. Electrical pumping was achieved by alternating 25 undoped coupled-well active regions with compositionally graded layers. The coupled-well region was essentially a four-level laser system, where a population inversion was achieved between the two excited states, $n = 3$ and $n = 2$. While the nonradiative phonon emission lifetime is short ($\tau_{32} \sim$ 4.3 ps at the electric fields used in the tunneling process), strong coupling to an adjacent GaInAs well gives an even shorter lower state lifetime ($\tau_{21} \sim 0.6$ ps) from optical phonons with nearly zero momentum transfer between the strongly overlapped and closely spaced $n = 2$ and $n = 1$ states. This enables the achievement of a population inversion between these states. (Note that the separation between the $n = 2$ and $n = 1$ states, E_{21}, is designed to be approximately equal to the optical phonon energy.)

The radiative efficiency of these QC devices is unavoidably low ($<10^{-3}$) because most of the electrons in the upper sub-band relax within ~1 ps through nonradiative optical phonon emission before radiative transitions with a lifetime in the nanosecond range have a chance to occur. An alternative configuration has been proposed, the type II interband cascade laser (T2ICL), which retains the advantages of cascaded tunneling injection and wavelength tuneability but eliminates the nonradiative phonon path between valence and conduction bands of opposite curvature. In addition, band structure engineering ideas discussed above are incorporated to suppress the unwanted Auger processes; the design results in a light in-plane hole mass and the removal of resonances between the energy gap and intervalence transitions.

See also

Semiconductor Materials: Group IV Semiconductors, Si/SiGe; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Band Structure and Optical Properties; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Faist J, Capasso F, Sirtori C, Sivco DL, Hutchinson AL and Cho AL (1994) Quantum cascade laser. *Science* 264: 553–556.

O'Reilly EP and Adams A (1994) Band structure engineering in strained semiconductor lasers. *IEEE Journal of Quantum Electronics* 30: 366–379.

Pidgeon CR, Ciesla CM and Murdin B (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Yablonovitch E and Kane EO (1988) Band structure engineering of semiconductor lasers for optical communications. *Journal of Lightwave Technology* 6: 1292–1299.

Youngdale ER, Meyer JR, Hoffman CA, Bartoli FJ, Grein CH, Young PM, Ehrenreich H, Miles RH and Chow DH (1994) Auger lifetime enhancement in InAs-GaInSb superlattices. *Applied Physics Letter* 64: 3160–3162.

Dilute Magnetic Semiconductors

R R Gałązka, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction, General Remarks

Semimagnetic semiconductors (SMSs) are a group of solids at the interface between semiconductors and magnetic materials. Semimagnetic semiconductors, also referred to as dilute magnetic semiconductors, are semiconductor-based solid solutions where some of the cations are replaced by transition metals or rare earth elements. The crystallographic structure of the semiconductor is conserved; the lattice constant is a function of composition. From a magnetic point of view SMSs are disordered magnetic materials, since magnetic atoms are randomly distributed in the cation sublattice of the semiconductor compound. Table 1 presents the most complete list of bulk ternary SMSs. Quaternary SMSs are also investigated (e.g., PbSnMnTe or CdMnSeTe).

Generally speaking, in SMSs there coexist two interrelated and interacting subsystems: mobile delocalized charge carriers and localized magnetic moments connected with paramagnetic ions. Electronic (mainly optical) properties of SMSs have been the subject of intensive studies since early in 1970. Due to the strong spin exchange interaction between mobile carriers and localized magnetic moments (the exchange constant of the sp-d interaction is of the order of 1 eV for II–VI SMSs) significant changes of the band structure and behavior of the carriers were observed. A number of new physical phenomena were discovered, such as giant Faraday rotation (now important in magnetic field sensors and other applications in, for example, the automobile industry based on this strong magneto-optical effect), the magnetic field induced metal–insulator transition, the bound magnetic polaron and quantum thermomagnetic oscillations. Investigations of the influence of the electron subsystem on the magnetic properties of SMSs are connected with the observation of a ferromagnetic phase in PbSnMnTe induced by carrier concentration.

Molecular beam epitaxy (MBE) grown GaAs with Mn opened a new chapter of III–V SMSs showing ferromagnetic behavior with the highest transition temperature $T_c = 110\,\mathrm{K}$. In 1987 the first paper devoted to layered structure and magnetic properties of low-dimensional (LD) SMSs was published starting intensive studies of superlattices and other LD structures made of SMSs.

Energy Band Structure, Optical and Magneto-optical Properties of SMSs

Replacing cations such as Cd or Hg with a paramagnetic ion such as Mn in the same crystallographic structure does not markedly disturb the semiconductor properties of the material. The energy gap changes (increases) but the conduction and valence bands conserve their symmetry and character as in nonmagnetic semiconductor mixed crystals. The spin momentum of paramagnetic ions is connected with the 3d or 4f shell, for transition metals or rare earth elements, respectively. The energy level of 3d or 4f electrons lies below the top of the valence band and thus has negligible influence on its shape or that of the bottom of the conduction band. Thus all basic semiconductor optical properties connected with the band symmetry and topology are the same as for a typical semiconductor. These are described, for example, in articles on the lead salts (MS641) involving direct optical transitions at the L-point of the Brillouin zone, and on mercury cadmium telluride (MS639) with direct optical transitions at the Γ-point. The only influence of the magnetic subsystems on electrons comes from the spin exchange interaction between mobile carriers and localized magnetic ions. This interaction can be represented by a Heisenberg term

$$H_{\mathrm{ex}} = \bar{\sigma} \sum_i \bar{S}_i J(\mathbf{r} - \mathbf{R}_i) \qquad [1]$$

where $\bar{\sigma}$ and $\bar{S}$ are spin operators of the band electron and magnetic ion, respectively, the summation is over all lattice sites occupied by the magnetic ions, and J is the exchange constant.

Assuming mean field and virtual crystal approximations this term can be rewritten in the form which is periodic with a lattice constant:

$$H_{\mathrm{ex}} = x \langle \bar{S} \rangle_{\mathrm{av}} \bar{\sigma} \sum_R J(\mathbf{r} - \mathbf{R}) \qquad [2]$$

where x is the molar fraction of magnetic ions, $\langle \bar{S} \rangle_{\mathrm{av}}$ is the average over all magnetic ions and is directly

Table 1 Composition and crystal structure of ternary bulk SMSs

Material	*Crystal structure*	*Composition range*
II–VI (Mn)		
$Zn_{1-x}Mn_xS$	ZB	$0 < x \leq 0.10$
	W	$0.10 < x \leq 0.45$
$Zn_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.30$
	W	$0.30 < x \leq 0.57$
$Zn_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.86$
$Cd_{1-x}Mn_xS$	W	$0 < x \leq 0.45$
$Cd_{1-x}Mn_xSe$	W	$0 < x \leq 0.50$
$Cd_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.77$
$Hg_{1-x}Mn_xS$	ZB	$0 < x \leq 0.37$
$Hg_{1-x}Mn_xSe$	ZB	$0 < x \leq 0.38$
$Hg_{1-x}Mn_xTe$	ZB	$0 < x \leq 0.35$
II–V (Mn)		
$(Cd_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.18$
$(Zn_{1-x}Mn_x)_3As_2$	tetr.	$0 < x \leq 0.15$
IV–VI (Mn, Eu, Gd)		
$Pb_{1-x}Mn_xS$	RS	$0 < x \leq 0.05$
$Pb_{1-x}Mn_xSe$	RS	$0 < x \leq 0.17$
$Pb_{1-x}Mn_xTe$	RS	$0 < x \leq 0.12$
$Sn_{1-x}Mn_xTe$	RS	$0 < x \leq 0.40$
$Ge_{1-x}Mn_xTe$	RH	$0 < x \leq 0.18$
	RS	$0.18 < x \leq 0.50$
$Pb_{1-x}Eu_xSe$	RS	$0 < x \leq 0.04$*
$Pb_{1-x}Eu_xTe$	RS	$0 < x \leq 0.32$*
$Sn_{1-x}Eu_xTe$	RS	$0 < x \leq 0.013$
$Pb_{1-x}Gd_xTe$	RS	$0 < x \leq 0.11$*
$Sn_{1-x}Gd_xTe$	RS	$0 < x \leq 0.09$*
II–VI (Fe, Co, Cr)		
$Zn_{1-x}Fe_xS$	ZB	$0 < x \leq 0.26$*
$Zn_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.22$*
$Zn_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.01$*
$Cd_{1-x}Fe_xS$	W	$0 < x < 0.09$*
$Cd_{1-x}Fe_xSe$	ZB	$0 < x < 0.20$
$Cd_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.06$*
$Hg_{1-x}Fe_xS$	ZB	$0 < x \leq 0.037$*
$Hg_{1-x}Fe_xSe$	ZB	$0 < x \leq 0.20$*
$Hg_{1-x}Fe_xTe$	ZB	$0 < x \leq 0.02$*
$Zn_{1-x}Co_xS$	ZB	$0 < x \leq 0.15$*
$Zn_{1-x}Co_xSe$	ZB	$0 < x \leq 0.10$*
$Zn_{1-x}Co_xTe$	ZB	$0 < x < 0.06$*
$Cd_{1-x}Co_xS$	W	$0 < x \leq 0.064$*
$Cd_{1-x}Co_xSe$	W	$0 < x \leq 0.082$*
$Cd_{1-x}Co_xTe$	ZB	$0 < x \leq 0.04$
$Hg_{1-x}Co_xS$	ZB	$0 < x \leq 0.02$
$Hg_{1-x}Co_xSe$	ZB	$0 < x \leq 0.047$*
$Zn_{1-x}Cr_xSe$	ZB	$0 < x \leq 0.005$*
$Zn_{1-x}Cr_xTe$	ZB	$0 < x \leq 0.003$*
$Cd_{1-x}Cr_xS$	W	$0 < x \leq 0.002$*

ZB – zinc blende, W – wurtzite, RS – rock salt, RH – rombohedral. Asterisks denote SMSs with the highest reported *x* value; tetr, tetragonal structure.

Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. In: Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

related to the magnetization of the sample:

$$M = N_0 g \mu \langle \bar{S} \rangle_{\mathrm{av}} \qquad [3]$$

N_0 is the number of unit cells per unit volume, g is the Lande factor for magnetic ions, and μ is the Bohr magneton.

The Heisenberg term must be added to the effective-mass Hamiltonian to solve the energy eigenvalue problem for SMSs. Realizing this was the turning point in understanding their properties and provided a basis for this group of alloys to be distinguished from other semiconductor mixed crystals. Since that time SMSs have become the subject of intensive studies in Europe and elsewhere.

The exchange interaction depends on the ion spin and the exchange constant. The spin of Mn is the highest for transition metals and equals five Bohr magnetons, and whilst a typical value for the exchange coupling between magnetic ions $J_{\mathrm{d-d}}$ is 10^{-3} eV, the exchange constant for electrons and holes with magnetic ions $J_{\mathrm{sp-d}}$ is about 1 eV in SMSs.

The exchange interaction is also strongly temperature and magnetic field dependent. As the macroscopic magnetization of a sample is proportional to the thermodynamic average value of the magnetic ion spin, to a good approximation we can replace the ion spin operators in the Hamiltonian by their average value, calculated or taken from measurements of the magnetization. From typical magnetic functions we can thus obtain information on the electron behavior.

In semiconductors, an external magnetic field acts on both the orbital motion and the spin of electrons producing Landau quantization and spin splitting, to an extent dependent on the effective mass of the carriers. The exchange interaction acts on their spin only and is, in effect, mass independent. Because of this, and also the large value of $J_{\mathrm{sp-d}}$, the band structure in turn changes drastically under the influence of an external magnetic field and depends strongly on temperature. From Figure 1 we can see that the very typical structure of the degeneracy of the light and heavy hole valence bands disappears in SMS when a magnetic field is applied. The change is accompanied by a drastic increase of hole mobility like that observed in low-dimensional structures. In SMS such as $Hg_{1-x}Mn_xTe$ or $Hg_{1-x}Mn_xSe$ (for $x \leq 0.07$) (Figure 2), the magnetic field produces an overlap of the valence and conduction bands, an effect never observed in nonmagnetic semiconductors.

Impurity levels are also influenced by exchange interactions: the ionization energy of an acceptor

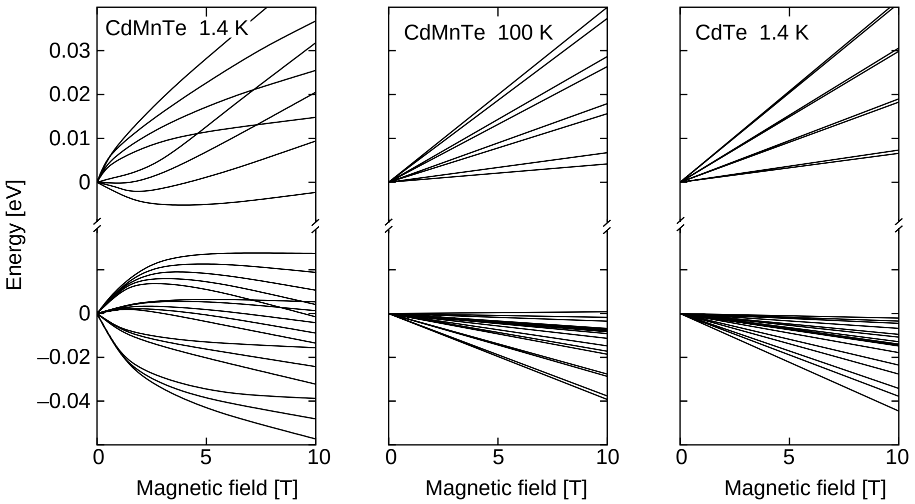

Figure 1 Conduction and valence band quantization in an external magnetic field for $Cd_{0.95}Mn_{0.05}Te$ and CdTe. Spin splitting, which is very weak in CdTe, is the main effect in CdMnTe. Notice the almost equal splitting of the heavy- and light-hole bands in CdMnTe. The influence of exchange interaction is visible even at 100 K. The picture of CdTe over this range of temperature is practically unchanged. Reproduced from Gałązka RR (1987) Semimagnetic semiconductors. *Europhysics News* 18(6): 90–92.

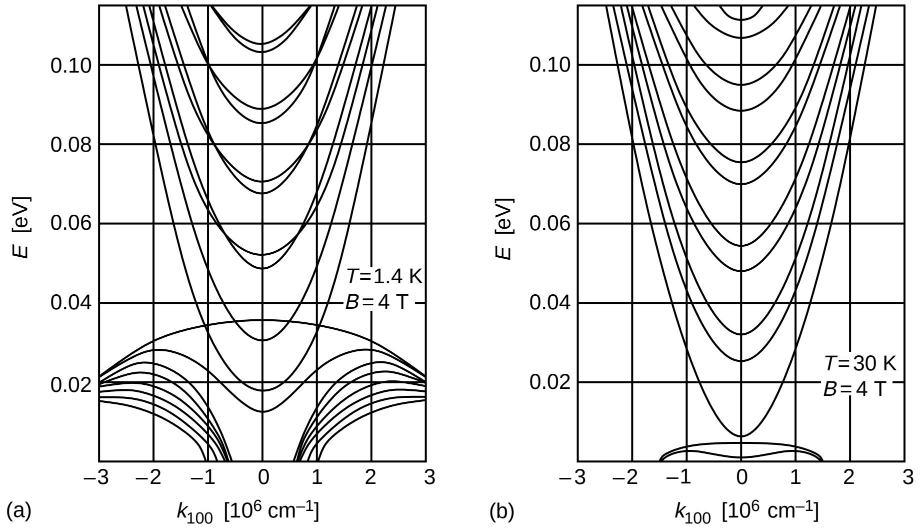

Figure 2 Energy band structure of $Hg_{1-x}Mn_xTe$, $x = 0.02$, calculated at 14 K (a) and 30 K (b) in a magnetic field of 4 T. After Gałązka RR (1978) Semimagnetic semiconductors based on HgMnTe and CdMnTe. Reproduced from Wilson BLM (ed) ICPS 14th, Edinburgh Institute of Physics, Conference series no. 43, 1978, p.133–140.

decreases and its wavefunction becomes anisotropic under the influence of a magnetic field. Both effects produce a giant negative magnetoresistance and a field-induced insulator–metal transition. Indeed all (particularly optical) properties susceptible to being changed by a magnetic field are very different in SMSs from those of standard semiconductors: the Faraday effect., Shubnikov–de Haas oscillations, interband magnetoabsorption, luminescence and spin flip Raman scattering (see below) become strongly dependent on temperature and magnetic field.

In the absence of a magnetic field, the average value of the ion spins is zero, the magnetization is zero too, and SMSs should behave as typical non-magnetic semiconductors. Whereas this is true for delocalized band electrons, when an electron is localized around an impurity, a bound magnetic polaron can be observed. This concept was first introduced to explain the behavior of magnetic semiconductors (Eu chalcogenides) and then later was applied also to SMSs. A localized electron can produce a spontaneous magnetization within the range of its wavefunction via the exchange interaction. Local ferromagnetic ordering in this range produces a Stokes shift in spin-flip Raman scattering even in the absence of an external magnetic field and other related phenomena. Local fluctuations of magnetization and a detailed knowledge of electronic states are essential for a theoretical description of this effect. The bound magnetic polaron is a subtle example of the feedback between electronic and magnetic subsystems present in SMS.

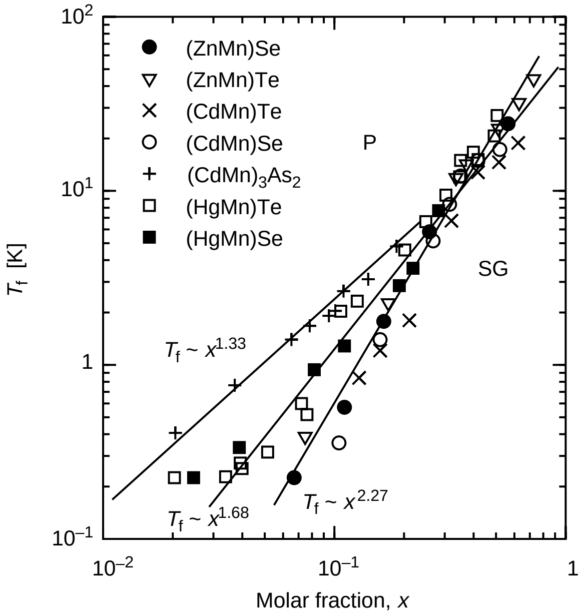

Figure 3 Magnetic phase diagram for the listed SMSs. P indicates the paramagnetic region, SG indicates the spin-glass region. Reprinted in part from Gałązka RR (1995) Influence of electron subsystem on magnetic properties of semimagnetic semiconductors. Reproduced with permission from Tróc R, Morkowski J and Szymczak M (eds) Proceedings of the International Conference of Magnetism, Warsaw 1994. *Journal of Magnetism and Magnetic Materials*, 113–116. Copyright Elsevier Science.

Magnetic Properties

Magnetic properties of SMS connected with the exchange interaction between paramagnetic ions are similar to those of other dilute magnetic materials. The broad range of possible magnetic ion concentrations (up to $x = 0.80$ in $Zn_{1-x}Mn_xTe$ and $x = 0.70$ for $Cd_{1-x}Mn_xTe$) allows us to observe the evolution of magnetic interactions in the same crystallographic structure as a function of magnetic ion concentration. A continuous transition from diamagnetic behavior (for $x \leq 0.005$) of the host semiconductor, to paramagnetism, spin-glass (for $x \geq 0.02$) and finally antiferromagnetic order has been observed. Type III antiferromagnetic ordering is observed for $x > 0.6$ indicating that the magnetic elementary cell contains two elementary cells of the cation sublattice. From a crystallographic point of view the magnetic subsystem is disordered; the distribution of magnetic atoms on the cation sublattice is random.

A spin-glass-like state is observed in SMSs in the broadest range of paramagnetic ion concentrations ($0.02 \leq x \leq 0.60$) below and above the percolation threshold (Figure 3). Indeed the temperature dependence of the magnetic susceptibility below and above the transition temperature and their dynamic properties at low and high magnetic fields suggest rather the existence of two spin-glass phases arising from the competition between spin-glass and clustering behavior. Below the percolation threshold ($x \leq 0.17$ for fcc lattice) the mechanism responsible for a spin-glass phase is even less clear, although, for such a dilute system it must be long-range to produce freezing of the spins.

The antiferromagnetic order has been inferred from specific heat and magnetic susceptibility measurements, and additional information concerning this phase has come from neutron diffraction studies: only a certain fraction of the total number of magnetic ions (e.g., about 50% for CdMnTe, $x = 0.65$) is well ordered; the rest remain in a disordered spin-glass phase. Thus a mixed phase (antiferromagnetic and spin-glass coexisting together) rather than a truly antiferromagnetic phase is observed.

The influence of carrier concentration on magnetic properties has been extensively investigated in metallic alloys. Several IV–VI semiconductors like SnTe can exhibit quite high metallic-like carrier concentrations. In the quarternary alloy PbSnMnTe, the concentration of holes can be varied by annealing in the range 10^{20} to 5×10^{21} cm^{-3}. In $Pb_{0.25}Sn_{0.72}Mn_{0.03}Te$ an abrupt transition from the paramagnetic to ferromagnetic phase has been found for $p > 3 \times 10^{20}$ cm^{-3}. The Curie temperature is a function of hole concentration, but the Mn ions conserve their magnetic moment of five Bohr magnetons. The magnetic phase diagram for this first ferromagnetic SMS (Figure 4) is thus three dimensional (T,x,n) in contrast to the

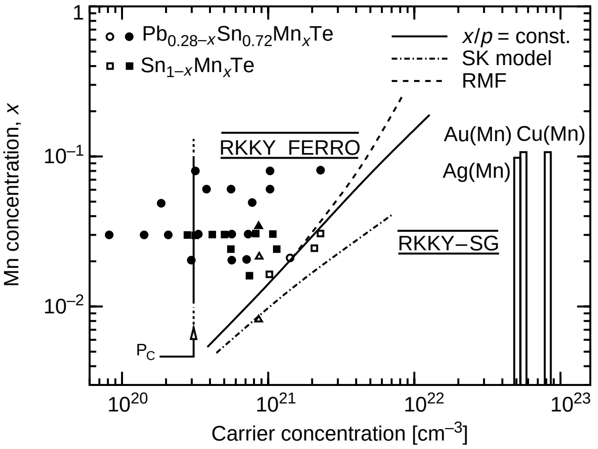

Figure 4 Magnetic phase diagram for PbSnMnTe (•) and SnMnTe (▪). The ferromagnetic phase is observed in the region above the critical carrier concentration p_c and the straight line defined by the equation x/p = const. Away from this region a spin-glass state is observed. Calculated phase boundaries between the ferromagnetic and spin-glass state are also shown for two theoretical models: SK: Sherrington–Kirkpatrick model and RMF: random mean field model. Canonical metallic spin-glasses are also shown. Reproduced from de Jonge WJM, Story T, Swagten HJM and Eggerkamp PJT (1992) Carrier-induced breakdown of ferromagnetism. *Europhysics Letters* 17: 631–636.

two-dimensional diagram (T,x) usually presented for magnetic alloys. Even a low concentration of spin polarized carriers when optically pumped can change the magnetization of HgMnTe owing to the strong exchange interaction of carriers with magnetic ions.

Low-Dimensional (LD) SMSs: Magneto-optical and Spintronic Effects

Modern technology such as MBE makes it possible to obtain quasi-3D SMSs in the whole range of x from 0 to 1. It also gives broad possibilities to create different combinations of semiconductor, SMS, and magnetic materials. In addition the dimensionality of LD structures influences the properties of both electronic and magnetic subsystems. Further there is almost always some strain in the system which also influences the properties of superlattices and other LD structures. Studies of LD SMSs not only extend the material base but open a new area of dimension-dependent phenomena and interactions in SMS systems.

Among many effects observed in LD SMS structures we will mention only a few phenomena, where exchange interaction is essential and which cannot be observed in nonmagnetic LD semiconductors. In all cases the giant (magneto-optical) Zeeman splitting caused by the exchange interaction plays a key role.

In superlattices made of ZnSe/ZnMnSe with a small amount of Mn, the band offset of conduction and valence bands is negligible at zero external magnetic field. In the presence of an external field the conduction and valence bands of ZnMnSe split creating a spin-superlattice because the spin-split states allow only one spin configuration. The possibility of the existence of such a spin superlattice was earlier suggested for SMSs and later experimentally observed.

If the superlattice, quantum well or heterostructure is made of SMS and non-SMS material (e.g. ZnTe/CdMnSe or CdTe/CdMnTe) one can exploit the drastic differences in the Zeeman splitting occurring in different layers to pinpoint the localization in space of specific electronic states. If one (hole or electron) or both states involved in optical transitions originate from the SMS layer, this state will show a very strong Zeeman splitting in the external magnetic field. Thus the SMS component can be used as a 'marker' in order to bring out the general properties of the wavefunction distribution.

Excitonic Zeeman splitting in non-magnetic quantum wells with SMS barriers can be applied to studies of the shape of interface profile and the influence of the growth conditions on the shape of the interface. If the interface is not completely flat, even a small fraction of magnetic atoms in the well close to the interface can significantly enhance the Zeeman exciton splitting in the quantum well. The sensitivity of this method allows one to detect the sharpness of an interface with one monolayer accuracy.

The MBE technique opens up possibilities of creating hybrid structures with SMSs, useful for spintronic devices such as spin filters, spin transistors and tunneling magnetoresistance sensors. Very importantly, it has been realized that SMSs are highly suitable to form the ferromagnetic spin injecting electrodes in such systems where simple metallic electrodes fail because of the large electrical resistance mismatch between the electrode and the semiconductor in which the spin is being processed. A further possibility is to fabricate hybrid metallic ferromagnetic–SMS structures. In such a case one can make use of the magnetic field exerted by the metallic ferromagnet positioned very close (on an atomic scale) to the SMS layer or quantum structure, thus affecting the spin splitting in the latter. One could use such a fringe field to shape the band edges in such a fashion that an additional localization of the carriers in a two-dimensional quantum well is induced leading to formation of quantum wires or dots. In other applications a spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. The optically pumped conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins(3d in the case of Mn) through the core–conduction-electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization has decayed away and presents an interesting avenue towards information storage.

The physics of SMSs is currently a well-established part of solid state physics. In the investigations of bulk crystals and LD structures many problems are solved and new problems are waiting further research. Particularly important seems to be investigation of layers and LD structures because of the broad technological possibilities and because of the new important factor (i.e., dimensionality) whereby spin-dependent phenomena can be intentionally regulated.

List of Units and Nomenclature

g	Lande factor for paramagnetic ion
H_{ex}	Heisenberg spin exchange Hamiltonian
$J_{\mathrm{d-d}}$	exchange constant for coupling between paramagnetic ions (transition metals)
$J(\bar{r} - \bar{R})$	exchange constant
$J_{\mathrm{sp-d}}$	exchange constant for electron or hole coupling with paramagnetic ions
LD	low dimensional
M	magnetization
MBE	molecular beam epitaxy
N_0	number of unit cells per unit volume
SMS	semimagnetic (diluted magnetic) semiconductor
$\langle \bar{S} \rangle_{\mathrm{av}}$	average value of spin operator
T_{c}	Curie temperature
V_{h}	hybridization energy
x	molar fraction of paramagnetic ions
$\boldsymbol{\mu}$	Bohr magneton
$\bar{\boldsymbol{\sigma}}, \bar{S}$	spin operators

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Lead Salts; Mercury Cadmium Telluride. **Semiconductor Physics:** Spin Transport and Relaxation in Semiconductors; Spintronics.

Further Reading

Bauer G, Pascher H and Zawadzki W (1992) Magneto-optical properties of semimagnetic lead chalcogenides. *Semiconductor Science and Technology* 7: 703–723.

Dietl T (1994) Magnetic semiconductors. Moss TS (ed.) *Handbook of Semiconductors*, vol. 3b, pp. 1251–1342. Amsterdam, The Netherland: North-Holland.

Dietl T (2002) Ferromagnetic semiconductors. *Semiconductor Science and Technology* 17: 377–392. (Special issue: Semiconductor Spintronics, guest ed Hideo Ohno).

Dobrowolski W, Kossut J and Story T (2003) Magnetic semiconductors–new bulk materials and low-dimensional quantum structures. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials, II–VI and IV–VI.* vol. 15, pp. 289–378. Amsterdam: Elsevier North-Holland.

Furdyna JK (1991) In: Averous M and Balkanski M (eds) *Semimagnetic Semiconductors and Diluted Magnetic Semiconductors*. New York: Plenum.

Furdyna JK and Kossut J (eds) (1988) Semiconductors and Semimetals. *Diluted Magnetic Semiconductors*, vol. 25, 470 pp. San Diego, CA: Academic Press.

Gałązka RR and Kossut J (1980) Narrow-gap semimagnetic semiconductors. In: Zawadzki W (ed.) *Lecture Notes in Physics*, vol. 133, pp. 245–265. Berlin: Springer Verlag.

Gałązka RR and Kossut J (1982) Semimagnetic semiconductors. In: Madelung O, Schulz M and Weiss H (eds) *Landolt-Börnstein New Series*, vol. III/17b, pp. 302–318. Berlin: Springer Verlag.

Gałązka RR, Kossut J and Story T (1999) Semimagnetic semiconductors. In: Rössler U (ed.) *Landolt-Börnstein New Series*, vol. III/41B, pp. 650–721. Berlin: Springer Verlag (Update to chapter 5, Vol. III/17b).

Kossut J and Dobrowolski W (1993) Diluted magnetic semiconductors. In: Buschow KHJ (ed.) *Handbook of Magnetic Materials*, vol. 7, pp. 231–305. Amsterdam, The Netherlands: North Holland.

Kossut J and Dobrowolski W (1997) Properties of diluted magnetic semiconductors. In: Capper P (ed.) *Narrow-Gap II–VI Compounds for Optoelectronics and Electromagnetic Applications*, pp. 401–429. London: Chapman and Hall.

Kossut J (2001) Low dimensional structure of diluted magnetic (semimagnetic) semiconductors – a subjective review. *Acta Physica Polonica, A* 100 (supplement): 111–138.

Ohno H, Matsukura F and Ohno Y (2002) Semiconductor spin electronics. *Japan Society of Applied Physics (JSAP International)* 5: 4–13.

Story T (2003) Optoelectronic properties. In: Monasreh MO (series ed.) *Optoelectronic Properties of Semiconductors and Superlattices* vol. 18, Khokholov D. (volume ed.) *Lead Chalcogenides: Physics and Applications, Semimagnetic Semiconductors Based on Lead Chalcogenides*, pp. 385–425. New York, London: Taylor and Francis Books Inc.

GaAs Based Compounds

J J Finley and J P R David, University of Sheffield, Sheffield, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Background

Gallium Arsenide (GaAs) is an important semiconductor that has come to dominate the field of optoelectronics by virtue of its favorable electro-optical properties and the ease by which it can be controllably modified by extrinsic means; combining it with its large family of related alloys ($Al_xGa_{1-x}As$, $In_xGa_{1-x}As$, $GaAs_xP_{(1-x)}$, $(In_xGa_{(1-x)})_yAsN_{1-y}$) and via the growth of hetero-structures with reduced dimensionality (Quantum Wells-2D, Wires-1D, and Dots-0D). In such low dimensional semiconductor nanostructures, quantum mechanical effects are directly exploited to customize and dramatically enhance electro-optical properties. Such investigations of the fundamental electro-optical properties of GaAs-based compounds has resulted in the emergence of several new device concepts such as the Quantum Cascade laser, led to the discovery of new physical phenomena like the fractional Quantum Hall effect, and may pave the way for an entirely new class of opto-electronic devices with true quantum functionality.

In the following, we explore the fundamental physical and electro-optical properties of GaAs and its related alloys and illustrate the considerable impact this remarkable family of materials has had on modern semiconductor devices.

Fundamental Physical Properties

The crystal structure of GaAs (Figure 1 – inset) consists of two interpenetrating face centred cubic lattices that accommodate the group III and V atoms respectively. The sub-lattices of Ga and As are shifted from each other by $a/4$ along the [111] body diagonal, where a is the lattice constant, forming the well-known zincblende crystal structure. For stochiometric GaAs, the lattice constant is $a = 0.56536(1)$ nm at room temperature which translates to a Ga–As nearest neighbor separation (bond length) of 0.2308 nm.

Figure 1 summarizes the variation of the lattice constant for a number of GaAs related III–V semiconductor alloys, plotted as a function of their fundamental electronic bandgap. Particularly relevant is AlAs, which also crystallizes in the zincblende form and has a lattice constant (a(AlAs) $= 0.5660(2)$ nm) almost identical to that of GaAs ($\Delta a < 1\%$). The sole fact that the $Al_xGa_{(1-x)}As$ alloy can be almost perfectly lattice matched to GaAs over its entire compositional range, whilst exhibiting near continuously variable electronic and optical properties, is the principal reason for the key role GaAs-related materials play in semiconductor opto-electronic devices. We continue by investigating the electronic and optical properties of the $Al_xGa_{(1-x)}As$ family of materials and discover why they are particularly suited to opto-electronic device applications.

Summary of Electronic Properties of GaAs and $Al_xGa_{1-x}As$

The optical properties of semiconductors are determined principally by inter-band radiative transitions involving electronic states in the vicinity of the conduction (CB) and valence band (VB) extrema. Consequently, in the following we focus on electronic properties of GaAs close to these points and substantiate our discussion using experimental data obtained from optical spectroscopy.

GaAs Bandstructure

GaAs is a direct-gap semiconductor ($E_g = 1.42$ eV at 300 K), both the conduction band minimum and

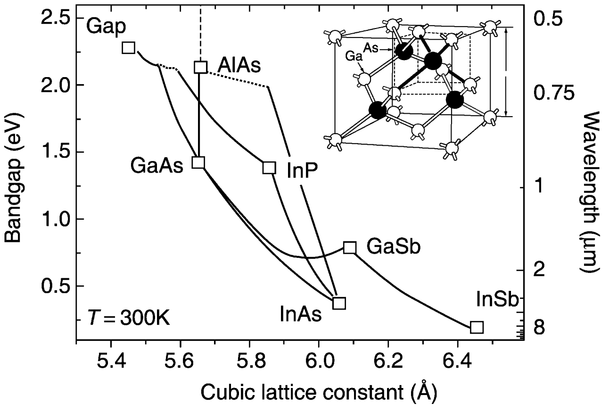

Figure 1 Fundamental bandgap of selected compounds within the III–V semiconductor family plotted as a function of the cubic lattice constant – a. The inset depicts the zincblende crystal structure within the GaAs unit cell. Data obtained from various sources, see for example Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16 and Loffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru

valence band maximum occurring at the center of the Brillouin zone (Γ-point). Figure 2 summarizes our discussion of the electronic structure of GaAs, depicting the bandstructure schematically within the first Brillouin zone. The valence electron configurations of Ga and As atoms result in the band edge CB states being two-fold degenerate due to the electron spin and characterized by an almost isotropic effective mass of $m_\Gamma^* = 0.068\,m_o$. For GaAs heterostructures with reduced dimensionality, the very weak anisotropy of the conduction band states can generally be ignored when calculating the lowest energy confined electron states but has a significant impact on higher quantized levels and must, therefore, be included. In addition to the fundamental CB minimum at the zone center or Γ-point of the Brillouin zone, satellite minima occur at the six equivalent X-points ($k = (1/a, 0, 0)$) and eight equivalent L-points ($k = (1/a, 1/a, 1/a)$). The X-point and L-point minima lie $\Delta E_{\Gamma X} = 0.49$ eV and $\Delta E_{\Gamma L} = 0.32$ eV above the Γ-minimum at low temperatures ($T = 4.2$ K) and are characterized by strongly direction dependent effective masses parallel (‖) or perpendicular (⊥) to the principal axes. Electronic states in the vicinity of these minima are termed the X and L-valleys respectively and have a significant impact on high-energy electron transport in devices such as HEMTs and in avalanche carrier multiplication processes, but do not generally play a significant role in the optical properties of the binary – GaAs. However, satellite CB minima can strongly influence the optical properties of GaAs–AlAs superlattices and QWs that can exhibit a type-II band alignment for GaAs layer thicknesses < 6 nm. In this case, the electrons (holes) are localized in the AlAs (GaAs) layer and the nature of the AlAs X-states dominates the optical properties.

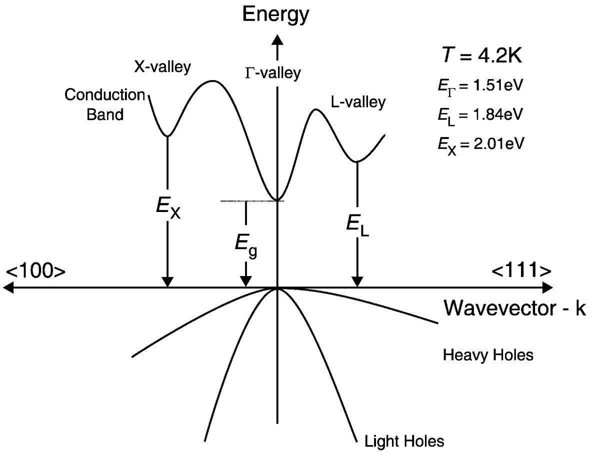

Figure 2 Schematic representation of the bandstructure of GaAs showing just a single conduction band over whole Brillouin zone in Γ-L and Γ-X directions only and neglecting the split off band that does not significantly influence electrical or optical properties.

Table 1 Low temperature GaAs effective masses for electrons (upper panel) and holes (lower panel). Values marked with * are inferred from measurements on GaAs based heterostructures and values marked ** are calculated

	Crystallographic axis	*Effective Mass (m_o)*
CB-Γ minimum	*Isotropic*	0.068(2)
CB-X minimum*	‖ [100]	1.9
	⊥ [100]	0.19
CB-L minimum**	‖ [111]	1.9
	⊥ [111]	0.075
VB-HH	[111]	0.72(3)
	[100]	0.33(6)
VB-LH	[111]	0.08(3)
	[100]	0.09(4)

The valence band (VB) states of GaAs are four-fold degenerate at the bandedge and, characteristic of most zincblende semiconductors the valence bandstructure, is significantly more strongly anisotropic than the CB states discussed above. Away from the Γ-point, the four-fold degeneracy of the valence band states is lifted into 2 doubly degenerate sets of bands termed the heavy (HH) and light hole (LH) bands by virtue of their zone center effective masses (Table 1). The HH band is considerably more strongly anisotropic than the LH band and is characterized by a much larger effective mass along the [100] direction when compared with [001], $m_h^*[111] > m_h^*[001]$. The anisotropy of the LH states is much weaker than for the HH band, with approximately equivalent effective masses along [100] and [111] crystallographic directions. Typical values for the CB and VB effective masses are summarized in Table 1, obtained using a variety of experimental techniques at low temperature.

Temperature Dependence

Quantitative values given above relate specifically to the bandstructure at liquid helium temperatures ($T = 4.2$ K). Increasing temperature results in a narrowing of many of the conduction–valence band gaps discussed above and lowering of the effective masses. This temperature dependence of the fundamental bandstructure arises from the anharmonic dilation of the lattice and from interactions with phonons, the population of which becomes significant at elevated temperature. In the simplest picture, one can visualize the effect of increasing temperature

Table 2 Varshnii coefficients for the Γ, L, and X CB minima in GaAs relative to the top of the VB. For the Γ minimum a number of experimental works are available and the value cited is the average

	$\alpha (10^{-4}$ eV/K)	β (K)
Γ point	5.45	210
X point	7.20	205
L point	6.05	204

as a shrinking of the first Brillouin zone as the lattice constant increases, and with a subsequent increase of the band curvature leading to ~5–10% lighter effective masses at room temperature than at 4.2 K. In addition, the dramatic increase of the phonon population at elevated temperatures results in a narrowing of the fundamental bandgaps. Phenomenologically, the temperature dependence of the fundamental bandgaps can be well described by the well-known Varshnii relation

$$E_g(T) = E_g^{T=0\text{ K}} - \frac{\alpha T^2}{(T+\beta)} \qquad [1]$$

where α and β are experimentally determined constants which are sensitive to the bandgap of interest (e.g. Γ, X, or L, etc.) as summarized in Table 2.

GaAs/Al$_x$Ga$_{(1-x)}$As – The Most Significant Alloy

Incorporation of aluminum into the GaAs crystal to form Al$_x$Ga$_{(1-x)}$As results in a number of profound and important modifications of the electronic properties compared with pure GaAs. Aluminum is more strongly electronegative than Gallium and consequently the electron charge distribution along the Al–As bonds are more strongly localized towards the group-III site than for Ga–As bonds which qualitatively results in a widening of the fundamental bandgap with increasing aluminum composition (x).

The variation of the Al$_x$Ga$_{(1-x)}$As energy gaps for the Γ, X, and L conduction band minima at low temperature are summarized in Figure 3 as a function of the alloy composition up to $x \sim 0.8$. The data of Figure 3 reflect the results of optical emission and absorption measurements and are most reliable for more dilute compositions since crystalline AlAs and high Al content Al$_x$Ga$_{(1-x)}$As oxidize rapidly under atmospheric conditions.[1] With increasing Al content, the Γ-minimum shifts rapidly to higher energy whilst the shift of the L and X-minima is significantly less pronounced. At $x \sim 0.4$ the Γ and X minima intersect and for $x > 0.4$, the lowest energy conduction band minimum shifts to the X-point and the fundamental bandgap becomes indirect. This effect is responsible for the short wavelength cutoff of Al$_x$Ga$_{(1-x)}$As/GaAs based multi-quantum well LEDs at $\lambda \sim 630$ nm since the transition to an indirect bandgap is accompanied by a drastic reduction of the radiative efficiency. Shorter wavelength LED operation region of the spectrum is possible by moving to other related alloys in the III–V family such as $(\text{Al}_x\text{Ga}_{1-x})_{0.5}\text{In}_{0.5}\text{P}$ (520–630 nm) and $(\text{GaAs})_{0.6}\text{P}_{0.4}$ (~600 nm). We return to discuss the optical properties of Al$_x$Ga$_{(1-x)}$As below in Section 4.

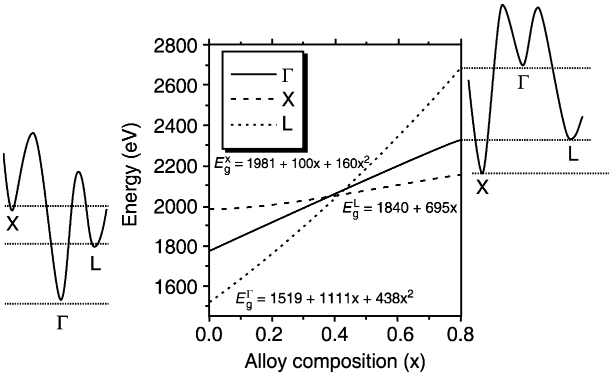

Figure 3 The absolute energy (at 4.2 K) of the Γ, X and L CB minima for Al$_x$Ga$_{(1-x)}$As as a function of aluminum composition up to $x \sim 0.8$. Higher compositions are not given due to the well-known unreliability of experimental data due to instability of high Al content Al$_x$Ga$_{(1-x)}$As due to oxidation. Polynominals describing each curve are also plotted. Data taken from the following sources: Γ-minimum, Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in Al$_x$Ga$_{1-x}$As. *Physical Review* 38: 3263–3268, X-minimum, Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in Al$_x$Ga$_{1-x}$As alloys. *Physical Review* B45: 10951, and L-minimum, Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in Al$_x$Ga$_{1-x}$As. *Journal of Physical Chemistry*, L335.

As expected, the pronounced modifications of the bandgap induced by Al incorporation are accompanied by modifications of the carrier effective masses and temperature dependence of the bandgap. These compositional dependence of the Γ, X, and L conduction band and HH and LH valence band effective masses are summarized in Table 3.

It should be stressed that the relationships given in Table 3 should be considered to be accurate at the

---

[1] The ease by which high Al content Al$_x$Ga$_{(1-x)}$As can be oxidized is of great technological benefit for device fabrication. Examples include the formation of current apertures in injection lasers, Al$_x$Ga$_{(1-x)}$As etch stops etc. For a summary Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Table 3 Compositional variation of the electron and hole effective masses for $Al_xGa_{(1-x)}As$ taken from various sources: Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E , Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society*, 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331

	Crystallographic Axis	*$Al_xGa_{(1-x)}As$ Effective Mass (m^0)*
CB-Γ minimum	*Isotropic*	$0.0682 + 0.0835x$
CB-X minimum	∥ [100]	$1.9 - 1.02x$
	⊥ [100]	$0.19 + 0.06x$
CB-L minimum	∥ [111]	$1.9 - 0.58x$
	⊥ [111]	$0.075 + 0.075x$
VB-HH	DOS	$0.62 + 0.14x$
VB-LH	DOS	$0.087 + 0.09x$

~5–10% level, since reliable experimental data only exists for the GaAs binary endpoint and more dilute alloy compositions ($x < 0.3$).

The Varshnii coefficients of the temperature dependence of the bandgap vary with $x (0 < x < 0.5)$ according to $\alpha(x) = 5.5 + 3.35x (10^{-4}$ eV/K) and $\beta(x) = 225 + 88x$(K) respectively. Recent photoreflectance measurements have indicated that these simple linear relationships do not hold well over the entire compositional range with evidence for some bowing of $\alpha(x)$ and $\beta(x)$ closer to the AlAs binary endpoint.

GaAs/$Al_xGa_{(1-x)}$As Hetero-interfaces

Modern epitaxial growth techniques now enable the routine fabrication of semiconductor heterostructures consisting of a well-defined sequence of atomically precise layers of different materials, which together manipulate the opto-electronic properties to suit a particular device or application. Of particular technological relevance is the GaAs/$Al_xGa_{(1-x)}$As hetero-interface since both materials are lattice matched for $0 < x < 1$ (Figure 1), and multiple layers of GaAs and $Al_xGa_{(1-x)}$As can be readily deposited without introducing defects into the crystal which seriously degrade optical performance. In this section, we summarize the most salient features of the GaAs/$Al_xGa_{(1-x)}$As hetero-interface, with particular emphasis on the electronic properties before continuing to discuss the optical properties of $Al_xGa_{(1-x)}$As in Section 4.

The most important parameters that characterize a semiconductor hetero-interface are the magnitude and sign of the conduction (ΔE_c) and valence band (ΔE_v) offsets. These parameters determine how effectively electrons and hole are localized within specific epilayers and are related to the total difference in the bandgap ($\Delta E_g = E_g^{AlGaAs} - E_g^{GaAs}$) via $\Delta E_g = |\Delta E_c + \Delta E_v|$. Taking the mean of 21 separate optical and electrical investigations between 1985 and 1995, the currently accepted value for the band-offset ratio is $\Delta E_c/\Delta E_v \approx 64/36$, with an associated variance over the body of the data of approximately ±8%. It should be stressed that the majority of the experimental investigations of the band offsets at the GaAs–$Al_xGa_{(1-x)}$As hetero-interface have been performed for alloy compositions $x < 0.4$ and much less experimental data exists regarding the variation of $\Delta E_c : \Delta E_v$ with higher Al composition. A rather old review (*c*. 1992), collating data available, at the time has indicated that the absolute energies of the conduction and valence band offsets vary with composition as $\Delta E_c = (800 \pm 30)x$ meV for ($x < 0.4$) and $\Delta E_c = (510 \pm 40)x \, (0 < x < 1)$. It is widely assumed that the band offsets do not change appreciably with temperature.

Optical Properties of GaAs and Significant Alloys

Optical Absorption

Monochromatic light of energy $E_p = \hbar\omega$ incident on the surface of GaAs may stimulate interband optical transitions in which a photon is absorbed and an electron from the valence band is promoted into the conduction band. In processes involving a single photon, the photon momentum is entirely negligible compared with the electron and optical transitions are vertical in k-space.

Following excitation, the electron–hole (e–h) pairs relax extremely rapidly ($\tau_{LO} \sim 150$ fs) by sequential emission of multiple longitudinal optical (LO) phonons until their energy becomes close to the bandedge and further LO-phonon emission is forbidden by energy conservation requirements. This initial phase of the carrier relaxation can be extremely rapid, the LO-phonon scattering time in GaAs being due to the semi-polar nature of the crystal. After this initial rapid relaxation phase, the electron and hole populations relax principally via acoustic phonon emission ($\Delta E < 2$ meV) over much longer timescales ($\sim 10-300$ ps), which depend upon the lattice temperature, carrier type and density. Finally, the electron and hole populations form thermal distributions via carrier–carrier and Auger scattering over longer timescales up ~ 500 ps. During these final relaxation and thermalization phases, e–h pairs are depleted from the conduction and valence bands by interband radiative recombination.

For energies less than the bandgap, the semiconductor normally appears transparent, although the presence of impurities and defect levels can result in a weak absorption 'tail' below the gap. Optical absorption by impurities results in only electrons in the conduction band or holes in the valence band. The wavelength of light that just causes the creation of an electron–hole pair is given by $\lambda = hc/E_g$, where h is Planck's constant, c the speed of light and E_g the fundamental bandgap. As described in Section 3.1, this corresponds to approximately $\lambda = 870$ nm at room temperature, depending on the type and level of the doping. Light with wavelength much shorter than this, corresponding to energies much larger than E_g, will create electron–hole pairs higher up the band structure with higher energy and be strongly absorbed. These carriers, however, will rapidly lose this excess energy by interacting with the lattice and relax to the minima of the conduction and valence bands.

The optical absorption is usually quantified by the absorption coefficient, α, which is a function of the photon energy or wavelength and is shown for GaAs in Figure 4. The fraction of the light intensity that is absorbed in the semiconductor after passing through a thickness x is given by $f = (1 - \exp(-\alpha x))$ and the fraction of light intensity that is transmitted is then simply $\exp(-\alpha x)$. α is usually expressed in units of cm^{-1} and at energies above the bandgap has values typically between 10^4–$10^5\ \text{cm}^{-1}$. Light with wavelength corresponding to energies above the bandgap is usually absorbed very efficiently (>90%) within a few microns. Increasing the bandgap, say by adding aluminum, would have the effect of making the material more transparent at a given wavelength.

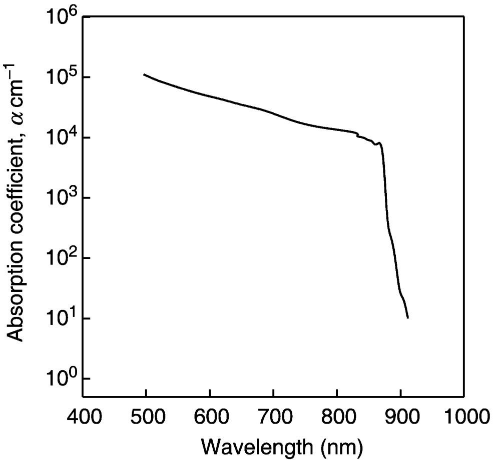

Figure 4 Variation of the absorption coefficient, α, as a function of wavelength in GaAs at room temperature.

Optical absorption is an important process as it allows light to be converted to free carriers, and hence a current in the semiconductor. This forms the basis of optical detection by most types of semiconductor photodetectors.

Refractive Index

The speed (more precisely the phase velocity) of light, of all wavelengths, in a vacuum is constant at $\sim 3.10^8\ \text{m s}^{-1}$. The speed of light in a medium such as a semiconductor is slower due to the microscopic interaction of the electromagnetic wave with the crystal. The refractive index, R_I, is defined as the ratio of the speed in a vacuum to the speed in the semiconductor. Increasing the density of the medium decreases the speed. Changing the semiconductor, for example by adding aluminum, not only increases the bandgap as discussed above, but also decreases R_I for a given wavelength.

At wavelengths corresponding to an energy just below the bandgap of the semiconductor, the material is essentially transparent, i.e., the absorption coefficient is low.

Light can be made to travel in a waveguide comprising this semiconductor, provided that it was surrounded by semiconductor material of a lower R_I. Light is totally internally refracted at the interface between the GaAs and $\text{Al}_x\text{Ga}_{1-x}\text{As}$ and so cannot escape. The semiconductors with a smaller R_I usually have a larger bandgap, so are even more transparent at the operating wavelength and help to physically confine the electrons and holes within the GaAs.

The R_I can be changed by injecting charge (electrons or holes) into the semiconductor, by changing the temperature (which changes the bandgap), and by applying an electric field. This forms the basis of operation of devices such as optical modulators. Unfortunately it also can cause 'chirping', i.e., the unwanted change in light intensity or wavelength as the charge or temperature changes.

The R_I which is a function of the wavelength, is sometimes expressed as a complex number:

$$R_I = n + ik \qquad [2]$$

where n is the real part and k is the imaginary part called the extinction coefficient, which is related to the absorption coefficient. Typical room temperature values of n for GaAs vary from about 3.37 at ~1.55 μm to 3.87 at 620 nm.

Common GaAs Based Alloys for Novel Technological Applications

GaInP/AlGaInP – 'Red, Orange and Green Emission'

The alloy $(Al_xGa_{1-x})_{0.52}In_{0.48}P$ is becoming technologically important as it has the same lattice parameter as GaAs for all values of x, so can grow lattice matched on GaAs. This alloy system offers the largest direct bandgap of virtually any III–V (excluding the Nitride family), so has attracted considerable interest for visible lasers and LEDs in the wavelength range 530 nm–670 nm. At room temperature the direct Γ bandgap increases almost linearly from ~1.85 eV for $Ga_{0.52}In_{0.48}P$ to ~2.3 for $(Al_{0.5}Ga_{0.5})_{0.52}In_{0.48}P$. For values of $x > 0.5$, the bandgap becomes indirect as in $Al_xGa_{1-x}As$ ($x > 0.4$) when the X-level indirect bandgap becomes lower than the Γ-level direct bandgap. Earlier work in this alloy suggested that the point at which it became indirect was at $x = 0.66$. The X-level bandgap appears to be relatively insensitive to the aluminum composition and has a value of ~2.3 eV at room temperature.

GaAs/GaInAsN – '1.3 μm and 1.55 μm for telecommunication applications'

One limitation of GaAs is that most of the compounds that are lattice matched to it, such as $Al_xGa_{1-x}As$ or $(Al_xGa_{1-x})_{0.52}In_{0.48}P$, have bandgap energies that are large, making them unsuitable for telecommunication applications which require narrow bandgap materials capable of operating in the range 1.3 μm to 1.55 μm. Adding indium (In) to GaAs can decrease the bandgap but also causes strain as the lattice parameter increases, which limits the amount of indium and total thickness of the structure.

Recently, it has been shown that while adding In to GaAs increases the lattice parameter, adding N to GaAs decreases it. Consequently GaInAsN can be made to be lattice matched to GaAs by controlling the In and N content. Unlike conventional alloy semiconductors, adding In or N both result in a reduction of the bandgap energy. Furthermore, GaInAsN is found to be a direct bandgap material with a type I band lineup when combined with $Al_xGa_{1-x}As$, which is a wide bandgap material also lattice matched to GaAs. These properties make GaInAsN ideally suited for lasers and there have been several reports of devices operating at 1.3 μm in the literature.

See also

Semiconductor Physics: Quantum Wells and GaAs-based Structures.

References

Bosio C, Staehli JL, Guzzi M, Burri G and Logan RA (1988) Direct-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$. *Physical Review* 38: 3263–3268.

Brozel MR and Stillman GE (eds) (1996) *Properties of Gallium Arsenide*, INSPEC, emis Datareviews Series No.16.

Data drawn from the following sources: Lautenschlager Γ-P, Garriga M, Logothetidis S and Cardona M (1997) *Physical Review* B35: 9174 and Grilli E, Guzzi M and Zamboni R (1992) *Physical Review* B45: 1638. Thurmond CD (1975) L-minimum. *Journal Electrochemical Society*, 122: 1133 and Aspines DE (1976) L-Minimum. *Physical Review* B14: 5331.

Guzzi M, Grilli E, Oggioni S, Staehli JL, Bosio C and Pavesi L (1992) Indirect-energy-gap dependence on Al concentration in $Al_xGa_{1-x}As$ alloys. *Physical Review* B45: 10951.

Henning JCM, Ansens JPM and Roksnoer PJ (1986) Spectroscopic determination of L6 conduction-band minima in $Al_xGa_{1-x}As$. *Journal of Physical Chemistry*, L335.

Ioffe Institute, on NSM semiconductor data repository at http://www.ioffe.rssi.ru/SVA/NSM/Semicond/index.html

Landolt-Börnstein (1986) *Numerical data and functional relationships in science and technology*. In: Madelung O (ed.) *Semiconductors*, vol 22. Berlin: Springer Verlag.

Nakwaski W (1995) InAs, AlAs and their ternary compounds. *Physica* B210: 1–25.

Varshni YP (1967) Temperature dependence of the energy gap in semiconductors. *Physica* B34: 149.

Group IV Semiconductors, Si/SiGe

D J Paul, University of Cambridge, Cambridge, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

SiGe is now a mature materials system with electronic circuits on the market place. While no optical devices are available yet, there is a substantial amount of research in the field. In this article, the fundamental materials properties will be reviewed. As SiGe is a strained system, we shall start with the basic material properties as this determines the Ge content and layer thicknesses which are available for use. The band structure and electronic behavior will then be reviewed before the optical properties of the material.

The main drive for SiGe research in optical systems is the ability to use the extensive Si processing knowledge and to integrate a number of components onto a Si chip. Conventional Si transistors may also be used as driver circuits for the optical components, reducing chip counts in systems and therefore reducing costs. As Si processing is substantially cheaper than other semiconductor materials, if SiGe devices can be made on Si processing lines, they have the potential to be substantially cheaper than comparable III–V or II–VI devices. The SiGe material properties that will be reviewed in this article should be considered as an expansion of bulk silicon.

Material Properties and Strain

Ge has a 4.2% larger lattice constant that Si (Table 1). When growing epitaxial films of germanium or the alloy $Si_{1-x}Ge_x$ $(0 < x < 1)$ on a silicon substrate, there exists a maximum thickness called the critical thickness above which it costs too much energy to strain additional layers of material into coherence with the substrate. Instead defects appear, in this case misfit dislocations, which act to relieve the strain in the epitaxial film. Relaxed Si, Ge or SiGe films form diamond, diamond and zincblende lattices respectively while strained heterolayers below the critical thickness form a tetragonal symmetry arrangement. The strain therefore will substantially modify the electronic and optical properties of the material as will be discussed later.

Most SiGe films are produced by either chemical vapor deposition or molecular beam epitaxy. Both these techniques allow the deposition of films at low temperatures (typically 670–970 K) which is essential to prevent strained layers relaxing and forming defects. Chemical vapor deposition uses a number of different gases such as silane, disilane, dichlorosilane, and germane (occasionally with a hydrogen carrier gas) as the main species, while phosphine or arsine are used as n-type dopants and diborane is the typical p-type dopant. With the high melting point of both silicon and germanium, molecular beam epitaxy requires electron beam evaporators for the host materials silicon and germanium while effusion cells of antimony (occasionally phosphorus) and boron are used for n- and p-type dopants, respectively.

There are a number of models used to calculate the equilibrium critical thickness. The most often quoted is the Matthews and Blakeslee model (Figure 1). Experimentally it has been shown that defect-free films with thicknesses above the critical thickness may be grown (Figure 1) although these layers are metastable and will relax if temperatures above the

Table 1 Materials properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk relaxed* $Si_{1-x}Ge_x$
Lattice		Diamond	Diamond	Zincblende
Lattice constant	a_o (nm)	0.5431	0.5657	$0.5431 + 0.0226x$
Lattice mismatch	η_o (%)	0	4.17	$0.04115x - 0.02500x^2 + 0.05826x^3 - 0.03379x^4$
Density	(kg m^{-3})	2329	5323	$2329 + 2994x$
Sound velocity	m s^{-1} (300 K)	8440	4900	$8440 - 3540x$
Poisson s ratio	ν	0.280	0.273	$0.280 - 0.007x$
Melting point	(K)	1685	1210	
Elastic moduli	c_{11} (GPa)	165.8	128.5	
Elastic moduli	c_{12} (GPa)	63.9	48.3	
Elastic moduli	c_{44} (GPa)	79.6	66.8	
Thermal conductivity	κ (W m^{-1} K^{-1}) (300 K)	140	60	
Dielectric constant	ε	11.9	16.2	$11.9 + 4.1x$

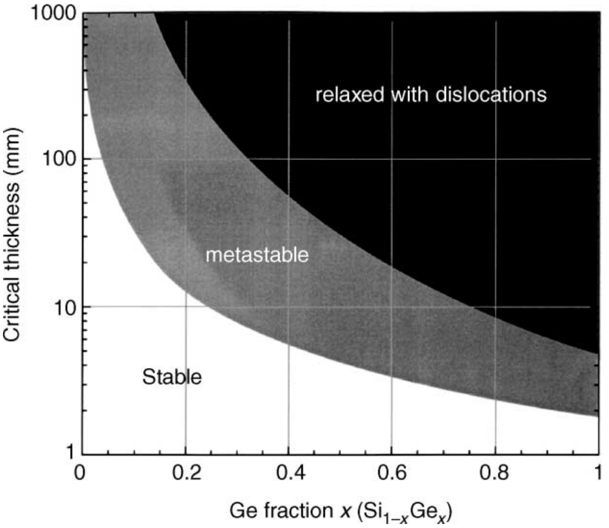

Figure 1 The critical thicknesses for growing a strained $Si_{1-x}Ge_x$ epitaxial layer on a bulk silicon substrate. The white area corresponds to the region below the equilibrium critical thickness as calculated with the Matthews and Blakeslee model while the gray region corresponds to a metastable layer which may relax after thermal treatments. The black area will have dislocations and strain relaxation using any growth technique or temperature.

growth temperature are applied to the heterolayers. For any manufacture, the layers cannot be metastable as high-temperature processes are normally employed in processing. The Matthews and Blakeslee equilibrium critical thickness, h_c for strained-$Si_{1-x}Ge_x$ on bulk Si is given by

$$h_c = \frac{0.55}{x}\ln[10h_c] \qquad [1]$$

while an approximate value can be obtained from

$$h_c \approx 1.7793x^{-1.2371}(\mathrm{nm}) \qquad [2]$$

Figure 1 clearly demonstrates that for thick SiGe films, low Ge contents must be used.

There is also a substantial interest in producing low-defect relaxed $Si_{1-y}Ge_y$ films or strained-Si on a relaxed $Si_{1-y}Ge_y$ substrate which can provide bandgaps and band discontinuities not available to pseudomorphically grown alloy layers. Ideally the relaxed $Si_{1-y}Ge_y$ substrate is grown on a Si substrate as this allows integration with conventional complementary metal oxide semiconductor (CMOS) electronics. To date no high-quality bulk single-crystal $Si_{1-y}Ge_y$ substrates are available due to the strong segregation of Si and Ge phases as liquid SiGe crystallizes. Growing a thick $Si_{1-x}Ge_x$ layer on Si well above the critical thickness so that it relaxes produces a threading dislocation density of about 10^{12} cm^{-2} which is far too large to produce device-quality material for electronics or optics. By growing a linearly graded $Si_{1-y}Ge_y$ with the grading rate sufficiently low (typically 10% per μm), the strain relaxation mechanism changes from a nucleation relaxation mechanism to a glide relaxation mechanism resulting in dislocations with comparatively long misfit segments. The defect densities are typically 10^5–10^6 cm^{-2} (Figure 2) although recent thermal processes and other techniques have now reduced this to 100 cm^{-2}.

Substantial work has appeared in the literature on self-organized growth of islands. Such islands are of interest in many materials systems as the quantum confinement they may offer can potentially enhance luminescence properties in numerous opto-electronic devices. The dots are formed by the Stranski–Kranstanow growth mode in a strained system. Planar epitaxial layers grown under stress conditions can lower their elastic energy by morphological changes. For the growth of strained-Ge or $Si_{1-x}Ge_x$ on bulk Si, this corresponds to the formation of islands when the growth parameters in the system are correctly set. While the islands may have well-defined shapes, the formation of the dots relies on random nucleation followed by ripening and coalescence. These mechanisms have produced large distribution of sizes (up to 30% of the mean size), densities and positions. It is quite common to have distributions of two or more sizes of islands and the densities are typically of order of 10^8–10^9 islands per cm^2. While the heights of these islands may be as low as 3 nm, the

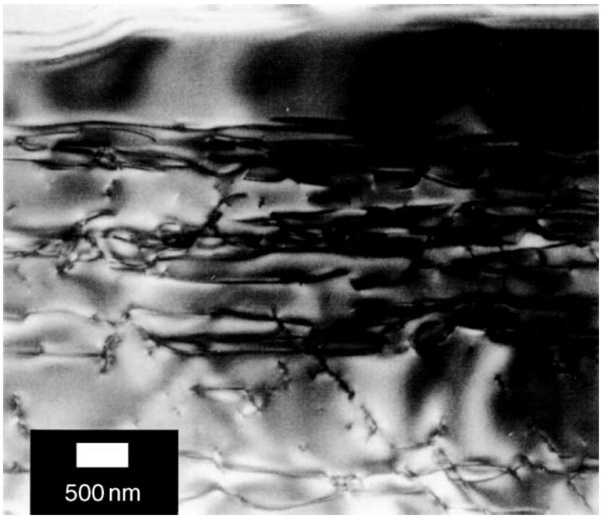

Figure 2 A transmission electron micrograph of a strain relaxation buffer grown by DERA, Malvern. The dense region of dislocations corresponds to an approximately 4 μm thick $Si_{1-y}Ge_y$ layer which is graded from $y = 0$ to $y = 0.23$. Above this is a 1 μm thick $Si_{0.77}Ge_{0.23}$ buffer which has no dislocation segments threading to the surface.

widths of the islands are typically 20 nm or greater, which is not yet small enough to produce the quantum confinement properties useful for significant improvements in opto-electronic devices.

Electronic Properties and Band Structure

As many of the optical properties and responses are derived or result from the band structure of materials, this section will concentrate on the band structure and electronic properties of SiGe layers. The bandgap of SiGe whether strained or unstrained is always below that of bulk Si. This potentially allows applications at longer wavelengths than comparable Si devices. Figure 3 shows the appropriate bandgaps for both the unstrained and strained $Si_{1-x}Ge_x$ grown on relaxed Si. For the unstrained case with a Ge fraction $x < 0.85$, the band structure is Si-like with six Δ-valleys forming the conduction band (Figure 4a). For $x > 0.85$ the band structure is Ge-like with four L-valleys. When strain is applied, both the conduction band and the valence band are split. Figure 4b shows the splitting of the conduction band with the Δ_4 valleys lower in energy and the Δ_2 valleys higher. In addition the valence band is split with the heavy-hole band being higher in energy than the light-hole band (Figure 3). The split-off band is also moved further away from the light and heavy hole bands as the strain in the system is increased. The indirect energy gap for the strained $Si_{1-x}Ge_x$ at 4.2 K as measured by photoluminescence is given by

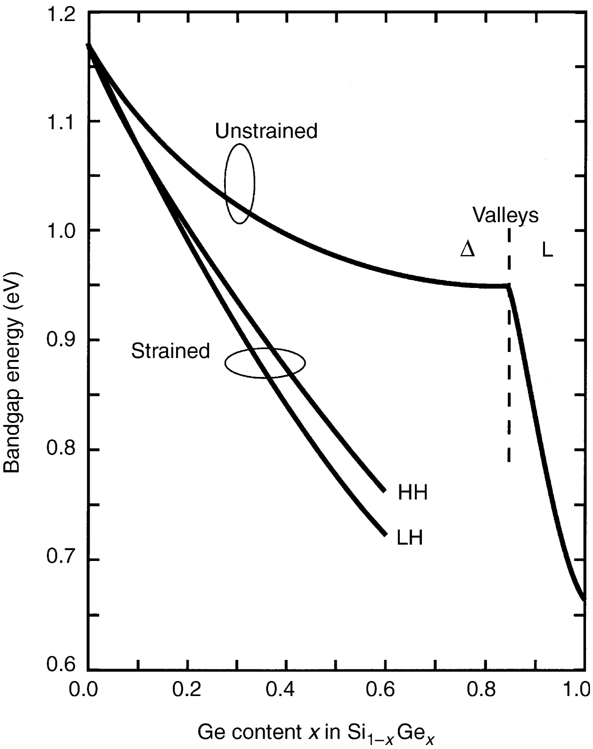

Figure 3 The bandgap energy for strained $Si_{1-x}Ge_x$ on a relaxed Si substrate and unstrained $Si_{1-x}Ge_x$. For the unstrained there is a discontinuity at $x \sim 0.85$ where the band structure changes from Si-like with Δ-valleys to Ge-like with L valleys (see Figure 4).

$$E_g = 1.155 - 0.43x + 0.0206x^2\,[\text{eV}] \text{ for } x < 0.85 \quad [3]$$

$$E_g = 2.010 - 1.27x\,[\text{eV}] \text{ for } x > 0.85 \quad [4]$$

while the direct bandgap is given by

$$E_g = 3.395 - 1.287x + 0.153x(1-x)\,[\text{eV}]\,(300\,\text{K}) \quad [5]$$

Due to the critical thickness, measurements of the bandgap of strained $Si_{1-x}Ge_x$ grown pseudomorphically on bulk Si substrates has been limited to low Ge concentrations. Using photoluminescence measurements at 4.2 K, Robbins has obtained a value for the exciton bandgap for strained $Si_{1-x}Ge_x$ layers for $x < 0.25$ to be

$$E_g = 1.155 - 0.874x + 0.376x^2\,[\text{eV}] \quad [6]$$

The bandgap for any strained $Si_{1-x}Ge_x$ layer grown on a relaxed $Si_{1-y}Ge_y$ substrate is shown in Figure 5. If the active strained $Si_{1-x}Ge_x$ layer is compressively strained then the conduction band is split into Δ_4 valleys being lower in energy and the Δ_2 valleys being higher (Figure 4b). For tensile strain (such as strained Si on a relaxed $Si_{1-y}Ge_y$ substrate), the Δ_2 valleys are lower in energy and the Δ_4 valleys are higher (Figure 4b).

The band alignments for typical heterolayer structures are shown in Figure 6. Pseudomorphically grown strained $Si_{1-x}Ge_x$ layers are only useful for forming quantum wells in the valence band (Figure 6a). Photoluminescence measurements do suggest a type I band alignment for these structures (quantum wells in both the conduction and valence bands) although it was later shown that the material is type II but due to charges producing band bending, the photoluminescence wrongly suggests a type I transition. For a quantum well for electrons, strained Si layers are grown on a relaxed $Si_{1-y}Ge_y$ substrate. In addition, quantum wells in the valence band can be produced for strained $Si_{1-x}Ge_x$ layers with high Ge content grown on relaxed $Si_{1-y}Ge_y$ substrates for $x > y$. In almost all cases the band alignments are type II although a small region exists where theory suggests type I is possible but this has not been demonstrated experimentally at present.

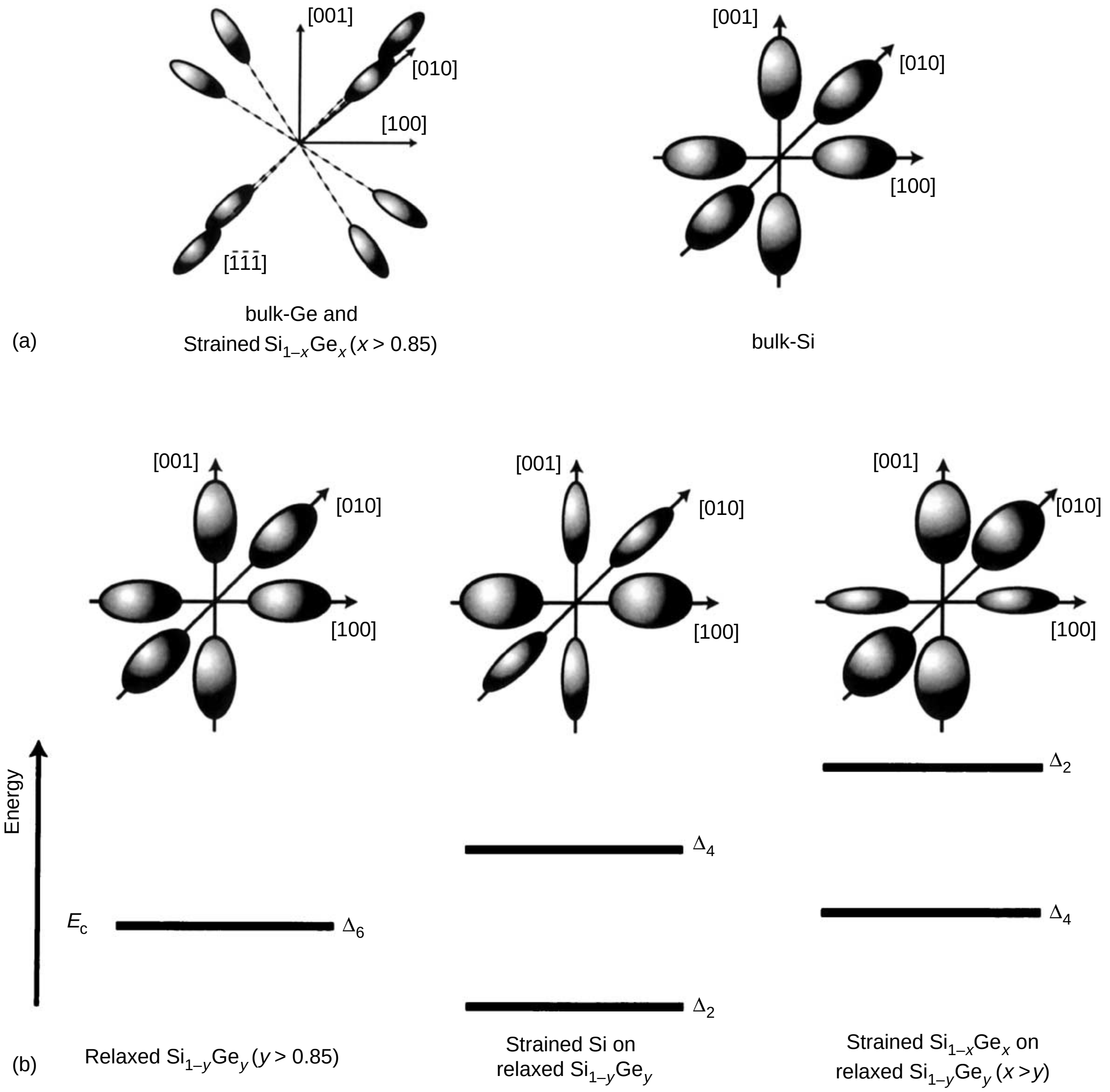
[001]
[010]
[100]
[$\bar{1}\bar{1}\bar{1}$]
(a)
bulk-Ge and
Strained $Si_{1-x}Ge_x$ ($x > 0.85$)
[001]
[010]
[100]
bulk-Si
[001]
[010]
[100]
Energy
E_c
Δ_6
Δ_4
Δ_2
Δ_2
Δ_4
(b)
Relaxed $Si_{1-y}Ge_y$ ($y > 0.85$)
Strained Si on
relaxed $Si_{1-y}Ge_y$
Strained $Si_{1-x}Ge_x$ on
relaxed $Si_{1-y}Ge_y$ ($x > y$)

Figure 4 (a) A schematic diagram of the valleys for Ge-like and Si-like material. (b) A schematic diagram of the valley structures and energy levels of the conduction bands for relaxed, tensile strained (strained Si on relaxed $Si_{1-y}Ge_y$) and compressively strained (strained $Si_{1-x}Ge_x$ on relaxed $Si_{1-y}Ge_y$ ($x > y$)) material.

The valence band discontinuity is given by

$$\Delta E_v = (0.74 - 0.06y)(x - y)\,[\mathrm{eV}] \qquad [7]$$

while the conduction band discontinuity for strained Si on a relaxed $Si_{1-y}Ge_y$ substrate is

$$\Delta E_c = 0.6y\,[\mathrm{eV}] \qquad \text{for } x < 0.6 \qquad [8]$$

The complete set of ΔE_c and ΔE_v values for all x and y are shown in Figures 7 and 8 respectively. Table 2 gives some of the electronic data on Si, Ge, and relaxed $Si_{1-x}Ge_x$ including mobilities, bandgaps, and effective masses.

The mobilities for electrons and holes in relaxed $Si_{1-x}Ge_x$ layers are below the bulk Si and Ge layers due to alloy scattering (Table 2). For strained layers, mobility enhancements have been demonstrated in a number of different systems. Electron mobility enhancements result from the splitting of the valleys reducing intervalley scattering. Hole mobility enhancements result from the splitting of the valence band and reduction of effective masses. The hole mobilities for pseudomorphically grown strained $Si_{1-x}Ge_x$ quantum wells have shown values as high as 17 000 $cm^2\,V^{-1}\,s^{-1}$ at a carrier density of $2 \times 10^{11}\,cm^{-2}$ for $Si_{0.87}Ge_{0.13}$ layers. Higher Ge contents reduce the mobility due to alloy scattering. At room temperature, only modest mobility improvements of about 20% have been observed giving 220 $cm^2\,V^{-1}\,s^{-1}$ at the high carrier

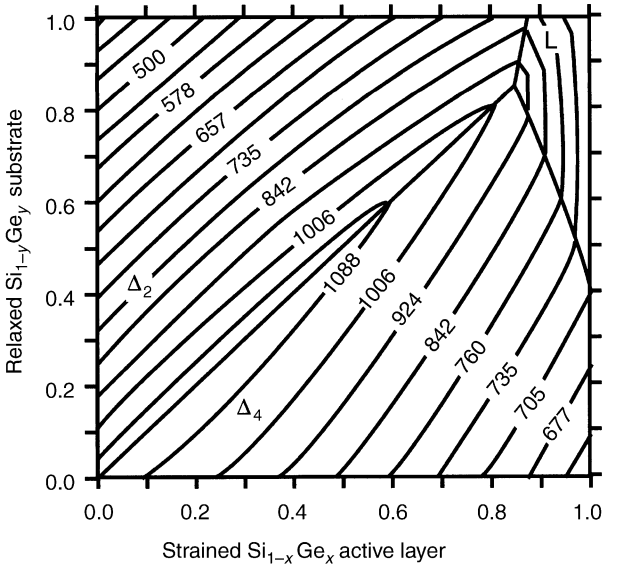

Figure 5 The bandgap energy (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Theoretical curves from Riger and Vogl have been calibrated to the experimental data from pseudomorphic layers.

densities required for electronic devices of about $2 \times 10^{12}\ cm^{-2}$. To remove alloy scattering, pure strained Ge channels grown on relaxed $Si_{1-y}Ge_y$ substrates have shown low-temperature mobilities of $55\,000\ cm^2\ V^{-1}\ s^{-1}$ for densities of $5 \times 10^{11}\ cm^{-2}$ with the room-temperature values reaching $1665\ cm^2\ V^{-1}\ s^{-1}$. For electrons, a strained Si layer on relaxed $Si_{1-y}Ge_y$ substrates ($0.2 < y < 0.3$) also reduces the alloy scattering in the channel and produces much higher low-temperature mobilities of $390\,000\ cm^2\ V^{-1}\ s^{-1}$ for densities of $4 \times 10^{11}\ cm^{-2}$ while room-temperature values have peaked at $2830\ cm^2\ V^{-1}\ s^{-1}$. All these values correspond to electrons or holes at heterointerfaces. Enhanced mobilities have also been observed for strained Si metal oxide semiconductor field effect transistors (MOSFET) where electron mobility enhancements of 75% compared to the bulk Si devices has been reported on relaxed $Si_{0.8}Ge_{0.2}$ although intrinsic transconductances are only increased by 60%. For holes, a strained Si layer on a $Si_{0.8}Ge_{0.2}$ relaxed substrate has demonstrated a 50% mobility enhancement over a bulk Si MOSFET.

The effective mass for electrons in strained Si layers on relaxed $Si_{1-y}Ge_y$ substrates for $0.1 < y < 0.4$ has produced a very small spread of effective masses with values of 0.19 m_0 (where m_0 is the free electron mass) for the longitudinal mass as measured by cyclotron resonance and the temperature dependence of Shubnikov–de Haas oscillations. The electron effective masses are very similar to the bulk values for strained Si quantum wells. Hole effective masses in pseudomorphically strained $Si_{1-x}Ge_x$ layers have produced a large spread of data even for similar alloy and strain values mainly due to the large nonparabolicity of the valence band (Table 3) and many of the measured samples having metastable $Si_{1-x}Ge_x$ quantum wells. The in-plane hole mass is theoretically expected to decrease as the strain in the system is increased but to date the experimental values have been much lower than theoretically calculated values. Experimental data (Table 3)

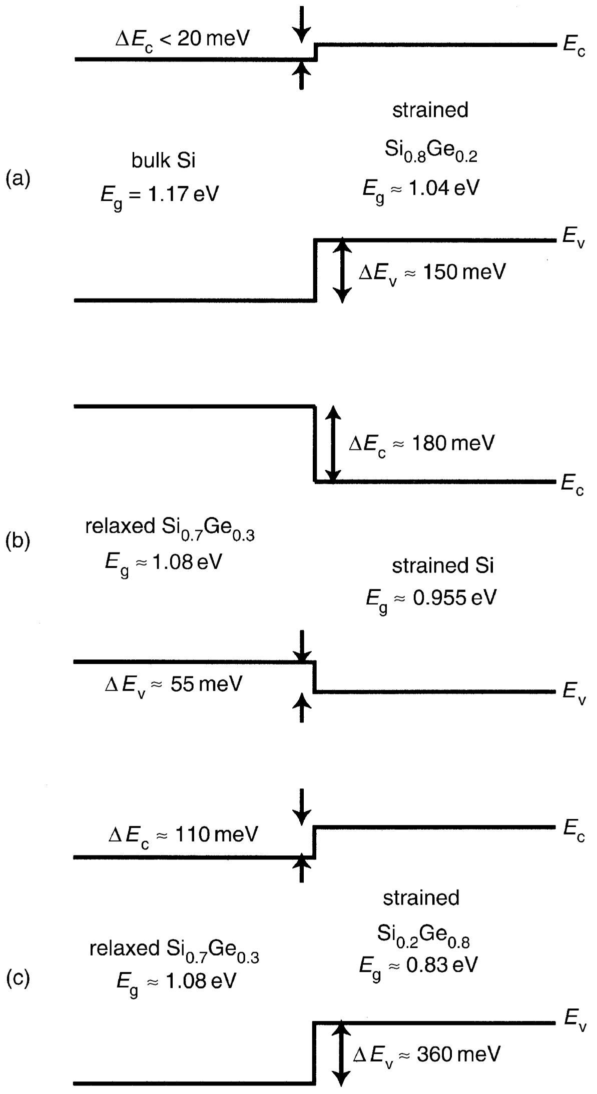
(a)
$\Delta E_c < 20$ meV
E_c
strained
$Si_{0.8}Ge_{0.2}$
$E_g \approx 1.04$ eV
bulk Si
$E_g = 1.17$ eV
E_v
$\Delta E_v \approx 150$ meV
(b)
$\Delta E_c \approx 180$ meV
E_c
relaxed $Si_{0.7}Ge_{0.3}$
$E_g \approx 1.08$ eV
strained Si
$E_g \approx 0.955$ eV
$\Delta E_v \approx 55$ meV
E_v
(c)
$\Delta E_c \approx 110$ meV
E_c
strained
$Si_{0.2}Ge_{0.8}$
$E_g \approx 0.83$ eV
relaxed $Si_{0.7}Ge_{0.3}$
$E_g \approx 1.08$ eV
E_v
$\Delta E_v \approx 360$ meV

Figure 6 A schematic diagram of some typical SiGe heterolayer structures showing the bandgap, conduction band discontinuity, ΔE_c, and the valence band discontinuity, ΔE_v. (a) shows a quantum well in the valence band using a pseudomorphic strained $Si_{0.8}Ge_{0.2}$ layer; (b) shows a quantum well in the conduction band using a strained Si layer on a relaxed $Si_{0.7}Ge_{0.3}$ substrate; and (c) shows a quantum well in the valence band using a high Ge content layer on a relaxed substrate with lower Ge content.

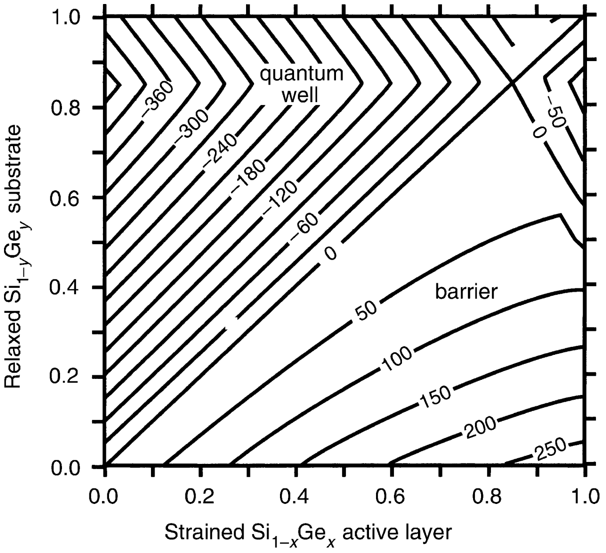

Figure 7 The conduction band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to potential barriers to electrons while negative values represent quantum wells. The general trends and unknown experimental regimes are from the theoretical work of Rieger and Vogl while the quantum well data for $0.1 < y < 0.4$ and $x = 0$ has been recalibrated from experimental work. No experimental data are known to allow calibration of the barrier regime.

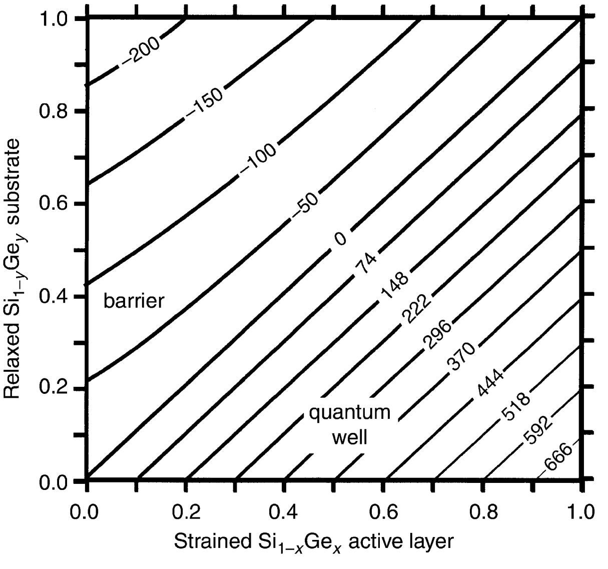

Figure 8 The valence band discontinuity (in meV) for strained $Si_{1-x}Ge_x$ grown on a relaxed $Si_{1-y}Ge_y$ substrate. Positive numbers correspond to quantum wells for holes while negative values represent potential barriers. The values from theory in the quantum well region agree well with experimental values but little experimental data exist to compare the barrier data from the theorertical values (plot from Rieger and Vogl) with experiment.

suggests that strained $Si_{0.88}Ge_{0.12}$ layers have a heavy effective mass of about $0.3\,m_0$ compared to the bulk value of $0.55\,m_0$.

Optical Properties

Due to the problems of producing high-quality single crystals of $Si_{1-x}Ge_x$, there are relatively little data on the optical properties of relaxed $Si_{1-x}Ge_x$ in the literature. The situation is even worse for strained $Si_{1-x}Ge_x$ as the layer thicknesses are limited to below 30 nm for Ge contents above 0.1, a value much smaller than the penetration depths in the system at most optical frequencies. There is, however, substantial interest in the optical properties, particularly the refractive index, as they suggest that SiGe can be used as a waveguide on Si substrates.

For the relaxed $Si_{1-x}Ge_x$ layers, most of the published work in the field has been from crystals grown by liquid phase epitaxy using melts of indium or tin. The samples had free hole concentrations of $10^{17}\,cm^{-3}$ along with indium or tin impurities and details can be found in the reviews by Humlícek. Some of the data are also from thick epitaxial layers grown well above the critical thickness where the threading dislocation density is likely to be about $10^{12}\,cm^{-2}$. The data below 2 eV were obtained by absorption measurements, those between 2 and 5.6 eV have been obtained using ellipsometry and the values between 6 and 12 eV were obtained by reflectance measurements.

Figures 9 and 10 show the real and imaginary parts of the complex dielectric function defined as $\varepsilon = \varepsilon_1 + i\varepsilon_2$ where $i = \sqrt{-1}$ for a number of different Ge contents. Figures 11 and 12 show the real and imaginary parts of the complex refractive index defined by $N = n + ik = \sqrt{\varepsilon}$. The peak starting at 3.395 eV and decreasing as the Ge content is increased corresponds to the E_1 transition across the direct gap and may be used to find the Ge content using the expression

$$x = 4.707 - \sqrt{6.538E_1 - 0.0397} \qquad [9]$$

The refractive index is high owing to the strong electronic interband absorption in the visible and ultraviolet regions of the spectrum.

The other important optical property of the material is the absorption. The absorption coefficient is defined as $K = 4\pi k/\lambda$ where k is the extinction coefficient for a wavelength λ. The intensity of a wave traversing a distance d in a homogeneous medium of extinction coefficient, k, is attenuated by a factor $\exp(-Kd)$. Therefore $1/K$ is the penetration depth. The absorption coefficient is shown in Figure 13.

Table 2 Electronic properties of Si, Ge and relaxed $Si_{1-x}Ge_x$

Element		*Si*	*Ge*	*Bulk unstrained* $Si_{1-x}Ge_x$
Direct bandgap	(eV) at 300 K	3.4	0.80	$3.395 - 1.287x + 0.153x(1 - x)$
Indirect bandgap	(eV) at 300 K	1.11 (D)	0.664 (L)	$1.155 - 0.43x + 0.0206x^2$ $(x < 0.85)$ $2.010 - 1.27x$ $(x > 0.85)$
Spin-orbit splitting	Δ_{so} (meV)	44	290	Discontinuity at $x \sim 0.85$
Electron mobility	μ_e ($cm^2\ V^{-1}\ s^{-1}$) at 300 K	1450	3900	$\mu_e = 1466.2 - 6425.9x + 6601.8x^2 + 7716.5x^3 - 10944x^4 (x < 0.6)$ $\mu_e = 2.9235 \times 10^{-28} e^{71.105x} + 4790.1 - 14246x + 11117x^2 (x > 0.6)$
Hole mobility	μ_h ($cm^2\ V^{-1}\ s^1$) at 300 K	505	1800	$\mu_h = 188.42 - 935.7x + 5455.5x^2 - 8553.7x^3 + 5751.7x^4$ (all x) ($N_A = 1 - 50 \times 10^{16}\ cm^{-3}$)
Longitudinal electron mass	m_l^*	0.1905	0.082	Discontinuity at $x \sim 0.85$
Transverse electron mass	m_t^*	0.9163	1.58	Discontinuity at $x \sim 0.85$
Heavy-hole mass	m_{hh}^*	0.537	0.284	Discontinuity at $x \sim 0.85$
Light-hole mass	m_{lh}^*	0.153	0.044	Discontinuity at $x \sim 0.85$
Spin–orbit mass	m_{so}^*	0.234	0.095	Discontinuity at $x \sim 0.85$

Table 3 A comparison of effective masses in modulation-doped samples as measured by the temperature dependence of the Shubnikov–de Haas oscillations (SdH) or by cyclotron resonance (CR) at liquid helium temperatures. All samples have strained $Si_{0.88}Ge_{0.12}$ or $Si_{0.87}Ge_{0.13}$ quantum wells grown on bulk Si substrates. Therefore only the heavy-hole band should be populated at these low temperatures and strain values. ℓ is the mean free path which is the mean scattering distance of a hole in the system and is a quantitative measure of disorder

	Mobility ($cm^2\ V^{-1}\ s^{-1}$)	p ($\times 10^{11}\ cm^{-2}$)	ℓ *(nm)*	m^* (m_0) *SdH*	m^* (m_0) *CR*
DERA	6560	4.4	71	0.304	0.26
NRC	6800	5.9	86	0.31	–
IBM	6000	3.7	60	0.34	0.18
Warwick	9600	2.5	78	0.27	0.23, 0.29
Warwick	10500	2.1	78	0.27	0.24, 0.30

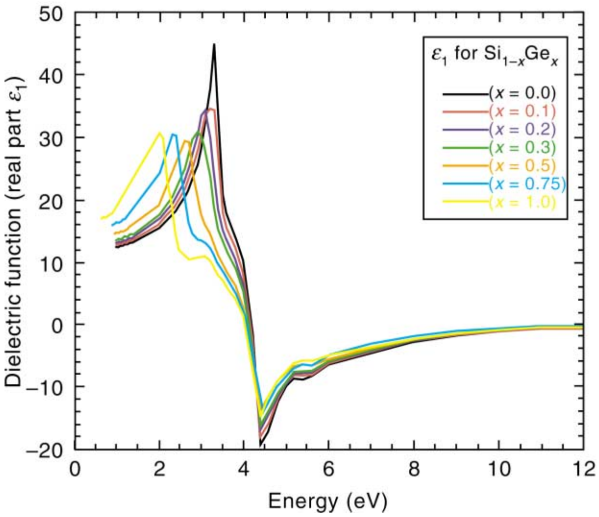

Figure 9 The real part (ε_1) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

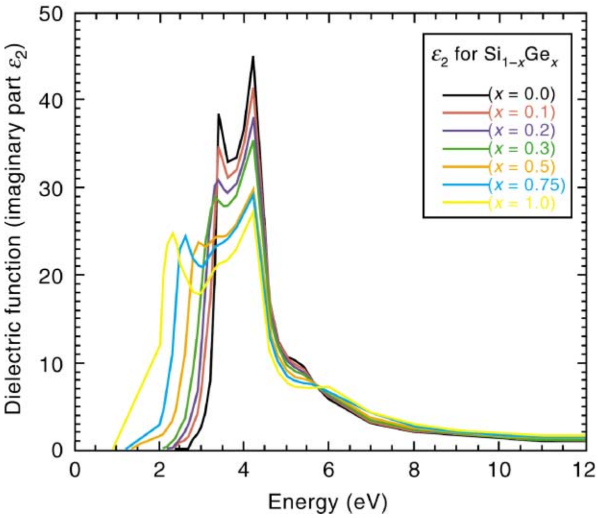

Figure 10 The imaginary part (ε_2) of the complex dielectric function $\varepsilon = \varepsilon_1 + i\varepsilon_2$.

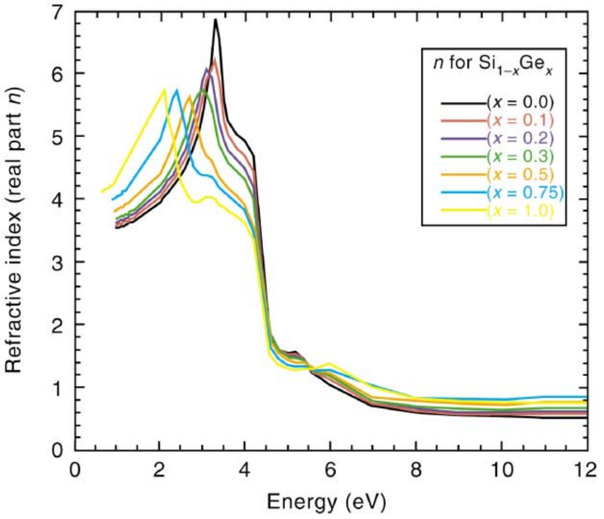

Figure 11 The real part (n) of the complex refractive index $N = n + ik$.

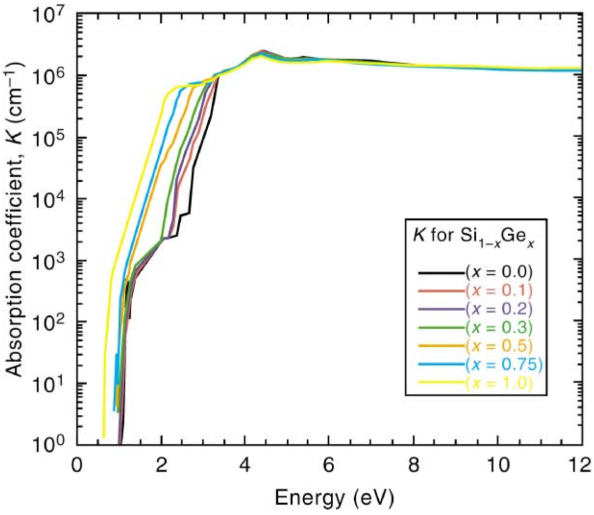

Figure 13 The absorption coefficient K.

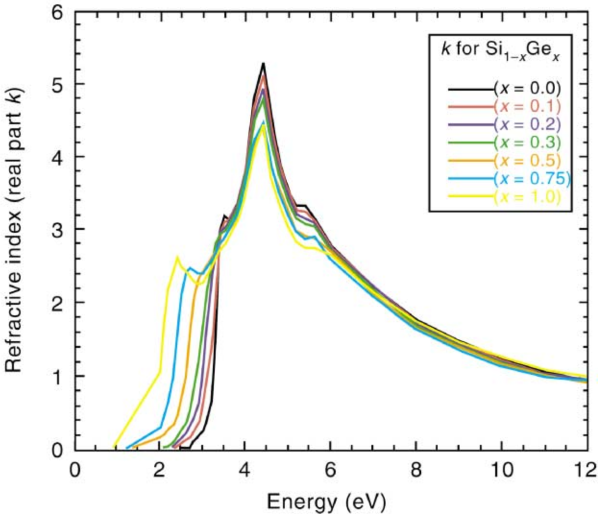

Figure 12 The imaginary part (k) of the complex refractive index $N = n + ik$.

Determining the optical properties of strained $Si_{1-x}Ge_x$ layers has been substantially more difficult. The major problem is that the absorption coefficient is about 10–20 cm^{-1} around the fundamental bandgap (Figure 6) and therefore any shift due to strain effects becomes very difficult to observe using absorption techniques in thin SiGe layers. Additional complexity results as the substrate material has a fundamental bandgap close to the strained layer. Therefore photoluminescence is one of the few methods for extracting experimentally the optical changes due to strain.

The splitting of the conduction (Figure 8) and valence bands (Figure 9) has already been discussed. The photoluminescence signals will also shift with strain for the lowest lying bandgap energy. Almost all the experimental data for strained layers have been obtained for pseudomorphically grown strained $Si_{1-x}Ge_x$ alloys on bulk Si substrates and the data agree well with the bandgaps shown in Figures 3 and 5. The quantum well photoluminescence signal is found to shift by −1.4 meV per kbar of pressure. After removal of nonhydrostatic components, a value of −1.52 meV per kbar for the pure hydrostatic shift has been reported.

Conclusions

A review has been presented of the fundamental properties of relaxed and strained $Si_{1-x}Ge_x$ alloys. The materials properties were first reviewed and demonstrate the limitations imposed by the critical thickness. The electronic properties including the bandstructure were then discussed and demonstrate the lowered bandgap for $Si_{1-x}Ge_x$ alloy compositions compared to bulk Si. This property has been used in $Si_{1-x}Ge_x$ p-i-n photodetectors to move the absorption to longer wavelengths in an attempt to compete with III–V photodetectors for the important 1.3 μm and 1.55 μm fiber optic frequencies. In the final section the optical properties of relaxed and strained $Si_{1-x}Ge_x$ material were reviewed. The difference in refractive indices has opened research into SiGe on Si waveguides.

List of Units and Nomenclature

Absorption coefficient (K)	[cm^{-1}]
Band gap energy (E_g)	[eV]
Conduction band discontinuity (ΔE_c)	[meV]
Critical thickness	[nm]
Density	[kg m^{-3}]
Dielectric constant (ε)	
Effective mass [m_0 where $m_0 = 9.11 \times 10^{-31}$ *kg*]	
Elastic moduli (c_{ij})	[GPa]
Extinction coefficient (k)	[cm^{-1}]
Lattice constant (a_0)	[nm]
Mean free path (l)	[nm]
Melting point	[K]
Mobility (μ)	[cm^2 V^{-1} s^{-1}]
Pressure	[kbar]
Refractive index (N)	
Sheet carrier densities	[cm^{-2}]
Sound velocity	[m s^{-1}]
Temperature	[K]
Thermal conductivity (κ)	[W m^{-1} K^{-1}]
Valence band discontinuity (ΔE_v)	[meV]
Wavelength (λ)	[nm]

See also

Semiconductor Physics: Band Structure and Optical Properties; Outline of Basic Electronic Properties; Recombination Processes.

Further Reading

Humlícek J, Garriga M, Alonso MI and Cardona M (1989) Optical spectra of $Si_{1-x}Ge_x$ alloys. *Journal of Applied Physics* 65(7): 2827–2832.

Humlícek J, Lukes F and Schmidt E (1991) Silicon-germanium alloys (Si_xGe_{1-x}). In: Palik ED (ed.) *Handbook of Optical Constants of Solids*, vol. II, pp. 607–636. San Diego, CA: Academic Press.

Kasper E and Lyutovich K (eds) *Properties of Silicon Germanium and SiGe:Carbon*. London: INSPEC IEE.

Paul DJ (1999) Silicon-germanium strained layer materials in microelectronics. *Advanced Materials* 11(3): 191–204.

Rieger MM and Vogl P (1993) Electronic-band parameters in strained $Si_{1-x}Ge_x$ alloys on $Si_{1-y}Ge_y$ substrates. *Physical Review B* 48(19): 14276–14287.

Robbins DJ, Canham LT, Barnett SJ, Pitt AD and Calcott P (1992) Near band gap photoluminescence from pseudomorphic $Si_{1-x}Ge_x$ single layers on silicon. *Journal of Applied Physics* 71(3): 1407–1414.

III-Nitrides

K P O'Donnell, University of Strathclyde, Glasgow, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction to InGaN

The optical properties of wurtzite GaN are of interest mainly to band-structure theorists. Luminescence from pure InN has probably never been detected. It is therefore surprising that the combination of these binary III–N compounds produces a solid solution, InGaN, which has revolutionized blue–green optoelectronics in the last 10 years, and may form the basis of a new lighting technology to replace the tungsten lamp in the first half of the twenty-first century.

InGaN is also the material basis of the blue laser diode, which will be used for future generations of optical data storage and for many other applications that require a compact source of short-wavelength coherent light. The market for nitride light emitters has been growing at a rate of ~20% a year since the introduction of the first blue nitride light-emitting diode (LED) by Nichia Chemical in 1993. The first commercial lasers, with lifetimes in excess of 10 000 hours, were released in 1998.

So what is InGaN? The naïve response–that it is a semiconductor alloy–ignores the theoretical demonstration, based on simple equilibrium thermodynamics, that InN is insoluble in GaN at the temperatures usual for growth (using metalorganic vapor phase epitaxy, MOVPE). If we imagine a cation lattice, on which we distribute Ga and In atoms, in a fixed numerical proportion, in order to form a solid solution of given composition, it is thermodynamically probable that the In atoms will associate to form agglomerations in certain regions of the lattice. The size, shape, and properties of the resulting concentration fluctuations have been a topic of intense speculation.

Practically speaking, InGaN layers are grown either as rather thin (<3 nm) quantum wells (QWs) immersed in GaN or AlGaN or as thicker epilayers (10–500 nm) on top of thick GaN templates. Only the QWs are of use in devices, whereas the composition and structure of the thick epilayers are easier to measure in research applications. It is dangerous to suppose that the material retains the same properties when grown in thicker layers. Recently, a third class of sample has emerged, in the form of self-organized 'quantum boxes' (originally French *boîtes*), which we shall call QBs. These samples, grown at low temperatures by molecular beam epitaxy (MBE) comprise pancake-shaped disks of uncertain composition, about 50 nm across and 3–5 nm high.

Optical Spectra of Devices

There have been several 'generations' of nitride LED. We refer here to the Nichia Blue, Extrablue (actually turquoise) and Green diodes which set the scene in 1996, and the Amber diode which first appeared in 1998, although not yet available commercially. The spectral properties of these devices have been stable while their efficiencies have improved in the intervening years. The basic design appears to incorporate a single quantum well. Up to a point, the wavelength tuning depends upon the incorporation of different amounts of indium in the well. (Green and amber devices may have the same amount.)

In what follows, we prefer to specify composition by quoting the indium nitride content, x, according to the formula unit

$$\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N} = x\mathrm{InN} + (1-x)\mathrm{GaN} \qquad [1]$$

Electroluminescence (EL) spectra of blue and green diodes are shown in Figure 1. Spectra are broad (compared to k_BT at room temperature) with exponential tails to high and low photon energies. The linewidth tends to increase as the device peak wavelength shifts further to the red.

Whereas it is difficult to measure directly the optical absorption spectrum of an ultrathin layer of material, an equivalent measurement can be made on an LED by monitoring the photocurrent (PC) generated by incident monochromatic light. PC spectra of the blue and green diode are shown in Figure 1, together with the absorption and photoluminescence (PL) spectra of a thick violet-light-emitting epilayer.

The spectra of Figure 1 transform into each other by 'stretching' the x-axis, using the bandgap energy of GaN as an anchor point. Comparison of many pairs of spectra produces evidence of a *universal scaling behavior*, which relates the details of each spectrum to a single controlling parameter. In the case of epilayers, this is undoubtedly the InN fraction, as we shall see later.

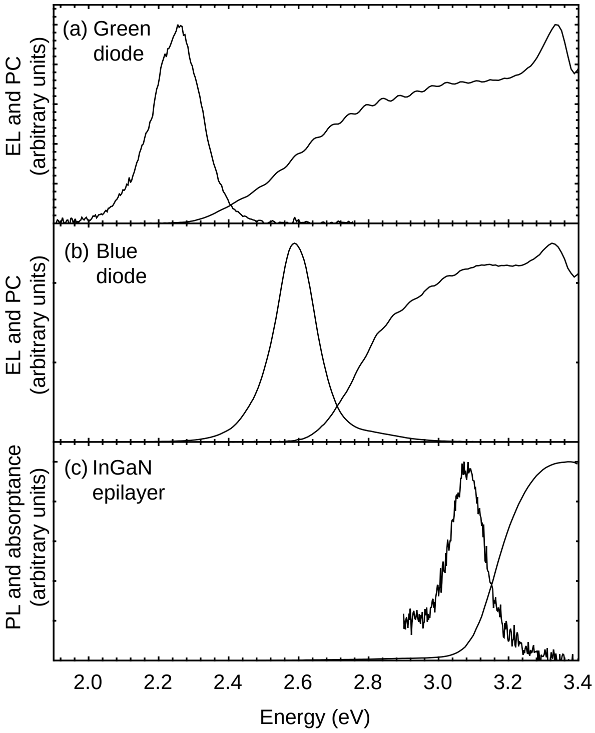

Figure 1 EL and PC (quasi-absorption) spectra of nitride LEDs, compared to PL and absorption spectra of a thick InGaN epilayer.

In order to describe the form of the absorption/excitation curve, we introduce the formal expression

$$\alpha = \frac{\alpha_0}{1 + \exp\left(\dfrac{E_g - E}{\Delta E}\right)} \qquad [2]$$

where α_0 is the plateau absorption, observed experimentally, but not yet theoretically explained, E_g is the (effective) bandgap energy of the sample, E is the photon energy, and ΔE is an energy broadening term, similar to the well-known Urbach parameter.

Optical Energy Relationships

The spectral universality pointed out in the previous section requires the existence of certain relationships between the optical energies. We list the defining energies as the bandgap, E_g, the peak energy of EL or PL, E_p, and the difference between these energies, measured for a particular sample, which we call the Stokes shift:

$$\mathrm{SS} = E_g - E_p \qquad [3]$$

(While the energy broadening tends to increase in a regular fashion with the Stokes shift, there are irregularities in both ΔE and the PL linewidths of samples from different sources. Hence ΔE and PL

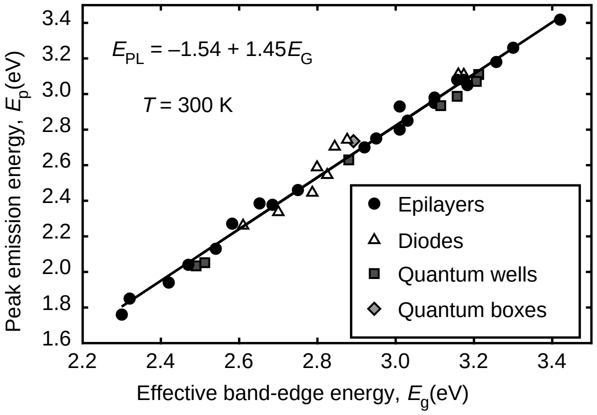

Figure 2 Linear dependence of luminescence peak energy on effective bandgap.

linewidth appear to be local rather than universal parameters.)

The relationship between E_g and E_p for a wide range of emitters is found to be linear (Figure 2):

$$E_p = 1.45E_g - 1.54 \qquad [4]$$

The bandgap energy and luminescence peak energy coincide at $1.54/(1.45 - 1) = 3.42$ eV, which corresponds to the energy bandgap of pure GaN. The simple linear dependence of E_p upon E_g implies a linear increase of the Stokes shift energy with increasing detuning from the GaN band edge.

Dependence of Optical Energies on *x*, the InN Fraction, for Thick Epilayers

As stated above, it is usual to measure the composition of thick layers, rather than quantum wells, for reasons of experimental convenience. Several techniques can be used to this end, including measurement of the lattice constants by X-ray diffraction (XRD) (one asymmetric and one symmetric reflections are required to take into account the strain state of the layer), Rutherford backscattering spectrometry (RBS), wavelength-dispersive (preferable) or energy-dispersive X-ray fluorescence (XRF), sputtered ion mass spectroscopy (SIMS), and extended X-ray absorption fine structure (EXAFS).

All usable techniques provide comparable results if carried out with sufficient care. We plot the peak energy of PL emission against the measured indium nitride fraction for a range of epilayer samples in Figure 3. Once more, the dependence appears to be linear in a restricted composition range. The best-fit line to a linear function $E_p(x)$ has an x-intercept which is close to the bandgap of GaN. The dependence of the optical energies on the indium nitride fraction can now be summarized by three interdependent linear relations, which are valid in the composition range $0 < x < 0.4$:

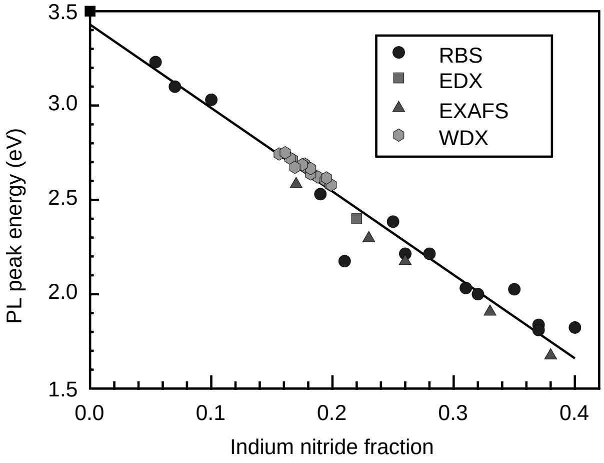

Figure 3 Linear dependence of the PL peak energy on the InN content of InGaN epilayers.

$$E_p = 3.4 - 4.4x$$
$$E_g = 3.4 - 3.0x \qquad [5]$$
$$SS = 1.4x$$

The measurements described above rely on separate determinations of composition and an optical energy for each sample. Since all samples are to some extent inhomogeneous, it would be preferable to make such measurements at one and the same time. This can be done by using an electron probe microanalyzer, modified to allow simultaneous cathodoluminescence (CL) spectroscopy. CL is analogous to PL. A recent study of samples with $0.05 < x < 0.25$ has yielded the result:

$$E_c = 3.40(1) - 3.91(5)x \qquad [6]$$

with error bars (shown) that are about ten times smaller than those of the previous work.

Bandgap of InN

It is dangerous to extrapolate experimentally determined relationships beyond the range of measurements used to derive them.

The bandgap of an alloy is best represented by a quadratic function of x, with the deviation from linearity described by a bowing parameter. The above measurements suggest a small bowing parameter for the InGaN alloy system. If we ignore bowing and equate E_p with E_c, the InN bandgap obtained by setting $x = 1$ is calculated to be 0.71 eV. This is certainly an underestimate, but is similar to recent values obtained experimentally for layers of pure InN grown by Molecular Beam Epitaxy. Before the year 2000, the generally accepted value of the InN bandgap was close to 2 eV.

Origin of the Stokes Shift

Stokes shifts originate in systematic energy loss during the excitation–luminescence cycle. While the bandgap, as defined above, is the mean energy of electron–hole pairs across the bandgap, the PL peak is the modal energy of electron–hole recombination. The large Stokes shift in the InGaN system is probably due to the influence of giant internal electric fields that separate the carriers prior to recombination.

In the absence of a microscopic model, it is difficult to provide a quantitative description of energy loss in the excitation–emission cycle. Experiment shows that the shift can be as large as ~ 0.6 eV in an epilayer with $x = 0.4$. If such a layer were 100 nm thick and uniform in composition, it would be strained by about 4% when pseudomorphic to a GaN template, with an estimated vertical electric field of 6×10^6 V m^{-1}. Unit strain would correspond to a 15-eV shift (equivalent to a piezoelectric deformation potential). While these numbers are large, they seem reasonable in the light of sophisticated theoretical calculations of the magnitude of the piezoelectric effect in InGaN. But it is an unfortunate experimental fact that layers grown with different thicknesses, and presumably under different states of strain, may show the same value of the Stokes shift.

If the material has composition fluctuations, accompanied by strain variations, the recombination energy will be much more sensitive to the local environment. One can imagine a situation in which recombination always takes place at the lowest available energy, as carriers diffuse under the action of local strain/electric fields. This situation would be rather similar to that which pertains in quantum wells with spatial energy disorder, caused by well width fluctuations. A theoretical treatment of localization in three dimensions would be very difficult.

In any case, the physical situation of excitation and recombination in the presence of an electric field is that sketched in Figure 4.

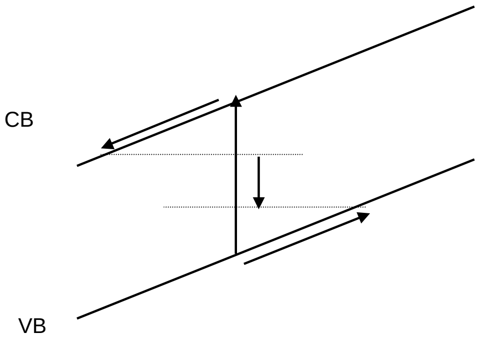

Figure 4 Excitation–emission cycle in the presence of an electric field. CB and VB are conduction band and valence band, respectively.

Quantum Wells and Quantum Boxes

The incorporation of quantum wells has several advantages for light-emitting devices. In nitrides, the expected increase of oscillator strength for radiative transitions, due to the reduced dimensionality, is mitigated to some extent by the opposing effect of the internal electric fields, but the quantum well still acts as an energy sink for mobile excitations and as a site for carrier localization.

Quantum confinement can also act as a wavelength tuning mechanism in QW devices. The effect in blue-emitting nitride devices is rather small until the QW width falls below 3 nm. It should be clear that a reduction in the photon energy of recombination with increasing well width will accompany vertical separation of carriers in a piezoelectric model. Usually, as mentioned above, wavelength tuning in nitride devices is accomplished by changing the composition. What should we make of the fact that amber and green diodes apparently have the same composition? It would appear that the well widths are different. On the other hand, the amber emitters have anomalously narrow linewidths, which may imply that some other mechanism is in force. Here, we must let the matter rest, in the absence of detailed structural information.

Quantum boxes should be distinguished from quantum dots (QDs). The former are mesoscopic objects that form spontaneously under certain growth conditions, which usually involve the influence of strain. The best-characterized example in semiconductor physics is the Stranski–Krastanow growth of InGaAs structures on GaAs. While GaN in AlGaN forms recognizable nanocrystals, InGaN forms less well-defined structures of indeterminate shape, some 30–50 nm in lateral extent and 3–10 nm high. QDs, on the other hand, are truly zero-dimensional objects, in the sense that carrier motion is quantized simultaneously in three spatial dimensions. The utilization of reduced dimensionality to enhance recombination probabilities in active devices is only valid when QB structures meet the criteria of low dimensionality: the carrier motion must be restricted to a distance smaller than the appropriate Bohr radius.

Structure of InGaN Epilayers

The state of strain of InGaN epilayers is usually determined by XRD. Both symmetric and asymmetric reflections are required to calculate the actual values of the out-of-plane and in-plane lattice constants, c and a, respectively. These can then be compared with those predicted for a relaxed alloy by Vegard's law, for example:

$$c(\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}) = xc_{\mathrm{InGaN}} + (1-x)c_{\mathrm{GaN}} \qquad [7]$$

The composition itself, as well as the state of strain, can be determined from measurements of c and a with one additional assumption: that the Poisson ratio of the alloy also obeys Vegard's law. If the composition is known from an independent chemical measurement, this assumption can be tested.

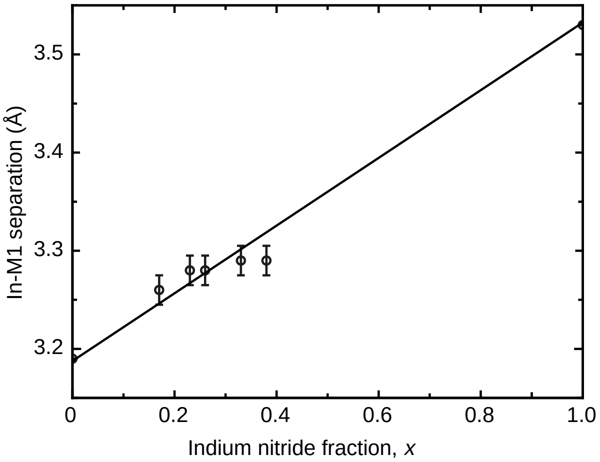

Figure 5 EXAFS-determined local structure parameters of thick InGan films.

InGaN epilayers are usually grown on thick GaN formers. In pseudomorphic growth, the constraint, $a(\text{layer}) = a_{\mathrm{GaN}}$ applies. This is sometimes assumed to be the case when layers are thinner than about 50 nm, but nothing should be taken for granted when dealing with this material. Each layer is different. A similar warning applies to the assumption that thick InGaN layers will be relaxed. It has recently been shown that partially relaxed and pseudomorphic regions may actually coexist within the same InGaN film, which nevertheless has a composition that is uniform throughout. Other films may show a concentration gradient due to compositional pulling during growth.

The local structure of InGaN films has been measured using In-edge EXAFS. Using a tetrahedral two-shell model that does not distinguish the wurtzite and cubic crystal structures, we find that the In–N1 bond length is more or less independent of the alloy composition in the range $0.17 < x < 0.38$ of samples measured, having a value of 2.11(1) Å (compared to 2.16 Å in pure InN). The M1 shell of mixed In and Ga atoms is located on average at about 3.28 Å, as shown in Figure 5.

The straight line on the figure connects the values of the a lattice parameters (of GaN at $x = 0$ and InN at $x = 1$) and hence represents the predictions of Vegard's law for a relaxed or bulk alloy. The measured In–M1 separations are less dependent on x than Vegard predicts.

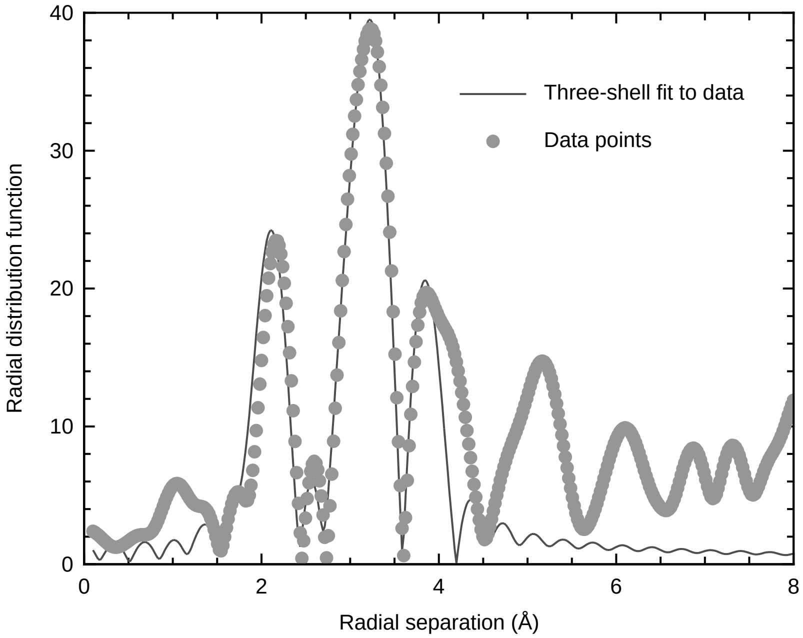

Figure 6 Fourier transform of the EXAFS spectrum yields the radial distribution function of an InGaN QB sample.

Local Structure of InGaN QW and QB

There has been a great deal of speculation in the literature concerning the local structure of InGaN alloys in the active layers of light-emitting devices. Much of this discussion paddles in the shallows of invalid comparison, as researchers extrapolate their findings from the results of measurements and calculations that can be made to those that cannot. One example is the large theoretical effort devoted to the defect physics of GaN, which is supposed somehow to 'explain' the optical properties of InGaN. ('Let us first suppose that all the horses in the race are spherical.')

We will try to avoid invalid comparisons in the brief discussion that follows. Recently, a Strathclyde University/Daresbury Laboratories/CHEA-CNRS Valbonne collaboration has succeeded in measuring the local structure of an InGaN QB sample, grown on a thick GaN template by MBE. The radial distribution function (RDF) around the target In probe atom is shown in Figure 6, together with a three-shell fit to the data.

The first shell of neighbors comprises four N atoms at 2.10(2) Å, the second shell is pure Ga at 3.22(3) Å, and the third shell pure In at 3.56(3) Å. These results suggest that there are two distinct In environments in such samples. One resembles the InGaN alloy, encountered in the epilayer samples, while the other is an environment of pure InN.

On the basis of these results, it is extremely likely that InGaN phase-separates in QB growth to form a two-component mixture comprising an In-poor and an In-rich phase. The manner in which these phases interpenetrate is a matter for speculation.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Infrared Lattice Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Gil B (ed.) (1998) *Group III Nitride Semiconductor Compounds, Physics and Applications*. Series on Semiconductor Science and Technology 6. Oxford, UK: Oxford Science Publications.

Gil B (ed.) (2002) *Low Dimensional Nitride Semiconductors*. Oxford, UK: Oxford Science Publications.

Nakamura S and Fasol G (2000) *The Blue Laser Diode*, 2nd edn. Berlin: Springer-Verlag.

Large Gap II–VI Semiconductors

J F Donegan, Trinity College Dublin, Dublin, Ireland

© 2005, Elsevier Ltd. All Rights Reserved.

Background

The chemical reaction of group IIA (Be, Mg), IIB(Zn, Cd) and group VI (O, S, Se, Te) atoms produces solid crystalline materials which are classified as semiconductors. The archetypal compound is ZnSe and this material will be used as the main example in this article. These are termed II–VI binary (involving two elements) compounds. It is also possible to grow alloy materials where more than one atom from each group is used; an example is $Zn_xCd_{1-x}Se$ where x is the proportion of Zn in the material. In general, there is a random arrangement of the Zn and Cd atoms within the crystal structure. Both binary and alloy semiconductor materials are single crystals so that the atoms are arranged in a periodic three-dimensional system called a lattice. The lattice may be represented by a unit cell; simply by repeating the pattern of the unit cell, we can generate the entire crystal lattice.

At the atomic level, the zinc (selenium) atom is bonded to 4 selenium (zinc) atoms which are arranged at the corners of a tetrahedron. The unit cell formed by this tetrahedral arrangement is termed the zincblende structure and is illustrated in Figure 1a. The dimension shown is called the lattice constant a and is characteristic of a particular II–VI semiconductor material. Some II–VI materials crystallize with a different unit cell which is termed the wurtzite structure, though again the atoms are bonded to four others in a tetrahedral arrangement as shown in Figure 1b. In this case, two lattice constants denoted a and c are required to characterize the unit cell. Some of the II–VI materials can occur in either of these two structures depending on how they were grown. The lattice constants of several II–VI compounds are given in Table 1.

II–VI Materials Growth

All of the II–VI family of semiconductors have high melting points due to the strong bonding that exists between the atomic species; values are given in

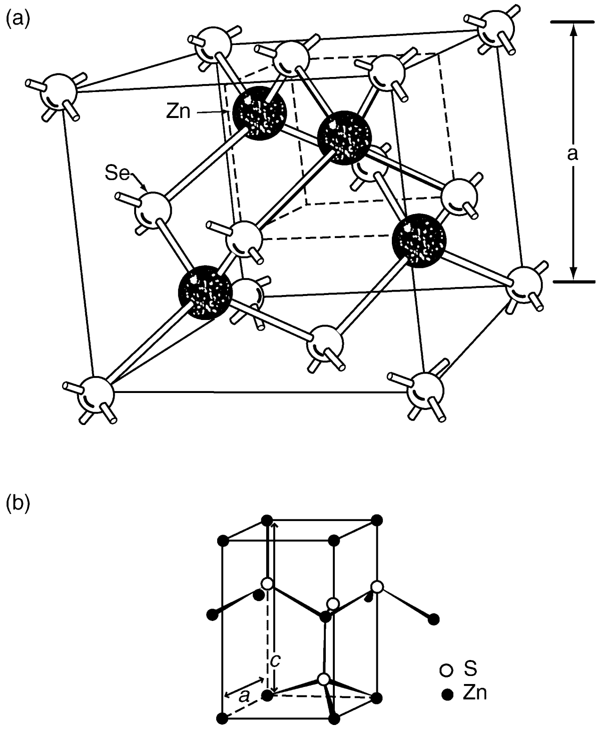

Figure 1 (a) Zincblende lattice; (b) wurtzite lattice. *a* and *c* are the lattice constants as given in Table 1.

Table 1. The crystals can be grown either by cooling a liquid (often called a melt) or by condensation from a gas (vapor). Crystal growth is a very precise and difficult process requiring the control of a large number of variables most notably temperature and pressure. As many of the II–VI materials can exist in either the zincblende or wurtzite structures, producing single crystals with only one structure is often quite difficult. Many of the crystals are 'twinned'; in a zincblende structure, a twin produces a wurtzite structure in one plane. The photograph in Figure 2 shows a ZnSe single crystal. This crystal took several days to grow indicating that very stable growth conditions must be maintained over long periods of time. The vapor phase technique takes advantage of the high vapor pressure of the II–VI compounds that allows for the growth of large single crystals at temperatures much below the melting point of these materials. Vapor growth is helped also by the congruent evaporation of the two constituents into the vapor state in a dissociated form. These two vapor constituents also recombine stoichiometrically into the solid crystal. The melting points and vapor pressures at the melting points of some II–VI semiconductors are given in Table 1.

It is also possible to grow materials in a layer-by-layer fashion onto an existing substrate; this is termed epitaxial growth. Recent developments have allowed the growth to proceed one atomic layer after the other. As a result, materials can be grown that are less than 100 Å in size. At these small sizes, the fundamental electronic and optical behavior of these materials is altered due to quantum mechanical effects. These materials are termed quantum wells. Growth temperatures in these cases are much below the melting point of these materials and allow for uniform single crystals without any 'twin' formation. In molecular beam epitaxy (MBE), atomic beams of the constituents impinge onto a heated substrate where they combine to form the crystalline material. Growth of ZnSe onto III–V substrates such as GaAs is well developed. The quality or purity of the single crystals can often be deduced from their optical properties as we will see later.

The addition of small amounts of other atoms into the crystal during the growth procedure is called doping. Semiconductors are materials whose electrical conductivity can be drastically altered by the addition of small amounts of dopants. This ability to control the electrical conductivity is the key to their use in the high-speed devices used in modern optical communications systems (*see* **Semiconductor Physics:** Impurities and Defects). In ZnSe, doping with Cl produces n-type material while doping with N produces p-type material. The ability to form a pn junction (a thin region between n and p type material) is the precursor for the production of optical devices such as light-emitting diodes (LEDs) or laser diodes (LDs) based on II–VI semiconductors. Doping also changes the optical properties of II–VI materials.

All materials are defective to some degree, and these II–VI materials are no exception. The number and type of defects are determined by thermodynamic and purity considerations during the growth process. There are native defects such as vacancies (missing atoms) and dislocations (extra or missing layers of atoms). There are also impurities present; foreign atoms due to the impurities in the starting growth materials. These impurities have an important effect on the ability of the material to emit light. Dopants and impurities may help to increase the amount of light emitted while dislocations usually reduce the emission efficiency in II–VI materials.

Optical Properties of II–VI Semiconductors: Basic Concepts

Introduction

When light is incident on a solid material, its electric field interacts with the electric field within the atoms comprising the solid. This interaction leads to a

Table 1 Structural, thermodynamic and optical properties of II–VI compound semiconductors

Material	*Lattice constant (Å)*	*Crystal structure*	*Melting point (mp) (°C)*	*Vapor pressure at mp (atm)*	*Refractive index*	*Bandgap (0 K)*	*Bandgap (300 K)*
ZnO	*a* 3.250 *c* 5.207	W	1975	22	2.0	3.44	3.37
ZnS	*a* 3.820 *c* 6.260	W	1718–1850	3.7	2.8	3.91	3.80
	a 5.410	Z			2.4	3.83	3.66
ZnSe	*a* 5.668	Z	1520	1.0	2.8	2.82	2.67
ZnTe	*a* 6.104	Z	1300	0.6	3.1	2.39	2.25
CdS	*a* 4.133 *c* 6.710	W	1405–1475	3.8	2.5	2.58	2.42
	a 5.832	Z			–	–	–
CdSe	*a* 4.299 *c* 7.015	W	1239–1264	0.4	2.6	1.84	1.74
	a 6.057	Z			–	1.77	–
CdTe	*a* 6.477	Z	–	–	2.75	1.60	1.49
MgS	*a* 5.89	Z	–	–	2.6	–	4.5
MgSe	*a* 5.62	Z	–	–	2.6	–	3.6
BeTe	*a* 5.627	Z	–	–	–	–	2.8

Data taken from many sources mentioned in the bibliography but primarily from Bhargava, R. (1997) (ed.) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers, London, Table 1 of Chapter 5.4 with permission.

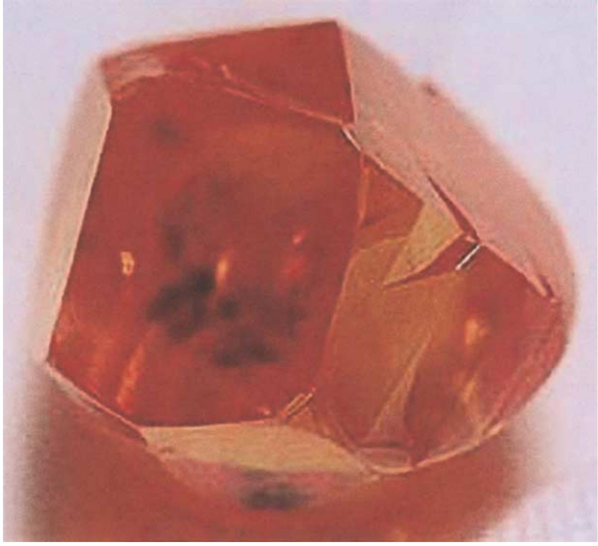

Figure 2 ZnSe single crystal grown by vapor transport. Photograph courtesy of NASA.

number of effects that we can observe. The light will be reflected or scattered from the surface of the solid, depending on whether the surface is smooth or rough on a length-scale similar to the wavelength of light. The materials of interest here are high-quality crystals with a smooth reflective surface. When light enters a material, it is refracted, i.e., it changes direction within the material. The simplest example of this is when a stick is placed in a jug of water, it appears to be bent. It is of course the light that is bent due to its different interaction with the air and with water. Snell's law states that:

$$\frac{\sin \alpha_{\mathrm{i}}}{\sin \alpha_{\mathrm{tr}}} = n \qquad [1]$$

where α_i is the angle of incidence of the light and α_{tr} is the angle of refraction. *n* is known as the refractive index of the material and is characteristic of it. As well as being refracted, any light that enters the crystal is either absorbed within the material or transmitted through the material to emerge on the other side. If the light is absorbed the material will look colored if the absorption takes place in the visible region where our eyes are sensitive. The ZnSe crystal in Figure 2 is such a material; the orange color is due to the absorption of blue light from the white light shining on the material. Other materials look transparent; ZnO is such an example. Following absorption, some materials produce light of their own; the color of the light is characteristic of the material and this process is termed luminescence. Light may also be scattered within a crystal by imperfections and by vibrations of the atoms in the crystal.

When light enters a solid material, the electric field polarizes the atoms within the structure. The positive and negative charges within the atoms move slightly from the equilibrium positions. All of the optical effects mentioned above are related to the induced polarization. As the constituent atoms in the structure are unique, the induced polarization and the resulting optical properties are specific to particular materials, and this makes optical characterization a very important tool for the investigation of semiconductor materials.

Reflection at the Air-to-Material Interface

Our experience of looking at reflection from glass tells us that these are weakly reflecting materials; the same is true of single crystal II–VI materials such as that in Figure 2. The interface between air and an optical material is a fundamental aspect of optical physics. The amount of light reflected can be determined from Maxwell's equations. The simplest equation is found where light is incident perpendicular to the surface of the material:

$$R = \frac{(n-1)^2}{(n+1)^2} \qquad [2]$$

where R is the reflection coefficient and n is the refractive index of the material. The value of n is characteristic of the particular material. Values of n for several II–VI materials are given in Table 1. The refractive index n for all these II–VI materials is not constant but varies slowly with the wavelength of the light impinging on the material. In Figure 3, we show the wavelength dependence of the refractive index of ZnSe.

Transmission and Absorption of Light within II–VI Materials

Once the light has passed through the interface with the air, it can either be transmitted through the material to emerge on the other side at a second material–air interface or be gradually absorbed within the body of the material. The ZnSe sample in Figure 2 is illuminated with white light, but it appears orange in color to our eyes, i.e., the orange component within the white light is transmitted through the material to our eye. The other components in the light, mostly green and blue, are absorbed within the material. To obtain more quantitative information, we take a spectrum of the absorption within the material. We illuminate the sample with light of different wavelengths and determine how much is transmitted or absorbed at each wavelength. The transmission spectrum of ZnSe is shown in Figure 4.

The visible spectral region extends from 0.4 to 0.7 microns and we see that there is a large change in transmission around 0.5 microns that is responsible for the color of the ZnSe crystal as seen in Figure 2. This spectral region around the onset of a large reduction in transmission at 0.5 microns is termed the bandgap of the material. As in the case of the refractive index, the value is unique to the material. II–VI materials are termed wide bandgap materials as they possess larger values of the bandgap energy than for Si, Ge and the III–V semiconductor materials. If we shine light of wavelength 0.5 microns onto ZnSe material, the transmission is very low. Under these conditions, the material would look black to our eyes. Once the light has been

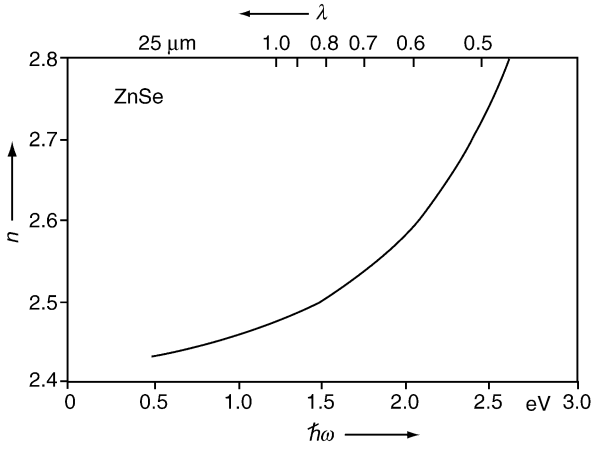

Figure 3 Refractive index n of ZnSe versus wavelength and photon energy at $T = 300\,K$. Adapted from Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II–VI and II–VII Compounds, Semimagnetic Semiconductors*, Berlin: Springer-Verlag, Figure 52, with permission.

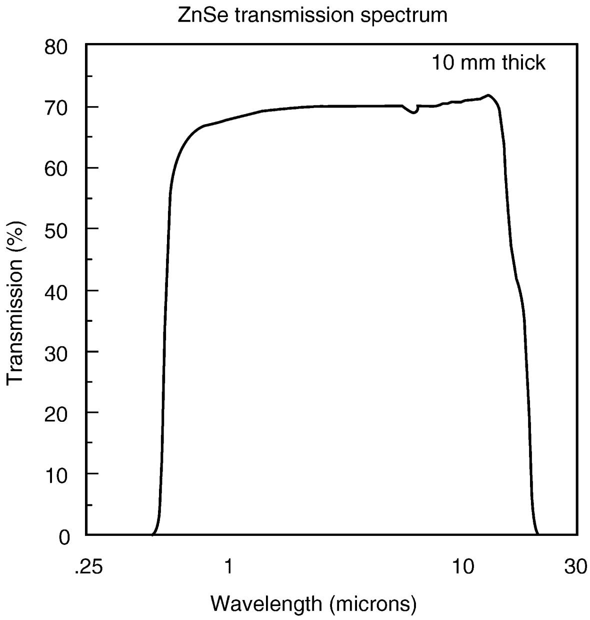

Figure 4 ZnSe transmission spectrum versus wavelength. Courtesy of II–VI Incorporated.

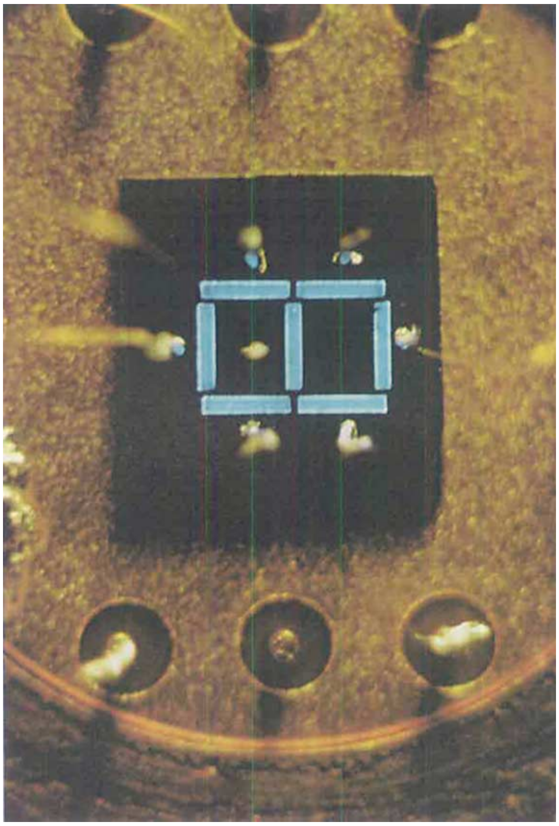

Figure 5 A seven-segment II–VI LED with blue emission taken at room temperature. Reproduced with permission from Hagerott M, Jeon H, Nurmikko AV, *et al.* (1992) Indium tin oxide as transparent electrode material for ZnSe-based blue quantum well light emitters. *Applied Physics Letters* 60: 2825–2827.

absorbed by the II–VI semiconductor, it can in turn produce light of its own. This light, known as luminescence, is again characteristic of the particular material and is the basis for the use of such materials in light-emitting diodes (LEDs) and laser diodes (LDs). In Figure 5 we show the luminescence from a II–VI LED. We see also from Figure 4 that at very long wavelengths, the ZnSe material also exhibits a large reduction in light transmission. This spectral region is termed the far infrared and the light absorbed at this spectral region produces vibrations of the atoms within the material, referred to as lattice absorption.

Light Scattering

Within the bulk of the material, small areas may have a different arrangement of atoms or may contain a substantial number of foreign atoms known as impurities. A large grouping of these impurities is known as an inclusion. In these regions, the optical properties will be different from the bulk of the material. Differences in both absorption coefficients and refractive indices can be found. These different refractive indices cause the light passing through the material to be scattered. Poor-quality material will often contain many inclusions. Even to the eye, these crystals will not look highly transparent. Changes in the refractive indices can also occur in highly pure materials due to vibrations of the atoms in the crystals. The lattice vibrations change the distance between the atoms in the material which alters the electric fields within the material and thereby changes the optical properties. Light can also be scattered due to these vibrations; experimental measurements of light scattering can provide useful information on the vibrations in these materials.

Optical Properties of II–VI Semiconductors: More Advanced Concepts

Light–Semiconductor Interactions

Light is an oscillating electromagnetic field whose interaction with semiconductor materials is in terms of the complex dielectric constant $\varepsilon_c = n_c^2$ where n_c is the complex refractive index defined as $n_c = n - ik$. k is known as the extinction coefficient and the value of k determines the amount of absorption in a material. The optical processes of reflection and absorption in semiconductors are governed by this complex dielectric constant. It turns out that there is a strong relationship between these processes. If the absorption spectrum is determined at all frequencies, then we can determine the refractive index spectrum at all frequencies by what is known as the Kramers–Kronig relations. From this description, we can now refine eqn [2] presented above:

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \qquad [3]$$

If the material is transparent at a particular wavelength, then $k = 0$ and we can determine the refractive index n. This is the case for values given in the table. The technique of spectroscopic ellipsometry can be used to determine the optical response of materials over a large wavelength range. This technique relies upon the change from linear polarization to elliptical polarization for light reflected from a material. From these measurements, the dielectric response can be determined.

If we look again at Figure 3, we see that there is a strong change in sample transmission around 0.5 μm. This corresponds to an energy of 2.18 eV and is the bandgap energy of ZnSe. According to the details presented above, transitions are now being made from the full valence to the empty conduction bands of the semiconductor material. Once we have exceeded the bandgap energy, we can make further transitions between the bands. The spectrum shown in the figure is a broad band, reflecting the change from sharp atomic transitions to broad transitions in the solid state. Light with a wavelength greater than 0.5 μm does not have sufficient energy to effect the transitions between the bands. These wavelengths are not absorbed in the material. The material is transparent and the value of the extinction coefficient k is zero. The orange color of the crystal in Figure 2 corresponds to this transparent spectral region.

Cooling samples to low temperatures close to absolute zero has a profound effect on the optical properties of materials, particularly the absorption spectrum. In Figure 6, we show the absorption spectrum of a ZnSe sample at a temperature of 2 K. The absorption spectrum is determined from the measured transmission by the following equation

$$\alpha = \frac{1}{L} \ln \frac{I_0}{I_t} \qquad [4]$$

where α is the absorption coefficient, L the length of the sample, I_0 the light intensity incident on the sample and I_t the light intensity transmitted through the sample. In addition to the broad band spectrum there are now a number of sharp spectral lines near the bandgap energy.

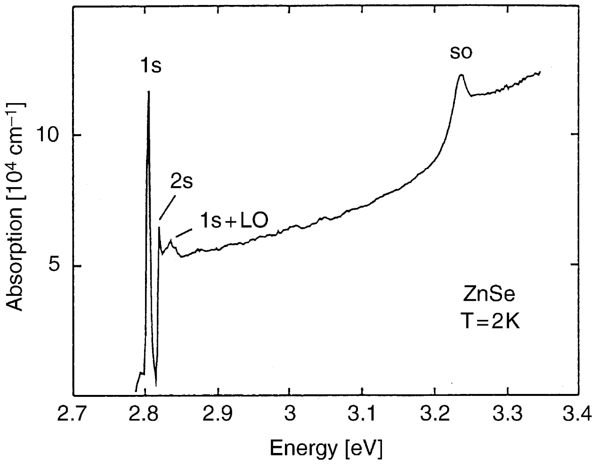

Figure 6 Low-temperature optical absorption spectrum of a ZnSe layer showing the 1S and 2S exciton features, the 1S + 1LO-phonon transition and the split-off (SO) exciton. Reproduced with permission from Bhargava R (ed.) (1997) *Wide Bandgap II–VI Semiconductors*, The Institute of Electrical Engineers: London, Figure 2 of Chapter 4.1.

These lines are due to exciton formation. An exciton is an electron–hole pair bound together by their mutual Coulomb interaction. It is the solid state equivalent of the Hydrogen atom (*see* **Semiconductor Physics:** Excitons). The bandgap energy at 2 K has increased from its value at 300 K. This is due to the thermal changes in the interatomic distance and the vibrational properties of the materials. Values of the bandgap energies determined from the position of the exciton lines in absorption spectra at low and at room temperatures are given in Table 1. At low temperatures, the luminescence is dominated by excitons, and we observe sharp line transitions. The width of the exciton transition is an excellent indicator of the quality of the material, and is often used by crystal growers to determine the quantity of defects in their materials.

It is possible for excitons and for free carriers (those that are not bound together) to recombine without giving out light. This nonradiative process usually involves the creation of local vibrations of the crystal lattice in the vicinity of defects in the material. We can define the radiative efficiency η_r as follows

$$\eta_r = \frac{R_r}{R_r + R_{nr}} \qquad [5]$$

where R_r is the radiative transition rate and R_{nr} is the nonradiative transition rate. In order to develop new optical devices based on II–VI semiconductors, it is essential that the value of R_{nr} be as small as possible. As R_{nr} is directly related to the concentration of defects in these materials, one can see why it is essential to make semiconductors with as few defects as possible. We can study these effects using a pulsed light source. Carriers are created during the pulse of light and recombine after the pulse has stopped. If we measure the luminescence intensity versus time, we can deduce the lifetime, which is given as

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}} \qquad [6]$$

where τ_r is the radiative and τ_{nr} the nonradiative lifetime, respectively. By varying the carrier density we can separately determine the contribution to the lifetime from radiative and nonradiative effects. In II–VI materials at room temperature, lifetimes are of the order of 1 ns.

Quantum Wells, High Carrier Density Effects, and Lasers

Recent developments in the growth of semiconductor materials has allowed the production of samples with monolayer accuracy (3–5 Å). If we grow a layered structure with materials of high bandgap on either side of a material of lower bandgap, we can observe quantum mechanical effects. These materials are called quantum wells (*see* **Semiconductor Physics:** Quantum Wells and GaAs-Based Structures). As an example, a 75 Å ZnSe quantum well can be grown between layers of ZnSSe. Due to the small size of the active material, quantum mechanical effects come into play, most notably an increase in the bandgap energy as the width of the quantum well decreases. The effect of placing this material between ZnSSe layers that have a larger bandgap is to confine the electrons and holes to the small region of the well. In this small region, they are less likely to encounter defects and so quantum well materials usually have a higher radiative efficiency than bulk II–VI materials which are grown by standard techniques. In Figure 7, we show the exciton absorption and luminescence from a ZnSe quantum well at low temperature.

If we increase the excitation level in these materials, we can eventually cause them to operate as lasers. At high carrier densities, nonradiative transitions become saturated and radiative efficiency increases. As the refractive indices in II–VI materials are quite large, the reflectivity of the two ends of the sample are usually sufficient to act as the mirrors for the laser cavity. All the II–VI semiconductor materials have exhibited laser action due to their direct bandgap. Most of the early studies of lasing were carried out at low temperature. In recent times, lasing action has been observed at room temperature in CdZnSe quantum well materials and there are attempts at present to develop commercial II–VI lasers for use in optical storage devices and flat panel displays. In Figure 8 we show the transition from luminescence through gain and into lasing action for a CdZnSe quantum well material.

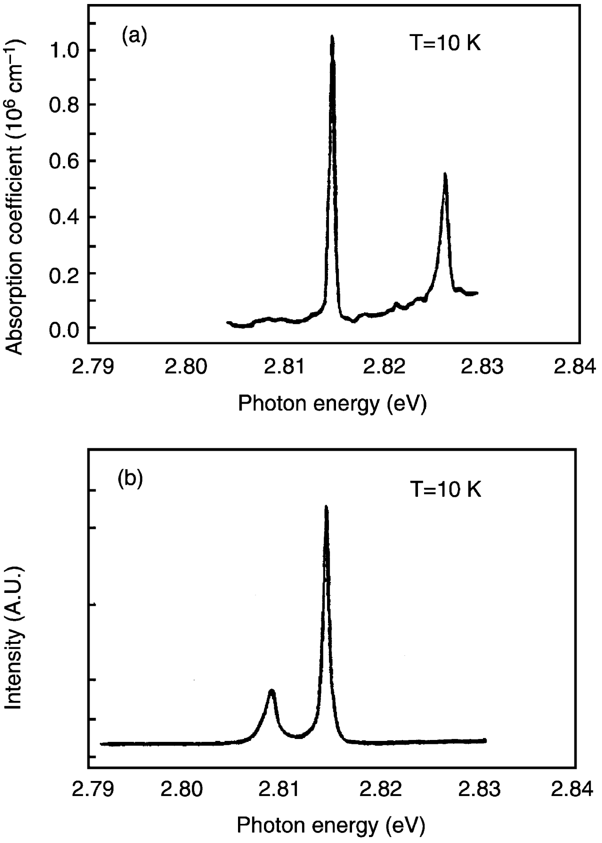

Figure 7 (a) Absorption and (b) luminescence from a single ZnSe quantum well at $T = 10$ K in the vicinity of the $n = 1$ quantum well exciton transition. Reprinted from Nurmikko AV, Jeon H, Gunshor RL and Han J (1996) II–VI lasers – directions. *Journal of Crystal Growth*, 159: 644–651 with permission from Excerpta Media Inc.

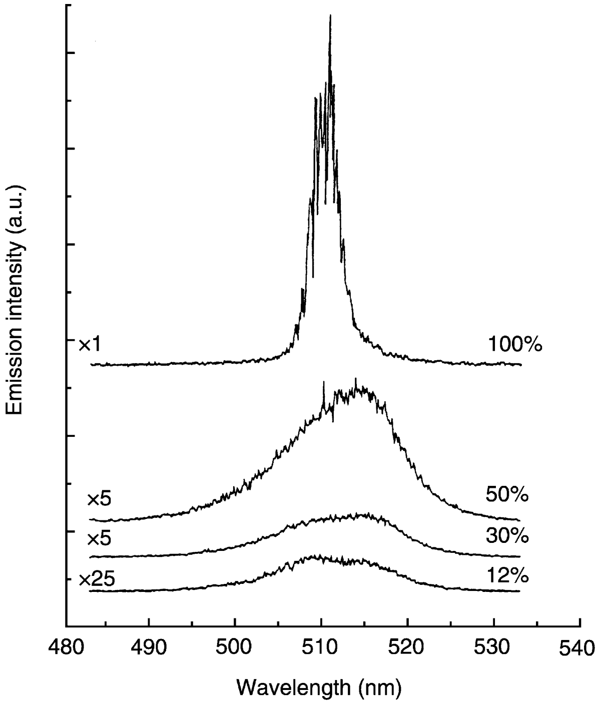

Figure 8 Emission spectra above and below lasing for a 40 Å $Cd_{0.28}/Zn_{0.72}Se$/ZnSe system at room temperature. 100% corresponds to 1.1 times lasing threshold. Reprinted from Donegan JF, Jordan C, Rees P, Logue F, Heffernan JF and Hegarty J (1996) A study of internal loss in CdZnSe/ZnSe multiple quantum well materials. *Journal of Crystal Growth*, 159: 653–656 with permission from Excerpta Media Inc.

Dopants and Impurities

The key to the large-scale use of semiconductor materials is the large change in the electrical conductivity of the materials with the incorporation of small amounts of other elements. This process is known as doping and the foreign elements are known as dopants. The recent development of II–VI laser diodes in ZnSe materials is based on the p-n diode junction. Electrons injected into the n-type material recombine with holes injected into the p-type material at the junction between these two materials. A particular type of luminescence transition unique to semiconductor materials is the donor–acceptor pair (DAP) recombination. As donor and acceptor atoms are located in specific crystalline sites with respect to one another, we observe that the recombination energy is dependent on separation. Pairs that are close together often exhibit sharp lines while those far apart merge into a broad band. In Figure 9 we show the DAP recombination in ZnSe at 1.6 K.

It is also possible to incorporate foreign atoms in II–VI materials that have a small effect on the electrical properties but significantly change the optical properties. Such foreign atoms are termed impurities to distinguish them from dopants. An example is Mn^{2+} in ZnS. The Mn ions take the place of a small number of Zn atoms in the lattice. The luminescence from such a material is a broad emission in the visible region due to optical transitions on the Mn ions. ZnS:Mn materials are used as phosphors in flat-panel displays.

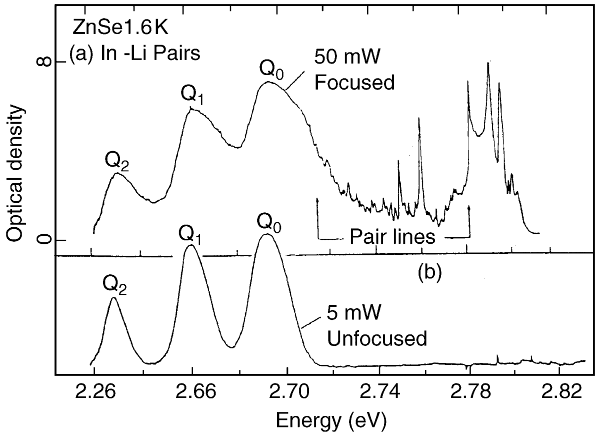

Figure 9 DAP recombination spectra of In-Li pairs in ZnSe showing pair lines, pair band (Q_0) and LO-phonon replicas (Q_1 and Q_2) at two different intensities (a) and (b). Adapted with permission from Merz JL, Nassan K and Shiever JW (1973) Pair spectra and the shallow acceptors in ZnSe. *Physical Review B* 8: 1444–1452.

Effects of Temperature, Pressure, Electric, and Magnetic Fields

The optical properties of II–VI materials are subject to variation due to temperature changes, external pressure and fields. Increasing temperature is found to lead to a decrease in the bandgap of II–VI materials. This decrease is due to the combined effects of an increase in the lattice spacing with temperature and the electron–phonon interaction. As the population of phonons is temperature dependent, this effect increases strongly as temperature increases and is the major contributor to the change in bandgap with temperature. A second effect of increasing the temperature is the change in the luminescence properties. In particular, it is found that the luminescence efficiency decreases and this is reflected in a reduced luminescence lifetime at room temperature compared with low-temperature values. Looking again at eqn [6], it is the nonradiative lifetime that changes rapidly with temperature and results in the decrease in the luminescence efficiency.

Pressure changes may be hydrostatic or uniaxial. In the hydrostatic case, the atoms are all brought closer together, and for the II–VI materials, this results in an increase in the bandgap energy. On the other hand, a uniaxial stress results in a decrease in the bandgap in the direction of the stress and an increase in the transverse directions. In II–VI materials with a wurtzite structure, the situation is more complex with two longitudinal and four transverse deformations possible.

External electric and magnetic fields can also be applied to II–VI semiconductors and thereby the optical properties can be changed. Electric fields polarize the materials and can lead to a reduction in the optical strength of transitions due to this polarization-induced carrier separation. Magnetic fields, on the other hand, lead to a splitting of the conduction and valence bands into Landau levels due to the circular motion of the carriers around the direction of the field.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Adachi S and Taguchi T (1991) Optical Properties of ZnSe. *Physical Review B* 43: 9569–9577.

Bhargava R (ed.) (1997) *Wide Bandgap II-VI Semiconductors*. London: Institute of Electrical Engineers.

Madelung E (ed.) (1982) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 17b: Physics of II-VI and I-VII Compounds, Semimagnetic Semiconductors*. Berlin: Springer-Verlag.

Madelung E (ed.) (1987) *Numerical Data and Functional Relationships in Science and Technology, New Series, Group III: Crystal and Solid State Physics, vol. 22a: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds*. Berlin: Springer-Verlag.

Mar H (1992) Optical properties of wide bandgap II-VI semiconductors. In: Ruda HE (ed.) *Widegap II-VI Compounds for Opto-electronic Applications*, pp. 167–188. London: Chapman and Hall.

Nurmikko AV and Gunshor RL (1993) Physics and application of II-VI quantum confined structures. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 229–262. Berlin: Akademie Verlag.

Pankove J (1975) *Optical Processes in Semiconductors*, 1st edn. New York: Dover Publications.

Stanley RP and Hegarty J (1993) Physics of excitons in wide bandgap II-VI quantum wells. In: Henneberger F, Schmitt-Rink S and Göbel EO (eds) *Optics of Semiconductor Nanostructures*, pp. 263–290. Berlin: Akademie Verlag.

Yu PY and Cadona M (1996) *Fundamentals of Semiconductors*. Berlin: Springer Verlag.

II–VI Compounds 1999. Ninth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 214–215, 2000.

II–VI Compounds 1997. Eighth International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 184–185, 1998.

II–VI Compounds 1995. Seventh International Conference. Elsevier, Netherlands. *Journal of Crystal Growth* vol. 159, 1996.

Lead Salts

G Bauer and G Springholz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The binary IV–VI compounds formed from Pb and group-VI elements like S, Se, and Te are among the oldest known semiconducting materials, and have been used for many years for electronic and optoelectronic device applications. The first ever reported solid state diode was made from single crystalline PbS and its rectifying properties were exploited in early radio receivers. Later on, interest in IV–VI semiconductors shifted to mid-infrared opto-electronic device applications such as photon detectors operating in the 3–14 μm wavelength range, taking advantage of the narrow energy gap between the valence and conduction band of the IV–VI compounds. Soon after the demonstration of lasing in III–V semiconductor diode structures, the first mid-infrared p-n junction laser was reported using $Pb_{1-x}Sn_xTe$, and since then efficient mid- and far-infrared IV–VI compound diode lasers have been fabricated, finding their main applications for remote sensing of gaseous pollutants in trace gas sensing devices, for toxic gas analysis systems, for human breath analysis in medical diagnostics, and for fabrication process control. Recently, IV–VI multiquantum-well structures have also attracted a lot of attention for their potential as efficient thermoelectric devices.

Basic Properties

In lead salt compounds, both the valence band maximum and the conduction band minimum occur at the L-points of the Brillouin zone. Thus these many-valley semiconductors have a direct gap, but not in the center of the Brillouin zone. In addition the small values of the energy gaps ($E_g < 410\,\mathrm{meV}$ at $T = 300\,\mathrm{K}$) lead to a strong $\mathbf{k}\cdot\mathbf{p}$ interaction which in turn results in a strong nonparabolicity of the $E(k)$ dispersion relations. The small energy gaps also lead to small effective masses of electrons and holes, and the strong spin–orbit interaction (Pb is a very heavy element) leads to large values of the g-factors.

The bandgaps of the lead compounds have a positive temperature coefficient ($\mathrm{d}E_g/\mathrm{d}T > 0$) and a negative pressure coefficient ($\mathrm{d}E_g/\mathrm{d}p < 0$), i.e., the gaps increase with temperature and decrease with hydrostatic pressure. These anomalous properties are related to peculiarities of the band structure, among them the fact that in the lead compounds the Pb 6s levels are below the top of the valence band, and thus are occupied, in contrast to the group IV, III–V and II–VI compounds.

Almost all optical device applications of the IV–VI semiconductors are based on the materials subgroup of cubic lead chalcogenides PbTe, PbSe, and PbS, and their quasibinary alloys with other chalcogenide such as compounds, like PbSnTe, PbSnSe, or ternary alloys like PbEuTe, PbSrTe, PbEuSe, PbSrSe, etc. These compounds all crystallize in the rocksalt structure with sixfold coordination and lattice constants in the 5.9 to 6.5 Å range. Mechanically, the lead salt compounds are much softer than the tetrahedrally bonded semiconductors, which has significant implications for device fabrication processes.

A further unusual property of the lead compounds is their huge static dielectric constant ε_0. In particular PbTe is close to a structural phase transition from the NaCl structure to a rhombohedral modification with ferroelectric properties. GeTe and SnTe are indeed ferroelectric with Curie temperatures of about 700 K and 145 K respectively, and also PbGeTe and PbSnTe become ferroelectric when the Ge or Sn content exceeds 1% or 20%, respectively. PbTe exhibits a softening of the transverse optic phonon mode with decreasing temperature. Although PbTe remains cubic, the temperature dependence of the static dielectric constant $\varepsilon_0 \propto C/(T-\theta)$ can be associated with an extrapolated negative Curie temperature θ of −75 K. As a consequence, the static dielectric constant increases strongly with decreasing temperature, reaching a value of $\varepsilon_0 = 1350$ at 4 K. This is the reason for the strong increase of the static dielectric constant at low temperatures as shown in Table 1, in which important properties of PbTe, PbSe, and PbS are summarized.

Also, the high-frequency dielectric constants ε_∞ of the lead salts are extremely high, which leads to refractive indices of about 4.4 for PbS, to 5.0 for PbSe and to about 6.0 for PbTe at $T = 300$ K, for photon energies just below the respective energy gaps.

As a result of the huge static dielectric constants, the hydrogen-like impurity binding energies are in the range of μeV, and the carrier scattering at ionized impurities is effectively screened. Consequently, the low-temperature carrier mobilities are high, exceeding 10^6 cm^2/Vs in PbTe. For doping of the lead salts two alternative methods are used, namely (1) adjustment of the stoichiometry of the compounds, or (2) doping with various kinds of impurities. Because the electronic states of these vacancies are resonant within the conduction or valence bands, each metal (Pb or Sn) vacancy acts as a doubly charged acceptor, and each chalcogen (Te, Se, or S) vacancy acts as a doubly charged donor. As a result, background doping levels below 10^{17} cm^{-3} are difficult. For doping with impurities, atoms from group IIIa (In, Ga, Tl), group Va (As, Sb, Bi), as well as group Ib elements (Cu, Ag) are used. These dopants are amphoteric p- or n-type, depending on whether they are incorporated substitutionally on metal or chalcogen lattice sites.

Table 1 Physical properties of some of binary IV–VI semiconductors

Parameter	*PbTe*	*PbSe*	*PbS*
Lattice constant a_0 (Å)	6.462	6.124	5.936
Melting point (°C)	930	1080	1113
Energy gap E_g			
at 300 K (meV)	319	278	410
at 77 K (meV)	217	176	307
at 4 K (meV)	190	147	286
Static dielectric constant ε_0			
at 300 K	414	210	169
at 77 K	1000	227	181
at 4 K	1350	280	–
High-frequency dielectric constant ε_∞	34	21	17
TO phonon mode frequency	18 (4 K) 33 (300 K)	37	65

Table 2 Expansion coefficients (10^{-6} K^{-1})

Compound	*Expansion coefficient*
PbS	20.3
PbSe	19.4
PbTe	19.8
GaAs	6
Si	2.6
BaF_2	18.8

Other important aspects of the lead salts are their rather large linear thermal expansion coefficients of around 20×10^{-6} K^{-1} which are about a factor of three larger than that of GaAs, and six times larger than that of Si as shown in Table 2. The thermal conductivity is much lower than that of other semiconductors, which is favorable for their use in thermoelectric devices but causes problems in the removal of the dissipated heat in IV–VI compound diode lasers.

Band Structure

A main property of the electronic band structure of the lead salt compounds is the small and direct energy gap of less than 410 meV (see Table 1). Since the maxima of the valence band and the minima of the conduction band are located at the same wavevector in reciprocal space, optical photons can be directly absorbed or emitted at the band edges. This makes the lead salts very well suited for opto-electronic applications such as infrared photon detectors or emitters. The band extrema are located at the edges in the <111> direction of the Brillouin zone (L-points). In contrast to most other semiconductors, the conduction and valence bands are nearly mirror symmetric with almost equal effective masses for the electrons and holes. Because of the narrow energy gaps, the energy bands are strongly nonparabolic and the effective masses of electrons as well as holes are rather small (0.02–0.08 m_0 for the transverse masses,

Table 3 Band edge values of effective masses (in units of m_0) and g-factors (absolute values), $T = 2$ K

Parameter	*PbTe*	*PbSe*	*PbS*
m_t^{CB}	0.0207	0.037	0.081
m_l^{CB}	0.0213	0.0687	0.108
m_t^{VB}	0.0241	0.036	0.076
m_l^{VB}	0.263	0.066	0.108
g_t^{CB}	16.7	32.5	–
g_l^{CB}	66.1	44.1	12
g_t^{VB}	13.5	30.6	–
g_l^{VB}	65.6	–	13

Table 4 Pressure and temperature coefficients of the energy gaps

	PbTe	*PbSe*	*PbS*
dE_g/dp(eV Pa^{-1})	-7.5×10^{-11}	-9.1×10^{-11}	-9.1×10^{-11}
dE_g/dT (meV K^{-1})	0.457	0.457	0.462

see Table 3). In addition, the bands are anisotropic, i.e., the Fermi surfaces are elongated ellipsoids of revolution around the $<111>$ axes, characterized by a longitudinal and transverse effective mass m_l and m_t parallel and perpendicular to the four-fold $<111>$ directions.

For PbTe the effective mass anisotropy is very large with a ratio of about 10 between longitudinal and transverse masses, whereas for PbSe and PbS this ratio is only 1.8 and 1.2, respectively (Table 2). Furthermore, in the alloy system $PbSe_{1-x}Te_x$ the energy gap shows a bowing, but not with a minimum value but rather a maximum gap for x at about 0.7. This is in contrast to the behavior found in direct-gap zincblende semiconductors: the bandgap $E_g(x)$ of an $A_{1-x}B_x$ alloy can generally be represented by $E_g(x) = (1-x)E_g(A) + xE_g(B) - bx(1-x)$ where b denotes the bowing coefficient. For III–V and II–VI compounds $b > 0$; whereas, for PbSeTe alloys $b < 0$.

Apart from the above-mentioned negative pressure coefficients of the direct gaps and their positive temperature coefficient which are tabulated in Table 4, there is a further intriguing anomaly in the chemical trends of the fundamental gap at the L-points: it decreases from PbS to PbSe, but increases from PbSe to PbTe, i.e., the direct gaps do not decrease monotonically with increasing atomic number of the anions. This decrease of the energy bandgap with decreasing temperature is substantial: i.e. it decreases from room temperature to 4 K by about 130 meV (see Table 1). This represents a large relative change with respect to the absolute value of the bandgap, and thus the emission of lead salt based diode lasers can be tuned over a broad wavelength range just by changing the operation temperature. This effect is actually utilized for spectroscopic applications of these lasers.

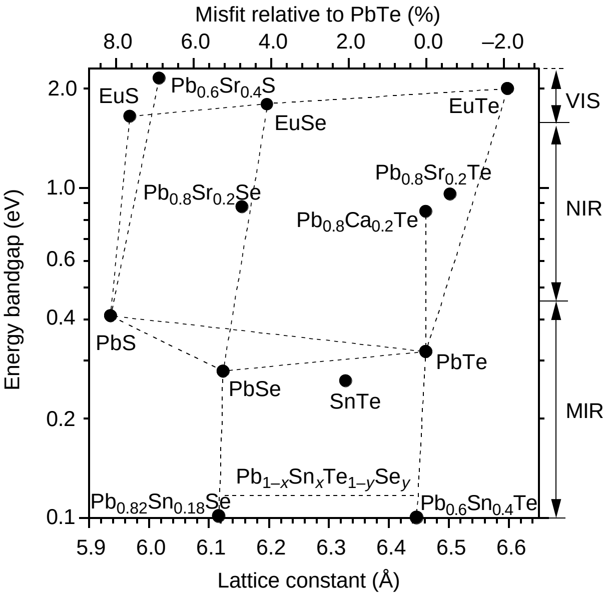

Figure 1 Energy bandgap versus lattice constant for several IV–VI compounds and alloys.

For adjustment of the fundamental absorption edge of infrared detectors as well as the adjustment of the wavelength emission of lead salt diode lasers, alloying of the lead salt compounds with other chalcogenide compounds is used. As illustrated in Figure 1, for this purpose many different ternary and quaternary alloys have been used. Important alloy systems for far-infrared applications are the lead–tin chalcogenide alloys. The tin chalcogenides (SnS, SnSe, SnTe) form single-phase pseudobinary alloy systems with the lead salts for all compositions. They have similar band structures as compared to the lead salt compounds; however, the arrangement of valence and conduction bands is exchanged. As a result, with increasing Sn content the energy bandgap of the ternary lead–tin chalcogenide alloys decreases and a zero bandgap semiconductor is reached for a certain ternary composition. For an operation temperature of 77 K, the bandgap is zero for a Sn content of $x_{Sn} \sim 40\%$ in the case of $Pb_{1-x}Sn_xTe$, and of $x_{Sn} \sim 20\%$ in the $Pb_{1-x}Sn_xSe$ case.

Optical Properties

The strong fundamental absorption in the lead salt compounds reflects the particularly high joint density of states at the direct gap at the L-points of the Brillouin zone. The main advantage of the lead compounds with respect to other direct narrow-gap semiconductors like InSb, or HgCdTe alloys, with band extremes at the center, i.e., the Γ-point of the Brillouin zone, is related to the differences in the effective masses. Whereas the transverse effective masses of the IV–VI compounds have comparable values to the effective masses of the narrow-gap III–V and II–VI compounds, the longitudinal effective masses of the lead compounds are much larger. Thus the mass anisotropy together with the four-fold valley degeneracy at the L-point of the Brillouin zone leads to the substantially higher joint density of states and corresponding large absorption constants. Due to the rapid change of $\alpha(\omega)$ in the fundamental absorption range, a peak appears in the refractive index, $n(\omega)$, as shown in Figure 2 for PbTe, PbSe and for PbS. The shift of the peak with increasing temperature to higher photon energies reflects the increasing energy gaps.

In two-dimensional structures the changes of the absorption constant with frequency are particularly steep for energies which correspond to the onset of absorption between quantum confined electric subbands in the valence and conduction band. Such step-like features have indeed been observed in the absorption of PbTe/PbEuTe multiquantum-well (MQW) structures by Yuan *et al.* In Figure 3 the absorption constants and refractive indices are shown for a series of three MQW samples with different PbTe well widths. The steps in the absorption coefficients for the interband transitions associated to the longitudinal valleys with their main axis parallel to the [111] growth direction, denoted by 'l' as well as those associated by the oblique valleys, denoted by 'o' are accompanied by corresponding cusp-like enhancements of the refractive index. These cusps are of importance for the proper design of MQW laser structures.

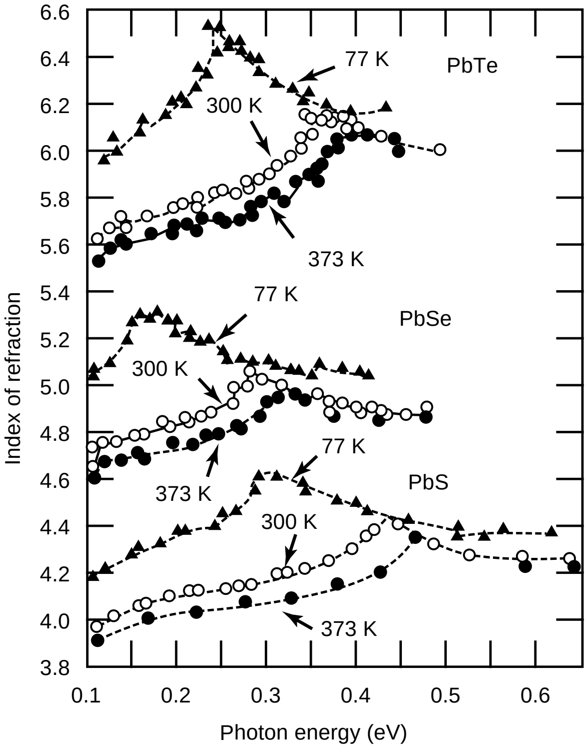

Figure 2 Refractive indices of PbTe, PbSe, and PbS as a function of photon energy at 77, 300, and 373 K. Reproduced with permission from Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337. Copyright 1965, American Physical Society.

Because of the extremely high values of the refractive indices, the lead salts can easily be combined with materials of much lower refractive indices like EuTe, EuSe, or BaF_2 for efficient multilayer Bragg mirror structures. The high reflectivity of Bragg mirrors is caused by constructive interference of electromagnetic waves reflected at the subsequent interfaces of a multilayer structure. To obtain constructive interference, all interfaces have to be parallel, and, depending on the refractive indices of the media inside the Bragg mirror, the distance between subsequent interfaces should be an even or odd multiple of one quarter of the optical wavelength λ. In practice, stacks of two alternating layers with different refractive indices n_1 and n_2 and with a thickness equal to a quarter optical wavelength are used.

If for these $\lambda/4$ layers $Pb_{0.93}Eu_{0.07}Te$ and EuTe are used, a high index contrast of 68% is obtained and therefore very high reflectivities can be achieved already by a very small number of layer pairs. It was shown that a Bragg mirror with 3.5 periods of $EuTe/Pb_{0.93}Eu_{0.07}Te$ layer pairs for a target wavelength of 3.8 μm has only a total thickness as small as 2.55 μm. For comparison with results obtained for other material combinations it is useful to relate the stop bandwidth $\Delta\lambda$ to the center wavelength λ_T of the Bragg mirror. For $EuTe/Pb_{0.93}Eu_{0.07}Te$ this relative stop bandwidth $\Delta\lambda/\lambda_T$ is as large as 50%. An even wider stop bandwidth of 59% was demonstrated for a mirror with PbTe and EuTe $\lambda/4$ layers. This value represents, to the best of our knowledge, the highest relative stop bandwidth obtained for molecular beam epitaxy (MBE) grown Bragg reflectors.

Infrared Lasers

Among the infrared diode lasers, those based on IV–VI compounds play a particularly important role. They cover a wavelength range from 2.5 μm to about 30 μm. In fact, up to now most commercially

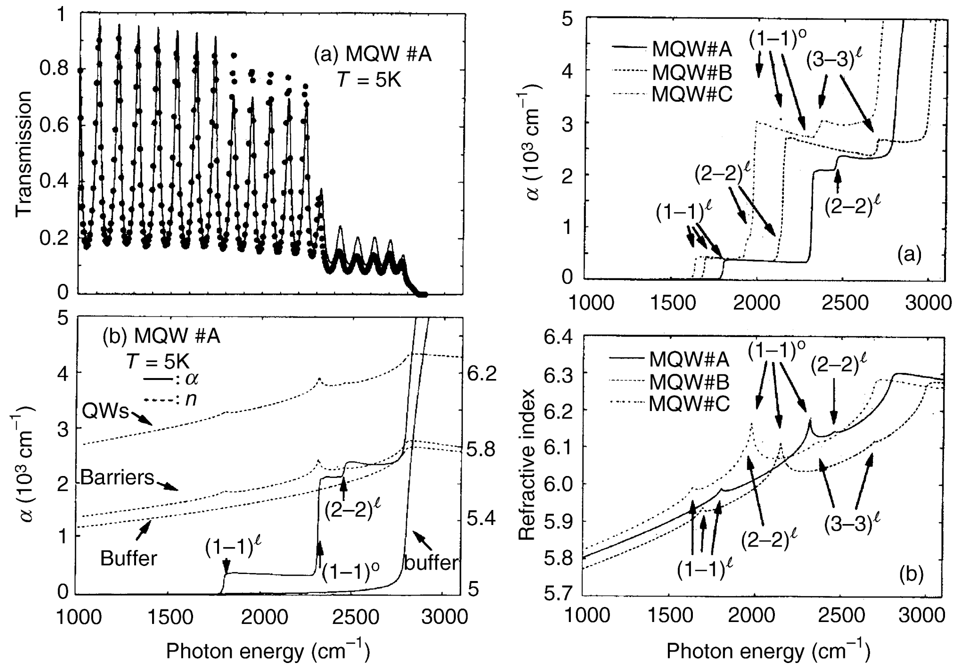

Figure 3 Left panel: (a) transmission versus energy for a PbTe/Pb$_{1-x}$Eu$_x$Te multiquantum-well sample A (6.2 nm/62.1 nm, $x = 0.026$, 46 periods). Dots: experimental data; full line: calculation based on an envelope function model. (b) Frequency dependence of the absorption constant (full line) and of the refractive index (dotted lines). The arrows indicate interband transitions between electric sub-bands associated with valleys with main axis parallel to the [111] growth direction (l) and oblique to it (o). Right panel: (a) Comparison of the absorption constant of three PbTe/PbEuTe MQW samples (B: 9.3 nm/55.4 nm, $x = 0.03$, 30 periods; C: 11.8 nm/48.6 nm, $x = 0.024$, 30 periods). (b) Refractive index versus energy. Steps in the absorption constant due to onset of interband transitions are associated with cusps in the refractive index. Reproduced with permission from Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PbTe/Pb$_{1-x}$Eu$_x$Te multiple-quantum-well structures. *Physical Review B* 49: 5476–5489, Copyright 1994, American Physical Society.

available mid- and far-infrared laser diodes are made from IV–VI compounds. The main material properties which are advantageous in this respect are (i) the electronic band structure with mirror-like bands at the L-point of the Brillouin zone and the resulting high joint density of states; (ii) the lower nonradiative Auger recombination rate in IV–VI compounds as compared to III–V and II–VI materials, and (iii) the high dielectric constant and the resulting weak detrimental influence of defects on nonradiative recombination losses. At wavelengths longer than 3 μm, IV–VI interband diode lasers are superior to their III–V counterparts for reaching high operating temperatures. So far IV–VI compound lasers hold extreme high cw operating temperature at wavelengths larger than 3 μm. With PbSe/PbSrSe diode lasers pulsed operation up to 60 °C was reported, whereas cw operation up to 223 K was achieved with separate confinement buried PbTe/PbEuSeTe heterostructure laser. Embedding PbTe quantum wells as the active medium in PbEuTe barriers between two dielectric Bragg mirrors optically pumped pulsed lasing was observed up to 65 °C. The recently developed III–V based quantum cascade lasers are rapidly improving and are becoming important competitors for IV–VI lasers. The commercial IV–VI lasers, however, are still based on conventional designs like homojunction, heterojunction, or single and multiple heterojunction lasers.

A unique property of lead salt laser diodes is the ease of tuning of the emission wavelength due to the strong change of bandgap of the IV–VI materials with increasing temperature. This, together with their narrow linewidths, makes these lasers ideally suited for spectroscopic applications, including research in molecular spectroscopy of ions or radicals in heterodyne and acoustic spectroscopy, or Doppler spectroscopy of molecular beams. So far the main applications are in time resolved combustion analysis in the automotive industry, in the monitoring of emissions of power plants, in exact monitoring of gas compositions in chemical plants, and spatially and time resolved *in situ* analysis.

Apart from the significant advances with IV–VI edge emitting diode lasers, vertical cavity surface

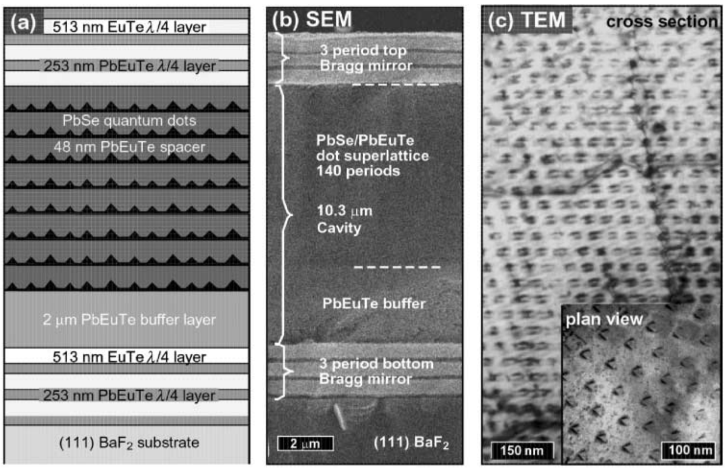
(a)
513 nm EuTe λ/4 layer
253 nm PbEuTe λ/4 layer
PbSe quantum dots
48 nm PbEuTe spacer
2 μm PbEuTe buffer layer
513 nm EuTe λ/4 layer
253 nm PbEuTe λ/4 layer
(111) BaF2 substrate
(b) SEM
3 period top Bragg mirror
PbSe/PbEuTe dot superlattice 140 periods
10.3 μm Cavity
PbEuTe buffer
3 period bottom Bragg mirror
2 μm
(111) BaF2
(c) TEM
cross section
plan view
150 nm
100 nm

Figure 4 (a) Schematic cross-section of a PbSe quantum-dot VCSEL structure; (b) corresponding cross-sectional scanning electron micrograph; and (c) cross-sectional and plan-view transmission electron micrographs of a PbSe/PbEuTe dot superlattice sample reference sample. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

emitting lasers (VCSELs) have been successfully produced with optically pumped pulsed mode operation up to 50 °C. These lasers have attractive features like low-divergence circular output beam, low active volumes, and low thresholds.

High-finesse vertical cavity structures have also made it possible to realize mid-infrared surface-emitting PbSe/PbEuTe quantum dot lasers. As shown in Figure 4 in this case a self-organized PbSe/PbEuTe quantum dot superlattice was introduced into the active cavity region between Bragg mirrors consisting of three λ/4 layer pairs of 513 nm EuTe alternating with 253 nm $Pb_{0.94}Eu_{006}Te$. Optically pumped VCSEL emission at $\lambda = 4.2–3.9$ μm was achieved up to 150 K. The PbSe dots exhibit a face-centered cubic three-dimensional stacking in the superlattice. Lasing occurs simultaneously at the $m = 28^{\text{th}}$ and 29^{th} order cavity modes, which shifts with increasing temperature to the $m = 29^{\text{th}}$ and $m = 30^{\text{th}}$ mode as shown in Figure 5. This simultaneous stimulated two-mode laser operation is a result of the inhomogeneously broadened quantum-dot gain spectrum, due to PbSe dot size fluctuations.

Infrared Detectors

The photoconductive properties of PbS and PbSe have been used for the detection of infrared radiation already since about 1930. These detectors are still mass produced either by vacuum deposition or precipitation from aqueous solution in the presence of an oxidizing agent. The polycrystalline films deposited on glass or similar substrates are

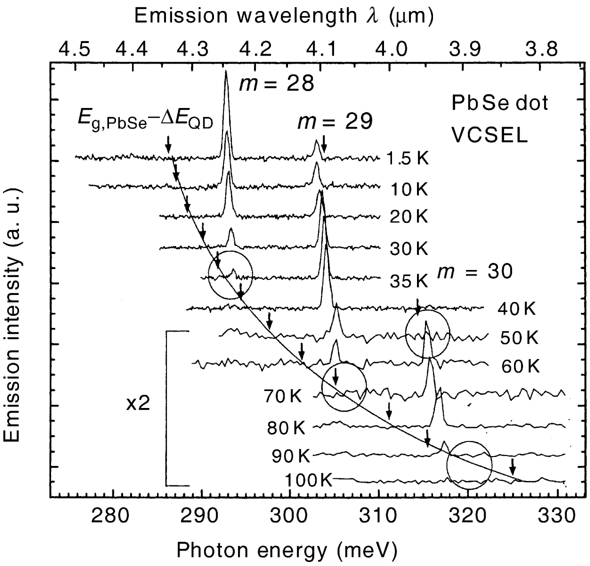

Figure 5 PbSe dot VCSEL emission spectra at various temperatures between 1.5 and 100 K for a pump power twice the threshold, showing the switching of the laser emission to higher cavity modes as the temperature increases. The arrows and the dashed line indicate the low-energy edge of the quantum-dot gain spectrum given by $E_{g,\mathrm{PbSe}}(T) + \Delta E_{\mathrm{QD}}$, where $\Delta E_{\mathrm{QD}} = 141$ meV. Reproduced with permission from Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Midinfrared surface emitting PbSe/PbEuTe quantum dot lasers. *Applied Physics Letters* 79: 1225–1227. Copyright 2001, American Institute of Physics.

usually about 1 μm thick with grain sizes of about 1 μm.

Apart from polycrystalline thin-film detectors also single crystalline films have been used for infrared detection. In particular, the photoconductive response of a sequence of n- and p-doped PbTe multilayers (so-called nipi structures) was shown to come close to that for the theoretical limit for the detectivity D^* of an ideal photodetector at temperatures between 77 and 90 K. Recently, lead selenide detectors have been used in advanced focal plane array (FPA) detectors and are commercially available as linear or bilinear FPAs with up to 526 detector elements. These elements are coupled to CMOS multiplexer readout circuits. The PbSe detector elements are thermoelectrically cooled, with entire power requirement for the FPAs and the cooler of less than 5 W. These structures are a low-cost alternative to the more elaborate FPAs based on HgCdTe technology.

The importance of thermal imaging by using focal plane arrays of many detector elements has recently stimulated further work with IV–VI epitaxial layers. In particular, the possibility of using Si substrates, as first shown by Zogg *et al.*, has brought a new impetus to the field of IV–VI infrared devices. With the use of Si wafers it is possible to integrate the detector of an infrared camera which consists of thousands of pixels to the read-out electronics which is usually based on standard silicon CMOS devices. In this case, the active lead salt infrared detector elements are grown by molecular beam epitaxy on CaF_2/BaF_2 buffer layers deposited on the Si (111) substrates after the fabrication of the electronic circuitry for the read-out. Conventional HgCdTe focal plane technology relies on hybrid structures with In bump bonds between the infrared devices and the Si substrate for each of the pixels. Due to the rather high dielectric constants of the IV–VI compounds as compared to HgCdTe, the high-frequency operation of IV–VI detector structures is limited to about 100 MHz. This upper frequency limit is, however, sufficient for infrared focal plane arrays. With IV–VI active epilayer thicknesses of few μm internal quantum efficiencies close to 100% and external efficiencies of about 60% have been achieved. Using materials like PbS, PbTe, PbEuSe, PbEuTe, or PbSnSe, the cutoff wavelength of the detector array can be varied from 4 to 14 μm at $T = 77$ K. The sensitivity of these detector structures is limited by generation-recombination noise for $T < 100$ K, whereas for higher temperatures these sensors are diffusion limited. The variation of quantum efficiencies between different pixels was shown to be less than about 3% for a 256 pixel device. Since all growth steps for these infrared sensors are performed at temperatures below 450 °C, a processing compatibility exists with standard Si-VLSI technology.

Summary

The peculiar structural, electronic, and optical properties of IV–VI semiconductors as compared to other semiconductor materials are a consequence of the ten valence electrons per atomic pair instead of the eight valence electrons typical for the tetrahedrally bonded group IV, III–V, and II–VI semiconductors. The most important group of IV–VI materials is the so-called lead chalcogenide (lead salt) compounds and their ternary and quaternary alloys. Their electronic band structure with the narrow direct minimum gap not at the center but at the edges of the Brillouin zone (L-points) exhibits a number of unique properties. The minimum energy gap increases with increasing temperature and decreases with hydrostatic pressure. The refractive indices are particularly large. As compared to narrow-gap III–V or II–VI semiconductors the Auger recombination rates are by two orders of magnitude smaller. All these properties provide the lead salts with great advantages for achieving infrared p-n junction lasing operation at comparatively high temperatures. Furthermore, efficient microcavity designs with extremely high reflectivities of 99% were realized with only three layer periods of combination like $PbSrTe/BaF_2$ or PbEuTe/EuTe due to their extremely high refractive index contrast. Further applications of the lead compounds are high sensitive infrared detectors, including focal plane array detectors which can be fabricated with comparatively low cost technologies, as well as thermoelectric devices.

List of Units and Nomenclature

Absorption *constant*	α
Absorption index	κ
Composition	x
Damping parameter	Γ_D
Detectivity	D^*
Dielectric function	ε
Energy	E
Energy gap	E_g
Free electron mass	m_0
High frequency dielectric constant	ε_∞
Kramers Kronig relations	KKR
Longitudinal effective mass	m_l
Longitudinal g-factor	g_l
Pressure	p
Refractive index	n
Static dielectric constant	ε_0
Temperature	T
Transverse effective mass	m_t
Transverse g-factor	g_t

See also

Nonlinear Optics, Basics: Kramers–Krönig Relations in Nonlinear Optics.

Further Reading

Bauer G (1980) Magnetooptical properties of IV–VI compounds. In: Zawadzki W (ed.) *Narrow Gap Semiconductors Physics and Applications*, pp. 470–494. Berlin: Springer Verlag.

Bauer G and Krenn H (1985) Lead selenide, lead telluride. In: Palik ED (ed.) *Handbook of Optical Constants*, pp. 517–523 and 535–546. New York: Academic Press.

Felix CL, Bewley WW, Vurgaftman I, *et al.* (2001) Low-threshold optically pumped $\lambda = 4.4$ μm vertical cavity surface emitting laser with a PbS quantum well active region. *Applied Physics Letters* 78: 3770–3772.

Holloway G (1980) Thin film IV–VI semiconductor photodiodes. In: Hass G and Francombe MF (eds) *Physics of Thin Films*, vol. 11, pp. 105–203. New York: Academic Press.

Jantsch W (1983) Dielectric properties and soft modes in semiconducting (PbSnGe)Te. *Springer Tracts in Modern Physics* 99: 1–50.

Katzir A, Rosman R, Shani Y, *et al.* (1989) Tuneable lead salt lasers. In: Cheo PK (ed.) *Handbook of Solid State Lasers*, pp. 227–347. New York: Marcel Dekker.

Khoklov D (2003) Lead chalcogenioles: Physics and applications. In: Manasreh MO (ed.) *Optoelectronic Properties of Semiconductors and Superlattices*, vol. 18. New York: Taylor and Francis.

Lucovsky G, Martin RM and Burstein E (1974) Structure, bonding and lattice dynamical properties of the IV–VI compounds. In: Rabii S (ed.) *Physics of IV–VI Compounds and Alloys*, pp. 93–102. New York: Gordon and Breach.

Nimtz G (1993) IV–VI compounds. In: Hellwege K and Madelung O (eds) *Numerical Data and Functional Relationships in Science and Technology.* Landolt-Börnstein, new series, group III, vol. 17, Part F, pp. 152–180. Berlin: Springer Verlag.

Partin DL and Heremans J (1994) Growth of narrow bandgap semiconductors. In: Moss TS and Mahajan S (eds) *Handbook on Semiconductors*, vol. 3A, pp. 369–450. Amsterdam: North-Holland.

Preier H (1990) Physics and applications of IV–VI compound semiconductor lasers. *Semiconductor Science and Technology* 5: S12–S20.

Ravich YI, Efimova BA and Smirnov IA (1970) Semiconducting Lead Chalcogenides. New York: Plenum.

Schwarzl T, Heiss W and Springholz G (1999) Ultra-high-finesse IV–VI microcavities for the midinfared. *Applied Physics Letters* 75: 1246–1248.

Springholz G, Shi Z and Zogg H (1998) Molecular beam epitaxy of IV–VI semiconductors. In: Liu AWK and Santos MB (eds) *Heteroepitaxial Thin Film Systems*, pp. 621–688. Singapore: World Scientific.

Springholz G, Schwarzl T, Heiss W, *et al.* (2001) Mid-infrared surface-emitting PbSe/PbEuTe quantum dot-lasers. *Applied Physics Letters* 79–81: 1225.

Tacke M (1995) New developments and applications of tunable IR lead salt lasers. *Infrared Physics and Technology* 36: 447–463.

Yuan S, Springholz G, Bauer G and Kriechbaum M (1994) Electronic and optical properties of PBTe/$\mathrm{PB}_{1\text{-}x}\mathrm{Eu}_x\mathrm{Te}$ multiple quantum well structures. *Physical Review* B49: 5476–5489.

Zemel JN, Jensen JD and Schoolar RB (1965) Electrical and optical properties of epitaxial films of PbS, PbSe, PbTe and SnTe. *Physical Review* 140: A330–A337.

Zogg H, Fach A, John J, *et al.* (1995) Photovoltaic IV–VI on Si infrared sensor arrrays for thermal imaging. *Optical Engineering* 34: 1964–1969.

Mercury Cadmium Telluride

M B Reine, BAE SYSTEMS, Lexington, MA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mercury-cadmium telluride (HgCdTe or $\mathrm{Hg}_{1-x}\mathrm{Cd}_x$ Te) is a pseudobinary alloy semiconductor that crystallizes in the zincblende structure. The energy bandgap $E_G(x,T)$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe. The bandgap passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K). This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range. Direct allowed valence-to-conduction band transitions are responsible for large values of the optical absorption coefficient, which enable quantum efficiencies approaching 100% to be achieved in practical device designs. Long minority carrier lifetimes result in low thermal noise, and permit high performance to be achieved at operating temperatures that are the highest reported among infrared detectors of comparable cutoff wavelengths. HgCdTe is well-established today as the most broadly applicable high-performance infrared detector material.

This article summarizes the fundamental properties of this versatile alloy semiconductor, and relates the material properties to its success as an infrared photoconductive and photovoltaic detector material.

Historical Background

It is noteworthy that HgCdTe is probably the first example of a semiconductor material that was intentionally created to satisfy well-defined theoretical device requirements. The importance of the infrared atmospheric transmission window at 8–12 μm was well known for thermal imaging systems, which enable night vision by imaging emitted infrared radiation from the scene rather than reflected visible radiation. The only available infrared detectors with the required sensitivity at 8–12 μm were extrinsic photoconductors such as copper-doped germanium, in which the photoinduced transitions were between a donor or acceptor level and the adjacent band extremum. However, such extrinsic detectors required cooling to prohibitively low temperatures, such as 10–15 K, which made thermal imaging systems bulky and impractical.

Theory showed that the reasons for such a low operating temperature were fundamental to the extrinsic detector material itself – essentially a short carrier lifetime that resulted in a large thermal generation rate and a correspondingly large thermal noise. Theory also showed that an intrinsic detector, in which the photoinduced transitions were direct (vertical) transitions between the valence band and the conduction band, could have significantly longer carrier lifetimes. This meant that an intrinsic detector, with a bandgap of 0.1 eV that matched the 8–12 μm band, could achieve the required sensitivity at much higher operating temperatures, as high as 77 K, that of liquid nitrogen. These theoretical and practical considerations motivated the search in the late 1950s for a semiconductor with a direct bandgap of 0.1 eV.

The first report of the synthesis of the semimetal HgTe and the wide-bandgap semiconductor CdTe to form the semiconductor alloy HgCdTe was published in 1959 by the Royal Radar Establishment in Malvern, England. This seminal paper reported both photoconductive and photovoltaic response at wavelengths extending to 12 μm, and made the understated observation that this material showed promise for intrinsic infrared detectors.

Soon thereafter, working under a US Air Force contract with the objective of devising an 8–12 μm background-limited semiconductor infrared detector that would operate at temperatures as high as 77 K, the group led by Paul Kruse at the Honeywell Corporate Research Center in Hopkins, Minnesota developed a modified Bridgman crystal growth technique for HgCdTe. In 1962 they reported both photoconductive and photovoltaic detection in rudimentary HgCdTe devices.

Fundamental Material Properties

Three fundamental properties determine the degree to which a semiconductor will make a successful infrared detector. First, the energy bandgap must be approximately equal to the smallest photon energy to be detected. There can be severe cooling penalties if the bandgap is significantly smaller than the smallest photon energy of interest. Second, the optical absorption coefficient must be large enough so that nearly all incident photons of interest are converted to electron–hole pairs for practical values of device thickness. Third, the inherent recombination mechanisms must allow long carrier lifetimes to be realized so that thermal generation rates, and hence thermal noise, can be minimized for practical devices and operating conditions.

The three key features that make HgCdTe the nearly ideal infrared detector material – tailorable energy band gap over the 1–30 μm range, large optical absorption coefficients that enable high quantum efficiency, and favorable inherent recombination mechanisms that lead to high operating temperature – are direct consequences of the energy band structure of this zincblende semiconductor. They apply no matter whether the device embodiment is a photoconductor, a photodiode or a metal–insulator–semiconductor detector. In addition, HgCdTe has many other favorable material properties, summarized in this article, that are technologically important for realizing practical infrared detectors. Tables 1 and 2 summarize various material properties for $Hg_{1-x}Cd_xTe$. **Table 1** summarizes material properties for $Hg_{1-x}Cd_xTe$ that vary with alloy composition x and temperature T: lattice constant a, energy band gap E_G, cutoff wavelength λ_{CO} (= 1.24/E_G), intrinsic carrier concentration n_i, conduction band effective mass ratio m_C/m_0, conduction band g-factor g_C, static and high frequency dielectric constants $\varepsilon_S/\varepsilon_0$ and $\varepsilon_\infty/\varepsilon_0$, index of refraction $n_R = (\varepsilon_\infty/\varepsilon_0)^{1/2}$, electron and heavy hole mobilities μ_e and μ_{hh}, mobility ratio b, radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$, and typical lifetimes $\tau_{typical}$ observed in n-type HgCdTe of low carrier concentration ($<1 \times 10^{15} cm^{-3}$).

Table 1 Material properties for selected compositions of $Hg_{1-x}Cd_xTe$, including the binary components HgTe and CdTe, and several technologically important alloy compositions. The radiative and Auger-1 lifetimes τ_{RAD} and $\tau_{Auger\text{-}1}$ are calculated for n-type HgCdTe with $N_D = 1 \times 10^{15}\ cm^{-3}$

Property	*HgTe*	*$Hg_{1-x}Cd_xTe$*						*CdTe*
x	0	0.194	0.205	0.225	0.31	0.44	0.62	1.0
a (Å)	6.461	6.464	6.464	6.464	6.465	6.468	6.472	6.481
T (K)	77	77	77	77	140	200	250	300
E_G (eV)	−0.261	0.073	0.091	0.123	0.272	0.474	0.749	1.490
λ_{CO} (μm)		16.9	13.6	10.1	4.6	2.6	1.7	0.8
n_i (cm^{-3})		1.9×10^{14}	5.8×10^{13}	6.3×10^{12}	3.7×10^{12}	7.1×10^{11}	3.1×10^{10}	4.1×10^{5}
m_C/m_0		0.006	0.007	0.010	0.021	0.035	0.053	0.102
g_C		−150	−118	−84	−33	−15	−7	−1.2
$\varepsilon_S/\varepsilon_0$	20.0	18.2	18.1	17.9	17.1	15.9	14.2	10.6
$\varepsilon_\infty/\varepsilon_0$	14.4	12.8	12.7	12.5	11.9	10.8	9.3	6.2
n_R	3.79	3.58	3.57	3.54	3.44	3.29	3.06	2.50
μ_e (cm^2/V-s)		4.5×10^5	3.0×10^5	1.0×10^5				
μ_{hh} (cm^2/V-s)		450	450	450				
$b = \mu_e/\mu_{hh}$		1000	667	222				
τ_{RAD} (μs)		16.5	13.9	10.4	11.3	11.2	10.6	
$\tau_{Auger\text{-}1}$ (μs)		0.45	0.85	1.8	40	450	4.8×10^3	
$\tau_{typical}$ (μs)		0.4	0.8	1	7			

Table 2 Material properties for $Hg_{1-x}Cd_xTe$ that are independent of or relatively insensitive to alloy composition

Property	*Symbol*	*Value*
Kane matrix element	E_P	19 eV
Split-off band energy	Δ	0.93 eV
Heavy-hole effective mass ratio	m_{hh}/m_0	0.40–0.53
Valence band offset	E_v(HgTe) − E_v(CdTe)	0.35–0.55 eV

Energy Band Gap and Band Structure

The energy band gap $E_G(x, T)$ of $Hg_{1-x}Cd_xTe$ varies continuously, and nearly linearly, with alloy composition parameter x, ranging from 1.6 eV for the wide-gap semiconductor CdTe to −0.3 eV for the semimetal HgTe, as shown in Figure 1a. Also plotted in Figure 1a is the cutoff wavelength $\lambda_{CO}(x, T)$ ($\approx hc/E_G(x, T)$, where h is Planck's constant and c is the speed of light), defined as that wavelength at which the photoresponse falls to 50% of its peak value. The energy bandgap $E_G(x, T)$, defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$, passes through zero at an alloy composition of $x = 0.16$ (for $T = 0$ K), as illustrated in Figure 1b. This broad range of energy bandgap provides a continuum of materials whose absorption edge can be tailored to provide high-performance quantum detectors for wavelengths over the 1–30 μm spectral range.

Optical Absorption Coefficient

Optical absorption coefficient data for several alloy compositions of $Hg_{1-x}Cd_xTe$ are plotted versus wavelength in Figure 2. Several trends are evident. The absorption strength generally decreases as the gap becomes smaller due both to the decrease in the conduction band effective mass and to the $\lambda^{-1/2}$ dependence of the absorption coefficient on wavelength λ. There is an exponential tail at energies just below the bandgap energy.

Carrier Lifetime Mechanisms

There are two fundamental mechanisms that determine the minority carrier lifetime in defect-free HgCdTe: radiative recombination and Auger recombination (Auger-1 in n-type and Auger-7 in p-type). Because they involve only interactions among electrons and holes in the valence and conduction bands, and do not require the intermediary of defect states, these mechanisms are regarded as inherent to the material itself, depending only on the band structure, the donor or acceptor concentration, and the temperature. The important advantage that HgCdTe enjoys as a high-performance detector material is that these fundamental mechanisms provide, for practically achievable

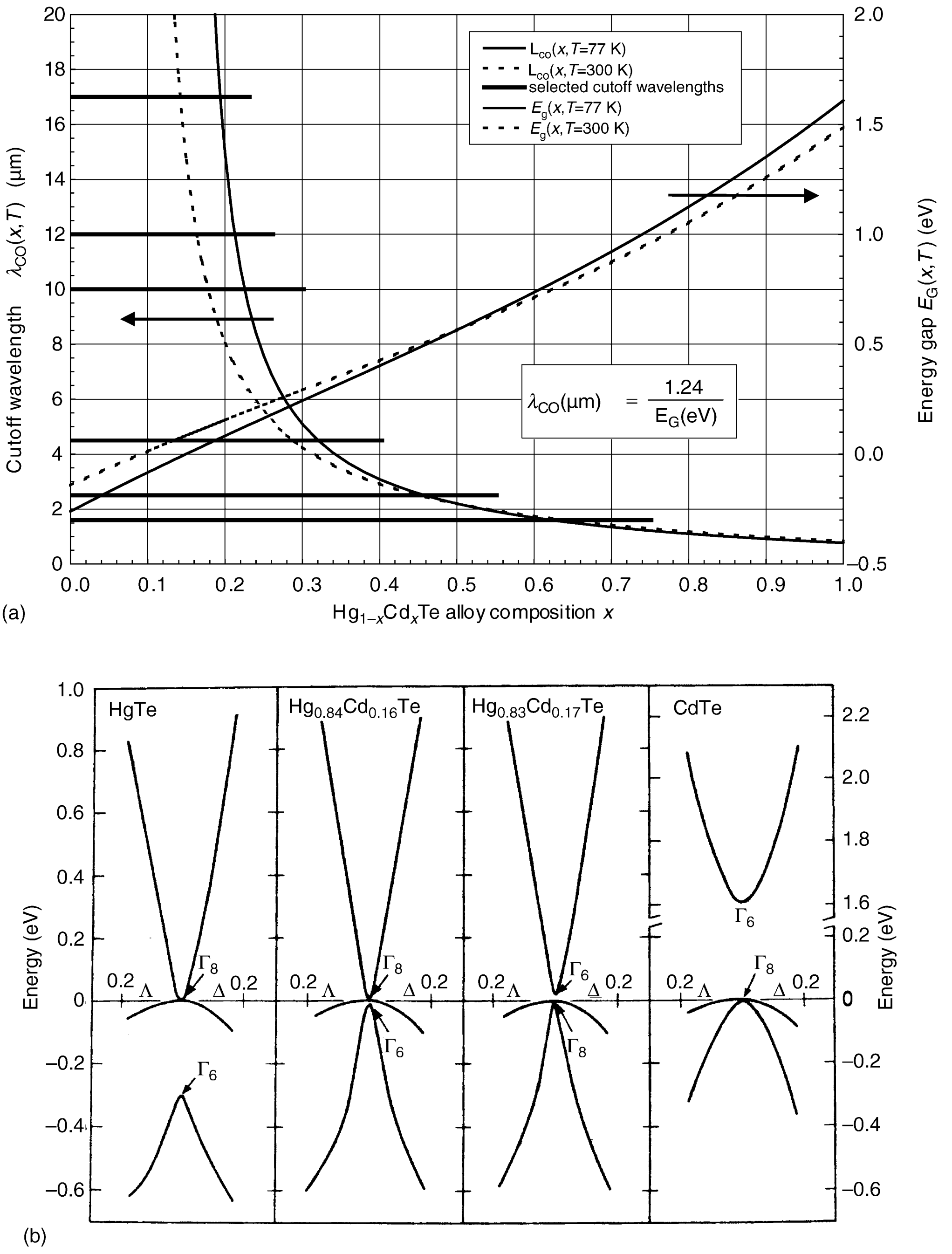

Figure 1 (a) Energy bandgap $E_G(x,T)$ and cutoff wavelength $\lambda_{CO}(x,T)$ for $Hg_{1-x}Cd_xTe$, plotted versus alloy composition parameter x, for temperatures of 77 K (solid curves) and 300 K (dashed curves). Horizontal bars indicate important cutoff wavelengths for HgCdTe detector technology. (b) Band diagrams near the Γ point calculated for HgTe, for two compositions of $Hg_{1-x}Cd_xTe$, and for CdTe, illustrating the transition from negative to positive energy bandgap. The energy bandgap is defined at the difference between the Γ_6 and Γ_8 band extrema at $\Gamma = 0$. Part (b) reproduced with permission from Chadi DJ and Cohen ML (1973) Electronic structure of $Hg_{1-x}Cd_xTe$ alloys and charge-density calculations using representative K points. *Physical Review B* 7: 692–699.

carrier concentrations, lifetimes that are long enough (and hence thermal generation rates that are low enough) that background-limited sensitivity is achieved at temperatures substantially higher than for other classes of infrared detectors, such as extrinsic detectors (As-doped Si, Cu-doped Ge) and quantum-well infrared photodetectors (QWIPs).

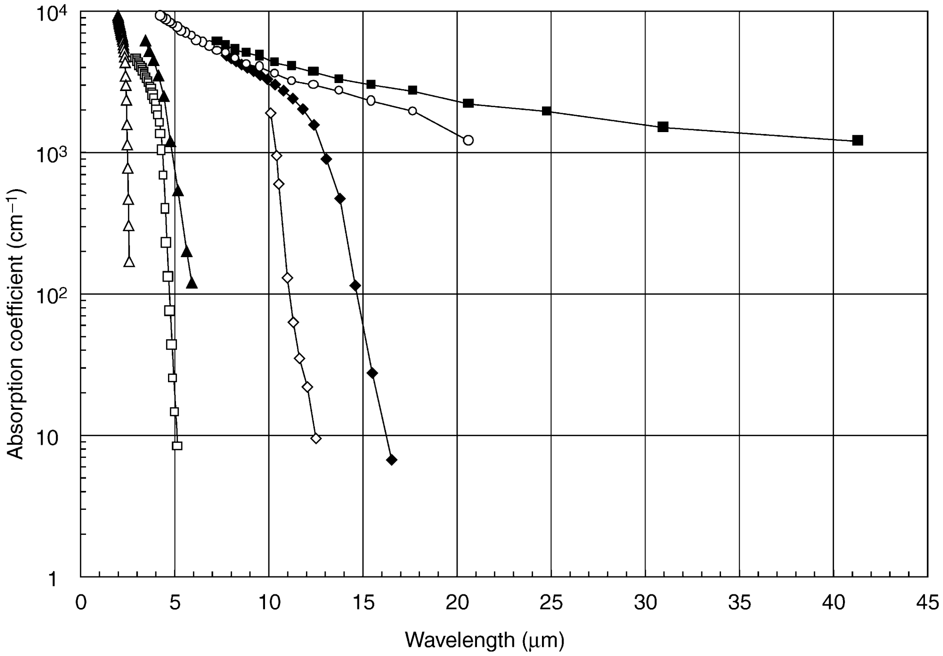

Figure 2 Optical absorption coefficient data for several $Hg_{1-x}Cd_xTe$ alloy compositions, for photon energies near the fundamental absorption edge, plotted versus wavelength.

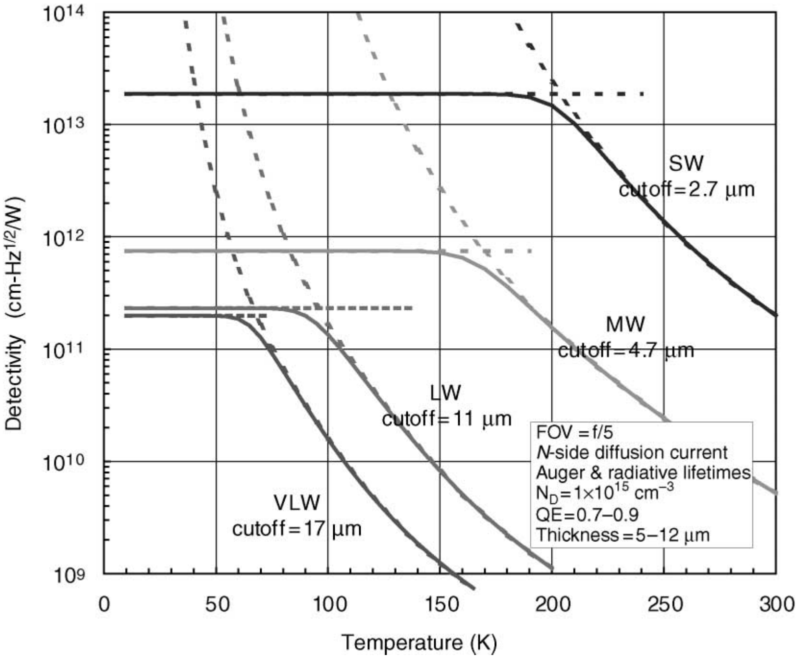

Figure 3 Calculated detectivity (D^*) for $Hg_{1-x}Cd_xTe$ infrared detectors for four technologically important wavelength regions, plotted versus operating temperature.

Figure 3 illustrates the sensitivities and operating temperatures that can be achieved for HgCdTe infrared detectors for four important wavelength regions. Detectivity, D^*, is the normalized signal-to-noise ratio that is the key figure of merit for the sensitivity of an infrared detector. At lower temperatures, detector thermal noise is negligible, and D^* is limited by detector noise due to fluctuations in the arrival rate of photons from the thermal room-background radiation. This limit is referred to as the BLIP (background-limited infrared photodetector) limit, and is indicated in Figure 3 by the horizontal dashed lines. As detector temperature increases, the detector thermal noise increases exponentially, and eventually overcomes the background noise, causing D^* to decrease exponentially for further increases in temperature. Detector thermal noise is proportional to the thermal generation rate, which is inversely proportional to the carrier lifetime.

n-Type and p-Type Doping

HgCdTe can be made n-type or p-type by a number of relatively convenient methods, at carrier concentrations required for high-performance n-type photoconductor and p-n junction photodiode architectures. Common donors are indium and iodine. Arsenic has become the most commonly used acceptor dopant, although copper and gold are used in some cases. Some HgCdTe photodiode designs still employ native metal-vacancy point defects as acceptors, although the clear trend is toward the use of extrinsic acceptor doping to avoid the strong Shockley–Read recombination associated with the Hg vacancy. Residual electrically active impurity concentrations are generally less than $1 \times 10^{14}\,\mathrm{cm}^{-3}$, which allows controllable doping at low-$10^{14}\,\mathrm{cm}^{-3}$ concentrations.

Crystal Growth Methods

The first growth methods for HgCdTe were bulk growth. Liquid phase epitaxy (LPE) began to be developed around 1975, followed by metalorganic chemical vapor deposition (MOCVD) and molecular beam epitaxy (MBE) in the early 1980s. Bulk growth, LPE, MOCVD, and MBE techniques have been refined through the years, and each is still in use today. Each technique has found application for certain types of HgCdTe devices. Bulk growth methods are used for production quantities of n-type wafers, with diameters of 12–20 mm, for LW and VLW photoconductive arrays, which are more tolerant of dislocations. LPE is in use today for production quantities of photovoltaic (PV) HgCdTe arrays, and has also been used for certain photoconductive HgCdTe arrays. Vapor-phase epitaxy (VPE) methods such as MBE and MOVPE are in use for engineering and prototype quantities. Both run-to-run reproducibility and control as well as intra-wafer uniformity of HgCdTe alloy composition are well in hand, meeting or exceeding requirements.

LPE-based one-layer and two-layer photodiode technology is well established today, offering the lowest defect densities and highest operabilities for production quantities of high performance FPAs. The VPE methods, MBE and MOVPE, will probably not entirely replace LPE in the near future, but will continue to improve, particularly with respect to lower defect densities, offering *in situ* growth of advanced bandgap-engineered detector designs that are unwieldy or impossible with LPE, such as dual-band detectors, avalanche photodiodes, and planar buried junctions with *in situ* CdTe passivation. MBE has shown the potential for rapidly switching from growth of one HgCdTe alloy composition to another in successive growth runs, without the necessity of preparatory calibration growth runs. This compositional agility, important for the rapid and cost-effective adjustment from product to product, will be further enhanced by the ongoing efforts to implement and improve *in situ* real-time monitors and feedback controls.

Substrates for Epitaxial Growth

The lattice mismatch between HgTe and CdTe is small, approximately 0.3%. This allows epitaxial growth of high-quality HgCdTe films on IR-transparent CdTe or nearly-lattice-matched IR-transparent $\mathrm{Cd}_{1-z}\mathrm{Zn}_z\mathrm{Te}$ ($z \approx 0.04$) substrates, with dislocation densities in the mid-$10^4\,\mathrm{cm}^{-2}$ range. It also allows the *in situ* growth of various isotype and anisotype heterojunctions with tolerably low or negligible densities of misfit dislocations. This has led to an increasing number of bandgap-engineered HgCdTe photodiode structures, such as dual-band or two-color detector arrays and Auger-suppressed photodiodes with the potential for greatly increased operating temperature.

Sapphire and silicon are IR-transparent substrates that are less costly, that are available in much larger areas, and that are more rugged than CdTe and CdZnTe. Both sapphire and silicon can be used for epitaxial growth of HgCdTe films with dislocation densities that are acceptably low (mid-$10^6\,\mathrm{cm}^{-2}$) for many important photodiode applications, such as for the MW and SW spectral ranges, and for the LW spectral range for high background photon fluxes. LPE growth of HgCdTe($x = 0.4$) is done on 3-inch diameter sapphire substrates with MOCVD-grown CdTe buffer layers. MBE growth of HgCdTe($x = 0.3$) on 4-inch diameter (211) silicon substrates, with ZnTe/CdTe buffer layers, has recently been reported.

Dielectric Constant

The relatively low dielectric constant ($\varepsilon_S = 18\varepsilon_0$) of HgCdTe, in contrast to those of PbTe and PbSnTe, permits low junction capacitance. This is important for fast response in laser pulse detectors where small $R-C$ time constants are needed. It is also important for suppressing preamplifier noise below the detector noise, and thereby achieving high detectivity (D^*), in photodiodes operating at low temperature and low background fluxes, where the junction resistance is sufficiently high that the junction capacitance becomes an important contribution to the junction impedance at frequencies of interest.

Surface Passivation

Passivation of the exposed HgCdTe surfaces is critical for both device performance and long-term stability. The favorable surface properties of HgCdTe have enabled practical surface passivation schemes to be developed for both n-type photoconductors and p-n junction photodiodes of both planar and mesa configurations.

n-type HgCdTe photoconductors are commonly passivated with a native oxide that accumulates the surface, resulting in surface recombination velocities as low as $50\ \mathrm{cm\,s^{-1}}$, along with low $1/f$ noise (knee frequencies less than 50 Hz). The surface accumulation layer also provides a layer of surface electrons with high mobility, though lower than the bulk electron mobility, such that the shunt conductance of this layer is tolerably low compared to the bulk conductance of the detector itself. Several methods have been used for growing this native oxide, including plasma deposition and electrochemical (anodic) oxidation.

HgCdTe photovoltaic detectors require quite different passivation because usually both n-type and p-type surfaces are exposed as well as the depletion region. The CdTe/HgCdTe heterostructure is an important passivation for HgCdTe photovoltaic detectors and arrays. CdTe passivation has enabled arrays of HgCdTe photodiodes with negligible $1/f$ noise (knee frequencies less than 1 Hz), radiation hardness well in excess of the 20–40 kRad(Si) usually encountered in Earth-resource space missions, and excellent stability to thermal bake and to exposure to moisture.

CdTe has a number of important advantages as a passivant for HgCdTe junctions. It is a ‘native’ passivant that is chemically compatible with HgCdTe. Adhesion is excellent. The valence band edge of CdTe is approximately 0.4 eV below that of HgTe, thereby allowing repulsive barriers to form in the underlying HgCdTe for both electrons in the conduction band and holes in the valence band. Both mesa and planar photodiodes have been successfully passivated with CdTe. A wide variety of methods has been reported for deposition of CdTe passivating layers, including MBE, MOVPE, e-beam evaporation, sputtering, and hot wall epitaxy.

Electron-to-Hole Mobility Ratio

The small values for the conduction band effective mass ratio lead to large values for the electron mobility, as large as $4.5 \times 10^5\ \mathrm{cm^2/V\text{-}s}$ for $x = 0.195$ at 77 K. The heavy hole effective mass ratio is quite large, approximately 0.5 for all alloy compositions, leading to low values for the heavy hole mobility, on the order of $500\ \mathrm{cm^2/V\text{-}s}$. This gives large values for the electron-to-heavy hole mobility ratio, b, as high as 1000 for $x = 0.194$ at 77 K.

These large values of b were of great benefit to n-type HgCdTe photoconductors. Most such devices have small active areas, on the order of $50 \times 50\ \mu\mathrm{m}^2$, and the drift length for minority carriers (holes) becomes comparable to the interelectrode spacing at modest bias voltages. In this case, when the contacts have a high recombination velocity, the photoconductive gain saturates to the value $b/2$. High values of b in HgCdTe allowed sufficiently high photoconductive gains to be achieved to raise the detector noise above the preamp noise, an essential condition for achieving BLIP sensitivity.

HgCdTe Infrared Detector Configurations

Three different types of infrared detectors can be realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode, also referred to as the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. These are illustrated in Figure 4.

In all three types of HgCdTe detector – PC, PV, and MIS – excess electron–hole pairs are photogenerated by the same mechanism: valence-to-conduction band absorption of infrared radiation with photon energy greater than the bandgap energy E_G. All three types can achieve high quantum efficiencies, approaching 100% for well-designed devices. All three types have the same fundamental recombination mechanisms (Auger-1 and radiative) that determine the highest possible operating temperature to achieve a given sensitivity (signal-to-noise ratio).

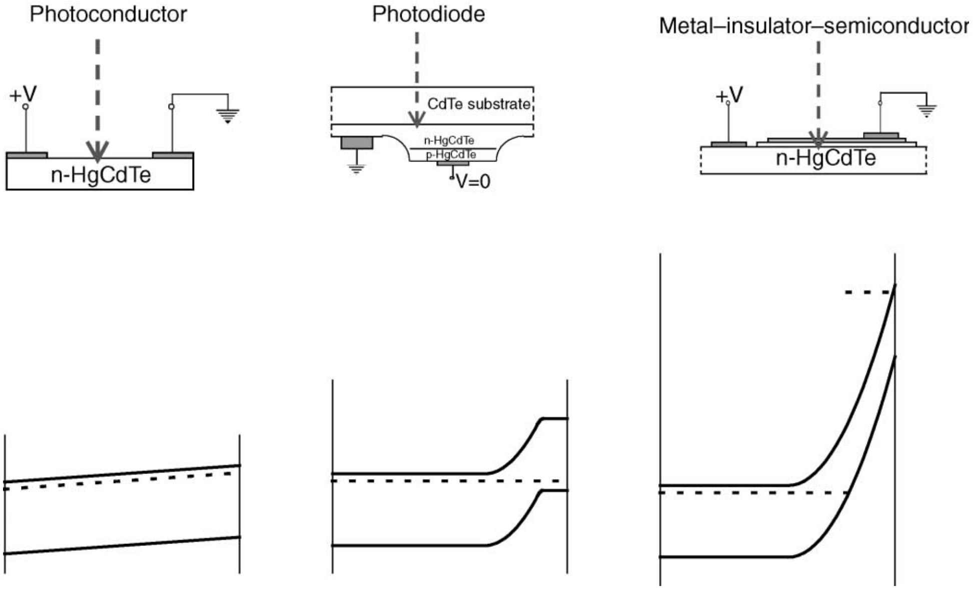

Figure 4 Three different types of infrared detectors have been realized in the $Hg_{1-x}Cd_xTe$ alloy semiconductor: the photoconductive (PC) detector, the photodiode or the photovoltaic (PV) detector, and the metal–insulator–semiconductor (MIS) detector. HgCdTe PC and PV detector arrays are being manufactured for a wide variety of applications. The MIS detector was abandoned in the mid-1980s because it placed unachievable demands on material quality.

The differences among PC, PV, and MIS HgCdTe detectors are due to the way in which the photogenerated electron–hole pairs are manifested electrically in the terminal characteristics of the device. In the PC detector, the excess electron–hole pairs are sensed as a small increase in the conductivity of what is basically a two-terminal resistor. In the PV detector, the excess electron–hole pairs are sensed as a photocurrent in a short-circuited p-n junction photodiode. In the MIS detector, the excess electron–hole pairs are sensed as a small change in the voltage due to minority photocarriers filling a transient charge storage well that has been biased to deep depletion.

PC HgCdTe Detectors

The HgCdTe PC detector is a low-resistance two-terminal device. Resistances are typically 25–100 ohms per square. A dc bias current is applied to convert the conductivity change to an observable voltage change. The device operates under near-equilibrium conditions, with small dc bias voltages typically on the order of 0.1 V. The dc bias electric field within the photoconductor is quite small, generally in the 20–50 $\mathrm{V\,cm^{-1}}$ range. The low-resistance device requires a low-noise bipolar preamplifier, usually external to the dewar. Each PC detector element requires one lead through the dewar wall, thus limiting the practical number of PC elements in an array to about 200–300.

The basic HgCdTe photoconductor is a rectangular area, photolithographically defined in an n-type HgCdTe layer approximately 8–15 μm thick, with two ohmic contacts on opposite edges. Surface passivation is straightforward: a thin native oxide strongly accumulates the surface of n-type HgCdTe, thereby reducing the surface recombination velocity to negligibly small values.

Many useful variations on the simple rectangular geometry are possible because of the favorable material properties of HgCdTe. Several contact geometries have been devised to minimize the recombination of photocarriers at the contacts. Serpentine designs have been used to increase detector resistance and reduce bias power dissipation, especially important issues for large-area and very long-wavelength detectors. The SPRITE (signal processing in the element) design reduces the number of electrical connections and dewar leads for scanned arrays by performing both detection and integration within an elongated HgCdTe bar.

PV HgCdTe Detectors

The HgCdTe photodiode is a more complicated device than the photoconductor, requiring both n-type and p-type layers, and having a surface depletion region that makes stringent demands on surface passivation technology.

The PV detector is a high-resistance p-n junction device. The photodiode is operated very close to equilibrium conditions, either at zero bias voltage or at a small reverse bias voltage whose magnitude is several *kT*, typically 20–30 mV. The absorber layer is usually n-type HgCdTe about 5–15 μm thick, with the thicker values required for longer-wavelength radiation. A p-n junction is formed by a thin p-type layer. A wide variety of junction formation methods are used. Wide-gap-p on narrow-gap-n heterojunctions are formed by two-layer LPE. Planar p-on-n or n-on-p junctions are formed by arsenic or boron implantation into n-type or p-type layers grown by LPE or MBE. Vertical-geometry p-on-n junctions are formed by a damage mechanism during ion beam milling. Both single-junction and multijunction devices are grown *in situ* by vapor phase epitaxial methods such as MBE and MOVPE. Surface passivation is generally accomplished by CdTe. The high-resistance PV device can match well with low-noise silicon CMOS preamplifiers that can be integrated with two-dimensional arrays to form large back-illuminated hybrid focal plane arrays (FPAs). Typical array formats for HgCdTe FPAs are 256 × 256 and 480 × 640, and can be as large as 1024 × 1024 and 2048 × 2048.

MIS HgCdTe Detectors

The metal–insulator–semiconductor (MIS) detector, also referred to as a photocapacitor, is usually formed on an n-type HgCdTe absorber layer. The insulator of choice is a thin native oxide. The gate electrode is a thin semitransparent metal film.

In contrast to the PC and PV detectors, the MIS detector operates in a strongly nonequilibrium mode.

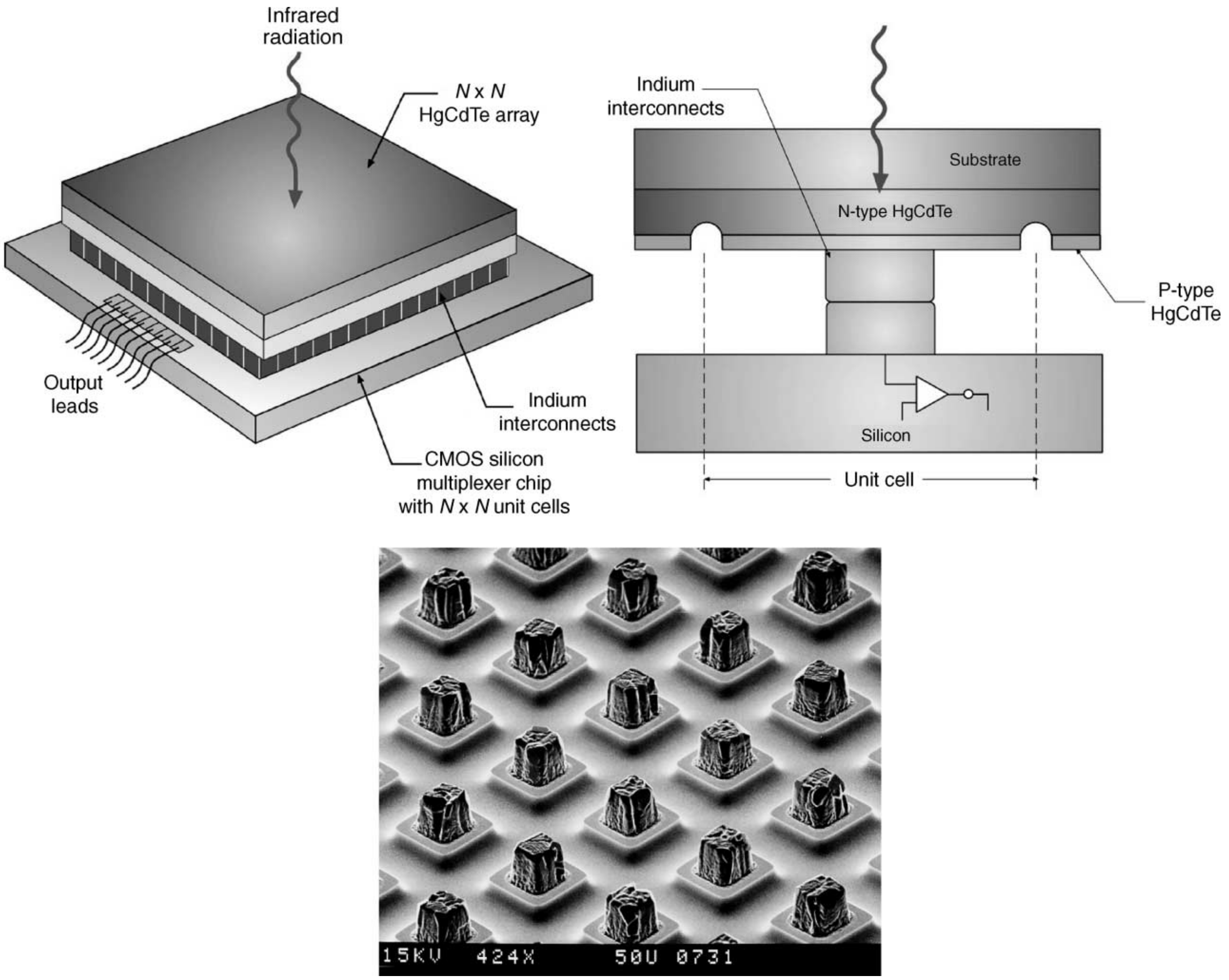
Infrared radiation
N x *N* HgCdTe array
Output leads
Indium interconnects
CMOS silicon multiplexer chip with *N* x *N* unit cells
Indium interconnects
Substrate
N-type HgCdTe
P-type HgCdTe
Silicon
Unit cell
15KV 424X 50U 0731

Figure 5a Back-illuminated bump-interconnected HgCdTe FPA architecture. The SEM photo shows an LPE P-on-n HgCdTe mesa photodiode array with 60 × 60 μm^2 unit cells and 33 × 33 μm^2 mesa (junction) areas.

A bias voltage pulse of several volts is applied across the capacitor to drive the HgCdTe surface into deep depletion. Excess electron–hole pairs are photogenerated in the absorber layer and diffuse to the edge of the depletion region, where the holes are collected in the charge well. After a period of time, called the integration time, the voltage across the capacitor is sensed, giving a measure of how much charge was collected, and the cycle is repeated.

Because of the nonequilibrium operation of the MIS detector, much larger electric fields are set up in the depletion region than in the p-n junction, resulting in defect-related tunneling dark current that is orders of magnitude larger than the fundamental dark current. The MIS detector requires much higher material quality than PV or PC detectors, which still has not been achieved. For this reason, all development of the HgCdTe MIS detector was abandoned in the mid-1980s.

Two-Dimensional HgCdTe Infrared Focal Plane Arrays

Among the most important applications of HgCdTe are large two-dimensional electronically scanned hybrid arrays, referred to as focal plane arrays (FPAs). A hybrid HgCdTe FPA consists of a two-dimensional HgCdTe photovoltaic detector array that is interfaced electrically, thermally and mechanically with a matching two-dimensional array of input circuits in a silicon CMOS ROIC chip. Each HgCdTe detector element in the array has one electrical connection to its matching input circuit node in the silicon ROIC.

The thermal expansion coefficient of HgCdTe is sufficiently close to that of silicon to allow several technologically viable hybrid arrangements of HgCdTe detector arrays and silicon multiplexer chips.

There are two main types of hybrid HgCdTe FPAs being developed and manufactured today: the back-illuminated bump-interconnected configuration, shown in Figure 5a, and the front-illuminated or 'loophole' configuration shown in Figure 5b. Each configuration successfully solves the problem of the thermal expansion mismatch between silicon and HgCdTe in different ways, and each requires fundamentally different HgCdTe photodiode designs and processing.

In the back-illuminated configuration, illustrated in Figure 5a, the HgCdTe detector array is bump-mounted onto the silicon ROIC by cold-welded or thermally reflowed indium interconnects. Incident radiation reaches the detectors through an infrared-transparent substrate such as CdTe, CdZnTe, sapphire, or silicon.

In the front-illuminated (or 'loophole') FPA architecture, shown in Figure 5b, the thin HgCdTe layer is epoxied to the silicon ROIC chip, and undergoes elastic deformation when the FPA is cooled. This allows large arrays to be made without need for engineering the thermal expansion of the silicon ROIC chip. Low-temperature processing techniques

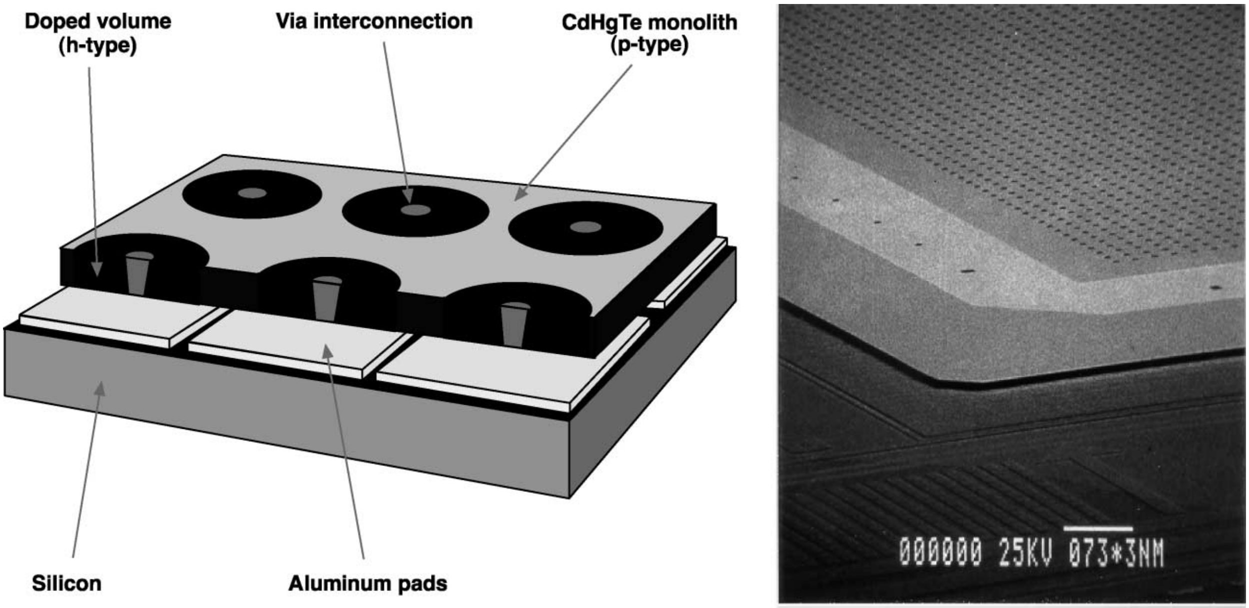

Figure 5b Front-illuminated loophole HgCdTe FPA architecture. The SEM photo shows one corner of a FPA. The metalized via holes are on 20 μm centers.

for junction formation and passivation are required because of the presence of the epoxy.

Conclusions and Trends

There has been enormous progress in HgCdTe material science and device technology over the past 45 years. This progress has been fueled by the many military and space applications for which HgCdTe PC and PV infrared detectors provide nearly ideal solutions. It is important to recognize that the progress so far in HgCdTe detector technology has been possible only through advances in basic materials growth and processing technology.

Further progress can be expected. The fundamental performance limits of HgCdTe photodiodes have not yet been reached. Continued reduction of material defects, both grown-in and process-induced, will reduce thermal generation rates, increase junction impedance, and allow higher operating temperatures for a number of important applications. Continued development of the *in situ* VPE growth methods – MBE and MOVPE – will allow bandgap-engineered heterojunction devices of increasing quality and complexity. Continued development of VPE growth on alternative substrates such as silicon will reduce the cost of two-dimensional arrays, and will make larger arrays practical.

List of Units and Nomenclature

BLIP	Background limited infrared photodetector
CdTe	Cadmium telluride
CdZnTe	Cadmium zinc telluride
CMOS	Complementary metal-oxide-semiconductor
FPA	Focal plane array
HgCdTe	Mercury cadmium telluride
$Hg_{1-x}Cd_xTe$	Mercury cadmium telluride
IR	Infrared
LPE	Liquid phase epitaxy
LW	Long wavelength
LWIR	Long wavelength infrared
MBE	Molecular beam epitaxy
MCT	Mercury cadmium telluride
MIS	Metal-insulator-semiconductor
MOCVD	Metal-organic chemical vapor epitaxy
MOVPE	Metal-organic vapor phase epitaxy
MW	Medium wavelength
MWIR	Medium wavelength infrared
PC	Photoconductive
PV	Photovoltaic
QWIP	Quantum well infrared photodetector
ROIC	Readout integrated circuit
SPRITE	Signal processing in the element
SW	Short wavelength
SWIR	Short wavelength infrared
VLW	Very long wavelength
VLWIR	Very long wavelength infrared
VPE	Vapor phase epitaxy

See also

Semiconductor Physics: Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties.

Further Reading

Capper P (ed.) (1994) *Properties of Narrow Gap Cadmium-based Compounds*. EMIS Datareviews Series No. 10. London: INSPEC, Institution of Electrical Engineers. Reviews and compilations of fundamental and technological properties of HgCdTe and CdTe/CdZnTe/CdTeSe.

Capper P (ed.) (1997) *Narrow-Gap II–VI Compounds for Optoelectronic and Electromagnetic Applications*. London: Chapman and Hall. Twenty review chapters on growth techniques, materials characterization, and device applications of HgCdTe.

Capper P and Elliott CT (eds) (2000) *Infrared Detectors and Emitters: Materials and Devices*. Boston, MA: Kluwer Academic Publishers. Review chapters on HgCdTe materials, photoconductive and nonequilibrium HgCdTe detectors, photovoltaic HgCdTe detectors, and HgTe-CdTe superlattice detectors.

Rogalski A, Adamiec K and Rutkowski J (2000) *Narrow Gap Semiconductor Photodiodes*. Washington, DC: SPIE Press.

Willardson RK and Beer AC (eds) (1981) *Defects, (Hg,Cd)Se, (Hg,Cd)Te*. Semiconductors and Semimetals, vol. 16. New York: Academic Press. Review chapters on magneto-optical properties of HgCdTe and nonlinear optical effects in HgCdTe.

Willardson RK and Beer AC (eds) (1981) *Mercury Cadmium Telluride*. Semiconductors and Semimetals, vol. 18. New York: Academic Press. Seven review chapters: history of HgCdTe, preparation of high-purity Hg, Cd and Te, crystal growth of HgCdTe, Auger recombination in HgCdTe, and photoconductive, photovoltaic and metal-insulator-semiconductor HgCdTe detectors.

Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures

Y-S Huang, National Taiwan University of Science and Engineering, Taipei, Taiwan

F H Pollak, Brooklyn College of the City University of New York, Brooklyn, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Modulation spectroscopy is a versatile and powerful optical technique for obtaining valuable information relevant to a large variety of semiconductor systems including bulk/thin films, surfaces/interfaces [semiconductor/air (vacuum), semiconductor/electrolyte, semiconductor/metal, semiconductor/semiconductor (homo- and heterojunctions)], micro- and nanostructures (quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots), the effects of growth/processing, as well as the characterization of actual device structures [heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical-cavity surface-emitting (VCSEL) quantum well lasers, etc.].

Modulation spectroscopy deals with the measurement and interpretation of changes in the optical response of a sample which are effected by the modification of the measurement conditions. This can easily be accomplished by periodically modulating either (1) the measurement conditions themselves ('internal' modulation) or (2) some parameters applied to the sample ('external' modulation). The observed normalized changes are usually small so that the difference signals are closely related to a derivative of the absolute spectrum with respect to the modifying parameters. The derivative nature of modulation spectroscopy emphasizes structure localized in the photon energy region of the interband (intersub-band) transitions of semiconductors (semiconductor microstructures) and suppresses uninteresting background effects. As a result, weak features that may not be detected in the absolute spectra are often enhanced. Because of this derivative-like nature a large number of sharp spectral features may be observed, even at room temperature.

The ability to perform a lineshape fit is one of the great advantages of modulation spectroscopy. Since for the modulated signal the features are localized in photon energy it is possible to account for the lineshapes to yield accurate values of energies and broadening parameters of the associated interband transitions. For example, the energies of various interband (intersub-band) transitions may be evaluated to within a few meV, even at room temperature. Thus, the effects of static external perturbations such as electric and magnetic fields, temperature, hydrostatic pressure, uniaxial stress, etc., can be conveniently studied.

A particularly useful form of modulation spectroscopy is electromodulation (EM) since it is sensitive to surface/interface electric fields and can be performed in contactless modes that require no special mounting of the sample. Under appropriate conditions the EM spectrum may exhibit above bandgap oscillatory features, called Franz–Keldysh oscillations (FKOs), which are a direct measure of the relevant electric field (built-in and/or applied).

Two of the most widely used forms of EM are photoreflectance (PR) and contactless electroreflectance (CER). In PR, modulation of the built-in electric field is caused by photo-excited electron–hole pairs created by a pump source of ~3–5 mW power of either (1) an internally modulated laser diode or (2) a mechanically chopped dc laser. The modulating frequency is typically ~100–200 Hz. CER utilizes a condenser-like system consisting of a front wire grid electrode with a second metal electrode separated from the first electrode by insulating spacers, which are ~0.1 mm larger than the sample thickness. The sample is placed between these two capacitor plates. Thus, there is nothing in direct contact with the front surface of the sample. The probe beam is incident through the front wire grid. EM is achieved by apply an ac voltage (~1 kV peak to peak at ~200 Hz) across the electrodes.

Instrumentation

Shown in Figure 1 is a schematic drawing of the experimental arrangement for a general 'external' modulated reflectance experiment (except for PR). Light from an appropriate light source (xenon arc or tungsten halogen lamp) passes through a monochromator. The exit intensity at wavelength λ, $I_o(\lambda)$, is focused onto the sample by means of a lens (or mirror). The modulation (electric field, stress, temperature) is applied to the sample at frequency Ω_m. The reflected light is collected by a second lens (mirror) and is focused onto an appropriate detector

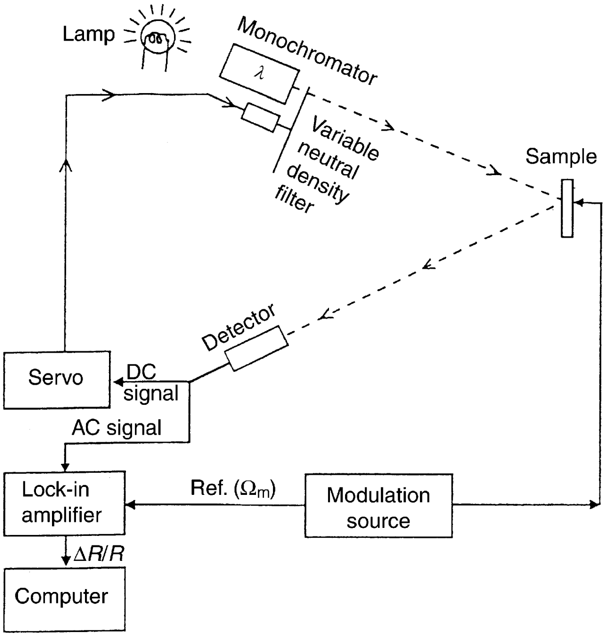

Figure 1 Schematic representation of external modulation experimental apparatus.

(photomultiplier or photodiode). For the sake of simplicity the two lenses (mirrors) are not shown.

The light striking the detector contains two signals: the dc is given by $I_o(\lambda)R(\lambda)$, where $R(\lambda)$ is the dc reflectance of the material while the modulated value (at frequency Ω_m) is $I_o(\lambda)\Delta R(\lambda)$, where $\Delta R(\lambda)$ is the change in reflectance produced by the modulation source. The ac signal from the detector, proportional to $I_o(\lambda)\Delta R(\lambda)$, is measured by a lock-in amplifier. Typically $I_o\Delta R$ is 10^{-4}–10^{-6} of I_oR. In order to evaluate the quantity of interest, i.e., the relative change in reflectance $\Delta R/R$, a normalization procedure must be used to eliminate the uninteresting common feature $I_o(\lambda)$. This can be done in several ways, one of which is shown in Figure 1, where the normalization is performed by a variable neutral density filter (VNDF) connected to a servo mechanism. The dc signal from the detector, which is proportional to $I_o(\lambda)R(\lambda)$, is fed into the servo which moves the VNDF in such a manner as to maintain a constant $I_o(\lambda)R(\lambda)$, i.e., $I_o(\lambda)R(\lambda) = C$. Under these conditions the ac signal $I_o(\lambda)\Delta R(\lambda) = C\Delta R(\lambda)/R(\lambda)$. Subsequently, the signal to the lock-in amplifier is proportional to the quantity of interest, i.e., $\Delta R(\lambda)/R(\lambda)$.

Bulk/Thin-Film Material

Modulation spectroscopy has been used to study the properties of bulk/epitaxial layered semiconductors such as ordering in $GaInP_2$, piezoelectric fields in strained GaN, the influence of isotopes, the properties of low-temperature grown GaAs, the space charge region, carrier concentration, the nature of the band bending (carrier type), impurity effects, deep levels, alloy composition, the influence of static external or internal perturbations such as temperature, external uniaxial stress, lattice-mismatch strain, hydrostatic pressure, the effects of crystal growth/processing/annealing procedures, and amorphous and microcrystalline materials.

The dashed lines in Figure 2 are the piezoreflectance (PZR) spectra in the region of the direct gap of ReS_2 in the temperature range $25\,K < T < 450\,K$. The data at low temperatures exhibit two well-resolved features labeled E_1^{ex} and E_2^{ex}. The solid lines are lineshape fits to the first derivative of a Lorentzian profile which yields the energies indicated by arrows and also the related broadening parameters. The solid/open squares and circles in Figure 3 are the temperature dependence of E_1^{ex}/E_2^{ex} of ReS_2 and $ReSe_2$, respectively. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively. The Varshni equation is:

$$E_i^{ex}(T) = \frac{E_i^{ex}(0) - \alpha_i T^2}{\beta_i + T} \qquad [1]$$

where $i = 1$ or 2, $E_i^{ex}(0)$ is the excitonic transition energy at 0 K and α, β are the Varshni coefficients. The O'Donnel–Chen expression is:

$$E_i^{ex}(T) = E_{i0} - S_i\langle\hbar\Omega_i\rangle\left(\frac{\coth\langle\hbar\Omega_i\rangle}{2kT} - 1\right) \qquad [2]$$

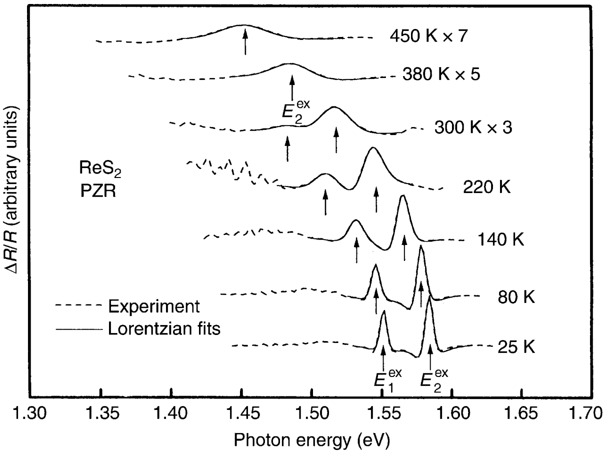

Figure 2 The PZR spectra (dashed lines) in the region of the direct gap of ReS2 in the temperature range $25\,K < T < 450\,K$. The solid lines are lineshape fits yielding the energies indicated by the arrows.

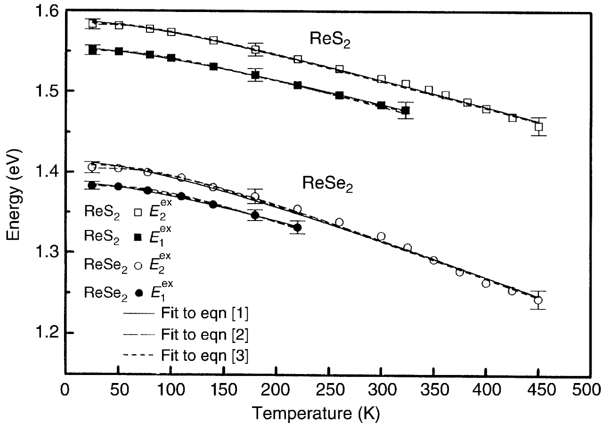

Figure 3 The temperature dependence of the excitonic transition energies of ReS$_2$ and ReSe$_2$. The solid, dot-dashed, and dashed lines in the figure are fits to the Varshni, O'Donnel–Chen, and Bose–Einstein expressions, respectively.

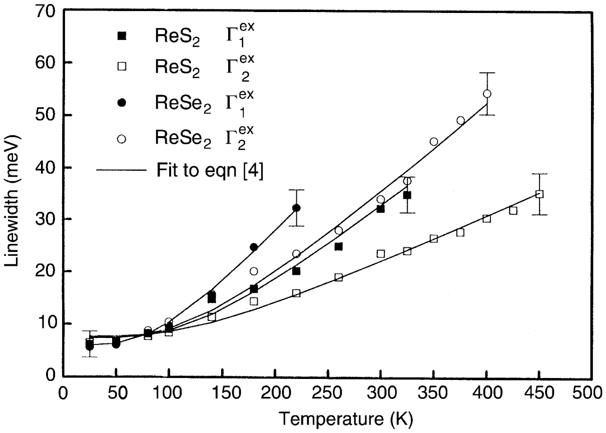

Figure 4 The solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS$_2$ and ReSe$_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression.

where E_{i0} is the excitonic transition energy at 0 K, S_i is a dimensionless coupling constant related to the strength of the electron–phonon interaction, and $\langle \hbar\Omega_i \rangle$ is the average phonon energy. The Bose–Einstein-type expression can be written as

$$E_i^{ex}(T) = E_{iB} - a_{iB}\left[1 + \frac{2}{\exp(E_{ap}/kT) - 1}\right] \qquad [3]$$

where a_{iB} represents the strength of the electron (exciton)–phonon interaction, and E_{ap} corresponds to the average phonon energy.

The temperature dependence of the broadening parameters, $\Gamma(T)$, is due only to the electron–phonon interaction and can also be described by a Bose–Einstein-type equation. For the lowest lying direct gap only the longitudinal optic (LO) phonon is involved. In Figure 4 the solid/open squares and circles are the temperature dependence of the broadening parameters of E_1^{ex}/E_2^{ex} of ReS$_2$ and ReSe$_2$, respectively. The solid lines are fits to the Bose–Einstein-type expression:

$$\Gamma_i(T) = \frac{\Gamma_i(0) + \Gamma_{i,ep}}{\exp(E_{i,LO}/kT) - 1} \qquad [4]$$

where $i = 1$ or 2. The first term of eqn [4] corresponds to broadening mechanisms due to intrinsic lifetime, electron–electron interaction, impurity, dislocation, and alloy scattering effects. The parameter Γ_{ep} is an electron (exciton)–phonon coupling coefficient and E_{LO} is the zone-center LO phonon energy. The obtained values of Γ_{ep} for these materials have been compared with those for the direct gap in GaAs and ZnSe.

Surfaces/Interfaces

Almost since the inception of modulation spectroscopy it has been recognized that EM could be used as an effective probe of surfaces and interfaces. For EM this is true not only because of the sharp, derivative-like spectral features but also due to its sensitivity to electric fields, i.e., FKOs. In addition, PR and CER are of considerable interest and usefulness since they are contactless, employ low light levels, require no special mounting of the sample, and can be performed in any transparent ambient including ultrahigh vacuum (UHV).

Of particular usefulness in the study of Fermi level pinning at surfaces are UN(SIN)/UP(SIP) and δ-doping configurations. The UN(SIN)/UP(SIP) structures are made by fabricating an undoped layer of thickness L ($\sim$1000 Å) on a buried doped n$^+$(p$^+$) buffer on a doped n$^+$(p$^+$) substrate. In the n$^+$(p$^+$) buffer/substrate the Fermi level occurs near the conduction (valence) band edge. At the surface the Fermi level is pinned at some value, V_F. Therefore, there exists in the undoped region, which has a small broadening parameter, a large, almost constant electric field, F. A similar constant field can be created by placing a δ-doping layer a distance L from the surface.

The relation between F, as observed from the FKOs, and V_F is given by:

$$FL = V_F - (kT/q) - \mathrm{SCC} - V_p(T) \qquad [5]$$

where the second, third, and fourth terms on the right-hand side are the Debye length, space charge layer, and temperature dependent photovoltaic effect corrections, respectively.

The properties of the metal–semiconductor interface have been reported by a number of investigators using PR and/or ER. Jonker *et al.* have demonstrated that an epitaxial Fe film on GaAs (UN structure) provides a ferromagnetic contact, suppresses midgap state formation, and does not pin the Fermi level. Four samples were investigated: a native oxide reference sample, the same sample after sulfur passivation, Al/GaAs-(2 × 4), and Fe/GaAs-(2 × 4). The width of the undoped GaAs layer was 1070 Å for the first three samples and 1500 Å for the last sample. These results are shown in Figure 5a–d, respectively. The fields obtained from the FKOs are 58 kV/cm, 61 kV/cm, 55 kV/cm and 60 kV/cm. For the first three samples the measured barrier height is about 0.64 eV (not corrected for the photovoltaic effect) while for the Fe/GaAs material it is about 40% higher (~0.9 eV). Another significant feature of the Fe/GaAs spectrum is the large number of FKOs (in relation to the other samples) extending to about 2 eV. The significant reduction in damping indicates a substantial increase in carrier lifetimes.

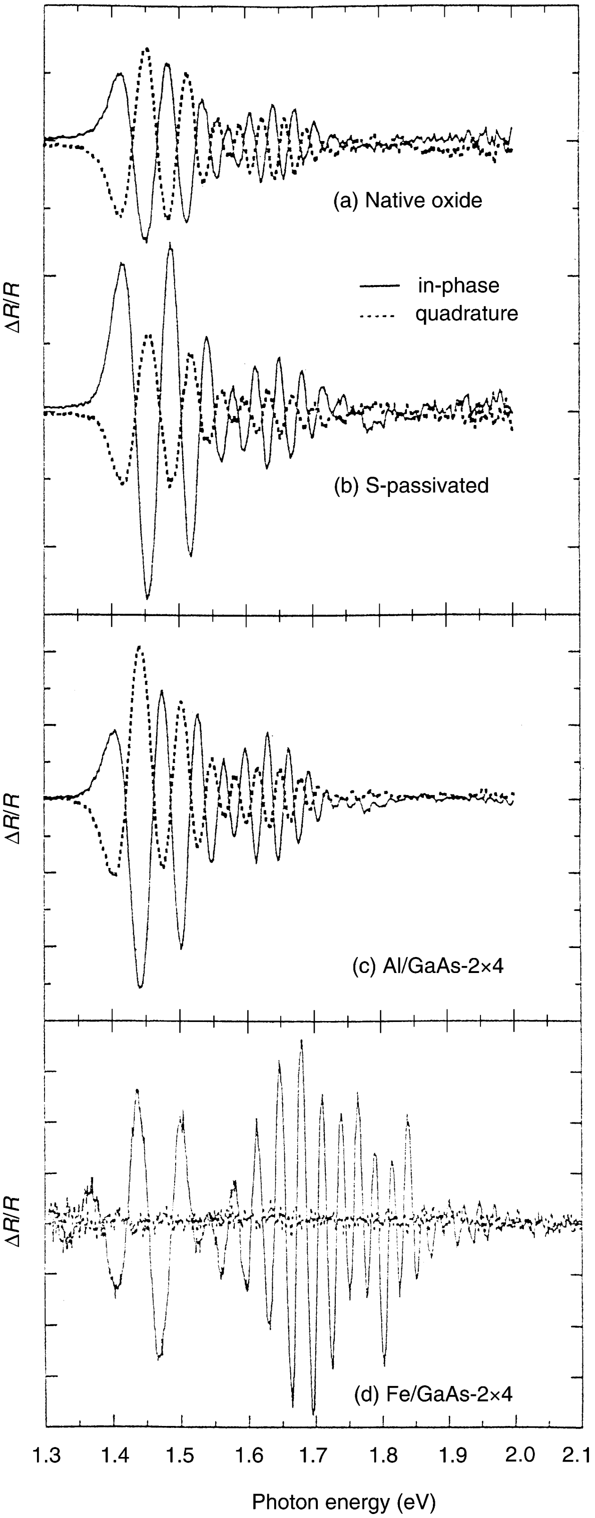

Figure 5 The PR spectra at 300 K for (a) a native oxide reference sample, $F = 58$ kV/cm; (b) the same sample following sulfur passivation, $F = 61$ kV/cm; (c) Al/GaAs-(4 × 2) sample, $F = 55$ kV/cm, and (d) Fe/GaAs-(2 × 4) sample, $F = 60$ kV/cm. The thickness of the undoped spacer region is 1070 Å for (a)–(c) and 1500 Å for (d).

Micro- and Nanostructures

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental aspects of compositional single quantum wells (SQWs)/multiple quantum wells (MQWs), superlattices (SLs), quantum wires, and quantum dots (QDs). Various phenomena can be studied, including band offset, well and barrier widths, excitons, strain, coupling and decoupling between wells, miniband formation, two-dimensional electron gas (2DEG) effects, zone-folding in short-period SLs, built-in electric fields, etc. Work on a large variety of GeSi, III–V and II–VI systems has been reported, including both lattice-matched and strain-layer configurations. In addition the effects of various external perturbations can be evaluated.

The properties of a vertically coupled InAs/GaAs QD-based laser structure fabricated by a self-assembled technique have been investigated by CER (300 K and 20 K). Signals have been observed from all the relevant regions of the sample. The dotted lines are the CER spectra at 300 K and 20 K in Figure 6a,b, respectively. The solid lines are lineshape fits which yield the energies indicated by arrows. The low-energy features, labeled QD_0 (lateral coupling), QD_1 (vertical coupling) and QD_2 (vertical coupling), originate from the QDs while the resonances designated W_1 and W_2, correspond to the first conduction to first heavy- (1C-1H) and light-hole (1C-1L) transitions, respectively, in the QW (approximately one monolayer) formed by the

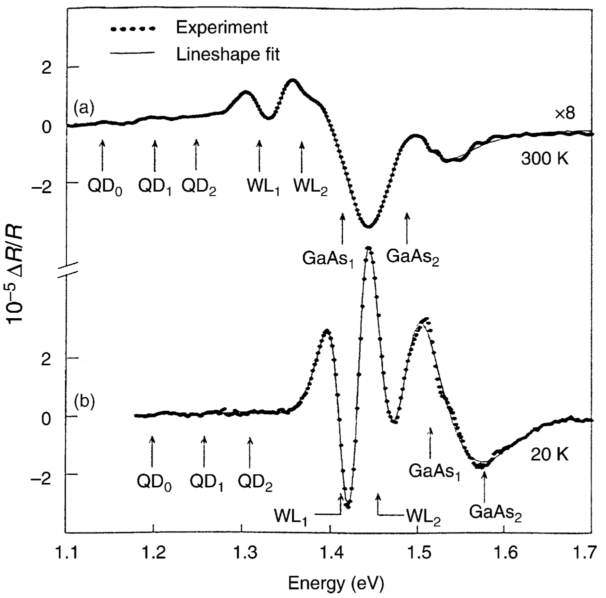

Figure 6 CER spectra (dotted lines) of an InAs/GaAs quantum dot-based laser structure at (a) 300 K and (b) 20 K. The solid lines are a lineshape fit to a first derivative Gaussian.

wetting layer (WL). Although the oscillator strength of the QD features is small compared to the WL resonances, they are clearly visible even at 300 K. The GaAs$_1$ and GaAs$_2$ features are attributed to GaAs-related transitions. The former at both 300 K and 20 K corresponds to the bandgap energy of GaAs and probably comes from the undoped 400 Å-thick GaAs layers which envelop the QD region. The GaAs$_2$ feature is most likely related to the p-doped 6000 Å-thick GaAs cap layer. The heavy doping of this layer induces a Burstein–Moss blue shift of the bandgap.

Device Structures

The methods of PR and CER are valuable tools in the evaluation of important device parameters for structures such as heterojunction bipolar transistors (HBTs), pseudomorphic high electron mobility transistors (PHEMTs), edge-emitting and vertical cavity surface-emitting (VCSEL) quantum well lasers, multiple quantum well infrared detectors, solar cells, superlattice optical mirrors, resonant tunneling structures, metal-oxide-semiconductor (MOS) configurations, among others.

The dashed lines in Figure 7a,b are the 300 K CER spectra of a GaInP$_2$(emitter)/GaAs(collector) HBT (fabricated by metalorganic chemical vapor deposition) for incident light with electric field vector *E*||[110] (polarization A) and *E*||[1$\bar{1}$0] (polarization B), respectively. Both collector and emitter signals exhibit well-pronounced FKOs. The solid lines are least-squares fits to a function which contains Lorentzian broadened electro-optic functions, yielding 1.873 ± 0.001 eV and 1.885 ± 0.001 eV for the GaInP$_2$ bandgaps for polarizations A and B, respectively, as indicated by the arrows. The ordering parameter of 0.3 deduced from the differences in the measured bandgaps for the two polarizations was consistent with transmission electron microscope measurements. The electric fields evaluated from the emitter and collector region FKOs were compared with a computer simulation of the field profiles using a comprehensive, self-consistent model, including the photovoltaic effect. The deduced emitter and collector doping densities were in good agreement with the intended growth conditions.

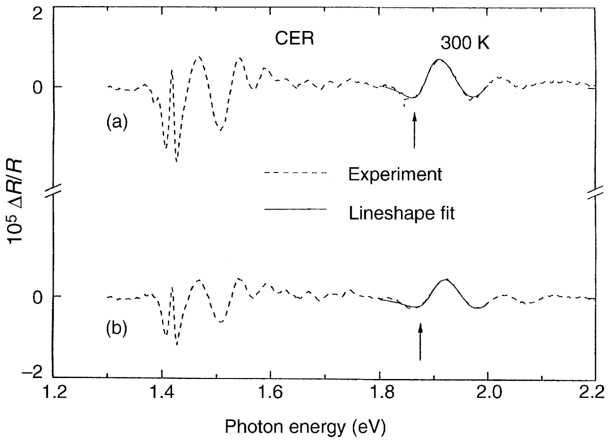

Figure 7 Room-temperature CER spectra (dashed lines) for a GaInP/GaAs HBT sample fabricated by metalorganic chemical vapor deposition for incident light polarized (a) ***E***||[110] and (b) ***E***||[1$\bar{1}$0]. The solid lines are a lineshape fit which yields the bandgap energies denoted by arrows.

Summary

Modulation spectroscopy is a very powerful tool for investigating many of the fundamental and applied aspects of a wide variety of semiconductors and semiconductor micro- and nanostructures, including bulk/thin films, surfaces/interfaces, single quantum wells, multiple quantum wells, superlattices, quantum wires, quantum dots, etc. In addition the effects of various perturbations such as alloy composition, temperature, electric and magnetic fields, hydrostatic pressure, internal and/or external stain, processing, etc., can be evaluated. Recently these methods have also been applied for the characterization of actual device structures such as HBTs, PHEMTs, quantum well lasers, etc.

See also

Semiconductor Materials: Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures; Surface Photovoltage Spectroscopy of Semiconductors.

Further Reading

Aspnes DE (1980) Modulation spectroscopy/electric field effects on the dielectric properties of semiconductors. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 109–154. Amsterdam, The Netherlands: North-Holland.

Glembocki OJ and Shanabrook BV (1992) Photoreflectance spectroscopy of microstructures. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 222–289. New York: Academic Press.

Pollak FH (1994) Modulation spectroscopy of semiconductors and semiconductor microstructures. In: Balkanski M (ed.) *Handbook on Semiconductors*, vol. 2, pp. 527–635. Amsterdam, The Netherlands: North Holland.

Pollak FH (1995) Electromodulation spectroscopy of semiconductor surfaces and interfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 175–209. Amsterdam, The Netherlands: North-Holland.

Pollak FH (1996) Modulation spectroscopy of quantum wells and superlattices. In: Bhattacharya P (ed.) *Properties of III–V Quantum Wells and Superlattices*: *EMIS Data Reviews Series #15*, pp. 232–240. London: INSPEC.

Pollak FH (1998) Modulation spectroscopy of the group III-nitrides. In: Gil B (ed.) *Group III Nitride Semiconductor Compounds*, pp. 158–181. Oxford, UK: Clarendon.

Pollak FH (2001) Modulation spectroscopy characterization of semiconductors and semiconductor microstructures. In: Buschow KH, Cahn W, Flemmings MC, *et al.* (eds) *Encyclopedia of Materials: Science and Technology*, vol. 6, pp. 5743–5748. Amsterdam, The Netherlands: Elsevier.

Ramdas AK and Rodriguez S (1992) Piezospectroscopy of semiconductors. In: Seiler DG and Littler CL (eds) *Semiconductors and Semimetals*, vol. 36, pp. 137–220. New York: Academic Press.

Quantum Dots

R J Warburton, Heriot-Watt University,
Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Quantum dots are nanometer-sized clusters of semiconductor material. An island of low-bandgap material is embedded in a sea of high-bandgap material so that both electrons and holes are spatially confined in all three directions. In an unstructured semiconductor, the electron energy levels of the individual atoms hybridize to form energy bands. In a quantum dot, however, there are discrete energy levels. This can be understood from the band structure of the unstructured material where there is a so-called dispersion relation connecting the energy, E, and the momentum, k. The confinement potential of a quantum dot quantizes the k-values, giving rise to discrete energy levels. If the energy levels of a quantum dot are separated by more than the typical energy level broadenings and by more than the thermal energy then the quantum dot will behave quite differently to the bulk material. The quantum dot will exhibit atom-like properties – discrete energy levels with small degeneracies – despite the fact that it is embedded in a semiconductor matrix. It is this ability to prepare a variety of atom-like systems in a semiconductor environment which has driven the research into semiconductor quantum dots.

In some sense, nanosized quantum dots have been the subject of research for quite some time. Color filters, for instance, consist of nanosized particles dispersed in a glass matrix, and this is a very mature technology. Also, lithography of layered semiconductor systems already possessing confinement in one dimension has been pursued for at least 15 years. However, developments in the past few years have been rapid, all based on the discovery of self-assembly of quantum dots directly in the growth of semiconductor heterostructures. These quantum dots can be produced with standard materials in standard growth systems, and this allows an easy integration of quantum dots into devices such as diode lasers. Furthermore, the self-assembled quantum dots are of extremely high quality and this is of great value not just for devices but also for fundamental physics experiments, aimed at studying or exploiting the coherence of excitons confined in the quantum dots. For these reasons, this short review concentrates on the properties of self-assembled quantum dots, giving just brief descriptions of other systems. It should be noted that the bibliography is extremely limited and in no way does justice to the large number of research groups which have made valuable contributions to this field. Instead, the bibliography is simply a list of suggestions for finding out more about this fascinating research area.

Quantization in Quantum Dots

Quantum dots are on the one hand large enough that the bulk band structure of the host material can be used to describe the material's electronic properties. In the effective mass approximation, this amounts to giving both electrons and holes a quadratic dispersion on momentum, just as for a free electron, but with an effective mass, typically about 0.1 for the electrons, and 0.3 for the holes. On the other hand, quantum dots are small enough that confinement effects are important. For instance, the energy of an exciton in a nanometer-sized quantum dot is a strong function of the quantum dot size. As the dots become smaller, confinement effects become larger, and the energy separating electron and hole levels increases. This does not increase indefinitely, however. There is a finite-size energy barrier to the matrix surrounding the dot such that in the limit of small dot sizes, the wavefunction spreads out into the barrier region and confinement effects become very weak. Ideally, an electron in a quantum dot has a confinement energy of at least several meV, and an energy barrier to the surrounding matrix of several 100 meV so that confinement effects are visible even at room temperature.

The motivation for the study of quantum dots is to combine the atom-like properties with the flexibility of a solid-state environment. This is perfectly illustrated by considering a quantum dot as the gain medium for a laser. In a conventional semiconductor diode, a band of levels has to be inverted in order to achieve population inversion. This is costly in terms of the threshold current and the temperature stability. In a quantum dot, inversion can be achieved simply by inverting the population of the two-fold degenerate ground state, which is clearly trivial. Furthermore, a strong confinement energy and a large barrier to the surrounding matrix imply that this population inversion is stable against rises in temperature, implying better room-temperature performance. This example also exemplifies the motivation for investigating quantum dots using well-developed inorganic semiconductor technology as in this case the dots can be integrated into established device geometries. Generally speaking, the important concept in quantum dot physics is spatial confinement of electrons in all three directions, giving rise to discrete energy levels.

Lithography for Quantum Dots

Developments in semiconductor growth technology, largely during the 1970s and 1980s, enabled thin layers of different semiconductors to be grown on top of each other. There are two principal growth techniques, molecular beam epitaxy (MBE), in which a heated substrate is bombarded with pure species in a high vacuum, and metal organic chemical vapor deposition (MOCVD), in which chemical precursors flow over a heated substrate and react. In the simplest structure, termed a quantum well, a thin layer of low-bandgap material is sandwiched between high-bandgap material. The discontinuities in the potentials at the interface give rise to confined electronic states for both carrier types, electrons and holes. The quantum well is an example of a two-dimensional system: there is confinement in one direction, the growth direction, but free dispersion in the other two directions. To make a quantum dot, it is necessary to confine the electron and hole in all three directions. An obvious approach is to take a quantum well, with perfect confinement in one direction, and process it laterally with lithographic techniques to complete the quantization. Note that an analogy can be made to the physics of the quantum Hall effect, where a two-dimensional electron gas is subjected to a very high magnetic field. The magnetic field quantizes the lateral motion in a similar way to the electrostatic potential in a quantum dot.

This approach to the formation of optically active quantum dots obviously has some advantages, notably that the shape and position of the quantum dots can be chosen in the lithography. However, there are some significant problems associated with these techniques. As most of these problems do not plague the properties of self-assembled quantum dots, self-assembled quantum dots are more suitable for advanced photonic devices.

A number of different lithographic techniques has been used to process quantum-well heterostructures into quantum dots. For research applications, electron-beam lithography is the method of choice. The sample is coated with a resist, for instance PMMA, and exposed with electrons in an area with nanometer size. A protective layer, either a metal or insulator, is transferred to the exposed area by developing the resist, by depositing the protective layer material, and by removing the unwanted material with a lift-off step. The wafer is then etched, typically in a dry environment, to produce a free-standing column.

The photoluminescence of these quantum dots has been measured by various research groups, and although the results vary quite widely there are some systematic trends. The photoluminescence efficiency decreases rapidly for pillars smaller than a few microns. This is somewhat improved at low temperature, where reasonably strong emission can be observed down to 100-nm diameter pillars. For pillars with micron diameters, the lateral confinement effects are weak, and it is debatable if true quantum dots have been achieved. There are a number of effects which degrade the emission. First, pinning of the Fermi energy by surface states generates large electric fields near the surface, which can ionize electron–hole pairs. This is well known for quantum wells: if a GaAs quantum well for example is placed less than about 25 nm from the surface, the photoluminescence is very weak because of the surface electric field. In an etched pillar, these effects are clearly magnified by the new surface area generated. Secondly, the etching process damages the material to a greater or lesser degree, and the damaged region contains centers for nonradiative recombination.

Other lithography techniques have also been used to generate semiconductor quantum dots for optical experiments. An elegant idea is to deposit small regions of a highly strained film onto the sample surface. In this way, a laterally varying strain film in an underlying quantum well can be set up, generating quantum dots through the effect of strain on the band gap. (This idea has also been married with self-assembly, by using Stranski–Krastanow-grown InP stressors on a GaAs surface, generating lateral confinement in an underlying InGaAs quantum well.) A further idea is to generate a quantum dot by deliberately growing a quantum well with rough interfaces. In such a system, there will inevitably be regions which confine an electron–hole pair through localization. Both of these approaches clearly circumvent the problems generated by the exposed surface in an etched quantum dot. However, they generate small lateral confinement potentials, typically a few meV. This is of value in a research environment where fundamental experiments can be carried out at low temperature, but in a possible device at room temperature, where the thermal energy is 25 meV, quantization effects will not be discernible.

Self-Assembled Quantum Dots

A solution to the difficulties inherent in quantum dots made by lithographic techniques began to emerge in the early 1990s. It was discovered that quantum dots can be produced directly in the semiconductor growth by depositing a material onto a substrate with a substantially smaller lattice constant. The most well-known example is the growth of InAs (lattice constant 6.05 Å) on GaAs (lattice constant 5.65 Å) where In-rich quantum dots form. Because billions of dots form spontaneously in this process, the dots are referred to as self-assembled quantum dots. Since the early experiments, considerable progress has been made, such that the state-of-the-art quantum dots possess many of the hoped-for properties: they are coherent (meaning that they contain no dislocations), highly homogeneous (meaning that fluctuations in shape and composition from dot to dot are small), and have large quantum efficiencies for the emission of light.

Growth of Self-Assembled Quantum Dots

When growing one material on top of another, three principal growth mechanisms have been identified: Frank–van der Merwe, Volmer–Weber, and Stranski–Krastanov. In the Frank–van der Merwe mode, material is deposited layer by layer, as in the growth of the lattice-matched pair GaAs and AlAs. In the Volmer–Weber mode, island formation occurs: the material does not wet the surface because it is energetically unfavorable. Another possibility arises if there is a lattice mismatch. In this case, the epilayer is strained, and the strain energy obviously increases with increasing epilayer thickness. In the Stranski–Krastanow mode, the initial growth is layer by layer, as in Frank–van der Merwe, but beyond a certain thickness, islands form, as shown schematically in Figure 1. On the one hand, the islands limit the strain energy, because some strain relaxation is allowed, which is not possible in a thin film. On the other hand, island formation results in an increase of the

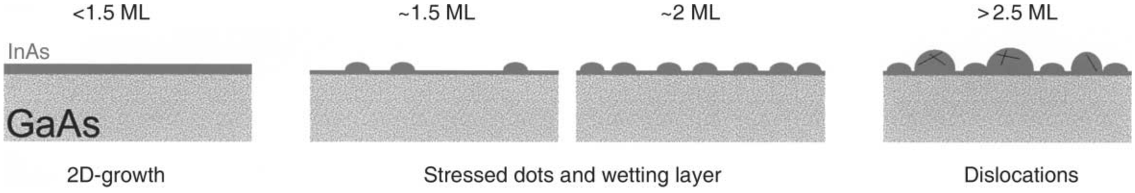

Figure 1 A schematic of the Stranski–Krastanow mode. The material deposited has a substantially larger lattice constant than the substrate, for example InAs on GaAs. Initially, a thin and uniform layer is deposited. At the critical thickness (1.5 monolayers for InAs on GaAs), islands spontaneously form. Further growth results in an increase in the number of islands. At large coverages, large and dislocated islands appear.

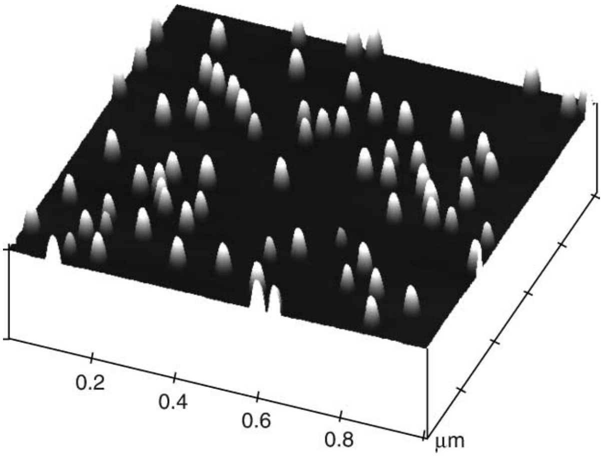

Figure 2 An atomic force micrograph of InAs quantum dots on a GaAs substrate. The InAs dots were grown by MBE at a growth temperature of 530 °C. The image was taken by Axel Lorke.

surface energy. This means that there is a size at which the total energy is minimized.

It is now well established that many combinations of semiconductors grow in the Stranski–Krastanow mode, and this can be utilized for self-assembly of quantum dots. The most studied example is InAs on GaAs where high-quality quantum dots can be grown with both MBE and MOCVD. Dots form at the so-called critical thickness of 1.5 monolayers, corresponding to just 4 Å of material. For thicknesses less than the critical thickness, the covering is not completely uniform; there are island-like structures, elongated along the $[0\bar{1}1]$ direction. At the critical thickness dots form, leaving a thin InAs layer (the wetting layer) between the dots. As more InAs is added, the density of dots increases with only small changes in the dot size. As an example, Figure 2 shows InAs dots on a GaAs substrate.

The dots grown in this Stranski–Krastanow mode are free of dislocations, which is an essential prerequisite in III–V compounds for a high quantum efficiency for optical emission. Furthermore, the dots are remarkably homogeneous. Statistical fluctuations during growth will always give rise to a distribution in dot size and height. Nevertheless, the fluctuations in dot size can be as small as ~10%. The dots are randomly arranged in the lateral plane if the growth proceeds on a flat substrate, and densities are between 10^9 and 10^{11} cm^{-2}. The shape and composition of the dots depend on the growth parameters. For instance, InAs/GaAs dots can have facets along particular crystal directions, or they can be lens-shaped without facets, depending on the growth technique, and in particular the growth temperature. In fact, measuring the shape and composition is a very challenging task, particularly for buried dots.

Stranski–Krastanow growth has turned out to be a robust phenomenon for III–V materials; quantum dots have been produced in this way for a number of material combinations. The most important condition is that the deposited material has a substantially larger lattice constant than the substrate. InAs forms dots not just on GaAs but also on InP. The InAs/InP dots are particularly interesting because they emit at the technologically important 1.5 µm wavelength. InP can also form the dot material by using GaInP substrates where the bandgap is pushed up into the red region of the spectrum. Dots can also be grown with II–VI materials; the growth of CdSe on ZnSe is analogous to the growth of InAs on GaAs. These II–VI dots emit in the green. Finally, dots can also be formed in the nitrides, for instance by depositing InGaN on an AlGaN surface. In fact, the high emission efficiency of GaN, despite the large defect density, has been attributed to the formation of quantum dot-like structures in the active layers.

It is a challenge to arrange the dots in the plane of the substrate. An easy route with self-assembly has not been found to accomplish this. Nevertheless, dots can be encouraged to grow at particular positions by pre-patterning the substrate. This obviously limits the separation of the dots in any such array to the limits set by lithography. However, self-assembly does provide a means to order the dots in the growth direction: if the separation between successive layers is small enough (less than 15 nm for InAs/GaAs) then the dots grow directly above one another, as shown in Figure 3. In a new layer, the residual strain field from the dots in the layer below is sufficient to seed the dots. In this way, vertical stacks of dots can be built up. In fact, it has been argued that the dots tend to become more laterally ordered with each layer. The point is that if two dots are close together in one layer, their strain fields overlap, and provide only one seeding center in the subsequent layer.

Optical Properties

The main interest in quantum dot physics is the quantization in all three spatial directions, giving rise to atomic-like energy levels. For a system of self-assembled quantum dots, it is important to measure the fundamental bandgap, the energetic separation between the confined states both for electrons and for holes, the barriers to the electron continuum and hole continuum, and the oscillator strengths for the various interband transitions. All this information can be gleaned with optical spectroscopy.

Perhaps the simplest experiment to perform, if not to understand, is photoluminescence in which electron–hole pairs are excited with a laser beam.

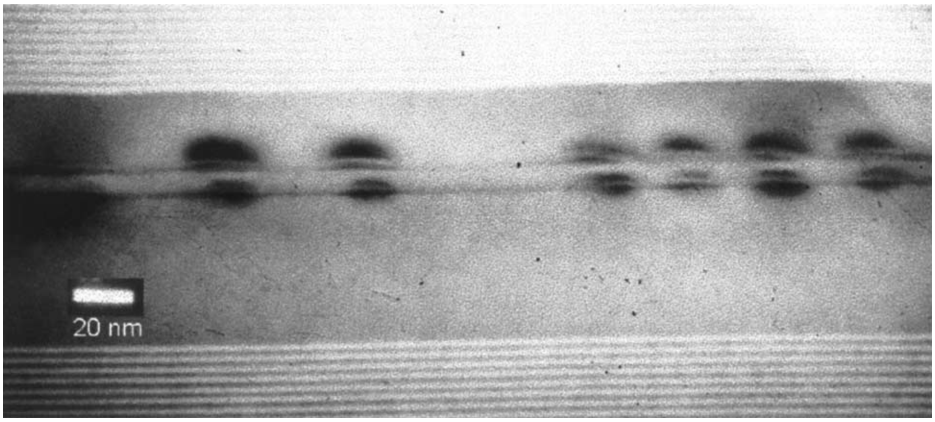

Figure 3 A transmission electron micrograph of two layers of InAs quantum dots in GaAs showing how the dots in the upper layer lie directly above the dots in the first. The image was taken by Gio Medeiros-Ribeiro.

The laser energy can be tuned to lie just above the bandgap of the wetting layer in which case the electron–hole pairs are excited in the wetting layer itself. Alternatively, if the laser energy is larger than the bandgap of the barrier material, electron–hole pairs are excited in the barrier. In both cases, the carriers relax into the dots, and then down the ladder of states in the dot, so that the dot electron and hole ground states are occupied. Emission is detected when an electron and a hole recombine. As such, the experiment measures only the energy of the ground-state exciton. However, if the pump intensity is increased such that electron–hole pairs are generated faster than they can recombine, the excited states of the dots are also occupied and recombination can also occur from excited states. An example of such an experiment is shown in Figure 4. At low pump powers, only the ground state emission is detected; at higher pump powers, the first excited state emerges, and at higher powers still, the second excited state also emerges, and so on. The spectra are broadened by the inhomogeneous broadening, i.e., from fluctuations from dot to dot, but the broadening is small enough that the excited state emission can be easily distinguished from the ground state emission. For these particular quantum dots, the energetic separation between the ground and excited state emission is 15 meV, and this corresponds to the sum of the electron and hole confinement energies. The separation between the other emissions is also about 15 meV, implying that the states are approximately equally spaced, implying in turn that the confining potential is approximately parabolic. Furthermore, in this experiment, emission from the wetting layer can also be made out, and this enables the electron and hole confining potentials to be estimated.

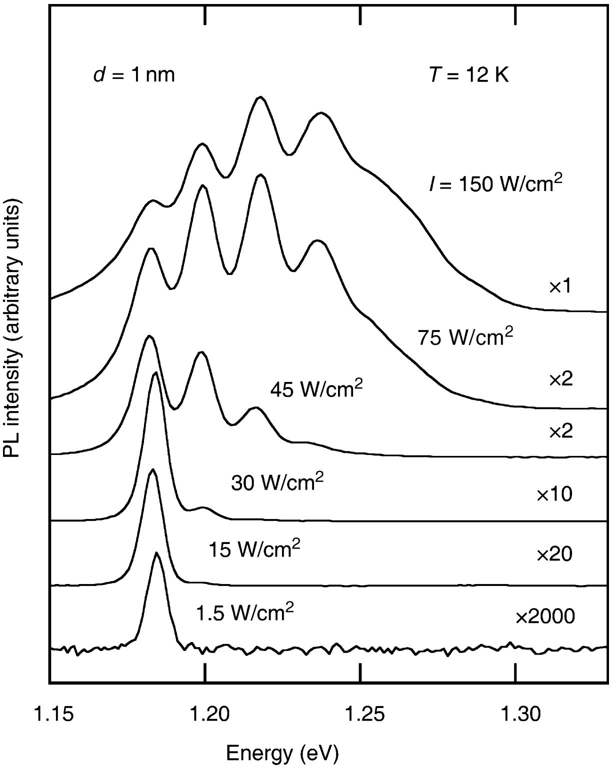

Figure 4 Photoluminescence from an ensemble of quantum dots as a function of laser excitation power. The measurement temperature was 12 K. The quantum dots are induced in an InGaAs/GaAs quantum well by a strain field generated by InP stressors on the sample surface. The stressors are grown in the Stranski–Krastanow mode, and are 1 nm away from the quantum well. At low excitation power, only the ground state emission is observed. As the power increases, emission from excited states is also observed through state filling. Reprinted with permission from Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review* 51: 13868. Copyright 1995 American Physical Society.

Photoluminescence invariably yields the sum of electron and hole energies. In order to measure just the electronic properties, it is possible to occupy the dots with, say, two electrons (so that in each dot the two-fold degenerate ground state is fully occupied) and to excite electric dipole transitions between the electron levels. This has been accomplished with so-called charge-tunable heterostructures, with which electrons can be controllably loaded into the dots from an electron reservoir. The interlevel absorption lies in the far-infrared region of the spectrum; for InAs/GaAs quantum dots for instance, the transition is at a wavelength of 25 μm (an energy of 50 meV), as shown in Figure 5. The results of this experiment also point to an approximately parabolic confining potential from the fact that as electrons are added to the dots, the energy of the far-infrared absorption does not change much. (In the limit of a perfect parabolic potential, a famous result known as Kohn's theorem states that the long-wavelength absorption should show no energy dependence on the electron occupation.)

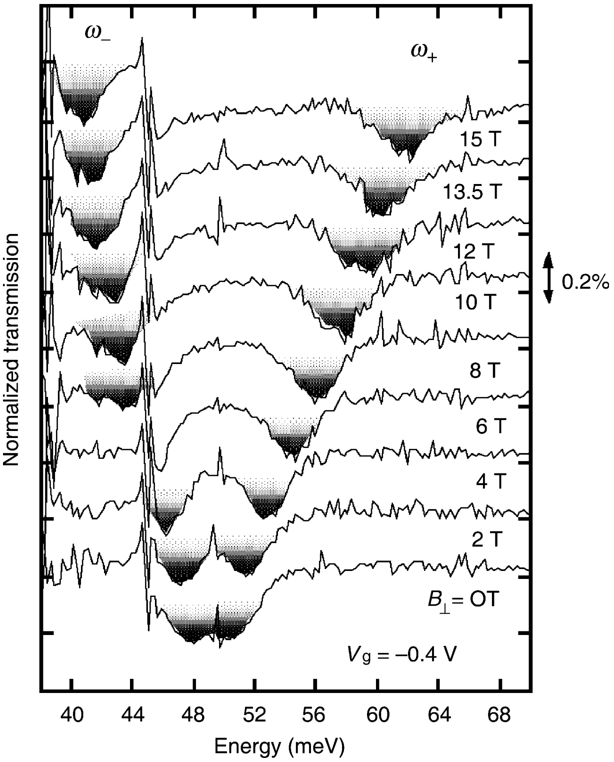

Figure 5 Transmission in the far infrared of an ensemble of InAs/GaAs quantum dots where each quantum dot is occupied with two electrons. The data were taken at a temperature of 4.2 K. The shaded transmission minima correspond to absorption from the quantum dots. The transitions are between the electron ground state and the first excited state. Two peaks are observed even at zero magnetic field, suggesting that the dots are slightly oval in shape. In a magnetic field, the two peaks move apart, which is a consequence of the Zeeman effect. The strong feature at 45 meV arises from an electronic interaction with an interface phonon. Reprinted with permission from Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron-electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197. Copyright 1996 European Physical Society.

In self-assembled quantum dots, the confining potential can be thought of as very steep in the growth direction, and relatively shallow in the lateral plane. In other words, the dots are essentially two-dimensional disks. The electron ground state is, in analogy with atomic physics, s-like, and the excited state p-like. However, because of the disk-like nature of the dots, the p-state is four-fold degenerate, and not six-fold degenerate as in a conventional atom. One way to demonstrate this is to perform interband absorption experiments on charge-tunable structures, as shown in Figure 6. The first transition corresponds to the transition from the hole s-state to the electron s-state, and disappears when the dots are occupied with two electrons. This is simply because the Pauli principle forbids the transition once the final level is fully occupied. The second transition corresponds to the transition from the hole p-state to the electron p-state, and disappears when the dots are occupied with four electrons, confirming the degeneracies of

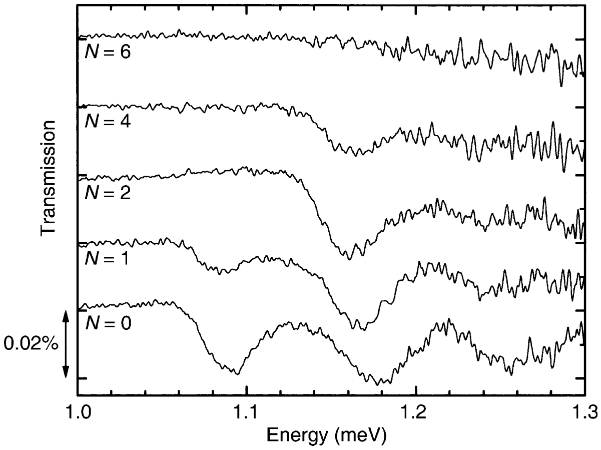

Figure 6 Transmission in the near infrared of an ensemble of InAs/GaAs quantum dots at 4.2 K. Spectra are shown for different values of *N*, the electron occupation. At $N = 0$, there are three transitions corresponding to transitions from a valence state to an electron state. These are the s–s, p–p and d–d transitions, where the states are labeled in analogy to atomic physics. At $N = 2$, the electron ground state is fully occupied, and the s–s transition disappears because it is blocked, a consequence of the Pauli principle. Similarly, the p–p transition is blocked at $N = 6$ when both the electronic s and p states are fully occupied. Reprinted with permission from Warburton RJ, Dürr CS, Karrai K *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282. Copyright 1997 American Physical Society.

the confined states. This experiment, unlike a photoluminescence experiment, also yields a value for the oscillator strength. The result can be understood simply in terms of the overlap integral between the hole and electron states. In other words, these dots are in the strong confinement regime where the confinement energies are larger than the exciton binding energy. In the other limit, the weak confinement regime, the exciton binding energy dominates over the confinement energies, and it makes sense only in this limit to think of an exciton moving as a coherent entity throughout the dot. In the weak confinement regime, the oscillator strength is actually larger than in the strong confinement regime.

An important issue in interpreting photoluminescence experiments is the rate at which a highly excited electron–hole pair can relax to the ground state of a quantum dot. Relaxation in the continuum is known to occur on a picosecond time-scale: electrons and holes relax by emitting LO-phonons if possible; and if not, acoustic phonons. Capture by a dot also occurs quickly (apart from in low-density samples where capture may take a few hundred picoseconds). It was initially thought that relaxation of carriers once captured by a quantum dot would be hindered by the discrete nature of the density of states. The argument is that the energy separation between the levels is too large for acoustic phonon emission; and there is no reason why the separation between the levels should match the LO-phonon energy (which has a very weak dispersion). Nonetheless, experiments have shown that relaxation from excited dot states to the ground state is rapid, occurring on a picosecond time-scale, pretty much as for quantum wells. Auger processes can be important, whereby an electron in a dot loses energy by promoting an electron in the continuum to a higher energy. But relaxation can also be fast without a large population of mobile carriers in the wetting layer, and this is thought to arise through multiphonon processes. The radiative lifetime is typically 1 ns for self-assembled quantum dots, depending to some extent on the size and material system. As for quantum wells, relaxation occurs much faster than recombination.

Lasers and Self-Assembled Quantum Dots

The motivation for using quantum dots as the gain medium for semiconductor lasers is that they should have advantageous properties over quantum wells. These include a lower threshold current, better temperature stability, and a higher modulation frequency. All of these possible advantages stem from the quantization. In a quantum dot, there are discrete levels each with low degeneracy so that population inversion is easy to achieve. Also, there are large potential barriers to the continuum of states in the surrounding matrix, and an energy separation between levels in the dot at least comparable to the thermal energy at room temperature. This should make a quantum dot device relatively insensitive to temperature. The threshold current of a semiconductor laser usually behaves as $I = I_0 e^{T/T_0}$ where T is the temperature, T_0 is the characteristic temperature, and I_0 the low-temperature threshold current. Generally speaking, lasers are desirable with a small I_0 and a large T_0.

The main obstacle to realizing the benefits of full spatial confinement in self-assembled quantum dots is the inhomogeneity. For a particular lasing wavelength only a fraction of the dots can contribute to the lasing signal. Additionally, the dot densities are typically about 10^{11} cm^{-2} for these applications, implying that in the layer plane there is more unoccupied space than dot material. All in all, the gain at a particular wavelength is limited. Nevertheless, it would appear that even with inhomogeneous dot ensembles the advantages inherent in zero-dimensional states can be exploited.

In addition to the density of states, there is an important difference in the nature of the gain spectrum between quantum well and quantum dot lasers. In a quantum well, gain exists over a range of energies through band filling. If an electron–hole pair is stimulated to emit by the lasing field, then the vacancies will be quickly filled by carrier–carrier scattering in the bands, maintaining a quasi-thermal equilibrium in the conduction and valence bands. In quantum dots at low temperature, capture into a particular dot is a random process, and given the rapid relaxation and the high barriers to the continuum, an electron–hole pair is very likely to emit from the dot where it was created. In other words, the system is not in thermal equilibrium. At higher temperatures, thermal escape out of the dots becomes likely, enabling the system to come into thermal equilibrium. The transition into thermal equilibrium will depend critically on the barrier height, and with appropriate design may take place close to, or even above, room temperature.

Quantum dot lasers have now been realized, and emit continuous wave at room temperature. Progress has been rapid in the past few years. In fact the best quantum dot lasers now have the lowest I_0 and the highest T_0 of any semiconductor laser diode.

Even if the intrinsic advantages of quantum dots are largely offset by the inhomogeneous broadening for laser applications, there are some other advantages. The InAs/GaAs quantum dot emission wavelength is significantly higher than that of InGaAs/GaAs quantum wells, and with subtle changes in the growth can be increased to 1.3 μm, corresponding to one of the low-loss windows in optical fiber. Also, much higher powers have been realized with quantum dot lasers than with quantum well lasers. In quantum well lasers, device failure occurs at high current, often due to degradation of the end mirrors. It would appear that the carrier localization in quantum dots is sufficiently strong to prevent diffusion to the end facets, allowing the devices to withstand higher currents.

Colloidal Quantum Dots

Progress has been made recently in making inorganic quantum dots with techniques from organic chemistry. Organometallic precursors are employed. The growth temperature is typically below 200 °C which has been found to give material with quite high crystalline quality. A crucial development has been the use of various organic groups to passivate the surface so that the quantum dots have bright photoluminescence. The II–VIs, being more polar than the III–Vs, are easier to produce with these techniques, and most work has concentrated on CdS and CdSe quantum dots which emit in the visible. Nevertheless, it is also possible to produce III–V quantum dots with this technology, and although the bulk GaAs bandgap lies in the near-infrared, the quantum confinement effects can be so high as to push the photoluminescence into the visible. Furthermore, HgCdTe has also been developed as a material system for quantum dots, and the emission can be tuned to the all-important 1.5 μm.

The spectral width of emission from these colloidal quantum dots tends to be very large because of a large spread in particle size. However, techniques are being developed to isolate a specific size. With a selective precipitation technique for instance, the spectral width can be reduced considerably to the extent that a fine structure can be observed in the optical response.

The advantage of these quantum dots lies in the possibility of mass-production for applications such as solar cells. The ease with which one material system can be used to cover a large spectral bandwidth is already attracting a lot of interest, for instance in using passivated and water-soluble quantum dots as biological labels. By using different materials, a huge spectral range can be covered with this technology, as illustrated in Figure 7. Of course, there are still challenges to be met in this field. The quantum dots tend to degrade in the course of time, losing their high quantum efficiency for emission, presumably as a result of changes to the surface structure. Improvements have been made recently by capping the dots with a high-bandgap semiconductor, such as ZnS, which hinders interactions of the exciton with the surface states. The emission from organic-based dots also tends to show temporal fluctuations in wavelength, revealed experimentally from the emission of a single dot. This is thought to come about through the charging and discharging of traps in the vicinity of the quantum dot. Fundamentally, there are difficulties injecting carriers across an interface into the colloidal dots. However, progress can be expected in all these areas, and these quantum dots may well find applications in novel photonic devices.

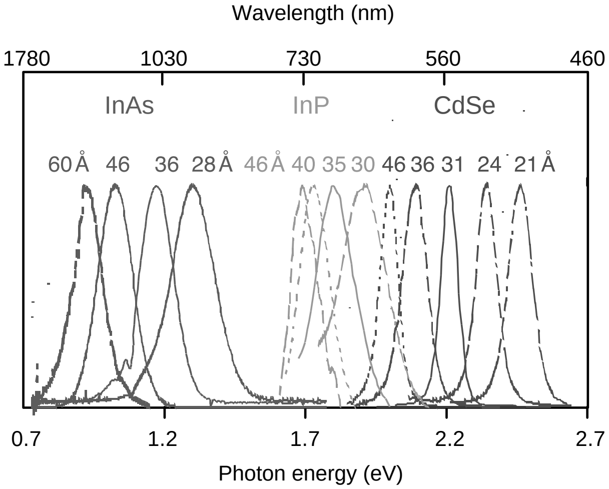

Figure 7 Room temperature emission spectra of surfactant-coated colloidal quantum dots. The quantum dots consist of InAs (left), InP (middle), and CdSe (right). Selection of the size has been used to give a particular wavelength in each band. Reprinted (abstracted/excerpted) with permission from Bruchez M Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013. Copyright 1998 American Association for the Advancement of Science.

Single-Dot Spectroscopy

No matter how quantum dots are prepared, there is always a distribution in emission energy. This constitutes an inhomogeneous broadening mechanism. For self-assembled InAs dots on GaAs, the lowest emission width from an ensemble of dots is about 20 meV. This energy is large compared to several other energies in the system. For example, the binding energy of a charged exciton, or of an exciton–exciton complex (a biexciton), is about 2 meV in this system. Also, the width of single dot emission is two to three orders of magnitude smaller than the ensemble width at low temperature. The width of single-dot emission is inversely related to the time over which the exciton maintains its coherence, and this should be as long as possible for future applications of quantum dots which manipulate the phase of the excitonic wavefunction. In order to investigate the properties of quantum dots in detail it is clearly necessary to measure the optical response of a single quantum dot rather than the response of a large ensemble of dots. The development of this field has been aided by the techniques developed for single-molecule spectroscopy which was first reported in the late 1980s and has subsequently become almost a standard experimental technique.

Typical densities of self-assembled quantum dots correspond to 100 per μm^2. Given that the emission wavelength is around 1 μm, it is clear that an optical technique with subwavelength spatial resolution is required to excite and detect a signal from just one quantum dot. The techniques employed include cathodoluminescence, where a focused electron beam provides the spatial resolution, and near-field optics where a subwavelength-sized aperture in a metal mask determines the spatial resolution. In the latter case, the aperture is either permanently connected to the sample surface, or it is formed on a tapered optical fiber positioned just a few nanometers away from the sample surface. An example is shown in Figure 8: with a confocal microscope the spatial resolution is not high enough to isolate spectroscopically a single dot. With increased spectral resolution, however, just a small number of peaks remain, each corresponding to the emission from a single dot. A further technique is to etch nanometer-sized mesas in the sample such that each mesa contains on average just one quantum dot. In some cases, single quantum dot emission has been detected with a conventional optical microscope (usually a confocal microscope) by adjusting the growth conditions to give particularly low dot densities. This can be achieved with colloidal nanocrystals by spinning a dilute suspension of the dots onto a substrate, a technique which has been frequently used in single-molecule spectroscopy. In each case, the photoluminescence signal from a single quantum dot is not large, and a detector capable of counting individual photons with a high quantum efficiency is required. A cooled silicon CCD camera is very convenient, giving not only a very low dark count and high quantum efficiency, but also a huge multiplexing advantage. However, the response of a Si detector tails off in the visible, and decreases rapidly at about 1050 nm in the near infrared, so it is by no means a universal solution.

The most striking feature of the emission from a single inorganic quantum dot is its spectral purity. Linewidths as small as a few μeV (a few hundred MHz) have been reported at low temperature for

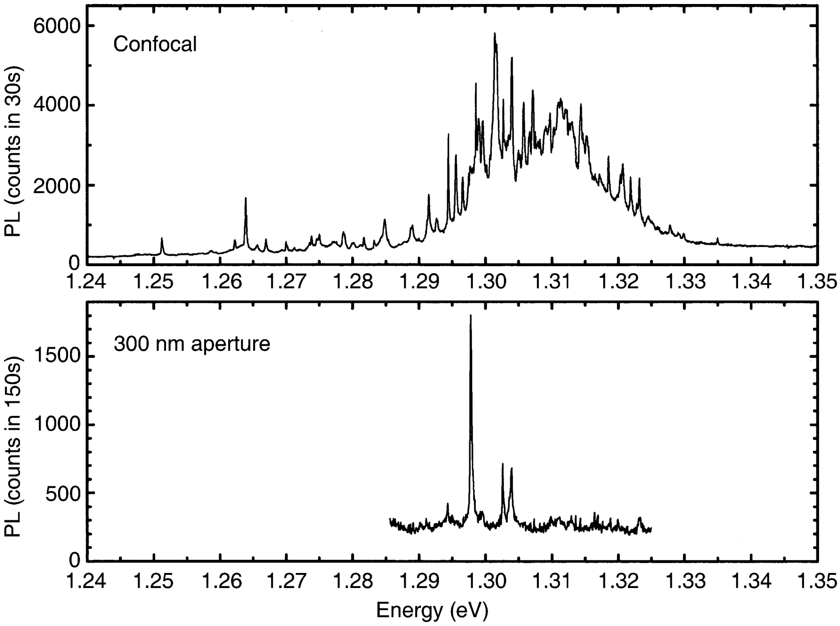

Figure 8 Emission from quantum rings from a 1 μm^2 area of the sample, and from a 0.07 μm^2 area of the sample, both taken at 4.2 K. The quantum rings are generated with a Stranski–Krastanow growth procedure using InAs and GaAs. The first spectrum was taken with a low-temperature confocal microscope; the second with a near-field technique. It can be seen how large statistical fluctuations arise in the first spectrum, because the number of rings probed is not sufficient to give a smooth distribution function. In the second spectrum, just a few peaks can be discerned above the noise level, each arising from the emission from an individual quantum ring.

quantum dots formed in a localizing potential in a quantum well. These linewidths can be interpreted as de-phasing times of the excitons. The maximum possible de-phasing time is obviously the radiative lifetime, in which case the line is homogeneously broadened. This limit has been observed at low temperature on natural quantum dots, but not at elevated temperature: as the temperature increases, scattering of the excitons by phonons decreases the coherence time, observed through a broadening of the emission.

The sharp lines observed in single-dot spectroscopy give access to fine structure on a meV energy scale. Most of this fine structure arises through Coulomb interactions between the electrons and holes. For example, two recent experiments have explored the emission as a function of electron occupation, and exciton population. In both cases, new lines emerge when a single dot is populated by more than one electron–hole pair.

Perspectives

The field of semiconductor quantum dots has made rapid advances in the past decade largely through the discovery of self-assembly in the growth of lattice-mismatched inorganic semiconductors. The most immediate application of these defect-free nanostructures is as the gain medium for semiconductor laser diodes, the aim being to reduce the threshold currents beyond those achievable with quantum wells. This approach is already proving highly successful. However, there are also other possibilities for novel photonic applications of quantum dots. These include memory elements for image processing, detectors and sources for single photons, and perhaps even elements for the coherent manipulation of wavefunctions.

See also

Semiconductor Materials: Band Structure Engineering; GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Excitons; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Bayer M, Stern O, Hawrylak P, Fafard S and Forchel A (2000) Hidden symmetries in the energy levels of excitonic artificial atoms. *Nature* 405: 923.

Bimberg D, Grundmann M and Ledentsov NN (1998) *Quantum Dot Heterostructures*. Chichester, UK: Wiley.

Bruchez M, Jr, Moronne M, Gin P, Weiss S and Alivisatos AP (1998) Semiconductor nanocrystals as fluorescent biological labels. *Science* 281: 2013.

Empedocles SA and Bawendi MG (1997) Quantum-confined Stark effect in single CdSe nanocrystalline quantum dots. *Science* 278: 2114.

Fricke M, Lorke A, Kotthaus JP, Medeiros-Ribeiro G and Petroff PM (1996) Shell structure and electron–electron interaction in self-assembled InAs quantum dots. *Europhysics Letters* 36: 197.

Gaponenko SV (1998) *Optical Properties of Semiconductor Nanocrystals*. Cambridge, UK: Cambridge University Press.

Leonard D, Pond K and Petroff PM (1994) Critical layer thickness for self-assembled InAs islands on GaAs. *Physical Review B* 50: 11687.

Lipsanen H, Sopanen M and Ahopelto J (1995) Luminescence from excited states in strain-induced $In_xGa_{1-x}As$ quantum dots. *Physical Review B* 51: 13868.

Tersoff J, Teichert C and Lagally MG (1996) Self-organization in growth of quantum dot superlattices. *Physical Review Letters* 76: 1675.

Warburton RJ, Dürr CS, Karrai K, *et al.* (1997) Charged excitons in self-assembled semiconductor quantum dots. *Physical Review Letters* 79: 5282.

Warburton RJ, Schäflein C, Haft D, *et al.* (2000) Optical emission from a charge-tunable quantum ring. *Nature* 405: 926.

Weisbuch C and Vinter B (1991) *Quantum Semiconductor Structures*. San Diego, CA: Academic Press.

Xie Q, Madhukar A, Chen P and Kobayashi NP (1995) Vertically self-organized InAs quantum box islands on GaAs(100). *Physical Review Letters* 75: 2542.

Type-II Quantum Wells and Superlattices

I Vurgaftman and J R Meyer, Naval Research Laboratory, Washington, DC, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

When thin layers of two semiconductor materials are combined to form a heterojunction, quantum well (QW), or superlattice, the properties of the resulting ‘quantum heterostructure’ are to a large extent governed by the alignment of the conduction and valence bands at the interfaces. Usually the overall conduction-band (CB) minimum lies in the same material as the overall valence-band (VB) maximum, as for example in the GaAs/$Al_{0.3}Ga_{0.7}As$ QW illustrated in Figure 1a. In such a type-I structure,

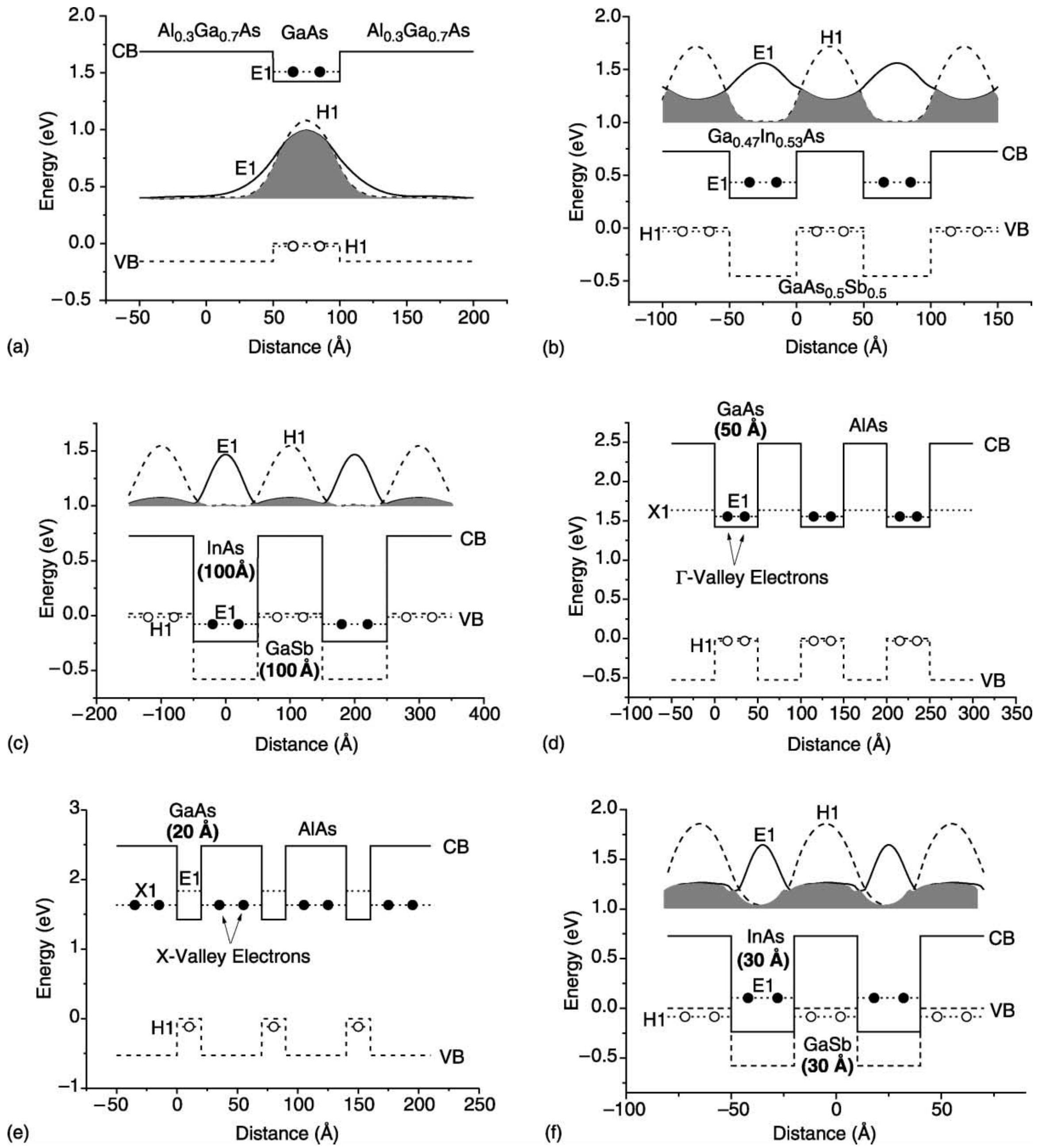

Figure 1 Conduction band (CB) and valence band (VB) profiles along with selected sub-band energies and wavefunctions for: (a) 50 Å $GaAs/Al_{0.3}Ga_{0.7}As$ type-I quantum well (GaAs substrate); (b) 50 Å/50 Å $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ type-II staggered-gap superlattice (InP substrate); (c) 100 Å/100 Å InAs/GaSb type-II broken-gap semimetallic superlattice (GaSb substrate); (d) 50 Å/50 Å GaAs/AlAs type-I superlattice (GaAs substrate); (e) 20 Å/50 Å GaAs/AlAs type-II superlattice with indirect band alignment both in real and momentum spaces; and (f) 30 Å/30 Å InAs/GaSb type-II semiconducting superlattice. E1, H1, and X1 represent the first sub-band in the Γ-valley conduction band, Γ-valley valence band, and X-valley conduction band, respectively. The overlap between the electron and hole wavefunctions is shown in gray.

the electrons and holes both reside mostly in the GaAs and their wavefunctions overlap almost perfectly as shown in the figure.

There are cases, however, when the CB minimum is in one layer while the VB maximum is in the other. For this so-called type-II alignment, which is exemplified by the $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattice of Figure 1b, the electrons and holes tend to separate spatially. It should not be surprising that this significantly influences many of the basic optical and electronic properties. Just one example is the band-edge absorption, whose strength is proportional to the square of the wavefunction overlap integral.

In contrast to the staggered type-II heterostructure of Figure 1b, another possibility is the broken-gap type-II alignment of the InAs/GaSb superlattice shown in Figure 1c. Since the CB minimum in InAs actually lies below the VB maximum in GaSb, this combination becomes a semimetal even though the bulk constituents are both semiconductors.

The nature of the band alignment can also be strongly influenced by shifts of the electron and hole energy levels induced by quantum confinement. Figure 1d illustrates a type-I GaAs/AlAs superlattice with relatively thick GaAs layers. Notice that the same superlattice becomes type-II if, as in Figure 1e, the GaAs layers are thin enough that the GaAs Γ-valley electron level (center of the Brillouin zone) is pushed to a higher energy than the AlAs X-valley level (boundary of the Brillouin zone). Similarly, quantum confinement can turn the broken-gap (semimetallic) type-II InAs/GaSb superlattice of Figure 1c into a staggered (semiconducting) type-II superlattice as in Figure 1f.

Although many different type-II heterostructures have been identified over the last three decades, only a few are likely to have long-term technological importance for optics or electronics. The system that has received the most attention in terms of both fundamental scientific studies and opto-electronic device development is InAs/GaSb and related alloy combinations such as InAs/GaInSb. Following a brief overview of wider-gap type-II heterostructures, this article will concentrate primarily on that interesting narrow-gap system.

Wide-Gap Type-II Heterostructures

The best-known wide-gap type-II heterostructure is the GaAs/AlAs superlattice of Figure 1e. The band lineup is type-II provided the GaAs thickness is no greater than ≈10 monolayers (28 Å) and the AlAs layer is no thinner than the GaAs layer. Since the CB minimum is in the AlAs X-valley and the VB maximum is in the GaAs Γ-valley, this structure has an indirect energy gap in both real space and momentum space. In type-II GaAs/AlAs, the X-valley states at the CB minimum form one or more superlattice minibands, which display relatively low resistance at small biases, negative differential velocity, etc. A related structure is the GaAs/AlAs/GaAs/AlAs/GaAs double-barrier intervalley resonant tunneling diode, in which resonant or sequential phonon-assisted tunneling proceeds via intervalley transitions whenever the Fermi level in the GaAs emitter lines up with the X-valley minimum in AlAs. Structures containing AlGaAs do not necessarily need quantum confinement to establish a type-II alignment, since it occurs naturally in $Al_xGa_{1-x}As$/AlAs heterojunctions with $x > 0.3$.

The device potential of wide-gap type-II heterostructures is rather limited, owing to their low luminescence efficiencies (due in large part to the doubly indirect band alignment) and low mobilities (due to the heavy X-valley effective mass). A few proposals have focused on modulation of the band alignment between the type-I and type-II regimes. The $Ga_{0.47}In_{0.53}As/GaAs_{0.5}Sb_{0.5}$ superlattices of Figure 1b and related structures have been investigated as alternative active regions for semiconductor lasers emitting in the important 1.3–2 μm wavelength range.

InAs/GaSb/AlSb-Based Type-II Superlattices and Quantum Wells

Molecular beam epitaxy (MBE) was used to fabricate the first InAs/GaSb-based superlattices in 1977, only a few years after the initial GaAs/AlGaAs superlattice growths. It was quickly demonstrated that for relatively thick layers the band alignment is broken-gap type-II, with 100–150 meV overlap between the conduction band minimum in InAs and the valence band maximum in GaSb. In this semimetallic superlattice, GaSb valence-band electrons spill over into the InAs and leave behind holes (see Figure 1c). It was also observed that at equal layer thicknesses of ≈90 Å quantum confinement induces a semimetal-to-semiconductor transition (Figure 1f), and that the gap becomes increasingly positive when the thickness is reduced further. This is no mere scientific curiosity, since being able to tune in a wide range of energy gaps spanning 0 to >0.5 eV has enormous practical consequences for infrared (IR) emitters and detectors. Such a capability cannot be replicated by any other III–V heterostructure system.

Another happy accident is that the lattice constants of InAs (6.058 Å) and GaSb (6.096 Å) are matched closely enough to allow them to be grown together by MBE or metalorganic chemical vapor deposition (MOCVD). A third binary member of the 'antimonide' or '6.1-Å' family of growth-compatible heterostructures is AlSb (6.136 Å), which is valuable in that it serves as a quantum barrier to both InAs conduction states and GaSb valence states. The extended family includes a variety of ternary alloys (GaInSb, GaAlSb, InAsSb, AlAsSb, etc.), whose energy gaps as a function of lattice constant are given by the curves in Figure 2, and also at least three quaternary alloys: GaInAsSb, AlGaAsSb, and InAlAsSb. Taken together, this menagerie of materials

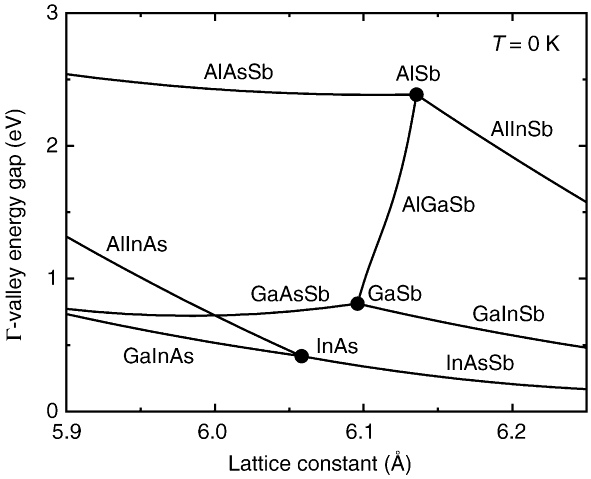

Figure 2 Direct energy gaps at zero temperature for III–V compound semiconductors and their ternary alloys with lattice constants in the vicinity of 6.1 Å.

provides the type-II IR device designer with an exceptional degree of configurational flexibility.

Type-II IR structures are typically grown on either a GaSb or InAs substrate. Since even a small amount of strain build-up can produce dislocation defects that seriously degrade the performance, one usually balances the net tensile strain (associated with layers having a lattice constant slightly larger than that of the substrate) with a compensating amount of net compressive strain (from layers whose lattice constant is slightly smaller than that of the substrate).

Band structures for complicated type-II layering configurations have been calculated by a variety of methods, including $k{\cdot}p$, tight-binding, and empirical pseudopotentials. While all reproduce the main features of the electron and hole sub-bands and dispersion relations near the center of the Brillouin zone, the relative merits of the different approaches remain controversial. Due to the small energy gaps, band-mixing effects are especially important in this system.

One consequence of the small type-II antimonide energy gap is that the electron and hole bands repel each other and both effective masses can be lighter than in any other III–V semiconductor (e.g., smaller than in the bulk constituents). This leads to high mobilities at room temperature, and promising possibilities for high-electron-mobility transistors (HEMTs). Also of interest are resonant tunneling diodes (RTDs) based on InAs/AlSb/GaSb/AlSb/InAs and related structures, in which interband tunneling via the GaSb valence states results in negative differential resistance with a very high peak-to-valley ratio. Since these devices are primarily electronic rather than optical, we will not discuss them further here.

While interfaces play an important role in any semiconductor heterostructure, they are especially crucial in InAs/GaSb and related type-II systems because the two materials on opposite sides of the interface have neither a common cation nor a common anion (unlike GaAs/AlGaAs, for example, in which both wells and barriers have As anions). By timing the shutter sequences in the growth, one can in principle end the InAs on an In layer and begin the GaSb on Sb, to produce InSb-like interface bonds. Alternatively, one can end the InAs on As and begin the GaSb on Ga to form GaAs-like bonds. Detailed characterizations by a variety of methods have shown that the device properties and even the band structure can be relatively sensitive to whether the interface bonds are InSb-like or GaAs-like.

With regard to opto-electronic devices, a potential limitation of InAs/GaSb superlattices is that in order to reach a long wavelength (small energy gap), the layers must be rather thick (e.g., ≈90 Å for zero gap). In such a structure the overlap between the electron and hole wavefunctions becomes quite small (see Figures 1c and 1f), and consequently the interband optical interactions are weak. However, Mailhiot and Smith pointed out that this issue can be circumvented to a large extent by employing $Ga_{1-x}In_xSb$ (with $x \leq 0.4$) rather than GaSb for the hole QWs. Compressive strain in the GaInSb produces a large splitting of the heavy and light hole sub-bands, along with a substantial increase of the type-II offset energy. The result is that much thinner layers (with large wavefunction overlaps) may be employed in conjunction with energy gaps corresponding to long IR wavelengths. In fact, the overlap for a short-period InAs/GaInSb superlattice such as that illustrated schematically in Figure 3 can be large enough to yield interband optical interaction strengths comparable to those in bulk materials and type-I superlattices.

While practical constraints on the minimum reproducible layer thickness set a lower wavelength limit on the order of 2 μm, the upper limit governed by uniformity considerations is quite high, perhaps in the $\lambda = 25-100$ μm range. This makes type-II antimonide quantum heterostructures a natural choice for mid-IR to far-IR photodetectors and lasers.

Type-II IR Photodetectors

Because IR detectors sense the radiant black-body emission from warm objects rather than relying on reflected visible light, they can be used for 'night vision.' Both photoconductive (PC) and

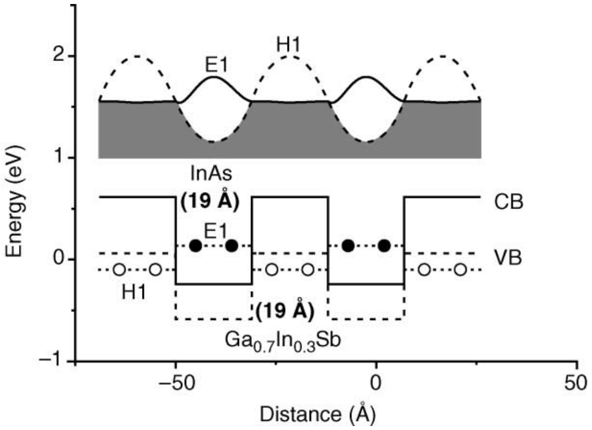

Figure 3 Conduction and valence band profiles along with electron (E1) and hole (H1) sub-band energies and wavefunctions for a 19 Å/19 Å $InAs/Ga_{0.7}In_{0.3}Sb$ type-II superlattice (GaSb substrate) with an energy gap equal to that of the InAs/GaSb superlattice in Figure 1d.

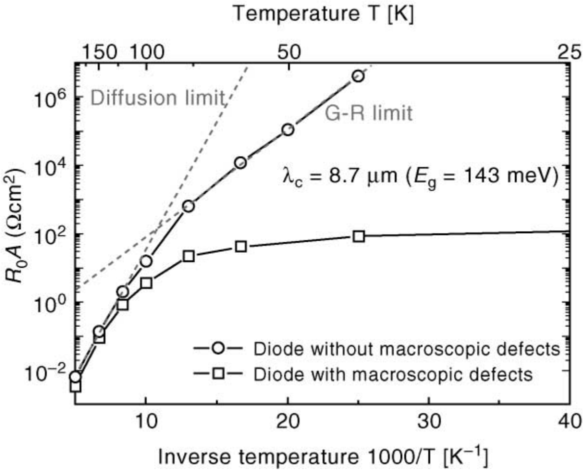

Figure 4 Temperature dependence of the dynamic impedance at zero bias R_0A for two different InAs/GaInSb superlattice diodes with cutoff wavelengths of 8.7 μm. The generation-recombination (G-R) and diffusion-limited R_0A values are indicated by the dashed lines. Reproduced with permission from Bürkle *et al.* (2000) *MRS Symp. Proc.*, 607, pp. 77–82.

photovoltaic (PV) IR detectors are now being developed with type-II active regions, usually consisting of simple InAs/GaInSb superlattices. The PV geometry, in which a reverse bias is applied to a p–n junction, is ultimately preferred since it is easily adaptable to pixelation in a 2D array with many discrete detector elements. That allows an IR camera to construct a detailed thermal picture of any given scene.

An optimized PV photodetector must have high absorption in the active region, unimpeded vertical transport to assure collection of the photogenerated carriers, and a low dark current. The thin layers of a type-II InAs/GaInSb active region help to satisfy the first two criteria, since they assure both a large wavefunction overlap (Figure 3) and rapid electron and hole tunneling through the superlattice minibands.

The dark current in a high-quality diode tends to be dominated by generation-recombination currents at low temperatures and diffusion currents at higher *T*. However, trap-assisted tunneling via dislocations and other macroscopic defects can significantly increase the dark current. Figure 4 plots the resistance–area product (R_0A) for diodes with and without macroscopic defects. Surface leakage resulting from inadequate sidewall passivation can also lead to a substantial contribution that is proportional to the perimeter of the mesa. The diffusion current scales as $1/\tau^{1/2}$, where τ is the recombination lifetime. To date, the defect-mediated Shockley–Read lifetimes of <100 ns have been significantly shorter than in competing HgCdTe IR materials (>1 μs). However, Auger lifetimes are substantially enhanced, as will be discussed further in the context of type-II IR lasers. In an Auger event, the energy and momentum of the recombining electron–hole pair is transferred to a third carrier, which scatters to a higher-lying state in either the same or a different band. Because three different carriers must interact simultaneously, the recombination rate scales with the cube of the carrier density. Usually Shockley–Read recombination dominates at lower temperatures and Auger processes at higher *T*.

Typically the n-side of a type-II diode is heavily doped (10^{18} cm^{-3} range) and relatively thin, while most of the absorption occurs on the lightly doped (10^{16} cm^{-3} range) p-side. A standard figure of merit is the detectivity (D^*), which is essentially the signal-to-noise ratio in a device of standard size. For 77 K operation, type-II PV detectors with cutoff wavelengths of 7.5 μm and 12 μm achieved D^* of 1×10^{12} cm Hz$^{1/2}$/W and 5×10^{10} cm Hz$^{1/2}$/W, respectively. Those values are nearly as high as typical results for the much more mature HgCdTe FPA technology.

It is anticipated that type-II detectors will be particularly attractive at very long wave IR (VLWIR), at which it becomes increasingly difficult to maintain adequate control over the HgCdTe compositional uniformity. Recent PC detectors operated out to 22 μm, and PV devices to 16 μm. Due in part to the strong suppression of Auger recombination, type-II detectors are also expected to be advantageous at noncryogenic temperatures where future systems will increasingly operate.

Type-II IR Lasers

At visible and near-IR wavelengths out to ≈2 μm, the QW diode laser is by now a mature technology that dominates the commercial marketplace due to its low cost (<$1 per laser in some cases), compactness, and reliability. However, the technical challenges become much more severe in the mid-IR spectral region (3–5 μm). Before the advent of the type-II laser, no interband III–V semiconductor device emitting at $\lambda \geq 3$ μm had ever operated at room temperature. Alternatives are the lead salt laser, which is limited to very low-output powers, and the quantum cascade laser (QCL) that is based on optical transitions between two electron QW sub-bands rather than interband electron–hole processes. Mid-IR lasers are needed for high-power military applications such as IR countermeasures against heat-seeking missiles and for the sensitive detection of trace chemicals using the spectroscopic 'fingerprint' that distinguishes each species from any other.

From the 1960s to the late 1990s, conventional III–V double heterostructure (DH) and type-I QW lasers for the mid-IR never achieved the desired levels of performance because they failed to overcome certain fundamental limitations. Those included:

1. inadequate band offsets, which at higher temperatures made it difficult for the QW or DH active regions to retain the injected electrons and holes;
2. rapid Auger nonradiative decay, which led to excessive threshold current densities at higher temperatures; and
3. restricted spectral range, limited to wavelengths shorter than the cutoff for bulk $InAs_{1-x}Sb_x$.

However, it has more recently become apparent that all of these limitations can be either minimized or removed by employing a type-II InAs/GaSb-based active region.

For example, electrical confinement is far more effective because the GaSb conduction band confines InAs electrons and the InAs valence band confines GaSb holes (see Figure 1f). Auger recombination is also significantly suppressed, in part because the holes are nearly as light in the plane as the electrons, making it much more difficult to conserve both momentum and energy in the three-carrier process. Nearly an order-of-magnitude suppression of the room-temperature Auger rate has been reported, which translates directly into a similar reduction of the threshold current density. And as was discussed above in connection with IR photodetectors, very long wavelengths are readily attainable because the energy gap in a type-II active region can be arbitrarily small. Type-II lasers have already operated cw at $\lambda > 7$ μm, and cryogenic operation out to wavelengths as long as 100 μm has been projected theoretically.

In 1986 the first type-II semiconductor laser, which emitted at $\lambda = 1.86$ μm from single liquid-phase epitaxy (LPE)-grown heterojunctions, was demonstrated. Since the mid-1990s, 3–4 μm GaInAsSb/InAs-based single-interface type-II diodes have been developed that operate up to 205 K in pulsed mode.

In 1995, lasing at $\lambda = 3.2$–3.5 μm was reported from type-II $InAs/Ga_{0.75}In_{0.25}Sb$ superlattice diodes grown by MBE. Subsequent work on these devices extended the temperature range to 255 K in pulsed mode and 180 K for cw operation, and the emission wavelength to 4.3 μm.

While the wavefunction overlap and hence the gain for a two-constituent type-II superlattice active region can be quite favorable, penetration of the electron wavefunctions into the thin GaInSb layers is in fact so great that a wide miniband forms along the growth axis. The electrons then have strong energy dispersion in all three dimensions, whereas present-generation near-IR and visible semiconductor lasers nearly always have QW active regions with 2D carriers. The stepped quasi-2D density of states yields a much higher differential gain per injected carrier at threshold, and also narrower spectral emission lines than are attainable with a 3D electron population. The most straightforward approach to blocking the formation of a miniband in the superlattice is to add a high electron barrier between the periods, e.g., InAs/GaInSb/AlSb. However, the drawback of that approach is that the wavefunction overlap becomes much smaller than in the simple superlattice, resulting in significantly lower gain.

On the other hand, if a second electron well is added on the other side of the hole well, the resulting structure with four constituents per period (e.g., InAs/GaInSb/InAs/AlSb) combines the large wavefunction overlap of the two-constituent superlattice with 2D electron and hole confinement. Figure 5 illustrates band profiles, energy levels, and wavefunctions for the 'W' laser, which takes its name from the shape of the conduction-band profile.

Optically pumped type-II W lasers display high-power pulsed operation at temperatures well above ambient. Wavelengths have spanned the range 2.6–7.3 μm, and mid-IR vertical-cavity surface-emitting lasers (VCSELs) were also fabricated. Optical pumping is viewed primarily as a means of achieving high cw or quasi-cw output powers in the mid-IR, and in that regard W lasers have exceeded all other semiconductor approaches with 1.5 W peak output power being obtained at 71 K for 100 μs

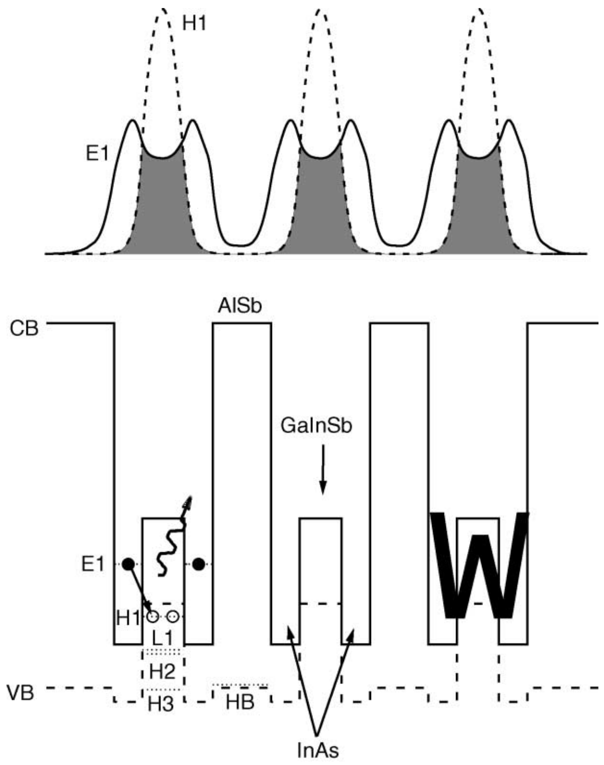

Figure 5 Conduction and valence band profiles along with selected sub-band energies and wavefunctions for InAs/$Ga_{0.7}In_{0.3}Sb$/InAs/AlSb type-II 'W' structure (GaSb substrate). Electron (E1), heavy hole (H1, H2, H3), light hole (L1), and barrier heavy hole (HB) sub-bands are indicated.

pulses with a 10% duty cycle. The average output at 82 K for that device was up to 360 mW for 20 μs pulses and a 50% duty cycle.

Cw operation at $\lambda = 3.0$ μm was observed nearly to room temperature (290 K), and even a 6.1 μm emitter operated cw to 210 K. The cw output power of 540 mW/facet from a device with 80 W QW periods was the highest ever reported for 78 K operation in the mid-IR. Despite the rather high output power, however, the differential power conversion efficiency was only 2.5% at 78 K and fell to 1% at 140 K. Much higher efficiencies were obtained using the optical pumping injection cavity (OPIC) approach, in which the active region was surrounded by semiconductor quarter-wave mirror stacks that formed an etalon cavity resonant at the 2.1 μm pump wavelength. For pulsed operation, OPIC devices with uncoated facets and only 10 active QWs yielded lower thresholds, lower internal losses, and higher power conversion efficiencies than any previous results at $T \geq 200$ K. Figure 6 illustrates that the efficiencies were 7.1% at 220 K for one device and 4% at 275 K for another.

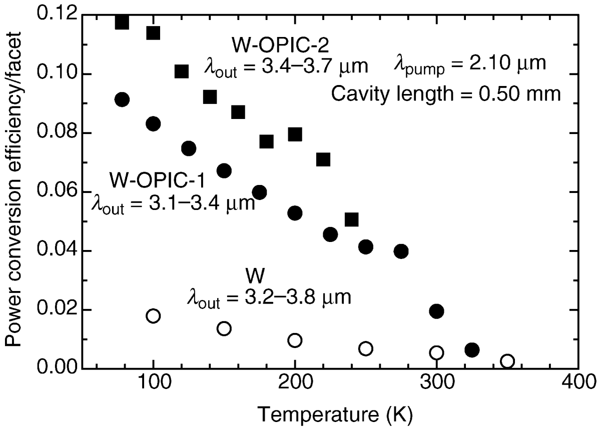

Figure 6 Power conversion efficiency per uncoated facet as a function of temperature for two different W-OPIC lasers emitting in the wavelength ranges of 3.1–3.4 μm (filled circles) and 3.4–3.7 μm (filled squares), respectively. For comparison, conversion efficiencies for a typical 'W' laser without an OPIC emitting in the same spectral range (open circles) are also shown. All of the lasers were optically pumped in pulsed mode by a 2.1 μm Ho:YAG laser, and the cavity length for all three devices was 0.5 mm.

An alternative approach to maximizing the pump absorbance despite a small number of QWs is the 'integrated absorber' (IA) concept. In this approach, the W QWs are separated by intermediate-gap GaInAsSb absorbing layers which donate optically generated electrons and holes to the active wells. At 80 K with 35 μs pulses at 2.5% duty cycle, an IA laser with only five QWs separated by ≈1600-Å-thick absorber layers had a power conversion efficiency of 9.8% and peak output power of 2.1 W. Thus far the IA has not done as well as the OPIC at high temperatures, due mostly to an inadequate valence band offset between the absorber layers and the active QWs.

The beam quality for optically pumped type-II lasers was substantially improved by using the angled-grating distributed-feedback (α-DFB) approach pioneered at shorter wavelengths. Since both the grating and the facets are needed to produce optical feedback, only a narrow angular cone is fed back into the tilted cavity. For a pump-stripe width of 50 μm, the near-diffraction-limited output beam had 15 times narrower angular divergence (1.4° full width at half-maximum (FWHM)) than the double-lobed beam emitted from a conventional Fabry–Perot laser occupying an unpatterned portion of the same bar. Even more promising results have been reported using the photonic-crystal distributed-feedback (PCDFB) laser, in which the grating is defined on a two-dimensional rectangular lattice that is tilted with respect to the facets as in the α-DFB. For a pump-stripe width of 200 μm, the beam quality for the direction along the laser stripe was four times the diffraction limit.

Although optical pumping is the simplest way to inject electrons and holes into the active region, type-II diode lasers will ultimately be much more inexpensive, convenient, compact, and energy efficient. However, generalization of the W configuration to electrical pumping is not entirely straightforward. This is because holes injected from the p-type cladding layer may have poor growth-axis transport, and hence hole populations in the various GaInSb QWs could be highly nonuniform. Whereas the holes in a superlattice are required to tunnel through only one InAs electron well at a time (see Figure 3), the barrier for the structure in Figure 5 consists of two InAs layers in addition to the AlSb barrier. To overcome this problem the AlSb was replaced with a specially designed quaternary $Al_{0.15}Ga_{0.85}As_{0.05}Sb_{0.95}$ barrier, which raised the L1 light-hole sub-band to within ≈100 meV of the H1 heavy-hole valence band maximum. That permitted thermally excited light holes to provide the needed tunneling transport to all of the QWs. Another key design element was the insertion of relatively thick (0.6 μm) undoped $Al_{0.35}Ga_{0.65}As_{0.03}Sb_{0.97}$ 'broadened-waveguide' layers on each side of the active region, in order to minimize free-carrier absorption in the doped $Al_{0.9}Ga_{0.1}As_{0.07}Sb_{0.93}$ cladding layers. Some of the devices also contained InAs/AlSb 'hole-blocking' layers, which prevented hole leakage into the n-type broadened waveguide layer.

A pulsed $\lambda = 3.27$ μm W-diode of this type with 10 active QW periods achieved the first room temperature operation (to 310 K) of an electrically pumped III–V interband laser in the mid-IR. Another device with five QWs operated in cw mode to 200 K, the highest $T_{\max}$ to date for any interband III–V laser beyond 3 μm. The threshold at 78 K was 63 A/cm^2, and Figure 7 shows that the cw output was 140 mW at 78 K and 20 mW at 180 K. Following optimization of the turn-on voltages, series resistances, and internal losses, cw diodes with one–three QWs are projected to operate in the thermo-electric cooler range and ultimately at room temperature.

A more exotic variation on the type-II diode emitter is the interband cascade laser (ICL) that was first proposed by RQ Yang. Like the intersub-band QCL, the ICL generates multiple photons per injected electron by making an optical transition at each step of a staircase-like QW structure. While the threshold current density for the QCL is high because of the inherently short lifetimes associated with intersub-band phonon scattering, the interband configuration of the ICL eliminates that nonradiative relaxation path. Early projections that the ICL should display low thresholds, high cw operating temperatures, high output powers, and high differential quantum efficiencies (well above the conventional limit of one photon/electron) have all been confirmed experimentally.

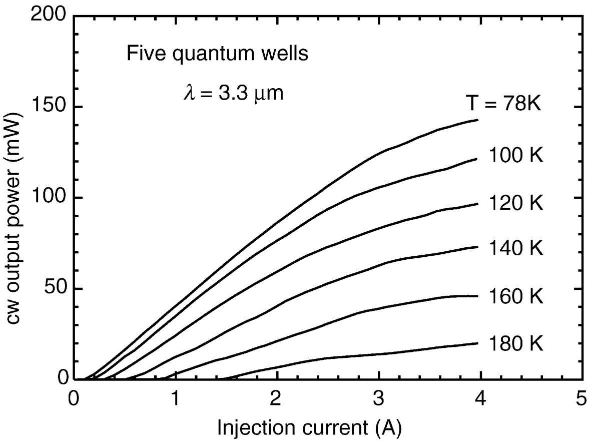

Figure 7 Cw output power as a function of injection current at temperatures between 78 K and 180 K for a 'W' diode laser with a 2-mm-long cavity and 100-μm-wide stripe emitting at a wavelength of 3.3 μm.

Figure 8 illustrates conduction and valence band profiles for a typical single stage of the ICL staircase. While this example employs a W configuration, in which two InAs electron QWs (19 Å and 13 Å) surround the active GaInSb hole QW (32 Å), some designs employ only a single InAs active well. Electrons tunneling into the InAs QWs from the preceding injection region at left emit a photon by making spatially indirect radiative transitions to the GaInSb valence band. They next tunnel into the valence states of the adjacent GaSb and GaInSb hole wells, whose function is to provide a thick barrier to prevent electron leakage from the active InAs QWs directly to the 70 Å InAs well that begins the next injection region. From the final GaInSb hole QW (40 Å), the carriers undergo near-elastic interband scattering into conduction states of the 70 Å InAs QW. Electrons traversing the superlattice miniband of the injection region, which in this case consists of eight digitally graded InAs wells, finally tunnel into the active electron QWs of the next period. This series of events is repeated at each step of the ICL staircase.

Following the first ICL demonstration in 1997, there have been reports of threshold current densities 13 A/cm^2 at 80 K, maximum operating temperatures of up to 300 K for pulsed operation and 150 K for cw (with epitaxial-side-up mounting), pulsed output powers of 6 W/facet and slope quantum efficiencies exceeding 600%. A cw wall plug efficiency of 22%

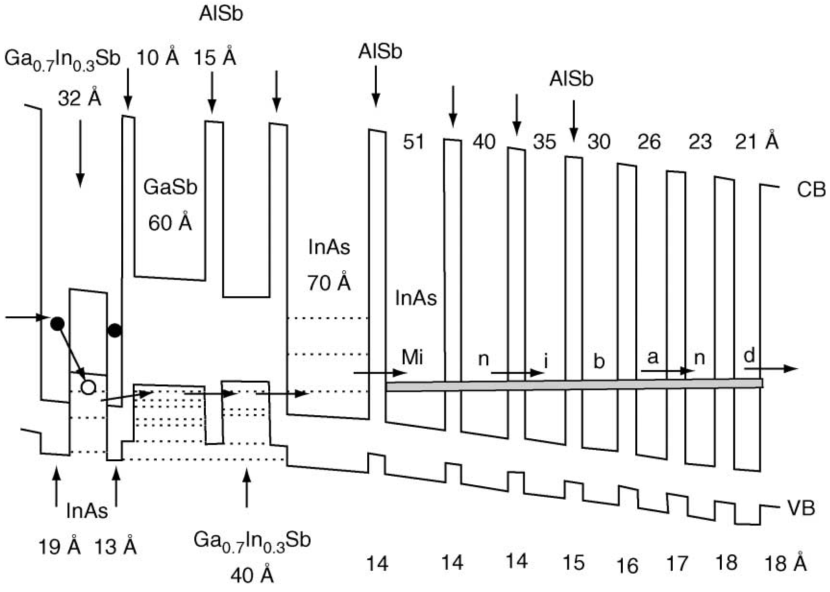

Figure 8 Conduction and valence band profiles for one period of an interband cascade laser with a W active region and three hole wells. Sub-band energy levels are also indicated.

at 80 K was recently reported. While substantial further improvements are expected, the primary disadvantage of the ICL is its complexity.

Despite numerous dramatic advances by type-II mid-IR lasers since their first demonstration in 1995, the performance has not yet progressed to the level that will be required to support a robust commercial technology. One key issue is the cw operating temperature, which must increase to at least the thermo-electric cooler range if the potential market is to extend beyond niche applications. Higher-temperature operation will result automatically once the number of QWs, and hence the threshold current density, can be reduced. Further suppression of the Auger decay rate would also be highly beneficial if it is proven possible.

The laser output power and efficiency are governed by the internal loss, which in type-II QWs tends to be dominated by resonant free-hole intersub-band transitions. The details of the valence sub-band structure are not well known at present, although recent experiments have apparently identified a number of the transition energies. As in the case of the threshold, reducing the number of QWs is highly beneficial since it reduces the number of free carriers that can absorb the lasing photons. The optically pumped OPIC and IA structures discussed above achieved dramatic reductions in the internal losses by following that strategy.

The beam quality attainable from a semiconductor laser is governed by the linewidth enhancement factor (LEF), which is proportional to the variation in the refractive index induced by a small change in the carrier density. A large positive value of the LEF leads to filamentation in broad-area lasers as tiny peaks in the optical intensity cause carrier depletion and a local increase in the refractive index, which in turn stimulates further focusing of the output beam. The output then consists of an incoherent superposition of many rapidly diverging filaments, resulting in a beam quality that is several orders of magnitude worse than the diffraction limit. While theoretical LEFs for optimized type-II W structures can be in the 1.5–2 range even at room temperature, recent Hakki–Paoli measurements have yielded values in the 2.5–5 range. This is possibly due to additional inhomogeneous broadening of the gain spectrum induced by interface roughness.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Mailhiot C and Smith DL (1989) Long-wavelength infrared detectors based on strained InAs-GaInSb type-II superlattices. *Journal of Vacuum Science and Technology* A7: 445–449.

Meyer JR, Malin JI, Vurgaftman I, Hoffman CA and Ram-Mohan LR (1997) Antimonide-based quantum heterostructure devices. In: Manasreh MO (ed.) *Strained*

Quantum Wells and Their Applications, pp. 235–272. Amsterdam, The Netherlands: Gordon and Breach.

Olafsen LJ, Vurgaftman I and Meyer JR (in press) Antimonide mid-IR lasers. In: Choi HK (ed.) *Long-Wavelength Semiconductor Lasers*. New York: Wiley.

Ram-Mohan LR, Vurgaftman I, Meyer JR and Dossa D (2000) Wavefunction engineering: a new paradigm in quantum nanostructure modeling. In: Nalwa HS (ed.) *Handbook of Nanostructured Materials and Nanotechnology*, vol. 2, pp. 707–739. New York: Academic Press.

Razeghi M (1998) Current status and future trends of infrared detectors. *Optoelectronics Review* 6: 155–194.

Vurgaftman I, Felix CL, Aifer EH and Meyer JR (2000) Interband III–V mid-IR semiconductor lasers. In: Perera AGU and Liu HU (eds) *Handbook of Thin Film Devices*, vol. 2, pp. 171–193. San Diego, CA: Academic Press.

Vurgaftman I, Felix CL, Bewley WW, *et al.* (2001) Mid-IR “W” lasers. *Philosophical Transactions of the Royal Society A* 359: 489–504.

Vurgaftman I, Meyer JR and Ram-Mohan LR (2001) Band structure parameters for III–V compound semiconductors and their alloys. *Applied Physics Reviews, Journal of Applied Physics* 89: 5815–5875.

Wilson BA (1988) Carrier dynamics and recombination mechanisms in staggered-alignment heterostructures. *IEEE Journal of Quantum Electronics* 24: 1763–1777.

Yang RQ, Bradshaw JL, Bruno JD, Pham JT and Wortman DE (2002) Mid-IR type-II interband cascade lasers. *IEEE Journal of Quantum Electronics* (Feature Section on Mid-IR Quantum Cascade Lasers) 38: 559–568.

SEMICONDUCTOR PHYSICS

Contents

Outline of Basic Electronic Properties
Band Structure and Optical Properties
Excitons
Impurities and Defects
Infrared Lattice Properties
Light Scattering
Polarons
Quantum Wells and GaAs-Based Structures
Recombination Processes
Spin Transport and Relaxation in Semiconductors; Spintronics
Surface Photovoltage Spectroscopy of Semiconductors

Outline of Basic Electronic Properties

C R Pidgeon, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

We briefly introduce the main concepts underlying the detailed treatment of specialist aspects of semiconductor physics in ensuing sections, viz: the free electron model, density of states, Fermi-Dirac statistics, energy bands and energy gaps, effective mass, electrons and holes, intrinsic and extrinsic conductivity, direct and indirect interband optical transitions, excitons and low-dimensional systems.

Introduction

A proper treatment of optoelectronic devices needs to include the physical properties of the materials from which the devices are made, and this subject is covered under the section of the Encyclopedia entitled ‘Semiconductor Fundamentals’. This includes sections on the basic physics common to all semiconductors useful in optoelectronic applications, the experimental techniques used to investigate the optical properties of these semiconductors and the more detailed properties of the most widely used semiconductors, and semiconductor systems. The properties of greatest importance here are those of electrons in crystalline solids, but the interaction between the electrons and the atoms of the crystal lattice is also significant. The combination of these microscopic effects is responsible for determining the macroscopic optical and electrical properties of semiconductors, and in this section we give the necessary background to these topics.

The classical treatment is extremely useful in describing the electrical conductivity of solids; the electrons are regarded as charged particles obeying Newton's laws of motion, and electrical resistance is assumed to arise from collisions with the vibrating lattice and/or impurities. However, the classical theory does not give a method of predicting the natural frequencies, or energies, of these electronic oscillators. To do this we need to use quantum mechanics, and the elements of this are reviewed here. We start with the free electron theory of metals, where the electrons are treated as particles constrained to move in a potential box. For convenience we begin by considering the situation in one dimension (1D).

Free Electron Theory of Metals

Free electrons in a 1D potential box of length L in the x-direction are described by a plane wave function that is normalized within the box, $\psi = (1/L)e^{ikx}$, and satisfies the Schrödinger equation in the x-direction:

$$\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + E\psi = 0 \qquad [1]$$

where E is the electron energy. This gives the familiar parabolic $E-k$ dispersion relation (see Figure 1), $E = \hbar^2k^2/2m$ (with electron momentum $p = \hbar k$). In order to evaluate physical problems it is necessary to define the boundary conditions for the wave function. We assume fixed boundary conditions (i.e., $\psi = 0$ at $x = 0, L$) leading to the quantization condition $k = n\pi/L$ where n is a positive integer.

Density of States

Physical properties of solids depend on the number of allowed k-values per unit range of k (or E-values per unit range of E), a quantity referred to as the density of states, D(k) or D(E) – for example, the number of allowed states between k and $k + dk = D(k)dk$. From the fixed boundary condition ($k = n\pi/L$, the number of allowed values of k per unit range of k in 1D is just $D(k) = L/\pi$. In terms of energy, using the $E-k$ dispersion relation, we can write (including the factor 2 for two possible spin states):

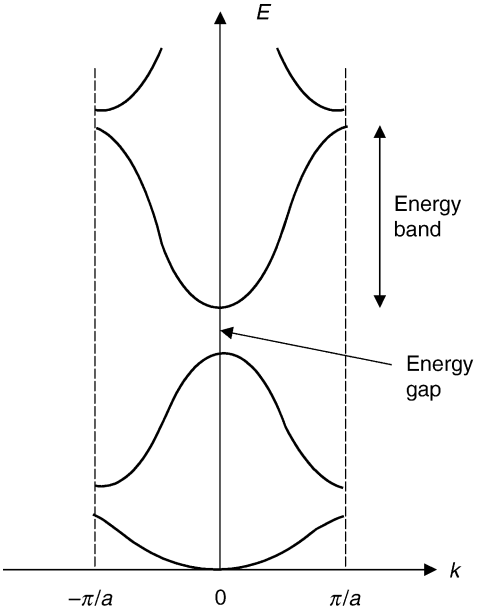

Figure 1 Schematic representation of energy bands in one dimension, characterized by gaps at the high symmetry points of the Brillouin zone.

$$\begin{aligned} D(E)dE &= D(k)(dk/dE)dk \\ &= (L/2\pi)(2m/\hbar^2)^{1/2}E^{-1/2}dE \qquad [2] \end{aligned}$$

i.e., the density of states depends on $E^{-1/2}$ for the 1D situation.

For a cube of side L, in three dimensions we generalize the identity for fixed boundary conditions:

$$\frac{L^2k^2}{\pi^2} = n_x^2 + n_y^2 + n_z^2 = R^2 \qquad [3]$$

where $k_x = n_x\pi/L$ etc., and the states fill up the positive quadrant of a sphere of radius ($R = Lk/\pi$) in k-space. Thus, the number of states in a given volume of k-space is numerically equal to that volume, so that the number of states between R and $R + dR$ in the positive quadrant (hence 1/8th) of a spherical shell of radius R in k-space is

$$D(R)dR = 4\pi R^2 dR(1/8) \qquad [4]$$

i.e., (1/8)th of the surface area of the shell multiplied by the thickness of the element. Changing the variable to k, the number of states between k and $k + dk$ is

$$D(k)dk = D(R)\frac{dR}{dk}dk = \frac{V}{2\pi^2}2k^2dk \qquad [5]$$

where the volume $V = L^3$, and a factor of 2 has again been included to account for the two possible spin states of the electron. This in turn, through the relation between E and k, can be expressed in terms of the density of states per unit energy so that the number of states between E and $E + \mathrm{d}E$ is given by:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{2V}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}}\mathrm{d}E \qquad [6]$$

In two dimensions, by the same argument:

$$\frac{L^2k^2}{(\pi)^2} = n_x^2 + n_y^2 = R^2 \qquad [7]$$

Therefore $D(R)\mathrm{d}R = (2\pi R/4)\,\mathrm{d}R$, where the factor 1/4 implies the positive quarter of the 2D ring between R and $R + \mathrm{d}R$ in k-space, $A = L^2$ and

$$D(k)\mathrm{d}k = D(R)\frac{\mathrm{d}R}{\mathrm{d}k}\mathrm{d}k = \frac{A}{\pi}k\mathrm{d}k \qquad [8]$$

From the E–k relation we have the number of states between E and $E + \mathrm{d}E$ as:

$$D(E)\mathrm{d}E = D(k)\frac{\mathrm{d}E}{\mathrm{d}k}\mathrm{d}E = \frac{A}{\pi}\left(\frac{m}{\hbar^2}\right)\mathrm{d}E \qquad [9]$$

Thus, in 2D systems we have the important result that $D(E)$ is a constant.

Fermi-Dirac Distribution of Electrons

At zero temperature an atom exists in what is known as the ground state. All the electrons rest in the lowest energy states allowed to them by the Pauli principle, and fill the allowed states up to an energy known as the Fermi energy, E_F. Above absolute zero of temperature this is no longer the case, and thermal energy allows some electrons to exist in higher energy states. The probability $f(E)$ that a state of energy E will be occupied for an ideal electron gas in thermal equilibrium (thermal energy kT) is governed by Fermi-Dirac statistics:

$$f(E) = \frac{1}{1 + \exp\left(\frac{E - E_\mathrm{F}}{kT}\right)} \qquad [10]$$

For a metal at room temperature the difference between the Fermi energy and the energy of the conduction band minimum is much greater than the thermal energy, $(E_\mathrm{F} - E_\mathrm{co}) \gg kT$; this is also true for a strongly n-type semiconductor (see below) and the distribution is said to be degenerate.

Energy Bands and Energy Gaps: Semiconductors

In the presence of the (perturbing) periodic crystal potential, $V(r)$, the above simplicity is lost: this is the so-called ‘nearly free electron model’, and the perturbed Schrödinger equation is now:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V(x)\psi = E\psi \qquad [11]$$

for the x-direction. This topic is discussed in all undergraduate texts on solid state physics, but we list the main results of the model as follows:

1. The electrons that we are concerned with are confined within a periodic lattice. Consequently the wave function ψ is modulated with the periodicity of the lattice. By symmetry arguments the correct wave function can be shown to be:

   $$\psi_k(x) = \frac{1}{\sqrt{L}}\mathrm{e}^{ikx}u_k(x) \qquad [12]$$

   where $u_k(x)$ is a function that has the symmetry of the lattice. $\psi_k(x)$ is known as the Bloch function and is a satisfactory solution to the Schrödinger equation. Because of the periodic nature of the Bloch equation, it is always possible to select the wavevector index k of any Bloch function to lie within the range $\pi/\mathrm{a} < k < \pi/a$. This region is referred to as the first Brillouin zone, and the representation where the E–k dispersion is mapped back entirely into the first Brillouin zone is called the ‘reduced zone scheme’ (Figure 1).
2. Regions of E appear where solutions $\psi_k(x)$ are not allowed, corresponding to Bragg reflections of the electron waves by the lattice; these are the so-called energy gaps (forbidden energy regions) which separate energy bands (allowed energy regions). In 3D these discrete bands of energy form the conduction and valence bands within materials. The forbidden energy gap separates these bands. An additional quantum number n is required to label the individual energy bands corresponding to successive Brillouin zones in the wavevector, $2n\pi/a$ (where $n = 0, 1\ldots$). The position of the Fermi level determines the electron population of the energy bands, and the character of the solid (see below).
3. The parabolic E–k dispersion relation is retained near $k = 0$, and at other symmetry points of the Brillouin zone, but with the substitution of an effective mass, m^* (see below) for the free electron mass, i.e., $E = \hbar^2k^2/2m^*$ for $k \ll \pi/a$. This is referred to as the effective mass approximation.

The nearly free electron model is one limit for the description of the formation of energy bands and energy gaps in a solid. In real crystals the energy bands derive from, and retain the symmetry of, the parent atomic states of the host atoms. In this limit, the model that starts out from the wave functions of the free atoms is known as the tight binding model or the LCAO (linear combination of atomic orbitals) approximation. As neutral separated atoms are brought together to form a crystal, their wave functions overlap and the Coulomb interaction between the atom cores and the electrons splits the atomic energy levels, spreading them into bands. Each state of given quantum number of the free atom is spread in the crystal into a band of energies; a given s, p, d-state in N separate atoms will produce the corresponding N-orbital s, p, d-energy band in the solid. Including the factor of 2 for spin, a full s-band would contain 2N electrons, a full p-band 6N electrons and so on. The width of the band is proportional to the strength of the overlap interaction between neighboring atoms. Each energy band will not have the same energy as any other band over any substantial range of the wavevector.

Metals, Insulators and Semiconductors

The position of the Fermi level distinguishes whether the solid is a metal, insulator or semiconductor. Generally those bands that derive from full atomic orbital states will be fully occupied with electrons, those that derive from empty atomic states lead to empty energy bands, and those from partially full atomic states lead to partially full energy bands. The highest band of those that are filled with electrons is called the valence band; the next highest band is called the conduction band (the fundamental energy gap, E_G, is the valence/conduction band separation, referred to from now on as simply the energy gap). For a metal, E_F is in the conduction band at all temperatures; for an insulator, E_F is close to the middle of the energy gap. Semiconductors are a class of crystals with small energy gaps; they are insulators at $T = 0\,K$, but are conductors at finite temperatures due to either thermal excitation of electrons across the gap or doping, n- or p-type (see below).

Effective Mass

Consider what happens to an electron when an external electric field F is applied. Suppose the electron is initially in a state k. When the field has acted on the electron for a small time dt, it has gained an energy:

$$\mathrm{d}E = \mathrm{e}Fv\mathrm{d}t = (\mathrm{e}F/\hbar)(\mathrm{d}E/\mathrm{d}k)\mathrm{d}t \qquad [13]$$

since the group velocity of the de Broglie wave (energy, $E = \hbar\omega$) is $v = \hbar^{-1}(\mathrm{d}E/\mathrm{d}k)$. Or simplifying we have: $\mathrm{d}k/\mathrm{d}t = \mathrm{e}F/\hbar$. This gives for the resulting acceleration:

$$\begin{aligned} a &= \mathrm{d}v/\mathrm{d}t = (1/\hbar)(\mathrm{d}^2k/\mathrm{d}t^2)(\mathrm{d}k/\mathrm{d}t) \\ &= (\mathrm{e}F/\hbar^2)(\mathrm{d}^2E/\mathrm{d}k^2) \end{aligned} \qquad [14]$$

Comparing this result for the acceleration of a free electron of mass m under the influence of F, $a = \mathrm{e}F/m$, it follows that the electron behaves as if it had an effective mass, m^*, given by

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{\mathrm{d}^2E}{\mathrm{d}k^2} \qquad [15]$$

Thus, the electron effective mass m^* gives a measure of the conduction band curvature near the minimum at $k = 0$. We note that the same result is achieved by differentiating twice the 'nearly free electron' E–k relation, $E = \hbar^2k^2/2m^*$.

Electrons and Holes

Consider the situation where a single electron is excited either optically or thermally from the top of the filled valence band to the bottom of the conduction band of a 1D lattice, and consider the collective behavior of the valence band when an external electric field is applied. Denoting the charge of an electron by −e and the velocities of the electrons by v_i, we may write for the current associated with all electrons in the completely filled valence band in the absence of an external field:

$$\mathrm{I} = -\mathrm{e}\sum_i v_i = -\mathrm{e}[v_j + \sum_{i\neq j} v_i] = 0 \qquad [16]$$

Thus, if the electron j were missing, we should have current

$$I' = -\mathrm{e}\sum_{i\neq j} v_i = \mathrm{e}v_j \qquad [17]$$

Applying an external field F, the rate of change of the current due to the field is:

$$\mathrm{d}I'/\mathrm{d}t = \mathrm{e}(\mathrm{d}v_j/\mathrm{d}t) = -\mathrm{e}^2F/m_j^* \qquad [18]$$

Since the holes are at the top of the valence band, m_j^* is negative and the right hand side of the equation is positive – i.e., a positive external field F gives a positive change of current at the top of the valence band. Thus, a band in which an electron is missing behaves as a 'positive hole' with an effective mass $m_h = -m_e$, and positive charge +e. Since m_e is negative near the top of the valence band, m_h is positive.

Impurities, Intrinsic and Extrinsic Conductivity

The key point about semiconductors from a device point of view is that one can control their conductivity. The basic method is by doping the host crystal with impurities either of valence greater than the host (donors giving excess electrons, called n-type) or of valence less than the host (acceptors giving excess holes, called p-type). The conductivity is referred to as intrinsic conductivity (from intrinsic carrier concentration, N_i) and extrinsic conductivity (from extrinsic carrier concentration, N_e) for undoped and doped material, respectively.

The impurities are chosen to have very low binding energies, so that the electrons (holes) are substantially ionized at room temperature to provide n-type (or p-type) conductivity. These shallow impurities can be treated using the simple Bohr theory of the hydrogen atom, substituting the electron orbiting around the proton in the hydrogen atom under the Coulomb potential e/r, by the electron in the crystal (electron effective mass m^*) orbiting round the ionized (donor/acceptor) impurity atom under the Coulomb potential $e/\varepsilon r$ in the crystal – where ε is the static dielectric constant of the crystal. As with the hydrogen atom, this gives a set of discrete energy levels:

$$E = \frac{m^*/m_o}{\varepsilon^2}\frac{13.6}{n^2}\text{eV} \qquad [19]$$

where 13.6 eV is the binding energy of the H-atom and n is the principal (Bohr) quantum number. The binding energy is then

$$E_B = \frac{m^*/m_o}{\varepsilon^2} \times 13.6\ \text{eV} \qquad [20]$$

For typical semiconductors (Ge, Si...) we have $m^*/m_o \sim 0.05$, $\varepsilon \sim 16$. Thus, E_B is in the range of a few meV (i.e., much less than E_G, and kT at room temperature). The energy level scheme for a doped semiconductor is shown schematically in real space (at room temperature) in Figure 2.

The total conductivity will be $\sigma = \sigma_i + \sigma_e$, where the intrinsic conductivity $\sigma_i = n_i e^2 \tau/m^*$ and the extrinsic conductivity $\sigma_e = n_e e^2 \tau/m^*$. τ is the so-called carrier scattering time. The intrinsic carrier concentration is related to temperature by $n_i = n_0 \times \exp(-E_G/2kT)$, where n_0 is a constant. In the extrinsic regime ($n_e \gg n_i$), at temperatures such that $kT \gg E_d$ (so that the impurities are all ionized), the conduction (valence) carrier concentration $n_e \to n_d$ (or n_a), the donor (or acceptor) concentration.

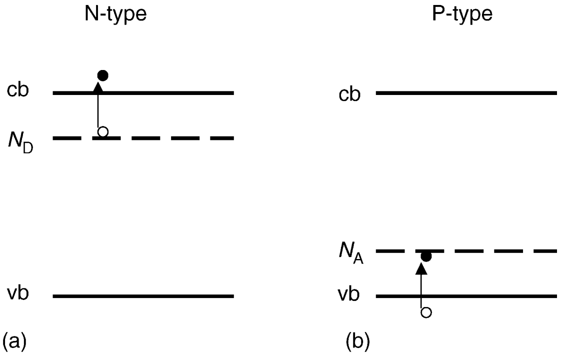

Figure 2 Doping in semiconductors. In n-type material (a), carriers are excited thermally from the donor states to the conduction band where they can contribute to the conductivity. In p-type material (b), electrons are effectively 'captured' by the acceptor states, leaving behind holes in the valence band which can also contribute to conduction. N_D and N_A are the concentration of donor and acceptor impurities, respectively.

Infrared Absorption – Interband Optical Transitions, Excitons

Apart from thermal promotion of electrons and holes via interband (valence to conduction band) transitions responsible for intrinsic conductivity, it is possible to directly excite such transitions optically by subjecting the sample to radiation with $\hbar\omega > E_G$. If the lowest conduction band minimum is situated vertically above the valence band maximum (usually at $k = 0$), then the optical transition process is referred to as a direct transition. If the lowest conduction band minimum is situated at some other point in k-space than that of the valence band maximum, the process is referred to as an indirect transition; this requires the participation of optical phonons (i.e., quantized lattice vibrations) to satisfy momentum conservation, and is therefore a weaker (higher order) process. The semiconductors are respectively referred to as Direct Gap and Indirect Gap Semiconductors. Simplified energy band diagrams for direct and indirect transitions are shown in Figure 3.

The optical spectrum corresponding to the onset of interband electric dipole transitions is referred to as the absorption edge. The most widely used indirect gap materials are the group IV semiconductors Si and Ge. The best known direct gap materials are the III–V's (InSb, GaAs, GaP, etc.), II–VI's (HgTe, CdTe, ZnSe, etc.) and IV–VI's (lead salts PbTe, PnSe, SnTe) and alloy combinations of all these. An important additional effect, which becomes progressively stronger as the energy gap becomes larger in a semiconductor is the Coulomb interaction between the electrons and holes created by interband optical transitions. The electron is raised into an excited state but remains bound in a hydrogen-like orbit around the positive hole creating a two particle excitation called an exciton. The composite particle can then move throughout the crystal. In an ideal crystal, therefore, the absorption spectrum should consist of a series of discrete lines corresponding to the ground and excited exciton states, followed at somewhat high energies by the continuum associated with interband electronic transitions. The energy levels of the exciton spectrum measured from the beginning of the ionization continuum will be given by the Bohr formula:

$$E_{ex} = \frac{\mu}{\varepsilon^2}\frac{13.6}{n^2}\text{eV} \qquad [21]$$

where μ is the reduced mass given by: $1/\mu = 1/m_e + 1/m_h$ (m_e, m_h are electron and hole effective masses). A schematic diagram of the absorption edge is shown in Figure 4 (dashed line – in the absence of excitons; solid line – in the presence of excitons).

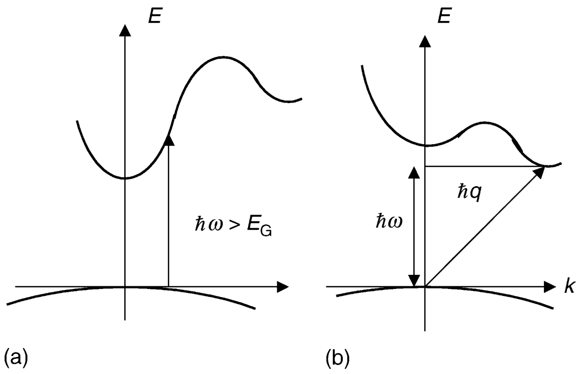

Figure 3 (a) Direct and (b) indirect gap semiconductors. Direct gap transitions involve only the absorption or emission of a single photon and are therefore very nearly 'vertical' transitions in k-space. On the other hand, indirect transitions have a 'two-stage' nature, involving the participation of both a photon and an optical phonon that carries sufficient momentum $\hbar q$ to allow the conduction band electron created in the absorption process to have a large non-zero wavevector.

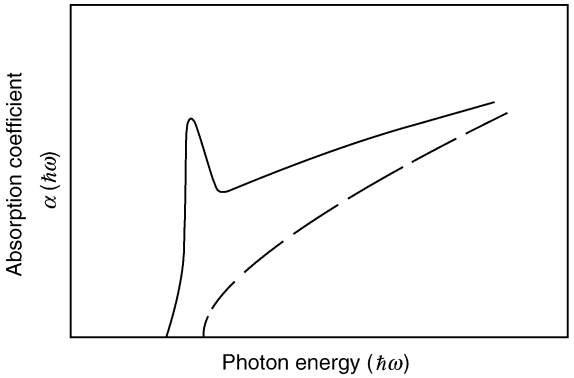

Figure 4 Absorption edge spectra without (dashed curve) and with (solid curve) the excitonic effect.

Conduction and Valence Band Parameters in Typical Semiconductors

As has been described earlier in real semiconductors the conduction and valence bands retain the symmetry of the parent atomic states from which they derive. For example, the conduction and valence bands of silicon can be regarded as deriving from 3s and 3p electrons, and it is generally true that for all group IV, III–V and II–VI semiconductors commonly used in electronic applications the conduction band has s-like symmetry and the valence band has p-like symmetry. This means that the valence band has the additional complication that it is triply degenerate (c.f. the p_x, p_y, p_z components of the hydrogen atom p-orbital); and further, in the presence of spin-orbit coupling (familiar from atomic physics) part of this degeneracy is raised, resulting in doubly degenerate (at $k = 0$) so-called heavy hole (hh) and light hole (lh) bands, and a band split-off by the spin-orbit splitting, Δ:

$$E_{hh} = -\hbar^2k^2/2m_{hh}\ ;\ E_{lh} = -\hbar^2k^2/2m_{lh}\ ;$$
$$E_{so} = -\Delta - \hbar^2k^2/2m_{so} \qquad [22]$$

The resulting conduction and valence bands near $k = 0$ are shown in Figure 5.

Low Dimensional Systems: Quantum Wells

So far we have been concerned almost entirely with the bulk properties of crystalline solids. Such properties can normally be specified in terms of coefficients that are independent of the shape and size of a particular specimen. Thus, for example the electrical

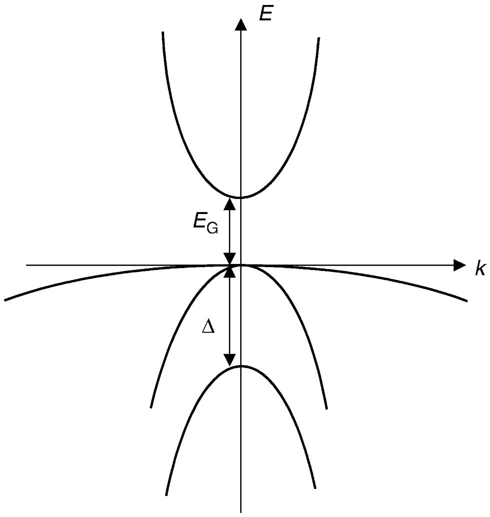

Figure 5 Energy bands (s-like conduction and p-like valence bands) of a direct gap semiconductor near the Brillouin zone center.

conductivity and resistivity, or the optical absorption coefficient and refractive index are properties only of the material and not of the sample dimensions. When one or more dimensions of a solid are reduced sufficiently, the properties are no longer given by these bulk coefficients. The sample is then described as a low-dimensional system (LDS). Such systems form the basis of many optoelectronic devices. LD systems can be classified according to the number of dimensions that are small: thin films (and heterojunctions) are two-dimensional since only the thickness is reduced, and these form so-called quantum wells; fine wires are one-dimensional since two dimensions are reduced, forming quantum wires; dots or specks are zero-dimensional, forming quantum dots.

From an optoelectronic point of view, the most important example is that of electrons confined to a semiconductor film or junction; this system is called the two-dimensional electron gas (2-DEG) and the structure a quantum well. Thus, it is possible to create a narrow electric potential junction region so that the electrons are restricted in 1D (say z) according to the quantum mechanics of a particle in an infinite potential well (described earlier), but 'free' within the xy junction plane.

See also

Semiconductor Materials: Band Structure Engineering; Quantum Dots. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Infrared Lattice Properties; Quantum Wells and GaAs-Based Structures; Recombination Processes.

Further Reading

Hook JR and Hall HE (1991) *Solid State Physics*, 2nd edn. Chichester, UK: John Wiley and Sons, Inc.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. Chichester, UK: John Wiley and Sons, Inc.

Seeger K (1989) *Semiconductor Physics: an Introduction*, 4th edn. Berlin: Springer.

Band Structure and Optical Properties

W Zawadzki, Polish Academy of Sciences, Warsaw, Poland

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The purpose of this article is to describe the influence of the electronic energy band structure of semiconductors on their optical properties. By the band structure one understands the energy–wavevector dispersion relation $\varepsilon(\mathbf{k})$ for the electron energy bands and the relative positions of band maxima and minima in the Brillouin zone.

The main problem in calculating the band structure of a semiconductor is that of the periodic potential of the lattice in which the electrons move. It is, in fact, a self-consistent problem since the moving electrons partly screen the potential. Different approximations have been developed to deal with the question, and in all of them the symmetry of the lattice is of fundamental importance. Thus, in the so-called empty-lattice approximation one deals exclusively with the symmetry and periodicity of the lattice without specifying the potential. This can give qualitative ordering and symmetry of the bands but no quantitative results. In the opposite limit one uses the tight binding approximation, in which the bands are constructed from the atomic states of separate atoms. This method gives quite a good description of lower (valence) bands but poor results for higher (conduction) bands. A powerful way to describe real energy bands is obtained by various forms of pseudopotential calculation, where one approximates the potential near actual atoms by simple parametrized forms and then adjusts the parameters to fit experimental (mostly optical) data. The pseudopotential methods give a good overall description of various bands in the entire Brillouin zone, but they do not provide sufficiently precise results near the band extrema. A semi-empirical way to serve the latter purpose is called $\mathbf{k}\cdot\mathbf{p}$ theory which we describe below. Various methods of band structure calculation are reviewed in books on solid state physics as outlined in the suggestions for Further Reading at the end of this article.

In the majority of important semiconductor materials (Si, Ge, many III–V and II–VI compounds) the maxima of the valence bands are at the center of the Brillouin zone (Γ point). The minima of the conduction bands in Si and Ge are not at the Γ point. This means that the fundamental optical interband transitions in these materials are indirect (in the ε–$\mathbf{k}$ diagram), i.e., they require phonon assistance. On the other hand, the minima of conduction bands in important III–V and II–VI compounds are at the Γ point, so that the interband optical transitions are direct and they do not require phonon assistance.

Both systems are utilized in opto-electronic devices, particularly for detectors. However, for emitters (light-emitting diodes and lasers) we are mostly concerned with the second case, on which we concentrate here.

Bloch States

The Bloch theorem states that if the potential $V(\mathbf{r})$ in which the electron moves is periodic with the periodicity of the lattice, then the solutions $\Psi(\mathbf{r})$ of the Schrödinger wave equation

$$\left[\frac{\mathbf{p}^2}{2m_0} + V(\mathbf{r})\right]\Psi(\mathbf{r}) = \varepsilon\Psi(\mathbf{r}) \qquad [1]$$

are of the form $\Psi(\mathbf{r}) = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})u_k(\mathbf{r})$, where $u_\mathrm{k}(\mathbf{r})$ is periodic with the periodicity of the direct lattice, and $\mathbf{k}$ is the wavevector ($\hbar\mathbf{k}$ is the pseudomomentum). The proof of this theorem can be found, for example, in undergraduate textbooks on solid state physics.

k·p Theory

The underlying idea of semi-empirical **k·p** theory is to describe the energy bands in the vicinity of a given point of the Brillouin zone (usually near a band extremum). Symmetry properties are used to minimize the number of unknown band parameters (effective mass, energy gap, etc.) which are then determined by experiment. The initial Schrödinger equation for an electron in the periodic potential $V(\mathbf{r})$ of the crystal lattice reads

$$[\mathbf{p}^2/2m_0 + V(r) + H_{\mathrm{so}}]\Psi = \varepsilon\Psi \qquad [2]$$

where m_0 is the free electron mass and H_{so} is the spin–orbit interaction. Since $H_{\mathrm{so}}(\mathbf{r})$ is also periodic with the lattice periodicity, the solutions of eqn [2] are given by the Bloch states above.

We use **k·p** perturbation theory, as follows, to describe the energy bands near $\mathbf{k} = 0$ (i.e., the so-called Γ point of the Brillouin zone). Since we are interested in small values of $\mathbf{k}$, we expand our unknown cell periodic functions, $u_{l\mathbf{k}}(\mathbf{r})$, in terms of the set of Γ-point functions, $u_{lo}(\mathbf{r})$, whose symmetry we know. One defines a representation $\Phi_{l\mathbf{k}} = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})u_{lo}(r)$, where $u_{lo}(\mathbf{r})$ is the periodic $\mathbf{k}$-independent function satisfying the equation

$$[\mathbf{p}^2/2m_0 + V + \mathrm{H}_{\mathrm{so}}]u_{lo} = \varepsilon_{lo}u_{lo} \qquad [3]$$

whose solution we know in terms of the band-edge energies, ε_{lo}. This representation is sometimes referred to as the Luttinger and Kohn (LK) representation. The u_{lo} functions are in general mixtures of spin up and spin down states because of the spin–orbit interaction. One looks for the solutions of eqn [2] in the form

$$\Phi_{n\mathbf{k}} = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})\sum_l c_l^{(n)}(\mathbf{k})u_{lo}(\mathbf{r}) \qquad [4]$$

where the sum is over all bands and $c_l^{(n)}(\mathbf{k})$ are the coefficients to be determined. Inserting the above form into eqn [2], performing the operation $\mathbf{p}^2$ (i.e., operating twice on the Bloch product function with the operator $\mathrm{p} = -\mathrm{i}\hbar\nabla$) and using the property [3] one eliminates the unknown potential $V(\mathbf{r})$. Multiplying the resulting equation from the left by $u^*_{l'o}$, integrating over the unit cell Ω and using the orthonormality of u_{lo}, one finally obtains

$$\sum_l\left[(\varepsilon_{lo} - \varepsilon')\delta_{l'l} + \frac{\hbar}{m_0}\mathbf{k}\cdot\mathbf{p}_{l'l}\right]c_l^{(n)}(\mathbf{k}) = 0 \qquad [5]$$

for $l' = 1, 2, 3, \ldots$ Here $\delta_{l'l}$ is the Kronecker delta function, $\varepsilon' = \varepsilon - \hbar^2k^2/2m_0$, and $\mathbf{p}_{l'l}$ is the so-called matrix element of the momentum given by $\mathbf{p}_{l'l} = (1/\Omega)\int_\Omega u^*_{l'o}\mathbf{p}u_{lo}\,\mathrm{d}\mathbf{r}$. The index l' runs over all energy bands.

Equation [5] formulates the famous **k·p** theory. The nondiagonal part $\mathbf{k}\cdot\mathbf{p}_{l'l}$ can be eliminated by the perturbation procedure, and the method is referred to as **k·p** perturbation theory. In the second order of perturbation the bands are separated and in each band the carriers are described by a dispersion relation $\varepsilon_l = (\hbar^2/2)\mathbf{k}(\widehat{1/m^*_{lo}})\mathbf{k}$ where $(\widehat{1/m^*_{lo}})$ is an inverse effective mass tensor at the band edge in question. The inverse mass is generally a 3×3 tensor, but for cubic materials in the center of the Brillouin zone it is a scalar, so that $\varepsilon_l(\mathrm{k}) = \hbar^2\mathrm{k}^2/2\mathrm{m}^*_{lo}$ (where $1/\mathrm{m}^*_{lo} \equiv 1/\mathrm{m}_0 + (2/\mathrm{m}_0^2)\,\Sigma_{l\neq l'}\,(\mathrm{p}^2_{ll'})/[\varepsilon_{l0} - \varepsilon_{l'0}]$). We then say that the band is spherical and parabolic. The second order of perturbation is a good approximation if the energy ε (as counted from a nondegenerate band edge) is small compared to forbidden energy gaps. For the triply degenerate p-like valence band one has to do degenerate perturbation theory and treat the bands together as a 3×3 matrix equation.

In the same approximation the wavefunction for a given band is: $\Psi_{l\mathbf{k}}(\mathbf{r}) = \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r})\,u_{lo}(\mathbf{r})$. In the absence of spin–orbit coupling, H_{so}, the u_{lo} function would just have the symmetry of the parent band (i.e., labeled S for the s-like conduction band, and X, Y or Z for the triply degenerate p-like valence band – each with a single spin up or spin down component). In the presence of H_{so}, the triply degenerate valence band states become mixtures of X, Y and Z with mixed spin character and part of the degeneracy is raised.

In atomic notation the presence of H_{so} has transformed the basis from the (l, s) to the (J, m_J) representation. This results in the well-known light and heavy hole bands (degenerate at $k = 0$) and the spin–orbit split-off band.

Under the influence of a radiation field, of vector potential $\mathbf{A}'$ and frequency ω, the optical transition probability for an electron to be raised from state Ψ_i (initial) to Ψ_f (final) is proportional to the square of $\mathrm{M} = (e/mc)\langle\Psi_f|\mathbf{A}'\cdot\mathbf{p}|\Psi_i\rangle$ i.e., it is determined by the same momentum matrix element, p_{fi}, which governs the effective mass. For interband transitions one immediately gets the selection rule $\Delta k = 0$ (i.e., direct transitions). The polarization selection rules are determined by the $u_{l0}(r)$ components of the initial and final states, through the momentum matrix elements.

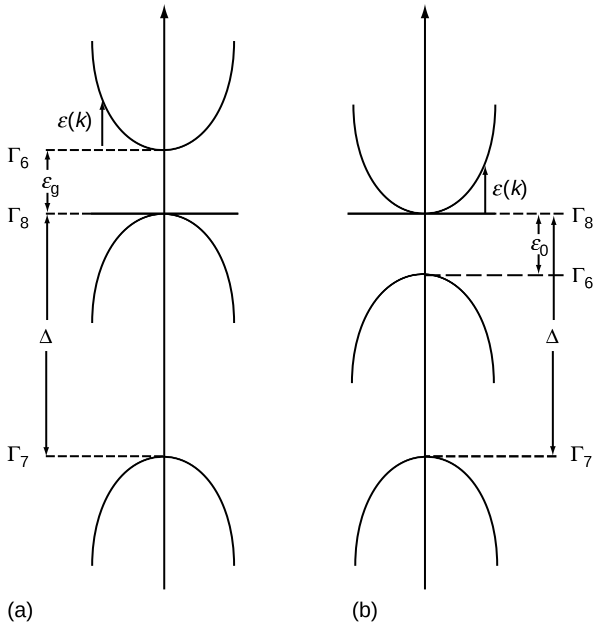

Figure 1 Three-level model of band structure near the Γ point of the Brillouin zone: (a) InSb-type semiconductors; (b) HgTe-type semiconductors.

Narrow-Gap Semiconductors

Semiconductors such as InSb and $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ (with $x < 0.3$) have small energy gaps and are referred to as narrow gap semiconductors (NGSs). They are commonly used in opto-electronic applications which may be accessing energies in the conduction or valence band which are a significant fraction of the energy gap. Under these circumstances it is not valid to cut off the expansion to order k^2, and one has to deal with so-called nonparabolic bands. Thus, in NGSs (or indeed in any situation where the energy becomes an appreciable fraction of ε_{g}) the above procedure is not valid and an alternative approximation is in order. Following semidegenerate perturbation theory one includes exactly in eqn [5] a finite number of bands (near each other in energy) and neglects distant bands. This is referred to as the Kane model, and the energy bands near $k = 0$ are shown for InSb-type III–V compounds and for HgTe-type II–VI compounds in Figure 1. In this case we include only the conduction and valence band in our set of states (eqn [5]). Including spin and degeneracy of the Γ_8 symmetry the set [5] then has eight equations. They can be solved analytically and the resulting energies are given by

$$\varepsilon'(\varepsilon' + \varepsilon'_{\mathrm{g}})(\varepsilon' + \varepsilon'_{\mathrm{g}} + \Delta) - \left(\varepsilon' + \varepsilon'_{\mathrm{g}} + \frac{2\Delta}{3}\right)\kappa^2 k^2 = 0 \quad [6]$$

where $\varepsilon' = \varepsilon - \hbar^2 k^2/2m_0$. We see that for this restricted set of states the only unknown parameters are the energy gap, ε_{g}, the spin–orbit splitting energy, Δ, and $\kappa = -\mathrm{i}(\hbar/m_0)\langle S|p_z|Z\rangle$ (note that this is the only non-vanishing matrix element of the momentum). The fourth energy is $\varepsilon' = -\varepsilon_{\mathrm{g}}$. All four energies are spin degenerate. In NGSs one may neglect the free electron term, i.e., set $\varepsilon' \approx \varepsilon$. For $\varepsilon \ll \varepsilon_{\mathrm{g}} + (2/3)\Delta$ the resulting quadratic equation for the conduction and the light-hole bands is

$$\varepsilon(1 + \varepsilon/\varepsilon_{\mathrm{g}}) = \hbar^2 k^2/2m_0^* \quad [7]$$

where $1/m_0^* = (4\kappa^2/3\hbar^2\varepsilon_{\mathrm{g}})(\varepsilon_{\mathrm{g}} + 2\Delta/3)/(\varepsilon_{\mathrm{g}} + \Delta)$ defines the effective mass at the band edge. The root for the heavy holes is $\varepsilon = -\varepsilon_{\mathrm{g}}$ i.e., this band is not correctly described within the three-level model. The bands given by eqn [6] are spherical but nonparabolic. For $\varepsilon \ll \varepsilon_{\mathrm{g}}$ one recovers the standard dispersion, $\varepsilon = \hbar^2 k^2/2m_0^*$.

For HgTe-type materials, setting the zero of energy at the Γ_8 edge and replacing ε_{g} by $-\varepsilon_0$ (cf. Figure 1), one obtains

$$\varepsilon'(\varepsilon' + \varepsilon_0)(\varepsilon' + \Delta) - \left(\varepsilon' + \frac{2\Delta}{3}\right)\kappa^2 k^2 = 0 \quad [8]$$

For $\varepsilon' \ll (2/3)\Delta$ the dispersion [7] is valid with $1/m_0^* = 4\kappa^2/3\hbar^2\varepsilon_0$. In this case the conduction and the heavy hole bands ($\varepsilon \equiv 0$) are degenerate at $k = 0$, i.e., the thermal gap is zero.

The important property of the above model is that it holds also in the limit of $\varepsilon_{\mathrm{g}} \to 0$ (i.e., for mixed $\mathrm{Hg}_{1-x}\mathrm{Cd}_x\mathrm{Te}$ crystals). The effective mass in eqn [7] tends to zero, but the dispersion described by eqn [6] is valid and it gives $\varepsilon = (2/3)\kappa k$, i.e., a linear band.

The electron and hole wavefunctions in NGSs resulting from semidegenerate perturbation theory are given by eqn [4], in which the sum runs over all the included states. Thus they involve truly **k**-dependent amplitudes of the Bloch states, cf. eqn [1]. In addition, these functions are mixtures of spin-up and spin-down states. These features have important consequences for optical and dc transport phenomena.

If the conduction band minimum is not at the Γ point (Ge and Si), there are at least two different matrix elements of momentum and the resulting band is ellipsoidal.

Effective Mass

The nonparabolic dispersions $\varepsilon(k)$ in NGSs require generalizations of important relations. The momentum mass $\widehat{m}^*$ is introduced as a connection between the psuedomomentum and the velocity: $\hbar\mathbf{k} = \widehat{m}^*\mathbf{v}$. Since $v_i = \delta\varepsilon/\delta(\hbar k_i)$, one obtains for a spherical band

$$\frac{1}{m^*} = \frac{1}{\hbar^2 k}\frac{\mathrm{d}\varepsilon}{\mathrm{d}k} \qquad [9]$$

The above mass is a scalar, but it depends on the energy (or momentum). This mass is more useful than the one usually defined in textbooks, relating force to acceleration. The latter is not a scalar even for a spherical band. However, for the standard parabolic band, $\varepsilon = \hbar^2k^2/2m_0^*$, both masses are the same and are equal to m_0^*. The mass [9] is related via velocity to the electric current, so that it enters naturally in transport phenomena (also free carrier optics). In particular, this mass enters the definition of mobility $\mu = q\tau/m^*$, where q is the charge and τ is the relaxation time. It enters the density of states (see below). Finally, it is the momentum mass [9] which is measured in cyclotron resonance.

For the dispersion [7] (the so-called two-level model), the mass [9] in the conduction band is

$$m^*(\varepsilon) = m_0^*\left(1 + \frac{2\varepsilon}{\varepsilon_g}\right) \qquad [10]$$

Again, for $\varepsilon/\varepsilon_g \ll 1$ the energy dependence of the mass is negligible, which represents the standard regime. Figures 2 and 3 illustrate energy dependences of the electron masses in InSb and HgSe, as measured by various optical and transport methods. The increase of the mass with energy is very strong in both cases.

Density of States

The density of states in energy space is calculated beginning with pseudomomentum space. For the spherical band one obtains

$$\rho(\varepsilon) = \frac{k^2}{\pi^2}\frac{\mathrm{d}k}{\mathrm{d}\varepsilon} = \frac{m^*k}{\pi^2\hbar^2} \qquad [11]$$

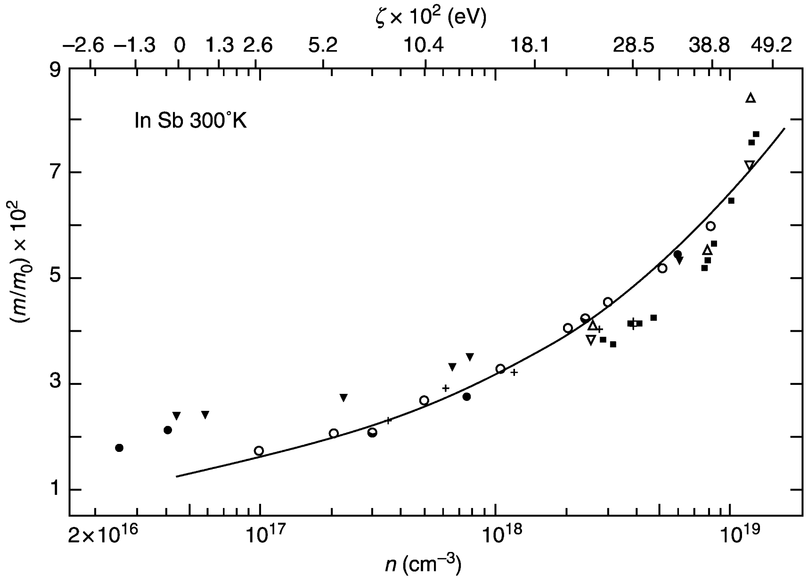

Figure 2 Electron effective mass in InSb at room temperature versus free electron concentration. The solid line, calculated for the two-level model, represents the mass value at the Fermi energy, as indicated on the upper abscissa. The experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

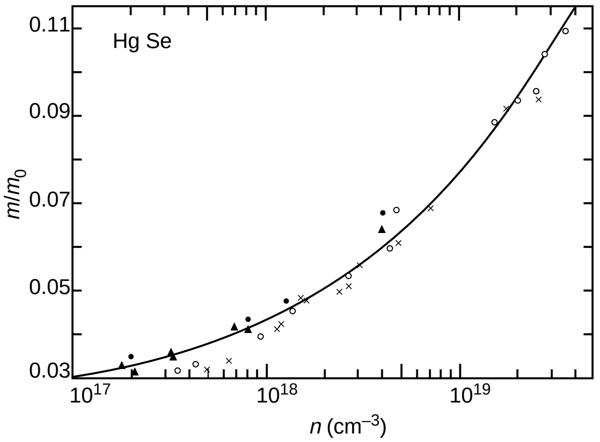

Figure 3 Electron effective mass in zero-gap material HgSe versus free electron concentration. The solid line is calculated for the three-level model. Experimental results of various authors are also indicated. Reproduced from Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

The momentum mass [9] enters naturally into this important quantity. For the parabolic case $m^* = \text{const.}$ and $k \sim \varepsilon^{1/2}$, so that the standard result is recovered. In the general case one should use eqn [10] for $m^*(\varepsilon)$ and determine k from eqns [6] or [7].

Electron–photon Interaction

The electron–photon interaction can be introduced into the theory on two levels. If one replaces in the initial equation [2] the momentum $\mathbf{p}$ by $\mathbf{p} + (e/c)\mathbf{A}'$, the interaction term is obtained in the scalar form $H_{\text{int}} = (e/m_0c)\mathbf{A}'\cdot\mathbf{p}$, where $\mathbf{A}'$ is the vector potential of radiation. The wavefunctions to be used for the calculation of matrix elements are of the form [4], i.e., they include the periodic LK amplitudes. In fact, the matrix elements of momentum for optical transitions are identical to those of the $\mathbf{k}\cdot\mathbf{p}$ theory, as pointed out earlier. Since (as noted above) the wavefunctions for all bands have the same form [4], differing only by the coefficients $c_l^{(n)}(\mathbf{k})$, the matrix elements for interband and intraband optical transitions have the same form.

The electron–photon interaction can also be introduced directly to the $\mathbf{k}\cdot\mathbf{p}$ theory of eqn [5] replacing $\hbar\mathbf{k}$ by $\hbar\mathbf{k} + (e/c)\mathbf{A}'$. If the free electron term $\hbar^2k^2/2m_0$ is neglected, the interaction matrix involving $\mathbf{A}'$ terms is a number matrix. In this procedure the wavefunctions for initial and final states are simply columns and rows of $c_l^{(n)}$ coefficients and LK amplitudes no longer come into play. Here the $\mathbf{p}_{l'l}$ elements of $\mathbf{k}\cdot\mathbf{p}$ theory determine directly the electron–photon interaction. Both procedures described above give the same results.

Quantum Wells

If charge carriers are placed in a quantum well described by a potential $U(z)$, the motion in the z-direction is quantized while the motion in the x–y plane remains free. One takes the same LK basis (cf. eqn [3]), but the general form of solution is given by

$$\Psi = \sum_l f_l(r)u_{l0}(r) \qquad [12]$$

since, in the presence of an external potential, $\mathbf{k}$ is not a good quantum number and the envelope functions $f_l(\mathbf{r})$ are unknown. If $U(z)$ is a slowly varying function within the unit cell the potential appears only on the diagonal of the set [5] and $\hbar\mathbf{k}$ is replaced by $\mathbf{p}$. One obtains

$$\sum_l \left[\left(\varepsilon_{l0} + \frac{\mathbf{p}^2}{2m_0} + U - \varepsilon\right)\delta_{l'l} + \frac{\mathbf{p}_{l'l}\cdot\mathbf{p}}{m_0}\right] f_l(\mathbf{r}) = 0 \qquad [13]$$

One can now eliminate the nondiagonal terms applying the standard Luttinger and Kohn scheme of second-order perturbation theory and arrive at the decoupled band equations with effective masses. This results in the standard quantization due to electric (and magnetic, if present) fields, and wavefunctions are of the form $\Psi_{lm} = f_m u_{l0}$ for the m-th sub-band of the l-th band.

For NGSs this procedure is a poor approximation and one should proceed as before including exactly a finite number of bands (cf. eqn [12]). This results in a set of coupled differential equations for the envelope functions $f_l(\mathbf{r})$. In some important cases one can find the eigenenergies without finding the functions by using the semiclassical approximation (the so-called WKB method).

We shall omit the calculations with coupled differential equations by using a semiclassical procedure also in another sense. Namely, we shall generalize the nonparabolic dispersion relation [7] to include the presence of external fields. To this end we observe that, including the potential $U(\mathbf{r})$ on the diagonal of eqn [13], one replaces $-\varepsilon$ by $-\varepsilon + U$. Further, if a magnetic field is introduced to the $\mathbf{k}\cdot\mathbf{p}$ theory one replaces $\hbar\mathbf{k}$ by $\mathbf{P} = \mathbf{p} + (e/c)\mathbf{A}$, where $\mathbf{A}$ is the vector potential of the magnetic field. Thus the semiclassical equation resulting from eqn [7] is

$$\left[\frac{\mathbf{P}^2}{2m_0^*} - (\varepsilon - U)\left(1 + \frac{\varepsilon - U}{\varepsilon_g}\right)\right]\Psi = 0 \qquad [14]$$

It can be seen that for $\varepsilon - U \ll \varepsilon_g$ one recovers the standard one-band approximation mentioned above. However, below we consider the general situation described by eqn [14].

Let us first consider the case of no magnetic field, i.e., $\mathbf{P} = \mathbf{p}$. For $U(\mathbf{r}) = U(z)$ one can separate the variables by looking for solutions in the form $\Psi = \exp(\mathrm{i}k_x x + \mathrm{i}k_y y)\Phi(z)$. One can now transform eqn [14] into

$$\left[\frac{1}{2m_0^*}p_z^2 - \frac{1}{\varepsilon_g}(\varepsilon - \varepsilon_\perp - U)(\varepsilon_g + \varepsilon + \varepsilon_\perp - U)\right]\Phi(z) = 0 \qquad [15]$$

where

$$\varepsilon_\perp = \frac{\varepsilon_g}{2} + \left[\left(\frac{\varepsilon_g}{2}\right)^2 + \varepsilon_g \frac{\hbar^2 k_\perp^2}{2m_0^*}\right]^{1/2} \qquad [16]$$

in which $k_\perp^2 = k_x^2 + k_y^2$. Equation (15) is suitable for the WKB semiclassical quantization with which one determines the eigenenergies ε by one integration.

For a magnetic field applied along the z-direction (parallel electric and magnetic fields) the effects of both fields are separated and the form of eqn [15] is valid, with $\hbar^2 k_\perp^2/2m_0^*$ in eqn [18] replaced by $\hbar\omega_c(n + 1/2) \pm (1/2)|g_0^*|\mu_B B$. Here $\omega_c = eB/m_0^* c$ is the cyclotron frequency, $n = 0, 1, 2, \ldots$ is the Landau quantum number, g_0^* is the spin splitting factor, and μ_B is the Bohr magneton. This results in nonequal spacing of Landau levels for a fixed value of B and in nonlinear B dependence of the Landau energies as functions of B (similarly but not identically to the bulk case). These features are illustrated by the cyclotron resonance experiment and theory shown in Figure 4.

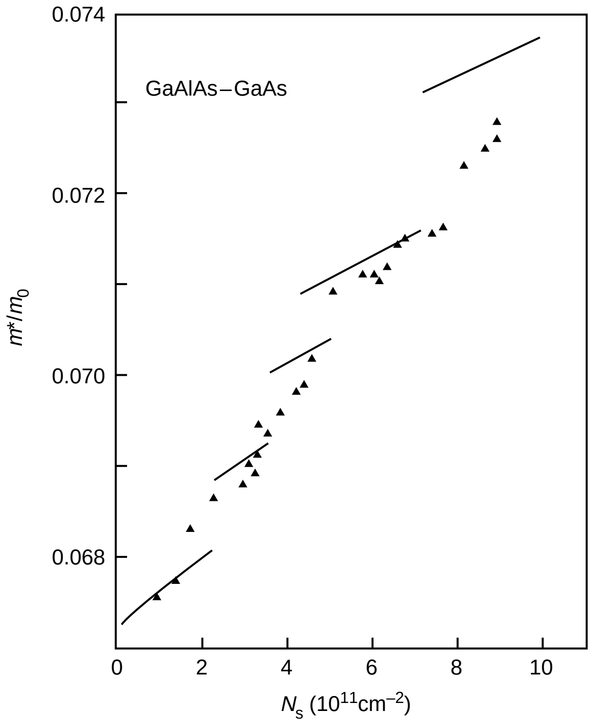

Figure 4 Electron effective mass versus two-dimensional electron density in GaAs/GaAlAs heterostructures, as measured by far infrared cyclotron resonance. The solid lines are calculated using the effective two-level model for GaAs. Reproduced from Zawadzki W, *et al.* (1994) Cyclotron emission study of electron masses in GaAs-GaAlAs heterostructures. *Semiconductor Science and Technology* 9: 320–328.

Selection Rules for Intersub-band Optical Transitions

For the one-band effective mass approximation the electron–photon interaction is $H_{\text{int}} \sim \mathbf{A}'{\cdot}\mathbf{p}$ (cf. above), and the wavefunctions for the initial and final sub-bands are $\Psi_m = \exp(\mathrm{i}k_x x + \mathrm{i}k_y y)\Phi_m(z)$, in which the $\Phi_m(z)$ must be orthogonal to each other since they describe different energies. It is then clear that the matrix element $\langle fn|\mathbf{A}'{\cdot}\mathbf{p}|in\rangle$ is finite only if $\mathbf{A}'$ is polarized along the z-direction (also called the growth direction) since for $\mathbf{A}'$ parallel to x or y the integral over z is $\langle \Phi_{m'}|\Phi_m\rangle = 0$ giving a vanishing matrix element. Experimentally this is a serious problem for spectroscopic applications involving intraband (i.e., intersub-band) transitions in n-type semiconductors, since it is not possible to access these using normal incidence radiation.

This selection rule is somewhat relaxed for p-type semiconductors as a result of the mixed spin states of the complex valence band. In addition the narrow-gap band structure introduces interesting new possibilities into the selection rules. The wavefunctions have the form [12] and for the light polarizations A'_x or A'_y the momentum operators p_x or p_y act also on the periodic amplitudes $u_{l0}(\mathbf{r})$, which leads to a nonvanishing transition probability for intersub-band transitions. This has indeed been observed in narrow-gap materials.

See also

Magneto-Optics: Interband Magnetoabsorption, Cyclotron Resonance, Spin Flip Raman Scattering. **Semiconductor Materials:** Mercury Cadmium Telluride; Quantum Dots. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Ashcroft NW and Mermin ND (1976) *Solid State Physics*. Holt, Rinehart and Winston.

Kane EO (1980) Band structure of narrow gap semiconductors. In: Zawadzki W (ed.) *Narrow Gap Semiconductors. Physics and Applications*, pp. 13–31. Berlin: Springer.

Luttinger JM and Kohn W (1955) Motion of electrons and holes in perturbed periodic fields. *Physical Review* 97: 869–883.

Zawadzki W (1974) Electron transport phenomena in Sm all-gap semiconductors. *Advances in Physics* 23: 435–512.

Zawadzki W (1984) Electric and magnetic quantisation of two-dimensional systems – elementary theory. In: Bauer G, *et al.* (eds) *Two-Dimensional Systems, Heterostructures, and Superlattices*, pp. 2–11. Berlin: Springer.

Zawadzki W (1991) Intraband and interband magneto-optical transitions in semiconductors. In: Landwehr G and Rashba EI (eds) *Landau Level Spectroscopy*, pp. 483–512. Amsterdam: North-Holland.

Excitons

I Galbraith, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

In semiconductors electron–hole pairs can lower their energy by correlating their motion in an exciton state. Such states dominate the bandedge optical properties of direct-gap semiconductors. The resultant absorption peaks are affected by quantum confinement, electric, and magnetic fields and carrier scattering.

Introduction

In the simplest, single-particle picture, the ground state of a semiconductor consists of a full valence band and an empty conduction band. The lowest lying excitation above the ground state in such a scheme is produced by excitation of one electron from the valence band to the conduction band. To achieve this optically requires a photon energy greater than the bandgap, E_g. However in many semiconductor materials this simple picture is unable to explain the observed absorption spectrum in the neighborhood of the fundamental bandgap. The origin of the discrepancy lies in the neglect of the Coulomb interaction between the negatively charged excited electron and the hole which is left in the valence band.

The term exciton refers to a two-particle excitation, consisting of a bound electron and hole. Such excitations dominate the bandedge optical spectra of most direct gap semiconductors, especially at low temperatures. In particular there exists a series of hydrogen-like bound states lying below the bandedge. These states are bound by energies

$$E_b^{3D} = \frac{\mu e^4}{8\mathrm{h}^2 \varepsilon_r^2 \varepsilon_0^2} \frac{1}{n^2} \qquad [1]$$

where $\mu = m_e m_h/(m_e + m_h)$, is the reduced mass, $m_{e(h)}$ the electron (hole) effective mass, e is the electronic charge, h is Planck's constant, ε_0 is the permittivity of free space, ε_r the relative permittivity of the material and $n = 1, 2, 3, \ldots \infty$ is the principal quantum number.

The binding energy of the lowest lying exciton state varies considerably from one semiconductor to another being, e.g., 0.5 meV in InSb and over 60 meV in Zn0. Similarly the spatial extent of the 1 s electron-hole relative wave function or effective Bohr radius is given by

$$a_0^{3D} = \frac{\hbar}{\sqrt{2\mu E_b^{3D}}} \qquad [2]$$

and ranges from 750 μm in InSb to ≈2 nm in Zn0. Typically, in semiconductors the exciton spatial extent is much larger than the lattice constant of around 0.6 nm, and such excitons are classed Wannier excitons. At the opposite limit, which occurs in many molecular materials, the exciton is localized around a particular atomic site and is termed a Frenkel exciton. Wannier excitons are characterized by the hydrogen-like quantum numbers of their relative motion and an overall center of mass momentum which describes the wave-like motion of the bound entity through the crystal.

Another type of exciton is often referred to as a bound exciton. This consists of an electron–hole pair bound to a neutral impurity center. Such excitons are localized around the impurity and cannot move through the crystal in the same way as a free exciton.

Optical Properties

Excitons manifest themselves primarily as strong modifications to the bandedge optical properties of semiconductors. In particular the bound states give rise to discrete absorption lines at lower energy than the bandgap energy. Several peaks corresponding to the different hydrogen-like bound states

(1s, 2s, etc.) can be seen provided the lines are sufficiently narrow. A theoretical calculation of this is shown in Figure 1.

Associated with these absorption changes are concomitant changes in the refractive index. Contributions to the spectral width of the absorption peaks are characterized as either homogeneous or inhomogeneous. Homogeneous broadening refers to broadening associated with the finite lifetime of a particular state, e.g., due to phonon scattering. Inhomogeneous broadening is due to non-uniformity in the system, e.g., in a sample having a spatially non-uniform strain field.

There is also enhancement of the absorption at photon energies above the bandgap due to the residual influence of the Coulomb attraction on the electron-hole scattering states. Although not bound, electrons and holes in these states still have an enhanced probability of being found close together. This is called the Sommerfeld Enhancement Factor (see Figure 1).

It is important to note that although there may be an excitonic peak in the absorption spectrum this does not imply that excitons form a stable population in such a sample. Often, especially at room temperature, the lifetime of an optically generated exciton is determined by the scattering time with optical phonons which may be very short (<1 ps). In this case the excitons are quickly ionized and the quasi-thermal equilibrium which is formed is dominated by essentially plasma-like behavior. In photoluminescence experiments, where carriers are generated high in the band and the spectrum of the resulting emitted luminescence is measured, peaks associated with the free exciton are often observed and taken as a signature of the presence of excitons. This view has been challenged recently by theoretical calculations which show that a photoluminescence peak at the exciton energy can be also be produced from an uncorrelated plasma if proper account of the the Coulomb interaction is taken. This issue remains controversial.

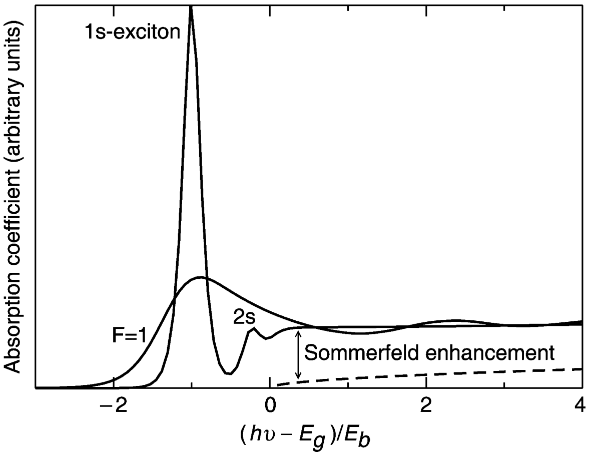

Figure 1 Calculated bulk absorption spectra including (solid-line) and neglecting (dashed line) the Coulomb interaction. The Sommerfeld enhancement is seen as photon energies above the bandgap. Also shown is the spectrum with an applied electric field of one exciton Rydberg per Bohr radius.

In bulk samples, experiments such as absorption and reflectivity are complicated by the existence of polariton effects. An exciton-polariton is a quantum mixture of the propagating photon inside the semiconductor sample and the material excitation. Where the photon dispersion and the exciton energy dispersion meet there is an anti-crossing and this is manifested in, for example, the appearance of two lines in the reflectivity spectrum.

It was hoped during the 1970s that laser action in wide bandgap material such as ZnSe would be possible based on excitonic lasing. In this process a quasi-equilibrium exciton population would form the injected excitation and stimulated recombination would occur with the associated emission of a scattering partner which would take up the necessary momentum to ensure overall momentum conservation. This leads to a light emission wavelength below the absorption edge which is advantageous for minimizing losses. In fact, this process has only been seen at low temperature for scattering with LO-phonons, electrons and other excitons. Each mechanism has its own characteristic lasing threshold and temperature dependence.

Influence of Quantum Confinement

The advent of Molecular Beam Epitaxy enabled the growth of semiconductor layers of thickness similar to the electronic de Broglie wavelength. By sandwiching a low bandgap semiconductor between two high bandgap layers, one can make quantum well structures in which the electronic motion is effectively confined to the plane of the layer. In such structures the excitonic properties are also radically altered by this confinement. An example of the absorption spectrum of a ZnCdSe quantum well embedded in ZnSe is shown in Figure 2. The heavy and light hole excitons correspond to different spin quantum numbers for the electron and hole making up the exciton.

The exciton binding energy is enhanced by a factor of 4 compared to the bulk case and the spatial extent of the 1s wavefunction is also reduced by half:

$$E_b^{2D} = \frac{\mu e^4}{8\mathrm{h}^2 \varepsilon_r^2 \varepsilon_0^2} \frac{1}{(n - \frac{1}{2})^2} \qquad n = 1, 2, 3\ldots \qquad [3]$$

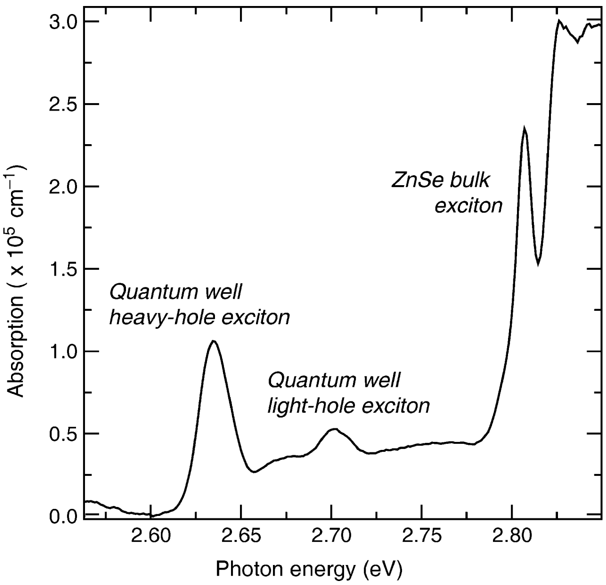

Figure 2 Heavy and light hole exciton absorption spectrum for ZnCdSe quantum wells embedded between ZnSe barriers. Also shown is the bulk exciton of the ZnSe barriers.

This enhancement can be understood since the particles are confined to lie in the plane and are hence closer together on average than in the three-dimensional case. In practice, a four-fold enhancement in the binding energy is never observed since the quantum wells are neither infinitely deep nor infinitely thin. Both these limitations lead to a reduction in the actual enhancement from 4 to about 2, depending on the structure. For example, the binding energy in a realistic quantum well corresponds to the bulk value of the well material for very wide wells and to the bulk value of the barrier material for very narrow wells. In between the value is enhanced by the quantum confinement. The maximum enhancement typically occurs when the well width is around half the bulk Bohr radius for the well material.

Accurate calculation of the exciton binding energy is complicated by the fact that the quantum confinement influences the band structure, making it strongly anisotropic and splitting heavy and light holes. A complete inclusion of these effects is at the limit of current computational power due to the need to calculate the electron-hole interaction matrix elements. Approximate schemes, including variational approaches, have proved successful in fitting a variety of samples, including GaAs and ZnSe based quantum wells. The limitation to such calculations is often the requirement for accurate values of the material parameters such as effective masses and energy band offsets.

The Sommerfeld enhancement of the absorption into the continuum states of quantum wells leads to a doubling of the absorption at the bandedge. This enhancement reduces only slowly as the photon energy is increased.

If multiple quantum wells with very thin barrier layers are used, the electronic states in the neighboring wells couple together to form a superlattice. The superlattice dispersion in the growth direction behaves like a heavy mass and the resulting exciton is three-dimensional but strongly anisotropic.

More extreme quantum confinement is possible making quantum wires, by etching or selective growth of quantum well samples. In these quasi-one dimensional structures, as in quantum wells, the exciton binding energy is again increased further. The Coulomb problem in one-dimension is pathological in that the ground-state binding energy diverges logarithmically, and higher states are pairwise degenerate. However, accounting for the finite cross-section of realistic wires eliminates this problem and recovers a finite binding energy. One consequence of this is that there is no simple enhancement factor which corresponds to the four-fold enhancement in going from bulk to two-dimensional excitons. The density of states in one-dimensional systems diverges at the bandedge, implying the existence of an infinite absorption in a perfect quantum wire. However, the influence of excitonic correlations cancels this divergence, resulting in an enhanced, but not divergent, bandedge absorption feature.

The ultimate limit in quantum confinement can be achieved in quantum dots, where the carriers are confined in all three directions. In this case the term exciton should be used with some care as there are no bandedge states to compare with. The discrete energy levels dictated by the confinement are affected by the Coulomb interaction and the energy is renormalized – in some cases substantially. The absorption spectrum should consist of a series of sharp atomic-like transitions corresponding to s-s, p-p, d-d,... transitions and indeed luminescence spectra obtained from single quantum dots do display such fine structure. Even in the best controlled material systems, such as InAs quantum dots, the spectra are strongly inhomogeneously broadened due to variations in the dot size, shape, and alloy composition.

In large quantum wells, where the width is larger than the exciton Bohr radius the center-of-mass part of the wavefunction can be confined by the barriers. This leads to a quantized motion for the exciton as a whole, rather than for the individual electron and hole separately. One implication of this is that for such excitons the center-of-mass momentum is no longer a good quantum number.

The above effects are a controlled case of disorder induced effects in semiconductors. This is an area which is only now beginning to be understood. The presence of local potential fluctuations will influence the energy spectrum of excitons in that vicinity. Clearly both the depth and spatial extent of these fluctuations will be important – very short range (less than Bohr radius), shallow potential fluctuations will not have any influence on the exciton. While deep potentials can lead to the center of mass confinement above.

Influence of Electric and Magnetic Fields

Just as an externally applied electric field affects the absorption edge via the Franz–Keldysh Effect, broadening the edge itself and inducing oscillations in the continuum part of the spectrum, so, in the case of an excitonic bandedge, there is substantial modification due to the presence of the field. At low fields, when the Coulomb interaction is strong compared to the potential drop across the exciton diameter, the exciton will be little affected by the applied field. At higher fields the exciton will become gradually more polarized by the field until the exciton is field ionized. This ionization process is manifested in the field broadening of the exciton line (see Figure 1).

Of more technological importance is the influence of an electric field applied perpendicular to the plane of a quantum well. The field forces the electron and hole in opposite directions leading to a reduction in the exciton binding energy. This would lead to a blue shift in the exciton absorption peak were it not for an even larger reduction in the single particle electron and hole energy levels within the quantum wells. Overall a strong red-shift of the excitonic absorption peak is observed with increasing field. This is accompanied by a reduction in the oscillator strength due to the reduced electron-hole overlap. This is termed the Quantum Confined Stark Effect and is the basis for the operation of some semiconductor electro-optic modulators. An electric field applied in the plane of the quantum well gives rise to Franz–Keldysh oscillations analogous to the bulk case.

The influence of magnetic fields on excitons is more subtle and richer than the electric field. This is because the magnetic field seeks to induce a circular cyclotron orbit motion on the carriers, but this is influenced by their mutual Coulomb interaction. Two regimes exist where either the cyclotron energy:

$$\hbar\omega_c = \hbar(eB/\mu) \tag{4}$$

or the exciton binding energy dominates. In the former strong field case, one can treat the Coulomb interaction as a small perturbation on the electron and hole states in the presence of the magnetic field. In exciton states, which have a finite magnetic moment at zero field, we find a linear Zeeman shift in the magneto-exciton energy. In the other case we must use the magnetic field as the perturbation which produces a mixing of the zero-field exciton states. For example, for the ground state the magnetic field induces an admixture of the p-like excited state with the s-like ground state. The total angular momentum of the mixed state is proportional to B and as the energy of a magnetic dipole in a magnetic field is also proportional to B the shift of the magneto-exciton is proportional to B^2. This distinction between linear and quadratic shifts has been used to identify the nature of carrier populations (excitons or unbound free carriers) in a variety of samples. Clearly there exists an intermediate regime where both the magnetic and Coulomb energies are comparable and in this case the precise energy shifts need to be evaluated numerically.

In quantum well samples the orientation of the magnetic field with respect to the confinement direction is crucial in determining the physics. When the field is perpendicular to the confinement plane the cyclotron orbits exist as before and essentially the same phenomenon as in the bulk emerge. For other orientations, the behavior is much more complex and beyond the scope of this article.

Exciton Scattering

An electron and hole in an excitonic bound state execute a correlated motion which can be disturbed by scattering of either partner. This may lead to the ionization of the exciton and the destruction of the correlation or, alternatively, it may change the center of mass momentum of the exciton as a whole. Almost all of the important scattering mechanisms for excitons arise from the charged nature of the electron and hole. Via the Coulomb interaction, excitons can scatter with lattice phonons, other excitons, free carriers, and impurities. Each of these scattering processes has its own regime of dominance, dependent on, for example, temperature or material quality.

When the scattering rate with other particles is much larger than the recombination rate for electrons and holes, a quasi-equilibrium state is reached. The detailed nature of such an interacting electron/hole plasma remains a a long-standing open question. The reasons for this can be traced to interplay between the intrinsic Coulomb interactions within the plasma, which give rise to both bound and scattering states, and Pauli exclusion arising from the fermionic nature of the electrons and holes. This leads to a complex phase diagram which encompasses, e.g., electron-hole droplets, the ionized electron-hole plasma, biexciton phase, and the exciton gas which, some suggest, may undergo Bose–Einstein condensation. The parameters of this phase diagram are carrier density, temperature, and the semiconductor material parameters. Interest in this essentially fundamental question has remained high, stimulated by the stream of technological benefits that even partial answers have brought.

In this context it is worth mentioning that there have been two theoretical approaches to exciton physics which each have their advantages and problems. One approach is to treat excitons as bosonic quasi-particles and derive results from Hamiltonians formulated using bosonic operators. This approach has the appeal of simplicity and produces acceptable results in the low-density regime. It does, however, omit a key feature of the constituent particles making up an exciton namely the fermionic nature of electrons and holes. A more rigorous approach based around electron and hole operators has been followed but this is considerably more complex and numerically involved.

Excitonic Molecules

Just as two hydrogen atoms can lower their total energy by forming a bound hydrogen molecule, so two excitons can make a bound complex which has lower energy than the two isolated excitons. Such complexes are called biexcitons and are mostly important in high quality material and at low temperatures. Their binding energy is less than that of the exciton by a factor of about 0.1–0.3, depending on the ratio of electron to hole effective masses.

Three particle complexes have also been observed consisting of two electrons bound to a hole. These are termed trions or charged excitons. Their binding energy lies intermediate between excitons and biexcitons.

See also

Semiconductor Materials: GaAs Based Compounds; Large Gap II–VI Semiconductors; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Bastard G (1998) Wave Mechanics Applied to Semiconductor Heterostructures. New York: Halstead John Wiley.

Haug H and Koch SW (1993) *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. Singapore: World Scientific.

Klingshirn CF (1997) *Semiconductor Optics*. Berlin: Springer.

Rashba EI and Sturge MD (eds) (1982) *Excitons*. Amsterdam: North Holland.

Schmitt-Rink S, Chemla DS and Miller DAB (1989) Linear and nonlinear optical properties of semiconductor quantum wells. *Advances in Physics* 38: 89–188.

Ueta M, Kanzaki H, Kobayashi K, Toyozawa Y and Hanamura E (1986) *Excitonic Processes in Solids*. Berlin: Springer.

Yu PY and Cardona M (1999) *Fundamentals of Semiconductors*. Berlin: Springer.

Impurities and Defects

K A Prior, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A simple definition of a defect is any feature that disturbs the periodicity of the perfect lattice. This definition is sufficiently broad to encompass a whole range of different entities, which are normally divided into three categories based on their dimensionality, or extent in real space.

0D defects are those which have an extent of the order of only a few atomic spacings at most in any direction. They include all cases where one, or at most a few, atoms are removed or displaced, or are replaced by atoms of a different type.

1D defects are linear features within the crystal that extend over many atomic dimensions. The most important type of 1D defect is the dislocation and in a single crystal it is possible for a dislocation to extend from one side of the crystal to the other.

2D defects include a variety of different features that occur primarily in one plane. Any junction between two different types of crystal in an epitaxially grown layer (a heterointerface) fits this description, as of course does the sample surface. There are also crystal defects which exist in two dimensions of which the most important is the stacking fault.

Finally, there are also 3D defects. This category includes voids and inclusions, which are both major perturbations in a crystal structure. In most pure crystals grown by modern techniques these are eliminated and they will not be discussed further.

In the sections below the main types of defect will be discussed according to the scheme described above. Despite the substantial research effort to understand the basic properties of many semiconductors, and of course the huge effort devoted to producing semiconductor devices, it remains true that the defect structure of most semiconductors is still poorly understood, with the exceptions of silicon and gallium arsenide. This is especially true in the case of the nitride semiconductors with the wurtzite structure which have recently been the subject of much commercial interest. In the following sections examples will mainly be taken from gallium arsenide.

Lastly, it is important to note that defects of different dimensionality interact. Several possible interactions are discussed in the last section.

0D Defects

Dopants

There are two basic types of 0D defects to consider: chemical defects, or impurities, and native or point defects caused by missing or misplaced atoms. Table 1 gives the major types of 0D defect found in compound semiconductors together with the symbol used in the common notation for point defects and impurities.

Impurity atoms are those which are introduced either accidentally during the production of the bulk crystal or epitaxial layer, or deliberately by the crystal grower in precisely controlled amounts to modify some property of the material. Usually, this means that the material added is a dopant, and has been added to create a surplus of either free electrons or holes in the semiconductor. Typically, the amounts of dopants added are between 10^{-6} and 10^{-3} mole fraction, giving doping levels between $10^{22}\ \mathrm{m}^{-3}$ and $10^{25}\ \mathrm{m}^{-3}$.

Table 1 Impurities and point defects in a compound *AB* with the zincblende lattice structure

Symbol	*Description*
X_A^n, X_B^n	Impurity atom, *X*, on *A* or *B* site with charge *n*
X_i^n	Interstitial impurity atom with charge *n*
V_A^n, V_B^n	Vacancy on *A* or *B* site with charge *n*
A_i^n, B_i^n	Interstitial *A* or *B* atom with charge *n*
A_B, B_A	*A* atom on *B* site or *B* atom on *A* site (antisite)

In general, any atom added to a material will introduce a localized state into the band structure of the semiconductor. This state may rarely occur in either the conduction or valence band, but is more likely to be located in the bandgap. If the localized energy level is within a few kT of the valance or conduction band edge, then the state is an electron acceptor or donor. If it is much further from the band edge then it is usually known as a deep level. In this case its main effect on the electrical properties will be to act as a trap for electrons or holes, and introduce noise into the current flow. Its effect on the optical properties of the material will be to reduce the radiative recombination rate for band-to-band transitions by acting as an alternate recombination pathway for either radiative or nonradiative transitions.

It is generally assumed that dopant atoms occur only on the site of the appropriate sublattice and are well behaved. While it is true that some elements are well behaved substitutional dopants, such as sulfur in gallium arsenide, which can be denoted $\mathrm{S}_{\mathrm{As}}^{+}$ when acting as a singly charged donor, the behavior of other species is more complex. In many III–V compounds, silicon is an amphoteric dopant, occupying sites on both the metal and nonmetal sublattices which ionize as shown:

$$\mathrm{Si}_{\mathrm{Ga}}^{0} = \mathrm{Si}_{\mathrm{Ga}}^{+} + \mathrm{e}^{-}$$

$$\mathrm{Si}_{\mathrm{As}}^{0} = \mathrm{Si}_{\mathrm{As}}^{-} + \mathrm{h}^{+}$$

In molecular beam epitaxy (MBE) growth on GaAs(100) the ratio between the donor $\mathrm{Si}_{\mathrm{Ga}}^{+}$ and the acceptor $\mathrm{Si}_{\mathrm{As}}^{-}$ varies as the concentration of doping is increased. At low levels silicon is a donor, while at high concentrations the fraction which is an acceptor increases rapidly. The semiconductor is therefore compensated, effectively limiting the maximum concentration of free carriers.

Another type of behavior occurs with beryllium in III–V semiconductors and lithium in II–VI semiconductors. Here, the dopant atom is so small that it can easily be displaced from the lattice site. Thus, for example, in InGaAs, it is energetically favorable for the acceptor $\mathrm{Be}_{\mathrm{In}}^{-}$ atom to move from the lattice site to an interstitial position:

$$\mathrm{Be}_{\mathrm{In}}^{-} = \mathrm{V}_{\mathrm{In}}^{-} + \mathrm{Be}_{\mathrm{i}}^{+} + \mathrm{e}^{-}$$

The interstitial Be ion is a donor and causes some compensation, but even worse, is highly mobile. It can move a substantial fraction of a micron in an epitaxial device and in so doing completely destroys a carefully engineered abrupt p-n junction. It can be seen therefore that dopant interactions with the host lattice are important in determining device performance.

Native Defects

A large variety of native defect species are possible in semiconductors. Interstitial atoms and vacancies are found in both the diamond structure (silicon) and the zincblende structure (most III–V and II–VI semiconductors). In zincblende structures there is also the possibility of placing an atom on the wrong atomic site, or sublattice; for example in GaAs a Ga atom may be found on an As site. This type of defect is known as an antisite.

Bulk semiconductor crystals are grown under conditions close to equilibrium. In the long periods of time needed to grow the crystals, equilibrium concentrations of point defects are established. Epitaxial growth techniques (in particular molecular beam epitaxy, MBE and metalorganic chemical vapor deposition, MOCVD) operate under conditions far from equilibrium and the point defect concentrations may differ markedly from the equilibrium values. Some semiconductor processing operations can also increase point defect concentrations. Etching semiconductors with ion beams or energetic plasmas is a very efficient way of creating in the subsurface region high concentrations of point defects well above the equilibrium values.

It is possible to anneal crystals in the vapor of one of the constituent elements to alter the point defect concentrations. Interstitial and vacancy concentrations can be changed relatively rapidly over the course of several hours (interstitials) or days (vacancies) as these species are much more mobile than dopant atoms. Usually, annealing is done to reduce point defect concentrations, but on occasions, for example with the alloy HgCdTe, it is done to increase the concentration of a type of defect, in this case mercury vacancies which are used to dope the material.

Like dopant atoms, point defects can also introduce defect levels into the band structure of the semiconductor, and some defects, for example the gallium vacancy $\mathrm{V}_{\mathrm{Ga}}^{n-}$, exist in multiple charge states. In some materials, the native defects act either as dopants or as compensating traps, and are present in concentrations that can severely limit the maximum obtainable doping levels. The gallium vacancy, for example, is believed to limit the maximum doping level obtainable in n-type GaAs, and its concentration increases dramatically as the concentration of n-type dopant is increased.

It should be noted that the concentrations of the various types of defect are not independent of one another, and at any given growth (or annealing) temperature and applied partial pressures of the constituent elements, the equilibrium concentrations of the defects are fixed. Examples of the interactions between the types of defect may be seen by considering the displacement of a gallium atom from its lattice site to create a vacancy and an interstitial (the Frenkel reaction):

$$\mathrm{Ga}_{\mathrm{Ga}}^{0} = \mathrm{Ga}_{\mathrm{i}}^{0} + \mathrm{V}_{\mathrm{Ga}}^{0}$$

and the interaction of an arsenic interstitial with a gallium vacancy to create an arsenic antisite:

$$\mathrm{As}_{\mathrm{i}}^{0} + \mathrm{V}_{\mathrm{Ga}}^{0} = \mathrm{As}_{\mathrm{Ga}}^{0}$$

The relative importance of the types of defect varies between semiconductors. In compounds where one of the elements is very volatile, then high concentrations of vacancies on that particular sublattice are to be expected. Antisites occur only at very low concentrations in the ionic II–VI compounds, as the energy involved in placing an atom on the wrong sublattice is very high. However, in the more covalent III–V compounds, there is not such a large energy penalty, and it has been suggested that in GaAs the $\mathrm{As}_{\mathrm{Ga}}$ antisite may be the most common point defect.

In addition to the simple defect types described above, there are other possibilities. One of these is that complexes of defects may form. The best known are the complexes that occur between dopant atoms and vacancies, for example the $[\mathrm{Te}_{\mathrm{As}}\mathrm{V}_{\mathrm{Ga}}]^{-}$ complex found in Te doped GaAs (Figure 1). In this material, partial compensation of the n-type dopant occurs with the formation of gallium vacancies, as described above:

$$\frac{1}{2}\mathrm{As}_2(\nu) = \mathrm{V}_{\mathrm{Ga}}^{0} + \mathrm{As}_{\mathrm{As}}$$

$$\mathrm{V}_{\mathrm{Ga}}^{0} = \mathrm{V}_{\mathrm{Ga}}^{2-} + 2\mathrm{h}^{+}$$

These vacancies are mobile and diffuse in the material. In the vicinity of $\mathrm{Te}_{\mathrm{As}}^{+}$ they experience a strong Coulombic attraction and the complex forms when the vacancy and dopant atom are on nearest neighbor sites:

$$\mathrm{V}_{\mathrm{Ga}}^{2-} + \mathrm{Te}_{\mathrm{As}}^{+} = [\mathrm{V}_{\mathrm{Ga}}\mathrm{Te}_{\mathrm{As}}]^{-}$$

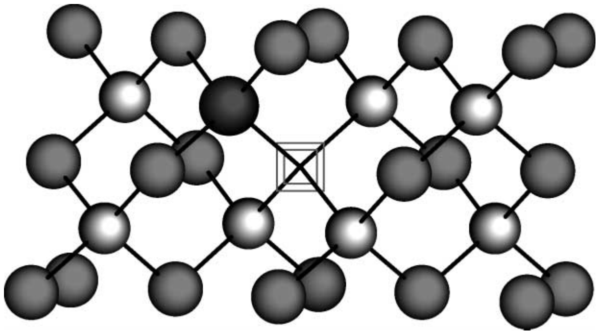

Figure 1 Diagram of the tellurium atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the tellurium atom is shown in dark gray.

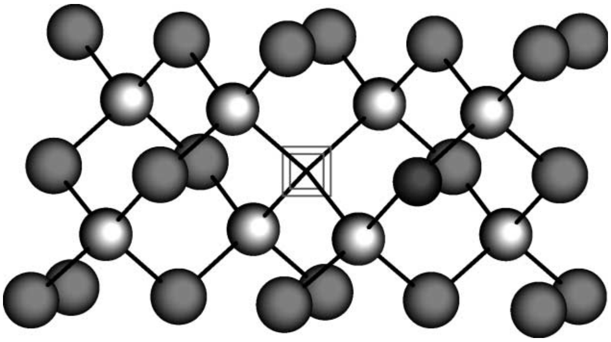

Figure 2 Diagram of the silicon atom–vacancy complex in GaAs. Gallium and arsenic atoms are shown in mid-gray and white, respectively, while the silicon atom is shown in dark gray.

A similar complex also forms in Si doped GaAs: $[\mathrm{Si_{Ga}V_{Ga}}]^-$ (Figure 2), although here both species exist on the same sublattice, and consequently occur as next nearest neighbors in the complex.

In the description above it is assumed that all atoms which are located on a particular site are centered on that site. However, it is possible for a dopant atom to replace an atom of the host semiconductor, but not occupy the lattice site. *Ab initio* calculations of the bonding around the dopant atom suggest that it can be displaced away from one of its nearest neighbors towards the other three atoms (Figure 3). The dopant can be considered as having one bond weaker than the other three.

In the case of n- and p-type doped semiconductors the corresponding defects are known as DX and AX centers respectively. DX centers occur in many semiconductors but are particularly important in AlGaAs, where Si (or other n-type dopant) atoms form DX centers that partially compensate the material and reduce the electron mobility. However, irradiation with white light excites the DX center and converts it back to a normal dopant centered on the lattice site. In AlGaAs the DX center is the stable species and the dopant atom is a higher energy state. Conversion back to the DX center requires surmounting an energy barrier and thus the normal dopant is metastable. The irradiated semiconductor has a lower concentration of DX and a higher doping concentration than before, and consequently has a higher conductivity than the unirradiated material. This results in persistent photoconductivity in the irradiated material, which is a characteristic feature of semiconductors containing DX centers. At low temperatures, this effect can persist for many hours.

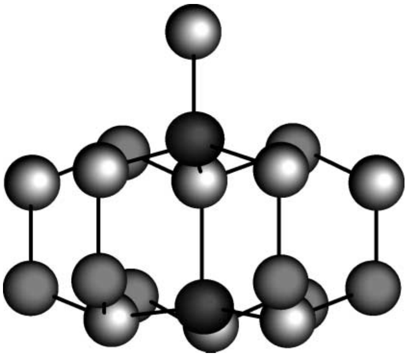

Figure 3 A GaAs crystal shown with [111] oriented vertically. Two Ga atoms have been substituted by silicon atoms which are shown in dark gray. One Si atom is on a substitutional site (top atom) and the other is displaced along [111] away from one of its nearest neighbors (bottom atom). This is one of the structures suggested for the DX center.

1D Defects

Within any crystal, there must exist a set of planes where the interplanar bonding is weakest. In both the diamond and zincblende crystal structures these are the {111} planes. When a crystal is subjected to a shear stress above its elastic limit, then the material accommodates that stress by changing shape, alternately breaking and reforming bonds across these planes where the bonding is weakest. The resulting movement is easily observable under a microscope in many materials and is known as slip or glide; the planes on which this happens are known as slip or glide planes.

Movement on a slip plane does not happen over the whole crystal at once, and there is always a boundary between the material in which slip has occurred and material where it has not. This boundary is the dislocation. A section through a piece of material where slip has occurred is shown in Figure 4. The curved line DD′ marks the dislocation and slip has occurred in the region marked DQD′. Because it marks a boundary, the dislocation may not end within the crystal and the only permitted ways that a dislocation may terminate are to finish at the surface of the crystal (as shown) or to meet another dislocation, forming either a closed loop or a junction.

Across the slip plane, the material has moved by a certain amount, which is, typically, a lattice vector within the crystal. The direction of slip is known as the Burgers vector of the dislocation, and it is a constant for any one dislocation. The Burgers vector at any point may be determined by describing a simple loop in the crystal passing through adjacent lattice points lying in one plane. If a dislocation lies within the loop then the ends will not close, and the displacement needed to close the loop is the Burgers vector. This can be seen in Figure 4 on the left-hand face of the crystal around point D, where the Burgers vector is parallel to the line of the dislocation. It can also be seen on the right hand face of the crystal at point D′, where the Burgers vector, which has not changed direction, now lies perpendicular to the dislocation line. Dislocations where the Burgers vector is parallel or perpendicular to the dislocation line are known as screw and edge dislocations, respectively.

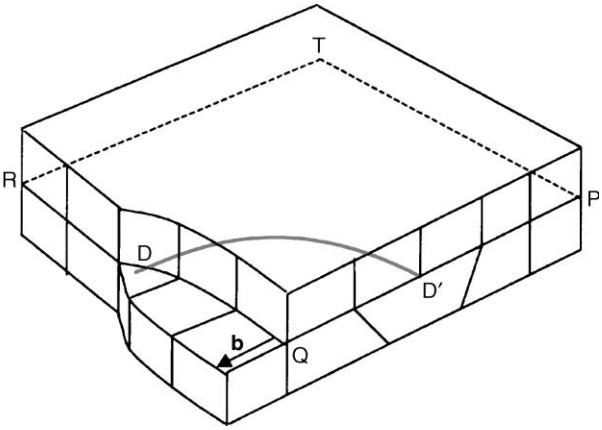

Figure 4 Diagram of a crystal showing a region where slip has occurred. For clarity only the lattice points are shown. The slip plane is PQRT and the slipped region is DQD′. The curve DD′ is the line of the dislocation and has a Burgers vector **b**.

Examples of edge and screw dislocations are shown in Figures 5 and 6, respectively. Edge and screw dislocations are considered to be 'pure' types of dislocations. If the Burgers vector makes any other angle to the dislocation line, the dislocation is referred to as a mixed dislocation. In semiconductors mixed dislocations are the most common type. There are many different ways of describing dislocations; however, the most common one is to give the angle between the dislocation line and the Burgers vector. In this system screw and edge dislocations are referred to as 0° and 90° dislocations.

At the core of the dislocation there is a small region, typically only one or two atoms wide, where the atoms have highly distorted bonding. Around this

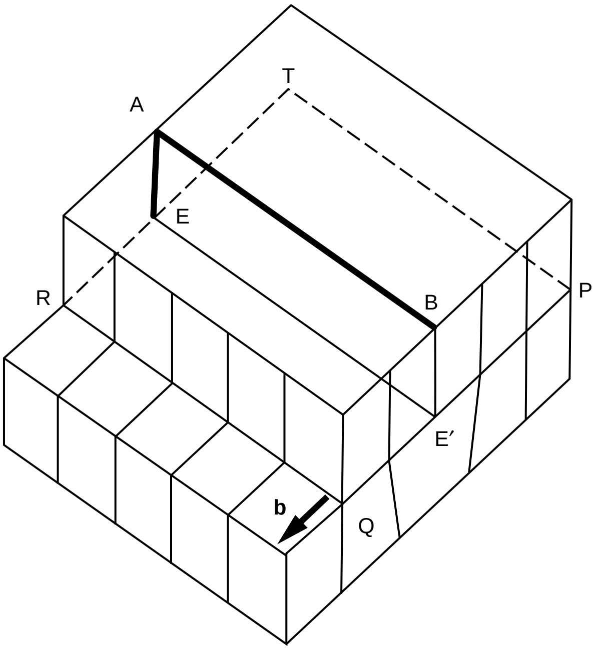

Figure 5 A schematic diagram of an edge dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the edge dislocation is EE′, which can be thought of as terminating a half-plane of atoms denoted ABEE′.

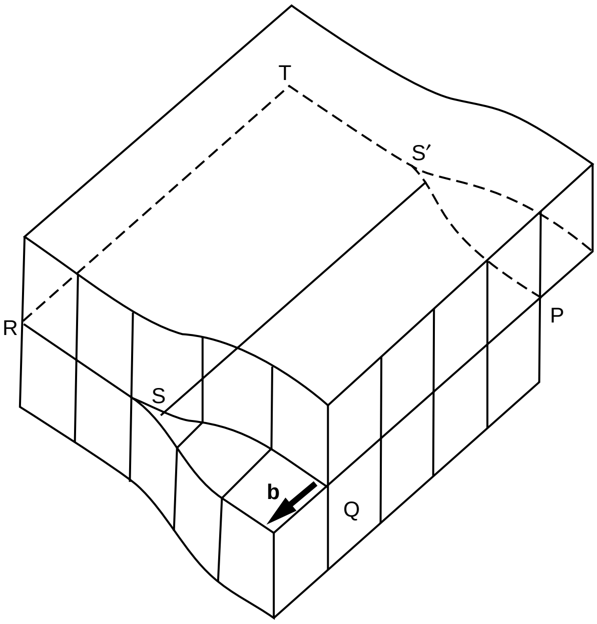

Figure 6 A schematic diagram of a screw dislocation. Only the lattice points are shown for clarity. The slip plane is PQRT and contains the Burgers vector denoted **b**. The line of the screw dislocation is SS′.

core the atoms adopt the normal (tetrahedral) bonding, although they are strained for many nanometers around the dislocation. The strain field around the dislocation means that low dislocation densities are very easy to detect using a technique such as X-ray topography, which images the strain introduced locally in a set of crystal planes by a dislocation. Using this technique, individual dislocations may be determined within crystals. Currently, it is possible to obtain silicon crystals which are essentially dislocation free, while the best GaAs crystals have dislocation densities of $<10^5$–$10^6\ \mathrm{m}^{-2}$. In other materials the densities may be 100–1000 times higher.

The highly distorted bonding at the core of the dislocation inhibits dislocation motion. Under the action of an external stress, dislocations normally move on their glide plane to relieve the stress and deform the material plastically. This is dislocation glide. In covalent semiconductors, such movement means breaking and reforming strong bonds, and does not occur in silicon below the crystal growth temperature. Dislocations are also immobile in GaAs at room temperature, although they can be induced to move in softer, more ionic II–VI semiconductors. Dislocation densities are therefore usually considered 'frozen in' at room temperature.

In addition to macroscopic shear forces, each dislocation experiences the stresses introduced by other dislocations and point defects present in the crystal, and is either attracted or repelled by them. At high densities, therefore, the motion of dislocations is inhibited by the presence of the surrounding defects – the dislocations are said to be pinned.

Generation and movement of dislocations occurs in most semiconductors easily at their growth temperature, and is important in the relief of strain introduced in growing a material on a substrate with a different lattice constant. Edge dislocations always introduce, or remove, extra planes of atoms (Figure 5), and so an array of dislocations can add or remove atoms to accommodate the strain between two semiconductor layers of different lattice constant. These are known as mismatch dislocations and they occur frequently in epitaxial growth when strain relief has occurred.

Typically, growth in MBE or MOCVD occurs on the (001) surface. Mismatch dislocations lie on the $\{111\}$ planes, which intersect this plane, and at low densities occur in crystals regularly spaced along $[110]$ and $[1\bar{1}0]$ directions. The Burgers vectors are also of $\langle 110\rangle$ type and are typically at 60° to the dislocation line. The edge component of this vector, projected on to the (001) plane, gives the amount of strain relief to be expected.

Despite the distorted bonding that exists around the dislocation core, there are thought to be remarkably few broken bonds there. An exception to this is thought to be when the dislocation line is not straight, but kinked, the atoms at the kinks having broken bonds. As a result, nonradiative transitions involving dislocations are not strong. They are important, however, during operation of lasers and LEDs. Dislocations can be seen in electroluminescence of light-emitting devices as dark lines. These so-called dark-line defects (DLDs) are the result of nonradiative recombination occurring at dislocations. Not only do they reduce the efficiency of the device, but the energy released by the recombination of the electron and hole is deposited at the dislocation core as thermal energy. Under these conditions, dislocation movement becomes easy and DLDs propagate, eventually destroying the device.

2D Defects

In a simple description of metallic structures, the hexagonal close packed structure is described as a sequence of hexagonal sheets with the stacking pattern *ABABABAB*..., while that of the face centered cubic structure is *ABCABCA*... In a similar manner, we can describe the wurtzite crystal structure by the stacking sequence *ABABABAB*..., although the layers are no longer close packed and the stacking units comprise one metal and one nonmetal atom. The corresponding stacking sequence for the zincblende structure then becomes *ABCABCA*... These structures are shown in Figures 7a and b, respectively. It can be seen that the two atom units are oriented perpendicular to the stacking sequence, and it is possible to rotate the structure about the vertically oriented bonds between any two given planes. A rotation of 180° would place all the atoms back in registry but on the wrong site. For example, in a zincblende crystal, a rotation after the third layer in the sequence *ABCABCA*... would result in *ABCBCAB*... This means that the top part of the crystal is related to the bottom half by a rotation of 180°, and the two parts are said to be twinned. The twin plane extends throughout the crystal, and is a two-dimensional defect. Both above and below the twin plane, the stacking sequence is the usual *ABC*... found in the zincblende structure; however, at the junction there is a region denoted *BCB*. This region is three planes of the wurtzite stacking sequence, and is an example of what is called a stacking fault. In zincblende crystals, stacking faults introduce small regions of the wurtzite structure, while the reverse happens in wurtzite crystals.

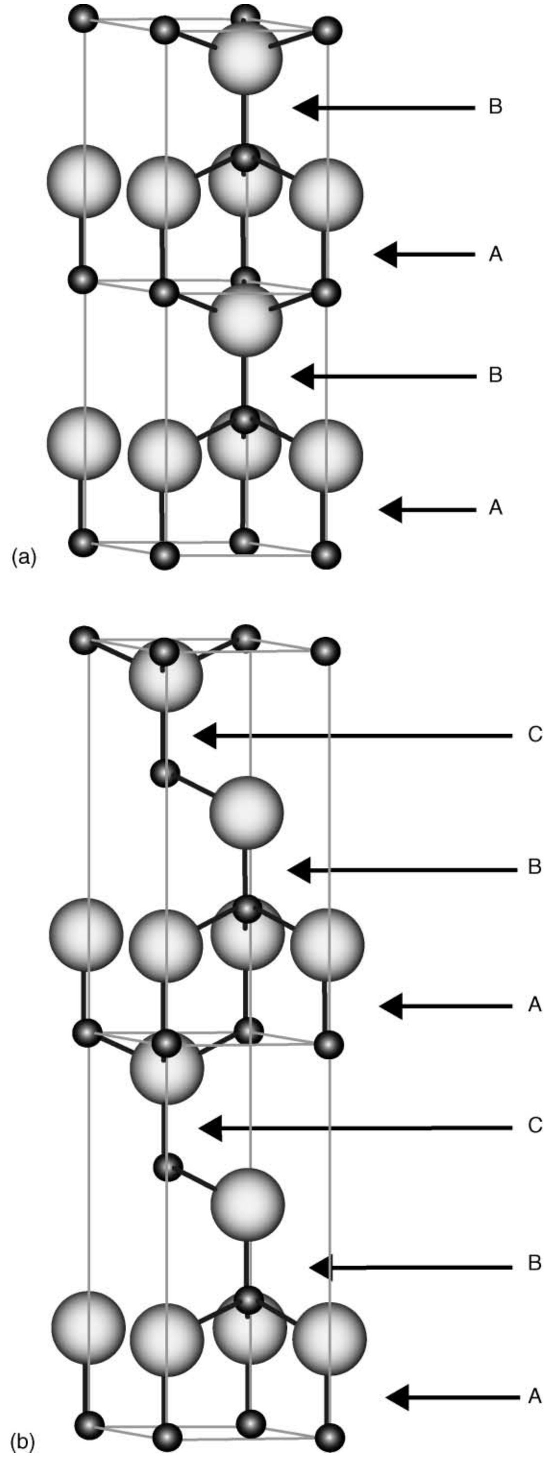

Figure 7 Diagrams of the stacking sequence of layers for two different semiconductor structures. In both cases the metal atoms are small and dark while the larger pale ones are the nonmetal atoms. The structures are: (a). The wurtzite crystal structure. The atoms which occupy the A, B and C sites are marked. Here the vertical direction in which the planes stack is the [0001] direction. (b). The zincblende crystal structure. The atoms which occupy the A and B sites are marked and the vertical direction corresponds to one of the ⟨111⟩ directions in the crystal.

The density of stacking faults in a material is primarily a function of the ease of placing the crystal planes in the wrong order. Under nonequilibrium growth conditions, this can occur easily for many different materials, but for slow growth near equilibrium, the prevalence of stacking faults is governed by the stacking fault energy. This is a measure of the wrong stacking sequence occurring and is also related to the energy of transformation between the zincblende and wurtzite crystal structures. For GaAs, the energy is high and stacking faults are rare, while for II–VI semiconductors the stacking fault energies are low and stacking faults are relatively common.

Typically, stacking faults are observed in TEM planar view micrographs, and on (001) oriented epitaxial samples are clearly visible as triangular regions lying on {111} planes, as described above, nucleated at the interface between the substrate and epitaxial layer.

While stacking faults are accidental introductions in a crystal structure, other 2D defects are introduced deliberately. Heterojunctions between different semiconductor alloys are an obvious example and are responsible for most of the useful properties of advanced semiconductor devices. However, by far the most important 2D crystal defect is the surface of the crystal. This creates a high density of broken bonds, each of which can introduce a defect level in the band structure. Surfaces usually lower their free energy by moving atoms from the bulk equilibrium positions to new sites and forming extra bonds in the plane of the surface. This removes some of the defect levels, but it is still possible to have surface defect densities as large as $10^{18}\,\mathrm{m}^{-2}$ which are associated with enhanced rates of nonradiative recombination of electrons and holes.

Interactions of Defects

Earlier it was assumed that the Burgers vector associated with a dislocation is a lattice vector. The energy of a dislocation per unit length is proportional to the magnitude of the Burgers vector squared:

$$E \propto b^2$$

In some crystals, however, it is possible for the displacement to be composed of two smaller steps, neither of which is a complete lattice vector:

$$\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$$

where $\mathbf{b}$ is a lattice vector and $\mathbf{b}_1$ and $\mathbf{b}_2$ are the smaller displacements. In this case, the total energy

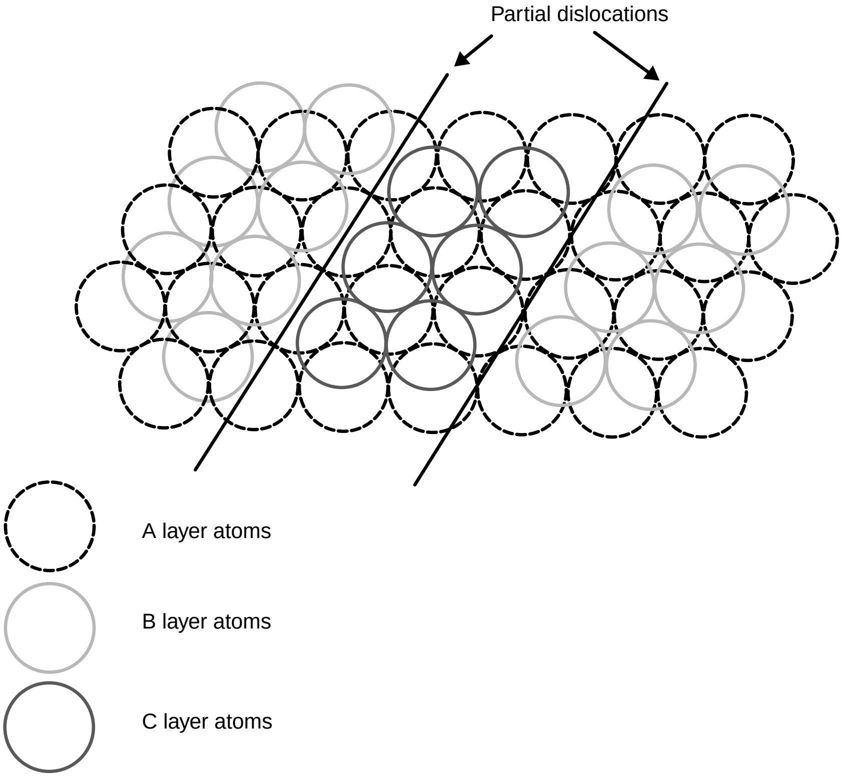

Figure 8 An example of a close packed structure where a region with a stacking fault lies between two partial dislocations.

can be reduced if the dislocation splits into two:

$$E_1 + E_2 \propto b_1^2 + b_2^2 < E$$

These dislocations are known as *partial* dislocations. A normal dislocation splits into two partial dislocations if it is energetically favorable to do so. Figure 8 shows an example in a close packed structure, which illustrates the point without the problem of introducing two different types of atoms. Two partial dislocations run through the top layer. The material lying between them has a different stacking sequence (AC) from the material around it (AB) and must contain a stacking fault. This shows that a 2D defect, such as a stacking fault, is bounded by two 1D defects, in this case partial dislocations.

There also exists an intimate relationship between point defects and dislocations, which may be seen by examining the edge dislocation shown in Figure 5. The edge dislocation may be considered as an extra half-plane of atoms, ABEE′ which terminates along line EE′. The last row of atoms in the plane can be removed from the sheet by one of two processes. Firstly, the atoms at the dislocation core can become interstitials and diffuse away from the dislocation. Alternatively, vacancies can diffuse towards the dislocation. In either case the result is the same; the dislocation line EE′ moves upwards towards AB. The opposite sequence of events (diffusion of interstitials towards the dislocation or vacancies away from it) is also possible and results in a movement of line EE′ downwards. Such processes are known as dislocation climb. Climb is not as common as glide, but it is important as the dislocation acts as a source or sink for point defects. This process contributes to the reduction in radiative efficiency in LEDs and lasers, particularly as the devices age.

See also

Semiconductor Physics: Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Beanland R, Kiely CJ and Pond RC (1992) Dislocations in heteroepitaxial films. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1149–1225. Amsterdam: North-Holland.

Fukuda M (1991) Reliability and Degradation of Semiconductor Lasers and LEDs. Boston, MA: Artech House.

Hirth JP and Lothe J (1982) Theory of Dislocations. New York: Wiley.

Hull D (1965) Introduction to Dislocations. Oxford, UK: Pergamon Press.

Hurle DTJ (1999) A comprehensive thermodynamic analysis of native point defect and dopant solubilities in gallium arsenide. *Journal of Applied Physics* 85: 6957–7022.

Kelly A, Groves GW and Kidd P (2000) Crystallography and Crystal Defects, revised edn. Chichester, UK: Wiley.

Kossevich AM (1999) The Crystal Lattice. Berlin: Wiley-VCH.

Malloy KJ and Khachaturyan K (1993) DX and related defects in semiconductors. In: Willardson RK and Beer AC (eds) *Semiconductors and Semimetals*, pp. 235–291. New York: Academic Press.

Schubert EF (1993) Doping in III–V semiconductors. In: Ahmed H (ed.) *Cambridge Studies in Semiconductor Physics and Microelectronic Engineering*, vol. 1. Cambridge, UK: Cambridge University Press.

Skowronski M (1992) Impurity- and defect-centres in compound semiconductors. In: Mahajan S (ed.) *Handbook on Semiconductors. Materials, Properties and Preparation*, vol. 3a, pp. 1343–1398. Amsterdam: North-Holland.

Infrared Lattice Properties

T J Parker and S R P Smith, University of Essex, Colchester, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The vibrations of atoms in solids were first studied theoretically at the beginning of the twentieth century to explain the observed temperature dependence of the heat capacity. Following earlier work by Einstein and Debye, Born and von Karman showed in 1912 that the vibrational properties of a linear chain of alternate light and heavy atoms connected by springs account very well for the observed thermal behavior of solids. It was shown many years later that, if the atoms are alternately positively and negatively charged, the model also gives a good description of the optical properties.

In this article the vibrational properties of a linear diatomic chain of atoms are first described. The model gives a good description of the basic vibrational properties of solids with two atoms per unit cell, such as Si, Ge, and the binary III–V (e.g., GaAs) and II–VI (e.g. CdTe) semiconductors, and explains the origin of optic and acoustic branches of the phonon dispersion curves. This leads naturally to an understanding of the basic far-infrared reflectivity of binary semiconductors. The model is then further developed to account for the effects of impurities and the optical properties of alloys.

The far-infrared properties of semiconductors are then reviewed. We begin with the group IV semiconductors Si and Ge in which there is no first-order interaction with infrared radiation, and explain the origins of second-order interactions, and the effects of disorder and impurities. Examples of the first-order reflection spectra of bulk single crystal binary semiconductors and their alloys are then presented. Localized vibrational modes and phonon combination bands are discussed next. We finish with a brief review of more exotic excitations that occur in layered semiconductors: confined optic modes, surface phonon polaritons, guided waves, and interface modes.

Background Theory of Phonons in Semiconductor Materials

Phonons in a Diatomic Linear Chain Lattice

Consider a line of atoms of alternating type A and B, separated by a distance a, as shown in Figure 1. Suppose that the nth A atom is displaced by a small amount $u_{\mathrm{A},n}$ from its equilibrium position. The forces attempting to return it to its equilibrium position will be proportional to the distance it has moved relative to the neighboring atoms. If we assume that the only significant force involves the nearest neighbor B atoms n and $n+1$, the equation of motion for the nth A atom is

$$M_{\mathrm{A}}\frac{\mathrm{d}^2u_{\mathrm{A},n}}{\mathrm{d}t^2} = C(u_{\mathrm{B},n} + u_{\mathrm{B},n+1} - 2u_{\mathrm{A},n}) \qquad [1]$$

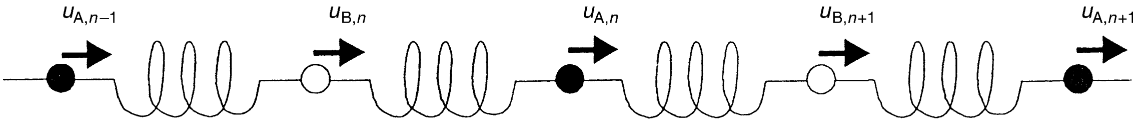

Figure 1 Atomic displacements in a diatomic linear chain.

where M_A is the mass of the atom, and C is the Hooke's law nearest-neighbor force constant. There is a similar equation

$$M_B \frac{d^2 u_{B,n}}{dt^2} = C(u_{A,n-1} + u_{A,n} - 2u_{B,n}) \qquad [2]$$

for the motion of the nth B atom. We assume that the solution of the equations of motion has the wavelike form

$$\begin{aligned} u_{A,n} &= U_A \exp\{i(qan - \omega t\} \\ u_{B,n} &= U_B \exp\{i(qan - \omega t\} \end{aligned} \qquad [3]$$

in which U_A and U_B are amplitude factors, ω is the angular frequency, and q is the wavevector, so that qa is the phase factor by which the phase of the wave changes from one site n to the next site $n+1$. The equations of motion are satisfied so long as the angular frequency ω is given by

$$\omega^2 = \frac{C}{M_A M_B}\Big[M_A + M_B \pm \sqrt{(M_A + M_B)^2 - 4M_A M_B \sin^2(qa/2)}\Big] \qquad [4]$$

There are two possible solutions here because there are two ions in the unit cell. A plot of ω as a function of q is shown in Figure 2a, with q in the range $-\pi/a$ to $+\pi/a$. In the lower branch, the *acoustic* phonon branch, the vibrations at long wavelength (equivalent to small q) are essentially the same as ordinary sound waves. The A and B atoms vibrate in phase, and at long wavelength there is a linear relation between ω and q:

$$\omega \approx v_s q \qquad [5]$$

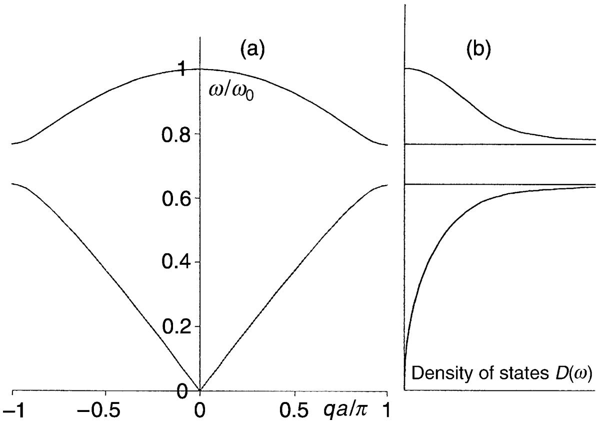

Figure 2 (a) Phonon frequency ω versus wavevector, q, in a one-dimensional diatomic lattice; (b) three-dimensional density of states $D(\omega)$.

where $v_s = \omega_0 a/2$ is the velocity of sound of the acoustic phonons.

The upper branch in Figure 2a, which is of more interest in the present context, is called the *optic* phonon branch. Here the A and B atoms vibrate in antiphase, and because the A and B atoms have equal but opposite charges there is an electric field associated with the vibration. The optic phonon vibrations will consequently affect the propagation of electromagnetic waves through the medium – hence the name – although the vibration frequencies (around 10 THz) are in the far infrared rather than the visible part of the electromagnetic spectrum. At this frequency, the infrared wavelength λ is about 30 μm. The propagation of the infrared will be affected only by phonons whose wavevector q matches the infrared wavevector $k = 2\pi/\lambda$. Since λ is around 10^4 times larger than the typical lattice spacing a in semiconductor materials, it follows that the optical properties of the material may be largely understood merely by considering the long-wavelength phonons with wavevector $q \approx 0$. Thus, in the diatomic case, only the optic phonon of wavevector $q \approx 0$ and frequency

$$\omega_0 = \sqrt{\frac{2C(M_A + M_B)}{M_A M_B}} \qquad [6]$$

is of importance to the propagation of electromagnetic radiation.

Many semiconductor materials are of the diatomic AB type – for example group III–V materials such as GaAs, and II–VI semiconductors such as CdTe. Their phonons are described in detail by three-dimensional versions of the simple linear chain model discussed above. The most basic semiconductor materials, Si and Ge, are clearly monatomic, but their crystal structure is identical to the zincblende structure of GaAs, and they show two phonon bands as in Figure 2a. However, because there is only one type of atom, these materials differ from the diatomic materials in that the optic phonon branch has no electric dipole moment. In Si and Ge, there is no dielectric anomaly around the frequency ω_0 of the type discussed in the following section. In these materials, the only significant optic effects occur when phonons combine in pairs with equal but opposite wavevector q, so that the resultant has very small wavevector that can match the infrared wavevector $k = 2\pi/\lambda$. The significant factor here is the density of states $D(\omega)$, representing the number of phonon states per unit frequency range. In Figure 2b, we show $D(\omega)$ for the diatomic lattice. This is obtained as

$$D(\omega) = D(q)\frac{dq}{d\omega} \qquad [7]$$

where $D(q)$ is the density of states with respect to the wavevector, which is proportional to q^2 for a three-dimensional lattice. Note that $D(\omega)$ has peaks at frequencies corresponding to the zone boundary phonon frequencies because the dispersion curves are flat in these regions. These features are known as critical points, and they have well-defined shapes. Their spectroscopic importance is discussed later.

In bulk three-dimensional crystals, both transverse and longitudinal optic and acoustic phonons occur, each represented by different branches of the phonon dispersion curves, and the transverse branches in cubic materials are doubly degenerate. In three dimensions, more elaborate theoretical models are required, such as the rigid ion model and the shell model, and improved results can sometimes be achieved by including next-nearest-neighbor atoms in the model. As an example, the phonon dispersion curves of GaAs are shown in Figure 3.

It should be noted that the dispersion curves, and hence the associated optical properties, differ for propagation in different crystallographic directions, e.g., [100], [110], and [111].

Born–Huang Optic Phonon Model

The infrared lattice properties of materials at long wavelength are usually described by using the relative permittivity ε (often called the 'dielectric constant'), which is defined through the relation

$$D = \varepsilon\varepsilon_0 E \qquad [8]$$

between the electric field E in a medium and the displacement field D (here, ε_0 is the permittivity of free space). D is the sum of the field E and the polarization P of the medium:

$$D = P + \varepsilon_0 E \qquad [9]$$

P has contributions from lattice displacements (phonons), free electrons (plasmons), and the basic atomic polarizabilities. Since these quantities depend on frequency, ε is also a function $\varepsilon(\omega)$ of angular frequency ω. The charges of the A and B sites are taken as $+Q$ and $-Q$ respectively, and their equations of motion [1] and [2] are extended to include the electric forces $\pm EQ$. As a result of these considerations, the dielectric function can be written in the form

$$\varepsilon(\omega) = \varepsilon_\infty\left[1 + \frac{\omega_L^2 - \omega_T^2}{\omega_T^2 - \omega^2}\right] = \varepsilon_\infty\frac{\omega_L^2 - \omega^2}{\omega_T^2 - \omega^2} \qquad [10]$$

Here, ε_∞ is the high-frequency dielectric constant due to the atomic polarizabilities, and ω_L and ω_T are frequencies that are close to the natural vibrational frequency ω_0 that occurs in the absence of electric field effects; the difference $(\omega_L^2 - \omega_T^2)$ is of order $\omega_0^2 Q^2/(Ca^3\varepsilon_0)$.

Figure 4 shows a plot of $\varepsilon(\omega)$ as a function of ω. Note that $\varepsilon(\omega)$ diverges at $\omega = \omega_T$, and that $\omega = 0$ at $\omega = \omega_L$ ($\omega_L/\omega_T = 1.25$ in Figure 4). ω_T is called the transverse optic (TO) phonon frequency: it is the resonance frequency of propagation of transverse

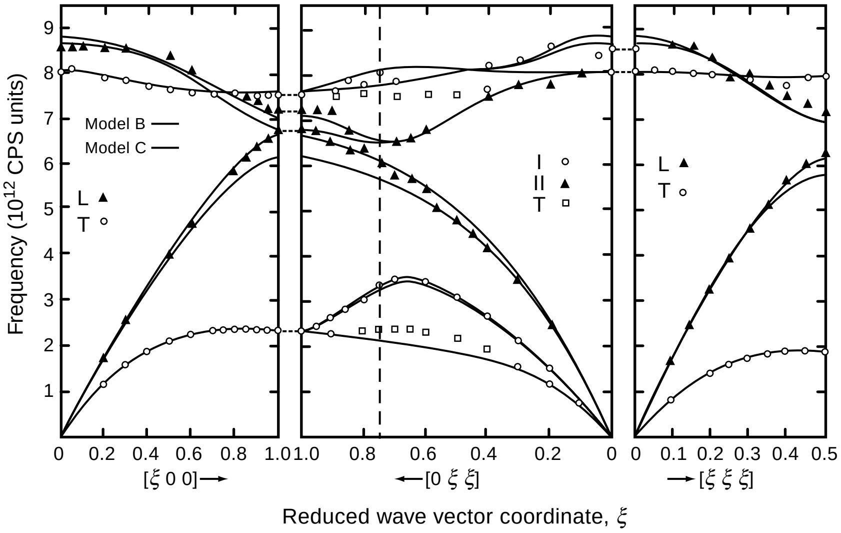

Figure 3 Phonon dispersion curves of GaAs determined by neutron spectrometry (points) compared with calculation (dotted and solid lines). Reproduced with permission from Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

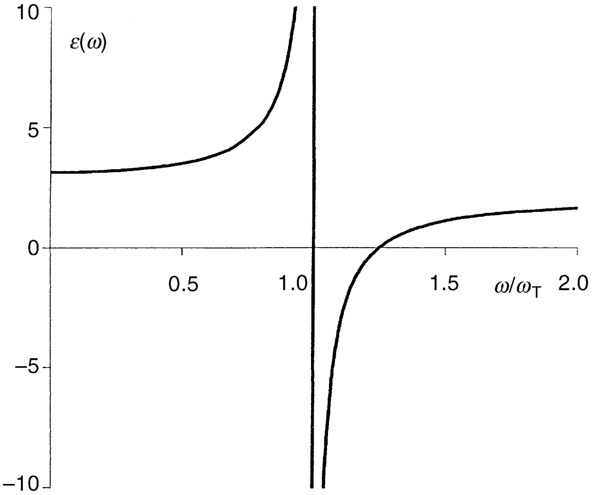

Figure 4 Dielectric function $\varepsilon(\omega)$ as a function of angular frequency, ω.

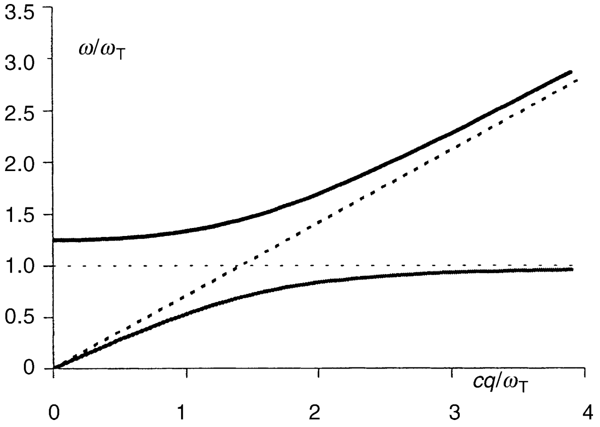

Figure 5 The dispersion curves for phonon polariton modes (solid lines) for a diatomic linear chain of atoms showing the gap in bulk phonon modes between the frequencies ω_T and ω_L.

electromagnetic waves in the material. ω_L is called the longitudinal optic (LO) phonon frequency: it is the frequency of the longitudinal electrostatic wave. Physically, the dielectric constant determines the velocity v of propagation of electromagnetic waves in the material

$$v^2 = \frac{\omega^2}{q^2} = \frac{c^2}{\varepsilon(\omega)} \qquad [11]$$

where c is the velocity of light. Equation [11] has two solutions for ω as a function of wavevector: the two solutions are shown in Figure 5. Note that there is no solution with ω in the range between ω_T and ω_L. The excitations in the region where the graphs are curved are called 'polaritons' – a combination of polar phonon and photon.

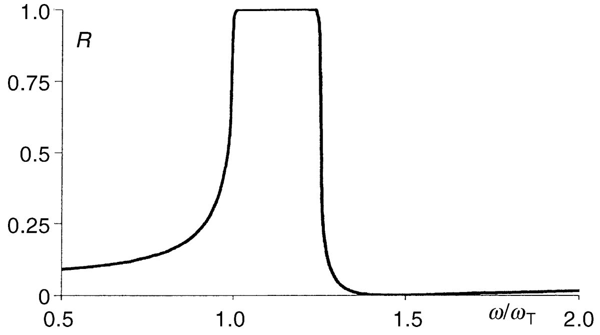

Figure 6 Calculated reflectivity plotted versus angular frequency for a diatomic linear chain of atoms for the case of zero damping.

The refractive index n of the material is the ratio of v, the velocity of light in the medium, to c, the velocity in free space; thus $n = \sqrt{\varepsilon}$. n is often found from a measurement of the optical reflectivity R. At normal incidence

$$R = \left| \frac{1 - n}{1 + n} \right|^2 \qquad [12]$$

A graph of R is shown in Figure 6; note that the reflectivity is 100% for ω in the range $\omega_T < \omega < \omega_L$.

Impurity Phonon Modes

Impurity atoms in crystal lattices can significantly affect the infrared properties. The simplest case is when an impurity atom has the same binding but a different mass. Consider the monatomic, one-dimensional case, in which the impurity atom has a mass $M + \delta M$, where M is the normal mass. Let the impurity be at the site $n = 0$. The equations of motion are

$$\begin{aligned} -M\omega^2 u_n &= C(u_{n-1} + u_{n+1} - 2u_n) \;\; \text{for } n \neq 0 \\ -(M + \delta M)\omega^2 u_0 &= C(u_{-1} + u_1 - 2u_0) \;\; \text{for } n = 0 \end{aligned} \qquad [13]$$

These equations have a solution provided that ω satisfies the equation

$$N^{-1} \frac{\delta M}{M} \sum_q \frac{\omega^2}{\omega_q^2 - \omega^2} = 1 \qquad [14]$$

where N is the number of atoms in the linear chain, and ω_q the unperturbed normal mode frequencies. This equation has solutions ω that lie close to the bulk phonon frequencies ω_q. However, if $\delta M < 0$ (i.e. the impurity atom is lighter than the bulk atoms), one additional mode is possible at a frequency ω that is higher than the maximum frequency $\omega_0 = \sqrt{C/M}$ of the bulk phonons. This is the so-called impurity or localized vibrational mode (LVM – see Figure 7).

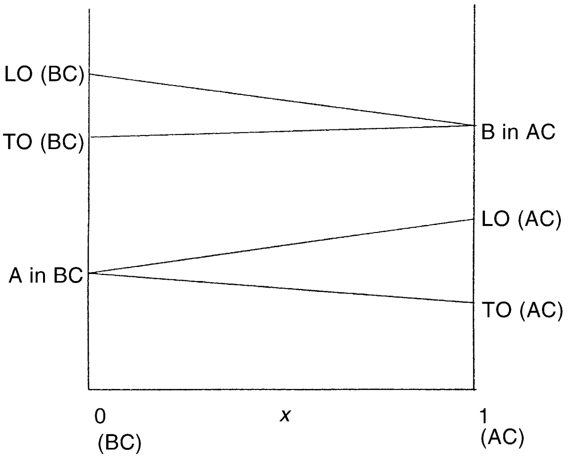

Figure 7 Representation of TO and LO phonon frequencies in a mixed crystal $A_xB_{1-x}C$ plotted versus composition x.

The impurity atom, being lighter than the material atoms, vibrates at a higher frequency than the bulk frequencies; the amplitude of the vibration is greatest at the impurity site, and decreases the further away one goes from the impurity site. If the impurity atom is heavier than the bulk atoms ($\delta M > 0$), there is no distinguishable local mode. In the case of diatomic lattices, an analogous situation pertains: if the impurity is lighter than the bulk atoms, an impurity mode is distinguishable above the top of the optic phonon band; if the impurity is heavier, there is the possibility that a distinct impurity mode may appear in the gap between the acoustic and optic phonon branches.

Mixed Crystals

Mixed compound semiconductor crystals are in widespread use because the bandgap can be adjusted by varying the composition. For example, the ternary material $A_xB_{1-x}C$ is an alloy of a fraction x of the material AC with a fraction $1 - x$ of the material BC. A typical example is $Al_xGa_{1-x}As$. The infrared lattice properties of these materials are classified as 'one-mode' or 'two-mode' depending on how many distinct optic phonon frequencies are observed. The distinction can easily be understood in the context of the discussion above of impurity modes. If an A atom in the bulk material BC has a distinct local impurity optic phonon mode frequency, and similarly for a B impurity in AC, then there will usually be two distinct optic phonon modes throughout the whole composition range $0 < x < 1$: one mode is AC-like, and the other BC-like, as shown in Figure 7. On the other hand, if neither A in BC, nor B in AC, provide distinct optic phonons, there is only one distinguishable optic phonon in the alloy $A_xB_{1-x}C$.

Review of Far-Infrared Properties of Semiconductors

Group IV Semiconductors

We begin with Si and Ge, both of which crystallize in the diamond structure. As mentioned earlier, there is no first-order interaction between far infrared radiation and the TO mode. However, there are several other mechanisms that result in absorption. The dipole moment of the zone center TO phonon in an ionic solid is an example of a symmetry property that, in group theory, is described by a set of selection rules for each different crystal lattice. Other mechanisms that can lead to the creation of a dipole moment are:

Disorder

Disorder destroys the translational symmetry, and hence the selection rules. Consequently, phonon modes that, in a perfect lattice, are not infrared active will become active. The most general case occurs if the crystal becomes completely disordered, in which case all phonon modes become dipole-active and the far-infrared absorption spectrum resembles the phonon density of states. This is illustrated in Figure 8, in

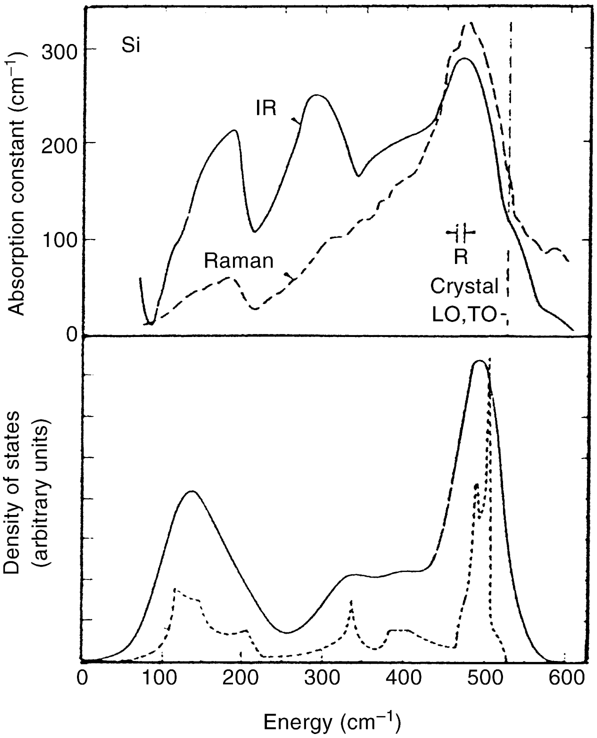

Figure 8 Top: infrared absorption coefficient (solid line) and Raman spectrum (dashed line) of amorphous Si. Bottom: calculated density of states (solid line), and broadened density of states (dashed line). Reproduced with permission from Bradskyn H and Lunio (1974) Infrared vibrational spectra of amorphous Si and Ge. *Physical Review B* 9: 1646, ADS.

which measured infrared and Raman spectra of amorphous Si are compared with the calculated one-phonon density of states.

Phonon combination bands

A combination of the lattice distortions due to two phonons excited simultaneously can produce a resultant displacement of the ions in the lattice with a dipole moment. Such phonon combinations can be excited simultaneously by a far-infrared photon, and they are observed as weak bands in the infrared absorption spectrum. Apart from the dipole moment selection rule, it is also necessary for wavevector conservation to occur. Since $q \sim 0$ for the photon, the two phonons must either be at the zone center, or have equal and opposite wavevectors. An additional consideration is that, if the process is to lead to observable absorption, it should have a high probability as determined by the density of states, so observable absorption features usually arise from critical-point phonon combinations. An example of phonon combination bands in Si is shown in Figure 9. Assignment of the features in the spectrum is usually possible by inspection of the phonon dispersion curves at the zone boundaries, a procedure known as critical-point analysis. Alternatively, the two-phonon density of states can be calculated, taking into account the selection rules, and then compared with the measured spectrum. Similar considerations apply to higher-order phonon combination bands, e.g., three or four phonons, etc., but the strength of the combination bands usually decreases rapidly as the order of interaction increases.

The presence of impurities

As mentioned above, dipole-active localized vibrational modes (LVMs) can occur if the sample is doped with lighter elements as this destroys the local translational symmetry. For instance, an LVM occurs if Si is doped with B, as shown in Figure 10. The LVM frequency is proportional to the ratio $(C/M)^{1/2}$ as seen earlier where C is the force constant and M is the isotopic mass of the substituted atom. As such modes have narrow linewidths, the features in the spectrum due to different isotopes of the same element are easily distinguished, as shown in Figure 10.

Group III–V and II–VI Semiconductors: Bulk Crystals

The group III–V and group II–VI semiconductors crystallize in either the cubic (zincblende) structure, or the hexagonal (wurtzite) structure. Many mixed crystal materials, such as $Al_xGa_{1-x}As$, are grown for

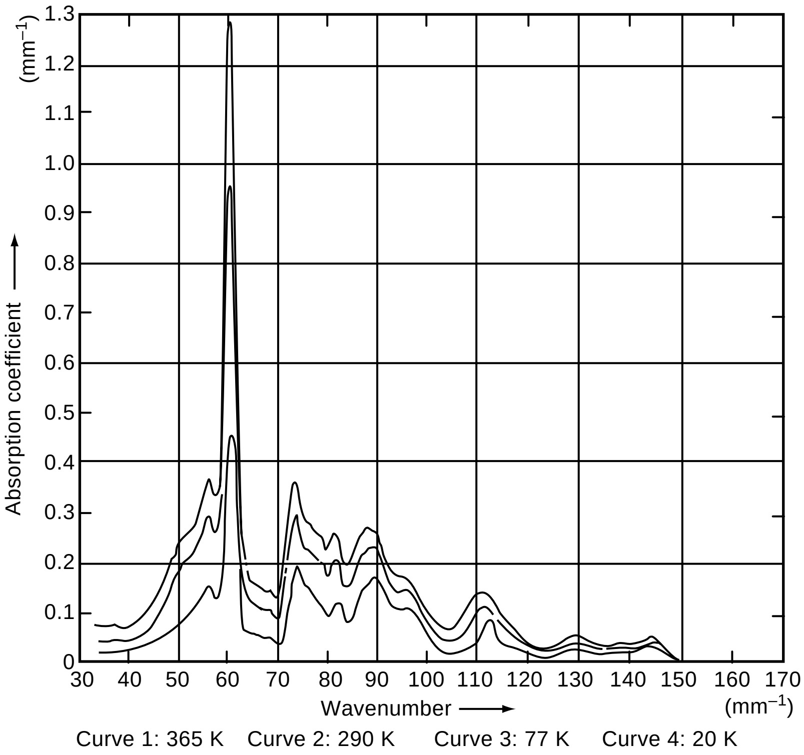

Figure 9 Measured lattice combination bands in a single crystal of silicon at 290 K (top), 77 K (middle), and 20 K (bottom). (Note that 100 mm^{-1} = 1000 cm^{-1}.) Reproduced with permission from Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

electronic applications that require materials with bandgaps not found naturally. Furthermore, a host of low-dimensional structures, such as epilayers, multiple quantum wells (MQWs) and superlattices, are fabricated by molecular beam epitaxy (MBE) or metalorganic vapor phase epitaxy (MOVPE) to create devices that exploit the novel properties which occur when one or more of the physical dimensions are small.

Reflectivity from a bulk single crystal

In the reststrahl frequency range bulk phonon modes cannot propagate and incident radiation is perfectly reflected, i.e., $R = 1$, as shown in the calculated spectrum in Figure 6. Figures 4 and 6 represent the ideal case of a harmonic crystal, in which there is no interaction between the phonon modes, so that all modes have infinite lifetime or, equivalently, spectral features have zero linewidth, corresponding to zero damping. This is close to the behavior of a crystal at absolute zero but, in practice, the phonon modes are coupled, so each mode will decay into combinations of other phonon modes. This leads to finite phonon lifetimes and finite linewidths of all features in the measured spectra. This damping has the effect of

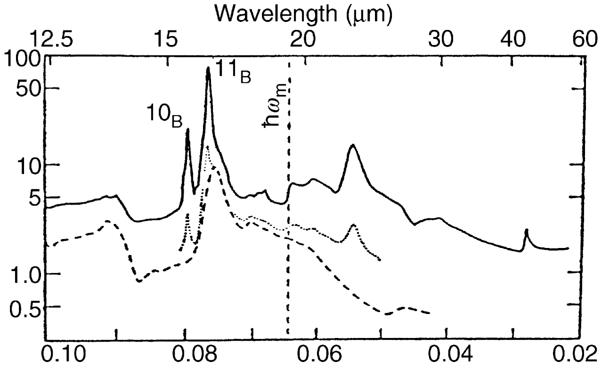

Figure 10 The far-infrared absorption spectrum of Si doped with B (^{10}B + ^{11}B). Reproduced with permission from Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

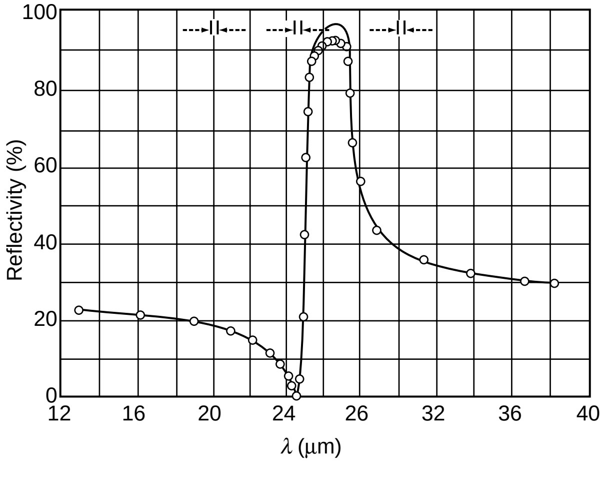

Figure 11 The reststrahlen band reflectivity of GaP, measured (points) and calculated from a damped simple harmonic oscillator (line). Reproduced with permission from Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

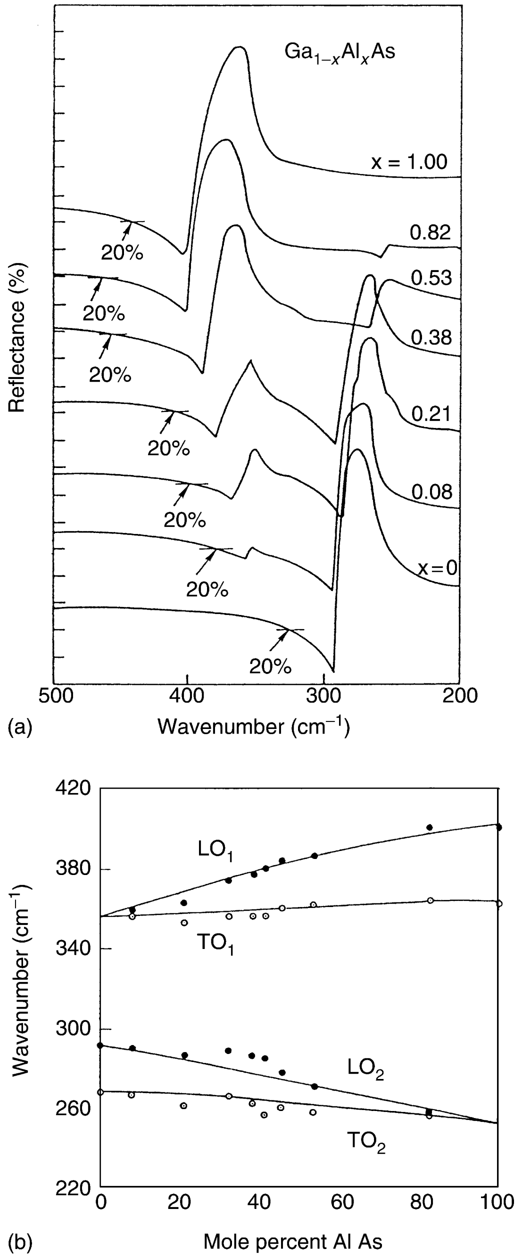

Figure 12 (a) Reflection spectra for GaAs, AlAs, and five Ga$_{1-x}$Al$_x$As mixed crystals. The curves have been displaced for clarity. (b) TO and LO mode frequencies obtained analytically from the spectra, plotted versus alloy composition. Reproduced with permission from Ilegens M and Pearson GL (1970) Infrared reflection spectra of Ga$_{1-x}$AL$_x$As mixed crystals. *Physical Review B* 1: 1576.

rounding off features in the reflectivity compared with Figure 6, and it can be represented analytically by the equation:

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{(\varepsilon(0) - \varepsilon_\infty)\omega_T^2}{\omega_T^2 - \omega^2 - i\gamma\omega} \qquad [15]$$

This differs from eqn [10] only by the addition of the damping coefficient, γ, which takes account in a phenomenological way of the interactions between the phonon modes. Here $\varepsilon(0)$ is the dielectric function at zero frequency, and $\varepsilon(0) - \varepsilon_\infty$ is the oscillator strength (i.e., the dipole strength) of the mode. We shall consider γ again later.

The reflectivity of GaP is shown in Figure 11, and this should be compared with Figure 6. In pure diatomic crystals like the simple ionic solids, e.g., NaCl and KCl, or the binary semiconductors, e.g., GaAs and CdTe, this is the only region of high reflectivity at frequencies below the visible or ultraviolet, and it is often described as the reststrahlen band, from the German term for residual ray.

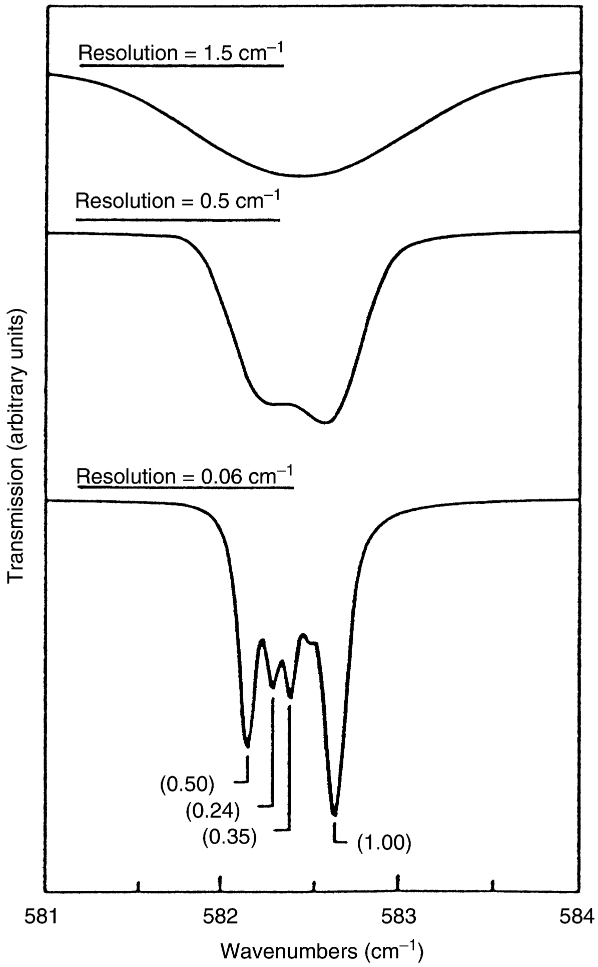

Figure 13 Localized vibrational mode (LVM) absorption band for ^{12}C in GaAs measured at 80 K. The numbers in brackets are estimated band strengths. Reproduced with permission from Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

Reflectivity from mixed crystals of binary semiconductors

The spectra of mixed crystals were discussed above and an example is shown in Figure 12a, which shows the reflection spectra of GaAs, AlAs and five $Ga_{1-x}Al_xAs$ mixed crystals. The TO and LO phonon frequencies were obtained from the spectra using a Kramers–Kronig analysis and classical dispersion theory. Changes in the mode frequencies with composition are interpreted using the so-called random element isodisplacement (REI) model. In Figure 12b, the TO and LO frequencies are plotted versus composition, x, and it can be seen that this system displays typical two-mode behavior, following closely the behavior shown in Figure 7.

Localized vibrational modes due to impurities

LVMs have been widely investigated in binary semiconductors because they offer a useful diagnostic tool for the investigation of impurities. In addition to the identification of isotopic species, high-resolution transmission spectroscopy can be used to determine the sites of the impurity atoms in the lattice. In Figure 13 the LVM absorption band for ^{12}C in GaAs is shown at different resolutions. The ^{12}C atoms are substitutional impurities on the As sites and the different components of the band arise from the different nearest-neighbor isotopic arrangements

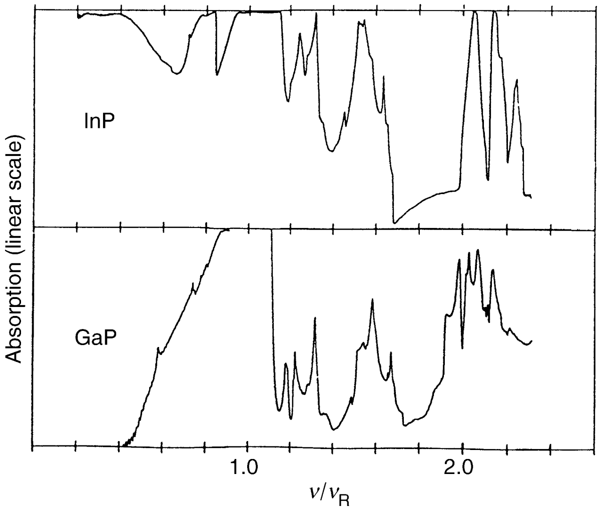

Figure 14 High-resolution far-infrared absorption spectra of GaP and InP with the frequency scale normalized in each case to the reststrahl frequency ($\nu/\nu_R = 1$ corresponds to $\omega = \omega_T$). Reproduced with permission from Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

surrounding the different ^{12}C atoms in the lattice. Note that ^{12}C is much lighter than either of the host lattice atoms, Ga or As, as required theoretically.

Phonon combination bands

Weak features due to the simultaneous excitation of two or more phonons are observed in the spectra of all semiconductors, and there are two excitation mechanisms. The first is direct excitation of two or more phonons by an incident photon via nonlinear terms in the dipole moment. The second process is indirect: the incident photon first excites a zone center TO phonon and the TO phonon then decays into a combination of phonons elsewhere in the Brillouin zone, a process described as anharmonic decay of the TO phonon. Mathematically, it is described by a complex frequency-dependent damping coefficient, $\gamma(\omega)$, in the oscillator equation [15]. Clearly, the latter mechanism is not available in Si and Ge as the TO phonons are not dipole-active. Anharmonicity is very important in strongly ionic crystals like the alkali halides (i.e., group I–VII compounds like NaCl), leading to prominent phonon combination bands, but it becomes progressively less important in semiconductors as the interatomic bonding becomes less ionic and more covalent, which is the case in the sequence through group II–VI (e.g., semiconductors like CdTe or ZnSe), group III–V (e.g., GaAs or InP), and group IV (Ge and Si), which are entirely covalent. Nearly all two-phonon combinations are dipole-active in zinc-blende structure crystals with the result that the transmission spectrum either side of the reststrahlen band is very rich in weak phonon combination bands, as shown in Figure 14 for GaP and InP.

Layered Semiconductor Structures

Confined optical phonons

Semiconductor superlattices are fabricated for a variety of electronic and optical applications, and the

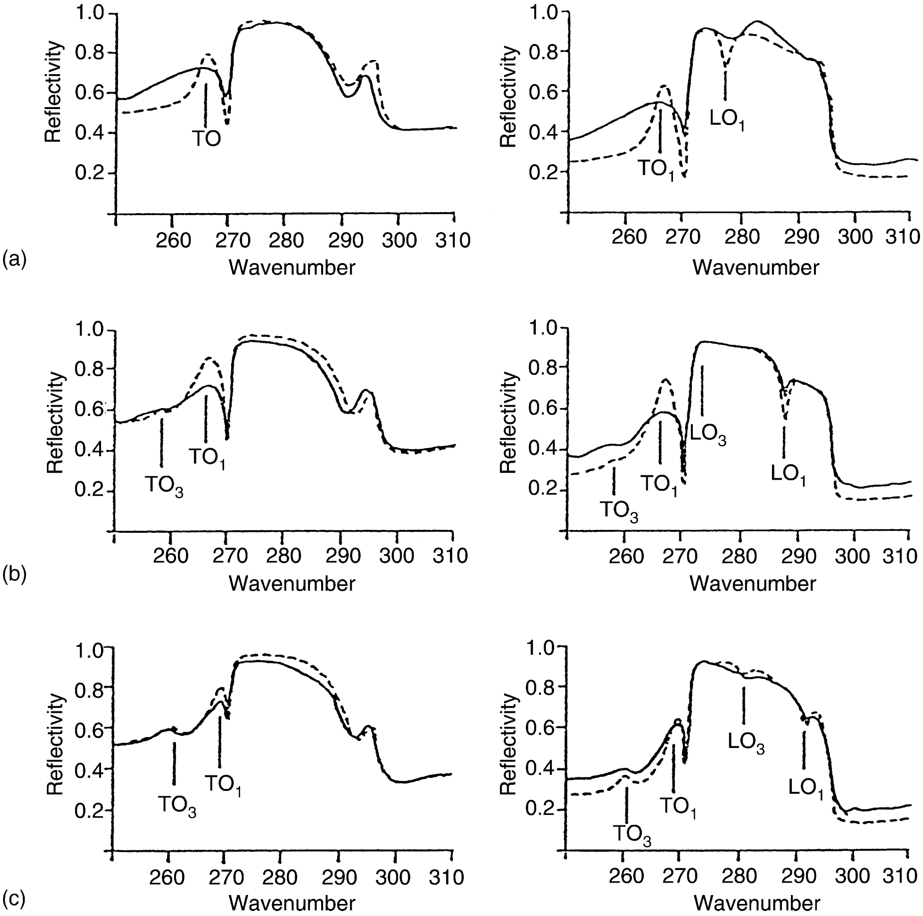

Figure 15 Measured (solid lines) and calculated (dashed lines) oblique incidence (45°) reflectivity spectra at 77 K of three short-period $(GaAs)_n/(AlAs)_n$ superlattices on GaAs substrates: n denotes the number of monolayers. Left curves: s-polarization, right curves: p-polarization. Note that the strong reststrahl reflectivity is from the GaAs substrates and the fine structure is due to confined modes. The subscripts 1, 2, and 3 denote the orders of the confined modes. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

quality of the interfaces in such structures is important as it determines the amount of electronic scattering at the interfaces. Some structures are fabricated from alternate layers of materials such as GaAs and AlAs, in which there is no overlap between the frequencies of the optical branches of the phonon dispersion curves. This has the effect that optical phonons excited in one type of layer, e.g., in GaAs, cannot propagate into the other type of layer, i.e., AlAs, and vice versa, so that the optical phonon modes are confined in their respective layers. The confined modes behave like standing waves, with frequencies determined by the thicknesses of the layers and the bulk optical phonon dispersion curves. Thus, if the superlattice structure is of poor quality, and the interfaces are broad due to alloying, rather than abrupt, there will be a change in the effective widths of the layers. This leads to small shifts in the frequencies of the confined phonons which can be measured and used to obtain a quantitative estimate of the amount of interface broadening. An example is shown in Figure 15. The degree of interface broadening determined from these spectra was 1.4 atomic spacings, in agreement with estimates using other techniques.

Surface phonon polaritons and guided waves

The reststrahl reflection band (Figure 6) occurs because there is a gap between ω_T and ω_L in the polariton dispersion curves (Figures 4 and 5), with no solutions for bulk phonon modes. Surface phonon polaritons can, however, propagate in this frequency band, but they have wavevectors that are larger than those, ω/c, of far-infrared radiation, so special techniques are required to observe them. The necessary wavevector enhancement can be obtained in two ways. Prism coupling can be used, as shown in Figure 16. When the angle of incidence of the incident beam at the base of the prism exceeds the critical angle for total internal reflection an evanescent wave with advanced wave vector, $q > \omega/c$, is set up at the base of the prism. This wave can couple to a surface polariton, and the energy used to excite the polariton results in a dip in the reflectivity at the frequency of the surface polariton, hence the name of the technique: attenuated total reflection (ATR). An example is shown in Figure 16. Alternatively, wavevector enhancement can be achieved by placing a grating on the surface of the sample. The incident photon wavevector is then enhanced by $+2m\pi/d$, where d is the grating period and m is an integer.

Interface phonons

Under certain conditions phonons can propagate along the interface between two different media that are in intimate contact, as for instance, in the case of thin epilayers on substrates, or in multilayered

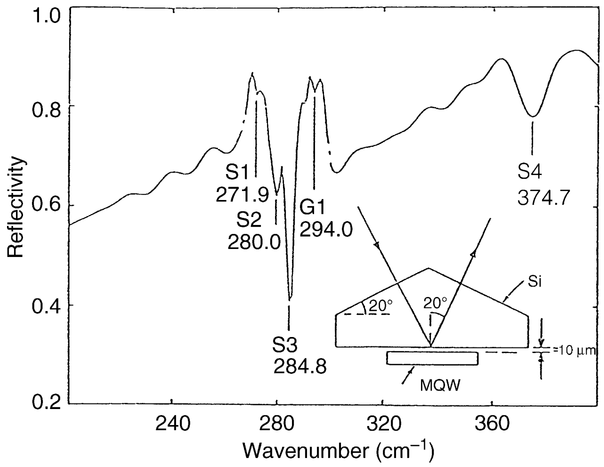

Figure 16 ATR reflection spectrum with dips due to surface phonon polaritons (S) and guided waves (G) in a GaAs/Al$_{0.38}$ Ga$_{0.65}$As MQW. Inset: schematic diagram of the ATR stage. Reproduced with permission from Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

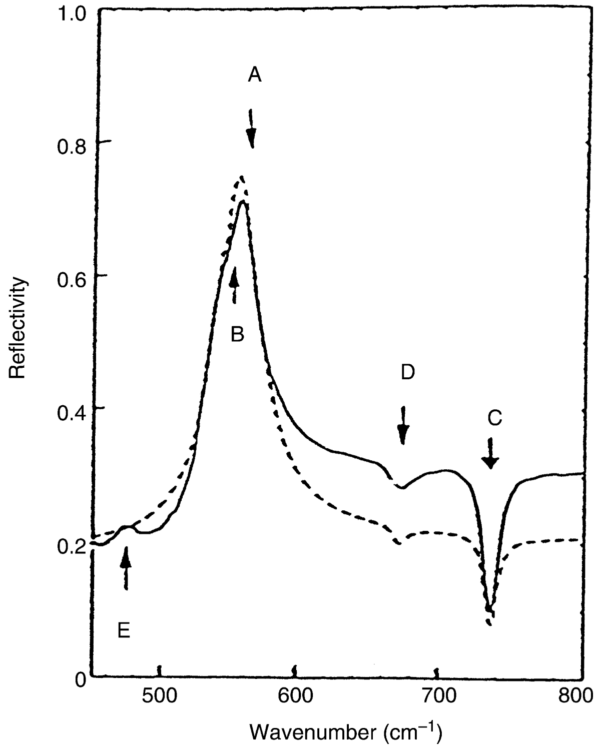

Figure 17 Far-infrared reflectivity of a GaN epilayer on a GaAs substrate at 290 K with dips due to a Brewster mode at C and a Berreman mode at D. Reproduced with permission from Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

structures. There are two common sets of criteria for the existence of such interface modes. The first, leading to so-called Brewster modes, is when the numerator of the Fresnel reflection coefficient passes through zero due to impedance matching in the two media. The second, leading to a Berreman mode, is when the real part of the dielectric constant has a zero crossing in one of the media at a frequency where it is negative in the other medium. Both modes are observed as dips in the reflectivity. A reflection spectrum for a GaN epilayer on a GaAs substrate is shown in Figure 17. The two interface modes occur at frequencies close to the frequencies of GaN LO phonons; the Berreman mode is close to the GaN LO phonon frequency at the base of the GaN epilayer, which is shifted down by about 50 cm^{-1} from the bulk GaN value (735 cm^{-1}) due to disorder arising from the large lattice mismatch between the epilayer and the substrate. Both interface modes couple strongly to free carriers (i.e., plasmons) propagating normal to the interface, and thus act as useful probes of the structural and electronic properties of the epilayer. No other technique is available for probing the electronic properties of the epilayer for propagation normal to the layer.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Band Structure and Optical Properties; Light Scattering; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Angress JF, Goodwin AR and Smith SD (1965) A study of the vibrations of boron and phosphorus in silicon by infrared-red absorption. *Proceedings of the Royal Society A* 287: 64.

Barker AS Jr and Sievers AJ (1975) Optical studies of the vibrational properties of disordered solids. *Reviews of Modern Physics* 47 (Suppl. 2): S1.

Berreman DW (1963) Infrared absorption at longitudinal optic frequency in cubic crystal films. *Physical Review* 130: 2193.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattices*. Oxford: Clarendon Press.

Dumelow T and Tilley DR (1993) Optical properties of semiconductor superlattices in the far infrared. *Journal of the Optical Society of America* 10: 633.

Dumelow T, Hamilton AA, Parker TJ, Tilley DR, Foxon CT, Hilton D and Moore KJ (1990) Far infrared measurements of bulk and surface phonons in GaAs/AlAs superlattices. *International Journal of Infrared and Millimeter Waves* 11: 901.

Johnson FA (1959) Lattice absorption bands in silicon. *Proceedings of the Physical Society* 73: 265.

Kittel C (1996) *Introduction to Solid State Physics*, 7th edn. New York: John Wiley.

Kleinman DA and Spitzer WG (1960) Infrared lattice absorption of GaP. *Physical Review* 118: 110.

Koteles ES and Datars WR (1976) Two-phonon absorption in InP and GaP. *Solid State Communications* 19: 221.

Mirjalili G, Parker TJ, Farjami Shayesteh S, Bulbul MM, Smith SRP, Cheng TS and Foxon CT (1998) Far infrared and Raman analysis of phonons and phonon interface modes in GaN epilayers on GaAs and GaP substrates. *Physical Review B* 57: 4656.

Newman RC (1974) *Studies of Crystal Defects*. London: Taylor and Francis.

Theis WM, Bajaj KK, Litton CW and Spitzer WG (1982) Direct evidence for the site of substitutional carbon impurity in GaAs. *Applied Physics Letters* 41: 70.

Waugh JLT and Dolling G (1963) Crystal dynamics of gallium arsenide. *Physical Review* 132: 2410.

Light Scattering

M Balkanski, Université Pierre et Marie Curie, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

When a monochromatic light beam passes through a transparent medium, a small part of the light is scattered out of the incident direction with a change in frequency. The scattering processes are usually classified into three groups. The light scattering is named:

Brillouin scattering when the light is scattered with a small frequency shift that varies continuously with scattering angle.

Raman scattering when the light is scattered with a relatively large frequency shift that is independent of scattering angle. The possibility of observing a given transition, however, depends on the orientation of the crystal relative to the polarization of the incident light.

Rayleigh scattering when the light is scattered without frequency shift. In this case the scattering is elastic with the incident and scattered frequencies equal: $\omega_i = \omega_s$. Rayleigh scattering is particularly useful in the study of critical phenomena or aspects related to the size and polarizability of particles.

Brillouin Scattering

Brillouin scattering is caused by the interaction of light with the acoustic modes of vibration of the crystal. Consider an acoustic wave of frequency ω_a propagating with velocity $\pm v$ as shown in Figure 1. Light with incident frequency ω_i interacts strongly with acoustic waves satisfying the Bragg condition $\omega_a = 2\omega_i(v/c)n\sin(\theta/2)$, where n is the refractive index of the material and θ is the angle of deviation of the scattered beam. Since the acoustic waves are moving with velocity $\pm v$ the scattered light suffers a Doppler shift in frequency and exhibits a frequency doublet at the angle θ given by the Brillouin equation

$$\omega_i \pm \omega_a = \omega_i \pm 2\omega_i(v/c)n\sin(\theta/2)$$

In terms of the quantum picture, Brillouin scattering corresponds to scattering of incident photons of frequency ω_i and wavevector $\mathbf{k}_i$ into scattered photons of frequency ω_s and wavevector $\mathbf{k}_s$ with the emission or absorption of an acoustic phonon of frequency ω_a and wavevector $\mathbf{q}$. The predicted Brillouin spectrum is shown in Figure 2. It consists of Stokes $(\omega_s = \omega_i - \omega_a)$ and anti-Stokes $(\omega_s = \omega_i + \omega_a)$ components corresponding to phonon emission and absorption, respectively.

The frequency shift $|\omega_i - \omega_s|$ for a well-defined scattering angle θ directly gives the acoustic phonon frequency and velocity. The width Γ of the Brillouin component of the spectrum is a measure of the damping or attenuation of the wave.

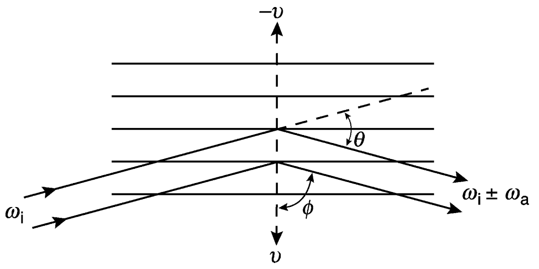

Figure 1 Light scattering at angle ϕ and Bragg reflection at angle θ by moving thermal waves.

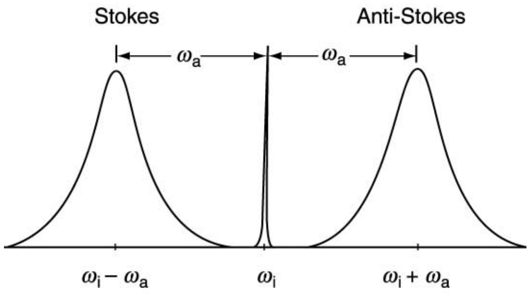

Figure 2 Schematic diagram of a Brillouin spectrum.

Raman Scattering

The essential difference between Brillouin and Raman scattering is that in Raman scattering the incident light beam is scattered with relatively large frequency shift independent of the scattering angle. The same basic considerations apply to Raman scattering that apply to Brillouin scattering. The Raman spectrum has Stokes and anti-Stokes branches corresponding to the emission and absorption, respectively, of an elementary excitation. A variety of elementary excitations are important. They include optical phonons and, in the case of magnetic materials, magnons. Also of interest are electronic excitations such as intraband single-particle excitations, interband excitations, and collective excitations (plasmons).

Raman Scattering by Phonons

An inelastic light scattering event involves the destruction of a photon of frequency ω_i incident from a light source, the creation of a scattered photon of frequency ω_s and the creation or destruction of an optical phonon of frequency ω_o.

Destruction of a photon involves interaction of the radiation field with matter by which a pair of free carriers is created in the semiconductor corresponding to a virtual transition between the valence and conduction bands. The excited state, free electron in the conduction band, interacts with the lattice via the optical deformation potential. This interaction results in the creation or annihilation of an optical phonon of frequency ω_s. When the resulting process corresponds to the creation emission of a phonon, the frequency of the scattered photon is $\omega_s = \omega_i - \omega_a$ and is referred to as the Stokes component of the spectrum. When the process is the annihilation or absorption of a phonon the scattered frequency is $\omega_s = \omega_i + \omega_a$, referred to as the anti-Stokes component.

These two first-order scattering events can be described in terms of Feynman diagrams as shown in Figure 3.

Considering these elementary processes one can calculate, using time-dependent perturbation theory, the Raman amplitude and hence the differential scattering cross-section. The remarkable feature here is that for incident frequencies approaching the frequencies of an electronic excited state of the material the Raman amplitude diverges. This leads to the phenomenon of resonant Raman scattering.

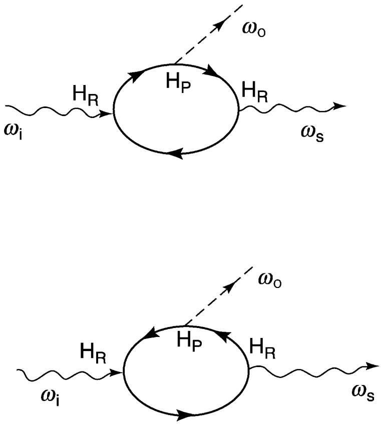

Figure 3 Feynman diagrams for two first-order scattering events.

The Raman scattering efficiency, which is the scattering cross-section per illuminated area can be calculated numerically. Numerical estimates give efficiencies typically of order 10^{-6} or 10^{-7}.

In polar semiconductors the macroscopic electric field leads to an additional contribution to the scattering cross-section for the longitudinal optical (LO) phonon modes. The macroscopic field is a consequence of the Coulomb field of the ionic charges and gives rise to an electron–phonon interaction.

There is no simple relationship between the scattering intensities for transverse optical (TO) and LO phonons in polar semiconductors.

When the probing frequency coincides with the transverse optical mode frequency, i.e., for the regime $cq \cong \omega_{\mathrm{TO}}$ (where c is the light velocity, q the photon wavevector and ω_{TO} the transverse optical phonon frequency) in polar semiconductors it is necessary to take into account the polariton nature of the coupled photon–TO-phonon modes and their associated macroscopic electrical field. The dispersion relation of the coupled photon–TO mode, called a polariton, consists of two branches: an upper branch ω_+, in which the electro-optical contribution dominates, and the lower branch ω_-, which contains essentially the mechanical contribution. This leads to a contribution to the Raman amplitude from both mechanical and electro-optical origins. An exception occurs for the lower polariton branch with $\omega = \omega_{\mathrm{TO}}$; then the electro-optic contribution is very small compared to the mechanical contribution. If $cq \gg \omega_{\mathrm{TO}}$ as in the upper polariton branch, destructive interference between the mechanical and electro-optic contributions can occur.

Selection Rules in Raman Scattering

Not all elementary excitations in semiconductors scatter light. The Raman active modes are determined by selection rules established using group-theoretical methods. The various normal modes in a given crystal correspond to various symmetries of the vibrations of the atoms in the crystal and are characterized by the irreducible representations of the space group of the crystal lattice. One can show that a normal mode can participate in a first-order Raman transition if and only if its irreducible representation is the same as one of the irreducible representations that occur in the reduction of the representation of the Raman tensor.

An important result of group theory is the rule of mutual exclusion which states that, in crystals with a center of inversion, excitations that are active in the first-order infrared spectrum are inactive in the first-order Raman spectrum, and conversely, excitations that are active in the first-order Raman spectrum are inactive in the first-order infrared spectrum. In particular, the first-order spectra due to optical phonons in NaCI are infrared active, but Raman inactive, whereas in Si they are infrared inactive, but Raman active. This difference is related to the fact that each atomic site in NaCI is a center of inversion and the active optical phonons have odd parity whereas each midpoint between two nearest-neighbor atomic sites in Si is a center of inversion and the active optical phonons have even parity.

In a crystal with the zinc blend structure, there is no center of inversion, and the crystal is both infrared and Raman active. The only non-vanishing elements of the Raman tensor $R(\hat{e})$ with $\hat{e} \| z$ are the xy and yx elements. To observe the Raman effect of an LO phonon propagating in the z-direction, one can arrange the polarization of the incident light parallel to the x-axis and observe the scattered light with its polarization parallel to the y-axis or vice versa. An analogous set of constraints applies to TO phonons. A typical spectrum is illustrated in Figure 4.

For higher-order Raman processes in which several phonons participate, it is the product of the irreducible representations of the phonons involved that must be the same as the irreducible representation of the Raman tensor.

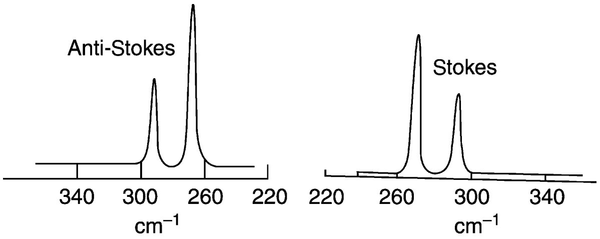

Figure 4 Raman spectrum for optical phonons.

Geometrical Aspects of First-Order Raman Scattering

The observation of light scattering by optical phonons in transparent crystals is usually done in a geometry in which the linearly polarized incident light beam is directed along, say, the x-axis and the scattered beam is observed along the y-axis (see Figure 5). When the crystal is not transparent, as is often the case in semiconductors, the observation is made in the backscattering geometry in which the scattered beam is observed in the opposite direction to the incident beam.

The scattering geometry affects the range of phonon wavevectors that is accessible in first-order Raman scattering. The condition of wavevector conservation is

$$\mathrm{K}_\mathrm{i} - \mathrm{K}_\mathrm{s} = \pm \mathbf{q}$$

where $\mathbf{q}$ is the phonon wavevector and the plus (minus) sign refers to the Stokes (anti-Stokes) process. The Stokes geometry is shown in Figure 6 and satisfies the following relation:

$$q^2 = k_\mathrm{i}^2 + k_\mathrm{s}^2 - 2k_\mathrm{i}\, k_\mathrm{s} \cos\theta$$

Forward scattering is characterized by $\theta = 0$ and a minimum value of q given, for isotropic media, by

$$q_\mathrm{min} = \frac{n(\omega_\mathrm{i})\ \omega_\mathrm{i} - n(\omega_\mathrm{s})\omega_\mathrm{s}}{c}$$

where $n(\omega_\mathrm{i})$ and $n(\omega_\mathrm{s})$ are the refractive indices of the crystal for the incident and scattered light, respectively, and $k_{\mathrm{i(s)}} = n(\omega_{\mathrm{i(s)}})\omega_{\mathrm{i(s)}}/c$.

Back scattering is characterized by the maximum value of q when $\theta = 180°$ and is given by

$$q_\mathrm{max} = \frac{n(\omega_\mathrm{i})\omega_\mathrm{i} - n(\omega_\mathrm{s})\omega_\mathrm{s}}{c}$$

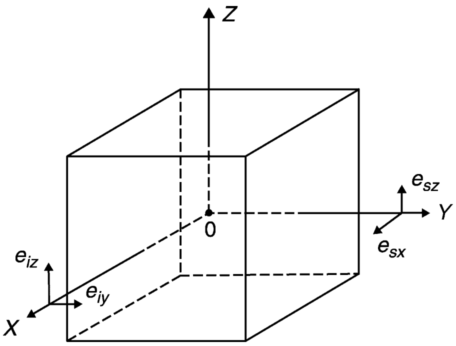

Figure 5 Geometry for light-scattering experiments in transparent crystals.

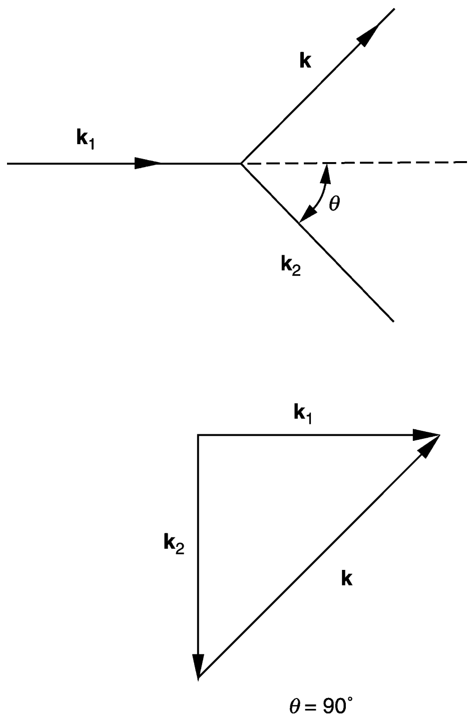

Figure 6 Diagram for Stokes processes.

For typical light-scattering experiments in the visible region the range of the incident wavevector is $0 < k_\mathrm{i} < 10^6\ \mathrm{cm}^{-1}$. This implies that for first-order scattering processes the accessible range of $\mathbf{q}$ under conditions of wavevector conservation is small compared to a nonzero reciprocal lattice wavevector. Light-scattering experiments yield the frequencies of optical modes at essentially the center of the Brillouin zone.

The energy and momentum conservation rules have to be modified when the lifetime of the crystal excitations are strongly limited by their decay into other crystal excitations. Momentum conservation breaks down in imperfect crystals, in solids lacking translational symmetry like amorphous materials, and in crystals which are opaque to incident and scattered light.

In those cases where the incident and scattered waves are damped inside the scattering volume, such as occur in small-gap semiconductors that are opaque at the light frequencies involved, $\mathbf{k}_\mathrm{i}$ and $\mathbf{k}_\mathrm{s}$ are complex. The inelastic scattering is due to excitations having a range of wavevector

$$\Delta q = |\mathrm{Im}\ \mathbf{k}_\mathrm{i}| + |\mathrm{Im}\ \mathbf{k}_\mathrm{s}|$$

about $\mathbf{q} = \mathrm{Re}(\mathbf{k}_\mathrm{i} - \mathbf{k}_\mathrm{s})$. Effects associated with such a wavevector uncertainty have been reported in

Raman scattering spectra of III–V semiconductor compounds.

Second-Order Raman Scattering

Second-order Raman scattering involves two phonons rather than a single phonon as in first-order scattering. The possible second-order processes are the following: two phonons may be created, giving a Stokes component in the scattered light, or one is created and the other destroyed, giving a Stokes or anti-Stokes component, or both may be destroyed giving an anti-Stokes component. For each of these cases there are the possibilities of a pair of first-order electron–phonon interactions or a single second-order electron–phonon interaction.

In the case of second-order Raman scattering the wavevector conservation condition is $\mathbf{q}_1 \pm \mathbf{q}_2 \cong 0$, where $\mathbf{q}_1$ and $\mathbf{q}_2$ are the wavevectors of the phonons involved. This condition places no restriction on the magnitudes of the individual wavevectors, other than $|q_1| \cong |q_2|$, which is in contrast to first-order scattering where $q \cong 0$.

The possible phonon frequency pairings associated with second-order processes are $\omega_1 + \omega_2$ (combination band), $\omega_1 - \omega_2$ (difference band), and $2\omega_1$, (overtone band). The second-order scattering spectrum covers a broad range of frequencies. The overtone spectrum provides a measure of the phonon density of states.

Resonant Light Scattering

General Formulation

Resonant Raman scattering occurs when the incident or scattered photon energy is close to the energy of an intermediate electronic state relative to the ground electronic state. Certain terms in the Raman amplitude then diverge leading to a very large scattering cross-section. Divergences occur in the Stokes spectrum when $\hbar\omega_\mathrm{i} = \hbar\omega_\mathrm{a}$, and $\hbar\omega_\mathrm{s} = \hbar\omega_\mathrm{b}$: the photon energies *resonate* with the excitation energies of the intermediate states a or b. If a and b are the same state, resonance occurs nearly simultaneously for both ω_i and ω_s, and gives rise to a particularly strong enhancement of the scattering efficiency. It is thus clear that the resonance phenomenon is specific to the nature of the intermediate state, and its investigation leads to basic information concerning the electronic states of the system. We shall now examine the resonance behavior in several cases where the intermediate states are carrier Bloch states, free excitons, or bound excitons.

Resonance at the Fundamental Bandgap

The divergent behavior of the Raman amplitude is associated with the factor $(\hbar\omega_\mathrm{i} - E_\mathrm{g})^{-1}$ where E_g is the lowest direct bandgap of the material.

Resonance at Free Exciton States

The Raman intensity as a function of the incident photon energy has a Lorentzian line shape in the range of a single intermediate state. It is centered at the resonance frequency and has a width determined by the lifetime of the intermediate state. The distinction between scattering at resonance and scattering off resonance in the wings of the Lorentzian is that at resonance the intensity is determined by the exciton lifetime, whereas off resonance it is determined by the frequency separation from resonance. Since in many instances excitons have very long lifetimes, they can lead to a very large Raman cross-section at resonance.

Resonance at Bound Exciton States

Scattering induced by the presence of impurities involving bound exciton states has the following essential features: (a) observation of sharp resonance at the energies of the bound exciton which form discrete levels below the free exciton resonance; (b) dependence upon impurity concentration; (c) involvement of LO phonons of wavevector of the order of the inverse of the impurity state radius. Resonance at bound exciton energies has been observed in impurity-induced resonant Raman scattering by LO phonons in CdS. The resonance is centered at the absorption peak for the impurity state and has approximately Lorentzian shape. If the impurity gives rise to a localized vibrational mode, bound exciton resonance in Raman scattering due to the local mode can occur.

Anharmonic Effects on Raman Spectra

In a perfect crystal whose vibrations are harmonic and which is transparent to the incident and scattered light, the Raman lineshape is a delta function centered on the optical phonon frequency. Experimentally one finds that even in crystals of very high quality, the Raman line is broadened into a roughly Lorentzian shape with a width that increases with increasing temperature. This width can be attributed to anharmonic terms in the vibrational Hamiltonian.

The temperature dependence of the line width is determined by the phonon occupation factors. As the temperature approaches zero, the phonon occupation factor approaches zero also, and the width approaches the value due to the zero-point motion of the nuclei. At temperatures above the Debye temperature, the width due to cubic anharmonicity becomes proportional to T. Quartic anharmonicity imparts a T^2 dependence to the width which becomes significant at very high temperatures.

In addition to line broadening, anharmonicity causes a shift of the frequency of peak intensity to lower values as the temperature increases. Contributions to the shift proportional to T and T^2 arise in the high-temperature regime from cubic and quartic anharmonicity, respectively.

Light Scattering due to Electronic Excitations

Light Scattering by Plasmons

Plasmons are similar to LO phonons in that they have a macroscopic electric field associated with them. The field modulates the electric susceptibility and gives rise to Raman scattering. The difference in frequency of the incident and scattered light is equal to the plasma frequency ω_p specified by $\omega_p = e^2 n/\varepsilon_0 \varepsilon_\infty m^*$, where n is the electron concentration, m^* is their effective mass, and ε^∞ is the high-frequency dielectric constant.

When the plasma frequency is close to the LO phonon frequency, there is a plasmon–phonon interaction via the macroscopic electric fields that leads to coupled modes and forces the frequencies apart.

Light Scattering Associated with Interband Transitions

The plasmons dealt with in the preceding section are an intraband type of excitation. Also of interest are interband transitions in which a carrier undergoes a transition from one band to a different band during the scattering process. An example is provided by p-type silicon in which an electron makes a transition from an occupied state in the light-hole band to an unoccupied state in the heavy-hole band via the conduction band as shown in Figure 7.

For the case of isotropic bands the scattering efficiency spectrum at 0 K has sharp edges. If the anisotropy of the valence bands is taken into account, the sharp edges of the spectrum are smoothed out.

The experimental Raman spectrum for p-type Si exhibits a broad continuum due to the electronic transitions but in the same spectral range there are also Raman active optical phonons which superpose on the electronic continuum and distort it. The resulting distortion of the spectrum is associated with resonant interference that occurs between the discrete phonon line and the electronic continuum.

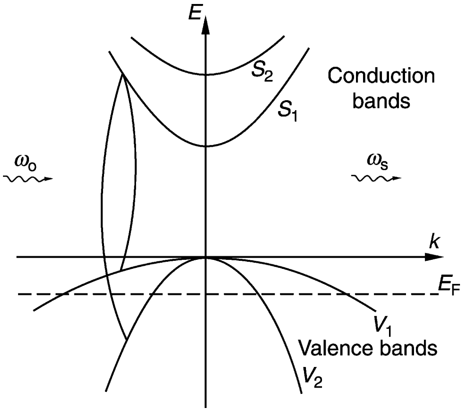

Figure 7 Electronic transitions occurring in intervalence band scattering.

See also

Scattering: Raman Scattering. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Balkanski M (1972) In: Abeles F (ed.) *Optical Properties of Solids*. Amsterdam: North-Holland.

Balkanski M, Jain KP, Beserman R and Jouanne M (1975) *Phys. Rev.* B12: 4328.

Birman JL (1974) *Encyclopedia of Physics*, vol XXV 12b. Berlin: Springer-Verlag.

Born M and Huang K (1954) *Dynamical Theory of Crystal Lattice*. Oxford, UK: Oxford University Press.

Brillouin L (1922) *Ann. Physique* 17: 88.

Buchler S and Burstein E (1974) *Phys. Rev. Letter* 33: 908.

Daman TC and Shah J (1971) *Phys. Rev. Letter* 27: 1506.

Hayes W and Loudon R (1978) Scattering of Light by Crystals. New York: John Wiley.

Haro E, Balkanski M, Wallis RF and Wanser KH (1986) *Phys. Rev.* B34: 5358.

Kanehisa MA, Wallis RF and Balkanski M (1982) *Phys. Rev.* B25: 7619.

Wallis RF and Balkanski M (1986) Many-Body Aspects of Solid Spectroscopy. Amsterdam: North-Holland.

Yu PY, Pilkhun MH and Evangelisti F (1978) *Solid State Communications*, vol. 25, p. 371.

SEMICONDUCTOR PHYSICS

Polarons

J T Devreese, Universiteit Antwerpen, Antwerp, Belgium

© 2005, Elsevier Ltd. All Rights Reserved.

The Polaron Concept

If an electron (or hole) is placed in a polarizable medium (like an ionic crystal or a polar semiconductor), it induces a deformation of polarization field around itself. On the other hand, the electron is attracted to the potential well of this disturbed polarization field. Thus, a feedback loop between the electron and the polarization field arises, which leads to a highly correlated state of the charge carrier and the polarization field. A conduction electron (or hole) together with its self-induced polarization in an ionic crystal or in a polar semiconductor forms a quasiparticle, which is called a polaron. The polaron concept was introduced by Landau in 1933.

The physical properties of the polaron differ from those of the band carrier. In particular, the polaron is characterized by its binding (or self-) energy E_0, effective mass m^* and by its response to external electric and magnetic fields (e.g., dc mobility and optical absorption coefficient).

If the spatial extension of the polaron is large compared to the lattice parameters of the solid, the latter can be treated as a polarizable continuum. This is the case of a 'large polaron', to which most of this article is confined.

The polaron characteristics are determined by the coupling between an electron and the long-wavelength longitudinal optical (LO) phonons. The strength of this coupling is expressed in terms of the constant introduced by Fröhlich:

$$\alpha = \frac{e^2}{\hbar c}\sqrt{\frac{m_b c^2}{2\hbar\omega_{LO}}}\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_0}\right) \qquad [1]$$

In this definition, ω_{LO} is the frequency of LO phonons, c is the velocity of light, and ε_∞ and ε_0 are, respectively, the electronic and the static dielectric constant of the polar crystal.

In order to find the coupling constant according to eqn [1], it is crucial to know the electron (hole) band mass m_b. This unknown parameter can, in principle, be determined from experiment. However, in ionic crystals and polar semiconductors, experiment measures the polaron effective mass m^* rather than the band mass, because of the polaron effect. The problem of determining the band mass is therefore far from trivial and should be tackled on the basis of the results of polaron theory. In Table 1 the coupling constants are indicated for a number of crystals.

Landau and Pekar were the first to investigate the self-energy and the effective mass of the strong-coupling polaron, when $\alpha \gg 1$. As seen from Table 1, in substances with large values of the electron–phonon coupling constant (like alkali halides, TlBr, CdF_2, $SrTiO_3$) α ranges between 2.5 and 4.5. Though, strictly speaking, polarons in such materials are of intermediate coupling, their properties can be approached within the strong-coupling approximation.

The cases when $\alpha \ll 1$ (like InSb, InAs, GaAs) are referred to as the weak-coupling regime. Polarons in many other substances with $\alpha < 1$ (e.g., GaP, CdTe, ZnSe, CdSe, CdS, CdO) are well described in the weak-coupling approximation.

In the weak-coupling regime, the states of the polaron can be imagined as those of a band carrier perturbed by the electron–phonon interaction. Fröhlich provided in 1954 the first weak-coupling perturbation theory results for the polaron characteristics:

$$E_0 = -\alpha\hbar\omega_{LO} \qquad [2]$$

$$m^* = \frac{m_b}{1 - \alpha/6} \qquad [3]$$

Feynman formulated the polaron problem in the Lagrangian (path integral) form of quantum mechanics and obtained an upper bound for the polaron self-energy at all values of the α, which at weak and strong coupling gave accurate limits. Over the years the Feynman all-coupling model

Table 1 Electron–phonon coupling constants

Material	α	*Material*	α
InSb	0.02	AgCl	1.8
InAs	0.052	KI	2.5
GaAs	0.068	TlBr	2.55
GaP	0.201	KBr	3.05
CdTe	0.286	RbI	3.16
ZnSe	0.43	CdF_2	3.2
CdSe	0.49	KCl	3.5
CdS	0.52	CsI	3.67
CdO	0.74	RbCl	3.81
AgBr	1.6	$SrTiO_3$	4.5

for the polaron has been proven to be in many respects the most successful approach to the polaron problem.

Optical Properties and Internal Structure of Polarons

Optical Absorption of Polarons at Weak Coupling

At zero temperature and in the weak-coupling limit, the optical absorption is due to the following elementary polaron scattering process, schematically shown in Figure 1. An incoming photon is absorbed by a polaron. The polaron emits a phonon during the absorption process and takes recoil energy from the incident light. At zero temperature, the absorption coefficient for absorption of light with frequency Ω can be expressed in terms of elementary functions in two limiting cases: in the region of comparatively low frequencies ($\hbar(\Omega - \omega_{\mathrm{LO}})/\zeta \ll 1$)

$$\Gamma(\omega) = \frac{1}{\varepsilon_0 nc}\frac{2^{1/2}N^{2/3}\alpha}{(3\pi^2)^{1/3}}\frac{e^2}{(\hbar m_{\mathrm{b}}\omega_{\mathrm{LO}})^{1/2}}\frac{\omega-1}{\omega^3}\Theta(\omega-1) \quad [4]$$

and in the high-frequency region ($\hbar(\Omega-\omega_{\mathrm{LO}})/\zeta \gg 1$)

$$\Gamma(\omega) = \frac{1}{\varepsilon_0 nc}\frac{2Ne^2\alpha}{3m_{\mathrm{b}}\omega_{\mathrm{LO}}}\frac{(\omega-1)^{1/2}}{\omega^3}\Theta(\omega-1) \quad [5]$$

Here $\omega = \Omega/\omega_{\mathrm{LO}}$, ε_0 is the dielectric permittivity of the vacuum, n is the refractive index of the medium, N is the concentration of polarons, and ζ is the Fermi level for electrons. The step function

$$\Theta(\omega-1) = \begin{cases} 1 & \text{if } \omega > 1 \\ 0 & \text{if } \omega < 1 \end{cases} \quad [6]$$

reflects the fact that at zero temperature absorption of light accompanied by emission of a phonon can occur only if the energy of a photon is larger than that of a phonon ($\omega > 1$). In the weak-coupling limit, according to eqn [4], the absorption spectrum consists of a one-phonon line.

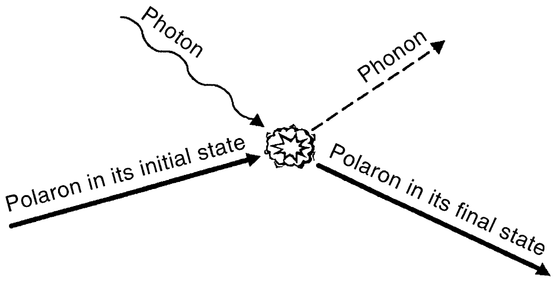

Figure 1 The elementary polaron scattering process leading to absorption of an incoming photon and to generation of an outgoing phonon.

Experimentally, the one-phonon line has been clearly seen for free polarons in the infrared absorption spectra of CdO-films (see Figure 2). In CdO, which is a weakly polar material with $\alpha \approx 0.74$, the optical absorption band is observed in the spectral region between 6 and 20 μm (above the LO phonon frequency). As seen from Figure 2, this optical absorption band can be attributed to the weak-coupling polaron absorption as described by eqn [4].

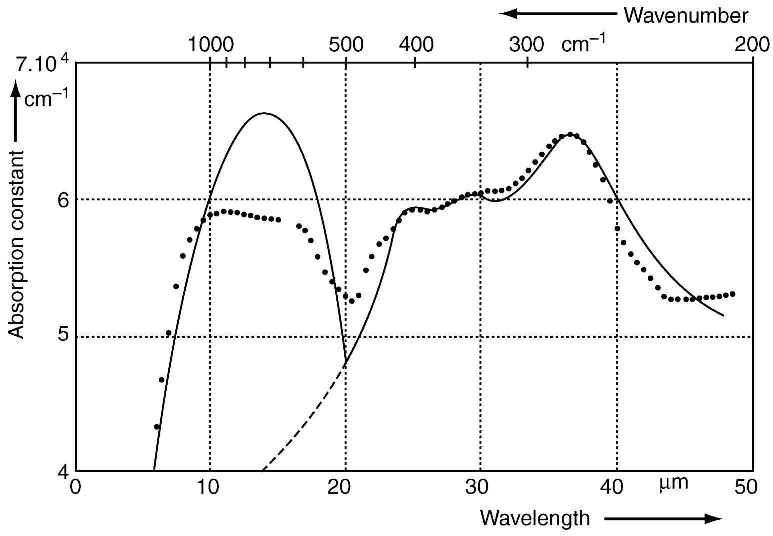

Figure 2 Optical absorption spectrum of a CdO film with the carrier concentration $N = 5.9 \times 10^{19}$ cm^{-3} at $T = 300$ K. The theoretical results are shown with (solid curves) and without (dashed curves) the polaron contribution of eqn [4], and compared to the experimental data (solid dots). (Reproduced with permission from Finkenrath H, Uhle N and Waidelich W (1969) The influence of phonons and polarons on the infrared absorption of cadmium oxide*1. *Solid State Communications* 7: 11–14.)

With strengthening coupling, it becomes possible that absorption of a photon is accompanied by emission of two, three, ..., K phonons, thus giving rise to two-, three-, ..., K-phonon sub-bands in the optical absorption spectra. At high temperatures, absorption of a photon can be accompanied not only by emission, but also by absorption of one or more phonons.

Optical Absorption of Polarons at Arbitrary Coupling

The theoretical absorption spectrum of a single large polaron (at all electron–phonon coupling strengths) is given by the expression

$$\Gamma(\omega) = \frac{1}{n\varepsilon_0 c}\frac{\mathrm{Im}\Sigma(\omega)}{[\omega - \mathrm{Re}\Sigma(\omega)]^2 + [\mathrm{Im}\Sigma(\omega)]^2} \qquad [7]$$

The so-called memory function $\Sigma(\omega)$ contains the dynamics of the polaron and depends on α and ω. This function is related to the impedance function $Z(\omega)$ of the polaron through $-\mathrm{i}Z(\omega) = \omega - \Sigma(\omega)$.

Optical Absorption of Polarons at Strong Coupling

The absorption of light by free large polarons was treated by Kartheuser, Evrard and Devreese using the adiabatic strong-coupling method by Landau and Pekar. The polaron ground state in this scheme has the energy

$$E_0 = -\frac{\alpha^2}{3\pi}\hbar\omega_{\mathrm{LO}} = -0.106\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [8]$$

If the lattice polarization is allowed to relax or adapt to the electronic distribution of the excited electron (which itself then adapts its wavefunction to the new potential, etc., leading to a self-consistent final state), the so-called relaxed excited state (RES) results. Its energy is

$$E_{\mathrm{RES}} = -0.041\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [9]$$

It was argued by Kartheuser, Evrard and Devreese, that for a sufficiently large α ($\alpha > 3$), the (first) RES of a polaron is a stable state, which can participate in optical absorption transitions. This idea was important in order to understand the optical absorption spectrum in the strong-coupling regime. In particular, the following scenario of a transition, which leads to a zero-phonon peak in the absorption by strong-coupling polaron, can be suggested. If the frequency of a photon is equal to

$$\Omega_{\mathrm{RES}} \equiv \frac{E_{\mathrm{RES}} - E_0}{\hbar} = 0.065\alpha^2\omega_{\mathrm{LO}} \qquad [10]$$

then the electron goes from the ground state (following the analogy with the hydrogen atom, let us call it '1s') to an excited state ('2p'), while the lattice polarization in the final state is adapted to the '2p' state of the polaron.

Each RES of the polaron can be accompanied by 1, 2, 3, ... free optical phonons in the lattice. Such states are called scattering states (ScS). They are shown in Figure 3. Starting with the photon energy

$$\Omega_{\mathrm{RES}} + \omega_{\mathrm{LO}} \qquad [11]$$

a transition of the polaron towards the first ScS, belonging to the RES, becomes possible. In the course of this transition, while absorbing a photon, the polaron goes from the ground state to the RES with emission of one extra phonon. This process is called one-phonon sideband absorption.

The one-, two-, ..., K-phonon sidebands give rise to the broad structure of the absorption coefficient. These transitions occur without lattice relaxation. The (unstable) polaron state, in which the lattice polarization corresponds to the electron ground state, while the electron is excited ('a frozen lattice'), is referred to as the *Franck–Condon* (FC) state of a polaron. Within the adiabatic approach, the energy of the lowest FC state is

$$E_{\mathrm{FC}} = \frac{\alpha^2}{9\pi}\hbar\omega_{\mathrm{LO}} = 0.035\alpha^2\hbar\omega_{\mathrm{LO}} \qquad [12]$$

The superposition of the one-, two-, ..., K-phonon sidebands is centered at the FC transition frequency

$$\Omega_{\mathrm{FC}} \equiv \frac{E_{\mathrm{FC}} - E_0}{\hbar} = 0.141\alpha^2\omega_{\mathrm{LO}} \qquad [13]$$

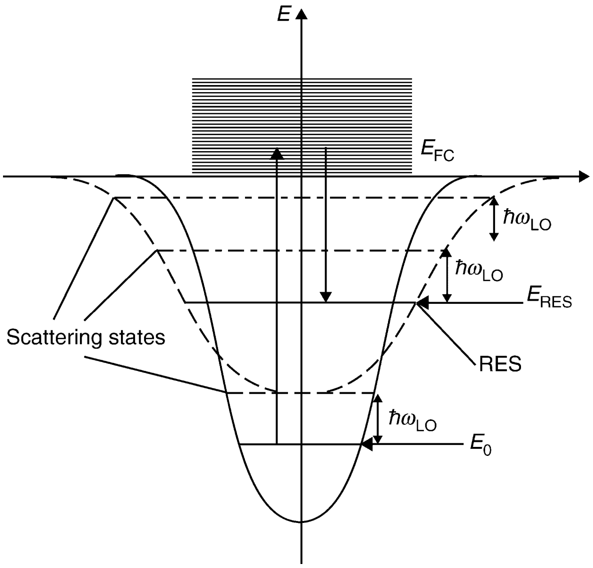

Figure 3 Internal excitations of a polaron at strong coupling: E_0 – the ground state, E_{RES} – the (first) relaxed excited state; the Franck–Condon states (E_{FC}) are shown with a set of horizontal lines above the zero energy. In fact, both the Franck–Condon states and the relaxed excited states lie in the continuum and, strictly speaking, are resonances.

The optical absorption spectrum at strong coupling is, thus, characterized by the following features.

(a) There is an intense absorption peak ('zero-phonon line'), which corresponds to a transition from the ground state to the first RES at Ω_{RES}.
(b) At $\Omega_{\mathrm{RES}} + 1$, a phonon sideband structure appears (namely, a K-phonon sideband at $\Omega_{\mathrm{RES}} + K$). The maximum of this structure occurs at the FC transition frequency Ω_{FC}.

The qualitative behavior predicted by Kartheuser, Evrard and Devreese, namely, an intense zero-phonon (RES) line with a broader sideband at the high-frequency side, was revealed using a general all-coupling expression for the optical absorption coefficient [7] at $\alpha = 5, 6, 7$.

Internal Structure of Polarons at Arbitrary Coupling

The optical absorption spectra calculated according to the theory of Devreese *et al.* are displayed in Figure 4 for values of α in the range from 1 to 6.

The absorption spectrum for $\alpha = 1$ consists of a one-phonon line, similar to the weak-coupling result. The absorption peak for $\alpha = 3$ might correspond to a one-phonon peak at this intermediate coupling. It was also argued that part of the oscillator strength in the absorption for $\alpha = 3$ is due to lattice relaxation in the final states, be it to a smaller extent than for $\alpha = 5$.

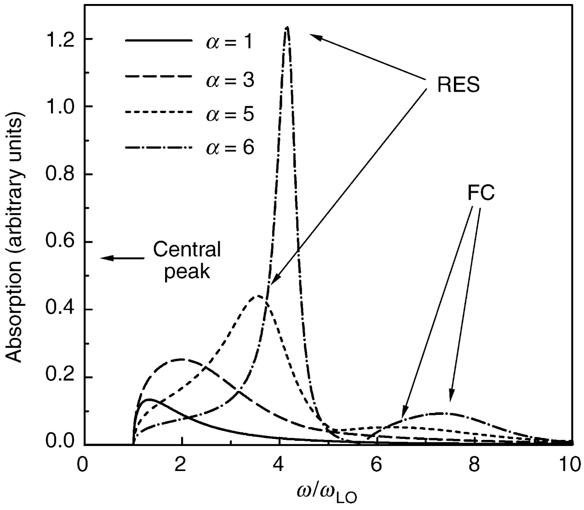

Figure 4 Polaron optical absorption spectra at zero temperature, calculated within the path integral method (according to eqns [11a], [12], [13] of the paper (Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381)) for different values of the electron–phonon coupling constant α. The peaks are labeled in the picture according to their physical origin as discussed in the text. A δ-like central peak is schematically shown by a vertical line.

It is remarkable that from [7] the three different kinds of polaron excitations appear in the calculated optical absorption spectra for polarons at $\alpha \geq 5$:

- scattering states where, e.g., one real phonon is excited (the structure starting at $\omega = 1$);
- relaxed excited state (RES);
- Franck–Condon (FC) states.

Besides that, at zero temperature, the optical absorption spectrum for one polaron contains a central peak [$\sim\delta(\omega)$] at the origin. For nonzero temperature, this central peak smears out and the optical absorption spectrum consists of both a broad envelope and an anomalous Drude-type low-frequency component.

For example, in Figure 4 the main peak of optical absorption for $\alpha = 5$ at $\omega = 3.51$ is interpreted as due to transitions to RES. The shoulder at the low-frequency side of the main peak is attributed to one-phonon transitions to scattering states. The structure centered at about $\omega = 6.3$ is attributed to a FC band. As seen from Figure 4, when increasing the electron–phonon coupling constant to $\alpha = 6$, the RES peak at $\omega = 4.3$ becomes very intense compared with the FC peak centered at $\omega = 7.5$.

The free-carrier polaron effects were revealed through measurements of optical constants, of cyclotron resonance and of mobility in semiconductors and insulating photoconductors. Experimentally, only the scattering states have been seen for free polarons. However, the full structure of eqn [7] has been revealed through cyclotron resonance measurements.

Magneto-absorption of Polarons

Polaron Cyclotron Resonance

A polaron in a magnetic field is referred to as a *magnetopolaron*. The most powerful technique to study the electron–phonon coupling in semiconductors is cyclotron resonance. Changing the magnetic field B allows one to vary the cyclotron frequency $\omega_c = eB/m_b c$. When it is brought closer to resonance with the phonon frequency ω_{LO}, a strongly renormalized magnetopolaron state is formed. This results in a splitting of the cyclotron resonance, which is a measure of the electron–phonon (polaron) coupling constant α.

A theory of the polaron effects in cyclotron resonance was proposed by Larsen. In particular, the variational approach of Larsen is based on an intermediate-coupling theory to calculate the energy levels (modified Landau levels) of a polaron in a magnetic field. The polaron mass is then defined from the energy differences between the polaron (Landau-) energy levels.

A rigorous way to find the polaron cyclotron mass is to calculate the magneto-optical absorption spectrum of the polaron (the quantity which is actually measured) and to define the polaron mass, in the same way as the experimentalists do, from the peak positions in the spectrum. The magneto-optical absorption of polarons for all α and ω_c at $T = 0$ was calculated by Peeters and Devreese. They extended the memory-function formalism described above to study the response of a polaron in a magnetic field. The magneto-absorption was then obtained from

$$\lim_{\varepsilon \to 0} \mathrm{Re} \frac{1}{z - \omega_c - \Sigma(z)} \qquad z = \omega + \mathrm{i}\varepsilon \qquad [14]$$

where $\omega_c = eB/m_b$ is the cyclotron frequency for a rigid-lattice band mass m_b, and $\Sigma(z)$ is the memory function, which takes into account all the polaron internal states as well as all the Landau levels and depends on the coupling constant α and on the magnetic field. The explicit results for $\Sigma(z)$ depend on the optimal parameters of the anisotropic Feynman polaron model, which are found by a variational calculation of the polaron ground-state energy.

At $T = 0$ and when $\omega < 1$, it follows that $\mathrm{Im}\Sigma(\omega) = 0$. According to eqn [14], the position of the cyclotron resonance line is determined by the equation $\omega - \omega_c - \Sigma(\omega) = 0$. If solutions of this equation are denoted as $\omega = \omega_c^*$ then the polaron cyclotron mass m_p^* is calculated from the relation $\omega_c^* = eB/m_p^*$.

Cyclotron Resonance of Polarons in Silver Halides

In pioneering experimental studies, Brown and coworkers have combined mobility experiments and cyclotron resonance measurements to clearly demonstrate the polaron effect in AgBr. From a theoretical plot of polaron mobility versus band mass compared to experimental Hall mobility data at a given temperature, they estimate the band mass. This allows them to calculate α and the polaron mass m^*. This value of m^* can then be compared to the measured cyclotron mass in order to test the polaron theory. For AgBr, the experimental value $m^*/m_e = 0.27 \pm 0.01$ obtained from the cyclotron resonance at millimeter wave frequencies and at 18 K is shown to agree well with the value $m^*/m_e = 0.27 \pm 0.05$ determined by comparison of mobility experiment and theory.

Precise cyclotron mass measurements in AgBr and AgCl covered the range from zero magnetic field to 16 T (see Figure 5). Several polaron theories were compared in analyzing these experimental data. It should be pointed out that the weak-coupling theories (Rayleigh–Schrödinger perturbation theory, Wigner–Brillouin and its improvements) fail (and are all off by at least 20% at 16 T) to describe the

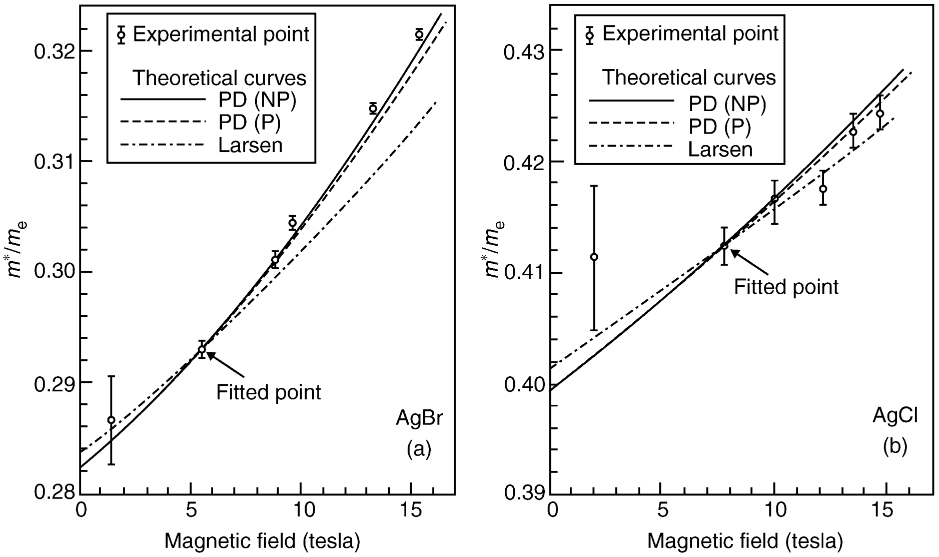

Figure 5 The polaron cyclotron mass in AgBr (a) and in AgCl (b): comparison of experiment and theory. Larsen: (Larsen D (1974) *Journal of Physics C* 7: 2877–2889); PD: (Peeters FM and Devreese JT (1986) *Physical Review B* 34: 7246–7259) (P) – with parabolic band, (NP) – with corrections of the two-band Kane model. In each case the band mass was adjusted to fit the experimental point at 525 GHz. (Reproduced with permission from Hodby JW, Russell G, Peeters F, Devreese JT and Larsen DM (1987) Cyclotron resonance of polarons in the silver halides. *Physical Review Letters* 58: 1471–1474). ©(1987) American Physical Society.

experimental data for the silver halides. The variational approach by Larsen underestimates the polaron cyclotron mass by 2% at 15.3 T. It turns out that the magneto-absorption calculated by Peeters and Devreese leads to the best quantitative agreement between theory and experiment as was analyzed for AgBr and AgCl. This analysis provides a clear confirmation of the Fröhlich description of the polaron in the case where weak-coupling approximations are adequate.

Cyclotron Resonance of Polarons in CdTe

The early infrared transmission study of hydrogen-like shallow donor impurity states in *n*-CdTe was reported by Cohn, Larsen and Lax (see Figure 6). By studying the Zeeman splitting of the $(1s \rightarrow 2p, m = \pm 1)$ transition in the Faraday configuration at magnetic fields up to ~160 kOe, they performed the first quantitative determination of polaron shifts of the energy levels of a bound electron. The experimental data were shown to be in good agreement with the weak-coupling theory of the polaron Zeeman effect. In this comparison, however, the value $\alpha = 0.4$ was used instead of $\alpha = 0.286$, which comes from the definition [1]. Similarly, the value $\alpha \sim 0.4$ was suggested to explain the measured variation of the cyclotron mass with magnetic field in CdTe. This discrepancy gave rise to a discussion where important contributions have been, in particular, due to Harper and Zawadzki.

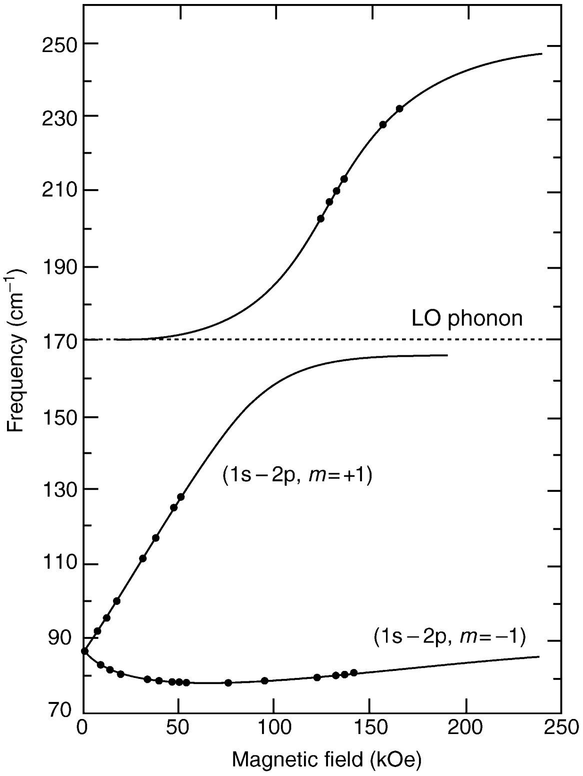

Figure 6 Plot of the experimentally determined magnetic field dependence of the $(1s \rightarrow 2p, m = \pm 1)$ transition frequencies in n-CdTe. The solid lines represent the theoretical results for the effects of the electron–LO-phonon interaction upon the shallow donor impurity levels. α treated as a fitting parameter is taken to be 0.4. (Reproduced with permission from Cohn DR, Larsen DM and Lax B (1972) Polaron Zeeman effect in CdTe. *Physical Review B* 6: 1367–1374). ©(1972) American Physical Society.

Grynberg *et al.* applied the far-infrared photoconductivity technique to study the energy spectrum of shallow In donors in CdTe layers and obtained experimental data over the energy range relevant to study the magnetopolaron effect, as shown in Figure 7. An overall good agreement is found between experiment and the theoretical approach, where the electron–phonon interaction is treated within a second-order improved Wigner–Brillouin perturbation theory and a variational calculation is performed for the lowest-lying donor states ($1s$, $2p^{\pm}$, $2s$, $2p_z$, $3d^{-2}$, $4f^{-3}$). It is to be stressed that this agreement is achieved with the coupling constant $\alpha = 0.286$ as follows from eqn [1] with the known material parameters of CdTe.

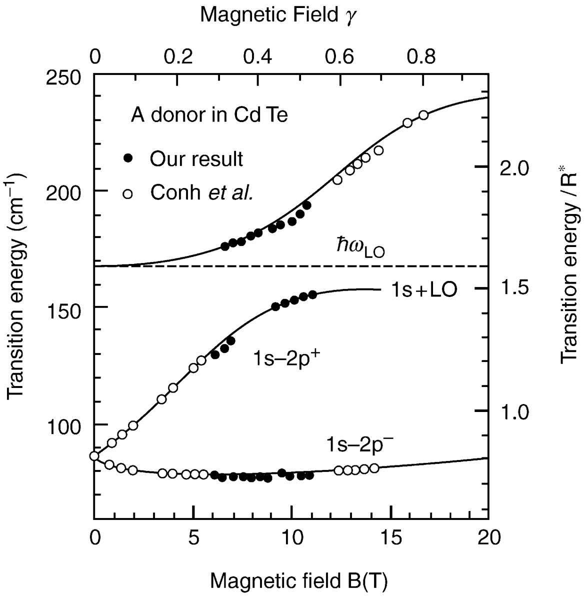

Figure 7 Plot of the experimentally determined magnetic field dependence of the $1s \rightarrow 2p^{\pm 1}$ transition energies in CdTe layers grown by molecular beam epitaxy. The solid lines represent the results of the calculation described in the text without any fitting parameters. The solid dots are the experimental data of the present work and the open circles represent the data of Cohn, Larsen and Lax. (Reproduced with permission from Grynberg M, Huant S, Martinez G, *et al.* (1996) Magnetopolaron effect on shadow indium donors in CdTe. *Physical Review B* 54: 1467–1470). ©(1996) American Physical Society.

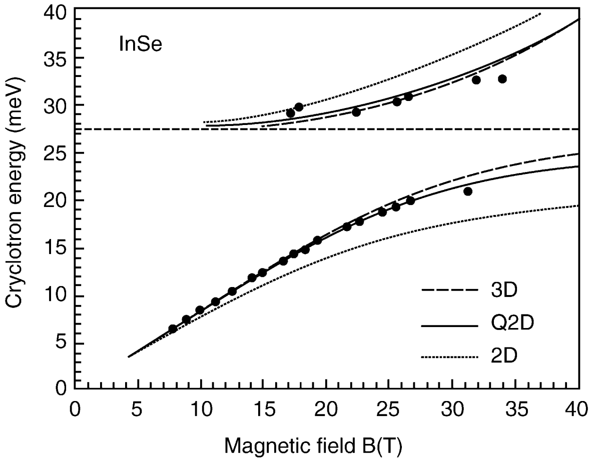

Figure 8 The cyclotron resonance position plotted as a function of magnetic field for InSe. (Reproduced with permission from Nicholas RJ, Watts M, Howell DF, *et al.* (1992) Cyclotron resonance of both magnetopolaron branches for polar and neutral. *Physical Review B* 45: 12144–12147). ©(1992) American Physical Society.

Polarons in the Layered Compound InSe and in Superlattices

Nicholas *et al.* provided a clear demonstration of the polaron coupling by the cyclotron resonance in a two-dimensional electron gas (2DEG), which naturally occurs in the polar semiconductor InSe. One clearly sees, over a wide range of magnetic fields ($B = 18-34$ T), the two distinct magnetopolaron branches separated by as much as 11 meV ($\sim 0.4\omega_{\mathrm{LO}}$) at resonance (Figure 8). The lines show the results of theoretical calculations for coupling to the LO phonon in bulk (3D), sheet (2D) and after correction for the quasi-2D systems at $\alpha = 0.29$ (for motion perpendicular to the *c*-axis). The agreement is reasonable for the 3D case, and is better for the quasi-2D system, where the finite spatial extent of the 2D electron gas in the symmetric planar layer is taken into account.

For GaAs/Al$_x$Ga$_{1-x}$As quantum wells and superlattices, the polaron effect is found to decrease the energy of the shallow donor states at low magnetic fields and to lead to a resonant splitting of the energies at high magnetic fields. The results are in very good agreement with available experimental far-infrared optical data in the whole magnetic-field range.

Optical Properties of Quantum Dots: Effects of the Polaron Interaction

In order to interpret the phonon-assisted optical transitions in semiconductor quantum dots, a theory was developed by Fomin, Gladilin, Devreese *et al.* comprising the exciton interaction with both adiabatic and Jahn–Teller phonons, and also the external nonadiabaticity (pseudo-Jahn–Teller effect). The effects of nonadiabaticity of the exciton–phonon system are shown to lead to a significant enhancement of phonon-assisted transition probabilities and to multiphonon optical spectra which are considerably different from the Franck–Condon progression (see Figure 9). The calculated relative intensity of the phonon satellites and its temperature dependence compare well with the experimental data on the photoluminescence of CdSe quantum dots, both colloidal and embedded in glass. These results clearly demonstrate a breakdown of the adiabatic approximation when describing the polaron interaction in quantum dots.

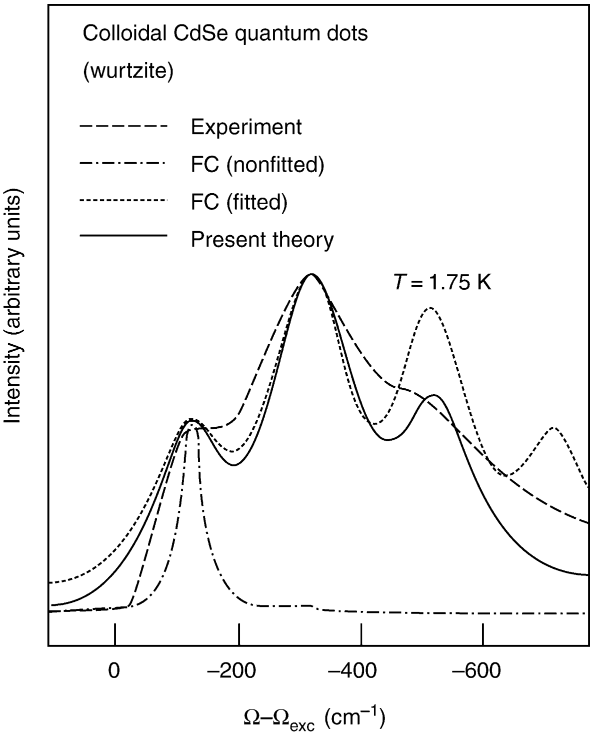

Figure 9 Fluorescence spectra of CdSe quantum dots with wurtzite structure at the average radius $\langle R \rangle = 1.25$ nm. The dashed line represents the experimental data of Nirmal M, Murray CB, Norris DJ and Bawendi MG (1993) *Z. Phys. D* 26: 361–363, the dot-dashed line displays a Franck–Condon progression with the Huang–Rhys parameter $S = 0.06$ calculated using the adiabatic approximation, the dotted line shows another Franck–Condon progression with the Huang–Rhys parameter $S = 1.7$, which is obtained by fitting the ratio of one-phonon and zero-phonon peak heights to the experimental value, and the solid line results from the nonadiabatic theory. (Reproduced with permission from Fomin VM, Gladilin VN, Devreese JT, *et al.* (1988) Photoluminescence of spherical quantum dots. *Physical Reviews B* 57: 2415–2425). ©(1998) American Physical Society.

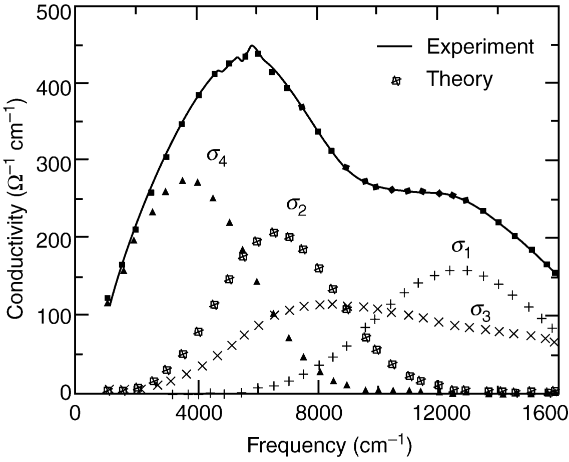

Figure 10 Experimental and theoretical conductivity data at 250 K for $Pr_2NiO_{4.22}$ and four contributions to the theoretical curves resulting from the transitions between the following polaron states: σ_1 – small → small; σ_2 – large → small; σ_3 – small → large; σ_4 – large → large. (Reproduced with permission from Eagles DM, Lobo RPSM and Gervais F (1995) Infrared absorption in oxides in the presence of both large and small polarons. *Physical Review B* 52: 6440–6450). ©(1995) American Physical Society.

Small Polarons

An electron or a hole trapped by its self-induced atomic (ionic) displacement field in a region of linear dimension, which is of the order of the lattice constant, is called a ‘small polaron’. As distinct from large polarons, small polarons appear due to short-range forces.

Experimentally small-polaron effects have been analyzed, e.g., in KCl, LiF, NiO, MnO, TiO_2, $BaTiO_3$, $SrTiO_3$, $LaCoO_3$. More recently, Alexandrov and Mott (1996) surveyed both the principles and the main results of the small-polaron theory in the context of the (bi)polaronic approach to describe the physics of high-T_c superconductors.

Spectroscopic manifestations of both large and small polarons have been found by Eagles *et al.* in the infrared optical absorption spectra of $Pr_2NiO_{4.22}$. The infrared reflectivity spectra on a crystal of this material were measured between 30 and 20 000 cm^{-1}. Mid-infrared conductivity data for frequencies between 1000 and 16 000 cm^{-1} were obtained by Kramers–Kronig transformation of those reflectivity spectra. The resulting optical absorption spectrum was quantitatively interpreted (see Figure 10) using the assumption that in $Pr_2NiO_{4.22}$ two types of polarons, small and large, coexist but do not mix with each other. The contribution due to the transitions between the large-polaron states, which dominates the absorption in the low-frequency region, was calculated using the arbitrary coupling theory by Devreese *et al.* for the coupling constant $\alpha = 3$.

In summary, the polaron concept has been proven to be one of the corner-stones of modern solid-state optics.

Acknowledgments

This work has been supported by the BOF NOI (UA-UIA), GOA BOF UA 2000, IUAP, FWO-V projects G.0287.95, 9.0193.97, G.0274.01 and the W.O.G. WO.025.99N (Belgium).

See also

Semiconductor Materials: Quantum Dots.

Further Reading

Alexandrov AS and Mott N (1996) *Polarons and Bipolarons*. Singapore: World Scientific.

Devreese JT (ed.) (1972) *Polarons in Ionic Crystals and Polar Semiconductors*. Amsterdam: The Netherlands North-Holland.

Devreese JT (1996) Polarons. In: Trigg GL (ed.) *Encyclopedia of Applied Physics*, vol. 14, pp. 383–413. Weinheim: VCH.

Devreese JT, De Sitter J and Goovaerts M (1972) Optical absorption of polarons in the Feynman–Hellwarth–Iddings–Platzman approximation. *Physical Review B* 5: 2367–2381.

Feynman RP (1955) Slow electrons in a polar crystal. *Physical Review* 97: 660–665.

Kartheuser E, Evrard R and Devreese J (1969) Mechanism of absorption of light by free continuum polarons. *Physical Review Letters* 22: 94–97.

Kuper CG and Whitfield GD (eds) (1963) *Polarons and Excitons*. Edinburgh, UK: Oliver and Boyd.

Landwehr G and Rashba EI (eds) (1991) *Landau Level Spectroscopy*, 2 vols. Amsterdam, The Netherlands: North-Holland.

Peeters FM and Devreese JT (1986) Magneto-optical absorption of polarons. *Physical Review B* 34: 7246–7259.

Pekar SI (1951) *Issledovanija po Ekektronnoj Teorii Kristallov*. Moscow: Gostekhizdat (in Russian); German translation: (1954) *Untersuchungen über die Elektronentheorie der Kristalle*. Berlin: Akademie Verlag.

Quantum Wells and GaAs-Based Structures

P Blood, Cardiff University, Cardiff, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

A quantum well is a potential minimum within a semiconductor structure which is sufficiently thin to localize charge carriers on a length-scale similar to their de Broglie wavelength which, for an electron in GaAs at room temperature, is about 30 nm. When electrons are localized in this way their electronic and optical properties are determined by quantum mechanical aspects of their behavior which are not apparent in larger-scale structures. The potential well is usually formed by a sandwich of a thin layer of narrow-gap semiconductor between two layers of a wider gap material as depicted in Figure 1. Typical quantum wells have widths of about 5 nm and the key to their routine production has been the development of advanced epitaxial semiconductor crystal growth techniques such as metalorganic vapor phase epitaxy (MOVPE) and molecular beam epitaxy (MBE). The most significant application of these structures has been in opto-electronic devices, especially laser diodes, and in this article I describe how electrons behave in quantum wells with particular reference to optical properties.

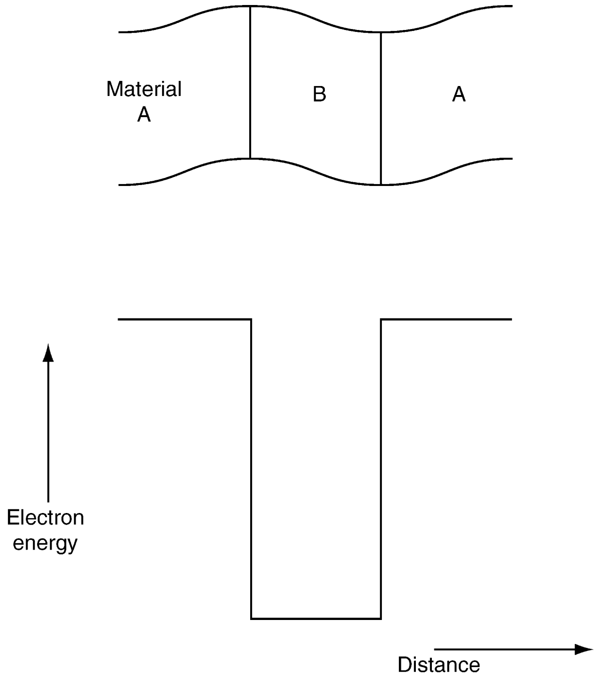

Figure 1 Electron energy diagram illustrating the formation of a potential well by sandwiching a layer of narrow gap semiconductor material (B) between two wider gap layers (A).

The optical properties of semiconductors are usually considered in terms of transitions of electrons between quantum mechanical energy levels, as in the Bohr atom. An alternative description is to consider the spatial displacement of electrons in the oscillating electric field of the electromagnetic radiation. These two descriptions are linked through Schrödinger's equation. For each energy state there is an associated wavefunction which determines the spatial probability distribution of electrons in the state. Consequently, a change in quantum state implies a change in spatial distribution and energy of the electron system. Both approaches are used: here I adopt the viewpoint of transitions between energy states.

We begin with a quantum mechanical description of electrons in a 'bulk' material, as appears in many textbooks. Here electrons are constrained by a three-dimensional potential well of extent given by the sample dimensions, typically millimeters or centimeters and therefore large compared with the de Broglie wavelength. Then we examine what happens when the size of the sample is reduced in one direction: in the limiting case of a quantum well electrons are able to move in only two directions. Quantum confinement is also possible in two or three dimensions to produce a quantum wire or quantum dot, respectively, and it is easy to use the concepts developed in this article to determine the properties of such systems.

Electron States in Bulk Material

We first consider the behavior of an electron in one direction (the x-direction) within a potential well with large dimensions (side length L), then generalize this to three directions (see Figure 2). The wavefunction ψ_n and energy E_n (measured with respect to the bottom of the well) are given by solutions of Schrödinger's equation, which within the well is:

$$\frac{\hbar^2}{2m}\frac{d^2\psi_n}{dx^2} = E_n \qquad [1]$$

It is assumed that the potential is infinitely deep compared with the energy of the electrons.

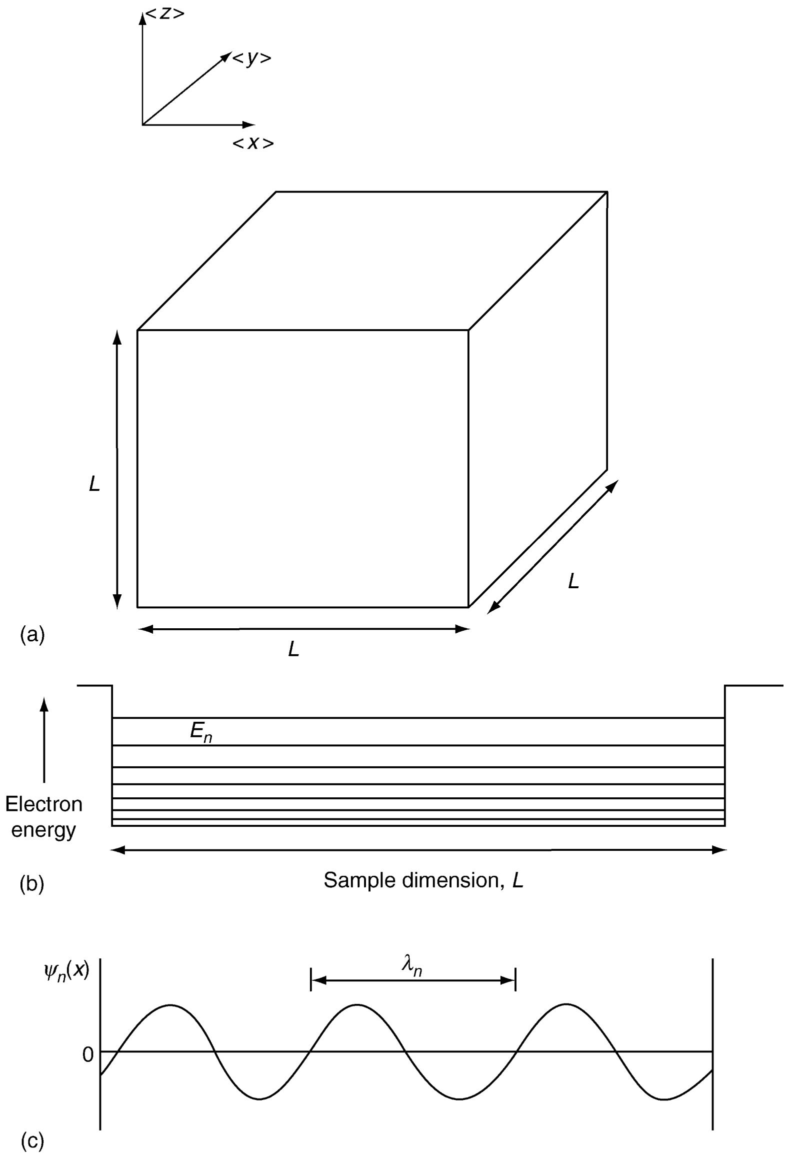

Figure 2 The properties of electrons in a large cubic sample (a) can be obtained by solving Schrödinger's equation for the potential well formed by the sample in each direction (b). This potential well keeps the electrons within the material. The wavefunctions $\psi(x)$ shown in part (c) satisfy cyclic boundary conditions as described in the text, leading to a series of electron energy levels, E_n, given by eqn [3], shown in (b) of the figure.

Within the sample the electrons are described by plane waves of the form $\psi_n = A_n \sin(k_n x)$, where k_n is a wavevector ($= 2\pi/\lambda_n$, where λ_n is the wavelength), and substitution into eqn [1] gives the corresponding energy eigenvalues as

$$E_n = \frac{\hbar^2 k_n^2}{2m} \qquad [2]$$

where m is the mass of the electron. [Substituting $E_n = k_B T$ at room temperature (0.025 eV) and using an effective mass of $0.067m_0$ (for GaAs) in eqn [2] gives $k_n = 2.1 \times 10^8\ \mathrm{m}^{-1}$ and $\lambda = 30$ nm.] Applying cyclic boundary conditions which permit traveling-wave solutions of eqn [1] the wavelength must satisfy the condition $\lambda_n = nL$, i.e., $k_n = 2\pi n/L$, as illustrated in Figure 2, where n is an integer. The dimensions of a typical sample are much greater than the 'size' of an electron so n can be a very large number. The energy eigenvalues are therefore given by

$$E_n = \frac{\hbar^2}{2m}\left(\frac{2\pi n}{L}\right)^2 \qquad [3]$$

This treatment can be extended by solving eqn [1] for motion in three orthogonal directions, x, y, and z. In this case electron motion is described by a wavevector **k** comprising components k_x, k_y, k_z, each of which satisfies the cyclic boundary condition in the respective direction. We take the sample to be of dimension L in each direction with no loss of generality. Thus (representing unit vectors by $\langle \mathbf{x} \rangle$, etc.)

$$\begin{aligned} \mathbf{k} &= k_x\langle \mathbf{x}\rangle + k_y\langle \mathbf{y}\rangle + k_z\langle \mathbf{z}\rangle \\ &= \frac{2\pi n_x}{L}\langle \mathbf{x}\rangle + \frac{2\pi n_y}{L}\langle \mathbf{y}\rangle + \frac{2\pi n_z}{L}\langle \mathbf{z}\rangle \end{aligned} \qquad [4]$$

the energy is

$$E = \frac{\hbar^2}{2m}\left\{k_x^2 + k_y^2 + k_z^2\right\} \qquad [5]$$

and the wavefunction takes the vector form

$$\psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}\cdot\mathbf{r}) \qquad [6]$$

$\psi_n^2(\mathbf{r})$ represents the probability of an electron being at the location **r**. Each allowed electron state obtained by solution of Schrödinger's equation is specified by a unique combination of the numbers $(n_x\, n_y\, n_z)$, which take both positive and negative values corresponding to plane waves traveling in positive and negative directions. The Pauli exclusion principle states that only one electron of each spin can occupy a given quantum state so the amplitude of each wavefunction is normalized such that

$$\int \psi_k^2(\mathbf{r})\,\mathrm{d}\mathbf{r} = \int [A_k \sin(\mathbf{k}\cdot\mathbf{r})]^2\,\mathrm{d}\mathbf{r} = 1 \qquad [7]$$

where the integrals are evaluated over the volume of the sample. Since the amplitude of these solutions is constant throughout the sample electrons may be anywhere in the sample with equal probability. The motion of an *individual* electron of momentum $\hbar\mathbf{k}$ is represented by a wavepacket formed by combination of a number of wavefunctions having similar values of **k**.

Because the sample dimensions are large the energies of the states allowed by Schrödinger's equation take a series of very closely spaced values for increasing integer values of n as shown in Figure 2b. Substituting a typical sample size of $L = 1$ mm into eqn [3], the $n = 1$ and 2 states are separated by only 7×10^{-11} eV in GaAs and these values are very small compared with thermal energies: at room temperature, $k_\mathrm{B}T = 0.025$ eV. The allowed values of **k** and E are discrete and there is a finite but very large number of allowed states in the sample. Since these are very closely spaced we can regard **k** and E as continuous variables and the allowed energy states form a continuum. We can calculate the number of allowed energy states dN in a small energy interval, dE, and hence determine the density of states in energy $g = (\mathrm{d}N/\mathrm{d}E)$ (the number per unit energy interval) at any value of energy.

Electron States in Quantum Wells

Now imagine a sample where the length of one side is reduced in the z-direction (L_z), as illustrated in Figure 3. Gradually the allowed values of k_z become more widely spaced as a consequence of the boundary conditions (eqn [4]). Eventually the width of the potential well becomes similar to the wavelength of the plane-wave state corresponding to the lowest energy level in the well. In these circumstances electron motion is not possible in the z-direction and the wavefunctions become standing waves with boundary conditions $n(\lambda_n/2) = L_z$ in an infinitely deep potential (where n_z takes only positive values) and the amplitude of the wavefunction outside the well must be zero. The lowest energy state, given by eqn [5] with $n_x = n_y = n_z = 1$, no longer lies near the bottom of the potential well. Since the energies associated with motion in the x and y directions remain very small (of order 10^{-10} eV derived earlier) because L_x and L_y are large, the energy of the lowest state is effectively determined by the z-dimension because it is very small, so eqn [5] gives

$$E_{n_z} = \frac{\hbar^2}{2m}\left(\frac{n_z\pi}{L_z}\right)^2 \qquad [8]$$

for the lowest energy state in the well. For $L_z = 5$ nm (and $m = 0.067$) E_1 is 0.22 eV above the bottom of the well. Equation [8] shows that this energy can be changed by choice of the well thickness. A simple interpretation of this behavior is as follows. When L_z becomes very small it is no longer possible for the z-component of the wavefunction of the lowest energy state to satisfy the boundary condition of zero amplitude at opposite sides of the sample. The condition can only be satisfied by wavefunctions which have a smaller wavelength, and consequently a higher energy.

Electrons can occupy states defined by all values of $(n_x\, n_y\, n_z)$, thus for $n_z = 1$ there is a continuum of allowed states at increasing energies corresponding to increasing values of $(n_x\, n_y)$ and corresponding to motion in the (x, y) plane (Figure 4a). Since L_x and L_y are both large these states are closely spaced and form a continuum. There is a similar continuum of energy

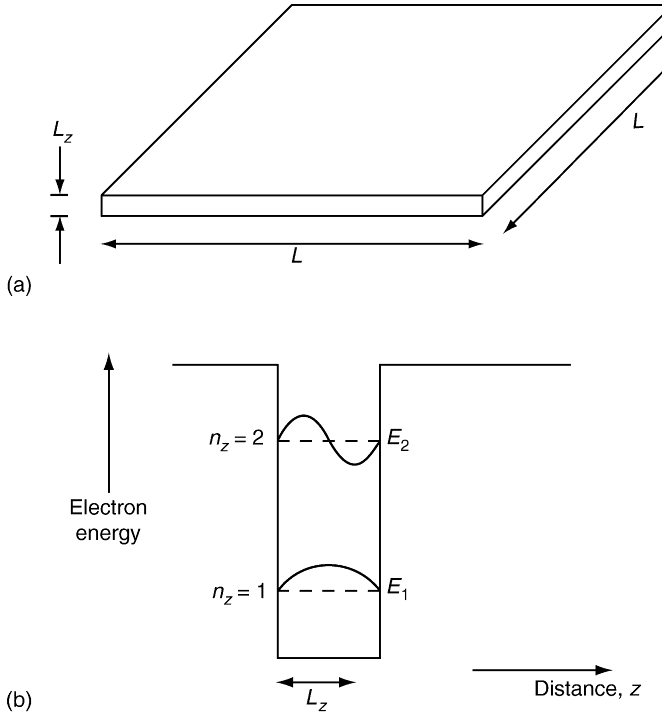
L_z
L
L
(a)
$n_z = 2$
E_2
Electron energy
$n_z = 1$
E_1
L_z
Distance, z
(b)

Figure 3 Electron energy diagram (b) for a sample (a) in which one dimension, L_z, is very small. The electron motion in the z-direction is constrained and the associated energy levels become widely spaced, defined by the integers $n_z = 1, 2, \ldots$ in eqn [8]. The wavefunctions in the z-direction, $\psi(z)$, are also illustrated for the first two electron levels.

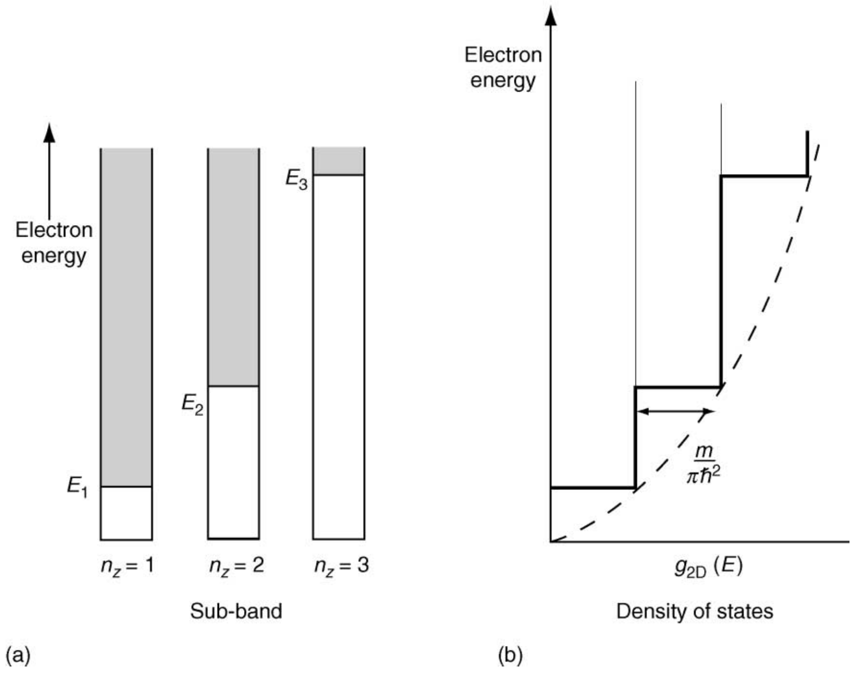
Electron energy
E_1
E_2
E_3
$n_z = 1$
$n_z = 2$
$n_z = 3$
Sub-band
(a)
Electron energy
$\frac{m}{\pi \hbar^2}$
$g_{2D}(E)$
Density of states
(b)

Figure 4 (a) Electron energy diagram showing the energy states associated with unconstrained motion in the (x, y) plane for each sub-band formed by localizing the electrons in the z-direction, defined by $n_z = 1, 2, 3$, etc. When all these allowed states are summed at any energy and expressed as a number of states per unit energy interval we obtain the density of states function $g_{2D}(E)$ shown in (b). This has a series of steps corresponding to each sub-band edge.

states above the energy levels for $n_z = 2, 3$, etc. Thus the energy level diagram is a series of sub-bands, each defined by n_z and with a continuum of states associated with motion in two dimensions in the (x, y) plane, as shown in Figure 4. The number of continuum energy states in a given sub-band in a small energy interval gives the density of states for both spin directions

$$g_{2D}(E) = \frac{m}{\pi\hbar^2} \qquad [9]$$

per unit energy interval per unit sample area.

The density of states is independent of n_z and therefore the same for all sub-bands, and is independent of energy. The density of states function is therefore a series of steps of height given by eqn [9] at each sub-band energy as illustrated in Figure 4b. The density of states per unit area within a given sub-band is independent of L_z as a consequence of dealing with a two-dimensional system.

In real samples the quantum well is not infinitely deep: typically the well depth is in the range 100–400 meV, consequently electrons are not totally confined to the well but are able to penetrate into the barrier material by quantum mechanical tunneling. Solution of Schrödinger's equation for a finite well (treated in many quantum mechanics textbooks) again yields a series of sub-bands, finite in number and with an energy spacing smaller than that given by eqn [8]. In a typical GaAs well the lowest energy state may be about 100 meV from the bottom of the well. The density of states in each sub-band, which is due to the (x, y) motion, remains unchanged. The wavefunction in the z-direction $F(z)$ comprises a sine wave-like standing wave within the well and an exponentially decaying wave in the barrier representing tunneling (Figure 5). The wavefunction is therefore made up of (x, y) plane-wave components and the localized z component and is of the form

$$\Psi_k(\mathbf{r}) = A_k \sin(\mathbf{k}_{xy} \cdot \mathbf{r}_{xy}) F(z) \qquad [10]$$

where $\mathbf{k}_{xy}$ and $\mathbf{r}_{xy}$ are the wavevector and position vector in the (x, y) plane. This wavefunction is again normalized according to eqn [7] such that each state is occupied by one electron of a given spin direction. We see from Figure 5 that we cannot regard the electron as being localized in the well: the electron distribution in the z-direction is specified by the probability distribution $F^2(z)$.

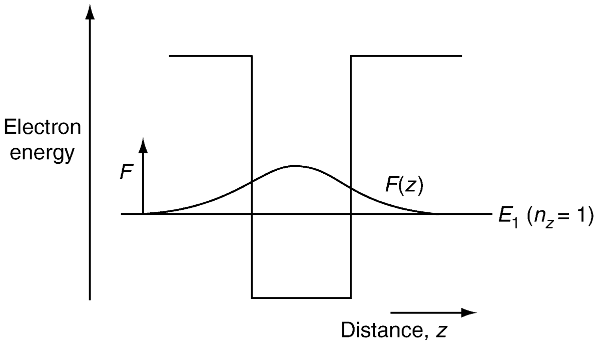

Figure 5 Illustration of the envelope wavefunction $F(z)$ in the z-direction for an electron localized in the $n_z = 1$ sub-band of a well of finite depth. The electron is able to penetrate the barrier by tunneling and this is represented by the decaying part of the wavefunction outside the well.

Occupancy of States in Quantum Wells

The density of states function describes the energy distribution of allowed states in the quantum well. Whether or not particular states are occupied by electrons is determined by the electron concentration and the temperature through the thermodynamic properties of the system. In most cases electron–electron scattering is sufficiently rapid to bring the electron population into an internal equilibrium. Furthermore the interaction between the electrons and the crystal lattice is able to redistribute energy between the two systems so that the lattice and the electron distribution have the same temperature. The electron distribution can therefore be described by Fermi–Dirac statistics for which the probability of a state at energy E being occupied with an electron is

$$f(E) = \frac{1}{1 + \exp\left\{\dfrac{E - E_f}{k_B T}\right\}} \qquad [11]$$

where E_f is the chemical potential (which can be equated with the Fermi level for electrons), T is the lattice temperature and all energies are positive quantities measured with respect to the same arbitrary zero. The energy distribution of electrons in the well, $n(E)$, is then the product of the density of states function, $g_{2D}(E)$, and the occupation probability, $f(E)$, as depicted in Figure 6. The total number of electrons, n, is the integral over energy. Combining eqns [9] and [11], for a single sub-band gives:

$$n_a = \frac{mk_B T}{\pi\hbar^2} \ell n\left\{1 + \exp\left(-\frac{E_n - E_f}{k_B T}\right)\right\} \qquad [12]$$

per unit area, single sub-band.

We cannot specify the carrier density in the well 'per unit volume' since the extent of the distribution in the z-direction is not defined because electrons

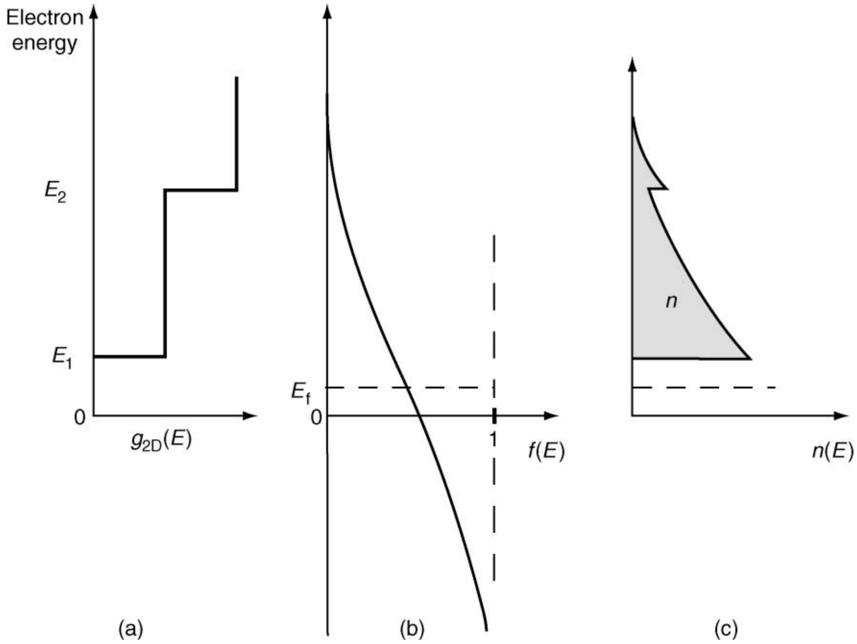

Figure 6 (a) The density of states function for a quantum well, with the value of $g_{2D}(E)$ on the horizontal axis indicating how the density of states varies with increasing electron energy. (b) is an illustration of the probability that a state is occupied by an electron, showing how $f(E)$ varies with increasing electron energy, given by eqn [11]. The actual density of electrons in a given energy interval at any energy $n(E)$ is given by the product of the density of available states and the probability that the state is occupied as shown in (c). The integral of $n(E)$ over energy gives the total density of electrons given for each sub-band by eqn [12].

may tunnel into the barrier material. The electron distribution is properly specified as a number per unit area having a probability distribution in the z-direction given by the function $F(z)$. Where more than one sub-band is populated the total electron density is obtained by adding the contributions from each sub-band each given by eqn [12] with the same Fermi energy and the appropriate sub-band energy. We always use Fermi factors to specify the probability of occupation of a state by an *electron*. The probability that a state is empty is $(1 - f)$.

Formation of Quantum Wells

A quantum well is formed by sandwiching a narrow gap semiconductor layer (E_{g1}) between two layers of wider gap material (E_{g2}). To avoid the formation of defects and dislocations, which have a deleterious effect on many properties of the structure, the constituent materials must have the same lattice parameter. Figure 7a is an energy band diagram for an n-type double heterostructure with layers of micron dimensions. At each interface there are discontinuities in the conduction and valence bands, ΔE_c and ΔE_v, respectively, which account for the difference in the band gaps of the two materials, ΔE_g. While the value of ΔE_g is known for any system, the manner in which this difference is apportioned between the two band edges is not known a priori and the values of ΔE_c and ΔE_v must be obtained from experiment. The discontinuities are often expressed as fractions Q_c, Q_v of the band gap difference:

$$\Delta E_c = Q_c \Delta E_g, \qquad \Delta E_v = Q_v \Delta E_g \qquad [13]$$

In equilibrium the Fermi energy is constant across the diagram and there is a band-bending region on each side of the discontinuity such that the conduction band edges return to their equilibrium energy values at large distances from the interface. The extent of these regions is determined by the discontinuity and the doping level and is typically 200–1000 nm. At each single interface a triangular well is formed which localizes electrons and whose precise shape is determined by the doping density, the carrier distribution and discontinuity through simultaneous solution of Schrödinger's equation and Poisson's equation. Single heterointerfaces only localize the majority carrier. Quantum confinement of both types of carrier can be achieved by a very thin narrow gap semiconductor layer.

When the layer of narrow gap material is made thinner than the band-bending regions, the

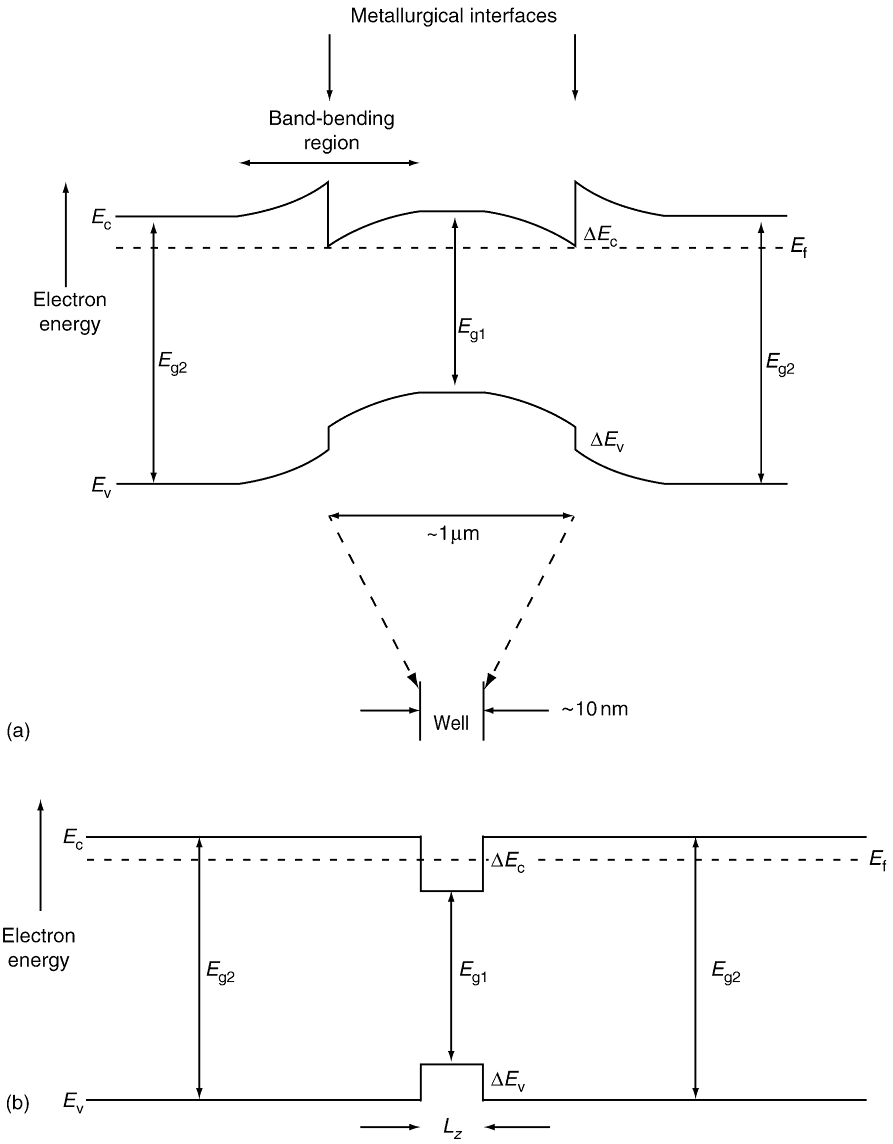

Figure 7 Electron energy diagrams showing the formation of a quantum well as a very thin double heterostructure made up of materials of band gaps E_{g1} and E_{g2}. When the layers are thick the energy diagram is influenced by band bending at the interfaces as shown in (a). When the narrow gap layer is made thinner than the band-bending distances a rectangular well is formed in the conduction and valence bands with depths determined by the respective band offset as shown in (b). On the small distance scale of the well the bands are flat in each region of the structure.

conduction band edge in this layer is not able to regain its bulk equilibrium position and the two triangular wells coalesce as illustrated in Figure 7b. When the well is sufficiently thin, say L_z less than 20 nm, the band-bending across this layer is negligibly small and the well becomes effectively rectangular. This well accumulates electrons from the surrounding barrier material and its potential relative to the Fermi level rises such that there is band bending in the barrier material on each side of the well inhibiting further accumulation. Since the band bending in the barrier occurs on a distance of hundreds of nanometers (being determined by the doping density) it is not apparent over the distance scale of the diagram of the quantum well, which is on the order of 10 nm. Consequently, as shown in Figure 7b, the well can be drawn as a rectangle to a good degree of approximation. This thin, narrow gap layer localizes both electrons and holes. The form of the well is determined by the respective band discontinuities, ΔE_c and ΔE_v, and the thickness of the layer.

Potential wells formed in this way are the basis for most quantum-confined device structures.

Other forms are possible. In certain material combinations ΔE_c or ΔE_v may be negative, meaning that, at the interface, both discontinuities are in the same sense and only one carrier type is confined in the thin narrow gap material. These are known as 'type II' structures. If there is an electric field across the well due to doping or strain (piezoelectric) effects the well becomes triangular. Wells can also be engineered with steps in the potential profile and multiple well systems can be grown in which the states are quantum mechanically coupled. All these variants provide opportunities to engineer the properties of the structure.

One important development of the rectangular quantum well deserves mention. It is possible for the well material to have a different lattice parameter to that of the surrounding barrier material provided that the strain energy can be accommodated elastically within the structure. This means that the strain energy in the layer must be below the energy necessary for formation of dislocations and this translates into an upper limit on the layer thickness (the critical thickness) for a given mismatch. Incorporation of elastic strain is significant because relaxing the lattice match requirement widens the choice of well and barrier materials and because strain modifies the properties of the electronic states in the quantum well and these effects can be used to advantage in the design of devices.

Here we concentrate on the rectangular potential well. Such wells are widely used, particularly in opto-electronic devices, and this structure provides the basic concepts which lie at the heart of all quantum-confined systems. We consider the 'model' lattice-matched quantum well system: a GaAs well bounded by AlGaAs barriers.

GaAs/AlGaAs Quantum Well Structures

As Al is substituted for Ga in the $Al_xGa_{1-x}As$ alloy system the direct band gap increases from 1.424 eV in GaAs ($x = 0$) to 3.018 eV in AlAs ($x = 1$) (Figure 8) while the lattice parameter remains unchanged. A quantum well is formed using a narrow GaAs layer sandwiched between barrier layers of $Al_xGa_{1-x}As$ having a composition chosen to give the desired well depth. The whole structure is grown on a GaAs substrate. The band discontinuity ratio between GaAs and $Al_xGa_{1-x}As$ is independent of alloy composition (ΔE_c and ΔE_v are constant fractions of ΔE_g as x is varied) with $Q_c = 0.66$ and $Q_v = 0.33$. For a typical barrier composition of $x = 0.3, \Delta E_g = 0.374$ eV and ΔE_c and ΔE_v are 0.247 eV and 0.127 eV, respectively.

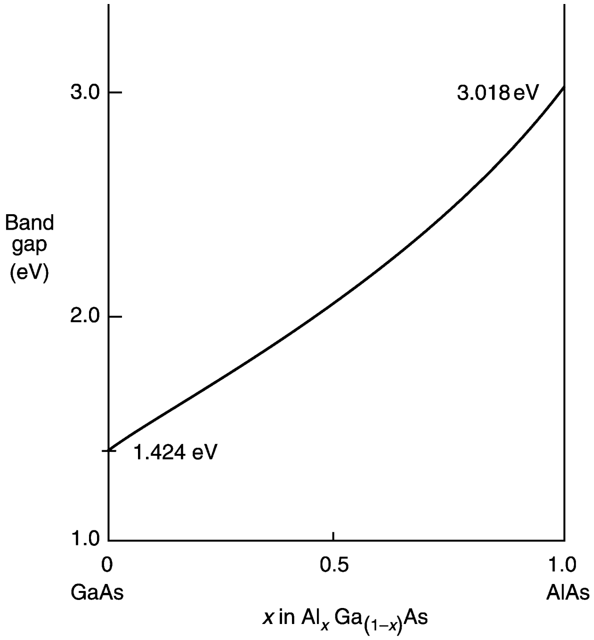

Figure 8 Variation of the direct band gap of $Al_xGa_{(1-x)}As$ with Al content (x).

Optical Properties of AlGaAs/GaAs Quantum Wells

General Principles

Figure 9 illustrates a transition of an electron between an occupied state at E_1 in the $n_z = 1$ sub-band in the conduction band and an empty state (i.e., a hole) at E_2 in the $n_z = 1$ sub-band in the valence band, resulting in emission of a photon of energy $h\nu = E_1 - E_2$ as required by energy conservation. This is the process of luminescence (or spontaneous emission), which requires external excitation such as illumination (photoluminescence) or biasing a p-n junction (electroluminescence, as in a light-emitting diode). Alternatively, incident light may promote an electron from an occupied state in the valence band to an empty state in the conduction band with the consumption of the energy of a photon. This is the process of optical absorption.

At typical temperatures the electrons and holes are at energies near to their respective band edges so the photon energy $h\nu$ is close to the separation of the $n_z = 1$ sub-bands. Since these energies are determined by the well width it is possible to change the photon energy for emission or absorption by changing the well width. As the well width is reduced the energy separation increases and the wavelength of light emitted by the structure is reduced. This ability to engineer the emission wavelength by simply adjusting

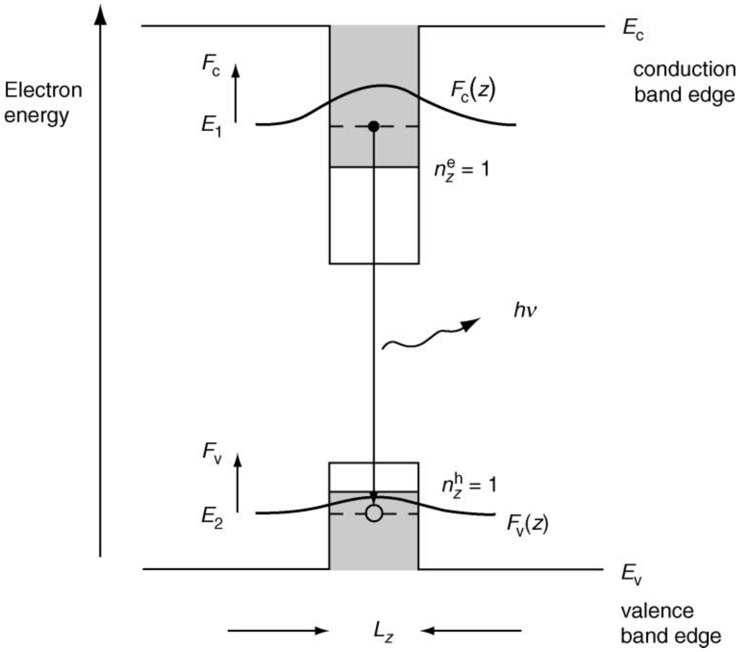

Figure 9 Illustration of a downward optical transition of an electron from a state in the $n_z = 1$ conduction sub-band to an empty state (hole) in the $n_z = 1$ valence sub-band resulting in emission of a photon of energy $h\nu = E_1 - E_2$. The spatial probability distribution of the electron and hole is specified by their respective envelope functions $F_c(z)$ and $F_v(z)$ and the probability of the transition occurring is proportional to their overlap integral, eqn [14].

the well width, without changing the chemical composition of the constituent materials, is one of the major attractions of quantum-confined structures. The shortest wavelength possible for a given material combination is determined by the band gap of the wide gap barrier material alongside the well.

The strength of the luminescence signal or the absorption is determined by the rates at which the electronic transitions occur. In general these rates are determined by the following intrinsic factors.

1. *The quantum mechanical probability of an electron making a transition between a full and an empty state*. This depends upon the 'matrix element' for the transition, which is a fundamental property of the well material and the overlap of the envelope functions. The electron and hole in the initial and final states must be in the same region of space and if the spatial distributions of electrons and holes, determined by F_c and F_v, do not overlap the probability for an electron to make a transition to the hole state is very low. The overlap is almost complete in a rectangular well (see Figure 9), but is only partial in a triangular well. The transition probability is proportional to

$$I_{\text{overlap}} = \left\{ \int F_v(z) F_c(z)\, dz \right\}^2 \qquad [14]$$

2. *The probability of occupation of initial and final states by electrons, specified by the Fermi factors* (eqn [11]). The initial state must be occupied and the Pauli exclusion principle prohibits an electron entering a final state which is already occupied. Thus for a downward transition resulting in emission of light the rate is proportional to $f_c(1 - f_v)$.
3. *The density of initial and final states* (eqn [9]). The more states there are within a given small energy interval, the greater the emission rate per unit energy.
4. *Photon distribution*. For processes such as optical absorption, which are 'induced' by the presence of a photon, the transition rate is also proportional to the photon density in the region of the well. Because the electrons are not wholly localized in the well, and because the spatial extent of the optical field is larger than that of the carriers, the coupling between the photon field (intensity $I_{ph}(z)$) and the carrier probability distributions is expressed as

$$I_{\text{coupling}} = \frac{\left\{ \int F_v(z) [I_{ph}(z)]^{1/2} F_c(z)\, dz \right\}^2}{\int I(z)\, dz} \qquad [15]$$

5. *Optical mode density*. The emission rate also depends upon the density of optical modes

available to receive the recombination radiation. In most situations the emission occurs into the cavity of free space which is very large compared with the wavelength of the radiation so the modes are very closely spaced and the cavity can accept radiation at any transition energy (this is the well-known Raleigh–Jeans theory of cavity radiation). However, in some special situations the cavity has dimensions comparable with the wavelength of the radiation (microcavities) and then the modes are widely spaced. In such a microcavity, the optical cavity effectively controls the electronic transitions.

Optical Absorption

Figure 10 is an optical absorption spectrum for a GaAs quantum well measured with light incident normal to the plane of the quantum wells. A series of steps can be seen which correspond to absorption at increasing photon energy between increasing orders of sub-bands, $E_{1e} \rightarrow E_{1h}$, $E_{2e} \rightarrow E_{2h}$, etc. Normally $n_z = 1$ to $n_z = 2$ transitions are forbidden in rectangular wells. At each step there is a pair of peaks (not always resolved) and above each step the absorption is roughly constant, following the step-like density of states function sketched in Figure 4. The peaks are due to the formation of excitons which are weakly bound electron–hole pairs. In bulk materials excitons are only observed at low temperatures whereas in quantum wells the localization increases the binding energy (E_{bX}, typically 6–8 meV) such that they are observed at room temperature. The excitonic peak is at an energy E_{bX} below the sub-band separation. There are in fact two different kinds of electrons in the valence band with different masses so there are pairs of $n_z = 1, 2, 3, \ldots$ sub-bands due to the different confinement energies for the different valence electrons. Spectra of this kind provided the first evidence for the distinctive sub-band structure of quantum well systems produced by carrier confinement. Absorption measurements provide data on the energy spacing of the sub-bands in the well, and it is possible to determine the well width and depths, from which the band offsets can also be determined.

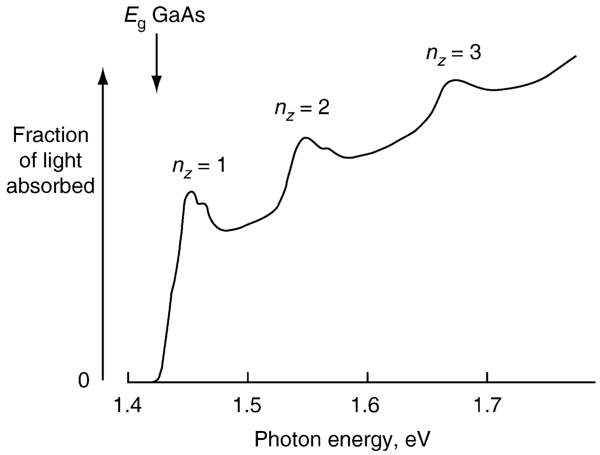

Figure 10 Absorption spectrum of a multiple quantum well structure having 10 nm GaAs wells measured at room temperature. The absorption edges corresponding to transitions between $n_z = 1, 2$, and 3 pairs of sub-bands can be identified. The peaks are due to absorption by formation of excitons.

The strength of the absorption is also of interest because it provides a measure of the optical matrix element if the Fermi factors f_c and f_v are known. If the incident optical beam is very weak such that the absorption process excites very few electrons then the system remains close to thermal equilibrium so the upper states are empty ($f_c = 0$) and the lower states are full ($f_v = 1$).

In *bulk material* the change in photon flux over a small distance Δx is proportional to the flux Φ and the distance traveled

$$\Delta\Phi = -\Phi\alpha\Delta x \qquad [16]$$

where α is the absorption coefficient which has units L^{-1}. This results in an exponential decrease of photon flux with distance traveled through the material

$$\Phi(x) = \Phi_0 \exp(-\alpha x) \qquad [17]$$

For a quantum well structure with light incident in the z-direction perpendicular to the plane of the quantum well, the fraction of light absorbed by a single well by transitions between a single sub-band pair (e.g., transitions between the $n_z = 1$ conduction and valence sub-bands) is independent of the thickness of the well because the density of states for a single sub-band is independent of well thickness. The strength of the absorption cannot be expressed by an absorption coefficient but by the fraction of light absorbed per well:

$$\Delta\Phi = -\gamma_w\Phi \qquad [18]$$

This arises because the quantum well is effectively a sheet in the z-direction: changing its thickness does not change the density of states per unit area (eqn [9]) and we cannot specify the electron concentration along the z-direction. This behavior also occurs because the well is much thinner than the wavelength of the light and it is not possible to specify the variation of photon flux on a distance scale smaller than the wavelength. The barrier material surrounding the well has a wider band gap and does not absorb at the photon energies absorbed by the well. The well thickness does affect the sub-band spacing and at a fixed photon energy the absorption increases as the well width is decreased as light is absorbed by transitions between more sub-bands. Equation [18] applies individually to each sub-band because the density of states is the same for each sub-band.

In a multiple well system, the fraction of light transmitted through N independent wells is

$$T_N = \{1 - \gamma_w\}^N \qquad [19]$$

so the fraction of light absorbed by such a system is

$$\frac{\Delta\Phi}{\Phi} = -\left[1 - \{1 - \gamma_w\}^N\right] = -N\gamma_w \qquad [20]$$

which is proportional to the number of wells when $\gamma_w << 1$.

The fraction of light absorbed by transitions between a single pair of sub-bands for a rectangular GaAs quantum well is predicted to be about 0.01, and a similar value has been obtained from absorption measurements. In a rectangular well the envelope function overlap is complete; however, in a triangular well of the same material the absorption is reduced by the smaller overlap of the envelope functions.

Photoluminescence and Spontaneous Emission

We turn now to discuss the optical emission from quantum wells as a result of external excitation. An external light source with photon energy above the effective band gap produces excess electron–hole pairs by excitation of electrons from the valence band. These rapidly lose their excess energy and thermalize to the sub-band edges to take up Fermi energy distributions. The electrons subsequently recombine to vacant valence band states by spontaneous emission. The low energy edge of this spectrum corresponds to the sub-band separation and the shape of the spectrum at higher energies corresponds to the thermal distribution of carriers in the bands. At very low excitation intensities where the excess electron population is small the luminescence may be due to recombination of electrons within an exciton and occur at an energy E_{bX} below the sub-band separation. Generally the sub-band separation is several $k_B T$ so only the lowest sub-band is populated and photoluminescence is not seen from higher sub-bands. In this respect absorption measurements provide a more complete characterization of the structure.

While the energy position of features in the photoluminescence spectrum provides useful information, it is difficult to use the strength of the photoluminescence signal to provide quantitative information about the rate of optical transitions within the material because the excitation power and volume within the sample must be known and a known fraction of emitted light must be collected and measured with a calibrated detector. The relative intensity of photoluminescence under standard conditions does provide comparative information and the linewidth of the spectrum at low intensity provides a qualitative indication of 'quality' (e.g., well width variations in the structure). A quantitative determination of the internal recombination rate can be obtained by measuring the decay of the luminescence signal following short-pulse excitation. Fuller discussion can be found in books dealing with the characterization of semiconductor structures.

Concluding Remarks

We have shown that an understanding of the electronic structure of quantum wells follows naturally from quantum mechanical treatments of electrons in bulk materials. When one of the dimensions of the structure is reduced to be similar to the de Broglie wavelength, quantization of the electron motion in this direction becomes apparent. The sub-bands which are formed have a profound effect on the electronic and optical properties when their spacing exceeds thermal energies such that only one band is significantly populated with carriers. The density of states function is a series of steps, each corresponding to a sub-band, and this produces characteristic step features in the optical absorption spectrum. The effective band gap of the structure is controlled by the width of the well. Quantum confinement also increases the binding energy of excitons compared with bulk materials and under low excitation conditions absorption and luminescence spectra are dominated by excitonic features even at room temperature.

One of the most successful application areas of quantum well structures has been in laser diodes. Figure 11 shows experimental data for laser emission wavelengths as a function of well width, showing the dramatic reduction which can be achieved simply by changing the width of the quantum well. Furthermore the step density of

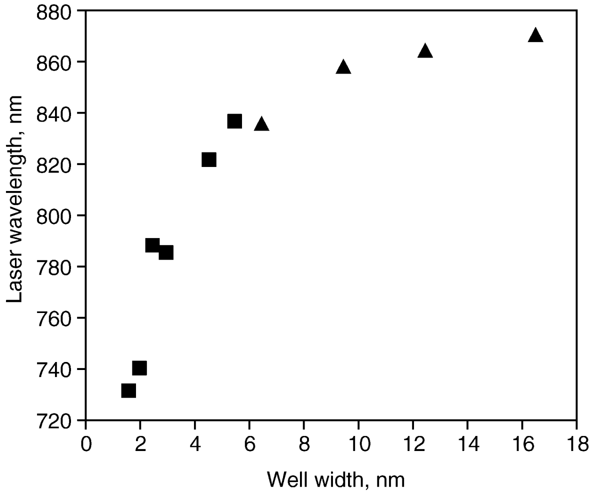

Figure 11 Measured values of the emission wavelength of GaAs/AlGaAs quantum well lasers as a function of well width, showing how the wavelength can be engineered by choice of the well width. Data from: squares: Woodbridge K, Blood P, Fletcher ED and Hulyer PJ (1984) Short wavelength (visible) GaAs quantum well lasers grown by molecular beam epitaxy. *Applied Physics Letters* 45: 16–18; and triangles: Chen HZ, Ghaffari A, Morkoç H and Yariv A (1987) Effect of substrate tilting on molecular beam epitaxial grown AlGaAs/GaAs lasers having very low threshold current densities. *Applied Physics Letters* 51: 2094–2096.

states function, characteristic of quantum wells, is one of the fundamental reasons for the reduction in threshold current of quantum well lasers relative to bulk GaAs devices.

We have considered GaAs because it is a model material system for production of quantum well structures. The concepts developed in this chapter are, however, quite general and can be applied to other material systems, other forms of quantum well and quantum wires and dots.

List of Units and Nomenclature

E	energy
E_f	Fermi energy
E_g	bandgap
F	envelope wavefunction
$f(E)$	Fermi–Dirac distribution function
g_{2D}	two-dimensional density of states $[\text{energy}]^{-1}$ $[\text{area}]^{-1}$
$\hbar$	Planck's constant divided by 2π
I_{ph}	optical field intensity
$\mathbf{k}$	wavevector
k_B	Boltzmann's constant
m	electron mass
n	carrier density, expressed per unit area in quantum well structures
Q_c, Q_v	heterostructure band offset ratios
$\mathbf{r}$	position vector
t	time
T_N	fraction of light transmitted at normal incidence through N quantum wells
$\langle\mathbf{x}\rangle$	unit vector along x-direction, similarly for y, z directions
α	optical absorption coefficient in a bulk material, per unit length
γ	fraction of light absorbed by a quantum well at normal incidence to the plane of the well
λ	wavelength
Φ	photon flux (photons per unit time crossing unit area)
Ψ	electronic wavefunction
ν	light frequency

See also

Semiconductor Materials: Band Structure Engineering; Quantum dots; Type-II Quantum Wells and Superlattices. **Semiconductor Physics:** Excitons; Outline of Basic Electronic Properties.

Further Reading

Bastard G (1988) *Wave Mechanics Applied to Semiconductor Heterostructures*. New York: Les Editions de Physique, Halstead Press.

Bimberg D, Grundmann M and Ledentsov NN (1999) *Quantum Dot Heterostructures*. Chichester, UK: Academic Press.

Blood P (1991) Heterostructures in lasers. In: Morgan DV and Williams RH (eds) *Physics and Technology of Heterojunction Devices*. Stevenage, UK: Peter Peregrinus.

Blood P (2000) On the dimensionality of optical absorption, gain and recombination in quantum-confined structures. *Journal of Quantum Electronics* 36: 354–362.

Blood P and Orton JW (1992) *The Electrical Characterisation of Semiconductors: Majority Carriers and Electron States*. London: Academic Press.

Casey HC Jr and Panish MB (1978) *Heterostructure Lasers*. San Diego, CA: Academic Press.

Dingle R, Gossard AC and Weigmann W (1975) Direct observation of superlattice formation in semiconductor heterostructures. *Physical Review Letters* 34: 1327–1330.

Eisberg R and Resnick R (1985) *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn. New York: John Wiley.

Hook JR and Hall HE (1991) *Solid State Physics*. Chichester, UK: John Wiley.

Kelly MJ (1995) *Low-Dimensional Semiconductors*. Oxford, UK: Clarendon Press.
Orton JW and Blood P (1990) *The Electrical Characterisation of Semiconductors: Measurement of Minority Carrier Properties*. London: Academic Press.
Perkowitz S (1993) *Optical Characterisation of Semiconductors*. London: Academic Press.
Wolfe CM, Holonyak N Jr and Stillman GE (1989) *Physical Properties of Semiconductors*. Englewood Cliffs, NJ: Prentice Hall.

Recombination Processes

P T Landsberg, The University of Southampton, Southampton, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In studies of recombination there occur considerable complications; many are associated with interactions between electrons, electrons and phonons, excitons, bi-excitons, impurity centers, etc. Most of these are not required in the present exposition. Modern work normally assumes that the basics of recombination physics are understood and the present exposition offers an appropriate outline.

Basic Assumptions

Electron–Electron Interactions for Electrons in Bands

The recombination problem in semiconductors is greatly complicated by the interaction of the electrons with each other. This allows one to speak only of the quantum states of the semiconductor crystal as a whole. However, as in metals, so also in semiconductors, a simplified picture is successful. In this, the electron interactions, and other interactions, are first neglected, but are later taken into account as a perturbation. Thus, the electrons are first treated as if they can move through each other; the fact that they collide and deflect each other by virtue of their Coulomb interaction is treated as a perturbation. The transitions are still described within the framework of the single-particle states of the unperturbed problem.

The Effect of Electron–Boson Interactions

Our first approximation is to neglect most (but not all) electron interactions. Later we take into account the two-electron transitions that arise. This two-particle recombination process is referred to as Auger recombination and its inverse as the generation of current carriers by impact ionization. In addition, electrons interact with the radiation and lattice fields and emit or absorb photons or phonons. These electron–boson interactions result in transitions of single electrons in an energy band scheme. We shall attach a superfix ‘S’ (for single-electron) to the recombination coefficients for such processes. These are then denoted by B^{S} or T^{S} depending whether only bands are involved or whether traps are also involved. They are illustrated in Figure 1, where the solid horizontal lines represent the conduction and valence band edges and the dashed line refers to traps. Note that the two-electron transitions do not have the superfix ‘S’, while n and p refer to the electron and hole concentrations; N_0 and N_1 are the concentration of unoccupied and occupied trap states, respectively.

Electron–Electron Interaction in Traps

Electron–electron interactions can at least formally be taken into account in connection with electrons trapped in a center: the spectrum is a function of the number, $r\,(=1,2,3,\ldots,M)$, of electrons captured. The ‘irremovable’ electrons can be included with the ion core. Given that the center captured r electrons (say), it can still be in a variety of quantum states, and they will be denoted by the symbol ℓ, yielding a set of quantum states (ℓ, r) for a center. The energy of such a state, divided by kT to make it dimensionless, is denoted by $\eta(\ell, r)$, yielding the canonical partition function

$$Z_r = \sum_{\ell} \exp[-\eta(\ell, r)] \qquad [1]$$

The location of a center in space will here be considered to be of no significance.

Centers can capture several electrons. This brings in a need for the chemical potentials (or Fermi levels) for r-electron centers. They are here denoted by γ (when divided by kT) and the suffix ‘eq’ denotes an equilibrium value. Thus, the equilibrium probability of finding an r-electron center in state ℓ is given by:

$$P(\ell, r)_{\mathrm{eq}} = \frac{\exp[r\gamma_{\mathrm{eq}} - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_{\mathrm{eq}}) Z_S} \qquad [2]$$

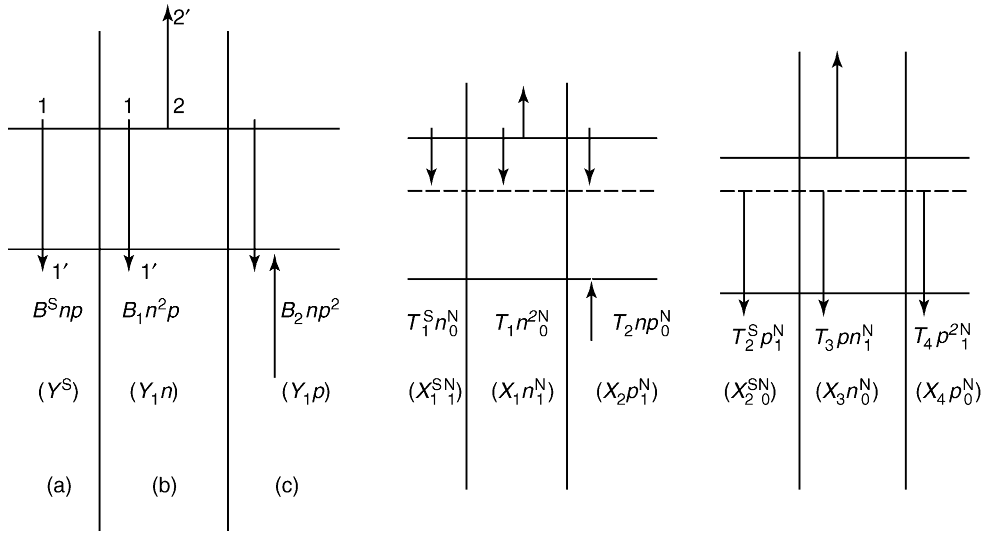

Figure 1 Definition of recombination coefficients. Transition rates per unit volume are stated with each process, and, in brackets, for the reverse process. Thus B_1, B_2, $T_1 - T_4$ refer to Auger processes; B^s, T_1^s, T_2^s refer to single-electron recombination; Y^s, X_1^s, X_2^s refer to carrier generation processes; Y_1, Y_2, $X_1 - X_4$ refer to impact ionization processes. Arrows indicate transitions made by electrons.

One can identify the dominant energies from optical or electrical experiments.

To understand the properties of these centers, suppose we can arrange for the equilibrium Fermi level to rise from the valence band to the conduction band. At first practically no electrons are captured ($r = 0$). As the Fermi level rises the states corresponding to $r = 1$ begin to appear. The states ℓ of the center which correspond to the ground states in equilibrium are always more highly populated than the states ℓ corresponding to excited states, so that we can often confine attention to them. As the Fermi level rises, the ground states for $r = 2$ become more important, and there may now be hardly any centers which have captured fewer electrons. If a larger number of electrons cannot be captured, by the time the Fermi level reaches the conduction band, then states of the center with $r > 2$ can (normally) be neglected as unstable. That this is a satisfactory picture is our second approximation.

Assumptions for Nonequilibrium Statistics

The now much simplified recombination problem is still complex because of the many states available to electrons in bands and centers. A key simplification arises from the fact that it is often possible to talk about a small number of groups of quantum states: let $I\,(= 1, 2, \ldots)$ label quantum states in a group i, let $J\,(= 1, 2, \ldots)$ label quantum states in a group j, etc. Within each group it is supposed that the transitions are much more rapid than they are between groups. In that sense electrons in each group are in equilibrium among themselves. Their quasi-equilibrium is then characterized by what is called a quasi-Fermi level. The dimensionless version, obtained by dividing it by kT, is denoted by γ_i, γ_j, etc., for groups i, j, etc. That this is reasonable is our third assumption. Thus γ_c, γ_v are quasi-Fermi levels which refer to the quasi-equilibrium of the conduction and valence band, respectively, and γ_I to an I-electron center. Groups of states with different quasi-Fermi levels are not in equilibrium with each other.

Recombination problems can now be discussed by neglecting transitions within one (quasi-equilibrium) group, because they proceed at exactly the same rate as their reverse transitions. The number of transition types to be considered is thereby greatly reduced. Away from equilibrium we shall adopt eqn [3] instead of [2] in order to allow for distinct quasi-Fermi levels:

$$P(\ell, r) = \frac{\exp[r\gamma_r - \eta(\ell, r)]}{\sum_{s=0}^{M} (\exp s\gamma_S) Z_S} \qquad [3]$$

Equation [3] is not always correct, but is an approximation arising from the quasi-Fermi level assumption. Among the improved theories are, for example, the ‘cascade’ theories.

The assumption of a quasi-Fermi level for each state of charge of a center implies that all the excited states of an *r*-electron center are populated according to eqn [2].

The Main Recombination Rates

General Results and The Two-Band Case

Recombination and its converse, generation, consist of a transition of an electron from one state to another. The observed rate is the net rate of recombination and is the algebraic sum of the recombination and generation processes. During these processes both total energy and total momentum must be conserved. This is achieved by creation or absorption of photons, or phonons, excitation of secondary electrons, etc.

The transition probability per unit time, S_{IJ}, for a single-electron transition from state I in a band to state J requires that state I should be occupied, with probability p_I say, and state J should be vacant with probability q_J say. The general expression for the average rate of the transition from I to J then takes the form $p_I S_{IJ} q_J$. For the reverse process electron state J has to be occupied and state I vacant. Thus the rate of this reverse process is $p_J S_{JI} q_I$. The net recombination rate per unit volume of the process I to J can then be written as:

$$u_{IJ} = (p_I S_{IJ} q_J - p_J S_{JI} q_I)V^{-1} \qquad [4]$$

By the principle of detailed balance, this expression vanishes at equilibrium. If one puts $X_{JI} \equiv S_{IJ}p_I q_J / S_{JI} p_J q_I$, one has $u_{IJ} = p_J S_{JI} q_I (X_{JI} - 1)V^{-1}$. In equilibrium $X_{JI} \to (X_{JI})_{\text{eq}} = 1$, and the recombination rate is zero.

One can say a little more if one assumes that states I and J are in conduction or valence bands, each with a quasi-Fermi level (divided by kT to make it dimensionless). If these are denoted by γ_{e} and γ_{h}, then $p_I = [1 + \exp(\eta_I - \gamma_{\text{e}})]^{-1}$ for a conduction band. For the total recombination rate per unit volume between the bands i and j one finds:

$$u_{ij} \equiv \left[\sum_{I\in i}\sum_{J\in j} p_J S_{JI} q_I\right]\left[\exp\left(\frac{|e|\phi}{kT}\right) - 1\right]V^{-1} \qquad [5]$$

where ϕ is essentially the difference between the quasi-Fermi levels: $\gamma_{\text{e}} - \gamma_{\text{h}}$. The first factor depends on the bands involved and it has been assumed that the transition probability ratio S_{IJ}/S_{JI} is independent of the excitation.

A transition rate, when multiplied by the charge of current carriers, is a current, and when divided by the area of the surface involved, is a current density. If i and j denote the states of the conduction and valence bands of a semiconductor, excitation independence may often be assumed, and one then expects a current density proportional to $\exp(|e|\phi/kT) - 1$. This is characteristic of the current through pn junctions, metal semiconductor junctions, etc. In these configurations the Fermi level difference between the ends of the device determines the voltage across it. When radiation is involved, however, excitation dependence of some of the parameters introduced above (e.g. S_{JI}) tends to spoil this simple story.

The Case of Defects

So far we have considered only two bands. When a trap is involved matters are rather different. Because of the interactions among the electrons on a center it is not possible to talk of the same level being occupied or vacant. Consequently, identification of forward and reverse processes in terms of levels becomes impossible. Instead one deals with a center, say an *r*-electron center, as a whole; we then need the probability that a given center is an *r*-electron center. For example, the capture of an electron converts an $(r-1)$-electron center into an *r*-electron center. Thus the *p*'s and *q*'s must be replaced by more complicated expressions. It is convenient to denote u_{ij} by u_{cv} in the simple two-band case and its structure is given (with a sign change) by a recombination coefficient:

$$u_{\text{cv}} = B^{\text{S}}np[1 - \exp(\gamma_{\text{h}} - \gamma_{\text{e}})] \equiv B^{\text{S}}np - Y^{\text{S}} \qquad [6]$$

using Figure 1a. Here, n and p are the electron and hole concentrations. Auger effects in Figures 1b and 1c can also be included, at least formally. Then B^{S} has to be replaced in eqn [6] by $B^{\text{S}} + B_1 n + B_2 p$. Analogous replacements are found for recombination processes involving defects.

In the simplest case of one type of localized defect one finds a steady-state recombination rate per unit volume of the form:

$$u = \frac{np - (np)_0}{\tau_{n0}(p + p_1) + \tau_{p0}(n + n_1)} \qquad [7]$$

Here τ_{n0}, τ_{p0} are parameters with the dimension of time and p_1, n_1 are parameters with the dimension of concentration. The recombination increases with the defect concentration, which is actually in the denominators of τ_{n0} and τ_{p0}. This is an old and much used result associated with the names of W Shockley and W T Read.

Radiative Transitions

Spontaneous and Stimulated Emission

Quantum theory was initiated by Planck's law for black-body radiation at temperature T. This gives the spatial energy density as a function of frequency ν in this system as

$$f(\nu, T) = \frac{8\pi\nu^2}{c^3}\frac{h\nu}{\exp(h\nu/kT) - 1} \qquad [8]$$

For low frequencies one finds the Rayleigh–Jeans law which makes [8] proportional to kT in agreement with the classical equipartition theorem. For high temperatures [8] diverges, as required.

The result [8] may also be obtained by writing

$$N = [A + Bf(\nu, T)]N_{\mathrm{u}} \qquad [9]$$

for the emission rate of photons by atoms, where N_{u} is the number of atoms in the upper of two states which are separated by the energy $h\nu$. The first term on the right-hand side is due to the normal decay of an excited state ('spontaneous emission'). The second term refers to additional emission ('stimulated emission') induced by the radiation of frequency $h\nu$ itself.

The first factor in [8] is due to the density of states and the second factor is due to the fact that the energy is considered. If one considers the number of photons of energy $h\nu$ at temperature T one comes up with

$$N_\nu = \left[\exp\frac{h\nu - \mu}{kT} - 1\right]^{-1} \qquad [10]$$
(Bose–Einstein distribution)

Here μ is a possible chemical potential of the radiation which is non-zero only in nonequilibrium situations such as in a semiconductor laser.

In a pn junction the two bulk materials several diffusion lengths away from the junction are approximately in equilibrium even if a modest current is flowing. On one side one has then just one quasi-Fermi level, say μ_I, and on the other side one has just one quasi-Fermi level, say μ_J. Then the difference

$$\mu \equiv \mu_I - \mu_J = q\varphi \qquad [11]$$

corresponds to the applied voltage φ. This is in agreement with eqn [10], in that $\varphi = 0$ implies no current and hence thermal equilibrium is possible.

Statistical mechanics teaches one the rule that, in equilibrium, occupation probabilities of individual quantum states are always less for states of higher energies. This ensures that the total energy of a quantum system converges, the occupation probability p being a kind of convergence factor. However, if one is away from equilibrium, the above rule can be suspended and one can have population inversion.

For $h\nu = \mu$ there is trouble with [10] because the steady-state photon occupation diverges. This does not correspond to a 'death ray', but is the result of imperfect modeling; for example, the leakage of photons from the cavity may have been neglected. A formula of type [10] is also needed in connection with solar cells.

Donor–Acceptor Pair Recombination

A striking demonstration of radiative donor–acceptor transitions in GaP at low temperature (1.6 K) was revealed by sharp lines (Figure 2) at photon energies $h\nu_i$ given by

$$h\nu_i = E_{\mathrm{G}} - E_{\mathrm{A}} - E_{\mathrm{D}} + e^2/\varepsilon R_i\,[-E] \qquad [12]$$

where R_i is the distance between the ith-order nearest neighbors.

The discrete nature of the peaks is due to the fact that the impurities involved settle in general on lattice sites so that only definite separations R_i are possible. The energy gap is E_{G}; the value of $E_{\mathrm{A}} + E_{\mathrm{D}}$ can be inferred from the experimental lines by extrapolation to $R_i \to \infty$. The energy term E is sometimes neglected.

The ZnS phosphors were the first materials in which donor–acceptor radiation was hypothesized. However, it is hard to control its stoichiometry and its impurity content, and it has relatively large carrier mass and hence relatively large impurity activation energies. In such cases spectra such as those shown in Figure 2 are hard or impossible to obtain. This applies in general to many II–VI compounds.

Quantum Efficiency

The quantum efficiency is the radiative recombination as a fraction of the total recombination. It can also be regarded as the average number of electrons produced per incident photon. Consider an intrinsic semiconductor, i.e., one in which the electron and hole concentrations are equal. Then with the notation of Figure 1 the radiative band–band recombination rate is $B^{\mathrm{s}}n^2$. The nonradiative rate is $(B_1 + B_2)n^3$. We shall add another nonradiative rate, An say, proportional to the injected carrier density $n \sim p$. The quantum efficiency is then

$$\eta = \frac{B^{\mathrm{s}}n^2}{An + (B_1 + B_2)n^3 + B^{\mathrm{s}}n^2} \qquad [13]$$

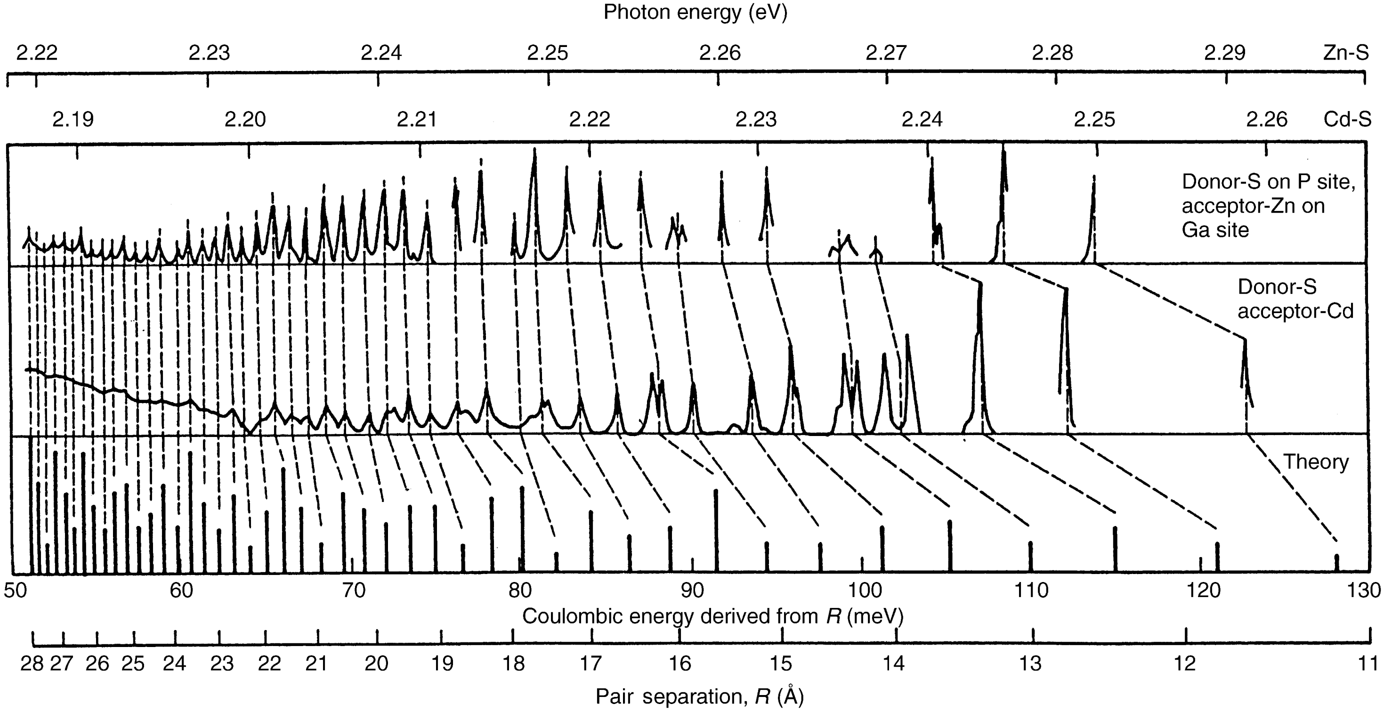

Figure 2 Comparison of the positions and intensities of the sharp line spectra at 1.6 K corresponding to both ZnS and CdS acceptor–donor pairs with the predicted pair distribution. The lower scales show the pair separation (R) and the Coulombic energy derived from R. The emission energy scales for the two measured spectra are shown at the top. Reproduced with Gershenzon M, Logan RA, Nelson DF, *et al.* (1968) *Proceedings of the International Conference on Luminescence*. Budapest, Hungary: Akademia Kiada.

Traps are here neglected, and one finds a maximum

$$\eta_{\max} = \left\{1 + \frac{2}{B^s} A^{1/2}(B_1 + B_2)^{1/2}\right\}^{-1} \qquad [14]$$

For a high-quality epitaxial AlGaAs/GaAs double heterostructure of great purity we may take

$$A \sim 0.5 \times 10^6\ \mathrm{s}^{-1},\ \ B \sim 10^{-10}\ \mathrm{cm}^3\ \mathrm{s}^{-1},$$
$$B_1 + B_2 \sim 10^{-29}\ \mathrm{cm}^6\ \mathrm{s}^{-1} \qquad [15]$$

whence $n_1 \sim 2 \times 10^{17}\ \mathrm{cm}^{-3}$, $\eta_{\max} \sim 0.96$.

Detailed Balance

It is clear that the radiative recombination rate from a material should be obtainable in terms of its optical 'constants', the absorption coefficient $\alpha(\lambda)$ and the refractive index $\mu(\lambda)$. Both are functions of the wavelength. This connection can be formalized by comparing the optical absorption process with the emission process and this can be done by using the principle of detailed balance. This equates the rate of disappearance by absorption of photons with the rate of production of photons by radiative recombination. The radiative recombination rate can thus be obtained as an integral over the optical functions. This relationship, pioneered by van Roosbroeck and Shockley in 1954, has been used a great deal because the optical quantities are often known with some accuracy.

This result corresponds to balancing the rate $B^s np$ and the rate Y^s in Figure 1a. Analogous detailed balance results are obtainable for the other pairs of processes in Figure 1. Thus, the Auger recombination rate can be related to an integral over the impact ionization rate. However, this connection is less useful since impact ionization data are generally less well known than the optical absorption information.

Nonradiative Processes

Auger Effects

The Auger effect was discovered in 1925 in gases by Pierre Auger. An atom is ionized in an inner shell. An electron drops into the vacancy from a higher orbit and a second electron takes up the energy which is used to eject it from the atom. In solids the effect is roughly analogous. One of its characteristics is that it is not radiative. Since radiation is what is seen in many experiments, the Auger effect is suspected if and when there is less radiation than expected in the first place. It is rather harder to investigate than radiative transitions.

The effect has proved to be important as it limits the performance of semiconductor lasers, light-emitting diodes and solar cells, and it can be crucial in transistors and similar devices whose performance is governed by lifetimes. When heavy doping is required, as it is in the drive towards microminiaturization, its importance tends to increase, since the Auger recombination rate behaves roughly as n^2p or p^2n (see Figure 1) compared with the radiative rate which behaves more like np. The inverse process is impact ionization, and is important in the photodiode, the impact avalanche transit time (IMPATT) diode, and hot electron devices.

Impact ionization can be regarded as an autocatalytic reaction of order one:

$$\mathrm{e} \rightarrow 2\mathrm{e} + \mathrm{h} \quad \text{or} \quad \mathrm{h} \rightarrow 2\mathrm{h} + \mathrm{e} \qquad [16]$$

i.e., one extra particle is produced of the type present in the first place. This is a key feature for impact-induced nonequilibrium phase transitions in semiconductors.

In the theory one needs wavefunctions for four states of two electrons which are relevant after the many-electron problem has been reduced to a two-electron problem. In a matrix element calculation the four wavevectors imply a 12-fold integration in $\boldsymbol{k}$-space to cover all possible states of the two electrons. However, momentum and energy conservation usually reduce this to an eight-fold integration. Such a calculation is hard, for it requires (1) good wavefunctions and (2) accurate integrations.

Here we shall merely concentrate on the broad principles. In addition, the many-electron nature of the problem implies that another approximation is often inherent in the treatment apart from the use of perturbation theory. This is clear if one considers that the electron interactions are screened twice: once by an exponential screening factor and a second time by the dielectric constant. So there is some double counting, and the treatment of the effect as uncorrelated electronic transitions mediated by screened Coulomb interactions is another approximation. The collective effects that enter require more sophisticated field theoretic methods which take care of electron–hole correlations, plasmon effects and the effect of free excitons, the so-called excitonic Auger effect. These calculations broadly confirm the results obtained by the simpler methods used for energy gaps large compared with the plasmon energies, provided the high-frequency dielectric constant is used and doping is not too heavy. Significant corrections are required for the narrow gap lead compounds, for example, and a considerable computational effort is required.

Let us now consider lifetimes for high carrier concentrations, as limited by band–band processes. They are best studied by looking at the emitted radiation. In this connection we recall the striking rise of the threshold current density J in a semiconductor laser as the material is changed from a direct material like GaAs to an indirect material, by mixing it with a compound such as GaP. Figure 3 shows this spectacular rise for $\mathrm{GaAs}_{1-x}P_x$ at 77 K near $x = 0.38$. It is due to the drop in the radiative transition probabilities, as the substance becomes indirect, thus requiring a higher current density for threshold. So we should start with band–band processes in direct materials as most favorable for the experimentally accessible radiative transitions.

The essential point here is that the injected carrier density behaves as:

$$n_{\mathrm{inj}} = \frac{J\tau}{qd} \sim \frac{(5 \times 10^3\ \mathrm{A\ cm^{-2}})(10^{-9}\ \mathrm{s})}{(1.6 \times 10^{-19}\ \mathrm{C})(10^{-4}\ \mathrm{cm})} \sim 10^{17}\ \mathrm{cm^{-3}} \qquad [17]$$

where J/q is the particle current density at threshold, τ is the lifetime of the carriers, and d is the thickness of the active layer. As active layers are made thinner n_{inj} increases to $10^{19}\ \mathrm{cm^{-3}}$ or so, far in excess of the defect concentration in the (undoped) material. This brings in band–band Auger effects. These have also been invoked to explain the undesirable increase in J with temperature. Thus the interest in direct band–band Auger effects in III–V compounds is fuelled by the need for better and smaller opto-electronic devices which work at long wavelengths (1.3–1.5 μm). It matches the interest in indirect band–band Auger and impurity Auger effects in silicon due to the importance of heavy doping in VLSI (very large scale integration) and transistor technology. Figure 4 shows a typical Auger process, called CHHS, as the conduction band (C) and the split-off band (S) are involved. Two relevant states are in the heavy hole band (H). State 2′ is referred to as the Auger carrier, as it has more kinetic energy than the others.

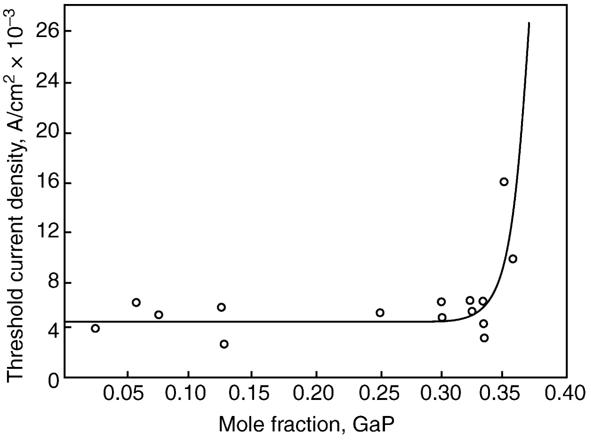

Figure 3 Lowest values of laser threshold current density at 77 K as a function of mole fraction of GaP for $\mathrm{Ga(As}_{1-x}P_x)$ junction lasers. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. *Solid-State Electronics* 9: 735–742, with permission from Elsevier.

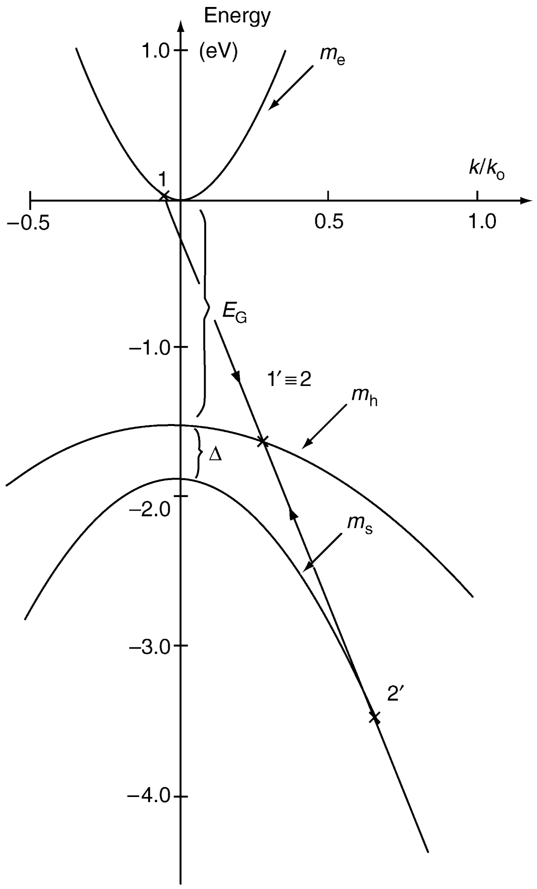

Figure 4 Conduction band, heavy hole and split-off valence bands of GaAs, all treated as parabolic with $\hbar^2 k_0^2/2m_h \equiv E_G - \Delta$. x denotes a quadruplet of states for a most probable transition. The two states in the heavy hole band are not shown separately. The arrows indicate electron transitions. Reproduced from Neill CJ, Stillman GE, Sirkis MD, *et al.* (1966) Gallium arsenide-phosphide: Crystal, diffusion and laser properties. Solid-State Electronics 9: 735–742, with permission from Elsevier.

Impact Ionization

The CHHS and CHCC processes and their inverses can be written as

$$\mathrm{e_C} + 2\mathrm{h_H} \leftrightarrow \mathrm{h_s} \quad 2\mathrm{e_C} + \mathrm{h_H} \leftrightarrow \mathrm{e_C} \qquad [18]$$

where the suffix refers to the band. Viewed from left to right these are Auger processes. Viewed from right to left they are autocatalytic and impact ionizations.

Such processes are important for the impact-induced nonequilibrium phase transition in semiconductors. Also reaction rates with autocatalytic elements imply nonlinear equations in the concentrations and this gives rise to much interesting behavior as regards stability and bifurcation phenomena.

Momentum and energy conservation are very restrictive conditions on the four states involved in [18], and it is easily seen that they cannot normally be satisfied if in a direct band semiconductor the recombining electron drops from the band minimum to the valence band maximum. This effect raises all the kinetic energies of possible processes. The Auger electron (on the right-hand sides of [18]) must also have a minimum energy in order that the impact-ionization process can proceed. This leads to an activation energy for the process. For Figure 4, for example, and in the case of nondegeneracy, the energetic hole has a kinetic energy at threshold of

$$E_{\mathrm{th}} = \frac{m_{\mathrm{e}} + 2m_{\mathrm{h}}}{m_{\mathrm{e}} + 2m_{\mathrm{h}} - m_{\mathrm{s}}}(E_{\mathrm{G}} - \Delta) \quad \text{(CHSS)} \qquad [19]$$

As the band gap increases we see that E_{th} goes up and so the Auger rate for simple parabolic bands decreases. Values of E_{th} for other transitions can be obtained from equation [19].

The total kinetic or threshold energy E_{th} can be converted to an activation energy by subtracting the basic energy which must under all conditions be part of the Auger particle energy. In the case of eqn [19] one finds for $E_G > \Delta$:

$$\begin{aligned} E_{\mathrm{a}} &= E_{\mathrm{th}} - (E_{\mathrm{G}} - \Delta) \\ &= \frac{m_{\mathrm{s}}}{m_{\mathrm{e}} + 2m_{\mathrm{h}} - m_{\mathrm{s}}}(E_{\mathrm{G}} - \Delta) \end{aligned} \qquad [20]$$

This can result in a strong temperature dependence in accordance with an Arrhenius factor $\exp(-E_{\mathrm{th}}/kT)$. However, this is again lost, and the direct band–band Auger effect will certainly be important and only weakly temperature dependent, if $E_{\mathrm{G}} \sim \Delta$. This can occur, for example, for InAs, GaSh and their solid solutions.

The connection between Auger processes and impact ionization has also been formalized. Of all Auger quadruplets of states in a set of nondegenerate bands, the most probable Auger transition involves the quadruplet, which yields the threshold for impact ionization. The three crosses in Figure 4 indicate such a quadruplet (the central cross represents two states).

Identification of Auger Effects

How does one know that an Auger effect has occurred? In pure but highly excited materials there is the original solid-state Auger effect which leads to a lifetime broadening of the electronic states at the band edge. For a semiconductor this effect causes fuzziness in the band edge which is a contributory factor to the overall bandgap shrinkage. The mechanism is illustrated in Figure 5, which shows how the lifetime of a vacant electron state near the band edge is shortened by the Auger processes within this band. Under normal conditions this effect is small in a semiconductor. But under degenerate conditions the effect leads by lifetime broadening to a low-energy tail in emission.

The blanket term ‘Auger effect’ for all Coulombically excited two-particle transitions has been used here. In semiconductor device work the ‘Auger effect’ has become associated with transitions in which one electron bridges an energy gap. But there is no need to limit the concept in this way. Its competition with radiative effects means that it is often detrimental and a full discussion of means of suppressing it has recently been given by Pidgeon *et al.*

More spectacular and more convincing is the *detection* of the energetic Auger electron or hole. For this purpose one may look for the weak luminescence emitted when this carrier recombines radiatively. This has been done for the band–band process in Si and for the band–impurity process

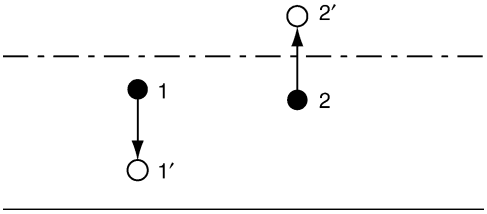

Figure 5 The filling of a state *k* by the Auger effect. The arrows indicate electron transitions.

in GaAs. It leads to so-called $2E_G$-emission. Its rate-limiting step is the rate of the Auger process per unit volume, B_1n^2p, which populates the high-energy level. The radiative process then proceeds at a rate $B'_1n^2p^2$ per unit volume with $B'_1 \sim 10^{-54}$–10^{-51} cm^9 s^{-1}.

More usual is the identification of the band–band Auger recombination mechanism from the minority carrier lifetime τ, which behaves as

$$\frac{1}{\tau_p} = B_1n^2 \quad \text{or} \quad \frac{1}{\tau_n} = B_2P^2 \qquad [21]$$

where B_1 and B_2 are Auger coefficients when the Auger particle is an electron or hole, respectively. A more complete expression allows also for trapping processes.

The departure from parabolic bands plays an important part for the energetic Auger particle (in state 2′). Theory shows that when this effect is taken into account, it tends to lower the band–band Auger coefficient. In fact in GaAs the CHCC process is ruled out altogether at 0 K for nonparabolic bands, suggesting that it should be an ideal material for radiative transitions. However, the possibility of phonon participation brings the effect back again, though it is still comparatively weak. The difficulty of conserving electron energy and momentum, which gives rise to the activation energies (eqn [20]) is greatly alleviated if phonons can take up some of the momentum in a direct gap semiconductor.

We have seen that the activation energy barrier against the band–band Auger effect can be overcome by a suitable disposition ($E_G = \Delta$) of the three direct bands and by phonon participation. It can also be overcome if there is an indirect minimum near a Brillouin zone edge at about half the direct energy gap.

Acknowledgments

The author wishes to acknowledge support from NATO under their grant PST.CLG 975758.

See also

Semiconductor Physics: Band Structure and Optical Properties; Excitons; Impurities and Defects; Infrared Lattice Properties; Outline of Basic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Fraser DA (1986) *The Physics of Semiconductor Devices*, 4th edn. Oxford, UK: Clarendon Press.

Hangleiter A (1985) Experimental proof of impurity Auger recombination in silicon. *Physical Review Letter* 55: 2976–2978.

Harrison D, Abram RA and Brand S (1999) Characteristics of impact ionization rates in direct and indirect gap semiconductors. *Journal of Applied Physics* 85: 8186.

Landsberg PT (1991) (paperback edn., 2003) *Recombination in Semiconductors*. Cambridge, UK: Cambridge University Press.

Nimtz G (1980) Recombination in narrow gap semiconductors. *Physics Reports* 63: 265.

Pidgeon CR, Ciesla CM and Murdin BN (1997) Suppression of non-radiative processes in semiconductor mid-infrared emitters and detectors. *Progress in Quantum Electronics* 21: 361–419.

Spin Transport and Relaxation in Semiconductors; Spintronics

M E Flatté, University of Iowa, Iowa City, IA, USA

D D Awschalom, University of California,
Santa Barbara, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Control of the generation and transport of coherent electron spin in semiconductors suggests new ways to probe the fundamentals of quantum decoherence in solids and to explore device applications that rely on coherence, including quantum computation. The focus here will be on the physical phenomena that govern spin transport and relaxation, with a particular emphasis on those phenomena that are amenable to direct manipulation. Manipulation of material properties in general has been taken to its most advanced level in the design of semiconductor electronic devices, in which operations are performed by perturbing the motion of electrons through their electric charge; hence, ‘spintronics’ is a term introduced to describe both the manipulation of electrons by their spin and control of their spin properties, such as *g*-factors.

The study of the evolution of spin coherence first requires the generation and detection of spin coherence. Here the ultrafast optical technique for generating and detecting spin-polarized populations will be emphasized. The next logical component to treat will be the persistence of spin coherence, followed by the transport of spin coherence from one region of a material to another. Finally the direct ultrafast manipulation of spin-polarized populations will be described. These components are likely to play a prominent role in any attempt to apply the fundamental physical phenomena of spin coherence to specific device technologies.

Optical Generation of Spin-Polarized Distributions

Optical pulses with definite circular polarization can be used to generate spin-polarized distributions of excited electrons in crystals. In order for this process to work one of the energy bands involved in the optical transition must be characterized by a substantial spin–orbit interaction. The selection rules for circularly polarized photons require a change in the orbital angular momentum of the electron projected along the propagation direction of the photon. For the optical transitions of an atom in free space this azimuthal orbital angular momentum l_z must increase or decrease by $\hbar$, permitting the coupling of the p-state with $l_z = 1$ to an s-state for left circular polarization, and of the p-state with $l_z = -1$ to that s-state for right circular polarization. Thus the optical selection rules pick out a particular orbital angular momentum projection.

These selection rules by themselves do not explain how spin-polarized populations are generated, for both spin directions are possible for each value of l_z. In the presence of spin–orbit interaction, however, the energy of a state with a particular l_z is correlated with the spin orientation. The six p-states, which are degenerate in the absence of spin–orbit interaction, split into two groups: one of four degenerate states and the other of two degenerate states. The higher-energy group of four degenerate states consists of two states with **L** and **S** parallel and two states with a one-third probability for **L** and **S** to be antiparallel and a two-thirds probability for **L** and **S** to be perpendicular. Thus for a transition involving the four degenerate states, a circularly polarized photon (that selects a particular initial l_z) is three times more likely to generate a carrier with spin parallel to the initial l_z than antiparallel to it.

In a solid, although full rotational symmetry is lost, the band symmetries preserve some of these selection rules. In crystals of tetrahedral symmetry, such as zincblende semiconductors, the selection rules for transitions from the valence to the conduction band are identical to the selection rules between p and s states for the atom in free space. For example, the valence heavy-hole states are those for which **L** and **S** are parallel. Thus by setting the optical photon energy to the fundamental gap, the resulting excited conduction electrons and valence holes will be spin polarized. In quantum wells, where the energy of the heavy hole is split from that of the light hole by quantum confinement (and, sometimes, strain), fully spin-polarized carrier distributions are possible. Commonly used group IV, III–V, and II–VI semiconductors have a sufficient spin–orbit interaction to permit the straightforward optical generation of spin-polarized distributions. Shown in Figure 1 are the selection rules

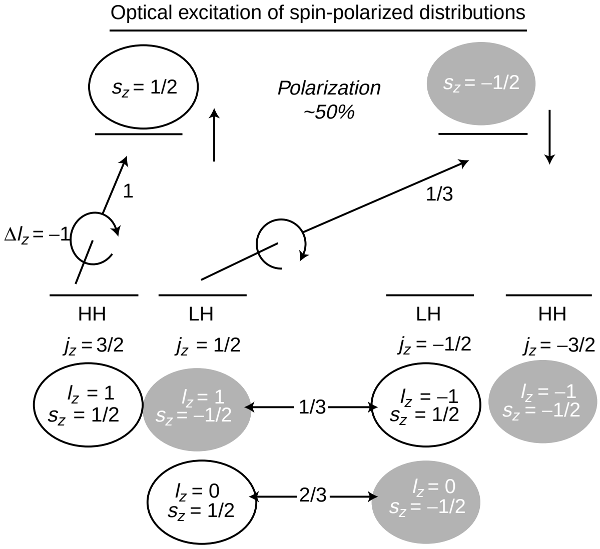

Figure 1 Selection rules for bulk zincblende crystals. Circularly polarized light changes the azimuthal orbital quantum number l_z by $+1$ or -1. Here $\Delta l_z = -1$ is shown. The orbital and spin quantum numbers for the two degenerate conduction band edge states and the four degenerate valence band edge states are shown – unshaded areas correspond to spin up and shaded to spin down. A transition with $\Delta l_z = -1$ couples one heavy-hole state to the conduction band up-spin state, and one light-hole state to the conduction band down-spin state. The orbital angular momentum of the heavy-hole state is entirely parallel to its spin angular momentum. In contrast only one-third of the light-hole state has antiparallel orbital and spin angular momentum. The rest of the light-hole state has perpendicular orbital and spin angular momentum. Thus the oscillator strength of the transition creating the conduction up-spin state is three times larger than that of the transition creating the conduction down-spin state, yielding a 50% spin polarized distribution. In quantum wells for which the heavy–light splitting is larger than the linewidth of the light source driving the transition, only the heavy-hole transition is excited and the spin polarization approaches 100%.

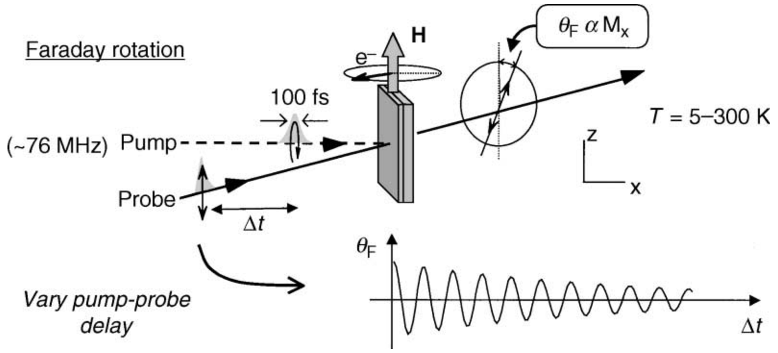

Figure 2 Diagram of the experimental geometry for time-resolved Faraday rotation. A pulse of circularly polarized light illuminates the sample, followed by a probe of linearly polarized light. The linearly polarized light can be considered as a sum of equal parts left and right circularly polarized light. The spin-polarization of the electron population (along the photon propagation direction) induces a difference in index of refraction for left and right circularly polarized light (circular dichroism). This produces a phase change between the left and right polarized components of the probe, which is measured as a rotation (θ_F) of the linear polarization direction. Monitoring the angle of rotation of the probe as a function of time directly indicates the spin polarization along the propagation axis of the photon as a function of time. Reproduced with permission from Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer

which permit circularly polarized optical pulses to generate spin-polarized carriers in the conduction and valence band.

The evolution of a coherent spin-polarized population can be monitored through optical means as well. If the optical transitions used to generate the spin-polarized population are pumped hard enough, the transitions will be bleached. By measuring the polarization dependence of this bleaching the decay of the spin-polarized population can be determined. This can be done either by monitoring the transmission of time-delayed probe pulses of circularly polarized light, or by monitoring the rotation angle (Faraday rotation) of a pulse of linearly polarized light. The first of these depends directly on the imaginary component of the spin-polarized dielectric function, whereas the second depends directly on the real components. Shown in Figure 2 is the geometry for observing Faraday rotation from a spin-polarized distribution in the presence of an applied magnetic field perpendicular to the spin polarization.

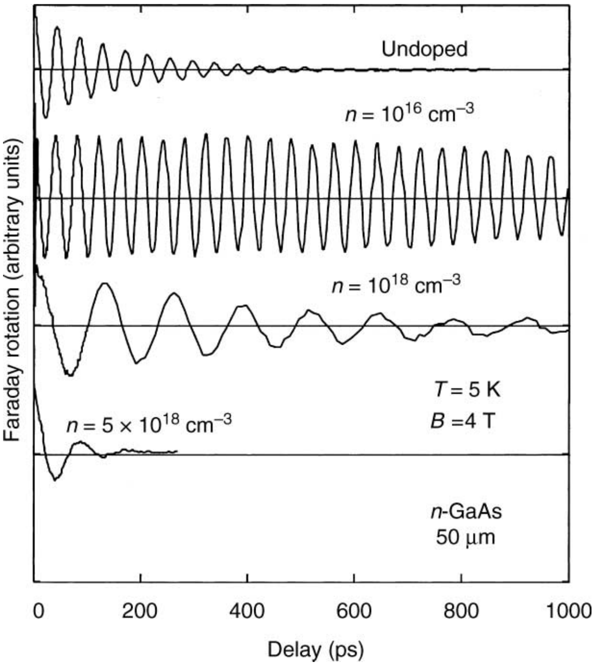

Figure 3 Time-resolved Faraday rotation measurements on bulk GaAs at several doping levels, indicating a sensitive doping dependence of the transverse spin coherence time T_2. T_2 for ideal doping densities can exceed 150 ns at 5 K. The magnetic field is 4 T, and the sample is 50 μm thick. Reproduced with permission from Kikkawa JM and Awschalom DD (1998) Resonant spin amplification in *n*-type GaAs. *Physical Review Letters* 80: 4313–4316.

Mobile Electron Spin Time Evolution: Decoherence and Precession

Two natural phenomena in the time evolution of a coherent spin-polarized population are decay of that population (decoherence) and rotation of the macroscopic magnetization in the presence of an applied magnetic field (precession). The time-scale of decay of the spin coherence can be very much longer than time-scales for the decay of other coherent phenomena in solids. Typical orbital coherence times, for example, are 100 fs, whereas spin coherence times can exceed microseconds. In Figure 3 are shown time-resolved Faraday rotation data for three bulk GaAs crystals with different doping levels. The oscillations are characteristic of the spin precession, and the oscillation frequency is the Larmor frequency of the electron spin in this crystal. The crystal with the middle doping level, 10^{16} cm^{-3}, has oscillations that persist for the longest time, and therefore this material has the longest spin coherence time.

The dominant mechanism for the decay of these spin polarized distributions works via the same spin–orbit interaction that permits the generation and detection of these spin-polarized distributions. For mobile electrons in homogeneous environments, the mechanisms of spin coherence relate to the crystal momentum-dependence of the eigenstates of the electronic system. Because of time reversal invariance the electronic states with zero momentum (at the Γ point in the Brillouin zone) are doubly degenerate. The electronic states for momentum $\mathbf{k}$ near the Γ point occur in nearly degenerate pairs. Although it is common usage to refer to these eigenstates as degenerate spin eigenstates (correct in the absence of spin–orbit interaction), with spin–orbit interaction the states are not spin eigenstates and are not necessarily degenerate. As a consequence, when a spin-polarized carrier is scattered by ordinary, non-spin-dependent scattering from one momentum state $\mathbf{k}$ to another $\mathbf{k}'$, and then back again to $\mathbf{k}$, the effective orientation of the carrier spin will have rotated by an angle dependent on the intermediate momentum state $\mathbf{k}'$. For a population of carriers in thermal equilibrium this process will lead to misorientation of the individual spin moments from the mean, and thus a decoherence of the spin population. This process is characterized by an increasing decoherence rate with increasing orbital scattering. Thus the dirtier a material is made, the faster the spin orientations decohere.

In crystals with inversion symmetry the pairs of nearly degenerate states are precisely degenerate. Lack of inversion symmetry, and the consequential energy splitting, can be viewed as an effective (crystal) magnetic field, in which the approximate spin eigenstates precess. The magnitude and direction of this effective field depends on the orbital momentum $\mathbf{k}$, and because of time reversal invariance, the effective fields for $\mathbf{k}$ and $-\mathbf{k}$ are opposite. Thus an initially spin-polarized population will dephase due to experiencing different effective magnetic fields. Ordinary spin-independent scattering can now be seen as randomly changing the precession axis for the spin in this effective field. Once the information about the precession axis orientation has been lost the dephasing is not reversible and the spin orientations have decohered. As the precession angle increases linearly between scattering events, the longer the time between such scattering events the faster the spin population decoheres. Thus for this process the cleaner a material is made, the faster the spin orientations decohere. This precessional decoherence process dominates for mobile conduction electrons in bulk and quantum-well III–V semiconductors near room temperature.

The influence of the spin–orbit interaction is far stronger for the valence electrons than the conduction electrons for two reasons. First, as the effective magnetic fields come from relativistic transformations of internal electric fields, and the electric fields are largest near the nucleus, the valence electrons, which are closer to the nucleus than the conduction electrons, experience a stronger spin–orbit interaction. Second, the valence electrons have a dominant p-like character, whereas the conduction electrons have dominant s-like character (and thus lack orbital angular momentum to be coupled to the spin). Thus the spin coherence times of valence electrons (or holes) are commonly much shorter than the coherence times of the conduction electrons. Spin-flip exchange between conduction electrons and valence holes can therefore be the dominant source of decoherence for conduction electrons. This process can dominate at low temperatures in clean III–V semiconductors, and even up to room temperature in III–V quantum wells grown along the unusual (110) growth direction.

As a consequence one would expect the spin coherence times of electrons in n-doped semiconductors to be longer than those of electrons in p-doped or undoped semiconductors (in which excess holes must also be optically generated). This expected trend is present in Figure 3, for the n-doped materials have longer spin coherence times than the semi-insulating material. However, as the effect of the spin–orbit interaction increases with increasing momentum $\mathbf{k}$, the spin coherence times of the more heavily doped material are shorter than those of the lightly doped material. The doping corresponding to the longest spin coherence times may lie near the metal–insulator transition, although the theory of spin coherence times on the insulating side of that transition is not as well developed as on the metallic side.

Transport and Manipulation of Spin Coherence

An essential component of the motion of spin coherence in semiconductors is the response of carriers to electric fields. Outside of very unusual electronic states in semiconductors, such as the quantum Hall state, there is no spin–charge separation of elementary carriers. The carriers in the semiconductors will move under the influence of an electric field, and if they happen to be spin-polarized, then a spin current will exist. Although there are numerous possible regimes of spin transport (at least as many as there are regimes of charge transport), only two will be presented here, both in the low-field diffusive limit.

Mobilities and diffusion constants are the material properties that describe the low-field charge transport in a homogeneous material. Spin transport can be described similarly, and the spin mobility and diffusion constant can be related to those of charge motion. The relationship between the two emerges from the nature of carrier packets that can be maintained under low-field conditions. Deviations from local charge neutrality cause very large space-charge fields in the semiconductor, and thus under low-field conditions approximate local charge neutrality must be maintained. Shown in Figure 4 are charge and spin packets in an n-doped semiconductor. The charge polarization packet, incorporating extra conduction electrons and valence holes, is shown in Figure 4a. The electron spin polarization packet shown in Figure 4b is qualitatively different from the charge packet as well as the hole spin polarization packet (shown in Figure 4c). Only the electron spin polarization packet does not incorporate both conduction electrons and valence holes.

The difference in packet structure profoundly changes the mobility and diffusion properties of the packet. For packets consisting of both electrons and holes the low-conductivity carrier species dominates the packet motion (in this case, the holes). For the electron spin packet, however, majority carriers dominate the packet motion. Thus motion of this spin packet is very rapid, and also sensitive to the higher Fermi temperature of electrons relative to holes. Experimentally this has been seen by optically creating a local spin-polarized packet of conduction electrons (an electron spin packet) in a thin sample of n-doped GaAs. The application of a lateral electric field drags that spin packet, and an optical probe measures the transport distance (thus the mobility) and the packet spread (yielding the diffusion constant). Such an approach is very similar in concept to the electrical probe of charge packet motion developed by Shockley and Haynes to measure minority hole mobility and diffusion in n-doped Ge.

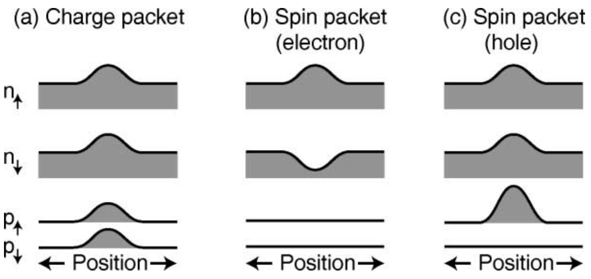

Figure 4 Inhomogeneous carrier packets in n-doped semiconductors in the low-field regime. The charge polarization packet (a) is constructed of approximately equal numbers of electrons and holes. The electron spin polarization packet (b) can be constructed without any excess holes, whereas the hole spin polarization packet (c) requires both excess electrons and excess holes. The presence of minority carriers in a packet significantly alters the mobility and diffusion properties of the packet.

The picture of approximately charge-neutral packets is very helpful for describing spin motion in homogeneous materials, but most semiconductor devices are based on inhomogeneous configurations of semiconducting material. For example, the most ubiquitous charge-based semiconductor device is the p–n diode, whose nonlinear current–voltage characteristics have been used as building blocks for numerous other devices. Such devices rely on the ability of a current to traverse an interface between two dissimilar materials – either different in doping density or different in host structure. Experiments in metallic ferromagnetic systems indicated that, whereas charge current was easily passed from one ferromagnetic metal to another through an interface, the same was not always true for the spin current. Differences in band structure or interface roughness are usually expected in such systems to strongly diminish the efficiency of spin current transfer across the interface.

Semiconductors, however, offer a possible new regime. The bulk band structures of most III–V semiconductors are very compatible – the variations in bandgap may be substantial, but the orbital wavefunctions are very similar. Current growth techniques permit the fabrication of atomically smooth interfaces between two semiconductors, so long as the lattice constants are almost identical. Thus one might expect the spin current to be transferred across such interfaces without tremendous degradation.

The transfer process for spin current across the GaAs/ZnSe interface has been investigated in detail. Here the ZnSe material was n-doped, and GaAs layers that were n-doped, p-doped, and semi-insulating were investigated. Electron spin polarization was generated optically in the GaAs layer. For n-doped GaAs in the absence of an electric field only a small signal for spin-polarized electrons in the ZnSe was detected. An applied electric field, however, which drove the electrons within the GaAs towards the interface, produced a substantial spin-polarized population in the ZnSe. This indicates that spin-polarized packets can be created in one semiconductor, and then controllably moved into a second semiconductor with an applied electric field. This ability may become a key component of manipulating spin in semiconductors.

One reason the p–n diode is a paradigm of charge-based semiconducting electronics is that it exhibits the crucial phenomenon of minority spin injection. Manipulation of minority spin injection permits great control of charge current, for its amplitude is very sensitive to barriers (such as those introduced by gates). Spin-polarized minority carrier injection has not been shown yet experimentally, however the complementary process of minority carrier extraction has. This process corresponds to optically generating spin-polarized electrons in p-doped GaAs, where the electrons are minority carriers. The built-in electric field of the p–n junction drives these electrons out of the GaAs into the n-doped ZnSe with great efficiency. Thus one may imagine that spin-polarized minority carrier injection is plausible in these systems.

A hybrid of these two situations is the spin-polarized light-emitting diode. Here spin-polarized carriers (either electrons or holes) are electrically injected into the insulating region of p–i–n structure. In the insulating region they recombine with unpolarized carriers of the other species and emit light. Due to the selection rules described before, the emitted light is circularly polarized, and the degree of circular polarization is an indication of the spin-polarization of the recombining carriers. This device demonstrates that large numbers of spin-polarized carriers can be accumulated in the insulating region, although the spin-polarized carriers are not minority carriers.

Methods of manipulating electron spin coherence externally without electric fields are also under investigation. One demonstrated technique relies on the shift of energy levels in a semiconductor under illumination by laser light with frequencies below the fundamental bandgap. When the laser light is circularly polarized this ‘ac Stark effect’ shifts the energy levels of spin-up states and spin-down states differently, generating an effective magnetic field. Rotations of electron spin by angles greater than $\pi/2$ have been demonstrated in this ultrafast effective magnetic field in semiconductor quantum wells.

Transfer of Spin Coherence from Mobile Electrons to Other Spin Systems

Another fundamental phenomenon in spin coherence is the transfer of spin coherence from one system to another. If one spin system has an extremely long coherence time, this transfer can be viewed as establishing a reservoir of spin coherence that can be drawn on at later times. Explorations of transfer phenomena have focused on establishing spin coherence in a system that is coupled to the conduction electrons, such as core electron polarization or nuclear polarization. The polarization of these other spin systems can be detected because, when polarized, they exert an effective field on the conduction electrons. For example, when the nuclear system is polarized it exerts an effective field on the electrons through the hyperfine coupling.

It is also possible to polarize these other systems indirectly through their interaction with the conduction electron system. When the conduction electron system is driven far from equilibrium by optically pumping the spin polarization, the coupled spin systems also become spin polarized. In the case of the nuclear spin system this effect is the Overhauser effect. While continually generating conduction electron spin polarization with optical pumping the polarization of the nuclear system begins to rise with a time-scale characterized by the coupling between the nuclear spins and the conduction electron spins (hyperfine coupling). Because of the substantial difference between the nuclear spin coherence times parallel to and perpendicular to the applied magnetic field, the effective nuclear polarization is generated parallel to the applied magnetic field independent of the orientation of the conduction electron spin polarization. In bulk GaAs at 5 K the percentage polarization achieved was a few per cent, yielding an effective field through the hyperfine interaction of ~ 0.4 T. These nuclear polarizations and effective fields are increased over an order of magnitude in certain quantum wells, suggesting that nanostructure design can be used to control the effective fields. The induced nuclear polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. If the optical pumping is turned off, the conduction electron spin polarization decays away rapidly, whereas the nuclear spin polarization persists for time-scales characteristic of the nuclear longitudinal spin coherence time, on the order of minutes. This nuclear hyper-polarization, and its manipulation through an optically pumped time-dependent conduction electron spin polarization, suggests a possible general technique of ultrafast all-optical nuclear magnetic resonance in solids.

A similar spin-coherence storage effect has been obtained for core electron polarization in Mn-doped ZnSe quantum wells. Again, the optically pumped conduction electron spin polarization generates a nonequilibrium spin polarization of the paramagnetic core electron spins (3d in the case of Mn) through the core-conduction electron coupling. This core electron spin polarization persists long after the conduction electron spin polarization (and, in undoped systems, long after the conduction electron density) has decayed away. The magnetic polarization of the paramagnetic Mn spins dephases over the very long time-scales determined by spin–lattice relaxation, and thus is an interesting avenue towards information storage.

The above systems are not ferromagnetic – the spin systems involved in the process are paramagnetic spin systems coupled to the conduction electrons, which are 'super-polarized' through the conduction electron spin polarization. An alternative approach to manipulating the conduction electrons has recently been demonstrated using a thin layer of a ferromagnetic material placed on top of n-doped GaAs. Here the presence of this ferromagnetic material, even if only a few nanometers thick, leads to a remarkable enhancement of the nuclear spin polarization in the GaAs. For example in the presence of a mere 1000 G applied magnetic field, the ferromagnetic layer generates an effective field on the electrons from the nuclei of up to 9000 G. This effective field also depends on the orientation of the magnetism in the ferromagnetic layer. This intriguing effect has yet to be fully explored, but may yield a new type of magnetic 'gate' on the mobile electrons in a semiconductor without a direct form of electrical spin injection.

The Future of Semiconductor Spintronic Devices

Semiconductor charge-based electronics (e.g., transistors and integrated circuit chips) and magnetoelectronics (e.g., hard disk read heads) are two of the most successful electronic technologies of the twentieth century. A union of elements of these two technologies is anticipated in the area of semiconductor spintronics. Immediate approaches involve placing hybrid magnetic elements within known semiconductor devices. Semiconductor spintronic devices, however, are unlikely to compete with the most successful devices from each of these two areas – the semiconductor transistor is an extremely effective and efficient device, as is the magnetoelectronic hard disk read head. As has occurred with many other new technologies perhaps the most important applications of semiconductor spintronics have not even been envisaged yet.

Despite this, one significant application without a dominant technology suggests itself – that of quantum computation. Recent advances in the theoretical understanding of quantum information technology, including novel algorithms, error-correcting codes, and plausible scalable device architectures, suggest that calculations involving quantum mechanical coherent states are possible. Here the extremely long spin coherence times in semiconductor systems, and the demonstrated ability to manipulate that coherence on ultrafast optical times, indicate that spins in semiconductors (whether electron or nuclear) may constitute a good basis for a physical realization of a quantum computer.

See also

Quantum Optics: Quantum Computing with Atoms. **Semiconductor Physics:** Outline of Basic Electronic Properties.

Further Reading

Abragam A (1961) *Principles of Nuclear Magnetism*. Oxford, UK: Oxford University Press.

Awschalom DD and Kikkawa JM (1999) Electron spin and optical coherence in semiconductors. *Physics Today* 52: 33–38.

Awschalom DD, Samarth N and Loss D (eds) (2002) *Semiconductor Spintronics and Quantum Computation*. New York: Springer.

Ivchenko EL and Pikus GE (1997) *Superlattices and Other Heterostructures*, 2nd edn. New York: Springer.

Jeffries CD (1963) *Dynamic Nuclear Orientation*. New York: Wiley.

Meier F and Zachachrenya BP (eds) (1984) *Optical Orientation*. Amsterdam, The Netherlands: North-Holland.

Slichter CP (1992) *Principles of Magnetic Resonance*, 3rd edn. New York: Springer.

Wolf SA, Awschalom DD, Buhrman RA, *et al.* (2001) Spintronics: a spin-based electronics vision for the future. *Science* 294: 1488–1495.

Yafet Y (1963) g factors and spin-lattice relaxation of conduction electrons. In: *Solid State Physics*, vol. 14. New York: Academic Press.

Surface Photovoltage Spectroscopy of Semiconductors

L Kronik, Weizmann Institute of Science, Rehovot, Israel

Y Shapira, Tel-Aviv University, Tel-Aviv, Israel

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Surface photovoltage spectroscopy (SPS) is a well-established contactless technique for semiconductor characterization, which relies on analyzing illumination-induced changes in the surface voltage. SPS traces its origins to the pioneering work of Brattain and Bardeen in the early 1950s, and was extended into a powerful spectroscopic tool by Gatos, Lagowski and Balestra in the early 1970s. It has been used as an extensive source of surface and bulk information on various semiconductors and semiconductor interfaces. In the following, we present the basic theory behind SPS, its experimental setup, and a range of its applications.

Principles of Operation

In general, a surface is defined as a boundary of media with different physical properties. The periodic structure of an ideal crystalline semiconductor results in the appearance of allowed energy bands separated by forbidden energy gaps. In particular, semiconductors are characterized by a nearly full 'valence band', separated by a forbidden bandgap from a higher lying, nearly empty 'conduction band.' Electron transport takes place within the latter and hole transport takes place within the former. The mere termination of this periodic structure at the surface, as well as the chemical changes in the surface associated with this termination, allow for the formation of surface-localized electronic states within the semiconductor bandgap. Electrons or holes in these local states are referred to as 'trapped carriers', as opposed to the 'free carriers' in the conduction and valence bands.

The appearance of surface-localized states induces charge transfer between bulk and surface in order to establish thermal equilibrium between the two. The free carrier density in the vicinity of the surface therefore deviates from its equilibrium value in the bulk. This results in a surface space charge region (SCR), which is electrically non-neutral, implying a nonzero electric field in it and therefore a potential drop across it. Thus, even under equilibrium conditions the surface potential, V_s, is different from the electric potential far away in the bulk. As a specific example which illustrates these concepts, consider an n-type semiconductor with a depleted surface. The majority carriers are electrons and some of them have been trapped in surface states, such that their concentration in the vicinity of the surface is smaller than its equilibrium value but larger than that of the holes (see Figure 1a). The potential drop across the SCR is manifested by the bending of the semiconductor bands, which is such that electrons are repelled from the surface and holes are attracted to it, due to the trapped surface electrons. We note that, by definition, the energy band edge is lower the higher the electrical potential, so that a positive V_s corresponds to downward-bent bands.

Generally, the photovoltaic effect comprises an illumination-induced change in the equilibrium potential distribution and is typically the result of some charge transfer and/or redistribution within the device due to the incident illumination. This is the basic principle behind devices ranging from photodiode detectors to photovoltaic solar panels. A specific variant of the photovoltaic effect is the *surface* photovoltaic effect, which is at the center of this text. It is important to note that the formation of a surface photovoltage (SPV) occurs only if carrier photogeneration *per se* is followed by net charge redistribution. Usually, no significant driving force for such redistribution is found beyond the SCR and the underlying bulk remains quasineutral. Thus, we are interested primarily in the surface SCR.

The SPV mechanism depends strongly on whether the incident photon energy is larger or smaller than the semiconductor bandgap (denoted below as the super-bandgap and sub-bandgap, respectively). Dominant SPV mechanisms are demonstrated below for a depleted n-type surface. The dominant mechanism by which super-bandgap illumination results in surface photovoltage generation is shown in Figure 1b. The electric field in the SCR causes excess electrons to be swept away from the surface and excess holes to be swept towards it. This serves to reduce the density of surface-trapped electrons and decreases the band-bending. In a second mechanism, shown in Figure 1c, either electrons or holes are preferentially trapped at surface defects. This effectively charges the surface and thus modifies the

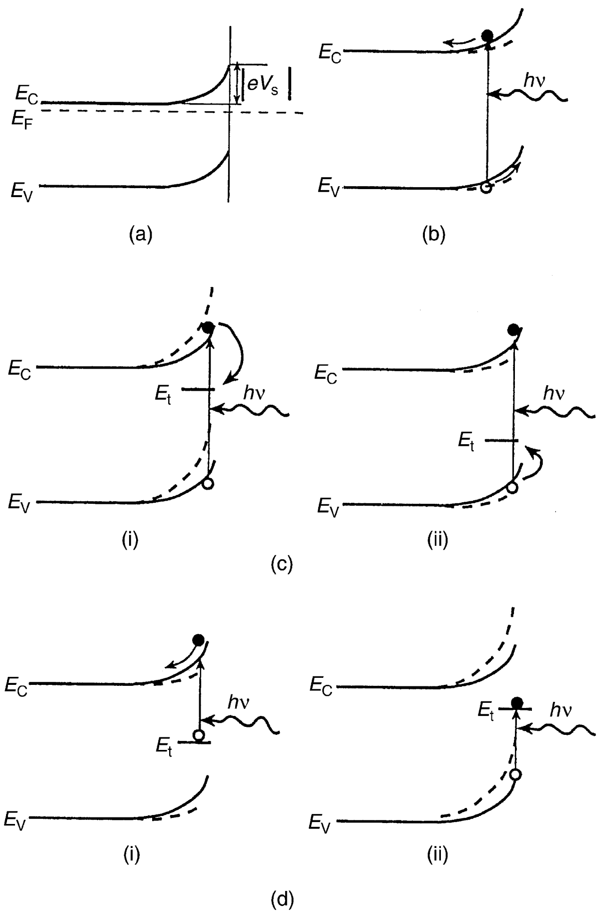

Figure 1 Schematic band diagrams of the surface space charge region at a depleted n-type semiconductor surface under different conditions: (a) thermal equilibrium; (b) super-bandgap illumination with carrier separation under an electric field; (c) super-bandgap illumination with preferential trapping of: (i) electrons, (ii) holes; (d) sub-bandgap illumination with excitation of trapped: (i) electrons, (ii) holes. In all diagrams, solid and dashed lines indicate band positions in the dark and under illumination, respectively. Straight arrows denote carrier generation and curved arrows denote carrier trapping. Adapted from Gal D, Mastai Y, Hodes G and Kronik L (1999) Bandgap determination of semiconductor powders via surface photovoltage spectroscopy. *Journal of Applied Physics* 86: 5573. Reproduced by permission of the American Institute of Physics.

surface potential. This second mechanism is usually apparent in bulk samples only if the equilibrium surface band-bending is fairly small, but increases in importance in polycrystalline materials with decreasing crystallite size, because of the increasing surface to volume ratio.

The most common mechanism for sub-bandgap SPV involves the direct modification of the surface charge, and hence potential, by excitation of trapped carriers, as shown in Figure 1d. Illumination by photons with energy $h\nu > E_c - E_t$ may produce electron transitions from a surface state at an energy E_t into the conduction band, where they are quickly swept to the semiconductor bulk by means of the surface electric field. Hence, the negative surface charge is reduced and the band-bending decreases. Conversely, illumination by photons with energy $h\nu > E_t - E_v$ may produce electron transitions from the valence band into a surface state situated at an energy E_t above the valence band maximum, E_v (which are equivalent to hole transitions from the surface state to the valence band). Such transitions increase the surface negative charge and therefore the surface band-bending. For a SPV to develop in this case, it is necessary to have a significant diffusion of the excess holes into the bulk and/or a significant recombination of electrons and holes inside the SCR, as there is no field-assisted driving force of holes into the bulk.

A typical SPV spectrum shows the SPV as a function of incident photon energy. Its most important features – the ones to watch for when attempting to interpret an SPV spectrum – are the sharp slope changes ('knees') associated with an abrupt onset of an additional carrier excitation mechanism. Consider, as an illustrative example, the top SPV curve in the right-hand part of Figure 2 (the rest of the figure is discussed below). In this curve, two 'knees' are shown. The first 'knee' (marked by an arrow) occurs at sub-bandgap photon energies and corresponds to the onset of carrier excitation from surface states (as in Figure 1d). The second 'knee' occurs in the vicinity of the bandgap energy and corresponds to the onset of band-to-band carrier excitation (as in Figure 1b,c). The proper interpretation of the energy position of these 'knees', their sign (negative or positive SPV change), and the relation of both to the SPV mechanism at work, is elaborated below in the 'Applications' sections.

So far, our discussion has been devoted solely to a study of semi-infinite, uniform bulk samples. However, actual semiconductor samples may be thin, i.e., not much larger than the diffusion length and/or the absorption length, so that excess carriers may be present near their back part as well. Many optically and electronically interesting structures (advanced detectors, transistors, lasers) consist of a multilayer structure, which has several buried interfaces.

Any change in the potential drop across a buried surface will also be measurable at the front surface, because potential drops are additive. For example, the photo-induced potential drop across a photodiode is measured at its contacts, which can be many microns away from its active region. Thus, the sensitivity of SPV measurements to a certain region in the semiconductor is limited only by the

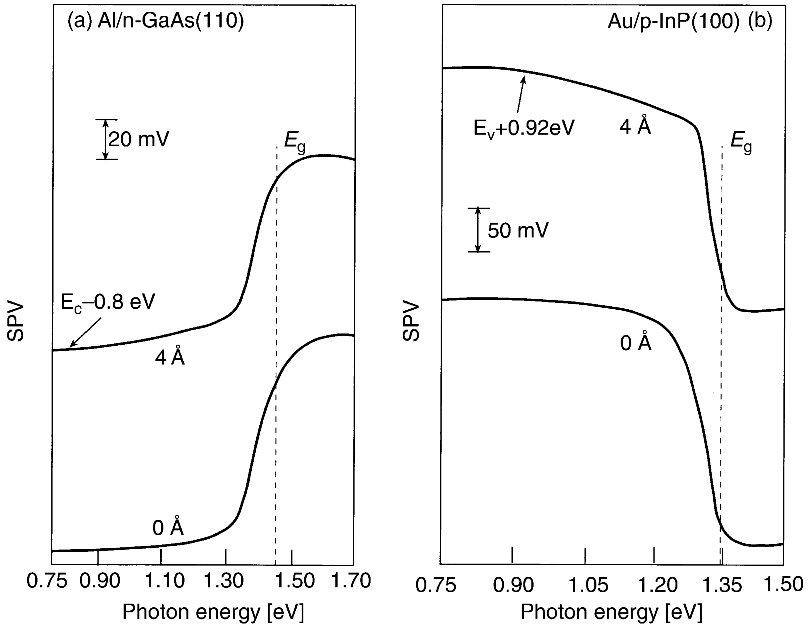

Figure 2 SPV spectra of: (a) an n-type GaAs(110) surface, UHV cleaved (top curve) and with an Al overlayer (bottom curve); (b) a p-type InP(110) surface, UHV cleaved (top curve) and with an Au overlayer (bottom curve). Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Burstein L, Bregman J and Shapira Y (1991) Characterization of interface states at III–V compound semiconductor–metal interfaces. *Journal of Applied Physics* 69: 2312. Reproduced by permission of the American Institute of Physics.

absorption length of the photons, i.e., by the possibility of introducing an excess of free carriers in the region of interest.

Experimental Details

SPV measurements are nontrivial because the surface potential is a *built-in* potential, rather than an external potential. In other words, the surface potential is not equal to the difference in Fermi levels between the front and back surfaces. Therefore, it cannot be simply measured using some form of voltmeter. Moreover, in the case of a free surface, the application of any contact to indirect electrical measurements of the built-in voltage (e.g., current–voltage or capacitance–voltage analyses) will invariably alter the surface properties and hence the quantity under measurement. Consequently, most of the prevalent techniques for measuring the surface potential (or at least changes of it upon illumination) are based on capacitive coupling schemes that do not require a direct electrical contact.

The two most commonly used methods rely on bringing a metallic probe close to the free surface, but never touching it, thus forming a parallel-plate capacitor. An alternating current, from which surface potential changes can be calculated, is then excited by (a) vibrating the metallic probe, which results in alternating capacitance, or (b) using 'chopped' illumination. The two approaches are known as the Kelvin probe and the ac-SPV schemes, respectively. In principle, they provide similar information, but their somewhat different physical principles result in several relative strengths and weaknesses, which make the method of choice application-dependent. In a nutshell, the Kelvin probe is more suitable for following phenomena with large time constants (in particular surface states with long relaxation times) and does not perturb the true band-bending of the free surface. The ac-SPV technique is more suitable for following fast relaxation phenomena, allows for a systematic modification of the surface band-bending in the dark as an extra degree of freedom, and does not require an ohmic back-contact.

Both Kelvin probe and ac-SPV measurements may also be employed in a scanning mode, which allows for recording lateral variations in the SPV, by scanning a small probe or a well-focused light beam, respectively, across the surface. Both approaches are typically limited to a lateral resolution roughly of the order of ~1 μm. SPV measurements can also be performed by a limited modification of two high-resolution techniques for topographic mapping of surfaces that have emerged in the 1980s: scanning tunneling microscopy (STM) and atomic force microscopy (AFM). This opens new horizons for SPV mapping with a resolution on the nanometer scale or even the atomic scale.

A schematic view of a complete, generic SPS setup is given in Figure 3. The sample under study is typically placed inside a metallic box, which serves as both a dark box and a Faraday cage. This box must be such that it effectively includes the probe.

Clearly, for spectroscopic purposes one needs to generate the SPV signal using a broadly tunable light source. The default choice would therefore be the use of a 'white-light' source in conjunction with a monochromator. Because the typical super-bandgap absorption coefficient is typically orders of magnitude larger than the sub-bandgap coefficient, SPS is vulnerable to the spurious contribution of high-order diffraction peaks and stray light. Use of a monochromator where these effects are as small as possible is therefore highly recommended. For some applications, use of an auxiliary light source (e.g., white light, laser light, ultraviolet light, etc.) allows for a 'photobias' of the system and/or for additional quantitative analysis.

The actual recording of an SPV spectrum is coordinated by a real-time computer program. This program controls the monochromator motor and thus wavelength changes, operates the control electronics of the probe, reads the measurement results by interfacing with the probe read-out electronics via an analog-to-digital converter, and finally stores the obtained spectrum for subsequent display and analysis.

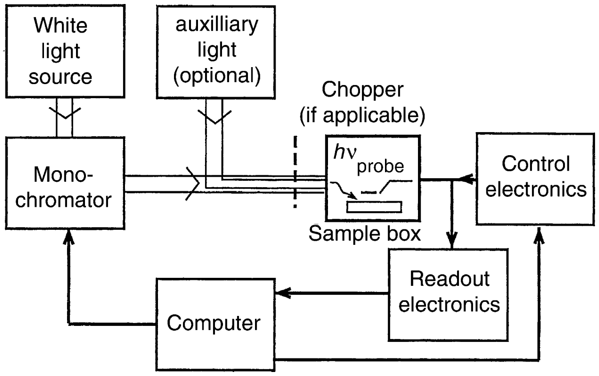

Figure 3 Schematic block diagram of a generic SPS setup. Adapted from Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the cross-roads of physics, chemistry and electrical engineering. *Surface and Interface Analysis* 21: 954. New York: John Wiley and Sons, Inc.

Applications I: Uniform Semiconductors

An elementary application of SPS is the approximate determination of the semiconductor bandgap. This determination is based on the large increase in absorption coefficient near the bandgap energy, E_g, found in most semiconductors. This increase brings about a significant change of the SPV signal, which is easily identified as a sharp change in slope of the SPV curve and is often the most significant one in a given spectrum. An example is shown in Figure 2, where the slope changes related to the bandgaps of GaAs and InP are clearly identified. Indeed, the bandgaps of most known semiconductors have been analyzed using this approach.

The use of SPS for extraction of E_g is at heart nothing more than an emulation of an absorption spectrum. However, as opposed to, e.g., transmission spectroscopy, SPS does not require light collection. It can therefore be performed on arbitrarily thick samples (or on layers sandwiched within a heterostructure – see below) and does not require the sample to be removed from the substrate or grown on a transparent one. It is also inherently insensitive to reflection and scattering and is thus eminently useful for micro- and nanocrystallites.

One must bear in mind that the value of E_g obtained from an SPS plot by inspection is only approximate. Indeed, detailed experimental comparisons between SPV and absorption spectra reveal that the two are often similar, but never identical. Thus, the nominal bandgap is nearly always found within the onset of the largest SPV signal, but is usually relatively broad, the exact position of E_g within it is by no means obvious, and the error in E_g is often ~0.1 eV. However, it has been shown that this error can be reduced significantly – often by one to two orders of magnitude – by a careful quantitative analysis of the experimental data.

We now turn our attention to determining the type (p or n) of the semiconductor. This may be achieved with the help of the sign of the knee associated with the onset of the super-bandgap SPV. Most semiconductor surfaces are depleted, which means that the bands of p-type semiconductors are bent downwards towards the surface, whereas the bands of n-type semiconductors are bent upwards. Since super-bandgap illumination typically tends to decrease the surface band-bending, this would result in a positive SPV in n-type semiconductors and a negative SPV in p-type semiconductors. For example, the SPV spectra of n-type GaAs and p-type InP, shown in Figure 2, clearly feature opposite onset signs which obey the above rules.

While in many cases the type of the semiconductor is known a priori, this is not always the case and SPS can become very useful in determining the semiconductor type. For example, SPS was used for fast and nondestructive verification of the semiconductor type in GaN films, where the p-doping is known to be nontrivial and subject to compensation by its native n-type doping.

We note that under conditions where the SPV is dominated by trapping of carriers, as shown in Figure 1c, the sign convention described above clearly does not apply. For example, etched CdSe quantum dot films were found to exhibit a p-type response in a humid ambient and an n-type response in a dry ambient, on account of preferential trapping of holes or electrons in a dry or humid ambient, respectively.

A key strength of SPS is in the determination of the energy position and type of surface states. Because photons of sufficient energy may excite charge carriers from a surface state to a band or vice versa, we expect a knee in the SPV spectrum whenever the photon energy exceeds the threshold energy of a certain transition. If the latter involves excitation of electrons from the surface state to the conduction band (Figure 1d(i)),a positive change in the surface charge is induced, and hence a positive SPV is expected. Conversely, excitation of holes to the valence band (Figure 1d(ii)) makes the surface charge more negative and a negative SPV is expected. Thus, the combination of the SPV threshold energy and slope sign makes the determination of the approximate position of the surface state within the bandgap possible.

As an illustrative example, consider how SPS can be used to monitor metal-induced surface states, as shown in Figure 2. After cleaving in ultrahigh vacuum, neither GaAs nor InP surfaces display any knees associated with deep surface states, as expected for these surfaces. However, upon deposition of a very thin metallic overlayer, distinct sub-bandgap knees emerge. These knees, associated with positive and negative SPV changes at the Al/GaAs and Au/InP interfaces, respectively, indicate the formation of surface states situated ~0.8 eV below the conduction band edge and ~0.92 eV above the valence band edge, respectively.

Because SPS is contactless and nondestructive, it can be applied not only to clean surfaces, but also to real ones, in practically any ambient. This makes SPS highly suitable for a direct correlation of surface electronic structure with chemical treatments. One such area of active research is the study of surface passivation treatments. An illustrative example, in which the passivating effects of various liquids on a specific p-InP surface (denoted as (100)) was studied *in situ*, is shown in Figure 4. The negative SPV change at ~1.3 eV found in all spectra shown is clearly due to the onset of band-to-band transitions in the InP sample, with its sign indicating a p-type sample, as appropriate. The SPV spectrum of the water-exposed sample (which is similar to that obtained in air (not shown)) features two surface states: the negative SPV slope change, at ~1 eV, is due to a photo-induced population of a surface state situated ~1 eV above the valence band edge. The positive SPV slope change at ~1.2 eV is due to a photo-induced depopulation of a surface state situated ~1.2 eV below the conduction band edge. Interestingly, exposure to HF eliminates the latter state, but not the former. The opposite is true for exposure to H_2O_2. Finally, exposure to a Na_2S solution results in removal of both surface states, in agreement with the known passivating action of this solution. This is just one example of many illustrating the strength of SPS for revealing the link between chemical/physical treatments and device-affecting electronic properties, already at the material level.

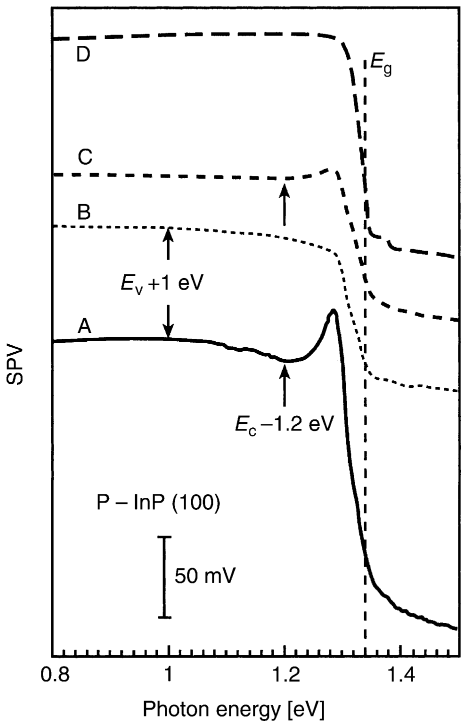

Figure 4 SPV spectra of the p-InP(100) surface in air and after successive immersion in (from top to bottom in the lower energy part): A – 18 MΩ H_2O; B – 0.4% HF; C – 30% H_2O_2; D – a 1.3 M Na_2S solution. Arrows denote the onset of optically induced transitions and labels denote the corresponding defect level. Adapted from Bastide S, Gal D, Cahen D and Kronik L (1999) Surface photovoltage measurements in liquids. *Review of Scientific Instruments* 70: 4032. Reproduced by permission of the American Institute of Physics.

While we have so far considered only surface states, SPS is sensitive to bulk states too, as long as charge excited from or trapped in them can be separated (typically under the electric field of a space charge region). Accordingly, SPS has been used for studying, e.g., Cr-related defects in GaAs or Fe-related defects in InP. The simplest distinction between surface and bulk states in SPS is obtained by examining whether the pertinent spectral feature disappears after surface (e.g., etching) or bulk (e.g., doping) treatment, respectively.

We note that in recent years, SPS has also been applied successfully along the lines outlined here for the study of bandgaps, doping type, and surface and bulk defects in the emerging field of organic semiconductors and organic treatments of semiconductors, as well as for nanostructures, such as fullerenes and nanotubes.

Additional, quantitative analyses of the SPV spectra allow for a determination of densities, thermal and optical transition rates, and the spatial site of gap states. Such analyses are usually based on the SPV response to a systematic modification of one or more external parameters, e.g., time (for transient responses), illumination intensity, temperature, etc.

Applications II: Multilayer Structures

Beyond the study of surface properties, the application of SPS to heterojunctions, multilayer structures, and actual device structures has been most rewarding in recent years. The SPS of multilayer structures is illustrated by Figure 5. This figure features the SPV spectra of a ZnO:Al/ZnO/CdS/Cu(In,Ga)Se_2 layered structure, used for the fabrication of thin-film solar cells. Distinct spectral knees at ~1 eV, ~2.4 eV, and ~3.1 eV, corresponding to the bandgaps of Cu(In,Ga)Se_2, CdS, and ZnO, respectively, are clearly observed. The unequivocal identification of all three bandgaps confirms the above-mentioned theoretical prediction of SPS sensitivity to buried interfaces. Interestingly, the magnitude of the SPV signal is seen to increase with increasing annealing time (in air, at 200 °C), in agreement with the increase in overall solar cell efficiency with such annealing.

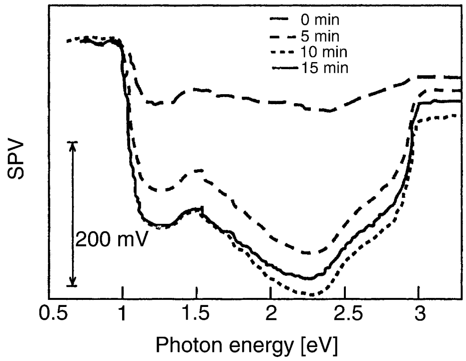

Figure 5 SPV spectra of ZnO:Al/ZnO/CdS/CIGS structures for various annealing times, in air, at 200 °C. Adapted from Gal D, Beier J, Moons E, *et al.* (1995) Band diagram and band line-up of the polycrystalline Cds/Cu(In,Ga)se_2 heterojunction and its response to air annealing. *AIP Conference Proceedings* 353: 453. Reproduced by permission of the American Institute of Physics.

Another feature of growing importance in SPS is that it allows for the characterization of quantum wells and other low-dimensional heterostructures. An illustrative example of the SPS of a multiple-quantum-well (MQW) GaAs/AlGaAs structure is shown in Figure 6a. The spectral knees at ~1.4 eV and 1.78 eV correspond to the onset of band-to-band absorption in the GaAs epilayer and the AlGaAs cap layer, respectively. In between these two knees, the shape of the SPV spectrum resembles typical absorption spectra of MQW structures and the observed MQW-related maxima agree well with calculated energy values of the hole level-electron level transitions denoted as 1HH-1E and 2HH-2E transitions. Note that before etching, a nonnegligible sub-bandgap SPV signal was apparent at photon energies below 1.38 eV, indicating the presence of optically active gap states. Since no such signal was observed after etching, these states were located at the external AlGaAs surface and were not associated with the MQW region. Moreover, the SPV spectrum in the MQW absorption energy range has not changed at all (other than undergoing a uniform shift) after etching, showing that the signal SPV in the MQW absorption range is not interfered by surface effects. Thus, all allowed hole–electron transitions have been resolved, at room temperature.

As SPV analyses of multilayer structures offer the possibility of performing contactless and nondestructive electrical and optical characterization of buried interfaces, they have found significant uses in the field of process monitoring and quality control of actual device structures. This is because the possibility of assessing the quality of a layered structure before any actual device processing takes place presents an opportunity for significant savings in time and money. Recent applications include the identification of defective batches and prediction of ultimate performance of solar cells and x-ray detectors, the prediction of energy levels and lasing wavelength in modern laser heterostructures, the assessment of bandgap narrowing and ultimate gain of heterojunction bipolar transistors, and the

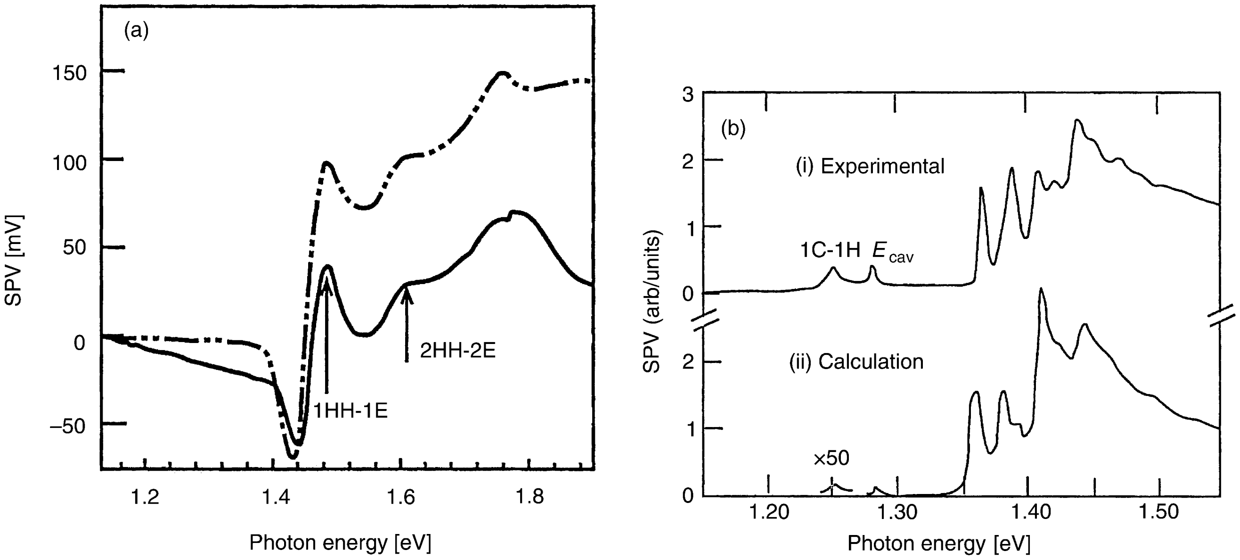

Figure 6 (a) SPV spectra of a GaAs/AlGaAs MQW sample before (solid curve) and after (dashed curve) etching. Adapted from Bachrach-Ashkenasy N, Kronik L, Shapira Y, *et al.* (1996). Surface photovoltage spectroscopy of quantum welts and superlattices. *Applied Physics Letters* 68: 879. Reproduced by permission of the American Institute of Physics. (b) Room temperature experimental SPV spectra (i) and calculated photo-excited carrier spectra (ii) of an InGaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser structure. Adapted from Huang YS, Malikova L, Pollak FH, *et al.* (2000). Surface photovoltage spectroscopy, photoreflectance, and reflectivity characterization of an InGaAs/GaAs/GaAlAs vertical-cavity surface-emitting laser including temperature dependence. *Applied Physics Letters* 77: 37. Reproduced by permission of the American Institute of Physics.

characterization of the electronic structure of high electron mobility transistors and vertical cavity surface emitting lasers.

As an example of such studies, consider the recent characterization of an In-GaAs/GaAs/AlGaAs vertical-cavity surface-emitting laser, shown in Figure 6b. It is readily observed that SPS resolves both the gain providing 1C-1H transition and the lasing-wavelength-related cavity mode, as well as a rich oscillatory structure above ~1.37 eV due to interference effects from the distributed Bragg reflector stacks. The measured SPV spectrum is also in very good agreement with the calculated dependence of the photo-excited carrier density on photon energy, also shown in Figure 6b. In fact, SPS is the first contact-less technique to resolve all of these features in one measurement, and it also allows for a temperature-resolved determination of these quantities.

Summary

SPS features a theory that is relatively straightforward (at least at the qualitative level), and offers typical experimental setups that make the method contactless, nondestructive, ambient insensitive, and capable of any lateral resolution. It can provide a detailed picture of the electronic structure at surfaces, interfaces, heterojunctions, and even complicated multilayer structures. The wide range of SPS applicability also makes it easy to compare and link changes in electronic structure to changes in chemical/physical properties on the one hand, and electrical/optical device properties and performance on the other.

See also

Semiconductor Materials: Modulation Spectroscopy of Semiconductors and Semiconductor Microstructures. **Semiconductor Physics:** Band Structure and Optical Properties; Impurities and Defects; Outline of Basic Electronic Properties; Quantum Wells and GaAs-Based Structures.

Further Reading

Brattain WH and Bardeen J (1953) Surface properties of germanium. *Bell System Technical Journal* 32: 1–41.

Brillson LJ (1982) The structure and properties of metal–semiconductor interfaces. *Surface Science Reports* 2: 123–326.

Gatos HC and Lagowski J (1973) Surface photovoltage spectroscopy – a new approach to the study of high-gap semiconductor surfaces. *Journal of Vacuum Science and Technology* 10: 130–135.

Kronik L and Shapira Y (1999) Surface photovoltage phenomena: theory, experiment, and applications. *Surface Science Reports* 37: 1–206.

Kronik L and Shapira Y (2001) Surface photovoltage spectroscopy of semiconductor structures: at the crossroads of physics, chemistry and electrical engineering. *Surface and Interface Analysis (special issue on Optical Techniques)* 21: 954–965 (main source for the present text).

Lüth H (1993) *Surfaces and Interfaces of Solids*, 2nd edn. Berlin: Springer-Verlag.

Many A, Goldstein Y and Grover NB (1971) *Semiconductor Surfaces*, 2nd edn. Amsterdam, The Netherlands: North-Holland.

Pollak FH (2000) Modulated electric fields produce semiconductor spectra. *Laser Focus World* 36: S9.

SOLITONS

Contents

Bright Spatial Solitons
Optical Fiber Solitons, Physical Origin and Properties
Soliton Communication Systems
Temporal Solitons

Bright Spatial Solitons

C Conti and G Assanto, University ‘Roma Tre’, Rome, Italy

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical spatial solitons, once considered an exotic and rather abstract concept belonging to the realm of mathematics or mathematical physics, nowadays are pursued both experimentally and theoretically in terms of their rich dynamics and variegated features, with important perspectives in applications to signal processing and switching, routing, filtering and – generally – light-by-light controlling. In an intuitive description of a bright spatial soliton, a self-focusing response balances natural diffraction of a light beam, allowing for its invariant propagation in transverse size and intensity. Spatial solitons can be regarded as the stationary or time-independent self-guided ‘eigenwaves’ of Maxwell equations in a material system with a nonlinear optical response. Basically, ‘bright’ spatial solitons (we will not address ‘dark’ cases here, i.e., a dip on an illuminated background) bear the same relationship to nonlinear optics in space as temporal, or fiber, solitons do in time. While both temporal and spatial solitons are special cases of light localization in the $(3+1)$ dimensions (light bullets) encompassing time and two transverse coordinates in space in addition to propagation; the spatial case requires particular attention compared to the fiber inasmuch as, in general, two transverse coordinates (2D) complement the direction of propagation. In other words, the scenario is potentially one-dimension richer than for ‘standard’ temporal solitons, where light confinement, i.e., the effect of the nonlinearity, takes place along a single coordinate. In this respect, the latter case finds a direct counterpart with spatial solitons in planar waveguides, i.e., $(1+1)$D solitons, where light confinement occurs linearly in one spatial dimension, providing group-velocity dispersion in time is replaced by beam diffraction in space, respectively.

When both transverse dimensions are involved, i.e. when a light beam propagates in bulk, in the simplest case of Kerr self-focusing, the nonlinearity can become overcritical and cause filamentation and catastrophic collapse. Additional mechanisms (such as saturation, nonlocality, or higher-order effects) or inherently different nonlinear responses (parametric or molecular), need therefore to be involved in the description of spatial solitons. In several instances, the mathematics differs from the nonlinear Schrödinger equation originally invoked to define a 1D bright optical soliton, leading to nonintegrable systems and to a wider class of solutions, called solitary waves. Since solitary waves constitute the vast majority of the observed spatial light-localization phenomena, we will refer to spatial solitons even in those cases when their model is nonintegrable, defining them as non-diffracting beams in a self-focusing material. While retaining the distinction between 1D and 2D cases, in the following sections we will discuss bright spatial solitons and their features, grouping them on the basis of the underlying nonlinear mechanism and corresponding model, trying to provide a direct link with experimental observations. In doing so, we will leave out discussions on dark solitons and modulationally unstable phenomena, as well as more involved subjects ranging from higher-order to vector, vortex, photonic-crystal, and incoherent spatial solitons.

The General Model of Bright Spatial Solitons

Assuming the paraxial approximation to be valid, the general structure of the evolution equations which lead to spatial solitons propagating along z takes the so-called Foch–Leontovich form:

$$2\mathrm{i}k\frac{\partial A}{\partial z}+\nabla_{\perp}^{2}A=-2k^{2}\frac{\Delta n}{n_0}A \qquad [1]$$

where A is the complex amplitude of the linearly polarized electric field, $k=\omega/c=2\pi n_0/\lambda$, with λ the wavelength, $\omega=2\pi\nu$ the pulsation, n_0 the bulk refractive index, and $\Delta n(\mathbf{r})$ the index change induced by the nonlinear effect. In anisotropic materials, both n_0 and Δn may depend on the polarization of the envelope A. The latter is normalized such that the optical intensity is $I=|A|^2$. Hence the whole optical field is

$$E_{\mathrm{opt}}=\mathrm{Re}\left[\sqrt{\frac{2Z_0}{n_0}}A\mathrm{e}^{-\mathrm{i}\omega t+\mathrm{i}kz}\right] \qquad [2]$$

with Z_0 being the vacuum impedance.

Hereafter we will neglect material losses and assume $\Delta n(\mathbf{r})$ to be real-valued. In the simplest case it is a function of the field intensity I calculated in $\mathbf{r}$. Typical models are the Kerr medium with $\Delta n=n_2 I$, or the saturable Kerr with $\Delta n=n_2 I/(1+I/I_s)$ and I_s a saturation value; they are particular cases of local nonlinear media defined by a relation of the type $F[\Delta n(\mathbf{r}),I(\mathbf{r})]=0$, with F a 'well-behaved' function.

There is also the possibility that the refractive index modification depends on the electromagnetic perturbation (e.g. the intensity) in a finite region, as happens with nonlocal solitons. In the latter case, F needs to be substituted by an integro-differential system of equations, i.e., it represents a functional between $\Delta n(\mathbf{r})$ and $I(\mathbf{r})$.

Relevant exceptions to eqn [1] are spatial solitons due to a quadratic nonlinearity, where a system of two coupled equations must be taken into consideration, and a description in terms of a refractive index change only holds in an approximate, and by no means physical, sense. Quadratic solitons will be discussed later. Here, instead of embarking into a general classification, we will describe the relationships to some representative cases. In this simple framework, spatial solitons are defined as solutions of the type $A=a\exp(\mathrm{i}\beta z)$, with a real-valued and satisfying the equation:

$$\frac{1}{2k^2}\nabla_{\perp}^{2}a+\frac{\Delta n(I)}{n_0}a=\frac{\beta}{k}a \qquad [3]$$

Equation [3] is the z-independent counterpart of eqn [1]. It has the form of the stationary Schrödinger equation (with a field-dependent Hamiltonian operator), where the kinetic energy and the potential are given by the diffraction term and by the nonlinear correction to the refractive index, respectively. Its solutions correspond to solutions of eqn [1] due to the balance between diffraction and nonlinearity. β is the nonlinear correction to the dispersion relation $k=n_0\omega/c$, i.e., it yields the resulting longitudinal wavevector $K=k+\beta$. Solitary solutions are often classified in terms of their dimensionality. Solutions of eqn [3], which depend on either 1 or 2 transverse coordinates, define 1D or 2D solitons, respectively. 1D solitons refer to situations in which a bulk nonlinear medium is illuminated with a highly elliptic beam (such that the diffraction along a transversal dimension, denoted by y, is negligible), or a planar waveguide is excited with a linear mode fixing the transverse profile in one of the dimensions (denoted by y). The nonlinear dynamics and the corresponding steady-state solutions are then described by eqns [1] and [3], with $\partial_y=0$.

Spatial Solitons in Kerr Media

1D Kerr Spatial Solitons and the Nonlinear Schrödinger Equation

The simplest example of a spatial soliton is an optical beam diffracting in one transverse dimension in the presence of the so-called Kerr effect, i.e., a refractive index that increases linearly with the local intensity I:

$$n=n_0+n_2 I \qquad [4]$$

where $n_2(>0)$ is measured in $m^2\,W^{-1}$ and has typical values of 10^{-20} for silica and 10^{-17} for semiconductors at photon energies below half-bandgap, i.e., leaving aside two photon absorption and higher-order effects. For 1D Kerr spatial solitons, eqn [1] becomes

$$2ik\frac{\partial A}{\partial z}+\frac{\partial^2 A}{\partial x^2}=-2k^2\frac{n_2}{n_0}|A|^2A \qquad [5]$$

Equation [5] is known as the nonlinear Schrödinger equation (NLS), and is the basic model in describing temporal soliton, i.e., nondispersing pulses propagating in optical fibers, provided the variable x is intended as the temporal coordinate. The importance of eqn [5] is due to several remarkable, in some sense unique, mathematical properties, and to the fact that it can be considered a fundamental model for diverse nonlinear phenomena. This circumstance is often referred to as the 'universality of the NLS'.

In the framework of optical spatial solitons, the NLS relevance is moderated by the fact that most experimentally explored environments involve processes significantly more complicated than the Kerr effect. Nevertheless, it settles the basic principles and constitutes a reference point in optical soliton assessment.

Trying to gain some physical insight, the starting point is the beam spreading when $n_2 = 0$ in eqn [5], i.e., the linear regime corresponding to low intensities. The diffraction is inherent to the spectrum of plane waves in a finite beam (along x). To underline the role of the nonlinear response, one can consider the NLS without the diffraction term (the second derivative with respect to x), obtaining:

$$A(x,z)=\sqrt{I(x)}\exp\left[ik\frac{n_2}{n_0}I(x)z\right] \qquad [6]$$

Equation [6] shows how the nonlinearity yields a phase-front curvature related to the sign of n_2: the medium is self-focusing (a converging beam) for $n_2 > 0$ or self-defocusing (a diverging beam) for $n_2 < 0$ (see Figure 1). In the absence of diffraction, i.e., for a plane wave, the Kerr effect modifies the field phase by increasing ($n_2 > 0$) or decreasing ($n_2 < 0$) the local refractive index (and hence the effective optical path) proportional to the intensity. For a beam of finite transverse extent, nonlinearity and diffraction act together. In a self-focusing medium, the concave phase-front due to diffraction can be balanced by the nonlinearity, and the natural tendency of a beam to spread can therefore be compensated. For this to happen, the length-scales associated to the diffraction, i.e., the Rayleigh range $L_R = kw_0^2$, and the Kerr effect must be comparable. A characteristic length $L_N = n_0/(kn_2I_0)$, over which the nonlinear phase shift equals 1, can be associated

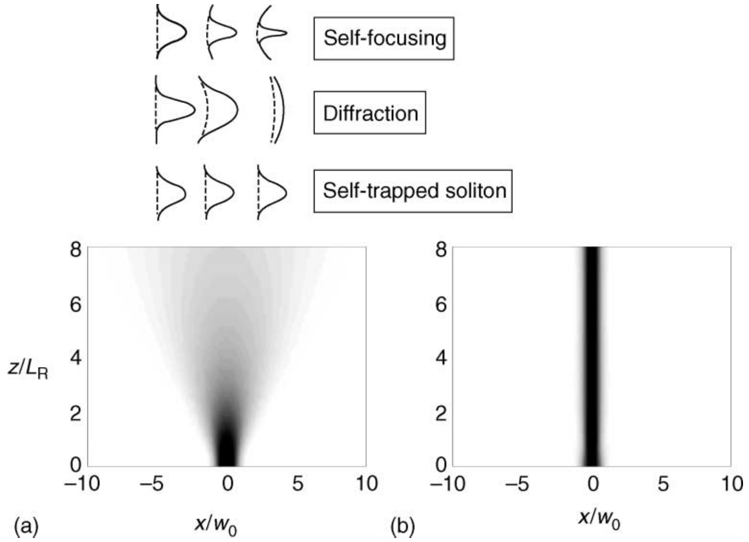

Figure 1 Kerr spatial solitons. Top: sketch of the mechanism leading to Kerr soliton formation. In the absence of diffraction, self-focusing of a Gaussian beam provides a converging phase-front (dashed lines). Conversely, diffraction by itself causes beam divergence. When these two effects act jointly, they can balance each other and soliton generation can be observed, with a flat phase-front. (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6:1419–1426). Bottom: beam propagation method used to simulate the propagation of a one-dimensional Gaussian beam by integrating the nonlinear Schrödinger equation, for (a) low-power input, (b) input high enough to generate a self-trapped solitary wave. The transverse coordinate x is expressed in units of the beam waist, the propagation coordinate z in Rayleigh lengths.

to the nonlinearity, with I_0 the peak intensity across the beam profile. When the two lengths L_R and L_N are nearly equal, i.e., $w_0^2 I_0 \approx n_0/(k^2 n_2)$, the balance between the phase-front curvatures, due to diffraction and nonlinearity, yields a self-localized nondiffracting beam. Such a situation is described by the exact solutions of eqns [1] and [3], which takes the form:

$$A(x,z)=a(x)e^{i\beta z} \quad \text{with} \quad a(x)=\sqrt{I_0}\,\mathrm{sech}\left(x\sqrt{\frac{k^2 n_2}{n_0}I_0}\right) \qquad [7]$$

and $\beta = n_2 I_0/2n_0$ is determined by the peak intensity or, equivalently, by the power P_1 carried by the soliton per unit of transverse wavefront along y:

$$P_1=\int_{-\infty}^{\infty} |A|^2\,dx=\sqrt{\frac{4n_0 I_0}{k^2 n_2}}$$

Note that the overall phase-front is flat, and the spatial extension in width yields $1/(n_2 k^2 I_0)^{1/2}$, as anticipated. The relationship between the soliton width – which determines the strength of diffraction – and the intensity – responsible for the nonlinear phase shift – is known as the existence curve, and is characteristic of the specific nonlinear mechanism.

Equation [7] is the fundamental soliton of the NLS and the paradigmatic 1D spatial soliton. Its sech shape specifically results from the type of nonlinearity; as discussed below, other effects (including parametric generation) can yield different profiles.

Figure 1 shows the simulation of a 1D Kerr soliton generated from a Gaussian input. Self-trapping is attained when the peak intensity or, equivalently, the power, reaches a threshold value depending on experimental details, the size of the nonlinear response, and the beam profile.

Figure 2 shows experimental results on the formation of a 1D Kerr soliton in a semiconductor (AlGaAs, $n_0 \approx 3.5$) waveguide. At low power (center panel) the output beam exhibits diffraction, whereas the soliton (bottom) reproduces the input beam profile (top panel) at the end facet of the sample. These results were obtained at $\lambda = 1.55$ μm, using pulsed laser beams, with temporal pulse durations of about 1 ps and peak powers of about 500 W; the generated 1D solitons had waists of tens of microns.

In a physically intuitive sense, self-trapping is due to the refractive index increase across the beam profile. Such an increase gives rise to a graded index optical waveguide able to confine all the plane wave components with angles of propagation within the numerical aperture of the input beam. Large spatial-frequency components k_x cannot be trapped and are emitted as radiation during soliton formation. It is possible to experimentally verify that this self-induced waveguide is also able to guide a weak probe uncorrelated to the soliton, provided it is above the cutoff for dielectric confinement. The use of spatial solitons for intensity-controlled or all-optical lightwave circuitry can therefore be envisaged. The first experiments on optical 1D Kerr spatial solitons date back to 1985, and were performed in carbon disulfide. Since then, 1D Kerr spatial solitons have been reported in glasses, semiconductors, and organic materials.

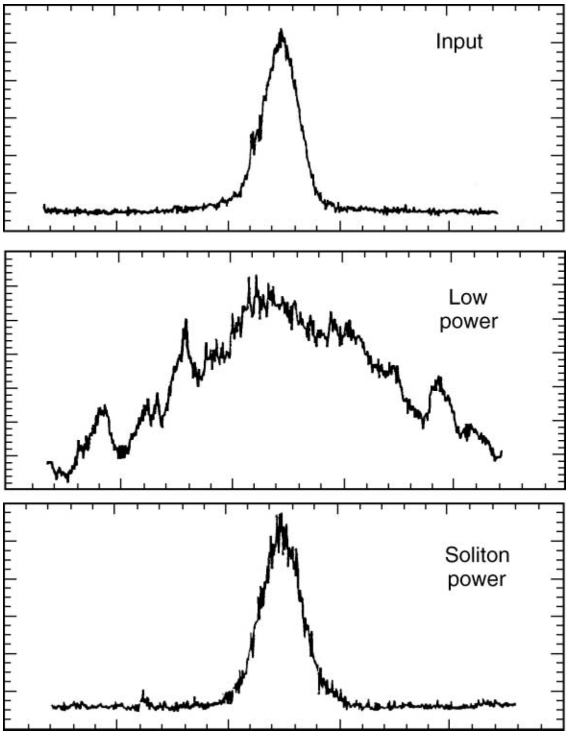

Figure 2 1D Kerr soliton in a planar AlGaAs waveguide. Transverse beam profiles at the input (top) and output facets of a planar waveguide in AlGaAs, at low (center) and high (bottom) power, showing diffraction and soliton formation, respectively. (Adapted with permission from Aitchison JS, Al-hemyari K, Ironside CN, Grant RS and Sibbett W (1992) Observation of spatial solitons in AlGaAs waveguides. *Electronic Letters* 28: 1879.)

Kerr Spatial Solitons in Two Transverse Dimensions: Saturable and Nonlocal Media

When the dynamics of the optical beam encompasses two transverse dimensions, the idealized discussion above no longer applies. As long as the medium is described by the simple Kerr law (eqn [4]), 2D propagation does not admit stable self-trapped stationary beams. Because of this, two alternatives are available: (i) the beam diffracts; or (ii) the beam progressively self-focuses until it undergoes catastrophic collapse: in the absence of additional limiting phenomena, such as plasma formation or nonparaxial/vectorial coupling, its waist keeps shrinking with a corresponding unbound increase in peak intensity (beam power is conserved). Which regime actually applies, depends on whether the excitation is lower or higher than a self-focusing critical power $P_c = (0.61\lambda)^2/8n_0n_2$. The latter corresponds to a diffraction-driven divergence equal to the critical angle for total internal reflection in the induced waveguide. The self-focusing instability, first predicted by Kelly in 1965, is related to the so-called Derrick's theorem: particularly related to spatial solitons in a positive ($n_2 > 0$) Kerr medium, it states that the equilibrium between diffraction and nonlinearity in 2D is unstable.

The situation is markedly different when saturation is present: self-focusing arrests at the maximum value of the nonlinear refractive index increase, and stable self-trapped beams are allowed. The simplest model for a saturating nonlinearity yields the following stationary-state equation:

$$\frac{1}{2k^2}\nabla_\perp^2 a + \frac{n_2}{n_0}\frac{I}{1+I/I_s}a = \frac{\beta}{k}a \qquad [8]$$

Its solutions, while retaining a bell-shaped profile, depart from the purely Kerr model corresponding to $I_s \longrightarrow \infty$, and can only be obtained numerically.

Gases, and specifically sodium vapors, were the first media in which spatial solitons of a saturable Kerr-like nonlinearity were observed in 1974. Historically, these were the first 2D optical spatial solitons.

Most recent developments on optical spatial solitons involve more complicated environments and nonlinear responses. In the next sections we will consider those most investigated experimentally: photorefractive solitons, nonlocal solitons in liquid crystals and parametric solitons.

Spatial Solitons in Photorefractive Media

In recent years, research on spatial optical solitons in photorefractives has been among the most active. Self-trapping in photorefractive materials was first predicted in 1992 and observed in 1993. However, a clear assessment of this result came with the demonstration of the so-called steady-state screening photorefractive solitons in 1994.

The principle underlying the photorefractive response can be described as a feedback mechanism relating to the optical induction of a static electric field **E**. In specific materials, such as semiconductors or ferroelectrics, a light beam generates free charges which diffuse and drift under an externally applied voltage. Owing to free carriers and donor/acceptor sites, the resulting net spatial charge produces a static electric field which, in turn, modifies the refractive index n_0 via the electro-optic effect, phenomenologically described by

$$\Delta n_{ij} = -\frac{1}{2}n_0^3 r_{ijk}E_k - \frac{1}{2}n_0^3\varepsilon_0^2(\varepsilon_r - 1)^2 g_{ijkl}E_kE_l \qquad [9]$$

Equation [9] is a series expansion of the index change in powers of the static electric field: all terms are present if the medium is noncentrosymmetric, otherwise only even-powered terms need be included. In essence, eqn [9] entails the feedback mediated by the static electric field on the optical wave.

The feedback leads to stable solutions under appropriate conditions. These are steady-state solitons if the inhomogeneous light-induced index change makes a suitable waveguide for the beam. The three main types of such self-localized beams are screening, photovoltaic, and centrosymmetric photorefractive solitons. Boundary conditions determine the induced static field to be inserted in eqn [9] and in the propagation eqn [1]. In the simplest cases we consider, the light-wave is linearly polarized and the tensorial nature of the index perturbation in eqn [9] can be neglected.

The equations encompassing most of the physics of the photorefractive effect constitute the Kukhtarev model and can be shown as:

$$\begin{aligned} &s(I + I_b + I_{dark})(N_d - N_d^i) - \gamma_r\rho N_d^i = 0 \\ &\nabla\cdot\mathbf{J} = \nabla\cdot[q\mu\rho\mathbf{E} + k_B T\mu\nabla\rho + \beta_{ph}(N_d - N_d^i)I\mathbf{c}_{opt}] = 0 \\ &\nabla\cdot\mathbf{E} + \frac{q}{\varepsilon_S}(\rho + N_A - N_d^i) = 0 \end{aligned} \qquad [10]$$

The relevant quantities in eqns [10] are: the current density **J**; the free charges (electrons) density ρ; the density of ionized (total) donors N_d^i (N_d), the density of negatively charged sites N_A, the photo-ionization cross-section s, the background density I_b (due to background laser light), the dark irradiance I_{dark} (sI_{dark} is the dark generation rate, typically $I_b \gg I_{dark}$), the recombination rate γ_r, the electron mobility μ, the low frequency dielectric constant $\varepsilon_S = \varepsilon_0\varepsilon_r$, the Boltzmann constant k_B, the versor $\mathbf{c}_{opt}$ of the optic axis, and the temperature T. In the second of eqn [10], the pertinent element β_{ph} of the photovoltaic tensor needs be taken into account only for the photovoltaic solitons described below. The above must be completed by boundary conditions for the bias V – the line integral of

E – and the geometry of the metal contacts. Typically, $N_d \approx 10^{18}\ \text{cm}^{-3} \gg N_a$, $I_{\text{dark}} \approx 10^{-3}-10^{-6}\ \text{W/cm}^2$, $I_b \approx 10\ \text{W/cm}^2$, $\varepsilon_r \approx 10^3$ and solitons are observed in the visible.

Equations [10] determine the field **E** in terms of the optical intensity I in steady-state, and **E** is inserted in eqn [9] to yield the optically induced index change. A major signature of photorefractive solitons is the cumulative process they rely on: the space-charge field **E** builds up in time along with the distribution of carriers, until equilibrium is reached. Hence the phenomenon does not require high optical intensities but, conversely, a sufficiently long time (of the order of minutes or hours). Photorefractive solitons are typically observed at modest excitations, as low as a few microwatts, with typical waists of tens of microns.

Screening and Photovoltaic Solitons

When an externally applied bias lowers the refractive index of a crystal through its electro-optic response, a light beam can screen this effect, defining a self-waveguide supporting a spatial soliton. In a noncentrosymmetric photorefractive, the beam raises the conductivity in the illuminated region, locally lowering the voltage drop compared to dark regions. The induced space-charge screens the applied static field, and it may bring the refractive index back to its unbiased value. This can be seen as a voltage divider formed by a small optically controlled resistance and two large ones corresponding to the dark regions. The resulting solitary waves are called screening photorefractive solitons. For them to be stable (in time), additional illumination of intensity I_b at the same wavelength is required on the crystal in order to increase the conductivity of the (dark) regions surrounding the soliton. The charge accumulation in the beam-tails would otherwise flatten the induced index distribution and destroy the soliton.

The simplest description of screening solitons is provided by the saturating relationship between the electric field E and the light intensity I:

$$E = -\frac{V}{L}\frac{1}{1 + I/I_b} \qquad [11]$$

to be used in eqn [9] to get the index perturbation via the electro-optic effect. L is the width across which the voltage is applied. Since the quadratic term in eqn [9] (weighed by the tensor g_{ijkl}) can be neglected in noncentrosymmetric crystals:

$$\Delta n \approx -\frac{1}{2} n_0^3 r_{\text{eff}} E \qquad [12]$$

where r_{eff} is the relevant component of the electro-optics tensor; typically $n_0 \approx 2$, $r_{\text{eff}} \approx$ tens or hundreds of pm/V, and the overall index perturbation is of the order of $10^{-4}-10^{-3}$, with $V/L \approx 10^3$ V/cm. This description applies to 1D solitons; for 2D solitons, sometimes referred to as 'needle' solitons, the whole set of constitutive equations must be solved numerically. Nevertheless, in most cases the saturating model, even if not rigorous, describes well solitons and related phenomena both theoretically and experimentally, especially when nonlocal effects can be ignored.

Figure 3 shows the propagation of a 2D spatial screening soliton (top) compared to the linear case (bottom). The diffractive regime is realized by artificially 'turning off' the nonlinearity acting on the applied voltage (when $V = 0$ in eqn [11] then $\Delta n = 0$, corresponding to no induced perturbation).

Another class of photorefractive spatial solitary waves is that of photovoltaic solitons. In this case the electric field in the crystal has an additional component through the photovoltaic effect, the latter being relevant in materials such as $LiNbO_3$. To sustain bright solitons the index perturbation due to the photovoltaic current, which induces the space-charge field, must be positive. This is the case for Cu:KNSBN, where 2D photovoltaic solitons were first observed.

The space-charge field, due to the photovoltaic effect, can be derived from the Kukhtarev model:

$$E = -E_{\text{ph}}\frac{I/I_b}{1 + I/I_b} \qquad [13]$$

with E_{ph} is a constant depending on the material ($E_{\text{ph}} \approx 10^3$ V/cm) and I_b is the dark-irradiance controlled by the background laser light. Typical

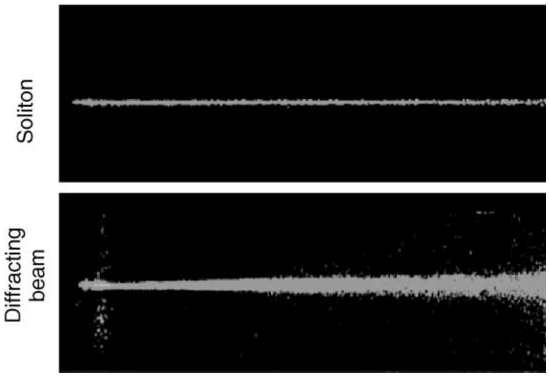

Figure 3 2D photorefractive soliton. Top view photograph of a 10 mm wide spatial soliton propagating in strontium barium niobate (top), and for comparison, the same beam diffracting when the nonlinearity if 'turned off' (bottom). (Adapted with permission from Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426).

geometries account for the tensorial nature of the effect, and are arranged to make I_b of the same magnitude of the beam intensity I. If $E_{ph} > 0$, as in Cu:KNSBN, eqn [13] in eqn [12] shows that a positive index change of the saturating type can be induced and support bright solitons. This model is successfully used for both 1D and 2D bright photovoltaic spatial solitons.

Photorefractive Solitons in Centrosymmetric Media

Most experiments on screening solitons were performed in ferroelectric noncentrosymmetric crystals. When brought above the Curie temperature T_c, such materials exhibit a phase-transition to a centrosymmetric paraelectric state. Crystals such as Potassium Lithium Tantalate Niobate have room-temperature phase-transitions, conveniently enabling experimental investigations. In paraelectrics, the second-order electro-optic tensor r_{ijk} is zero because of symmetry, but the photorefractive effect is still available through the third-order coefficient g_{ijkl} in eqn [9]. By properly choosing the orientation, eqn [12] can then be replaced by

$$\Delta n = \frac{1}{2} n_0^3 \varepsilon_0^2 (\varepsilon_r - 1)^2 g_{\text{eff}} E^2 \qquad [14]$$

Operation near T_c greatly enhances the relevant electro-optic coefficient $g_{\text{eff}} > 0$ ($g_{\text{eff}} \approx 0.1\ \text{m}^4\ \text{C}^{-2}$), making the observation of photorefractive screening solitons possible. Using eqn [11] in eqn [14] the model is saturating and works for both 1D and 2D bright solitons.

Thanks to the low powers involved and a mature crystal technology, the photorefractive effect and related soliton phenomena have contributed several impressive results in optical morphology of nonlinear light propagation, including incoherent and white light excitations.

Spatial Solitons in Nonlocal Media, Liquid Crystalline Media

A link, such as the Kerr law (eqn [4]), between the index perturbation and the optical intensity is 'local', because Δn at a given point (x, y, z) only depends on the intensity at the same location. Otherwise the medium is said to be 'nonlocal', e.g., one of those described by the general Kukhtarev model (eqn [10]) although, sometimes well approximated by eqns [11] or [13].

There are various mechanisms yielding nonlocality, mainly through spatial diffusion of the index perturbation far from the excitation. Remarkable nonlocal effects are expected when the spatial features of the beam, e.g., its waist, are significantly smaller than those of the induced polarization.

One of the most important effects of nonlocality is the stabilization against the catastrophic self-focusing in 2D, as experimentally investigated by Suter and Blasberg in 1992. Evidence of a highly nonlocal behavior has been very recently reported in liquid crystals, where the nonlinearity can be thermal or re-orientational. In the former case the index perturbation is much wider than the beam waist due to thermal diffusion associated to (absorption mediated) heat transfer. In the latter case the elastic properties of the medium produce a widening of the self-waveguide far from the beam axis. In nematic liquid crystal (NLC), the re-orientational response stems from the optically induced tilt of the elongated molecules towards the linearly polarized field versor. The light forced re-orientation increases the refractive index, thus creating the self-guiding conditions for spatial solitons. However, since in the nematic phase the molecules are linked together in a wide-scale order, the region (kernel) affected by this process is determined by the elastic properties of the medium and can extend to tens of microns.

An effective approach to describe spatial solitons in a highly nonlocal medium (where the kernel is much larger than the beam) postulates that the punctual relationship between Δn and I is replaced by a link between Δn and the beam power P. In such a way the transverse integral of the intensity is fully embedded in the index perturbation.

Assuming $n = n(P, r)$ with the beam mainly subject to the refractive perturbation near its axis, one can write:

$$n^2(r, P) = n_0^2(P) - r^2 \gamma(P) \qquad [15]$$

with $n_0^2(P) \cong n_0^2$ the on-axis value. $\gamma(P)$ depends on the specific geometry and determines the wavefront convexity at the origin. The overall profile is basically determined by the constitutive relationships of the medium. The propagation equation can be shown as

$$2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A = \frac{k^2}{n_0^2} \gamma^2(P) r^2 A \qquad [16]$$

Equations like [16] also describe the transverse profile of a beam linearly propagating in optical fibers, as well as the quantum harmonic oscillator (interpreting z as time). The soliton is its nondiffracting solution, a particular stationary case of a more general Gaussian beam with waist variable in propagation.

For intensities:

$$I(z) = \frac{P}{\pi w^2(z)} e^{-r^2/w^2(z)} \qquad [17]$$

the spot-size obeys:

$$\frac{w^2(z)}{w_0^2} = 1 + \left[\frac{1}{k_0^2 w_0^4 \gamma^2(P)} - 1\right] \sin^2\left[\frac{n_0 z}{\gamma(P)}\right] \qquad [18]$$

Therefore, for a Gaussian excitation, the waist oscillates along z between the values w_0 – in general determined by the launch conditions – and $w_m = w_0 \gamma(P_S)/\gamma(P)$. P_s is the 'soliton power' for which $w_m = w_0$ and $k^{02} w_0^4 \gamma^2(P_S) = 1$, hence for $P < P_s$ or $P > P_s$ the beam pulsates cyclically with z (breathes). Such a dynamic balance between diffraction and nonlinearity stems from the spectral broadening associated to an increased beam shrinking: when the spectrum widens sufficiently, diffraction starts to prevail and the beam spreads again. This process repeats itself in propagation.

The stationary solution for $P = P_s$ corresponds to a point in the existence curve of waist versus power, i.e.:

$$k_0 w_0^2 \gamma(P_S) = 1 \qquad [19]$$

Such exact balance is critical, and experimentally observed solitary waves in highly nonlocal media appear as breathers.

Figure 4 displays the intensity profile of a soliton propagating in nematic liquid crystals, as acquired by a digital camera placed on the top of the cell and retrieving the scattered light (see Figure 5). The waist oscillation in the self-trapped beam and the dependence of the period on excitation are apparent.

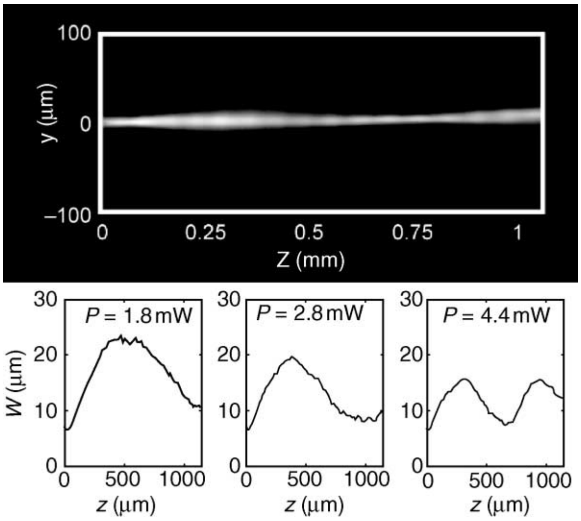

Figure 4 Nonlocal breather. (Top) top view photograph of a spatial soliton in NLC. Beam waist oscillations are apparent in propagation, as graphed (bottom panels) versus z for various excitations.

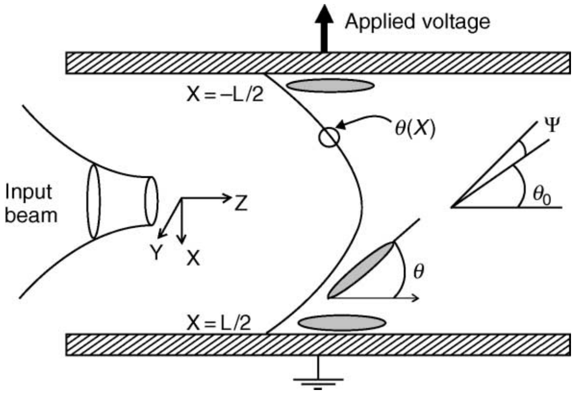

Figure 5 Sketch of a liquid crystal cell for the observation of nonlocal solitons. $\theta(x)$ is the orientation of the main axis of the ellipsoidal molecules, fixed by an external bias. The alignment is planar, with molecules parallel to the cell facets. The optical beam propagates in the center of the cell, where $\theta(x) = \theta_0$, and induces an additional re-orientation ψ. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

A specific highly nonlocal nonlinear system is formed by a linearly polarized beam in NLC with molecules pre-tilted (by an external electric or magnetic field) at $\theta_0 = \pi/4$ with respect to the input polarization. As sketched in Figure 5, the light wave induces an additional tilt Ψ. At the lowest order of approximation, the relevant set of equations can be shown as:

$$\begin{aligned} 2ik\frac{\partial A}{\partial z} + \nabla_\perp^2 A &= -k^2 \frac{n_a^2}{n_0^2} \Psi A \\ K\nabla_\perp^2 \Psi - G\Psi + \frac{n_a^2}{2cn_0}|A|^2 &= 0 \end{aligned} \qquad [20]$$

with K the NLC elastic constant ($K \approx 10^{-11}$ N, for NLC of the E7 type), $n_0 \approx 1.5$, n_a^2 the optical anisotropy ($n_a \approx 1$), i.e., the difference between relative permittivities along the two principal axes, G a constant determined by the specific geometry of the NLC cell and the pre-tilt angle θ_0. Since, for $K = 0$, eqn [20] represents a Kerr medium, it can be considered one of the simplest nonlocal generalizations of local models for spatial solitons. The length $R_c = \sqrt{K/G}$ quantifies the diffusion of the index perturbation away from the beam axis, such that Ψ decays according to the modified Bessel function $K_0(r/R_c)$. Typically, $R_c \approx 20$ μm and solitons are observed in the visible and near-infrared.

Solitary wave profiles are found with the ansatz $A = a\exp(\mathrm{i}\beta z)$, obtaining:

$$\begin{aligned} &\frac{\nabla_\perp^2 a}{2k^2} + \frac{n_a^2}{2n_0^2}\Psi a = \frac{\beta}{k}a \\ &R_c^2\nabla_\perp^2\Psi + \frac{R_c^2 n_a^2}{2Kcn_0}a^2 = \Psi \end{aligned} \qquad [21]$$

No analytical solutions of eqns [21] are known in 2D, hence the profiles have to be derived numerically. Conversely, it can be shown that 1D sech^2 solutions exist (see next section on parametric solitons). From eqn [21], the whole family can be parameterized by a dimensionless $\alpha = 1/(2k\beta R_c^2)$. Since $(1/2k\beta)^{1/2}$ measures the soliton waist (see eqn [7]), α relates the beam size and the extension of the induced perturbation. Small α solutions access the highly nonlocal regime, as shown in Figure 6, where the ratio R between the index perturbation and the soliton waist is graphed. The insets show sample profiles of a and ψ in two limits, respectively. On the other hand, when $K = 0$, $R_c = 0$, and $\alpha \to \infty$, the local limit yields an index perturbation comparable in size to the beam waist.

It is worth pointing out that a model such as eqn [16], also derived in the framework of plasma physics, predicts unconditionally stable self-trapped beams. Nonlocal solitons realize an energy minimum, thus preventing beam collapse such as in saturable (Kerr) media. The oscillations observed in experiments can be regarded as fluctuations around the lowest energy state.

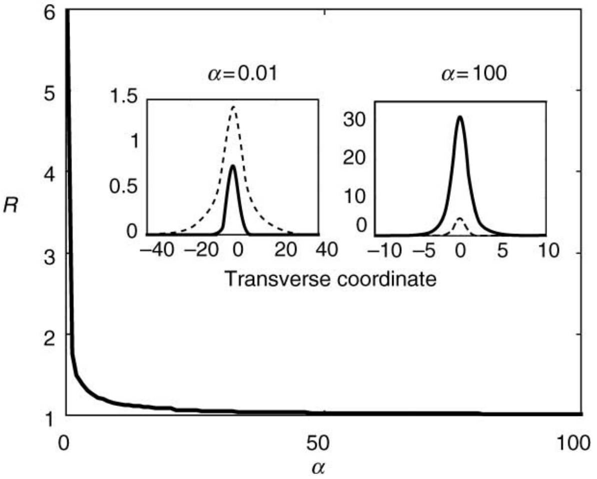

Figure 6 Spatial solitons versus nonlocality. Graph of the ratio *R* between the waist of the induced index perturbation in a nonlocal medium and the soliton waist versus parameter α. The insets show (in normalized units) soliton (solid line) and perturbation profiles (dashed line) for two different α, i.e., in the local ($\alpha = 100$) and highly nonlocal ($\alpha = 0.01$) regimes. (Adapted from Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901).

Parametric Spatial Solitons in Quadratic Media

Solitons due to parametric processes, such as second-harmonic generation, belong to a different class with respect to previous sections. In fact, no refractive index modification can be physically associated to parametric effects. Parametric solitons are characterized by a power spectral density peaked around at least two carrier harmonic frequencies. It is the interaction between these components which provides the diffraction beating mechanism. Their existence was predicted in the early 1970s by Sukhorukov and Karamzin, while the first experimental evidence was reported in 1995.

The simplest process to address is second-harmonic generation (SHG): a fundamental frequency (FF) wave at ω interacts with a component at 2ω. The latter can be spontaneously generated in noncentrosymmetric crystals (e.g., $\mathrm{LiNbO_3}$, $\mathrm{KNbO_3}$, BBO, KTP, etc.) or externally seeded.

For plane waves, this process is ruled by:

$$\begin{aligned} &\mathrm{i}\frac{\mathrm{d}A_1}{\mathrm{d}z} + \frac{\omega d_{\mathrm{eff}}}{c}A_1^*A_2\,\mathrm{e}^{-\mathrm{i}\Delta kz} = 0 \\ &\mathrm{i}\frac{\mathrm{d}A_2}{\mathrm{d}z} + \frac{\omega d_{\mathrm{eff}}}{c}A_1^2\,\mathrm{e}^{\mathrm{i}\Delta kz} = 0 \end{aligned} \qquad [22]$$

with $\Delta k = k_1 - 2k_2 = 2\omega/c(n_1 - n_2)$ the phase mismatch, d_{eff} the nonlinear coefficient, n_2 and n_1 the indices at SH and FF, respectively, and

$$E_{\mathrm{opt}} = \mathrm{Re}\left[\sqrt{\frac{2Z_0}{n_1}}A_1\mathrm{e}^{-\mathrm{i}\omega t+\mathrm{i}k_1z} + \sqrt{\frac{2Z_0}{n_2}}A_2\mathrm{e}^{-2\mathrm{i}\omega t+\mathrm{i}k_2z}\right]$$

the overall electric field. d_{eff} relates to the quadratic susceptibility $d_{\mathrm{eff}}^{(2)}$ (measured in $\mathrm{mV^{-1}}$) through to $d_{\mathrm{eff}} = d_{\mathrm{eff}}^{(2)}\sqrt{2Z_0}/(n_1\sqrt{n_2})$, and typically $d_{\mathrm{eff}} \approx 10^{-10}\,\mathrm{m^2\,W^{-1/2}}$.

Equations [22] show that, in general, the harmonics with amplitudes A_1 and A_2 affect each other's phase and intensity during propagation. The solution with constant moduli, i.e.:

$$\begin{aligned} &A_1 = \sqrt{I_1}\exp\left[\mathrm{i}\frac{\omega}{c}\Delta n_1 z\right] \\ &A_2 = \sqrt{I_2}\exp\left[\mathrm{i}\frac{2\omega}{c}\Delta n_2 z\right] \end{aligned} \qquad [23]$$

with I_1 and I_2 the respective intensities, provides a 'pure' phase variation and 'apparent' index changes Δn_1 and Δn_2 on either wave. The latter satisfy:

$$\begin{aligned} \Delta n_1 &= d_{\mathrm{eff}}^2 \frac{I_2 - I_1/2}{n_1 - n_2} \\ \Delta n_2 &= d_{\mathrm{eff}}^2 \frac{I_1}{2I_2} \frac{I_2 - I_1/2}{n_1 - n_2} \end{aligned} \qquad [24]$$

with the constraint (for plane waves to exist):

$$I_1 = 2I_2 - 2\frac{n_1 - n_2}{d_{\mathrm{eff}}}\sqrt{I_2} \qquad [25]$$

Equations [24] encompass a great deal of the physics in self-trapping via SHG. When eqn [25] is satisfied (at least approximately), the effect of SHG is exclusively on the phase. Assuming eqn [24] is valid for (wide) beams, the propagation in the presence of diffraction can be described by two coupled equations of the type [1] with relevant perturbations (eqn [24]). This shows that the sign of the term $(2I_2 - I_1)/(n_1 - n_2)$ is crucial, because focusing at both harmonics ($\Delta n_{1,2} > 0$) requires $(2I_2 - I_1)/(n_1 - n_2) > 0$.

When just a fundamental frequency (ω) is launched, solitons are generated for $\Delta k > 0$: a small 2ω component is provided via SHG and $I_1 > 2I_2$. In fact, since SHG ($\omega + \omega = 2\omega$) is followed by difference frequency generation ($2\omega - \omega = \omega$, DFG), for any finite mismatch $\Delta k > 0$, a portion of the FF travels for a while at 2ω before it is back-converted: this is commonly named 'cascading' and, as a consequence, part of the FF undergoes a positive phase shift. The converse happens in the DFG + SHG process. Finally, if an SH beam with a small FF seed is input, solitons form for $\Delta k < 0$ being $2I_2 > I_1$. From a physical point of view, these mechanisms create the phase-front distortion required to counteract diffraction. Figure 7 is an artist's sketch on the formation of an SHG soliton.

Equations [24] are approximately valid and rely on a plane-wave insight, but they point out the wealth of possibilities offered by three-wave interactions for light transverse localization via parametric processes. The full set of coupled equations describing parametric solitons from SHG differs from eqn [1]:

$$\begin{aligned} \mathrm{i}\frac{\mathrm{d}A_1}{\mathrm{d}z} + \frac{1}{2k_0 n_1}\nabla_\perp^2 A_1 + \frac{\omega d_{\mathrm{eff}}}{c} A_1^* A_2 \mathrm{e}^{-\mathrm{i}\Delta k z} &= 0 \\ \mathrm{i}\frac{\mathrm{d}A_2}{\mathrm{d}z} + \frac{1}{4k_0 n_2}\nabla_\perp^2 A_2 + \frac{\omega d_{\mathrm{eff}}}{c} A_1^2 \mathrm{e}^{\mathrm{i}\Delta k z} &= 0 \end{aligned} \qquad [26]$$

as it includes diffraction terms added to eqn [22].

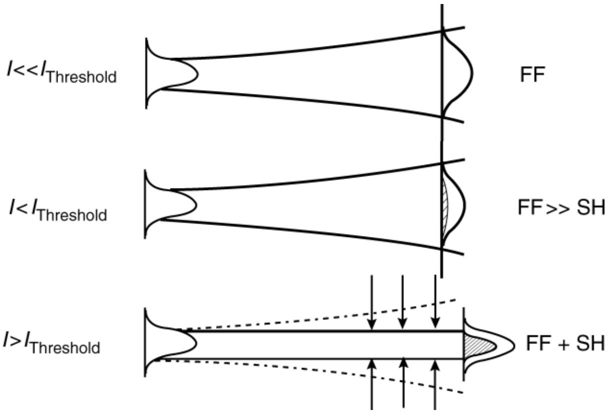
$I \ll I_{\mathrm{Threshold}}$
FF
$I < I_{\mathrm{Threshold}}$
FF>> SH
$I > I_{\mathrm{Threshold}}$
FF + SH

Figure 7 Parametric solitons via SHG. Intuitive description of soliton formation with an FF Gaussian input. At low powers (top) and for a small amount of generated SH (center) linear diffraction dominates. Above threshold parametric gain efficiently counteracts diffraction, originating a spatial soliton via quadratic cascading.

The self-trapped solutions are found as in eqn [23], with

$$\begin{aligned} A_1 &= a_1(x,y)\exp(\mathrm{i}\beta z) \\ A_2 &= a_2(x,y)\exp(\mathrm{i}\Delta k z)\exp(2\mathrm{i}\beta z), \quad \text{and} \\ a_{1,2}(x,y) &= \sqrt{I_{1,2}(x,y)} \end{aligned} \qquad [27]$$

These show that, even in the presence of diffraction, SHG can give rise to pure phase effect if the profiles satisfy:

$$\begin{aligned} \frac{1}{2k_0 n_1}\nabla_\perp^2 a_1 + \frac{\omega d_{\mathrm{eff}}}{c} a_1 a_2 &= \beta a_1 \\ \frac{1}{8k_0 n_2}\nabla_\perp^2 a_2 + \frac{\omega d_{\mathrm{eff}}}{2c} a_1^2 &= \beta + \frac{\Delta k}{2} a_2 \end{aligned} \qquad [28]$$

which are the space-dependent counterpart of eqn [25]. Again, β is a nonlinear contribution to the FF wavevector, and determines the corresponding correction $\beta_2 = 2\beta + \Delta k$ at SH. β is univocally determined by the power of the two harmonics in the solutions of eqn [28].

In the particular case:

$$\Delta k = \beta\frac{n_1 - 4n_2}{2n_2} \qquad [29]$$

1D parametric solitons exhibit profiles

$$\begin{aligned} a_1 &= \sqrt{\frac{n_1}{n_2}}\frac{3c}{4\omega d_{\mathrm{eff}}}\beta\,\mathrm{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}}x\right)^2 \\ a_2 &= \frac{3c}{4\omega d_{\mathrm{eff}}}\beta\,\mathrm{sech}\left(\sqrt{\frac{\beta k_0 n_1}{2}}x\right)^2 \end{aligned} \qquad [30]$$

and the powers per unit of transverse wavefront along y at the two components are:

$$\begin{aligned} P_1 &= \int_{-\infty}^{\infty} a_1^2(x)\mathrm{d}x = \frac{n_1}{n_2}\frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2} \\ P_2 &= \int_{-\infty}^{\infty} a_2^2(x)\mathrm{d}x = \frac{3\sqrt{2}c^2}{4\omega^2 d_{\mathrm{eff}}^2\sqrt{k_0 n_1}}\beta^{3/2} \end{aligned} \qquad [31]$$

with the total soliton power $P_1 + P_2$, directly related to β.

Equation [29] implies that, for $n_1 > 4n_2$ and therefore $\Delta k > 0$, 1D parametric solitons exhibit the sech^2 profile. Otherwise, no analytical solutions are known even in 1D, and numerical techniques need to be employed. The same applies to 2D soliton profiles in all instances.

Notably, eqns [28] are essentially identical to eqns [21] for solitons in NLC. Therefore, Figure 4 provides the features of 2D parametric solitons as well, with 'a' replaced by 'a_1', and 'Ψ' by 'a_2'. For parametric solitons $\alpha = 4n_2(1 + \Delta k/2\beta)/n_1$, showing that at phase-matching ($n_2 = n_1$, $\Delta k = 0$) $\alpha = 4$, while large α's correspond to large Δk or small β's, e.g., low power and very wide solitons. The sech^2 profile for 1D solitons corresponds to $\alpha = 1$. 1D and 2D soliton solutions exist for any $\alpha > 0$.

The model for SHG solitons is more complicated than the previously discussed Kerr or saturable cases. The stability question can be formulated as the constrained-minimum problem of a specific functional, the Hamiltonian H. In general, for a given soliton power (the sum of the powers at the two harmonics) two branches exist for H. They correspond to two solitary solutions, differing in the ratio between the two powers. The solution with $P_2 > P_1$ corresponds to higher H, and is unstable via parametric amplification. When a harmonic field propagates in the presence of energy at a sub-harmonic (e.g., a small amount of FF), it tends to break up and transfer energy to it. Looking at Figure 4, this corresponds to small α's, only attainable when $\Delta k < 0$. As a result, quadratic solitons with a relevant SH content are unstable and thus, more difficult to observe.

The generation of a parametric soliton with a Gaussian input at FF, has a threshold (see Figure 7). This is because the input has to generate sufficient SH in order to produce and 'feel' the nonlinear phase-shift. A numerical simulation of eqn [26] shows the formation of a parametric spatial soliton, as displayed in Figure 8. As shown in (a), if only a small second-harmonic is generated by an FF excitation, diffraction dominates. Above threshold, as in (b), a stable self-trapped two-color beam is generated.

Two significant experimental results are reproduced in Figures 9 and 10. The photograph in Figure 10 shows a soliton propagating in periodically poled KTP (PPKTP) over five Rayleigh ranges, obtained by imaging the out-of-plane scattered light. Parametric solitons are typically observed with the FF in the near-infrared and the SH in the visible.

Figure 10 shows the output spots at FF and SH of a beam propagating in PPKTP for various FF excitations, until a soliton is generated (right end).

We close this section by observing that more general models must be considered to describe the whole physics of solitons in media with a quadratic

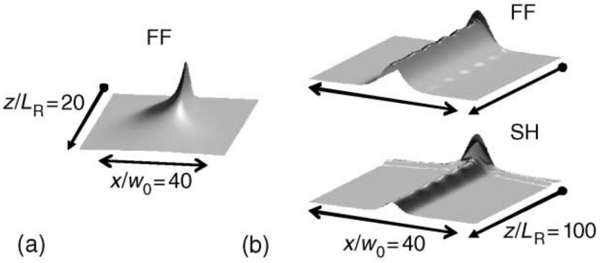

Figure 8 Parametric solitons via SHG. 1D numerical simulations using the beam propagation method. (a) An FF Gaussian excitation diffracts at low powers; (b) at high FF powers, after a transient both harmonics propagate undiffracted as a parametric spatial soliton. (Adapted with permission from Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396).

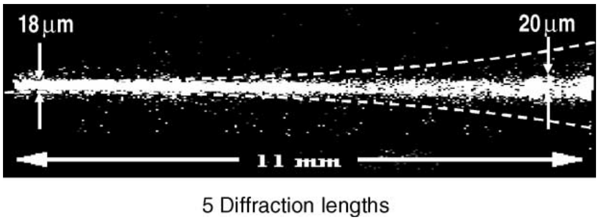

Figure 9 2D parametric solitons via SHG. Top view photograph of a soliton in PPKTP (20 ps pulses at $\lambda \approx 983$ nm and pulse energy of about 0.3 µJ). The dashed lines illustrate the diffractive behavior in the absence of nonlinear effects. (Adapted with permission from Malendevich R, Jankovic L, Polyakov S, Fuerst R, Stegeman G, Bosshard C and Gunter P (2002) Two-dimensional type I quadratic spatial solitons in KNbO3 near non-critical phase-matching. *Optics Letters* 27: 631–673.)

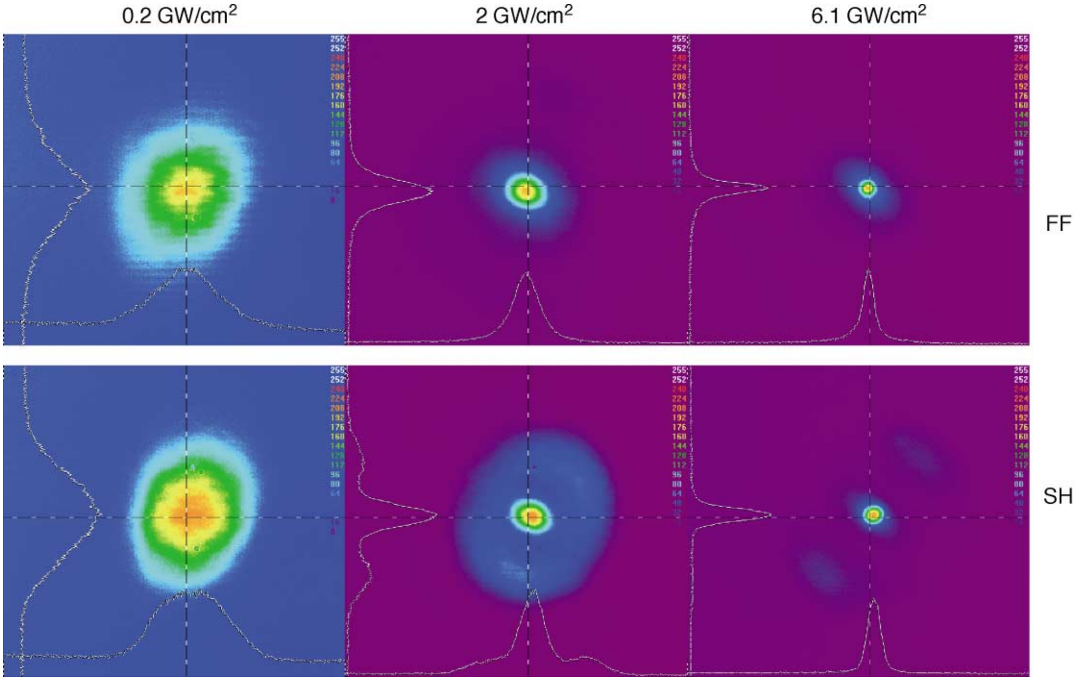

Figure 10 2D parametric solitons via SHG. Beam spots at FF (top row) and SH (bottom) for various input intensities in a PPKTP crystal. (Adapted with permission from Stegeman GI, Jankovic L, Kim H, *et al.* (2003) Generation of, and interactions between quadratic spatial solitons in non-critically-phase-matched crystals. *Journal of Nonlinear Optical Physics and Materials*. 12: 1–20).

response. We mention, for example, the subtleties associated to anisotropy (walk-off), nondegenerate three wave interactions, additional Kerr and higher-order effects.

Conclusions

Bright spatial solitons are widespread, and have been reported in several materials using various wavelengths and geometries. The brief overview presented above was intended to introduce the variety of basic mechanisms underlying their generation and propagation, but could not even touch upon the wealth of additional properties and potential applications of individual and multiple spatial solitons, their interactions and collisional properties, their stability and their control with other linear and nonlinear light waves, as well as with additional nonoptical effects (temperature, voltage, acoustic, and magnetic fields, etc.). Fascinating phenomena involve bright spatial solitons and are the subject of current research. Among them we like to mention all-optical logic and signal readdressing, beam shaping and filtering, limiting, sub-wavelength spatial compression, and digital imaging. To learn more on the subject, we encourage the interested readers to look up the vast literature covering both fundamental and applied aspects of optical spatial solitons.

List of Units and Nomenclature

Angular frequency [s^{-1}]	ω
Beam waist [m]	w
Boltzmann constant [J K^{-1}]	k_B
Current density vector [A m^{-2}]	$\mathbf{J}$
Density of ionized donors [m^{-3}]	N_d^i
Dielectric constant of vacuum [F m^{-1}]	ε_0
Director angle perturbation [rad]	ψ
Director orientation angle [rad]	θ_0
Electric field amplitude, normalized [W$^{1/2}$ m^{-1}]	a, A
Electric field amplitude, optical frequency [Vm^{-1}]	E_{opt}
Electric field vector, low frequency [V m^{-1}]	$\mathbf{E}$
Electron charge [C]	q
Electron density [m^{-3}]	ρ
Electron mobility [m^2 V^{-1} s^{-1}]	μ
Electro-optic second-order tensor [m V^{-1}]	r, r_{ijk}

Electro-optic third-order tensor [m^4 C^{-2}]	$g_{\mathrm{eff}}, g_{ijkl}$
Element of the photovoltaic tensor [C m^3 J^{-1}]	β_{ph}
Kerr nonlinear coefficient [m^{2}W^{-1}]	n_2
Low frequency permittivity [F m^{-1}]	ε_S
Medium elastic constant [N]	K
Medium forcing constant [N m^{-2}]	G
Non local decay distance [m]	R_c
Nonlinear length [m]	L_N
Nonlinear wavevector correction [m^{-1}]	β, β_2
Optic axis versor	$\mathbf{c}_{\mathrm{opt}}$
Optical background intensity [Wm^{-2}]	I_b
Optical dark intensity [Wm^{-2}]	I_{dark}
Optical index anisotropy	n_a
Optical intensity [Wm^{-2}]	I
Optical saturation intensity [Wm^{-2}]	I_S
Photo-ionization cross-section [m^2 J^{-1}]	s
Photovoltaic index perturbation constant [V m^{-1}]	E_{ph}
Power [W]	P
Power per wavefront unit [W m^{-1}]	P_1, P_2
Propagation coordinate [m]	z
Rayleigh length [m]	L_R
Recombination rate [m^3 s^{-1}]	γ_r
Refractive index	n
Refractive index perturbation	Δn
Relative dielectric constant	ε_r
Sample width [m]	L
Second order effective susceptibility [m V^{-1}]	$d^{(2)}_{\mathrm{eff}}$
Second order nonlinear coefficient [m W$^{-1/2}$]	d_{eff}
Self-focusing power [W]	P_c
Soliton power [W]	P_S
Speed of light in vacuum [ms^{-1}]	c
Static dielectric constant [F m^{-1}]	ε_s
Temperature [K]	T
Total density of negatively charged sites [m^{-3}]	N_A
Total donor density [m^{-3}]	N_d
Transverse coordinates [m]	x, y, r
Transverse wave-vector [m^{-1}]	k_x
Vacuum impedance [Ω]	Z_0
Wavelength [m]	$\lambda = 2\pi n/k = 2\pi/k_0$
Wavevector in medium [m^{-1}]	$k = n\omega/c$
Wavevector in vacuum [m^{-1}]	$k_0 = \omega/c$

See also

Babinet's Principle. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Nonlinear Optics, Applications:** Self-focusing and Related Effects (Solitons and Multiphoton Absorption). **Nonlinear Optics, Basics:** Cascading; Photorefraction.

Further Reading

Assanto G and Peccianti M (2003) Spatial solitons in nematic liquid crystals. *IEEE Journal of Quantum Electronics* 39: 13–21.

Assanto G and Stegeman GI (2002) Simple physics of quadratic spatial solitons. *Optics Express* 10: 388–396.

Assanto G, Peccianti M and Conti C (2003) Nematicons: optical spatial solitons in nematic liquid crystals. *Optics and Photonics News* 14: 44–48.

Buriak AV, Di Trapani P, Skryabin DV and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

Chiao RY, Garmire E and Townes CH (1964) Self-trapping of optical beams. *Physics Review Letters* 13: 479–482.

Conti C, Peccianti M and Assanto G (2003) Route to nonlocality and observation of accessible solitons. *Physics Review Letters* 91: 073901.

Conti C, Peccianti M and Assanto G (2004) Observation of optical spatial solitons in a highly nonlocal medium. *Physics Review Letters* 92: 113902.

Crosignani B and Salamo G (2002) Photorefractive solitons. *Optics and Photonics News* 13: 38–41.

Kivshar Y and Agrawal G (2003) *Optical Solitons: From Fibers to Photonic Crystals*. New York: Academic Press.

Stegeman GI, Christodoulides DN and Segev M (2000) Optical spatial solitons: historical perspectives. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1419–1426.

Torner L and Sukhorukov AP (2002) Quadratic solitons. *Optics and Photonics News* 13: 42–47.

Various Authors (1998) In: Kajzar F and Reinisch R (eds) *Beam Shaping and Control with Nonlinear Optics*. New York: Plenum Press.

Various Authors (2001) In: Boardman AD and Sukhorukov AP (eds). *Soliton-driven Photonics*. London: Kluwer Academic.

Various Authors (2002) In: Trillo S and Torruellas WE (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Optical Fiber Solitons, Physical Origin and Properties

G Millot and P Tchofo-Dinda, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In the last few decades there has been considerable effort made in understanding energy-localization phenomena in many areas of physics. Optical fibers constitute excellent media where such phenomena are routinely observed by means of relatively simple experimental setups. A well-known example of those energy-localization phenomena is the optical soliton, which corresponds to a light pulse that can propagate in optical fibers over relatively long distance without experiencing significant degradations. From a fundamental point of view, the high stability of the optical soliton results from a delicate balance between the fiber chromatic dispersion and the nonlinear refractive (or Kerr) index. The idea of using optical solitons as information bits in high-speed communication systems was first proposed in 1973, and then demonstrated experimentally in 1980. After these discoveries, interest in optical soliton transmission started to increase. In an ideal fiber, optical solitons can be modeled approximately by the nonlinear Schrödinger equation, whose solutions are well known. In this article we present the main types of soliton behavior and the fundamental physical processes that are involved in soliton stability.

The Nonlinear Schrödinger Equation

In optical fibers, there are many nonlinear phenomena that are difficult to understand by direct analysis of the nonlinear polarization (e.g., the soliton effect) but that can be readily understood by use of the nonlinear Schrödinger equation (NLSE). Let us consider an optical pulse that propagates in the z-direction, in a dielectric Kerr medium such as an optical fiber, with frequency ω_0, wavenumber k_0, and an envelope $\mathcal{A}$:

$$E = \mathcal{A}\exp[\mathrm{i}(\mathbf{k}_0\cdot\mathbf{r} - \omega_0 t)] \qquad [1]$$

We assume that the wave envelope $\mathcal{A}$ varies slowly in time and space compared with the fast variations of the carrier wave ω_0. In this situation, the Fourier spectrum of the wave is restricted to frequencies and wavenumbers close to $\omega = \omega_0$ and $k = k_0$, respectively. Then, the wavenumber k can be expanded in a Taylor series around k_0, as follows:

$$k(\omega) = k_0 + \left(\frac{\partial k}{\partial \omega}\right)_{\omega_0}(\omega - \omega_0) + \frac{1}{2}\left(\frac{\partial^2 k}{\partial \omega^2}\right)_{\omega_0}(\omega - \omega_0)^2 + \left(\frac{\partial k}{\partial |\mathcal{A}|^2}\right)_{\mathcal{A}_0=0}|\mathcal{A}|^2 + \cdots \qquad [2]$$

The nonlinear term $Q = (\partial k/\partial |\mathcal{A}|^2)_{\mathcal{A}_0=0}$ arises from the optical Kerr effect, as $k = n\omega/c = k(\omega, |\mathcal{A}|^2)$.

Setting $K = k - k_0$, $\Omega = \omega - \omega_0$, $\beta_1 = (\partial k/\partial\omega)_{\omega_0}$ and $\beta_2 = (\partial^2 k/\partial\omega^2)_{\omega_0}$, one can rewrite eqn [2] as

$$K = \beta_1\Omega + \frac{1}{2}\beta_2\Omega^2 + Q|\mathcal{A}|^2 \qquad [3]$$

where β_1 is the inverse of the group velocity and β_2 is the group velocity dispersion (GVD) coefficient. Equation [3] represents the nonlinear dispersion relation for the envelope. Using a Fourier transform method in which K is replaced by a spatial operator $-\mathrm{i}\,\partial/\partial z$, and Ω by a temporal operator $\mathrm{i}\,\partial/\partial T$, one can apply eqn [3] on the envelope $\mathcal{A}$ to obtain the following equation:

$$\mathrm{i}\frac{\partial \mathcal{A}}{\partial z} + \mathrm{i}\beta_1\frac{\partial \mathcal{A}}{\partial T} - \frac{1}{2}\beta_2\frac{\partial^2 \mathcal{A}}{\partial T^2} + Q|\mathcal{A}|^2\mathcal{A} = 0 \qquad [4]$$

In practice it is more convenient to make use of the quantity $A = (\alpha A_{\mathrm{eff}})^{1/2}\mathcal{A}$, for which $|A|^2$ represents the wave power. The parameter A_{eff} is the effective core area, which takes into account the variation of the light intensity across the transverse section of the fiber. The value of A_{eff} strongly depends on the mode distribution. Using a system of coordinates that moves at the group velocity v_g, i.e., replacing T by $t = T - z/v_g$, eqn [4] reduces to the basic NLSE:

$$\mathrm{i}\frac{\partial A}{\partial z} - \frac{1}{2}\beta_2\frac{\partial^2 A}{\partial t^2} + \gamma|A|^2 A = 0 \qquad [5]$$

where γ is the nonlinear coefficient defined by $\gamma = Q/(\alpha A_{\mathrm{eff}}) = \omega_0 n_2/(cA_{\mathrm{eff}})$ and expressed in $\mathrm{W}^{-1}\mathrm{m}^{-1}$. This equation is formally analogous to the Schrödinger equation in quantum mechanics but with a nonlinear potential. Furthermore, in this basic NLSE, all higher-order terms that describe effects such as the fiber losses, third-order dispersion, or stimulated Raman scattering, are neglected. The NLSE for pulse propagation in optical fibers was first derived in 1973 by Hasegawa and Tappert. It has also found applications in other areas of physics, such as in plasma physics, fluid dynamics and electrical transmission lines. From a fundamental point of view, the basic NLSE involves the combined effects of the group velocity dispersion (GVD) and self-phase modulation (SPM), which are described by the second and third terms in eqn [5], respectively. Each of these two effects is analyzed below.

Group Velocity Dispersion

The individual effect of the GVD on pulse propagation can be obtained by setting the nonlinear coefficient to zero ($\gamma = 0$). In this case, eqn [5] reduces to

$$\mathrm{i}\frac{\partial A}{\partial z} = \frac{1}{2}\beta_2 \frac{\partial^2 A}{\partial t^2} \qquad [6]$$

In the frequency domain, eqn [6] becomes

$$\mathrm{i}\frac{\partial \tilde{A}}{\partial z} = -\frac{1}{2}\beta_2 \omega^2 \tilde{A} \qquad [7]$$

where $\tilde{A}$ represents the Fourier transform of A. The solution of eqn [7] is given by

$$\tilde{A}(z,\omega) = \tilde{A}(0,\omega)\exp(\mathrm{i}\beta_2\omega^2 z/2) \qquad [8]$$

As eqn [8] shows, the GVD changes the phase of each spectral component of the pulse by an amount that depends on the propagated distance and the frequency of the spectral component. But this effect does not modify the power spectrum: $(|\tilde{A}(z,\omega)|^2 = |\tilde{A}(0,\omega)|^2)$. This phase change can modify the shape of the pulse envelope. For example, a Gaussian pulse with a width of δ_0 at $z = 0$, $A(z=0,t) = \exp\{(-t^2/(2\delta_0^2)\}$, will transform into

$$A(z,t) = [\delta_0^2/(\delta_0^2 - \mathrm{i}\beta_2 z)]^{1/2}\exp[-t^2/2(\delta_0^2 - \mathrm{i}\beta_2 z)] \qquad [9]$$

at a distance z. In other words, the pulse width increases from δ_0 to $\delta_0[1+(z/L_\mathrm{D})^2]^{1/2}$. The parameter $L_\mathrm{D} = \delta_0^2/|\beta_2|$ defines a dispersion length, that is, a length scale over which the dispersive effects become important. At $z = L_\mathrm{D}$ the Gaussian pulse broadens by a factor of $\sqrt{2}$. Thus, the GVD causes the spectral components of the pulse to move at different velocities, which broadens the pulse, as shown by Figure 1.

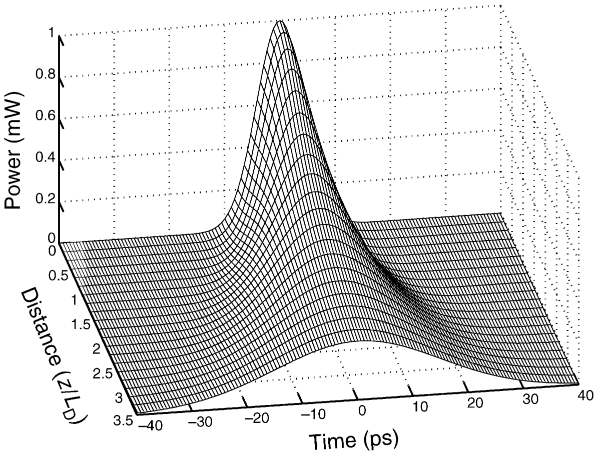

Figure 1 Temporal evolution of the power of a Gaussian pulse, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 1$ mW, $\beta_2 = -20\ \mathrm{ps}^2\ \mathrm{km}^{-1}$, $\gamma = 3\times10^{-3}\ \mathrm{m}^{-1}\ \mathrm{W}^{-1}$ and $\delta_0 = 10$ ps. The dispersion length is $L_\mathrm{D} = 1.8$ km.

Another important effect of the GVD comes from the time dependence of the pulse phase, which one can easily obtain by rewriting eqn [9] as $A(z,t) = |A(z,t)|\exp[\mathrm{i}\phi(z,t)]$. It turns out that the GVD induces an instantaneous variation of the frequency across the pulse, called a frequency chirp. For the initially unchirped Gaussian pulse, the GVD induces a chirp given by

$$\delta\omega = -\frac{\partial\phi}{\partial t} = \frac{t\beta_2(z/L_\mathrm{D})}{|\beta_2|\delta_0^2[1+(z/L_\mathrm{D})^2]} \qquad [10]$$

As eqn [10] illustrates, the GVD-induced frequency chirp depends on the sign of β_2, and varies linearly across the pulse.

Self-Phase Modulation

Similarly to dispersion effects, nonlinear effects in optical fibers require a minimum propagation distance to develop significantly. The distance over which nonlinear effects become important is defined as the nonlinear length L_NL, given by $L_\mathrm{NL} = 1/(\gamma P_0)$, where P_0 is the peak power.

Let us consider the propagation of a light pulse in a fiber of length L, such that $L_\mathrm{NL} < L \ll L_\mathrm{D}$. In this particular case, the effects of dispersion can be neglected, and then the NLS equation becomes

$$\mathrm{i}\frac{\partial A}{\partial z} + \gamma|A|^2 A = 0 \qquad [11]$$

It results from this equation that $|A(z,t)|^2 = |A(z=0,t)|^2$. This means that the pulse shape remains unchanged during propagation. The solution of eqn [11] is given by

$$A(z,t) = A(0,t)\exp[\mathrm{i}\Phi_\mathrm{NL}(z,t)] \qquad [12]$$

where $\Phi_{\mathrm{NL}}(z,t) = \gamma|A(0,t)|^2 z$. Thus, the Kerr effect causes the phase of the wave envelope to vary during the propagation. The nonlinear phase shift Φ_{NL} increases with the propagated distance z. The maximum phase shift $\Phi_{\mathrm{NL}}^{\max}$ occurs at the pulse center and is given by $\Phi_{\mathrm{NL}}^{\max} = \gamma|A(0,0)|^2 z = \gamma P_0 z$. One can see that the nonlinear length represents the propagation distance at which $\Phi_{\mathrm{NL}}^{\max} = 1$. Since the nonlinear phase shift Φ_{NL} depends on the initial pulse shape, it varies nonlinearly with time t. This implies that the instantaneous optical frequency differs across the pulse from its central value ω_0. The frequency chirp is given by

$$\delta\omega(t) = \omega(t) - \omega_0 = -\frac{\partial\Phi_{\mathrm{NL}}}{\partial t} = -\frac{\partial|A(0,t)|^2}{\partial t}\gamma z \qquad [13]$$

The modification of the phase of the wave envelope by itself is called self-phase modulation. The frequency chirp increases with the propagated distance, thus implying that new frequencies are continuously generated as the pulse propagates through the fiber. This spectral broadening depends on the initial pulse shape. Figure 2 shows the variation of Φ_{NL} and the corresponding induced frequency chirp across the pulse at $z = L_{\mathrm{NL}}$ for the case of an initial Gaussian pulse.

The SPM-induced chirp exhibits several interesting features. $\delta\omega$ is negative near the leading edge (red shift) and is positive near the trailing edge (blue shift). The chirp is linear (and increases) in the central region of the pulse. In fact, the chirp $\delta\omega$ at a given propagation distance z depends on the initial chirp. As Figure 3 shows, at a sufficiently long propagated distance, the SPM-induced spectrum exhibits several peaks due to interference effects across the pulse. Indeed as Figure 2 shows, the same frequency chirp occurs at two distinct times associated with two distinct nonlinear phase shifts.

The spectrum width can be estimated by calculating the maximum value $\delta\omega_{\max}$ of the frequency chirp. For an initial unchirped Gaussian pulse one obtains

$$\delta\omega_{\max} \approx \frac{0.86\gamma P_0 z}{\delta_0} \qquad [14]$$

Thus, the spectral width increases linearly with the propagated distance z and the input peak power P_0, and is inversely proportional to the initial pulse width δ_0. The SPM can be detrimental for optical communication systems. The SPM broadens the spectrum of the signal and makes it then more susceptible to GVD effects. Note that in wavelength division multiplexing (WDM) systems, the phase modulation of a given channel can come not only from the power in the considered channel itself (SPM) but also from the power in a different channel of the WDM system (CPM, i.e., cross-phase modulation). By appropriate dispersion compensation, the impact of SPM and CPM can be significantly reduced. On the other hand, the SPM, combined with dispersive effects, leads to very useful effects such as pulse compression, or the generation of optical solitons (described below).

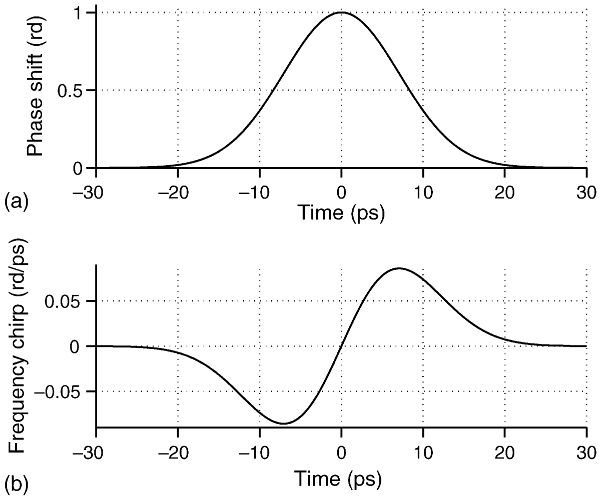

Figure 2 Temporal variation of the nonlinear phase shift (a) and the corresponding frequency chirp induced by SPM (b), for a Gaussian pulse.

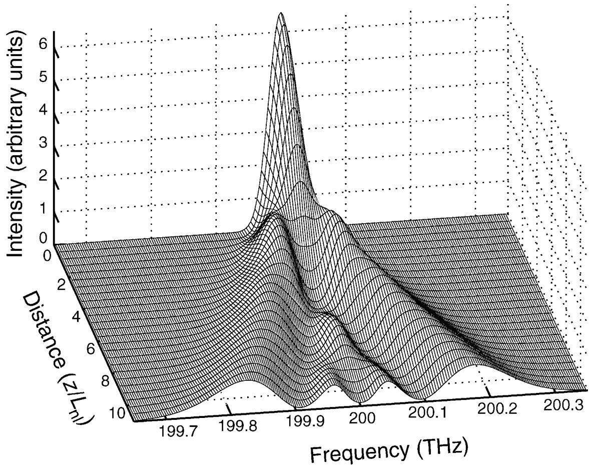

Figure 3 Evolution of the SPM-induced spectral broadening of an initially unchirped Gaussian pulse, as a function of the propagation distance. Here the normalized distance is z/L_{NL}.

Optical Solitons

Bright Solitons

Soliton propagation results fundamentally from a delicate balance between two phenomena: GVD and SPM. As we mentioned above, the GVD causes the spectral components of the pump to move at different velocities, thus leading to a temporal broadening of the pulse as shown in Figure 1 in the case of a Gaussian input pulse. Moreover, eqn [10] shows that the effect of GVD on a propagating Gaussian pulse is associated with a linear frequency chirp $\delta\omega$ and a parabolic temporal phase, proportional to β_2. On the other hand, eqn [12] shows that the SPM induces a frequency chirp across the pulse proportional to the temporal derivative of the pulse intensity. The chirp induced by SPM is then linear only in the central region of the pulse as long as the initial pulse shape is nonparabolic. As a consequence a Gaussian input pulse cannot propagate in a distortion-free manner since the GVD for anomalous dispersion ($\beta_2 < 0$) is able to completely cancel the intensity-dependent phase shift from SPM only over the central region of the pulse. On the other hand we show below that, in the anomalous dispersion regime, the exact cancellation of the nonlinear and dispersive temporal phase components (and corresponding chirps) is obtained in the case of the propagation of a hyperbolic secant pulse, which corresponds to the fundamental bright optical fiber soliton.

We consider the propagation of an unchirped hyperbolic secant pulse at a wavelength of 1.55 μm in a standard telecommunications fiber with $\beta_2 = -21 \times 10^{-3}\ \mathrm{ps^2\ m^{-1}}$ and $\gamma = 1.2 \times 10^{-3}\ \mathrm{W^{-1}\ m^{-1}}$. The input pulse at $z = 0$ is given by $A(0,t) = \sqrt{P_0}\ \mathrm{sech}\,(t/\delta_0)$ with $\delta_0 = 10$ ps and $P_0 = 175$ mW. With this choice of parameters the dispersion length is equal to the nonlinear length, i.e. $L_D = L_{NL} = 4.76$ km. In fact, these conditions correspond to those which would support a propagating soliton. However, let us consider first how SPM and GVD modify the incident pulse when acting independently. Figure 4 shows numerically computed results from the solutions of eqns [12] and [8] to illustrate the effects of nonlinear and dispersive evolution over a range of propagation distances. The figure shows the pulse intensity as well as the time-dependent phase and the corresponding frequency chirp. The evolution is clearly very different in the presence of only nonlinear or dispersive effects. Nonlinear SPM alone induces no change in the temporal intensity of the pulse, but the magnitudes of the SPM-induced phase and chirp increase with propagation distance. In contrast, GVD induces changes in both the intensity and the phase of the pulse as it propagates, and the temporal broadening with propagation is associated with a significant change in the functional forms of the phase and frequency chirp. In particular, whilst for large propagation distances it can be seen that GVD induces a parabolic temporal phase and corresponding linear chirp, this is clearly not the case for small propagation distances. This result strongly differs from that obtained for a propagating Gaussian pulse for which the temporal phase is always parabolic. In the context of soliton propagation, the results in Figure 4b are of particular significance, since they clearly show that at $z = L_D/10$ where neither SPM nor GVD noticeably modify the pulse intensity profile, the SPM and GVD-induced phases and frequency chirps have qualitatively similar forms, yet opposite sign. The split-step Fourier method demonstrates that over a sufficiently small propagation distance (e.g., $\delta z = L_D/10$) the effects of nonlinearity and dispersion can be considered to act independently on a propagating pulse. This suggests that, over the propagation segment ($0 < z < \Delta z$), the phases are capable of cancelling each other across the entire pulse profile. It is obvious that, if the effects cancel over this distance, then the propagation in the following segment will begin from the same initial condition of an unchirped pulse which was the case at $z = 0$. The nonlinear and dispersive phase components will therefore continue to cancel over successive iterations for an arbitrary length of fiber, and we thus obtain a simple physical interpretation to explain the distortion-free propagation of an optical soliton. Thus, the combined effects of GVD and SPM can prevent the occurrence of a frequency chirp across the pulse. In this situation the pulse propagates with a constant temporal profile. Such light pulses are called optical solitons.

Mathematically, the basic NLSE, eqn [5], belongs to a remarkable class of integrable nonlinear equations, which can be solved exactly for an arbitrary initial condition by means of the inverse scattering method. The solutions depend strongly on the sign of dispersion, and the relative importance of dispersive and nonlinear effects, measured by a dimensionless parameter N:

$$N^2 = \frac{L_D}{L_{NL}} = \frac{\gamma P_0 \delta_0^2}{|\beta_2|} \qquad [15]$$

In the anomalous dispersion regime ($\beta_2 < 0$) the NLS equation admits envelope-soliton solutions. The soliton order is given by the parameter N. Only the fundamental soliton, defined by $N = 1$, can propagate in the fiber with a constant profile, as schematically represented in Figure 5.

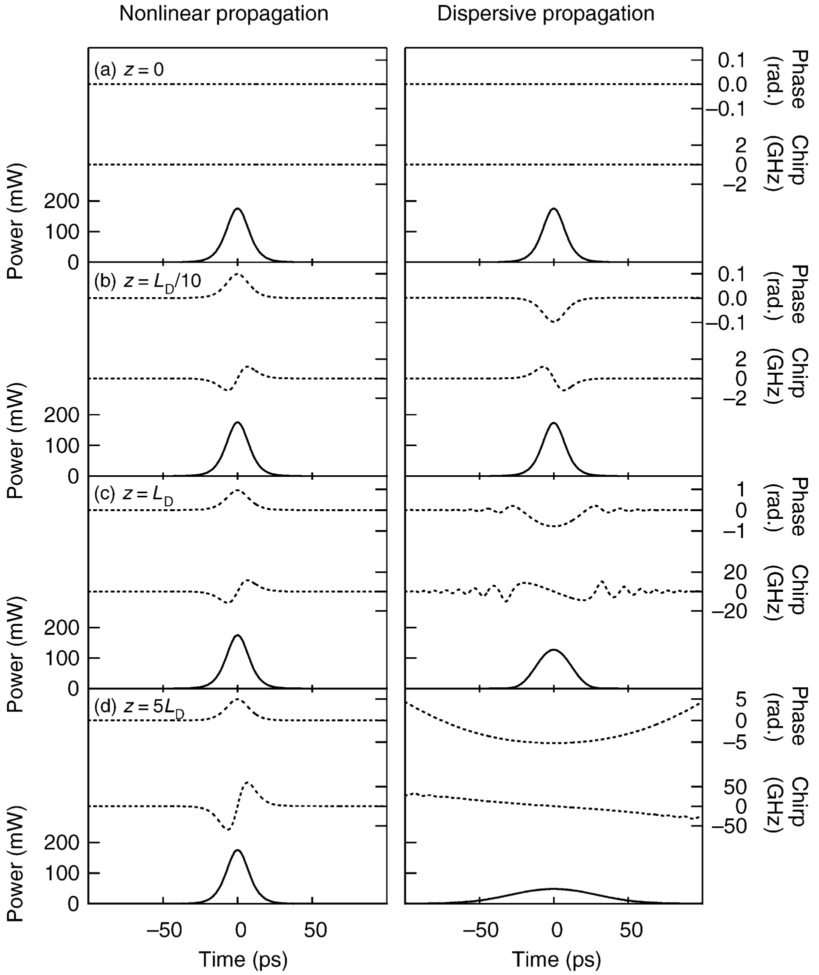

Figure 4 Numerical results showing nonlinear (left column) and dispersive (right column) propagation in an optical fiber, with parameters as described in the text, and for propagation distances as indicated. The solid lines show the pulse intensity (left axis) whilst the dashed lines show the calculated phase (right axis, top) and chirp (right axis, bottom). Constant phase offsets have been removed for clarity.

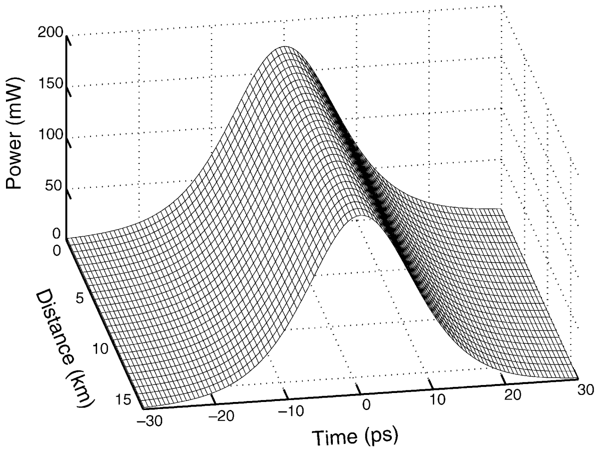

Figure 5 Temporal evolution of the power of a fundamental bright soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 67$ mW, $\beta_2 = -20\ \text{ps}^2\ \text{km}^{-1}$, $\gamma = 3 \times 10^{-3}\ \text{m}^{-1}\ \text{W}^{-1}$ and $\delta_0 = 10$ ps. The nonlinear length is $L_{NL} = 5$ km.

For the fundamental soliton, dispersion and nonlinearity exactly balance each other in a way such that neither the pulse shape nor the pulse spectrum change during the pulse propagation. The fundamental soliton ($N = 1$) is given by

$$A(z,t) = \sqrt{P_0}\ \text{sech}(t/\delta_0)\ \exp(\text{i}\gamma P_0 z/2) \qquad [16]$$

where the peak power required to generate this soliton is determined by $L_D = L_{NL}$, and given by

$$P_0 = \frac{|\beta_2|}{\gamma\delta_0^2} \qquad [17]$$

Note that the phase of this bright soliton remains constant across the entire pulse. For a standard single-mode fiber (SMF), at 1.55 μm wavelength, typical values of the dispersion and nonlinear parameters are $\beta_2 = -21\ \text{ps}^2\ \text{km}^{-1}$ and $\gamma = 1.2 \times 10^{-3}\ \text{W}^{-1}\ \text{m}^{-1}$, respectively.

So, the peak power required to propagate a fundamental soliton with pulse width $\delta_0 = 10$ ps, is $P_0 = 175$ mW. This soliton is represented in Figure 5. A further reduction of P_0 can be obtained by using dispersion-shifted fibers (DSF), for which the dispersion can be significantly reduced to values of the order of 1 ps^2 km^{-1}. In this situation, the power required to create a soliton reduces to $P_0 = 8.3$ mW. Thus, fundamental solitons can be generated at power levels that are easily obtainable from conventional laser diodes. The first experimental observation of soliton propagation in optical fibers was reported in 1980 by Mollenauer, Stolen and Gordon.

Note that all the higher-order solitons ($N > 1$) execute a recurrent deformation during their propagation, as illustrated in Figure 6 in the case of the second-order soliton ($N = 2$).

Dark Solitons

In the normal dispersion regime ($\beta_2 > 0$) the chirps induced by GVD and SPM reinforce each other and thus make the hyperbolic secant pulse more susceptible to temporal broadening. Thus, in the normal dispersion regime bright solitons cannot propagate. However, in this case the NLS equation, eqn [5], admits a hole-soliton solution. In other words, the intensity profile of this soliton exhibits a dip in a uniform background, as represented in Figure 7.

Mathematically, this hole-soliton, commonly called a dark soliton, takes the following form:

$$A(z,t) = \pm\sqrt{P_0}\,\tanh(t/\delta_0)\,\exp(\mathrm{i}\gamma P_0 z) \qquad [18]$$

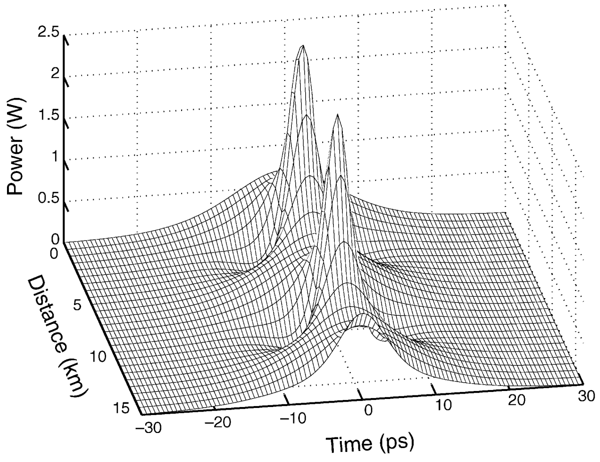

Figure 6 Temporal evolution of the power of a second-order soliton, as a function of the propagation distance. The simulation parameters are the following: input peak power $P_0 = 267$ mW. The fiber parameters are the same as in Figure 5. The nonlinear length is $L_{\mathrm{NL}} = 1.25$ km.

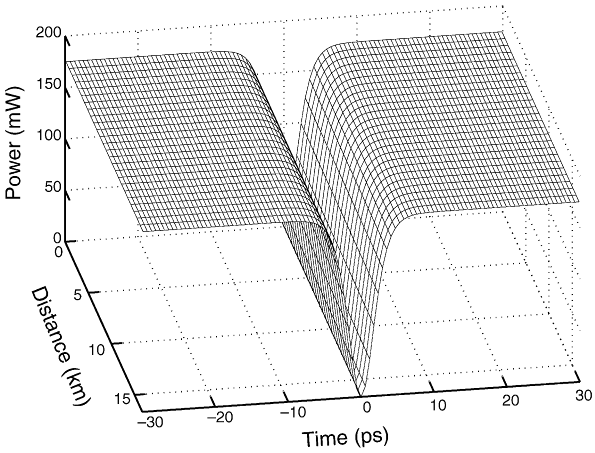

Figure 7 Temporal evolution of the power of a dark soliton, as a function of the propagation distance, for an input peak power $P_0 = 67$ mW.

As eqn [18] shows, the amplitude of the dark soliton is given by an odd function of time. Thus its phase exhibits an abrupt π jump at the dip center, thus implying that dark solitons are chirped. The phase between each jump is constant. Note that there exist other soliton solutions that are similar to the dark soliton, but for which the intensity at the dip center does not fall to zero and the phase variation is more gradual and smaller. These solutions are often called gray solitons.

A common feature of bright and dark solitons is their robustness. This property is very important for ensuring practical applications in optical communications. Moreover, optical solitons emerge unchanged from collision processes. However, dissipative perturbations such as material loss or Raman scattering can destroy such solitons.

Dispersion-Managed Solitons

As we mentioned above, the fundamental soliton is based on a delicate balance between dispersion and nonlinearity, through the relation $L_{\mathrm{D}} = L_{\mathrm{NL}}$. In addition, the existence of this soliton requires that only these two effects are present in the fiber. But in practice, additional linear and nonlinear effects such as fiber losses, Raman scattering, or third-order dispersion, are always present in the fiber. In particular, the fiber losses cause the soliton power to continually decrease during the propagation. The use of lumped amplifiers along the transmission line allows one to compensate for the power loss, but does not prevent the occurrence of chirp. As a consequence, the fundamental soliton does not survive over long distances (a few tens of kilometers), without in-line control. Moreover, this soliton is not stable against phenomena such as CPM and FWM processes that occur unavoidably in WDM systems. The CPM between the pulses in a given channel and the pulses in other channels induce a frequency chirp (in the considered channel), which is converted into amplitude variations by the fiber dispersion. The FWM induces energy-exchange processes between different channels, which lead to pulse distortions.

A new transmission scheme, called *dispersion management*, was recently proposed to resolve the above-mentioned limitations. Basically, the dispersion-management technique utilizes a transmission line with a periodic dispersion map, such that each period is built up by two types of fibers, generally with different lengths and opposite group-velocity dispersion. Whereas the fundamental soliton requires total cancellation of the chirp at each propagation distance, the dispersion-managed (DM) soliton induces a small variation of the chirp, with a zero mean value over each period of dispersion management.

The DM soliton exhibits major differences with respect to the fundamental bright soliton. First, whereas the fundamental bright soliton possesses a smooth profile (as Figure 5 shows), the profile of the DM soliton exhibits an internal oscillatory structure, with several side-lobes on each side of the soliton's center. A typical representation of a DM-soliton profile is shown in Figure 8, at the free-chirp point of the dispersion map. Another outstanding difference is the following: whereas the fundamental soliton propagates with a constant profile, the DM soliton executes a periodic deformation, with a period that coincides with that of the dispersion map. The minimum value of the soliton's width occurs at the free-chirp point. From this point, the soliton's width can execute several oscillations within one dispersion map, with a peak value that can be quite large compared with the width at the free-chirp point. Figure 9 illustrates the propagation of a DM soliton within one dispersion map of a lossless DM line.

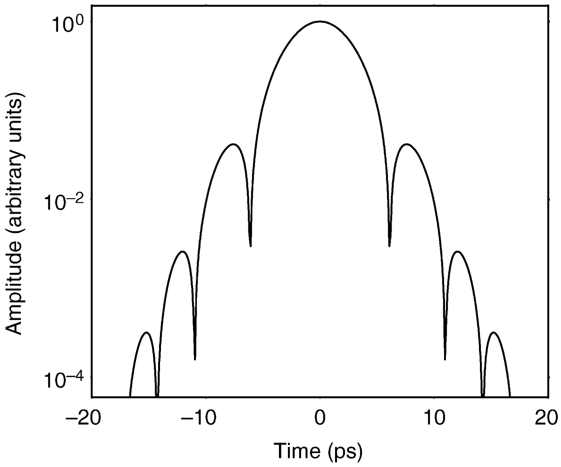

Figure 8 Plot showing the profile of a DM soliton at the free-chirp point of a typical DM line made up of 20.5 km of anomalous-dispersion fiber [$\beta_2 = -1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}], and 19.5 km of normal-dispersion fiber [$\beta_2 = 1.5 \times 10^{-3}$ ps^2 m^{-1}, $\gamma = 1.4 \times 10^{-3}$ W^{-1} m^{-1}].

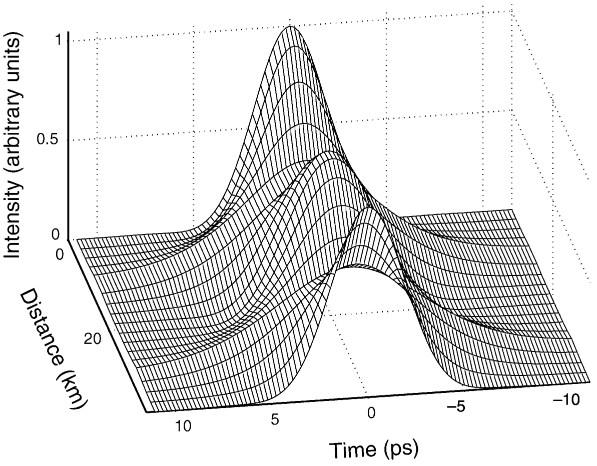

Figure 9 Plot showing the evolution of the DM soliton's profile as a function of the propagation coordinate *z*.

Recent experiments have demonstrated that in this new transmission scheme the pulse propagation becomes much more stable than fundamental solitons against linear and nonlinear effects such that SPM, FWM, CPM, or third-order dispersion. The DM transmission lines are currently subject to intense research with a view both to upgrade the capacity of existing terrestrial networks, and to design submarine fiber systems with ultrahigh capacity.

Modulational Instability

In optical fibers, the interaction between nonlinear and dispersive effects may lead to modulational instability (MI), a phenomenon in which a continuous or quasicontinuous wave undergoes a modulation of its amplitude or phase in the presence of noise or any other small perturbation. The perturbation can originate from quantum noise (spontaneous-MI) or from a frequency shifted signal wave (induced-MI). MI was observed for the first time by Tai, Hasegawa and Tomita in 1986 for a single pump wave propagating in a standard nonbirefringent fiber (scalar MI). Scalar MI occurs only when the group velocity dispersion (GVD) is negative (anomalous dispersion regime). Modulational instability in eqn [5] is examined through a linear stability analysis (LSA) of its steady state solution

$$A = \sqrt{P_0}\exp[\mathrm{i}\gamma P_0 z] \qquad [19]$$

Equation [19] shows that the power remains constant through propagation along the fiber. The linear stability of the steady state solution is examined by looking into the system in the presence of small amplitude and phase perturbations *a;* that is, we consider

$$A = (\sqrt{P_0} + a)\exp[\mathrm{i}\gamma P_0 z] \qquad [20]$$

Then, the linearization of the NLSE with respect to *a* yields

$$\frac{\partial a}{\partial z} + \frac{\mathrm{i}}{2}\beta_2\frac{\partial^2 a}{\partial t^2} = \mathrm{i}\gamma P_0(a + a^*) \qquad [21]$$

Then one can assume for the perturbation a modulational ansatz with wavenumber K and frequency Ω, of the form

$$a(z,t) = u\cos(Kz - \Omega t) + \mathrm{i}v\sin(Kz - \Omega t) \qquad [22]$$

Substitution of eqn [22] into eqn [21] leads to the following dispersion relation for the perturbation

$$K^2 = \frac{\beta_2^2\Omega^2}{4}\left[\Omega^2 + \frac{4\gamma P_0}{\beta_2}\right] \qquad [23]$$

The MI phenomenon manifests itself by an exponential growth of the amplitude of the perturbation. This behavior occurs when the wavenumber K possesses a nonzero imaginary part. The dispersion relation [23] clearly shows that MI strongly depends on the sign of the dispersion. For normal dispersion ($\beta_2 > 0$), K is always real and the steady state is stable, whereas in the anomalous dispersion regime, MI appears for $\Omega < \Omega_c = \sqrt{4\gamma P_0/|\beta_2|}$. The importance of the phenomenon is measured by a power gain defined by

$$g(\Omega) = 2|\mathrm{Im}(K)| = |\beta_2|\Omega\sqrt{\Omega_c^2 - \Omega^2} \qquad [24]$$

Figure 10 shows the MI-gain spectra at three power levels with fiber parameters appropriate for standard silica fibers at $\lambda_0 = 1.319\ \mu\mathrm{m}$.

The peak value of the gain, $g_{\max} = 2\gamma P_0$, occurs at a frequency

$$\Omega_{\mathrm{opt}} = \sqrt{\frac{2\gamma P_0}{|\beta_2|}} \qquad [25]$$

called the optimal modulational frequency. In practice, when the perturbation corresponds to quantum noise, the spectral components that are amplified by the MI-process are close to Ω_{opt}. Here, as shown by Figure 11, the MI phenomenon leads to two sidebands that are situated symmetrically with respect to the pump, at frequency detunings of $\approx \pm\Omega_{\mathrm{opt}}$ (see Figure 10 for a pump peak power $P_0 = 3$ W).

In the time domain, induced-MI leads to the break-up of the quasi-cw pump wave into a train of ultrashort pulses. The repetition rate of the pulses is determined by the frequency detuning between the signal and the pump (i.e. the modulational frequency). The temporal shape of these ultrashort pulses depends not only on the powers

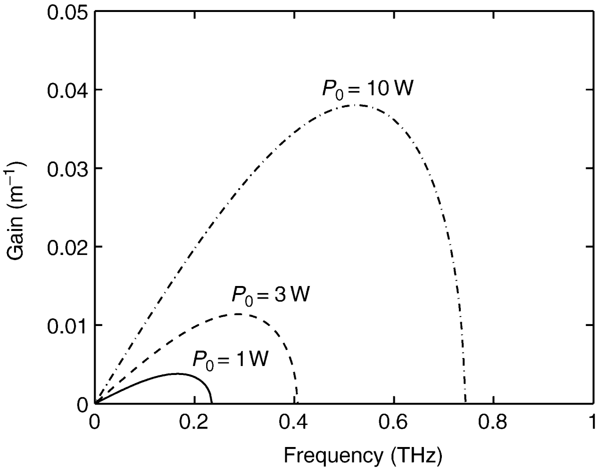

Figure 10 Spectral profiles of MI gain g for a pump peak power $P_0 = 1$ W (solid), 3 W (dashed) and 10 W (dot-dashed). The fiber parameters are $\beta_2 = -3.5\ \mathrm{ps}^2\ \mathrm{km}^{-1}$ and $\gamma = 1.9\times10^{-3}\ \mathrm{m}^{-1}\mathrm{W}^{-1}$.

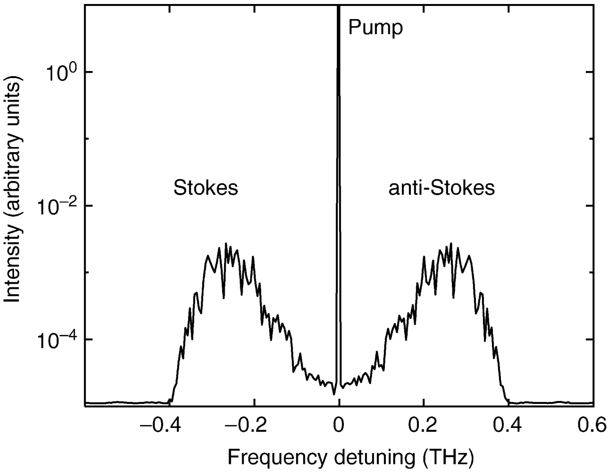

Figure 11 Spontaneous modulational instability spectrum (i.e., parametrically amplified spontaneous noise), as obtained for a peak pump power $P_0 = 3$ W and a fiber length $L = 1$ km. The fiber parameters are the same as in Figure 10.

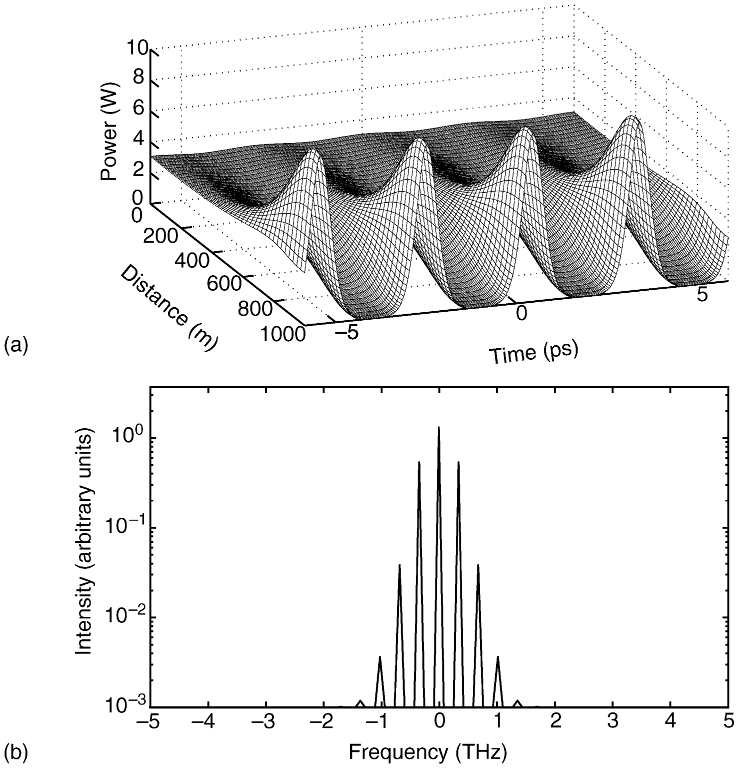

Figure 12 (a) Theoretical temporal evolution of the power for induced-MI, as a function of the propagation distance. (b) Output power spectrum in logarithmic intensity scale. The pump (signal) power is 3 W (0.5 mW), the pump-signal frequency detuning is 0.34 THz, and the fiber length is 1 km. The fiber parameters are the same as in Figure 10.

of the input waves but also on the modulational frequency. In particular, trains of bright and dark solitons with a terahertz repetition rate have been generated by MI in optical fibers. Figure 12a shows theoretically the evolution of a small initial modulation.

As Figure 12a shows, MI causes a small initial perturbation to evolve into a periodic pulse train, whose period is inversely proportional to the frequency spacing Δf between the pump and signal waves. Thus the technique of induced-MI can be exploited for generating high-repetition-rate ultrashort pulse trains. As shown in Figure 12b, such pulse train generation manifests itself in the frequency domain by the growth of a cascade of sidebands. The number of harmonics ($n\Delta f$) is determined by the temporal shape of the generated pulses and the initial conditions at the fiber input.

Cross-phase modulation (CPM) between two modes can be used to extend the instability domain to the normal dispersion regime (positive GVD). This is achieved by using orthogonally polarized pump waves. This CPM-induced MI is called vector modulational instability. The optical field evolutions are described by two coupled NLS equations. Several experiments involving two polarization modes, or two spatial modes, have been done in order to observe MI in the normal GVD regime.

Conclusion

The invention of soliton transmission, predicted in 1973 and experimentally demonstrated in 1980, is one of the major discoveries made in the area of optical telecommunication. Although from the period of invention of optical solitons up to very recent years the main line of research activities was based on the concept of the ‘classical soliton’, which represents an exact balance between the fiber group-velocity dispersion and its intensity-dependent refractive index, most recent experimental and theoretical developments focus on dispersion-managed solitons. It is now widely recognized that DM solitons in fiber-optic links are the most promising way both to upgrade the capacity of existing terrestrial telecommunication networks and to design submarine fiber systems. Most of the current research and development activities in this context aim to develop a non-empirical methodology of designing ultrahigh-speed DM fiber systems.

List of Units and Nomenclature

Dispersion length L_D [m]
Group velocity dispersion coefficient β_2 [$\mathrm{ps^2\,km^{-1}} = 10^{-27}\,\mathrm{s^2\,m^{-1}}$]
Nonlinear length L_{NL} [m]
Nonlinear parameter γ [$\mathrm{W^{-1}\,m^{-1}}$]
Nonlinear refractive index n_2 [$\mathrm{m^2\,W^{-1}}$]
Optical intensity I [$\mathrm{GW\,cm^{-2}} = 10^9\,\mathrm{W\,cm^{-2}}$]

CPM: cross-phase modulation
DM Soliton: dispersion-managed soliton
FWM: four-wave mixing
GVD: group velocity dispersion
LSA: linear stability analysis
MI: modulational instability
NLS (E): nonlinear Schrödinger (equation)
SPM: self-phase modulation
WDM: wavelength division multiplexing

See also

Fiber and Guided Wave Optics: Dispersion; Light Propagation; Nonlinear Effects (Basics). **Lasers:** Optical Fiber Lasers. **Scattering:** Scattering Theory. **Solitons:** Soliton Communication Systems.

Further Reading

Ablowitz MJ and Clarkson PA (1991) *Solitons, Evolution Equations and Inverse Scattering*. Cambridge, UK: Cambridge University Press.
Agrawal GP (1997) *Fiber-Optic Communication Systems*. New-York: Wiley-Interscience.
Agrawal GP (2001) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Agrawal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego, CA: Academic Press.
Hasegawa A (1989) *Solitons in Optical Fibers*. New-York: Springer Verlag.
Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford, UK: Clarendon Press.
Remoissenet M (1999) *Waves Called Solitons, Concepts and Experiments*, 3rd edn. Berlin: Springer.
Taylor JR (ed.) (1992) *Optical Solitons – Theory and Experiment*. Cambridge, UK: Cambridge University Press.
Zakharov VE and Wabnitz S (eds) (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Les Ulis (France) and Berlin (Germany): EDP Sciences and Springer.

Soliton Communication Systems

M Karlsson and P Andrekson, Chalmers University of Technology, Gothenburg, Sweden

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In 1973 Hasegawa and Tappert proposed that optical soliton pulses in fibers could exist. A soliton is a propagating wave packet that is localized in the sense that it does not spread its energy during propagation, and with the additional property that it is so stable that it can collide with other solitons and emerge unaffected with respect to energy, shape, and momentum after the collision. Solitons are based on some kind of nonlinearity in the system, and for optical fibers the weak Kerr nonlinearity (which makes the refractive index increase in proportion to the optical intensity) can counteract the pulse broadening induced by group-velocity dispersion (GVD). The two effects can form a stable balance in the form of a soliton pulse, which then propagates without changing shape.

Due to the lack of short-pulse laser sources at wavelengths above 1.3 μm, and low-loss silica fibers, it took another seven years for optical soliton pulses to be experimentally verified. In an experiment by Mollenauer, Stolen, and Gordon in 1980, soliton pulse transmission over 700 m fiber was demonstrated. During the 1980s soliton research aimed towards the use of solitons as information carriers in optical communications, and in 1990 the first data transmission experiment using solitons (2.8 Gb/s over 23 km of fiber) were reported by Iwatsuki *et al.*

During the 1990s, soliton-based communication systems have matured, an important reason being the development of the erbium-doped fiber amplifier (EDFA), which made high power levels commercially available. The most recent developments have been towards the use of solitons in alternating dispersion maps (so-called dispersion management), and such systems have reached performance levels near commercialization.

It should be stressed that fiber soliton research has been almost exclusively devoted to single-mode fibers at a wavelength of 1550 nm. In principle, solitons can also be used to counteract mode dispersion in multimode fibres, but that generally requires unrealistically high peak powers, and is therefore regarded to be of less interest. Similarly, solitons with wavelengths around 1300 nm in single-mode fibers have been considered, but found to be of less interest due to the lack of low-noise optical amplifiers in this spectral range. Another kind of soliton, the so-called dark soliton, can also be generated in single-mode fibers. However, dark solitons have only been used as information carriers in a very limited number of experiments, and will also be omitted from this discussion.

In the present review we will discuss both theoretical and experimental aspects of soliton transmission. We will distinguish between conventional solitons, which have constant dispersion during transmission, and dispersion-managed solitons (although solitary waves would be the proper mathematical name) for which the dispersion varies periodically during transmission.

Conventional Fiber Solitons

Group-velocity dispersion means that the group velocity of the light changes with the optical frequency, and it is usually defined as the second derivative of the propagation constant β with respect to the angular frequency ω, and denoted β_0'' [ps^2 km^{-1}]. Standard single-mode fibers have a $\beta_0'' \approx -20$ [ps^2 km^{-1}] around a wavelength of 1550 nm, whereas dispersion-shifted fibers (DSFs) have β_0'' close to zero in the 1550 nm region. The dispersion is defined with respect to sign, so that the GVD is said to be normal if the group velocity decreases with frequency (i.e., $\beta_0'' > 0$), and anomalous if the group velocity increases with frequency (i.e., $\beta_0'' < 0$). A pulse propagating in a dispersive fiber will broaden with propagation distance at a rate proportional to the product of β_0'' and the spectral width of the pulse. In a communication system such a broadening is devastating, since neighboring pulses will interfere and disperse power into adjacent bit slots. The problem is increased for high data rates (as this requires short pulses with broad spectra) and/or long transmission distances.

The nonlinearity in optical fibers manifests itself as a dependence of the refractive index with power, and is known as the optical Kerr effect. For pulse propagation the Kerr effect will give rise to self-phase modulation (SPM), which means that the pulse will get a phase modulation proportional to its own intensity. Mathematically speaking, the amplitude of an optical pulse $u(0, t)$ [W$^{1/2}$] launched at $z = 0$ will evolve according to $u(z,t) = u(0,t) \exp(-\mathrm{i}\gamma z|u(0,t)|^2)$, where γ [W^{-1} km^{-1}] is the nonlinear coefficient of the fiber, z [km] is the propagation distance along the fiber, t [ps] is the time coordinate, and $|u(z, t)|^2$ [W] is the power level of the light. The value of γ varies between different fiber types due to the dependence on the core area of the fiber, but for standard single-mode fibers, in the 1550 nm region, it is approximately 2.2 km^{-1} W^{-1}.

Taking both self-phase modulation and dispersion into account, one can derive a propagation equation for light, known as the nonlinear Schrödinger (NLS) equation. Of particular importance for the solution to the NLS equation is solitons, which is a solution that propagates without dispersive broadening. The soliton solution arises as a result of a cancellation of the dispersion and the self-phase modulation effects, and it can only be formed when the dispersion is anomalous. For normal dispersion the nonlinearity and the dispersion will instead act together to increase the spectral broadening of the pulse. The soliton solution of duration t_0 and dispersion β_0'' reads

$$u_{\mathrm{sol}}(z,t) = \sqrt{\frac{1}{\gamma L_{\mathrm{D}}}}\,\mathrm{sech}\left(\frac{t}{t_0}\right)\exp\left(-\mathrm{i}\frac{z}{2L_{\mathrm{D}}}\right) \qquad [1]$$

where the dispersive length is $L_{\mathrm{D}} = t_0^2/|\beta_0''|$ [km]. The soliton peak power P_{s} [W] and energy E_{s} [J] can be expressed as $P_{\mathrm{s}} = |u_{\mathrm{sol}}(z,0)|^2 = |\beta_0''|/\gamma t_0^2$ and $E_{\mathrm{s}} = 2P_{\mathrm{s}}t_0 = 2|\beta_0''|/\gamma t_0$.

The fact that solitons are not broadened by dispersion will make them ideally suited as information carriers in a communication system. However, the solitons are affected by various perturbations that will affect the balance by dispersion and nonlinearity. Unless proper care is taken in the design of the system, those perturbations can destroy the performance of the communication system. Next, we will review the most important perturbations and design issues of soliton-based communication systems.

Solitons in the Presence of Third-Order Dispersion

The third-order dispersion (TOD), or the dispersion slope, is modeled with the derivative of the GVD, β_0'' [ps^3 km^{-1}], which is typically 0.1 ps^3 km^{-1} in standard and dispersion-shifted fibers. For solitons in dispersion-shifted fibers the effect of the dispersion slope is increasingly important the closer the solitons are to the zero dispersion wavelength.

When a soliton is placed close to the zero-dispersion wavelength, it will emit radiation in the normal dispersion regime of the spectrum, lose energy, and recoil further into the anomalous dispersion region so that the radiation stops and a stable soliton is formed. As a design criterion $\beta_0'''/(|\beta_0''|t_0) < 0.24$ have been suggested for this phenomenon to provide negligible energy loss.

Solitons in the Presence of Amplification and Loss

The most important property that has been neglected in the derivation of the NLS equation for optical pulses is the effect of loss. The loss will decrease the power and hence the nonlinearity so that the balance sustaining the soliton is no longer maintained. As a result of the loss (which is of the order of 0.2 dBkm^{-1} in standard transmission fibers) the soliton pulse will be broadened. Partly, this can be circumvented by requiring the launched peak power of the soliton after each amplifier to be such that the path-average power between the amplifiers equals the soliton power, P_s. The peak power P_peak falls off as $\exp(-\alpha z)$, where $\alpha = 0.042\ \mathrm{km}^{-1}$ is the fiber loss coefficient, and the average peak power over one amplifier span is then $P_\mathrm{peak}\ (1 - \exp(-\alpha L_a))/\alpha L_a = P_\mathrm{peak}(G-1)/G\ \ln(G) = P_\mathrm{s}$, where $G = \exp(\alpha L_\mathrm{a})$ is the gain of each amplifier and L_a [km] is the amplifier spacing.

Nevertheless the periodic fluctuation of power and width over each amplifier length may cause the solitons to become unstable and lose power to dispersive wave radiation. To minimize this loss a useful design criterion is that the amplifier length L_a should be smaller than the dispersive length, L_D.

Sources of Timing Jitter

Another transmission obstacle is the various sources of random movement of the pulse in the bit slot, i.e., timing jitter of the pulses in the data transmission link. There are various sources of timing jitter, e.g., soliton interaction, Gordon–Haus, acoustic, and wavelength division multiplexing collision induced jitter.

Soliton interactions

Solitons, being nonlinear pulses, will interact with adjacent pulses in the pulse train. The interaction between solitons of the same polarization state and wavelength is phase sensitive so that in-phase solitons will attract each other whereas out-of-phase solitons will repel each other. To reduce this effect the solitons must be sufficiently spaced in the data stream. A typical design condition is that the pulse width of the solitons are one fifth of the bit slot, or one fifth of the inverse bit rate. A straightforward way of reducing the interaction is to have alternating amplitudes of the soliton in the pulse train, and the solitons can then be packed more densely.

Orthogonally polarized solitons interact substantially less, since it is the intensity overlap that causes the interaction, rather than the amplitude overlap as for copolarized solitons. In polarization multiplexed systems where adjacent pulse have orthogonal polarization it is usually enough with half the above separation, i.e., a pulse duration of 2.5 less than the bit separation.

Gordon–Haus jitter

Noise from the in-line amplifiers will give rise to a small jitter in the carrier frequency of each soliton, which, via the dispersion, will alter the group velocity and hence affect the arrival time of each pulse. This is known as the Gordon–Haus effect, and it has to be accounted for in long-distance systems. The variance of the timing jitter is proportional to the ratio of the power spectral density of the noise to the energy of the soliton, and also proportional to the transmission distance cubed. The fact that Gordon–Haus jitter grows cubically with distance makes it particularly important at transoceanic lengths, typically exceeding 1 Mm.

Acoustic jitter

The electrostriction nonlinearity in the fiber gives rise to a mechanical pressure proportional to the optical intensity, which in turn modifies the refractive index of the fiber. In particular, an intense optical pulse like a soliton will give rise to a pressure (acoustic) wave moving radially outwards from the fiber center. The pulses in the wake of this wave will experience a randomly changing local refractive index, and hence (just as for Gordon–Haus jitter) a randomly changing carrier frequency that transforms into a timing jitter. The variance of this jitter will grow as the transmission distance to the power of four.

WDM Considerations

Wavelength division multiplexing (WDM) is another way of increasing the bit-rate of soliton systems. That is, several frequency channels are used for solitons transmission. The problem with WDM transmission using solitons stems mainly from collisions between solitons in the different wavelength channels. In a perfectly ideal NLS equation solitons would collide elastically without changing carrier wavelength. However, the presence of losses and amplification may cause the collisions to be asymmetric if they occur around an amplifier, and the result will be a frequency displacement and a concomitant timing jitter.

Soliton Control

Soliton control is the common name for methods to control the soliton parameters such as wavelength and position, mainly in order to reduce the deleterious timing jitter effects. There are two different approaches: passive and active control.

Passive soliton control has been suggested in the form of filters that are inserted along the transmission path. This helps to keep the soliton wavelength fixed. In this way not only Gordon–Haus and acoustic jitter can be remedied, but also interaction jitter and

WDM-collision-induced jitter. A problem with this kind of filtering is that it defines a spectral region with excess gain, in which amplifier noise will grow excessively. A way around that problem is to slightly shift the center wavelength of the filters along the transmission path. In that way the solitons will follow the frequency shift, but the linear noise will not. Such sliding filter experiments have demonstrated 8 × 10 Gb/s WDM soliton transmission over 10 Mm.

Active control usually acts in the time domain by using phase or amplitude modulators to retime and reshape the solitons. Using this technique, 10 Gbit/s over unlimited distances has been demonstrated. However, this kind of active reshaping of the pulses suffers from the same drawbacks as conventional electronic regeneration, i.e., incompatibility with WDM, complexity, and high cost.

Dispersion-Managed Solitons

Introduction

In the last years of the 1990s, solitons have become substantially more attractive through the rapidly emerging strategy of improving the performance of soliton transmission with dispersion management (DM). While the DM strategy, which involves altering the local dispersion between a large positive and a large negative GVD such that the average GVD is small, has long been used in linear systems, it was only relatively recently appreciated that the same technique, if properly implemented, gives rise to several very striking improvements over conventional soliton transmission systems. While DM-solitons are clearly nonlinear pulses they are by no means classical solitons.

From a commercial viewpoint, however, the most important benefit with using DM-solitons is that they, in principle, can use the already installed conventional fibers (with zero dispersion at 1300 nm) allowing a much more cost-effective upgrade together with dispersion-compensating fibers (DCF) or chirped fiber gratings, which can be located together with in-line amplifiers at certain intervals in the link.

In Figure 1, an example of breathing in spectral and temporal widths of a DM-soliton can be seen, together with its dispersion map at the bottom.

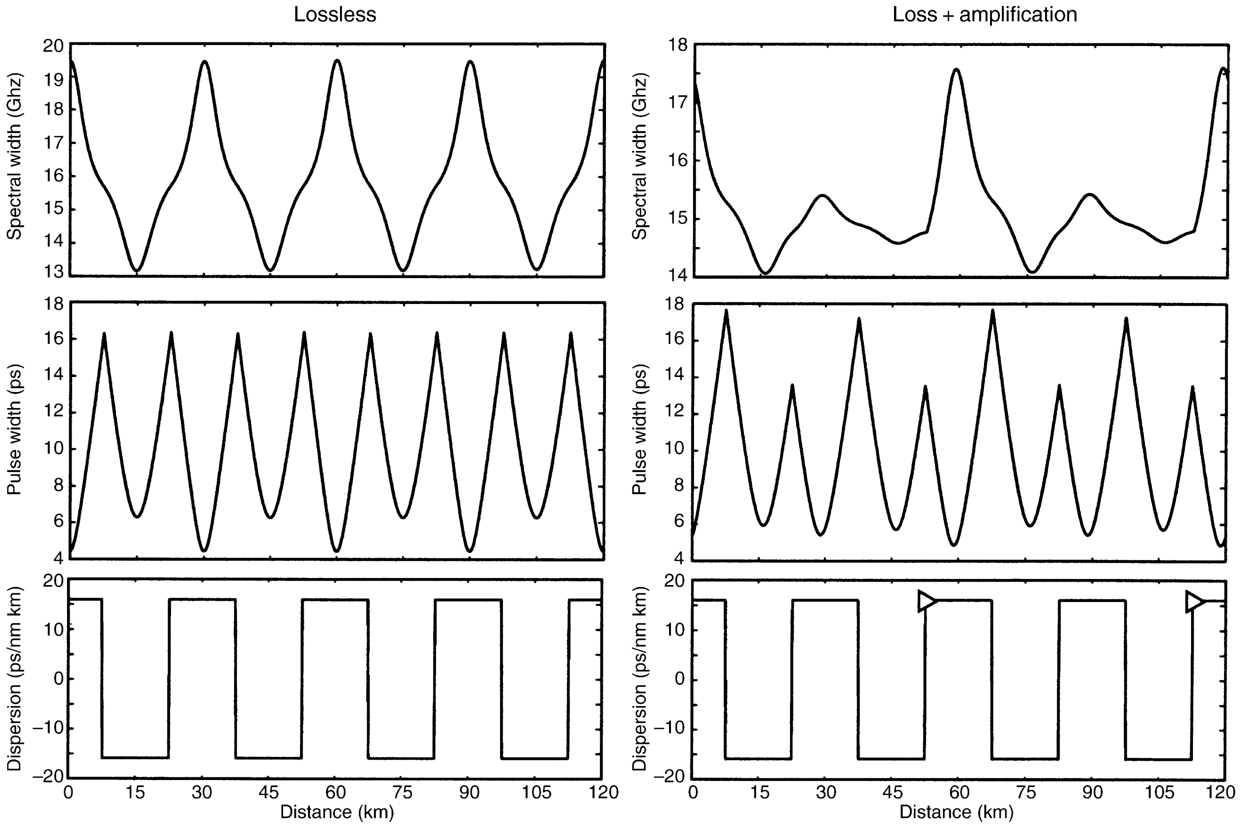

Figure 1 Plot of the spectral width (top), temporal width (middle), dispersion map (bottom) of a dispersion-managed soliton in the lossless case (left column) and the lossy case that is periodically amplified (right column). The positions of the amplifiers are denoted by the small triangles in the plot of the dispersion map.

Two cases are shown; one without fiber attenuation (left column) and one including fiber losses and periodic, lumped amplification every 60 km. One may note the strictly periodic behavior of both the spectral and temporal widths in both cases, and that the evolution of the spectral width is significantly different in the lossy case. This is because a change of the pulse spectrum requires nonlinearities and high power, and the spectral broadening is therefore highest directly after the amplifiers.

Properties

DM-solitons emerged from extensive simulation work, and it was quite surprising to many researchers that the simulations revealed such stable and strictly periodic pulses. The periodicity follows that of the dispersion variation of the system. Usually the system under study has a dispersion map, i.e., a plot of how the GVD varies along the transmission path. The map has a characteristic period (typically 10–100 km) and an average GVD which is significantly lower than the local GVD in the map.

As a result of the massive simulation work done by many groups the following properties of DM-solitons have been found:

- The pulse width, chirp, and spectral width oscillates periodically in the dispersion map. There are two points within the period at which the pulse is chirp-free, and those correspond to local minima of the pulse width. One of those points is the global minimum width, referred to as the 'shortest pulse' below.
- A central parameter that is useful for the characterization of DM-solitons is the map strength, $S = (L_1|\beta_1''| + L_2|\beta_2''|)T_{\text{FWHM}}^2$, where L is the length, β'' is the dispersion, T_{FWHM} is the minimum pulse width in the full width half maximum sense, and the subscripts 1 and 2 refer to the two fibers in the dispersion map. In this example we thus assume the map to consist of two fibers, one with normal and one with anomalous dispersion. Physically the map strength is the number of dispersive lengths the pulse propagates in one period. DM-solitons have been found for map strengths ranging from $S = 0$ (which is the same as conventional solitons) to $S \approx 12$, although this upper limit is a transition regime in which the pulses radiate and perfect periodic evolution never arises.
- DM-solitons have been found for anomalous, normal, and zero average dispersion β'', defined as $\beta_0'' = (L_1\beta_1'' + L_2\beta_2'')/(L_1 + L_2)$. Normal average dispersion is only possible for map strengths above 3.9.
- The shortest pulse shape ranges from hyperbolic secant at $S = 0$ to Gaussian for higher map strengths, and this is also evident from the time-bandwidth product, which increases with S from 0.32 (sech-shape) at $S = 0$ to 0.44 (Gaussian shape) and even higher for large values of S. In addition, the shortest pulse has oscillating tails in the wings.
- The energy of a DM-soliton pulse is enhanced relative to a soliton with the same average dispersion and pulse width.
- The interaction between DM-solitons is less than that of conventional solitons, and an optimum map strength exists that minimizes the interaction.

The fact that DM solitons can work for an average net zero GVD and even for normal dispersion, was a striking and unexpected difference from conventional solitons. This can be understood by the fact that the spectrum is widest in the anomalous dispersion regime (i.e. positive values of the dispersion parameter D, cf. Figure 1), and it has been conjectured that the average dispersion, weighted by the spectral width of the DM-soliton, should always be anomalous. This is then a generalization of the fact that conventional solitons must always have a constant anomalous dispersion.

The technical improvements with DM solitons over conventional ones are numerous. The signal-to-noise ratio is improved since the DM-solitons have a larger peak power than the corresponding conventional solitons. DM-solitons have less Gordon–Haus and acoustic timing jitter, since the system average GVD is much smaller in these systems. A very important added benefit appears in WDM systems. Due to the alteration between large positive and negative GVD along the path, the jitter induced from WDM-soliton collisions is greatly reduced. This, in turn, allows for very dense WDM which will improve the spectral efficiency substantially.

An important practical consequence of using DM-solitons is that they reduce (or eliminate) the need for in-line soliton control such as synchronous modulation or sliding filters. Yet soliton control methods are still applicable and will give improvement in terms of signal-to-noise ratio for DM-solitons as well.

Quite impressive circulating loop experiments including WDM have been reported. For example, 51 densely packed WDM channels each operating at 20 Gb/s were transmitted over 1000 km with 100 km sections of standard fiber, clearly demonstrating the strength of the DM soliton technique.

Quasi-linear Transmission Intrachannel Impairments

As a rule, map strengths in the range 1–8 make best use of the unique features of DM-solitons. This means that the use of installed standard fiber becomes difficult at very high bit-rates (say beyond 30–40 Gb/s) as shorter pulses require a more rapidly (sub-km) varying dispersion map to maintain a proper S-value. If this cannot be maintained, i.e. for S values significantly above 10, one reaches the quasi-linear regime, in which the dispersion significantly dominates the transmission.

In the quasi-linear regime, there are a couple of nonlinear transmission impairments that need to be accounted for and analyzed in more detail. They are the so-called intra-channel effects; intra-channel four-wave mixing (ICFWM), and intra-channel cross-phase modulation (ICXPM), and arise due to the nonlinear interaction between two neighboring pulses. Four-wave mixing (FWM) and cross-phase modulation (XPM) are usually effects associated with WDM transmission. However, the fact that the pulses are chirped and broadened, will cause different frequency components from neighboring pulses within the same wavelength channel to overlap in time, thereby causing FWM and/or XPM within the same channel.

ICFWM arises for two neighboring pulses that, via four-wave mixing (FWM), creates new frequency components that in the time domain will give rise to a new pulse (commonly referred to as a ghost pulse), next to the two. The ghost pulse will then give rise to intersymbol interference and reduction of the eye opening. ICFWM is most prominent for large map strengths and high power.

ICXPM can be viewed as pulse-to-pulse interaction, and physically, it manifests as the frequency shift of one pulse induced by the presence of a neighboring pulse, which, by the dispersion transforms into a timing jitter. The effect can be minimized by selecting proper map strength and pre-chirp of the pulses.

Finally it should be emphasized that these intra-channel impairments will affect also non-soliton systems (e.g., non-return-to-zero (NRZ) modulated systems) if the nonlinearities are significant.

Experiments and Field Trials

Soliton Pulse Sources

When doing soliton experiments, be it conventional or DM-solitons, particular importance is placed on the properties of the pulse source, as it sets the lower limit of the system performance. A high bit-rate soliton pulse source needs to produce low-chirp, low timing-jitter pulses with proper duration (in the ps regime), repetition rate (10–40 GHz), and shape.

One possible choice is gain-switched (GS) laser diodes, possibly with an external cavity for tunability. However, they suffer the drawback of producing pulses that are strongly chirped, asymmetric and often too wide.

For laboratory experiments, fiber ring lasers (FRL) are very attractive, as they provide wavelength and pulse width tunability, besides meeting the above demands. Their drawback is that they are bulky, need active stabilization, and sometimes also temperature control to achieve long-term stability.

Finally it appears quite clear that electro-absorption modulators (EAM), which can be integrated or not with a distributed feedback (DFB) laser, are very useful and simple sources for soliton transmission. While such sources were developed for linear NRZ systems, they have now proven to be near ideal in soliton systems as well. Although EAMs are not commercially available at 40 GHz yet, they are likely to be so soon.

Special considerations need to be taken in DM-soliton systems, however, as the launch condition is different from that for conventional solitons. The pulses should have a linear chirp such that it fits seamlessly in the periodically induced chirp variation along the link. This can be achieved by incorporating a proper length of fiber (or chirped fiber grating) in the transmitter once the overall dispersion map is known.

Loop Experiments

In order to investigate really long distances (megameters) of transmission, loop experiments were developed in the early 1990s. This means that the data pulses are injected in a loop consisting of transmission fiber and amplifier, and then left to propagate a number of laps corresponding to a certain transmission distance. Acousto-optic switches are used to switch the pulse train in and out from the loop at proper time intervals. The drawback of loop experiments is that they may be poor models of reality when it comes to things like dispersion variation along the fiber, polarization mode dispersion (PMD), or various kinds of drift that may arise. In addition, a real system has more options to fine tune, e.g., amplifiers along the transmission line. However, as long as these drawbacks are recognized, loop experiments are very powerful indeed, and invaluable in laboratory evaluations of long-distance transmission.

Field Trials

In the field, there are many transmission link design restrictions and fiber properties which make the systems far from optimal. The actual fiber parameters are nonperiodic with propagation distance as the systems are straight lines rather than relatively short loops. Both loss (in particular when including many contacts and splices along the link) and dispersion might vary significantly along a fiber span. In addition, it may not be possible to tailor the dispersion map and amplifier locations to reach an optimal state. All these examples of nonidealities justify the need for field experiments.

Several soliton field experiments have been conducted over the past few years in Japan by NTT, in the US by MCI/Pirelli, as well as in Europe by ACTS projects, and the results from those experiments are discussed below. It is a good indication that solitons are indeed foreseen as very interesting candidates in commercial systems. Again, it is not very easy to compare the results as the situation in each case differs. All of the systems operated in the 1550 nm range, used optical time division demultiplexing with a 10 Gb/s electronic base rate, and the average loss/km ranged from 0.24 to 0.33 dB/km. Dispersion-shifted fiber was always used, apart from in two cases, where standard fiber was used. One field trial has used DCF for dispersion compensation, which makes this the only DM-soliton field experiment to date.

Polarization multiplexing, for which adjacent pulses have orthogonal polarization, have been used in a few experiments and this serves mainly to allow the use of relatively wide pulses, which in turn allows for larger amplifier spans. Polarization-multiplexing, however, is not as useful if the PMD of the system is high, as then the orthogonally polarized pulses would start to drift statistically in time relative to each other, thereby creating intersymbol interference and increasing the soliton interaction. In the 40 Gb/s cases and above, PMD was found to be the main capacity limiting factor.

Most of the recent experiments used a mode-locked fiber ring laser as a source, mainly because these provide excellent pulse quality as well as tunability in terms of wavelength and pulse width. Other experiments used either gain-switched lasers or electro-absorption modulators.

Future soliton field trials are expected to (1) take advantage of the now well-known strategy of improving soliton transmission performance with dispersion-management, this being very attractive for upgrading existing fiber plants, (2) implement dense WDM (in fiber lines that do not contain dispersion-shifted fibers) to boost aggregate capacity, (3) utilize different forms of in-line control, particularly at high bit rates, and (4) further address the implications of PMD and techniques to combat it. The interesting trade-off between wavelength and time division multiplexing (WDM–TDM) for optimization of overall aggregate capacity will depend on the details of the fiber line parameters.

Future Outlook

To conclude, we note that the motivation for using solitons as information carriers has changed over the years. The property of being resistant to dispersive broadening was originally the main feature, but this was considered less important when dispersion-compensating fibers became commercially available. Instead, this led to the development of the dispersion-managed soliton. The advantages of the DM-soliton over linear transmission are features like the large power which gives a high signal-to-noise ratio. On the other hand, the difference between linear and nonlinear pulses is becoming increasingly fuzzy, and perhaps the distinction should be made between return-to-zero (RZ) and NRZ modulation rather than between linear and nonlinear transmission.

It is nevertheless interesting to note that solitons are now not only considered for oceanic systems but also for shorter terrestrial systems. There are still several challenges and opportunities remaining in order to take full advantage of solitons and to reach a better understanding. Further work is also needed on WDM soliton- and very high-speed TDM soliton-systems. The use of DM-solitons is a very recently established technique and there are thus many issues to consider. These include studies of robustness to deviations of optimum conditions, e.g., improper pulse launch conditions, the impact of nonperiodic dispersion maps and of PMD (both of which are difficult to study in loop experiments), and intrachannel effects. Nevertheless, solitons have now reached a level of maturity such that commercialization seems very near.

List of Units and Nomenclature

Amplifier gain	G	[−]
Amplifier separation	L_a	[km]
Average dispersion	β_0''	[ps^2 km^{-1}]
Dispersive length	L_D	[km]
Fiber loss coefficient	α	[km^{-1}]
Fiber nonlinear coefficient	γ	[W^{-1} km − 1]
Group velocity dispersion (GVD)	β_0''	[ps^2 km^{-1}]

Map strength	S	[–]
Optical field amplitude	$u(z, t)$	[$\mathrm{W}^{1/2}$]
Pulse duration	t_0	[ps]
Soliton energy	E_s	[J]
Soliton peak power	P_s	[W]
Third-order dispersion	β_0'''	[$\mathrm{ps}^3\ \mathrm{km}^{-1}$]
Time	t	[ps]
Transmission distance	z	[km]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics). **Solitons:** Temporal Solitons.

Further Reading

Agarwal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. San Diego: Academic Press.

Agarwal GP (2001) *Nonlinear Fiber Optics*, 3rd edn. San Diego: Academic Press.

Agarwal GP (2001) *Applications of Nonlinear Fiber Optics*. San Diego: Academic Press.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. Oxford: Clarendon Press.

Iannone E, Matera F, Mecozzi A and Settembre M (1998) *Nonlinear Optical Communication Networks*. New York: Wiley.

Mollenauer LF, Gordon JP and Mamyshev PV (1997) Solitons in high bit-rate long-distance transmission. In: Kaminow IP and Koch TL (eds). *Optical Fiber Telecommunications III A*. San Diego: Academic Press.

Smith NJ and Doran NJ (1995) Evaluating the capacity of phase modular-controlled long-haul soliton transmission. *Optical Fiber Technology* 1: 218–235.

Suzuki K and Nakazawa M (1995) Recent progress in optical soliton communication. *Optical Fiber Technology* 1: 289–308.

Taylor JR (ed) (1992) *Optical Solitons–Theory and Experiment*. Cambridge: Cambridge University Press.

The final report of the ACTS-MIDAS project, Jan 1999, can be downloaded at *www.infowin.org/ACTS/RUS/PROJECTS/FINAL-REPORTS/fr-053.pdf*.

Turitsyn SK, Mezentsev VK and Shapiro EG (1998) Dispersion-managed solitons and optimization of the dispersion management. *Optical Fiber Technology* 4: 384–452.

Temporal Solitons

S Trillo, University of Ferrara, Ferrara, Italy

A Tonello, Université de Bourgogne, Dijon, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The fact that wave propagation phenomena are intrinsically dispersive, that is localized disturbances tend to spread, is part of our experience. Less well known is the fact that the spreading of wavepackets can be counteracted by the nonlinear response of the host medium. Solitons are wavepackets that exploit this principle, realizing a perfect balance of dispersion and nonlinearity that allows them to self-trap and travel undistorded, or at least to recover their shape periodically, for long (virtually unlimited) distances. More specifically, here we deal with optical temporal solitons constituted by envelopes, or pulses with relatively narrow bandwidths around a carrier optical frequency, whose temporal profile remains unchanged (either strictly or on average) upon propagation. In contrast with spatial solitons predicted earlier and such that nonlinearities are exploited to compensate for diffractive beam spreading even in the two transverse dimensions (i.e., in bulk samples, for details the reader is referred to **Solitons:** Bright Spatial Solitons), temporal soliton trapping is intrinsically a one-dimensional phenomenon typically observed in optical fibers where diffraction is compensated by the standard fiber guidance mechanism. Temporal solitons were first predicted and observed in single-mode fibers (SMFs). Their applications in digital optical communications have increased due to their capability to behave as nearly ideal bits. Meanwhile, however, the concept of temporal soliton has been considerably generalized by investigating new settings and effects such as polarization, parametric amplification (*see* **Nonlinear Optics, Basics:** $\chi^{(2)}$-Harmonic Generation), gratings (*see* **Fiber Gratings**), stimulated scattering (*see* **Scattering:** Scattering Phenomena in Optical Fibers), and fiber lasers (*see* **Lasers**: Optical Fiber Lasers), etc. Here we briefly review the essential features of these different families of temporal solitons. In particular, in the section on Fiber Temporal Solitons below, we introduce the basic model for fiber solitons, whereas we discuss their application to optical communications in the section on Solitons in Optical Communications below. In the sections on Multi-component Temporal Solitons, and Temporal Solitons in Bragg Gratings below, we introduce significant generalizations of the concept involving multicomponent and slow-light grating solitons, respectively. Finally, in the section on Solitons on Dissipative nonlinear optics we discuss issues related to solitons in strongly dissipative systems. Note that, throughout the paper, the word soliton is used in the aforementioned general sense, which does not necessarily require the underlying model to be integrable (*see* the section on Fiber Temporal Solitons below,), as implied by the more restrictive and rigorous mathematical meaning of soliton. We further focus our discussion to bell-shaped solitons (for dark solitons, holes on a continuous-wave background, *see* **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), which can be excited by means of standard laser pulses.

Fiber Temporal Solitons

An optical pulse is described by the complex electric-field envelope $E(z,t)$ that modulates the carrier frequency ω_0. Its evolution along the z axis of a silica SMF is described with good approximation by the following universal model for dispersive and weakly nonlinear systems, known as the nonlinear Schrödinger equation (NLSE),

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \gamma |E|^2 E = 0 \qquad [1]$$

In eqn [1] the second term is responsible for dispersive spreading of pulses ruled by the group-velocity dispersion

$$k' = \left.\frac{\mathrm{d}^2 k}{\mathrm{d}\omega}\right|_{\omega_0}$$

(GVD, known also as chromatic or second-order dispersion), while the first-order dispersion

$$k' = \left.\frac{\mathrm{d}k}{\mathrm{d}\omega}\right|_{\omega_0}$$

has been removed in eqn [1] by introducing the retarded time $t = T - k'z$ measured in the pulse frame traveling at the natural group-velocity $V \equiv k'^{-1}$ of light, T being the physical time in the laboratory frame. The cubic term in eqn [1] is responsible for self-phase modulation (SPM) that stems from the instantaneous optical Kerr effect characterized by the effective coefficient $\gamma = k_0 n_{2I}/A_{\mathrm{eff}} = (\omega_0/cA_{\mathrm{eff}})n_{2I}$, where n_{2I} is the nonlinear index of silica, i.e., the nonlinear refractive index change divided by the intensity which has produced it. The real-world total electric field can be recovered as $\mathcal{E}(z,r,t) = \mathrm{Re}[f(r)E(z,t)\exp(\mathrm{i}k_0 z - \mathrm{i}\omega_0 T)]$, where the transverse modal profile $f(r)$ has been averaged out in eqn [1] to yield an effective fiber area A_{eff}.

The mechanism behind soliton formation can be understood as due to the instantaneous frequency changes along the pulse caused by the Kerr SPM, compensating those due to the group velocity dispersion (GVD) in the anomalous case ($k'' < 0$, at carrier wavelengths longer than 1.3 μm in standard SMFs) where the GVD induces a blue-shift on the pulse leading edge and a red-shift on the pulse trailing edge (in the absence of such compensation the pulse broadens because anomalous GVD means that blue-shifted frequencies travel faster than red-shifted frequencies). Such mechanism of compensation becomes ideal in the case of a soliton waveform. This concept finds its mathematical support in the fact that eqn [1] is a conservative (Hamiltonian) model with the remarkable property to possess infinite conserved quantities, a rather exceptional feature which makes it exactly integrable. Among the solutions, the following fundamental (so-called $N = 1$, see below) bright temporal soliton exists whenever $k'' < 0$:

$$\begin{aligned} E(z,t) = {} & \frac{1}{\sqrt{\gamma L_{\mathrm{d}}}}\,\mathrm{sech}[\eta(t - \mathrm{d}k'z - t_0)] \\ & \times \exp\left[-\,\mathrm{i}\delta\omega t + \frac{\mathrm{i}}{2}\left(\frac{1}{L_{\mathrm{d}}} + k''\delta\omega^2\right)z + \mathrm{i}\phi_0\right] \end{aligned} \qquad [2]$$

where t_0 and ϕ_0 are arbitrary initial position and phase, and $L_{\mathrm{d}} = (k''|\eta^2)^{-1}$ is the characteristic distance after which the GVD causes a significant pulse broadening, the so-called dispersion length. Temporal solitons [2] are characterized by two independent parameters, namely the duration $1/\eta$ and the frequency detuning $\delta\omega$, and possess several remarkable features:

(i) they travel with invariant shape by realizing the condition $N \equiv L_{\mathrm{d}}/L_{\mathrm{nl}} = 1$, i.e., exact balance between the two length scales associated with eqn [1], namely L_{d} and the nonlinear length $L_{\mathrm{nl}} = (\gamma P_1)^{-1}$ associated with the peak power $P_1 = (\gamma L_{\mathrm{d}})^{-1} = |k''|\eta^2/\gamma$ needed for the input pulse [2] to be a soliton;

(ii) in any given fiber with fixed GVD k'' and nonlinearity γ, a larger soliton power P_1 is required for a shorter duration, i.e., narrower solitons are taller;

(iii) solitons experience a phase retardation as a whole (particle), which increases with P_1, i.e., taller solitons have larger shifts;

(iv) as a consequence of Galileian invariance, a frequency shift $\delta\omega$ from ω_0 induces the soliton velocity to change from k'^{-1} to $(k' + \mathrm{d}k')^{-1}$, with $\mathrm{d}k' = \mathrm{d}k''\delta\omega$;

(v) unlike other waves in nonlinear media, collisions between solitons with different velocities are strictly elastic, as shown in the example of Figure 1a; and

(vi) solitons are extremely robust: for instance they can be excited with input peak powers differing from P_1, whereas a weak attenuation $\alpha\,(\alpha L_\mathrm{d} \ll 1)$ induces the soliton to reshape adiabatically, i.e., the peak power decreases, yet the pulse remains locally a soliton by adjusting its width, as displayed in Figure 1b.

More general (and cumbersome) solutions of eqn [1] describe periodic soliton evolutions, such as the interaction of in-phase adjacent input pulses at the same frequency (see Figure 1c), or a single input pulse with sech-shape but peak powers corresponding to higher integer values of N, so-called breathers because of their ability to recover periodically the initial shape after 'breathing' (see Figure 1d for $N = 3$) over the distance $\pi L_\mathrm{d}/2$.

Solitons [2] describe stable propagation-invariant pulses under ideal conditions (lossless case, relatively long pulses and short SMFs). In practical settings, the importance of solitons stems from their robustness against real-world perturbing phenomena described by terms that in turn break the integrability of the underlying model [1]. Situations of this kind arise, for instance, when one accounts for stimulated Raman scattering whose effect is to down-shift continuously the central frequency of ultra-short solitons, or in optical communication systems treated below.

Solitons in Optical Communications

Solitons [2] represent ideal bits for RZ (return-to-zero) optical data sequences. However, fibers have unavoidable losses of about 0.2 dB/km at the usual operating wavelength $\lambda_0 = 1550$ nm. To ensure long-haul connections, the optical signal must be periodically amplified, commonly using erbium-doped fiber amplifiers (EDFAs). The pulse peak power hence

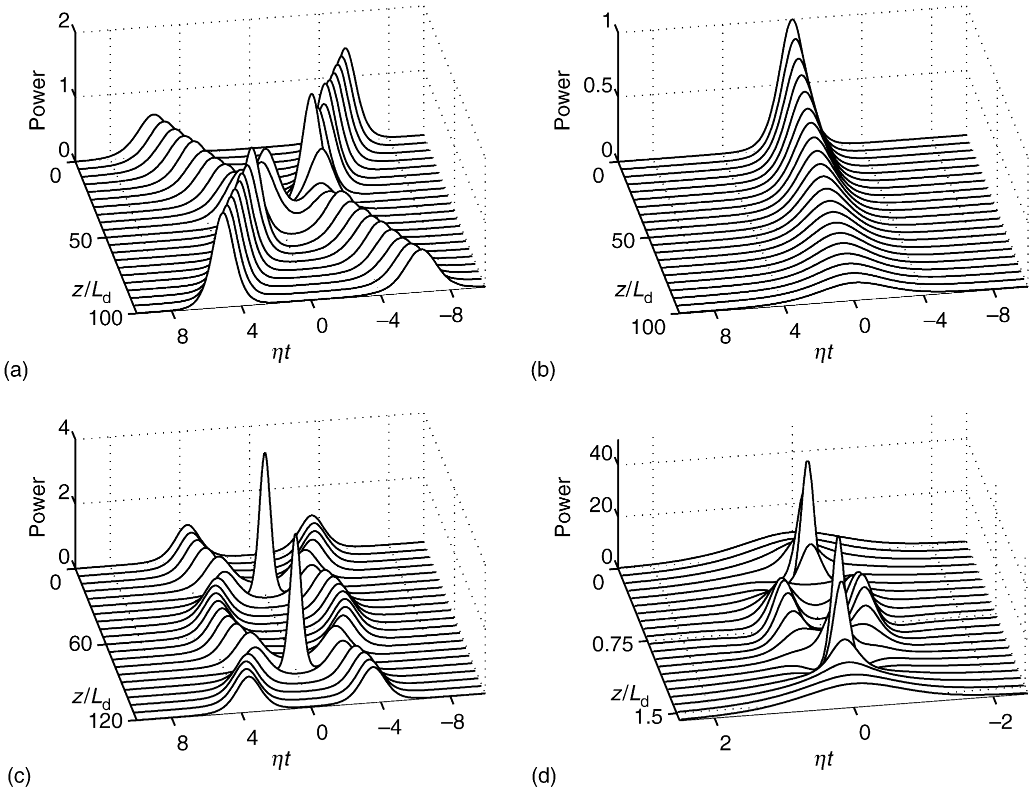

Figure 1 Fundamental features of fiber (NLSE) solitons: (a) elastic collision of two $N = 1$ solitons with different values of parameter pairs (frequency–velocity and amplitude–width); (b) adiabatic soliton damping in the presence of losses; (c) two-soliton bound state (in-phase interacting solitons of equal amplitudes); (d) evolution of a $N = 3$ soliton breather along one period of its evolution. Here the power $|E|^2$ is scaled in units of fundamental soliton power P_1, while distance z and time t are reported in units of L_d and initial pulse-width $1/\eta$, respectively.

undergoes both large exponential decay, because of distributed losses and abrupt lumped amplification under the action of EDFAs. Therefore, the stable local balance between GVD and nonlinearity cannot be rigorously sustained in the fiber. If both losses and amplifiers are accounted for, the mathematical model for pulse envelope propagation along M spans becomes:

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2 E = -\mathrm{i}\alpha E + \mathrm{i}[\exp(\alpha Z_\mathrm{A}) - 1] \times \sum_{n=1}^{M} \delta(z - nZ_\mathrm{A})E(nZ_\mathrm{A}, t) \qquad [3]$$

where the RHS contains the perturbing terms arising form distributed losses (α) and lumped amplifications (described by the comb with spacing Z_A).

The analysis of nonintegrable systems, such as eqn [3], can be carried out by appropriate transformations that allow the recovery of a reduced integrable model. The key result is that, under the limitations set by an appropriate length scale, the solitons of the reduced system can be close to those of the original problem. For temporal solitons, such transformation was first called 'the guiding center soliton', envisaging a similarity with the guiding center motion of a charged particle in a magnetic field. The main outcome is that eqn [3] can be reduced to the standard NLSE [1], clearly showing that soliton-like propagation still occurs in the presence of losses (even large, $\alpha L_\mathrm{d} \gg 1$) and rapid periodic amplification. The appropriate scale for these systems is the amplifier spacing Z_A, that must be much shorter than L_d. To reproduce an average behavior similar to the lossless case, it is sufficient to multiply the input condition for an enhancement factor that counteracts the effect of fiber losses. Therefore soliton-like solutions do exist even in fiber systems with loss management, though on average (over several fiber spans) rather than locally.

Another issue that makes solitons suitable as optical bits, lies in their intrinsic aptness to be controlled either through passive or through active systems. Solitons tend to follow their relative center of mass in the temporal or spectral domain. For instance, a powerful technique to overcome one of the major problems, i.e., the growth of noise due to amplified spontaneous emission, may be obtained by deploying periodically along the line optical passband filters. Mollenauer and co-workers have demonstrated that, by sliding the filters progressively (i.e., changing slightly the central peak from filter to filter), one can guide a shape-invariant soliton which adapts its carrier frequency while effectively suppressing the noise. Similar control techniques involve amplitude or phase modulators and a consistent amount of work has focused on the problem of all optical soliton regeneration.

To exploit the idea of soliton in optical communications, however, we should consider the fact that ultra-short optical pulses, suitable for high transmission rates, require extremely low fiber GVD to accomplish the stable balance between dispersion and nonlinearity for a fixed pulse-width and available power. For instance, in order to transmit at 40 Gbit/s, we should achieve a fixed GVD coefficient $D = -(2\pi c/\lambda_0^2)k'' \simeq 0.25$ ps/(km nm) for the whole link.

On the other hand, in many optical systems the fiber nonlinearity is not exploited at all but rather considered as a perturbing effect and a source of penalty that limits the maximum transmission distance. In such an approach the desired cumulated GVD should be as close as possible to zero to avoid errors coming from inter-symbol interference caused by dispersive broadening. Dispersion-shifted fibers may ensure low GVD only in a narrow spectral region, thus limiting the transmission bandwidth. Conversely GVD can be reduced to zero on average, simply combining fiber spans with opposite GVD. The periodic alternance of fiber types may also carry several benefits, e.g., limiting the impact of resonant nonlinear interactions such as four-wave mixing (*see* **Fiber and Guided Wave Optics:** Nonlinear Effects (Basics) for its definition). The different architectures of dispersion management proposed with the twofold scope of limiting nonlinear effects and reducing the cumulated GVD, have immediately stimulated the investigation of their impact on soliton propagation. In the early schemes, the attention was focused on the so-called adiabatic dispersion management, where the GVD was changed proportionally to the soliton power to improve soliton system performances. In contrast with previous works, nonadiabatic map profiles were introduced in the experiments by Suzuki and co-workers to compensate the cumulated dispersion by locating periodically dispersion compensating fibers (DCFs). In net contrast, Doran and co-workers have suggested the use of a combination of fibers with alternate normal and anomalous GVD (i.e., a periodic stepwise variation of GVD), predicting the existence of solitons with a peak power larger than the equivalent classical soliton power, i.e., dispersion-managed solitons (DMSs) The periodic stepwise GVD map is not only conceptually different from earlier designs (with exponentially decaying GVD), but has opened the field to real application due to its simplicity of implementation. The innovative discovery that special pulse waveforms may propagate in periodically dispersion managed fiber links has triggered the interest of several groups, leading to a complete characterization of the problem, both theoretically and experimentally.

Similar to the case of guiding center solitons, a DMS usually shows a solitonic behavior. However, the DMS has a Gaussian shape surrounded by oscillating tails (see Figure 8 in **Solitons:** Optical Fiber Solitons, Physical Origin and Properties), thus being more suitable, due to its faster decaying tails, to be closely packed in a bit sequence in comparison with a sech-soliton (eqn [2]) of the same FWHM. We show an example of DMS transmission line in the top frame of Figure 2, where we sketched the variation of the GVD of a standard SMF periodically balanced by shorter segments of DCFs with opposite GVD. We report, in the middle frame, the DMS pulsewidth and chirp and in the bottom frame the pulse peak power. As shown, the macroscopic soliton parameters are characterized by large fluctuations: the local minima of pulsewidth correspond to chirp-free points and at such points one observes either a minimum or a maximum of the pulse spectral bandwidth. Nonetheless, all these pulse parameters have periodical evolutions that are fingerprints of an invariant and soliton-like pulse when observed at integer multiples of Z_A. The exact periodicity preserves the pulse waveform and hence the carried information. In DMS transmission we assist at the threefold compensation of energy, GVD and nonlinearity. Similarly to the guiding center solitons, the threefold compensation has an average, rather than pointwise, connotation. However, unlike guiding center solitons, for DMSs the definition of average should take into account the different relevance of pulse peak power in the map. Consequently the nonlinearity is compensated in a weighted manner, and it is not surprising in this context (though astonishing if placed back in the context of the NLSE [1] which requires anomalous GVD), that DMSs have been observed even when the uniform average GVD is weakly normal.

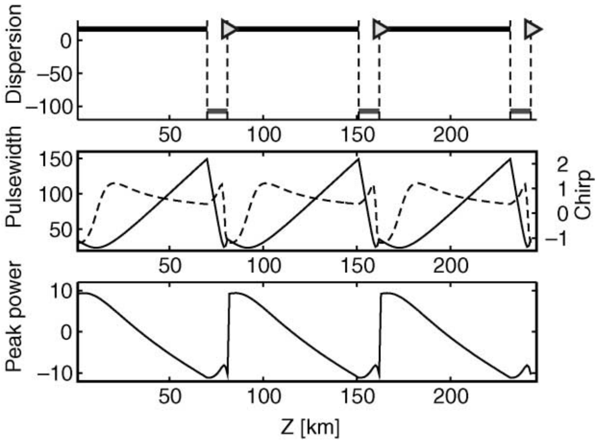

Figure 2 Evolution of soliton parameters along three periods of a fiber link with lumped amplifiers and dispersion management. Top frame: GVD map D(z) [ps/(nm km)] and EDFAs (triangles) location; Middle frame: soliton chirp (in units of 1000 ps^{-2}, dashed line), and pulse-width (in ps, solid line) vs. distance. Bottom frame: DMS peak power (dBm) vs. distance.

The net difference between the local and average GVD allows simultaneous control of different detrimental phenomena: for instance four-wave mixing is reduced by the large local GVD, whereas a small average GVD limits the impact of timing jitter due to amplified spontaneous emission (Gordon–Haus effect). As a result, DMSs perform better than conventional solitons.

Figure 3 shows a typical high-speed (20 Gb/s) single-channel experiment, and the achieved performance in terms of quality factor Q versus propagation distance, with the output after 8100 km. We recall that the Q factor measures the bit-error-rate (an error rate lower than 10^{-9} corresponds to a Q higher than $\sim$ 16 dB). DMSs also allow the control of collisions between solitons that pertain to different channels (we recall that solitons at different carrier frequencies go at different velocities, and hence collide), thus enhancing performances of standard or dense WDM systems.

Multi-Component Temporal Solitons

So far we have discussed temporal solitons described by a single envelope. However, the same mechanism of mutual balance between nonlinearity and dispersion can also be effective for multiple pulses, leading to trap simultaneously multicomponent (or so-called vector) solitons. Solitons of this type are, for instance, those due to the interaction of polarization modes in isotropic or birefringent Kerr media, higher-order modes in large-core (multimode) fibers, or mixing of pulses at different carrier frequencies that interact parametrically (i.e., without net exchange of energy with the medium). The latter case entails a broad class of solitons since virtually any parametric four-photon mixing process in cubic media (e.g., nondegenerate or partially degenerate four-wave mixing, third harmonic generation, ...), or three-photon mixing process in quadratic media (sum- and difference-frequency generation, second-harmonic generation (SHG), optical rectification, etc.) is allowed to sustain them. This opens up the field to an entirely new class of materials, namely noncentrosymmetric (quadratic) materials which usually have stronger nonlinear responses if compared with silica fibers. SHG can be taken as a paradigm to

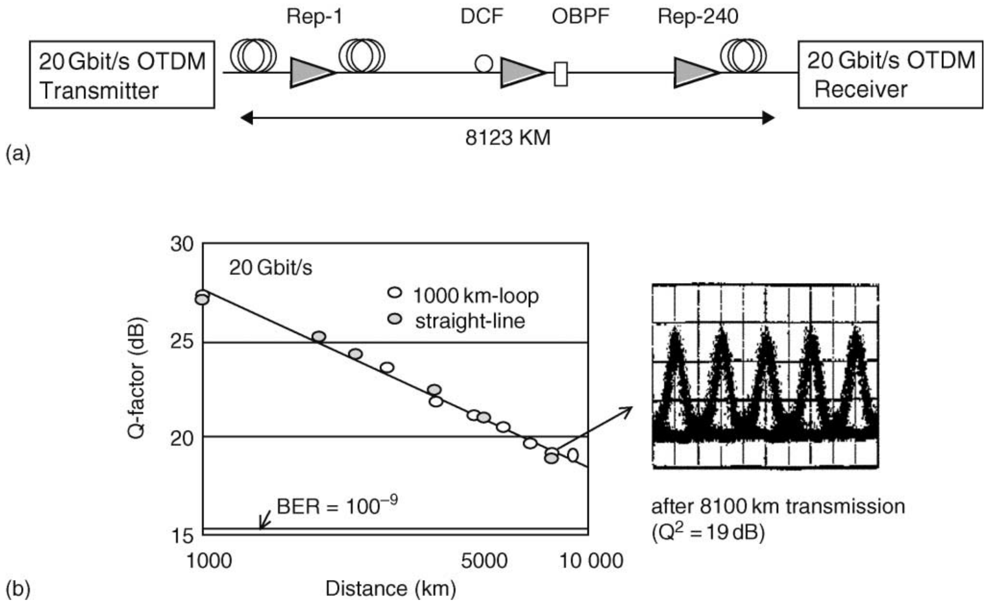

Figure 3 (a) Schematic diagram of the straight-line experiment with optical time-division multiplexing (OTDM) of soliton at 20 Gbit/s; (b) Measured *Q* factor vs. distance comparing experimental results obtained in the straight-line and loop configuration (loop length of 1000 km), respectively. The waveform received after 8100 km is also shown. With permission from Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929. © 2003 IEEE.

understand the main features of these multicomponent solitons. In this case, the evolution of the interacting envelopes E_j at carrier frequencies $j\omega_0$, $j = 1, 2$, is ruled by the coupled equations:

$$\left(\mathrm{i}\frac{\partial}{\partial z} - \frac{k_1''}{2}\frac{\partial^2}{\partial t^2}\right)E_1 + \chi E_2 E_1^* \exp(\mathrm{i}\Delta k z) = 0, \tag*{[4]}$$

$$\left(\mathrm{i}\frac{\partial}{\partial z} + \mathrm{i}\delta k'\frac{\partial}{\partial t} - \frac{k_2''}{2}\frac{\partial^2}{\partial t^2}\right)E_2 + \chi E_1^2 \exp(-\mathrm{i}\Delta k z) = 0,$$

where the subscript $j = 1, 2$ indicates that the relative quantities evaluated at frequency $j\omega_0$, $t = T - k'_1 z$, $k''_{1,2}$ are GVD coefficients, $\delta k' \equiv k'_2 - k'_1$ is the group-delay, χ is the effective nonlinear coefficient, and $\Delta k = k_2 - 2k_1$ is the wave-number (or phase) mismatch. In spite of the fact that eqns [4] are not integrable, their stable soliton solutions can be found (numerically) in a wide range of parameters as symbiotic phase-locked pulses $E_{1,2}(z,t)$ that travel undistorted. For instance, SHG solitons exist either in phase-matching ($\Delta k = 0$) or far-off phase-matching (with some limitations for $\Delta k > 0$), and, unlike fiber solitons, also in normally dispersive media ($k''_{1,2} > 0$). As for fiber solitons, their width is related to their peak power, since the basic mechanism is the mutual balance between the dispersive broadening (that depend on the pulsewidth) and nonlinear phase-shifts (that depend on power), which occur due to repeated up-conversion and down-conversion processes, often referred to as cascading. A SHG soliton is robust enough to permit the observation of trapping, even from an input that differs considerably from the soliton (e.g., from the fundamental only), or in the presence of a relatively strong group-delay which tends to rip apart the two-component soliton, inducing walk-off between the two envelopes.

In contrast with fibers, in this case, experiments are necessarily made in relatively short bulk samples, thus resulting in a limited number of dispersion lengths. Usually broad (quasi-planewave) beams are conveniently employed in order to make diffraction ineffective.

The main experimental problem is that, in crystals commonly employed for SHG, the dynamics of typical laser pulses (duration above 100 fs) is fully dominated by the group-delay $\delta k'$. In other words, the material GVD becomes relevant only for pulses as short as a few fs. Therefore, the observation of solitons with longer pulses has been required to tailor the dispersion, which can be accomplished by means of pulses with tilted phase front. As displayed in Figure 4, successful compression of 200 fs (FWHM) tilted pulses was observed at relatively large negative Δk in a BBO bulk crystal with effective zero group-delay and enhanced (anomalous) GVD.

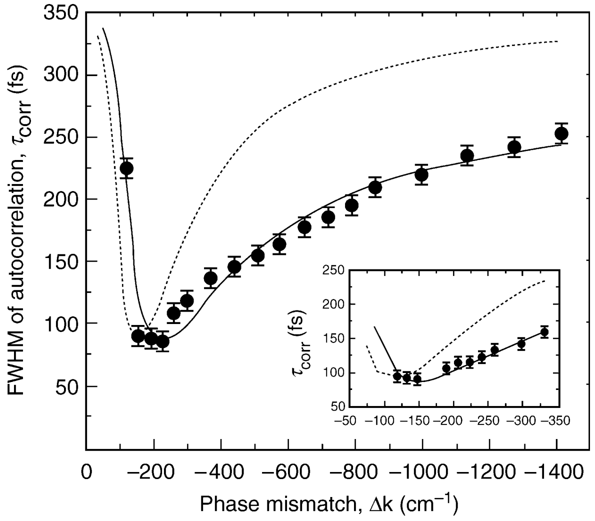

Figure 4 Soliton fomation via SHG: the circles show the output pulse duration (full width at half maximum (FWHM) of autocorrelation $\tau_{\rm corr}$) vs. mismatch Δk, measured after propagation of an input 200 fs tilted pulse (input $\tau_{\rm corr} = 284$ fs) with 13.4 GW/cm^2 peak intensity, in a 7 mm BBO crystal (inset: same with 10.4 GW/cm^2). Superimposed (dotted line) is the corresponding theoretical value obtained from eqn [4]. The solid curve that fits better the data accounts for the additional effect of the material Kerr effect. With permission from Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physical Review Letters* 81: 570–573. © 1998 APS.

The measured compression is a signature of soliton formation, as clearly indicated by the example of simulated pulse dynamics reported in Figure 5: the pulse at second-harmonic is generated and then trapped with the input one as a two-component bound state. Radiation and oscillations occur, in this case, because the launching conditions (single envelope) are far enough from the two-component soliton.

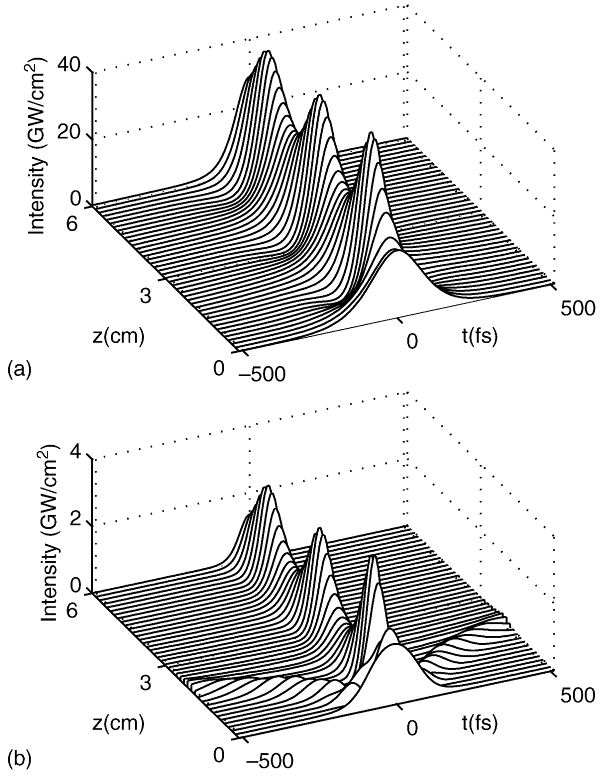

Figure 5 Dynamics of temporal soliton formation ruled by eqn [4]. A compressed soliton is formed via strongly mismatched ($\Delta k = -100\ \text{cm}^{-1}$) SHG from a 200 fs (FWHM) input pulse at fundamental carrier frequency, in a crystal with $\chi = 1.4 \times 10^{-4}\ \text{W}^{-1/2}$, effective dispersion $\delta k' = 0$ (group–velocity matching) and $k_1'' = -1.8\ \text{ps}^2/\text{m}$, $k_2'' = -0.75\ \text{ps}^2/\text{m}$.

Temporal Solitons in Bragg Gratings

A linear periodic structure, such as a Bragg grating with pitch Λ, couples counter-propagating waves when their frequency lies close to the Bragg frequency $\omega_{\rm B}$ (vacuum wavelength $\lambda_{\rm B} = 2n_0\Lambda$) defined by the resonance condition $k(\omega_{\rm B}) = \pi/\Lambda$. Such a type of structure supports a different type of temporal solitons, so-called gap solitons (GSs) since they originate from a linear dispersion relationship of the type shown in Figure 6a, characterized by a stop-band (gap) centered around ωB. The gap is characterized by a unitary grating reflectivity, as shown in Figure 6b, and the forward linear propagation is therefore forbidden. However, if the incoming wave is intense enough to induce the refractive index to increase through the optical Kerr effect, the Bragg resonance (and the whole gap) is shifted downwards (in frequency) allowing for self-transparency at the operating frequency. GSs can be understood as bell-shaped envelopes which are allowed to travel along the grating due to a high-intensity core that bleaches the reflectivity, while the high reflectivity seen by the tails takes the pulse together. GSs are expected to be slow because the slope of the dispersion curve tends to zero when approaching the gap, while its curvature results in the grating GVD (huge if compared with material GVD in real-world units), which is exactly balanced by the nonlinearity when the GS is formed. This naive portrait is supported by a more thorough analysis starting from the coupled-mode equations obeyed by the forward $E_+(z,T)$ and backward $E_-(z,T)$ envelopes at frequency $\omega_{\rm B}$, propagating along a (e.g., fiber) Bragg grating with Kerr effect (inducing SPM and

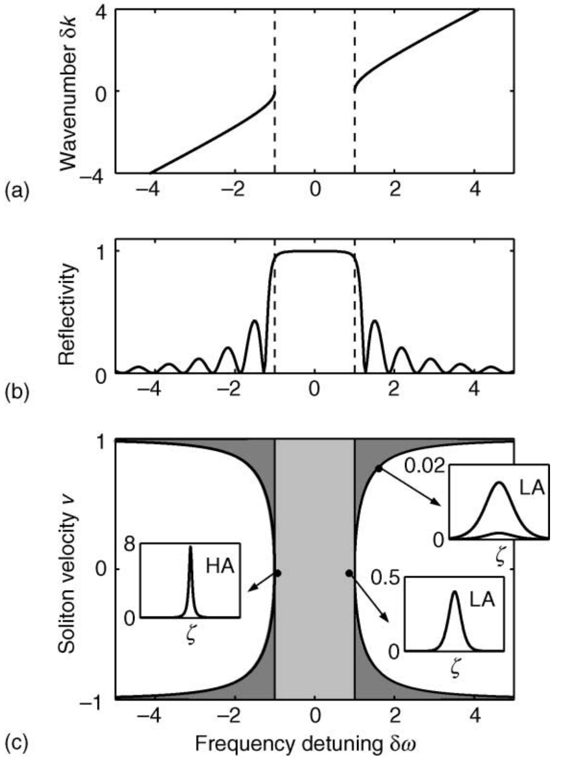

Figure 6 (a) Bragg linear dispersion relationship in normalized (dimensionless) units: wave-number shift $\delta k = (k(\omega) - \pi/\Lambda)/\Gamma$ vs. frequency detuning $\delta\omega = k'(\omega - \omega_B)/\Gamma$; (b) linear grating reflectivity vs. $\delta\omega$ for a grating figure of merit $\Gamma L = 4$; (c) existence domain (whole shaded area) of GSs in the normalized detuning–velocity plane ($\delta\omega, v \equiv k V_{sol}$). The light shaded area corresponds to the stop-band $|\delta\omega| < 1$ (inner region between dashed lines in a–b). The insets show samples of GS intensity profiles vs. the moving frame coordinate $\zeta = (1 - v^2)(z - V_{sol}T)$. Zero-velocity GSs have symmetric envelopes ($|E_+| = |E_-|$) and range from high-amplitude (HA) to low-amplitude (LA) across the gap. The third inset shows a moving GS in the experimentally accessible region, where it retains LA but has a stronger component in the direction of motion.

cross-phase modulation):

$$\mathrm{i}\left(k'\frac{\partial E_\pm}{\partial T} \pm \frac{\partial E_\pm}{\partial z}\right) + \Gamma E_\mp + \gamma\left(|E_\pm|^2 + 2|E_\mp|^2\right)E_\pm = 0 \qquad [5]$$

The linear dispersion associated with eqn [5] is as in Figure 6a, with a stop-band $|\omega - \omega_B| \leq \Gamma/k'$ proportional to the index corrugation depth through the Bragg coupling coefficient Γ. In the nonlinear regime, GS solutions of eqn [5] differ qualitatively from conventional fiber solitons [2]:

(i) first, GSs are indeed slow since the two envelopes can travel locked with any soliton velocity V_{sol} such that $|v| \equiv |k'V_{sol}| \leq 1$, i.e., lower than the natural group velocity k'^{-1};

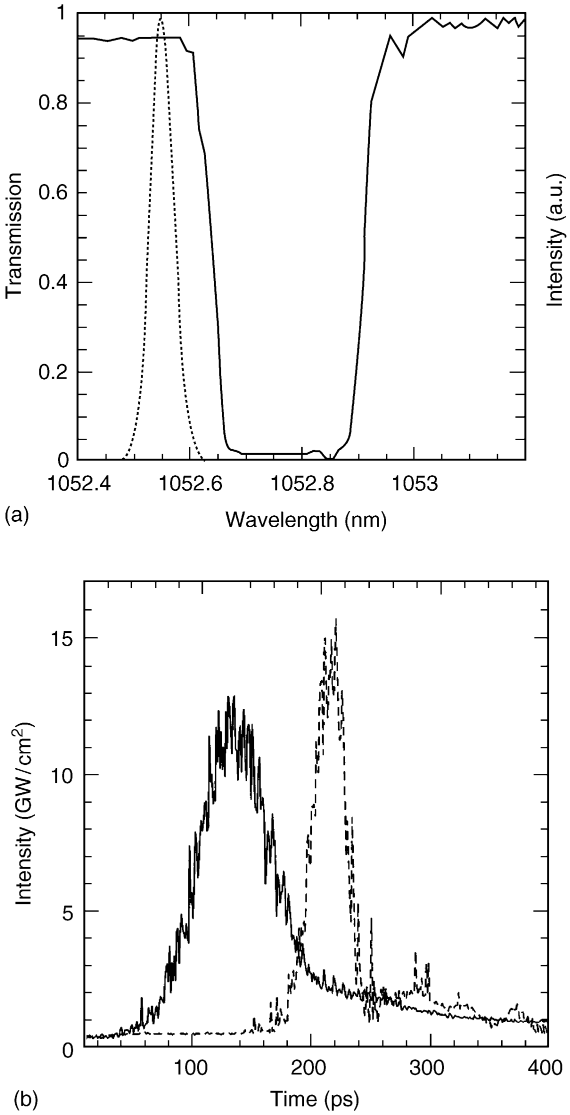

Figure 7 Gap solitons in a fiber Bragg grating: (a) measured pulse spectrum compared with the linear grating transmission, (b) compressed (15 ps) and delayed pulse transmitted at high intensity (dashed line), compared with the pulse transmitted at frequency far from the Bragg resonance (solid line) where it retains its input width of 80 ps. With permission from Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physical Review Letters* 76: 1627–1630. ©1996 APS.

(ii) GSs exist only in a narrow bandwidth, i.e., the shaded area in Figure 6c, which corresponds to the gap seen in the soliton frame (it reduces to the conventional stop-band for still solitons and gets wider for faster GSs); and

(iii) the gap is strongly asymmetric since lower intensity is required to induce transparency close to the high-frequency band-edge where, as a consequence, GSs have low amplitudes and are stable (the amplitude grows moving towards the low-frequency bandedge, where GSs become unstable).

Experimentally, GSs have been typically observed with pulses at frequency close to the blue bandedge of fibers (see Figure 7a) and AlGaAs waveguides with built-in Bragg gratings of a few centimeters. The typical signature of GSs, illustrated in Figure 7b, is the enhanced pulse transmission observed under the operating conditions of Figure 7a when the input intensity is sufficiently high. The transmitted pulse experiences pulse compression and a significant group-delay (measured with respect to the transit-time at frequency far off the stop-band), witnessing its nature of slow-light solitons.

Finally, we emphasize that grating solitons can also be supported in the case where the two modal envelopes propagate in the forward direction along a grating whose period Λ is long enough to be resonant with the modal phase slippage, thus inducing efficient linear coupling between them. In this case, the structure has a gap in wave-number, and solitons are slow in the frame traveling at the average modal group-velocity. Soliton formation is expected to lead to grating-induced compensation of the group-delay between the modes. Experiments on such types of structures have been carried out only recently.

Solitons in Dissipative Nonlinear Optics

Temporal solitons can also be observed in nonlinear optical interactions that involve changes of the state (energy level) of matter as in inverted media or scattering phenomena. For instance, early studies in the 1960s have demonstrated that coherent pulses with energy above a critical level can pass through resonant (two-level) atoms without being absorbed, a phenomenon denoted as self-induced transparency. Such pulses are temporal solitons solutions of the Maxwell–Bloch model which describes the atom–light interaction.

Also scattering phenomena occurring in optical fibers, such as Raman and Brillouin scattering, possess three-wave solitons, in which the pump and scattered (Stokes) light waves are coupled to a phonon field of the acoustic or optical branch (for Brillouin or Raman, respectively).

Perhaps more important from the point of view of the modern applications, are fiber lasers composed by spans of active and/or passive fiber closed to form a recirculating loop. These types of structures can often be described by distributed (averaged over several round-trips) models, which take the form of a dissipative perturbed NLSE, known as the Ginzburg–Landau equation (GLE):

$$\mathrm{i}\frac{\partial E}{\partial z} - \frac{k''}{2}\frac{\partial^2 E}{\partial t^2} + \chi|E|^2 E = \mathrm{i}gE + \mathrm{i}g_{\mathrm{nl}}|E|^2 E + \mathrm{i}q|E|^4 E + \frac{\mathrm{i}}{\Omega_{\mathrm{f}}^2}\frac{\partial^2 E}{\partial t^2} \qquad [6]$$

where g is the excess (net) gain ($g > 0$) or loss ($g < 0$), g_{nl} and q describe the nonlinear gain mechanism and its saturation (usually introduced to control the mode-locking), and Ω_{f} is linked to spectral filtering. Several soliton (pulse) solutions of the GLE [6] can be found even when the parameters (g, g_{nl}, q, Ω_{f}) are such as to constitute a strong perturbation of the NLSE. Unlike NLSE solitons [2], pulse solutions of the GLE [6] are usually chirped (i.e., possess a nonlinear phase profile), and have fixed amplitude for a given value of the parameters. The set of soliton waveforms is generally richer and includes, for instance, flat-top stationary pulses, two-pulse bound-states, solitons moving at fixed velocity, etc. Although instabilities of the background (for $g > 0$) or of the pulse itself can affect their observability, large islands of stability in the parameter space can be found. The relevant role of stable solitons in mode-locking and intra-cavity pulse formation mechanisms has been demonstrated in many different fiber laser geometries, for which we refer the reader to the more specialized papers listed as Further Reading at the end of this article. We point out that some schemes, for instance the stretched pulse-fiber laser developed at MIT, involve a basic configuration (alternating GVD within the cavity length) which is quite similar to that seen in the framework of DMSs, thus exhibiting the same basic physical mechanism (strong compression and stretching of the pulse along each round-trip which play the role of the map period for a DMS).

List of Units and Nomenclature

BBO	β-barium borate
DCF	dispersion compensating fiber
DMS	disperion managed soliton
EDFA	Erbium doped fiber amplifier
GLE	Ginzburg–Landau equation
GS	gap soliton
GVD	group-velocity dispersion
NLSE	nonlinear Schrödinger equation
OTDM	optical time-division multiplexing
SHG	second-harmonic generation

SPM	self-phase modulation
WDM	wavelength division multiplexing

Amplifier spacing L_A [m]
Bragg frequency ω_B [Hz]
Bragg period Λ [m]
Central wavelength λ_0 [nm]
Coupling coefficient Γ [m^{-1}]
Dispersion coefficient D [ps nm^{-1} km^{-1}]
Dispersion length L_d [m]
Effective area A_{eff} [m^2]
First-order dispersion k' [s/m]
Fundamental soliton power P_1 [W]
Gain (linear) g [m^{-1}]
Gain (nonlinear) g_{nl} [m^{-1}]
Group-delay δ [s/m]
Group-velocity $V = k'^{-1}$ [m/s]
Group velocity dispersion k'' [ps^2 m^{-1}]
Inverse pulse-width η [s^{-1}]
Light velocity in vacuum c [m/s]
Loss coefficient α [m^{-1}]
Nonlinear coefficient (fiber) γ [m^{-1} W^{-1}]
Nonlinear coefficient (quadratic bulk materials) χ [W$^{-1/2}$]
Nonlinear refractive index n_{2I} [m^2 W^{-1}]
Nonlinear length L_{nl} [m]
Refractive index (linear) n_0
Retarded time (in the moving frame) t [s]
Spectral filtering coefficient Ω_f [Hz m$^{1/2}$]
Time in the laboratory frame T [s]
Wave-number mismatch Δk [m^{-1}]

See also

Fiber and Guided Wave Optics: Nonlinear Effects (Basics); Nonlinear Optics. **Fiber Gratings**. **Lasers:** Optical Fiber Lasers. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Optical Communication Systems:** Wavelength Division Multiplexing. **Scattering:** Scattering Phenomena in Optical Fibers. **Solitons:** Bright Spatial Solitons; Optical Fiber Solitons, Physical Origin and Properties.

Further Reading

Ablowitz MJ, Biondini G and Ostrovsky LA (eds) (2000) Focus issue: optical solitons: perspectives and applications. *Chaos* 10: 471–640.

Agrawal GP (1995) *Nonlinear Fiber Optics*, 2nd edn. London: Academic Press.

Agrawal GP and Kivshar YS (2003) *Optical Solitons*. London: Academic Press.

Akhmediev NN and Ankiewicz A (1997) *Solitons: Nonlinear Pulses and Beams*. London: Chapman and Hall.

Buryak AV, Di Trapani P, Skryabin D and Trillo S (2002) Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. *Physics Reports* 370: 63–235.

De Sterke CM, Eggleton BJ and Sipe JE (2001) In: Trillo S and Torruellas W (eds) *Spatial Solitons*. Berlin: Springer-Verlag.

Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R and Piskarskas A (1998) Observation of temporal solitons in second-harmonic generation with tilted pulses. *Physics Review Letters* 81: 570–573.

Eggleton BJ, Slusher RE, de Sterke CM, Krug PA and Sipe JE (1996) Bragg grating solitons. *Physics Review Letters* 76: 1627–1630.

Forysiak W, Knox FM and Doran NJ (1994) Average soliton propagation in periodically amplified systems with stepwise dispersion-profiled fiber. *Optics Letters* 19: 174–176.

Hasegawa A (2000) Soliton-based optical communications: An overview. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1161–1172.

Hasegawa A and Kodama Y (1995) *Solitons in Optical Communications*. London: Oxford University Press.

Hasegawa A and Matsumoto M (2002) *Optical Solitons in Fibers*. Berlin: Springer-Verlag.

Haus HA (2000) Mode-locking of lasers. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1173–1185.

Mollenauer LF and Mamyshev PV (1998) Massive wavelength-division multiplexing with solitons. *IEEE Journal of Quantum Electronics* 34: 2102.

Nakazawa N, Kubota H, Sahara A, Suzuki K and Yamada E (2000) Ultrahigh-speed long-distance TDM and WDM soliton transmission technologies. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 363.

Porsezian K and Kuriakose VC (2003) *Optical Solitons: Theoretical and Experimental Challenges*. Berlin: Springer-Verlag.

Slusher RE and Eggleton BJ (eds) (2003) *Nonlinear Photonics Crystals*. Berlin: Springer-Verlag.

Suzuki M and Edagawa N (2003) Dispersion-managed high-capacity ultra-long-haul transmission. *Journal of Lightwave Technology* 21: 916–929.

Suzuki M, Morita I, Edagawa N, Yamamoto S, Toga H and Akiba S (1995) Reduction of Gordon-Haus jitter by periodic dispersion compensation in soliton transmission. *Electronic Letters* 31: 2027–2028.

Turitsyn SK, Aceves AB, Jones CKRT, Zharnitsky V and Mezentsev VK (1999) Hamiltonian averaging in soliton-bearing systems with a periodically varying dispersion. *Physics Review E* 59: 3843–3846.

Wabnitz S, Kodama Y and Aceves AB (1995) Control of optical soliton interactions. *Optics and Fiber Technology* 1: 187–217.

Zakharov VE and Wabnitz S (1999) *Optical Solitons: Theoretical Challenges and Industrial Perspectives*. Berlin: Springer-Verlag.

SPECTROSCOPY

Contents

Absolute Optical Frequency Metrology
Fourier Transform Sepctroscopy
Hadamard Spectroscopy and Imaging
Nonlinear Laser Spectroscopy
Raman Spectroscopy
Second-Harmonic Spectroscopy
Single Molecule Spectroscopy

Absolute Optical Frequency Metrology

S T Cundiff, National Institute of Standards and Technology and University of Colorado, Boulder, CO, USA

L Hollberg, National Institute of Standards and Technology, Boulder, CO, USA

Published by Elsevier Ltd.

Introduction

Visible light occupies a portion of the electromagnetic spectrum between frequencies of approximately 400 terahertz (THz) to 750 THz, where 1 THz is 10^{12} oscillations per second. Optical radiation, which includes adjacent infrared and ultraviolet portions of the spectrum, covers a frequency range of approximately 15–1500 THz. The techniques required for measuring frequencies in this range are qualitatively different from those currently employed in the radio frequency (RF) and microwave portions of the spectrum, where high-speed electronic counters make a direct measurement possible.

The frequency of an oscillatory phenomenon is the number of oscillations that occurs in unit time, or alternatively, the inverse of the period of one oscillation. Thus, to understand frequency, we must understand time, which is intrinsically related to phase. Only a time interval, the time elapsed between two events, has fundamental physical meaning. The ‘time’ that we are most familiar with, the time of day (including the date), is based on an arbitrarily chosen starting point. The starting point chosen in current international agreements can be traced to noon on December 31st, 1899. Furthermore, the time defined by those agreements, Universal Coordinated Time (UTC), is periodically adjusted relative to atomic time (discussed below) to compensate for irregularities in the rotation of the Earth.

Knowledge of the transition frequencies (energies) in simple atoms gives detailed information on the structure of the atoms. This can be used to test our understanding of the fundamental interactions in nature by comparing the measured frequencies with predictions of fundamental physical theories, in this case quantum electrodynamics (QED). Beautiful results along these lines are exemplified by precision measurements of transition frequencies in hydrogen and helium. In addition, the ability to measure and precisely control laser frequencies promises tremendous advances in the performance of the next generation of atomic clocks and frequency standards. These have potential applications to navigation and communication systems.

The measurement of frequency and time intervals also involves a choice, that of the unit of time, i.e., the definition of the second. The original definition of the second was based on the rotation of the Earth. However, not only is it irregular, but it is slowing down. Both of these facts limit its utility as a frequency standard. Once atomic clocks were highly developed, the second was redefined in 1967 to be the time it takes for the $F = 4$, $m_F = 0 \rightarrow F = 3$, $m_F = 0$ transition in the hyperfine structure of the ground state of ^{133}Cs to undergo 9192 631 770 oscillations. Thus an ‘absolute’ measurement of time interval, and hence frequency, must be directly connected to this defining frequency. The basic structure of an atomic clock is shown in Figure 1.

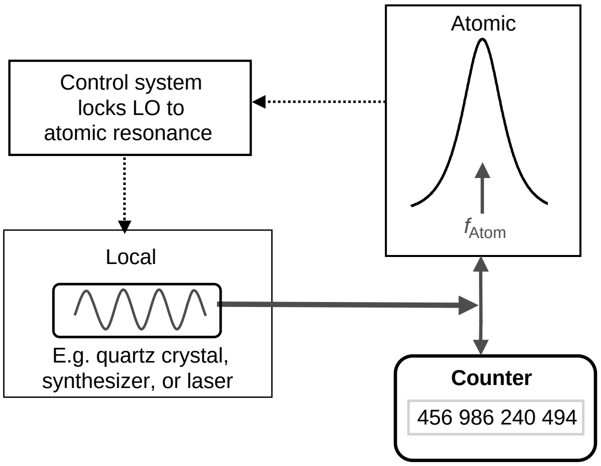
Atomic
Control system locks LO to atomic resonance
f_{Atom}
Local
E.g. quartz crystal, synthesizer, or laser
Counter
456 986 240 494

Figure 1 An atomic clock has three basic components: an atomic resonance, an oscillator to probe and lock to the atomic resonance, and a counter that records the number of oscillations and thus displays the time interval since some chosen starting time.

Frequency standards (or clocks) are characterized by three properties: accuracy, stability, and reproducibility. Accuracy describes how well the natural and fundamental atomic frequency is produced by the atomic frequency standard. Stability describes the frequency fluctuations on short timescales (an accurate but unstable standard produces a frequency that fluctuates around a constant mean that is the atomic frequency). Reproducibility between standards addresses whether or not two implementations of the standard produce exactly the same frequency. Atomic clocks are intrinsically accurate because the frequency of the transitions they are based on is determined by identical atoms in nature. The largest inaccuracies are usually due to environmental perturbations such as tiny variations in the local magnetic field.

A careful analysis of all of these issues, together with technical considerations, yielded the choice of cesium for the current definition of the second. Other common frequency standards include quartz crystal oscillators, hydrogen masers, and rubidium vapor cell standards. Quartz oscillators are inexpensive and have good stability, but in comparison to atomic standards they suffer from accuracy and reproducibility issues. Hydrogen masers have the best stability, but also poor accuracy and reproducibility compared to cesium. Vapor cell rubidium standards can be small and inexpensive but have poorer accuracy than cesium beam standards due to cell and buffer gas effects.

There are several motivations for using optical sources rather than RF or microwaves to measure time intervals and frequencies. Obviously, a higher oscillation frequency can divide time into smaller units and this can provide higher measurement precision. It is thus natural to use the highest oscillation frequency that can be precisely counted to measure time and frequency. Not long after lasers were invented, suggestions were made that it would be possible to make optical frequency standards using lasers locked to narrow atomic resonances.

A basic feature of optical transitions is that they have a very high *Q* factor (the ratio of the center frequency to the linewidth). This allows them to be measured to very high precision (current measurements are good to a few parts in 10^{15}). It may be possible to use this high precision to build improved atomic clocks that use optical transitions instead of the microwave transition currently being used. Candidates for optical standards include both trapped ions (Hg^+, Yb^+, Sr^+, and In^+ are under investigation) and laser-cooled neutral atoms (Ca, Sr, Mg, Ba, and Ag). The high precision is also utilized for fundamental tests of quantum electrodynamics by comparing transition frequencies (usually of hydrogen or helium atoms) to first-principle calculations. For comparison Figure 2 shows the relevant transitions and energy levels in the Cs microwave standard and the Hg^+ optical frequency. It is even possible to use the precision of atomic clocks for high-sensitivity tests of variation in the fundamental constants with time. Finally, because the speed of light is constant length measurements based on the wavelength of light with a known frequency, the precision of length metrology ultimately rests on optical frequency metrology.

An obvious question, which might occur to many in the field of optics, is why a standard spectrometer does not provide absolute frequency measurements. All spectrometers measure wavelength, not frequency, based on interference between light traveling two different path lengths. Thus, knowledge of the path length is needed to determine the wavelength. However, since length is now defined using the speed of light, the definitions become circular unless a source of light with known absolute frequency is available. In addition, frequency measurement is intrinsically more precise and reliable than mechanical length measurement.

Recently, there has been a significant breakthrough in optical-frequency metrology and optical clocks by using mode-locked lasers that generate optical pulses with durations of a few femtoseconds. Before discussing optical frequency metrology based on mode-locked lasers, we will review earlier techniques to provide the necessary background. After describing

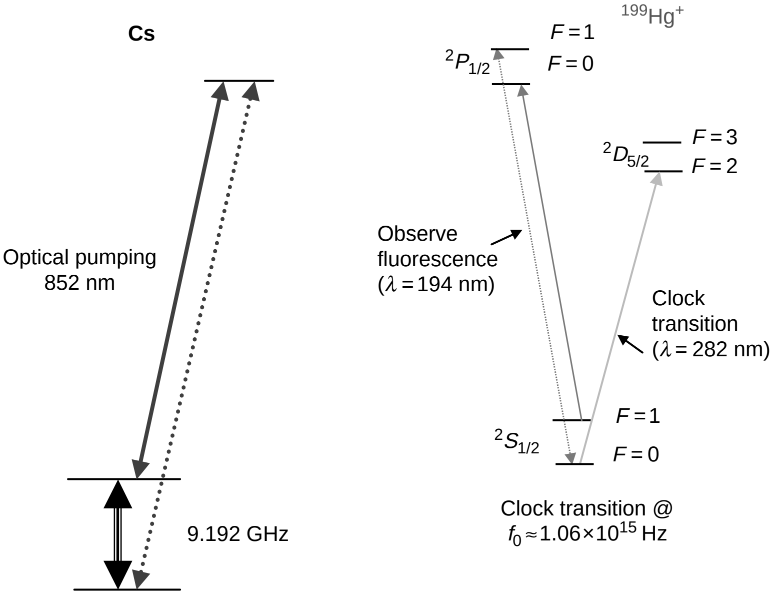

Figure 2 Simplified energy level diagram of Hg^+ and Cs.

methods that use mode-locked lasers, we will then provide a summary of current measurements and standards. Finally we will briefly describe the outlook for the future of absolute optical frequency metrology.

Background

To imagine being able to measure the frequency of light we need to have laser sources with high coherence, like that of radio waves, where the phase is stable enough to be measured. Even though the first beat-note between two lasers was demonstrated in the 1960s, most lasers were not coherent (stable) enough that the optical phase could be tracked for much longer than about 10 to 100 ns. After many years of research and inspiring technological achievements, Hall, Chebotayev, Bergquist and others were able to stabilize the frequency of tunable laser sources to the point that lasers could indeed be considered coherent sources, with the phase continuously measurable and even predictable for times as long as seconds. Once the lasers were stable enough to be counted, there was still the essential problem that no counter was fast enough to actually count optical frequencies.

With spectrally pure laser sources it was possible to conceive of measuring optical frequencies by extending the nonlinear methods that are used to generate and measure RF and microwave frequencies. By using simple nonlinear components such as diodes, it is easy to generate high harmonics of an input frequency in the RF and microwave regions of the spectrum. At least in principle, and by direct analogy, it seemed that it should be possible to multiply coherent sine waves from radio frequencies up through the microwave region to reach the THz frequencies of far-infrared lasers, and then to use nonlinear optical methods to multiply on up to the infrared (IR) and eventually to the visible region of the spectrum. The concept is simple enough, $f \rightarrow 100\times f \rightarrow 1000\times f \rightarrow 2000\times f \rightarrow 4000\times f\ldots$ until you reach an optical frequency. Multiplication up to the microwave region works quite well, but unfortunately the nonlinear optical processes in the far-infrared, IR, and optical regions are typically quite inefficient; the power at the second harmonic is typically only 10^{-5} times the square of the fundamental power in watts. Just multiplying a single-frequency laser by a factor of two can be a technological challenge. After reaching a microwave frequency of about 50 GHz we still require a multiplication factor of 10^4 to reach the visible at 500 THz. Since we are usually forced to multiply in the IR and optical by factors of 2, we have $2^N = 10^4$, which means we require $N \approx 13$ multiplication stages. This is approximately the number of stabilized laser oscillators that are required to span the frequency gap from the microwave to the visible. Notwithstanding these technological limitations, Evenson and collaborators did a groundbreaking demonstration of a harmonic frequency chain in 1972 that connected a microwave atomic frequency standard and a stabilized HeNe laser in the IR at 88 THz. With a concerted multiperson effort, more lasers, and 10 years of research, Jennings and co-workers finally succeeded in extending the harmonic optical frequency chain to the visible. That original optical frequency chain from the RF to the visible is diagramed in Figure 3. In the succeeding 18 years, only three other harmonic frequency chains were built world-wide that connected the microwave standards to optical frequencies.

An alternative scheme for measuring optical frequencies is the optical bisection method, a conceptually simple technique proposed by Telle, Meschede and Hänsch, for dividing down from

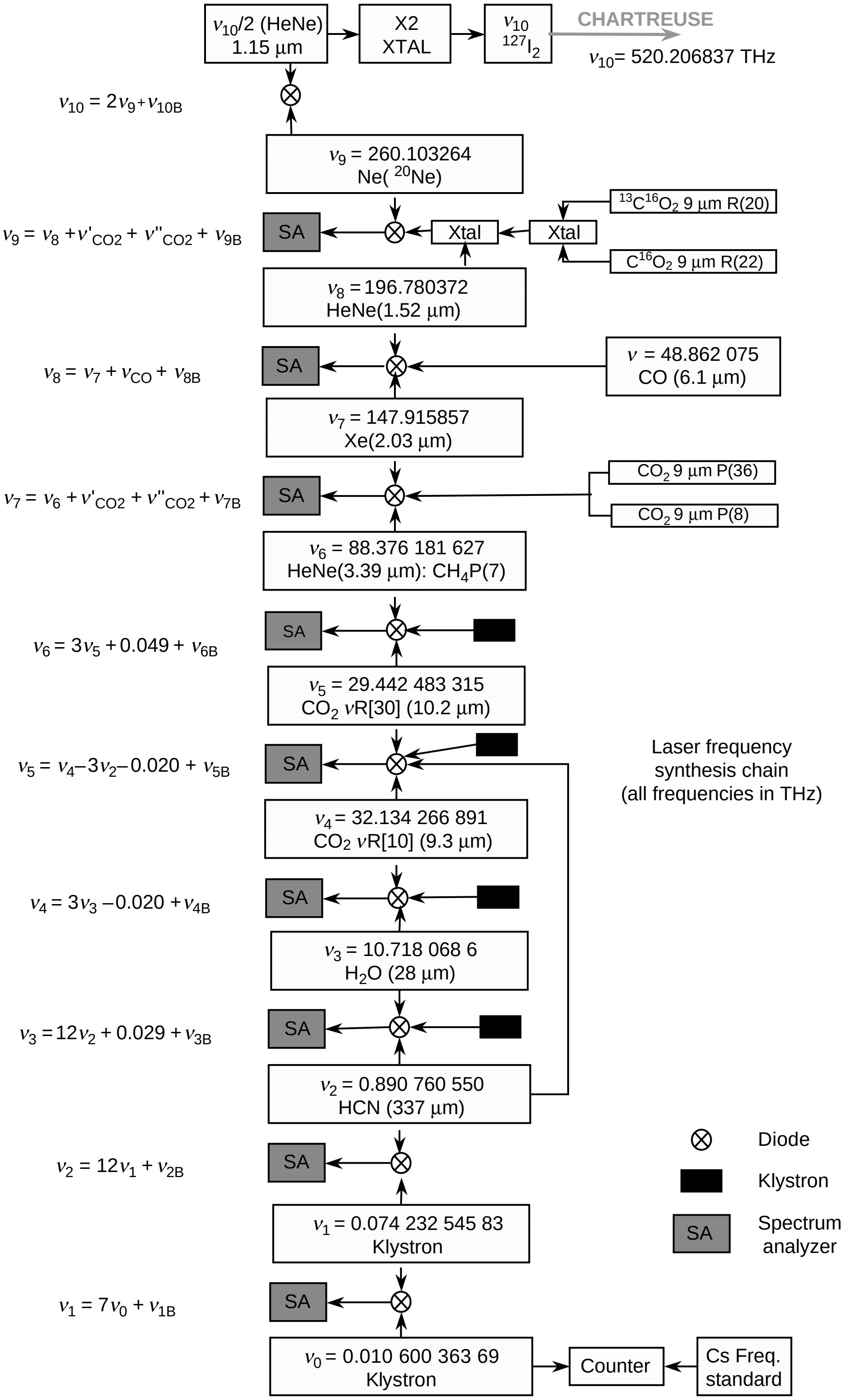
ν10/2 (HeNe) 1.15 μm
X2 XTAL
ν10 127I2
CHARTREUSE
ν10= 520.206837 THz
ν10 = 2ν9+ν10B
ν9 = 260.103264 Ne(20Ne)
13C16O2 9 μm R(20)
Xtal
Xtal
C16O2 9 μm R(22)
ν9 = ν8 +ν'CO2 + ν''CO2 + ν9B
SA
ν8 = 196.780372 HeNe(1.52 μm)
ν = 48.862 075 CO (6.1 μm)
ν8 = ν7 + νCO + ν8B
SA
ν7 = 147.915857 Xe(2.03 μm)
CO2 9 μm P(36)
CO2 9 μm P(8)
ν7 = ν6 +ν'CO2 + ν''CO2 + ν7B
SA
ν6 = 88.376 181 627 HeNe(3.39 μm): CH4P(7)
ν6 = 3ν5 + 0.049 + ν6B
SA
ν5 = 29.442 483 315 CO2 νR[30] (10.2 μm)
ν5 = ν4–3ν2–0.020 + ν5B
SA
Laser frequency synthesis chain (all frequencies in THz)
ν4 = 32.134 266 891 CO2 νR[10] (9.3 μm)
ν4 = 3ν3 –0.020 +ν4B
SA
ν3 = 10.718 068 6 H2O (28 μm)
ν3 = 12ν2 + 0.029 + ν3B
SA
ν2 = 0.890 760 550 HCN (337 μm)
ν2 = 12ν1 + ν2B
SA
Diode
Klystron
SA Spectrum analyzer
ν1 = 0.074 232 545 83 Klystron
ν1 = 7ν0 + ν1B
SA
ν0 = 0.010 600 363 69 Klystron
Counter
Cs Freq. standard

Figure 3 Schematic of the first frequency chain used to measure optical frequencies relative to the Cs standard.

optical frequencies to microwaves. In fact, the bisection method divides an arbitrary optical frequency interval in half by utilizing second-harmonic generation and sum-frequency-mixing, and forcing the condition that $2f_3 = f_1 + f_2$. Thus, the laser at frequency $f_3 = (f_1 + f_2)/2$ bisects the frequency interval between the lasers at f_1 and f_2. The optical bisection method has the following two significant advantages over the multiplication method: the phase noise is decreased in successive divisions (rather than increased by multiplications), and the system could be constructed out of one convenient and reproducible laser technology, say diode lasers in the near-visible. In principle, any optical frequency could be measured by successive operations of the optical bisector. Optical frequency intervals as large as 8 THz have been measured with this method, but even though it is feasible, no bisection system has been built to date that connects the optical region all the way to countable microwave frequencies. The main limitation is the same as that facing the multiplication scheme: it still requires lots of stabilized laser sources to divide by factors of 2 from 10^{15} to 5×10^{10} Hz. The problem can be reduced somewhat by taking advantage of an optical-frequency comb generator based on an electro-optic modulator inside a resonant optical cavity. These systems, developed by Korougi and others, generate many coherent microwave sidebands on a laser so that optical frequency intervals on the order of 5 THz can be measured.

Even combining all of these ideas and techniques, a convenient system to count optical frequencies remained elusive. This has now changed with the revolutionary new optical frequency combs based on mode-locked femtosecond lasers.

Optical Frequency Metrology with Femtosecond Combs

Recently, the introduction of mode-locked lasers into the field of absolute optical frequency metrology has resulted in an important advance. The extremely complex frequency chains described above can be replaced by a single mode-locked laser if it produces pulses with sufficiently large bandwidth. Although the underlying ideas are not new (they were originally discussed by Hänsch and Chebotayaev in the 1970s) it is only recently that mode-locked lasers with the required characteristics were developed and the concept carefully tested.

The dramatic simplification of a frequency chain that uses a mode-locked laser means that absolute optical-frequency measurements can now be made by a single person as compared to the team of approximately 10 highly trained scientists required to run the previous chains. This has resulted in a large number of measurements being reported in the last 2 years. The simplicity also allows longer averaging times and a greater number of measurements to be used in determining the final reported frequency. In addition, since a simpler system has fewer sources of error, it has improved the quality of measurements made in a given amount of time.

Mode-locked lasers produce ultrashort pulses of light. For a typical modern high-quality mode-locked laser, the pulse duration is around 10 fs (1 fs = 10^{-15} second), corresponding to three or four optical cycles in the near-infrared, where these lasers typically operate. The current best mode-locked lasers can produce pulses that are shorter than two optical cycles in duration (5–6 fs). Since the spectrum of these lasers is complicated, the bandwidth is significantly larger than it would be for a smooth pulse such as one with a Gaussian or hyperbolic-secant temporal intensity profile.

At first sight, using mode-locked lasers for optical-frequency metrology seems counter-intuitive, because optical-frequency metrology clearly requires very well-defined frequencies in contrast to the broad frequency spectrum of ultrashort pulses. This apparent conundrum is resolved by the fact that a mode-locked laser actually produces a very regularly spaced train of ultrashort pulses. The ultrashort pulses emitted by a mode-locked laser are replicas of a steady-state pulse that circulates inside the cavity of the laser. Every time the internal pulse impinges on the output coupler, which is a partially reflective mirror, a portion of it is transmitted, producing the output. The timing between these output pulses is determined purely by the time, τ_c, for the intracavity pulse to make one round trip, which is typically between 1 and 20 ns. The spectrum of a train of pulses is a regularly spaced 'comb' of frequencies spaced by the repetition rate of the train, $f_{rep} = 1/\tau_c$. If all of the pulses are identical, then these comb lines are just integer multiples of f_{rep}. However, due to dispersion, the group and phase velocities are different inside the cavity of a mode-locked laser. This leads to a pulse-to-pulse shift in the phase between the carrier and envelope (which we call the carrier–envelope phase; ϕ_{ce}, the pulse-to-pulse shift of ϕ_{ce} is designated by $\Delta\phi_{ce}$). The presence of $\Delta\phi_{ce}$ results in a rigid shift of all the comb lines by an amount $\delta = 2\pi\Delta\phi_{ce}/f_{rep}$. Thus the frequencies of the comb lines are given by $\nu_n = nf_{rep} + \delta$, where n is a large integer of order 10^6. The essential point of this equation is that it gives optical frequencies in terms of RF frequencies (f_{rep} and δ) that can easily be measured with conventional electronics and

(a) Time domain

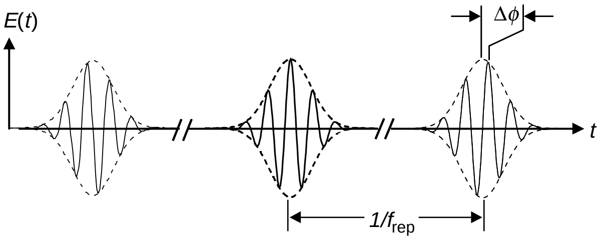

(b) Frequency domain

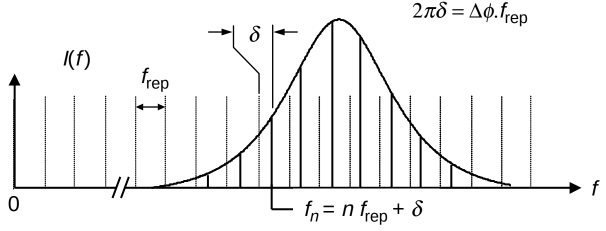

Figure 4 Time–frequency correspondence and relationship between $\Delta\phi$ and δ. (a) In the time domain, the relative phase between the carrier and the envelope evolves from pulse to pulse by the amount $\Delta\phi$. (b) In the frequency domain, the elements of the frequency comb of a mode-locked pulse train are spaced by f_{rep}. The entire comb is offset from integer multiples of f_{rep} by an offset frequency $\delta = \Delta\phi f_{\mathrm{rep}}/2\pi$.

compared to the microwave cesium frequency standard. The relationship between the time and frequency domains is shown in Figure 4.

Given such a comb of frequencies, a heterodyne measurement readily yields the frequency difference between the unknown optical frequency of a narrow-band source (typically a single frequency laser) and the optical frequency of a nearby comb line. Thus, the absolute unknown frequency can be determined if the absolute frequencies of the comb lines are known. This requires measurement of f_{rep} and δ. The repetition rate, f_{rep}, is easily measured with a fast photodiode. Measurement of δ requires somewhat greater effort.

The development of a 'self-referencing' method to easily measure δ is the key enabling breakthrough that has brought about the recent revolution in optical-frequency metrology. It is called self-referencing because it provides a direct measure of δ with no other optical frequencies as input. It works by comparing frequencies that differ by a factor of two (typically these lie in the wings of the spectrum) using second-harmonic generation. Specifically, the frequency difference between the second harmonic of a comb line n and the comb line $2n$ is given by $2f_n - f_{2n} = 2(nf_{\mathrm{rep}} - \delta) - (2nf_{\mathrm{rep}} - \delta) = \delta$. Such a frequency difference is easily measured using a heterodyne beat. This technique is shown schematically in Figure 5. Other variations on this basic idea are possible; for example, rather than frequency doubling the comb itself, it is also possible to extract the same information by comparing the comb lines to a single-frequency laser and its second harmonic.

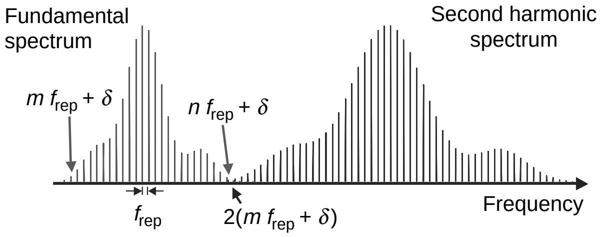

Figure 5 Schematic of the self-referencing technique for determining the offset frequency of the frequency comb produced by a femtosecond mode-locked laser. The frequency difference between the fundamental comb and its second harmonic is the offset frequency. In the region where overlap occurs, this is easily detected as a heterodyne beat.

Implementation of the self-referencing technique requires a pulse spectrum that spans a factor of two in frequency (an optical octave). Although a laser has recently been demonstrated that produces such a broad spectrum, they are not commonly available. However, the development of microstructured fiber has made it easy to broaden the spectrum of ordinary mode-locked lasers so that it spans an octave, as was first demonstrated by Ranka and co-workers at Bell Labs. The broadening occurs because optical nonlinearity in the fiber results in self-phase modulation of the laser pulse. Microstructure fiber achieves guiding by surrounding the core region with microscopic air holes, as compared to regular fiber that uses doping to produce a difference in the index of refraction. The much larger difference in the index of refraction in microstructure fiber allows a smaller core region to be used, which increases the effective nonlinearity. It also makes it possible to modify the dispersion such that the group velocity dispersion goes through zero for light with a wavelength close to 800 nm. This is crucial because group-velocity dispersion in ordinary fiber causes ultrashort pulses to temporally spread, which lowers the peak power and hence nonlinearity. It is possible to implement self-referencing techniques using less bandwidth if higher orders of nonlinearity are used; for example, if the difference between the second harmonic of the high-frequency end and third harmonic of the low-frequency end of the comb are used, only a half-octave bandwidth is required.

Given the ability to measure both f_{rep} and δ, and as long as its frequency lies within the comb spectrum, it is possible to measure the optical frequency of any optical source by measuring the heterodyne beat with a nearby comb line. Thus, an absolute optical frequency measurement comes down to measuring three RF frequencies. It is often advantageous to actively control either one or both of f_{rep} and δ to hold them to a fixed frequency. This requires real-time adjustment of some parameters of the mode-locked laser to affect the necessary feedback. Controlling f_{rep} is straightforward; it is achieved by simply changing the cavity length of the laser, typically by mounting a mirror on a piezoelectric transducer. Controlling δ is more challenging, as it requires a differential change between the cavity group delay and phase delay. This has been achieved by adjusting the pump power, which in turn controls the energy of the intracavity pulse, and/or by tilting the end mirror of the cavity in lasers that employ prism dispersion compensators. Since the optical spectrum is spread out across this mirror, tilting it can be thought of as a frequency-dependent change in the cavity length, with the optical frequency corresponding to the pivot point of the tilt experiencing no length change, while those on opposite sides of the pivot point see changes with opposite signs.

In addition to the three RF frequencies, it is necessary to know the large integer n that describes the comb line that is beat against the unknown source. If a priori knowledge about the unknown source is available with precision somewhat better than $f_{\text{rep}}/2$ or better, n can readily be determined. For typical mode-locked lasers, standard commercial wavemeters can provide this level of precision. In the absence of additional outside information, n can also be determined with some effort and precision measurements by changing f_{rep} and δ in a systematic fashion. In the end, the simple formula $f_n = nf_{\text{rep}} + \delta$ gives the optical frequency of any arbitrary nth mode of the femtosecond optical comb. With this established, the frequency of any single-mode laser within the octave span of the comb can be determined simply by measuring the frequency of the beat-note generated in a photodiode between the cw laser and the known frequencies of the comb.

The diagram in Figure 6 shows how a femtosecond comb acts as an optical frequency synthesizer, which provides frequencies that are known relative to the input frequency (provided by a hydrogen maser in the figure). A typical series of measurements is shown in Figure 7 for the NIST Hg^+ single ion standard. The inset shows the results of numerous measurements for the frequency of Hg^+ relative to the NIST primary cesium standard.

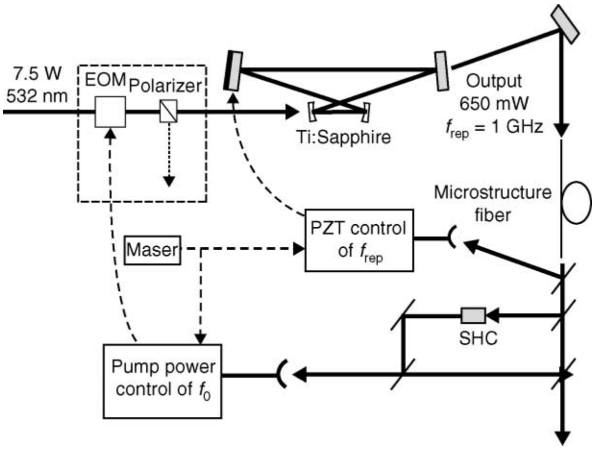

Figure 6 Schematic example of how the femtosecond comb is configured to make optical frequency measurements relative to the reference frequency provided by a hydrogen maser (Maser) that is in turn calibrated to a cesium atomic frequency standard. The repetition rate of the pulses from the Ti:Sapphire laser are detected in a photodiode and are controlled using a piezoelectric transducer (PZT) that changes the length of the laser cavity. The microstructure fiber and second harmonic crystal (SHC), provide the 'self-referencing' beat-note signal that is used to control the offset frequency, δ, by changing the power of the pump laser with an electro-optic modulator (EOM). This 'self-referenced' system then provides a set of modes with known frequencies spanning the visible. An unknown laser frequency can then be measured by making a beat-note with one of the modes of the comb.

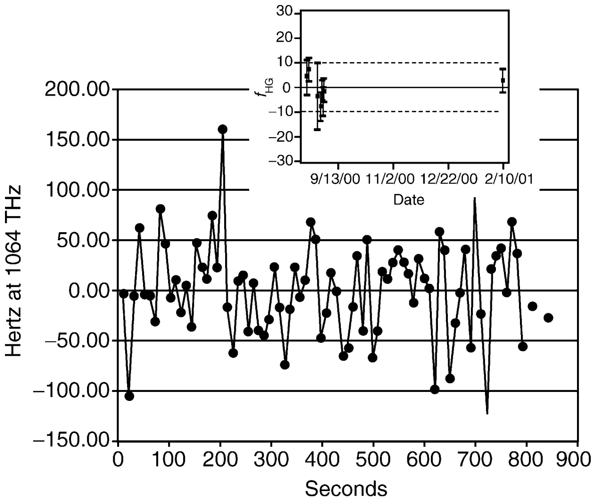

Figure 7 Actual frequency measurements of the beat-note between a hydrogen maser stabilized optical frequency comb (as shown in Figure 6) and the NIST Hg^+ optical frequency standard. The scatter in the data for counter gate time of 10 seconds corresponds to a fractional frequency uncertainty of about 4×10^{-14}, consistent within the instability in the maser. The inset shows the results from numerous datasets of measurement of the frequency of the Hg^+ optical standard relative to the cesium primary standard at NIST. The frequency data are plotted relative to 1064 721 609 899 142.6 Hz. The dotted lines give estimates of systematic uncertainties in lieu of a complete accuracy evaluation.

Table 1 Summary of optical frequency measurements of a selection of molecular, atomic, and ionic transitions. Recent measurements listed for all and additional measurements are given for a few to indicate the level of agreement

Atom/molecule/ion	*Wavelength (nm)*	*Frequency (THz)*	*Uncertainty*		*Year*	*First author and institution*
			± *Hz*	*Frac.*		
O_sO_4	10318.0	29.054 057 446 660	50	2×10^{-12}	1985	Clarion, LPTF
		29.054 057 446 579	10	3×10^{-13}	1999	Ducos, LPTF
CH_4	3392.2	88.376 181 627 000	50 000	6×10^{-10}	1972	Evenson, NBS
		88.373 149 028 553	200	2×10^{-11}	1998	Ering, PTB
Rb−2 photon	778.11	385.285 142 367 000	8000	2×10^{-11}	1993	Nez, ENS
		385.285 142 374 800	3000	8×10^{-12}	2000	Diddams, JILA
Sr^+	674.03	444.779 904 409 540	200	4×10^{-13}	1999	Bernard, NRC
Ca	657.45	455.986 240 495 150	8	10×10^{-14}	2003	Helmcke, PTB
		455.986 240 494 158	26	6×10^{-14}	2000	Udem, NIST
I_2 a_{15} of $R(127)$ 11-5	632.99	473.612 340 492 000	74 000	2×10^{-10}	1983	Jennings, NBS
I_2 a_{16} of $R(127)$ 11-5	632.99	473.612 353 604 800	1200	3×10^{-12}	2000	Ye, JILA
I_2 a_{10} of $R(56)$ 32-0	532.24	563.260 223 480 000	70 000	1×10^{-10}	1995	Jungner, JILA
		563.260 223 514 000	5000	9×10^{-12}	2000	Diddams, JILA
Yb^+	435.51	688.358 979 230 931	6	9×10^{-15}	2001	Stenger, PTB
H 2*S*-8*S*	389.01	770.649 561 581 100	5900	8×10^{-12}	1997	de Bouroir, ENS
Hg^+	281.57	1064.721 609 899 140	10	9×10^{-15}	2000	Udem, NIST
In^+	236.54	1267.402 452 899 920	230	2×10^{-13}	2000	von Zanthier, MPQ
H 1*S*-2*S*	121.57	2466.061 413 187 100	46	2×10^{-14}	2000	Holzwarth, MPQ

Summary of Measurements/Standards

Starting with the first measurement of the frequency of the methane-stabilized HeNe laser in 1972, a new laser frequency measurement was reported every year or so using harmonic frequency chains. These are summarized in Table 1. An indication of the history of frequency measurements of stable laser references is given in Figure 8. The accuracy, in terms of fractional frequency uncertainty, improved from 10^{-10} in 1972 to 10^{-12} in 1999, at which time the femtosecond optical frequency combs came on line and there was a dramatic improvement in the precision as well as the number of measurements that were completed. It is impossible to predict with certainty but some have projected that optical standards can ultimately reach uncertainties as small as 10^{-18}.

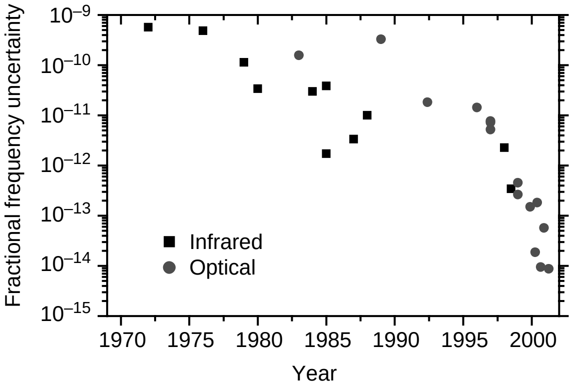

Figure 8 Progress in the accuracy of optical frequency measurements.

Outlook

Three separate technologies have now reached a level of maturity that it is possible to build high-performance optical-frequency standards and clocks. The essential achievements are: laser cooling and trapping of atoms (first proposed by Wineland and Dehmelt, and Hänsch and Schawlow), highly stabilized narrow-linewidth cw lasers, and femtosecond optical frequency combs. Combining these key ingredients, we can construct an optical atomic clock as shown schematically in Figure 9.

Here, for comparison with Figure 1, the cold atoms or single ion provide the narrow atomic resonance, the cw laser serves as the local oscillator to probe the resonance, and the femto-comb serves as the counter. Optical frequency standards of the future are expected to provide orders of magnitude better stability and improved accuracy over the existing atomic frequency standards that now use microwave transitions in atoms. As described in the previous section, frequency combs produced by femtosecond lasers can directly measure the frequency of a stable laser locked to an atomic transition relative to a known microwave frequency standard. This gives fundamental information about the atomic energy levels, and structure, and allows comparisons between different elements. However, to take

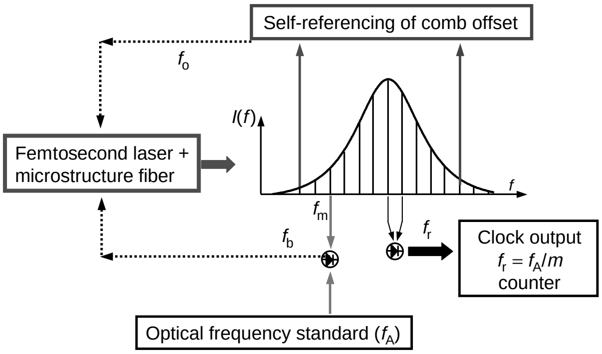
Self-referencing of comb offset
f_o
$I(f)$
Femtosecond laser + microstructure fiber
f
f_m
f_b
f_r
Clock output
$f_r = f_A/m$
counter
Optical frequency standard (f_A)

Figure 9 Simplified schematic of an optical clock.

advantage of the high stability of the optical references we run the system as an optical clock (Figure 9) where the stable laser and the femto-comb are locked to the atomic resonance, and the clock output comes as pulses at the repetition rate (e.g., 1 GHz) of the femtosecond mode-locked laser. With a judicious choice of control parameters it is possible to have the pulse repetition frequency, i.e., the clock output, at an exact subharmonic of the optical transition frequency. It is intriguing to note that a portable optical clock could measure time and length at the same time.

See also

Instrumentation: Spectrometers. **Quantum Electrodynamics:** Quantum Theory of the Electromagnetic Field.

Further Reading

Barnett JE (1998) Time's Pendulum. New York: Plenum.

Bergquist JC, Jefferts SR and Wineland DJ (2001) Time measurement at the millennium. *Physics Today* March: 37–42.

Cundiff ST, Ye J and Hall JL (2001) Optical frequency synthesis based on mode-locked lasers. *Reviews of Scientific Instruments* 72: 3749–3771.

Diddams SA, Jones DJ, Ye J, *et al.* (2000) Towards the ultimate control of light. *Optics and Photonics News* October: 16–22.

Diddams SA, Udem T, Bergquist JC, *et al.* (2001) An optical clock based on a single trapped $^{199}\mathrm{Hg}^+$ ion. *Science* 293: 825–828.

Diels JC and Rudolph W (1996) Ultrashort Laser Pulse Phenomena. San Diego, CA: Academic Press.

Hall JL (2000) Optical frequency measurement: 40 years of technology revolutions. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 1136–1144.

Hall JL and Ye J (2001) A new era of frequency standards and optical frequency measurements. *Optics and Photonics News* February: 45–50.

Hänsch TW and Walther H (1999) Laser spectroscopy and quantum optics. *Reviews of Modern Physics* 71: S242–S252.

Hollberg L, Diddams SA, Curtis EA, Oates CW and Fox RW (2001) Optical frequency standards for clocks of the future. In: The International Society for Optical Engineering. *Harnessing Light: Optical Science and Metrology at NIST.* Bellingham, DC. *Proceedings of SPIE* 4450: 54–59.

Jennings DA, Evenson KE and Knight DJE (1986) Optical frequency measurements. *Proceedings of the IEEE* 74: 168–179.

Jespersen J and Fitz-Randolph J (1977) From Sundials to Atomic Clocks. New York: Dover.

Jones T (2000) Splitting the Second: The Story of Atomic Time. Bristol, UK: Institute of Physics.

Luiten AN (2001) Frequency Measurement and Control: Advanced Techniques and Future Trends. Berlin: Springer-Verlag.

Menoni CS (ed.) (2001) Feature section on optical frequency synthesis: a new tool for precision optical frequency metrology. *IEEE Journal of Quantum Electronics* 37: 1481–1513.

Quinn TJ (1999) Practical realization of the definition of the metre. *Metrologia* 36: 211–244.

Fourier Transform Spectroscopy

T Fromherz, Johannes Kepler Universität, Linz, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fourier transform spectroscopy (FTS) has emerged as one of the most powerful spectroscopic techniques since the first instruments became commercially available in the late 1960s. With this method, high spectral resolution, high wavenumber accuracy, broad spectral range, high optical throughput, and high signal to noise ratio can be achieved simultaneously. Because of these combined advantages that will be discussed in more detail throughout this article, especially in the far- (10–500 cm^{-1}) and mid-infrared (FIR and MIR) spectral region (500–5000 cm^{-1}), FTS has become the method of choice for the most sensitive spectral investigations. In this article, an overview of the setup, the working principle, and the computation of spectra from the measured interferograms will be given. Finally, by comparing FTS to dispersive spectroscopic methods, the conditions under which FTS can be exploited to advantage, will be discussed.

Spectrometer Setup

Almost all commercially available Fourier transform (FT) spectrometers are based on a Michelson interferometer. A typical layout of this interferometer is sketched in Figure 1. A collimated beam with spectral density $S(\sigma)$ entering the spectrometer is divided into two beams by a beamsplitter. Here, σ denotes the wavenumber that is related to the wavelength (λ) of the radiation by $\sigma = 1/\lambda$. As shown in Figure 1, one part of the radiation is reflected by the beamsplitter and propagates through the compensator plate to a fixed mirror. The mirror reflects the beam through the compensator plate back onto the beamsplitter. The other part of the radiation beam is transmitted through the beamsplitter and its substrate, and hits a moveable mirror from which it is reflected back onto the beamsplitter. At the beamsplitter, both beams are divided again. As shown in Figure 1, two beams, one propagating to the source, the other propagating to the detector, finally result. Both beams contain a contribution that was reflected by the fixed mirror and a second contribution that was reflected by the moveable mirror. In Figure 1, lines indicate the paths of the beams schematically. In the case of a reflection on the beamsplitter, the line type is changed in Figure 1, whereas for a transmission, the line type remains unchanged. The widths of the lines in Figure 1 indicate the size of the field amplitudes and, therefore, are reduced for each reflection/transmission on the beamsplitter. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors in Figure 1 are offset laterally.

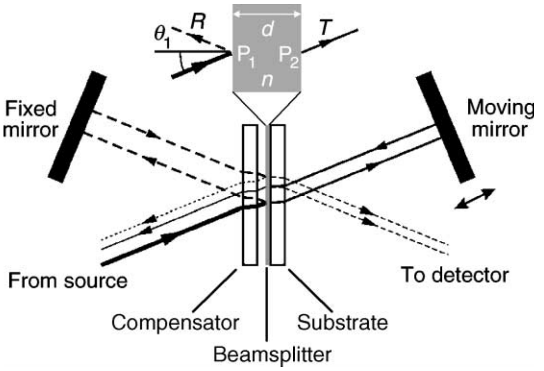

Figure 1 Schematic sketch of the layout of a Michelson interferometer. The paths of the rays are indicated by lines and arrows. At a reflection on the beamsplitter, the line type is changed, whereas for a transmission it remains unchanged. The widths of the lines indicate the size of the field amplitudes. For clarity, the lines indicating the beams reflected into themselves by the fixed and moveable mirrors, are offset laterally. The inset shows a magnified cross-section of the beamsplitter. In the inset, the geometrical details used for calculating the phase difference between *R* and *T* are defined.

By changing the position of the moveable mirror along the beam axis, a difference between the optical paths from the beamsplitter to the fixed and moveable mirror is generated. The compensator consists of the same material and has the same thickness as the beamsplitter substrate and, therefore, minimizes the influence of the dispersion of the beamsplitter substrate on the optical path difference.

Fourier Analysis and Interferometry

To calculate the intensities of the combined beams falling on the detector and on the source, we start by considering the phase difference between the reflected and transmitted beams leaving the beamsplitter. For simplicity, we ignore the compensator and beamsplitter substrate, since ideally, they generate the same phases in both arms of the interferometer, and therefore, do not contribute to the phase difference. Thus the beamsplitter is regarded as a freestanding, dielectric slab with thickness d. The complex reflection (R) and transmission (T) coefficients, including the effect of multiple internal reflections within the beamsplitter layer, are given by:

$$\begin{aligned} R &= r(1 - e^{i2\gamma})/(1 - r^2 e^{i2\gamma}) \\ T &= e^{i\gamma}(1 - r^2)/(1 - r^2 e^{i2\gamma}) \end{aligned} \qquad [1]$$

Here, r is the Fresnel reflection coefficient that describes the ratio of the moduli and the phase difference between incoming and reflected electrical field for a electromagnetic wave incident under a angle Θ_1 on the surface of a dielectric with dielectric constant n, and $\gamma = 2\pi\sigma nd\cos(\Theta_2)$ is the phase difference evaluated at P_1 and P_2 in Figure 1 for a single pass of the electromagnetic wave through the dielectric slab under the angle Θ_2 (by Snellius law, $\sin(\Theta_1) = n\sin(\Theta_2)$). Since n is real, so is r and, therefore, the phase difference $\Delta\varphi$ between R and T is calculated from eqn [1] according to:

$$\begin{aligned} \frac{R}{T} &= \frac{|R|}{|T|} e^{i\Delta\varphi_{RT}} = \frac{r}{1 - r^2}(e^{-i\gamma} - e^{i\gamma}) \\ &= -2i\frac{r}{1 - r^2}\sin(\gamma) \end{aligned} \qquad [2]$$

i.e., $\Delta\varphi_{RT} = -\pi/2$. With this result, we are able to calculate the phase difference between the two rays falling on the detector and the two rays falling back on the source. The geometry considered in the following is sketched in Figure 1. The coefficients R and T connect the phases and magnitudes of reflected and transmitted electric field amplitudes at the points P_1 and P_2. From these points, the radiation beams travel back and forth to the fixed and moveable mirrors, thereby accumulating phases according to: $\varphi_f = 2\pi 2x_f\sigma$ and $\varphi_m = 2\pi 2(x_f + \Delta x)\sigma$, where we have defined $\Delta x = x_m - x_f$ and x_m, x_f denote the distances of moveable and fixed mirrors from the points P_1 and P_2. The two planewaves falling on the detector interact twice with the beamsplitter. Each planewave is once reflected at and once transmitted through the beamsplitter. Therefore, the beamsplitter adds no additional phase difference to the two planewaves and the total phase difference is given by $\Delta\varphi = \varphi_m - \varphi_f$. The situation is different for the two beams falling back on the source. Here, one beam is reflected at both interactions with the beamsplitter, whereas the other is transmitted at both interactions. According to eqn [2], between these two beams an additional phase shift of π results. Quantitatively, the total field amplitude at the detector, neglecting an arbitrary phase, is given by $E_D = E_0RT(e^{i\varphi_m} + e^{i\varphi_f})$, and the corresponding intensity is:

$$\begin{aligned} I_D \propto E_D E_D^* &= 2|E_0|^2|R|^2|T|^2\{1 + \mathrm{Re}[e^{i(\varphi_m-\varphi_f)}]\} \\ &= 2|E_0|^2|R|^2|T|^2\{1 + \cos(4\pi\sigma\Delta x)\} \end{aligned} \qquad [3]$$

The sum of the field amplitudes falling back on the source is $E_S = E_0(R^2e^{i\varphi_m} + T^2e^{i\varphi_f})$. For the intensity one gets:

$$I_S \propto E_S E_S^* = |E_0|^2\{|R|^4 + |T|^4 + 2\mathrm{Re}[R^2T^{*2}e^{i(\varphi_m-\varphi_f)}]\} \qquad [4]$$

Denoting the complex coefficients R and T by their modulus and their phase $R = |R|e^{i\varphi_R}$, $T = |T|e^{i\varphi_T}$ and using eqn [2], one gets:

$$I_S \propto |E_0|^2\{|R|^4 + |T|^4 + 2|R|^2|T|^2\cos(4\pi\sigma\Delta x + \pi)\} \qquad [5]$$

Both intensities, I_D and I_S contain a constant component and a component that varies with $\cos(4\pi\sigma\Delta x)$. The varying parts have the same amplitude $2|E_0|^2|R|^2|T^2|$ but are out of phase by π. Consequently, a maximum in I_D corresponds to a minimum in I_S and the sum of both intensities is independent of the mirror displacement Δx.

In most of the commercially available spectrometers, only I_D is used for spectral measurements. Usually, the detector signal is amplified by an AC coupled electronics. Therefore, in a measurement of $I_D(\Delta x)$, the constant part of the signal is suppressed. Assuming a spectral density $S(\sigma)$ of the source and an ideal detector with a responsivity R_D linear in the intensity and independent of the wavenumber σ, the detector signal $D(\Delta x)$ is given by:

$$D(\Delta x) = R_D\int_0^\infty S(\sigma)\cos(4\pi\sigma\Delta x)\mathrm{d}\sigma \qquad [6]$$

The information about the spectrum is contained in $S(\sigma)$ and can be extracted from the interferogram $D(\Delta x)$ by a cosine Fourier transformation. The transformation has to be done numerically and, therefore, it is more convenient to use a complex, fast Fourier transform algorithm. For that purpose, eqn [6] is modified by using the complex expansion of the cosine $\cos(x) = (e^{ix} + e^{-ix})/2$:

$$\begin{aligned} D(\Delta x) &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma + \int_0^\infty S(\sigma)e^{-i4\pi\sigma\Delta x}\mathrm{d}\sigma\right) \\ &= \frac{R_D}{2}\left(\int_0^\infty S(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma + \int_{-\infty}^0 S(-\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma\right) \end{aligned} \qquad [7]$$

Extending $S(\sigma)$ to negative wavenumbers according to:

$$S'(\sigma) = \begin{cases} S(\sigma) & \text{for } \sigma \geq 0 \\ S(-\sigma) & \text{for } \sigma < 0 \end{cases} \qquad [8]$$

leads to the following expression for the interferogram:

$$D(\Delta x) = \frac{R_D}{2}\int_{-\infty}^\infty S'(\sigma)e^{i4\pi\sigma\Delta x}\mathrm{d}\sigma \qquad [9]$$

and the spectral density $S'(s)$ can be obtained by the complex Fourier transform according to:

$$S'(\sigma) = \frac{4}{R_D}\int_{-\infty}^\infty D(\Delta x)e^{-i4\pi\sigma\Delta x}\mathrm{d}(\Delta x) \qquad [10]$$

In the following, we will skip the prime and S will refer to the spectral density extended to negative values of σ.

To calculate the spectral density according to eqn [10], one has to use an interferogram extending from $-\infty$ to $+\infty$ measured at infinitely small sampling distances $\mathrm{d}(\Delta x)$. In practise, of course, the interferogram is measured only between finite mirror displacements $-\Delta x_{\max}$ and $\Delta x_{\max}$, at sampling points equally spaced by ξ.

The effect of the finite mirror displacement on the resulting spectral density S can be included into eqn [10] by multiplying the interferogram $D(\Delta x)$ with a function that vanishes for $|\Delta x| > \Delta x_{\max}$. As an example, we use the boxcar function

$\theta(\Delta x + \Delta x_{\max})\theta(\Delta x_{\max} - \Delta x)$, where $\theta(x)$ is the Heavyside step function with values 0 and 1 for $x < 0$ and $x > 0$, respectively, and calculate the output spectrum for a spectral density $S_1(\sigma)$ consisting of a single line at wavenumber $\sigma_0 : S_1(\sigma) = (\delta(\sigma - \sigma_0) + \delta(\sigma + \sigma_0))/2$ (note that the spectral density is extended to negative wavenumbers according to the definitions given above). According to eqn [9], the interferogram is $D_1(\Delta x) = (R_D/2)\cos(4\pi\sigma_0\Delta x)$. Multiplying the interferogram with the boxcar function and Fourier transforming the product back using eqn [10], the resulting spectral density S_1^{box} is:

$$S_1^{\text{box}} = 2\Delta x_{\max}\left\{\frac{\sin[4\pi(\sigma_0 - \sigma)\Delta x_{\max}]}{4\pi(\sigma_0 - \sigma)\Delta x_{\max}} + \frac{\sin[4\pi(\sigma_0 + \sigma)\Delta x_{\max}]}{4\pi(\sigma_0 + \sigma)\Delta x_{\max}}\right\} \qquad [11]$$

Therefore, the approximation to a monochromatic spectrum is a sum of $\text{sinc}(z_\pm)$ $(= \sin(z_\pm)/z_\pm)$ functions where $z_\pm = 4\pi(\sigma_0 \pm \sigma)\Delta x_{\max}$. The lineshape function S_1^{box} is plotted around σ_0 in Figure 2 by the full line. Compared to the monochromatic spectrum S_1, the lines of S_1^{box} have a finite linewidth with a fullwidth at half maximum (FWHM) of approximately $0.3/\Delta x_{\max}$. As a general rule in spectroscopy, two spectral lines of equal intensity and linewidth have to be separated at least by their FWHM, in order to be clearly distinguishable. According to this definition, the limit of resolution of a FT spectrometer $\Delta\sigma_{\min}$ is inversely proportional to the maximum mirror displacement and given by $\Delta\sigma_{\min} = 0.3/\Delta x_{\max}$.

A less precise, but more intuitive definition of the resolution refers to the difference of the phases of the interferograms corresponding to two monochromatic lines separated by $\Delta\sigma$. The two lines are defined to be resolvable if the phases of their interferograms differ by more than 2π between the mirror positions $-\Delta x_{\max}$ and $\Delta x_{\max}$, i.e., the beating in the superposition of the interferograms shows at least one period. It is easily seen that, according to this definition $\Delta\sigma_{\min} = 1/(2\Delta x_{\max})$, a result in good agreement but slightly larger than that obtained by the more precise definition given above.

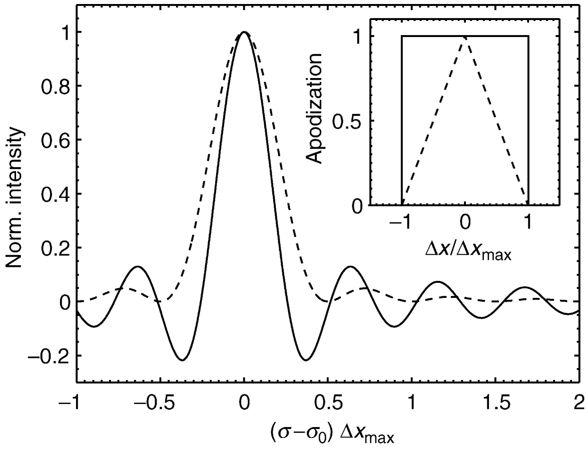

Figure 2 Normalized lineshape functions for boxcar (full line) and triangular apodization (broken line). In the inset, the corresponding apodization functions are shown.

As shown in Figure 2 by the full line, the sinc function has negative side-lobes with magnitudes of up to 22% of the maximum value. For some spectroscopic applications this side lobes are unwanted since they lead to a suppression of smaller peaks separated by approximately the resolution limit of the spectrometer. Thus, in these cases, it appears to be reasonable to multiply the interferogram by a (apodization) function different from the natural boxcar function. Usually, triangular or trapezoidal functions vanishing for $|\Delta x| > \Delta x_{\max}$ are used. By Fourier transformation of these functions, it can be shown that no negative side lobes are generated. However, these functions have a larger (up to ~50%) peak FWHM than the sinc function. For comparison, in Figure 2, also the lineshape of a monochromatic spectrum using a triangular apodization function $A(\Delta x) = \theta(\Delta x + \Delta x_{\max})\theta(\Delta x_{\max} - \Delta x)(1 - |\Delta x/\Delta x_{\max}|)$ is shown by the broken line. In this case, the negative sidelobes are strongly suppressed but the FWHM is increased.

Generally, the output spectrum $S^{\text{A}}(\sigma)$ corresponding to an arbitrary spectral density $S(\sigma)$ and an apodization function $A(\Delta x)$ is given by the convolution $S^{\text{A}}(\sigma) = 2[S \otimes A^{\text{FT}}](\sigma)$, where the lineshape function A^{FT} is the Fourier transform of $A(\Delta x)$:

$$A^{\text{FT}}(\sigma) = \int_{-\infty}^{\infty} A(\Delta x)\mathrm{e}^{-4\pi\sigma\Delta x}\mathrm{d}(\Delta x) \qquad [12]$$

and

$$[f \otimes g](t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t - \tau)\mathrm{d}\tau \qquad [13]$$

Discrete Sampling

In order to discuss the effect of the discrete, equally spaced sampling points with a mutual distance of ξ on the output spectrum, Δx is replaced by $x_n = n\xi$ in eqn [10], where n is an integer and runs from $-N_{\max}$ to $N_{\max}$ and $N_{\max}\xi = \Delta x_{\max}$. Instead of the continuous interferogram $D(\Delta x)$, a discrete interferogram:

$$D_n = \frac{R_D}{2}\int_{-\infty}^{\infty} S(\sigma)\mathrm{e}^{\mathrm{i}4\pi n\xi\sigma}\mathrm{d}\sigma \qquad [14]$$

results. From eqn [14], it is readily seen (by shifting the integration variable) that replacing $S(\sigma)$ with $S[\sigma + m/(2\xi)]$, where m is an integer, does not change the discrete interferogram D_n. Keeping in mind that according to our definition $S(-\sigma) = S(\sigma)$, it is clear that $S[-\sigma + m/(2\xi)]$ also leads to the same interferogram as $S(\sigma)$. Therefore, to be able to transform the interferogram to the spectral density $S(\sigma)$ for a wavenumber band $[0, \sigma_{\max}]$ unambiguously, one has to make sure, by electronically and/or optical filtering, that no wavenumbers outside this range contribute to the interferogram and that the sampling interval ξ is small enough so that $\sigma_{\max} < -\sigma_{\max} + 1/(2\xi)$, i.e., $\sigma_{\max} < 1/(4\xi)$. In Figure 3, a schematic sketch of a spectral density $S(\sigma)$, consisting of two spectral bands shown by the full and dashed lines, is plotted. For the spectral band shown by the full line, the sampling interval ξ is small enough so that in the interval $[0, 1/(4\xi)]$ no ambiguity arises. For the spectral band shown by the dashed line, the interferogram appears to be under-sampled: ξ is too large, and, therefore, the part of the spectral band at negative wavenumbers shifted by $+1/(2\xi)$ falls into the interval $[0, 1/(4\xi)]$ (shown by the dotted line in Figure 3). Thus, for the sampling interval ξ, the interferogram for the sum of spectral bands, shown by the full and the dashed lines in Figure 3, is the same as for the sum of the spectral bands shown by the full and dotted lines. Since the interferograms are the same, the two spectra are indistinguishable at sampling intervals ξ. This ambiguity of spectra for an under-sampled interferogram is called aliasing. Comparing the correct sampling interval $\xi < 1/(4\sigma_{\max})$ with the period of the interferogram D_n corresponding to a spectrum consisting of a single line at $\sigma_{\max}$ ($D_n = R_D \cos(4\pi n \xi \sigma_{\max})$), it follows that the interferogram has to be sampled more than two times per period in order to obtain an unambiguous result (Nyquist Theorem).

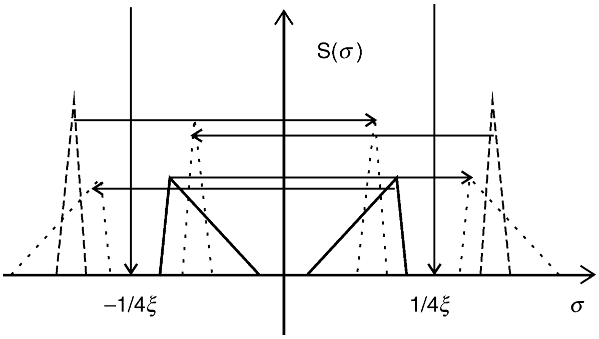

Figure 3 Schematic sketch of aliasing for under-sampled interferograms. For spectral density shown by the full line, the sampling interval ξ is chosen correctly and the $1/(2\xi)$ wavenumber ambiguity indicated by the arrows in the plot does not result in erroneous features in the spectral range $[0, 1/(4\xi)]$. The situation is different for the spectral band sketched by the dashed line: For this band, the sampling interval ξ is too large, and consequently, the $1/(2\xi)$ wavenumber ambiguity maps the line from the negative wavenumber region into the range $[0, 1/(4\xi)]$. The mapped spectral band is indicated by the dotted line.

Mathematically, the spectral density is calculated from the discrete interferogram D_n by a discrete Fourier transformation:

$$S_m = \frac{2}{R_D N} \sum_{n}^{N} D_n e^{-2\pi i nm/N} \qquad [15]$$

where N is the number of sampling points ($N = 2N_{\max} + 1$) and the wavenumbers corresponding to S_m are given by $\sigma_m = m/2\xi N$.

In order to obtain a high accuracy in the sampling intervals ξ, usually a reference interferogram of a monochromatic laser line (for example from a HeNe laser) is measured simultaneously with the interferogram of the actual experiment. By counting, for example, the zeros with positive slope in the interferogram of the laser line, uniformly spaced sampling points can be achieved with high accuracy. If, in an experiment many interferograms have to be averaged, the absolute position of the sampling points is important, in order to ensure that only points of equal mirror displacement are averaged. Therefore, in some spectrometers, the interferogram of a broadband source is measured simultaneously with the interferograms of the laser and the experiment and the maximum of the interferogram of the broadband source is taken as the origin for counting the zeros of the laser interferogram.

Phase Correction

Ideally, the interferogram is expected to be symmetrically around the position of equal optical path in the two arms of the interferometer. However, in practise this is seldom the case. Many reasons can lead to an asymmetric interferogram, the most obvious being that none of the discrete sampling points of the interferogram coincides with the mirror position for equal optical path in the two interferometer arms. Let Δx_{m} be the measured mirror displacement that differs from the real mirror displacement Δx by x_0, i.e., $\Delta x = \Delta x_{\mathrm{m}} + x_0$. In that case the measured interferogram is:

$$D(\Delta x_{\mathrm{m}}) = \frac{R_D}{2} \int_{-\infty}^{\infty} S(\sigma) e^{i[4\pi\sigma(\Delta x_{\mathrm{m}} + x_0)]} d\sigma \qquad [16]$$

and the Fourier transformation, according to eqn [10] with Δx_{m} as the integration variable, results in

$S(\sigma)e^{i4\pi x_0 \sigma}$, that is, in a spectrum with both real and imaginary parts. Also, mirror or compensator misalignment can result in a wavenumber-dependent optical path at a fixed mirror position. Finally, for rapid-scan interferometers, in that the movable mirror is scanned at a constant velocity v, the electronic filtering and amplification of the detector signal will also add different phases, depending on the signal frequency ν that is correlated to the wavenumber σ by $\nu = 2\sigma v$. In general, as long as the phase φ depends only on the wavenumber and not on the mirror displacement, the measured interferogram is:

$$D(\Delta x_m) = \frac{R_D}{2} \int_{-\infty}^{\infty} S(\sigma) e^{i[4\pi\sigma\Delta x_m + \varphi(\sigma)]} d\sigma \qquad [17]$$

and by the Fourier transformation, according to eqn [10], one obtains $S(\sigma)e^{i\varphi(\sigma)}$. There are several methods implemented in the software of commercially available Fourier spectrometers to eliminate the phase from the Fourier transform. In most experimental cases, in that the measured interferograms correspond to spectra with $S(\sigma) \geq 0$ for all $\sigma \in [0, \sigma_{max}]$, these methods usually give the correct results. However, care has to be taken if spectra contain positive and negative components. This can be the case, for example, if the transmission of a sample is periodically modulated by externally applied voltage or a laser light and the detector signal is amplified by a lock-in technique before it is fed into the electronics of the Fourier transform spectrometer. If the modulation increases the sample absorbance in one part of the spectral range and decreases it in a different part, then the measured interferogram corresponds to a spectrum with both negative and positive regions. In the following, first the phase correction methods for spectra with positive values only are discussed before it is argued, why these methods might fail for spectra with positive and negative parts.

For spectra with only positive values, the most simple method to eliminate the unknown phase is to measure the interferogram symmetrically around $\Delta x_m = 0$ and calculate $S(\sigma)e^{i\varphi(\sigma)}$ according to eqn [10]. Including the effect of finite scan length, the output spectrum becomes $S^A = 2[S(\sigma)e^{i\varphi(\sigma)}] \otimes A(\sigma)^{FT}$. Under normal experimental conditions, the phase $\varphi(\sigma)$ does not vary over the peak FWHM of $A(\sigma)^{FT}$ and, therefore, the phase term can be put outside the integral of the convolution and $S^A = 2e^{i\varphi(\sigma)}[S(\sigma) \otimes A(\sigma)^{FT}]$ results. Since $A(\Delta x_m) \geq 0$, and chosen to be symmetric around $\Delta x_m = 0$, the convolution is real and positive and the true spectrum with instrumental resolution $2[S(\sigma) \otimes A(\sigma)^{FT}]$ can be obtained by calculating the modulus of S^A without any knowledge about $\varphi(\sigma)$. However, for several reasons this method for phase correction is not ideal. First of all, since the maximum spectral resolution of a Fourier transform spectrometer is inversely proportional to the maximum mirror displacement, for a given spectrometer layout with a fixed maximal mirror scan length L, the maximal achievable spectral resolution is reduced by a factor of two if scans symmetric around $\Delta x = 0$ have to be performed. Secondly, for signals closer to or smaller than the noise level of the spectrometer, the random noise will be entirely positive by taking the modulus. Hence, for small signals, the noise level will be increased relative to the signal level by taking the modulus.

Therefore, usually the 'Mertz phase correction method' is used. This method is based on the assumption that the phase $\varphi(\sigma)$ is a slowly varying function of σ. In this case, $\varphi(\sigma)$ can be determined with low spectral resolution, i.e., by a short, symmetric scan of length $2L_1$ around $\Delta x = 0$. It can be shown that from this scan the phase can be obtained according to:

$$\varphi(\sigma) = \arctan\left\{ \frac{\mathrm{Im}[S(\sigma)e^{i\varphi(\sigma)} \otimes A^{FT}(\sigma)]}{\mathrm{Re}[S(\sigma)e^{i\varphi(\sigma)} \otimes A^{FT}(\sigma)]} \right\} + \pi \text{ (if denominator is negative)} \qquad [18]$$

Here, $S(\sigma)e^{i\varphi(\sigma)} \otimes A(\sigma)^{FT}$ results from the Fourier transformation of the short interferogram measured between $-L_1$ and L_1 and the apodization function $A(\Delta x_m)$ is adjusted to $-L_1$ and L_1. The addition of π for negative denominators is a consequence of the assumption that $S(\sigma)$ is positive everywhere. Under this assumption, a negative denominator in eqn [18] corresponds to a phase angle $|\varphi(\sigma)| > \pi/2$, that is mapped into $[-\pi/2, \pi/2]$ by the arctan-function. To get the correct phase angle, π has to be added to the phase angle, thereby extending the range of $\varphi(\sigma)$ to $[0, 2\pi]$.

With $\varphi(\sigma)$ known, the interferogram has to be recorded only for mirror displacements $\Delta x_m \geq 0$, since the information about the interferogram for $\Delta x_m < 0$ is contained in the phase spectrum and the interferogram for $\Delta x_m \geq 0$. Therefore, with this method, the totally available mirror displacement length can be used for increasing the spectral resolution.

However, for spectra containing positive and negative spectral bands, the Mertz phase correction usually fails, due to the addition of π if the denominator in eqn [18] becomes negative. For spectra with positive and negative bands, the sign of the denominator is determined by both $S(\sigma)$ and $\varphi(\sigma)$. Applying the Mertz algorithm in a negative region of the spectrum with a vanishing physical phase angle, results in a calculated phase angle of $\varphi = \pi$. As a consequence, the negative spectral band appears as a positive band in the output spectrum calculated according to the Mertz algorithm.

To avoid erroneous phase-flips, the addition of π in eqn [18] for negative denominators can be omitted and the allowed values for $\varphi(\sigma)$ are reduced to $[-\pi/2, \pi/2]$. The corresponding phase correction method is usually referred to as the 'Mertz signed method'. However, it is immediately clear from the discussion above, that in this case any physical phase angle outside the range $[-\pi/2, \pi/2]$ results to an erroneous π-jump of the phase. According to eqn [16], a physical phase shift larger than $\pi/2$ can, for example, be the consequence of a large offset x_0, i.e., of an inaccurate alignment of Δx_m and Δx. At $\Delta x = 0$, all contributions to the interferogram add constructively and the height of the interferogram is proportional to the area under the spectral curve $S(\sigma)$, as can be readily seen by setting $\Delta x = 0$ in eqn [10]. Therefore, for $S(\sigma) > 0$ the position $\Delta x_\mathrm{m} = 0$ is usually estimated by the global maximum of the measured interferogram. However, for spectra with both positive and negative bands, in most cases the interferogram does not exhibit a maximum at $\Delta x = 0$ (for positive and negative bands of equal height and width, the interferogram vanishes at $\Delta x = 0$), and estimating $\Delta x_\mathrm{m} = 0$ by an extremum of the interferogram results, therefore, in a large offset x_0. Thus, in order to be able to use the 'Mertz signed' algorithm, $\Delta x_\mathrm{m} = 0$ has to be determined separately in a calibration experiment in that the condition $S(\sigma) > 0$ is fulfilled. However, even for accurately determined $\Delta x_\mathrm{m} = 0$, erroneous phase jumps can appear if lines of different intensity and opposite sign are spaced approximately equal to the resolution used for determining the phase spectrum $\varphi(\sigma)$.

In conclusion, there are no general strategies to completely exclude phase jumps in the Fourier transformation of interferograms corresponding to spectra with positive and negative bands. The experimentalist has to check from case to case if lines with strange shapes appear in the output spectra, and modify the phase correction accordingly. More elaborate strategies for phase correction, that are beyond the scope of this article, are reported in literature.

Example of Experimental Results

To illustrate the features of FTS discussed in the previous paragraphs by a practical example, in Figure 4 the results of a typical measurement in the MIR between 500 cm^{-1} and 5000 cm^{-1} are shown. The experiments were performed with a BRUKER IFS113 spectrometer using a silicon carbide glowbar as a source for the infrared radiation, a potassium bromide beamsplitter, and a liquid-nitrogen cooled mercury cadmium telluride detector. The spectrometer used for the experiments is set up in a vacuum chamber. The two interferograms shown in Figure 4a were measured at two different pressures (broken line: 10 mbar; full line: 300 mbar), i.e., at two different levels of absorption of the atmospheric gas in the interferometer chamber (for clarity, the interferograms are offset vertically). The abscissa axis is broken in Figure 4a and shows portions of the interferograms around zero mirror displacement and around the maximum mirror displacement of 2.5 cm (chosen for 0.2 cm^{-1} spectral resolution). For higher chamber pressures, the increased absorption of the atmospheric gas results in an increased amplitude of the wiggles of the interferogram apparent for large mirror displacements in Figure 4a. The amplitude of these wiggles is approximately three orders of magnitude smaller than the detector signal observed at the center-burst of the interferogram around zero mirror displacement. However, in the corresponding spectra (Figure 4b) calculated from the interferograms after phase correction, the absorption of the atmospheric gas is clearly evident by a strong increase of the dense and narrow absorption lines in the spectral regions 1200 cm^{-1}–2200 cm^{-1}, 2300 cm^{-1}–2400 cm^{-1}, and 3200 cm^{-1}–4000 cm^{-1} (note that in Figure 4b the spectra are offset vertically by 0.1 for clarity). The lines observed between 2300 cm^{-1}–2400 cm^{-1} are due to absorption of the CO_2 molecule and are shown in more detail in Figure 4c from 2335 cm^{-1} to 2345 cm^{-1}. In order to simulate the contribution of the CO_2 lines to the interferogram shown in Figure 4a, two neighboring lines in Figure 4c were fitted to a doublet of Lorentz-absorption lines (indicated by the symbols in Figure 4c). The corresponding interferogram was calculated by Fourier transformation of the Lorentz lines and is shown in Figure 4d. It consists of a rapid oscillating, cosine-like signal (see inset for an expanded plot between 0.6 and 0.602 cm) with an amplitude modulation corresponding to the difference of the resonance wavenumbers of the two Lorentz lines. The ordinate axes of Figures 4a,d have the same arbitrary units and, therefore, the amplitudes of the interferograms shown in these two panels can be directly compared. From Figure 4a, d we conclude that the two CO_2 lines contribute approximately with only one in 10^5 parts to the center-burst of the measured interferogram, i.e., theses lines absorb only a fraction of 10^{-5} of the integrated (over the wavenumbers) intensity of

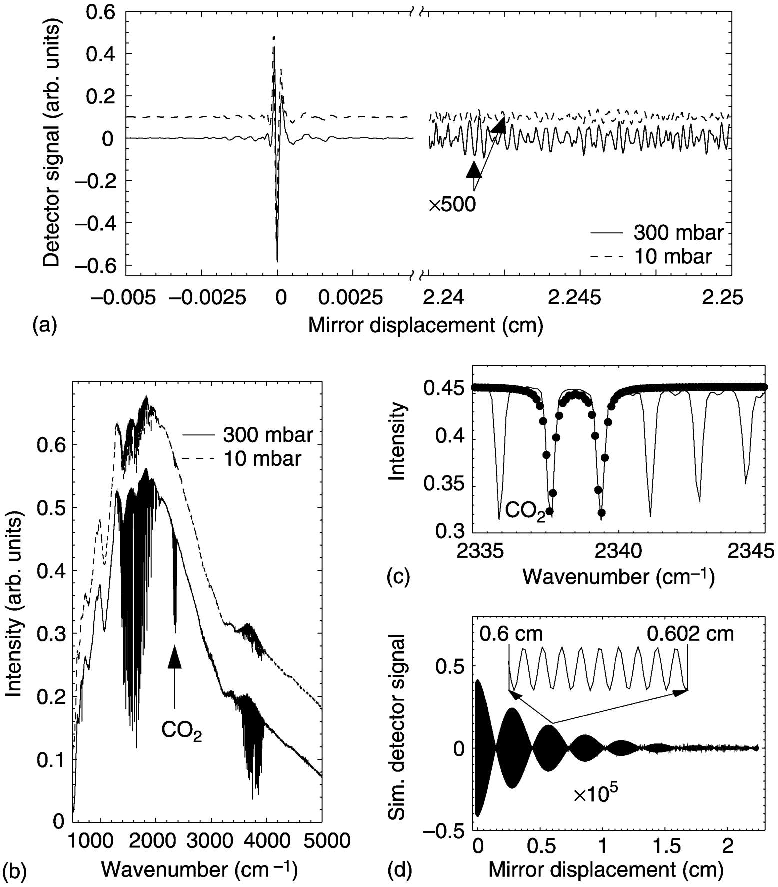

Figure 4 Measurement of the absorption of atmospheric gas for different pressures in the interferometer chamber (broken lines: 10 mbar, full lines: 300 mbar). (a) Interferograms (note the break in the axis of the mirror displacement). For clarity, the interferograms are offset vertically by 0.1 and 0.0002 on the left and right side of the axis break. On the right side of the axis break, the interferograms are amplified by a factor 500. (b) Spectra corresponding to the interferograms shown in (a). The spectra are offset vertically by 0.1 for clarity. (c) Part of the CO_2 absorption lines on an expanded wavenumber axis. The symbols indicate a fit of two Lorentz lines to neighbouring CO_2 absorption lines. The fit is used for calculating the interferogram shown in (d). (d) Simulated interferogram corresponding to the two CO_2 lines shown by the symbols in (c). The arbitrary units of the ordinate axis are the same as those used in (a) but the interferogram is amplified by a factor 10^5. In the inset, the interferogram is shown on an enlarged scale between 0.6 and 0.602 cm mirror displacement.

the incident radiation. Despite this small fraction, they are extremely well resolved with a large signal-to-noise ratio in Figures 4b,c, thus demonstrating the ability of FTS to measure spectra simultaneously with high resolution and over a large spectral range. This ability is one of the advantages (the so-called multiplex advantage) inherent in FTS and will be discussed in more detail in the following paragraph.

Advantage of FTS Over Dispersive Spectrometer

The two main advantages of FTS over dispersive spectroscopic methods, using, for example, a grating for the spectral decomposition of an incoming radiation, are directly connected with the working principle of FT spectrometers; high optical throughput and multiplexed measurement over the whole spectral range of interest. These advantages are discussed in the following, in more detail.

For a fair comparison of the optical throughput, spectrometers with equal spectral resolving power ρ have to be compared. Here, $\rho = \sigma_{\max}/\Delta\sigma$, where $\sigma_{\max}$ is the maximum wavenumber analyzed and $\Delta\sigma$ is the spectral resolution. For an ideal Fourier transform spectrometer, the limit for $\sigma_{\max}$ depends on the divergence angle of the beams propagating in the spectrometer. As sketched in Figure 5, the divergence angle α is determined by the radius r_a of the entrance aperture and the focal length f of the collimating lens by $\alpha \approx r_a/f$. This angle causes a phase shift between the rays propagating parallel to and those propagating under the angle of α inclined to the optical axis

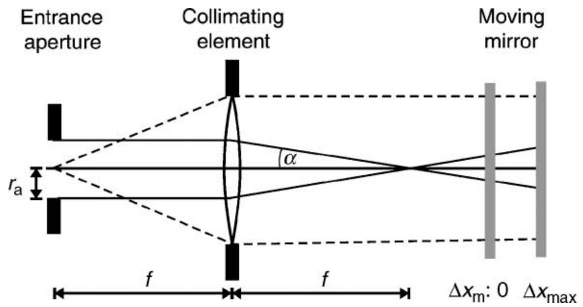

Figure 5 Entrance aperture and collimating element for an interferometer. For simplicity, the beamsplitter and fixed mirror are omitted in the sketch. The extreme rays entering the interferometer are indicated by the full and broken lines. Due to divergence α, a monochromatic light with wavenumber σ_0 produces all periods in the range $[4\pi\sigma_0, 4\pi\sigma_0/\cos(\alpha)]$ in the interferogram.

through the interferometer. The phase difference φ_{div} depends on the mirror displacement according to $\varphi_{div} = 4\pi\sigma\Delta x((1/\cos(\alpha)) - 1)$ and can be approximated for small α by $\varphi_{div} \approx 2\pi\sigma\Delta x\alpha^2$. It can be shown, that as long as φ_{div} does not exceed approximately π, the influence of the divergence on the interferogram can be tolerated, i.e., $\sigma_{max} \approx 1/(2\Delta x_{max}\alpha^2)$. Using $\Delta\sigma \approx 1/(2\Delta x_{max})$ for the resolution, one obtains for the resolving power of a FT spectrometer, $\rho = (f/r_a)^2$.

The optical throughput Θ describes the allowable energy per unit time that the system can let through and is related to area πr_a^2 of the aperture and the solid angle Ω of the collimating (or focusing) optics by $\Theta = \pi r_a^2 \Omega$. Using $\Omega = A_{FTS}/f^2$, where A_{FTS} is the clear area of the collimating lens (for a collimating mirror, A_{FTS} is the area of the mirror projected on a plane perpendicular to the beam direction), results in $\Theta_{FTS} = \pi A_{FTS} r_a^2/f^2 = \pi A_{FTS}/\rho$. For a grating spectrometer, $\Theta_{grating} = hA_G/f\rho$. (see respective chapter of this encyclopedia) where h is the height of the slit, f the focal length of the collimating optics, and A_G the projected area of the grating. Generally, h/f is smaller than 1/20 in up-to-date grating spectrometers. Thus, for the same resolving power ρ and similar instrument size, FT spectrometer can offer a more than $\pi f/h \approx 60$ times larger energy gathering capability. Therefore, for measurements with weak signals in that the detector noise is the dominant noise source, the large throughput of a FTS greatly increases the spectral signal to noise ratio that can be achieved for a fixed measuring time.

The other major principal advantage compared to dispersive methods relies on the simultaneous (multiplexed) gathering of information about all spectral bands $\Delta\sigma$ over a broad spectral range from σ_{min} and σ_{max}. For a grating spectrometer using a single detector element, the measurement time T_m available for recording the spectrum in that range has to be distributed between the $M = (\sigma_{max} - \sigma_{min})/\Delta\sigma$ spectral bands so that for a single spectral band the available integration time is T_m/M and the integrated signal increases proportional to T_m/M. If the dominant noise source is the detector noise which is independent of the signal level, then the noise will be proportional to $(T_m/M)^{1/2}$ and, therefore, the signal to noise ratio for the grating spectrometer $(S/N)_G$ is:

$$(S/N)_G \propto (T_m/M)^{1/2} \qquad [19]$$

For a FTS the situation is different, since it detects in the band $[\sigma_{min}, \sigma_{max}]$ all M small bands over the whole measurement time T_m via their contribution to the interferogram. So the integrated signal in a small band $\Delta\sigma$ is proportional to T_m. If the noise is again assumed to be random and independent of the signal level, for the FTS:

$$(S/N)_{FTS} \propto T_m^{1/2} \qquad [20]$$

results. For the same detector, assumed in the grating spectrometer and in the FTS, the proportionality constants are the same in eqns [19] and [20] and, therefore

$$\frac{(S/N)_{FTS}}{(S/N)_G} = M^{1/2} \qquad [21]$$

Equation [21] shows that the multiplex advantage becomes increasingly important for measurements of broad spectral bands with high resolution. In this case, an enhancement of the $(S/N)_{FTS}$ over $(S/N)_G$ by 2–3 orders of magnitude can be achieved. However, it has to be noted that the multiplex advantage of a FTS can be exploited only in cases where the detector noise is the dominant source of noise. In the visible and near infrared spectral region, low noise detectors, that allow single photon detection, are available. With these detectors, the spectral noise is dominated by the photon noise that is proportional to the square root of the intensity. It is evident that in this case the multiplexed detection is not an advantage, since for M spectral bands of width $\Delta\sigma$ contributing to the interferogram at a given mirror position, the noise is enhanced by $M^{1/2}$ compared to the noise that would be measured for a single band (here, it is assumed that all M bands contribute with the same intensity to the interferogram). Consequently, in this case, the noise enhancement by $M^{1/2}$ cancels the beneficial factor $M^{1/2}$ in eqn [21] and with single-detector grating and FT spectrometers, the same S/N ratio is achieved for a given measurement time T_m. In addition, using linear detector arrays with dispersive spectrometers, the multiplex advantage increases the S/N ratio for the dispersive spectrometer. However, detector arrays and low-noise detectors are mainly available for spectral regions with $\sigma > 6000\ \text{cm}^{-1}$, i.e., in the near-infrared and visible spectral region. In the FIR and MIR region ($\sigma < 6000\ \text{cm}^{-1}$) in that such detectors are absent, the multiplexing inherent in FTS makes it the method of choice if low-noise, broadband, and high-resolution spectroscopic data are required.

List of Units and Nomenclature

Angle of incidence for radiation on beamsplitter	Θ_1
Angle of propagation of radiation in the beamsplitter layer	Θ_2
Beam divergence angle	α
Beam splitter thickness (cm)	d
Complex reflection coefficient	R, r
Complex transmission coefficient	T, t
Detector Signal normalized to detector area (A cm^{-2})	D
Difference between real and measured mirror displacement (cm)	x_0
Dirac delta-function	δ
Discrete interferogram (A cm^{-2})	D_n
Discrete spectral density (W cm^{-1})	S_m
Distance from beamsplitter to fixed mirror (cm)	x_f
Distance from beamsplitter to movable mirror (cm)	x_m
Electric field amplitude of incident radiation (V/m)	E_0
Electric field amplitude of radiation falling back on source (V/m)	E_S
Electric field amplitude of radiation reaching detector (V/m)	E_D
Focal length (cm)	f
Fourier transform of apodization function *A* (cm)	A^FT
General apodization function	A
Heavyside step function	θ
Intensity falling back on source (W cm^{-2})	I_S
Intensity on detector (W cm^{-2})	I_D
Maximum mirror displacement (cm)	Δx_max
Maximum wavenumber (cm^{-1})	σ_max
Measured mirror displacement (cm)	Δx_m
Measured single line spectral density for a finite scan length (W cm^{-1})	S_1^box
Measured single line spectral density for a finite scan length using the apodization function *A* (W cm^{-1})	S^A
Measurement time for complete spectrum (s)	T_m
Minimum wavenumber (cm^{-1})	σ_min
Mirror displacement	Δx
Mirror velocity (cm s^{-1})	v
Modulation frequency (s^{-1})	ν
Number of sampling points of the interferogram	N
Number of spectral bands	M
Optical throughput (cm^2)	Θ
Optical throughput for FTS (cm^2)	Θ_FTS
Optical throughput for grating spectrometer (cm^2)	Θ_grating
Phase accumulation due to optical path to fixed mirror	φ_f
Phase accumulation due to optical path to moveable mirror	φ_m
Phase angle	γ
Phase angle	φ
Phase difference	$\Delta\varphi$
Phase difference between beams reflected at and transmitted through beamsplitter	$\Delta\varphi_\text{RT}$
Phase difference due to beam divergence	φ_div
Phase jump for reflection at beamsplitter	φ_R
Phase jump for transmission through beamsplitter	φ_T
Projected area of collimating element of FTS (cm^2)	A_FTS
Projected area of grating (cm^2)	A_G
Radius of entrance aperture (cm)	r_a
Refractive index	n
Responsivity (A W^{-1})	R_D
Sampling interval (cm)	ξ
Signal to noise ratio for FTS	$(S/N)_\text{FTS}$
Signal to noise ratio for grating spectrometer	$(S/N)_\text{G}$
Single line spectral density (W cm^{-1})	S_1
Solid angle	Ω
Spectral bandwidth (cm^{-1})	$\Delta\sigma$
Spectral density (W cm^{-1})	S
Spectral resolving power	ρ
Wave length (μm)	λ
Wave number (cm^{-1})	σ, σ_0
Wave number resolution (cm^{-1})	$\Delta\sigma_\text{min}$

See also

Interferometry: Overview; Phase Measurement Interferometry.

Further Reading

Bell RJ (1972) *Introductory Fourier Transform Spectroscopy*. New York, London: Academic Press.

Born M and Wolf E (1970) *Principles of Optics*. New York: Pergamon Press.

Griffiths PR (1975) *Chemical Infrared Fourier Transform Spectroscopy*. New York, London, Sydney, Toronto: John Wiley & Sons.
Hecht E (1987) *Optics*. Reading, MA: Addison-Wesley.
Mertz L (1965) *Transformations in Optics*. New York, London, Sydney: John Wiley & Sons.
Saptari V (2004) *Fourier-Transform Spectroscopy Instrumentation Engineering*. Bellingham: SPIE Press.

Hadamard Spectroscopy and Imaging

R A DeVerse, Plain Sight Systems, Inc., Kailua-Kona, HI, USA

R M Hammaker, Kansas State University, Manhattan, KS, USA

W G Fateley, Plain Sight Systems, Inc., Manhattan, KS, USA

F B Geshwind and A C Coppi, Plain Sight Systems, Inc., Hamden, CT, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Consider the task of measuring the spectra of a source, the spectra of an array of contiguous sources, or simply the broadband irradiance of an array of sources. These measures are commonly known as spectrometry, spectral imaging, and imaging, respectively. The sensitivity at which spectrometric, spectral imaging, and imaging systems can be operated is often dominated by noise associated with the detection system and the ability of the optics employed to deliver sufficient photonic flux to the detector subsystem to overcome this noise limitation. The ideal case is where the optical system can be fashioned in such a way as to deliver greater signal to the detector in order to maximize the signal-to-noise ratio (SNR) for a given finite source irradiance and detector noise.

Let us consider the number of discrete spectral resolution elements of spectrometers, spatio-spectral resolution elements of spectral imagers, and spatial resolution elements of imaging systems as being the N quantities of interest we would like to measure. We wish to measure these N quantities in time T with the greatest accuracy and precision as determined by the maximum SNR possible given a finite source flux and a fixed detector noise floor. One option might be to use a multi-element detection system (multi-channel detection system) with N detectors, so each detector is measuring one of the N quantities for the full time T in order to maximize the available integration time. However, in practice, these detectors can be problematic or simply not available for many applications when compared to single-element detectors. When a single-element detector is employed in spectrometry, spectral imaging, or imaging, the measurement methodology required becomes N measurements, each for a time T/N, so that each of the N quantities is measured separately or alternatively, one can record N measurements, each for time T/N, for various groups of the N quantities. The measurement procedure for various groups of the N quantities requires an encoding procedure to choose which of the N quantities to include in each of the N measurements and a decoding procedure to recover the values for each of the N individual quantities. When combinations of N quantities are measured, the signal at the single detector element is increased and the procedure is called multiplexing. Multiplexing may be defined as measuring the sum of groups of quantities rather than measuring each quantity separately with the primary goal of increasing SNR. Multiplexing can and does increase SNR when the detection system of the measuring device is operating in the detector noise limit. The detector noise limit is defined as the noise of the detector output that is independent of the amplitude of the signal. Photon-noise-limited detector systems can benefit from multiplexing strategies, however, one must be careful not to realize a multiplex disadvantage when the source of noise is not detector noise limited.

Fourier Transform Multiplexing

In optical spectrometry a common multiplexing method employs the Michelson interferometer in a Fourier transform (FT) spectrometer where, in the best case, as much as half of the spectral energy of the source entering the aperture is incident upon the single detector element for the duration of the measurement. The output of the detector is recorded as an interferogram, which is the detector response as a function of optical retardation (twice the path difference between the arms of the Michelson interferometer). The recorded interferogram is decoded using a Fourier transform algorithm to yield the desired spectrum. FT spectrometers are routinely employed in near-infrared, mid-infrared, and far-infrared spectrometry. In Raman spectrometry, FT spectrometers are used in the near-infrared spectral region typically using an Nd:YAG laser providing an excitation wavelength of 1064 nm. FT techniques are also employed in nuclear magnetic resonance (NMR) spectrometry and mass spectrometry.

Hadamard Transform Multiplexing

Hadamard transform multiplexing in dispersive optical spectrometry can be accomplished by use of an encoding mask at the focal plane of a dispersive spectrometer, that is operable to select different combinations of spectral resolution elements that are allowed to pass to the single detector element for each of the *N* required measurements. The specification of which ones of the *N* quantities are included in each of the *N* measurements of groups of the *N* quantities is called a weighing design. The term weighing design originally referred to a procedure for accurately determining the mass of a number of objects by weighing them in groups rather than one at a time. The optimal weighing designs and corresponding optimal encoding masks are specified by Hadamard mathematics. The Hadamard matrix or **H** matrix has all elements of +1 and −1 and provides the optimal weighing design for a two-pan balance or an optical instrument having two detector elements for the +1 and −1 measures. The Hadamard matrices specify which quantities go to each balance pan or detector element.

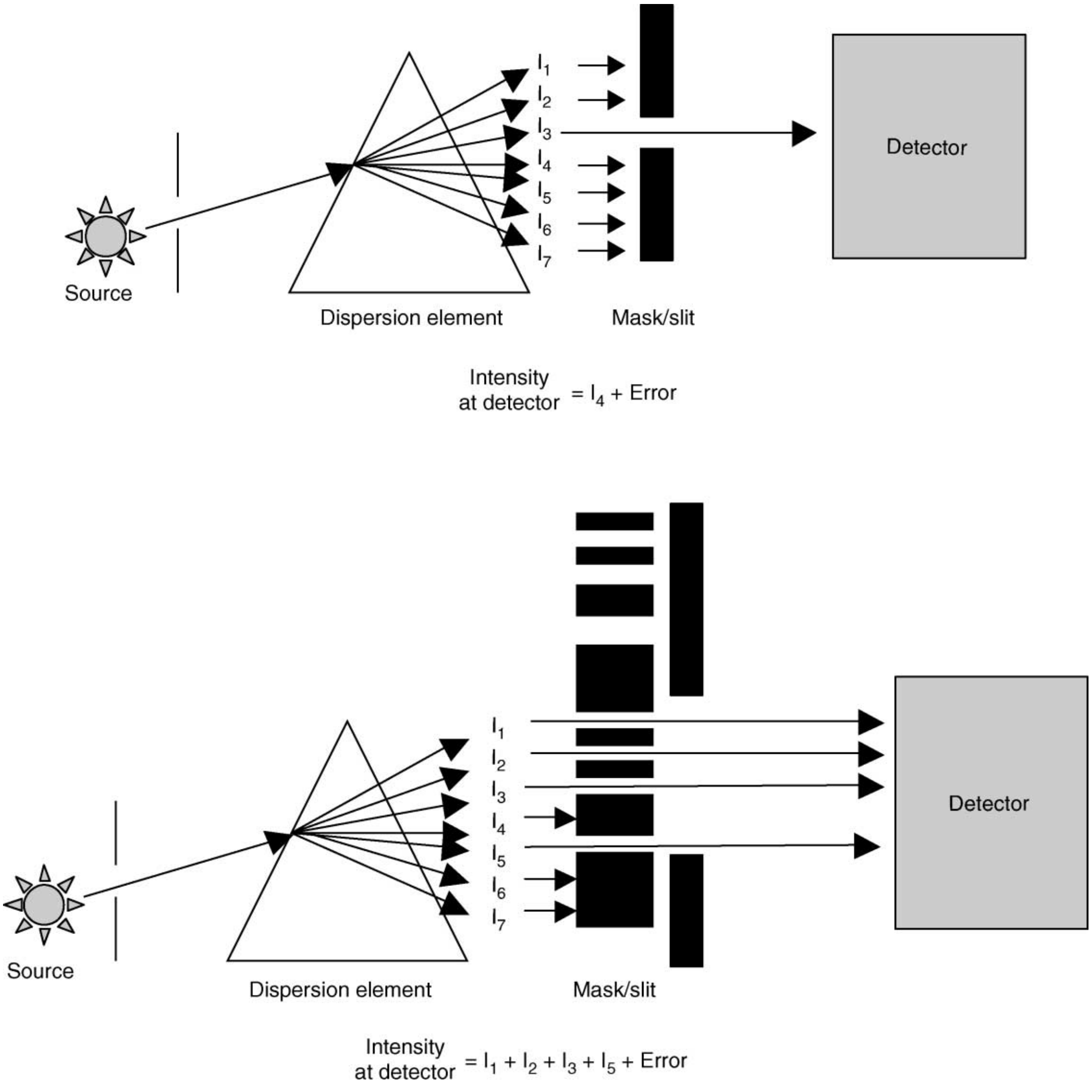

Figure 1 Simple schematic of a conventional and a Hadamard transform spectrometer. The top shows a conventional spectrometer where only one of seven spectral resolution elements ($N = 7$) impinge upon the single element detector for each of *N* measurements. The bottom shows a Hadamard transform spectrometer where $N = 7$. One more than half of the spectral resolution elements impinge upon the single element detector for each of the *N* measurements.

One and Two Detector Hadamard Encodement Matrices

In the case of a dispersive spectrometer with a single detector element, the optimal weighing designs and corresponding optimal encoding masks are provided by the simplex matrix or **S** matrix derived from the **H** matrix. The type of **S** matrix that is most convenient for optical applications is the cyclic or left circulant **S** matrix where only the first row needs to be known to generate the complete cyclic **S** matrix of N columns and N rows. Each row is obtained by shifting the previous row by a left-shift registration in which the element in the far left column moves to the far right column of the next row and each of the other $N-1$ elements moves one column to the left. Each of the N rows of the cyclic **S** matrix specifies one of the N encodements and each of the N columns specifies one of the N spectral resolution elements. The elements of the cyclic **S** matrix are 1 and 0 instead of the +1 and −1 found in the **H** matrix implementation. Spectral resolution elements allowed and not allowed to reach the single detector element are given as 1's and 0's, respectively. Each row and each column of the cyclic **S** matrix is a different sequence of $(N+1)/2$ elements of 1's and $(N-1)/2$ elements 0's so that each encodement in the row of the cyclic **S** matrix has a different combination of $(N+1)/2$ spectral resolution elements that are incident upon the single detector element. Just as the positions of the 1's in a row show which spectral resolution elements are included in a particular encodement, the positions of the 1's in a column show which encodements include that particular spectral resolution element. The result is that provided the detector noise is independent of its output signal, the SNR is increased by a factor of $(N+1)/(2\sqrt{N})$, approximately $(\sqrt{N})/2$, when compared to the method of measuring each spectral resolution element separately as in a raster scan performed by a classical scanning dispersive spectrometer.

Multiplexing in dispersive optical spectrometry as introduced in the previous paragraph is called Hadamard transform spectrometry (HTS) (see Figure 1). It is a unique combination of dispersive and multiplexing spectrometries. The choice of transparent or opaque for the elements of a multislit array or Hadamard encoding mask provides an encoding of spectral information that may be decoded into a conventional spectrum by Hadamard mathematics. In a typical dispersive spectrometer, the source energy is collected and separated into its individual spectral resolution elements by a spectral separator and then collected and focused for spatial presentation on a focal plane. Unlike a scanning dispersive spectrometer, which possesses a single exit slit, the Hadamard transform (HT) spectrometer employs a multislit array as a Hadamard encoding mask. This arrangement allows for the simultaneous measurement of a multitude of spectral resolution elements at a single-element detector and produces a multiplexing spectrometer.

Theory of Hadamard Multiplexing

The operation of the HT spectrometer can be described using matrix algebra with column vectors [**D**] and [**I**] and the N-order square matrices $\mathbf{S}_N$, $\mathbf{S}_N^{-1}$, $\mathbf{S}_N^{T}$, and $\mathbf{T}_N$. [**D**] is a column vector of N rows containing the detector responses for the N encodements. [**I**] is a column vector of N rows containing the N spectral resolution elements to be determined. $\mathbf{S}_N$ is the cyclic **S** matrix of order N that is the basis for construction and design of the Hadamard encoding mask. $\mathbf{S}_N^{-1}$ is the inverse of $\mathbf{S}_N$ and is given by $(2/(N+1))\mathbf{T}_N$. $\mathbf{T}_N$ is a matrix of +1 and −1

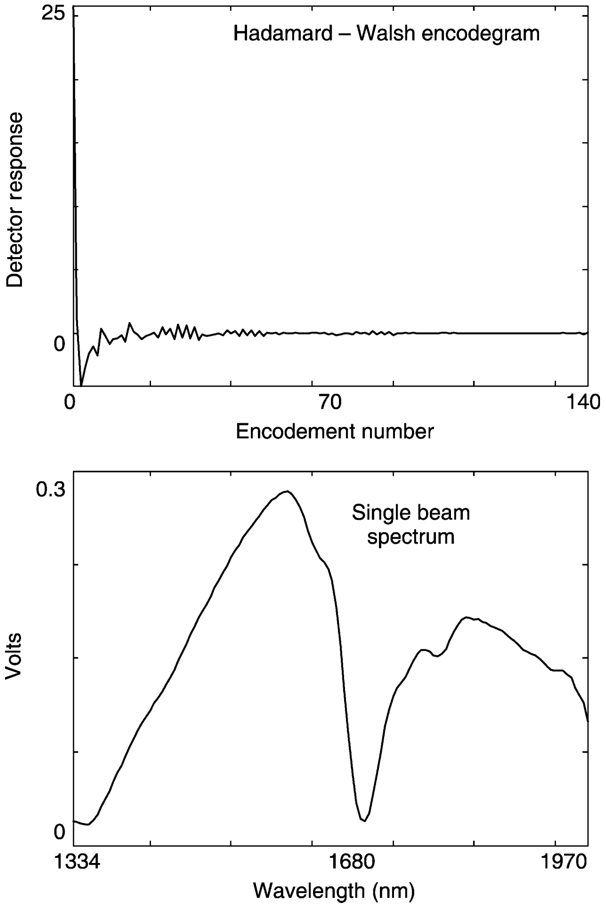

Figure 2 An encodegram (top plot) and the corresponding single-beam spectrum (bottom plot) as a result of decoding the encodegram using Hadamard mathematics.

elements obtained from S_N^T, the transpose of S_N, with T_N having +1's and −1's where S_N^T has 1's and 0's, respectively. A cyclic matrix is required to be symmetric so that an interchange of columns and rows can recover the original matrix and S_N^T then becomes equal to S_N. The encoding of the radiation by the HT spectrometer is given by $[\mathbf{D}] = S_N[\mathbf{I}]$ and the resulting primary data are a record of detector response versus encodement number. The plot of detector response versus encodement number is called an encodegram (analogous to the interferogram produced by an FT spectrometer). The decoding of the encodegram to recover the conventional spectrum using a Hadamard transform can be described by $[\mathbf{I}] = S_N^{-1}[\mathbf{D}]$, obtained by left multiplying $[\mathbf{D}] = S_N[\mathbf{I}]$ by S_N^{-1}. In practice, it is not necessary to invert S_N because there is a fast Hadamard transform (FHT) algorithm available when N is of the form $N = 2^w - 1$, w being an integer. There are three known methods for generating the first row of cyclic **S** matrices of order $N = 2^w - 1$. Consider the example of $N = 7$ for $w = 3$ where the first row of S_7 is 1 1 1 0 1 0 0. Figure 2 shows an encodegram and the result of decoding using Hadamard mathematics. Figure 3 contains the matrices and matrix equations for $N = 7$.

Hadamard Encoding Masks

The key component of any HT spectrometer is the Hadamard encoding mask (Figure 4). Two choices for the Hadamard encoding mask are the moving mechanical mask and the stationary electro-optical mask. The moving mask for N spectral resolution elements requires $2N - 1$ mask elements that are either completely open with a transmittance of 1 or completely closed with a transmittance of 0. The stationary mask for N spectral resolution elements requires only N mask elements that are selected to be either transparent: transmittance $T_t = 1$ or <1 or opaque: transmittance $T_o = 0$ or >0. The resulting SNR is dependent upon $\Delta T = T_t - T_o$, which has its optimal value of unity only for an ideal mask which is $T_t = 1$, $T_o = 0$. The advantage of the moving mask is

$$[D] = \begin{matrix} D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \\ D_6 \\ D_7 \end{matrix} = \begin{matrix} I_1 + I_2 + I_3 + I_5 \\ I_1 + I_2 + I_4 + I_7 \\ I_1 + I_3 + I_6 + I_7 \\ I_2 + I_5 + I_6 + I_7 \\ I_1 + I_4 + I_5 + I_6 \\ I_3 + I_4 + I_5 + I_7 \\ I_2 + I_3 + I_4 + I_6 \end{matrix} \qquad S_7 = \begin{matrix} 1&1&1&0&1&0&0 \\ 1&1&0&1&0&0&1 \\ 1&0&1&0&0&1&1 \\ 0&1&0&0&1&1&1 \\ 1&0&0&1&1&1&0 \\ 0&0&1&1&1&0&1 \\ 0&1&1&1&0&1&0 \end{matrix} \qquad [I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix}$$

$$T_7 = \begin{matrix} 1&1&1&-1&1&-1&-1 \\ 1&1&-1&1&-1&-1&1 \\ 1&-1&1&-1&-1&1&1 \\ -1&1&-1&-1&1&1&1 \\ 1&-1&-1&1&1&1&-1 \\ -1&-1&1&1&1&-1&1 \\ -1&1&1&1&-1&1&-1 \end{matrix} \qquad S_7^{-1} = (1/4)T_7 = \begin{matrix} (1/4)(1\ 1\ 1 -1\ 1 -1 -1) \\ (1/4)(1\ 1 -1\ 1 -1 -1\ 1) \\ (1/4)(1 -1\ 1 -1 -1\ 1\ 1) \\ (1/4)(-1\ 1 -1 -1\ 1\ 1\ 1) \\ (1/4)(1 -1 -1\ 1\ 1\ 1 -1) \\ (1/4)(-1 -1\ 1\ 1\ 1 -1\ 1) \\ (1/4)(-1\ 1\ 1\ 1 -1\ 1 -1) \end{matrix}$$

$$[I] = \begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \\ I_6 \\ I_7 \end{matrix} = S_7^{-1}[D] = \begin{matrix} (1/4)(D_1 + D_2 + D_3 - D_4 + D_5 - D_6 - D_7) \\ (1/4)(D_1 + D_2 - D_3 + D_4 - D_5 - D_6 + D_7) \\ (1/4)(D_1 - D_2 + D_3 - D_4 - D_5 + D_6 + D_7) \\ (1/4)(-D_1 + D_2 - D_3 - D_4 + D_5 + D_6 + D_7) \\ (1/4)(D_1 - D_2 - D_3 + D_4 + D_5 + D_6 - D_7) \\ (1/4)(-D_1 - D_2 + D_3 + D_4 + D_5 - D_6 + D_7) \\ (1/4)(-D_1 + D_2 + D_3 + D_4 - D_5 + D_6 - D_7) \end{matrix}$$

For example: Note $I_1 = (1/4)\,[\,(D_1 + D_2 + D_3 + D_5) - (D_4 + D_6 + D_7)\,]$ so

$$I_1 = (1/4)\,\{\,[\,(I_1 + I_2 + I_3 + I_5) + (I_1 + I_2 + I_4 + I_7) + (I_1 + I_3 + I_6 + I_7) +$$
$$(I_1 + I_4 + I_5 + I_6)\,] - [\,(I_2 + I_5 + I_6 + I_7) + (I_3 + I_4 + I_5 + I_7) +$$
$$(I_2 + I_3 + I_4 + I_6)\,]\,\} = (1/4)\,\{\,4I_1\,\} = I_1$$

Figure 3 Matrices and matrix equations (note since $\mathbf{S}_7$ is symmetric $\mathbf{S}_7^T = \mathbf{S}_7$).

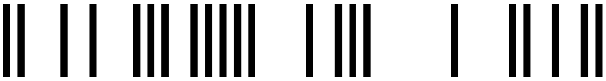

Figure 4 A linear encodement of length $N = 43$ generated by using a quadratic residue construction.

no optical transmission problems. However, it suffers the disadvantage of having moving parts and position repeatability problems. For the case of the stationary mask, it has the advantage that it has no moving parts and no positioning repeatability problems. The disadvantage of the stationary mask is the optical transmission problem where $T_t < 1$ and $T_o > 0$. The moving Hadamard encoding mask of $2N - 1$ mask elements is changed to the next encodement by translation of the Hadamard encoding mask to the left past the aperture by a single mask element. For $N = 7$ and $2N - 1 = 13$, the assignment of 1 and 0 to positions 1 through 13 is [1 1 1 0 1 0 0 1 1 1 0 1 0]. For example, the first encodement corresponding to the first row of S_7 in Figure 3 is from positions 1 through 7 as [1 1 1 0 1 0 0] and the fourth encodement corresponding to the fourth row of S_7 in Figure 3 is from positions 4 through 10 as [0 1 0 0 1 1 1].

2D Encoding Masks for Multiplexing in Spectral Imaging and Imaging

If the conventional one-dimensional (1D) Hadamard encoding mask is folded in some manner, the result is the generation of a two-dimensional (2D) Hadamard encoding mask (Figure 5). Similar to the 1D Hadamard encoding masks that are used for spectral mutiplexing, 2D Hadamard encoding masks can be used for the spatial multiplexing done in Hadamard transform spectral imaging (HTSI) and Hadamard transform imaging (HTI). It is important to note that spectrometry as HTS, spectral imaging as HTSI, and imaging as HTI all realize the same theoretical SNR improvement based on the number of spectral, spatio-spectral, or spatial resolution elements (N), respectively. Consider the folding of a 1D Hadamard encoding mask for $N = 15$ into a 2D Hadamard encoding mask of three rows and five columns (Figure 6). A cyclic S_{15} has a first row of [0 0 0 1 0 0 1 1 0 1 × 0 1 1 1 1] so the 2D Hadamard encoding mask has a first row of [0 0 0 1 0], a second row of [0 1 1 0 1], and a third row of [0 1 1 1 1] from columns 1 through 5, 6 through 10 and 11 through 15, respectively, from S_{15}.

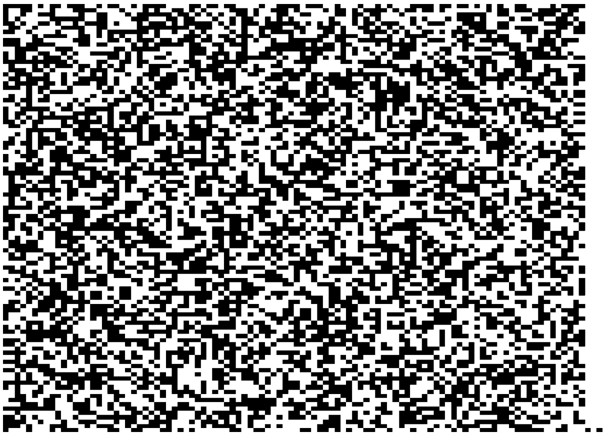

Figure 5 A two dimensional encodement of length $N = 10{,}007$ in a 100×100 format generated by quadratic residue construction. The remaining 7 elements are shown as the last 7 elements in the bottom row on the right. In practice these 'extra' data points can be masked off and used as a measure of system noise characteristics.

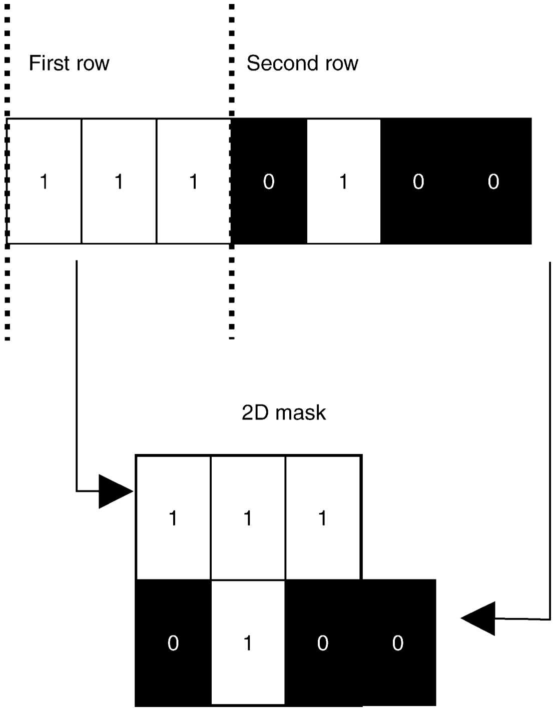

Figure 6 Spectrometry to imagery: The folding of a ID encoding mask that can be used for spectrometry is used to generate a 2D encoding mask that can be used for imaging. Note that using a symmetric aperture window where only 3×2 elements are used, the mask element that is not included in the windowed aperture can be used as a measure of the dark noise of the system, however, a more conventional encodement matrix might be one that can be divided into an integer number of rows and columns such as $N = 15$ where one can construct a 5×3 2D encodement mask for imagery and use a symmetric aperture that uses the entire length of the encodement matrix.

Hadamard Mathematics

Consider a general multiplexing problem where one seeks to measure N quantities, $X = \{x_0, x_1, \ldots, x_{N-1}\}$, in the presence of noise, and wishes to determine a vector function $F(X)$, such that the measurement of $F(X)$ yields the best estimate of X. Conservation of energy and superposition principles dictate a restriction to linear functions. The principle of rank suggests that we restrict our measurements to N-dimensional linear combinations. Determining the best linear combination is the challenge. In the case of constant error for each measurement, an answer is provided by Hotelling's theorem, which proves Hadamard matrices are optimal for such measurements. Hadamard matrices uniquely achieve the optimal bound for root-mean-square measurement error. A Hadamard matrix is a matrix $\mathbf{H}$ for which $\mathbf{H}_{i,j} = \pm 1$, and $\mathbf{H}^{-1} = \frac{1}{N}\mathbf{H}^{\mathrm{T}}$. The definitions of matrix inverse, transpose, and multiplication dictate that the rows of the matrix $\mathbf{H}$ form an orthogonal basis with each basis element having norm $\sqrt{N}$. The norms of the rows of $\mathbf{H}$ follow from the entries of the matrix that are ± 1, and place no additional restriction on $\mathbf{H}$. The columns of $\mathbf{H}$ are also an orthogonal basis with these same norms. The rows (columns) of the matrix $\frac{1}{\sqrt{N}}\mathbf{H}$ form an orthonormal basis. From these few statements and definitions, one can already derive estimates of the multiplex advantage. For example, one seeks to measure X and in one case measures $Y_1 = X + \varepsilon$, where ε is a random noise or error that has a Gaussian distribution with a mean of 0 and variance σ^2. In another case, suppose that one measures $Y_2 = HX + \varepsilon$. In this second case, one would estimate

$$X_2 = \frac{1}{N}\mathbf{H}^{\mathrm{T}}(\mathbf{H}X + \varepsilon) = X + \left(\frac{1}{\sqrt{N}}\right)^2 \mathbf{H}^{\mathrm{T}}\varepsilon$$
$$= X + \frac{1}{\sqrt{N}}\varepsilon_1$$

where ε_1 is random noise with the same characteristics as ε. It follows that the measurement error in the second case has a standard deviation σ that is less than in the former case, with an improvement by the factor $\frac{1}{\sqrt{N}}$.

Hadamard Computations

The ± 1 Hadamard matrices can be applied to two detector optical systems, and 0,1 matrices can be applied to single detector systems. In a two detector system $\mathbf{H}^{\mathrm{T}}$ is proportional to $\mathbf{H}^{-1}$ and one can invert these transforms. Suppose that one has chosen a particular Hadamard matrix $\mathbf{H}$, with which to conduct a multiplexed measurement. One can show that multiplication of any row or any column by -1 does not affect the efficacy of $\mathbf{H}$ in improving SNR. Hence one may assume that the first row and first column of $\mathbf{H}$ are all $+1$. From $\mathbf{H}$, we can form a new $(N-1) \times (N-1)$ matrix $\mathbf{S}$ by deleting the first row and column of $\mathbf{H}$, and then changing each 1 to a 0, and each -1 to a 1. This new matrix $\mathbf{S}$, an $\mathbf{S}$-matrix, is the one corresponding to $\mathbf{H}$ that is to be used for single-element detector optical systems. It is known that $\mathbf{S}$ matrices are optimal for 0,1 matrices just as Hadamard matrices are optimal for ± 1 matrices, with signal to noise improvement by a factor $\frac{\sqrt{N-1}}{2}$, where here N is the dimension of $\mathbf{H}$.

Letting $\mathbf{I}$ be the identity matrix, and $\mathbf{J}$ be the matrix whose entries are all 1, it follows that $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \frac{N}{4}(\mathbf{I} + \mathbf{J})$, where N is the dimension of $\mathbf{H}$. From this we can ascertain that a multiplexed system using the matrix $\mathbf{S}$, multiplied by the encoded parameters of $\mathbf{S}^{\mathrm{T}}$ recovers the original signal to within the addition of an offset. In practice this is what is done to recover the original X. In certain cases, there is also a fast algorithm for computing $\mathbf{S}^{\mathrm{T}}Y$, without ever explicitly computing $\mathbf{S}^{\mathrm{T}}$.

Constructions for Hadamard Matrices

It is known that if $\mathbf{H}$ is an $N \times N$ Hadamard matrix, then N must be 1, 2, or a multiple of 4 and so nontrivial S-matrices must be of size one less than a multiple of 4. It is conjectured that Hadamard matrices exist for each N which is a multiple of 4. Some constructions of Hadamard matrices are enumerated below.

Walsh–Hadamard Construction

A Hadamard matrix of size $N = 2^K$ can be constructed by sampling the first N Walsh functions at $0, 1/N, 2/N, 3/N, \ldots, (N-1)/N$, where the Walsh functions are defined by $W_0(x) = 1$, for $0 \le x < 1$, $W_0(x) = 0$ otherwise, and the recurrence relations, for each non-negative integer n:

$$\begin{aligned} W_{2n}(x) &= W_n(2x) + W_n(2x-1), \text{ and} \\ W_{2n+1}(x) &= W_n(2x) - W_n(2x-1) \end{aligned} \qquad [1]$$

Tensor Product Construction

Walsh is a special case of iterating this construction, starting with the matrix $\mathbf{H}_1 = [1]$. In general, if $\mathbf{H}_{\mathrm{N}}$ is a Hadamard matrix of size $N \times N$, then:

$$\mathbf{H}_{2N} = \begin{bmatrix} H_N & H_N \\ H_N & -H_N \end{bmatrix} \qquad [2]$$

is a Hadamard matrix of size $2N \times 2N$.

Quadratic Residue Construction

This construction is an example of a cyclic S-matrix: one for which each row is a cyclic shift of the previous row. Given a prime number $p = 4n + 3$, define, for $i = 0, \ldots, p-1$, the values $\mathrm{S}_{1,i} = 1$ if there is some integer k, such that the remainder of dividing k^2 by p is i, and $\mathrm{S}_{1,i} = 0$ otherwise. Taking the cyclic matrix with the first row given by $\mathrm{S}_{1,i}$, yields a Hadamard S-matrix.

Other constructions are possible. Maximal length shift register codes, and a construction involving quadratic residues of twin primes are two examples.

Fast Hadamard Transforms

The fast Walsh transform can be deduced from eqn [1]. The basic idea is that, in eqn [1], the function W_n is specified by $\log_2 n$ recurrence relations. For this reason, a Walsh transform (or its transpose) can be computed in $O(n \log_2 n)$ steps. The algorithm goes as follows, for a sequence of length $N = 2^n$:

(i) begin with $Y_{0,0}, Y_{0,1}, \ldots, Y_{0,N-1} = Y_0, Y_1, \ldots, Y_{N-1}$
(ii) for $i = 1, \ldots, n-1$, let $Y_{i,k} = Y_{i-1,2k} + Y_{i-1,2k+1}$, and $Y_{i,n/2+k} = Y_{i-1,2k} - Y_{i-1,2k+1}$, for $k = 0, \ldots, N/2$
(iii) output $X_i = Y_{n-1,i}$

Then the resulting X_i is the inverse Walsh transform of Y_i.

Note that cyclic S matrices that come from maximal length shift registers are known to be permutations of the Wzalsh–Hadamard S matrices, and hence there is also a fast algorithm for inverting these, albeit involving some tedious permutations.

Optics for Hadamard Encoded Apertures

Hadamard optical systems utilize spatially encoded apertures that can be employed either at the entrance aperture of the optical system, the exit aperture or both. They have the common attribute that they encode the available aperture spatially where the spatial resolution elements that make up the encodement dictate the spectral, spatio-spectral, or spatial resolution elements that propagate through the optical system, including diffractive optical elements and on to the sensor or exit aperture. These masks have some spatial extent that places special requirements on the optics of the system. As the encodement mask grows either by longer length encodements with fixed subapertures or as the subaperture dimension grows for a fixed encodement length, the spatial resolution elements making up the subapertures in the encodement mask depart from the optical axis. When the resolution elements depart from the optical axis or paraxial condition, it becomes important to employ optics that can image the off-axis resolution elements without inducing excessive aberrations that degrade the performance and cripple the advantages gained by HT multiplexing.

Typically the optical path for conventional monochromators begins with a source that is focused onto an aperture plane that has a large aspect ratio aperture know as a slit. This slit is often very small in extent in the dispersion plane compared to the other extent in the spatial plane. However, it is not required that this aspect ratio is large. If the aspect ratio is close to 1 then simple spherical optics can be employed that perform well as long as the departure from the optical axis is kept to a minimum. However, most monochromators have a large aspect ratio in order to increase the opportunity to maximize throughput, and detectors must be able to 'see' the large extent of the slit aperture. The light entering the slit aperture is then dispersed and focused onto an exit slit aperture. Monochromators are only required to perform well on the optical axis and do not typically employ optics that can manage rays that depart from the optical axis in the plane of dispersion as required by HT multiplexing instruments. To employ encoding techniques the optical system is required to utilize optical performance attributes normally found only in imaging and spectral imaging systems. This requirement is driven by the extent of the encoding mask. The extent of the encoding mask is governed by the diffraction limit of the wavelengths within the bandpass, the encodement length N, and the attributes of the optical system.

In a conventional dispersive spectrometer the radiation from a source is collected and separated into its individual spectral resolution elements by a spectral separator such as a diffraction grating or prism and then is collected and focused for spatial

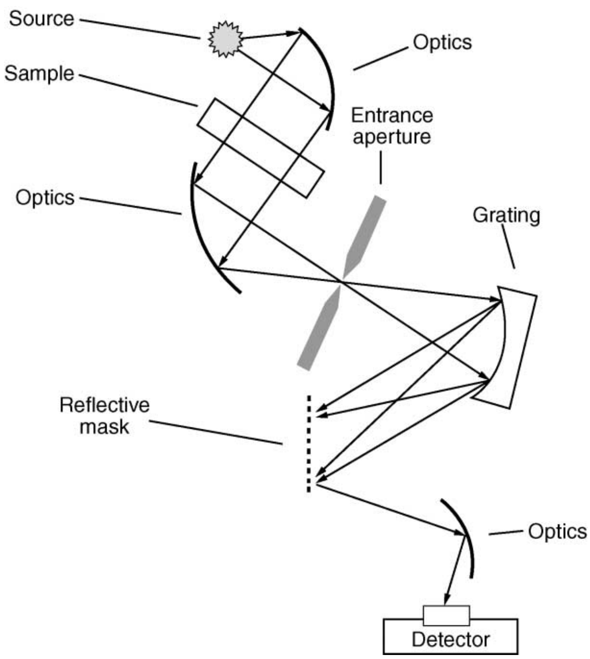

Figure 7 Example implementation of a Hadamard transform spectrometer with reflective optics. Light from a source passes through a sample, is transferred into an entrance aperture, bounced off of a concave grating, and then a reflective mask, and finally transferred to a detector.

presentation on a focal plane. The dispersive spectrometer uses a single exit slit to select one spectral resolution element at any given time, out of N possible spectral resolution elements for measurement by the detector. The Hadamard transform spectrometer (HTS, Figure 7) uses an array of slits (i.e., a mask) at the focal plane to select one more than half, $(N+1)/2$, of the spectral resolution elements at the focal plane for measurement by the detection system. The optical challenge to effect an HT multiplexing spectrometer is to collect all of the spatially distributed individual bandpass images of the entrance slit and transfer them to as small a detector as possible. One desires to keep the area of the detector at a minimum as the noise of many detectors increases with the square of the area. If the optics are able to illuminate a single detector element with all of the available light impinging upon the focal plane containing the spatially distributed images of the slit for each of the N bandpass resolution elements, a multitude of spectral resolution elements can be measured simultaneously using a single detector element. This arrangement results in a multiplexing spectrometer. The recovery of N spectral resolution elements requires measuring the detector response for N different encodements of $(N+1)/2$ open mask elements. The raw data are recorded as the detector response versus encodement number and is called an encodegram. Hadamard transformation of the encodegram yields the spectrum.

The History of Applied Hadamard Multiplexing

The Hadamard transform instruments developed in the 1960s and 1970s employed moving masks. Significant problems such as misalignment and jamming associated with a moving mask led to a reputation of poor reliability and contributed to a dormant period in the development of HTS and HTI. Interest was rekindled in the 1980s using stationary Hadamard encoding masks based on liquid crystal (LC) technology. The first generation 1D stationary Hadamard encoding mask was a cholesteric LC with $N = 127$ mask elements and used polarization as its operating phenomenon. Two parallel polarizers and rotation or lack of rotation of the polarized radiation generated the opaque and transparent states, respectively. The second generation 1D stationary Hadamard encoding mask was fabricated using a polymer dispersed liquid crystal (PDLC) material with $N = 255$ mask elements and used light scattering as its operating phenomenon. The PDLC contained LC droplets dispersed in a polymer matrix whose index of refraction matched the index of refraction in one direction in the birefringent LC droplet. Alignment of the LC droplets optical axis under an applied voltage removed discontinuities in index of refraction at the polymer matrix/LC interface to generate a good transparent state while random orientation of LC droplets in the polymer matrix generated the opaque state from light scattering by the discontinuities in index of refraction at the polymer matrix/LC droplet interface. A 2D stationary Hadamard encoding mask was also based on LC technology. A ferro-electric liquid crystal positioned between a pair of polarizers with perpendicular orientation operated as an electro-optic half-wave plate when a +value of applied voltage rotated the plane of polarization by 90 degrees to produce the transparent state and a −value of applied voltage left the plane of polarization unaltered to produce the opaque state.

Development based on stationary Hadamard encoding masks continued in the 1990s and a 2D moving Hadamard encoding mask was also fabricated and used to perform imaging in the near-infrared and mid-infrared spectral regions. Note that the mid-infrared spectral region is not generally accessible via Hadamard encoding masks based on

LC technology, since any LC material is expected to have strong absorption bands in the mid-infrared spectral region. From the mid-1990s to the present, the stationary Hadamard encoding mask of choice for the visible and near-infrared spectral regions has been the digital micromirror device (DMD), a device based on micro-optoelectromechanical systems (MOEMS) technology and developed by Texas Instruments for projector display applications. One DMD format incorporates 508 800 micromirrors in an 848 column by 600 row array that is 14.4 mm wide by 10.2 mm high. Each individual micromirror is 16 μm square and adjacent micromirrors are separated by a 1 μm gap. The micromirrors are individually addressable and rotate by +10 or −10 degrees about the diagonal axis to produce binary 'on' and 'off' states. The on state has T_t determined by the mirror reflectivity and approaches 1 while the off state approaches $T_o = 0$. The ideal condition of on and off is not realized due to diffraction of the light off of the small and periodic features of the micromirror device. The DMD is an array of spatial resolution elements that may be selected as groups of super-resolution elements or as individual resolution elements consisting of a single micromirror. The DMD resolution elements are realized as spectral resolution elements in the spectrometer with the columns attributed to the frequency or wavelength dimension and the rows attributed to the slit height dimension. The DMD resolution elements are utilized as spatio-spectral resolution elements in the imaging spectrograph with the columns as the frequency or wavelength dimension and the rows as a vertical spatial dimension with the horizontal spatial dimension being accessed, if desired, by translating the sample relative to the imaging spectrograph. The DMD resolution elements are spatial resolution elements in the imager with the columns for the horizontal dimension and the rows for the vertical dimension and the frequency or wavelength dimension provided by other instrumentation. If a photo-acoustic detection system is present then the depth dimension of the sample may also be accessed by changing the modulation frequency used in the photo-acoustic detection system.

Some important features of HTS to keep in mind are:

(i) it is a dispersive technique using a spectral separator;
(ii) it is a multiplexing technique using a single-element detector;
(iii) it uses a Hadamard encoding mask (multislit array) in a focal plane;
(iv) it sends one more than half the resolution elements to the single-element detector in each encodement;
(v) it uses a number of encodements equal to the number of resolution elements desired and the number of mask elements (pixels) in the stationary Hadamard encoding mask (a moving Hadamard encoding mask has $2N - 1$ mask elements);
(vi) it has each encodement containing a different combination of one more than half the resolution elements;
(vii) it has as its primary data the encodegram, a record of detector response versus encodement number; and
(viii) it uses a FHT of the encodegram to decode the encodegram and generate the spectrum or image.

Note that all of these, except for item (i), also apply to HTSI and HTI. The DMD promises to be the best Hadamard encoding mask yet developed for the visible and near-infrared spectral regions. However, its potential applications in spectrometry and imaging are by no means limited to HT techniques since the information corresponding to any micromirror in the DMD may be included in or excluded from any measurement as desired by the investigator. An instrument with no moving parts, other than the micromirrors in the DMD, promises to provide a compact and robust instrument for operation in potentially hostile environments ranging from process control to outer space. It is our belief that the combination of a DMD with a single-element detector may provide an important advance in spectroscopic instrumentation and that instrumentation based on the DMD may lead to a host of environmental, industrial, medical, and military applications.

See also

Fourier Optics. **Imaging:** Multiplex Imaging. **Optical Communication Systems:** Optical Time Division Multiplexing. **Spectroscopy:** Fourier Transform Spectroscopy.

Further Reading

Bellamy MK, Mortensen AN, Hammaker RM and Fateley WG (1997) Chemical mapping in the mid- and near-IR spectral regions by Hadamard transform/FT-IR spectrometry. *Applied Spectroscopy* 51: 477–486.

Fateley WG, Hammaker RM, Paukstelis JV, *et al.* (1993) Application of a two-dimensional Hadamard encoding mask for the imaging of thin-layer chromatography plates by laser-induced fluorescence or surface-enhanced Raman scattering and for use with a photoacoustic detector to generate three-dimensional photoacoustic images. *Applied Spectroscopy* 47: 1464–1470.

Fateley WG, Hammaker RM, DeVerse RA, Coifman RR and Geshwind FB (2002) The other spectroscopy: demonstration of a new de-dispersion imaging spectrograph. *Vibrational Spectroscopy* 29: 163–170.

Hammaker RM, Bohlke AP, Freeman RD, *et al.* (1996) Hadamard transform Raman spectrometry. In: Laserna JJ (ed.) *Modern Techniques in Raman Spectroscopy*, pp. 143–226. Chichester: Wiley.

Hammaker RM, DeVerse RA, Asunskis DJ and Fateley WG (2002) Hadamard transform near-infrared spectrometers. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, vol. 1, pp. 453–460. Chichester: Wiley.

Harwit M and Sloane NJA (1979) *Hadamard Transform Optics*. New York: Academic Press.

Nonlinear Laser Spectroscopy

P Ewart, University of Oxford, Oxford, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction: Linear and Nonlinear Spectroscopy

Light propagates in transparent or weakly absorbing media by inducing the atoms or molecules in the material to oscillate in response to the electromagnetic field of the light. The induced dipole oscillations set up a polarization field **P** which radiates the wave in the propagation direction. The induced polarization **P** is proportional to the amplitude of the electric field **E** in the incident light:

$$\mathbf{P} = \varepsilon_0 \chi^{(1)}(\omega)\mathbf{E} \qquad [1]$$

where ε_0 is the permittivity of free space and $\chi^{(1)}(\omega)$ is the first-order, frequency-dependent, susceptibility of the material. This interaction of the light with the medium generally results in some absorption of energy from the field and a shift in the phase of the propagating light. Both the phase delay and the absorption depend on the frequency ω (and wavelength λ) of the light. The phase shifting is responsible for dispersion – the variation, as a function of frequency, of the speed (phase velocity) of light in the medium. The dispersion and absorption effects are described by the real and imaginary parts, respectively, of the complex susceptibility:

$$\chi(\omega) = \chi_{\mathrm{Re}}(\omega) + \mathrm{i}\chi_{\mathrm{Im}}(\omega) \qquad [2]$$

This susceptibility parameter describes the optical properties of the medium. Linear spectroscopy uses the frequency-dependent property of the $\chi(\omega)$ to create spectra by absorption or dispersion of specific frequencies that are resonant with transitions between internal energy states of the molecules. The electric field in the incident light is considered to be weak when it leaves the atomic and molecular properties essentially unaffected. Nonlinear optical spectroscopy, however, is based on the ability of intense laser light to modify the optical properties of the medium through which it propagates. These effects are also frequency dependent and so can be used to generate spectra. When the strength of the electric field becomes comparable with the electric fields within the atoms or molecules (typically $\sim 10^{11}\ \mathrm{V\,m^{-1}}$) then the induced polarization becomes nonlinear and is described by a power series expansion:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega)\mathbf{E}_1 + \chi^{(2)}(\omega)\mathbf{E}_1\mathbf{E}_2 + \chi^{(3)}(\omega)\mathbf{E}_1\mathbf{E}_2\mathbf{E}_3 + \text{higher-order terms}] \qquad [3]$$

$\chi^{(2)}(\omega)$ and $\chi^{(3)}(\omega)$ are the second- and third-order susceptibilities respectively. The real and imaginary parts of these susceptibilities (and of the higher-order terms) describe observable effects arising from absorption and dispersion. The associated interactions involve two and three photon processes, respectively. In general, nonlinear interactions involve multiphoton processes and will be described by the appropriate term in the expansion of the induced polarization.

Broadly speaking, nonlinear spectroscopy is based on detection of these multiphoton interactions by either incoherent or coherent processes. Incoherent processes include fluorescence and ionization where, following the multiphoton interaction, photons or electrons respectively are emitted spontaneously in all directions. Coherent processes, on the other hand, arise from the coupling of the input driving fields via the nonlinear polarization of the medium.

The waveform of the polarization oscillations is a distorted form of the driving field. Fourier analysis of these nonlinear oscillations would reveal extra frequencies not present in the incident field. When a suitably phased array of dipoles is established, the nonlinear optical process is detected by emission of coherent directional light. The efficiency with which this coherent signal beam is generated is crucially dependent upon the correct phasing of the radiating molecules. This process is known as phase matching and is achieved by geometrical arrangements of the input beams or by modifying the refractive index to counteract dispersion between the generating polarization wave and the radiated signal wave.

Apart from the general division of nonlinear spectroscopy according to whether the detection uses incoherent or coherent emission, the techniques are characterized as frequency or time domain methods, according to how the signal is generated and detected. In most cases the signals recorded in either domain may be related by a Fourier transformation to an equivalent signal in the other.

The advantages of nonlinear spectroscopic techniques lie in their ability to provide high spectral, temporal, and spatial resolution. In particular, it is possible, using a variety of nonlinear optical processes, to eliminate inhomogeneous broadening of the spectral lines recorded. Second, when coherent signal beams are generated they can be readily distinguished from incoherent background emissions, thus permitting high signal-to-noise ratios. Third, the use of short laser pulses, down to the order of 10^{-15} s, allows high temporal resolution. High spatial resolution is achieved by generating signals from interaction regions defined by crossed and focused beams. Although the coefficients of successive terms in the power series expansion of the nonlinear polarization decrease by many orders of magnitude with increasing order, signal strengths may be made comparable to those of linear spectroscopy. Strong signals are obtained by the use of resonant enhancement of the nonlinear susceptibility, high field strengths from laser sources, and the use of sensitive photodetectors.

Second-Order Nonlinear Effects

Weak light fields leave unaffected the optical properties of the medium through which they propagate. When, however, the second-order susceptibility $\chi^{(2)}$ contributes significantly to the induced polarization, the complex refractive index includes a field-dependent term:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(2)}(\omega)\mathbf{E}_1]\mathbf{E}_2 \qquad [4]$$

Owing to the symmetry properties of $\chi^{(2)}$, second-order nonlinear effects are absent in isotropic media or those possessing a center of inversion symmetry. They may, however, be observed in anisotropic materials such as crystalline solids where they are the basis of frequency mixing schemes leading to second-harmonic generation, sum- and difference-frequency generation, and optical parametric amplification and oscillation. The symmetry of gas phase media, such as atomic vapors, may be broken by application of an external electric or magnetic field. Resonantly enhanced second-harmonic generation has been achieved in alkali metal vapors in the presence of a magnetic field giving potential for spectroscopy. Symmetry breaking is achieved also on solid surfaces where sum-frequency generation has been used to study surface morphology, monatomic layers, and catalytic effects of adsorbed materials.

Third-Order Nonlinear Spectroscopy

In gas phase or isotropic solid and liquid media, the first nonlinear term is of third order and leads to a refractive index that is dependent on the field intensity $|\mathbf{E}|^2$:

$$\mathbf{P} = \varepsilon_0[\chi^{(1)}(\omega) + \chi^{(3)}(\omega)\mathbf{E}^2]\mathbf{E} \qquad [5]$$

Intensity-dependent refractive index effects lead to a wide range of nonlinear optical phenomena. In general, the polarization contains terms that are the product of three input fields generating Fourier components at sum and difference frequencies. Phase matching constraints will generally determine which of the possible interactions will be observed. A wide range of physical effects contribute to $\chi^{(3)}(\omega)$. The simplest of these nonlinear effects is saturation of the absorption.

Linear absorption is defined by the situation where the intensity decays exponentially with propagation distance x, according to Beer's Law:

$$I(x) = I(0)\exp\{-\alpha_0(\omega)x\} \qquad [6]$$

where $I(0)$ and $I(x)$ are the intensity values at $x = 0$ and x, respectively, and $\alpha_0(\omega)$ is the weak field absorption coefficient. The fractional power absorbed in this case is, therefore:

$$[I(0) - I(x)]/I(0) = \alpha_0(\omega)x \qquad [7]$$

and is independent of the intensity. When the incident intensity increases, until the rate of stimulated absorption from the lower level equals the spontaneous emission rate out of the upper level, the medium is said to be saturated. The intensity is then I_{sat}. A simple rate-equation analysis shows that the absorption coefficient in the presence of a strong light field is:

$$\alpha(\omega) = \frac{\alpha_0(\omega)}{[1 + I/I_{sat}]} \qquad [8]$$

The absorption is then no longer independent of the light intensity.

Saturation Spectroscopy

In the case of a homogeneously broadened transition the effect of a saturating monochromatic light field is to reduce the absorption across the entire line profile. This is shown in Figure 1a where the frequency of the incident light w_L is detuned from line center ω_0. The degree of saturation and the effect on the absorption coefficient depends on the intensity I and the detuning $\Delta\omega = \omega_L - \omega_0$.

For a transition dominated by inhomogeneous broadening a monochromatic light field interacts strongly only with a sub-group of atoms. This subgroup is determined by those having a resonance frequency within the spectral range covered by the homogeneous width and centered on the incident frequency. Saturation in this case results in 'hole burning' as shown in Figure 1b. A dip in the absorption coefficient $\alpha(\omega)$ is produced centered on ω_L of width given by the homogeneous width.

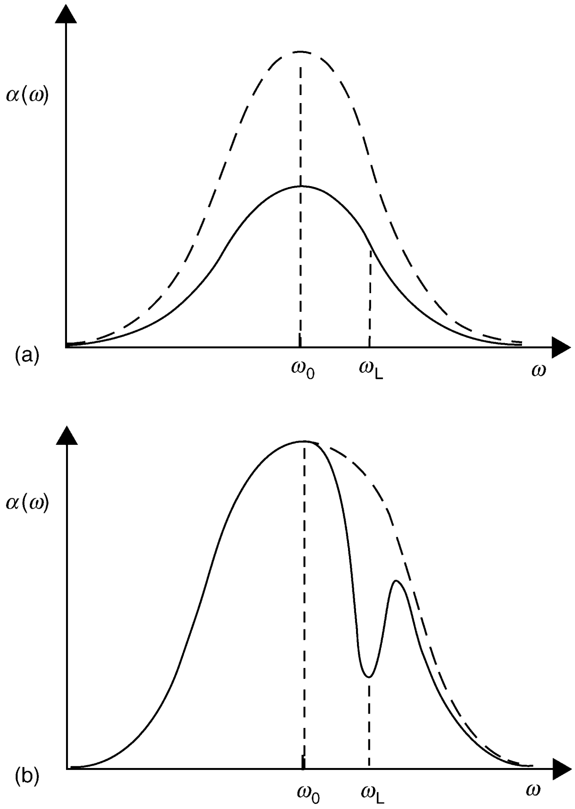

Figure 1 Absorption lineshapes for (a) homogeneously broadened and (b) inhomogeneously broadened lines centered at ω_0. Dashed lines in both figures show lineshape measured by weak probe. Solid lines show lineshape measured by weak probe in presence of a monochromatic, saturating beam at frequency $\omega_L \neq \omega_0$.

Saturation affects the population of both upper and lower levels involved in an optical transition and so affects both absorption and amplification (gain). The phenomenon shows itself as a narrow dip in the spectral profile of lasers operating on an inhomogeneously broadened gain line (e.g., a HeNe gas laser). These 'Lamb dips', as they are called, are the result of saturation by the laser mode which has a very narrow spectral width. Only those atoms in the Maxwell–Boltzmann velocity distribution are saturated, whose velocity results in a Doppler shift of the light that brings it into resonance, i.e., to within the atom's homogeneous linewidth of the resonance frequency. When the laser mode frequency is detuned from atomic line center two velocity subgroups of atoms, within the distribution, are saturated corresponding to positive and negative Doppler shifts for counter-propagating fields of the mode. When the mode frequency is tuned to atomic line center, the counter-propagating fields interact with the same (zero) velocity subgroup. The saturation effect is thereby enhanced, reducing the gain by depletion of the population inversion, leading to a dip in output intensity at line center. The width of the dip is the homogeneous width of the transition (Figure 2).

The variation of output intensity around the Lamb dip is used to provide an error signal in one scheme for laser frequency stabilization. A feedback system is employed to adjust the cavity length, L, in order to maintain the mode frequency, ν, exactly on atomic line-center ($\nu = mc/2L$ where m is an integer). This scheme uses gain saturation of the amplifying medium, whereas an alternative uses saturated absorption. In this scheme, a cell of gas having a narrow absorption line, within the gain profile of the laser, is placed inside the cavity. When the laser mode is tuned to absorption, line-center saturation reduces the absorption loss, resulting in a narrow peak in the laser emission. Drifting off this peak can be used as an error signal, as before, to stabilize the mode frequency.

Saturated absorption spectroscopy uses the same mechanism by placing the absorbing specimen outside the laser cavity. The effect is detected by monitoring the transmission of a weak probe beam in the presence of a strong counter-propagating 'pump' beam. Both probe and pump beams are

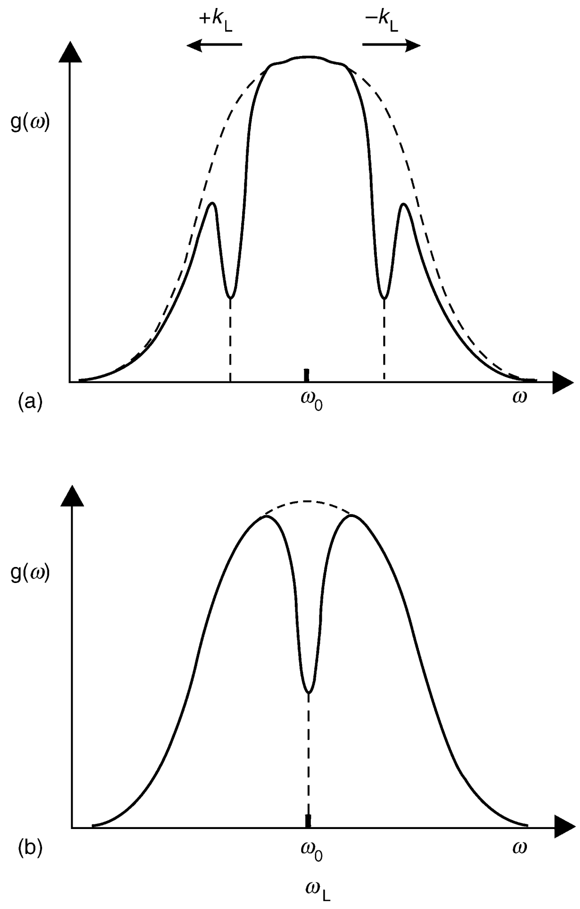

Figure 2 (a) Holes burned in gain profile by two counter-propagating monochromatic waves at $\omega_L \neq \omega_0$. (b) 'Lamb dip' in gain profile $g(\omega)$ (or absorption profile $\alpha(\omega)$) when $\omega_L = \omega_0$.

derived from the same laser using a suitable beamsplitter and, thus, have identical frequency. When tuned to line-center of the absorbing sample, the transmission of the probe increases owing to saturation by the strong pump beam interacting with the same subgroup of the inhomogeneously broadened medium (Figure 3).

The elimination of the first-order Doppler effect, in gaseous media, using this technique results in very high-resolution spectra. The residual linewidth is determined by homogeneous processes such as lifetime broadening, collisional (or pressure) broadening and power (or saturation) broadening. The latter effect is inevitable, as saturation is necessary, but may be controlled so as to contribute very little to the overall broadening. Power broadening becomes significant only when the laser intensity is appreciably in excess of the saturation intensity. Another limiting effect is transit time broadening: moving atoms spend only a limited time interacting with the narrow beams as they pass through. The consequent limit on the interaction time leads to conjugate broadening of the frequency of the transition as a result of the uncertainty principle.

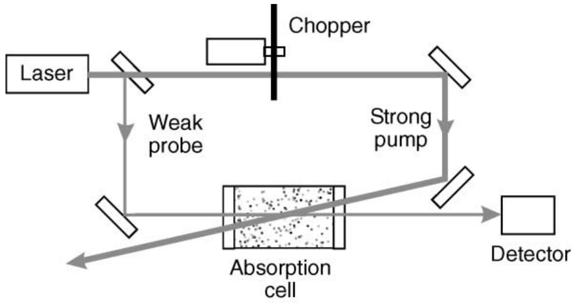

Figure 3 Schematic diagram of saturated absorption spectroscopy experiment. The saturating pump beam is chopped or modulated and the signal is processed using phase-sensitive detection to improve the signal-to-noise ratio.

The general problem of the spectral lineshape of signals in saturation spectroscopy is complex. Apart from the effects mentioned above, the lineshape will be affected by relativistic effects (second-order Doppler shifts), velocity-changing collisions, recoil effects from photon absorption and emission, velocity-dependent transit time broadening, optical pumping and laser lineshape effects.

Saturation spectroscopy is widely used in high-resolution spectroscopy for measurement of atomic and molecular transition frequencies, time and frequency standards using stabilized lasers, measurement of fundamental constants such as the Rydberg constant and the fine structure constant. Dynamical effects such as atomic collisions are also studied using saturation, whereby the relaxation of hole burning is measured in the time or frequency domain.

Polarization Spectroscopy

An important variation of saturation spectroscopy uses the change induced by a saturating beam in the real part of the complex refractive index, i.e., dispersion rather than absorption. Specifically, the technique of polarization spectroscopy employs a saturating beam that is circularly polarized to effect optical pumping on the degenerate lower level of a transition. The degenerate states may be identified, for example, with magnetic substates based on orientation of the intrinsic atomic or molecular angular momentum and labeled by the quantum numbers m_J. In equilibrium, the populations of all these degenerate substates will be equal and the optical properties of the system will then be isotropic. In particular, the polarization state of a linearly polarized probe beam will be unaffected by passage through the medium. This probe is totally rejected by an analyzer (a linear polarizer set to transmit only light in an orthogonal plane). The linear polarized beam is a linear combination of two oppositely circular polarization states, σ^+ and σ^-, and both components experience the same complex refractive index. In the presence of a circularly polarized saturating pump beam, the distribution of population amongst the degenerate substates is disturbed from its equilibrium by optical pumping. The result will be an excess of population in higher or lower m_J states (a condition known as orientation) and this leads to optical anisotropy in the medium. The refractive index for σ^+ light is now different from that for σ^- light and a relative phase shift is introduced between these components of the linearly polarized probe resulting in elliptical polarization. A component of this elliptically polarized light will then be transmitted by the analyzer and is detected. When the strong pump beam is closely counter-propagating to the probe, then both beams interact with the same velocity subgroup as in Doppler-free saturation spectroscopy. The inhomogeneous (Doppler) broadening is thus significantly reduced. The technique provides a signal above an essentially 'zero' background and is thus capable of a higher signal-to-noise ratio than saturated absorption spectroscopy. How close to 'zero' the background is depends crucially on the degree of extinction by the analyzer. High-quality polarizer/analyzer devices can achieve extinction ratios of 1:10^{-8}. Stress-induced birefringence in the windows of the cell containing the sample is a common limiting factor (Figure 4).

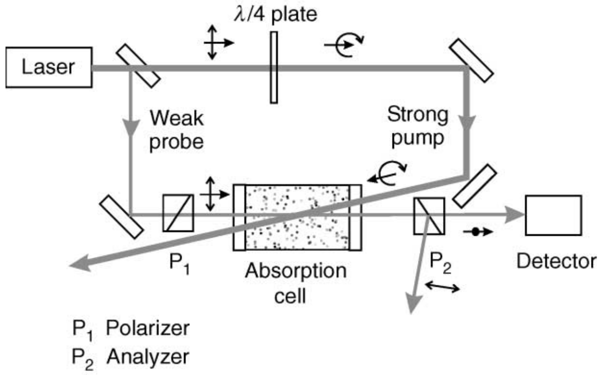

Figure 4 Schematic diagram of polarization spectroscopy experiment. Circularly polarized pump beam may be modulated for phase-sensitive detection of transmitted probe beam.

A probe laser of intensity I_{probe} is transmitted through a sample of length l with an intensity $I_{\text{PS}}(\omega)$ given by the following expression, that takes account of different contributions to the detected signal:

$$I_{\text{PS}}(\omega) = I_{\text{probe}}\left[\xi + \varphi^2 + b^2 + \frac{1}{2}\frac{b\Delta\alpha_\pm l}{(1+x^2)} + \frac{1}{2}\varphi\Delta\alpha_\pm l\frac{x}{(1+x^2)} + \left(\frac{1}{4}\Delta\alpha_\pm l\right)^2\frac{1}{(1+x)^2}\right] \quad [9]$$

where $x = 2(\omega_0 - \omega)/\gamma$ is the detuning from the resonance frequency ω_0 normalized by the halfwidth γ of the collision broadened line. The frequency-dependent terms provide absorptive (Lorentzian) and dispersive contributions to the total lineshape. The first three, frequency independent, terms in eqn [9] are responsible for a constant background signal. They originate from residual transmission of the polarizers (ξ), their accidental uncrossing (φ), and a possible residual birefringence (b) from optical elements between both polarizers in the probe beam path. ($\xi + b^2$), normally is in the range 10^{-6}. The differential absorption coefficient for left and right circularly polarized light, $\Delta\alpha_\pm$, at resonance ($x = 0$), is given by:

$$\Delta\alpha_\pm(x=0) = \Delta\alpha_\pm^0 = \alpha^+ - \alpha^- = \alpha_0 S_0 \Delta C_{JJ_1} \quad [10]$$

where α_0 is the unsaturated absorption coefficient, S_0 is the saturation parameter for the pump wave ($I_{\text{pump}}/I_{\text{sat}}$), and ΔC_{JJ_1} are Clebsch–Gordan coefficients for the respective transition and coupling case.

The dispersion-shaped component may be made to dominate by deliberately uncrossing the polarizer/analyzer combination to a small extent. This technique enhances the resolution of the transition frequency. Linewidths are limited by residual Doppler broadening owing to finite crossing angles of pump and probe beams, collisional broadening, transit time broadening, etc.

Applications of the technique have included Doppler-free spectroscopy of atomic transitions for precision measurements of fundamental constants, molecular spectroscopy and detection of atomic, molecular, and radical species in flames. By employing a different laser to accomplish the induced optical anisotropy from that used to probe the medium, molecular spectra can be simplified by the technique of 'polarization labeling'. Signals are obtained only when the probe and pump excite transitions sharing a common level and the particular transitions observed are sensitive to the selection rules for ΔJ, governing both pump and probe transitions. In this way P- and R-branch transitions can be distinguished.

Advantages of polarization spectroscopy include high sensitivity (10^2–10^3 times that of saturation spectroscopy), high signal-to-noise, simplification of molecular spectra by polarization labeling, dispersion profiles for enhanced resolution, and for frequency stabilization techniques.

Multiphoton Absorption Spectroscopy

Multiphoton absorption is the term used to describe a process in which an atom or molecule makes a single transition between two of its allowed energy levels by absorbing the energy from more than a single photon. Sequential transitions, by which successive photons are absorbed via single-photon transitions between stationary energy levels, take place provided photons arrive within the lifetime of the intermediate state. Multiphoton transitions take place via 'virtual' energy levels. The lifetime of a virtual level is determined by the uncertainty principle to be the inverse frequency detuning from the nearest stationary energy level. Thus a high photon flux is usually required for multiphoton effects to be observed (Figure 5).

The simplest multiphoton effect, two-photon absorption, may be observed using strong atomic or molecular electronic transitions, even with modest laser power. The absorption is detected by observing the fluorescence from the spontaneous, radiative decay of the excited state. Two-photon absorption spectroscopy may be made Doppler-free by arranging absorption of one photon from each of two counterpropagating beams. The moving atoms experience opposite Doppler shifts $\mathbf{k}\cdot\mathbf{v}$ from each beam. The transition energy is thus:

$$E = (\hbar\omega + \mathbf{k}\cdot\mathbf{v}) + (\hbar\omega - \mathbf{k}\cdot\mathbf{v}) = 2\hbar\omega \qquad [11]$$

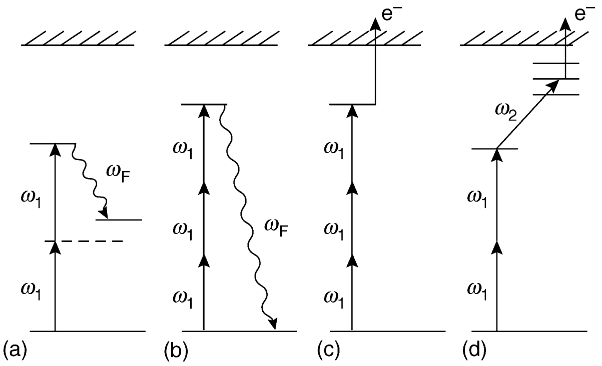

Figure 5 Multiphoton absorption spectroscopy. (a) Two-photon absorption using fluorescence detection of radiative decay at ω_F. (b)Three-photon absorption spectroscopy using fluorescence detection of radiative decay at ω_F. (c) Three-photon absorption spectroscopy using ionization detection following photoionization, collisional or field ionization. (d) Two-photon resonant multiphoton ionization spectroscopy; laser at frequency ω_1 is fixed and laser at ω_2 is tuned over transitions to high-lying levels.

Absorption of two photons from a single beam is prevented by choosing a transition between levels of the same parity and angular momentum quantum number, i.e., $\Delta J = 0$ and using circularly polarized light. Absorption of two σ^+ or σ^- photons gives $\Delta J = \pm 2$, which would be forbidden, e.g., on a $n^2S - n'^2S$ transition in an alkali metal. This method of Doppler-free spectroscopy has the advantage that it involves all the atoms and not just the zero-velocity subgroup, as in saturation or polarization spectroscopy. In principle, it is also possible to use three-photon absorption, with the momentum of the wavevectors arranged to sum to zero, to achieve Doppler-free absorption on a transition between opposite parity states.

Instead of observing fluorescence following multiphoton absorption, the excited atom or molecule may be ionized by atomic collisions or by applying a d.c. electric field. Excitation of Rydberg levels (i.e., having high value of the principal quantum number n) by two- or three-photon absorption is readily accomplished. The sensitivity of this technique is enhanced by using two lasers. One laser at frequency ω_1 is tuned to a two-photon transition (energy $= 2\hbar\omega_1$) and the second laser ω_2 is tunable to excite transitions from the two-photon intermediate state to higher-lying levels (energy $= 2\hbar\omega_1 + \hbar\omega_2$). Thus, energy levels normally requiring excitation in the vacuum UV may be reached using visible frequency lasers. In such cases ionization detection may be more efficient than fluorescence since the nonradiative decay rate of these highly excited levels may exceed the fluorescence decay rate. The ion signal produced may be significantly amplified by the process of space charge amplification. A weak discharge established in the atomic vapor creates a space charge around the cathode. A single ion moving slowly through the space charge cloud acts to cancel the local field allowing many (up to 10^6) electrons to escape to the anode with a resulting amplification of 10^6 in the 'photoionization current'.

Multiphoton ionization spectroscopy is a sensitive method of recording highly excited energy levels. The sensitivity is enhanced in the technique of resonance ionization spectroscopy (RIS): two lasers are used such that the frequency of one laser is fixed close to a two-photon allowed transition and the second is tuned to excite a range of higher excited states. Absorption of an even number of photons allows excitation of levels having the same parity as the ground state and thus not accessible to normal absorption spectroscopy. In combination with a mass spectrometer to detect the ions produced by the multiphoton absorption resonance ionization mass spectrometry (RIMS), affords additional species selectivity and sensitivity. RIMS finds application in high-resolution spectroscopy of rare isotopic species and is the basis for a scheme for isotope separation.

Coherent Third-Order Processes

An important class of spectroscopic techniques exploits the coherent interaction of three input waves via the nonlinear response of the medium. Specifically, the third-order term has the form:

$$\mathrm{P}_i^{(3)}(\omega_4,\vec{r}) = \chi_{ijkl}^{(3)}(\omega_1,\omega_2,\omega_3,\omega_4)\mathrm{E}_j(\omega_1,\vec{r}) \times \mathrm{E}_k(\omega_2,\vec{r})\mathrm{E}_l(\omega_3,\vec{r}) \qquad [12]$$

where $\mathrm{E}_m(\omega_n)$ are complex field amplitudes and the subscript $m = i, j, k, l$ refers to each of the Cartesian coordinates x, y, z of the electric vector. $\chi_{ijkl}^{(3)}$ is a fourth-rank tensor. In condensed phase media, the third-order nonlinearity is readily observed, even with modest laser intensities. In the gas phase, although $\chi_{ijkl}^{(3)}(\omega)$ is small relative to $\chi^{(1)}$, its effect is enhanced when the incident fields are resonant and this nonlinear response may be used for spectroscopy. Equation [12] shows that the induced polarization couples three waves to radiate a fourth and describes a class of interactions known collectively as four-wave mixing.

The physics of four-wave mixing interactions may be visualized as the formation and read-out of a volume hologram. The three input beams are denoted as two pumps and one probe. The process involves the interference of one pump with the probe to form a spatially periodic grating pattern in the electric field distribution. The nonlinear response of the medium via $\chi^{(3)}$ leads to a spatial modulation of the refractive index – a laser-induced grating. The second pump beam generates a signal by scattering off this grating in a manner analogous to Bragg scattering of X-rays from regular crystal planes. The condition for generating a coherent scattered beam is equivalent to matching the momentum (or wavevectors) of the interacting photons, i.e., phase matching.

Degenerate Four-Wave Mixing (DFWM) Spectroscopy

The simplest case of four-wave mixing spectroscopy involves pump and probe beams of identical (i.e., degenerate) frequency to generate a signal at the same frequency. Spectra are generated as the frequency is tuned through atomic or molecular resonances where the interaction is strongly enhanced. In this degenerate case, the pump and probe interference pattern is stationary and the remaining pump beam scatters off this grating with the same frequency. As with all parametric processes, the interaction is phase matched by a suitable geometrical arrangement of the beams. A probe intersecting with counter-propagating pump beams leads to a signal that counter-propagates with the probe. This arrangement provides ‘Doppler-free’ lineshapes, since only those atoms with essentially zero velocity along the line of intersection of the beams contribute to the signal. Alternatively, pump and probe beams may be arranged to cross at an angle (usually small) in the forward direction – the ‘forward geometry’, with the resulting signal propagating at the same small angle to the incoming beams. Signals in this case are stronger, since all velocity classes may contribute but the lineshape is Doppler broadened (Figure 6).

An additional feature of the counter-propagating pump geometry is that the phase-matched signal arises from terms that involve products of fields with the complex conjugate of the incident

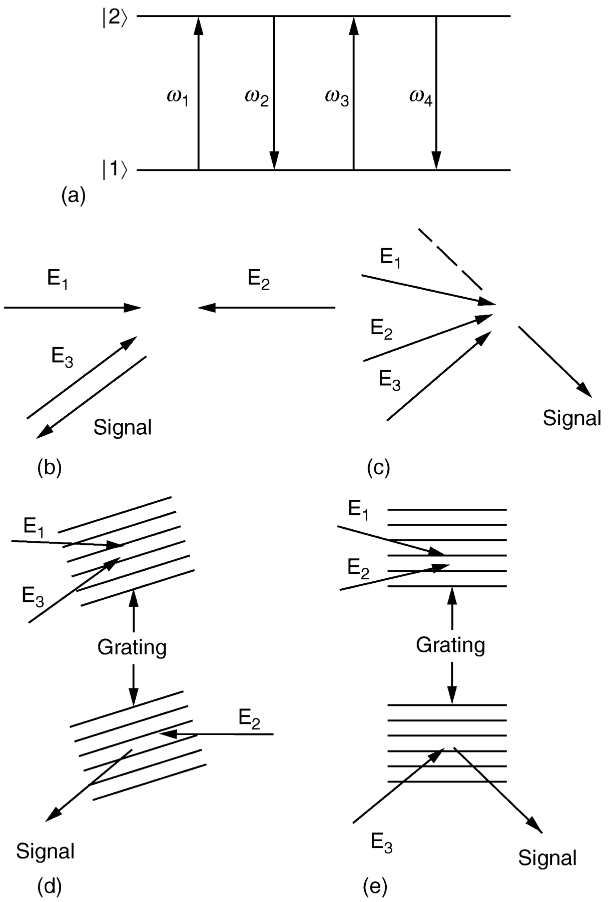

Figure 6 Degenerate four-wave mixing: (a) all input frequencies are degenerate, $\omega_1 = \omega_2 = \omega_3$ and give signal at the same frequency, ω_4. Phase matching geometries are shown in (b) counter-propagating pump (phase conjugate) and (c) forward geometry. (d) and (e) show the grating formation and read-out respectively.

probe field. The induced polarization radiating the signal wave has a component given by:

$$\mathbf{P}^{(3)}_{\text{signal}}(\omega) = \chi^{(3)}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}_{\text{pump}}(\omega)\mathbf{E}^{*}_{\text{probe}}(\omega) \qquad [13]$$

The backward propagating signal appears as a phase conjugated reflection of the incident probe wave. A probe wave that has propagated through a phase-distorting medium thus generates a signal with an initial phase distortion that will be undone as it retraces the path through the distorting medium. This phase conjugating property finds application in adaptive optics but also lends an advantage when DFWM is used as a spectroscopic probe in hostile or distorting environments, such as flames or plasmas. The counter-propagating pump geometry is often referred to as the phase conjugating geometry.

The intensity of the DFWM signal beam may be calculated using perturbation theory to derive $\chi^{(3)}(\omega)$ arising from saturated absorption. In the perturbation approximation, the probe beam is assumed to be weak and interacts with stationary atoms in the presence of counter-propagating, saturating pump fields. The signal is found from the phase-matched terms in the wave radiated from the induced polarization. When integrated over the atomic resonance line, the signal intensity, $I^{\text{int}}_{\text{sig}}$, is given by analytic expressions in the two limiting regimes of weak and strong saturation by the pump:

$$I^{\text{int}}_{\text{sig}} = \frac{|\mu_{12}|^8 N^2 k^2 T_1^2 T_2^3}{[1 + 4I_{\text{pump}}/I_{\text{sat}}(0)]^{5/2}} I^2_{\text{pump}} I_{\text{probe}} \qquad [14]$$

when $I_{\text{pump}} \ll I_{\text{sat}}(0)$ and

$$I^{\text{int}}_{\text{sig}} \propto |\mu_{12}|^3 N^2 k^2 T_1^{-1/2} T_2^{1/2} I^{-1/2}_{\text{pump}} I_{\text{probe}} \qquad [15]$$

when $I_{\text{pump}} \gg I_{\text{sat}}(0)$.

In these relations μ_{12} is the atomic dipole moment, N the atomic number density, $I_S(0) = \hbar^2/T_1T_2|\mu_{12}|^2$ is the line-center saturation intensity, and T_1 and T_2 are the longitudinal and transverse relaxation times, respectively. Extensions of the basic theory accommodate the effects of moving absorbers (Doppler effect) in both the phase-conjugate and forward phase matching geometry.

These results are derived using perturbation theory, by assuming that the probe field is weak. In practice the signal is optimized when pump and probe are of equal intensity and approximately equal to the saturation intensity. This situation violates the perturbation theory assumption but the signal intensity may be derived using nonperturbative analytical methods, although the resulting expression is more complicated. Alternatively, the quantum mechanical equations for the nonlinear susceptibility can be solved numerically to find the signal level for arbitrary intensity of the incident fields. Optimum signal-to-noise ratio in the spectra is achieved using approximately saturating intensities.

The sensitivity of resonant DFWM and the coherent nature of the signal beam suits the process to spectroscopy of nonfluorescing or trace molecular species, such as radicals created in chemical reactions. Equation [15] shows that the DFWM signal becomes relatively insensitive to collisional effects when the incident lasers saturate the transition. This feature of DFWM gives it an additional advantage over laser-induced fluorescence in optical diagnostic applications where quantitative evaluation of fluorescence intensity is hampered by collisional quenching. The state and species selectivity afforded by resonant DFWM is exploited in diagnostics of combustion and plasma processes, where the reduction of Doppler broadening allows high-resolution spectra from high-temperature gases. The relative intensity of spectral lines is used to derive temperature and concentration information with time and space resolution.

Additional selectivity is provided by using two input frequencies in two-color four-wave mixing (TC-FWM), to probe molecular excited state dynamics. The TC-FWM process uses resonant enhancement on two coupled transitions, i.e., the pump beams excite one transition and the probe and signal are generated on another transition having a common level with that excited by the pumps.

Coherent Raman Spectroscopy

The classical theory of spontaneous Raman scattering shows that a nonresonant pump of frequency ω_{p} wave can excite a resonance ω_{m} in a molecule with a nonlinear polarizability $\beta(x)$. (x is the spatial coordinate of the internuclear potential.) The wave resulting from this inelastic scattering is a Stokes wave of frequency $\omega_{\text{S}} = \omega_{\text{p}} - \omega_{\text{m}}$. The propagation of the Stokes wave is governed by the complex Raman susceptibility (Figure 7a):

$$\chi_{\text{Raman}}(\omega_{\text{s}}) = \chi'_{\text{Raman}}(\omega_{\text{s}}) + \mathrm{i}\chi''_{\text{Raman}}(\omega_{\text{s}}) \qquad [16]$$

This generated Stokes wave may then couple with the pump to drive the molecular resonance coherently leading to amplification of the Stokes wave at the expense of the pump. This process of stimulated Raman scattering (SRS), provides gain if the amplification exceeds the loss due to absorption, scattering, etc. There is, therefore, a definite threshold intensity that the pump must satisfy to observe SRS. The SRS

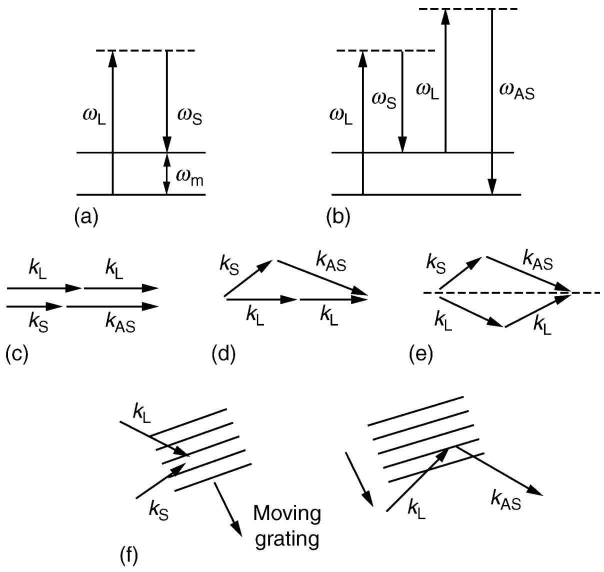

Figure 7 Coherent Raman spectroscopy. (a) Spontaneous Raman scattering of input pump ω_L producing Stokes wave ω_S in inelastic scattering. (b) Coherent anti-Stokes Raman scattering (CARS). Phase matching geometries for CARS is shown for a nondispersive medium in (c) and dispersive media in (d) and (e). The planar boxcars geometry in (e) may be folded along the dotted line (folded boxcars geometry) to separate the signal into an orthogonal plane from the pump beams to reduce scattered pump light reaching the detector.

gain $g(\omega_s)$ at the Stokes frequency ω_s is given by:

$$g(\omega_s) = -\frac{k_s}{2n_s^2}|E(\omega_L)|^2\chi''_{\text{Raman}}(\omega_s) \qquad [17]$$

where k_s and n_s are the wavevector and refractive index for the Stokes wave, $E(\omega_L)$ is the field amplitude at the pump laser frequency ω_L, and $\chi''_{\text{Raman}}(\omega_s)$ is the imaginary part of the complex Raman susceptibility:

$$\chi''_{\text{Raman}}(\omega_s) = \frac{-\varepsilon_0 N\left(\dfrac{\partial\beta}{\partial x}\right)^2\gamma}{32m_e\omega_v\left\{(\omega_m-(\omega_L-\omega_S))^2+\dfrac{\gamma^2}{4}\right\}} \qquad [18]$$

In this expression $\partial\beta/\partial x$ is the variation of polarizability β with internuclear displacement x, N is the molecular population density, and γ is the linewidth due to damping. Spectroscopy based on SRS provides signals that are many orders of magnitude stronger than those from spontaneous Raman scattering, and is widely used to study spectra of complex molecules having no absorption spectra in the visible region.

A four-wave mixing interaction, involving a pump and probe beam of different frequencies, generates a nonstationary grating. The signal generated by scattering of a second pump beam off this moving grating is Doppler-shifted to lower or higher frequency. These signal beams correspond respectively to Stokes or anti-Stokes waves in Raman scattering. A probe beam at the Stokes frequency will thus interact with a pump beam to generate a coherent signal at the anti-Stokes frequency; a process known as coherent anti-Stokes Raman scattering (Figure 7b). A probe at the anti-Stokes frequency correspondingly gives the process of coherent Stokes Raman scattering (CSRS). Phase matching is necessary for both CARS and CSRS and is satisfied by an appropriate arrangement of the incident wavevectors (Figure 7c,d and e).

CARS has proven to be a valuable spectroscopic technique in a wide variety of applications. A CARS spectrum is generated by scanning the frequency of the probe, i.e., the Stokes laser. The signal strength is given by

$$I_4 = I_{\text{CARS}} = \frac{\omega_4^2}{n_1^2 n_2 n_4 c^4 \varepsilon_0^2} I_1^2 I_2 |\chi_{\text{CARS}}|^2 l^2 \times\left(\frac{\sin(\Delta k l/2)}{\Delta k l/2}\right)^2 \qquad [19]$$

where I_1 and I_2 are the intensity of the pump and Stokes beams, respectively, l is the interaction length, and n_i are the refractive indices at frequency ω_i. The maximum signal is generated when the phase mismatch $\Delta k = 0$. The third-order susceptibility, $\chi^{(3)} = \chi_{\text{CARS}}$, in general consists of a nonresonant part $\chi_{\text{nr}}^{(3)}$, which is independent of the frequency of the exciting beams, and a Raman-resonant contribution $\chi_{\text{r}}^{(3)}$:

$$\chi^{(3)} = \chi_{\text{nr}}^{(3)} + \chi_{\text{r}}^{(3)} \qquad [20]$$

The lineshape in coherent Raman spectroscopy is affected by interference between resonant and nonresonant contributions. This effect is exploited to derive the concentration of the resonant (Raman active) species relative to that of the background of known density. A more common application of CARS, however, is for temperature measurement. The dependence of the signal intensity on $|\chi_{\text{CARS}}|^2$, and therefore on the square of the molecular population of the initial state, yields a spectrum from which the temperature may be derived by assuming a Maxwell–Boltzmann distribution over the state populations.

A particularly useful application of CARS employs a narrow linewidth pump laser and a broad bandwidth Stokes laser. CARS signals are generated simultaneously on all frequencies such that the corresponding Stokes frequency is available from the broadband probe. The spectrum recorded with a single laser pulse can thus provide population information on a wide range of molecular states, giving time- and space-resolved measurement of temperature and concentration.

CARS and CSRS spectroscopy is applied to both gas and condensed phase media. Owing to the higher density in the latter, signals may be strong in spite of the two-photon nature of the resonant interaction.

Time Domain Four-Wave Mixing Spectroscopy

Four-wave mixing spectroscopy may also be executed in the time domain. The three incident waves may arrive as pulses separated in time by variable delays. The interaction is, therefore, mediated by the atom or molecule's induced polarization or coherence. DFWM, CARS, and CSRS processes may thus be used to study ultrafast relaxation since the signal intensity gives a measure of the coherence decay in the intervening delay between the pulses. These processes are intimately related to coherent transient spectroscopy using measurements of free induction decay, photon echoes, and stimulated photon echoes.

Laser-Induced Grating Spectroscopy

The four-wave mixing processes discussed above involve a spatial modulation of the molecular population and a consequent modulation of the refractive index for resonantly interacting waves. The signals are radiated by the coherent oscillations of the molecular populations that decay rapidly. Relaxation processes result in transfer of energy from the excited molecular states, leading to a temperature and density grating in the bulk medium with a period Λ, given by:

$$\Lambda = \lambda_{\text{pump}}/(2 \sin \theta_1) \qquad [21]$$

where θ_1 is the angle between the interfering beams of wavelength λ_{pump}. A nonresonant wave of wavelength λ_{probe} may be scattered by this laser-induced grating provided it is incident on the grating at the angle θ_2 satisfying the Bragg condition:

$$\theta_2 = \sin^{-1}(\lambda_{\text{probe}}/2\Lambda) \qquad [22]$$

The periodic temperature/density grating, when induced by a short duration laser pulse, is accompanied by two oppositely propagating acoustic waves. The reflectivity of Bragg scattering is temporally modulated as the acoustic waves traverse the stationary grating at the local sound speed. Measurement of the period of these oscillations yields the sound speed from which the temperature may be derived. The stationary grating and acoustic waves decay by diffusive and damping processes, dependent upon the bulk medium properties. Measurement of the decay time of the grating induced in gas phase media yields the gas pressure if the gas dynamic parameters are known (Figure 8).

Laser-induced grating spectroscopy (LIGS), has been widely employed for studies of condensed phase relaxation processes. Applications have also been developed for gas phase studies using resonant absorption to enhance the thermal grating component. A grating may, however, also be induced in response to the electrostrictive effect of the high electric field in the interference pattern of two intense pump waves. Such electrostrictive gratings do not require resonant absorption and their spectroscopic utility centers on the study of relaxation effects in the molecular medium.

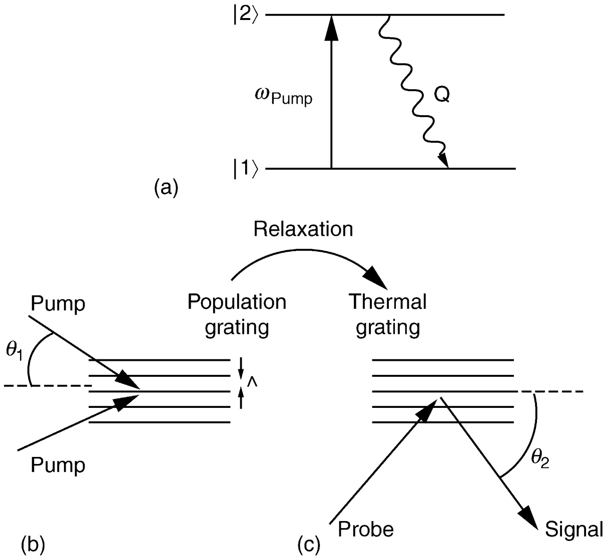

Figure 8 Laser-induced thermal grating spectroscopy. (a) Absorption of energy from pump lasers creates excited molecular population in grating pattern shown in (b). Spacing of laser induced grating Λ is set by angle of intersection of pumps. Quenching, *Q*, of excited population by collisions transfers heat to bulk medium and forms thermal grating. (c) Probe incident at Bragg angle θ_2 is scattered to form signal.

List of Units and Nomenclature

Formulae are expressed in SI units

c	Speed of electromagnetic wave *in vacuo*
$\mathbf{E}$	Electric field amplitude
$g_0(\omega)$, $g(\omega)$	Gain coefficient at line center, at frequency ω
h	Planck's constant
I	Intensity of electromagnetic wave
I_{sat}	Saturation intensity of medium
$\mathbf{k}$, k	Wave vector, scalar value of wave vector ω/c
m_e	Mass of electron
n_m	Refractive index of medium for mth beam
N	Atomic or molecular population density
$\mathbf{P}$	Polarization field amplitude
T_1	Longitudinal relaxation time
T_2	Transverse relaxation time
α_0, $\alpha(\omega)$	Absorption coefficient at line center, at frequency ω
$\beta(x)$	Molecular polarizability as function of internuclear separation x
ε_0	Permittivity of free space
μ_{12}	Dipole moment for transition between energy levels 1 and 2
ν	Frequency of electromagnetic wave
$\chi^{(n)}$	nth order susceptibility
ω	Angular frequency of electromagnetic wave

See also

Coherent Transients: Coherent Transient Spectroscopy in Atomic and Molecular Vapours; Foundations of Coherent Transients in Semiconductors; Ultrafast Studies of Semiconductors. **Fiber and Guided Wave Optics:** Nonlinear Optics. **Imaging:** Adaptive Optics. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation; Four-Wave Mixing; Kramers–Kronig Relations in Nonlinear Optics. **Phase Control:** Phase Conjugation and Image Correction. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces. **Scattering:** Raman Scattering; Stimulated Scattering. **Spectroscopy:** Raman Spectroscopy; Second-Harmonic Spectroscopy.

Further Reading

Clark RJH and Hester RE (eds) (1988) Quantitative CARS spectroscopy. In: *Advances in Non-Linear Spectroscopy*, pp. 193–251. New York: John Wiley.

Corney A (1986) *Atomic and Laser Spectroscopy*. Oxford, UK: Oxford University Press.

Demtröder W (1996) *Laser Spectroscopy*, 2nd edn. Berlin: Springer-Verlag.

Dreier T and Ewart P (2002) Coherent techniques for measurements with intermediate concentrations. In: Kohse-Höinghaus K and Jeffries JB (eds) *Applied Combustion Diagnostics*. London: Taylor and Francis.

Eesley GL (1981) *Coherent Raman Spectroscopy*. London: Pergamon.

Eichler HJ, Günter P and Pohl DW (1986) *Laser Induced Dynamic Gratings*. Berlin: Springer-Verlag.

Feld MS and Letokhov VS (eds) (1980) *Coherent Nonlinear Optics, Recent Advances*. Berlin: Springer-Verlag.

Fisher RA (ed.) (1983) *Optical Phase Conjugation*. San Diego, CA: Academic Press.

Levenson MD and Kano SS (1987) *Introduction to Nonlinear Laser Spectroscopy*. Revised edn. San Diego, CA: Academic Press.

Raman Spectroscopy

R Withnall, University of Greenwich, Chatham, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Inelastic light scattering, the optical analogue of Compton scattering, had been predicted to occur by Smekal in 1923, but it was Chandresekar Venkataraman Raman who provided the first experimental demonstration of the phenomenon in February 1928. Only a few months later in 1928 the Russian scientists, Landsberg and Mandelstam observed inelastic light scattering from a quartz crystal. In spite of initial claims to the contrary, this was the same effect that had been observed by Raman, although the Russian scientists did not acknowledge this, preferring to call the effect 'combinatorial scattering'. It was not until the 1970s that the effect became known as Raman scattering everywhere, including Russia.

As typically practised, Raman spectroscopy involves laser excitation of a sample and measurement of the wavelength and intensity distribution of the scattered Raman light. When a microscope is used for delivering laser excitation and/or collecting the inelastically scattered light, the technique is referred to as Raman microscopy. However, it is important to recognize that, apart from the differences in sampling configuration, there are no fundamental differences between Raman microscopy and Raman spectroscopy; the terms merely identify the different sampling techniques.

Since its invention in 1960, major advances in the technology of the laser have occurred, resulting in the wide variety of laser light sources that are currently available for Raman spectroscopy. Characteristic properties of laser systems for Raman microscopy will be described in this chapter, following an introduction to the technique and its applications in the next section below. Then, the commonly used laser sources for Raman microscopy are categorized according to the wavelength regions of their emissions, along with their merits and drawbacks for each application. The chapter will consider mainly continuous wave laser sources, because pulsed laser beams that are tightly focused with microscope optics give very high irradiance that would destroy most samples.

Raman Microscopy

The first experimental Raman microscopes were developed in the mid-1970s by two independent groups, and it was not long before the first generation of Raman microscopes were subsequently commercialized. Many of these early instruments simply consisted of commercial optical microscopes coupled to Raman spectrometers, as shown in Figure 1. It was recognized that the use of a microscope for sampling in this way can facilitate the Raman examination of tiny particles of micrometer dimensions.

The optical lay-out in Figure 1 shows an infinity corrected microscope. The advantage of using this, rather than a standard tube length microscope, is that its length can be extended in order to incorporate additional optical elements that are required for coupling to the Raman excitation and collection optics. In the optical layout, the incident laser beam is spatially filtered by a pinhole, D_1, in order to remove the diffraction rings and give a point source. The laser light is then partially reflected by beamsplitter, B_1, and focused on to the sample by the microscope objective. The back-scattered Raman radiation is collected by the same microscope objective, partially transmitted by the beamsplitter, and directed into the entrance slit of the Raman spectrometer by means of coupling optics. The accurate location of an aperture, D_2, at the focal point improves the spatial resolution and allows depth profiling of transparent samples. The two pinholes, D_1 and D_2, act as spatial filters and are referred to as confocal diaphragms. The Raman microscope is described as being confocal, because out-of-focus light, collected from outside the focal volume, is rejected.

The use of the beamsplitter, in the optical path of early Raman microscope designs, gives rise to significant losses of the precious Raman

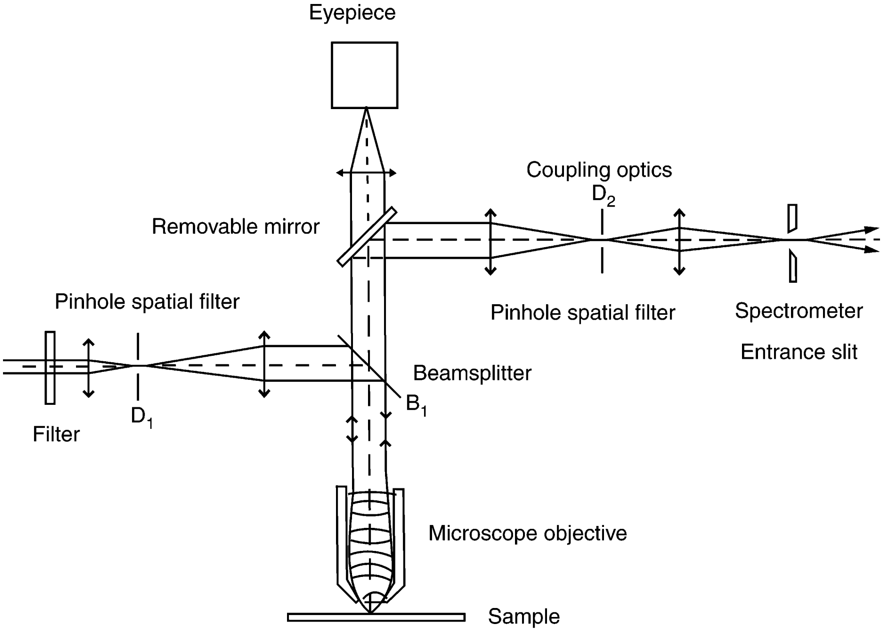

Figure 1 Diagram of the optical layout of a visible Raman microscope (Reproduced from Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy. Development and Applications*, 2. London: Academic Press).

scattered light. If a 50:50 beamsplitter is used then half of the incoming laser light is lost by transmission at B_1 and, more significantly, half of the Raman scattered light is lost by reflection at B_1. As long as a higher laser power can be employed, it is better to use a 90:10 beamsplitter that transmits 90% of the light and reflects 10%; then, only 10% of the incoming laser light is directed by B_1 on to the sample, but 90% of the Raman scattered light is transmitted by B_1 towards the detector and only 10% of the Raman light is lost by reflection at B_1. The efficiency of a beamsplitter can be defined as the product of the reflectivity at the laser wavelength and the transmission at the Raman-shifted wavelength. Thus, conventional 50:50 and 90:10 beamsplitters would have efficiencies of 25% and 9%, respectively. Recently, however, holographic notch filters have been developed which can be used as higher efficiency beamsplitters. These typically have a laser rejection contrast ratio of better than 10^4, reflecting ca. 90% of the laser excitation and transmitting ca. 90% of the Raman scattered light, thus giving an efficiency of ca. 81%. Due to these considerations, higher throughput, commercial Raman microscope designs usually incorporate a holographic notch filter.

The importance of Raman microscopy stems from the fact that it is the only microanalytic method available today, by use of which it is possible to identify, or characterize, small particles of micrometer dimensions *in situ*. It is also advantageous that no sample preparation is required when performing Raman microscopy.

The fundamental limit of the lateral spatial resolution (i.e., for a diffraction limited focus) is the separation at which the maximum of one Airy disc function just touches the first minimum of an adjacent Airy disc, given by:

$$\text{Lateral spatial resolution} = \frac{1.22\lambda}{2\text{NA}} \qquad [1]$$

where λ is the wavelength of the light and NA is the numerical aperture of the microscope objective.

The axial Airy disc function determines the resolution in the longitudinal direction. A good estimate of the axial resolution limit for low numerical aperture is given by:

$$\text{Axial spatial resolution} = \frac{2\lambda}{(\text{NA})^2} \qquad [2]$$

When the sample is heterogeneous and exhibits fluorescence that is not evenly distributed within the sample, a region of the sample can be selected with the microscope that shows the minimum amount of fluorescence. If the fluorescence is intrinsic to the sample itself, then it is possible to use the shift subtract technique and/or temporal discrimination between the fluorescence and the Raman scattering. The latter can be achieved by Kerr gate fluorescence rejection, in which a pulse of light is used to close a Kerr gate shutter. The problem with this approach, however, is that it uses a laser pulse, and even pulses of modest energy will have extremely high irradiance when focused to a tiny spot by a microscope objective. Such laser pulses would inevitably destroy the sample under a microscope.

Another approach for reducing fluorescence is to use near infrared excitation, which is low enough in energy so that absorption cannot occur to promote electronic transitions. The technique of Fourier Transform Raman (FTR) spectroscopy often offers the best chance of obtaining Raman spectra from fluorescent samples for this reason. Low energy excitation of wavelength equal to 1064 nm, provided by a Nd:YAG laser, is normally used for FTR. Disadvantages are that the Raman scattering efficiency is low relative to that obtained with visible excitation, due to the ν^4 dependency (ν^4 refers to Rayleigh scattering of the Raman light). This is compensated to some extent by employing a high throughput (Jacquinot advantage) interferometer. However, this works better for macro rather than micro samples. This is because the coupling of a microscope to an FT Raman spectrometer has a fundamental drawback; the microscope is throughput limited, so the high throughput advantage of the FTR spectrometer cannot be realized in FTR microscopy. The image of the Jacquinot stop (the large circular aperture which is the entrance to the interferometer) at the sample plane typically has a diameter of a few hundred μm, which is much larger than particles with diameters of ca. 1 μm having similar dimensions to the waist diameter of the 1064 nm light excitation at its diffraction-limited focus. Consequently the throughput advantage of the interferometer is not fully exploited and there is a trade-off of spatial resolution with signal-to-noise at the detector. For this reason, the typical spatial resolution that is achieved is in the range of 15–100 μm rather than 4 μm, as claimed in the literature. Bruker has commercialized an FTR microscope by coupling an optical microscope to an FTR spectrometer with near infrared optical fibers.

An approach that offers more promise for reducing fluorescence and achieving spatial resolution close to the diffraction limit is to use near infrared excitation in conjunction with dispersive Raman microscopy. For example, semiconductor lasers operating from 785 to 852 nm can be used in conjunction with sensitive multichannel silicon-based CCD arrays. If longer excitation wavelengths are necessary, in order to overcome the problems with fluorescence, then diode lasers can be used that emit at longer wavelengths in conjunction with dispersive Raman spectrometers equipped with multichannel near infrared (NIR) detectors, e.g., Ge- or InGaAs-array detectors.

The applications of Raman microscopy cover many areas, including material science, the earth sciences, environmental science, biology and medicine, forensic science, and even the analysis of artworks. These areas are too wide ranging to be described in detail here, but the interested reader is referred to the Further Reading section at the end of this article.

Characteristics of Laser Sources

The laser is an excellent light source for Raman spectroscopy to such an extent that the terms 'Raman spectroscopy' and 'laser Raman spectroscopy' are synonymous for all but the most esoteric experiments, for example, with synchotron sources. Indeed, the advent of the laser was the stimulation for the renaissance of Raman spectroscopy in the 1960s, given its special properties, such as monochromaticity, high intensity, beam collimation, and coherence.

The characteristics of laser light and the advantages it offers to Raman microscopy are now considered.

Beam Quality

The light inside a laser tube is formed from a number of standing waves having distinct vibrational modes. There are a limited number of these modes transverse to the beam and these are characterized by the TEM_{pq} number (where p and q can be 0, 1, 2,…) where TEM is an acronym for 'transverse electromagnetic'. When a laser is operating in its fundamental TEM_{00} mode, light rays are reflected on axes between the end mirrors of the laser cavity. The '00' indicates that there are no nodes in the beam profile (Figure 2a), and the laser beam has a Gaussian intensity profile in the radial direction:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \qquad [3]$$

where $I(r)$ is the irradiance at a radial distance r from the axis of the beam, I_0 is the axial irradiance, and w is the beam radius.

For higher-order modes, a number of nodes are observed in the beam profile, which arise from off-axis light rays being reflected between the end mirrors, for example, the TEM_{11} mode has two nodes which are mutually perpendicular (Figure 2b).

The propagation characteristics of a Gaussian beam can be fully defined, either by the diameter of the beam waist or by the far-field divergence. Consequently, it is only necessary to know the diameter of the beam waist ($2w_0$), or the diameter of the beam ($2w_z$) at a longitudinal distance z from the waist, in order to determine the propagation characteristics of the beam:

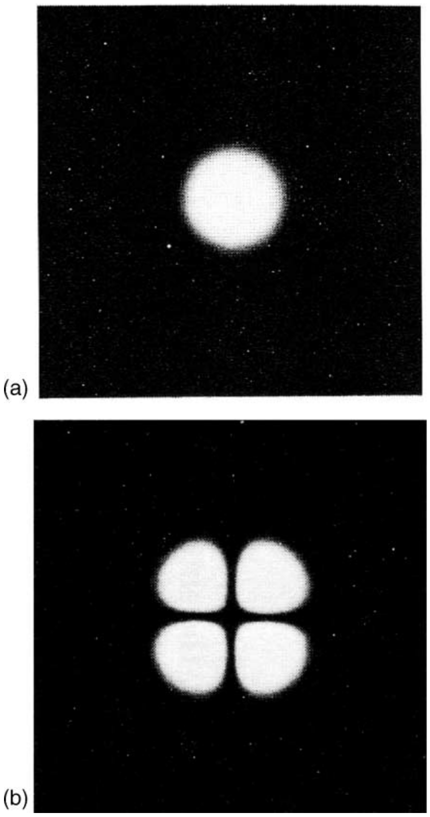

Figure 2 Transverse electromagnetic modes formed with confocal concave mirrors, (a) TEM_{00}, and (b) TEM_{11}. (Reproduced with permission from Young M (1977) *Optics and Lasers: an Engineering Approach.* Berlin, Heidelberg: Springer-Verlag

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_R}\right)^2} \qquad [4]$$

$$R(z) = z\left[1 + \left(\frac{z_R}{z}\right)^2\right] \qquad [5]$$

where the quantity $z_R = \pi w_0^2/\lambda$ is known as the Rayleigh range of the beam, λ is the wavelength of the laser radiation, and $R(z)$ is the radius of curvature of the wavefront at a distance z from the beam waist.

The wavefront is planar at the minimum beam waist and the Rayleigh range is the distance from the beam waist to the location at which the wavefront is most curved (Figure 3), the region from the waist to

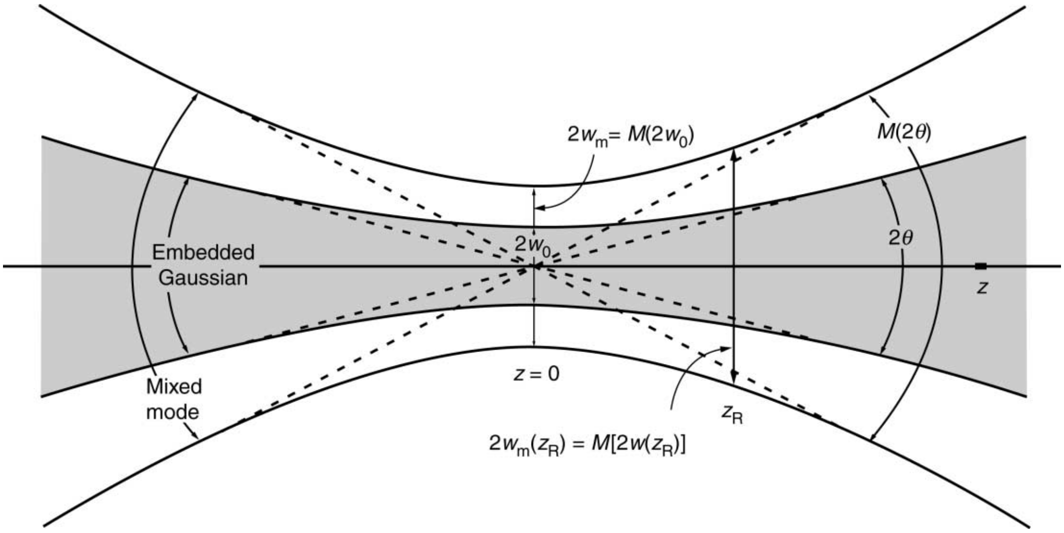

Figure 3 The Rayleigh range and embedded Gaussian (adapted from http://beammeasurement.mellesgriot.com/tut_m2.asp).

the Rayleigh range being the near field. Beyond approximately ten times the Rayleigh range, in the far field, the beam diverges as a cone with approximately straight sides. It can be seen by substituting $z = z_R$ into eqn [4] that the beam diameter at the Rayleigh range is $\sqrt{2}$ times the waist diameter.

Unfortunately laser beams do not conform to pure Gaussian functions and therefore they do not propagate according to the above equations. As a result, a dimensionless beam propagation parameter was devised in the early 1970s. This parameter, known as the M^2 factor or the 'times diffraction limit', is based on the brightness theorem, which states that for any laser beam the product of the beam diameter and the far-field divergence is a constant. Thus, M^2 is defined as the ratio of the laser beam's multimode diameter-divergence product to the ideal diffraction-limited (TEM_{00}) beam diameter-divergence product:

$$M^2 = \left(\frac{2w_m\,\Theta_m}{2w_0\,\theta_0}\right) \qquad [6]$$

where Θ_m is the laser beam's multimode divergence, θ_0 is the theoretical diffraction-limited divergence, $2w_m$ is the laser beam's multimode waist diameter, and $2w_0$ is the ideal diffraction-limited beam waist diameter.

Alternatively, M^2 can be defined as the square of the ratio of the multimode beam diameter ($2w_m$) to the diffraction-limited beam diameter ($2w_0$):

$$M^2 = \left(\frac{2w_m}{2w_0}\right)^2 \qquad [7]$$

The intensity of the beam has a Gaussian distribution in the radial direction, but the accepted definition of beam diameter is the distance at which the intensity has fallen to $1/e^2$ (i.e., ca. 13.5%) of its axial intensity, as can be seen by setting $r = w$ in eqn [3] above. This definition only applies properly to Gaussian beams; for other beam profiles the diameter can be calculated using a second moment measurement of the entire beam.

The real beam, then, can be treated as Gaussian by substituting an artificial wavelength $M^2\lambda$ into the equations that apply to an ideal diffraction-limited TEM_{00} beam. The ideal TEM_{00} Gaussian can be superimposed as an 'embedded Gaussian' on the optical diagram of the real multimode beam (Figure 3). It can be seen that the multimode beam has a waist diameter and a divergence that are both M times larger than those of the embedded Gaussian.

The M^2 factor has a value of unity for an ideal diffraction-limited TEM_{00} beam, and values of greater than unity for all other beams; it can be as high as several hundred for a distorted beam of poor quality. This is an important parameter where a tightly focused laser spot (e.g., confocal Raman microscopy) or low divergence (e.g., remote sensing) are required.

Polarization

Although polarization is not an inherent property of laser light, lasers provide plane polarized light when the end windows of the laser tube are mounted at Brewster's angle (θ_B):

$$\theta_B = \tan^{-1} n \qquad [8]$$

where n is the refractive index of the window material for the appropriate wavelength of laser light. For this angle, reflected light is completely plane polarized and consequently the laser light that is generated is also plane polarized. Brewster windows are often used inside the laser cavity in order to reduce reflection losses. This is important in low-gain lasers such as HeNe lasers where laser action could be prevented by reflection losses. Air-cooled ionic gas lasers typically deliver highly linearly polarised output with ratios exceeding 1000:1.

The polarization property of the exciting light is used a great deal in Raman spectroscopy. In combination with the polarization properties of the Raman scattered light, it can be used to determine the depolarization ratios, ρ, of Raman bands in the solution state. In solid state studies it can be used to determine the degree of orientation, for example, in polymeric fibers or films, and in single crystal studies it can be used to determine the symmetry classes of the Raman active vibrations. In such studies a linear polarizer is used to improve the degree of plane polarization of the laser light, and a half-wave plate is used to rotate the plane of polarization by 90°.

Plane polarized laser light can be converted to circularly polarized laser light which is required for Raman optical activity (ROA) studies. For example, plane polarized 514.5 nm laser light provided by an argon ion laser can be converted to circularly polarized light by an electro-optic modulator. As CCD detectors are usually employed in present-day ROA experiments and they have relatively long sampling times, slow polarization modulation is required which can be achieved by periodically rotating a quarter-wave plate or by applying a square quarter-wave voltage to a Pockels cell.

Whereas optical isomers or enantiomers give identical Raman spectra, they are distinguished in the ROA experiment by the sign of the ROA signal. Indeed, the sign can be used to determine the absolute configuration of a chiral molecule. Also the enantiomeric excess of mixtures of enantiomers can be determined in such experiments, with obvious applications in the pharmaceutical industry, and the conformations of biological molecules, for example, proteins, nucleic acids, sugars, and viruses, can be determined.

Up until now, the ROA technique has been practised by no more than a handful of research groups world-wide, but it is expected that this situation will change in the near future, since the first commercial ROA spectrometer (the ChiralRAMAN spectrometer) has recently been launched by Biotools Inc. This spectrometer makes use of a solid state laser as the source of the polarized laser light.

Beam Divergence

Referring to Figure 3, the full divergence angle, Θ, for the fundamental TEM$_{00}$ Gaussian beam is given by:

$$\Theta = \lim_{z\to\infty} \frac{2w(z)}{z} = \frac{2\lambda}{\pi w_0} \qquad [9]$$

From eqn [4] it can be seen that the spot size increases linearly with the longitudinal distance z and, from eqn [9], that it diverges at a constant cone angle, Θ. It can also be seen from eqn [9] that the smaller the spot size, w_0, the greater the beam divergence angle, Θ.

Indeed, a highly divergent beam is a problem with edge-emitting laser diodes due to the small dimensions of the light source. Furthermore, it is much smaller in the vertical direction than in the horizontal, giving rise to an elliptical output beam that diverges much more rapidly in the vertical direction than in the horizontal. The highly divergent, elliptical beam can be corrected, to an extent, with a cylindrical lens, but the inherent problem of a small, elliptical source can never be completely corrected.

In contrast, vertical-cavity surface-emitting semiconductor lasers (VCSELs) do not have the same limitations on beam divergence because the cavity is in the vertical direction and the light emission occurs from the surface, which has a much larger area than the light source of an edge emitter. Indeed, frequency-doubled green and blue semiconductor lasers are now available, having beams with M^2 values of less than 1.1 and beam divergences of a few milliradians. In terms of these beam characteristics, it appears that these lasers are beginning to challenge the argon ion laser.

Emission Linewidth

The laser linewidth limit, $\Delta\nu_L$, is given by the Schawlow–Townes expression:

$$\Delta\nu_L = \frac{2\pi h\nu(\Delta\nu_c)^2}{P_{out}} \qquad [10]$$

where $\Delta\nu_c$ is the linewidth of the passive resonator, $h\nu$ is the photon energy, and P_{out} is the laser power output.

The power output, P_{out}, is equal to the number of photons, N_p, in the resonator times the energy per photon, $h\nu$, divided by the photon lifetime, τ_c:

$$P_{out} = \frac{N_p h\nu}{\tau_c} \qquad [11]$$

Furthermore, the width $\Delta\nu$ of the resonance curve, at which the intensity has fallen off to half the maximum value, is given by:

$$\Delta\nu_c = (2\pi\tau_c)^{-1} \quad [12]$$

Substituting for P_{out} and $\Delta\nu_c$ from eqns [8] and [9] into eqn [10] gives the limit for the linewidth, $\Delta\nu_L$:

$$\Delta\nu_L = \frac{\Delta\nu_c}{N_p} \quad [13]$$

The linewidth of a laser is characterized by its Q factor:

$$Q = \frac{\nu}{\Delta\nu} \quad [14]$$

where ν is the frequency of the laser line and $\Delta\nu$ is the linewidth.

Other useful relationships involving linewidth parameters are:

$$Q = \frac{\lambda}{\Delta\lambda} \quad [15]$$

and

$$\Delta\tilde{\nu} = -\frac{\Delta\lambda}{\lambda^2} \quad [16]$$

where λ and $\Delta\lambda$ are the wavelength and linewidth (in units of metres), respectively, and $\Delta\tilde{\nu}$ is the linewidth expressed as a wavenumber (in units of cm^{-1}).

The Q factor can be as high as 10^8 which is of great use in high resolution spectroscopy.

Visible Lasers

Helium-Neon Laser

The helium-neon laser is continuously pumped electrically using a high dc voltage of up to 1 kV. The gain medium consists of a gas mixture of about 10 parts helium to each part of neon at a pressure of about 3 Torr. The gain of this medium is extremely low, hence the requirement for Brewster-angle windows to eliminate reflection loss of the light polarized in the plane that includes the axis of the laser and the normal to the Brewster window. Due to this requirement, the light output is necessarily plane polarized.

The pumping excites helium atoms by electron impact, and resonant energy transfer to neon atoms then occurs via collisions of the gaseous atoms (Figure 4). This creates a population inversion in the neon atoms, enabling the laser transition to occur at 632.8 nm. Following emission, the neon atoms decay nonradiatively down to their metastable $2p^5 3s^1$ level from which they decay back down to the ground state via collisional de-excitation with the walls of the tube. This mechanism restricts the maximum achievable power output to ca. 50 mW, because of the need to depopulate the metastable neon atoms by wall collisions.

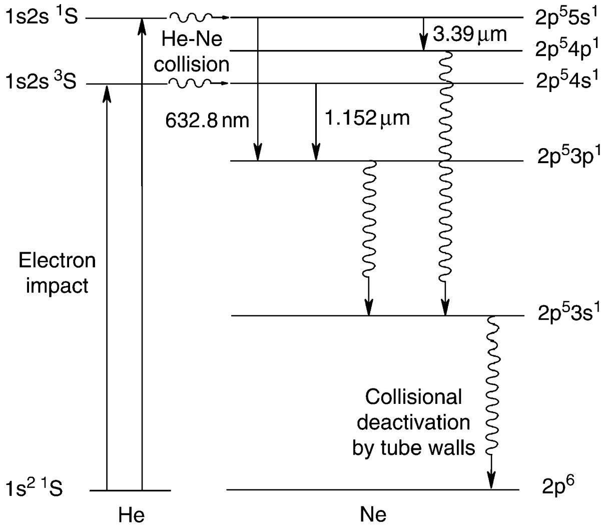

Figure 4 Energy levels of the He-Ne laser.

An advantage of using helium-neon lasers for Raman spectroscopy is that they generally operate in the fundamental TEM_{00} mode, which is critically dependent on the ratio of the length of the tube to its diameter. For this reason they are designed to have small tube diameters of around a few millimetres and lengths of 0.15–0.50 m, the small diameter also aiding collisional de-activation with the walls due to the relatively large surface to volume ratio of the tube. Another advantage of this laser for Raman spectroscopy is that the linewidth of the 632.8 nm emission line is ca. 1.5 GHz.

A disadvantage of the helium-neon laser is that it emits a large number of spontaneous emission lines originating from the excited neon atoms. These plasma emissions are observed in the Raman spectrum as sharp lines unless they are filtered out, for example, by a pre-monochromator or an interference filter. It should be mentioned, in passing, that these plasma lines are not always unwanted, because they can be very useful for wavenumber calibration of the Raman spectrometer.

The low output power of <50 mW is not generally a disadvantage for Raman microscopy since higher laser powers can often cause photolytic or thermal degradation of the sample. This is due to the high irradiance at the sample when the laser light is focused to a tight spot by the microscope objective. The excitation wavelength of 632.8 nm is suitable for combined use of the laser with a silicon-based CCD detector in Raman spectroscopy, because large Stokes shifts in the 3000–4000 cm^{-1} region lie well within the quantum efficiency curve of this detector. It is also possible, though less common, to operate the helium-neon laser on weaker transitions that include the green 543.5 nm line.

Argon Ion Laser

The argon ion laser was one of the first lasers to be discovered following the invention of the laser and up until now it has been used extensively for Raman spectroscopy, among many other applications. Excitation is provided by a continuous electrical discharge, and because of the high energy required to ionize the argon atoms and then promote the ions to an excited state, the efficiency of the laser is very small (ca. 0.1%). In spite of this, once population inversion has been achieved, the gain of the laser is very high and output powers of up to 25 W can be obtained for the strong lines at 488.0 and 514.5 nm and up to 50 W for multiline operation.

An advantage of the argon ion laser is that it can provide emission at more than 35 discrete wavelengths, the strongest of which are listed in Table 1. These lie in the green, blue, and near ultraviolet regions of the spectrum, a number of the ultraviolet lines only being obtained on the larger frame argon ion lasers. Discrete laser emission lines are selected by tuning a prism or grating inside the cavity. The argon ion laser can also be operated in multiline mode, for example, for pumping dye or Ti:sapphire lasers.

Table 1 Continuous wave argon ion laser wavelengths and output powers for the Coherent Inc. Innova 300 argon ion laser system

Wavelength (nm)	*Power (mW)*
228.9	30
238.3	100
244.0	400
248.3	180
257.3	750
275.4	5
300–305.5	20
333.4	40
333.8	30
335.8	20
351.1	200
351.4	60
363.8	240
454.5	140
457.9	420
465.8	180
472.7	240
476.5	720
488.0	1800
496.5	720
501.7	480
514.5	2400
528.7	420

The five lines below 260 nm are frequency doubled.

The linewidths of the argon ion emission lines at 488.0 and 514.5 nm are around 4.0 GHz. The high temperature laser tube has a diameter of approximately 12 mm and a length in the range of 0.1 to 1.8 m. A 240 V three phase power supply and ca. 35 A are required for pumping a medium power 5 W argon ion laser. The tube also requires water cooling at a flow rate of ca. 10 L/min and a pressure of 25 psi because of the large amount of heat dissipation.

For Raman microscopy smaller, air-cooled argon ion lasers, which only require a 240 V single phase power supply, can be used to provide output of a few hundred milliwatts on the 488.0 and 514.5 nm lines. For this application, the excitation is provided by argon ion lasers, which are designed to operate in the TEM_{00} mode.

Until recently, the argon ion laser has been a workhorse as the most common excitation source for Raman spectroscopy, but it is now losing ground to solid state lasers. The principle reasons for this are that the latter are much more efficient and consequently do not in general have three phase power and water-cooling requirements. Nevertheless, argon ion lasers still have a niche when high excitation powers on the order of watts are required in conjunction with a good beam quality ($M^2 < 1.1$), for example, for Raman spectroscopy of gaseous samples. Another advantage of argon ion lasers over solid state lasers is their multiline capability.

Krypton Ion Laser

The krypton ion laser has useful discrete emission lines in the near ultraviolet, blue, yellow, red, and near infrared regions of the electromagnetic spectrum; the strongest lines of a Coherent Inc. Innova 400 krypton ion laser are listed in Table 2.

The argon ion and krypton ion lasers are close relatives, thus the large frame krypton ion laser has similar power and water-cooling requirements to those mentioned above for the argon ion laser. Also, like the argon ion laser, air-cooled models with lower power output are available. The characteristics of the two lasers are also similar; for example, the linewidths of the krypton ion transitions at 530.9, 568.2, and 647.1 nm are about 4.0 GHz, the laser tube has similar dimensions (0.1–2.0 m in length) and the laser can be operated in either TEM_{00} or multimode.

The krypton ion laser is even less efficient than the argon ion laser, consequently lower-power outputs of

Table 2 Continuous wave krypton ion laser wavelengths and output powers for the Coherent Inc. Innova 400 krypton ion laser system

Wavelength (nm)	*Power (mW)*
206.5	4
234	8
337.5–356.4	2000
406.7	900
413.1	1800
415.4	280
468.0	500
476.2	400
482.5	400
520.8	700
530.9	1500
568.2	1100
647.1	3500
676.4	900
720.8	45
752.5	1200
793.1–799.3	300

The 206.5 and 234 nm lines are frequency doubled.

up to about 20 W can be achieved when operated in multiline mode.

Mixed argon and krypton ion laser tubes are also commonly used that provide laser lines originating from both argon ion and krypton ion transitions. As with the individual argon ion and krypton ion lasers, sharp plasma lines, due to spontaneous emission from the rare gas ions, are observed in the Raman spectrum unless they are filtered out. If interference filters are used for this, each laser line requires its own filter that is tailored to the specific emission wavelength.

HeCd Laser

Laser lines at 325 and 442 nm, both capable of providing milliwatts of output power, can be obtained from the helium-cadmium laser. These emission lines result from electronic transitions in free cadmium atoms.

It is obviously advantageous to use blue, rather than longer wavelength, excitation (e.g., provided by the 442 nm line of the HeCd laser) for off-resonance Raman spectroscopy, due to the ν^4 dependence of the Raman light scattering efficiency, provided that the efficiencies of the illumination/collection optics and Raman spectrometer throughput are optimized in the blue region of the spectrum. Unfortunately, far more samples fluoresce when excited with blue light than with red or near infrared radiation, which explains why red or near infrared lasers (e.g., HeNe, semiconductor lasers) are more commonly used for general Raman applications.

Near Infrared Lasers

Although the traditional laser systems for Raman spectroscopy have been the argon ion, krypton ion, and helium-neon lasers for discrete excitation wavelengths, the diode laser has gained popularity over recent years.

The principal advantage of near infrared semiconductor lasers for Raman microscopy is that they generally excite less fluorescence in the Raman spectrum than visible lasers due to their longer wavelength. Commonly available wavelengths are 670, 785, 830, and 852 nm and, of these, the first two can be used in conjunction with silicon-based CCD detectors to give Stokes Raman shifts over the whole range of fundamental vibrations, up to 4000 cm^{-1}. However, for excitation wavelengths of 830 and 852 nm, only Stokes Raman bands in the fingerprint region can be detected with a silicon-based CCD detector because higher wavenumber Raman shifts approach the bandgap of the silicon semiconductor ($\lambda >$ ca. 1050 nm).

The disadvantages of diode lasers are that they cannot supply high power of narrow linewidth in single-mode and they are susceptible to mode hopping which gives rise to a shift in the excitation wavelength. This latter drawback is particularly disadvantageous for Raman spectroscopy because it results in a shift in the wavenumber positions of the Raman bands. When good beam quality is not required, broad stripe, high power laser diodes can be used; these can have output powers that are greater than 1 W but their emission linewidths are equal to ca. 2 nm, which is far too wide for Raman spectroscopic applications. These laser diodes find applications as pumps of solid state lasers, for example, Nd:YAG, Nd:YVO_4 lasers, however. Additionally, amplified stimulated emission (ASE) gives an unwanted background signal that can be very broad (ca. 20–30 nm) and 0.1–1% of the intensity of the laser line. This necessitates the use of bandpass filters in order to reduce this unwanted background.

Although laser diodes having an output of less than 200 mW can operate in TEM_{00} and in single longitudinal modes, mode hopping can occur due to optical feedback or fluctuations in environmental factors and this causes severe broadening of the linewidth. The sensitivity to optical feedback can be greatly reduced, hence the laser linewidth can be narrowed appreciably, by confining the frequency of the photons either in an internal cavity containing a small diffraction grating or in an external cavity. The former types of laser are sometimes called 'distributed feedback' (DFB) lasers because the feedback is distributed over the length of the grating, rather than occurring all at once at a mirror. The wavelength that is fed back is determined by the period of the grating. Usually, a DFB laser has a grating fabricated into the entire length of the laser. A variation referred to as a distributed Bragg reflector (DBR) laser has a distinct grating fabricated into the substrate on each side of the active area. The external cavity semiconductor laser (ECSL) is fabricated by placing the laser diode in a separate resonator like a conventional gas or solid-state laser. The DBR and ECSL lasers are now discussed.

Distributed Bragg Reflector Lasers

The DBR laser consists of a grating on each side of the active region (Figure 5); these gratings act as mirrors having a reflectivity that is optimized at one particular wavelength, in addition to narrowing the laser linewidth. At present, DBR lasers are only commercially available having an excitation wavelength of 852 nm.

The emission linewidth is less than 4 MHz, which is suitable for the vast majority of Raman spectroscopic applications and the laser operates in single TEM_{00} mode. The disadvantages of DBR lasers are that the maximum output power is restricted to around 150 mW, so an optical isolator is usually necessary in order to prevent external facet damage caused by external optical feedback, and only an excitation wavelength of 852 nm is currently available.

External-Cavity Semiconductor Lasers (ECSL)

ECSLs provide higher output power (up to ca. 1 W) and a wider range of excitation wavelengths (630 nm to around 850 nm), as well as being less expensive than DBR lasers. They use the semiconductor chip only as the gain medium and employ an external grating, both as the frequency selector and the reflective mirror. Specific excitation wavelengths are becoming widely adopted in Raman microscopy employing single grating spectrograph in conjunction with CCD detection, such as 670, 785, 830, and 852 nm. The ECSLs have emission linewidths as low as a few MHz, but in general they span the range 2 MHz–30 GHz, and can operate in a nearly diffraction-limited transverse mode.

Three different cavity designs employing a diffraction grating have been discussed in the literature: Littrow type-I, Littrow type-II, and Littman configurations. In the Littrow type-I design, the laser diode light is incident on the grating at an angle of incidence equal to $\theta_{\mathrm{Littrow}}$ (Figure 6a). The diffracted light in first order is re-directed back to the laser diode to provide feedback and the output is the zeroth order (i.e., specularly reflected) light. A disadvantage of this design is that the bandwidth of the grating is relatively large because a single pass geometry is used. In the Littrow type-II design, light from the rear facet of the laser diode is collimated by a lens and directed on to a grating at an angle of incidence equal to $\theta_{\mathrm{Littrow}}$ (Figure 6b). The diffracted light in first order is re-directed back to the laser diode, in a similar fashion to the Littrow type-I design, in order to provide optical feedback, and the output is the laser light emitted from the front facet. Disadvantages of the Littrow type-II design are the requirement for

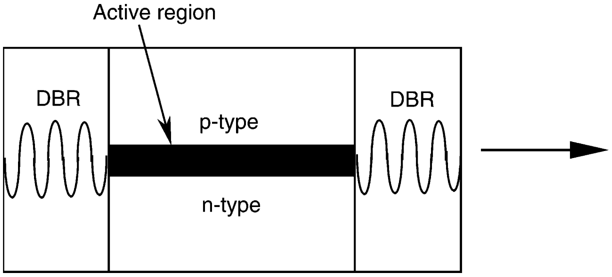

Figure 5 Diagram of the cavity of the DBR laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

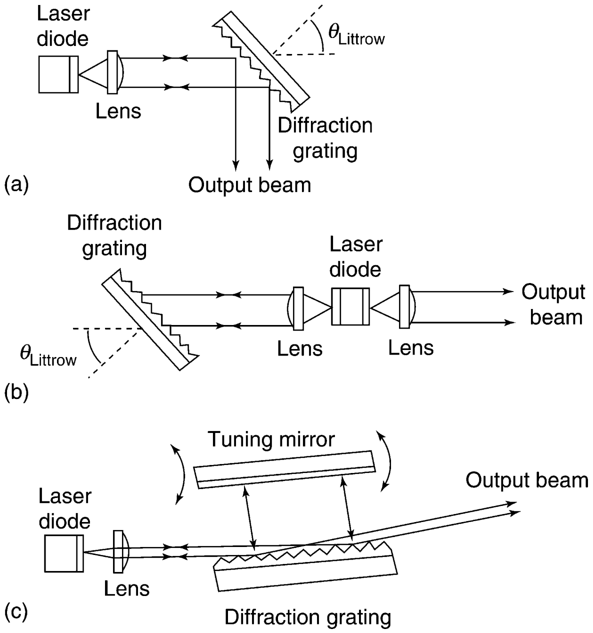

Figure 6 Diagrams of the cavities of (a) Littrow type-I, (b) Littrow type-II, and (c) Littman external cavity diode lasers (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

custom fabrication of antireflection coatings on both facets and the sensitivity to external optical feedback. In the Littman design, the laser diode light is collimated by a lens and directed on to a grating at a grazing angle of incidence. The diffracted light is reflected by a mirror and diffracted back to the laser diode by the grating in order to provide optical feedback (Figure 6c). An advantage of this design is that the grating bandwidth is less than half that of the Littrow designs due to the double pass geometry, but a disadvantage is that the external cavity length is longer, resulting in a narrower spacing of the cavity modes. Another advantage of this design is that the tuning is accomplished by rotating the mirror instead of the grating, thus the alignment of the output beam is not altered when tuning. For Littrow type I and II and Littman ECSL designs, filtering of the ASE is required, as it can have an intensity of ca. 0.1 to 1% of the intensity of the laser line. Thus, these designs necessitate the incorporation of a bandpass filter (having a rejection of better than 10^{-4} at the ASE).

Nd:YAG Laser

Under normal conditions, the Nd:YAG laser oscillates on a transition in the near infrared at 1064 nm, and this is the excitation line that is used in most commercial Fourier Transform Raman spectrometers. The gain medium is a crystal of yttrium aluminium garnet ($Y_3Al_5O_{12}$, YAG) doped with ca. 1.0 mol% Nd^{3+} cations that substitute Y^{3+} ions in the cubic YAG lattice.

The Nd:YAG laser is a four-level system (Figure 7), which has high gain and low threshold due to the narrow fluorescent linewidth of the laser transition. Absorption bands of the Nd^{3+}ions around 808 nm conveniently match the energy of commercially available high-power multimode diode lasers that serve as the pump. The Nd^{3+} ions decay nonradiatively to the upper laser level, the $^4F_{3/2}$ state, thereby creating a population inversion. This is because the lower laser level, the $^4I_{11/2}$ state, has no appreciable thermal population at room temperature, since it is $>2000\ cm^{-1}$ above the $^4I_{9/2}$ ground state, and the $^4F_{3/2}$ excited state has a relatively long lifetime of 230 μs.

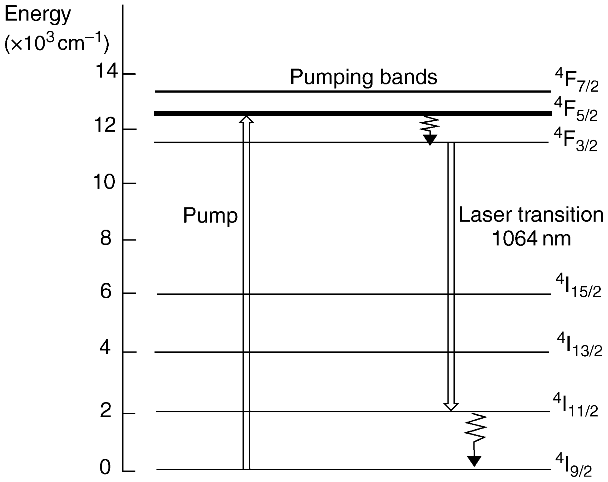

Figure 7 Energy levels of the Nd:YAG laser.

The high thermal conductivity of Nd:YAG enables the laser to operate at high power in either continuous wave (CW) or *Q*-switched modes, and the diode pumped solid state (DPSS) variety of the cw Nd:YAG laser can currently achieve an output power in excess of 20 W on the 1064 nm line. The laser can be designed to operate in the fundamental TEM_{00} mode and the full width at half maximum (FWHM) linewidth of the spontaneous emission of the 1064 nm laser transition is 120 GHz (ca. 0.45 nm). Advantages of the diode pumped Nd:YAG laser are that it is air-cooled and can be operated at 240 V single phase. Also, being all solid state and having a small footprint, it is robust and portable. The lifetime of the laser is dependent on the laser diodes used for pumping, but some Nd:YAG laser designs enable these to be replaced in the field.

The frequency-doubled Nd:YAG laser emitting green light, having a wavelength of 532 nm, is also commonly used nowadays for dispersive Raman spectroscopy, and an output power of 10 W on the 532 nm laser line is provided by commercially available frequency-doubled diode-pumped Nd:YAG lasers.

It is worth mentioning that Nd:YAG lasers have also been fabricated that emit either at 946 or 1330 nm, by suppressing the strong emission at 1064 nm and optimising the optics for the desired wavelength.

Nd:YVO$_4$ and Nd:YLF Lasers

An intracavity, frequency-doubled DPSS $Nd:YVO_4$ laser has recently become commercially available, and it has some advantages over the Nd:YAG laser including a larger stimulated emission cross-section and a higher absorption coefficient (along the extraordinary direction of the birefringent crystal). It has the same emission wavelength as the frequency-doubled Nd:YAG laser, i.e., 532 nm, and $Nd:YVO_4$ is the material of choice for cw end-pumped lasers having around 5 W output power. This is because the diode laser pump beam is tightly focused in the end-pumped system, but a small waist diameter cannot be retained over a distance of more than a few millimeters, consequently a high absorption coefficient and gain are very beneficial.

The gain medium of the Nd:YLF laser consists of a crystal of yttrium lithium fluoride (YLF) that is doped with Nd^{3+} ions on the Y^{3+} cation sites. Unlike the Nd:YAG and Nd:YVO_4 lasers, the emission does not occur at 1064 nm; instead the $^4F_{3/2}-^4I_{11/2}$ emission occurs at wavelengths of either 1047 or 1053 nm, depending on the polarization that is selected. The former is due to extraordinary polarized light, whereas the latter is due to ordinary polarization, and either of these emission wavelengths can be selected using an intracavity polarizer. The Nd:YLF laser can offer benefits in *Q*-switched operation when the longer fluorescence lifetime (480 μs) of Nd^{3+} ions in the $^4F_{3/2}$ state enables a higher energy to be stored for the same number of pump laser diodes.

Ti:Sapphire Laser

The Ti:sapphire laser is tunable over the approximate range of 670–1070 nm with the peak of the gain curve at ca. 800 nm. In the gain medium, ca. 0.1% by weight Ti^{3+} is doped into a crystal of sapphire grown by the Czochralski method. The Ti^{3+} ions substitute for Al^{3+} ions in the Al_2O_3 of the sapphire and the laser emission is due to the $^2E-^2T_2$ transition of the Ti^{3+} cation, which has a $3d^1$ valence electronic configuration (Figure 8). The laser has a large stimulated emission cross-section but the fluorescence lifetime of the upper laser level (^2E state) is quite short (3.2 μs), thus the laser is usually laser (e.g., by argon ion or frequency-doubled Nd:YAG or Nd:YVO_4 lasers) rather than flashlamp pumped because a very high pump flux is required.

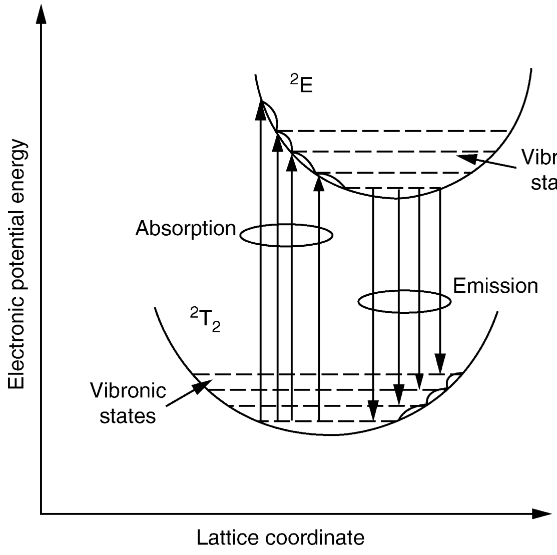

Figure 8 Energy levels of the Ti:sapphire laser (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

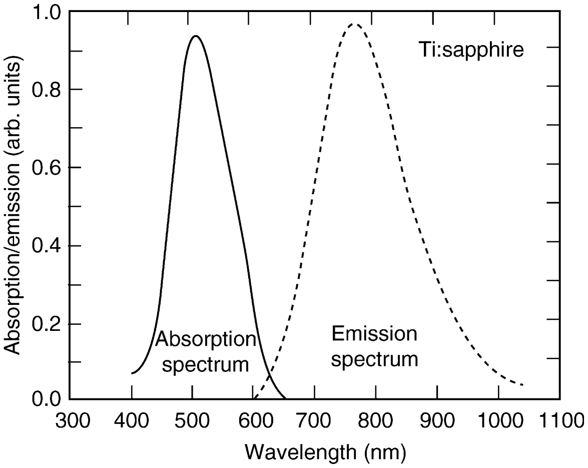

Figure 9 Absorption and emission spectra of the Ti^{3+} ion in sapphire. Al_2O_3 (adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley).

The absorption and emission bands are broad and widely separated, due to the vibronic coupling between the Ti^{3+} host and the Al_2O_3 lattice (Figure 9). The lower laser level is any one of the vibronic levels of the $^2T_2$ state. Following the laser transition, the Ti^{3+} ions decay from the upper vibronic levels of the $^2T_2$ state down to the lower vibronic levels. Hence, this is a four-level laser.

The broad spontaneous emission linewidth of the $^2E-^2T_2$ laser transition is around 100 THz, the ouput power can be as high as 50 W, and the laser can be operated in either TEM_{00} or multimodes.

UV Lasers

UV laser sources offer numerous advantages over visible laser sources for Raman spectroscopy. A major consideration is that many analytes have absorptions in the near UV, making them amenable to resonance Raman spectroscopic studies. The signal enhancement (up to 10^6), that can be achieved in resonance Raman spectra, results in a large increase in detection sensitivity. Furthermore, some vibrational modes, which normally give rise to weak bands in the off resonance Raman spectrum, can show strong enhancement in the UV excited resonance Raman spectrum. A good example of this is the amide II band of peptides and proteins, which is due to a combination of N–H in plane bending and C–N stretching modes of the peptide linkage. This band is normally weak in the Raman spectrum but strong in the infrared spectrum; however, the amide II band can show strong enhancement in the UV excited resonance Raman spectrum and this can be useful for determining secondary structure in proteins. A further benefit of UV excited Raman spectroscopy is that fluorescence can often be avoided as it tends to occur at a lower energy, outside the Stokes Raman spectral window.

It can be advantageous to use cw rather than pulsed excitation for UV resonance Raman spectroscopy, if the analyte has a long excited state lifetime relative to the pulsewidth of the pulsed laser. This is because the concentration of analyte molecules in the ground state is significantly depleted during pulsed excitation, due to Raman saturation giving a lower signal to noise ratio in the Raman spectrum than for the case of cw excitation. It has been found that pulse energy flux densities should be less than 1 mJ/cm^2 and flowing or spinning samples should be used, in order to ensure that nonlinear phenomena and saturation effects do not occur.

For UV Raman microscopy, where the laser beam is focused into a small spot of micrometer dimensions on the sample, it is essential to avoid using pulsed lasers delivering high peak powers. This is because such pulsed lasers can cause dielectric breakdown, and even at lower peak powers they can cause nonlinear effects and Raman saturation phenomena. Consequently, with UV Raman microscopy, one is restricted to the use of cw or quasicontinuous laser excitation.

It is only recently that high thoughput UV Raman microspectrometers have been developed. It has been demonstrated for visible Raman spectroscopy that a significantly higher throughput can be achieved by employing a single stage spectrograph with a notch filter instead of a double or triple monochromator. In the UV region, blocking the Rayleigh scattering is a problem because notch filters are not currently available. However, it has been discovered that this can be overcome by introducing two modifications to the design of the optical layout of a visible Raman microscope. First, two novel dielectric longpass filters were used instead of a notch filter in order to reject the elastically scattered light. Second, an all-reflecting Cassegrain microscope objective was used, instead of a lens, in order to block the specular reflection, and thereby further reduce the Rayleigh scattering background. These design modifications have enabled a single stage spectrograph with a holographic grating to be used to disperse the Stokes Raman radiation in a UV Raman microspectrometer (Figure 10).

In the optical layout of Figure 10 it can be seen that the laser excitation is not focused by the Cassegrain microscope objective because an epi-illumination configuration is not employed. Instead the laser is focused by a separate lens and directed on to the sample with a turning prism located directly underneath the Cassegrain objective. This minimizes loss of laser beam throughput or scattered light throughput to the spectrometer at the expense of spatial resolution, since the laser light is focused by a longer focal length lens than the microscope objective. An additional difference between the optical layouts of the visible and UV Raman microscopes shown in Figures 1 and 10, respectively, is that pinhole spatial filters are not shown in the latter. However, better axial spatial resolution could be achieved if this UV Raman microscope was made confocal by introducing a pinhole and lens.

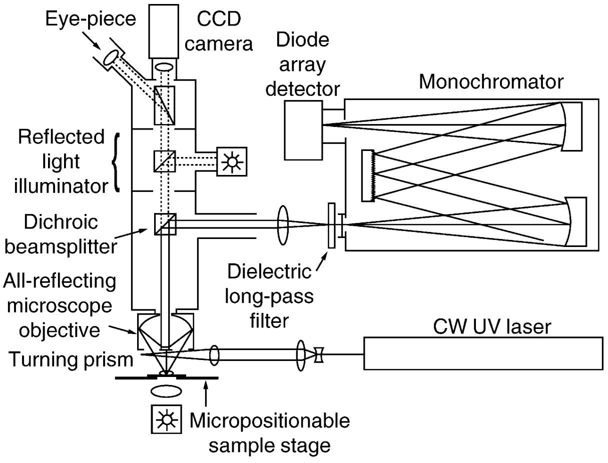
Eye-piece
CCD camera
Diode array detector
Monochromator
Reflected light illuminator
Dichroic beamsplitter
All-reflecting microscope objective
Turning prism
Dielectric long-pass filter
CW UV laser
Micropositionable sample stage

Figure 10 Diagram of the optical layout of a UV Raman microscope (adapted from Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86).

Frequency-Doubled Argon Ion Laser

The frequency doubled argon ion laser is a popular choice among UV laser sources. The cw UV laser contains a nonlinear optical beta-barium borate (BBO) crystal, which is located within the laser cavity; this crystal frequency doubles the strong Ar^+ lines to give five UV lines below 260 nm which are listed in Table 1 for the Coherent Inc. Innova 300 Ar^+ ion laser system. Of these UV lines, the 228.9 nm line is almost ideal for resonance Raman enhancement of tyrosine and trpytophan residues in proteins and the 244.0 nm line is well suited for studying tyrosinate groups. The 244.0 nm line has also been used to excite selectively UV resonance Raman spectra from spatially resolved areas of biological samples within the nucleus of a single cell, from DNA in particular, using low laser powers and short acquisition times to keep sample damage to a minimum under the microscope.

Frequency-Doubled Krypton Ion Laser

Like the cw frequency-doubled argon ion laser mentioned above, the cw frequency-doubled krypton ion laser contains a nonlinear optical BBO crystal, which is located within the laser cavity; this crystal frequency doubles the strong Kr^+ lines to give 206.5 and 234 nm UV lines which are listed in Table 2 for the Coherent Inc. Innova 400 Kr^+ ion laser system. The 206.5 nm laser line is useful for exciting resonance Raman spectra within the $\pi-\pi^*$ amide transition of the peptide backbone of proteins, and the enhanced protein amide vibrational bands can be used to determine the protein secondary structure.

HeCd Laser

The 325 nm HeCd excitation line, in the near ultraviolet, can be used for combined micro-Raman/photoluminescence studies of, for example, semiconductors. These lasers are suitable for low-power applications, typically having output powers in the 1–100 mW range, and they can be designed to operate in TEM_{00} single-mode or multimode.

Quasi-CW Mode-Locked Ti:Sapphire Laser

The second harmonic of the mode-locked Ti:sapphire laser (ca. 350–500 nm), third harmonic (ca. 233–333 nm), and fourth harmonic (ca. 200–250 nm), are all available with conventional nonlinear crystals.

Typically, the Ti:sapphire crystal is pumped by a cw argon ion laser, for example, as the excitation source for a UV Raman microscope. This quasicontinuous laser system produces 2–3 ps pulses at a 76 MHz repetition rate and is continuously tunable over the 200–300 nm range. The typical power output is ca. 50 mW at 250 nm, ca. 10 mW at 240 nm, and 1 mW at 230 nm. In future, it is anticipated that a frequency-doubled Nd:YAG laser will be increasingly favored as the pump, creating an all-solid-state laser source.

Conclusions

Laser systems for Raman microscopy have been described in this article. Although the differences in the terms 'Raman spectroscopy' and 'Raman microscopy' only refer to whether a macro or micro sampling configuration is used, this does have a bearing on the types of laser systems that are employed. This is because the latter imposes a restriction on using low duty pulsed lasers with high pulse peak power, as both spatial and temporal confinement of the laser excitation can lead to dielectric breakdown of the sample, nonlinear phenomena, or Raman saturation. Thus, for this reason, one is limited to cw lasers or quasi-cw lasers for Raman microscopy whereas there is no such restriction for Raman spectroscopy.

The quality of the laser beam, measured by the M^2 value or 'times diffraction limit' influences the ultimate spatial resolution that can be achieved in Raman microscopy.

In the visible region, argon ion, krypton ion, and frequency-doubled Nd:YAG lasers have good beam quality and high power, with helium-neon and helium-cadmium lasers also having good beam quality but lower power; this makes these lasers good light sources for visible Raman microscopy.

For FT Raman microscopy using Nd:YAG excitation, there is a lack of fluorescence interference from a wide variety of samples, but sensitivity and spatial resolution are compromised in comparison to visible dispersive Raman microscopy.

Red and near infrared semiconductor lasers operating in TEM_{00} single mode can have good beam quality, approaching that of gas lasers, but they are power limited. Obviously, when good beam quality is not necessary, for example, for optically pumping other laser sources, high-power diode lasers can be used in multimode operation. Solid state red and near infrared laser sources are becoming very popular for routine Raman microscopic analysis due to their compactness, robustness, and economical power consumption.

Another reason for the popularity of the red and near infrared diode laser sources for Raman microscopy is that fluorescence is less of a problem due to the lower energy of the light, compared to conventional green 514.5 nm excitation. Although there is a penalty, due to the dependency of the scattering efficiency on the ν^4 term, present day silicon-based CCD arrays can exhibit excellent detection quantum efficiency of Stokes shifted Raman radiation, across the whole vibrational spectrum, when using standard 670 and 785 nm diode lasers and, at least across the fingerprint region, when using 830 and 852 nm diode lasers.

For ultimate spatial resolution in the Raman microscopic experiment, short wavelength excitation is advantageous because it is fundamentally possible to focus to a smaller diffraction-limited laser spot. In this regard, it would be preferable to use UV Raman microscopy where fluorescence is also not as problematic as it is in the visible region. However, in spite of recent improvements in UV laser sources such as the intracavity frequency-doubled argon and krypton ion cw lasers, there are current instrumental limitations on the throughput efficiency of the UV Raman microscope imposed by the lack of availability of suitable notch filters and low stray light and high throughput optics.

It is anticipated that in the future all-solid-state, quasi-cw, diode pumped, mode-locked Ti:sapphire lasers operating on their third or fourth harmonics will become increasingly popular for providing tunable UV excitation affording good beam quality. It is now also possible to fabricate hollow cathode metal ion deep UV lasers, 10–15 cm in length, 2–4 cm in diameter, weighing 50–100 g and requiring only 2–3 W of electrical power. These lasers have low beam quality (M^2 equal to ca. 18), but they make possible the development of portable UV fluorescence imaging and Raman microprobes for geobiological exploration of terrestrial and extraterrestrial environments.

See also

Fourier Optics. **Imaging:** Infrared Imaging. **Introductory Article:** Early Development of the He-Ne Laser (Title TBC). **Lasers:** Noble Gas Ion Lasers; Semiconductor Lasers. **Nonlinear Optics, Applications:** Raman Lasers. **Scattering:** Raman Scattering.

Further Reading

Andrews DL (1990) *Lasers in Chemistry* (2nd edition). Berlin, Heidelberg: Springer-Verlag.

Asher SA, Bormett RW, Chen XG, *et al.* (1993) UV resonance Raman spectroscopy using a new cw laser source: Convenience and experimental simplicity. *Applied Spectroscopy* 47: 628–633.

Bell SEJ, Bourguignon ESO, O'Grady A, Villaumie J and Dennis AC (2002) Extracting Raman spectra from highly fluorescent samples with "Scissors" (SSRS, Shifted-Subtracted Raman Spectroscopy). *Spectroscopy Europe* 14: 17.

Best SP, Clark RJH and Withnall R (1992) Non-destructive pigment analysis of artefacts by Raman microscopy. *Endeavour* 16: 66.

Boustany NN, Manoharan R, Dasari RR and Feld MS (2000) Ultraviolet resonance Raman spectroscopy of bulk and microscopic human colon tissue. *Applied Spectroscopy* 54: 24–30.

Delhaye M, Barbillat J, Aubard J, Bridoux M and Da Silva E (1996) Instrumentation. In: Turrell G and Corset J (eds) *Raman Microscopy: Developments and Applications*, chap. 3. London: Academic Press.

Delhaye M and Dhamelincourt P (1975) Raman microprobe and microscope with laser excitation. *Journal of Raman Spectroscopy* 3: 33.

Derbyshire A and Withnall R (1999) Pigment analysis of portrait miniatures using Raman microscopy. *Journal of Raman Sprectroscopy* 30: 185.

Gordeyev SA, Nikolaeva GYu, Prokhorov KA, Withnall R, Dunkin IR and Shilton SJ (2001) Super-selective polysulfone hollow fiber membranes for gas separation: assessment of molecular orientation by Raman spectroscopy. *Laser Physics* 11: 82–85.

Hayward CL, Best SP, Clark RJH, Ross NL and Withnall R (1994) Polarised single crystal Raman spectroscopy of sinhalite, $MgAlBO_4$. *Spectrochimica Acta* 50A: 1287–1294.

Hendra P, Jones C and Warnes G (1991) *Fourier Transform Raman Spectroscopy*. London: Ellis Horwood.

Hitz B (2003) Solid state lasers are gunning for argon ion's place. *Photonics Spectra* 54–58.

Holtz JSW, Bormett RW, Chi Z, *et al.* (1996) Applications of a new 206.5 nm continuous wave laser source: UV Raman determination of protein secondary structure and CVD diamond material properties. *Applied Spectroscopy* 50: 1459–1468.

http://www.btools.com

Koechner W and Bass M (2003) *Solid-State Lasers*. New York: Springer-Verlag.

Landsberg G and Mandelstam L (1928) Eine neue Erschienung bei der Lichtzerstreaung in Krystallen. *Naturwiss* 16: 557.

Matousek P, Towrie M, Stanley A and Parker AW (1999) Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. *Applied Spectroscopy* 53: 1485.

Nakashima S, Okumura H, Yamamoto T and Shimidzu R (2004) Deep-ultraviolet Raman microspectroscopy: Characterization of wide-gap semiconductors. *Applied Spectroscopy* 58: 224–229.

Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86.

Pallister DM and Morris MD (1993) In: Morris MD (ed.) *Microscopic and Spectroscopic Imaging of the Chemical State*, chap. 1. New York: Marcel Dekker.

Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy,* 1. Chichester, UK: Wiley.

Pan M-W, Benner RE, Johnson CW and Smith LM (2000) Near-IR lasers rejuvenate Raman spectroscopy. *Optoelectronics World* S5–S10.

Raman CV and Krishnan KS (1928) A new type of secondary radiation. *Nature* 121: 501.

Rosasco GJ (1980) Raman microprobe spectroscopy. In: Clark RJH and Hester RE (eds) *Advances in Infrared and Raman Spectroscopy*, vol. 7. Chichester: Wiley.

Rosasco J, Etz ES and Cassatt WA (1975) The analysis of discrete fine particles by Raman spectroscopy. *Applied Spectroscopy* 29: 396.

Schawlow AL and Townes CH (1958) Infrared and optical masers. *Physics Review* 112: 1940.

Smekal A (1923) Sur Quantentheorie der Dispersion. *Naturwiss* 11: 873.

Storrie-Lombardi MC, Hug WF, McDonald GD, Tsapin AI and Nealson KH (2001) Hollow cathode ion lasers for deep ultraviolet Raman spectroscopy and fluorescence imaging. *Review of Scientific Instruments* 72: 4452–4459.

Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy: Development and Applications*, 2. London: Academic Press.

Young M (1977) *Optics and Laser: An Engineering Approach*. Berlin, Heidelberg: Springer-Verlag.

Yu Y-M, Nam S, Byungsung O, *et al.* (2002) Resonant Raman scattering in ZnS epilayers. *Materials Chemistry and Physics* 78: 149–153.

Second-Harmonic Spectroscopy

J I Dadap and T F Heinz, Columbia University, New York, NY, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The demonstration of second-harmonic generation (SHG), one of the early experiments with pulsed lasers, is considered to have marked the birth of the field of nonlinear optics. The importance of second-harmonic generation, and the related phenomena of sum- and difference-frequency generation, as methods for producing new frequencies of coherent radiation was recognized immediately and has in no way diminished over the years. Associated with these applications came an interest in the fundamental nature of the second-order nonlinear response, and some spectroscopic studies were undertaken with this motivation. These investigations were of rather limited scope and were not developed as a general probe of materials. In contrast, the third-order nonlinear optical interactions have given rise to a panoply of significant spectroscopic measurement techniques, including, to name a few, coherent Raman spectroscopy, pump-probe and other four-wave mixing measurements, two-photon absorption and hole burning measurements, photon echoes, and Doppler-free spectroscopy.

A central factor in the lack of spectroscopic applications of the second-order nonlinear response lies in a fundamental symmetry constraint. Unlike the third-order nonlinearity, which is present in all materials, only materials lacking a center of inversion give rise to an (electric-dipole) allowed second-order nonlinear response. As we demonstrate in this article, however, it is this property that makes SHG, as well as the sum-frequency generation process discussed elsewhere in this encyclopedia, such a remarkable probe of surfaces and interfaces. For centrosymmetric materials, the second-order nonlinear response is strong only at interfaces, where only a monolayer or so of material contributes to the response. This yields a purely optical probe (with incident and emitted photons) that gives an inherent surface sensitivity comparable to the best electron spectroscopies, such as electron diffraction, Auger-electron spectroscopy, photoemission spectroscopy, and scanning tunneling microscopy, the workhorses of surface science. Neither linear nor third-order nonlinear optical measurements provide this inherent interface sensitivity. The possibility of probing interfaces by purely optical means permits one to take advantage of several important features of optical radiation. Perhaps most noteworthy is the possibility of probing buried interfaces of every sort, from solid/solid to liquid/liquid, as well as the more conventional cases of solid/vapor, liquid/vapor, and liquid/solid. In addition, laser-based techniques offer unequalled capabilities for spectral or temporal resolution.

In this article, we present a brief overview of SH spectroscopy for the study of surfaces and interfaces. The article is divided into two principal sections: an introductory section that presents some of the fundamentals of SHG and the behavior at interfaces; and a discussion of the different types of information that the method can yield, as illustrated by some representative examples. A brief section on the experimental technique links these two parts.

Fundamentals of Surface SHG

Basic Notions

Within the domain of classical optics, the response of a material to electromagnetic radiation is described by an induced polarization **P** that varies linearly with the electric field strength, **E**. This situation reflects the fact that the strength of electric fields for light encountered under conventional conditions is minute compared to that of the electric fields binding atoms and solids together. The electric field binding an atom may be estimated as $E_a \sim 1\ \mathrm{V/Å} = 10^{10}\ \mathrm{V/m}$. To reproduce this field strength by an electromagnetic wave requires an irradiance of $I_a \sim 10^{17}\ \mathrm{W/m^2}$, so the validity of this linear approximation in classical optics is clear. The advent of the laser, with its capability for producing light beams with very high optical power and irradiance, has permitted us to probe and exploit the nonlinear response of materials, giving rise to the field of nonlinear optics.

The next-order material response beyond the linear approximation consists of an induced nonlinear source polarization $\mathbf{P}^{(2)} = \varepsilon_0 \boldsymbol{\chi}^{(2)} : \mathbf{EE}$ that depends quadratically on the driving electric field **E**, described by a second-order nonlinear susceptibility $\boldsymbol{\chi}^{(2)}$. For excitation by a single laser at a frequency ω, this response gives rise to SHG at a frequency of 2ω, as well as optical rectification with a dc induced-polarization. For beams at two distinct frequencies of ω_1 and ω_2, the processes of sum- or difference-frequency generation (SFG or DFG) become operative with the production of the new frequencies of $|\omega_1 \pm \omega_2|$. For a relationship of the given form, which assumes only a spatially local quadratic dependence of the polarization on the driving electric field, one can show that $\boldsymbol{\chi}^{(2)}$ must vanish within any centrosymmetric medium. This is a direct consequence of the fact that $\boldsymbol{\chi}^{(2)}$ as a quantity describing a centrosymmetric medium should not change under an inversion operation, while the quantities $\mathbf{P}^{(2)}$ and $\mathbf{E}$ as polar vectors will change sign. The interface specificity of the SHG technique for centrosymmetric media follows from this general observation: only at interfaces is the bulk symmetry broken and the SHG process allowed.

Anharmonic Oscillator Model

Before we embark on our discussion of surface SHG, we first briefly consider a simple classical model of SHG based on an anharmonic oscillator. Although this is not a realistic treatment of the response of materials, it helps to illustrate the characteristic features of the surface SHG process and its spectroscopic attributes. In the anharmonic oscillator model, the medium is treated as a collection of electrons bound to fixed ion cores. While the usual Lorentz model only includes the effect of a linear restoring force, to describe second-order response, we generalize this description to include a leading-order anharmonic restoring force. For simplicity, we consider motion of the anharmonic oscillator in one spatial dimension, in which case the classical equation of motion for the displacement, x, of the electron can then be written as

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + 2\gamma \frac{\mathrm{d}x}{\mathrm{d}t} + \omega_0^2 x - bx^2 = -\frac{e}{m} E(t) \qquad [1]$$

Here $-e$ and m denote, respectively, the electron charge and mass, and $E(t)$ is the optical driving field. The (angular) frequency ω_0 defines the resonance of the harmonic component of the response, and γ represents a phenomenological damping rate for the oscillator. The nonlinear restoring force corresponds to the term containing the parameter b, a material-dependent constant. We are interested in the solution of this equation for a monochromatic driving field at frequency ω with amplitude $E(\omega)$ which we write as

$$E(t) = E(\omega)\exp(-\mathrm{i}\omega t) + \text{c.c.} \qquad [2]$$

An exact solution to eqn [1] in the presence of this electric field is not possible. Let us, however, consider a regime in which we do not excite the oscillator too strongly, and the effect of the anharmonic term remains relatively small compared to the harmonic one. In this case, we may solve the problem perturbatively in powers of the driving field. We find a linear response at the fundamental frequency and a nonlinear response at the SH frequency, which we can write in terms of the corresponding induced dipole moments $p = -ex$:

$$p(t) = p^{(1)}(\omega)\exp(-\mathrm{i}\omega t) + p^{(2)}(2\omega)\exp(-\mathrm{i}2\omega t) + \text{c.c.} \qquad [3]$$

We can then write the amplitudes for the harmonically varying dipole moments as

$$p^{(1)}(\omega) = \alpha^{(1)}(\omega)E(\omega) \qquad [4]$$

$$p^{(2)}(2\omega) = \alpha^{(2)}(2\omega = \omega + \omega)E(\omega)E(\omega) \qquad [5]$$

where $\alpha^{(1)}$, the linear polarizability, and $\alpha^{(2)}$, the second-order polarizability (or first hyperpolarizability) for SHG are given, respectively, by

$$\alpha^{(1)}(\omega) = \frac{e^2}{m}\frac{1}{D(\omega)} \qquad [6]$$

$$\alpha^{(2)}(2\omega = \omega + \omega) = \frac{(e^3/m^2)b}{D(2\omega)D^2(\omega)} \qquad [7]$$

Here we have introduced the resonant response at the fundamental frequency by $D(\omega) \equiv \omega_0^2 - \omega^2 - 2\mathrm{i}\gamma\omega$, with a corresponding relation for the response $D(2\omega)$ at the SH frequency. A similar derivation can, of course, be applied to obtain the corresponding material response for the SFG or DFG processes.

The spectral dependence of the polarizabilities $\alpha^{(1)}$ and $\alpha^{(2)}$ is illustrated in Figure 1 for the anharmonic oscillator model with a resonance frequency of ω_0. In the linear case, one sees the expected behavior of the loss, associated with the $\mathrm{Im}[\alpha^{(1)}]$, and dispersion associated with $\mathrm{Re}[\alpha^{(1)}]$. A single resonance in the linear response is seen for the driving frequency ω near the oscillator frequency ω_0. For the case of SHG, on the other hand, we observe both resonances when the fundamental frequency $\omega \approx \omega_0$ and when the SH

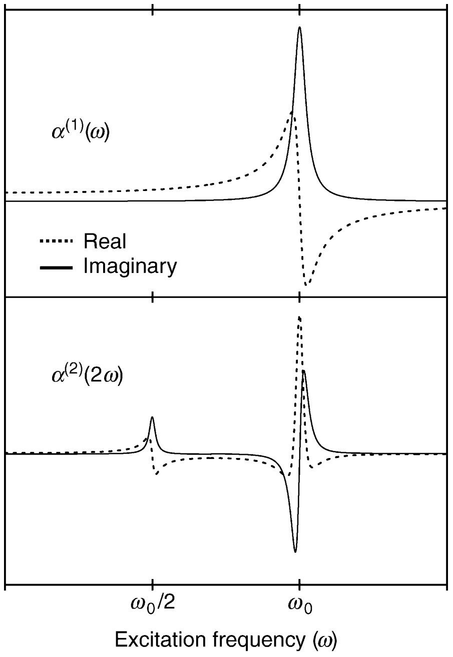

Figure 1 Real and imaginary components of the linear polarizability $\alpha^{(1)}(\omega)$ (top panel) and the second-order nonlinear polarizability $\alpha^{(2)}(2\omega = \omega + \omega)$ (bottom panel) plotted versus frequency according for the anharmonic oscillator model. For the SH response, the resonance features at $\omega = \omega_0$ and $\omega = \omega_0/2$ correspond, respectively, to single-photon and two-photon resonances.

frequency $2\omega \approx \omega_0$. In quantum mechanical language introduced below, these peaks can be considered as arising, respectively, from matching of the fundamental and second-harmonic photon energies with the transition energy in the material.

In SHG spectroscopy, one normally measures the SH power as a function of the laser frequency. The SH field will scale with the nonlinear material response, i.e., with $\alpha^{(2)}(2\omega = \omega + \omega)$, so that the SH power will be proportional to $|\alpha^{(2)}(2\omega = \omega + \omega)|^2$. Figure 2 shows the corresponding spectra in the vicinity of the second-harmonic resonance. An important aspect of this type of spectroscopy, which is also present in other nonlinear spectroscopies, such as coherent Raman measurements, is the role that may be played by a nonresonant background. This may arise either from the off-resonant response of the system under study or from a coherent background from another material. In either case, the effect can be described by adding a spectral flat background response to the nonlinearity. Now when we detect the optical power, we then measure a quantity proportional to $|\alpha^2(2\omega = \omega + \omega) + k|^2$, where k represents the nonresonant background. The presence of k obviously elevates the baseline response near a resonance. In addition, however, depending on the relative phase of k, this extra term can introduce significant changes in the observed lineshapes. Figure 2 illustrates this effect, which must always be borne in mind in the interpretation of experimental spectra.

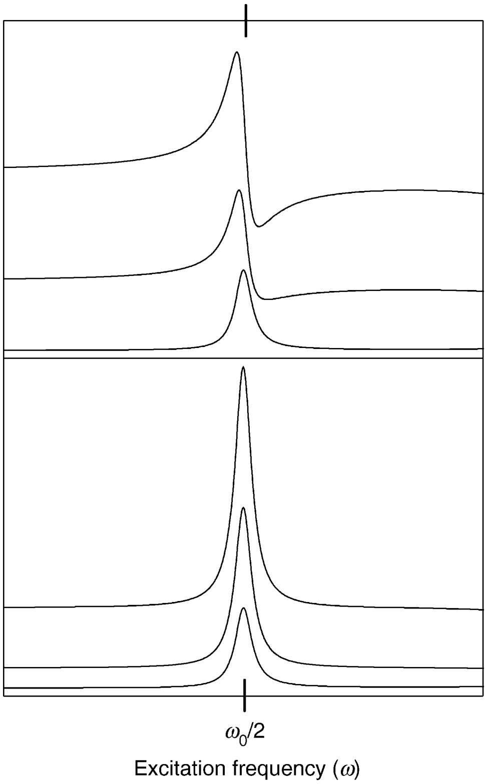

Figure 2 SH spectra $I(2\omega)$ versus pump laser frequency ω corresponding to Figure 1 in the vicinity of the two-photon resonance, with and without the presence of a nonresonant (frequency-independent) background: $I(2\omega) \sim |\alpha^{(2)}(2\omega = \omega + \omega) + k|^2$, where k represents a constant background and parametrized through $k = m|\alpha^{(2)}(2\omega = \omega_0)|$ and m is constant factor. Top panel: $m = 0, 0.5, 1$ (bottom to top curves). Bottom panel: $m = 0, 0.5i, i$ (bottom to top curves). In the top panel, the upper two lineshapes were displaced vertically to distinguish the three curves.

For the purposes of studies of surfaces and interfaces, the key property of second-order nonlinear processes is the fact that they exhibit, for centrosymmetric bulk media, an inherent interface specificity. This feature can be readily understood in the context of the nonlinear oscillator model in which the SH polarization scales with the material parameter b. Since this term is associated with a potential that varies as x^3, it is clear that the value of the b parameter can be taken as a measure of the material's departure from inversion symmetry. At the interface, such a term can be present, reflecting the inherent asymmetry at a boundary, while it must necessarily be absent in the bulk of a centrosymmetric material.

Quantum-Mechanical Description of SHG

A correct description of SHG naturally requires a quantum mechanical treatment of the material response to the optical field. For our present purposes, we consider a localized entity with a nonlinear response, such as a non-centrosymmetric molecule at a surface. We consider below how these individual units can be added together to yield the surface nonlinear response of the material. Within this picture, we write the induced nonlinear dipole moment $\mathbf{p}^{(2)}$ of each molecule in terms of the driving electric field $\mathbf{E}$ as

$$\mathbf{p}^{(2)}(2\omega) = \boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega) : \mathbf{E}(\omega)\mathbf{E}(\omega) \qquad [8]$$

This relation is simply the generalization of eqn [5] to include a full 3-dimensional description, where $\mathbf{p}^{(2)}$ and $\mathbf{E}$ are vectors and the second-order nonlinear polarizability $\boldsymbol{\alpha}^{(2)}$ is a third-rank tensor. The quantum mechanical description enters in how we relate the response function $\boldsymbol{\alpha}^{(2)}(2\omega = \omega + \omega)$ to the underlying properties of the material.

We can obtain an expression for the second-order nonlinear polarizability by application of second-order perturbation theory with the light–matter interaction treated as the perturbation. Within the electric-dipole approximation, the interaction Hamiltonian can be taken as $H_{\text{int}} = -\boldsymbol{\mu}\cdot\mathbf{E}(t)$, where $\boldsymbol{\mu} = -e\mathbf{r}$ denotes the electric-dipole operator and $\mathbf{E}(t)$ the electric field of the driving laser beam. Using the standard density-matrix formalism for second-order perturbation theory, one calculates for the Cartesian components $\alpha^{(2)}_{ijk}$ of the tensor $\boldsymbol{\alpha}^{(2)}$ as

$$\alpha^{(2)}_{ijk}(2\omega = \omega + \omega) = -\frac{1}{\hbar^2}\sum_{g,n,n'}\left[\frac{(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}}{(2\omega - \omega_{ng} + \mathrm{i}\Gamma_{ng})(\omega - \omega_{n'g} + \mathrm{i}\Gamma_{n'g})} + \cdots\right]\rho^{(0)}_g \qquad [9]$$

In this expression, the letters g, n', and n represent energy eigenstates $|g\rangle$, $|n'\rangle$, and $|n\rangle$ of the system, with $\rho^{(0)}_g$ corresponding to the thermal population for differing available ground states g. As illustrated in Figure 3, the SHG process can be regarded as involving a series of three transitions: two transitions associated with the absorption of two pump photons, each of energy $\hbar\omega$, and a transition associated with emission of a second-harmonic photon of energy $2\hbar\omega$. These transitions occur through the electric-dipole operator $\boldsymbol{\mu}$ and are characterized by the matrix elements $(\mu_i)_{gn}$. In eqn [9], the energy denominators involve the energy differences $\hbar\omega_{ng} \equiv E_n - E_g$ and widths $\hbar\Gamma_{ng}$ for transitions between eigenstates $|n\rangle$ and $|g\rangle$, and similarly for other combinations of states. In addition to the term indicated explicitly in eqn [9], there are seven additional terms with different Cartesian coordinates in the matrix elements and/or frequency denominators.

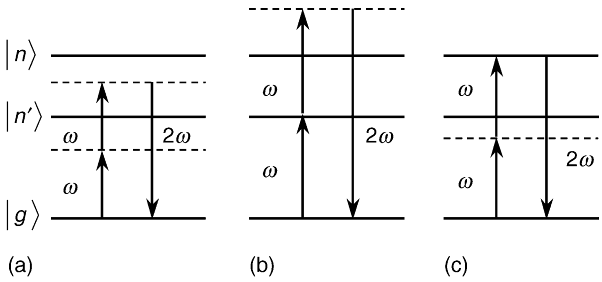

Figure 3 Quantum mechanical description of SHG process for the situations of (a) nonresonant, (b) single-photon resonant, and (c) two-photon resonant interactions. SHG occurs through transition from the initial state electronic state $|g\rangle$ and a first intermediate state $|n'\rangle$, then to a second intermediate state $|n\rangle$, followed by a return to the initial state. The process is just a parametric conversion of two photons at frequency ω into one photon at frequency 2ω.

The frequency denominators in the eight terms of eqn [9] introduce a resonant enhancement in the nonlinearity when either the fundamental frequency ω or the SH frequency 2ω coincides with a transition from a ground state $|g\rangle$ to one of the intermediate states $|n'\rangle$ or $|n\rangle$. Figure 3a shows the situation for a nonresonant nonlinear response, while Figures 3b,c illustrate, respectively, single- and two-photon resonances. The numerator in the perturbation theory expression consists of products of the three dipole matrix elements of the form $(\mu_i)_{gn}(\mu_j)_{nn'}(\mu_k)_{n'g}$. These terms reflect the structure and symmetry of the material that is built into the third-rank tensor $\alpha^{(2)}_{ijk}(2\omega = \omega + \omega)$.

The Surface Nonlinear Response

As a model of the surface nonlinearity, let us consider a monolayer of oriented molecules. The surface nonlinear response accessible to a macroscopic measurement is given, under the neglect of local-field effects, simply by summing the response of the individual molecules. The resulting surface nonlinear susceptibility tensor $\chi^{(2)}_{\mathrm{s},ijk}$ connecting the induced nonlinear sheet polarization to the driving electric field can then be expressed as

$$\chi^{(2)}_{\mathrm{s},ijk} = \frac{N_{\mathrm{s}}}{\varepsilon_0}\langle T_{i\lambda}T_{j\mu}T_{k\nu}\rangle\alpha^{(2)}_{\lambda\mu\nu} \qquad [10]$$

Here N_s denotes the adsorbate surface density, $\alpha^{(2)}_{\lambda\mu\nu}$ the nonlinear polarizability of the molecule expressed in its own coordinate system, $T_{i\lambda}$ the transformation tensor from the molecule's coordinate system to the laboratory frame, $\langle\,\rangle$ an ensemble average over the orientation of the different molecules in the monolayer, and $\varepsilon_0 = 8.85 \times 10^{-12}$ F/m the permittivity of free space. This expression for the surface nonlinear susceptibility has been successful in describing properties of adsorbed molecules, including their coverage, orientation, and spectroscopic features.

For the surfaces of materials such as semiconductors or metals with delocalized electronic states, a somewhat different formulation of the nonlinear response is appropriate. A modified version of eqn [9] can be obtained using the relevant band states. The underlying concepts involved in the surface nonlinear response are, however, similar to those presented for the molecular case.

Regardless of the details of the model describing the surface nonlinear susceptibility, its tensor properties must reflect the symmetry of the interface. In that respect $\chi^{(2)}_{s,ijk}$ is quite analogous to the bulk nonlinear response, $\chi^{(2)}_{ijk}$, in a non-centrosymmetric medium. In the absence of any symmetry constraints, $\boldsymbol{\chi}_s^{(2)}$ will exhibit $3 \times 3 \times 3 = 27$ independent elements for SFG and DFG. For SHG, the order of the last two indices has no significance, and the number of independent elements is reduced to 18. If the surface exhibits a certain in-plane symmetry, then the form of $\boldsymbol{\chi}_s^{(2)}$ will be simplified correspondingly. For the common situation of an isotropic surface, for example, SHG possesses three allowed elements of $\boldsymbol{\chi}_s^{(2)}$ and may be denoted as $\chi^{(2)}_{s,\perp\perp\perp}$, $\chi^{(2)}_{s,\perp\|\|}$, $\chi^{(2)}_{s,\|\perp\|} = \chi^{(2)}_{s,\|\|\perp}$, where $\perp$ corresponds to the direction of the surface normal and $\|$ to an in-plane direction. The nonvanishing elements of $\boldsymbol{\chi}_s^{(2)}$ for other commonly encountered surface symmetries are summarized in several textbooks and reviews listed below.

Radiation Properties for Surface SHG

In order to probe and extract information from interfaces through surface SHG, it is necessary to understand how radiation interacts with the relevant media and gives rise to the experimentally observable signals. Here we present the principal results for the usual case of a spatially homogeneous planar interface. Nonplanar and inhomogeneous geometries, which are also of considerable interest, will be considered briefly under the heading of applications of the SHG technique below.

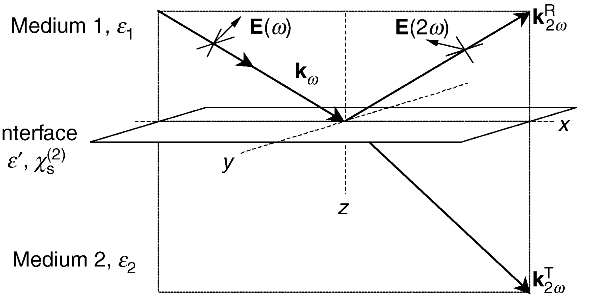

Figure 4 Schematic representation of SHG at the interface between two centrosymmetric media. The pump field $\mathbf{E}(\omega)$ at frequency ω with corresponding wavevector $\mathbf{k}_\omega$ is incident on the interface from medium 1. The SH field $\mathbf{E}(2\omega)$ at frequency 2ω is emitted in directions described by the wavevectors $\mathbf{k}^R_{2\omega}$ (reflected in medium 1) and $\mathbf{k}^T_{2\omega}$ (transmitted in medium 2). The linear dielectric constants of media 1, 2, and the interface are denoted, respectively, by ε_1, ε_2, and ε'. The nonlinear response of the interface ($z = 0$) is given by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}_s^{(2)}$.

The SHG process is coherent in nature. Thus, for the planar geometry, a pump beam impinging on the interface will give rise to a well-collimated SH radiation emerging in distinct reflected and transmitted directions. To describe this situation in a more detailed fashion and provide formulas for the radiation efficiency, we introduce a general description of SHG by a planar interface excited by a plane wave, as shown in Figure 4. The nonlinear response of the interface described by the surface nonlinear susceptibility tensor $\boldsymbol{\chi}_s^{(2)}$ is incorporated through a nonlinear source polarization $\mathbf{P}^{(2)}$ of

$$\begin{aligned}\mathbf{P}^{(2)}(2\omega) &= \mathbf{P}_s^{(2)}(2\omega)\delta(z)\\ &= \varepsilon_0\boldsymbol{\chi}_s^{(2)}(2\omega = \omega + \omega) : \mathbf{E}(\omega)\mathbf{E}(\omega)\delta(z)\end{aligned} \quad [11]$$

localized at the interface ($z = 0$). In addition to the strong nonlinear response at the interface, one must also generally consider the nonlocal nonlinear response of the bulk media. This response is much weaker than the symmetry-allowed interfacial response, but is permitted even in centrosymmetric materials. Since it is present in a much larger volume, however, its cumulative effect, while generally weaker than the surface response, may be of comparable magnitude. For simplicity, however, we neglect this constant background response in our discussion in this article, but a complete discussion can be found in the referenced works. The overall SH response is naturally also influenced by the linear optical properties of the surrounding media. The two bulk media and the interfacial region are characterized, respectively, by frequency-dependent dielectric functions ε_1, ε_2, and ε', as shown in Figure 4. Again for simplicity, and in accordance with most applications, we take the linear response to be isotropic.

Figure 4 shows the incoming pump radiation and the reflected and transmitted SH beams. The directions of these beams are determined by conservation of the component of the momentum parallel to the interface in the SHG process, i.e., $2\mathbf{k}_{\omega,\|} = \mathbf{k}^{\mathrm{R}}_{2\omega,\|} = \mathbf{k}^{\mathrm{T}}_{2\omega,\|}$. Thus the reflected and transmitted beams both remain in the plane of incidence, with directions governed by the so-called nonlinear Snell's law. For bulk media without dispersion between the frequencies of ω and 2ω, the reflected and transmitted beams simply maintain the same angle of incidence as the pump beam. However, for dispersive bulk media, the directions of the beams are, however, altered, as can be understood immediately from the expression $2k_\omega/k_{2\omega} = n_\omega/n_{2\omega}$, where n denotes the refractive index of the medium. The surface SHG process does, nonetheless, differ in an important respect. In contrast to the situation for bulk media, the interfacial region is of negligible thickness, so phase shifts can obviously be disregarded. The issue of phase matching, which is of great significance for SHG from bulk media, consequently does not enter into the surface SHG problem.

By application of the Maxwell equations with the indicated nonlinear source polarization, one can derive explicit expressions for the SH radiation in terms of the linear and nonlinear response of the materials and the excitation conditions. For the case of the nonlinear reflection, the irradiance for the SH radiation can be written as

$$I(2\omega) = \frac{\omega^2 \sec^2\theta |\mathbf{e}'(2\omega)\cdot\boldsymbol{\chi}_{\mathrm{s}}^{(2)} : \mathbf{e}'(\omega)\mathbf{e}'(\omega)|^2 [I(\omega)]^2}{2\varepsilon_0 c^3 [\varepsilon_1(2\omega)]^{1/2} \varepsilon_1(\omega)} \qquad [12]$$

where $I(\omega)$ denotes the irradiance of the pump beam at the fundamental frequency; c is the speed of light in vacuum; θ is the angle of emission of the SH beam with respect to the surface normal. The vectors $\mathbf{e}'(\omega)$ and $\mathbf{e}'(2\omega)$ represent the polarization vectors $\hat{\mathbf{e}}_1(\omega)$, and $\hat{\mathbf{e}}_1(2\omega)$, respectively, after they have been adjusted to account for the linear propagation of the waves to the interface. More specifically, we may write $\mathbf{e}'(\omega) = \mathbf{F}_{1\to2}\hat{\mathbf{e}}_1(\omega)$. Here the $\mathbf{F}_{1\to2}$ describes the relationship between the electric field $E\hat{\mathbf{e}}_1$ in medium 1 (propagating towards medium 2), which yields a field $E\mathbf{e}'$ at the interface. For light incident in the x–z plane as shown in Figure 4, $\mathbf{F}_{1\to2}$ is a diagonal matrix whose elements are $F^{xx}_{1\to2} = 2\varepsilon_1 k_{2,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, $F^{yy}_{1\to2} = 2k_{1,z}/(k_{1,z} + k_{2,z})$, and $F^{zz}_{1\to2} = 2(\varepsilon_1\varepsilon_2/\varepsilon')k_{1,z}/(\varepsilon_2 k_{1,z} + \varepsilon_1 k_{2,z})$, where the quantity $k_{i,z}$ denotes the magnitude of the z-component of the wavevector in medium i at the relevant wavelength (ω or 2ω) (Figure 4).

A few observations may be made about eqn [12], which describes the relationship between the typical experimental observables of the pump irradiance, $I(\omega)$, and the irradiance of the emitted SH reflection, $I(2\omega)$, from the interface. First, one can immediately recognize the quadratic relationship between the pump irradiance and the SH irradiance that is expected for a second-order nonlinear optical effect. Second, the SH radiation is determined not only by the nonlinear optical response, but by the linear optical properties of the relevant media. Consequently, obtaining the surface nonlinear susceptibility tensor $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ from experiment requires knowledge or measurement of the linear response. (It should be noted, however, that the interfacial linear properties always enter into the response in a fixed manner and can be incorporated into an effective surface nonlinear response.) The linear response can significantly enhance or reduce the efficiency of the SHG process. The use of a total-internal reflection geometry, for example, provides a significant enhancement in the SH radiation through the linear optical response of the media. Third, while full knowledge of the linear and nonlinear response will obviously allow one to predict the strength of the SH radiation for any set of conditions, a simple measurement of the SH response does not permit one to determine $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$, even if the linear optical properties are known. The determination of the tensor $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ requires probing the material with different polarization states and/or angles of incidence. Protocols for making these measurements have been discussed extensively in the literature. For the general case of a material with in-plane anisotropy, such as the surface of a crystal, measurements of the response as the crystal is rotated about its surface normal are also very useful in a complete determination of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$. Fourth, the simple measurement of the SH intensity as assumed here obviously eliminates information about the overall phase of the $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$. This phase information is often of considerable interest and importance. For example, if the response of the interface is dominated by an oriented molecular monolayer, the sign of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ will be inverted if the monolayer is reflected through the plane of the surface, i.e., the phase of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ yields information on the polar ordering of the layer. To obtain such phase information, one must employ a scheme using interference with a known SH reference. This approach, which is sketched briefly in the next section, is implemented quite commonly in experiment.

Experimental Considerations

The merits of the SHG technique for selective probing of interfaces are clear. On the other hand, one needs to be able to detect the optical signal associated with a second-order nonlinear response of a material that arises from an effective thickness of just a monolayer or so. Consequently, it is important to consider various factors that aid in optimizing the strength and detectability of the SH radiation. To discuss how best to achieve this end, let us recast eqn [12] in terms of more convenient experimental parameters. We express the SH signal, S, in units of photons/s and describe the laser excitation by its pulse energy E_{p}, pulse width τ, and repetition rate R. For irradiation of an area A of the sample, we then obtain

$$S = \frac{\omega \sec^2 \theta E_{\mathrm{p}}^2 R |\chi_{\mathrm{eff}}^{(2)}|^2}{4\hbar \varepsilon_0 c^3 \tau A} \qquad [13]$$

where $|\chi_{\mathrm{eff}}^{(2)}| \equiv |\mathrm{e}'(2\omega) \cdot \boldsymbol{\chi}_{\mathrm{s}}^{(2)} : \mathrm{e}'(\omega)\mathrm{e}'(\omega)| / [\varepsilon_1^{1/2}(2\omega)\, \varepsilon_1(\omega)]^{\frac{1}{2}}$, whose typical value is of the order of 10^{-21} m^2/V. From this relation, we see that increasing either the pulse energy or repetition rate of the laser will enhance our SH count rate. These parameters are, however, often limited by the characteristics of the available laser, as well as possible damage of the sample. For a given pulse energy and repetition rate, the SH signal is also seen to increase as the laser pulse duration or sample area is decreased. In practice, adequate SH signals can generally be obtained from high-power Q-switched lasers (with nanosecond pulses) with a fairly large focal spot. A preferred solution in most cases is, however, provided by modelocked lasers with high repetition rate and pulses of sub-picosecond duration. Surface SH signals obtained, for example, using a femtosecond mode-locked Ti:sapphire oscillator, are typically in the range of 10^2–10^6 photons/s. Although these signals are relatively weak, they are easily detectable with photomultiplier tubes (PMT) and appropriate electronics.

The basic arrangement for surface SHG measurements is shown in Figure 5. It makes use of the directionality and frequency selectivity of the SHG process not only to detect the desired signal with high efficiency, but to discriminate strongly against background signals, including the reflected laser beam. In the setup, irradiation of the sample involves a pulsed laser source, LS; a halfwave plate, λ, and polarizer, P, for controlling the input polarization. Spectral control is achieved with a filter F1 that passes the fundamental radiation in the pump beam, but rejects other wavelengths, particularly spurious SH radiation arising from other optical components; and a color filter F2 that passes SH, but rejects the reflected fundamental radiation. In addition, a spectrometer is commonly employed to limit the detection bandwidth further. An analyzer A serves to select the desired SH polarization. The SH signal is then measured by a high-quantum efficiency detector and the signal is monitored by processing electronics, typically a photon counter, a gated integrator, or a lock-in amplifier, depending on the strength of the signal and the repetition rate of the laser. The laser source may be either of fixed wavelength or tunable for spectroscopic measurements.

There are other important experimental configurations for surface SHG that extend beyond the basic SH reflection geometry, shown schematically in Figure 6a. A common improvement (Figure 6b) is the implementation of a SH reference in which a small fraction of the pump radiation is directed into a crystal with a bulk second-order nonlinear response. This approach provides a reference against which to compare the measured signal, thus compensating appropriately for pulse-to-pulse variation in duration,

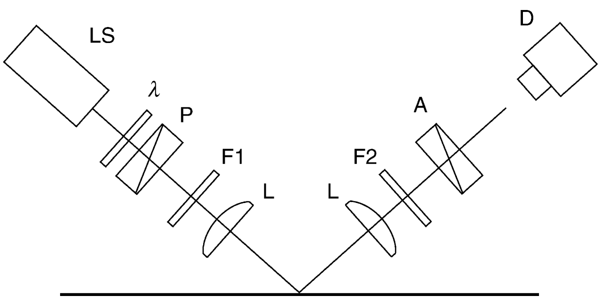

Figure 5 Schematic representation of a basic surface experimental setup for surface SHG. It is comprised of laser source (LS), halfwave plate (λ), polarizer (P), lenses (L), spectral filters (F1, F2), analyzer (A), and detector (D).

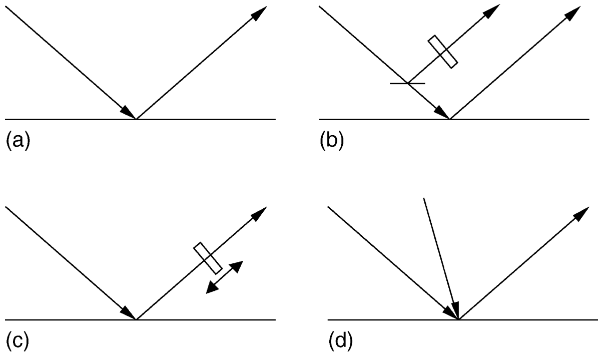

Figure 6 Various schematic representations of different experimental schemes for surface SH measurements. Panel (a) represents the basic scheme as in Figure 5. Panel (b) describes use of a separate SH source as a reference for normalizing the surface SH signals against laser fluctuation and drift. Panel (c) represents the arrangement for a measurement of the phase of the surface SH signal by mixing the fields from the surface and an external source SH radiation. Panel (d) describes the configuration for a time-resolved measurement using the pump-probe scheme.

energy, and spatial profile, as well as for long-term drifts. This type of normalization (against a material with a suitably flat spectral response) is essential for spectroscopy measurements in which the laser frequency is scanned. In addition, use of a calibrated reference is the best way of determining the surface nonlinear response in absolute terms.

It may also be desirable to measure the phase of the signal. For this type of an experiment, a reference field, $\mathbf{E}_{\mathrm{ref}}(2\omega)$, of similar amplitude and known phase response is generated along the path of the surface SH signal $\mathbf{E}(2\omega)$ (Figure 6c). The two collinear SH beams of the same polarization interfere at the detector. The SH power scales as $|\mathbf{E}_{\mathrm{ref}}(2\omega) + \mathbf{E}(2\omega)|^2 = |\mathbf{E}_{\mathrm{ref}}|^2 + 2E_{\mathrm{ref}}E\cos\varphi + |\mathbf{E}|^2$, where φ is the relative phase of the two fields. The relative phase φ may be varied by translating a reference plate along the path of the beam and making use of the dispersion of the ambient air or by inserting another control phase-shifting element. A variant of this phase measurement method is the homodyne scheme in which the reference electric field amplitude $\mathbf{E}_{\mathrm{ref}}(2\omega)$ is significantly stronger than that of the signal amplitude $\mathbf{E}(2\omega)$. The measured SH power then scales as $|\mathbf{E}_{\mathrm{ref}}|^2 + 2E_{\mathrm{ref}}E\cos\varphi$, which yields a signal that varies linearly with the desired SH field strength from the sample.

In the SHG technique, a short-pulse laser is used to interrogate the sample and thus the measurement provides high time-resolution. To make use of this capability, we can add a pump pulse to excite the sample away from equilibrium. This provides the opportunity to examine ultrafast surface dynamics induced by excitation of the sample with a suitable pump laser pulse. For such measurements (Figure 6d), the pump pulse is generally derived from the same laser source as the probe for the SHG measurement, although one or both of these pulses may undergo frequency conversion before being applied to the surface. The critical element is that time synchronism is maintained. The dynamics can then be followed by repeating the SHG measurement at various time delays relative to pump excitation, as controlled by an optical delay line. The ultimate time resolution of this technique is limited only by the duration of the laser pulses and may be as short as a few femtoseconds with state-of-the-art modelocked lasers.

Applications of Surface SHG

The technique of surface SHG is extremely flexible in application, since it relies only on basic symmetry properties for its remarkable surface or interface specificity. SHG measurements can provide quite varied information about the interface, depending on the mode of application of the technique and the material system under study. In this section, we attempt to illustrate the principal types of measurements that can be performed with SHG and the information that can be obtained thereby. For systems with molecular adsorbates, SHG studies can yield information on the density and orientation of molecular species, and can typically do so in real time. For crystalline surfaces, the symmetry and order of the surface can be examined through studies of tensor properties of the surface nonlinear susceptibility. Spectroscopic studies of electronic transitions are also of great value. These can be accomplished for both solids and molecular systems through the same experimental approach of tuning the frequency of the laser source. In all of these investigations, one can achieve very high time resolution, down into the femtosecond range, by the use of pump-probe techniques. In addition, as we discuss below, SHG provides interesting additional capabilities for systems with applied electric and magnetic fields. Finally, the method permits one to probe spatially inhomogeneous systems.

In addition to this diverse range of information that can be provided by surface SHG measurements, the technique is unique in the extremely wide range of material and chemical systems to which it can be – and, indeed, has been – applied. These span the gamut from solids under ultrahigh vacuum to solid/solid, solid/liquid, liquid/vapor, and even liquid/liquid interfaces. Since our examples were drawn to illustrate the type of information obtainable, rather than the range of material systems investigated, the reader is unfortunately obliged to refer to other more detailed treatments of surface SHG measurements to appreciate fully this important facet of the technique.

Adsorbate Density

The remarkably high sensitivity of the SHG technique to the nature of the surface is illustrated by the data in Figure 7, which shows the dramatic change in the SH response from a clean Si(111)-7 × 7 surface upon adsorption of atomic hydrogen. Indeed, it is possible to detect well below 1% of a monolayer (ML) of adsorbed hydrogen. A useful point to note is that the change in $\chi_{\mathrm{s}}^{(2)}$ is linear with adsorbate density for modest values of the coverage, θ. Thus, it is quite easy to obtain a relative gauge of the adsorbate coverage in this regime.

The behavior of the SH response to adsorbates at low coverages can often be described by a simple linearized model of the surface nonlinear susceptibility of $\chi_{\mathrm{s}}^{(2)} = (1 - a\theta)\chi_{\mathrm{s},1}^{(2)} + \theta\chi_{\mathrm{s},2}^{(2)}$, where

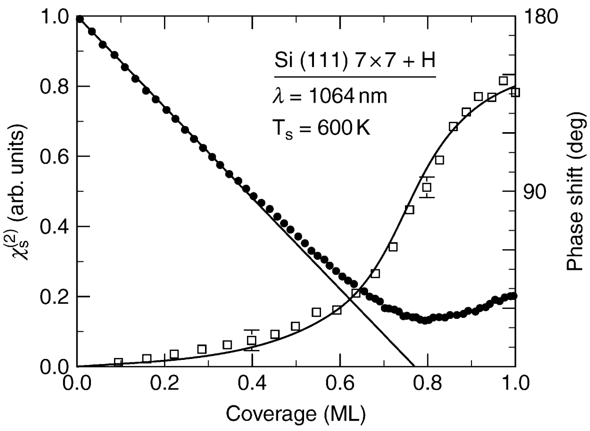

Figure 7 Dependence of the magnitude (full circles) and the phase (open squares) of the nonlinear susceptibility $\chi_s^{(2)}$ for a Si(111)-7 × 7 surface as a function of the coverage of adsorbed atomic hydrogen using a fundamental photon energy 1.17 eV. Reproduced with permission from Höfer U (1996) Nonlinear optical investigations of the dynamics of hydrogen interaction with silicon surfaces. *Applied Physics A* 63: 533.

$\chi_{s,i}^{(2)}$ $(i = 1, 2)$ are complex quantities and a is a constant. This relation describes the surface nonlinear susceptibility as arising from a contribution $\chi_{s,1}^{(2)}$ from the clean surface and a contribution $\chi_{s,2}^{(2)}$ from a saturated monolayer $(\theta = 1)$ of the adsorbate. The term $-a\theta\chi_{s,1}^{(2)}$ then represents the modification of the surface nonlinearity arising from interaction with the adsorbate. For the case of hydrogen adsorption of Figure 7, this is the dominant term, since hydrogen atoms would have no nonlinearity on their own. In other systems, however, the response can be considered as arising essentially from the nonlinear response of aligned adsorbed molecules. The simple linearity of this relation of $\chi_s^{(2)}$ with adsorbate coverage obviously neglects important effects such as adsorbate–adsorbate chemical interactions, as well as local-field effects and the possible influence of differing adsorption sites. All of these factors have been considered in the literature. Nonetheless, the linearity of response with adsorbate coverage is often found to be a good approximation, even for fairly high adsorbate coverages. From the point of view of simply following adsorption dynamics with SHG, it is not, we note, necessary to have any a priori knowledge of the relation between the adsorbate coverage and the SH response. This relation can be established empirically in static measurements and then used advantageously to follow the dynamics of surface processes.

Surface Symmetry and Molecular Orientation

The surface SHG process is described by the third-rank tensor $\boldsymbol{\chi}_s^{(2)}$. The tensorial properties of this response reflect the symmetry of the surface. Other than for a few cases of high symmetry, such a third-rank tensor gives a distinct signature of surfaces with crystalline order compared to surfaces, such as disordered ones, exhibiting effective in-plane isotropy. The form of the $\boldsymbol{\chi}_s^{(2)}$ tensor can be probed both by measurements that vary the polarization of the fundamental and second-harmonic beams and by measurements in which the plane of incidence is changed with respect to the orientation of the surface, generally by rotating the sample about its normal with a fixed experimental geometry.

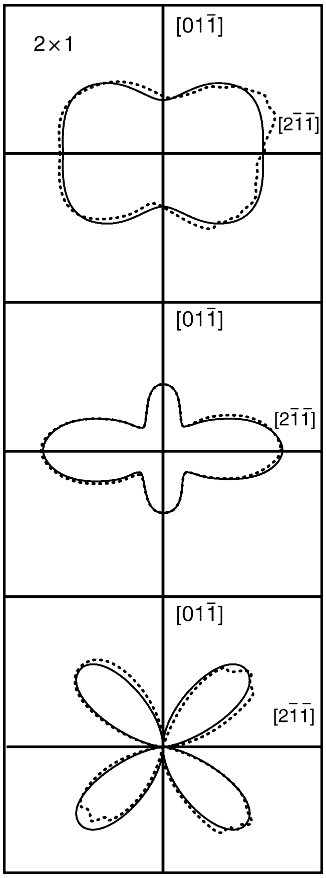

Figure 8 Polarization dependence of SHG for normal incidence excitation of the cleaved Si(111)-2 × 1 surface. The polar plot of the SH signal versus the pump polarization orientation is shown for the case of no polarization analysis (top panel), for the SH analyzer along the [2$\bar{1}\bar{1}$] crystallographic direction of the surface (middle panel), and along the [01$\bar{1}$] direction (bottom panel). The solid curves are fits to theory assuming that the surface has *m* symmetry. Reproduced with permission from Heinz TF, Log MMT and Thomson WA (1985) Study of Si(111) surfaces by optical second-harmonic generation: reconstruction and surface phase-transformation. *Physical Review Letters* 54: 63.

Figure 8 illustrates the capability of surface SHG measurements to probe surface symmetry. The figure shows SHG polarization dependence for a Si(111) surface cleaved in ultrahigh vacuum. While the underlying bulk symmetry suggests that the surface might exhibit 3m symmetry, the analysis of data indicates that only a single mirror *m* plane is present. This lowered symmetry corresponds to the metastable 2×1 reconstruction induced by the crystal cleavage. By heating the sample, one can observe a phase transformation to the equilibrium 7×7 reconstruction and the emergence of the corresponding 3m surface symmetry.

The analysis of surface or interface symmetry by SHG relies only on the form of the nonlinear susceptibility tensor, i.e., which elements are independent and nonvanishing. Within the context of a microscopic model of the surface response, however, the relative values of the tensor elements of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ can also yield valuable information about the nature of the surface. This approach has been developed extensively in the context of the analysis of molecular orientation at surfaces. The basic notion behind this treatment relies on the knowledge of the microscopic molecular nonlinear polarizability $\alpha_{\lambda\mu\nu}^{(2)}$ and the measured surface nonlinear susceptibility $\chi_{\mathrm{s},ijk}^{(2)}$ to infer information about the orientational average that relates these quantities. As discussed above in conjunction with eqn [10], under the neglect of local-field corrections, the relationship between these two quantities involves the ensemble average of the coordinate transformation tensor from the molecular to the laboratory coordinate systems $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Hence a knowledge of the quantities $\alpha_{\lambda\mu\nu}^{(2)}$, N_{s}, and $\chi_{\mathrm{s},ijk}^{(2)}$ allows us to infer information about the orientational distribution of the molecules as described by the quantity $\langle T_{i\lambda} T_{j\mu} T_{k\nu} \rangle$. Even in the absence of knowledge of the magnitudes of the nonlinear polarizabilities and the adsorbate density, one may still obtain useful information for molecules with a simple form of $\alpha_{\lambda\mu\nu}^{(2)}$.

An important example of this analysis of molecular orientation concerns the case of molecules with a dominant nonlinear polarizability along a single axis, i.e., $\boldsymbol{\alpha}^{(2)} = \alpha_{z'z'z'}^{(2)} \hat{\mathbf{z}}' \hat{\mathbf{z}}' \hat{\mathbf{z}}'$. For the usual situation of in-plane isotropy, one can then show that the nonlinear susceptibility elements are given by $\chi_{\mathrm{s},\perp\perp\perp}^{(2)} = N_{\mathrm{s}} \langle \cos^3 \theta \rangle \alpha_{z'z'z'}^{(2)} / \varepsilon_0$ and $\chi_{\mathrm{s},\|\perp\|}^{(2)} = \chi_{\mathrm{s},\|\|\perp}^{(2)} = \frac{1}{2} N_{\mathrm{s}} \langle \cos\theta \times \sin^2\theta \rangle\ \alpha_{z'z'z'}^{(2)} / \varepsilon_0$, where θ denotes the angle between the surface normal and the z' molecular axis. It should be noted that these orientational moments, reflecting their origin in a third-rank tensor, give a signature of the polar ordering, since a sign change in the elements of $\boldsymbol{\chi}_{\mathrm{s}}^{(2)}$ follows if the molecular orientation is inverted. Also, when only one element of $\boldsymbol{\alpha}^{(2)}$ is present, we can form ratios of the experimentally measurable ratio of tensor elements that depend only on the molecular orientational factors, such as $R \equiv (2\chi_{\mathrm{s},\|\perp\|}^{(2)} + \chi_{\mathrm{s},\perp\perp\perp}^{(2)}) / \chi_{\mathrm{s},\perp\perp\perp}^{(2)} = \langle \cos\theta \rangle / \langle \cos^3\theta \rangle$. For a narrow distribution of the orientations centered at θ_0, this ratio yields directly the molecular orientation angle through $R \approx \sec^2 \theta_0$.

This type of molecular orientation analysis has been applied to a wide range of material systems. The method, while powerful, relies on several assumptions, including an adequate knowledge of the molecular nonlinear polarizability, the lack of significant local field corrections, the availability of information about the effective linear dielectric constant in the interfacial region, and a suitable method of eliminating any substrate nonlinear response in the measurement. The method is, consequently, best suited for adsorbed layer of molecules with strong and well-defined nonlinear response, present at relatively modest coverages. The use of resonant excitation is a valuable adjunct in these studies, since it generally simplifies the form and strengthens the molecular nonlinear response. An analogous approach to the analysis of molecular orientation can be accomplished with infrared–visible sum-frequency generation. These measurements often provide a superior method of analysis, since the molecular response and its symmetry are better defined by resonant excitation of a molecular vibration.

Surface and Interface Spectroscopy

Surface SHG can be readily adapted to yield spectroscopic information on surface transitions. As indicated above, both resonances at the fundamental and second-harmonic frequencies are manifested in the SHG response. The examination of resonances at the SH frequency is often convenient, since it permits the application of a relatively intense pump radiation at a frequency well separated from that of the resonance. Such SHG spectroscopy measurements have been widely applied to examine electronic transitions at surfaces and interfaces of solids, as well as in adsorbed molecules under a wide variety of conditions.

An example of SHG spectroscopy of electronic transitions at a buried solid/solid interface is presented in Figure 9. For the epitaxial CaF_2/Si(111) interface, distinctive new electronic transitions, associated with interface states, are observed that are absent from the response of either of the bulk materials. These excitations occur in a spectral range where the bulk silicon sample is highly absorbing, which would render the identification of the new interface transitions very difficult using conventional linear optics. The buried nature of the interface naturally also constrains application of conventional surface probing techniques involving electron

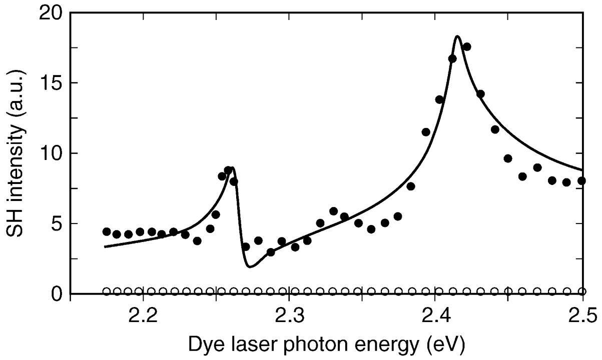

Figure 9 SHG spectrum (solid dots) for the CaF_2/Si(111) interface. For comparison, the open circles represent the corresponding signals for the SiO_2/Si(111) interface, which is seen to give a very weak signal in this frequency range. The solid line is a theoretical fit based on a two-dimensional direct transition of the interface states, together with an excitonic feature slightly below the band gap. Reproduced with permission from Heinz TF, Himpsel FJ, Palange E and Burstein E (1989) Electronic transitions at the CaF_2/Si(111) interface probed by resonant three-wave-mixing spectroscopy. *Physical Review Letters* 63: 644.

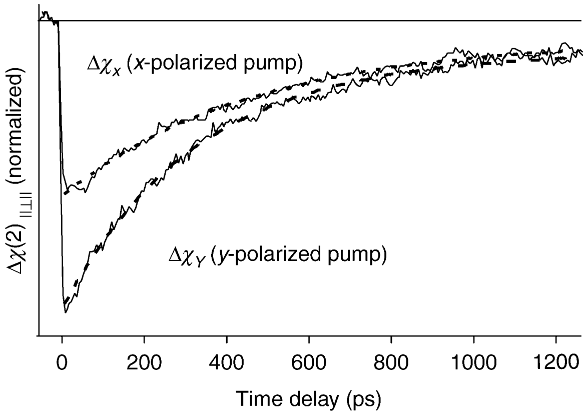

Figure 10 Rotational relaxation of Coumarin 314 molecules at the air–water interface subsequent to photoexcitation. The initial anisotropy in the molecular orientational distribution is created by linearly polarized pump radiation in two orthogonal directions in the surface. Reproduced with permission from Zimdars D, Dadap JI, Eilsenthal KB and Heinz TF (1999) Anisotropic orientational motion of molecular adsorbates at the air–water interface. *Journal of Physical Chemistry B* 103: 3425–3433.

irradiation or emission. Two additional general observations are appropriate. First, the lineshape analysis for such measurements must take into account the fact that the SHG response derives from both the real and imaginary parts of $\chi_s^{(2)}$, as we discussed above. Second, in the absence of any other knowledge, SHG spectroscopic data contain an ambiguity about whether the resonances correspond to the fundamental or SH photon energy. This ambiguity, however, can be resolved in a straightforward fashion by complementing the SHG measurement with a sum-frequency measurement in which one frequency is scanned and one is held fixed, as was demonstrated in the study associated with Figure 9.

Time-Resolved Measurements

One of the fascinating frontiers that can be addressed by SHG measurements is probing the ultrafast dynamics of surfaces and interfaces by the pump-probe method. This possibility is exemplified by a variety of studies ranging from the dynamics of laser-driven phase transitions, charge-transfer, solvation, phonon dynamics, and orientational and torsional dynamics.

An example of applying time-resolved SHG to examine ultrafast molecular dynamics at the liquid/vapor interface of water is presented in Figure 10. In these studies, the rotational motion of dye molecules at the interface, which probes the local environment, is followed in real time. This is accomplished by photo-exciting the dye molecules at the interface. Because polarized pump radiation is used, an anisotropic orientational distribution of excited-state molecules is established. Since the excited molecules have a different SHG response, a change in the signal is observed immediately. Subsequently, the anisotropic orientational distribution relaxes to the equilibrium configuration. The time evolution of the rotational anisotropy is followed by detecting the SH of a probe laser pulse as a function of the delay time. Through a comparison of the results for different initial anisotropic distributions (produced by pump beams with differing polarizations), one may deduce rates for both in-plane and out-of-plane orientational relaxation. This example illustrates the possibilities of time-resolved SHG measurements for probing the detailed dynamics of molecules at an interface on the ultrafast timescale.

Spatially Resolved Measurements

Up to this point, our discussions have centered on spatially homogenous surfaces and interfaces. Surface SHG has also been exploited to probe the lateral variation of inhomogeneous surfaces with the micron to sub-micron resolution characteristic of optical microscopy techniques. Spatial resolution may be achieved simply by detecting the nonlinear response with a focused laser beam that is scanned across the surface. A large area of the surface may also be illuminated and the emitted nonlinear radiation imaged. A wide range of applications of this imaging capability has emerged, including the probing of biological materials, such as cells and tissues, to probing spatially varying electric fields and magnetic domains. Within the biological context, SH microscopy has recently become an active research direction. In these applications, while probing with SHG typically senses non-centrosymmetric structures in bulk samples as opposed to interfaces, most of the aspects of SHG presented in this article still apply. To further improve the spatial resolution of SHG microscopy, researchers have also employed near-field techniques for spatial resolutions below the diffraction limit.

Complementary approaches to the real-space imaging just described have also been developed based on SH diffraction. In this scheme, periodically modulated surfaces are established holographically on a surface by utilizing two interfering laser beams. In addition to the reflected and transmitted SH beams, additional beams are then generated at well-defined angles as dictated by the grating period. This approach has been used to monitor the relaxation of a monolayer grating of adsorbates in time, which yields precise information on the rate of adsorbate surface diffusion.

Probing Electric Fields at Interfaces

For centrosymmetric media, the SHG process shows a high degree of sensitivity to the presence of electric fields. This property follows from the fact that an applied electric field can break the inversion symmetry, just as an interface does. The resulting process, termed electric-field induced SHG (or EFISH), has been known for many years. From a phenomenological standpoint, it can usually be described as a nonlinear response of the form $\mathbf{P}^{(3)}(2\omega) = \varepsilon_0 \boldsymbol{\chi}^{(3)} : \mathbf{E}(\omega)\mathbf{E}(\omega)\mathbf{E}$, where $\mathbf{E}$ is the electric field being probed and $\mathbf{E}(\omega)$ is the laser pump field. Within the context of surfaces and interfaces, there are many circumstances where probing such electric fields is of interest and importance. The SHG technique provides high sensitivity, good spatial resolution, accessibility to buried interfaces, and, in the pump-probe implementation, extremely high time resolution. In addition, with appropriate measurements of the polarization dependence and homodyne detection, the full vector character of the electric field can be determined.

The behavior of electric-field induced SHG is similar to that arising from the asymmetry of an interface. Indeed, it is formally equivalent to the response of an interface provided that the electric field is present in (or radiates from) a region of less than a wavelength's thickness. The applications of this measurement capability are quite diverse, ranging from studies of charging and acid/base reactions at liquid/solid interfaces to probing fields in metal-oxide-semiconductor structures and Schottky barriers to the measurement of freely-propagating terahertz radiation. In Figure 11, SHG with appropriate polarization control and homodyne detection scheme has been applied to the measurement of the vector character of electric fields present in a dc-biased dipole structure on silicon. By spatially scanning the laser beam, a two-dimensional electric field map can be produced. In a pump/probe scheme, the same measurements can be made with femtosecond time resolution.

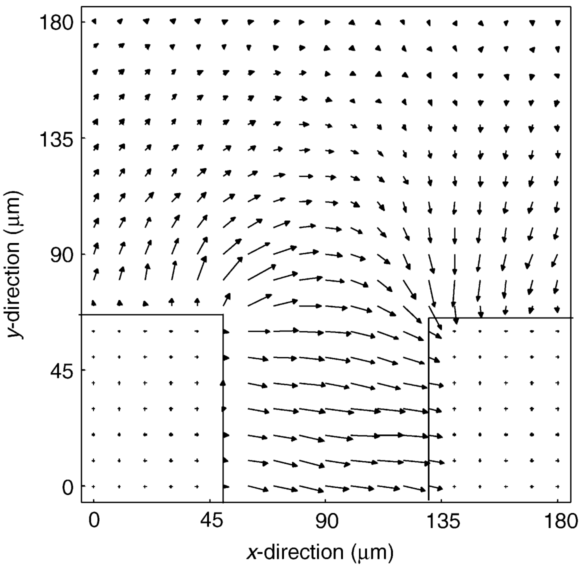

Figure 11 Vector map of the electric field in Si-on-sapphire substrate as deduced by SHG measurements. A voltage is imposed on the rectangular metallic electrodes, which are separated by 80 μm. Reproduced with permission from Dadap JI, Shan J, Weling AS, Misewich JA, Nahata A and Heinz TF (1999) Measurement of the vector character of electric fields by optical second-harmonic generation. *Optics Letters* 24: 1059–1061.

Probing Magnetization at Interfaces

In contrast to the application of an external electric field (a polar vector), the presence of an external magnetic field (an axial vector) does not break the inversion symmetry of a material. Thus, SHG retains its interface selectivity for magnetic effects. SHG measurements, as an interface-specific probe, thus provide an excellent complement to the linear magneto-optical Kerr effect (MOKE), which is sensitive to magnetic effects within the penetration depth of the light.

The ability of the SHG technique to sense the magnetization at a surface is illustrated in Figure 12 for a film of epitaxially-grown magnetic Co on nonmagnetic substrate Cu. A hysteresis loop of the magnetization versus the strength of an applied magnetic field is shown, where the magnetization is gauged from the SHG response. Through measurements of the response as a function of the thickness of

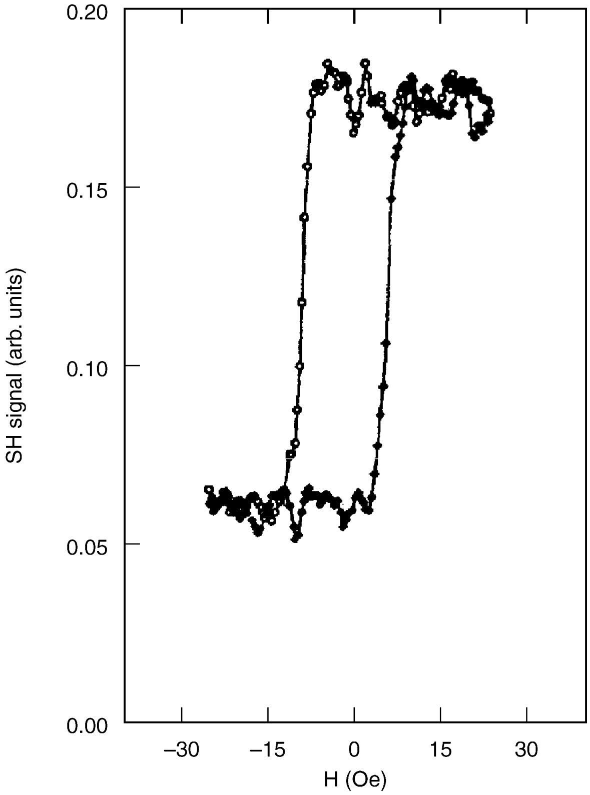

Figure 12 SH response of a Co film on a Cu(001) substrate as a function of the applied magnetic field. A hysteresis loop corresponding to the surface response is observed. Reproduced with permission from Wierenga HA, de Jong W, Prins MWJ *et al.* (1995) Interface magnetism and possible quantum-well oscillations in ultrathin Co/Cu films observed by magnetization induced second harmonic generation. *Physical Review Letters* 74: 1462–1465.

the magnetic film, the researchers demonstrated that the SH response reflected the interface properties, rather than those of the bulk film.

Probing Micro- and Nanoscale Structures

In addition to the microscopy studies described above that probe spatially inhomogeneous surfaces, there has been considerable interest in SHG from a variety of strongly textured surfaces and small particles. From the standpoint of the optical response of materials, particular interest has been attached to studies of metallic surfaces with surface roughness. For appropriate roughness of low-loss materials, such as silver, very significant local-field enhancements can be achieved. Since these field enhancements enter into the strength of the SH emission nonlinearly, overall enhancements in the SH power by several orders of magnitude have been demonstrated. This provides both a useful method of study of such surface enhancement effects and a scheme for amplifying the relatively weak surface nonlinear response.

Beyond these problems associated with roughened surfaces, considerable attention has recently been directed towards the application of the SHG technique for probing micro- and nano-scale particles. The fundamental principle of the interface selectivity of the SHG process for such systems comprised of centrosymmetric materials remains unaltered. The interaction of these particles with the radiation fields is, however, significantly modified. For the case of particles with dimensions of several optical wavelengths, the overall SHG signal strength may be comparable to that from a corresponding area of a planar surface. The radiation pattern will, of course, be strongly modified and will generally be diffuse in character. While this makes detection (and background discrimination) somewhat more difficult than for the planar geometry, many SHG measurements of surface and interface properties have been carried out in this regime. As the particle size becomes progressively smaller, particles with overall inversion symmetry begin to exhibit appreciable cancellation of the SH emission. Indeed, in the limit of vanishing size, full cancellation is expected. This tendency is, however, offset by the increasing surface-to-bulk ratio and the possibility of increasing the number of particles probed. It has been demonstrated experimentally that the surfaces of particles in the nanometer length scale can indeed be probed by SHG.

From the point of view of applications, this extension of the SHG technique to probe micro- and nano-structured materials opens up a rich variety of new possibilities. Results have been reported on metal and semiconductor nanoparticles in suspension, as well as on colloids of minerals. In addition, liquid droplets have been investigated. Self-assembling structures such as liposomes and vesicles have been examined. For these structures, many of the same capabilities of SHG as demonstrated for the planar geometry, such as probing of adsorption and charging effects, have been demonstrated. More uniquely, molecular transport *through* the liposome membrane could also be observed and characterized. The possibility of extension of these measurements to biological systems is of obvious interest and promise.

Concluding Remarks

In this article, we provided a brief review of surface SHG as a technique for interface-specific probing of materials. In addition to discussing some of the fundamental ideas underlying the method, we outlined the wide variety of information that may be obtained from such measurements, ranging from surface symmetry and molecular ordering to interface spectroscopy and ultrafast dynamics. While several aspects of the method have now become quite well established, developments in widely available mode-locked lasers with excellent performance characteristics have greatly enhanced the ease of making surface SHG measurements in a wide class of material systems. In addition, new developments and applications of the method have continued to emerge apace. Of particular note in this regard are applications to probe electric fields and magnetization at interfaces, as well as advances in the use of the method for interface-specific probing of micro- and nanoscale structures. The sharp increase in applications of SHG for probing systems of biological importance constitutes another important development, and one with much promise for the future.

See also

Materials for Nonlinear Optics: Organic Nonlinear Materials. **Nonlinear Optics, Basics:** $\chi^{(2)}$–Harmonic Generation. **Physical Applications of Lasers:** Sum-Frequency Generation at Surfaces.

Further Reading

Bennemann KH (ed.) (1998) *Nonlinear Optics in Metals*. Oxford, UK: Clarendon.

Bloembergen N (1965) *Nonlinear Optics*. New York: WA Benjamin.

Boyd RW (1992) *Nonlinear Optics*. New York: Academic.

Brevet PF (1997) *Surface Second Harmonic Generation*. Lausanne, Switzerland: Presses Polytechniques et Universitaires Romandes.

Butcher PN and Cotter D (1990) *The Elements of Nonlinear Optics*. Cambridge, UK: Cambridge.

Corn RM and Higgins DA (1994) Optical second-harmonic generation as a probe of surface-chemistry. *Chemical Review* 94: 107–125.

Dadap JI and Heinz TF (2001) Nonlinear optical spectroscopy of surfaces and interfaces. In: Moore JH and Spencer ND (eds) *Encyclopedia of Chemical Physics and Physical Chemistry*, pp. 1089–1125. Bristol, UK: IOP.

Eisenthal KB (1996) Liquid interfaces probed by second-harmonic and sum-frequency spectroscopy. *Chemical Review* 96: 1343–1360.

Heinz TF (1991) Second-order nonlinear optical effects at surfaces and interfaces. In: Ponath H-E and Stegeman GI (eds) *Nonlinear Surface Electromagnetic Phenomena*, pp. 353–416. Amsterdam: North Holland.

Liebsch A (1995) Electronic excitations at metal surfaces. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 479–532. Amsterdam: Elsevier.

Lüpke G (1999) Characterization of semiconductor interfaces by second-harmonic generation. *Surface Science Report* 35: 75–161.

McGilp JF (1995) Optical characterisation of semiconductor surfaces and interfaces. *Progress in Surface Science* 49: 1–106.

McGilp JF (1999) Second-harmonic generation at semiconductor and metal surfaces. *Surface Review Letters* 6: 529–558.

McGilp JF, Patterson CH and Weaire DL (eds) (1995) *Epioptics: Linear and Nonlinear Optical Spectroscopy of Surfaces and Interfaces*. Berlin, New York: Springer.

Prasad PN and Williams DJ (1991) *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. New York: Wiley.

Reider GA and Heinz TF (1995) Second-order nonlinear optical effects at surfaces and interfaces: recent advances. In: Halevi P (ed.) *Photonic Probes of Surfaces*, pp. 413–478. Amsterdam: Elsevier.

Richmond GL (1997) Vibrational spectroscopy of molecules at liquid/liquid interfaces. *Analytical Chemistry* 69: A536–A543.

Richmond GL, Robinson JM and Shannon VL (1988) Second harmonic generation studies of interfacial structure and dynamics. *Progress in Surface Science* 28: 1–70.

Shen YR (1984) *The Principles of Nonlinear Optics*. New York: Wiley.

Shen YR (1989) Optical second harmonic-generation at interfaces. *Annual Review of Physical Chemistry* 40: 327–350.

Single Molecule Spectroscopy

X Michalet and S Weiss, University of California, Los Angeles, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Textbooks in physics, chemistry, and biology abound in diagrams depicting single-particle, single-molecule or single-enzyme properties, reactions or interactions with their environment. Such pictures and the numbers associated with them, however, were, until recently, extrapolated from measurements performed on large ensembles of molecules, which only yield quantities averaged out over time, space, conformations and/or the local environment. Advances in detector sensitivity and improvements in instrument design do now provide scientists with tools to probe single molecules with light, and monitor their photophysical properties with exquisite sensitivity and time resolution. This spectroscopic information is used to explore the molecular structure, conformational dynamics, local environment, and intermolecular interactions. The collected data generate a wealth of new information on single-molecule physical chemistry or biomechanics, but also in polymer gel physics or glass dynamics and other domains of solid state physics.

This article reviews the field of single-molecule spectroscopy. First, the principles underlying single-molecule spectroscopy approaches are introduced, and this is followed by a discussion of the different spectroscopic techniques available to the experimentalist and what kind of information can be extracted with each of them. It also discusses important photophysical properties of molecules that influence the outcome of single-molecule spectroscopic measurements. The next section presents some applications of these techniques, focusing on physical properties of materials and biological applications of single-molecule spectroscopy. The last part briefly reviews recent developments and outlines the future prospects of single-molecule spectroscopy.

Principles of Single-Molecule Spectroscopy

Motivation

Measurements performed on ensembles of molecules only yield averaged-out information. This averaging process is of several kinds: average over molecular heterogeneity (e.g., chemical composition, oxidation state, or spatial conformation) and average over different or changing local environments. Moreover, dynamic processes studied by ensemble measurements cannot sort out molecules exhibiting fast or slow reaction times, although such differences are expected from the stochastic nature of any chemical kinetics. The interest in single-molecule analysis is for these reasons manifold. At low temperature, being able to study a single molecule and its evolution under the influence of various external parameters allows for precise tests of molecular quantum electrodynamics. At higher temperature, it gives access to the temporal evolution of different nanodomains of a bulk material in which single molecules are embedded as reporters of their environment. In biology and biochemistry, the building blocks of any process are individual molecules (genes, enzymes, motors, filaments, etc.); the ability to fully understand their kinetic behavior at the single-molecule level is thus of utmost importance. Finally, reliable observation of single molecules is a prerequisite of the full development of nanotechnology and quantum computing.

Historical Outlook

Single-molecule experiments were first performed on isolated ion channels by the patch clamp technique, which allows the recording of single-ion translocation through a pore-like protein embedded in a fluid membrane. Scanning tunneling and atomic force microscope observations later opened the way to investigations of nanometer-scale objects on surfaces. Only recently have optical methods reached the sensitivity required to detect, image, manipulate, and follow the spectroscopic evolution of single molecules on surfaces, in solids, liquids, and biological membranes as well as inside live cells.

Single-molecule spectroscopy (SMS) was initially demonstrated in a system of pentacene molecules embedded in a *p*-terphenyl host crystal at liquid helium temperature. Absorption (leading to fluorescence excitation) spectroscopy in this system takes advantage of the very narrow zero-phonon absorption line (ZPL) of the rigid pentacene molecule (Figure 1). The position in laser-frequency space of each molecule's ZPL indeed depends on the local fields, allowing for precise molecular selectivity. However this approach is limited to a few molecule–host pairs and requires cryogenic techniques, due to broadening of the ZPL above 10 K.

Room temperature studies started with total internal reflection (TIR) microscopy and standard flow-cytometry methods in the mid-1980s. In the latter case, single molecules randomly passing through the excitation volume gave rise to sudden bursts of photons emitted during their brief transit. Initially developed for rapid DNA sequencing, this simple approach is now widely used for more sophisticated studies in its confocal version (Figure 2), which will be detailed in the following.

The first image of a single molecule at room temperature was later obtained using scanning near-field optical microscopy (Figure 3), rapidly followed by similar achievements using far-field confocal microscopy, at the expense of a lower spatial resolution, but with much more ease of use. TIR and wide-field fluorescence imaging are now used to image and study the two-dimensional and three-dimensional diffusion trajectories of single fluorescent molecules (Figure 4), making a come-back after the pioneering studies of the early 1980s.

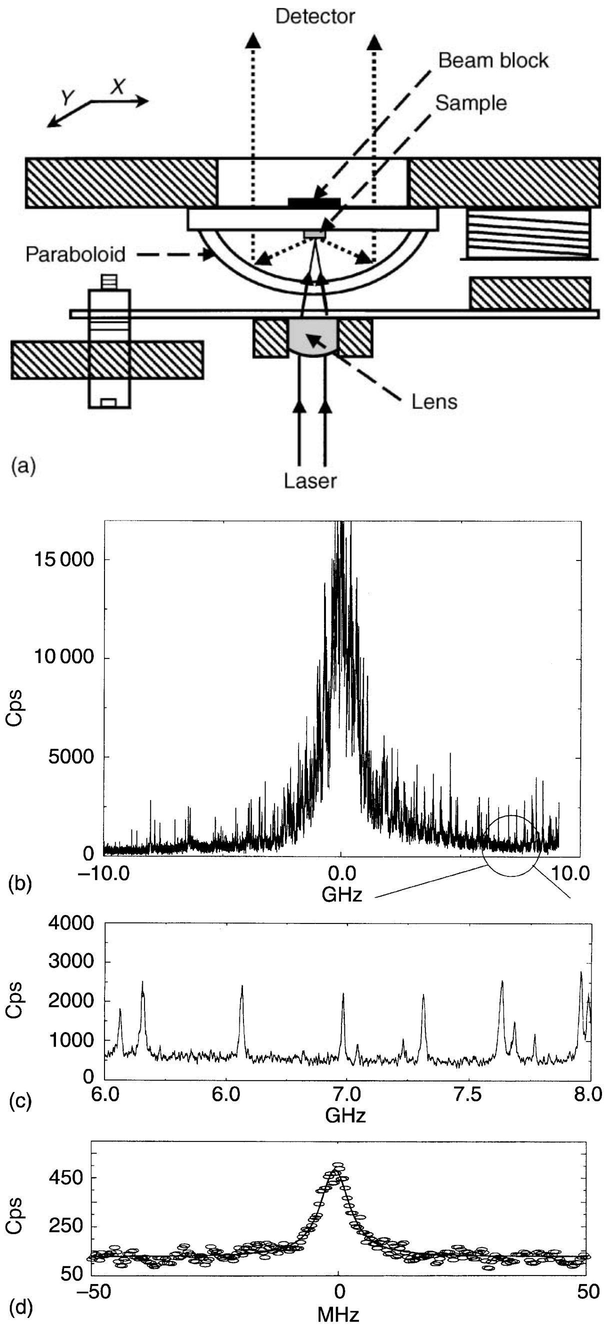

Figure 1 Fluorescence excitation spectra of pentacene in *p*-terphenyl at 1.5 K. (a) Schematic of the cryogenic setup used. A laser beam is focused onto the sample through a lens into the cryogenic chamber. The sample can be moved in the *XY* plane, with a magnet and coil pair. A beam block rejects the incident laser light (black arrow). A paraboloidal mirror collects the emitted fluorescence (dashed arrows) and redirects it towards the detector (photomultiplier tube). The laser frequency (linewidth 3 MHz) is tuned around the absorption line center at 592.321 nm. (b) Broad scan showing the inhomogeneously broadened line shape of pentacene molecules embedded in a *p*-terphenyl crystal at 1.5 K. The sharp lines correspond to spectra of individual molecules (see panel c) (cps: counts per second). (c) Detail of the previous spectrum showing spectra of several individual molecules. The variability of intensities is due to the different positions of the molecule in the excitation beam. (d) Individual molecule spectrum obtained at lower excitation power exhibiting a Lorentzian shape (solid line) centered at 592.407 nm with a lifetime-limited width of 7.8 MHz (intensity is measured in counts-per-second for all three spectra). (a): adapted from Ambrose WP, Basché T and Moerner WE (1991) Detection and spectroscopy of single pentacene molecules in a *p*-terphenyl crystal by means of fluorescence excitation. *Journal of Chemical Physics* 95: 7150–7163. (b)–(d): after Moerner WE (1994) Examined nanoenvironments in solids on the scale of a single, isolated impurity molecule. *Science* 265: 46–53. Copyright 1994, The American Association for the Advancement of Science.

Requirements for Single-Molecule Sensitivity

Observing single molecules may be as easy as finding a needle in a haystack if background, signal and noise are not carefully optimized. Simple considerations illustrated in Figure 5a help measure the technical challenge bearing on single-molecule optical detection. Whatever the type of detected signal (fluorescence, Raman scattering, or others), it is characterized by an absorption cross-section σ (probability of photon absorption) and a quantum yield Q (number of emitted photons per absorbed photon). The optical setup transmits an excitation power P at a frequency ν, resulting in a signal rate $s = EQ\sigma P/(Ah\nu)$ expressed in counts per second (or Hz). This rate is proportional to P/A where A is the cross-section of the excitation beam in the plane of the molecule. E is the collection efficiency of the setup. The environment adds a background rate b per unit volume and unit power, and the detector contributes a dark count rate D. Neglecting noise from the readout electronics, a measurement of duration τ will yield the following signal-to-noise ratio (SNR):

$$\mathrm{SNR} = \frac{s\tau}{\sqrt{(s + bVP + D)\tau}} \qquad [1]$$

where V is the excited volume. The denominator is the root mean square of all contributions to the measured count rate, considered shot-noise limited. The SNR obviously increases at larger integration time τ. Another important quantity is the signal-to-background ratio (SBR):

$$\mathrm{SBR} = \frac{EQ\sigma}{bVAh\nu} \qquad [2]$$

These ratios can be increased by improving the collection efficiency E or decreasing the excitation volume V. A larger excitation power or a longer integration time will improve the SNR without affecting the SBR. Typical values for the above

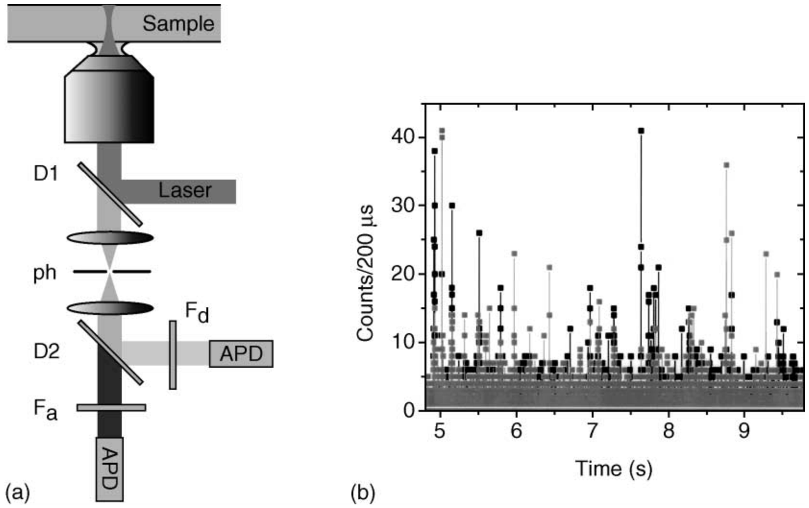

Figure 2 Single-molecule fluorescence burst detection. (a) Schematic of the confocal setup used for single-molecule fluorescence detection. A collimated laser beam is focused through a high numerical aperture (NA) objective in a liquid cell. Molecules diffusing through the diffraction-limited excitation volume (~1 fl) emit photons during their passage, which are collected by the same objective. A dichroic mirror (D1) separates excitation and emission wavelengths. The emitted light is focused on a pinhole (ph) in order to reject out-of-focus background light, and finally recollimated onto two different avalanche photodiodes (APDs) after spectral separation by a dichroic mirror (D2). Filters specific for each channel (F_a: acceptor channel, F_d: donor channel) may be used to improve spectral separation. (b) 5 s time trace of recorded bursts for a 30 pM solution of doubly labeled DNA in 20 mM sodium phosphate buffer. The DNA constructs are used for distance measurement (detailed in **Figure 9**) and have a TMR fluorophore attached to one end of the DNA molecule and a CY5 fluorophore attached to the *n*th base from the end. The time traces correspond to $n = 12$. Gray: TMR fluorescence (donor); black: CY5 fluorescence (acceptor). The laser excitation (0.6 mW, 514 nm) is focused 10 µm within the solution through an oil immersion objective (numerical aperture: 1.4). For each detected photon, the APD generates a pulse, which is recorded by a computer-embedded acquisition board. Counting these events into 200 µs bins, the resulting time trace exhibits distinct bursts distributed randomly in time, with duration and amplitude corresponding to a Brownian diffusion of individual molecules through the excitation volume. (b): is adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

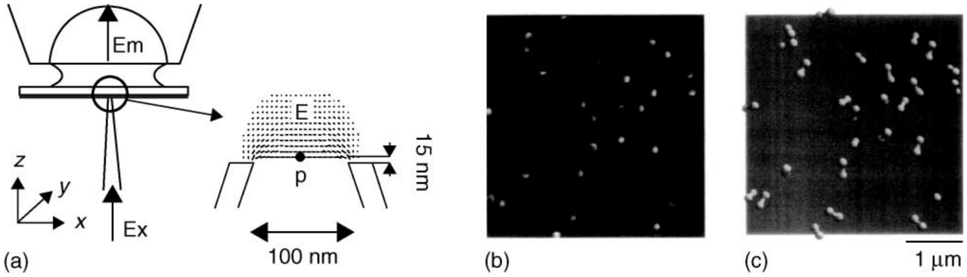

Figure 3 Scanning near-field imaging of single molecules at room temperature. (a) Schematic of a scanning near-field optical microscope setup used for single-molecule imaging. An aluminum-coated tapered optical fiber (raster-scanned at nanometer distance from the sample) serves as a waveguide for laser excitation. Shear-force feedback keeps the tip at a constant distance from the sample, resulting in a signal used for nanometer-resolution topographic reconstruction of the scanned area. The excitation volume and the corresponding local evanescent electric field are detailed in the expanded view. Fluorescence light emitted by individual molecules is collected by an oil immersion, high NA objective and recorded by an avalanche photodiode. (b) $4 \times 4\ \mu m^2$ area of a polymethylmethacrylate (PMMA) coated coverslip on which a diluted methanol solution of a lipophilic carbocyanine dye, $diIC_{12}$, has been deposited. Scanning was performed with randomly polarized light. (c) Orientation of the corresponding individual emitting dipole determined by modeling the field distribution (expanded view in a), simulation of the absorption intensity with polarized excitation and comparison with data obtained at two different excitation polarizations. (a): Adapted and (b), (c) reprinted with permission from Betzig E and Chichester RJ (1993) Single molecules observed by near-field scanning optical microscopy. *Science* 262: 1422–1425. Copyright the American Association for the Advancement of Science.

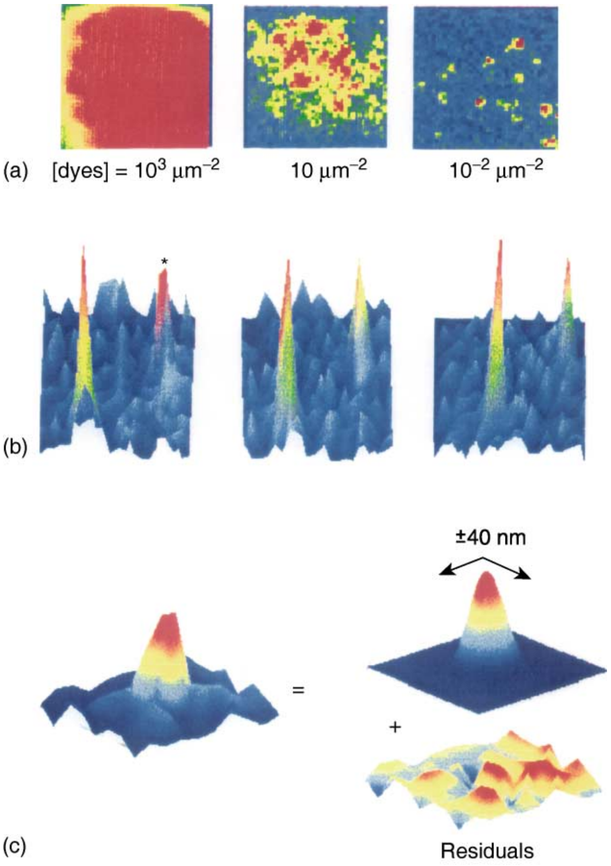

Figure 4 Single molecule diffusion in two dimensions. (a) 6.8×6.8 µm^2 images of fluorescently labeled lipids (tetramethylrhodamineinthiocarbamoyl, linked to 1,2-dihexadecanoyl-*sn*-glycero-3-phospho-ethanolamine, DHPE) in a lipid double layer membrane (palmitoyloleoylphosphocholine, POPC) at different concentration. From left to right: 10^3 dyes/µm^2, 10 dyes/µm^2, 10^{-2} dyes/µm^2. Each image was obtained with an epifluorescence microscope, illuminated by a defocused 514 nm Ar laser line (~60 kW/cm^2), and 5 ms integration on a nitrogen cooled CCD camera. Color scale: blue = 0, $d = 60$ counts). Image intensity at 10^3 dyes/µm^2 was divided by 60. (b) Sequence of three 5.4×5.4 µm^2 images of two labeled lipids observed at 105 ms intervals (actual acquisition rate was one frame every 35 ms, integration time: 5 ms). (c) Subregion of a 2.4×2.4 µm^2 membrane area showing the peak marked by a star in B. The nonlinear least-square fit of a two-dimensional Gaussian profile to the peak yields values of the fluorescence intensity $I = 168 \pm 25$ counts, width $r = 480 \pm 80$ nm and two-dimensional position with 40 nm resolution. Adapted from Schmidt T, Schutz GJ, Baumgartner W, *et al.* (1996) Imaging of single molecule diffusion. *Proceedings of the National Academy of Sciences USA* 93: 2926–2929. Copyright 1996, The Academy of Sciences.

parameters in the case of a confocal microscopy study of freely diffusing fluorescein isothiocyanate (FITC) in water, using avalanche photodiode detectors (APD) are: $E = 10^{-2}$, $Q = 0.85$, $\sigma = 2.8 \times 10^{-16}\,\mathrm{cm}^2$, $A = 0.1\,\mu\mathrm{m}^2$, $\lambda = c/\nu = 525\,\mathrm{nm}$, $bV = 10\,\mathrm{Hz}/\mu\mathrm{W}$, $D = 100\,\mathrm{Hz}$. With an excitation power of 100 µW and a 1 ms integration time, a count rate of 629 counts/ms can be obtained, of which 1 comes from the background sources, leading to a SNR ~25 and SBR ~629. Although these are acceptable figures, they have to be balanced by the fact that single molecules have a finite lifespan. Single molecules in an oxygen-rich environment typically emit on the order of 10^6 photons before irreversible degradation (photobleaching). With the above parameters, this amounts to a total emission of ~3.2 ms. This is enough to observe freely diffusing molecules during their transit time of a few hundred microseconds. For a fixed molecule, however, it sets a stringent limit on the total duration of a single-molecule observation, and all efforts are put on reducing the excitation volume and intensity, as well as background sources. Similar considerations hold on the detector side, where a trade-off has to be found between time-resolution and quantum efficiency.

In addition to SNR and SBR issues, care has to be taken to ensure that the collected signal originates from a single molecule. Two different strategies can be envisioned:

(i) work at low concentration, such that at most one molecule is present in the excitation volume (Figure 5b), or equivalently, minimize the excitation volume;
(ii) use a selective excitation or emission detection protocol, such that only one molecule is excited or detected within the sampled volume (Figure 5c).

The first case has already been illustrated in Figure 2, in which a solution of fluorophores at low concentration yields separate bursts corresponding most of the time to transits of individual molecules. The second strategy can be adopted for the example illustrated in Figure 1, in which the ZPL of individual molecules lying in the wings of the ensemble spectrum are well separated. Tuning the laser to the peak absorption frequency of a molecule will excite this molecule only, allowing for a separate study of this molecule.

In summary, a set of six criteria can be defined for SMS:

1. The observed density of emitter is compatible with the known concentration of individual molecules and scales with the original concentration.
2. The observed intensity level is consistent with that of a single emitting molecule.
3. Each immobilized emitter has a well-defined absorption or emission dipole (for organic dyes usually linear, but for other emitters such as a fluorescent semiconductor nanocrystal, possibly circular).

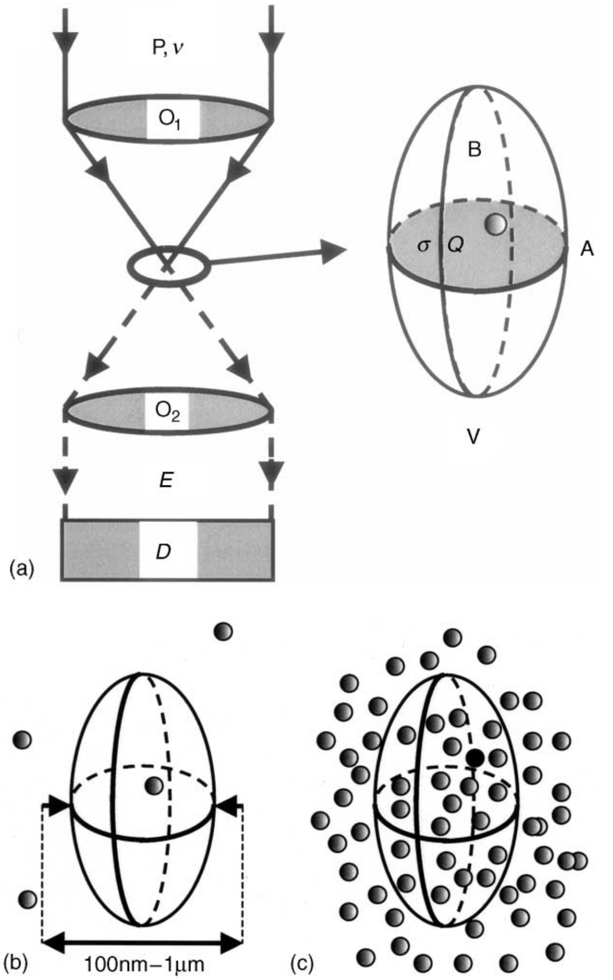

Figure 5 Signal, background and noise issues in single-molecule detection. (a) In a single-molecule spectroscopic experiment, excitation light (power P, frequency ν) is focused on the sample through an optical element O_1 (objective lens, fiber tip), exciting a diffraction-limited volume of section A, represented by an ellipsoid in the expanded view. The absorption process is characterized by an absorption cross-section σ and the emission by a quantum yield Q. The background contribution per watt of incident power is B, whereas the detector will have a dark count rate D. The signal is collected by an optical element O_2 (possibly confounded with O_1), and further guided to the detector, with an overall efficiency E. (b) The signal of a single molecule can be separated from that of other molecules if the probability of finding two or more molecules in the probed volume is negligible. This can be obtained either by a low concentration, or by reducing the excitation volume. (c) If each molecule in the excited volume can be characterized by a different spectral property (which may or may not depend on its exact environment), it is possible to use very selective excitation or detection techniques to probe just one molecule in the sampled volume (the black sphere in the ellipsoid).

4. Fluorescence emission exhibits only two levels (on/off) exemplified in blinking or photobleaching.
5. If there are more emission levels due to spectral jumps (see the section on fluorescence), blinking and spectral jumps are correlated.
6. Antibunching of the emitted photons permits us to test whether the detected signal comes from a single quantum emitter or more (see the section on biological studies below).

Types of Spectroscopy

Fluorophores

A variety of fluorescence emitters (or fluorophores) have been investigated with single-molecule spectroscopy techniques. Fluorescent organic molecules, green fluorescent proteins (GFP) and other biological fluorescent proteins, conjugated polymers (J-aggregates), light-harvesting complexes (e.g., B-phycoerythrin) comprising several fluorophores acting effectively as a single quantum emitter, or semiconductor nanocrystals, are but a few examples of systems which have been extensively studied. Each of these systems exhibits fluorescence based on specific processes, which cannot be described in detail here. A simple picture valid for single fluorescent dyes will be presented instead. It captures, however, the essential ingredients needed to describe the different parameters accessible with single-molecule spectroscopy.

Fluorescence Emission

Most single-molecule experiments use one-photon fluorescence excitation because of its relatively large cross-section ($\sim 10^{-16}\,\mathrm{cm}^2$). It also allows for a simple rejection of the excitation light: the emitted photons have a lower energy because of losses in intramolecular relaxation processes (the so-called Stokes shift). Excitation and emission processes are conveniently represented by a Jablonski diagram (Figure 6a) depicting the initial, final, and intermediate electronic and vibrational states of the molecule. Upon excitation to the first excited state, the molecule has a certain probability to be shelved in a nonemitting triplet state for up to milliseconds, resulting in 'dark' or off-states (Figure 6b,c). This results in a saturation of the emitted count rate R (Figure 6d) according to:

$$R = R_s \frac{I/I_s}{1 + I/I_s}, \qquad [3]$$

where the saturation intensity $I_s \sim 10$–$100\,\mathrm{kW/cm^2}$ and $R_s \sim 1$–$20\,\mathrm{MHz}$ for typical fluorophores. Incidentally, this puts a limit in SNR enhancement due to saturation of the emitted signal (triplet bottleneck). Other fluorescence emitting 'single molecules' like semiconductor nanocrystals exhibit similar dark state intervals, although for different reasons (Auger ionization or surface trapping of carriers). Finally, spectral jumps are also observed at room temperature (Figure 6e). This phenomenon results in a shift of absorption and emission maxima due to the changing environment of the molecule, or a sudden conformational change. The corresponding changes in excitation and emission spectra are noticed by sudden fluctuations in the emitted intensity

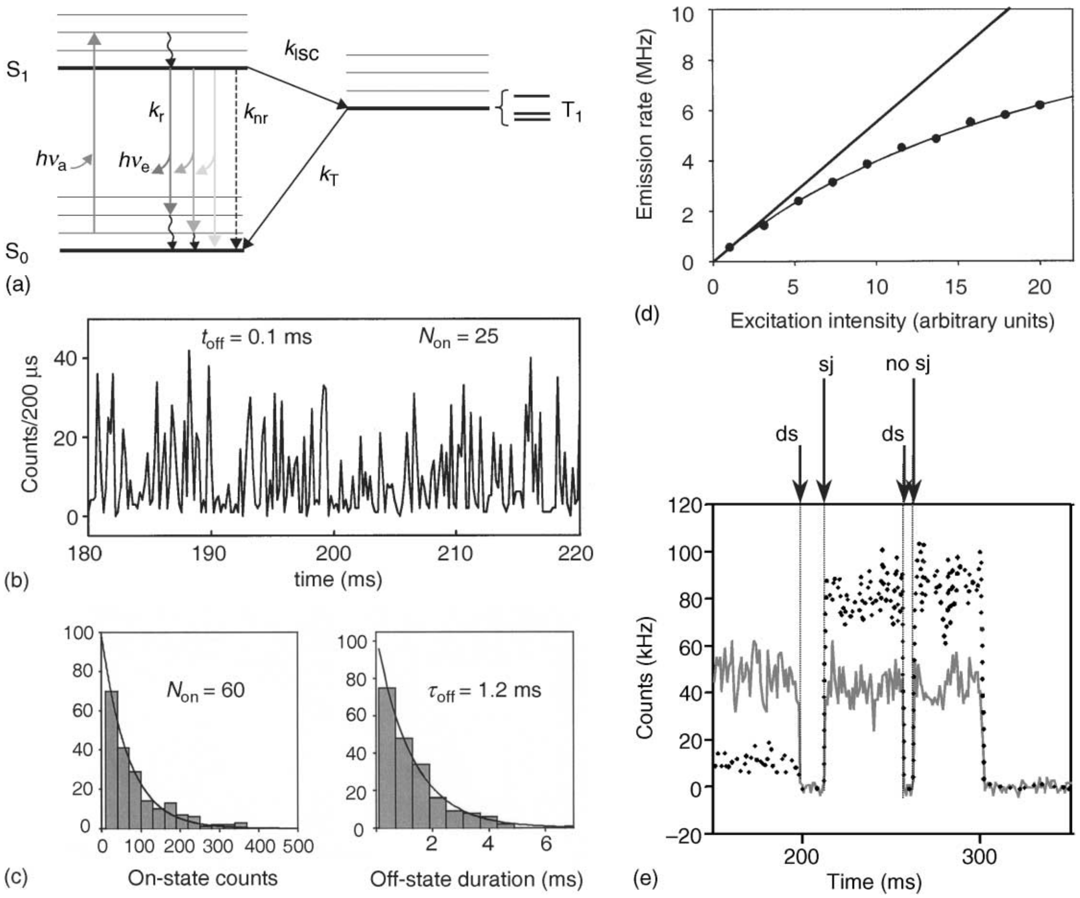

Figure 6 Single-molecule fluorescence. (a) Jablonski diagram for fluorescence (at room temperature). Upon absorption of a photon of energy $h\nu_a$ close to the resonance energy $E_{S_1} - E_{S_0}$, a molecule in a vibronic sublevel of the ground singlet state S_0 is promoted to a vibronic sublevel of the lowest excited singlet state S_1. Nonradiative, fast relaxation brings the molecule down to the lowest sublevel in picoseconds. Emission of a photon of energy $h\nu_e < h\nu_a$ (radiative rate k_r) can take place within nanoseconds and bring the molecule back to one of the vibronic sublevels of the ground state. Alternatively, collisional quenching may bring the molecule back to its ground state without photon emission (nonradiative rate k_{nr}). A third type of process present in organic dye molecules is intersystem crossing to the first excited triplet state T_1 (rate k_{ISC}). Relaxation from this excited state back to the ground state is spin-forbidden and thus the lifetime of this state is in the order of microseconds to milliseconds. Relaxation to the ground state takes place either by photon emission (phosphorescence) or nonradiative relaxation. The fluorescence lifetime is defined by $\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$. (b) Fluorescence time trace (40 ms excerpt) of a Texas Red (TR) molecule attached to a 20-nucleotides-long single-stranded DNA molecule immobilized on a silanized glass coverslip. The confocal setup is similar to that depicted in Figure 2. Excitation: 514 nm, circularly polarized, 6 kW cm^{-2}, binning: 200 μs. On- and off-states (the latter corresponding to the molecule being in the triplet state) alternate stochastically. (c) Histograms of on-state count (left) and off-state duration (right) for the time trace partly shown in (b). The decay parameters of the exponential fits (solid line) are indicated. $t_{off} = 1.2$ ms is the triplet state lifetime. $N_{on}^{-1} = k_{ISC} = 1.7\%$. (d) Intensity dependence of the emission rate of a single Texas Red molecule. The data departs from a linear dependence because of saturation due to intersystem crossing. The fitted curve yields a saturation rate of $R_S = 6.5$ MHz and a triplet lifetime $t_{off} = (k_{ISC}\ R_S)^{-1} = 10$ μs. The difference between this value and the value obtained in (c) illustrates the heterogeneity uncovered by single-molecule measurements. (e) Spectral jumps of a single Texas Red molecule attached to a single-stranded DNA molecule. ds: dark state, sj: spectral jump, no sj: dark state not followed by a spectral jump. Cases where spectral jumps are observed without noticeable dark states may still correspond to dark states with durations shorter than the time resolution of the experiment. (b)–(d) Adapted from Ha T, Enderle T, Chemla DS, *et al.* (1997) Quantum jumps of single molecules at room temperature. *Chemical Physics Letters* 271: 1–5. Copyright 1997, Elsevier Science B. V. (e) Adapted from Ha T, Enderle T, Chemla DS and Weiss S (1996) Dual-molecule spectroscopy: molecular rules for the study of biological macromolecules. *IEEE Journal of Selected Topics in Quantum Electronics* 2: 1115–1127. Copyright 1996 IEEE.

(due to wavelength dependence of the absorption cross-section).

Fluorescence Lifetime

A fluorescence photon is emitted within nanoseconds after absorption of the excitation photon. The decay law is usually mono-exponential, characterized by a lifetime $\tau = \Gamma^{-1}$ (Figure 6a). However, the lifetime depends on the emission spectrum (which can change due to spectral shifts) and the local environment. For instance, the proximity to a dielectric or metallic surface markedly changes it via perturbations of the intramolecular transition matrix elements. Thus, fluorescence lifetime is a sensitive probe of the environment of the molecule (Figure 7). Lifetime measurements require a pulsed laser source and time-correlated single-photon detection and analysis techniques.

Fluorescence Polarization

The efficiency of photon absorption by a fluorescent molecule is proportional to $(\boldsymbol{E}\cdot\boldsymbol{\mu})^2$ where $\boldsymbol{E}$ represents the local electric field and μ the molecule's absorption dipole moment. For a fixed fluorophore, modulation of the excitation polarization thus allows the determination of the absorption dipole orientation (the emission dipole being generally nearly parallel to it). This information in turn allows the determination of the spatial orientation of the molecule. For a mobile molecule, however, more information is needed, since the emission dipole has time to tumble significantly during the few nanoseconds between absorption and emission. The emission polarization is needed to fully recover the relevant information, as illustrated in Figure 8. Different experimental schemes can be used, besides the near-field optical one described in Figure 3 and the confocal one described in Figure 8. In particular, it is important to recover the projection of the polarization on more than two orthogonal axes. Approaches taking advantage of aberrations induced by index mismatch in high numerical aperture (NA) objectives, or using TIR excitation with alternating polarizations, or using amplitude or phase masks, allow a complete determination of this orientation.

Fluorescence Resonance Energy Transfer

A special case of the local influence of the environment on fluorescence is encountered in fluorescence resonance energy transfer (FRET, see Figure 9a), first theoretically explained by Förster in the late 1940s. In case of the presence of a nearby molecule (acceptor) having an absorption spectrum overlapping that of the donor emission, there is a possibility that the energy absorbed by the donor is transferred nonradiatively to the acceptor. The efficiency E of this transfer is given by:

$$E = \left(1 + \left(\frac{r}{R_0}\right)^6\right)^{-1} \qquad [4]$$

where r is the distance between the two emitting centers and R_0 is the Förster radius (in Å):

$$R_0 = a(\kappa^2 n^{-4} Q_D J(\lambda))^{1/6} \qquad [5]$$

In this complex expression, $a = 9.78 \times 10^3$, n is the medium refractive index, Q_D the donor fluorescence quantum yield in the absence of the acceptor and $J(\lambda)$ is a spectral overlap integral (in $\mathrm{M}^{-1}\ \mathrm{cm}^3$). κ^2 is a geometric factor, which averages out to 2/3 in the case of freely rotating dyes. A typical order of magnitude for R_0 is a few nanometers, which gives the range of the energy transfer. FRET is thus very sensitive to the donor–acceptor distance, which makes it a useful molecular ruler. Experimentally, the efficiency is measured as a function of the recorded donor (F_D) and acceptor (F_A) emissions (dual-channel measurement) via:

$$E = \left(1 + \alpha \frac{F_D}{F_A}\frac{Q_A}{Q_D}\right)^{-1} \qquad [6]$$

where Q_A and Q_D are the fluorescence quantum efficiencies of the donor and acceptor, respectively, and α is the ratio of the acceptor and donor channel detection efficiencies. An illustration of this utilization is given in Figure 9b–d, where different predefined distances between donors and acceptors are set by rigid double-stranded DNA molecules. The corresponding donor–acceptor distances are readily measured on diffusing molecules, with subnanometer precision.

Alternatively, the transfer efficiency can be measured via the fluorescence lifetimes as:

$$E = 1 - \frac{\tau_{D(A)}}{\tau_{D(0)}} \qquad [7]$$

where $\tau_{D(A)}$ and $\tau_{D(0)}$ are the donor lifetimes in the presence or absence of an acceptor, respectively.

Other Spectroscopies

Other types of spectroscopic techniques can be used at the single-molecule level. Fluorescence can for instance be excited via a two-photon absorption process, whereby a laser excitation with half the photon energy needed to attain the excited state

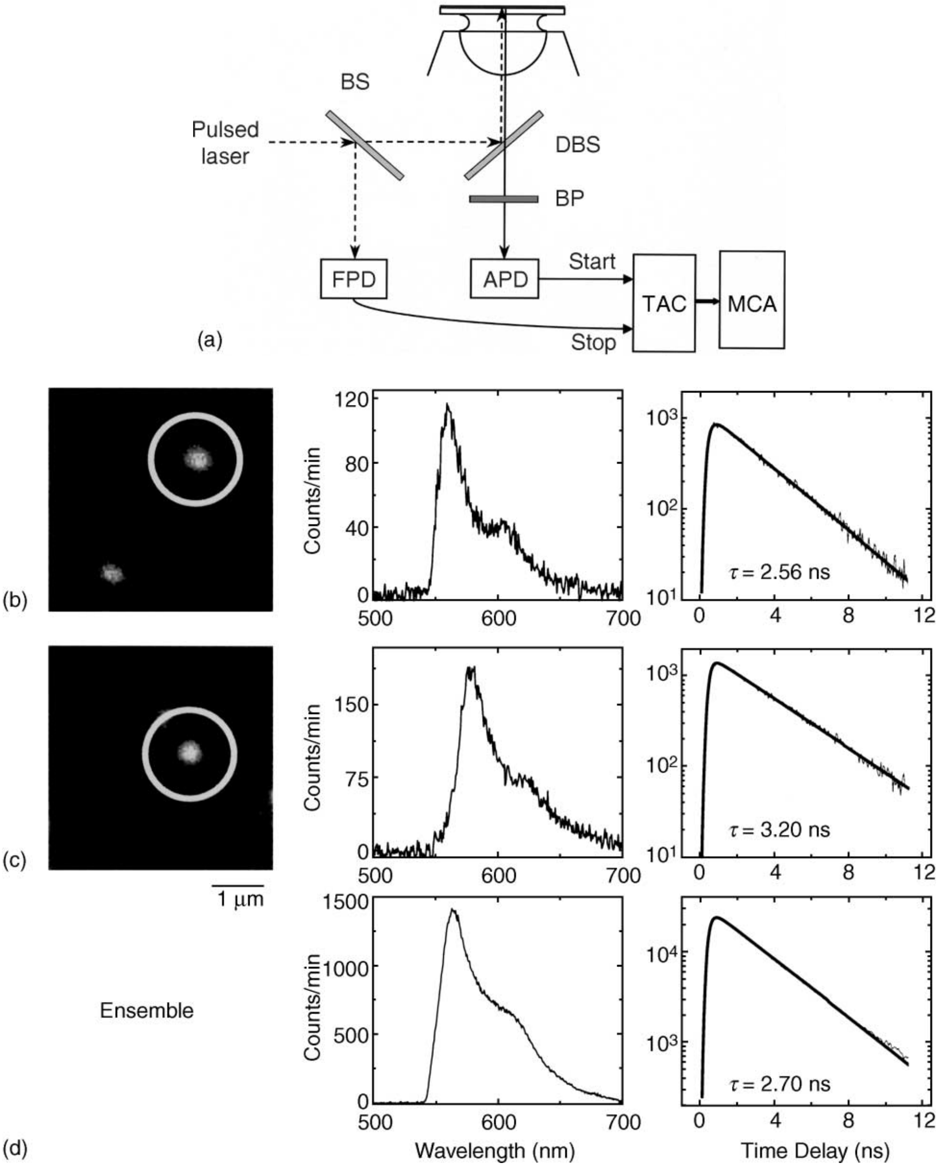

Figure 7 Lifetime measurements. (a) Schematic of a lifetime measurement setup. BS: beamsplitter used to deflect part of the incident pulse towards a fast photodiode (FPD); DBS: dichroic beam splitter used to spectrally separate incident from emitted light; BP: bandpass filter further rejecting stray light; APD: avalanche photodiode. For each detected photon, the APD emits a pulse, which is fed to the start input of a time-to-amplitude converter (TAC). A stop pulse from the laser immediately follows it (regularly emitted at a rate of τ_L^{-1}), allowing the measurement of the separation between both pulses. A multichannel analyzer (MCA) histograms the arrival times, resulting in the curves shown in the right part of Figure 2b–d. (b) and (c) Fluorescence images (left panels), corresponding spectra (middle), and fluorescence decays (right) for two molecules located at a PMMA–air interface. The peak emission wavelength λ_p was 560 nm in (b) and 578 nm in (c). Lifetimes were fitted to a single exponential (dotted curves), with decay times of $\tau = 2.56$ ns (b) and 3.20 ns (c). A ensemble measurement of spectrum and fluorescence decay averaged over several hundred of molecules is shown in (d). The ensemble values are $\lambda_p = 565$ nm and $\tau = 2.70$ ns. (b)–(d): Adapted from Macklin JJ, Trautman JK, Harris TD and Brus LE (1996) Imaging and time-resolved spectroscopy of single molecules at an interface. *Science* 272: 255–258. Copyright 1996, The American Association for the Advancement of Science.

is used. However, due to the very low cross-section and quadratic dependence in the incident energy (two simultaneous photons are needed for excitation), an excitation power several orders of magnitude larger than that for one-photon excitation is needed. In compensation, excitation takes place in a much more reduced volume of the sample, reducing the out-of-focus background contribution. The higher power required increases the probability of photobleaching due to the increased probability of

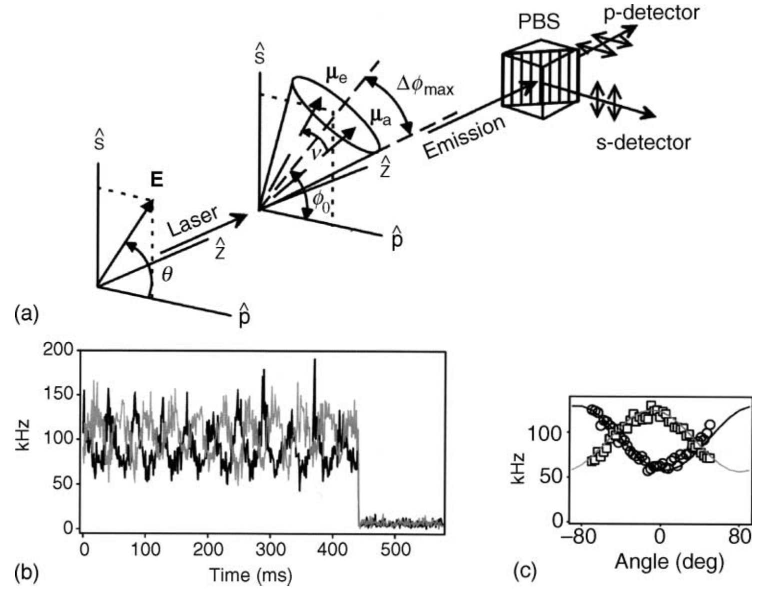

Figure 8 Polarization Measurements. (a) Experiment schematic. **E** is the electric field, making an angle θ with the *p* polarization axis. The excitation propagates along the *z* axis, which is also the collection axis. $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_e$ are the absorption and emission dipole moments, initially aligned. ν represents the rotational diffusion of the emission dipole during the excited lifetime. The dipole is supposed to be confined in a cone positioned at an angle ϕ_0 projected on the (*s*, *p*) plane, and having a half-angle $\Delta\phi_{max}$. A polarizing beam-splitter (PBS) splits the collected emission into two signals I_s and I_p, which are simultaneously recorded by APDs. (b) Simultaneously recorded I_s (black) and I_p (gray) of a molecule rapidly rotating in a liquid. (c) Same data as in (b), but averaged over 11 on-periods. The fit corresponds to a $\Delta\phi_{max}$ close to 90° (freely rotating molecule) and permits the determination of the constrained rotational diffusion parameters. Adapted from Ha T, Glass J, Enderle T, *et al.* (1998) Hindered rotational diffusion and rotational jumps of single molecules. *Physical Review Letters* 80: 2093–2096. Copyright 1998, The American Physical Society.

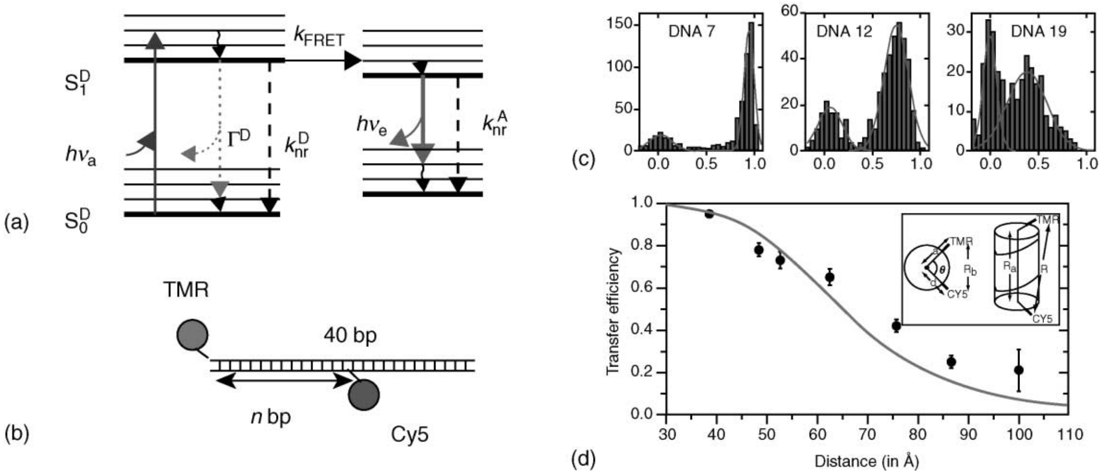

Figure 9 Single pair fluorescence resonance energy transfer. (a) Jablonski diagram for FRET. Fluorescence energy transfer involves two molecules: a donor D and an acceptor A whose absorption spectrum overlaps the emission spectrum of the donor. Excitation of the acceptor to the lowest singlet excited state is a process identical to that previously described for single-molecule fluorescence (**Figure 6a**). In the presence (within a few nm) of a nearby acceptor molecule, donor fluorescence emission is largely prevented by energy transfer to the acceptor by dipole–dipole interaction. The donor exhibits fluorescent emission following the rules outlined in **Figure 6a**. (b) DNA *n* constructs used for the FRET distance study. Tetramethylrhodamine (TMR) is attached to one end of the DNA and CY5 is attached to the *n*th base from the end ($n = 7$, 12, 14, 16, 19, 24 and 27). (c) FRET histograms extracted from time traces for DNAs 7, 12, and 19, as shown in **Figure 2b**. Double Gaussian fits extract numbers for the mean (width) of the higher efficiency peak of 0.95 ± 0.05, 0.75 ± 0.13, and 0.38 ± 0.21, respectively. The peak around 0 efficiency corresponds to bleached or inactive acceptor molecules. (d) Mean FRET efficiencies extracted from FRET histograms plotted as a function of distance for DNA 7, 12, 14, 16, 19, 24, and 27. Distances are calculated using the known B-DNA double helix structure. The solid line corresponds to the expected Förster transfer curve with $R_0 = 65$ Å. (b)–(c) Adapted from Deniz AA, Dahan M, Grunwell JR, *et al.* (1999) Single-pair fluorescence resonance energy transfer on freely diffusing, molecules: Observation of Förster distance dependence and subpopulations. *Proceedings of the National Academy of Sciences USA* 96: 3670–3675. Copyright 1999, The National Academy of Sciences.

photochemical degradation in the long-lived triplet state and the interplay of multiphoton ionization processes.

Although Raman scattering cross-sections ($\sim 10^{-30}$ cm^2) are orders of magnitude lower than the fluorescence cross-section, they can be enormously amplified if the single molecule is trapped in a 'hot spot' of a metallic nanoparticle, giving rise to surface-enhanced Raman scattering (SERS). Highly resolved vibrational information can be recovered, exhibiting much more stable emission than fluorescence. Moreover, this technique can be applied to nonfluorescent molecules. Similarly to fluorescence, the technique can be extended to nonlinear (two-photon) regimes, as in surface-enhanced hyper-Raman scattering and surface-enhanced anti-Stokes Raman scattering. However, since they require the single molecule to be located on rare sites of a metallic particle, these methods have only been applied to fundamental chemical studies of molecular species.

Finally, extending the range of the FRET 'molecular ruler', electron transfer allows measuring distances between an electron donor and an acceptor within a few ångströms (Å), via opening of a nonradiative channel for the relaxation of the electron in the excited state. Lifetime measurements ($\tau = 1/\Gamma = (k_r + k_{nr})^{-1}$ as defined in Figure 6) yield distance information via the electron transfer rate E_{ET}:

$$E_{ET} = k_0 e^{-\beta r}, \qquad [8]$$

where β^{-1} is of the order of one Å and k_0 of the order of 10^{15} s^{-1} are constants characterizing the donor–acceptor pair.

Applications of Single-Molecule Spectroscopy

Single-molecule spectroscopy has already proven to be a useful tool in various domains. The following examples are but a few of an expanding number of applications of SMS.

Molecular Physics

The realization that the fluorescence of a single molecule could be detected with a good SNR over extended periods of time, especially embedded in a crystalline matrix at cryogenic temperature, has led to high-resolution experiments aimed at determining molecular energy levels and transitions. In particular, studies of transitions between sublevels of the triplet state by electron spin resonance (ESR) using superimposed radio-frequency pulses have permitted a precise test of the quantum description of such systems. The dc and ac Stark effect (the shift of a transition energy in the presence of a fixed or alternating electric field) has also been studied in several systems, allowing the symmetry and deformations of the molecule to be determined. The method has been extended to study the photophysics of luminescent polymers (J-aggregates), light-harvesting complexes, green fluorescent proteins (GFP) or semiconductor nanocrystals, and other fundamental systems.

Material Science Studies

Single-molecule photophysical properties are extremely sensitive to their local (nanometer-sized) environment (Figure 1). A precise knowledge of these properties allows using single molecules as sensitive probes of their environment, as illustrated by experiments designed to study the local dynamics of a polymer matrix at the onset of the glass transition (Figure 10).

Molecules of the organic fluorophore Rhodamine 6G (R6G) dispersed within a thin (250 nm) poly(methacrylate) film were observed at temperatures slightly above the melting temperature of the polymer, using a confocal microscope detecting the *s* and *p* polarization of the emitted fluorescence (Figure 8a). Observations are performed under nitrogen atmosphere and at low excitation power (20 nW) in order to prevent oxidation, thus allowing long (several hours) observation time. Each observed R6G molecule exhibits a slow rotational diffusion, sometimes accompanied by lateral or transverse diffusion on a much longer time-scale (Figure 10a). Computation of the autocorrelation function of the emission polarization shows a multi-exponential decay, pointing to a distribution of time-scales of rotational diffusion. A closer look at subwindows of the time trace reveals, however, that the decay comes closer to mono-exponential behavior at shorter time-scales (Figure 10b,c). This study demonstrates that an individual molecule probes an increasing number of different environments over time. At long time-scales, the observable characteristics (such as the autocorrelation function) are similar to those measured on an ensemble of molecules, as expected from the ergodicity hypothesis. At short time-scale, the molecule reveals the local homogeneous and stable characteristic of its nanoenvironment.

Biophysical and Biological Studies

SMS has allowed a reassessment of long-standing questions in biophysics, biochemistry, and biology, by giving scientists the possibility to look at individual building blocks of biological processes: DNA, RNA, proteins and other biomolecules or biomolecular

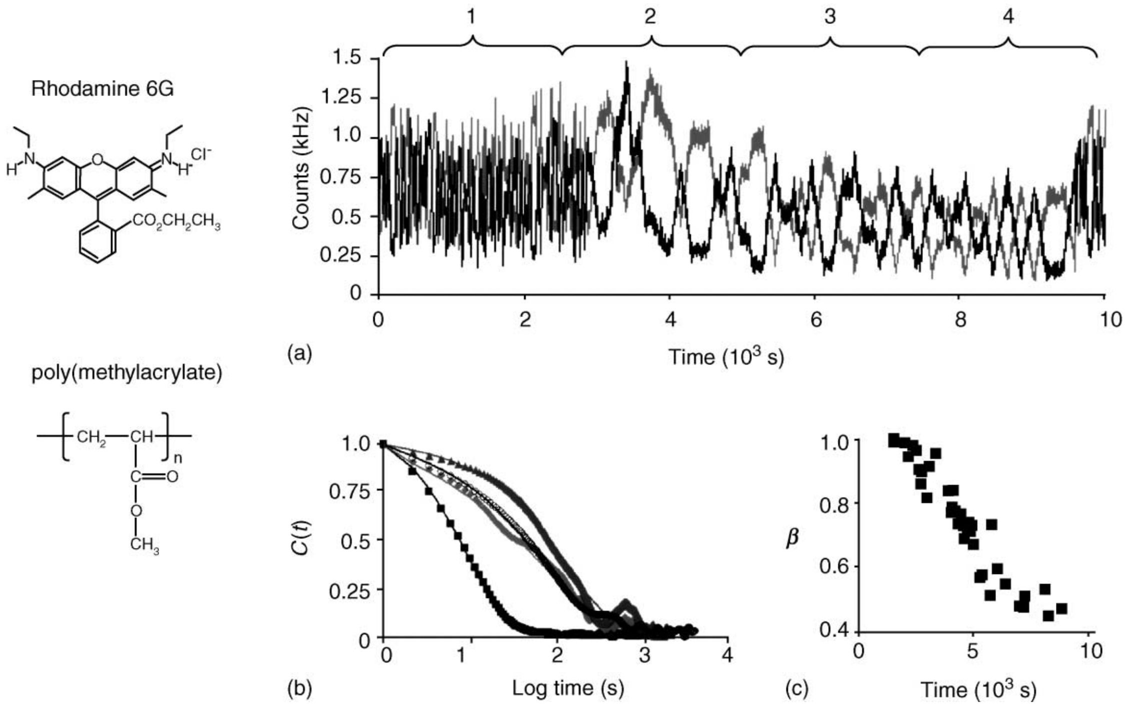

Figure 10 Polymer dynamics near the glass transition. (a) Fluorescence intensity of a single long-lived Rhodamine 6G molecule in poly(methylacrylate) (PMA) at 291 K detected at orthogonal polarizations (gray: *s* and black: *p* polarizations). The melting temperature of PMA is $T_g = 281$ K. The time trace integration is 1 s. This time trace shows the anticorrelation of both polarizations, characteristic of rotational diffusion of the molecule. Different regimes are observed along the time trace (1–4), which could be due to either a modification of the local 'cage' in which the molecule rotates, or escape of the molecule to another nearby nano-environment. The autocorrelation function of polarization computed over the whole trace (not shown) is well fitted by a model adapted to the nonexponential relaxation of polymers, the Kohlrausch–Williams–Watt (KWW) model:

$$\mathrm{KWW}(t) \propto \mathrm{e}^{-(t/\tau)^{\beta}}.$$

with $\tau = 88$ s, $\beta = 0.46$. (b) Correlation functions of subsets 1–4 of the full trace. 1 (squares): $\tau = 15$ s, $\beta = 1.0$. 2 (triangles): $\tau = 112$ s, $\beta = 0.77$. 3 (open circles): $\tau = 90$ s, $\beta = 0.59$. 4 (diamonds): $\tau = 64$ s, $\beta = 0.59$. The β exponent is closer to 1, showing that the rotational diffusion is an unhindered one at short time-scales, converging to the bulk exponent at larger time-scale due to ergodicity. This experiment illustrates changes of environment as the observed molecule diffuses and rotates through the polymer mesh, changes that are hidden in ensemble measurements. (c) Values of β as a function of photochemical survival time in PMA at $T_g + 5$ K. Adapted from Deschenes L and Vanden Bout DA (2001) Single-molecule studies of heterogeneous dynamics in polymer melts near the glass transition. *Science* 292: 255–258. Copyright 2001, The American Association for the Advancement of Science.

complexes formed by association of a few of these molecules.

To simplify the task, some biological molecules are intrinsically fluorescent in their active state, as the cholesterol oxidase (COx) enzyme of Brevibacterium sp., an enzyme that catalyzes the oxidation of cholesterol by oxygen. As a result, it is relatively easy to study the dynamics of this enzyme at the single-molecule level using a confocal setup with enzymes immobilized in an agarose gel (Figure 11). The intervals during which the enzyme is active ('on' state) are followed by intervals during which it is inactive ('off' states). The fluorescence recording of a single enzyme thus appears as a 'blinking' time trace, as shown in Figure 11b. Relevant kinetic parameters can be extracted from the distribution of 'on' and 'off' times. Apart from the usual heterogeneity between enzymes of the same batch characterized by different kinetic parameters for the catalysis reaction, this experiment reveals a dynamic heterogeneity, which shows up for each individually studied enzyme as a tendency for short 'on' times to follow short 'on' times, and for long 'on' times to follow long 'on' times (Figure 11c). This variable behavior of a single enzyme is reminiscent of the variable rotational dynamics of Rhodamine 6G molecules in a polymer matrix (see the previous section), and can be studied only thanks to single-molecule techniques.

Not all biological molecules can be studied in such a convenient way using their intrinsic fluorescence. Several labeling schemes are currently available, which allow quantitative, site-specific labeling of proteins, nucleic acid or lipid molecules with few biochemical side effects.

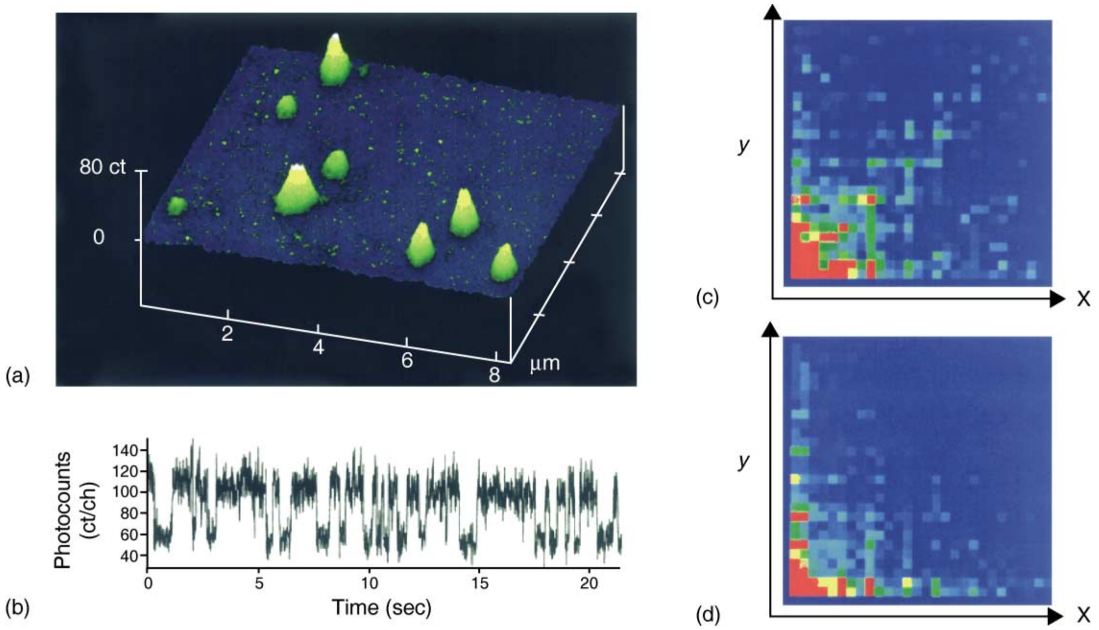

Figure 11 Single-enzyme dynamics. (a) Fluorescence image of single CO_x molecules immobilized in a 10-μm-thick film of agarose gel of 99% pH 7.4 buffer solution (8 μm × 8 μm, 4 min scan, 500 nW excitation at 442 nm). The emission is from the fluorescent active site, FAD, which is tightly bound to the center of CO_x. Each individual peak is attributed to a single CO_x molecule. The intensity variation between the molecules is due to different vertical positions in the excitation volume. (b) Real-time observation of enzymatic turnovers of a single CO_x molecule catalyzing oxidation of cholesterol molecules. This panel shows a portion of an emission intensity trajectory recorded in 13.1 ms per channel. The trajectory was recorded with a cholesterol concentration of 0.2 mM and saturated oxygen concentration of 0.25 mM. The emission exhibits stochastic blinking behavior as the flavin adenine dinucleotide (FAD), the CO_x fluorescent active site toggles between oxidized (fluorescent) and reduced (nonfluorescent) states, each on–off cycle corresponding to an enzymatic turnover. More than 500 on–off cycles are recorded in this trajectory. (c) The 2D conditional probability distribution for on-times (x and y) of two adjacent turnovers as derived from the trajectories of 33 CO_x molecules with 2 mM 5-pregene-3β-20α diol substrate, a derivative of cholesterol. The scales of the x- and y-axes are from 0 to 1 s. A subtle diagonal feature is present, indicative of a memory effect. (d) The 2D conditional histogram for two on-times separated by 10 turnovers for the CO_x molecules in (c). The diagonal feature vanishes because the two on-times become independent of each other at the 10-turnover separation. The color code in (c) and (d) represents the occurrence (z-axis from 350 (red) to 0 (purple). Adapted from Lu HP, Xun L and Xie XS (1998) Single-molecule enzymatic dynamics. *Science* 282: 1877–1882. Copyright 1998, The American Association for the Advancement of Science.

An example of this versatility has been provided by the labeling of the central part of a rotary protein, F_1-ATPase. In this experiment (not shown), the molecular rotor was labeled with a single fluorophore whose orientation, detected by emission polarization measurements, directly reported the angle of the rotor with respect to the shaft. The small size of the fluorophore guaranteed that the protein motion would not be hindered, contrary to previous experiments, which used much larger reporters (micron-size latex beads or fluorescent actin filaments). This experiment reproduced the previous results, showing that the rotor performed 120° steps (recent work hints at substeps of 90° and 30°).

One of the most promising domains where SMS can be of great importance is the study of protein folding. Doubly labeled proteins can be monitored as they undergo conformational changes, taking advantage of the distance dependence of fluorescence energy transfer (demonstrated for a model DNA system in the section on FRET, Figure 9).

Figure 12 shows one of many examples of such a study, performed on diffusing molecules in solution, a configuration that suppresses uncontrolled effects of nearby surfaces. The enzyme chymotrypsin inhibitor 2 (CI2) is a model system for protein folding, believed to have two clearly distinct folding states, which can be controlled by the concentration of denaturant guanidinium chloride. Using single-pair FRET (spFRET) techniques, it is possible to sort out the proportion of folded and unfolded molecules present at different denaturant concentration (Figure 12b). These measurements yield the same titration curve as obtained by ensemble measurements (Figure 12b) and can give access, under some simplifying assumptions, to the energy landscape of the folding reaction. The additional information provided by spFRET is

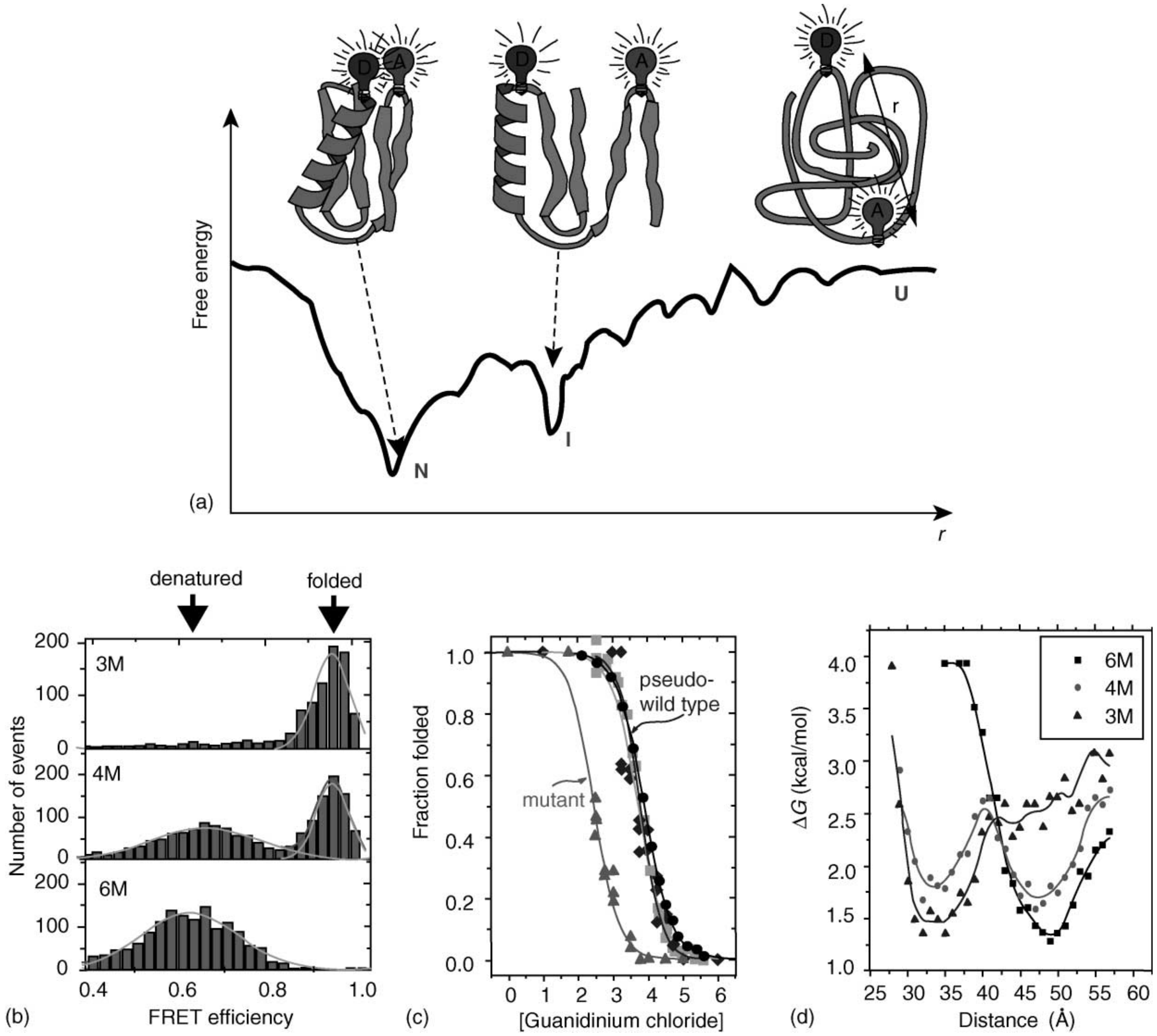

Figure 12 Protein folding. (a) Energy landscape for a protein folding reaction represented as a cartoon of the free energy as a function of distance *r* between donor and acceptor. The rugged energy landscape has a funnel shape with three extrema: U (unfolded state), I (intermediate state) and N (folded or native state). In this labeling scheme, donor and acceptor are close to one another in the native, folded state (N), resulting in a high FRET efficiency. There is reduced energy transfer (larger distance) in the denatured, unfolded state (U). (b) Single-molecule protein folding data. Single-pair FRET (spFRET) histograms of the pseudo-wild type enzyme chymotrypsin inhibitor 2 (PWT CI2) at 3, 4, and 6M denaturant (guanidinium chloride). The FRET efficiency exhibits two peaks, one for folded molecules (high FRET) and one for unfolded molecules (low FRET). Population is transferred from folded to unfolded at higher denaturant concentration, as expected from the protein labeling schematically depicted in (a). (c) Ensemble and single molecule denaturation curves for PWT and K17G mutant CI2. Symbols correspond to data; lines showing sigmoidal fits. PWT CI2: ensemble tryptophan fluorescence experiment (circles), ensemble FRET (squares), spFRET (diamonds); K17G mutant CI2 spFRET (triangles). (d) Free energy functions for PWT CI2 at 3, 4, and 6M denaturant. The solid lines represent a smoothing of the data and are only meant to guide the eye. (a) Adapted from Weiss S (2000) Measuring conformational dynamics of biomolecules by single-molecule fluorescence spectroscopy. *Nature Structural Biology* 7: 724–729. Copyright 2000, Nature Structural Biology. (b)–(d) Adapted from Deniz AA, Laurence TA, Beligere GS, *et al.* (2000) Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proceedings of the National Academy of Sciences USA* 97: 5179–5184. Copyright 2000, The National Academy of Sciences.

the number of molecules in the two different states, providing direct evidence of the two-states model hypothesized from ensemble measurements.

Other biological systems have been addressed using the spFRET approach. Intramolecular spFRET (used in the CI2 study) allows monitoring conformational changes, whereas by labeling two different molecules, intermolecular spFRET permits the detection of association and dissociation events. For instance, spFRET has revealed transient intermediate states in

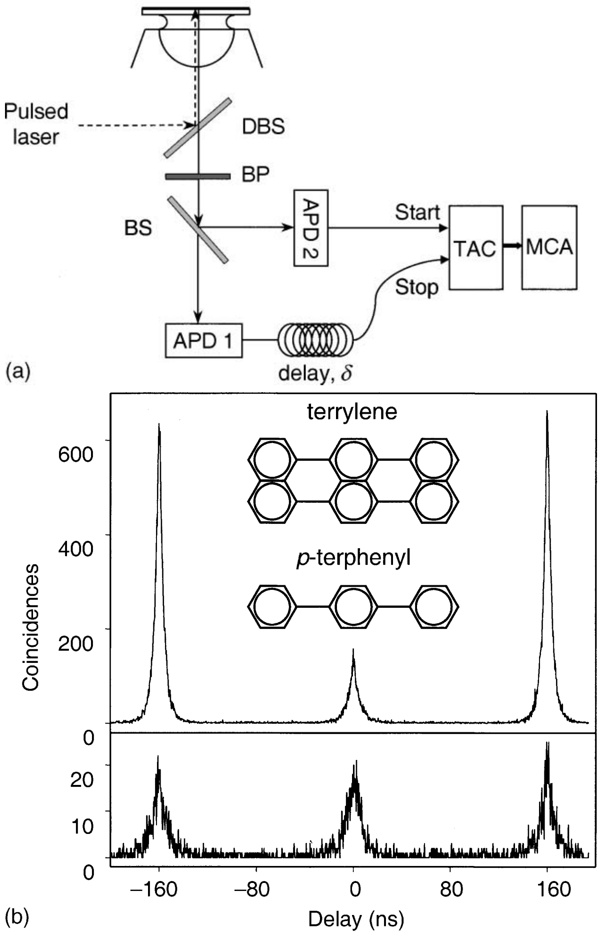

Figure 13 Single-photon source at room temperature. (a) Antibunching measurement setup. The basic elements are similar to those of a lifetime measurement setup (Figure 7a) except that the emitted filtered signal (BP: bandpass filter) is split in two with a 50/50 beamsplitter (BS). Two avalanche photodiodes (APD) detect the arrival of photons. One APD pulse is used as the start signal for a time-to-amplitude converter (TAC), which is stopped by the pulse coming from the second APD. Since the TAC response time is finite, it cannot detect simultaneously arriving pulses. A delay time δ is thus added to the stop signal, such that two photons impinging simultaneously on the two APDs will generate pulses with a separation of δ. If the start APD detects a photon from the previous excitation, the measured delay will be $160\,\text{ns} + \delta$, whereas if the opposite happens, the delay will be $\delta - 160\,\text{ns}$. (b) Results of measurements on a terrylene molecule (upper chemical structure) embedded in a p-terphenyl (lower chemical structure) at room temperature. Upper panel: histogram of delay measurements shifted by an amount δ. The histogram corresponds to 120 s of recording on an isolated emitting spot. The residual central peak corresponding to coincidence of photon arrival at APD 1 and 2 is the result of background contamination. Lower panel: histogram acquired over 300 s on a background spot. In both measurements, the laser pulse width was 35 ps, each pulse being separated by 160 ns. The lifetime of terrylene is measured to be 3.8 ns, which sets the width of the observed peaks. Excitation is performed at 532 nm. (b): Adapted with permission from Lounis B and Moerner WE (2000) Single photons on demand from a single molecule at room temperature. *Nature* 407: 491–493. Copyright 2000, Nature Publishing Group.

the Tetrahymena ribozyme, which had remained unnoticed in ensemble studies.

Quantum Optics

Low-temperature spectroscopic studies of single molecules have turned into a benchmark for testing predictions of quantum mechanics. In particular, one effect expected for a single quantum emitter is the absence of simultaneous emission of two photons: a fluorescent system which is excited at time t_0 has to emit a photon before it can return to the ground state, and be re-excited at a later time t_1. In conditions where the excitation takes place immediately as the electron returns to the ground state, the time difference $t_1 - t_0$ is of the order of the excited state lifetime, of the order of a few nanoseconds. This antibunching phenomenon was first observed in atoms and ions, and rapidly confirmed by low-temperature SMS. This observation is possible either using continuous wave or using pulsed excitation. In fact, nothing prevents this effect from occurring at room temperature, as demonstrated in Figure 13 for pulsed excitation in which excitation photons are provided as well-separated bunches of a few hundred femtoseconds duration.

In this experiment, a system originally used at low temperature (terrylene in a crystal of p-terphenyl) is excited with a repetition rate of 6.25 MHz. The fluorescence emitted by a single molecule is detected in a confocal setup, split in two and directed to two different detectors (Figure 13a). In this way, the time interval between each pair of photons detected by the two detectors can be precisely measured and histogrammed. The result of this time-correlated single-photon counting approach is the absence of coincidence of two photons (Figure 13b), besides the residual counts due to background contamination.

Since the quantum efficiency of this system is high, the experimental setup can be seen as a periodic single-photon source, with an extremely low probability for the simultaneous emission of two photons. This is precisely the kind of device that is needed for quantum cryptography, in which information encoded in the polarization of a beam should not be sensitive to a 'beamsplitter' attack. Future efforts will undoubtedly build on such 'turnstile' photon sources to move forward into optical quantum computing applications.

Perspectives

SMS is a mature discipline, which will certainly prove to be most useful in domains beyond those previously mentioned. Its relative ease of implementation will certainly attract more scientists of other disciplines to this methodology, each of them having specific goals in mind that are difficult to foresee. The principal advantages of SMS outlined at the beginning of this article will remain, but improvements, especially towards simultaneous acquisition of all the fluorescence parameters described earlier are promising. New detectors will permit the combination of the high time resolution of single-photon counting devices with the large field of view and spectral resolution allowed by two-dimensional detectors. SMS will be associated with single molecule manipulation techniques, to correlate applied forces and molecular conformations. New ways of controlling local fields (electric, magnetic, or others) would take advantage of the non-invasiveness, high-temporal and spatial resolution of SMS to get a direct feedback of events at the nanometer scale. Undoubtedly, SMS will be a major tool in the future of nanotechnology.

List of Units and Nomenclature

Å	Ångström
cm^2	square centimeter
fl	femtoliter
K	kelvin
ms	millisecond
MHz	megahertz
μW	microwatt
nm	nanometer

APD	Avalanche photodiode
DNA	Deoxyribonucleic acid
FRET	Fluorescence resonance energy transfer
NA	Numerical aperture
PMMA	Polymethylmethacrylate
SERS	Surface-enhanced Raman scattering
SMS	Single-molecule spectroscopy
SNR	Signal-to-noise ratio
SNE	Signal-to-background ratio
TIR	Total internal reflection
ZPL	Zero phonon line

See also

Scattering: Raman Scattering.

Further Reading

Basché T, Moerner WE, Orrit M and Wild UP (1997) Single-Molecule Optical Detection, Imaging and Spectroscopy. Weinheim: VCH.

Craig DP and Thirunamachandran T (1998) Molecular Quantum Electrondynamics. An Introduction to Radiation Molecule Interactions. Mineola: Dovers.

Forkey JN, Quinlan ME and Goldman YE (2000) Protein structural dynamics by single-molecule fluorescence polarization. *Progress in Biophysics and Molecular Biology* 74: 1–35.

Kneipp K, Katrun H, Kneipp I, *et al.* (1999) Surface-enhanced non-linear Raman scattering at the single-molecule level. *Chemical Physics* 247: 155–162.

Marcus RA and Sutin N (1985) Electron transfers in chemistry and biology. *Biochimica et Biophysica Acta* 811: 265–322.

Mehta AD, Rief M, Spudich JA, Smith DA and Simmons RM (1999) Single-molecule biomechanics with optical method. *Science* 283: 1689–1695.

Moerner WE and Orrit M (1999) Illuminating single molecules in condensed matter. *Science* 283: 1670–1676.

Nie S and Zare RN (1997) Optical detection of single molecules. *Annual Review of Biophysics and Biomolecular Structure* 26: 567–596.

Plakhotnik T, Donley EA and Wild UP (1997) Single-molecule spectroscopy. *Annual Review of Physical Chemistry* 48: 181–212.

Tamarat P, Maali A, Lounis B and Orrit M (2000) Ten years of single-molecule spectroscopy. *Journal of Physical Chemistry A* 104: 1–16.

Weiss S (1999) Fluorescence spectroscopy of single biomolecules. *Science* 283: 1676–1683.

Xie XS and Trautman JK (1988) Optical studies of single molecules at room temperature. *Annual Review of Physical Chemistry* 49: 441–480.

T

TERAHERTZ TECHNOLOGY

Contents

Coherent Terahertz Sources
Terahertz Physics of Semiconductor Heterostructures

Coherent Terahertz Sources

L Wang, Chinese Academy of Sciences, Beijing, China

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Terahertz (THz) radiation is usually referred to as an electromagnetic wave with frequencies ranging from 1 to a few terahertz ($1\,\mathrm{THz} = 10^{12}\,\mathrm{Hz}$). In the electromagnetic spectrum, terahertz radiation lies between the infrared and microwave, as shown in Figure 1. The techniques for generation and detection of THz radiation bridge photonics and electronics in the sense of the concepts and the techniques discussed below.

Physical quantities corresponding to 1 THz are listed as follows;

- Frequency, 1 THz
- Wavelength, 300 μm
- Wavenumber, 33 cm^{-1}
- Energy, 4.1 meV
- Temperature, 48 K

In this spectral region, there exist many rich physical, chemical, and biological phenomena. For example, many phonon resonance and other elementary excitations in condensed matter fall in this region; many molecular vibrations and rotations occur at THz frequency; conformation-related collective vibrational modes in macromolecules, especially in bio-molecules such as proteins and DNA, also lie in THz frequency. However, the development of THz technique lags far behind that of photonics and electronics, due to lack of feasible, reliable, and economic coherent THz sources. The situation has been rapidly changing since the 1990s. Benefiting from advanced material science and the ultrafast laser techniques, various coherent THz sources are being developed and commercialization is soon to follow. We will introduce these coherent THz sources and their main features, and discuss their generation mechanisms. More details can be found in the listed books and articles in the Further Reading section at the end of this article.

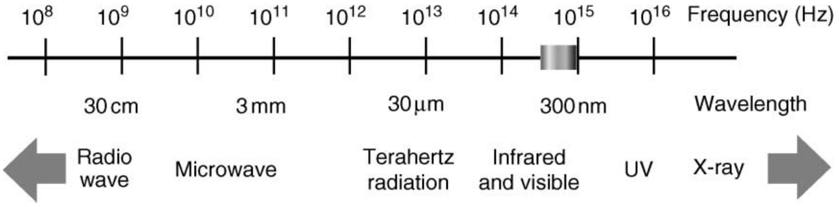

Figure 1 The electromagnetic spectrum with THz radiation sitting between the microwave and the infrared.

Coherent THz Radiation

When a time varying polarization, $\mathbf{P}(\mathbf{r},t)$, is generated in a medium and oscillates at THz frequency, it will radiate a THz wave, $\mathbf{E}(\mathbf{r},t)$, according to Maxwell's equation:

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r},t) + \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E}(\mathbf{r},t) = -\mu_0 \frac{\partial^2}{\partial t^2}\mathbf{P}(\mathbf{r},t) \qquad [1]$$

or

$$\nabla \times \nabla \times \mathbf{E}(\omega) - \frac{\omega^2}{c^2}\mathbf{E}(\omega) = \omega^2 \mu_0 \mathbf{P}(\omega) \qquad [2]$$

in frequency domain, for a monochromatic wave or a Fourier component of a wave with a finite bandwidth. The properties of THz radiation are determined by the polarization source $\mathbf{P}(\mathbf{r},t)$. Incoherent THz sources exist in blackbody radiation, and are also detected in astronomy observations. If the phase evolution of the THz wave keeps in step, both spatially and temporally, it is called coherent radiation. A coherent THz source is more powerful and useful in spectroscopy, material characterization, imaging, and many other applications. It is clear that a coherent polarization source is first required for generation of a coherent THz wave. In practice, the working media for THz emitting can be free electrons, quasi-free electrons in solids, or bound electrons and other charge oscillations, such as plasmon oscillation, lattice vibration, etc. Classified by the excitation properties, the emitting process can be of resonance or nonresonance. Coherent polarization sources with different features can be created using various excitation techniques under different principles and device configurations, which lead to different coherent THz sources. Here we focus on coherent THz sources, primarily based on photonic techniques, and working in the spectral range from 1 THz to a few THz.

Gas THz Laser

Molecules have many rotational and vibrational energy levels spaced at THz frequency. For polar molecules, the radiative transition between these levels can radiate THz wave. Following the standard laser technology, molecules pumped either optically or electronically can be used as laser gain media for coherent THz wave generation. For THz gas lasers, the very low THz photon energy poses a problem for lasing, because the coherent radiative transition can be disrupted by thermal depopulation and dephasing. An effective population inversion is difficult to set up and maintain. Therefore, it is important to carefully choose working medium and energy levels, as well as the way to excite the molecules. A number of media has been used for THz laser, such as methanol and HCN vapor, working in both continuous wave (CW) and pulsed modes. For example, methanol lasers operate under excitation of the rocking and asymmetric deformation modes of methanol excited by a CO_2 infrared laser. HCN laser is pumped electrically, which needs a large voltage and current to excite the molecules, a long cavity to obtain sufficient gain, and dedicated control of the temperature in the cavity for stable operation.

In general, THz gas lasers can be tuned discretely in thousands of lasing lines, ranging from a few tens to a few hundreds of micrometers, with power output from μW to mW. THz gas lasers have existed since the early 1970s, and are the only commercialized THz lasers to date. They are still subject to development, for less bulky volume, enhanced tunability and efficiency, and reduced costs.

Free Electron THz Laser

A free electron laser (FEL) uses free electrons as working medium, rather than bound atomic or molecular states in a conventional laser. A typical FEL consists of three parts: (i) an electron source that generates an electron beam with high current; (ii) an accelerator to raise the electron energy; and (iii) a lasing cavity. After the electron beam is generated, electrons are first accelerated to a relativistic velocity, typically with energy of hundreds of MeV, and then enter the cavity consisting of end mirrors and a spatially periodic magnets array called a wiggler (Figure 2). Due to the Lorentzean force imposed by the periodic magnetic field, the high-energy electrons move in the cavity along a sinusoidal path, and emit coherent radiation with a wavelength determined by the spacing between magnets and the electron velocity, as well as the magnetic induction. With the feedback provided by the cavity mirrors, electrons are accelerated or decelerated continuously by the optical field, and are bunched via the resonant interaction. The collective motion of the electron bunches radiates powerful coherent synchrotron radiation. In a standard configuration, the wavelength of the radiation can be expressed as:

$$\lambda = \frac{\Lambda}{2\gamma^2}[1 + (k\Lambda B_0)^2/2] \qquad [3]$$

where Λ is the spatial period of the wiggler magnets, γ the Lorentz factor of the electron beam, B_0 the peak magnetic induction, and k a constant. With a suitable

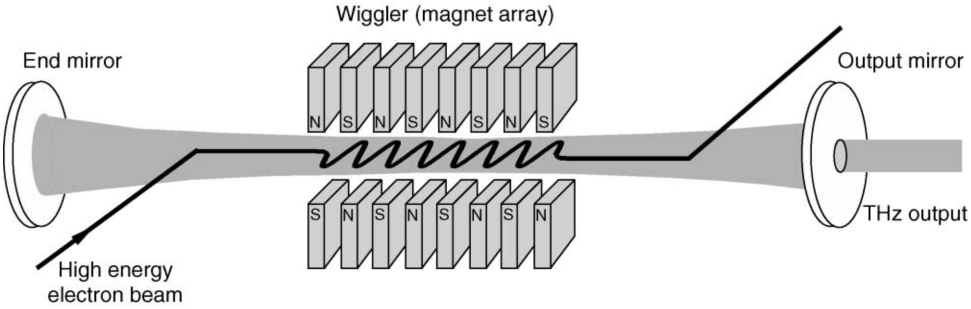

Figure 2 A schematic of THz generation in a free electron laser.

arrangement of the magnet array, as well as other related factors, the facility can work in THz spectral region.

An FEL generates tunable, coherent, high-power THz radiation. It is the most powerful coherent source of THz radiation available. Up to 50 W of average THz power has been generated in the Jefferson Lab recently. The free electron THz laser is becoming an important tools for studying THz radiation and its interactions with condensed matter and biological materials. On the other hand, the FEL is a complicated and expensive facility, and needs dedicated maintenance.

Semiconductor THz Laser

In the twentieth century, semiconductor devices have achieved tremendous success, both in electronic and photonic regime, such as transistors and diode lasers. The semiconductor industry has resulted in revolutionary changes to our world and everyday life. To fill the THz gap between electronics and photonics, great efforts have been made to meet the demand in compact, economic, tunable, and highly efficient THz sources. Encouraging advances have been made since we stepped into the new millennium.

In a semiconductor quantum well, the energy difference between a pair of sublevels can be artificially designed at THz frequency. As early as 1971, the concept of far infrared lasers based on intersub-band transition in semiconductor superlattices was proposed. Although the sublevels can be populated by electrical or optical pumping, the low photon energy at THz frequency causes serious difficulties for lasing, because the two levels sit so close to each other that many thermal processes could destroy the population inversion. Various designs and configurations have been proposed to tackle the problem. An attractive idea is so-called quantum cascade heterostructure. In this kind of device, the electrons are injected into a series of coupled quantum wells electrically. When the electrons are driven by the biased voltage, they can make radiative transition between a pair of sub-bands in a well. Subsequently, electrons enter the next well through resonant tunneling, and so on. However, this kind of device needs the materials of very high quality and suffers severely from fast depopulation of the excited states, even at very low temperatures.

As the advance in material science and unremitting efforts in pursuing the dedicated design of the device structures continued, the real breakthrough came in 2001, when Köhler *et al.* at Pisa demonstrated THz lasing in a quantum cascade heterostructure. Many improvements have been made since then; for example, the idea of using phonon resonance to selectively deplete the population of the lower sub-band has been successful, so that much stabler and robuster population inversion can be set up. The operation temperature has been raised to 136 K, well above the liquid nitrogen temperature, which opens the way for semiconductor THz lasers stepping into many more daily applications. These semiconductor devices are pumped with low voltages and currents at mA levels, and generate narrow bandwidth radiation with mW output at a frequency of a few THz. It is also probable that a four-level lasing system could even working at room temperature.

Another promising device is the p-germanium (Ge) laser developed recently. The Ge laser operates through the electrical excitation of hot holes in p-doped Ge. The laser cavity is formed by polishing the surfaces of the Ge crystal. These lasers could run at 5% duty cycles, with several mW output power. The output wavelength can be continuously tuned from 1 to 4 THz. At the moment, the devices have to operate at 20–30 K and consume about 10–20 W for refrigeration.

It is safe to predict that the feasibility, low cost, and compact semiconductor THz laser will emerge in the near future.

Beside the THz lasers, coherent THz radiation can also be generated by coherently exciting suitable media in other ways. Since the mid-1990s, the development of ultrafast laser techniques has led to two important pulsed coherent THz sources, i.e., photoconductive antenna and optical rectification THz sources. A unique feature of the pulsed coherent sources is broad bandwidth, which is useful in spectroscopic applications. Figure 3 is a typical setup for pulsed THz generation and spectroscopic study, using photoconductive antenna or optical rectification as THz sources.

THz Photoconductive Antenna (PCA)

Different from a conventional antenna working in radio and microwave frequency that are driven directly by electrical current, a photoconductive antenna for THz wave generation works with ultrafast pulsed lasers. When the photoconductive gap is irradiated by an ultrafast laser pulse, photocarriers are generated in the semiconductor. These photocarriers are driven by a DC voltage bias and form a current transient. The antenna couples the electromagnetic field associated with the time varying current into free space, and produces a THz pulse. The THz pulse usually has a duration of a few picoseconds, with a broad frequency bandwidth centered at THz frequency. The output THz power scales with both the optical pulse power and the DC bias field:

$$E_{\mathrm{THz}} \propto \frac{\partial J}{\partial t} \propto \frac{\partial^2 n_{\mathrm{e}}}{\partial t^2} E_{\mathrm{bias}} \qquad [4]$$

The output power, frequency bandwidth, and polarization of the THz radiation depend on the photoconductive features of the semiconductor layer and the geometric structure of the antenna. For an optimal performance, the substrate should have a suitable bandgap for optical interband excitation, a high electron mobility and short carrier lifetime to support an ultrafast current transient, and high dielectric breakdown threshold and to sustain high bias voltage. High-power PCAs have been demonstrated with semi-insulating GaAs, ion-implanted GaAs, and In-GaAs. In the aspects of antenna structures, coplanar strip lines and large aperture emitters are the most effective. Average output power from a good system can reach 30–40 μW.

A broad bandwidth is a special characteristic of a THz source, which is particularly desirable in the spectroscopic study as a coherent probe source. From the property of Fourier transform, we know that the shorter the THz pulse duration, the broader the spectral band. PCAs made of GaAs typically have a useful bandwidth extending from ~100 GHz to 3 THz, which can be extended to 4 THz by tuning the pumping laser wavelength close to the bandedge. Up to 6 THz bandwidths have been reported for PCAs. The pulse duration and, therefore the achievable bandwidth, are ultimately limited by carrier mobility and TO phonon resonant absorption which is around 8.3 THz in GaAs.

In many applications, THz radiation is coupled to free space from the antenna using a closely attached silicon hemispherical lens. This practice increases the system's output and also provides control of the radiation pattern. The radiation pattern for the common dipole antenna is essentially dipolar, with a weak quadrupole component perpendicular to the bias field. In the far-field, the THz beam has a Gaussian cross-section with high-frequency components concentrated in the center.

As the first practical pulsed coherent THz source, PCA has been widely used for many applications in

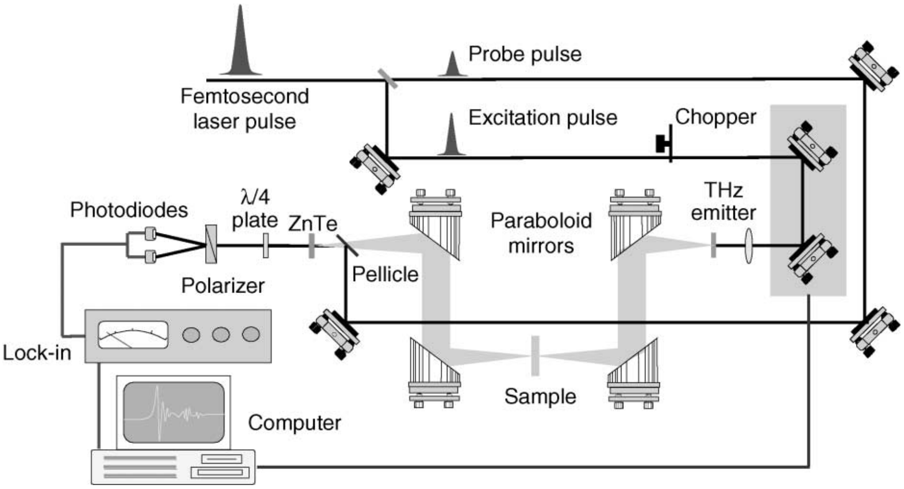

Figure 3 A typical setup for broadband pulsed THz generation and spectroscopic study with a femtosecond laser as excitation source.

scientific research and material characterization. Combined with commercial ultrafast optical fiber lasers, full functional and self-contained THz spectroscopy systems with PCAs as emitters, have been demonstrated and are under commercialization.

Optical Rectification

Optical rectification is a second-order nonlinear optical effect originating from susceptibility $\chi^{(2)}$, and was first observed in 1962 using two intense monochromatic nanosecond laser pulses. In this process, the interaction between laser and nonlinear optical crystal produces a frequency difference polarization:

$$P(0) = \chi^{(2)}E(\omega_1)E(-\omega_1) \qquad [5]$$

When a nonlinear optical crystal is irradiated by an ultrafast laser pulse, the different frequency components in the laser pulse will be coupled with each other by $\chi^{(2)}$, to induce frequency difference polarizations as:

$$P(\Omega) = \chi^{(2)}(\omega_1 - \omega_2 = \Omega)E(\omega_1)E(-\omega_2) \qquad [6]$$

A femtosecond laser pulse usually has a spectral linewidth of at least a few nm, the induced polarization $P(\Omega)$ has a finite frequency distribution centered at THz frequency in the far field, and results in THz radiation $E_{\mathrm{THz}} \propto \partial^2 P(t)/\partial t^2$.

THz radiation in free space by optical rectification was obtained in the early 1990s. One of the differences from a PCA device is that the THz output by optical rectification scales with both the optical pulse power and the DC bias field, while the THz radiation generated by optical rectification solely results from the incident laser. The optimal performance depends on several factors. The generation efficiency first depends on the magnitude of the $\chi^{(2)}$ and phase-matching condition between the optical and THz pulses. The effective magnitude of the $\chi^{(2)}$ coefficient varies with the crystal cut and orientation. Up to now, the most popular material is ZnTe, for its physical durability and excellent phase matching, when a Ti:sapphire femtosecond laser is used as the excitation source. With a moderately focused optical pump beam and a ZnTe crystal, nW T-ray average power can be generated by optical rectification. In general, the generation efficiency may be increased by increasing the laser power, but there are two factors limiting the effect. First, the incident laser power cannot exceed the damage threshold. Second, other second-order nonlinear processes may compete with the optical rectification, such as second-harmonic generation of the pump beam at high optical flux.

For THz radiation generated by optical rectification, the ultimate bandwidth is primarily determined only by the linewidth of the pump laser pulse. Because it is a nonresonance process in most cases, optical rectification can reach broader bandwidths than PCA. Using ultrashort optical pulses and thin nonlinear crystal, THz pulses have been generated with spectra extending to the mid-infrared, well beyond the phonon band of the materials. Like the photoconductive antenna, optical rectification has become a popular technique to generate coherent broadband THz radiation in various applications. Beside the conventional bulk inorganic crystals, many different media have been used for THz generation, such as organic crystal DAST, biased quantum wells, periodically poled LiNbO3 (PPLN), poled polymers, super-conducting thin films, etc.

Other Coherent THz Sources

When two light sources with slightly different wavelengths interact together with a nonlinear optical crystal, a beating polarization will be generated and will radiate a THz wave. It is called difference frequency generation, or three-wave mixing. Optical parametric processing is another way to generate a THz wave. In this process, a near-infrared pump beam generates a second NIR idler beam in a nonlinear crystal, and THz radiation can be generated from the beating of the pump and the idler. The merit of this technique is the continuous frequency tunability of the THz output, which is valuable in spectroscopic applications, and relatively cheaper nanosecond lasers can be used. If the nonlinear crystal is further placed in a cavity and the idler beam is amplified by feedback from end mirrors, a THz optical parametric oscillator is formed. A number of improvements have been made since the mid-1990s with this kind of device. More recently, THz parametric generation with an injection seeded idler beam has shown a reduced linewidth ($\Delta v/v \approx 10^{-4}$) and a peak THz power of over 100 mW for 3.4 ns pulses. A narrow linewidth combined with reasonable tunability make this kind of THz source attractive in spectroscopic studies.

Besides using photonic techniques, coherent THz radiation can also be generated by electronic techniques through increasing the output frequency of the microwave devices. For example, an electrically driven microwave generator, backward wave oscillator, can generate CW THz radiation up to 2 THz. A backward wave oscillator, running under optimal conditions, can provide up to 300 mW of polarized, narrowband radiation with a limited tunability about 30% of its central frequency.

See also

Imaging: Infrared Imaging. **Lasers:** Free Electron Lasers. **Optical Parametric Devices:** Overview. **Semiconductor Physics:** Infrared Lattice Properties. **Terahertz Technology:** Terahertz Physics of Semiconductor Heterostructures.

Further Reading

Grüner G (1998) *Millimeter and Submillimeter Wave Spectroscopy of Solids*, *Vol. 74 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Kaiser W (1988) *Ultrashort Laser Pulses and Applications*, *Vol. 60 of Topics in Applied Physics*. Berlin: Springer-Verlag.

Khurgin JB (1994) Optical rectification and terahertz emission in semiconductors excited above the bandgap. *Journal of the Optical Society of America B* 11: 2492–2501.

Rullière C (1998) *Femtosecond Laser Pulses*. Berlin: Springer-Verlag.

Saldin EL, Schneidmiller EA and Yurkov MV (2000) *The Physics of Free Electron Lasers*. *Advanced Texts in Physics*. Berlin: Springer-Verlag.

Siegel PH (2002) Terahertz technology. *IEEE Transactions on Microwave Theory and Techniques* 50: 910–928.

Tredicucci A, Gmachl C, Capasso F, *et al.* (2001) Novel quantum cascade devices for long wavelength IR emission. *Optics Materials* 17: 211–217.

Tsen KT (2001) *Ultrafast Physical Processes in Semiconductors*, *Vol. 67 of Semiconductors and Semimetals*. New York: Academic Press.

Terahertz Physics of Semiconductor Heterostructures

R Bratschitsch and K Unterrainer, Technische Universität Wien, Vienna, Austria

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

The terahertz (THz) regime of the electromagnetic spectrum holds promise for a large number of important applications in areas like environmental sensing, quality control, medicine, military applications, astronomy, etc. The development of THz devices has been mainly hindered by the lack of compact sources. In this article we overview the types of tabletop semiconductor sources that are available right now. We focus on the use of pulsed THz sources that are optically excited by femtosecond laser pulses. The generated bursts of THz radiation are extremely useful for time-resolved THz spectroscopy of semiconductor heterostructures.

THz Generation from Semiconductors

The first freely propagating THz pulses using femtosecond lasers were generated in 1984 when Auston *et al.* fabricated a Hertzian dipole antenna. It mainly consists of an electrically biased transmission line structure mounted on a semiconductor, i.e. a photoconductive switch that is triggered by a femtosecond laser pulse (Figure 1). The generated transient photocurrent $J(t)$ emits radiation according to $E(t) \propto \partial J(t)/\partial t$. Since 1989, sapphire half-spheres or refractive silicon lenses have been attached to these dipole antennas, which made it possible to produce collimated THz beams for use in everyday spectroscopic experiments. Today generation and detection of THz radiation by semiconductor photoconductive antennas is widely used. However, most of the antennas are limited to the low-frequency regime of the THz spectrum since they usually have a high-frequency cutoff of about 1–2 THz.

THz generation from semiconductor surfaces was introduced by Zhang in 1990. There are three different physical mechanisms that are responsible for the surface emission of THz radiation. One is the transient current, i.e., the acceleration of photogenerated carriers in the semiconductor surface electric field. This process is accompanied by field induced optical rectification which is often called instantaneous polarization, where the electron–hole pairs are polarized in the dc electric field. Finally, bulk optical rectification as a source of THz radiation can be observed in noncentrosymmetric crystals.

The surface emission intensity can be enhanced by using an externally applied electric field instead of the intrinsic electric field. Until recently, the use of laser amplifiers that optically pump these so-called large-aperture antennas has led to the brightest ultrafast tabletop THz emitters.

Recently, Sarukura and co-workers found that the application of a magnetic field dramatically increases the THz emission from bulk InAs. This led to a re-examination of Zhang's work, in which he reported the magnetic field enhancement in bulk GaAs,

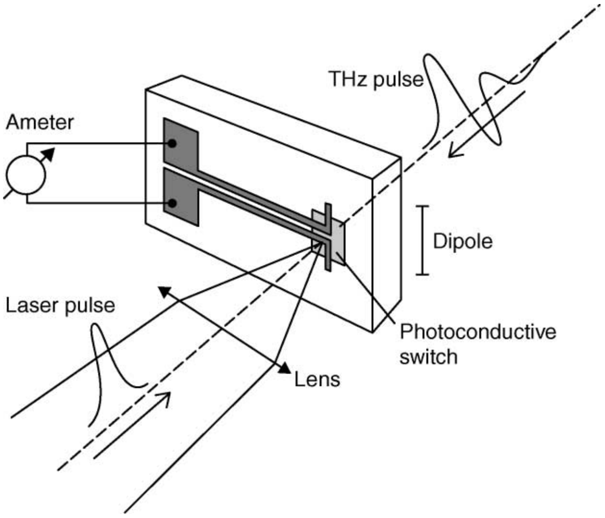

Figure 1 Photoconducting dipole antenna. Reproduced with permission from Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

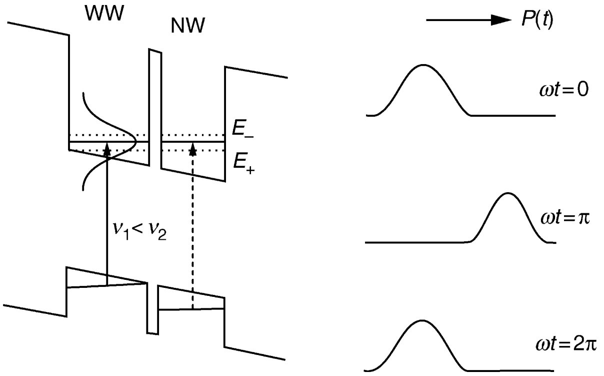

Figure 2 Schematic band diagram of an asymmetric double quantum well structure. A short laser pulse with a spectral width larger than the intersub-band splitting can excite an electronic wavepacket which tunnels back and forth between both wells. Reproduced with permission from Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216. Copyright (1992) by the American Physical Society.

and to a large number of studies of different types of other semiconductors in various magnetic field geometries. Currently, there is no unified model of the enhancement mechanism(s).

The use of a semiconductor heterostructure by Roskos *et al.* in 1992 (a coupled double quantum well) marked the advent of a new class of semiconductor THz emitters that can be deliberately designed using bandgap engineering. In these devices the quantum nature of electrons in coupled potential wells is used to generate electron wavepackets that tunnel back and forth between different potential wells (Figure 2). The coherent charge oscillation is triggered by a femtosecond laser pulse that has a spectral width that is larger than the separation between the quantized states in the well. In this way a coherent superposition of these levels is created. THz emission from these quantum beats has also been observed from charge oscillations due to light and heavy hole excitons in a quantum well by Planken and co-workers. The concept of quantum beats can even be transferred to the high-THz regime (tens of THz) which was successfully demonstrated by Bonvalet and co-workers. By using an InAs step quantum well they generated THz radiation at a frequency of 30 THz.

Bloch oscillations, one of the fundamental concepts of solid state physics, were predicted in 1934 by Zener. He stated that electrons in a periodic potential under an applied electric field should perform oscillations. Using a quantum mechanical picture, Bloch oscillations can be described as a quantum beat of Wannier–Stark states and in 1993 Waschke *et al.* detected THz emission from these oscillations.

These heterostructure THz sources are inherently frequency tunable since the energy level spacing can be altered by an applied electric field that tilts the quantum wells. The Bloch frequency is given by:

$$\omega_{\mathrm{Bloch}} = \frac{eFd}{\hbar} \qquad [1]$$

where F is the applied electric field and d is the superlattice period. Unfortunately, these devices suffer from their low emission power.

An alternative concept is the use of collective excitations in semiconductor bulk material or heterostructures. Dekorsy and co-workers showed in 1995 that coherent phonons emit THz radiation when excited by femtosecond laser pulses. In the same year Sha and co-workers studied the absorption changes within an electrically biased GaAs p-i-n structure and found distinct oscillations which they ascribed to another collective phenomenon, namely coherent plasma oscillations of the photogenerated carriers (Figure 3). This explanation was supported by the square-root dependence of the observed frequencies on the photogenerated carrier density which is predicted by a simple oscillator model:

$$\omega_{\mathrm{p}} = \sqrt{\frac{ne^2}{\varepsilon\varepsilon_0 m^*}} \qquad [2]$$

where n is the carrier density, ε the dielectric constant, and m^* the effective mass of the carriers. Fischler and co-workers confirmed this finding and in particular investigated the damping behavior of the plasmons. In 1997, Kersting *et al.* detected THz radiation from

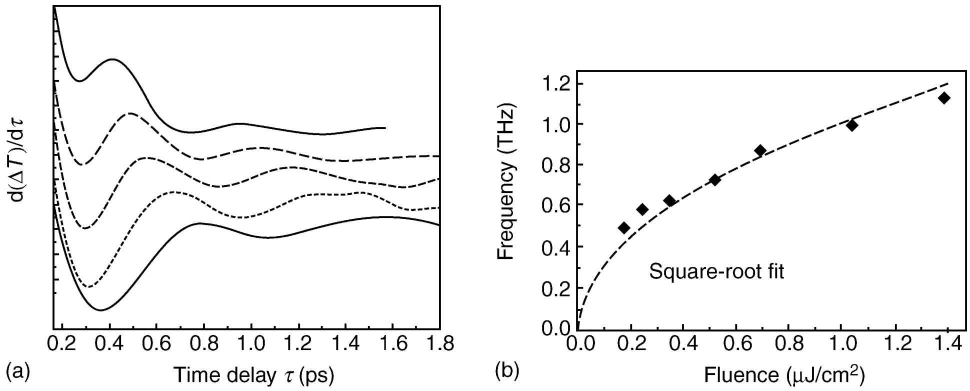

Figure 3 (a) Time-derivative of the charge in electro-absorption versus pump-probe time delay for decreasing fluences (top to bottom respectively); (b) the frequency of the oscillations as a function of fluence. The dashed curve is a square-root function fit. Reproduced with permission from Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273. Copyright (1995) by the American Physical Society.

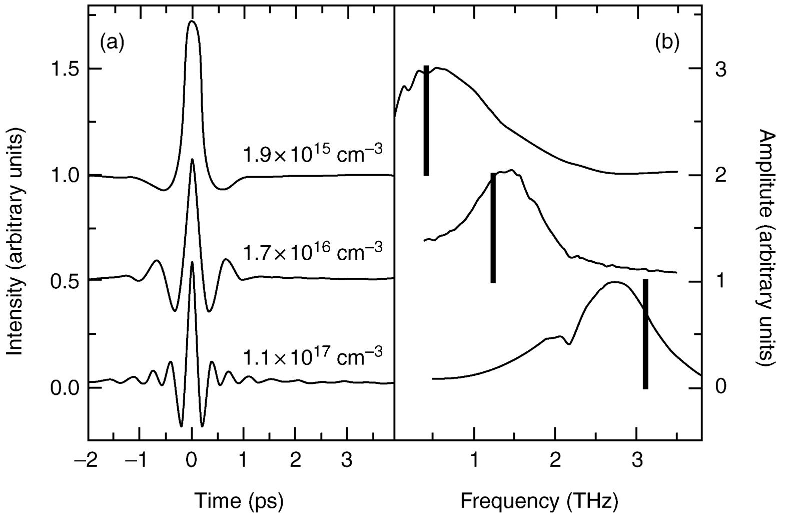

Figure 4 (a) THz autocorrelation traces for three different doping densities; (b) THz spectra; lines denote the frequencies calculated by formula [2]. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmas in n-doped GaAs. *Physical Review B* 58: 4553. Copyright (1998) by the American Physical Society.

these p-i-n structures and identified heating of the photogenerated carriers as the origin for the strong damping of the plasma oscillations resulting in a low emission intensity.

This problem can be circumvented by using coherent plasma oscillations of extrinsic charge carriers. These electrons are cold compared to the photo-excited carriers and should undergo less scattering. Kersting *et al.* showed that THz pulses are indeed emitted from coherent three-dimensional (3D) plasma oscillations in n-doped GaAs epilayers. The THz pulses are temporally and spatially coherent with intensities up to 100 nW.

Figure 4 shows the THz emission from three GaAs epilayer samples which have been doped with densities 1.9×10^{15}, 1.7×10^{16}, and $1.1 \times 10^{17}\,\mathrm{cm}^{-3}$. The curves are recorded using an autocorrelation technique: 13 nJ, 100 fs laser pulses at a repetition rate of 80 MHz and a central wavelength of 800 nm are split into two pulses with a Michelson interferometer. Every visible pulse generates a THz pulse in the transmission or pseudo-reflection direction. After collimating and focusing of the THz radiation by parabolic mirrors, the interference of the two THz pulses can be recorded by changing the time delay between the visible pulses (Figure 5). Assuming negligible nonlinear interactions between the two excitation processes, the linear autocorrelation function can be recorded. The spectrum is calculated by simply Fourier transforming (Figure 4b). The experiments are performed at room temperature, and the whole setup is purged with dry nitrogen gas to prevent

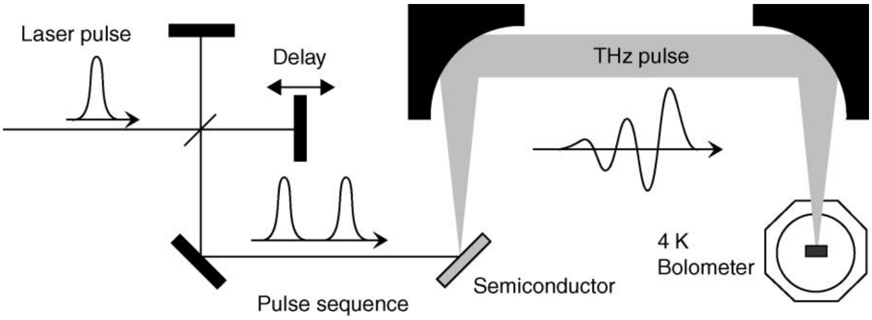

Figure 5 THz autocorrelation setup. Reproduced with permission from Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553. Copyright (1998) by the American Physical Society.

water vapor absorption. Although the autocorrelation traces are recorded at the same optical excitation density of $n_{ex} = 2 \times 10^{16}\ \mathrm{cm}^{-3}$, the emission frequency changes significantly. In addition, the autocorrelation trace of the highly doped structure ($1.1 \times 10^{17}\ \mathrm{cm}^{-3}$) shows multiple oscillations while in the case of the sample with the lowest doping ($1.9 \times 10^{15}\ \mathrm{cm}^{-3}$) the slow oscillation is damped out rather quickly. The frequency behavior of the three different samples can be analyzed from the spectra shown in Figure 4b. The bars indicate the frequency calculated with [2] and setting n equal to the doping density determined by standard Hall measurements. Obviously, the calculated frequencies match very well the experimentally determined ones. This indicates that the THz emission results from a collective oscillation of the extrinsic electrons within the epilayer. This is also supported by the observation that a variation of the excitation density of the photogenerated carriers over two orders of magnitude does not change the THz emission frequency. This is in clear contrast to the experiments with undoped semiconductors where the emission frequency shows a square-root dependence on the optically excited carrier density.

Excitation of the coherent plasma oscillation results from the ultrafast screening of the surface electric field by the femtosecond laser pulse. Before excitation, the doping electrons are confined between the surface electric field and the undoped GaAs substrate (Figure 6). A short laser pulse with an energy above the bandgap of the semiconductor generates electron–hole pairs that screen this surface electric field for a short time. The doping electrons inside the material respond to the change in potential at the surface of the sample leading to the oscillation behavior. A detailed model calculation of the excitation mechanism can be found in Kersting *et al.* (1998). The oscillation frequency – the eigenfrequency of the electron system – is simply the plasma frequency given by the simple formula [2].

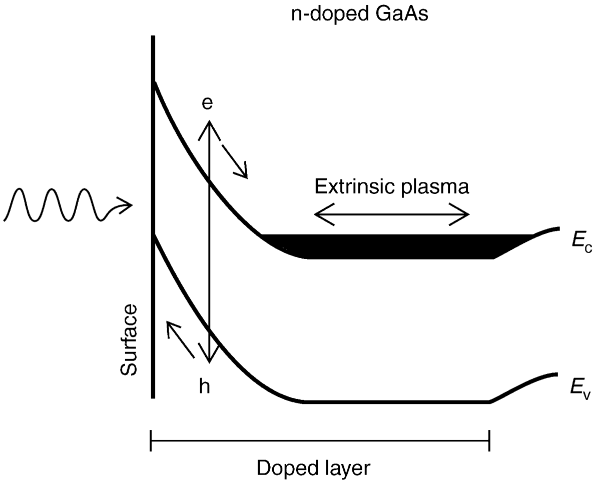

Figure 6 Schematic drawing of the 3D plasmon excitation. After Kersting R, Bratschitsch R, Heyman JN and Unterrainer K (1999) Unpublished.

There are two mechanisms that are responsible for the damping of the plasma oscillations. If the doping density is lower than $10^{16}\ \mathrm{cm}^{-3}$ the coherence time of the 3D plasmon is mainly limited by optical phonon scattering. This can be verified by measuring the plasmon damping times at different temperatures and comparing the results with mobility data which show the same tendency. Since optical phonon scattering is suppressed at low temperatures, the THz emission intensity can be substantially increased by cooling the sample. In the high-doping-density regime $>10^{16}\ \mathrm{cm}^{-3}$ the observed damping times are shorter than those expected due to the optical phonon scattering mechanism alone. This is because in this regime impurity scattering is the dominant scattering mechanism.

One way around this problem is the use of a remotely doped structure. In this case the scatterers are not in the same region as the electrons that perform the plasma oscillation (Figure 7). The higher mobility should lead to a longer decoherence time.

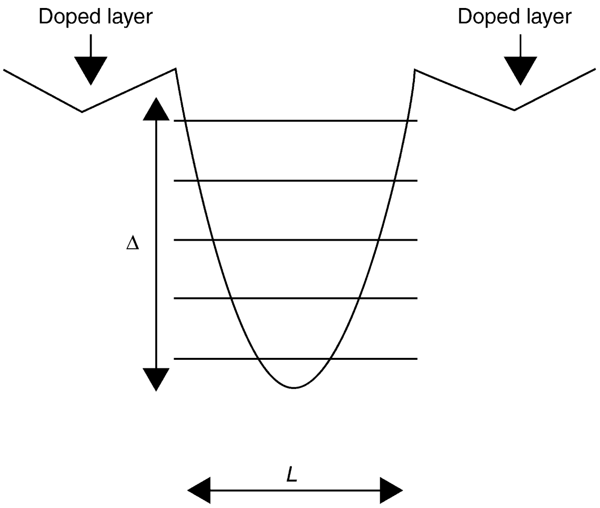

Figure 7 Schematic drawing of a modulation-doped PQW. The doped layers are outside the quantum well.

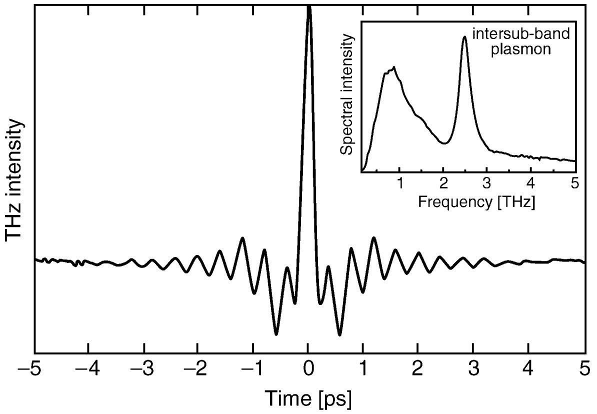

Figure 8 THz autocorrelation signal of a 1400 Å PQW excited by 780 nm laser pulses ($T = 5$ K). Inset: Fourier transform of the recorded autocorrelation. Reproduced with permission from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent THz emission from optically pumped intersubband plasmons in parabolic quantum wells. In: *Ultrafast phenomena XII*. Springer series in Chemical Physics, vol. 66, pp. 204. New York: Springer-Verlag.

For this purpose Bratschitsch *et al.* used modulation doped GaAs/AlGaAs parabolic quantum wells (PQWs), with widths L in the range of 1200–2000 Å and carrier sheet densities n_s of 1.7×10^{11}–5×10^{11} cm^{-2}. The electrons confined within the parabolic potential form a so-called quasi-3D electron system. Figure 8 shows the THz autocorrelation trace of a modulation doped PQW ($L = 1400$ Å, $n_s = 5 \times 10^{11}$ cm^{-2}) excited by 780 nm laser pulses. The density of the optically generated carriers is kept well below the carrier density due to the modulation doping inside the PQW.

The spectrum of the emitted THz radiation (inset of Figure 8) consists of two components: a broad background and a narrow line (FWHM: 0.3 THz) with a center frequency of 2.55 THz. The low-frequency broadband component is found for all the different PQW samples and is independent of the PQW sample structure. It can also be observed at room temperature while the narrowband component completely vanishes. The origin of the broadband component is due to THz generation at the surface of the sample.

The THz emission at 2.55 THz can be observed within a wide range of excitation wavelengths (815–760 nm) above the 819 nm GaAs bandgap ($T = 5$ K). When varying the excitation density over one order of magnitude ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}), the emission frequency does not change. The narrowband emission results from the oscillation of the carriers inside the PQW. The observed frequency corresponds to the intersub-band plasmon of the PQW. The expected frequency of the intersub-band plasmon is given by

$$\omega_0 = \sqrt{\frac{8\Delta}{L^2 m^*}} \qquad [3]$$

where Δ is the energetic depth of the well, L is the well width, and m^* the effective mass. For the 1400 Å wide PQW the formula yields 2.2 THz for the intersubband plasmon frequency.

The excitation mechanism for the intersub-band plasmon is due to screening of the surface depletion field by the electron–hole pairs injected by the ultrafast laser pulse. In this way the electrons inside the quantum well experience a kick and begin to oscillate with their eigenfrequency (Figure 9). This is supported by the fact that the oscillation can be excited over a large wavelength range of the femtosecond pulses. The onset of intersub-band plasmon emission appears at 815 nm near the 819 nm GaAs bandgap ($T = 5$ K) and can be observed to up to 760 nm. This large range implies that the excitation mechanism is clearly a nonresonant phenomenon, in contrast to the THz quantum beat experiments. Due to ultrafast field screening, a symmetric nanostructure can emit optically driven THz radiation.

The width of the intersub-band plasmon emission line is 0.3 THz (full width at half maximum; FWHM) for the 1400 Å PQW and even 0.2 THz for the 2000 Å PQW sample. The linewidth is a factor of 5 smaller than the smallest linewidth measured for a 3D plasmon emitter. The width of the intersub-band emission is nearly unaffected when increasing the excitation wavelength from near bandgap to approximately 60 nm above bandgap. However, it shows a linear dependence ($5.6 \times 10^{10} - 6.2 \times 10^{11}$ cm^{-2}) on

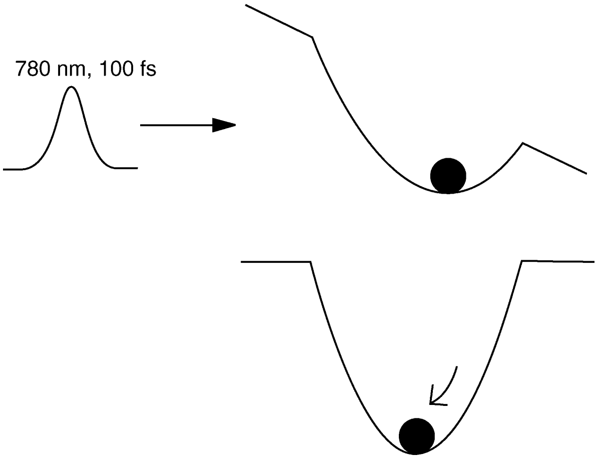

Figure 9 Symbolic drawing of the excitation mechanism: an oscillation of the carriers in the parabolic potential is initiated by ultrafast field screening. Reproduced from Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501, with permission from the American Institute of Physics.

the number of optically excited carriers. These optically injected carriers are added to those from the modulation doping and cause a faster dephasing of the intersubband plasmon oscillation. This linear dependence due to carrier–carrier scattering has also been observed in Wolter's experiments dealing with the dephasing of Bloch oscillations in semiconductor superlattices. The linewidth stays constant up to about 50 K (0.33 THz = 1.36 meV) and then starts to broaden until the sample temperature reaches 150 K where the PQW emission is buried by the surface-generated THz emission. The broadening of the line between $T = 50$ K and $T = 150$ K can be explained by the onset of optical phonon scattering in addition to interface roughness scattering which is the dominant scattering mechanism at low temperatures. The onset and linear increase of the line broadening has also been observed with Fourier Transform Infrared (FTIR) absorption and THz time-domain cross-correlation spectroscopy measurements by Kersting and co-workers.

The combination of the designability of the transition frequency, the narrowband emission, and the absence of any processing of the sample make modulation-doped PQWs attractive and easy-to-use THz emitters.

Other THz emission experiments of heterostructures include the optically pumped emission of THz radiation from grating-coupled intrasub-band (2D) plasmons in a doped single quantum well observed by Sekine and co-workers or optically excited THz emission from coherent plasmons in doped superlattices detected by Bratschitsch and co-workers.

Few-Cycle THz Spectroscopy of Semiconductor Quantum Structures

There are a large number of studies that use few-cycle THz radiation for the investigation of semiconductor bulk materials. The complex index of refraction in the THz regime gives insight into the conductivity (i.e., carrier mobility) of the material. The free carriers that change the index of refraction by Drude absorption can be provided either by doping (n-doped silicon or GaAs) or by photo-excitation. An optical pump-THz probe experiment with undoped GaAs reveals the different intervalley scattering times.

There are compelling reasons to use THz radiation to excite and probe carriers in semiconductor quantum structures. The photon energies are comparable to the sub-band spacings, carrier kinetic energies, and phonon energies. In addition, THz radiation does not generate minority carriers so that the experiments directly probe free carriers rather than excitons.

Heyman *et al.* investigated a modulation-doped GaAs/AlGaAs multiple quantum well ($d = 510$ Å, 10×) with a transition energy between the two lowest sub-bands of 1.5 THz and a carrier density of $n_s = 2.75 \times 10^{10}\ \mathrm{cm}^{-2}$. On top of the sample an aluminum Schottky gate is evaporated and the well is contacted with AuGe alloyed ohmic contacts. The wells can be depleted of carriers by applying a negative gate voltage to the Schottky gate (−10 V gate bias). By modulating the voltage between 0 V and −10 V the response of the 2D electron system can be measured. Otherwise, the recorded signal is dominated by the contribution of the bulk GaAs substrate that dramatically chirps the THz pulse due to its frequency-dependent index of refraction. The THz pulses are coupled into the cleaved edge of the quantum well sample so that the electric field is parallel to the growth direction in order to excite the intersub-band transition. The electrons' response is measured with a cross-correlation technique. Here the THz radiation transmitted through the quantum well is mixed with a short THz pulse emitted from low-temperature grown GaAs to get the desired time resolution in the experiment. This allows the detection of the transmitted signal in amplitude and phase. Figure 10 shows the cross-correlation signal of the QW. The signal from the carriers rises during the first 2 ps in response to the THz field and then continues to oscillate at a constant frequency of 1.5 THz. Due to the femtosecond time resolution in this experiment it is possible to observe the phase of the electrons even during the process of excitation. Initially, the incident THz pulse generates a coherent superposition of states in the first and second sub-band. After the THz driving pulse is over

(>2 ps) the carriers continue to radiate at the intersub-band transition frequency, and the signal is damped out by the free induction decay. Since the amplitude and phase of the radiated electric field is recorded, not only the absorption due to the QW electrons but also the phase shift associated with the absorption can be extracted (Figure 11).

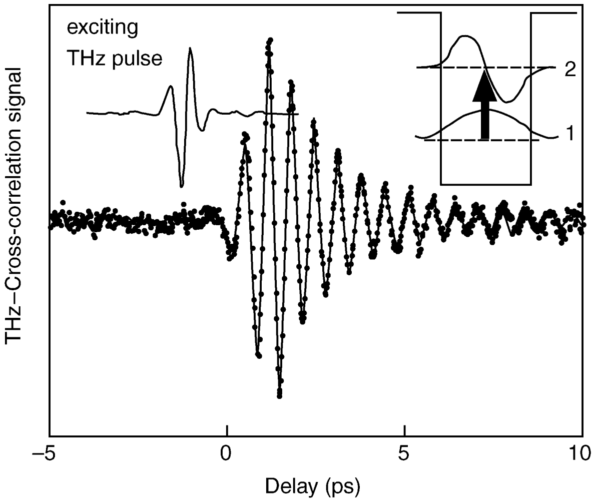

Figure 10 Cross-correlation signal obtained by electrically modulating the charge density in a 510 Å wide quantum well. The THz pulse generates a coherent superposition of the $n = 1$ and $n = 2$ sub-band states. The resulting oscillating polarization radiates and is damped out by the free induction decay. The upper trace shows the incident THz pulse. Reprinted from Unterrainer K, Kersting R, Bratschitsch R, *et al.* (2001) Few-cycle THz spectroscopy of semiconductor quantum structures. *Physica E* 9: 76, with permission from Elsevier.

This experiment showed the response of the QW electrons to radiation which is resonant to the intersub-band transition. Using the same experimental technique Kersting *et al.* investigated the response of electrons to a non-resonant excitation of the intersub-band transition. Instead of a square QW, a modulation-doped parabolic quantum well is used in this experiment which has the advantage that the transition frequency is independent of the carrier concentration (Kohn's theorem) and the applied electric field. Furthermore the THz radiation is coupled to the intersub-band transition via a metallic Schottky grating. Figure 12 shows the response of the QW electrons to resonant THz radiation. The electrons follow the driving pulse and continue to radiate at the intersub-band transition frequency. The response to a nonresonant THz pulse is shown in Figure 13. Initially, the electrons inside the QW follow the driving pulse but when the driving pulse has ended a phase jump occurs and the electrons lock to their eigenfrequency and slowly lose their coherence. This behavior can be understood by using time-dependent perturbation theory to solve Schrödinger's

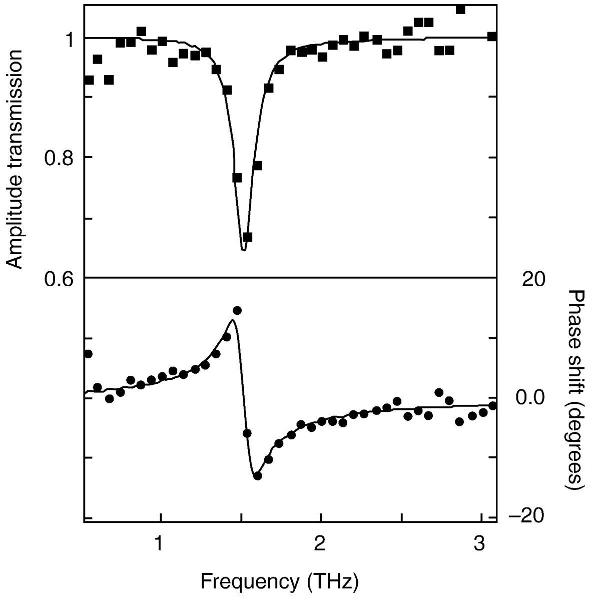

Figure 11 Change in the absorption coefficient and index of refraction in the quantum well due to the quantum well electrons. Solid points are calculated from the time-domain data. Reproduced from Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644, with permission from the American Institute of Physics.

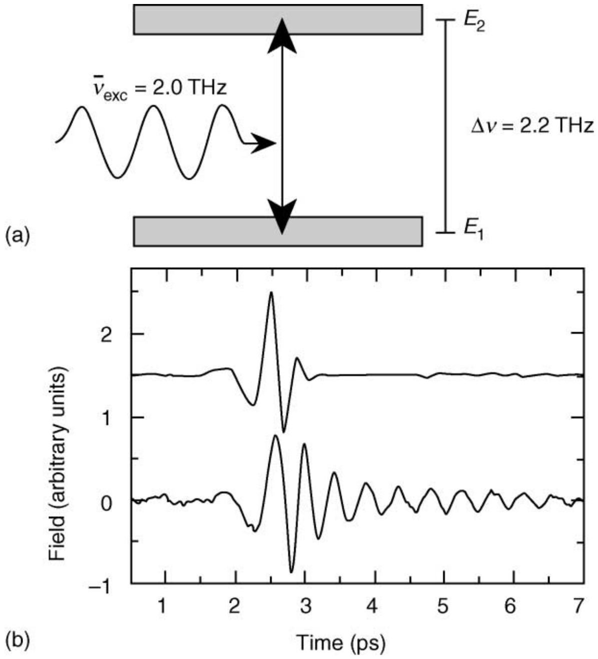

Figure 12 Resonant excitation: (a) band diagram of an intersub-band transition of 2.2 THz and driving THz field at 2.0 THz; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

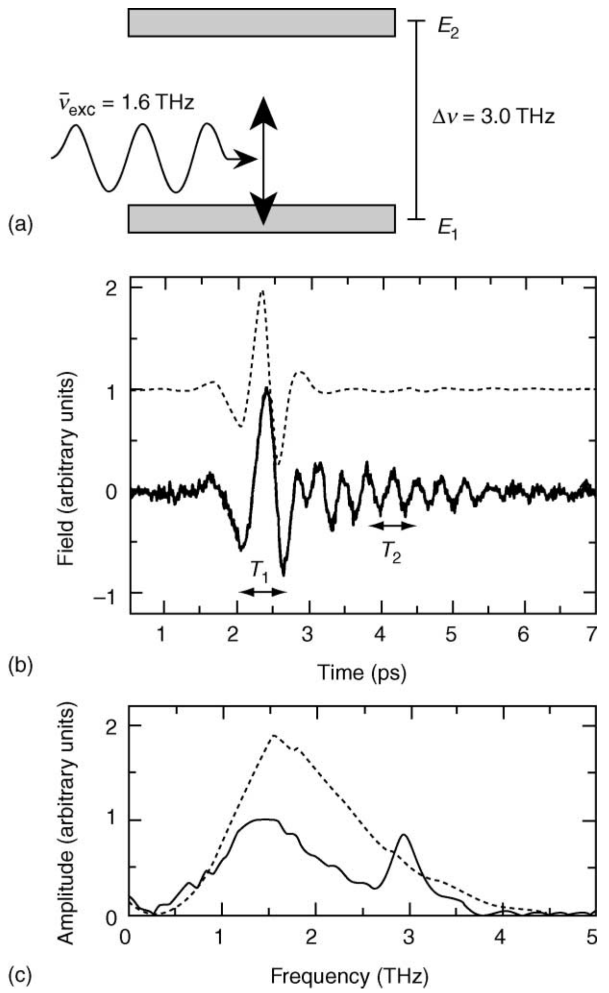

Figure 13 Nonresonant excitation: (a) band diagram of an intersub-band transition of 3.0 THz and driving THz field at 1.6 THz; the oscillation frequency changes from T_1 to T_2; (b) exciting THz pulse (upper curve) and cross-correlation signal (lower curve); (c) spectra of the exciting THz pulse (dotted curve) and cross-correlation signal (solid curve). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

equation for a two-level system with an electromagnetic driving field. While the two-level system is treated quantummechanically, the dipole Hamiltonian is used for the incoming THz radiation: $H = H^0 + H'(t)$ and $H'(t) = \mu E(t)$. Making the ansatz $\Psi(t) = a_1(t)\Psi_1 + a_2(t)\Psi_2$ one gets a system of two coupled equations for $a_1(t)$ and $a_2(t)$. They can be decoupled by the condition that the population in the lower level is close to one $(||a_1^2|| \approx 1)$ since the THz radiation field is not very strong. Figure 14 shows the result of the model calculation that reproduces the change in frequency as well as the observed phase jump after the driving pulse is over.

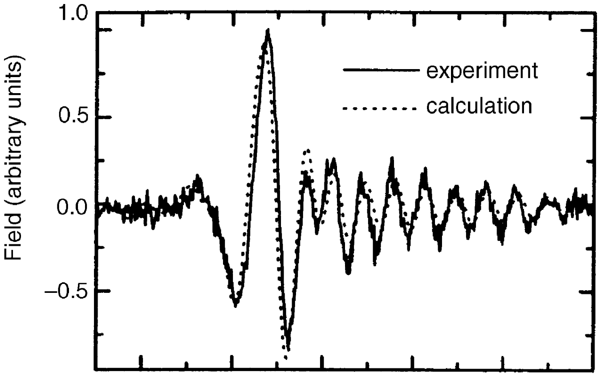

Figure 14 Model calculation of the polarization (dotted curve) and experimental data (solid line). Reproduced from Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272, with permission from the Optical Society of America.

These experiments are all performed in the linear regime but with nonadiabatic excitation conditions (half-cycle THz pulses). The excellent time resolution together with sufficiently high THz intensities can open up a previously unexplored field of quantum optics beyond the slowly varying envelope approximation.

Conclusions

The recent progress in the development of semiconductor sources together with room temperature THz detection techniques (free-space electro-optic sampling) paves the way for the commercialization of THz technology. In particular, spectrally sensitive THz imaging seems to have great potential in medicine and quality control of products.

See also

Semiconductor Materials: GaAs Based Compounds. **Semiconductor Physics:** Outline of Basic Electronic Properties; Quantum Wells and GaAs-based Structures.

Further Reading

Auston DH, Cheung KP and Smith PR (1984) Picosecond photoconducting Hertzian dipoles. *Applied Physics Letters* 45: 284.

Bonvalet A, Nagle J, Berger V, Migus A, Martin J-L and Joffre M (1996) Femtosecond infrared emission resulting from coherent charge oscillations in quantum wells. *Physical Review Letters* 76: 4392.

Bratschitsch R, Kersting R, Müller T, Strasser G, Unterrainer K, Fischler W and Höpfel RA (1999) Coherent THz plasmons in GaAs/AlGaAs superlattices. *Physica B* 272/1-4: 375.

Bratschitsch R, Müller T, Kersting R, Strasser G and Unterrainer K (2000) Coherent terahertz emission from optically pumped intersubband plasmons in parabolic quantum wells. *Applied Physics Letters* 76: 3501.

Chamberlain M (ed.) (1999) *Terahertz Spectroscopy and Applications*, SPIE Proceedings (Europto Series), vol. 3828, pp. 1–454.

Dekorsy T, Auer H, Waschke C, Bakker H, Roskos H, Kurz H, Wagner V and Grosse P (1995) Emission of submillimeter electromagnetic waves by coherent phonons. *Physical Review Letters* 74: 738.

Gornik E and Kersting R (2001) Coherent THz emission in semiconductors. In: Tsen KT (ed.) *Ultrafast Physical Processes in Semiconductors*, vol. 25, pp. 396. New York: Academic Press.

Heyman J, Kersting R and Unterrainer K (1998) Time-domain measurement of intersubband oscillations in a quantum well. *Applied Physics Letters* 72: 644.

Kersting R, Unterrainer K, Strasser G, Kauffmann H and Gornik E (1997) Few-cycle thz emission from cold plasma oscillations. *Physical Review Letters* 79: 3038.

Kersting R, Heyman J, Strasser G and Unterrainer K (1998) Coherent plasmons in n-doped GaAs. *Physical Reviews B* 58: 4553.

Kersting R, Bratschitsch R, Strasser G, Unterrainer K and Heyman J (2000) Sampling a THz dipole transition with sub-cycle time-resolution. *Optics Letters* 25: 272.

Nun MC and Orenstein J (1997) Terahertz time-domain spectroscopy. In: Gruener G (ed.) *Millimeter Wave Spectroscopy of Solids*. Berlin: Springer.

Roskos H, Nuss M, Shah J, Leo K, Miller D, Fox A, Schmitt-Rink S and Köhler K (1992) Coherent submillimeter-wave emission from charge oscillations in a double-well potential. *Physical Review Letters* 68: 2216.

Sarukura N, Ohtake H, Izumida S and Liu Z (1998) High average power radiation from femtosecond laser-irradiated InAs in a magnetic field and its elliptical polarization characteristics. *Journal of Applied Physics* 84: 654.

Sekine N, Yamanaka K, Hirakawa K, Voßebürger M, Haring-Bolivar P and Kurz H (1999) Observation of terahertz radiation from higher-order two-dimensional plasmon models in GaAs/AlGaAs single quantum wells. *Applied Physics Letters* 74: 1006.

Sha W, Smirl A and Tseng W (1995) Coherent plasma oscillations in bulk semiconductors. *Physical Review Letters* 74: 4273.

Shah J (1999) *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. Springer Series in Solid-State Sciences, vol. 115. Berlin: Springer.

Waschke C, Roskos H, Schwedler R, Leo K, Kurz H and Köhler K (1993) Coherent submillimeter-wave emission from bloch oscillations in a semiconductor superlattice. *Physical Review Letters* 70: 3319.

Wolter F, Martini R, Tolk S, *et al.* (1999) Influence of carrier–carrier scattering on intraband dephasing. *Superlattice and Microstructure* 26: 93.

Zhang X-C, Hu B, Darrow J and Auston D (1990) Generation of femtosecond electromagnetic pulses from semiconductor surfaces. *Applied Physics Letters* 56: 1011.

TIME-RESOLVED FLUORESCENCE

Contents

Laser Applications
Measurements in Polymer Science

Laser Applications

B Valeur, Conservatoire National des Arts et Métiers, Paris, France

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Fluorescence is extensively used in physical, chemical, material, biological, and medical sciences, for investigating the structure and dynamics of matter or living systems. After excitation of a fluorescent sample by a short pulse of light, the fluorescence decay is, in the simplest case, a single exponential whose time constant is the lifetime of the excited state of the fluorescent species (i.e., the average time during which these species stay in the excited state). However, interactions of a fluorescent species with its micro-environment often perturb the fluorescence decay so that information on this micro-environment can be obtained. More specifically, when intermolecular photophysical processes (such as electron transfer, proton transfer, energy transfer, excimer, or exciplex formation) occur during the lifetime of the excited state, competition with normal de-excitation results in a fluorescence decay that contains the rate constants for those processes. Consequently, the fluorescence decay is generally a sum of discrete exponentials, or a more complicated function; for example, sometimes the system is characterized by a distribution of decay times. Thus it is of major interest to recover the parameters characterizing the δ-pulse response of a fluorescent sample (i.e., the response to an iznfinitely short pulse of light expressed as the Dirac function δ). With this aim, two time-resolved techniques, pulse fluorometry and phase-modulation fluorometry, are usually employed.

In pulse fluorometry, the sample is excited by a short pulse of light, and the fluorescence response is the δ-pulse response of the sample convoluted by the instrument response. In phase-modulation fluorometry, the sample is excited by modulated light at variable frequency, and the harmonic response of the sample obtained is the Fourier transform of the δ-pulse response. The first technique works in the time domain, and the second technique in the frequency domain. Pulse fluorometry and phase-modulation fluorometry are theoretically equivalent, but the principles of instrumentation are different. In both techniques, lasers are currently used as sources of pulsed or modulated light.

It is the aim of this article to explain the principles of pulse and phase fluorometries using lasers as light sources, and to compare these techniques.

General Principles of Time-Resolved Fluorometry

The principles of time-resolved fluorometry are illustrated in Figure 1. For any excitation function $E(t)$, the fluorescence response $R(t)$ of the sample is the convolution product of this function by the δ-pulse response $I(t)$:

$$R(t) = E(t)\otimes I(t) = \int_{-\infty}^{t} E(t')I(t-t')\,\mathrm{d}t' \qquad [1]$$

In pulse fluorometry, the sample is excited by a short pulse of light and the fluorescence response is recorded as a function of time. If the duration of the pulse is not short with respect to the time constants of the fluorescence decay, the fluorescence response is the convolution product given by eqn [1]: the fluorescence intensity increases, goes through a maximum, and becomes identical to the true δ-pulse response $i(t)$ as soon as the intensity of the light pulse becomes negligible. In this case, data analysis for the determination of the parameters characterizing the δ-pulse response requires deconvolution of the fluorescence response.

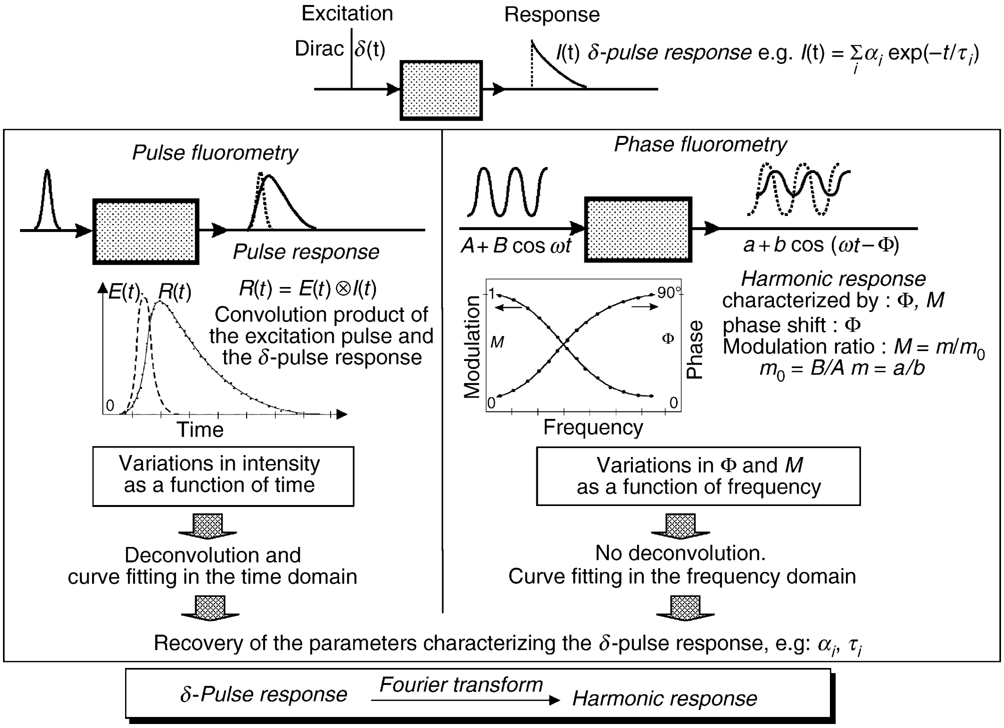
Excitation
Dirac δ(t)
Response
I(t) δ-pulse response e.g. I(t) = Σ_i α_i exp(−t/τ_i)
Pulse fluorometry
Pulse response
E(t), R(t)
R(t) = E(t) ⊗ I(t)
Convolution product of the excitation pulse and the δ-pulse response
0
Time
Variations in intensity as a function of time
Deconvolution and curve fitting in the time domain
Phase fluorometry
A + B cos ωt
a + b cos (ωt − Φ)
Harmonic response characterized by : Φ, M
phase shift : Φ
Modulation ratio : M = m/m0
m0 = B/A m = a/b
Modulation
1
M
0
90°
Φ
Phase
0
Frequency
Variations in Φ and M as a function of frequency
No deconvolution.
Curve fitting in the frequency domain
Recovery of the parameters characterizing the δ-pulse response, e.g: α_i, τ_i
δ-Pulse response
Fourier transform
Harmonic response

Figure 1 Principles of time-resolved fluorometry.

In phase fluorometry, the sample is excited by a sinusoidally modulated light at high frequency. The fluorescence response, which is the convolution product (eqn [1]) of the δ-pulse response by the excitation function, is sinusoidally modulated at the same frequency, but delayed in phase and partially demodulated with respect to the excitation. The modulation ratio M is defined as the ratio of the modulation depth m (AC/DC ratio) of the fluorescence and the modulation depth of the excitation m_0 (see Figure 1). The phase shift Φ and the modulation ratio M characterize the harmonic response of the system. These parameters are measured as a function of the modulation frequency. No deconvolution is necessary because the data are directly analyzed in the frequency domain.

The relationship between the harmonic response and the δ-pulse response can be obtained as follows. When the excitation function is sinusoidal, i.e., of the following form:

$$E(t) = E_0[1 + m_0 \exp(i\omega t)] \qquad [2]$$

where ω is the angular frequency $(= 2\pi f)$, the response of the system can be calculated using eqn [1]. It can then be shown that

$$M \exp(-j\Phi) = \int_0^\infty i(t) \exp(-i\omega t)\, dt \qquad [3]$$

where $i(t)$ is the normalized δ-pulse response according to

$$\int_0^\infty i(t)\, dt = 1 \qquad [4]$$

This important expression shows that the harmonic response expressed as $M \exp(-j\Phi)$ is the Fourier transform of the δ-pulse response.

It is convenient to introduce the sine and cosine transforms P and Q of the δ-pulse response:

$$P = \int_0^\infty i(t) \sin(\omega t)\, dt \qquad [5]$$

$$Q = \int_0^\infty i(t) \cos(\omega t)\, dt \qquad [6]$$

If the δ-pulse response is not normalized according to eqn [4], then eqns [5] and [6] should be replaced by

$$P = \frac{\int_0^\infty I(t) \sin(\omega t)\, dt}{\int_0^\infty I(t)\, dt} \qquad [7]$$

$$Q = \frac{\int_0^\infty I(t) \cos(\omega t)\, dt}{\int_0^\infty I(t)\, dt} \qquad [8]$$

Since eqn [3] can be rewritten as $M \cos \Phi - iM \times \sin \Phi = Q - iP$, it is easy to show that the phase shift and the modulation ratio are given by

$$\Phi = \tan^{-1}\left(\frac{P}{Q}\right) \qquad [9]$$

$$M = [P^2 + Q^2]^{1/2} \qquad [10]$$

For a single exponential decay, the δ-pulse response is

$$I(t) = \alpha \exp(-t/\tau) \qquad [11]$$

where τ is the decay time and α is the pre-exponential factor or amplitude. The phase shift and relative modulation are related to the decay time by

$$\tan \Phi = \omega\tau \qquad [12]$$

$$M = \frac{1}{(1 + \omega^2\tau^2)^{1/2}} \qquad [13]$$

For a multi-exponential decay with n components, the δ-pulse response is

$$I(t) = \sum_{i=1}^{n} \alpha_i \exp(-t/\tau_i) \qquad [14]$$

Note that the fractional intensity of component i, i.e., the fractional contribution of component i to the total steady-state intensity, is

$$f_i = \frac{\int_0^\infty I_i(t)\, dt}{\int_0^\infty I(t)\, dt} = \frac{\alpha_i\tau_i}{\sum_{i=1}^{n} \alpha_i\tau_i} \qquad [15]$$

with, of course, $\sum_{i=1}^{n} f_i = 1$.

Using eqns [7] and [8], the sine and cosine Fourier transforms, P and Q, are given by

$$P = \frac{\omega \sum_{i=1}^{n} \frac{\alpha_i\tau_i^2}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \omega \sum_{i=1}^{n} \frac{f_i\tau_i}{1 + \omega^2\tau_i^2} \qquad [16]$$

$$Q = \frac{\sum_{i=1}^{n} \frac{\alpha_i\tau_i}{1 + \omega^2\tau_i^2}}{\sum_{i=1}^{n} \alpha_i\tau_i} = \sum_{i=1}^{n} \frac{f_i}{1 + \omega^2\tau_i^2} \qquad [17]$$

It should be noted that the optimum frequency for decay time measurements using either the phase shift or the modulation ratio is, according to eqns [12] and [13], such that $\omega\tau$ is close to 1, i.e., $f \approx 1/(2\pi\tau)$. Therefore, for decay times of 10 ps, 1 ns and 100 ns, the optimum frequencies are about 16 GHz, 160 MHz and 1.6 MHz, respectively.

In the case of a single exponential decay, eqns [12] and [13] provide two independent ways of measuring the decay time:

– by phase measurements:

$$\tau_\Phi = \frac{1}{\omega} \tan \Phi \qquad [18]$$

– by modulation measurements:

$$\tau_M = \frac{1}{\omega}\left(\frac{1}{M^2} - 1\right)^{1/2} \qquad [19]$$

The values measured in these two ways should, of course, be identical and independent of the modulation frequency. This provides two criteria to check whether an instrument is correctly calibrated, by using a lifetime standard whose fluorescence decay is known to be a single exponential.

It should be noted that the measurement of a decay time is fast (a fraction of a second) for a single exponential decay since a single frequency suffices. It is also of interest that a significant difference between values obtained by means of eqns [18] and [19] is compelling evidence for nonexponentiality of the fluorescence decay.

Pulse Fluorometers

The most widely used technique in time domain is the time-correlated single-photon counting technique, preferably called single-photon timing technique (SPT). Time-gated systems are less popular. Streak cameras offer a very good time resolution (a few picoseconds or less) but the dynamic range is smaller than that of the single photon-timing technique. The instruments that provide the best time resolution (a few tens of femtoseconds) are based on fluorescence up-conversion, but they are very expensive and not commercially available. Because of space limitations, only the single-photon timing technique will be presented below.

Principle of the Single-Photon Timing Technique

The basic principle of this technique relies on the fact that the probability to detect a single photon at time t, after an exciting pulse, is proportional to the fluorescence intensity at that time. After timing and recording the single photons following a large number of exciting pulses, the decay of the fluorescence intensity is reconstructed.

Figure 2 shows a schematic diagram of a conventional single-photon counting instrument using a pulsed laser as an excitation source. An electrical pulse associated to the optical pulse is generated (e.g. by a photodiode or the electronics associated to the excitation source) and routed – through a constant-fraction discriminator – to the start input of the time-to-amplitude converter (TAC). Meanwhile, the sample is excited by the optical pulse and emits fluorescence. The optics is tuned (e.g., by means of a neutral density filter) so that the photomultiplier detects no more than one fluorescence photon for each exciting pulse. The corresponding electrical pulse is routed – through a constant-fraction discriminator – to the stop input of the TAC. The latter generates an output pulse whose amplitude is directly proportional to the delay time between the start and the stop pulses. The height analysis of this pulse is achieved by an analog-to-digital converter and a multichannel

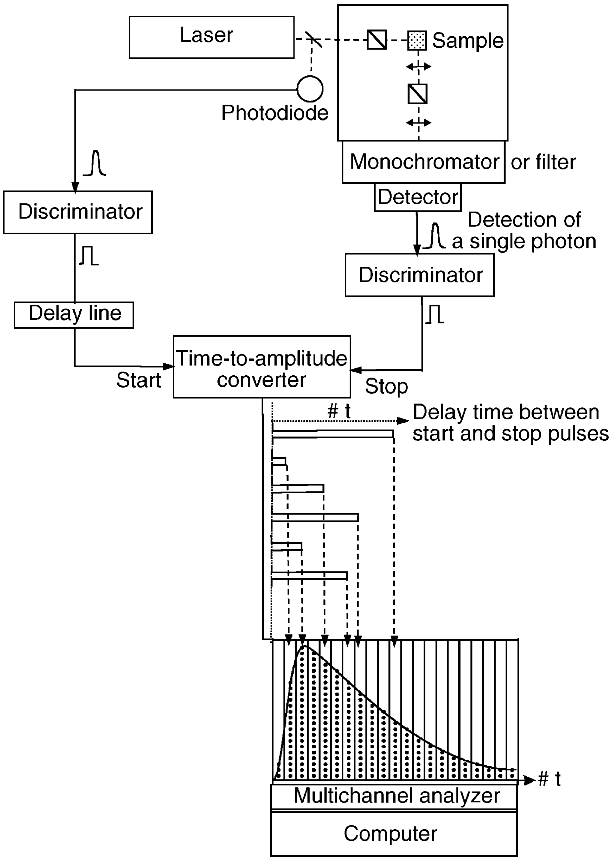

Figure 2 Schematic diagram of a single-photon timing fluorometer.

analyzer (MCA), which increases by one the contents of the memory channel corresponding to the digital value of the pulse. After a large number of excitation and detection events, the histogram of pulse heights represents the fluorescence decay curve.

Obviously, the larger the number of events, the better the accuracy of the decay curve. The required accuracy depends on the complexity of the δ-pulse response of the system; for instance, a high accuracy is, of course, necessary for recovering a distribution of decay times.

When deconvolution is required, the time profile of the exciting pulse is recorded under the same conditions by replacing the sample with a scattering solution (e.g., a suspension of colloidal silica (Ludox)).

It is important to note that the number of detected fluorescence photons must be kept much smaller than the number of exciting pulses ($<0.01-0.05$ stops per pulse), so that the probability of detecting two fluorescence photons per exciting photon is negligible. Otherwise, the TAC will take into account only the first fluorescence pulse and the counting statistics will be distorted: the decay will appear shorter than it is in reality. This effect is called the 'pile-up effect'.

Laser Sources for the Single-Photon Timing Technique

Mode-locked dye lasers and titanium-doped sapphire lasers are well suited to the single-photon timing technique: in fact, they can generate pulses over broad wavelength ranges, and the pulse widths are in the femtosecond and picosecond ranges with a high repetition rate. Flash lamps running in air, or filled with gas, can also be used, but the range of excitation wavelengths is limited; they deliver nanosecond pulses, and the repetition rate is not high (10^4–10^5 Hz). Dye lasers require the handling of large volumes of dye solutions which must be replaced when the dye is bleached. In contrast, the Ti:sapphire laser is a solid-state laser that does not have this drawback and delivers shorter light pulses. Moreover, long operational lifetime, operational simplicity, reliability, and broad tunability are distinct advantages. For these reasons, Ti:sapphire lasers have become more popular, and are described below.

Ti:sapphire lasers

Titanium-doped sapphire crystals are obtained by replacing a small fraction of the Al atoms of the host (Al_2O_3) with Ti atoms. When the concentration of Ti is low, the optical quality and the mechanical properties of Ti:sapphire crystals are comparable to those of ruby crystals in ruby lasers. The absorption spectrum of Ti:sapphire crystals exhibits a maximum at 490 nm and a shoulder at 540 nm. Therefore, pumping can be achieved by an Argon ion laser or a green solid-state laser. The emission spectrum ranges from 600 to 1,100 nm. Such a broad emission spectrum results from electronic coupling between the electronic levels of the 3d electrons with the lattice vibrations. Since excited-state absorption is not possible in the spectral region of the pump and the fluorescence spectrum, the tuning range is broad: it extends from 680 to 1,100 nm, i.e., almost over the entire fluorescence spectrum.

The width of the pulses that can be generated by Ti:sapphire lasers ranges from a few tens of femtoseconds to about 100 picoseconds. Commercially available Ti:sapphire lasers can operate either in the femtosecond configuration or in the picosecond configuration. In time-resolved fluorescence experiments using the single-photon timing technique, the instrument response is limited by the time response of the photomultiplier; the shortest width of this response that can be measured with the faster detectors (i.e., microchannel plate photomultipliers) is about 30–40 ps (see below). Therefore, there is no reason to operate the Ti:sapphire laser in the femtosecond configuration. In the picosecond configuration, the pulse width is $<2-100$ ps.

Figure 3 shows the six-mirror folded cavity design for the picosecond configuration in a Tsunami

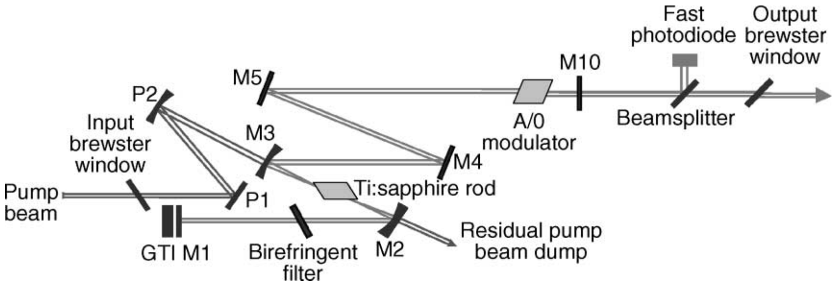

Figure 3 Beam path of the laser cavity of the Tsunami Ti:sapphire laser (Spectra Physics) operating in the picosecond configuration (Adapted from the Tsunami brochure with permission of Spectra Physics).

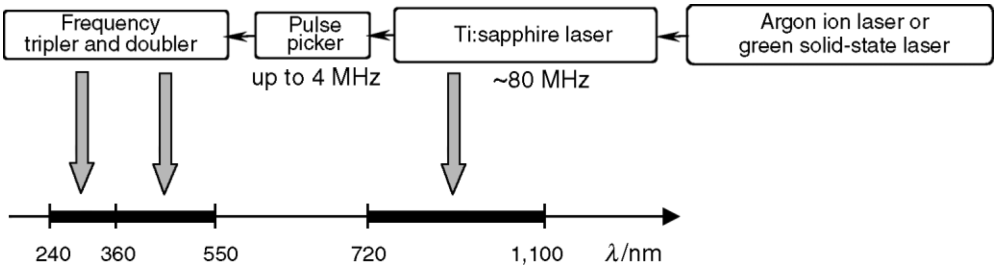
Frequency tripler and doubler
Pulse picker
up to 4 MHz
Ti:sapphire laser
~80 MHz
Argon ion laser or green solid-state laser
240
360
550
720
1,100
λ/nm

Figure 4 Schematics of the excitation system in time-resolved fluorescence experiments.

Ti:sapphire laser (Spectra Physics). A Brewster's angle acousto-optic modulator, driven by a low-power rf source, insures an 80 MHz repetition rate. Consequently, in contrast to standard passive mode-locking systems, the laser can operate for long periods without dropouts or shut-downs. Gires–Tournois Interferometer (GTI) mirror spacing is used for tuning the laser cavity. Wavelength tuning is achieved by means of a two-plate birefringent filter placed within the laser cavity at Brewster's angle. The tuning range depends on the mirror sets that are used. With broadband cavity mirrors, the entire tuning range of Ti:sapphire is covered, i.e., from below 700 nm to greater than 1,000 nm, and changes in wavelength can be made in a matter of minutes.

When pumping with a 10 W argon ion laser, the average output power is about 1.5 W at 800 nm and the pulse energy is about 15 nJ. Argon ion lasers can now be advantageously replaced by cw green solid-state lasers (532 nm) in which the gain medium ($Nd:YVO_4$) is pumped by diodes, and a LBO crystal placed inside the cavity ensures frequency doubling. These lasers offer several advantages: the head is compact and does not require external cooling water; it is insensitive to temperature, and thermal equilibration is very rapid.

Since the excited-state lifetime of fluorescent compounds can be as long as several hundreds of nanoseconds, the repetition rate of 80 MHz is too high to let the fluorescence of long lifetime samples vanish before a new exciting pulse is generated. Therefore, it must be limited to a few MHz, thanks to a pulse picker using a Bragg-angle acousto-optic modulator synchronized to the mode-locked 80 MHz pulse train. Selection of pulses from the train can thus be achieved at a maximum pulse selection rate of 8 MHz.

For excitation of fluorescent samples, wavelengths in the visible and UV regions are desirable. Frequency doublers and triplers using LBO and BBO crystals, respectively, have been designed with this aim. The possible tuning ranges are shown in Figure 4. When pumping with a 10 W argon ion laser, the average output powers are about 300 mW and 120 mW for frequency doubling and tripling, respectively.

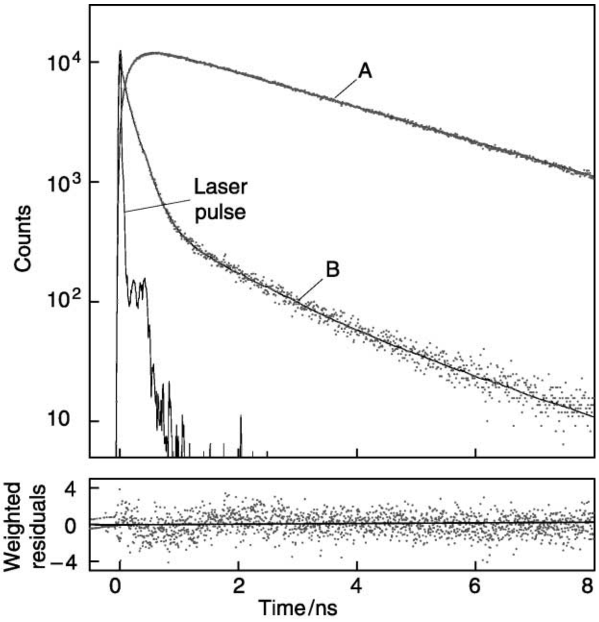

Figure 5 Example of data obtained by the single-photon timing technique using a Spectra-Physics setup composed of a titanium-sapphire Tsunami laser pumped by an argon ion laser, a pulse selector, and doubling (LBO) and tripling (BBO) crystals. The detector is a Hamamatsu MCP photomultiplier R3809U. The sample is a solution of DANS1 (a derivative of dansylamide) in a mixture CH_3CN/H_2O (60:40 v/v) at pH 0.8. Excitation wavelength: 286 nm. Channel width: 9.8 ps. A: Observation through a bandpass filter (300–370 nm) of the protonated form that undergoes photoinduced deprotonation. B: Observation through a cut-off filter ($\lambda > 400$ nm) of the neutral form; the observed risetime corresponds to the formation of the neutral form from the protonated form. Global analysis of the two curves leads to two time constants: 183 ± 9 ps and 3.01 ± 0.03 ns (Global $\chi^2_R = 1.26$) (Métivier R, Leray I and Valeur B, unpublished results. Métivier R, doctoral thesis, Ecole Normale Supérieure de Cachan, 2003).

With Ti-sapphire lasers and microchannel plate photomultipliers, decay times as short as 10 ps can be measured.

The results of an experiment are shown in Figure 5.

Pulsed diode lasers

Picosecond diode laser heads offer an interesting alternative to mode-locked lasers and are much less expensive. They can produce light pulses as short as 50–90 ps with repetition rates from single shot to 40–80 MHz. Peak powers up to 500 mW can be obtained. However, the main disadvantage is the absence of tunability, and the number of wavelengths is limited: 375, 400, 440, 635 to 1,550 nm. No wavelength below 375 nm is presently available.

Phase-Modulation Fluorometers

In the frequency domain, the phase and modulation measurements can be taken by using either a cw laser and an optical modulator, or the harmonic content of a pulsed laser.

Phase Fluorometers Using a cw Laser and an Optical Modulator

The optical modulator is generally a Pockels cell. Low-power cw lasers are sufficient and not as expensive as pulsed lasers, so that one can afford several lasers of this type in order to get various excitation wavelengths. Available wavelengths from cw lasers are given in Table 1. A Xenon lamp can also be used; however, the optical modulator works better with a cw laser.

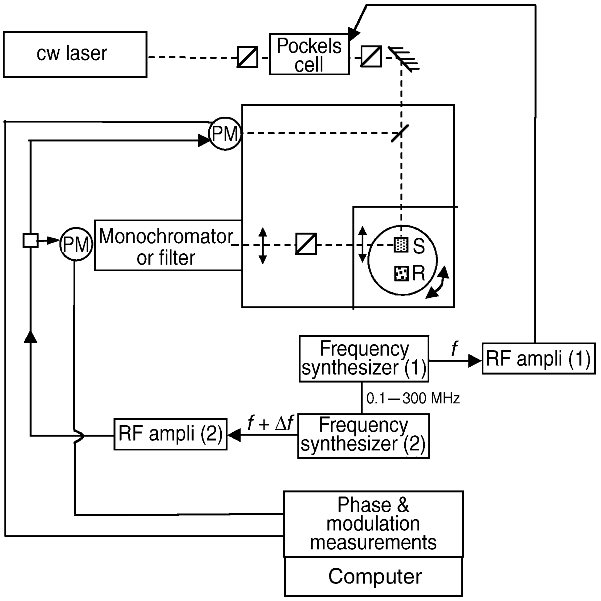

Figure 6 Schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell (S: sample; R: reference; PM: photomultiplier).

Figure 6 shows the schematic diagram of a multifrequency phase-modulation fluorometer using a cw laser and a Pockels cell. A beamsplitter reflects a few percent of the incident light towards a reference photomultiplier (via, or not, a cuvette containing a reference scattering solution). The fluorescent sample and a reference solution (containing either a scatter or a reference fluorescent compound) are placed in a rotating turret. The emitted fluorescence or scattered light is detected by a photomultiplier through a monochromator or an optical filter. The Pockels cell is driven by a frequency synthesizer and the photomultiplier response is modulated by varying the voltage at the second dynode by means of another frequency synthesizer locked in phase with the first one. The two synthesizers provide modulated signals that differ in frequency by a few tens of Hz in order to achieve cross-correlation (heterodyne detection). This procedure offers excellent accuracy because the phase and modulation information contained in the signal is transposed to the low-frequency domain, where phase shifts and modulation depths can be measured with a much better accuracy than in the high-frequency domain.

Table 1 Examples of available wavelengths from low-power cw lasers

Laser	*Wavelength (nm)*
He–Cd	325; 442.5
He–Ne	543.3; 594; 611.9; 632.8
Ar^{+a}	457.9; 476.5; 488.0; 514.5

[a] High-power argon ion lasers can also provide wavelengths of 351.1 and 363.8 nm.

Practically, the phase delay ϕ_R and the modulation ratio m_R of the light emitted by the scattering solution are measured with respect to the signal detected by the reference photomultiplier. Then, after rotation of the turret, the phase delay ϕ_F and the modulation ratio m_F for the sample fluorescence are measured with respect to the signal detected by the reference photomultiplier. The absolute phase shift and modulation ratio of the sample are then $\Phi = \phi_F - \phi_R$ and $M = m_F/m_R$, respectively.

Figure 7 shows an example of an experiment.

Phase Fluorometers Using the Harmonic Content of a Pulsed Laser

The type of laser source that can be used is exactly the same as that for single-photon timing pulse fluorometry (see above). Such a laser system that delivers pulses in the picosecond range, with a repetition rate of a few MHz, can be considered as an intrinsically modulated source. The harmonic content of the pulse train – which depends on the

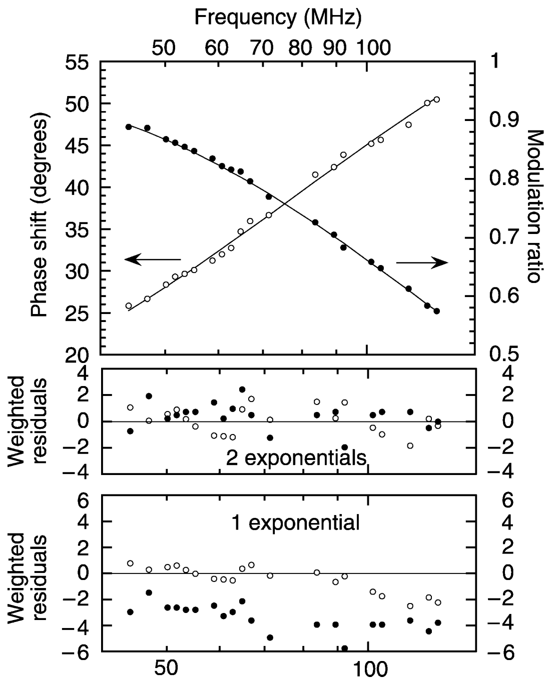

Figure 7 Example of data obtained by the phase and modulation technique using a fluorometer is equipped with a cw Omnichrome He–Cd laser (442 nm) and a Pockels cell operating at frequencies ranging from 0.1 to 200 MHz. The sample is a solution of a calcium complex of a crowned merocyanine (DCM-crown) in acetonitrile (observation through a Balzers bandpass filter centered at 581 nm). The solid line corresponds to the best fit with a biexponential decay. Weighted residuals are shown for the best fit with one and two exponentials. The time constants are 1.87 ± 0.04 and 0.35 ± 0.09 ns with fractional intensities of 0.85 and 0.15, respectively. The value of χ_r^2 is 1.01. (Redrawn from Martin MM, Plaza P, Meyer YH, *et al.* (1996) Steady-state and picosecond spectroscopy of Li^+ or Ca^{2+} complexes with a crowned merocyanine. Reversible photorelease of cations. *Journal of Physical Chemistry* 100: 6879–6888, with permission from the American Chemical Society).

pulse width (as illustrated in Figure 8) – extends to several gigahertz.

For high-frequency measurements, normal photomultipliers are too slow, and microchannel plate photomultipliers are required. However, internal cross-correlation is not possible with the latter and an external mixing circuit must be used.

The time resolution of a phase fluorometer, using the harmonic content of a pulsed laser and a microchannel plate photomultiplier, is comparable to that of a single-photon timing instrument using the same kind of laser and detector.

Data Analysis

In both pulse and phase fluorometries, the most widely used method of data analysis is based on a nonlinear least-squares method. The basic principle of this method is to minimize a quantity which expresses the mismatch between data and fitted function. This quantity is the reduced chi-square χ_r^2, defined as the weighted sum of the squares of the deviations of the experimental response $R(t_i)$ from the calculated responses $R_c(t_i)$:

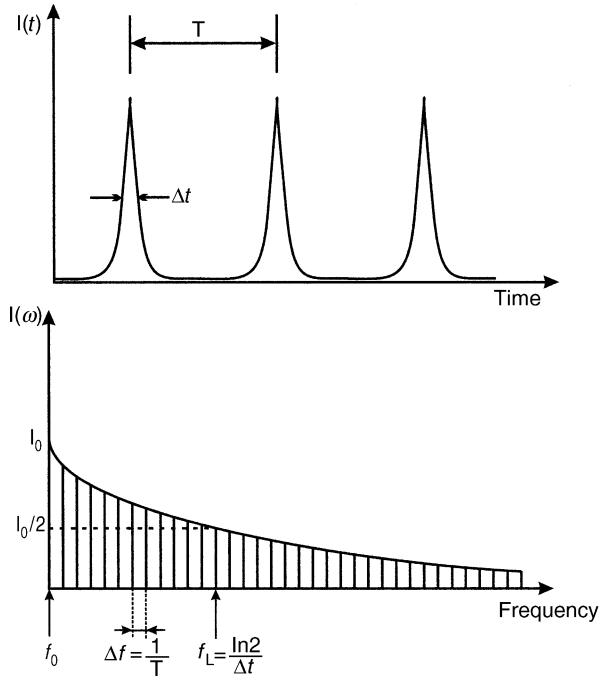

Figure 8 Harmonic content of a pulse train. Example of numerical values: $\Delta f = 4$ MHz; $\Delta t = 4$ Ps; $f_L = 44$ GHz.

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\left[\frac{R(t_i) - R_c(t_i)}{\sigma(i)}\right]^2 \qquad [20]$$

where N is the total number of data points and $\sigma(i)$ is the standard deviation of the ith data point, i.e., the uncertainty expected from statistical considerations (noise). ν is the number of degrees of freedom ($\nu = N - p$, where p is the number of fitted parameters). The value of χ_r^2 should be close to 1 for a good fit.

In the single-photon timing technique, the statistics obey the Poisson distribution and the expected deviation $\sigma(i)$ is approximated to $[R(t_i)]^{1/2}$ so that eqn [20] becomes

$$\chi_r^2 = \frac{1}{\nu}\sum_{i=1}^{N}\frac{[R(t_i) - R_c(t_i)]^2}{R(t_i)} \qquad [21]$$

In phase fluorometry, no deconvolution is required: curve fitting is performed in the frequency domain, directly using the variations of the phase shift Φ and the modulation ratio M as functions of the modulation frequency. Phase data and modulation data can be analyzed separately, or simultaneously. In the latter case the reduced chi-squared is given by

$$\chi_r^2 = \frac{1}{\nu}\left[\sum_{i=1}^{N}\left[\frac{\Phi(\omega_i) - \Phi_c(\omega_i)}{\sigma_\Phi(\omega_i)}\right]^2 + \sum_{i=1}^{N}\left[\frac{M(\omega_i) - M_c(\omega_i)}{\sigma_M(\omega_i)}\right]^2\right] \qquad [22]$$

where N is the total number of frequencies. In this case, the number of data points is twice the number of frequencies, so that the number of degrees of freedom is $\nu = 2N - p$.

In addition to the value of χ_r^2, it is useful to display graphical tests. The most important of them is the plot of the weighted residuals defined as

$$W(t_i) = \frac{R(t_i) - R_c(t_i)}{\sigma(i)} \qquad [23]$$

where $\sigma(i) = [R(t_i)]^{1/2}$ for single-photon counting data. The fit is satisfactory when the weighted residuals are randomly distributed around zero.

Concluding Remarks

From the instrumental point of view, the latest generations of pulse and phase fluorometers both use pulsed lasers (usually Ti:sapphire lasers) and microchannel plate detectors; only the electronics is different. Since the instrument response is mainly limited by the time response of the detector, the time resolution is the same for both techniques. Moreover, the optical module is identical so that the total cost of instrumentation is similar.

There are, however, significant differences between pulse and phase fluorometers from the methodological point of view. In pulse fluorometry, the single-photon timing technique offers various advantages: direct visualization of the fluorescence decay; high sensitivity; outstanding dynamic range and linearity; and well-defined statistics (Poisson distribution) allowing the accurate weighing of each point in data analysis. In phase-modulation fluorometry, the main advantage is that no deconvolution is necessary, while this operation is often necessary in pulse fluorometry and requires great care in recording the instrument response, especially for very short decay times. Moreover, the short acquisition time for phase shift and modulation ratio measurements at a given frequency is a distinct advantage.

See also

Ultrafast Laser Techniques: Pulse Characterization Techniques.

Further Reading

Alcala JR, Gratton E and Jameson DM (1985) A multifrequency phase fluorometer using the harmonic content of a mode-locked laser. *Analytical Instrumentation* 14: 225–250.

Demas JN (1983) *Excited-State Lifetime Measurements*. New York: Academic Press.

Jameson DM, Gratton E and Hall RD (1984) The measurement and analysis of heterogeneous emissions by multifrequency phase and modulation fluorometry. *Applied Spectroscopy Reviews* 20: 55–106.

Lakowicz JR (ed.) (1991) *Topics in Fluorescence Spectroscopy, Vol. 1: Techniques*. New York: Plenum Press.

Lakowicz JR, Laczko G and Gryczynski I (1986) 2 GHz frequency-domain fluorometer. *Review of Scientific Instruments* 57: 2499–2506.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Valeur B (2002) *Molecular Fluorescence. Principles and Applications*, chap. 6. Weinheim, Germany: Wiley-VCH.

Measurements in Polymer Science

T A Smith, University of Melbourne, Melbourne, Australia

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

In this article we explore the use of ultrashort pulsed lasers in time-resolved fluorescence measurements for the study of the dynamics of excimer formation and energy transfer/migration in synthetic aromatic polymers. The temporal profile of monomer and excimer emission, and the use of time-resolved fluorescence polarization approaches are discussed.

The use of lasers in chemistry has most often exploited the high degree of spectral purity achievable with these devices, in both photochemical and high-resolution spectroscopic applications. Through the use of lasers, we also have the ability to generate pulses of light of ultrashort duration (a few femtoseconds) and of incredibly high peak powers (petawatts), characteristics significantly beyond those achievable with conventional (incoherent) light sources. This enables the kinetics of the many photophysical and photochemical processes that occur on ultrafast time-scales with high efficiency within polymers, to be determined through various time-resolved measurements.

These processes are dealt with in detail in specialist photochemistry textbooks and so knowledge of the terms associated with photochemistry and photophysics is assumed (e.g., electronic configurations, orbitals, spin, multiplicity, etc.).

Following the absorption of light, an isolated chromophore will be in one of its excited electronic and vibrational states. Once formed, the excited state of a molecule will only survive for a limited time, with the molecule losing its excess energy to return to a preferred lower energy state (electronic configuration). It can do this through a number of spontaneous, radiative, and nonradiative deactivation processes, each with its own characteristic rate constant k_{ij}. Nonradiative pathways include internal conversion (IC), such as intramolecular vibrational redistribution, and vibrational coupling to the bath, and inter-system crossing (ISC) between singlet and triplet manifolds. Radiative pathways involve the emission of a photon of light, in the form of fluorescence or phosphorescence.

The time the molecule spends in the excited state is determined by the rates of all the pathways that contribute to the overall deactivation of that state. The lifetime, τ, of a singlet state is therefore defined as:

$$\tau = \frac{1}{k_{\mathrm{ISC}} + k_{\mathrm{IC}} + k_{\mathrm{Rad}} + \cdots} = \frac{1}{\sum k_{ij}} \qquad [1]$$

In systems of higher chromophore concentration, such as aromatic-containing polymers, additional, bimolecular pathways are available through which the excess excited state energy may be dissipated, making it rare that a single fluorescence 'lifetime' is observed from polymer systems. The observation of complex (multi- or nonexponential) fluorescence decay profiles sometimes invokes the use of the terms 'average lifetimes' or 'lifetime distributions'.

The complex nonexponential fluorescence decay behavior observed, particularly from polymer systems, has been variously attributed to one or more of the following factors: molecular aggregation (e.g., excimer/exciplex equilibria); acid–base equilibria (e.g., proton transfer); one-, two-, or three-dimensional energy transfer and energy migration (and variations in the singlet energy migration length); electron transfer; molecular motion (translational diffusion, rotational relaxation); vibrational or solvent relaxation; concentration quenching; fluorescent impurities, etc. The kinetics of excimer formation are further complicated by heterogeneity of kinetically distinct monomer and/or excimer species, configurational and conformational influences upon excimer formation, and the time dependence of energy trapping (excimer formation) through nonequilibrium diffusion-controlled processes. Here, we will concentrate on excimer formation and energy migration in synthetic aromatic polymers, as shown schematically in Figure 1.

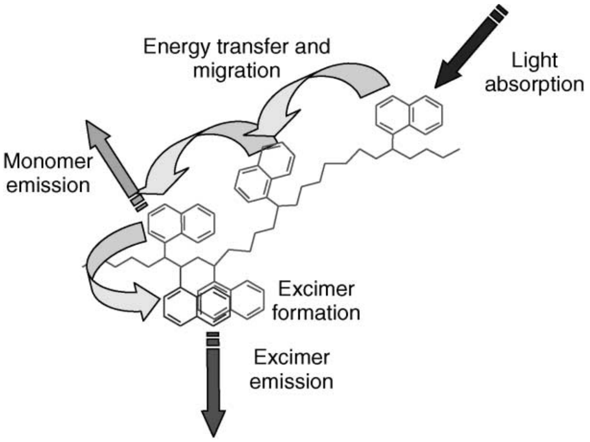

Figure 1 Schematic representation of some of the photophysical processes in polymers.

Ultrafast Laser Spectroscopic Techniques

A wide range of experimental techniques and instrumentation is available for the study of time-resolved processes in polymers. The most widely used methods rely on the detection of emission from the polymer. Fluorescence-based techniques are popular for investigating time-dependent processes in polymers, since the time-scales involved in many of the photochemical and photophysical processes discussed correspond with the time-scales on which fluorescence occurs (normally on the subpicosecond to nanosecond time-scales). Phosphorescence usually occurs on significantly longer time-scales and therefore is less likely to be competitive with fluorescence or other radiative or nonradiative deactivation pathways under normal conditions. Fluorescence-based techniques are attractive for reporting on a number of time-dependent processes since fluorescence is inherently multidimensional, it can be readily resolved temporally, and fluorescence detection can be extremely sensitive, achieving single photon and single molecule detection levels.

Amongst the most commonly used fluorescence-based techniques are: time-correlated single photon counting (TCSPC, or single photon timing), fluorescence transient digitization, and ultrafast streak cameras. Methods such as transient absorption (flash photolysis) can also provide some details regarding the identity of short-lived species formed following photo-excitation. Other ultrafast laser techniques have been applied to the study of time-dependent processes in polymers in order to achieve higher temporal resolution. So-called pump–probe methods provide the ultimate in time-resolution and can be combined with fluorescence detection through the use of Kerr shutters or ultrafast fluorescence upconversion techniques. Other, more complex pump–probe methods, such as transient grating and multiple pulse photon echo measurements, are potentially very powerful tools in the study of time-resolved processes in polymers.

The generation of short pulses of light can be achieved through numerous mechanisms, including active or passive mode-locking and Q-switching techniques, and superradiance processes (e.g., in nitrogen lasers). Lasers capable of producing short (~ns) pulses of light by Q-switching are usually characterized by high energies per pulse (from μJ up to many hundreds of mJ/pulse) at low (~Hz–kHz) pulse repetition rates. Lasers producing pulses of ultrashort duration (often in the picosecond or femtosecond region) through mode-locking techniques, characteristically operate at high pulse repetition rates (~MHz) and with lower energies per pulse (~nJ/pulse). It should be noted that even pulses of a few nJ energy delivered in ~100 fs, still have peak powers of tens of kW. These high peak powers enable the generation of a range of excitation wavelengths suitable for the study of polymer systems through nonlinear optical wavelength conversion processes, such as second-harmonic generation (SHG) and optical parametric oscillation (OPO) devices.

Excimer Formation

Excimers (or *exci*ted-state di*mers*) are formed when an aromatic molecule, while in its excited electronic state, $^1M^*$, interacts with a second molecule of the same chemical type in its ground electronic state, ^1M. Excimer formation is often observed in concentrated molecular systems such as concentrated solutions and molecular crystals; however, it can also occur in dilute solutions of polymers containing aromatic chromophores, or in the confined environment of solid polymer systems such as homopolymer films.

The structure of excimers in solution is generally thought to be a face-to-face, coplanar, sandwiched, or in some cases, partially eclipsed arrangement of the two constituent aromatic moieties, which attempt to adopt the lowest possible energy configuration, with a separation of 3–4 Å, as depicted in Figure 2. A range of possible excimer configurations may exist rather than just the two extremes.

A feature of excimers is that they have a dissociative ground state (Figure 3), i.e., they do not exist prior to the absorption of light by one monomer chromophore and thus cannot be formed directly without the coming together of the two constituent chromophores within the lifetime of the $^1M^*$. Excimers, being of lower energy than the corresponding excited monomer species and with this dissociative ground state, usually exhibit characteristically

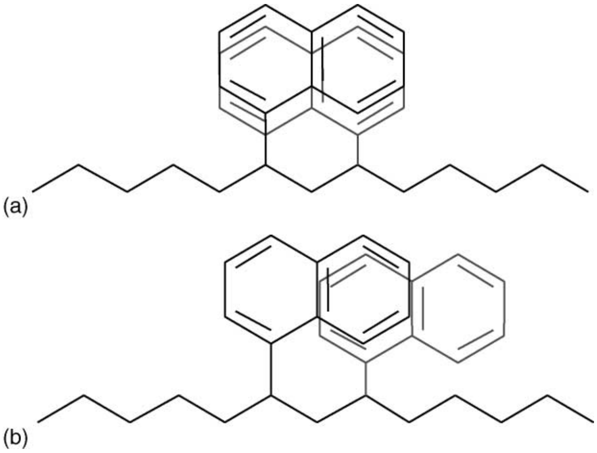

Figure 2 (a) Sandwiched and (b) partially eclipsed excimer configurations.

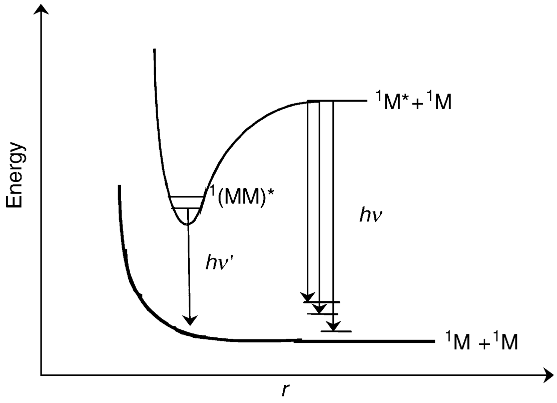

Figure 3 Schematic potential energy diagram for excimer formation and decay. *r* represents the monomer–monomer separation.

broad, structureless emission to the red of the monomer emission (Figure 4). Upon the emission of fluorescence from an excimer, the complex loses its excess energy and dissociates back to two ground state monomer units (i.e., 2 ^{1}M). Excimers may also dissociate back to a noninteracting pair of ^{1}M* and ^{1}M.

Excimer formation of simple aromatic molecules in solution is often modeled on the kinetic scheme proposed by Birks and co-workers (Figure 5). From Figure 5, if a δ-pulse excitation, represented by $h\nu_{\mathrm{abs}}$, is used to produce an initial concentration of excited singlet monomer molecules, $[^1\mathrm{M}^*]_\mathrm{o}$, then such a scheme results in the following rate equations:

$$\frac{\mathrm{d}[^1\mathrm{M}^*]}{\mathrm{d}t} = k_{\mathrm{MD}}[^1(\mathrm{MM})^*] - X[^1\mathrm{M}^*] \qquad [2]$$

$$\frac{\mathrm{d}[^1(\mathrm{MM})^*]}{\mathrm{d}t} = k_{\mathrm{DM}}[^1\mathrm{M}][^1\mathrm{M}^*] - Y[^1(\mathrm{MM})^*] \qquad [3]$$

where $X = k_{\mathrm{FM}} + k_{\mathrm{NM}} + k_{\mathrm{DM}}[^1\mathrm{M}^*] = k_\mathrm{M} + k_{\mathrm{DM}}[^1\mathrm{M}]$, $Y = k_{\mathrm{FD}} + k_{\mathrm{ND}} + k_{\mathrm{MD}} = k_\mathrm{D} + k_{\mathrm{MD}}$.

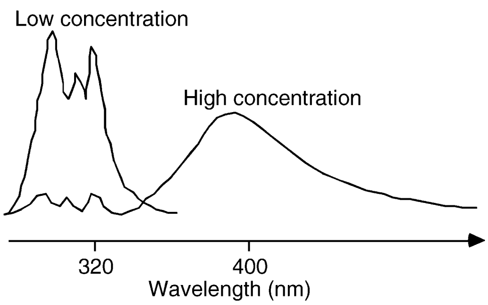

Figure 4 Schematic representation of monomer and excimer emission.

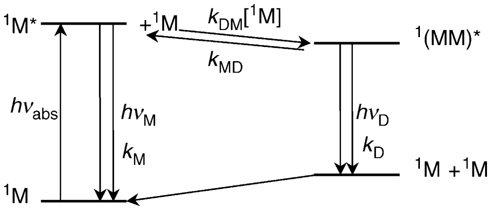

Figure 5 Kinetic scheme for excimer formation where $h\nu_\mathrm{M}$ and $h\nu_\mathrm{D}$ represent the emission of photons by the monomer and excimer species respectively, k_M and k_D comprise the sum of the various rate constants for all deactivation pathways for the monomer and excimer respectively (excluding k_{MD} and k_{DM}) (e.g., $k_\mathrm{M} = k_{\mathrm{FM}} + k_{\mathrm{NM}}$ where k_{FM} and k_{NM} are the radiative (fluorescence) and nonradiative deactivation rate constants respectively of the monomer).

Solving these rate equations and applying the initial conditions $[^1\mathrm{M}^*]_t = [^1\mathrm{M}^*]_\mathrm{o}$ and $[^1(\mathrm{MM})^*]_\mathrm{o} = 0$ at $t = 0$, yields expressions given below for the decay of monomer and excimer fluorescence intensity with time. The emission intensity (monitored in a spectral region in which only the monomer $[^1\mathrm{M}^*]$ species emits), will have the temporal form given by eqn [4]:

$$\begin{aligned} i_\mathrm{M}(t) &= k_{\mathrm{FM}}\frac{[^1\mathrm{M}^*]}{[^1\mathrm{M}^*]_\mathrm{o}} \\ &= \left[k_{\mathrm{FM}}\frac{(\lambda_2 - X)}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) + A\exp(-\lambda_2 t)\right] \end{aligned} \qquad [4]$$

where

$$A = \frac{(X - \lambda_1)}{(\lambda_2 - X)},$$

$$\lambda_{1,2} = \frac{1}{2}\left[X + Y \mp \left\{(Y - X)^2 + 4k_{\mathrm{MD}}k_{\mathrm{DM}}[^1\mathrm{M}^*]\right\}^{1/2}\right]$$

The emission intensity collected from a spectral region in which the excimer ($^1[\mathrm{MM}]^*$) species only emits, will have the temporal form given by eqn [5]:

$$\begin{aligned} i_\mathrm{D}(t) &= k_{\mathrm{FD}}k_{\mathrm{DM}}\frac{[^1\mathrm{D}^*]}{[^1\mathrm{M}^*]_\mathrm{o}} \\ &= \left[\frac{k_{\mathrm{FD}}k_{\mathrm{DM}}[^1\mathrm{M}]}{(\lambda_2 - \lambda_1)}\right]\left[\exp(-\lambda_1 t) - \exp(-\lambda_2 t)\right] \end{aligned} \qquad [5]$$

If excimer formation follows Birks' scheme, then the decay profile of the monomer fluorescence should therefore be described by the sum of two exponential decay terms, and the excimer emission by an exponential growth term and a decay term (Figure 6). Only two decay constants, λ_1 and λ_2, should be required to fit both the monomer and excimer fluorescence profiles successfully. This is often not observed, which has led to the introduction of more complex interpretations.

In synthetic aromatic polymers, the effective local concentration of the aromatic molecules can be high while the concentration of polymer chains in solution may be very low. This means that excimer formation can be significant in such polymers, even in dilute solution. In fact, in some polymers the emission is dominated by excimers over monomers. The fact that the excimers are of lower energy than the corresponding excited monomer species also means that they can act as effective traps of excitation energy that might be migrating/relocating within a polymer chain as discussed below.

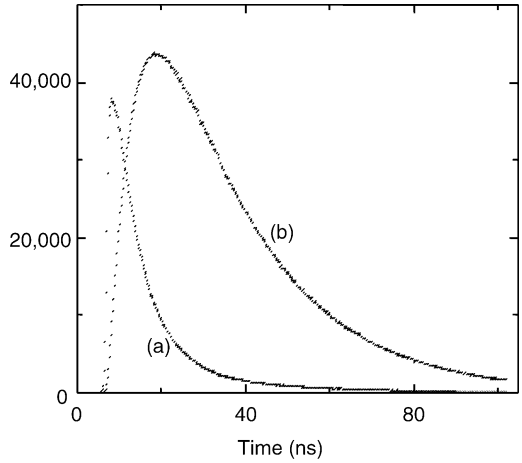

Figure 6 Examples of (a) monomer and (b) excimer fluorescence decays.

Excimer formation in solutions of aromatic polymers is generally thought to proceed via a complicated route that may involve a combination of electronic energy transfer and migration, segmental rotation, and chromophore diffusion. It is believed that energy migrates along a polymer chain until sites suitable for excimer formation are encountered, at which point the energy is trapped. A number of mechanisms for the formation of suitable excimer sites in polymers are then conceivable. These include:

(i) interactions between chromophores belonging to different chains (this *inter*molecular excimer formation is highly concentration dependent);
(ii) interactions between two chromophores far removed from each other, but on the same chain. This necessitates mainchain conformational changes and side group motions during the excited state lifetime, so as to bring both molecules together;
(iii) near neighbor interactions (c.f. Hirayama's '$n = 3$' rule). The configuration required for excimer formation need not exist prior to excitation provided that the two constituent groups can attain the required geometry within the lifetime of the excited chromophore; or
(iv) preformed sites, suitable for excimer formation, present in the polymer before the initial absorption of light occurs. These may be formed by interactions between nearest neighbor chromophores or between two chromophores on loops of the same, or different chains in cases where chain motion is restricted.

The first two of these mechanisms depend upon segmental motion of the polymer chain and the third mechanism may depend upon rotation of the chromophore/s to adopt the close proximity and optimum orientational configuration required for excimer formation. In solution, this may be, at least partially, diffusion-controlled, causing a dominant complicating factor in the kinetics of excimer formation in polymeric systems. One approach is therefore to substitute a time-dependent expression for the rate coefficient of excimer formation, $k_{\mathrm{DM}}(t) = A + B/\sqrt{t}$, (based on the Einstein–Smoluchowski diffusion theory), for k_{DM}, into the rate equations. Such an approach produces the following expression for the time dependence of the monomer fluorescence, $I_{\mathrm{M}}(t)$ in the presence of diffusion-controlled excimer formation and the absence of excimer back dissociation:

$$I_{\mathrm{M}}(t) = A\exp\left\{-at - b\sqrt{t}\right\} \qquad [6]$$

In some cases, such as when excimer back dissociation cannot be neglected, more complex functional forms have been derived, such as:

$$I_{\mathrm{M}}(t) = A\exp\left\{-at - b\sqrt{t}\right\} + B\exp\left\{-t/\tau\right\} \qquad [7]$$

Similar expressions can be derived for the time dependence of the excimer fluorescence, $I_{\mathrm{E}}(t)$. More detailed expressions for the term $k(t)_{\mathrm{DM}}$, involving the complete form of the diffusion rate coefficient, have also been used in such approaches.

A related approach is the concept of 'convolution kinetics', in which the intimate dependence of excimer formation upon the monomer decay is emphasized. Once formed, the excimer is assumed to decay exponentially. The excimer fluorescence temporal profile is predicted to take the form:

$$I_{\mathrm{E}}(t) = k_1(t)I_{\mathrm{M}}(t)\otimes\exp\left(\frac{-t}{\tau}\right) \qquad [8]$$

where $\otimes$ represents the convolution integral and $I_{\mathrm{E}}(t)$ and $I_{\mathrm{M}}(t)$ are the measured decay profiles in the excimer and monomer emission regions respectively.

Electronic Energy Transfer, Energy Migration and Trapping

In polymers the high local concentration of chromophores makes additional deactivation pathways competitive with the other photophysical processes.The excess energy may transfer from the excited fluorophore of interest to a lower energy chromophore. Three of the more common possible mechanisms, through which energy transfer can occur, are:

(i) *Radiative* ('*trivial*') – involves the reabsorption of donor emission by an acceptor chromophore.

(ii) *Resonance* (*Dipole–Dipole*, *Förster*) – long-range (15–100 Å), nonradiative, single-step energy transfer process arising from Coulombic (dipole–dipole) interactions between an excited-state donor chromophore, D*, and an initially unexcited acceptor, *A*. Förster derived a statistical expression for the (first-order) rate constant, for such a process, considering three-dimensional, dipole–dipole interactions between stationary molecules:

$$k(R)_{\mathrm{D}^*\rightarrow\mathrm{A}} = \frac{9000\kappa^2\phi_{\mathrm{o}}\ln 10}{128\pi^5\eta^4 N\tau_{\mathrm{D}}R^6}\int_0^\infty \frac{f_{\mathrm{D}}(\bar{\nu})\varepsilon_{\mathrm{A}}(\bar{\nu})}{\bar{\nu}}\mathrm{d}\bar{\nu} = \frac{1}{\tau_{\mathrm{D}}}\left(\frac{R_{\mathrm{o}}}{r_{\mathrm{D}}}\right)^6 \qquad [9]$$

where R is the donor–acceptor distance, ϕ_{o} is the quantum yield for emission from the donor in the absence of acceptor molecules, η is the refractive index of the solvent, τ_{D} is the lifetime of the donor excited state (reduced by the onset of electronic energy transfer), N is Avogadro's number, κ is a constant dependent on the mutual orientation of the donor and acceptor molecules, and the integral term is a measure of the spectral overlap between the donor emission profile, $f_{\mathrm{D}}(\bar{\nu})$ (in quanta cm^{-1} and normalized to unit area) and the acceptor absorption profile, $\varepsilon_{\mathrm{A}}(\bar{\nu})$ (converted to molar extinction coefficients and measured on a wavenumber scale). R_{o} is the critical transfer distance, at which the probability of resonance transfer, $k(R)_{\mathrm{D}^*\rightarrow\mathrm{A}}$, is equal to the probability of spontaneous decay, $1/\tau_{\mathrm{D}}$.

(iii) *Exchange mechanism* – also a nonradiative, single-step process but involves the exchange of electronic excitation when the donor and acceptor are close enough so that their electron clouds overlap. Consequently it only operates over small distances. Dexter has derived a rate expression for exchange energy transfer, again assuming that the donor and acceptor molecules are stationary:

$$k_{\mathrm{D}^*\rightarrow\mathrm{A}} = \left(\frac{2\pi}{\hbar}\right)Z^2\int f_{\mathrm{D}^*}(\bar{\nu})\varepsilon_{\mathrm{A}}(\bar{\nu})\mathrm{d}\bar{\nu} \qquad [10]$$

where $Z^2 = K^2\exp(-2R/L)$ and K and L are simply constants.

It has been clearly demonstrated that energy transfer, from pendant aromatic monomer units to chemically bound acceptor chromophores, can occur with high efficiency and is influenced significantly by polymer structure and conformation.

In addition to energy transfer between D* and A (being different chemical species), more than one exchange of excitation energy between chemically and spectroscopically identical chromophores, attached by covalent bonds to a polymer chain, can occur. A single step between identical units is still termed energy transfer, while more than one such step in a sequence would constitute intramolecular energy migration.

Energy migration is a long-range process, thought to occur via a series of energy transfer steps occurring between chromophores, either: (i) along the polymer chain (often the adjacent groups); (ii) across loops in a single polymer chain; or (iii) across different polymer chains. The migration continues until either a trap site is encountered or one of the chromophores excited in the migration process deactivates spontaneously. Each step is believed to occur with a probability defined by the exchange or Förster mechanisms, but since the energy transfer occurs between donor and acceptor groups, which are of the same type, the spectral overlap integral involved in both mechanisms will usually be small. The distances over which such energy transfer can occur therefore cannot be large and typically takes values of <10–20 Å. It is worth noting that many authors consider the energy to migrate as an incoherent exciton (where the term exciton is used in a much looser sense than is customary among solid-state physicists).

Long-range energy migration can only occur by energy transfer from the initially excited chromophore to ground state chromophores of the same, or lower, energy. If energy transfer is favorable between one chromophore and another of the same energy, then back transfer to the initially excited chromophore is also highly probable. When there is a change in the emission and absorption dipole orientations, caused by rotation of the chromophores relative to one another on the time-scale of the energy transfer step, or the transfer is to a chromophore of significantly lower energy, then the probability of energy transfer back to the original chromophore might be reduced. The lower energy chromophores on the polymer chain may therefore act as 'traps' of the excitation originally absorbed by a higher energy chromophore, even in the absence of excimer energy trap sites.

Various approaches have been proposed for modeling the kinetics of energy migration. One can consider the time-dependent nature of each energy transfer step of the migration process, and introduce a time-dependent rate coefficient for each step of the energy migration process, $k_{\mathrm{DD}}(t)$. The form for $k_{\mathrm{DD}}(t)$ by Sienicki *et al.*, is given by eqn [11], which leads to nonexponential functions predicted for the ensemble averaged probability that an originally excited donor molecule is still excited at time t:

$$k_{\mathrm{DD}}(t) = \left(\frac{\Delta}{n}\right)\gamma_{\Delta} c_{\mathrm{D}} \Gamma(1-\Delta/n)\left(\frac{t}{\tau_{\mathrm{OD}}}\right)^{\Delta/n-1} \qquad [11]$$

where $\Delta(=1, 2, 3)$ is the dimensionality of the medium, $n(=6, 8, 10, \ldots)$ is the order of the interaction, τ_{OD} is the lifetime of the donor fluorescence, $\Gamma(x)$ is the gamma function, γ_{Δ} is a constant dependent upon Δ and c_{D} is the donor concentration.

Energy migration has also been considered as a diffusion-like process, and the Smoluchowski diffusion theory adopted. The rate coefficient for energy migration, $k(t)_{\mathrm{EM}}$, is then expressed in a form similar to the time-dependent rate coefficient for partially diffusion-controlled reactions (eqn [12]). In this type of treatment, the rate of change of the concentration of excited monomer chromophores, $^1\mathrm{M}^*$, is assumed to be given by eqn [13]:

$$k(t)_{\mathrm{EM}} = 4\pi N'_{\mathrm{A}} R D \rho\left[1 + R\rho(\pi D t)^{-1/2}\right] \qquad [12]$$

$$\frac{\mathrm{d}[^1\mathrm{M}^*]}{\mathrm{d}t} = I_{\mathrm{o}} - \left(\frac{1}{\tau_{\mathrm{o}}} + k(t)_{\mathrm{EM}}\right)[^1\mathrm{M}^*] \qquad [13]$$

In these expressions, τ_{o} is the reciprocal of the sum of the rate constants for radiative and nonradiative deactivation, ρ is the reaction probability per collision, R is the interaction distance, and N'_{A} is the number density of the energy acceptors or traps ($N'_{\mathrm{A}} = NC_{\mathrm{A}}$, with the concentration of acceptors, C_{A}, in mol cm^{-3}). D is the mutual diffusion coefficient which can be replaced by a migration coefficient, Λ, R can be replaced by the Förster critical transfer distance, R_{o}, and the reaction probability, ρ, can be set to 0.5 (since R_{o} is the distance at which the probability of energy transfer is 0.5). This leads to a simplified expression for $k(t)_{\mathrm{EM}}$, which can be substituted into eqn [13], to produce an expression (eqn [14]) for the monomer decay:

$$I(t) = A\exp\left(-\left(\frac{1}{\tau_{\mathrm{o}}} + 2\pi N'_{\mathrm{A}} R_{\mathrm{o}} \Lambda[^1\mathrm{M}^*]\right)t - 2N'_{\mathrm{A}} R_{\mathrm{o}}^2 (\pi\Lambda)^{1/2}[^1\mathrm{M}^*]\sqrt{t}\right) \qquad [14]$$

Yokota and Tanimoto have also developed a statistical representation of Förster's theory and derived an expression for dipole–dipole transfer in a fluid medium in which the distribution of excited-state donor and acceptor molecules is determined, both by diffusion and by the decay and transfer of the donor excited state. According to the Yokota and Tanimoto approach, the decay law is approximated by the expression:

$$I(t) = A\exp\left\{\frac{-t}{\tau_{\mathrm{o}}} - 2B\delta\left(\frac{-t}{\tau_{\mathrm{o}}}\right)^{1/2}\right\} \qquad [15]$$

where $\delta = [\mathrm{C}]/[\mathrm{C}]_{\mathrm{o}}$ (C_{o} is the critical molar concentration of acceptor molecules given by: $[\mathrm{C}_{\mathrm{o}}] = 3000/2\pi^{3/2} N R_{\mathrm{o}}^3$, and B is a complex, time-dependent function given by: $B = (1 + 10.87x + 15.5x^2/1 + 8.743x)^{3/4}$ with $x = D\alpha^{-1/3} t^{2/3}$, $\alpha = R_{\mathrm{o}}^6/\tau_{\mathrm{o}}$ and D is the 'diffusion coefficient' representing energy migration.

Emission depolarization is often exploited in the study of energy migration in polymers. Only molecules with some vector component of their absorption transition dipole lying parallel to the plane of polarization of the incident radiation, will absorb incoming radiation. In the absence of angular motion of the absorbing chromophore within the lifetime of the excited state, emission from the same chromophore will also be highly polarized; however, upon each step in the energy migration process, some loss of polarization will be incurred (Figure 7). The extent of depolarization of emission is usually referred to by the term fluorescence anisotropy, $r(t)$ (eqn [16]):

$$r(t) = \frac{I_{\parallel}(t) - I_{\perp}(t)}{I_{\parallel}(t) + 2I_{\perp}(t)} \qquad [16]$$

The rate of energy migration can be estimated directly through observation of the time-dependent decay of $r(t)$ in the absence of other fluorescence depolarizing processes, such as molecule rotation (Figure 8).

One of the important properties which characterizes excitation energy transport in a system of n identical molecules, which are distributed randomly in space, is the Green function, $G(r,t)$. The Green function can be thought of as the time-dependent probability of finding an excitation at position r on the polymer chain at time t. It has been shown that the overwhelming contribution to fluorescence polarization is due to fluorescence from the sites that were initially excited. Thus, $G^{\mathrm{s}}(r,t)$ is directly related to the fluorescence anisotropy if the polarization memory of secondary excited molecules can be neglected. The $r(t)$ for donor–donor energy transfer is then given by:

$$r(t) = \alpha r_{\mathrm{rot}}(t) G^{\mathrm{s}}(t) \qquad [17]$$

in which r_{rot} describes the anisotropy in the absence of energy transfer, i.e., in the presence of chromophore

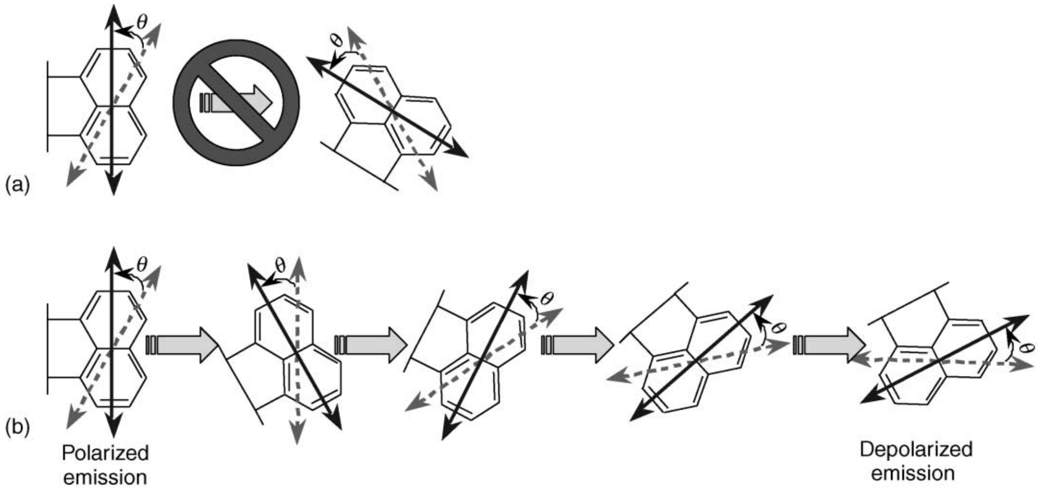

Figure 7 Schematic representation of energy migration. θ is the inherent angle between the donor emission (dotted arrow) and acceptor absorption (solid arrow) transition dipoles of the chromophore (a) orthogonal donor emission and acceptor absorption dipoles – no energy transfer/migration, emission depolarized only due to θ. (b) some vector component of the absorption dipole of the acceptor is parallel with that of the donor emission dipole – this and subsequent energy transfer steps can occur resulting in loss of emission polarization.

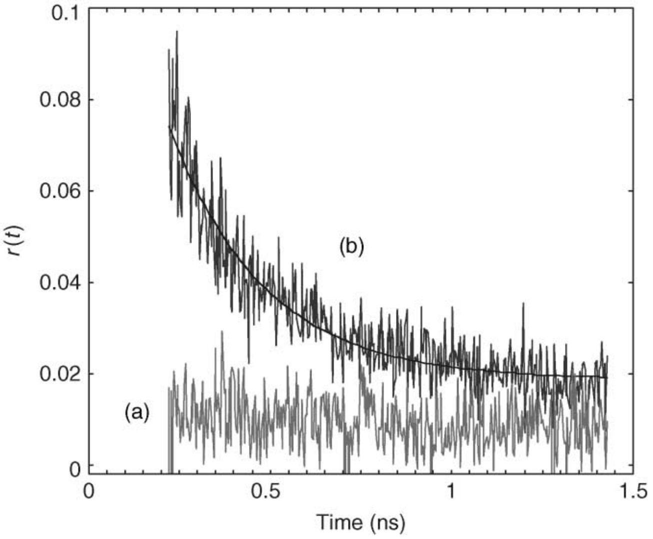

Figure 8 Temporal decays of $r(t)$ from (a) a polymer in which emission depolarization due to energy migration is complete on ultrafast time-scales, and (b) a polymer in which the rate of energy migration has been inhibited. Emission collected from polymers in frozen glass to exclude polymer motion contributions to emission depolarization.

rotation only, and α is the anisotropy from time-independent sources of depolarization.

The determination of the functional form of $G^s(r, t)$, resulting from excitation energy migration in polymer systems, is a challenging theoretical problem to which a number of approaches have been applied. The Gochanour, Anderson and Fayer approach has had some success in arriving at an expression for $G^s(r, t)$, represented as the time decay of emission polarization curves.

Polymeric systems that undergo energy migration, yet contain no excimer forming or other nonmonomeric, energy-trapping sites, are rare. In the presence of singlet energy traps, the problem can be treated as being analogous to energy transfer (via the Förster mechanism), following a diffusive process in a similar manner to that described above.

Fredrickson and Frank have derived expressions for the temporal decays of the monomer and excimer fluorescence when one-dimensional electronic singlet energy migration is the major excimer-forming path, and segmental rotation and chromophore diffusion are considered. The Fredrickson and Frank model assumes inherently that excimers are formed predominantly via a nearest-neighbor mechanism, where rapid energy migration is at least an order of magnitude faster than any competing rotationally sampled pathways. The expressions for the fluorescence decay of both the monomer, $I_M(t)$, and excimer species, $I_E(t)$, are highly nonexponential, with the expression for $I_M(t)$ being given by eqn [18]:

$$I_M(t) = Q_{fm}k_m(1-q)^2 \times \exp[(4q^2W - k_m - k_{rot})t]\operatorname{erfc}[2q(Wt)^{1/2}] \quad [18]$$

In eqn [18], Q_{fm} is the quantum efficiency for monomer fluorescence and k_m is a rate constant, which encompasses the total rate of monomer decay by radiative and nonradiative pathways in the absence of energy migration and rotational sampling. Similarly, k_{rot} represents the rate at which an excited monomer is transformed into an excimer by segmental rotation. The parameter W represents the rate constant for nearest-neighbor electronic energy transport between the donor (monomer) chromophores, averaged over both the transition moment orientation and the interchromophore separation associated with each dyad.

Energy transfer from pendant aromatic monomer units to a specific chemically bound acceptor chromophore can occur in some polymers following energy migration. This phenomenon is known as the 'antenna effect', which has commercial applications such as in the sensitization of photo-oxidation–reduction, i.e. reactions, in the stabilization of polymers against photodegradation or in the modeling of photosynthesis by light harvesting. The kinetic behavior of such processes is yet to be modeled definitively, but the approaches discussed above (and others) have been used with varying degrees of success to date.

List of Units and Nomenclature

Ångstrom (Å)	10^{-10} m
chromophore	An aromatic molecule that absorbs (and may emit) light
exciton	An excited state localized on a particular chromophore at any given time
femtoseconds	10^{-15} s
intermolecular	Occurring between molecules or polymer chains
intramolecular	Occurring within a single molecule or polymer chain
petawatts	10^{15} W
δ-pulse excitation	Excitation by an infinitely short laser pulse

See also

Chemical applications of lasers: Non-linear spectroscopies. **Lasers:** Excimer Laser. **Optical Parametric Devices:** Optical Parametric Oscillators (Pulsed). **Spectroscopy:** Second Harmonic Spectroscopy.

Further Reading

Andrews DL and Demidov AA (1999) *Resonance Energy Transfer.* New York: Wiley.

Birks JB (1970) *Photophysics of Aromatic Molecules.* London: Wiley-Interscience.

Fleming GR (1986). *Chemical Applications of Ultrafast Spectroscopy*, vol. 13. New York: Oxford University Press.

Fredrickson GH and Frank CW (1983) Nonexponential transient behavior in fluorescent polymeric systems. *Macromolecules* 16: 572–577.

Ghiggino KP and Smith TA (1993) Dynamics of energy migration and trapping in photoirradiated polymers. *Progress in Reaction Kinetics* 18: 375–436.

Gochanour CR, Andersen HC and Fayer MD (1979) Electronic excited state transport in solution. *Journal of Chemical Physics* 70: 4254–4271.

Guillet JE (1985) *Polymer Photophysics and Photochemistry: An Introduction to the Study of Photoprocesses in Macromolecules*. New York: Cambridge University Press.

Lakowicz JR (1999) *Principles of Fluorescence Spectroscopy*, 2nd edn. New York: Plenum.

O'Connor DV and Phillips D (1984) *Time-Correlated Single Photon Counting*. London: Academic Press.

Sienicki K, Itagaki H and Mattice WL (1989) On the theory of concentration depolarization of fluorescence in one- and two-component systems for multipole interactions in one, two, and three dimensional medium. *Journal of Chemical Physics* 91: 4515–4521.

Van der Meer BW, Coker G and Chen S-YS (1994) *Resonance Energy Transfer: Theory and Data*. New York: VCH.

Van Hecke GR and Karukstis KK (1998) *A Guide to Lasers in Chemistry.* Boston, MA: Jones and Bartlett.

Yokota M and Tanimoto O (1967) Effects of diffusion on energy transfer by resonance. *Journal of the Physics Society of Japan* 22: 779–784.

TOMOGRAPHY

Contents

Optical Coherence Tomography
Tomography and Optical Imaging

Optical Coherence Tomography

S A Boppart, University of Illinois, Urbana, IL, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical coherence tomography (OCT) is an emerging imaging technique for a wide range of biological, medical, and material investigations. OCT was initially developed for imaging biological tissue because it permits the imaging of tissue microstructure *in situ*, yielding micron-scale image resolution without the need for excision of a specimen and tissue processing. OCT is analogous to ultrasound B-mode imaging except that it uses low-coherence light rather than sound and performs cross-sectional imaging by measuring the backscattered intensity of light from structures in tissue. The OCT image is a gray-scale or false-color two-dimensional representation of backscattered light intensity in a cross-sectional plane. The OCT image represents the differential backscattering contrast between different tissue types on a micron scale. Because OCT performs imaging using light, it has a one to two order-of-magnitude higher spatial resolution than ultrasound and does not require contact with the specimen or sample.

OCT was originally developed and demonstrated in ophthalmology for high-resolution tomographic imaging of the retina and anterior eye. Because the eye is transparent and is optically accessible, it is well suited for diagnostic OCT imaging. OCT is promising for the diagnosis of retinal disease because it can provide images of retinal pathology with 10 micron resolution, almost one order-of-magnitude higher than previously possible using ultrasound. Recently, OCT has been applied for imaging in a wide range of nontransparent tissues. In tissues other than the eye, the imaging depth is limited by optical attenuation due to scattering and absorption. Ophthalmic imaging is typically performed at 800 nm wavelengths. However, because optical scattering decreases with increasing wavelength, OCT imaging in nontransparent tissues is possible using longer near-infrared wavelengths. In most tissues, imaging depths of 2–3 mm can be achieved using a system detection sensitivity of 100 to 110 dB. Imaging studies have been performed in a wide range of biological, medical, and surgical specialties including developmental biology, cardiology, gastroenterology, urology, and neurosurgery. High-resolution OCT using short coherence length, short pulsed light sources has also been demonstrated and axial resolutions less than 2 μm have been achieved. High-speed real-time OCT at image acquisition rates of 4 to 8 frames per second have also been demonstrated. OCT has been extended to perform Doppler imaging of blood flow and birefringence imaging to investigate tissue injury. Different imaging delivery systems including transverse imaging catheters and forward imaging devices have been developed to enable internal body OCT imaging. Most recently, an OCT catheter has been combined with endoscope-based delivery to perform *in vivo* imaging in animal models and human patients.

This chapter will provide an overview of the OCT technology, beginning with the basic principles of operation. Technological advancements over the last decade have enabled OCT to transition from laboratory- to clinical-based imaging. These enabling advancements will be discussed, followed by representative applications in biology, medicine, surgery, and materials.

Principles of Operation

OCT is based on optical ranging, the high-resolution, high dynamic range detection of backscattered light. In contrast to ultrasound, the velocity of light is extremely high. Therefore, the echo time delay of reflected light cannot be measured directly and interferometric detection techniques must be used. One method for measuring echo time delay is to use low coherence interferometry or optical coherence domain reflectometry. Low-coherence interferometry was first developed for measuring reflections in fiber optics and optoelectronic devices.

The echo time delay of reflected light is measured by using a Michelson-type interferometer (Figure 1). The light reflected from the specimen or sample is interfered with light that is reflected from a reference path of known path length. Interference of the light reflected from the sample arm and reference arm of the interferometer can occur only when the optical path lengths of the two arms match to within the coherence length of the optical source. As the reference arm optical path length is scanned, different echo delays of backscattered light from within the sample are measured. The interference signal is detected at the output port of the interferometer, electronically band-pass filtered, demodulated, digitized, and stored on a computer. The position of the incident beam on the specimen is typically scanned in the transverse direction and multiple axial measurements are performed. This generates a two-dimensional data array that represents the optical backscattering through a cross-sectional plane in the specimen (Figure 2). The logarithm of the backscatter intensity is then mapped to false-color or gray-scale and displayed as an OCT image. The interferometer in an OCT instrument can be implemented using a fiber optic coupler and beam-scanning can be performed with small mechanical galvanometers in order to yield a compact and robust system (Figure 3).

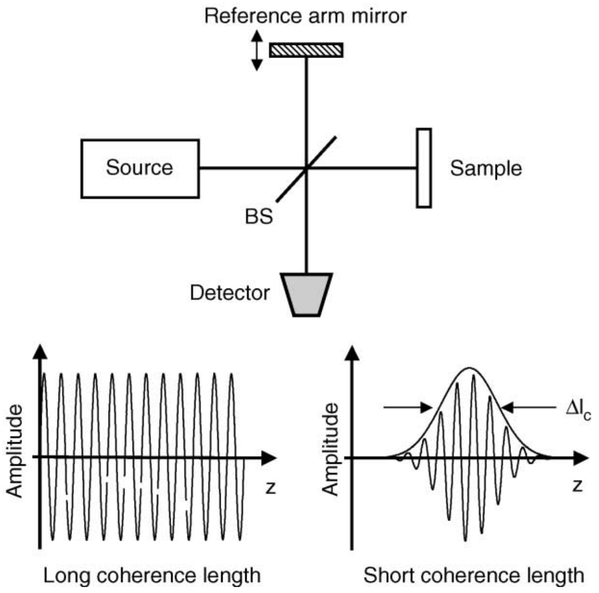

Figure 1 Schematic illustrating the concept of low-coherence interferometry. Using a short coherence-length light source and a Michelson-type interferometer, interference fringes are observed only when the path lengths of the two interferometer arms are matched to within the coherence length of the optical source. Abbreviations: BS, beam splitter; Δl_c, coherence length.

In contrast to conventional microscopy, the axial resolution in OCT images is determined by the coherence length of the light source. The axial point spread function of the OCT measurement as defined by the signal detected at the output of the

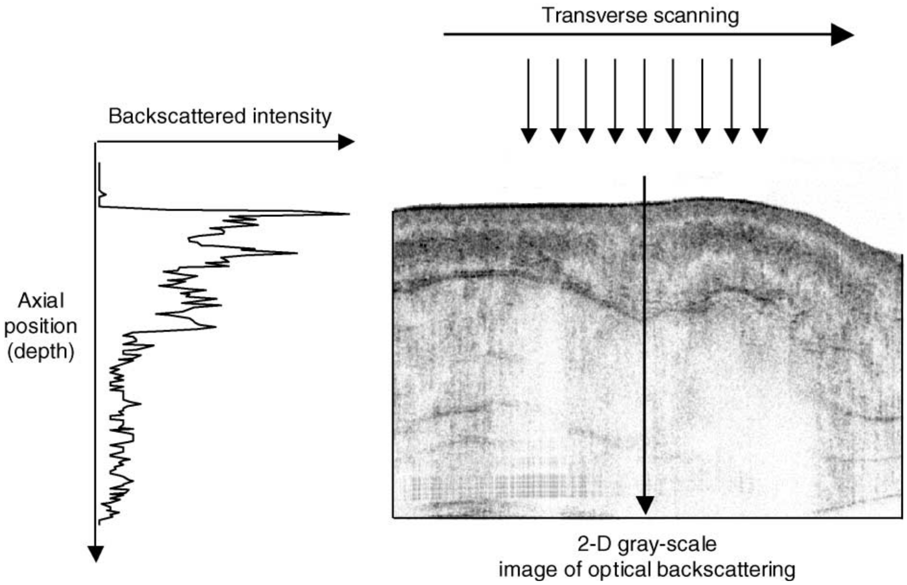

Figure 2 OCT image generation. The OCT image is typically acquired by performing axial measurements of optical backscatter at different transverse positions on the specimen and displaying the resulting two-dimensional data set as a gray-scale (or false-color) image.

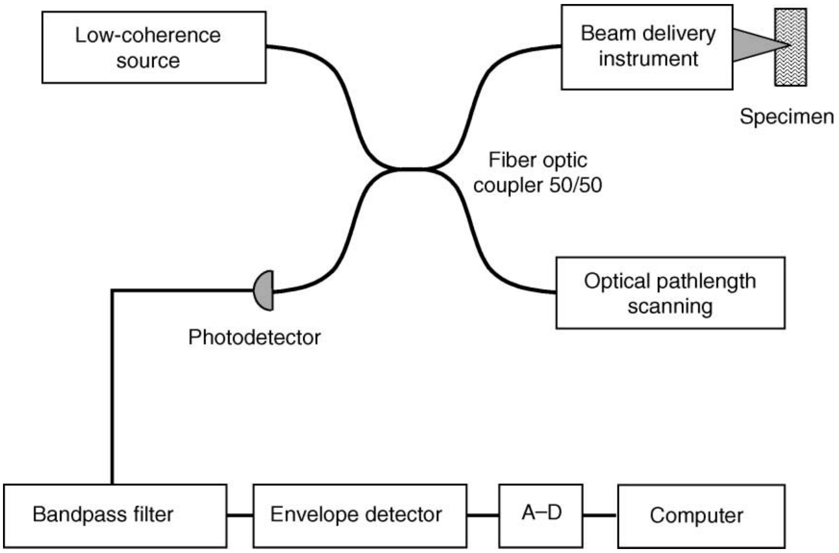

Figure 3 Schematic representation of an OCT system implemented using fiber optics. The Michelson interferometer is implemented using a fiber-optic coupler (beam splitter). The sample arm contains a beam-delivery instrument and the reference arm contains a mechanism for rapidly scanning the optical pathlength. The output of the interferometer is detected with a photodiode, filtered, demodulated, analog-to-digital (A–D) converted, and processed by a computer.

interferometer is the electric-field autocorrelation of the source. The coherence length of the light is the spatial width of the field autocorrelation and the envelope of the field autocorrelation is equivalent to the Fourier transform of its power spectrum. Thus, the width of the autocorrelation function, or the axial resolution, is inversely proportional to the width of the power spectrum. For a source with a Gaussian spectral distribution, the axial resolution Δz is given:

$$\Delta z = \frac{2\ln 2}{\pi} \cdot \frac{\lambda^2}{\Delta\lambda}$$

where Δz and $\Delta\lambda$ are the full-widths-at-half-maximum of the autocorrelation function and power spectrum respectively and λ is the source central wavelength. Figure 4 illustrates the dependence of the coherence length (axial resolution) on the bandwidth of the optical source. To achieve high axial resolution requires broad bandwidth optical sources. Resolution is also improved by using shorter wavelengths, however, shorter wavelengths are scattered and absorbed more in biological tissue.

The transverse resolution in an OCT imaging system is determined by the focused spot size in analogy with conventional microscopy and is given by:

$$\Delta x = \frac{4\lambda}{\pi} \cdot \frac{f}{d}$$

Where d is the spot size on the objective lens and f is its focal length. High transverse resolution can be obtained by using a large numerical aperture and focusing the beam to a small spot size. The transverse resolution is also related to the depth of focus or the confocal parameter $2z_R$ (two times the Raleigh range).

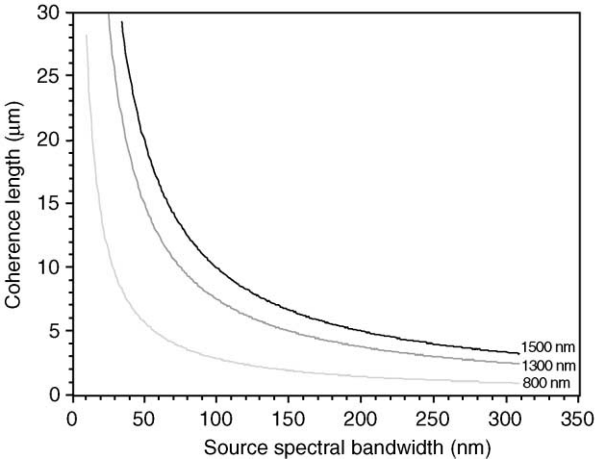

Figure 4 Dependence of coherence length (axial resolution) on the bandwidth of the optical source. Curves are shown for imaging wavelengths of 800 nm, 1300 nm, and 1500 nm. To achieve micron-scale resolution imaging, broad spectral bandwidths are necessary.

$$2z_R = \frac{\pi \Delta x^2}{2\lambda}$$

Thus, increasing the transverse resolution results in a reduced depth of field. Typically, the confocal parameter or depth of focus is chosen to match the desired depth of imaging. Increased resolution may also be obtained by using a higher numerical aperture lens or objective and spatially tracking the focus through the specimen.

Finally, the detection signal-to-noise is given by the optical power backscattered from the sample divided by the noise equivalent bandwidth:

$$\mathrm{SNR} = 10 \log\left(\frac{\eta}{\hbar\omega} \frac{P_{\mathrm{SAM}}}{\mathrm{NEB}} \right)$$

Depending upon the desired signal to noise performance, incident powers of 5–10 mW are typically required for OCT imaging of 250–500 square pixel images at several frames per second. If lower data acquisition speeds or signal-to-noise can be tolerated, power requirements can be reduced accordingly.

Technological Developments

Since the inception of OCT in the early 1990s, there has been rapid technological developments aimed at improving the imaging resolution, acquisition rate, and methods for beam delivery to the tissue or sample. Investigators have also explored other imaging methods using the principles of OCT to extract information from the tissue or sample. Some of these methods have included acquiring optical Doppler signals from moving scatterers or structures, obtaining images based on the polarization state of the returned light, and extracting spectroscopic information based on the local absorption or scattering properties of the tissue.

Optical Sources / High-Resolution Imaging

The majority of OCT imaging systems to date have used superluminescent diodes (SLDs) as low coherence light sources. SLDs are manufactured with a similar structure to that of laser diodes, but commonly end facets are angle-cleaved to suppress lasing. These device structures do not support oscillation modes and generate output based on amplified spontaneous emission, resulting in emission spectra that are broader than laser diodes. SLDs are commercially available at a range of wavelengths including 800 nm, 1.3 μm, and 1.5 μm and are attractive because they are compact, have high efficiency, and low noise. However, output powers are typically limited to only a few milliwatts which limits fast real-time acquisition rates, and the available bandwidths are relatively narrow, permitting imaging with 10–15 micron resolution. Recent advances in short-pulse solid-state laser technology make these sources attractive for OCT imaging in research applications. Femtosecond solid-state lasers can generate tunable, low-coherence light at powers sufficient to permit high-speed OCT imaging. Short pulse generation has been achieved across the full wavelength range in titanium:sapphire ($\mathrm{Ti{:}Al_2O_3}$) from 0.7 μm to 1.1 μm and over more limited tuning ranges near 1.3 μm and 1.5 μm in chromium: forsterite ($\mathrm{Cr^{4+}{:}Mg_2SiO_4}$) and chromium:yttrium-aluminum-garnet ($\mathrm{Cr^{4+}}$:YAG) lasers, respectively. OCT imaging with resolutions of 1 μm and 5 μm has been demonstrated at 800 nm and 1.3 μm respectively using $\mathrm{Ti{:}Al_2O_3}$ and $\mathrm{Cr^{4+}{:}Mg_2SiO_4}$ sources. More compact and convenient sources such as superluminescent fiber sources, are currently under investigation. The titanium:sapphire laser technology is routinely used in multi-photon microscopy applications for its high peak intensities to enable multi-photon absorption and subsequent emission of fluorescence from exogenous fluorescent contrast agents. Combined OCT and multi-photon microscopy has been used to provide complementary image data using a single optical source.

Fast Scanning

The short-pulse solid-state laser technology not only provides broad spectral bandwidths for high-resolution OCT imaging, but also higher output powers to enable fast real-time OCT imaging. Higher incident powers are required to maintain equivalent signal-to-noise ratios when scanning at a faster rate. Linearly translating a reference arm mirror is problematic at high rates and provides axial scan frequencies of approximately 100 Hz, depending on the mirror size and the translating galvanometer. Several investigators have utilized rotating glass cubes, piezoelectric modulators, and multi-pass optical cavities to increase axial scan rates while maintaining scan ranges of 1–2 mm. An optical delay based on the principles used in femtosecond pulse shaping has been demonstrated for OCT. This delay line spectrally disperses the reference arm beam with a grating. The dispersed beam is then focused by a lens on to a rotating mirror mounted on a galvanometer. The mirror, located in the Fourier-transform plane of the lens, imparts a wavelength-dependent phase shift on the light. Subsequently, when re-coupled back into the interferometer, this phase-shift is equivalent to a time-delay in the time domain. The use of high-speed resonant galvanometers has permitted axial scan rates as high as 8 kHz over scan ranges of several millimeters. Depending on the image pixel size (number of axial scans within each image), this scan rate can provide video-rate OCT imaging (30 frames per second). An example of high-speed *functional* OCT imaging is shown in Figure 5. Images of a

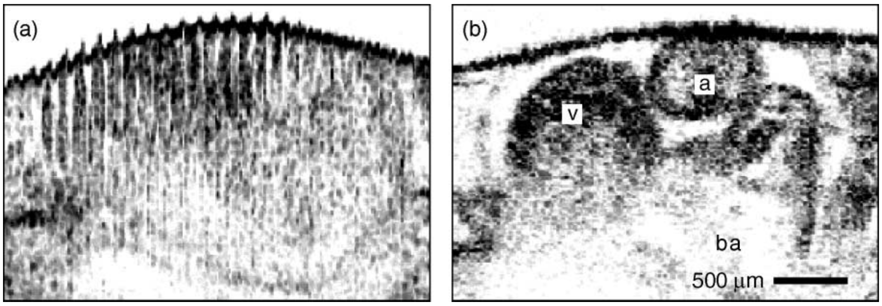

Figure 5 Comparison of slow and fast OCT imaging of a beating *Xenopus* (African frog) tadpole heart. The OCT image in (a) was acquired in 30 seconds and shows oscillations and blurring from the beating heart. The image on the right (b) was acquired in 250 ms (4 frames per second), fast enough to capture the anatomy of the heart with minimal motion artifacts. Abbreviations: a, atrium; ba, bulbous arteriosus; v, ventricle. (Images reprinted with permission from Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.)

beating *Xenopus laevis* (African frog) tadpole were compared at slow (30 s per image) and at fast (250 ms per image) acquisition rates. Artifacts due to cardiac motion were present at slow acquisition rates, but were minimized with acquisition rates of 4 frames per second. Functional cardiac parameters could be measured and high-speed processes such as chamber and valve function could be visualized in real-time.

Depth-priority scanning, as described above, is performed by rapidly varying the optical delay in the reference arm and collecting a single axial scan before translating the beam laterally and repeating this depth-scanning. An alternative method for generating OCT images is with transverse-priority scanning. This is equivalent to optical sectioning in confocal and multi-photon microscopy. Three-dimensional OCT imaging using transverse-priority scanning can be obtained by stepping the position of the reference arm mirror after each *en face* image is acquired. The OCT images produced from transverse-priority scanning can be correlated with confocal or multi-photon microscopy images, which are acquired with the same scanning method. This method can also utilize higher numerical aperture objective lenses to provide high transverse resolutions since a large depth of focus is not needed as in depth-priority OCT scanning. The combination of OCT with high-numerical aperture objectives has been termed optical coherence microscopy.

Doppler Optical Coherence Tomography

Conventional OCT detects the amplitude of the optical backscatter versus position within the sample. Doppler OCT is a technique which has been applied to measuring fluid flow within small capillary tubes and *in vivo* vessels in the skin, retina, and heart of small animal models and humans. This technique is based on digitally sampling the interference fringes that are produced from a moving scatterer within the sample or specimen and determining the Doppler frequency shift in the signal due to the moving scatterer. Figure 6 illustrates two- and three-dimensional optical Doppler data for fluid flow through a cylindrical tube. A two-dimensional cross-sectional profile of fluid flow through a silicon tube with a 600 μm inner-diameter is plotted in Figure 6. The acquired data (points) closely match the predicted profile (line) for these tube dimensions. Below the plot, one method of obtaining a three-dimensional fluid flow profile is illustrated. While imaging the tube in cross-section, the fluid flow velocity was altered with a perfusion pump. The detection filter bandwidth of the OCT electronics served a windowing function, detecting only flow velocities with frequencies within the detection bandwidth. For an unknown flow profile, the center frequency of the detection filter can be shifted to map the flow velocities present within the sample.

Beam Delivery

The OCT imaging technology is modular in design and a variety of optical instruments can be used to deliver the OCT beam to the tissue or sample. Because OCT is fiber-optic based, single optical fibers can be used to deliver the OCT beam and collect the reflected light. The OCT technology can readily be integrated into existing optical instruments such as research and surgical microscopes, ophthalmic slit-lamp biomicroscopes, and hand-held imaging probes.

Imaging penetration is determined by the optical absorption and scattering properties of the tissue or specimen. The imaging penetration for OCT ranges from tens of millimeters for transparent tissues such as the eye to less than 3 mm in highly-scattering tissues such as skin. To image highly-scattering tissues deep within the body, novel beam-delivery

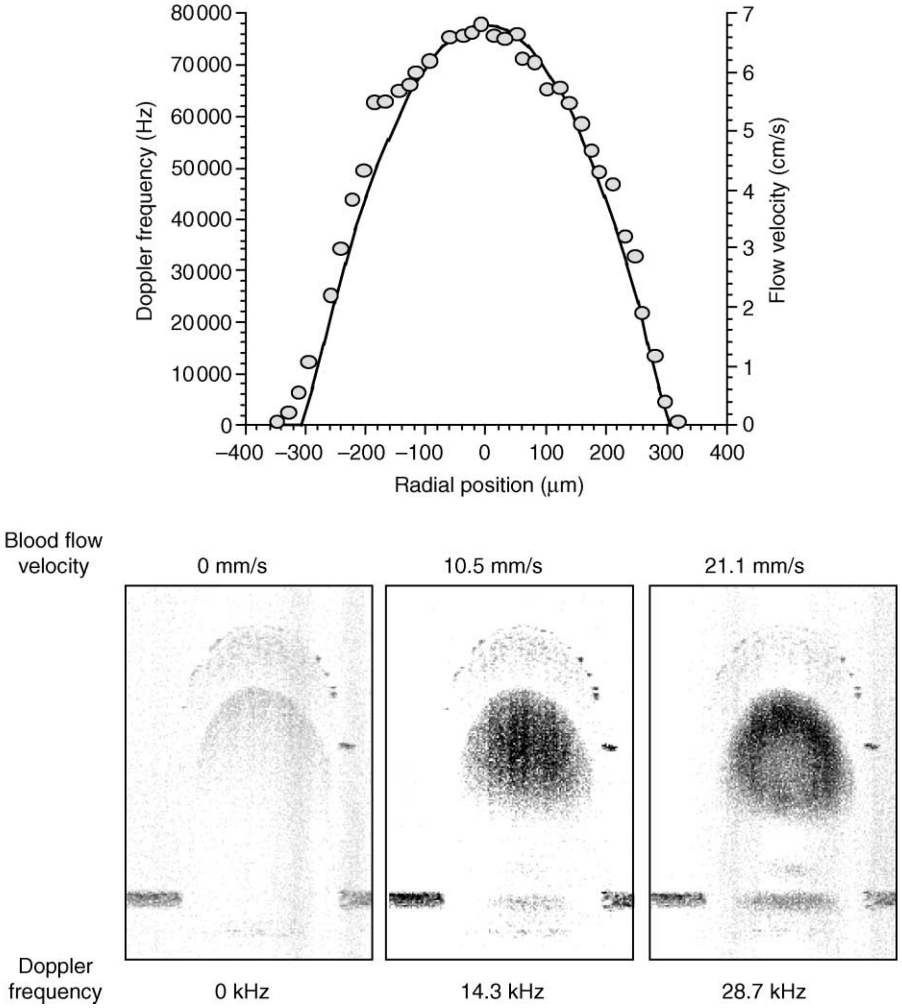

Figure 6 Doppler OCT. A two-dimensional plot of the fluid-flow velocity profile of flowing blood through a 600 μm inner-diameter tube is shown. Data points acquired using OCT closely follow the theoretical laminar flow profile (line). Two-dimensional Doppler OCT images of blood flow through the tube are shown below the plot. As fluid flow increases, only fluid-flow velocities within the bandwidth of the detection electronics are detected. This represents one method by which 3-D fluid flow can be imaged.

instruments have been developed to relay the OCT beam to the site of the tissue to be imaged. An OCT catheter has been developed for insertion into biological lumens such as the gastrointestinal tract. Used in conjunction with endoscopy, the 1 mm diameter catheter can be inserted through the working channel of the endoscope for simultaneous OCT and video imaging. Minimally-invasive surgical procedures utilize laparoscopes, which are long, thin, rigid optical instruments that permit video-based imaging within the abdominal cavity. Laparoscopic OCT imaging has been demonstrated by passing the OCT beam through the optical elements of a laparoscope. Deep solid-tissue imaging is possible with the use of fiber-needle probes. Small (400 μm diameter) needles housing a single optical fiber and micro-optic elements can be inserted into solid tissues and rotated to acquire OCT images. Recently, microfabricated micro-electro-optical-mechanical systems (MEOMS) technology has been used to miniaturize the OCT beam scan mechanism.

Applications

Developmental Biology

OCT has been demonstrated in the field of developmental biology as a method to perform high-resolution, high-speed imaging of developing morphology and function. Cellular-level imaging is possible, providing a non-invasive technique for visualizing cellular processes such as mitosis and migration. Imaging studies have been performed on several standard biological animal models commonly employed in developmental biology investigations including *Rana pipiens* (Leopard frog), *Xenopus laevis*

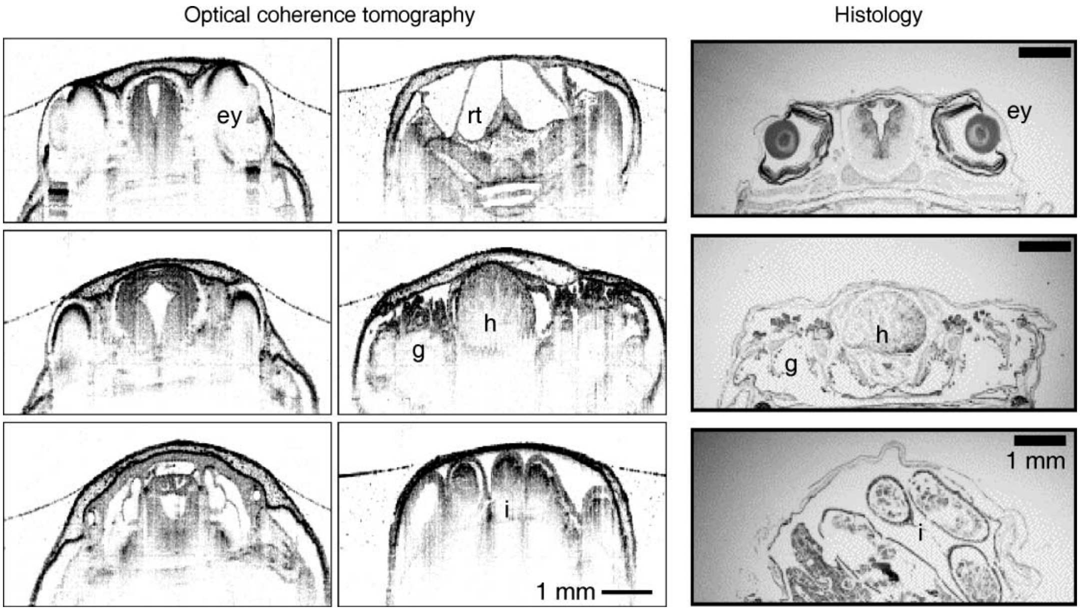

Figure 7 OCT images and corresponding histology of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole. OCT images in the left and middle columns were acquired with the OCT beam incident from the dorsal and ventral sides of the specimen, respectively. The corresponding histology in the right column illustrates strong correlations between OCT images and the actual tissue morphology. Abbreviations: ey, eye; g, gills; h, heart; i, intestines; rt, respiratory tract. (Images reprinted with permission from Boppart SA, Brezinski ME, Bouma BE, *et al.* (1996) Investigation of developing embryonic morphology using optical coherence tomography. *Dev. Biol.* 177: 54–64.)

(African frog), and *Brachydanio rerio* (zebrafish) embryos and eggs, and the murine (mouse) model.

A series of cross-sectional images acquired *in vitro* from the dorsal and ventral sides of a Stage 49 (12 day) *Rana pipiens* (Leopard frog) tadpole is shown in Figure 7. Features of internal architectural morphology are clearly visible in the images. The image of the eye differentiates structures corresponding to the cornea, lens, and iris. Internal morphology not accessible in one orientation due to the specimen size or shadowing effects can be imaged by reorienting the specimen and scanning in the same cross sectional image plane. With the OCT beam incident on the ventral side, images of the respiratory tract, ventricle of the heart, internal gills, and gastrointestinal tract were acquired.

These images can be compared with corresponding histology (Figure 7). Histological images are acquired by euthanizing the specimen, immersing the specimen in a chemical fixative, and physically sectioning thin (2–5 micron-thick) slices using a microtome. The slices are placed on a microscope slide, selectively stained to highlight particular features, and viewed with light microscopy. The correlations between OCT and histology images are strong, suggesting that OCT images can accurately represent the *in vivo* specimen morphology. The potential exists to repeatedly image specimens to quantify organo- and morphogenesis throughout development. Technologies such as OCT are likely to become increasingly important in functional genomics, relating genetic features to the morphology and function in living specimens.

OCT images represent the optical backscatter intensity from regions within the tissue or sample. Because OCT relies on the inherent optical scattering changes to produce imaging contrast, no exogenous contrast agents or fluorophores are necessary. This permits long-term sequential imaging of development *in vivo* without loss of specimen viability. Repeated images of a developing zebrafish embryo within its egg beginning immediately after fertilization and up until hatching have been demonstrated without loss of specimen viability or without developmental abnormalities. In this example, the zebrafish egg and embryo were semi-transparent and the use of OCT significantly complemented observations made using light microscopy. By imaging subtle differences in backscattering intensity, interfacial structural layers millimeters deep within specimens can be clearly delineated.

Previous OCT images have characterized morphological features within biological specimens. These structures are static even though they may have been acquired from *in vivo* specimens. *In vivo* imaging in living specimens, particularly in larger organisms and for medical diagnostic applications, must be performed at high speeds to eliminate motion artifacts within the images. Functional OCT imaging is the quantification of *in vivo* images which yield information characterizing the functional properties of the organ system or organism. High speed OCT permits both the positioning and manipulation of specimens as well as imaging in real time and is a powerful technology for functional imaging in developmental biology animal models.

Studies investigating normal and abnormal cardiac development have been frequently limited by an inability to access cardiovascular function within the intact organism. OCT has been demonstrated for the high-resolution assessment of structure and function in the developing *Xenopus laevis* (African frog) cardiovascular system (Figure 5). The morphology of the *in vivo* cardiac chambers is clearly delineated. Image acquisition rates are fast enough to capture the cardiac chambers in mid-cycle. With this capability, images can be acquired at various times during the cardiac cycle. These frames can be displayed in real-time to produce a movie illustrating the dynamic, functional behavior of the developing heart. OCT, unlike technologies such as computed tomography and magnetic resonance imaging, provides high-speed *in vivo* imaging, allowing quantitative dynamic activity, such as ventricular ejection fraction, to be assessed.

Cellular Imaging

Although previous studies have demonstrated *in vivo* OCT imaging of tissue morphology, most have imaged tissue at ~10–15 μm resolutions, which does not allow differentiation of cellular structure. The ability of OCT to identify the mitotic activity, the nuclear-to-cytoplasmic ratio, and the migration of cells has the potential to not only impact the fields of cell and developmental biology, but also impact medical and surgical disciplines for the early diagnostics of disease such as cancer.

The *Xenopus laevis* (African frog) tadpole has been used to demonstrate the feasibility of OCT for high-resolution *in vivo* cellular and subcellular imaging. Many of the cells in this common developmental biology animal model are rapidly dividing and migrating during the early growth stages of the tadpole, providing an opportunity to image dynamic cellular processes. Three-dimensional volumes of high-resolution OCT data have been acquired from these specimens throughout development. From this 3-D data set, cells undergoing division were identified and tracked in three dimensions. In a similar manner, 3-D data sets were acquired to track single melanocytes (neural crest cells) as they migrated through the living specimens. The ability of OCT to characterize cellular processes such as mitosis and migration not only are of interest in cell and developmental biology, but also have relevance for cancer diagnostics and tumor metastasis.

An example of cellular-level OCT imaging in these specimens is shown in Figure 8. This composite image (0.83 × 1 mm, 1800 × 1000 pixels) was acquired using a titanium:sapphire laser with a broad bandwidth (~260 nm). The axial and transverse resolution in this image is 1 μm and 5 μm, respectively (see Figure 4). Because the high transverse resolution reduced the depth-of-focus to 49 μm, separate OCT images were first acquired with the focus at different depths within the specimen. These images were then assembled to produce the composite image shown in Figure 8. This type of image construction is similar to C-mode ultrasound. Cellular features including cell membranes, nuclei, and nuclear morphology are clearly observed.

Medicine – Imaging Barrett's Esophagus

OCT performs *in situ* imaging and has the potential to be used as a screening technique of early pathological changes in patients. OCT imaging has the potential to be a general diagnostic in many organ systems,

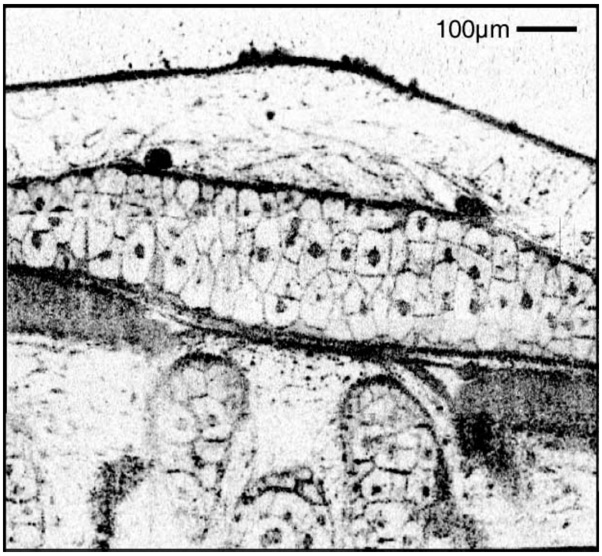

Figure 8 Cellular-level OCT imaging using a broad bandwidth titanium:sapphire laser. Axial and transverse image resolutions are 1 μm and 5 μm, respectively, enabling the visualization of cell membranes, nuclear morphology, and sub-cellular organelles. Cellular processes such as mitosis and migration can be visualized in real-time in living specimens. (Image reprinted with permission from Drexler W, Morgner U, Kartner FX, *et al.* (1999) *In vivo* ultrahigh resolution optical coherence tomography. *Optics Letters* 24: 1221–1223.)

particularly where current methods of screening by excisional biopsy are limited. Barrett's esophagus is a condition where the cells of the distal esophagus undergo a metaplastic change, resembling cells of the lower gastrointestinal tract. This condition is believed to be caused by chronic gastroesophageal reflux. Several studies have demonstrated that Barrett's esophagus is associated with a 30–125 times increased risk of developing adenocarcinoma. For this reason, endoscopic surveillance of Barrett's epithelium every 12–18 months is recommended. Endoscopic screening currently involves random four-quadrant biopsies every 1–2 cm along the length of suspect mucosa. However, excisional biopsy is prone to sampling errors and small foci of carcinoma or dysplasia may be missed. Because of the imprecision and high cost associated with screening, new methods are being developed to assess patients at increased risk. Endoscopic ultrasound catheters have been used for imaging the gastrointestinal tract with 50–100 μm resolution. However, these resolutions are insufficient to resolve early epithelial changes that occur in Barrett's esophagus and the pre-malignant changes that lead to adenocarcinoma of the esophagus.

The ability of OCT to differentiate normal and pathologic tissue is a central question being addressed by many research groups. Comparisons with histology show good correlations between OCT images and histological findings. The image resolution of conventional OCT (10–15 μm) is sufficient to differentiate architectural but not cellular morphology. Still, endoscopic OCT resolution can differentiate normal from Barrett's epithelium in real-time based on differences in epithelial architecture. Crypt- and gland-like structures that disrupt the relatively uniform layers of squamous epithelium can be readily identified, enabling differentiation between normal and Barrett's epithelium (Figure 9). The ability to differentiate normal from Barrett's epithelium suggests that the OCT technology could be used for screening applications.

Oncology – Identifying Tumors and Tumor Margins

OCT has been used to differentiate between the morphological structure of normal and cancerous tissue for a wide-range of tumors. The use of OCT to identify tumors and tumor margins *in situ* will represent a significant advancement for medical or image-guided surgical applications. OCT has been demonstrated for the detection of brain tumors and their margins with normal brain parenchyma, suggesting a role for guiding surgical resection. A hand-held surgical imaging probe was constructed for this application. The compact and portable probe permits OCT imaging within the surgical field while

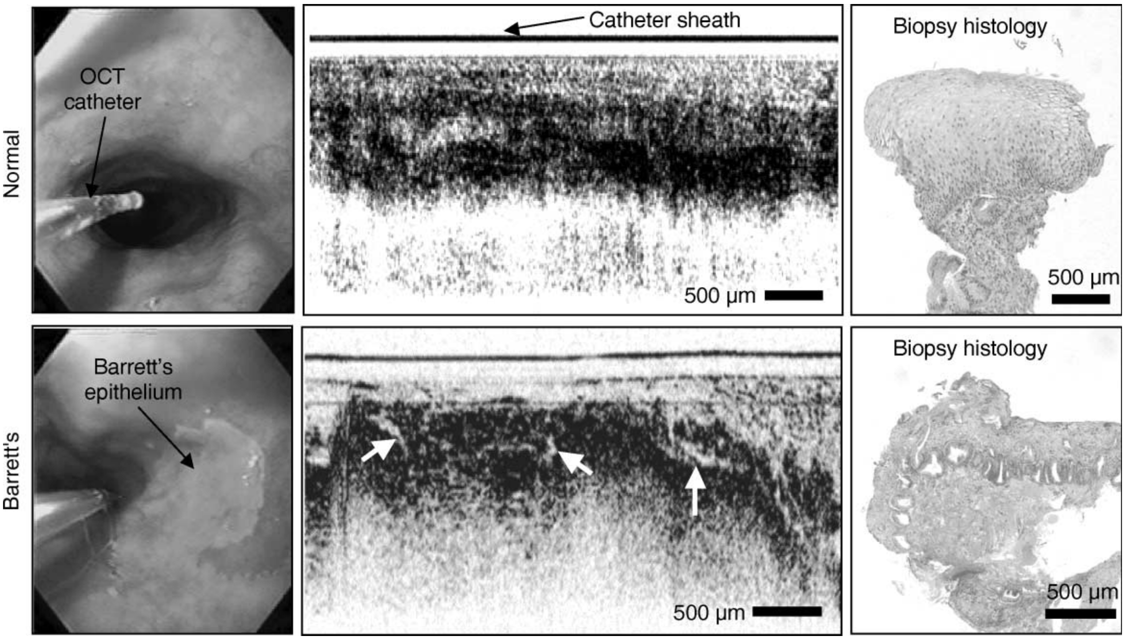

Figure 9 Endoscopic OCT imaging in humans. An OCT catheter was inserted down the working channel of an endoscope. Guided by video imaging, regions of normal and Barrett's esophagus were imaged. Corresponding biopsy histology is shown. OCT can identify Barrett's esophagus by the presence of gland- and crypt-like structures (arrows). (Images reprinted with permission from Li XD, Boppart SA, Van Dam J, *et al.* (2000) Optical coherence tomography: advanced technology for the endoscopic imaging of Barrett's esophagus. *Endoscopy* 32: 921–930.)

the OCT instrument can be remotely located in the surgical suite.

Figure 10 shows a specimen of outer human cerebral cortex with metastatic melanoma. The OCT images in Figure 10a,b were acquired through the tumor. These original images were threshold segmented to identify regions of high backscatter within the tumor. The original images were then overlaid with the segmented data and are shown in Figure 10c,d. The OCT images show increased optical backscattering in the region of the larger tumor (arrows). Smaller tumor lesions also appear within the image (arrows). A shadowing effect is observed below each tumor site due to the increased optical backscatter and the subsequent loss of optical power penetrating beneath the tumor. In Figure 10a,c, the boundary of the tumor can be identified. In Figure 10b,d, the tumor is identified below the surface of normal cortex. The histology in Figure 10e,f confirms the presence and relative size of the tumor.

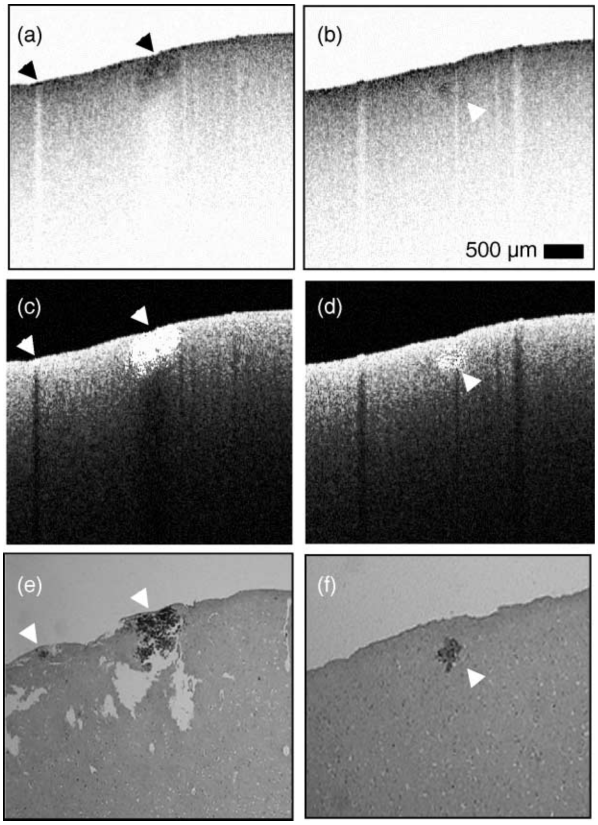

Figure 10 Cancer diagnostics. OCT imaging was used to image metastatic melanoma in an *ex vivo* specimen of human cortex (brain). Raw OCT images (a,b) were threshold-segmented (c,d) to highlight regions of increased optical backscatter and to identify tumor margins. OCT images correlate well with histological findings (e,f). (Images reprinted from Boppart SA, Brezinski ME, Pitris C and Fujimoto JG (1998) Optical coherence tomography for neurosurgical imaging of human intracortical melanoma. *Neurosurgery* 43: 834–841, with permission from Lippincott Williams and Wilkins.)

The image resolutions used to acquire the images in Figure 10 were as high as 16 μm, higher than any current ultrasound, CT, or MRI intraoperative imaging technique. This allowed the tumor-cortex interface and the extent of tumor below the surface to be defined with high resolution. At higher imaging resolutions, it may be possible to image individual tumor cells which have migrated away from the central tumor. OCT represents a new high-resolution optical imaging technology that has the potential for identifying tumors and tumor margins on the micron scale and in real-time. OCT offers imaging performance not achievable with current imaging modalities and may contribute significantly toward the surgical resection of neoplasms.

Image-Guided Surgery

The repair of vessels and nerves is necessary to restore function following traumatic injury. Although the repair of these sensitive structures is performed with the aid of surgical microscopes and loupes to magnify the surgical field, surgeons are limited to the *en face* view that they provide. A technique capable of subsurface, three-dimensional, micron-scale imaging in real-time would permit the intraoperative monitoring of microsurgical procedures. The capabilities of OCT for the intraoperative assessment of microsurgical procedures have been demonstrated. High-speed OCT imaging was integrated with a surgical microscope to performed micron-scale three-dimensional imaging on microsurgical specimens.

OCT has been used to image *in vitro* peripheral nerves and identify individual fascicles. Longitudinal tracking of the spatial orientation of rabbit peripheral nerve fascicles is demonstrated in Figure 11. Representative cross-sectional images of the peripheral nerve are shown in Figure 11a–d. For each slice, one fascicle was manually segmented, colored white, and tracked through the acquired volume of data. Forty images at 100 μm spacing were assembled for the 3-D projections shown in Figure 11e,f. The horizontally- and vertically-rotated projections of the peripheral nerve dramatically reveal the twisted path of the segmented fascicle along the longitudinal axis of the nerve. In addition, a branch in an unsegmented fascicle is observed in Figure 11f. The use of OCT to acquire multiple cross-sectional images and three-dimensionally reconstruct the peripheral nerve offers the opportunity to determine the relative diameters of individual fascicles and to longitudinally track their spatial orientation. OCT may also be useful at identifying and grading the degree of injury in nerves during surgical repair. These results have shown how 2-D OCT images and 3-D OCT projections can

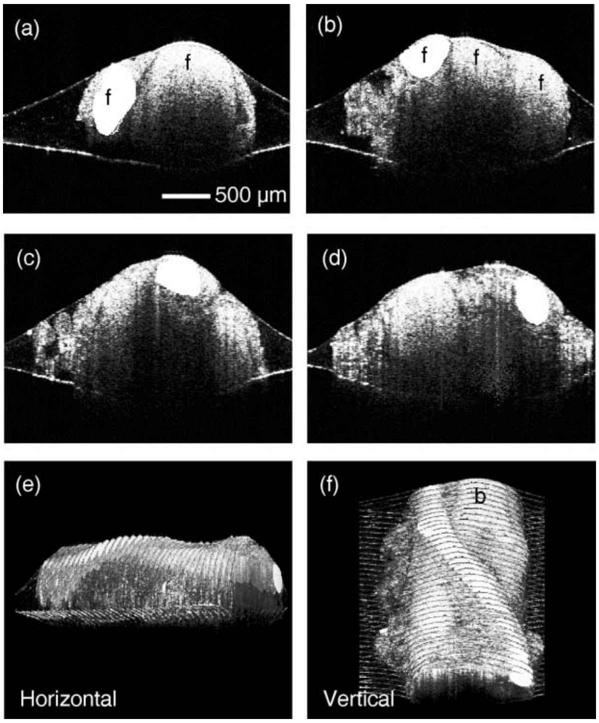

Figure 11 OCT image-guided microsurgery. Multiple cross-sectional OCT images were acquired through a resected rabbit peripheral nerve (a–d). A single nerve fascicle (f) was identified and segmented longitudinally along the length of the nerve. Three-dimensional reconstruction projections of the nerve in (e,f) reveal the tortuous path of the segmented fascicle as well as reveal a bifurcation (b). Intraoperative OCT can guide surgical repair and potentially identify regions of damaged tissue. (Images reprinted from Boppart SA, Bouma BE, Pitris C, *et al.* (1998) Intraoperative assessment of microsurgery with three-dimensional optical coherence tomography. *Radiology* 208: 81–86, with permission from the Radiology Society of North America.)

provide diagnostic feedback to assess microsurgical anastomoses. This previously unavailable diagnostic ability offers the potential to directly impact and improve patient outcome by incorporating high-speed, high-resolution intraoperative image-guidance during microsurgical procedures.

Surgical intervention requires visualization to identify tissue morphology, precision to avoid sensitive tissue structures, and continuous feedback to monitor the extent of the intervention. OCT may provide the technological advancements to improve the operative procedure. The feasibility of OCT to perform image-guided surgical intervention has been investigated. OCT has been used to monitor laser ablation therapy in real-time and may enable more precise control of laser delivery. An argon-laser ablation sequence of fresh *ex vivo* rat rectus abdominis muscle is shown in Figure 12. A pair of blood-filled vessels were located 1.5 mm below the tissue surface and centered within the OCT image. Because blood has a higher absorption coefficient than muscle at the 514 nm argon laser wavelength, coagulation and thermal heating of the blood occurred prior to thermal damage to the overlying tissue. The accompanying histology reveals both the thermally-damaged overlying tissue and coagulated blood present within the vessel pair.

These examples demonstrate the use of OCT for guiding and monitoring surgical intervention. The laser is only a representative interventional surgical technique for a wide range of instruments and techniques including scalpels, electrosurgery, radio-frequency, microwaves, and ultrasound ablation. OCT imaging was performed at 8 frames per second, fast enough to capture dynamic changes in the optical properties of the tissue during thermal ablation. These image sequences provided interesting insight into ablation mechanisms for a variety of tissue types. OCT can monitor the extent of thermal injury below the surface of the tissue by imaging the changes in optical backscatter. OCT imaging can therefore provide empiric information for dosimetry to minimize the extent of collateral injury. The use of OCT for guiding surgical interventions has the potential to improve intraoperative monitoring and more effectively control interventional procedures.

Materials

While the majority of OCT applications have been in the fields of biology and medicine. OCT has also been demonstrated in the non-biological areas of materials investigation, optical data storage, and microfluidic devices. The highly-scattering or reflecting optical properties of many materials prohibits deep imaging penetration using OCT. Many material defects, however, originate near the surface or at interfacial boundaries, making the use of OCT a possibility for inspection and quality control. OCT has been used to identify subsurface defects in ceramics and polymer composites. The optical ranging capabilities of OCT through scattering materials has been utilized for increasing the data storage capacity by assembling multiple layers of optically-accessible data.

The advancement of microfabrication techniques has led to increasingly complex microfluidic and bioMEM (biological micro-electro-mechanical) systems. Microstructures within microfluidic systems range from 10–1000 μm, within both the imaging depth and resolution of OCT. In addition, microfluidic systems are typically fabricated from transparent or semi-transparent substrates, facilitating imaging penetration to deeper three-dimensional features.

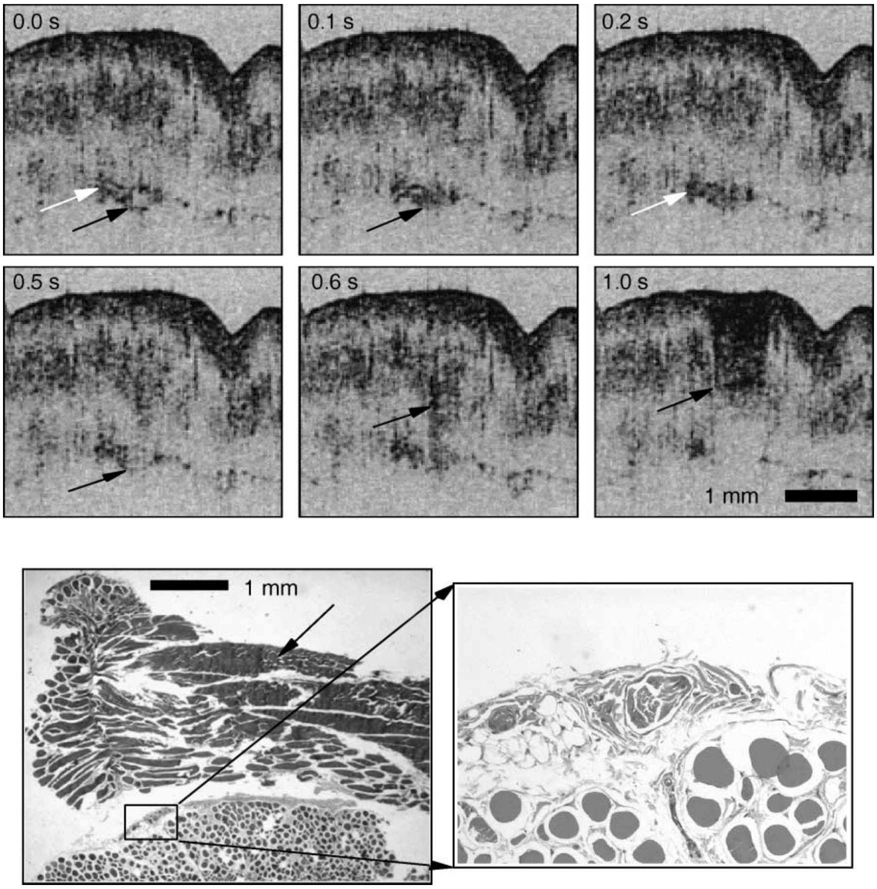

Figure 12 OCT image-guided laser ablation. High-speed OCT image sequence of argon-laser ablation of blood vessels in *ex vivo* rat muscle tissue. Thermal heating within the blood vessels (arrows, 0.0–0.5 s) induces coagulation prior to injury to superficial muscle tissue (0.6–1.0 s). Accompanying histology shows coagulated blood within the vessels.

OCT is not only capable of imaging three-dimensional microstructures at micron-scale resolutions, but is also capable of obtaining dynamic functional data from microfluidic systems. Microfluidic devices made from polymeric materials are attractive since they provide flexibility in the design of the channels, can be manufactured in large volumes, and are mostly bio-compatible. A three-dimensional imaging technique, such as OCT, will be instrumental in the analysis of these devices and the monitoring of device performance.

Conclusions

The capabilities of OCT offer a unique and informative means of imaging biological specimens and non-biological samples. The non-contact nature of OCT and the use of low-power near-infrared radiation for imaging causes few harmful effects on living cells or damage to materials. OCT imaging does not require the addition of fluorophores, dyes, or stains in order to improve contrast in images. Instead, OCT relies on the inherent optical contrast generated from variations in optical scattering and index of refraction. These factors permit the use of OCT for extended imaging over the course of hours, days, or weeks. OCT permits the cross-sectional imaging of tissue and samples and enables *in vivo* structure to be visualized in opaque specimens or in specimens too large for high-resolution confocal or light microscopy.

Imaging at cellular and subcellular resolutions with OCT is an important area of ongoing research. The *Xenopus* (African frog) developmental animal model has been commonly used because its care and handling are relatively simple while allowing cells with a high mitotic index to be assessed. Many of the cells observed were as large as 100 μm in diameter, but ranged in size down to a few microns, below the resolution of most OCT systems. Imaging human cells *in vivo* remains a challenge since differentiated human cells are 10 to 20 microns in size, too small for most OCT systems. With further advances in OCT technology, improved discrimination and imaging of more detailed structures should be possible. New laser sources at other wavelengths in the near-infrared can enhance tissue contrast as well as potentially provide functional information since tissue scattering and absorbance properties in specimens are wavelength dependent. Short coherence length short pulse laser sources have been used to achieve higher axial resolutions on the order of 1 to 3 microns. Unfortunately, unlike superluminescent diode source, these high-speed and high-resolution systems utilize femtosecond lasers that are relatively complex and costly. Developing compact and portable optical sources at near-infrared wavelengths, broad spectral bandwidths, and high output powers is an area of active research.

Optical coherence tomography provides high-resolution morphological, functional, and cellular information of biological, medical, and material specimens and samples. OCT represents a multifunctional investigative tool that not only complements many of the existing imaging technologies available today, but also is poised to become a major optical imaging modality.

List of Units and Nomenclature

Cr^{4+}:Mg_2SiO_4	chromium:forsterite
Cr^{4+}:YAG	chromium:yttrium-aluminum-garnet
CT	computed tomography
d	beam diameter
f	focal length
MRI	magnetic resonance imaging
NEB	noise equivalent bandwidth
OCT	optical coherence tomography
P_{SAM}	power from sample
SNR	signal-to-noise ratio
SLD	superluminescent diode
Ti:Al_2O_3	titanium:sapphire
z_R	Rayleigh range
3-D	three-dimensional
Δx	transverse resolution
$\Delta\lambda$	spectral bandwidth
Δz	axial resolution
η	quantum efficiency
λ	center wavelength

See also

Coherence: Coherence and Imaging; Overview. **Coherent Lightwave Systems**. **Interferometry:** Overview. **Fiber and Guided Wave Optics:** Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging. **Terahertz Technology:** Coherent Terahertz Sources. **Tomography:** Tomography and Optical Imaging. **Microscopy:** Confocal Microscopy; Imaging multiple Photon Fluorescence Microscopy; Overview.

Further Reading

Boppart SA, Tearney GJ, Bouma BE, *et al.* (1997) Noninvasive assessment of the developing *Xenopus* cardiovascular system using optical coherence tomography. *Proceedings of the National Academy of Sciences* 94: 4256–4261.

Boppart SA, Bouma BE, Pitris C, *et al.* (1998) *In vivo* cellular optical coherence tomography imaging. *Nature Medicine* 4: 861–864.

Boppart SA, Deutsch TF and Rattner DW (1999) Optical imaging technology in minimally invasive surgery. *Surgical Endoscopy* 13: 718–722.

Boppart SA, Brezinski ME and Fujimoto JG (2000) Optical Coherence Tomography Imaging. In: Tuan RS and Lo CW (eds) *Methods in Molecular Biology, Vol. 135: Developmental Biology Protocols*, vol. 1, pp. 217–233. Totowa, NJ: Humana Press, Inc.

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker Inc.

Chen Z, Zhao Y, Srinivas SM, *et al.* (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

Fujimoto JG, Brezinski ME, Tearney GJ, *et al.* (1995) Biomedical imaging and optical biopsy using optical coherence tomography. *Nature Medicine* 1: 970–972.

Fujimoto JG, Pitris C, Boppart SA and Brezinski ME (2000) Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy. *Neoplasia* 2(1–2): 9–25.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Masters BR (ed.) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography*. SPIE Milestone Series, Vol. MS 165, Thompson BJ (general ed.) Bellingham, WA: SPIE Optical Engineering Press.

Puliafito CA, Hee MR, Schuman JS and Fujimoto JG (1995) Optical Coherence Tomography of Ocular Diseases. Thorofare, NJ: Slack, Inc.

Schmitt M (1999) Optical coherence tomography (OCT): A review. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1205–1215.

Tadrous PJ (2000) Methods for imaging the structure and function of living tissues and cells: 1. Optical coherence tomography. *Journal of Pathology* 191(2): 115–119.

Tearney GJ, Brezinski ME, Bouma BE, *et al.* (1997) *In vivo* endoscopic optical biopsy with optical coherence tomography. *Science* 276: 2037–2039.

Welzel J (2001) Optical coherence tomography in dermatology: a review. *Skin Research and Technology* 7(1): 1–9.

Tomography and Optical Imaging

Z Chen, University of California, Irvine, CA, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

'Tomos' means section, in Greek. A tomogram is a sectional image of the internal structure of a three-dimensional object. One of the challenges for tomographic imaging of biological tissue using visible or near-infrared light is tissue optical scattering. When photons propagate inside a tissue, they undergo multiple scattering. The mean free path of photons in the skin, for example, is 5 μm. If one focuses the light 1 mm below the skin surface, most photons undergo multiple scattering and never reach the target point. Less than one out of a million photons will be able to reach the target, undergo a single backscattering, and return to the detector (Figure 1). Most photons that are backscattered and reach the detector are multiple scattered photons that carry little information on the target point. The challenge for tomographic imaging of tissues is to separate the single backscattered photons, that carry information about the target point, from multiple scattered photons.

Several optical tomography techniques have been developed. Each uses different techniques to discriminate the multiple scattering photons from the minimally scattered photons for imaging constructions. The confocal microscope, for example, uses a spatial filtering technique to reject multiple scattering photons because multiple scattered photons are more spread out in their angular distribution than the minimally scattered photons. Time-of-flight imaging uses a time gating technique to discriminate diffuse photons because single backscattered photons arrive earlier than diffuse photons. Optical coherence tomography (OCT) uses coherent gating to discriminate single scattered photons from multiple scattered photons. These gating techniques provide good discrimination of minimally scattered photons from multiple scattered photons and can form high-resolution sectional images, albeit with a limited imaging depth. Diffuse photons can propagate deep into the tissue and these photons can also be used for optical tomography. If a modulated photon density wave is propagated inside the tissue, the scattering and absorption properties of the tissue will modulate the phase and amplitude of the photon density wave. However, the resolution of optical tomography, based on diffuse photons, is usually much lower than that of the tomography constructed from single or minimally scattered photons. Three of the most important parameters for optical tomography are contrast, spatial resolution, and image depth.

Most optical tomography uses absorption, scattering, and fluorescent properties of materials as contrast. OCT, for example, uses variations in tissue scattering as contrast for tomographic imaging. However, in many instances, and especially during the early stages of disease, the change in tissue scattering properties between normal and diseased tissues is small and difficult to measure. One of the great challenges for extending applications of OCT is to find more contrast mechanisms that can provide physiological information in addition to morphological structure. This article will focus on optical tomography technologies that use nonconventional contrast mechanisms to obtain tomographic images. These include optical Doppler tomography (ODT), optical polarization tomography, and optical second-harmonic tomography. These optical tomography technologies are extensions of OCT with contrast mechanisms that can provide physiological information, in addition to morphological structure. ODT, also known as Doppler OCT, combines the Doppler principle with OCT to obtain high-resolution tomographic images of tissue structure and blood flow simultaneously. Optical polarization tomography, also known as polarization sensitive OCT (PS-OCT), combines polarization sensitive

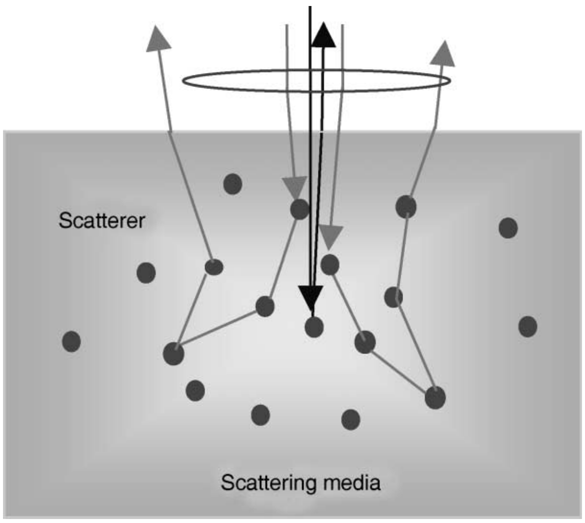

Figure 1 Optical imaging in scattering media.

detection with OCT, to determine tissue birefringence. Optical second-harmonic tomography, also known as second-harmonic optical coherence tomography (SH-OCT), combines second-harmonic generation (SHG) with coherence gating to obtain images with molecular contrast. These functional extensions of OCT provide clinically important information on tissue physiology, such as tissue blood perfusion, oxygen saturation, hemodynamics, and structural remodeling. Each provides several potential clinical applications, such as vasoactive drug screening, tissue viability and burn depth determination, tumor angiogenesis studies, tumor diagnosis, bleeding ulcer management, and ocular pathology evaluation.

Optical Doppler Tomography

ODT combines the Doppler principle with OCT to obtain high-resolution tomographic images of static and moving constituents in high scattering media. When light backscattered from a moving particle interferes with the reference beam, a Doppler frequency shift (f_D) occurs in the interference fringe:

$$f_D = \frac{1}{2\pi}(\mathbf{k}_s - \mathbf{k}_i) \cdot \mathbf{v} \qquad [1]$$

where $\mathbf{k}_i$ and $\mathbf{k}_s$ are wavevectors of incoming and scattered light, respectively, and $\mathbf{v}$ is the velocity vector of the moving particle (Figure 2). Since ODT measures the backscattered light, assuming the angle between flow and sampling beam is θ, the Doppler shift equation is simplified to:

$$f_D = \frac{2V \cos\theta}{\lambda_0} \qquad [2]$$

where λ_0 is the vacuum center wavelength of the light source.

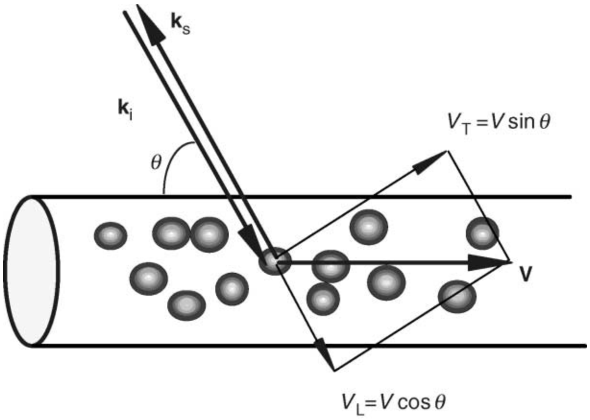

Figure 2 Schematic of flow direction and probe beam angle.

The optical system of ODT is similar to that of OCT. The primary difference is in signal processing. Figure 3 illustrates an ODT instrument that uses a fiber optic Michelson interferometer with a broadband light source. Light from a broadband partial coherence source is coupled into a fiber interferometer by a 2 × 2 fiber coupler and then split equally into reference and target arms of the interferometer. Light backscattered from the turbid sample is coupled back into the fiber and forms interference fringes with the light reflected from the reference arm. A rapid-scanning optical delay line is used for group delay and axial scanning. Because this delay line can decouple the group delay from the phase delay, an electro-optical phase modulator is introduced to produce a stable carrier frequency. The interference fringe intensity signal is amplified, bandpass filtered, and digitized with a high-speed analog-to-digital converter. The signal processing is carried out at the same time as data are transferred to the computer, and real-time display can be accomplished with the use of a digital signal processing board.

To understand the signal processing of ODT, let us look at the fringe signal due to the moving particles.

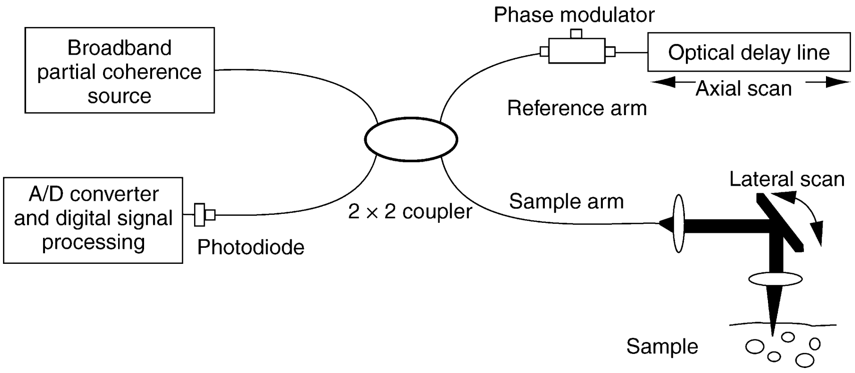

Figure 3 Schematic of ODT system consists of a fiber-based Michelson interferometer with a partially coherent light source.

If we denote $U(t)$ as a complex-valued analytic signal of a stochastic process representing the field amplitude emitted by a low coherent light source and $\bar{U}(\nu)$ as the corresponding spectral amplitude at optical frequency ν, the amplitude of a partially coherent source light coupled into the interferometer at time t, is written as a harmonic superposition:

$$U(t) = \int_0^\infty \bar{U}(\nu) e^{2\pi i\nu t}\, d\nu \qquad [3]$$

Because the stochastic process of a partially coherent light source is stationary, the cross spectral density of $\bar{U}(\nu)$ satisfies:

$$\langle \bar{U}^*(\nu)\bar{U}(\nu')\rangle = S_0(\nu)\delta(\nu-\nu') \qquad [4]$$

where $S_0(\nu)$ is the source power spectral density, and $\delta(\nu-\nu')$ is the Dirac delta function. Assuming that light couples equally into the reference arm and sample arm with spectral amplitude of $\bar{U}_0(\nu)$, the light coupled back to the detector from the reference, $\bar{U}_r(\nu)$, and sample, $\bar{U}_s(\nu)$, are:

$$\bar{U}_r(\nu) = e^{2\pi i\nu(2L_r+L_d)/c} K_r(\nu) e^{i\alpha_r} \bar{U}_0(\nu) \qquad [5]$$

$$\bar{U}_s(\nu) = e^{2\pi i\nu(2L_s+L_d)/c} K_s(\nu) e^{i\alpha_s} \bar{U}_0(\nu) \qquad [6]$$

where L_r and L_s are the optical pathlengths from the beamsplitter to the reference mirror and sample, respectively; L_d is the optical pathlength from the beamsplitter to the detector; and $K_r(\nu)e^{i\alpha_r(\nu)}$ and $K_s(\nu)e^{i\alpha_s(\nu)}$ are the amplitude reflection coefficients of light backscattered from the reference mirror and turbid sample, respectively.

The total power detected at the interferometer output is given by a time-average of the squared light amplitude:

$$P_d(\tau) = \langle |U_s(t) + U_r(t+\tau)|^2 \rangle \qquad [7]$$

where τ is the depth scan time delay. Combining harmonic expansions for $U_s(t)$ and $U_r(t)$ and applying eqn [4] when calculating the time-average, the total power detected is a sum of three terms representing reference (I_r), sample (I_s), and the interference fringe intensity ($\Gamma_{\mathrm{ODT}}(\tau)$):

$$\begin{aligned} P_d(\tau) &= \int_0^\infty (P_r(\nu) + P_s(\nu) + P_{\mathrm{ODT}}(\nu))\, d\nu \\ &= I_r + I_s + \Gamma_{\mathrm{ODT}}(\tau) \end{aligned} \qquad [8]$$

with

$$P_r(\nu) = S_0(\nu)|K_r(\nu)|^2 \qquad [9]$$

$$P_s(\nu) = S_0(\nu)|K_s(\nu)|^2 \qquad [10]$$

$$\begin{aligned} P_{\mathrm{ODT}}(\nu) = {} & 2S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu(\Delta_d/c+\tau) + \alpha_s(\nu) - \alpha_r(\nu)] \end{aligned} \qquad [11]$$

and

$$I_r = \int_0^\infty P_r(\nu)\, d\nu \qquad [12]$$

$$I_s = \int_0^\infty P_s(\nu)\, d\nu \qquad [13]$$

$$\Gamma_{\mathrm{ODT}}(\tau) = \int_0^\infty P_{\mathrm{ODT}}(\nu)\, d\nu \qquad [14]$$

When there is a moving particle, light scattered from this particle is equivalent to a moving phase front; therefore, Δ_d in eqn [11] can be written as:

$$\Delta_d = \Delta + 2\bar{n}V_z t \qquad [15]$$

where Δ is the optical pathlength difference between light in the sampling and reference arms; V_z is the velocity of a moving particle parallel to the probe beam; and $\bar{n}$ is the refractive index of flow media.

To simplify the computation, we assume α_s and α_r are constants over the source spectrum and can be neglected. The spectral domain fringe signal, $P_{\mathrm{ODT}}(\nu)$, is simplified to:

$$\begin{aligned} P_{\mathrm{ODT}}(\nu) = {} & 2S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu((\Delta + 2\bar{n}V_z t)/c + \tau)] \end{aligned} \qquad [16]$$

The corresponding time domain signal, $\Gamma_{\mathrm{ODT}}(\tau)$, is given by:

$$\begin{aligned} \Gamma_{\mathrm{ODT}}(\tau) = {} & 2\int_0^\infty S_0(\nu)K_r(\nu)K_s(\nu) \\ & \times \cos[2\pi\nu((\Delta + 2\bar{n}V_z t)/c + \tau)]\, d\nu \end{aligned} \qquad [17]$$

A comparison of eqns [16] and [17] shows that there is a Fourier transformation relation between spectral domain and time domain signals. Consequently, there are two methods to acquire the ODT signal: the time domain method and the spectral domain method.

In the time domain method, a delay line is incorporated in the reference arm to generate a delay. A spectrogram analysis or phase-resolved algorithm is then used to determine the Doppler frequency shift. In the spectral domain method, the reference mirror is fixed, and there is no depth scan ($\tau = 0$). The spectral domain fringe signal, $P_{\mathrm{ODT}}(\nu)$, is obtained either by a spectrometer at the detection arm, or by a frequency sweeping light source. The time domain signal, $\Gamma_{\mathrm{ODT}}(\tau)$, is determined from the spectral domain signal by a Fourier transformation.

Spectrogram Method

The spectrogram method uses time–frequency analysis to calculate the Doppler shift. Signal processing algorithms to obtain structural and velocity images from recorded temporal interference fringe intensity using the spectrogram method, are illustrated in Figure 4.

The spectrogram is an estimate of the power spectrum of the temporal interference fringe intensity ($\Gamma_{\mathrm{ODT}}(\tau_i, f_j)$) in the ith time delay window. The power spectrum of the temporal interference fringe at the ith pixel, corresponding to time delay τ_i in the structural and velocity images, is calculated by a short-time fast Fourier transformation (STFFT) or a wavelet transformation:

$$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) = |\mathrm{STFFT}(f_m; \Gamma_{\mathrm{ODT}}(\tau))|^2 \qquad [18]$$

where f_m is the discrete frequency value. A tomographic structural image is obtained by calculating the value of the power spectrum at the phase modulation frequency (f_0). Fluid flow velocity is determined from the Doppler frequency shift (f_{D}), which is the difference between the carrier frequency established by the optical phase modulation (f_0) and the centroid (f_{c}) of the measured power spectrum at the ith pixel:

$$V_{\mathrm{ODT}}(i) = \frac{\lambda_0 f_{\mathrm{D}}}{2\bar{n}\cos\theta} = \frac{\lambda_0(f_{\mathrm{c}} - f_0)}{2\bar{n}\cos\theta} \qquad [19]$$

where we have assumed, $\mathbf{k}_{\mathrm{s}} = -\mathbf{k}_{\mathrm{i}}$ and θ is the angle between $\mathbf{k}_{\mathrm{i}}$ and $\mathbf{v}$. The centroid of the measured power spectrum is determined by:

$$f_{\mathrm{c}} = \sum_m f_m \hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) \Big/ \sum_m \hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m) \qquad [20]$$

Figure 5 shows the *in vivo* structural and blood flow images from a chick chorioallantoic membrane. The lumen wall, chorion membrane, and yolk sac membrane are observed in structural image (a). In the velocity image (b), static regions ($V = 0$) appear dark, while blood flowing at different velocities appears as different brightnesses on the gray scale. The velocity profile taken from a horizontal cross-section passing through the center of the vessel is shown in (c).

Phase-Resolved ODT Method

Although spectrogram methods allow simultaneous imaging of *in vivo* tissue structure and flow velocity, the velocity sensitivity is limited for high-speed imaging. When STFFT or wavelet transformation is used to calculate flow velocity, the resolution is determined by the window size of the Fourier transformation for each pixel. The minimum detectable Doppler frequency shift, f_{D}, varies inversely with the STFFT window size. Because pixel acquisition time is proportional to the STFFT window size, the image frame rate is limited by velocity resolution. Furthermore, spatial resolution is also proportional to the STFFT window size. Therefore, a large STFFT window size increases velocity resolution while decreasing imaging speed and spatial resolution. This coupling between velocity sensitivity, spatial resolution, and imaging speed prevents the spectrogram method from achieving simultaneously both high imaging speed and high velocity sensitivity, which are essential for measuring flow in small blood vessels where flow velocity is low.

Phase-resolved ODT overcomes the compromise between velocity sensitivity and imaging speed by using the phase change between sequential scans to construct flow velocity images. The phase information of the fringe signal can be determined from the complex analytical signal $\tilde{\Gamma}(t)$, which is determined through analytic continuation of the measured interference fringe function, $\Gamma(t)$, using a Hilbert transformation:

$$\tilde{\Gamma}(t) = \Gamma(t) + \frac{i}{\pi} P \int_{-\infty}^{\infty} \frac{\Gamma(\tau)}{\tau - t}\,\mathrm{d}\tau = A(t)\mathrm{e}^{i\phi(t)} \qquad [21]$$

where P denotes the Cauchy principle value, i is the complex number, and $A(t)$ and $\phi(t)$ are amplitude and phase of $\tilde{\Gamma}(t)$, respectively. Because the interference signal $\Gamma(t)$ is quasi-monochromatic, the complex

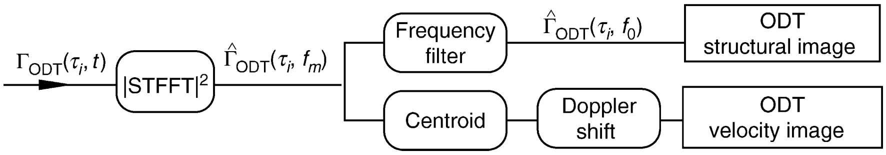
$\Gamma_{\mathrm{ODT}}(\tau_i, t)$
|STFFT|^2
$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_m)$
Frequency filter
$\hat{\Gamma}_{\mathrm{ODT}}(\tau_i, f_0)$
ODT structural image
Centroid
Doppler shift
ODT velocity image

Figure 4 Signal processing algorithms for ODT structural and velocity images.

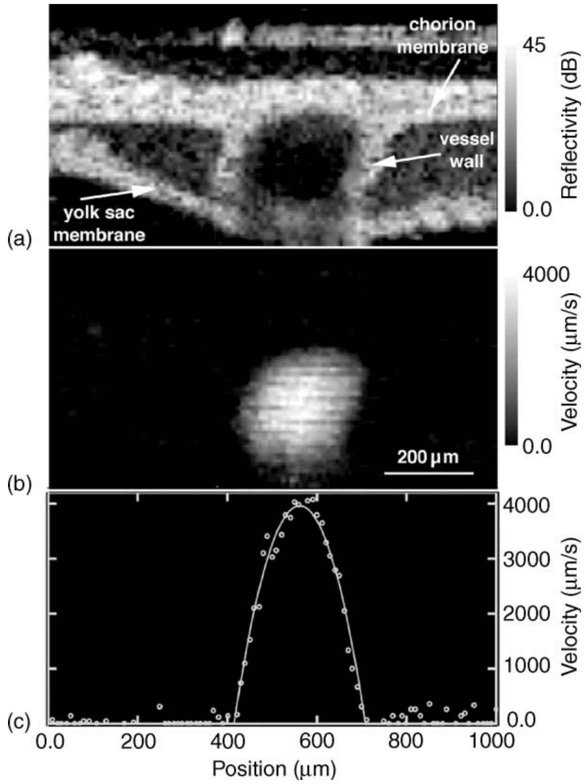

Figure 5 ODT images of *in vivo* blood flow in a CAM vein. (a) Structural image, (b) velocity image, and (c) velocity profile taken from a horizontal cross-section passing through the center of the vein, where the open circles are experimental data and the solid line is a parabolic fit. Reproduced with permission from Chen Z, MilnerTE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

analytical signal can be determined by:

$$\tilde{\Gamma}(t) = 2\int_0^{\infty}\int_0^{\tau}\Gamma(t')\exp(-2\pi i v t')\,\mathrm{d}t' \times \exp(2\pi i v t)\,\mathrm{d}v \qquad [22]$$

where τ is the time duration of the fringe signal in each axial scan.

The Doppler frequency shift (f_n) at nth pixel in the axial direction is determined from the average phase shift between sequential A-scans. This can be accomplished by calculating the phase change of sequential scans from the individual analytical fringe signal:

$$f_n = \frac{\Delta\phi}{2\pi T} = \frac{1}{2\pi T}\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\left[\tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_{j+1}(t_m)}{\mathrm{Re}\tilde{\Gamma}_{j+1}(t_m)}\right) - \tan^{-1}\left(\frac{\mathrm{Im}\tilde{\Gamma}_j(t_m)}{\mathrm{Re}\tilde{\Gamma}_j(t_m)}\right)\right] \qquad [23]$$

Alternatively, the phase change can also be calculated by the cross-correlation method:

$$f_n = \frac{1}{2\pi T}\tan^{-1}\left(\frac{\mathrm{Im}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}{\mathrm{Re}\left[\sum_{m=(n-1)M}^{nM}\sum_{j=1}^{N}\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)\right]}\right) \qquad [24]$$

where $\tilde{\Gamma}_j(t_m)$ and $\tilde{\Gamma}^*_j(t_m)$ are the complex signals at axial time t_m corresponding to the jth A-scan and its respective conjugate; $\tilde{\Gamma}_{j+1}(t_m)$ and $\tilde{\Gamma}^*_{j+1}(t_m)$ are the complex signals at axial time t_m corresponding to the next A-scan and its respective conjugate; M is an even number that denotes the window size in the axial direction for each pixel; N is the number of sequential scans used to calculate the cross correlation; and T is the time duration between A-scans. Because T is much longer than the pixel time window within each scan used in the spectrogram method, high-velocity sensitivity can be achieved.

Phase-resolved ODT decouples spatial resolution and velocity sensitivity in flow images and increases imaging speed by more than two orders of magnitude, without compromising spatial resolution and velocity sensitivity. In addition, because two sequential A-line scans are compared at the same location, speckle modulations in the fringe signals cancel each other and, therefore, will not affect the phase difference calculation. Consequently, the phase-resolved method reduces the speckle noise in the velocity image. Furthermore, if the phase difference between sequential frames is used, the velocity sensitivity can be further increased. Real-time imaging with velocity sensitivity on the order of 10 µm/s has been demonstrated.

In addition to digital processing of the fringe signal using Hilbert transformation, the complex analytical signal can also be achieved through hardware implementation. Optical Hilbert transformation using polarization optics has been implemented for real-time phase-resolved ODT imaging. Real-time ODT imaging using hardware demodulation of the ODT signal has also been demonstrated.

Spectral Domain Phase-Resolved ODT Method

A schematic diagram of a spectral domain ODT system is shown in Figure 6. The signal from the Michelson interferometer is directly coupled to a

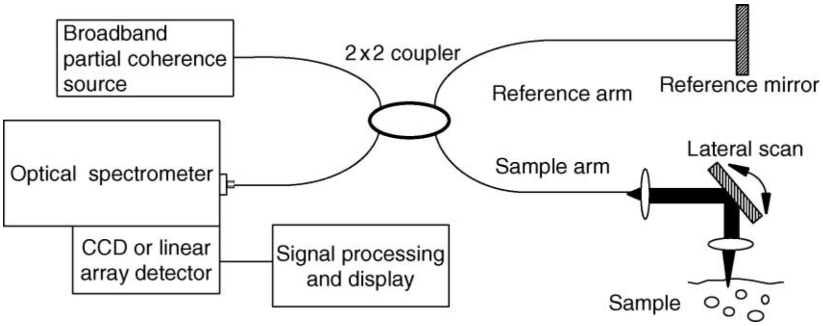

Figure 6 Schematic diagram of a spectral domain ODT instrument.

spectrometer that records the spectral fringe pattern, $P_{\mathrm{ODT}}(\nu)$. The temporal interference fringe can be calculated by a Fourier transformation of the spectral fringe pattern (eqn [17]). The Doppler shift can then be determined from the phase shift between sequential scans using the phase-resolved ODT algorithm (eqns [23] and [24]).

Spectral domain phase-resolved ODT has the advantage that no optical delay line is required. In addition, the spectral domain method can achieve high signal to noise ratio over time domain ODT. Since parallel detection can be implemented with a high-speed spectrometer, high-speed imaging acquisition is possible. Finally, because the dynamic range of the phase-resolved ODT depends on the speed of the line scans, spectral domain ODT has the advantage in terms of imaging speed and velocity dynamic range.

Transverse Flow Velocity and Doppler Angle Determination

One of the limitations of using the Doppler shift to determine the flow is that the technique is only sensitive to longitudinal flow velocity (Figure 2). If one knows the flow direction, Doppler shift measurement can fully quantify the flow. However, in many biological cases, where flow direction is not known, Doppler shift measurement alone is not enough to fully quantify the flow. Furthermore, there are many clinical cases, such as ocular bloodflow, where vessels are in the plane perpendicular to the probing beam. When flow direction is perpendicular to the probing beam, the Doppler shift is not sensitive to transverse bloodflow (V_{T}). Therefore, a method to measure transverse flow velocity is essential.

Standard deviation of the Doppler spectra can be used to determine the transverse flow. The technique is based on the fact that ODT imaging uses a relatively large numeric aperture lens in the sampling arm. The beam from different sides of the edges will produce different Doppler shifts, f_1 and f_2, as indicated in Figure 7. Consequently, the Doppler spectra will be broadened by the transverse flow. If the incident beam has a Gaussian spectral profile and contributions from Brownian motion and other sources that are independent of the macroscopic flow velocity are included, a linear relation between standard deviation of the Doppler spectra and transverse flow velocity, $V_{\mathrm{T}} = V \sin\theta$, can be derived:

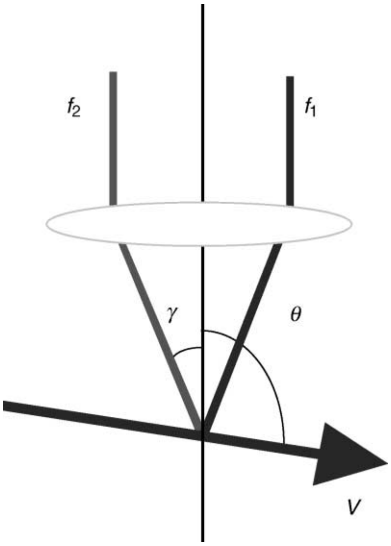

Figure 7 Effect of numerical aperture and transverse flow velocity on Doppler bandwidth.

$$\sigma = \frac{\pi V \sin\theta NA_{\mathrm{eff}}}{8\lambda} + b \qquad [25]$$

where b is a constant, and NA_{eff} is the effective numeric aperture. The standard deviation can be determined from the measured analytical fringe signal:

$$\sigma^2 = \frac{1}{2\pi^2 T^2} \times \left(1 - \frac{\left| \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} \tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m) \right|}{\frac{1}{2} \sum_{m=(n-1)M}^{nM} \sum_{j=1}^{N} [\tilde{\Gamma}_j(t_m)\cdot\tilde{\Gamma}^*_j(t_m) + \tilde{\Gamma}_{j+1}(t_m)\cdot\tilde{\Gamma}^*_{j+1}(t_m)]} \right) \quad [26]$$

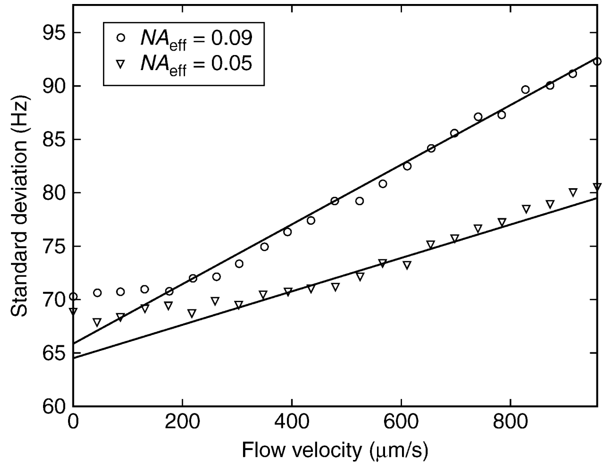

Figure 8 Standard deviation as a function of flow velocity for two different numeric apertures. Reproduced with permission from Ren *et al.* (2002) *Optics Letters* 27: 409–411.

The measured standard deviation as a function of transverse flow velocity is shown in Figure 8. As predicted, the Doppler bandwidth is a linear function of transverse flow velocity above a certain threshold level. The effective numerical aperture of the optical objective in the sample arm determines the slope of this dependence. This result indicates that standard deviation can be used to determine the transverse flow velocity. Since both longitudinal and transverse flow velocity (V_L and V_T) can be measured by the Doppler shift and standard deviation, respectively, flow direction can be determined from a single measurement of the Doppler fringe signal.

Applications

Due to its exceptionally high spatial resolution and velocity sensitivity, several clinical applications of ODT have been demonstrated, including screening vasoactive drugs, monitoring changes in image tissue morphology and hemodynamics following pharmacological intervention and photodynamic therapy, evaluating the efficacy of laser treatment in port wine stain (PWS) patients, assessing the depth of burn wounds, mapping cortical hemodynamics for brain research, imaging ocular bloodflow, and

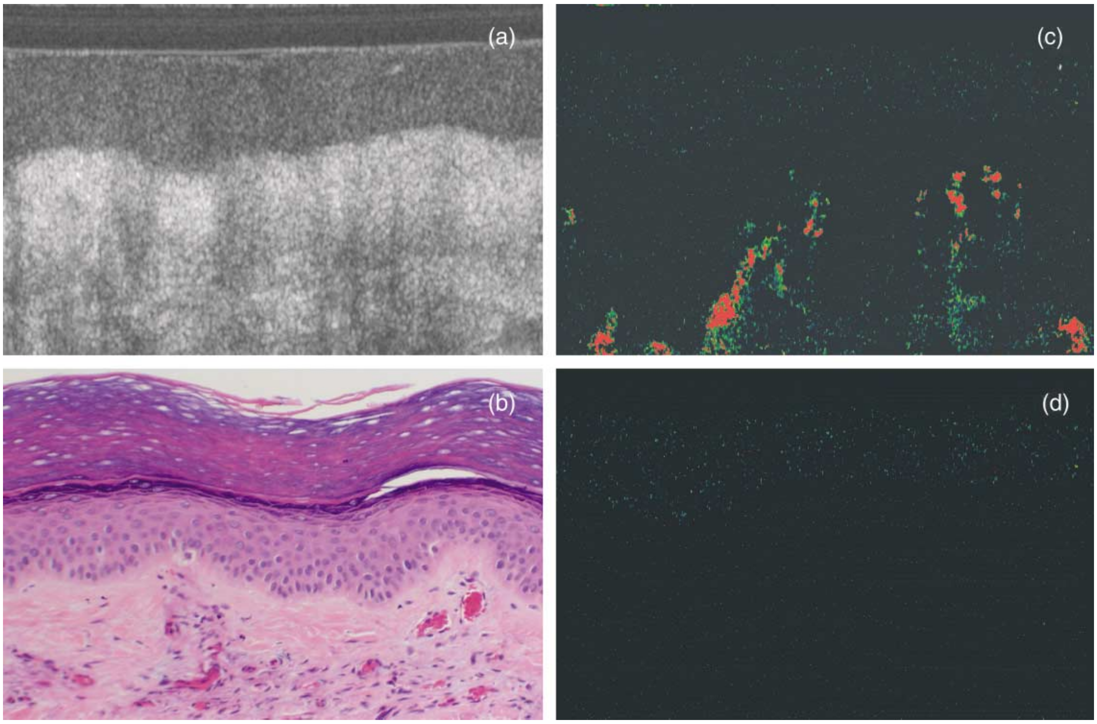

Figure 9 ODT images taken *in situ* from PWS human skin. (a) Structural image, (b) histological section, (c) image before laser treatment, and (d) image after laser treatment (scale bar 100 μm). Reproduced with permission from Nelson *et al.* (2001) *Archives of Dermatology* 137: 741–744. Copyright © 2001, American Medical Association. All rights reserved.

mapping bloodflow in gastrointestinal tracts. Figure 9 shows ODT structural and flow velocity images of a patient with PWS before and after laser treatment, respectively. For comparison, a histology picture taken at the same site is included. The vessel location from the ODT measurement and histology agree very well. Furthermore, the destruction of the vessel by laser can be identified, since no flow appears on the Doppler flow image after laser treatment.

Three-dimensional images of a microvascular network can be obtained by stacking 2D scans together. It is known that the microvasculature of mammary tumors has several distinct differences from normal tissues. Mapping of a 3D cancer microvascular network could provide additional information for cancer diagnosis. Figure 10 shows multiple blood vessels imaged from a patient with an abnormal microvascular network. Different colors represent different signs of the Doppler shift, which depends on the angle between the direction of flow and probing beam.

ODT is also a powerful tool to study flow dynamics in microfluidic channels. In addition to imaging flow velocity profile, ODT can also be used to measure osmotic mobility, quantify size of the scattering particle, and study flow dynamics of microfluids in microchannels of different materials, geometry, and surface treatment.

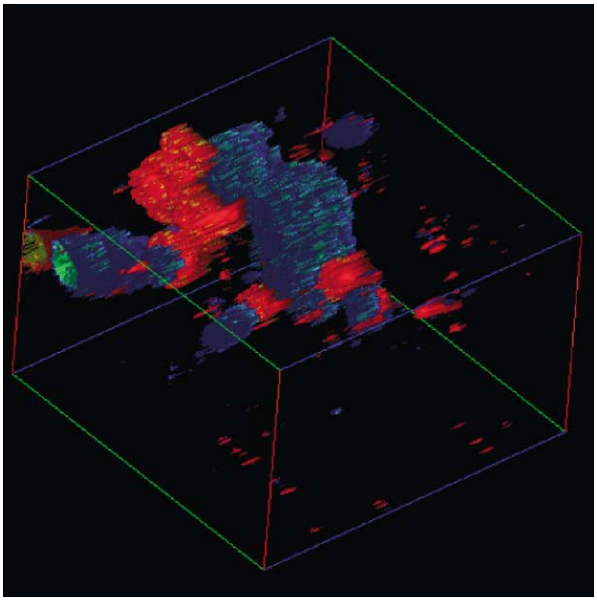

Figure 10 Three-dimensional ODT images of multiple blood vessels in human skin from a patient with a PWS birthmark. Reproduced with permission from Zhao *et al.* (2001) *IEEE J. Select Topics Quantum Electro.* 7: 931–935. © 2001, IEEE.

Polarization Sensitive OCT

PS-OCT combines polarization sensitive detection with OCT to determine tissue birefringence. PS-OCT can obtain enhanced image contrast and additional physiological information by studying polarization properties of biological tissues. Many biological tissues, such as tendon, muscle, nerve, bone, cartilage, and skin, exhibit birefringence properties. Consequently, PS-OCT can provide contrast mechanisms that are directly related to the physiological conditions of these tissues.

The polarization state of a light beam can be described by a Stokes vector **S** as:

$$\mathbf{S} = \begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix} = \begin{bmatrix} a_x^2 + a_y^2 \\ a_x^2 - a_y^2 \\ 2a_x a_y \cos\varphi \\ 2a_x a_y \sin\varphi \end{bmatrix} \qquad [27]$$

where I, Q, U, V are Stokes parameters, a_x and a_y are amplitudes of two orthogonal components of the electric vector, and φ represents the phase difference between the two components.

The effect of an optical device on the polarization of light can be characterized by a 4×4 Mueller matrix. The matrix acts on the input state, $\mathbf{S}_1$, to give the output state, $\mathbf{S}_2$:

$$\mathbf{S}_2 = \begin{bmatrix} I_2 \\ Q_2 \\ U_2 \\ V_2 \end{bmatrix} = \mathbf{MS}_1$$

$$= \begin{bmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ M_{10} & M_{11} & M_{12} & M_{13} \\ M_{20} & M_{21} & M_{22} & M_{23} \\ M_{30} & M_{31} & M_{32} & M_{33} \end{bmatrix} \begin{bmatrix} I_1 \\ Q_1 \\ U_1 \\ V_1 \end{bmatrix} \qquad [28]$$

In the Poincaré sphere representation, a polarization state, **S**, can be represented by a point (Q, U, V). The Poincaré sphere representation provides a convenient method to evaluate changes of Stokes vectors. Determination of either the Stokes vector or the Mueller matrix allows full quantification of birefringence properties of the biological sample.

A fiber-based, high-speed PS-OCT system is shown in Figure 11. The system is similar to the ODT system, except a polarization modulator and two orthogonal detection channels are included. Due to the coherence detection intrinsic to OCT, backscattered light that

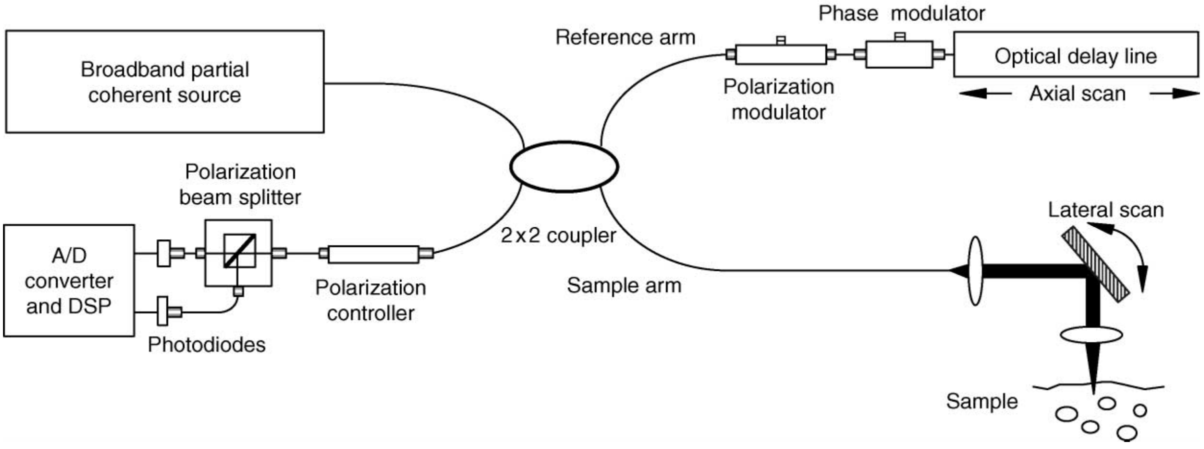

Figure 11 Fiber-based PS-OCT system. DSP: digital signal processing.

originates from incident light in different polarization states can be determined by controlling polarization states in the reference arm. Modulating the reference light in four orthogonal polarization states allows coherent detection of backscattered light from the sample under equivalent illumination in four different polarization states. The choice of orthogonal polarization states in the Poincaré sphere is important because it ensures that the birefringence measurements will be independent of the orientation of the optical axis in the sample.

For every polarization state controlled by the polarization modulator, the A-scan signals corresponding to the two orthogonal polarization diversity channels are digitized. The coherence matrix can be calculated from the complex electrical field vector from these two channels:

$$J = \begin{bmatrix} \langle E_{\mathrm{H}}^*(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{H}}^*(t)E_{\mathrm{V}}(t)\rangle \\ \langle E_{\mathrm{V}}^*(t)E_{\mathrm{H}}(t)\rangle & \langle E_{\mathrm{V}}^*(t)E_{\mathrm{V}}(t)\rangle \end{bmatrix} = \begin{bmatrix} J_{\mathrm{HH}} & J_{\mathrm{HV}} \\ J_{\mathrm{VH}} & J_{\mathrm{VV}} \end{bmatrix} \qquad [29]$$

where E_{H}, E_{V} are the components of the complex electric field vector corresponding to horizontal and vertical polarization channels, respectively, and E_{H}^*, E_{V}^* are their conjugates, respectively. The Stokes vector can be derived from the coherence matrix:

$$\begin{aligned} S_0 &= J_{\mathrm{HH}} + J_{\mathrm{VV}} & S_2 &= S_0^{-1}(J_{\mathrm{HV}} + J_{\mathrm{VH}}) \\ S_1 &= S_0^{-1}(J_{\mathrm{HH}} - J_{\mathrm{VV}}) & S_3 &= iS_0^{-1}(J_{\mathrm{VH}} - J_{\mathrm{HV}}) \end{aligned} \qquad [30]$$

where S_1, S_2, and S_3 are the normalized coordinates of the Stokes vector in the Poincaré sphere, characterizing the polarization state of the backscattered light, and S_0 is the module of the Stokes vector characterizing light intensity. The Stokes vector, corresponding to each of the input polarization states for the *n*th pixel in the *j*th A-scan, can be calculated as:

$$\begin{aligned} S_{0,n} &= \sum_{m=(n-1)M}^{nM} [\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) + \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)] \\ S_{1,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} [\tilde{\Gamma}_j^{\mathrm{H}}(t_m)\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m) - \tilde{\Gamma}_j^{\mathrm{V}}(t_m)\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)] \\ S_{2,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Re}(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{\mathrm{V}}(t_m)) \\ S_{3,n} &= S_{0,n}^{-1} \sum_{m=(n-1)M}^{nM} 2\,\mathrm{Im}(\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)\tilde{\Gamma}_j^{V}(t_m)) \end{aligned} \qquad [31]$$

where M is an integer number that determines the size of the average in the axial direction for each pixel; $\tilde{\Gamma}_j^{\mathrm{H}}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}}(t_m)$ are complex signals detected from the two orthogonal polarization channels at axial time t_m, for the *j*th A-scan; and $\tilde{\Gamma}_j^{\mathrm{H}^*}(t_m)$ and $\tilde{\Gamma}_j^{\mathrm{V}^*}(t_m)$ are their conjugates, respectively.

In order to measure the birefringence properties of the sample accurately, four states of light polarization are generated for each lateral pixel. For each polarization state, one A-line scan is performed. Therefore, a total of four A-line scans are used to calculate the Stokes vector and phase retardation images simultaneously.

For a sample with an assumed linear birefringence, there exists two eigenwaves that are polarized along the projected fast and slow axes of the sample normal to the propagation direction of incident light. The Stokes vectors of these eigenwaves determine a rotation axis in the equator plane of a Poincaré sphere.

The effect of birefringence is to rotate the Stokes vector about this axis through an angle that is equal to the phase retardation of the sample. Conversely, the rotation axis can be determined from the known polarization states of incident and backscattered light at the sample location. The phase retardation image, which characterizes the accumulated birefringence distribution in the sample, is calculated by the rotation of the Stokes vectors in the Poincaré sphere.

The simultaneous Stokes vectors and phase retardation images of a rat muscle are shown in Figure 12. Stokes vectors, corresponding to four different polarization states, are shown in the top figure. From Stokes vector images, the phase retardation image can be calculated as shown in the bottom figure. The banded structure in the phase retardation image indicates the accumulated tissue birefringence distribution in the rat muscle. In addition, differential optical axis and birefringence images can also be calculated.

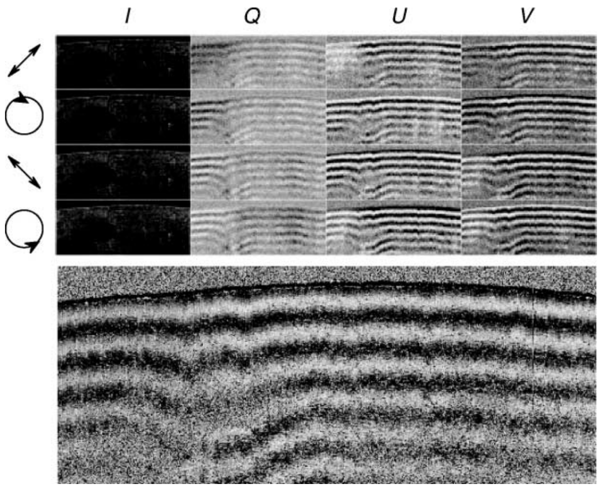

Figure 12 Stokes vectors images (top) and phase retardation image (bottom) in fresh rat muscle. The image area is 2 mm × 5 mm. Reproduced with permission from Ren *et al.* (2003) *SPIE Proceedings* 4956: 320–327.

There are a number of clinical applications for PS-OCT, including burn depth determination and retinal fiber nerve layer thickness evaluation. Figure 13 shows *in vivo* structural and phase retardation images from rat skin burned at 100°C for 10 seconds. Although the structure image shows very little contrast, the phase retardation image clearly shows the contrast between the normal and burned regions. As light propagates into the tissue, the retardation increases due to tissue birefringence. For normal tissue, it only takes less than 20 μm to cause a phase retardation of 90°. However, for the burned tissue, it takes much thicker tissue to cause the same phase retardation. The decrease of birefringence in thermally damaged tissue can clearly be seen in PS-OCT scans in comparison to normal rat skin.

Second Harmonic OCT

SH-OCT combines SHG with coherence gating for high-resolution tomographic imaging of tissue structures with molecular contrast. SHG is the lowest-order nonlinear optical process where the second-order nonlinear optical susceptibility is

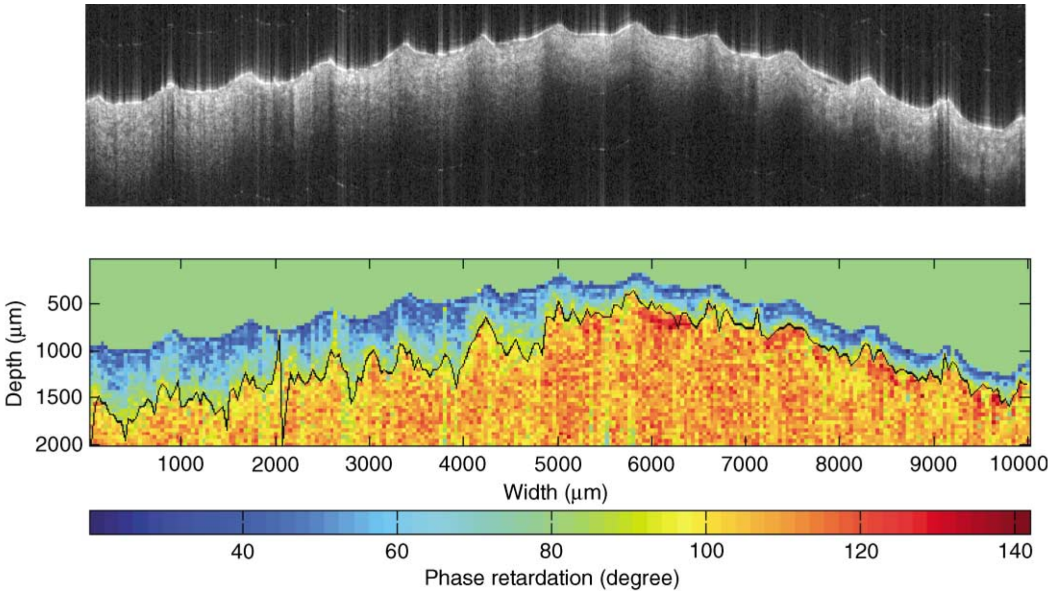

Figure 13 *In vivo* structure (top) and phase retardation image (bottom) from rat skin burned at 100°C for 10 seconds. The black contour line in the phase retardation image demarcates the depth at which 90° phase retardation has been reached with respect to the incident polarization (image size 2 mm × 10 mm).

responsible for the generation of light at second-harmonic frequency. Because second-order nonlinear optical susceptibility is very sensitive to electronic configurations, molecular structure and symmetry, local morphology, and ultrastructures, SHG provides molecular contrast for the coherence image.

The schematic diagram for SH-OCT is shown in Figure 14. A broadband short pulse laser is used as a light source in a Michelson interferometer. The light is split into the reference arm and the sample arm. In the reference arm, a thin nonlinear crystal is used to convert the input radiation to SH photons. Both SH and fundamental waves are then reflected by a metal mirror (M1) mounted on a motorized translation stage, which acts as the delay line in this SH-OCT system. A backscattered SH wave from the sample recombines with the SH wave from the reference arms to form interference fringe. The SH interference fringe signal is detected by a photomultiplier tube (PMT) after passing through a short-pass filter (F1). The fundamental interference fringe signal is detected by a photodiode (PD) after passing through a long-pass filter (F2).

A SH-OCT image of a rat tendon is shown in Figure 15. Compared with conventional OCT performed at fundamental wavelengths, SH-OCT offers enhanced molecular contrast and spatial resolution. It is also an improvement over existing SHG scanning microscopy technology as the intrinsic coherence gating mechanism enables the detection and discrimination of SH signals generated at deeper locations. The enhanced molecular contrast of SH-OCT extends conventional OCT's capability for detecting small changes in molecular structure. SH-OCT is promising for the diagnosis of cancers and other diseases at an early stage when changes in tissue and molecular structure are small.

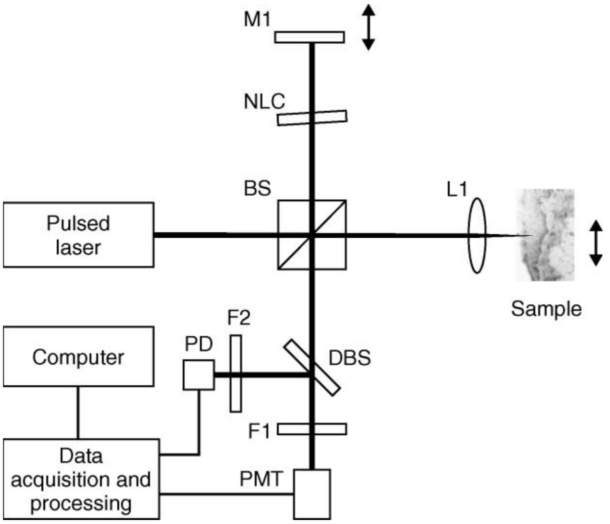

Figure 14 Schematic of experiment set-up for SH-OCT. M1, mirror; NLC, nonlinear crystal; BS, broadband nonpolarization beamsplitter; DBS, dichroic beamsplitter; L1, lens; F1–F2, filters; PD, photo diode; PMT, photomultiplier tube.

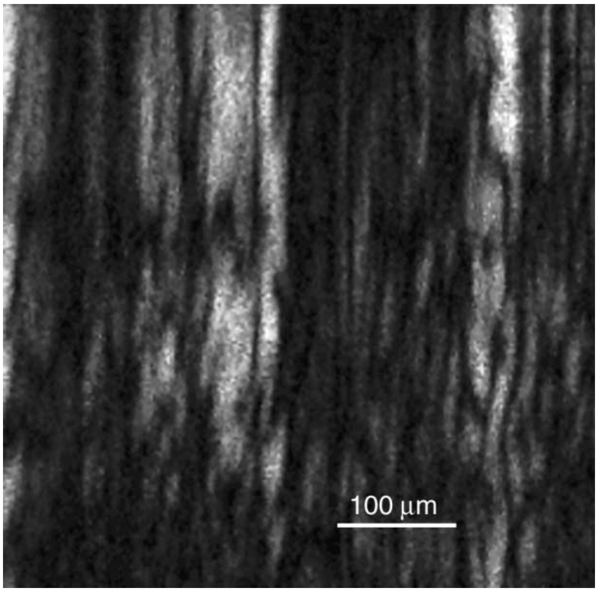

Figure 15 SH-OCT image of a rat tendon; image size 250 μm × 250 μm. Scale bar 100 μm. Image reprinted with permission from Jiang Y *et al.* (2004) Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Conclusions

Optical tomography is a rapidly developing imaging technology with many potential applications. Through different contrast enhancement mechanisms, ODT, PS-OCT, and SH-OCT provide clinically important physiological information that is not available in the structure image. Given the noninvasive nature and exceptionally high spatial resolution, optical tomography can simultaneously provide tissue structure, blood perfusion, birefringence, and other physiological information and have great potential for basic biomedical research and clinical medicine.

Acknowledgments

I would like to thank many of my colleagues who have contributed to the functional OCT project at the Beckman Laser Institute and Center for Biomedical Engineering at UCI, particularly the students and postdoctoral fellows. I also want to acknowledge the research grants awarded from the National Institutes of Health (EB-00293, NCI-91717, RR-01192 and EB-00255), the National Science Foundation (BES-86924), the Whitaker

Foundation (WF-23281) and the Defense Advanced Research Program Agency (Bioflip program). Institutional support from the Air Force Office of Scientific Research (F49620-00-1-0371), and the Beckman Laser Institute Endowment are also gratefully acknowledged. Dr Chen's e-mail address is zchen@bli.uci.edu.

List of Units and Nomenclature

OCT	Optical coherence tomography
ODT	Optical Doppler tomography
PS-OCT	Polarization sensitive OCT
SH-OCT	Second harmonic OCT
STFFT	Short time fast Fourier transformation

See also

Coherence: Coherence and Imaging; Overview. **Fiber and Guided Wave Optics:** Nonlinear Optics; Optical Fiber Cables. **Imaging:** Imaging Through Scattering Media; Interferometric Imaging; Photon Density Wave Imaging. **Interferometry:** Overview. **Microscopy:** Confocal Microscopy; Imaging Multiple Photon Fluorescence Microscopy. **Polarization:** Introduction; Matrix Analysis. **Tomography:** Optical Coherence Tomography; Tomography and Optical Imaging.

Further Reading

Bouma BE and Tearney GJ (eds) (2001) *The Optical Coherence Tomography Handbook*. New York: Marcel Dekker, Inc.

Chen Z (2003) Functional optical coherence tomograpy. In: Hwang NHC and Woo LY (eds) *Frontiers in Biomedical Engineering*, pp. 345–364. New York: Kluwer Academic Publishers.

Chen Z, Milner TE, Srinivas S, Wang XJ, Malekafzali A, van Gemert MJC and Nelson JS (1997) Noninvasive imaging of *in vivo* blood flow velocity using optical doppler tomography. *Optics Letters* 22: 1119–1121.

Chen Z, Zhao Y, Srinivas SM, Nelson S, Prakash N and Frostig RD (1999) Optical Doppler tomography. *IEEE Journal of Selected Topics in Quantum Electronics* 5(4): 1134–1142.

de Boer JF and Milner TE (2002) Review of polarization sensitive optical coherence tomography and Stokes vector determination. *Journal of Biomedical Optics* 7: 359–371.

Fercher AF and Hizenberger CK (2002) Optical coherence tomography. *Progress in Optics* 44: 215–296.

Huang D, Swanson EA, Lin CP, *et al.* (1991) Optical coherence tomography. *Science* 254: 1178–1181.

Izatt JA, Kulkarni MD, Yazdanfar S, Barton JK and Welch AJ (1997) *In vivo* bidirectional color Doppler flow imaging of picoliter blood volumes using optical coherence tomography. *Optics Letters* 22: 1439–1441.

Jiang Y, Tomov I, Wang Y and Chen Z (2004) Second harmonic optical coherence tomography. *Optics Letters* 29: 1090–1092.

Jiao S and Wang LV (2002) Jones-matrix imaging of biological tissues with quadruple-channel optical coherence tomography. *Journal of Biomedical Optics* 7: 350–358.

Masters BR and Thompson BJ (eds) (2001) *Selected Papers on Optical Low-Coherence Reflectometry & Tomography*. SPIE Milestone Series, vol. MS 165. Bellingham, WA: SPIE Optical Engineering Press.

Tuchin VV (ed.) (2004) *Coherent-Domain Optical Methods for Biomedical Diagnostics, Environmental and Material Science*, vol. 2. Boston, MA: Kluwer Academic Publishers.

Vo-Dinh T (ed.) (2003) *Biomedical Photonics Handbook*. New York: CRC Press.

ULTRAFAST LASER TECHNIQUES

Contents

Generation of Femtosecond Pulses
Pulse Characterization Techniques

Generation of Femtosecond Pulses

D T Reid, Heriot-Watt University, Edinburgh, UK

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Optical pulses from femtosecond lasers represent some of the shortest physical events ever observed and their high instantaneous powers allow them to be focused to intensities that can heat and compress matter to resemble the extreme conditions found in a nuclear explosion. The extreme units of time and power needed to describe this area of physics can easily distract from the simple and elegant techniques by which femtosecond pulses can be generated. The review presented here is divided into two main sections. The first explains the theory of pulse generation and shaping and the second concentrates on actual femtosecond laser sources in common use today.

Theory of Femtosecond Pulse Generation

Active Mode-locking

In an inhomogeneously broadened laser system, optical frequencies within the emission bandwidth of the material can experience gain simultaneously at discrete frequencies corresponding to the longitudinal modes of the laser resonator. In the absence of any mechanism to coherently couple energy between these modes, the laser will operate with a continuous-wave (cw) output containing a narrow band of frequencies whose roundtrip cavity gain is highest. An alternative operating regime exists in which each mode is related to its neighbor by a fixed phase relationship and it is this mode-locked condition that is used to generate a periodic sequence of ultrashort optical pulses.

To understand how mode-locking works, consider the frequency separation between adjacent longitudinal resonator modes which can be expressed as

$$\Delta\omega = 2\pi c/l \qquad [1]$$

where l is the optical roundtrip length of the resonator and c is the speed of light. If the nth longitudinal mode has an amplitude E_n then the total optical field can be denoted in complex notation as

$$E(t) = \sum_n E_n \exp \mathrm{i}[(\omega_0 + n\Delta\omega)t + n\Delta\varphi] \qquad [2]$$

with ω_0 being the center frequency of the output and $\Delta\varphi$ being the phase difference between adjacent modes. Figure 1 illustrates how a simple pulse can be produced by coherently adding together adjacent modes. In Figure 1a, three modes with equal amplitudes and the same initial phase ($\Delta\varphi = 0$) are interfered to give a pulse. As further modes are added (Figure 1b) the pulse duration decreases.

Coupling of energy between adjacent cavity modes can be achieved in practice by modulating either the

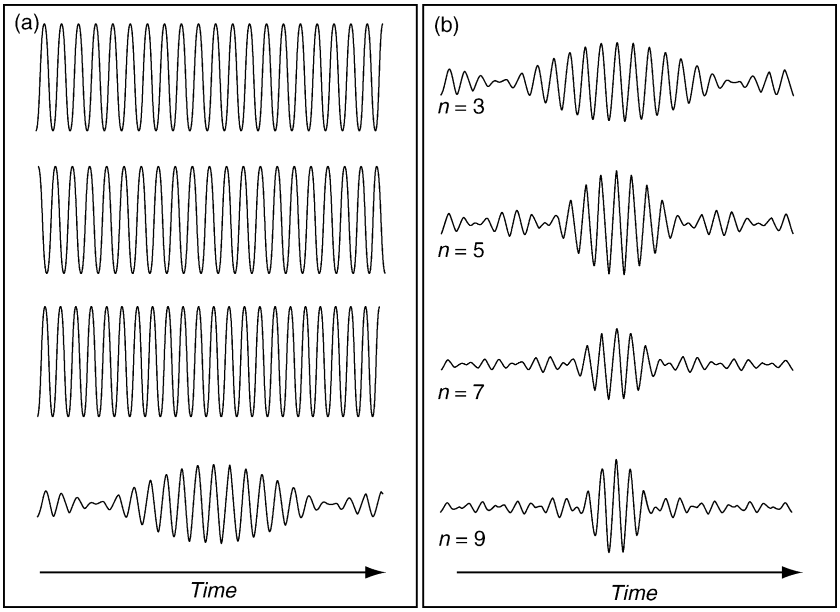

Figure 1 The principles of mode-locking: (a) three adjacent modes add coherently to produce a pulse; (b) shorter pulses require a larger number of *n* coupled modes.

intracavity loss (amplitude modulation (AM) mode-locking) or the intracavity phase (frequency modulation (FM) mode-locking). A common practical amplitude modulator is an acousto-optic modulator and for operation in the visible and near-infrared, fused silica is a common acousto-optic modulator material. Lithium-niobate devices based on the Pockels effect are the most commonly used phase modulators and are attractive for mode-locked fiber-laser applications because fiber-pigtailed waveguide devices are readily available.

Active mode-locking has been employed in a wide variety of laser systems including argon ion, dye, erbium-doped fiber, semiconductor, and solid-state lasers. Unlike passive mode-locking, no optical nonlinearity is required which means that actively mode-locked lasers can be operated with low intracavity powers. The technique is particularly suitable for mode-locked sources which require active synchronization to an external clock source.

A form of 'optical' active mode-locking exists in which gain modulation resulting from optical pumping using another mode-locked laser is the mode-locking mechanism. Known as synchronously pumped mode-locking, this technique has been used to obtain ultrashort pulses from oscillators incorporating gain media of laser or nonlinear optical materials. The technique is most appropriate to laser gain materials which have a short-lived fluorescence lifetime and whose gain therefore exhibits a significant change when the pump light is modulated at high frequencies. Synchronous pumping has been used to generate mode-locked pulses from color-center lasers, dye lasers, and optical parametric oscillators.

Passive Mode-locking

To produce the shortest optical pulses, passive mode-locking, a technique based on exploiting intracavity optical nonlinearities, must be employed. Unlike the active technique, the strength of the mode-locking action increases as the propagating pulses shorten and their peak intensities rise. One consequence of this nonlinear behavior is that the passive technique commonly generates and mode-locks entirely new cavity frequencies which were not originally present during cw operation.

Passive mode-locking has been implemented using a wide range of different nonlinear effects, but in every technique the mode-locking process can be described in terms of dynamic loss saturation combined, in certain systems, with dynamic gain saturation. The principal mode-locking element is a saturable absorber whose loss decreases as the incident pulse intensity increases. A saturable absorber can take the form either of a resonant component such as a solid-state semiconductor device or a dye jet with a fixed absorption wavelength, or a nonresonant component that replicates a saturable-absorber-like action using self-focusing or self-phase-modulation effects. Saturable absorbers can be categorized as either slow saturable absorbers or fast saturable absorbers, depending on whether their recovery time is longer or shorter than the pulse duration, respectively. The pulse-shaping mechanisms are different in each case and will be described separately.

Slow saturable absorbers are physical devices, commonly dye molecules in solution or semiconductor-doped glass, which rely on resonant excitation to produce an excited electronic state whose transmission is greater than that of the ground state. When an optical pulse is incident on a slow absorber, its leading edge is absorbed and creates an excited state which is relatively transparent to the trailing edge. In a laser medium whose gain is readily saturated, a pulse propagating first through the absorber and then through the gain medium experiences a 'gain window' with a finite duration which strongly shapes the pulse (see Figure 2a,b).

A fast saturable absorber is required to realize mode-locking in laser media which are not easily saturated. The time-dependent gain and loss profiles of the absorber and the gain medium are again depicted in Figure 2c,d and refer to an absorber with both a rapid response and recovery time. In this system, the dynamics of the absorber and gain medium no longer represent the limiting factor influencing the pulse duration, and other effects such as linear and nonlinear dispersion become the dominant pulse-shaping mechanisms.

Pulse Shaping by Material Dispersion

Intense pulses propagating in a mode-locked laser experience strong linear and nonlinear shaping effects associated, respectively, with group-velocity dispersion and self-phase-modulation. Contributions to the net intracavity dispersion arise from the gain medium, the cavity mirrors, other elements such as intracavity prisms, even the air in the beam path. Dispersive effects in materials originate as a result of a frequency-dependent dielectric susceptibility:

$$P(\omega) = \varepsilon_0 \chi^{(1)}(\omega)E(\omega) + \varepsilon_0 \chi^{(2)}(\omega)E(\omega)^2 + \varepsilon_0 \chi^{(3)}(\omega)E(\omega)^3 + \cdots \qquad [3]$$

which is due to physical resonances (poles) in the electronic response to an applied E-field.

Linear dispersion

Linear dispersion refers to the variation of the refractive index associated with the real part of the $\chi^{(1)}$ term, namely,

$$n_0 = \sqrt{1 + \mathrm{Re}\{\chi^{(1)}\}} \qquad [4]$$

and leads to a wavelength-dependent group velocity which is responsible for pulse broadening and break-up in optical systems. Dispersive effects are best analyzed in terms of optical phase which is related to the refractive index by

$$\varphi(\omega) = \frac{n_0(\omega)\omega L}{c} \qquad [5]$$

where L is the medium length and c is the vacuum speed of light. The local variation of the spectral

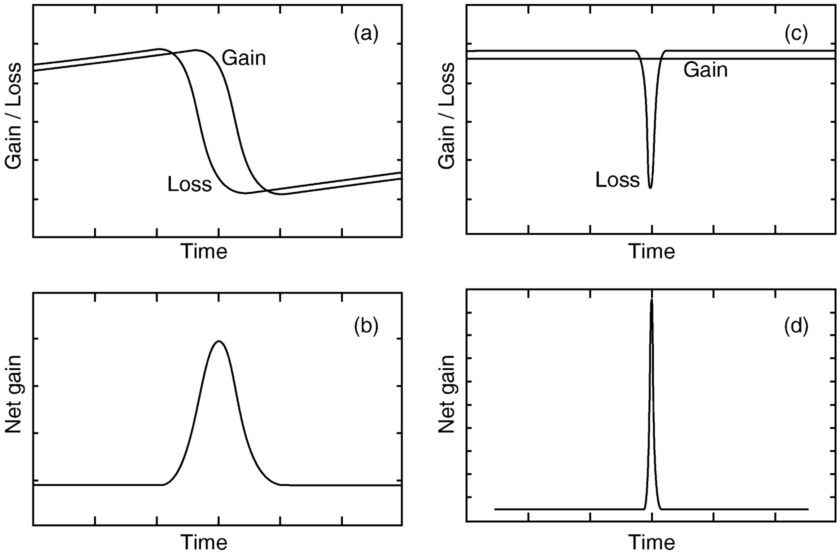

Figure 2 Saturable absorber action: (a) intracavity gain and loss profiles for a laser comprising a slow saturable absorber and a rapid relaxation time gain medium, and (b) the net gain (gain − loss) for this system; (c) gain and loss profiles for a laser comprising a near-instantaneous saturable absorber and a slow relaxation time gain medium, and (d) the net gain for this system.

phase due to an optical medium can be represented as the Taylor series

$$\varphi(\omega) = \varphi(\omega_0) + (\omega - \omega_0)\frac{\mathrm{d}\varphi}{\mathrm{d}\omega}\bigg|_{\omega_0} + \frac{1}{2}(\omega - \omega_0)^2\frac{\mathrm{d}^2\varphi}{\mathrm{d}\omega^2}\bigg|_{\omega_0} + \left|\frac{1}{6}(\omega - \omega_0)^3\frac{\mathrm{d}^3\varphi}{\mathrm{d}\omega^3}\right\|_{\omega_0} + \cdots \qquad [6]$$

where ω_0 is the center frequency of the incident light. When an ultrashort pulse propagates through the medium, only the quadratic and higher-order terms in [6] have a direct influence on the shape of the pulse in time. The $(\omega - \omega_0)^2$ term is responsible for adding linear chirp which refers to a time-varying linear increase or decrease in the instantaneous optical frequency across the pulse. Obtaining the shortest pulses from a mode-locked laser requires the quadratic and higher-order terms in the roundtrip intracavity spectral phase to be zero. In practice this means minimizing the material group-velocity dispersion of the gain medium by using a short but highly doped laser crystal and then identifying additional optical elements whose group-velocity dispersion is equal in magnitude but opposite in sign to that of the crystal. Optical components for compensating group-velocity dispersion fall into two categories: those which use bulk optics and geometry to achieve a wavelength-dependent path length, and others based on interference effects within a dielectric optical coating.

Geometrical methods of dispersion compensation are commonly based on pairs of diffraction gratings or prisms and in both cases the geometry of the systems means that the group delay is wavelength dependent and can be made greater for longer wavelengths than for shorter ones. Analytic expressions describing the quadratic and cubic spectral phase of prism and diffraction grating pairs are summarized in Table 1. Dispersion compensation using prisms has been extremely successful and 11-fs pulses were reported using this approach. The simultaneous compensation of arbitrary second- and third-order spectral phase is, however, difficult and involves identifying a suitable prism material, apex spacing, and tip insertion. The expressions listed in Table 1 indicate that the form of the spectral phase depends on the refractive index of the prism material and its wavelength derivatives, while the apex separation, l, also contributes to a large extent. A further contribution to the spectral phase comes from the glass encountered by the beam passing through the prism tips which contributes a positive dispersion described by the equations for a bulk material given in Table 1. In general, an iterative solution of the equations given in Table 1 is needed to determine the geometry of a prism pair suitable for producing a desired spectral phase profile.

Laser cavity mirrors can be designed with chirped multilayer dielectric coatings which possess a chosen group-velocity dispersion profile. A Bragg mirror with smaller layer periods at the surface and larger periods deeper in the coating provides negative group-velocity dispersion because the longer wavelengths experience larger group delays within the coating than the shorter ones. These chirped mirrors provide constant reflectivity across their stopband and therefore only modify the spectral phase of a pulse, leaving the intensity unchanged. The design of chirped mirrors has been refined in recent years to enhance their operating bandwidth and to reduce undesirable modulations in the group-delay response. The technique has been used to obtain 6.5-fs pulses directly from a Ti:sapphire oscillator and to achieve compression of amplified pulses to below 5 fs.

Nonlinear dispersion

Nonlinear effects arise in all materials due to the $\chi^{(3)}$ term in [3]. In a transparent medium the refractive index due to this net susceptibility is approximately

$$n \approx n_0 + \frac{\chi^{(3)}E^2}{2n_0} = n_0 + n_2 I \qquad [7]$$

where I is the optical intensity. The importance of the $\chi^{(3)}$ susceptibility is therefore that it causes the refractive index to become intensity-dependent, and significant changes in the refractive index can be induced by the high peak intensities of ultrashort pulses.

Self-phase-modulation (SPM) resulting from the nonlinear refractive index is essential in generating the shortest pulses from modern ultrafast lasers and is caused by the time-varying intensity profile of the pulse. The higher-intensity components induce a larger nonlinear refractive index change and therefore experience a greater phase shift than the weaker components. The resulting frequency shift across the pulse can be shown to be

$$\Omega(t) = -\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = -\frac{\omega n_2 L}{c}\frac{\mathrm{d}I(t)}{\mathrm{d}t} \qquad [8]$$

which describes a redshift of the pulse leading edge and a blueshift of the trailing edge. Only the pulse center frequency remains unchanged and SPM therefore broadens the pulse spectrum by redistributing energy from the center to the wings. Across the center of the pulse, SPM leads to an approximately linear positive chirp that, with appropriate optics, can be

Table 1 Expressions for the second and third-order dispersion of common optical systems

Bulk material, thickness t 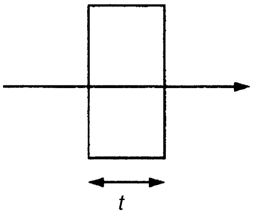	$\frac{d^2\varphi}{d\omega^2} = t\frac{\lambda^3}{2\pi c^2}\frac{d^2n}{d\lambda^2}$ $\frac{d^3\varphi}{d\omega^3} = -t\frac{\lambda^4}{4\pi^2c^3}\left(3\frac{d^2n}{d\lambda^2} + \lambda\frac{d^3n}{d\lambda^3}\right)$
Double-pass prism-pair, apex separation l, angular deviation within ray of β 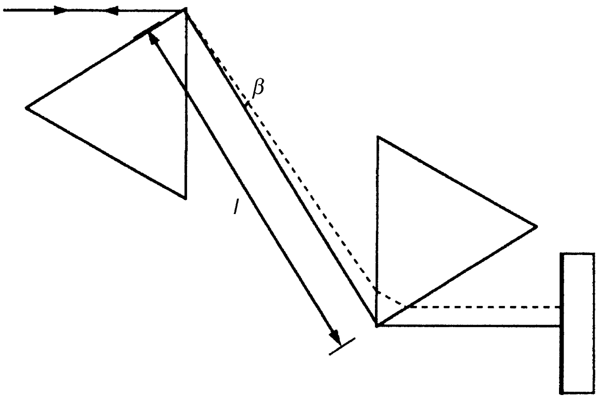	$\frac{d^2\varphi}{d\omega^2} = \frac{\lambda^3}{2\pi c^2}\frac{d^2P}{d\lambda^2}$ $\frac{d^3\varphi}{d\omega^3} = -\frac{\lambda^4}{4\pi^2c^3}\left(3\frac{d^2P}{d\lambda^2} + \lambda\frac{d^3P}{d\lambda^3}\right)$ $\frac{d^2P}{d\lambda^2} = 4l\sin\beta\left[\frac{d^2n}{d\lambda^2} + (2n - 1/n^3)\left(\frac{dn}{d\lambda}\right)^2\right] - 8l\cos\beta\left(\frac{dn}{d\lambda}\right)^2$ $\frac{d^2P}{d\lambda^3} = l\cos\beta\left[\left(24/n^3 - 48n\right)\left(\frac{dn}{d\lambda}\right)^3 - 24\frac{dn}{d\lambda}\frac{d^2n}{d\lambda^2}\right] + l\sin\beta\left[\begin{matrix}\left(\frac{dn}{d\lambda}\right)^3\left(12/n^6 + 12/n^4 + 8/n^3 - 16/n^2 + 32n\right) \\ +\left(24n - 12/n^3\right)\frac{dn}{d\lambda}\frac{d^2n}{d\lambda^2} + 4\frac{d^3n}{d\lambda^3}\end{matrix}\right]$
Double-pass grating-pair, separation l, angle of incidence γ, line spacing d 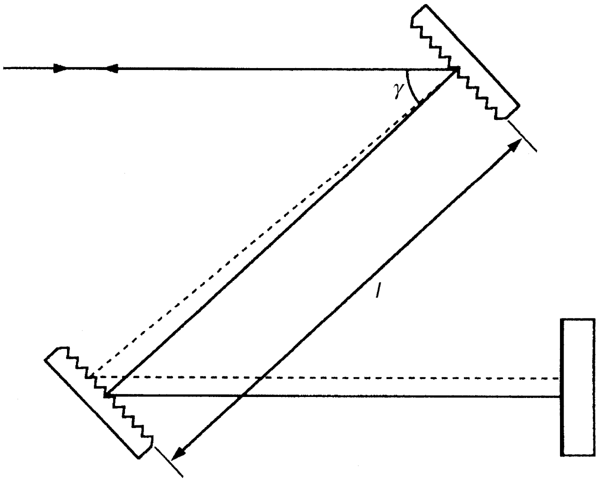	$\frac{d^2\varphi}{d\omega^2} = l\frac{\lambda^3}{\pi c^2d^2}\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-3/2}$ $\frac{d^3\varphi}{d\omega^3} = -6l\frac{\lambda^4}{c^3d^2}\left(1 + \frac{\lambda}{d}\sin\gamma - \sin^2\gamma\right)\left[1 - \left(\frac{\lambda}{d} - \sin\gamma\right)^2\right]^{-5/2}$

corrected using an equal but opposite amount of negative second-order dispersion. The combination of spectral broadening using SPM and subsequent chirp-removal using linear dispersion is the basis for optical pulse compression techniques that have been applied effectively to create pulses which exist for only a few carrier-wave periods.

When the pulse propagates in an environment where SPM and linear dispersion are well balanced and evenly distributed, a soliton will evolve which, in the absence of loss, can propagate indefinitely without changing shape. The nonlinear Schrödinger equation predicts that lasers in which the dominant pulse-shaping mechanisms are negative group-velocity dispersion and self-phase-modulation will produce soliton-like pulses with a $\text{sech}^2(t)$ intensity profile. In experimental measurements of ultrafast lasers the pulse shape is commonly assumed to be $\text{sech}^2(t)$ and the chirp of the pulses is then estimated by measuring their duration–bandwidth product, $\Delta\nu\Delta\tau$, which is a dimensionless quantity that can be treated as a figure-of-merit for ultrashort pulses. For unchirped pulses the duration–bandwidth product is independent of their duration or spectral bandwidth but is sensitive to their shape: chirp-free Gaussian pulses have $\Delta\nu\Delta\tau = 0.441$ while $\text{sech}^2(t)$ pulses have $\Delta\nu\Delta\tau = 0.315$.

Sources of Femtosecond Pulses

Dye Lasers

The first practical passively mode-locked ultrafast lasers were those based on a gain medium of a thin (~10 μm) organic dye jet and incorporating a slow saturable absorber jet at a separate cavity focus. The most common combination was Rhodamine 6G in ethylene glycol (gain jet) and 3,3-diethyloxadicarbocyanine iodide in ethylene glycol (absorber jet) which produced pulses at a wavelength of 620 nm. The most successful approach was colliding-pulse mode-locking in which two counter-propagating pulses within a ring cavity are synchronized to arrive simultaneously at the absorber jet, so achieving greater saturation of the absorber and allowing the pulses to be transmitted with a loss significantly smaller than in the single-pulse case. The first sub-100-fs optical pulses were reported in 1981 from a colliding-pulse mode-locking dye laser and later work reported pulse durations as low as 27 fs but the configuration was inefficient and produced low average output powers. Other approaches such as synchronous pumping and hybrid mode-locking produced higher output powers but the maintenance-intensive nature of dye lasers and the hazardous nature of many of the chemicals used led to their gradual replacement by more convenient solid-state alternatives, based principally on Ti:sapphire.

Color-Center Lasers

A new source of femtosecond pulses in the near-infrared was reported in 1984. Coined the soliton laser the system was a synchronously pumped KCl:Tl color-center laser tunable from 1.4 to 1.6 μm and modified to allow direct generation of pulses as short as 50 fs, substantially shorter than the 8 ps achieved in synchronously pumped operation alone. Femtosecond operation was achieved by coupling a length of polarization-preserving anomalously dispersive fiber to the main laser cavity using a beamsplitter and retroreflector arrangement. This coupled-cavity arrangement later became known as additive pulse mode-locking (APM) and produced short pulses when the optical length of the fiber was adjusted to be an integral multiple of the main cavity length. The long upper-state lifetime of the laser-active centers in the KCl:Tl gain medium allowed particularly stable APM operation, but mode-locking of this kind was also successfully employed in an actively stabilized NaCl:OH^- color-center laser to produce 110-fs pulses. The need to maintain color-center crystals at liquid-nitrogen temperatures limited the convenience of these lasers but until the advent of Ti:sapphire they remained the only solid-state sources capable of femtosecond operation.

Ti:sapphire Lasers

The creation in 1986 of a new broadband laser material – titanium-doped sapphire or Ti:sapphire – began a revolution in ultrafast laser sources that still continues today. Unlike other contemporary solid-state materials for mode-locked lasers such as Nd:YAG, Ti:sapphire is a vibronic gain medium which exhibits a strong electron-phonon coupling between the titanium ion transition and the host lattice. This vibronic coupling leads to a broad continuum of possible transition energies and results in a fluorescence bandwidth covering 670–1050 nm, making Ti:sapphire capable of supporting extremely short femtosecond pulses. The long fluorescence lifetime of Ti:sapphire means that, unlike dye lasers, femtosecond mode-locking requires a passive method using a fast saturable absorber and two methods have prevailed and are discussed below.

The first passively mode-locked Ti:sapphire lasers were based on the APM technique in which mode-locking is achieved by a fast-saturable absorber action due to interference between the intracavity pulse and a self-phase-modulated replica propagating in an auxiliary cavity. An important breakthrough was made in 1990 when it was reported that an APM Ti:sapphire laser continued to operate even when the auxiliary cavity was blocked. This discovery was attributed to an entirely new mode-locking effect known as Kerr-lens mode-locking (KLM) or self-mode-locking in which the presence of a Kerr lens within the gain medium during mode-locked operation changes the mode focusing within the cavity so that, compared to cw operation, mode-locked pulses experience higher gain. Two configurations of KLM exist and are known as soft aperture mode-locking, where the presence of the Kerr lens increases the gain by improving the overlap between the pump and laser modes, and hard-aperture mode-locking, in which a physical aperture such as a slit or the edge of a prism is adjusted to introduce greater loss for cw operation. Because KLM is based on a nonresonant nonlinearity, Ti:sapphire lasers can be mode-locked at any wavelength in their gain bandwidth using this technique. Recent advances in controlling the intracavity group-velocity dispersion characteristics have led to the generation of sub-5-fs pulses directly from a Ti:sapphire oscillator. The principal drawback of the KLM method is that it is not self-starting because mode-locked operation must be seeded by an intense noise spike or other short fluctuation. Different starting methods have been applied successfully including mirror tapping, acousto-optic modulation, and mode-dragging, and once mode-locking has been initiated it is generally stable until some external perturbation disturbs the intracavity beam.

An alternative method for obtaining femtosecond pulses has been pioneered largely in parallel with the KLM efforts and is based on using resonant nonlinearities in semiconductors as fast saturable absorbers. The approach is to fabricate a semiconductor multiple-quantum-well (MQW) device whose bandgap has been chosen to match the laser wavelength and whose absorption can be saturated at high fluences. Initially these devices were utilized in an APM geometry in which the MQW device was situated in an auxiliary cavity but it was later realized that an equivalent monolithic configuration could be obtained by sandwiching the MQW layer between two high-reflectivity mirrors (see Figure 3, inset) to form a Fabry–Perot structure and the resulting element was known as an antiresonant Fabry–Perot saturable absorber. The first application of a such a semiconductor saturable absorber mirror (SESAM) for mode-locking Ti:sapphire was reported in 1995 where self-starting operation was demonstrated. A typical modern Ti:sapphire laser design is shown in Figure 3 and illustrates how a saturable absorber mirror can be integrated into the cavity. Further design enhancements such as reducing the finesse of the Fabry–Perot cavity resulted in SESAMs with a broader wavelength response which enabled sub-10-fs self-starting operation.

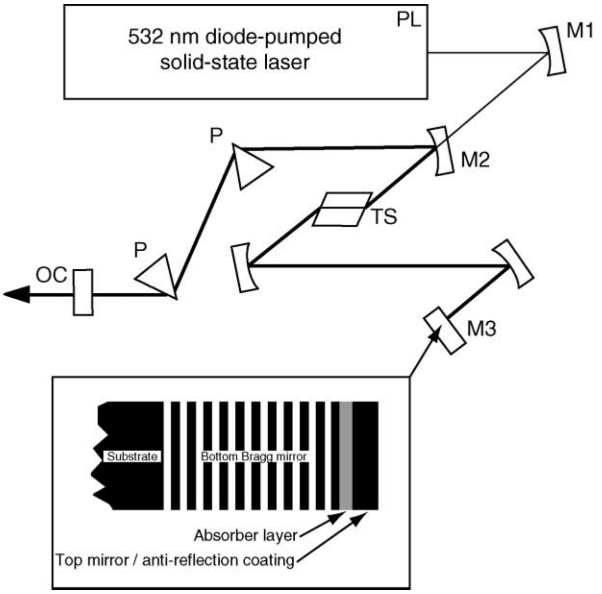

Figure 3 Typical configuration of a modern femtosecond Ti:sapphire laser. PL, pump laser; M1, pump focusing mirror; M2, high-reflectivity cavity mirror (800 nm); TS, Ti:sapphire crystal, typically, 2–10 mm long; M3, SESAM mirror; P, dispersion-compensating prisms; OC, output coupler. Inset: a simplified schematic of a saturable-absorber Bragg mirror.

Fiber Lasers

Ultrafast fiber lasers, particularly those based on silica doped with rare earth Er^{3+}, Nd^{3+}, or Yb^{3+}ions, have enjoyed renewed attention in recent years as efforts to scale their average output powers and pulse energies to practical levels have proved successful. Four main mode-locking strategies have been applied to these systems: active mode-locking which is generally restricted to picosecond generation; intensity-dependent feedback using a nonlinear amplifying loop mirror; polarization APM; and semiconductor saturable absorber mode-locking. The upper-state lifetime of rare-earth-doped fibers is long and therefore a fast saturable absorber mechanism is necessary for passive mode-locking, limiting the choice of mode-locking elements to those based on Kerr or semiconductor nonlinearities.

The development of mode-locked fiber sources has concentrated on Er-doped lasers because of the compatibility of their output wavelengths with the 1.5 μm optical communication window in standard silica fiber. Passively mode-locked designs based on the nonlinear loop mirror work by exploiting the Kerr effect in an intracavity fiber Sagnac interferometer containing a gain section. With suitable control of the intracavity polarization, such a nonlinear amplifying loop mirror (NALM) can be set up to transmit light of high intensities but reflect low intensities because of the differential phase shift induced between the two interferometer arms. By adding the loop mirror to a unidirectional ring cavity a figure-of-eight laser is formed (Figure 4) which is capable of producing femtosecond soliton pulses with average powers of around 1 mW. Drawbacks of the figure-of-eight laser include the presence of spectral sidebands at the shortest pulse durations and a tendency for multiple pulse operation. Mode-locking can also be achieved

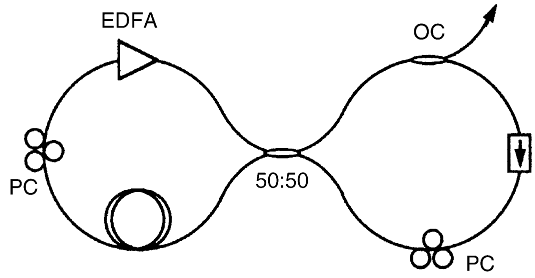

Figure 4 Schematic of a figure-of-eight femtosecond fiber laser. EDFA, erbium-doped fiber amplifier gain section; OC, output coupler; PC, polarization controller.

using a variation on the APM technique and significant average output powers have been demonstrated from stretched-pulse-APM lasers, with up to 90 mW being reported at pulse energies of 2.25 nJ.

Semiconductor saturable absorber mode-locking has been demonstrated in fiber lasers using a variety of configurations including bulk InGaAsP on InP and saturable Bragg reflectors based on InGaAs/InP MQWs on a AlAs/GaAs mirror structure. Fiber lasers mode-locked using semiconductor saturable absorbers are self-starting and exhibit stable mode-locking, normally at the fundamental cavity frequency. The narrow bandwidth of saturable absorber devices at near-infrared wavelengths makes the generation of sub-100-fs pulses difficult.

Semiconductor Sources

Diode lasers and external-cavity lasers based on semiconductor optical amplifiers (SOAs) are of importance because of their ability to produce ultrashort pulses from a compact system at wavelengths of relevance to optical fiber communications. The two principal methods used to generate ultrashort pulses from such devices are mode-locking with an external cavity or gain-switching (equivalently known as *Q*-switching) in a bow-tie waveguide structure. Subpicosecond pulse durations are now routinely produced using dispersion compensation and/or nonlinear pulse compression techniques and, by combining these strategies with Er:fiber amplification, 20-fs pulses have been produced. The mode-locking and gain-switching techniques each have their individual advantages. Gain-switching provides a simple and compact approach offering high average output power, but significant timing jitter and poor spectral quality can accompany pulses produced in this way. Mode-locked systems are physically larger than gain-switched devices because they require an external cavity and additional optical components, but shorter and higher quality pulses are achievable.

Other Common Solid-State Laser Sources

Certain solid-state gain media offer the potential for direct diode pumping, miniaturization and operation at wavelengths or average powers not accessible with other sources. Systems based on Nd or Cr ion transitions have been shown to be capable of ultrashort picosecond or femtosecond operation. Femtosecond mode-locking (175 fs) has been reported in Nd:glass which has one of the broadest fluorescence spectra of the Nd materials. Crystals with active ions of Cr^{2+}, Cr^{3+}, and Cr^{4+} have broad transitions covering wavelengths in the 2.0 μm, 800 nm and 1.4 μm regions and femtosecond operation has been demonstrated in Cr^{4+}:YAG at 1.52 μm, Cr^{4+}:Mg_2SiO_4 (Cr:forsterite) at 1.25 μm, Cr^{3+}: $LiSrAlF_4$ (Cr:LiSAF) at 850 nm, and Cr^{3+}:$LiSrGaF_6$ (Cr:LiSGAF). Direct diode pumping of Cr-ion-based lasers is attractive because of the opportunities for producing low-noise, compact, and efficient femtosecond sources, but the limited power available from high-quality pump diodes results in low intracavity powers which make KLM operation difficult. As a result, stable operation is better achieved using a physical saturable absorber, and SESAM mode-locking has been applied successfully to many of these lasers.

Sources Based on Nonlinear Frequency Conversion

Laser sources alone are unable to provide ultrashort pulses continuously tunable from the ultraviolet to the mid-infrared, and large gaps exist at wavelengths where no broadband laser materials are available. Harmonic generation using Ti:sapphire can be used to access some parts of the visible and ultraviolet regions, but infrared and certain visible wavelengths cannot be generated in this way and coverage is limited by the ability to tune the source itself. The nonlinear production of infrared wavelengths requires parametric down-conversion using a synchronously pumped optical parametric oscillator. Ti:sapphire pumped synchronously pumped optical parametric oscillators have produced, either directly or using harmonic generation, outputs from the green to beyond 7 μm.

Sources of Amplified Ultrashort Pulses

Modern ultrafast oscillators can produce pulses with peak powers of >1 MW, average powers of several watts, and cavity frequencies from 4 MHz to 2 GHz, but many applications require pulses with higher intensities or lower pulse repetition frequencies. Additional amplification stages can be used to satisfy these requirements and these are commonly implemented in Nd:glass for picosecond pulses and Ti:sapphire for femtosecond pulses. In order to avoid damage and unwanted nonlinear effects caused by high pulse intensities within the amplifier system, most practical sources now apply the chirped-pulse amplification (CPA) approach in which a low-energy femtosecond seed pulse is stretched to sub-nanosecond durations to reduce its peak power then amplified and finally compressed to its original duration. Using the chirped-pulse amplification technique, pulses with peak powers as high as 1.5 petawatt have been generated.

Two possible amplifier geometries are used which are categorized either as multipass or regenerative amplifiers. In the multipass configuration the injected pulse is refocused many times through the same gain crystal using a system of mirrors. The gain crystal is pumped above the saturation fluence in order to extract the maximum stored energy from the medium per pass. The multipass geometry is more commonly used for high-average-power operation and critical alignment is required to maximize the overlap between the pump mode profile and the multiple beams intersecting the gain medium. The alternative regenerative amplifier is essentially a stable laser resonator which includes a Pockel cell and a polarizing beamsplitter to enable pulses to be switched into and out of the cavity. The gain per pass is considerably lower than in the multipass configuration but regenerative systems produce superior beam quality and their performance can be optimized without affecting the beam pointing of the output. Commercially available Ti:sapphire amplifiers commonly use the regenerative method and are available with pulse energies of up to 1 mJ at repetition frequencies in the 1–300 kHz range.

Further Reading

Agrawal GP (1995) *Nonlinear Fiber Optics*. San Diego, CA: Academic Press.

Duling IN (1995) *Compact Sources of Ultrashort Pulses*. Cambridge University Press.

Fork RL, Martinez OE and Gordon JP (1984) Negative dispersion using pairs of prisms. *Optics Letters* 9: 150–152.

Keller U, Weingarten KJ, Kartner FX, *et al.* (1996) Semiconductor saturable absorber mirrors for femtosecond to nanosecond pulse generation. *Solid-State Lasers, IEEE Selected Topics in Quantum Electronics* 2: 435–453.

Siegman AE (1986) *Lasers*. University Science Books.

Spence DE, Kean PN and Sibbett W (1991) 60-fsec pulse generation from a self-mode-locked Ti-sapphire laser. *Optics Letters* 16: 42–44.

Szipocs R, Ferencz K, Spielmann C and Krausz F (1994) Chirped multilayer coatings for broad-band dispersion control in femtosecond lasers. *Optics Letters* 9: 201–203.

Yariv A (1989) *Quantum Electronics*. New York: Wiley.

Pulse Characterization Techniques

D J Kane, Southwest Sciences, Inc., Santa Fe, NM, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Ultrashort optical pulses have very short time durations, typically less than a few tens of picoseconds. As a result, these pulses are spectrally broad. Because the index of refraction of materials is a function of wavelength, different wavelengths of light travel at different speeds in optical materials, causing the properties of ultrashort optical pulses to change as they propagate. However, the shape of the pulse can influence how the pulse itself interacts with materials. Thus, the goal of ultrafast laser pulse measurement is to obtain not only the intensity profile of the pulse, but also the actual variation of the frequencies that make up the pulse. This is called measuring the intensity and phase of the pulse, respectively.

Even though the development of techniques to characterize ultrashort optical pulses has not been easy, a myriad of pulse measurement techniques have been developed. In this chapter, we will confine our discussion to the most well-known pulse and generally accepted pulse characterization methods in an effort to provide a basic working knowledge of pulse measurement techniques and a fundamental understanding of the principles behind pulse measurement. After a brief discussion of the mathematical representation of ultrashort optical pulses, we will discuss pulse measurement in (roughly) chronological order, starting with autocorrelation methods. The next section introduces a useful method for the measurement of the relative phase between two pulses called spectral interferometry (SI). Following SI, we introduce the notion of the time-frequency representation of ultrashort optical pulses where techniques such as frequency-resolved optical gating (FROG) are discussed. Following the time-frequency section, a relatively new technique known as spectral phase interferometry for direct electric field reconstruction (SPIDER) is discussed.

Mathematical Representation of an Optical Pulse

The time-dependent variations of the pulse are embodied in the pulse electric field, $A(t)$, which can be written:

$$A(t) = \mathrm{Re}[E(t)e^{i\omega_0 t}] \qquad [1]$$

where ω_0 is the carrier frequency and Re refers to the real part. While we can use $A(t)$ as it stands for calculations, it is much easier to remove the rapidly varying ω_0 part, $e^{i\omega_0 t}$, and use a slowly varying envelope together with a phase term that contains only the frequency variations, not the rapidly varying carrier frequency:

$$E(t) = [I(t)]^{1/2} e^{i\varphi(t)} \qquad [2]$$

where $I(t)$ and $\varphi(t)$ are the time-dependent intensity and phase of the pulse. (Note that $E(t)$ is complex.) The frequency variation, $\Omega(t)$, is the derivative of $\varphi(t)$ with respect to time:

$$\Omega(t) = -\mathrm{d}\varphi(t)/\mathrm{d}t \qquad [3]$$

The pulse field can be written equally well in the frequency domain by taking the Fourier transform of eqn [2]:

$$\tilde{E}(\omega) = [\tilde{I}(\omega)]^{1/2} e^{-i\phi(\omega)} \qquad [4]$$

where $\tilde{I}(\omega)$ is the spectrum of the pulse, and $\phi(\omega)$ is its phase in the frequency domain. The spectral phase contains time versus frequency information; that is, the derivative of the spectral phase with respect to frequency yields the time arrival of the frequency – the group delay.

Obtaining the intensity and phase, $I(t)$ and $\varphi(t)$ (or $\tilde{I}(\omega)$ and $\phi(\omega)$) is called full characterization of the pulse. Common phase distortions include linear chirp, where the phase (either the time domain or frequency domain) is parabolic. When the frequency is increasing in time, the pulse is said to have positive linear chirp; negative linear chirp is when the high frequencies lead the lower frequencies. Higher-order chirps are common, but for these, differentiation between spectral and temporal chirp is required because spectral phase and temporal phase are not interchangeable.

Autocorrelation

Traditionally, we measure events using shorter events. Unfortunately, for the ultrafast researcher, shorter events do not exist and modern electronics are not fast enough. Therefore, because the shortest event we have is the event we wish to measure, traditional pulse measurement methods use the pulse itself to determine the approximate duration of the pulse in question giving birth to the ubiquitous intensity autocorrelation (see Figure 1). While autocorrelations cannot be used to fully characterize ultrashort optical pulses, the methods used in autocorrelations are fundamental to all pulse measurement schemes.

The intensity autocorrelation is measured by combining a pulse and a delayed replica of a pulse in a nonlinear medium such as a second harmonic generation (SHG) crystal. A pulse is sent onto a beamsplitter to produce the two replicas. One pulse is delayed relative to the other and both are focused together into a SHG crystal. As the delay of one pulse relative to the other is varied, the intensity of the second-harmonic signal is recorded. The intensity autocorrelation is not limited to using SHG; any nonlinearity may be utilized. Indeed, two-photon absorption in a semiconductor LED or photodiode often acts as a convenient nonlinearity. Sometimes a third-order nonlinearity is used to provide some direction of time information about the pulse. When the generation of the signal involves phase matching, such as second harmonic generation, care must be taken to use a thin crystal. Typically, 10 μm to 1 mm thick SHG crystals are used, depending on the bandwidth of the pulse to be measured. The exact thickness depends on the SHG crystal, the pulse width, and the requirements of the measurement.

Regardless of the nonlinearity used, an autocorrelation yields only the approximate duration and

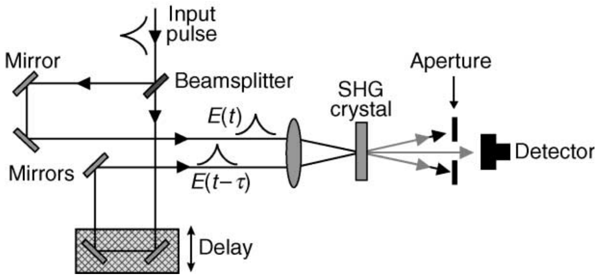

Figure 1 An intensity autocorrelator used to determine the approximate duration of a pulse. The pulse is split into two replicas that are sent into delay lines. The outputs from both delay lines are focused into a doubling crystal. The second harmonic output is monitored as a function of delay between the two pulses. A plot of the signal versus time is the intensity autocorrelation of the pulse.

shape of the pulse. The structure of the pulse is smeared, producing a smooth, featureless profile for even a complex pulse. Thus, the intensity autocorrelation alone does not determine the intensity profile of the pulse, $I(t)$. Only the interaction of the intensity of the pulse is recorded, producing no phase information for the pulse. Some qualitative information can be gleaned from the spectrum and the autocorrelation, but usually only that the pulse is chirped, and higher-order chirps and complex pulse structures elude such an analysis.

Another useful form of intensity autocorrelation is the single-shot autocorrelator. In this case, fairly large beams are used, and these beams are set to intersect at an angle of 2Φ, tilting the pulse fronts, so that delay is mapped onto a spatial coordinate (see Figure 2). The beams are then focused into a nonlinear medium using a cylindrical lens. The interaction region in the SHG crystal is imaged onto a linear array or CCD camera. Typically, when second-harmonic generation is used, the SHG crystal is oriented for Type I phase matching. The time window is proportional to $\Delta w(\sin\Phi)/\Delta v_g$, where Δw is the beam waist at the SHG crystal, and v_g is the group velocity of the pulse in the crystal. Like all SHG interactions, care must be exercised to insure that the phase matching is sufficient to mix the entire bandwidth of the pulse.

Jean-Claude Diels improved the intensity autocorrelation by the development of the interferometric autocorrelation. In this configuration, a Michelson interferometer is typically used so that the beams are collinear when arriving at the SHG crystal, which allows the beams to interfere. As a result, fringes appear on the autocorrelation. By examining the shape and extent of the fringes, some information about the pulse chirp can be obtained. While this does add to the information provided by an intensity autocorrelation, the added information is only the spectrum of the second harmonic. Unfortunately the addition of the second harmonic spectrum does not provide enough information for full retrieval of the intensity and phase.

Even though autocorrelation alone does not completely determine pulse intensity and phase, it is a very simple and useful technique to determine approximate pulse duration. Sometimes, cross-correlation can be used to determine the duration of a long pulse, if a shorter one is available, or a pulse too weak (or at an inappropriate wavelength) to measure with autocorrelation. The addition of the spectrum and/or the second harmonic spectrum can provide more,

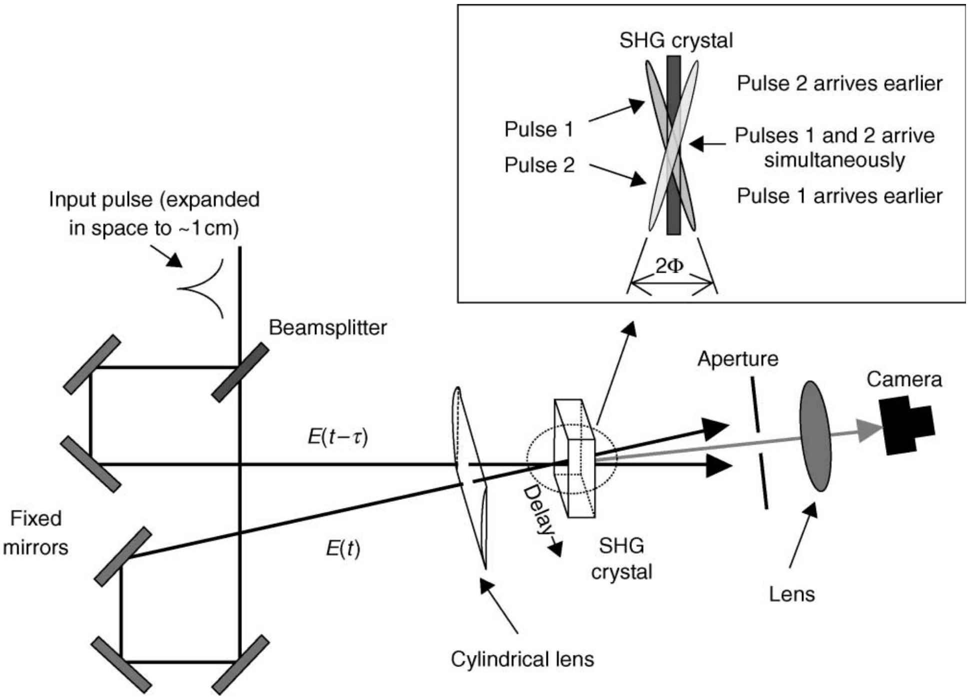

Figure 2 A single shot intensity autocorrelator. For this type of autocorrelator, the beam size is large, on the order of 1 cm in diameter, and the pulse fronts are tilted with respect to each other in order to map relative delay to position. The two beams are focused using a cylindrical lens and the output is recorded using an array detector. The inset shows how delay is mapped into position. Inset: the tilted pulse fronts cause time delay to be mapped spatially.

albeit, incomplete information about pulse chirp. More importantly, the experimental techniques used by correlation methods are fundamental to all pulse measurement methods.

Spectral Interferometry – Relative Phase Measurements

While determination of the absolute intensity and phase of an ultrashort laser pulse is difficult, determination of the relative phase is not. Invented in the late 1800s, spectral interferometry (SI), measures the relative phase between two pulses. SI, a linear technique, can also be modified (albeit, a some-what complex modification) to be self-referencing for full characterization of the ultrashort optical pulse (See the section on SPIDER below). In addition, when combined with a full pulse characterization technique, SI, because it is a linear technique, provides a method by which very weak optical pulses can be characterized (see below).

Typically, an SI apparatus uses a Mach–Zender interferometer (see Figure 3). A pulse is split into two replicas; one is sent through a region or medium to measure and the other through a known path. The two pulses remain separated by some known time and are sent, collinearly, into a spectrometer; no time scanning is required. An analysis of the fringe pattern yields the relative phase between the two pulses. In other words, SI provides $\phi_{\text{unk}}(\omega) - \phi_{\text{ref}}(\omega)$, where $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase versus frequency. The spectrum of the two pulses in the frequency domain is

$$\tilde{I}_{\text{SI}}(\omega) = |\tilde{E}_0(\omega) + \tilde{E}_{\text{unk}}(\omega)|^2$$

$$\begin{aligned}\tilde{I}_{\text{SI}}(\omega) = {} & \tilde{I}_0(\omega) + \tilde{I}_{\text{unk}}(\omega) + 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)} \\ & \times \cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau)\end{aligned} \qquad [5]$$

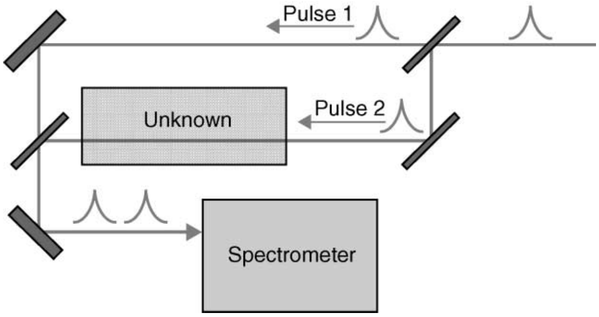

Figure 3 A Mach–Zender interferometer is the typical arrangement for spectral interferometry. One path acts as the reference arm and the sample to be measured is placed in the other arm. Unlike standard interferometry, the delay is fixed in a spectral interferometry experiment.

where $\tilde{I}_{\text{SI}}(\omega)$ is the output spectrum, $\tilde{E}_0(\omega)$ is the reference pulse electric field, $\tilde{E}_{\text{unk}}(\omega)$ is the unknown pulse electric field, $\tilde{I}_0(\omega)$ is the reference pulse spectrum, $\tilde{I}_{\text{unk}}(\omega)$ is the unknown pulse spectrum, $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase, $\phi_0(\omega)$ is the reference pulse phase, and τ is the delay between the two pulses. To facilitate extracting the phase difference from the SI output, τ is chosen to yield fringes in the sum spectrum. The spectral fringes, which have a period inversely proportional to the optical path difference between the two beams (see Figure 4), contain all of the phase difference information.

Figure 4 shows the steps required to obtain the phase difference between the two pulses. The first step is to subtract out the spectra of the individual pulses in order to isolate the spectral interferogram, $S(\omega)$ (Figure 4a), where

$$S(\omega) = 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \qquad [6]$$

By Fourier transforming $S(\omega)$, we obtain

$$\Im^{-1}[S(\omega)] = f(t-\tau) + f(-t-\tau) \qquad [7]$$

where $f(t)$ is the correlation product between the reference and the unknown electric field (see Figure 4b). The Fourier transform of eqn [7] multiplied by a Heaviside function, $\Theta(t)$ (to remove $f(-t-\tau)$), recovers the amplitude and phase of $f(\omega)$, the Fourier transform of $f(t-\tau)$, which contains the relative spectral phase between the unknown pulse and the unmodified pulse. Some care, however, needs to be taken to correctly remove the linear phase due to the time delay between the two pulses, $\omega\tau$. Also, for best results, the spectrometer should be well calibrated and care should be taken to properly window the spectrum before taking the Fourier transform.

Spectral interferometry is a straightforward and simple technique that always provides a relative phase. For this very reason, care must be taken to make sure the beams are mode-matched as well as collinear; also, interferometric stability must be maintained over the course of the measurement. Furthermore, SI is nongating; that is, CW background in the laser can add to the interferogram, masking transient effects. Nevertheless, spectral interferometry is a useful technique that has been applied to the measurement of the linear and nonlinear spectral phase introduced by optical fibers. More recently, SI has been applied to phase-locking and to phase-resolved pump probe experiments.

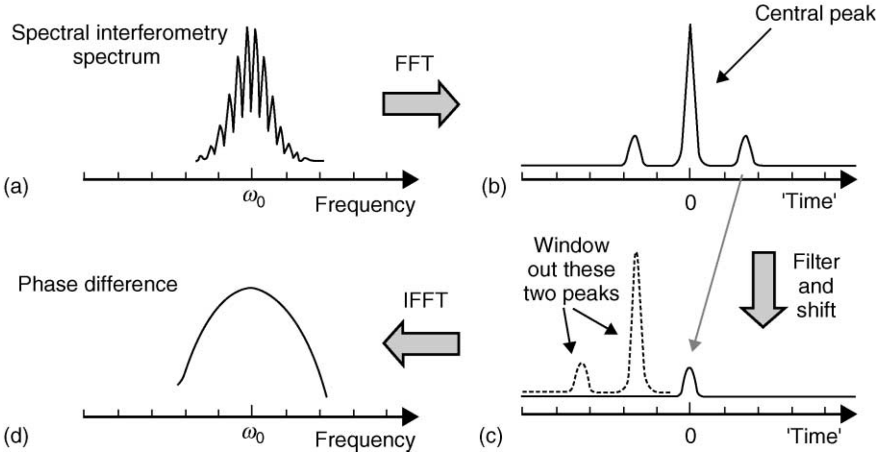

Figure 4 The steps required for the analysis of spectral interferograms. Part (a) shows a sample SI interferogram. The dominant frequency of the fringes is the delay multiplied by the speed of light. The actual phase differences of interest are the perturbations on this frequency. Part (b) is the Fourier transform of the SI interferogram in Part (a). (Note that this is not the true time domain.) The central peak contains only spectral information. All of the phase information is contained in the satellite peaks. The next step is to mask out the central peak and one of the satellite peaks (Part (c)). The phase of the inverse Fourier transform of Part (c) yields the relative phase between the two pulses (Part (d)).

Time-Frequency Representation

In 1971, E.B. Treacy laid the foundations for a revolution in pulse measurement by introducing the idea of measuring the intensity versus time for different spectral splices of a pulse. Because his measurements provided both time and frequency information simultaneously, these measurements could be thought of as applying in a hybrid time-frequency domain. At first, this may seem confusing, but similar methods have been used to visualize sounds patterns (a musical score, for example) and speech. A time-frequency plot of this type, where spectral slices of a pulse are plotted versus time, is called a sonogram. The mathematical formalism of a sonogram is:

$$S_{\mathrm{E}}(\Omega, T) = \left| \int_{-\infty}^{\infty} \tilde{E}(\omega) h(\omega - \Omega) \mathrm{e}^{-\mathrm{i}\omega T} \mathrm{d}\omega \right|^2 \qquad [8]$$

where $\tilde{E}(\omega)$ is the electric field of the pulse to be measured in the frequency domain and $h(\omega - \Omega)$ is a frequency gate that varies with frequency. The magnitude squared of the inverse Fourier transform of spectral slices yields the sonogram.

In 1991, Chilla and Martinez showed that the sonogram could be used to reconstruct the full intensity and phase of the pulse. That is, under certain conditions, the approximate group delay could be determined as a function of frequency from the sonogram by finding the peak time arrival of each spectral slice; integration of the group delay yields the spectral phase, which together, with the pulse spectrum, provides the intensity and phase of the pulse in the frequency domain. They labelled this technique as frequency domain phase measurement (FDPM).

An experimental diagram of an FDPM apparatus is shown in Figure 5. The pulse is split into two replicas, and one of the pulse replicas is spectrally filtered. The spectrally filtered pulse is cross-correlated with the original pulse to find the 'time arrival' of the spectrally filtered pulse – defined as the peak of the cross-correlated pulse. Because the spectrally filtered pulse has a much longer time duration than the original pulse, the small perturbation caused by the finite length of the original pulse is neglected.

The main difficulty of the Chilla–Martinez method is that the frequency gate must be very narrow, reducing the filtered pulse energy significantly, which greatly reduces the measured signal strength. If the spectral phase is not well behaved, the filtered pulse may not be much longer than the original pulse. In addition, for a given frequency, the sonogram may have two or more peaks in time, that is, the group delay may not be a function. However, a method called 2D phase retrieval, discussed in the next section, can alleviate many of these issues.

A spectrogram is a relative of the sonogram. Rather than determining the time arrival of spectral slices of a pulse, a spectrogram is obtained when the spectrum of time slices of the pulse are measured. The spectrogram is experimentally much easier to obtain than the sonogram; however, the measured quantity is not

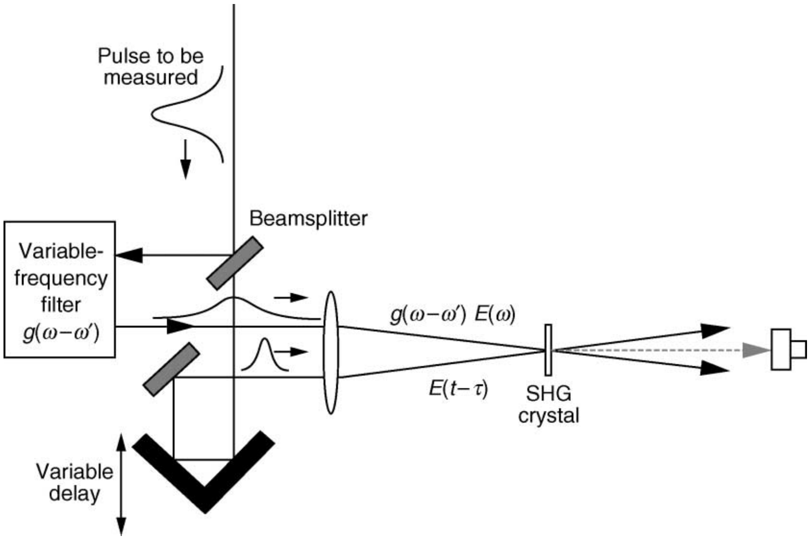
Pulse to be measured
Beamsplitter
Variable-frequency filter $g(\omega-\omega')$
$g(\omega-\omega')\ E(\omega)$
$E(t-\tau)$
SHG crystal
Variable delay

Figure 5 Measuring a sonogram requires determining the time arrival of spectral slices of the pulse. The pulse is split into two replicas using a beamsplitter. One replica is spectrally filtered using a tunable filter. The other pulse is cross-correlated with the spectrally filtered pulse to determine the time arrival.

quite as mathematically useful as the sonogram because the time slicing of the pulse occurs in the time domain. Thus, if a narrow time gate is used, the time domain phase is obtained. The time domain phase together with the intensity of the pulse, $I(t)$, provides the full intensity and phase of the pulse. Unfortunately, the intensity profile is not as readily measurable as the pulse spectrum. Consequently, to obtain the full intensity and phase of a pulse from its spectrogram requires an iterative 2D phase retrieval algorithm. This is the basis of a pulse measurement technique called frequency-resolved optical gating (FROG).

Frequency-Resolved Optical Gating

Frequency-resolved optical gating (FROG), developed by Kane and Trebino, measures the spectrum of a particular temporal component of the pulse (see Figures 6 and 7) by spectrally resolving the signal pulse in an autocorrelation-type experiment using an instantaneously responding nonlinear medium. As shown in Figure 6, FROG involves splitting a pulse and then overlapping the two resulting pulses in an instantaneously responding $\chi^{(3)}$ or $\chi^{(2)}$ medium. Even though any instantaneous nonlinear interaction may be used to implement FROG, perhaps the most intuitive is the polarization-gating configuration. In this case, induced birefringence, due to the electronic Kerr effect, is used as the nonlinear-optical process. In other words, the 'gate' pulse causes the $\chi^{(3)}$ medium, which is placed between two crossed polarizers, to become slightly birefringent. The polarization of the 'gated' probe pulse (which is cleaned up by the first polarizer) is rotated slightly by the induced birefringence allowing some of the 'gated' pulse to leak through the second polarizer. This is referred to as the signal. Because most of the signal emanates from the region of temporal overlap between the gate pulse and the probe pulse, the signal pulse contains the frequencies of the 'gated' probe pulse within this overlap region. The signal is then spectrally resolved, and the signal intensity is measured as a function of wavelength and delay time τ. The resulting trace of intensity versus delay and frequency is a spectrogram, a time- and frequency-resolved transform that intuitively displays time-dependent spectral information of a waveform.

The spectrogram can be expressed as:

$$S_{\mathrm{E}}(\omega,\tau)=\left|\int_{-\infty}^{\infty}E(t)g(t-\tau)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t\right|^2 \qquad [9]$$

where $E(t)$ is the measured pulse's electric field, $g(t-\tau)$ is the variable-delay gate pulse, and the subscript E on S_{E} indicates the spectrogram's dependence on $E(t)$. The gate pulse $g(t)$ is usually somewhat shorter in length than the pulse to be measured, but not infinitely short. This is an important point: an infinitely short gate pulse yields only the intensity $I(t)$ and conversely, a CW gate yields only the spectrum $I(\omega)$. On the other hand,

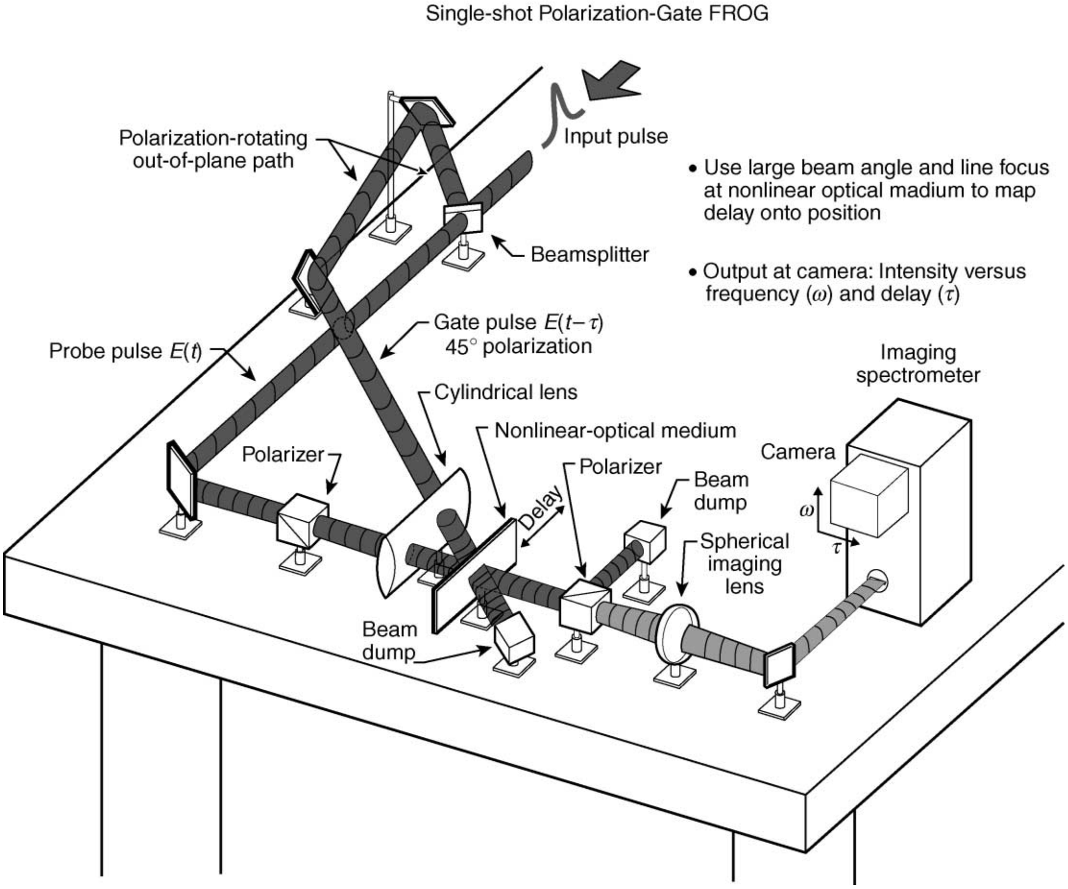

Single-shot Polarization-Gate FROG

- Use large beam angle and line focus at nonlinear optical medium to map delay onto position
- Output at camera: Intensity versus frequency (ω) and delay (τ)

Input pulse
Polarization-rotating out-of-plane path
Beamsplitter
Gate pulse $E(t-\tau)$ 45° polarization
Probe pulse $E(t)$
Imaging spectrometer
Cylindrical lens
Nonlinear-optical medium
Camera
Polarizer
Polarizer
Beam dump
ω
τ
Delay
Spherical imaging lens
Beam dump

Figure 6 Measuring the spectrogram of a pulse is easier than measuring its sonogram – a spectrogram is a spectrally resolved autocorrelation. In this figure, the optical Kerr–effect is used (polarization-gate) as the nonlinearity. (Adapted with permission from Kane DJ and Trebino R (1993) Single shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating. *Optics Letters* 18(10): 823–825.)

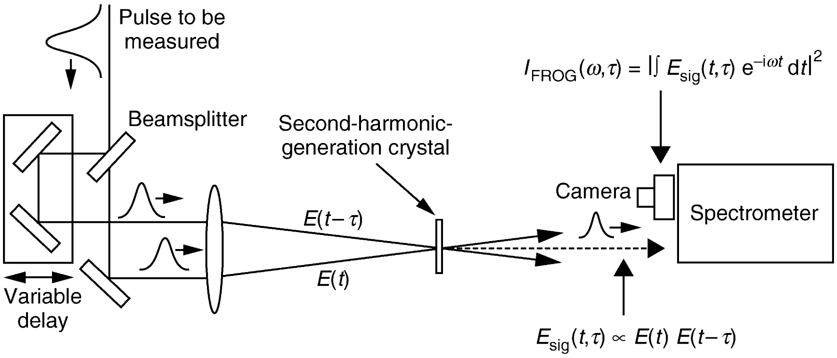

Pulse to be measured
$I_{\text{FROG}}(\omega,\tau) = |\int E_{\text{sig}}(t,\tau)\, e^{-i\omega t}\, dt|^2$
Beamsplitter
Second-harmonic-generation crystal
Camera
Spectrometer
$E(t-\tau)$
$E(t)$
Variable delay
$E_{\text{sig}}(t,\tau) \propto E(t)\, E(t-\tau)$

Figure 7 This figure shows a schematic of an SHG FROG device. The SHG signal from the autocorrelation is spectrally resolved. SHG FROG is very simple and sensitive, but it has a direction-of-time ambiguity. For example, if the pulse has chirp, only the magnitude of the chirp is determined – the sign of the chirp remains unknown.

a finite-length gate pulse yields the spectrum of all of the finite pulse segments with duration equal to that of the gate. While the phase information remains lacking in each of these short-time spectra, having spectra of an infinitely large set of pulse segments compensates for this loss. The spectrogram has been shown to nearly uniquely determine both the intensity $I(t)$ and phase $\varphi(t)$ of the pulse, even if the gate pulse is longer than the pulse to be measured (although if the gate is too long, sensitivity to noise and other practical problems arise).

In FROG, when using optically induced birefringence as the nonlinear effect, the signal pulse is given by:

$$E_{\mathrm{sig}}(t,\tau) \propto E(t)|E(t-\tau)|^2 \qquad [10]$$

So the measured signal intensity $I_{\mathrm{FROG}}(\omega,\tau)$, after the spectrometer is:

$$I_{\mathrm{FROG}}(\omega,\tau) = \left| \int_{-\infty}^{\infty} E(t)|E(t-\tau)|^2 \mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t \right|^2 \qquad [11]$$

We see that the FROG trace is a spectrogram of the pulse $E(t)$ although the gate, $|E(t-\tau)|^2$, is a function of the pulse itself.

To see that the FROG trace essentially uniquely determines $E(t)$ for an arbitrary pulse, it is first necessary to observe that $E(t)$ is easily obtained from $E_{\mathrm{sig}}(t,\tau)$. Then it is simply necessary to write eqn [11] in terms of $E_{\mathrm{sig}}(t,\Omega)$, the Fourier transform of the signal field $E_{\mathrm{sig}}(t,\tau)$, with respect to delay variable τ. We then have what appears to be a more complex expression, but one that will give us better insight into the problem:

$$I_{\mathrm{FROG}}(\omega,\tau) = \left| \int_{-\infty}^{\infty} E_{\mathrm{sig}}(t,\Omega)\mathrm{e}^{-\mathrm{i}\omega t-\mathrm{i}\Omega\tau}\mathrm{d}t\,\mathrm{d}\Omega \right|^2 \qquad [12]$$

Equation [12] indicates that the problem of inverting the FROG trace $I_{\mathrm{FROG}}(\omega,\tau)$, to find the desired quantity $E_{\mathrm{sig}}(t,\Omega)$, is that of inverting the squared magnitude of the two-dimensional (2D) Fourier transform of $E_{\mathrm{sig}}(t,\Omega)$. This problem, which is called the 2D phase-retrieval problem, is well known in many fields, especially in astronomy, where the squared magnitude of the Fourier transform of a 2D image is often measured. At first glance, this problem appears unsolvable; after all, much information is lost when the magnitude is taken. It is well known that the one-dimensional (1D) phase retrieval problem is unsolvable (for example, infinitely many pulse fields give rise to the same spectrum). Intuition fails in this case, however; two- and higher-dimensional phase retrieval essentially always yields unique results.

The Simplified FROG – GRENOUILLE

A simplified version of the FROG device, known as grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE), can be constructed using the doubling crystal as both the gating medium and the spectrometer (see Figure 8). In this case, the phase matching condition in certain thick doubling crystals causes a wavelength dependent output from the doubling crystal. A cylindrical lens is used in a Fourier transform configuration to image the wavelength variation onto a CCD camera. A cylindrical lens set at 90 degrees to the Fourier transform lens images the time axis onto the CCD camera. A Fresnel biprism replaces the traditionally used Mach–Zender interferometer to produce two pulses propagating at an angle with respect to each other. Presently, the GRENOUILLE technique works with BBO, in the wavelength region around 500 nm–1200 nm, and with proustite in the wavelength region around 1200 nm–2000 nm.

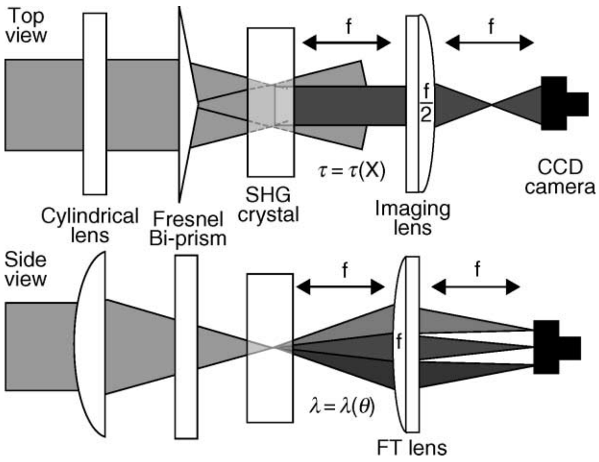

Figure 8 Figure showing a schematic of a GRENOUILLE FROG device which uses a thick doubling crystal as both the gating nonlinearity and the spectrometer. Like single-shot autocorrelators, a cylindrical lens focuses a spatially large input beam into an SHG crystal in only one dimension. Unlike most autocorrelators, the Mach–Zender interferometer is replaced with a Fresnel biprism that forces each half of the input beam to propagate, at an angle, toward the SHG crystal. An imaging cylindrical lens images the spatial dependence of the delay onto a CCD camera. A Fourier transform cylindrical lens maps the angular dependence of wavelength from the SHG crystal to a spatial coordinate on the CCD camera. (Adapted with permission from O'Shea P, Kimmel M, Xun G and Trebino R (2001) Highly simplified device for ultrashort-pulse measurement. *Optics Letters* 26(12): 932–934.)

FROG Inversion Algorithms

An iterative 2D phase retrieval algorithm is required to extract the pulse information from the measured FROG trace (see Figures 9 and 10). This algorithm converges to a pulse that minimizes the difference

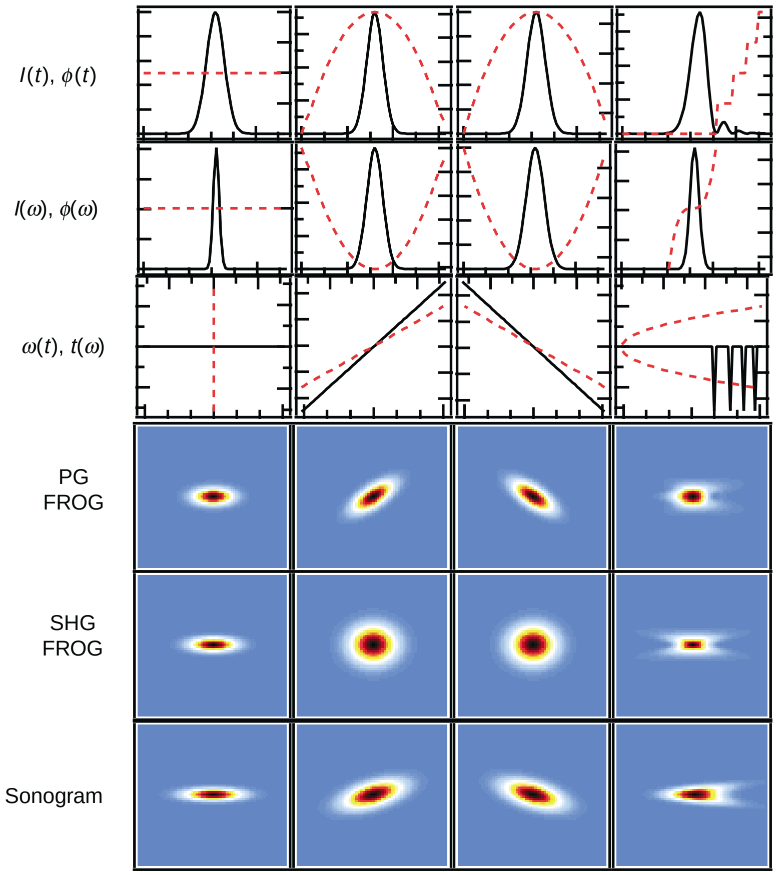

Figure 9 Example FROG traces and sonograms are shown for four different pulses: a transform limited pulse, a positively chirped pulse, and a negatively chirped pulse. The top series of plots are the time domain representation of the pulse intensity (solid line) and phase (dashed line). The second row of four plots is the frequency domain representation of the sample pulse's intensity and phase. The next four plots show the instantaneous frequency and the group delay. The vertical axis is frequency and the horizontal axis is time. The remaining rows show the PG FROG traces, the SHG FROG traces and sonograms of the pulse shown in the first two rows.

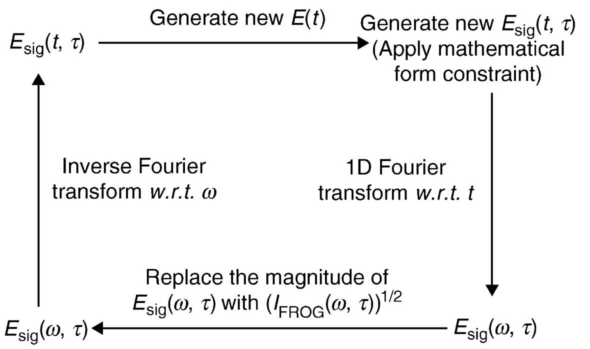
$E_{sig}(t, \tau)$
Generate new $E(t)$
Generate new $E_{sig}(t, \tau)$ (Apply mathematical form constraint)
Inverse Fourier transform *w.r.t.* ω
1D Fourier transform *w.r.t.* t
Replace the magnitude of $E_{sig}(\omega, \tau)$ with $(I_{FROG}(\omega, \tau))^{1/2}$
$E_{sig}(\omega, \tau)$
$E_{sig}(\omega, \tau)$

Figure 10 Phase retrieval algorithm for the inversion of FROG spectrograms. Start with an initial guess for the pulse to generate an initial $E_{sig}(t,\tau)$. A 1D Fourier transform generates the FROG trace. The next step is to replace the magnitude of the calculated FROG trace with the square root of the measured FROG trace. Inverse Fourier transform with respect to ω to produce the new signal field and generate a new guess for $E(t)$. Interestingly, it is only the step that produces $E(t)$ from $E_{sig}(t,\tau)$ that differentiates all the FROG algorithms.

between the measured and the calculated FROG trace. While this aspect of FROG has been its Achilles heel in the past, in reality, new generalized projections algorithms (together with faster computers) converge quickly and can track pulse changes at rates up to 30 Hz and beyond, making FROG a true, real-time pulse measurement technique. Indeed, excellent algorithms for the analysis of FROG traces in real-time are commercially available.

The original FROG inversion algorithm, commonly referred to as the vanilla algorithm, is simple and iterates quickly, but tends to stagnate, giving erroneous results, especially for geometries that use a complex gate function such as SHG or self-diffraction. An improved algorithm, incorporating several different algorithms including brute force minimization, was developed to alleviate stagnation programs at the expense of both speed and convergence time. By the mid-1990s, a significant advance in both speed and stability was made by the addition of a numerical method called generalized projections. This algorithm determines the next guess by constructing a projection – minimizing the error (distance) between the FROG electric field, $E_{\text{sig}}(t, \tau)$, obtained immediately after the application of the intensity constraint, and the FROG electric field calculated from the mathematical form constraint.

The first generalized projections algorithm used a standard minimization procedure to find the electric field for the next iteration, which can still be slow. For the most common FROG geometries, PG and SHG, a different algorithm that determines the next iteration directly has been developed. This algorithm, called the principal components generalized projections (PCGP) algorithm, converts the generalized projections algorithm into an eigenvector problem. Using this algorithm, pulse measurement rates of at least 30 Hz have been achieved.

The goal of the phase retrieval algorithm is to find the $E(t)$ that satisfies two constraints. The first is the FROG trace itself which is the magnitude squared of the 1D Fourier transform of $E_{\text{sig}}(t, \tau)$:

$$I_{\text{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\text{sig}}(t, \tau) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \right|^2 \qquad [13]$$

The other constraint is the mathematical form of the signal field, $E_{\text{sig}}(t, \tau)$, for the nonlinear interaction used. The signal forms for a variety of FROG beam geometries are:

$$E_{\text{sig}}(t, \tau) \propto \begin{Bmatrix} E(t)|E(t-\tau)|^2 & \text{PG FROG} \\ E(t)^2 E^*(t-\tau) & \text{SD FROG} \\ E(t)E(t-\tau) & \text{SHG FROG} \\ E(t)^2 E(t-\tau) & \text{THG FROG} \end{Bmatrix} \qquad [14]$$

where PG is polarization gate, SD is self-diffraction, SHG is second harmonic generation, and THG is third harmonic generation FROG.

All FROG algorithms work by iterating between two different data sets: the set of all signal fields that satisfy the data constraint, $I_{\text{FROG}}(\omega, \tau)$, and the set of all signal fields that satisfy eqn [14]. The difference between the FROG algorithms is how the iteration between the two sets is completed. In the case of generalized projections, the $E(t)'$s are chosen such that the distance between the $E(t)$ on the magnitude set and the $E(t)$ on the mathematical form set is minimized. This is accomplished by minimizing the following equation:

$$Z = \sum_{i,j=1}^{N} \left| E^{(k)}_{\text{sig(DC)}}(t_i, \tau_j) - E^{(k+1)}_{\text{sig(MF)}}(t_i, \tau_j) \right|^2 \qquad [15]$$

where $E^{(k)}_{\text{sig(DC)}}(t_i, \tau_j)$ is the signal field generated by the data constraint, and $E^{(k+1)}_{\text{sig(MF)}}(t_i, \tau_j)$ is the signal field produced from one of the beam geometry equations in eqn [14]. For the normal generalized projection, the minimization of Z is completed using a standard steepest descent algorithm; the derivative of Z with respect to the signal field is computed to determine the direction of the minimum. The computation of the derivatives are tedious; they are tabulated elsewhere.

An alternative to an algorithm that minimizes eqn [15] is the PCGP algorithm. This algorithm converts the computation of the next guess into an eigenvector problem, reducing the computation of the next guess to simple matrix–vector multiplications. This algorithm works for both the PG and SHG beam geometries, it is robust and the fastest FROG algorithm. Indeed, the PCGP algorithm was used in the original real-time FROG work by D.J. Kane.

Self Checks in FROG Measurements

Unlike any other pulse measurement techniques, FROG can provide a great deal of feedback about both the quality of the measurement (systematic errors) and the quality of the algorithm's performance. The most common check for convergence is the FROG trace error together with a visual comparison between the retrieved FROG trace and the measured FROG trace. The FROG trace error is given by:

$$G = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^{N} \left| I_{\text{FROG}}(\omega_i, \tau_j) - \alpha I^{(k)}_{\text{FROG}}(\omega_i, \tau_j) \right|^2} \qquad [16]$$

where α is a renormalization constant, I_{FROG} is the measured trace, and $I^{(k)}_{\text{FROG}}(\omega, \tau)$ is computed from the retrieved electric field. Typically, the FROG trace error of a PG measurement should be less than 2% for a 64 × 64 pixel trace, while the FROG trace error of a 64 × 64 pixel SHG FROG trace should be about 1% or less. Acceptable FROG trace errors decrease as FROG trace size increase for smaller FROG traces. These values are only rules of thumb only;

for example, acceptable retrievals of large and very complicated FROG traces can produce larger FROG trace errors.

In a good FROG measurement, the spectrum of the retrieved pulse should faithfully reproduce the salient features of the pulse's measured spectrum. SHG FROG even provides an additional check called the frequency marginal. The sum of an SHG FROG trace along the time axis yields the autoconvolution of the pulse's spectrum, providing two ways the FROG measurement can be checked. First, the autoconvolution of an independently measured spectrum can be compared to the sum of the FROG trace along the time axis, providing an indication of how well the measurement was made. For example, if the doubling crystal was too thick in the pulse measurement, the FROG trace's frequency marginal will be narrower than the autoconvolution of the measured spectrum. Second, comparing the autoconvolution of the retrieval pulse spectrum with the FROG trace marginal can provide a test of algorithm convergence in addition to a test of the measurement. Phase matching problems appear as a mismatch between the FROG trace frequency marginal and the autoconvolution of the retrieved spectrum.

Because FROG is a spectrally resolved autocorrelation, summing any FROG trace along the frequency axis yields the autocorrelation of the measured pulse. This autocorrelation can be compared to an independently measured autocorrelation, or a comparison can be made between the frequency sum of the FROG trace and the autocorrelation calculated from the retrieved pulse to determine algorithm convergence and the quality of the measurement.

Measuring Pulses Directly – SPIDER

SPIDER is a novel technique that measures the pulse directly – no iterative phase retrieval method is required. To accomplish this feat, spectral interferometry is conducted on two pulse replicas that are shifted in frequency with respect to one another. The original pulse is split into two replicas. The two pulse replicas are sent into a phase modulator that shifts the center frequency of each pulse slightly. The frequency-shifted pulses are sent into a spectrometer and a standard spectral interferometry analysis yields the derivative of the phase. Integration of the phase derivative, together with the spectrum of pulse, provides the full intensity and phase.

Thus, analysis of the spectral interferogram gives:

$$\Delta\phi = (\phi(\omega_1) - \phi(\omega_2))\Delta\omega \qquad [17]$$

where $\Delta\phi$ is the measured phase difference from the spectral interferogram, $\Delta\omega$ is the difference in the center frequency between the two pulses, and $\phi(\omega)$ is the pulse phase as a function of frequency. Consequently, the phase difference must be divided by the frequency difference between the two pulses to determine the true derivative. It is also important to subtract out the linear phase difference resulting from the delay between the two pulses (the delay produces the fringes in the resulting interferogram) as this integrates to a linear chirp.

Because phase modulators are too slow to be used in the femtosecond region, a different method is used to shift the frequency of the two pulses – they are mixed with a highly chirped pulse (see Figure 11).

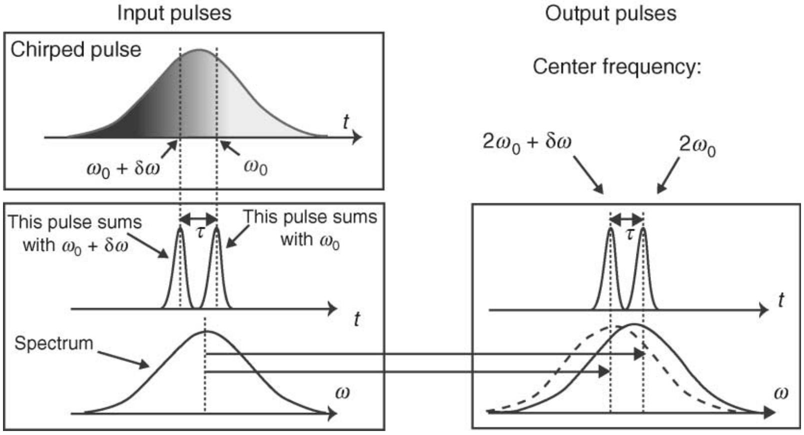
Input pulses
Chirped pulse
t
$\omega_0 + \delta\omega$
ω_0
This pulse sums with $\omega_0 + \delta\omega$
τ
This pulse sums with ω_0
t
Spectrum
ω
Output pulses
Center frequency:
$2\omega_0 + \delta\omega$
$2\omega_0$
τ
t
ω

Figure 11 Figure showing how a SPIDER device works. The highly chirped pulse has a frequency that varies with time. Mixing the chirped pulse with two time delayed replicas produces two pulses, separated in time, that have slightly different frequencies, but are otherwise identical. By interfering the two pulses in a spectrometer, the relative phase between $\omega_0 + \delta\omega$ portions of the pulse and ω_0 portions of the pulse are determined.

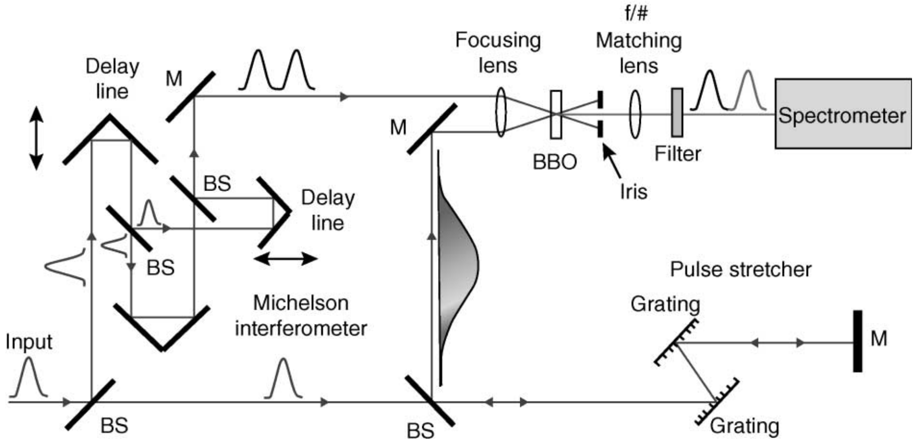

Figure 12 SPIDER apparatus. The input pulse is first split into two pulses. One pulse is stretched in a pulse stretcher. The other pulse is sent into a Michelson interferometer to produce two time-delayed pulses. The two time-delayed pulses are mixed with the stretched pulse to produce two pulses at a slightly different center frequency. The two pulses are sent into a spectrometer to produce a spectral interferogram. Analysis of the spectral interferogram yields the derivative of the pulse phase.

Because the pulses are slightly time delayed, each one aligns with a slightly different frequency in the chirped pulse. The chirped pulse must be so highly chirped that the pulse frequency does not change significantly over the duration of the pulse to be measured.

A schematic of a SPIDER device for the measurement of femtosecond laser pulses is shown in Figure 12. The input pulse is first split into two replicas. The first replica is sent to a pulse stretcher to produce to highly chirped pulse. The other replica is sent into a Michelson interferometer (or a simple étalon may sometimes suffice) to create two time-delayed replicas. The time-delayed replicas are mixed with the chirped pulse. Because each replica is mixed with a slightly different frequency in the chirped pulse, each replica produces a pulse with a slightly different center frequency. When the time-delayed replicas are spectrally resolved, each frequency of the pulse interferes with a frequency-shifted portion of the same pulse. Thus, the measured phase difference is proportional to the derivative of the phase of the original pulse.

Measuring Weak Pulses

Because virtually all the pulse measurement techniques, that do not require a reference pulse, use a nonlinearity, measuring weak ultrafast pulses directly is difficult. Fortunately, because most weak pulses are generated from a stronger pulse, spectral interferometry can be used to determine the relative phase between the weaker pulse and the stronger pulse (for regions where there is spectral overlap between the two pulses). If the stronger pulse is characterized using another pulse measurement technique, then the weaker pulse phase can be determined from the relative phase measurement and the known phase of the stronger pulse. For example, the Trebino group termed the combination of FROG and SI as TADPOLE (Temporal Analysis by Dispersing a Pair Of Light E-fields).

Which Pulse Measurement Method Should I Use?

Choosing a pulse measurement method depends on the required measurement as well as the precision and accuracy required. All the pulse measurement techniques have advantages and disadvantages, but excellent measurements have been made with all of them. Autocorrelation is adequate when all that is required is the approximate pulse duration without any phase information. Spectral interferometry is quite useful for measuring transient changes in a sample in pump-probe experiments. FROG is perhaps the most commonly used and general-purpose pulse measurement technique, providing a great deal of feedback about the quality of the pulse measurement. Inexpensive and convenient FROG devices, together with software, are available commercially, adding to their acceptance. While SPIDER is experimentally complex and does not have the cross checks that FROG has, it provides the quickest answer, requiring no iterative phase retrieval algorithm to obtain pulse intensity and phase.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant numbers DMI-9801116 and DMI-0091454. The author would also like to acknowledge Rick Trebino for the use of figures from his lectures.

List of Units and Nomenclature

$A(t)$	Full electric field
$E(t)$	Complex electric field without carrier frequency
$E_{\mathrm{sig}}(t, \tau)$	Complex electric field of the FROG trace versus time and time delay
$E_{\mathrm{sig}}(t, \Omega)$	Fourier transform of $E_{\mathrm{sig}}(t, \tau)$ with respect to τ
$\tilde{E}(\omega)$	Complex electric field in the frequency domain
$f(t)$	Correlation product
$f(\omega)$	Fourier transform of the correlation product
$g(t - \tau)$	time gate or filter
G	FROG trace error
$h(\omega - \Omega)$	frequency gate or filter
$I(t)$	Intensity profile of the electric field in the time domain
$\tilde{I}(\omega)$	spectral intensity
$I_{\mathrm{FROG}}(\omega, \tau)$	Intensity of the FROG trace
$\Theta(t)$	Heaviside function $= 0$ $t < 0$, $= 1$ $t \geqq 0$
$\phi(\omega)$	Frequency domain phase
$S_{\mathrm{E}}(\omega, \tau)$	Spectrogram
$S_{\mathrm{E}}(\Omega, T)$	Sonogram
$S(\omega)$	Spectral interferogram
t, T	variables for time
τ	Fourier transform
v_{g}	group velocity
$\varphi(t)$	Time domain phase
Δw	Beam waist
Z	Distance metric between the FROG signal field generated by the data constraint and the FROG signal field generated by the mathematical form constraint
Φ	Geometric angle between two beams
$\Delta\phi$	Phase difference
$\Pi^{(n)}$	nth order nonlinear susceptibility
ω	Primary variable for frequency
$\Delta\omega$	Difference in center frequency between two pulses
Ω	Secondary variable for frequency
$\Omega(t)$	Instantaneous frequency

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels JC and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena*. San Diego, CA: Academic Press Inc.

Dorrer C, Belabas N, Likforman J-P and Joffre M (2000) Spectral resolution and sampling issues in Fourier-transform spectral interferometry. *Journal of the Optical Society B* 17: 1795.

Kane DJ, Weston J and Chu K-CJ (2003) Real-time inversion of polarization gate frequency-resolved optical gating spectrograms. *Applied Optics* 42: 1140–1144.

Iaconis C and Walmsley IA (1999) Self-referencing spectral interferometry for measuring ultrashort optical pulses. *IEEE Journal of Quantum Electronics* 35: 501–509.

Lepetit L, Cheriaux G and Joffre M (1995) Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. *Journal of the Optical Society of America B* 12: 2467–2474.

Rulliere C (ed.) (2000) *Femtosecond Laser Pulses*. New York: Springer Verlag.

Shuman TM, Anderson ME, Bromage J, Iaconis C, Waxer L and Walmsley IA (1999) Real-time SPIDER: ultrashort pulse characterization at 20 Hz. *Optics Express* 5: 134–143.

Treacy EB (1971) Measurement and interpretation of dynamic spectrograms of picosecond light pulses. *Journal of Applied Physics* 42: 3848–3858.

Trebino R (2000) *Frequency Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses*. Boston, MA: Kluwer Academic Publisher.

Trebino R, DeLong KW, Fittinghoff DN, Sweetser JN, Krunbugel MA, Richman BA and Kane DJ (1997) Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. *Review of Scientific Instruments* 68: 3277–3295.

Various authors (1999) Feature issue on ultrashort-laser-pulse measurement and applications. *IEEE Journal of Quantum Electronics* 35: 4.

ULTRAFAST TECHNOLOGY

Contents

Femtosecond Chemical Dynamics: Gas-Phase
Femtosecond Condensed Phase Spectroscopy: Structural Dynamics
Ultrafast Illumination and Processing

Femtosecond Chemical Dynamics: Gas-Phase

M Dantus, Michigan State University, East Lansing, MI, USA

E J Brown, McDaniel College, Westminster, MD, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Chemical reactions are at the heart of chemistry; the quest to understand them has sparked research and study along many different avenues. The fundamental dynamics involved in bond breakage, bond formation, and transition states take place on a very fast time-scale, usually in the tens to hundreds of femtoseconds, where a femtosecond (fs) is equal to 10^{-15} seconds. Therefore, ultrashort laser pulses are required to measure these fundamental processes as well as the rotational and vibrational motion before and after the reaction. The whole set of measurements can then be used for mapping a potential energy surface for the reaction. As a result of these studies, one learns the key details that are required to develop a picture of how chemical reactions occur. The measurement of these fundamental dynamics, using ultrashort laser pulses, is the subject of this article.

To illustrate how long it takes to form or break a chemical bond, one needs to consider the following analogy. The time it takes to break a chemical bond is approximately a tenth of a trillionth of a second (10^{-13} s). During such a short time, light can travel only 0.03 mm in vacuum. This time is so short that it cannot be measured electronically. Its measurement requires the use of ultrashort laser pulses, with pulse durations of 50 femtoseconds or shorter. The development of methods aimed at measuring the fundamental steps involved in chemical reactions culminated in the development of femtosecond transition state spectroscopy (FTS), a method developed in the 1980s by Zewail and co-workers at Caltech. This work allows researchers to observe bond breakage and formation, vibrational and rotational dynamics, and transition state configurations, as the reaction occurs. The Nobel Prize in Chemistry was awarded in 1999 to Ahmed Zewail for this development. Since the 1980s, femtosecond lasers have been used to study molecular dynamics and chemical dynamics on very fast time-scales. The field, known as femtochemistry, has grown quickly and diversified to involve hundreds of scientists around the world. This article discusses the measurement of ultrafast dynamics measured in gas-phase samples. The presentation encompasses some of the measurement techniques and the fundamental values to be measured and concludes with an outlook towards the future.

From Nanosecond to Picosecond to Femtosecond Reaction Kinetics

The quest for shorter light pulses to study chemical reactions predates the invention of the laser. In the early days, electric discharges were used to measure fast dynamics in the millisecond (10^{-3} s) and then the microsecond (10^{-6} s) time-scales. With the invention of the laser, measurements improved drastically with temporal resolution going from the nanosecond (10^{-9} s) to the picosecond (10^{-12} s) time-scales. Since the early days, it was recognized that the best time resolution was obtained when pairs of light pulses were used – one to initiate the reaction and the second to probe it. These pump-probe measurements have progressed with the laser technology almost to the single femtosecond level.

The measurements being considered here go beyond the measurement of fast reaction kinetics. When a chemical reaction occurs in a beaker, one usually measures a statistical rate of reaction that involves diffusion of reagents in a solvent. These kinetic measurements, which involve an ensemble of molecules, are much slower and do not reveal the fundamental steps of the reaction such as bond formation. The introduction of ultrafast pulses allowed, for the first time, the measurement of the motion of the atoms in real time, as if the motion was frozen by ultrafast flash photography. Capturing this motion helps to reveal the internal forces, described by a potential energy surface, that act on the atoms in the molecule during the chemical reaction. This concept is illustrated in Figure 1, where we see the reagents and the products for the reaction. Typically, the starting and ending point of a chemical reaction are very well known and the compounds are very well characterized. However, how the reagents become the products is not always known. Chemists typically deduce a reaction mechanism based on a large body of experiments where different parameters including the structure of the reagents are modified to evaluate their effect on the outcome of the reaction. Unfortunately in most cases, reaction mechanisms, even when consistent with all available experimental evidence, may not accurately reflect how a particular chemical reaction takes place.

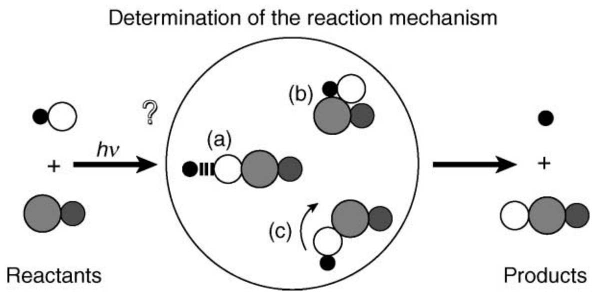

Figure 1 A sketch of a chemical reaction, AB + CD → A + BCD. The reactants and products are well known, but it is not known how they change from reactants to products. This mechanism may involve a transition state which is (a) a linear complex with vibrational motion, (b) a nonlinear complex, (c) a bent complex with rotational motion, or many other possibilities and combinations. The elucidation of the mechanism and transition state dynamics are the goals of femtosecond time-resolved studies.

The ideal method to make the determination involves direct observation. As mentioned earlier, direct observation requires femtosecond laser pulses, just like a fast camera shutter is required to take pictures of fast-moving objects. The concept of a femtosecond pump-probe measurement is illustrated in Figure 2. The chemical reaction is initiated at time $t = 0$ fs by the pump laser. The pump laser provides the energy required for initiating the chemical transformation. The probe laser, delayed in time, probes the formation of the product. Figure 2 illustrates a number of pump-probe measurements

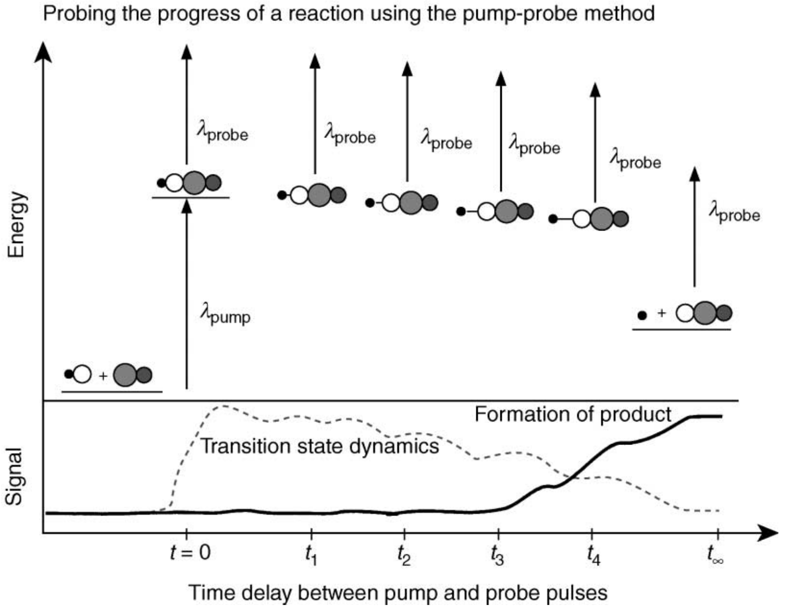
Probing the progress of a reaction using the pump-probe method
Energy
λprobe
λprobe
λprobe
λprobe
λprobe
λprobe
λpump
+
+
Signal
Formation of product
Transition state dynamics
t = 0
t1
t2
t3
t4
t∞
Time delay between pump and probe pulses

Figure 2 Energy schematic of the reaction of AB + CD → A + BCD assuming vibrational motion in a linear transition state. The pump pulse provides enough energy to excite the reactants to the transition state at $t = 0$. This transition state is probed by the second pulse. This pulse arrives at different time delays, from no delay ($t = 0$), to increasing delays (t_1, t_2,…), to infinite delay (t_∞). The signal varies depending on the time delay between the two pulses and these changes in signal intensity correspond to vibrational and rotational information about the transition state. To obtain a transient as shown at the bottom, the time delay between the pump and probe pulses is scanned. The dashed line shows the transient that would be obtained when the transition state is monitored; the solid line shows the transient that would be obtained when the product is monitored.

each obtained at a different time delay between the pump and probe laser pulses. When a series of pump-probe measurements are gathered as a function of time delay, one obtains a transient that contains a record for the evolution of the reagents to product as a function of time.

The Formation of Coherent Wave Packets and their Motion

The illustration shown in Figure 2 shows atoms in a molecule moving as classical particles. Because of their small size, atoms behave quantum mechanically and this has some implications on the observed motion. A prototypical pump-probe measurement is illustrated for this purpose in Figure 3. For resonant excitation, a photon is absorbed by a molecular system, originally in the ground electronic state (V_0). This absorption causes a vertical transition to an excited electronic state (V_1). Consistent with Heisenberg's uncertainty principle, a pulse that has a short temporal duration has a large spectral bandwidth; for most molecular systems and ultrashort pulses, the frequency spread of the pulse is wider than the vibrational energy spacing in the excited state. Thus, a number of quantum vibrational states are populated on V_1 by the absorption of a short laser pulse. When a short pulse excites a number of states, a coherent superposition of states is formed. One can understand the formation of a wave packet by simply realizing that in the short time the laser pulse excites the system there is no opportunity for the atoms in the molecule to move. For this reason the initial wave packet is very similar to the ground state atomic arrangement.

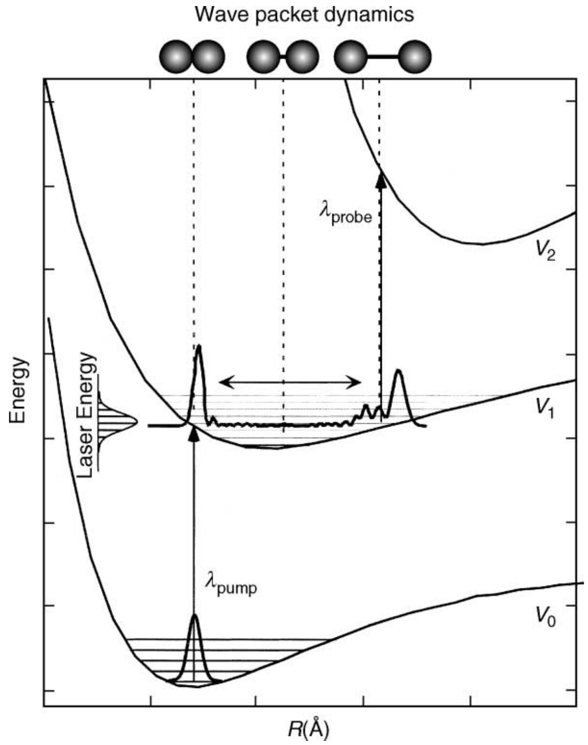

Figure 3 Pump-probe experiment. The pump pulses excite the molecular system from the ground state V_0 to excited state V_1. The large spectral bandwidth of femtosecond pulses allows for a number of vibrational levels in V_1 to be populated. A wave packet is formed that vibrates on V_1; the classical motion of the particle is shown as well at the top. In this case the energy of the probe pulse will excite the molecules to V_2 when the bond is stretched (right side of oscillation); however, when the bond is compressed, the energy is not resonant with any excitation and the probe pulse will not be absorbed.

When a short laser pulse excites a molecule to a bound electronic state, the resulting wave packet, $\Psi(t)$, is a sum of all the possible states φ_n in V_1 with amplitudes, a_n, $\Psi = \sum_n a_n \varphi_n$. The amplitudes, a_n, depend on the overlap between the initial (ground state) and final (excited state) spatial overlap. This overlap indicates that the excitation process is faster than the motion of atoms in a molecule. Based on this approximation for the overlap integral, the amplitudes are replaced by the Franck–Condon overlaps to obtain $\Psi = C\sum_n \langle \varphi_n | \varphi_0^{(0)} \rangle \varphi_n$. The wave packet on V_1 evolves in time according to $\Psi(t) = \sum_n a_n \exp[-\mathrm{i}E_n t/\hbar]\varphi_n$, where E_n is the energy that corresponds to the state φ_n and $\hbar$ is Planck's constant divided by 2π. When a pulse is very short in time, the coherent superposition of states closely resembles the initial distribution of atoms in the molecule. The motion from inner to outer turning point of the quantum mechanical wave packet is illustrated in Figure 3. One can imagine this wave packet as a classical particle oscillating on the excited potential energy curve with the periodic nature of the molecular vibrations. When the pulse is very long and only one state is populated, the dynamics can no longer be observed, as is the case with conventional frequency-resolved spectroscopy.

Molecular dynamics can be inferred from high-resolution frequency-resolved spectroscopy; however, the spectra can become congested and the data analysis becomes more complicated for large molecules, for high temperatures, or in solutions. Molecular dynamics are obtained from frequency-resolved spectra by taking the Fourier transform of the data. It seems more intuitive to consider these dynamics in the time domain where the dynamics can be observed directly as 'snapshot' taken by the femtosecond pulses. Ultrafast time-resolved spectroscopy has other advantages – high peak intensities allow for multiphoton excitation and other nonlinear processes, the selection of the detection wavelength discriminates between different species and results in an increased signal-to-noise ratio for intermediate short-lived species, collisions take place on a longer time-scale and do not affect these measurements, and the characteristics of the ultrafast laser pulse (such as phase, frequency components, and chirp) can be manipulated to control the dynamics and the yield of chemical reactions.

Methods for Measuring Femtosecond Dynamics

Pump-Probe Method

Pump-probe techniques use a pair of pulses separated in time, one to initiate the reaction and the second to determine its progress. The first femtosecond pulse with wavelength λ_{pump} excites a molecule from its (bound) ground electronic state (V_0) to an excited (V_1) state. After a variable time delay, a second femtosecond pulse with wavelength λ_{probe} causes a transition to another excited state (V_2) as shown in Figures 3 and 4.

The progress of the reaction dynamics is followed through the measurement of signal resulting from the absorption of the probe pulse by the system. There are different detection modes depending on the potential energy surfaces reached by the pump and probe pulses. In one case (Figure 4a), the first excited state (V_1) is nonfluorescent but the second one (V_2) fluoresces and this fluorescence is measured. In another case (Figure 4b), state V_1 is fluorescent but state V_2 is not; in this case the depletion of the fluorescence by the probe is measured. Finally (Figure 4c), excited states V_1 and V_2 may be dissociative; therefore, the fluorescence detected is emitted by the products. In any of these cases, as the wave packet on the intermediate state propagates (e.g., vibrates, dissociates), the transition probability varies (i.e., the ability of the probe pulse to be absorbed by the molecule changes). Therefore, the amount of laser-induced fluorescence (LIF) emitted varies with the pump-probe time delay. In some cases, the probe pulse produces ions and there are a number of methods to detect the electrons or the ions with great sensitivity.

Measurement of rotational dynamics

Measurement of rotational dynamics requires a vectorial frame of reference and this is achieved with linearly polarized laser beams. Excitation by the pump pulse selects an initial population distribution described by $\cos^2\theta$ where θ is the angle between the transition dipole and the pump laser polarization vector. As the rotational wave packet dephases due to the distribution of the population in different rotational levels, the probe pulse probes the transient alignment of the molecules in space. As the molecules rotate away from the initial alignment, the signal decreases. In Figure 5 rotational dynamics are depicted schematically. First, a single (classical) molecule is considered (top panel). As the molecule rotates, its dipole changes from being aligned with the polarization vector of the laser field

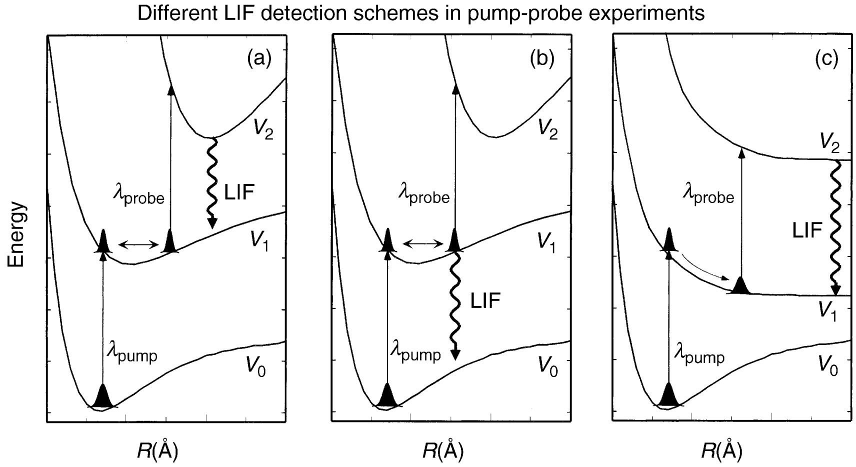
Different LIF detection schemes in pump-probe experiments
(a)
(b)
(c)
Energy
V_2
V_1
V_0
λ_{probe}
λ_{pump}
LIF
R(Å)
R(Å)
R(Å)

Figure 4 Different pump-probe experiments.(a) Bound-to-bound transitions from $V_0 \rightarrow V_1$ and from $V_1 \rightarrow V_2$. The V_2 state fluoresces and this increase in LIF from V_2 is measured as signal. (b) Bound-to-bound transition from $V_0 \rightarrow V_1$ where V_1 fluoresces. When the probe pulse causes an excitation from $V_1 \rightarrow V_2$, a loss of LIF signal is observed. The depletion of fluorescence is measured in this case. (c) Bound-to-repulsive surface transition from $V_0 \rightarrow V_1$. The energy of the probe causes an excitation to V_2 at a particular molecular distance. Fluorescence of a product fragment is measured.

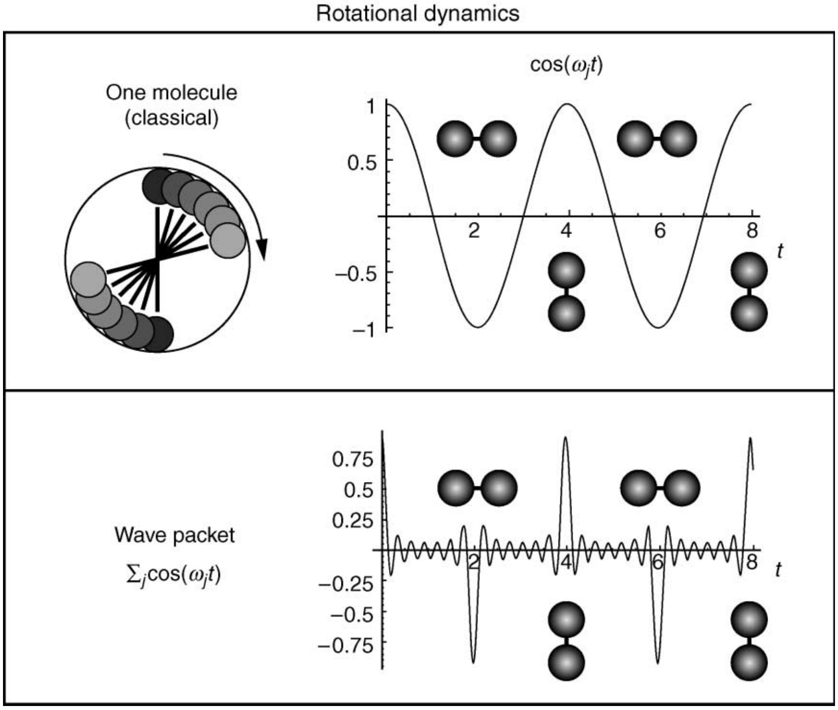

Figure 5 Rotational motion of molecules. Excitation by a linearly polarized laser pulse initially selects molecules that are well aligned ($\cos^2\theta$) with its polarization. Considering only one molecule evolving in time (top), the rotational motion in the molecule causes it to move away from the initial alignments and then eventually return to it. The signal will decrease when the molecule is not aligned with the polarization of the probe laser and will increase when aligned well. The signal looks like a cosine function that depends on the rotational frequency of the molecule (ω_j). For a rotational wave packet, the rotational dynamics are similar except that a number of rotational levels are populated; thus, the signal depends on the summation of many cosine functions and the features in the data become sharper.

(assumed vertical) to being perpendicular. The rotational motion can be described by a cosine function. The rotational frequency ω_j is proportional to the rotational level and the rotational constant of the molecule. Because the rotational constant is inversely proportional to the angular momentum, larger molecules rotate slower than smaller molecules. When several rotational levels are contained in a rotational wave packet the motion produces sharper features (bottom panel). The sharp features are called rotational recurrences or revivals. Note that the opposite trend is observed if the probe pulse is polarized perpendicular to the pump pulse; the signal increases as the molecules rotate away from the initial alignment and decreases as the alignment recurs. The time-resolved data for each polarization arrangement (parallel and perpendicular) contain both an isotropic component (vibrational motion) and an anisotropic component (rotational motion). In rotational anisotropy measurements, data from these two polarization arrangements are manipulated mathematically to yield just the rotational motion; rotational dynamics and rotational information such as the spectroscopic rotational constants and the rotational energy partitioning in the reactants and products can be extracted from these measurements.

Three-Pulse Four-Wave Mixing Method

Some methods are more complicated than pump-probe technique but they are more powerful. Three-pulse four-wave mixing (FWM) is one such method and is briefly presented here. Time-resolved four-wave mixing techniques can be understood based on the formation of transient gratings by two of the incident lasers. When two laser beams cross, the spatial modulation of their electric fields varies due to constructive and destructive interference as shown in Figure 6. The molecules in the interaction region experience varying electric field intensities according to their position and this leads to the formation of a transient grating of polarized molecules in space. The transient grating formation can be probed easily by the detection of Bragg scattering of a third laser beam, also known as homodyne detection; here, the signal beam is scattered in the phase-matching direction $\mathbf{k}_{\text{signal}}$ as indicated in Figure 6a. Alternatively, probing can be achieved by detection of changes in the intensity of a fourth beam that propagates in the

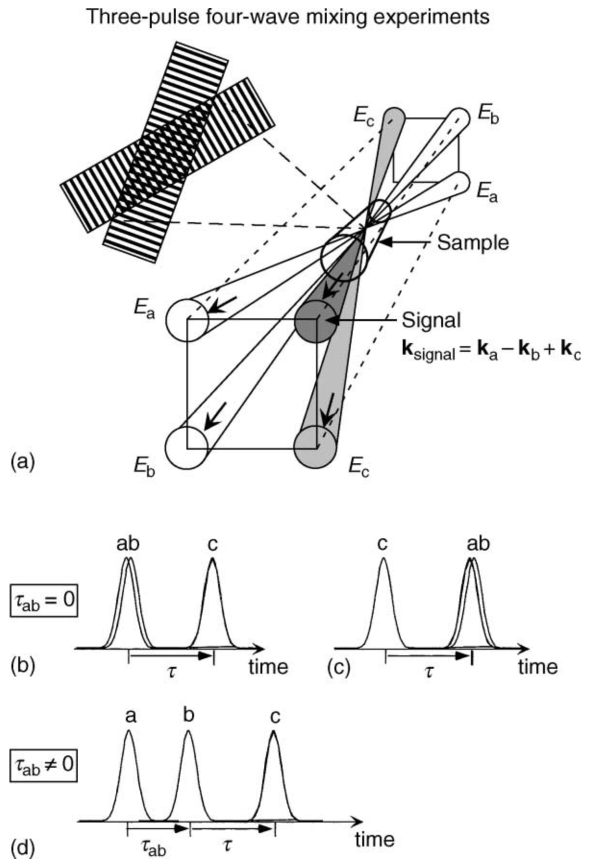

Figure 6 (a) FWM experimental configuration. Two of the beams cross which forms a grating and the third beam scatters off the grating to generate a fourth beam (the signal). The three laser pulses are noted as electric fields E_a, E_b, and E_c. These three laser pulses can be delayed or advanced in time relative to each other giving rise to a number of different pulse sequences. (b) Pulse sequence with fields E_a and E_b overlapped in time ($\tau_{ab} = 0$). Field E_c follows at a time delay τ. (c) Pulse sequence in which E_c precedes E_a and E_b by time delay τ. For both of these cases, τ is scanned in time to obtain the transients. (d) Pulse sequence in which no pulses are overlapped in time and are applied in the order E_a, E_b, and E_c. Time delay τ_{ab} can be set to specific values while τ is scanned in time.

same direction as the signal; this is called heterodyne detection. The medium that constitutes the gratings is composed of the sample molecules, which in the gas phase move freely. Therefore as time evolves, the molecular dynamics (rotation and vibration) modulate the transient grating and hence the signal.

Three-pulse four-wave mixing is a nonlinear spectroscopic method that combines the interaction of three laser pulses in a phase-matched geometry with a well-defined time sequence of the pulses. The signal from these measurements arises from a third-order polarization resulting from the interaction of the three electric fields. Three-pulse FWM is similar to the pump-probe technique in that a preparation step is followed, after some variable time delay, by a probing step. However, three-pulse FWM allows for a greater degree of control over the preparation and probing processes.

When the lasers are in resonance with an electronic state, FWM measurements can provide dynamic information from the ground and excited electronic states. When two pulses are incident on the sample at $t = 0$ and a third pulse probes the dynamics at a later time τ (as in Figure 6b), the measurements provide information that is very similar to that from a pump-probe measurement. However, having three pulses allows for other different setups. For example, when one pulse arrives at the sample first and the other two pulses follow after a time delay (as in Figure 6c), the resulting signal provides dynamic information from the excited state. This type of measurement also provides a measurement of the electronic coherence of the sample, i.e., for how long the molecules remember the phase of the incident laser pulses. This information is important if one wants to take advantage of the phase of laser pulses to cause quantum mechanical interference in the molecular sample. These measurements are known as photon echo and are much more common in condensed-phase samples. There are numerous pulse sequences possible; only three examples are shown in Figure 6.

FWM measurements do not depend on resonance excitation to an electronic excited state. In the absence of such resonance, the grating is formed in the ground state and ground state vibrational and rotational dynamics are observed. These measurements, also known as transient grating (TG), depend on the impulsive Raman scattering process. By changing the time delay (in Figure 6b) between the electric fields that generate the grating and the third (probe) beam, the signal is modulated by the molecular dynamics. Analysis of the time-dependent signal yields the rotational and vibrational spectroscopic values from the sample being studied.

Studies of Reaction Dynamics

Dissociation on a Repulsive Surface

In 1987, Zewail and co-workers studied the photodissociation of cyanogen iodide (ICN) using femtosecond lasers; the first direct observation of a chemical reaction from reactants to products was conducted. The excitation and decay of the reactants (ICN), the formation and decay of the intermediates (I–CN), and the formation of the products (I + CN) were observed in this experiment.

The pump pulse (306 nm) excited ground state cyanogen iodide to a repulsive excited state of ICN which dissociated into I and CN fragments (both in the ground state). The probe pulse when set to 388.5 nm was resonant with an excitation of the CN fragment; therefore, the reaction was monitored by detecting the laser-induced fluorescence (LIF) of this CN^* fragment using a probe pulse at different time delays (as shown in Figure 4c). The wavelength of the probe pulse could be detuned to different wavelengths for making observations at different lengths of the I–C bond. The results of these experimental studies showed that ICN dissociated in about 200 fs and that the transition state lived for only about 50 fs. The data from the wavelength detuning studies were used to map the excited state potential of ICN both in time and in space; this data analysis is called inversion to the potential.

A schematic representation of the ICN experiment is shown in Figure 7. The potential energy curves are shown in Figure 7a. The pump laser initiates the chemical reaction. Three different probe wavelengths are depicted, the first probing early in the reaction, the second at an intermediate distance, and the final one probing the free products. Notice that the wavelength of the probe determines the location in the potential energy curve that is probed. This concept allows one to study reagents as they are transformed into products. The signals obtained from the three different probe wavelengths are shown in Figure 7b.

Reactions Involving Crossing of Potential Energy Surfaces

For many chemical reactions, the progress of the reaction requires changes of the electronic state. This can take place where potential energy surfaces cross. The best-known example of reaction dynamics involves the experiment on the photodissociation of NaI by Zewail and co-workers. This experiment is depicted in Figure 8. The pump pulse excites the wave packet from ground state (ionic) to an excited state (covalent) where it begins to oscillate. The ground and excited states overlap with an avoided crossing in this molecule (ionic versus covalent nature, as shown in Figure 8a); when the wave packet approaches this avoided crossing, it splits into two partial waves. One wave stays on the excited state while the other ‘escapes’ onto the dissociative surface and separates into Na and I. The partial wave on the excited state again oscillates and then splits at the avoided crossing and the pattern continues. The time delay of the probe pulse shows the oscillations of the wave packet on the excited surface when detecting the activated complex $[NaI]^{*\ddagger}$ (thin line), while the data of the product fragments (free Na atoms detected by using the D-line absorption) (thick line) show a step-wise accumulation in the signal corresponding to the sequential formation of products during each oscillation near the avoided crossing region of the surfaces (see Figure 8b).

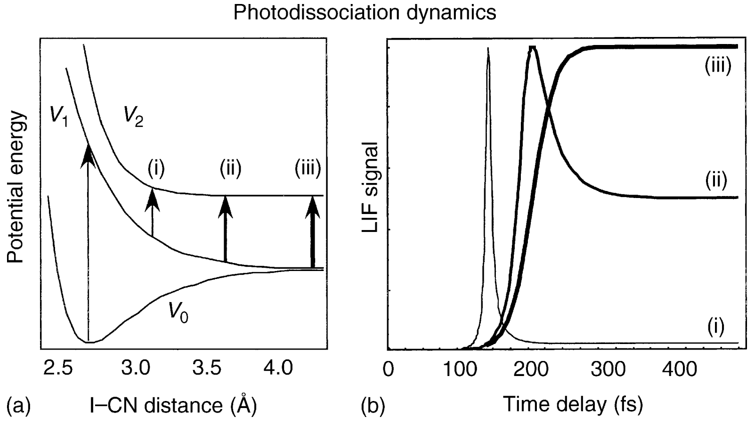

Figure 7 Photodissociation of ICN. (a) Potential energy surface schematic showing the initial excitation by the pump laser to a dissociative surface. The reaction could be monitored at different I–CN distances by using different probe wavelengths as shown by three examples here, (i), (ii) and (iii), from shorter to longer distances. (b) At short distances (long probe wavelengths), the signal is observed at shorter time delays only. The signal decreases at later time delays because the internuclear distance continues to lengthen as the dissociation occurs over time. When the distance increases as monitored by shorter probe wavelengths, the signal rises to a maximum at longer time delays reflecting the dissociation of the reaction in progress. At ‘infinite’ I–CN distance (i.e., the molecule is dissociated and only free I atoms and CN fragments remain), the signal reaches a maximum and remains there because the reaction is complete.

Chemical Dynamics with Multiple Pathways

The pump-probe concept can be used to probe more complex reactions. To date, a great number of different chemical reactions have been studied (see Further Reading). Among the most important reactions we list ring opening, isomerization, proton transfer, and concerted chemical reactions. We illustrate these measurements with experiments on the photodissociation of dihalogenated alkanes. Upon high-energy excitation (12 eV), the methylene iodide (CH_2I_2) molecule dissociates and forms molecular iodine in excited states D′ and f as well as producing I atoms (main channel). The detection wavelength used in the experiment can distinguish between these products; D′ fluorescence can be collected at 342 nm and f fluorescence is collected at 272 nm. The Dantus group has studied the dynamics of this type of concerted molecular elimination reaction extensively. The photodissociation reaction occurs by a multiphoton excitation (three-photons from the pump beam at 310 nm) reaching a dissociative state where the CH_2 fragment and excited molecular iodine are formed. The 620 nm probe pulse depletes the fluorescence from the excited I_2 molecule (either D′ or f state) as shown in Figure 4b; this fluorescence depletion is measured and analyzed to determine the dynamics of the reaction. The object of this experimental work was to determine if the photodissociation mechanism for forming molecular iodine followed a step-wise, synchronous concerted, or asynchronous concerted mechanism as shown in Figure 9.

The time transients reveal coherent vibrational motion of the I_2 product. Analysis of the transition state dynamics shows that the breaking of the carbon–halogen bond and the formation of the halogen–halogen bond occurs within 50 fs. Further, the iodine molecule is formed with a very hot rotational distribution indicating that the symmetry of the molecule is broken during the process; the torque imparts a large amount of rotational motion. Overall, the results of these studies show that the photodissociation process of CH_2I_2 follows an asynchronous concerted reaction mechanism and corresponds to a symmetry barrier that exists in C_{2v} molecules preventing the formation of the interhalogen bond. The existence of this symmetry barrier was explored by using CH_2ICl, a molecule that does not have C_{2v} symmetry; as expected, the halogen molecule (ICl) was formed with a cold rotational distribution indicating that it was able to follow a synchronous concerted mechanism. The dissociation times were also studied on the family of compounds CX_2Y_2 and R–CHI_2 to investigate the influence of thermodynamics and reduced mass changes on the reaction.

The Dynamics of Bond Formation

Most chemical reactions that are studied with femtosecond spectroscopy are unimolecular reactions (photodissociation); however, bimolecular reactions (photo-association) are interesting and chemically relevant in trying to understand how chemical reactions occur. In the photo-association process, a pair of atoms or molecules cooperatively absorbs the photon undergoing a free-to-bound transition and forming a bond. When the ultrashort pulse initiates the reaction in femtosecond photo-association spectroscopy, a clearly defined initial time for the reaction is given

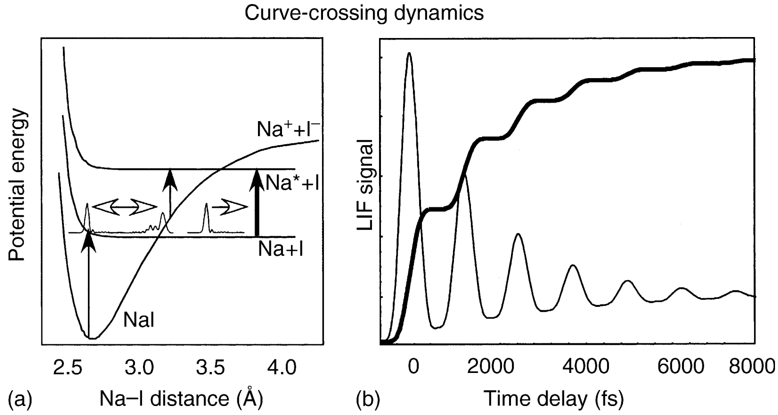

Figure 8 Photodissociation of sodium iodide. (a) Potential energy curve schematic showing the curve-crossing between the ground (ionic) and excited (covalent) states. As the wave packet oscillates on the excited curve, it splits into two wave packets at the curve crossing. (b) The thin line shows the vibrations of the excited complex as it oscillates on the excited surface. The signal intensity decreases because some of the molecules have dissociated to form Na and I atoms when the wave packet splits. The thick line shows the stepwise increase in free Na atom signal each time the wave packet oscillates and splits.

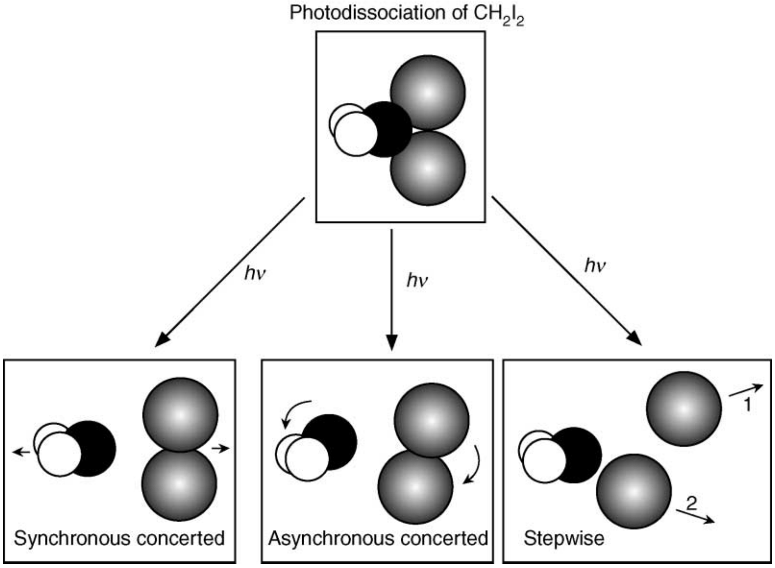

Figure 9 Possible mechanisms for the photodissociation of CH_2I_2 following multiphoton excitation. The molecule could break both C–I bonds and form an I–I bond all at the same time (synchronous concerted mechanism); this would result in very little rotational motion in the products. The molecule could begin to break one C–I bond slightly before breaking the second one and forming the I–I bond, but all three events still occur within the duration of the laser pulse (asynchronous concerted mechanism). This would result in significant rotational motion in the products. The molecule could also lose one I atom and at some later time lose the other I atom; these I atoms could collide later to form I_2 (stepwise mechanism).

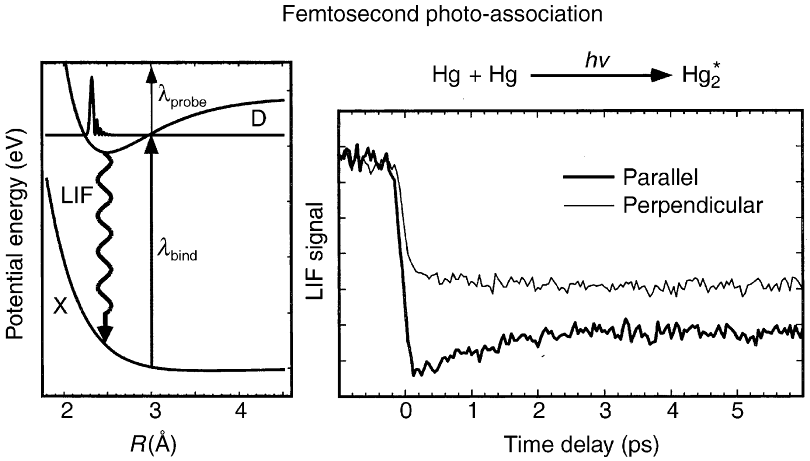

Figure 10 Photo-association of mercury dimers. The pump (bind) pulse is absorbed by two ground-state Hg atoms to form Hg_2^* on the excited state which is fluorescent. The probe pulse depletes this fluorescence causing the LIF signal to decrease at $t = 0$. Analysis of the transient reveals that the time of association is <60 fs. Parallel and perpendicular pump-probe data were examined to obtain the rotational distribution of this reaction.

and the alignment conditions of the reaction can be examined by taking advantage of the polarization of the lasers. The femtosecond photo-association spectroscopy technique was used by the Dantus group to study the formation of an excited mercury dimer. Their experiment is depicted in Figure 10. Two ground-state mercury atoms absorbed the pump (bind) pulse (at 312 nm) forming a wave packet in the excited D state. The probe pulse depleted the fluorescence from the newly formed molecules. The time of association was determined to be within the laser pulse duration (<60 fs). The dimers were found to have a narrow rotational distribution which together with the wavelength of the binding pulse reflects the impact parameter (geometry) of the collision. Calculations reveal that the ‘optimal’

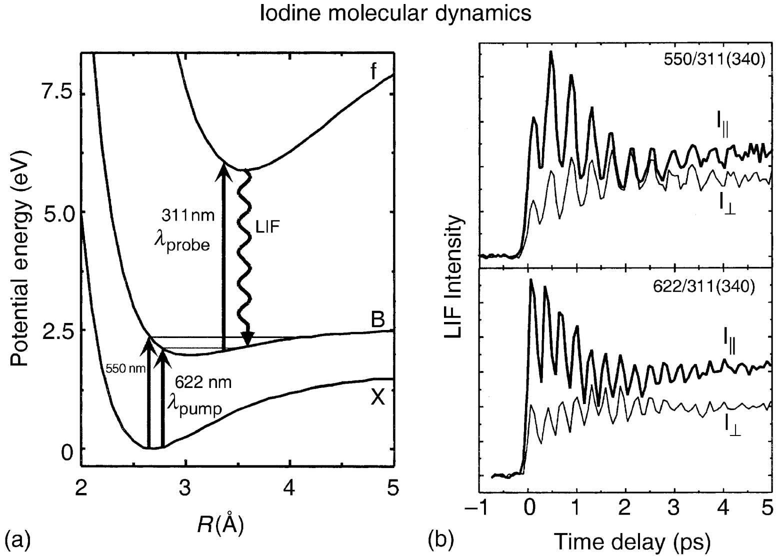

Figure 11 Molecular dynamics of iodine using two different pump laser wavelengths. (a) The pump laser, either at 550 or 622 nm, causes an excitation from the ground state to the B excited state. The choice of wavelength determines the vibrational levels which are excited on the B state. In either case, the probe at 311 nm causes an excitation to the f excited state which fluoresces around 340 nm. (b) Parallel and perpendicular pump-probe transients were obtained using each pump wavelength. At higher pump energies (550 nm), the molecule vibrates more slowly (higher vibrational level) and is reflected by the 410 fs vibrational period observed in the data. At lower pump energies (622 nm), the molecule vibrates faster and corresponds to the 300 fs vibrational period observed in the data.

binding pulse for this association reaction is 10 fs; such short pulses would overlap the entire absorption band from the repulsive ground state to the bound D state.

Bimolecular reactions can also be studied by preparing a low-temperature cluster that contains the reagents of interest. For example, a cluster containing one HI and one CO_2 molecule can be used to study the reaction $H + CO_2 \rightarrow OH + CO$. Initiation of the reaction takes place when the pump laser cleaves the H–I bond and the H atom attacks the CO_2 molecule present in the cluster. The reaction dynamics are probed as a function of time delay between pump and probe pulses as for other pump-probe experiments. The femtosecond dynamics of this reaction have been studied independently by Wittig's and Zewail's research groups.

Studies of Molecular Dynamics

I_2-Pump-Probe, Rotation and Vibration (Bound Surface)

Aside from observing chemical reactions, using ultrafast spectroscopy allows for studying the motion of isolated molecules to elucidate the internal motions of molecules themselves and to extract spectroscopic parameters. Iodine is one of the most frequently used molecules in femtosecond spectroscopy. Here we highlight only a small subset of the femtosecond studies on I_2 that have been conducted by Zewail's and other research groups. Using a pump beam at 550 nm, ground state I_2 is excited to the B state; probing with 310 nm excites the molecule to an even higher bound potential (f state) from which fluorescence at 340 nm can be obtained (as shown in Figure 11). By changing the pump-probe time delay, oscillations on the B excited state are clearly visible and show a vibrational period of approximately 410 fs. Excitation at 622 nm produces a wave packet lower in the potential well, with a vibrational period of approximately 300 fs. The differences observed for parallel and perpendicular probing correspond to the early rotational motion of the molecules. Fourier transfer analysis of the data yields the energy of each vibrational state that forms part of the vibrational wave packet. Different pump wavelengths can be used to access different vibrational levels such that vibrational frequencies and rotational constants can be obtained for a wide range of vibrational levels of the potential. This information can then be used to refine the B excited state potential energy curve of iodine.

Different Measurements Possible with Three-Pulse FWM

Ground and excited state molecular dynamics

Molecular dynamics can also be examined with four-wave mixing techniques. Once again we use iodine as a model system to illustrate that unlike the previous experimental transients of iodine obtained with pump-probe measurements, both excited and ground state dynamics of I_2 are revealed with FWM. Materny, Kiefer, and co-workers have carried out coherent anti-Stokes Raman spectroscopy and degenerate four-wave mixing measurements on this molecule using two pulses overlapped in time followed by a third pulse after a time delay; both ground and excited state dynamic are observed in the Fourier transform of data. Dantus and co-workers explored all the pulse sequences possible with three-pulse FWM using molecular iodine as a model system (see below).

Double-sided Feynman diagrams and third-order response functions can be used to describe the interaction of the three pulses with the molecular system in three-pulse four-wave mixing. These mathematical tools are described in several books on nonlinear optical spectroscopy (see Further Reading). Figure 12 illustrates how different pulse sequences reveal different molecular dynamics. The transient grating method (Figure 12a) shows both ground (160 fs oscillations) and excited state (310 fs oscillations) wave packet dynamics. When data collection occurs over long time delays, the transient grating signal persists with a similar time constant to that of the lifetime of the excited state. The photon echo pulse sequence (shown in Figure 12b) reveals only excited state dynamics (310 fs oscillations). Note that at longer time delays, the signal decays exponentially with a time constant that reflects the rate of electronic coherence dephasing. This fundamental decay parameter indicates how long molecules remember the phase of the electric field that acted on them.

Off-resonance transient grating molecular dynamics

Off-resonance transient grating can be used to study the molecular motion of molecules in the ground state. The pulse sequence used is shown in Figure 6b. With this method it is possible to examine the rotational and vibrational dynamics of diatomic and polyatomic molecules. The vibrational dynamics belong to the Raman-active modes. From the rotational recurrences, one can obtain rotational constants with great accuracy. Dantus and co-workers have studied carbon dioxide at room temperature and in flames. At high temperatures, rotational recurrences narrow because more rotational levels are populated and the rotational recurrence moves towards longer times because of centrifugal distortion. Further studies on carbon dioxide and carbon disulfide have shown that strong fields can impart an increase in initial alignment (due to torque along the polarization vector of the laser) as well as molecular deformation such as bending (due to electronic state mixing). Heterodyne detected off-resonance transient grating data for CO_2 at room temperature are shown in Figure 13. The data closely resemble the schematic illustration in Figure 5 where rotational dynamics were introduced. The data in Figure 13 have been analyzed to yield the B and D rotational constants which give the structural parameters of the molecules and the centrifugal distortions, respectively. The type of analysis given here for CO_2 can be done on molecules

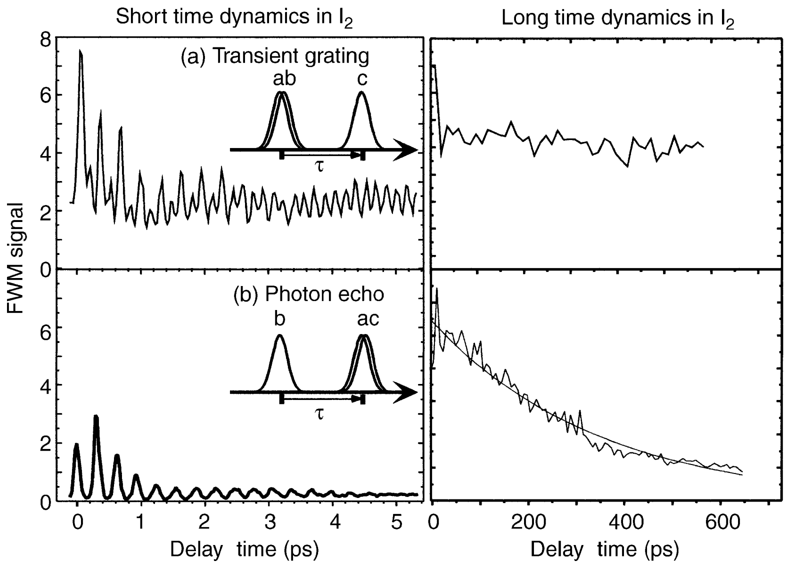

Figure 12 FWM signal in molecular iodine using two pulse sequences at short and long delay times. (a) Transient grating – E_a and E_b overlapped in time followed by E_c. Similar amounts of ground (208 cm^{-1}) and excited (105 cm^{-1}) state dynamics are observed with short delay times. In the transients, these correspond to oscillations of 160 fs and 307 fs respectively. (b) Photon echo – E_a and E_c overlapped in time but preceded by E_b. Excited state dynamics are observed in the transient with short delay times. At long time delays, the decrease in the signal reflects (a) the lifetime of the excited state or (b) the rate of the electronic coherence dephasing.

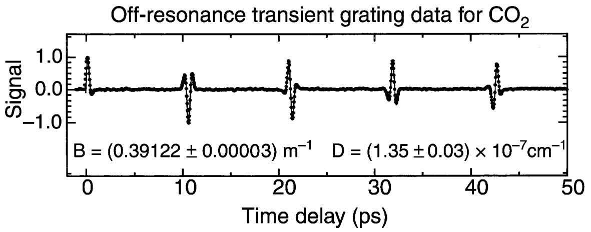

Figure 13 Heterodyne detected off-resonance transient grating on carbon dioxide at room temperature. The experimental data (dots) reflect the rotational motion of the CO_2 molecule. The data have been analyzed (the fit is shown as a line that goes through the experimental data points) to obtain the rotational constant (B) and the centrifugal distortion constant (D).

and molecular fragments resulting from chemical reactions. This method can also be used for molecular identification.

Perspectives for the Future

Most of this article has focused on the observation of ultrafast dynamics involved in chemical reactions. Presently, it is possible to generate pulses with durations as short as 5 fs. Such short pulses are capable of observing motion involving frequencies near 3000 cm^{-1}. This implies that one can observe the dynamics of most chemical bonds. The availability of shorter pulses will allow one to observe electronic dynamics. Advances in laser technology have also produced ultrashort pulses in the infrared region as well as in the vacuum ultraviolet and X-ray region. The wide spectrum available permits experiments that follow specific chemical bonds or specific transitions in elements. This in turn gives a wide range of possibilities for probing ultrafast chemical reaction dynamics.

Having pulses that are shorter than the duration of the dynamics they probe allows one to consider influencing the evolution of a chemical reaction as it is taking place. Controlling chemical reactions with lasers is not a new theme; however, the availability of ultrashort laser pulses has given new power to these efforts. Here we present a brief sample of some of the current efforts.

Control Using Three-Pulse Four-Wave Mixing

Coherent control of chemical reactions depends on the relative phase of two different laser pulses that interact with the sample in order to optimize the transfer of population between electronic states. The three-pulse four-wave mixi\ng technique can be utilized to attain coherent control; the specific timing between the pulses can be used to achieve near-unity or near-zero values of the transfer of population for molecules that interacted with the two electric fields. The Dantus group has also used three-pulse FWM to demonstrate control over optimizing the population transfer between two electronic states in iodine. The populations of the ground and excited states are controlled based on the time delay between the first two pulses τ_{ab} in the pulse sequence shown in Figure 6d. The transient is obtained as a function of the time delay between the second and third pulses, which probes the vibrational dynamics of the population; this experiment is illustrated in Figure 14. For Figure 14a using $\tau_{ab} = 460$ fs, only excited-state dynamics are observed; while for Figure 14b using $\tau_{ab} = 614$ fs, primarily ground-state dynamics are observed. This observation is confirmed by the Fourier transform of the time-resolved transient. The top transient has a vibrational frequency of 108 cm^{-1}, corresponding to the vibrational frequency of the excited B state. The bottom transient has a vibrational

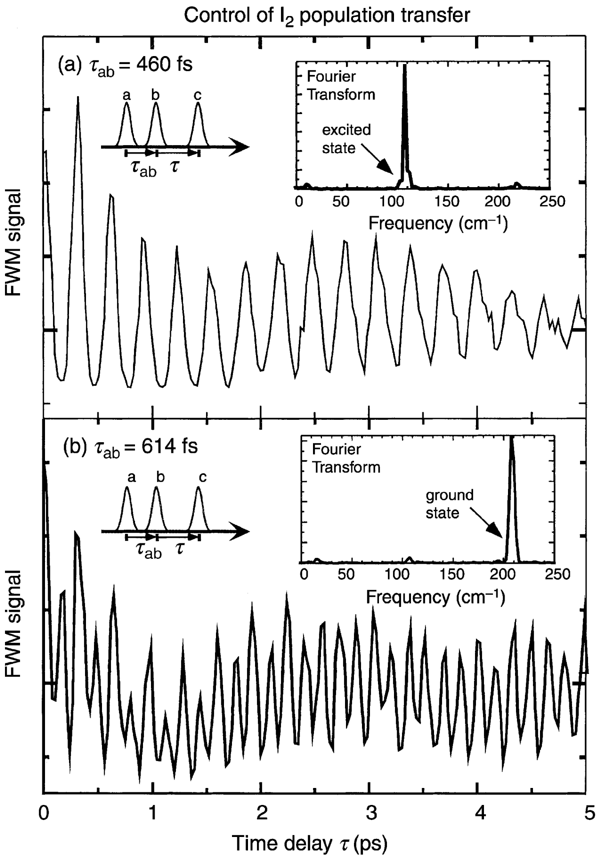

Figure 14 FWM signal in molecular iodine demonstrating control over observation of molecular dynamics. By using the pulse sequence shown, τ_{ab} was defined to be either (a) 460 fs or (b) 614 fs while τ was scanned in time. As is clearly evident in the data and the FFT, excited-state dynamics alone (307 fs oscillations) are observed with $\tau_{ab} = 460$ fs and ground-state dynamics (160 fs oscillations) are observed with $\tau_{ab} = 614$ fs.

frequency of 208 cm^{-1}, corresponding to the vibrational frequency of the ground state of iodine molecules.

Control Using Feedback and Evolutionary Algorithm

For complex molecules, one seldom knows the precise sequence of pulses or phase modulation of a given pulse that can optimize the yield of a chemical reaction. In these cases, a pulse-shaping device which manipulates the phase and amplitude of frequencies within an ultrashort pulse can be used. The device is part of a learning loop that includes a computer and an experimental setup. The learning algorithm running in the computer generates different pulses to be evaluated experimentally. Signal from the experiments is sent back to the computer for further refinement. The learning algorithm iteratively arrives at the optimum pulse. This method, first introduced by Rabitz, was first tested by Wilson. Gerber and co-workers have demonstrated automated control over the photodissociation reaction of organometallic compounds using an evolutionary algorithm with feedback. The organometallic compounds chosen were iron pentacarbonyl ($Fe(CO)_5$) and dicarbonylchloro(η^5-cyclopentadienyl)iron ($CpFe(CO)_2Cl$). For the first molecule, multiphoton absorption of femtosecond pulses at 800 nm caused ionization and fragmentation resulting in numerous ionic species – Fe^+, $Fe(CO)^+$, $Fe(CO)_2^+$, $Fe(CO)_3^+$, $Fe(CO)_4^+$, and $Fe(CO)_5^+$. These species were detected by mass spectrometry; the measured signal was used as feedback for the evolutionary algorithm that adjusts the phase of the femtosecond laser pulse. The $Fe^+$and $Fe(CO)_5^+$ exit channels were used to demonstrate the control available over branching ratios with this automated control method. Maximization (5.0) of the $Fe(CO)_5^+/Fe^+$ ratio was achieved with very short laser pulses while minimization (0.057) of this ratio was achieved with long (ps) pulses. These results were found to correspond well to changing the pulse duration of bandwidth-limited pulses. The optimization procedure did not attempt to control the other ionic species. This same automated optimization procedure was used to control the branching ratio of $CpFeCOCl^+/FeCl^+$ for the second organometallic compound. The maximum branching ratio was 4.9; the minimum was 1.2; and bandwidth-limited pulses yielded 2.4. Unlike the simpler carbonyl compound, simple changes to the pulse duration of a laser pulse could not optimize this branching ratio. The evolutionary algorithm with the feedback from the mass spectrometer allowed for these tailored pulses to be obtained and optimized the desired exit channel without any knowledge of the specific molecular system. The full potential of shaped femtosecond pulses has yet to be realized. It is possible that shaped femtosecond pulses will play an important role in future time-resolved experiments where they will be used to improve time resolution or to access potential energies that could not be reached with simpler pulses. Pulse shapers will also play a role in the preparation of ultrashort pulses with single femtosecond or attosecond duration.

List of Units and Nomenclature

Å	angstrom (10^{10} Å = 1 meter)
Attosecond	10^{18} as = 1 second
B	rotational constant
cm^{-1}	wavenumbers
D	centrifugal distortion constant
electron volts	1 eV = 8065 cm^{-1}
fs	femtosecond (10^{15} fs = 1 second)
FTS	femtosecond transition state spectroscopy
FWM	four-wave mixing
k	wavevector
LIF	laser-induced fluorescence
nm	nanometer (10^9 nm = 1 meter)
ps	picosecond (10^{12} ps = 1 second)
R(Å)	internuclear distance (given in angstroms)
$t = 0$	initiation of reaction (pump and probe are overlapped in time)
τ	time delay between pump and probe pulses being applied
TG	transient grating
ω_j	rotational frequency

See also

Chemical Applications of Lasers: Non-Linear Spectroscopies; Pump and Probe Studies of Femtosecond Kinetics. **Coherent Control:** Experimental; Theory. **Nonlinear Optics, Applications:** Phase Matching. **Nonlinear Optics, Basics:** Four-Wave Mixing; Ultrafast and Intense-Field Nonlinear Optics; $\chi^{(3)}$ – Third-Harmonic Generation. **Spectroscopy:** Nonlinear Laser Spectroscopy; Raman Spectroscopy. **Ultrafast Laser Techniques:** Generation of Femtosecond Pulses. **Ultrafast Technology**: Femtosecond Condensed Phase Spectroscopy: Structural Dynamics.

Further Reading

Bandrauk AD, Fujimura Y and Gordon RJ (eds) (2002) *Laser Control and Manipulation of Molecules*, vol. 821 Washington, DC: ACS Press.

Chergui M (ed.) (1996) *Femtochemistry: Ultrafast Chemical and Physical Processes in Molecular Systems*. Singapore: World Scientific.

Dantus M (2001) Coherent nonlinear spectroscopy: from femtosecond dynamics to control. *Annual Reviews of Physical Chemistry* 52: 639–679.

De Schriber FC, De Feyter S and Schweitzer G (eds) (2001) *Femtochemistry: With the Nobel Lecture of A. Zewail*. Weinheim, Germany: Wiley-VCH.

Douhal A and Santamaria J (eds) (2002) *Femtochemistry and Femtobiology*. Singapore: World Scientific.

Goswami D (2003) Optical pulse shaping approaches to coherent control. *Physics Reports* 374: 385–481.

Grimberg I, Lozovoy VV, Dantus M and Mukamel S (2002) Femtosecond three pulse spectrocopies in the gas phase: density matrix representation. *Journal of Physical Chemistry* 106: 697–718.

Manz J and Woste L (eds) (1995) *Femtosecond Chemistry*, vols. I and II. Heidelberg, Germany: Verlag Chemie.

Mukamel S (1995) Principles of Nonlinear Optical Spectroscopy. New York: Oxford University Press.

Sundstrom V (ed.) (1997) *Femtochemistry and Femtobiology*. Singapore: World Scientific.

Zewail AH (ed.) (1994) *Femtochemistry – Ultrafast Dynamics of the Chemical Bond*, vols. I and II. Singapore: World Scientific.

Femtosecond Condensed Phase Spectroscopy: Structural Dynamics

E T J Nibbering, Max Born Institut, Berlin, Germany

© 2005, Elsevier Ltd. All Rights Reserved.

Ultrafast Chemistry

Chemistry means rearrangement of atoms in/between molecules. Chemical bonds are broken and/or formed, leading to a change in chemical composition, from reactants to products, during a reaction. Following the temporal behavior of the reaction all the way from reactant – via possible intermediates – towards the product state(s), amounts to deciphering the molecular processes that lead to these chemical reaction dynamics. Molecular rearrangements from reactant to product may encompass many orders of magnitude in time (see Figure 1). Typically the larger and more involved the rearrangements are, the more time it takes for a molecular species to reach the final state. Rearrangements of biomolecular structure occur in micro- to milliseconds (spontaneous folding of small proteins consisting of several hundreds of amino acid units), or longer (DNA multiplication, protein synthesis in ribosomes). In these conformational dynamics, hundreds to thousands of chemical bonds change their nature. Bimolecular reaction dynamics in liquid solution occur on time-scales of nanoseconds or longer, being controlled by the relatively slow diffusional motions of the reaction partners to each other. Many elementary processes in chemistry, however, occur on much faster time-scales. In fact, when one considers the dynamical event of a single-bond rearrangement, the relocation of the (relatively few) atoms appears to take place as an ultrafast event. Bond fission, hydrogen and proton transfer, electron transfer, and cis/trans-isomerizations have been found to occur on femto- to picosecond time-scales ('femtochemistry'). The dynamics of these elementary processes are not only determined by the energy landscapes of the reacting partners. The energy landscapes (potential energy surfaces) are determined by molecular parameters such as relative orientation (distances, angles) of the reaction partners, the energy levels of the reactant, intermediate and product species, and – very important – the energy barriers represented by transition states. When considering condensed phase reaction dynamics, the important role of the surrounding solvent has been recognized. The solvent shells modulate energy levels of the reactant, intermediate and product states, through electrostatic interactions. In the liquid phase the solvent fluctuations lead through these interactions to fluctuating energy levels of the reaction partners. In addition, the solvent may facilitate chemical reactions through energy exchange with the reaction partners, leading to efficient dissipation of excess energy, making chemical reactions often irreversible. The processes that are extremely dominated by these solvent interactions (such as molecular collisions, electronic and vibrational dephasing, vibrational relaxation) are also often found to occur on femto- to picosecond time-scales. The outcome of a bond fission may be a null effect, when the dissociating fragments are forced by the surrounding solvent to recombine and relax ('cage effect').

Spectroscopy has been a prime tool to obtain insight into these key molecular processes. Reactant, intermediate, and products contribute to the spectra with their respective molecular resonances. With spectral domain spectroscopy one may, in principle, obtain a precise determination of structural information. Fluctuations of energy levels of these species by the solvent interactions, and finite lifetimes, contribute to a broadening of these molecular resonances. Using steady-state spectroscopy one

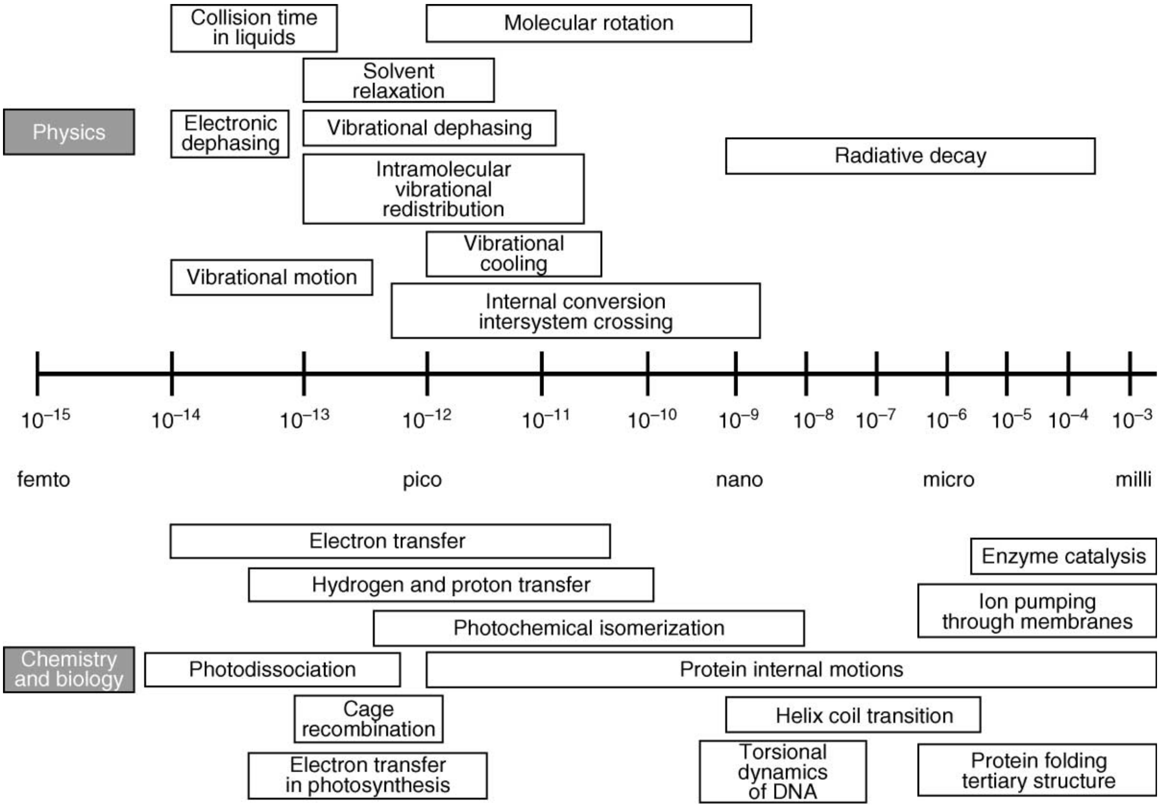
Collision time in liquids
Molecular rotation
Solvent relaxation
Physics
Electronic dephasing
Vibrational dephasing
Radiative decay
Intramolecular vibrational redistribution
Vibrational cooling
Vibrational motion
Internal conversion intersystem crossing
10^{-15}
10^{-14}
10^{-13}
10^{-12}
10^{-11}
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
femto
pico
nano
micro
milli
Electron transfer
Enzyme catalysis
Hydrogen and proton transfer
Ion pumping through membranes
Photochemical isomerization
Chemistry and biology
Photodissociation
Protein internal motions
Cage recombination
Helix coil transition
Torsional dynamics of DNA
Electron transfer in photosynthesis
Protein folding tertiary structure

Figure 1 Chart showing the ultrafast time-scales of fundamental physical and chemical processes (inspired by charts of Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press and Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694).

could try to determine these effects through a lineshape analysis. Using this approach, the results have to be analyzed with an a priori defined model representing the dynamics. The disadvantage lies in the averaging over all time-scales inherent in spectral domain spectroscopy, where details of transient states may be lost in this averaging procedure. In contrast, in experiments designed to grasp the molecular dynamics in real time, the temporal resolution is the prerequisite quantity in observation of these processes. In time-resolved experiments one can trigger a chemical reaction at a well-defined point in time, after which one can follow the conformational changes from reactant – via possible intermediates – to the product, and identify any possible transition states along the reaction pathways.

During the last century the technological advances in time-resolved spectroscopy have prompted significant breakthroughs in the study of chemical reaction dynamics. Initially the temporal resolution was given by the duration of flashes from light bulbs (‘flash’ spectroscopy) or by rapid mixing of reaction partners (stopped-flow technique), as developed by Nobel Laureates Eigen, Norrish, and Porter, and many others. These techniques allow a temporal resolution of at best milliseconds. Already, up to the early 1960s, it was understood that elementary reaction dynamics occur on much faster time-scales. Fourier spectroscopy may improve time resolution down to microseconds or slightly less. With the advent of pulsed laser sources with ever-increasing temporal resolution, a wide arena of time-resolved spectroscopic techniques has emerged. Nanosecond laser systems are known since the early 1960s, picosecond pulses can be generated since the 1970s, the first femtosecond laser system was reported in 1981, and currently laser engineering groups are breaking into the attosecond domain. The majority of pulsed laser systems operate at visible and near-infrared wavelengths. This means that usually chemical events in ultrafast time-resolved spectroscopy are initiated by electronic excitation (‘photochemistry’). Subsequent molecular rearrangements have then to be followed by probing the electronic states through electronic resonances (UV/VIS pump–UV/VIS probe or UV/VIS four-wave mixing

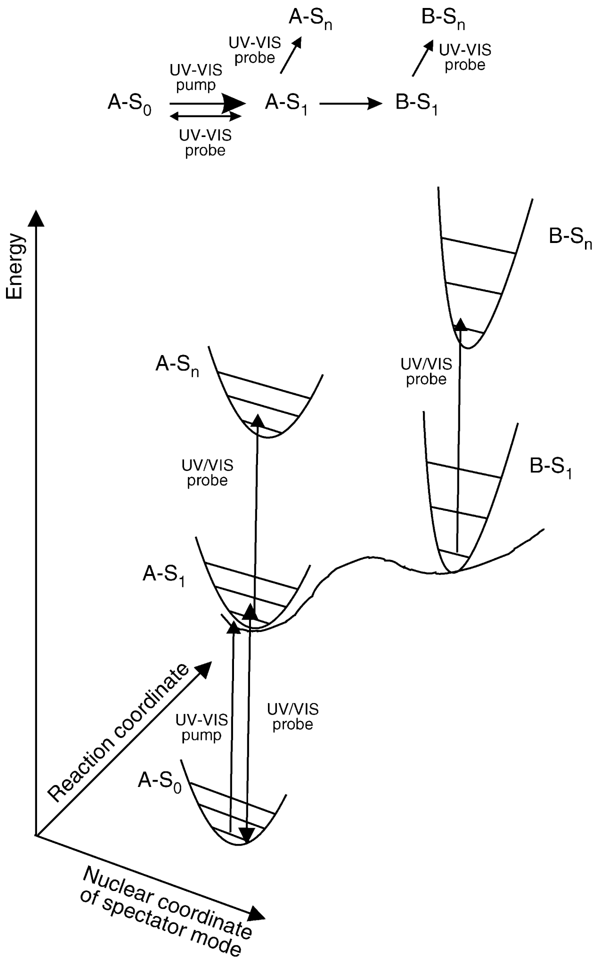
A-Sn
B-Sn
UV-VIS probe
UV-VIS probe
UV-VIS pump
A-S0
A-S1
B-S1
UV-VIS probe
Energy
B-Sn
A-Sn
UV/VIS probe
UV/VIS probe
B-S1
A-S1
Reaction coordinate
UV-VIS pump
UV/VIS probe
A-S0
Nuclear coordinate of spectator mode

Figure 2 Ultrafast time-resolved electronic spectroscopy. A UV/VIS pump pulse excited an electronic resonance, and a probe pulse determines, depending on the tuning in the UV/VIS population changes in the reactant ground or excited states or in the product state.

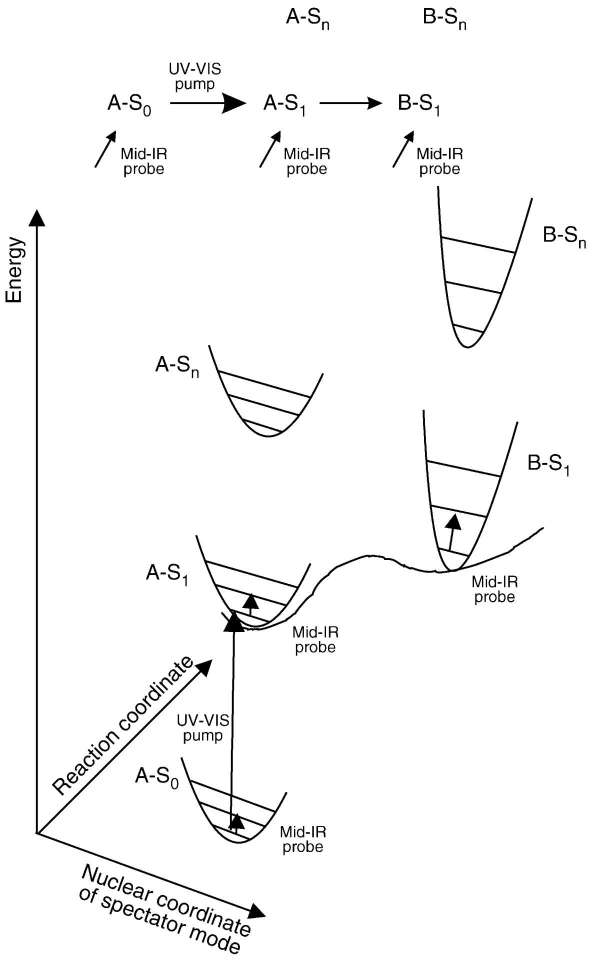
A-Sn
B-Sn
UV-VIS pump
A-S0
A-S1
B-S1
Mid-IR probe
Mid-IR probe
Mid-IR probe
Energy
B-Sn
A-Sn
B-S1
A-S1
Mid-IR probe
Mid-IR probe
Reaction coordinate
UV-VIS pump
A-S0
Mid-IR probe
Nuclear coordinate of spectator mode

Figure 3 Ultrafast time-resolved infrared spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with an infrared probe pulse by inspection of marker modes in the different electronic states.

spectroscopy) (Figure 2), or via vibrational resonances (UV/VIS pump–IR/Raman probe spectroscopy) (see Figures 3 and 4).

Laser Technology

Since the beginning of the 1990s, a major advancement in femtosecond laser technology has emerged by the discovery of Kerr lens mode locking in laser oscillators with Ti:sapphire as lasing material. Ultrastable and ultrashort laser pulses tunable between 700 and 1,000 nm can be generated with durations as short as 6 fs. The second advancement is due to the ability of Ti:sapphire as laser amplifying material, using the method of chirped pulse amplification. Pulses with output powers up to several W, with repetition rates in the kHz-regime, are now routinely generated. These developments facilitate the efficient generation of ultrashort pulses tunable from the UV to the mid-infrared, and even beyond to the THz-regime using schemes based on nonlinear optics. Frequency conversion of the fundamental output of an amplified Ti:sapphire laser system is then achieved with nonlinear processes such as self-phase modulation and supercontinuum generation, n^{th}-order harmonic generation ($n = 2, 3, 4, \ldots$), sum and difference frequency generation, or parametric generation and amplification. As a result, one can now almost arbitrarily tune the excitation and probe wavelengths of the applied laser pulses. Additional parameters, that one may alter, are pulse duration (from less than 10 fs up to several ps) and pulse energy (in the mJ range or less; typically one uses for

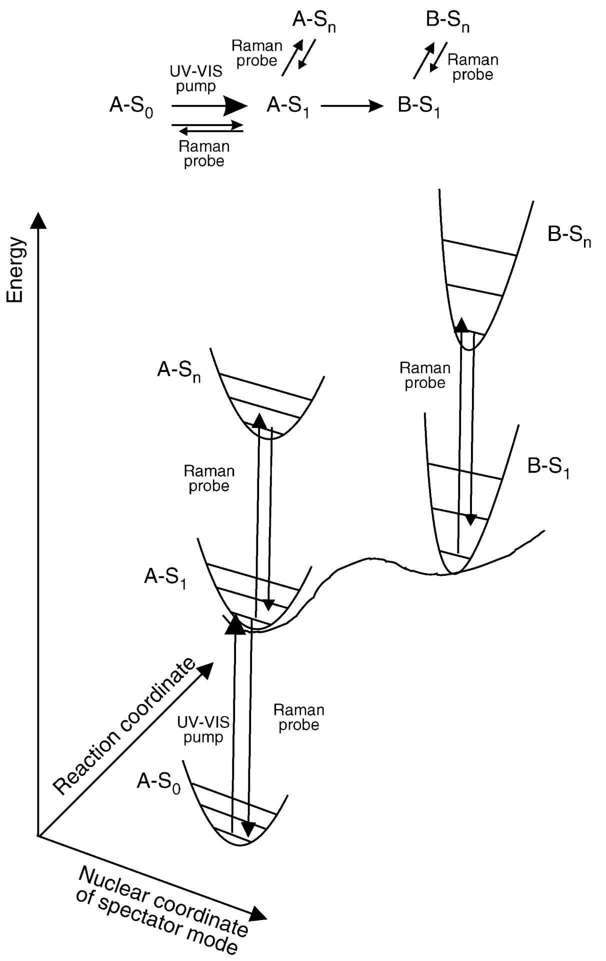

Figure 4 Ultrafast time-resolved Raman spectroscopy. A UV/VIS pump pulse excites an electronic resonance, after which the populations are followed with (resonance) Raman probe pulse by inspection of marker modes in the different electronic states.

condensed phase spectroscopy pulse energies in the range of nJ to μJ). More advanced pulse manipulation schemes use amplitude and phase masks. The latter aspect holds the promise of amplitude and phase control of molecular excitation, feeding the ultrafast chemist's dream of laser control of molecular reaction dynamics.

Ultrafast Electronic Spectroscopic Techniques

Pulsed laser sources enable a variety of experimental approaches to determine transient states. Already, since the development of nanosecond laser sources, time-resolved pump–probe and four-wave mixing spectroscopy have been widely used. In a first approach, often one uses these forms of spectroscopy to probe in real-time populations of reactant, transient, and product states as function of pulse delay. With the advent of pico- and femtosecond laser sources, coherence properties of the material response have been investigated as well.

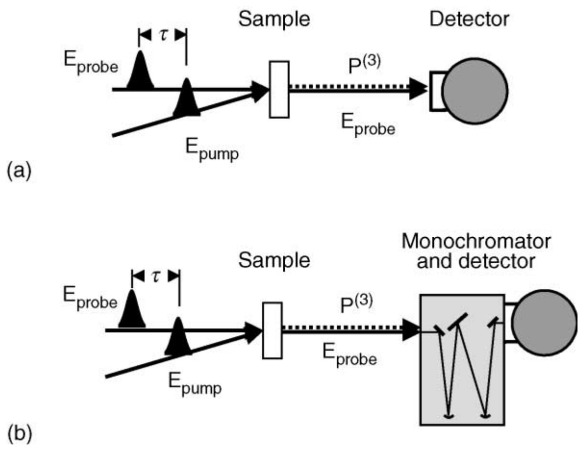

Figure 5 Pump–probe technique: A pump pulse induces an absorbance change in a sample, that is subsequently measured by a time delayed probe pulse, either spectrally integrated (a) or spectrally resolved (b).

The following techniques have been widely used in ultrafast electronic condensed phase spectroscopy.

UV/VIS Pump–UV/VIS Probe Spectroscopy

In this technique a 'pump' pulse resonant to a transition promotes the molecules to an electronic excited state, and the time-dependent populations are determined by measurement of the transmission of a time-delayed 'probe' pulse (Figure 5). The absorbance change $\Delta A(\tau) = -^{10}\log[T(\tau)/T(\tau = 0)]$ (with T for transmission change and τ the pulse delay) is a direct indication of populations of states. Here one can tune the probe pulse to the same transition as the pump pulse, and one measures an absorbance decrease due to a bleach of ground state population (i.e., fraction of molecules being excited), and due to molecules in the excited state stimulated back to the ground state. One could also tune the probe pulse to transitions between the transient state and higher lying states. For electronic spectroscopy, often transitions between the ground and first excited state ($S_0 \rightarrow S_1$) overlap with excited state absorptions ($S_1 \rightarrow S_n$). One has to rely then on numerical analysis procedures such as singular value decomposition and decay associated spectra, that correlate spectral features with temporal behavior.

The time resolution of the experiment is given by the cross-correlation between pump and probe pulses. In the case of extremely short pulses one has to consider the effects of group velocity dispersion between different frequency components due to different travel times through the samples, leading to pulse temporal broadening. For the case of pump and probe tuned to different wavelengths, the time-resolution is usually dominated by group velocity mismatch between pump and probe pulses.

When ultrashort laser pulses are used, often the applied spectral bandwidth enables simultaneous excitation of several vibrational sublevels in vibronic transitions. Due to the phase-relationship of the applied laser pulses, these vibrational sublevels are prepared in states that are well-defined in temporal phase with respect to each other. Such a coherent superposition of vibrational eigenstates is nothing less than a vibrational wavepacket, that will evolve in time. The occurrence of vibrational wavepacket motions leads to oscillatory modulations of the pump–probe signals.

Time-Resolved Fluorescence Spectroscopy

Here again a 'pump' pulse excites an electronic resonance. The transient excited states are now probed through their spontaneous fluorescence emission, either by use of time-correlated single photon counting or by fluorescence upconversion with a gating pulse in a nonlinear medium (Figure 6). Single-photon counting has a time resolution of at best a few picoseconds. Fluorescence upconversion can be performed with a time resolution of 100–200 fs. The advantage lies in the fact that the measured signals are more likely due to the specific excited states only, and not corrupted by other contributions.

Electronic Four-Wave Mixing Spectroscopy

In time-resolved degenerate electronic four-wave mixing spectroscopy, where all the three applied laser pulses (with wavevectors $\mathbf{k}_1$, $\mathbf{k}_2$, and $\mathbf{k}_3$) are tuned to the same electronic resonance, one can detect a nonlinear signal in the phase-matched direction $\mathbf{k}_S = \mathbf{k}_3 + \mathbf{k}_2 - \mathbf{k}_1$ (Figure 7). The first laser pulse generates a coherent superposition between the electronic ground and excited states. This electronic coherence evolves in time during the period (denoted as coherence time τ), until the second pulse converts this coherence in a population frequency grating in the electronic ground and excited states. After a second time period of free evolution (denoted as population or waiting time T), this frequency grating is again transformed into an electronic coherence, which after some time evolution generates a macroscopic polarization in the phase-matched directions. This macroscopic polarization is either directly measured by a time-integrating detector (in which case one measures the intensity signal proportional to the absolute square of the nonlinear polarization), or one uses another light pulse that mixes as local oscillator with the nonlinear polarization (and the signal is proportional to the amplitude of the polarization). The time resolution is determined by the third-order autocorrelation of the applied pulses.

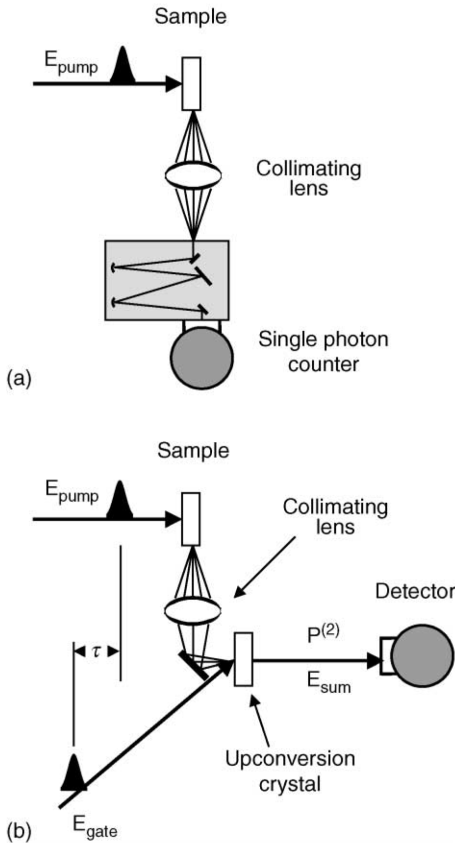

Figure 6 Time-resolved emission detection: The fluorescence emission is either temporally resolved by time-correlated single photon counting (a) or by up-conversion with a gating pulse (b).

Depending on the delay between the three input pulses, and which delay one varies, different dynamics can be measured. When one performs an experiment with τ scanned and T set at zero, one measures the electronic coherence decay (due to electronic dephasing) time in a two-pulse photon echo experiment. When, on the other hand, τ is set to zero, and T is scanned, one measures the evolution of a population grating, and electronic lifetimes can be deduced. In a three-pulse photon echo experiment both pulse delays are scanned, and the signals give insight into both electronic coherence and frequency grating decay (due to electronic dephasing and spectral diffusion, respectively). A novel extension of four-wave mixing technique is the three-pulse echo peak shift (3PEPS). In a 3PEPS measurement, the delay time τ_{max} is

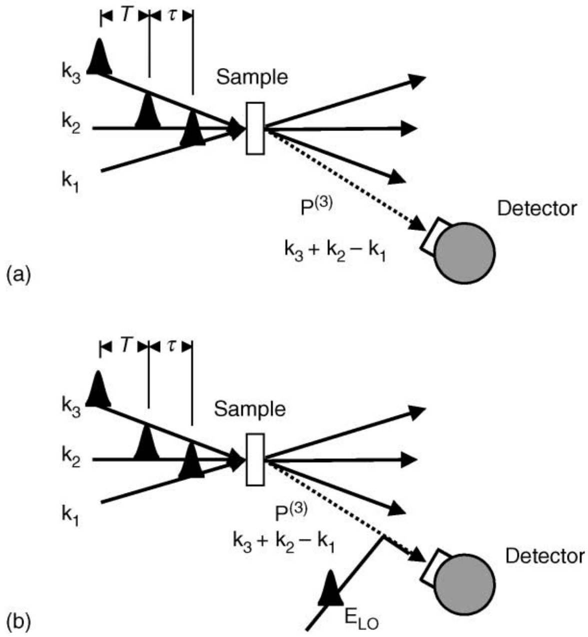

Figure 7 Beam geometry for time-resolved four-wave mixing (grating/photon echo) spectroscopy: The nonlinear signal is transmitted by the sample in a phase-matched direction and its intensity signal (a) or its amplitude by interference with a local oscillator field (b) is detected by a time-integrating detector.

determined, where the echo signal has a maximum along the coherence delay τ, as a function of the population delay T. This method provides direct insight into the frequency fluctuation correlation function that governs the linear and nonlinear signals. Extension of the method to multi-color four-wave mixing spectroscopy has been demonstrated.

Electronic Condensed Phase Spectroscopy: Femtochemistry and Solvation Dynamics

In the narrow sense of meaning, 'femtochemistry' deals with the real-time observation of nuclear motions during chemical reactions. Quantum mechanics teaches us that this is only possible when molecular systems are prepared in vibrational superpositions. Only in the case of quantum states describing vibrational wavepackets evolving in time the narrow-defined positions of nuclei allow for a real-time description of molecular rearrangements, or in other words: dynamics. This has been demonstrated in the gas-phase landmark experiments by Nobel Laureate Zewail (Figure 8).

In the condensed phase at room temperature, however, often such well-defined quantum states are hard to prepare, since the ensembles of molecular systems exist in mixed states. Coherent laser excitation on an ensemble will then prepare mixed states as well, and then information that one obtains from the experiments is less explicit. In addition, room temperature liquid solutions induce extremely fast fluctuations in the states of the molecules under study and as a consequence the electronic and vibrational coherences show ultrafast decaying behavior. This limits the temporal window with which one can follow the evolution of vibrational wavepackets.

Most of the ultrafast spectroscopic experimental work performed on photochemistry in the condensed phase, being described as 'femtochemistry', only deals with the time-scales of changes in electronic states (usually indicated with reaction rates typified with exponentially decaying or rising functions). In this regard it is better to speak of 'kinetics' rather than 'dynamics', although both terms are used in condensed phase spectroscopy. Due to the relative broad electronic spectra of condensed phase molecules masking any structural detail, it is hard to derive nuclear motions in real time, and as such it is difficult to make any statements about dynamics of molecular structures during chemical reactions.

Electronic resonances have also been used to probe the dynamical interactions between a nonreactive solute and the surrounding fluctuating solvent. The idea behind these experiments is to achieve information about the time-scales of solvent fluctuations on (and solvent motions due to a change in) the electronic charge distribution in a molecule (as induced by an electronic excitation). Usually the solvent motions are characterized with a frequency fluctuation correlation function (Figure 9). This information is highly relevant for the case of reacting molecules, where electronic motions accompany the relocations of nuclei, even in electron transfer reactions, where the nucleic rearrangements are modest. Solvent motions have been shown to often control reaction rates. Solvent rearrangement, as a response to a change in electronic charge distribution, is known as solvation dynamics. Experimental probes for solvent motional fluctuations are electronic coherence decay (electronic dephasing) and population frequency grating decay (spectral diffusion) that can be measured in two- and three-pulse photon echoes. Solvent adaptation to new charge distributions can be followed with the time-dependent Stokes-shift in fluorescence emission, as well as in three-pulse photon echoes (in so-called three-pulse echo peak shift measurements described above). Ultimately these experiments resolve the temporal characteristics of the frequency fluctuation correlation function.

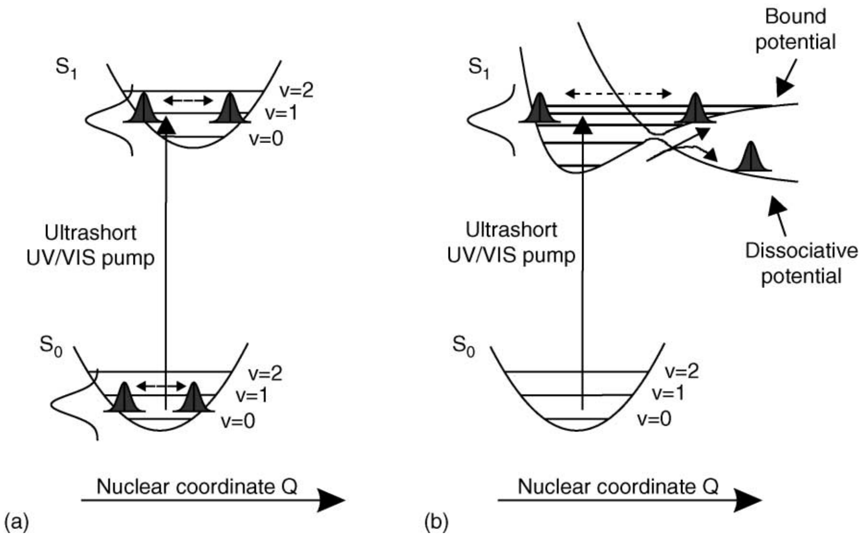

Figure 8 Ultrafast excitation with broadband pulses creates coherent superpositions of vibrational eigenstates, i.e., vibrational wavepackets. These wavepackets are generated in electronic excited and ground states (a), and their time-dependence can be determined in pump–probe and four-wave mixing experiments. When level crossing occurs to another electronic state, the vibrational wavepacket may be observed to evolve along the reaction coordinate (b).

Ultrafast Vibrational Spectroscopy in Photochemistry: Structural Dynamics

Vibrational spectroscopy has, despite a smaller signal sensitivity due to smaller transition cross-sections, several advantages over electronic spectroscopy. Vibrational transitions have typically smaller bandwidths than electronic transitions, due to longer dephasing times (the exception to the rule: O–H/O–D stretching bands of hydrogen bonded hydroxyl groups). Specific signal contributions are thus easier tractable in vibrational spectroscopy than in electronic spectroscopy. Vibrational bands can often be correlated to specific vibrational motions by inspection of the transition frequencies, (e.g., O–H, N–H, and C–H stretching bands can be found in the 3000 cm^{-1} range, C=O and C=N stretching modes are located near 1,600–1,750 cm^{-1}, O–H, N–H, and C–H bending vibrations between 1,500–1,650 cm^{-1}, C–O stretching mode around 1,200–1,300 cm^{-1} etc.), and, in particular, the frequency range between 1,000 and 1,800 cm^{-1} is called the fingerprint region because of this reason. Measurement of vibrational bands thus leads to identification of particular vibrational motions, and conclusions can be drawn on specific structural motifs in the molecules. Vibrational bands can be infrared (IR) or Raman active, sometimes both, and both IR and Raman techniques have been widely applied in structural determination studies ranging from small molecules to larger biomolecular systems. In ultrafast photochemistry one thus excites the molecules with a UV/VIS pump pulse, and an IR/Raman probe pulse follows the outcome of the chemical reaction by inspection of vibrational bands of reactant, transients, and products.

Vibrational spectroscopy has the potential of revealing site-specific information if the marker modes are due to nuclear motions of specific molecular side-groups. For instance, hydrogen bonding induces marked shifts of O–H, N–H, C=O, and C=N bands. Observation of changes in spectral shifts reveals important information on hydrogen bond interactions (weakening/strengthening or even hydrogen bond cleavage). Infrared spectroscopy is able to probe small molecular species in solution that typically have their electronic resonances in the far-UV (making electronic spectroscopy impossible, since normally the solvent would absorb this radiation) (Figure 10). For instance, in acid–base neutralization reactions, where the acidity of a so-called photoacid is switched on by use of a UV pulse, a VIS probe pulse would only be able to probe the photoacid (or its conjugate photobase), thus revealing only when a proton leaves the photoacid. An infrared pulse can, besides probing vibrational resonances of the photoacid, also probe bands of a small base (or its conjugated acid) indicating when a proton arrives at the base.

In the case where the vibrational normal modes do not allow such a structural insight into site-specific groups, one can make a comparison of the

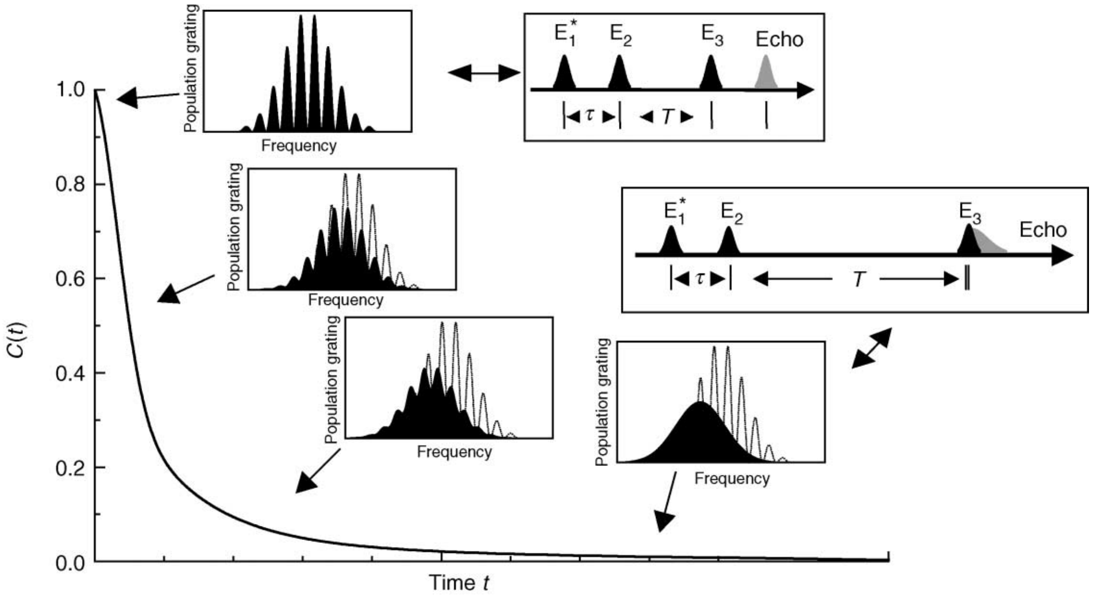

Figure 9 Solute–solvent interactions as characterized by the transition frequency fluctuation correlation function $C(t)$. This quantity can be measured in photon echo and time-resolved fluorescence experiments. In a three-pulse photon echo experiment phase information as given by the excited state frequency grating created with the first two interactions, is washed out due to solvent fluctuations and Stokes shifts to lower frequencies due to solvent rearrangement. A similar plot can be drawn for the ground state frequency grating.

experimentally observed vibrational bands with predictions made by quantum chemical calculations. When a full correspondence between experiment and theory is possible, one can make statements about the three-dimensional structure. With current quantum chemical calculational routines, such as density functional theory, medium-sized molecules are routinely calculated for electronic ground-state conformations. Reliable results of transient and product states in electronic excited states can be obtained with the routine *ab initio* complete active space self-consistent field (CASSCF), albeit for mid-size molecules of at most on the order of about 20 atoms. New developments in numerical procedures, such as time-dependent density functional theory (TD-DFT) may prove fruitful in the calculation of larger molecular systems in electronic excited states.

During recent years, femtosecond IR spectroscopy has been used in photo-induced chemical reactions ranging from excited state hydrogen and proton transfer, transformations of photochromic switches, excited state charge transfer, and cis–trans isomerizations. Here an ultrashort UV/VIS ‘pump’ pulse promotes the molecule to an electronic excited state, and the reaction is followed by a measurement of the absorbance change of a ‘probe’ pulse tuned in the mid-IR region where the vibrational marker modes sensitive to structural changes have their resonances. Experimentally, one performs spectrally resolved transient IR spectroscopy. Femtosecond IR parametric devices deliver pulses with bandwidths of 150 cm^{-1} or more. In order to be able to observe shifts as small as the linewidths of IR-active vibrations, one usually measures the IR absorbance change with a detector after spectral dispersion with a monochromator. As a side-effect of this spectral dispersion ground-state bleach signals often appear to grow in at negative pulse delay with the dephasing time of the transition. This effect, known as ‘perturbed free induction decay’ is a common feature of spectrally resolved nonlinear pump–probe spectroscopy of bleached transitions with dephasing times much longer than the time resolution of the experiment. The time resolution of the experiment is given by the cross-correlation between the UV/VIS-pump and IR-probe pulse (about 100–200 fs), and is typically dominated by group velocity mismatch in samples with thicknesses of about 100 μm.

In principle, the same approach can be followed by probing Raman-active vibrations. In this case the spectral resolution is not only determined by the monochromator through which the spontaneous Raman emission dispersed, but also by the bandwidth of the gating pulse by which the Raman effect is induced. As a result, UV/VIS pump–Raman probe

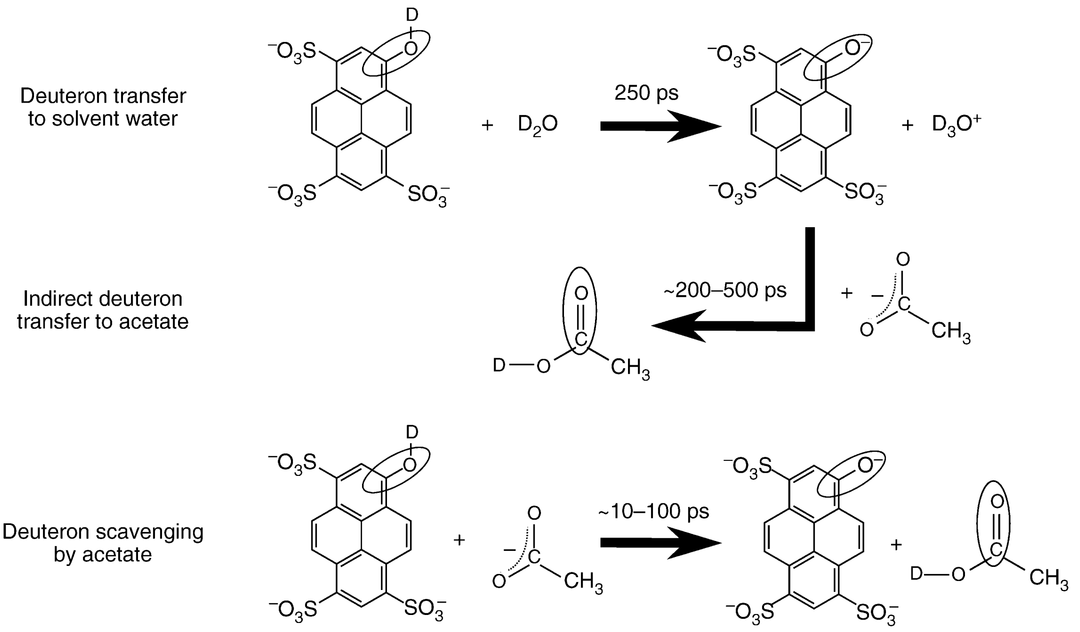
Deuteron transfer to solvent water
+ D2O
250 ps
+ D3O+
Indirect deuteron transfer to acetate
~200–500 ps
Deuteron scavenging by acetate
~10–100 ps
D
O
−O3S
SO3−
O
C
CH3
D—O

Figure 10 In excited state acid–base neutralization experiments, the proton dissociation can be followed on two sides with ultrafast infrared spectroscopy. Vibrational marker modes of photoacid and conjugated photobase indicate when the proton leaves the acid. The marker mode of acid formed by proton pick-up by the accepting base is a direct measure of the arrival time of the proton at the base. Depending on relative concentrations, either the indirect proton transfer via the solvent or direct scavenging of the proton by the base dominates the dynamics.

spectroscopy has a temporal resolution of around 1 ps (a compromise between spectral and time resolution). Due to the even weaker cross-sections of Raman transitions, one often uses resonance enhancement by tuning the gating pulse close to or resonant with an electronic transition of the state that is probed. The advantage is that one usually only observes Raman bands of the state under inspection. The drawback is that fluorescence resulting from the resonant electronic excitation often inhibits detection of Raman bands in extended spectral ranges.

Raman spectroscopy offers an additional insight into how chemical reactions evolve. By comparison of the intensities of anti-Stokes and Stokes lines of a particular vibration, it is possible to derive time-dependent changes in the excitation level of this vibration. One can draw conclusions on whether particular modes initially drive a chemical reaction (‘state transition promoting modes’), or only get excited after the transition is made, by taking up the excess energy released by the reaction (‘accepting modes’). In such a way, insight is obtained on the energy flow inside a molecular system (intramolecular vibrational redistribution, abbreviated as IVR) and vibrational energy dissipation to the surrounding solvent (vibrational cooling). Infrared spectroscopy is less powerful in revealing this, since red-shifted transient absorption of vibrational bands may either mean that the particular mode is highly excited (the red-shift is a consequence of the diagonal anharmonicity of the vibration), or it may mean that other modes are highly excited and these cause a red-shift of the marker mode under inspection (the red-shift is then due to off-diagonal anharmonic coupling with other modes).

Equilibrium Structural Dynamics in the Electronic Ground State

Existence of anharmonic coupling between vibrational modes means that the vibrational motions are not decoupled from each other. These couplings always exist, since otherwise no IVR and vibrational cooling (no relaxation) would occur. When the couplings are significantly large, excitation of one specific mode will induce a significant instantaneous shift of other vibrations anharmonically coupled to this particular mode. Estimation of the magnitude of these couplings should lead to a determination of the curvatures of potential energy surfaces along the respective coordinates. With current state-of-the-art femtosecond infrared technology (with time resolution of about 150 fs or less) it is possible to excite

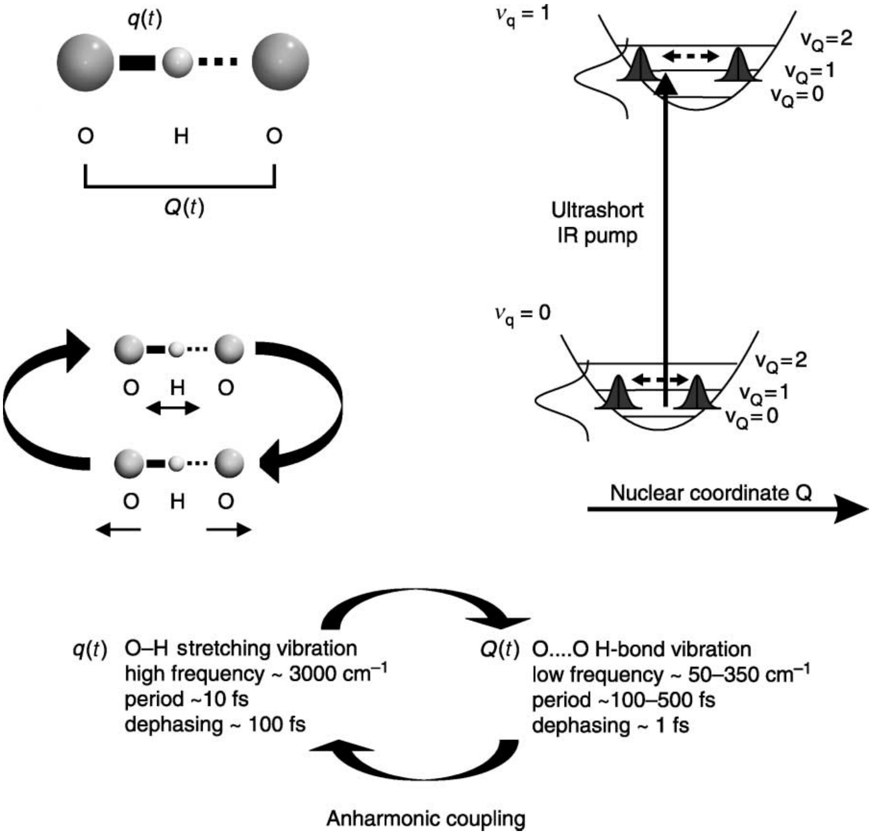

Figure 11 In hydrogen-bonded hydroxyl groups a strong anharmonic coupling exists between the hydroxyl stretching and modes modulating the hydrogen-bond distance. A Born–Oppenheimer-like separation of time-scales of vibrational motion between the high-frequency O–H/O–D stretching and the low-frequency hydrogen bond modes enables a description with potential energy surfaces for the low-frequency mode as a function of the quantum state of the hydroxyl stretching vibration. The formation of low-frequency vibrational wavepackets in ultrafast IR spectroscopy of O–H/O–D stretching modes is then fully analogous to wavepacket formation in ultrafast electronic spectroscopy.

infrared-active vibrations and to probe the same or other vibrations (either in the IR or by the Raman process), enabling insight into the anharmonic couplings between these vibrational modes.

Direct anharmonic couplings are probed in IR-pump–IR-probe and IR photon echo spectroscopy (Figure 11). In the case of peptides, the excitonic (Davydov) couplings between amide I vibrations of the amino acid units have been the subject of extensive study since these lead to spatial information of the relative orientations of the different amide I modes, and thus, of the spatial orientation of the amino acid units inside the peptide. Exploring the effects of the delay between the IR pulses on the observed signals, give insight into the fluctuations of these orientational features, or in other words, into structural dynamics of peptides. In the case of hydrogen bonded O–H/O–D stretching vibrations, the anharmonic couplings with underdamped low-frequency modes modulating the hydrogen bond distance, lead to marked modulations of observed pump–probe and echo signals as function of pulse delay. These experiments reveal which modes couple strongly to the hydrogen stretching oscillator.

Fluctuations in couplings between vibrational modes (in particular with those of the solvent) induce dephasing of the vibrational coherences. The dephasing and spectral diffusion dynamics of vibrational transitions, e.g., the O–H stretch vibration of HOD dissolved in D_2O, have been determined by IR photon echo spectroscopy. From these studies the time-scales of structural memory decay of hydrogen bond networks can be derived. Couplings between vibrational modes ultimately lead to population relaxation through intramolecular vibrational redistribution effects and vibrational cooling. Ultrafast two-color IR-pump–IR-probe reveals whether excitation of one IR-active vibration is followed by population transfer into another IR-active mode.

With IR-pump–Raman-probe, one can estimate whether Raman-active vibrations are transiently excited. In the case of hydrogen-bonded O–H/O–D stretching bands, an extremely short population relaxation time T_1 (below 1 ps) is observed, as well as an ultrarapid dephasing time T_2 (200 fs or less). These values are definitely longer for other vibrations, from which it has been concluded that the strong anharmonicities in hydrogen-bonded vibrational systems enable dephasing and relaxation processes to have such a rapid impact.

Future Developments

Future activities in the determination of dynamically evolving molecular structures in the condensed phase, may involve the aforementioned application of amplitude and phase controlled excitation pulses in optimal control of chemical reactions. Vibrational spectroscopy may be used as a tool in determination of evolving structures after electronic excitation by a shaped pulse tuned in the UV/VIS. Infrared pulses could also be shaped in amplitude and phase, and may be used in the exploration of steering of chemical reactions in the electronic ground state. Multidimensional vibrational spectroscopy is an object of extensive research. Extension of the method to determination of structure of transient states will lead to new results on molecular rearrangements, e.g., polypeptides reaching a new equilibrium after inducing a UV-excitation induced geometric change of a photochromic switch incorporated in the peptide structure. Much activity is currently also being put in the development of ultrafast structure resolving techniques such as electron diffraction, X-ray diffraction, and X-ray spectroscopy. While still extremely demanding at a technological level, final implementation of these latter techniques will open up new areas of ultrafast chemical dynamics.

See also

Chemical Applications of Lasers: Pump and Probe Studies of Femtosecond Kinetics.

Further Reading

Cerullo G and De Silvestri S (2003) Ultrafast optical parametric amplifiers. *Review of Scientific Instruments* 74(1): 1–18.

de Boeij WP, Pshenichnikov MS and Wiersma DA (1998) Ultrafast solvation dynamics explored by femtosecond photon echo spectroscopies. *Annual Review of Physical Chemistry* 49: 99–123.

Elsaesser T and Bakker HJ (eds) (2002) *Ultrafast Hydrogen Bonding Dynamics and Proton Transfer in the Condensed Phase*. Dordrecht, Germany: Kluwer.

Elsaesser T and Kaiser W (1991) Vibrational and vibronic relaxation of large polyatomic molecules. *Annual Review of Physical Chemistry* 42: 83–107.

Fayer MD (ed.) (2001) *Ultrafast Infrared and Raman Spectroscopy*. New York: Marcel Dekker.

Fleming GR (1986) *Chemical Applications of Ultrafast Spectroscopy*. Oxford, UK: Clarendon Press.

Fleming GR and Cho M (1996) Chromophore-solvent dynamics. *Annual Review of Physical Chemistry* 47: 109–134.

Fleming GR and Hänggi P (eds) (1993) *Activated Barrier Crossing: Applications in Physics, Chemistry and Biology*. Singapore: World Scientific.

Kaiser W (ed.) (1988) *Ultrashort Laser Pulses and Applications*. Berlin: Springer.

Mukamel S (1995) *Principles of Nonlinear Optical Spectroscopy*. Oxford, UK: Oxford University Press.

Nibbering ETJ and Elsaesser T (2004) Ultrafast vibrational dynamics of hydrogen bonds in the condensed phase. *Chemical Reviews* 104: 1887–1914.

Owrutsky JC, Raftery D and Hochstrasser RM (1994) Vibrational relaxation dynamics in solutions. *Annual Review of Physical Chemistry* 45: 519–555.

Sundström V (ed.) (1996) *Nobel Symposium: Femtochemistry and Femtobiology: Ultrafast Reaction Dynamics at Atomic-Scale Resolution*. Singapore: World Scientific.

Wiersma DA (ed.) (1994) *Femtosecond Reaction Dynamics*. Amsterdam: North-Holland.

Zewail AH (2000) Femtochemistry: atomic-scale dynamics of the chemical bond. Nobel lecture as appeared in *Journal of Physical Chemistry A* 104(24): 5660–5694.

Web-information: http://www.nobel.se
Nobel web-site showing information on Nobel Laureates M Eigen, RGW Norrish and G Porter (The Nobel Prize in Chemistry 1967: 'for their studies of extremely fast chemical reactions, effected by disturbing the equilibrium by means of very short pulses of energy') and AH Zewail (The Nobel Prize in Chemistry 1999: 'for his studies of the transition states of chemical reactions using femtosecond spectroscopy').

Ultrafast Illumination and Processing

Y Fainman, University of California, San Diego, CA, USA

D M Marom, Bell Laboratories, Holmdel, NJ, USA

© 2005, Elsevier Ltd. All Rights Reserved.

Introduction

Mode-locked lasers emitting ultrashort pulses have expanded the range of available instruments and techniques to the scientific community. The unique properties of these ultrashort pulses – their high peak powers and short duration – enable such diverse applications as multiphoton microscopy, pump-probe experiments, and coherent control of molecular and atomic wavepackets, to name a few. Ultrashort pulses can also be used as a basis for ultrahigh data rate optical communications using time division multiplexing (TDM) in a fiber communication application. These ultrafast applications often require complementary real-time optical processors based on instantaneous nonlinear phenomena, since conventional electronic means are limited by the response time of electronic devices.

During the last few years, we have developed real-time optical processors utilizing nonlinear three- and four-wave mixing of signals originating from spatial or temporal channels for synthesizing, processing, and detecting ultrafast waveforms. We perform real-time optical signal processing that can be applied to data conversion between slow parallel channels in space and an ultrafast serial signal. A time-to-space mapping processor has been built for detection of ultrafast waveforms. It mixes two spatially inverted temporal frequency information waves in a three-wave mixing arrangement, generating a quasi-monochromatic spatial signal carrying the temporal image for detection by slower electronic means. A space-to-time mapping processor was constructed for generation of arbitrary ultrafast waveforms from an input ultrashort pulse. The waveform is a time-scaled replica of the employed input spatial image, converted by a four-wave mixing arrangement. By introducing only temporal information channels to our four-wave mixing processors, real-time optical processing of temporal waveforms was enabled. We have demonstrated time reversal experiments based on performing spectral phase conjugation and spectral inversion operations, achieving time reversal of the electrical field and of the complex amplitude waveform, respectively. The three- and four-wave mixing arrangements are performed within a spectral processing device – a free-space optical setup consisting of diffraction gratings and lenses – utilizing a $\chi^{(2)}$ nonlinear crystal. A cascaded second-order nonlinearity technique is used for four-wave mixing within the $\chi^{(2)}$ crystal (a frequency-sum generation process followed by a frequency-difference generation process satisfying the type-II noncollinear phase matching condition). However, a $\chi^{(3)}$ nonlinear crystal may be readily substituted without loss of generality (albeit with a likely smaller conversion efficiency).

This article reviews the spatial/temporal processing techniques for synthesizing, processing, and detecting ultrafast waveforms. The generation of spectrally decomposed waves within a spectral processing device is described in the following section. The subsequent section articulates on the employed nonlinear wave mixing processes, and the ensuing sections express the synthesis, processing, and detection of ultrafast waveforms.

Generation and Characteristics of Spectrally Decomposed Waves

Our wave mixing experiments are performed with spectrally decomposed waves, in which the spectral frequency content is spatially dispersed. When the pulse is decomposed with high resolution, the signal at every spatial location can be treated as quasi-monochromatic. This allows for increased interaction length within a nonlinear crystal, allowing us to use long crystals, whereas typically the interaction length is limited by temporal walkoff. In this section we analyze the creation of spectrally decomposed waves.

We introduce an ultrashort optical waveform, with a temporal envelope of $s(t/\tau)$, into the spectral processing device, where τ parameterized the waveform's duration. The waveform is propagating in free space towards a diffraction grating, at an incidence angle θ (see Figure 1). The input ultrashort waveform is characterized in its coordinate system (x_1, z_1) as:

$$E_{\text{pulse}}(x_1, z_1; t) = w\left(\frac{x_1}{L_1}\right) s\left(\frac{t - t_0 - z_1/c}{\tau}\right) \times \exp\left(\mathrm{i}\left(\omega_0 t - \frac{\omega_0}{c} z_1\right)\right) \qquad [1]$$

where $w(\cdot)$ defines the transversal field distribution with a scale size parameter L_1, t_0 is a time delay parameter, c is the speed of light in vacuum, and ω_0 is the center optical frequency. The waveform is

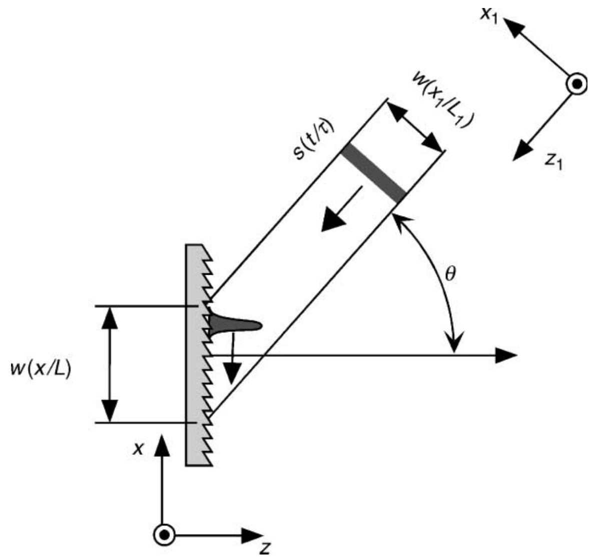

Figure 1 An ultrashort optical waveform, $s(t/\tau)$, propagating in free-space is incident upon a diffraction grating at angle θ. The transverse spatial mode of the pulse is stretched due to the projection onto the grating.

propagating in the z_1 direction at a group velocity and phase velocity of the speed of light in vacuum. The spatial confinement in the y_1 direction is omitted for brevity. We perform a time-domain Fourier transform (TFT) on eqn [1] to perform a frequency decomposition of the ultrafast waveform, yielding:

$$\tilde{E}_{\text{pulse}}(x_1, z_1; \omega) = \tau w\left(\frac{x_1}{L_1}\right) \times \exp\left(-\mathrm{i}\frac{\omega}{c}z_1\right)\tilde{s}(\tau(\omega-\omega_0)) \times \exp(-\mathrm{i}(\omega-\omega_0)t_0) \qquad [2]$$

where the tilde over the script denotes a FT relationship, defined for the TFT as $\tilde{s}(\omega)A\int s(t)\exp(-\mathrm{i}\omega t)\mathrm{d}t$. Next we perform a rotation of the coordinate system from (x_1, z_1) to (x, z), for compatibility with the coordinate system of the spectral processing device, yielding:

$$\tilde{E}_{\text{pulse}}(x, z; \omega) = \tau w\left(\frac{x\cos\theta - z\sin\theta}{L_1}\right) \times \exp\left(-\mathrm{i}\frac{\omega}{c}(x\sin\theta + z\cos\theta)\right) \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0) \qquad [3]$$

To find the incident field on the input grating of the spectral processing device, the field of eqn [3] is evaluated at $z = 0$. The effect of the grating diffraction can be modeled by adding the grating momentum, k_g, in the x direction to the k-vector that characterizes the propagation direction of the wave, resulting in:

$$\tilde{E}_{\text{input}}(x, 0; \omega) = \tau w\left(\frac{x\cos\theta}{L_1}\right) \times \exp\left(-\mathrm{i}\left(\frac{\omega}{c}\sin\theta - k_g\right)x\right) \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0) \qquad [4]$$

The grating momentum, k_g, and the incidence angle, θ, are chosen such that the center frequency ω_0 will diffract in the direction of the optical axis of the system. Setting $\omega_0/c\cdot\sin\theta = k_g$ and substituting the grating's k-vector, $k_g = 2\pi/\Lambda$, where Λ is the grating period, yields $\sin\theta = \lambda_0/\Lambda \equiv \alpha$, where λ_0 is the optical wavelength corresponding to the center carrier frequency. To further simplify the notation, we define a new spatial width scaling parameter $L = L_1/\cos(\theta)$, accounting for the aperture size increase due to the projection onto the diffraction grating. With these substitutions, the expression takes on the more familiar form, given by

$$\tilde{E}_{\text{input}}(x; \omega) = \tau w\left(\frac{x}{L}\right)\exp\left(-\mathrm{i}\frac{(\omega-\omega_0)\alpha x}{c}\right) \times \tilde{s}(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0) \qquad [5]$$

We may perform an inverse TFT on eqn [5] to characterize the input signal in the time domain, yielding:

$$E_{\text{input}}(x; t) = w\left(\frac{x}{L}\right)s\left(\frac{t - t_0 - \alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \qquad [6]$$

Equation [6] describes the input waveform scanning across the fixed aperture at velocity of c/α in the x-direction.

The spectrally decomposed wave is generated by performing a spatial Fourier transform (SFT) employing a lens of focal length f on the input wave (either frequency domain representation, eqn [4], or time domain representation, eqn [6]). Using the frequency domain representation, the spectrally decomposed wave U_{SDW} of the input waveform is given by

$$U_{\text{SDW}}(x'; \omega) = \tau\tilde{s}(\tau(\omega-\omega_0)) \times \exp(-\mathrm{i}(\omega-\omega_0)t_0)\int_{-\infty}^{\infty} w\left(\frac{x}{L}\right) \times \exp\left(-\mathrm{i}(\omega-\omega_0)\frac{\alpha x}{c}\right) \times \exp\left(-\mathrm{i}2\pi\frac{xx'}{\lambda_0 f}\right)\mathrm{d}x \qquad [7]$$

where we use the central wavelength in the Fourier transform kernel. Evaluating the integration of eqn [7], yields:

$$U_{\mathrm{SDW}}(x';\omega) = L\tau\tilde{s}\,(\tau(\omega-\omega_0))\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \qquad [8]$$

where the SFT of the input pupil function, $\tilde{w}(f_x)A\int w(x)\exp(-\mathrm{i}2\pi x f_x)\mathrm{d}x$, has mixed space-domain and time-domain frequency variables in its argument, characterizing the dispersion. A spectral component shifted away from the center frequency, $\Delta\omega = \omega - \omega_0$, will be centered at a displaced location $\Delta x' = -\Delta\omega\cdot\alpha f/\omega_0$, defining the spatial dispersion term $\partial x'/\partial\omega = -\alpha f/\omega_0$. One can see that in order to get fine localization of each spectral component, the spatial extent of the function $\tilde{w}(\cdot)$ must be minimized. This can be readily achieved by increasing the size of the input aperture, or the parameter L, as well as increasing α (grating's angular dispersion). Let us make the following observation on system resolution: The ultrafast pulse's spectral width is proportional to $1/\tau$, and its spatial extent is therefore proportional to $\alpha f/\omega_0\tau$ (multiplying by the spatial dispersion term). The spatial width of each spectral component is proportional to $cf/\omega_0\tau$. The dimensionless ratio of these two spatial widths, which is indicative of the number of resolvable spectral components, is $N \sim \alpha L/c\tau$, where N is defined as the system spectral resolution. When the spectral processing device is designed to satisfy the requirement of $N \gg 1$, then the spectrally decomposed wave can be considered quasi-monochromatic at every spatial position. In our experiments, we use $\lambda_0 = 800$ nm, gratings of 600 lines/mm frequency ($\Lambda = 1666$ nm), $L \approx 1$ cm, and $\tau \approx 100$ femtoseconds, yielding $\alpha \cong 0.5$ and $N \sim 160$, easily satisfying the high spectral resolution requirement. Let us make a second observation on system resolution using the time domain representation of eqn [6]: The spatial extent of the input aperture is proportional to L, whereas the instantaneous spatial extent of the scanning waveform is proportional to $\tau c/\alpha$ (see Figure 2). The dimensionless ratio of these two spatial widths is again $N \sim \alpha L/c\tau$, exemplifying the concept of resolvable positions and demonstrating the conservation of the space-bandwidth product.

To further facilitate the ensuing analysis, let us assume that we operate in the high-resolution regime (satisfied with the values given above), and we seek to simplify the expressions defining the spectrally decomposed wave. The spectrally decomposed wave can be defined in the time domain by performing a SFT on eqn [6], yielding:

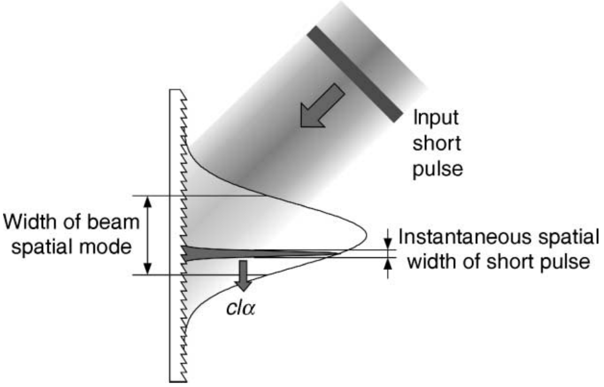
Input short pulse
Width of beam spatial mode
Instantaneous spatial width of short pulse
c/α

Figure 2 The input short pulse scans across the diffraction grating at velocity of c/α. The ratio of the width of the spatial mode of the beam to the instantaneous spatial width of the scanning waveform determines the system resolution. © 2001 Optical Society of America.

$$\begin{aligned} U_{\mathrm{SDW}}(x';t) &= \int w\left(\frac{x}{L}\right)s\left(\frac{t-t_0-\alpha x/c}{\tau}\right)\exp(\mathrm{i}\omega_0 t) \times \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}xx'\right)\mathrm{d}x \\ &\approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right) \times \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \end{aligned} \qquad [9]$$

where by account of the high-resolution approximation, the instantaneous location of the traveling signal waveform samples the value of the aperture function and removes the aperture dependence from the Fourier integral. This approximation decouples the aperture function from the temporal waveform, and illustrates that the duration of the spectrally decomposed wave is defined by the aperture function (time of flight through aperture of size L), and that the spatial dependence is a mapping of the incident waveform spectrum. A rotating wavevector in time is also present, on account of the traveling (moving) waveform on the input plane. Again, we may perform a TFT on eqn [9], defining the high-resolution limit of the spectrally decomposed wave, yielding:

$$U_{\mathrm{SDW}}(x';\omega) = \tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(-\mathrm{i}(\omega-\omega_0)t_0) \times \tilde{w}\left(\frac{\alpha L}{2\pi c}\left[\frac{\omega_0 x'}{\alpha f}+(\omega-\omega_0)\right]\right) \qquad [10]$$

One can compare eqns [8] and [10], as both characterize the spectrally decomposed wave in the

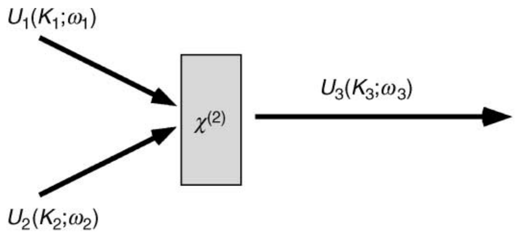

(a) Three-wave mixing in a $\chi^{(2)}$ crystal

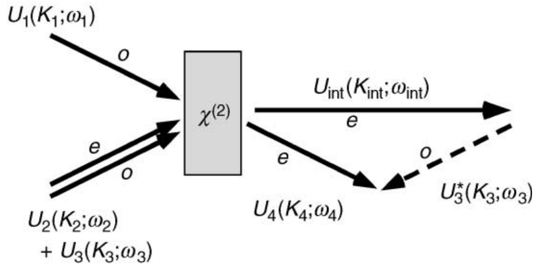

(b) Four-wave mixing by cascaded $\chi^{(2)}$ processes

Figure 3 (a) Three-wave mixing interaction: two input waves U_1 and U_2, generate a third wave U_3. (b) Four-wave mixing by cascaded second order nonlinearities: three input waves, U_1, U_2, and U_3, generate a fourth wave U_4. The first two waves generate an intermediate wave U_{int}, by a frequency-sum process, followed by a frequency-difference process between the intermediate wave and the third input wave.

temporal frequency domain, and observe that the difference lies with the argument of the spectral term. If we apply the high-resolution approximation to eqn [8], then each spectral component would occupy a small region in space, as defined by the extent of the aperture function's SFT. Substituting for $\omega - \omega_0$ by $-\omega_0 x'/\alpha f$ for the spectral argument of eqn [8], corresponding to the high-resolution limit, equates the two expressions (essentially the same approximation as was performed in deriving eqn [9]).

Nonlinear Wave Mixing With Nondepleting Pumps

A nonlinear crystal exhibiting a large nonlinear susceptibility $\chi^{(2)}$ is placed at the Fourier plane of the spectral processing device. We employ noncollinear phase matching in our experiments. Either type-I or type-II phase matching conditions can be used in three-wave mixing experiments, with the resultant frequency-sum wave propagating along the bi-sector direction (see Figure 3a). Since both input waveforms are at ω_0 frequency, the generated field will be at a doubled carrier frequency, $2\omega_0$, due to the frequency-sum process. Four-wave mixing experiments use type-II phase matching in the $\chi^{(2)}$ crystal, involving a frequency-sum process generating an intermediate wave followed by a frequency-difference process between the intermediate wave and the third input wave. The resultant wave will co-propagate with one of the input waves and at the same temporal carrier frequency (propagation direction and frequency ω_0), albeit at an orthogonal polarization state (see Figure 3b). The output signal can be extracted using a polarizer or a polarizing beam splitter. The interacting input waves propagate noncollinearly at the Fourier transform plane by offsetting the signals from each other at the input plane of the spectral processing device.

While it is possible, though often difficult, to obtain a precise analytic solution to the coupled mode equations governing the wave-mixing processes for the generated waves, it is beyond the scope of this article whose focus is on the signal processing functionality. We therefore limit the analysis to weak interactions with nondepleting input waves. Using the nondepleted input waves approximation, the nonlinear polarization term, which drives the generation of new waves, is constant and an ordinary differential equation characterizes the generation process.

The nonlinear polarization for the frequency-sum process is proportional to the product of the input waves, when expressed in the time domain: $P_{\text{NL}}^{\text{sum}}(x';t) = \chi_{\text{eff}}^{(2)} U_1(x';t)U_2(x';t)$. The nonlinear polarization for the frequency-difference process is proportional to the product of an input waves and a conjugated input wave: $P_{\text{NL}}^{\text{diff}}(x';t) = \chi_{\text{eff}}^{(2)} U_1(x';t)U_2^*(x';t)$.

The solution of the ordinary differential equation driving the generation of the new wave, takes into account the possible phase mismatch between the propagation of the nonlinear polarization and the generated wave within the nonlinear crystal medium. Since we are interested in the signal processing functionality only, we approximate the generated waves as being directly proportional to the driving nonlinear polarization. This approach is valid for obtaining the generated signal forms in the case of short nonlinear crystals (limited interaction length) satisfying the phase matching condition. In practice, the temporal and spatial bandwidths, as well as the crystal length, will introduce filtering effects which will slightly modify the output signals. However, it can be shown that when wave mixing with spectrally decomposed waves, especially mutually inverted waves, the phase matching requirement is well satisfied and the filtering effects are negligible for the employed crystal lengths. In our experiments, we typically used 2 mm long β-barium borate (BBO) crystals. Using the simplified method for obtaining the form of the generated waves, arrangements for synthesizing, processing, and detecting ultrafast waveforms can be explored.

Detection of Ultrafast Waveforms

Ultrafast optical waveforms are too rapid for acquisition by opto-electronic means alone, due to the latter's finite response time. An optical processor is usually required to convert a finite time window (typically in the tens to hundreds of ps range) to a stationary signal in space for detection by slower means. We mix two mutually inverted spectrally decomposed waves in the Fourier plane of a spectral processing device, and spatially Fourier transform the resultant wave to the output plane. As will be shown below, when one of the spectrally decomposed waves is generated by a transform limited short pulse, the output spatial signal carries the temporal information of the input waveform. We call this process a time-to-space converter.

Let two ultrafast waveforms, $s((t-t_0)/\tau)$ and $r(t/\tau)$, enter the spectral processing device from opposite directions (one strikes the grating at angle θ and the other at $-\theta$, see Figure 4). Each waveform will generate a spectrally decomposed wave in the spatial Fourier transform plane. The spectrally decomposed wave of the signal $s(t)$, as defined by eqn [9], yields the first input field:

$$U_1(x';t)\approx w\left(\frac{c(t-t_0)}{\alpha L}\right)\exp\left(-\mathrm{i}\frac{\omega_0(t-t_0)x'}{\alpha f}\right)\times\tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \qquad [11a]$$

whereas the inverted spectrally decomposed wave for $r(t/\tau)$ is given by

$$U_2(x';t)\approx w\left(-\frac{ct}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0 t x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}\omega_0 t) \qquad [11b]$$

Without loss of generality, we set the time delay parameter in eqn [11b] to zero, such that the parameter t_0 in the spectrally decomposed wave of the signal waveform is a measure of the relative time delay of the two waveforms. With full knowledge of the two input waves, U_1 and U_2, we can express the resultant waveform of the three-wave mixing process as:

$$U_3(x';t)\propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t)=w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right)\times\exp\left(\mathrm{i}\frac{\omega_0 t_0 x'}{\alpha f}\right)\tilde{s}\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{r}\left(\frac{\omega_0\tau x'}{\alpha f}\right)\times\exp(\mathrm{i}2\omega_0 t) \qquad [12]$$

Next we apply a SFT to propagate the wave U_3 to the output plane, yielding:

$$U_{\mathrm{out}}(x'';t)=\int U_3(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f/2}x'x''\right)\mathrm{d}x' =y\left(\frac{2\alpha x''}{c}-t_0\right)w\left(\frac{c(t-t_0)}{\alpha L}\right)w\left(-\frac{ct}{\alpha L}\right)\times\exp(\mathrm{i}2\omega_0 t_0) \qquad [13]$$

where $y(t)=s(t/\tau)\otimes r(-t/\tau)$, as the Fourier transform of a product is the convolution of their Fourier transforms (the $\otimes$-symbol denotes the convolution operation), and we use the spatial Fourier transform kernel at the doubled carrier frequency. The stationary spatial signal, $y(\cdot)$, is proportional to the convolution of the two input waveforms, where

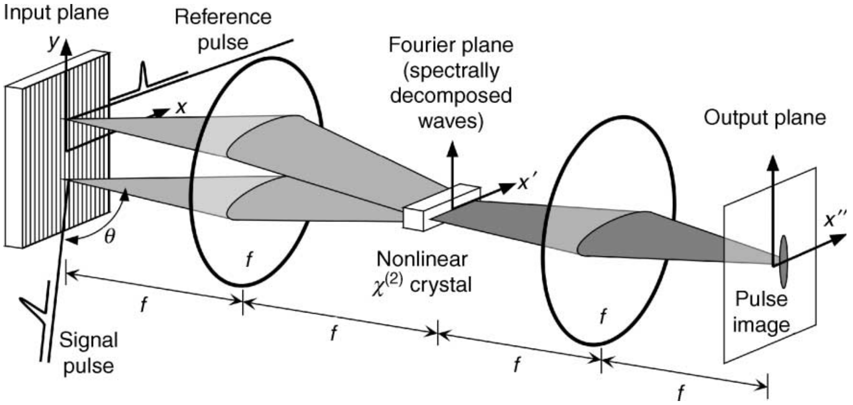

Figure 4 Arrangement for imaging ultrafast waveforms using a time-to-space conversion. The mutually inverted, spectrally decomposed waves of a signal waveform and that of a reference pulse are mixed in a nonlinear crystal placed at the Fourier plane. The generated wave is spatially Fourier transformed to the output plane, yielding a stationary image that carries the signal waveform information. © 2001 Optical Society of America.

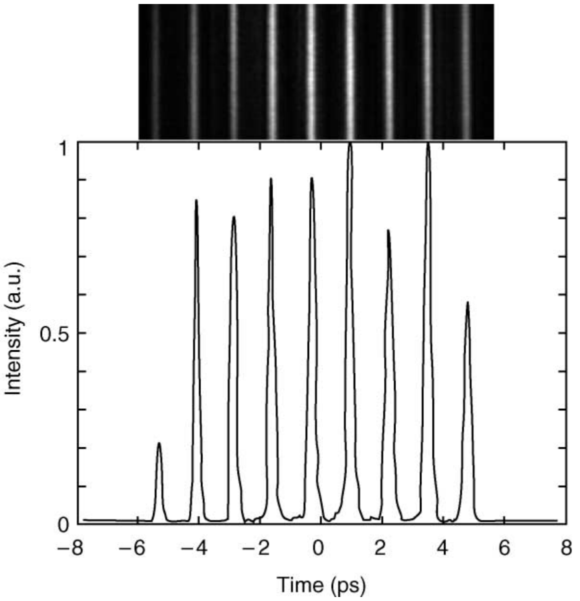

Figure 5 Output of a time-to-space converter. Top: image acquired by a CCD camera of the stationary image of a pulse sequence. Botttom: extracted time domain data from the image, showing inter-pulse time separation of 1.25 ps.

one of them is reversed in time. This mixing process was developed for time-to-space conversion by using a reference transform limited pulse for the reversed waveform. The resulting spatial image will be that of the signal waveform, albeit smoothed out due to the convolution operation with the reference pulse. The spatial signal is centered at $x'' = ct_0/2\alpha$, illustrating that the relative time delay between the two waveforms shifts the location of the output signal. The temporal characteristic is determined by the product of the time-scaled input apertures. The relative time delay, t_0, serves as an attenuation mechanism when the time windows are not perfectly aligned. When the time delay between the signal pulse and the reference pulse is too large, no output signal will emerge as the two signals do not overlap for the wave mixing process. Therefore, the time window of the spectral processing device is determined by the input aperture size, not the physical dimension of the nonlinear crystal. Consequently, we have been able to demonstrate high-resolution temporal imaging with time windows in the range of tens of picoseconds (see Figure 5). Additionally, as mentioned earlier, one of the significant advantages of this technique is the favorable phase matching condition with the spectrally decomposed wave and its inverted counterpart.

Processing of Ultrafast Waveforms

The simplest form of ultrafast waveform processing uses linear filtering, or placement of an element with transmittance characteristics that are frequency selective. However, a linear filter is a time invariant process and cannot perform more advanced processing operations of ultrafast waveforms such as interaction between ultrafast waveforms. For such processing applications, we mix three spectrally decomposed waves in the Fourier plane of a spectral processing device to generate a resultant output ultrafast waveform in real time. We illustrate advanced processing by performing a time reversal experiment using the spectral phase conjugation and spectral inversion techniques in four-wave mixing experiments.

Consider three ultrafast waveforms, $s_1(t - t_0/\tau)$, $s_2((t - t_1)/\tau)$ and $s_3(t/\tau)$, which enter a spectral processing device for generation of their respective spectrally decomposed waves. Let all three waves enter the spectral processing device from the same direction, such that the spatial dispersion is in the same direction for each signal (as defined by eqn [11a]), and generate the spectrally decomposed waves $U_1(x';t)$, $U_2(x';t)$, and $U_3(x';t)$, respectively (see Figure 6a). Since we utilize a cascaded second-order nonlinearity process, we let the first two waves interact first in a frequency-sum process, generating the intermediate field:

$$U_{\mathrm{int}}(x';t) \propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(\frac{c(t-t_1)}{\alpha L}\right) \times \exp\left(-\mathrm{i}\frac{\omega_0(2t-t_0-t_1)x'}{\alpha f}\right) \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \times \tilde{s}_2\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\mathrm{i}2\omega_0 t) \qquad [14]$$

It should be noted that upon a more rigorous analysis of the generated frequency-sum wave, phase matching is not well met across the temporal bandwidth of the waveform, as is the case when wave mixing mutually inverted spectrally decomposed waves. This will introduce time domain spectra filtering effects, but for simplicity these effects are ignored below. The generated frequency-sum wave interacts with the third input wave in a frequency-difference process, giving rise to the fourth wave:

$$U_4(x';t) \propto P_{\mathrm{NL}}^{\mathrm{diff}}(x';t) = w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(\frac{c(t-t_1)}{\alpha L}\right) \times w\left(\frac{ct}{\alpha L}\right) \exp\left(-\mathrm{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \times \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_2\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_3^{*}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\mathrm{i}\omega_0 t) \qquad [15]$$

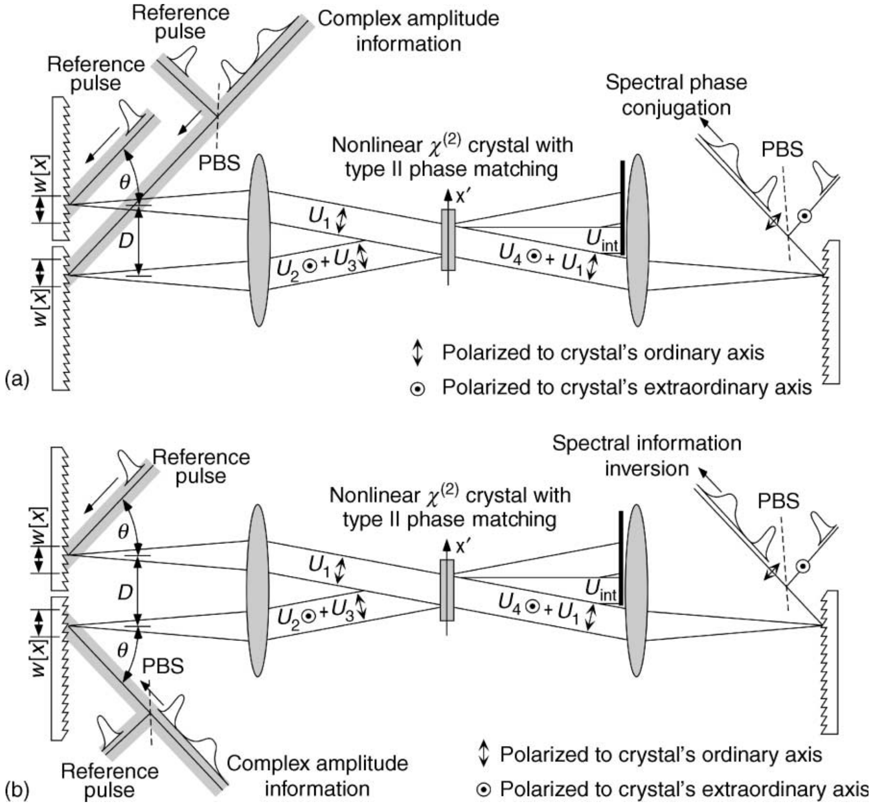
Reference pulse
Complex amplitude information
Reference pulse
Spectral phase conjugation
PBS
Nonlinear $\chi^{(2)}$ crystal with type II phase matching
PBS
x'
$w[x]$
θ
U_1
D
U_{int}
$U_2 \odot + U_3$
$U_4 \odot + U_1$
$w[x]$
Polarized to crystal's ordinary axis
Polarized to crystal's extraordinary axis
(a)
Reference pulse
Spectral information inversion
Nonlinear $\chi^{(2)}$ crystal with type II phase matching
PBS
$w[x]$
θ
x'
U_1
D
U_{int}
$U_2 \odot + U_3$
$U_4 \odot + U_1$
$w[x]$
θ
PBS
Reference pulse
Complex amplitude information
Polarized to crystal's ordinary axis
Polarized to crystal's extraordinary axis
(b)

Figure 6 Arrangements of processing ultrafast waveforms within a spectral processing device. (a) Three input waveforms are introduced from the same direction, and their spectrally decomposed waves interact via four-wave mixing. (b) The three input waveforms are introduced from opposite directions, giving rise to mutually inverted spectrally decomposed waves. These two arrangements can perform time reversal, the first by spectral phase conjugation and the second by spectral information inversion. © 2001 IEEE.

where we assume that the window function $w(\cdot)$ is real. Next we apply a SFT to propagate the wave U_4 to the output plane, yielding:

$$\begin{aligned} U_{\text{out}}(x'';t) &= \int U_4(x';t)\exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f}x'x''\right)\mathrm{d}x' \\ &= y\left(\frac{\alpha x''}{c}+t-t_0-t_1\right)w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad \times w\left(\frac{c(t-t_1)}{\alpha L}\right)w\left(\frac{ct}{\alpha L}\right)\exp(\mathrm{i}\omega_0 t_0) \end{aligned} \qquad [16]$$

where $y(t)=s_1(t/\tau)\otimes s_2(t/\tau)\otimes s_3^*(-t/\tau)$. The new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction (compare to the input signals scanning in the opposite direction along the input plane, eqn [6]). By placing another diffraction grating at this plane, the new ultrafast waveform can be recomposed and made to propagate once more in free space. The waveform information consists of a convolution between the first and second signals, followed by a convolution with the third signal, which is both conjugated and reversed. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the reversed and conjugated information of the third signal. Such capability can be very important for dispersion compensation in optical fiber communication links.

Let us now consider an alternate arrangement, where the second and third input signals are introduced to the spectral processing device from the opposite direction, and consequently have reversed spatial dispersion direction (see Figure 6b). Therefore, the first signal will generate a spectral decomposed wave $U_1(x';t)$, as defined by eqn [11a], whereas the second and third signals will give rise to the spectrally decomposed waves $U_2(x';t)$ and $U_3(x';t)$, with inverted spectra relative to that of U_1, as defined by eqn [11b]. The interaction of the first two waves gives rise to:

$$\begin{aligned} U_{\text{int}}(x';t) \propto P_{\text{NL}}^{\text{sum}}(x';t) &= w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\quad \times w\left(-\frac{c(t-t_1)}{\alpha L}\right)\exp\left(\mathrm{i}\frac{\omega_0(t_0+t_1)x'}{\alpha f}\right) \\ &\quad \times \tilde{s}_1\left(-\frac{\omega_0\tau x'}{\alpha f}\right)\tilde{s}_2\left(\frac{\omega_0\tau x'}{\alpha f}\right)\exp(\mathrm{i}2\omega_0 t) \end{aligned} \qquad [17]$$

which is identical to the wave mixing process of the time-to-space converter (see eqn [12]). However, in this case, we follow the first wave mixing process by a frequency-difference process, generating:

$$\begin{aligned} U_4(x';t) \propto P_{\mathrm{NL}}^{\mathrm{diff}}(x';t) &= w\left(\frac{c(t-t_0)}{\alpha L}\right) w\left(-\frac{c(t-t_1)}{\alpha L}\right) \\ &\times w\left(-\frac{ct}{\alpha L}\right) \exp\left(-\mathrm{i}\frac{\omega_0(t-t_0-t_1)x'}{\alpha f}\right) \tilde{s}_1\left(-\frac{\omega_0 \tau x'}{\alpha f}\right) \\ &\times \tilde{s}_2\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \tilde{s}_3^*\left(\frac{\omega_0 \tau x'}{\alpha f}\right) \exp(\mathrm{i}\omega_0 t) \qquad [18] \end{aligned}$$

Next we apply a SFT to propagate the wave U_4 to the output plane, yielding the output waveform:

$$\begin{aligned} U_{\mathrm{out}}(x'';t) &= \int U_4(x';t) \exp\left(-\mathrm{i}\frac{2\pi}{\lambda_0 f} x' x''\right) \mathrm{d}x' \\ &= y\left(\frac{\alpha x''}{c} + t - t_0 - t_1\right) w\left(\frac{c(t-t_0)}{\alpha L}\right) \\ &\times w\left(-\frac{c(t-t_1)}{\alpha L}\right) w\left(-\frac{ct}{\alpha L}\right) \exp(\mathrm{i}\omega_0 t_0) \qquad [19] \end{aligned}$$

where $y(t) = s_1(t/\tau) \otimes s_2(-t/\tau) \otimes s_3^*(t/\tau)$. As before, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction, and will propagate in free-space upon another grating diffraction. This time, however, the waveform information consists of a convolution between the first signal and a time-reversed second signal, followed by a convolution with the conjugated third signal. If we set the first and second signals to be transform limited pulses, then the resultant waveform will carry the conjugated information of the third signal. If we set the first and third signals to be transform limited pulses, then the resultant waveform will carry the reversed information of the second signal.

We have experimentally demonstrated the two time-reversal techniques: the first technique whereby the resultant signal is both reversed and conjugated is labeled 'spectral phase conjugation', as the signal's spectrum is conjugated by the frequency difference process, and the second technique, where the signal is purely reversed in time, is named 'spectral information inversion', as the spectrum is flipped about the carrier frequency (see Figure 7). By proper selection of the input waveforms, one can also obtain higher-level operations and functionality such as convolutions and correlations among the different signals. All these operations are time-variant, and cannot be performed by passive filtering alone.

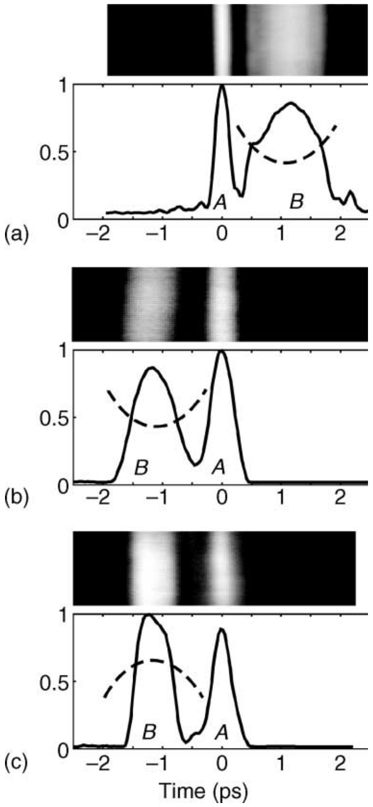

Figure 7 Time reversal experimental results. (a) Input pulse pair consisting of a transform limited pulse followed by a chirped pulse. (b) Time reversal by spectral information inversion technique. Chirped pulse leads transform limited pulse and maintains the sign of the chirp. (c) Time reversal by spectral phase conjugation technique. Chirped pulse leads transform limited pulse and flips the sign of the chirp. Data acquired with time-to-space imaging technique and sign of quadratic term (dashed line) determined by translating output plane.

Synthesis of Ultrafast Waveforms

Shaping of ultrafast waveforms is usually performed by filtering an input transform limited pulse by a fixed filter. Adaptive filtering can be performed by various means incorporating opto-electronic modulators such as liquid-crystal-based or acousto-optic-based spatial light moduloators, however the response time required for changing the waveform is typically limited to the ms and μs regime, respectively. For ultrafast reconfigurability, we utilize our spectral processing device and wave mixing, to achieve pulse shaping controlled by light, or a second optical signal.

Let the three input signals entering the spectral processing device be an ultrashort waveform, $s(t/\tau)$, and two additional spatial domain signals; one carrying spatial information $m(x/\chi)$, where χ is a scaling factor, and the second a point source (see Figure 8). Both spatial domain signals are illuminated

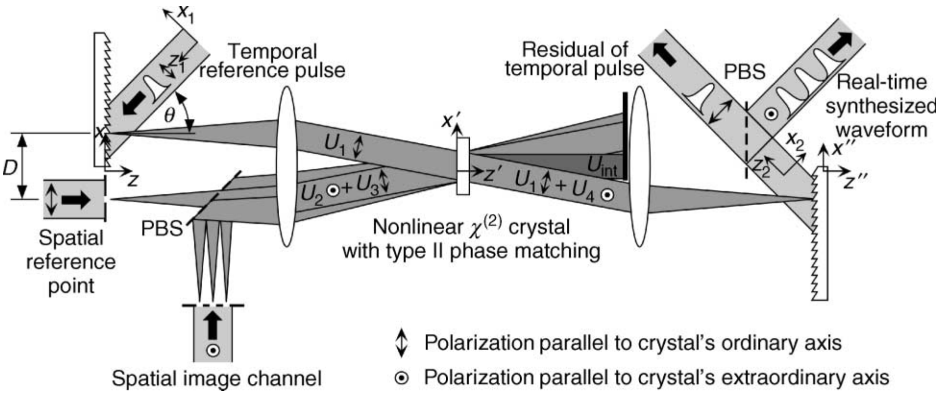
x_1
z_1
Temporal reference pulse
θ
x
z
D
Spatial reference point
PBS
Spatial image channel
U_1
U_2+U_3
x'
z'
U_1+U_4
U_{int}
Nonlinear $\chi^{(2)}$ crystal with type II phase matching
Residual of temporal pulse
PBS
Real-time synthesized waveform
x_2
z_2
x''
z''
Polarization parallel to crystal's ordinary axis
Polarization parallel to crystal's extraordinary axis

Figure 8 Arrangement for ultrafast waveform synthesis by wave mixing temporal and spatial information channels. Three input channels consist of a temporal channel, a spatial channel carrying information, and a second spatial channel consisting of a point source. The information from the spatial channel is transferred to the temporal channel via the wave mixing process. © 2001 IEEE.

by a monochromatic light source at frequency ω_0. The input ultrashort waveform gives rise to a spectrally decomposed wave $U_1(x';t)$, as described by eqn [11a]. The spatial information channel is SFT by a lens, generating the wave:

$$U_2(x';t) \propto \tilde{m}\left(\frac{\omega_0 \chi x'}{2\pi c f}\right) \exp(\mathrm{i}\omega_0 t) \qquad [20]$$

The signal from the point source (or $\delta(x)$) gives rise to a wave $U_3(x';t)$ with no spatial information (plane wave). As before, we utilize cascaded second-order nonlinearities for the wave mixing process. The interaction of the first two waves gives rise to the intermediate frequency-sum wave:

$$\begin{aligned} U_{\mathrm{int}}(x';t) &\propto P_{\mathrm{NL}}^{\mathrm{sum}}(x';t) \\ &= w\left(\frac{ct}{\alpha L}\right)\tilde{s}\left(-\frac{\omega_0 \tau x'}{\alpha f}\right)\tilde{m}\left(\frac{\omega_0 \chi x'}{cf}\right)\exp(\mathrm{i}2\omega_0 t) \end{aligned} \qquad [21]$$

The frequency-difference process with an informationless waveform serves to reduce the carrier frequency of resultant waveform back to ω_0. The down-converted wave U_4 is SFT to the output plane, yielding:

$$U_{\mathrm{out}}(x'';t) = y\left(\frac{\alpha x''}{c} + t\right) w\left(\frac{ct}{\alpha L}\right) \exp(\mathrm{i}\omega_0 t_0) \qquad [22]$$

where $y(t) = s(t/\tau) \otimes m(-ct/\alpha\chi)$. As in the processing of ultrafast waveforms section above, the new ultrafast waveform $y(\cdot)$ scans along the output plane at a velocity of $-c/\alpha$ in the x-direction. The synthesized ultrafast waveform can propagate in freespace after an additional grating diffraction. The waveform information, in this case, consists of a convolution between the input temporal signal and a time-scaled and reversed version of the spatial information channel. Since the input temporal channel is typically a transform limited pulse, the output waveform carries the spatial information image, mapped to the time domain. In analogy to the imaging of ultrafast waveforms section, we call this technique space-to-time conversion. We have generated experimentally numerous waveforms that are controlled by a space domain image, or mask (flat-top waveforms, Figure 9, and pulse sequences of Figure 5 were generated by a cylindrical lens array in the spatial information channel).

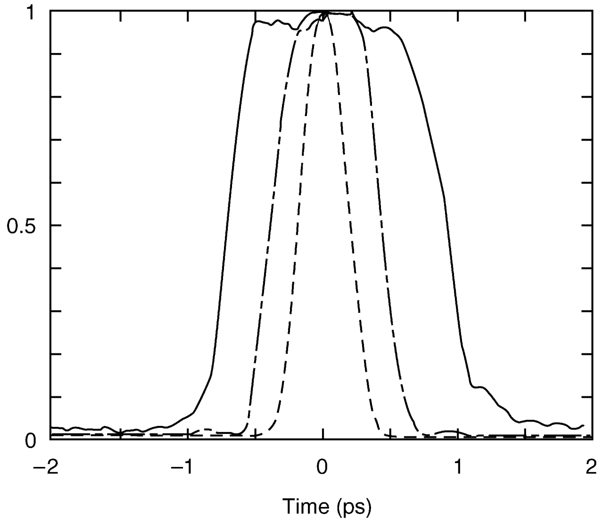

Figure 9 Synthesized flat-top waveforms via space-to-time conversion utilizing a variable slit in the spatial information channel. Increasing the width of the slit results in broader square-shaped pulses in real-time. Solid line: 2 ps pulse, dot-dash line: ~1 ps pulse, dashed line: ~300 fs pulse.

Conclusions

To properly utilize ultrafast waveforms in scientific and advanced technology applications requires instantaneous techniques from generation to processing and imaging of the waveforms. We have studied the nonlinear wave-mixing approach within a spectral processing device, and demonstrated the ability to perform all of these tasks. The wave mixing was performed in a second-order nonlinear crystal, using three-wave mixing for detection and four-wave mixing for synthesis and processing of ultrafast waveforms. The three-wave mixing process results in a second-harmonic output image, suitable for acquisition by silicon-based CCD imagers. The four-wave mixing processes result in an output waveform of an identical carrier frequency, as typically required in processing and synthesis applications. The high nonlinear coefficient of $\chi^{(2)}$ crystals enabled us to demonstrate these techniques with ultrashort pulse energies ranging from nJ to mJ levels. Further advantages afforded by the techniques are expanded time windows and favorable phase matching.

This review of the instantaneous optical signal processing capabilities with optical nonlinearities and interaction with ultrafast waveforms introduced and demonstrated the basic signal processing capabilities to manipulate optical signals in amplitude and phase. The same techniques can be used and expanded upon to achieve different signal processing abilities, which are beyond the scope of this article.

See also

Coherent Transients: Ultrafast Studies of Semiconductors. **Nonlinear Optics, Basics:** Ultrafast and Intense-Field Nonlinear Optics. **Ultrafast Laser Techniques:** Pulse Characterization Techniques. **Ultrafast Technology:** Femtosecond Chemical Dynamics: Gas-Phase; Femtosecond Condensed Phase Spectroscopy: Structural Dynamics; Ultrafast Illumination and Processing.

Further Reading

Diels J-C and Rudolph W (1996) *Ultrashort Laser Pulse Phenomena: Fundamentals, Techniques, and Applications on a Femtosecond Time Scale*. San Diego: Academic Press.

Marom DM, Panasenko D, Sun P-C and Fainman Y (2000) Femtosecond-rate space-to-time conversion. *Journal of the Optical Society of America B* 17: 1759–1773.

Marom DM, Panasenko D, Sun P-C, Mazurenko YY and Fainman Y (2001) Real-time spatial-temporal signal processing with optical nonlinearities. *IEEE Journal of Selected Topics in Quantum Electronics* 7: 683–693.

Miller DR, Murnane MM, Scherer NF, Weiner AM (eds) (2003) *Ultrafast Phenomena XIII, Proceeding of the 13th International Conference*. Berlin: Springer-Verlag.

Weiner AM (1995) Femtosecond optical pulse shaping and processing. *Progress in Quantum Electronics* 19: 161–238.

Weiner AM and Kanan M (1998) Femtosecond pulse shaping for synthesis, processing, and time-to-space conversion of ultrafast optical waveforms. *IEEE Journal of Selected Topics in Quantum Electronics* 4: 317–331.

EARLY HISTORY OF QUANTUM ELECTRONICS

C H Townes, University of California, Berkeley, CA, USA

© 2005, C H Townes. Published by Elsevier Ltd. All rights reserved.

All the ideas essential to making a laser were known before 1930, but there was no operating laser before 1960. So why didn't the laser come sooner? There are several reasons. One important impediment to the laser invention was that a combination of ideas from quantum mechanics and from electrical engineering was needed, and these two fields were not well mixed in the early days. Another is that, while some physicists recognized that amplification could occur if there was a population inversion of states, they did not consider the coherence of such amplifications and did not recognize its importance or usefulness. It was just something that in principle could happen, but was not very interesting. What was required was a combination of physics and electrical engineering thinking and recognition of the coherence of such amplification. In addition the importance of such a development had to be visualized and recognized in a way that it led to very devoted work towards its achievement.

It is striking that lasers grew out of the study of microwave spectra of molecules, for which engineering and quantum mechanics were both important and which helped orient thinking in the appropriate direction. That this origin was not accidental is convincingly demonstrated by the fact that three independent ideas for amplification by stimulated emission were generated about 1950. They were from Joe Weber, at the University of Maryland, from Nikolai Basov and Alexander Prokhovov at the Soviet Academy, and myself at Columbia University. Weber primarily wanted to point out the possibility, but didn't try to do it. In addition, his numbers were a bit off and no practical system was suggested. Basov and Prokhovov actually worked towards microwave amplification using a beam of molecules, as did I.

That stimulated amplification was recognized early, but not thought through, is illustrated by the fact that Prof. Richard Tolman, a theoretical physicist, wrote a discussion in 1924 of the net absorption of light by molecules, pointing out in particular that induced emission counteracted absorption and noting that if there were more molecules in the upper than in the lower state there could be 'negative absorption'. But, he wrote, 'This would usually be very small'. The Russian physicist Vitaly Ginsburg wrote me, after the maser and laser had appeared, that his professor, S.M. Levi, had been well aware of such effects back in the 1930s and had told him 'create an overpopulation at higher atomic levels and you will obtain an amplifier; the whole trouble is that it is difficult to create a substantial overpopulation of levels'.

The German physicist F.G. Houtermans said to me that in 1932, when told by a colleague of an unusual light intensity in a gaseous discharge, he had thought it might be a 'photon avalanche', i.e. multiplication of photons by stimulated emission.

Another Russian physicist, V.A. Fabricant, wrote a thesis in 1939 in which he discussed absorption and emission of light radiation in a gas and looked for 'negative absorption', or amplification. He did not discuss coherence or a resonant cavity, and was not able to achieve any amplification so his work was quickly forgotten. None of these early mentions of stimulated emission proposed how to actually get amplification, that it would be useful, nor clearly noted its coherence. Tolman did, however, write in 1927 that, 'We should expect radiation induced by an external field to be coherent with the radiation associated with that field'. I know of no proposal to actually make use of such amplification, until those made by microwave spectroscopists in the early 1950s.

Another clear indicator that even in the 1950s physicists and engineers did not think amplification by stimulated emission was particularly interesting nor useful is that during the $2\frac{1}{2}$ years when Jim Gordon, Herb Zeiger, and I were working on trying to obtain microwave amplification (the maser), a large number of scientists visited my laboratory, saw what we were doing, but no one bothered to also try to obtain such amplification. After the maser worked, it hit the newspapers and then very quickly became a popular and intense field of interest for a number of physicists.

My own drive to produce oscillators by stimulated emission came from my strong interest in obtaining sources of waves shorter than the few millimeters wavelength which could be produced by electronic devices, in order to extend the high resolution study of molecular spectra to wavelengths shorter than microwaves, down into the far infrared. My students and I worked on several possible schemes for producing waves shorter than those produced by klystrons or magnetrons – frequency multiplication by nonlinearities, electronic beams passing over surfaces of solid materials with resonances, and anything else I could think of. None worked very well.

In early 1950, I was asked by the Office of Naval Research to form and chair a committee which would examine possible research towards obtaining wavelengths down to one millimeter and shorter. I chose outstanding scientists and engineers in a variety of fields which might touch on this problem. We met together, and visited many pertinent laboratories and individuals interested in such research. Nothing very promising seemed to turn up. But we wanted, of course, to at least provide a report summarizing the situation as we saw it. Our last meeting was April 26, 1951, in Washington, D.C. Worrying over our lack of success, I woke up early before the meeting. Breakfast was not ready, so I walked over to nearby Franklin Park, sat on a bench in front of beautiful blooming azaleas, and puzzled over why neither I nor the Committee had found any promising solutions.

I went over all the ideas I had had previously. Molecules, of course, can produce high frequencies. But I had previously concluded, I thought wisely and rigorously, that one could not obtain intense radiation from them because radiation intensity was a function of temperature, and the temperature could not be very high without destroying the molecules. Suddenly I realized that they did not need to have a temperature in the usual sense – they need not be in temperature equilibrium. There could be more molecules in an upper than a lower excitation state, which could in principle produce indefinitely intense radiation. I pulled a piece of paper out of my pocket and wrote down the equations and numbers for such a case, using a molecule beam sent into a resonant cavity and ammonia molecules with which I was very familiar. My equations said one could get enough excited molecules, low enough loss in a resonant cavity, and it would work! Why hadn't I thought of it before?!

Back at Columbia, where I worked, and about 4 months later, the graduate student Jim Gordon agreed to work on trying to obtain such an oscillator using a beam of ammonia. I assured him that if it didn't work he could modify the experiment to do interesting spectroscopy and thus complete a thesis. But we both thought he had a good chance of making it work. And a young post doc working with me, Herb Zeiger, joined the effort.

Actually, I had first thought of obtaining stimulated emission from a molecular beam back in 1948, but simply as a demonstration of physical principles rather than as a useful amplification. Also, about a year later a young post doc, J.W. Trischka, working on molecular beams with Professors Rabi and Kusch, had also thought of demonstrating stimulated emission. We talked about it together, and he decided it wasn't worthwhile just demonstrating the effect because it wouldn't really prove any new physics. Neither of us at that time had recognized the real point and the possibility of useful amplification, which is one of the reasons mentioned above that the idea was delayed as long as it was.

It is perhaps important to emphasize again and to illustrate how out-of-the-way the use of stimulated emission was at that time for physicists, and how distant stimulated emission was from engineers. While we were working on the ammonia beam maser, Prof. L.H. Thomas, an outstanding physicist known for the 'Thomas Effect', frequently would run into me in the hallway at Columbia University and say that I didn't understand, and that my proposed ammonia oscillator could not work. However, I never got a clear explanation from him of what it was I didn't understand. And after we had been working on the ammonia maser system for about 2 years (a more-or-less normal time for a student thesis project), the Physics Department Chairman, Prof. Polycarp Kusch, and the previous chairman, Prof. I.I. Rabi, came into my office to object. They were excellent physicists, both received Nobel Prizes, and both were experts on molecular beams. They sat down in my office and said 'Look, Charlie, that is not going to work. We know it won't work and you know it won't work. You are wasting departmental money, and must stop!' Other people had also questioned what I was doing, frequently in particular whether the stimulated radiation would be coherent. So I had already thought over the situation many times. I had full notes from the quantum mechanics course I took as a student back in 1938, and could derive from them a proof of coherence. I felt also that I knew the quantitative numbers, such as the molecular beam intensity and the possible 'Q', or loss, in the cavity resonator, very well. I was also by then an Associate Professor, and the department chairman could not fire me simply because of stupidity. I replied to Rabi and Kusch 'No, I believe it has a reasonable chance of working and I'm going to continue!' Annoyed, they stomped out of the room.

About 2 months after the Rabi/Kusch incident, Jim Gordon dashed into the classroom where I was lecturing and said loudly 'It's working!' Most of the class then went up to the lab to see the new device. Rabi and Kusch were not against me; even though they were outstanding physicists, they probably just didn't quite comprehend the device. A couple of months after its successful operation, Kusch more or less apologized by saying 'Well, I should have realized that you probably know more about what you are doing than I do.'

After the maser successfully operated, there were other incidences showing the lack of appropriate focus of physicists on stimulated emission. I was a friend of Aage Bohr, the son of Niels Bohr, and in that connection was visiting him in Denmark. Niels Bohr asked me what research I was presently doing, so I told him about our new oscillator, the maser, and its remarkably pure frequency. He looked at me and said 'Oh no, that's not possible. You must be misunderstanding something.' I emphasized again what it was really doing, but he still seemed not to believe it could function that way. I presumed he was thinking in terms of the uncertainty principle and the finite time of passage of a molecule through the cavity, though I never was quite sure just why he felt it impossible. A similar thing happened shortly after that at a cocktail party in Princeton, where John Von Neumann asked the same question – what was my research at the moment? After telling him about the maser oscillator and the purity of frequency he reacted very similarly 'Oh no,' he said, 'That can't be right. You must be misunderstanding something.' After arguing a little more, he left to get another drink. Fifteen minutes later he came back, saying 'Hey, you are right!' He had understood, and wanted to talk much more about the maser and of possibly using excited semiconductors. Only after his death did I learn from his notebooks that he had considered exciting electrons in semiconductors with neutrons from a reactor, and had written Edward Teller about whether some experiments might be done with this to obtain intense light. However, he had not considered coherence, and Teller was apparently not interested enough to respond, so the matter was dropped.

The above account reminds me a bit of the amusing comments of Arthur Clark on 'Change'. He writes:

'People go through four states before any revolutionary development:

1. It's nonsense, don't waste my time
2. It's interesting, but not important
3. I always said it was a good idea
4. I thought of it first'.

It's clear that the world of physics was thinking very little in the direction of masers or lasers, and that many preconceptions as well as lack of interest stood in the way. My own experience with engineering at Bell Telephone Labs during World War II, in designing radar and electronic systems, plus my intense interest in obtaining short-wave oscillators, were clearly important in bringing me to the right ideas.

As masers became very interesting to the physics community and the field grew rapidly, my engineering experiences continued to be of help. I was well acquainted with the theoretical examination of noise in vacuum tube amplifiers by various individuals at Bell Labs, and recognized that the maser could provide much more sensitive amplification than could common electronic amplifiers, where discrete electronic charges produce the basic noise. On sabbatical in France, I worked on electron spin masers with Jean Combrisson and Arnold Honig who had the appropriate equipment. And then in Japan, Koichi Shimoda, Hidetoshi Takahashi, and I wrote a theoretical paper on the basic quantum noise of maser (or laser) amplification. The theory showed that maser amplifiers of microwaves should be about 100 times more sensitive than the existing electronic ones.

After 2–3 years of maser experiments and development I felt I wanted to move on to the shorter wavelengths for which I had generated the maser idea. Although I had first tried the idea at microwave frequencies because that seemed the easiest way to test out the general idea and the result had been exciting, I still wanted those shorter wavelengths. I had not come up with any great ideas of just how to get to much shorter wavelengths, which is why I waited several years after the maser worked before moving on. However, in 1957, 3 years after the first successful operation of the maser, I decided it was high time to simply figure out what was the best way I could imagine to move on into the infrared region and do it. A number of physicists had concluded that of course one couldn't make masers work at much shorter wavelengths, certainly not in the visible region, because spontaneous emission becomes so much faster as the wavelength is shortened and adequate inversion of population was not practical. But that was intuition, not quantitative science. As I wrote down equations for what might be done to move towards shorter wavelengths using a model of atomic or molecular excitation by radiation and a reasonably high Q resonator, it became clear that it was quite practical to move on even into the visible region. That was exciting! Why hadn't I, or someone else, looked at it carefully and quantitatively before that time?

I was at the time consulting at Bell Labs, with the assignment to spend a day there every 2 weeks and just talk with the Lab's scientific personnel. Since my former post doc and now brother-in-law Arthur Schawlow was then at Bell Labs, I of course talked with him. On telling him of my ideas for an 'optical maser', using a resonant cavity and excitation of atoms with optical radiation, he said he too was interested in that, and we decided to work together to optimize a system. It was Art who then suggested use of a Fabry–Perot resonator rather than the cavity with large holes I had used for a model, and that was an excellent addition. Why I did not think of that is a mystery, but Art had done his thesis at the University of Toronto in Fabry–Perot spectroscopy, and that might have been why the thought came to him.

Since Art Schawlow was participating, I decided the patent for the new 'optical maser' should belong to the Bell Labs (I already claimed ownership of the basic maser patent, which covered all wavelengths). Hence we recognized that the new idea must be kept confidential until Bell Labs lawyers had worked out and applied for an appropriate patent. This delay in public information sheds some additional light on how the scientific and technical world was thinking at the time. I had written down my original ideas for an 'optical maser' in my notebook and had it witnessed by my student Joe Giordmaine at Columbia University. I had also talked with a Columbia student Gordon Gould because he had been doing research with an intense light source and I wanted to know how much intensity he had in order to be sure I could get enough excited atoms and provide an oscillator at these short wavelengths. Except for these two persons, neither Art Schawlow nor I told anyone outside of Bell Labs about the 'optical maser', or laser idea until after Bell Labs had properly prepared its patent, which was about 11 months after my first notebook entry. For that entire time, no one has produced a record of any thoughts about extending the maser to optical wavelengths except Gordon Gould, who entered some ideas in his notebook about 1 month after I talked with him about the possibility of an optical maser, and 2 months after my original notebook entry. His notes were later to be the source of a long patent case.

The striking observation is that no one outside of Bell Labs except Gordon Gould, to whom I had explained my ideas, seems to have written or noted down anything about extending masers to these shorter wavelengths during the 11 months from my recognition that it could well be done until after the Schawlow and Townes paper on how to do it became available. After that there was considerable excitement and a number of ideas.

It is also noteworthy that, because of low general interest and competition, I did not publish a paper on how a maser might be made, but waited for publication until we demonstrated its operation, about 3 years after the idea had arrived. But after the maser worked the field became exciting and competitive. Hence, Schawlow and I thought we surely should publish a theoretical paper establishing the idea before taking the time to make a laser work. And indeed, there was much competition to build the first laser.

I myself helped a couple of my graduate students start work on trying to build an 'optical maser' or laser. But at about that time (1958), I was urged to undertake a job in Washington as Vice President and Director of Research of the Institute for Defense Analysis, an organization put together largely by the presidents of several universities to try to help advise the government on matters of science and technology. Sputnik had gone up the year before, and everyone was worried about the position of the U.S. with respect to Communist Russia, which seemed to be ahead particularly in some areas important to defense. I decided I should try to help, and accepted a two-year appointment in Washington. I recognized that this seriously distracted my attention from developing a laser quickly, but knew there were many others working towards lasers and so such devices would certainly be developed and our doing it just at Columbia University was neither critically important nor highly probable.

The field of masers and 'optical masers' or lasers was becoming so exciting and a bit scrambled that The Office of Naval Research asked me if I would organize a meeting on the subject. I did, with the help of a committee of many distinguished people in the field or closely related science and technology. And it was in a meeting of the Committee that we christened the field with the name 'Quantum Electronics'. This first international meeting on the subject was at Shawanga Lodge in New York State in September 1959. It made an occasion for the Russians Basov and Prokhorov to visit the United States (and my lab and home), and was about 6 months before Ted Maiman made the first working laser. There were many interesting discussions of masers and their coming operation at optical wavelengths.

It is significant to note that, while the maser grew out of basic research in universities (with Russian work at the Russian Academy) and industry had little to do with early masers, all the first lasers were made in industrial laboratories. This illustrates the sociology and the strengths and weaknesses of industrial and of academic laboratories. While masers (and from them lasers) originated from research on microwave spectroscopy of molecules, industrial laboratories believed the field of microwave spectroscopy had little to offer in the way of commercial results. Because of equipment available and the interest of physicists in industry, the field was initiated and pursued immediately after World War II in commercial laboratories – by myself at Bell Labs, my friends at the RCA Labs, at Westinghouse, and at General Electric. For lack of interest in industry, such work was soon shut down, except at Bell Labs, and it moved to universities. Bell Labs generously allowed me to continue such work, although they wanted me to do some 'more useful' engineering. Once the field had obvious commercial possibilities, commercial laboratories began to support it well and the clear importance of masers and lasers made industry very interested. Really interested industry can more easily put strong new resources and push harder on a field than can academic laboratories, where money has to be granted and professors have a variety of other assignments. The first laser was made to work, of course, by Ted Maiman at Hughes Research Laboratories. Ted had been a student of Willis Lamb at Stanford, and there worked on radio spectroscopy. The second type of laser, rather similar to Maiman's but using a different material, was made to work at the General Electric Laboratories by Peter Sorokin and Mirek Stevenson. Sorokin had been a student of Bloembergen at Harvard in microwave or radio spectroscopy and Stevenson was one of my students in microwave spectroscopy at Columbia University. The next type of laser, and one I particularly admire, was made by Ali Javan, William Bennett, and Don Herriott. Javan had been a student with me in microwave spectroscopy at Columbia University, Bill Bennett a student in the molecular beam group at Columbia working on radio spectroscopy, and Don Harriott was an optics specialist. All of these originators, with the exception of Herriott, had been working at universities in the field which originated the idea and were recently hired by industry. The next important laser, involving semiconductors, was invented by Robert Hall and collaborators at General Electric Labs. Note that every one of these early lasers was created in industry.

After the first laser was operated, my own students at Columbia quickly turned from trying to make a laser to using lasers to explore more physics, the normal university function. And I am delighted that masers and lasers have provided such excellent tools for research, as well as for commercial and medical applications.

After a few years of exploring new physics with lasers, I decided that since there were many excellent scientists doing such work, I should move into fields which it seemed to me were being relatively neglected. I moved to the University of California at Berkeley to look for molecules in interstellar space by microwave spectroscopy, and to do infrared astronomy. Very soon after initiating work at Berkeley, one of my students, Albert Cheung, not only discovered the first polyatomic molecules in space, he found powerful water masers. A while before our discovery of water and identification of its radiation as due to maser action, it had been deduced that some microwave radiation of OH must be due to maser action. And since then, many, many masers due to a wide variety of molecules have been found in astronomical sources as well as a few powerful lasers. Since deviations from thermal equilibrium are common in the very thin gases excited by powerful sources in space, we should have expected this, but didn't. And these masers in space could have been easily detected with radio technology available back in the 1930s if anyone had searched the microwave spectrum carefully.

Clearly, masers and lasers could have been discovered and used much sooner than they actually were. We simply were neither thinking nor looking in the right directions. And this raises the natural question – what important and more-or-less obvious ideas are we missing now because of our lack of imaginative exploration?

Further Reading

Basov N and Prokhorov A (1954) Application of molecular beams to radio spectroscopic studies of rotation spectra of molecules. *J. Experimental. Theoretical Physics U.S.S.R.* 27: 431–438.

Bloembergen N (1956) Proposal for a new type of solid state maser. *Physical Review* 104: 324–327.

Cheung A, Rank D, Townes C, Thornton D and Welch J (1969) Detection of water in interstellar regions by its microwave radiation. *Nature* 121: 626–628.

Combrisson J, Honig A and Townes C (1956) Utilization de la resonance de spins electroniques pour realiser un oscillateur ou un amplificateur en hyperfrequencies. *Comptes Rendues* 242: 2451–2453.

Gordon J, Zeiger H and Townes C (1954) Molecular microwave oscillator and new hyperfine spectrum of NH_3. *The Physical Review* 95: 282–284.

Hall R, Fenner G, Kingsley G, Soltys J and Carlson R (1962) Coherent light emission from GaAs junction. *Physical Review Letters* 9: 366–368.

Javan A, Bennett W and Herriot D (1961) Population inversion and continuous optical maser oscillation in a gas discharge containing a He–Ne mixture. *Physical Review Letters* 6: 106–110.

Knowles S, Mayer C, Cheung A, Rank D and Townes C (1969) Spectra, variability, size, and polarization of H_2O microwave emission sources in the galaxy. *Science* 163: 1055–1057.

Maiman T (1960) Stimulated optical radiation in ruby. *Nature* 187: 493–494.

Perkins F, Gold T and Salpeter E (1966) Maser action in interstellar OH. *Ap. J.* 145: 361–366.

Schawlow A and Townes C (1958) Infrared and optical masers. *Physical Review* 112: 1940–1949.

Shimoda K, Takahashi H and Townes C (1957) Fluctuations in amplification of quanta with application to maser amplifiers. *Journal of the Physical Society of Japan* 12: 686–700.

Sorokin P and Stevenson M (1960) Stimulated infrared emission from trivalent uranium. *Physical Review Letters* 5: 557–559.

Tolman R (1924) Weak quantization. *The Physical Review* 24: 287–295.

Tolman R (1927) Statistical mechanics with applications to physics and chemistry. The Chemical Catalog Co., New York, pp. 169.

Weber J (1953) Amplification of microwave radiation by substances not in thermal equilibrium. *Trans. Inst. Radio Engineers Professional Group on Electron Devices* 3: 1–4.

LET THERE BE COHERENT LIGHT: THE DISCOVERY OF THE LASER

T H Maiman, Vancouver, BC, Canada

© 2005, T H Maiman. Published by Elsevier Ltd. All rights reserved.

Editor's Note

The discovery of the first laser was a remarkable culmination of human creativity and ingenuity. However the achievement was finally made by one individual young scientist, Theodore Maiman, working independently of the mainstream and without the benefit of strong government or private financial support. With persuasive arguments, he convinced his management at the Hughes Research Laboratories to allow him to continue his work, albeit with limited resources. And then, working alone with his assistant he saw the light that now illuminates the world of laser science and engineering. This is the story of how thinking outside the box created a new field of science, whose developing technology has advanced the state of medicine and engineering and the quality of life in general.

Introduction

I am frequently asked how it was that I discovered the laser. Many assume that the concept evolved from some sudden, inspirational thought. It didn't happen that way.

It is dramatic and exciting to have a scientific discovery emerge from a dream or a vision that comes out of nowhere, but it seldom does. In reality, almost all scientific discoveries come from building on other, prior, scientific developments.

So it was with the laser. In 1916, Albert Einstein laid the foundation and conceived the basic underlying principles on which lasers are based. He formulated and explained the relations that govern the way that atoms and molecules absorb and emit radiation. He introduced the key concept of stimulated emission. Then, in the 1920s, physicists C.H. Füchtbauer and Rudolph Ladenburg added formulations that used the Einstein theory to go further, and tie the absorption of light in a material medium to the fundamental properties of its constituent atoms. But it was the Russian physicist A.V. Fabricant who first had the vision to propose the concept of a laser in 1940.

Unlike the Sun and other incoherent light sources, the laser produces light of one precise color. Its waves are emitted in phase with each other, and its radiation can be efficiently focused to a precise spot. In short, the laser produces coherent light, with the properties we were already used to with radio, television and microwave sources.

Fabricant specified the conditions needed for amplification of light via stimulated emission. He appreciated the concept of an inverted population and the concept of coupling such an inverted population medium to a resonant structure (resonator). He proposed using a gaseous electrical discharge to achieve laser action as one possibility. Later, he proposed the use of a helium discharge lamp to optically pump the gaseous form of cesium.

Scientists Purcell and Pound were the first to report evidence of net stimulated emission in 1951. Their observation, in the radio part of the spectrum, was a by-product of their pioneering developments in the techniques of nuclear induction, the foundation for magnetic resonance imaging.

The first achievement of a working device that utilized stimulated emission as its operating principle was a microwave oscillator/amplifier. That device used a beam of ammonia molecules as the working medium. The design details were worked out independently by groups at the Lebedev Institute in the Soviet Union and Columbia University in the USA. The device was dubbed a *maser*, which is an acronym for *microwave amplification by stimulated emission of radiation*.

In 1955, the year after I was awarded my Ph.D., I joined the Hughes Research Laboratory in Culver City, California. One of my major assignments was to head a project developing a miniature, liquid-nitrogen ruby maser for the US Army Signal Corps.

Race to the Light

Before starting work on the maser I had begun to germinate some ideas about the possibility of a laser. My concept so far was to try to use a solid material for the lasing medium, and to fabricate the potential laser material into a rod shape. In microwave thinking this configuration would be called a *dielectric wave-guide*. In optics it is known as a *light pipe*.

Abridged from "The Laser Odyssey" by Theodore H. Maiman. Copyright © 2000 Theodore H. Maiman. Reproduced with the permission of the author and Laser Press, Blaine, WA, USA.

I planned to put mirrors on each end of the rod to form a resonator.

Despite the advances in physics it represents, the maser does not in any way extend the coherent electromagnetic spectrum. Furthermore, its use as an amplifier turned out to be impractical because cryogenic temperatures were required for proper operation. There was strong motivation to push the frontiers of coherent electromagnetic radiation to higher frequencies, principally the promise of higher energies and greater concentrations. The reduction in wavelength from microwave to visible is a factor of about 10^4. This would lead to a consequent 10^4-fold reduction in the minimum achievable focal size, and a 10^8-fold increase in energy concentration delivered by the laser.

Work on the maser project therefore proved something of a distraction, and it was not until August 1959 that I was able to devote full effort to the consideration and analysis of my concepts for a laser.

Why would I be willing to enter such a race? There was rather formidable global competition already in play, well-funded and very competently manned. The answer lies in my knowledge of the proposals that had been floated about. Generally, the scientists who pursued these notions were not particularly secretive about what they were working on. They published and gave talks at conferences. On the whole, however, I found that the authors and conference presenters were only offering vague proposals. To be sure, this served as an exchange of information and stimulated ideas. But it didn't appear to me that anyone was close to the answer. In short, no viable laser concept was yet in existence.

As I look back, I was a little brash. I would be thrusting myself, in a sense, into a technological Olympics. The competition was of the best quality and of international scope. But my competitive spirit won out. The challenge of working in the top league of such an exciting project, that had so many questions and problems to resolve, was very compelling to me.

We know *now* that many kinds of lasers can be made. But back then, in 1959, we didn't know. We didn't even know with any confidence if it was really possible to make a laser at all. My strategy was to limit myself to potential solutions to the making of a laser that did not have appreciable distractions in the design. That way I could focus strictly on just the laser problem itself.

For various reasons I was reluctant to work on alkali-vapor or gas discharge systems. I chose to work with solid-state crystals. The main appeal that solids held for me was simplicity. By that, I mean simple in analysis and understanding, and simple in device conception.

In contrast to a gas discharge, the energy level diagram in an appropriate crystal is very limited. There are relatively few possibilities for the energy states, and by and large, the pertinent parameters for a potential laser candidate are amenable to a combination of calculations and relatively straightforward direct measurements.

Another advantage, in principle, to a solid crystal is its relatively high gain coefficient. By that, I mean the amplification in a given length of material is of reasonable proportions. This meant that the laser medium could be relatively small in size and short in extent, and I would not have the problem of developing or depending on the use of special mirrors. Indeed, my first laser used a crystal that was only 2 cm long.

I was also intrigued with the concept of a solid medium since I would not have to deal with vacuum pumps, impurity problems and gas handling apparatus, or complex mirror mechanisms. I could put simple silver mirror coatings directly on the crystal as I had done with my small ruby maser. In principle, a solid crystal laser could be designed to be very simple, compact and rugged.

My first choice to study and contemplate was that of a ruby crystal. Ruby is the result when a water-clear cast of aluminum oxide is doped with a small amount of chromium oxide impurity. It is the chromium that is responsible for the red color. The rubies used for devices are usually not natural gemstones, which have a chromium impurity level of about 0.5%. The chromium concentration in device rubies is around ten times less than this, and hence they display a lighter red color than gemstone ruby and are referred to as *pink ruby*.

Why choose ruby as a potential laser candidate? Most importantly, I was quite familiar with and fascinated by the interesting optical properties of the crystal. Ruby is a fluorescent mineral; if ultraviolet light is shone on a ruby, it will glow with deep red fluorescence. Furthermore, ruby also gives off a red glow when either blue or green light is shone upon it. It is these blue and green absorption bands that give ruby its red color. When a green photon impinges on and is absorbed by the ruby, a chromium impurity ion is raised from its ground state into a broad, excited band. Although the chromium ion has the possibility of radiating by spontaneous emission from that excited level, another process comes into play. The competing process uses the thermal vibrations of the crystal lattice to interact with the excited ion and deposit most of the excitation energy to another slightly lower excited level of the ruby chromium where it stays for a while. This latter process is much more probable and dominates.

The energy from this third level is radiated as spontaneous incoherent emission. This spontaneous emission is made up of red photons and is the observed fluorescence. The level where the fluorescence emanates is sometimes called a metastable level, since the chromium ions linger in that energy state for a comparatively long time before they radiate red photons.

I developed a model that could be mathematically analyzed and I set up kinetic equations to take into account the various mechanisms taking place in the fluorescent process. I also set up simple intuitive criteria for establishing the condition for laser action. This model and these equations have subsequently become a standard way for others to analyze crystal lasers.

I was able to determine which material parameters were important and relevant to a laser by getting the solution to the equations describing the model. Using known and estimated values for the pertinent parameters in ruby, I found that ruby would require a very bright pump lamp to excite the crystal sufficiently to allow laser action to take place. The brightness of a lamp is an important parameter for lasers. It is a measure not of the total power radiated by the lamp, but rather the power per unit area.

Obstacles and Solutions

Ruby has many desirable qualities. It is a very stable, very rugged crystal to work with. It has some broad pumping bands that make it helpful to get a reasonable energy transfer from an incoherent pumping lamp, and compared to some other crystals I considered (e.g., gadolinium-doped crystals), a working ruby laser would produce visible light.

However, there was some uncertainty over the crystal's measured quantum efficiency. The quantum efficiency for a fluorescent crystal is the number of fluorescent photons emitted, divided by the number of pumping photons. In the case of ruby, this is the number of red photons radiated out compared to the number of green photons absorbed. A report in November 1959 had suggested that ruby's quantum efficiency was only 1%. This would effectively mean that it would be impossible for ruby to achieve laser action. However, I was able to make my own measurements of ruby's fluorescent efficiency, and these showed that it was nearer 75%.

I now had enough information that I could proceed to work out an actual laser design. But how would I pump the ruby? What would be the ruby's shape and size?

I knew that I needed a very bright lamp. Among the brightest around was a high-pressure mercury arc lamp manufactured by General Electric. In addition to being one of the brightest available laboratory lamps, it had the advantage of radiating most of its energy in the green and blue-violet parts of the spectrum. This was a good match to what I needed for the ruby.

I drew up a paper design with the mercury arc lamp at the focus of an elliptic cylinder. A small ruby rod was to be mounted on the other focus of the elliptic cylinder. It is a property of an elliptically shaped reflector that a point of light placed at one focus of the ellipse will be imaged at the other focus. The elliptic cylinder that I had in mind would be highly polished on its inner surface to a mirror finish.

But I had a problem. My paper analysis of the design showed that although it should work, it would only do so by a slim margin. I studied the design in more detail, looking for ways to optimize and improve it, but I couldn't convince myself that it was anything but marginal.

I stewed over the problem at hand. I started to think about other ways to look at the problem. I was frustrated: I felt that, on the one hand I was very close to an answer, but on the other hand it was eluding me.

I went back to my analytical model. I pondered my options and decided to put the pumping lamp requirement for the ruby in a different form. I calculated the equivalent black-body temperature of a suitable pumping lamp. The temperature I calculated to have enough brightness capable of driving a ruby into laser action was close to 5000 K. By contrast an ordinary tungsten light bulb has a brightness temperature of about 2800 K.

Once I put the pumping requirement in terms of brightness temperature, I began to think in a different way. I remembered reading an article about photographic strobe lamps, a camera's flash mechanism. The article said that strobe lamps could reach brightness temperatures of 8000 K or more.

I now had my 'aha'!

Most scientists had been thinking in terms of a continuous laser, and that was certainly my thinking to start with. But why should I place such restrictions on myself? At this juncture I was only trying to demonstrate that coherent light could be made at all. Besides, a pulsed source is not only acceptable in many applications, it may even be preferable.

I went back to my analytical model, modified the equations to account for a pulsed light source, and then analyzed the results to guide me in the actual laser design. I had already determined that the most important parameter of the light source would be its brightness, that is, a lamp that maximized the power per unit area into the useful pump bands of the ruby.

My calculations also instructed me that this lamp brightness requirement is largely independent of the ruby crystal dimensions and its chromium concentration over a reasonably wide range. The model assumption was that of a three-level system (appropriate to ruby) where the lower laser-energy level is the ground state.

Next, I scoured every flashlamp catalogue that I could find. I calculated the luminous intensity for promising candidates and found that the highest intensity came from three General Electric spiral-shaped flashlamps. My calculations showed that I would have an adequate margin of safety, in terms of excess brightness intensity for the laser, of some two to three times, even with the smallest lamp.

Let there be 'Coherent' Light

Based on my most current calculations and measurements, I was becoming optimistic about the possibility of creating a laser. The tension was building and I started to have dreams that I actually did it. It was a wildly exciting thought!

Still, I was worried that I might be missing something. Instead of proceeding directly to the actual design, I decided to do an experiment that would check the validity of my model.

I made up a cube of pink ruby with each dimension equal to one centimeter in length, about the size and shape of a sugar cube. The crystal axis was perpendicular to one set of faces of the cube. I then placed the cube between two parallel plates, thus forming a microwave cavity. The cavity resonance was designed to be equal to the natural ground-state splitting of the ruby. These are the ground-state sublevels used in a ruby maser.

The purpose of this arrangement and the subsequent experiment was to monitor the ground state population of the ruby. My calculations had shown that it would be possible to substantially reduce the ruby ground state ion population necessary to make a laser, but I wanted tangible evidence to confirm and justify my optimism.

I was still harassed by the thought that no one had yet ever made a laser, although by now several teams of scientists had been diligently working away for nearly two years. Was there a fundamental insurmountable problem that I had missed?

I proceeded with the experiment using the ruby cube. As described above, one set of parallel faces of the cube served as the resonant microwave cavity. I connected one of the open (second set of) faces of the cube to one end of a polished quartz rod called a light pipe. The other end of the light pipe was connected to the flashlamp. The third set of faces of the cube was used to probe with selected wavelengths of light. I used the 'loading' on the parallel-plate microwave cavity to monitor the ground state population of the ruby, making use of the microwave properties of ruby that I had learned from my work on the ruby maser.

When I flashed the lamp, the cavity loading did indeed change. It decreased because, as expected, there were fewer chromium ions in the ground state. More important, the *magnitude* of the population change, 3%, was very close to the value I had predicted from my analysis of the experimental parameters. This was an extremely gratifying result since it gave further confirmation to my model and its analysis.

Now I was getting more excited. Since I couldn't think of anything else to check, it was time to proceed with the definitive design of a laser.

The Laser Design

My first thought was to use some modification of the elliptic-cylinder configuration that I considered for the mercury arc lamp as discussed in the previous section. The problem was, I would need a straight, rod-shaped flashlamp, but the straight lamps in the catalogues that I checked didn't have high enough intensity. Therefore, I resorted to a different design and stuck with what was available – the spiral-shaped lamps. I didn't want to get sidetracked at this time into the development of a special lamp.

It dawned on me that I would not have to devise a focusing arrangement, since the brightness of the radiation that is attainable at the focus of a mirror or lens can only approach, but not exceed, the brightness of the source. So, why not just place the crystal in close proximity to the lamp? That is, why not place the crystal *inside* the lamp helix?

To help gather the light, I placed a polished aluminum cylinder around the outside of the spiral. To fit the lamp, I used a pink ruby crystal in the shape of a right circular cylinder rod, about 1 × 2 cm. The ends of the ruby cylinder were polished flat, parallel to each other and perpendicular to the axis of the cylinder. For mirrors, I used evaporated silver.

I chose silver for the mirrors because it has the highest reflectivity of any metal at the deep red ruby laser wavelength. However, silver tarnishes quickly, and so with time, the thin output layer will change its transmission characteristics; it is not stable. To solve this problem, I had a thick layer of silver evaporated on both ends of the crystal and I scraped a tiny hole in the coating at one end. The laser beam would emanate from that coupling hole.

The Crucial Test

As I neared completion of my laser design, Bob Hellwarth, one of my colleagues in the Atomic Physics Department, asked, 'How will you know if it's working?'

At first, I was concerned that if I were stuck with a ruby crystal that deviated too much from optical perfection, these imperfections could prevent measurable laser behavior.

To get a better understanding of the processes, I elaborated on my existing analysis to account for an imperfect crystal. I found that if I were able to drive the ruby crystal reasonably far above the point of inverted population, or laser threshold, that I would see very substantial evidence of an inverted population and impressive laser behavior even with a poor crystal.

Just exactly what would I see?

I planned to monitor the red light emitted from the ruby through the hole in the output silver coating and expected to observe three kinds of laser behavior evidence: a reduction in the decay time of the fluorescent level; the bunching of the radiation into a beam; and a significant reduction of the spectral width of the red light.

The Laser Process

Laser action can be understood by following through the operational details in the above-mentioned ruby laser design. The process starts when the ruby crystal is excited by the flashlamp and chromium ions are excited into the metastable fluorescent level. The ions lose energy by radiating randomly red photons. This is the familiar red fluorescence (spontaneous radiation).

When the excitation level is great enough, that revered 'inverted population' condition is reached. In that case, more chromium ions are in the ruby's metastable upper fluorescent level than are in the ground state. Therefore, chromium ions can radiate by *stimulated* emission as well as by the normal spontaneous emission process.

Because the metastable level now empties much faster than it normally would, the fluorescent lifetime is reduced. That decreased lifetime can be observed by monitoring the red ruby glow with a photoelectric cell connected to an oscilloscope when the ruby is flashlamp-excited.

When the ruby is in the inverted population condition, as explained previously, it becomes an amplifier. The red photons are amplified as they progress through the crystal.

An important selection process starts to take place. The red photons are initially emitted in random directions. But the fluorescent photons that happen to be radiated at large angles to the mechanical axis of the ruby cylinder are lost through the sides of the crystal. On the other hand, photons radiated along the crystal axis, or at small angles with respect to the crystal axis, are in effect, trapped. They are reflected when they strike either one of the end mirrors and move back and forth through the crystal. As these axial photons move through the ruby they are amplified and consequently generate more photons in the same direction. They quickly become the dominant stimulated radiation from the excited metastable level.

Keep in mind that photons moving through the crystal are responsible for the stimulated emission. The axial photons pick up a following and march down that crystal axis. Consequently, the photons that emerge from the coupling hole in the output silver mirror are concentrated in a direction along, or nearly along, the crystal axis.

The red ruby fluorescent (spontaneous) emission extends over a distribution of frequencies in a curve that resembles a bell shape. The fluorescent photons are most concentrated at the center of that curve. When the inverted population condition exists, the top of the curve has the highest amplification. The consequence is (as with the discussion of beam angle) that photons near the center of the distribution are favored, since it is these photons that get amplified most in the back and forth transits, through the crystal from the multiple mirror reflections. It is this last process that explains why the frequency distribution, the linewidth of a laser, is so small.

I planned to vary the excitation to the flashlamp. In doing so, as explained above, I would expect to see a reduction of the fluorescent lifetime giving evidence of stimulated emission. As I varied the ruby excitation level, the fluorescent intensity should increase proportionately, as long as I was below threshold. But, when above threshold, small increases in excitation should make much bigger changes in the output since the detector and monochromator combination are more responsive to the smaller beam and narrower linewidth.

Do it!

It was the afternoon of May 16, 1960; it was time to confirm or deny all the fears of why 'ruby can't work' or why 'lasers can't be made to work.' No more new calculations, no more diversionary experiments. This was the moment of truth!

The laser head was mounted on a workbench. The flashlamp was connected to the power supply. The trigger electrode was connected to the spark coil (the mechanism that initiates the flash from the strobe lamp). The light output from the coupling hole in the end of the ruby was directed through the monochromator to a photomultiplier tube. The electrical signal from the photomultiplier was connected to an oscilloscope.

Irnee D'Haenens, my laboratory assistant, and I were the only ones performing and observing the experiment. We first took a test shot so that we could adjust the monitoring equipment. We turned the power supply up to about 500 volts. We fired the flashtube. Indeed, we observed a trace on the oscilloscope!

The trace was a recording of the red ruby fluorescence. The decay in the trace was about three milliseconds, the lifetime of the upper possible laser level. We made the appropriate adjustments to optimize the monitor display.

We continued. We progressively increased the supply voltage, each time monitoring and recording the light output trace. As we did so, the peak output increased proportionately to the energy input, the decay time remained the same. So far, so good.

But, when we got past 950 volts on the power supply, everything changed! The output trace started to shoot up in peak intensity and the initial decay time rapidly decreased.

Voilà. This was it! The laser was born!

The Light Fantastic

When Irnee and I observed the first laser go into action, Irnee was smiling and jumping up and down with glee. I was numb and emotionally drained from all the tension and excitement.

Imagine: in the 10 years prior to the laser, the coherent electromagnetic spectrum had been extended by perhaps a factor of five. Now, with the advent of the laser, there was a quantum jump in that spectrum of ten thousand!

The significance of my historic accomplishment didn't sink in right away. I'm not sure that it has yet.

THE HELIUM–NEON LASER

A Javan, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005, A Javan. Published by Elsevier Ltd. All rights reserved.

> The date, twelve of December nineteen sixty, and the time, 4:20 PM, are the date and the time of an event that triggered such an imortant segment of what has now evolved in modern optics. Here is how:

It had snowed heavily on that day. The laboratory at Murray Hill, New Jersey, the Bell Telephone Laboratory, BTL, had been closed at 4. What occurred was a triumph of our depth of understanding of the atomic spectra, the physics of the inter-atom collisions and electron impacts and exchange effects, the fascinating sequence of what takes place in a *gas discharge plasma* when the discharge is switched on. The triumph was also that of the art in spectroscopy, this time that of the colorful glow of a gas discharge plasma in a controlled mixture of helium and neon. Performing the spectroscopy, one could *effectively* watch and follow a set of fascinating physics effects in the discharge plasma, that would lead to an *inversion* to set-in in a specific transition between a high-lying pair of energy levels in neon, by the now well-known transition at 1.15 μm wavelength in neon. Helium atoms in a metastable state in the discharge plasma played the key role.

On that day: The laser had already been de-gassed for 5 days and nights at 600 °C under the high vacuum, to drive off the impurities prior to introducing the gas mixture in the laser-plasma quartz tubing and sealing it off. We had a way to optimize the mixture, at the exact mixture ratio for the optimum inversion and the laser gain at the 1.15 μm neon transition wavelength. The laser design had a *daring* interferometer length *one meter long*. Why *daring* at *one meter long*, I will have a chance to explain why below. The laser-mirrors at each end of the one-meter-long interferometer were situated internally in the laser, exposed to the gas mixture. Two fine micrometers at each end of the laser provided the fine control for the laser–mirror alignment control externally.

To convey the scene, we had known already, that the inversion and the laser gain at the 1.15 μm neon transition will be *higher* in the discharge '*afterglow*', the afterglow following *rapidly switching-off the discharge*. The inversion and the laser gain at the 1.15 μm wavelength, we knew, would last for as long as a millisecond in the afterglow. Certain that the inversion and the gain will be higher in the afterglow, we had set-up the electronics to switch on the discharge for about 10 milliseconds, and switching it off fast at a one microsecond switch-off time, and leaving it off for the same 10 milliseconds time interval, and repeating it at a low rep-rate, about 20 times a second on that day. Expecting that while searching for the laser signal, the laser will first break into oscillation in the afterglow where the laser-gain will be higher, knowing that pulling it into the glow afterward will be easy.

With me on that day I had two of the three good people who worked with me on the experiment. One of the two, Donald Herriott, was about to miss his car pool. He had his hand at the time on the laser–mirror tuning-alignment, the fine micrometer at one end of the laser interferometer on his side. I had my back towards the oscilloscope that would display the laser signal, aligning the detection system. My journey had begun over two years previously. I was *certain* on that day, and at that time, that possibly right then, or no later than hours later or by midnight, the moment would come for me and others that I had with me, to witness the onset of the event sought after so hard. And then … .

I had my back towards Don Herriott, urging him to stay a bit longer. I heard Don's voice, saying '*What is that*?'. I turned and looked. The laser signals in the discharge afterglow on the oscilloscope screen, synchronized to the gas discharge rep rate. I reached the laser interferometer fine-tuning micrometer on my end … a gentle tuning, in a mere few seconds time, optimized the afterglow signal, and then pulled the laser signal right into the glow. I loudly called on 'Ed' to switch on to CW. He, Edward Ballick, jumped over the equipment and the electronics on the floor in the lab, reached the switching electronics, switched it on to CW:

> The laser signal on the oscilloscope … switched to CW … a coherent light beam … CW … continuous … and at a degree of coherence … that as it proved later and I know then … to the limits that the laws of nature will permit. I looked at my watch, 4:20 PM.

The year 1960 was the beginning. Some months earlier, in the summer of that year, Theodore Maiman at Hughes Laboratories, in Malibu, CA, succeeded to operate an optically pumped laser, operating in short-duration laser pulses, the Ruby laser at 0.694 μm. That approach to the laser differed fundamentally from the discharge plasma approach. The two were different inventions and served different purposes. One, the *optically pumped*, converts the *optical energy* in the pump, to the laser light output. The other, *electrically pumped,* converts the *electrical energy* in the *electrical current* that excites the laser plasma, to the laser light output. The rest on how things evolved is history.

The Scene Before Lasers

Before lasers, optics had long passed its peaks. The last words had been written by Born and Wolf in 1957, in an impressive volume on Physical Optics. When I went through my graduate studies at Columbia University in the early fifties, there were no graduate courses offered in optics. The subject was dealt with incidentally in courses in electricity and magnetism. Action was all on the microwaves at the time. Optical spectroscopy had reached its peaks much earlier yet, before World War II. Condon and Shortly had written the last words on Theory of Atomic Spectra in 1935, amazingly, as it seems, in *hardly a decade* after the discoveries in quantum mechanics. As events had it, optics gave way to the microwaves after the war, and optical spectroscopy to the RF and microwave spectroscopy – and all this, a result of the discoveries in Radar at the microwave frequencies in World War II.

We owed everything we knew of the atomic spectra to the *gas discharges* used as spectroscopic light sources at the time, from the high-resolution spectra of the hydrogen atoms at one end of the periodic table, to that of the trans-uranium elements on the other. We need not be reminded, everyone knows, how much we owed then and do now, to the high-resolution spectra of the hydrogen atoms, all because of the early hydrogen discharge light sources and then at the beginning, what led to the discovery of the quantum mechanics, and the wave nature of the atoms. The excitement that followed has remained unparalleled in physics since. Within hardly a decade after as a result, the cleverly designed spectroscopic gas-discharge sources, of different makes and designs, exhaustively revealed every feature predicted by the quantum mechanics in the high-resolution spectra of every element we know in nature. Gas discharge light sources set the course at the time, as did the lasers after the events in 1960.

Optics had given way to the microwaves in the fifties after the war. The physics of the gas discharges, however, continued as an active research field after the war, although of short duration. Bell Telephone Laboratory had a most distinguished research department in gas discharges in the early fifties. Towards the late fifties, as it seemed, the gas discharges as a field in research, also seemed to have been exhausted. The payoff at the time had evolved in the colorful neon light sources and fluorescence lamps, a thriving industry all over the country.

In the summer of 1958, Bell Telephone Laboratory had dissolved their research department in gas discharges. They had transferred every one of their distinguished research people to the other research areas in the laboratory. Some good people had left earlier, already.

As the fate in gas discharge physics had it, I joined the Laboratory in September 1958 – with ZERO backgrounds in gas discharge physics. I knew physics.

How Things Evolved

It has been said that when time is ripe for a new idea, if one person misses it, the next one will not. Time was ripe in the late fifties for the optically pumped idea for the laser, but far from it for the *electrically pumped*. Optical pumping was an active field in research at the time. The French had introduced the 'pompage optic', the optical pumping. With the French pompage optic, one would by optical pumping, resonantly excite an atomic species from its ground into its excited electronic state, and perform a double-resonance at an RF frequency on a fine or hyperfine structure of the excited electronic state. There were other versions of the French pompage optic. Each version used the optical pumping to induce resonantly, an atomic species into an excited electronic state.

Important, and more so, in the fifties, in the second half, the discoveries of the three-level masers at the microwave frequencies had attracted much attention. The three-level masers are the counterparts at the microwaves of our now optically pumped lasers. In three-level masers, one drives to saturation a pair of energy levels at microwave energy-spacing in a paramagnetic solid, and allows the inversion and the gain to set-in in a transition to a low-lying third energy level at the microwave frequencies. Nicholas Blumbergen at Harvard had proposed the three-level maser idea in an important publication in 1956. I had my own original three-level maser ideas, and had them published in 1957. In my work I had discovered a fascinating new effect, and for the first time, that in the transition to a third level in a three-level maser, it will be possible to drive the maser into its self-oscillation, *without* requiring the inversion. The effect had led me to the idea of the Raman masers, requiring a '*pumping*' microwave field, followed by a Raman transition across a pair of energy levels at microwave energy-spacing. In my publication of the effect, I had also considered the pumping field at an optical frequency, instead of at the microwave.

The ideas were ‘in the air’ at the time, to extend the three-level maser approach into the optical frequencies. Yet, in the summer of 1958, on an occasion in June of that year, I think on the first Monday in June, *flashed* through my mind a far different approach, all because of the *helium metastables*. Ideas always come to mind in a *flash*, the rest follows.

Summer of 1958: A Bit of History

I think it was on the first Monday in June 1958. I had been on an occasion visiting the Bell Telephone Laboratory (BTL) on that day. They had approached me that I join their research staff at Bell. My work on masers – I had done as a part of my post-doctoral research position in physics at Columbia University in New York City – had in part overlapped the then ongoing work at Bell. They knew me, and I knew Bell.

At the visit meeting with Arthur Schawlow, whom I knew well, he gave me his news on that day, that he and Charles Townes, had submitted to the *Physical Review* a paper for publication proposing an optically pumped counterpart of a three-level maser at the optical frequencies – by-now, the well-known Schawlow–Townes optically pumped laser publication. Specially, I was attracted to Schawlow’s idea on a two-mirror Perot–Fabry-type interferometer as a resonator at the optical frequencies for the laser. The use of it in that publication was Schawlow’s idea.

My thoughts, it may seem paradoxical, right there and then moved far away from an optically pumped laser as a possibility. The helium atoms in a *metastable state* flashed through my mind. I will avoid explaining why or how, other than, of a three-level atom, two of the levels were now to belong to a *different* atomic species, the helium atom, and that there will be no optical pumping … and so much ‘*electrical energy*’ could be converted to and stored in the energetic metastable helium. A new possibility presented itself, with no resemblance to the optical pumping – an energy-exchange effect taking place resonantly, between the highly energetic helium metastables, and – as it evolved – another inert gas, the neon atoms in a helium–neon gas mixture. The resonant nature of the effect embodied everything we know of the wave nature of the atoms. I will have a chance to explain the effect shortly below.

Driving home at the end of the day on that day, a two hour drive from Murray Hill, New Jersey to uptown New York City, the thought of the metastable helium had crystallized before me the task to examine the effect in a helium-plasma in a gas discharge, in a gas mixture with the other *four* well-behaved *inert gases*, from neon through xenon. There are four.

Within weeks in June, searching through the literature I learned, that over the years, and mostly in the early fifties, the helium plasmas in a gas discharge in mixtures with *every other* inert gases had been the subject of most extensive research, *except in a mixture with neon*. The mixture with neon, for the reason that the reader will find below, had been thought of as *uninteresting*. The matter was left for me.

I had accepted in June already, to join the Bell research staff at Murray-Hill, and I did on September. From my work that entire summer, unfolded the entire theory for the helium–neon gas laser. There was also another far-reaching offshoot.

The Physics Effect

Next to the hydrogen atom, the helium is the most fundamental of all atoms in physics. Hydrogen has one electron. Its spectrum, detached from the complexities of the multi-electron atoms, carries the entire signature of the wave nature of the atom, in ways that are subject to accurate predictions and the measurements. The helium atom has two electrons. Its spectrum carries the entire signature of the physics effect of the entire multi-electron atoms, also in ways that are subject to accurate predictions and the measurements. In its metastable states, it carries special *dynamic features*, deeply fundamental, and special *only* to the helium metastables.

In a gas discharge in pure helium at a moderate gas pressure, the discharge color takes on a grayish-white with a tint of blue. Over 500 excited electronic states in helium have been identified – their excitation by the energetic electrons in the discharge, followed by their spontaneous radiative decays, the emissions of photons, accounts for the high-resolution spectra of the discharge grayish-white with a tint of blue in color. Of these 500, all of them, high lying in energy, are short-lived radiatively, with the exception, however, of two amongst them. The two have the lowest excitation energy. Their radiative decays to the ground electronic state are deeply forbidden; they cannot decay radiatively. The two form the well-known helium metastables, identified by their spectroscopic designations: $^3S_{1/2}$, the triplet S one half, and $^1S_0$, the singlet S zero – generally referred to in the literature as, the triplet S and the singlet S metastables. Of the two, the triplet S metastable has the lowest excitation energy. It lies at 19.8 electron volts (eV), above the helium ground state. The singlet S metastable lies higher, at 20.2 eV.

Helium atoms in their ground electronic state are chemically inert. A helium atom's two electrons form a closed shell. It cannot enter in a chemical reaction, neither with its own species, nor with any other, as we all know. The situation is, however, far different with every one of the 500 excited electronic states in a helium gas discharge. They are energetically high-reactive, except, and with the exception of the two metastables, they are all *short-lived*, and cannot play an important role.

The effect manifests itself differently in the long-lived metastables, and accounts for their *special dynamic features,* special only to the helium metastables. What follows outlines the effect.

Dynamic Features: Why so Special

By far the majority of the atoms in the helium gas discharge are in their inert, ground electronic states. Once in its energetic metastable state, a helium metastable in the gas discharge encounters in an impact – in a collision – with a helium atom in its ground electronic state. The encounter in the impact is of a short duration in our scale of time, but long in the atomic time-scale, of about 2×10^{-13} second, a fifth of a picosecond. What occurs is strictly wave-mechanical, dictated by the wave nature of the electrons in the atoms. During the encounter, the electron wave functions in the encountering atoms overlap momentarily. A momentary binding occurs. In no ways can one tell during the encounter which electron is which: Electrons are the properties of their wave functions – once overlapped, individual electrons lose their identity, in no way can one tell which is which. As the atoms fly apart, caused by the kinematics of the impact, an *electron exchange* occurs, *wave-mechanically.* The helium atom initially in its ground electronic state, will fly away from the encounter with one of its two electrons now in the *highly excited* energetic *metastable* state. The other, the one initially in the excited metastable state, emerges from the encounter in the *unexcited* ground electronic state.

At a gas pressure in the discharge of about one torr, a millimeter, the mean free-time between two impacts is about one microsecond, a very long time in the scale of the atoms. The helium atom that flew away from the encounter in the metastable state remains in its now metastable state, during the entire mean-free collision time. Once it encounters the next helium atom in its ground electronic state, the act repeats again, a momentary binding occurs, an electron exchange takes place, and a new metastable flying away. The effect repeats over and over again, at about every microsecond. A *diffusion* of the metastability through the helium in the discharge sets in. Depending on the diameter of the discharge tubing, it takes about a millisecond until the helium that emerges from the encounter in the metastable state reaches the walls of the discharge tubing. The act will then *terminate*, in a *most violent* collision effect with the wall. The energetic metastable helium *knocks out* an electron off the wall, as itself emerges from the collision in its ground electronic state – all intact.

With one exception, the encounter between a helium metastable with any species other than of its own kind, a helium, is energetically *highly violent*. The exception is the neon atoms. In encounters with all atomic species other than neon, the metastable helium *knocks out* an electron off the species, ionizes the species, and itself emerges from the encounter in its ground electronic state, intact. The ionizing impacts had attracted so much attentions in the early fifties, the effect in the helium gas discharge in mixtures specially with every other inert gases, had been exhaustively investigated, except with neon. Neon has its ionization potential higher than the two metastable excitation energies. It can not undergo an ionizing impact with either of the two metastables. The mixture with neon had thus been considered uninteresting, and was discarded.

What emerged by the end of the summer of 1958 was the entire theory for the helium neon laser – and the recognition of a deeply wave-mechanical effect, the resonant energy exchange taking place between the energetic helium metastable and an inert neon atom in the discharge gas mixture. In the mixture, a momentary binding occurs between the helium metastable encountering an inert neon atom in the gas discharge. Wave-function overlaps occur for the electrons in the encountering atoms. A *resonant* energy exchange takes place during the encounter. Helium metastable will fly away from the impact in its now ground electronic state, while neon emerges in one of its high-lying excited energy states, at the near-resonant with the helium metastable.

A special effect sets in in the energy exchange, as well. The kinetics of the impact in the encounter enters in satisfying the exact resonance in the energy exchange. The singlet S metastable flying away from the encounter after the energy exchange, will be slightly 'cooler' kinetically, and the triplet S flying away from the encounter after the energy-exchange, will be slightly 'hotter'. The kinetic energy of the impact enters in the energy exchange.

Optical Gain and Amplification Effect

Within weeks, early in the summer of 1958, investigating the energy transfer effect led to the prediction that there can be inversion and optical gain in two sets of transitions in neon, one as a result of the excitation transfer from the triplet S metastables, and the other a result of the transfer from the singlet S. By far the more challenging task that summer, however, was to estimate theoretically the *values* for the *gain and amplification factors* obtainable for the predicted inverted transitions, and more specifically, for the inverted transitions that will have the highest gain and amplification factors. There followed the theoretical estimates on which relied the entire experimental phase of the work, the planning of it commenced within weeks after my joining the Bell Telephone Laboratory on September of that year.

The set of transitions originating from the triplet S energy transfer to neon, occur in the near-red region of the spectrum, the 1 μm wavelength region. The strongest inverted transition at the highest predicted gain value, lies at 1.15 μm transition wavelength, the spectral laser line that first broke into self-oscillation (see the opening page in this article). The set of transitions originating from the singlet S energy transfer to neon occur in the red-yellow region of the spectrum. The strongest, although at less gain-value than 1.15 μm, occurs at the 0.633 μm in the red.

In the early phase of the investigation, it came to light also that the achievement of the inverted transitions with optical gains, was, interestingly, a *commune occurrence* in gaseous discharges in a variety of the atomic species, including in pure neon. However, the complexities in the excitations by the hot electrons of the high-lying energy levels, and in particular in the presence of the secondary excitation mechanism, and a list of other similar effects, in no way would allow a quantitative estimate of the gain and amplification factors in the inverted transitions obtainable in the gas discharge.

The reliance on the excitations of the high-lying energy levels in neon via the resonant energy transfer from the helium metastables, as per the physics effect I have outlined above, placed at my disposal the one gas discharge system that lent itself to the theoretical estimates of the predicted optical gains and the amplification factor for the inverted transitions.

Without a knowledge of the predicted values for the optical gain and the amplification factor, it would have been in no way possible to perform the experiment.

Experimental Phase

The Bell Telephone Laboratory, considered as a foremost research institution in the country at the time, operated at a $250 million yearly budget – in 1960s' dollars. Transistors had been invented at Bell in the forties. With the awareness at the time at Bell, that a light beam at the degree of frequency-purity and the coherence expected of the laser, will have far-reaching consequences on the telecommunication technology, without hesitation I received from Bell an open ticket to do what it takes – one key person in the administration understood in-depth my approach to the laser – via the gas discharge and the physics effects I had invoked. Within weeks following joining Bell in September, I had my planning underway for what came to be known in a short while later, as a most aggressive research in gas discharges – and at a time when Bell had already, in the summer of 1958, dismantled the entire research activities in gas discharges at Bell. That speaks a great deal for Bell at the time … at the time before the AT&T break up.

In lasers, the knowledge of the full length of the amplifying medium across which the optical gain obtainable at the laser transition will suffice to drive the laser into its self-oscillation, will dictate the full length of the laser interferometer resonator. The spacing between the two high-reflectance interferometer end-mirrors will have to accommodate the full length of the amplifying medium, thus dictating the interferometer full length. With respect to the mirror reflectivity obtainable at the time, one had to be content with the values at the highest close to 98%. The highest gain transition from my theoretical estimates, that of the neon transition at the 1.15 μm wavelength in the near-red, indicated an interferometer length of no less than 50 centimeters, or possibly as long as one meter – and there lay the challenge, achieving the required degree of parallelism needed for the interferometer end-mirrors at such spacing. One could show, from simple inspection, that the parallelism needed for the mirrors had to be at values better than one part in several-hundredths of a wavelength. This value applied to the use of the flat interferometer mirrors known to me at that time.

An alignment to such a degree of accuracy could well be obtainable at the time for Perot–Fabry-type interferometer mirrors at the mirror-spacing not exceeding about eight or ten centimeters at the most – where one would rely on the observation of the interference fringes using the available spectroscopic light sources to observe the fringes, a principle well known in physical optics. At longer mirror spacing, the spectroscopic light sources lacked the degree of coherence, the color purity, needed to observe fringes for the mirror alignment. The fringes will get wiped out.

These days, using a helium–neon laser light source available at every laboratory in optics, one can align a two-mirror interferometer at tens of meters mirror-spacing, or hundreds, or even as far apart as several miles in some experiments. Ideas are floating around now of the possibility, using gas lasers, to align mirrors placed on satellites in orbit, with the mirror on two different satellites at controlled spacing as far apart as thousands of miles – the experiment thought of is that of the detection of the gravity waves. Without a laser operating and available at the time, in no ways would it have been possible to align an interferometer at a *daring* mirror-spacing as far apart as one meter, or even half as long. One had to rely on an auto-collimator, an optical instrument that at the best would provide an alignment to within one or two wavelengths, nowhere sufficient to drive the laser into its self oscillation.

The challenge mapped the entire course for the experimental phase, foremost aimed at showing every feature of the predicted effects experimentally, from the energy transfer effect, to observing and measuring the gain and amplification effect in the predicted inverted transitions in neon. The feat was that of the art in spectroscopy. The observation of gain and amplification effect was a highlight in the course of the experiment. Up to that point, in optical spectroscopy, one always measured the spectral lines in either emission or in absorption. The feat was the first in spectroscopy, observing the spectral line of an atomic species not in absorption, but in amplification – the amplification of an external light source transmitted through the gas discharge in the experiment. The feat, yet, lay in the art in spectroscopy, throughout the experiments.

The event at 4:20 PM in December two years later, speaks for itself.

> It may seem now, as if one needed an operating helium–neon laser at the time, to align the laser mirrors for the first operating helium–neon laser. It seemed evident from the beginning, that the presence in the very first helium–neon laser, of the optical-gain at the predicted value in the predicted inverted transition, will guide achieving the alignment of its own mirrors, to the exact optical degree of the alignment accuracy needed for the self-oscillation. Driving the laser into its self-oscillation signaled the feat. The feat, however, was that of the power in spectroscopy, making it possible to predetermine the presence in the first helium–neon laser, of the optical gain, at the predicted optimum value needed for self oscillation.
>
> There were doubters and doubters, in the technical staff and the administration at Bell. Donald Herriott, William Bennett and Edward Ballick collaborated with me in the experiment.

The Impact

Within months of the publication of the work in January 1961 in the *Physical Review Letters*, walking through the long hallways at BTL where the laboratories were situated, one could see through the open doors in many laboratories down the long halways, the pink colors of long helium–neon gas discharges plasma tubing, with people in the labs searching for new laser transitions. BTL had revived the gas discharge research activities they had dismantled in the summer of 1958. Not surprisingly, the second set of new laser transitions driven into self-oscillation took place at Bell in 1962, that of the set of transitions in neon originating from the energy transfer from the helium singlet S metastable – with the most intense in the series, that of the by now well-known 0.6328 μm neon transition in the red. There followed the discoveries of the argon–ion gas laser and shortly thereafter the CO_2 gas laser, the two amongst the foremost in the many new gas lasers driven to self-oscillation at Bell.

Every gas laser discovered since the December 12 event, has enjoyed the use of a '*no-longer*' a '*daring*' '*long*' interferometer resonator, merely because one would use a He–Ne laser to *pre-align* the laser mirrors to within a small fraction of the wavelength needed, prior to searching for a new laser line in the gas discharge chosen for the search. The already aligned resonator removed the need to *predetermine* the presence of the gain and the amplification effect at the values needed for self-oscillation. Within a span of about a decade, virtually hundreds of different gases were driven to self-oscillation with this method, operating literally at thousands of different wavelengths, ranging from the far infrared, the 100 100*ηn* region, to the near-UV.

The helium–neon laser triggered so many of the key discoveries that followed. Within months, in the spring of 1961, one of the distinguished engineers at BTL, Rudolf Komphner, the discoverer of the backward tube oscillators at the microwave frequencies, began experimenting with the transmission through the improvised thin fiberoptic, of the output beam of a helium–neon laser at the 1.15 μm transition in the near-red, the near-IR, having his eyes on the potential someday in communication on the use of a continuously operating laser light beam at the near-IR frequencies. I recollect the very first ad that appeared on lasers nationally was a BTL ad in the spring of 1961, showing a cross-section of a beam of light, underneath it saying: '*Someday a Beam of Light This Size Will Carry Millions of Telephone Conversations Simultaneously*'. The ad speaks for itself.

In every gas laser, the active laser medium – being a gas – is highly homogeneous. As a result, with a reasonably good optics, one can obtain a nearly perfect Gaussian output beam. For this reason, the helium–neon lasers available commercially since the mid-sixties, have served as the indispensable tools for the alignment of the complex optics in every research laboratory in lasers, or in every manufacturer of the lasers in the country, or the manufacturers of the precise optical components, or the medical uses and so on.

In the applied areas, the helium–neon laser triggered so many of the key experiments in the early days. Philips at Eindhoven can be considered as the first discoverer of a video disc recorder. The model used a helium–neon red laser, at a most impressive fidelity and color resolution, in recording and in reading for the display. I recollect, in an exhibit shown to me at a Philips Lab, one groove on the recorder was that of a horse in a prairie, and the next groove, one '*micrometer away*' on the recorder, was that of a crowd in a park – so impressive at the time, with no cross talks. They had it packaged in a compact unit already. In other areas, all bar-code readings were done using a helium–neon laser, and some still do.

In the scientific area, from high-resolution spectroscopy at unprecedented accuracy, to the accurate clocks operating at optical frequencies, or in the measurements of the absolute frequencies of the accurate laser clocks, or the early experiments in relativity and so on, all are owed to the precision gas lasers. Helium–neon laser has played the key role, in the early days throughout, and continues today.

A note to the reader

In the early publications on laser, the device was generally known and referred to at the time as the 'Optical Maser'. The word laser became an accepted designation for the device in the late 1960s and the early 1970s.

Suggested Readings

Javan A (1961) In: Singer J (ed.) *Advances in Quantum Electronics*, pp. 18. Columbia University Press.

Javan A, Bennett WR, Jr. and Herriott DR (1961) Publications reporting operation of the first helium–neon laser. *Physical Review Letters* 6: 106.

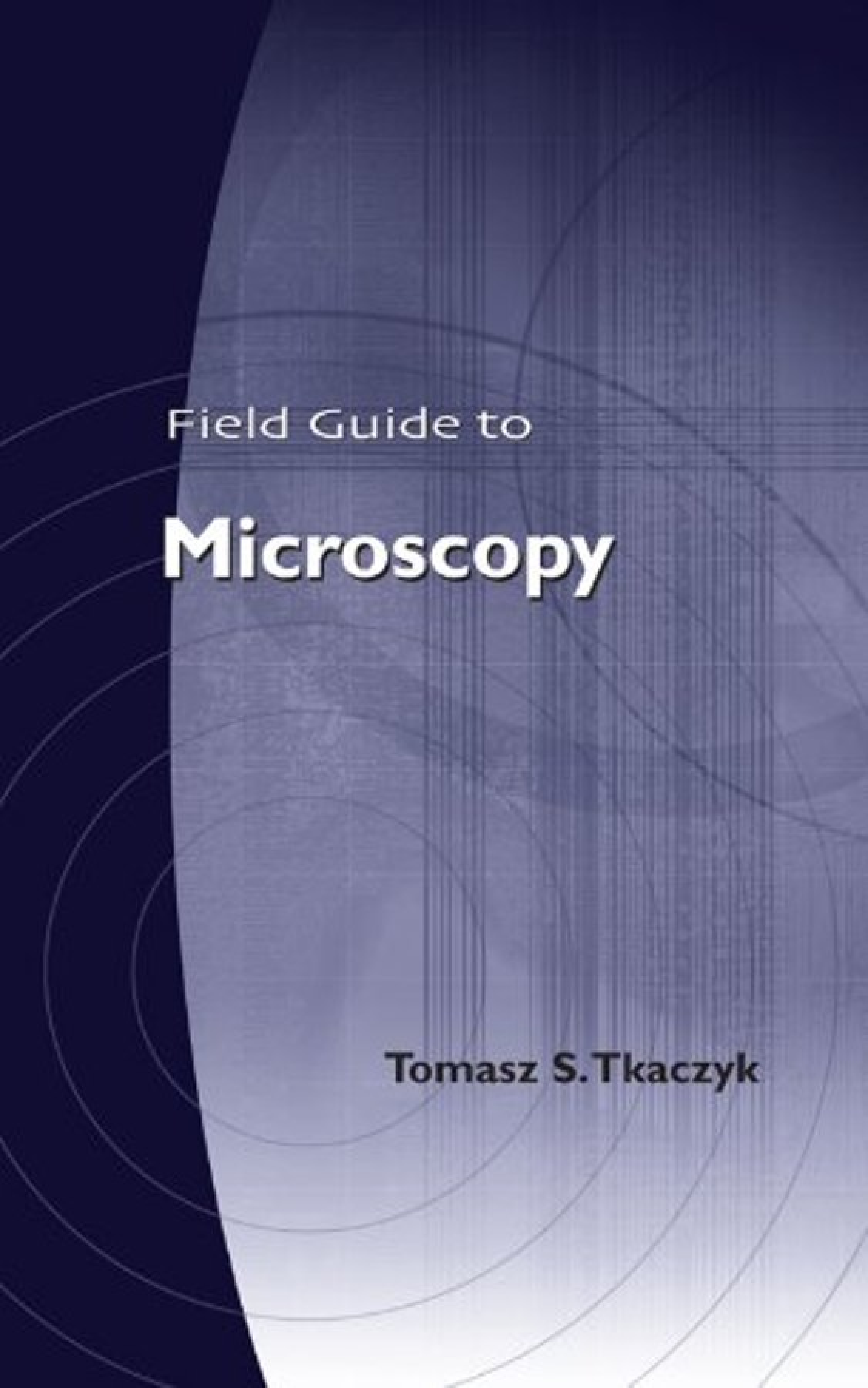

Field Guide to

Microscopy

Tomasz S. Tkaczyk

Field Guide to

Microscopy

Tomasz S. Tkaczyk

SPIE Field Guides
Volume FG13

John E. Greivenkamp, Series Editor

SPIE PRESS
Bellingham, Washington USA

Library of Congress Cataloging-in-Publication Data

Tkaczyk, Tomasz S.
Field guide to microscopy / Tomasz S. Tkaczyk.
p. cm. -- (The field guide series)
Includes bibliographical references and index.
ISBN 978-0-8194-7246-5
1. Microscopy--Handbooks, manuals, etc. I. Title.
QH205.2.T53 2009
502.8'2--dc22
2009049648

Published by

SPIE
P.O. Box 10
Bellingham, Washington 98227-0010 USA
Phone: +1 360.676.3290
Fax: +1 360.647.1445
Email: spie@spie.org
Web: http://spie.org

Copyright © 2010 The Society of Photo-Optical Instrumentation Engineers (SPIE)

All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means without written permission of the publisher.

The content of this book reflects the work and thought of the author. Every effort has been made to publish reliable and accurate information herein, but the publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Printed in the United States of America.

SPIE

Introduction to the Series

Welcome to the *SPIE Field Guides*—a series of publications written directly for the practicing engineer or scientist. Many textbooks and professional reference books cover optical principles and techniques in depth. The aim of the *SPIE Field Guides* is to distill this information, providing readers with a handy desk or briefcase reference that provides basic, essential information about optical principles, techniques, or phenomena, including definitions and descriptions, key equations, illustrations, application examples, design considerations, and additional resources. A significant effort will be made to provide a consistent notation and style between volumes in the series.

Each *SPIE Field Guide* addresses a major field of optical science and technology. The concept of these *Field Guides* is a format-intensive presentation based on figures and equations supplemented by concise explanations. In most cases, this modular approach places a single topic on a page, and provides full coverage of that topic on that page. Highlights, insights, and rules of thumb are displayed in sidebars to the main text. The appendices at the end of each *Field Guide* provide additional information such as related material outside the main scope of the volume, key mathematical relationships, and alternative methods. While complete in their coverage, the concise presentation may not be appropriate for those new to the field.

The *SPIE Field Guides* are intended to be living documents. The modular page-based presentation format allows them to be easily updated and expanded. We are interested in your suggestions for new *Field Guide* topics as well as what material should be added to an individual volume to make these *Field Guides* more useful to you. Please contact us at **fieldguides@SPIE.org.**

John E. Greivenkamp, *Series Editor*
College of Optical Sciences
The University of Arizona

The Field Guide Series

Keep information at your fingertips with all of the titles in the *Field Guide* series:

Field Guide to Geometrical Optics, John E. Greivenkamp (FG01)

Field Guide to Atmospheric Optics, Larry C. Andrews (FG02)

Field Guide to Adaptive Optics, Robert K. Tyson & Benjamin W. Frazier (FG03)

Field Guide to Visual and Ophthalmic Optics, Jim Schwiegerling (FG04)

Field Guide to Polarization, Edward Collett (FG05)

Field Guide to Optical Lithography, Chris A. Mack (FG06)

Field Guide to Optical Thin Films, Ronald R. Willey (FG07)

Field Guide to Spectroscopy, David W. Ball (FG08)

Field Guide to Infrared Systems, Arnold Daniels (FG09)

Field Guide to Interferometric Optical Testing, Eric P. Goodwin & James C. Wyant (FG10)

Field Guide to Illumination, Angelo V. Arecchi; Tahar Messadi; R. John Koshel (FG11)

Field Guide to Lasers, Rüdiger Paschotta (FG12)

Field Guide to Microscopy, Tomasz Tkaczyk (FG13)

Field Guide to Laser Pulse Generation, Rüdiger Paschotta (FG14)

Field Guide to Infrared Systems, Detectors, and FPAs, Second Edition, Arnold Daniels (FG15)

Field Guide to Optical Fiber Technology, Rüdiger Paschotta (FG16)

Preface to the Field Guide to Microscopy

In the 17th century Robert Hooke developed a compound microscope, launching a wonderful journey. The impact of his invention was immediate; in the same century microscopy gave name to "cells" and imaged living bacteria. Since then microscopy has been the witness and subject of numerous scientific discoveries, serving as a constant companion in humans' quest to understand life and the world at the small end of the universe's scale.

Microscopy is one of the most exciting fields in optics, as its variety applies principles of interference, diffraction, and polarization. It persists in pushing the boundaries of imaging limits. For example, life sciences in need of nanometer resolution recently broke the diffraction limit. These new super-resolution techniques helped name microscopy the method of the year by *Nature Methods* in 2008.

Microscopy will critically change over the next few decades. Historically, microscopy was designed for visual imaging; however, enormous recent progress (in detectors, light sources, actuators, etc.) allows the easing of visual constrains, providing new opportunities. I am excited to witness microscopy's path toward both integrated, digital systems and nanoscopy.

This Field Guide has three major aims: (1) to give a brief overview of concepts used in microscopy; (2) to present major microscopy principles and implementations; and (3) to point to some recent microscopy trends. While many presented topics deserve a much broader description, the hope is that this Field Guide will be a useful reference in everyday microscopy work and a starting point for further study.

I would like to express my special thanks to my colleague here at Rice University, Mark Pierce, for his crucial advice throughout the writing process and his tremendous help in acquiring microscopy images.

This Field Guide is dedicated to my family: my wife, Dorota, and my daughters, Antonina and Karolina.

Tomasz Tkaczyk
Rice University

Table of Contents

Glossary of Symbols **xi**

Basics Concepts **1**
Nature of Light 1
The Spectrum of Microscopy 2
Wave Equations 3
Wavefront Propagation 4
Optical Path Length (OPL) 5
Laws of Reflection and Refraction 6
Total Internal Reflection 7
Evanescent Wave in Total Internal Reflection 8
Propagation of Light in Anisotropic Media 9
Polarization of Light and Polarization States 10
Coherence and Monochromatic Light 11
Interference 12
Contrast vs Spatial and Temporal Coherence 13
Contrast of Fringes (Polarization and Amplitude Ratio) 15
Multiple Wave Interference 16
Interferometers 17
Diffraction 18
Diffraction Grating 19
Useful Definitions from Geometrical Optics 21
Image Formation 22
Magnification 23
Stops and Rays in an Optical System 24
Aberrations 25
Chromatic Aberrations 26
Spherical Aberration and Coma 27
Astigmatism, Field Curvature, and Distortion 28
Performance Metrics 29

Microscope Construction **31**
The Compound Microscope 31
The Eye 32
Upright and Inverted Microscopes 33
The Finite Tube Length Microscope 34

Infinity-Corrected Systems 35
Telecentricity of a Microscope 36
Magnification of a Microscope 37
Numerical Aperture 38
Resolution Limit 39
Useful Magnification 40
Depth of Field and Depth of Focus 41
Magnification and Frequency vs Depth of Field 42
Köhler Illumination 43
Alignment of Köhler Illumination 45
Critical Illumination 46
Stereo Microscopes 47
Eyepieces 48
Nomenclature and Marking of Objectives 50
Objective Designs 51
Special Objectives and Features 53
Special Lens Components 55
Cover Glass and Immersion 56
Common Light Sources for Microscopy 58
LED Light Sources 59
Filters 60
Polarizers and Polarization Prisms 61

Specialized Techniques **63**
Amplitude and Phase Objects 63
The Selection of a Microscopy Technique 64
Image Comparison 65
Phase Contrast 66
Visibility in Phase Contrast 69
The Phase Contrast Microscope 70
Characteristic Features of Phase Contrast 71
Amplitude Contrast 72
Oblique Illumination 73
Modulation Contrast 74
Hoffman Contrast 75
Dark Field Microscopy 76
Optical Staining: Rheinberg Illumination 77

Optical Staining: Dispersion Staining 78
Shearing Interferometry: The Basis for DIC 79
DIC Microscope Design 80
Appearance of DIC Images 81
Reflectance DIC 82
Polarization Microscopy 83
Images Obtained with Polarization Microscopes 84
Compensators 85
Confocal Microscopy 86
Scanning Approaches 87
Images from a Confocal Microscope 89
Fluorescence 90
Configuration of a Fluorescence Microscope 91
Images from Fluorescence Microscopy 93
Properties of Fluorophores 94
Single vs Multi-Photon Excitation 95
Light Sources for Scanning Microscopy 96
Practical Considerations in LSM 97
Interference Microscopy 98
Optical Coherence Tomography/Microscopy 99
Optical Profiling Techniques 100
Optical Profilometry: System Design 101
Phase-Shifting Algorithms 102

Resolution Enhancement Techniques 103
Structured Illumination: Axial Sectioning 103
Structured Illumination: Resolution Enhancement 104
TIRF Microscopy 105
Solid Immersion 106
Stimulated Emission Depletion 107
STORM 108
4Pi Microscopy 109
The Limits of Light Microscopy 110

Other Special Techniques 111
Raman and CARS Microscopy 111
SPIM 112

Array Microscopy 113

Digital Microscopy and CCD Detectors 114
Digital Microscopy 114
Principles of CCD Operation 115
CCD Architectures 116
CCD Noise 118
Signal-to-Noise Ratio and the Digitization of CCD 119
CCD Sampling 120

Equation Summary 122

Bibliography 128

Index 133

Glossary of Symbols

$a(x, y)$, $b(x, y)$	Background and fringe amplitude
AO	Vector of light passing through amplitude Object
A_{R}, A_{S}	Amplitudes of reference and sample beams
b	Fringe period
c	Velocity of light
C	Contrast, visibility
C_{ac}	Visibility of amplitude contrast
C_{ph}	Visibility of phase contrast
$C_{\mathrm{ph\text{-}min}}$	Minimum detectable visibility of phase contrast
d	Airy disk dimension
d	Decay distance of evanescent wave
d	Diameter of the diffractive object
d	Grating constant
d	Resolution limit
D	Diopters
D	Number of pixels in the x and y directions (N_x and N_y, respectively)
DA	Vector of light diffracted at the amplitude object
DOF	Depth of focus
DP	Vector of light diffracted at phase object
$\mathrm{D}_{\mathrm{pinhole}}$	Pinhole diameter
d_{xy}, d_z	Spatial and axial resolution of confocal microscope
E	Electric field
E	Energy gap
E_x and E_y	Components of electric field
F	Coefficient of finesse
F	Fluorescent emission
f	Focal length
f_{C}, f_{F}	Focal length for lines C and F
f_{e}	Effective focal length
FOV	Field of view
FWHM	Full width at half maximum
h	Planck's constant
h, h'	Object and image height
H	Magnetic field
I	Intensity of light
i, j	Pixel coordinates
I_{o}	Intensity of incident light

Glossary of Symbols (cont.)

I_{max}, I_{min}	Maximum and minimum intensity in the image
I, I_{t}, I_{r}	Irradiances of incident, transmitted, and reflected light
I_1, I_2, I_3 …	Intensity of successive images
I_0 , $I_{2\pi/3}$, $I_{4\pi/3}$	Intensities for the image point and three consecutive grid positions
k	Number of events
k	Wave number
L	Distance
l_{c}	Coherence length
m	Diffraction order
M	Magnification
M_{min}	Minimum microscope magnification
MP	Magnifying power
MTF	Modulation transfer function
M_{u}	Angular magnification
n	Refractive index of the dielectric media
n	Step number
N	Expected value
N	Intensity decrease coefficient
N	Total number of grating lines
n_{a}	Probability of two-photon excitation
NA	Numerical aperture
n_{e}	Refractive index for propagation velocity of extraordinary wave
n_{m}, n_{r}	Refractive indices of the media surrounding the phase ring and the ring itself
n_{o}	Refractive index for propagation velocity of ordinary wave
n_1	Refractive index of media 1
n_2	Refractive index of media 2
o, e	Ordinary and extraordinary beams
OD	Optical density
OPD	Optical path difference
OPL	Optical path length
OTF	Optical transfer function
OTL	Optical tube length
P	Probability
P_{avg}	Average power
PO	Vector of light passing through phase object

Glossary of Symbols (cont.)

PSF	Point spread function
Q	Fluorophore quantum yield
r	Radius
r	Reflection coefficients
r, m, and o	Relative, media, or vacuum
r_{AS}	Radius of aperture stop
r_{PR}	Radius of phase ring
r_λ	Radius of filter for wavelength λ
s, s'	Shear between wavefront object and image space
S	Pinhole/slit separation
$\mathrm{s}_\perp$	Lateral shift in TIR perpendicular component of electromagnetic vector
$\mathrm{s}_\parallel$	Lateral shift in TIR for parallel component of electromagnetic vector
S_{FD}	Factor depending on specific Fraunhofer approximation
SM	Vector of light passing through surrounding media
SNR	Signal-to-noise ratio
SR	Strehl ratio
t	Lens separation
t	Thickness
t	Time
t	Transmission coefficient
T	Time required to pass the distance between wave oscillations
T	Throughput of a confocal microscope
t_{c}	Coherence time
u, u'	Object, image aperture angle
V_{d}, V_{e}	Abbe number as defined for lines d, F, C or e, F′, C′
V_{m}	Velocity of light in media
w	Width of the slit
x, y	Coordinates
z	Distance along the direction of propagation
z	Fraunhofer diffraction distance
z, z'	Object and image distances
z_{m}, z_{o}	Imaged sample depth, length of reference path

Glossary of Symbols (cont.)

α	Angle between vectors of interfering waves
α	Birefringent prism angle
α	Grating incidence angle in plane perpendicular to grating plane
α_s	Visual stereo resolving power
β	Grating diffraction angle in plane perpendicular to grating plane
γ	Angle of fringe localization plane
γ	Convergence angle of a stereo microscope
γ	Incidence / diffraction angle from the plane perpendicular to grating plane
Γ	Retardation
δ	Birefringence
δ	Excitation cross section of dye
δz	Depth perception
Δ_b	Axial phase delay
Δf	Variation in focal length
$\Delta \mathbf{k}$	Phase mismatch
Δz, $\Delta z'$	Object and image separations
$\Delta\lambda$	Wavelength bandwidth
$\Delta\nu$	Frequency bandwidth
$\Delta\varphi$	Phase delay
$\Delta\varphi$	Phase difference
$\Delta\varphi$	Phase shift
$\Delta\varphi_{min}$	Minimum perceived phase difference
ε	Angle between interfering beams
ε	Dielectric constant, i.e., medium permittivity
η	Quantum efficiency
θ'	Refraction angle
θ_{cr}	Critical angle
θ_i	Incidence angle
θ_r	Reflection angle
λ	Wavelength of light
λ_p	Peak wavelength for the m^{th} interference order
μ	Magnetic permeability
ν	Frequency of light
ν	Repetition frequency
ξ	Propagation direction
ρ	Spatial frequency
σ	Molecular absorption cross section

Glossary of Symbols (cont.)

σ_{dark}	Dark noise
σ_{photon}	Photon noise
σ_{read}	Read noise
τ	Integration time
τ	Length of a pulse
τ	Transmittance
φ	Incident photon flux
φ	Optical power
φ	Phase difference generated by a thin film
φ_o	Initial phase
φ_o	Phase delay through the object
φ_p	Phase delay in a phase plate
φ_{TF}	Phase difference generated by a thin film
ω	Angular frequency
ω	Bandgap frequency
$\perp$ and $\|$	Perpendicular and parallel components of the light vector

Nature of Light

Optics uses two approaches, termed the **particle** and **wave models,** to describe the nature of light. The former arises from atomic structure, where the transitions between energy levels are quantized. Electrons can be excited into higher energy levels via external processes with the release of a discrete quantum of energy (a photon) upon their decay to a lower level.

The wave model complements this corpuscular theory and explains optical effects involving diffraction and interference. The wave and particle models can be related through the frequency of oscillations, with the relationship between quanta of energy and frequency given by

$$E = h\nu \text{ [in eV or J]},$$

where $h = 4.135667 \times 10^{-15}$ [eV·s] $= 6.626068 \times 10^{-34}$ [J·s] is **Planck's constant**, and ν is the **frequency of light** [Hz]. The frequency ν of a wave is the reciprocal of its period T [s]:

$$\nu = \frac{1}{T}.$$

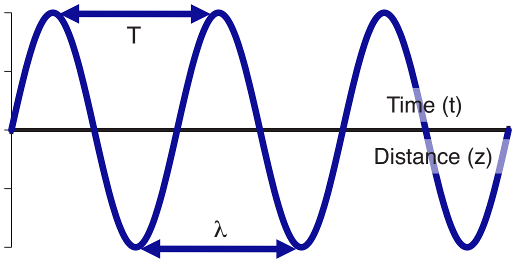

The period T [s] corresponds to one cycle of a wave or, if defined in terms of the distance required to perform one full oscillation, describes the **wavelength of light** λ [m]. The **velocity of light** in free space is 2.99792×10^8 m/s and is defined as the distance traveled in one period (λ) divided by the time taken (T):

$$c = \frac{\lambda}{T}.$$

Note that wavelength is often measured indirectly as time T required to pass the distance between wave oscillations.

The relationship for the velocity of light c can be rewritten in terms of wavelength and frequency as

$$c = \lambda\nu.$$

The Spectrum of Microscopy

The range of electromagnetic radiation is called the **electromagnetic spectrum**. Various microscopy techniques employ wavelengths spanning from **x-ray radiation** and **ultraviolet radiation** (UV) through the **visible spectrum** (VIS) to **infrared radiation** (IR). The wavelength in combination with the microscope parameters determines the resolution limit of the microscope (0.61λ/NA). The smallest feature resolved using light microscopy and determined by diffraction is approximately 200 nm for UV light with a high numerical aperture (for more details, see *Resolution Limit* on page 39). However, recently emerging super-resolution techniques overcome this barrier, and features are observable in the 20–50 nm range.

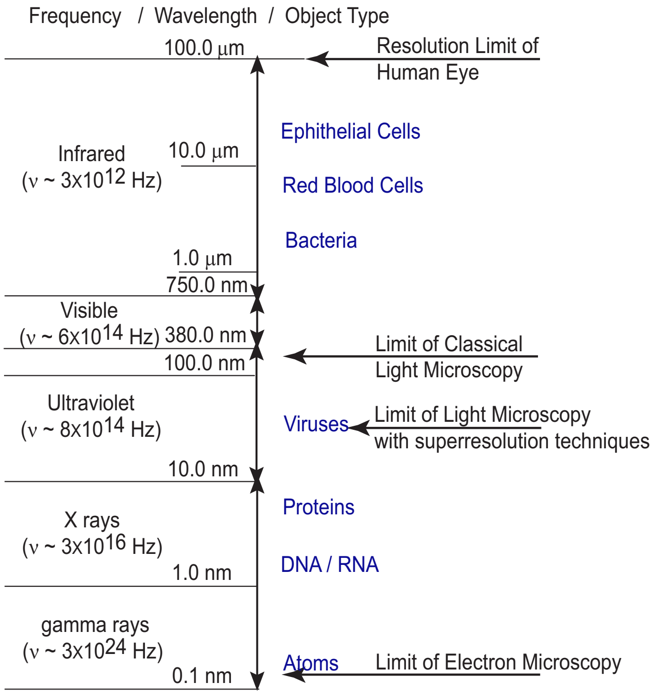
Frequency / Wavelength / Object Type
100.0 μm
Resolution Limit of Human Eye
Infrared (ν ~ 3×10^{12} Hz)
10.0 μm
Ephithelial Cells
Red Blood Cells
Bacteria
1.0 μm
750.0 nm
Visible (ν ~ 6×10^{14} Hz)
380.0 nm
Limit of Classical Light Microscopy
100.0 nm
Ultraviolet (ν ~ 8×10^{14} Hz)
Viruses
Limit of Light Microscopy with superresolution techniques
10.0 nm
X rays (ν ~ 3×10^{16} Hz)
Proteins
1.0 nm
DNA / RNA
gamma rays (ν ~ 3×10^{24} Hz)
0.1 nm
Atoms
Limit of Electron Microscopy

Wave Equations

Maxwell's equations describe the propagation of an electromagnetic wave. For homogenous and isotropic media, magnetic (*H*) and electric (*E*) components of an electromagnetic field can be described by the wave equations derived from Maxwell's equations:

$$\nabla^2 \vec{E} - \varepsilon_m \mu_m \frac{\partial^2 \vec{E}}{\partial t^2} = 0$$

$$\nabla^2 \vec{H} - \varepsilon_m \mu_m \frac{\partial^2 \vec{H}}{\partial t^2} = 0 ,$$

where ε is a **dielectric constant,** i.e., **medium permittivity** while μ is a **magnetic permeability**:

$$\varepsilon_m = \varepsilon_o \varepsilon_r \qquad \mu_m = \mu_o \mu_r$$

Indices r, m, and o stand for relative, media, or vacuum, respectively.

The above equations indicate that the time variation of the electric field and the current density creates a time-varying magnetic field. Variations in the magnetic field induce a time-varying electric field. Both fields depend on each other and compose an electromagnetic wave. Magnetic and electric components can be separated.

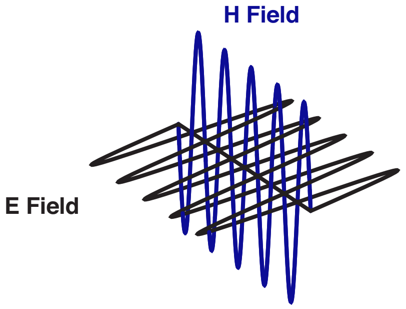

The electric component of an electromagnetic wave has primary significance for the phenomenon of light. This is because the magnetic component cannot be measured with optical detectors. Therefore, the magnetic component is most often neglected, and the electric vector $\vec{E}$ is called a light vector.

Wavefront Propagation

Light propagating through isotropic media conforms to the equation

$$E = A\sin\left(\omega t - kz + \varphi_o\right),$$

where t is time, z is distance along the direction of propagation, and ω is an **angular frequency** given by

$$\omega = \frac{2\pi}{T} = \frac{2\pi V_m}{\lambda}.$$

The term kz is called the **phase of light**, while φ_o is an initial phase. In addition, k represents the **wave number** (equal to $2\pi/\lambda$) and V_m is the velocity of light in media:

$$kz = \frac{2\pi}{\lambda} nz$$

The above propagation equation pertains to the special case when the electric field vector varies only in one plane.

The **refractive index** of the dielectric media n describes the relationship between the speed of light in a vacuum and in media. It is

$$n = \frac{c}{V_m},$$

where c is the speed of light in vacuum.

An alternative way to describe the propagation of an electric field is with an exponential (complex) representation:

$$E = A\exp\left[-i\left(\omega t - kz + \varphi_o\right)\right].$$

This form allows the easy separation of phase components of an electromagnetic wave.

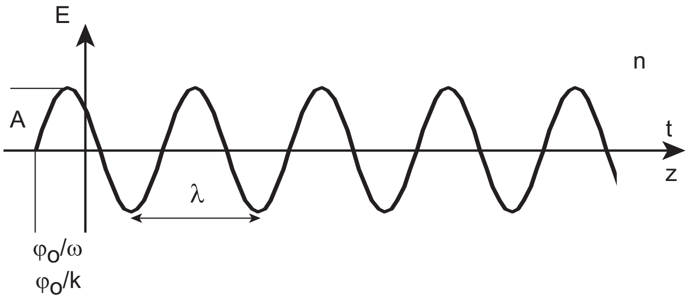

Optical Path Length (OPL)

Fermat's principle states that "The path traveled by a light wave from a point to another point is stationary with respect to variations of this path." In practical terms, this means that light travels along the minimum time path.

Optical path length (OPL) is related to the time required for light to travel from point P1 to point P2. It accounts for the media density through the refractive index:

$$t = \frac{1}{c}\int_{P1}^{P2} n\,ds$$

or

$$OPL = \int_{P1}^{P2} n\,ds\ ,$$

where

$$ds^2 = dx^2 + dy^2 + dz^2\ .$$

Optical path length can also be discussed in the context of the number of periods required to propagate a certain distance *L*. In a medium with refractive index *n*, light slows down and more wave cycles are needed. Therefore, OPL is an equivalent path for light traveling with the same number of periods in a vacuum:

$$OPL = nL\ .$$

Optical path difference (OPD) is the difference between optical path lengths traversed by two light waves:

$$OPD = n_1 L_1 - n_2 L_2\ .$$

OPD can also be expressed as a phase difference:

$$\Delta\varphi = 2\pi\frac{OPD}{\lambda}\ .$$

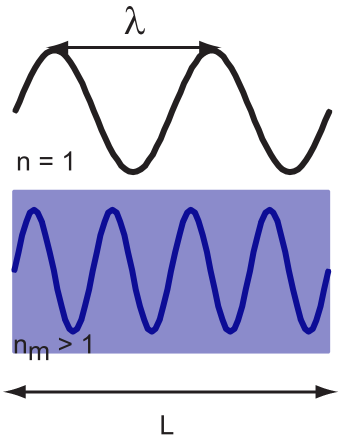

Laws of Reflection and Refraction

Light rays incident on an interface between different dielectric media experience reflection and refraction as shown.

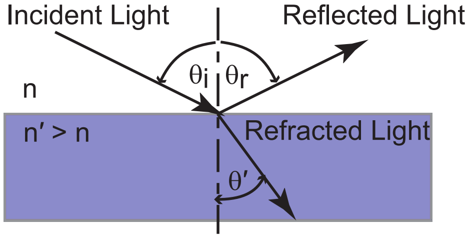

Reflection Law: Angles of **incidence** and **reflection** are related by

$$\theta_i = -\theta_r \, .$$

Refraction Law (**Snell's Law**): Incident and refracted angles are related to each other and to the refractive indices of the two media by Snell's Law:

$$n \sin\theta_i = n' \sin\theta' \, .$$

Fresnel reflection: The division of light at a dielectric boundary into transmitted and reflected rays is described for nonlossy, nonmagnetic media by the Fresnel equations:

Reflectance coefficients	Transmission Coefficients
$r_\perp = \frac{I_{r\perp}}{I_\perp} = \frac{\sin^2(\theta_i - \theta')}{\sin^2(\theta_i + \theta')}$ $r_\parallel = \frac{I_{r\parallel}}{I_\parallel} = \frac{\tan^2(\theta_i - \theta')}{\tan^2(\theta_i + \theta')}$	$t_\perp = \frac{I_{t\perp}}{I_\perp} = \frac{4\sin^2\theta' \cos^2\theta_i}{\sin^2(\theta_i + \theta')}$ $t_\parallel = \frac{I_{t\parallel}}{I_\parallel} = \frac{4\sin^2\theta' \cos^2\theta_i}{\sin^2(\theta_i + \theta')\cos^2(\theta_i - \theta')}$

t and r are transmission and reflection coefficients, respectively. I, I_t, and I_r are the irradiances of incident, transmitted, and reflected light, respectively. ⊥ and ‖ denote perpendicular and parallel components of the light vector with respect to the plane of incidence. θ_i and θ' in the table are the angles of incidence and reflection/refraction, respectively. At a normal incidence ($\theta' = \theta_i = 0$ deg), the Fresnel equations reduce to

$$r = r_\perp = r_\parallel = \frac{(n' - n)^2}{(n' + n)^2} \quad \text{and} \quad t = t_\perp = t_\parallel = \frac{4nn'}{(n' + n)^2} \, .$$

Total Internal Reflection

When light passes from one medium to a medium with a higher refractive index, the angle of the light in the medium bends toward the normal, according to Snell's Law. Conversely, when light passes from one medium to a medium with a lower refractive index, the angle of the light in the second medium bends away from the normal. If the angle of refraction is greater than 90 deg, the light cannot escape the denser medium, and it reflects instead of refracting. This effect is known as **total internal reflection** (TIR). TIR occurs only for illumination with an angle larger than the **critical angle**, defined as

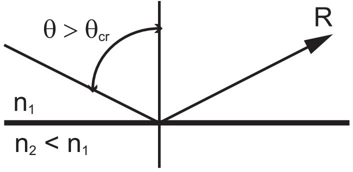

$$\theta_{cr} = \arcsin\left(\frac{n_1}{n_2}\right).$$

It appears, however, that light can propagate through (at a limited range) to the lower-density material as an **evanescent wave** (a nearly standing wave occurring at the material boundary), even if illuminated under an angle larger than θ_{cr}. Such a phenomenon is called **frustrated total internal reflection**. Frustrated TIR is used, for example, with thin films (TF) to build beam splitters. The proper selection of the film thickness provides the desired beam ratio (see the figure below for various split ratios). Note that the maximum film thickness must be approximately a single wavelength or less, otherwise light decays entirely. The effect of an evanescent wave propagating in the lower-density material under TIR conditions is also used in total internal reflection fluorescence (**TIRF**) microscopy.

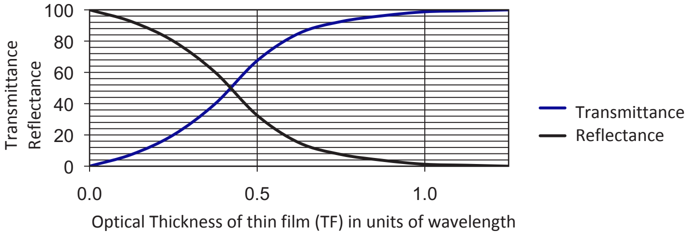

Evanescent Wave in Total Internal Reflection

A beam reflected at the dielectric interface during **total internal reflection** is subject to a small lateral shift, also known as the **Goos-Hänchen effect**. The shift is different for the parallel and perpendicular components of the electromagnetic vector.

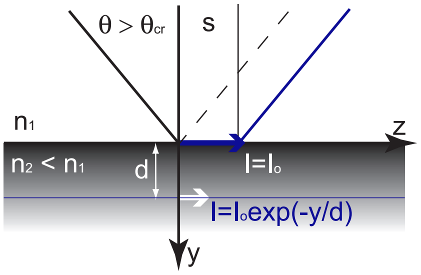

	Lateral shift in TIR
Parallel* component of e/m vector	$s_{\|} = \frac{\lambda n_1}{\pi n_2^2}\frac{\tan\theta}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
Perpendicular* component of e/m vector	$s_{\perp} = \frac{\lambda}{\pi n_1}\frac{\tan\theta}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
* to the plane of incidence	

The intensity of the illuminating beam decays exponentially with distance y in the direction perpendicular to the material boundary:

$$I = I_o \exp\left(-y/d\right).$$

Note that d denotes the distance at which the intensity of the illuminating light I_o drops by e. The decay distance is smaller than a wavelength. It is also inversely proportional to the illumination angle.

Decay distance (d) of evanescent wave	
Decay distance as a function of incidence and critical angles	$d = \frac{\lambda}{4\pi n_1}\frac{1}{\sqrt{\sin^2\theta - \sin^2\theta_c}}$
Decay distance as a function of incidence angle and refractive indices of media	$d = \frac{\lambda}{4\pi}\frac{1}{\sqrt{n_1^2\sin^2\theta - n_2^2}}$

Propagation of Light in Anisotropic Media

In anisotropic media the velocity of light depends on the direction of propagation. Common anisotropic and optically transparent materials include **uniaxial crystals**. Such crystals exhibit one direction of travel with a single propagation velocity. The single velocity direction is called the **optic axis** of the crystal. For any other direction, there are two velocities of propagation.

The wave in birefringent crystals can be divided into two components: ordinary and extraordinary. An **ordinary wave** has a uniform velocity in all directions, while the velocity of an **extraordinary wave** varies with orientation. The extreme value of an extraordinary wave's velocity is perpendicular to the optic axis. Both waves are linearly polarized, which means that the electric vector oscillates in one plane. The vibration planes of extraordinary and ordinary vectors are perpendicular. Refractive indices corresponding to the direction of the optic axis and perpendicular to the optic axis are $n_o = c/V_o$ and $n_e = c/V_e$ respectively.

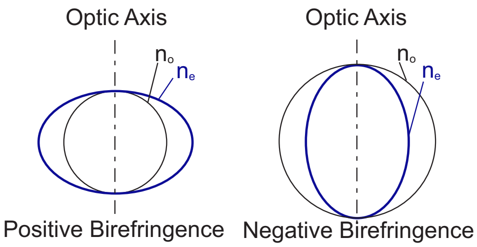

Uniaxial crystals can be positive or negative (see table). The **refractive index** $n(\xi)$ for velocities between the two extreme (n_o and n_e) values is

$$\frac{1}{n^2(\xi)} = \frac{\cos^2\xi}{n_o^2} + \frac{\sin^2\xi}{n_e^2}.$$

Positive Birefringence	$V_e \leq V_o$
Negative Birefringence	$V_e \geq V_o$

Note that the propagation direction ξ (with regard to the optic axis) is slightly off from the ray direction, which is defined by the Poynting (energy) vector.

Uniaxial Crystal	**Refractive index**	**Abbe Number**	**Wavelength Range** [μm]
Quartz	$n_o = 1.54424$ $n_e = 1.55335$	70 69	0.18–4.0
Calcite	$n_o = 1.65835$ $n_e = 1.48640$	50 68	0.2–2.0
The refractive index is given for the *D* spectral line (589.2 nm).			

(Adapted from Pluta, 1988)

Polarization of Light and Polarization States

The orientation characteristic of a light vector vibrating in time and space is called the **polarization of light**. For example, if the vector of an electric wave vibrates in one plane, the state of polarization is linear. A vector vibrating with a random orientation represents unpolarized light.

The wave vector E consists of two components E_x and E_y.

$$E(z,t) = E_x + E_y$$

$$E_x = A_x \exp\left[-i\left(\omega t - kz + \varphi_x\right)\right]$$

$$E_y = A_y \exp\left[-i\left(\omega t - kz + \varphi_y\right)\right].$$

The **electric vector** rotates periodically (the periodicity corresponds to wavelength λ) in the plane perpendicular to the propagation axis (z) and generally forms an elliptical shape. The specific shape depends on the A_x and A_y ratios and the phase delay between the E_x and E_y components, defined as $\Delta\varphi = \varphi_x - \varphi_y$.

Linearly polarized light is obtained when one of the components E_x or E_y is zero, or when $\Delta\varphi$ is zero or π. **Circularly polarized light** is obtained when $E_x = E_y$ and $\Delta\varphi = \pm\pi/2$. The light is called right circularly polarized (RCP) if it rotates in a clockwise direction or left circularly polarized (LCP) if it rotates counterclockwise when looking at the oncoming beam.

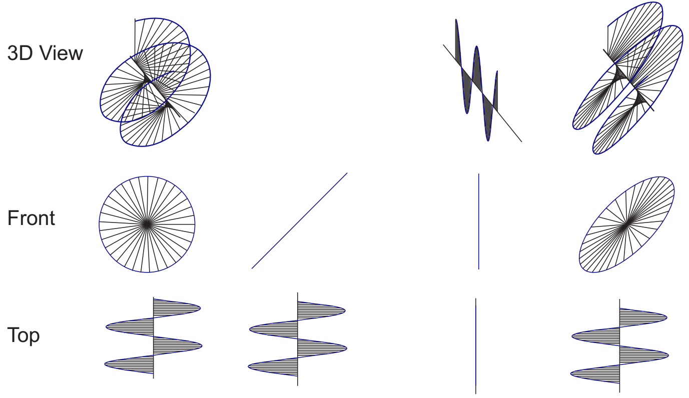

Polarization	Circular	Linear	Linear	Elliptical
Amplitudes	1 1	1 1	0 1	1 1
Δφ	π/2	0	0	π/4

Coherence and Monochromatic Light

An ideal light wave that extends in space at any instance to $\pm\infty$ and has only one wavelength λ (or frequency ν) is said to be **monochromatic**. If the range of wavelengths λ or frequencies ν is very small (around λ_o or ν_o, respectively), the wave is **quasi-monochromatic**. Such a wave packet is usually called a **wave group**.

Note that for monochromatic and quasi-monochromatic waves, no phase relationship is required, and the intensity of light can be calculated as a simple summation of intensities from different waves; phase changes are very fast and random, so only the average intensity can be recorded.

If multiple waves have a common phase relation dependence, they are **coherent** or **partially coherent**. These cases correspond to full- and partial-phase correlation respectively. A common source of a coherent wave is a laser where waves must be in resonance and therefore in phase. The average length of the wave packet (group) is called the **coherence length**, while the time required to pass this length is called **coherence time**. Both values are linked by the equations

$$\Delta\nu = \frac{1}{t_c} = \frac{V}{l_c}$$

where coherence length is

$$l_c = \frac{\lambda^2}{\Delta\lambda}.$$

The coherence length l_c and **temporal coherence** t_c are inversely proportional to the bandwidth $\Delta\lambda$ of the light source.

Spatial coherence is a term related to the coherence of light with regard to the extent of the light source. The fringe contrast varies for interference of any two spatially different source points. Light is **partially coherent** if its coherence is limited by the source bandwidth, dimension, temperature, or other effects.

Interference

Interference is a process of superposition of two coherent (correlated) or partially coherent waves. Waves interact with each other, and the resulting intensity is described by summing the **complex amplitudes** (**electric fields**) E_1 and E_2 of both wavefronts:

$$E_1 = A_1 \exp\left[-i\left(\omega t - kz + \varphi_1\right)\right]$$

and

$$E_2 = A_2 \exp\left[-i\left(\omega t - kz + \varphi_2\right)\right].$$

The resultant field is

$$E = E_1 + E_2 .$$

Therefore, the **interference** of the two beams can be written as

$$I = EE^*,$$

$$I = A_1^2 + A_2^2 + 2A_1A_2\cos\left(\Delta\varphi\right),$$

$$I = I_1 + I_2 + 2\sqrt{I_1I_2}\cos\left(\Delta\varphi\right),$$

$$I_1 = E_1E_1^*,\ \ I_2 = E_2E_2^*,\ \text{and}$$

$$\Delta\varphi = \varphi_2 - \varphi_1,$$

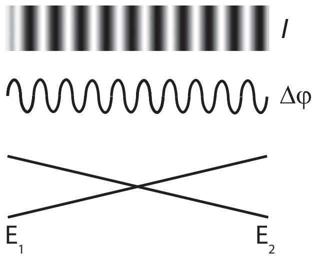

Propagating Wavefronts

where * denotes a conjugate function, I is the **intensity of light**, A is an amplitude of an electric field, and $\Delta\varphi$ is the phase difference between the two interfering beams. **Contrast** C (called also **visibility**) of the interference fringes can be expressed as

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}.$$

The fringe existence and visibility depends on several conditions. To obtain the interference effect,

- Interfering beams must originate from the same light source and be temporally and spatially coherent;
- The polarization of interfering beams must be aligned;
- To maximize the contrast, interfering beams should have equal or similar amplitudes.

Conversely, if two noncorrelated, random waves are in the same region of space, the sum of intensities (irradiances) of these waves gives a total intensity in that region: $I = I_1 + I_2$.

Contrast vs Spatial and Temporal Coherence

The result of interference is a periodic intensity change in space, which creates fringes when incident on a screen or detector. **Spatial coherence** relates to the contrast of these fringes depending on the extent of the source and is not a function of the phase difference (or OPD) between beams. The intensity of interfering fringes is given by

$$I = I_1 + I_2 + 2C\left(\text{Source Extent}\right)\sqrt{I_1 I_2}\cos\left(\Delta\varphi\right),$$

where C is a constant, depending on the extent of the source.

$C = 1$

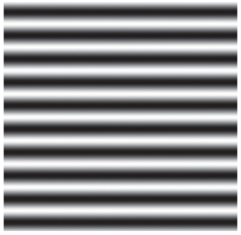

$C = 0.5$

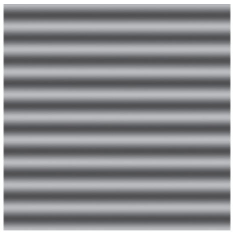

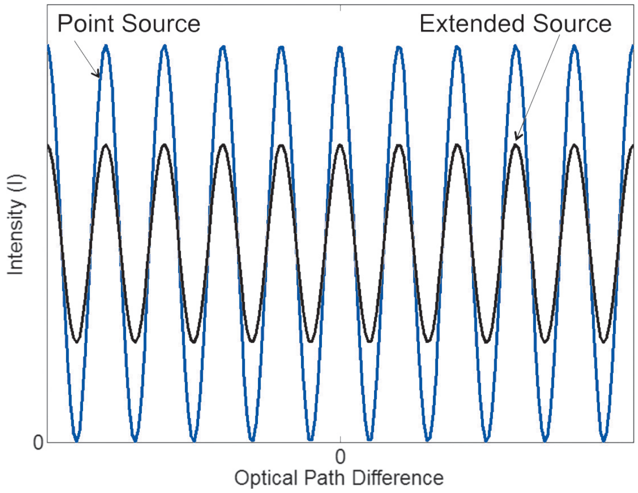

The spatial coherence can be improved through spatial filtering. For example, light can be focused on the pinhole (or coupled into the fiber) by using a microscope objective. In microscopy, spatial coherence can be adjusted by changing the diaphragm size in the conjugate plane of the light source.

Contrast vs Spatial and Temporal Coherence (cont.)

The intensity of the fringes depends on the OPD and the **temporal coherence** of the source. The fringe contrast trends toward zero as the OPD increases beyond the coherence length:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\, C\left(OPD\right)\cos\left(\Delta\varphi\right).$$

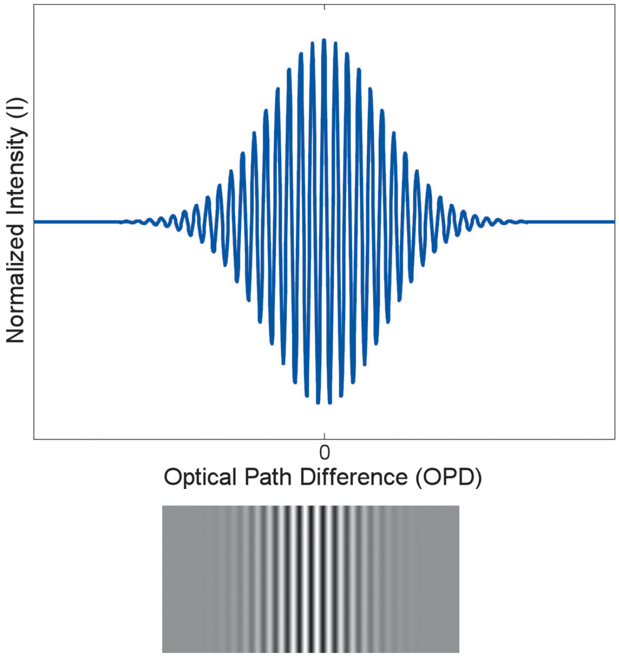

The spatial width of a fringe pattern envelope decreases with shorter temporal coherence. The location of the envelope's peak (with respect to the reference) and narrow width can be efficiently used as a gating mechanism in both imaging and metrology. For example, it is used in interference microscopy for optical profiling (white light interferometry) and for imaging using optical coherence tomography. Examples of short coherence sources include white light sources, luminescent diodes, and broadband lasers.

Long coherence length is beneficial in applications that do not naturally provide near-zero OPD, e.g., in surface measurements with a Fizeau interferometer.

Contrast of Fringes (Polarization and Amplitude Ratio)

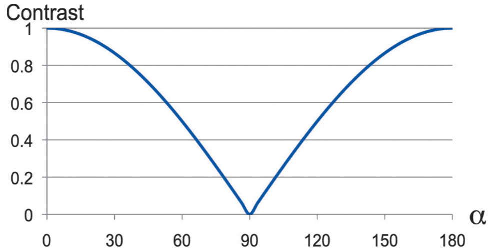

The contrast of fringes depends on the polarization orientation of the electric vectors. For linearly polarized beams, it can be described as $C = |\cos\alpha|$, where α represents the angle between polarization states.

Angle between Electric Vectors of Interfering Beams			
0	π/6	π/3	π/2
Interference Fringes			
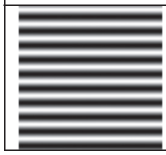	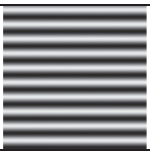	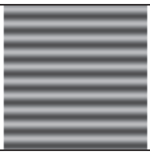	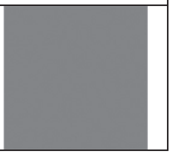

Fringe contrast also depends on the interfering beams' amplitude ratio. The intensity of the fringes is simply calculated by using the main interference equation. The contrast is maximum for equal beam intensities and for the interference pattern, defined as

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\Delta\varphi),$$

it is

$$C = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}.$$

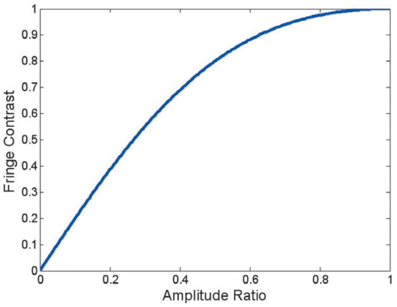

Ratio between Amplitudes of Interfering Beams			
1	0.5	0.25	0.1
Interference Fringes			
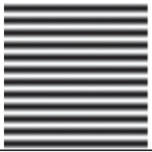	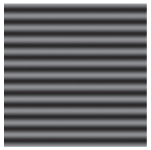	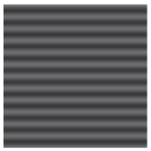	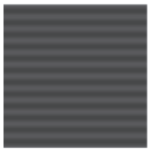

Multiple Wave Interference

If light reflects inside a thin film, its intensity gradually decreases and **multiple beam interference** occurs.

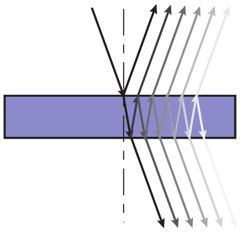

The intensity of reflected light is

$$I_r = I_i \frac{F\sin^2(\varphi/2)}{1+F\sin^2(\varphi/2)} ,$$

and for transmitted light is

$$I_t = I_i \frac{1}{1+F\sin^2(\varphi/2)} .$$

The coefficient of finesse F of such a resonator is

$$F = \frac{4r}{(1-r)^2} .$$

Because phase depends on the wavelength, it is possible to design selective **interference filters**. In this case, a dielectric thin film is coated with metallic layers. The peak wavelength for the m^{th} interference order of the filter can be defined as

$$\lambda_p = \frac{2nt\cos\theta'}{m-\left(\frac{\varphi_{TF}}{2\pi}\right)} ,$$

where φ_{TF} is the phase difference generated by a thin film of thickness t and a specific incidence angle. The **interference order** relates to the phase difference in multiples of 2π.

Interference filters usually operate for illumination with flat wavefronts at a normal incidence angle. Therefore, the equation simplifies as

$$\lambda_p = \frac{2nt}{m} .$$

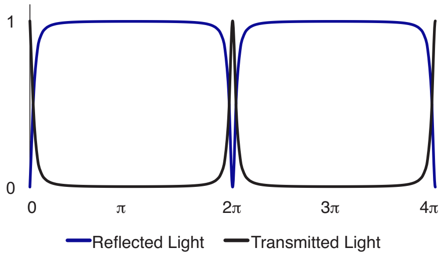

The **half bandwidth** (HBW) of the filter for normal incidence is

$$HBW = \frac{1-r}{m\pi}\lambda_p \ [\mathrm{m}] .$$

The peak intensity transmission is usually at 20% to 50%, or up to 90% of the incident light for metal-dielectric or multi-dielectric filters respectively.

Interferometers

Due to the high frequency of light, it is not possible to detect the phase of the light wave directly. To acquire the phase, interferometry techniques may be used. There are two major classes of interferometers, based on **amplitude splitting** and **wavefront splitting**.

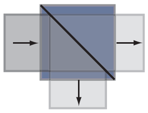

Amplitude Split

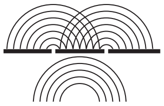

Wavefront Split

From the point of view of fringe pattern interpretation, interferometers can also be divided into those that deliver fringes, which directly correspond to the phase distribution, and those providing a derivative of the phase map (also called shearing interferometers). The first type is realized by obtaining the interference between the test beam and a reference beam. An example of such a system is the **Michelson interferometer**. **Shearing interferometers** provide the intereference of two shifted object wavefronts.

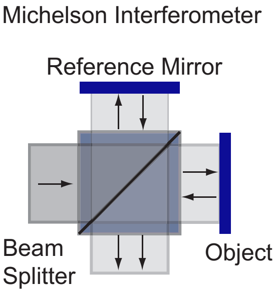
Michelson Interferometer
Reference Mirror
Beam Splitter
Object

There are numerous ways of introducing shear (linear, radial, etc.). Examples of shearing interferometers include the parallel or wedge plate, the grating interferometer, the **Sagnac**, and polarization interferometers (commonly used in microscopy). The **Mach-Zehnder interferometer** can be configured for direct and differential fringes, depending on the position of the sample.

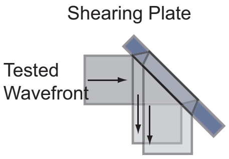
Shearing Plate
Tested Wavefront

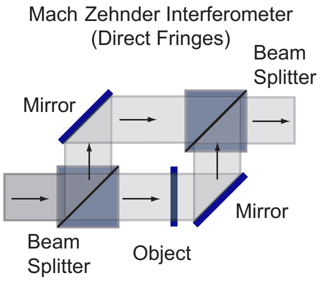
Mach Zehnder Interferometer
(Direct Fringes)
Beam Splitter
Mirror
Mirror
Beam Splitter
Object

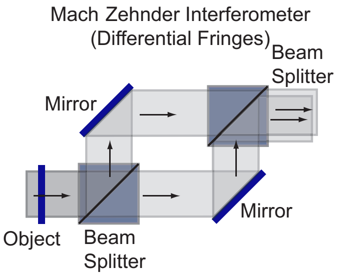
Mach Zehnder Interferometer
(Differential Fringes)
Beam Splitter
Mirror
Mirror
Object
Beam Splitter

Diffraction

The bending of waves by apertures and objects is called **diffraction of light**. Waves diffracted inside the optical system interfere (for path differences within the coherence length of the source) with each other and create diffraction patterns in the image of an object.

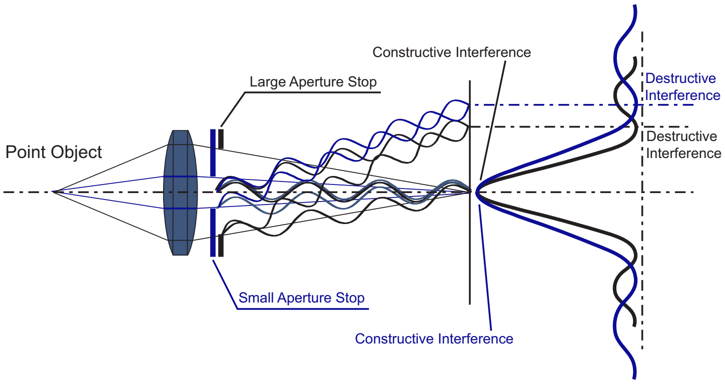
Constructive Interference
Large Aperture Stop
Point Object
Destructive Interference
Destructive Interference
Small Aperture Stop
Constructive Interference

There are two common approximations of diffraction phenomena: **Fresnel diffraction** (near-field) and **Fraunhofer diffraction** (far-field). Both diffraction types complement each other but are not sharply divided due to various definitions of their regions. Fraunhofer diffraction occurs when one can assume that propagating wavefronts are flat (collimated), while the Fresnel diffraction is the near-field case. Thus, Fraunhofer diffraction (distance z) for a free-space case is infinity, but in practice it can be defined for a region

$$z > S_{\mathrm{FD}} \frac{d^2}{\lambda},$$

where d is the diameter of the diffractive object, and S_{FD} is a factor depending on approximation. A conservative definition of the Fraunhofer region is $S_{\mathrm{FD}} = 1$, while for most practical cases it can be assumed to be 10 times smaller ($S_{\mathrm{FD}} = 0.1$).

Diffraction effects influence the resolution of an imaging system and are a reason for fringes (ring patterns) in the image of a point. Specifically, Fraunhofer fringes appear in the conjugate image plane. This is due to the fact that the image is located in the Fraunhofer diffracting region of an optical system's aperture stop.

Diffraction Grating

Diffraction gratings are optical components consisting of periodic linear structures that provide "organized" diffraction of light. In practical terms, this means that for specific angles (depending on illumination and grating parameters), it is possible to obtain constructive (2π phase difference) and destructive (π phase difference) interference at specific directions. The angles of constructive interference are defined by the diffraction grating equation and called **diffraction orders** (m).

Gratings can be classified as transmission or reflection (mode of operation), amplitude (periodic amplitude changes) or phase (periodic phase changes), or ruled or holographic (method of fabrication). Ruled gratings are mechanically cut and usually have a triangular profile (each facet can thereby promote select diffraction angles). Holographic gratings are made using interference (sinusoidal profiles).

Diffraction angles depend on the ratio between the grating constant and wavelength so various wavelengths can be separated. This makes them applicable for spectroscopic detection or spectral imaging. The **grating equation** is

$$m\lambda = d \cos\gamma (\sin\beta \pm \sin\alpha)$$

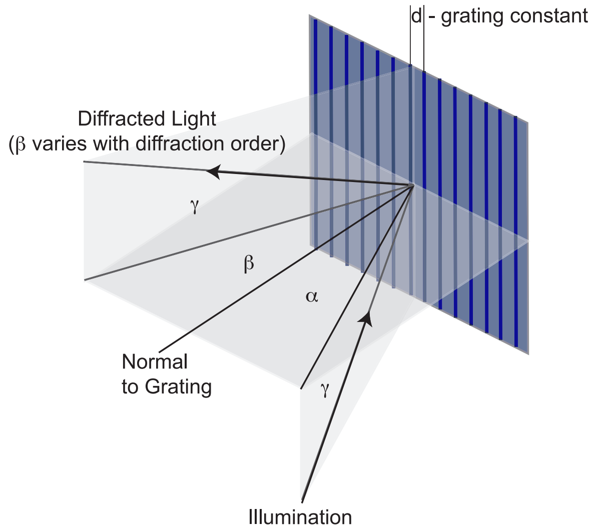

Diffraction Grating (cont.)

The sign in the diffraction grating equation defines the type of grating: A **transmission grating** is identified with a minus sign (–), and a **reflective grating** is identified with a plus sign (+).

For a grating illuminated in the plane perpendicular to the grating, its equation simplifies and is

$$m\lambda = d\left(\sin\beta \pm \sin\alpha\right).$$

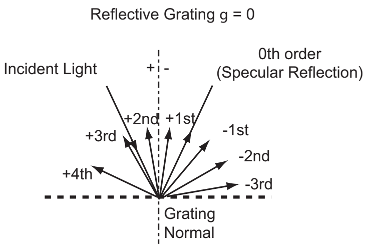
Reflective Grating g = 0
Incident Light
+ -
0th order
(Specular Reflection)
+2nd
+1st
+3rd
-1st
-2nd
+4th
-3rd
Grating
Normal

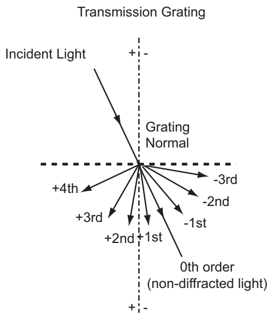
Transmission Grating
Incident Light
+ -
Grating
Normal
-3rd
+4th
-2nd
+3rd
-1st
+2nd
+1st
0th order
(non-diffracted light)
+ -

For normal illumination, the grating equation becomes

$$m\lambda = d\sin\beta.$$

The **chromatic resolving power** of a grating depends on its size (total number of lines N). If the grating has more lines, the diffraction orders are narrower, and the resolving power increases:

$$\frac{\lambda}{\Delta\lambda} = mN.$$

The **free spectral range** $\Delta\lambda$ of the grating is the bandwidth in the m^{th} order without overlapping other orders. It defines useful bandwidth for spectroscopic detection as

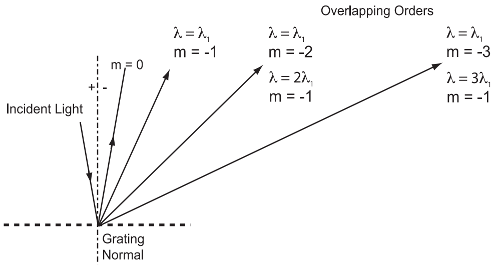
Overlapping Orders
$\lambda = \lambda_1$
m = -1
$\lambda = \lambda_1$
m = -2
$\lambda = 2\lambda_1$
m = -1
$\lambda = \lambda_1$
m = -3
$\lambda = 3\lambda_1$
m = -1
m = 0
+ -
Incident Light
Grating
Normal

$$\Delta\lambda = \frac{m+1}{m}\lambda_1 - \lambda_1 = \frac{\lambda_1}{m}.$$

Useful Definitions from Geometrical Optics

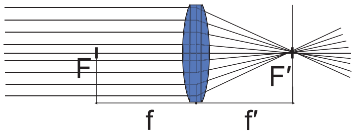

All optical systems consist of refractive or reflective interfaces (surfaces) changing the propagation angles of **optical rays**.

Rays define the propagation trajectory and always travel perpendicular to the wavefronts. They are used to describe imaging in the regime of geometrical optics and to perform optical design.

There are two **optical spaces** corresponding to each optical system: **object space** and **image space**. Both spaces extend from –∞ to +∞ and are divided into real and virtual parts. Ability to access the image or object defines the real part of the optical space.

Paraxial optics is an approximation assuming small ray angles and is used to determine first-order parameters of the optical system (image location, magnification etc.).

Thin lenses are optical components reduced to zero thickness and are used to perform first-order optical designs (see next page).

The **focal point** of an optical system is a location that collimated beams converge to or diverge from. Planes perpendicular to the optical axis at the focal points are called **focal planes**. **Focal length** is the distance between the lens (specifically, its principal plane) and the focal plane. For thin lenses, principal planes overlap with the lens.

Sign Convention: The common sign convention assigns positive values to distances traced from left to right (the direction of propagation) and to the top. Angles are positive if they are measured counterclockwise from normal to the surface or optical axis. If light travels from right to left, the refractive index is negative. The surface radius is measured from its vertex to its center of curvature.

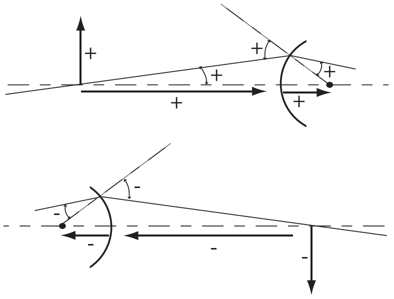

Image Formation

A simple model describing imaging through an optical system is based on thin lens relationships. A **real image** is formed at the point where rays converge.

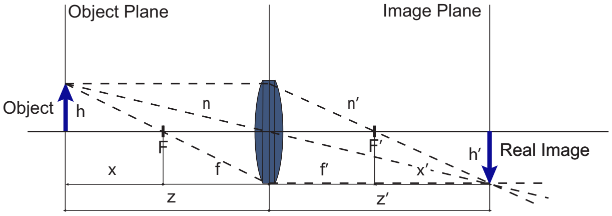
Object Plane
Image Plane
Object
h
n
n′
F
F′
h′
Real Image
x
f
f′
x′
z
z′

A **virtual image** is formed at the point from which rays appear to diverge.

For a lens made of glass, surrounded on both sides by air ($n = n' = 1$), the imaging relationship is described by the **Newtonian equation**

$$xx' = ff' \qquad \text{or} \qquad xx' = -f'^2 .$$

Note that Newtonian equations refer to the distance from the focal planes so, in practice, they are used with thin lenses.

Imaging can also be described by the **Gaussian imaging equation**

$$\frac{f}{z} + \frac{f'}{z'} = 1 .$$

The **effective focal length** of the system is

$$f_e = \frac{1}{\varphi} = -\frac{f}{n} = \frac{f'}{n'} ,$$

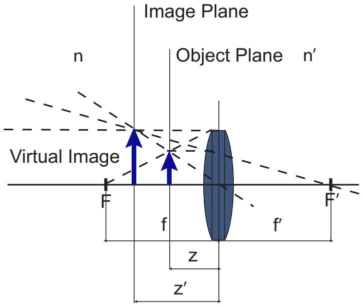
Image Plane
n
Object Plane
n′
Virtual Image
F
F′
f
f′
z
z′

where φ is the **optical power** expressed in diopters D [m^{-1}].

Therefore,

$$\frac{n'}{z'} = \frac{n}{z} + \frac{1}{f_e} .$$

For air, when n and n' are equal to 1, the imaging relation is

$$\frac{1}{z'} = \frac{1}{z} + \frac{1}{f_e} .$$

Magnification

Transverse magnification (also called **lateral magnification**) of an optical system is defined as the ratio of the image and object size measured in the direction perpendicular to the optical axis:

$$M = \frac{h'}{h} = -\frac{x'}{f'} = -\frac{f}{x} = \frac{z'}{z} = \frac{-f}{z-f} = \frac{-(z'-f')}{f'} .$$

Longitudinal magnification defines the ratio of distances for a pair of conjugate planes:

$$\frac{\Delta z'}{\Delta z} = \left(-\frac{f'}{f}\right) M_1 M_2 ,$$

where

$$\Delta z = z_2 - z_1$$
$$\Delta z' = z_2' - z_1'$$

and

$$M_1 = \frac{h_1'}{h_1}$$
$$M_2 = \frac{h_2'}{h_2} .$$

Angular magnification is the ratio of angular image size to angular object size and can be calculated with

$$M_u = \frac{u'}{u} = \frac{z}{z'} .$$

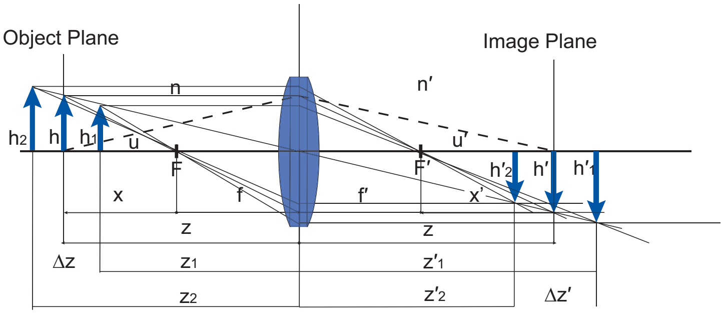

Stops and Rays in an Optical System

The primary stops in any optical system are the **aperture stop** (which limits light) and **field stop** (which limits the extent of the imaged object or the field of view). The aperture stop also defines the resolution of the optical system. To determine the aperture stop, all system diaphragms including the lens mounts should be imaged to either the image or the object space of the system. The aperture stop is determined as the smallest diaphragm or diaphragm image (in terms of angle) as seen from the on-axis object/image point in the same optical space.

Note that there are two important conjugates of the aperture stop in object and image space. They are called the **entrance pupil** and **exit pupil**, respectively.

The physical stop limiting the extent of the field is called the **field stop**. To find a field stop, all of the diaphragms should be imaged to the object or image space, with the smallest diaphragm defining the actual field stop, as seen from the entrance/exit pupil. Conjugates of the field stop in the object and image space are called the **entrance window** and **exit window**, respectively.

Two major rays that pass through the system are the **marginal ray** and the **chief ray**. The marginal ray crosses the optical axis at the conjugate of the field stop (and object) and the edge of the aperture stop. The chief ray goes through the edge of the field and crosses the optical axis at the aperture stop plane.

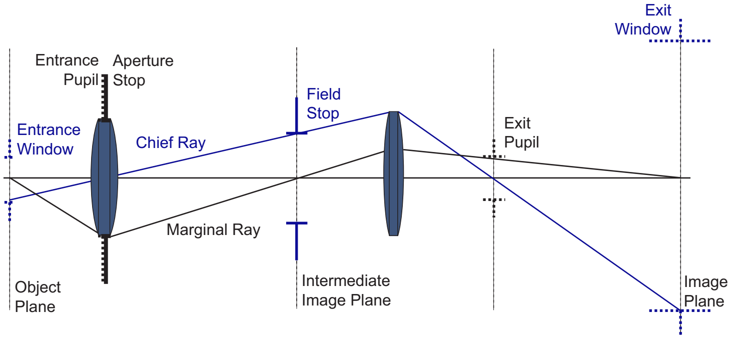

Aberrations

Individual spherical lenses cannot deliver perfect imaging because they exhibit errors called aberrations. All aberrations can be considered either **chromatic** or **monochromatic**. To correct for aberrations, optical systems use multiple elements, aspherical surfaces, and a variety of optical materials.

Chromatic aberrations are a consequence of the **dispersion** of optical materials, which is a change in the refractive index as a function of wavelength. The parameter-characterizing dispersion of any material is called the **Abbe number** and is defined as

$$V_{\mathrm{d}} = \frac{n_{\mathrm{d}} - 1}{n_{\mathrm{F}} - n_{\mathrm{C}}}.$$

Alternatively, the following equation might be used for other wavelengths:

$$V_{\mathrm{e}} = \frac{n_{\mathrm{e}} - 1}{n_{\mathrm{F'}} - n_{\mathrm{C'}}}.$$

λ [nm]	Symbol	Spectral Line
656	C	red hydrogen
644	C′	red cadmium
588	d	yellow helium
546	e	green mercury
486	F	blue hydrogen
480	F′	blue cadmium

In general, V can be defined by using refractive indices at any three wavelengths, which should be specified for material characteristics. Indices in the equations denote spectral lines. If V does not have an index, V_{d} is assumed.

Geometrical aberrations occur when optical rays do not meet at a single point. There are longitudinal and transverse ray aberrations describing the axial and lateral deviations from the paraxial image of a point (along the axis and perpendicular to the axis in the image plane), respectively.

Wave aberrations describe a deviation of the wavefront from a perfect sphere. They are defined as a distance (the optical path difference) between the wavefront and the reference sphere along the optical ray.

Chromatic Aberrations

Chromatic aberrations occur due to the dispersion of optical materials used for lens fabrication. This means that the refractive index is different for different wavelengths; consequently, various wavelengths are refracted differently.

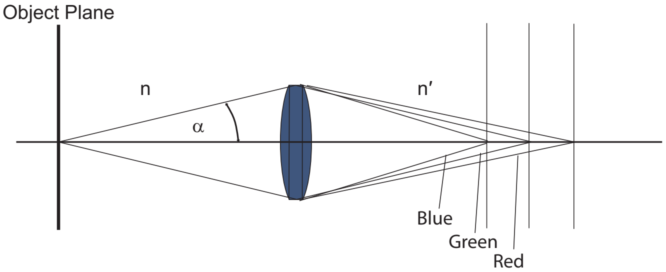

Chromatic aberrations include axial (longitudinal) or transverse (lateral) aberrations. **Axial chromatic aberration** arises from the fact that various wavelengths are focused at different distances behind the optical system. It is described as a variation in focal length:

$$\frac{\Delta f}{f} = \frac{f_F - f_C}{f} = \frac{1}{V} .$$

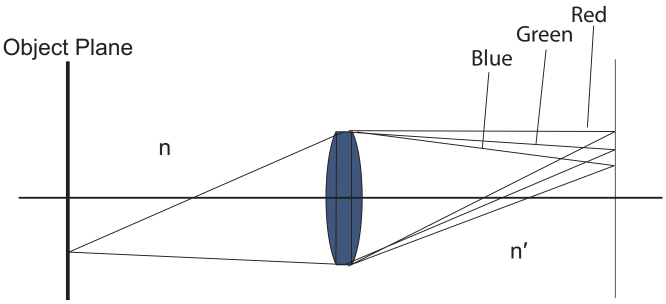

Transverse chromatic aberration is an off-axis imaging of colors at different locations on the image plane.

To compensate for chromatic aberrations, materials with low and high Abbe numbers are used (such as flint and crown glass). Correcting chromatic aberrations is crucial for most microscopy applications, but it is especially important for multi-photon microscopy. Obtaining multi-photon excitation requires high laser power and is most effective using short pulse lasers. Such a light source has a broad spectrum, and chromatic aberrations may cause pulse broadening.

Spherical Aberration and Coma

The most important wave aberrations are spherical, coma, astigmatism, field curvature, and distortion.

Spherical aberration (on-axis) is a consequence of building an optical system with components with spherical surfaces. It occurs when rays from different heights in the pupil are focused at different planes along the optical axis. This results in an axial blur. The most common approach for correcting spherical aberration uses a combination of negative and positive lenses. Systems that correct spherical aberration heavily depend on imaging conditions. For example, in microscopy a cover glass must be of an appropriate thickness and refractive index in order to work with an objective. Also, the media between the objective and the sample (such as air, oil, or water) must be taken into account.

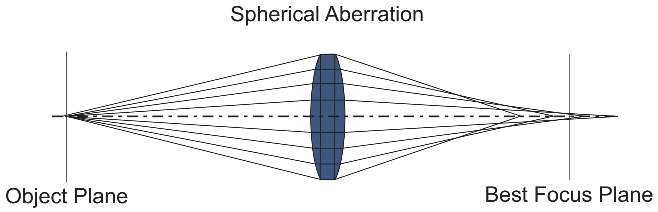

Coma (off-axis) can be defined as a variation of magnification with aperture location. This means that rays passing through a different azimuth of the lens are magnified differently. The name "coma" was inspired by the aberration's appearance, because it resembles a comet's tail as it emanates from the focus spot. It is usually stronger for lenses with a larger field, and its correction requires accommodation of the field diameter.

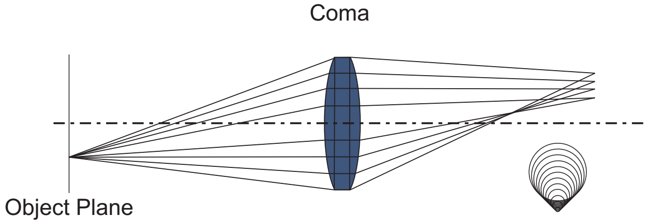

Astigmatism, Field Curvature, and Distortion

Astigmatism (off-axis) is responsible for different magnifications along orthogonal meridians in an optical system. It manifests as elliptical, elongated spots for the horizontal and vertical directions on opposite sides of the best focal plane. It is more pronounced for an object farther from the axis and is a direct consequence of improper lens mounting or an asymmetric fabrication process.

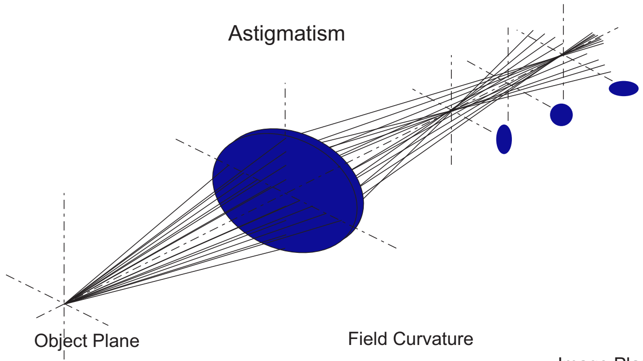

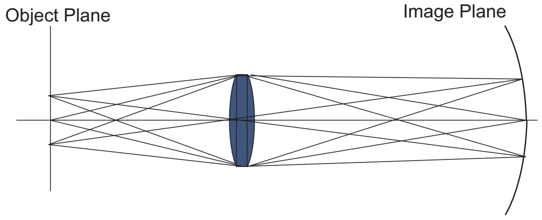

Field curvature (off-axis) results in a non-flat image plane. The image plane created is a concave surface as seen from the objective; therefore, various zones of the image can be seen in focus after moving the object along the optical axis. This aberration is corrected by an objective design combined with a tube lens or eyepiece.

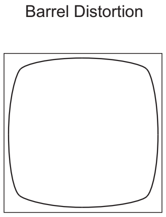

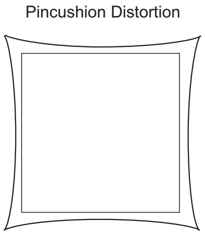

Distortion is a radial variation of magnification that will image a square as a pincushion or barrel. It is corrected in the same manner as field curvature. If preceded with system calibration, it can also be corrected numerically after image acquisition.

Performance Metrics

The major metrics describing the performance of an optical system are the **modulation transfer function** (MTF), the **point spread function** (PSF), and the **Strehl ratio** (SR).

The MTF is the modulus of the **optical transfer function** described by

$$OTF(\rho) = MTF(\rho)\exp\left[i\varphi(\rho)\right],$$

where the complex term in the equation relates to the phase transfer function. The MTF is a contrast distribution in the image in relation to contrast in the object as a function of spatial frequency ρ (for sinusoidal object harmonics) and can be defined as

$$MTF(\rho) = \frac{C(\rho)_{\text{image}}}{C(\rho)_{\text{object}}}.$$

The PSF is the intensity distribution at the image of a point object. This means that the PSF is a metric directly related to the image, while the MTF corresponds to spatial frequency distributions in the pupil. The MTF and PSF are closely related and comprehensively describe the quality of the optical system. In fact, the amplitude of the Fourier transform of the PSF results in the MTF.

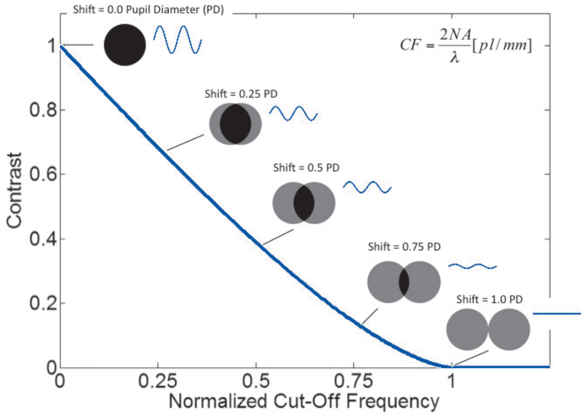

The MTF can be calculated as the autocorrelation of the **pupil function**. The pupil function describes the field distribution of an optical wave in the pupil plane of the optical system. In the case of a uniform pupil's transmission, it directly relates to the field overlap of two mutually shifted pupils where the shift corresponds to spatial frequency.

Performance Metrics (cont.)

The **modulation transfer function** has different results for coherent and incoherent illumination. For incoherent illumination, the phase component of the field is neglected since it is an average of random fields propagating under random angles.

For coherent illumination, the contrast of transferring harmonics of the field is constant and equal to 1 until the location of the spatial frequency in the pupil reaches its edge. For higher frequencies, the contrast sharply drops to zero since they cannot pass the optical system. Note that contrast for the coherent case is equal to 1 for the entire MTF range.

The **cutoff frequency** for an incoherent system is two times the cutoff frequency of the equivalent aperture coherent system and defines the **Sparrow resolution limit**.

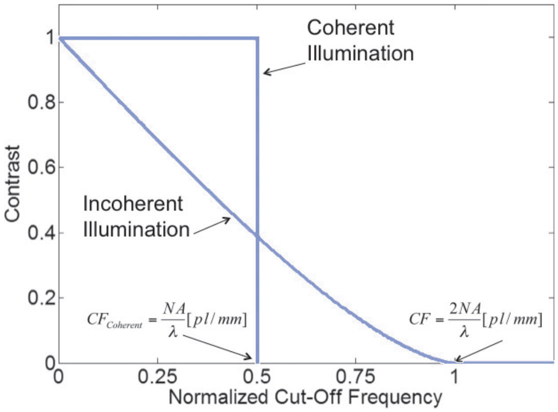

The **Strehl ratio** is a parametric measurement that defines the quality of the optical system in a single number. It is defined as the ratio of irradiance within the theoretical dimension of a diffraction-limited spot to the entire irradiance in the image of the point. One simple method to estimate the Strehl ratio is to divide the field below the MTF curve of a tested system by the field of the diffraction-limited system of the same numerical aperture. For practical optical design consideration it is usually assumed that the system is diffraction limited if the Strehl ratio is equal to or larger than 0.8.

The Compound Microscope

The primary goal of microscopy is to provide the ability to resolve the small details of an object. Historically, microscopy was developed for visual observation and, therefore, was set up to work in conjunction with the human eye. An effective high-resolution optical system must have resolving ability and be able to deliver proper magnification for the detector. In the case of visual observations, the detectors are the cones and rods of the retina.

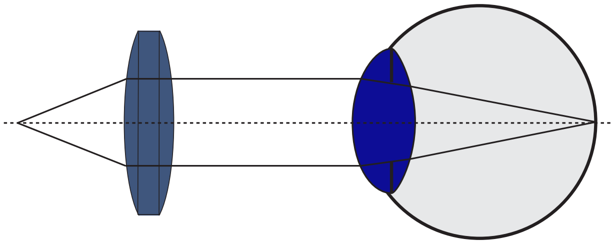

A basic microscope can be built with a single-element, short-focal-length lens (**magnifier**). The object is located in the focal plane of the lens and is imaged to infinity. The eye creates a final image of the object on the retina. The system stop is the eye's pupil.

To obtain higher resolution for visual observation, the **compound microscope** was first built in the 17th century by Robert Hooke. It consists of two major components: the **objective** and the **eyepiece**. The short-focal-length lens (objective) is placed close to the object under examination and creates a real image of an object at the focal plane of the second lens. The eyepiece (similar to the magnifier) throws an image to infinity, and the human eye creates the final image. An important function of the eyepiece is to match the eye's pupil with the system stop, which is located in the back focal plane of the microscope objective.

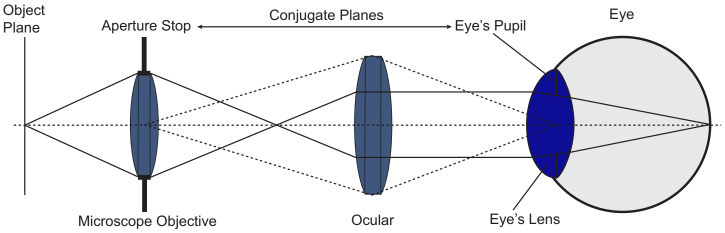

The Eye

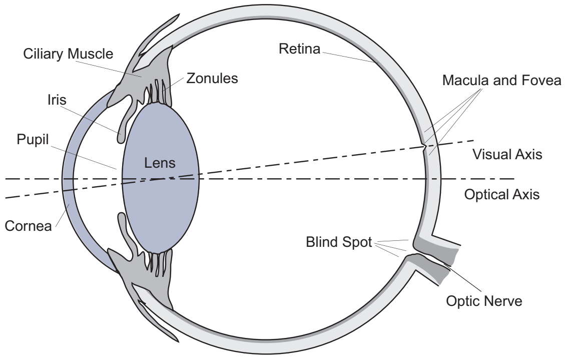

The **eye** was the first—and for a long time, the only—real-time detector used in microscopy. Therefore, the design of the microscope had to incorporate parameters responding to the needs of visual observation.

Cornea—the transparent portion of the sclera (the white portion of the human eye), which is a rigid tissue that gives the eyeball its shape. The cornea is responsible for two thirds of the eye's refractive power.

Lens—the lens is responsible for one third of the eye's power. Ciliary muscles can change the lens's power (accommodation) within the range of 15–30 diopters.

Iris—controls the diameter of the pupil (1.5–8 mm).

Retina—a layer with two types of photoreceptors: rods and cones. **Cones** (about 7 million) are in the area of the **macula** (~3 mm in diameter) and **fovea** (~1.5 mm in diameter, with the highest cone density), and they are designed for bright vision and color detection. There are three types of cones (red, green, and blue sensitive), and the spectral range of the eye is approximately 400–750 nm. **Rods** are responsible for night/low-light vision, and there are about 130 million located outside the fovea region.

It is arbitrarily assumed that the eye can provide sharp images for objects between 250 mm and infinity. A 250-mm distance is called the **minimum focus distance** or **near point**. The maximum eye resolution for bright illumination is 1 arc minute.

Upright and Inverted Microscopes

The two major microscope geometries are **upright** and **inverted.** Both systems can operate in reflectance and transmittance modes.

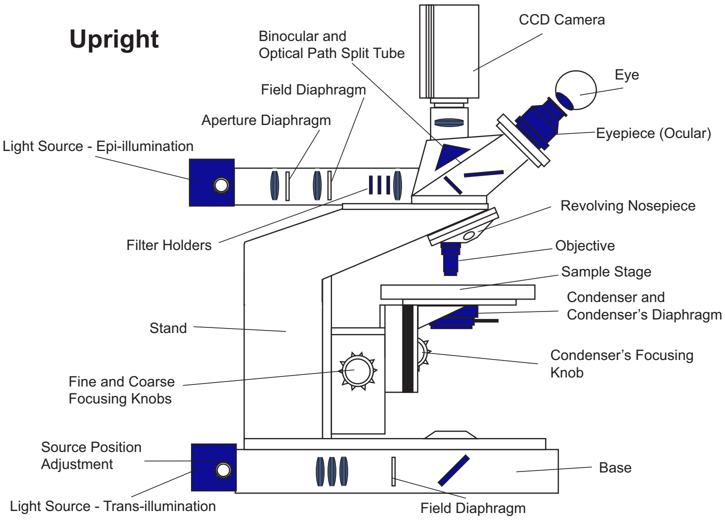
Upright
CCD Camera
Binocular and
Optical Path Split Tube
Eye
Field Diaphragm
Aperture Diaphragm
Eyepiece (Ocular)
Light Source - Epi-illumination
Revolving Nosepiece
Filter Holders
Objective
Sample Stage
Condenser and
Condenser's Diaphragm
Stand
Condenser's Focusing
Knob
Fine and Coarse
Focusing Knobs
Source Position
Adjustment
Base
Light Source - Trans-illumination
Field Diaphragm

The inverted microscope is primarily designed to work with samples in cell culture dishes and to provide space (with a long-working distance condenser) for sample manipulation (for example, with patch pipettes in electrophysiology).

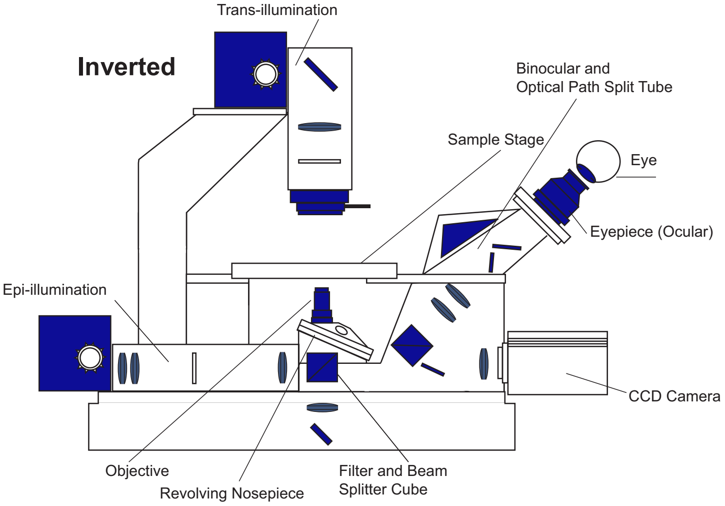
Trans-illumination
Inverted
Binocular and
Optical Path Split Tube
Sample Stage
Eye
Eyepiece (Ocular)
Epi-illumination
CCD Camera
Objective
Revolving Nosepiece
Filter and Beam
Splitter Cube

The Finite Tube Length Microscope

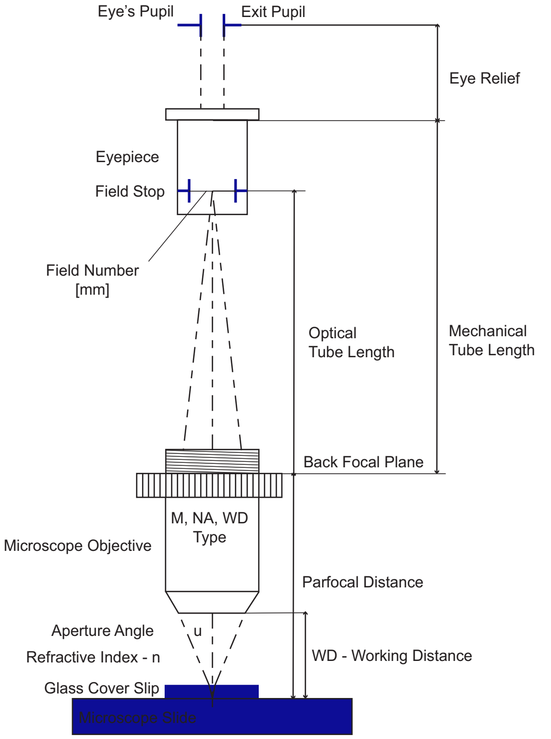
Eye's Pupil
Exit Pupil
Eye Relief
Eyepiece
Field Stop
Field Number [mm]
Optical Tube Length
Mechanical Tube Length
Back Focal Plane
M, NA, WD Type
Microscope Objective
Parfocal Distance
Aperture Angle
u
Refractive Index - n
WD - Working Distance
Glass Cover Slip
Microscope Slide

Historically, microscopes were built with a **finite tube length**. With this geometry, the microscope objective images the object into the tube end. This intermediate image is then relayed to the observer by an eyepiece. Depending on the manufacturer, different **optical tube lengths** are possible (for example, the standard tube length for Zeiss is 160 mm). The use of a standard tube length by each manufacturer unifies the optomechanical design of the microscope.

A constant **parfocal distance** for microscope objectives enables switching between different magnifications without defocusing. The **field number**, corresponding to the physical dimension (in millimeters) of the **field stop** inside the eyepiece, allows the microscope's field of view (FOV) to be determined according to

$$FOV = \frac{FieldNumber}{M_{\text{objective}}} \, [\text{mm}].$$

Infinity-Corrected Systems

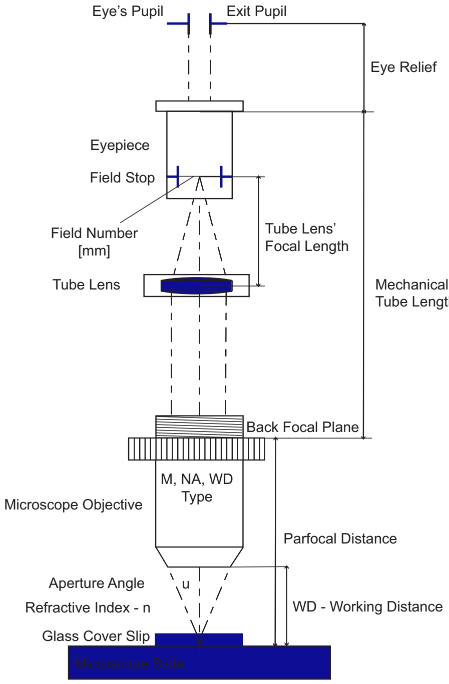
Eye's Pupil
Exit Pupil
Eye Relief
Eyepiece
Field Stop
Field Number [mm]
Tube Lens' Focal Length
Tube Lens
Mechanical Tube Length
Back Focal Plane
M, NA, WD Type
Microscope Objective
Parfocal Distance
Aperture Angle
u
Refractive Index - n
WD - Working Distance
Glass Cover Slip
Microscope Slide

Microscope objectives can be corrected for a conjugate located at infinity, which means that the microscope objective images an object into infinity. An infinite conjugate is useful for introducing beam splitters or filters that should work with small incidence angles. Also, an infinity-corrected system accommodates additional components like DIC prisms, polarizers. etc. The collimated beam is focused to create an intermediate image with additional optics, called a **tube lens**. The tube lens either creates an image directly onto a CCD chip or an intermediate image, which is further reimaged with an eyepiece. Additionally, the tube lens might be used for system correction. For example, Zeiss corrects aberrations in its microscopes with a combination objective-tube lens.

In the case of an **infinity-corrected objective**, the transverse magnification can only be defined in the presence of a tube lens that will form a real image. It is given by the ratio between the tube lens's focal length and the focal length of the microscope objective.

Manufacturer	Focal Length of Tube Lens
Zeiss	164.5 mm
Olympus	180.0 mm
Nikon	200.0 mm
Leica	200.0 mm

Telecentricity of a Microscope

Telecentricity is a feature of an optical system where the principal ray in object, image, or both spaces is parallel to the optical axis. This means that the object or image does not shift laterally, even with defocus; the distance between two object or image points is constant along the optical axis.

An optical system can be **telecentric** in

- Object space, where the entrance pupil is at infinity, and the aperture stop is in the back focal plane;
- Image space, where the exit pupil is at infinity, and the aperture stop is in the front focal plane; or
- Both (doubly telecentric), where the entrance and exit pupils are at infinity, and the aperture stop is at the center of the system, in the back focal plane of the element before the stop and front focal length of the element after the stop (afocal system).

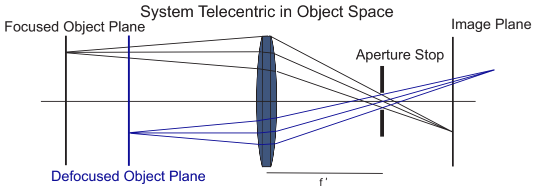
System Telecentric in Object Space
Focused Object Plane
Image Plane
Aperture Stop
Defocused Object Plane
f '

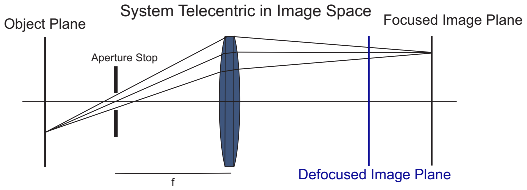
System Telecentric in Image Space
Object Plane
Focused Image Plane
Aperture Stop
Defocused Image Plane
f

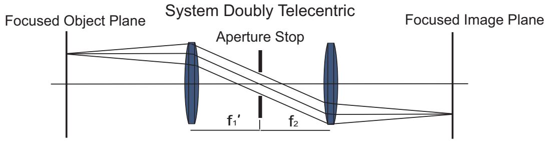
System Doubly Telecentric
Focused Object Plane
Focused Image Plane
Aperture Stop
f_1'
f_2

The aperture stop in a microscope is located at the back focal plane of the microscope objective. This makes the microscope objective telecentric in object space. Therefore, in microscopy, the object is observed with constant magnification, even for defocused object planes. This feature of microscopy systems significantly simplifies their operation and increases the reliability of image analysis.

Magnification of a Microscope

Magnifying power (MP) is defined as the ratio between the angles subtended by an object with and without magnification. The magnifying power (as defined for a single lens) creates an enlarged virtual image of an object. The angle of an object observed with magnification is

$$u' = \frac{h'}{z' - l} = \frac{h\left(f - z'\right)}{f\left(z' - l\right)}.$$

Therefore,

$$MP = \frac{u'}{u} = \frac{d_o\left(f - z'\right)}{f\left(z' - l\right)}.$$

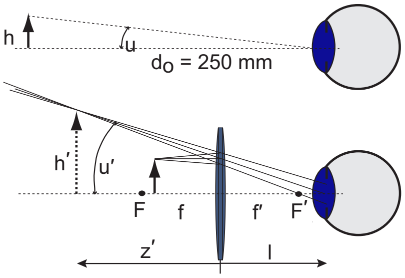

The angle for an unaided eye is defined for the **minimum focus distance** (d_o) of 10 inches or 250 mm, which is the distance that the object (real or virtual) may be examined without discomfort for the average population. A distance l between the lens and the eye is often small and can be assumed to equal zero:

$$MP \approx \frac{250\text{mm}}{f} - \frac{250\text{mm}}{z'}.$$

If the virtual image is at infinity (observed with a relaxed eye), $z' = -\infty$, and

$$MP \approx \frac{250\text{mm}}{f}.$$

The total magnifying power of the microscope results from the magnification of the microscope objective and the magnifying power of the eyepiece (usually 10×):

$$M_{\text{objective}} = \frac{OTL}{f_{\text{objective}}}$$

$$MP_{\text{microscope}} = M_{\text{objective}} MP_{\text{eyepiece}} = -\frac{OTL}{f_{\text{objective}}} \frac{250\text{mm}}{f_{\text{eyepiece}}}.$$

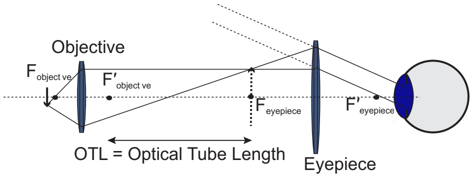

Numerical Aperture

The aperture diaphragm of the optical system determines the angle at which rays emerge from the axial object point and, after refraction, pass through the optical system. This acceptance angle is called the object space aperture angle. The parameter describing system throughput and this acceptance angle is called the **numerical aperture** (NA):

$$NA = n \sin u \, .$$

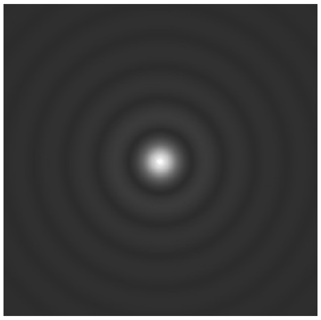

As seen from the equation, throughput of the optical system may be increased by using media with a high refractive index *n*, e.g., oil or water. This effectively decreases the refraction angles at the interfaces.

The dependence between the numerical aperture in the object space NA and the numerical aperture in the image space between the objective and the eyepiece NA′ is calculated using the objective magnification:

$$NA = NA' M_{\text{objective}} \, .$$

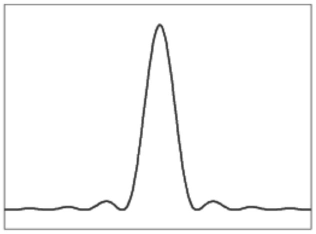

As a result of diffraction at the aperture of the optical system, self-luminous points of the object are not imaged as points but as so-called Airy disks. An **Airy disk** is a bright disk surrounded by concentric rings with gradually decreasing intensities. The disk diameter (where the intensity reaches the first zero) is

$$d = \frac{1.22\lambda}{n \sin u} = \frac{1.22\lambda}{NA} \, .$$

Media	Refractive Index
Air	1
Water	1.33
Oil	1.45–1.6 (1.515 is typical)

Note that the refractive index in the equation is for media between the object and the optical system.

Resolution Limit

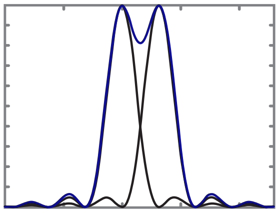

The lateral resolution of an optical system can be defined in terms of its ability to resolve images of two adjacent, self-luminous points. When two **Airy disks** are too close, they form a continuous intensity distribution and cannot be distinguished. The **Rayleigh resolution limit** is defined as occurring when the peak of the Airy pattern arising from one point coincides with the first minimum of the Airy pattern arising from a second point object. Such a distribution gives an intensity dip of 26%. The distance between the two points in this case is

$$d = \frac{0.61\lambda}{n \sin u} = \frac{0.61\lambda}{NA}.$$

The equation indicates that the resolution of an optical system improves with an increase in *NA* and decreases with increasing wavelength λ. For example, for $\lambda = 450$ nm (blue) and oil immersion $NA = 1.4$, the microscope objective can optically resolve points separated by less than 200 nm.

The situation in which the intensity dip between two adjacent self-luminous points becomes zero defines the **Sparrow resolution limit**. In this case $d = 0.5\lambda/NA$.

The **Abbe resolution limit** considers both the diffraction caused by the object and the *NA* of the optical system. It assumes that if at least two adjacent diffraction orders for points with spacing *d* are accepted by the objective, these two points can be resolved. Therefore, the resolution depends on both imaging and illumination apertures and is

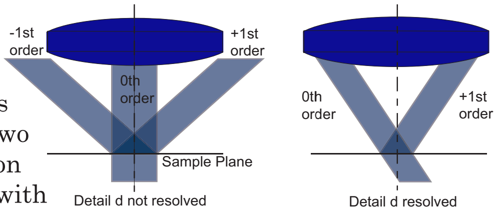

$$d = \frac{\lambda}{NA_{\text{objective}} + NA_{\text{condenser}}}.$$

Useful Magnification

For visual observation, the **angular resolving power** can be defined as 1.5 arc minutes. For unmagnified objects at a distance of 250 mm (the eye's minimum focus distance), 1.5 arc minutes converts to d_{eye} = 0.1 mm. Since microscope objectives by themselves do not usually provide sufficient magnification, they are combined with oculars or eyepieces. The resolving power of the microscope is then

$$d_{\text{mic}} = \frac{d_{\text{eye}}}{M_{\text{min}}} = \frac{d_{\text{eye}}}{M_{\text{obj}} M_{\text{eyepiece}}} .$$

In the Sparrow resolution limit, the minimum microscope magnification is

$$M_{\text{min}} = 2 d_{\text{eye}} NA / \lambda .$$

Therefore, a total minimum magnification M_{min} can be defined as approximately 250–500 NA (depending on wavelength). For lower magnification the image will appear brighter, but imaging is performed below the overall resolution limit of the microscope. For larger magnification, the contrast decreases and resolution does not improve. While a theoretically higher magnification should not provide additional information, it is useful to increase it to approximately 1000 NA to provide comfortable sampling of the object. Therefore, it is assumed that the **useful magnification** of a microscope is between 500 NA and 1000 NA. Usually, any magnification above 1000 NA is called **empty magnification**. The image size in such cases is enlarged, but no additional useful information is provided. The highest useful magnification of a microscope is approximately 1500 for an oil-immersion microscope objective with $NA = 1.5$.

Similar analysis can be performed for digital microscopy, which uses CCD or CMOS cameras as image sensors. Camera pixels are usually small (between 2 and 30 microns), and useful magnification must be estimated for a particular image sensor rather than the eye. Therefore, digital microscopy can work at lower magnification, and magnification of the microscope objective alone is usually sufficient.

Depth of Field and Depth of Focus

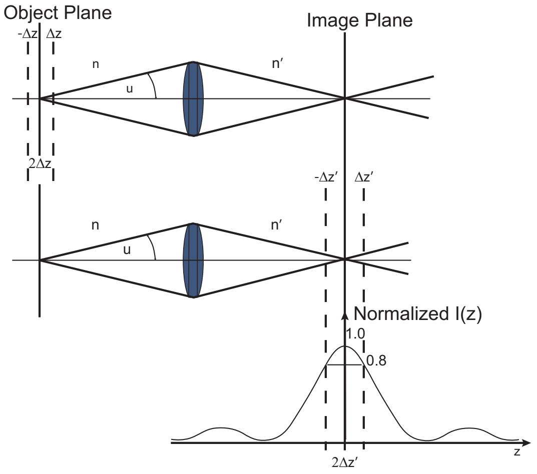
Object Plane
Image Plane
$-\Delta z$
Δz
n
n′
u
$2\Delta z$
$-\Delta z'$
$\Delta z'$
n
n′
u
Normalized I(z)
1.0
0.8
z
$2\Delta z'$

Two important terms are used to define axial resolution: **depth of field** and **depth of focus**. Depth of field refers to the object thickness for which an image is in focus, while depth of focus is the corresponding thickness of the image plane. In the case of a diffraction-limited optical system, the depth of focus is determined for an intensity drop along the optical axis to 80%, defined as

$$DOF = 2\Delta z' = \frac{n'\lambda}{NA'^2}.$$

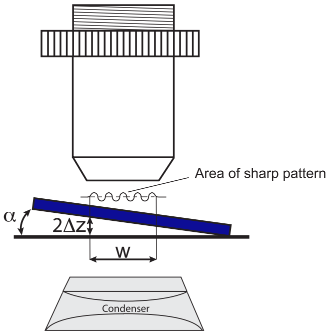
Area of sharp pattern
α
2Δz
w
Condenser

The relation between depth of field ($2\Delta z$) and depth of focus ($2\Delta z'$) incorporates the objective magnification:

$$2\Delta z' = M^2_{\text{objective}} 2\Delta z \frac{n'}{n},$$

where n' and n are medium refractive indexes in the object and image space, respectively.

To quickly measure the depth of field for a particular microscope objective, the grid amplitude structure can be placed on the microscope stage and tilted with the known angle α. The depth of field is determined after measuring the grid zone w while in focus:

$$2\Delta z = nw\tan\alpha\,.$$

Magnification and Frequency vs Depth of Field

Depth of field for visual observation is a function of the total microscope magnification and the NA of the objective, approximated as

$$2\Delta z = n\frac{0.5\lambda}{NA^2} + n\frac{340}{M_{\text{microscope}}NA}.$$

Note that estimated values do not include eye accommodation. The graph below presents depth of field for visual observation. Refractive index n of the object space was assumed to equal 1. For other media, values from the graph must be multiplied by an appropriate n.

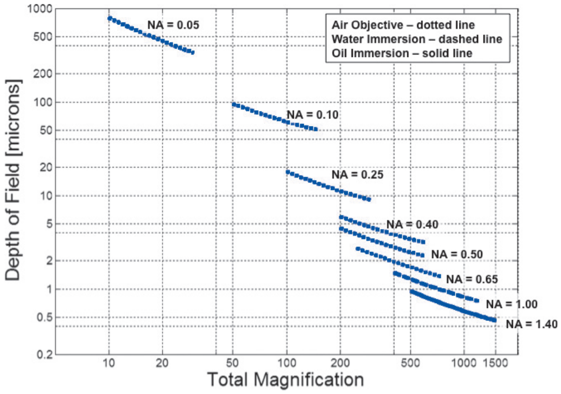

Depth of field can also be defined for a specific frequency present in the object because imaging contrast changes for a particular object frequency, as described by the modulation transfer function. The approximated equation is

$$2\Delta z = \frac{0.4}{\rho NA},$$

where ρ is the frequency in cycles per millimeter.

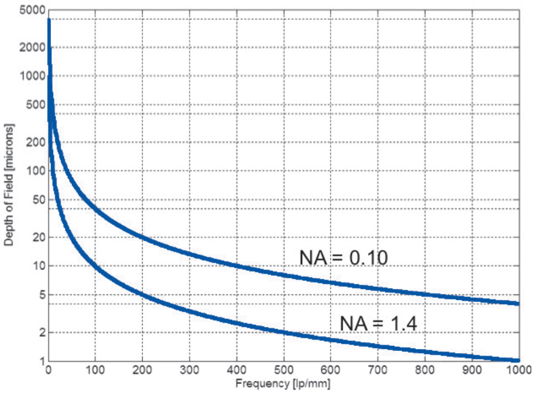

Köhler Illumination

One of the most critical elements in efficient microscope operation is proper set up of the illumination system. To assist with this, August Köhler introduced a solution that provides uniform and bright illumination over the field of view, even for light sources that are not uniform (e.g., a lamp filament). This system consists of a light source, a field-stop diaphragm, a condenser aperture, and collective and condenser lenses. This solution is now called **Köhler illumination** and is commonly used for a variety of imaging modes.

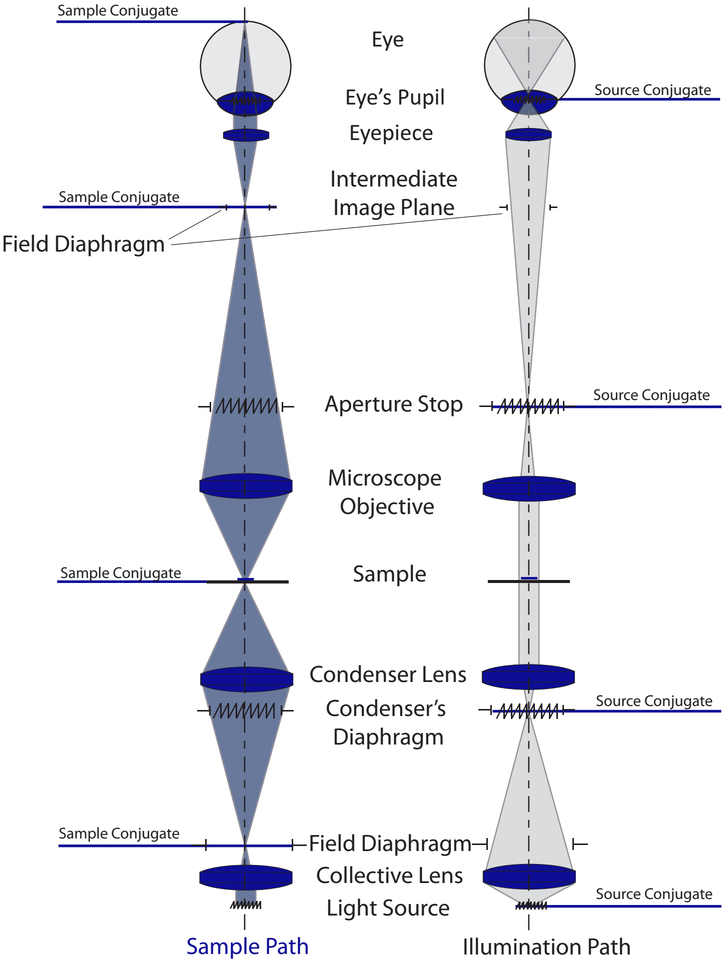
Sample Conjugate
Eye
Eye's Pupil
Source Conjugate
Eyepiece
Intermediate Image Plane
Sample Conjugate
Field Diaphragm
Aperture Stop
Source Conjugate
Microscope Objective
Sample Conjugate
Sample
Condenser Lens
Condenser's Diaphragm
Source Conjugate
Sample Conjugate
Field Diaphragm
Collective Lens
Light Source
Source Conjugate
Sample Path
Illumination Path

Köhler Illumination (cont.)

The illumination system is configured such that an image of the light source completely fills the condenser aperture.

Köhler illumination requires setting up the microscope system so that the field diaphragm, object plane, and intermediate image in the eyepiece's field stop, retina, or CCD are in conjugate planes. Similarly, the lamp filament, front aperture of the condenser, microscope objective's back focal plane (aperture stop), exit pupil of the eyepiece, and the eye's pupil are also at conjugate planes. Köhler illumination can be set up for both the transmission mode (also called **DIA**) and the reflectance mode (also called **EPI**).

The uniformity of the illumination is the result of having the light source at infinity with respect to the illuminated sample.

EPI illumination is especially useful for the biological and metallurgical imaging of thick or opaque samples.

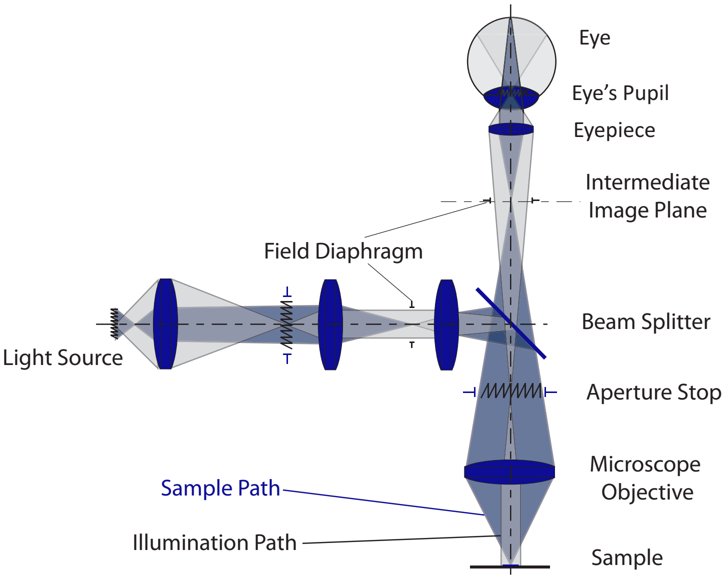
Eye
Eye's Pupil
Eyepiece
Intermediate Image Plane
Field Diaphragm
Beam Splitter
Light Source
Aperture Stop
Microscope Objective
Sample Path
Illumination Path
Sample

Alignment of Köhler Illumination

The procedure for **Köhler illumination** alignment consists of the following steps:

1. Place the sample on the microscope stage and move it so that the front surface of the condenser lens is 1–2 mm from the microscope slide. The condenser and collective diaphragms should be wide open.
2. Adjust the location of the source so illumination fills the condenser's diaphragm. The sharp image of the lamp filament should be visible at the condenser's diaphragm plane and at the back focal plane of the microscope objective. To see the filament at the back focal plane of the objective, a **Bertrand lens** can be applied (also see *Special Lens Components* on page 55). The filament should be centered in the aperture stop using the bulb's adjustment knobs on the illuminator.
3. For a low-power objective, bring the sample into focus. Since a microscope objective works at a parfocal distance, switching to a higher magnification later is easy and requires few adjustments.
4. Focus and center the condenser lens. Close down the field diaphragm, focus down the outline of the diaphragm, and adjust the condenser's position. After adjusting the *x*-, *y*-, and *z*-axis, open the field diaphragm so it accommodates the entire field of view.
5. Adjust the position of the condenser's diaphragm by observing the back focal objective plane through the Bertrand lens. When the edges of the aperture are sharply seen, the condenser's diaphragm should be closed to approximately three quarters of the objective's aperture.
6. Tune the brightness of the lamp. This adjustment should be performed by regulating the voltage of the lamp's power supply or, preferably, through the neutral density filters in the beam path. Do not adjust the brightness by closing the condenser's diaphragm because it affects the illumination setup and the overall microscope resolution.

Note that while three quarters of the aperture stop is recommended for initial illumination, adjusting the aperture of the illumination system affects the resolution of the microscope. Therefore, the final setting should be adjusted after examining the images.

Critical Illumination

An alternative to Köhler illumination is **critical illumination**, which is based on imaging the light source directly onto the sample. This type of illumination requires a highly uniform light source. Any source non-uniformities will result in intensity variations across the image. Its major advantage is high efficiency, since it can collect a larger solid angle than Köhler illumination and therefore provides a higher energy density at the sample. For parabolic or elliptical reflective collectors, critical illumination can utilize up to 85% of the light emitted by the source.

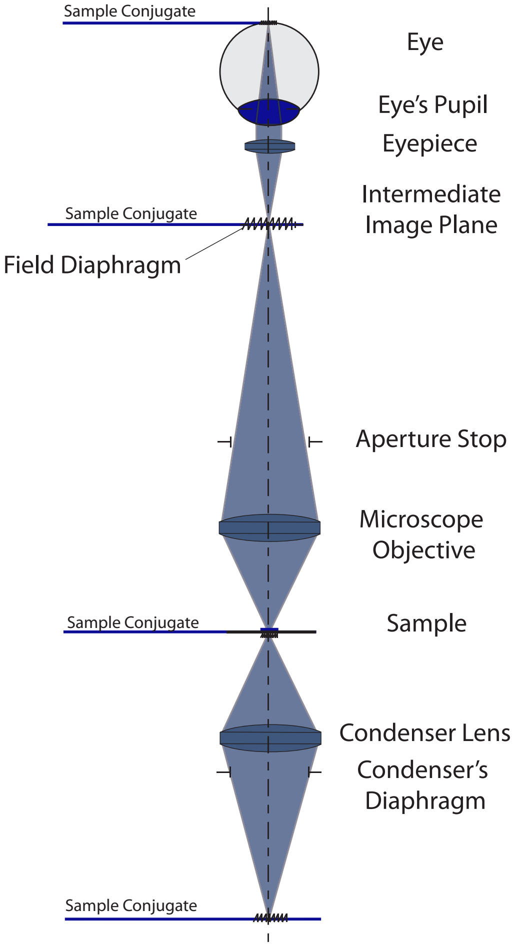
Sample Conjugate
Eye
Eye's Pupil
Eyepiece
Sample Conjugate
Intermediate Image Plane
Field Diaphragm
Aperture Stop
Microscope Objective
Sample Conjugate
Sample
Condenser Lens
Condenser's Diaphragm
Sample Conjugate

Stereo Microscopes

Stereo microscopes are built to provide depth perception, which is important for applications like micro-assembly and biological and surgical imaging. Two primary stereo-microscope approaches involve building two separate tilted systems, or using a common objective combined with a binocular system.

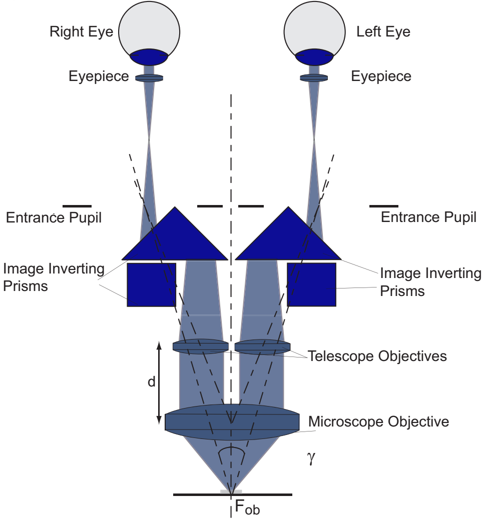
Right Eye
Left Eye
Eyepiece
Eyepiece
Entrance Pupil
Entrance Pupil
Image Inverting Prisms
Image Inverting Prisms
Telescope Objectives
d
Microscope Objective
γ
F_{ob}

In the latter, the angle of convergence γ of a stereo microscope depends on the focal length of a microscope objective and a distance *d* between the microscope objective and telescope objectives. The entrance pupils of the microscope are images of the stops (located at the plane of the telescope objectives) through the microscope objective.

Depth perception δz can be defined as

$$\delta z = \frac{250\alpha_s}{M_{\text{microscope}} \tan\gamma} [\text{mm}],$$

where α_s is a **visual stereo resolving power**, which for daylight vision is approximately 5–10 arc seconds.

Stereo microscopes have a convergence angle γ in the range of 10–15 deg. Note that γ is 15 deg for visual observation and 0 for a standard microscope. For example, depth perception for the human eye is approximately 0.05 mm (at 250 mm), while for a stereo microscope of $M_{\text{microscope}} = 100$ and $\gamma = 15$ deg, it is $\delta z = 0.5$ μm.

Eyepieces

The eyepiece relays an intermediate image into the eye, corrects for some remaining objective aberrations, and provides measurement capabilities. It also needs to relay an exit pupil of a microscope objective onto the entrance pupil of the eye. The location of the relayed exit pupil is called the **eye point** and needs to be in a comfortable observation distance behind the eyepiece. The clearance between the mechanical mounting of the eyepiece and the eye point is called **eye relief**. Typical eye relief is 7–12 mm.

An eyepiece usually consists of two lenses: one (closer to the eye) which magnifies the image, and a second working as a collective lens and also responsible for the location of the exit pupil of the microscope. An eyepiece contains a field stop that provides a sharp image edge.

Parameters like **magnifying power** of an eyepiece and **field number** (FN), i.e., field of view of an eyepiece, are engraved on an eyepiece's barrel: M × / FN. Field number and magnification of a microscope objective allow quick calculation of the imaged area of a sample (see also *The Finite Tube Length Microscope* on p. 34). Field number varies with microscopy vendors and eyepiece magnification. For 10× or lower magnification eyepieces, it is usually 20–28 mm while for higher magnification, wide-angle oculars can get down to approximately 5 mm.

The majority of eyepieces are **Huygens**, **Ramsden**, or derivations of them. The Huygens eyepiece consists of two plano-convex lenses with convex surfaces facing the microscope objective:

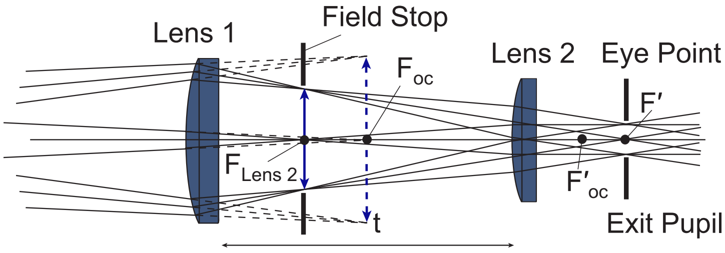

Eyepieces (cont.)

Both lenses are usually made with crown glass and allow for some correction of lateral color, coma, and astigmatism. To correct for color, the following conditions should be met:

$$f_1 \approx f_2 \quad \text{and} \quad t \approx 1.5 f_2 .$$

For higher magnifications, the exit pupil moves toward the ocular, making observation less convenient, and so this eyepiece is used only for low magnifications (≤10×). In addition, Huygens oculars effectively correct for chromatic aberrations, and therefore can be more effectively used with lower-end microscope objectives (e.g., achromatic).

The Ramsden eyepiece consists of two plano-convex lenses with convex surfaces facing each other. Both focal lengths are very similar, and the distance between lenses is smaller than f_2:

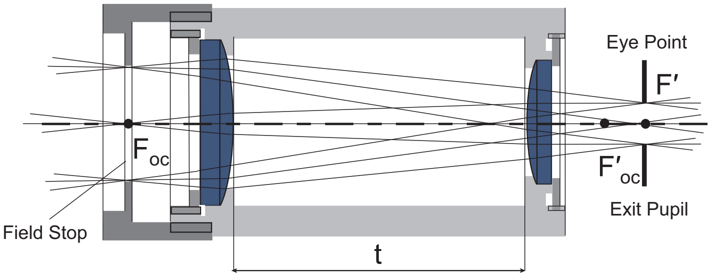

The field stop is placed in the front focal plane of the eyepiece. A collective lens does not participate in creating an intermediate image, so the Ramsden eyepiece works as a simple magnifier:

$$f_1 = f_2 \quad \text{and} \quad t < f_2 .$$

Compensating eyepieces work in conjunction with microscope objectives to correct for lateral color (apochromatic).

High-eye point oculars provide an extended distance between the last mechanical surface and the exit pupil of the eyepiece. They allow users with glasses to comfortably use a microscope. The convenient high eye-point location should be 20–25 mm behind the eyepiece.

Nomenclature and Marking of Objectives

Objective parameters include:

Objective correction—such as Achromat, Plan Achromat, Fluor, Plan Fluor, Apochromat (Apo), and Plan Apochromat.

Magnification—the lens magnification for a finite tube length, or for a microscope objective in combination with a tube lens. In infinity-corrected systems, magnification depends on the ratio between the focal lengths of a tube lens and a microscope objective. A particular objective type should be used only in combination with the proper tube lens.

Magnification	Zeiss Code
1×, 1.25×	Black
2.5×	Khaki
4×, 5×	Red
6.3×	Orange
10×	Yellow
16×, 20×, 25×, 32×	Green
25×, 32×	Green
40×, 50×	Light Blue
63×	Dark Blue
> 100×	White

Application—specialized use or design of an objective, e.g., H (bright field), D (dark field), DIC (differential interference contrast), RL (reflected light), PH (phase contrast), or P (polarization).

Tube length—an infinity corrected (∞) or finite tube length in mm.

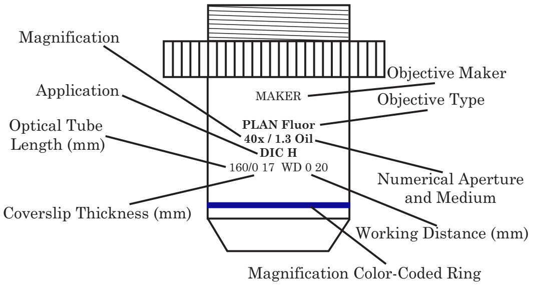
Magnification
Application
Optical Tube Length (mm)
Coverslip Thickness (mm)
Objective Maker
Objective Type
MAKER
PLAN Fluor
40x / 1.3 Oil
DIC H
160/0 17 WD 0 20
Numerical Aperture and Medium
Working Distance (mm)
Magnification Color-Coded Ring

Cover slip—the thickness of the cover slip used (in mm). "0" or "–" means no cover glass or the cover slip is optional, respectively.

Numerical aperture (NA) and medium—defines system throughput and resolution. It depends on the media between the sample and objective. The most common media are air (no marking), oil (Oil), water (W), or Glycerine.

Working distance (WD)—the distance in millimeters between the surface of the front objective lens and the object.

Objective Designs

Achromatic objectives (also called **achromats)** are corrected at two wavelengths: 656 nm and 486 nm. Also, spherical aberration and coma are corrected for green (546 nm), while astigmatism is very low. Their major problems are a secondary spectrum and field curvature.

Achromatic objectives are usually built with a lower NA (below 0.5) and magnification (below 40×). They work well for white light illumination or single wavelength use. When corrected for field curvature, they are called plan-achromats.

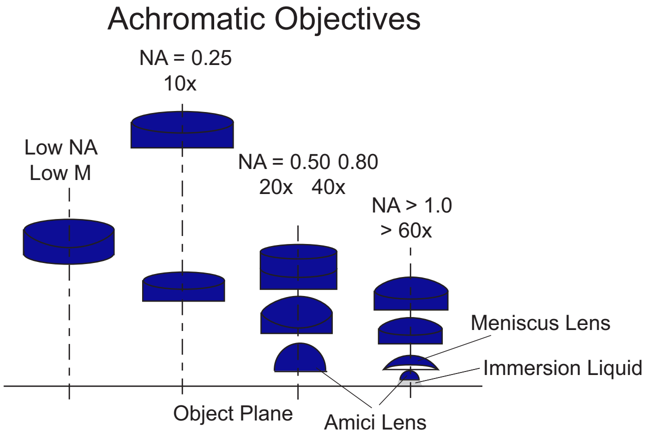

Fluorites or **semi-apochromats** have similar color correction as **achromats**; however, they correct for spherical aberration for two or three colors. The name "fluorites" was assigned to this type of objective due to the materials originally used to build this type of objective. They can be applied for higher NA (e.g., 1.3) and magnifications, and used for applications like immunofluorescence, polarization, or differential interference contrast microscopy.

The most advanced microscope objectives are **apochromats**, which are usually chromatically corrected for three colors, with spherical aberration correction for at least two colors. They are similar in construction to fluorites but with different thicknesses and surface figures. With the correction of field curvature, they are called plan-apochromats. They are used in fluorescence microscopy with multiple dyes and can provide very high NA (1.4). Therefore, they are suitable for low-light applications.

Objective Designs (cont.)

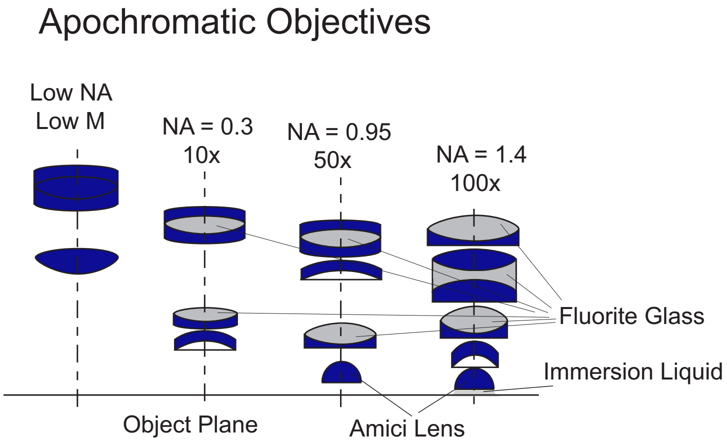
Apochromatic Objectives
Low NA
Low M
NA = 0.3
10x
NA = 0.95
50x
NA = 1.4
100x
Fluorite Glass
Immersion Liquid
Object Plane
Amici Lens

Type	Number of wavelengths for spherical correction	Number of colors for chromatic correction
Achromat	1	2
Fluorite	2–3	2–3
Plan-Fluorite	2–4	2–4
Plan-Apochromat	2–4	3–5

(Adapted from Murphy, 2001 and http://www.microscopyu.com/)

Examples of typical objective parameters are shown in the table below:

M	Type	Medium	*WD*	NA	*d*	DOF
10	Achromat	Air	4.4	0.25	1.34	8.80
20	Achromat	Air	0.53	0.45	0.75	2.72
40	Fluorite	Air	0.50	0.75	0.45	0.98
40	Fluorite	Oil	0.20	1.30	0.26	0.49
60	Apochromat	Air	0.15	0.95	0.35	0.61
60	Apochromat	Oil	0.09	1.40	0.24	0.43
100	Apochromat	Oil	0.09	1.40	0.24	0.43
Refractive index of oil is $n = 1.515$						

(Adapted from Murphy, 2001)

Special Objectives and Features

Special types of objectives include long working distance objectives, ultra-low-magnification objectives, water-immersion objectives, and UV lenses.

Long working distance objectives allow imaging through thick substrates like culture dishes. They are also developed for interferometric applications in Michelson and Mirau configurations. Alternatively, they can allow for the placement of instrumentation (e.g., micropipettes) between a sample and an objective. To provide a long working distance and high NA, a common solution is to use reflective objectives, or extend the working distance of a standard microscope objective by using a reflective attachment. While reflective objectives have the advantage of being free of chromatic aberrations, their serious drawback is a smaller field of view relative to refractive objectives. The central obscuration also decreases throughput by 15%–20%.

LWD Reflective Objective

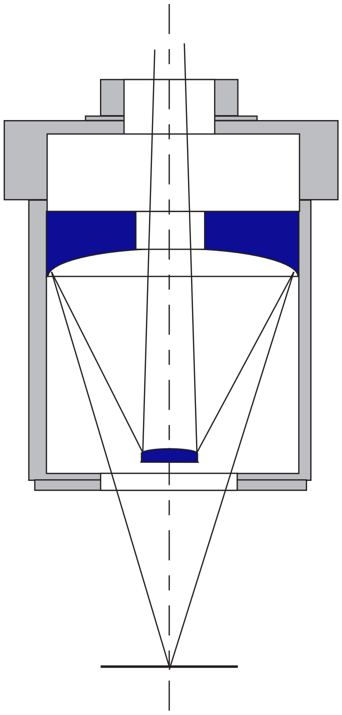

Special Objectives and Features (cont.)

Low magnification objectives can achieve magnifications as low as 0.5. However, in some cases they are not fully compatible with microscopy systems, and may not be telecentric in the image space. They also often require special tube lenses and special condensers to provide Köhler illumination.

Water immersion objectives are increasingly common, especially for biological imaging, because they provide a high NA and avoid toxic immersion oils. They usually work without a cover slip.

UV objectives are made using UV transparent, low-dispersion materials such as quartz. These objectives enable imaging at wavelengths from the near-UV through the visible spectrum, e.g., 240–700 nm. Reflective objectives can also be used for the IR bands.

Reflective Adapter to extend working distance WD

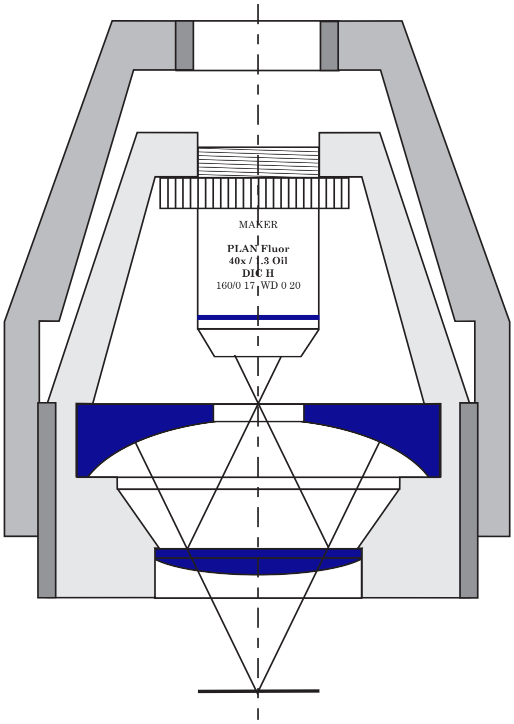

Special Lens Components

The **Bertrand lens** is a focusable eyepiece telescope that can be easily placed in the light path of the microscope. This lens is used to view the back aperture of the objective, which simplifies microscope alignment, specifically when setting up Köhler illumination.

To construct a high-numerical-aperture microscope objective with good correction, numerous optical designs implement the **Amici lens** as a first component of the objective. It is a thick lens placed in close proximity to the sample. An Amici lens usually has a plane (or nearly plane) first surface and a large-curvature spherical second surface. To ensure good chromatic aberration correction, an achromatic lens is located closely behind the Amici lens. In such configurations, microscope objectives usually reach an NA of 0.5–0.7.

To further increase the NA, an Amici lens is improved by either cementing a high-refractive-index meniscus lens to it or placing a meniscus lens closely behind the Amici lens. This makes it possible to construct well-corrected high magnification (100×) and high numerical aperture (NA > 1.0) oil-immersion objective lenses.

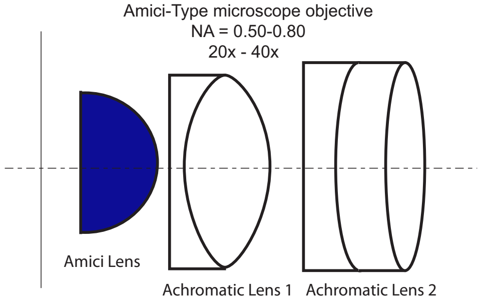
Amici-Type microscope objective
NA = 0.50-0.80
20x - 40x
Amici Lens
Achromatic Lens 1
Achromatic Lens 2

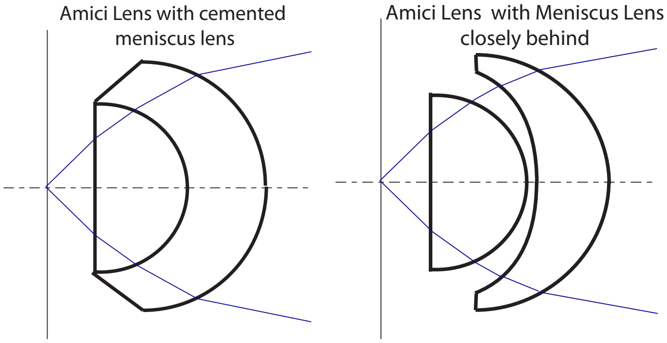
Amici Lens with cemented meniscus lens
Amici Lens with Meniscus Lens closely behind

Cover Glass and Immersion

The **cover slip** is located between the object and the microscope objective. It protects the imaged sample and is an important element in the optical path of the microscope. Cover glass can reduce imaging performance and cause spherical aberration, while different imaging angles experience movement of the object point along the optical axis toward the microscope objective. The object point moves closer to the objective with an increase in angle.

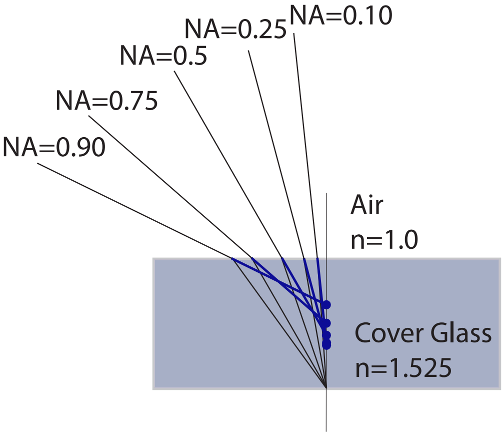

The importance of the cover glass increases in proportion to the objective's NA, especially in high-NA dry lenses. For lenses with an NA smaller than 0.5, the type of cover glass may not be a critical parameter, but should always be properly used to optimize imaging quality. Cover glasses are likely to be encountered when imaging biological samples. Note that the presence of a standard-thickness cover glass is accounted for in the design of high-NA objectives. The important parameters that define a cover glass are its thickness, index of refraction, and Abbe number. The ISO standards for refractive index and Abbe number are $n = 1.5255 \pm 0.0015$ and $V = 56 \pm 2$, respectively.

Microscope objectives are available that can accommodate a range of cover-slip thicknesses. An adjustable collar allows the user to adjust for cover slip thickness in the range from 100 microns to over 200 microns.

Cover Glass Grade Number	Thickness (in microns)
0	83–130
1	130–160
1.5	160–190
2	190–250

Cover Glass and Immersion (cont.)

The table below presents a summary of acceptable cover glass thickness deviations from 0.17 mm and the allowed thicknesses for Zeiss air objectives with different NAs.

NA of the Objective	Allowed Thickness Deviations (from 0.17 mm)	Allowed Thickness Range
<0.3	–	0.000–0.300
0.30–0.45	0.07	0.100–0.240
0.45–0.55	0.05	0.120–0.220
0.55–0.65	0.03	0.140–0.200
0.65–0.75	0.02	0.150–0.190
0.75–0.85	0.01	0.160–0.180
0.85–0.95	0.005	0.165–0.175

(Adapted from Pluta, 1988)

To increase both the microscope's resolution and system throughput, immersion liquids are used between the cover-slip glass and the microscope objective, or directly between the sample and objective. An **immersion liquid** effectively increases the NA and decreases the angle of refraction at the interface between the medium and the microscope objective. Common immersion liquids include oil (n = 1.515), water (n = 1.34), and glycerin (n = 1.48).

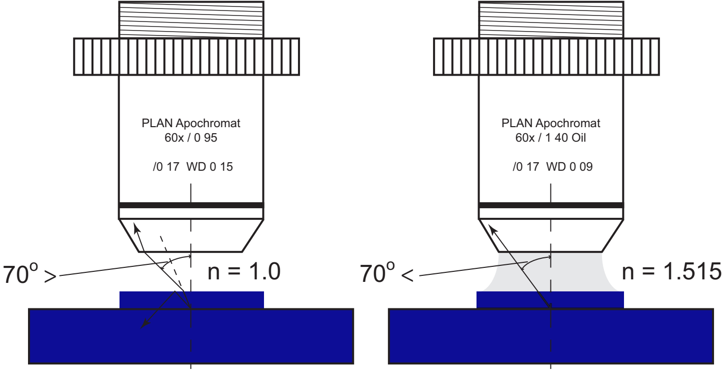

Water, which is more common, or glycerin immersion objectives are mainly used for biological samples, such as living cells or a tissue culture. They provide a slightly lower NA than oil objectives but are free of the toxicity associated with immersion oils. Water immersion is usually applied without a cover slip. For fluorescent applications it is also critical to use non-fluorescent oil.

Common Light Sources for Microscopy

Common light sources for microscopy include **incandescent lamps**, such as tungsten-argon and tungsten (e.g., quartz halogen) lamps. A **tungsten-argon lamp** is primarily used for bright-field, phase-contrast, and some polarization imaging. **Halogen lamps** are inexpensive and a convenient choice for a variety of applications that require a continuous and bright spectrum.

Arc lamps are usually brighter than incandescent lamps. The most common examples include xenon (XBO) and mercury (HBO) arc lamps. Arc lamps provide high-quality monochromatic illumination if combined with the appropriate filter. Arc lamps are more difficult to align, are more expensive, and have a shorter lifetime. Their spectral range starts in the UV range and continuously extends through visible to the infrared. About 20% of the output is in the visible spectrum, while the majority is in the UV and IR. The usual lifetime is 100–200 hours.

Common Mercury Lines	
Wavelength [nm]	**Color**
254	Far UV
366	Near UV
405	Violet
435	Deep Blue
546	Yellow-green
578	Yellow doublet

Another light source is the **gas-arc discharge lamp**, which includes mercury, xenon, and halide lamps. The **mercury lamp** has several strong lines, which might be 100 times brighter than the average output. About 50% is located in the UV range, and should be used with protective glasses. For imaging biological samples, the proper selection of filters is necessary to protect living cell samples and micro-organisms (e.g., UV-blocking filters/cold mirrors). The xenon arc lamp has a uniform spectrum, can provide an output power greater than 100 W, and is often used for fluorescence imaging. However, over 50% of its power falls into the IR; therefore, IR-blocking filters (hot mirrors) are necessary to prevent the overheating of samples.

Metal halide lamps were recently introduced as high-power sources (over 150 W) with lifetimes several times longer than arc lamps. While in general the metal-halide lamp has a spectral output similar to that of a mercury arc lamp, it extends further into longer wavelengths.

LED Light Sources

Light-emitting diodes (LEDs) are a new, alternative light source for microscopy applications. The LED is a semiconductor diode that emits photons when in forward-biased mode. Electrons pass through the depletion region of a p-n junction and lose an amount of energy equivalent to the bandgap of the semiconductor. The characteristic features of LEDs include a long lifetime, a compact design, and high efficiency. They also emit narrowband light with relatively high energy.

Wavelengths [nm] of High-power LEDs Commonly Used in Microscopy		**Total Beam Power [mW] (approximate)**
455	Royal Blue	225–450
470	Blue	200–400
505	Cyan	150–250
530	Green	100–175
590	Amber	15–25
633	Red	25–50
435–675	White Light	200–300

An important feature of LEDs is the ability to combine them into arrays and custom geometries. Also, LEDs operate at lower temperatures than arc lamps, and due to their compact design they can be cooled easily with simple heat sinks and fans.

The radiance of currently available LEDs is still significantly lower than that possible with arc lamps; however, LEDs can produce an acceptable fluorescent signal in bright microscopy applications. Also, the pulsed mode can be used to increase the radiance by 20 times or more.

LED Spectral Range [nm]	**Semiconductor**
350–400	GaN
400–550	$In_{1-x}Ga_xN$
550–650	$Al_{1-x-y}In_yGa_xP$
650–750	$Al_{1-x}Ga_xAs$
750–1000	$GaAs_{1-x}P_x$

Filters

Neutral density (ND) filters are neutral (wavelength wise) gray filters scaled in units of **optical density** (OD):

$$OD = \log_{10}\left(\frac{1}{\tau}\right),$$

where τ is the transmittance. ND filters can be combined to provide OD as a sum of the ODs of the individual filters. ND filters are used to change light intensity without tuning the light source, which can result in a spectral shift.

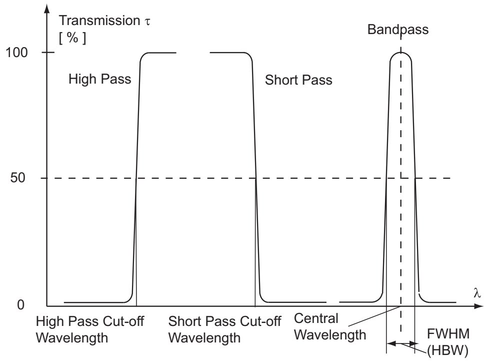

Color **absorption filters** and **interference filters** (also see *Multiple Wave Interference* on page 16) isolate the desired range of wavelengths: e.g. bandpass and edge filters.

Edge filters include short-pass and long-pass filters. Short-pass filters allow short wavelengths to pass and stop long wavelengths, and long-pass filters allow long wavelengths to pass while stopping short wavelengths. Edge filters are defined for wavelengths with a 50% drop in transmission.

Bandpass filters allow only a selected spectral bandwidth to pass and are characterized with a central wavelength and full-width-half-maximum (FWHM) defining the spectral range for transmission of at least 50%.

Color absorption glass filters usually serve as broad bandpass filters. They are less costly and less susceptible to damage than interference filters.

Interference filters are based on multiple beam interference in thin films. They combine between three to over 20 dielectric layers of $\lambda/2$ and $\lambda/4$, separated by metallic coatings. They can provide a sharp bandpass transmission down to the sub-nm range and full-width-half-maximum of 10–20 nm.

Polarizers and Polarization Prisms

Polarizers are built with birefringent crystals using polarization at multiple reflections, selective absorption (dichroism), form birefringance, or scattering.

For example, polarizers built with birefringent crystals (Glan-Thompson prisms) use the principle of total internal reflection to eliminate ordinary or extraordinary components (for positive or negative crystals).

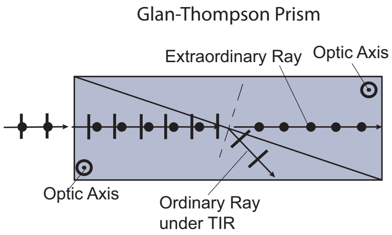
Glan-Thompson Prism
Extraordinary Ray
Optic Axis
Optic Axis
Ordinary Ray under TIR

Birefringent prisms are crucial components for numerous microscopy techniques, e.g., differential interference contrast. The most common birefringent prism is a **Wollaston prism**. It splits light into two beams with orthogonal polarization and propagates under different angles.

The angle between propagating beams is

$$\varepsilon = 2\left|n_e - n_o\right|\tan\alpha\ .$$

Both beams produce interference with fringe period *b*:

$$b = \frac{\lambda}{\varepsilon} = \frac{\lambda}{2\left|n_e - n_o\right|\tan\alpha}\ .$$

The localization plane of fringes is

$$\gamma = \frac{1}{2}\left(\frac{1}{n_e} + \frac{1}{n_o}\right)\alpha\ .$$

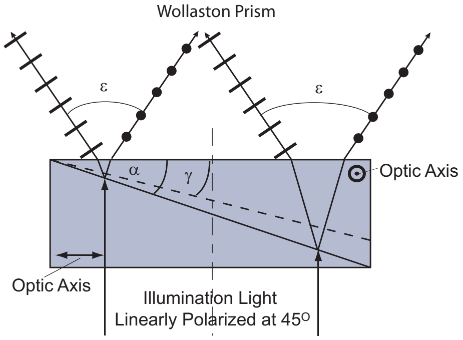
Wollaston Prism
ε
ε
α
γ
Optic Axis
Optic Axis
Illumination Light
Linearly Polarized at 45°

Polarizers and Polarization Prisms (cont.)

The fringe localization plane tilt can be compensated by using two symmetrical Wollaston prisms.

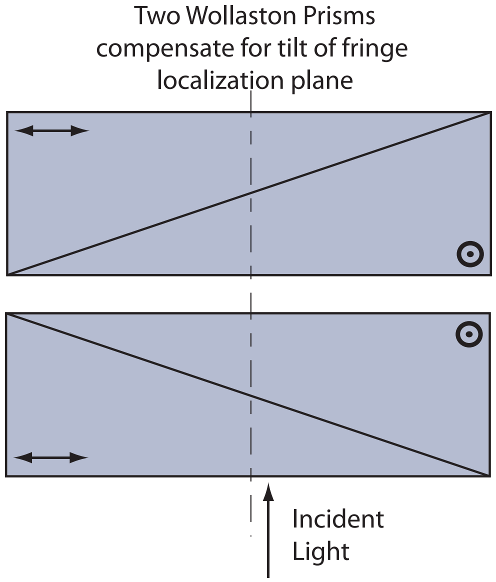
Two Wollaston Prisms compensate for tilt of fringe localization plane
Incident Light

Wollaston prisms have a fringe localization plane inside the prism. One example of a modified Wollaston prism is the **Nomarski prism**, which simplifies the DIC microscope setup by shifting the plane outside the prism, i.e., the prism does not need to be physically located at the condenser front focal plane or the objective's back focal plane.

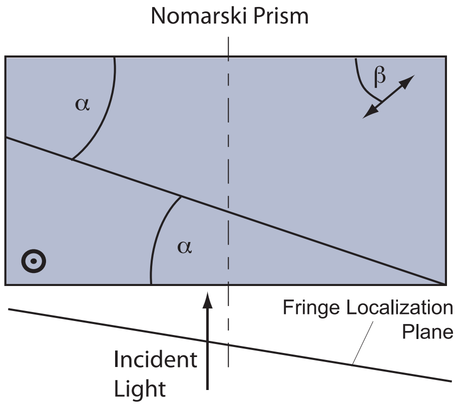
Nomarski Prism
α
β
α
Fringe Localization Plane
Incident Light

Amplitude and Phase Objects

The major object types encountered on the microscope are **amplitude** and **phase** objects. The type of object often determines the microscopy technique selected for imaging.

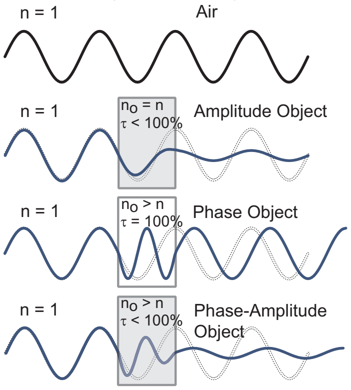

The amplitude object is defined as one that changes the amplitude and therefore the intensity of transmitted or reflected light. Such objects are usually imaged with bright-field microscopes. A stained tissue slice is a common amplitude object.

Phase objects do not affect the optical intensity; instead, they generate a phase shift in the transmitted or reflected light. This phase shift usually arises from an inhomogeneous refractive index distribution throughout the sample, causing differences in optical path length (OPL). Mixed amplitude-phase objects are also possible (e.g., biological media), which affect the amplitude and phase of illumination in different proportions.

Classifying objects as **self-luminous** and **non-self-luminous** is another way to categorize samples. Self-luminous objects generally do not directly relate their amplitude and phase to the illuminating light. This means that while the observed intensity may be proportional to the illumination intensity, its amplitude and phase are described by a statistical distribution (e.g., fluorescent samples). In such cases, one can treat discrete object points as secondary light sources, each with their own amplitude, phase, coherence, and wavelength properties.

Non-self-luminous objects are those that affect the illuminated beam in such a manner that discrete object points cannot be considered as entirely independent. In such cases, the wavelength and temporal coherence of the illuminating source needs to be considered in imaging. A diffusive or absorptive sample is an example of such an object.

The Selection of a Microscopy Technique

Microscopy provides several imaging principles. Below is a list of the most common techniques and object types:

Technique	Type of sample
Bright-field	Amplitude specimens, reflecting specimens, diffuse objects
Dark-field	Light-scattering objects
Phase contrast	Phase objects, light-scattering objects, light-refracting objects, reflective specimens
Differential interference contrast (DIC)	Phase objects, light-scattering objects, light-refracting objects, reflective specimens
Polarization microscopy	Birefringent specimens
Fluorescence microscopy	Fluorescent specimens
Laser scanning, Confocal microscopy, and Multi-photon microscopy	3D samples requiring optical sectioning, fluorescent and scattering samples
Super-resolution microscopy (**RESOLFT**, **4Pi**, **I^{5}M**, **SI**, **STORM**, **PALM**, and others)	Imaging at the molecular level; imaging primarily focuses on fluorescent samples where the sample is a part of an imaging system
Raman microscopy, **CARS**	Contrast-free chemical imaging
Array microscopy	Imaging of large FOVs
SPIM	Imaging of large 3D samples
Interference Microscopy	Topography, refractive index measurements, 3D coherence imaging

All of these techniques can be considered for transmitted or reflected light. Below are examples of different sample types:

Sample type	Sample Example
Amplitude specimens	Naturally colored specimens, stained tissue
Specular specimens	Mirrors, thin films, metallurgical samples, integrated circuits
Diffuse objects	Diatoms, fibers, hairs, micro-organisms, minerals, insects
Phase objects	Bacteria, cells, fibers, mites, protozoa
Light-refracting samples	Colloidal suspensions, minerals, powders
Birefringent specimens	Mineral sections, crystallized chemicals, liquid crystals, fibers, single crystals
Fluorescent specimens	Cells in tissue culture, fluorochrome-stained sections, smears and spreads

(Adapted from www.microscopyu.com)

Image Comparison

The images below display four important microscope modalities: **bright field**, **dark field**, **phase contrast**, and **differential interference contrast**. They also demonstrate the characteristic features of these methods. The images are of a blood specimen viewed with the Zeiss upright Axiovert Observer Z1 microscope. The microscope was set with an objective at 40×, NA = 0.6, Ph2, LD Plan Neofluor, and a cover slip glass of 0.17 mm. The pictures were taken with a monochromatic CCD camera.

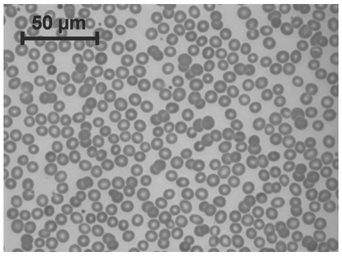 Bright Field	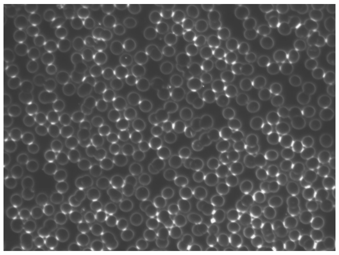 Dark Field
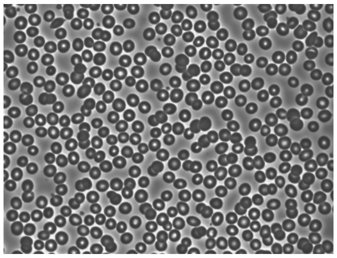 Phase Contrast	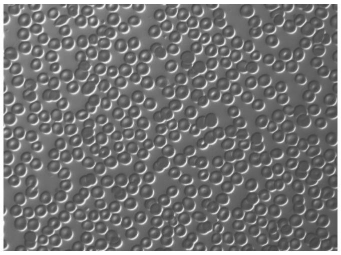 Differential Interference Contrast (DIC)

The bright-field image relies on absorption and shows the sample features with decreasing amounts of passing light. The dark-field image only shows the scattering sample components. Both the phase contrast and the differential interference contrast demonstrate the optical thickness of the sample. The characteristic halo effect is visible in the phase-contrast image. The 3D effect of the DIC image arises from the differential character of images; they are formed as a derivative of the phase changes in the beam as it passes through the sample.

Phase Contrast

Phase contrast is a technique used to visualize phase objects by phase and amplitude modifications between the direct beam propagating through the microscope and the beam diffracted at the phase object. An object is illuminated with monochromatic light, and a phase-delaying (or advancing) element in the aperture stop of the objective introduces a phase shift, which provides interference contrast. Changing the amplitude ratio of the diffracted and non-diffracted light can also increase the contrast of the object features.

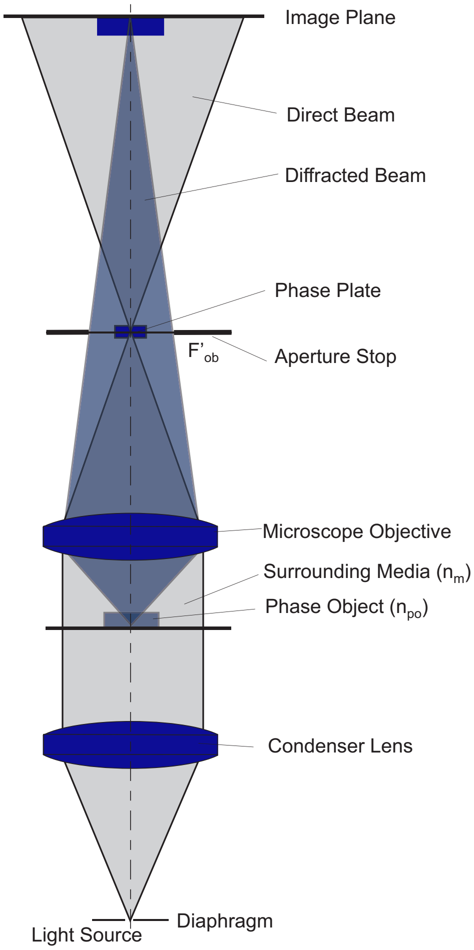
Image Plane
Direct Beam
Diffracted Beam
Phase Plate
F'_{ob}
Aperture Stop
Microscope Objective
Surrounding Media (n_m)
Phase Object (n_{po})
Condenser Lens
Diaphragm
Light Source

Phase Contrast (cont.)

Phase contrast is primarily used for imaging living biological samples (immersed in a solution), unstained samples (e.g., cells), thin film samples, and mild phase changes from mineral objects. In that regard, it provides qualitative information about the optical thickness of the sample. Phase contrast can also be used for some quantitative measurements, like an estimation of the refractive index or the thickness of thin layers. Its phase detection limit is similar to the **differential interference contrast** and is approximately $\lambda/100$. On the other hand, phase contrast (contrary to DIC) is limited to thin specimens, and phase delay should not exceed the depth of field of an objective. Other drawbacks compared to DIC involve its limited optical sectioning ability and undesired effects like halos or shading-off. (See also *Characteristic Features of Phase Contrast Microscopy* on page 71).

The most important advantages of phase contrast (over DIC) include its ability to image birefringent samples and its simple and inexpensive implementation into standard microscopy systems.

Presented below is a mathematical description of the phase contrast technique based on a vector approach. The phase shift in the figure (See page 68) is represented by the orientation of the vector. The length of the vector is proportional to amplitude of the beam. When using standard imaging on a transparent sample, the length of the light vectors passing through sample PO and surrounding media SM is the same, which makes the sample invisible. Additionally, vector PO can be considered as a sum of the vectors passing through surrounding media SM and diffracted at the object DP.

$$PO = SM + DP$$

If the wavefront propagating through the surrounding media can be the subject of an exclusive phase change (diffracted light DP is not affected), the vector SM is rotated by an angle corresponding to the phase change. This exclusive phase shift is obtained with a small circular or ring phase plate located in the plane of the aperture stop of a microscope.

Phase Contrast (cont.)

Consequently, vector *PO*, which represents light passing through a phase sample, will change its value to *PO′* and provide contrast in the image:

$$PO' = SM' + DP,$$

where *SM′* represents the rotated vector *SM*.

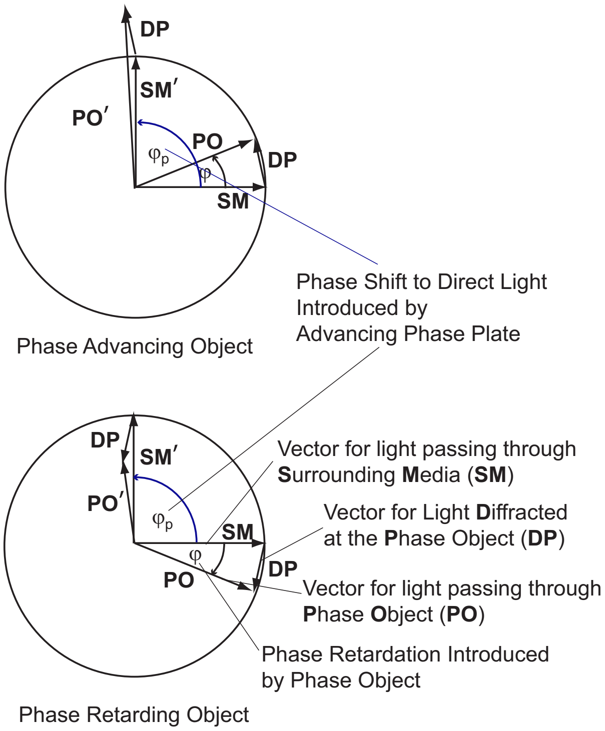
DP
SM′
PO′
PO
φ_p
φ
DP
SM
Phase Advancing Object
Phase Shift to Direct Light
Introduced by
Advancing Phase Plate
DP
SM′
PO′
φ_p
SM
φ
DP
PO
Vector for light passing through
Surrounding **M**edia (**SM**)
Vector for Light **D**iffracted
at the **P**hase Object (**DP**)
Vector for light passing through
Phase **O**bject (**PO**)
Phase Retardation Introduced
by Phase Object
Phase Retarding Object

Phase samples are considered to be **phase retarding objects** or **phase advancing objects** when their refractive index is greater or less than the refractive index of the surrounding media, respectively.

Phase plate	Object Type	Object Appearance
$\varphi_p = +\pi/2$ $(+90^o)$	phase-retarding	brighter
$\varphi_p = +\pi/2$ $(+90^o)$	phase-advancing	darker
$\varphi_p = -\pi/2$ (-90^o)	phase-retarding	darker
$\varphi_p = -\pi/2$ (-90^o)	phase-advancing	brighter

Visibility in Phase Contrast

Visibility of features in **phase contrast** can be expressed as

$$C_{\text{ph}} = \frac{I_{\text{media}} - I_{\text{object}}}{I_{\text{media}}} = \frac{|SM'|^2 - |PO'|^2}{|SM'|^2}.$$

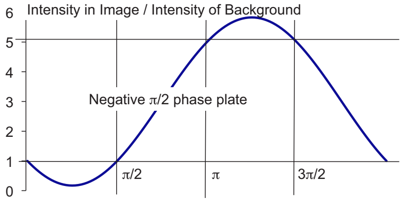

This equation defines the phase object's visibility as a ratio between the intensity change due to phase features and the intensity of surrounding media $|SM'|^2$. It defines the negative and positive contrast for an object appearing brighter or darker than the media, respectively. Depending on the introduced phase shift (positive or negative), the same object feature may appear with a **negative** or **positive contrast**. Note that C_{ph} relates to classical contrast C as

$$C = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} = \frac{|I_1 - I_2|}{I_1 + I_2} = |C_{\text{ph}}| \frac{|SM'|^2}{|SM'|^2 + |PO'|^2}.$$

Contrast	Phase Plate
-2φ	$\varphi_p = +\pi/2$ $(+90^\circ)$
$+2\varphi$	$\varphi_p = -\pi/2$ (-90°)

For phase changes in the 0–2π range, intensity in the image can be found using vector relations; small phase changes in the object $\varphi << 90$ deg can be approximated as $\pm 2\varphi$.

To increase the contrast of images, the intensity in the direct beam is additionally changed with beam attenuation τ in the phase ring. τ is defined as a transmittance, $\tau = 1/N$, where N is a dividing coefficient of intensity in a direct beam (intensity is decreased N times). The contrast in this case is

$$C_{\text{ph}} = -2\varphi / \sqrt{\tau} = -2\varphi\sqrt{N}$$

for a $+\pi/2$ phase plate, and

$$C_{\text{ph}} = +2\varphi / \sqrt{\tau} = +2\varphi\sqrt{N}$$

for $-\pi/2$ phase plate. The minimum perceived phase difference with phase contrast is

$$\Delta\varphi_{\text{min}} = \frac{C_{\text{ph-min}}\lambda}{4\pi\sqrt{N}}.$$

$C_{\text{ph-min}}$ is usually accepted at the contrast value of 0.02.

The Phase Contrast Microscope

The common phase-contrast system is similar to the bright-field microscope but with two modifications:

1. The condenser diaphragm is replaced with an annular aperture diaphragm.
2. A ring-shaped phase plate is introduced at the back focal plane of the microscope objective. The ring may include metallic coatings, typically providing 75% to 90% attenuattion of the direct beam.

Both components are in conjugate planes and are aligned such that the image of the annular condenser diaphragm overlaps with the objective phase ring. While there is no direct access to the objective phase ring, the anular aperture in front of the condenser is easily accessible.

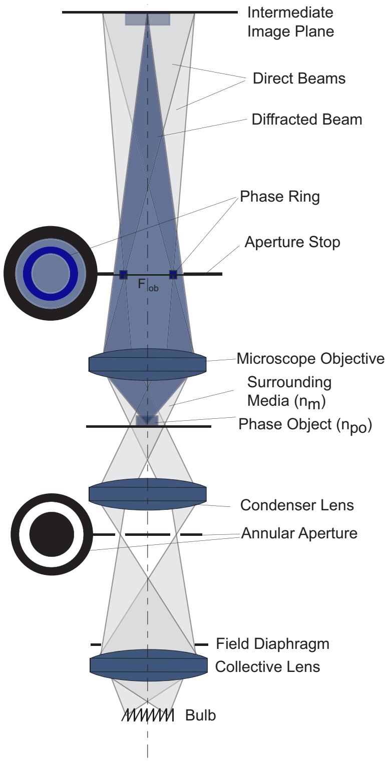
Intermediate Image Plane
Direct Beams
Diffracted Beam
Phase Ring
Aperture Stop
F ob
Microscope Objective
Surrounding Media (n$_m$)
Phase Object (n$_{po}$)
Condenser Lens
Annular Aperture
Field Diaphragm
Collective Lens
Bulb

The phase shift introduced by the ring is given by

$$\varphi_p = 2\pi \frac{OPD}{\lambda} = \frac{2\pi}{\lambda}(n_m - n_r)t ,$$

where n_m and n_r are refractive indices of the media surrounding the phase ring and the ring itself, respectively. t is the physical thickness of the phase ring.

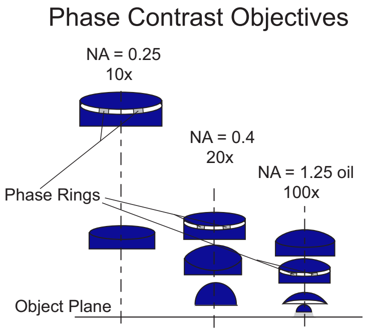
Phase Contrast Objectives
NA = 0.25
10x
NA = 0.4
20x
NA = 1.25 oil
100x
Phase Rings
Object Plane

Characteristic Features of Phase Contrast

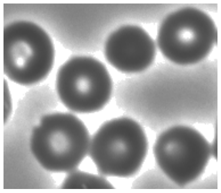

Images in phase contrast are dark or bright features on a background (positive and negative contrast, respectively). They contain undesired image effects called halo and shading-off, which are a result of the incomplete separation of direct and diffracted light. The **halo effect** is a phase contrast feature that increases light intensity around sharp changes in the phase gradient.

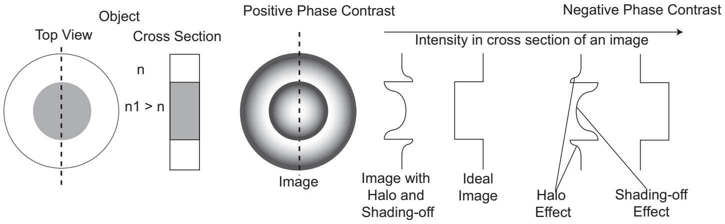
Object
Top View
Cross Section
n
n1 > n
Positive Phase Contrast
Image
Negative Phase Contrast
Intensity in cross section of an image
Image with Halo and Shading-off
Ideal Image
Halo Effect
Shading-off Effect

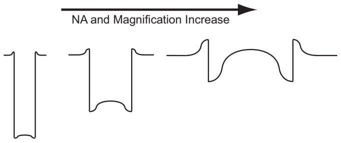
NA and Magnification Increase

The **shading-off effect** is an increase or decrease (for dark or bright images, respectively) of the intensity of the phase sample feature.

Both effects strongly increase with an increase in numerical aperture and magnification. They can be reduced by surrounding the sides of a phase ring with ND filters.

Lateral resolution of the phase contrast technique is affected by the annular aperture (the radius of phase ring r_{PR}) and the aperture stop of the objective (the radius of aperture stop is r_{AS}). It is

$$d = f'_{\text{objective}} \frac{\lambda}{(r_{\text{AS}} + r_{\text{PR}})},$$

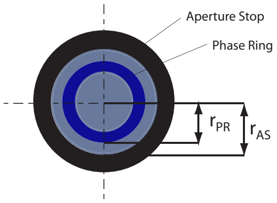
Aperture Stop
Phase Ring
r_{PR}
r_{AS}

compared to the resolution limit for a standard microscope:

$$d = f'_{\text{objective}} \frac{\lambda}{r_{\text{AS}}}.$$

Amplitude Contrast

Amplitude contrast changes the contrast in the images of absorbing samples. It has a layout similar to phase contrast; however, there is no phase change introduced by the object. In fact, in many cases a phase-contrast microscope can increase contrast based on the principle of amplitude contrast. The visibility of images is modified by changing the intensity ratio between the direct and diffracted beams.

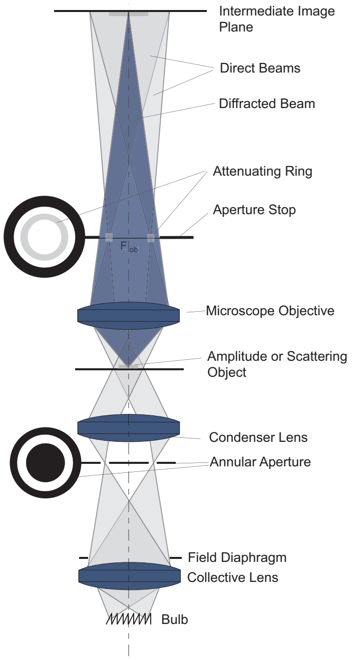
Intermediate Image Plane
Direct Beams
Diffracted Beam
Attenuating Ring
Aperture Stop
F_{ob}
Microscope Objective
Amplitude or Scattering Object
Condenser Lens
Annular Aperture
Field Diaphragm
Collective Lens
Bulb

A vector schematic of the technique is as follows:

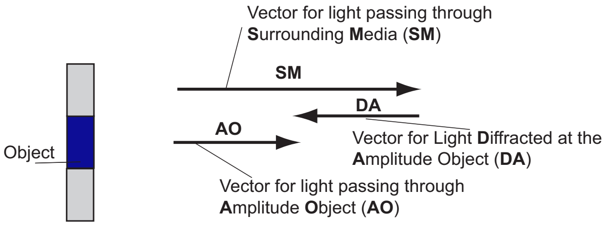
Vector for light passing through Surrounding Media (SM)
SM
DA
AO
Object
Vector for Light Diffracted at the Amplitude Object (DA)
Vector for light passing through Amplitude Object (AO)

Similar to visibility in phase contrast, image contrast can be described as a ratio of the intensity change due to amplitude features surrounding the media's intensity:

$$C_{\text{ac}} = \frac{I_{\text{media}} - I_{\text{object}}}{I_{\text{media}}} = \frac{2|SM||DA| - |DA|^2}{|SM|^2}\,.$$

Since the intensity of diffraction for amplitude objects is usually small, the contrast equation can be approximated with $C_{\text{ac}} = 2|DA|/|SM|$. If a direct beam is further attenuated, contrast C_{ac} will increase by factor of $N^{-1} = 1/\tau^{-1}$.

Oblique Illumination

Oblique illumination (also called **anaxial illumination**) is used to visualize phase objects. It creates pseudo-profile images of transparent samples. These reliefs, however, do not directly correspond to the actual surface profile.

The principle of oblique illumination can be explained by using either refraction or diffraction and is based on the fact that sample features of various spatial frequencies will have different intensities in the image due to a nonsymmetrical system layout and the filtration of spatial frequencies.

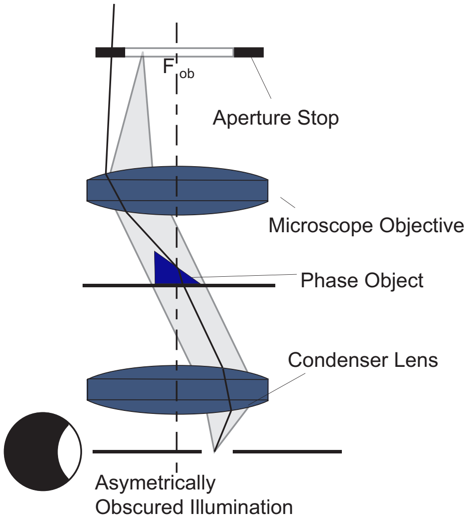

In practice, oblique illumination can be achieved by obscuring light exiting a condenser; translating the sub-stage diaphragm of the condenser lens does this easily.

The advantage of oblique illumination is that it improves spatial resolution over Köhler illumination (up to two times). This is based on the fact that there is a larger angle possible between the 0th and 1st orders diffracted at the sample for oblique illumination. In Köhler illumination, due to symmetry, the 0th and –1st orders at one edge of the stop will overlap with the 0th and +1st order on the other side. However, the gain in resolution is only for one sample direction; to examine the object features in two directions, a sample stage should be rotated.

Modulation Contrast

Modulation contrast microscopy (MCM) is based on the fact that phase changes in the object can be visualized with the application of multilevel gray filters.

The intensities of light refracted at the object are displayed with different values since they pass through different zones of the filter located in the stop of the microscope. MCM is often configured for oblique illumination since it already provides some intensity variations for phase objects. Therefore, the resolution of the MCM changes between normal and oblique illumination.

MCM can be obtained through a simple modification of a bright-field microscope by adding a slit diaphragm in front of the condenser and amplitude filter (also called the modulator) in the aperture stop of the microscope objective. The modulator filter consists of three areas: 100% (bright), 15% (gray), and 1% (dark). This filter acts in a similar way to the knife-edge in the Schlieren imaging techniques.

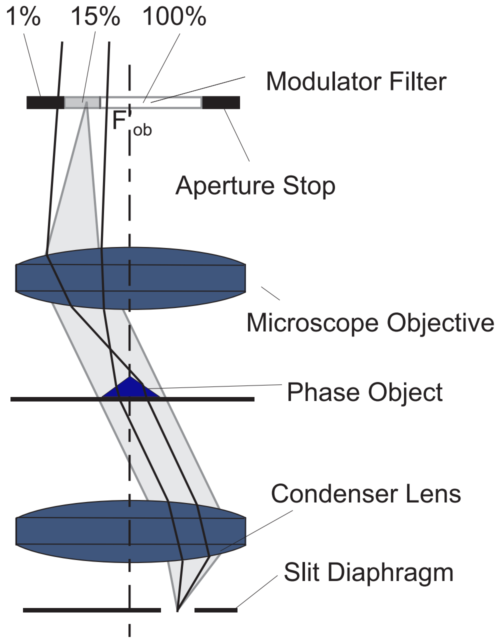
1%
15%
100%
Modulator Filter
F_{ob}
Aperture Stop
Microscope Objective
Phase Object
Condenser Lens
Slit Diaphragm

Hoffman Contrast

A specific implementation of MCM is **Hoffman modulation contrast**, which uses two additional polarizers located in front of the slit aperture. One polarizer is attached to the slit and obscures 50% of its width. A second can be rotated, which provides two intensity zones in the slit (for crossed polarizers, half of the slit is dark and half is bright; the entire slit is bright for the parallel position).

The major advantages of this technique include imaging at a full spatial resolution and a minimum depth of field, which allows optical sectioning. The method is inexpensive and easy to implement. The main drawback is that the images deliver a pseudo-relief image, which cannot be directly correlated with the object's form.

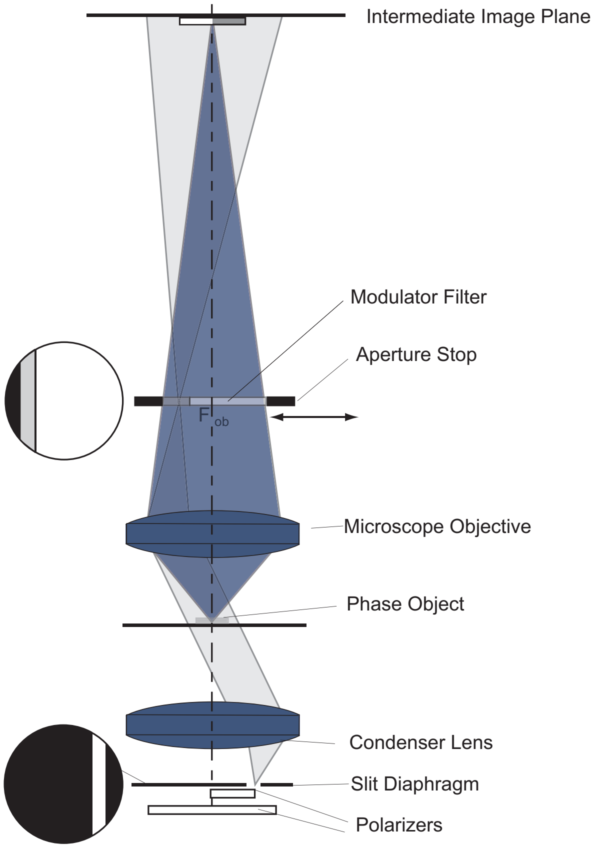
Intermediate Image Plane
Modulator Filter
Aperture Stop
F_{ob}
Microscope Objective
Phase Object
Condenser Lens
Slit Diaphragm
Polarizers

Dark Field Microscopy

In **dark-field microscopy**, the specimen is illuminated at such angles that direct light is not collected by the microscope objective. Only diffracted and scattered light are used to form the image.

The key component of the dark-field system is the condenser. The simplest approach uses an annular condenser stop and a microscope objective with an NA below that of the illumination cone. This approach can be used for objectives with NA < 0.65. Higher-NA objectives may incorporate an adjustable iris to reduce their NA below 0.65 for dark-field imaging. To image at a higher NA, specialized reflective dark-field condensers are necessary. Examples of such condensers include the **cardioid** and **paraboloid** designs, which enable dark-field imaging at an NA up to 0.9–1.0 (with oil immersion).

For dark-field imaging in epi-illumination, objective designs consist of external illumination, and internal imaging paths are used.

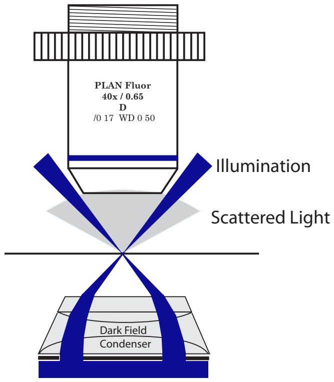
PLAN Fluor
40x / 0.65
D
/0 17 WD 0 50
Illumination
Scattered Light
Dark Field
Condenser

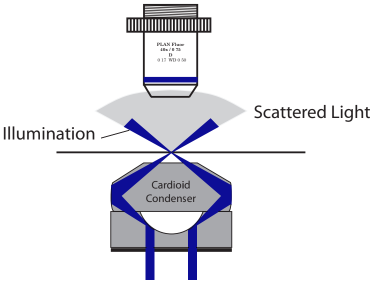
Scattered Light
Illumination
Cardioid
Condenser

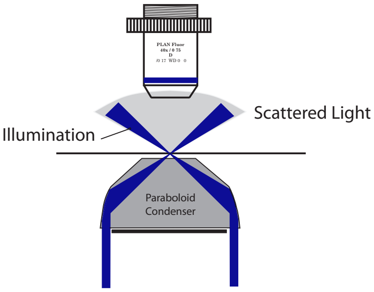
Scattered Light
Illumination
Paraboloid
Condenser

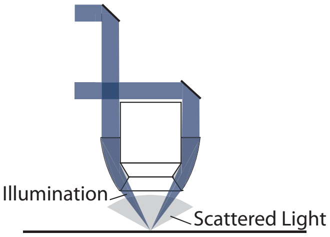
Illumination
Scattered Light

Optical Staining: Rheinberg Illumination

Optical staining is a class of techniques that uses optical methods to provide color differentiation between different areas of a sample.

Common Color Combinations	
Inner Filter	Annular Filter
Violet	Red
Blue	Yellow-Orange
Green	Red

Rheinberg illumination is a modification of dark-field illumination. Instead of an annular aperture, it uses two-zone color filters located in the front focal plane of the condenser. The overall diameter of the outer annular color filter is slightly greater than that corresponding to the NA of the system:

$$2r_{\lambda 2} > 2NAf'_{\text{condenser}}\,.$$

To provide good contrast between scattered and direct light, the inner filter is darker. Rheinberg illumination provides images that are a combination of two colors. Scattering features in one color are visible on the background of the other color.

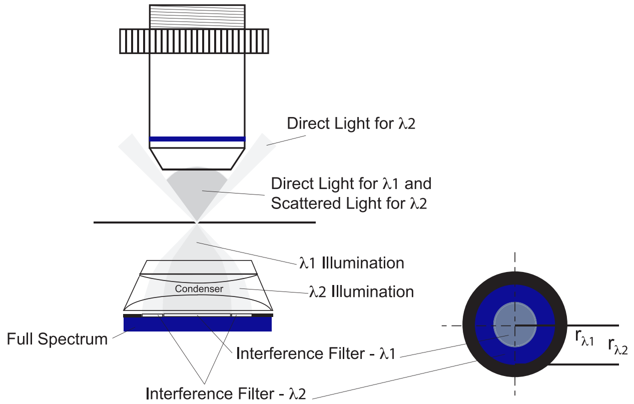
Direct Light for λ2
Direct Light for λ1 and
Scattered Light for λ2
λ1 Illumination
Condenser
λ2 Illumination
Full Spectrum
Interference Filter - λ1
Interference Filter - λ2
rλ1
rλ2

Color-stained images can also be obtained in a simplified setup where only one (inner) filter is used. For such an arrangement, images will present white-light scattering features on the colored background. Other deviations from the above technique include double illumination where transmittance and reflectance are separated into two different colors.

Optical Staining: Dispersion Staining

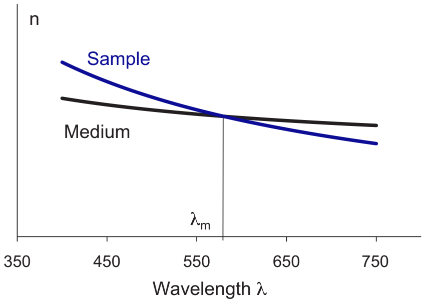

Dispersion staining is based on using highly dispersive media that matches the refractive index of the phase sample for a single wavelength λ_m (the matching wavelength). The system is configured in a setup similar to dark-field or oblique illumination. Wavelengths close to the matching wavelength will miss the microscope objective, while the other will refract at the object and create a colored dark-field image. Various configurations include illumination with a dark-field condenser and an annular aperture or central stop at the back focal plane of the microscope objective.

Color	λ Range [nm]	λ Mean [nm]
Violet	400–450	440
Blue	450–480	470
Blue-Green	480–510	495
Green	510–550	530
Green-Yellow	550–570	560
Yellow	570–590	580
Orange	590–630	605
Red	630–700	650

(Adapted from Pluta, 1989)

The annular-stop technique uses low-magnification (10×) objectives with a stop built as an opaque screen with a central opening, and is used for bright-field dispersion staining. The annular stop absorbs red and blue wavelengths and allows yellow through the system. A sample appears as white with yellow borders. The image background is also white.

The central-stop system absorbs unrefracted yellow light and direct white light. Sample features show up as purple (a mixture of red and blue) borders on a dark background. Different media and sample types will modify the color's appearance.

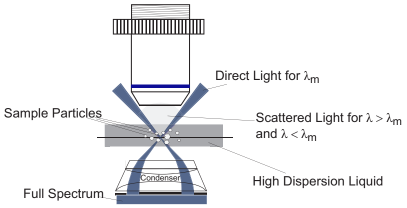
Direct Light for λ_m
Sample Particles
Scattered Light for $\lambda > \lambda_m$ and $\lambda < \lambda_m$
Condenser
High Dispersion Liquid
Full Spectrum

Shearing Interferometry: The Basis for DIC

Differential interference contrast (DIC) microscopy is based on the principles of shearing interferometry. Two wavefronts propagating through an optical system containing identical phase distributions (introduced by the sample) are slightly shifted to create an interference pattern. The phase of the resulting fringes is directly proportional to the derivative of the phase distribution in the object, hence the name "differential." Specifically, the intensity of the fringes relates to the derivative of the phase delay through the object (φ_o) and the local delay between wavefronts (Δ_b):

$$I \propto I_{\max} \cos^2\left(\Delta_b \pm \Delta\varphi_o\right)$$

or

$$I \propto I_{\max} \cos^2\left(\Delta_b \pm s\frac{d\varphi_o}{dx}\right),$$

where s denotes the shear between wavefronts, and Δ_b is the axial delay.

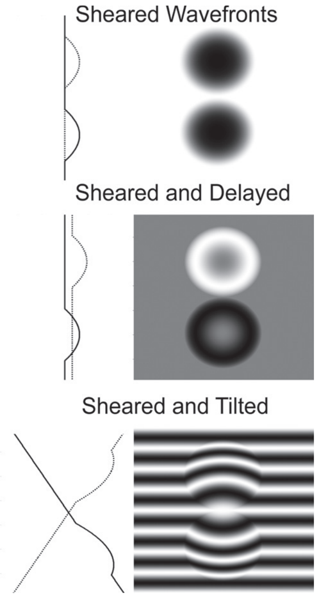

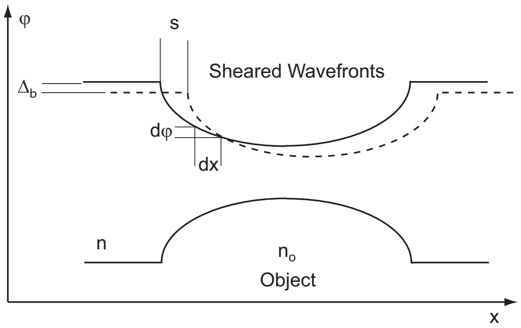

Wavefront shear is commonly introduced in DIC microscopy by incorporating a Mach-Zehnder interferometer into the system (e.g., Zeiss) or by the use of birefringent prisms. The appearance of DIC images depends on the sample orientation with respect to the shear direction.

DIC Microscope Design

The most common **differential interference contrast** (**Nomarski DIC**) design uses two **Wollaston** or **Nomarski birefringent prisms**. The first prism splits the wavefront into ordinary and extraordinary components to produce interference between these beams. Fringe localization planes for both prisms are in conjugate planes. Additionally, the system uses a crossed polarizer and analyzer. The polarizer is located in front of prism I, and the analyzer is behind prism II. The polarizer is rotated by 45 deg with regard to the shear axes of the prisms.

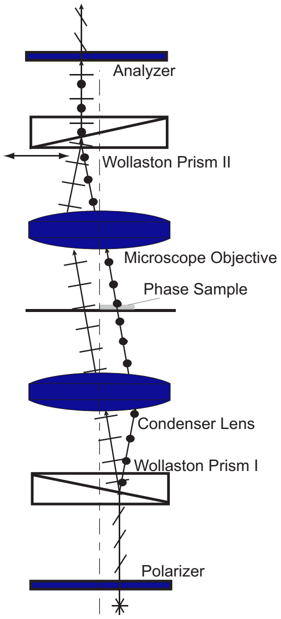

If prism II is centrally located, the intensity in the image is

$$I \propto \sin^2\left(\frac{\pi s \dfrac{d\varphi_o}{dx}}{\lambda}\right).$$

For a translated prism, phase **bias** Δ_b is introduced, and the intensity is proportional to

$$I \propto \sin^2 \frac{\pi\left(\Delta_b \pm s\dfrac{d\varphi_o}{dx}\right)}{\lambda}.$$

The sign in the equation depends on the direction of the shift. Shear s is

$$s = \frac{s'}{M_{\text{objective}}} = \varepsilon \frac{OTL}{M_{\text{objective}}},$$

where ε is the angular shear provided by the birefringent prisms, s' is the shear in the image space, and OTL is the optical tube length. The high spatial coherence of the source is obtained by the slit located in front of the condenser and oriented perpendicularly to the shear direction. To assure good interference contrast, the width of the slit w should be

$$w = \frac{\lambda f'_{\text{condenser}}}{4s}.$$

Low-strain (low-birefringence) objectives are crucial for high-quality DIC.

Appearance of DIC Images

In practice, shear between wavefronts in **differential interference contrast** is small and accounts for a fraction of a fringe period. This provides the differential character of the phase difference between interfering beams introduced to the interference equation.

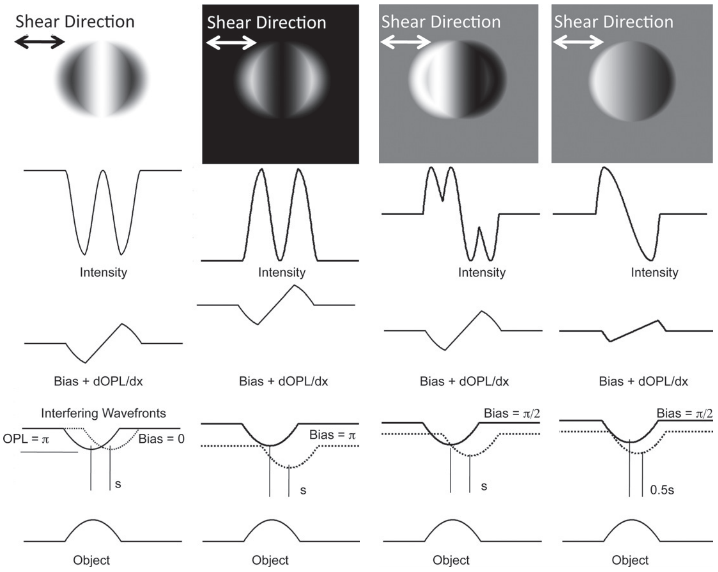
Shear Direction
Shear Direction
Shear Direction
Shear Direction
Intensity
Intensity
Intensity
Intensity
Bias + dOPL/dx
Bias + dOPL/dx
Bias + dOPL/dx
Bias + dOPL/dx
Interfering Wavefronts
OPL = π
Bias = 0
s
Bias = π
s
Bias = π/2
s
Bias = π/2
0.5s
Object
Object
Object
Object

Increased shear makes the edges of the object less sharp, while increased bias introduces some background intensity. Shear direction determines the appearance of the object.

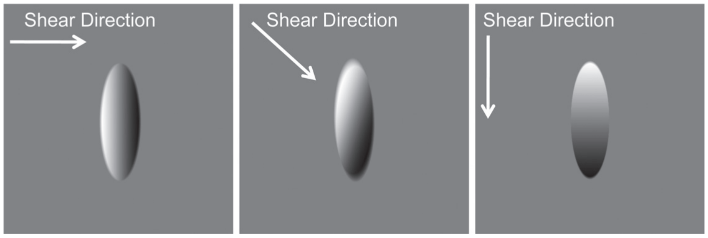
Shear Direction
Shear Direction
Shear Direction

Compared to phase contrast, DIC allows for larger phase differences throughout the object and operates in full resolution of the microscope (i.e., it uses the entire aperture). The depth of field is minimized, so DIC allows optical sectioning.

Reflectance DIC

A **Nomarski interference microscope** (also called a polarization interference contrast microscope) is a reflectance DIC system developed to evaluate roughness and the surface profile of specular samples. Its applications include metallography, microelectronics, biology, and medical imaging. It uses one Wollaston or Nomarski prism, a polarizer, and an analyzer. The information about the sample is obtained for one direction parallel to the gradient in the object. To acquire information for all directions, the sample should be rotated.

If white light is used, different colors in the image correspond to different sample slopes. It is possible to adjust the color by rotating the polarizer.

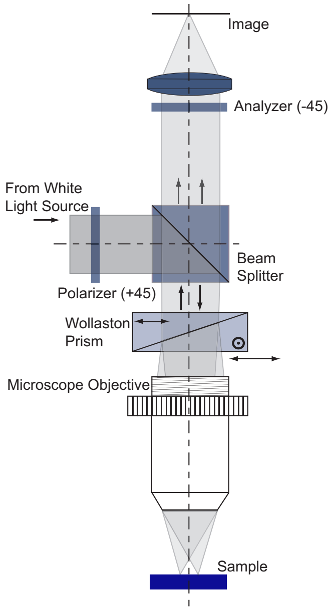
Image
Analyzer (-45)
From White Light Source
Beam Splitter
Polarizer (+45)
Wollaston Prism
Microscope Objective
Sample

Phase delay (**bias**) can be introduced by translating the birefringent prism. For imaging under monochromatic illumination, the brightness in the image corresponds to the slope in the sample. Images with no bias will appear as bright features on a dark background; with bias they will have an average background level, and slopes will vary between dark and bright. This way, it is possible to interpret the direction of the slope. Similar analysis can be performed for the colors from white-light illumination.

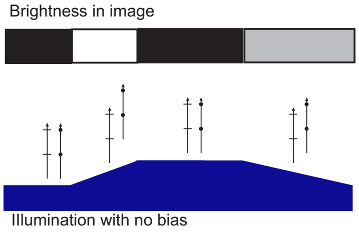
Brightness in image
Illumination with no bias

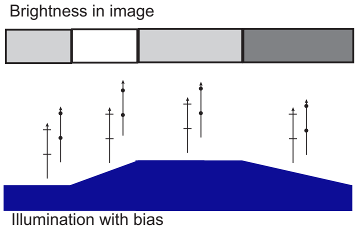
Brightness in image
Illumination with bias

Polarization Microscopy

Polarization microscopy provides images containing information about the properties of anisotropic samples. A polarization microscope is a modified compound microscope with three unique components: the **polarizer**, **analyzer**, and **compensator**. A linear polarizer is located between the light source and the specimen, close to the condenser's diaphragm. The analyzer is a linear polarizer with its transmission axis perpendicular to the polarizer. It is placed between the sample and the eye/camera, at or near the microscope objective's aperture stop. If no sample is present, the image should appear uniformly dark. The compensator (a type of retarder) is a birefringent plate of known parameters used for quantitative sample analysis and contrast adjustment.

Quantitative information can be obtained by introducing known retardation. Rotation of the analyzer correlates the analyzer's angular position with the retardation required to minimize the light intensity for object features at selected wavelengths. **Retardation** introduced by the compensator plate can be described as

$$\Gamma = \left(n_{\mathrm{e}} - n_{\mathrm{o}}\right) t \, ,$$

where t is the sample thickness, and subscripts e and o denote the extraordinary and ordinary beams. Retardation in concept is similar to the optical path difference for beams propagating through two different materials:

$$OPD = \left(n_1 - n_2\right) t = \left(n_{\mathrm{e}} - n_{\mathrm{o}}\right) t \, .$$

The phase delay caused by sample birefringence is therefore

$$\delta = \frac{2\pi OPD}{\lambda} = \frac{2\pi\Gamma}{\lambda} \, .$$

A polarization microscope can also be used to determine the orientation of the optic axis.

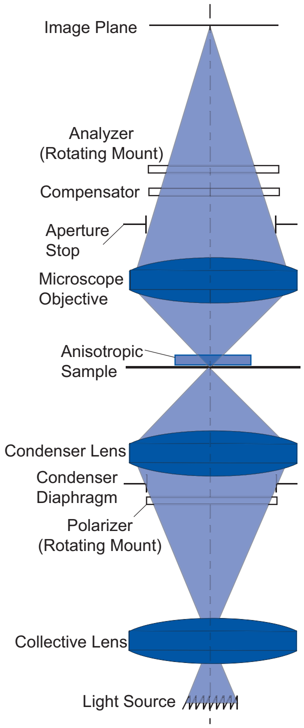
Image Plane
Analyzer (Rotating Mount)
Compensator
Aperture Stop
Microscope Objective
Anisotropic Sample
Condenser Lens
Condenser Diaphragm
Polarizer (Rotating Mount)
Collective Lens
Light Source

Images Obtained with Polarization Microscopes

Anisotropic samples observed under a **polarization microscope** contain dark and bright features and will strongly depend on the geometry of the sample. Objects can have characteristic elongated, linear, or circular structures:

- linear structures appear as dark or bright, depending on the orientation of the analyzer.
- circular structures have a "Maltese Cross" pattern with four quadrants of different intensities.

While polarization microscopy uses monochromatic light, it can also use white-light illumination. For broadband light, different wavelengths will be subject to different retardation; as a result, samples can provide different output colors. In practical terms, this means that color encodes information about retardation introduced by the object. The specific retardation is related to the image color. Therefore, the color allows for determination of sample thickness (for known retardation) or its retardation (for known sample thickness). If the polarizer and the analyzer are crossed, the color visible through the microscope is a complementary color to that having full wavelength retardation (a multiple of 2π, and the intensity is minimized for this color). Note that the two complementary colors combine into white light. If the analyzer could be rotated to be parallel to the analyzer, the two complementary colors will be switched (the one previously displayed will be minimized, while the other color will be maximized).

Polarization microscopy can provide quantitative information about the sample with the application of compensators (with known retardation). Estimating retardation might be performed visually or by integrating with an image acquisition system and CCD. This latter method requires one to obtain a series of images for different retarder orientations and calculate sample parameters based on saved images.

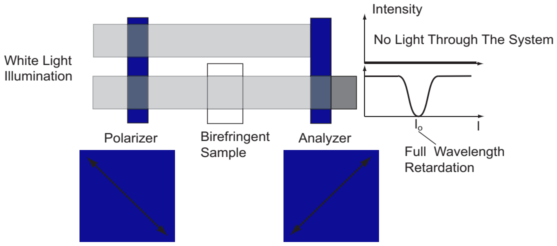

Compensators

Compensators are components that can be used to provide quantitative data about a sample's retardation. They introduce known retardation for a specific wavelength and can act as nulling devices to eliminate any phase delay introduced by the specimen. Compensators can also be used to control the background intensity level.

A **wave plate compensator** (also called a first-order red or red-I plate) is oriented at a 45-deg angle with respect to the analyzer and polarizer. It provides retardation equal to an even number of waves for 551 nm (green), which is the wavelength eliminated after passing the analyzer. All other wavelengths are retarded with a fraction of the wavelength and can partially pass the analyzer and appear as a bright red magenta. The sample provides additional retardation and shifts the colors towards blue and yellow. Color tables determine the retardation introduced by the object.

The **de Sénarmont compensator** is a quarter-wave plate (for 546 nm or 589 nm). It is used with monochromatic light for samples with retardation in the range of $\lambda/20$ to λ. A quarter-wave plate is placed between the analyzer and the polarizer with its slow ellipsoid axis parallel to the transmission axis of the analyzer. Retardation of the sample is measured in a two-step process. First, the sample is rotated until the maximum brightness is obtained. Next, the analyzer is rotated until the intensity maximally drops (the extinction position). The analyzer's rotation angle θ finds the phase delay related to retardation: $\Gamma_{\text{sample}} = 2\theta$.

A **Brace Köhler** rotating **compensator** is used for small retardations and monochromatic illumination. It is a thin birefringent plate with an optic axis in its plane. The analyzer and polarizer remain fixed while the compensator is rotated between the maximum and minimum intensities in the sample's image. Zero position (the maximum intensity) is determined for a slow axis being parallel to the transmission axis of the analyzer, and retardation is

$$\Gamma_{\text{sample}} = \Gamma_{\text{compensator}} \sin 2\theta \, .$$

Confocal Microscopy

Confocal microscopy is a scanning technique that employs pinholes at the illumination and detection planes so that only in-focus light reaches the detector. This means that the intensity for each sample point must be obtained in sequence through scanning in the x and y directions. An illumination pinhole is used in conjunction with the imaged sample plane and only illuminates the point of interest. The detection pinhole rejects the majority of out-of-focus light.

Confocal microscopy has the advantage of lowering the background light from out-of-focus layers, increasing spatial resolution, and providing the capability of imaging thick 3D samples, if combined with z scanning. Due to detection of only the in-focus light, confocal microscopy can provide images of thin sample sections. The system usually employs a photo-multiplier tube (PMT), avalanche photodiodes (APD), or a charge-coupled device (CCD) camera as a detector. For point detectors, recorded data is processed to assemble x-y images. This makes it capable of quantitative studies of an imaged sample's properties. Systems can be built for both reflectance and fluorescence imaging.

Spatial resolution of a confocal microscope can be defined as

$$d_{xy} \approx \frac{0.4\lambda}{NA},$$

and is slightly better than wide-field (bright-field) microscopy resolution. For pinholes larger than an Airy disk, the spatial resolution is the same as in a wide-field microscope. The **axial resolution** of a confocal system is

$$d_z \approx \frac{1.4n\lambda}{NA^2}.$$

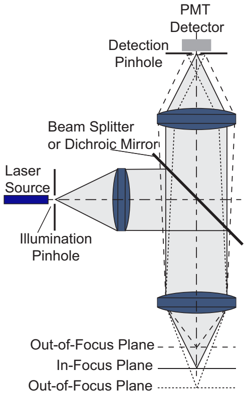

The optimum pinhole value is at the full width at half maximum (FWHM) of the Airy disk intensity. This corresponds to ~75% of its intensity passing through the system, minimally smaller than the Airy disk's first ring:

$$D_{\text{pinhole}} = \frac{0.5\lambda M}{NA}.$$

Scanning Approaches

A scanning approach is directly connected with the temporal resolution of confocal microscopes. The number of points in the image, scanning technique, and the frame rate are related to the signal-to-noise ratio SNR (through time dedicated to detection of a single point). To balance these parameters, three major approaches were developed: point scanning, line scanning, and disk scanning.

A **point-scanning confocal microscope** is based on point-by-point scanning using, for example, two mirror galvanometers or resonant scanners. Scanning mirrors should be located in pupil conjugates (or close to them) to avoid light fluctuations. Maximum spatial resolution and maximum background rejection are achieved with this technique.

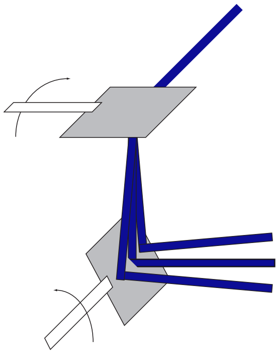

A **line-scanning confocal microscope** uses a slit aperture that scans in a direction perpendicular to the slit. It uses a cylindrical lens to focus light onto the slit to maximize throughput. Scanning in one direction makes this technique significantly faster than a point approach. However, the drawbacks are a loss of resolution and sectioning performance for the direction parallel to the slit.

Feature	**Point Scanning**	**Slit Slit Scanning**	**Disk Spinning**
z resolution	High	Depends on slit spacing	Depends on pinhole distribution
x,*y* resolution	High	Lower for one direction	Depends on pinhole spacing
Speed	Low to moderate	High	High
Light sources	Lasers	Lasers	Laser and other
Photobleaching	High	High	Low
QE of detectors	Low (PMT) Good (APD)	Good (CCD)	Good (CCD)
Cost	High	High	Moderate

Scanning Approaches (cont.)

Spinning-disk confocal imaging is a parallel-imaging method that maximizes the scanning rate and can achieve a 100–1000 speed gain over point scanning. It uses an array of pinholes/slits (e.g., Nipkow disk, Yokogawa, Olympus DSU approach). To minimize light loss, it can be combined (e.g., with the Yokogawa approach) with an array of lenses, so each pinhole has a dedicated focusing component. Pinhole disks contain several thousand pinholes, but only a portion is illuminated at one time.

	Throughput of a confocal microscope T [%]
Array of pinholes	$T_{\text{pinhole}} = 100 \times (D/S)^2$
Multiple slits	$T_{\text{slit}} = 100 \times (D/S)$
Equations are for a uniformly illuminated mask of pinholes/slits (array of microlenses not considered). D is the pinhole's diameter or slit's width respectively, while S is the pinhole/slit separation.	

Crosstalk between pinholes and slits will reduce the background rejection. Therefore a common pinhole/slit separation S is 5−10 times larger than the pinhole's diameter or the slit's width.

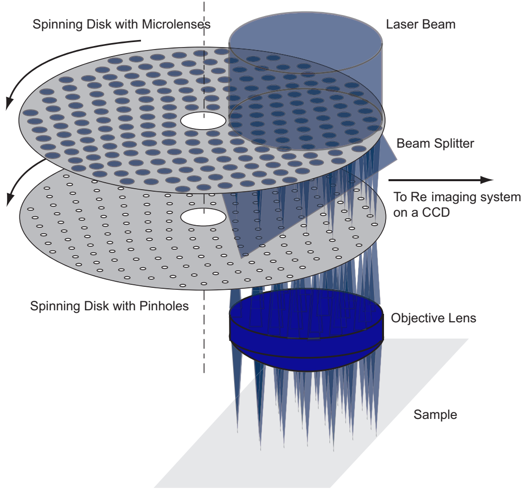
Spinning Disk with Microlenses
Laser Beam
Beam Splitter
To Re imaging system on a CCD
Spinning Disk with Pinholes
Objective Lens
Sample

Spinning-disk and slit-scanning confocal microscopes require high-sensitivity array image sensors (CCD cameras) instead of point detectors. They can use lower excitation intensity for fluorescence confocal systems; therefore, they are less susceptible to photobleaching.

Images from a Confocal Microscope

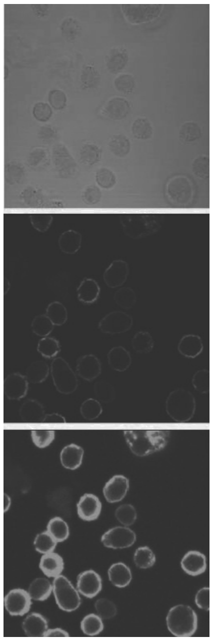

Laser-scanning confocal microscopy rejects out-of-focus light and enables optical sectioning. It can be assembled in reflectance and fluorescent modes. A 1483 cell line, stained with membrane anti-EGFR antibodies labeled with fluorecent Alexa488 dye, is presented here. The top image shows the bright-field illumination mode. The middle image is taken by a confocal microscope with a pinhole corresponding to approximately 2.5 Airy disks. In the bottom image, the pinhole was entirely open. Clear cell boundaries are visible in the confocal image, while all out-of-focus features were removed.

Images were taken with a Zeiss LSM 510 microscope using a 63× Zeiss oil-immersion objective. Samples were excited with a 488-nm laser source.

Another example of 1483 cells labeled with EGF-Alexa647 and proflavine obtained on a Zeiss LSM 510 confocal at 63× / 1.4 oil is shown below. The proflavine (staining nuclei) was excited at 488 nm, with emission collected after a 505-nm long-pass filter. The EGF-Alexa647 (cell membrane) was excited at 633 nm, with emission collected after a 650–710 nm bandpass filter. The sectioning ability of a confocal system is nicely demonstrated by the channel with the membrane labeling.

EGF-Alexa647 Channel (Red)

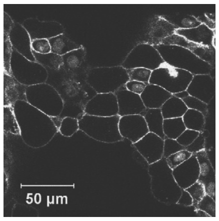

Proflavine Channel (Green)

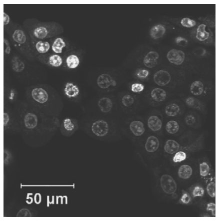

Combined Channels

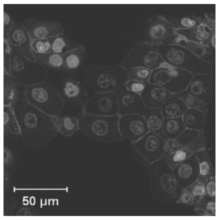

Fluorescence

Specimens can absorb and re-emit light through **fluorescence**. The specific wavelength of light absorbed or emitted depends on the energy level structure of each molecule. When a molecule absorbs the energy of light, it briefly enters an excited state before releasing part of the energy as fluorescence. Since the emitted energy must be lower than the absorbed energy, fluorescent light is always at longer wavelengths than the excitation light. Absorption and emission take place between multiple sub-levels within the ground and excited states, resulting in absorption and emission spectra covering a range of wavelengths. The loss of energy during the fluorescence process causes the **Stokes shift** (a shift of wavelength peak from that of excitation to that of emission). Larger Stokes shifts make it easier to separate excitation and fluorescent light in the microscope. Note that energy quanta and wavelength are related by $E = hc/\lambda$.

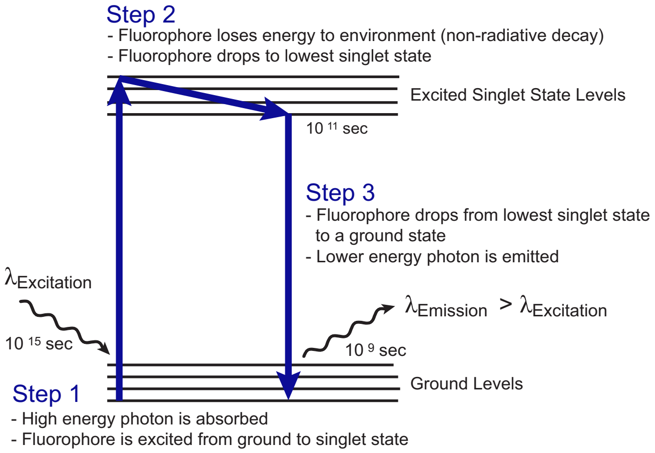
Step 2
- Fluorophore loses energy to environment (non-radiative decay)
- Fluorophore drops to lowest singlet state
Excited Singlet State Levels
10^{11} sec
Step 3
- Fluorophore drops from lowest singlet state to a ground state
- Lower energy photon is emitted
$\lambda_{Excitation}$
$\lambda_{Emission} > \lambda_{Excitation}$
10^{15} sec
10^{9} sec
Ground Levels
Step 1
- High energy photon is absorbed
- Fluorophore is excited from ground to singlet state

The fluorescence principle is widely used in fluorescence microscopy for both organic and inorganic substances. While it is most common for biological imaging, it is also possible to examine samples like drugs and vitamins.

Due to the application of filter sets, the fluorescence technique has a characteristically low background and provides high-quality images. It is used in combination with techniques like wide-field microscopy and scanning confocal-imaging approaches.

Configuration of a Fluorescence Microscope

A fluorescence microscope includes a set of three **filters**: an **excitation filter**, **emission filter**, and a **dichroic mirror** (also called a **dichroic beam splitter**). These filters separate weak emission signals from strong excitation illumination. The most common fluorescence microscopes are configured in epi-illumination mode. The dichroic mirror reflects incoming light from the lamp (at short wavelengths) onto the specimen. Fluorescent light (at longer wavelengths) collected by the objective lens is transmitted through the dichroic mirror to the eyepieces or camera. The transmission and reflection properties of the dichroic mirror must be matched to the excitation and emission spectra of the fluorophore being used.

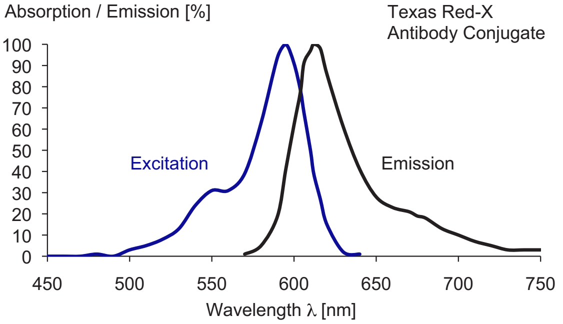

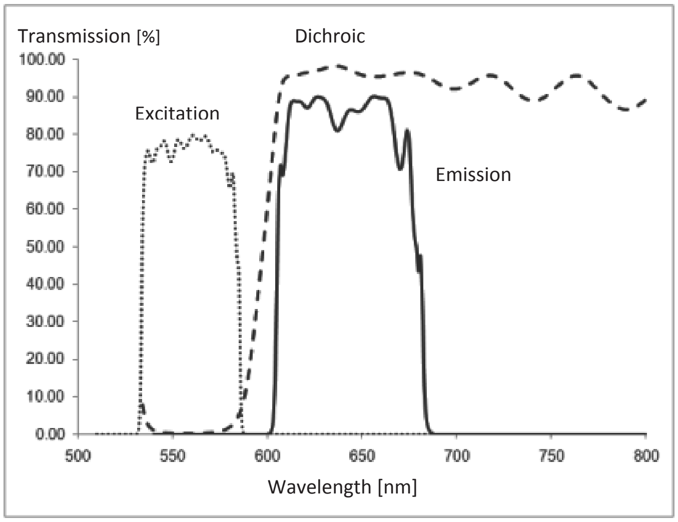

Configuration of a Fluorescence Microscope (cont.)

A sample must be illuminated with light at wavelengths within the excitation spectrum of the fluorescent dye. Fluorescence-emitted light must be collected at the longer wavelengths within the emission spectrum. Multiple fluorescent dyes can be used simultaneously, with each designed to localize or target a particular component in the specimen.

The **filter turret** of the microscope can hold several cubes that contain filters and dichroic mirrors for various dye types. Images are then reconstructed from CCD frames captured with each filter cube in place. Simultanous acquisition of emission for multiple dyes requires application of multiband dichroic filters.

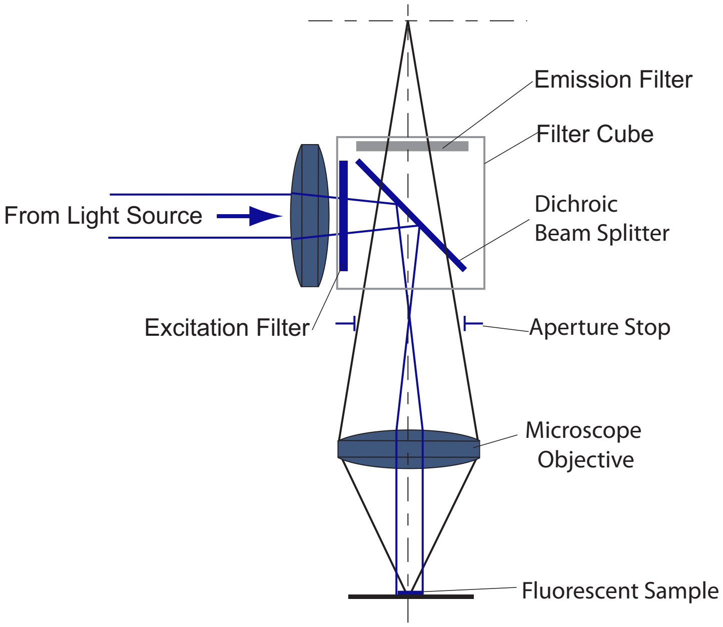
Emission Filter
Filter Cube
From Light Source
Dichroic Beam Splitter
Excitation Filter
Aperture Stop
Microscope Objective
Fluorescent Sample

Images from Fluorescence Microscopy

Fluorescent images of cells labeled with three different fluorescent dyes, each targeting a different cellular component, demonstrate fluorescence microscopy's ability to target sample components and specific functions of biological systems. Such images can be obtained with individual filter sets for each fluorescent dye, or with a multiband (a triple-band in this example) filter set, which matches all dyes.

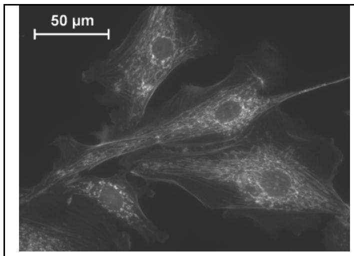

Triple-band Filter

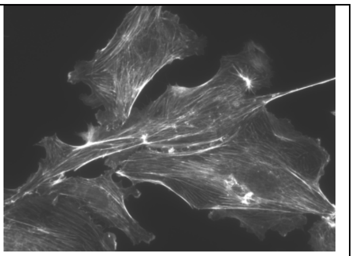

BODIPY FL phallacidin (F-actin)

MitoTracker Red CMXRos (mitochondria)

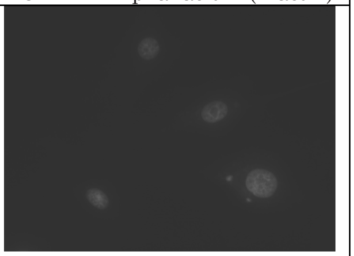

DAPI (nuclei)

Sample: Bovine pulmonary artery epithelial cells (Invitrogen FluoCells No. 1)

Components: Plan-apochromat 40×/0.95, MRm Zeiss CCD (1388×1040 mono)

Fluorescent labels:	Peak Excitation	Peak Emission
DAPI	358 nm	461 nm
BODIPY FL	505 nm	512 nm
MitoTracker Red CMXRos	579 nm	599 nm

Filter:	Excitation	Dichroic	Emission
Triple-band	395–415	435	448–472
	480–510	510	510–550
	560–590	600	600–650
DAPI	325–375	395	420–470
GFP	450–490	495	500–550
Texas Red	530–585	600	615LP

Properties of Fluorophores

Fluorescent emission F [photons/s] depends on the intensity of excitation light I [photons/s] and fluorophore parameters like the **fluorophore quantum yield** Q and **molecular absorption cross section** σ:

$$F = \sigma Q I \ ,$$

where the molecular absorption cross section is the probability that illumination photons will be absorbed by fluorescent dye. The intensity of excitation light depends on the power of the light source and throughput of the optical system. The detected fluorescent emission is further reduced by the quantum efficiency of the detector.

Quenching is a partially reversible process that decreases fluorescent emission. It is a non-emissive process of electrons moving from an excited state to a ground state.

Fluorescence emission and excitation parameters are bound through a photobleaching effect that reduces the ability of fluorescent dye to fluoresce. **Photobleaching** is an irreversible phenomenon that is a result of damage caused by illuminating photons. It is most likely caused by the generation of long-living (in triplet states) molecules of singlet oxygen and oxygen free radicals generated during its decay process. There is usually a finite number of photons that can be generated for a fluorescent molecule. This number can be defined as the ratio of emission quantum efficiency to bleaching quantum efficiency, and it limits the time a sample can be imaged before entirely bleaching. Photobleaching causes problems in many imaging techniques, but it can be especially critical in time-lapse imaging. To slow down this effect, optimizing collection efficiency (together with working at lower power at the steady-state condition) is crucial.

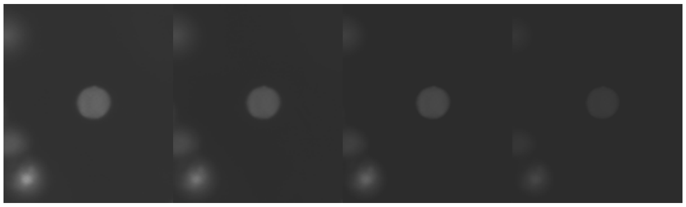

Photobleaching effect as seen in consecutive images

Single vs Multi-Photon Excitation

Multi-photon fluorescence is based on the fact that two or more low-energy photons match the energy gap of the fluorophore to excite and generate a single high-energy photon. For example, the energy of two 700-nm photons can excite an energy transition approximately equal to that of a 350-nm photon (it is not an exact value due to thermal losses). This is contrary to traditional fluorescence, where a high-energy photon (e.g., 400 nm) generates a slightly lower-energy (longer wavelength) photon. Therefore, one big advantage of multi-photon fluorescence is that it can obtain a significant difference between excitation and emission.

Fluorescence is based on stochastic behavior, and for single-photon excitation fluorescence, it is obtained with a high probability. However, multi-photon excitation requires at least two photons delivered in a very short time, and the probability is quite low:

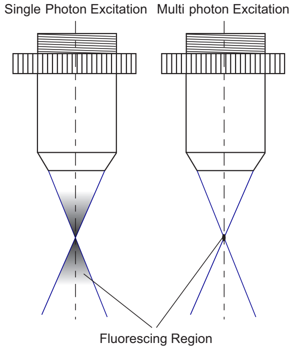

$$n_a \propto \delta\left(\frac{P_{\mathrm{avg}}^2}{\tau\nu^2}\right)\left(\pi\frac{NA^2}{hc\lambda}\right)^2.$$

δ is the excitation cross section of dye, P_{avg} is the average power, τ is the length of a pulse, and ν is the repetition frequency. Therefore, multi-photon fluorescence must be used with high-power laser sources to obtain favorable probability. Short-pulse operation will have a similar effect, as τ will be minimized.

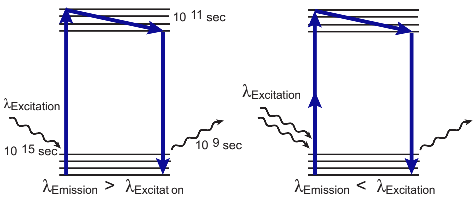

Due to low probability, fluorescence occurs only in the focal point. Multi-photon imaging does not require scanning through a pinhole and therefore has a high signal-to-noise ratio and can provide deeper imaging. This is also because near-IR excitation light penetrates more deeply due to lower scattering, and near-IR light avoids the excitation of several autofluorescent tissue components.

Light Sources for Scanning Microscopy

Lasers are an important light source for scanning microscopy systems due to their high energy density, which can increase the detection of both reflectance and fluorescence light. For laser-scanning confocal systems, a general requirement is a single-mode TEM_{00} laser with a short coherence length. Lasers are used primarily for point-and-slit scanning modalities. There are a great variety of laser sources, but certain features are useful depending on their specific application:

- Short-excitation wavelengths are preferred for fluorescence confocal systems because many fluorescent probes must be excited in the blue or ultraviolet range.
- Near-IR wavelengths are useful for reflectance confocal microscopy and multi-photon excitation. Longer wavelengths are less suseptible to scattering in diffused objects (like tissue) and can provide a deeper penetration depth.
- Broadband short-pulse lasers (like a Ti-Sapphire mode-locked laser) are particularily important for multi-photon generation, while shorter pulses increase the probability of excitation.

Laser Type	Wavelength [nm]
Argon-Ion	351, 364, 458, 488, 514
HeCd	325, 442
HeNe	543, 594, 633, 1152
Diode lasers	405, 408, 440, 488, 630, 635, 750, 810
DPSS (diode pumped, solid state)	430, 532, 561
Krypton-Argon	488, 568, 647
Dye	630
Ti-Sapphire*	710–920, 720–930, 750–850, 690–100, 680–1050

*High-power (1000mW or less), pulses between 1 ps and 100 fs

Non-laser sources are particularly useful for disk-scanning systems. They include arc lamps, metal-halide lamps, and high-power photo-luminescent diodes (See also pages 58 and 59).

Practical Considerations in LSM

A light source in laser-scanning microscopy must provide enough energy to obtain a sufficient number of photons to perform successful detection and a statistically significant signal. This is especially critical for non-laser sources (e.g., arc lamps) used in disk-scanning systems. While laser sources can provide enough power, they can cause fast photobleaching or photo-damage to the biological sample.

Detection conditions change with the type of sample. For example, fluorescent objects are subject to photobleaching and saturation. On the other hand, back-scattered light can be easily rejected with filter sets. In reflectance mode, out-of-focus light can cause background comparable or stronger than the signal itself. This background depends on the size of the pinhole, scattering in the sample, and overall system reflections.

Light losses in the optical system are critical for the signal level. Collection efficiency is limited by the NA of the objective; for example, a NA of 1.4 gathers approximately 30% of the fluorescent/scattered light. Other system losses will occur at all optical interfaces, mirrors, filters, and at the pinhole. For example, a pinhole matching the Airy disk allows approximately 75% of the light from the focused image point.

Quantum efficiency (QE) of a detector is another important system parameter. The most common photodetector used in scanning microscopy is the photomultiplier tube (PMT). Its quantum efficiency is often limited to the range of 10–15% (550–580 nm). In practical terms, this means that only 15% of the photons are used to produce a signal. The advantages of PMTs for laser-scanning microscopy are high speed and high signal gain. Better QE can be obtained for CCD cameras, which is 40–60% for standard designs and 90–95% for thinned, back-illuminated cameras. However, CCD image sensors operate at lower rates.

Spatial and temporal sampling of laser-scanning microscopy images imposes very important limitations on image quality. The image size and frame rate are often determined by the number of photons sufficient to form high-quality images.

Interference Microscopy

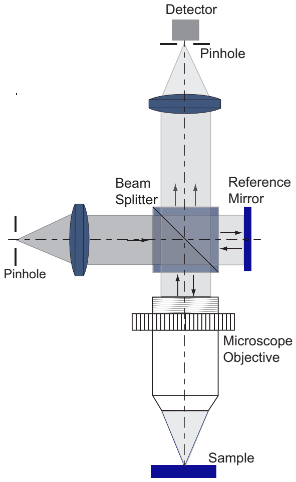

Interference microscopy includes a broad family of techniques for quantitative sample characterization (thickness, refractive index, etc.). Systems are based on microscopic implementation of interferometers (like the Michelson and Mach-Zehnder geometries), or polarization interferometers.

In interference microscopy, short coherence systems are particularly interesting and can be divided into two groups: **optical profilometry** and **optical coherence tomography** (OCT) [including **optical coherence microscopy** (OCM)]. The primary goal of these techniques is to add a third (z) dimension to the acquired data.
Optical profilers use interference fringes as a primary source of object height. Two important measurement approaches include **vertical scanning interferometry** (VSI) (also called **scanning white light interferometry** or **coherence scanning interferometry**) and **phase-shifting interferometry**. Profilometry techniques are capable of achieving nanometer-level resolution in the z direction while x and y are defined by standard microscope limitations.

Profilometry is used for characterizing the sample surface (roughness, structure, microtopography, and in some cases film thickness) while OCT/OCM is dedicated to 3D imaging, primarily of biological samples.
Both types of systems are usually configured using the geometry of the Michelson or its derivatives (e.g., Mirau).

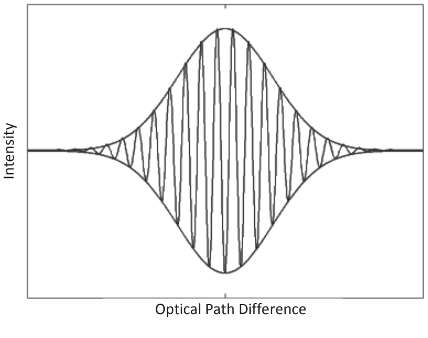

Optical Coherence Tomography/Microscopy

In early 3D coherence imaging, information about the sample was gated by the coherence length of the light source (**time-domain OCT**). This means that the maximum fringe contrast is obtained at a zero optical path difference, while the entire fringe envelope has a width related to the coherence length. In fact, this width defines the axial resolution (usually a few microns) and images are created from the magnitude of the fringe pattern envelope. **Optical coherence microscopy** is a combination of OCT and confocal microscopy. It merges the advantages of out-of-focus background rejection provided by the confocal principle and applies the coherence gating of OCT to improve optical sectioning over confocal reflectance systems.

In the mid-2000s, time-domain OCT transitioned to **Fourier-domain OCT** (FDOCT), which can acquire an entire depth profile in a single capture event. Similar to Fourier Transform Spectroscopy, the image is acquired by calculating the Fourier transform of the obtained interference pattern. The measured signal can be described as

$$I(k, z_o) = \int A_R A_S(z) \cos\left[k(z_m - z_o)\right] dz ,$$

where A_R and A_S are amplitudes of the reference and sample light, respectively, and z_m is the imaged sample depth (z_o is the reference length delay, and k is the wave number).

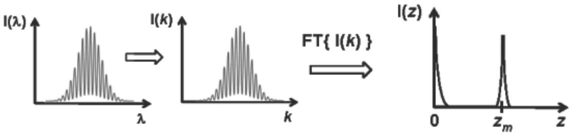

The fringe frequency is a function of wave number k and relates to the OPD in the sample. Within the Fourier-domain techniques, two primary methods exist: **spectral-domain OCT** (SDOCT) and **swept-source OCT** (SSOCT). Both can reach detection sensitivity and signal-to-noise ratios over 150 times higher than time-domain OCT approaches. SDOCT achieves this through the use of a broadband source and spectrometer for detection. SSOCT uses a single photodetector but rapidly tunes a narrow linewidth over a broad spectral profile. Fourier-domain systems can image at hundreds of frames per second with sensitivities over 100 dB.

Optical Profiling Techniques

There are two primary techniques used to obtain a surface profile using white-light profilometry: **vertical scanning interferomtry** (VSI) and **phase-shifting interferometry** (PSI).

VSI is based on the fact that a reference mirror is moved by a piezoelectric or other actuator, and a sequence of several images (in some cases hundreds or thousands) is acquired along the scan. During post-processing, algorithms look for a maximum fringe modulation and correlate it with the actuator's position. This position is usually determined with capacitance gauges; however, more advanced systems use optical encoders or an additional Michelson interferometer. The important advantage of VSI is its ability to ambiguously measure heights that are greater than a fringe.

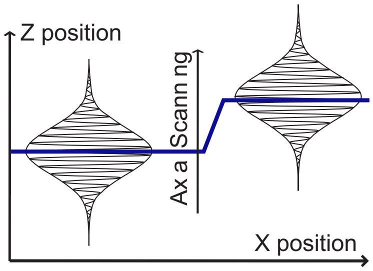

The PSI technique requires the object of interest to be within the coherence length and provides one order of magnitude or greater *z* resolution compared to VSI. To extend coherence length, an optical filter is introduced to limit the light source's bandwidth. With PSI, a phase-shifting component is located in one arm of the interferometer. The reference mirror can also provide the role of the phase shifter. A series of images (usually three to five) is acquired with appropriate phase differences to be used later in phase reconstruction. Introduced phase shifts change the location of the fringes. An important PSI drawback is the fact that phase maps are subject to 2π discontinuities. Phase maps are first obtained in the form of so-called phase fringes and must be unwrapped (a procedure of removing discontinuities) to obtain a continuous phase map.

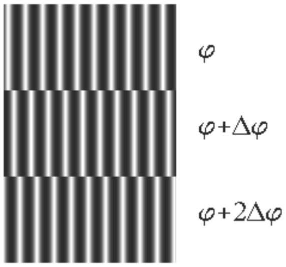

Note that modern short-coherence interference microscopes combine phase information from PSI with a long range of VSI methods.

Optical Profilometry: System Design

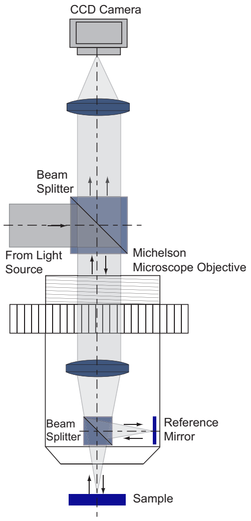
CCD Camera
Beam Splitter
From Light Source
Michelson Microscope Objective
Beam Splitter
Reference Mirror
Sample

Optical profilometry systems are based on implementing a Michelson interferometer or its derivatives into their design. There are three primary implementations: classical **Michelson** geometry, **Mirau**, and **Linnik**. The first approach incorporates a Michelson inside a microscope objective using a classical interferometer layout. Consequently, it can work only for low magnifications and short working distances. The Mireau objective uses two plates perpendicular to the optical axis inside the objective. One is a beam-splitting plate, while the other acts as a reference mirror. Such a solution extends the working distance and increases the possible magnifications.

Design	Magnification
Michelson	1× to 5×
Mireau	10× to 100×
Linnik	50× to 100×

The Linnik design utilizes two matching objectives. It does not suffer from NA limitation, but it is quite expensive and susceptible to vibrations.

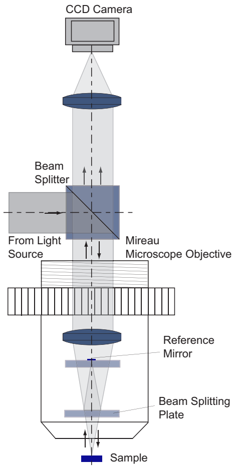
CCD Camera
Beam Splitter
From Light Source
Mireau Microscope Objective
Reference Mirror
Beam Splitting Plate
Sample

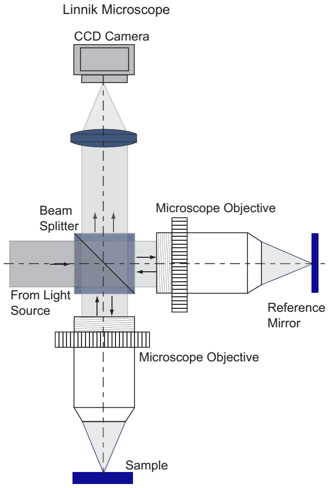
Linnik Microscope
CCD Camera
Beam Splitter
Microscope Objective
From Light Source
Reference Mirror
Microscope Objective
Sample

Phase-Shifting Algorithms

Phase-shifting techniques find the phase distribution from interference images given by

$$I(x,y) = a(x,y) + b(x,y)\cos(\varphi + n\Delta\varphi),$$

where $a(x, y)$ and $b(x, y)$ correspond to background and fringe amplitude, respectively. Since this equation has three unknowns, at least three measurements (images) are required. For image acquisition with a discrete CCD camera, spatial (x, y) coordinates can be replaced by (i, j) pixel coordinates.

The most basic PSF algorithms include the **three-**, **four-**, and **five-image** techniques.

I denotes the intensity for a specific image (1st, 2nd, 3rd, etc.) in the selected i, j pixel of a CCD camera. The phase shift for a three-image algorithm is π/2. Four- and five-image algorithms also acquire images with π/2 phase shifts.

$$\varphi = \arctan\left(\frac{I_3 - I_2}{I_1 - I_2}\right)$$

$$\varphi = \arctan\left(\frac{I_4 - I_2}{I_1 - I_3}\right)$$

$$\varphi = \arctan\left(\frac{2[I_2 - I_4]}{I_1 - I_3 + I_5}\right)$$

A reconstructed phase depends on the accuracy of the phase shifts. π/2 steps were selected to minimize the systematic errors of the above techniques. Accuracy progressively improves between the three- and five-point techniques. Note that modern PSI algorithms often use seven images or more, in some cases reaching thirteen.

After the calculations, the wrapped phase maps (modulo 2π) are obtained (arctan function). Therefore, unwrapping procedures have to be applied to provide continuous phase maps.

Common unwrapping methods include the line by line algorithm, least square integration, regularized phase tracking, minimum spanning tree, temporal unwrapping, frequency multiplexing, and others.

Structured Illumination: Axial Sectioning

A simple and inexpensive alternative to confocal microscopy is the **structured-illumination technique**. This method uses the principle that all but the zero (DC) spatial frequencies are attenuated with defocus. This observation provides the basis for obtaining the **optical sectioning** of images from a conventional wide-field microscope. A modified illumination system of the microscope projects a single spatial-frequency grid pattern onto the object. The microscope then faithfully images only that portion of the object where the grid pattern is in focus. The structured-illumination technique requires the acquisition of at least three images to remove the illumination structure and reconstruct an image of the layer. The reconstruction relation is described by

$$I = \sqrt{\left(I_0 - I_{2\pi/3}\right)^2 + \left(I_0 - I_{4\pi/3}\right)^2 + \left(I_{2\pi/3} - I_{4\pi/3}\right)^2}\ ,$$

where I denotes the intensity in the reconstructed image point, while I_0, $I_{2\pi/3}$, and $I_{4\pi/3}$ are intensities for the image point and the three consecutive positions of the sinusoidal illuminating grid (zero, 1/3 and 2/3 of the structure period).

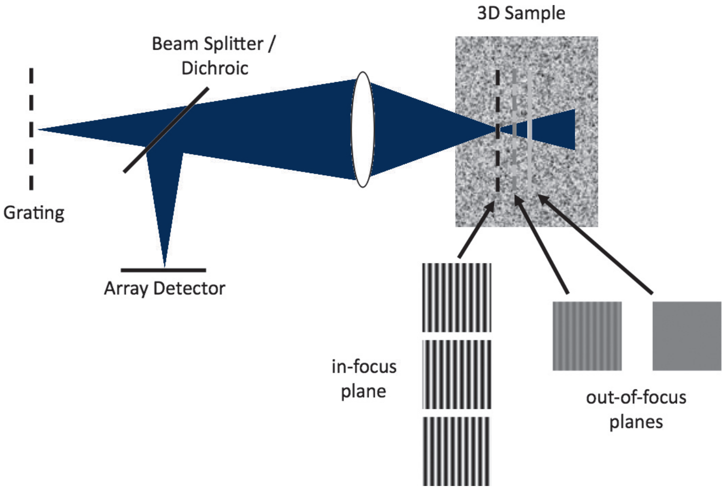

Note that the maximum sectioning is obtained for 0.5 of the objective's cut-off frequency. This sectioning ability is comparable to that obtained in confocal microscopy.

Structured Illumination: Resolution Enhancement

Structured illumination can be used to enhance the spatial resolution of a microscope. It is based on the fact that if a sample is illuminated with a periodic structure, it will create a low-frequency beating effect (Moiré fringes) with the features of the object. This new low-frequency structure will be capable of aliasing high spatial frequencies through the optical system. Therefore, the sample can be illuminated with a pattern of several orientations to accommodate the different directions of the object features. In practical terms, this means that system apertures will be doubled (in diameter), creating a new synthetic aperture.

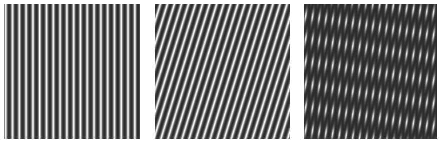

To produce a high-frequency structure, the interference of two beams with large illumination angles can be used. Note that the blue dots in the figure represent aliased spatial frequencies.

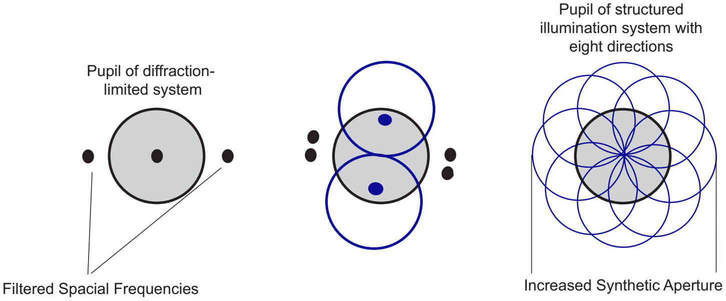
Pupil of diffraction-limited system
Pupil of structured illumination system with eight directions
Filtered Spacial Frequencies
Increased Synthetic Aperture

To reconstruct distribution, at least three images are required for one grid direction. In practice, obtaining a high-resolution image is possible with seven to nine images and a grid rotation, while a larger number can provide a more efficient reconstruction.

The linear-structured illumination approach is capable of obtaining two-fold resolution improvement over the diffraction limit. The application of nonlinear gain in fluorescence imaging improves resolution several times while working with higher harmonics. Sample features of 50 nm and smaller can be successfully resolved.

TIRF Microscopy

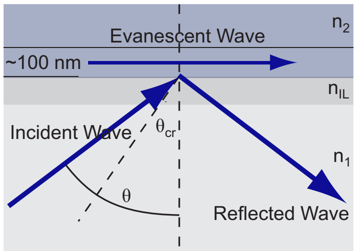

Total internal reflection fluorescence (TIRF) **microscopy** (also called evanescent wave microscopy) excites a limited width of the sample close to a solid interface. In TIRF, a thin layer of the sample is excited by a wavefront undergoing total internal reflection in the dielectric substrate, producing an **evanescent wave** propagating along the interface between the substrate and object.

The thickness of the excited section is limited to less than 100–200 nm. Very low background fluorescence is an important feature of this technique, since only a very limited volume of the sample in contact with the evanescent wave is excited.

This technique can be used for imaging cells, cytoplasmic filament structures, single molecules, proteins at cell membranes, micro-morphological structures in living cells, the adsorption of liquids at interfaces, or Brownian motion at the surfaces. It is also a suitable technique for recording long-term fluorescence movies.

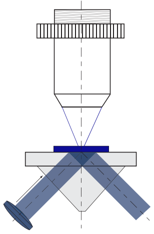

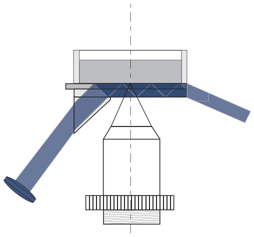

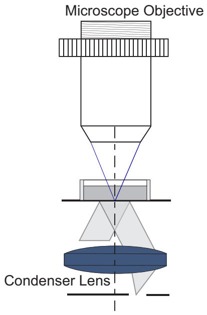

While an evanescent wave can be created without any layers between the dielectric substrate and sample, it appears that a thin layer (e.g., metal) improves image quality. For example, it can quench fluorescence in the first 10 nm and therefore provide selective fluorescence for the 10–200 nm region. TIRF can be easily combined with other modalities like optical trapping, multi-photon excitation, or interference. The most common configurations include oblique illumination through a high-NA objective or condenser and TIRF with a prism.

Solid Immersion

Optical resolution can be enhanced using **solid immersion** imaging. In the classical definition, the diffraction limit depends on the NA of the optical system and the wavelength of light. It can be improved by increasing the refractive index of the media between the sample and optical system or by decreasing the light's wavelength. The solid immersion principle is based on improving the first parameter by applying **solid immersion lenses** (SILs). To maximize performance, SILs are made with a high refractive index glass (usually greater than 2). The most basic setup for a SIL application involves a hemispherical lens. The sample is placed at the center of the hemisphere so the rays propagating through the system are not refracted (they intersect the SIL surface direction) and enter the microscope objective. The SIL is practically in contact with the object, but has a small sub-wavelength gap (<100 nm) between the sample and optical system; therefore, an object is always in an evanescent field and can be imaged with high resolution. Therefore, the technique is confined to a very thin sample volume (up to 100 nm) and can provide optical sectioning. Systems based on solid immersion can reach sub-100-nm lateral resolutions.

The most common solid-immersion applications are in microscopy, including fluorescence, optical data storage, and lithography. Compared to classical oil-immersion techniques, this technique is dedicated to imaging thin samples only (sub-100 nm), but provides much better resolution depending on the configuration and refractive index of the SIL.

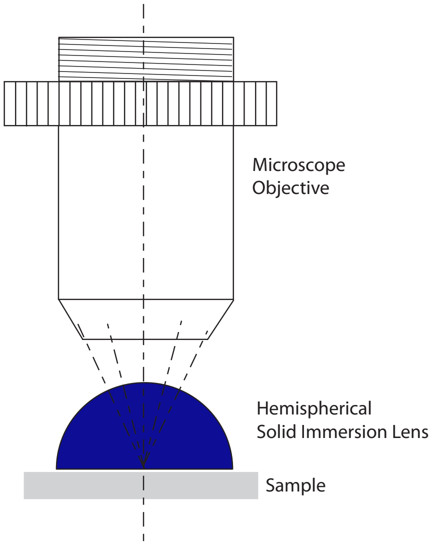

Stimulated Emission Depletion

Stimulated emission depletion (STED) microscopy is a fluorescence super-resolution technique that allows imaging with a spatial resolution at the molecular level. The technique has shown an improvement 5–10 times beyond the possible diffraction-resolution limit.

The STED technique is based on the fact that the area around the excited object point can be treated with a pulse (depletion pulse or STED pulse) that depletes high-energy states and brings fluorescent dye to the ground state. Consequently, an actual excitation pulse excites only the small sub-diffraction resolved area. The depletion pulse must have high intensity and be significantly shorter than the lifetime of the excited states. The pulse must be shaped, for example, with a half-wave phase plate and imaging optics to create a 3D doughnut-like structure around the point of interest. The STED pulse is shifted toward red with regard to the fluorescence excitation pulse and follows it by a few picoseconds. To obtain the complete image, the system scans in the *x*, *y*, and *z* directions.

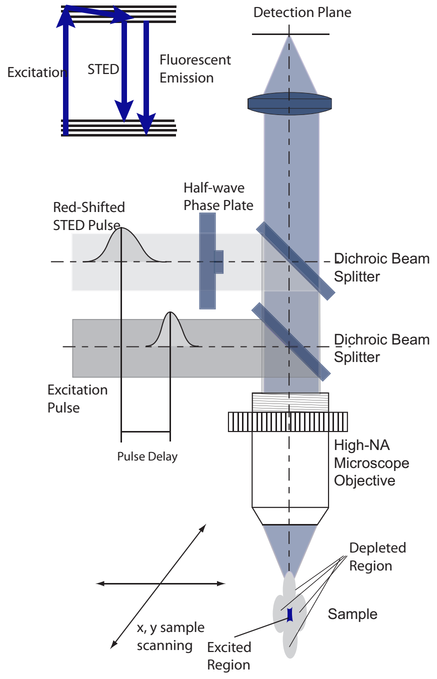
Excitation
STED
Fluorescent Emission
Detection Plane
Half-wave Phase Plate
Red-Shifted STED Pulse
Dichroic Beam Splitter
Dichroic Beam Splitter
Excitation Pulse
Pulse Delay
High-NA Microscope Objective
Depleted Region
Sample
x, y sample scanning
Excited Region

STORM

Stochastic optical reconstruction microscopy (STORM) is based on the application of multiple photo-switchable fluorescent probes that can be easily distinguished in the spectral domain. The system is configured in **TIRF** geometry. Multiple fluorophores label a sample, and each is built as a dye pair where one component acts as a fluorescence activator. The activator can be switched on or deactivated using a different illumination laser (a longer wavelength is used for deactivation—usually red). In the case of applying several dual-pair dyes, different lasers can be used for activation and only one for deactivation.

A sample can be sequentially illuminated with activation lasers, which generate fluorescent light at different wavelengths for slightly different locations on the object (each dye is sparsely distributed over the sample and activation is stochastic in nature). Activation for a different color can also be shifted in time. After performing several sequences of activation/deactivation, it is possible to acquire data for the centroid localization of various diffraction-limited spots. This localization can be performed with precision to about 1 nm. The spectral separation of dyes detects closely located object points encoded with a different color.

The STORM process requires 1000+ activation/deactivation cycles to produce high-resolution images. Therefore, final images combine the spots corresponding to individual molecules of DNA or an antibody used for labeling. STORM images can reach a lateral resolution of about 25 nm and an axial resolution of 50 nm.

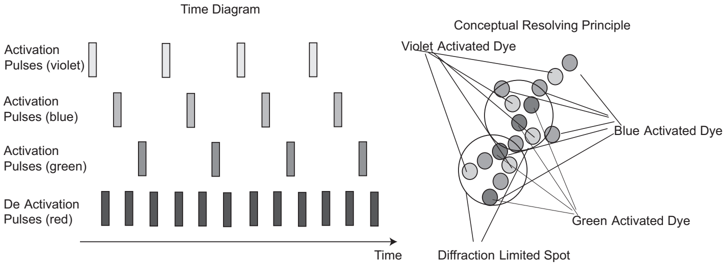
Time Diagram
Activation Pulses (violet)
Activation Pulses (blue)
Activation Pulses (green)
De Activation Pulses (red)
Time
Conceptual Resolving Principle
Violet Activated Dye
Blue Activated Dye
Green Activated Dye
Diffraction Limited Spot

4Pi Microscopy

4Pi microscopy is a fluorescence technique that significantly reduces the thickness of an imaged section. The 4Pi principle is based on coherent illumination of a sample from two directions. Two constructively interfering beams improve the axial resolution by three to seven times. The possible values of z resolution using 4Pi microscopy are 50–100 nm and 30–50 nm when combined with STED.

The undesired effect of the technique is the appearance of side lobes located above and below the plane of the imaged section. They are located about λ/2 from the object. To eliminate side lobes, three primary techniques can be used:

- Combine 4Pi with confocal detection. A pinhole rejects some of the out-of-focus light.
- Detect in multi-photon mode, which quickly diminishes the excitation of fluorescence.
- Apply a modified 4Pi system, which creates interference at both the object and detection planes (two imaging systems are required).

It is also possible to remove side lobes through numerical processing, if they are smaller than 50% of the maximum intensity. However, side lobes increase with the NA of the objective; for an NA of 1.4, they are about 60–70% of the maximum intensity. Numerical correction is not effective in this case.

4Pi microscopy improves resolution only in the z direction; however, the perception of details in the thin layers is a useful benefit of the technique.

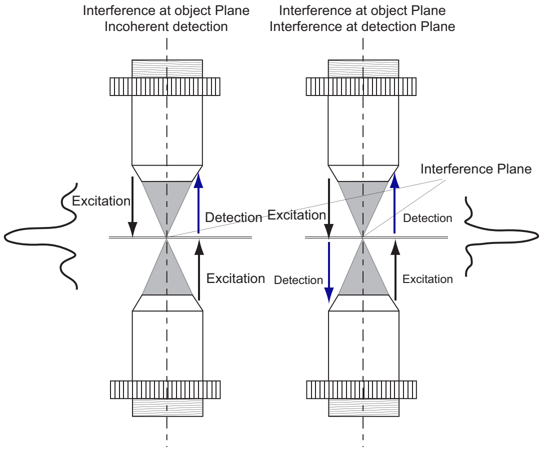
Interference at object Plane
Incoherent detection
Interference at object Plane
Interference at detection Plane
Interference Plane
Excitation
Detection
Excitation
Excitation
Detection
Detection
Excitation

The Limits of Light Microscopy

Classical light microscopy is limited by diffraction and depends on the wavelength of light and the parameters of an optical system (NA). However, recent developments in super-resolution methods and enhancement techniques overcome these barriers and provide resolution at a molecular level. They often use a sample as a component of the optical train [resolvable saturable optical fluorescence transitions (**RESOLFT**), saturated structured-illumination microscopy (**SSIM**), photoactivated localization microscopy (**PALM**) and **STORM**] or combine near-field effects (**4Pi, TIRF, solid immersion**) with far-field imaging. The table below summarizes these methods.

		Demonstrated Values	
Method	**Principles**	**Lateral**	**Axial**
Bright field	Diffraction	200 nm	500 nm
Confocal	Diffraction (slightly better than bright field)	200 nm	500 nm
Solid immersion	Diffraction / evanescent field decay	< 100 nm	< 100 nm
TIRF	Diffraction / evanescent field decay	200 nm	< 100 nm
4Pi / I5	Diffraction/ interference	200 nm	50 nm
RESOLFT (e.g.,STED)	Depletion, molecular structure of sample–fluorescent probes	20 nm	20 nm
Structured illumination (SSIM)	Aliasing, nonlinear gain in fluorescent probes (molecular structure)	25–50 nm	50–100 nm
Stochastic techniques (PALM, STORM)	Fluorescent probes (molecular structure), centroid localization, time-dependant fluorophore activation	25 nm	50 nm

Raman and CARS Microscopy

Raman microscopy (also called chemical imaging) is a technique based on inelastic **Raman scattering** and evaluates the vibrational properties of samples (minerals, polymers, and biological objects).

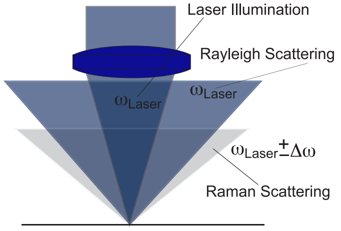

Raman microscopy uses a laser beam focused on a solid object. Most of the illuminating light is scattered, reflected, and transmitted, and preserves the parameters of the illumination beam (the frequency is the same as the illumination). However, a small portion of the light is subject to a shift in frequency: $\omega_{Raman} = \omega_{laser} \pm \Delta\omega$. This frequency shift between illumination and scattering bears information about the molecular composition of the sample. Raman scattering is weak and requires high-power sources and sensitive detectors. It is examined with spectroscopy techniques.

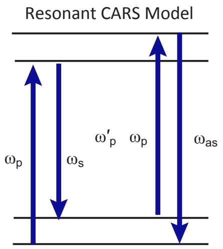

Coherent Anti-Stokes Raman Scattering (**CARS**) overcomes the problem of weak signals associated with Raman imaging. It uses a pump laser and a tunable Stokes laser to stimulate a sample with a four-wave mixing process. The fields of both lasers [pump field $E(\omega_p)$, Stokes field $E(\omega_s)$, and probe field $E'(\omega_p')$] interact with the sample and produce an anti-Stokes field [$E(\omega_{as})$] with frequency ω_{as}, so $\omega_{as} = 2\omega_p - \omega_s$.

CARS works as a resonant process, only providing a signal if the vibrational structure of the sample specifically matches $\omega_p - \omega_s$. It also must assure phase matching so that l_c (the coherence length) is greater than $\pi/\Delta\mathbf{k} = \pi/\left|\mathbf{k}_{as} - (2\mathbf{k}_p - \mathbf{k}_s)\right|$, where $\Delta\mathbf{k}$ is a phase mismatch. CARS is orders of magnitude stronger than Raman scattering, does not require any exogenous contrast, provides good sectioning ability, and can be set up in both transmission and reflectance modes.

SPIM

Selective plane illumination microscopy (SPIM) is dedicated to the high-resolution imaging of large 3D samples. It is based on three principles:

- The sample is illuminated with a **light sheet**, which is obtained with cylindrical optics. The light sheet is a beam focused in one direction and collimated in another. This way, the thin and wide light sheet can pass through the object of interest (see figure).
- The sample is imaged in the direction perpendicular to the illumination.
- The sample is rotated around its axis of gravity and linearly translated into axial and lateral directions.

Data is recorded by a CCD camera for multiple object orientations and can be reconstructed into a single 3D distribution. Both scattered and fluorescent light can be used for imaging.

The maximum resolution of the technique is limited by the longitudinal resolution of the optical system (for a high numerical aperture, objectives can obtain micron-level values). The maximum volume imaged is limited by the working distance of a microscope and can be as small as tens of microns or exceed several millimeters. The object is placed in an air, oil, or water chamber.

The technique is primarily used for bio-imaging, ranging from small organisms to individual cells.

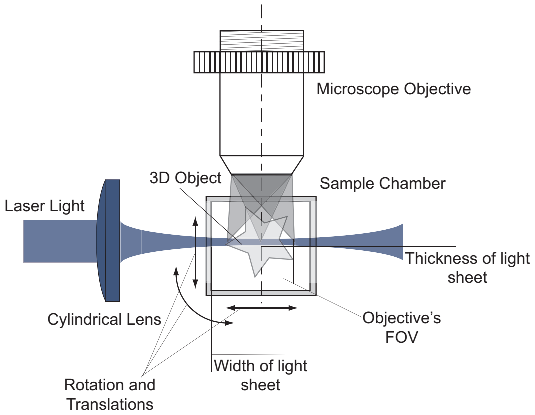
Microscope Objective
3D Object
Sample Chamber
Laser Light
Thickness of light sheet
Cylindrical Lens
Objective's FOV
Width of light sheet
Rotation and Translations

Array Microscopy

An **array microscope** is a solution to the trade-off between field of view and lateral resolution. In the array microscope, a miniature microscope objective is replicated tens of times. The result is an imaging system with a field of view that can be increased in steps of an individual objective's field of view, independent of numerical aperture and resolution. An array microscope is useful for applications that require fast imaging of large areas at a high level of detail. Compared to conventional microscope optics for the same purpose, an array microscope can complete the same task several times faster due to its parallel imaging format.

An example implementation of the array-microscope optics is shown. This array microscope is used for imaging glass slides containing tissue (histology) or cells (cytology). For this case, there are a total of 80 identical miniature 7×/0.6 microscope objectives. The summed field of view measures 18 mm. Each objective consists of three lens elements. The lens surfaces are patterned on three plates that are stacked to form the final array of microscopes. The plates measure 25 mm in diameter. Plate 1 is near the object at a working distance of 400 μm. Between plates 2 and 3 is a baffle plate. A second baffle is located between the third lens and the image plane. The purpose of the baffles is to eliminate cross talk and image overlap between objectives in the array.

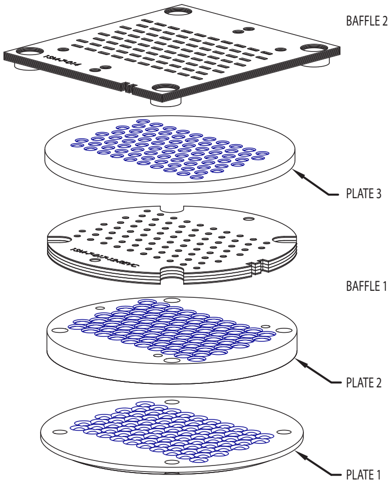

The small size of each microscope objective and the need to avoid image overlap jointly dictate a low magnification. Therefore, the array microscope works best in combination with an image sensor divided into small pixels (e.g., 3.3 μm).

Focusing the array microscope is achieved by an up/down translation and two rotations: a pitch and a roll.

Digital Microscopy

Digital microscopy emerged with the development of electronic array image sensors and replaced the microphotography techniques previously used in microscopy. It can also work with point detectors when images are recombined in post or real-time processing. Digital microscopy is based on acquiring, storing, and processing images taken with various microscopy techniques. It supports applications that require:

- Combining data sets acquired for one object with different microscopy modalities or different imaging conditions. Data can be used for further analysis and processing in techniques like hyperspectral and polarization imaging.
- Image correction (e.g., distortion, white balance correction, or background subtraction). Digital microscopy is used in deconvolution methods that perform numerical rejection of out-of-focus light and iteratively reconstruct an object's estimate.
- Image acquisition with a high temporal resolution. This includes short integration times or high frame rates.
- Long time experiments and remote image recording.
- Scanning techniques where the image is reconstructed after point-by-point scanning and displayed on the monitor in real time.
- Low light, especially fluorescence. Using high-sensitivity detectors reduces both the excitation intensity and excitation time, which mitigates photobleaching effects.
- Contrast enhancement techniques and an improvement in spatial resolution. Digital microscopy can detect signal changes smaller than possible with visual observation.
- Super-resolution techniques that may require the acquisition of many images under different conditions.
- High throughput scanning techniques (e.g., imaging large sample areas).
- UV and IR applications not possible with visual observation.

The primary detector used for digital microscopy is a CCD camera. For scanning techniques, a photomultiplier or photodiodes are used.

Principles of CCD Operation

Charge-coupled device (CCD) sensors are a collection of individual photodiodes arranged in a 2D grid pattern. Incident photons produce electron-hole pairs in each pixel in proportion to the amount of light on each pixel. By collecting the signal from each pixel, an image corresponding to the incident light intensity can be reconstructed.

Here are the step-by-step processes in a CCD:

1. The CCD array is illuminated for integration time τ.
2. Incident photons produce mobile charge carriers.
3. Charge carriers are collected within the p-n structure.
4. The illumination is blocked after time τ (full-frame only).
5. The collected charge is transferred from each pixel to an amplifier at the edge of the array.
6. The amplifier converts the charge signal into voltage.
7. The analog voltage is converted to a digital level for computer processing and image display.

Each pixel (**photodiode**) is reverse biased, causing photoelectrons to move towards the positively charged electrode. Voltages applied to the electrodes produce a potential well within the semiconductor structure. During the integration time, electrons accumulate in the potential well, up to the full-well capacity. The full-well capacity is reached when the repulsive force between electrons exceeds the attractive force of the well. At the end of the exposure time, each pixel has stored a number of electrons in proportion to the amount of light received. These charge packets must be transferred from the sensor, from each pixel to a single amplifier without loss. This is accomplished by a series of parallel and serial shift registers. The charge stored by each pixel is transferred across the array by sequences of gating voltages applied to the electrodes. The packet of electrons follows the positive clocking waveform voltage from pixel-to-pixel or row-to-row. A potential barrier is always maintained between adjacent pixel charge packets.

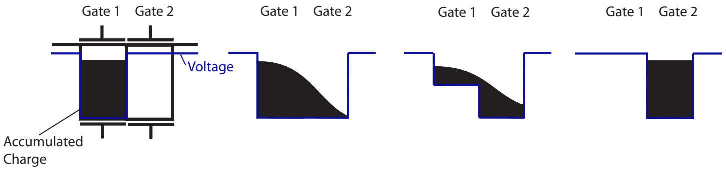

CCD Architectures

In a **full-frame architecture**, individual pixel charge packets are transferred by a parallel row shift to the serial register, then one-by-one to the amplifier. The advantage of this approach is a 100% photosensitive area, while the disadvantage is that a shutter must block the sensor area during readout, and consequently limits the frame rate.

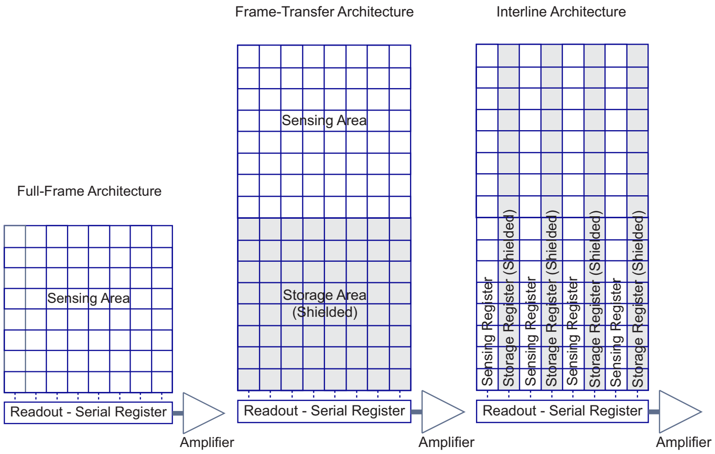
Full-Frame Architecture
Sensing Area
Readout - Serial Register
Amplifier
Frame-Transfer Architecture
Sensing Area
Storage Area
(Shielded)
Readout - Serial Register
Amplifier
Interline Architecture
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Sensing Register
Storage Register (Shielded)
Readout - Serial Register
Amplifier

The **frame-transfer architecture** has the advantage of not needing to block an imaging area during readout. It is faster than full-frame, since readout can take place during the integration time. The significant drawback is that only 50% of the sensor area is used for imaging.

The **interline transfer architecture** uses columns with exposed imaging pixels interleaved with columns of masked storage pixels. A charge is transferred into an adjacent storage column for readout. The advantage of such an approach is a high frame rate due to a rapid transfer time (< 1 ms), while again the disadvantage is that 50% of the sensor area is used for light collection. However, recent interleaved CCDs have an increased fill factor due to the application of microlens arrays.

CCD Architectures (cont.)

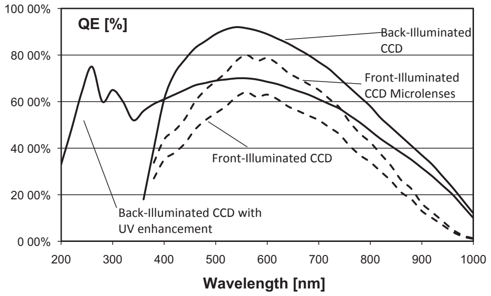

Quantum efficiency (QE) is the percentage of incident photons at each wavelength that produce an electron in a photodiode. CCDs have a lower QE than individual silicon photodiodes due to the charge-transfer channels on the sensor surface, which reduces the effective light-collection area. A typical CCD QE is 40–60%. This can be improved for back-thinned CCDs, which are illuminated from behind; this avoids the surface electrode channels and improves QE to approximately 90–95%.

Recently, **electron-multiplying CCDs** (EM-CCD), primarily used for biological imaging, have become more common. They are equipped with a gain register placed between the shift register and output amplifier. A multi-stage gain register allows high gains. As a result, single electrons can generate thousands of output electrons. While the read noise is usually low, it is therefore negligible in the context of thousands of signal electrons.

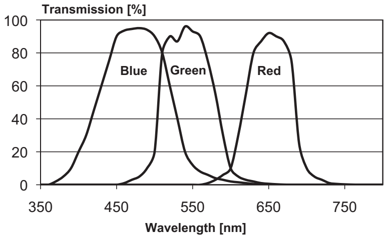

CCD pixels are not color sensitive, but use color filters to separately measure red, green, and blue light intensity. The most common method is to cover the sensor with an array of **RGB** (red, green, blue) filters, which combine four individual pixels to generate a single color pixel. The **Bayer mask** uses two green pixels for every red and blue pixel, which simulates human visual sensitivity.

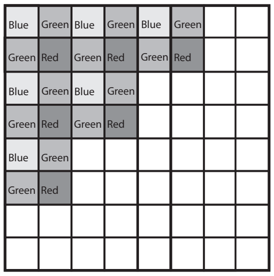

Blue	Green	Blue	Green	Blue	Green		
Green	Red	Green	Red	Green	Red		
Blue	Green	Blue	Green				
Green	Red	Green	Red				
Blue	Green						
Green	Red						

CCD Noise

The three main types of noise that affect CCD imaging are **dark noise**, **read noise**, and **photon noise**.

Dark noise arises from random variations in the number of thermally generated electrons in each photodiode. More electrons can reach the conduction band as the temperature increases, but this number follows a statistical distribution. These random fluctuations in the number of conduction band electrons results in a **dark current** that is not due to light. To reduce the influence of dark current for long time exposures, CCD cameras are cooled (which is crucial for the exposure of weak signals, like fluorescence, with times of a few seconds and more).

Read noise describes the random fluctuation in electrons contributing to measurement, due to electronic processes on the CCD sensor. This noise arises during the charge transfer, the charge-to-voltage conversion, and the analog-to-digital conversion. Every pixel on the sensor is subject to the same level of read noise, most of which is added by the amplifier.

Dark noise and read noise are due to the properties of the CCD sensor itself.

Photon noise (or **shot noise**) is inherent in any measurement of light, due to the fact that photons arrive at the detector randomly in time. This process is described by **Poisson statistics**. The probability P of measuring k events, given an expected value N is

$$P\left(k,N\right)=\frac{N^{k}e^{-N}}{k!}\,.$$

The standard deviation of the Poisson distribution is $N^{1/2}$. Therefore, if the average number of photons arriving at the detector is N, then the noise is $N^{1/2}$. Since the average number of photons is proportional to incident power, shot noise increases with $P^{1/2}$.

Signal-to-Noise Ratio and the Digitization of CCD

Total **noise** as a function of the number of electrons from all three contributing noises is given by

$$Noise\left[N_{\text{electrons}}\right] = \sqrt{\sigma^2_{\text{Photon}} + \sigma^2_{\text{Dark}} + \sigma^2_{\text{Read}}}\ ,$$

where

$\sigma_{\text{Photon}} = \sqrt{\varphi\eta\tau}$, $\sigma_{\text{Dark}} = \sqrt{I_{\text{Dark}}\tau}$, and $\sigma_{\text{Read}} = N_{\text{R}}$, I_{Dark} is dark current (electrons per second), and N_{R} is read noise (electrons, rms).

The quality of an image is determined by the **signal-to-noise ratio** (SNR). The signal can be defined as

$$Signal\left[N_{\text{electrons}}\right] = \varphi\eta\tau\ ,$$

where φ is the incident photon flux at the CCD (photons per second), η is the quantum efficiency of the CCD (electrons per photon), and τ is the integration time (in seconds). Therefore, SNR can be defined as:

$$SNR = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau + I_{\text{dark}}\tau + N^2_{\text{R}}}}\ .$$

It is best to use a CCD under photon-noise-limited conditions. If possible, it would be optimal to increase the integration time to increase the SNR and work in photon-noise-limited conditions. However, an increase in integration time is possible until reaching full-well capacity (saturation level):

$$SNR = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau}} = \sqrt{\varphi\eta\tau}.$$

The **dynamic range** can be derived as a ratio of full-well capacity and read noise. Digitization of the CCD output should be performed to maintain the dynamic range of the camera. Therefore, an analog-to-digital converter should support (at least) the same number of gray levels calculated by the CCD's dynamic range. Note that a high bit depth extends readout time, which is especially critical for large-format cameras.

CCD Sampling

The maximum spatial frequency passed by the CCD is one half of the sampling frequency—the **Nyquist frequency**. Any frequency higher than Nyquist will be aliased at lower frequencies.

Undersampling refers to a frequency where the sampling rate is not sufficient for the application. To assure maximum resolution of the microscope, the system should be optics limited, and the CCD should provide sampling that reaches the diffraction resolution limit. In practice, this means that at least two pixels should be dedicated to a distance of the resolution. Therefore, the maximum pixel spacing that provides the diffraction limit can be estimated as

$$d_{\text{pix}} = M\frac{0.61\lambda}{2NA},$$

where M is the magnification between the object and the CCD's plane. A graph representing how sampling influences the MTF of the system, including the CCD, is presented below.

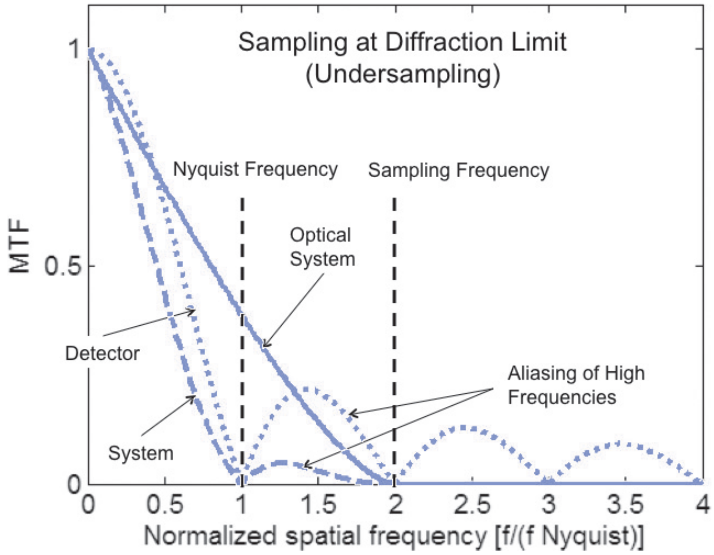

CCD Sampling (cont.)

Oversampling means that more than the minimum number of pixels, according to the Nyquist criteria, are available for detection, which does not imply excessive sampling of the image. However, excessive sampling decreases the available field of view of a microscope. The relation between the extent of field D, the number of pixels in the x and y directions (N_x and N_y, respectively), and the pixel spacing d_{pix} can be calculated from

$$D_x = \frac{N_x d_{\text{pix}-x}}{M}$$

and

$$D_y = \frac{N_y d_{\text{pix}-y}}{M} .$$

This assumes that the object area imaged by the microscope is equal to or larger than D. Therefore, the size of the microscope's field of view and the size of the CCD chip should be matched.

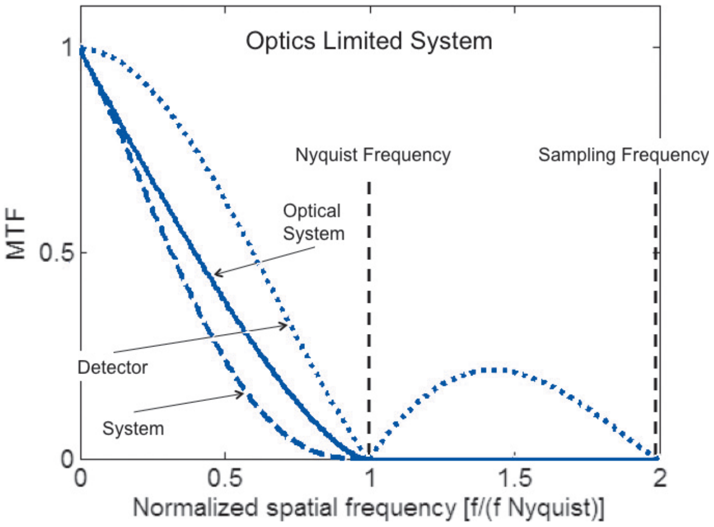

Equation Summary

Quantized Energy:

$$E = h\nu = hc/\lambda$$

Propagation of an electric field and wave vector:

$$E = A\sin\left(\omega t - kz + \varphi_{\mathrm{o}}\right) = A\exp\left[-i\left(\omega t - kz + \varphi_{\mathrm{o}}\right)\right]$$

$$\omega = \frac{2\pi}{T} = \frac{2\pi V_{\mathrm{m}}}{\lambda} \qquad E\left(z,t\right) = E_x + E_y$$

$$E_x = A_x \exp\left[-i\left(\omega t - kz + \varphi_x\right)\right] \qquad E_y = A_y \exp\left[-i\left(\omega t - kz + \varphi_y\right)\right]$$

Refractive index:

$$n = \frac{c}{V_{\mathrm{m}}}$$

Optical path length:

$$\mathrm{OPL} = nL$$

$$\mathrm{OPL} = \int_{P1}^{P2} n\,ds$$

$$ds^2 = dx^2 + dy^2 + dz^2$$

Optical path difference and phase difference:

$$\mathrm{OPD} = n_1 L_1 - n_2 L_2$$

$$\Delta\varphi = 2\pi\frac{\mathrm{OPD}}{\lambda}$$

TIR:

$$\theta_{\mathrm{cr}} = \arcsin\left(\frac{n_1}{n_2}\right)$$

$$I = I_{\mathrm{o}}\exp\left(-y/d\right)$$

$$d = \frac{\lambda}{4\pi n_1}\frac{1}{\sqrt{\sin^2\theta - \sin^2\theta_{\mathrm{c}}}}$$

Coherence length:

$$l_{\mathrm{c}} = \frac{\lambda^2}{\Delta\lambda}$$

Equation Summary (cont'd)

Two-beam interference:

$$I = EE^*$$

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(\Delta\varphi)$$

$$\Delta\varphi = \varphi_2 - \varphi_1.$$

Contrast:

$$C = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$

Diffraction grating equation:

$$m\lambda = d(\sin\beta \pm \sin\alpha)$$

Resolving power of diffraction grating:

$$\frac{\lambda}{\Delta\lambda} = mN$$

Free spectral range:

$$\Delta\lambda = \frac{m+1}{m}\lambda_1 - \lambda_1 = \frac{\lambda_1}{m}.$$

Newtonian equation:

$$xx' = ff'$$

$$xx' = -f'^2$$

Gaussian imaging equation:

$$\frac{n'}{z'} = \frac{n}{z} + \frac{1}{f_e}$$

$$\frac{1}{z'} = \frac{1}{z} + \frac{1}{f_e}$$

$$f_e = \frac{1}{\varphi} = -\frac{f}{n} = \frac{f'}{n'}$$

Transverse magnification:

$$M = \frac{h'}{h} = -\frac{x'}{f'} = -\frac{f}{x} = \frac{z'}{z} = \frac{-f}{z-f} = \frac{-(z'-f')}{f'}$$

Longitudinal magnification:

$$\frac{\Delta z'}{\Delta z} = \left(-\frac{f'}{f}\right)M_1 M_2$$

Equation Summary (cont'd)

Optical transfer function:

$$\mathrm{OTF}(\rho) = \mathrm{MTF}(\rho)\exp\left[i\varphi(\rho)\right]$$

Modulation transfer function:

$$\mathrm{MTF}(\rho) = \frac{C(\rho)_{\mathrm{image}}}{C(\rho)_{\mathrm{object}}}$$

Field of view of microscope:

$$\mathrm{FOV} = \frac{FieldNumber}{M_{\mathrm{objective}}}[\mathrm{mm}]$$

Magnifying power:

$$\mathrm{MP} = \frac{u'}{u} = \frac{d_o\left(f - z'\right)}{f\left(z' - l\right)}$$

$$MP \approx \frac{250\ \mathrm{mm}}{f}$$

Magnification of the microscope objective:

$$M_{\mathrm{objective}} = \frac{OTL}{f_{\mathrm{objective}}}$$

Magnifying power of the microscope:

$$\mathrm{MP}_{\mathrm{microscope}} = M_{\mathrm{objective}}\mathrm{MP}_{\mathrm{eyepiece}} = -\frac{\mathrm{OTL}}{f_{\mathrm{objective}}}\frac{250\,\mathrm{mm}}{f_{\mathrm{eyepiece}}}$$

Numerical aperture:

$$NA = n\sin u$$

$$NA = NA'M_{\mathrm{objective}}$$

Airy disk:

$$d = \frac{1.22\lambda}{n\sin u} = \frac{1.22\lambda}{NA}$$

Rayleigh resolution limit:

$$d = \frac{0.61\lambda}{n\sin u} = \frac{0.61\lambda}{NA}$$

Sparrow resolution limit:

$$d = 0.5\lambda/NA$$

Equation Summary (cont'd)

Abbe resolution limit:

$$d = \frac{\lambda}{NA_{\text{objective}} + NA_{\text{condenser}}}$$

Resolving power of the microscope:

$$d_{\text{mic}} = \frac{d_{\text{eye}}}{M_{\text{min}}} = \frac{d_{\text{eye}}}{M_{\text{obj}} M_{\text{eyepiece}}}$$

Depth of focus and depth of field:

$$DOF = 2\Delta z' = \frac{n'\lambda}{\mathrm{NA}'^2}$$

$$2\Delta z' = M^2_{\text{objective}} 2\Delta z \frac{n'}{n}$$

Depth perception of stereoscopic microscope:

$$\delta z = \frac{250\alpha_s}{M_{\text{microscope}} \tan\gamma} [\text{mm}]$$

Minimum perceived phase in phase contrast:

$$\Delta\varphi_{\min} = \frac{C_{\text{ph-min}}\lambda}{4\pi\sqrt{N}}$$

Lateral resolution of the phase contrast:

$$d = f'_{\text{objective}} \frac{\lambda}{\left(r_{\text{AS}} + r_{\text{PR}}\right)}$$

Intensity in DIC:

$$I \propto \sin^2\left(\frac{\pi s \dfrac{d\varphi_o}{dx}}{\lambda}\right)$$

$$I \propto \sin^2 \frac{\pi\left(\Delta_b \pm s\dfrac{d\varphi_o}{dx}\right)}{\lambda}$$

Retardation:

$$\Gamma = \left(n_e - n_o\right)t$$

Equation Summary (cont'd)

Birefringence:

$$\delta = \frac{2\pi OPD}{\lambda} = \frac{2\pi\Gamma}{\lambda}$$

Resolution of a confocal microscope:

$$d_{xy} \approx \frac{0.4\lambda}{NA},$$

$$d_z \approx \frac{1.4n\lambda}{NA^2}$$

Confocal pinhole width:

$$D_{\text{pinhole}} = \frac{0.5\lambda M}{NA}$$

Fluorescent emission:

$$F = \sigma QI$$

Probability of two-photon excitation:

$$n_{\text{a}} \propto \delta\left(\frac{P_{\text{avg}}^2}{\tau\nu^2}\right)\left(\pi\frac{NA^2}{hc\lambda}\right)^2$$

Intensity in FD-OCT:

$$I(k, z_{\text{o}}) = \int A_{\text{R}} A_{\text{S}}(z)\cos\left[k(z_{\text{m}} - z_{\text{o}})\right]dz$$

Phase-shifting equation:

$$I(x, y) = a(x, y) + b(x, y)\cos(\varphi + n\Delta\varphi)$$

Three-image algorithm (π/2 shifts):

$$\varphi = \arctan\left(\frac{I_3 - I_2}{I_1 - I_2}\right)$$

Four-image algorithm:

$$\varphi = \arctan\left(\frac{I_4 - I_2}{I_1 - I_3}\right)$$

Five-image algorithm:

$$\varphi = \arctan\left(\frac{2[I_2 - I_4]}{I_1 - I_3 + I_5}\right)$$

Equation Summary (cont'd)

Image reconstruction structure illumination (sectioning):

$$I = \sqrt{\left(I_0 - I_{2\pi/3}\right)^2 + \left(I_0 - I_{4\pi/3}\right)^2 + \left(I_{2\pi/3} - I_{4\pi/3}\right)^2}$$

Poisson statistics:

$$P(k, N) = \frac{N^k e^{-N}}{k!}$$

Noise:

$$\mathit{Noise}\left[N_{\text{electrons}}\right] = \sqrt{\sigma_{\text{Photon}}^2 + \sigma_{\text{Dark}}^2 + \sigma_{\text{Read}}^2}$$

$$\sigma_{\text{Photon}} = \sqrt{\varphi\eta\tau}\ ,\ \sigma_{\text{Dark}} = \sqrt{I_{\text{Dark}}\tau}\ ,\ \sigma_{\text{Read}} = N_{\text{R}}\ ,\ I_{\text{Dark}}$$

Signal-to-noise ratio (SNR):

$$\text{Signal}\left[N_{\text{electrons}}\right] = \varphi\eta\tau$$

$$\text{SNR} = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau + I_{\text{dark}}\tau + N_{\text{R}}^2}}$$

$$\text{SNR} = \frac{\varphi\eta\tau}{\sqrt{\varphi\eta\tau}} = \sqrt{\varphi\eta\tau}$$

Bibliography

M. Bates, B. Huang, G.T. Dempsey, and X. Zhuang, “Multicolor Super-Resolution Imaging with Photo-Switchable Fluorescent Probes.” *Science* **317**, 1749 (2007).

J. R. Benford, Microscope objectives, Chapter 4 (p. 178), in *Applied Optics and Optical Engineering, Vol. III,* R. Kingslake, ed., Academic Press, New York, NY (1965).

M. Born and E. Wolf, *Principles of Optics*, Sixth Edition, Cambridge University Press, Cambridge, UK (1997).

S. Bradbury and P. J. Evennett, *Contrast Techniques in Light Microscopy*, BIOS Scientific Publishers, Oxford, UK (1996).

T. Chen, T. Milster, S. K. Park, B. McCarthy, D. Sarid, C. Poweleit, and J. Menendez, “Near-field solid immersion lens microscope with advanced compact mechanical design.” *Optical Engineering* **45**(10), 103002 (2006).

T. Chen, T. D. Milster, S. H. Yang, and D. Hansen, “Evanescent imaging with induced polarization by using a solid immersion lens.” *Optics Letters* **32**(2), 124–126 (2007).

J.-X. Cheng and X. S. Xie, “Coherent anti-stokes raman scattering microscopy: instrumentation, theory, and applications.” *J. Phys. Chem. B* **108***,* 827-840 (2004).

M. A. Choma, M. V. Sarunic, C. Yang, and J. A. Izatt, “Sensitivity advantage of swept source and Fourier domain optical coherence tomography.” *Opt. Express* **11**, 2183–2189 (2003).

J. F. de Boer, B. Cense, B. H. Park, M.C. Pierce, G. J. Tearney, and B. E. Bouma, “Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography.” *Opt Lett* **28**, 2067–2069 (2003).

E. Dereniak, materials for SPIE Short Course on Imaging Spectrometers, SPIE, Bellingham, WA (2005).

E. Dereniak, *Geometrical Optics*, Cambridge University Press, Cambridge, UK (2008).

M. Descour, materials for OPTI 412, “Optical Instrumentation,” University of Arizona (2000).

Bibliography

D. Goldstein, *Polarized Light*, Second Edition, Marcel Dekker, New York, NY (1993).

D. S. Goodman, Basic optical instruments, Chapter 4, in Geometrical and Instrumental Optics, D. Malacara, ed., Academic Press, New York, NY (1988).

J. Goodman, *Introduction to Fourier Optics*, 3rd Edition, Roberts and Company Publishers, Greenwood Village, CO (2004).

E. P. Goodwin and J. C. Wyant, *Field Guide to Interferometric Optical Testing*, SPIE Press, Bellingham, WA (2006).

J. E. Greivenkamp, *Field Guide to Geometrical Optics*, SPIE Press, Bellingham, WA (2004).

H. Gross, F. Blechinger, and B. Achtner, *Handbook of Optical Systems*, *Vol. 4*: *Survey of Optical Instruments*, Wiley-VCH, Germany (2008).

M. G. L. Gustafsson, "Nonlinear structured-illumination microscopy: Wide-field fluorescence imaging with theoretically unlimited resolution." *PNAS*, **102**(37) 13081–13086 (2005).

M. G. L. Gustafsson, "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy." *Journal of Microscopy*, **198**(2), 82-87 (2000).

Gerd Häusler and Michael Walter Lindner, ""Coherence Radar" and "Spectral Radar"—New tools for dermatological diagnosis." *Journal of Biomedical Optics* **3**(1), 21–31 (1998).

E. Hecht, *Optics*, Fourth Edition, Addison-Wesley, Upper Saddle River, New Jersey (2002).

S. W. Hell, "Far-field optical nanoscopy." *Science* **316**, 1153 (2007).

B. Herman and J. Lemasters, *Optical Microscopy: Emerging Methods and Applications*, Academic Press, New York, NY (1993).

P. Hobbs, *Building Electro-Optical Systems: Making It All Work*, Wiley and Sons, New York, NY (2000).

Bibliography

G. Holst and T. Lomheim, *CMOS/CCD Sensors and Camera Systems*, JCD Publishing, Winter Park, FL (2007).

B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy." *Science* **319**, 810 (2008).

R. Huber, M. Wojtkowski, and J. G. Fujimoto, "Fourier domain mode locking (FDML): A new laser operating regime and applications for optical coherence tomography." *Opt. Express* **14**, 3225–3237 (2006).

Invitrogen, http://www.invitrogen.com

R. Jozwicki, *Teoria Odwzorowania Optycznego* (in Polish), PWN (1988).

R. Jozwicki, *Optyka Instrumentalna* (in Polish), WNT (1970).

R. Leitgeb, C. K. Hitzenberger, and A. F. Fercher, "Performance of Fourier-domain versus time-domain optical coherence tomography." *Opt. Express* **11**, 889–894 (2003).

D. Malacara and B. Thompson, Eds., *Handbook of Optical Engineering*, Marcel Dekker, New York, NY (2001).

D. Malacara and Z. Malacara, *Handbook of Optical Design*, Marcel Dekker, New York, NY (1994).

D. Malacara, M. Servin, and Z. Malacara, *Interferogram Analysis for Optical Testing*, Marcel Dekker, New York, NY (1998).

D. Murphy, *Fundamentals of Light Microscopy and Electronic Imaging*, Wiley-Liss, Wilmington, DE (2001).

P. Mouroulis and J. Macdonald, *Geometrical Optics and Optical Design*, Oxford University Press, New York, NY (1997).

M. A. A. Neil, R. Juškaitis, and T. Wilson, "Method of obtaining optical sectioning by using structured light in a conventional microscope." *Optics Letters*, **22**(24), 1905–1907 (1997).

Nikon Microscopy U, http://www.microscopyu.com/

Bibliography

C. Palmer (Erwin Loewen, First Edition), *Diffraction Grating Handbook*, Newport Corp. (2005).

K. Patorski, *Handbook of the Moiré Fringe Technique*, Elsevier, Oxford, UK(1993).

J. Pawley, Ed., *Biological Confocal Microscopy*, Third Edition, Springer, New York, NY (2006).

M. C. Pierce, D. J. Javier and R. Richards-Kortum, “Optical contrast agents and imaging systems for detection and diagnosis of cancer.” *Int. J. Cancer* **123**, 1979–1990 (2008).

M. Pluta, *Advanced Light Microscopy, Volume One: Principle and Basic Properties*, PWN and Elsevier, New York, NY (1988).

M. Pluta, *Advanced Light Microscopy Volume Two: Specialized Methods*, PWN and Elsevier, New York, NY (1989).

M. Pluta, *Advanced Light Microscopy, Volume Three: Measuring Techniques*, PWN, Warsaw, Poland; and North Holland, Amsterdam, Holland (1993).

E. O. Potma, C. L. Evans, and X. S. Xie, “Heterodyne coherent anti-Stokes Raman scattering (CARS) imaging.” *Optics Letters*, **31**(2), 241–243 (2006).

D. W. Robinson, G. T.Reed,, Eds., *Interferogram Analysis*, IOP Publishing, Bristol, UK (1993).

M. J. Rust, M. Bates, and X. Zhuang, “Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM).” *Nature Methods*, **3**, 793–796 (2006).

B. Saleh and M. C. Teich, *Fundamentals of Photonics*, SecondEdition, Wiley, New York, NY (2007).

J. Schwiegerling, *Field Guide to Visual and Ophthalmic Optics*, SPIE Press, Bellingham, WA (2004).

W. Smith, *Modern Optical Engineering,* Third Edition, McGraw-Hill, New York, NY (2000).

D. Spector and R. Goldman, Eds., *Basic Methods in Microscopy*, Cold Spring Harbor Laboratory Press, Woodbury, NY (2006).

Bibliography

Thorlabs website resources: http://www.thorlabs.com/

P. Török and F. J. Kao, Eds., *Optical Imaging and Microscopy*, Springer, New York, NY (2007).

Veeco Optical Library entry, http://www.veeco.com/pdf/Optical_Library/Challenges_in_White_Light.pdf

R. Wayne, *Light and Video Microscopy*, Elsevier (reprinted by Academic Press), New York, NY (2009).

H. Yu, P. C. Cheng, P. C. Li, F. J. Kao, Eds., *Multi Modality Microscopy*, World Scientific, Hackensack, NJ (2006).

S. H. Yun, G. J. Tearney, B. J. Vakoc, M. Shishkov, W. Y. Oh, A. E. Desjardins, M. J. Suter, R. C. Chan, J. A. Evans, I. K. Jang, N. S. Nishioka, J. F. de Boer, and B. E. Bouma, "Comprehensive volumetric optical microscopy *in vivo*." *Nature Med* **12**, 1429–1433 (2006).

Zeiss Corporation, http://www.zeiss.com

Index

4Pi, 64, 109, 110
4Pi microscopy, 109

Abbe number, 25
Abbe resolution limit, 39
absorption filter, 60
achromatic objectives, 51
achromats, 51
Airy disk, 38
Amici lens, 55
amplitude contrast, 72
amplitude object, 63
amplitude ratio, 15
amplitude splitting, 17
analyzer, 83
anaxial illumination, 73
angular frequency, 4
angular magnification, 23
angular resolving power, 40
anisotropic media, propagation of light in, 9
aperture stop, 24
apochromats, 51
arc lamp, 58
array microscope, 113
array microscopy, 64, 113
astigmatism, 28
axial chromatic aberration, 26
axial resolution, 86

bandpass filter, 60
Bayer mask, 117
Bertrand lens, 45, 55
bias, 80
Brace Köhler compensator, 85
bright field, 64, 65, 76

charge-coupled device (CCD), 115
chief ray, 24
chromatic aberration, 26
chromatic resolving power, 20
circularly polarized light, 10
coherence length, 11
coherence scanning interferometry, 98
coherence time, 11
coherent, 11
coherent anti-Stokes Raman scattering (CARS), 64, 111
coma, 27
compensators, 83, 85
complex amplitudes, 12
compound microscope, 31
condenser's diaphragm, 46
cones, 32
confocal microscopy, 64, 86
contrast, 12
contrast of fringes, 15
cornea, 32
cover glass, 56
cover slip, 56
critical angle, 7
critical illumination, 46
cutoff frequency, 30

dark current, 118
dark field, 64, 65
dark noise, 118
de Sénarmont compensator, 85
depth of field, 41
depth of focus, 41
depth perception, 47
destructive interference, 19

Index

DIA, 44
DIC microscope design, 80
dichroic beam splitter, 91
dichroic mirror, 91
dielectric constant, 3
differential interference contrast (DIC), 64–65, 67, 79
diffraction, 18
diffraction grating, 19
diffraction orders, 19
digital microscopy, 114
digitiziation of CCD, 119
dispersion, 25
dispersion staining, 78
distortion, 28
dynamic range, 119

edge filter, 60
effective focal length, 22
electric fields, 12
electric vector, 10
electromagnetic spectrum, 2
electron-multiplying CCDs, 117
emission filter, 91
empty magnification, 40
entrance pupil, 24
entrance window, 24
epi-illumination, 44
epithelial cells, 2
evanescent wave microscopy, 105
excitation filter, 91
exit pupil, 24
exit window, 24
extraordinary wave, 9
eye, 32, 46
eye point, 48
eye relief, 48
eye's pupil, 46
eyepiece, 48

Fermat's principle, 5
field curvature, 28
field diaphragm, 46
field number (FN), 34, 48
field stop, 24, 34
filter turret, 92
filters, 91
finite tube length, 34
five-image technique, 102
fluorescence, 90
fluorescence microscopy, 64
fluorescent emission, 94
fluorites, 51
fluorophore quantum yield, 94
focal length, 21
focal plane, 21
focal point, 21
Fourier-domain OCT (FDOCT), 99
four-image technique, 102
fovea, 32
frame-transfer architecture, 116
Fraunhofer diffraction, 18
frequency of light, 1
Fresnel diffraction, 18
Fresnel reflection, 6
frustrated total internal reflection, 7
full-frame architecture, 116

gas-arc discharge lamp, 58
Gaussian imaging equation, 22
geometrical aberrations, 25

Index

geometrical optics, 21
Goos-Hänchen effect, 8
grating equation, 19–20

half bandwidth (HBW), 16
halo effect, 65, 71
halogen lamp, 58
high-eye point, 49
Hoffman modulation contrast, 75
Huygens, 48–49

I^5M, 64
image comparison, 65
image formation, 22
image space, 21
immersion liquid, 57
incandescent lamp, 58
incidence, 6
infinity-corrected objective, 35
infinity-corrected systems, 35
infrared radiation (IR), 2
intensity of light, 12
interference, 12
interference filter, 16, 60
interference microscopy, 64, 98
interference order, 16
interfering beams, 15
interferometers, 17
interline transfer architecture, 116
intermediate image plane, 46
inverted microscope, 33
iris, 32

Köhler illumination, 43–45

laser scanning, 64
lasers, 96
laser-scanning microscopy (LSM), 97
lateral magnification, 23
lateral resolution, 71
laws of reflection and refraction, 6
lens, 32
light sheet, 112
light sources for microscopy, common, 58
light-emitting diodes (LEDs), 59
limits of light microscopy, 110
linearly polarized light, 10
line-scanning confocal microscope, 87
Linnik, 101
long working distance objectives, 53
longitudinal magnification, 23
low magnification objectives, 54

Mach-Zehnder interferometer, 17
macula, 32
magnetic permeability, 3
magnification, 21, 23
magnifier, 31
magnifying power (MP), 37
marginal ray, 24
Maxwell's equations, 3
medium permittivity, 3
mercury lamp, 58
metal halide lamp, 58
Michelson, 17, 53, 98, 100, 101

Index

minimum focus distance, 32
Mirau, 53, 98, 101
modulation contrast microscopy (MCM), 74
modulation transfer function (MTF), 29
molecular absorption cross section, 94
monochromatic light, 11
multi-photon fluorescence, 95
multi-photon microscopy, 64
multiple beam interference, 16

nature of light, 1
near point, 32
negative birefringence, 9
negative contrast, 69
neutral density (ND), 60
Newtonian equation, 22
Nomarski interference microscope, 82
Nomarski prism, 62, 80
non-self-luminous, 63
numerical aperture (NA), 38, 50
Nyquist frequency, 120

object space, 21
objective, 31
objective correction, 50
oblique illumination, 73
optic axis, 9
optical coherence microscopy (OCM), 98
optical coherence tomography (OCT), 98
optical density (OD), 60
optical path difference (OPD), 5
optical path length (OPL), 5
optical power, 22
optical profilometry, 98, 101
optical rays, 21
optical sectioning, 103
optical spaces, 21
optical staining, 77
optical transfer function, 29
optical tube length, 34
ordinary wave, 9
oversampling, 121

PALM, 64, 110
paraxial optics, 21
parfocal distance, 34
partially coherent, 11
particle model, 1
performance metrics, 29
phase-advancing objects, 68
phase contrast microscope, 70
phase contrast, 64, 65, 66
phase object, 63
phase of light, 4
phase retarding objects, 68
phase-shifting algorithms, 102
phase-shifting interferometry (PSI), 98, 100
photobleaching, 94
photodiode, 115
photon noise, 118
Planck's constant, 1

Index

point spread function (PSF), 29
point-scanning confocal microscope, 87
Poisson statistics, 118
polarization microscope, 84
polarization microscopy, 64, 83
polarization of light, 10
polarization prisms, 61
polarization states, 10
polarizers, 61–62
positive birefringence, 9
positive contrast, 69
pupil function, 29

quantum efficiency (QE), 97, 117
quasi-monochromatic, 11
quenching, 94

Raman microscopy, 64, 111
Raman scattering, 111
Ramsden eyepiece, 48
Rayleigh resolution limit, 39
read noise, 118
real image, 22
red blood cells, 2
reflectance coefficients, 6
reflected light, 6, 16
reflection law, 6
reflective grating, 20
refracted light, 6
refraction law, 6
refractive index, 4
RESOLFT, 64, 110

resolution limit, 39
retardation, 83
retina, 32
RGB filters, 117
Rheinberg illumination, 77
rods, 32

Sagnac interferometer, 17
sample conjugate, 46
sample path, 44
scanning approaches, 87
scanning white light interferometry, 98
Selective plane illumination microscopy (SPIM), 64, 112
self-luminous, 63
semi-apochromats, 51
shading-off effect, 71
shearing interferometer, 17
shearing interferometry, 79
shot noise, 118
SI, 64
sign convention, 21
signal-to-noise ratio (SNR), 119
Snell's law, 6
solid immersion, 106
solid immersion lenses (SILs), 106
Sparrow resolution limit, 30
spatial coherence, 11
spatial resolution, 86
spectral-domain OCT (SDOCT), 99
spectrum of microscopy, 2
spherical aberration, 27
spinning-disc confocal imaging, 88
stereo microscopes, 47

stimulated emission depletion (STED), 107
stochastic optical reconstruction microscopy (STORM), 64, 108, 110
Stokes shift, 90
Strehl ratio, 29, 30
structured illumination, 103
super-resolution microscopy, 64
swept-source OCT (SSOCT), 99

telecentricity, 36
temporal coherence, 11, 13–14
thin lenses, 21
three-image technique, 102
total internal reflection (TIR), 7
total internal reflection fluorescence (TIRF) microscopy, 105
transmission coefficients, 6
transmission grating, 20
transmitted light, 16
transverse chromatic aberration, 26
transverse magnification, 23
tube length, 34
tube lens, 35
tungsten-argon lamp, 58

ultraviolet radiation (UV), 2
undersampling, 120
uniaxial crystals, 9
upright microscope, 33
useful magnification, 40
UV objectives, 54

velocity of light, 1
vertical scanning interferometry (VSI), 98, 100
virtual image, 22
visibility, 12
visibility in phase contrast, 69
visible spectrum (VIS), 2
visual stereo resolving power, 47

water immersion objectives, 54
wave aberrations, 25
wave equations, 3
wave group, 11
wave model, 1
wave number, 4
wave plate compensator, 85
wavefront propagation, 4
wavefront splitting, 17
wavelength of light, 1
Wollaston prism, 61, 80
working distance (WD), 50

x-ray radiation, 2

Tomasz S. Tkaczyk is an Assistant Professor of Bioengineering and Electrical and Computer Engineering at Rice University, Houston, Texas, where he develops modern optical instrumentation for biological and medical applications. His primary research is in microscopy, including endo-microscopy, cost-effective high-performance optics for diagnostics, and multi-dimensional imaging (snapshot hyperspectral microscopy and spectro-polarimetry).

Professor Tkaczyk received his M.S. and Ph.D. from the Institute of Micromechanics and Photonics, Department of Mechatronics, Warsaw University of Technology, Poland. Beginning in 2003, after his postdoctoral training, he worked as a research professor at the College of Optical Sciences, University of Arizona. He joined Rice University in the summer of 2007.

Fundamental Optics

1

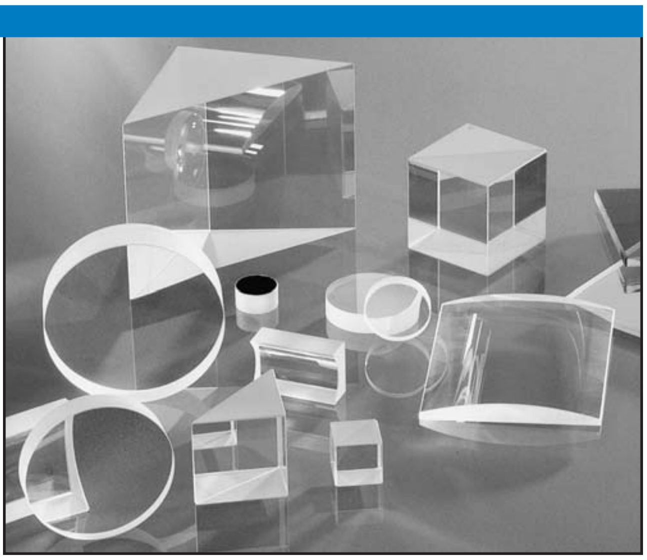

Introduction 1.2
Paraxial Formulas 1.3
Imaging Properties of Lens Systems 1.6
Lens Combination Formulas 1.8
Performance Factors 1.11
Lens Shape 1.17
Lens Combinations 1.18
Diffraction Effects 1.20
Lens Selection 1.23
Spot Size 1.26
Aberration Balancing 1.27
Definition of Terms 1.29
Paraxial Lens Formulas 1.32
Principal-Point Locations 1.36
Prisms 1.37
Polarization 1.41
Waveplates 1.46
Etalons 1.49
Ultrafast Theory 1.52

Introduction

The process of solving virtually any optical engineering problem can be broken down into two main steps. First, paraxial calculations (first order) are made to determine critical parameters such as magnification, focal length(s), clear aperture (diameter), and object and image position. These paraxial calculations are covered in the next section of this chapter.

Second, actual components are chosen based on these paraxial values, and their actual performance is evaluated with special attention paid to the effects of aberrations. A truly rigorous performance analysis for all but the simplest optical systems generally requires computer ray tracing, but simple generalizations can be used, especially when the lens selection process is confined to a limited range of component shapes.

In practice, the second step may reveal conflicts with design constraints, such as component size, cost, or product availability. System parameters may therefore require modification.

Because some of the terms used in this chapter may not be familiar to all readers, a glossary of terms is provided in *Definition of Terms*.

Finally, it should be noted that the discussion in this chapter relates only to systems with uniform illumination; optical systems for Gaussian beams are covered in *Gaussian Beam Optics*.

Engineering Support

CVI Melles Griot maintains a staff of knowledgeable, experienced applications engineers at each of our facilities worldwide. The information given in this chapter is sufficient to enable the user to select the most appropriate catalog lenses for the most commonly encountered applications. However, when additional optical engineering support is required, our applications engineers are available to provide assistance. Do not hesitate to contact us for help in product selection or to obtain more detailed specifications on CVI Melles Griot products.

THE OPTICAL ENGINEERING PROCESS

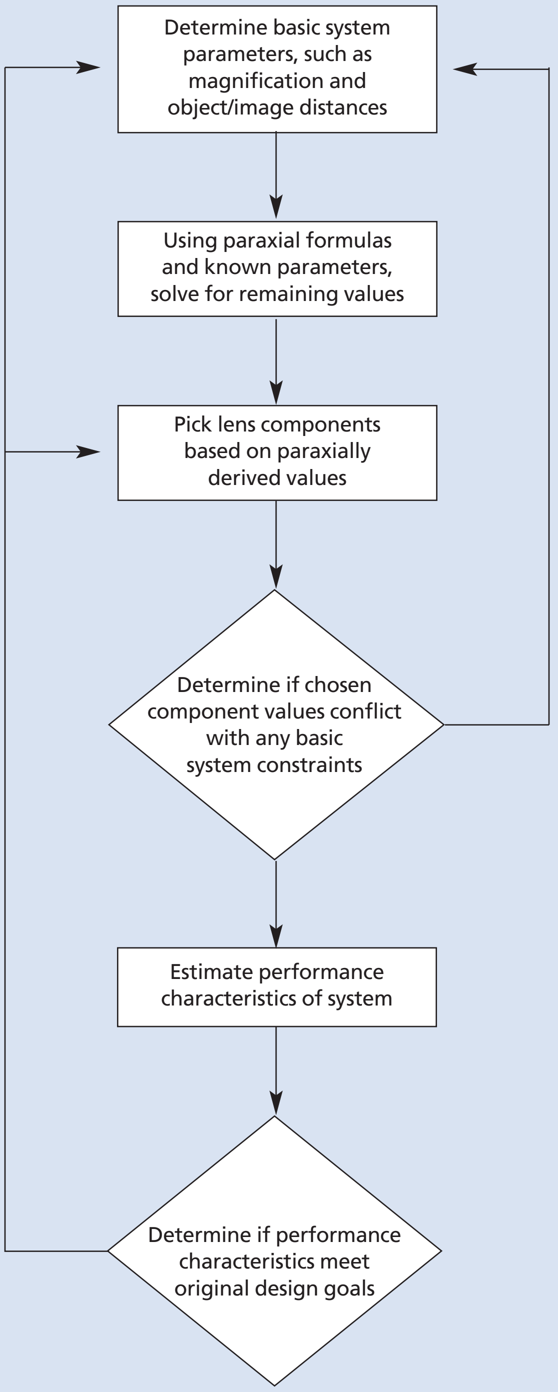
Determine basic system parameters, such as magnification and object/image distances
Using paraxial formulas and known parameters, solve for remaining values
Pick lens components based on paraxially derived values
Determine if chosen component values conflict with any basic system constraints
Estimate performance characteristics of system
Determine if performance characteristics meet original design goals

Fundamental Optics

Gaussian Beam Optics

Optical Specifications

Material Properties

Optical Coatings

Paraxial Formulas

Sign Conventions

The validity of the paraxial lens formulas is dependent on adherence to the following sign conventions:

For lenses: (refer to figure 1.1)

s is + for object to left of H (the first principal point)

s is − for object to right of H

s'' is + for image to right of H″ (the second principal point)

s'' is − for image to left of H″

m is + for an inverted image

m is − for an upright image

For mirrors:

f is + for convex (diverging) mirrors

f is − for concave (converging) mirrors

s is + for object to left of H

s is − for object to right of H

s'' is − for image to right of H″

s'' is + for image to left of H″

m is + for an inverted image

m is − for an upright image

When using the thin-lens approximation, simply refer to the left and right of the lens.

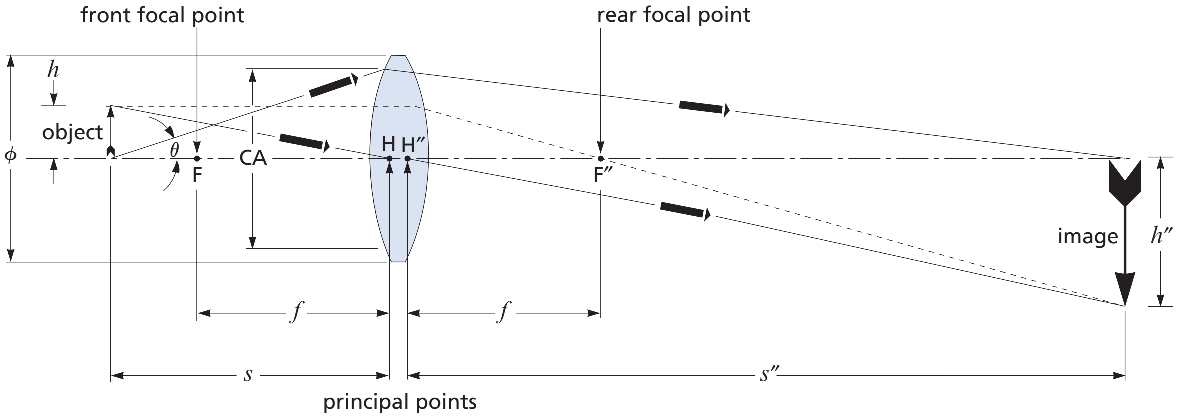

Note location of object and image relative to front and rear focal points.

ϕ = lens diameter

CA = clear aperture (typically 90% of ϕ)

f = effective focal length (EFL) which may be positive (as shown) or negative. f represents both FH and H″F″, assuming lens is surrounded by medium of index 1.0

m = $s''/s = h''/h$ = magnification or conjugate ratio, said to be infinite if either s'' or s is infinite

θ = arcsin (CA/2s)

s = object distance, positive for object (whether real or virtual) to the left of principal point H

s'' = image distance (s and s'' are collectively called conjugate distances, with object and image in conjugate planes), positive for image (whether real or virtual) to the right of principal point H″

h = object height

h'' = image height

Figure 1.1 **Sign conventions**

Typically, the first step in optical problem solving is to select a system focal length based on constraints such as magnification or conjugate distances (object and image distance). The relationship among focal length, object position, and image position is given by

$$\frac{1}{f}=\frac{1}{s}+\frac{1}{s''} \tag{1.1}$$

This formula is referenced to figure 1.1 and the sign conventions given in *Sign Conventions*.

By definition, magnification is the ratio of image size to object size or

$$m=\frac{s''}{s}=\frac{h''}{h}. \tag{1.2}$$

This relationship can be used to recast the first formula into the following forms:

$$f=m\frac{(s+s'')}{(m+1)^2} \tag{1.3}$$

$$f=\frac{sm}{m+1} \tag{1.4}$$

$$f=\frac{s+s''}{m+2+\dfrac{1}{m}} \tag{1.5}$$

$$s(m+1)=s+s'' \tag{1.6}$$

where ($s+s''$) is the approximate object-to-image distance.

With a real lens of finite thickness, the image distance, object distance, and focal length are all referenced to the principal points, not to the physical center of the lens. By neglecting the distance between the lens' principal points, known as the hiatus, $s+s''$ becomes the object-to-image distance. This simplification, called the thin-lens approximation, can speed up calculation when dealing with simple optical systems.

Example 1: Object outside Focal Point

A 1-mm-high object is placed on the optical axis, 200 mm left of the left principal point of a LDX-25.0-51.0-C ($f=50$ mm). Where is the image formed, and what is the magnification? (See figure 1.2.)

$$\frac{1}{s''}=\frac{1}{f}-\frac{1}{s}$$

$$\frac{1}{s''}=\frac{1}{50}-\frac{1}{200}$$

$$s''=66.7\ \mathrm{mm}$$

$$m=\frac{s''}{s}=\frac{66.7}{200}=0.33$$

or real image is 0.33 mm high and inverted.

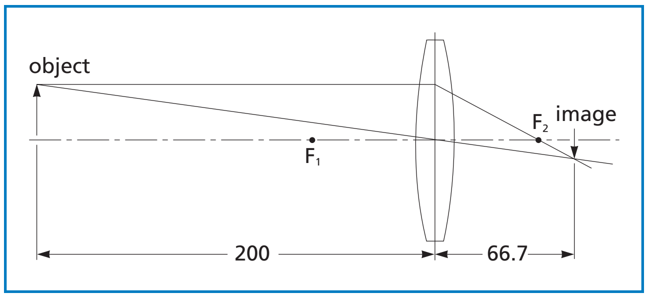

Figure 1.2 **Example 1** ($f=50$ mm, $s=200$ mm, $s''=66.7$ mm)

Example 2: Object inside Focal Point

The same object is placed 30 mm left of the left principal point of the same lens. Where is the image formed, and what is the magnification? (See figure 1.3.)

$$\frac{1}{s''}=\frac{1}{50}-\frac{1}{30}$$

$$s''=-75\ \mathrm{mm}$$

$$m=\frac{s''}{s}=\frac{-75}{30}=-2.5$$

or virtual image is 2.5 mm high and upright.

In this case, the lens is being used as a magnifier, and the image can be viewed only back through the lens.

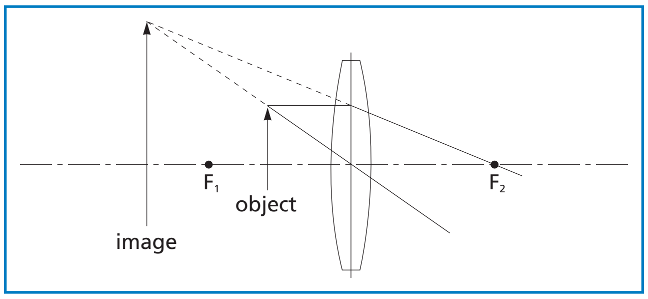

Figure 1.3 **Example 2** ($f=50$ mm, $s=30$ mm, $s''=-75$ mm)

Example 3: Object at Focal Point

A 1-mm-high object is placed on the optical axis, 50 mm left of the first principal point of an LDK-50.0-52.2-C ($f=-50$ mm). Where is the image formed, and what is the magnification? (See figure 1.4.)

$$\frac{1}{s''}=\frac{1}{-50}-\frac{1}{50}$$

$$s''=-25\ \mathrm{mm}$$

$$m=\frac{s''}{s}=\frac{-25}{50}=-0.5$$

or virtual image is 0.5 mm high and upright.

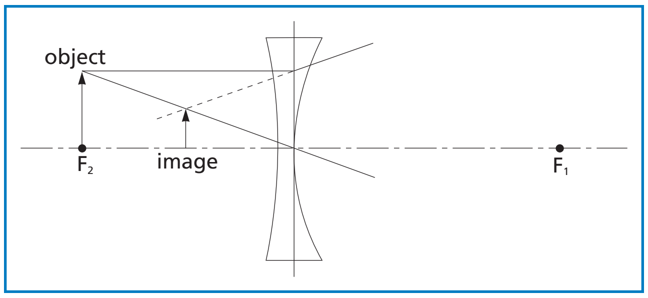

Figure 1.4 **Example 3** ($f = -50$ mm, $s = 50$ mm, $s'' = -25$ mm)

A simple graphical method can also be used to determine paraxial image location and magnification. This graphical approach relies on two simple properties of an optical system. First, a ray that enters the system parallel to the optical axis crosses the optical axis at the focal point. Second, a ray that enters the first principal point of the system exits the system from the second principal point parallel to its original direction (i.e., its exit angle with the optical axis is the same as its entrance angle). This method has been applied to the three previous examples illustrated in figures 1.2 through 1.4. Note that by using the thin-lens approximation, this second property reduces to the statement that a ray passing through the center of the lens is undeviated.

F-NUMBER AND NUMERICAL APERTURE

The paraxial calculations used to determine the necessary element diameter are based on the concepts of focal ratio (f-number or f/#) and numerical aperture (NA). The f-number is the ratio of the focal length of the lens to its "effective" diameter, the clear aperture (CA).

$$\text{f-number} = \frac{f}{\text{CA}} \cdot \qquad (1.7)$$

To visualize the f-number, consider a lens with a positive focal length illuminated uniformly with collimated light. The f-number defines the angle of the cone of light leaving the lens which ultimately forms the image. This is an important concept when the throughput or light-gathering power of an optical system is critical, such as when focusing light into a monochromator or projecting a high-power image.

The other term used commonly in defining this cone angle is numerical aperture. The NA is the sine of the angle made by the marginal ray with the optical axis. By referring to figure 1.5 and using simple trigonometry, it can be seen that

$$\text{NA} = \sin\theta = \frac{\text{CA}}{2f} \qquad (1.8)$$

and

$$\text{NA} = \frac{1}{2(\text{f-number})}. \qquad (1.9)$$

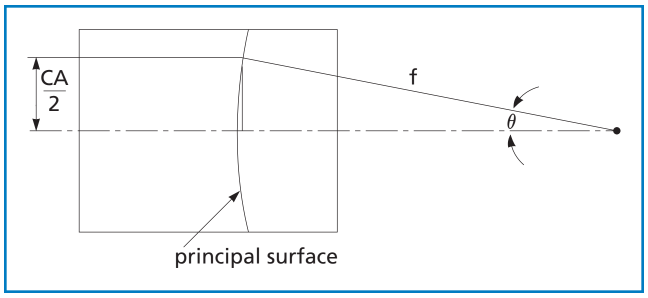

Figure 1.5 **F-number and numerical aperture**

Ray f-numbers can also be defined for any arbitrary ray if its conjugate distance and the diameter at which it intersects the principal surface of the optical system are known.

NOTE

Because the sign convention given previously is not used universally in all optics texts, the reader may notice differences in the paraxial formulas. However, results will be correct as long as a consistent set of formulas and sign conventions is used.

Imaging Properties of Lens Systems

THE OPTICAL INVARIANT

To understand the importance of the NA, consider its relation to magnification. Referring to figure 1.6,

$$\text{NA (object side)} = \sin\theta = \frac{\text{CA}}{2s} \qquad (1.10)$$

$$\text{NA}'' \text{ (image side)} = \sin\theta'' = \frac{\text{CA}}{2s''} \qquad (1.11)$$

which can be rearranged to show

$$\text{CA} = 2s \, \sin\theta \qquad (1.12)$$

and

$$\text{CA} = 2s''\sin\theta'' \qquad (1.13)$$

leading to

$$\frac{s''}{s} = \frac{\sin\theta}{\sin\theta''} = \frac{\text{NA}}{\text{NA}''}. \qquad (1.14)$$

Since $\frac{s''}{s}$ is simply the magnification of the system,

we arrive at

$$m = \frac{\text{NA}}{\text{NA}''}. \qquad (1.15)$$

The magnification of the system is therefore equal to the ratio of the NAs on the object and image sides of the system. This powerful and useful result is completely independent of the specifics of the optical system, and it can often be used to determine the optimum lens diameter in situations involving aperture constraints.

When a lens or optical system is used to create an image of a source, it is natural to assume that, by increasing the diameter (ϕ) of the lens, thereby increasing its CA, we will be able to collect more light and thereby produce a brighter image. However, because of the relationship between magnification and NA, there can be a theoretical limit beyond which increasing the diameter has no effect on light-collection efficiency or image brightness.

Since the NA of a ray is given by CA/2s, once a focal length and magnification have been selected, the value of NA sets the value of CA. Thus, if one is dealing with a system in which the NA is constrained on either the object or image side, increasing the lens diameter beyond this value will increase system size and cost but will not improve performance (i.e., throughput or image brightness). This concept is sometimes referred to as the optical invariant.

SAMPLE CALCULATION

To understand how to use this relationship between magnification and NA, consider the following example.

Example: System with Fixed Input NA

Two very common applications of simple optics involve coupling light into an optical fiber or into the entrance slit of a monochromator. Although these problems appear to be quite different, they both have the same limitation — they have a fixed NA. For monochromators, this limit is usually expressed in terms of the f-number. In addition to the fixed NA, they both have a fixed entrance pupil (image) size.

Suppose it is necessary, using a singlet lens, to couple the output of an incandescent bulb with a filament 1 mm in diameter into an optical fiber as shown in figure 1.7. Assume that the fiber has a core diameter of 100 μm and an NA of 0.25, and that the design requires that the total distance from the source to the fiber be 110 mm. Which lenses are appropriate?

By definition, the magnification must be 0.1. Letting $s+s''$ total 110 mm (using the thin-lens approximation), we can use equation 1.3,

$$f = m\frac{(s+s'')}{(m+1)^2}, \qquad \text{(see eq. 1.3)}$$

to determine that the focal length is 9.1 mm. To determine the conjugate distances, s and s'', we utilize equation 1.6,

$$s(m+1) = s + s'', \qquad \text{(see eq. 1.6)}$$

and find that $s = 100$ mm and $s'' = 10$ mm.

We can now use the relationship NA = CA/2s or NA″ = CA/2s'' to derive CA, the optimum clear aperture (effective diameter) of the lens.

With an image NA of 0.25 and an image distance (s'') of 10 mm,

$$0.25 = \frac{\text{CA}}{20}$$

and

$$\text{CA} = 5 \text{ mm}.$$

Accomplishing this imaging task with a single lens therefore requires an optic with a 9.1-mm focal length and a 5-mm diameter. Using a larger diameter lens will not result in any greater system throughput because of the limited input NA of the optical fiber. The singlet lenses in this catalog that meet these criteria are LPX-5.0-5.2-C, which is plano-convex, and LDX-6.0-7.7-C and LDX-5.0-9.9-C, which are biconvex.

Making some simple calculations has reduced our choice of lenses to just three. The following chapter, *Gaussian Beam Optics*, discusses how to make a final choice of lenses based on various performance criteria.

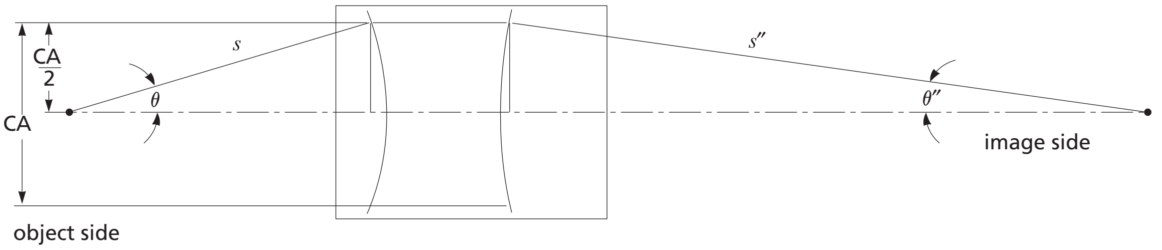
CA
$\frac{CA}{2}$
s
θ
s''
θ''
image side
object side

Figure 1.6 **Numerical aperture and magnification**

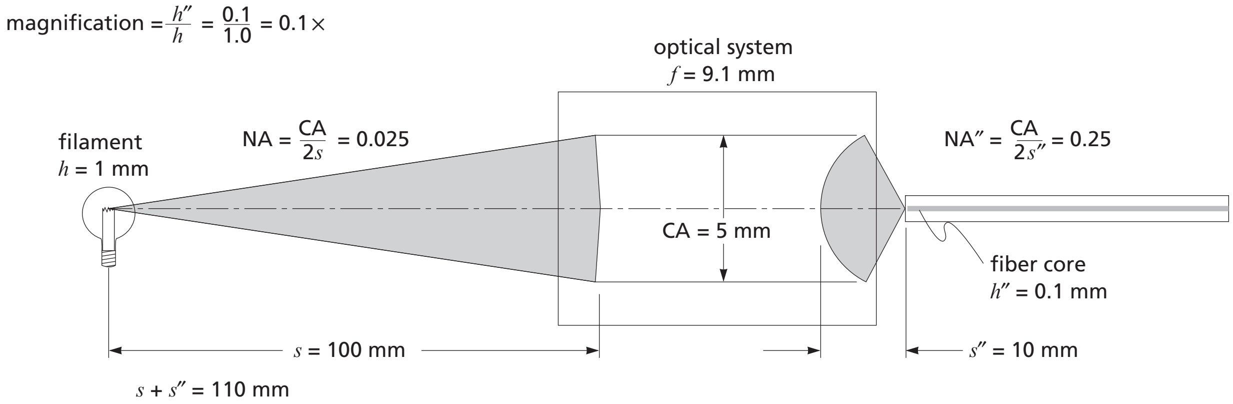
magnification = $\frac{h''}{h} = \frac{0.1}{1.0} = 0.1\times$
optical system
f = 9.1 mm
filament
h = 1 mm
NA = $\frac{CA}{2s}$ = 0.025
NA″ = $\frac{CA}{2s''}$ = 0.25
CA = 5 mm
fiber core
h'' = 0.1 mm
s = 100 mm
s'' = 10 mm
$s + s''$ = 110 mm

Figure 1.7 **Optical system geometry for focusing the output of an incandescent bulb into an optical fiber**

Lens Combination Formulas

Many optical tasks require several lenses in order to achieve an acceptable level of performance. One possible approach to lens combinations is to consider each image formed by each lens as the object for the next lens and so on. This is a valid approach, but it is time consuming and unnecessary.

It is much simpler to calculate the effective (combined) focal length and principal-point locations and then use these results in any subsequent paraxial calculations (see figure 1.8). They can even be used in the optical invariant calculations described in the preceding section.

EFFECTIVE FOCAL LENGTH

The following formulas show how to calculate the effective focal length and principal-point locations for a combination of any two arbitrary components. The approach for more than two lenses is very simple: Calculate the values for the first two elements, then perform the same calculation for this combination with the next lens. This is continued until all lenses in the system are accounted for.

The expression for the combination focal length is the same whether lens separation distances are large or small and whether f_1 and f_2 are positive or negative:

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}. \tag{1.16}$$

This may be more familiar in the form

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}. \tag{1.17}$$

Notice that the formula is symmetric with respect to the interchange of the lenses (end-for-end rotation of the combination) at constant *d*. The next two formulas are not.

COMBINATION FOCAL-POINT LOCATION

For all values of f_1, f_2, and d, the location of the focal point of the combined system (s_2''), measured from the secondary principal point of the second lens (H_2''), is given by

$$s_2'' = \frac{f_2(f_1 - d)}{f_1 + f_2 - d}. \tag{1.18}$$

This can be shown by setting $s_1 = d - f_1$ (see figure 1.8a), and solving

$$\frac{1}{f_2} = \frac{1}{s_1} + \frac{1}{s_2''}$$

for s_2''.

Symbols

f_c = combination focal length (EFL), positive if combination final focal point falls to the right of the combination secondary principal point, negative otherwise (see figure 1.8c).

f_1 = focal length of the first element (see figure 1.8a).

f_2 = focal length of the second element.

d = distance from the secondary principal point of the first element to the primary principal point of the second element, positive if the primary principal point is to the right of the secondary principal point, negative otherwise (see figure 1.8b).

s_1'' = distance from the primary principal point of the first element to the final combination focal point (location of the final image for an object at infinity to the right of both lenses), positive if the focal point is to left of the first element's primary principal point (see figure 1.8d).

s_2'' = distance from the secondary principal point of the second element to the final combination focal point (location of the final image for an object at infinity to the left of both lenses), positive if the focal point is to the right of the second element's secondary principal point (see figure 1.8b).

z_H = distance to the combination primary principal point measured from the primary principal point of the first element, positive if the combination secondary principal point is to the right of secondary principal point of second element (see figure 1.8d).

z_H'' = distance to the combination secondary principal point measured from the secondary principal point of the second element, positive if the combination secondary principal point is to the right of the secondary principal point of the second element (see figure 1.8c).

Note: *These paraxial formulas apply to* ***coaxial*** *combinations of both thick and thin lenses immersed in air or any other fluid with refractive index independent of position. They assume that light propagates from left to right through an optical system.*

COMBINATION SECONDARY PRINCIPAL-POINT LOCATION

Because the thin-lens approximation is obviously highly invalid for most combinations, the ability to determine the location of the secondary principal point is vital for accurate determination of d when another element is added. The simplest formula for this calculates the distance from the secondary principal point of the final (second) element to the secondary principal point of the combination (see figure 1.8b):

$$z = s_2'' - f\,. \tag{1.19}$$

COMBINATION EXAMPLES

It is possible for a lens combination or system to exhibit principal planes that are far removed from the system. When such systems are themselves combined, negative values of d may occur. Probably the simplest example of a negative d-value situation is shown in figure 1.9. Meniscus lenses with steep surfaces have external principal planes. When two of these lenses are brought into contact, a negative value of d can occur. Other combined-lens examples are shown in figures 1.10 through 1.13.

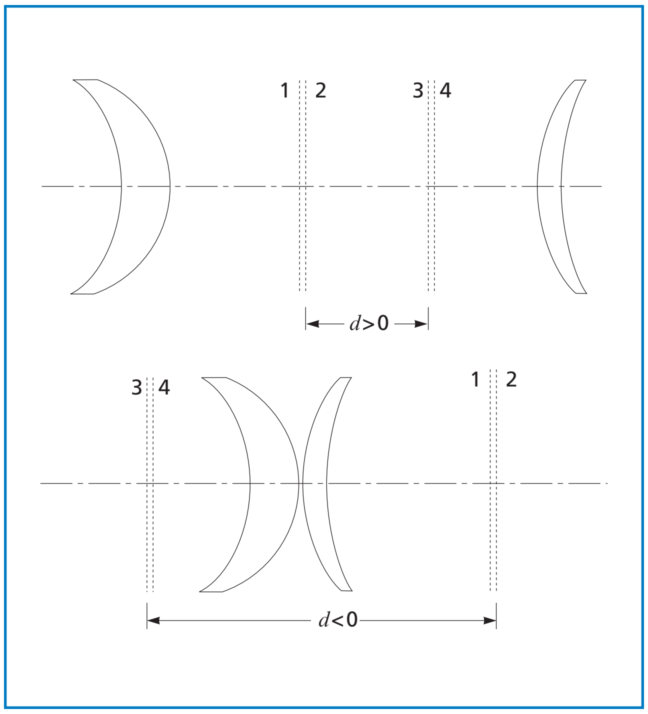

Figure 1.9 **"Extreme" meniscus-form lenses with external principal planes** (drawing not to scale)

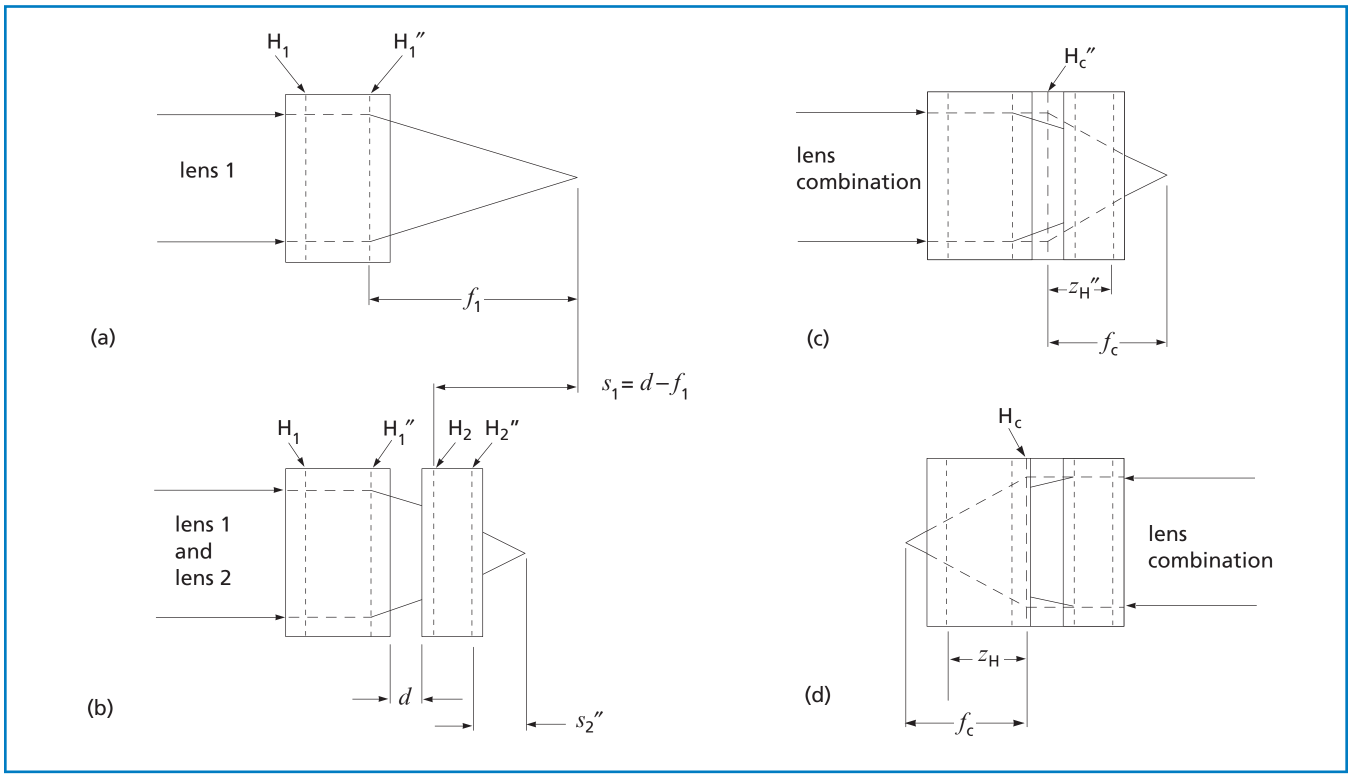

Figure 1.8 **Lens combination focal length and principal planes**

Fundamental Optics
Gaussian Beam Optics
Optical Specifications
Material Properties
Optical Coatings

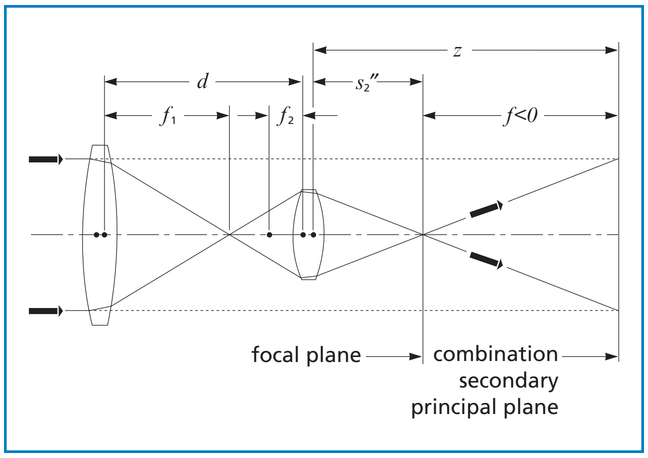

Figure 1.10 **Positive lenses separated by distance greater than $f_1 + f_2$:** f is negative and both s_2'' and z are positive. Lens symmetry is not required.

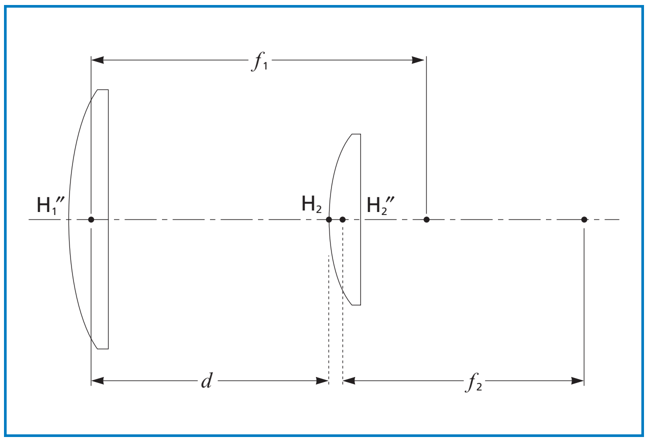

Figure 1.11 **Achromatic combinations:** Air-spaced lens combinations can be made nearly achromatic, even though both elements are made from the same material. Achieving achromatism requires that, in the thin-lens approximation,

$$d = \frac{(f_1 + f_2)}{2}.$$

This is the basis for Huygens and Ramsden eyepieces.

This approximation is adequate for most thick-lens situations. The signs of f_1, f_2, and d are unrestricted, but d must have a value that guarantees the existence of an air space. Element shapes are unrestricted and can be chosen to compensate for other aberrations.

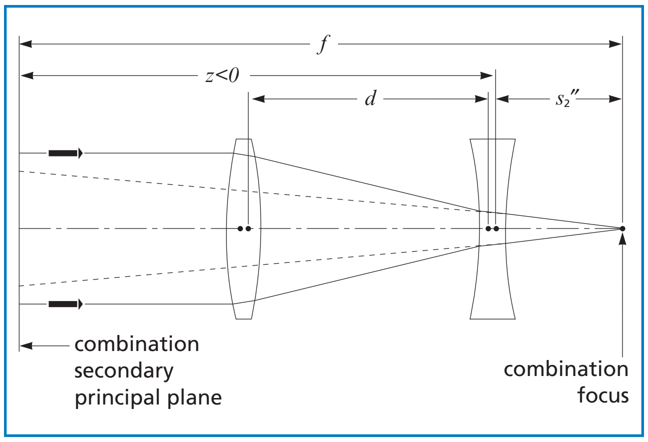

Figure 1.12 **Telephoto combination:** The most important characteristic of the telephoto lens is that the EFL, and hence the image size, can be made much larger than the distance from the first lens surface to the image would suggest by using a positive lens followed by a negative lens (but not necessarily the lens shapes shown in the figure). For example, f_1 is positive and $f_2 = -f_1/2$. Then f is negative for d less than $f_1/2$, infinite for $d = f_1/2$ (Galilean telescope or beam expander), and positive for d larger than $f_1/2$. To make the example even more specific, catalog lenses LDX-50.8-130.4-C and LDK-42.0-52.2-C, with $d = 78.2$ mm, will yield $s_2'' = 2.0$ m, $f = 5.2$ m, and $z = 43.2$ m.

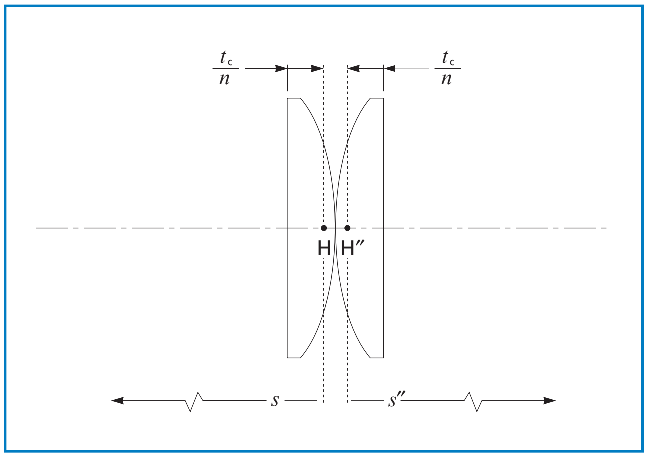

Figure 1.13 **Condenser configuration:** The convex vertices of a pair of identical plano-convex lenses are on contact. (The lenses could also be plano aspheres.) Because $d = 0$, $f = f_1/2 = f_2/2$, $f_1/2 = s_2''$, and $z = 0$. The secondary principal point of the second element and the secondary principal point of the combination coincide at H'', at depth t_c/n beneath the vertex of the plano surface of the second element, where t_c is the element center thickness and n is the refractive index of the element. By symmetry, the primary principal point of the combination is similarly located in the first element. Combination conjugate distances must be measured from these points.

Performance Factors

After paraxial formulas have been used to select values for component focal length(s) and diameter(s), the final step is to select actual lenses. As in any engineering problem, this selection process involves a number of tradeoffs, including performance, cost, weight, and environmental factors.

The performance of real optical systems is limited by several factors, including lens aberrations and light diffraction. The magnitude of these effects can be calculated with relative ease.

Numerous other factors, such as lens manufacturing tolerances and component alignment, impact the performance of an optical system. Although these are not considered explicitly in the following discussion, it should be kept in mind that if calculations indicate that a lens system only just meets the desired performance criteria, in practice it may fall short of this performance as a result of other factors. In critical applications, it is generally better to select a lens whose calculated performance is significantly better than needed.

DIFFRACTION

Diffraction, a natural property of light arising from its wave nature, poses a fundamental limitation on any optical system. Diffraction is always present, although its effects may be masked if the system has significant aberrations. When an optical system is essentially free from aberrations, its performance is limited solely by diffraction, and it is referred to as diffraction limited.

In calculating diffraction, we simply need to know the focal length(s) and aperture diameter(s); we do not consider other lens-related factors such as shape or index of refraction.

Since diffraction increases with increasing f-number, and aberrations decrease with increasing f-number, determining optimum system performance often involves finding a point where the combination of these factors has a minimum effect.

ABERRATIONS

To determine the precise performance of a lens system, we can trace the path of light rays through it, using Snell's law at each optical interface to determine the subsequent ray direction. This process, called ray tracing, is usually accomplished on a computer. When this process is completed, it is typically found that not all the rays pass through the points or positions predicted by paraxial theory. These deviations from ideal imaging are called lens aberrations.

The direction of a light ray after refraction at the interface between two homogeneous, isotropic media of differing index of refraction is given by Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad (1.20)$$

where θ_1 is the angle of incidence, θ_2 is the angle of refraction, and both angles are measured from the surface normal as shown in figure 1.14.

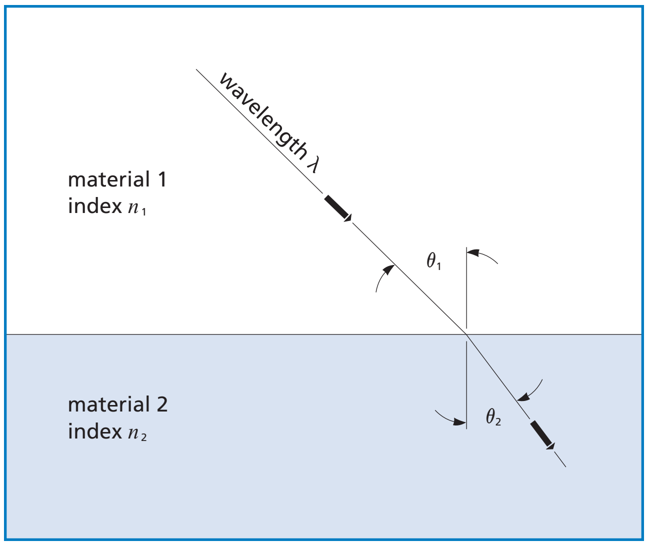

Figure 1.14 **Refraction of light at a dielectric boundary**

APPLICATION NOTE

Technical Assistance

Detailed performance analysis of an optical system is accomplished by using computerized ray-tracing software. CVI Melles Griot applications engineers are able to provide a ray-tracing analysis of simple catalog-component systems. If you need assistance in determining the performance of your optical system, or in selecting optimum components for your particular application, please contact your nearest CVI Melles Griot office.

Even though tools for the precise analysis of an optical system are becoming easier to use and are readily available, it is still quite useful to have a method for quickly estimating lens performance. This not only saves time in the initial stages of system specification, but can also help achieve a better starting point for any further computer optimization.

The first step in developing these rough guidelines is to realize that the sine functions in Snell's law can be expanded in an infinite Taylor series:

$$\sin\theta_1 = \theta_1 - \theta_1^3/3! + \theta_1^5/5! - \theta_1^7/7! + \theta_1^9/9! - \ldots \qquad (1.21)$$

The first approximation we can make is to replace all the sine functions with their arguments (i.e., replace $\sin\theta_1$ with θ_1 itself and so on). This is called first-order or paraxial theory because only the first terms of the sine expansions are used. Design of any optical system generally starts with this approximation using the paraxial formulas.

The assumption that $\sin\theta = \theta$ is reasonably valid for θ close to zero (i.e., high f-number lenses). With more highly curved surfaces (and particularly marginal rays), paraxial theory yields increasingly large deviations from real performance because $\sin\theta \neq \theta$. These deviations are known as aberrations. Because a perfect optical system (one without any aberrations) would form its image at the point and to the size indicated by paraxial theory, aberrations are really a measure of how the image differs from the paraxial prediction.

As already stated, exact ray tracing is the only rigorous way to analyze real lens surfaces. Before the advent of electronic computers, this was excessively tedious and time consuming. Seidel* addressed this issue by developing a method of calculating aberrations resulting from the $\theta_1{}^3/3!$ term. The resultant third-order lens aberrations are therefore called Seidel aberrations.

To simplify these calculations, Seidel put the aberrations of an optical system into several different classifications. In monochromatic light they are spherical aberration, astigmatism, field curvature, coma, and distortion. In polychromatic light there are also chromatic aberration and lateral color. Seidel developed methods to approximate each of these aberrations without actually tracing large numbers of rays using all the terms in the sine expansions.

In actual practice, aberrations occur in combinations rather than alone. This system of classifying them, which makes analysis much simpler, gives a good description of optical system image quality. In fact, even in the era of powerful ray-tracing software, Seidel's formula for spherical aberration is still widely used.

SPHERICAL ABERRATION

Figure 1.15 illustrates how an aberration-free lens focuses incoming collimated light. All rays pass through the focal point F″. The lower figure shows the situation more typically encountered in single lenses. The farther from the optical axis the ray enters the lens, the nearer to the lens it focuses (crosses the optical axis). The distance along the optical axis between the intercept of the rays that are nearly on the optical axis (paraxial rays) and the rays that go through the edge of the lens (marginal rays) is called longitudinal spherical aberration (LSA). The height at which these rays intercept the paraxial focal plane is called transverse spherical aberration (TSA). These quantities are related by

$$\text{TSA} = \text{LSA} \times \tan(u''). \qquad (1.22)$$

Spherical aberration is dependent on lens shape, orientation, and conjugate ratio, as well as on the index of refraction of the materials present. Parameters for choosing the best lens shape and orientation for a given task are presented later in this chapter. However, the third-

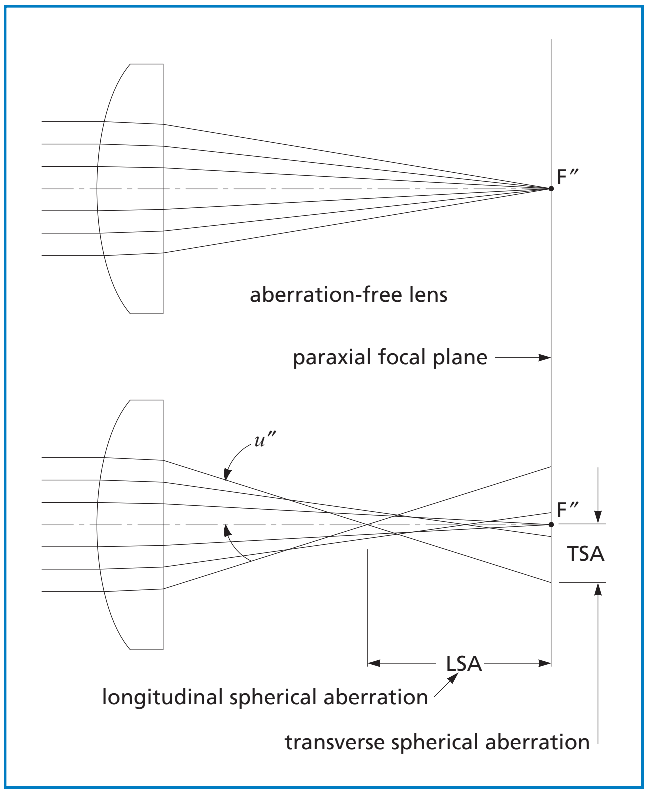
F″
aberration-free lens
paraxial focal plane
u″
F″
TSA
LSA
longitudinal spherical aberration
transverse spherical aberration

Figure 1.15 **Spherical aberration of a plano-convex lens**

order, monochromatic, spherical aberration of a plano-convex lens used at infinite conjugate ratio can be estimated by

---

* *Ludwig von Seidel, 1857.*

$$\text{spot size due to spherical aberration} = \frac{0.067\,f}{\text{f/\#}^3}. \qquad (1.23)$$

Theoretically, the simplest way to eliminate or reduce spherical aberration is to make the lens surface(s) with a varying radius of curvature (i.e., an aspheric surface) designed to exactly compensate for the fact that $\sin\theta \neq \theta$ at larger angles. In practice, however, most lenses with high surface accuracy are manufactured by grinding and polishing techniques that naturally produce spherical or cylindrical surfaces. The manufacture of aspheric surfaces is more complex, and it is difficult to produce a lens of sufficient surface accuracy to eliminate spherical aberration completely. Fortunately, these aberrations can be virtually eliminated, for a chosen set of conditions, by combining the effects of two or more spherical (or cylindrical) surfaces.

In general, simple positive lenses have undercorrected spherical aberration, and negative lenses usually have overcorrected spherical aberration. By combining a positive lens made from low-index glass with a negative lens made from high-index glass, it is possible to produce a combination in which the spherical aberrations cancel but the focusing powers do not. The simplest examples of this are cemented doublets, such as the LAO series which produce minimal spherical aberration when properly used.

ASTIGMATISM

When an off-axis object is focused by a spherical lens, the natural asymmetry leads to astigmatism. The system appears to have two different focal lengths.

As shown in figure 1.16, the plane containing both optical axis and object point is called the tangential plane. Rays that lie in this plane are called tangential, or meridional, rays. Rays not in this plane are referred to as skew rays. The chief, or principal, ray goes from the object point through the center of the aperture of the lens system. The plane perpendicular to the tangential plane that contains the principal ray is called the sagittal or radial plane.

The figure illustrates that tangential rays from the object come to a focus closer to the lens than do rays in the sagittal plane. When the image is evaluated at the tangential conjugate, we see a line in the sagittal direction. A line in the tangential direction is formed at the sagittal conjugate. Between these conjugates, the image is either an elliptical or a circular blur. Astigmatism is defined as the separation of these conjugates.

The amount of astigmatism in a lens depends on lens shape only when there is an aperture in the system that is not in contact with the lens itself. (In all optical systems there is an aperture or stop, although in many cases it is simply the clear aperture of the lens element itself.) Astigmatism strongly depends on the conjugate ratio.

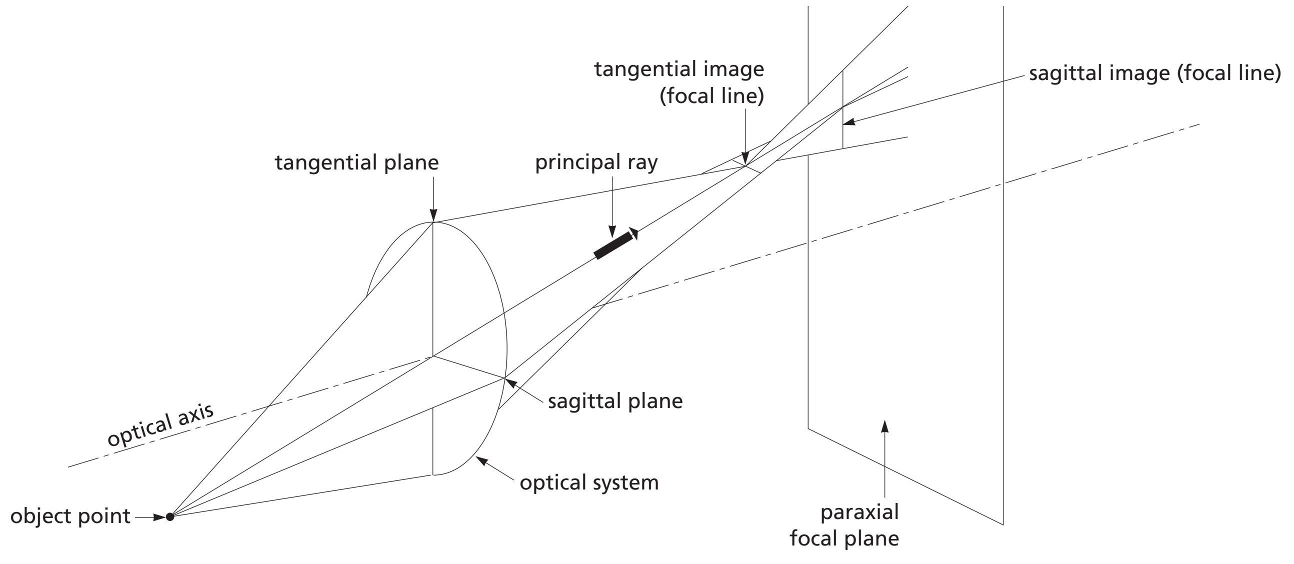

Figure 1.16 **Astigmatism represented by sectional views**

COMA

In spherical lenses, different parts of the lens surface exhibit different degrees of magnification. This gives rise to an aberration known as coma. As shown in figure 1.17, each concentric zone of a lens forms a ring-shaped image called a comatic circle. This causes blurring in the image plane (surface) of off-axis object points. An off-axis object point is not a sharp image point, but it appears as a characteristic comet-like flare. Even if spherical aberration is corrected and the lens brings all rays to a sharp focus on axis, a lens may still exhibit coma off axis. See figure 1.18.

As with spherical aberration, correction can be achieved by using multiple surfaces. Alternatively, a sharper image may be produced by judiciously placing an aperture, or stop, in an optical system to eliminate the more marginal rays.

FIELD CURVATURE

Even in the absence of astigmatism, there is a tendency of optical systems to image better on curved surfaces than on flat planes. This effect is called field curvature (see figure 1.19). In the presence of astigmatism, this problem is compounded because two separate astigmatic focal surfaces correspond to the tangential and sagittal conjugates.

Field curvature varies with the square of field angle or the square of image height. Therefore, by reducing the field angle by one-half, it is possible to reduce the blur from field curvature to a value of 0.25 of its original size.

Positive lens elements usually have inward curving fields, and negative lenses have outward curving fields. Field curvature can thus be corrected to some extent by combining positive and negative lens elements.

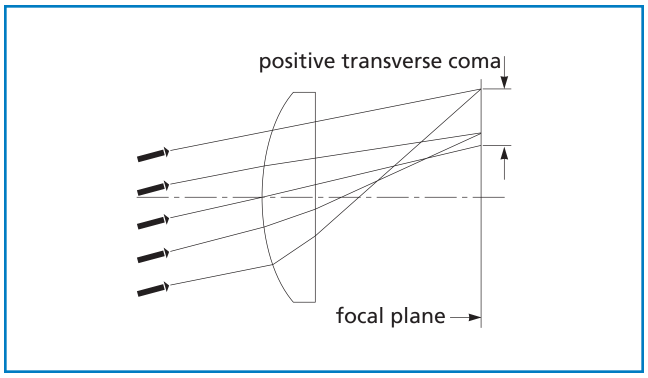

Figure 1.18 **Positive transverse coma**

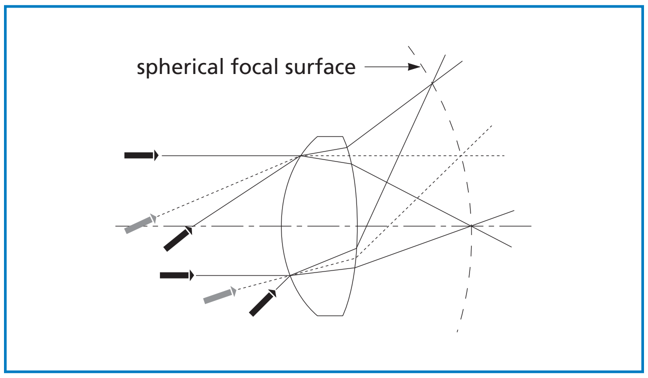

Figure 1.19 **Field curvature**

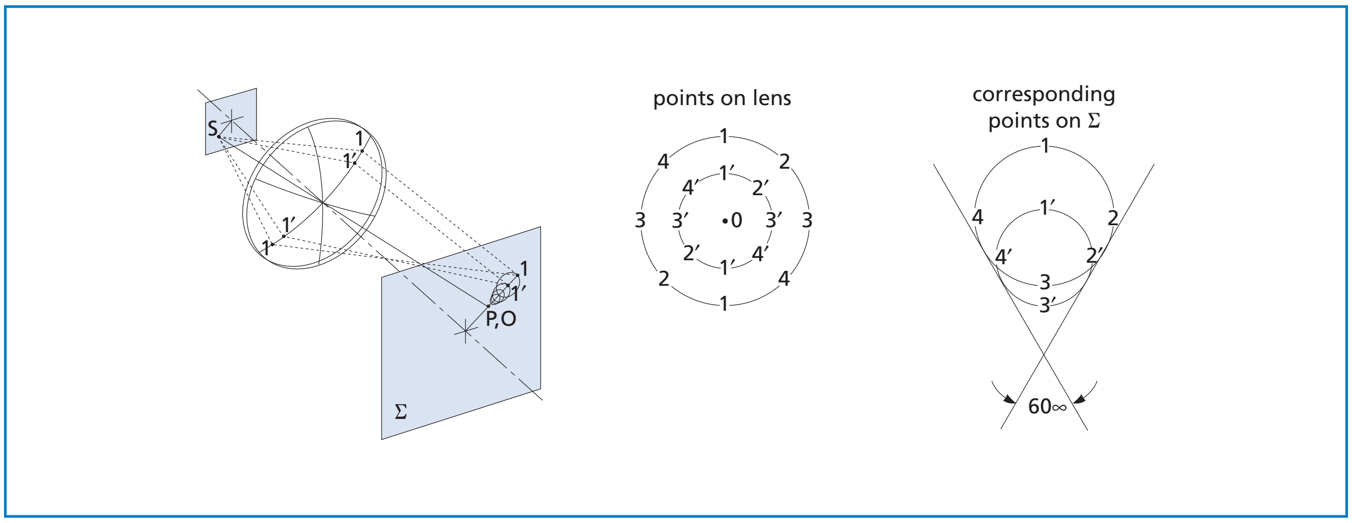

Figure 1.17 **Imaging an off-axis point source by a lens with positive transverse coma**

DISTORTION

The image field not only may have curvature but may also be distorted. The image of an off-axis point may be formed at a location on this surface other than that predicted by the simple paraxial equations. This distortion is different from coma (where rays from an off-axis point fail to meet perfectly in the image plane). Distortion means that even if a perfect off-axis point image is formed, its location on the image plane is not correct. Furthermore, the amount of distortion usually increases with increasing image height. The effect of this can be seen as two different kinds of distortion: pincushion and barrel (see figure 1.20). Distortion does not lower system resolution; it simply means that the image shape does not correspond exactly to the shape of the object. Distortion is a separation of the actual image point from the paraxially predicted location on the image plane and can be expressed either as an absolute value or as a percentage of the paraxial image height.

It should be apparent that a lens or lens system has opposite types of distortion depending on whether it is used forward or backward. This means that if a lens were used to make a photograph, and then used in reverse to project it, there would be no distortion in the final screen image. Also, perfectly symmetrical optical systems at 1:1 magnification have no distortion or coma.

CHROMATIC ABERRATION

The aberrations previously described are purely a function of the shape of the lens surfaces, and they can be observed with monochromatic light. Other aberrations, however, arise when these optics are used to transform light containing multiple wavelengths. The index of refraction of a material is a function of wavelength. Known as dispersion, this is discussed in *Material Properties*. From Snell's law (see equation 1.20), it can be seen that light rays of different wavelengths or colors will be refracted at different angles since the index is not a constant. Figure 1.21 shows the result when polychromatic collimated light is incident on a positive lens element. Because the index of refraction is higher for shorter wavelengths, these are focused closer to the lens than the longer wavelengths. Longitudinal chromatic aberration is defined as the axial distance from the nearest to the farthest focal point. As in the case of spherical aberration, positive and negative elements have opposite signs of chromatic aberration. Once again, by combining elements of nearly opposite aberration to form a doublet, chromatic aberration can be partially corrected. It is necessary to use two glasses with different dispersion characteristics, so that the weaker negative element can balance the aberration of the stronger, positive element.

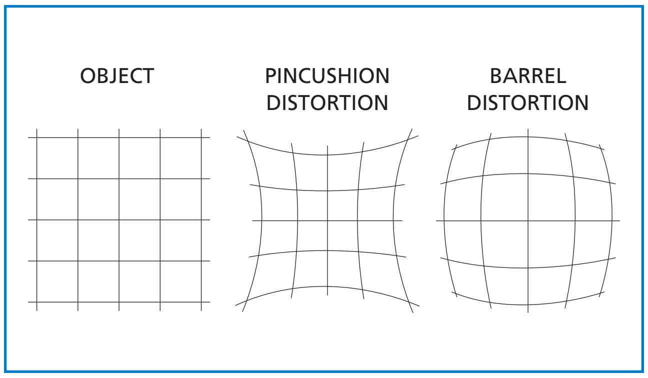

Figure 1.20 **Pincushion and barrel distortion**

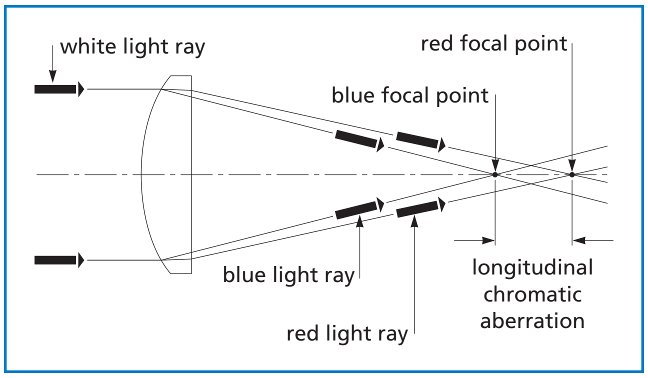

Figure 1.21 **Longitudinal chromatic aberration**

Variations of Aberrations with Aperture, Field Angle, and Image Height

Aberration	Aperture (ϕ)	Field Angle (θ)	Image Height (y)
Lateral Spherical	ϕ^3	—	—
Longitudinal Spherical	ϕ^2	—	—
Coma	ϕ^2	θ	y
Astigmatism	ϕ	θ^2	y^2
Field Curvature	ϕ	θ^2	y^2
Distortion	—	θ^3	y^3
Chromatic	—	—	—

LATERAL COLOR

Lateral color is the difference in image height between blue and red rays. Figure 1.22 shows the chief ray of an optical system consisting of a simple positive lens and a separate aperture. Because of the change in index with wavelength, blue light is refracted more strongly than red light, which is why rays intercept the image plane at different heights. Stated simply, magnification depends on color. Lateral color is very dependent on system stop location.

For many optical systems, the third-order term is all that may be needed to quantify aberrations. However, in highly corrected systems or in those having large apertures or a large angular field of view, third-order theory is inadequate. In these cases, exact ray tracing is absolutely essential.

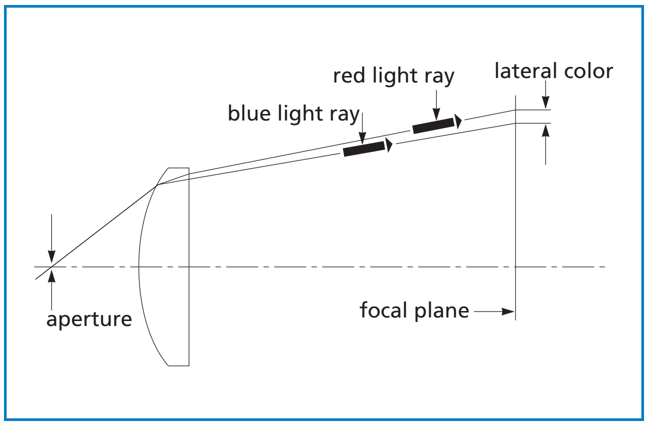

Figure 1.22 **Lateral Color**

APPLICATION NOTE

Achromatic Doublets Are Superior to Simple Lenses

Because achromatic doublets correct for spherical as well as chromatic aberration, they are often superior to simple lenses for focusing collimated light or collimating point sources, even in purely monochromatic light.

Although there is no simple formula that can be used to estimate the spot size of a doublet, the tables in *Spot Size* give sample values that can be used to estimate the performance of catalog achromatic doublets.

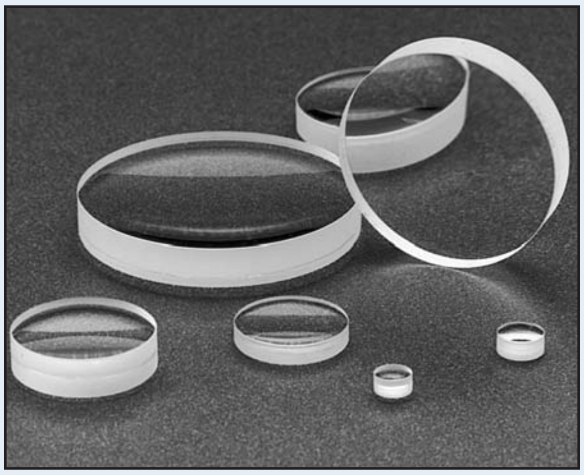

Lens Shape

Aberrations described in the preceding section are highly dependent on application, lens shape, and material of the lens (or, more exactly, its index of refraction). The singlet shape that minimizes spherical aberration at a given conjugate ratio is called best-form. The criterion for best-form at any conjugate ratio is that the marginal rays are equally refracted at each of the lens/air interfaces. This minimizes the effect of $\sin\theta \neq \theta$. It is also the criterion for minimum surface-reflectance loss. Another benefit is that absolute coma is nearly minimized for best-form shape, at both infinite and unit conjugate ratios.

To further explore the dependence of aberrations on lens shape, it is helpful to make use of the Coddington shape factor, q, defined as

$$q=\frac{(r_2+r_1)}{(r_2-r_1)}. \tag{1.24}$$

Figure 1.23 shows the transverse and longitudinal spherical aberrations of a singlet lens as a function of the shape factor, q. In this particular instance, the lens has a focal length of 100 mm, operates at f/5, has an index of refraction of 1.518722 (BK7 at the mercury green line, 546.1 nm), and is being operated at the infinite conjugate ratio. It is also assumed that the lens itself is the aperture stop. An asymmetric shape that corresponds to a q-value of about 0.7426 for this material and wavelength is the best singlet shape for on-axis imaging. It is important to note that the best-form shape is dependent on refractive index. For example, with a high-index material, such as silicon, the best-form lens for the infinite conjugate ratio is a meniscus shape.

At infinite conjugate with a typical glass singlet, the plano-convex shape ($q = 1$), with convex side toward the infinite conjugate, performs nearly as well as the best-form lens. Because a plano-convex lens costs much less to manufacture than an asymmetric biconvex singlet, these lenses are quite popular. Furthermore, this lens shape exhibits near-minimum total transverse aberration and near-zero coma when used off axis, thus enhancing its utility.

For imaging at unit magnification ($s = s'' = 2f$), a similar analysis would show that a symmetric biconvex lens is the best shape. Not only is spherical aberration minimized, but coma, distortion, and lateral chromatic aberration exactly cancel each other out. These results are true regardless of material index or wavelength, which explains the utility of symmetric convex lenses, as well as symmetrical optical systems in general. However, if a remote stop is present, these aberrations may not cancel each other quite as well.

For wide-field applications, the best-form shape is definitely not the optimum singlet shape, especially at the infinite conjugate ratio, since it yields maximum field curvature. The ideal shape is determined by the situation and may require rigorous ray-tracing analysis. It is possible to achieve much better correction in an optical system by using more than one element. The cases of an infinite conjugate ratio system and a unit conjugate ratio system are discussed in the following section.

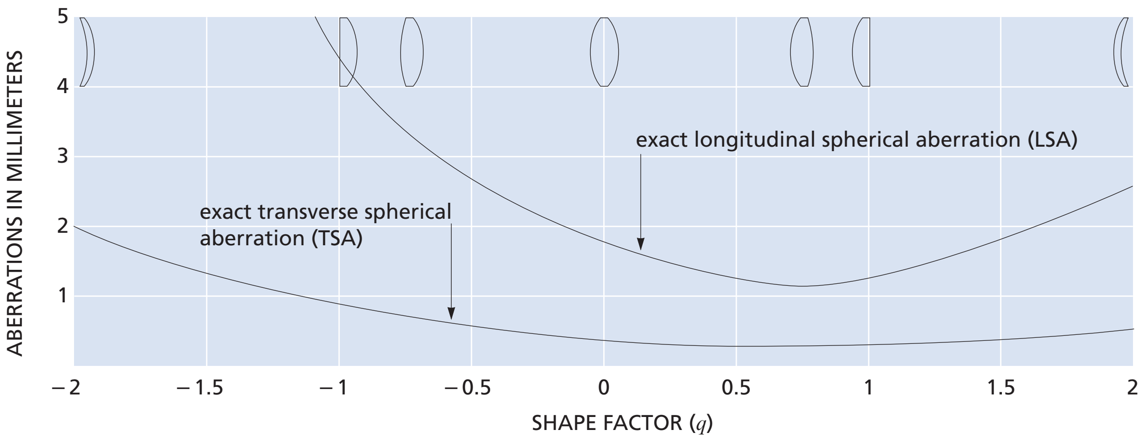

Figure 1.23 **Aberrations of positive singlets at infinite conjugate ratio as a function of shape**

Lens Combinations

INFINITE CONJUGATE RATIO

As shown in the previous discussion, the best-form singlet lens for use at infinite conjugate ratios is generally nearly plano-convex. Figure 1.24 shows a plano-convex lens (LPX-15.0-10.9-C) with incoming collimated light at a wavelength of 546.1 nm. This drawing, including the rays traced through it, is shown to exact scale. The marginal ray (ray f-number 1.5) strikes the paraxial focal plane significantly off the optical axis.

This situation can be improved by using a two-element system. The second part of the figure shows a precision achromat (LAO-21.0-14.0), which consists of a positive low-index (crown glass) element cemented to a negative meniscus high-index (flint glass) element. This is drawn to the same scale as the plano-convex lens. No spherical aberration can be discerned in the lens. Of course, not all of the rays pass exactly through the paraxial focal point; however, in this case, the departure is measured in micrometers, rather than in millimeters, as in the case of the plano-convex lens. Additionally, chromatic aberration (not shown) is much better corrected in the doublet. Even though these lenses are known as achromatic doublets, it is important to remember that even with monochromatic light the doublet's performance is superior.

Figure 1.24 also shows the f-number at which singlet performance becomes unacceptable. The ray with f-number 7.5 practically intercepts the paraxial focal point, and the f/3.8 ray is fairly close. This useful drawing, which can be scaled to fit a plano-convex lens of any focal length, can be used to estimate the magnitude of its spherical aberration, although lens thickness affects results slightly.

UNIT CONJUGATE RATIO

Figure 1.25 shows three possible systems for use at the unit conjugate ratio. All are shown to the same scale and using the same ray f-numbers with a light wavelength of 546.1 nm. The first system is a symmetric biconvex lens (LDX-21.0-19.2-C), the best-form singlet in this application. Clearly, significant spherical aberration is present in this lens at f/2.7. Not until f/13.3 does the ray closely approach the paraxial focus.

A dramatic improvement in performance is gained by using two identical plano-convex lenses with convex surfaces facing and nearly in contact. Those shown in figure 1.25 are both LPX-20.0-20.7-C. The combination of these two lenses yields almost exactly the same focal length as the biconvex lens. To understand why this configuration improves performance so dramatically, consider that if the biconvex lens were split down the middle, we would have two identical plano-convex lenses, each working at an infinite conjugate ratio, but with the convex surface toward the focus. This orientation is opposite to that shown to be optimum for this shape lens. On the other hand, if these lenses are reversed, we have the system just described but with a better correction of the spherical aberration.

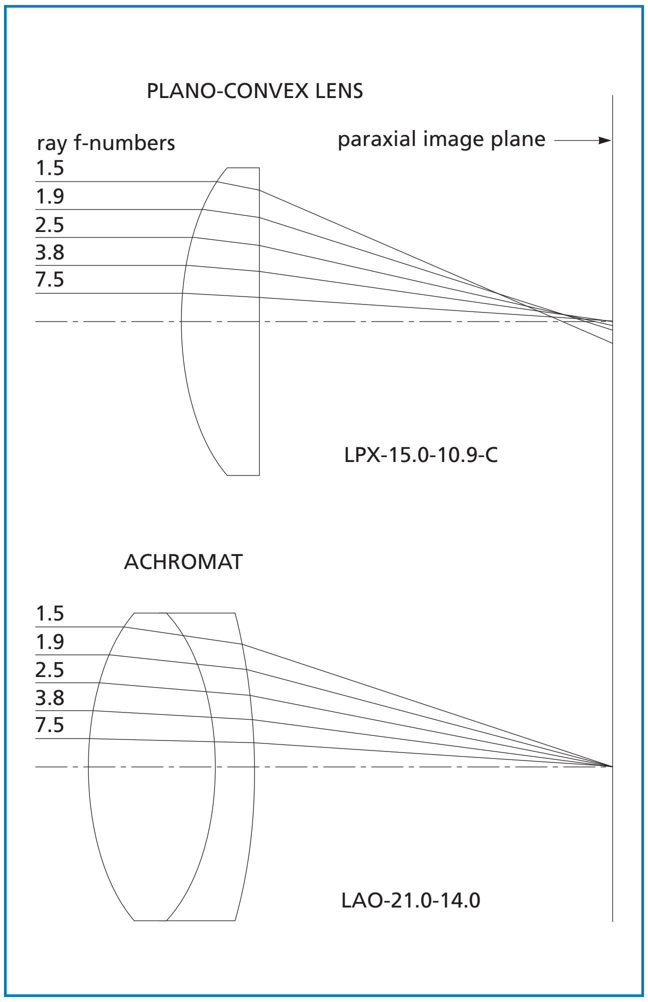

Figure 1.24 **Single-element plano-convex lens compared with a two-element achromat**

Previous examples indicate that an achromat is superior in performance to a singlet when used at the infinite conjugate ratio and at low f-numbers. Since the unit conjugate case can be thought of as two lenses, each working at the infinite conjugate ratio, the next step is to replace the plano-convex singlets with achromats, yielding a four-element system. The third part of figure 1.25 shows a system composed of two LAO-40.0-18.0 lenses. Once again, spherical aberration is not evident, even in the f/2.7 ray.

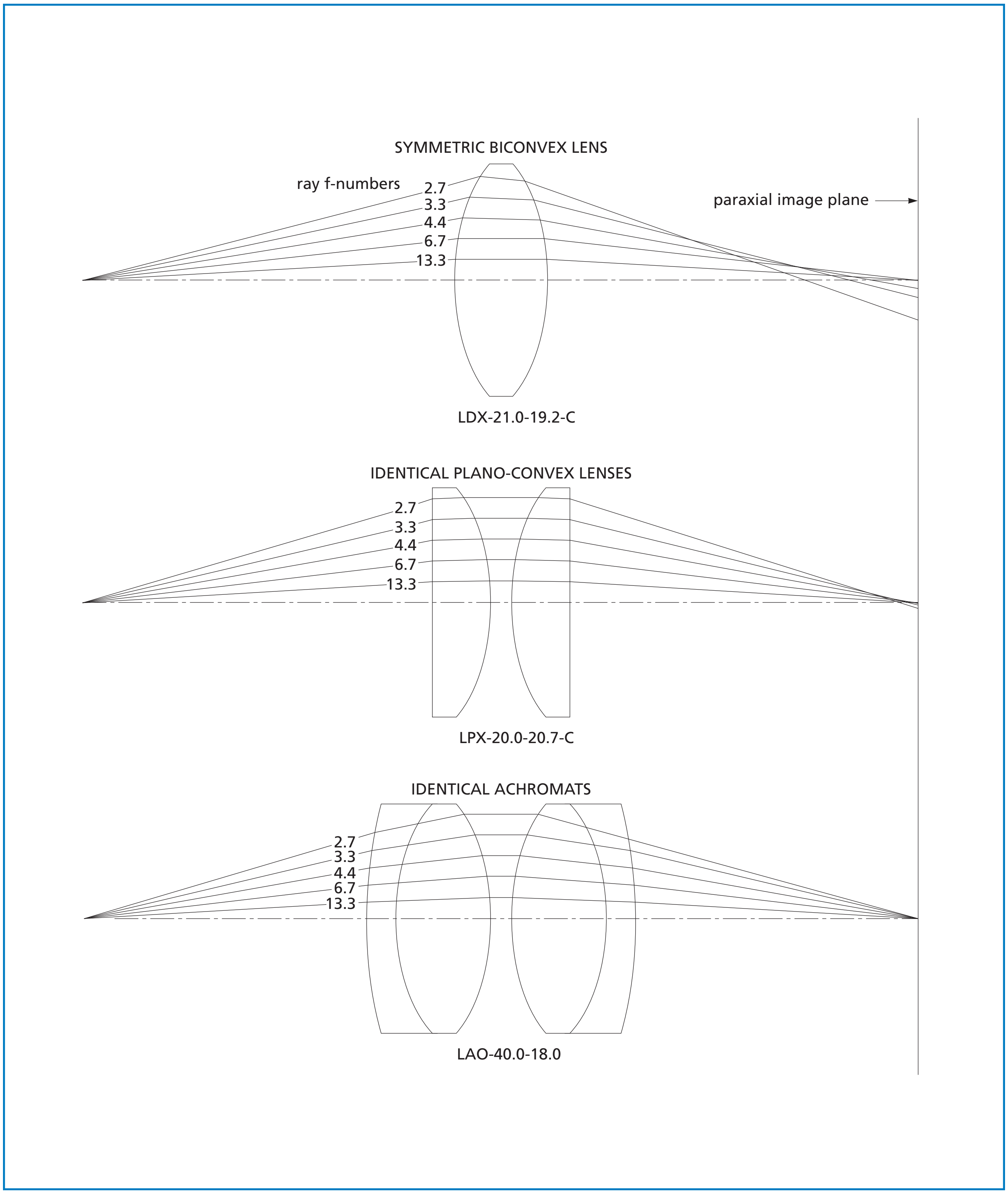
SYMMETRIC BICONVEX LENS
ray f-numbers
2.7
3.3
4.4
6.7
13.3
paraxial image plane
LDX-21.0-19.2-C
IDENTICAL PLANO-CONVEX LENSES
2.7
3.3
4.4
6.7
13.3
LPX-20.0-20.7-C
IDENTICAL ACHROMATS
2.7
3.3
4.4
6.7
13.3
LAO-40.0-18.0

Figure 1.25 **Three possible systems for use at the unit conjugate ratio**

Diffraction Effects

In all light beams, some energy is spread outside the region predicted by geometric propagation. This effect, known as diffraction, is a fundamental and inescapable physical phenomenon. Diffraction can be understood by considering the wave nature of light. Huygens' principle (figure 1.26) states that each point on a propagating wavefront is an emitter of secondary wavelets. The propagating wave is then the envelope of these expanding wavelets. Interference between the secondary wavelets gives rise to a fringe pattern that rapidly decreases in intensity with increasing angle from the initial direction of propagation. Huygens' principle nicely describes diffraction, but rigorous explanation demands a detailed study of wave theory.

Diffraction effects are traditionally classified into either Fresnel or Fraunhofer types. Fresnel diffraction is primarily concerned with what happens to light in the immediate neighborhood of a diffracting object or aperture. It is thus only of concern when the illumination source is close to this aperture or object, known as the near field. Consequently, Fresnel diffraction is rarely important in most classical optical setups, but it becomes very important in such applications as digital optics, fiber optics, and near-field microscopy.

Fraunhofer diffraction, however, is often important even in simple optical systems. This is the light-spreading effect of an aperture when the aperture (or object) is illuminated with an infinite source (plane-wave illumination) and the light is sensed at an infinite distance (far-field) from this aperture.

From these overly simple definitions, one might assume that Fraunhofer diffraction is important only in optical systems with infinite conjugate, whereas Fresnel diffraction equations should be considered at finite conjugate ratios. Not so. A lens or lens system of finite positive focal length with plane-wave input maps the far-field diffraction pattern of its aperture onto the focal plane; therefore, it is Fraunhofer diffraction that determines the limiting performance of optical systems. More generally, at any conjugate ratio, far-field angles are transformed into spatial displacements in the image plane.

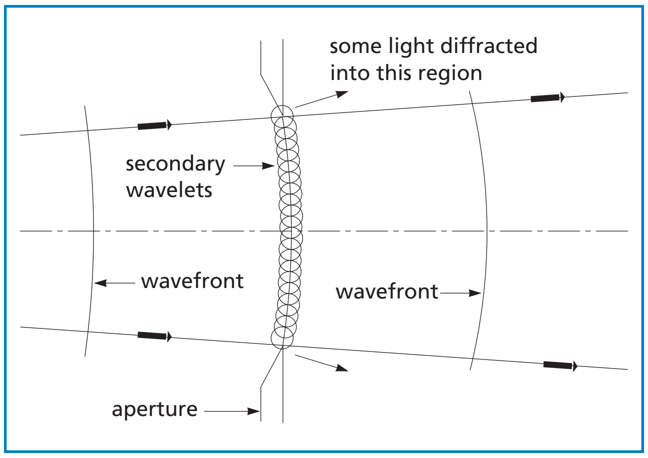

Figure 1.26 **Huygens' principle**

CIRCULAR APERTURE

Fraunhofer diffraction at a circular aperture dictates the fundamental limits of performance for circular lenses. It is important to remember that the spot size, caused by diffraction, of a circular lens is

$$d = 2.44\lambda(f/\#) \qquad (1.25)$$

where d is the diameter of the focused spot produced from plane-wave illumination and λ is the wavelength of light being focused. Notice that it is the f-number of the lens, not its absolute diameter, that determines this limiting spot size.

The diffraction pattern resulting from a uniformly illuminated circular aperture actually consists of a central bright region, known as the Airy disc (see figure 1.27), which is surrounded by a number of much fainter rings. Each ring is separated by a circle of zero intensity. The irradiance distribution in this pattern can be described by

$$I_x = I_0 \left[\frac{2J_1(x)}{x}\right]^2 \qquad (1.26)$$

where

I_0 = peak irradiance in the image

$J_1(x)$ = Bessel function of the first kind of order unity

$$= x\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^{2n-2}}{(n-1)!\,n!\,2^{2n-1}}$$

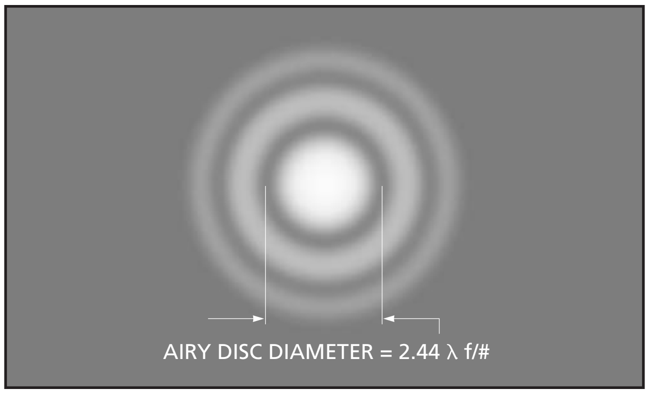

Figure 1.27 **Center of a typical diffraction pattern for a circular aperture**

Fundamental Optics

Gaussian Beam Optics

Optical Specifications

Material Properties

Optical Coatings

and

$$x = \frac{\pi D}{\lambda}\sin\theta$$

where

λ = wavelength

D = aperture diameter

θ = angular radius from the pattern maximum.

This useful formula shows the far-field irradiance distribution from a uniformly illuminated circular aperture of diameter D.

SLIT APERTURE

A slit aperture, which is mathematically simpler, is useful in relation to cylindrical optical elements. The irradiance distribution in the diffraction pattern of a uniformly illuminated slit aperture is described by

$$I_x = I_0\left[\frac{\sin x}{x}\right]^2 \tag{1.27}$$

where I_0 = peak irradiance in image

$$x = \frac{\pi w \sin\theta}{\lambda}$$

where λ = wavelength

w = slit width

θ = angular deviation from pattern maximum.

APPLICATION NOTE

Rayleigh Criterion

In imaging applications, spatial resolution is ultimately limited by diffraction. Calculating the maximum possible spatial resolution of an optical system requires an arbitrary definition of what is meant by resolving two features. In the Rayleigh criterion, it is assumed that two separate point sources can be resolved when the center of the Airy disc from one overlaps the first dark ring in the diffraction pattern of the second. In this case, the smallest resolvable distance, d, is

$$d = \frac{0.61\lambda}{\text{NA}} = 1.22\lambda(\text{f/\#})$$

ENERGY DISTRIBUTION TABLE

The accompanying table shows the major features of pure (unaberrated) Fraunhofer diffraction patterns of circular and slit apertures. The table shows the position, relative intensity, and percentage of total pattern energy corresponding to each ring or band. It is especially convenient to characterize positions in either pattern with the same variable x. This variable is related to field angle in the circular aperture case by

$$\sin\theta = \frac{\lambda x}{\pi D} \tag{1.28}$$

where D is the aperture diameter. For a slit aperture, this relationship is given by

$$\sin\theta = \frac{\lambda x}{\pi w} \tag{1.29}$$

where w is the slit width, π has its usual meaning, and D, w, and λ are all in the same units (preferably millimeters). Linear instead of angular field positions are simply found from $r=s''\tan\theta$ where s'' is the secondary conjugate distance. This last result is often seen in a different form, namely the diffraction-limited spot-size equation, which, for a circular lens is

$$d = 2.44\,\lambda(\text{f/\#}) \qquad \text{(see eq. 1.25)}$$

This value represents the smallest spot size that can be achieved by an optical system with a circular aperture of a given f-number, and it is the diameter of the first dark ring, where the intensity has dropped to zero.

The graph in figure 1.28 shows the form of both circular and slit aperture diffraction patterns when plotted on the same normalized scale. Aperture diameter is equal to slit width so that patterns between x-values and angular deviations in the far-field are the same.

GAUSSIAN BEAMS

Apodization, or nonuniformity of aperture irradiance, alters diffraction patterns. If pupil irradiance is nonuniform, the formulas and results given previously do not apply. This is important to remember because most laser-based optical systems do not have uniform pupil irradiance. The output beam of a laser operating in the TEM_{00} mode has a smooth Gaussian irradiance profile. Formulas used to determine the focused spot size from such a beam are discussed in *Gaussian Beam Optics*. Furthermore, when dealing with Gaussian beams, the location of the focused spot also departs from that predicted by the paraxial equations given in this chapter. This is also detailed in *Gaussian Beam Optics*.

Energy Distribution in the Diffraction Pattern of a Circular or Slit Aperture

	Circular Aperture			Slit Aperture		
Ring or Band	Position (x)	Relative Intensity (I_x/I_0)	Energy in Ring (%)	Position (x)	Relative Intensity (I_x/I_0)	Energy in Band (%)
Central Maximum	0.0	1.0	83.8	0.0	1.0	90.3
First Dark	1.22π	0.0		1.00π	0.0	
First Bright	1.64π	0.0175	7.2	1.43π	0.0472	4.7
Second Dark	2.23π	0.0		2.00π	0.0	
Second Bright	2.68π	0.0042	2.8	2.46π	0.0165	1.7
Third Dark	3.24π	0.0		3.00π	0.0	
Third Bright	3.70π	0.0016	1.5	3.47π	0.0083	0.8
Fourth Dark	4.24π	0.0		4.00π	0.0	
Fourth Bright	4.71π	0.0008	1.0	4.48π	0.0050	0.5
Fifth Dark	5.24π	0.0		5.00π	0.0	

Note: Position variable (x) is defined in the text.

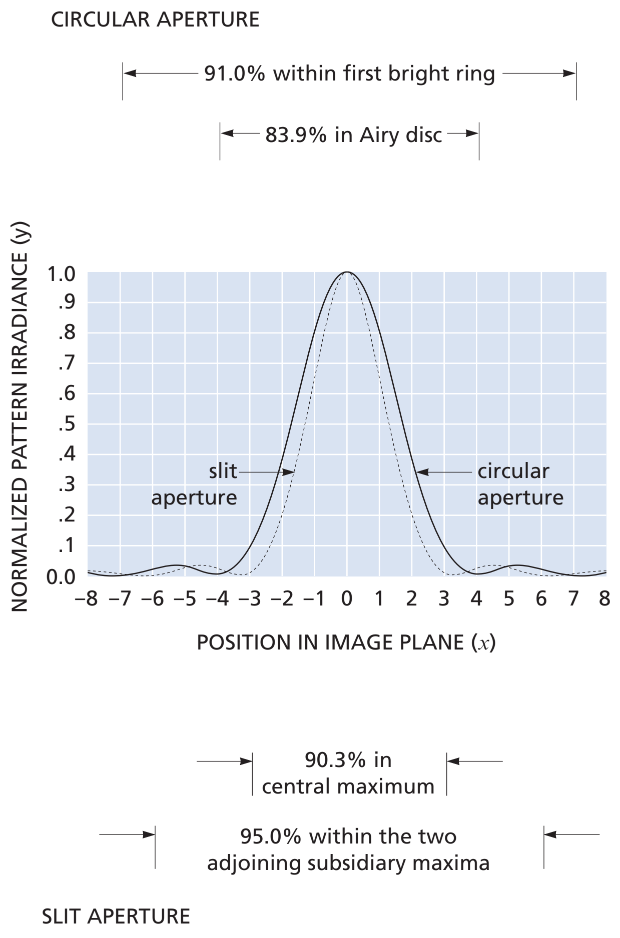

Figure 1.28 **Fraunhofer diffraction pattern of a singlet slit superimposed on the Fraunhofer diffraction pattern of a circular aperture**

Lens Selection

Having discussed the most important factors that affect the performance of a lens or a lens system, we will now address the practical matter of selecting the optimum catalog components for a particular task. The following useful relationships are important to keep in mind throughout the selection process:

- Diffraction-limited spot size = 2.44 λ f/#
- Approximate on-axis spot size of a plano-convex lens at the infinite conjugate resulting from spherical aberration $= \dfrac{0.067\,f}{\text{f}\,/\#^3}$
- Optical invariant $= m = \dfrac{\text{NA}}{\text{NA}''}$.

Example 1: Collimating an Incandescent Source

Produce a collimated beam from a quartz halogen bulb having a 1-mm-square filament. Collect the maximum amount of light possible and produce a beam with the lowest possible divergence angle.

This problem, illustrated in figure 1.29, involves the typical tradeoff between light-collection efficiency and resolution (where a beam is being collimated rather than focused, resolution is defined by beam divergence). To collect more light, it is necessary to work at a low f-number, but because of aberrations, higher resolution (lower divergence angle) will be achieved by working at a higher f-number.

In terms of resolution, the first thing to realize is that the minimum divergence angle (in radians) that can be achieved using any lens system is the source size divided by system focal length. An off-axis ray (from the edge of the source) entering the first principal point of the system exits the second principal point at the same angle. Therefore, increasing the system focal length improves this limiting divergence because the source appears smaller.

An optic that can produce a spot size of 1 mm when focusing a perfectly collimated beam is therefore required. Since source size is inherently limited, it is pointless to strive for better resolution. This level of resolution can be achieved easily with a plano-convex lens.

While angular divergence decreases with increasing focal length, spherical aberration of a plano-convex lens increases with increasing focal length. To determine the appropriate focal length, set the spherical aberration formula for a plano-convex lens equal to the source (spot) size:

$$\frac{0.067\,f}{\text{f}\,/\#^3} = 1\ \text{mm}.$$

This ensures a lens that meets the minimum performance needed. To select a focal length, make an arbitrary f-number choice. As can be seen from the relationship, as we lower the f-number (increase collection efficiency), we decrease the focal length, which will worsen the resultant divergence angle (minimum divergence = 1 mm/f).

In this example, we will accept f/2 collection efficiency, which gives us a focal length of about 120 mm. For f/2 operation we would need a minimum diameter of 60 mm. The LPX-60.0-62.2-C fits this specification exactly. Beam divergence would be about 8 mrad.

Finally, we need to verify that we are not operating below the theoretical diffraction limit. In this example, the numbers (1-mm spot size) indicate that we are not, since

$$\text{diffraction-limited spot size} = 2.44 \times 0.5\ \mu\text{m} \times 2 = 2.44\ \mu\text{m}.$$

Example 2: Coupling an Incandescent Source into a Fiber

In *Imaging Properties of Lens Systems* we considered a system in which the output of an incandescent bulb with a filament of 1 mm in diameter was to be coupled into an optical fiber with a core diameter of 100 µm and a numerical aperture of 0.25. From the optical invariant and other constraints given in the problem, we determined that $f = 9.1$ mm, CA = 5 mm, $s = 100$ mm, $s'' = 10$ mm, NA″ = 0.25, and NA = 0.025 (or f/2 and f/20). The singlet lenses that match these specifications are the plano-convex LPX-5.0-5.2-C or biconvex lenses LDX-6.0-7.7-C and LDX-5.0-9.9-C. The closest achromat would be the LAO-10.0-6.0.

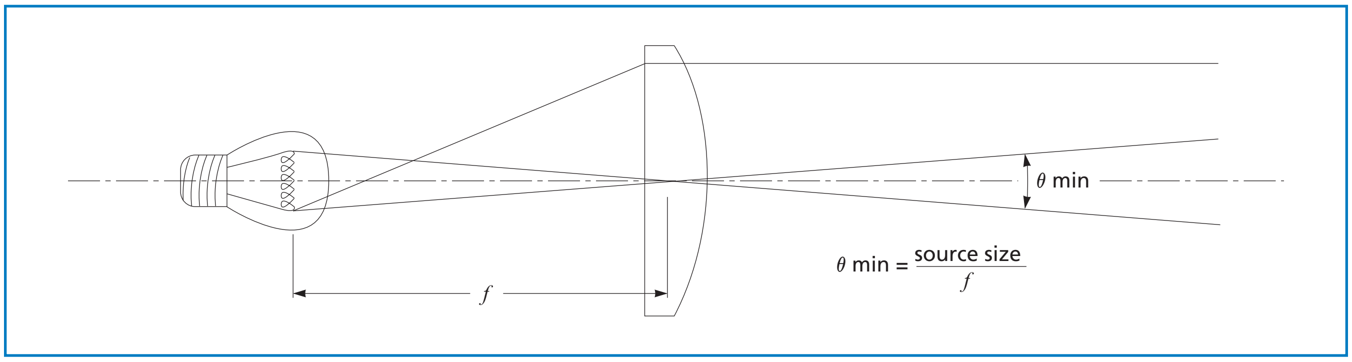

Figure 1.29 **Collimating an incandescent source**

We can immediately reject the biconvex lenses because of spherical aberration. We can estimate the performance of the LPX-5.0-5.2-C on the focusing side by using our spherical aberration formula:

$$\text{spot size} = \frac{0.067\ (10)}{2^3} = 84\ \mu\text{m}.$$

We will ignore, for the moment, that we are not working at the infinite conjugate.

This is slightly smaller than the 100-µm spot size we are trying to achieve. However, since we are not working at infinite conjugate, the spot size will be larger than that given by our simple calculation. This lens is therefore likely to be marginal in this situation, especially if we consider chromatic aberration. A better choice is the achromat. Although a computer ray trace would be required to determine its exact performance, it is virtually certain to provide adequate performance.

Example 3: Symmetric Fiber-to-Fiber Coupling

Couple an optical fiber with an 8-µm core and a 0.15 numerical aperture into another fiber with the same characteristics. Assume a wavelength of 0.5 µm.

This problem, illustrated in figure 1.30, is essentially a 1:1 imaging situation. We want to collect and focus at a numerical aperture of 0.15 or f/3.3, and we need a lens with an 8-µm spot size at this f-number. Based on the lens combination discussion in *Lens Combination Formulas*, our most likely setup is either a pair of identical plano-convex lenses or achromats, faced front to front. To determine the necessary focal length for a plano-convex lens, we again use the spherical aberration estimate formula:

$$\frac{0.067\ f}{3.3^3} = 0.008\ \text{mm}.$$

This formula yields a focal length of 4.3 mm and a minimum diameter of 1.3 mm. The LPX-4.2-2.3-BAK1 meets these criteria. The biggest problem with utilizing these tiny, short focal length lenses is the practical considerations of handling, mounting, and positioning them. Because using a pair of longer focal length singlets would result in unacceptable performance, the next step might be to use a pair of the slightly longer focal length, larger achromats, such as the LAO-10.0-6.0. The performance data, given in *Spot Size*, show that this combination does provide the required 8-μm spot diameter.

Because fairly small spot sizes are being considered here, it is important to make sure that the system is not being asked to work below the diffraction limit:

$$2.44 \times 0.5\ \mu\text{m} \times 3.3 = 4\ \ \mu\text{m}\,.$$

Since this is half the spot size caused by aberrations, it can be safely assumed that diffraction will not play a significant role here.

An entirely different approach to a fiber-coupling task such as this would be to use a pair of spherical ball lenses (LMS-LSFN series) or one of the gradient-index lenses (LGT series).

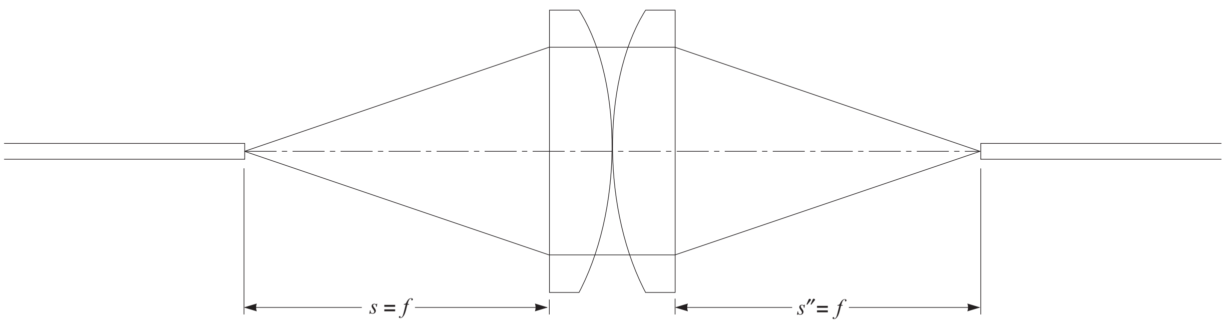

Figure 1.30 **Symmetric fiber-to-fiber coupling**

Example 4: Diffraction-Limited Performance

Determine at what f-number a plano-convex lens being used at an infinite conjugate ratio with 0.5-μm wavelength light becomes diffraction limited (i.e., the effects of diffraction exceed those caused by aberration).

To solve this problem, set the equations for diffraction-limited spot size and third-order spherical aberration equal to each other. The result depends upon focal length, since aberrations scale with focal length, while diffraction is solely dependent upon f-number. By substituting some common focal lengths into this formula, we get f/8.6 at $f = 100$ mm, f/7.2 at $f = 50$ mm, and f/4.8 at $f = 10$ mm.

$$2.44 \times 0.5\ \mu\text{m} \times \text{f}/\# = \frac{0.067 \times f}{\text{f}/\#^3}$$

or

$$\text{f}/\# = (54.9 \times f)^{1/4}.$$

When working with these focal lengths (and under the conditions previously stated), we can assume essentially diffraction-limited performance above these f-numbers. Keep in mind, however, that this treatment does not take into account manufacturing tolerances or chromatic aberration, which will be present in polychromatic applications.

APPLICATION NOTE

Spherical Ball Lenses for Fiber Coupling

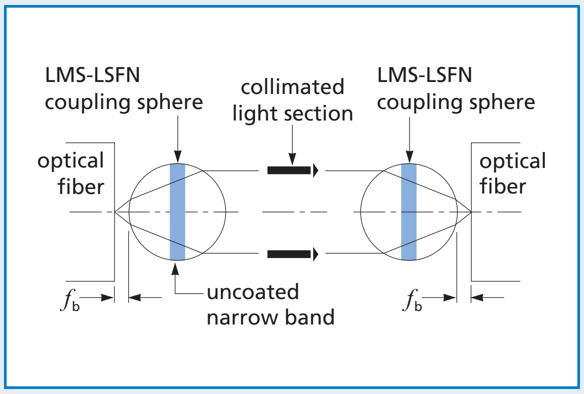

Spheres are arranged so that the fiber end is located at the focal point. The output from the first sphere is then collimated. If two spheres are aligned axially to each other, the beam will be transferred from one focal point to the other. Translational alignment sensitivity can be reduced by enlarging the beam. Slight negative defocusing of the ball can reduce the spherical aberration third-order contribution common to all coupling systems. Additional information can be found in "Lens Coupling in Fiber Optic Devices: Efficiency Limits," by A. Nicia, *Applied Optics*, vol. 20, no. 18, pp 3136–45, 1981. Off-axis aberrations are absent since the fiber diameters are so much smaller than the coupler focal length.

Spot Size

In general, the performance of a lens or lens system in a specific circumstance should be determined by an exact trigonometric ray trace. CVI Melles Griot applications engineers can supply ray-tracing data for particular lenses and systems of catalog components on request. In certain situations, however, some simple guidelines can be used for lens selection. The optimum working conditions for some of the lenses in this catalog have already been presented. The following tables give some quantitative results for a variety of simple and compound lens systems, which can be constructed from standard catalog optics.

In interpreting these tables, remember that these theoretical values obtained from computer ray tracing consider only the effects of ideal geometric optics. Effects of manufacturing tolerances have not been considered. Furthermore, remember that using more than one element provides a higher degree of correction but makes alignment more difficult. When actually choosing a lens or a lens system, it is important to note the tolerances and specifications clearly described for each CVI Melles Griot lens in the product listings.

The tables give the diameter of the spot for a variety of lenses used at several different f-numbers. All the tables are for on-axis, uniformly illuminated, collimated input light at 546.1 nm. They assume that the lens is facing in the direction that produces a minimum spot size. When the spot size caused by aberrations is smaller or equal to the diffraction-limited spot size, the notation "DL" appears next to the entry. The shorter focal length lenses produce smaller spot sizes because aberrations increase linearly as a lens is scaled up.

The effect on spot size caused by spherical aberration is strongly dependent on f-number. For a plano-convex singlet, spherical aberration is inversely dependent on the cube of the f-number. For doublets, this relationship can be even higher. On the other hand, the spot size caused by diffraction increases linearly with f-number. Thus, for some lens types, spot size at first decreases and then increases with f-number, meaning that there is some optimum performance point at which both aberrations and diffraction combine to form a minimum.

Unfortunately, these results cannot be generalized to situations in which the lenses are used off axis. This is particularly true of the achromat/aplanatic meniscus lens combinations because their performance degrades rapidly off axis.

Focal Length = 10 mm

	Spot Size (μm)*		
f/#	LDX-5.0-9.9-C	LPX-8.0-5.2-C	LAO-10.0-6.0
f/2	—	94	11
f/3	36	25	7
f/5	8	6.7 (DL)	6.7 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.3 (DL)

Diffraction-limited performance is indicated by DL.

Focal Length = 30 mm

	Spot Size (μm)*		
f/#	LPX-18.5-15.6-C	LAO-30.0-12.5	LAO-50.0-18.0 & MENP-18.0-4.0-73.5-NSF8
f/2	295	—	3
f/3	79	7	4 (DL)
f/5	17	6.7 (DL)	6.9 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.8 (DL)

Diffraction-limited performance is indicated by DL.

Focal Length = 60 mm

	Spot Size (μm)*			
f/#	LDX-50.0-60.0-C	LPX-30.0-31.1-C	LAO-60.0-30.0	LAO-100.0-31.5 & MENP-31.5-6.0-146.4-NSF8
f/2	816	600	—	—
f/3	217	160	34	4 (DL)
f/5	45	33	10	6.7 (DL)
f/10	13.3 (DL)	13.3 (DL)	13.3 (DL)	13.3 (DL)

Diffraction-limited performance is indicated by DL.

Aberration Balancing

To improve system performance, optical designers make sure that the total aberration contribution from all surfaces taken together sums to nearly zero. Normally, such a process requires computerized analysis and optimization. However, there are some simple guidelines that can be used to achieve this with lenses available in this catalog. This approach can yield systems that operate at a much lower f-number than can usually be achieved with simple lenses.

Specifically, we will examine how to null the spherical aberration from two or more lenses in collimated, monochromatic light. This technique will thus be most useful for laser beam focusing and expanding.

Figure 1.31 shows the third-order longitudinal spherical aberration coefficients for six of the most common positive and negative lens shapes when used with parallel, monochromatic incident light. The plano-convex and plano-concave lenses both show minimum spherical aberration when oriented with their curved surface facing the incident parallel beam. All other configurations exhibit larger amounts of spherical aberration. With these lens types, it is now possible to show how various systems can be corrected for spherical aberration.

A two-element laser beam expander is a good starting example. In this case, two lenses are separated by a distance that is the sum of their focal lengths, so that the overall system focal length is infinite. This system will not focus incoming collimated light, but it will change the beam diameter. By definition, each of the lenses is operating at the same f-number.

The equation for longitudinal spherical aberration shows that, for two lenses with the same f-number, aberration varies directly with the focal lengths of the lenses. The sign of the aberration is the same as focal length. Thus, it should be possible to correct the spherical aberration of this Galilean-type beam expander, which consists of a positive focal length objective and a negative diverging lens.

If a plano-convex lens of focal length f_1 oriented in the normal direction is combined with a plano-concave lens of focal length f_2 oriented in its reverse direction, the total spherical aberration of the system is

$$\mathrm{LSA} = \frac{0.272\, f_1}{\mathrm{f/\#}^2} + \frac{1.069\, f_2}{\mathrm{f/\#}^2}.$$

After setting this equation to zero, we obtain

$$\frac{f_1}{f_2} = -\frac{1.069}{0.272} = -3.93.$$

To make the magnitude of aberration contributions of the two elements equal so they will cancel out, and thus correct the system, select the focal length of the positive element to be 3.93 times that of the negative element.

Figure 1.32(a) shows a beam-expander system made up of catalog elements, in which the focal length ratio is 4:1. This simple system is corrected to about 1/6 wavelength at 632.8 nm, even though the objective is operating at f/4 with a 20-mm aperture diameter. This is remarkably good wavefront correction for such a simple system; one would normally assume that a doublet objective would be needed and a complex diverging lens as well. This analysis does not take into account manufacturing tolerances.

A beam expander of lower magnification can also be derived from this information. If a symmetric-convex objective is used together] with a reversed plano-concave diverging lens, the aberration coefficients are in the ratio of 1.069/0.403 = 2.65. Figure 1.32(b) shows a system of catalog lenses that provides a magnification of 2.7 (the closest possible given the available focal lengths). The maximum wavefront error in this case is only a quarter-wave, even though the objective is working at f/3.3.

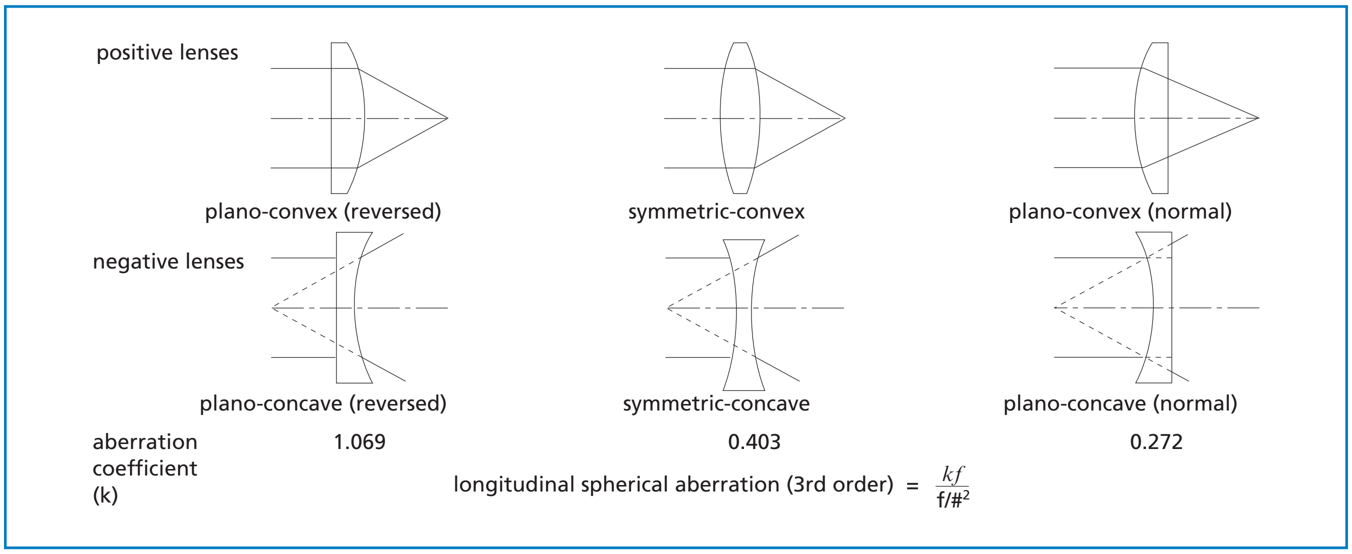

longitudinal spherical aberration (3rd order) = $\frac{kf}{\mathrm{f/\#}^2}$

Figure 1.31 **Third-order longitudinal spherical aberration of typical lens shapes**

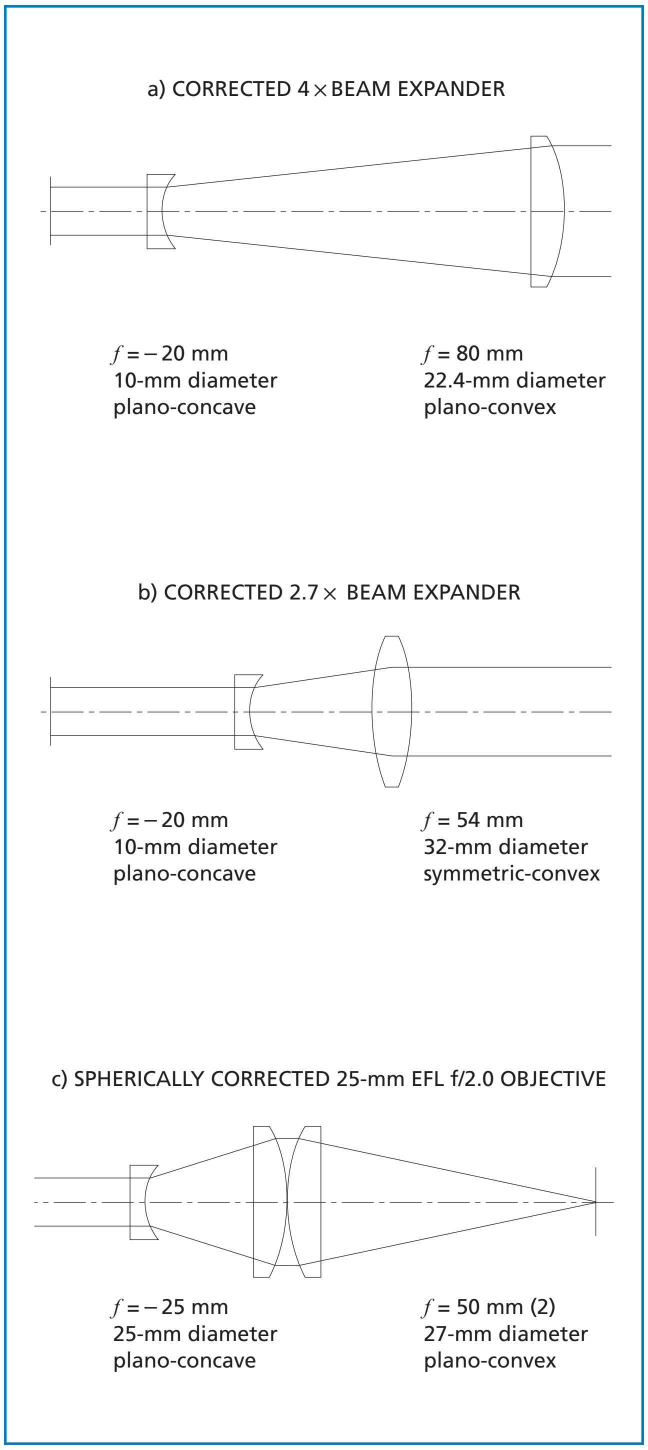
a) CORRECTED 4 × BEAM EXPANDER
f = – 20 mm
10-mm diameter
plano-concave
f = 80 mm
22.4-mm diameter
plano-convex
b) CORRECTED 2.7 × BEAM EXPANDER
f = – 20 mm
10-mm diameter
plano-concave
f = 54 mm
32-mm diameter
symmetric-convex
c) SPHERICALLY CORRECTED 25-mm EFL f/2.0 OBJECTIVE
f = – 25 mm
25-mm diameter
plano-concave
f = 50 mm (2)
27-mm diameter
plano-convex

Figure 1.32 **Combining catalog lenses for aberration balancing**

The relatively low f numbers of these objectives is a great advantage in minimizing the length of these beam expanders. They would be particularly useful with Nd:YAG and argon-ion lasers, which tend to have large output beam diameters.

These same principles can be utilized to create high numerical aperture objectives that might be used as laser focusing lenses. Figure 1.32(c) shows an objective consisting of an initial negative element, followed by two identical plano-convex positive elements. Again, all of the elements operate at the same f-number, so that their aberration contributions are proportional to their focal lengths. To obtain zero total spherical aberration from this configuration, we must satisfy

$$1.069\, f_1 + 0.272\, f_2 + 0.272\, f_2 = 0$$

or

$$\frac{f_1}{f_2} = -0.51.$$

Therefore, a corrected system should result if the focal length of the negative element is just about half that of each of the positive lenses. In this case, $f_1 = 425$ mm and $f_2 = 50$ mm yield a total system focal length of about 25 mm and an f-number of approximately f/2. This objective, corrected to 1/6 wave, has the additional advantage of a very long working distance.

Definition of Terms

FOCAL LENGTH (*f*)

Two distinct terms describe the focal lengths associated with every lens or lens system. The effective focal length (EFL) or equivalent focal length (denoted f in figure 1.33) determines magnification and hence the image size. The term f appears frequently in lens formulas and in the tables of standard lenses. Unfortunately, because f is measured with reference to principal points which are usually inside the lens, the meaning of f is not immediately apparent when a lens is visually inspected.

The second type of focal length relates the focal plane positions directly to landmarks on the lens surfaces (namely the vertices) which are immediately recognizable. It is not simply related to image size but is especially convenient for use when one is concerned about correct lens positioning or mechanical clearances. Examples of this second type of focal length are the front focal length (FFL, denoted f_f in figure 1.33) and the back focal length (BFL, denoted f_b).

The convention in all of the figures (with the exception of a single deliberately reversed ray) is that light travels from left to right.

FOCAL POINT (F OR F″)

Rays that pass through or originate at either focal point must be, on the opposite side of the lens, parallel to the optical axis. This fact is the basis for locating both focal points.

PRIMARY PRINCIPAL SURFACE

Let us imagine that rays originating at the front focal point F (and therefore parallel to the optical axis after emergence from the opposite side of the lens) are singly refracted at some imaginary surface, instead of twice refracted (once at each lens surface) as actually happens. There is a unique imaginary surface, called the principal surface, at which this can happen.

To locate this unique surface, consider a single ray traced from the air on one side of the lens, through the lens and into the air on the other side. The ray is broken into three segments by the lens. Two of these are external (in the air), and the third is internal (in the glass). The external segments can be extended to a common point of intersection (certainly near, and usually within, the lens). The principal surface is the locus of all such

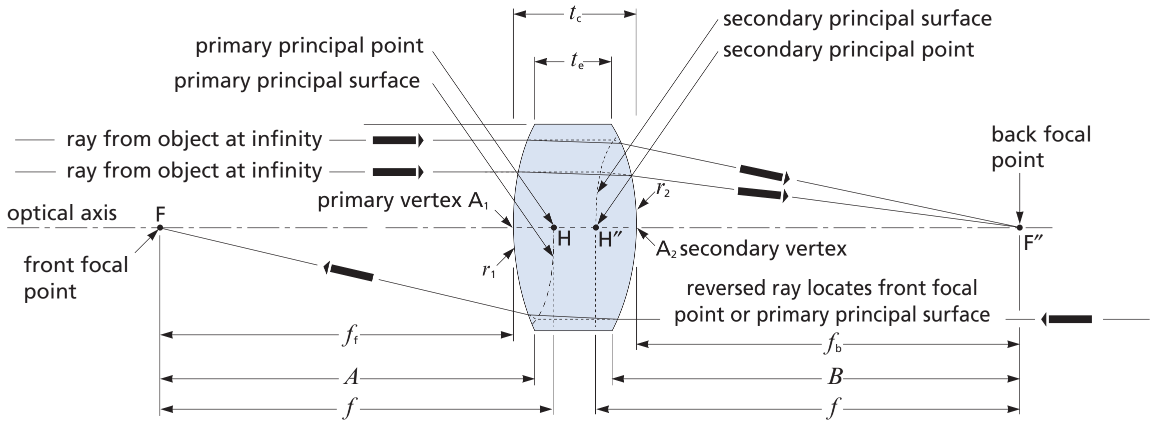
t_c
t_e
primary principal point
primary principal surface
secondary principal surface
secondary principal point
ray from object at infinity
ray from object at infinity
back focal point
optical axis
F
primary vertex A_1
r_2
H
H″
A_2 secondary vertex
F″
front focal point
r_1
reversed ray locates front focal point or primary principal surface
f_f
f_b
A
B
f
f

A = front focus to front edge distance

B = rear edge to rear focus distance

f = effective focal length; may be positive (as shown) or negative

f_f = front focal length

f_b = back focal length

t_e = edge thickness

t_c = center thickness

r_1 = radius of curvature of first surface (positive if center of curvature is to right)

r_2 = radius of curvature of second surface (negative if center of curvature is to left)

Figure 1.33 **Focal length and focal points**

points of intersection of extended external ray segments. The principal surface of a perfectly corrected optical system is a sphere centered on the focal point.

Near the optical axis, the principal surface is nearly flat, and for this reason, it is sometimes referred to as the principal plane.

SECONDARY PRINCIPAL SURFACE

This term is defined analogously to the primary principal surface, but it is used for a collimated beam incident from the left and focused to the back focal point F″ on the right. Rays in that part of the beam nearest the axis can be thought of as once refracted at the secondary principal surface, instead of being refracted by both lens surfaces.

PRIMARY PRINCIPAL POINT (H) OR FIRST NODAL POINT

This point is the intersection of the primary principal surface with the optical axis.

SECONDARY PRINCIPAL POINT (H″) OR SECONDARY NODAL POINT

This point is the intersection of the secondary principal surface with the optical axis.

CONJUGATE DISTANCES (s AND s'')

The conjugate distances are the object distance, s, and image distance, s''. Specifically, s is the distance from the object to H, and s'' is the distance from H″ to the image location. The term infinite conjugate ratio refers to the situation in which a lens is either focusing incoming collimated light or is being used to collimate a source (therefore, either s or s'' is infinity).

PRIMARY VERTEX (A_1)

The primary vertex is the intersection of the primary lens surface with the optical axis.

SECONDARY VERTEX (A_2)

The secondary vertex is the intersection of the secondary lens surface with the optical axis.

EFFECTIVE FOCAL LENGTH (EFL, f)

Assuming that the lens is surrounded by air or vacuum (refractive index 1.0), this is both the distance from the front focal point (F) to the primary principal point (H) and the distance from the secondary principal point (H″) to the rear focal point (F″). Later we use f to designate the paraxial EFL for the design wavelength (λ_0).

FRONT FOCAL LENGTH (f_f)

This length is the distance from the front focal point (F) to the primary vertex (A_1).

BACK FOCAL LENGTH (f_b)

This length is the distance from the secondary vertex (A_2) to the rear focal point (F″).

EDGE-TO-FOCUS DISTANCES (A AND B)

A is the distance from the front focal point to the primary vertex of the lens. B is the distance from the secondary vertex of the lens to the rear focal point. Both distances are presumed always to be positive.

REAL IMAGE

A real image is one in which the light rays actually converge; if a screen were placed at the point of focus, an image would be formed on it.

VIRTUAL IMAGE

A virtual image does not represent an actual convergence of light rays. A virtual image can be viewed only by looking back through the optical system, such as in the case of a magnifying glass.

F-NUMBER (f/#)

The f-number (also known as the focal ratio, relative aperture, or speed) of a lens system is defined to be the effective focal length divided by system clear aperture. Ray f-number is the conjugate distance for that ray divided by the height at which it intercepts the principal surface.

$$\mathrm{f/\#} = \frac{f}{\mathrm{CA}}. \qquad \textit{(see eq. 1.7)}$$

NUMERICAL APERTURE (NA)

The NA of a lens system is defined to be the sine of the angle, θ_1, that the marginal ray (the ray that exits the lens system at its outer edge) makes with the optical axis multiplied by the index of refraction (n) of the medium. The NA can be defined for any ray as the sine of the angle made by that ray with the optical axis multiplied by the index of refraction:

$$\mathrm{NA} = n\sin\theta. \qquad (1.30)$$

MAGNIFICATION POWER

Often, positive lenses intended for use as simple magnifiers are rated with a single magnification, such as 4 × . To create a virtual image for viewing with the human eye, in principle, any positive lens can be used at an infinite number of possible magnifications. However, there is usually a narrow range of magnifications that will be comfortable for the viewer. Typically, when the viewer adjusts the object distance so that the image appears to be essentially at infinity (which is a comfortable viewing distance for most individuals), magnification is given by the relationship

$$\text{magnification} = \frac{254\text{ mm}}{f} \qquad (f \text{ in mm}). \tag{1.31}$$

Thus, a 25.4-mm focal length positive lens would be a 10 × magnifier.

DIOPTERS

The term diopter is used to define the reciprocal of the focal length, which is commonly used for ophthalmic lenses. The inverse focal length of a lens expressed in diopters is

$$\text{diopters} = \frac{1000}{f} \qquad (f \text{ in mm}). \tag{1.32}$$

Thus, the smaller the focal length is, the larger the power in diopters will be.

DEPTH OF FIELD AND DEPTH OF FOCUS

In an imaging system, depth of field refers to the distance in object space over which the system delivers an acceptably sharp image. The criteria for what is acceptably sharp is arbitrarily chosen by the user; depth of field increases with increasing f-number.

For an imaging system, depth of focus is the range in image space over which the system delivers an acceptably sharp image. In other words, this is the amount that the image surface (such as a screen or piece of photographic film) could be moved while maintaining acceptable focus. Again, criteria for acceptability are defined arbitrarily.

In nonimaging applications, such as laser focusing, depth of focus refers to the range in image space over which the focused spot diameter remains below an arbitrary limit.

APPLICATION NOTE

Technical Reference

For further reading about the definitions and formulas presented here, refer to the following publications:

Rudolph Kingslake, *Lens Design Fundamentals*
(Academic Press)

Rudolph Kingslake, System Design
(Academic Press)

Warren Smith, *Modern Optical Engineering*
(McGraw Hill).

Donald C. O'Shea, *Elements of Modern Optical Design*
(John Wiley & Sons)

Eugene Hecht, *Optics* (Addison Wesley)

Max Born, Emil Wolf, *Principles of Optics*
(Cambridge University Press)

If you need help with the use of definitions and formulas presented in this guide, our applications engineers will be pleased to assist you.

Paraxial Lens Formulas

PARAXIAL FORMULAS FOR LENSES IN AIR

The following formulas are based on the behavior of paraxial rays, which are always very close and nearly parallel to the optical axis. In this region, lens surfaces are always very nearly normal to the optical axis, and hence all angles of incidence and refraction are small. As a result, the sines of the angles of incidence and refraction are small (as used in Snell's law) and can be approximated by the angles themselves (measured in radians).

The paraxial formulas do not include effects of spherical aberration experienced by a marginal ray — a ray passing through the lens near its edge or margin. All EFL values (f) tabulated in this catalog are paraxial values which correspond to the paraxial formulas. The following paraxial formulas are valid for both thick and thin lenses unless otherwise noted. The refractive index of the lens glass, n, is the ratio of the speed of light in vacuum to the speed of light in the lens glass. All other variables are defined in figure 1.33.

Focal Length

$$\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right)+\frac{(n-1)^2}{n}\,\frac{t_c}{r_1 r_2} \qquad (1.33)$$

where n is the refractive index, t_c is the center thickness, and the sign convention previously given for the radii r_1 and r_2 applies. For thin lenses, $t_c \cong 0$, and for plano lenses either r_1 or r_2 is infinite. In either case the second term of the above equation vanishes, and we are left with the familiar lens maker's formula:

$$\frac{1}{f}=(n-1)\left(\frac{1}{r_1}-\frac{1}{r_2}\right). \qquad (1.34)$$

Surface Sagitta and Radius of Curvature

(refer to figure 1.34)

$$r^2=(r-s)^2+\left(\frac{d}{2}\right)^2 \qquad (1.35)$$

$$s=r-\sqrt{r^2-\left(\frac{d}{2}\right)^2}>0 \qquad (1.36)$$

$$r=\frac{s}{2}+\frac{d^2}{8s}. \qquad (1.37)$$

An often useful approximation is to neglect $s/2$.

Symmetric Lens Radii ($r_2 = -r_1$)

With center thickness constrained,

$$r_1=(n-1)\left[f\pm\sqrt{f^2-\left(\frac{f t_c}{n}\right)}\right] \qquad (1.38)$$

$$=(n-1)\,f\left[1+\sqrt{1-\left(\frac{t_c}{nf}\right)}\right]$$

where, in the first form, the + sign is chosen for the square root if f is positive, but the − sign must be used if f is negative. In the second form, the + sign must be used regardless of the sign of f.

Plano Lens Radius

Since r_2 is infinite,

$$r_1=(n-1)\,f. \qquad (1.39)$$

Principal-Point Locations (signed distances from vertices)

$$\mathrm{A_2H''}=\frac{-r_2 t_c}{n(r_2-r_1)+t_c(n-1)} \qquad (1.40)$$

$$\mathrm{A_1H}=\frac{-r_1 t_c}{n(r_2-r_1)+t_c(n-1)} \qquad (1.41)$$

where the above sign convention applies.

For symmetric lenses ($r_2 = -r_1$),

$$\mathrm{A_1H}=-\mathrm{A_2H''}$$

$$=\frac{r_1 t_c}{2nr_1-t_c(n-1)}. \qquad (1.42)$$

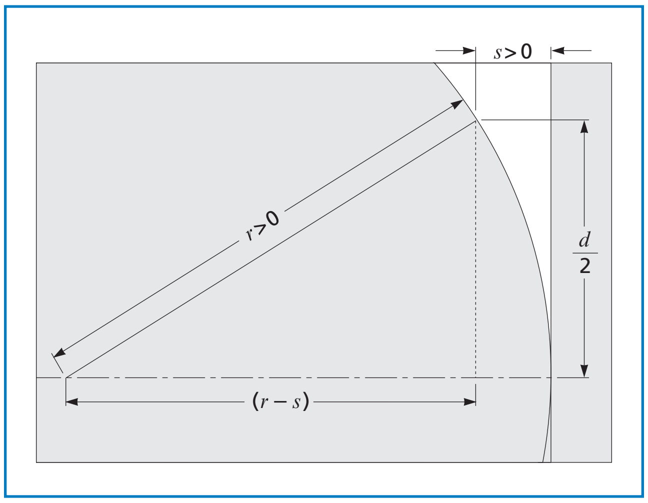

Figure 1.34 **Surface sagitta and radius of curvature**

If either r_1 or r_2 is infinite, l'Hôpital's rule from calculus must be used. Thus, referring to *Aberration Balancing*, for plano-convex lenses in the correct orientation,

$$\mathrm{A_1H} = 0$$

and

$$\mathrm{A_2H''} = -\frac{t_c}{n}. \qquad (1.43)$$

For flat plates, by letting $r_1 \rightarrow \infty$ in a symmetric lens, we obtain $\mathrm{A_1H} = \mathrm{A_2H''} = t_c/2_n$. These results are useful in connection with the following paraxial lens combination formulas.

Hiatus or Interstitium (principal-point separation)

$$\mathrm{HH''} = t_c\left\{1 - \frac{f}{n}\left[\frac{1}{f} - \frac{(n-1)^2}{n}\,\frac{t_c}{r_1 r_2}\right]\right\} \qquad (1.44)$$

which, in the thin-lens approximation (exact for plano lenses), becomes

$$\mathrm{HH''} = t_c\left(1 - \frac{1}{n}\right). \qquad (1.45)$$

Solid Angle

The solid angle subtended by a lens, for an observer situated at an on-axis image point, is

$$\begin{aligned}\Omega &= 2\pi(1-\cos\theta)\\ &= 4\pi\sin^2\left(\frac{\theta}{2}\right)\end{aligned} \qquad (1.46)$$

where this result is in steradians, and where

$$\theta = \arctan\left(\frac{\mathrm{CA}}{2s}\right) \qquad (1.47)$$

is the apparent angular radius of the lens clear aperture. For an observer at an on-axis object point, use s instead of s''. To convert from steradians to the more intuitive sphere units, simply divide Ω by 4π. If the Abbé sine condition is known to apply, θ may be calculated using the arc sine function instead of the arctangent.

Back Focal Length

$$\begin{aligned}f_\mathrm{b} &= f'' + \mathrm{A_2H''}\\ &= f'' - \frac{r_2 t_\mathrm{c}}{n(r_2 - r_1) + t_\mathrm{c}(n-1)}\end{aligned} \qquad (1.48)$$

where the sign convention presented above applies to $\mathrm{A_2H''}$ and to the radii. If r_2 is infinite, l'Hôpital's rule from calculus must be used, whereby

$$f_\mathrm{b} = f'' - \frac{t_\mathrm{c}}{n}. \qquad (1.49)$$

Front Focal Length

$$\begin{aligned}f_\mathrm{f} &= f - \mathrm{A_1H}\\ &= f + \frac{r_1 t_c}{n(r_2 - r_1) + t_\mathrm{c}(n-1)}\end{aligned} \qquad (1.50)$$

where the sign convention presented above applies to $\mathrm{A_1H}$ and to the radii. If r_1 is infinite, l'Hôpital's rule from calculus must be used, whereby

$$f_\mathrm{f} = f - \frac{t_\mathrm{c}}{n}. \qquad (1.51)$$

Edge-to-Focus Distances

For positive lenses,

$$\mathrm{A} = f_\mathrm{f} + s_1 \qquad (1.52)$$

and

$$\mathrm{B} = f_\mathrm{b} + s_2 \qquad (1.53)$$

where s_1 and s_2 are the sagittas of the first and second surfaces. Bevel is neglected.

Magnification or Conjugate Ratio

$$\begin{aligned}m &= \frac{s''}{s}\\ &= \frac{f}{s-f}\\ &= \frac{s''-f}{f}.\end{aligned} \qquad (1.54)$$

PARAXIAL FORMULAS FOR LENSES IN ARBITRARY MEDIA

These formulas allow for the possibility of distinct and completely arbitrary refractive indices for the object space medium (refractive index n'), lens (refractive index n''), and image space medium (refractive index n). In this situation, the EFL assumes two distinct values, namely f in object space and f'' in image space. It is also necessary to distinguish the principal points from the nodal points. The lens serves both as a lens and as a window separating the object space and image space media.

The situation of a lens immersed in a homogenous fluid (figure 1.35) is included as a special case ($n = n''$). This case is of considerable practical importance. The two values f and f'' are again equal, so that the lens combination formulas are applicable to systems immersed in a common fluid. The general case (two different fluids) is more difficult, and it must be approached by ray tracing on a surface-by-surface basis.

Lens Constant (k)

This number appears frequently in the following formulas. It is an explicit function of the complete lens prescription (both radii, t_c and n') and both media indices (n and n''). This dependence is implicit anywhere that k appears.

$$k=\frac{n'-n}{r_1}+\frac{n''-n'}{r_2}-\frac{t_c(n'-n)(n''-n')}{n'r_1r_2}. \tag{1.55}$$

Effective Focal Lengths

$$f=\frac{n}{k} \qquad f''=\frac{n''}{k}. \tag{1.56}$$

Lens Formula (Gaussian form)

$$\frac{n}{s}+\frac{n''}{s''} = k. \tag{1.57}$$

Lens Formula (Newtonian form)

$$xx''=f''f=\frac{nn''}{k^2}. \tag{1.58}$$

Principal-Point Locations

$$\mathrm{A_1H}=\frac{nt_c}{k}\left(\frac{n''-n'}{n'r_2}\right) \tag{1.59}$$

$$\mathrm{A_2H''}=\frac{-n''t_c}{k}\left(\frac{n'-n}{n'r_1}\right). \tag{1.60}$$

Object-to-First-Principal-Point Distance

$$s=\frac{ns''}{ks''-n''}. \tag{1.61}$$

Second Principal-Point-to-Image Distance

$$s''=\frac{n''s}{ks-n}. \tag{1.62}$$

Magnification

$$m=\frac{ns''}{n''s}. \tag{1.63}$$

Lens Maker's Formula

$$\frac{n}{f}=\frac{n''}{f''}=k. \tag{1.64}$$

Nodal-Point Locations

$$\mathrm{A_1N = A_1H + HN} \tag{1.65}$$

$$\mathrm{A_2N'' = A_2H'' + H''N''}. \tag{1.66}$$

Separation of Nodal Point from Corresponding Principal Point

$\mathrm{HN} = \mathrm{H''N''} = (n''-n)/k$, positive for N to right of H and N″ to right of H″.

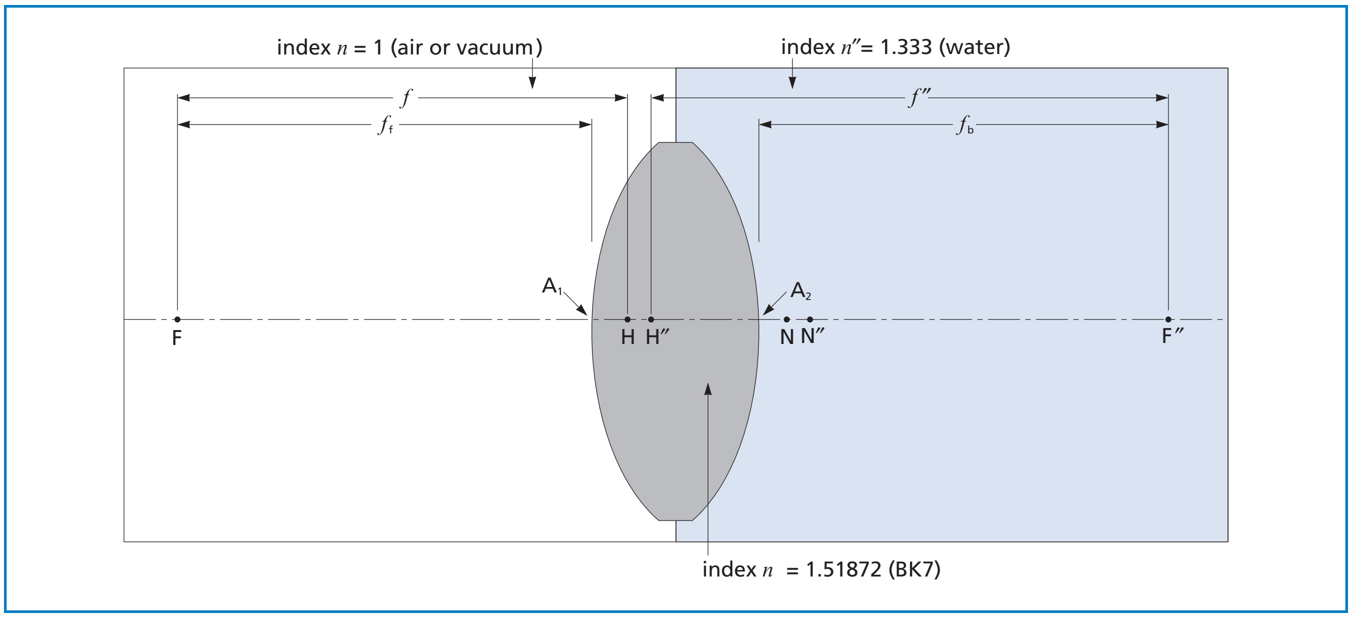
index n = 1 (air or vacuum)
index n''= 1.333 (water)
f
f_f
f''
f_b
A$_1$
A$_2$
F
H H″
N N″
F″
index n = 1.51872 (BK7)

Figure 1.35 **Symmetric lens with disparate object and image space indexes**

Back Focal Length

$$f_b = f'' + A_2H''. \qquad \textit{(see eq. 1.48)}$$

Front Focal Length

$$f_f = f - A_1H. \qquad \textit{(see eq. 1.50)}$$

Focal Ratios

The focal ratios are f/CA and f''/CA, where CA is the diameter of the clear aperture of the lens.

Numerical Apertures

$$n \sin\theta$$

$$\text{where } \theta = \arcsin\left(\frac{\text{CA}}{2s}\right)$$

and

$$n'' \sin\theta''$$

$$\text{where } \theta'' = \arcsin\left(\frac{\text{CA}}{2s''}\right).$$

Solid Angles (in steradians)

$$\Omega = 2\pi(1-\cos\theta) \qquad \textit{(see eq. 1.46)}$$

$$= 4\pi \sin^2\left(\frac{\theta}{2}\right)$$

$$\text{where } \theta = \arctan\left(\frac{\text{CA}}{2s}\right)$$

$$\Omega = 2\pi(1 - \cos\theta'')$$

$$= 4\pi \sin^2\left(\frac{\theta}{2}\right)$$

$$\text{where } \theta'' = \arctan\left(\frac{\text{CA}}{2s''}\right).$$

To convert from steradians to spheres, simply divide by 4π.

APPLICATION NOTE

For Quick Approximations

Much time and effort can be saved by ignoring the differences among f, f_b, and f_f in these formulas by assuming that $f = f_b = f_f$. Then s becomes the lens-to-object distance; s'' becomes the lens-to-image distance; and the sum of conjugate distances $s + s''$ becomes the object-to-image distance. This is known as the thin-lens approximation.

APPLICATION NOTE

Physical Significance of the Nodal Points

A ray directed at the primary nodal point N of a lens appears to emerge from the secondary nodal point N″ without change of direction. Conversely, a ray directed at N″ appears to emerge from N without change of direction. At the infinite conjugate ratio, if a lens is rotated about a rotational axis orthogonal to the optical axis at the secondary nodal point (i.e., if N″ is the center of rotation), the image remains stationary during the rotation. This fact is the basis for the nodal slide method for measuring nodal-point location and the EFL of a lens. The nodal points coincide with their corresponding principal points when the image space and object space refractive indices are equal ($n = n''$). This makes the nodal slide method the most precise method of principal-point location.

Principal-Point Locations

Figure 1.36 indicates approximately where the principal points fall in relation to the lens surfaces for various standard lens shapes. The exact positions depend on the index of refraction of the lens material, and on the lens radii, and can be found by formula. in extreme meniscus lens shapes (short radii or steep curves), it is possible that both principal points will fall outside the lens boundaries. For symmetric lenses, the principal points divide that part of the optical axis between the vertices into three approximately equal segments. For plano lenses, one principal point is at the curved vertex, and the other is approximately one-third of the way to the plane vertex.

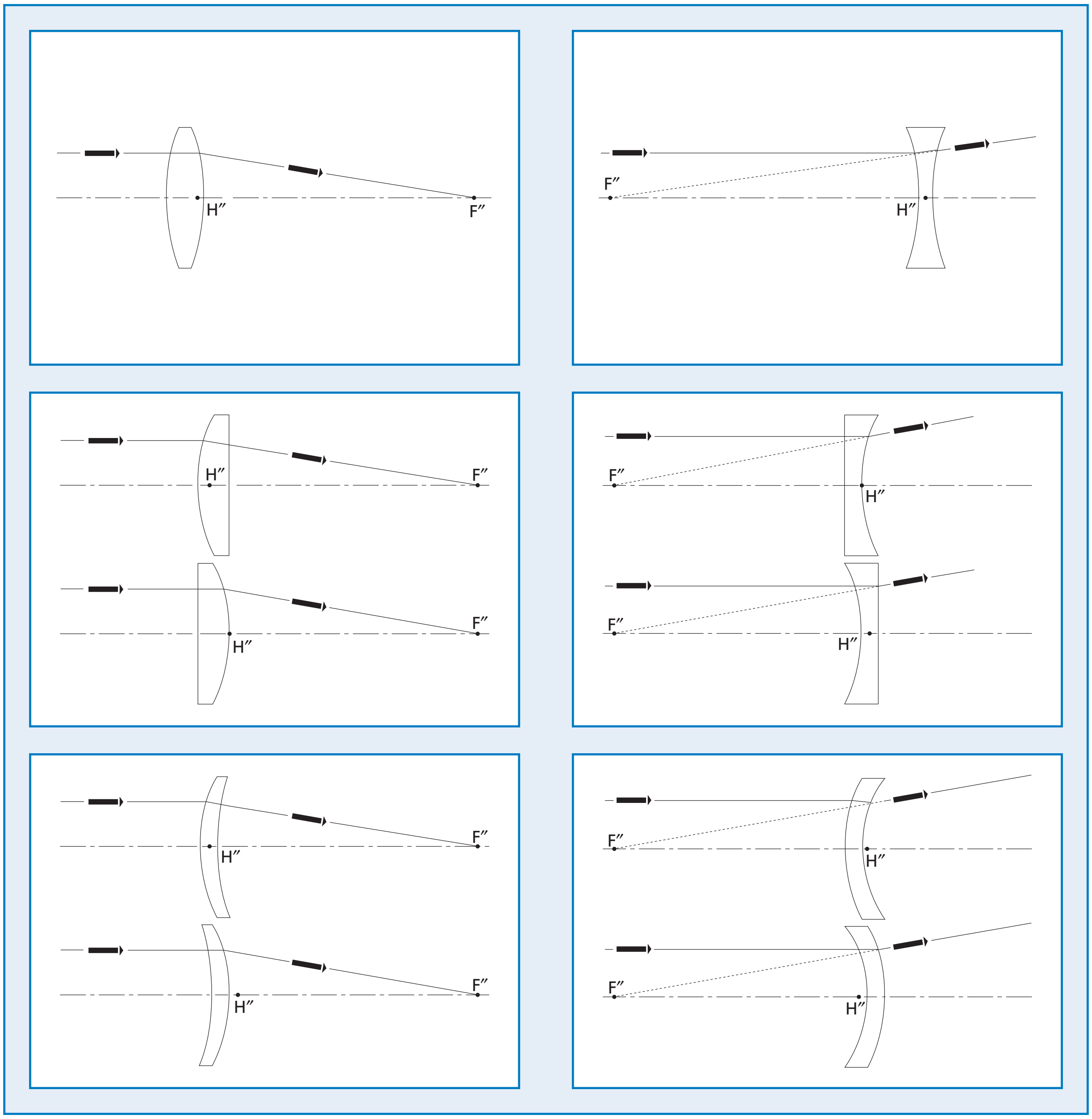

Figure 1.36 **Principal points of common lenses**

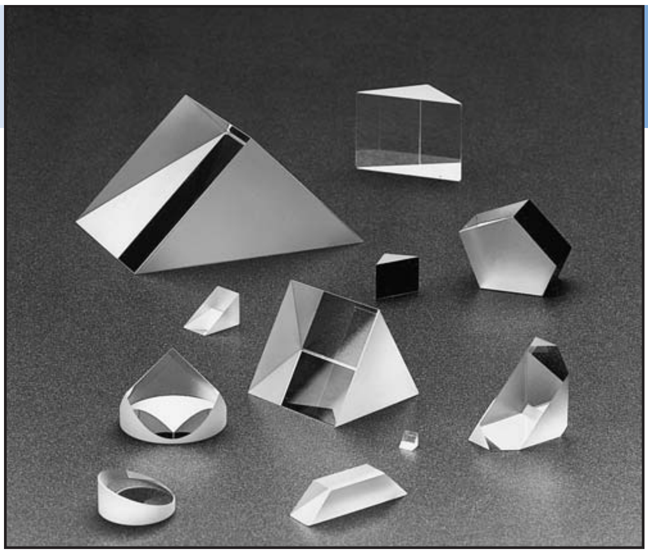

Prisms

Prisms are blocks of optical material whose flat, polished sides are arranged at precisely controlled angles to each other. Prisms may be used in an optical system to deflect or deviate a beam of light. They can invert or rotate an image, disperse light into its component wavelengths, and be used to separate states of polarization.

PRISM ORIENTATION

The orientation of a prism determines its effect on a beam of light or an image.

A viewer looks through a prism at an object and sees a virtual image (see Figure 1.37). This image may be displaced from the original object, or, if a dove prism is used, it may coincide with the object. Furthermore, image orientation may differ from the object; in the case of a right-angle prism, the image is reversed.

A real image (see Figure 1.38) can be formed only if imaging optics are present in the system. Without imaging optics, the image is virtual. A virtual image has the same orientation as the real image shown, but it can be viewed by the observer only by looking back through the prism system.

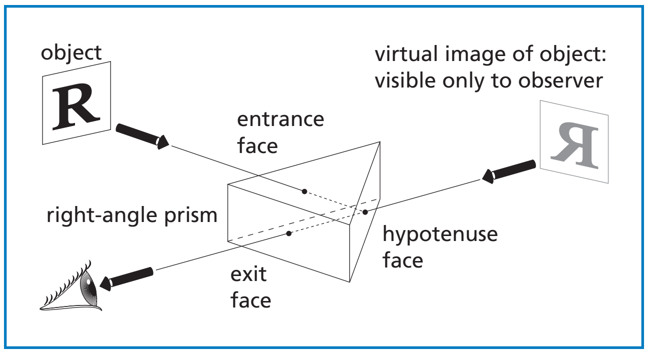

Figure 1.37 **Virtual imaging using a prism**

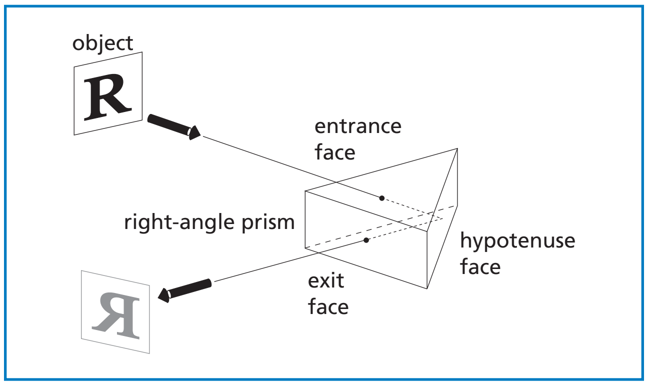

Figure 1.38 **Real imaging using a prism**

TOTAL INTERNAL REFLECTION

Rays incident upon a glass/air boundary (i.e., an internal reflection) at angles that exceed the critical angle are reflected with 100-percent efficiency regardless of their initial polarization state. The critical angle is given by

$$\theta_c(\lambda) = \arcsin \frac{1}{n_\lambda} \qquad (1.67)$$

and depends on the refractive index, which is a function of wavelength. If, at some wavelength, the refractive index should fall to less than $\sqrt{2} = 1.414$, the critical angle will exceed 45 degrees, and total internal reflection (TIR) will fail for a collimated beam internally incident at 45 degrees on the hypotenuse face of a right-angle prism. Reflectance decreases rapidly at angles of incidence smaller than the critical angle.

The index of BK7 is sufficiently high to guarantee the TIR of a collimated beam at 45 degrees internal incidence over the visible and near-infrared region. The possibility of significant TIR failure with convergent or divergent beams should be kept in mind if polarization is important. TIR can also fail if the hypotenuse face is not kept extremely clean. Even an almost invisible fingerprint can lead to TIR failure. An aluminum- or silver-coated hypotenuse is recommended for applications where the right-angle prism is frequently handled, or where convergent or divergent beams are used. There is a slight loss of reflectance at all internal angles with the coating, and no critical angle exists.

ABERRATIONS FOR PRISMS

Prisms will introduce aberrations when they are used with convergent or divergent beams of light. Using prisms with collimated or nearly collimated light will help minimize aberrations. Conjugate distances that include prisms should be long.

DISPERSING PRISMS

Dispersing prisms are used to separate a beam of white light into its component colors. Generally, the light is first collimated and then dispersed by the prism. A spectrum is then formed at the focal plane of a lens or curved mirror. In laser work, dispersing prisms are used to separate two wavelengths following the same beam path. Typically, the dispersed beams are permitted to travel far enough so the beams separate spatially.

A prism exhibits magnification in the plane of dispersion if the entrance and exit angles for a beam differ. This is useful in anamorphic (one-dimensional) beam expansion or compression, and may be used to correct or create asymmetric beam profiles.

As shown in Figure 1.39, a beam of width W_1 is incident at an angle α on the surface of a dispersing prism of apex angle A. The angle of refraction at the first surface, β, the angle of incidence at the second surface, γ, and the angle of refraction exiting the prism, δ, are easily calculated:

$$\beta = \sin^{-1}((\sin\alpha)\,/\,\eta)$$
$$\gamma = A - \beta$$
$$\delta = \sin^{-1}(\eta\sin\gamma)$$

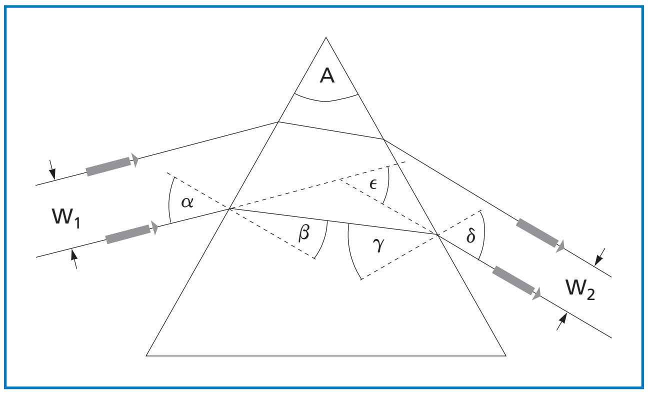

Figure 1.39 **Diagram of dispersing prism**

The beam deviation, ϵ, is of greatest importance. It is the angle the exit beam makes with its original direction.

$$\epsilon = \alpha + \delta - A \qquad (1.68)$$

The magnification W_2/W_1 is given by:

$$M = \frac{\cos\delta\cos\beta}{\cos\alpha\cos\gamma} \qquad (1.69)$$

The resolving power of a prism spectrometer angle α, the angular dispersion of the prism is given by:

$$\frac{d\delta}{d\lambda} = \left(\frac{\sin A}{\cos\delta\cos\beta}\right)\left(\frac{d\eta}{d\lambda}\right) \qquad (1.70)$$

If the spectrum is formed by a diffraction limited focal system of focal length f, the minimum spot size is dx ~ $f\lambda/W$. This corresponds to a minimum angular resolution $d\delta \sim \lambda/w$ for a beam of diameter w. The diffraction limited angular resolution at a given beam diameter sets the limit on the spectral resolving power of a prism. Setting the expression for $d\delta$ equal to the minimum angular resolution, we obtain:

$$RP = \frac{\lambda}{d\lambda} = \left(\frac{w_2\sin A}{\cos\delta\cos\beta}\right)\left(\frac{d\eta}{d\lambda}\right) \qquad (1.71)$$

where RP is the resolving power of the prism.

At a given wavelength, the beam deviation ϵ is a minimum at an angle of incidence:

$$\alpha_{\text{min dev}} = \sin^{-1}[\eta\sin(A/2)] \qquad (1.72)$$

where η is the prism index of refraction at that wavelength. At this angle, the incident and exit angles are equal, the prism magnification is one, and the internal rays are perpendicular to the bisector of the apex angle.

By measuring the angle of incidence for minimum deviation, the index of refraction of a prism can be determined. Also, by proper choice of apex angle, the equal incident and exit angles may be made Brewster's angle, eliminating losses for p-polarized beams. The apex angle to choose is:

$$A = \pi - 2\theta_B \qquad (1.73)$$

If, in addition, the base angles of the prism are chosen as Brewster's angle, an isosceles Brewster prism results.

Another use is illustrated next.

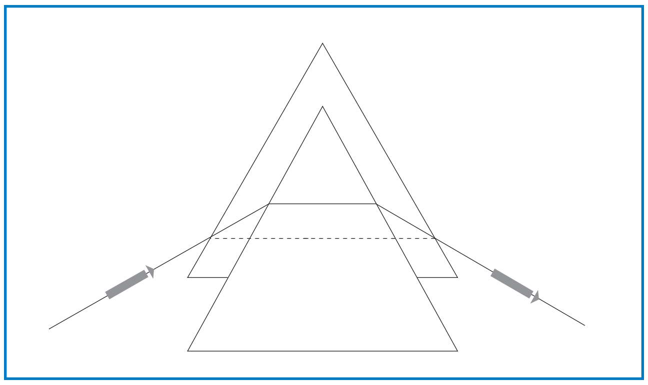

Figure 1.40 **Translation of a prism at minimum deviation**

At minimum deviation, translating a prism along the bisector of the apex angle does not disturb the direction of the output rays. See Figure 1.40. This is important in femtosecond laser design where intracavity prisms are used to compensate for group velocity dispersion. By aligning a prism for minimum deviation and translating it along its apex bisector, the optical path length in material may be varied with no misalignment, thus varying the contribution of the material to overall group velocity dispersion. Finally, it is possible to show that at minimum deviation

$$\mathrm{RP} = \left(b_2 - b_1\right)\frac{d\eta}{d\lambda} \qquad (1.74)$$

where the relevant quantities are defined in Figure 1.41.

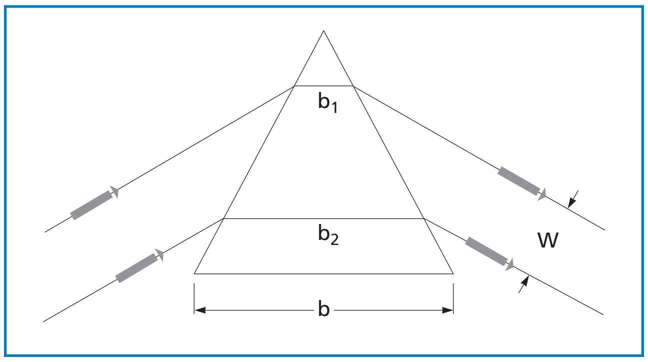

Figure 1.41 **Ray path lengths of a prism at minimum deviation**

If the beam is made to fill the prism completely, $b_1 = 0$, and $b_2 = b$, the base of the prism. So, we have the classical result that the resolving power of a prism spectrometer is equal to the base of the prism times the dispersion of the prism material.

As an example, consider CVI Melles Griot EDP-25-F2 prism, operating in minimum deviation at 590 nm. The angle of incidence and emergence are both then 54.09° and $d\eta/d\lambda$ is $-0.0854\ \mu\mathrm{m}^{-1}$ for F2 glass at 590 nm. If the 25-mm prism is completely filled, the resolving power, $\lambda/d\lambda$, is 2135. This is sufficient to resolve the Sodium D lines.

PELLIN BROCA PRISMS

In a Pellin Broca prism, an ordinary dispersing prism is split in half along the bisector of the apex angle. Using a right angle prism, the two halves are joined to create a dispersing prism with an internal right angle bend obtained by total internal reflection, as shown in Figure 1.42.

In principle, one can split any type of dispersing prism to create a Pellin Broca prism. Typically the Pellin Broca prism is based on an Isosceles Brewster prism. Provided the light is *p*-polarized, the prism will be essentially lossless.

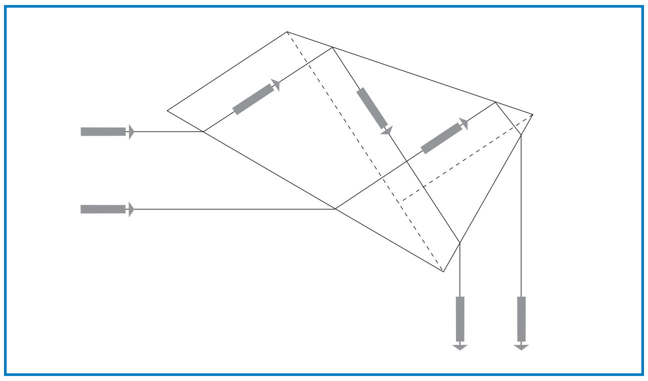

Figure 1.42 **One of the wavelengths deviates at exactly 90° to its initial direction**

Suppose wavelengths λ_1 and λ_2 are superimposed in a collimated beam, as at the output of a harmonic generating crystal, the diagram in Figure 1.42 suggests that it is always possible to find a rotation of the prism in its plane that ensures that one of the two wavelengths will operate at minimum deviation when refracting at the input face of the first of the half-dispersing prisms. This means that it will enter the right angle prism normal to one of its faces, be turned exactly 90°, be presented to the second half-dispersing prism in minimum deviation, and hence exit the Pellin Broca prism deviated at exactly 90° to its initial direction.

A simple dispersing prism always deviates the longer wavelength less than the shorter wavelength. In a Pellin Broca prism, whether the longer wavelength is deviated more or less depends on the orientation of the prism. This is an important consideration when designing a high power Pellin Broca beam separator, as shown in Figures 1.43 and 1.44.

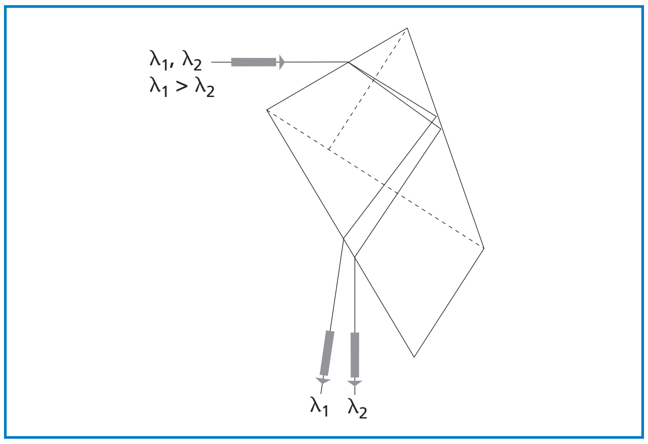

Figure 1.43 **Longer wavelength is deviated more than the shorter wavelength**

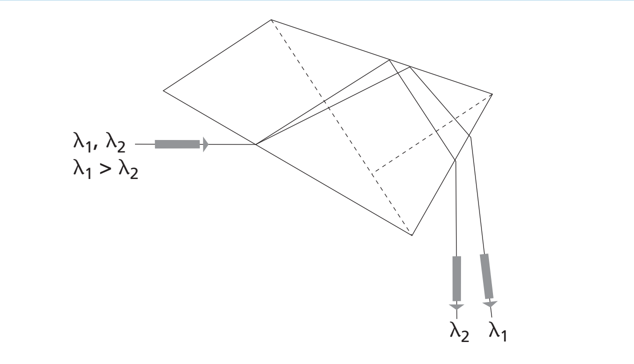

Figure 1.44 **Longer wavelength is deviated less than the shorter wavelength**

CVI Melles Griot offers Brewster angle Pellin Brocca prisms in a number of sizes and materials. BK7 prisms are used in the visible and near IR, and is the least expensive. UV-grade fused silica Pellin Broca prisms are used from 240 nm to 2000 nm. Excimer-grade prisms are used in the 180-nm to 240-nm region. Crystal-quartz Pellin Broca prisms are specifically designed for high-power Q-switched 266-nm laser pulses at fluence levels of 50 mJ/cm^2. Fused silica prisms track (i.e., suffer internal catastrophic damage) above this fluence, probably due to self-focusing.

PORRO PRISMS

A Porro prism, named for its inventor Ignazio Porro, is a type of reflection prism used to alter the orientation of an image. In operation, light enters the large face of the prism, undergoes total internal reflection twice from the 45° sloped faces, and exits again through the large face. An image traveling through a Porro prism is rotated by 180° and exits in the opposite direction offset from its entrance point, as shown in Figure 1.45. Since the image is reflected twice, the handedness of the image is unchanged. Porro prisms have rounded edges to minimize breakage and facilitate assembly.

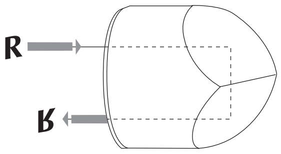

Figure 1.45 **Porro prisms retroreflect and invert the image**

Porro prisms are most often used in pairs, forming a double Porro prism, as shown in Figure 1.46. A second prism, rotated 90° with respect to the first, is placed such that the beam will traverse both prisms. The net effect of the prism system is a beam parallel to but displaced from its original direction, with the image rotated 180°. As before, the handedness of the image is unchanged.

Double Porro prism systems are used in small optical telescopes to reorient an inverted image and in many binoculars to both re-orient the image and provide a longer, folded distance between the objective lenses and the eyepieces.

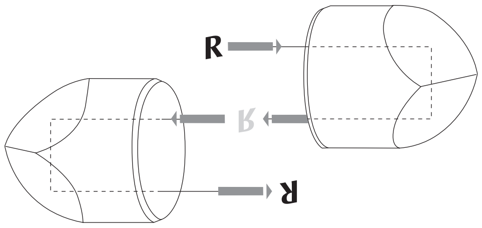

Figure 1.46 **Double porro prism results in a beam parallel to but displaced from its original direction, with the image rotated 180°**

Polarization

Polarization States

Four numbers are required to describe a single plane wave Fourier component traveling in the $+z$ direction. These can be thought of as the amplitude and phase shift of the field along two orthogonal directions.

1. CARTESIAN REPRESENTATION

In Cartesian coordinates, the propagation equation for an electric field is given by the formula

$$E(x,y,z)=(xE_x e^{i\phi_x}+yE_y e^{i\phi_y})e^{i(kz-\omega t)} \qquad (1.75)$$

where E_x, E_y, ϕ_x, and ϕ_y are real numbers defining the magnitude and the phase of the field components in the orthogonal unit vectors x and y. If the origin of time is irrelevant, only the relative phase shift

$$\phi=\phi_x-\phi_y \qquad (1.76)$$

need be specified.

2. CIRCULAR REPRESENTATION

In the circular representation, we resolve the field into circularly polarized components. The basic states are represented by the complex unit vectors

$$e_+=\left(1/\sqrt{2}\right)\left(x+iy\right) \text{ and} \qquad (1.77)$$

$$e_-=\left(1/\sqrt{2}\right)\left(x-iy\right) \qquad (1.78)$$

where e_+ is the unit vector for left circularly polarized light; for positive helicity light; for light that rotates counterclockwise in a fixed plane as viewed facing into the light wave; and for light whose electric field rotation obeys the right hand rule with thumb pointing in the direction of propagation.

e_- is the unit vector for right circularly polarized light; for negative helicity light; for light that rotates clockwise in a fixed plane as viewed facing into the light wave; and for light whose electric field rotation disobeys the right hand rule with thumb pointing in the direction of propagation.

APPLICATION NOTE

Polarization Convention

Historically, the orientation of a polarized electromagnetic wave has been defined in the optical regime by the orientation of the electric vector. This is the convention used by CVI Melles Griot.

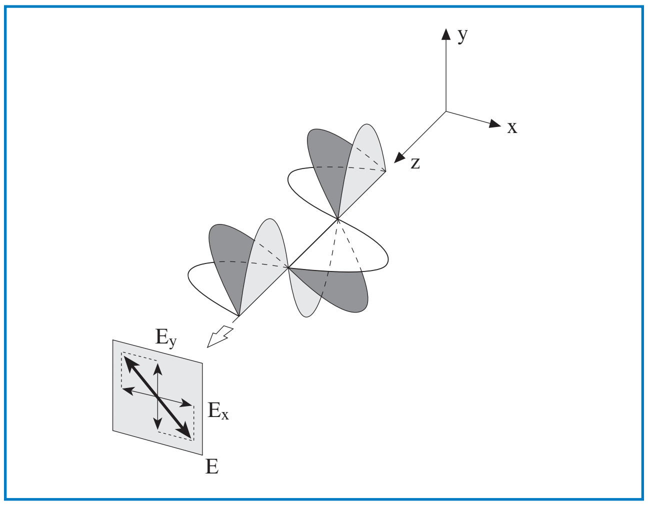

Figure 1.47 **Linearly polarized light. E_x and E_y are in phase**

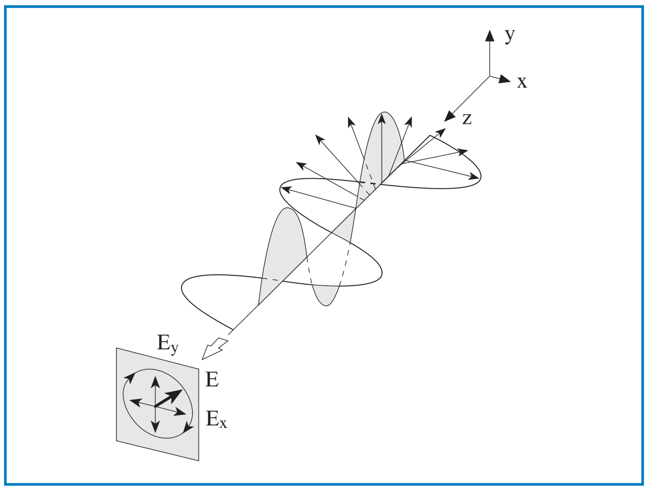

Figure 1.48 **Circularly polarized light. E_x and E_y are out of phase by angular frequency ω**

As in the case of the Cartesian representation, we write

$$E=(e_{+}E_{+}\mathrm{e}^{i\phi+}+e_{-}E_{-}\mathrm{e}^{i\phi-})\,e^{i(kz-\omega t)} \qquad (1.79)$$

where E_{+}, E_{-}, ϕ_{+}, and ϕ_{-} are four real numbers describing the magnitudes and phases of the field components of the left and right circularly polarized components. Note that

$$E_{+}=e_{-}\bullet E \qquad (1.80)$$

$$E_{-}=e_{+}\bullet E \qquad (1.81)$$

3. ELLIPTICAL REPRESENTATION

An arbitrary polarization state is generally elliptically polarized. This means that the tip of the electric field vector will describe an ellipse, rotating once per optical cycle.

Let a be the semimajor and b be the semiminor axis of the polarization ellipse. Let ψ be the angle that the semimajor axis makes with the x axis.

Let ξ and η be the axes of a right-handed coordinate system rotated by an angle $+\psi$ with respect to the x axis and aligned with the polarization ellipse as shown in Figure 1.49.

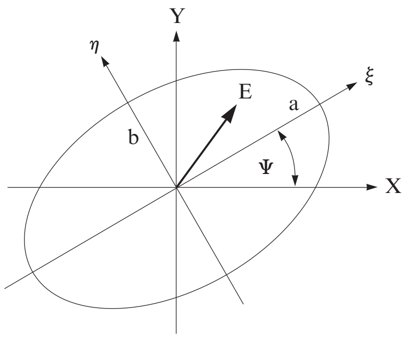

Figure 1.49 **The polarization ellipse**

The elliptical representation is:

$$E=\left(a\hat{\xi}+b\hat{\eta}\right)e^{i\delta_0}e^{i(kz-\omega t)} \qquad (1.82)$$

Note that the phase shift δ_o above is required to adjust the time origin, and the parameter ψ is implicit in the rotation of the ξ and η axes with respect to the x and y axes.

CONVERSIONS BETWEEN REPRESENTATIONS

For brevity, we will provide only the Cartesian to circular and Cartesian to elliptical transformations. The inverse transformations are straightforward. We define the following quantities:

$$g_1=E_x\cos\phi_x-E_y\sin\phi_y \qquad (1.83)$$

$$g_2=E_x\sin\phi_x+E_y\cos\phi_y \qquad (1.84)$$

$$g_3=E_x\cos\phi_x+E_y\sin\phi_y \qquad (1.85)$$

$$g_4=E_x\sin\phi_x-E_y\cos\phi_y \qquad (1.86)$$

$$u=\left(g_1^2+g_2^2\right)^{1/2} \qquad (1.87)$$

$$v=\left(g_3^2+g_4^2\right)^{1/2} \qquad (1.88)$$

$$\phi_{12}=\mathrm{atan}\,(g_1,g_2) \qquad (1.89)$$

$$\phi_{34}=\mathrm{atan}\,(g_3,g_4) \qquad (1.90)$$

In the above, atan(x,y) is the four quadrant arc tangent function. This means that atan(x,y) = atan(y/x) with the provision that the quadrant of the angle returned by the function is controlled by the signs of both x and y, not just the sign of their quotient; for example, if $g_2 = g_1 = -1$, then ϕ_{12} above is $5\pi/4$ or $-3\pi/4$, not $\pi/4$.

A. CARTESIAN TO CIRCULAR TRANSFORMATION

$$E_{+}=v/\sqrt{2} \qquad (1.91)$$

$$E_{-}=u/\sqrt{2} \qquad (1.92)$$

$$\phi_{+}=\phi_{34} \qquad (1.93)$$

$$\phi_{-}=\phi_{12} \qquad (1.94)$$

B. CARTESIAN TO ELLIPTICAL TRANSFORMATION

$$a=(v+u)/2 \qquad (1.95)$$

$$b=(v-u)/2 \qquad (1.96)$$

$$\psi=(\phi_{12}-\phi_{34})/2 \qquad (1.97)$$

$$\delta_0=(\phi_{12}+\phi_{34})/2 \qquad (1.98)$$

LINEAR POLARIZERS

A linear polarizer is a device that creates a linear polarization state from an arbitrary input. It does this by removing the component orthogonal to the selected state. Unlike plastic sheet polarizers which absorb the rejected beam (which turns into heat), cube polarizers and thin-film plate polarizers reflect the rejected beam, creating two usable beams. Still others may refract the two polarized beams at different angles, thereby separating them. Examples are Wollaston and Rochon prism polarizers.

Suppose the pass direction of the polarizer is determined by unit vector p. Then the transmitted field E_2, in terms of the incident field E_1, is given by

$$E_2 = p(p \bullet E_1) \tag{1.99}$$

where the phase shift of the transmitted field has been ignored.

A real polarizer has a pass transmission, $T_\parallel$, less than 1. The transmission of the rejected beam, $T_\perp$, may not be 0. If r is a unit vector along the rejected direction, then

$$E_2 = (T_\parallel)^{1/2} p(p \bullet E_1)e^{i\phi_\parallel} + (T_\perp)^{1/2} r(r \bullet E_1)e^{i\phi_\perp} \tag{1.100}$$

In the above, the phase shifts along the two directions must be retained. Similar expressions could be arrived at for the rejected beam. If θ is the angle between the field E_1 and the polarizer pass direction p, the above equation predicts that

$$T = T_\parallel \cos^2\theta + T_\perp \sin^2\theta \tag{1.101}$$

The above equation shows that, when the polarizer is aligned so that $\theta = 0$, $T = T_\parallel$. When it is "crossed", $\theta = \pi/2$, and $T = T_\perp$. The extinction ratio is $\epsilon = T_\parallel / T_\perp$. A polarizer with perfect extinction has $T_\perp = 0$, and thus $T = T_\parallel \cos^2\theta$ is a familiar result. Because $\cos^2\theta$ has a broad maximum as a function of orientation angle, setting a polarizer at a maximum of transmission is generally not very accurate. One has to either map the $\cos^2\theta$ with sufficient accuracy to find the $\theta = 0$ point, or do a null measurement at $\theta = \pm\pi/2$.

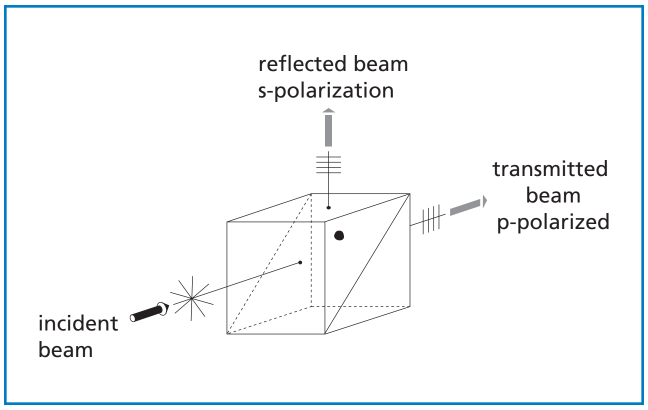

Figure 1.50 **At CVI Melles Griot, the DOT marks preferred input face. This is the tested direction for transmitted wavefront. Damage threshold is also higher for this orientation as well.**

Polarization Definitions

BIREFRINGENCE

A birefringent crystal, such as calcite, will divide an entering beam of monochromatic light into two beams having opposite polarization. The beams usually propagate in different directions and will have different speeds. There will be only one or two optical axis directions within the crystal in which the beam will remain collinear and continue at the same speed, depending on whether the birefringent crystal is uniaxial or biaxial.

If the crystal is a plane-parallel plate, and the optical axis directions are not collinear with the beam, radiation will emerge as two separate, orthogonally polarized beams (see Figure 1.51). The beam will be unpolarized where the beams overlap upon emergence. The two new beams within the material are distinguished from each other by more than just polarization and velocity. The rays are referred to as extraordinary (E) and ordinary (O). These rays need not be confined to the plane of incidence. Furthermore, the velocity of these rays changes with direction. Thus, the index of refraction for extraordinary rays is also a continuous function of direction. The index of refraction for the ordinary ray is constant and is independent of direction.

The two indexes of refraction are equal only in the direction of an optical axis within the crystal. The dispersion curve for ordinary rays is a single, unique curve when the index of refraction is plotted against wavelength. The dispersion curve for the extraordinary ray is a family of curves with different curves for different directions. Unless it is in a particular polarization state, or the crystalline surface is perpendicular to an optical axis, a ray normally incident on a birefringent surface will be divided in two at the boundary. The extraordinary ray will be deviated; the ordinary ray will not. The ordinary ray index n, and the most extreme (whether greater or smaller) extraordinary ray index n_e, are together known as the principal indices of refraction of the material.

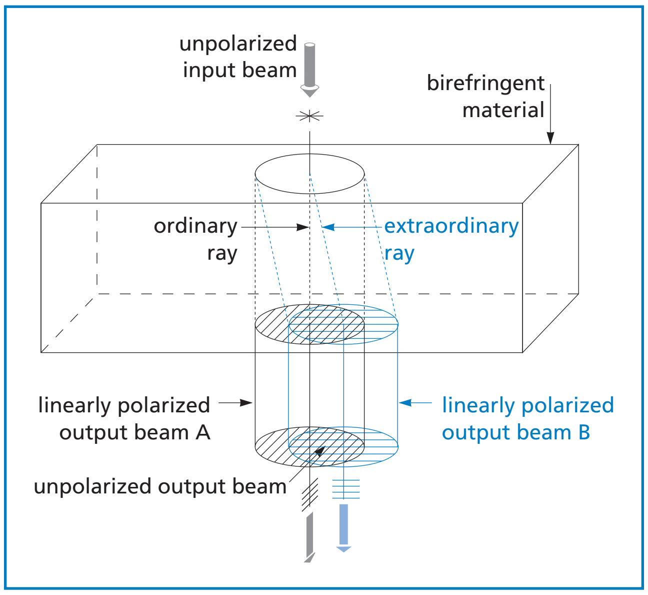

Figure 1.51 **Double refraction in a birefringent crystal**

If a beam of linearly polarized monochromatic light enters a birefringent crystal along a direction not parallel to the optical axis of the crystal, the beam will be divided into two separate beams. Each will be polarized at right angles to the other and will travel in different directions. The original beam energy, which will be divided between the new beams, depends on the original orientation of the vector to the crystal.

The energy ratio between the two orthogonally polarized beams can be any value. It is also possible that all energy will go into one of the new beams. If the crystal is cut as a plane-parallel plate, these beams will recombine upon emergence to form an elliptically polarized beam.

The difference between the ordinary and extraordinary ray may be used to create birefringent crystal polarization devices. In some cases, the difference in refractive index is used primarily to separate rays and eliminate one of the polarization planes, for example, in Glan-type polarizers. In other cases, such as Wollaston and Thompson beamsplitting prisms, changes in propagation direction are optimized to separate an incoming beam into two orthogonally polarized beams.

DICHROISM

Dichroism is selective absorption of one polarization plane over the other during transmission through a material. Sheet-type polarizers are manufactured with organic materials imbedded into a plastic sheet. The sheet is stretched, aligning molecules and causing them to be birefringent, and then dyed. The dye molecules selectively attach themselves to aligned polymer molecules, so that absorption is high in one plane and weak in the other. The transmitted beam is linearly polarized. Polarizers made of such material are very useful for low-power and visual applications. The usable field of view is large (up to grazing incidence), and diameters in excess of 100 mm are available.

POLARIZATION BY REFLECTION

When a beam of ordinary light is incident at the polarizing angle on a transmissive dielectric such as glass, the emerging refracted ray is partially linearly polarized. For a single surface (with n=1.50) at Brewster's angle, 100 percent of the light whose electric vector oscillates parallel to the plane of incidence is transmitted. Only 85 percent of the perpendicular light is transmitted (the other 15 percent is reflected). The degree of polarization from a single-surface reflection is small.

If a number of plates are stacked parallel and oriented at the polarizing angle, some vibrations perpendicular to the plane of incidence will be reflected at each surface, and all those parallel to it will be refracted. By making the number of plates within the stack large (more than 25), high degrees of linear polarization may be achieved. This polarization method is utilized in CVI Melles Griot polarizing beamsplitter cubes which are coated with many layers of quarter-wave dielectric thin films on the interior prism angle. This beamsplitter separates an incident laser beam into two perpendicular and orthogonally polarized beams.

THIN METAL FILM POLARIZERS

Optical radiation incident on small, elongated metal particles will be preferentially absorbed when the polarization vector is aligned with the long axis of the particle. CVI Melles Griot infrared polarizers utilize this effect to make polarizers for the near-infrared. These polarizers are considerably more effective than dichroic polarizers.

Polarizing thin films are formed by using the patented Slocum process to deposit multiple layers of microscopic silver prolate spheroids onto a polished glass substrate. The exact dimensions of these spheroids determine the optical properties of the film. Peak absorption can be selected for any wavelength from 400 to 3000 nm by controlling the deposition process. Contrast ratios up to 10,000:1 can be achieved with this method. Other CVI Melles Griot high-contrast polarizers exhibit contrasts as high as 100,000:1.

CALCITE

Calcite, a rhombohedral crystalline form of calcium carbonate, is found in various forms such as limestone and marble. Since calcite is a naturally occurring material, imperfections are not unusual. The highest quality materials, those that exhibit no optical defects, are difficult to find and are more expensive than those with some defects. Applications for calcite components typically fall into laser applications or optical research. CVI Melles Griot offers calcite components in two quality grades to meet those various needs.

There are three main areas of importance in defining calcite quality.

Spectral Properties

Trace amounts of chemical impurities, as well as lattice defects, can cause calcite to be colored, which changes absorption. For visible light applications, it is essential to use colorless calcite. For near-infrared applications, material with a trace of yellow is acceptable. This yellow coloration results in a 15-percent to 20-percent decrease in transmission below 420 nm.

Wavefront Distortion (Striae)

Striae, or streaked fluctuations in the refractive index of calcite, are caused by dislocations in the crystal lattice. They can cause distortion of a light wavefront passing through the crystal. This is particularly troublesome for interferometric applications.

Scatter

Small inclusions within the calcite crystal account for the main source of scatter. They may appear as small cracks or bubbles. In general, scatter presents a significant problem only when the polarizer is being used with a laser. The amount of scatter centers that can be tolerated is partially determined by beam size and power.

CVI MELLES GRIOT CALCITE GRADES

CVI Melles Griot has selected the most applicable calcite qualities, grouped into two grades:

Laser Grade

Calcite with a wavefront deformation of $\lambda/4$ at 633 nm or better due to striae only.

Optical Grade

Calcite with a wavefront deformation of 1λ to $\lambda/4$ at 633 nm due to striae only.

Waveplates

Waveplates use birefringence to impart unequal phase shifts to the orthogonally polarized field components of an incident wave, causing the conversion of one polarization state into another.

There are two types of birefringence. With linear birefringence, the index of refraction (and hence the phase shift) differs for two orthogonally polarized linear polarization states. This is the operation mode of standard waveplates. With circular birefringence, the index of refraction and hence phase shift differs for left and right circularly polarized components. This is the operation mode of polarization rotators.

STANDARD WAVEPLATES: LINEAR BIREFRINGENCE

Suppose a waveplate made from a uniaxial material has light propagating perpendicular to the optic axis. This makes the field component parallel to the optic axis an extraordinary wave and the component perpendicular to the optic axis an ordinary wave. If the crystal is positive uniaxial, $\eta_e > \eta_o$, then the optic axis is called the slow axis, which is the case for crystal quartz. For negative uniaxial crystals $\eta_e < \eta_o$, the optic axis is called the fast axis.

The equation for the transmitted field E_2, in terms of the incident field E_1 is:

$$E_2 = \mathrm{s}(\mathrm{s} \bullet E_1)e^{i\phi_s} + \mathrm{f}(\mathrm{f} \bullet E_1)e^{i\phi_f} \qquad (1.102)$$

where *s* and *f* are unit vectors along the slow and fast axes. This equation shows explicity how the waveplate acts on the field. Reading from left to right, the waveplate takes the component of the input field along its slow axis and appends the slow axis phase shift to it. It does a similar operation to the fast component.

The slow and fast axis phase shifts are given by:

$$\phi_s = \eta_s(\omega)\omega t/c = 2\pi\eta_s(\lambda)t/\lambda \qquad (1.103)$$

$$\phi_f = \eta_f(\omega)\omega t/c = 2\pi\eta_f(\lambda)t/\lambda \qquad (1.104)$$

where η_s and η_f are, respectively, the indices of refraction along the slow and fast axes, and *t* is the thickness of the waveplate.

To further analyze the effect of a waveplate, we throw away a phase factor lost in measuring intensity, and assign the entire phase delay to the slow axis:

$$E_2 = \mathrm{s}(\mathrm{s} \bullet E_1)e^{i\phi} + \mathrm{f}(\mathrm{f} \bullet E1) \qquad (1.105)$$

$$\phi = \phi_s - \phi_f = 2\pi\left[\eta_s(\lambda) - \eta_f(\lambda)\right]t/\lambda \qquad (1.106)$$

$$\phi = 2\,\pi\,\Delta\eta(\lambda)t/\lambda \qquad (1.107)$$

In the above, $\Delta\eta(\lambda)$ is the birefringence $\eta_s(\lambda)$ - $\eta_f(\lambda)$. The dispersion of the birefringence is very important in waveplate design; a quarter waveplate at a given wavelength is never exactly a half waveplate at half that wavelength.

Let E_1 be initially polarized along *x*, and let the waveplate slow axis make an angle θ with the *x* axis. This orientation is shown in Figure 1.52.

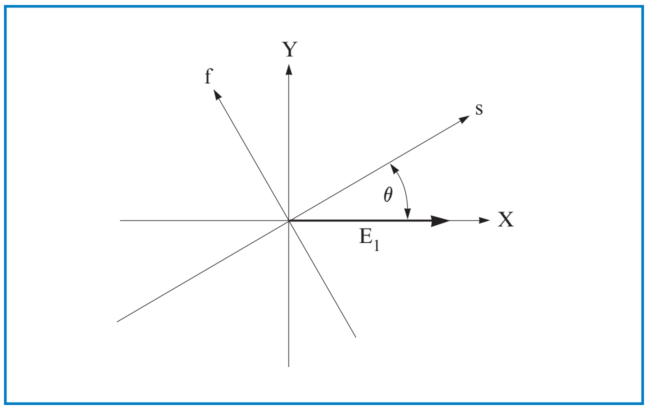

Figure 1.52. Orientation of the slow and fast axes of a waveplate with respect to an *x*-polarized input field

When the waveplate is placed between parallel and perpendicular polarizers the transmissions are given by:

$$T_{\parallel} \propto |E_2 x|^2 = 1 - \sin^2\ 2\theta \sin^2 \phi/2 \qquad (1.108)$$

$$T_{\perp} \propto |E_2 y|^2 = \sin^2\ 2\theta \sin^2 \phi/2 \qquad (1.109)$$

Note that θ is only a function of the waveplate orientation, and ϕ is only a function of the wavelength, the birefringence is a function of wavelength and the plate thickness.

For a full-wave waveplate:

$\phi = 2m\pi$, $T_{\parallel} = 1$, and $T_{\perp} = 0$, regardless of waveplate orientation.

For a half-wave waveplate:

$$\phi = (2m + 1)\pi,\ T_{\parallel} = \cos^2 2\theta, \text{ and } T_{\perp} = \sin^2 2\theta. \qquad (1.110)$$

This transmission result is the same as if an initial linearly polarized wave were rotated through an angle of 2θ. Thus, a half-wave waveplate finds use as a polarization rotator.

For a quarter waveplate,

$$\phi = (2m + 1)\pi/2 \text{ (i.e., an odd multiple of } \pi/2\text{).} \qquad (1.111)$$

To analyze this, we have to go back to the field equation. Assume that the slow and fast axis unit vectors s and f form a right handed coordinate system such that s x f = +z, the direction of propagation. To obtain circularly polarized light, linearly polarized light must be aligned midway between the slow and fast axes. There are four possibilities listed in the table below.

Phase Shift	Input Field Along $(\mathbf{s} + \mathbf{f})/\sqrt{2}$	Input Field Along $(\mathbf{s} - \mathbf{f})/\sqrt{2}$
$\phi = \pi/2 + 2m\pi$	RCP	LCP
$\phi = 3\pi/2 + 2m\pi$	LCP	RCP

Sometimes, waveplates described by the second line above are called 3/4 waveplates. For multiple order waveplates, CVI Melles Griot permits the use of either of the above classes of waveplates to satisfy the requirements of a quarter-wave waveplate.

MULTIPLE-ORDER WAVEPLATES

For the full-, half-, and quarter-wave waveplate examples given in standard waveplates, the order of the waveplate is given by the integer m. For $m > 0$, the waveplate is termed a multiple-order waveplate. For $m = 0$, we have a zero order waveplate.

The birefringence of crystal quartz near 500 nm is approximately 0.00925. Consider a 0.5-mm-thick crystal quartz waveplate. A simple calculation shows that this is useful as a quarter waveplate for 500 nm; in fact, it is a 37 $\lambda/4$ waveplate at 500 nm with $m = 18$. Multiple-order waveplates are inexpensive, high-damage-threshold retarders. Further analysis shows that this same 0.5mm plate is a 19 $\lambda/2$ half waveplate at 488.2 nm and a 10λ full-wave waveplate at 466.5 nm. The transmission of this plate between parallel polarizers is shown in Figure 1.53 as a function of wavelength. The retardance of the plate at various key points is also shown. Note how quickly the retardance changes with wavelength. Because of this, multiple-order waveplates are generally useful only at their design wavelength.

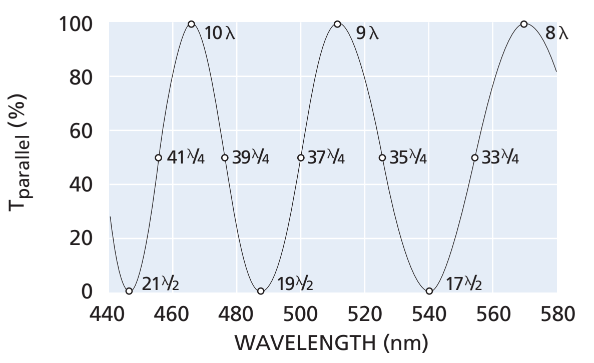

Figure 1.53 **Transmission of a 0.5-mm-thick crystal quartz waveplate between parallel polarizers**

ZERO-ORDER WAVEPLATES

As discussed above, multiple-order waveplates are not useful with tunable or broad bandwidth sources (e.g., femtosecond lasers). A zero-order waveplate can greatly improve the useful bandwidth in a compact, high-damage-threshold device.

As an example, consider the design of a broadband half-wave waveplate centered at 800 nm. Maximum tuning range is obtained if the plate has a single π phase shift at 800 nm. If made from a single plate of crystal quartz, the waveplate would be about 45 μm thick, which is too thin for easy fabrication and handling. The solution is to take two crystal quartz plates differing in thickness by 45 μm and align them with the slow axis of one against the fast axis of the other. The net phase shift of this zero-order waveplate is π. The two plates may be either air-spaced or optically contacted. The transmission of an 800-nm zero-order half-wave waveplate between parallel polarizers is shown in Figure 1.54 using a 0-10% scale. Its extinction is better than 100:1 over a bandwidth of about 95 nm centered at 800 nm.

CVI Melles Griot produces multiple order and zero order crystal quartz waveplates at any wavelength between 193 nm and 2100 nm. Virtually all popular laser wavelengths are kept in stock, and custom wavelength parts are available with short delivery time.

Mica waveplates are an inexpensive zero-order waveplate solution for low-power applications and in detection schemes.

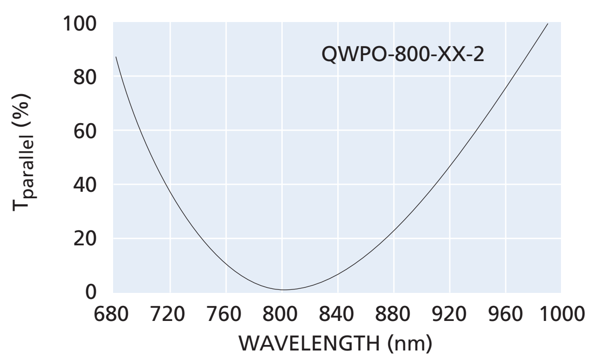

Figure 1.54 **Zero-order crystal quartz half-wave waveplate for 800 nm**

POLYMER WAVEPLATES

Polymer waveplates offer excellent angular field of view since they are true zero-order waveplates. Figure 1.55 compares the change in retardance as function of incidence angle for polymer and quartz waveplates. A polymer waveplate changes by less than 1% over a ±10° incidence angle.

Retardance accuracy with wavelength change is often of key concern. For example, an off-the-shelf diode laser has a center wavelength tolerance of ±10 nm. Changes with temperature and drive conditions cause wavelength shifts which may alter performance. These polymer waveplates maintain excellent waveplate performance even with minor shifts in the source wavelength. The temperature sensitivity of laminated polymer waveplates is about 0.15 nm/°C, allowing operation over moderate temperature ranges without significantly degrading retardance accuracy. A comparison of different waveplate types and their dependence on wavelength is shown in figure 1.56.

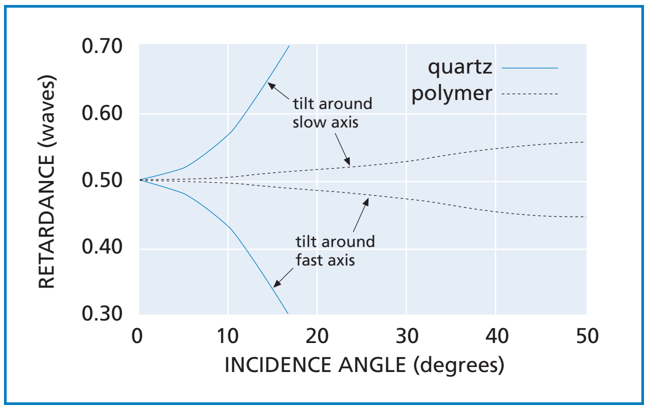

Figure 1.55 **Retardance vs incidence angle for quartz and polymer waveplates**

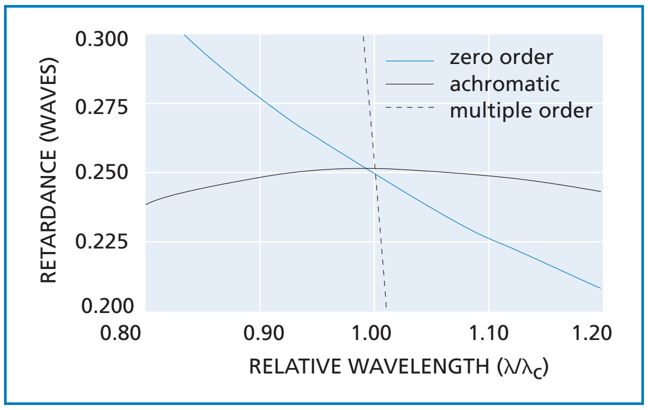

Figure 1.56 **Wavelength performance of common quarter wave retarders.**

ACHROMATIC WAVEPLATES

At 500 nm, a crystal quartz zero-order half-wave waveplate has a retardation tolerance of λ/50 over a bandwidth of about 50 nm. This increases to about 100 nm at a center wavelength of 800 nm. A different design which corrects for dispersion differences over the wavelength range is required for bandwidths up to 300 nm.

If two different materials are used to create a zero-order or low-order waveplate, cancellation can occur between the dispersions of the two materials. Thus, the net birefringent phase shift can be held constant over a much wider range than in waveplates made from one material. ACWP-series achromatic waveplates from CVI Melles Griot (see Figure 1.57) are comprised of crystal quartz and magnesium fluoride to achieve achromatic performance.

Three wavelength ranges are available in both quarter and half wave retardances. Retardation tolerance is better than λ/100 over the entire wavelength range.

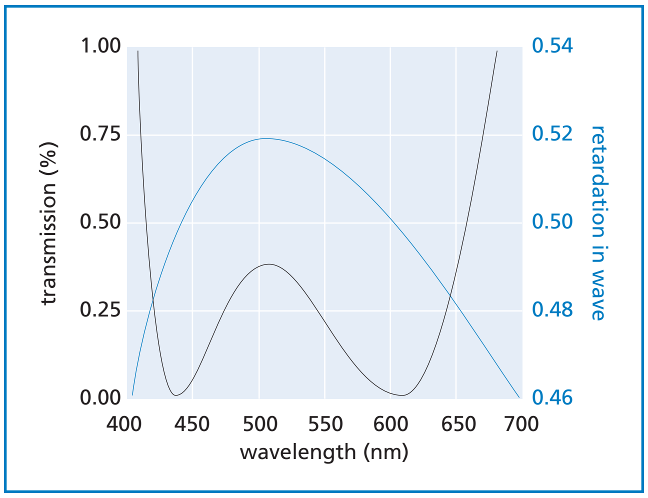

Figure 1.57 **ACWP-400-700-10-2**

For quarter-wave waveplates, perfect retardance is a multiple of 0.25 waves, and transmission through a linear polarizer must be between 33% and 67%. (In all but the shortest wavelength design, quarter-wave retardation tolerance is better than λ/100.) For half-wave waveplates, perfect retardance is 0.5 waves, while perfect transmission through a linear polarizer parallel to the initial polarization state should be zero. A high degree of achromatization is achievable by the dual material design. In addition, we manage low group velocity dispersion for ultrashort pulse applications through the use of thin plates.

DUAL-WAVELENGTH WAVEPLATES

Dual-wavelength waveplates are used in a number of applications. One common application is separation of different wavelengths with a polarizing beamsplitter by rotating the polarization of one wavelength by 90°, and leaving the other unchanged. This frequently occurs in nonlinear doubling or tripling laser sources such as Nd:YAG (1064/532/355/266).

One way to achieve the multiple retardation specifications is through careful selection of multiple-order waveplates which meet both wavelength and retardation conditions. This often results in the selection of a relatively high order waveplate. Therefore, these dual-wavelength waveplates operate best over a narrow bandwidth and temperature range.

Another approach is to combine two quartz waveplates with their optical axes orthogonal to one another, effectively creating a zero-order waveplate. In this configuration, the temperature dependence is a function of the thickness difference between the waveplates, resulting in excellent temperature stability. The retardation of the compound waveplate is also a function of the thickness difference enabling wide bandwidth performance.

Etalons

Etalons are most commonly used as line-narrowing elements in narrow-band laser cavities or as bandwidth-limiting and coarse-tuning elements in broadband and picosecond lasers. Further applications are laser line profile monitoring, diagnosis.

The etalons described in this section are all of the planar Fabry-Perot type. Typical transmission characteristics for this type etalon are shown in Figure 1.58.

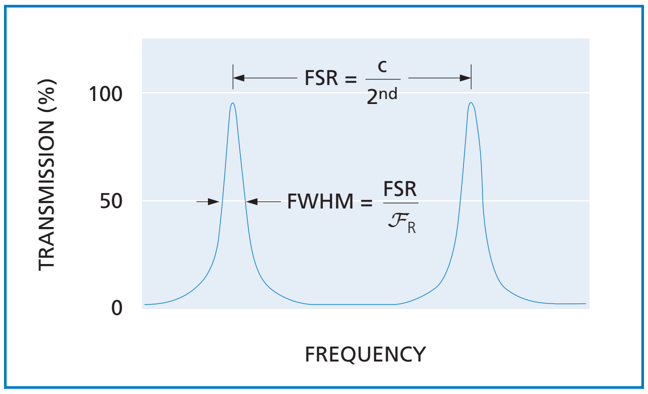

Figure 1.58 **Transmission characteristics of a Fabry-Perot etalon**

Air-Spaced Etalons consist of pairs of very flat plano-plano plates separated by optically contacted spacers. The inner surfaces of the plates are coated with partially reflecting coatings, the outer surfaces are coated with antireflection coatings.

Solid Etalons are made from a single plate with parallel sides. Partially reflecting coatings are then deposited on both sides. The cavity is formed by the plate thickness between the coatings.

Deposited Solid Etalons are a special type of solid etalon in which the cavity is formed by a deposited layer of coating material. The thickness of this deposited layer depends on the free spectral range required and can range from a few nanometers up to 15 micrometers. The cavity is sandwiched between the etalon reflector coatings and the whole assembly is supported on a fused-silica base plate.

Etalon plates need excellent surface flatness and plate parallelism. To avoid peak transmission losses due to scatter or absorption, the optical coatings also have to meet the highest standards.

For a plane wave incident on the etalon, the transmission of the etalon is given by:

$$T = \frac{I_{\text{trans}}}{I_{\text{inc}}} = \frac{1}{1+\frac{4R}{(1-R)^2}\sin^2(\delta/2)} \tag{1.112}$$

Here, R is the reflectance of each surface; δ is the phase shift

$$\delta = \frac{2\pi}{\lambda}\eta d\cos\theta \tag{1.113}$$

where,

η is the refractive index (e.g., 1 for air-spaced etalons)

d is the etalon spacing or thickness

θ is the angle of incidence

The free spectral range (FSR) of the etalon is given by

$$\begin{aligned} \text{FSR} &= \frac{c}{2\eta d} \text{ in Hz} \\ &= \frac{1}{2\eta d} \text{ in cm}^{-1} \\ &= \frac{\lambda^2}{2\eta d} \text{ in nm} \end{aligned} \tag{1.114}$$

The reflectivity finesse, $\mathcal{F}_R$ is given by

$$\mathcal{F}_R = \frac{\pi\sqrt{R}}{1-R} \tag{1.115}$$

Figure 1.59 shows the reflectivity finesse as a function of the coating reflectivity.

The bandwidth (FWHM) is given by

$$\text{FWHM} = \frac{\text{FSR}}{\mathcal{F}_R} \tag{1.116}$$

However, the above applies to theoretical etalons which are assumed to be perfect. In reality, even the best etalon will show defects that limit theoretically expected performance. Therefore, in a real etalon, the actual finesse will usually be lower than the reflectivity finesse.

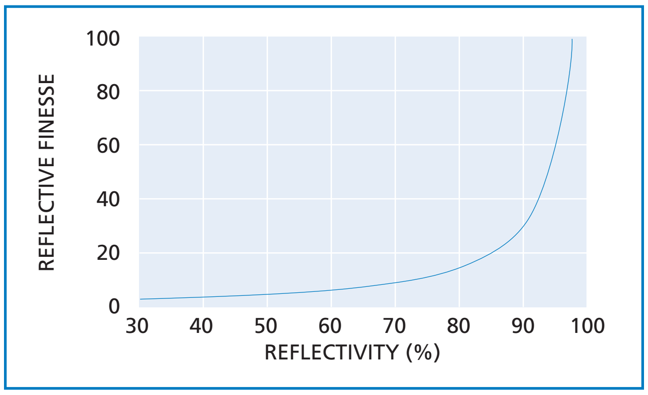

Figure 1.59 **Reflectivity finesse vs. coating reflectance of each surface**

$$\mathcal{F}_R = \frac{\pi\sqrt{R}}{1-R} \tag{1.117}$$

$$\mathcal{F}_S = \frac{M}{2}\,\frac{\lambda}{633\text{ nm}} \tag{1.118}$$

$$\mathcal{F}_\theta = \frac{\lambda}{d\tan^2\theta} \tag{1.119}$$

$$\mathcal{F}_D = \frac{\mathrm{CA}^2}{2d\lambda} \tag{1.120}$$

where $\mathcal{F}_R$ is the reflectivity finesse, $\mathcal{F}_s$ is the plate spherical deviation finesse coefficient, $\mathcal{F}_\theta$ is the incident beam divergence finesse coefficient, and $\mathcal{F}_d$ is the diffraction-limited finesse coefficient.

The defects that contribute to this reduction are as shown in Figure 1.60 (graphical representations are exagerated for clarification).

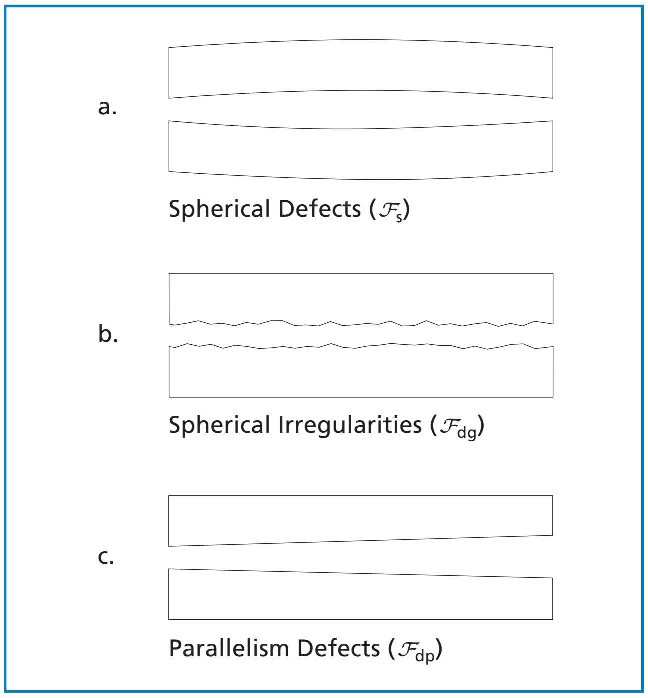
a.
Spherical Defects ($\mathcal{F}_s$)
b.
Spherical Irregularities ($\mathcal{F}_{dg}$)
c.
Parallelism Defects ($\mathcal{F}_{dp}$)

Figure 1.60 **Three types of defects contributing to the total defect finesse**

All three types of defects contribute to the total defect finesse $\mathcal{F}_d$:

$$\frac{1}{\mathcal{F}_d{}^2} = \frac{1}{\mathcal{F}_S{}^2} + \frac{1}{\mathcal{F}_{dg}{}^2} + \frac{1}{\mathcal{F}_{dp}{}^2} \tag{1.121}$$

The beam divergence also influences the actual finesse of an etalon.

Taking into account all these contributions, the effective finesse ($\mathcal{F}_e$) of an etalon (with $\mathcal{F}_R$ being the reflectivity finesse and $\mathcal{F}_\theta$ the divergence finesse) is:

$$\frac{1}{\mathcal{F}_e} = \sqrt{\frac{1}{\mathcal{F}_R{}^2} + \frac{1}{\mathcal{F}_D{}^2} + \frac{1}{\mathcal{F}_\theta{}^2} + \frac{1}{\mathcal{F}_S{}^2}} \tag{1.122}$$

The effective finesse a user sees when using the etalon depends not only on the absolute clear aperture, but also on the used aperture of the etalon, especially when a high finesse is required.

The examples below show how the effective finesse varies with plate flatness and clear aperture.

Example 1: Air-spaced etalon,

$R = 95\%$ (± 1%) at 633 nm

CA = 25 mm

Used aperture = 20 mm

Spacer (air gap) = 1 mm

Spherical / parallelism defects = <λ/20

Plate rms = 0.80 nm

Beam divergence = 0.1 mRad

$\mathcal{F}_R = 61$, $\mathcal{F}_e = 10$

Example 2: Air-spaced etalon,

Same parameters as example 1 except:

Used aperture = 5 mm

$\mathcal{F}_R = 61$, $\mathcal{F}_e = 40$ (± 4)

Example 3: Air-spaced etalon,

Same parameters as example 1 except:

Spherical / parallelism defects = λ/100

Plate rms = 0.40 nm

Beam divergence: 0.1mRad

$\mathcal{F}_R = 61$, $\mathcal{F}_e = 40$ (± 8)

These examples illustrate that, for large-aperture applications, it is important to use very high-quality plates to ensure a high finesse and good transmission values.

TUNING AN ETALON

Etalons can be tuned over a limited range to alter their peak transmission wavelengths. These techniques are:

Angle tuning or tilting the etalon: As the angle of incidence is increased, the center wavelength of the etalon can be tuned down the spectrum.

Temperature tuning: Primarily used for solid etalons, temperature-tuning changes both the actual spacing of the reflective surfaces via expansion and the index of refraction of the material, which changes the optical spacing. The tuning result can be given by

$$\frac{\partial(\mathrm{FSR})}{\partial T} = -(\mathrm{FSR})\left[\frac{1}{\mathrm{n}}\frac{\partial \eta}{\partial T} + \frac{1}{d}\frac{\partial d}{\partial T}\right] \tag{1.123}$$

Pressure tuning: Air-spaced etalons can be tuned by increasing the pressure in the cavity between the optics, thereby increasing the effective index of refraction, and thus the effective spacing.

The above examples illustrate how critical the optical surface flatness, plate parallelism and surface quality are to the overall performance of an etalon. At CVI Melles Griot we have developed sophisticated software that allows us to simulate all effects that influence the performance of an etalon. To order an etalon, FSR, finesse and used aperture are required.

Ultrafast Theory

The distinguishing aspect of femtosecond laser optics design is the need to control the phase characteristic of the optical system over the requisite wide pulse bandwidth. CVI Melles Griot has made an intensive theoretical study of these effects. Certain coating designs have been modified with control of the phase characteristics in mind. New proprietary designs have been created with desirable characteristics for femtosecond researchers. All optics in this section have been tested by researchers in the field and we are constantly fielding new requests.

Assume that the power, reflectivity, and polarization characteristics of a laser mirror are acceptable over the bandwidth of a femtosecond pulse. This means that, over the entire pulse bandwidth, a cavity mirror may have a reflectivity greater than 99.8%; a 50% beamsplitter may have a fairly constant reflection; a polarizer may maintain its rejection of one polarization with an acceptable transmission of the other. It is not enough, however, to simply preserve the power spectrum $S(\omega) = |E(\omega)|^2$ when dealing with femtosecond pulses. The phase relationship among the Fourier components of the pulse must also be preserved in order that the pulse not be broadened or distorted. What constraint on the performance of a mirror or transmissive optic does this imply?

Consider a general initial pulse shape $E_0(t)$. As a function of its Fourier components, it may be expressed as:

$$E_0(t) = \int E(\omega)^{-i\omega t}\, d\omega \qquad (1.124)$$

Suppose this pulse reflects off of a mirror. For this example, we assume the mirror is "ideal", and use the Fourier transform of its complex amplitude reflectance:

$$r(\omega) = re^{+i\Phi(\omega)} = re^{+i\omega t_\mathrm{d}} \qquad (1.125)$$

For this "ideal" mirror, r is a real constant equal to the amplitude reflectivity that is assumed constant over the pulse bandwidth. All phase effects have been assumed to be describable by a single phase shift $\phi(\omega)$ that is linearly proportional to frequency with proportionality constant t_d. The reflected pulse is then:

$$\begin{aligned} E_\mathrm{r}(t) &= r\int E(\omega)^{-i\omega(t-t_\mathrm{d})}\, d\omega \\ &= rE(t-t_\mathrm{d}) \end{aligned} \qquad (1.126)$$

Thus, provided the phase shift is linear in frequency over the pulse bandwidth, the reflected pulse is scaled by the amplitude reflectance r, and delayed in time by the constant group delay t_d. It is, otherwise, an undistorted replica of the original pulse.

Examined over a large enough bandwidth, no optical system will exhibit the constant group delay over frequency needed for perfect fidelity. In general, the phase shift near some center frequency ω_0 may be expanded in a Taylor series for frequencies near ω_0:

$$\begin{aligned} \Phi(\omega) = \Phi(\omega_0) + \Phi'(\omega_0)(\omega-\omega_0) + \\ \Phi''(\omega_0)(\omega-\omega_0)^2/2! + \\ \Phi'''(\omega_0)(\omega-\omega_0)^3/3!\ \ldots \end{aligned} \qquad (1.127)$$

These derivatives are, respectively, the group delay $\Phi'(\omega_0)$, the group velocity dispersion $\Phi''(\omega_0)$, and the cubic term $\Phi'''(\omega_0)$, evaluated at a center frequency ω_0. This expansion is heuristically useful, in an exactly soluble model, for the propagation of a transform-limited Gaussian pulse. Note, however, that for extremely short pulses the expansion above may be insufficient. A full numerical calculation may have to be performed using the actual phase shift function $\Phi(\omega)$. CVI Melles Griot will be happy to assist those interested in the modeling of real optical elements.

To illustrate pulse distortion due to the dependence of the group delay on frequency, consider what happens when an unchirped, transform-limited Gaussian pulse passes through a medium, or is incident on a mirror whose dominating contribution to phase distortion is non-zero group velocity dispersion. The field envelope of the pulse is assumed to be of the form:

$$E(t) = \exp\left[-\left(2\ \ln 2t^2/\tau_0^2\right)\right] \qquad (1.128)$$

where τ_0 is the initial pulse duration (FWHM of the pulse intensity). Let the pulse enter a medium or reflect off of a mirror with non-zero $\Phi''(\omega)$, measured in fsec^2 radians. (For a continuous medium-like glass, $\Phi''(\omega) = \beta''(\omega) \times z$ where $\beta''(\omega)$ is the group velocity dispersion (GVD) per centimeter of material, and z is the physical path length, in centimeters, traveled through the material.) The Gaussian pulse will be both chirped and temporally broadened by its encounter with group velocity dispersion. The power envelope will remain Gaussian; the result for the broadened FWHM is:

$$\tau_1 = \tau_0\left[1 + \left(4\ln 2\Phi''(\omega)/\tau_0^2\right)^2\right]^{1/2} \qquad (1.129)$$

This result, valid only for initially unchirped, transform-limited Gaussian pulses, is nevertheless an excellent model to study the effects of dispersion on pulse propagation. The graphs shown in Figure 1.61 represent the theoretical broadening from dispersion for initial pulse widths ranging from 10 to 100 femtoseconds.

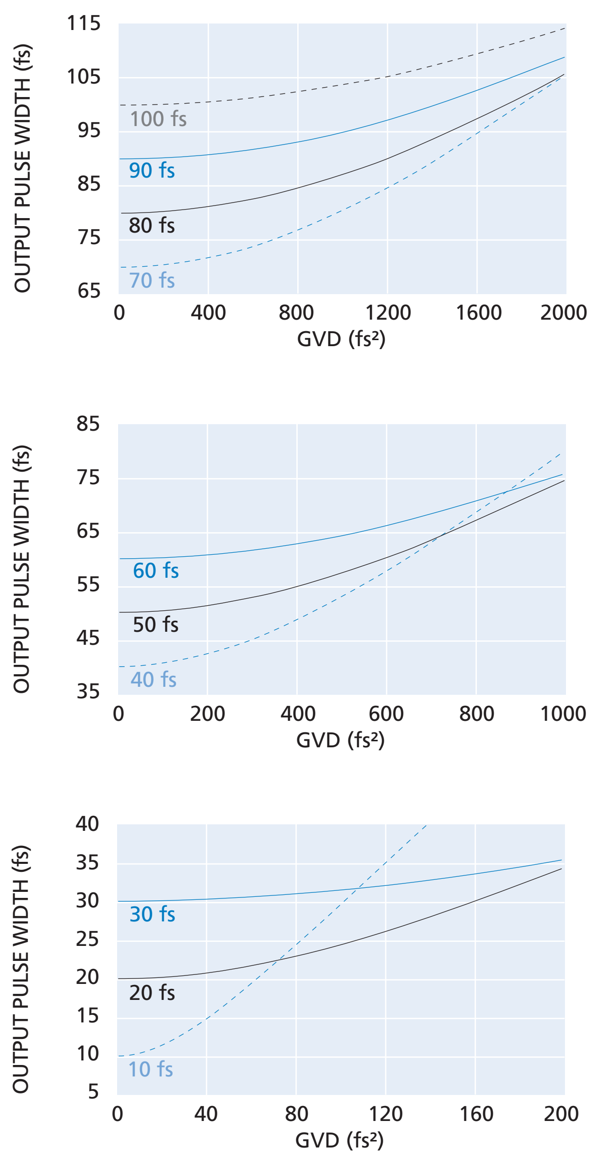

Figure 1.61 **Output pulse width vs. GVD**

GROUP-VELOCITY & CUBIC DISPERSION FOR VARIOUS OPTICAL MATERIALS

Figures 1.62 and 1.63 show the GVD and cubic dispersion respectively for some common used glasses. Some of the glasses can be used in the UV region. They should be useful in estimating material dispersion and pulse distortion effects. Please check these calculations independently before using them in a final design.

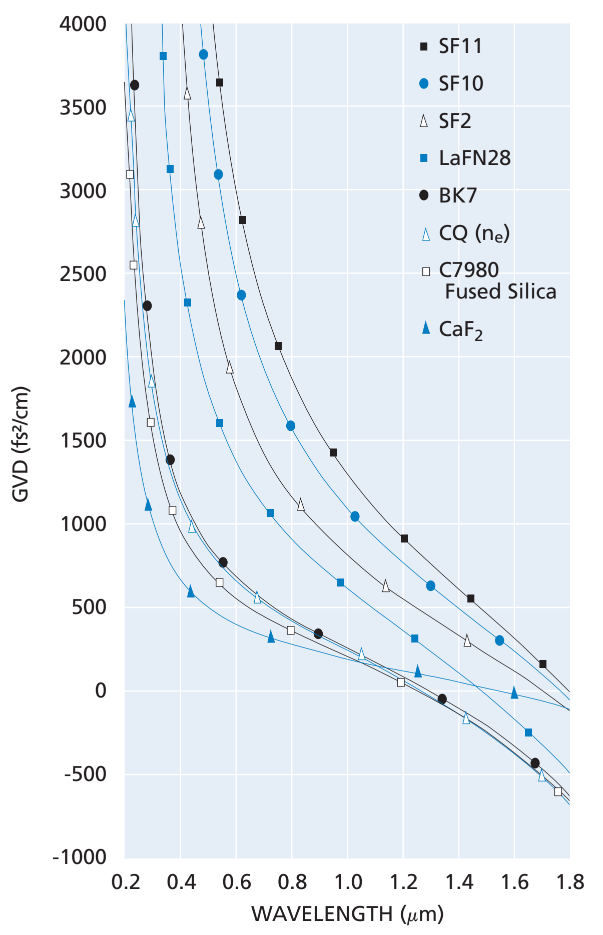

Figure 1.62 **GVD for common glasses**

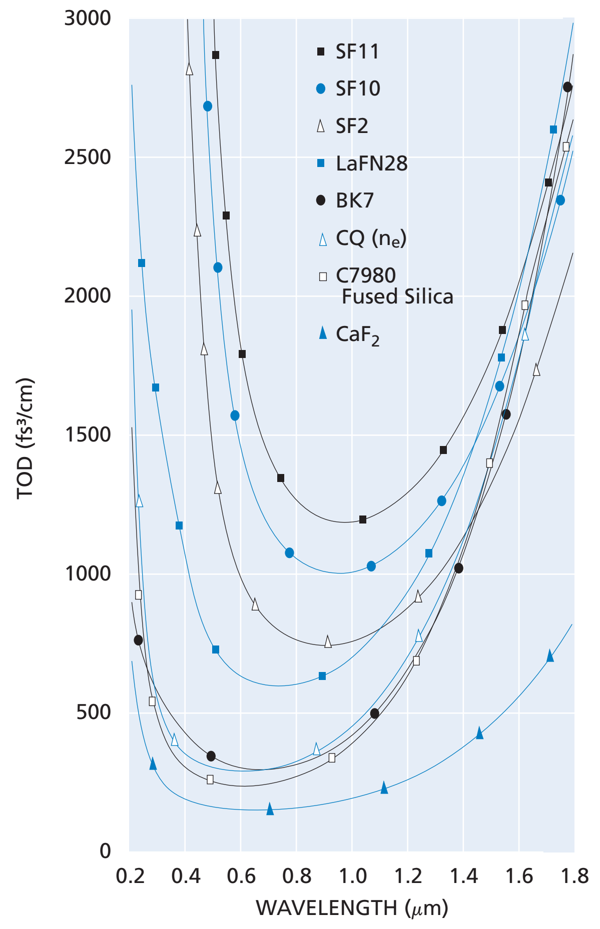

Figure 1.63 **Cubic dispersion for common glasses**

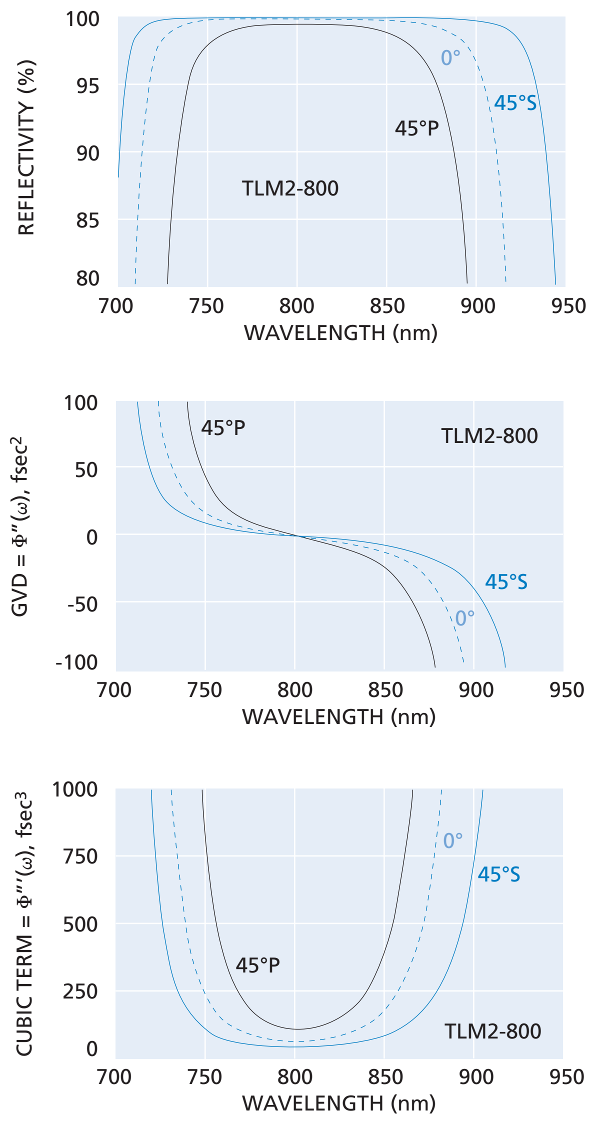

Figure 1.64 **Dispersion and reflectivity for mirrors TLM2-800-0 and TLM2-800-45**

DISPERSIVE PROPERTIES OF MIRRORS

CVI Melles Griot uses three basic designs; TLM1 mirrors for energy fluence greater than 100 mJ/cm^2, TLM2 mirrors for cw oscillators and low-fluence pulses, and TLMB mirrors which are a hybrid of the two. The reflectivity, GVD parameter, and cubic dispersion parameter for TLM2 high reflectors are shown in Figure 1.64. In these examples, the mirrors are centered at 800 nm and are designed for use at normal incidence and at 45 degrees. Note that, at the design wavelength, (a) GVD is zero, (b) the cubic term is minimized, and (c) at 45° incidence, the GVD of the p-polarization component is very sensitive to wavelength, while the GVD for s-polarization component is nearly zero over a broad wavelength range. Thus one should avoid using mirrors at 45° incidence with the p-polarization. On the other hand, at 45° incidence, s-polarization provides very broad bandwidth and minimizes pulse distortion problems and should be used whenever possible.

Ti:Sapphire and other femtosecond laser systems need prismless compensation of the built-in positive chirp encountered in the laser optical circuit. This becomes mandatory in industrial and biomedical applications where the laser must provide a compact, stable, and reliable solution.

Negative Group Velocity Dispersion Mirrors (TNM2) meet these needs with off-the-shelf availability, and can be employed both intracavity and extracavity to satisfy chirp control requirements.

In experiments using CVI Melles Griot TNM2 negative group velocity dispersion mirrors, 200-mW, 80-fsec pulses centered at 785 nm were achieved in a simple, prismless, Ti:Sapphire oscillator. The configuration is shown in Figure 1.65.

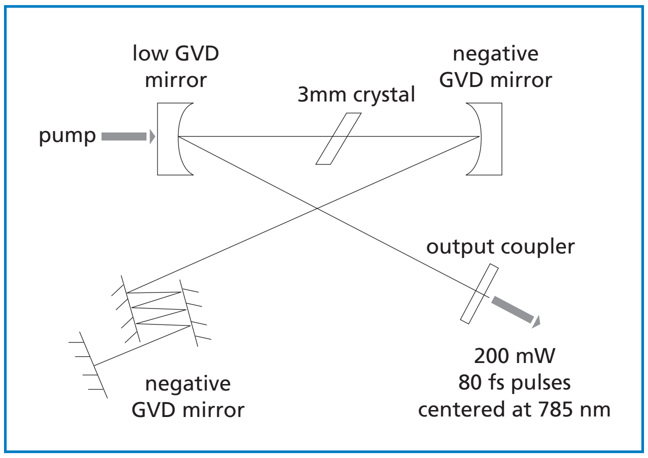

Figure 1.65 **Typical optical setup incorporating low GVD and Negative GVD mirrors in an ultrafast application**

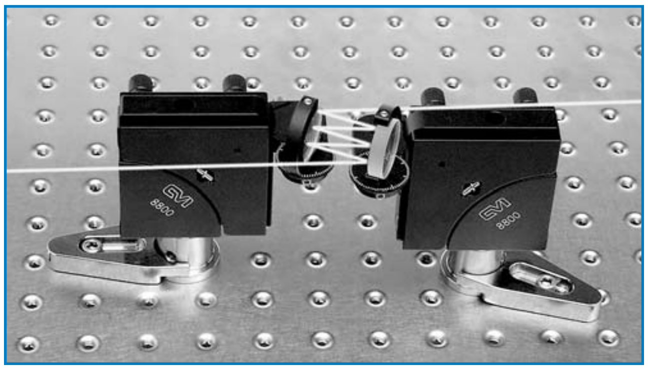

Figure 1.66 **Typical optical set-up of negative GVD mirrors**

OUTPUT COUPLERS AND BEAMSPLITTERS

Output-coupler partial reflectors and beamsplitters behave similarly; however, here is an additional consideration in their analysis. The behavior of the transmitted phase of the coating and the effect of material dispersion within the substrate on the transmitted beam have to be taken into account in a detailed analysis. In general, the coating transmitted phase has similar properties and magnitudes of GVD and cubic to the reflected phase. As usual, centering is important. As a beamsplitter, we recommend the 1.5 mm thick fused silica substrate PW-1006-UV. As an output coupler substrate, we recommend the 3.0 mm thick, 30 minute wedge fused silica substrate IF-1012-UV.

CVI Melles Griot has developed the TFPK Series Broadband Low Dispersion Polarizing Beamsplitters to satisfy requirements for very-high-power, short-pulse lasers. These optics are ideal for intracavity use in femtosecond regenerative amplifiers. The main emphasis is on linear phase characteristics. See Chapter 9 of Lasers, A. E. Siegman (University Science Books, Mill Valley, California, 1986), for a good discussion of linear pulse propagation.

In chirped pulse regenerative amplification, the pulse may have to pass through one or two polarizers twice per round trip. There can be 10 to 20 round trips before the gain is saturated and the pulse is ejected. At this stage the pulse is long (100–1000 psec); however the phase shift at each frequency must still be maintained to minimize the recompressed pulse width. The many round trips of the pulse in the regenerative amplifier put stringent requirements on the phase characteristics of the coatings.

Figure 1.67 shows the power transmission curves for both s- and p-polarization and the transmitted phase characteristics of the p component for a TFPK optimized at 800 nm. (Users may specify any wavelength

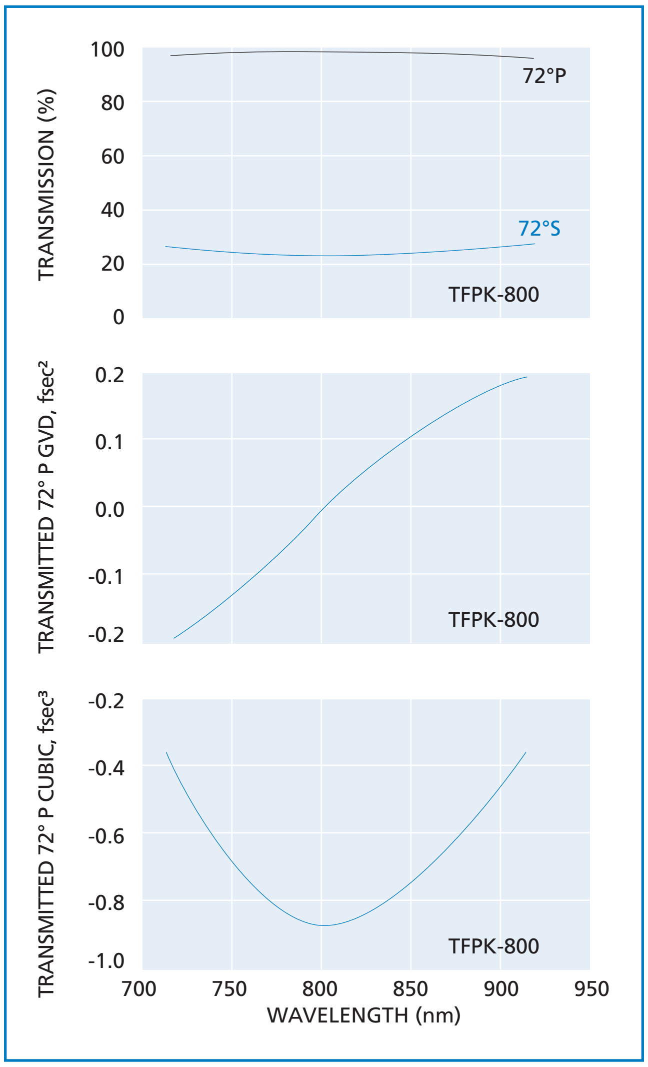

Figure 1.67 **Properties for one coated side of a TFPK polarizing beamsplitter optimized for 800 nm. Both sides are coated for these properties.**

Fundamental Optics
Gaussian Beam Optics
Optical Specifications
Material Properties
Optical Coatings

from 250 nm to 1550 nm.) The phase characteristics shown are the GVD and the cubic phase term. Not shown are the reflected phase characteristics for *s*-polarization; they are similar to the *p*-polarization transmission curves, and have the same low nonlinearity and broad bandwidth. Note that both sides of the optic have the coating whose properties are described in Figure 1.67. Therefore, the *s*- and *p*-polarization transmissions per surface should be squared in determining the specifications. The phase characteristics show that in all modes of operation, the TFPK polarizer performance is dominated by the substrate.

There are some subtleties associated with the TFPK. The near 72° angle has to be set properly and optimized. Some thought has to be given to mechanical clearances of the laser beam at such a steep incidence angle. The reflectivity for *s*-polarization is limited to 75%. Variant designs can increase this at a slight loss in bandwidth, increase in incidence angle, and increase in insertion loss for the transmitted *p*-polarized component.

The FABS autocorrelator beamsplitters from CVI Melles Griot are broadband, 50% all-dielectric beamsplitters. They are useful in many types of pump-probe experiments and in the construction of antiresonant ring configurations. They are essentially lossless and extremely durable. Both have advantages over partially reflecting metal coatings.

Power transmission curves for the *s*- and *p*-polarized versions of the FABS, along with the corresponding reflected phase characteristics for beamsplitters optimized at 800 nm, are shown in Figures 1.68 and 1.69. The linear pulse propagation properties of these beamsplitters are dominated by the substrate material dispersion. As with virtually all dielectric coated optics, the *s*-polarized version is broader than *p*-polarized version. CVI Melles Griot can produce FABS in other than 50:50 with excellent phase characteristics.

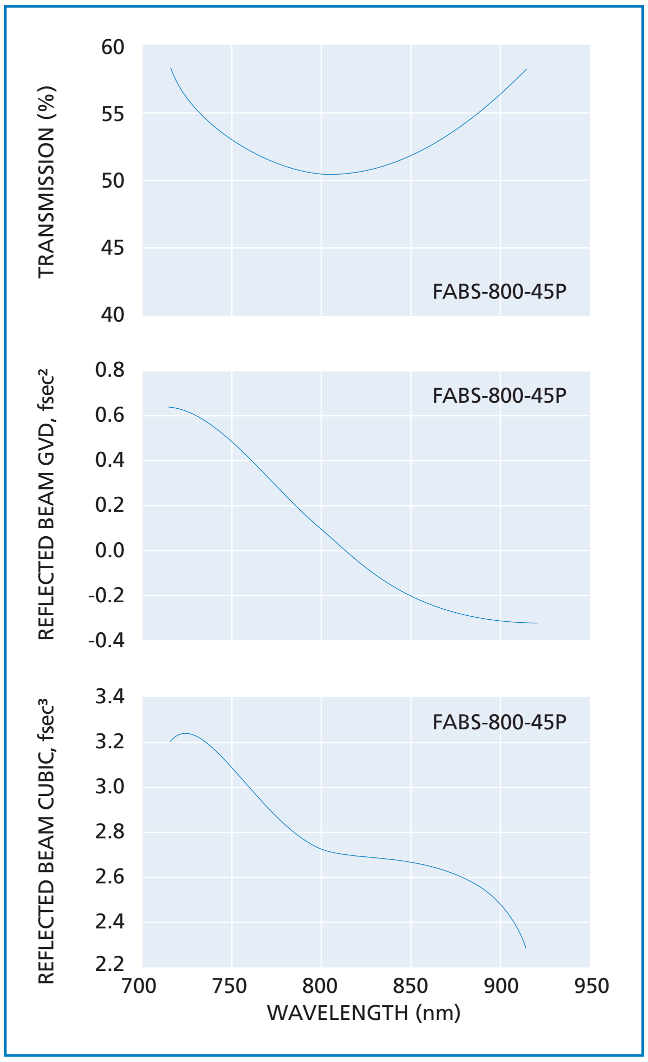

Figure 1.68 **Transmission characteristics for FABS series polarizers with *p*-polarized light**

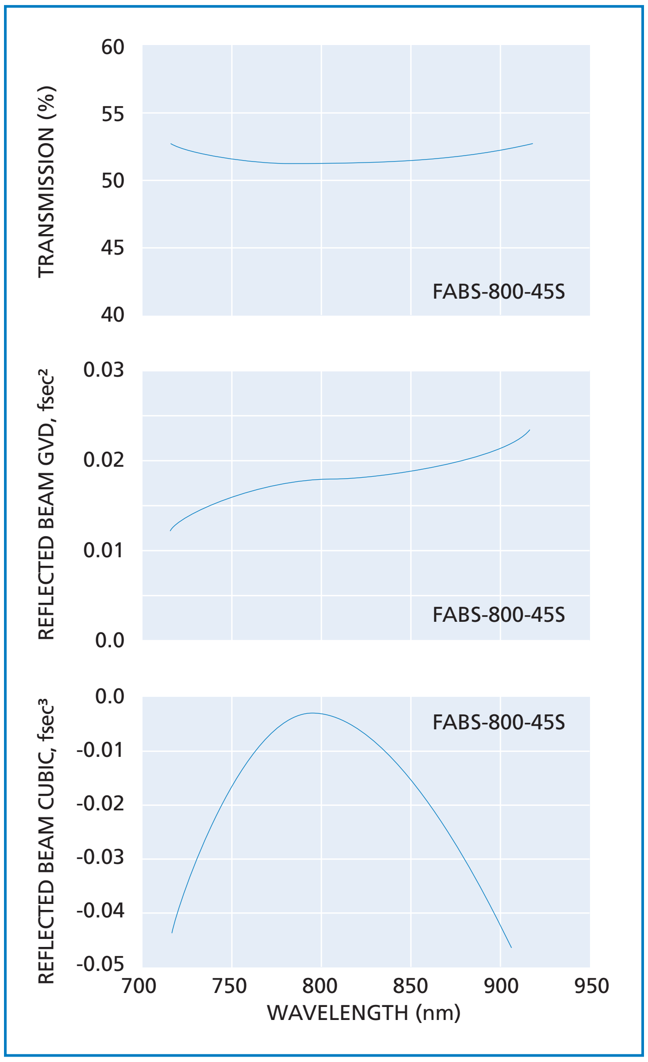

Figure 1.69 **Transmission characteristics for FABS series polarizers with *s*-polarized light**

ANTIREFLECTION COATINGS

All CVI Melles Griot antireflection coating designs work well in femtosecond operation as the forward-going phasor is the dominant contribution to the phase shift; the AR coating is very thin and simply “fixes” the small Fresnel reflection of the substrate.

PRISMS

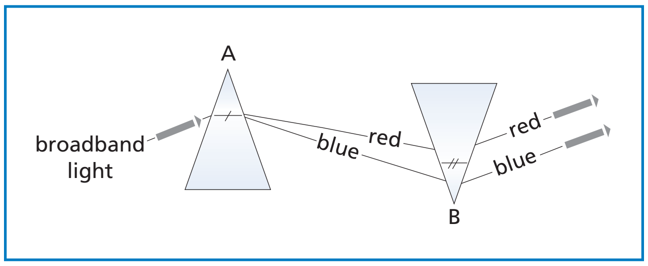

Figure 1.70 **Brewster prism**

Very-high-quality isosceles Brewster’s angle prisms for intra and extracavity use are available from CVI Melles Griot. The design of these prisms satisfies the condition of minimum loss due to entrance and exit at Brewster’s angle. To calculate GVD at Brewsters angle, refer to Figure 1.70 and use the following equation:

$$\mathrm{GVD} = \frac{d^2}{d\omega^2}\,\psi\bigg|_{\omega-\omega_l} \approx \frac{\lambda_l^3}{2\pi c^2}\left[\mathrm{L}\,\frac{d^2\eta}{d\lambda^2}\bigg|_{\lambda_l} - 4l\left(\frac{d\eta}{d\lambda}\bigg|_{\lambda_l}\right)^2 \right] \qquad (1.130)$$

where

η = refractive index of the prisms (assuming the same material)
l = tip to tip distance ($\overline{AB}$)
L = total avg. glass path
ψ = spectral phase of the electric field
$\omega_l \lambda_l = 2\pi_c$ (assumes Brewster prism at minimum deviation).

For more on the Ultrafast phenomena, see J.C. Diels and W. Rudolph, Ultrashort Laser Pulse Phenomena, Academic Press, 1996.

CVI MELLES GRIOT. YOUR SOURCE FOR

BUILD-YOUR-OWN, RAPID DELIVERY SOLUTIONS.

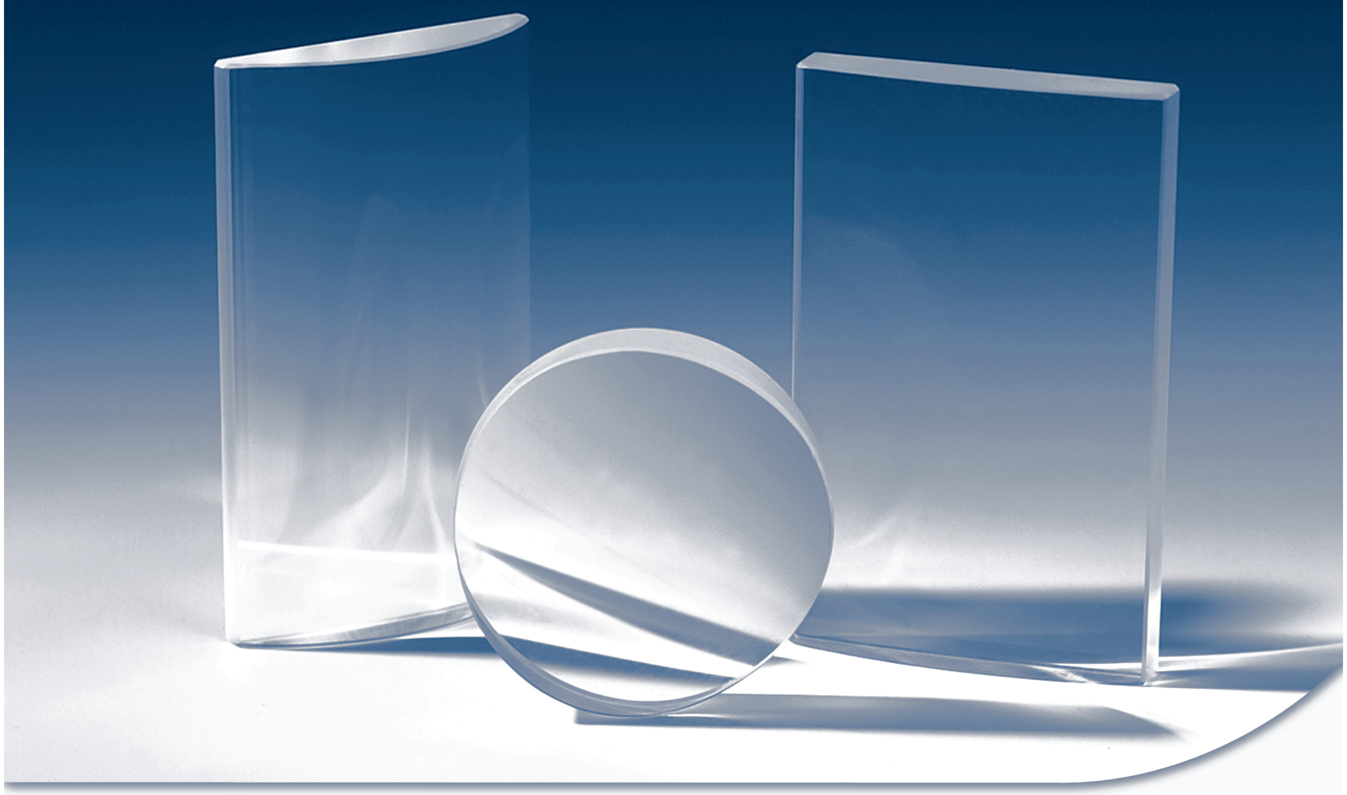

CVI Melles Griot offers two-week delivery of optics at catalog prices.

- 4000 uncoated substrates
- 2500 coating options
- A wide variety of in-stock substrates available (curves, flats, prisms, waveplates)
- Coatings from 193 nm to 2300 nm
- Your order ships in two weeks or less

How can we help you make your project a success?

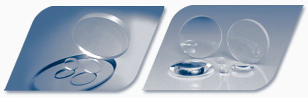

Lasers | Lenses | Mirrors | Assemblies | Windows | Shutters | Waveplates | Mounts

A GUIDE TO Understanding LAND SURVEYS

THIRD EDITION

Stephen V. Estopinal

A GUIDE TO UNDERSTANDING LAND SURVEYS

A GUIDE TO UNDERSTANDING LAND SURVEYS

Third Edition

STEPHEN V. ESTOPINAL

WILEY

JOHN WILEY & SONS, INC.

This book is printed on acid-free paper. ♾

Copyright © 2009 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: While the publisher and the author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor the author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information about our other products and services, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Estopinal, Stephen V. (Stephen Vincent)
A guide to understanding land surveys / Stephen V. Estopinal. – 3rd ed.
p. cm.
Includes index.
ISBN 978-0-470-23058-9 (cloth : alk. paper)
1. Surveying–Handbooks, manuals, etc. I. Title.
TA551.E88 2009
526.9–dc22
2008038625

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For my father, Eugene

CONTENTS

PREFACE xv

1 PURPOSE OF THE BOOK 1

2 REAL PROPERTY ON A ROUND PLANET 7

2.1. Ownership / 7
2.2. Title / 8
2.3. Boundaries / 8
2.4. Deed / 9
2.5. Identifying Boundaries / 9
  2.5.1. Geography / 9
  2.5.2. Latitude / 10
  2.5.3. Longitude / 12
  2.5.4. Astronomic Position / 14
  2.5.5. Geodetic Position / 14
  2.5.6. Flat-Earthers / 15
2.6. Corners / 17

3 GEOMETRY 19

3.1. Plane Geometry / 19
3.2. Land Point / 21
3.3. Land Line / 22

3.4. Straight Land Line / 23
3.5. Plumb Line / 24
3.6. Level / 24
3.7. Land Distances / 24
3.8. Elevation / 26
3.9. Land Area / 27
3.10. Horizontal Angles / 27
3.11. Vertical Angles / 29
3.12. Degrees, Minutes, and Seconds / 29
3.13. Maps or Plats / 29

4 DEFINING NORTH 33

4.1. True North / 33
4.2. Astronomic North / 34
4.3. Magnetic North / 34
4.4. Assumed North / 35
4.5. Grid North / 35
4.6. Directions / 35
4.7. Azimuths / 37
4.8. Bearings / 37

5 PROJECTION SYSTEMS 43

5.1. Projectionless Maps / 43
5.2. Tangent Plane Projections / 44
5.3. Advantages of Tangent Plane Systems / 46
5.4. Disadvantages of Tangent Plane Systems / 47
5.5. State Plane Projections / 48
5.6. Advantages of State Plane Projections / 51
5.7. Disadvantages of the State Plane Projection / 52

6 FUNDAMENTALS OF MEASUREMENTS 55

6.1. Accuracy and Precision / 56
6.2. Implied Precision / 56
6.3. Errors / 57
  6.3.1. Systematic Errors / 58
  6.3.2. Random Errors / 58
  6.3.3. Blunders / 58

6.4. All Measurements Include Errors / 58
6.5. Reduction of Errors / 59
6.6. Development of Standard Procedures / 60
6.7. Colonial Period / 62
6.8. Post–Civil War Period / 68
6.9. Beginning of the Modern Period / 69
  6.9.1. Compass Rule / 73
  6.9.2. Transit Rule / 77
  6.9.3. Crandall's Rule / 77
  6.9.4. Least Squares Adjustment / 77
6.10. Modern Period / 78
6.11. Random Traverse / 80
6.12. Elevations in the Modern Period / 84
6.13. Future of Surveying / 86

7 LAND RECORD SYSTEMS 89

7.1. Metes and Bounds System / 91
  7.1.1. Reading a Metes and Bounds Description / 95
7.2. U.S. Public Land System (USPLS) / 102
  7.2.1. Initial Point / 103
  7.2.2. Principal Meridian / 103
  7.2.3. Baseline / 104
  7.2.4. Township and Range Lines / 105
  7.2.5. Ideal Township / 105
  7.2.6. Ideal Section / 107
  7.2.7. Irregular Sections / 110
  7.2.8. USPLS General Instructions / 111
  7.2.9. Government Surveyor's Original Notes / 111
  7.2.10. Nonaliquot Division of a USPLS Section / 115
  7.2.11. Private Claims / 117
7.3. Platted Subdivision or Urban Systems / 118

8 THE GLOBAL POSITIONING SYSTEM (GPS) 123

8.1. Overview / 123
8.2. Fundamentals / 124
8.3. GPS and Land Surveys / 124
8.4. CORS Networks / 126

8.5. Practical Application / 127
8.6. Strengths / 128
8.7. Weaknesses / 129
8.8. Coping with Reality / 129

9 THE GEOGRAPHIC INFORMATION SYSTEM (GIS) REVOLUTION 131

9.1. Building a Foundation / 131
9.2. Sources of Information / 132
9.3. System Maintenance / 132
9.4. Potential Users / 133
9.5. Potential for Misapplication / 133

10 BOUNDARY RECOVERY 135

10.1. Rules of Evidence / 135
10.2. Sources of Evidence / 136
10.3. Office Procedure / 137
10.4. Field Procedure / 137
10.5. Rendering a Decision / 138
10.6. USPLS Boundary Recovery / 139
  10.6.1. Accessories to a Corner / 141
  10.6.2. Recovered Corners / 141
  10.6.3. Deciphering Government Land Office (GLO) Notes / 142
  10.6.4. Passing Calls / 142
  10.6.5. Perpetuated Corner / 143
  10.6.6. Corner by Common Report / 143
  10.6.7. Corner Reset by Best Evidence / 144
  10.6.8. Corner Reset by Proportionate Measurement / 144
10.7. Metes and Bounds Boundary Recovery / 144
10.8. Platted Subdivision Boundary Recovery / 145

11 EVALUATING SURVEY PLATS 147

11.1. Plat Evaluation Checklist / 147
11.2. Determining the Land Record System Used / 148
  11.2.1. Recognizing a Metes and Bounds Description / 149
  11.2.2. Recognizing a USPLS / 152
  11.2.3. Recognizing Platted Subdivisions / 152

11.3. Determining the Map Projection Used / 153
  - 11.3.1. Recognizing a Projectionless System / 153
  - 11.3.2. Recognizing a Tangent Plane Projection System / 154
  - 11.3.3. Recognizing a State Plane Projection System / 154

11.4. Evaluating the Age of the Survey / 155
11.5. Determining the Purpose of the Survey / 156
  - 11.5.1. Land Title Surveys / 156

11.6. Examining the Survey for Gross Discrepancies / 157
11.7. North Arrow / 157
11.8. Legal Description / 157
11.9. Date of the Survey / 158
11.10. Name of the Surveyor / 158
11.11. Signature and Seal / 159
11.12. Adjoining Properties / 159
11.13. Dimensions of All Sides / 160
11.14. Bearing or Angles / 160
11.15. Name of Client, Purpose of Survey / 161
11.16. Certification / 161
11.17. Limiting Words or Phrases / 162
11.18. Area / 162
11.19. Scale / 163
11.20. Comparing the Survey Plat with the Deed / 163
11.21. Examining the Survey Plat for Easements / 164
11.22. Examining the Survey Plat for Encroachments / 165
  - 11.22.1. Fences and Fence Lines / 165

11.23. Determining the Precision Required / 168
  - 11.23.1 “Class A. Urban Surveys” / 169
  - 11.23.2 “Class B. Suburban Surveys” / 169
  - 11.23.3 “Class C. Rural Surveys” / 170
  - 11.23.4 “Class D. Mountain and Marshland Surveys” / 170

11.24. Determining Needs Not Covered in a Survey Plat / 170
11.25. Contacting the Surveyor / 171

12 EXERCISES IN EVALUATING SURVEY PLATS 173

12.1. The Case of the Three Partners / 173
  - 12.1.1. Conclusion / 176

12.2. The Land Grabber / 181
12.3. Easement Surprise / 186
12.4. Excessive Problems / 189
12.5. Metes Meets Bounds / 192
12.6. The Square That Wasn't There / 196

13 WRITING LEGAL DESCRIPTIONS 203

13.1. Creating New Parcels / 204
13.2. Existing Parcels / 205
13.3. General Outline / 205
13.4. The Caption / 206
13.5. The Narrative / 207
13.6. Key Phrases / 208
    13.6.1 "Commencing At" / 208
    13.6.2 "Point-of-Beginning" / 208
    13.6.3 "Thence" and ";" / 209
    13.6.4 "In a —ly Direction" / 209
    13.6.5 "-Most" / 211
    13.6.6 "Along" / 211
    13.6.7 "A Distance Of" / 211
    13.6.8 "To" and "To a Point" / 213
    13.6.9 "That Same Point" / 214
    13.6.10 "Point of Curvature" and "Point of Tangency" / 214
    13.6.11 "From Whence Bears" and "Distant" / 216
    13.6.12 "Encompassing an Area Of" / 216
13.7. Deletions or Additions / 217
13.8. References / 218

14 EXERCISES IN WRITING DEED DESCRIPTIONS 219

14.1. Case 1 / 219
14.2. Case 2 / 223
14.3. Case 3 / 224

15 ETHICS AND PROFESSIONAL CONDUCT 227

15.1. Confidentiality / 228
15.2. Professional Courtesy / 228

15.3. Impartial Evaluation / 229
15.4. Promote Professionalism / 229

APPENDIX OF TABLES 231

GLOSSARY 237

INDEX 245

PREFACE

The first edition of *A Guide to Understanding Land Surveys* was published in 1989. The original work was intended to present vital information and insight to users of Property Boundary Surveys. In the intervening decades, advances in the science of measurement, computation, data management, and other disciplines have provided Land Surveyors with tools that are so powerful that they defy description.

Advances in science have resulted in the modern, well-prepared Boundary Survey of today. These Surveys and the Land Survey Plats they produce are far superior to that which was commonly presented in the late 20th century in many ways. The conceptions that the general public has about modern Land Surveys have not kept pace with the changes.

Misunderstanding and misconceptions about Land Surveys have been compounded by these great advances, not alleviated. The airways are flooded with advertisements touting simple and easy-to-use navigational devices. Anyone with an Internet connection can visit multiple web sites that provide aerial photography, maps, and a host of geographically based information. The general public has been given the impression that science has simplified surveying to merely navigating to the corners as if the locations of, and the distances between, all property corners were known.

Measurements of distances, directions, and location are but tools that the Land Surveyor uses. A Surveyor determines a precise location in order to *begin* the survey, not as an end in itself. Highly accurate measurements of angles and distances are used by the Surveyor to document boundary locations and assist in the recovery of property corners. Precise measurements form a foundation for the Surveyor's work, not an end.

This book continues to be a valuable source for land title attorneys, paralegals, realtors, government agents, and others who use the information gathered and presented by Land Surveyors. These professionals will find this work very useful in deriving the maximum benefit from the Survey Plats that they review in performing their duties.

Dispelling misconceptions and incorrect assumptions that exist in the general community will do much to reduce conflict and unnecessary litigation. In that regard, this book has great value. The reader will also be provided with an overview that will go beyond merely dodging pitfalls. The broad and general concepts presented in this book will also assist the reader in evaluating survey information and *communicating* his or her informational requirements to the Professional Land Surveyor.

Communication is perhaps the most important aspect of contact between users of Land Surveys and the Professional Land Surveyor. The Professional Land Surveyor will also find this book to be most useful in developing firm and clear relationships with their clients.

Burnside, Louisiana
May 2008

STEPHEN V. ESTOPINAL

A GUIDE TO UNDERSTANDING LAND SURVEYS

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 1

PURPOSE OF THE BOOK

This book was developed to meet the needs of attorneys, abstractors, realtors, land planners, entrepreneurs, and others who must use and evaluate the work of professional surveyors. Boundary survey plats or maps, property maps, topographical maps, survey reports, and property descriptions, among other things, are all products of the professional surveyor. Rapid increases in technology, data availability, land values, community/regional planning, and the modern propensity for litigation have meant that more and more nontechnical individuals have begun to use and rely on increasingly complex and technical land information.

The most significant of the modern advances in Land Information Systems (LISs) or Geographical Information Systems (GISs) has been the Internet. Virtually any tidbit of information that can be associated with a geographic position can now be made available to anyone with an Internet connection. Anybody and everybody who desires to explore land uses, owners, zoning, regulations, conditions, and so on can now have that data at their fingertips. Municipal, state, and federal archives; agencies; and other organizations maintain web sites where the visitor is free to explore databases without restriction or *professional assistance*. Without a clear understanding of the source, reliability, and *intended use* of that data, an unsophisticated user can cause more harm than good. Indeed, there is a real danger that the ease with which this information is obtained will facilitate misapplication of that data while at the same time circumventing professional assistance.

As an example of the sort of confusion that can result from misuse of the sophisticated GIS data banks available today, an electronic database may be compiled for the intended purpose of identifying properties for appraisal and taxation. This data would typically be presented on the Internet in the form of a map developed from aerial photography. The boundaries on that map were created by technicians who reconciled fence lines and other visible features with whatever deed documentation was available. A user might access that map and, because the computer allows it, obtain distances, directions, and quantities of the real property parcels as presented in that graphic representation. While presented by the computer as precise, the distances, directions, and acreage obtained in this example would be far from accurate.

The location, configuration, and description of real property parcels are just some of the multitude of important geographic questions people typically have. The recovery, documentation, and interpretation of this data are the professional specialties of the Land Surveyor. Once a person realizes the necessity of obtaining the services of a professional surveyor, they may have difficulty in communicating to the surveyor just what it is they need. From attorneys to private citizens, the word *survey* has many different meanings. For example, a determination of boundary location for the purpose of erecting a fence is a "survey." The determination of the elevation of a building for the purposes of obtaining flood insurance is a "survey." The recovery or remonumentation of the boundaries of a parcel of land for the purpose of an exchange of title is a "survey." Each of these examples is called a "survey," yet the scope of work, the responsibilities of the surveyor, and the *expectations* of the persons ordering the "survey" are all very different.

A better basic understanding of just what a surveyor does and does not do can make the difference between complete service and unsatisfactory results. Much litigation, confusion, and aggravation can be traced to a misunderstanding between the surveyor and the client. It is essential that the surveyor be informed of the exact purpose of the survey being requested. The completeness of the surveyor's report, the extent of his or her research, the areas examined, and other pertinent work vary greatly, depending on the type of survey that is being conducted.

Compounding the problem is the fact that the person requesting the services of a Professional Land Surveyor may not be the person using and interpreting the results of the work. This "third-party" use of a surveyor's maps or reports is rife with dangers, not the least of which happens when the user assumes that the surveyor did what the user wanted done when in reality the surveyor carried out the client's orders. For example, an owner of a vacant urban lot may request that a surveyor recover only the boundaries of that lot. Easements, servitudes, building restrictions, and other important title restrictions that impact on the enjoyment of that lot do not affect the boundaries of that lot. The surveyor would then recover or mark the boundaries and issue a report in the form of a drawing to the owner, showing the dimensions of the lot and the locations of the boundary markers. If, at a later date, the owner were to sell the lot, he or she could present this drawing to a buyer. The buyer, having no knowledge of the limited request of the previous owner, might then rely on the drawing produced as if it reflected all of the information about easements, servitudes, buildings, and other facts important to the buyer. The resulting lawsuit would charge that the surveyor neglected to perform the extensive work that the buyer needed when he or she complied with the limited request of the owner.

The grief, aggravation, and expense resulting from the previous scenario could easily be avoided if the buyer or the buyer's attorney confirmed that the drawing presented to them contained all of the information that they wanted to know. Third-party situations like this one have resulted in so many lawsuits that many surveyors now place explanations of the extent of the work done by them in producing the survey plat directly on the plat. Some state registration boards have even adopted a regulation or "standards of practice" that require such statements.

The Land Surveyor is an investigator—a detective more than anything else. The subject of the investigation is the location of the boundaries of real property. In the pursuit of the evidence necessary to determine, with relative certainty, the original location of a particular boundary line, the same rules of evidence apply as in any civil court. Surveyors not only measure angles and distances, but also perform extensive records research in private and public files in an effort to reveal as much information as possible about the *location of boundaries*.

This extensive research does not normally include many other factors involved in property rights. Title insurance is a different service from a boundary survey, although most title insurance policies include the requirement that a boundary survey be performed. Easements, servitudes, building restrictions, setbacks, or side clearances for new construction, flood zones, and regulatory zoning, among other things that limit the use and enjoyment of real property, are not essential to the recovery of boundaries. If you wish to have all, or any, of these items shown on a survey plat, you must tell your surveyor.

The surveyor will report evidence of these factors limiting the free enjoyment of a real property parcel whenever it is discovered in the normal course of boundary recovery. If an original subdivision plat includes an easement, the surveyor will normally report it. If an easement is granted separately from the subdivision recordation, the surveyor may not have the occasion to discover it. The records search by a Land Surveyor takes a much different route than that of a title examiner. Provide the surveyor with the complete title record if you wish it to be shown on the survey plat. Otherwise, a boundary survey will result in the recovery and a report (plat) on the real property limits of a particular parcel *and nothing else*.

This book is intended to bridge the ever-widening gap between the users of land boundary information and the producers of that information—professional Land Surveyors. To that end, the information in this book will be only as technical as necessary to convey general concepts and will be presented in such a way that every reader, from one experienced in the use of land boundary information to the novice, will gain a better understanding of the profession of land surveying and the products of that profession. The early chapters of this book provide a very important background to understanding real property surveys because they deal with the root causes of confusion and misunderstandings concerning boundary surveys and the survey plats or maps that are produced as a result of surveys.

Real property laws, traditions, and practices vary throughout the United States. This book will not attempt to address the specific details of every state but will describe in a general setting the common facts found throughout the United States. The reader is advised to determine how the particulars of real property laws and the principles and practices of boundary surveying in his or her area vary from the general concepts addressed in this book. Most local surveyors or professional surveying organizations are very happy to assist anyone in familiarizing themselves with local practices.

A large portion of this book is devoted to the writing, reading, and interpretation of legal descriptions of real property parcels. That portion will be devoted to the mechanics of legal descriptions and will not presume to advise the user on the appropriate form of legal description to use for a particular situation but, rather, will tutor the reader on the geometric and semantic aspects of describing real property parcels. The great differences between those regions of the United States known as “metes and bounds” areas and United States Public Land Survey (USPLS) areas will be explored in depth.

This book will serve both as a handy, daily reference guide and as a tutorial text. The reader may want to refer to portions of this book when ordering survey work or when evaluating survey plats, legal descriptions, or reports already received. Although this book will improve the reader's understanding of land surveys, it will not make the reader a Land Surveyor. Many highly technical aspects of measurements, computations, details, and equipment are only lightly referred to, if at all. Sophisticated research and recovery techniques, so vital to the recovery of boundaries, would make a book in itself. Your Professional Surveyor is the expert you must rely upon for these skills.

This book will not serve as a substitute for qualified legal assistance in the interpretation of real property rights or transfers. The survey plats and problems used as examples in this book are highly simplified versions of their more complicated actual counterparts. The laws concerning boundaries and real property rights differ from state to state and change over time. Some general rules are mentioned, but these are not to be interpreted as universal laws. The attorney is the professional who should be relied upon for the status of real property law in a particular area. The reader should make note of such regional variations as may exist when he or she is using this book.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 2

REAL PROPERTY ON A ROUND PLANET

2.1. OWNERSHIP

Land cannot be owned, at least not in the same sense as physical objects are owned. People cannot carry their land with them to new locations. Land cannot be manufactured or destroyed. If one were to excavate a great hole in a property parcel, the land might be rendered useless, but the real property parcel would still exist. "Reclamation" of submerged land is not manufacturing land but simply altering the surface condition of an existing area. The "sale" of a real property parcel does not result in a giant cookie-cutter stamping out of a hunk of land to be turned over to the purchaser. One cannot tour the country collecting scraps of land and assemble a ranch in California.

Real property ownership is not the ownership of land; it is the designation as the authority to control an area. It is the ownership of "rights." The right to build upon, improve, inhabit, cultivate, deny access to others, and countless other rights are what is owned.

Mankind is not the only species to demand authority or control over a land area. Territorial creatures, great and small, impose their authority over their range by force. The dawn of civilization brought with it the concept that people could associate themselves with a group larger than a family or tribe and collectively defend their territorial authority. City-states and then national states were formed around the concept that individuals within a state would contribute to the protection of the state in return for the "right" to inhabit, cultivate, use, and enjoy the land.

All of these rights are intangibles and flow from the “sovereign” or governing authority. The laws enacted by national governments determine the extent of personal real property rights. It is no coincidence that national governments that maximize individual real property rights over collective control generally prosper and those that restrict individual rights in favor of central control do not.

Peoples who have found themselves to be without a defending force are routinely forced from the land by those who do. It is the armed forces of a national government that establishes sovereignty and defends the laws that dictate which rights may be acquired by individuals in a certain area.

2.2. TITLE

The Constitution, laws, and statutes of the United States of America and the individual states allow the assignment of “real property” rights to private persons. Those persons entitled, under the law, to exercise the rights of real property ownership are said to “hold title” to a land parcel. It is these rights that are transferred in the sale of “land.” Persons may gain “title” to real property through written and unwritten transfers. However, only recorded, written title transfers are merchantable.

The collection of papers, recordings, and other documents that describe a transfer of the titled rights to a real property parcel are often referred to as the *title*. In fact, such documents are only evidence to the lawful assignment of title. Title, the collection of real property rights, is a concept in law and therefore incorporeal.

2.3. BOUNDARIES

Real property boundaries are the invisible limits of the intangible rights that constitute real property ownership. Corporeal possessions have clear limits. Unless there has been a terrible accident, a person has no difficulty discerning where her or his car ends and another’s begins. A person can exit his or her car without entering another’s.

This is not the case with real property. The limits of the rights of ownership cannot be seen. Physical evidence as to the location of that limit may be obvious, but the limit itself is an invisible and dimensionless concept. One must be on one side of a property boundary or the other. As soon as one crosses the limit of one set of rights, another set of rights is entered.

2.4. DEED

A deed is the written instrument that describes the assignment of the rights that are "owned" and the geographical limits of that assignment. A deed is not the same as a title, although the words are often used interchangeably. Although the limits of the rights of real property ownership are invisible, they must be identifiable. Every deed must communicate what rights are owned and where the boundaries or limits of those rights are. It is folly to attempt to transfer rights in real property without adequately identifying and communicating the location of the limits of the rights conveyed.

The Professional Land Surveyor is charged by society with the responsibility to recover, monument, quantify, and document the imperceptible limits of the intangible rights that constitute real property ownership.

2.5. IDENTIFYING BOUNDARIES

There has always existed the need to identify and to communicate locations on the surface of the earth. Stone Age men and women could relate to localities only by naming mountains, rivers, or other features of the terrain and then specifying a locality as being near a certain feature. This first use of landmarks required that the receiver of the information be as knowledgeable about the terrain as the giver.

Artificial landmarks were, and are, also used to identify boundaries. Territorial animals, including humans, mark their territories. Although this was adequate for identifying locations to a person who was at the boundary or familiar with the territory, the advent of long-range travel, especially sea travel, required a method of identifying locations that did not rely exclusively on terrain features. The development of the concept of the assignment of real property rights by law, instead of physical occupation, required a method of identifying locations in a way that could be written, recorded, and communicated between parties, even people who are in other locations.

2.5.1. Geography

Geography and real property boundaries both require the communication of locations, distances, and directions. The glaring difference between the two is that geography can be directly measured, but real property boundaries can only be defined.

Although the question of defining real property boundaries was adequately addressed early on through the use of artificial marks and other references to physical features, geography and navigation required advances in math and science that has taken centuries to perfect. Real property boundary descriptions have, generally and often imperfectly, followed the improvements in geography and navigation. Slowly, both are beginning to solidify behind the same scientific foundation. The history and application of that foundation is fundamental to understanding Land Surveys.

As the earth spins in orbit around the sun, its motion can easily be detected by any observer using no other instrument than the human eye. Unlike the moon or the sun, the stars clearly reveal the constant, circular motion of the planet. It was this motion that led to the development of today's worldwide system of identifying points on the surface of the earth.

In spite of the fact that some of the ancient observers may have even believed that the world was flat, the direction to the center of the apparent circular celestial motion was the very first geographical standard reference adopted by those attempting to describe real property boundaries. That standard direction of reference is called "north." This standard has survived for centuries and remains in use to this day, simply because it can be easily and precisely determined anywhere on earth.

2.5.2. Latitude

Ancient seafaring societies noticed that the motion of the stars was not only circular but also regular and predictable. One of the earliest observations was that particular stars were known to pass directly over the same place every night. Many societies came to identify stars by the earth locations or "latitudes" that the stars would seem to track. The Polynesians, to this day, identify certain stars by which island or island group those stars pass directly over. The reverse is also true: a certain location can be identified by the star that passes overhead.

In reality, the apparent motion of the stars is the result of the spinning of the earth. The apparent center of rotation that the ancients observed is an indication of this axis of rotation. "North" is so universally detectable that it was inevitable that the system of identifying locations that developed used this axis as the reference base. Figure 2.1 shows how the concept of "latitude" first developed. Note that the angle measured to determine the latitude of the observer's location is actually the angle between a horizontal line tangent to the level surface of the earth

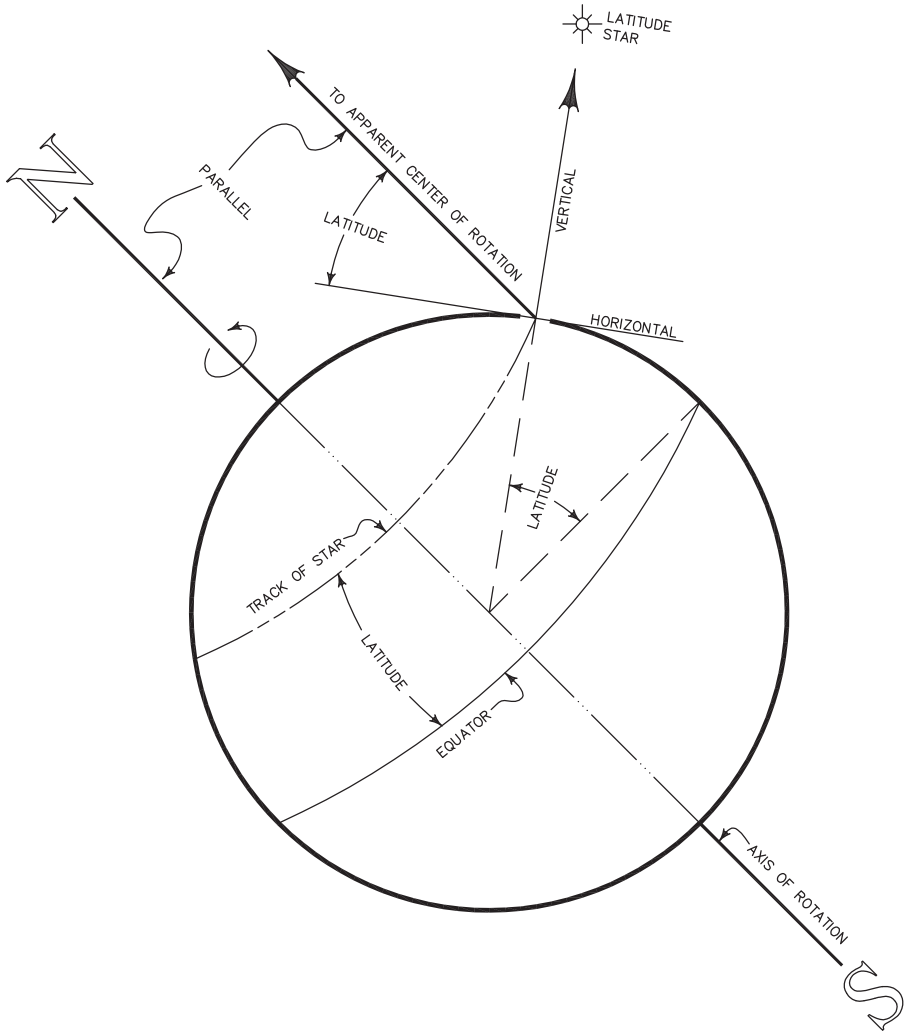
LATITUDE STAR
TO APPARENT CENTER OF ROTATION
VERTICAL
N
PARALLEL
LATITUDE
HORIZONTAL
LATITUDE
TRACK OF STAR
LATITUDE
EQUATOR
AXIS OF ROTATION
S

FIGURE 2.1

and the axis of rotation. Points on the surface at which the same angle or latitude is observed can be connected to form a curved line circling the earth.

These connections of points on the surface of the earth of equal latitude are called “latitude lines.” By definition, latitude lines are east–west lines, and, as shall be shown later, of all the latitude lines, only the equator (zero degrees latitude) is a “straight” line. The latitude of a particular point can be used as a measure of the distance north or south of the equator and is so easily measured that early explorers often sailed north or south to the latitude of their destination and then east or west to reach shore. This convenient measure of north–south distances often caused early maps to be quite accurate in reporting latitudes but woefully inaccurate in reporting east–west dimensions.

2.5.3. Longitude

The axis of the earth's rotation is, therefore, a naturally occurring reference that can define directions on the surface of the earth, as well as distances in a north–south direction. In order to identify a particular point along a line of latitude, the concept of *longitude* (shown in Figure 2.2) was developed.

Just as latitudes are a contiguous series of points with a common factor (the measure of the angle between the earth's axis and the horizon), longitudes are a series of contiguous points with the common factor of being aligned along the apparent axis of rotation. These north–south lines are sometimes referred to as "meridians" when they are used to define other lines or directions. Unlike latitudes, the reference for longitudes had to be arbitrarily chosen.

Throughout history many locations have been used as the origin of latitudes depending upon the map maker's preferences. Today, most of the world has accepted that zero degrees longitude is defined as passing through a particular point in Greenwich, England, and is called the "prime meridian." Longitudes are used to define or to measure distances in an east–west direction. The combination of a north–south distance and an east–west distance can be used to define a particular point on the surface of the earth. Theoretically, only one pair of latitude and longitude values can be associated with a specific surface point.

Although longitudes are reported in degrees, just as are latitudes, time is the factor actually measured by observers in determining the longitude of a particular location. In order to accomplish this measurement, an observer might note, for instance, that a particular star is directly over the prime meridian at a certain time and that that same star is directly over the longitude of the observer some time after that. The elapsed time is directly related to the distance from the prime meridian to the longitude of the observer. The earth completes approximately one rotation (360 degrees) in 24 hours. Therefore, the ratio of elapsed time to 24 hours is the same as the ratio of degrees longitude to 360 degrees. In actual practice, the stars observed, angles measured, and elapsed times noted are never as direct as this simple example.

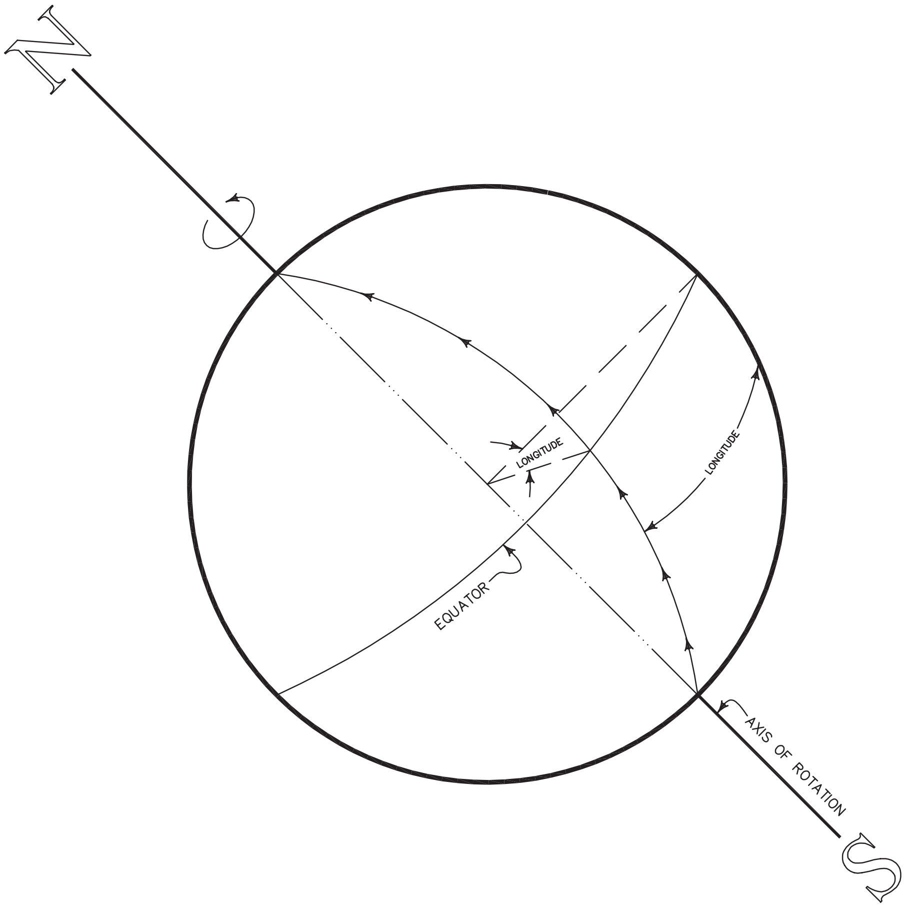

FIGURE 2.2

Latitudes and longitudes are reported as if they were angles measured at a point at the center of the earth, when, in fact, the measurements are of time and of angles and, until the advent of the earth-orbiting artificial satellite, made at the surface of the earth. This would not cause a major distortion if the earth were a sphere with no irregularities of surface or gravity. But precisely because the earth is not a sphere and because the observations are made on the surface of the earth, the shape of the earth and the distortions caused by the irregularities of the surface must be understood in detail. The impact of space-age technology will be discussed at length later.

2.5.4. Astronomic Position

With the first attempt to account for the irregularities of the earth's surface, a method was developed by which all measurements or observations were adjusted to account for elevation above sea level. This method developed a theoretical surface that covered the earth and that was everywhere perpendicular to the pull of gravity. The theory behind this assumed that, if the entire planet were covered with water, the planet would be a perfect sphere. The surface that resulted from this "reduction to sea level" is called a "geoid." The latitude and longitude based on celestial observations used to identify a specific location are called that location's "astronomic position." Astronomic positions are computed based on the geoid surface. Unfortunately, a geoid is not a sphere, nor is it a regular surface.

Gradually, the realization developed that the anomalies of the gravitational forces around the world produced a geoid that was "dimpled" and undulating. Astronomic positions are, necessarily, based on the pull of gravity. Astronomic latitudes and longitudes are not, therefore, perfectly circular curves or perfectly straight lines on the geoid surface. The positions reported by astronomic observations are distorted by the irregularities of the geoid.

Astronomic positions were quite adequate until the methods and precision of measurement became so refined that the irregularities of the earth's shape became detectable. Astronomic positions lack the precise predictability necessary for accurate navigation and mapping. The science of geodesy, the study of the size and shape of the planet Earth, and the science of cartography, the study of mapmaking, have developed methods to reduce the uncertainty inherent in astronomic positioning.

2.5.5. Geodetic Position

The need to compute precise distances between locations of known positions or to compute the position of a location based on distance and direction from a known position requires a regular, mathematically identifiable surface on which computations can be made. The geoid can be approximated by a theoretical shape or model based on regular geometrical formulae. The mathematical shape or model chosen for this approximation is called a "spheroid" or "ellipsoid" and forms a theoretical surface on which the values of latitude and longitude are computed for any earth location.

The latitude and longitude of a location based on an ellipsoid is called the "geodetic position." Imagine that the surface of the earth is transparent and that the surfaces known as the ellipsoid and the geoid are glass bulbs around the center of the earth. A man standing on a high mountain could shine a light straight down. Where that light hit the geoid would be the astronomic position of the man. Where that light hit the ellipsoid would be his geodetic position.

The geodetic position is what is reported for navigational and mapping purposes. For most localities and for many purposes, the differences between astronomic and geodetic positions are not significant. The ellipsoid used in a particular map or in a particular country is usually defined by national or regional statutes. Improvements in understanding the shape and size of the earth has caused the scientific community to redefine the ellipsoid used to create geodetic positions. The most recent change of any significance has resulted in the published geodetic positions of fixed earth locations that in some areas were 30 or more meters different from the old system. The standards for the United States of America and each of the member states will be discussed at length later in this book.[1]

Locations are reported by a longitude west or east of the prime meridian and a latitude north or south of the equator. Longitudes are always less than or equal to 180 degrees. Latitudes are always less than or equal to 90 degrees. The values reported are the apparent distances along the surface of the geoid, in the case of astronomic positions, or the ellipsoid, in the case of geodetic positions.

2.5.6. Flat-Earthers

The spherical reference systems of astronomic and geodetic latitude and longitude are perfectly applicable to defining locations on a planet or, more precisely, on the surface of a spheroid, as shown in Figure 2.3. However, the basic instincts of the human mind and the limited input of our senses tell us that the world is flat. Because the earth is so large compared to the range of our senses, we conceive of the surface of the earth as being two-dimensional. This concept of a two-dimensional surface is reinforced by our use of maps and our insistence on using two-dimensional geometric terms that describe features on the map when we are referring to features that are on the surface of the earth.

Consider the following trip: A woman on a Montana ranch begins a trip by walking due north for exactly one mile, then turns right and walks exactly one mile due east, turns right and walks exactly one mile due south, and then turns right and walks exactly one mile due west

---

[1] See section 5.6: Advantages of State Plane Projections.

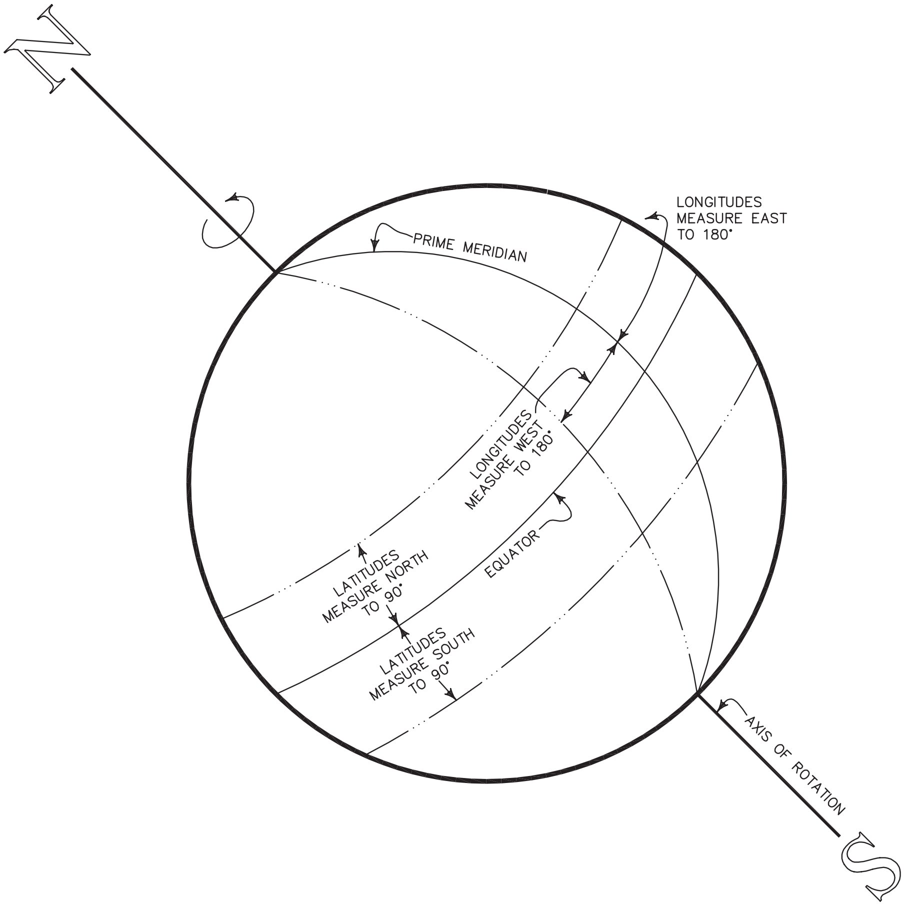
N
LONGITUDES MEASURE EAST TO 180°
PRIME MERIDIAN
LONGITUDES MEASURE WEST TO 180°
EQUATOR
LATITUDES MEASURE NORTH TO 90°
LATITUDES MEASURE SOUTH TO 90°
AXIS OF ROTATION
S

FIGURE 2.3

and stops. The question is, “Where is the woman in relation to her starting point?” If you say that she has returned to her starting point, then welcome to the flat-earth society!

Even after sketching such a trip out on paper, our instincts blind us to the fact that the trip took place on the curved surface of the earth and not a flat surface. The sketch on a sheet of paper is a two-dimensional representation of a three-dimensional event. In spite of discoveries, observations, knowledge, and measurements to the contrary, each of us persists in thinking of the surface of the earth as flat. This “flat-earth” mentality must be overcome in order to begin to understand land surveys and the maps or plats that represent these surveys.

All of this would be academic to most of us if it were not for the concept of ownership. The law recognizes the fact that persons can claim, as personal property, certain rights to portions of the planet. Real property law is that body of law devoted to defining the limits of these "rights of ownership" over portions of the earth. Real property boundaries are imaginary limits intended to define the extent and dimensions of such ownership. This ownership is not limited to the surface of an imaginary spheroid, nor is it limited to the surface of the earth.

The boundaries of real property rights or ownership are imaginary limits created by law in an attempt to control the actual, corporeal use of the land. The professional land surveyor serves as the bridge between the physical limits to the exercise of real property rights on the land and the imaginary concepts of ownership boundaries. The great value often attached to continued enjoyment of real property rights demands *permanence* and *recoverability* of the physical limits of real property boundaries. The boundaries cannot shift with every improvement in the science of geographic measurements.

It is people who must use the land. People are accustomed to straight lines, regular shapes, and consistency in the configuration of real property parcels. Geodetic positions can be defined, but geodetic lines of constant value do not follow straight lines of sight. Real property limits need to be defined and described in a way that is consistent with the flat-earth mind-set of people.

2.6. CORNERS

It is possible to mark a parcel that has straight land line segments on all sides. It is also possible for each of the four sides to be one mile in length. Instead of defining the boundaries of the parcel by marking each step of a walking man, the end points of each side could be marked and the boundary could be defined as a straight land line *between the end land points or corners.*

The desire for regular shapes, parallel lines, and constant directions for straight lines is resolved when the boundaries are so defined. For this reason, as well as others to be discussed later, the vast majority of boundaries in the United States are defined by the location of corners.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 3

GEOMETRY

In order to better understand exactly what it is that Professional Land Surveyors are presenting in their plats, maps, and reports, a closer look at the somewhat esoteric concepts of geometry and geodesy is necessary. Some of the words used in describing boundaries are common geometric terms and have generally accepted definitions that are *very different* from what is meant when these terms are used in real property situations. Geometry is consistent in that the terms used can describe two- or three-dimensional concepts. Maps are two-dimensional drawings, and the geometric terms describing map features are consistently confused with the three-dimensional surface features that the maps represent.

3.1. PLANE GEOMETRY

Consider these basic definitions of geometric terms:

Space—An extent or continuum in which objects can exist and have relative positions and directions.

Point—A specific, dimensionless location in space having a unique position.

Line—A series of contiguous points such that any point in the line is directly adjacent to two and only two other points in the line.

Straight line—Any line in which the route of measure of the shortest distance between any two points on the line lies entirely within the line.

Ray—A portion of a straight line originating or emanating from a point.

Line segment—A portion of a line beginning at one point and ending at another.

Distance—A measure of the separation of two points.

Length—The sum of all of the distances between consecutive points of a line segment.

Surface—Any array of contiguous points, having no thickness, that completely separates two distinct spaces.

Plane—A surface such that the shortest route between any two points on the surface lies entirely within the surface.

Area—The measure of a bounded surface.

Volume—The measure of a bounded space.

Angle—A measure of the relationship of two lines, rays, or line segments.

If one were to restrict the concepts of geometry to the two dimensions of a flat plane, then the three-dimensional terms of *space* and *volume* would be eliminated. The concept of *lines* also changes subtly. Straight lines in a two-dimensional system either intersect or are parallel. The concept of *angles* is rendered much simpler, and the entire system can be modeled quite easily by sketching on paper. There are also one-dimensional systems that are frequently used but rarely recognized as such. One particular one-dimensional system, the deed description, will be discussed at length in Chapter 13.

It is important to keep in mind that, in spite of the fact that one- and two-dimensional systems exist, our world is three-dimensional; therefore, whenever real property is located, measured, defined, or exchanged, a three-dimensional portion of our planet is involved. This fact requires that the geometric definitions already given be redefined to reflect what is really meant when these terms are used in describing real property. In order to assist the reader, when terms in this chapter are used that may confuse the real property definitions and the geometric definitions, the use of the adjective "land" will emphasize the real property definition—for example, "land line," "land boundary line," "land point," and so on when the real property definition is intended. In everyday practice, the adjective *land* is not used.

Anyone who has dealt with real property boundaries, title transfers, or surveys may have noticed that some of the words defined earlier in this chapter concerning dimensional geometry are often used in describing real property. These geometric terms, when used in real property descriptions, are misused in a curious mixing of two-dimensional and three-dimensional concepts. It is very important to understand how these words are misused, or, more accurately, redefined in real property descriptions before an appreciation of just what a boundary survey really is can be reached. The distinctions of what is conveyed, what is physically marked, what is drawn on a survey plat, and what is described in legal instruments are vitally important.

3.2. LAND POINT

An easy definition of a *land point* as used in real property descriptions would be something similar to "a specific location on the surface of the earth." However, it is much more than that. For instance, if a property corner were located on the side of a hill and the hill were removed by earthwork, the property corner or land point would remain. Its location would not be thought of as having "moved," although the surface of the earth would now be several feet lower than the original surface. If, however, a property corner were located in a swamp or a low spot and the land were later filled or leveled, the location of the property corner would not be altered as far as the real property definition of that parcel was concerned. Changes in elevation do not alter the locations of land points.

Figure 3.1 illustrates a real property land point that was marked (monumented) on the earth's surface in 1802. Years of erosion and earthwork have destroyed the monument and have lowered the surface several feet. In 1980, the land point was again monumented at a different location in space, several feet nearer the center of the earth, yet the location of the land point did not change as it is defined in land surveying terms.

What is called a "point" in real property terms is, in geometric terms, a ray emanating from the center of the earth and extending through the surface of the earth, where it is usually marked or defined, and beyond the earth as far as the legal concepts of "air space" and "reasonable control" extend. Real property descriptions call this ray a "point" because humankind's flat-earth instincts and, more importantly, the surveyor's plat depict the real property parcel as being a plane.

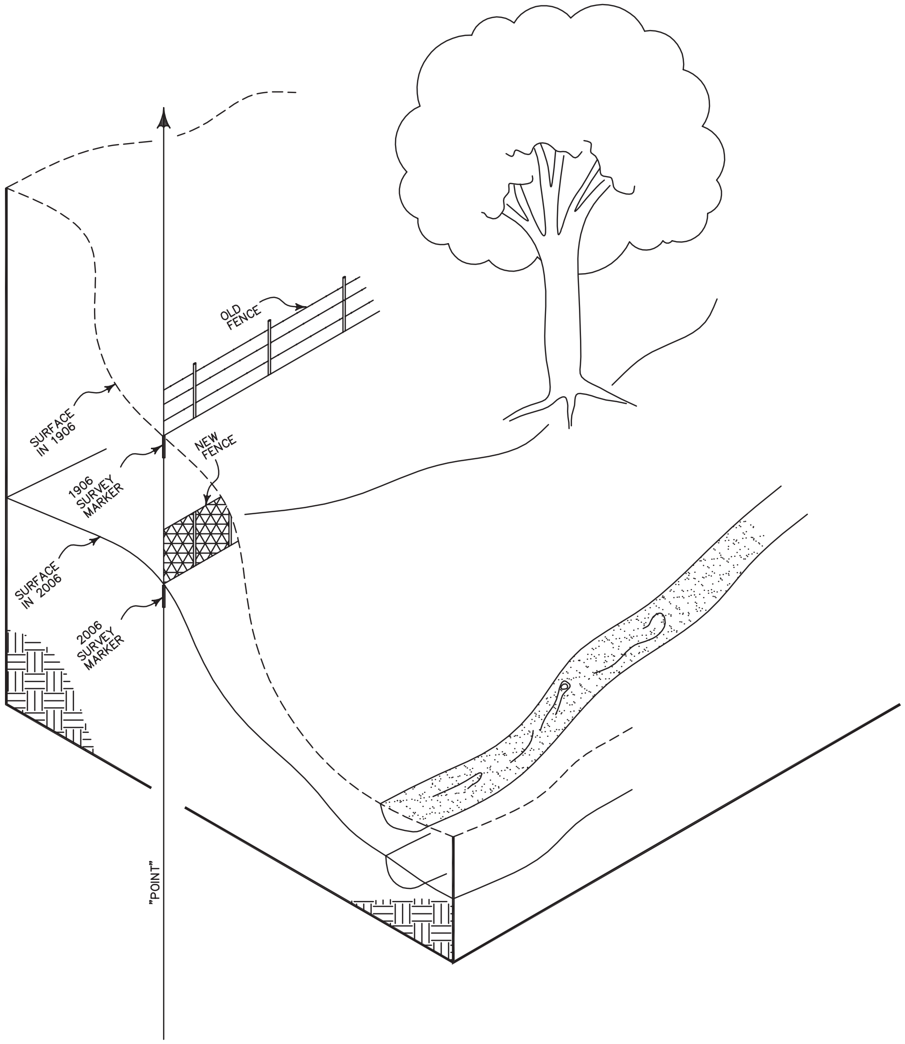
OLD FENCE
SURFACE IN 1906
1906 SURVEY MARKER
NEW FENCE
SURFACE IN 2006
2006 SURVEY MARKER
"POINT"

FIGURE 3.1 Land Point

3.3. LAND LINE

In real property descriptions, a "land line" is a series of land points. There are many different kinds of land lines, but the land line most commonly encountered by the layperson is the boundary or property line. Boundary or property land lines define the limits of rights or possession of land areas. Like real property land points, property land lines are not thought of as "moved" or "shifted" if changes of elevation take place. Land lines are imaginary, theoretical limits that have no thickness but do have direction and location. Because land lines are a series of land points, the strict geometric description of a land line is a vertical surface containing the center of the earth.

3.4. STRAIGHT LAND LINE

The direct, line-of-sight route between two distinct locations (land points) on the surface of the earth has been traditionally defined as a straight land line. Historically, straight land lines are marked on the ground by aligning all interior points of the land line between the end land points. Because land points are geometrically rays and land lines are geometrically surfaces, the procedure for establishing a straight land line results in the establishment of a *flat* vertical plane passing through the center of the earth and bounded by the rays that make up the end land points. Therefore, a straight land line is not altered by shifting terrain. A land line may travel up hills, down into valleys, and across rifts in the surface and still be straight. Any ray from the center of the earth that is contained in the plane of a straight line or in the surface of any land line is said to be a "point on the line."

Note that, in Figure 3.2, the surveyor's instrumentman has set up her theodolite such that the land point passes through, and is coincidental

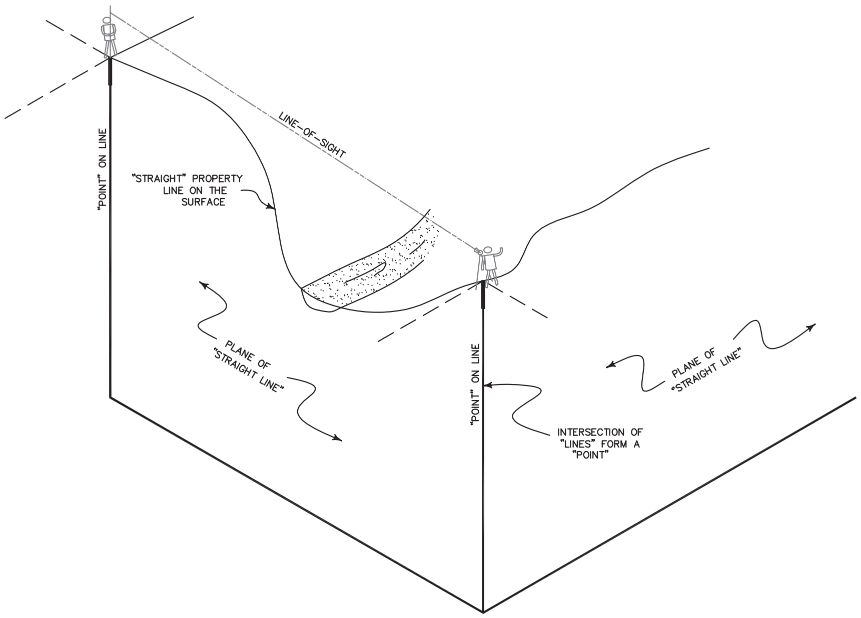
"POINT" ON LINE
LINE–OF–SIGHT
"STRAIGHT" PROPERTY LINE ON THE SURFACE
PLANE OF "STRAIGHT LINE"
"POINT" ON LINE
INTERSECTION OF "LINES" FORM A "POINT"
PLANE OF "STRAIGHT LINE"

FIGURE 3.2 Straight Land Line

with, the axis of rotation of the instrument (in surveying jargon, this is “occupying a point”) and is directing the rodman, who is holding a vertical pole (called a range pole), to a “point on line.” The vertical pole will be “on line” when the ray that is the land point passes through, and is coincidental with, the axis of the vertical range pole.

3.5. PLUMB LINE

In the real property definitions of both a point and a line, the concept of “up” and “down” or “above” and “below” were used. Many children’s programs deal with these concepts by simple demonstrations. Closer examinations of these instinctive terms have revealed subtle complexities, some of which need to be addressed now.

“Up” is away from the pull of gravity. “Down” is toward the pull of gravity. The pull of gravity is *not* always directly toward the center of the earth but toward the apparent center of gravitational attraction. A line that is everywhere aligned with the pull of gravity is said to be a “plumb line.” For most localities on the surface of the earth, the pull of gravity is very nearly toward the center of the earth. The effects of the difference are real, measurable, and, in some rare cases, affect the definitions of the location of real property lines.

3.6. LEVEL

A level surface is a surface such that each and every plumb line passing through that surface does so perpendicularly. The geoid, discussed earlier, is a level surface. “Flat as a mill pond” is a contradictory phrase. On a perfectly calm day, a mill pond is level, not flat. Notice that in Figure 3.3, the level surface ripples or undulates as deflections in the pull of gravity are influenced by irregularities in the earth’s landmasses.

3.7. LAND DISTANCES

Now that we see that real property points are geometrically rays and real property lines are geometrically surfaces (a “straight line” being a flat plane), let us examine how “distances”—that is, the separation of two land points—are defined. In geometry, a “distance” is the measure of the separation of two points in space. In Figure 3.4, it would be geometrically correct, for example, to define the distance between

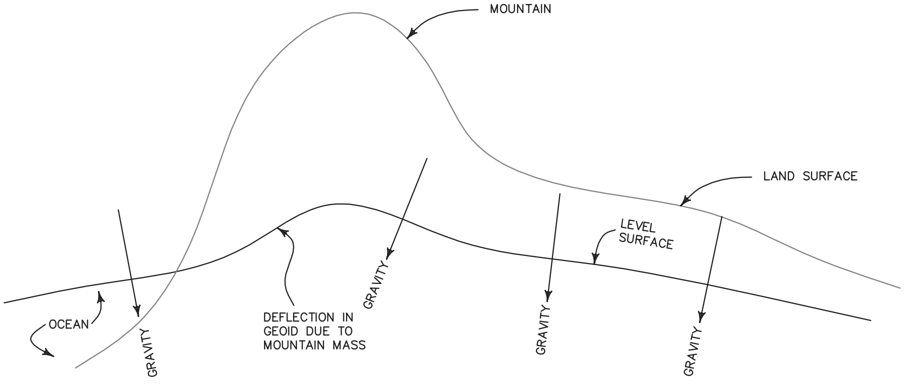
MOUNTAIN
LAND SURFACE
LEVEL SURFACE
GRAVITY
GRAVITY
GRAVITY
GRAVITY
OCEAN
DEFLECTION IN GEOID DUE TO MOUNTAIN MASS

FIGURE 3.3

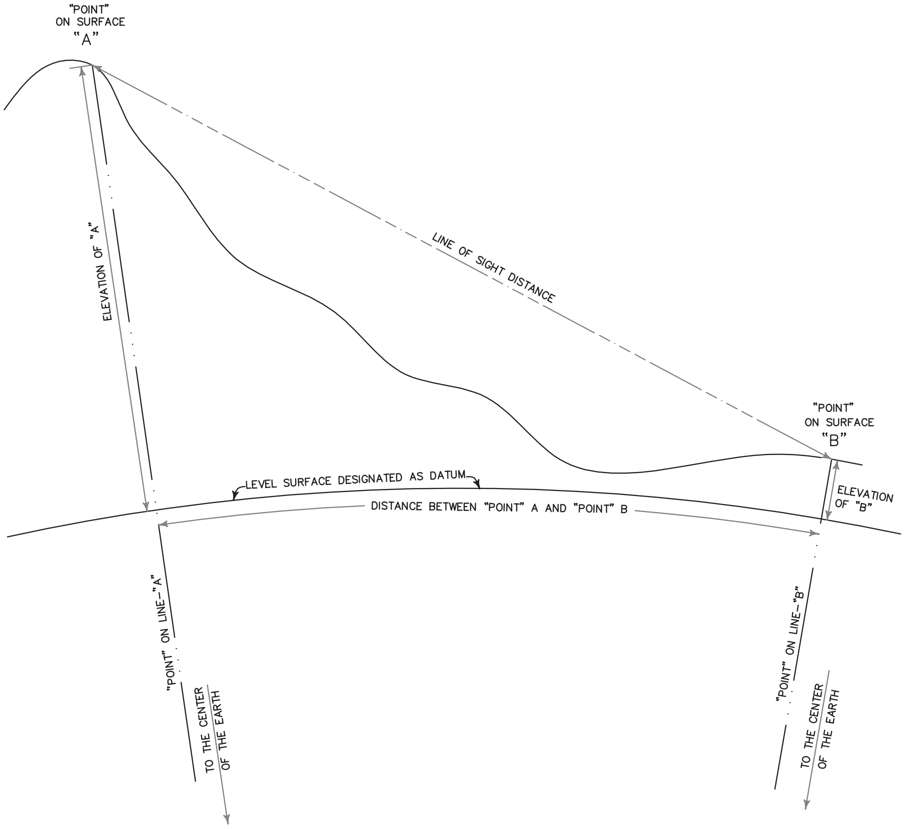
"POINT" ON SURFACE "A"
ELEVATION OF "A"
LINE OF SIGHT DISTANCE
"POINT" ON SURFACE "B"
LEVEL SURFACE DESIGNATED AS DATUM
ELEVATION OF "B"
DISTANCE BETWEEN "POINT" A AND "POINT" B
"POINT" ON LINE-"A"
TO THE CENTER OF THE EARTH
"POINT" ON LINE-"B"
TO THE CENTER OF THE EARTH

FIGURE 3.4

geometric point “A” on the surface of a mountain and geometric point “B” at the foot of that mountain as the measure of geometric line segment having “A” and “B” as end points. This is not the real property definition of the distance between these land points, because real property land points are actually rays.

Although the separation of two rays can be measured by the angle between them, and some private and many national boundaries are measured in this way, in real property terms, the distance between “A” and “B” is defined as being measured along a level surface from the intersection of one ray to the intersection of the other. Land distances are often referred to as the “horizontal” distance.

Notice that, under the real property definition of *distance,* changes in the topography do not alter the land distances. The mountain could be scraped away, but the land distance between point “A” and point “B” would remain the same. It is because of the need to reduce field measurements, which are geometric distances, to land distances, as well as the need to apply a multitude of other correction factors, that, except for very short distances, it is rare for the values shown on any survey plat to be the result of a single, direct measurement.

Distances are dependent on a standardized length. There is no naturally occurring object that is consistently and universally the same length always and everywhere, so a standard had to be created. The meter is the international standard unit of distance measurement. Although it is slowly, very slowly, being replaced by the meter, the U.S. survey foot is the most common unit of measure for real property boundaries in the United States. The U.S. survey foot is slightly longer than the U.S. standard foot, the difference being only about three-thousandths of a foot in one mile. This difference is significant only in long-range, very high-precision work. Other units of measure and the conversion factors for each are shown in Table A.1 in the appendix of tables.

3.8. ELEVATION

Distances can also be used to describe or measure the separation of two level surfaces. This sort of measure is usually identified as “elevation.” The most common use of elevation is in describing all surfaces from a common base or vertical datum. In Figure 3.4, the level surface shown might represent such a datum. The distance from the datum to the level surface containing point “B,” measured along a plumb line, would be the elevation of “B” as referenced to that datum.

Many different areas have established, through convention or municipal and state government, local vertical datums. These different datums are being rapidly replaced by the federal standard datum established by the National Geodetic Survey (NGS). This federal datum is known as the North American Vertical Datum of 1988 (NAVD 88). This replaced the NGS-defined National Geodetic Vertical Datum of 1929 (NGVD 29). Both datums are frequently and incorrectly called mean sea level (MSL). Neither is a sea level datum. Local mean sea level will vary greatly from the national datum.

The NGS has consistently emphasized that NAVD and NGVD are not sea level datums. NGS has also emphasized that NGVD is no longer supported and should not be used for surveying or engineering projects. In spite of this, many agencies, even federal regulations, will occasionally refer to NGVD or "mean sea level" or "MSL." Local sea level must be determined by observations of the tide. Such observations can render a sea level value in terms of NAVD.

Elevations are every bit as important in real property measurements as are horizontal distances and angles. More personal disasters have occurred because an individual's real property was below a certain elevation (a high-water mark, for example) than because of shortages in horizontal measure or improper horizontal location. In many regions of the country, there are significant changes in elevations taking place because of subsidence or heaving of the surface. Some changes in reported elevations are superficial—the result of improved measurement procedure. Many more of the changes are the result of actual surface movement. It is important, in regions of rapid changes, that elevation information be current.

3.9. LAND AREA

In the same way that real property distances are horizontal measurements usually defined at some datum, land areas are also defined as "horizontal" or "level" and are usually referenced to the same datum as the horizontal distances. The actual surface of a tract of ground—that is, the air-land interface, can be altered by earthwork. Leveling hills or plowing furrows alters the exposed surface of dirt, but land areas, as defined in real property terms, cannot be altered by earthwork.

3.10. HORIZONTAL ANGLES

Because real property lines are actually vertical planes, the intersection of these planes will form a vertical ray or a land point. A plane perpendicular to this vertical ray will also intersect with these

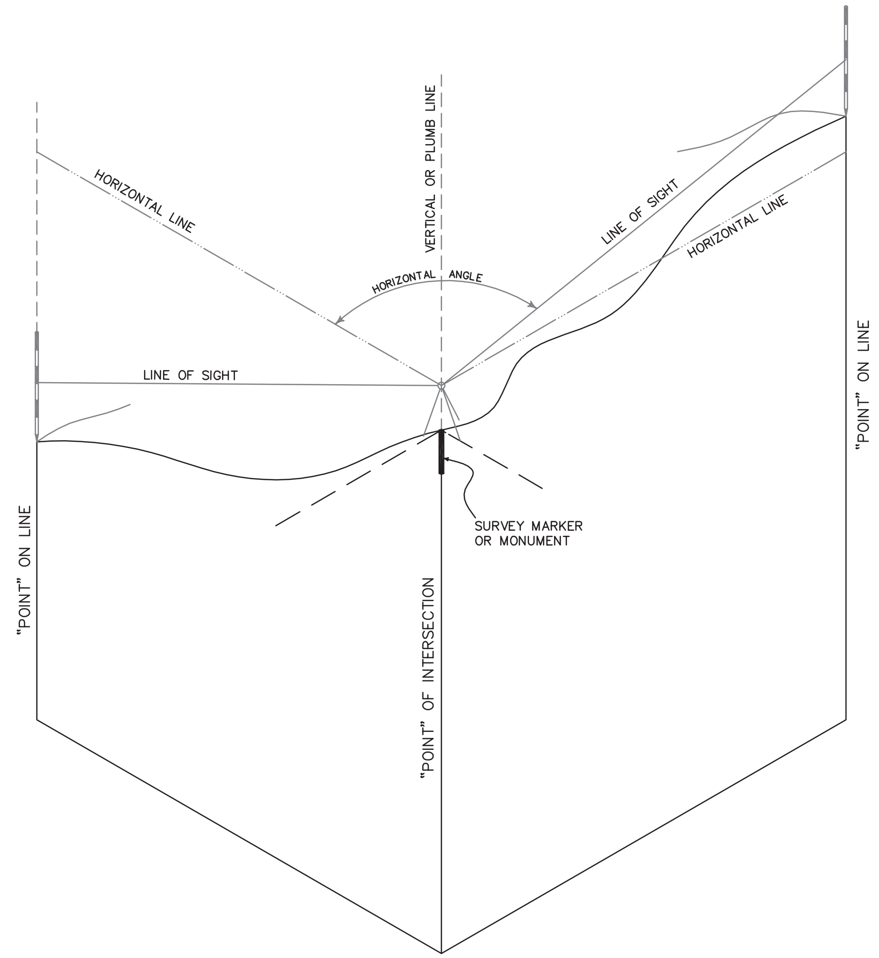
HORIZONTAL LINE
VERTICAL OR PLUMB LINE
LINE OF SIGHT
HORIZONTAL LINE
HORIZONTAL ANGLE
LINE OF SIGHT
"POINT" ON LINE
SURVEY MARKER
OR MONUMENT
"POINT" ON LINE
"POINT" OF INTERSECTION

FIGURE 3.5

vertical places, forming two "horizontal" geometric lines. It is the angle between these two "horizontal" lines that is reported as the angle of a property boundary intersection.

Many have found it convenient to refer to the angle between two such lines as if the measurement were an action that took place by turning from one line to another. When this sort of "reckoning" is used, angles are said to be positive, or to the right, if the direction of the imaginary "turn" is clockwise when observed from above, and negative, or to the left, if counterclockwise.[1] Consequently, the angle in Figure 3.5 could be said to be 90 degrees, or to the right, from “A” to “B,” but –90 degrees, or to the left, from “B” to “A.”

---

[1] The actual process of measuring angles is not at all dependent on the "direction" of the reckoning.

3.11. VERTICAL ANGLES

Vertical angles are measurements between geometric lines in a vertical plane. Because real property boundaries are vertical planes, vertical angles rarely find their way into property descriptions. Vertical angles are commonly measured in order to compute horizontal and vertical distances. Figure 3.6 illustrates a vertical angle.

3.12. DEGREES, MINUTES, AND SECONDS

All angles are based on a definition of a unit, not a standard. A point traversing a circle will, when observed from the center, appear to travel the same distance to complete one circuit, no matter where that circle or observer is in the universe. One could contact an alien on another planet and describe a “degree” as $^1/_{360}$ of a circle, and that alien could duplicate the standard, having never seen an example. However, no amount of verbal description would allow the alien to duplicate the U.S. survey foot; the alien would have to be given a physical example in order to duplicate it.

The relationship of the circumference with the radius of a circle is a natural constant. All angular measurement is defined in terms of a full circle. Table A.2 in the appendix of tables lists the most common definitions for angular measurements. In the United States, the most common definition of angular measurement is 360 degrees to a circle. The degree (noted by the symbol $^\circ$), is divided into minutes (noted by the symbol $'$), which is $^1/_{60}$ of a degree. The minute is divided into seconds (noted by the symbol $''$), which is $^1/_{60}$ of a minute.

3.13. MAPS OR PLATS

A map or plat is a report of a survey in the form of a drawing. The root cause of the confusion between the terms used in real property descriptions and between identical terms, with very different geometric meanings, is the widespread use of maps or plats.

Since the invention of paper, and perhaps even before that ancient time, people have communicated to one another the location of places on the surface of the earth by drawing lines on a flat surface. First as

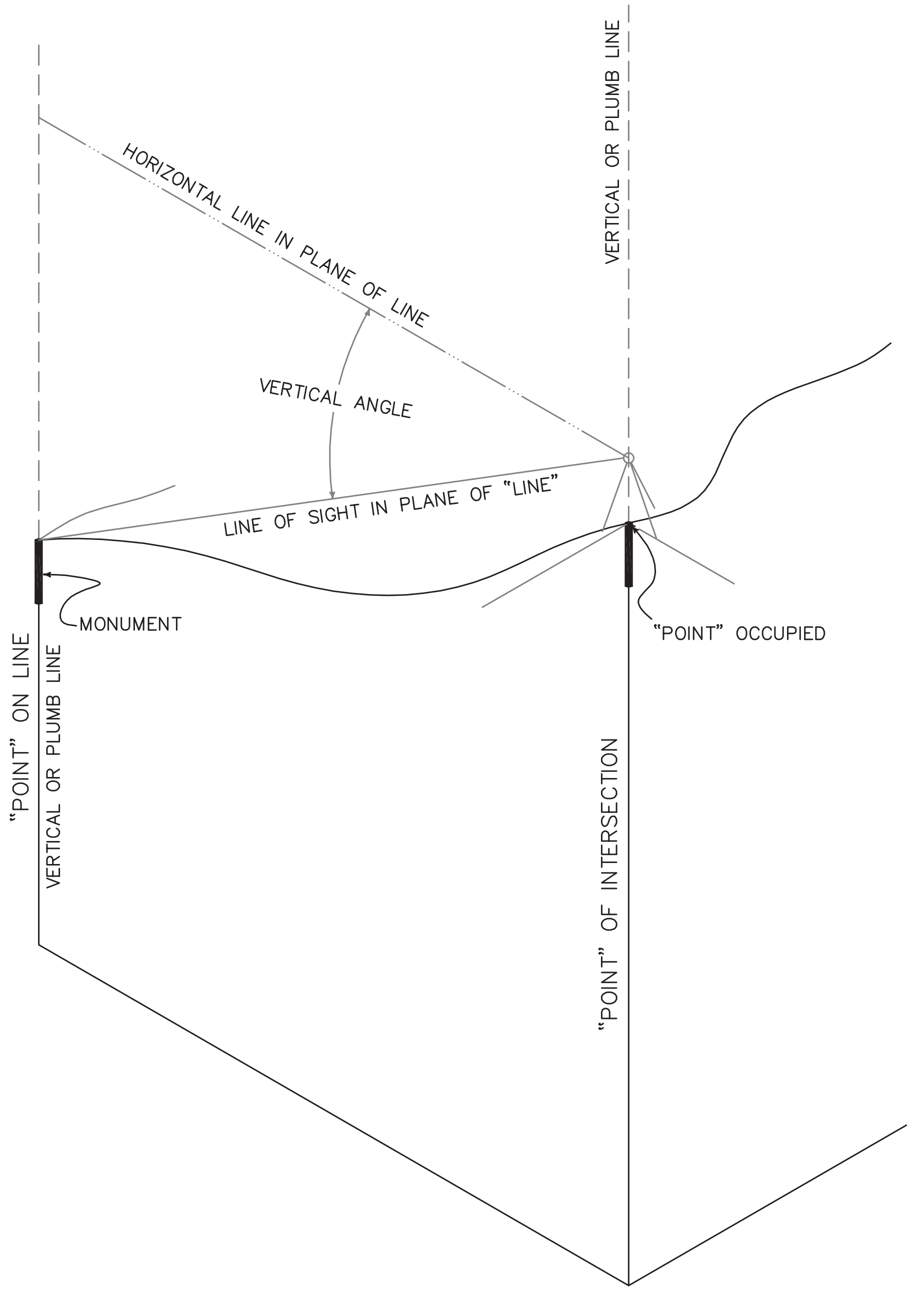
VERTICAL OR PLUMB LINE
HORIZONTAL LINE IN PLANE OF LINE
VERTICAL ANGLE
LINE OF SIGHT IN PLANE OF "LINE"
MONUMENT
"POINT" OCCUPIED
"POINT" ON LINE
VERTICAL OR PLUMB LINE
"POINT" OF INTERSECTION

FIGURE 3.6

lines scratched in the dirt and then as ink lines on paper, the map was developed as a communication aid. Even after the concept emerged of drawing line lengths and angles to scale, the map or plat was, is now, and always will be a two-dimensional representation of a three-dimensional reality.

A map is not a model. The interaction of ink lines drawn on a map is not the same as the interaction of real property land lines on the earth. Yet we persist in using the words and descriptions of the two-dimensional map as if it, and not the earth's surface, were the thing being described.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 4

DEFINING NORTH

Descriptions of real property boundaries often use terms that report the orientation of each of the boundary lines with a common line or datum. This orientation is called the “direction” of a boundary. The most common datum for boundary directions is “north.” Once again, the commingling of three-dimensional terms, two-dimensional instincts, and one-dimensional verbiage makes it necessary to separate and explain each concept in terms of what is really happening, what is shown, and what is described.

No concept in real property boundaries is so widely misunderstood as the term *north*. The misuse of this term in real property boundary definitions is so pervasive and so perfectly demonstrates the distortion of three-dimensional reality that a complete understanding of this single term and its many variations is required.

4.1. TRUE NORTH

True north is a three-dimensional term. In real property terms, straight land lines are defined as “true north” if the axis of rotation of the earth is entirely contained in the plane of that straight line. There are an infinite number of “true north” lines. All “true north” lines are straight land lines, but *none are parallel*.

4.2. ASTRONOMIC NORTH

Astronomic north was long believed to be true north. The apparent rotation of the stars, first observed by the ancients, is still used today as an accurate and universal method of determining the rotational axis of the earth. Unfortunately, the measurements of the angles between the horizon and the axis of rotation of the earth depend on an accurate determination of "down."

We must usually make the observations of the apparent movement of the stars at a point on the surface of the earth. This close proximity with the surface of the earth means that the pull of gravity, the "plumb line," will not always point toward the center of the earth. Land mass anomalies frequently deflect the plumb line from the vertical. Because these deflections distort celestial observations, astronomic north lines are not quite straight land lines but are meandering land lines that closely approximate true north. Astronomic north lines can be adjusted to "true north" by the application of regionally varying correction factors. These corrections are typically, but not always, very small.

4.3. MAGNETIC NORTH

Magnetic north is not directly connected to the rotational axis of the earth. The earth has a naturally occurring magnetic field focused on points near the Arctic and Antarctic regions. These points of focus, or poles, are not stationary, and the magnetic lines associated with these poles are far from straight and are not true land lines, for their location and intensities vary with altitude. These "lines of magnetic force" are easily interfered with or deformed by surrounding conditions. Alignment with these magnetic lines defines "magnetic north."

The corrections required to translate magnetic north, as observed at a point on the surface of the earth, to true north vary so rapidly that it is quite useless to attempt even a moderately accurate determination of true north based on magnetic observations. In spite of this, many old survey procedures were regularly based on magnetic lines and required the surveyor to adjust the directions observed on a compass by what was believed to be the variation of magnetic north from astronomic north, called "declination," or sometimes "deviation." The results of such practices will be discussed later.

4.4. ASSUMED NORTH

When the direction depicted as north is based on something other than observations intended to detect the axis of rotation of the earth, that direction is referred to as *assumed north*. Many, if not most, boundary surveys use an assumed north as the direction of orientation. Many survey plats fail to state the type of "north" used to define the directions on the plat. In such cases, it is almost certain that the north used was an assumed north.

It is also quite common for a survey plat, especially an old one, to report the base direction as "true north" when, in fact, the direction of orientation is an assumed north. *Assumed north lines are unique to each particular survey and have no fixed relationship to any other assumed north*. Chapter 12, which deals with the evaluation of survey plats, will demonstrate how to identify the type of "north" used in a survey plat.

4.5. GRID NORTH

Of all the "north" terms discussed in this chapter, *grid north* is the only two-dimensional term. If a flat plane is divided into a set of evenly spaced parallel straight lines intersecting at right angles (90 degrees) with another set of evenly spaced parallel straight lines, the resulting pattern is a "grid." "Grid north" could then be defined as a direction parallel with one of these sets.

This grid pattern (Figure 4.1) could be defined as having a specific relationship to portions of the surface of the earth. When this occurs, such a grid pattern can become an effective tool in mapping the earth's surface.

All maps, which are two-dimensional planes, depend in some way on a grid pattern to control locations of the drawings and symbols that represent land features. The science of cartography is the highly complex and technical study of the methods and concepts necessary to present the surface of the earth as a two-dimensional grid (map). (Later in this book, we will examine briefly some of these methods.) Unlike true north, astronomic north, and magnetic north, grid north lines are parallel.

4.6. DIRECTIONS

The various concepts of "north" discussed so far have dealt with land lines or planes that are oriented along the earth's axis of rotation. There

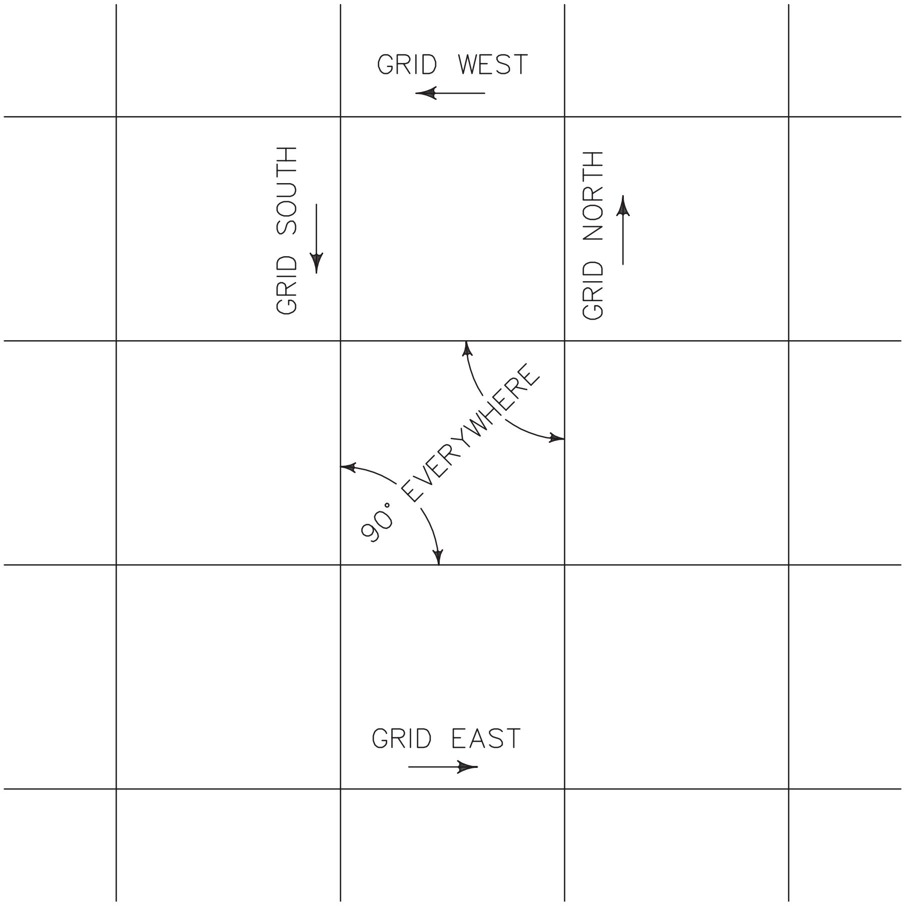

FIGURE 4.1

is an infinite group of land lines or planes that are not "north lines." These land lines can be, nevertheless, defined by their relationship to a particular "north." Azimuths and bearings are two universally accepted methods used to describe that relationship. Both of these methods are systematic reports of the angles formed between a land line and a meridian. The particular meridian used *must be defined*.

In the examples for azimuths as well as for bearings, the values report the relationship of land lines (flat planes) that intersect at a land point (vertical ray). The various definitions of direction are the same for each of the types of "north" involved. The methods of identifying directions and the verbiage employed are the same in two-dimensional systems.

Imagine a flat plane perpendicular to the line of intersection of a land line and a meridian. The vertical ray that is the line of intersection will appear as a geometric point on that imaginary plane. The vertical planes that are the land lines and the meridian will intersect the imaginary plane as geometric lines. The description of the relationship of the land line and the meridian has now been reduced to a two-dimensional image. This relationship is valid only *at a particular land point of intersection*. Reducing an intersection to a two-dimensional relationship for the purposes of describing the direction of a land line *does not eliminate the three-dimensional nature of the land lines involved*.

In Figure 4.2, the line A–C intersects the north or meridian line at point B. Because the meridian is a straight line at point B and line A–C is a straight line at point B, then only one of the angles created needs to be reported or measured to define the direction or orientation of line A–C. Because azimuths and bearings developed as an aid to navigation or travel, both have the ability to distinguish the direction "B" to "A" from "B" to "C." This one-dimensional aspect often causes confusion among users of land information.

4.7. AZIMUTHS

Azimuths are "measured" or "reckoned" from the north in most cases.[1] Some organizations "reckon" azimuths from the south. Azimuths are always said to be clockwise, meaning that the observer would "turn right" from the meridian to the line being reported. If an observer standing at "B" in Figure 4.2 were to face north, a right turn of 30 degrees would cause the observer to face "C." A right turn of 210 degrees (30 degrees + 180 degrees) from north would cause the observer to face "A." Therefore, line "A–C" could be reported as having an azimuth of 30 degrees or 210 degrees. *Both azimuths describe the same line*. Azimuths are always less than 360 degrees.

4.8. BEARINGS

Bearings are "reckoned" clockwise or counterclockwise[2] from the meridian and are always less than or equal to 90 degrees. In the case

---

[1] In actual practice, the measurement of angles involves a process that is independent of such "reckoning."

[2] "Clockwise" is defined as a turn to the right from the meridian, and "counterclockwise" is defined as a turn to the left.

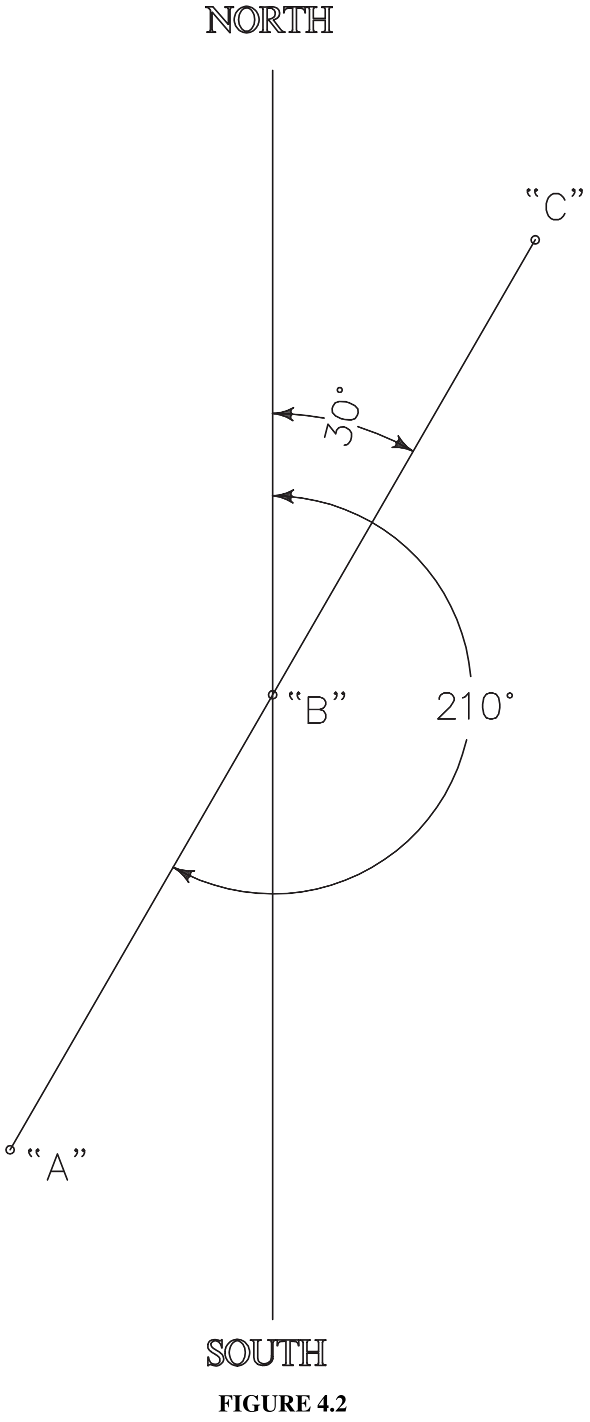
NORTH
"C"
30°
"B"
210°
"A"
SOUTH

FIGURE 4.2

of bearings, the reporter defines lines as being east or west of north, or east or west of south. This method is easily comprehended and widely favored by most users of real property information.

In Figure 4.3, land line “D–F” has been added to the Figure 4.2. If an observer at “B” were to face north, a right turn (clockwise or to the east) of 30 degrees would cause the observer to face “C.” The bearing to “C” is north 30 degrees east. If that same observer at “B” were facing south, a right turn (clockwise or to the west) of 30 degrees would cause the observer to face “A.” The bearing to “A,” therefore, is south 30 degrees west. The bearing of land line “A–C” could be reported as being north 30 degrees east or as being south 30 degrees west. *Both bearings define the same land line*. The distinction of when to use one rather than the other will be discussed in Chapter 13.

If the observer were to face north and turn left (counterclockwise or to the west) 10 degrees 15 minutes 22 seconds, then the observer would be facing “D.” The bearing to “D” is north 10 degrees 15 minutes 22 seconds west (written N 10° 15′ 22″ W). If the observer were to face south and turn left (counterclockwise or to the east) 10 degrees 15 minutes 22 seconds, then the observer would be facing “F.” The bearing to point “F” is S 10° 15′ 22″ E. Both bearings describe land line “D–F.”

If the magnitudes of the angles are the same, north–east bearings and south–west bearings describe the same line and are clockwise, or reckoned to the right. If the magnitudes of the angles are the same, north–west bearings and south–east bearings describe the same line and are counterclockwise, or reckoned to the left.

The concepts for bearings and azimuths discussed and sketched earlier are based on a relationship of lines *at a particular point*. In two-dimensional geometry, the meridian and all north–south grid lines are parallel, but, in the three-dimensional world of real property boundaries, meridian lines, such as longitudes, are not parallel with any other north–south line. Because straight real property lines are actually flat planes through the center of the earth, the following facts are true, although they may come as a shock to the “flat-earth” believers among us:

1. There cannot be two parallel straight lines.

   Because every straight land line is actually a plane containing the center of the earth, then every such plane must intersect. For any two straight lines, there are two places on the surface of the earth where these lines intersect; therefore, no two straight land lines can ever be parallel!

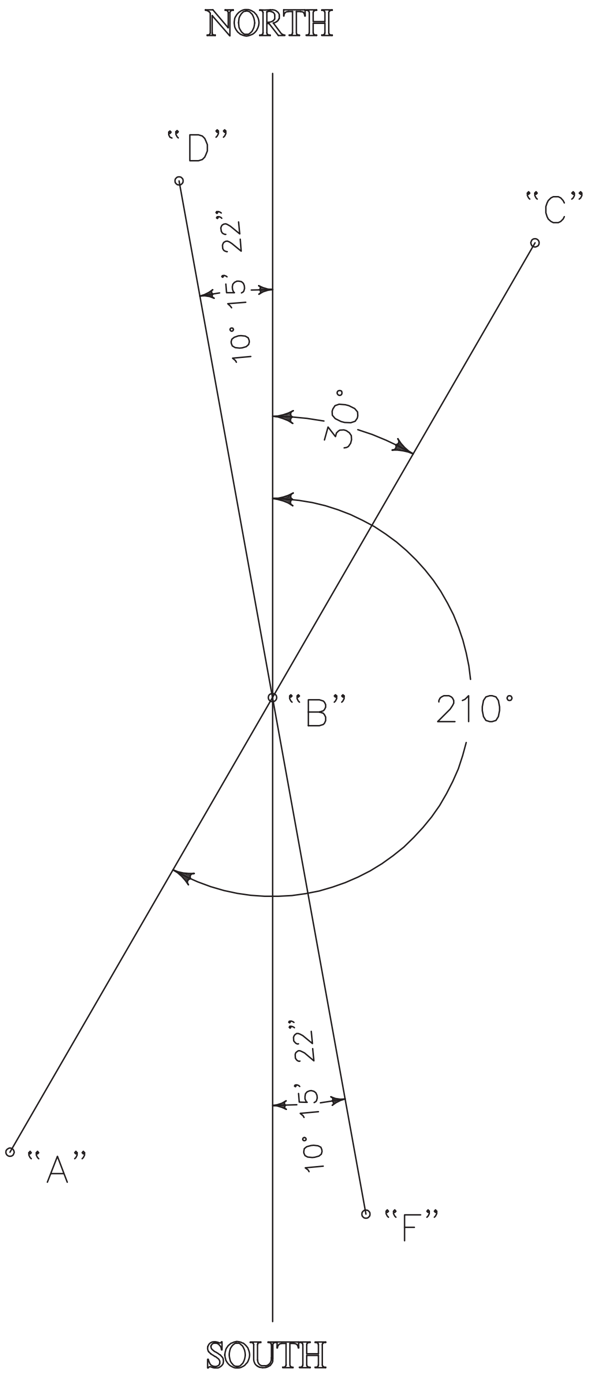
NORTH
"D"
"C"
10° 15' 22"
30°
"B"
210°
10° 15' 22"
"A"
"F"
SOUTH

FIGURE 4.3

2. The bearing of straight lines (except for north–south lines and the equator) vary throughout the line.

   Because none of the north–south lines are parallel, then a straight line that crosses several north–south lines must do so at a different angle for each north–south line.

A simple examination of a globe is all that is needed to confirm that north–south lines (lines of longitude) are not parallel (they meet at the poles). East–west lines (lines of latitude), however, are parallel but, except for the equator, none are straight land lines.

The consequences of these facts in documenting real property transfers are best demonstrated by an example. Let us return to the ranch in Montana. We might attempt to mark a square mile of land in many ways. First, if we were to mark every step of the woman walking in the scenario, we would notice several things. The north leg of the trip would be a straight land line segment, one mile long. The east leg of the trip would be a land line segment curving to the north, one mile long. The south leg of the trip would be a straight land line segment, one mile long. The west leg of the trip would be a land line segment

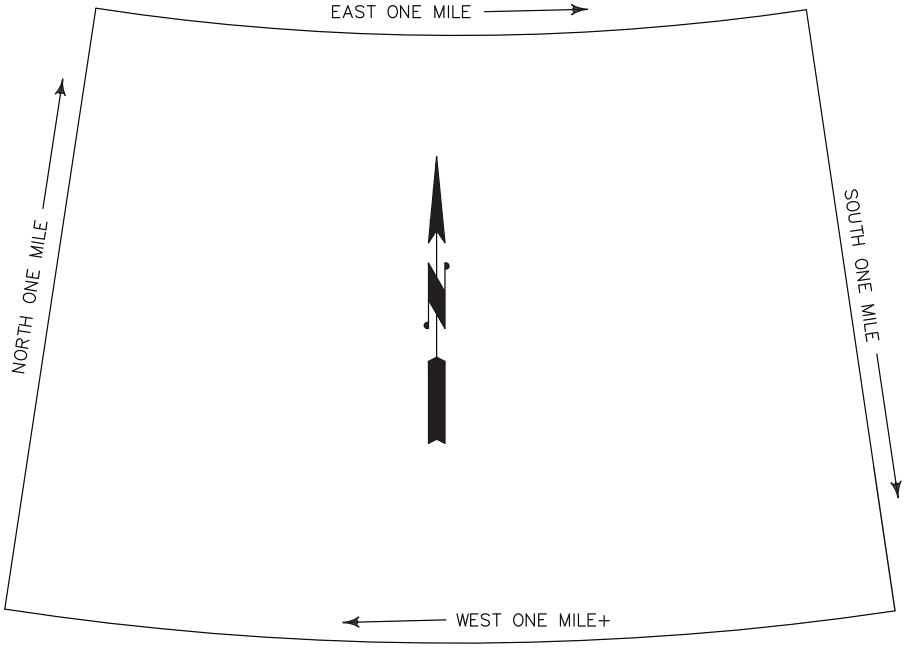

FIGURE 4.4

curving to the north, and, in order to return to the starting point, the walker would be required to travel slightly more than one mile.

In this example, such a closed circuit could represent a land parcel. This parcel could be described very accurately by reporting the directions and distances of the route. Figure 4.4 is an exaggerated drawing of such a parcel in Montana.

Although the parcel shown in Figure 4.4 is completely possible, the curved lines of the northern and southern boundaries are difficult to mark on the ground. By curving the lines on the figure, the actual shape of the parcel can be approximated. This is not possible with larger areas of land. The convenience of straight boundary lines,[3] as well as the human desire for regular shapes, reduces the acceptance of such a strict geodetic definition of real property parcels.

---

[3] The greatest convenience of straight land lines is that only the ends (corners) need to be marked to define clearly the location of that line on the ground.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 5

PROJECTION SYSTEMS

5.1. PROJECTIONLESS MAPS

The first attempts at drawing maps in a scaled and controlled manner where the distances, locations, and directions, as observed on the ground, were reproduced in miniature on a sheet of paper failed miserably. Model globes are excellent miniatures of the earth, but the scale required is far too small for practical application.

Except for very small land parcels, the first attempts at mapping land areas on flat sheets of paper (now called "projectionless mapping") resulted in inconsistencies of ever-increasing magnitude as the area being mapped was enlarged. This was because the earth's surface is not flat. However, to be able to talk about, draw, refer to, and think of real property parcels in two-dimensional terms by using maps, sketches, or plats is so necessary, basic, and instinctive that methods had to be developed that would be able to transfer the surface of the earth to a flat plane.

The systems that were developed are called "projection systems." There are many different versions of projection systems—too many to enumerate here—but the advantage of all projection systems is that, when applied, land surface areas can be mapped and described in two-dimensional terms. Straight lines can be parallel, and all straight lines will have constant bearings throughout. *Point, line, angle,* and other geometric terms can now be correctly used in their two-dimensional sense to designate real property features.

As long as the user is aware of the fact that a projection is involved, maps and property descriptions are readily developed and can, very accurately, reflect real property parcels. The few systems that are going to be discussed here are all "conformal" projection systems. This means that angles measured at the surface of the earth are presented on the projection (map) with minimum distortion.

5.2. TANGENT PLANE PROJECTIONS

The most common form of plane projection systems in use today is the tangent plane projection system. Although it is slowly being replaced by more sophisticated systems, most survey maps or plats are based on this simple system. This is especially true of small surveys and surveys performed before the World War II. Modern, large-tract surveys are less likely to be based on this system because of several disadvantages that affect accuracy and recoverability.

Figure 5.1 is a visualization of a tangent plane projection. Imagine that a flat plane of glass is balanced on a globe. The plane would touch the globe in only one spot. This is the "point of tangency." If one were

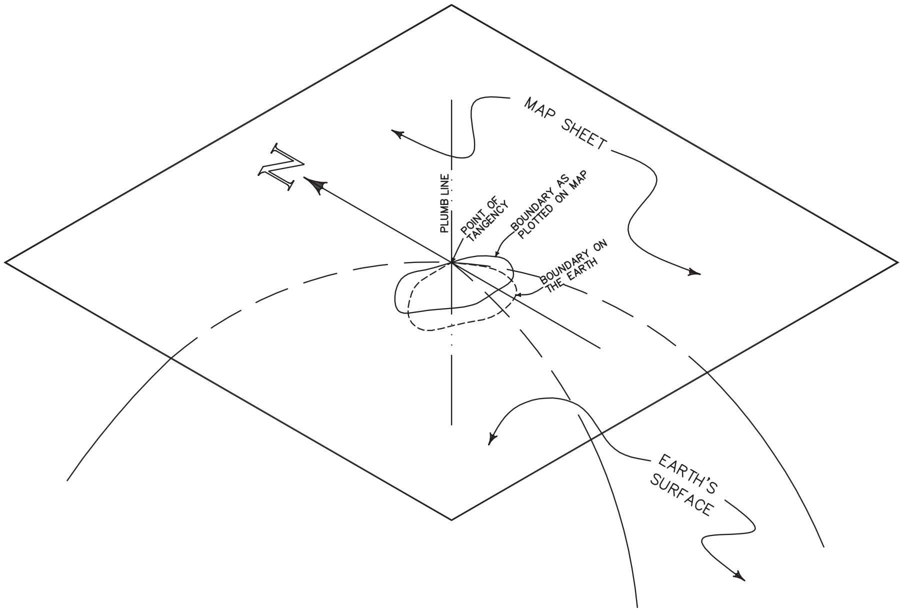

FIGURE 5.1 Tangent Plane Projection

to look down through the glass from directly above the globe, features on the globe would be visible. These features could then be traced on the glass in their apparent positions as observed from above.

The tracings on the glass would be a map of those portions of the globe visible from above. The features mapped would become increasingly distorted as the distance from the point of tangency increased. The area very near the point of tangency would have minimum distortion and, depending on size and precision requirements, could become a very serviceable two-dimensional map of that portion of the globe.

The procedures for establishing a tangent plane projection in order to survey and map a parcel of land roughly imitate the preceding example. There are many variations of this procedure—some of which will be discussed later—yet all tangent plane systems have certain common factors. Concepts or steps that are *italicized* in this discussion will be those things common to all forms of tangent plane systems.

In order to *establish the point of tangency*, the Land Surveyor may choose a location on or near the area to be mapped or surveyed. This location is the point where the Land Surveyor *orients the work to some form of directional control* that is recoverable, meaning physical marks on the ground, magnetic observations, celestial observations, or any procedure that defines a direction that can be repeated by another, at another time.

True north is widely claimed but almost never used as this control meridian. The most common form of directional control in boundary surveys of this type is an assumed north created by the recovery of a reported bearing. A "reported bearing" simply indicates that two or more marks set by a previous Land Surveyor were found and that the direction between these marks reported by the previous Land Surveyor was used as the directional control for the present work. Sometimes the control direction may be entirely "assumed"; that is, the Land Surveyor did not use a known or recoverable direction as a control.

The Land Surveyor then *measures the angles formed between lines of the survey and the control direction*. (This is very different from independently measuring a magnetic or astronomic direction for each line, as is the case in projectionless maps.) The angles so measured during the survey may then be drawn, two-dimensionally and to scale, on the map or plat. More commonly, the relationship of the land lines in the survey with the control direction are expressed by reporting the bearings of the lines on the map. When such bearings are used, they are *two-dimensional computations based upon angles measured from the control direction*. Bearings shown on tangent plane projections, and even those correctly based on true north, *do not represent the geodetic,* *astronomic, or true bearing of the land line*, except along the meridian that passes directly through the point of tangency. *The curvature of the earth and the convergence of the meridians are ignored*.

The distances measured during the survey are reduced to horizontal distances at a local datum. No adjustments are made to compensate for differences between distances between points on this level surface and distances between the same points on the flat plane of the projection. In short, the tangent plane projection system accounts for the curvature of the earth by ignoring it. Indeed, so little attention is given to the shape of the earth in performing tangent plane projection surveys that many people, even Land Surveyors, forget that every map or plat derived from measured angles and distances is the projection of a curved surface onto a flat plane.

5.3. ADVANTAGES OF TANGENT PLANE SYSTEMS

The advantages of a tangent plane system are many and compelling. Chief among these advantages is the ease with which such a system is established. The Land Surveyor has a wide choice of methods of directional control and complete freedom in choosing the point of tangency. This freedom reduces to a minimum the fieldwork required to map a particular parcel of land.

Knowledge of the elevation or geodetic position of the area being surveyed, in order to account for the earth's shape, is not necessary. Because the shape of the earth is ignored, there is no need to collect information on the geodetic positions of points within the parcel. This is a further reduction of the fieldwork required to survey and map a parcel of land.

The computations required are all two-dimensional and relatively simple. Field personnel do not need to master geometry beyond the high school level in order to understand and execute the data collection required. This reduces the formal training required of entry-level survey crew members, as well as the on-the-job training needed to attain adequacy. Because training requirements are low, salaries paid to field personnel are also typically low, especially when the responsibilities and the value of the work performed are considered. Consequently, the initial or "first-time" costs of tangent plane projection surveys are comparatively low.

Straight land lines have constant bearings. Parallel straight land lines exist and have the same bearing. Landowners and other laypersons quickly accept and understand the maps or plats produced by this system because the information conforms to their flat-earth instincts. The intricacies of defining land lines on the earth are greatly reduced and simplified.

Distances used in computations and shown on maps do not require a lot of "refining" and, in most cases, are simply a report of the level distance between land points at the average elevation of the site. Indeed, changes in elevation throughout the site are rarely great enough to be significant.

5.4. DISADVANTAGES OF TANGENT PLANE SYSTEMS

The disadvantages of tangent plane projection systems are as numerous, although not as obvious, as the advantages. Unless referenced to the same point of tangency, tangent plane surveys are like crackers floating in a bowl of soup: each is free of the other without any ties or a common reference base. This means that a survey based on the tangent plane projection will not contain any information about that particular survey's relationship to any other work that is not a part of the present survey.

Other surveys in the area, even those of adjacent properties, will disagree on the dimensions and directions of common lines. The directions between identical points on adjacent parcels will show different values. This lack of consistency of direction means that discrepancies, overlaps, disputes, or other irregularities will not be apparent by merely reviewing the survey maps of adjacent parcels. Potentially devastating conflicts in deeds may go unnoticed for generations until clear acts of possession take place.

Resurveys of real property parcels are entirely dependent on recovery of survey marks or monuments set during the original work. This greatly increases the time required to retrace previous work. Failure to recover at least part of the original work will result in a retracement that has a poor probability, at best, of remarking boundaries in their original locations. Because permanence of boundaries is an integral part of the rights of possession, this is a most serious flaw.

Tangent plane projection surveys are of little use to land planners, municipal utilities, and other governmental agencies, because the lack of a common reference base makes the task of relating several parcels to each other very difficult. Multipurpose land record systems, so vital to proper community planning and resource planning, can make little use of the work of private tangent plane surveys because of the lack of a common reference base.

Recovery of boundaries, resurveys, and resubdivision of existing parcels require extensive field recovery and frequent judgmental decisions by the Land Surveyor. This greatly increases the costs of resurvey or recovery work and increases the possibility of conflicts. In the long term, tangent plane projection surveys can be much more expensive than more sophisticated systems. This “hidden cost” limits the cost effectiveness of tangent plane projection surveys to very small, usually urban, parcels, where boundary lines are well established, clearly marked, and well maintained.

Because the area surveyed under a tangent plane system must be kept quite small, or the factors ignored by the system will result in unacceptable irregularities, this eliminates the tangent plane system for use in large-tract surveys, highway routing, and state- or county-level property accounting.

5.5. STATE PLANE PROJECTIONS

The rapid increases in land values, the need to mark real property boundaries permanently and precisely, the increasing governmental adaptation of Land Information System (LIS) and Geographical Information System (GIS) databases,[1] and many other factors have led to the wide use of projection systems that eliminate some of the disadvantages of the tangent plane system. This need was so prevalent that, late in the 19th century, the federal government of the United States began promoting the use of particular projection systems adapted to each state.

Known as State Plane Coordinates (SPC), the system is based on a rigorous mathematical translation of Geodetic Positions into a Grid Coordinate system. The adoption of SPC in private surveys gradually increased through the 20th century. Modern advances in Geodetic Positioning in the 21st century have made the control required for SPC-based surveys available to every Land Surveyor in every corner of the world. The specific version of SPC control that each state adapted is codified in that state’s law. The projection systems used vary greatly, but the benefits, application, advantages, and use of these systems is the same for every state.

One form of state plane projection is demonstrated in Figure 5.2. Instead of a rigid plane of glass, as in Figure 5.1, imagine that a clear, flexible sheet of plastic is wrapped to form a cylinder around a globe.

---

[1] See Chapter 9.

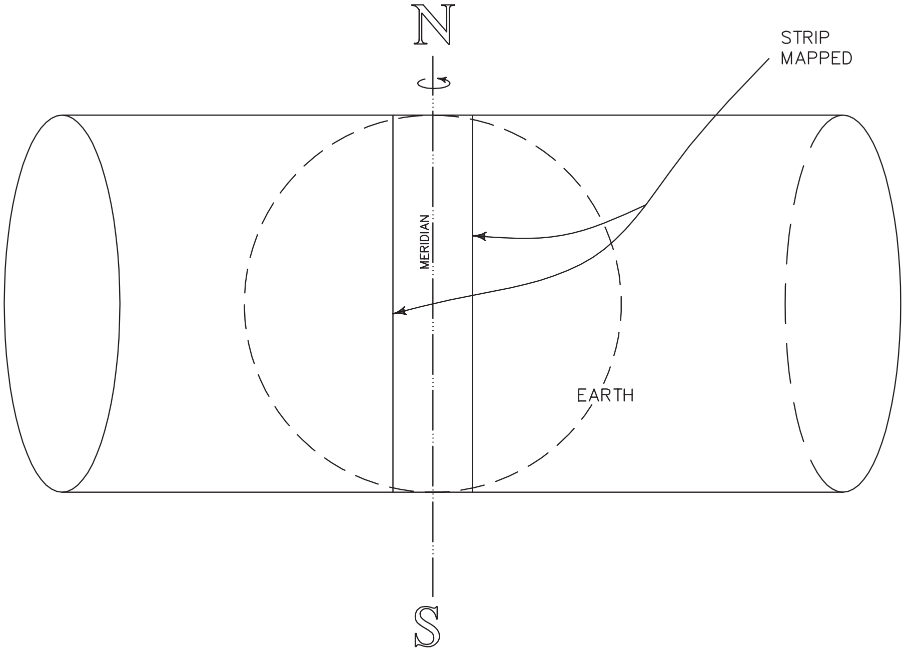

FIGURE 5.2 Transverse Mercator Projection

Then the sheet would touch the globe in a series of points instead of in one place. The features visible through the sheet could be traced onto the plastic, and, when unrolled, the sheet of plastic would form a flat plane on which a map of the globe would appear. The features on the map would become distorted as the distance to the line of contact was increased. The area near the line of contact would form a strip that would closely match the globe.

Let us further visualize that, instead of using a blank sheet of plastic, we begin with a clear sheet of plastic that has a rectangular grid pattern already established on it. This grid could consist of two sets of parallel lines intersecting at right angles, and each line would be defined by its distance from an established origin. The distance east is called the “x” distance. The distance north is called the “y” distance. When this grid sheet is wrapped around the globe, we could also deliberately align one of the grid lines with on of the north lines on the globe along the line of contact. Now, for each and every feature on the globe for which there exists a geodetic latitude and longitude measured in degrees, there can be found a corresponding grid location on the projection denned by the "x" distance and the "y" distance on the grid and measured in feet or meters, depending on the system.

Because the relationship of the sheet of plastic and the globe was rigorously controlled, then a specific set of such strips could be used to map the entire earth. In actual practice, this is what is done through mathematical manipulation of the geodetic positioning network. This procedure and innovative variations of it (using cones or other adaptations) have produced mapping systems that cover large portions of every state. Most states are typically served by two or three overlapping projection zones.

Maps and surveys developed in these systems are tied to the worldwide geodetic positioning system of latitudes and longitudes. The projection accounts for the convergence of the meridians, as well as the elevation of the parcel. The system mathematically and rigorously projects a portion of the earth's surface onto a theoretical flat plane. The portion of the earth mapped is typically a strip, about 150 miles or so wide, wrapped around the earth. Most states are divided into overlapping zones in order to obtain full coverage.

The geodetic monumentation network that supports these projections is maintained by the federal agency known as the National Geodetic Survey (NGS). This agency is responsible for the development and definition of the mathematical models used to approximate the earth. The foundation of this model has undergone significant renovation in recent years. The consequences and advantages of these modifications will be explored.

The development of the Global Positioning System (GPS)[2] using a constellation of satellites has greatly reduced the dependence on surface monumentation for precise positioning. Even greater advances, such as the continually operating receiver station (CORS) and the discoveries associated with that system, will be discussed at length later in this book.

All national mapping is also controlled by this geodetic network. The control network includes both horizontal control (geodetic positions) and vertical control (elevations). Any survey work that is performed as part of this network is automatically tied to any other work in the system, be it private or public surveys.

The procedure for establishing a state plane projection survey requires that the Land Surveyor recover one or more monuments of known geodetic position.[3] The latitude and longitude of the known location is “projected” to the flat state projection plane by a mathematical process. The location of that point on the projection plane is defined on the two-dimensional grid. The “x” value defines the distance east, whereas the “y” value defines the distance north. The process is reversible: If the “x” and “y” of a particular location are known on any state plane projection system, the corresponding geodetic latitude and longitude of that location can be computed.

---

[2] See Chapter 8.

[3] This process is being modernized greatly, as discussed in Chapter 8.

The Land Surveyor then measures the angles formed between land lines and the distances between land points in much the same way as in the case of the tangent plane projection. The major exception is that the computations systematically account for the curvature of the earth by converting level distances to grid distances and by defining all directions in terms of grid bearings. The adjustments to bearings and distances are “hidden” and usually not presented on the finished plat. The final map or plat produced will show the level ground distances measured, not the grid distances used during computations.

5.6. ADVANTAGES OF STATE PLANE PROJECTIONS

The advantages of a state plane projection system are numerous, if not apparent. The primary advantage is that, because surveys performed under these systems become, in a sense, a part of the national geodetic network, the boundary corners defined in the survey can be related to all other surveys in the system. Boundaries, therefore, need not rely solely on the recovery of marks within the survey in order for the Land Surveyor to learn of that parcel’s relationship to some other tract.

The recovery of boundary marks defined in state plane projections is greatly expedited by reducing considerably the area of search. This reduces the cost of resurveys, resubdivision, or recovery of existing parcels. Indeed, in areas where the use of state plane projection systems is widespread, even the cost of initial surveys is greatly reduced.

The information shown on adjacent tracts of land, and even remote tracts, can be related to the parcel surveyed. Surveys of adjacent properties will report (within the margin of error) the same bearings and distances for common lines. This makes the problem of discovering inconsistencies between adjacent tracts much simpler.

Municipal, county, and state-level projects and planning are more easily controlled and regulated. Highways, pipelines, utilities, and other public works are defined in relationship to every private parcel surveyed under the state plane system. This enhances the acquisition of rights-of-way by increasing the precision of the description of the areas acquired.

The Land Surveyor uses marks set by the NGS[4] or other agencies that are at published locations and known to all. Marks that define horizontal location are called "stations." Marks that define vertical locations (elevations) are called "benchmarks." Both stations and benchmarks are part of an international network of control points and define locations based on a particular datum.

In areas where the national control network is not well monumented,[5] the state plane projection system can still be effectively used by a simple modification. If knowledge of the exact latitude and longitude of a point is not available (usually because of lack of monumentation), the approximate latitude and longitude, perhaps scaled from a U.S. quadrangle map[6] or provided by a simple navigational GPS receiver, are all that is required to convert astronomic directions in a particular area to precise grid directions. All the adjustments of the state plane system can be applied, and only the exact "x" and "y" of the survey points are excluded from the computations.

5.7. DISADVANTAGES OF THE STATE PLANE PROJECTION

The lack of control stations or monuments of known "x" and "y" in the immediate area of the survey may require extensive work beyond the area of the survey simply to make the "tie to the network" necessary for complete implementation. This disadvantage is being quickly eliminated by the introduction of artificial satellites in position determination. The geodetic position of locations on the earth can now be determined very precisely by using relatively inexpensive radio receiving stations in remote areas. The perfection of this method of determining location will mean that the type of control required by state plane projection systems will be available everywhere.

The computations and adjustments required in state plane projection systems are slightly more complicated than those required in the tangent plane system. The curvature of the earth and the convergence of the meridians are accounted for instead of ignored.

The educational and training requirements of field personnel performing state plane projection surveys are more extensive. The field crews must possess or acquire knowledge of both the real property boundaries in the area of the survey and the geodetic positions of certain control points.

---

[4] For more information concerning the NGS control system, see www.ngs.noaa.gov.

[5] The latitude and longitude values obtained by navigational GPS units are accurate to within 30 meters or so and are not sufficiently precise for land surveys.

[6] For more information concerning USGS maps, see www.usgs.gov.

The additional training requirements and the additional fieldwork required sometimes result in the “first-time” survey costs of state plane projection surveys being more than those of tangent plane projection surveys of the same parcel. This is not true in cases where SPC-based surveys have been performed on adjacent tracts.

The few disadvantages cited indicate that, except for small urban parcels, the increase in costs for “first-time” boundary surveys are more than offset by the permanence, recoverability, and certainty of location that are rendered to real property boundaries documented by the state plane projection system.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 6

FUNDAMENTALS OF MEASUREMENTS

In the preceding chapters, we have made frequent reference to the *measurement* of angles and distances. It is very important to explain more fully the term *measurement*. Theoretically, there is only one true distance between two land points at a particular level surface. Likewise, there is only one acute angle formed between two rays or lines. The act of determining the values associated with these angles and distances is the act of measurement. Unfortunately, even the finest of measurements is only an estimate. Indeed, *all measurements are estimates!*

Measuring is not the same as counting. Counting is dependent on an indivisible fundamental unit, below which the thing or things being counted do not exist. The smallest unit of money is the penny. Even though we say that a dollar forms the basis of our monetary system, the penny is the smallest unit possible. Accountants do not measure the money in a bank; they count it. Empty your pockets and you either have 50 cents or you don't. Changes in quantity can only take place one penny at a time.

A great deal of time and effort has been and will continue to be expended by Land Surveyors in the pursuit of perfection in the *science* of geodetic measurements. Real property boundaries exist independently of measurements.

Boundaries are defined by the location of the real property corners.

Measurements of angles and distances are important because that is how we quantify, document, and describe boundaries. We will explore the procedures and histories of boundary-related geodetic measurements, but it is important to keep in mind that, while precise measurements are a very important part of boundary documentation, measurements *do not* define boundaries.

6.1. ACCURACY AND PRECISION

The terms *accuracy* and *precision,* when referring to measured dimensions, are not synonymous. "Accuracy" is an evaluation of the difference between a measured value and the "true" value. "Precision" is an evaluation of the procedure used to arrive at a particular value.

For example, assume that the distance between two particular points is exactly 500 feet. Two men are requested to determine the distance by whatever method they see fit. Mr. Jones stands over one point and, peering at the other, announces, "500 feet." Mr. Smith acquires a 100-foot measuring tape and carefully measures the distance several times and declares it to be 500.02 feet. Purely by luck, Mr. Jones's estimate was more accurate. Mr. Smith's estimate, because of procedure, was more precise.

6.2. IMPLIED PRECISION

Measurement data are usually presented by parameter (feet, miles, degrees, etc.) and numerical values. The form that the data are presented in contains a clue to the precision of the measurement procedure used to arrive at the value. This is only a clue, for unsophisticated reporters of measurement information can unwittingly change the implied precision of a reported measurement.

The nature of the parameter is important when implied precision is analyzed. A distance given in paces implies a precision far below that of a distance given in inches. A direction given in "points" of the compass (i.e., northeast) implies a precision below that of a direction given in degrees.

The division of the parameter is also significant, for it contains information about the smallest unit of measure used. A dimension reported as 22 feet implies that a foot was the smallest unit of measure. A dimension reported as 22 feet, 0 inches implies that an inch was the smallest unit of measure. The last figure given is said to be the "doubtful" figure in a measurement process. If a measurement were reported only to the nearest foot, then distances from just over 21 feet, 6 inches to just under 22 feet, 6 inches would all be reported as 22 feet. If measurements were reported to the nearest inch, then distances from just over 21 feet, $11^{1}/_{2}$ inches to just under 22 feet, $0^{1}/_{2}$ inches would be reported as 22 feet, 0 inches. The implied precision concept classifies 22 feet and 22 feet, 0 inches as two different measurements!

Fractions form another category under the implied precision rules. "$^{1}/_{2}$ foot" is not the same as "6 inches," nor is either one the same as "0.50 feet." The implied precision of $^{1}/_{2}$ foot is that the smallest unit of measure was $^{1}/_{2}$ of a foot. The implied precision of 6 inches is that the smallest unit of measure was an inch. The implied precision of 0.50 feet is that the smallest unit of measure was one-hundredth of a foot. One-fourth of a degree is not the same as 15 minutes, and 1 acre is not the same as 43,560 square feet when we are considering implied precision! Many a realtor has multiplied an area given in acres by 43,560 (square feet in an acre) to arrive at a square footage in order to determine a sale value. If an area is given in even acres, the square footage, under the implied precision rule, could be off by 21,780 square feet either way![1]

6.3. ERRORS

Measuring is typically dependent on the comparison of the thing being measured to a standard or known value. Variations in the accuracy of the standard value, the precision of the limits, as well as a host of other factors, introduce variables into the measurement process. These variables are called "errors." The word *error* has a very different meaning to the Land Surveyor or scientist than it does to the layperson. Measurement errors are a function of the level of refinement of the instruments, procedure and personnel conducting the work. Training, attention to detail, quality instrumentation, and a well-analyzed course of action will reduce the magnitude of not only individual measurements, but of the sum of the errors present in every survey.

There are three general categories of errors: systematic, random, and blunders.

---

[1] For a more detailed analysis on the theory of measurement, the reader is referred to R. B. Buckner, *Surveying Measurements and Their Analysis* (Rancho Cordova, CA: Landmark Enterprises, 1983); and E. M. Mikhail and G. Gracie, *Analysis and Adjustment of Survey Measurements* (New York: Van Nostrand Reinhold, 1981).

6.3.1. Systematic Errors

Systematic errors are inaccuracies in measurement that occur in the same direction every time that a measurement takes place. If, for example, a ruler were a fraction of an inch shorter than the length shown on it, then every measurement using that ruler would be off by the same amount and would mislead the user in the same direction. Objects would consistently be reported as longer than they really were.

6.3.2. Random Errors

Random errors are inaccuracies in measurement that occur in varying directions and magnitudes. Every time a carpenter measures a cut line, he or she must estimate the alignment of the end of the ruler with the end of the wood and the alignment of the marking pencil with the desired measurement. Each time these estimates are made, the carpenter might be off to the right one time and off to the left another. Random errors tend to be compensating; that is, they tend to cancel one another out when measurements are repeated several times.

6.3.3. Blunders

Blunders are not really errors in the strict scientific interpretation of the word. A blunder is the gross misinterpretation of a measurement due to a careless or mistaken observation. The carpenter meant to cut a 4-foot plank but marked the cut line at 3 feet. Blunders are mistakes. One thing was intended, but another was done. “Measure twice, cut once” is the old saying that addresses a type of blunder.

Blunders occur randomly, are usually large, and are rarely compensating. Blunders are what typically cause conflicts, confusion, and economic loss in the survey profession. Blunders are readily eliminated by the adoption of rigorous measurement procedures.

6.4. ALL MEASUREMENTS INCLUDE ERRORS

All measurements include errors. Imagine the following experiment. A classroom of high school students was instructed to attempt to measure the teacher’s desk using a feed store yardstick. The teacher requested that each student measure each dimension to the smallest fraction of an inch that he or she could estimate. As one would reasonably expect, some variations in the reported dimensions of the desk were reported.

The values for the width, for example, might have ranged from $40^1/_{32}$ to $40^7/_{32}$ inches. The average width reported would then have been $40^1/_8$ inches. One might report that the width of the desk was $40^1/_8$ inches, with a measurement error (uncertainty) of plus or minus $^3/_{32}$ of an inch.

The true width of the desk is not known and never will be known. The absolute, exact distance from the first molecule of desk on one end to the last molecule of desk at the other end is much finer than $^1/_{32}$ of an inch. The feed store yardstick is not a calibrated standard, and changes in humidity and temperature affect the length of the yardstick and the width of the desk. The variety of skill possessed by the many different measurers also influenced the results. The difference between the absolute width of the desk and the reported average width is the measurement error.

The students were engaged in a direct measurement. The standard (the yardstick) was applied directly to the desk. The beginning and end points were the limits of the value sought. This is the simplest act of measuring a quantity. If the students wished to reach a more refined value, they might have recorded humidity, pressure, temperature, and other factors and examined how variations in these factors affected the length of the yardstick. The students might have compared the length of the yardstick with a calibrated standard so that corrections for imperfections in the yardstick could be accounted for.

The problem of estimating the fractions of an inch might have been simplified by the addition of a vernier[2] or some other device that reduces the amount of interpretation involved. These are things that increase the precision of the measurement of the desk, but the fact remains that the measurement will never be absolute.

No matter how refined the process is there will remain a range of values that can result from correct and legitimate application of any measurement procedure.

6.5. REDUCTION OF ERRORS

The science of measurement is the study of methods or procedures that eliminate blunders, account for systematic errors, and reduce the effect of random errors.

---

[2] An auxiliary scale, mounted on the primary scale or ruler, that permits a direct reading of fractions (usually tenths) between divisions marked on the primary scale.

Some of the methods by which blunders can be eliminated are by repeating measurements, by careful observation, or by analysis of the results.

Some of the ways in which systematic errors can be accounted for include standardization of the equipment used, understanding the effects of the environment, and analysis of the results. Systematic errors must be detectable before methods to account for them are possible.

Some of the ways by which the effects of random errors are reduced include refinement of instrumentation (less estimation in interpreting readings), multiple observations, and analysis of results. How this is done in the modern land survey will be discussed later.

6.6. DEVELOPMENT OF STANDARD PROCEDURES

The history of land measurements is as ancient as the human race. The developments of methods of identifying locations discussed earlier were accompanied by attempts to perfect the measurement of angles and distances. Ancient societies that developed a centralized government soon employed people whose task it was to measure the area under the control of the central government. The earliest reports of these specialists come from Babylon, where pacing was extensively used to measure long distances. These professional pacers would report the distances between towns so that the king might know how many days it would take for his army to travel from place to place.

The Egyptians, long noted as master builders, used ropes of known length that were knotted at regular intervals in order to measure the fields and to lay out the great construction projects for which they are so well known. Many of the words we use today originate from the methods used to measure the land. The word *mile* is derived from the Latin *milia,* meaning “a thousand.” The Roman soldiers would count each time the right foot struck the ground while marching. Each thousand paces was a milia. The average man today will cover about 5 feet for each pace (two steps), or 5,000 feet for a “milia.”

Whether counting paces, turns of a wheel, camel paces, or any of a number of innovative means of measuring distances, the problem of a lack of a standard and inconsistency of results was a constant problem. The development of the Gunter’s chain, by Edmund Gunter circa 1620, was the first real advancement in the measurement of horizontal land distances. The concept of standard length, horizontal measure, correction for temperature, and sag quickly followed this development. The Gunter's chain remained state of the art for measuring horizontal distances in land boundary surveys until the twentieth century.

The modern surveyor's steel measuring tape, still called a "chain," improved only slightly the measuring process and accuracy. It was not until the 1970s that the advent of the electronic distance measuring instrument significantly improved the accuracy of distance measurement. Global Positioning System (GPS) equipment (detailed in Chapter 8) is just another means by which distances and directions are measured. GPS-derived distances and directions have one great advantage. The locations that are being measured do not need to be intervisible.

Even with these new and fantastically accurate devices, there still exists a measurement error that cannot be eliminated. It is for this reason that many states have developed a set of standard survey criteria listing the size of error acceptable for various grades of survey work. Table A.3 in the appendix of tables is an example of a typical set of measurement standards. As the science of measuring the earth advances, the standards by which the Land Surveyor must abide become increasingly stringent. However, there will always be a "margin of error" in every boundary survey.

Unlike distances, the measurement of angles developed to a high precision very early in history. The capability of the modern Land Surveyor to measure angles has improved only slightly during the latter part of this century. Although the Land Surveyor of years ago probably did not avail himself of the most sophisticated angular measurement devices possible, the capability was there. Indeed, land surveying angles were rarely measured to any great precision until the art of measuring distances improved to the point where real precision was possible.

Yet, even when calibrated metal tapes are applied directly between boundary markers and all of the appropriate corrections are made, there is a range of values that will be observed. The conditions that affect this range are sometimes beyond the control of the observer. Size, shape, and the nature and condition of the end points of the land line segment being measured are significant factors.

If the northeast corner of farmer Jones's property is marked by a six-inch square concrete post, four feet of which is above the ground and leaning southwest several inches, while the northwest corner is marked by a four-inch diameter post, also several inches out of plumb, then even the most carefully applied measurement procedures are not going to be more precise than plus or minus a few inches. Before beginning the measurement, the Land Surveyor has to estimate the upright location of each corner and the center point of that upright location. The condition of the corner monuments are not such that a more precise measurement can be made.

The measurement of distance is not the only kind of measurement required in real property surveys. The angle formed at the corners of property parcels are also measured. The condition of the end points, uncertainty of location, and limitations of the instruments involved mean that the angles measured, like the distances, have inherent measurement errors that must be understood and accounted for.

Boundary lines are, more often than not, subject to acts of possession or acquiescence. Fences, hedges, tree lines, or other physical obstructions frequently occupy the full length of a real property boundary. These features, while clearly indicating the general location of the boundary, make the direct measurement of the distance between corners quite difficult. The vast majority of boundary dimensions in the world are the result of indirect measurement. Dimensions are almost always the result of computations based on several measurements of angles and distances. In light of this, it is a constant source of wonder that so many real property boundary dimensions are the same from survey to survey.

6.7. COLONIAL PERIOD

Perhaps the best way to clarify the impact of measurement methods, procedure, accuracy, and reliability on modern real property parcels is by reviewing how the procedure and precision of the past compare with those today. For the purpose of illustration, let us imagine that a particular boundary in colonial America had each end of the line marked by six-inch square stone monuments. The Land Surveyor in 1770 may have been required to produce a survey plat (Figure 6.1) to assist in the identification of the parcel to be transferred to private hands.

The Land Surveyor probably would have used the following method to determine the distance between the corner markers: Because the line in question is a boundary line, there would probably be a fence or a hedgerow along the entire length of the line. The Land Surveyor would, therefore, clear a line parallel with the boundary at some convenient distance from the actual line. Temporary wooden pegs might have been set at each end of the cleared line opposite the stone corner markers. These pegs would have marked the ends of the segment to be directly measured.

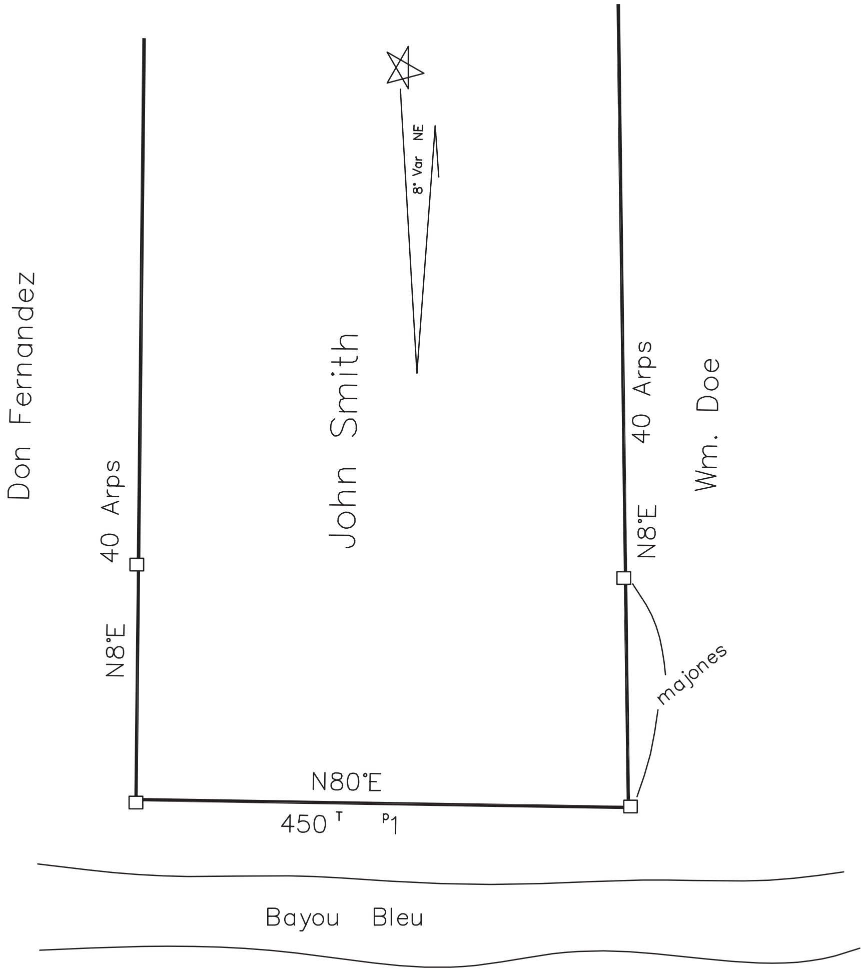
Don Fernandez
40 Arps
N8°E
John Smith
8° Var NE
40 Arps
N8°E
Wm. Doe
majones
N80°E
450 T P1
Bayou Bleu

FIGURE 6.1

Although a Gunter's chain might have been available, a device called a "compass" or a "toise" might have been used to measure the distance between the pegs of the offset line. Figure 6.2 provides a sample of how a toise might have been constructed. Depending on where in colonial America the land was located, the dimensions of the toise would have varied from about 6 to 6.4 feet. The device would usually have been made of wood by the Land Surveyor himself. The points might have been metal tipped or simply sharpened wood.[3]

---

[3] In other areas, a pole or a rod 16½ feet long and possibly capped with brass ends was commonly used to measure distances.

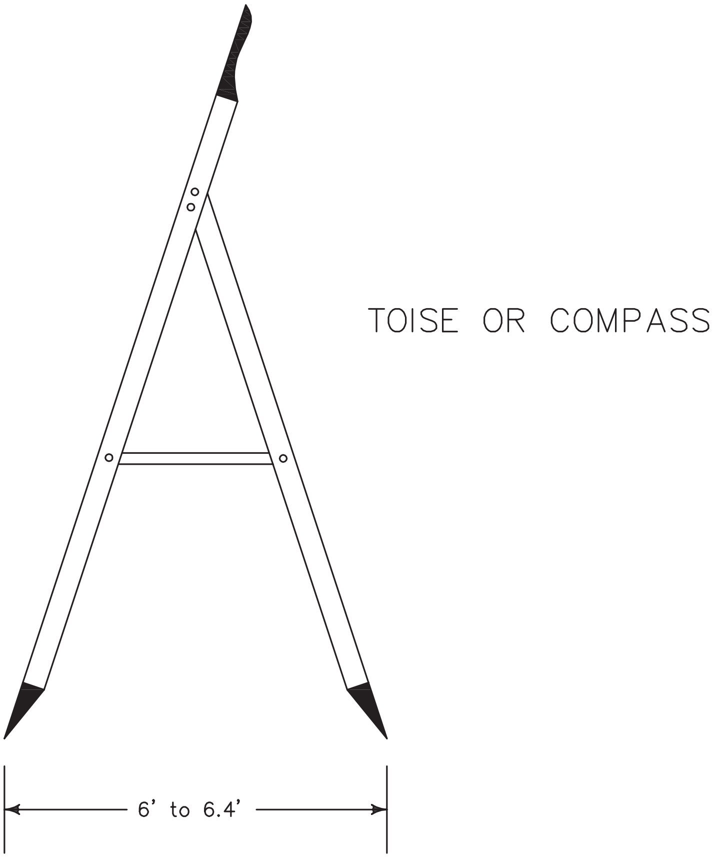

FIGURE 6.2

The Land Surveyor would have begun the measurement by placing one end of the toise at a peg and the other end of the toise on the ground in line with the far peg. The device would then have been rotated about the leading tip so that the trailing tip would be brought to the leading position and in line with the far peg. The process would be repeated, alternating the tips of the toise until the far peg was less than one rotation from the leading tip. This remaining portion would be measured with a yardstick, a foot rule or simply estimated.

An examination of this process reveals several sources of measurement error or inconsistencies. Any slight failure of the toise to be exactly the correct length would be repeated for every measurement. If, for instance, the distance between the tips were an eighth of an inch shorter than the six feet it was believed to be, then the surveyor would have reported a mile (5,280 feet) measured when only 5,270 feet, 10 inches had actually been traversed.

Each rotation of the toise would have to have been without any slippage. Any movement of the tips during the act of transferring the trailing tip to the leading position would introduce error. The work would often be done from horseback, adding another factor to the difficulty involved in simply rotating the device. Each tip would also have to be set perfectly in line with the end pegs, or additional error would be introduced.

The toise would have been set directly against the surface of the earth, so every rise and fall in the land would have added apparent distance to the measurement. The Land Surveyor would sometimes have attempted to account for this increase in distance by estimating the amount of extra turns the changes in elevation caused. The practice of "adding one for good measure"—that is, deliberately adding an unreported turn on the toise or adding one more length of rod to prevent "shorting" the buyer—was very common, especially in hilly country.

The beginning peg and ending peg would have been set in locations estimated to be directly across from the end points. This estimation would introduce error that could be quite large if the offset line were some distance from the boundary line. The Land Surveyor would have recognized that the measurements made by him were far from perfect, and so the slight error introduced by estimating the offset peg locations would be considered acceptable. In cases of original surveys, meaning surveys that created new parcels, the lines run might have been the actual boundaries.

In measuring one mile, the Land Surveyor would have to have made 880 rotations of a six-foot toise. It would have been highly possible, even probable, that occasionally a miscount might have occurred. Often, lines were measured several times to avoid miscount blunders, but just as often they were not. This factor alone introduced considerable doubt about just how much reliance could be placed on surveys conducted in this manner.

The direction of the boundary line would be determined by placing a magnetic compass at one offset peg and sighting the other. The bearing of each boundary line would be determined in the same way. The angles formed at the corners would not be measured in most cases.

The preceding scenario assumes that the end points or offset end points were intervisible. If the intervention of hills or woods made it impossible to see one corner post from the other, the problem of aligning the route of the offset line would be much more complicated.

In this case, a magnetic compass would be used to maintain line as well as to determine direction.

If the end points were not intervisible, the Land Surveyor would set a magnetic compass over one offset peg and send a rodman as far ahead on the line as could be seen, using the compass bearing believed correct for that boundary. The distance to the rodman would be measured as before, and the compass would be brought ahead to the rodman. The rodman would then be sent ahead again and the process repeated until the end peg was reached. The distance from the offset line at the end of the segment being measured would be noted, and corrections to the reported bearing would be made if the end offset was significantly different than the starting offset.

If the work being done was the creation of a new parcel, then the actual boundaries may have been traversed in a consecutive sequence such that the Land Surveyor would begin the work at one corner (the point-of-beginning) and measure the bearing and distance of each side, in order, returning to the point-of-beginning. If the new parcel was to be *a specific size* then one additional step would be added.

Because the new parcel being created represented a bounded area, the Land Surveyor would know that, if the measurements were perfect, the last line measured would return to the point-of-beginning. Of course, perfect measurements are impossible, so the last line measured would never return to the exact point-of-beginning. The most common method used to address this problem would have been the “field correction.”

After the Land Surveyor had satisfied himself that some gross blunder had not occurred, the field correction would be applied. This correction usually would consist of adjusting the location of one or more of the survey marks until the *distances* called for were satisfied. In Figure 6.3, the attempt to return to the point-of-beginning resulted in arriving at the point labeled “end.” In this case, the Land Surveyor might have applied the field correction by simply moving the point labeled “3” northeast 25 feet.

Each Land Surveyor had his own preferred method of applying field corrections, and, in most cases, these corrections were applied based on the Land Surveyor’s judgment, not on any scientific analysis.

The procedures outlined were representative of the type of work that was done in the latter part of the 18th century. The actual practice in any particular part of the country varied greatly. The notes, letters, and drawings of the Land Surveyors of that period and jurisdiction would be the best source of information on actual practice in that area.

Plats often hold clues to the procedure used. Because of the uncertainties involved, the distances reported by these early Land Surveyors

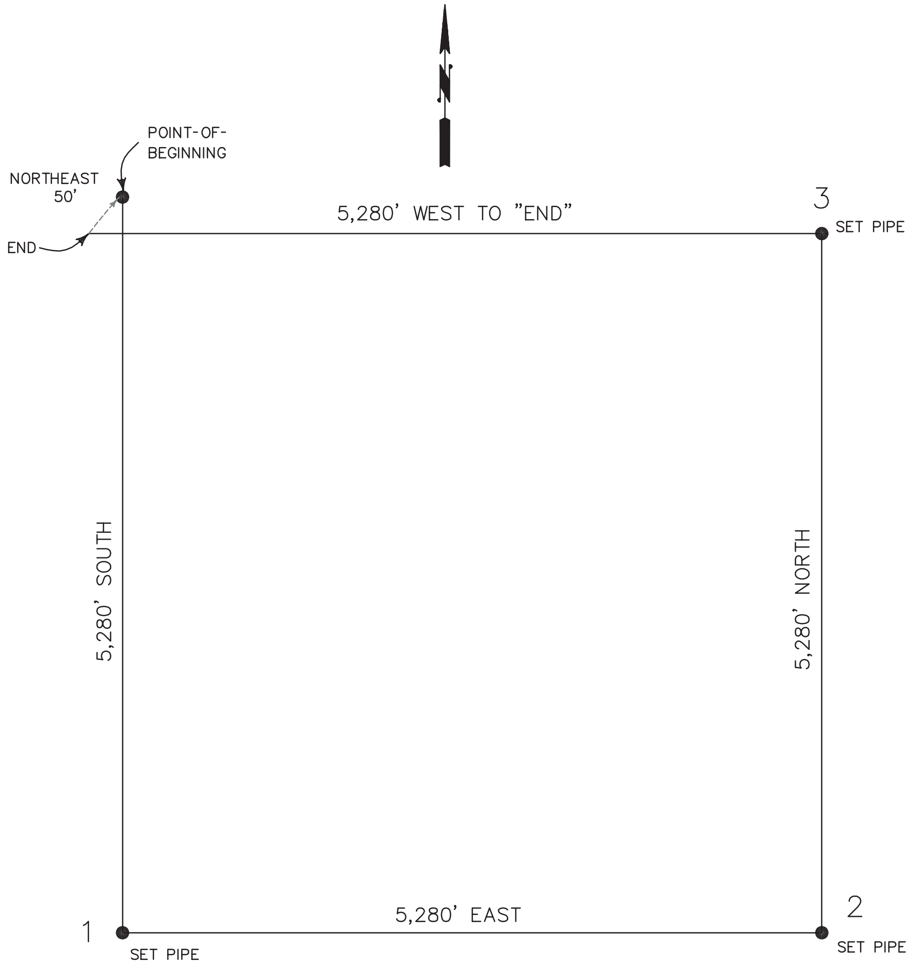
N
POINT-OF-BEGINNING
NORTHEAST 50'
END
5,280' WEST TO "END"
3
SET PIPE
5,280' SOUTH
5,280' NORTH
5,280' EAST
1
SET PIPE
2
SET PIPE

FIGURE 6.3

could not be expected to be more accurate than one foot in 100. The bearings shown are even less reliable because of the lack of precise sighting devices and reliance on magnetic north.

The concepts of errors in measured values were well known in the 18th century. The degree to which the methods of accounting for these errors were applied is a reflection of land values, craftsmanship, and skill or sophistication. Multiple observations might have been used to eliminate blunders. Skilled craftsmanship could have reduced the magnitude of systematic errors. But the greatest tool in the detection of errors, the analysis of results, was not very effective in the 18th century because the measurements of direction were so crude compared with the measurements of distances.

If one were to measure a closed area, such as a parcel of land, a mathematical analysis of the distances and directions reported could determine if the values presented were consistent with each other. During the 18th century, the crudeness of the directions observed, the laboriousness of the computations, and the low value of the land meant that these methods were rarely applied.

6.8. POST–CIVIL WAR PERIOD

If the imaginary boundary shown in Figure 6.1 were resurveyed a century later, there would have been only a few changes in the survey procedure. The existence of a fence line or a tree line along the boundary would have still required the use of the offset survey line. Pegs would have been set at estimated right angles to the boundary line a short distance from the stones just as 100 years ago.

The Land Surveyor of 1870 probably would have used a variation of the Gunter's chain, called the "two-pole chain," to measure the distance between the end pegs. The Gunter's chain was an iron chain of 100 links that was 66 feet long. Figure 6.4 is a detail of a portion of a Gunter's chain.

The figure is based on an antique that had tags on each link, which was not a common practice. Link tags were more commonly set at 10-link intervals or more. The two-pole chain was a shortened version of the Gunter's chain and was 33 feet long and had 50 links. A two-pole chain was laid out twice to measure one "chain" of distance.

The measurement of the distance began in the same way as the 1770 survey. One end of the chain was held over the starting peg, and the chain was stretched ahead, supported at the ends only. If the ground was

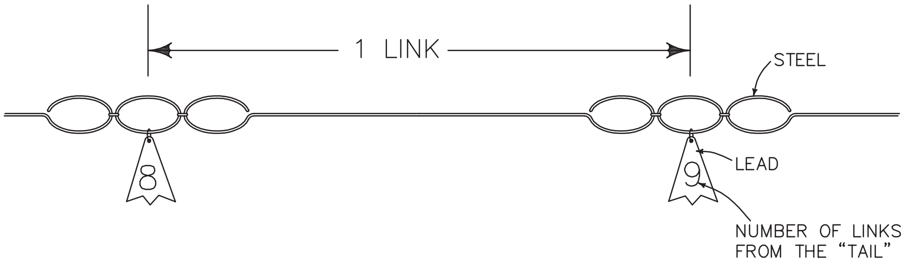

FIGURE 6.4

sloping, the ends were held such that the ends of the chain would be at roughly the same elevation. This was done so that only the horizontal distance was measured. The end of the chain was marked by a pin, and the chain was moved forward. Alignment was maintained by sighting the end points, if possible, or by sighting the bearing of the line with a magnetic compass.

The common sources of errors in the process were well known, and some attempt at compensating for these errors was often made. The chain length was subject to wear because the chain links rubbed together. Often, a standard chain was kept at the home office for comparison with the field chains. The sag of the chain supported at the ends varied with the force exerted by the chainmen. The marking of the chain ends by pins sometimes included guesswork as to where the ends actually were, especially in rough terrain. Failure to measure in a straight line also introduced error.

The sighting devices on the Land Surveyor's compasses were improved, and the introduction of the solar compass slightly reduced some of the uncertainty associated with magnetic directions. Generally speaking, the state of the art had advanced to the point where distances measured were accurate to one foot in 500. Directions were probably accurate to one degree, although failure to account correctly for magnetic deviations still rendered this part of the Land Surveyor's work the least accurate.

The slight improvement in determining directions also improved the effectiveness of analysis of results in detecting errors. Although mathematical analysis was possible, time, cost, and the specter of laborious computations meant that any analysis of results usually took the form of plotting the parcel to scale to determine if it "closed." The use of mathematical analysis did begin to increase with the approach of the 20th century. This is not to say that rigorous mathematical analysis of survey results did not take place prior to the 20th century; only that it was not commonly used. Figure 6.5 shows how that same parcel might have been platted in 1870.

6.9. BEGINNING OF THE MODERN PERIOD

By 1920, the introduction of the Surveyor's Transit[4] into everyday survey work had corrected some of the shortcomings that had hounded

---

[4] In America, this term is used to identify a theodolite that is read by use of a vernier scale and has a telescope that can be inverted (transited) without removing it from the instrument.

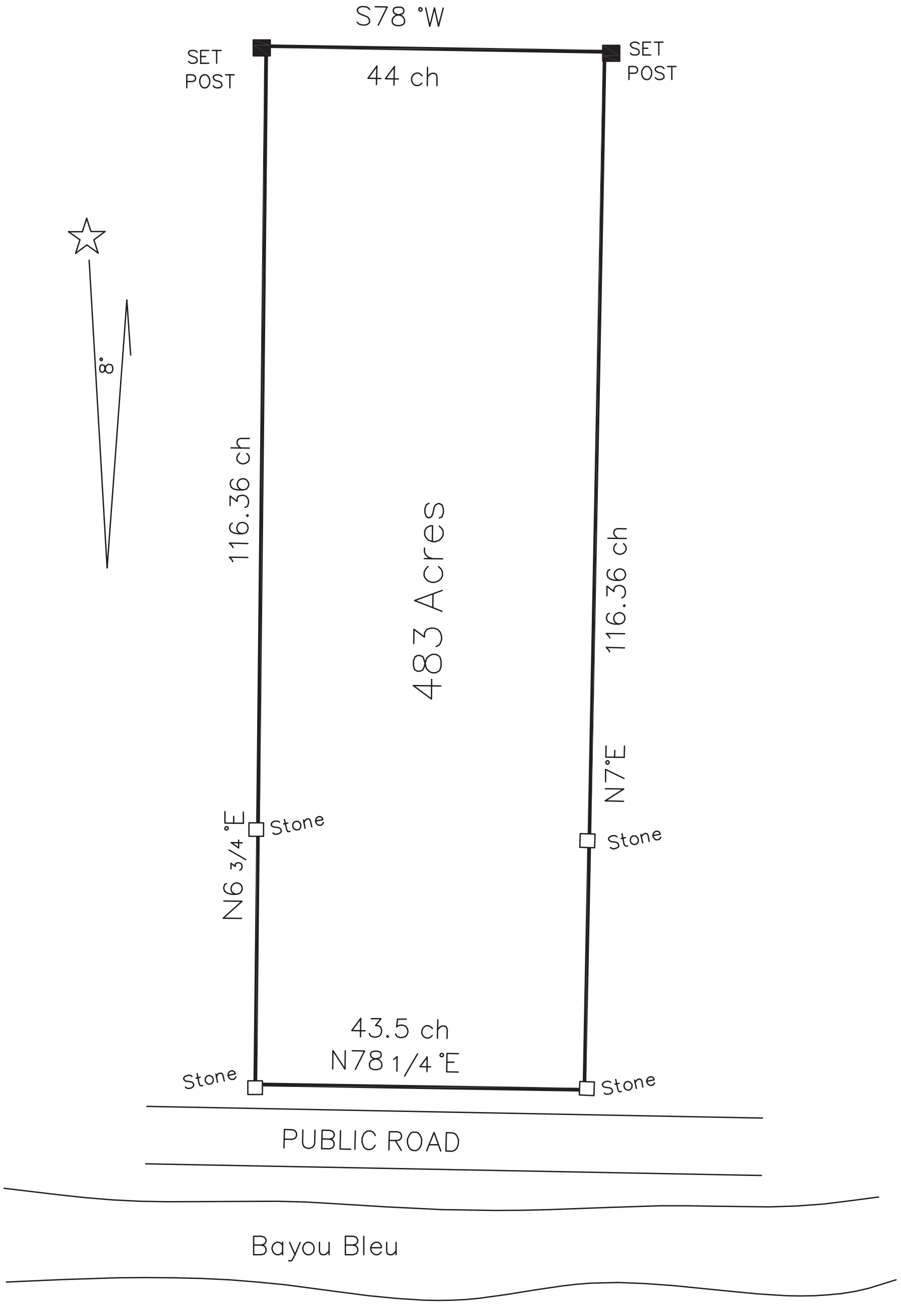
S78 °W
SET POST
44 ch
SET POST
8°
116.36 ch
483 Acres
116.36 ch
N7°E
N6 3/4 °E
Stone
Stone
43.5 ch
N78 1/4 °E
Stone
Stone
PUBLIC ROAD
Bayou Bleu

FIGURE 6.5

Land Surveyors in the past. Instead of observing independent bearings for every line of a boundary, the Land Surveyor now measured the angles *formed at the boundary corners*. This greatly increased the reliability of the reported shape of the parcel of land being surveyed, which greatly increased the accuracy of the acreage computed, but it did not improve the reliability of the magnetic direction used to orient the survey.

Distance measurement was much the same as in 1870, except that a steel Surveyor's Tape (still called a "chain") measured the distance in feet and decimals of a foot. The actual distance measured was still between offset points. The relationship of the offset points to the corner monumentation was now more accurately known because of the use of the transit to measure the angle to the actual corner. The steel tape was laid out and the lengths marked with chaining pins exactly as the two-pole chain had been used. Figure 6.6 shows how our imaginary parcel might have been platted in 1920.

The introduction of the transit enabled the Land Surveyor to measure accurately the angles formed by the intersecting boundaries of the parcel instead of simply measuring the independent bearings of each side. This increased the accuracy of the typical survey of the day to the point where deviations of less than one foot in 5,000 were common. For most work of the early twentieth century though, the accuracy of one foot in 1,000 for distances and one minute of arc for angles is about the norm.

The introduction of the transit to measure angles also meant that the mathematical analysis of the distances and angles measured became a very useful tool in determining the validity of the results of a survey. The concept of "closing a traverse" became widespread. If a parcel of land were measured by linking all of the lines around the parcel, then a Land Surveyor might compute the changes in distance and direction that took place if he or she were to travel from one corner to another in consecutive order. This "traverse" would return the Land Surveyor to the exact point-of-beginning if all of the angles and distances were error free (assuming a proper plane projection was used). The amount by which this traverse failed to return to the point-of-beginning would then be the sum of all the measurement errors: systematic, random, and blunder.

If a blunder occurred, then a large failure to close would have been noted (unless two blunders in opposite directions occurred, in which case one might hide the other). If the failure to close was within an acceptable range, then the Land Surveyor would "balance" the computations by simply distributing a "correction" to each angle and distance

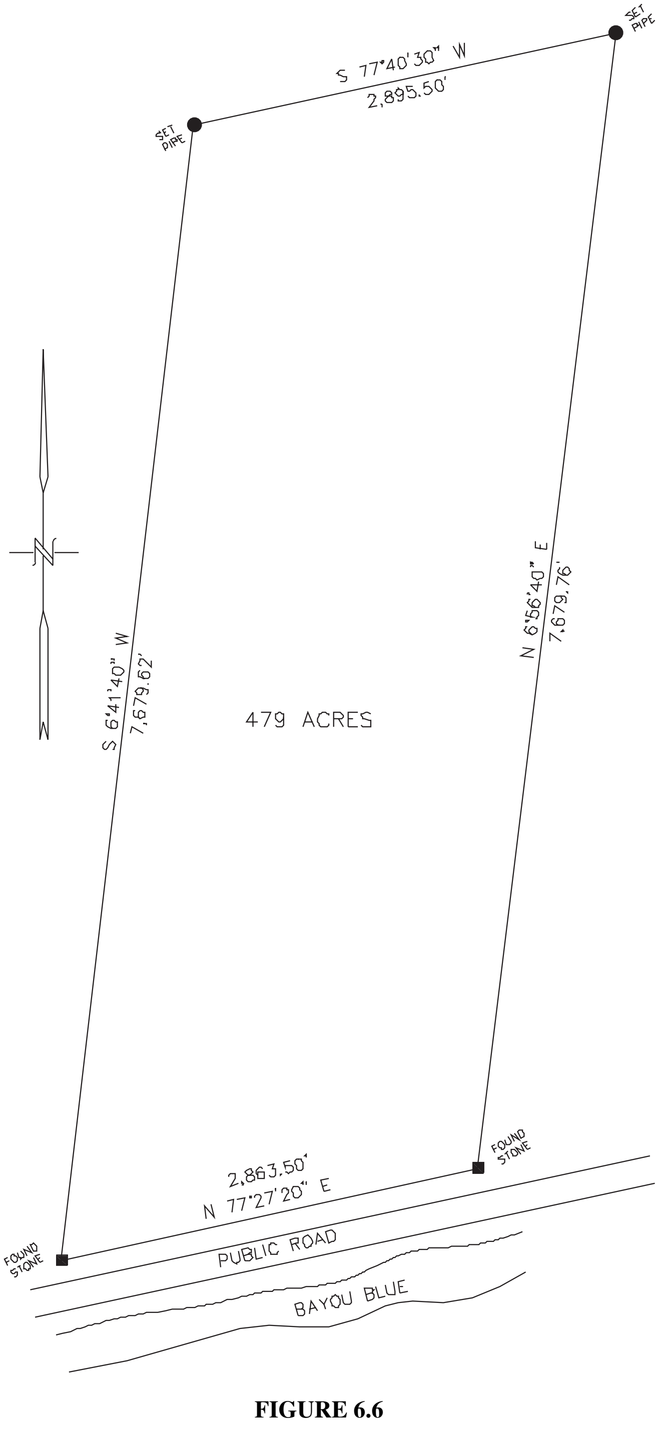
SET PIPE
S 77°40'30" W
2,895.50'
SET PIPE
S 6°41'40" W
7,679.62'
479 ACRES
N 6°56'40" E
7,679.76'
FOUND STONE
2,863.50'
N 77°27'20" E
FOUND STONE
PUBLIC ROAD
BAYOU BLUE

FIGURE 6.6

measured until the figure did "close" upon mathematical analysis. Variations of these methods of "balancing the traverse" are used today but with greater sophistication (we hope).

The most common methods of "balancing a traverse" generally available were, and still are, the compass rule, the transit rule, Crandall's rule, and the least squares adjustment. Notice that the first three methods are identified as rulesand the last is identified as an adjustment. All of these methods, with the exception of the least squares adjustment, require the use of a plane projection system, where each point can be defined by plane coordinates. Least squares adjustments can be applied to any set or group of measurements, regardless of the mapping or computational control used.

6.9.1. Compass Rule

The compass rule was, by far, the most popular method of balancing a traverse during the early and mid-20th century. It is still quite popular today and, under certain limited circumstances, is still a valid method.

In Figure 6.7, the measurements made by the Land Surveyor on the example triangular traverse are shown. It is known that the interior angles of a closed figure must add up to a particular value; in this case, 180 degrees. The sum of the measured values is 179 degrees, 58 minutes, 48 seconds, indicating an error of 1 minute, 12 seconds. The first step in applying the compass rule to this case would be to divide the total angular error by 3 and add this number to each of the three measured values. This adjustment of angular error is made *before* any other computations are made. Figure 6.8 shows the "adjusted" angles of the example traverse. Note that no correction is made to the angle from "north" to the line "A–B."

Using the adjusted angles, a plane projection bearing of each line is computed. These bearings are then used to compute the location of each of the corners based on the distance north (or south) from the point-of-beginning. This distance is sometimes called "latitude" when it is referring to the change from one point to the next. Similarly, the location of each comer based on the distance east (or west) is computed. This distance is sometimes called the "departure" when it is referring to the change from one point to the next.

All computations are two-dimensional and are made on the theoretical plane of the map projection, even though the three-dimensional terms of *latitude, longitude, north,* and *east,* among others, are frequently used.

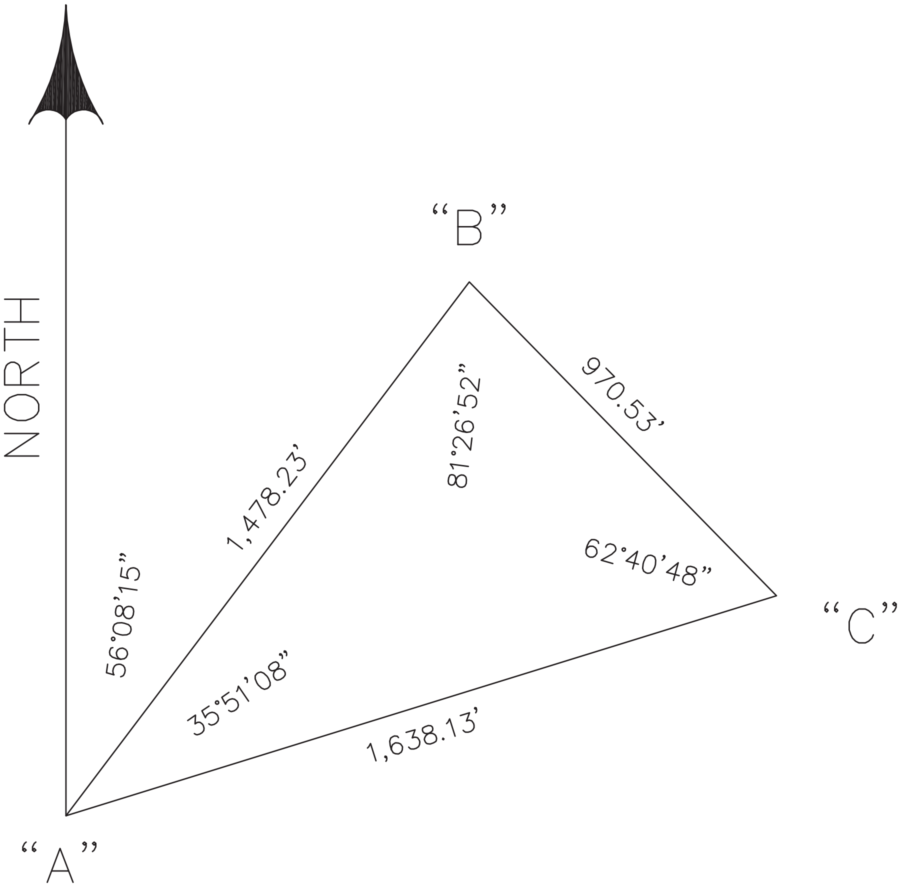

FIGURE 6.7

Consecutive computations of each point in the traverse will result in two sets of values for the point-of-beginning: the first set being the values assigned at the beginning of the computations and the second set being the values computed at the end. The difference between the original values and the computed values, or failure to close, has a north–south component and an east–west component. The compass rule states that a correction will be added to (or subtracted from) the computed coordinates of each point in proportion to the accumulated distance from the beginning, measured along each side, divided by the total distance around the traverse.

If, in the example traverse, point "A" were given the initial value of 0 feet north and 0 feet east, then the following table gives the initial computed values for the points shown, rounded off to the nearest one

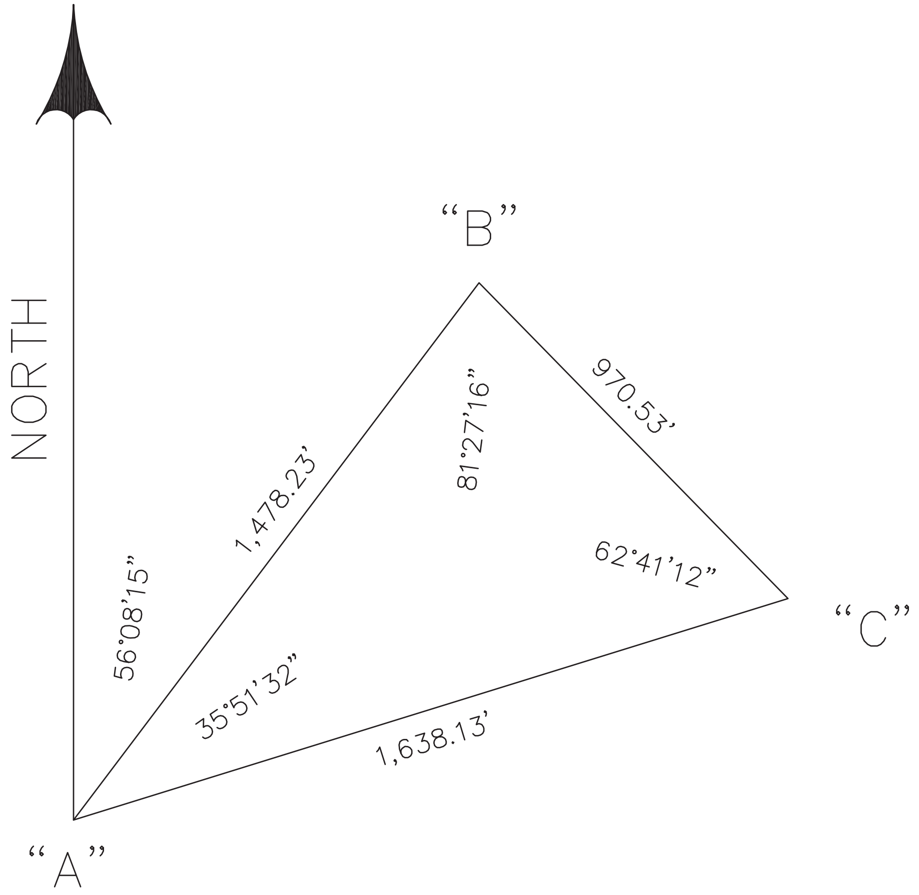
NORTH
"B"
"C"
"A"
1,478.23'
970.53'
1,638.13'
81°27'16"
62°41'12"
56°08'15"
35°51'32"

FIGURE 6.8

hundredth of a foot. A negative value in the "Feet North" column indicates south; in the "Feet East" column, it indicates west.

Point	Feet North	Feet East
"A"	0.00	0.00
"B"	823.67	1,227.49
"C"	−53.65	1,642.51
"A"	3.43	5.38

In order to "correct" the value for point "C," it is necessary to subtract 2.06 from the north coordinate [3.43 × (1,478.23 + 970.53)/(1,478.23 + 970.53 + 1,638.13)] and subtract 3.22 from the east coordinate [5.38 × (1,478.23 + 970.53)/(1,478.23 + 970.53 + 1,638.13)]. The

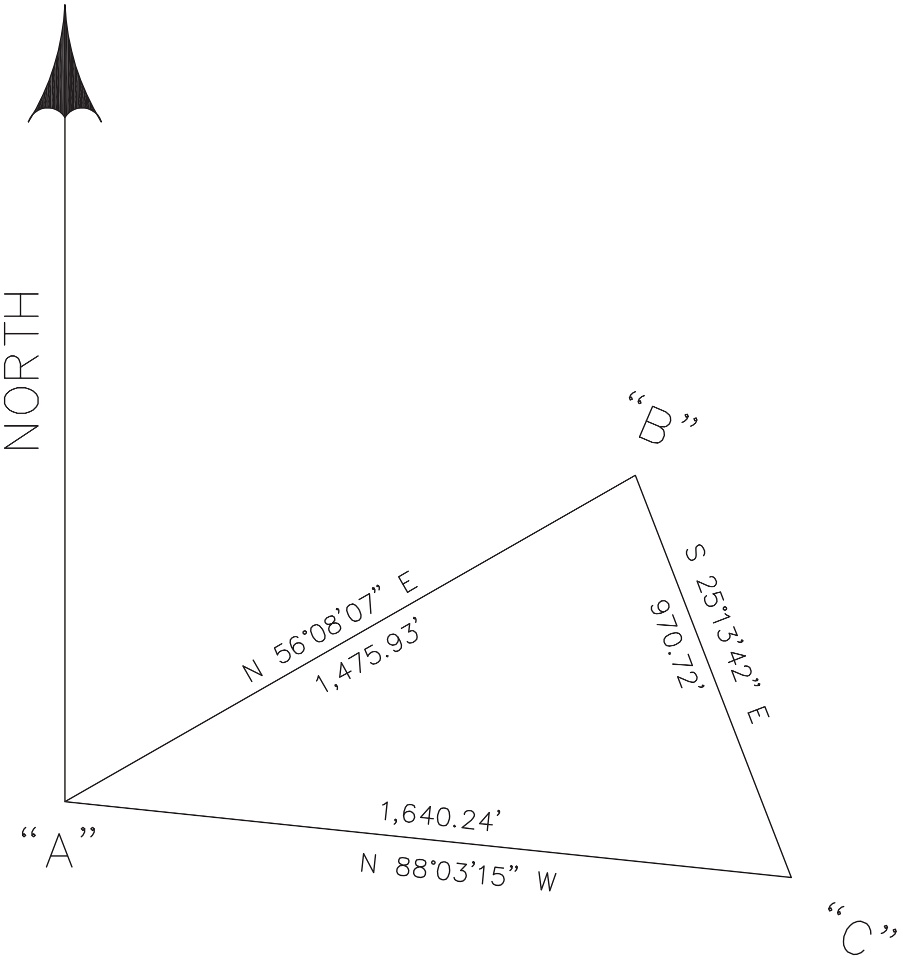

FIGURE 6.9 Adjusted Traverse

coordinate for "C" can be "adjusted" to north –55.71 feet and east 1,639.29 feet. Each point in the traverse is corrected in this manner.

Once the coordinates are adjusted, the distances and directions between each point can be *computed* based on the adjusted coordinate values. Figure 6.9 shows the example traverse after an application of the compass rule.

The *adjusted values* are used in all reports, other measurements, locations, and descriptions of the parcel being surveyed. Unlike the field correction, *no points are moved*. The only change that takes place is in the values assigned to the directions and distances of the lines.

One of the most common mistakes made in applying the compass rule is the failure to adjust the measured angles *before* the initial computation bearings and coordinates. This blunder, while resulting in an adjusted traverse, yields a different adjustment for the same traverse computed clockwise than counterclockwise. Another common misconception is that the compass rule adjusts the traverse based on the probability that certain measurements will have greater error involved than others. In truth, the compass rule, like all the "rules," only *makes things fit*.

6.9.2. Transit Rule

The transit rule also begins with an adjustment of the measured angles in the exact manner as the compass rule. Just as in the compass rule, the initial bearings of the sides of the traverse are computed from these adjusted angles. The coordinates of each point are computed using the measured distances and the adjusted bearings. The difference between the compass rule and the transit rule is in how the failure to close is proportioned to each coordinate.

The transit rule uses the ratio of the changes in north (or east to correct the east coordinate) for a particular line to the arithmetic total of the changes in north (or east for the east coordinate) and adjusts each latitude (or departure) of each line accordingly. This adjustment method was quickly recognized as one that produces inconsistent results, and so the use of this method has been, for the most part, discontinued.

6.9.3. Crandall's Rule

Crandall's rule was an attempt to apply the concepts of probability to error correction. According to this rule, errors in long distances are more probable than in short distances, and errors in distances in general are more probable than errors in angles. The application of this rule is a fairly complex computational procedure that essentially ignores angular errors and disproportionately adjusts distances. Crandall's rule almost always results in adjustments that are unnecessarily distorted, so the use of this procedure has generally been restricted to academicians.

6.9.4. Least Squares Adjustment

The least squares adjustment of a traverse is one of the many applications of the theory of least squares analysis of measured values. Unlike the rules cited earlier, the least squares adjustment is strictly based on the theory of the propagation of measurement errors. The theory states that, for any set of measured values, the best set of corrections to apply to the measured values is one such that the sum of the *squares* of all of the corrections is minimized. The least squares adjustment is the most commonly used method of adjusting measured values.

The simplest example of the least squares adjustment theory is the average. If the least squares theory is applied to a single set of things that are measured many times, calculus renders the arithmetic average as the least squares solution. If the distance between two points were measured many times, then the average (obtained by summing all of the measurements and dividing by the number of measurements made) would represent the most probable measured distance. Every person has, at one time or another, used a least squares adjustment without knowing it.

The least squares adjustment is applied directly to measured values and, in surveying, is simplest when the precision of angular measurements is comparable to the precision of the distance measurements. If the procedure used to measure the angles (distances) is much more precise than the procedure used to measure the distances (angles), then special steps within the least squares adjustment must be taken for the results to be valid.

Unlike the rules, the angles of the traverse are not adjusted prior to beginning the least squares procedure. Angles and distances are adjusted simultaneously, based on the theories of probability. The procedure renders consistent and reliable results in proportion to the quality of the measurements made.

6.10. MODERN PERIOD

From the mid-1970s to today, the increased use of electronic distance measuring devices (EDMs) has greatly increased the accuracy of the distances reported on modern surveys. Unlike the toise or the chain, the EDM does not rely on repetitively “laying out” a standard length. Most EDMs measure distances by emitting a laser light that is reflected back to the instrument by a special mirror. The distance to the reflector is determined by comparing the departing signal with the returning one.

The accumulation of error with every length of chain, so inherent with the old methods of measurement, is almost absent with the EDM. It is only necessary to be able to see from one end of the line segment being measured to the other in order to measure the distance.

The high precision of the EDM has matched the precision long possible in angular measurement. This "matching of precision" is perfectly suited to the least squares adjustment method of analyzing measured values. When the distance values were obtained by chaining, even when strict procedures were used, the angular values were usually more reliable. The Land Surveyor had to determine the relative degree of difference in reliability in order to use the least squares adjustment method properly. This difficulty, which discouraged the use of least squares, was eliminated by the EDM.

Distances of several thousand feet are measured as quickly and as easily as a few feet. This has also led to a more widespread acceptance and use of the state plane projection system by the Professional Land Surveyor. Full use of state plane system requires that the work be tied into control stations that may be miles from the job site. Before the accuracy and ease of long-distance measurement were provided by the EDM, one could hardly fault the Land Surveyor in private practice for opting not to spend several extra days on a job just to tie it to the state plane system.

GPS instrumentation has increased this long distance measurement capability to distances that are truly global. The near distance precision of GPS-based systems is marginally less than that of the EDM, but on the scale of miles it is unparalleled. This development alone has all but mandated the acceptance of state plane controlled surveys by the professional community.

The introduction of the computer has allowed the Land Surveyor to use formerly very cumbersome, but theoretically superior, computational methods of detecting error distribution and balancing traverses. Now, instead of using guesswork in identifying and correcting for errors, complex and sophisticated mathematical procedures are available to every Land Surveyor. The formerly time-consuming tasks of computing traverses, areas, distances, and a thousand other things are now performed in fractions of a second. This relief from tedium has allowed the Land Surveyor the opportunity to look at every problem from many different sides.

Modern surveys are performed with a degree of precision that was nearly impossible just a few decades ago. Accuracies of better than one in 70,000 are now commonplace. Angles and distances can be measured with equivalent degrees of accuracy. This results in a greater consistency or reliability of computed dimensions. This is perhaps best illustrated by the standard deviations for typical modern instrumentation listed in Table A.5 in the appendix of tables.

The future of survey measurements—that is, the tools and procedures—that will be available to the Land Surveyor are unlimited.

Already, unimagined combinations of many differing aspects of the electronic era are happening that effect many of the ways that data can be collected.

Three-dimensional digital scanning, which combines photography with special control and digital archiving, is only one of many promising new technologies. The collection of data is being automated to the point that it is almost overwhelming. The effort, considerations, and judgment that the Professional Land Surveyor must exercise today require an ever-increasing level of education and wisdom.

6.11. RANDOM TRAVERSE

In order to illustrate better the impact of these modern advances on the profession of surveying, let us recreate a typical boundary survey of a simple fictitious parcel of land. Figure 6.10 shows the corner monuments of the parcel in their relative positions. The methods used to recover these corners will be discussed in detail in Chapter 10. The corners are the limits of the parcel, and the task at hand is to *measure* the distances between the corners, the angles formed, and the area of the parcel. Knowing how this is typically done today will greatly enhance your understanding of boundary surveys.

Because the property involved is valuable and because of the expected use of the property, the Land Surveyor, in our example, has decided to use the local state plane coordinate system to control and map the survey The Land Surveyor may have previously established a "control traverse" between two National Geodetic Survey (NGS) stations, "Billy" and "Bob," as a part of another job. The latitude and longitude of "Billy" and "Bob" are published by NGS and are the results of measurements that have been *adjusted to conform* to the national network. The latitude and longitude of these stations have been mathematically translated to the appropriate state plane grid coordinates (x's and y's), called the "values" of the stations.

If NGS stations were not available in reasonable proximity to the property, the modern Land Surveyor would use GPS instruments and procedure to establish stations near the project that would be the foundation for the state plane referenced data collected during the survey.

The prior work was one of high precision and was rigorously adjusted to the national network. This means that the measurements of the traverse that were used to develop the "x" and "y" values for each traverse station were strictly controlled and that these values were *adjusted* to conform to the published values of "Billy" and "Bob."

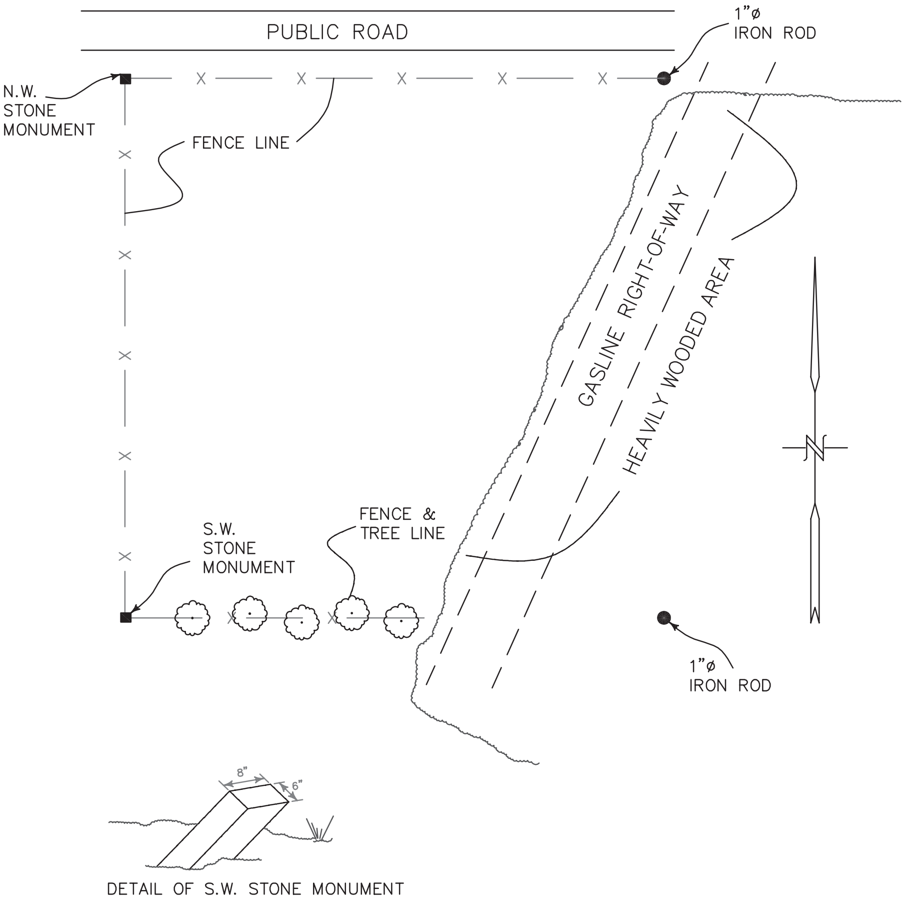
PUBLIC ROAD
1"ø
IRON ROD
N.W.
STONE
MONUMENT
FENCE LINE
GASLINE RIGHT-OF-WAY
HEAVILY WOODED AREA
N
FENCE &
TREE LINE
S.W.
STONE
MONUMENT
1"ø
IRON ROD
8"
6"
DETAIL OF S.W. STONE MONUMENT

FIGURE 6.10

Consequently, the distances and angles actually measured in the traverse are not exactly the same as the distances and angles indicated by the assigned values of the stations. This variation between "measured" and "reported" is very small because of the high precision involved. Figure 6.11 shows the relationship of this traverse to the parcel being surveyed. The traverse stations are often called "random," because they were chosen for reasons of convenience, observation, permanence, and other factors. Traverse stations do not directly represent real property boundary corners.

The land between the corner monuments consists of hedgerows, fence lines, heavy woods, hilly terrain, and other obstacles that prevent direct measurement between the corners. Therefore, the Land Surveyor must establish a secondary control traverse that takes advantage of

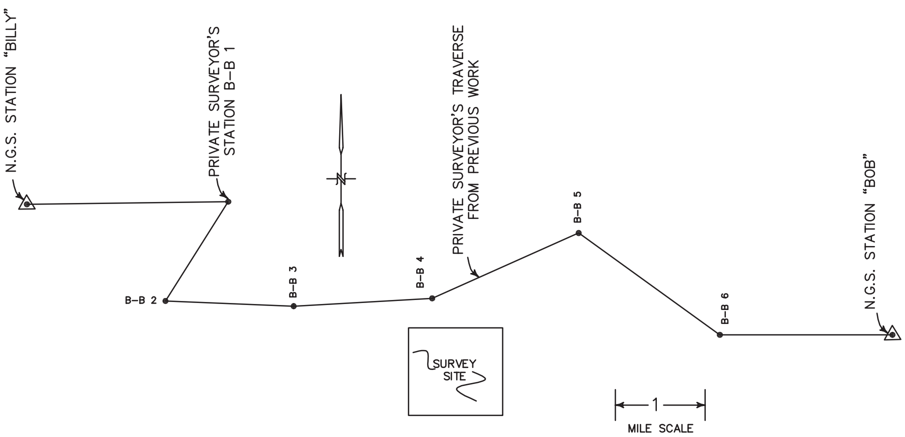

FIGURE 6.11 Random Traverse

roads, power lines, pipelines, cleared fields, or other features that allow for unobstructed lines of sight.

The points at which the secondary traverse changes direction are typically called "control stations" or "random stations." The locations of these control stations are deliberately chosen so that features that are to be recorded are near or, at least, visible from one or more control stations. Figure 6.12 shows the secondary control traverse for this example.

Angles 1 through 6 were measured using procedures and equipment that resulted in quantities that were expected to vary from the actual value less than 5 seconds either way. Distances 1 through 5 were measured using procedures and equipment that resulted in quantities that were accurate to within 1 in 70,000. The coordinate values of each of the secondary stations are then computed and adjusted using the least squares method of traverse adjustment to conform to the values of stations B–B 4 and B–B 5. Once again, the adjustment must be within certain limits to be consistent with the level of precision required.

The corner monuments are then tied into the secondary traverse by measuring the angle and distance from the secondary stations to each corner monument. These ties must be multiple and redundant to avoid the possibility of blunder. *Multiple* means that the angles and distances must be measured several times. *Redundant* means that separate sets of measurements must be made wherein any one of the sets is sufficient to compute a value for the corner. The coordinate value for each corner is computed separately for each set of measurements, and the results are compared for consistency. The results are usually combined and

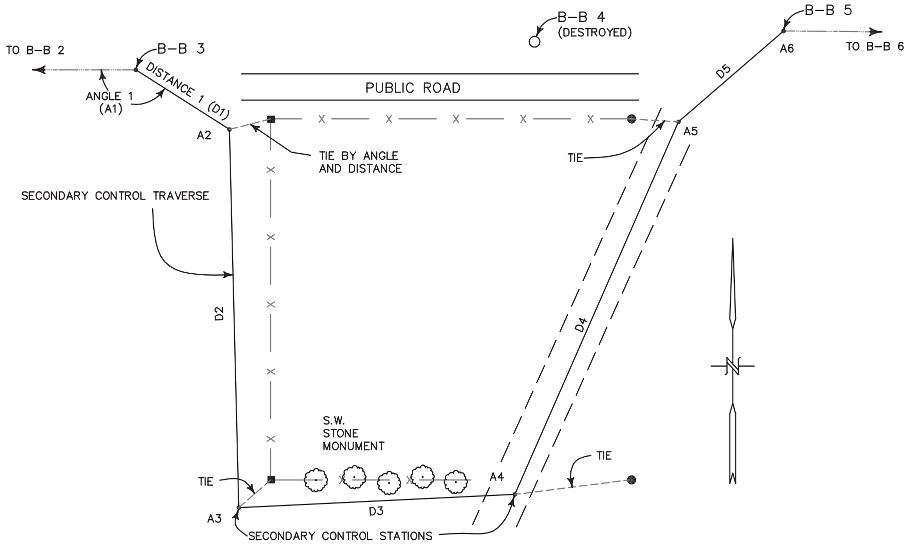
B–B 4
(DESTROYED)
B–B 5
TO B–B 2
B–B 3
A6
TO B–B 6
DISTANCE 1 (D1)
D5
ANGLE 1
(A1)
PUBLIC ROAD
A2
A5
TIE BY ANGLE
AND DISTANCE
TIE
SECONDARY CONTROL TRAVERSE
D2
D4
N
S.W.
STONE
MONUMENT
TIE
TIE
A4
A3
D3
SECONDARY CONTROL STATIONS

FIGURE 6.12 Secondary Traverse

averaged to produce the accepted coordinates for each of the corner monuments.

Once the coordinates for the corner monuments have been determined, it is a simple matter to *compute* the bearings and distances between each of the corners. The area of the parcel is also computed directly from the coordinates. The state plane projection requirements for accounting for the curvature of the earth and the conversion to and from grid distances are made during the process of the computations so that a map of the parcel can be produced that shows the grid bearings of the lines, the level surface distances between the corners, and the level surface area of the parcel.

It is important to note that the preceding example produced quantities for directions and distances that were based on a series of measurements that had been repeatedly adjusted in an attempt to account for errors. The condition of the corner stones was such that an interpretation was required to determine the former, upright location of the stones. If a second Land Surveyor were to clear the line between two of the stone corner monuments of the example parcel and directly measure the distance between them, the distance reported as a result of the direct measurement would be expected to vary slightly from the distance reported by the first Land Surveyor. Moreover, if a third Land Surveyor were to measure the same distance between the same two monuments, directly, the results of that third measurement would vary slightly from the others.

If anything is learned from this example, it is that angles and distances reported on all survey plats and maps are the results of measurements and are never absolutely "correct."

The modern survey, as in the preceding example, has reduced the uncertainties of the angles, distances, and areas reported by the Land Surveyor to a minimum. Compared to the methods used during the late 19th and early 20th centuries, when most of the titles of private property were developed, the modern survey is vastly more accurate and, consequently, more reliable. The laws governing real property transfers recognize the uncertainty of measurements made by Land Surveyors and operate effectively in spite of these ancient deficiencies.

This short history of measurement systems emphasizes the third,[5] and most compelling, reason for the canon that boundaries are defined by the location of the corners.

If boundaries were defined by directions and distances, every improvement in the science of measurements would move boundaries!

6.12. ELEVATIONS IN THE MODERN PERIOD

Land Surveyors have always been aware of the role that changes in elevations (vertical distances) play in measuring and mapping real property parcels. Until very recently, the interest in the elevation of a property parcel was rarely associated with a national datum. Landowners were interested only in relative elevations. Was the property hilly or flat? Was the land subject to frequent flooding? These and other questions relating to relative elevation were rarely part of the real property deed documentation and were usually reserved for the inquiry of sophisticated buyers.

The process used to measure the differences in elevation has changed little in the last century, in spite of all of the other advances in technology. The greatest change has been one of regulation, not science. Federal, state, and local regulatory authorities have intervened in the process of real estate transactions. Federal flood insurance programs, lending agency policies, and state and local regulations have added requirements for information that go far beyond the traditional documentation of real property parcels. The absolute elevation of structures relative to a specific datum is now a common requirement of many regulatory agencies.

---

[5] The first reason that corners define boundaries is ancient tradition. The second is the need for regular shapes on an irregular planet.

Before an agency can require elevation data pertaining to a real property parcel, that agency must establish or specify the datum to be used. The federally defined North American Vertical Datum of 1988 (NAVD 88) (see section 3.8) has become almost universal as the datum of choice for all regulatory bodies. Many regulatory agencies refer to NAVD (formerly National Geodetic Vertical Datum [NGVD]) elevations as a requirement without understanding the least bit what is involved. The process of measuring the NAVD value of the elevation of a structure and the foundation of that data need to be understood if that information is going to be used in a beneficial way.

In the past, the NGS had established a series of “bench marks,” which are simply specific locations that have a published elevation associated with them. A typical bench mark is a brass disk clamped onto the end of a very long rod (often 150 feet long). The rod is driven into the ground or otherwise fixed to the earth at what is hoped to be a stable site.

The differences in elevation between all of these bench marks are measured using differential leveling procedures that are a sophisticated version of the Land Surveyor’s ordinary level and leveling procedure. These differences in elevation are measured and remeasured frequently. New bench marks were constantly being added, and old bench marks are frequently being destroyed or disturbed. Furthermore, the entire network of bench marks had to be constantly reevaluated and updated.

In the near future, NGS will provide elevations through a network of continually operating receiver stations (CORS). The CORS network will be updated regularly and available to anyone with an Internet connection and a GPS receiver. These stations will eventually replace the bench mark system of establishing elevations. Local and project elevation control will still be supported by bench marks.

The most common method used by the private Land Surveyor in determining the NAVD 88 elevation of a location is differential leveling, using a Land Surveyor’s level and level rod. The modern Land Surveyor’s level is a device fixed with a telescopic sight that can be mounted on a tripod and adjusted so that the line of sight through the telescope at the center crosshair is perpendicular to the plumb line at the instrument. The level rod is a graduated ruler that, when set vertically on a surface, displays the vertical distances along the rod from that surface.

The process of differential leveling begins at a bench mark or location of known elevation. The level is set up a short distance from the bench mark (usually less than 150 feet), and the level rod is set vertically on the bench mark. The observer then peers through the level’s scope and records the value seen on the level rod at the horizontal line of a center crosshair. This value is the measured distance between the surface and the line of sight. If the earth were flat, the value observed on the rod would also be the difference between the elevation of the level and the elevation of the bench mark. The elevation of the level would equal the elevation of the bench mark plus the rod reading.[6]

The rod is then moved to a new location nearer the site to be surveyed (still less than 150 feet from the level) and set vertically on a new surface. The observer again records the value shown on the rod through the sight of the level. This second value is subtracted from the just determined elevation of the level, producing (theoretically) the elevation of the second surface.

The level is then moved ahead, and the process is repeated, using the newly determined elevation of the second surface or "turning point" to compute the elevation of the level at the new location. This process is repeated, in leapfrogging fashion, until the location to be measured is reached.

This laborious and tedious process of differential leveling requires at least 17 instrument setups to "carry" elevations one mile. The initial network of level runs needed by NGS to establish the federal bench mark system consisted of hundreds of thousands of miles!

Unfortunately, elevation information, so laboriously collected, is not permanent. Aside from the changes in elevations wrought by redefining the datum, many regions are experiencing vertical movement in the land itself. Heaving and subsidence of the earth's surface is a common phenomenon in many areas. The elevation of a building is not a constant value but changes with time. Even in areas where there is very little vertical movement, the significance of an elevation value will change.

Land development will change the storm water runoff patterns, often subjecting land to an increased frequency of flooding. The maps used by federal, state, and local agencies to determine flood hazard are often updated and revised, based on changes in the watershed. Elevation information must be current to be of any value.

6.13. FUTURE OF SURVEYING

Several decades ago, indeed within my professional lifetime, the science of measuring angles and distances on the surface of the earth was not much improved beyond what it was millennia ago. The measurement of angles was dependent on sighting distant objects and interpreting etchings on a circular plate. The measurement of distances was dependent upon sequentially placing an object of known length on the ground and summing the results.

---

[6] The world isn't flat, of course. For that and other reasons, the distance from the level to the rod is kept small. The myriad complications associated with precise differential leveling is far beyond the scope of this book.

The electronic revolution brought a gigantic improvement in not only the science of measurement, but the computations, record keeping, documentation, drafting, and every aspect of land surveying.

The Professional Land Surveyor of today has been empowered with measurement, computational, and recording tools that free him or her to commit more time to research, investigation, and the art of recovering real property boundaries.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 7

LAND RECORD SYSTEMS

Many territorial animals identify land parcels by marking the boundaries with their scent, scrapes, or other physical signs that tell other animals, "This is mine—keep out." The human race is no different from other territorial animals in this respect. The method of marking boundaries most frequently used by humans consists of marking boundary corners with physical objects, such as posts, rocks, pipes, or iron rods, and marking the lines between the corners with fences, ditches, roadways, or tree lines. This is fine for demonstrating to others the physical limits of a parcel of land. This is not adequate, however, when the need arises to define or describe that parcel to others who cannot visit each boundary corner.

The private ownership of land and the statutes that regulate the transfer of that ownership require that each and every separate parcel, public or private, be capable of being uniquely distinguished from all other parcels in writing. The systems of identification used in each state or region vary but usually take some form of one of three basic systems of identification for original tracts. They are the "metes and bounds" system, the rectangular United States Public Lands System, and the Platted Subdivision system.

In many places, or because of interstate ownership of property, variations of these systems are used simultaneously. Each system will be described in detail.

The first system to be employed in the United States was the "metes and bounds" system and is the basis for property identification in most areas settled and in private hands before the Revolutionary War. This system originated at the dawn of civilization and grew to favor in Europe during the 18th century, when the increase of privately owned (and taxed) land prompted governments to develop more accurate cadastre records. The metes and bounds system requires that privately owned land be identified by naming the adjoining owners and physical limits (bounds) of a property, and that a report of the dimensions (metes) be included in the description of the property.

The metes and bounds system of identifying or describing real property has often been erroneously referred to as the "legal description" when, in fact, any proper description based on any of the systems of identifying land parcels is a "legal" description.

The additional requirement that all land transactions be in writing and recorded with some governmental body to assist the local tax collector in developing a record of property owners (and tax debtors) meant that a general description of the property that simply named all adjoiners (bounds) was insufficient. The area (therefore the value) was a part of the record that was most important to the tax collector.

The second system, which affects the majority of the land area in the United States is the United States Public Land System (USPLS). The USPLS is a more regulated, formal system of identifying land parcels than the old European metes and bounds system.

In 1784, a congressional committee, chaired by Thomas Jefferson, proposed a rectangular division that became the Ordinance of 1785 and the foundation of the USPLS. It is very important to remember that the prime purpose of the USPLS was to divide U.S. public land quickly into easily and uniquely identified parcels so that they could be sold and taxed. Although based on field surveys, the precision (accuracy) of these surveys was of the lowest priority.

The third system (platted subdivision) is used throughout the United States and comes into play where large tracts have been subdivided into several smaller lots. In these cases, reference to the lot and/or block numbers or letters shown on a "subdivision plat" is the means by which a land parcel is identified. The subdivision plat is a plan, usually recorded, that specifies the location and dimensions of several parcels of land. A subdivision plat may divide a tract of land that was originally defined by metes and bounds, the USPLS, or another subdivision plat.

In spite of the possibility that any one of the preceding systems for identifying land is sufficient for sale and tax purposes, the tendency throughout the United States has been to add to or supplement USPLS and subdivision plat descriptions with a metes and bounds description for each sale of property. This dual use of property descriptions has led—can only lead—to instances of confusion in cases where the two descriptions conflict. Much of the confusion could he eliminated if land title users would realize that metes and bounds descriptions of USPLS or platted parcels were intended to *supplement* the description, not *supplant* it.

7.1. METES AND BOUNDS SYSTEM

The first settlers of the New World began to claim, occupy, and divide up parcels of land based on the European system of defining or describing land according to its relative position to well-known landmarks and by identifying the adjoining landowners. Although the large land settlement companies usually claimed land based on latitudinal boundaries, the precision needed to define smaller, individual holdings by astronomic position was not possible. It was much simpler, quicker, and more distinctive to choose natural boundaries to designate the limits of large tracts. Rivers, being the transportation arteries at the time, figured heavily in these early boundary definitions. Other natural features were also used and were usually chosen for their distinctive appearance or durability.

Of course, it was difficult to find rivers, streams, or rock formations in just the right locations to use as landmarks when the limits of a parcel of a particular size were being defined. This meant that artificial landmarks or monuments had to be set by the owners as a substitute for natural ones. Again, the limits of the parcel were defined by the actual location of particular corner landmarks or monuments.

In many cases of areas of long settlement, the construction of fences, hedgerows, or other obvious physical signs of the parcel limits made the task of marking the boundaries less difficult. The primary means of distinguishing one parcel from another became the "bounds" system. Land was held and worked by individuals on a regular basis. One need only name the individuals who owned, and usually lived on, the land surrounding a particular parcel to define where it was and what were its limits. This "bounds" system did not require that the size of the parcel be determined in order to define it.

In order to describe *roughly* the size of land parcels and assist in defining the location of that parcel, the concept of "metes" was used, along with the naming of natural landmarks, artificial monuments, and adjacent owners. The business of measuring between corners, outlined earlier, was taken on by anyone and everyone. The corners or limits of the land were physical marks on the ground. The “metes” were the measurements between those marks. Often, a landowner would simply pace off a section of land, set boundary stones, and sell the land so marked to another. The stones, neighboring landowners, and the paced distances would all be recorded in the sale, but the land would be occupied, cultivated, and possessed according to the location of the *corner monumentation*. This record of the distances (metes) and physical limits (bounds) of a land parcel constituted a “metes and bounds” description.

The metes and bounds description of a land parcel is, in essence, the words that draw a picture. It takes the form of an imaginary trip, leading the reader from a relatively well-known landmark to and around the parcel being defined. The naming of landmarks or monuments at each corner, along with the identification of each adjoining parcel, is an essential part of the description.

The metes and bounds description is so ingrained in the land title community and dovetails so well with the laws regulating the transfer of titles that, even in areas where more precise land identification systems are used, the metes and bounds description is often used to supplement land record system descriptions. A complete boundary description using the metes and bounds system will incorporate the following:

1. A commencing point that is well known, easily found, permanent, recoverable, recognizable, and preferably public in origin. The purpose of the commencing point is to clarify the location of the parcel in relationship to some universally recognized landmark.
2. A point-of-beginning that is a part of the property being described. The purpose of the point-of-beginning is to emphasize that the limits of the parcel itself are to follow in the dialogue. The point-of-beginning is used purely for descriptive purposes and is not the point where a Land Surveyor “begins” a survey.
3. A report of the physical objects (monuments) that mark the location of ends, and sometimes areas, along each line.
4. A report of the contiguous owners along each line or land record identification of the contiguous parcels (bounds).
5. A direction, usually a bearing, for each line. In some cases, the angle formed at a corner may be reported in lieu of a bearing. In the cases of curved or meandering boundaries, the appropriate words necessary to describe the configuration of the line are used.

6. A distance between each corner (metes).
7. The area of the parcel.[1]
8. Reference to the particular plat or map of survey that forms the basis for the description.[2]

Metes and bounds descriptions are, of necessity, redundant. A particular boundary will have several requirements placed on it. A boundary line will be reported as having a particular adjoiner, having certain monuments as its limits, having a certain length and direction, and as combining with all of the other boundaries to form an enclosed parcel of a particular area. It is not difficult to see that, given the inaccuracies of measurements, many of these elements might be contradictory.

The history of the development of the metes and bounds description shows that, for the most part, the physical limits of a parcel controlled possession and use of the land. Because of these factors, the practice, in most areas, has settled on a "hierarchy of calls" that rank the elements of a metes and bounds description. This ranking of calls is based on the rules of evidence and is commonly broadened to assist in interpretation of all forms of descriptions and boundary recovery.

The most common hierarchy of calls is as follows:

Natural monuments[3]
Artificial monuments[4]
Directions
Distances[5]
Areas

---

[1] The area is not a necessary part of a metes and bounds description.

[2] All metes and bounds descriptions must be based on a survey. The survey may have been crude, unplatted, and performed by the individuals involved in the sale, but, unless someone "measured" the distances and visited the bounds, a metes and bounds description is impossible.

[3] Natural monuments may consist of rivers, rock formations, trees, and other distinct features. Features made by humans, such as canals, levees, highways, or mounds, may be considered as natural monuments under certain conditions.

[4] Objects placed for the specific purpose of marking boundary corners, such as posts, concrete posts, iron rods, pipes, or other physical markers, constitute artificial monuments only when they are correctly set and undisturbed. Many Land Surveyors' markers represent locations that are not boundary controlled. These markers are not "artificial monuments" in the context of this listing.

[5] In Louisiana, the status of distances and directions is reversed. In some states, these items are combined.

Any element of a description of a parcel may be rejected or overruled, based on a review of the best evidence available. The courts have usually held that the most important and overriding factor in the interpretation of property descriptions is *the intent of the parties*.[6]

All of the words of a deed are to be considered so that evidence that best demonstrates the intention of the buyer and seller will prevail in most cases. With this in mind, the writer of a metes and bounds description should ensure that the *intention of the parties is clearly expressed*.

The significance of this ranking will be evident during the examination of the process of corner recovery. The charge that society gives to the Land Surveyor is to recover the boundaries where they exist on the ground. The wonderful precision of measurement and location that is possible today is of little use if an incorrect location for a boundary corner is used. A poor measurement to the correct corner is much superior to a precise measurement to a false corner. In real property boundaries, it is the *physical location of the corners that is important*.

During the early years of the settlement of the United States, when metes and bounds descriptions were the primary system of identifying and locating land parcels, the fact that every copy of a document had to be handwritten limited the use of survey plats to sketches and caused many to shorten descriptions to the point that the intentions of the parties became obscure. The dimensions of the parcels were often ignored or reduced to a report of adjoiners only.

Often, the dimensions reported in these early descriptions were the results of estimations by the landowners, not Land Surveyors. Measurements by amateurs, and indeed some Land Surveyors, most commonly report distances between monuments as being greater than the actual distance. Rough terrain, poor chaining techniques, meandering routes, and other error-developing factors more often misled all into believing a greater distance was covered than was actually traversed. Because vast areas of land were rapidly being settled, the shortage of skilled Land Surveyors forced many to depend on themselves to measure their land. Even today, in order to save a few dollars, many land transfers take place without a land boundary survey to verify the dimensions of a parcel.

The situation of possessing according to physical marks should be distinguished from the case of a transfer of property where relying on specific dimensions is the intention of the parties. If a sale is not followed by possession and it was the clear intent of the parties to transfer a land parcel of specific size and shape, then any blunders in marking the limits of such a parcel can be corrected at any time. The intent to transfer a specific-sized parcel can occur only in cases of resubdivision. Original private claims and USPLS parcels cannot fall into this category because *possession* was required in the first case and the *corners set by the government surveyor* controlled in the latter.

---

[6] C. M. Brown, W. G. Robillard, D. A. Wilson, *Evidence and Procedures for Boundary Location*, 5th ed. Hoboken, NJ: John Wiley & Sons, 2006.

7.1.1. Reading a Metes and Bounds Description

The most difficult part of writing a metes and bounds description is deciding the degree of detail required to convey the intentions of the parties. The natural instincts are to limit the verbiage as much as possible so that the resulting descriptions are not overly "wordy." Unfortunately, this trend has led to many misinterpretations in the past. Instead of rehashing the general rules, perhaps a review of some typical parcels and how their metes and bounds descriptions might be created would better clarify the need for specificity.

In Figure 7.1, the iron post shown is reported by the Land Surveyor as monumenting the intersection of the line between the parcels known as the "Brownell" estate and the "Willis" estate with the northernmost right-of-way line of "Patterson Highway." Each of the corners of the "Johnston" property is monumented by six-inch square concrete posts. The Willis estate, in this example, is the parent tract from which the Johnston property was formed. It is the intention of the parties that the parcel sold be 200 feet wide, 500 feet deep, and commence 388 feet from the western boundary of the Willis estate, as measured along the highway, and that the sidelines of the parcel be parallel with the western boundary of the parent estate.

The plat of the survey by Jackson clearly reflects these intentions and provides some additional information. The large iron post is a well-known (in 1923) and accepted feature that qualifies as an artificial monument. The six-inch square concrete posts are simply survey markers set by Jackson. These may mature into artificial monuments, provided sufficient acts of possession and reliance on these posts take place. There is a discrepancy on the survey plat in the form of an incorrectly computed area. Assuming that all of the other dimensions are correct, the area should be 2.21 acres (rounded off to the nearest one-hundredth of an acre).

A typical metes and bounds description of the period might read:

> . . . a 2.3 acre parcel of land measuring 200 feet wide by 500 feet deep between equal and parallel lines and beginning 388 feet from the western boundary of

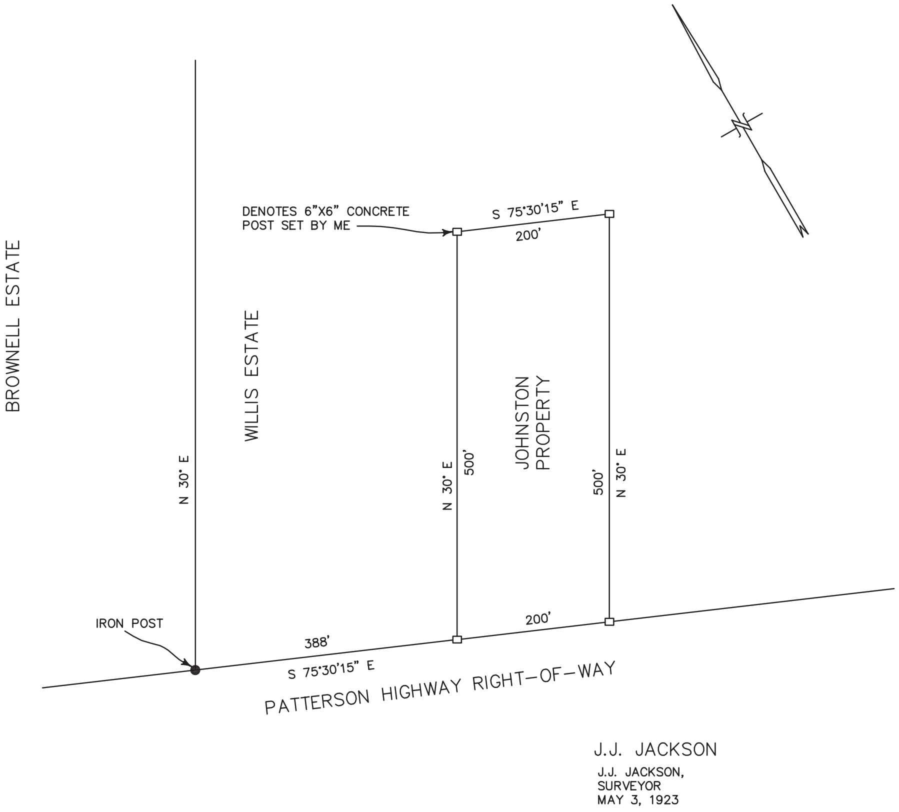
BROWNELL ESTATE
WILLIS ESTATE
JOHNSTON PROPERTY
DENOTES 6"X6" CONCRETE POST SET BY ME
S 75°30'15" E
200'
N 30° E
N 30° E
500'
N 30° E
500'
IRON POST
388'
200'
S 75°30'15" E
PATTERSON HIGHWAY RIGHT-OF-WAY
J.J. JACKSON
J.J. JACKSON,
SURVEYOR
MAY 3, 1923

FIGURE 7.1

> the Willis Estate, bounded in the front by the Patterson Highway and on the sides as well as in the rear by the property of the seller. . . .

Note the lack of reference to the survey plat by Jackson. If such a reference had been made, then the several discrepancies between the description and the intention of the parties might be clarified. Even if such a reference were made, there would always exist the possibility that the survey plat might not survive. For this reason, metes and bounds descriptions should be written as if all other documents reflecting the intentions of the parties were going to be destroyed.

Figure 7.2 shows one interpretation of the description superimposed on the original survey plat. The scale is distorted to emphasize the discrepancies.

The description does not reflect the intentions of the parties to the sale. Later surveys that recover the concrete posts set by Jackson will

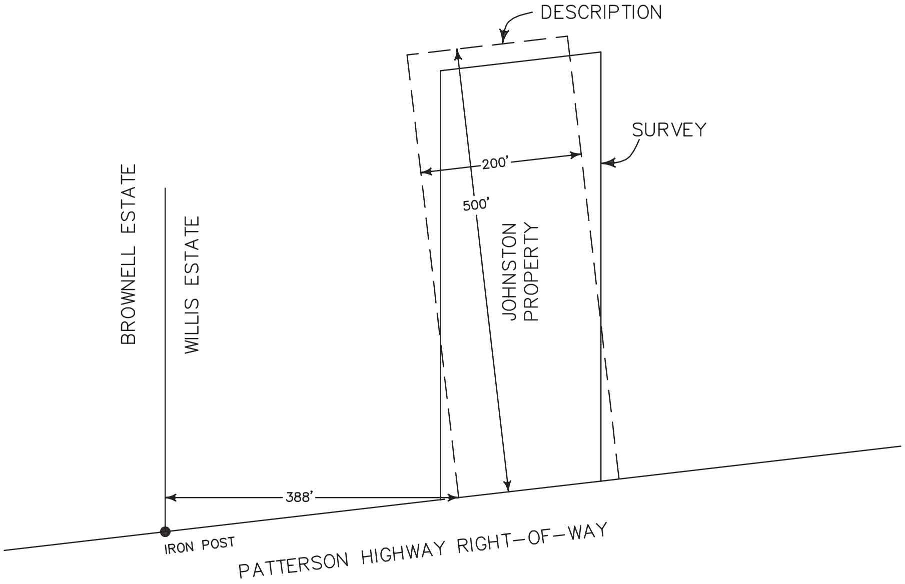

FIGURE 7.2

require additional evidence and be accompanied by sufficient possessive acts before it can be stated with confidence that they do indeed represent the intended limits of the Johnston property. Even if the survey plat by Jackson were to surface, it might be argued that it was not relevant because it was not part of the written instruments included in the transaction. Now consider the following description:

> . . . commencing at the large iron post[7] found at the intersection of the boundary line between the Brownell Estate and the Willis Estate and the northernmost right-of-way line of Patterson Highway[8]; thence, in an easterly direction[9] along said northernmost right-of-way line of Patterson Highway,[10] South 75 degrees

---

[7] Reference to the "iron post" fixes the commencement point to a recoverable terrain feature. If, at some later date, the location of the boundary between the Brownell and Willis estates were to become unclear, the location of the Johnston property would not be cast in doubt.

[8] The reference to the intersection of the boundary line and the right-of-way line of the highway is used not only to document the location of the iron post, lest it be disturbed, but also to emphasize that the right-of-way for Patterson Highway is relevant to the measurements.

[9] The words "in an easterly direction" are used to assist the reader in interpreting whether the south–east or the north–west version of the bearing that follows is appropriate.

[10] This removes any doubt about the line of measurement of the 388 feet. It is clear that it was not the intention of the parties that the 388 feet be measured at right angles to the boundary line.

> 30 minutes 15 seconds East, a distance of[11] 388 feet to the Point-of-Beginning[12]; thence, in a northerly direction parallel with the boundary line between the Brownell Estate and the Willis Estate,[13] North 30 degrees East, a distance of 500 feet; thence, in an easterly direction parallel with the northernmost right-of-way line of Patterson Highway, South 75 degrees 30 minutes 15 seconds East, a distance of 200 feet; thence, in a southerly direction parallel with the boundary line between the Brownell Estate and the Willis Estate, South 30 degrees West,[14] a distance of 500 feet to the northernmost right-of-way line of Patterson Highway; thence, in a westerly direction along said northernmost right-of-way line of Patterson Highway,[15] North 75 degrees 30 minutes 15 seconds East, a distance of 200 feet to the Point-of-Beginning, encompassing an area of 2.3 acres[16] and all as more fully described on a plat of survey by J. J. Jackson, Surveyor dated May 3, 1923[17] . . .

This description is somewhat longer than the first example, but it presents a much more accurate picture of the transaction. The inclusion of the survey plat as part of the description also gives a clue to the source of the information given in the description.

Figure 7.3 represents a survey plat of the same Johnston property 55 years later. The changes that time has wrought are not as great as might have been expected. Two of the original concrete posts are still standing and, from all indications, were not disturbed during all that time. It is known that Abraham Johnston built the house in 1926 and has lived there ever since. The acceptance and use of these concrete posts by the possessor of the parcel has converted them from

---

[11] The use of “a distance of” assists the reader in distinguishing distances along the described boundary line and other dimensions that maybe necessary, such as a radius for a curve, and so on.

[12] If one were to insert the phrase “to a 6-inch square concrete post” here, then the distance “388 feet” might be interpreted as an estimation and the concrete post might be considered an artificial monument, without any acts of possession. Blunders by Jackson in setting the posts may not be correctable. Adequate possession with reliance on the post would make this a moot point.

[13] By calling for the sidelines to be parallel to the western boundary, any measurement blunder by Jackson should not alter the location of the line.

[14] Notice that the survey plat indicates the direction to be “North 30 degrees East.” The survey plat is a two-dimensional drawing; therefore, “North 30 degrees East” and “South 30 degrees West” define the same line.

[15] This leaves no doubt that frontage on Patterson Highway is intended for this parcel.

[16] The incorrect area is repeated in the description but is of no consequence in the absence of any evidence that it was the controlling factor in the intention of the vendor or vendee.

[17] The reference to the survey plat in the description has the effect of including the plat as an integral part of the description. Discrepancies between the written description and the figures on the plat are normally resolved in favor of the plat, unless clear and convincing evidence to the contrary is discovered.

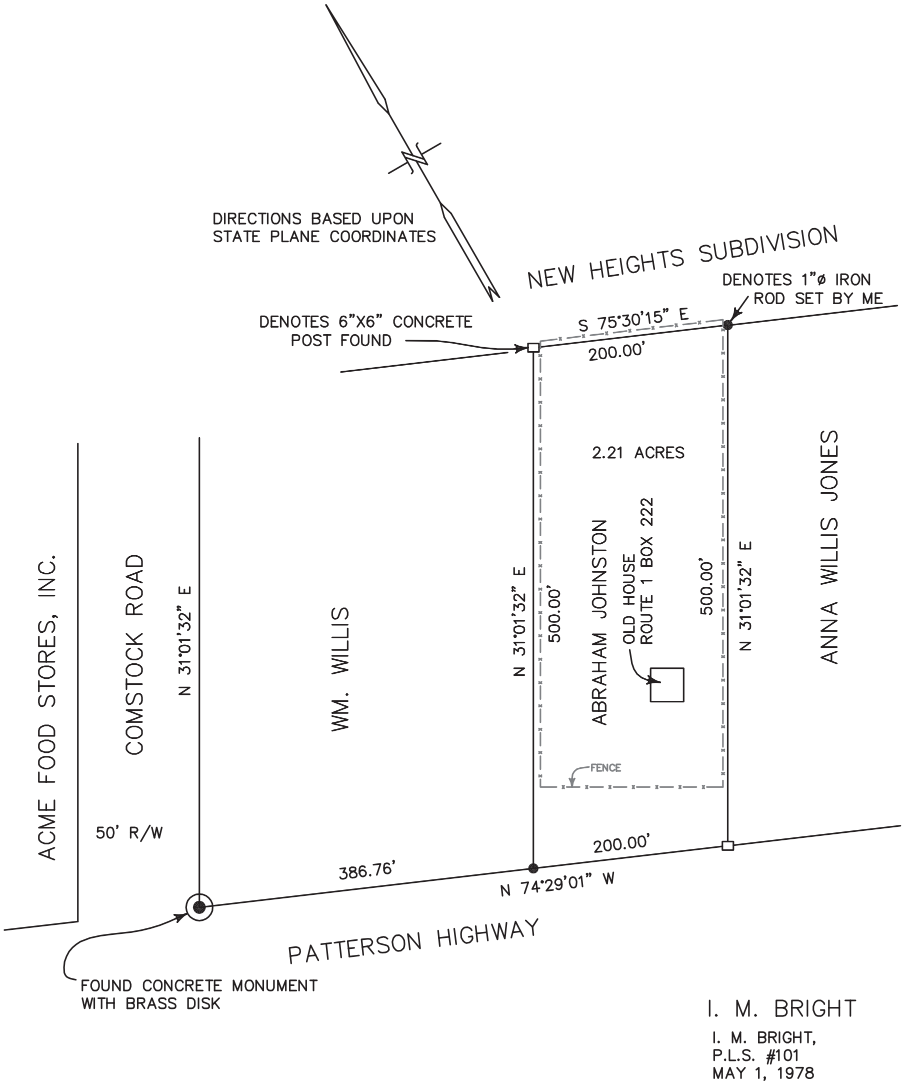
DIRECTIONS BASED UPON
STATE PLANE COORDINATES
N
NEW HEIGHTS SUBDIVISION
DENOTES 1"ø IRON
ROD SET BY ME
DENOTES 6"X6" CONCRETE
POST FOUND
S 75°30'15" E
200.00'
2.21 ACRES
ABRAHAM JOHNSTON
OLD HOUSE
ROUTE 1 BOX 222
500.00'
N 31°01'32" E
500.00'
N 31°01'32" E
ANNA WILLIS JONES
ACME FOOD STORES, INC.
COMSTOCK ROAD
N 31°01'32" E
WM. WILLIS
FENCE
50' R/W
200.00'
386.76'
N 74°29'01" W
PATTERSON HIGHWAY
FOUND CONCRETE MONUMENT
WITH BRASS DISK
I. M. BRIGHT
I. M. BRIGHT,
P.L.S. #101
MAY 1, 1978

FIGURE 7.3

mere survey markers, attempting to demonstrate the intent of the parties, into artificial monuments that clearly define the limits of the parcel.

The easternmost line of Comstock Road is believed to be the original line between the Brownell and Willis estates. So many changes have occurred that a new metes and bounds description is required in order for the property to be recognized. The application of the hierarchy of calls, as well as the professional judgment of Bright, will be noted during the development of the updated metes and bounds description.

> . . . commencing at a brass disk set in concrete found at the intersection of the easternmost right-of-way line of Comstock Road and the northernmost right-of-way line of Patterson Highway[18]; thence, in an easterly direction along said northernmost right-of-way line of Patterson Highway, South 74 degrees 29 minutes 01 seconds East,[19] a distance of 386.76 feet to a 1 inch diameter iron rod on the easternmost boundary line of the property of William Willis and the Point-of-Beginning[20]; thence, in a northerly direction along said easternmost boundary line of the property of William Willis,[21] North 31 degrees 01 minutes 32 seconds East, a distance of 500.00 feet to a 6 inch square concrete post on the southernmost boundary line of New Heights Subdivision[22]; thence, in an easterly direction along said southernmost boundary line of New Heights Subdivision, South 74 degrees 29 minutes 01 seconds East a distance of 200.00 feet to a 1 inch diameter iron rod on the westernmost boundary line of the property of Anna Willis Jones; thence, in a southerly direction along said westernmost boundary of the property of Anna Willis Jones, South 31 degrees 01 minutes 32 seconds West a distance of 500.00 feet to a 6 inch square concrete post on the northernmost right-of-way line of Patterson Highway; thence, in a westerly direction along said northernmost right-of-way line of Patterson Highway, North 74 degrees 29 minutes 01 seconds West, a distance of 200.00 feet to the Point-of-Beginning, encompassing an area of 2.21 acres and all as more fully described on a plat of survey by I. M. Bright, P.L.S. dated May 1, 1978 . . .

---

[18] This commencement point may or may not be in the identical location as the large iron post found in the 1923 description. The differences in the commencement distance that follows may be the result of a slight change in the accepted position for the line between the original estates or may indicate a measurement discrepancy between Bright and Jackson.

[19] The route of the description is in an easterly direction. Therefore, the bearing must be south–east, not north–west. The difference in magnitudes is not significant. The modern survey is oriented to state plane directions; the 1923 survey does not report a bearing base.

[20] The call is to the iron rod in this case, because nearby original concrete posts enabled Bright to recover and remonument the southwest and northeast corners of the tract. The original concrete posts have clearly become artificial monuments, marking with certainty the limits of the Johnston property.

[21] There is no longer the need to specify the line as parallel to the original boundary between the Brownell and Willis estates.

[22] The fact that the distances are now reported to the one-hundredth of a foot is an indication of the precision of the measurements. The size of the end monuments (6 inches × 6 inches and 1 inch o.d.) are such that a variation of 0.10 feet or so ought to be expected. Bright chose to report the same metes as Jackson.

In the day of hand-copied records and manual typewriters, the preceding description would have been condemned as much too verbose. In the modern world of copying machines and word processors, there is no need to sacrifice clarity of intent by reducing the length of a metes and bounds description. Some may still insist that the exact same idea can be expressed in fewer words. If we trim any mention of "New Heights Subdivision," then the description would read as follows:

> . . . commencing at a brass disk set in concrete found at the intersection of the easternmost right-of-way line of Comstock Road and the northernmost right-of-way line of Patterson Highway; thence, in an easterly direction along said northernmost right-of-way line of Patterson Highway, South 74 degrees 29 minutes 01 seconds East, a distance of 386.76 feet to a 1 inch diameter iron rod on the easternmost boundary line of the property of William Willis and the Point-of-Beginning; thence, in a northerly direction along said easternmost boundary line of the property of William Willis, North 31 degrees 01 minutes 32 seconds East, a distance of 500.00 feet to a 6 inch square concrete post; thence, in an easterly direction, South 74 degrees 29 minutes 01 seconds East a distance of 200.00 feet to a 1 inch diameter iron rod on the westernmost boundary line of the property of Anna Willis Jones; thence, in a southerly direction along said westernmost boundary of the property of Anna Willis Jones, South 31 degrees 01 minutes 32 seconds West a distance of 500.00 feet to a 6 inch square concrete post on the northernmost right-of-way line of Patterson Highway; thence, in a westerly direction along said northernmost right-of-way line of Patterson Highway, North 74 degrees 29 minutes 01 seconds West, a distance of 200.00 feet to the Point-of-Beginning, encompassing an area of 2.21 acres and all as more fully described on a plat of survey by I. I M. Bright, P.L.S. dated May 1, 1978 . . .

At first glance, there doesn't appear to be any difference between the two descriptions. Let us assume that the second version is used and that the northernmost concrete post, as well as the iron rod, are destroyed. Later, when New Heights subdivision's southern boundary is monumented and found to be 501 feet from Patterson Highway, who owns the strip between the 500-foot rear line of the Johnston property and New Heights subdivision? The boundary action needed to settle this aggravating, petty discrepancy might have been avoided by the use of a complete, albeit wordy, metes and bounds description.

7.2. U.S. PUBLIC LAND SYSTEM (USPLS)

The great American philosopher, Will Rogers, once said:

> "The trouble ain't so much the things we don't know—it's the things we do know that just ain't so."

I know of no instance where this thought is more applicable than in USPLS states. A close examination of the origin and history of the USPLS will lead us to a better appreciation of the ingenuity—and pitfalls—associated with this system.

At the end of the Revolutionary War, the United States was a new nation without a firm revenue base. Taxes were unthinkable; yet it takes money to operate a federal government. The one thing that the fledgling country did have was land. From the original state cessions in 1781 to the purchase of Alaska in 1867, approximately 1,807,682,000 acres of land were acquired by the federal government, and the faster this land could be turned into hard cash, the better. The prime, yet unwritten, directive to the committee that created the Ordinance of 1785 was to divide public lands into marketable portions. In order to accomplish this, the following things, listed by priority, were necessary.

1. The parcels had to marked on the ground so that the new landowners could take possession of their land.
2. The parcels created had to be uniquely identifiable.
3. The system of identification had to be one that required little formal education to use, if not understand.
4. Further division of the created parcels had to be simple, universal, systematic, and based on the location of the monumented corners, and could not require additional survey work by the government surveyor.
5. The time and paperwork required to create and subdivide each parcel had to be kept to a minimum.
6. The system could not depend on prior knowledge of the terrain in order to implement it.
7. The quality or arability of the land had to be reported to determine relative value.
8. The dimensions of the sides of the parcel and other measurements had to be reported to *facilitate the recovery of the corner markers*, as well as estimates of acreage.

Notice the priority given to measuring the dimensions of the newly created parcels. Land lines are defined by the corners *set by the government surveyor*, not by the reported dimensions or directions. The USPLS fulfilled these requirements completely and in the required order of priority. It is important to keep this order of priority in mind when dealing with USPLS parcels.

As the USPLS was perfected over the years, slight changes in the techniques, numbering order, monumentation, and other minor alterations took place. The 20 states that do not operate under the USPLS are Connecticut, Delaware, Georgia, Hawaii, Kentucky, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, North Carolina, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Vermont, Virginia, and West Virginia. The other 30 states are USPLS states, and the exact version of division used in each area vary according to the instructions in force, the government surveyor, the terrain, local customs, and other factors.

7.2.1. Initial Point

As so often occurs in surveying, the theory is simple, but the application is complex. The theory of the USPLS was based on a rectangular division of land along the cardinal directions. The first step in the division of an area of land, such as a state, under the USPLS, was the establishment of the "initial point" for that division. Although some attempt was made to locate initial points at a specific latitude (a list of initial points was made, reporting both latitude and longitude of the various initial points[23]), the initial point was the actual location, on the ground, as monumented by the government surveyor. The reported position existed only to assist in the recovery of the true location.

7.2.2. Principal Meridian

Having chosen and monumented the initial point, the government surveyor established the "principal meridian" by traversing north and south from the initial point. Depending on the instructions at the time, a post or other monument was set at least every six miles or as often as every half mile. The principal meridians were run with particular care, and every attempt was made to run the line "true" north, resulting in an approximation of astronomic north being marked on the ground.

---

[23] J. G. McEntyre, *Land Survey Systems*. New York: John Wiley & Sons, 1978.

7.2.3. Baseline

Lines were also laid out east and west from the initial point and monumented with the same frequency as the principal meridian. This “curved” line is called the “baseline” and is actually a series of short chords between the monuments set by the government surveyor. As with the principal meridian, every attempt was made to follow the “true” or cardinal direction, resulting in an approximation of the astronomic latitude of the initial point.

Figure 7.4 demonstrates what should have been the result of the establishment of a typical initial point, principal meridian, and baseline.

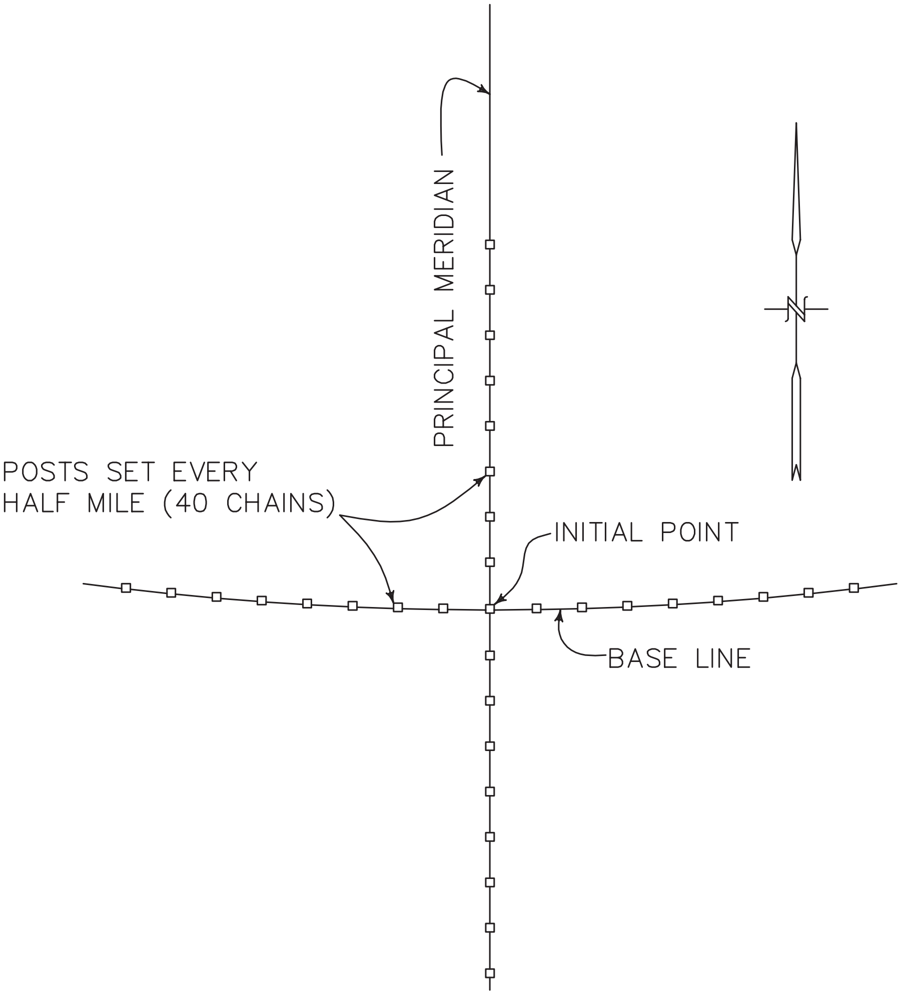
PRINCIPAL MERIDIAN
POSTS SET EVERY HALF MILE (40 CHAINS)
INITIAL POINT
BASE LINE
N

FIGURE 7.4

Additional or "correction" east–west lines may have been monumented at regular intervals along the principal meridian with the same care as the baseline so that a correction could be made for the convergence of the meridians. It was well known by the designers that a rigid rectangular division of land based on the cardinal directions was impossible, but the inaccuracies of measurement, the scientific unsophistication of the ultimate user of the system, the need for haste, and the need for a simple method of record keeping outweighed the requirements of accuracy.

7.2.4. Township and Range Lines

After the establishment of the principal meridian and the base line, additional east–west lines were monumented at 6-mile intervals along the principal meridian. These are known as "township" lines. Additional north–south lines were monumented along the base line, and later along correction lines, at 6-mile intervals.[24] These are known as "range lines." The theoretical result of these steps is shown in Figure 7.5.

Each east–west row of "6-mile squares" is identified by numbering its position north or south of the base line. Each north–south column of "squares" is identified by its position east or west of the principal meridian. Thus, every square, called a "township," is uniquely identifiable by a simple report of its township number and range number.

Notice that township 8 south, range 4 east has been highlighted in Figure 7.5. This is abbreviated to "T8S, R4E."

The township and range lines were not run with quite the same degree of care as were the principal meridian and the base line. Of course, none of the townships were exact squares measuring six miles on a side. This fact is not important to the division and inventory of the land, and, although not forgotten, it was set aside by the planners of the USPLS because it was considered relatively unimportant.

7.2.5. Ideal Township

Further division of each township was accomplished by running lines that were, more or less, north–south and east–west at one-mile intervals and monumented every half mile. Each resulting "square mile" is called

---

[24] The interval between correction lines varied from every fourth to every sixth township line, depending on the instructions in force at the time.

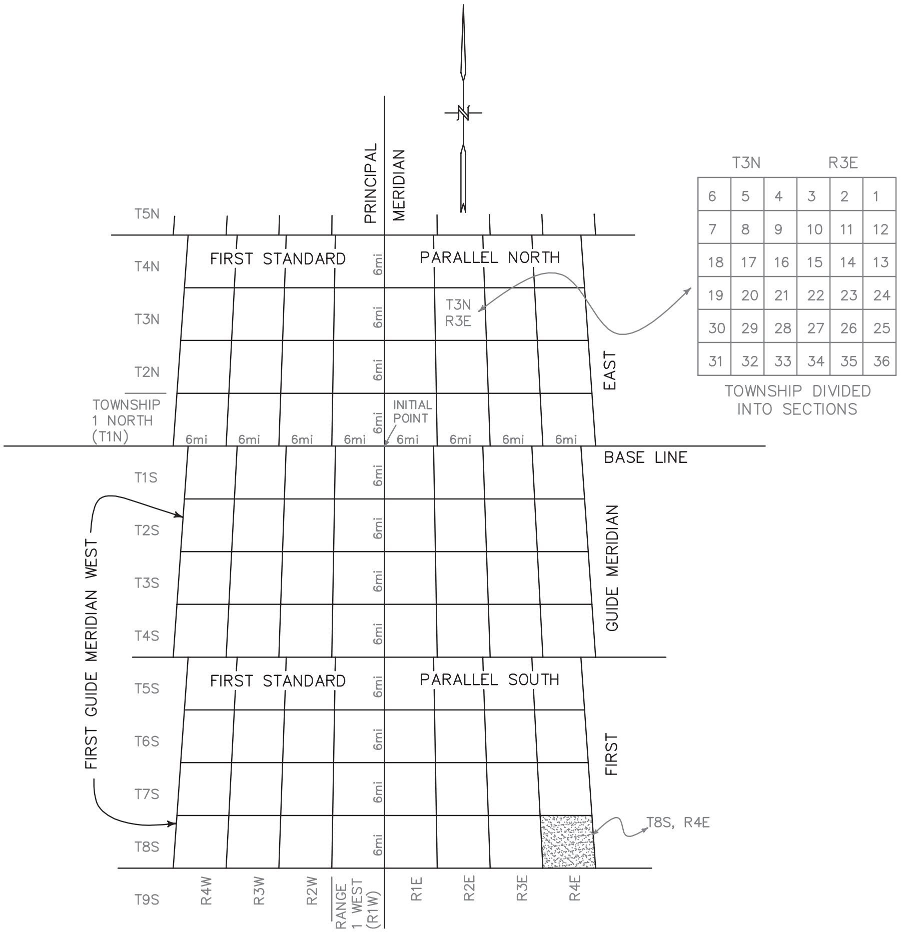
N
PRINCIPAL MERIDIAN
T3N R3E
6 5 4 3 2 1
7 8 9 10 11 12
18 17 16 15 14 13
19 20 21 22 23 24
30 29 28 27 26 25
31 32 33 34 35 36
TOWNSHIP DIVIDED INTO SECTIONS
T5N
T4N
FIRST STANDARD
PARALLEL NORTH
T3N
T3N R3E
T2N
EAST
TOWNSHIP 1 NORTH (T1N)
INITIAL POINT
6mi
BASE LINE
T1S
T2S
T3S
T4S
GUIDE MERIDIAN
FIRST GUIDE MERIDIAN WEST
T5S
FIRST STANDARD
PARALLEL SOUTH
T6S
FIRST
T7S
T8S, R4E
T8S
T9S
R4W
R3W
R2W
RANGE 1 WEST (R1W)
R1E
R2E
R3E
R4E

FIGURE 7.5

a "section" and identified by a number. The work of monumentation was typically begun at the southeast corner of each township and progressed to the north and west until the entire township was monumented. The discrepancies that resulted from the closure of the meridians, errors of measurement, and undiscovered blunders were accounted for by allowing the western tier and northern row of sections to deviate from the "square mile" template.

All other sections were *reported* as squares measuring one mile to a side. Of all the over 2 million sections monumented by the Government Land Office (GLO), it is safe to say that *none* is a square measuring

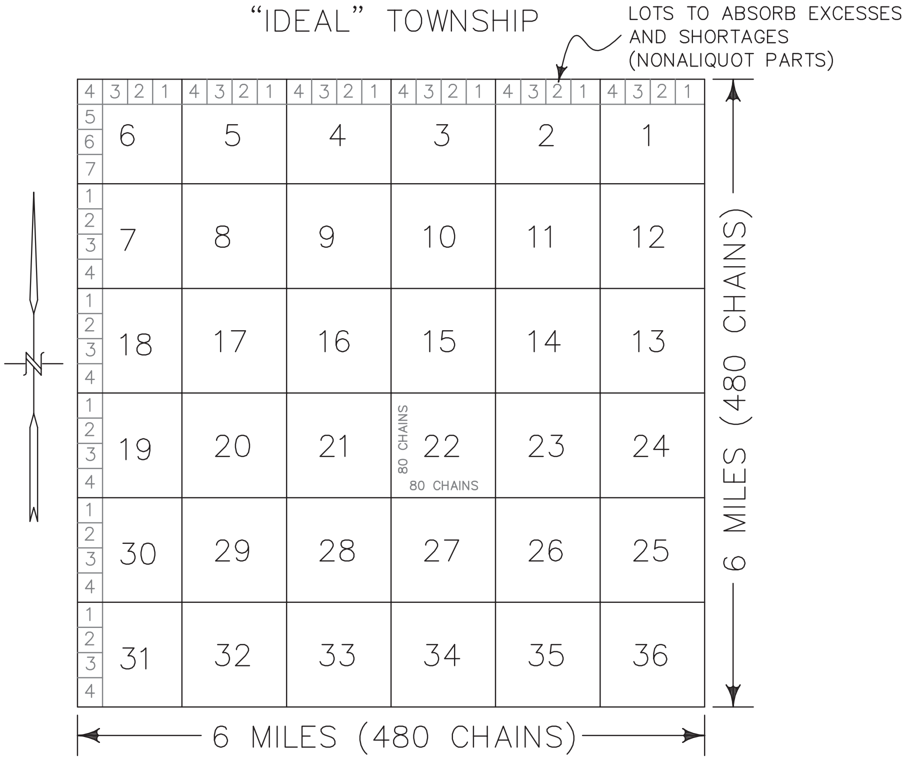
"IDEAL" TOWNSHIP
LOTS TO ABSORB EXCESSES AND SHORTAGES (NONALIQUOT PARTS)
4 3 2 1
5 6 7
1 2 3 4
6 5 4 3 2 1
7 8 9 10 11 12
18 17 16 15 14 13
19 20 21 22 23 24
30 29 28 27 26 25
31 32 33 34 35 36
80 CHAINS
80 CHAINS
N
6 MILES (480 CHAINS)
6 MILES (480 CHAINS)

FIGURE 7.6

one mile to a side and few are even close, especially when the precision of modern surveys are considered. Still, the "ideal section" must be discussed in order to explain the USPLS. Figure 7.6 is that of the "ideal township," divided into "ideal sections."

7.2.6. Ideal Section

The division of the "ideal section" into its aliquot parts is something that has been widely sketched and described as *if the dimensions of the sides* were an important part of the process. This is not the case. The beauty of the USPLS is the very fact that it is *independent* of dimensions. It is a system that depends entirely on the actual locations of the corners set by the government surveyor, and *nothing else*.

Figure 7.7 shows the division of a section and how each parcel is identified. The dimensions are not shown, because they are *not essential* to the process and tend to only mislead the novice into believing that

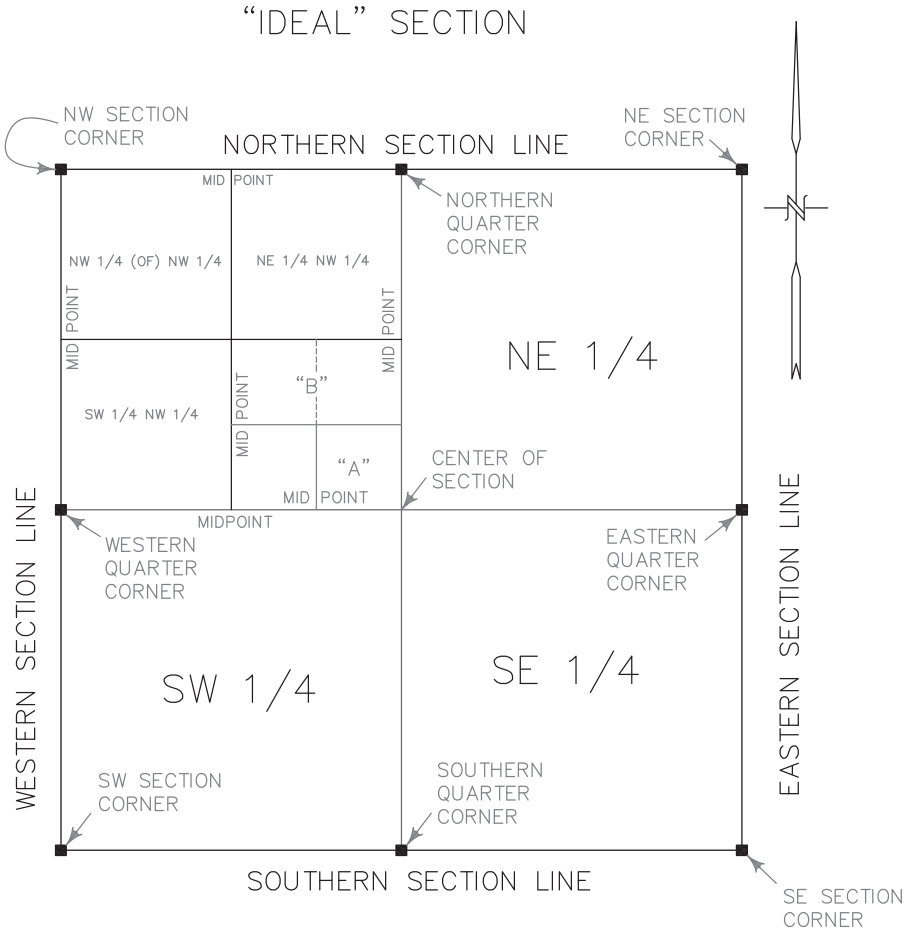
"IDEAL" SECTION
NW SECTION CORNER
NORTHERN SECTION LINE
NE SECTION CORNER
MID POINT
NORTHERN QUARTER CORNER
NW 1/4 (OF) NW 1/4
NE 1/4 NW 1/4
MID POINT
MID POINT
NE 1/4
"B"
MID POINT
SW 1/4 NW 1/4
MID POINT
"A"
CENTER OF SECTION
MID POINT
WESTERN SECTION LINE
MIDPOINT
WESTERN QUARTER CORNER
EASTERN QUARTER CORNER
EASTERN SECTION LINE
SW 1/4
SE 1/4
SW SECTION CORNER
SOUTHERN QUARTER CORNER
SOUTHERN SECTION LINE
SE SECTION CORNER
N

ALL POINTS LABELED "CORNERS" WERE SET BY THE GOVERNMENT SURVEYOR

FIGURE 7.7

sections and their aliquot parts should measure the same as the example in the explanation.

The monuments set by the government surveyor at what was typically reported as 1-mile intervals (80 chains) are the section comers. Often monuments were also set by the government surveyor at, reportedly, half-mile intervals (40 chains). These are called "quarter corners." The linchpin of the USPLS is the actual location, on the ground, of the monuments set by the government surveyor. The reported directions and distances can be, and usually are, incorrect or, at best, imprecise. The physical spot on the ground monumented by the government surveyor is, and can only be, in one place, nowhere else. That location can be lost, obscured, misreported, and confused but *never moved*. No matter how imprecise the measurements or how gross the blunders are that may have occurred, the corner monumented can mark only one land point on the earth, and that land point is exactly where it is. Because the actual location of the monuments marking the corners is so vital, extraordinary emphasis must be placed on the recovery of these locations.

Straight land lines between opposite quarter corners divide the section into a southeast quarter (SE$^1/_4$), a northeast quarter (NE$^1/_4$), a southwest quarter (SW$^1/_4$), and a northwest quarter (NW$^1/_4$). The intersection of these lines is a land point and is called the "center of the section." Further division is accomplished by connecting straight lines between the midpoints of opposing sides of the larger tract. The description of the resulting parcel is identified by listing the divisions made in the reverse order in which they were made. Tract "A" in Figure 7.7 is the southeast quarter of the southeast quarter of the northwest quarter of the section. Tract "B" is the north half of the southeast quarter of the northwest quarter of the section.

If Figure 7.7 was that of section 26, in the fourth township south of the base line and the fifth range west of the sixth principal meridian, then tract "A" could be uniquely identified by the single line "SE$^1/_4$ SE$^1/_4$ NW$^1/_4$ Sec. 26, T.4S., R.5W., 6th P.M." Tract "B" would be abbreviated as "N$^1/_2$ SE$^1/_4$ NW$^1/_4$ Sec. 26, T.4S., R.5W., 6th P.M." Note the lack of commas in the list up to the section number. If a comma were inserted in the description of "B," for example, the result might be "N$^1/_2$, SE$^1/_4$ NW$^1/_4$ Sec. 26, T.4S., R.5W., 6th P.M." This could be interpreted as calling for the north half *and* the southeast quarter of the northwest quarter of the section!

The dimensions of tract "A" or "B" in the example are not known and are not important in the description. What is known about the dimensions of these parcels is that, if section 26 were an "ideal" section, with each side measuring 80 chains, then tract "A" would be a square measuring 10 chains on a side and encompassing 10 acres. Tract "B" would be a rectangle measuring 10 chains on the east and west and 20 chains on the north and south and encompassing 20 acres.

Of course, section 26 is not an "ideal" section; therefore, tract "A" will not be exactly 10 acres, nor will tract "B" be exactly 20 acres. A recovery and careful survey of the corners set by the government surveyor is required before a more precise knowledge of the dimensions of these parcels is possible.

7.2.7. Irregular Sections

The pattern of sections, and even townships, had to be interrupted because of several factors, such as existing private claims, large rivers, lakes, state boundaries, national boundaries, the sea, and other irregularities both natural and human made. This resulted in fractional sections, where the reported configuration deviated from the typical pattern. Aliquot division is still possible in these "irregular" sections by the application of certain rules design to isolate the irregularity.

7.2.7.1. Meander Lines Rivers, lakes, the shore of the sea, and other natural barriers that arose in the course of the monumentation of the section lines were accounted for by the "meander line." Figure 7.8 illustrates the application of the meander line along a navigable river. The limit of private ownership along a navigable river is defined by the waterway. It is not practical to measure each minute and sinuous turn of the waterway; therefore, a method that approximates the waterway is used.

In the example, the government surveyor encountered the river while running west along the line between sections 5 and 8. A meander comer was established on this line *at a convenient point* on or near the bank of the river. Additional meander points were established at points along the river, and the directions and distances between these points were measured so that the approximate shape of the bank of the river could be plotted. This resulted in the line labeled as the "meander line." All distances and directions are to the meander points, but the private ownership extends beyond the meander line to the waterway.

Each state has differing laws that define the limit of private ownership along navigable rivers, navigable lakes, the sea, non-navigable lakes, rivers, and streams. These boundaries are known collectively as "riparian boundaries."[25]

Although a general knowledge of the rules of the establishment and division of USPLS lands is important, each section was separately surveyed, monumented, and sold. Because of this, each section is unique in some way. The recovery of the corner monuments and knowledge of the boundaries of any section or its aliquot parts requires special knowledge of that particular section.

---

[25] Additional rules and definitions dealing with irregular sections and the aliquot division of USPLS sections constitute a book in itself. The reader desiring greater detail is referred to John G. McEntyre, *Land Survey Systems*. New York: John Wiley & Sons, 1978.

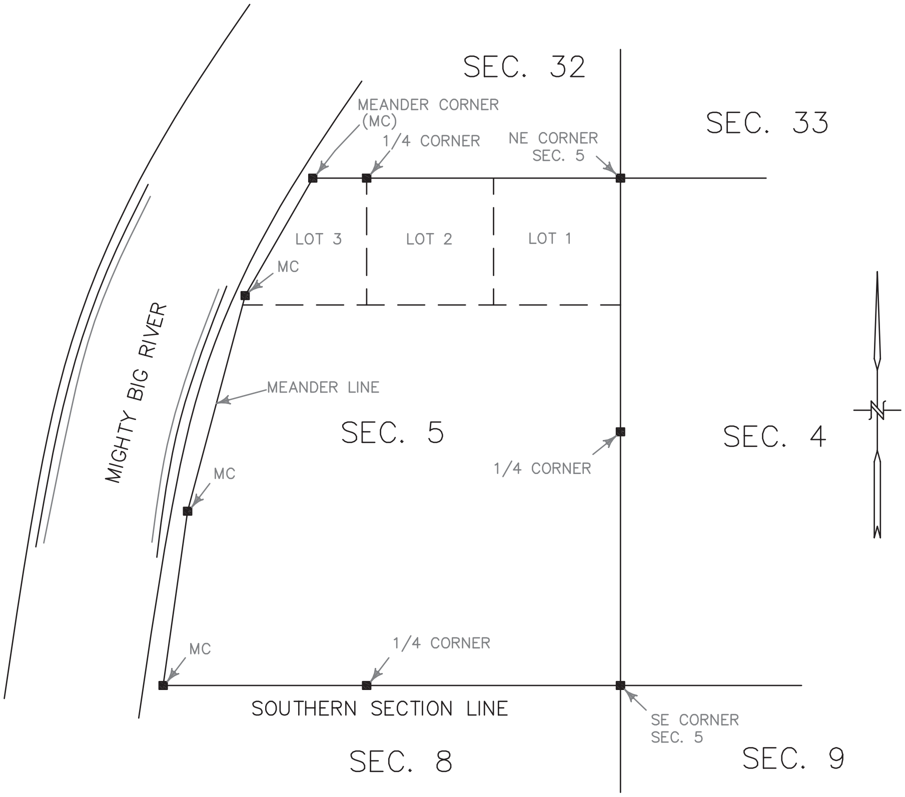
SEC. 32
SEC. 33
MEANDER CORNER (MC)
1/4 CORNER
NE CORNER SEC. 5
LOT 3
LOT 2
LOT 1
MC
MIGHTY BIG RIVER
MEANDER LINE
SEC. 5
SEC. 4
N
1/4 CORNER
MC
MC
1/4 CORNER
SOUTHERN SECTION LINE
SE CORNER SEC. 5
SEC. 8
SEC. 9

ALL POINTS LABELED "CORNERS" WERE SET BY THE GOVERNMENT SURVEYOR

FIGURE 7.8

7.2.8. USPLS General Instructions

Depending on the year that the work was done, each government deputy surveyor was issued a set of general instructions. These instructions directed the details of measurement, monumentation, observations to be recorded, and method of division, among other things. Subsequent instructions might alter or expand upon previous ones. The instruction told the deputy surveyor how he *should* conduct the surveys; how the deputy surveyor actually performed the surveys is another matter.

7.2.9. Government Surveyor's Original Notes

Each government surveyor kept a record of the distances, directions, observations, monumentation, and other pertinent data in what is called the "surveyor's notes." Typically, these notes were not written as work was progressing but were transcribed from rough notes, usually kept on a slate, at the end of each line. The notes were the best record of what was done, how it was done, the order in which it was done, what was measured, what was encountered along the way, and how monumentation was accomplished.

These notes were used to create the township plats that mapped each township, showing the sections and/or lots that were to be sold to private owners. The government surveyor's notes and the township plat could only reflect the directions and dimensions reported at the time of the original survey. The actual configuration of any township, section, or lot was dictated by the location of the corners, not the reported dimensions.

This significant difference is best explained by an example. In Figure 7.9, the solid lines represent the reported route of the government surveyor and the reported location of the monuments set by the government surveyor. The small circles represent the actual location of the monuments. The dashed lines represent the actual section lines.

The dramatic differences between the reported and actual locations of the section corners in Figure 7.9 are not exaggerations. The

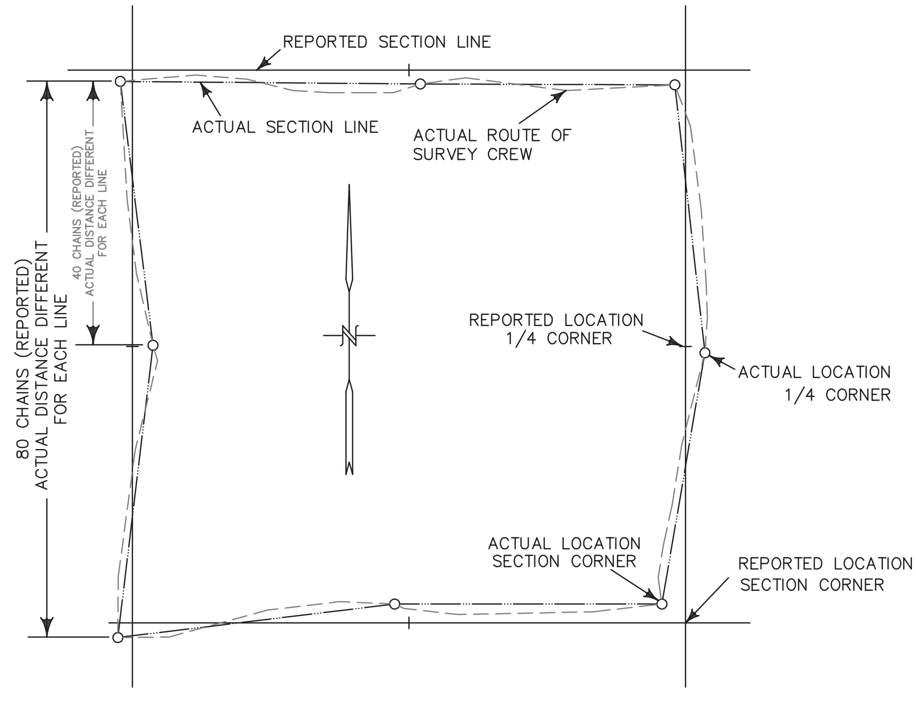
REPORTED SECTION LINE
ACTUAL SECTION LINE
ACTUAL ROUTE OF SURVEY CREW
80 CHAINS (REPORTED) ACTUAL DISTANCE DIFFERENT FOR EACH LINE
40 CHAINS (REPORTED) ACTUAL DISTANCE DIFFERENT FOR EACH LINE
N
REPORTED LOCATION 1/4 CORNER
ACTUAL LOCATION 1/4 CORNER
ACTUAL LOCATION SECTION CORNER
REPORTED LOCATION SECTION CORNER

FIGURE 7.9

discrepancies between the reported and actual dimensions in this example, and greater, were common, especially in those states of early surveys. The division of the section, being dependent on section corner and quarter corner location, is accomplished in spite of the inaccuracies of the original reports of direction and distance. As a rule, *all* four quarter corners and at least one section corner *must* be recovered before any aliquot part of a section can be made.

For example, Figure 7.10 highlights the SE$^1/_4$ SE$^1/_4$ SE$^1/_4$ SE$^1/_4$ of the section. If the original survey information were relied upon, one might believe that this parcel formed a square 330 feet (5 chains) on a side and encompassed 2.5 acres. An improperly prepared metes and bounds supplemental description of this parcel might read, in part, "... commencing and Point-of-Beginning at the southeast Section corner; thence, West 330 feet; thence, North 330 feet; thence, East 330 feet; thence, South 330 feet to the Point-of-Beginning, encompassing an area of 2.5 acres. ..."

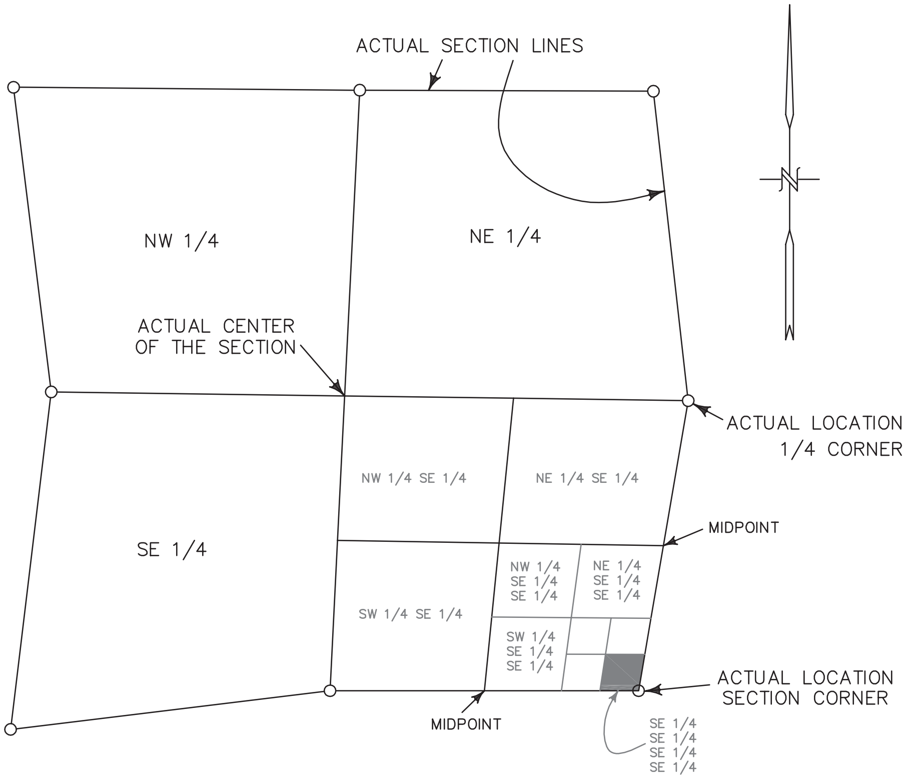
ACTUAL SECTION LINES
NW 1/4
NE 1/4
ACTUAL CENTER OF THE SECTION
ACTUAL LOCATION 1/4 CORNER
NW 1/4 SE 1/4
NE 1/4 SE 1/4
MIDPOINT
SE 1/4
NW 1/4 SE 1/4 SE 1/4
NE 1/4 SE 1/4 SE 1/4
SW 1/4 SE 1/4
SW 1/4 SE 1/4 SE 1/4
ACTUAL LOCATION SECTION CORNER
MIDPOINT
SE 1/4 SE 1/4 SE 1/4 SE 1/4
N

FIGURE 7.10

The novice might argue that one need only recover the southeast section corner, and, by measuring west, north, east, and south the 330 feet called for, this parcel could be monumented. This is simple but quite incorrect, as the very large discrepancies in Figure 7.11 clearly demonstrate.

In order for the described SE$^1/_4$ SE$^1/_4$ SE$^1/_4$ SE$^1/_2$ of the section to be monumented, all four quarter corners must be recovered, as well as the southeast section corner.

A "straight line" must be run between opposite quarter corners to establish the center of the section. The midpoint of each of the four sides of the resulting southeast quarter must then be set and a "straight line" run between opposite midpoints to establish the center of the southeast quarter.

The midpoints of all four sides of the resulting southeast quarter of the southeast quarter must then be established and a "straight line" run between these opposite midpoints to establish the center of the southeast quarter of the southeast quarter of the section.

The midpoints of all four sides of the resulting southeast quarter of the southeast quarter of the southeast quarter must be established and a "straight line" run between opposite midpoints to establish the

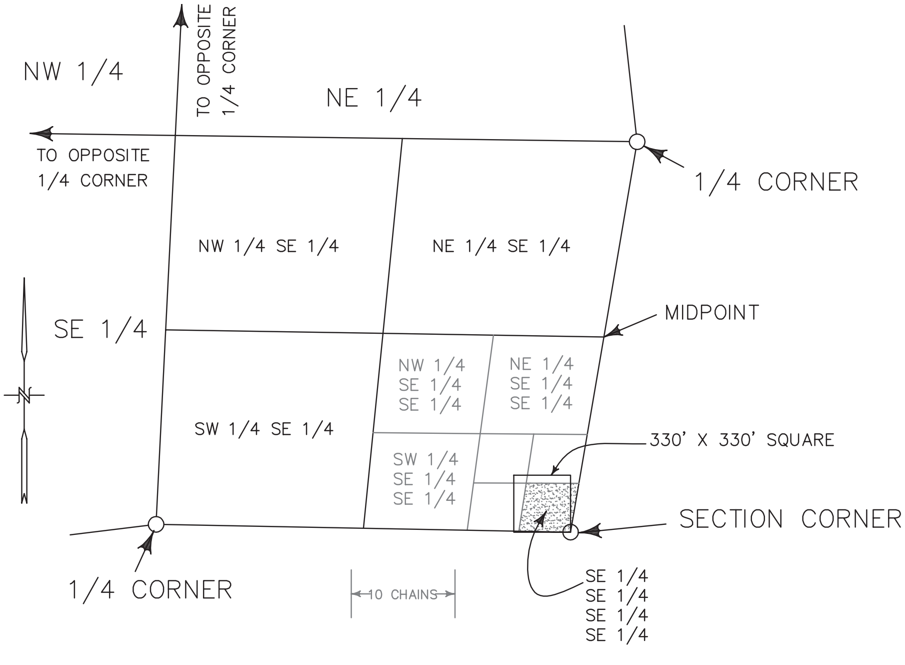
TO OPPOSITE 1/4 CORNER
NW 1/4
NE 1/4
TO OPPOSITE 1/4 CORNER
1/4 CORNER
NW 1/4 SE 1/4
NE 1/4 SE 1/4
MIDPOINT
SE 1/4
N
NW 1/4 SE 1/4 SE 1/4
NE 1/4 SE 1/4 SE 1/4
SW 1/4 SE 1/4
330' X 330' SQUARE
SW 1/4 SE 1/4 SE 1/4
SECTION CORNER
1/4 CORNER
10 CHAINS
SE 1/4 SE 1/4 SE 1/4 SE 1/4

FIGURE 7.11

center of the southeast quarter of the southeast quarter of the southeast quarter of the section. Only this procedure will result in the correct monumentation of the described parcel.

Most of the steps that established the midpoints of land lines and the “runs” between opposite sides can be, and almost always are, the results of computations, as opposed to actual field monumentation. The only field work that is absolutely necessary in this case is the recovery of all four quarter corners of the section, the southeast section corner, and the monumentation of the three remaining corners of the parcel. Survey plats or maps of USPLS sections or their aliquot parts that report exactly the same distances and directions as the original government plat are almost certainly the result of incorrect work.

If the reader believes that Figure 7.9 offers an overly pessimistic view of a typical section, instead of a regular occurrence, now is a good time to recall the conditions under which most of this work was done.

The bulk of the work was done under contracts that paid by the mile. Nineteenth-century work was typically contracted at less than 10 dollars per mile. Out of this fabulous sum, the government surveyor had to provide for the payroll and equip the crews with chains, compasses, tents, food, mules, wagons, brush knives, axes, posts, firearms, and other necessities. Whatever remained was the government surveyor contractor’s profit. It may be possible that some of the surveys were hurried or that some of the work didn’t follow the general instructions to the letter. Perhaps there may even be one or more instances of fraudulent field notes being returned to the surveyor general’s office.

Considering the wild condition of the land, the rough terrain, the crudeness of the instrumentation and techniques, the hostility of the indigenous population, and the demand for haste, one can only be amazed that any dimensions and directions reported were even close. The wisdom of relying on the actual location of the original monumentation over the purported dimensions becomes more evident as the examination of the physical limitations of the original surveys is reviewed.

7.2.10. Nonaliquot Division of a USPLS Section

Care must be taken when regular sections are being divided that the intention to divide along lines that are not aliquot parts is not confused with those that are. The best assurance of this is the creation of a subdivision plat, the assignment of lot numbers to the parcels created, and the recording of the plat. The written instruments that document the sale could stress the intent to deviate from the normal process of

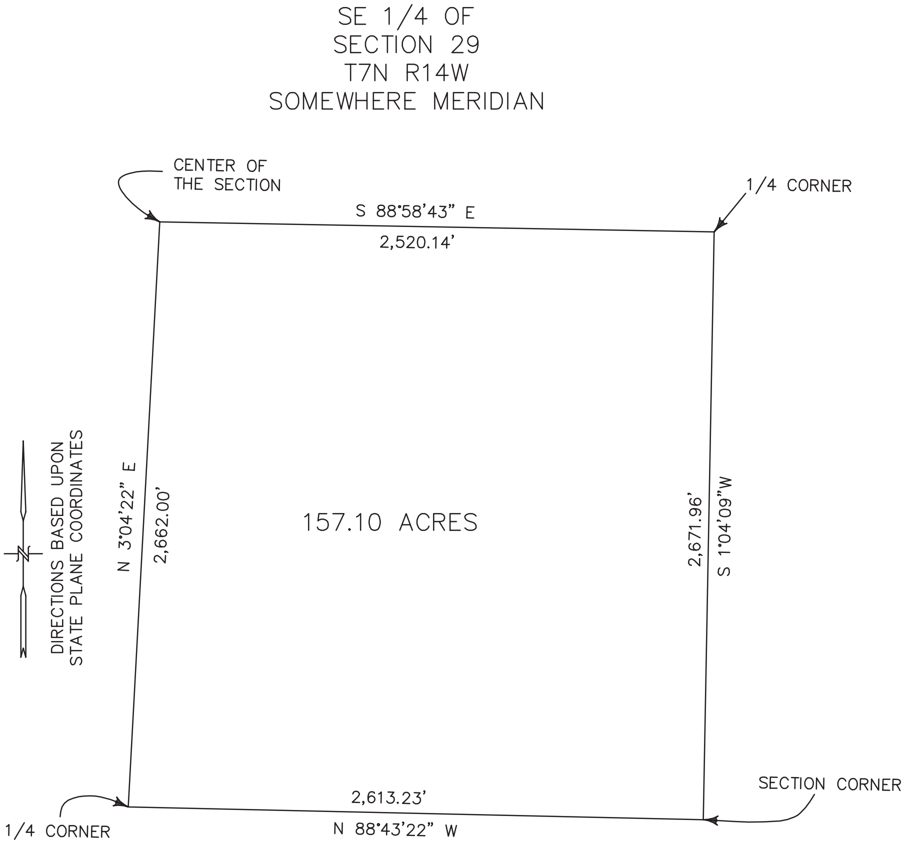
SE 1/4 OF
SECTION 29
T7N R14W
SOMEWHERE MERIDIAN
CENTER OF THE SECTION
1/4 CORNER
S 88°58'43" E
2,520.14'
DIRECTIONS BASED UPON STATE PLANE COORDINATES
N 3°04'22" E
2,662.00'
157.10 ACRES
2,671.96'
S 1°04'09" W
SECTION CORNER
2,613.23'
N 88°43'22" W
1/4 CORNER

FIGURE 7.12

dividing regular sections. Finally, all reference to fractions, such as $^1/_2$ or $^1/_4$, should be avoided. The use of examples may better clarify situations that dictate deviation from the aliquot methods.

Figure 7.12 is of the SE$^1/_4$ of Section 29, T7N; R14W Somewhere principal meridian. Naturally, there has never been a survey conducted on this quarter section, since the government surveyor set the corners in 1890. Everyone "knows"[26] that this quarter section is 40 chains to a side and contains 160 acres. The dimensions shown in Figure 7.12 are the actual dimensions that would be revealed by a modern survey.

The owners have two sons, and they want to give each an equal portion of the property. Suppose number one son is deeded the N$^1/_2$ SE$^1/_4$ Sec. 29 and number two son is deeded the S$^1/_2$ SE$^1/_4$ Sec. 29. Figure 7.13 shows the resulting division.

---

[26] Remember Will Rogers.

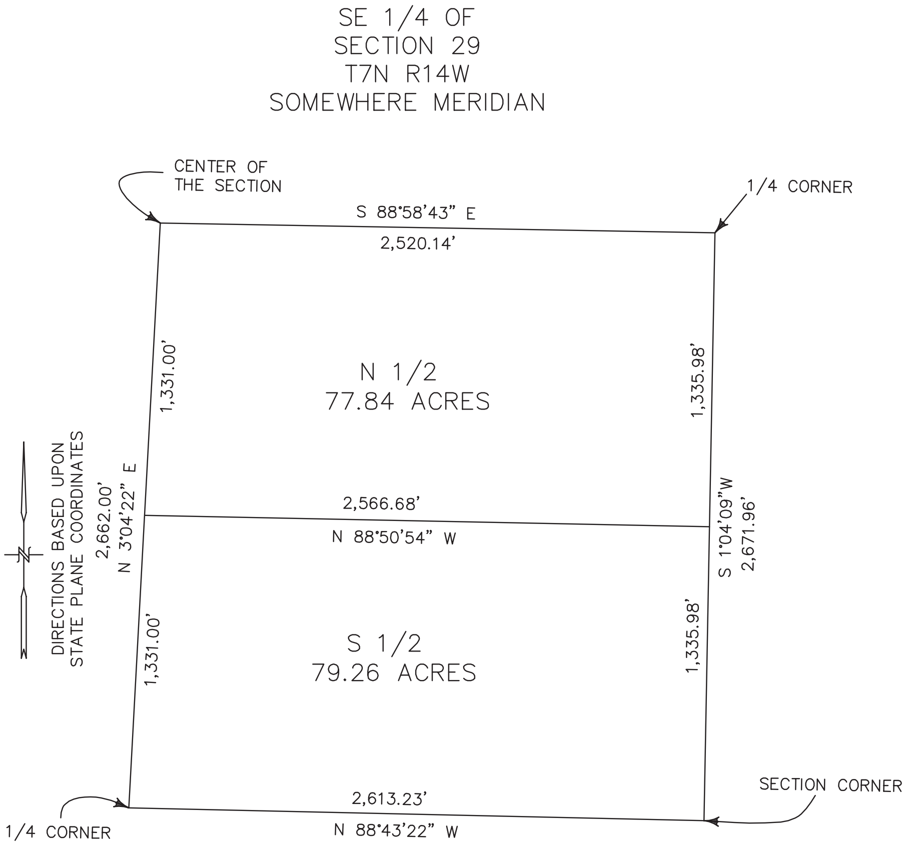
SE 1/4 OF
SECTION 29
T7N R14W
SOMEWHERE MERIDIAN
CENTER OF THE SECTION
1/4 CORNER
S 88°58'43" E
2,520.14'
N 1/2
77.84 ACRES
1,331.00'
1,335.98'
DIRECTIONS BASED UPON STATE PLANE COORDINATES
2,662.00'
N 3°04'22" E
2,566.68'
N 88°50'54" W
S 1°04'09"W
2,671.96'
1,331.00'
1,335.98'
S 1/2
79.26 ACRES
SECTION CORNER
2,613.23'
N 88°43'22" W
1/4 CORNER

FIGURE 7.13

The division is not equal. The situation could be even more complicated if number one son were to measure south 1,320 feet from the north line of the SE$^1/_4$, fence the property, and begin cultivation, confident that he is in possession of his 80 acres. In reality, he possesses 77 acres, because the fence is 10–15 feet north of the correct location. This seems trivial until number two son sells the land to a manufacturing company that builds a warehouse 5 feet from the fence.

7.2.11. Private Claims

More often than not, especially in the Mississippi Valley, the government surveyor encountered lands that had been claimed and settled prior to any surveys. In most cases, the private claim was just bypassed, as if it were a large lake. This disruption of the "regular" pattern caused the formation of irregular sections. The section numbering system was adhered to, as nearly as possible, in these cases, with some notable exceptions. The treaty that concluded the Louisiana Purchase required the United States to honor all existing land grants and claims and to conduct surveys of these parcels so that the U.S. Congress could confirm these claims.

This resulted in some very interesting township plats. Surveys of heavily populated areas often produced entire townships where not one regular section occurred. Rarely did the U.S. Congress or the government surveyor make any attempt to rectify preexisting boundary disputes arising from the overlap of private claims and royal land grants. Each confirmed claim was given its own section number, causing some townships to have well over 100 sections.

Plate 7.1 is a photo-reduction of a township plat that was approved by the surveyor general of Louisiana on August 19, 1853. The field work and notes that were used to develop this plat were begun in 1829 and completed in 1853. Note the meander line along the Mississippi River as well as the absence of a regular section. The table of contents reports 5,165.58 acres of private claims and only 323.41 acres of public lands in this township.

Divisions of irregular or "fractional" sections are provided for in the regulations controlling the division of public lands, but such highly irregular sections as are shown in Plate 7.1 defy any semblance of regular division. In many areas, where such settled conditions existed long before the USPLS implementation, the owners, attorneys, and even the courts continued to use the metes and bounds system of property identification. It is very common, in southeast Louisiana, for example, for a real property chain of title to be devoid of any reference whatsoever to section number, township, or range. These areas, although officially in USPLS states, are acting as metes and bounds areas.

7.3. PLATTED SUBDIVISION OR URBAN SYSTEMS

The metes and bounds system and the USPLS discussed so far are systems that are most effective when they are used for large tracts of land. The original land grants of the colonies and the USPLS sections typically transferred land for some agricultural use. Both systems allow for division into very small parcels, but they lose effectiveness or become unwieldy as the size of the parcel decreases. This is especially true for the USPLS. The formal USPLS rules only specify the aliquot division process to the one-sixteenth of a section. The division of

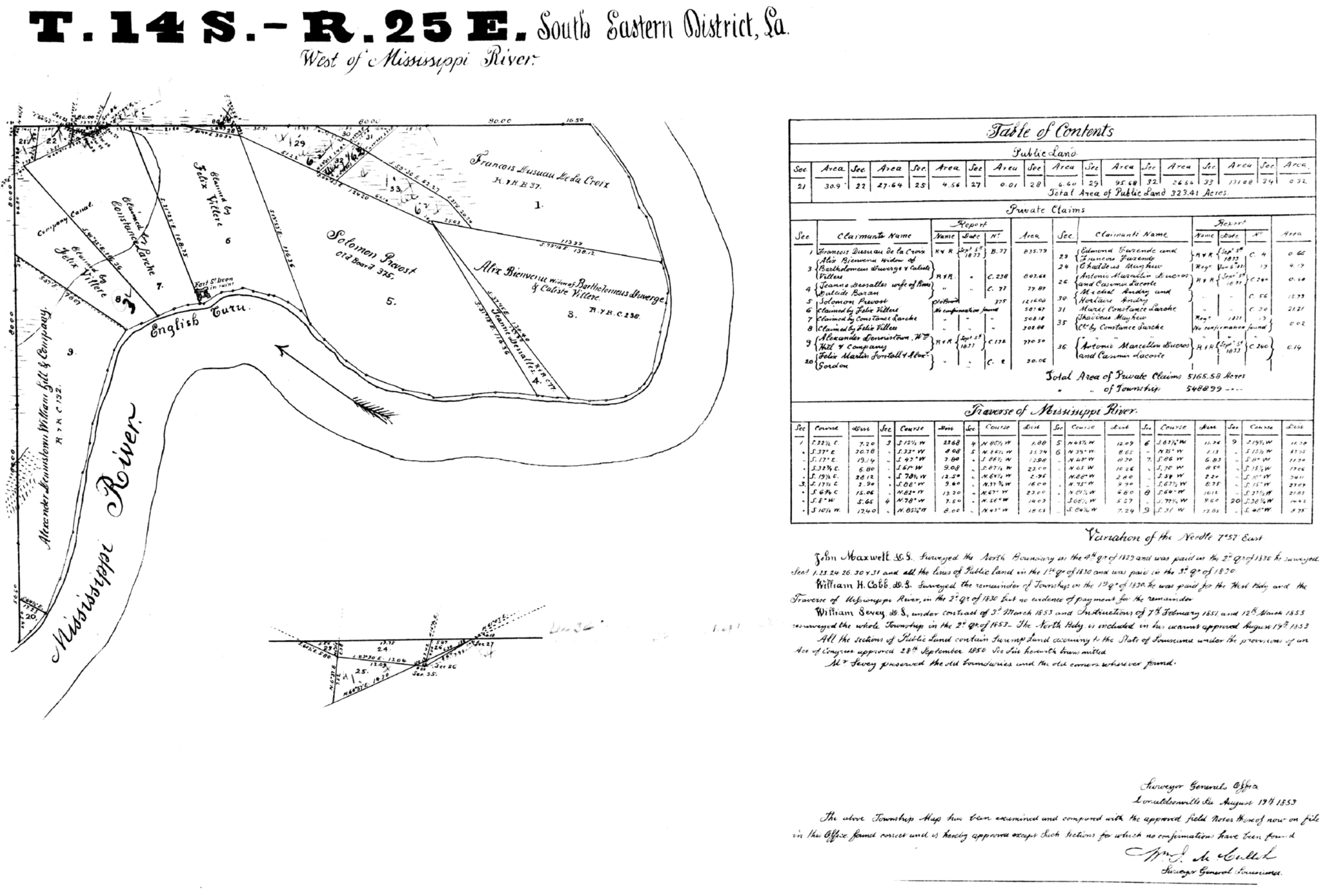
T. 14 S. - R. 25 E. South Eastern District, La.
West of Mississippi River.
Francois Dusuau De La Croix
R. & H. B. 37.
1.
Solomon Prevost
Old Board 375.
5.
Alix Bienvenu widow of Bartholomeus Duverge & Caliste Villere.
3.
R. & H. C. 238
Jeanne Dessalles
4.
Claimed by Felix Villere
6
Claimed by Constance Larche
7.
Claimed by Felix Villere
8
Company Canal
Alexander Dennistoun William Hill & Company
R & H C 132.
9
English Turn
Fort St Leon
Mississippi River.
20
22
29
33
24
25
Sec 26
Sec 27
Sec 35
80.00

Table of Contents

Public Land

Sec	Area	Sec	Area	Sec	Area	Sec	Area	Sec	Area	Sec	Area	Sec	Area	Sec	Area	Sec	Area
21	30.9	22	27.64	25	4.56	27	0.01	28	6.60	29	95.68	32	26.56	33	131.08	34	0.32

Total Area of Public Land 323.41 Acres.

Private Claims

Sec	Claimants Name	Report			Area	Sec	Claimants Name	Report			Area
		Name	Date	No.				Name	Date	No.	
1	Francois Dusuau de la Croix	R & R	Sept 5th 1833	B. 37	835.79	23	Edmond Fazende and Francois Fazende	R & R	Sept 5th 1833	C. 4	0.66
3	Alix Bienvenu widow of Bartholomeus Duverge & Caliste Villere	R & R.	„	C. 238	807.62	24	Chaddeus Mayhew	Regr	Jan 6th 1821	19	4.19
4	Jeanne Dessalles wife of Pierre Duluide Baran	„	„	C. 77	77.87	26	Antonio Marcellin Ducros and Casimir Lacoste	R & R	Sept 5th 1833	C. 240	0.90
5	Solomon Prevost	Old Board		375	1216.00	30	Michel Andry and Hortaire Andry	„	„	C. 56	15.73
6	Claimed by Felix Villere	No confirmation found			587.67	31	Marie Constance Larche	„	„	C. 30	21.21
7	Claimed by Constance Larche	„	„	„	508.18	35	Thaddeus Mayhew Cd by Constance Larche	Regr No confirmation found	1821	19	0.02
8	Claimed by Felix Villere	„	„	„	302.88						
9	Alexander Dennistoun, Wm Hill & Company	R & R	Sept 5th 1833	C. 132	770.50	36	Antonio Marcellin Ducros and Casimir Lacoste	R & R	Sept 5th 1833	C. 240	0.14
20	Felix Martin Fontall & Alexr Gordon	„	„	C. 2	20.06						

Total Area of Private Claims 5165.58 Acres
„ „ of Township 5488.99 „

Traverse of Mississippi River.

Sec	Course	Dist	Sec	Course	Dist	Sec	Course	Dist	Sec	Course	Dist	Sec	Course	Dist	Sec	Course	Dist
1	S.32½ E.	7.20	3	S.12½ W	22.68	4	N.80½ W	1.88	5	N.43½ W	12.09	6	S.63½° W	11.76	9	S.19½ W	11.38
„	S.37° E	30.78	„	S.33° W	8.08	5	N.66½ W	15.74	6	N.39° W	8.65	„	N.75° W	1.13	„	S.13½ W	33.95
„	S.17° E.	19.14	„	S.47° W	7.80	„	S.86½ W	12.98	„	N.48° W	8.70	7	S.86 W	6.83	„	S.8° W	11.70
„	S.32¾ E.	6.80	„	S.61° W	9.08	„	S.87½ W	23.00	„	N.65 W	10.26	„	S.70 W	8.50	„	S.15½ W	17.06
„	S.19½ E.	28.12	„	S.78½ W	12.50	„	N.64½ W	2.96	„	N.88° W	2.80	„	S.54 W	2.20	„	S.10° W	24.11
3	S.17½ E	2.90	„	S.88° W	9.60	„	N.77¾ W	16.00	„	N.75° W	9.90	„	S.67½ W	8.75	„	S.15° W	27.87
„	S.67¾ E	15.06	„	N.82° W	13.20	„	N.67° W	23.00	„	N.81½ W	6.80	8	S.64° W	10.12	„	S.27½ W	21.85
„	S.8° W	5.65	4	N.78° W	7.50	„	N.56° W	14.07	„	S.68½ W	5.57	„	S.72¼ W	9.60	20	S.38¾ W	14.42
„	S.10½ W.	17.40	„	N.85¼° W	8.00	„	N.45° W	18.05	„	S.84¼ W	7.24	9	S.31 W	12.65	„	S.48° W	8.75

Variation of the Needle 7°57' East

John Maxwell D.S. Surveyed the North Boundary in the 4th Qr of 1829 and was paid in the 2d Qr of 1830 he surveyed Secs 1.23.24.26.30 & 31 and all the lines of Public land in the 1st Qr of 1830 and was paid in the 3d Qr of 1830.

William H. Cobb D.S. Surveyed the remainder of Township in the 1st Qr of 1830 he was paid for the West Bdy and the Traverse of Mississippi River, in the 2d Qr of 1830 but no evidence of payment for the remainder

William Sevey D.S., under contract of 3d March 1853 and Instructions of 7th February 1851 and 12th March 1853 resurveyed the whole Township in the 2d Qr of 1853 - The North Bdy is included in his account approved August 19th 1853

All the sections of Public Land contain Swamp Land accruing to the State of Louisiana under the provisions of an Act of Congress approved 28th September 1850 See List herewith transmitted

Mr Sevey preserved the old boundaries and the old corners wherever found.

Surveyor Generals Office
Donaldsonville La August 19th 1853

The above Township Map has been examined and compared with the approved field Notes thereof now on file in this Office found correct and is hereby approved except such sections for which no confirmations have been found

Wm J. McCulloh
Surveyor General Louisiana

PLATE 7.1

USPLS sections into smaller fractions is only an extension of these rules and can become quite clumsy (see Figure 7.7).

Very early in the metes and bounds system, and later in the USPLS, landowners often wished to divide their large tracts into smaller parcels. One of the ways in which this was frequently done was to produce a subdivision plan or plat in which the smaller parcels were identified by a lot number or some other identification. This plat was then recorded in the local conveyance office, and any title transfers would refer to the subdivision name and lot number.

This system has many advantages. In urban situations, where public streets are created, the addition of the concept blocks greatly reduces the verbiage needed to describe any particular lot. The plat can be used by the owner as a sales aid. Buyers can readily see on the plat, if not on the ground, the configuration of the property being purchased. The tax collector can use the plat to develop a record of the ownership and relative worth of the properties. Deeds that refer to lot and block identifiers may be supplemented by a metes and bounds description, which refers to street right-of-way corners and adjoining lots to confirm the lot and square.

The platted system of real property identification has one very important distinguishing feature that greatly separates it from the metes and bounds system and the USPLS. Unlike the previous two systems discussed, the platted subdivision does *not necessarily depend on monumentation or possession* to create boundaries to the property. Before local governments began to exercise their authority, it was not uncommon for owners to draw a platted subdivision of their property solely based on the old metes and bounds system or USPLS description. It may be that *none* of the corners in the subdivision was ever monumented. It is just as possible that the platted subdivision was properly performed and that all of the newly created corners were monumented. The consequences of both cases will be discussed at length in Chapter 10.

Platted subdivisions should be separated into two additional categories. Subdivisions in which all of the lots were platted and recorded at one time are very different from subdivisions in which the lots were created one at a time. If the former is true, none of the lots can claim "seniority of title."[27] If the lots are created and sold one at a time and some chronological sequence can be established for the transfer of title, then a seniority of title can be established.

---

[27] "Seniority of title" refers to the concept that, in the sequential division of land, the parcel sold first must conform, as closely as possible, to the monumentation and dimensions called for in the deed. Parcels sold at a later date must yield to the senior deed in cases of conflict.

Survey markers set at the corners of a platted subdivision do not have the same ranking as corner monumentation by the government surveyor in the USPLS, nor does a survey marker set hold the same status as a corner called for in the metes and bounds description. A platted subdivision usually expresses the intent to convey a parcel of property of specific dimensions. The survey markers, if any are set, reflect a corporeal attempt to place on the ground the limits intended by the survey plat. Any blunder in the setting of these survey marks can be corrected at any time, provided that correction is geometrically possible and sufficient acquisitive acts have not taken place.

Frequently, only the street rights-of-way were monumented by the original subdivider. This created the blocks or squares of the subdivision but did not monument the individual lots. In general, it is public reliance on the monumented location of the street rights-of-way that will, in most cases, causes any excesses or deficiencies to be isolated within each block. It is because of this public reliance that block corner monumentation usually ranks above lot corners in importance during boundary recovery.

In order for a Land Surveyor's marker to mature into the kind of artificial monument that is referred to in the hierarchy of calls, reliance on that marker and the rules of acquisitive prescription must be satisfied. A private Land Surveyor's marker must satisfy several conditions. "Not every cross upon a stone does a corner make, nor each iron pipe a monument" is an amusing but wise phrase.

Unlike USPLS or original private claims, platted subdivisions typically create land parcels that are quite small. Because of their small size, along with the fact that most platted subdivisions are relatively modern, it is more likely that resurveys of individual lots will discover the dimensions to be identical, or at least very close to, the dimensions reported on the original subdivision plat. More often than not, this is because the typically short distances encountered in modern platted subdivisions reduce measurement errors to values that are less than the inherent errors produced by centering on corner monuments.

For example, let us assume that a particular lot in a platted subdivision is reported to be 200 feet from an intersection, 100 feet wide and 200 feet deep. The block corners were originally monumented by one-inch diameter iron pipes and the lot corners by $^3/_4$ diameter iron pipes. A Land Surveyor, recovering the monuments and measuring 200.02 feet and 99.99 feet, respectively, will, in all likelihood, report the original 200 and 100 feet, because the variation from the record is within normal standards of precision.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 8

THE GLOBAL POSITIONING SYSTEM (GPS)

8.1. OVERVIEW

The Global Positioning System (GPS) in operation around the world first began as a U.S. military navigational program. The American constellation of satellites is still operated, maintained, and supported by the United States Air Force.

Other nations, or groups of nations, have begun to establish their own navigational systems based on the U.S. model. All of the systems were, and are, primarily navigational. The use of these systems for precise, centimeter-level positioning was completely unanticipated.

Initially, the U.S. government, for security reasons, would scramble signals or not release frequencies and codes in an effort to reduce the positioning precision of nonmilitary GPS receivers. The ingenuity of the private sector was not to be denied. The methods and equipment developed were able to achieve land survey–level precision by essentially eavesdropping on the signals without the ability to "read" all that was being sent.

The modern Land Surveyor uses signals sent from every positioning satellite constellation in orbit without regard to which government(s) is maintaining that system. Land Surveyors and users of survey information must have a broad understanding of the capabilities and liabilities associated with this very powerful addition to the Land Surveyor's toolbox.

8.2. FUNDAMENTALS

Unlike all of the other systems established to navigate the earth, GPS is not a surface-based system. Until the space age, every mechanism used to map and navigate the earth was anchored on the surface of the planet.

Latitudes were based on the rotation of the earth as determined by observations of the stars made while standing on the ground. Longitudes were established by arbitrarily naming a location the starting point and then measuring the time differentials between when certain stars were vertical at an unknown site and the arbitrarily chosen meridian.

The positioning system chosen (latitude, longitude, and elevation) was spherical and particularly useful for observers that were on, or very near, the surface of the earth.

GPS has approached the problem of positioning from a radically new direction. The system is based on the center of the earth, not a location on the surface. The coordinate system used is a three-dimensional rectangular system. One axis of the system coincides with the axis of rotation of the earth. The other two axes are fixed at defined, but essentially arbitrary locations. All of the axes are mutually perpendicular.

GPS receivers use the signals from the satellites to compute the receiver's location in terms of an x, y, and z. It is only through a very complicated and rigorous set of computations that these rectangular spatial coordinates are translated into the spherical geodetic positions we use.

The translations to elevation are even more complicated. The geodetic positions are on a mathematically defined surface, the ellipsoid, and a function of distance from the center of the earth. Elevations are on a surface, the geoid, which undulates according to the gravitational attraction of the planet. It is only through extensive and comprehensive measurements that the relationship between the gravitational surface and the theoretical surface can be mapped.

This is a continuing process. As the observations become more numerous, the mapping of the differences provides the Land Surveyor with the information needed to precisely measure elevation differences over greater distances. Some areas of the earth have been mapped more densely than others, so GPS derived heights are not yet uniformly precise. This will change as use increases.

8.3. GPS AND LAND SURVEYS

GPS broadcasts signals laden with data that capable receivers use to compute the receiver's position. The precision of these locations are to within 20 meters or so without any additional information. This is certainly not precise enough for land surveys. The addition of a corrections-receiving radio to the GPS receiver can improve the level of precision to about five meters or so. Land surveys require the level of precision to be a few centimeters. That precision is not yet available for a single receiver.

The deliberate signal variations and natural anomalies that perturb the location signals as they enter the atmosphere on the way to a receiver are so unpredictable as to render centimeter-grade raw positioning very difficult. The solution that the survey profession hit upon was nothing less than brilliant.

Let us suppose that one receiver is placed on a point of known geodetic position (base station) and others placed at locations of unknown geodetic positions (rover stations). If the ones placed at the unknown locations are not too far from the one that is at a known position, then the distortions of the signals from a common satellite should be about the same for each receiver. The errors in the computed position caused by ambient conditions would be the same for both. Therefore, the differences in the values computed for the position of each must be the result of the relative distances between the receivers. In this manner, the geographical relationship between the receivers can be very precisely computed.

It is in this differential mode that the Land Surveyor has adapted what was intended as a navigational aid into a mechanism by which distances and directions can be measured between points on the earth's surface that are not intervisible. The only requirement is that the receivers are operating at the same time, under similar atmospheric conditions, in locations with a clear view of the sky, and they are observing enough of the same satellites.

Horizontal distances of several miles can now be measured with a precision comparable to line-of-sight electronic instrumentation but with greatly reduced expenditures of effort and time. Depending on the level of precision required, the distances measured can be continental.

Fifty miles from the base station is usually considered as the outer range of reliable horizontal differential measurements. Vertical differences are much more susceptible to distortion and the distance from the base station is reduced to five miles. Both of these limits are functions of range, procedure, and equipment. The future promises pronounced improvements in the range at which reliable horizontal and vertical data can be collected.

The adaptation of GPS techniques to land surveys is rapidly advancing. There are presently a multitude of variations and improvements to the differential application of GPS technology for land surveying purposes. (Those interested in a more thorough explanation are referred to Jan Van Sickle, *GPS for Land Surveyors,* 3rd ed. Boca Raton, FL: CRC Press, 2008, as an excellent starting point.)

8.4. CORS NETWORKS

GPS technology, as adapted by the Land Surveyor, is dependent on at least one receiver's being at a known location. It often occurs that the effort required to recover and occupy a station of known location that is GPS observable and within range of the area to be surveyed is significant. The idea arose that instead of hunting up NGS control stations with a good view of the sky, the Land Surveyor could determine the geodetic position of a permanently mounted receiver antenna in a central location.

The rover stations then become the only portable part of the operation. The base station could be housed in a secure building, like the home office, further reducing the personnel required and time expended performing a set of GPS measurements. As long as the base is receiving satellite data at the same time as the rovers, land survey grade differential measurements are possible.

From this came the concept of continually operating receiver stations (CORS). In this case, the base station is never turned off. Rover stations could then be operated when conditions were right and the data compared to the base later. Better yet, a network of CORS with overlapping ranges could be used by a single rover to develop a position that was not dependent on one fixed location.

Today, CORS systems have been established throughout the United States and elsewhere. Local, state, and other organizations, in cooperation with the National Geodetic Survey (NGS) have systems in operation that will allow any GPS operator to submit the digital data downloaded from their receiver to the NGS for processing. This system is Internet based and maintained by NGS under what is known as the Online Positioning User Service (OPUS).

This powerful and useful service offered by NGS has done as much as or more than any other technology of the 21st century to advance the capability of the Professional Land Surveyor. Geodesy and boundary surveying have become closer in goals and capabilities. Procedures, data collection, and map projections systems have been somewhat standardized as a result of superior results, not mandates. Indeed, the many standards, regulations, and rules established by governing authorities or professional organizations lag far behind the rapidly advancing state of technology.

8.5. PRACTICAL APPLICATION

The survey example in section 6.11 was based on the availability of the NGS horizontal control stations “Billy” and “Bob.” The network of NGS control stations throughout the United States was typically concentrated along the coastlines and transportation corridors. The vast majority of the land area was not well supported by control stations.

More often than not, the Land Surveyor did not have monuments of known geographic position to rely on within a reasonable distance of the project. This is no longer the case. The ability of the GPS receivers to measure the distances between locations that are many miles apart to a level of precision that, over such spans, was heretofore unheard of provides NGS supported geographic positions to every Land Surveyor everywhere in the United States or the world.

Figure 8.1 is an example of how high-quality State Plane Coordinate (SPC) data can be developed for surveys that are in remote locations. The remainder of the survey work follows the same procedure as the random traverse outlined in section 6.11.

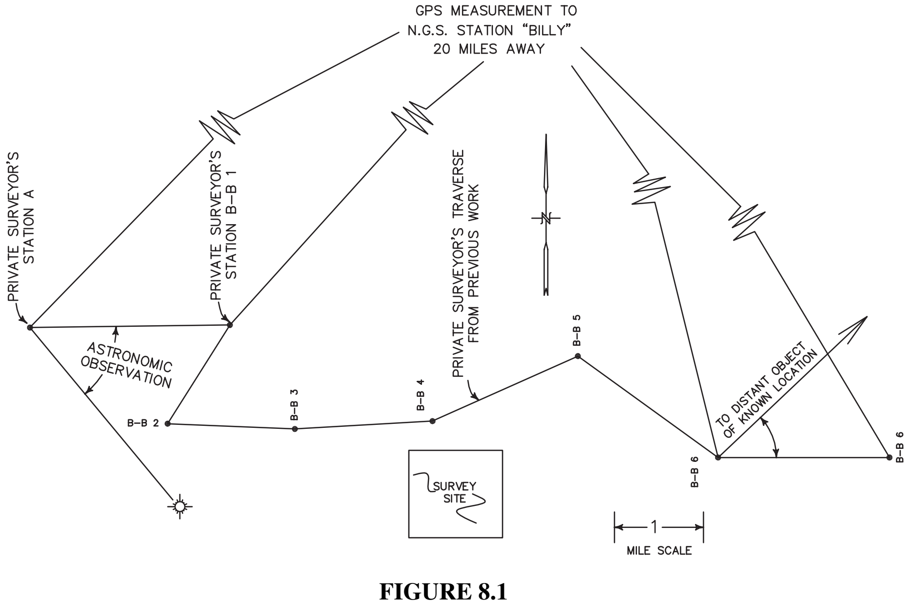
GPS MEASUREMENT TO
N.G.S. STATION "BILLY"
20 MILES AWAY
PRIVATE SURVEYOR'S STATION A
PRIVATE SURVEYOR'S STATION B-B 1
PRIVATE SURVEYOR'S TRAVERSE FROM PREVIOUS WORK
N
ASTRONOMIC OBSERVATION
B-B 2
B-B 3
B-B 4
B-B 5
B-B 6
TO DISTANT OBJECT OF KNOWN LOCATION
B-B 6
SURVEY SITE
1
MILE SCALE

FIGURE 8.1

Notice that the bearings between the control stations have been verified by a method independent of GPS technology. Independent verification is an economic alternative to the multiple (at least three) GPS observations recommended to develop sufficient redundancy of observations. Even small horizontal positional errors between stations in close proximity can produce directional errors that can be magnified during the traversing process, so redundancy in developing positional data or directional observations that are not dependent on stations in close proximity are used.

Many states, perhaps soon to be all states, are creating CORS networks. One variation of the GPS differential measurement procedures, known as Real Time Kinematic (RTK) positioning uses radio communications between the base station and the rover to develop position solutions that are displayed on the rover as soon as the solution is computed. The operator is provided with the receiver location within seconds: “in real time.”

RTK has proven to be a very effective topographic data collection and construction control stake-out tool. This procedure requires a full and uninterrupted view of the satellite constellation to be effective. However, even when used in close proximity to the control stations, RTK is not as precise as are other positioning methods. The CORS network developers are investigating a radiophone version of RTK for general use.

Positioning technology is advancing with almost daily developments and improvements. When it is appropriate to use one or more of these powerful additions to the Land Surveyor’s options is a professional decision that will be based on more than just the client’s needs. In the case of recorded real property transactions and activities, the land boundary information obtained by the Land Surveyor must be capable of assimilation into the community’s record and recovery systems.

8.6. STRENGTHS

The great advantages presented by GPS technology are ones of scale and sustainability. Great distances between land points are no obstacle. No longer must locations be intervisible in order to measure between them. The adoption and use of national geographic positioning systems, such as SPC, is within the reach of every survey and Land Surveyor.

Locations recovered and documented using reliable nationally supported positioning criterions are more easily recovered and can be compared to data developed by other Land Surveyors at different times.

This data can be directly imported into local, state, and federal Geographical Information Systems (GIS) used for everything from land record keeping to infrastructure maintenance.

8.7. WEAKNESSES

The ease with which GPS observations develop precise geographic positions often causes a user to become overreliant on the numbers that come out of the computer. The temptation to use a position obtained from a single GPS observation and "run with it" is great.

Notice that in the example in section 8.5 the Land Surveyor had four locations that were processed and tied together by standard line-of-sight (LOS) measurements into a network. If one of the GPS observed locations contained large errors due to a regional anomaly, the inconsistencies between computed and LOS measured distances and directions would alert the Land Surveyor to a problem. Manufacturers of GPS receivers correctly tote the very small deviations that are attained with "95 percent certainty."

In other words, 5 percent of the time the data has location errors that exceed the published level of precision. Statically, in a set of four measurements that are "95 percent certain" the odds are one in six that one of the four contains an error that exceeds the 95 percent deviation. The magnitude of any particular 5 percent deviation may be significant. For this reason, no horizontal or vertical position derived from a *single* GPS session should be considered sufficient for land boundary survey control.

8.8. COPING WITH REALITY

The constraints that face a Land Surveyor during a land boundary survey are several. The time, effort, and expense associated with the degree of precision obtained by survey procedures must be economically and legally justified. Once the level of precision has been determined by the Land Surveyor, the equipment and procedures are tailored to meet that need.

In the GPS world, the level of precision is relatively high, with the exception of an occasional anomaly. Sometimes, for reasons that have not been adequately explained at the time of this writing, the data set from a GPS observation will provide a solution that is far beyond the expected error limits. These "fliers" are rare but can be devastating.

Prudent GPS procedures require a level of redundancy in observation that will identify and eliminate such fliers from the control network. There are many methods of obtaining redundant observations but the most effective and economical combine GPS observations with LOS measurements.

For example, GPS data can be collected at three intervisible stations. This is most effective when the data sets are collected on different days. LOS measurements can then be made between the stations recording both distances and angles. If the LOS results are consistent with the computed GPS positions, the data set is verified. If there is a flier, the remaining two stations that are in agreement can be used.

LOS verification of the direction between two intervisible GPS stations by astronomic or distant sight points is another very effective means of verifying GPS results. LOS verification of the distance between two GPS stations is a must, but distance alone is insufficient to verify GPS results. LOS distances combined with angular or bearing measurements are both required to confidently verify location data.

Sun or Polaris observations are highly accurate and very simple. The geodetic azimuths are easily reduced to SPC bearings. The precision of the directions obtained by astronomic measurements is commonly repeatable to within 5 seconds or so. This is more precise than the directions derived by inversing between high-precision GPS stations that are less than 2,000 feet apart.

Distant sight points, such as television or radio transmission towers, are usually associated with published geodetic positions that have been determined to a high degree of precision. The bearing to these locations can be computed and used to verify the relative positions of GPS stations within the survey network. Distant sight points often become "permanent back sights" for Land Surveyors working on extended projects in an area.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 9

THE GEOGRAPHIC INFORMATION SYSTEM (GIS) REVOLUTION

9.1. BUILDING A FOUNDATION

The profound increase in the sheer amount of land information that has become available to everyone has been nothing less than a tsunami. Users of land information are now confronted with the task of assimilating, cataloging, and storing all this information in a way that it can be recovered and used.

The systems that have been developed to control land information are collectively known as Geographic Information Systems (GIS). It is "geographic" because the information is linked to the earth. This is not just geography; this is everything that can be associated with a location.

Technically speaking, every map is a geographic information system. Any agency that has had to study or manage land-related matters would develop maps on which pipelines, streets, crops, population, or whatever was being investigated would be indicated by symbols. The list of things that can be mapped is endless, and any information that can be mapped is geographic information. The electronic process is analogous to a base map over which selected transparencies containing the information of interest are placed.

The big improvement in the electronic GIS programs is that anybody can use it. The data can be stored in many different locations by the owning agency and summoned by users in far-off locations with ease.

The linchpin, the keystone, the foundation, the one piece of information that links it all together is *location*. In the United States, the base map of an effective GIS program is geodetic positioning controlled by the National Geodetic Survey (NGS)-supported North American Datum of 1983 (NAD 83) geodetic model. Jurisdictions will often convert the geodetic positions of latitude and longitude into the more user-friendly State Plane Coordinate (SPC) system. But, nevertheless, precise and reliable positioning must be available in order to initiate an effective GIS program.

9.2. SOURCES OF INFORMATION

The Electronic Age has prompted the many different owners of geodetic information (maps) to convert that information into digital format. First by digitizing existing paper scaled maps and then by developing maps by computer-aided drafting (CAD) programs, what were once paper drawings stacked in filing cabinets slowly became replaced by digital data sets that could be viewed only by computer programs.

The precision and reliability of the data found in these data sets are very much a function of the means and methods used to develop them. One of the primary methods of developing digital data from a paper map is known as “digitization.” In this process the paper drawings are placed on a pad laced with sensors and the lines on the paper are traced. Digitized electronic maps can never be more precise than the original paper plan.

Digitized maps can then have information added to the electronic version that was not a part of the original paper map. Municipal, county, state, and federal agencies often place the data associated with their function over digitized topographic maps, aerial photographs, or other sources of control that mimic geodetic control as the base when, in fact, a digitized map is the base map.

The distances, locations, areas, elevations, and bearings computed from a GIS program can be only as precise as the original source of that information.

9.3. SYSTEM MAINTENANCE

GIS programs must be maintained to be of any use. This may be best explained by example. The Public Works Sewerage Division of a municipality instituted a GIS database for all of the facilities in their system. The first step was to collect all of the existing maps of the sewer lines they own. These maps were digitized and then electronically scaled and rotated to fit a street map (also digitized) of the municipality. The locations of the lines in this initial layer are no more precise than the old paper maps.

In order to improve the precision of the GIS data, it was decided to determine the geodetic position of all manholes and visible structures. The cost of using a survey crew to do this work was prohibitive. Aerial photography was used instead, and the geodetic position of items determined by scaling or overlaying the photo onto the base map. This reduced the locative uncertainty from tens of feet to a few feet. Even though this was not as precise as it could be, it was a great improvement over what had existed.

Ideally, anytime repair work was done on the system, a survey crew would precisely determine the geodetic position of the exposed pipes. However, because of cost, handheld *navigational gps* units were employed to collect the new data. These radio adjusted units provided locations with a precision of less than five feet. This was about the same level of precision as that provided by the aerial photography.

New work by developers or under facilities improvement contracts required surveyed geodetic positions, converted to SPCs, for all new structures. This was made easier when the municipality adopted a continually operating receiver station (CORS).[1]

It was in this way that the example agency developed and continues to maintain the GIS database of their facilities.

9.4. POTENTIAL USERS

It may be simpler to list data users that would not utilize a GPS data source. If the item of interest involves "where" in any of its attributes, it is an item that can be, and soon will be, a part of a GIS database somewhere. From soil types to rainfall patterns to ownerships to demographics, data will be, and is being, complied in a GIS digital format. It is vitally important that the persons directing the locations and recordation of these data be aware of the need for a firm foundation, a valid base map, on which this data is to be presented.

9.5. POTENTIAL FOR MISAPPLICATION

GIS is a powerful tool and, as it is with many powerful tools, if mishandled can cause much grief. The user need not be aware of all of the nuances of the data collection procedures, but a firm understanding of the *limitations* of the particular fragment of the GIS data base is vital.

---

[1] See section 8.4.

An inquiry made electronically by "clicking" on a property line, for example, may result in a answer of a distance to the one-hundredth of a foot and a direction to the second, yet the GIS data may have originated with a digitized paper map wherein the original drafting precision was dozens of feet one way or another. GIS data often use SPCs or other rectangular map projections. Even distances that are precisely determined on the earth's surface will be adjusted to some degree because of the requirements of the map projection in use.[2]

Users of GIS sources must know where the data comes from and apply the results of a query accordingly. Most people, particularly the novice, have the disconcerting habit of believing anything that a computer spits out. GIS users must constantly strive to wisely assess sources and application.

---

[2] See section 5.5.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 10

BOUNDARY RECOVERY

Contrary to popular belief, Land Surveyors do not establish boundary lines. Boundary lines can only be established by the landowners, although disputes may be settled by adjudication.

The act of subdividing a parcel of land by a private Land Surveyor acting as an agent of the property owner, or the monumentation of section corners by the government surveyor in the case of federal lands, is the first step in the process of establishing the boundaries of newly created parcels. These are the "original steps" that all descendants in title, and the Land Surveyors working for them, are bound to follow.

The Land Surveyor is charged with the task of recovering those boundaries, not establishing them. The original boundaries can be established only once when the owner divides the lands. All surveys after that original establishment are boundary recoveries.

A boundary recovery by a Land Surveyor constitutes a *professional opinion* and is no more binding on the property owners than the advice of any other professional. Every property owner has the right to reject a purported boundary recovery if that owner so desires.

10.1. RULES OF EVIDENCE

Boundary recoveries are opinions rendered based on the results of an investigation. The rules of evidence that are applied to that investigation are not very different from the rules of evidence in a civil proceeding. Every item discovered during an investigation is evidence. There are no exclusionary rules for a Land Surveyor. Parole evidence, hearsay, private documents, recorded documents, and anything else may be considered or may point the way to further investigation and discoveries. Physical evidence is generally more reliable than testimony, written evidence is given more credence than parole, and recorded evidence is given more weight than private, and so on.

The hierarchy of calls (see section 7.1) used to assist in the interpretation of deed descriptions also guides the Land Surveyor in evaluating boundary evidence. Natural and artificial monuments are physical evidence. A survey plat may be a physical thing, but the measured angles, distances, and computed areas shown on that plat are the recorded observations (testimony) of the previous Land Surveyor. Neighbors and landowners often offer undocumented (and sometimes biased) opinions with regard to the location of boundaries. The Land Surveyor weighs all of the evidence and, based on the quality and quantity of the evidence, renders an opinion on the location of a boundary.

10.2. SOURCES OF EVIDENCE

Title examinations are usually restricted to recorded documents. The deed document and recorded survey plat are far from the only source of evidence used by the Land Surveyor in recovering real property boundaries. Because the location of real property boundaries is defined by the location of the real property *corners*, boundary recovery is, in fact, *corner recovery*. The primary source of evidence is the physical features discovered by the Land Surveyor in the field. Prior to conducting a field search for evidence, the Land Surveyor will use several other sources of evidence in an attempt to narrow the area of search for corner monumentation.

The deed document and previous surveys are obvious sources of evidence. Land Surveyors who have practiced in an area for a long time will have extensive files that may include deeds and plats of adjacent properties. The county records are also an important source of information on contiguous tracts of land. Land Surveyors can often be seen looking through the public property records book. They are not looking for a chain of title; they are searching for evidence pertaining to the location of the corners. Adjacent landowners often have plats or other documents that add details about the location of real property corners.

10.3. OFFICE PROCEDURE

Of all the questions commonly asked of the Land Surveyor, “Where do you start?” has to rank as number one.

The Land Surveyor is usually told the current land record identification of the particular parcel to be surveyed. That identification may be a lot number and subdivision name, an aliquot part of a section, a metes and bounds description, or just “the Jones’s land.”

As in a Greek tragedy that begins at the end before it returns to the beginning, the Land Surveyor must trace back through the public records and his or her private files to the original tract from which the present parcel was created. This “paper trail” may require many hours of research before any fieldwork is started. This step is greatly shortened if the Land Surveyor has previously surveyed other parcels of the same original tract. The value of prior knowledge in a particular area is a greater influence on the “regionalization” of private Land Surveyors than any other factor.

The theoretical starting point for any survey is the same whether a USPLS, metes and bounds, or platted subdivisions are involved. That theoretical starting point is the *record containing the creation of the original parcel*.

In the USPLS, that starting point is the government surveyor’s original notes. In the case of the metes and bounds states, that starting point is the original land grant. In the case of the platted subdivision, that starting point is the original subdivision plat. In cases where the platted subdivision was not monumented, recorded, or public use and acceptance of the created streets has not taken place, the starting point reverts to the parent tract.

10.4. FIELD PROCEDURE

Once the Land Surveyor has gathered as much recorded and other written information as possible pertaining to the location of the real property corners, the field search for evidence of boundaries can begin. Each discovery is used to narrow the search for other corners.

The most difficult corner to find is usually the first corner. After the first corner is found, the area of search is greatly narrowed for subsequent corners, facilitating their recovery. Potential artificial corner monumentation must be compared with the record for consistency before it is accepted. Among the questions that need to be answered are, “Who set the monument?,” “Has it been disturbed?,” “Is its location consistent with other monuments?," "Is it of the same material as called for in the plat?," and so on.

Suspected corner locations are examined for evidence of the monumentation of record. The advent of the electronic metal detector has been as great an advancement in the land surveying profession as the computer and the electronic distance measuring device. Large areas can be searched electronically for iron-based monumentation, and the area excavated can be greatly reduced.

However, in the final analysis, the most important tool in the recovery of boundaries is the shovel. There are very few corner monuments that survive standing above the ground. The threat of earth moving equipment, brush cutting machinery, and other hazards—unintentional and otherwise—has led to the practice of burying boundary monumentation in order to ensure that the location of that monument is undisturbed.

The Land Surveyor may establish a control traverse first (see section 6.11), and/or measure from known corners or accessories to isolate the search areas, compute theoretical positions for the corners, and then intensify the search in those areas. Recovery of corners that were originally documented by state plane coordinates is faster and more reliable when the control traverse used in the recovery attempt is also based upon the state plane projection system.

There is no secret formula for boundary recovery, and the choice of measurement procedure made by the Land Surveyor is in many ways incidental to the success of the boundary recovery. A successful recovery is one that finds the natural or artificial monuments at the corners of a real property parcel.

10.5. RENDERING A DECISION

Rarely is every corner monument found in any boundary survey. The ravages of time and the activity of agriculture or land development frequently destroy many of the artificial monuments placed by the original surveyor. This does not mean that the location of the corner is lost.

The recovery of accessories to the corner reported by the original Land Surveyor is a recovery of the monumentation for that corner, even if the artificial monument at the corner is gone. Often, the location of buildings, tree or fence lines, or other terrain features in relation to the corner being sought were reported on previous surveys of the subject or adjacent tracts.

Certainty of recovery is directly proportional to the relative relationship of reported terrain features to the corner being sought. Recovery of the original and undisturbed monument set at the corner is absolute. Other features, such as witness trees or buildings that are a known distance and direction from the corner can be used to recover a corner.

Features that are nearer a corner, geographically and chronologically, are more likely to lead to a correct corner recovery than those that are more remote from the corner being sought. A known corner one mile away is not nearly as reliable as a witness tree 10 feet from a corner. A survey plat claiming to have recovered an original monument produced one year after that original corner was set is usually more reliable than one produced 100 years later.

The time and effort spent in searching for physical evidence of the location of a corner is the dominant factor. The greatest source of dispute among Land Surveyors about the location of real property corners can usually be traced to a failure to discover all of the physical evidence relevant to a corner location.

The tendency to simply "lay out" the deed dimensions of a parcel or to resort to mathematical computations that theoretically set the location of a corner often causes much confusion when, at a later date, additional physical evidence reveals the true location of the real property boundaries.

A Land Surveyor's professional ability is a function of his or her thoroughness in the collection of physical evidence more than any other single factor. Excellence in mathematics, beautiful plats, sophisticated electronics, or professional appearance are meaningless unless a Land Surveyor's boundary recoveries are based on all of the pertinent evidence that can be found.

Rendering a decision on the recovery of a property corner is an art. Formal education, continuing education, and extensive study will shorten the time it takes to excel in the art of corner recovery, but only experience in a particular geographical region will develop the skill needed for consistent and certain corner recovery. It is both proper and wise to request that a Land Surveyor inform you of the evidence and reasoning that led to the Land Surveyor's decision on the location of recovered corners.

10.6. USPLS BOUNDARY RECOVERY

Any survey of a USPLS section or *any aliquot part thereof* begins with the recovery of section corners and quarter corners. The best record of where any corner was set is the original notes of the government surveyor. Plate 10.1 is a photocopy of facing pages from a typical government surveyor's notebook.

Although most government surveyors' notes are arranged in this form, some notes may not be. The first column lists the direction of the line being "run." This line was typically a magnetic observation that was "corrected" for the local magnetic deviation. The "correction" involved marking on the ground a "true north" (actually astronomic north) line by observing the star Polaris at the headquarters of the government surveyor. The reading of a compass, usually very large and considered more "precise" than the surveyor's field compass, was then compared to the "true north" line, and the deviation of the needle from the "true north" line was recorded. This is why the deviation or declination recorded in many government surveyors' notes are reported to the nearest minute when the compass used in the actual work could only be read to the nearest quarter degree. Of course, the survey work was usually many miles away from where the magnetic "deviation" was measured.

The second column lists the distances along that line that terrain features or other items of note that were encountered as measured from the beginning of the land lines. These terrain features encountered along the survey route are known as "passing calls."

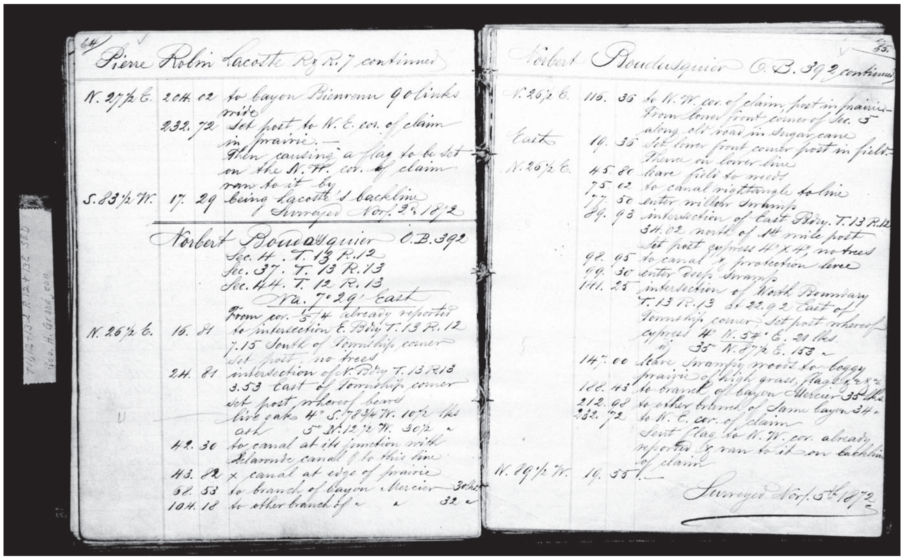

64

Pierre Robin Lacoste Rz R. 7 continued

N. 27½ E.	204.	02	to bayou Bienvenu 90 links wide
	232.	72	Set post to N. E. cor. of claim in prairie.— Then causing a flag to be set on the N. W. cor. of claim ran to it by
S. 83½ W.	17.	29	being Lacoste's backline

Surveyed Nov. 2d 1872

Norbert Boudousquier O.B. 392
Sec. 4. T. 13 R. 12
Sec. 37. T. 13 R. 13
Sec. 44. T. 12 R. 13
Va. 7°29' East

			From cor. 1/5/4 already reported
N. 26½ E.	16.	81	to intersection E. Bdy T. 13 R. 12 7.15 South of Township corner set post, no trees
	24.	81	intersection of N. Bdy T. 13 R. 13 3.53 East of Township corner set post whereof bears live oak 4" S. 78¾ W. 10½ lks ash 5" N. 12½ W. 30½ "
	42.	30	to canal at its junction with Delaronde canal ∥ to this line
	43.	82	× canal at edge of prairie
	68.	53	to branch of bayou Mercier 30 lks.
	104.	18	to other branch of " " 32 "

Norbert Boudousquier O.B. 392 continued 65.

N. 26½ E.	116.	36	to N. W. cor. of claim, post in prairie— From lower front corner of Sec. 5 along old road in sugar cane
East	19.	36	Set lower front corner post in field— Thence on lower line
N. 26½ E.	45.	80	leave field to weeds
	75.	02	to canal right angle to line
	77.	50	enter willow swamp
	89.	93	intersection of East Bdy. T. 13 R. 12 34.02 north of 1st mile post Set post cypress 4" × 4", no trees
	98.	95	to canal & protection levee
	99.	30	enter deep swamp
	141.	25	intersection of North Boundary T. 13 R. 13 at 22.92 East of Township corner, Set post whereof cypress 4" N. 54° E. 21 lks. " 35" N. 87½ E. 153 "
	147.	00	leave swampy woods to boggy prairie of high grass, flags &c &c
	188.	43	to branch of bayou Mercier 35 lks.
	212.	98	to other branch of same bayou 34 "
	232.	72	to N. E. cor. of claim Sent flag to N. W. cor. already reported & ran to it on backline of claim
N. 89½ W.	19.	55	

Surveyed Nov. 5th 1872

PLATE 10.1

Passing calls are frequently very useful in recovering corners set by the government surveyor, for they provide evidence of the location of the corner that is much closer than all other evidence, except for witness monuments (trees). If, for example, the notes indicated that at 38 chains from a found section corner the government surveyor crossed a small creek (in survey jargon, “a creek is called for at 38 chains”) and set a quarter corner at 40 chains, the first place to begin the search for the quarter corner is 2 chains past the creek, not 40 chains from the found corner.

The remainder of the page is reserved for a narrative describing and/or identifying the beginning of the line, the features encountered or “calls,” the end of the line, the corners set, and the *accessories to each corner*.

10.6.1. Accessories to a Corner

Accessories to a corner consist of physical objects near the corner to which distances and directions are noted for the purpose of future restoration of the corner. Accessories were most often natural objects, such as trees, and were sometimes referred to as “witness marks” or “bearing marks.” Accessories are as much a part of the corner monumentation as the corner post that may have been set by the government surveyor.

Witness trees are the most common form of accessory to a corner. These trees were chosen for their proximity to the mark, usually less than one chain (66 feet) away, as well as their size and species. In order to distinguish the accessory from other natural objects, the surveyor would scribe on the object specific letters or numbers.

Ideally, three or more accessories accompanied each corner and were chosen so that they encircled the corner. In the narrative portion of the notes, the surveyor lists the distance and direction *from* the corner *to* the accessory. Because the accessories are a part of the corner monumentation, a recovery of the accessories is a recovery of the original comer monumentation.

10.6.2. Recovered Corners

Corners that are reset by the use of accessories or by the discovery of the original, *undisturbed* monument are called “recovered corners.” Corners reset by use of the passing calls should not be considered as “recovered,” for passing calls rarely were in close proximity to the corner.

10.6.3. Deciphering Government Land Office (GLO) Notes

Line 10 of Plate 10.1 is the beginning of a line of a private claim. The line begins at the corner common to sections 1, 4, and 5. This is clearly an irregular section.

The variation of the compass reported by the government surveyor earlier in this field book is 7 degrees 29 minutes east, yet the directions observed for the claim boundary line are reported in increments of one-quarter of a degree. This inconsistency of precision may indicate that the instrument and personnel used to measure the magnetic variation were not the same as those used to measure this particular line.

Later recovery of the line (reported as bearing north 26$^1/_2$ degrees east) showed the actual geodetic bearing at this point to be north 25 degrees 32 minutes 14 seconds east (plus or minus 5 seconds). This difference of only 1 degree between the 1872 magnetically based observation and the modern (1985) geodetic direction for the same line is well within the range of reasonably expected values. Depending upon regional factors, the variation between the historic values and the more precisely determined values could be in excess of a degree or more.

According to these notes, the line intersected the northern boundary of the township at a point 24.81 chains along the line and 3.53 chains east of the township corner. A post was set to monument this intersection, and two bearing trees were reported. The first bearing tree is a 4-inch diameter live oak, 10$^1/_2$ links away at a bearing of south 78$^3/_4$ west. The second is a 5-inch diameter ash, 30$^1/_2$ links away at a bearing of north 12$^1/_2$ degrees west. The post and the trees are the monumentation of this particular point in the form of corner monumentation and accessories.

10.6.4. Passing Calls

At 42.30 chains along this line, the government surveyor encountered a canal "at its intersection with the Delaronde canal." These reports of features encountered along a line are referred to as "passing calls." This particular passing call is important, because both canals may remain long after the post and trees that monument the last point are gone. Passing calls, when contradictory to recovered corner monumentation, will to yield to the corner and its accessories as proof of the location of a corner. This is only logical. Witnesses to an act who are nearest the action will carry more credibility than someone who is far away.

The art of recovering witness trees is highly refined in certain regions. Even a pattern of three or more stump holes that match the bearing and distances of the original witness trees is considered strong, if not conclusive, evidence of corner recovery.

Unfortunately, in many areas, the activities of nature and human beings have erased all evidence of an original corner. This does not necessarily mean that the corner is "lost." Although the government surveyor's notes recount the features that were geographically near the corner, other notes, survey plats, and events in the chain of title contain information that may be chronologically near the setting of the corner.

10.6.5. Perpetuated Corner

A section corner may be perpetuated through a chain of events leading up to features that remain at the time of a recovery attempt.

For example, let us assume that a section corner and accessories are established in 1835. In 1898, a private Land Surveyor, failing to find the original post, recovers the accessories and remonuments the section corner, according to the GLO notes, with an iron bar.

In 1925, a private survey plat is made that notes the iron bar and several brick buildings, and shows several dimensions that relate the location of the buildings to the iron bar.

In 1988, all that remains are the foundations of several of the buildings. These buildings have become, through perpetuation, accessories that monument the location of the original corner.

Corners that are reset by the discovery of a documented chain of evidence linking existing monumentation to the original monumentation by the government surveyor are called "perpetuated corners."

10.6.6. Corner by Common Report

It may be that the trail of evidence leading from the government surveyor's monumentation to the modern monumentation of a section corner is incomplete.

A monument may have been accepted by an entire community as an original section corner without the documentation necessary to justify labeling it as a perpetuated corner. This acceptance by the community may be so complete that all of the aliquot parts created in each of the contiguous sections hinge on that particular monument. If any alternate location for the comer were to be asserted, the whole community would be thrown into confusion.

It is for this reason that, in the absence of better evidence, such monumentation may be accepted as marking the corner. This is known as a "corner by common report."

10.6.7. Corner Reset by Best Evidence

The most common type of corner or boundary recovery made by the Land Surveyor is based upon a location of a corner by the best available evidence. Passing calls, the relationship to other corners, evidence to the location of the corner not found in the government survey notes, and other sources of evidence often are sufficient to reset a monument at a corner location with certainty that the remonumentation is at the same location as the original corner.

10.6.8. Corner Reset by Proportionate Measurement

Finally, if *absolutely no creditable evidence remains* of the original location of a section corner, certain USPLS rules can be applied to "reset" the missing corner.

Unfortunately, these rules, known as "restoration by proportionate measurement," are widely published by the U.S. government as well as possibly every book on surveying ever published, except this one. The word *unfortunately* is used, because the promotion of the rules governing the restoration of lost corners has encouraged many Land Surveyors to resort improperly to proportionate measurement long before they have exhausted all of the possible sources of evidence.[1]

The recovery of the original section and quarter corners may not constitute a recovery of the boundaries or even assist in that recovery if actions took place, without regard to the sections or their aliquot parts, that reestablished the property boundaries. If a boundary agreement or subdivision of the land took place without reliance on the original section corners, or if a boundary was established by adjudication, then the section lines remain but *cease to be boundary lines*.

10.7. METES AND BOUNDS BOUNDARY RECOVERY

The recovery of property boundary lines under the metes and bounds system of land identification varies from that of the USPLS in a few subtle but important ways. Although section corners in the USPLS can never move, the boundary corners of the metes and bounds system are subject to change.

---

[1] Division of Cadastral Survey, Bureau of Land Management, U.S. Department of the Interior, Stock Number 024-011-00012-7, *Restoration of Lost or Obliterated Corners and Subdivision of Sections—A Guide for Surveyors,* 1979.

Because the monuments involved mark only boundary corners of contiguous lands and do not have the dual purpose of serving as control for other parcels, the adjoining owners are free to move or reestablish their common boundary lines. Boundary agreements, both formal and implied, as well as adjudication, can and will define new locations for property corners.

The paper trail from the present to the original parcel will still be present in the metes and bounds system, but the formal notes, regimented monumentation, and consistent use of accessories at each boundary corner are often absent. This places more reliance on old survey plats and descriptions than does the USPLS recovery.

The parcels created under the metes and bounds system are not aliquot parts of the parent tract. The chronology of the creation and selling of divisions of a parcel become very important. In this system, the intentions of the parties to a division and sale of land parcels are not based on a regimented procedure; rather, each division is a unique, individual act. Often, it is quite difficult to decipher the true intent of the parties. These, as well as other complications, makes the recovery of boundaries in the metes and bounds system a true art form.

The same rules of evidence apply as in any boundary recovery. Generally speaking, the closer an object is, chronologically and geographically, to an original boundary corner, the more creditable that object is as a source of evidence to the location of the original corner.

Old maps, plats, descriptions, and accounts, both recorded and unrecorded, may produce evidence pertinent to the recovery of a corner. Corners may be recovered, reset by best evidence, accepted by common report, perpetuated, or reset by proportional measurement, just as in USPLS corner recovery.

10.8. PLATTED SUBDIVISION BOUNDARY RECOVERY

Modern platted subdivisions present the simplest cases of boundary recovery. The requirements of most municipalities or local governments for the development and recordation of subdivisions produce well monumented, clearly defined parcels with typically slight discrepancies between recorded dimensions and measured dimensions. Subdivisions that are well settled and well monumented, and that have clearly delineated boundaries rarely produce boundary conflicts.

Older subdivisions or subdivisions that were not settled and occupied may harbor discrepancies that will become evident only after an attempt has been made to posses several lots of the subdivision.

The recovery of boundaries in subdivisions must begin with the limits of the subdivision block or lot involved. This recovery typically, but not always, involves public street rights-of-way. How much of the block is under physical possession (fences, houses, roadways, etc.) is of great importance to the recovery of lot boundaries. Rarely will any Land Surveyor venture beyond block corners to verify boundary locations if the possession lines are in agreement with the deed dimensions.

The metes and bounds description that accompanies many deeds to lots in platted subdivisions will be considered by most Land Surveyors as a supplement to the recorded plat. Discrepancies between the recorded plat and the metes and bounds supplemental description generally will be settled in favor of the platted data, unless it can be shown to be geometrically absurd.

Just as with all boundary recovery, metes and bounds corners can be recovered by finding the monument set by the original Land Surveyor, recovery of perpetuated monumentation, reset by best evidence, accepted by common report, or reset by proportionate measurement.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 11

EVALUATING SURVEY PLATS

Most of the preceding chapters have dealt, in a broad way, with the history and methods of boundary control and boundary recovery. Boundary recovery is the responsibility of the Professional Land Surveyor. The general information given so far will enable the reader to examine, evaluate, and use survey plats better. As in any profession, the ability, training, and competence of individual Land Surveyors will vary. Even the best and most competent Land Surveyors will occasionally make a mistake, miss a judgment call, or simply overlook something. Never assume that the Land Surveyor is always correct. Each survey plat should be examined carefully for discrepancies or inconsistencies. The following checklist is designed to assist the reader in such an evaluation. Each of the items in the list will be explained in detail.

11.1. PLAT EVALUATION CHECKLIST

1. Determine the land record system used.
2. Determine the map projection used.
3. Evaluate the age of the survey.
4. Determine the purpose of the survey.
5. Examine the survey plat for gross discrepancies and completeness.
    a. North arrow and bearing base.

   b. Legal description.
   c. Date of the survey.
   d. Name and address of the Land Surveyor.
   e. Signature and seal of the Land Surveyor.
   f. Adjoining properties.
   g. Dimensions of all sides of the property.
   h. Bearings of all sides or angles at each corner.
   i. Name of the client or purpose of the survey.
   j. Certification.
   k. Limiting notes or phrases.
   l. Area
   m. Scale.
6. Compare the survey plat with the deed for consistency.
7. Examine the survey plat for indications of easements.
8. Examine the survey plat for indications of encroachments.
9. Determine the level of precision.
10. Determine needs not covered in the plat.
11. Contact the Land Surveyor to discuss any excessive discrepancies or additional requirements.

11.2. DETERMINING THE LAND RECORD SYSTEM USED

It is not enough to know if the parcel in question is located in a metes and bounds state or in a USPLS state in order to determine the land record system that was used to create a particular parcel.

A platted subdivision may exist in any state. A metes and bounds division of a tract of land may be disguised as a platted subdivision by the seller's use of "lot numbers" in the sales. A reference to a portion of a section may refer to something other than an aliquot division, according to USPLS rules. Both platted subdivisions and USPLS parcels may have a metes and bounds description included in the deed as a supplement to the legal description.

The importance of the differences between the various systems is only realized when the actual dimensions and locations of real property boundaries of a particular parcel are substantially different from the information found in the deed. The procedure for rectifying such discrepancies will vary according to the land record system used.

Lots created in a platted subdivision do not follow the same rules as lots that developed sequentially in a metes and bounds system. The words *south half* in a USPLS area mean something quite different from those same words in a metes and bounds system.

11.2.1. Recognizing a Metes and Bounds Description

The key to recognizing a metes and bounds controlled description is the lack of direct reference to the control principles of either of the other systems. Even if a parcel is located in a USPLS section and the section number, township, range, and principal meridian are a part of the deed, the parcel may still be a metes and bounds division of a USPLS section. Similarly, the presence of a lot number and subdivision name does not necessarily eliminate the possibility that a metes and bounds system is the controlling land record system at work in a particular situation.

Figure 11.1 is an example of a metes and bounds division of a USPLS regular Section. The Frank James property can only be described by a metes and bounds definition. This particular situation assumes that the line between sections 5 and 8 and the real property boundary line between P. Short and R. McInnis are the same. The line between P. Short and W. Bonney (720 feet east of the section corner) may be one-quarter of the distance to the quarter corner (a $^1/_{64}$ corner is ideally 660 feet east of the section corner). The description of this parcel may identify it as a portion of the NW $^1/_4$ of section 8, and so on. Even if the bounds called for are section lines for certain sides, the departure from the USPLS recognized method of division causes this to be a metes and bounds described parcel.

If the parties that created the parcel shown as the "Frank James" property intend for it to be bounded by R. McInnis, W. Bonney, and P. Short, it is important that the real property boundary lines be stressed in the metes and bounds description. If only the section line or the $^1/_{64}$ line is called for in the metes and bounds description, confusion may result if the section and $^1/_{64}$ lines are later found to be separate from the real property ownership boundaries.

The series of subdivision plats shown in Figure 11.2 indicate that the original owner of the Smith estate sequentially created and sold portions of the property up to May 1, 1932. The deed to each of the lots contains a reference to the lot number, as well as a metes and bounds description. The fact that the May 1, 1932, survey plat is a recorded document does not cause the various titles created from June 3, 1911, to May 1, 1932, to be anything other than a metes and bounds controlled description. The titles for each lot were created at different

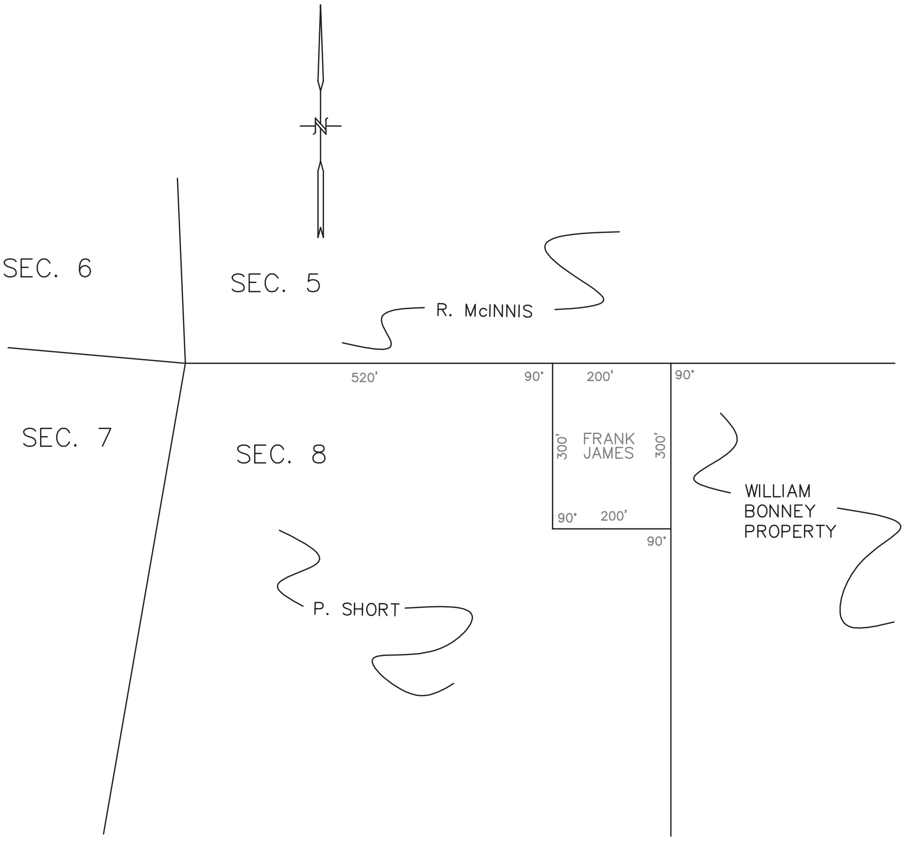
N
SEC. 6
SEC. 5
R. McINNIS
SEC. 7
SEC. 8
520'
90°
200'
90°
300'
FRANK JAMES
300'
90°
200'
90°
WILLIAM BONNEY PROPERTY
P. SHORT

FIGURE 11.1

times, and a precise chronology of that creation can be proven. In order for the division to be considered a platted subdivision, the May 1, 1932, plat must have been created and recorded *before* any of the lots were sold.

If, however, the June 3, 1911, plat had numbered that portion identified as the “Remainder of Smith Estate” as lot number 2, then this would constitute a platted subdivision. The significance of the difference would only be brought to light if the original Smith estate were not 900 feet wide, as the June 3, 1911, survey states it is. If the June 3, 1911, plat were a platted subdivision, then lot 1 and lot 2 would share proportionally the shortages (possession to the contrary excepted).

If the original Smith estate were 850 feet wide, for instance, then the sequence of events portrayed in Figure 11.2 would result in lot 4 being only 200 feet wide.

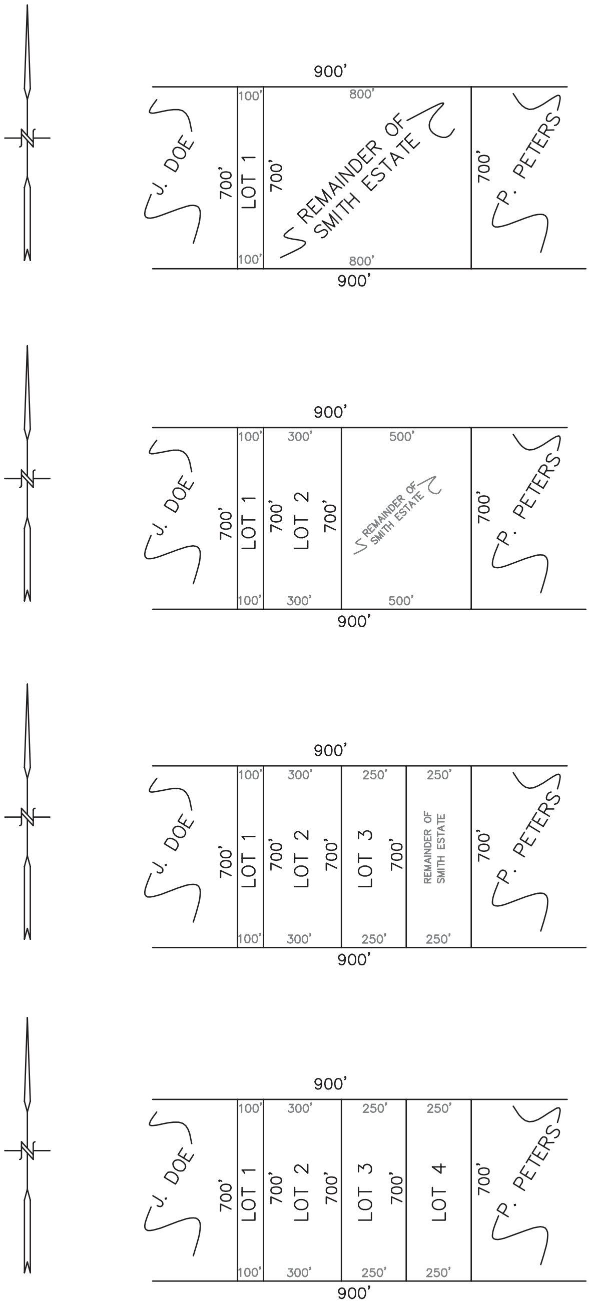
900’
100’
800’
J. DOE
700’
LOT 1
700’
REMAINDER OF SMITH ESTATE
700’
P. PETERS
100’
800’
900’
900’
100’
300’
500’
J. DOE
700’
LOT 1
700’
LOT 2
700’
REMAINDER OF SMITH ESTATE
700’
P. PETERS
100’
300’
500’
900’
900’
100’
300’
250’
250’
J. DOE
700’
LOT 1
700’
LOT 2
700’
LOT 3
700’
REMAINDER OF SMITH ESTATE
700’
P. PETERS
100’
300’
250’
250’
900’
900’
100’
300’
250’
250’
J. DOE
700’
LOT 1
700’
LOT 2
700’
LOT 3
700’
LOT 4
700’
P. PETERS
100’
300’
250’
250’
900’

PLAT SHOWING LOT 1
OF THE SMITH ESTATE
JUNE 3, 1911

PLAT SHOWING LOTS 1 & 2
OF THE SMITH ESTATE
JUNE 15, 1920

PLAT SHOWING
LOTS 1, 2, & 3
OF THE SMITH ESTATE
DECEMBER 5, 1925

PLAT SHOWING
LOTS 1, 2, 3, & 4
OF THE SMITH ESTATE
MAY 1, 1932

FIGURE 11.2

The description of lot 4 would call for the boundary on the east to be that of the western line of the P. Peters property and the boundary on the west to be the eastern line of lot 3. Because lot 3 existed long before lot 4 was created, then lot 3 must be granted all of the width called for, even though the remainder of the Smith estate was not 250 feet wide, as it was believed to be. These subtle nuances, interpreting intent, rely heavily on the system of identifying parcels. For this reason, it is quite important that the system in effect be properly identified.

11.2.2. Recognizing a USPLS

The use of the USPLS must include, as a minimum, the section number, the township number and quadrant, the range number and quadrant, and the principal meridian. In some cases, the principal meridian may not be necessary if the state and county are reported *and* the state or county is served by only one principal meridian. Divisions of sections into smaller parcels must be USPLS lots, in the case of irregular or fractional sections, or must be created by the application of the rules of division governing the aliquot parts of a section.

Divisions of a USPLS section, regular or not, that do not conform to the USPLS regulations must be by platted subdivision or metes and bounds. Division of a USPLS lot is not provided for in the regulations. Sections, or their aliquot parts, must be divided by halves or quarters, based on the *actual length of the sides*, not the theoretical length of the "ideal section." Special rules apply for irregular or fractional sections.

Divisions that are based on the intention to transfer a specific area of land cannot be based on the USPLS rules for the division of a section. A call for one-half of a section is not a transfer based upon a specific area. Because sections are never square, the phrases "south half of the acreage," "south 320 acres," and "$S^1/_2$" all have different meanings. Only the call for the "$S^1/_2$" is a USPLS division.

11.2.3. Recognizing Platted Subdivisions

Naturally, before the rules governing platted subdivisions can apply, a platted subdivision must have been created by someone. It is not absolutely necessary that the platted subdivision be recorded in the public records for it to be valid. It is not even necessary that *all* of the regulatory steps required by the local governing authority be complied with for the platted subdivision to be valid. The determination of the validity of a platted subdivision is a matter of law, and discussion of it is beyond the scope of this book. All of the following discussions about a platted subdivision will assume validity.

In order for a parcel to exist under the platted subdivision system, that parcel must have a unique identifying label. This usually takes the form of a lot "number" or "letter." Lots may also be called "parcels," "plots," or "tracts," among other things. The distinguishing feature that denotes a platted subdivision is the creation of two are more parcels of specific size from one or more existing parcels *where the division is based on a drawing of the parent tract*. The drawing may be the result of a boundary survey conducted at the time of the creation of the subdivision or simply lines drawn on an old map.

Divisions that intend to transfer parcels based on the location of distinctive terrain features are metes and bounds transfers, even if a plat is produced to assist in the transaction. Divisions that adhere to the requirements of the USPLS are not platted subdivisions, even if a survey and a plat are produced and recorded for the parties involved.

The use of a subdivision "name" is quite common, especially in instances in which a large number of lots are involved. The use of squares or blocks is also common but not necessary. The creation of new streets or public rights-of-way is also a common feature of the platted subdivision of a parcel of land. Modern platted subdivisions are usually easily identified because of the many requirements placed on the subdivider by the local governing authority.

11.3. DETERMINING THE MAP PROJECTION USED

The map projection used will have significant relevance to the degree of precision that may be reasonably expected of the information found on a particular survey plat or map. Directions, for example, are exceptionally sensitive to the distortions inherent in the tangent plane projection—so much so that bearings reported on plats using this projection system cannot be reliably compared with plats of adjacent tracts or even other plats of the same property.

11.3.1. Recognizing a Projectionless System

All projection systems commonly used in survey plats require that the bearings shown be based on *angles measured* at the intersection of land lines. Maps that report bearings of the boundary lines as *independently observed magnetic or astronomic directions* are rarely based upon any projection system at all. The most common example of a "projectionless" map is the U.S. government township plat. These plats are simply a graphic rendition of the government surveyor's notes. Each bearing of each line is an independently observed direction (usually magnetic and "corrected" to "true north").

These directions are not two-dimensional; therefore, the novice will easily be misled. Land lines are defined as "straight" between corners, but all of the intricacies described in Chapter 5 apply. Surveys that were performed by "running" the boundary lines with a compass are usually "projectionless" surveys. The distances and directions of such surveys are typically of the lowest order of precision. This is not to say that these survey plats are without value. It may be that the value of the land being surveyed or the purpose of the survey (e.g., cutting of timber) may not warrant the higher cost of more precise surveys.

11.3.2. Recognizing a Tangent Plane Projection System

Tangent plane projection plats are the most common form of survey plats. The earmark of this type of plat is the use of a general reference to a meridian. Notes on the plat that state that the directions are based on "true north" or some other meridian are strong indications that the map is a tangent plane projection. Notes that state that the directions are based on the reported bearing of a recovered line almost exclusively indicate tangent plane projections.

Plats of very small parcels, such as urban or suburban lots, are often tangent plane projections. In the case of a platted subdivision, the directions shown on the plats of individual lots are frequently based on the same projection system as the subdivision plat. In these cases, the directions shown on each lot plat should be consistent with similar plats of other lots in that same subdivision.

11.3.3. Recognizing a State Plane Projection System

State plane projection plats almost always state on the face of the plat that the directions are "state plane" or "based upon grid north." The appearance of state plane coordinates (SPCs) in the form of "x" and "y" values indicates a state plane projection. The SPC that is appropriate for use in an area is typically established by an act of the legislature of each state.

Prior to 1983, all state plane projections were based on the North American Datum of 1927 (NAD 27). After 1983, a new datum was introduced by NGS to replace NAD 27. This new datum is known as the North American Datum of 1983 (NAD 83). During a transitional period, maps or plats may have been based on either NAD 27 or the new NAD 83. After the end of the transition period, only NAD 83 will be allowed. The implementation of NAD 83 is individually legislated in each state.

Today, NAD 27 is no longer maintained by the federal government, and its use in modern surveys is unwise. Computer programs that translate between NAD 27 to NAD 83 are available, but those translations are somewhat *imprecise.* The many distortions in the NAD 27 network were not entirely modeled in the translation programs. A translation program may produce values to the thousandths of a foot, but in reality the translations may be off a foot or more depending on the source of the original NAD 27 value.

11.4. EVALUATING THE AGE OF THE SURVEY

It is vital that the survey data that is being relied upon is the result of work that is contemporaneous with the action that is taking place. Evaluating the age of the survey is a little more involved than simply reading the date of the survey plat. The "age" of the survey is also a reference to the methods employed by the Land Surveyor who conducted the work. The precision of the work, and thereby the confidence that can be placed in the accuracy of the results, is directly related to the "age" of the survey. As has been shown, government land surveys are not very precise when compared with the general level of capabilities of the survey profession during the same period. This may also be true of certain Land Surveyors when compared with their peers.

Even late 20th-century work may have been performed using old methods because of some special circumstance. A boundary line "run" by magnetic compass in 1980 is not more precise than one run in 1910. A distance measured using rigidly controlled procedures and a calibrated steel measuring tape in 1920 could be more accurate than a poorly controlled chaining in 1970. A telephone conversation with the Land Surveyor who performed the work is the best way to determine the "age" of the survey. If that is not possible, the general history of survey procedures in the region may be obtained from local Land Surveyors.

Most lending agencies have strict rules regarding the acceptable amount of elapsed time between the survey and the mortgage loan. These agencies have learned, through experience, that changes in the possession and use of the land can be quite rapid. Buildings can be erected or expanded. Fences can announce the commencement of acquisitive possession. There are a host of activities or occurrences that can alter the use and enjoyment (and thereby the title) of real property.

11.5. DETERMINING THE PURPOSE OF THE SURVEY

The mere existence of a survey plat is an indication that a contract, verbal or written, was formed between a Land Surveyor and persons interested in some feature of the parcel involved. *If you did not request the survey, the possibility that the Land Surveyor performed the work you want is very small.*

The variety and degrees of complexity in land surveys are such that many of the particular items of interest to you, such as easements or flood hazard, may not have been reviewed by the Land Surveyor, because the person who requested the work was not concerned about those particular details.

If the survey plat being reviewed was not ordered for the specific purpose to which it is being applied, then the survey information on that plat must not be important to the users. If a specific reason for the performance of a survey exists, the Land Surveyor must be completely aware of that purpose so that she or he may tailor the work to suit the needs given. Often, items that are of particular concern are not discoverable by a survey alone.

11.5.1. Land Title Surveys

Surveys that are part of a title examination for the purpose of a change in ownership of a parcel of land require coordination between the Land Surveyor and the title attorney. These particular surveys are known as "land title surveys."

To avoid confusion over the shared responsibilities in a land title survey, the American Land Title Association (ALTA) and the National Society of Professional Surveyors (NSPS), a member organization of the American Congress on Surveying and Mapping (ACSM), have jointly established and published the criteria for land title surveys. Called "Minimum Standard Detail Requirements for ALTA/ACSM Land Title Surveys", the publication is revised from time to time to reflect changes in technology and law. Most state regulatory authorities have also adopted similar standards. In many cases, these standards reflect many of the points or edicts of the ALTA/ACSM publication.

Title attorneys will often request "ALTA surveys" out of habit and tradition. ALTA/ACSM contracted surveys are very specific and well thought-out. Certifications, actions, division of responsibility, and standards are spelled out. Persons requesting “ALTA surveys” are well advised to obtain *the most current version* of the document *and read it!*

11.6. EXAMINING THE SURVEY FOR GROSS DISCREPANCIES

The surveying profession must recruit from the human race. This severe limitation means that Land Surveyors are fallible and may, from time to time, omit vital information or miss obvious discrepancies on the face of a survey plat or within a report. The most difficult blunders to detect in the survey profession occur during the final printing or typing of a survey plat or report.

Examine every survey plat with a critical eye. Take special joy in detecting spelling errors, misprints, obvious omissions, or the transposition of numbers or letters. Distances or bearings that don’t seem possible deserve an immediate response from the Land Surveyor. If the discrepancy is the result of a refinement or a correction of a former value, the Land Surveyor will tell you. Most of the items listed in this category are self-explanatory.

11.7. NORTH ARROW

The north arrow shown on a survey plat serves a twofold purpose. Primarily, the north arrow serves the purpose of orienting the plat. The north arrow also serves as an indicator of the bearing base used by the Land Surveyor in conducting his or her work. Changes in bearing base or the use of an imprecise meridian will be reflected in changes in the values presented for the directions of lines. These changes are not relocations of the lines, just an assignment of new values. The absence of a north arrow makes the development of a metes and bounds description quite difficult.

Considered by most Land Surveyors as an essential item on any survey plat, the absence of a north arrow is sometimes common in cases of surveys in very dense, urban areas. If the survey plat that you are using lacks a north arrow and you feel that it needs to be shown, call the Land Surveyor and have it added.

11.8. LEGAL DESCRIPTION

Depending on the land record system in use, the survey plat must show the information necessary to identify the parcel. This does not mean that the complete verbiage of a proper metes and bounds description should be shown on the face of a survey plat in a metes and bounds land records system. A general or abbreviated version of the description will suffice to identify the parcel. The words, sketches, and lines drawn on the survey plat itself will complete the identification.

In platted subdivisions and USPLS records systems, the survey plat caption ought to be in the form of lot, square, subdivision name, county and state or aliquot part, section, township, range, principal meridian (may be omitted), county, and state.

Rules and regulations that require that the “legal description” be shown on the face of a survey plat are referring to the land identification requirements of the system in use. Unfortunately, some have interpreted this requirement to mean that a complete metes and bounds description must be printed on the face of every survey plat. The absolute impracticality of this misinterpretation will be better demonstrated in the exercises in writing metes and bound descriptions found in Chapter 13.

11.9. DATE OF THE SURVEY

The date shown on the survey plat serves to “fix in time” the data represented on the plat. Just as a photograph will present images that are locked in time, a survey plat contains the three dimensions of physical space as well as the fourth dimension of time. As survey plats age, the data shown on the plat may become obsolete. Buildings, fences, waterways, and rights in property all may be altered by time. A photograph of a 2-year-old girl may not accurately portray that same person 30 years later. Changes over time are to be expected in land parcels as well.

11.10. NAME OF THE SURVEYOR

Most survey firms produce survey plats that are drawn on standard paper that has the border, company logo, address, and other information already printed on it. It is surprising how many times this can lead to misunderstanding or oversight. In every survey firm, the responsibility for the collection and interpretation of data rests with one person. That person must be a Professional Land Surveyor registered in the state where the land is located. That individual’s name must be shown on the survey plat.

Throughout this book, the need for communication between the user of a survey plat and the Land Surveyor who performed the work is stressed. This is possible only if the Land Surveyor is identified and the means of contacting her or him are clearly presented. A map that does not identify an individual as the Land Surveyor responsible for the drawing is not a survey plat.

11.11. SIGNATURE AND SEAL

The signature and seal have a special importance in land surveying. This formal act of “signing” and “sealing” is a warranty by the Land Surveyor that the information presented is the result of investigations, research, and measurements faithfully conducted according to the standards and procedures of the surveying community existing at the time that the survey was conducted.

This is a guarantee that the surveyor *believes* that the information is a reasonably accurate report of the facts. Deliberate misrepresentations, distortions, misleading information, deletion of relevant facts, or fraud are pledged not to be present. This is a commitment of faithful service made to all who may see the plat. *It is not a promise of infallibility*.

11.12. ADJOINING PROPERTIES

The identification of all of the adjoining properties is normally consistent with the land record system used to identify the subject parcel. Metes and bounds plats usually identify the owners of adjacent properties by name. USPLS-based plats often show the entire section, with emphasis on the aliquot part or government lot that is being surveyed. Platted subdivision lots may refer to only the lot numbers of the neighboring lots. Platted subdivisions that consist of squares or blocks, as well as lots, may only refer to the square or block boundaries instead of the lots immediately adjacent to the lot being surveyed. It is also possible that the adjoining lots may be identified by a combination of systems.

In all cases, the property is bounded on all sides by something. Be it a private owner, a public right-of-way, or the international seas, those adjoiners need to be identified on a survey plat.

11.13. DIMENSIONS OF ALL SIDES

The dimensions of all the boundary lines of the parcel surveyed should be shown. If the dimension of a particular side cannot be measured, then an estimate and an explanation for the lack of a measured dimension should be shown. The parameters of the dimensions should be consistent. If the measurements are reported in U.S. survey feet, tenths and hundredths, then all sides should be in U.S. survey feet, tenths and hundredths. If directions are reported by bearings, then the bearings of all sides should be shown.

An exception to this rule occurs in cases in which, for historical reasons or for clarity, dimensions that are reported in the deed record but are no longer in use are shown on the plat, along with the modern measured value. A frequent form of this exception is use of the notations "actual," meaning the dimension measured during the present survey, and "deed" or "title," meaning the recorded or previously reported dimension for that same line segment.

11.14. BEARING OR ANGLES

The bearings shown on a survey plat must also be consistent in both parameters (e.g., degrees, minutes, seconds) and meridian. A bearing shown to the nearest degree should not appear where other bearings are shown to the nearest second. The precision of the measurement implied in the first case is much less than that implied in the latter. For example, a bearing of north 25 degrees east must be shown as "north 25 degrees 00 minutes 00 seconds east" in order to be entirely consistent with a bearing shown elsewhere as "north 27 degrees 15 minutes 33 seconds west." Naturally, presenting bearings in degrees, minutes, and seconds in one place and grads, for instance, in another, is a gross violation of the "consistent parameter" rule.

It is also awkward to define directions of lines by a combination of angles and bearings. Clarity and consistency are best served when one method or another is consistently used throughout a survey plat.

A common historical exception to this general principle is the "implied direction or dimension." Often, particularly subdivisions involving small parcels, lines are parallel or angles are clearly 90 degrees and dimensions are repetitive. In these cases, omission of bearings, angles or dimensions were frequently accepted as minor or unimportant.

Plats of urban or suburban home sites are usually restricted to "legal size" paper ($8^1/_2$" × 14"), thereby limiting the amount of information that can be reasonably shown in such a small space. The use of "implied direction" is one way in which much of the clutter was eliminated from small drawings. Almost all of the modern standard detail requirements promulgated by regulatory authorities prohibit implied directions or dimensions.

11.15. NAME OF CLIENT, PURPOSE OF SURVEY

The name of the person for whom the survey was performed or a particular statement expressing the purpose of the survey can be an important indication of the extent of the work performed. Only the person named on the survey has a reasonable right to expect the survey to report the information that he or she requested.

A survey performed for a particular purpose can only be reasonably used for that purpose. It is foolish to assume that a survey performed at the request of someone else will contain all of the information you want to see. It is equally foolish to assume that a survey performed for one particular purpose will fulfill all of the requirements of an alternate need.

For example, a survey that was performed to confirm compliance with zoning restrictions may show the dimensions of a building that are not adequate for computation of the area of the floor space of that building. Zoning restrictions ignore recessed doorways; floor space computations do not.

11.16. CERTIFICATION

Special attention must be paid to the certification that appears on a survey plat. The phrase "certified correct" that was so popular in the past has been largely dropped by the land surveying profession. This phrase has been found to be much too general and subject to interpretations never dreamed of by Land Surveyors. "Certified correct" means not only that all of the information shown is correct (how accurate must a dimension be to be "correct"?) but also that *all* of the information that the reader desires is shown. This interpretation requires that the Land Surveyor know, at the time a survey is conducted, the needs of some unknown, future user of the survey plat. This is an impossible task to fulfill; consequently, Land Surveyors have begun to *explain*, in the certification statement, as well as in special notes, the extent of work performed and the limits of the warranty promised by that work.

11.17. LIMITING WORDS OR PHRASES

The most pronounced of the recent developments in the relationship between the Professional Land Surveyor and his or her client has been the increased use of words or phrases on survey plats that spell out what was and what was not performed by the Land Surveyor in the development of that plat. This clarification of the duties performed by the Land Surveyor has caused some protest from the users of survey plats who did not realize that the plats that they had been using all along actually had these limitations.

Land Surveyors are not title abstractors. Land Surveyors do not, in the normal course of business, search the deed record or chain of title for easements, restrictions, zoning, or other legal instruments that may affect a parcel of land. Those parties who had incorrectly believed that Land Surveyors were abstracting titles were quite upset when they were informed that the discovery and interpretation of recorded information rests with the legal profession.

The discovery of visible acts, on the ground, that encumber a property, whether recorded or not, is the responsibility of the Land Surveyor. The attorney evaluates the written or legal factors encompassed in the title to real property; the Land Surveyor evaluates the corporeal factors encompassed in the land itself. The combination of these two services constitutes a complete record of a real property parcel.

Statements on the survey that explain the extent of work performed by a Land Surveyor are there to ensure that no one mistakes a survey plat for an abstract of the deed record.

If it is important to the purpose of a particular plat that this information be shown, the Land Surveyor must be presented with such written records as the user desires to be shown. Limiting phrases can be removed if the information not normally searched for by the Land Surveyor is provided. These phrases are as much an integral part of the survey plat as are the bearings and distances, yet it is amazing how many persons, who would never contemplate asking a Land Surveyor to misrepresent a dimension, demand the removal of these important informational phrases.

11.18. AREA

The area of a parcel is based on a computation that involves several sets and combinations of measured values. For this reason, the area called for in a deed or reported in a survey plat holds the lowest ranking in the hierarchy of calls. Sales of property, however, often define the value of a parcel of land in terms of the area. Ironically, the one dimension considered to be the least reliable is often depended upon to set the most important aspect of any parcel of land—its monetary value.

Many Land Surveyors will omit the area of a parcel unless that information is specifically requested by the client. This omission is not critical in most cases of surveys of small, rectangular, urban, or suburban lots in which the computation of the area is obvious or the value of the property is not directly determined by the area. In cases of large or complicated parcels, however, the area is an important part of the information that is relevant to real property parcels.

11.19. SCALE

The scale shown on survey plats represents the relationship of the dimensions on the original plat to the dimensions of the survey. Most modern minimum standard detail requirements call for a statement of the scale and a bar scale on every survey plat.

Often, the survey plats that are handled by clients are *copies or reproductions* of the original drawing. Almost all reproduction processes will result in *distortions* of the plat. Paper will stretch or shrink in reaction to changes in humidity and many other factors. Such changes can have a dramatic impact on the fidelity of the scale on any plat. It is for these reasons that scaling dimensions on a plat are strictly reserved for *estimating* values. If a dimension is important, one should *never* depend on a scaled measurement of that dimension.

In the case of individual residential survey plats, the primary recipient of the plat is a homeowner who is usually unaware of the dangers of scaled values. For this reason, many Land Surveyors deliberately did not report the scale used to produce their plats. Land Surveyors are only too happy to compute or measure dimensions that are of interest to the user of a survey plat.

11.20. COMPARING THE SURVEY PLAT WITH THE DEED

There often are inconsistencies between deed information and survey data. These inconsistencies are, normally, the necessary result of the improved or updated information about a particular parcel of land. The Land Surveyor must make several judgmental decisions during the interpretation of boundary evidence discovered during a survey. It is very common for surveys of small, modern, urban home or business sites to agree exactly in every way with the information shown in the deed. This is because short distances are involved, rigid subdivision codes have been applied, and frequent street rights-of-way prevent the accumulation of small measurement errors into detectable proportions.

It is more common, particularly in cases of less-than-modern work, for the information contained in a deed to be slightly different from the information shown on a survey plat. Corrections to distances are quite common and may be the result of improved measuring techniques, corner interpretation, superior recovery methods, or simply new evidence, among other things.

Directions (bearings) are even more likely to vary from the deed information than are distances. Except in cases of state plane projections, the accumulation of inconsistencies found in adjacent tangent plane projection surveys or the even more pronounced variations of magnetic meridians cause the directions shown on many survey plats to be suitable for little else than computing the angles formed at boundary line intersections.

Variations between state plane directions reported on a deed and state plane directions recovered during a particular survey should reflect only accumulated measurement error. The amount of acceptable variation depends on the level of precision needed. Changes in the state plane projection datum may cause a change in the values associated with the state plane directions (bearings) of boundary lines. The survey plat should report the datum associated with the directions shown.

11.21. EXAMINING THE SURVEY PLAT FOR EASEMENTS

Easements are a nonpossessory interest held by one party in the real property of another whereby the first party is accorded partial use of said property for a specific purpose.[1] These rights that others may have can flow from one of two sources: the easement might be the result of a written agreement or the result of use. Easements may take one of two forms: they may be obvious (apparent), or they may be hidden.

Easements that are obvious, with physical indications of their existence on the ground, will be discovered by the Land Surveyor. Easements that are hidden may not be discovered by the Land Surveyor.

---

[1] American Congress on Surveying and Mapping and the American Society of Civil Engineers, *Definitions of Surveying and Associated Terms,* 2005 (rev.).

Hidden easements that are written and recorded will be discovered by *a title examination only*.

Land Surveyors do not normally perform complete title examinations unless they are requested to do so by the client. Most Land Surveyors are not in the title insurance business. If you want to have all of the easements or other statutory restrictions associated with a real property parcel shown on a survey plat, *you must inform the Land Surveyor of their existence*.

Title insurance companies are responsible for the majority of survey requests in the United States. Surveys performed for title insurance companies will not duplicate the title research work of the insurer. The Land Surveyor will report all *apparent* easements, some of which may not be discoverable by a title search, and the title insurance company will report all recorded easements, some of which may not be discoverable in the field.

Compare each survey plat with the title records for evidence of hidden, but recorded, easements. If a survey plat does not report an easement that is in the title record, it is wise to inform the Land Surveyor so that he or she may show this information on the survey plat. *Never rely on a survey plat as the absolute and sole source of land title information*. The survey plat is a field report of physical limits and conditions of a real property parcel. Title examinations and opinions are the professional preview of attorneys, not Land Surveyors.

11.22. EXAMINING THE SURVEY PLAT FOR ENCROACHMENTS

Encroachments are physical objects (obstructions) that invade upon the rights of another.[2] Encroachments are evaluated by the degree to which the free use and enjoyment of a real property parcel is affected. Few have any problem with recognizing and evaluating the impact of the encroachment of buildings, walkways, roadways, and other structures normally contained entirely within the boundaries of a designated parcel. The acquisitive act of a building extending across a property line is very clear; this is not so with fences.

11.22.1. Fences and Fence Lines

Fences and fence lines deserve particular discussion here because, unlike all of the other forms of structures mentioned, they are normally intended to be "on the boundary." Boundary land lines are not physical things and, as we have seen, are defined to have a length but not a width.

---

[2] Id.

Fences are very real things and do have width. It is quite impractical to expect a fence to be "on the boundary line" without some portion of the fence being across the boundary line. There is much debate about what is measured when the Land Surveyor is measuring a fence location.

Figure 11.3 is an example of what has been offered by some as a way of interpreting the term *fence line*. This interpretation defines the facing of the fence as that which constitutes the "fence line". It is the barbed wire, or the chain link, or the 1 × 6 facing, not the supporting fence posts, that obstruct movement across fence lines. Therefore, these obstructions must form the "fence line." This logical theory has a very fundamental flaw.

Let us assume that a landowner wants to construct a redwood fence on the boundary line between himself and a hostile neighbor. If the definition demonstrated in Figure 11.3 were to apply, then the owner would center the posts four inches from the boundary line (the thickness of the 1 × 6 facing plus the 2 × 4 stringers plus $^1/_2$ the diameter of the fence posts). Setting the posts and bolting on the stringers can be accomplished without crossing the boundary line. Nailing on the 1 × 6 facing, as well as regular maintenance, would require the owner of the fence to trespass onto the neighbor's land. Under this definition of "fence line" no one could construct or maintain a fence "on line" without the full and continual cooperation of the adjoiner.

However, if we were to define the fence line as being the centerline of the supporting fence posts, then half of all fence posts would be across the boundary line. There would be land between the boundary line and the facing of the fence that could be maintained only by the neighboring landowner. Complicating this situation is the fact that most fence posts are not perfectly vertical. Where is the fence line in these cases? Should it be at the ground? Should it be at the top of the fence? It is obvious that practicality demands toleration of a certain amount of encroachment, as well as trespass, in order for a boundary fence to exist.

It is for these reasons of practicality that the rigid definition of the location of a "fence line" demonstrated in Figure 11.3 is being replaced by one that defines a fence as being "on line" when any visible part of the fence is on line. Under this "liberal" definition of a fence line, fences may be constructed without continual trespass for maintenance and without the fear that a slight encroachment of facing or post could

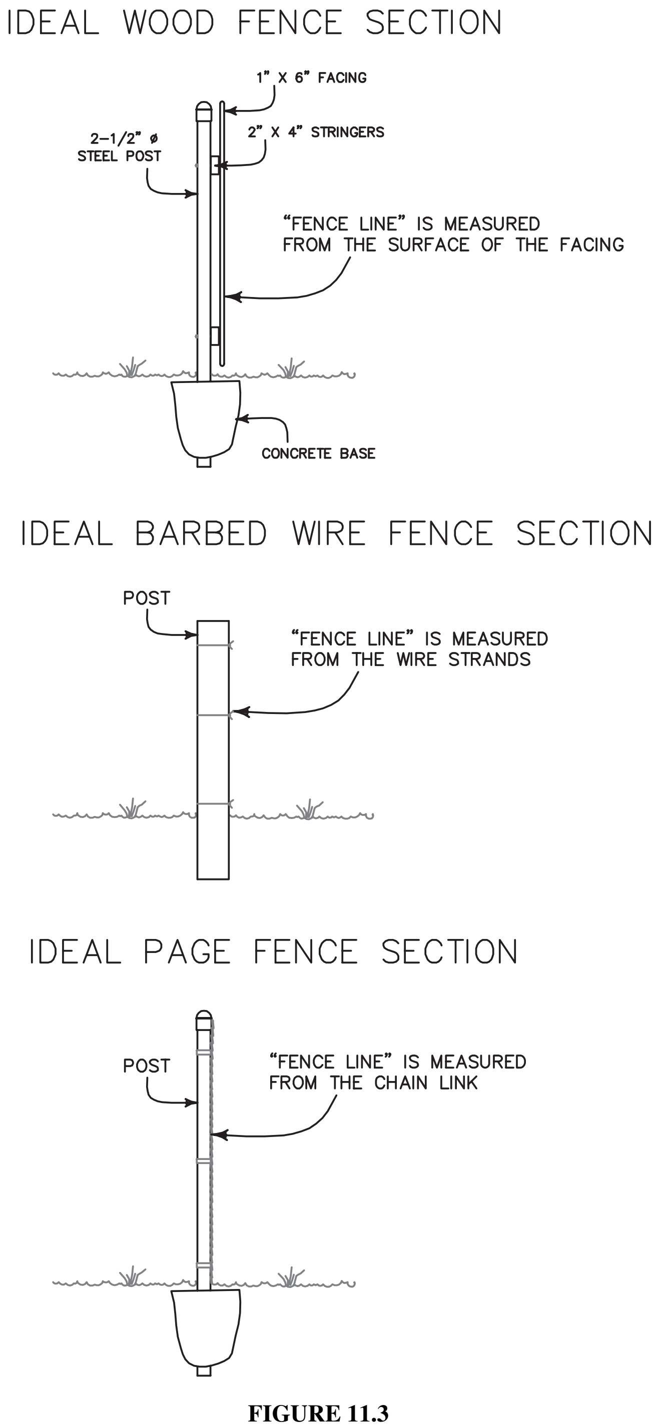

IDEAL WOOD FENCE SECTION

IDEAL BARBED WIRE FENCE SECTION

IDEAL PAGE FENCE SECTION

1" X 6" FACING
2–1/2" ø
STEEL POST
2" X 4" STRINGERS
"FENCE LINE" IS MEASURED FROM THE SURFACE OF THE FACING
CONCRETE BASE
POST
"FENCE LINE" IS MEASURED FROM THE WIRE STRANDS
POST
"FENCE LINE" IS MEASURED FROM THE CHAIN LINK

FIGURE 11.3

be interpreted as an acquisitive act. Each locality may differ in this interpretation, based on regional court decisions.

11.23. DETERMINING THE PRECISION REQUIRED

The ACSM developed a set of classes of surveys based on expected land use. Published as the *Minimum Standard Detail Requirements for ALTA/ACSM Land Title Surveys*, these general classifications have been mirrored, to varying degrees, by many state surveying organizations or regulatory boards. Early versions of this publication also listed several procedural and closure requirements for each class. These requirements are somewhat technical in nature, applicable only during the measurement and computational phases of a survey, and very difficult to assess without field notes or working papers.

Modern revisions of the ALTA/ACSM publication have dropped the “Class of Survey” designations, but most State and local regulators have not.

The ALTA and the National Society of Professional Surveyors (a member organization of the ACSM) has since significantly revised the initial publication and produced the *2005 Minimum Standard Detail Requirements for ALTA/ACSM Land Title Surveys*. The most recent version of the *Minimum Standards* is available on the web site www.acsm.net.

Among the revisions is the adoption of a single *Allowable Relative Positional Accuracy* standard. That standard is “0.07 feet (or 20 mm) +50 ppm.” While the deviations quoted are well within the capabilities of the modern Land Surveyor, the vague, inconsistent, and undefined terms contained in the *Allowable Relative Positional Accuracy* are significant, and the publication is certain to be amended further. The concept that “accuracy” rather than “precision” can be predetermined is contrary to reality. The revised publication is not included in this work.

State regulatory authorities may not have amended their rules and regulations to accommodate the changes. Therefore, many aspects of the former regulation need to be discussed even though the 2005 publication has dropped the concept of class of surveys. The list that follows presents the class of surveys, as previously defined by the ACSM publication, as well as this author’s interpretation of what the *effect* of the recommended procedure and computational requirements would be. The historical ACSM definition is set within quotation marks.

11.23.1. “Class A. Urban Surveys”

“Surveys of land lying within or adjoining a City or Town, this would also include the surveys of Commercial and Industrial Properties, Condominiums, Townhouses, Apartments, and/or other multi-unit developments, regardless of geographic location.”

Generally speaking, the actual relative positions of distinct features within a “Class A” survey may vary from the position published on the plat by a maximum distance obtained by the equation (Variation = 0.03 feet + D/15,000), where “D” is the distance between features. For example, if the distance between two boundary monuments is shown to be 1,347.56 feet on a survey plat, the acceptable variation would be about 0.12 feet (0.03 + 1,347.56/15,000). Therefore, the actual distance between monuments may be anywhere from 1,347.44 feet to 1,347.68 feet and still be within the tolerances of the “Class A” category.

Similarly, the bearings shown on the survey plat have an allowable variation of about 15 seconds (relative to the meridian or bearing base and depending on the length of the line). A published bearing of north 27 degrees 34 minutes 05 seconds east would be within tolerances if the actual bearing were anywhere between north 27 degrees 33 minutes 50 seconds east to north 27 degrees 34 minutes 20 seconds east. This, of course, does not apply when *different bearing bases are used*.

11.23.2. “Class B. Suburban Surveys”

“Surveys of lands lying outside urban areas, this land is used almost exclusively for single family residential use or residential subdivisions.”

This is, by far, the most common class of survey. The result of the application of the recommended tolerances for the procedures and computations of the “Class B” survey are positional accuracies or variations of about 0.05 + D/l0,000 feet and directional variations of about 20 seconds. The dimension of 1,347.56 feet given in the previous example would be considered within tolerances of a “Class B” survey if the actual value were anywhere between 1,347.38 feet and 1,347.74 feet. The bearing of north 27 degrees 34 minutes 05 seconds east would be considered within tolerances if the actual bearing (relative to the same base) were anywhere between north 27 degrees 33 minutes 45 seconds east and north 27 degrees 34 minutes 25 seconds east.

11.23.3. "Class C. Rural Surveys"

"Surveys of land such as farms and other undeveloped land outside the suburban areas which may have potential for future development" define Class C Surveys.

The tolerances under "Class C" are about (0.07 + D/7,500) for positions and about 30 seconds for directions. The acceptable range of values for our example dimension of 1,347.56 feet broadens to between 1,347.31 feet and 1,347.81 feet, and the directional range varies between north 27 degrees 33 minutes 35 seconds east and north 27 degrees 34 minutes 35 seconds east.

11.23.4. "Class D. Mountain and Marshland Surveys"

"Surveys of land which normally lie in remote areas with difficult terrain and usually have limited potential for development" constitute Class D Surveys.

The requirements for a "Class D" survey roughly translate into acceptable positional variations of 0.10 + D/5,000 feet and acceptable directional variations of 40 seconds. The acceptable limits of our examples would then be 1,347.19 feet and 1,347.93 feet for position, and north 27 degrees 33 minutes 25 seconds east and north 27 degrees 34 minutes 45 seconds east for directions.

These example variations are not presented as absolute limits but merely as guidelines intended to give the reader a "feel" for the range of reported dimensions that one should reasonably expect for each "class" of survey. Users should evaluate their own tolerance for deviations based on an assessment of the particulars of a given survey. Tolerances more restrictive than "Class A" are possible but prohibitively expensive. An old country surveyor's cliché best remembered when one is deciding on the tolerances required is: "We're measurin' a fence line, not buildin' a violin."

All of the positional tolerances ever presented assume that it is possible to *know* the *true* dimension of a given line. A challenge to a dimension presented on a survey plat will be reduced to comparing the *procedures used* by different Land Surveyors in arriving at a value.

11.24. DETERMINING NEEDS NOT COVERED IN A SURVEY PLAT

Every user of land survey data has his or her particular needs and expectations, as dictated by the purpose of the survey. The list of items in the ALTA/ACSM Minimum Standard Detail Requirements (Appendix A) represents the minimum requirements usually associated with the typical boundary survey. It is very possible that your particular needs will exceed the minimum requirements or be beyond the limits expressed by special notes or the certification.

Chief among the needs not commonly covered by the typical survey plat is the existence and location of all easements or servitudes affecting the parcel. Even in the absence of restrictive words or phrases, it is very wise to ensure the inclusion of recorded easements or servitudes by providing the Land Surveyor with an abstract of the title or, at the very least, by comparing the survey plat with the deed and title information yourself. The Land Surveyor will be delighted to add any information contained in the chain of title, even if it is not apparent on the ground.

11.25. CONTACTING THE SURVEYOR

It is vital that, before using the information found on any survey plat, the user of the information and the developer of that information have a clear and concise agreement on the limits of research and the purpose of the plat. This is not difficult if the survey plat was developed from a direct request by the user. Full communication, at the time the survey is requested, will provide this important link. If the user is not the same person as the one who is ordering the survey, full communication is not possible until the user contacts the Land Surveyor directly.

The user should simply pick up the telephone and call the Land Surveyor who developed the plat. The user should tell the Land Surveyor how the plat is to be used and should ask if the Land Surveyor anticipated that particular use. The user should question the Land Surveyor about *anything* that appears to be odd or unusual.

If the information is correct, the Land Surveyor will happily explain it. If the information cannot be easily explained or is incorrect, the Land Surveyor will gratefully revise the information and send an updated version of the plat or report. Land Surveyors want their work to be reviewed as carefully, as critically, and as often as possible *before* reliance is placed on the information presented. Often, the only critical review possible is in the Land Surveyor's own office. When external critical review is possible, the Land Surveyor welcomes it. The saddest words ever heard, usually after an expensive undertaking, are, "I thought that looked funny!"

Certifications or limiting words and phrases on survey plats that identify the plat as not providing all of the information that is desired do not render the plat useless. Often, the desired information can be added or the limiting phrase reworded if the necessary information is provided to the Land Surveyor. Plats that report that easement information is not shown, for example, may have that limiting phrase removed if the Land Surveyor is provided with all recorded easement information.

It is a great disservice to future users of a survey plat if any Land Surveyor were to remove the limiting words from a survey plat without also eliminating the need for the phrase.

If the Land Surveyor cannot be contacted, it is foolhardy to use any survey plat developed by that Land Surveyor for anything other than a general graphic aid for the parties involved. The Land Surveyors who are active in a particular area can provide insight on the abilities, typical procedure, and reasonable expectations of accuracy of deceased or retired Land Surveyors.

Often, the files of inactive Land Surveyors are sold or given to active local professionals for safekeeping and referral. The state or local professional surveyor's association can be contacted for assistance in locating such files.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 12

EXERCISES IN EVALUATING SURVEY PLATS

The following exercises are designed to clarify the recommended guidelines for evaluating survey plats. These guidelines are very general. Several states have developed, either through their state board of registration or through their professional societies, a minimum standard of practice. Such states may require information in addition to that shown as typical in these example plats.

None of these example plats is presented as an ideal example of a survey plat. All of the examples are drawn from actual cases but are greatly simplified, and appropriate name and location changes have been made. The strengths and weaknesses of each example will be discussed so that the reader may ponder them and apply the thoughts expressed to real survey plats. The numbers and letters in parentheses refer to the checklists mentioned in Chapter 11. Each of the exercises will take the form of a scenario. The reader is not expected to agree completely with each and every step in the scenario. Indeed, part of the learning experience is reflecting on and pondering how the reader may have proceeded differently.

12.1. THE CASE OF THE THREE PARTNERS

Ms. Cazes has the deed to lot 2 of section 11, township 3 north, range 5 west, Walnut County, Nebraska. The deed simply refers to the property as “lot 2, T3S, R5W, Walnut County Nebraska.” Along with the deed,

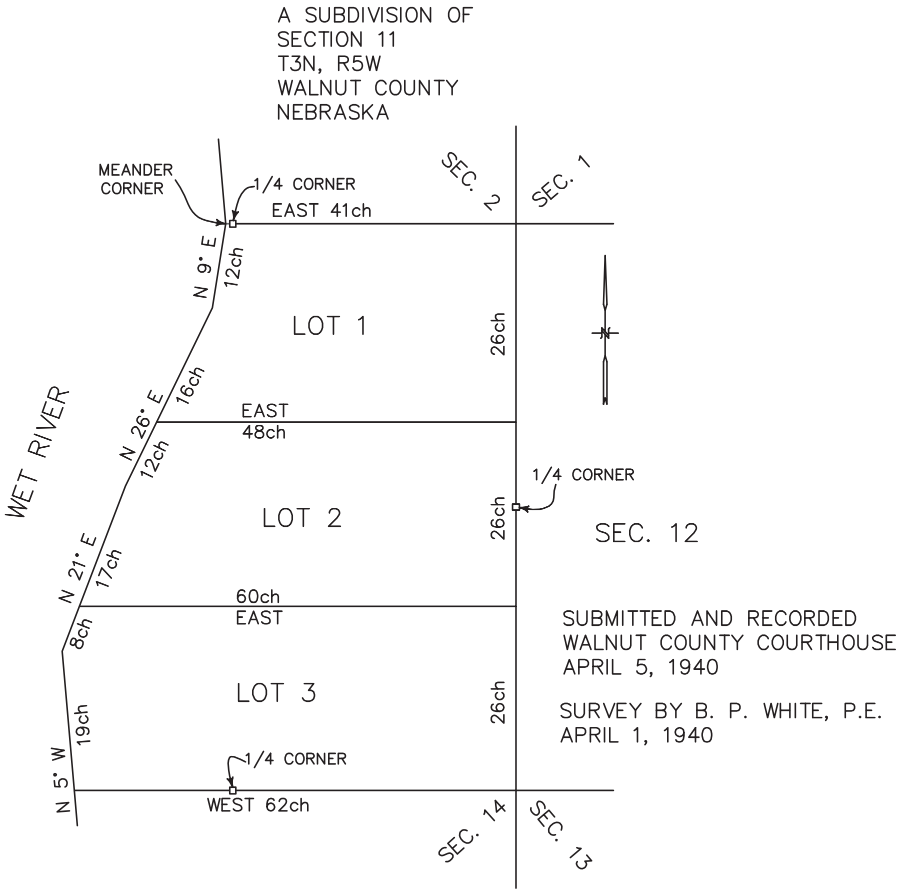
A SUBDIVISION OF
SECTION 11
T3N, R5W
WALNUT COUNTY
NEBRASKA
MEANDER CORNER
1/4 CORNER
EAST 41ch
SEC. 2
SEC. 1
N 9° E
12ch
LOT 1
26ch
N
WET RIVER
N 26° E
16ch
EAST
48ch
12ch
1/4 CORNER
LOT 2
26ch
SEC. 12
N 21° E
17ch
60ch
EAST
8ch
SUBMITTED AND RECORDED
WALNUT COUNTY COURTHOUSE
APRIL 5, 1940
LOT 3
26ch
SURVEY BY B. P. WHITE, P.E.
APRIL 1, 1940
N 5° W
19ch
1/4 CORNER
WEST 62ch
SEC. 14
SEC. 13

FIGURE 12.1

which she inherited from her father, is the map shown in Figure 12.1. Ms. Cazes's father was one of the original three partners who together owned section 11. The record indicates that the three caused the property to be subdivided and that each received one of the lots.

Ms. Cazes now wishes to sell lot 2. Evaluate the original plat and assess its usefulness in such a sale.

1. The presence of a section, township, or range might cause the reader to think that the controlling land record system here is the USPLS. This is only partly true. If lots 1, 2, and 3 were not formed by the government surveyor or by the application of the USPLS rules, then the lots were formed by a platted subdivision. The exterior boundaries of the subdivision, being the boundaries of section 11, fall under the recovery rules of the USPLS. The

lots do not. The land record system for the individual lots is the platted subdivision system.

2. The use of the cardinal directions, as well as the fact that all directions are reported to the nearest degree, suggest that the bearing of each line was independently observed. This is a projectionless map. The plat is simply a drawing of the directions and distances observed for each line.
3. The date indicates that the work was done in 1940. All distances are rounded off to the nearest chain (66 feet) in all directions, to the nearest degree. This indicates a crudeness of measurement that is more typical of 1840 than 1940.
4. The purpose of the survey is obviously to document a division of the section into three lots. The dimensions at the east boundary indicate a desire to divide the lots by near equal division of this side.
5. a. The north arrow is shown with no reference to a bearing base. The crudeness of the values suggests that a magnetic compass was used, and it is very unlikely that a resurvey of any of the lots will duplicate the directions reported on the subdivision plat.
   b. The legal description of any of the lots is "lot——of the subdivision of section 11, T3N; R5W, Walnut County, Nebraska, recorded April 5, 1940" and is shown on the survey plat. The absence of a reference to a principal meridian will not be of any consequence if only one principal meridian serves all of Walnut County.
   c. The date of the survey is shown, along with the recordation date. This must be verified by examining the records.
   d. The name of the "Land Surveyor" is shown, but not the address.
   e. The survey plat is not signed or sealed, but the age of the document indicates that it probably followed the accepted procedure for a subdivision of land at that time and, if recorded, will probably defeat any attempt to refute its authenticity.
   f. All adjoiners are shown by section numbers or as the Wet River. If the Wet River were navigable at the time of the original government survey, this would be a riparian boundary.
   g. All dimensions are shown and are unusually even. The dimension of 80 chains for the east boundary of the subdivision is particularly suspicious.

   h. The bearings of all sides are shown, and they, too, are suspiciously close to the directions typically called for in the government surveys.
   i. The name of the client is absent, but the purpose of the survey is obviously to document the creation of three lots.
   j. There is no certification statement.
   k. There are no limiting words or phrases. The drawing is one of such little detail that one could reasonably expect that buildings, fences, driveways, and other features exist but are not shown.
   l. The area is not shown.
   m. The scale of the drawing is not shown.
6. A comparison with the deed is consistent with the plat. The deed does not contain a metes and bounds supplementary description. The original government township plat of T3N, R5W, dated 1866, reports the exact same bearings and distances for the exterior of section 11 as does the subdivision plat.
7. There are no easements shown, but, considering the age of the plat, this is not surprising.
8. No improvements are shown at all. This may be an indication that the improvements were present but deemed unimportant or that the land may actually have been vacant.
9. The accuracy standards, as indicated by the dimensions shown, are so crude as to fall below any of the modern “classes” of surveys.
10. There is a wealth of information that is not addressed that would be of importance in a sale of this property. The deed record must be reviewed (title examination) for recorded easements, rights-of-way, possible partitions, and so on. Buildings, bounding fence lines, corner monumentation, and other physical features need to be examined. Clearly, this parcel must be resurveyed, and an abstract of the title record must be made prior to any sale. As with any purchase of real property, a title insurance policy is a must.
11. After attempts have been made to contact B. P. White, it is discovered that he is deceased.

12.1.1. Conclusion

The fact that the bearings and distances reported for section 11 by the government surveyor and those reported by B. P. White for the exterior of the subdivision of section 11 are exactly the same is conclusive proof that White did not resurvey or recover the boundaries of section 11 prior to the drafting of the subdivision plat. It is highly probable that White simply drafted the subdivision from the government plat and never even visited the property. This is a "paper survey" and serves no other purpose than to indicate the intended division of the section. Lot 2 must be resurveyed by recovering as many of the original section, quarter, and meander corners as possible.

A survey of lot 2, section 11, T3N, R5W, Walnut County, Nebraska, was ordered, and Figure 12.2 is a copy of the resulting survey plat.

By applying the same examination procedure as before, we can discover the following information:

1. The external boundaries of Section 11 are reported by J. Ford as being recovered. The distance between the NE corner of section 11 and the east $^1/_4$ corner is only 6.7 feet less than the original government survey's 40 chains (2,640 feet). The distance between the SE corner of section 11 and the east $^1/_4$ corner is 41.39 feet less than 40 chains. The line between the NE and SE corners of section 11 is not a straight line. All of these signs indicate a recovery of the east line of section 11 based on a proper recovery of the corners.
2. The presence of an "x" and "y" at the NE corner of section 11, as well as the note "Grid Bearings NAD 27" indicate that the state plane projection system for Nebraska is being used. Ford's references could be elaborated in order to define better the projection system used. Nebraska law may require special wording or notes when state plane coordinates (SPCs) are being used.
3. The survey plat indicates directions and distances that are defined to a second of an arch and to a hundredth of a foot.
4. Because we requested the survey, the purpose is known. When the survey was requested, a full explanation of the purpose and expected use of the survey was made to Ford.
5. a. The north arrow and bearing base are shown.
    b. The legal description "lot 2 of section 11, T3N, R5W, 6th P.M. Walnut County, Nebraska" is shown on the survey plat.
    c. The plat is dated "January 18, 1987."
    d. The name and address of the Land Surveyor is shown.
    e. The plat is signed and sealed.

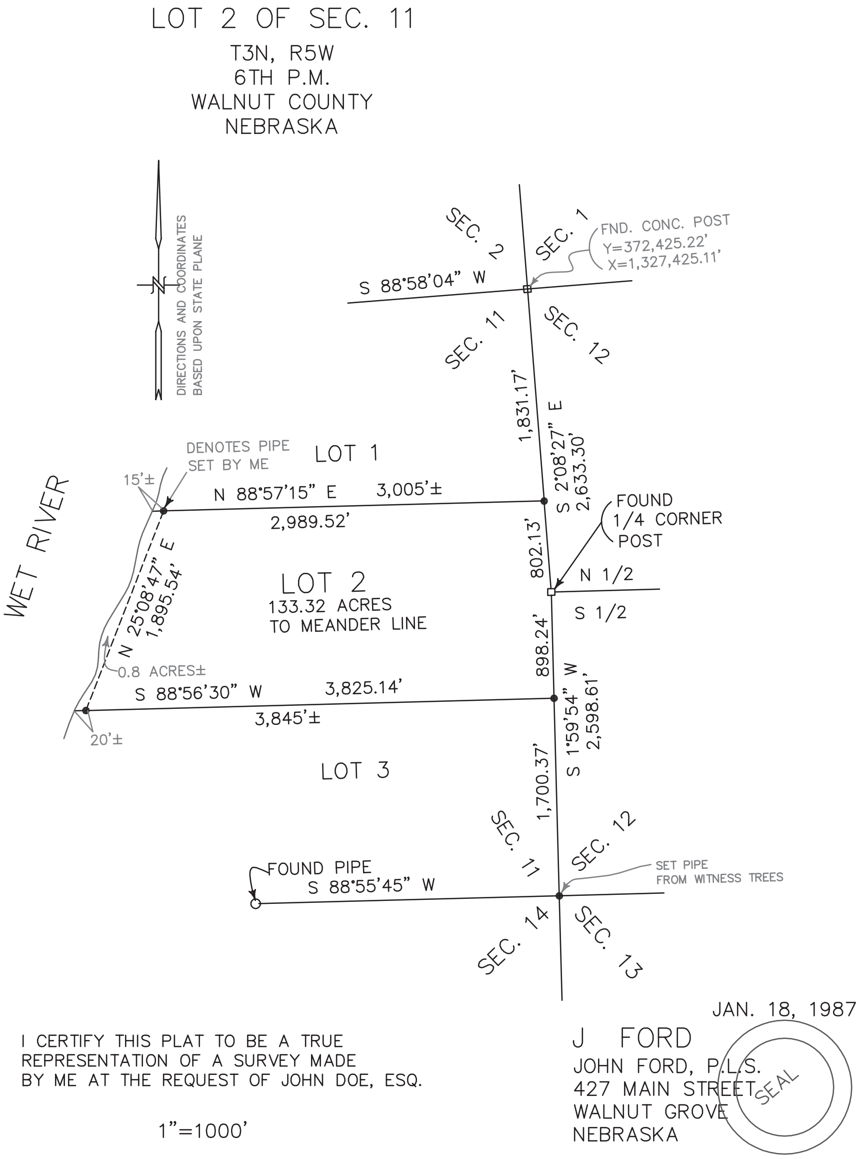
LOT 2 OF SEC. 11
T3N, R5W
6TH P.M.
WALNUT COUNTY
NEBRASKA
DIRECTIONS AND COORDINATES BASED UPON STATE PLANE
N
SEC. 2
SEC. 1
FND. CONC. POST
Y=372,425.22'
X=1,327,425.11'
S 88°58'04" W
SEC. 11
SEC. 12
1,831.17'
S 2°08'27" E
2,633.30'
DENOTES PIPE SET BY ME
LOT 1
15'±
N 88°57'15" E
3,005'±
2,989.52'
WET RIVER
FOUND 1/4 CORNER POST
802.13'
N 1/2
S 1/2
N 25°08'47" E
1,895.54'
LOT 2
133.32 ACRES
TO MEANDER LINE
0.8 ACRES±
898.24'
S 88°56'30" W
3,825.14'
3,845'±
20'±
S 1°59'54" W
2,598.61'
LOT 3
1,700.37'
SEC. 11
SEC. 12
FOUND PIPE
S 88°55'45" W
SET PIPE FROM WITNESS TREES
SEC. 14
SEC. 13
JAN. 18, 1987
I CERTIFY THIS PLAT TO BE A TRUE REPRESENTATION OF A SURVEY MADE BY ME AT THE REQUEST OF JOHN DOE, ESQ.
J FORD
JOHN FORD, P.L.S.
427 MAIN STREET
WALNUT GROVE
NEBRASKA
SEAL
1"=1000'

FIGURE 12.2

f. The adjoining properties are shown by lot number, natural feature, or by USPLS designation as applicable.

g. All dimensions of all sides are shown. The north line of lot 2 shows a total length of 3,005 feet, *plus or minus*, meaning "more or less." This is appropriate because of the indefinite nature of riverbanks. This distance is shown on the subdivision plat as 48 chains (3,168 feet). The distance between the NE corner of lot 2 and the NW meander corner of lot 2 is shown as 2,989.52 feet. The distance from the meander corner to the Wet River is shown as 15 feet, *plus or minus*. The sum of these two distances is 3,004.52 feet. This is consistent with the shown total of 3,005 feet, plus or minus. It would be scientifically incorrect to express the distance as "3,004.52 feet, plus or minus," because this would imply that the magnitude of the uncertainty was in fractions of a foot, whereas the plat clearly expresses the magnitude of uncertainty in feet (15 feet, plus or minus). Similarly, 20 feet, plus or minus, added to 3,825.14 feet is consistent with the total 3,845 feet, plus or minus. The subdivision plat reports this second distance as 62 chains (4,092 feet).

h. The bearings of all lines, including the new meander line (north 25 degrees 08 minutes 47 seconds east), are shown. The fact that the bearings are expressed to the nearest second suggests that a high degree of precision may be accorded to these values.
i. The name of the client is shown. The purpose of the survey is not stated.
j. The certification states that the information shown on the survey plat accurately reflects the results of the survey performed by Ford. It does not claim that the survey was performed to any particular standard, nor does it warrant any of the information as "correct."
k. There are no limiting words or phrases on the face of the plat.
l. The area is shown to be 133.32 acres to the meander line. This means that the strip of land between the meander line and the Wet River is not included in the acreage figure shown. This meander line is a new line established by Ford and is not to be confused with a boundary line or the original meander line by the government surveyor.
m. The scale is shown as "1" = 1,000'." This may be checked by measuring a few dimensions with a ruler. Paper may stretch in one direction more than another, so any check should include east–west as well as north–south distances.

6. The deed, referring only to a lot number, is consistent with the survey plat. The original subdivision plat is not entirely

consistent with the survey plat. In most cases, the deed will refer to a subdivision plat, making that plat a part of the deed. The failure of the deed to refer to the subdivision plat, in this case, is not of any consequence. The discrepancies between the subdivision plat and the survey plat of lot 2 should be explained by the Land Surveyor.

7. There are no easements shown. There is no note by the Land Surveyor indicating whether a search for recorded easements was made by his office. This needs clarification.
8. There are no fence lines, buildings, or other structures shown, and there is no note indicating that the property is vacant. This needs to be clarified.
9. The presence of SPCs, distances specified to the one-hundredth of a foot, and bearings specified to the second indicate a high order of precision. This needs to be verified.
10. A review of the title abstract does not indicate any recorded easements, servitudes, rights-of-way, or any other encumbrances to the property. The absence of fences should be explained, as well as the lack of buildings. Perhaps a statement to the effect that these things do not exist is appropriate.
11. The Land Surveyor can now be contacted to respond to the list of questions that have been raised. From this conversation, we learn that the section and quarter corners all have a long history and are well recognized, through perpetuation or recovery of accessories, as the original corners. The work was performed and controlled under the state plane projection system for Nebraska (south zone), North America Datum of 1927. Some of the data was developed over several years of fieldwork in the adjoining sections, although the monumentation of the north and south boundary lines of lot 2 and the Wet River was performed in response to the survey request. All work conformed to the requirements of a “Class A” survey.

Ford has affirmed that the survey plat was developed with the intent that it be used in a transfer of title. Ford also has reported that he believes that the original subdivision plat was an “office” survey that simply divided, on paper, the government version of the dimensions of section 11. His dimensions and directions for lot 2 are his professional opinion of the actual boundaries of lot 2, based on his interpretation of the intent of the subdivider and the constraints of the true configuration of section 11. It is not known if the Wet River shifted east or if the original subdivision plat incorrectly reported the river's location. Ford agreed to add to the plat, in parentheses, old dimensions and directions shown on the subdivision plat to clarify that the existence of the discrepancies is known.

Ford reported that he did not perform a title search for recorded easements or other title restrictions. The property is vacant, wooded lowlands, and subject to inundation. The construction of a reservoir, upstream on the Wet River, is under way and may change the condition of the land.

We have had all our questions about the land answered. A modern survey plat and close communication with the Land Surveyor who developed the plat have made it possible to sell the parcel with a better understanding of just what was being exchanged.

There are some who may balk at accepting a survey of lot 2 that is, apparently, at such variance with the subdivision plat. Those so concerned would do well to reflect on the fact that, if the original subdivision plat version of lot 2 were simply copied, there would be an overlap of 163–247 feet on the east boundary and over 16 feet on both the north and south boundaries. The buyer would be paying for 10.5 acres of land that does not exist or is, at best, under contention. A modern survey of a parcel of land should be ordered if one really wants to know the truth.

Land Surveyors are like doctors or lawyers in this respect: if the surveyor's professional opinion and explanation of that opinion are not satisfactory, then a second opinion should be sought. The very worst thing that one can do is to attempt to persuade the Land Surveyor to distort the survey plat to match the deed. Land cannot be created by words on paper. A Land Surveyor is consulted because of his or her special knowledge of measurements and boundary recovery. Don't reject the Land Surveyor's opinion just because it "rocks the boat," but don't blindly accept that which cannot be explained to your satisfaction either.

12.2. THE LAND GRABBER

Let us assume that you intend to purchase a parcel of land on New York State highway number 10, in Dutch County, near the town of Dutchville, New York. The owner bought the property in 1978, at which time it was surveyed by the owner's cousin (now deceased), who was a Land Surveyor from Albany. The papers presented to you by the owner describe the property as follows:

> . . . That portion or parcel of land bounded on the West by the land of Robert Beckmann, on the North by the lands of Gilbert Hirt, on the East by the lands of John Peters as defined by the old Cromwell Grant line, and on the South by New York State Highway #10 and containing approximately 5 and one half acres commencing and point of beginning at the old stone borne located on the Cromwell Grant line at the North side of New York State Highway Number 10; thence, S 25 E 50′ to the centerline of Highway 10; thence, S 65 W 350′; thence, N 25 W 50′; thence N 25 W 700′; thence, N 65 E 350′; thence, S 25 E 700′ to the point of beginning. . .

This is a poor description, at best, but it is all that the owner has. The local Land Surveyor is Mr. J. D. Times, who is highly recommended by the owner as being "the only man for the job."

In your initial conversation with Times, you advise him that you intend to purchase the property for investment purposes and that you are in the process of obtaining title insurance. Times says that if you send him copies of "all title papers and previous surveys" in your possession, as well as a copy of the results of the title examination by your title insurance company, he will perform the survey.

The title insurance company representative had previously stated that the survey could be done and the plat forwarded to them while the title examination was under way, and you mention this to Times. The sale has a strict time limit, so you urge Times to start immediately. Times agreed to start as soon as a copy of the deed is delivered.

The survey is completed, and several copies of the survey plat are delivered to you. Retaining one copy of the survey plat, you forward the rest to the title insurance company. Figure 12.3 shows the survey plat received by you.

While the title insurance company is completing its work, you decide to evaluate the survey plat, using the checklist that appeared in Chapter 11.

1. The land record system is a metes and bounds system.
2. There is no clear indication of whether a map projection has been used (meaning angles were measured) or whether the plat is simply the result of observed bearings.
3. The survey appears to be the result of recent work. The bearings seem to be imprecise because the minutes and seconds of direction are not shown. This may not be vitally important,

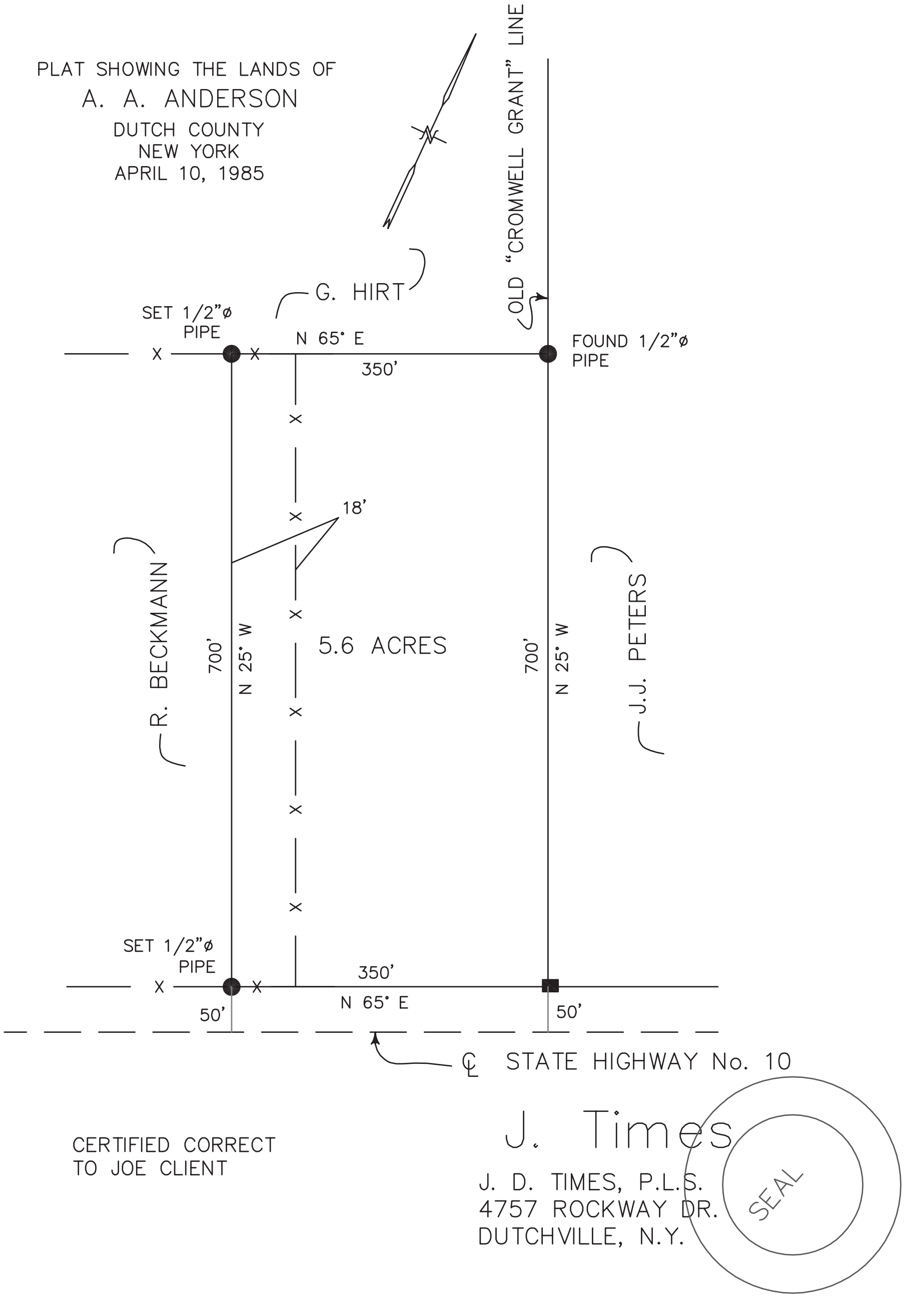
PLAT SHOWING THE LANDS OF
A. A. ANDERSON
DUTCH COUNTY
NEW YORK
APRIL 10, 1985
OLD "CROMWELL GRANT" LINE
G. HIRT
SET 1/2"ø PIPE
N 65° E
350'
FOUND 1/2"ø PIPE
18'
R. BECKMANN
700'
N 25° W
5.6 ACRES
700'
N 25° W
J.J. PETERS
SET 1/2"ø PIPE
350'
N 65° E
50'
50'
℄ STATE HIGHWAY No. 10
J. Times
CERTIFIED CORRECT
TO JOE CLIENT
J. D. TIMES, P.L.S.
4757 ROCKWAY DR.
DUTCHVILLE, N.Y.
SEAL

FIGURE 12.3

considering the small size of the parcel. The distances are all reported to the nearest foot. This may be the result of imprecise work, or it may simply be that Times did not show the tenths and hundredths of a foot if those figures were zero.

4. The purpose of the survey was discussed in depth between you and Times.
5. a. A north arrow is shown, but the bearing base is not defined.
   b. The caption of the plat describes the parcel as "The Lands of A. A. Anderson, Dutch County, New York." This may not qualify as a "legal description," depending on the cadastral records of Dutch County.
   c. The date of the survey is shown, and the work appears to be modern, except for the lack of precision reported in the bearings and distances.
   d. The name and address of the Land Surveyor are shown.
   e. The signature and seal of the Land Surveyor are shown.
   f. All adjoiners are identified.
   g. All dimensions are shown and appear to be consistent.
   h. All bearings are shown. The bearing base or method of determining the bearings should be clarified.
   i. The name of the client is shown.
   j. The plat is "certified correct" to you. This particular wording of the certification is extremely broad and is becoming very unpopular among Land Surveyors. Your understanding of just what was done by Times would be improved by a more precise certification.
   k. There are no limiting words or phrases, in spite of the fact that there is an apparent encroachment of some 18 feet along the western boundary line.
   l. The area is shown as 5.6 acres, but it is not explained whether that includes the encroachment or excludes it.
   m. There is no scale indicated.
6. The directions and distances are exactly the same as the deed description furnished to Times. As in the old "B" westerns when things were "too quiet," this might be a case of things being "too good," but, considering the small size of the parcel, it is quite probable that a match of distances exists within the acceptable range of precision.

7. There are no indications of any easements, nor is there any note about the extent to which easements were searched for. The conversation with Times when the survey was ordered clearly indicated that Times would not perform a title examination.
8. There is an extensive fence encroachment along the entire west boundary line with R. Beckmann. This is a serious encroachment that encumbers $^1/_3$ of an acre.
9. The property is quite valuable and subject to commercial development. It is not possible to tell from the survey plat what the accuracy standards were for this survey.
10. The title records indicate a 10-foot easement "across the entire rear or northerly line" of the parcel. This information should be added to the plat. Clarification is required on the encroachment. Times needs to investigate, if he hasn't already done so, the titles to the Beckmann and the Hirt properties.
11. You contact Times with a list of questions resulting from the examination of the survey plat and receive the following clarifications. The survey was performed by "side shots" from a random traverse around the parcel. The bearings shown are based upon the reported direction of the "Cromwell Grant" line, as monumented by the stone borne and $^1/_2$-inch-diameter iron pipe (therefore, the plat is a tangent plane projection). Times estimates that the distances shown are accurate to plus or minus $^1/_2$ inch and that the direction shown is accurate to plus or minus 20 seconds. (This is consistent with a "Class B" survey.)

Times reported that there are no visible indications of any easements or other encumbrances to the property except the apparent encroachment of a fence line, which is shown. The fence is not very old, according to Times, and it appears that Beckmann has fenced in more than he has "a right to." Times recommends that the boundary with Beckmann be settled before any purchase is made. The title insurance company reports that they will insure title to that portion of the property that is not in conflict (332 feet wide). Times also agreed to add the easement found during the title search to the plat.

The Land Surveyor cannot, and should not, pass judgment on whether or not Beckmann has "a right" to the strip of land between the fence line and the deed line. Beckmann's deed may not reflect a written claim to the 18-foot strip, but there may have been a transfer of rights based on acquisitive prescription.

Land Surveyors do not establish boundaries. Times had performed all the work that he is qualified to perform. He has uncovered and documented evidence of a conflict between the lines of possession and the written deed, thereby providing you with all of the pertinent facts about the parcel that you need in order to make a decision. Times cannot warrant that Beckmann's possession can be ended. The boundary is in doubt, and only an agreement between owners or a court decision can settle it.

You may purchase the property and attempt to settle the question about the boundary. You may require the owner to settle the question about the boundary before any purchase is made. You may decline to purchase the property. These options should be considered in consultation with *your* attorney, not the title insurance company's attorney. Times may appear on your behalf, testify to the facts discovered by him, offer an opinion about the boundary, and many other things, but he cannot establish the boundary contrary to Beckmann's possession or consent. The survey plat can only serve as documentation of the extent of the conflict.

12.3. EASEMENT SURPRISE

Let us assume that you are a real estate investor contemplating the purchase of a large building in an industrial area. The present owner purchased the property in 1965, and a copy of the survey performed at that time is shown in Figure 12.4. The owner states that he has not made any improvements to the property since the purchase.

Referring back to the checklist, you determine the following:

1. The parcel is a part of a platted subdivision.
2. The note that all corners are 90 degrees indicates that the plat is a tangent plane projection.
3. The survey is dated and appears to conform to the standards of practice during the 1960s.
4. The survey was for the purpose of assisting in the documentation of a sale.
5. a. North is shown, but no bearings are shown.
   b. The description shown is based on the subdivision plat and is sufficient to identify the plat.
   c. The date of the plat indicates that it is over 40 years old.

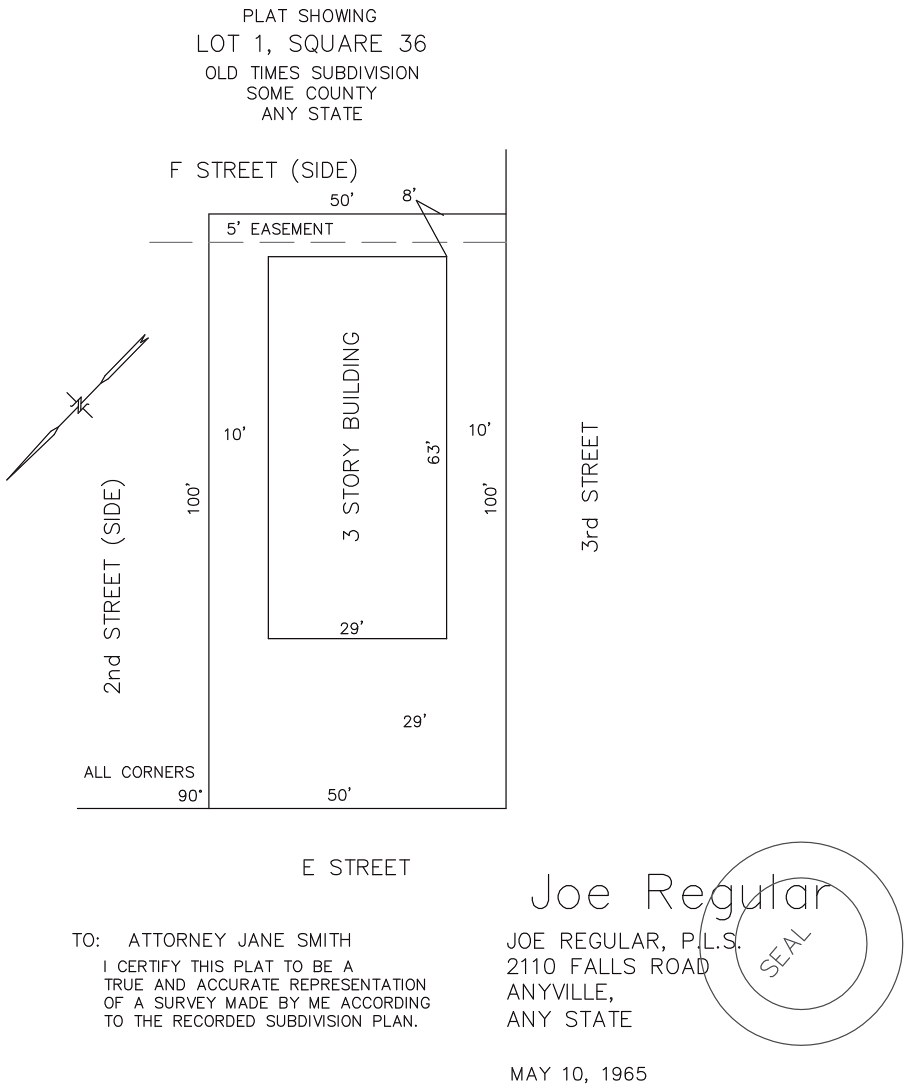
PLAT SHOWING
LOT 1, SQUARE 36
OLD TIMES SUBDIVISION
SOME COUNTY
ANY STATE
F STREET (SIDE)
50'
8'
5' EASEMENT
3 STORY BUILDING
10'
63'
10'
100'
100'
2nd STREET (SIDE)
3rd STREET
29'
29'
ALL CORNERS
90°
50'
E STREET
Joe Regular
TO: ATTORNEY JANE SMITH
I CERTIFY THIS PLAT TO BE A
TRUE AND ACCURATE REPRESENTATION
OF A SURVEY MADE BY ME ACCORDING
TO THE RECORDED SUBDIVISION PLAN.
JOE REGULAR, P.L.S.
2110 FALLS ROAD
ANYVILLE,
ANY STATE
SEAL
MAY 10, 1965

FIGURE 12.4

d. The name and address of the Land Surveyor are shown.

e. The signature and seal of the Land Surveyor are shown.

f. The streets bounding the square are shown. This is a common and excepted means of defining the boundaries of a lot in a platted subdivision in many communities.

g. The dimensions of all sides are shown. The fact that the tenths and hundredths of a foot are not shown may be the result of an attempt at brevity, as far as the lot boundary lines are concerned. The lack of fractions in building offset lines and dimensions is more likely to be the result of imprecise work.

   h. The angles at the corners are reported to be 90 degrees.
   i. The name of the client is shown.
   j. The certification is present and warrants that the drawing faithfully reflects what J. Regular discovered during his survey.
   k. The certification contains the limiting words "According to the Recorded Subdivision Plan."
   l. The area is not shown.
   m. The scale of the plat is not shown.

6. The deed only refers to the parcel by naming the lot, square, subdivision, county, and state. The lack of a metes and bounds supplementary description removes any possible check for blunders in listing the identification of the lot.
7. The plat shows a five-foot easement across the east side of the parcel.
8. There are no apparent encroachments.
9. The implied precision of the plat requires clarification.
10. The limiting phrase by Regular is notice that the *only* record that Regular relied on for data concerning lot 1 was the recorded subdivision plan and the discoveries, measurements, and observations made by him during the survey.
11. When he was contacted to clarify the questions raised by the checklist, Regular confirmed that he did not perform any search of the deed record for anything other than information on the locations of the exterior boundaries of the lot. The subdivision plat indicated the easement line shown without reference to the party enjoying the rights of that easement. Regular reported all distances and angles to be consistent with the accuracy requirements of a "Class B" survey. The building is constructed of a rough, shingled exterior and is in disrepair. Regular recommended that you obtain title insurance, which you decline to do (it costs too much).

Unknown to all in this scenario was that a 10-foot easement had been acquired by a major gas service company along the entire east boundary of the lot. The easement was acquired only two years after the subdivision was created, one year before the construction of the building. The hidden 2-foot encroachment is a ticking time bomb that will hinder the marketability of this property as soon as it is discovered.

12.4. EXCESSIVE PROBLEMS

John Mason, a local developer and home builder, decided to build a "spec" house on lot 1, block C of East Kingston Heights subdivision, local county, our state. Mason is the owner and developer of both Kingston Heights subdivision and East Kingston Heights subdivision. Work progressed rapidly in both subdivisions. As part of the construction loan to Mason, the lender, Easy Money Inc., required that an ALTA/ACSM land title survey be performed at the end of construction. The work was completed on the house in July of 1986, and Figure 12.5 shows the survey plat requested and used by Easy Money Inc., along with a complete title examination by Nit Pic Title Inc.

An application of the evaluation checklist to this plat reveals the following:

1. The land record system governing this parcel is the platted subdivision system.
2. The mapping system is probably some form of a conformal projection. It may be a tangent plane or state plane projection, but, for a parcel this small, it is not very important to distinguish which one is in use.
3. The survey is fairly modern, both chronologically and procedurally.
4. This survey was done for the purpose of fulfilling the boundary survey requirements of an ALTA/ACSM land title survey.
5. a. The north arrow is shown without a direct reference to a bearing base.
    b. The legal description is shown on the plat as "lot 1, block C, East Kingston Heights subdivision, as recorded in book 324 folio 15, local county, our state." This description is consistent with the land record system in force.
    c. The date of the survey is shown.
    d. The name and address of the Land Surveyor are shown.

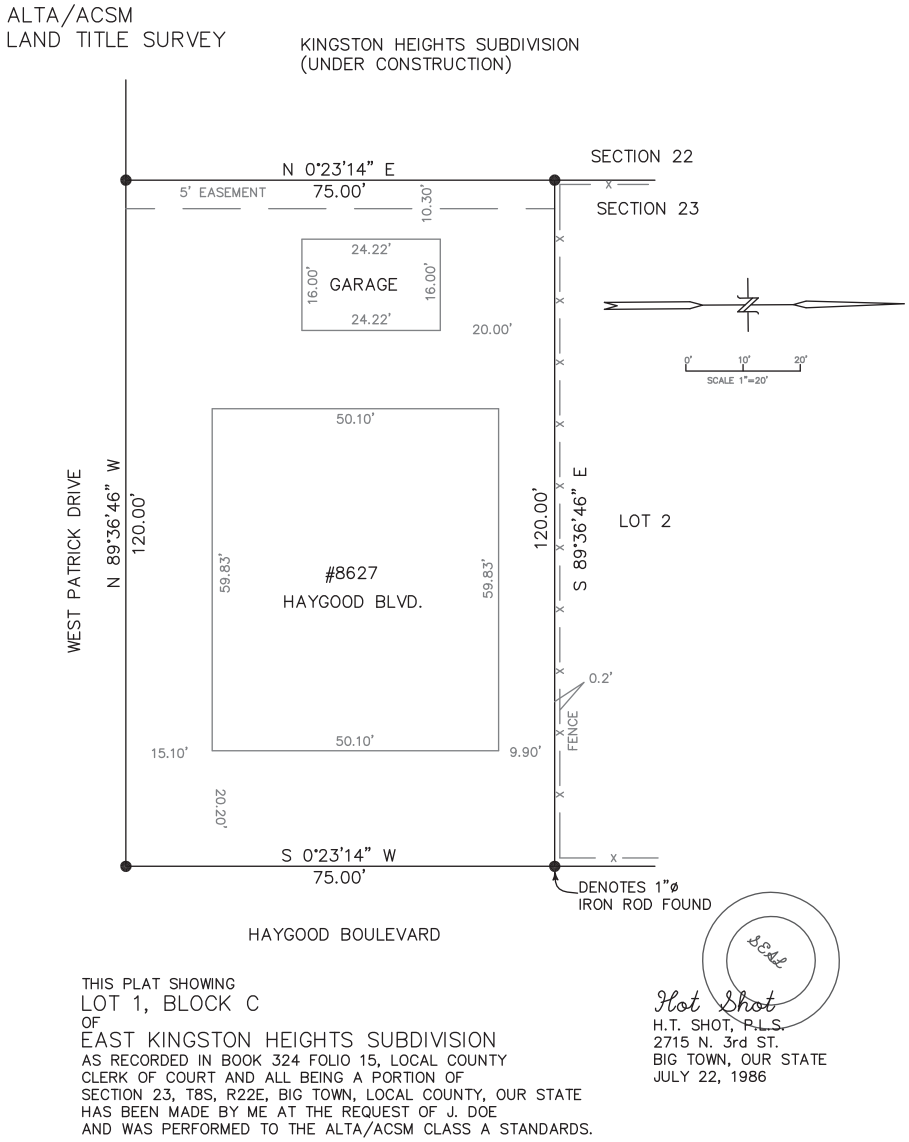
ALTA/ACSM
LAND TITLE SURVEY
KINGSTON HEIGHTS SUBDIVISION
(UNDER CONSTRUCTION)
SECTION 22
SECTION 23
N 0°23'14" E
75.00'
5' EASEMENT
10.30'
24.22'
GARAGE
16.00'
16.00'
24.22'
20.00'
0'
10'
20'
SCALE 1"=20'
50.10'
WEST PATRICK DRIVE
N 89°36'46" W
120.00'
59.83'
#8627
HAYGOOD BLVD.
59.83'
120.00'
S 89°36'46" E
LOT 2
0.2'
FENCE
50.10'
15.10'
9.90'
20.20'
S 0°23'14" W
75.00'
DENOTES 1"ø
IRON ROD FOUND
HAYGOOD BOULEVARD
SEAL
THIS PLAT SHOWING
LOT 1, BLOCK C
OF
EAST KINGSTON HEIGHTS SUBDIVISION
AS RECORDED IN BOOK 324 FOLIO 15, LOCAL COUNTY
CLERK OF COURT AND ALL BEING A PORTION OF
SECTION 23, T8S, R22E, BIG TOWN, LOCAL COUNTY, OUR STATE
HAS BEEN MADE BY ME AT THE REQUEST OF J. DOE
AND WAS PERFORMED TO THE ALTA/ACSM CLASS A STANDARDS.
Hot Shot
H.T. SHOT, P.L.S.
2715 N. 3rd ST.
BIG TOWN, OUR STATE
JULY 22, 1986

FIGURE 12.5

e. The signature and seal of the Land Surveyor are shown.
f. The adjoining properties are shown as street rights-of-way, lot number, or section number.
g. The dimensions of all sides are shown.
h. The bearings of all sides are shown.

i. The name of the client and the purpose of the survey are stated.
j. The certification is not worded in exact accordance with the requirements of an ALTA/ACSM land title survey.
k. The reworded certification contains the implication of limiting the records search by H. T. Shot to the recorded subdivision plat.
l. The area is not shown.
m. The scale is noted to be “1” = 20’” and is accompanied by a bar scale.

6. The subdivision is new, and a prior deed for this particular lot did not exist. The survey plat of lot 1 block C is consistent with the subdivision plat.
7. There is an easement across the entire western end of the lot. The title examination did not reveal any other easements or other noteworthy items of record.
8. There are no encroachments. The fence along the northern boundary of the lot is not considered by Nit Pic Title Inc. to be an encumbrance.
9. The survey standard of “Class A,” as defined by the “Minimum Standard Detail Requirements for ALTA/ACSM Land Title Surveys,” is of the highest order of accuracy for boundary surveys.
10. The certification should be changed to meet the ALTA/ACSM standard certification.
11. This step was not taken by Easy Money Inc.

Mrs. Gracie A. McCoy purchased this house in August 1986. The mortgage company that financed this purchase, Cut Rate Loans Inc., used the July 22, 1986, survey by H. T. Shot, P.L.S. to document this sale. Nit Pic Title Inc. has the local reputation of being extremely meticulous, Easy Money Inc. is known to be a very cautious lender, and H. T. Shot’s reputation is excellent. Cut Rate Loans did not require a new title examination, nor was an updated survey requested.

Eventually, Kingston Heights subdivision was also completed. The house and lot directly west of 8627 Haygood Blvd. was purchased by A. M. Schorr in January of 1987. On February 2, 1987, McCoy and Schorr agreed mutually to fund a fence along the line dividing their properties. The survey marks on Schorr's eastern boundary were visible, so the fencing company used these marks to build a wood fence along Schorr's eastern line.

McCoy became dissatisfied with the size of her garage, razed it, and rebuilt the garage to the dimensions of approximately 24 feet × 24 feet. Mrs. McCoy became dissatisfied with the entire house and sold it to R. R. Robert. This sale was also financed by Cut Rate Loans. The appraiser for Cut Rate Loans inspected the property in May of 1987 and noticed that the new garage was over seven feet east of the west fence line. No other construction had taken place. No survey was required by the buyer or the lender for this new sale.

In August 1988, R. R. Robert agreed to sell this same house (at a great profit) to Schorr's widowed mother-in-law, B. Axer. The loan for this sale was handled by Easy Money Inc., which immediately ordered a resurvey of this lot, in spite of the existence of a recent ALTA/ACSM land title survey on this property, dated July 22, 1986. This survey is shown in Figure 12.6

The reader may evaluate this survey plat, as well as contemplate the many solutions to the problem presented. The savings realized by not requiring a boundary survey for each exchange of title cannot compare with the losses that will result from the complicated lawsuit that Robert is about to bring against all involved. The rule of thumb in real property transfers is, "The land survey is too expensive when it is more than the cost of possible losses or the value of the land."

12.5. METES MEETS BOUNDS

Abraham Savings and Loan (AS&L), located in Capital City, east state is in the process of providing a mortgage loan to Mr. William Cheatham. Mr. Cheatham provided a copy of a survey plat, showing the property to be mortgaged. Figure 12.7 shows that survey plat.

You are a loan officer with AS&L and have been assigned this loan. Using the checklist from Chapter 9, you discover the following:

1. The land record system is a metes and bounds system. Two separate parcel histories (Johnston estate and Mills tract) must be examined for title certification in order to insure prime lien on the property. A copy of the survey plat is sent to the title insurance company, along with a request for insurance.
2. This is clearly a tangent plane projection survey.

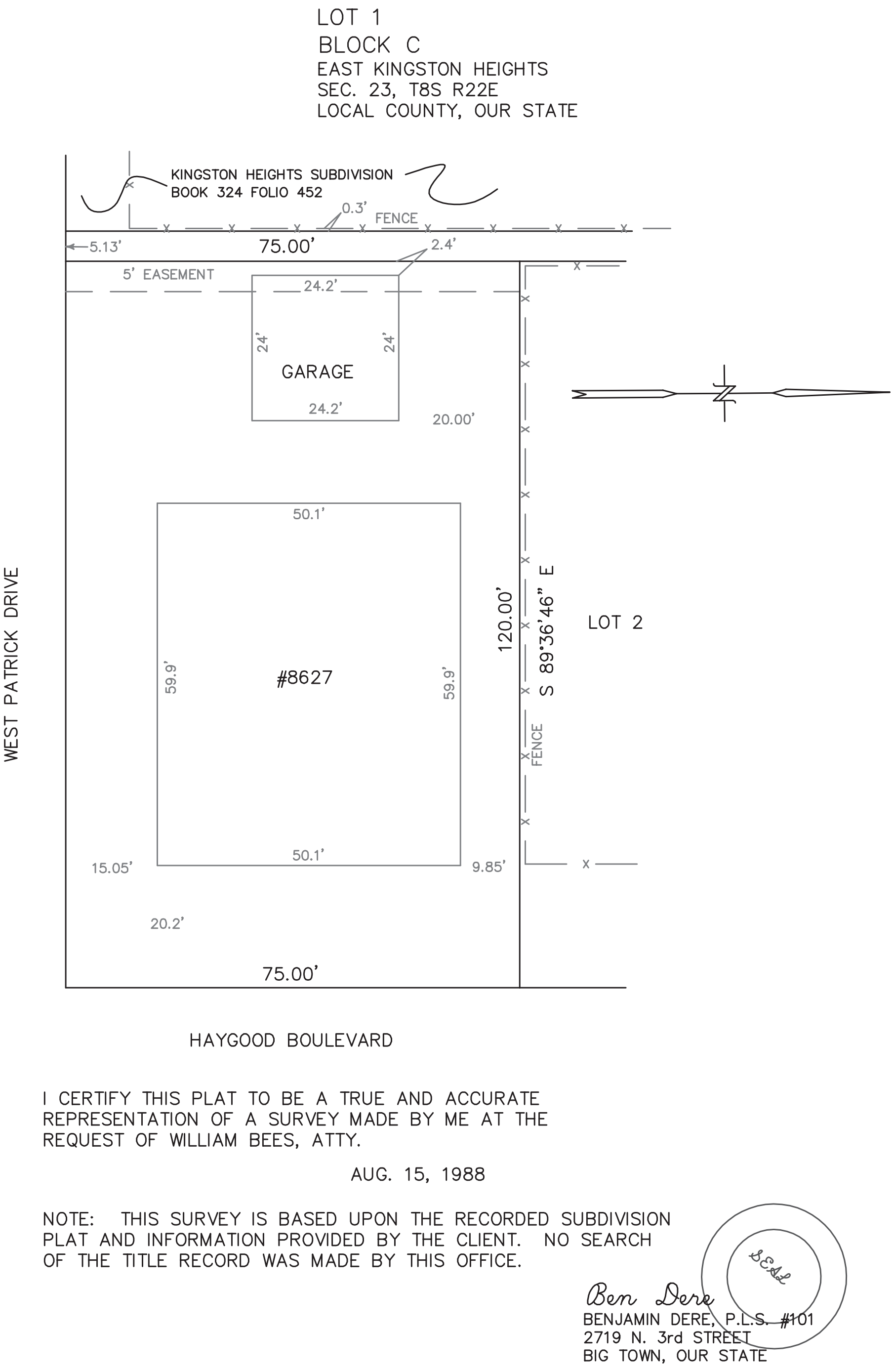

LOT 1
BLOCK C
EAST KINGSTON HEIGHTS
SEC. 23, T8S R22E
LOCAL COUNTY, OUR STATE

KINGSTON HEIGHTS SUBDIVISION
BOOK 324 FOLIO 452
0.3'
FENCE
5.13'
75.00'
2.4'
5' EASEMENT
24.2'
24'
24'
GARAGE
24.2'
20.00'
50.1'
59.9'
#8627
59.9'
50.1'
120.00'
S 89°36'46" E
FENCE
LOT 2
WEST PATRICK DRIVE
15.05'
9.85'
20.2'
75.00'
HAYGOOD BOULEVARD

I CERTIFY THIS PLAT TO BE A TRUE AND ACCURATE REPRESENTATION OF A SURVEY MADE BY ME AT THE REQUEST OF WILLIAM BEES, ATTY.

AUG. 15, 1988

NOTE: THIS SURVEY IS BASED UPON THE RECORDED SUBDIVISION PLAT AND INFORMATION PROVIDED BY THE CLIENT. NO SEARCH OF THE TITLE RECORD WAS MADE BY THIS OFFICE.

Ben Dere
BENJAMIN DERE, P.L.S. #101
2719 N. 3rd STREET
BIG TOWN, OUR STATE

FIGURE 12.6

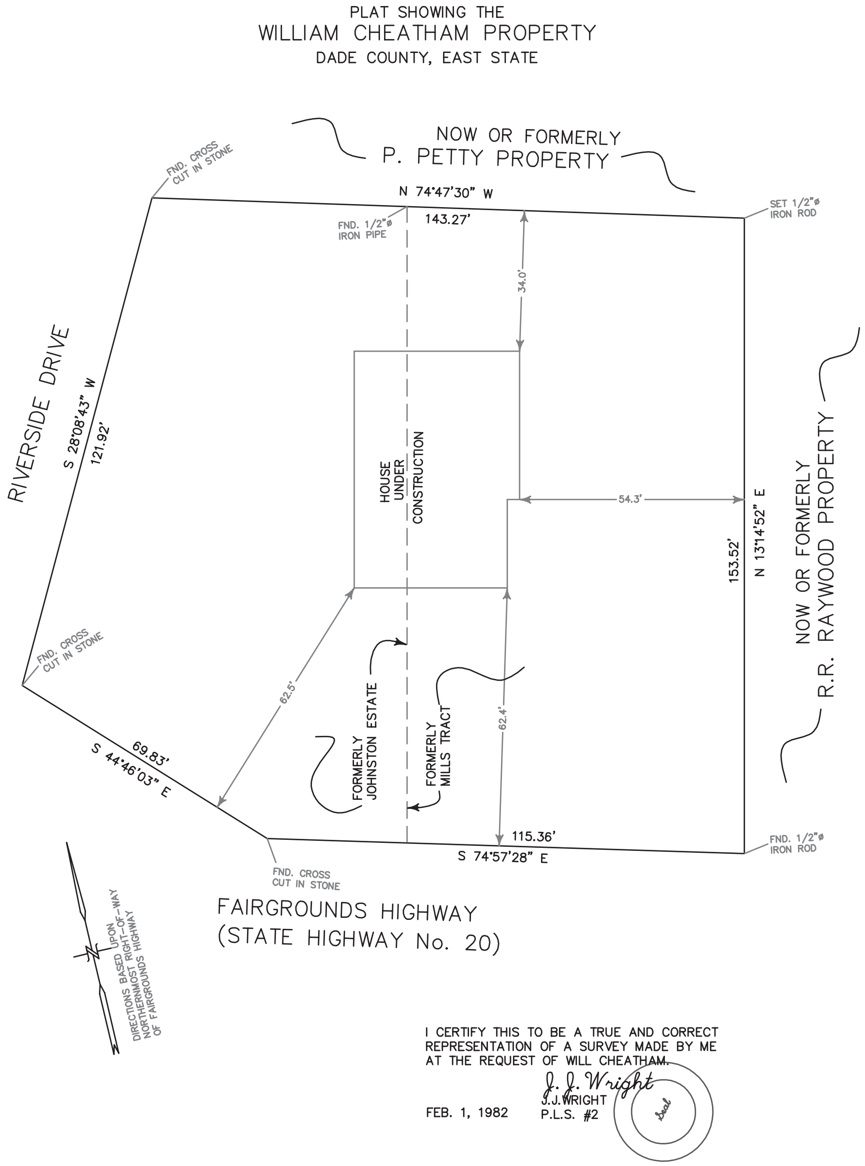

PLAT SHOWING THE
WILLIAM CHEATHAM PROPERTY
DADE COUNTY, EAST STATE

NOW OR FORMERLY
P. PETTY PROPERTY
FND. CROSS CUT IN STONE
N 74°47'30" W
143.27'
FND. 1/2"ϕ IRON PIPE
SET 1/2"ϕ IRON ROD
34.0'
RIVERSIDE DRIVE
S 28°08'43" W
121.92'
HOUSE UNDER CONSTRUCTION
54.3'
153.52'
N 13°14'52" E
NOW OR FORMERLY
R.R. RAYWOOD PROPERTY
FND. CROSS CUT IN STONE
62.5'
FORMERLY JOHNSTON ESTATE
FORMERLY MILLS TRACT
62.4'
69.83'
S 44°46'03" E
115.36'
S 74°57'28" E
FND. 1/2"ϕ IRON ROD
FND. CROSS CUT IN STONE
FAIRGROUNDS HIGHWAY
(STATE HIGHWAY No. 20)
N
DIRECTIONS BASED UPON NORTHERNMOST RIGHT-OF-WAY OF FAIRGROUNDS HIGHWAY
Seal

I CERTIFY THIS TO BE A TRUE AND CORRECT REPRESENTATION OF A SURVEY MADE BY ME AT THE REQUEST OF WILL CHEATHAM.

J. J. Wright
J.J.WRIGHT
P.L.S. #2

FEB. 1, 1982

FIGURE 12.7

3. The precision of the work seems to be modern.
4. The purpose of the survey appears to be documentation that the house under construction is within the property lines. The fact that a house is shown as "under construction" may be of serious concern if this is to be a first mortgage.

5. a. The north arrow is shown, and the bearing base is defined as the “call” direction for the Fairgrounds Highway (assumed north).
   b. The legal description is “the William Cheatham Property, Dade County, east state,” along with the bounds shown.
   c. The date is shown.
   d. The name of the Land Surveyor is shown, but no address is given.
   e. The signature and seal of the Land Surveyor are shown.
   f. All adjoiners are shown.
   g. All dimensions are shown.
   h. All directions are shown.
   i. The name of the client is shown.
   j. The certification is very general and gives no hint about the intended purpose of the survey.
   k. There are no limiting words or phrases.
   l. The area is not shown.
   m. The scale is not shown.
6. The deed presented to you by Cheatham describes the property as being bounded by Riverside Drive, Fairgrounds Highway, the property of B. B. Mills, and the property of P. Petty, and more particularly described as:

   > . . . Commencing at the corner of Fairgrounds Highway and Riverside Drive and True Point-of-Beginning; thence, South 75 degrees East along Fairgrounds Highway, a distance of 32 feet no inches; thence, North 13 degrees 15 minutes East, a distance of 153 feet no inches; thence, North 75 degrees West for 65 feet no inches; thence, fronting on Riverside Drive 122 feet no inches and a second fronting of 70 feet no inches to the point-of-beginning. . .

   The deed, though poorly written, clearly describes only a portion of the property depicted on the survey plat. No reference is made to the inclusion of the Mills tract.
7. No easements are shown.
8. No encroachments are shown onto or from the Cheatham property.
9. Accuracy standards are not stated.

10. Discrepancies between deed and the survey plat are profound and must be explained.
11. A telephone call to the Land Surveyor (the east state board of registration provided an address and a telephone number) reveals that the survey plat was prepared to assist in the resubdivision of the Mills tract and the Johnston estate into one parcel. The local county ordinances require the introduction of a plat of resubdivision, along with proof of ownership by the parties requesting the resubdivision. The county, upon approval, records a signed copy of the survey plat as the official plat of resubdivision.

The title insurance company reported that Cheatham's title to the Johnston estate was free and clear. The Mills tract was purchased by Cheatham Enterprises, Inc. and carries a mortgage in favor of First Bucks International for $125,000. There is no record of a resubdivision combining the two parcels.

12.6. THE SQUARE THAT WASN'T THERE

In 1997 you were requested by Cut Rate Loans Inc., a mortgage company, to examine the title on a parcel of land identified as "lots 1 through 22, square 16 Grandsire Farms, section 8, T4N R21W, this county, some state." The purpose of the title examination was for an act of sale, 70 percent financed by Cut Rate Loans, from the estate of O. Fogey to B. C. Jingwaski. Fogey acquired the property in 1908 from Night Investments, Inc., which had purchased all of section 8, T4N R21W from the U.S. government in 1899. Figure 12.8 shows the subdivision plat of Grandsire Farms that was attached to the 1908 conveyance.

The subdivision of Grandsire Farms had only been partially developed at the time. Some streets had been improved, some squares had been resubdivided, and some areas were untouched woodlands. Jingwaski acknowledged that he was aware that square 16 and the surrounding street rights-of-way were heavily wooded and undeveloped. This area had become a rapidly developing suburban area of Cin City, and the sale price for square 16 was $800,000.

The examination of the conveyance record revealed only two exchanges involving square 16: The first was the sale by the U.S. government to Night Investments of all of Section 8, and the second was

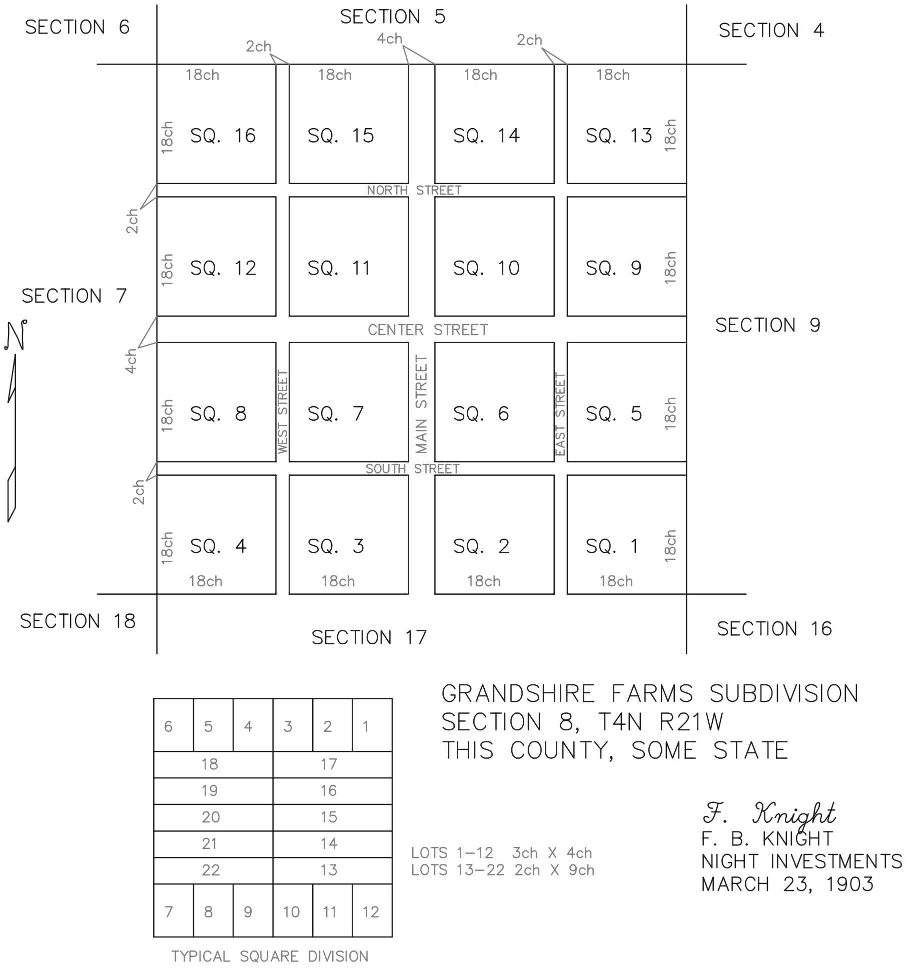
SECTION 6
SECTION 5
SECTION 4
2ch
4ch
2ch
18ch
SQ. 16
SQ. 15
SQ. 14
SQ. 13
NORTH STREET
SQ. 12
SQ. 11
SQ. 10
SQ. 9
SECTION 7
SECTION 9
CENTER STREET
N
SQ. 8
SQ. 7
SQ. 6
SQ. 5
WEST STREET
MAIN STREET
EAST STREET
SOUTH STREET
SQ. 4
SQ. 3
SQ. 2
SQ. 1
SECTION 18
SECTION 17
SECTION 16
6 5 4 3 2 1
18 17
19 16
20 15
21 14
22 13
7 8 9 10 11 12
TYPICAL SQUARE DIVISION
GRANDSHIRE FARMS SUBDIVISION
SECTION 8, T4N R21W
THIS COUNTY, SOME STATE
LOTS 1–12 3ch X 4ch
LOTS 13–22 2ch X 9ch
F. Knight
F. B. KNIGHT
NIGHT INVESTMENTS
MARCH 23, 1903

FIGURE 12.8

the sale by Night Investments to Fogey of lots 1 through 22, Square 16 "being all of Square 16." The record was as clean as you had ever seen.

An evaluation of the subdivision plat, using the checklist from Chapter 9, provides the following:

1. The land record system is a recorded subdivision plat.
2. The presence of cardinal directions (north, south, east, and west) and the lack of any indication that angles were measured indicate that this is a projectionless map.

3. The plat is dated March 23, 1903, but, because all of the distances are to the nearest chain and all of the directions are cardinal, you suspect that the last work done on this parcel was by a U.S. government surveyor prior to 1899.
4. The purpose of the plat (there was great doubt that a survey was ever conducted) is to create Grandsire Farms subdivision.
5. a. The north arrow is without reference to a bearing base.
   b. The legal description is indicated by the numbers in designated lots and squares. The legal description of Grandsire Farms is on the face of the plat.
   c. The plat is dated 1903.
   d. The name and address of the "surveyor" are shown.
   e. There is a signature but no seal.
   f. Adjoining properties are shown.
   g. All dimensions are shown. The exterior dimensions are identical to a regular USPLS section.
   h. Bearings are shown on the section boundaries only.
   i. The name of the client is shown, and the purpose of the plat is obvious.
   j. There is no certification.
   k. There are no limiting words or phrases.
   l. No areas are shown.
   m. There is no scale shown on the plat.
6. The deed only refers to lot and square numbers. There is no supplemental description.
7. Street rights-of-way are the only features shown on the plat.
8. No improvements are shown. You are familiar with the area and know that it is free of any fences, buildings, or cultivation.
9. You feel that $800,000 worth of land in a rapidly developing area adjacent to Cin City calls for a "Class A" survey.
10. The 1903 subdivision plat does not fulfill very many of the requirements of an ALTA/ACSM land title survey.
11. The "surveyor" has been dead for 32 years.

You advised Cut Rate Loans that you would require that a new survey be performed on the property being sold. Your surveyor quoted a cost of $8,450.00 to perform an ALTA/ACSM land title survey. Cut Rate

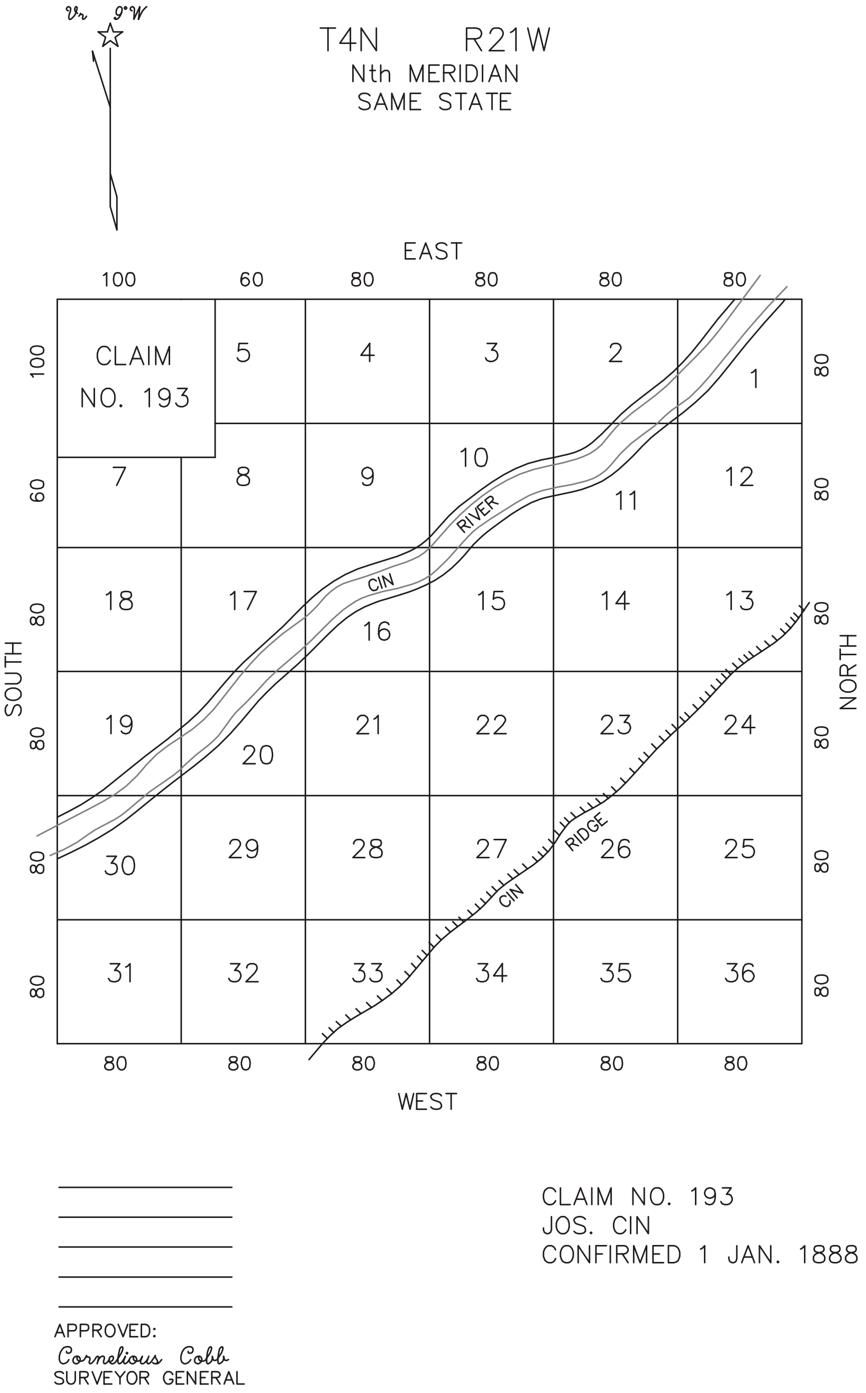
Vr 9°W
T4N R21W
Nth MERIDIAN
SAME STATE
EAST
100
60
80
80
80
80
CLAIM NO. 193
5
4
3
2
1
7
8
9
10
11
12
18
17
16
15
14
13
19
20
21
22
23
24
30
29
28
27
26
25
31
32
33
34
35
36
CIN RIVER
CIN RIDGE
SOUTH
NORTH
WEST
100
60
80
80
80
80
CLAIM NO. 193
JOS. CIN
CONFIRMED 1 JAN. 1888
APPROVED:
Cornelious Cobb
SURVEYOR GENERAL

FIGURE 12.9

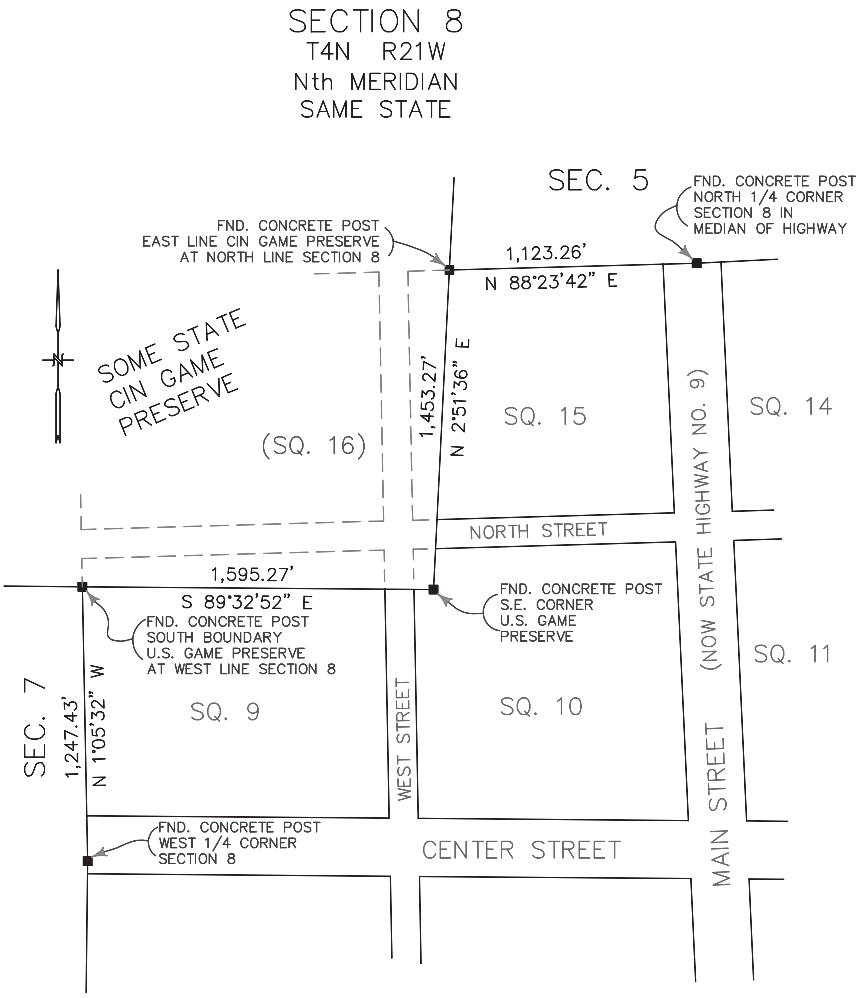
SECTION 8
T4N R21W
Nth MERIDIAN
SAME STATE
SEC. 5
FND. CONCRETE POST NORTH 1/4 CORNER SECTION 8 IN MEDIAN OF HIGHWAY
FND. CONCRETE POST EAST LINE CIN GAME PRESERVE AT NORTH LINE SECTION 8
1,123.26'
N 88°23'42" E
SOME STATE CIN GAME PRESERVE
1,453.27'
N 2°51'36" E
SQ. 15
SQ. 14
(SQ. 16)
(NOW STATE HIGHWAY NO. 9)
NORTH STREET
1,595.27'
S 89°32'52" E
FND. CONCRETE POST S.E. CORNER U.S. GAME PRESERVE
FND. CONCRETE POST SOUTH BOUNDARY U.S. GAME PRESERVE AT WEST LINE SECTION 8
SQ. 11
SEC. 7
1,247.43'
N 1°05'32" W
SQ. 9
WEST STREET
SQ. 10
MAIN STREET
FND. CONCRETE POST WEST 1/4 CORNER SECTION 8
CENTER STREET

I CERTIFY THAT THIS PLAT IS A TRUE AND ACCURATE REPRESENTATION OF MY ATTEMPT TO SURVEY SQUARE 16, GRAND SHIRE FARMS AS SHOWN ON THE SUBDIVISION PLAT BY F.B. KNIGHT cir 1903

JUNE 7, 1990
C. Brushcut
C.W. BRUSHCUT, P.L.S.
SOME STATE

FIGURE 12.10

Loans and Jing-waski both knew that the land was heavily wooded and completely vacant. Both decided that there was no possibility of any adverse possession and declined to commission a survey. (Who ever heard of paying $8,500 for some guy to stick pipes in the ground in the middle of nowhere?)

Last week, Bigbucks Development, Inc. signed a purchase agreement with Jingwaski for all of square 16 Grandsire Farms for the sum of $1,250,000. Bigbucks Development contracted with a surveying and engineering firm to survey square 16 and to design a single-family home-site development. Bigbucks Development directed the Land Surveyor to send the survey of square 16 to you to be used in the title examination for their purchase from Jingwaski. Today, the Land Surveyor, C. W. Brushcut, presented you with Figures 12.9 and 12.10. Figure 12.9 shows the U.S. township plat for township 4 north, range 21 west. Figure 12.10 shows his survey plat of square 16.

"Call me when you have a chance," said Brushcut as he left.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 13

WRITING LEGAL DESCRIPTIONS

The definitions of "legal description" and "land description" are sometimes thought to be interchangeable. However, there is a subtle difference between the two, and neither definition is exactly what the modern use of the phrase "legal description" has come to mean.

A legal description is "a description recognized by law which definitely locates property by reference to government surveys, coordinates systems or recorded maps; a description which is sufficient to locate the property without oral testimony." A land description is one that defines "the location of a parcel of land described in a narrative format." A metes and bound description is "a description of a parcel of land by references to course and distances around the tract, or by reference to natural or record monuments."[1]

The use of the phrase "legal description" has evolved to mean the body of words and phrases used within a real property deed to depict a specific parcel of land in a way that identifies that parcel, not only in terms of the land record system used to create the parcel, but also in terms of delineating the size, shape, and location of the parcel so specifically that recovery of the boundaries of that parcel is possible using only the written description.

In other words, the modern lexicon has combined "legal description" with a "metes and bounds description" to mean the combination of phrases that are to be included in the deed. This particular “legal” or “land” description takes the same form in every land record system and, because we will only concern ourselves with descriptions within deeds, will be called the “deed description” hereafter.

---

[1] American Congress on Surveying and Mapping and the American Society of Civil Engineers, *Definitions of Surveying and Related Terms,* 2005 (rev.).

This chapter will be devoted to developing the art of writing deed descriptions that are clear, complete, concise, and consistent (with the land record system in use and the intentions of the parties creating the title).

When called upon to determine intent, courts will consider all of the evidence presented in the deed description, as well as other evidence.[2] A well-written deed description will reflect the intent of the parties and will assist some future court in its deliberations. It is imperative that a deed description fully and unambiguously convey the *intent of the parties*.

There will be only one recommended form of deed description covered in this chapter, and no attempt will be made to discuss the portions of the deed that cover the specific rights conveyed, conditions of the sale, and so forth. Many other forms of deed descriptions in use across the United States are very regionally oriented and are being gradually replaced by this more universal and flexible form. The method proposed here will enable persons with little or no trigonometric or geometric background to write excellent deed descriptions.

The first step in writing a deed description is to evaluate the intentions of the parties to the transaction. In most cases, this intent is to transfer certain rights to an existing parcel. In other cases, a new parcel is being created. Each of these situations has its own particularities that may alter slightly the composition of the deed description.

13.1. CREATING NEW PARCELS

Divisions of a real property into two or more new parcels, combining existing parcels into one, or rearranging the boundaries of existing parcels are all acts of resubdivision. A resubdivision is most easily accomplished by a complete land survey and the development of a subdivision plat by a Land Surveyor. Although other land record systems allow for resubdivision of land, the platted subdivision accomplishes this so clearly and effectively that most local city and county governments now require platted subdivisions for all divisions of real property within their jurisdictions.

---

[2] J. S. Grimes, *Clark on Surveying and Boundaries*, 4th ed. Bobbs-Merrill Co., Inc.: Indianapolis & New York, 1976.

A new survey should be performed to collect the information about the property needed in order to decide upon division. Old survey plats will not reflect accurately present conditions or discover changes in boundaries. Current information is vitally important. Without it, the parties to the transaction may be misled.

Divisions of real property based on ancient data risk severe conflicts when, at a later date, the resulting parcels are possessed and it is discovered that the parent tract was not the size and shape it was believed to be. The creation of a platted subdivision affords the owners an opportunity to discover and to correct any dimensional discrepancy that may have existed in the old deed information. Once in possession of complete and accurate information on the actual present configuration of the parent tract, the parties involved in the resubdivision can develop their informed intent and take action without excessive risk of conflict among the assigns.

13.2. EXISTING PARCELS

Most often, a real property transaction involves an existing parcel that is already described in a deed. The current deed may or may not accurately describe the property. Changes in boundary lines may have occurred, or updated information about the boundaries may render portions of the old deed information obsolete. If a deed is to report accurately all of the rights transferred, then it is important that these rights be known. The ALTA/ACSM land title survey procedure[3] is designed to discover all of these rights and to detect possible changes.

13.3. GENERAL OUTLINE

All deed descriptions take the form of a five-part essay consisting of (1) caption, (2) narrative, (3) deletions or additions, (4) summation, and (5) references. Part (2) may be absent if information, beyond the caption, is not known. Parts (3) and (5) may be absent, and the order of (3), (4), and (5) is sometimes rearranged. Nevertheless, deed descriptions that fulfill all of the five parts have proven to be the least likely to be misinterpreted at a later date.

---

[3] See section 11.23.

13.4. THE CAPTION

The caption usually takes the form of identification that is most basic to the land record system appropriate to the parcel and is almost always the legal description of the parcel. In United States Public Land Survey (USPLS) and platted subdivisions, the caption is always distinct from the rest of the description.

In the metes and bounds system, the caption and the narrative may merge if caution is not exercised. To avoid this possibility, the phrase "and more particularly described as follows" or words to that effect are sometimes added at the end of the caption to alert the reader that the narrative section of the deed description is about to begin. The caption should also be segregated from the body of the deed description as a separate paragraph. Here are some examples of possible captions:

> That parcel of land designated as Lot 1, Square 12, Timber Estates, Waterproof, Louisiana, more particularly described as follows: . . . (platted subdivision)
>
> That parcel of land known as the "Brown Estates" located in Winston County, Kentucky bounded on the north by J. R. Westly, on the east by Philip Jones, on the south by D. Davis and on the west by Kentucky State Highway Number 5, more particularly described as follows: . . . (metes and bounds)
>
> That parcel of land designated as the NE$^1/_4$ of Section 2, T5S R4E, Fourth Principal Meridian, Hayseed County, Illinois, described as: . . . (USPLS)

The common factor in each of these examples is that the *caption alone is sufficient to identify the parcel without oral testimony*, based on the prevailing land record system. The land record system in the examples is noted within parentheses following the caption.

The identification in the caption is not normally negated by a contradiction occurring in the narrative portion of the deed description, unless it can be shown that the caption contains an error. If, for example, in a platted subdivision, the caption identifies a parcel by lot, square, and subdivision, according to a recorded subdivision plat, and the narrative portion reports a contradictory distance to a street corner, the caption (and the recorded plat that is an integral part of the caption by reference) will usually be assumed to be correct, unless additional evidence to the contrary has been found.

13.5. THE NARRATIVE

The narrative portion of the deed description always takes the form of a complete metes and bounds description, beginning with a commencement point, through a point-of-beginning, around the parcel, and returning to the point-of-beginning. The narrative forms separate paragraphs that consist of a single sentence.

The distinctions of calls for specific boundaries, bearings, distances, landmarks (monuments), and other factors discussed in Chapter 7 dealing with metes and bounds land record systems apply in a deed description. The narrative portion of a deed description is the same for all land record systems and is dependent on a survey's having been performed on the parcel being described.

Divisions of property that take place without a survey, such as the aliquot division of a section, should not have a narrative portion in the deed description.

As many of the key factors of a metes and bounds description as may exist must be presented in the narrative portion of a deed description, and, on the off chance that the reader did not read that section or does not remember it, the key factors will be repeated here.

The metes and bounds narrative must contain the following:

1. A commencing point that is well known, easily found, durable, recoverable, recognizable, and preferably public in origin.
2. A point-of-beginning that is a distinct part of the parcel being described.
3. A report of the physical objects (monuments) that mark the location of the ends or are found along each line.
4. A report of the owners or parcel identifiers of the contiguous properties along each line.
5. A direction for each line. This can be in the form of a bearing, an angle, or, in the case of curved lines or meandering lines, a description of the configuration of the line.
6. A length of each line segment, usually in the form of a distance between corners.

There are several recommended key words or phrases that should be used in the narrative of a deed description that are intended to assist the reader in identifying which of the preceding items is about to be reported or to clarify some other aspect of the description. Custom in a particular area may mean that a different combination of key words or phrases will be used to express the intent of the parties. In these areas, the introduction of the recommended system should be made gradually or with slight modifications for the sake of continuity.

How corner monumentation is identified is also significant in illustrating the intentions of the parties. Each region has developed certain preferred objects used to monument corners. Terrain, agriculture, local industries, geology, and other factors unique to a region all influence the materials used in the monumentation of real property corners.

The format recommended here is easily adapted to a wide variety of regional customs and practices involving identification, monumentation, and description of real property. Once this format becomes a habit, reading or writing the narrative portion of a deed description will become the simplest part of working with real property identification and boundaries.

13.6. KEY PHRASES

13.6.1. "Commencing At"

"Commencing at" normally are the first two words of the narrative portion of a deed description. The reader is alerted to the fact that items following this phrase will be a physical object and/or a distinct theoretical land point that, in the opinion of the writer, fulfills most, if not all, of the requirements enumerated for commencement points.

The commencement point must be just that—a land point. It would not be proper to state, for instance, "Commencing at the northernmost right-of-way line for U.S. Highway 66." The reader is not told where along the northernmost right-of-way line the land point is located. It would be correct to state, "Commencing at the intersection of the northernmost right-of-way line of U.S. Highway 66 and the easternmost right-of-way line of U.S. Highway 15."

The second version clearly designates a land point that any reader of the deed could easily find. Physical objects that monument the commencement point should also be "called for" in the narrative, as well as any state plane coordinates (SPCs) or other pertinent data.

13.6.2. "Point-of-Beginning"

The point-of-beginning brackets the narrative that traces or "traverses" along the boundaries of the parcel being described. This statement prepares the reader to begin sketching a mental picture of the parcel. The point-of-beginning is fundamental to the narrative portion of a deed description and, as with the commencement point, must be a true land point. Physical monuments, as well as theoretical land points, can serve as points-of-beginning. It is quite common for the monument, theoretical location, and SPCs of a point-of-beginning to be included in the narrative.

Although it is sometimes combined with the commencement point, the point-of-beginning cannot be dispensed with. "Commencing and point-of-beginning at . . ." is used when both are the same point, as in a comer lot. The second appearance of the words "point-of-beginning" in a narrative announces the termination of the traverse encompassing the parcel. When it is mentioned the second time, there is no further elaboration of any distinctive features of the point-of-beginning.

13.6.3. "Thence" and ";"

The narrative in a deed description is similar to a Faulkner sentence. From "commencing at" to references, the entire essay comprises one sentence. The word *thence* is used to announce that a new line segment is about to begin. The boundary line segment description always ends with a semicolon (";") giving the phrase that describes a single boundary line segment the appearance "thence, xxxxxxx;."

In this way, the very long narrative can be divided into "digestible" pieces that stand out to the reader. It is important to exclusively reserve these two items (thence and the semicolon) for this particular use. Applications of semicolons or the word *thence* for another purpose anywhere else in the narrative portion of a deed description will certainly cause confusion.

13.6.4. "In a —ly Direction"

The eight general directions of "norther*ly,*" "northeaster*ly,*" "easter*ly,*" "southeaster*ly,*" "souther*ly,*" "southwester*ly,*" "wester*ly,*" and "northwester*ly,*" are especially useful to persons who are not particularly familiar with reading and interpreting bearings. Directions with the suffix *-ly* eliminate confusion over which form of a bearing is applicable for describing a particular line. Examine Figure 13.1.

If the traverse being described includes the line segment from point "A" to point "B," an examination of the north arrow clearly indicates that this requires a move to the south. The phrase "in a *southerly* direction" prepares the reader to expect the bearing "*south* 5 degrees

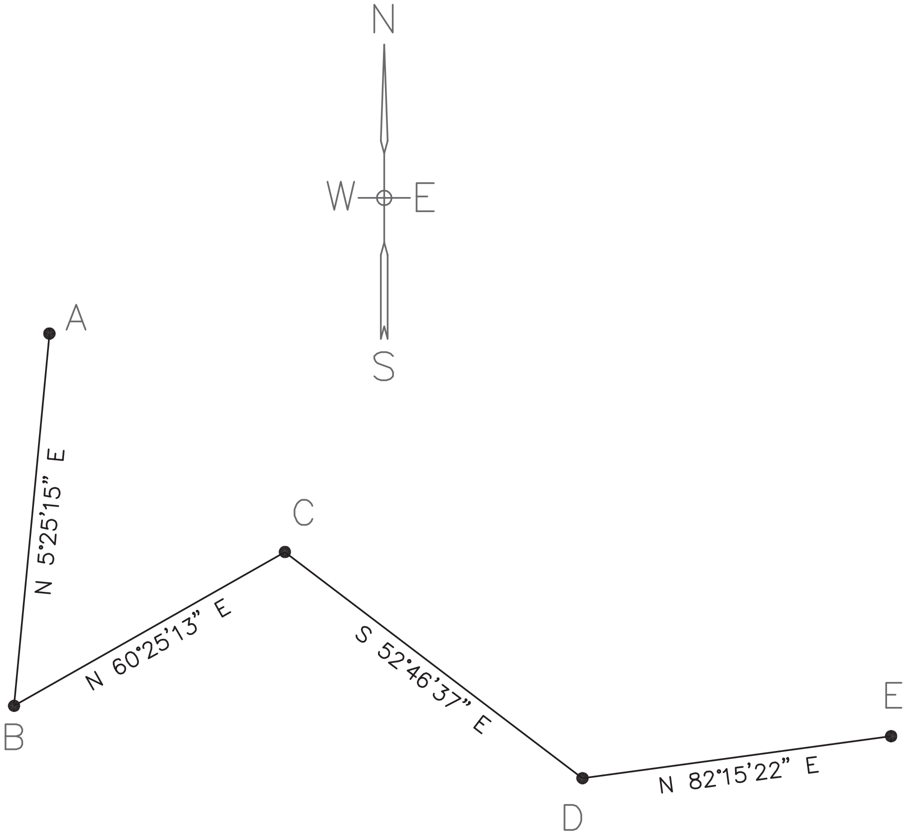

FIGURE 13.1

25 minutes 15 seconds west" instead of "*north* 5 degrees 25 minutes 15 seconds east," which is printed in Figure 13.1.

Similarly, from point "B" to point "C" would call for the use of the phrase "in a *northeasterly* direction" to indicate that the bearing to use is "*north* 60 degrees 215 minutes 13 seconds east." The line segment from point "C" to point "B" would call for the phrase "in a *southwesterly* direction" to indicate that the bearing to use for that direction is "*south* 60 degrees 25 minutes 13 seconds *west*."

Notice that, if in describing the line from point "C" to point "B" the writer had decided to use "in a *southerly* direction" instead of "in a *southwesterly* direction," the choice of bearings would still be "*south* 60 degrees 25 minutes 13 seconds west." The selection of "in a *westerly* direction" also results in "south 60 degrees 25 minutes 13 seconds *west*."

The use of these general directions allows a great deal of freedom of choice. The only limitation in choosing a *-ly* direction occurs when the lines are close to the cardinal directions. Lines that are nearly north–south should use "northerly" or "southerly"; lines that are nearly east–west should use "easterly" or "westerly."

13.6.5. "-Most"

The suffix *-most* is used with the same eight general directions as the suffix *-ly*. The purpose of the general directions is to distinguish between similar or paired items in a way that is obvious by an inspection of the property or its survey plat.

As with the *-ly* directions, there is wide leeway given in the choice of directions using the suffix *-most*. The majority of occasions that require these words can be satisfied with the cardinal directions.

In Figure 13.2, "northernmost," "northeasternmost," "easternmost," and "southeasternmost" all indicate the Jones and Smith side of U.S. Highway 77. Of the four choices, "northeasternmost" and "easternmost" are the most obvious, and equally preferred, selections in this example.

13.6.6. "Along"

The key word *along* indicates that the boundary line being described is congruent with some other previously established line. Whether the previously established line is straight, curved, or meandering, the key word *along* indicates that the boundary line being described is exactly superimposed on the previous line at every point.

For example, "... ; thence, in a southerly direction, *along the line between section 5 and section 6* ... ;" declares the described boundary to be the section line, and "... ; thence, in a northerly direction, *along the Snake River* ... ;" leaves no doubt that this is a riparian boundary.

13.6.7. "A Distance Of"

The phrase "a distance of" is sometimes considered to be unnecessary. However, deed descriptions that refer to distances that are not lengths of boundary line segments can become very confusing if this key phrase is eliminated.

For example, examine the excerpt "... ; thence, in a southerly direction, along a curved line, having the radius center to the east and a radius of 200.00 feet, 350.00 feet..." or "... ; thence, in a southerly

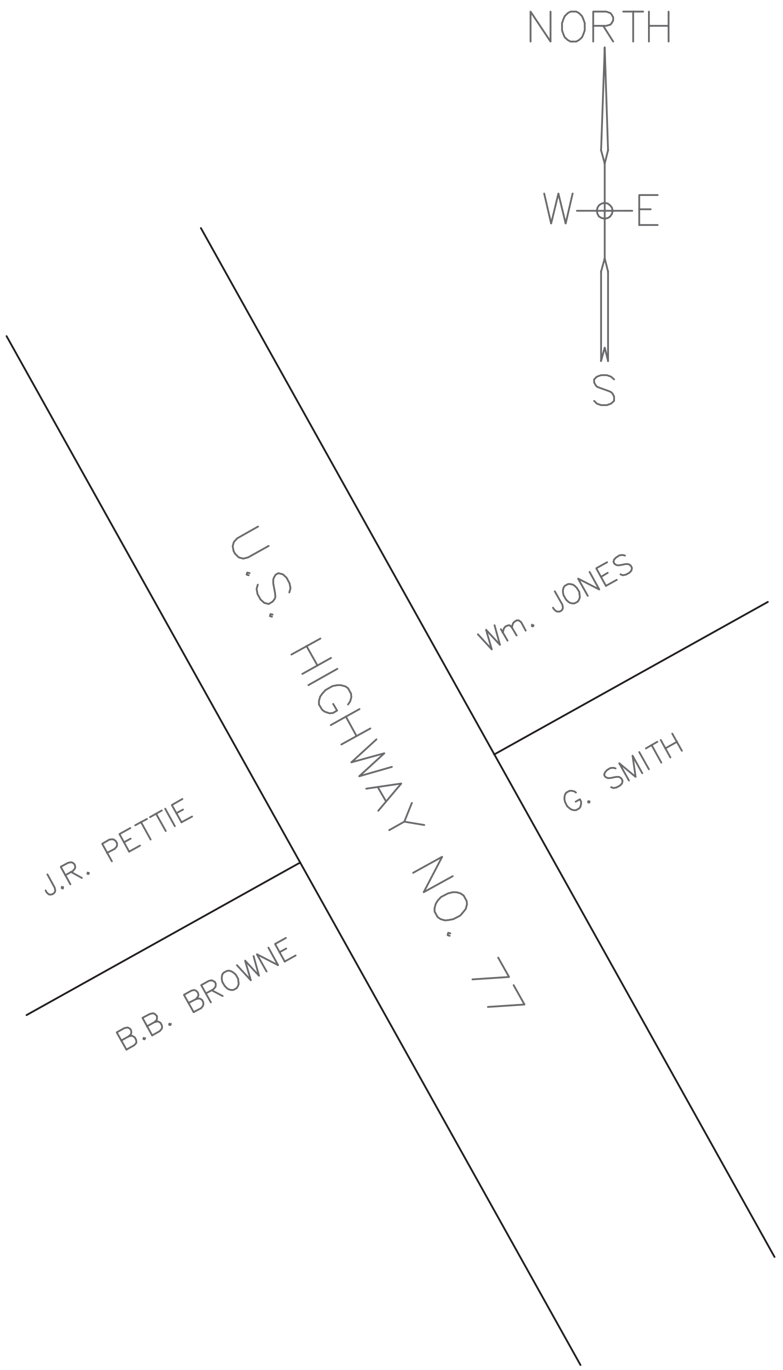
NORTH
W
E
S
U.S. HIGHWAY NO. 77
Wm. JONES
G. SMITH
J.R. PETTIE
B.B. BROWNE

FIGURE 13.2

direction, along a curved line having a radius of 200.00 feet and the radius center to the east, 350.00 feet . . ." is not as clear as ". . . ; thence, in a southerly direction, along a curved line, having the radius center to the east and a radius of 200.00 feet, *a distance of* 350.00 feet . . ."

13.6.8. "To" and "To a Point"

The key word *to* is of special significance in demonstrating the intention of the parties in a transfer of titles. *To* announces that the boundary line segment is being terminated and that the corner description that *immediately* follows the word *to* is considered by the writer to control the length of the boundary line segment. The words "to a point" indicate that the limit is not set by a physical feature but by a theoretical location. Many other supplementary descriptions of the corner, or accessories to the corner, may follow the primary one. This distinction is best explained by example:

". . . 723.32 feet *to the easternmost boundary of the Jones estate*, that same point being monumented by a 1-inch diameter iron pipe; . . ." clearly designates the "Jones estate" (an adjoining property) as the limit of that line segment. By using this phrase, the writer is stating that, if, for some reason, the one-inch diameter iron pipe was found to be short of the Jones estate or the distance called for was insufficient to reach the Jones estate, the limit of the line segment would not be the pipe or the distance but would continue on to (or stop at) the easternmost boundary line of the Jones estate.

". . . 723.32 feet *to a point* monumented by a 1-inch diameter iron pipe; . . ." indicates that the intention of the parties was to set the length of the line at *a specific distance* (723.23 feet), and, if the pipe were later to be found to be short of, or in excess of, that distance, the distance would prevail (in the absence of possession or other convincing and contradictory evidence). In this instance, the pipe is just a marker and has not "achieved" the status of an artificial monument.

". . . 723.32 feet *to* a 1-inch diameter iron pipe; . . ." indicates, according to the intent of the writer, that the iron pipe is a artificial monument and that it, not the distance, is the controlling factor.

The exception to this "rule" occurs in cases in which real property boundaries are named in any part of the corner description. The mention of a real property boundary usually indicates the intention to limit the line at that real property boundary, even if calls for markers or other corner identification occurs first.

In any case, the order is not exclusive and binding. As with any part of the deed description, other evidence of intent may be used to refute a contention based upon order of appearance of corner descriptions. In areas where the custom of writing deed descriptions is well established and does not follow the pattern that is recommended here, it may require some time to introduce this method.

13.6.9. "That Same Point"

The corner to a real property parcel is a theoretical location that may refer to several physical or theoretical things to define that location. Any corner may have several different items in the list of these things. A corner may be monumented by a surveyor's marker, it may be part of a common boundary line, it may have several accessories, the SPC values may be known, or other features may need to be described.

The phrase "that same point" prevents the reader of a deed description from confusing the features of one corner with those of another. An example of the use of this phrase is ". . . to a stone borne located on the line between sections 30 and 31, *that same point* having the Virginia state plane coordinates (south zone) of $y = 234{,}543.55$ feet, $x = 1{,}323{,}445.39$ feet; thence, . . . "

13.6.10. "Point of Curvature" and "Point of Tangency"

There are two types of intersections possible between lines when one or both of the lines are curved. A line—even another curved line—may simply intersect with a curved line, or it may meet that curved line in a very special way, known as a "tangent" intersection. Figure 13.3 shows several combinations of intersections involving curved lines.

The situations labeled "tangent intersection," "compound curve," and "reverse curve" are all intersections where the radius center of the curve has a special relationship to the point of intersection. In the straight line tangent intersection (case 2), the radius center is located on a line from the intersection that is at right angles to the straight line. In the case of the compound curve and the reverse curve (cases 4 and 5, respectfully), the radius centers are on a straight line drawn through the point of intersection.

The occurrence of tangent intersections in deed descriptions can include the use of the phrase "point of curvature" when a straight line ends and a curved line begins or "point of tangency" when a curved line ends and a straight one begins.

The phrase "point of compound curvature" is used when a curved line of one radius intersects with a curved line of a different radius, but the same direction.

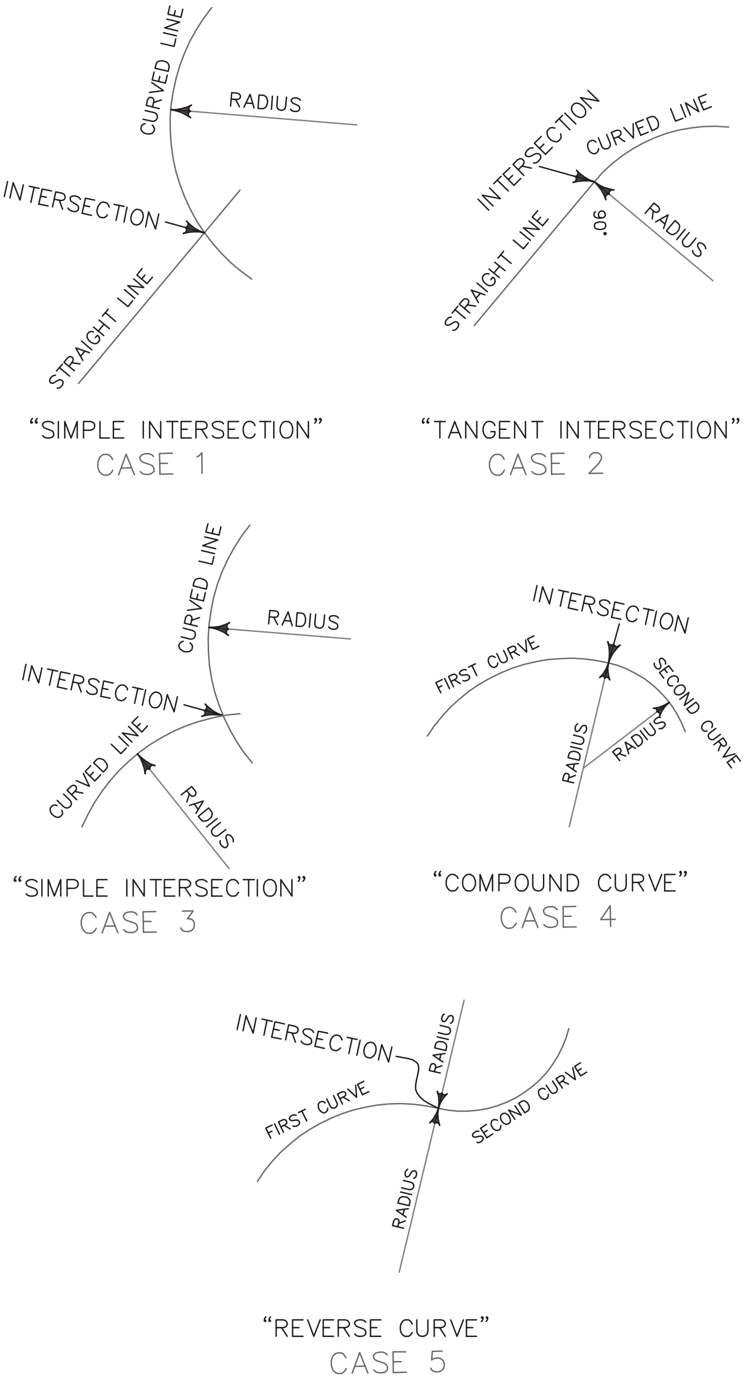
CURVED LINE
RADIUS
INTERSECTION
STRAIGHT LINE
INTERSECTION
CURVED LINE
RADIUS
90°
STRAIGHT LINE
CURVED LINE
RADIUS
INTERSECTION
CURVED LINE
RADIUS
INTERSECTION
FIRST CURVE
SECOND CURVE
RADIUS
RADIUS
INTERSECTION
RADIUS
FIRST CURVE
SECOND CURVE
RADIUS

"SIMPLE INTERSECTION"
CASE 1

"TANGENT INTERSECTION"
CASE 2

"SIMPLE INTERSECTION"
CASE 3

"COMPOUND CURVE"
CASE 4

"REVERSE CURVE"
CASE 5

FIGURE 13.3

The phrase "point of reversed curvature" is used when a curved line intersects with another line curving in the opposite direction at a point of tangency. Although all of the intersections discussed here occur at a point of tangency, the key phrase "point of tangency" is used in a deed description only when the description of a curved line ends at a point of tangency with a straight line.

These phrases are used only when there is a special orientation of circular curved lines with previous or subsequent lines and they are not used when simple intersections occur, such as cases 1 and 3 in Figure 13.3.

Some have argued that such "terms of art" should not be used in deed descriptions because most readers of the description do not know the definition of a tangent intersection. This may be true. However, the additional information provided by identifying tangent intersections is integral to the relationship of the boundaries and should not be omitted for fear that the reader is ignorant.

Readers who do not know terms in a deed description will consult a Land Surveyor or a dictionary and be educated. Deed descriptions that have important data excluded are incomplete and subject to possible misinterpretation.

13.6.11. "From Whence Bears" and "Distant"

The phrases "from whence bears" and "distant" are used exclusively to designate the location of accessories to a corner. The particular wording of the segments that describe accessories is deliberately "old-fashioned" so that accessory descriptions do not become confused with boundary line segments. For example, ". . . to an iron post *from whence bears* a large oak at north 23 degrees east and 23 feet distant, thence; . . ." makes the accessory a part of the deed description. In general, naming accessories in deed descriptions has not been routinely practiced.

13.6.12. "Encompassing an Area Of"

The phrase "encompassing an area of" should be placed directly after the second occurrence of the words "point-of-beginning" to emphasize that the area expressed is that of the entire parcel and does not account for any changes in area caused by the deletions or additions noted later in the deed description. For example, ". . . to the point-of-beginning; encompassing an area of 234.43 acres . . ." leaves no doubt that the narrative is intended to describe a parcel containing that area. If the statement of area is removed from the narrative portion of the deed description, then confusion can occur between area before and area after accounting for deletions or additions.

13.7. DELETIONS OR ADDITIONS

The deletions or additions portion of a deed description usually lists all of the easements, servitudes, rights-of-way, restrictions, or other modifications to the free, clear enjoyment of the parcel described.

Sometimes, more commonly in rural areas, this portion of the description actually entirely eliminates portions of the described tract from the deed. This practice can cause confusion and is becoming less common as local governments enact stricter and more formal codes or ordinances controlling the resubdivision of real property. The reader is urged to resubdivide and plat any deletion or addition to a parcel that completely changes all of the real property rights of that deletion or addition.

This deletions or additions portion of a deed description is clearly set apart from the narrative portion by the use of, in bold or capital letters, the heading "LESS THAN AND EXCEPT" "IN ADDITION" or "SUBJECT TO."

These are not found in every deed description. Unfortunately, if this heading appears on a separate page from the narrative portion of the deed description, it may become separated from the rest of the deed documents. For this reason, it is recommended that the summation and references always follow the deletions or additions portion of the deed description. If this procedure is followed, the reader will always know that a complete deed description is present if a caption to the references is present.

Some examples of the use of deletions or additions in deed descriptions follow:

> **"Subject To:**
>
> A 5-foot easement adjoining and along the entire rear or northerly portion of the property for the purpose of installation and maintenance of a telephone cable . . . "
>
> **"Subject To:**
>
> An easement of ingress and egress, 20-feet wide, adjoining and along the western boundary for the use of the owner of lot 3 and more particularly described as follows; . . ."

"Less Than and Except:

That portion of the property known as "Devils Well" and more particularly described as follows; . . ."

13.8. REFERENCES

The references portion of the deed description is used to "call for," thereby including in the deed, other documents, such as a survey plat, subdivision plat, maps, transactions, and the like. Items so included in the references will usually be considered as appearing in the deed description in full. A metes and bounds deed description *must* include a reference to the survey on which it is based.

A reference to a survey plat will normally have the effect of causing everything on the plat to become an integral and valid part of the deed. Sometimes the list of previous exchanges involving the parcel is also made in the references. Some examples of references follow:

> ". . . and all as more fully described on the plat of survey by T. T. Thomas, P.L.S., dated February 2, 1986."
>
> ". . . being the same property as acquired by the present owner from Sunshine Realty on December 4, 1944, and recorded in . . ."
>
> ". . . being the same property created by the subdivision plan of Oak Hills subdivision by T. Swift, L.S. and recorded in . . ."

The purpose of a deed description is to express, clearly and as completely as possible, the intent of the parties conducting the exchange of title. Information that is relevant to that intent ought to appear in the deed description in one of the categories already mentioned.

A properly written deed description will provide courts that are called on to settle a boundary dispute the information necessary to eliminate deed discrepancies as an object of contention. A little extra time and effort in writing deed descriptions will benefit any assigns or heirs by reducing confusion and litigation.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 14

EXERCISES IN WRITING DEED DESCRIPTIONS

The following exercises in writing deed descriptions will present a sketch of a fictitious real property parcel, special conditions, and the author's recommended deed description. In reviewing the recommended deed description, you may wish to evaluate the survey plat using the checklist in Chapter 12 and note whether the description reflects the circumstances accurately.

14.1. CASE 1

Bill Wright, the owner of section 11, T22N R43W, Hog County, farm state, wishes to sell, to Mike Farmer, the "southeast 160 acres" of section 11. Figure 14.1 shows the only drawing of the section in Wright's possession. Wright acquired section 11 from Crystal Shine on November 10, 1942.

Farmer confirms that he wants to buy the "SE$^1/_4$ of Section 11" because he needs "a square plot 2,640 feet on a side." Of course, the "SE$^1/_4$," the "southeast 160 acres," and "a square plot 2,640 feet on a side" are three different divisions of section 11.

The deed description of the SE$^1/_4$ of section 11 can be made without a survey and would be "the SE$^1/_4$ sec. 11, T22N, R43W, Nth PM, Hog County, farm state and being a portion of the same property acquired by Bill Wright, the present seller, from Crystal Shine on November 10, 1942."

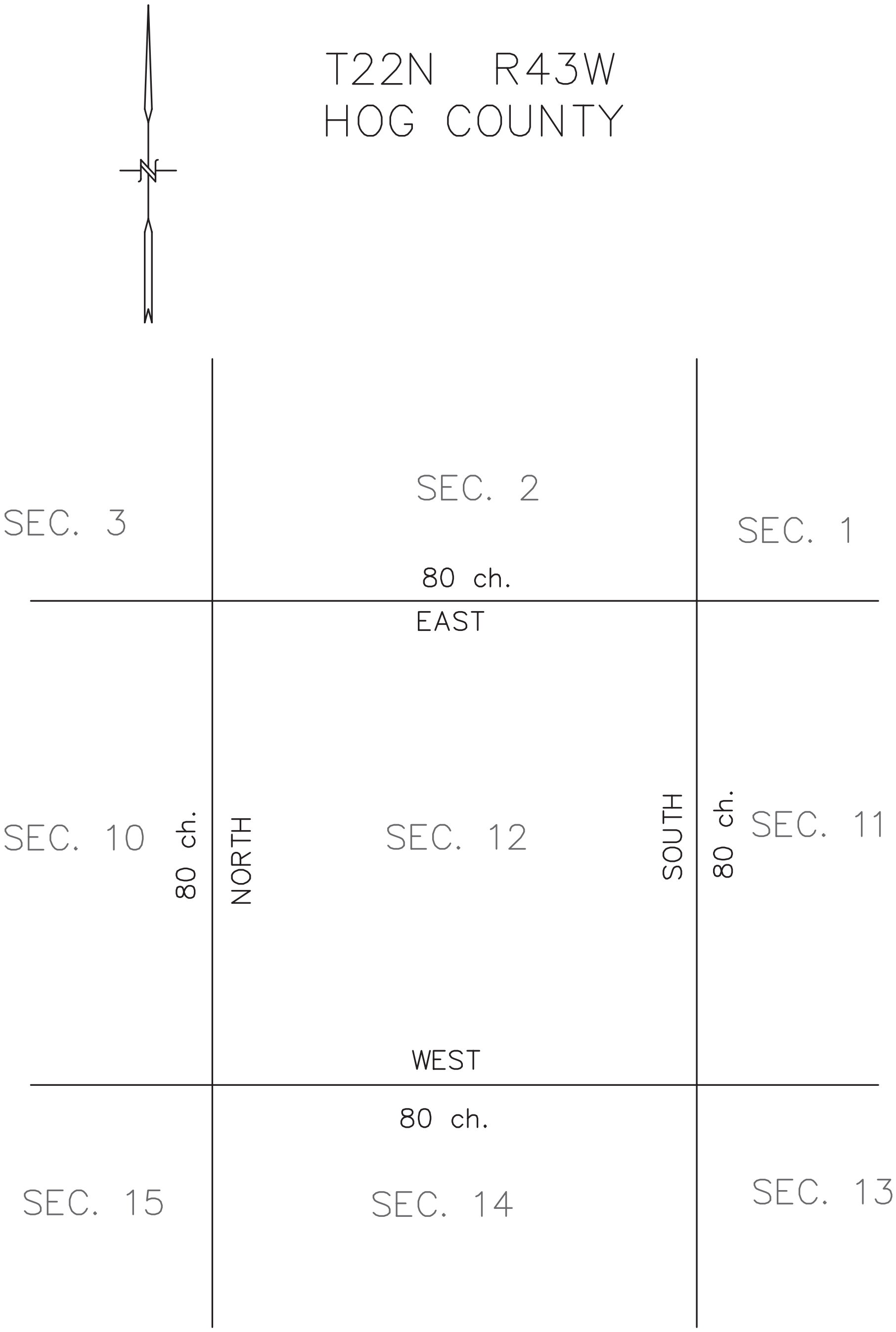
T22N R43W
HOG COUNTY
N
SEC. 3
SEC. 2
SEC. 1
80 ch.
EAST
SEC. 10
80 ch.
NORTH
SEC. 12
SOUTH
80 ch.
SEC. 11
WEST
80 ch.
SEC. 15
SEC. 14
SEC. 13

FIGURE 14.1

It would be quite improper to quarter section 11 based on the "ideal section" template and to add a narrative describing the SE$^1/_4$ as beginning at the section comer and measuring 40 chains (2,640 feet) to a side. If anything is certain about section 11, it is that the section is *not* 80 chains north, east, south, and west. Farmer wisely orders a survey of the SE$^1/_4$ of section 11. Figure 14.2 is a rendering of that survey.

An abstract of the deed record (title) does not indicate any recorded easements, servitudes or rights-of-way. The surveyor, Ben Round, reports that South Road was built and has been maintained by the local farmer's co-op. The size of the right-of-way shown is based upon a written agreement on file at the co-op and is signed by Wright and others. Now the deed description can be written. Wright and Farmer both know exactly what is to be transferred. The description might read as follows:

> That parcel or portion of land designated as the SE$^1/_4$ of section 11, T22N, R43W, Nth PM, Hog County, farm state, and more particularly described as follows:
>
> Commencing and point-of-beginning at the southeast corner of section 11, that same point being monumented by a 6-inch square concrete monument; thence, in a westerly direction along the line between sections 11 and 14, that same line being the centerline of the South Road right-of-way, north 89 degrees 52 minutes 14 seconds west, a distance of 2,712.43 feet, to the southern quarter corner of section 11 and a 2-inch diameter iron pipe; thence, in a northerly direction along the line between the southeast quarter and the southwest quarter of section 11, north 0 degrees 42 minutes 12 seconds west, a distance of 2,688.21 feet, to the center of section 11 and a 1-inch diameter iron bar; thence, in an easterly direction along the line between the southeast quarter and the northeast quarter of section 11, south 88 degrees 02 minutes 48 seconds east, a distance of 2,592.34 feet, to the eastern quarter corner of section 11 and a 2-inch diameter pipe; thence, in a southerly direction along the line between sections 11 and 12, south 3 degrees 23 minutes 42 seconds east, a distance of 2,610.36 feet, to the point-of-beginning; encompassing an area of 161.02 acres:
>
> SUBJECT TO: a right-of-way, 40 feet in width, adjacent to, and along, the entire southern line of the property herein described, for the purpose of the maintenance of a roadway:
>
> All as more fully described on a plat of survey by Ben A. Round, land surveyor, dated June 3, 1988, and being a portion of the same property acquired by Bill Wright, the present owner, from Crystal Shine on November 10, 1942.

PLAT SHOWING THE
SOUTHEAST QUARTER OF
SECTION 11, T22N, R43W
Nth P.M.
HOG COUNTY
FARM STATE

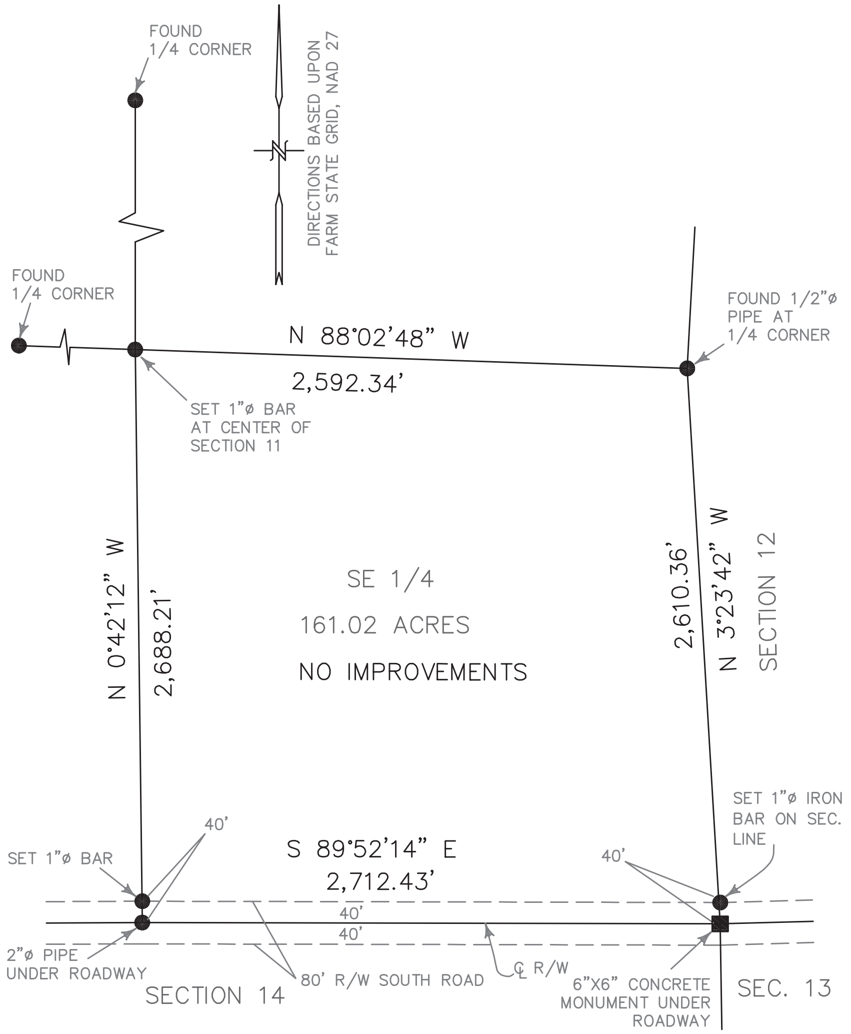
FOUND
1/4 CORNER
DIRECTIONS BASED UPON
FARM STATE GRID, NAD 27
N
FOUND
1/4 CORNER
N 88°02'48" W
2,592.34'
FOUND 1/2"ø
PIPE AT
1/4 CORNER
SET 1"ø BAR
AT CENTER OF
SECTION 11
N 0°42'12" W
2,688.21'
SE 1/4
161.02 ACRES
NO IMPROVEMENTS
2,610.36'
N 3°23'42" W
SECTION 12
SET 1"ø IRON
BAR ON SEC.
LINE
40'
SET 1"ø BAR
S 89°52'14" E
2,712.43'
40'
40'
40'
2"ø PIPE
UNDER ROADWAY
SECTION 14
80' R/W SOUTH ROAD
℄ R/W
6"X6" CONCRETE
MONUMENT UNDER
ROADWAY
SEC. 13

I CERTIFY THIS PLAT TO BE A TRUE AND
ACCURATE REPRESENTATION OF A SURVEY
PERFORMED BY ME AT THE REQUEST OF
MIKE FARMER.

Ben A. Round
BEN A. ROUND, L.S.
JUNE 3, 1988

FIGURE 14.2

14.2. CASE 2

Iron Town, Bay County, Maryland, was created by a subdivision plan by S. Blake, county surveyor. The plan was dated May 3, 1920, and recorded in the Bay County conveyance office in book 5, page 2. Figure 14.3 shows a portion of that plan. Blake noted on the plan that "square corners were marked by a 2-inch diameter cast iron pipe." Write a deed description for lots 1 and 3, square "B."

The description of lot 1 might be as follows:

> That parcel of land designated as lot 1, square B, Iron Town subdivision, Bay County, Maryland and all as more fully described as follows:
>
> Commencing and point-of-beginning at the intersection of the northeasternmost right-of-way line of 33rd Avenue and the northwesternmost right-of-way line of Main Street, that same point being monumented by a 2-inch diameter cast iron pipe; thence, in a northerly direction along said northeasternmost right-of-way line of 33rd Avenue, north 51 minutes 07 minutes 46 seconds west, a distance of 170 feet, to the southeasternmost right-of-way line of the Westline Railroad, that same point being monumented by a 2-inch diameter cast iron pipe; thence, in an easterly direction along said southeasternmost right-of-way line of the Westline Railroad, north 34 degrees 47 minutes 06 seconds east, a distance of

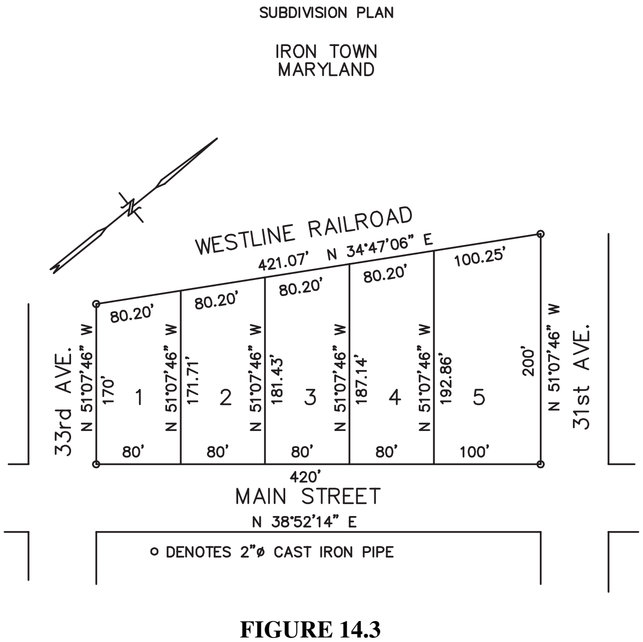
SUBDIVISION PLAN
IRON TOWN
MARYLAND
WESTLINE RAILROAD
421.07' N 34°47'06" E
80.20'
80.20'
80.20'
80.20'
100.25'
33rd AVE.
N 51°07'46" W
170'
1
N 51°07'46" W
171.71'
2
N 51°07'46" W
181.43'
3
N 51°07'46" W
187.14'
4
N 51°07'46" W
192.86'
5
200'
N 51°07'46" W
31st AVE.
80'
80'
80'
80'
100'
420'
MAIN STREET
N 38°52'14" E
o DENOTES 2"ø CAST IRON PIPE

FIGURE 14.3

80.20 feet, to the line between lots 1 and 2; thence, in a southeasterly direction along said line between lots 1 and 2, south 51 degrees 07 minutes 47 seconds east, a distance of 175.71 feet, to the northwesternmost right-of-way line of Main Street; thence, in a southwesterly direction along said southeasternmost right-of-way line of Main Street, south 38 degrees 52 minutes 14 seconds west, a distance of 80 feet, to the point-of-beginning; all as more fully described on the plan of subdivision for Iron Town subdivision by S. Blake, county surveyor, dated May 3, 1920, recorded in book 5, page 2 of the Bay County conveyance office.

The description of Lot 3 might read as follows:

That parcel of portion of land designated as lot 3, square B, Iron Town subdivision, Bay County, Maryland, more particularly described as follows:

Commencing at the intersection of the easternmost right-of-way line of 33rd Avenue and the northernmost right-of-way line of Main Street, that same point being monumented by a 2-inch diameter cast iron pipe; thence, in an easterly direction along said northernmost right-of-way line of Main Street, north 38 degrees 52 minutes 14 seconds east, a distance of 160 feet, to the line between lots 2 and 3; thence, in a northerly direction along said line between lots 2 and 3, north 51 degrees 07 minutes 47 seconds west, a distance of 181.43 feet, to the southernmost right-of-way line of the Westline Railroad; thence, in an easterly direction along said southernmost right-of-way line of the Westline Railroad, north 34 degrees 47 minutes 06 seconds east, a distance of 80.20 feet, to the line between lots 3 and 4; thence, in a southerly direction along said line between lots 3 and 4, south 51 degrees 07 minutes 47 seconds east, a distance of 187.14 feet, to the northernmost right-of-way line of Main Street; thence, in a westerly direction along said northernmost right-of-way line of Main Street, south 38 degrees 52 minutes 14 seconds west, a distance of 80 feet, to the point-of-beginning; all as more fully described on the plan of subdivision of Iron Town subdivision, by S. Blake, county surveyor, dated May 3, 1920, and recorded in the Bay County conveyance office in book 5, page 2.

14.3. CASE 3

T. R. Grubb wants to give his son, R. T. Grubb, a portion of his farm property in Hardtack, Georgia. Figure 14.4 is a drawing, made by the senior Grubb, of the property to be conveyed. The Grubbs refuse to have a survey made on the property, because “Dem survey folks don’t know nuttin’. We gotta move dere markers every time dey come by.”

Write a description of the parcel conveyed.

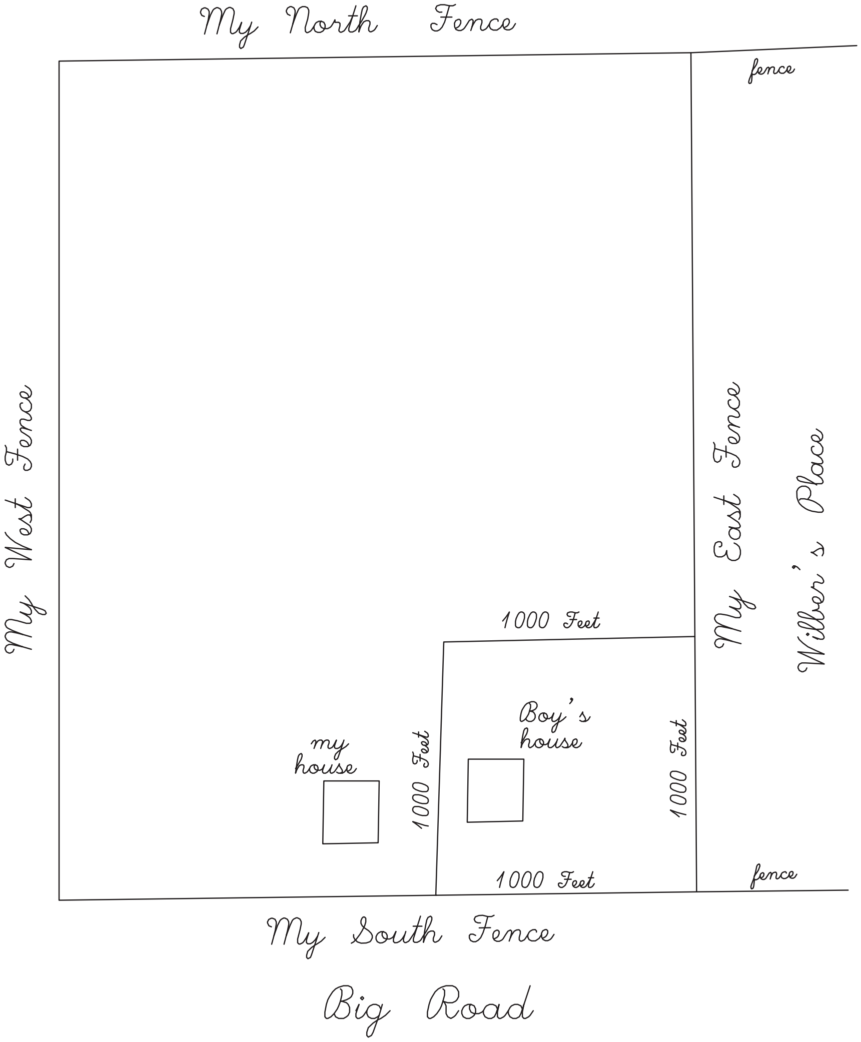
My North Fence
fence
My West Fence
My East Fence
Wilber's Place
1000 Feet
Boy's house
my house
1000 Feet
1000 Feet
1000 Feet
fence
My South Fence
Big Road

FIGURE 14.4

Perhaps a description of the son's parcel would be as follows:

That portion of land located in Hardtack, Hardtack County, Georgia, bounded on the south by Big Road, on the east by the property of T. T. Wilber, and on the north and west by the property of T. R. Grubb, and more particularly described as follows:

Commencing and point-of-beginning at the intersection of the westernmost fence line enclosing the property of T. T. Wilber and the fence line along

> the northernmost side of Big Road; thence, in a westerly direction along said northernmost fence line of Big Road, a distance of 1,000 feet, more or less, to an intersecting fence line enclosing the property herein conveyed; thence, in a northerly direction, along said fence line enclosing the property herein conveyed, a distance of 1,000 feet, more or less, to an intersecting fence line enclosing the property herein conveyed; thence, in an easterly direction along said fence line enclosing the property herein conveyed, a distance of 1,000 feet, more or less, to the westernmost property line of T. T. Wilber and a fence line; thence, in a southerly direction along said westernmost property line of T. T. Wilber, a distance of 1,000 feet, to the point-of-beginning; and all as more fully described on a sketch by T. R. Grubb, hereto attached.

This last exercise demonstrates that the absence of a recent land title survey and a proper subdivision plan does not prevent the creation of deed description that almost appears adequate. Unfortunately, the distances are doubtful, the angles at the corners are unknown, the corners are poorly monumented, the lines are dependent on possibly meandering fences, and the area is unknown. Pity the surveyor who, after 40 years and after the area has been cleared, is charged with the recovery of this particular parcel of land.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

CHAPTER 15

ETHICS AND PROFESSIONAL CONDUCT

The Professional Land Surveyor's registration boards of most states have published codes of ethical conduct. Every Land Surveyor should obtain and be familiar with the codes of conduct published by each of the states where he or she is registered. Procedural rules and regulations along with certification requirements are generally similar for all states. The most common mandates are:

1. The Land Surveyor must be in responsible charge and perform adequate oversight and review of any work certified by him or her. (Certification of work done by others and not under his or her direct supervision is strictly prohibited).
2. The Land Surveyor shall avoid all conflicts of interest, fraud, misrepresentation, and all other dishonest activities.
3. The Professional Land Surveyor is required to maintain a high degree of competency in the practice of his or her profession. (Many states have adapted rules and regulations for required continuing education to maintain licensure as a Land Surveyor.)
4. The Land Surveyor must assure that all employees are well trained and properly supervised.
5. The Land Surveyor must exercise care and sound judgment in the exercise of professional services without regard to compensation. (Charity work must be afforded the same care and

professionalism as compensated projects. Contracts that have been underestimated cannot be "short cut" to meet the budget.)

6. In the event that blunders, discrepancies, or significant mistakes are discovered by the Land Surveyor, even if it involves his or her own work, the Land Surveyor shall immediately give notice of the situation to his or her client and/or others that may be impacted. The Land Surveyor shall conduct such corrective acts as may be required to mitigate the injury to others.
7. The Land Surveyor shall not undertake projects beyond his or her area of expertise.
8. The Land Surveyor shall not undertake projects beyond his or her resources.

Not all of the ethical bounds for the Land Surveyor are as obvious as the few preceding examples. Some exceptions follow.

15.1. CONFIDENTIALITY

The Land Surveyor should maintain the confidential nature of the Land Surveyor-client relationship. While not as binding as the lawyer-client or doctor-patient confidentiality, the Land Surveyor-client relationship is such that what is discovered during a survey performed for a client may be the property of the Land Surveyor, but it may not be disseminated without the permission of the client. The nature and volatility of real estate activity is such that property rights, land values, and business contracts can be adversely impacted if information provided to a landowner or financier were made public.

Of course, the knowledge of real property boundary locations and other conditions discovered by the Land Surveyor during any survey will be melded into the body of his or her experience and expertise and drawn on during other work.

15.2. PROFESSIONAL COURTESY

When requested, the Land Surveyor should provide such information as he or she may have pertaining to the location of real property boundaries to other Land Surveyors working in the vicinity. The greatest single cause of boundary disputes is the failure to obtain all of the information available concerning boundary locations. It is very rare for two Land

Surveyors, working from the same complete set of data, to arrive at different opinions as to the boundary locations. Land Surveyors who readily share information with others practicing in the area minimize the occurrences of boundary disputes.

15.3. IMPARTIAL EVALUATION

The Land Surveyor shall evaluate the evidence of boundary locations and other pertinent information without prejudice toward or against any party. The Land Surveyor is not to distort information or selectively consider facts with the intent to increase benefit to his or her client. Boundary decisions must be based on an impartial consideration of the evidence without regard to the identity of the landowners.

This impartiality extends to the witness stand. When employed as an expert witness, the Land Surveyor is expected to render his or her opinion based on the facts as he or she has developed them without consideration of the effects on the litigants. When employed as an expert, the Land Surveyor should avoid the concept of “their side” and “our side”.

15.4. PROMOTE PROFESSIONALISM

The Land Surveyor is expected to exercise professional conduct at all times, both in the practice of the profession and in personal behavior. The Land Surveyor shall not participate in the denigration of another licensed Land Surveyor, nor attempt to supplant another Land Surveyor in a business relationship.

The Land Surveyor shall report any unprofessional activity or violations of the law, rules of the board, or rules of conduct to the appropriate authority.

The Land Surveyor shall council, tutor, or assist other Land Surveyors seeking to improve technical and professional skills.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

APPENDIX OF TABLES

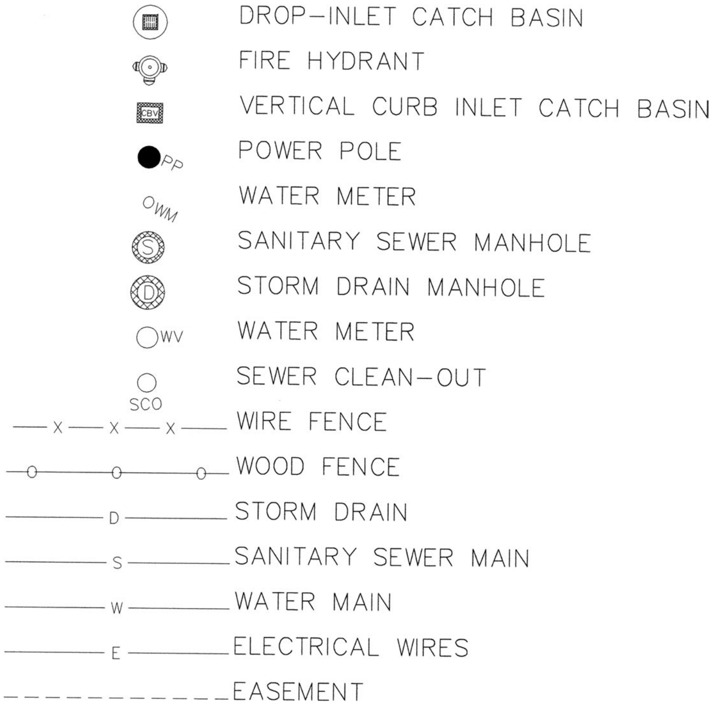
DROP–INLET CATCH BASIN
FIRE HYDRANT
CBV
VERTICAL CURB INLET CATCH BASIN
PP
POWER POLE
WM
WATER METER
S
SANITARY SEWER MANHOLE
D
STORM DRAIN MANHOLE
WV
WATER METER
SCO
SEWER CLEAN–OUT
X X X
WIRE FENCE
WOOD FENCE
D
STORM DRAIN
S
SANITARY SEWER MAIN
W
WATER MAIN
E
ELECTRICAL WIRES
EASEMENT

FIGURE A.1 Typical Map Symbols

TABLE A.1 Select Units of Linear and Square Measure

1 U.S. survey foot	= $\frac{1200}{3937}$ meters
1 U.S. standard foot	= 0.3048 meters
1 line	= $\frac{1}{8}$ inch[a]
1 foot	= 12 inches
1 yard	= 3 feet
1 rod, pole, perch	= $16\frac{1}{2}$ feet[b]
1 chain	= 66 feet
1 chain	= 100 links
1 chain	= 4 rods, poles, perches
1 mile	= 80 chains
1 mile	= 5,280 feet
1 nautical mile	= 6,080.27 feet[c]
1 rood	= $\frac{1}{4}$acre
1 acre	= 43,560 square feet
1 acre	= 10 square chains
1 square mile	= 640 acres
1 compass, row	= 6 feet[a]
1 French foot	= 1/180 arpent
1 toise	= $\frac{1}{30}$arpent
1 toise	= 6.3944 feet[a]
1 arpent	= 191.994 feet (GLO definition)
1 arpent	= 191.83 feet
1 arpent square measure	= 0.8448 acres[a]
1 vara	= $33\frac{1}{3}$ inches[a]
1 hand	= 4 inches
1 span	= 9 inches
1 cubit	= 18 inches
1 step	= $2\frac{1}{2}$ feet[a]
1 pace	= 5 feet[a]
1 fathom	= 6 feet
1 furlong	= 10 chains
1 league	= 3 miles[a]

[a] These units may vary greatly, depending on local custom.

[b] Feet listed in this table are U.S. survey feet, unless otherwise noted.

[c] This distance is a function of the spheroid in use and will vary. Some report this distance as 6,076.10 feet.

TABLE A.2 Common Angular Units of Measure

1 circle	= 360 degrees
1 circle	= 400 grads
1 circle	= 32 points
1 circle	= 6,400 mils
1 degree	= 60 minutes
1 minute	= 60 seconds
1 degree of arc	= 17.45 feet in 1,000 feet[a]
1 minute of arc	= 0.29 feet in 1,000 feet[a]
1 second of arc	= 0.005 feet in 1,000 feet[a]
1 second of latitude at the equator	= 101 feet[a]
1 minute of latitude at the equator	= 6,046 feet[a]

[a] Approximate distance only.

TABLE A.3 Accuracy Specifications and Positional Tolerances

	A	B	C	D	
Condition	Urban	Suburban	Rural	Mountain or Marshland	
Unadjusted closure (maximum allowable)	1:15,000	1:10,000	1:7,500	1:5,000	Traverse loop or between control monuments
Angular closure (maximum allowable)	10″ N	15″ N	25″ N	30″ N	N = number of angles in traverse
Accuracy of bearing in relation to source (maximum allowable)	± 15 sec.	± 20 sec.	± 30 sec.	± 40 sec.	1/sine angle = denominator in error of closure (approximately)
Linear distances accurate to: (maximum allowable)	± 0.05 ft. per 1,000 ft.	± 0.1 ft. per 1,000 ft.	± 0.15 ft. per 1,000 ft.	± 0.2 ft. per 1,000 ft.	Sine angle × 1,000 (approximately) where = accuracy of bearing
Positional error of any monument (maximum)	.03′ + AC/15,000	.05′ + AC/10,000	.07′ + AC/7,500	.1′ + AC/5,000	AC = length of any course
Calculation of area—accurate and carried to nearest——decimal place of acre	.0001	.0001	.001	.001	To 1 acre
	.001	.001	.01	.01	To 10 acres
	.01	.01	.1	.1	To 100 acres
	.1	.1	.2	.3	To 1,000 acres
Elevations for boundaries controlled by tides, contours, rivers, etc., accurate to:	± .03 ft.	± .05 ft.	± 0.1 ft.	± 0.2 ft.	Based on accepted local datum

(*Continued*)

TABLE A.3 (*Continued*)

Condition	A Urban	B Suburban	C Rural	D Mountain or Marshland	
Location of improvements, structures, paving, etc. (tie measurements)	± 0.1 ft.	± 0.2 ft.	± 0.5 ft.	± 1 ft.	
Scale of maps sufficient to show detail but no less than	1″ = 200’	1″ = 400’	1″ = 1,000’	1″ = 2,000’	
Positional error in map plotting not to exceed: (applies to original map only)	5 ft.	10 ft.	25 ft.	50 ft.	Generally $\frac{1}{40}$th inch; national map accuracy calls for $\frac{1}{50}$th inch
	1″ = 200’	1″ = 400’	1″ = 1,000’	1″ = 2,000′	
Adjusted mathematical closure of survey (minimum)	0.03′ or 1:50,000	0.05′ or 1:50,000	0.07′ or 1:50,000	0.1′ or 1:50,000	Larger value allowed

TABLE A.4 Typical Instrument Precisions

Type of Instrument (Typical)	Smallest Unit Directly Read	Direct Read Uncertainty in 1,000 Feet
Marine compass	$11\frac{1}{4}$ degrees	196.0 feet
Surveyor's compass	15 minutes	4.4 feet
Builder's transit	1 minute	0.29 feet
Mountain transit	30 seconds	0.15 feet
Surveyor's theodolite	10 seconds	0.05 feet[a]
Control survey theodolite	1 second	0.005 feet[a]
Electronic distance meters	0.01 feet	0.03 feet[a]

[a] The errors involved with pointing and centering of target and instrument are much greater than the minimum direct read of the instrument itself.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

GLOSSARY OF SELECT TERMS

(These definitions are condensed)

Abstract of title (or deed) A summary of conveyances, exchanges, easements, or other legal instruments affecting the property rights of a particular parcel.

Accessories to a corner Natural physical objects, such as trees, rock formations, ledges, and other features, that are at a known distance and/or direction from a corner. Accessories are part of the corner monumentation.

Adjoiner That parcel of land that shares a common boundary with another.

Adjusted value The value assigned to a measured quantity after the application of corrections designed to account for measurement errors.

Aliquot parts A division of a USPLS section following a specified procedure.

Angle A measure of the relationship of two intersecting lines.

Area The measure of the bounded surface formed by the intersection of real property boundaries and a particular vertical datum.

Astronomic Values assigned to direction or position based on measurements made of the relative positions of heavenly bodies.

Azimuth A definition of the direction of a line based on the clockwise angle formed between that line and a certain pole of a meridian.

Balancing a traverse A procedure, or procedures, for distributing the accumulated measurement errors of a traverse among the observed values in order to obtain computational consistency.

Baseline A series of points established for the expressed purpose of locating other features or lines.

Bearing A definition of the direction of a line based on the clockwise or counterclockwise angle formed between that line and either pole of a meridian.

Benchmark An object, natural or artificial, in a relatively stable location, that is at a known elevation relative to a particular vertical datum.

Blunder A mistake or an incorrect assessment of a measured value associated with a gross misinterpretation of facts.

Boundary line An imaginary line of demarcation between two adjoining land parcels, distinguishing a separation of real property rights, which may or may not be physically marked.

Bureau of Land Management (BLM) An agency of the U.S. government, formerly known as the Government Land Office (GLO), which is responsible for the survey of public lands, among other things.

Cadastre An official map of a political region delineating size, location, ownership, and land values for the purpose of assessing taxes.

Center of a section A point in a USPLS section determined by the intersection of two straight lines drawn between opposing quarter corners.

Chain (1) A length of measure equal to 66 U.S. survey feet; (2) a surveyor's measuring tape; (3) the act of measuring a linear distance.

Chord A straight line drawn between the ends of a curved line segment.

Compass (1) A device for detecting the earth's magnetic field and aligning with the lines of force; (2) a device for laying out a specific distance or marking an arc; (3) a unit of measure, usually defined as 6 feet.

Contour A series of lines on a map connecting points of equal elevation.

Control A series of vertical and/or horizontal survey marks, data, measurements, maps, photographs, or other acts specifically performed to serve as the foundation or datum for future surveys.

Coordinate system A method of identifying a particular point in two or three dimensions by a systematic listing of the distances from defined baselines or origins.

Corner A point of intersection of real property boundary lines, which may or may not be monumented.

Course The direction of a line segment. In some states, the course also includes the length of a line segment.

Datum A basis or measurement foundation on which a location can be defined or referenced either vertically, horizontally, or both.

Deed A written instrument that conveys rights or interests in real property.

Deed description That part of a deed that identifies and describes the relevant real property parcel.

Departure The change in the distance from a meridian experienced in moving from one end of a line segment to another.

Easement A right held by one party to the land of another.

Electronic distance meter (EDM) A device that measures distances by the use of electromagnetic radiation.

Elevation The distance above or below a vertical datum.

Equator (1) Zero degrees latitude; (2) an imaginary plane passing through the earth and perpendicular to the axis of rotation.

Error The difference between the measured and actual values for a certain quantity.

Error of closure The failure of the result of a set of measured values to agree mathematically with the theoretical result.

Fee simple Unlimited rights of real property ownership restricted only by the laws of the United States, state, and local governmental jurisdictions.

Field notes The written notes, sketches, and computations of a surveyor taken during and at the site of a survey.

Geoid A theoretical surface that is everywhere perpendicular to plumb at every point.

GIS (Geographical Information System) A database and information recovery system indexed by location.

Government Land Office See Bureau of Land Management (BLM).

GPS (Global Positioning System) A network of radio transmitting navigational satellites and specialized receivers used to determine the geodetic location of the receiver.

Grade The slope of a surface or structure.

Grid An imaginary network of evenly spaced parallel and perpendicular two-dimensional lines.

Horizontal A plane, perpendicular to the plumb line at a particular point.

Land description The exact location of a real property parcel in terms of the controlling land record system.

Legal description The description of a real property parcel sufficient to identify that parcel uniquely without oral testimony.

Latitude (1) The distance along a meridian; (2) the north or south change in distance experienced in moving from one end of a line segment to another.

Level (1) A surface that is everywhere perpendicular to plumb; (2) an instrument for measuring differences in elevation.

Line A series of contiguous points forming a vertical geometric surface and extending from the center of the earth up through the land surface and into space.

LIS (Land Information System) A database of information related to the land.

Longitude The distance between two meridians.

Map A graphic, two-dimensional representation of the surface of the earth.

Map projection A systematic method, accounting for the curvature of the earth, that mathematically reduces surface location information into two-dimensional data.

Mean sea level (MSL) The average elevation of the sea over a 19-year period. MSL is often confused with the North American Vertical Datum (NAVD).

Measurement An estimation of a quantity or a distance based on the systematic application of a standardized procedure or device.

Metes and bounds description A description formed by sequentially reciting the courses and adjoiners of a real property parcel.

Meridian A north–south line used to reference lines of a survey.

Monument The physical object that indicates the location of a point, station, or real property corner.

More or less A phrase indicating a crude or uncertain value for a quantity.

National Geodetic Survey (NGS) The agency of the U.S. government that is responsible for development and maintenance of benchmarks and stations for navigation and mapping.

National Geodetic Vertical Datum, 1927 (NGVD) A vertical datum established by the National Geodetic Survey in 1927 that defined elevations published for use on federal maps and regulations, now obsolete.

North Aligned with the axis of the earth's rotation and in the direction of that particular pole designated as "north."

North American Datum 1983 (NAD 83) The horizontal datum established by the National Geodetic Survey to define geodetic positioning for navigational and mapping purposes.

North American Vertical Datum, 1988 (NAVD 88) The vertical datum established by the National Geodetic Survey that defines elevations published for use on federal maps and regulations and replaced NGVD.

Observation The noting of a condition, which usually is the result of an act of measurement.

Plane A flat surface such that the shortest route between any two locations on the surface is entirely contained within the surface.

Plat A map, prepared by a land surveyor, usually for a specific legal purpose.

Plumb Aligned with the pull of gravity.

Platted subdivision description A description based on a map or plan, usually recorded, identifying a real property parcel by the letter or number designation found on that map or plan.

Point (1) A specific location; (2) a vertical geometric ray (line) originating at the center of the earth and extending up through the surface into space.

Point-of-Beginning The first point encountered in the narrative portion of a deed description that is a part of the real property boundary itself.

Point-of-Commencement The first point described in the narrative portion of a deed description that is well known and used to direct the reader to the location of the real property. The commencement point may not be a point on the real property boundary being described.

Point of compound curvature A point on a circular curved line at which a curve of one radius length ends and a curve of another radius length begins, and occurring at a point on the extension of a straight line drawn between the two radius centers.

Point of curvature A point at which a straight line begins a circular curve and is at right angles to the radius of the curve at that point.

Point of reverse curvature A point on a curved line, occurring on a line drawn between the radius centers, at which a curve in one direction ends and a curve in the opposite direction begins.

Pole *See* Rod.

Quarter corner That corner, set by the government surveyor, between section corners.

Random error Incidental errors occurring as a result of observational imprecision.

Random traverse A traverse in which the location of stations is chosen for accessibility and inter-visibility and does not have a constant relationship to any real property boundaries.

Range line A north–south line used to divide public lands.

Recovered corner A real property corner that has been verified by the discovery of the original monument, accessories, or other physical evidence.

Right-of-way Land granted (usually to the governing authority) by deed, servitude, or easement for the construction of an infrastructure. Rights-of-way may grant limited property rights or full property rights.

Riparian Pertaining to the banks of a body of water.

Rod (1) $16^1/_2$ feet; (2) a wooden pole (rod) being 16.5 feet long and used to measure horizontal distances.

Section The smallest division of land monumented by the U.S. government public lands for sale.

Servitude *See* Easement.

Spiral curve A curved line of constantly varying radius.

Station A point established by survey procedures used to determine the location of other features.

Straight line A vertical plane containing the center of the earth. A line established by the line of sight between two points.

Surface The separation of two distinct spaces. The interface between the earth and the atmosphere.

Land surveying The art and science of measuring, marking, recovering, and mapping the relative positions or locations of terrain features and real property boundaries.

Systematic error Errors occurring consistently, regularly, and of the same algebraic sign as the result of a measurement condition.

Theodolite An instrument designed to measure precisely vertical and horizontal angles.

Title The exclusive right to the use and enjoyment of a particular parcel of real property.

Township A division of public lands generally 6 miles wide and 6 miles long and containing 36 sections.

Township line An east–west line used to divide public lands.

Transit In the United States, a transit is a theodolite having a vernier read scale for measuring horizontal and vertical angles and having a scope that is capable of being inverted.

Traverse A systematic series of stations in which the direction and length of line segments formed by consecutive stations is measured.

Vernier An etched ruler or scale that is marked such that, when it is aligned with another ruler or scale, divisions much smaller than marked on either scale can be read directly.

Vertical Aligned with the pull of gravity.

Witness mark A physical feature that is at a known distance and direction from a corner.

A GUIDE TO UNDERSTANDING LAND SURVEYS
STEPHEN V. ESTOPINAL
Copyright © 2009 by John Wiley & Sons,Inc

INDEX

A

Abstractors, 162
Accessories, to a corner, 141
Accuracy, 56
  determining standard of, 168
ACSM, 156
ALTA, 156
Angles:
  clockwise, 28
  degrees, minutes, and seconds, 29
  geometric, 20
  horizontal, 27
  latitude, 13, 16
  longitude, 13, 16
  vertical, 29
Area, land, 27
Astronomic position, 14
Azimuth, definition of, 37

B

Bearing, definition of, 37
Benchmark, 52, 85
Boundary, 8
  corner, 17
  metes and bounds recovery, 144
  platted subdivision recovery, 145
  real property, 8
  recovery, 135
  USPLS recovery, 145

C

Call:
  hierarchy of, 93
  passing, 142
Cartography, 14
Chain:
  Gunter's, 68
  modern, 71
Control, traverse, 80
Coordinates, state plane, 48
Corner:
  by best evidence, 144
  by common report, 143
  by proportionate measurement, 144
  perpetuated, 143

Corner (*Continued*)
  quarter, 108
  section, 108
Corners, recovered, 141
Correction:
  compass rule, 73
  Crandall's rule, 77
  field, 66
  least squares, 77
  transit rule, 77

D

Datum:
  mean sea level (MSL), 27
  NAVD, 27
  NGVD, 27
  North American Datum of 1927 (NAD 27), 132
  North American Datum of 1983 (NAD 83), 132
  vertical, 26
Declination:
  magnetic, 34
  USPLS, 140
Deed, 9
  comparing to plat, 163
Degree, definition of, 29
Description:
  additions and deletions, 217
  caption, 206
  curve, 214
  deed, 204
  key phrases, 208
  legal, 90
  metes and bounds, 91
  narrative, 207
  outline, 205
  references, 218
Deviation:
  magnetic, 34
  USPLS, 140
Direction:
  azimuth, 37
  bearing, 37
Distance:
  geometric, 20
  horizontal, 24
  land, 24
  latitude, 10
  longitude, 12
  standard, 26

E

Easement, 164
EDM, 78
Elevation, 26
  differential, 85
  turning point, 86
Ellipsoid, 14
Encroachment, 165
Equator, 11
Error, 57
  blunder, 58
  random, 58
  systematic, 58
Evaluating:
  age of survey, 155
  purpose of survey, 156
  survey discrepancies, 157
Evidence:
  rules of, 135
  sources of, 136

F

Fences, 165
Foot:
  U.S. standard, 26
  U.S. survey, 26

G

Geodesy, 14
Geodetic position, 14

Geodetic Positioning System, *see* Global Positioning System
Geographic Information System, 131
Geoid, 14
Geometry:
  basic, 19
  land title, 20
  plane, 19
Global Positioning System, 123

H

Hierarchy of calls, 93
Horizontal distance, 24

I

Intent of the parties, 94

L

Land:
  area, 22
  distance, 24
  record systems, 89
  information system, 48
Land record system, determining, 148
Land title surveys, "Minimum Standard Detail Requirements for ALTA/ACSM," 156
Latitude, 10
  astronomic, 14
  geodetic, 14
Length, geometric, 20
Level surface, 24
Line, 19
  baseline, 104
  boundary, 55
  geometric, 19
  land, 24
  latitude, 10
  longitude, 12
  meander, 110
  plumb, 24
  property, 55
  range, 105
  straight, 20
  straight land, 23
  township, 105
Longitude:
  astronomic, 14
  geodetic, 14

M

Magnetic:
  declination, 30
  deviation, 30
  north, 30
Map:
  definition, 19
  projectionless, 43
  projection systems, 43
  state plane, 48
  tangent plane, 44
Measurement:
  accuracy, 56
  fundamentals, 55
  precision, 56
Meridian:
  prime, 12
  principal, 103
Minutes, definition of, 29
Monument:
  artificial, 93
  natural, 93

N

NAVD, 27
NGVD, 27

North:
  arrow, 157
  assumed, 35
  astronomic, 34
  grid, 35
  magnetic, 34
  magnetic declination, 34
  origin, 10
  “true,” 33
Notes, USPLS, 111, 140

O

Ownership, 7

P

Plat:
  bearings or angles, 160
  certification, 161
  checklist, 147
  class of survey (ALTA/ACSM), 168
  comparing to deed, 163
  dimensions, 160
  easements, 164
  encroachments, 165
  evaluation checklist, 147
  fence line, 165
  legal description, 157
  limiting words or phrases, 162
  north arrow, 157
  resubdivision, 204
  scale, 163
Platted subdivision, 118
Plus or minus, 178
Point, 19
  -of-beginning, 92
  commencing, 92
  geometric, 19
  initial, 103
  land, 21
  occupying, 24
  on line, 24
Position:
  astronomic, 14
  geodetic, 14
Precision:
  implied, 56
Prime meridian, 12
Projection systems: 43
  state plane, 48
  tangent plane, 44
Property, Real, 7

Q

Quarter, corner, 108

R

Random Traverse, 80
Riparian boundaries, 110

S

Scale, plat or map, 163
Sea level, datum, 27
Seconds, definition of, 29
Section:
  center, 109
  corner, 109
  ideal, 107
  irregular, 110
  nonaliquot division, 115
  quarter, 108
  USPLS, 102
Spheroid, 14
Standard procedure, 60
Station:
  control, 79
  random, 80
  traverse, 82
Straight line, 23
Survey, boundary, 156

T

Title:
  land, 8
  transfers, 8
Title insurance, 3
Toise, 63
Township:
  ideal, 105
  USPLS, 102
Transit, 69
Traverse:
  balancing, 73
  closing, 73
  compass rule, 73
  Crandall's rule, 77
  least squares adjustment, 77
  random, 80
  station, 80
  transit rule, 77

U

USPLS, 102
  accessories to a corner, 141
  aliquot parts, 108
  baseline, 104
  Division, 109
  general instructions, 111
  initial point, 103
  meander corner, 110
  meander line, 110
  principal meridian, 103
  private claims, 117
  quarter corner, 107
  range, 105
  section, 107
  section corner, 109
  surveyor's notes, 111
  Township, 105, 106
  township plat, 118
  witness trees, 141

V

Vertical, datum, 26

W

Witness trees, USPLS, 141

tler

ectronics
nd
notonics

HANDBOOK OF NANOPHYSICS

6

Nanoelectronics and Nanophotonics

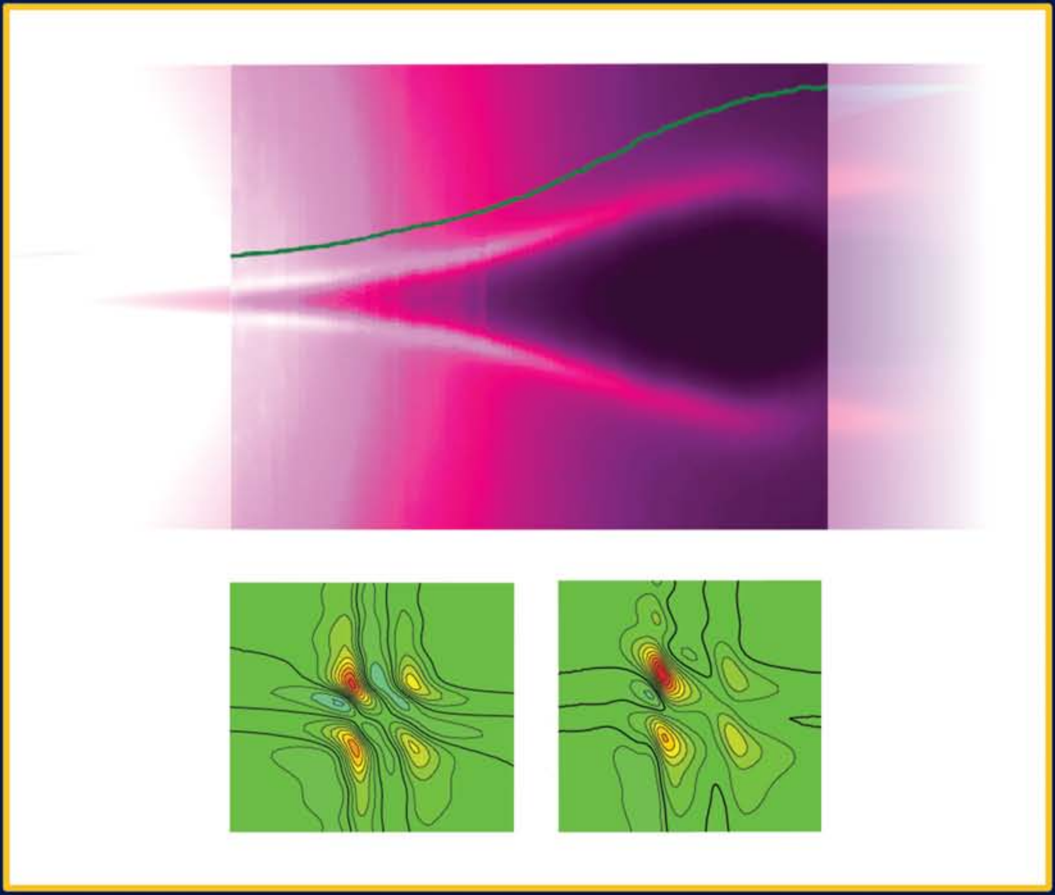

Edited by
Klaus D. Sattler

CRC Press
Taylor & Francis Group

8/4/10 4:30 PM

HANDBOOK OF NANOPHYSICS

6

Handbook of Nanophysics

Handbook of Nanophysics: Principles and Methods

Handbook of Nanophysics: Clusters and Fullerenes

Handbook of Nanophysics: Nanoparticles and Quantum Dots

Handbook of Nanophysics: Nanotubes and Nanowires

Handbook of Nanophysics: Functional Nanomaterials

Handbook of Nanophysics: Nanoelectronics and Nanophotonics

Handbook of Nanophysics: Nanomedicine and Nanorobotics

HANDBOOK OF NANOPHYSICS

6

Nanoelectronics and Nanophotonics

Edited by

Klaus D. Sattler

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2011 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-7550-2 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Handbook of nanophysics. Nanoelectronics and nanophotonics / editor, Klaus D. Sattler.
p. cm.
"A CRC title."
Includes bibliographical references and index.
ISBN 978-1-4200-7550-2 (alk. paper)
1. Nanoelectronics--Handbooks, manuals, etc. 2. Nanophotonics--Handbooks, manuals, etc. I. Sattler, Klaus D. II. Title.

TK7874.84.H36 2010
621.381--dc22 2010001108

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface ix

Acknowledgments xi

Editor xiii

Contributors xv

PART I Computing and Nanoelectronic Devices

1 Quantum Computing in Spin Nanosystems **1**-1
Gabriel González and Michael N. Leuenberger

2 Nanomemories Using Self-Organized Quantum Dots **2**-1
Martin Geller, Andreas Marent, and Dieter Bimberg

3 Carbon Nanotube Memory Elements **3**-1
Vincent Meunier and Bobby G. Sumpter

4 Ferromagnetic Islands **4**-1
Arndt Remhof, Andreas Westphalen, and Hartmut Zabel

5 A Single Nano-Dot Embedded in a Plate Capacitor **5**-1
Gilles Micolau and Damien Deleruyelle

6 Nanometer-Sized Ferroelectric Capacitors **6**-1
Nikolay A. Pertsev, Adrian Petraru, and Hermann Kohlstedt

7 Superconducting Weak Links Made of Carbon Nanostructures **7**-1
Vincent Bouchiat

8 Micromagnetic Modeling of Nanoscale Spin Valves **8**-1
Bruno Azzerboni, Giancarlo Consolo, and Giovanni Finocchio

9 Quantum Spin Tunneling in Molecular Nanomagnets **9**-1
Gabriel González and Michael N. Leuenberger

10 Inelastic Electron Transport through Molecular Junctions **10**-1
Natalya A. Zimbovskaya

11 Bridging Biomolecules with Nanoelectronics **11**-1
Kien Wen Sun and Chia-Ching Chang

PART II Nanoscale Transistors

12 Transistor Structures for Nanoelectronics **12**-1
Jean-Pierre Colinge and Jim Greer

13 Metal Nanolayer-Base Transistor **13**-1
André Avelino Pasa

14 ZnO Nanowire Field-Effect Transistors **14**-1
Woong-Ki Hong, Gunho Jo, Sunghoon Song, Jongsun Maeng, and Takhee Lee

15 C_{60} Field Effect Transistors **15**-1
Akihiro Hashimoto

16 The Cooper-Pair Transistor **16**-1
José Aumentado

PART III Nanolithography

17 Multispacer Patterning: A Technology for the Nano Era **17**-1
Gianfranco Cerofolini, Elisabetta Romano, and Paolo Amato

18 Patterning and Ordering with Nanoimprint Lithography **18**-1
Zhijun Hu and Alain M. Jonas

19 Nanoelectronics Lithography **19**-1
Stephen Knight, Vivek M. Prabhu, John H. Burnett, James Alexander Liddle, Christopher L. Soles, and Alain C. Diebold

20 Extreme Ultraviolet Lithography **20**-1
Obert R. Wood II

PART IV Optics of Nanomaterials

21 Cathodoluminescence of Nanomaterials **21**-1
Naoki Yamamoto

22 Optical Spectroscopy of Nanomaterials **22**-1
Yoshihiko Kanemitsu

23 Nanoscale Excitons and Semiconductor Quantum Dots **23**-1
Vanessa M. Huxter, Jun He, and Gregory D. Scholes

24 Optical Properties of Metal Clusters and Nanoparticles **24**-1
Emmanuel Cottancin, Michel Broyer, Jean Lermé, and Michel Pellarin

25 Photoluminescence from Silicon Nanostructures **25**-1
Amir Sa'ar

26 Polarization-Sensitive Nanowire and Nanorod Optics **26**-1
Harry E. Ruda and Alexander Shik

27 Nonlinear Optics with Clusters **27**-1
Sabyasachi Sen and Swapan Chakrabarti

28 Second-Harmonic Generation in Metal Nanostructures **28**-1
Marco Finazzi, Giulio Cerullo, and Lamberto Duò

29 Nonlinear Optics in Semiconductor Nanostructures **29**-1
Mikhail Erementchouk and Michael N. Leuenberger

30 Light Scattering from Nanofibers **30**-1
Vladimir G. Bordo

31 Biomimetics: Photonic Nanostructures **31**-1
Andrew R. Parker

PART V Nanophotonic Devices

32 Photon Localization at the Nanoscale **32**-1
Kiyoshi Kobayashi

33 Operations in Nanophotonics **33**-1
Suguru Sangu and Kiyoshi Kobayashi

34 System Architectures for Nanophotonics **34**-1
Makoto Naruse

35 Nanophotonics for Device Operation and Fabrication **35**-1
Tadashi Kawazoe and Motoichi Ohtsu

36 Nanophotonic Device Materials **36**-1
Takashi Yatsui and Wataru Nomura

37 Waveguides for Nanophotonics **37**-1
Jan Valenta, Tomáš Ostatnický, and Ivan Pelant

38 Biomolecular Neuronet Devices **38**-1
Grigory E. Adamov and Evgeny P. Grebennikov

PART VI Nanoscale Lasers

39 Nanolasers **39**-1
Marek S. Wartak

40 Quantum Dot Laser **40**-1
Frank Jahnke

41 Mode-Locked Quantum-Dot Lasers **41**-1
Maria A. Cataluna and Edik U. Rafailov

Index **Index**-1

Preface

The *Handbook of Nanophysics* is the first comprehensive reference to consider both fundamental and applied aspects of nanophysics. As a unique feature of this work, we requested contributions to be submitted in a tutorial style, which means that state-of-the-art scientific content is enriched with fundamental equations and illustrations in order to facilitate wider access to the material. In this way, the handbook should be of value to a broad readership, from scientifically interested general readers to students and professionals in materials science, solid-state physics, electrical engineering, mechanical engineering, computer science, chemistry, pharmaceutical science, biotechnology, molecular biology, biomedicine, metallurgy, and environmental engineering.

What Is Nanophysics?

Modern physical methods whose fundamentals are developed in physics laboratories have become critically important in nanoscience. Nanophysics brings together multiple disciplines, using theoretical and experimental methods to determine the physical properties of materials in the nanoscale size range (measured by millionths of a millimeter). Interesting properties include the structural, electronic, optical, and thermal behavior of nanomaterials; electrical and thermal conductivity; the forces between nanoscale objects; and the transition between classical and quantum behavior. Nanophysics has now become an independent branch of physics, simultaneously expanding into many new areas and playing a vital role in fields that were once the domain of engineering, chemical, or life sciences.

This handbook was initiated based on the idea that breakthroughs in nanotechnology require a firm grounding in the principles of nanophysics. It is intended to fulfill a dual purpose. On the one hand, it is designed to give an introduction to established fundamentals in the field of nanophysics. On the other hand, it leads the reader to the most significant recent developments in research. It provides a broad and in-depth coverage of the physics of nanoscale materials and applications. In each chapter, the aim is to offer a didactic treatment of the physics underlying the applications alongside detailed experimental results, rather than focusing on particular applications themselves.

The handbook also encourages communication across borders, aiming to connect scientists with disparate interests to begin interdisciplinary projects and incorporate the theory and methodology of other fields into their work. It is intended for readers from diverse backgrounds, from math and physics to chemistry, biology, and engineering.

The introduction to each chapter should be comprehensible to general readers. However, further reading may require familiarity with basic classical, atomic, and quantum physics. For students, there is no getting around the mathematical background necessary to learn nanophysics. You should know calculus, how to solve ordinary and partial differential equations, and have some exposure to matrices/linear algebra, complex variables, and vectors.

External Review

All chapters were extensively peer reviewed by senior scientists working in nanophysics and related areas of nanoscience. Specialists reviewed the scientific content and nonspecialists ensured that the contributions were at an appropriate technical level. For example, a physicist may have been asked to review a chapter on a biological application and a biochemist to review one on nanoelectronics.

Organization

The *Handbook of Nanophysics* consists of seven books. Chapters in the first four books (*Principles and Methods*, *Clusters and Fullerenes*, *Nanoparticles and Quantum Dots*, and *Nanotubes and Nanowires*) describe theory and methods as well as the fundamental physics of nanoscale materials and structures. Although some topics may appear somewhat specialized, they have been included given their potential to lead to better technologies. The last three books (*Functional Nanomaterials*, *Nanoelectronics and Nanophotonics*, and *Nanomedicine and Nanorobotics*) deal with the technological applications of nanophysics. The chapters are written by authors from various fields of nanoscience in order to encourage new ideas for future fundamental research.

After the first book, which covers the general principles of theory and measurements of nanoscale systems, the organization roughly follows the historical development of nanoscience. *Cluster* scientists pioneered the field in the 1980s, followed by extensive work on *fullerenes*, *nanoparticles*, and *quantum dots* in the 1990s. Research on *nanotubes* and *nanowires* intensified in subsequent years. After much basic research, the interest in applications such as the *functions of nanomaterials* has grown. Many bottom-up and top-down techniques for nanomaterial and nanostructure generation were developed and made possible the development of *nanoelectronics* and *nanophotonics*. In recent years, real applications for *nanomedicine* and *nanorobotics* have been discovered.

MATLAB® is a registered trademark of The MathWorks, Inc. For product information, Please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA, 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Acknowledgments

Many people have contributed to this book. I would like to thank the authors whose research results and ideas are presented here. I am indebted to them for many fruitful and stimulating discussions. I would also like to thank individuals and publishers who have allowed the reproduction of their figures. For their critical reading, suggestions, and constructive criticism, I thank the referees. Many people have shared their expertise and have commented on the manuscript at various stages. I consider myself very fortunate to have been supported by Luna Han, senior editor of the Taylor & Francis Group, in the setup and progress of this work. I am also grateful to Jessica Vakili, Jill Jurgensen, Joette Lynch, and Glenon Butler for their patience and skill with handling technical issues related to publication. Finally, I would like to thank the many unnamed editorial and production staff members of Taylor & Francis for their expert work.

Klaus D. Sattler
Honolulu, Hawaii

Editor

Klaus D. Sattler pursued his undergraduate and master's courses at the University of Karlsruhe in Germany. He received his PhD under the guidance of Professors G. Busch and H.C. Siegmann at the Swiss Federal Institute of Technology (ETH) in Zurich, where he was among the first to study spin-polarized photoelectron emission. In 1976, he began a group for atomic cluster research at the University of Konstanz in Germany, where he built the first source for atomic clusters and led his team to pioneering discoveries such as "magic numbers" and "Coulomb explosion." He was at the University of California, Berkeley, for three years as a Heisenberg Fellow, where he initiated the first studies of atomic clusters on surfaces with a scanning tunneling microscope.

Dr. Sattler accepted a position as professor of physics at the University of Hawaii, Honolulu, in 1988. There, he initiated a research group for nanophysics, which, using scanning probe microscopy, obtained the first atomic-scale images of carbon nanotubes directly confirming the graphene network. In 1994, his group produced the first carbon nanocones. He has also studied the formation of polycyclic aromatic hydrocarbons (PAHs) and nanoparticles in hydrocarbon flames in collaboration with ETH Zurich. Other research has involved the nanopatterning of nanoparticle films, charge density waves on rotated graphene sheets, band gap studies of quantum dots, and graphene foldings. His current work focuses on novel nanomaterials and solar photocatalysis with nanoparticles for the purification of water.

Among his many accomplishments, Dr. Sattler was awarded the prestigious Walter Schottky Prize from the German Physical Society in 1983. At the University of Hawaii, he teaches courses in general physics, solid-state physics, and quantum mechanics.

In his private time, he has worked as a musical director at an avant-garde theater in Zurich, composed music for theatrical plays, and conducted several critically acclaimed musicals. He has also studied the philosophy of Vedanta. He loves to play the piano (classical, rock, and jazz) and enjoys spending time at the ocean, and with his family.

Contributors

Grigory E. Adamov
Open Joint-Stock Company
Central Scientific Research Institute of Technology "Technomash"
Moscow, Russia

Paolo Amato
Numonyx

and

Department of Materials Science
University of Milano-Bicocca
Milano, Italy

José Aumentado
National Institute of Standards and Technology
Boulder, Colorado

Bruno Azzerboni
Department of Matter Physics and Electronic Engineering
Faculty of Engineering
University of Messina
Messina, Italy

Dieter Bimberg
Institut für Festkörperphysik
Technische Universität Berlin
Berlin, Germany

Vladimir G. Bordo
A.M. Prokhorov General Physics Institute
Russian Academy of Sciences
Moscow, Russia

Vincent Bouchiat
Nanosciences Department
Centre National de la Recherche Scientifique
Néel-Institut
Grenoble, France

Michel Broyer
Laboratoire de Spectrométrie Ionique et Moléculaire
Centre National de la Recherche Scientifique
Université de Lyon
Villeurbanne, France

John H. Burnett
Atomic Physics Division
National Institute of Standards and Technology
Gaithersburg, Maryland

Maria A. Cataluna
Division of Electronic Engineering and Physics
School of Engineering, Physics and Mathematics
University of Dundee
Dundee, United Kingdom

Gianfranco Cerofolini
Department of Materials Science
University of Milano-Bicocca
Milano, Italy

Giulio Cerullo
Dipartimento di Fisica
Politecnico di Milano
Milano, Italy

Swapan Chakrabarti
Department of Chemistry
University of Calcutta
Kolkata, West Bengal, India

Chia-Ching Chang
Department of Biological Science and Technology
National Chiao Tung University
Hsinchu, Taiwan

and

Institute of Physics
Academia Sinica
Taipei, Taiwan

Jean-Pierre Colinge
Tyndall National Institute
University College Cork
Cork, Ireland

Giancarlo Consolo
Department of Matter Physics and Electronic Engineering
Faculty of Engineering
University of Messina
Messina, Italy

Emmanuel Cottancin
Laboratoire de Spectrométrie Ionique et Moléculaire
Centre National de la Recherche Scientifique
Université de Lyon
Villeurbanne, France

Damien Deleruyelle
Centre National de la Recherche Scientifique
Institut Matériaux Microélectronique Nanosciences de Provence
Universités d'Aix Marseille
Marseille, France

Alain C. Diebold
College of Nanoscale Science and Engineering
University at Albany
Albany, New York

Lamberto Duò
Dipartimento di Fisica
Politecnico di Milano
Milano, Italy

Mikhail Erementchouk
NanoScience Technology Center

and

Department of Physics
University of Central Florida
Orlando, Florida

Marco Finazzi
Dipartimento di Fisica
Politecnico di Milano
Milano, Italy

Giovanni Finocchio
Department of Matter Physics
and Electronic Engineering
Faculty of Engineering
University of Messina
Messina, Italy

Martin Geller
Experimental Physics

and

Center for Nanointegration
Duisburg-Essen
University of Duisburg-Essen
Duisburg, Germany

Gabriel González
NanoScience Technology Center

and

Department of Physics
University of Central Florida
Orlando, Florida

Evgeny P. Grebennikov
Open Joint-Stock Company
Central Scientific Research Institute
of Technology "Technomash"
Moscow, Russia

Jim Greer
Tyndall National Institute
University College Cork
Cork, Ireland

Akihiro Hashimoto
Department of Electrical and Electronics
Engineering
Graduate School of Engineering
University of Fukui
Fukui, Japan

Jun He
Department of Chemistry
Institute for Optical Sciences

and

Centre for Quantum Information
and Quantum Control
University of Toronto
Toronto, Ontario, Canada

Woong-Ki Hong
Department of Materials Science
and Engineering
Gwangju Institute of Science
and Technology
Gwangju, Korea

Zhijun Hu
Center for Soft Matter Physics
and Interdisciplinary Research
Soochow University
Suzhou, China

Vanessa M. Huxter
Department of Chemistry
Institute for Optical Sciences

and

Centre for Quantum Information
and Quantum Control
University of Toronto
Toronto, Ontario, Canada

Frank Jahnke
Institute for Theoretical Physics
University of Bremen
Bremen, Germany

Gunho Jo
Department of Materials Science
and Engineering
Gwangju Institute of Science
and Technology
Gwangju, Korea

Alain M. Jonas
Institute of Condensed Matter
and Nanosciences
Division of Bio- and Soft Matter
Catholic University of Louvain
Louvain-la-Neuve, Belgium

Yoshihiko Kanemitsu
Institute for Chemical Research
Kyoto University
Uji, Kyoto, Japan

Tadashi Kawazoe
Department of Electrical Engineering
and Information Systems
School of Engineering
The University of Tokyo
Tokyo, Japan

Stephen Knight
Office of Microelectronics Programs
National Institute of Standards
and Technology
Gaithersburg, Maryland

Kiyoshi Kobayashi
Department of Electrical Engineering
and Information Systems
The University of Tokyo
Tokyo, Japan

and

Core Research of Evolutional Science
and Technology
Japan Science and Technology

and

Department of Electrical and Electronic
Engineering
University of Yamanashi
Kofu, Japan

Hermann Kohlstedt
Christian-Albrechts-Universität Zu Kiel
Faculty of Engineering Nanoelectronics
Kiel, Germany

Takhee Lee
Department of Materials Science
and Engineering
Gwangju Institute of Science
and Technology
Gwangju, Korea

Jean Lermé
Laboratoire de Spectrométrie Ionique et Moléculaire
Centre National de la Recherche Scientifique
Université de Lyon
Villeurbanne, France

Michael N. Leuenberger
NanoScience Technology Center

and

Department of Physics
University of Central Florida
Orlando, Florida

James Alexander Liddle
Center for Nanoscale Science and Technology
National Institute of Standards and Technology
Gaithersburg, Maryland

Jongsun Maeng
Department of Materials Science and Engineering
Gwangju Institute of Science and Technology
Gwangju, Korea

Andreas Marent
Institut für Festkörperphysik
Technische Universität Berlin
Berlin, Germany

Vincent Meunier
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Gilles Micolau
Institut Matériaux Microélectronique Nanosciences de Provence
Centre National de la Recherche Scientifique
Universités d'Aix Marseille
Marseille, France

Makoto Naruse
National Institute of Information and Communications Technology
Koganei, Japan

and

Department of Electrical Engineering and Information Systems
School of Engineering
The University of Tokyo
Tokyo, Japan

Wataru Nomura
School of Engineering
The University of Tokyo
Tokyo, Japan

Motoichi Ohtsu
Department of Electrical Engineering and Information Systems
School of Engineering
The University of Tokyo
Tokyo, Japan

Tomáš Ostatnický
Faculty of Mathematics and Physics
Department of Chemical Physics and Optics
Charles University
Prague, Czech Republic

Andrew R. Parker
Department of Zoology
The Natural History Museum
London, United Kingdom

and

School of Biological Science
University of Sydney
Sydney, New South Wales, Australia

André Avelino Pasa
Laboratório de Filmes Finos e Superfícies
Departamento de Física
Universidade Federal de Santa Catarina
Santa Catarina, Brazil

Ivan Pelant
Institute of Physics
Academy of Sciences of the Czech Republic
Prague, Czech Republic

Michel Pellarin
Laboratoire de Spectrométrie Ionique et Moléculaire
Centre National de la Recherche Scientifique
Université de Lyon
Villeurbanne, France

Nikolay A. Pertsev
A.F. Ioffe Physico-Technical Institute
Russian Academy of Sciences
St. Petersburg, Russia

Adrian Petraru
Christian-Albrechts-Universität Zu Kiel
Faculty of Engineering Nanoelectronics
Kiel, Germany

Vivek M. Prabhu
Polymers Division
National Institute of Standards and Technology
Gaithersburg, Maryland

Edik U. Rafailov
Division of Electronic Engineering and Physics
School of Engineering, Physics and Mathematics
University of Dundee
Dundee, United Kingdom

Arndt Remhof
Division of Hydrogen and Energy
Department of Environment, Energy and Mobility
Swiss Federal Laboratories for Materials Testing and Research
Dübendorf, Switzerland

Elisabetta Romano
Department of Materials Science
University of Milano-Bicocca
Milano, Italy

Harry E. Ruda
Centre for Advanced Nanotechnology
University of Toronto
Toronto, Ontario, Canada

Amir Sa'ar
Racah Institute of Physics

and

The Harvey M. Kruger Family Center
for Nanoscience and Nanotechnology
The Hebrew University of Jerusalem
Jerusalem, Israel

Suguru Sangu
Device and Module Technology
Development Center
Ricoh Company, Ltd.
Yokohama, Japan

Gregory D. Scholes
Department of Chemistry
Institute for Optical Sciences

and

Centre for Quantum Information
and Quantum Control
University of Toronto
Toronto, Ontario, Canada

Sabyasachi Sen
Department of Chemistry
JIS College of Engineering
Kolkata, West Bengal, India

Alexander Shik
Centre for Advanced Nanotechnology
University of Toronto
Toronto, Ontario, Canada

Christopher L. Soles
Polymers Division
National Institute of Standards
and Technology
Gaithersburg, Maryland

Sunghoon Song
Department of Materials Science
and Engineering
Gwangju Institute of Science
and Technology
Gwangju, Korea

Bobby G. Sumpter
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Kien Wen Sun
Department of Applied Chemistry
National Chiao Tung University
Hsinchu, Taiwan

Jan Valenta
Faculty of Mathematics and Physics
Department of Chemical Physics
and Optics
Charles University
Prague, Czech Republic

Marek S. Wartak
Department of Physics and Computer
Science
Wilfrid Laurier University
Waterloo, Ontario, Canada

Andreas Westphalen
Department of Physics and Astronomy
Institute for Condensed Matter Physics
Ruhr-Universität Bochum
Bochum, Germany

Obert R. Wood II
GLOBALFOUNDRIES
Albany, New York

Naoki Yamamoto
Department of Physics
Tokyo Institute of Technology
Tokyo, Japan

Takashi Yatsui
School of Engineering
The University of Tokyo
Tokyo, Japan

Hartmut Zabel
Department of Physics and Astronomy
Institute for Condensed Matter Physics
Ruhr-Universität Bochum
Bochum, Germany

Natalya A. Zimbovskaya
Department of Physics and Electronics
University of Puerto Rico
Humacao, Puerto Rico

and

Institute for Functional Nanomaterials
University of Puerto Rico
San Juan, Puerto Rico

I

Computing and Nanoelectronic Devices

1 **Quantum Computing in Spin Nanosystems** *Gabriel González and Michael N. Leuenberger* **1**-1
Introduction • Qubits and Quantum Logic Gates • Conditions for the Physical Implementation of Quantum Computing • Zeeman Effects • Atom–Light Interaction • Loss–DiVincenzo Proposal • Quantum Computing with Molecular Magnets • Semiconductor Quantum Dots • Single-Photon Faraday Rotation • Concluding Remarks • Acknowledgments • References

2 **Nanomemories Using Self-Organized Quantum Dots** *Martin Geller, Andreas Marent, and Dieter Bimberg*....... **2**-1
Introduction • Conventional Semiconductor Memories • Nonconventional Semiconductor Memories • Semiconductor Nanomemories • A Nanomemory Based on III–V Semiconductor Quantum Dots • Capacitance Spectroscopy • Charge Carrier Storage in Quantum Dots • Write Times in Quantum Dot Memories • Summary and Outlook • Acknowledgments • References

3 **Carbon Nanotube Memory Elements** *Vincent Meunier and Bobby G. Sumpter*... **3**-1
Introduction • CNFET-Based Memory Elements • NEMS-Based Memory • Electromigration CNT-Based Data Storage • General Conclusions • Acknowledgments • References

4 **Ferromagnetic Islands** *Arndt Remhof, Andreas Westphalen, and Hartmut Zabel*.. **4**-1
Introduction • Background • State of the Art • Summary and Discussion • References

5 **A Single Nano-Dot Embedded in a Plate Capacitor** *Gilles Micolau and Damien Deleruyelle*................................ **5**-1
Introduction • Studied Configuration: Geometry and Notations • Metallic Dot • Semiconductor Dot: Theoretical Approach • Finite Element Modeling of a Silicon Quantum Dot • Conclusion • Appendix A: A Numerical Validation of the Semi-Analytical Approach • References

6 **Nanometer-Sized Ferroelectric Capacitors** *Nikolay A. Pertsev, Adrian Petraru, and Hermann Kohlstedt* **6**-1
Introduction • Fundamentals of Ferroelectricity • Deposition and Patterning • Characterization of Ferroelectric Films and Capacitors • Physical Phenomena in Ferroelectric Capacitors • Future Perspective • Summary and Outlook • References

7 **Superconducting Weak Links Made of Carbon Nanostructures** *Vincent Bouchiat* **7**-1
Introduction • Superconducting Transport through a Weak Link • Superconducting Transport in a Carbon Nanotube Weak Link • Nanotube-Based Superconducting Quantum Interferometers • Graphene-Based Superconducting Weak Links • Fullerene-Based Superconducting Weak Links • Concluding Remarks • Acknowledgment • References

8 **Micromagnetic Modeling of Nanoscale Spin Valves** *Bruno Azzerboni, Giancarlo Consolo, and Giovanni Finocchio* **8**-1
Introduction • Background • Computational Micromagnetics of Nanoscale Spin-Valves: State of the Art • Conclusions and Future Perspective • References

9 **Quantum Spin Tunneling in Molecular Nanomagnets** *Gabriel González and Michael N. Leuenberger* **9**-1
Introduction • Spin Tunneling in Molecular Nanomagnets • Phonon-Assisted Spin Tunneling in Mn$_{12}$ Acetate • Interference between Spin Tunneling Paths in Molecular Nanomagnets • Incoherent Zener Tunneling in Fe$_8$ • Coherent Néel Vector Tunneling in Antiferromagnetic Molecular Wheels • Berry-Phase Blockade in Single-Molecule Magnet Transistors • Concluding Remarks • Acknowledgment • References

10 Inelastic Electron Transport through Molecular Junctions *Natalya A. Zimbovskaya* **10**-1
Introduction • Coherent Transport • Buttiker Model for Inelastic Transport • Vibration-Induced Inelastic Effects • Dissipative Transport • Polaron Effects: Hysteresis, Switching, and Negative Differential Resistance • Molecular Junction Conductance and Long-Range Electron-Transfer Reactions • Concluding Remarks • References

11 Bridging Biomolecules with Nanoelectronics *Kien Wen Sun and Chia-Ching Chang* **11**-1
Introduction and Background • Preparation of Molecular Magnets • Nanostructured Semiconductor Templates: Nanofabrication and Patterning • Self-Assembling Growth of Molecules on the Patterned Templates • Magnetic Properties of Molecular Nanostructures • Conclusion and Future Perspectives • References

1

Quantum Computing in Spin Nanosystems

1.1 Introduction **1**-1
1.2 Qubits and Quantum Logic Gates **1**-2
1.3 Conditions for the Physical Implementation of Quantum Computing **1**-3
1.4 Zeeman Effects **1**-3
Weak External Field • Strong External Field
1.5 Atom–Light Interaction **1**-4
Jaynes–Cummings Model
1.6 Loss–DiVincenzo Proposal **1**-7
RKKY Interaction
1.7 Quantum Computing with Molecular Magnets **1**-9
1.8 Semiconductor Quantum Dots **1**-12
Classical Faraday Effect • Quantum Faraday Effect
1.9 Single-Photon Faraday Rotation **1**-14
Quantifying the EPR Entanglement • Single Photon Faraday Effect and GHZ Quantum Teleportation • Single-Photon Faraday Effect and Quantum Computing
1.10 Concluding Remarks **1**-20
Acknowledgments **1**-20
References **1**-20

Gabriel González
University of Central Florida

Michael N. Leuenberger
University of Central Florida

1.1 Introduction

The history of quantum computers begins with the articles of Richard Feynman who, in 1982, speculated that quantum systems might be able to perform certain tasks more efficiently than would be possible in classical systems (Feynman, 1982). Feynman was the first to propose a direct application of the laws of quantum mechanics to a realization of quantum algorithms. The fundamentals of quantum computing were introduced and developed by several authors after Feynman's idea. A model and a description of a quantum computer as a quantum Turing machine was developed by Deutsch (1985). In 1994, Shor introduced the quantum algorithm for the integer-number factorization and in 1997, Grover proposed the fast quantum search algorithm (Grover, 1997). Later on, Wooters and Zurek proved the noncloning theorem, which puts definite limits on the quantum computations, but Shor's work challenges all that with the quantum error correction code (Shor, 1995). In the last years, the development of quantum computing has grown to enormous practical importance as an interdisciplinary field, which links the elements of physics, mathematics, and computer science. Currently, various physical models of quantum computers are under intensive study. Several types of elementary quantum computing devices have been developed based on atomic, molecular, optical, and semiconductor physics and technologies.

Before contemplating the physical realization of a quantum computer, it is necessary to decide how information is going to be stored within the system and how the system will process that information during a desired computation. In classical computers, the information is typically carried in microelectronic circuits that store information using the charge properties of electrons. Information processing is carried out by manipulating electrical fields within semiconductor materials in such a way as to perform useful computational tasks. Presently it seems that the most promising physical model for quantum computation is based on the electron's spin. A strong research effort toward the implementation of the electron spin as a new information carrier has been the subject of a new form of electronics based on spin called *spintronics*. Experiments that have been conducted on quantum spin dynamics in semiconductor materials demonstrate that electron spins have several characteristics that are promising for quantum computing applications. Electron spin states possess the following advantages: very long relaxation time in the absence of external fields, fairly long decoherence time $\tau_d \approx 1\,\mu s$, and the possibility of easy spin manipulation by an external magnetic field. These characteristics are very promising since longer decoherence times relax constraints on the switching speeds of quantum gates necessary for reliable error correction. Typically quantum gates are required to switch 10^4 times faster than the loss of qubit coherence. Spin coherent transport over lengths as large as $100\,\mu$m have been reported in semiconductors. This makes electron spin a perfect candidate as an information carrier in semiconductors (Adamowski et al., 2005).

The spin of particles exhibiting quantum behavior is specially suitable for the construction of quantum computers. The electron spin states can be used to construct qubits and logic operations in different ways. They can be constructed either directly with the application of a magnetic field or indirectly with the application of symmetry properties of the many-electron wave function (namely by the resulting singlet or triplet spin states). Traditionally, in nanoelectronic devices, the charge of the electrons has been used to carry and transform information, which makes the interaction of spintronic devices with charge-based devices important for compatibility with existing classical computing schemes. Overall, we can say that spin nanosystems are good candidates for the physical implementation of quantum computation.

1.2 Qubits and Quantum Logic Gates

We know that the information stored in a classical computer can take one of the two values, i.e., 0 or 1, with probability 0 or 1 each. Quantum bits or qubits are the quantum mechanical analogue of classical bits. In contrast, the qubit can be defined as a quantum state vector in a two-dimensional Hilbert space. Suppose $|0\rangle$, $|1\rangle$ forms a basis for the Hilbert space, then the qubit can be expressed as the superposition of the two states as

$$|\psi\rangle = c_0\,|0\rangle + c_1\,|1\rangle, \tag{1.1}$$

where c_0 and c_1 are complex numbers and the modulus squared of each complex number represents the probability to obtain the qubit $|0\rangle$ or $|1\rangle$, respectively. Additionally, they must satisfy the normalization condition $|c_0|^2 + |c_1|^2 = 1$. Contrary to the classical bit, the quantum bit takes on a continuum of values, which are determined by the probability amplitudes given by c_0 and c_1. If we perform a measurement on qubit $|\psi\rangle$, we obtain either outcome $|0\rangle$ with probability $|c_0|^2$ or outcome $|1\rangle$ with probability $|c_1|^2$. However, if the qubit is prepared to be exactly equal to one of the states of the computational basis, i.e., $|\psi\rangle = |0\rangle$ or $|\psi\rangle = |1\rangle$, then we can predict the exact result of the measurement with probability 1. This nondeterministic characteristic between the general state of the qubit and the precise result of the measurement in the basis state plays an essential role in quantum computations. To carry out a quantum computation, we require at least a two-qubit state, i.e., the states of a two-particle quantum system. The two-qubit states can be constructed as tensor products of the basis states $|0\rangle$, $|1\rangle$. The two-qubit basis consists of the states $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$, where in the shorthanded notation $|0\rangle \otimes |1\rangle \equiv |01\rangle$, etc. implied. An arbitrary two-qubit state has the form

$$|\Psi\rangle = c_0\,|00\rangle + c_1\,|00\rangle + c_2\,|10\rangle + c_3\,|11\rangle, \tag{1.2}$$

where the normalization condition takes the form $|c_0|^2 + |c_1|^2 + |c_2|^2 + |c_3|^2 = 1$. Another basic characteristic for quantum computing is the so-called *entanglement*. The two quantum two-level systems can become *entangled* by interacting with each other. This means that we cannot fully describe one system independently of the other. For example, suppose that the state $(|01\rangle - |10\rangle)/\sqrt{2}$ gives a complete description of the whole system. Then a measurement over the first subsystem forces the second subsystem into one of the two states $|0\rangle$ or $|1\rangle$. This means that one measurement over one subsystem influences the other, even though it may be arbitrarily far away.

Qubits can be transformed by unitary transformations (observables) $\mathcal{U}$ that play the role of quantum logic gates and which transforms the initial qubit into a final qubit according to

$$|\Psi_{\mathrm{f}}\rangle = \mathcal{U}\,|\Psi_{\mathrm{i}}\rangle. \tag{1.3}$$

Depending on the type of qubit on which they operate, we deal with either a 2 × 2 or a 4 × 4 matrix. For example, the quantum NOT gate is defined as

$$\mathcal{U}_{\mathrm{NOT}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{1.4}$$

The effect of the quantum logic gate (4) in the one-qubit state given in Equation 1.1 is to exchange the probability amplitudes i.e.,

$$\mathcal{U}_{\mathrm{NOT}} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_0 \end{pmatrix}. \tag{1.5}$$

An example of an important quantum logic gate that operates on a two-qubit state is the so-called controlled-NOT gate $\mathcal{U}_{\mathrm{CNOT}}$, for which the first qubit is the control qubit and the second qubit is the target qubit. The controlled-NOT gate transforms the two-qubit basis states as follows:

$$\begin{gathered} \mathcal{U}_{\mathrm{CNOT}}\,|00\rangle = |00\rangle, \quad \mathcal{U}_{\mathrm{CNOT}}\,|01\rangle = |01\rangle, \\ \mathcal{U}_{\mathrm{CNOT}}\,|10\rangle = |11\rangle, \quad \text{and} \quad \mathcal{U}_{\mathrm{CNOT}}\,|11\rangle = |10\rangle, \end{gathered} \tag{1.6}$$

which means that the CNOT (conditionalNOT) gate changes the second qubit if and only if the first qubit is in state $|1\rangle$. The matrix representation of the CNOT gate is

$$\mathcal{U}_{\mathrm{CNOT}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{1.7}$$

It has been shown that the set of logic operations, which consists of all the one-qubit gates and the single two-qubit gate $\mathcal{U}_{\mathrm{CNOT}}$ is universal in the sense that all unitary transformations on

N-qubit states can be expressed by different compositions of the set of universal gates (DiVincenzo, 1995).

Using the concepts of superposition and entanglement, we can describe another important characteristic of quantum computation, which is known as *quantum parallelism*. Quantum parallelism is based on the fact that a single unitary transformation can simultaneously operate on all the qubits in the system. In a sense, it can perform several calculations in a single step. In fact, it can be proved that the computing power of a quantum computer scales exponentially with the number of qubits, whereas a classical computer the scale is only linear.

1.3 Conditions for the Physical Implementation of Quantum Computing

The successful implementation of a quantum computer has to satisfy some basic requirements. These are known as the DiVincenzo criteria and can be summarized in the following way (DiVincenzo, 2001).

1. *Physical realizability of the qubits*. We need to find some quantum property of a scalable physical system in which to encode our information so that it lives long enough to enable us to perform computations.
2. *Initial state preparation*. It should be possible to precisely prepare the initial qubit state.
3. *Isolation*. We need a controlled evolution of the qubit; this will require enough isolation of the qubit from the environment to reduce the effects of decoherence.
4. *Gate implementation*. We need to be able to manipulate the states of individual qubits with reasonable precision, as well as to induce interaction between them in a controlled way, so that the implementation of gates is possible. Also, the gate operation time τ_s has to be much shorter than the decoherence time τ_d, so that $\tau/\tau_d \ll r$, where r is the maximum tolerable error rate for quantum error correction schemes to be effective.
5. *Readout*. We must be able to accurately measure the final qubit state.

The conditions listed above put certain limitations on the quantum computing technology. For example, the complete isolation of the qubit with respect to the environment disables the read/write operations. Therefore, some slight interaction of the quantum system and the environment is necessary. On the other hand, this interaction leads to decay and decoherence processes, which reduce the performance of the quantum computer.

In the decay process, the quantum system jumps in a very short time to a new state, releasing part of its energy to the environment. The decay is characterized by the relaxation time, which for the spin states can be very long. Decoherence is a more subtle process because the energy is conserved but the relative phase of the computational basis is changed. As a result of the decoherence the qubit changes as follows:

$$|\psi\rangle = c_0|0\rangle + e^{i\theta}c_1|1\rangle, \tag{1.8}$$

where the real number θ denotes the relative phase. The appearance of the nonzero relative phase results due to the coupling between the quantum system and the environment and can lead to significant changes in the measurement process. The ratio of the decoherence time to the elementary operation time τ_s, i.e., $R = \tau_d/\tau_s$, is an approximate measure of the number of computation steps performed before the coupling with the environment destroys the qubit. This ratio changes abruptly for different quantum computing schemes. For example, $R = 10^3$ for the electron states in quantum dots, $R = 10^7$ for nuclear spin sates, and $R = 10^{13}$ for trapped ions.

An important factor that should always be kept in mind when constructing quantum computers is the scalability of the device. We should be able to enlarge the physical device to contain many qubits and still fulfill the DiVincenzo requirements described above.

1.4 Zeeman Effects

An external magnetic field superimposed on an atom perturbs its state in a definite way. The Hamiltonian for such an atom may be divided into the operator $\mathcal{H}_0$ for the unperturbed atom and the perturbation operator $\mathcal{H}'$ due to the magnetic field. The external magnetic field $\vec{H}$ causes the vectors $\vec{L}$ and $\vec{S}$ to precess about its direction. We shall examine the two cases where the magnetic field is weak and is strong.

Let us first write down the perturbation operator explicitly. Consider an electron in the simplest possible atom (hydrogen) rotating around the nucleus. The electron orbit can be regarded as a current loop. The current is the charge per unit time past any fixed point on the orbit, therefore

$$j = \frac{e}{T} = \frac{ev}{2\pi r} = \frac{e(p/m_e)}{2\pi r}, \tag{1.9}$$

where p and m_e are the momentum and mass of the electron, respectively. The magnitude of the magnetic moment of the loop is the current times the enclosed area, i.e., $\mu_{orb} = \pi r^2 j$, therefore

$$\vec{\mu}_{orb} = \frac{e}{2m_e}\vec{r} \times \vec{p} = \frac{\mu_B}{\hbar}\vec{L}, \tag{1.10}$$

where we have introduced the Bohr magneton, $\mu_B = e\hbar/2m_e$. We do not have a good semiclassical picture for spin and therefore we can only conclude that the spin magnetic moment is

$$\vec{\mu}_{sp} = g_s\frac{\mu_B}{\hbar}\vec{S}, \tag{1.11}$$

where g_s is called the gyromagnetic factor and its value is approximately equal to 2. Therefore, the total magnetic moment is

$$\vec{\mu} = \frac{\mu_B}{\hbar}(\vec{L} + 2\vec{S}). \tag{1.12}$$

The perturbation energy caused by the magnetic field is given then by

$$\mathcal{H}' = -\vec{\mu} \cdot \vec{H} = \frac{\mu_B}{\hbar}(\vec{J} + \vec{S}), \tag{1.13}$$

where the plus sign resulted because the charge of the electron is $-e$.

1.4.1 Weak External Field

In a hydrogen atom, the electron circles around the proton; but in a system fixed to an electron, the proton circles around the electron and generates an inner magnetic field at the position of the electron given by

$$\vec{H}_{\text{in}} = \frac{\mu_0 e}{4\pi m_p r^3}\vec{L}. \tag{1.14}$$

Equation 1.14 gives the following energy for the electron's spin

$$\mathcal{H}_{\text{so}} = \vec{\mu}_{\text{sp}} \cdot \vec{H}_{\text{in}} = \frac{e^2}{4\pi\epsilon_0 r^3 m^2 c^2}\vec{S} \cdot \vec{L}, \tag{1.15}$$

which is called the spin–orbit interaction. We interpret the spin–orbit interaction as an internal Zeeman effect because it splits the energy levels without an external magnetic field. Let the external field be weak compared with the effective internal field. Since the Larmor frequency is proportional to the magnetic field, the triangle $\vec{L}\,\vec{S}\,\vec{J}$ in Figure 1.1 rotates about $\vec{J}$ considerably faster than the precession around $\vec{H}$. Therefore, the coupling of the vectors $\vec{L}$, $\vec{S}$, and $\vec{J}$ in the triangle is not disrupted.

Let us now find the correction to the energy due to the external magnetic field. Equation 1.13 involves the vectors $\vec{J}$ and $\vec{S}$, if we consider that the z-axis coincides with the direction of the external magnetic field, then, the only projections of these two vectors that contribute to the energy are the ones along the z-axis. Thus, the mean value of $\mathcal{H}'$ is equal to

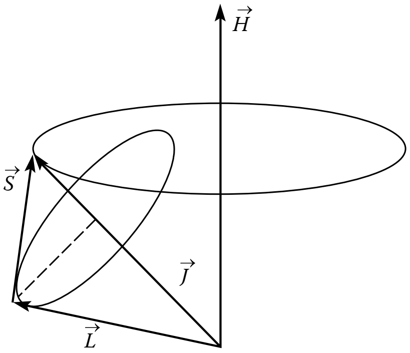

FIGURE 1.1 Schematic of the spin–orbit coupling.

$$\langle \mathcal{H}' \rangle = \frac{\mu_B H J_z}{\hbar}\left(1 + \frac{\vec{S} \cdot \vec{J}}{J^2}\right), \tag{1.16}$$

and using

$$\frac{\vec{S} \cdot \vec{J}}{J^2} = \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)} = g \tag{1.17}$$

we get

$$\langle \mathcal{H}' \rangle = \mu_B H g m_j \quad \text{for } m_j = -j, -j+1, \ldots, j. \tag{1.18}$$

Thus, the energy levels with a given J are split into as many levels as there are different projections of $\vec{J}$ on the magnetic field, i.e., $2j + 1$. The factor g is called the Landé factor and takes a given value for a given L and S and each corresponding value of J. The Zeeman effect in a weak magnetic field is called anomalous.

1.4.2 Strong External Field

We shall now consider the opposite extreme case, when the external field is strong compared with the internal field, so that the coupling between the vectors $\vec{L}, \vec{S}, \vec{J}$ is disrupted. This can be explained by the fact that $\vec{S}$ precesses twice as fast $\vec{L}$. Then, from the classical analogy, each of the vectors $\vec{S}$ and $\vec{L}$ precesses independently about the magnetic field. For this case, the correction to the energy is given by

$$\langle \mathcal{H}' \rangle = \frac{\mu_B e H}{\hbar}(L_z + 2S_z). \tag{1.19}$$

In Equation 1.19, L_z and S_z are the projections of the orbital angular momentum and spin along the z-axis, respectively, and are given by

$$L_z = \hbar m_l \quad \text{for } m_l = -l, \ldots, l, \quad S_z = \hbar m_s, \quad \text{for } m_s = \pm\frac{1}{2}. \tag{1.20}$$

The projections S_z and L_z are changed by unity, therefore all the levels in Equation 1.19 are equidistant. Of course, certain values of $\langle \mathcal{H}' \rangle$ may be repeated several times if the sum $L_z + 2S_z$ assumes the same value. This Zeeman effect in a strong magnetic field is called the normal Zeeman effect.

1.5 Atom–Light Interaction

Though an atom has infinitely many energy levels, we can have under certain assumptions a two-level atom. These assumptions are (1) the difference of the energy levels approximately matches the energy of the incident photon, (2) the selection rules allow transition of the electrons between the two levels, and (3) all other energy levels are sufficiently detuned in frequency separation with respect to the incoming frequency such that there is no transition to these levels.

Consider that we apply the two-level approximation to a simple atom where the interaction with visible light involves a single electron. The corresponding wave function of the system is

$$|\psi\rangle = c_0\,|0\rangle + c_1\,|1\rangle, \tag{1.21}$$

where $|0\rangle$ ($|1\rangle$) corresponds to the nonexcited (excited) state and the coefficients $c_0(c_1)$ represent the probability amplitude to find the system in either state, respectively. The probability amplitudes satisfy the normalization condition $|c_0|^2 + |c_1|^2 = 1$. The wavelength of visible light, typical for atomic transitions, is about a few thousand times the diameter of an atom. Therefore, there is no significant spatial variation of the electric field across an atom, and $\vec{E}(\vec{r},t) \approx \vec{E}(t)$ can be taken as independent of the position (dipole approximation). Consistent with this long-wavelength approximation is that the magnetic field $\vec{H}$ is approximately zero, i.e., $\vec{H} \approx 0$, so that the interaction energy between the field and the electron of the atom is given by $\mathcal{H}' = -e\vec{r}\cdot\vec{E}$, hence the total Hamiltonian is

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}', \tag{1.22}$$

where $\mathcal{H}_0$ represents the unperturbed atom system, i.e., $\mathcal{H}_0\,|0\rangle = \varepsilon_0\,|0\rangle$ and $\mathcal{H}_0\,|1\rangle = \varepsilon_1\,|1\rangle$, where ε_0 and ε_1 are the energy values for the nonexcited and excited states, respectively.

Seeking a solution to the Schrödinger equation

$$\mathcal{H}\,|\psi(t)\rangle = i\hbar\frac{\partial\,|\psi(t)\rangle}{\partial t}, \tag{1.23}$$

in the following form

$$|\psi(t)\rangle = c_0(t)e^{-i\varepsilon_0 t/\hbar}\,|0\rangle + c_1(t)e^{-i\varepsilon_1 t/\hbar}\,|1\rangle, \tag{1.24}$$

we get

$$\mathrm{i}\hbar\frac{c_l(t)}{\mathrm{d}t} = \sum_{k=0,1} c_k(t)\mathrm{e}^{-\mathrm{i}(\varepsilon_k-\varepsilon_l)t/\hbar}\langle l\,|\,\mathcal{H}'\,|\,k\rangle, \quad \text{for } l = 0, 1. \tag{1.25}$$

Representing the light with frequency ν by means of the complex electric field vector E_0 in the form

$$\vec{E}(t) = \frac{1}{2}(\vec{E}_0\mathrm{e}^{-\mathrm{i}\nu t} + \vec{E}_0^{\star}\mathrm{e}^{\mathrm{i}\nu t}), \tag{1.26}$$

and using the fact that the diagonal elements of the interacting Hamiltonian are zero due to the parity of the eigenfunction, i.e., $\langle k\,|\,\mathcal{H}'\,|\,k\rangle = 0$, we end up with the following coupled differential equations

$$\frac{\mathrm{d}c_0(t)}{\mathrm{d}t} = \frac{i}{2\hbar}\left[\vec{E}_0\cdot\langle 0\,|\,\vec{d}\,|\,1\rangle\mathrm{e}^{-\mathrm{i}(\omega_{10}+\nu)t} + \vec{E}_0^{\star}\cdot\langle 0\,|\,\vec{d}\,|\,1\rangle\mathrm{e}^{-\mathrm{i}(\omega_{10}-\nu)t}\right]c_1(t), \tag{1.27}$$

$$\frac{\mathrm{d}c_1(t)}{\mathrm{d}t} = \frac{i}{2\hbar}\left[\vec{E}_0^{\star}\cdot\langle 1\,|\,\vec{d}\,|\,0\rangle\mathrm{e}^{\mathrm{i}(\omega_{10}+\nu)t} + \vec{E}_0\cdot\langle 1\,|\,\vec{d}\,|\,0\rangle\mathrm{e}^{\mathrm{i}(\omega_{10}-\nu)t}\right]c_0(t), \tag{1.28}$$

where $\omega_{10} = (\varepsilon_1 - \varepsilon_0)/\hbar$ and $\langle 0|\vec{d}|1\rangle = \langle 1|\vec{d}|0\rangle^{\star} = \langle 0|e\vec{r}|1\rangle$.

For the case when $\nu \approx \omega_{10}$, we can drop the terms containing $\exp[\pm(\omega_{10} + \nu)t]$ in Equation 1.28 since they oscillate rapidly in time and can be neglected with respect to the near resonant terms, i.e., the terms of the form $\exp[\pm(\omega_{10} - \nu)t]$. This approximation is called the rotating wave approximation (RWA), and the coupled differential equations become

$$\begin{aligned} \frac{\mathrm{d}c_0(t)}{\mathrm{d}t} &= \frac{\mathrm{i}\Omega_{\mathrm{R}}^{\star}}{2}\mathrm{e}^{-\mathrm{i}\Delta t}c_1(t), \\ \frac{\mathrm{d}c_1(t)}{\mathrm{d}t} &= \frac{\mathrm{i}\Omega_{\mathrm{R}}}{2}\mathrm{e}^{\mathrm{i}\Delta t}\,c_0(t), \end{aligned} \tag{1.29}$$

where

$\Omega_{\mathrm{R}} = \vec{E}_0\cdot\langle 1|\vec{d}|0\rangle/\hbar$ is known as the Rabi frequency
$\Delta = (\omega_{10} - \nu)$ is the so-called detuning

The general solution for Equation 1.29 for strictly monochromatic fields, i.e., $\Omega_{\mathrm{R}} = |\Omega_{\mathrm{R}}|\mathrm{e}^{\mathrm{i}\phi} = \text{constant}$, is given by

$$\begin{aligned} c_0(t) &= \frac{1}{\Omega}(\mu_1\mathrm{e}^{\mathrm{i}\mu_2 t} - \mu_2\mathrm{e}^{\mathrm{i}\mu_1 t}), \\ c_1(t) &= \frac{\mathrm{i}\Omega_{\mathrm{R}}}{\Omega}\mathrm{e}^{\mathrm{i}\Delta t/2}\sin\left(\frac{\Omega t}{2}\right), \end{aligned} \tag{1.30}$$

where

$$\begin{aligned} \mu_{1,2} &= -\frac{\Delta}{2} \pm \sqrt{\Delta^2 + \Omega_{\mathrm{R}}^2}, \\ \Omega &= \mu_1 - \mu_2 = \sqrt{\Delta^2 + \Omega_{\mathrm{R}}^2}. \end{aligned} \tag{1.31}$$

Equation 1.31 gives the transition probabilities from the nonexcited to the excited state

$$|c_0(t)|^2 = \left(\frac{\Omega_{\mathrm{R}}}{\Omega}\right)^2\sin^2\left(\frac{\Omega t}{2}\right), \quad |c_1(t)|^2 = 1 - |c_0(t)|^2, \tag{1.32}$$

which oscillate with frequency Ω (Rabi flopping frequency) between levels ε_0 and ε_1. Any two-level system can be represented by a 2 × 2 matrix Hamiltonian and hence can be expressed in terms of Pauli matrices. For example, choosing the energy zero to be half way between the excited and nonexcited state $|0\rangle$ and $|1\rangle$, we have the following matrix Hamiltonian for this system

$$\mathcal{H}_0 = \frac{\hbar\omega}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \frac{\hbar\omega}{2}\sigma_z, \tag{1.33}$$

where

$\hbar\omega = \varepsilon_1 - \varepsilon_0$

σ_z is a Pauli matrix

Therefore, we can write the total Hamiltonian of the atom–light interaction in the RWA as

$$\mathcal{H} = \frac{\hbar}{2}\begin{pmatrix} \omega & -\Omega_{\mathrm{R}}\mathrm{e}^{\mathrm{i}\Delta t} \\ -\Omega_{\mathrm{R}}^{\star}\mathrm{e}^{-\mathrm{i}\Delta t} & -\omega \end{pmatrix} = \frac{\hbar\omega}{2}\sigma_z - \frac{\hbar}{2}(\Omega_{\mathrm{R}}\mathrm{e}^{\mathrm{i}\Delta t}\sigma_+ + \Omega_{\mathrm{R}}^{\star}\mathrm{e}^{-\mathrm{i}\Delta t}\sigma_-), \tag{1.34}$$

where

$$\sigma_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \sigma_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \tag{1.35}$$

1.5.1 Jaynes–Cummings Model

So far, we have used a semiclassical description of the interaction between a two-level atom and an electric field. A quantum description of the interaction would require a quantization of the radiation field. This description is known as the *Jaynes–Cummings model*. Consider the sourceless electromagnetic field in a cubic cavity of volume $V = L \times L \times L$. Working with the Coulomb gauge for which the vector potential satisfies the requirement

$$\nabla \cdot \vec{A} = 0, \tag{1.36}$$

and, since there are no sources, it satisfies the homogeneous wave equation

$$\nabla^2\vec{A} - \frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2} = 0. \tag{1.37}$$

Then the fields are fully specified by the vector potential in the form

$$\vec{E} = -\frac{\partial\vec{A}}{\partial t} \quad \text{and} \quad \vec{B} = \nabla \times \vec{A}. \tag{1.38}$$

We can expand the vector potential in the form

$$\vec{A}(\vec{r},t) = \frac{1}{\sqrt{\epsilon_0 V}}\sum_{\vec{k}}\left[\vec{A}_{\vec{k}}(t)\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} + \vec{A}_{\vec{k}}^{\star}(t)\mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{r}}\right]. \tag{1.39}$$

Equation 1.36 implies that $\vec{A}_{\vec{k}}$ and $\vec{A}_{\vec{k}}^{\star}$ are perpendicular to the vector of propagation $\vec{k}$, therefore we can rewrite Equation 1.39 in the following form:

$$\vec{A}(\vec{r},t) = \frac{1}{\sqrt{\epsilon_0 V}}\sum_{\vec{k}}\left[\vec{\epsilon}_{\vec{k}}A_{\vec{k}}(t)\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} + \vec{\epsilon}_{\vec{k}}^{\star}A_{\vec{k}}^{\star}(t)\mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{r}}\right], \tag{1.40}$$

where $\vec{\epsilon}_{\vec{k}} = (e_1, e_2)$ is called the polarization vector, which is perpendicular to the direction of propagation and $A_{\vec{k}}(t)$ is a time dependent amplitude that varies harmonically with time, i.e., $\dot{A}_{\vec{k}} = -\mathrm{i}\omega_{\vec{k}}A_{\vec{k}}$, where $\omega_{\vec{k}} = \sqrt{k_x^2 + k_y^2 + k_z^2} = |\vec{k}|c$. Assuming that the field is periodic in space, and the lengths of the periods in three perpendicular directions are equal to the dimensions of the cubic box, then $k_j = 2n_j\pi/L$ for $j = x, y, z$ and n_j are integers of any sign. Using Equation 1.38 we get

$$\vec{E}(\vec{r},t) = \frac{i}{\sqrt{\epsilon_0 V}}\sum_{\vec{k}}\nu_{\vec{k}}\left[\vec{\epsilon}_{\vec{k}}A_{\vec{k}}\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} + \vec{\epsilon}_{\vec{k}}^{\star}A_{\vec{k}}^{\star}\mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{r}}\right], \tag{1.41}$$

and

$$\vec{H}(\vec{r},t) = \frac{i}{\sqrt{\epsilon_0 V}}\sum_{\vec{k}}\left[(\vec{k}\times\vec{\epsilon}_{\vec{k}})A_{\vec{k}}\mathrm{e}^{\mathrm{i}\vec{r}\cdot\vec{r}} - (\vec{k}\times\vec{\epsilon}_{\vec{k}}^{\star})A_{\vec{k}}^{\star}\mathrm{e}^{-\mathrm{i}\vec{r}\cdot\vec{r}}\right]. \tag{1.42}$$

The classical Hamiltonian for the field is given by

$$\mathcal{H}_{\mathrm{E-M}} = \frac{\epsilon_0}{2}\int_V \left(|\vec{E}|^2 + c^2|\vec{H}|^2\right)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \tag{1.43}$$

Substituting Equations 1.41 and 1.42 into Equation 1.43 one gets

$$\mathcal{H}_{\mathrm{E-M}} = \sum_{\vec{k}}\omega_{\vec{k}}^2A_{\vec{k}}A_{\vec{k}}^{\star}. \tag{1.44}$$

If we introduce the variable

$$A_{\vec{k}} = \frac{1}{\sqrt{2}}\left(Q_{\vec{k}} + \mathrm{i}\frac{P_{\vec{k}}}{\omega_{\vec{k}}}\right), \tag{1.45}$$

then Equation 1.44 becomes

$$\mathcal{H}_{\mathrm{E-M}} = \frac{1}{2}\sum_{\vec{k}}(Q_{\vec{k}}^2\omega_{\vec{k}}^2 + P_{\vec{k}}^2). \tag{1.46}$$

Equation 1.46 corresponds to the sum of independent harmonic oscillators. This suggests that each mode of the field is dynamically equivalent to a mechanical harmonic oscillator. The canonical quantization of the field consists of the substitution of the variables $Q_{\vec{k}}$ and $P_{\vec{k}}$ for operators, which fulfill the commutation relation

$$[Q_{\vec{k}}, P_{\vec{k}'}] = \mathrm{i}\hbar\delta_{\vec{k}\vec{k}'}, \tag{1.47}$$

this means that

$$[A_{\vec{k}}, A_{\vec{k}'}^{\dagger}] = \frac{\hbar}{\omega_{\vec{k}}}\delta_{\vec{k}\vec{k}'}. \tag{1.48}$$

Introducing the creation and annihilation operators

$$a_{\vec{k}} = \sqrt{\frac{\omega_{\vec{k}}}{\hbar}}A_{\vec{k}} \quad \text{and} \quad a_{\vec{k}}^{\dagger} = \sqrt{\frac{\omega_{\vec{k}}}{\hbar}}A_{\vec{k}}^{\dagger}, \tag{1.49}$$

we can write the electromagnetic field Hamiltonian in the form

$$\mathcal{H}_{\mathrm{E-M}} = \sum_{\vec{k}}\left(\hbar\omega_{\vec{k}} a_{\vec{k}} a^{\dagger}_{\vec{k}} + \frac{1}{2}\hbar\omega_{\vec{k}}\right). \tag{1.50}$$

The total Hamiltonian that describes the interaction of an atom with the quantized electromagnetic field can be written in the form

$$\mathcal{H} = \sum_{i}\varepsilon_i\,|i\rangle\langle i| + \sum_{\vec{k}}\hbar\omega_k a^{\dagger}_{\vec{k}} a_{\vec{k}} - \sum_{i,j}\langle i|\,\vec{d}\,|j\rangle\,|i\rangle\langle j|\cdot\vec{E}, \tag{1.51}$$

where ϵ_i is the energy corresponding to the state $|i\rangle$ of the atom and we have dropped the constant terms corresponding to the zero point energy.

The electric field operator is given by

$$\vec{E} = i\sum_{\vec{k}}\vec{\epsilon}_{\vec{k}}\sqrt{\frac{\hbar\omega_{\vec{k}}}{\epsilon_0 V}}\left[a_{\vec{k}}\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} - a^{\dagger}_{\vec{k}}\mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{r}}\right]. \tag{1.52}$$

Making the dipole approximation, i.e., $\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}} \approx 1$, and assuming that we are dealing with a two-level atom, i.e., $i, j = 0, 1$, then we can write Equation 1.51 as

$$\mathcal{H} = \frac{\hbar\omega_{10}}{2}\sigma_z + \sum_{\vec{k}}\hbar\omega_{\vec{k}} a^{\dagger}_{\vec{k}} a_{\vec{k}} + \sum_{\vec{k}}(g^{\star}_{\vec{k}}\sigma_+ - g_{\vec{k}}\sigma_-)(a_{\vec{k}} - a^{\dagger}_{\vec{k}}), \tag{1.53}$$

where $g_{\vec{k}} = i\sqrt{\hbar\omega_k/\epsilon_0 V}\,\vec{d}_{01}\cdot\vec{\epsilon}_{\vec{k}}$. The scalar product between the dipole moment and the polarization vector yields a complex number that can be written as $\vec{d}_{01}\cdot\vec{\epsilon}_{\vec{k}} = |\vec{d}_{01}\cdot\vec{\epsilon}_{\vec{k}}|\,\mathrm{e}^{\mathrm{i}\phi}$. This allows one to choose the phase ϕ in such a way that $g_{\vec{k}} = g^{\star}_{\vec{k}}$. If the atom interacts only with one mode of the electromagnetic field, then we can drop the sum over the vector of propagation to get

$$\mathcal{H} = \frac{\hbar\omega_{10}}{2}\sigma_z + \hbar\omega a^{\dagger}_k a_k + g_k(\sigma_+ - \sigma_-)(a_k - a^{\dagger}_k) = \mathcal{H}_0 + \mathcal{H}'. \tag{1.54}$$

In the interaction picture, i.e., $\mathcal{H}_{\mathrm{I}} = \mathrm{e}^{-\mathrm{i}\mathcal{H}_0 t/\hbar}\,\mathcal{H}\,\mathrm{e}^{\mathrm{i}\mathcal{H}_0 t/\hbar}$, the Hamiltonian of interaction will have the form

$$\begin{aligned}\mathcal{H}'_{\mathrm{I}} = g_k\Big(&\sigma_+ a_k \mathrm{e}^{-\mathrm{i}(\omega_{10}-\omega)t} + \sigma_- a^{\dagger}_k\,\mathrm{e}^{\mathrm{i}(\omega_{10}-\omega)t} - \sigma_+ a^{\dagger}_k\,\mathrm{e}^{\mathrm{i}(\omega_{10}+\omega)t}\\ &-\sigma_- a_k \mathrm{e}^{-\mathrm{i}(\omega_{10}+\omega)t}\Big),\end{aligned} \tag{1.55}$$

using the RWA, i.e., dropping the rapidly oscillating terms containing $\mathrm{e}^{\pm\mathrm{i}(\omega_{10}+\omega)t}$, we end up with

$$\mathcal{H}_{\mathrm{I}} = \frac{\hbar\omega_{10}}{2}\sigma_z + \hbar\omega a^{\dagger}_k a_k + g_k(\sigma_+ a_k \mathrm{e}^{-\mathrm{i}\Delta t} + \sigma_- a^{\dagger}_k \mathrm{e}^{\mathrm{i}\Delta t}), \tag{1.56}$$

where $\Delta = \omega_{10} - \omega$ is the detuning. The Hamiltonian of Equation 1.56 corresponds in the Schrödinger picture to

$$\mathcal{H} = \frac{\hbar\omega_{10}}{2}\sigma_z + \hbar\omega a^{\dagger}_k a_k + g_k(\sigma_+ a_k + \sigma_- a^{\dagger}_k). \tag{1.57}$$

Equation 1.57 is the Jaynes–Cummings model.

1.6 Loss–DiVincenzo Proposal

The first proposals for quantum computing made use of cavity quantum electrodynamics (QED), trapped ions, and nuclear magnetic resonance (NMR). All of these proposals benefit from long decoherence times due to a very weak coupling of the qubits to their environment. The long decoherence times have led to big successes in achieving experimental realizations. A conditional phase (CPHASE) gate was demonstrated early on in cavity QED systems. The two-qubit controlled-NOT gate has been realized in single-ion and two-ion versions. The most remarkable realization of the power of quantum computing to date is the implementation of Shor's algorithm to factor the number 15 in a liquid-state NMR quantum computer to yield the known result 5 and 3. However, these proposals may not be scalable and therefore do not meet the DiVincenzo criteria. The requirement for scalability motivated the Loss–DiVincenzo proposal for a solid state quantum computer based on electron spin qubits (Loss and DiVincenzo, 1998). The spin of an electron in a quantum dot can point up or down with respect to an external magnetic field; these eigenstates, $|\uparrow\rangle$ and $|\downarrow\rangle$, correspond to the two basis states of the qubit. The electron trapped in a quantum dot, which is basically a small electrically defined box that can be filled with electrons, can be defined by metal gate electrodes on top of a semiconductor (GaAs/AlGaAs) heterostructure. At the interface between GaAs and AlGaAs conduction band, electrons accumulate and can only move in the lateral direction. Applying negative voltages to the gates locally depletes this two-dimensional electron gas underneath. The resulting gated quantum dots are very controllable and versatile systems, which can be manipulated and probed electrically. When the size of the dot is comparable to the wavelength of the electrons that occupy it, the system exhibits a discrete energy spectrum, resembling that of an atom.

Initialization of the quantum computer can be achieved by allowing all spins to reach their equilibrium thermodynamic ground state at a low temperature T in an applied magnetic field, so that all the spins will be aligned if the condition $|g\mu_{\mathrm{B}}H| \gg k_{\mathrm{B}}T$ is satisfied (where k_{B} is Boltzmann constant).

To perform single-qubit operations, we can apply a microwave magnetic field on resonance with the Zeeman splitting, i.e., with a frequency $f = \Delta E_Z/h$ (h is Planck's constant). The oscillating magnetic component perpendicular to the static magnetic field H results in a spin nutation. By applying the oscillating field for a fixed duration, a superposition of $|\uparrow\rangle$ and $|\downarrow\rangle$ can be created. This magnetic technique is known as electron spin resonance (ESR). In the Loss–Divincenzo proposal, two-qubit operations can be carried out purely electrically by varying the gate voltages between neighboring dots. When the barrier is high, the spins are decoupled. When the interdot barrier is pulsed low,

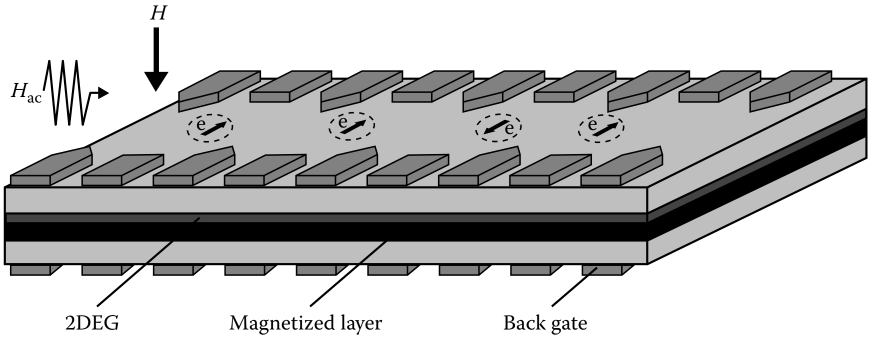

FIGURE 1.2 Theoretical proposal by Loss–DiVincenzo for quantum computing using quantum dots and electric gates.

an appreciable overlap develops between the two electron wave functions, resulting in a nonzero Heisenberg exchange coupling J (see Figure 1.2). The Hamiltonian describing this time dependent process is given by

$$\mathcal{H}(t) = J(t)\vec{S}_n \cdot \vec{S}_{n+1}. \tag{1.58}$$

Equation 1.58 is sometimes referred to as the direct interaction.

The evolution of the quantum state is described by the propagator given by $U(t) = \mathcal{T}\exp[-i\int\mathcal{H}(t)\mathrm{d}t/\hbar]$, where $\mathcal{T}$ is the time ordering operator. If the exchange is pulsed on for a time τ_s such that $\int J(t)\mathrm{d}t/\hbar = J_0\tau_s/\hbar = \pi$, the states of the two spins will be exchanged. This is the SWAP operation. Pulsing the exchange for a shorter time $\tau_s/2$ generates the square root of SWAP operation, which can be used in conjunction with single-qubit operator to generate the controlled-NOT gate.

A last crucial ingredient requires a method to read out the state of the spin qubit. This implies measuring the spin orientation of a single electron. Therefore, an indirect spin measurement is proposed. First, the spin orientation of the electron is correlated with its position, via spin to charge conversion. Then an electrometer is used to measure the position of the charge, thereby revealing its spin. In this way, the problem of measuring the spin orientation has been replaced by the much easier measurement of charge.

The Loss–DiVincenzo ideas have influenced an enormous research effort aimed at implementing the different parts of the proposal and has been quickly followed by a series of alternative solid state realizations for trapped atoms in optical lattices that may also be scalable. It should also be stressed that the efforts to create a spin qubit are not purely application driven. If we have the ability to control and read out a single electron spin, we are in a unique position to study the interaction of the spin with its environment. This may lead to a better understanding of decoherence and will also allow us to study the semiconductor environment using the spin as a probe.

1.6.1 RKKY Interaction

The RKKY interaction is a long-range magnetic interaction that involves nearest-neighbor ions as well as magnetic atoms that are further apart; this interaction is sometimes called indirect interaction. The basis of this interaction lies in an exchange interaction, which was proposed by Ruderman and Kittel (1954), and extended by Kasuya (1956) and Yosida (1957), and now known as the RKKY interaction. This interaction refers to an exchange energy written as

$$\mathcal{H}_{\text{exch}} = -J(r)\sum_i \vec{s}_i \cdot \vec{S}, \tag{1.59}$$

where

The exchange parameter $J(r)$ falls off rapidly with distance r between the center of a localized magnetic ion and electron spin
$\vec{s}_i$ is the spin state of a conduction electron
$\vec{S}$ is the spin of a localized magnetic ion

The minus sign in Equation 1.59 is related to the Pauli exclusion principle (lowest energy of electron occupancy). Below some critical magnetic ordering temperature, the itinerant or localized spins may condense into an ordered array, i.e., ferromagnetic or antiferromagnetic. As in the case of atomic scattering, any disorder within this array will cause additional electron scattering. This scattering may be elastic, in which case there is no change in energy or spin-flip, or it may be inelastic, in which case the spin state of the conduction electrons changes. Usually RKKY interaction is of significance only in compounds with a high concentration of the magnetic atom. Thus, localized ions may start to interact indirectly via the conduction electrons.

Now, consider the case in which there exist two localized spins at lattice points $\vec{R}_n$ and $\vec{R}_m$. By the interaction between spin S_{mz} localized at $\vec{R}_m$ and the spin density of conduction electrons polarized by spin S_{nz} localized at $\vec{R}_n$, the following interaction between the spins $\vec{S}_n$ and $\vec{S}_m$ is found:

$$\mathcal{H}_{\text{exch}} = -9\pi\frac{J^2}{\varepsilon_F}\left(\frac{N_e}{N}\right)^2 F\left(2k_F\,|\vec{R}_n - \vec{R}_m|\right)S_{mz}S_{nz}, \tag{1.60}$$

where

$$F(x) = \frac{-x\cos(x) + \sin(x)}{x^4}. \tag{1.61}$$

Recently, the optical RKKY interaction between two spins was introduced as a means to produce an effective exchange interaction (Piermarocchi et al., 2002). Similar to the Loss–DiVincenzo scheme, in this scheme, the two qubits are defined by the excess electrons of semiconductor quantum dots. Instead of using the direct exchange interaction between the two electrons, the exchange interaction is indirectly mediated by the itinerant electrons of virtual excitons that are optically excited in the host material, which can be made of bulk, quantum well, or quantum wire structures. This scheme has the advantage that two-qubit gates can be performed on the femtosecond timescale due to the possibility of using ultrafast laser optics. The Coulomb interaction between the photoexcited itinerant electrons and the localized electrons in the two quantum dots contains direct and indirect terms. While the direct terms give rise to state renormalization of the localized electrons, the exchange terms lead to an effective Heisenberg interaction of the form

$$\mathcal{H}_{\mathrm{ORKKY}} = -J_{12}\mathbf{S}_1 \cdot \mathbf{S}_2 \propto P_{12}\frac{H_C^2 H_X^2}{\delta^3}P_{12}, \qquad (1.62)$$

which is calculated in fourth-order perturbation theory using the diagram depicted in Figure 1.3.

P_{12} is the projection operator on the two-spin Hilbert space of the two localized spins. The control Hamiltonian

$$\mathcal{H}_{\mathrm{C}} = \sum_{\mathbf{k},\sigma} \frac{\Omega_{\mathbf{k},\sigma}(t)}{2} \mathrm{e}^{-\mathrm{i}\omega P\sigma} c^{\dagger}_{\mathbf{k},-\sigma} h^{\dagger}_{-\mathbf{k},\sigma} + h.c. \qquad (1.63)$$

describes the creation of the virtual excitons by means of an external laser field that is detuned by the energy δ from the continuum states of the host material, or more precisely from the exciton 1s level. $c^{\dagger}_{\mathbf{k},\sigma}$ and $h^{\dagger}_{\mathbf{k},\sigma}$ are electron and hole creation operators, respectively. The RKKY interaction between a localized electron in the quantum dot and the itinerant electrons in the host material is given in second quantization by

$$\mathcal{H}_X = -\frac{1}{V}\sum_{\alpha,\alpha',\mathbf{k},\mathbf{k}'} J(\mathbf{k},\mathbf{k}')\mathbf{S}_i \cdot \mathbf{s}_{\alpha,\alpha'} c^{\dagger}_{\mathbf{k},\alpha} c_{\mathbf{k}',\alpha'}. \qquad (1.64)$$

The predicted exchange interaction $J_{12}(\mathbf{R})$ (see Figure 1.2) can be of the order of 1 meV, which is of the same order as the Heisenberg interaction in the Loss–DiVincenzo scheme.

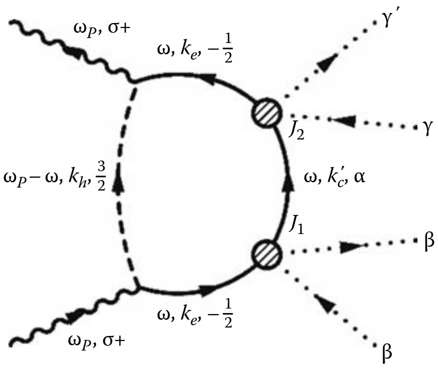

FIGURE 1.3 Effective spin–spin interaction for the localized electrons in the dots 1 and 2 (indicated by dotted lines) induced by a photoexcited electron–hole pair (the solid and dashed lines, respectively). The indices β and γ denote the spin states of the electrons localized in the dots. The photon propagator is depicted by a wavy line. (From Piermarocchi, C. et al., *Phys. Rev. Lett.*, 89, 167402, 2002.)

1.7 Quantum Computing with Molecular Magnets

Shor and Grover demonstrated that a quantum computer can outperform any classical computer in factoring numbers (Shor, 1997) and in searching a database (Grover, 1997) by exploiting the parallelism of quantum mechanics. Recently, the latter has been successfully implemented (Ahn et al., 2000) using Rydberg atoms. Leuenberger and Loss (2001) proposed an implementation of Grover's algorithm using molecular magnets (Friedman et al., 1996; Thomas et al., 1996; Sangregorio et al., 1997; Thiaville and Miltat, 1999; Wernsdorfer et al., 2000); their spin eigenstates make them natural candidates for single-particle systems. It was shown theoretically that molecular magnets can be used to build dense and efficient memory devices based on the Grover algorithm. In particular, one single crystal can serve as a storage unit of a dynamic random access memory device. Fast ESR pulses can be used to decode and read out stored numbers of up to 10^5, with access times as short as 10^{-10} s. This proposal should be feasible using the molecular magnets Fe_8 and Mn_{12}.

Suppose we want to find a phone number in a phone book consisting of $N = 2^n$ entries. Usually it takes $N/2$ queries on average to be successful. Even if the N entries were encoded binary, a classical computer would need approximately $\log_2 N$ queries to find the desired phone number (Grover, 1997). But the computational parallelism provided by the superposition and interference of quantum states enables the Grover algorithm to reduce the search to one single query (Grover, 1997). This query can be implemented in terms of a unitary transformation applied to the single spin of a molecular magnet. Such molecular magnets, forming identical and largely independent units, are embedded in a single crystal so that the ensemble nature of such a crystal provides a natural amplification of the magnetic moment of a single spin. However, for the Grover algorithm to succeed, it is necessary to find ways to generate arbitrary superpositions of spin eigenstates. For spins larger than ½, this turns out to be a highly nontrivial task as spin excitations induced by magnetic dipole transitions in conventional ESR can change the magnetic quantum number m by only ±1. To circumvent such physical limitations, it was proposed to use multifrequency coherent magnetic radiation that allows the controlled generation of arbitrary spin superpositions. In particular, it was shown that by means of advanced ESR techniques, it is possible to coherently populate and manipulate many spin states simultaneously by applying one single pulse of a magnetic a.c. field containing an appropriate number of matched frequencies. This a.c. field creates a nonlinear response of the magnet via multiphoton

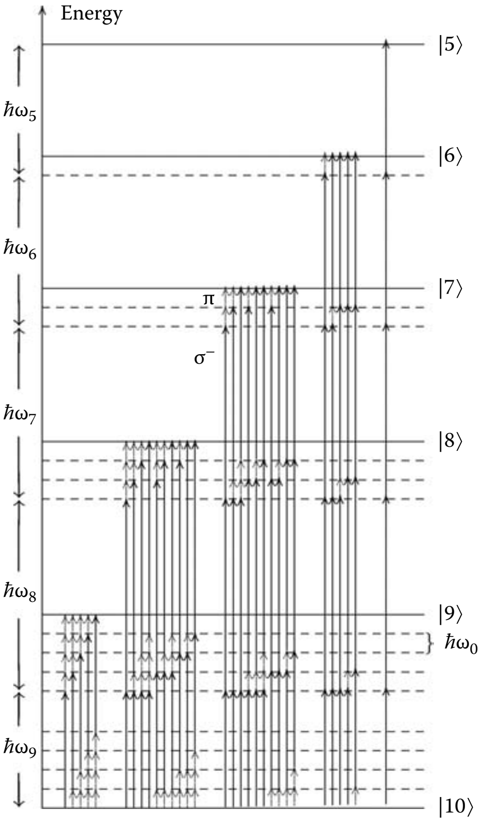

FIGURE 1.4 Feynman diagrams $\mathcal{F}$ that contribute to $S^{(5)}_{m,s}$ for $s = 10$ and $m_0 = 5$ describing transitions (of fifth order in V) in the left well of the spin system (see Figure 1.5). The solid and dotted arrows indicate σ^- and π transitions governed by Equations 1.66 and 1.65, respectively. We note that $S^{(j)}_{m,s} = 0$ for $j < n$, and $S^{(j)}_{m,s} \ll S^{(n)}_{m,s}$ for $j > n$.

absorption processes involving particular sequences of σ and π photons, which allows the encoding and, similarly, the decoding of states. Finally, the subsequent read-out of the decoded quantum state can be achieved by means of pulsed ESR techniques. These exploit the nonequidistance of energy levels, which is typical of molecular magnets.

Molecular magnets have the important advantage that they can be grown naturally as single crystals of up to 10–100 μm length containing about 10^{12}–10^{15} (largely) independent units so that only minimal sample preparation is required. The molecular magnets are described by a single-spin Hamiltonian of the form $\mathcal{H}_{\text{spin}} = \mathcal{H}_{\text{a}} + V + \mathcal{H}_{\text{sp}} + \mathcal{H}_{\text{T}}$ (Leuenberger and Loss, 1999, 2000a,b), where $\mathcal{H}_{\text{a}} = -AS_z^2 - BS_z^4$ represents the magnetic anisotropy ($A \gg B > 0$). The Zeeman term $V = g\mu_{\text{B}}\mathbf{H} \cdot \mathbf{S}$ describes the coupling between the external magnetic field $\mathbf{H}$ and the spin $\mathbf{S}$ of length s. The calculational states are given by the $2s + 1$ eigenstates of $\mathcal{H}_{\text{a}} + g\mu_{\text{B}}H_zS_z$ with eigenenergies $\varepsilon_m = -Am^2 - Bm^4 + g\mu_{\text{B}}H_zm$, $-s \le m \le s$. The corresponding classical anisotropy potential energy $E(\theta) = -As\cos^2\theta - Bs\cos^4\theta + g\mu_{\text{B}}H_zs\cos\theta$ is obtained by the substitution $S_z = s\cos\theta$, where θ is the polar spherical angle. We have introduced the notation $m, m' = m - m'$. By applying a bias field H_z such that $g\mu_{\text{B}}H_z > E_{mm'}$, tunneling can be completely suppressed and thus $\mathcal{H}_{\text{T}}$ can be neglected (Leuenberger and Loss, 1999, 2000a,b). For temperatures below 1 K, transitions due to spin–phonon interactions ($\mathcal{H}_{\text{sp}}$) can also be neglected. In this regime, the level lifetime in Fe_8 and Mn_{12} is estimated to be about $\tau_{\text{d}} = 10^{-7}$ s, limited mainly by hyperfine and/or dipolar interactions (Leuenberger and Loss, 2001).

Since the Grover algorithm requires that all the transition probabilites are almost the same, Leuenberger and Loss (2001) and Leuenberger et al. (2003) propose that all the transition amplitudes between the states $|s\rangle$ and $|m\rangle$, $m = 1, 2, \ldots, s-1$, are of the same order in perturbation V. This allows us to use perturbation theory. A different approach uses the magnetic field amplitudes to adjust the appropriate transition amplitudes (Leuenberger et al., 2002). Both methods work only if the energy levels are not equidistant, which is typically the case in molecular magnets owing to anisotropies. In general, if we choose to work with the states $m = m_0, m_0 + 1, \ldots, s-1$, where $m_0 = 1, 2, \ldots, s-1$, we have to go up to nth order in perturbation, where $n = s - m_0$ is the number of computational states used for the Grover search algorithm (see below) to obtain the first nonvanishing contribution. Figure 1.4 shows the transitions for $s = 10$ and $m_0 = 5$. The nth-order transitions correspond to the nonlinear response of the spin system to strong magnetic fields. Thus, a coherent magnetic pulse of duration T is needed with a discrete frequency spectrum $\{\omega_m\}$, say, for Mn_{12} between 20 and 300 GHz and a single low-frequency 0 around 100 MHz.

The low-frequency field $\mathbf{H}_z(t) = H_0(t)\cos(\omega_0 t)\mathbf{e}_z$, applied along the easy-axis, couples to the spin of the molecular magnet through the Hamiltonian

$$V_{\text{low}} = g\mu_{\text{B}}H_0(t)\cos(\omega_0 t)S_z, \tag{1.65}$$

where $\hbar\omega_0 \ll \varepsilon_{m_0} - \varepsilon_{m_0+1}$ and $\mathbf{e}_z$ is the unit vector pointing along the z-axis. The π photons of V_{low} supply the necessary energy for the resonance condition (see below). They give rise to virtual transitions with $\Delta m = 0$, that is, they do not transfer any angular momentum (see Figure 1.4). The perturbation Hamiltonian for the high-frequency transitions from $|s\rangle$ to virtual states that are just below $|m\rangle$, $m = m_0, \ldots, s-1$, given by the transverse fields $\mathbf{H}_\perp^-(t) = \sum_{m=m_0}^{s-1} H_m(t)[\cos(\omega_m t + \Phi_m)\mathbf{e}_x - \sin(\omega_m t + \Phi_m)\mathbf{e}_y]$, reads

$$\begin{aligned} V_{\text{high}}(t) &= \sum_{m=m_0}^{s-1} g\mu_{\text{B}}H_m(t)[\cos(\omega_m t + \Phi_m)S_x - \sin(\omega_m t + \Phi_m)S_y] \\ &= \sum_{m=m_0}^{s-1} \frac{g\mu_{\text{B}}H_m(t)}{2}\left[e^{i(\omega_m t + \Phi_m)}S_+ + e^{-i(\omega_m t + \Phi_m)}S_-\right], \end{aligned} \tag{1.66}$$

with phases Φ_m (see below), where we have introduced the unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ pointing along the x- and y-axis, respectively. These transverse fields rotate clockwise and thus produce left circularly polarized σ^- photons, which induce only transitions in the left well (see Figure 1.5). In general, absorption (emission) of σ^- photons gives rise to $\Delta m = -1$ ($\Delta m = +1$) transitions, and vice versa in the

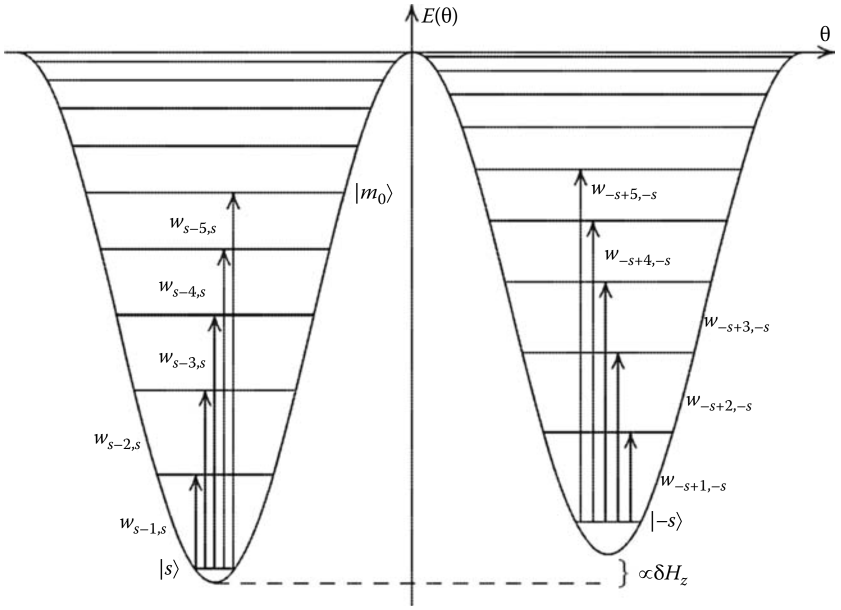

FIGURE 1.5 Double-well potential seen by the spin due to magnetic anisotropies in Mn_{12}. Arrows depict transitions between spin eigenstates driven by the external magnetic field $\mathbf{H}$.

case of σ^+ photons. Anticlockwise rotating magnetic fields of the form $\mathbf{H}_\perp^+(t)=\sum_{m=m_0}^{s-1} H_m(t)[\cos(\omega_m t+\Phi_m)\mathbf{e}_x+\sin(\omega_m t+\Phi_m)\mathbf{e}_y]$ can be used to induce spin transitions only in the right well (see Figure 1.5). In this way, both wells can be accessed independently.

Next we calculate the quantum amplitudes for the transitions induced by the magnetic a.c. fields (see Figure 1.4) by evaluating the S-matrix perturbatively. The jth-order term of the perturbation series of the S-matrix in powers of the total perturbation Hamiltonian $V(t) = V_{\mathrm{low}}(t) + V_{\mathrm{high}}(t)$ is expressed by

$$S_{m,s}^{(j)}=\left(\frac{1}{i\hbar}\right)^j \prod_{k=1}^{j-1}\int_{-\infty}^{\infty}\mathrm{d}t_k\int_{-\infty}^{\infty}\mathrm{d}t_j\Theta(t_k-t_{k+1})$$
$$\times\, U(\infty,t_1)V(t_1)U(t_1,t_2)V(t_2)\ldots V(t_j)U(t_j,-\infty), \tag{1.67}$$

which corresponds to the sum over all Feynman diagrams $\mathcal{F}$ of order j, and where $U(t,t_0)=\mathrm{e}^{-\mathrm{i}(\mathcal{H}_{\mathrm{a}}+g\mu_{\mathrm{B}}\delta H_z)(t-t_0)/\hbar}$ is the free propagator, $\Theta(t)$ is the Heavyside function. The total S-matrix is then given by $S=\sum_{j=0}^{\infty}S^{(j)}$. The high-frequency virtual transition changing m from s to $s-1$ is induced by the frequency $\omega_{s-1} = \omega_{s-1,s} - (n-1)\omega_0$. The other high frequencies ω_m, $m = m_0, \ldots, s-2$, of the high-frequency fields H_m mismatch the level separations by ω_0, that is, $\hbar\omega_m = \varepsilon_m - \varepsilon_{m+1} + \hbar\omega_0$ (see Figure 1.4). As the levels are not equidistant, it is possible to choose the low and high frequencies in such a way that $S_{m,s}^{(j)}=0$ for $j < n$, in which case the resonance condition is not satisfied, that is, energy is not conserved. In addition, the higher-order amplitudes $|S_{m,s}^{(j)}|$ are negligible compared to $|S_{m,s}^{(j)}|$ for $j > n$. Using rectangular pulse shapes, $H_k(t) = H_k$, if $-T/2 < t < T/2$, and 0 otherwise, for $k = 0$ and $k \geq m_0$, one obtains ($m \geq m_0$)

$$S_{m,s}^{(n)}=\sum_{\mathcal{F}}\Omega_m\frac{2\pi}{i}\left(\frac{g\mu_{\mathrm{B}}}{2\hbar}\right)^n\frac{\prod_{k=m}^{s-1}H_k\mathrm{e}^{-\mathrm{i}\Phi_k}H_0^{m-m_0}p_{m,s}(\mathcal{F})}{(-1)^{q\mathcal{F}}q\mathcal{F}!r_s(\mathcal{F})!\omega_0^{n-1}}$$
$$\times\,\delta^{(T)}\left(\omega_{m,s}-\sum_{k=m}^{s-1}\omega_k-(m-m_0)\omega_0\right), \tag{1.68}$$

where $\Omega_m=(m-m_0)!$, is the symmetry factor of the Feynman diagrams $\mathcal{F}$ (see Figure 1.4), $q\mathcal{F}=m-m_0-r_s(\mathcal{F})$, $p_{m,s}(\mathcal{F})=\prod_{k=m}^{s}\langle k|S_z|k\rangle^{r_k(\mathcal{F})}$, $r_k(\mathcal{F})=0,1,2,\ldots,\leq m-m_0$ is the number of π transitions directly above or below the state $|k\rangle$, depending on the particular Feynman diagram $\mathcal{F}$, and $\delta^{(T)}(\omega)=\frac{1}{2\pi}\int_{-T/2}^{+T/2}\mathrm{e}^{\mathrm{i}\omega t}\mathrm{d}t=\sin(\omega T/2)/\pi\omega$ is the delta-function of width $1/T$, ensuring overall energy conservation for $\omega T \gg 1$. The duration T of the magnetic pulses must be shorter than the lifetimes τ_{d} of the states $|m\rangle$ (see Figure 1.5). In general, the magnetic field amplitudes H_k must be chosen in such a way that perturbation theory is still valid and the transition probabilities are almost equal, which is required by the Grover algorithm. According to Leuenberger and Loss (2001), the amplitudes H_k do not differ too much between each other due to the partial destructive interference of the different transition diagrams shown in Figure 1.5. Leuenberger et al. (2002) show that the transition probabilities can be increased by increasing both the magnetic field amplitudes and the detuning energies under the condition that the magnetic field amplitudes remain smaller than the detuning energies. In this way, both high-multiphoton Rabi oscillation frequencies and small quantum computation times can be attained. This makes both methods (Leuenberger and Loss, 2001; Leuenberger et al., 2002) very robust against decoherence sources.

In order to perform the Grover algorithm, one needs the relative phases φ_m between the transition amplitudes $S_{m,s}^{(n)}$, which is determined by $\Phi_m = \Sigma_{k=s-1}^{m+1}\Phi_k + \varphi_m$, where Φ_m are the relative phases between the magnetic fields $H_m(t)$. In this way, it is possible to read-in and decode the desired phases Φ_m for each state $|m\rangle$. The read-out is performed by standard spectroscopy with pulsed ESR, where the circularly polarized radiation can now be incoherent because only the absorption intensity of only one pulse is needed. We emphasize that the entire Grover algorithm (read-in, decoding, read-out) requires three subsequent pulses each of duration T with $\tau_d > T > \omega_0^{-1} > \omega_m^{-1} > \omega_{m,m\pm1}^{-1}$. This gives a "clock-speed" of about 10 GHz for Mn_{12}, that is, the entire process of read-in, decoding, and read-out can be performed within about 10^{-10} s.

The proposal for implementing Grover's algorithm works not only for molecular magnets but for any electron or nuclear spin system with nonequidistant energy levels, as is shown by Leuenberger et al. (2002) for nuclear spins in GaAs semiconductors. Instead of storing information in the phases of the eigenstates $|m\rangle$ (Leuenberger and Loss, 2001), Leuenberger et al. (2002) use the eigenenergies of $|m\rangle$ in the generalized rotating frame for encoding information. The decoding is performed by bringing the delocalized state $(1/\sqrt{n})\Sigma_m |m\rangle$ into resonance with $|m\rangle$ in the generalized rotating frame. Although such spin systems cannot be scaled arbitrarily, large spin s (the larger a spin becomes, the faster it decoheres and the more classical its behavior will be) systems of given s can be used to great advantage in building dense and highly efficient memory devices.

For a first test of the nonlinear response, one can irradiate the molecular magnet with an a.c. field of frequency $\omega_{s-2,s}/2$, which gives rise to a two-photon absorption and thus to a Rabi oscillation between the states $|s\rangle$ and $|s-2\rangle$. For stronger magnetic fields, it is in principle possible to generate superpositions of Rabi oscillations between the states $|s\rangle$ and $|s-1\rangle$, $|s\rangle$ and $|s-2\rangle$, $|s\rangle$ and $|s-3\rangle$, and so on (see also Leuenberger et al., 2002).

1.8 Semiconductor Quantum Dots

In the following, we mainly focus on quantum dots made of III–V semiconductor compounds with zincblende structure, like GaAs or InAs. The electronic bandstructure of a three-dimensional semiconductor with zincblende structure is illustrated in Figure 1.6. The bands are parabolic close to their extrema, which are all located at the Γ point. The conduction (c) states have orbital s symmetry and are spin degenerate. The valence (v) band consists of three subbands: the heavy-hole (hh), the light-hole (lh), and the split-off (so) band. The v-band states have orbital p symmetry. The bottom of the c band and the top of the v band are split by the band-gap energy E_{gap}. The v-band states with different j ($j = \frac{1}{2}$ for the so-band, $j = \frac{3}{2}$ for the hh and lh band) are split by Δ_{so} in energy due to spin–orbit interaction.

The hh states have the angular momentum projections $J_z = \pm\frac{3}{2}$ and the lh states $J_z = \pm\frac{1}{2}$. For finite electron wavevectors $k \neq 0$, and the hh and lh subbands split into two branches according to the different curvatures of the energy dispersion, which implies different effective masses of heavy and light holes. The v-band states with spin can be written in terms of the orbital angular momentum basis by using the Clebsch–Gordon coefficients which gives us

$$\text{Heavy hole}\begin{Bmatrix} \left|\frac{3}{2},\frac{3}{2}\right\rangle = |1,1\rangle|\uparrow\rangle \\ \left|\frac{3}{2},-\frac{3}{2}\right\rangle = |1,-1\rangle|\downarrow\rangle \end{Bmatrix}, \tag{1.69}$$

$$\text{Light hole}\begin{Bmatrix} \left|\frac{3}{2},\frac{1}{2}\right\rangle = \frac{1}{\sqrt{3}}|1,1\rangle|\downarrow\rangle + \sqrt{\frac{2}{3}}|1,0\rangle|\uparrow\rangle \\ \left|\frac{3}{2},-\frac{1}{2}\right\rangle = \sqrt{\frac{2}{3}}|1,0\rangle|\downarrow\rangle + \frac{1}{\sqrt{3}}|1,-1\rangle|\uparrow\rangle \end{Bmatrix}, \tag{1.70}$$

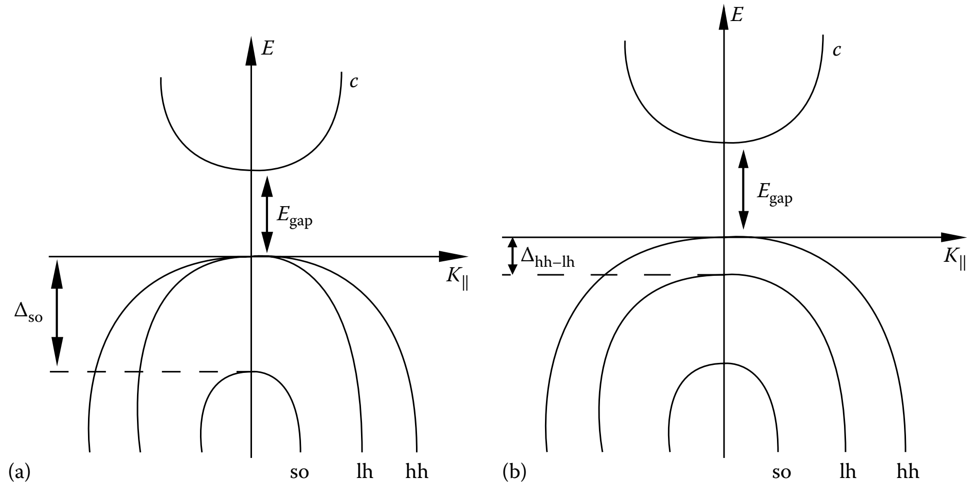

FIGURE 1.6 Electronic band structure in the vicinity of the Γ point for (a) a three-dimensional crystal and (b) a quantum well. The conduction and valence bands are shown as a function of the wavevector.

$$\text{Split off}\ \left\{\begin{array}{c} \left|\tfrac{1}{2},\tfrac{1}{2}\right\rangle = \tfrac{1}{\sqrt{3}}\left|1,0\right\rangle\left|\uparrow\right\rangle + \sqrt{\tfrac{2}{3}}\left|1,1\right\rangle\left|\downarrow\right\rangle \\ \left|\tfrac{1}{2},-\tfrac{1}{2}\right\rangle = \sqrt{\tfrac{2}{3}}\left|1,-1\right\rangle\left|\uparrow\right\rangle + \tfrac{1}{\sqrt{3}}\left|1,0\right\rangle\left|\downarrow\right\rangle \end{array}\right\}. \qquad (1.71)$$

Quantum confinement along the crystal axis quantizes the wavevector component, consequently the hh and lh states of the lowest subband are split by an energy $\Delta_{\text{hh-lh}}$ at the Γ point. Uniaxial strain in the semiconductor crystal can also lift the degeneracy of the heavy and light holes, and thus define the spin quantization axis. If we have a spherically symmetric quantum dot known as *colloidal quantum dots*, then we can have degeneracy between the heavy and light hole band, i.e., $\Delta_{\text{hh-lh}} = 0$.

Via photon absorption, an electron in a *v*-band state can be excited to a *c*-band state. Such interband transitions are determined by optical selection rules. The source or the optical transition rules are due to spin–orbit interaction. The electron–hole pair created with an interband transition is called an exciton. The electron and hole of an exciton form a bound state due to the Coulomb interaction, similar to that of a hydrogen atom. We refer to the system of two bound excitons as a biexciton.

1.8.1 Classical Faraday Effect

Michael Faraday first observed the effect in 1845 when studying the influence of a magnetic field on plane-polarized light waves. Light waves vibrate in two planes at right angles to one another, and passing ordinary light through certain substances eliminates the vibration in one plane. He discovered that the plane of vibration is rotated when the light path and the direction of the applied magnetic field are parallel. In particular, a linearly polarized wave can be decomposed into right and left circularly polarized waves where each wave propagates with different speeds. The waves can be considered to recombine upon emergence from the medium; however, owing to the difference in propagation speed, they do so with a net phase offset, resulting in a rotation of the angle of linear polarization.

The Faraday effect occurs in many solids, liquids, and gases. The magnitude of the rotation depends upon the strength of the magnetic field, the nature of the transmitting substance, and Verdets constant, which is a property of the transmitting substance, its temperature, and the frequency of light. The relation between the angle of rotation of the polarization and the magnetic field in a diamagnetic material is

$$\beta = \mathcal{V}Hd, \qquad (1.72)$$

where

β is the angle of rotation
$\mathcal{V}$ is the Verdet constant for the material
H is the magnitude of the applied magnetic field
d is the length of the path where the light and magnetic field interact (see Figure 1.7)

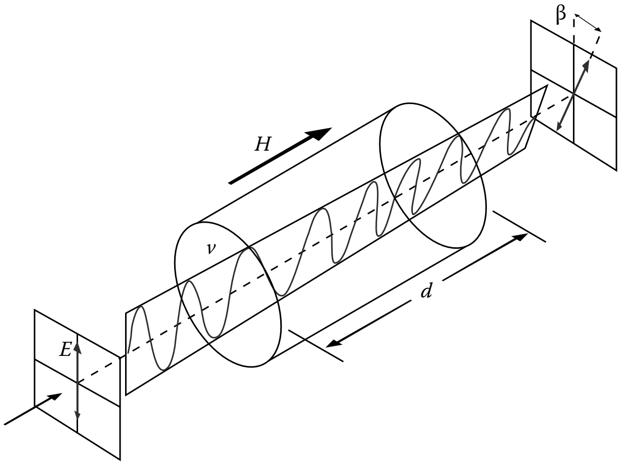

FIGURE 1.7 The figure shows how the polarization of a linearly polarized beam of light rotates when it goes through a material exposed to an external magnetic field.

A positive Verdet constant corresponds to an anticlockwise rotation when the direction of propagation is parallel to the magnetic field and to a clockwise rotation when the direction of propagation is antiparallel. The Faraday effect is used in spintronics research to study the polarization of electron spins in semiconductors.

1.8.2 Quantum Faraday Effect

The Faraday effect is expected to emerge in low dimensional systems such as semiconductor quantum dots in which the spin states of the electron in the conduction band and the light and heavy hole in the valence band provide a system where different circular polarizations of light couple differently during the process of virtual absorption. The quantum analogue of the Faraday effect does not require an external magnetic field because it is created by selection rules (one circular polarization interacts with the heavy hole band while the other circular polarization interacts with the light hole band) and by the Pauli exclusion principle (the absorption of a right polarized wave is excluded because the allowed transition state between bands have the same spin) (Leuenberger et al., 2005b). In particular, we will be interested in the quantum Faraday effect in a semiconductor colloidal two-level quantum dot system. We can have a two-level system in a colloidal quantum dot where the heavy hole and the light hole bands are degenerate at the Γ point. This two-level system is achieved by the valence and the conduction band under certain assumptions: (1) the split-off band can be ignored since typical split-off energies are around 10^2 meV, thus bringing the energy level out of resonance with the single photon; (2) under the appropriate doping and thermal conditions, it can be assumed that the top of the valence band is filled with four electrons, while there is one excess electron in the conduction band; and (3) the energy of the electromagnetic wave is taken to be slightly below the effective band-gap energy, so that the transition from the top of the valence band to the bottom of the conduction band is the strongest transition by far (Leuenberger, 2006).

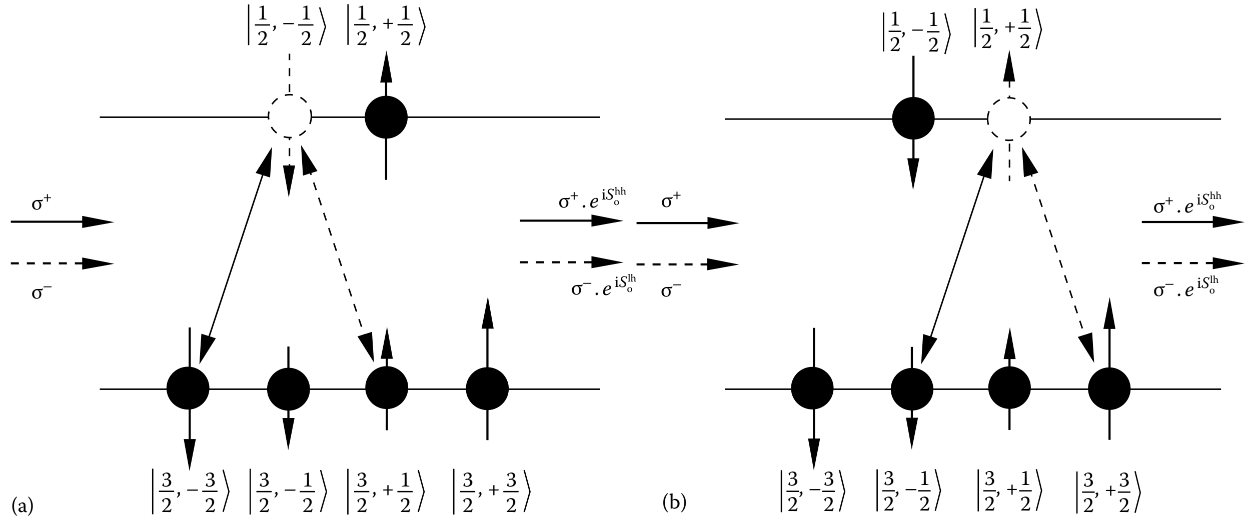
$\left|\frac{1}{2},-\frac{1}{2}\right\rangle$
$\left|\frac{1}{2},+\frac{1}{2}\right\rangle$
σ^+
σ^-
$\sigma^+ . e^{iS_o^{hh}}$
$\sigma^- . e^{iS_o^{lh}}$
$\left|\frac{3}{2},-\frac{3}{2}\right\rangle$
$\left|\frac{3}{2},-\frac{1}{2}\right\rangle$
$\left|\frac{3}{2},+\frac{1}{2}\right\rangle$
$\left|\frac{3}{2},+\frac{3}{2}\right\rangle$
(a)
(b)

FIGURE 1.8 The solid circles represent filled states and dashed circles represent empty states. The electron in the conduction band interacts with the right or left circularly polarized electromagnetic wave allowing virtual transitions between the conduction and valence band states. After the interaction, the RCP and LCP electromagnetic wave acquire different phases, which produce the rotation of the incident wave.

Interestingly, the idea of the conditional single-photon Faraday rotation first developed by Leuenberger et al. (2005b) and already patented by the authors has been copied (Hu et al., 2008), which demonstrates the importance of this scheme.

We now turn our focus to the optical selection rules. In Figure 1.8 we show how different circular polarizations of the photon interact with the two-level system. The conduction band state has one electron with spin either up ($|\frac{1}{2},\frac{1}{2}\rangle$) or down ($|\frac{1}{2},-\frac{1}{2}\rangle$). The presence of a spin up (down) electron in the conduction band forces the virtual transitions of the incident electromagnetic wave to couple $|\frac{3}{2},-\frac{3}{2}\rangle$ ($|\frac{3}{2},-\frac{1}{2}\rangle$) or $|\frac{3}{2},\frac{1}{2}\rangle$ ($|\frac{3}{2},\frac{3}{2}\rangle$) heavy hole and light hole states depending if it is a right or left circularly polarized photon. Examination of the coefficients given in Equations 1.69 and 1.70 shows that the matrix elements of the heavy hole band and the matrix element involving the light hole band are different, leading to different phases and causing a change in the refractive index, which induces the Faraday rotation effect. The Faraday rotation is clockwise if the spin is up, and counterclockwise if the spin is down. The complex refractive index $(\eta)\tilde{}$ can be calculated as follows:

$$\tilde{\eta}^2 = 1 - \frac{e^2 N}{\epsilon_0 m_e} \sum_{i,j} \frac{f_{ji}}{(\omega^2 - \omega_{ji}^2) + i\gamma\omega}, \tag{1.73}$$

where

$|e|$ is the electron's charge
N is the electron number density
ϵ_0 is the permittivity of free space
m_e is the mass of the electron
ω_{ji} is the transition frequency from the ith state to the jth state
γ is the full-width at half maximum (assumed to be much smaller than the detuning energy)
ω is the frequency of the incident electromagnetic wave
f_{ji} is the oscillator strength given by

$$f_{ji} = \frac{2}{m\hbar\omega_{ji}} |\langle j|(-e)\vec{r}\cdot\vec{E}|i\rangle|^2. \tag{1.74}$$

This quantum Faraday rotation can be used to measure the spin polarization of charge carriers in quantum wells (Kikkawa et al., 1997; Kikkawa and Awschalom, 1998, 1999). Recently, an experiment based on the quantum Faraday rotation was able to measure the spin state of a single excess electron inside a quantum dot (Berezovsky et al., 2006), which is crucial for the read-out of a qubit in a quantum computing scheme (see below).

1.9 Single-Photon Faraday Rotation

The single-photon Faraday effect (SPFE) is similar to the quantum Faraday effect but involves only the nonresonant interaction of a single photon with the quantum dot two-level system and can result in an entanglement of the photon with the spin of the extra electron (Leuenberger et al., 2005b; Leuenberger, 2006). The SPFE can be described by the Jaynes–Cummings Hamiltonian

$$\begin{aligned}\mathcal{H} = {} & \hbar\omega\left(a^\dagger_{\sigma^+}a_{\sigma^+} + a^\dagger_{\sigma^-}a_{\sigma^-}\right) + \hbar\omega_{\mathrm{hh}}\sigma_{\frac{3}{2}v} + \hbar\omega_{\mathrm{hh}}\sigma_{-\frac{3}{2}v} \\ & + \hbar\omega_{\mathrm{lh}}\sigma_{\frac{1}{2}v} + \hbar\omega_{\mathrm{lh}}\sigma_{-\frac{1}{2}v} + \hbar\omega_e\sigma_{\frac{1}{2}c} + \hbar\omega_e\sigma_{-\frac{1}{2}c} + \hbar g_{\frac{3}{2}v,\frac{1}{2}c} \\ & \times\left(a^\dagger_{\sigma^-}\sigma_{\frac{3}{2}v,\frac{1}{2}c} + a_{\sigma^-}\sigma_{\frac{1}{2}c,\frac{3}{2}v} + a^\dagger_{\sigma^+}\sigma_{-\frac{3}{2}v,-\frac{1}{2}c} + a_{\sigma^+}\sigma_{-\frac{1}{2}c,-\frac{3}{2}v}\right) \\ & + \hbar g_{\frac{1}{2}v,\frac{1}{2}c}\left(a^\dagger_{\sigma^-}\sigma_{\frac{1}{2}v,-\frac{1}{2}c} + a_{\sigma^-}\sigma_{-\frac{1}{2}c,\frac{1}{2}v} + a^\dagger_{\sigma^+}\sigma_{-\frac{1}{2}v,-\frac{1}{2}c} + a_{\sigma^+}\sigma_{\frac{1}{2}c,-\frac{1}{2}v}\right).\end{aligned} \tag{1.75}$$

Equation 1.76 describes how the right circularly polarized (σ^+) and left circularly polarized (σ^-) components of a linearly polarized photon field interact with the degenerate quantum dot levels in a cavity. It is assumed that the spin of the excess electron in the conduction band is up ($\uparrow$). The evolution of the system will be given by the state vector

$$\begin{aligned}|\Psi(t)\rangle = \frac{1}{\sqrt{2}}\Big[&C_{\uparrow\mathrm{hh}}(t)|\uparrow,\mathrm{hh}\rangle + C_{\uparrow\sigma^+}(t)|\uparrow,\sigma_z^+\rangle \\ &+ C_{\uparrow\mathrm{lh}}(t)|\uparrow,\mathrm{lh}\rangle + C_{\uparrow\sigma^-}(t)|\uparrow,\sigma_z^-\rangle\Big],\end{aligned} \tag{1.76}$$

where $|\uparrow,\mathrm{hh}\rangle$ and $|\uparrow,\mathrm{lh}\rangle$ are the states in which the quantum dot is in an excited state with a heavy-hole exciton or a light-hole exciton, and the spin of the excess electron in the conduction band is up. $|\uparrow,\sigma_z^+\rangle$ and $|\uparrow,\sigma_z^-\rangle$ are the states in which the quantum dot is in the ground state with the photon present in the cavity, and the spin of the extra electron in the conduction band is up.

In order to solve for the phase accumulated for RCP and LCP components of the linearly polarized photon during the interaction with the quantum dot in the nanocavity, we must find the time evolution of the probability amplitudes $C_{\uparrow\sigma+}(t)$ and $C_{\uparrow\sigma-}(t)$, respectively. Assuming the following initial conditions, $C_{\uparrow\mathrm{hh}}(0) = 0$, $C_{\uparrow\sigma+}(0) = 1$, $C_{\uparrow\mathrm{lh}}(0) = 0$, and $C_{\uparrow\sigma-}(0) = 1$, then the probability amplitudes of interest are given by (Seigneur et al., 2008)

$$C_{\uparrow\sigma^+}(t) = \mathrm{e}^{-\mathrm{i}\Delta t/2}\left[\cos\left(\frac{\Omega_{\frac{3}{2}}t}{2}\right) + \frac{\mathrm{i}\Delta}{\Omega_{\frac{3}{2}}}\sin\left(\frac{\Omega_{\frac{3}{2}}t}{2}\right)\right] \tag{1.77}$$

$$C_{\uparrow\sigma^-}(t) = \mathrm{e}^{-\mathrm{i}\Delta t/2}\left[\cos\left(\frac{\Omega_{\frac{1}{2}}t}{2}\right) + \frac{\mathrm{i}\Delta}{\Omega_{\frac{1}{2}}}\sin\left(\frac{\Omega_{\frac{1}{2}}t}{2}\right)\right], \tag{1.78}$$

where $\Omega_{\frac{3}{2}}^2 = \Delta^2 + 4g_{\frac{3}{2}v,\frac{1}{2}c}^2$ and $\Omega_{\frac{1}{2}}^2 = \Delta^2 + 4g_{\frac{1}{2}v,\frac{1}{2}c}^2$, with $g_{\frac{3}{2}v,\frac{1}{2}c}$ and $g_{\frac{1}{2}v,\frac{1}{2}c}$ being the coupling strength involving the heavy-hole electron and the light-hole electron, respectively, with $\Delta = \omega - \omega_{\mathrm{ph}}$ being the detuning frequency and $g_{\frac{3}{2}v,\frac{1}{2}c} = \sqrt{3}g_{\frac{1}{2}v,\frac{1}{2}c}$.

Rewriting the complex coefficients given in Equations 1.77 and 1.78 in the form $C_{\uparrow\sigma^+}(t) = |C_{\uparrow\sigma^+}(t)|\exp[\mathrm{i}S_0^{\mathrm{hh}}]$ and $C_{\uparrow\sigma^-}(t) = |C_{\uparrow\sigma^-}(t)|\exp[\mathrm{i}S_0^{\mathrm{lh}}]$, we obtain an expression for the phase accumulated during the interaction of the right and left circularly polarized component with the heavy-hole and the light-hole band, i.e.,

$$S_0^{\mathrm{hh}} = \tan^{-1}\left[\frac{\Delta}{\Omega_{\frac{3}{2}}}\tan\left(\frac{\Omega_{\frac{3}{2}}t}{2}\right)\right] \quad \text{and} \quad S_0^{\mathrm{lh}} = \tan^{-1}\left[\frac{\Delta}{\Omega_{\frac{1}{2}}}\tan\left(\frac{\Omega_{\frac{1}{2}}t}{2}\right)\right]. \tag{1.79}$$

To determine the entanglement of the single photon with the electron, we know that the electron–photon state after the interaction reads

$$|\psi_{\mathrm{ep}}^{(1)}\rangle = \mathrm{e}^{\mathrm{i}S_0^{\mathrm{hh}}}|\psi_{\mathrm{hh}}^{(1)}\rangle + \mathrm{e}^{\mathrm{i}S_0^{\mathrm{lh}}}|\psi_{\mathrm{lh}}^{(1)}\rangle, \tag{1.80}$$

where

$$|\psi_{\mathrm{hh}}^{(1)}\rangle = \frac{1}{\sqrt{2}}\left(\alpha|\uparrow\rangle|\sigma_{(z)}^+\rangle + \beta|\downarrow\rangle|\sigma_{(z)}^-\rangle\right) \tag{1.81}$$

represents the photon scattering off a heavy hole, and

$$|\psi_{\mathrm{lh}}^{(1)}\rangle = \frac{1}{\sqrt{2}}\left(\alpha|\uparrow\rangle|\sigma_{(z)}^-\rangle + \beta|\downarrow\rangle|\sigma_{(z)}^+\rangle\right) \tag{1.82}$$

represents the photon scattering off a light hole. Let

$$|\varphi\rangle = \cos\varphi|\leftrightarrow\rangle + \sin\varphi|\updownarrow\rangle \tag{1.83}$$

be the photon state with linear polarization that is rotated by φ around the z-axis with respect to the state $|\leftrightarrow\rangle$ of linear polarization in x direction. Changing to circular polarization states, Equation 1.83 can be rewritten as

$$\begin{aligned}|\varphi\rangle &= \frac{\cos\varphi}{\sqrt{2}}\left(|\sigma_{(z)}^+\rangle + |\sigma_{(z)}^-\rangle\right) + \frac{\sin\varphi}{\mathrm{i}\sqrt{2}}\left(|\sigma_{(z)}^+\rangle - |\sigma_{(z)}^-\rangle\right) \\ &= \frac{1}{\sqrt{2}}\left(\mathrm{e}^{-\mathrm{i}\varphi}|\sigma_{(z)}^+\rangle + \mathrm{e}^{\mathrm{i}\varphi}|\sigma_{(z)}^-\rangle\right).\end{aligned} \tag{1.84}$$

Using the representation of this linear polarization with $\varphi = \pm(S_0^{\mathrm{hh}} - S_0^{\mathrm{lh}})/2 = \pm S_0/2$, we find that Equation 1.10 becomes

$$\begin{aligned}|\psi_{\mathrm{ep}}^{(1)}(T)\rangle &= \alpha|\uparrow\rangle\left(\mathrm{e}^{\mathrm{i}S_0^{\mathrm{hh}}}|\sigma_{(z)}^+\rangle + \mathrm{e}^{\mathrm{i}S_0^{\mathrm{lh}}}|\sigma_{(z)}^-\rangle\right)\Big/\sqrt{2} \\ &\quad + \beta|\downarrow\rangle\left(\mathrm{e}^{\mathrm{i}S_0^{\mathrm{lh}}}|\sigma_{(z)}^+\rangle + \mathrm{e}^{\mathrm{i}S_0^{\mathrm{hh}}}|\sigma_{(z)}^-\rangle\right)\Big/\sqrt{2} \\ &= \mathrm{e}^{\mathrm{i}\left(S_0^{\mathrm{hh}} + S_0^{\mathrm{lh}}\right)/2}\left(\alpha|\uparrow\rangle|-S_0/2\rangle + \beta|\downarrow\rangle|+S_0/2\rangle\right).\end{aligned} \tag{1.85}$$

From this equation, we draw the conclusion that the accumulated CPHASE shift S_0 corresponds to a single-photon Faraday rotation around the z-axis by the angle $S_0/2$ due to a single spin. Let j be an integer. For $S_0/2 = (2j + 1)\pi/4$, the electron–photon state $|\psi_{\mathrm{ep}}^{(1)}\rangle$ is maximally entangled. If j is even,

$$|\psi_{\mathrm{ep}}^{(1)}\rangle = \alpha|\uparrow\rangle|\nwarrow\!\!\searrow\rangle + \beta|\downarrow\rangle|\swarrow\!\!\nearrow\rangle. \tag{1.86}$$

If j is odd,

$$|\psi_{\mathrm{ep}}^{(1)}\rangle = \alpha|\uparrow\rangle|\swarrow\!\!\nearrow\rangle + \beta|\downarrow\rangle|\nwarrow\!\!\searrow\rangle. \tag{1.87}$$

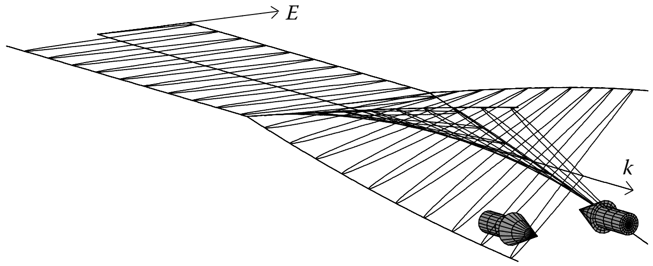

FIGURE 1.9 The conditional Faraday rotation of the linear polarization of the photon depends on the spin state $|\psi_e\rangle = \alpha\,|\uparrow\rangle + \beta\,|\downarrow\rangle$. If the spin state is up (down), the linear polarization turns (counter) clockwise. The photon and the spin get maximally entangled if the Faraday rotation angle is $S_0/2 = \pi/4$. This entanglement between photon polarization and electron spin can be used as an optospintronic link to transfer quantum information from electron spins to photons and back.

For $S_0/2 = 2j\pi/4$, the electron–photon state $|\psi_{\mathrm{ep}}^{(1)}\rangle$ is not entangled. If j is even,

$$|\psi_{\mathrm{ep}}^{(1)}(T)\rangle = \alpha\,|\uparrow\rangle\,|\leftrightarrow\rangle + \beta\,|\downarrow\rangle\,|\leftrightarrow\rangle. \tag{1.88}$$

If j is odd,

$$|\psi_{\mathrm{ep}}^{(1)}(T)\rangle = \alpha\,|\uparrow\rangle\,|\updownarrow\rangle - \beta\,|\downarrow\rangle\,|\updownarrow\rangle. \tag{1.89}$$

This means the entanglement oscillates continually between 0 and 1 if no decoherence is present. The entanglement process between the spin of the electron and the polarization of the photon during the conditional Faraday rotation is visualized in Figure 1.9.

1.9.1 Quantifying the EPR Entanglement

Rewriting Equation 1.85 in the basis states of the linear polarization in x and y directions yields

$$\left|\psi_{\mathrm{ep}}^{(1)}\right\rangle = \alpha\,|\uparrow\rangle\left(\cos\frac{S_0}{2}\,|\leftrightarrow\rangle - \sin\frac{S_0}{2}\,|\updownarrow\rangle\right) + \beta\,|\downarrow\rangle\left(\cos\frac{S_0}{2}\,|\leftrightarrow\rangle + \sin\frac{S_0}{2}\,|\updownarrow\rangle\right). \tag{1.90}$$

The entanglement between the photon and the spin of the excess electron can be calculated by means of the von Neumann entropy (Bennett et al., 1996):

$$E[\psi_{\mathrm{ep}}^{(1)}] = -\mathrm{Tr}_{\mathrm{e}}[\mathrm{Tr}_{\mathrm{p}}\,\rho_{\mathrm{ep}}\log_2(\mathrm{Tr}_{\mathrm{p}}\,\rho_{\mathrm{ep}})] \tag{1.91}$$

or the normalized linear entropy (Silberfarb and Deutsch, 2004)

$$E_{\mathrm{L}}[\psi_{\mathrm{ep}}^{(1)}] = 2[1 - \mathrm{Tr}_{\mathrm{e}}(\mathrm{Tr}_{\mathrm{p}}\,\rho_{\mathrm{ep}})^2]. \tag{1.92}$$

We will use the linear entropy in our calculation. The density matrix is given by

$$\rho_{\mathrm{ep}} = \begin{pmatrix} \rho_{\uparrow\uparrow} & \rho_{\uparrow\downarrow} \\ \rho_{\downarrow\uparrow} & \rho_{\downarrow\downarrow} \end{pmatrix}, \tag{1.93}$$

where

$$\rho_{\uparrow\uparrow} = \begin{pmatrix} |\alpha|^2\cos^2\frac{S_0}{2} & -2|\alpha|^2\sin S_0 \\ -2|\alpha|^2\sin S_0 & |\alpha|^2\sin^2\frac{S_0}{2} \end{pmatrix}, \tag{1.94}$$

$$\rho_{\uparrow\downarrow} = \begin{pmatrix} \alpha\beta^*\cos^2\frac{S_0}{2} & 2\alpha\beta^*\sin S_0 \\ -2\alpha\beta^*\sin S_0 & -\alpha\beta^*\sin^2\frac{S_0}{2} \end{pmatrix}, \tag{1.95}$$

$$\rho_{\downarrow\uparrow} = \begin{pmatrix} \alpha^*\beta\cos^2\frac{S_0}{2} & -2\alpha^*\beta\sin S_0 \\ 2\alpha^*\beta\sin S_0 & -\alpha^*\beta\sin^2\frac{S_0}{2} \end{pmatrix}, \tag{1.96}$$

$$\rho_{\downarrow\downarrow} = \begin{pmatrix} |\beta|^2\cos^2\frac{S_0}{2} & 2|\beta|^2\sin S_0 \\ 2|\beta|^2\sin S_0 & |\beta|^2\sin^2\frac{S_0}{2} \end{pmatrix}, \tag{1.97}$$

and thus by $\rho_{\mathrm{ep}} =$

$$\begin{pmatrix} |\alpha|^2\cos^2\frac{S_0}{2} & -2|\alpha|^2\sin S_0 & \alpha\beta^*\cos^2\frac{S_0}{2} & 2\alpha\beta^*\sin S_0 \\ -2|\alpha|^2\sin S_0 & |\alpha|^2\sin^2\frac{S_0}{2} & -2\alpha\beta^*\sin S_0 & -\alpha\beta^*\sin^2\frac{S_0}{2} \\ \alpha^*\beta\cos^2\frac{S_0}{2} & -2\alpha^*\beta\sin S_0 & |\beta|^2\cos^2\frac{S_0}{2} & 2|\beta|^2\sin S_0 \\ 2\alpha^*\beta\sin S_0 & -\alpha^*\beta\sin^2\frac{S_0}{2} & 2|\beta|^2\sin S_0 & |\beta|^2\sin^2\frac{S_0}{2} \end{pmatrix} \tag{1.98}$$

in the basis $|\uparrow\rangle\,|\leftrightarrow\rangle$; $|\uparrow\rangle\,|\updownarrow\rangle$; $|\downarrow\rangle\,|\leftrightarrow\rangle$; $|\downarrow\rangle\,|\updownarrow\rangle$. Since the measures for entanglement must be independent of the chosen basis, all of the measures involve the trace of ρ_{ep}. Taking the trace over the photon states yields

$$\mathrm{Tr}_{\mathrm{p}}\,\rho_{\mathrm{ep}} = \begin{pmatrix} |\alpha|^2 & \alpha\beta^*\cos S_0 \\ \alpha^*\beta\cos S_0 & |\beta|^2 \end{pmatrix}. \tag{1.99}$$

So we obtain the entanglement

$$E_{\mathrm{L}}[\psi_{\mathrm{ep}}^{(1)}] = 2\left(1 - |\alpha|^4 - |\beta|^4 - 2|\alpha|^2|\beta|^2\cos^2 S_0\right). \tag{1.100}$$

With the parametrization $\alpha = \cos\frac{\eta}{2}e^{-i\chi/2}$, $\beta = \sin\frac{\eta}{2}e^{-i\chi/2}$ we obtain

$$E_{\mathrm{L}}[\psi_{\mathrm{ep}}^{(1)}] = 2\left(1 - \cos^4\frac{\eta}{2} - \sin^4\frac{\eta}{2} - \frac{1}{2}\sin^2\eta\cos^2 S_0\right). \tag{1.101}$$

This result is plotted in Figure 1.10.

1.9.2 Single Photon Faraday Effect and GHZ Quantum Teleportation

The theory of quantum teleportation was first developed by Bennett et al. (1993) and later experimentally verified by Bouwmeester et al. (1997). This scheme relies on EPR pairs,

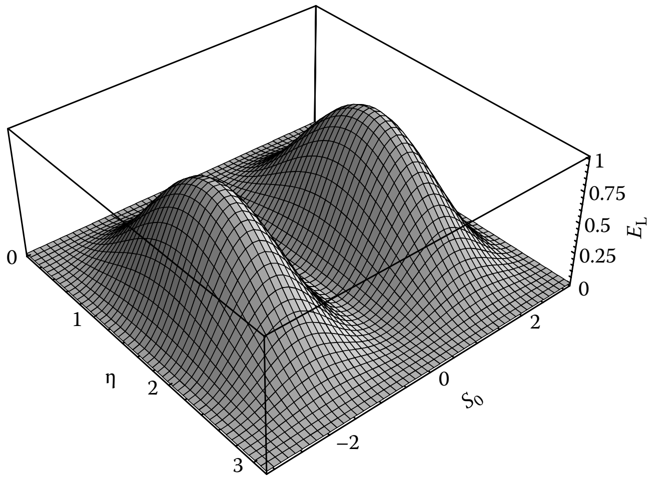

FIGURE 1.10 Entanglement of the photon with the spin of the excess electron.

i.e., pairs of entangled particles, which can be entangled in, for example, the polarization degree of freedom in the case of photons or the spin degree of freedom in the case of electrons or holes. Leuenberger et al. (2005b) show that all the three particles that take part in the teleportation process are entangled in a so-called Greenberger–Horne–Zeilinger (GHZ) state (Greenberger et al., 1989, 1990).

We will perform now a GHZ teleportation. In this GHZ quantum teleportation, the photon (qubit 2) interacts first with the electron spin of the destination (qubit 3) and then with the electron spin of the origin (qubit 1). We identify qubit 1, 2, and 3 with the qubits used by Bennett et al. (1993). Qubit 1 is with Alice, qubit 2 with Charlie, and qubit 3 with Bob. We start from the state $|\Psi^{(1)}_{\mathrm{pe}'}(t_A)\rangle = |{\leftrightarrow}\rangle\,|{\leftarrow}'\rangle$. After interaction in Bob's microcavity, we obtain

$$
\begin{aligned}
|\Psi^{(1)}_{\mathrm{pe}'}(t_{\mathrm{C}})\rangle &= \frac{1}{\sqrt{2}}\left(|\sigma_z^{+}\rangle\,|\uparrow'_y\rangle + |\sigma^{-}_{(z)}\rangle\,|\downarrow'_y\rangle\right) \\
&= \frac{1}{\sqrt{2}}\left(|⤡\rangle\,|⤢\rangle|\downarrow'\rangle\right).
\end{aligned} \tag{1.102}
$$

This hybrid photon–spin entangled state corresponds to the shared EPR state of qubit 2 and 3 in the original version of the teleportation (EPR teleportation). Note that there is no way to distinguish between a single-spin or a single-photon Faraday rotation because we do not have any preferred basis defined by α and β. We have not yet used qubit 1 at all. Alice's photon can be stored for as long as it maintains its entanglement with Bob's spin. This step is like the distribution of the EPR pair in EPR teleportation. Instead of performing now a Bell measurement on qubit 1 and 2 to complete the EPR teleportation, we let the photon (qubit 2) interact with Alice's spin (qubit 1), giving rise to a GHZ state in the hybrid spin–photon–spin system. After interaction of the photon in Alice's microcavity, we obtain

$$
\begin{aligned}
|\psi^{(1)}_{\mathrm{epe}'}(t_{\mathrm{C}}+T)\rangle = \frac{1}{\sqrt{2}}\Big[\alpha\,|\uparrow\rangle\Big(&e^{iS_0^{\mathrm{hh}}}\,|\sigma^{+}_{(z)}\rangle\,|\uparrow'_y\rangle \\
&+e^{iS_0^{\mathrm{lh}}}\,|\sigma^{-}_{(z)}\rangle\,|\downarrow'_y\rangle\Big) \\
&+\beta\,|\downarrow\rangle\Big(e^{iS_0^{\mathrm{lh}}}\,|\sigma^{+}_{(z)}\rangle\Big|\uparrow'_y\Big\rangle \\
&+e^{iS_0^{\mathrm{hh}}}\,|\sigma^{-}_{(z)}\rangle\,|\downarrow'_y\rangle\Big)\Big].
\end{aligned} \tag{1.103}
$$

Now we change to the S_z representation of Bob's spin and to the linear polarization representation of Charlie's photon, which yields ($t_{\mathrm{C}} = 0$)

$$
\begin{aligned}
|\psi^{(1)}_{\mathrm{epe}'}(T)\rangle = \frac{e^{i\left(S_0^{\mathrm{hh}}+S_0^{\mathrm{lh}}\right)/2}}{\sqrt{2}}\big(&\alpha|\uparrow\rangle\,|-S_0/2-\pi/4\rangle|\uparrow'\rangle \\
&+\alpha\,|\uparrow\rangle|-S_0/2+\pi/4\rangle|\downarrow'\rangle \\
&+\beta\,|\downarrow\rangle|+S_0/2-\pi/4\rangle|\uparrow'\rangle \\
&+\beta\,|\downarrow\rangle|+S_0/2+\pi/4\rangle|\downarrow'\rangle.
\end{aligned} \tag{1.104}
$$

Choosing $S_0 = \pi/2$, we obtain

$$
\begin{aligned}
|\psi^{(1)}_{\mathrm{epe}'}(T)\rangle = \frac{1}{\sqrt{2}}\Big[&|{\updownarrow}\rangle\Big(-\alpha\,|\uparrow\rangle\,|\uparrow'\rangle + \beta|\downarrow\rangle\,|\downarrow'\rangle\Big) \\
&+|{\leftrightarrow}\rangle\Big(\alpha\,|\uparrow\rangle\,|\downarrow'\rangle + \beta|\downarrow\rangle\,|\uparrow'\rangle\Big)\Big].
\end{aligned} \tag{1.105}
$$

[From this equation, it becomes obvious that we can produce all the four Bell states between qubit 1 (Alice's electron spin) and qubit 3 (Bob's electron spin).] Now we change to the S_x representation of Alice's spin:

$$
\begin{aligned}
|\psi^{(1)}_{\mathrm{epe}'}(T)\rangle = \frac{1}{\sqrt{2}}\,|{\updownarrow}\rangle\Big[&|{\leftarrow}\rangle\Big(-\alpha\,|\uparrow'\rangle+\beta\,|\downarrow'\rangle\Big) \\
&+|{\rightarrow}\rangle\Big(-\alpha\,|\uparrow'\rangle-\beta\,|\downarrow'\rangle\Big)\Big] \\
+\frac{1}{\sqrt{2}}\,|{\leftrightarrow}\rangle\Big[&|{\leftarrow}\rangle\Big(\beta|\uparrow'\rangle+\alpha|\downarrow'\rangle\Big) \\
&+|{\rightarrow}\rangle\Big(\beta|\uparrow'\rangle-\alpha|\downarrow'\rangle\Big)\Big].
\end{aligned} \tag{1.106}
$$

The difference between our method and the original version of the teleportation (Bennett et al., 1993) is that we have entangled qubit 2 (the photon) with qubit 1 (Alice's electron spin) by means of the electron–photon interaction, which is not done in the original version. That is why we do not need Bell measurements. So after measuring the photon polarization state (qubit 2) and the spin state of Alice's electron (qubit 1), the spin state of the electron of the destination gets projected onto $-\alpha|\uparrow'\rangle + \beta|\downarrow'\rangle$, $-\alpha|\uparrow'\rangle - \beta|\downarrow'\rangle$, $\beta|\uparrow'\rangle + \alpha|\downarrow'\rangle$, or $\beta|\uparrow'\rangle - \alpha|\downarrow'\rangle$ with equal probability. This corresponds exactly to the outcome of the original version of the teleportation. However, we do not use any Bell measurements.

Let us check what happens if there is no interaction between Charlie's photon and Alice's spin. Then we get

$$|\psi^{(1)}_{\mathrm{epe}'}(T)\rangle = \frac{1}{\sqrt{2}}\Big[|\nwsearrow\rangle\big(\alpha|\uparrow\rangle|\uparrow'\rangle + \beta|\downarrow\rangle|\uparrow'\rangle\big) + |\neswarrow\rangle\big(\alpha|\uparrow\rangle|\downarrow'\rangle + \beta|\downarrow\rangle|\downarrow'\rangle\big)\Big]. \tag{1.107}$$

In this case, we would have to apply a Bell measurement in order to perform the teleportation.

1.9.3 Single-Photon Faraday Effect and Quantum Computing

Recently, the *destructive probabilistic* CNOT gate for all-optical quantum computing proposed by Knill et al. (2001), which relies on postselection by measurements, has been implemented experimentally (O'Brien et al., 2003; Nemoto and Munro, 2004). Based on a proposal by Pittman et al. (2001), experiments have demonstrated that this destructive CNOT gate and a quantum parity check can be combined with a pair of entangled photons to produce a *nondestructive probabilistic* CNOT gate (Pittman et al., 2002a,b, 2003; Gasparoni et al., 2004; Zhao et al., 2005). It has been shown theoretically that deterministic quantum computing with postselection is feasible for quantum systems where the qubits are represented by several degrees of freedom of a single photon (Cerf et al., 1998). An interferometric approach to linear-optical deterministic CNOT gate between the polarization and momentum of a single photonic qubit has recently been demonstrated experimentally (Fiorentino and Wong, 2004). A general scheme to teleport a quantum state through a universal gate has been given by Gottesman and Chuang (1999). This idea gave Raussendorf and Briegel the inspiration to develop the one-way quantum computer (Raussendorf and Briegel, 2001; Raussendorf et al., 2003), which performs quantum computations by measurement-induced teleportation processes on an entangled network of qubits. The name one-way quantum computing refers to the irreversibility of the computations due to the measurement processes.

Leuenberger developed a scheme for fault-tolerant quantum computing with the excess electron spins in quantum dots based on the SPFE (Leuenberger, 2006). This scheme shares a similarity with the one-way quantum computing scheme in the sense that it makes use of measurement processes to perform quantum computations. Both the one-way quantum computing scheme and Leuenberger's scheme are fully deterministic and scalable to an arbitrary number of qubits. However, while the one-way quantum computing scheme is irreversible, Leuenberger's scheme is fully reversible. In addition, the one-way quantum computing scheme needs to maintain the coherence of the whole cluster of qubits and is therefore more sensitive to the environment than single isolated qubits, since in a cluster the decoherence of one qubit leads also to decoherence of neighboring qubits that are entangled to it. Therefore, Leuenberger's scheme is more robust in this respect.

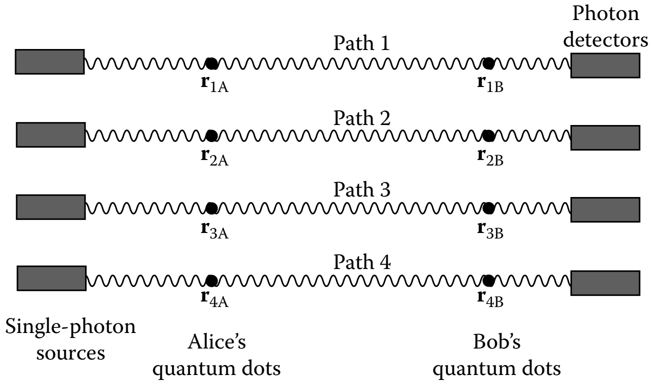

FIGURE 1.11 The nonresonant interaction of a photon with Alices and Bobs quantum dot produces the CPHASE gate required for universal quantum computing. This CPHASE gate applies to the case of the noncoded spins, where each microcavity contains only a single quantum dot with a single excess electron.

We describe now Leuenberger's quantum computing scheme (see Figure 1.11). DiVincenzo showed that the CNOT gate and a single qubit rotation are sufficient for universal quantum computing (DiVincenzo, 1995). In the framework of the conditional Faraday rotation, it is easier to work with the implementation of the CPHASE gate than with the implementation of the CNOT gate. The CPHASE gate

$$u_{\mathrm{CPHASE}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{1.108}$$

is equivalent (similar) to the CNOT gate (see Equation 1.7), i.e., they can be transformed into each other by means of a basis transformation. The CPHASE gate shifts the phase only if both spins point down. Let us define two persons Alice and Bob. Both of them have one photonic crystal, in which n noninteracting quantum dots are embedded. The single excess electrons of Alice's quantum dots are in a general single-spin state $|\psi\rangle_A = \alpha|\uparrow\rangle_A + \beta|\downarrow\rangle_A$, where the quantization axis is the z-axis. Bob's spins are in a general single-spin state $|\psi\rangle_B = \gamma|\uparrow\rangle_B + \delta|\downarrow\rangle_B$. The photons that interact with both Alice's and Bob's quantum dots are initially in a horizontal linear polarization state $|\leftrightarrow\rangle$.

Since there is no need to detect optically the spin states in transverse direction (Leuenberger et al., 2005a), our quantum dots can be nonspherical. So each photon can virtually create only a heavy-hole exciton on each quantum dot (see Figure 1.6). The strong selection rules imply that only a $\sigma^+_{(z)}(\sigma^-_{(z)})$ photon can interact with the quantum dot if the excess electron's spin is up (down). This leads to a conditional Faraday rotation of the linear polarization of the photon depending on the spin state of the quantum dot. The photonic crystal ensures that the photon's propagation direction is always in z-direction, perpendicular to the quantum dot plane.

In Leuenberger's implementation of the CPHASE gate, a single photon interacts sequentially with Alice's and then with Bob's quantum dot. The spin on Alice's quantum dot is prepared in the state $|\psi_A(0)\rangle = \alpha|\uparrow\rangle_A + \beta|\downarrow\rangle_A$. So we start with the electron–photon state $|\psi_{Ap}(0)\rangle = (\alpha|\uparrow\rangle_A + \beta|\downarrow\rangle_A)|\leftrightarrow\rangle$. After the interaction with Alice's quantum dot, the resulting electron–photon state is

$$|\psi_{Ap}(T)\rangle = \left(\alpha|\uparrow\rangle_A|\nwarrow\!\!\!\!\searrow\rangle + \beta|\downarrow\rangle_A|\swarrow\!\!\!\!\nearrow\rangle\right)/\sqrt{2}, \tag{1.109}$$

which is maximally entangled.

We let the photon interact also with Bob's quantum dot, which yields

$$\begin{aligned}|\psi_{ApB}(2T)\rangle &= |\updownarrow\rangle\left(-\alpha\gamma|\uparrow\rangle_A|\uparrow\rangle_B + \beta\delta|\downarrow\rangle_A|\downarrow\rangle_B\right)\\ &\quad + |\leftrightarrow\rangle\left(\alpha\delta|\uparrow\rangle_A|\downarrow\rangle_B + \beta\gamma|\downarrow\rangle_A|\uparrow\rangle_B\right)\\ &= \frac{1}{\sqrt{2}}\Big[|\nwarrow\!\!\!\!\searrow\rangle\big(-\alpha\gamma|\uparrow\rangle_A|\uparrow\rangle_B + \beta\delta|\downarrow\rangle_A|\downarrow\rangle_B\\ &\quad + \alpha\delta|\uparrow\rangle_A|\downarrow\rangle_B + \beta\gamma|\downarrow\rangle_A|\uparrow\rangle_B\big)\\ &\quad + |\swarrow\!\!\!\!\nearrow\rangle\big(\alpha\gamma|\uparrow\rangle_A|\uparrow\rangle_B - \beta\delta|\downarrow\rangle_A|\downarrow\rangle_B\\ &\quad + \alpha\delta|\uparrow\rangle_A|\downarrow\rangle_B + \beta\gamma|\downarrow\rangle_A|\uparrow\rangle_B\big)\Big].\end{aligned} \tag{1.110}$$

If the linear polarization of the photon is measured in the $\nwarrow\!\!\!\!\searrow$ axis, we obtain the two results

$$\begin{aligned}|\psi_{AB}(2T)\rangle &= -\alpha\gamma|\uparrow\rangle_A|\uparrow\rangle_B + \beta\delta|\downarrow\rangle_A|\downarrow\rangle_B\\ &\quad + \alpha\delta|\uparrow\rangle_A|\downarrow\rangle_B + \beta\gamma|\downarrow\rangle_A|\uparrow\rangle_B,\end{aligned} \tag{1.111}$$

$$\begin{aligned}|\psi_{AB}(2T)\rangle &= \alpha\gamma|\uparrow\rangle_A|\uparrow\rangle_B - \beta\delta|\downarrow\rangle_A|\downarrow\rangle_B\\ &\quad + \alpha\delta|\uparrow\rangle_A|\downarrow\rangle_B + \beta\gamma|\downarrow\rangle_A|\uparrow\rangle_B\end{aligned} \tag{1.112}$$

with equal probability, where the CPHASE shift is π if both spins are up or both spins are down, respectively. For deterministic quantum computing, we need to choose either Equation 1.111 or 1.112 to be the correct phase gate. Let us choose Equation 1.112 to be the correct implementation. The measurement of the photon's polarization state must be shared through a classical channel between Alice and Bob, in order for the CPHASE gate to be deterministic. Then Equation 1.111 can be transformed into Equation 1.112 by two local single-qubit phase shifts $\sigma_{A,z}\sigma_{B,z}$ applied to Alice's and Bob's qubit, where σ_z is a Pauli matrix. This scheme for implementing the CPHASE gate is similar to the scheme of quantum teleportation in the sense that the measurement outcome of the photon's polarization needs to be shared between Alice and Bob through a classical channel. This scheme for implementing the CPHASE gate is a generalization of the scheme of quantum teleportation in the sense that it takes not only one local unitary operation but two unitary operations to reconstruct the desired two-qubit state, which makes sense, because we perform a two-qubit manipulation, as opposed to a single qubit manipulation in the case of quantum teleportation.

Thus, we can produce deterministically the CPHASE gate, which is equivalent to the CNOT gate in the basis $\{|\uparrow\rangle_A|\leftarrow\rangle_B, |\uparrow\rangle_A|\rightarrow\rangle_B, |\downarrow\rangle_A|\leftarrow\rangle_B, |\downarrow\rangle_A|\rightarrow\rangle_B\}$, where $|\leftarrow\rangle = (|\uparrow\rangle + |\downarrow\rangle)/\sqrt{2}$ and $|\rightarrow\rangle = (|\uparrow\rangle - |\downarrow\rangle)/\sqrt{2}$ in the S_x representation. This can be proven by writing down the mappings

$$\begin{aligned}|\uparrow\rangle_A|\leftarrow\rangle_B &= \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(-|\uparrow\rangle_B + |\downarrow\rangle_B\right) = -|\uparrow\rangle_A|\rightarrow\rangle_B,\\ |\uparrow\rangle_A|\rightarrow\rangle_B &= \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(-|\uparrow\rangle_B - |\downarrow\rangle_B\right) = -|\uparrow\rangle_A|\leftarrow\rangle_B,\\ |\downarrow\rangle_A|\leftarrow\rangle_B &= \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right) = |\downarrow\rangle_A|\leftarrow\rangle_B,\\ |\downarrow\rangle_A|\rightarrow\rangle_B &= \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right) = |\downarrow\rangle_A|\rightarrow\rangle_B,\end{aligned} \tag{1.113}$$

corresponding to the CPHASE gate in Equation 1.111, and

$$\begin{aligned}|\uparrow\rangle_A|\leftarrow\rangle_B &= \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right) = -|\uparrow\rangle_A|\leftarrow\rangle_B,\\ |\uparrow\rangle_A|\rightarrow\rangle_B &= \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\uparrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right) = |\uparrow\rangle_A|\rightarrow\rangle_B,\\ |\downarrow\rangle_A|\leftarrow\rangle_B &= \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right) = |\downarrow\rangle_A|\rightarrow\rangle_B,\\ |\downarrow\rangle_A|\rightarrow\rangle_B &= \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B - |\downarrow\rangle_B\right)\\ &\to \frac{1}{\sqrt{2}}|\downarrow\rangle_A\left(|\uparrow\rangle_B + |\downarrow\rangle_B\right) = |\downarrow\rangle_A|\leftarrow\rangle_B,\end{aligned} \tag{1.114}$$

corresponding to the CPHASE gate in Equation 1.112.

For universal quantum computing, the CPHASE gate described above and single-qubit operations are sufficient (Barenco et al., 1995; DiVincenzo, 1995). The single-qubit operations on Alice's and Bob's spins can be implemented by means of the optical Stark effect (Gupta et al., 2001). In contrast to Imamoglu et al.'s two-qubit gate (Imamoglu et al., 1999), which relies on a shared mode volume for the exchange of a single virtual photon between the two qubits, this scheme requires only local mode volumes for each qubit separately and is therefore fully scalable to an arbitrary number of qubits. Since the strong interaction between a quantum dot and a cavity inside a photonic crystal has already been experimentally observed (Yoshie et al., 2004), it would be possible to build a fully scalable quantum network inside a photonic crystal hosting qubits, which can be for example spins of excess electrons in semiconductor quantum dots or N-V center states in diamond.

1.10 Concluding Remarks

Quantum computers have excited many researchers because they would perform a kind of parallel processing that would be extremely effective for certain tasks, such as searching databases and factoring large numbers. Also, quantum computers can be used to simulate or model quantum systems and could bring huge advances in physics, chemistry, and nanotechnology. Recently, diamond has become very attractive for solid state electronics. Pure diamond is an electrical insulator, but on doping, it can become a semiconductor. The particular impurity that researchers are interested in is the nitrogen-vacancy (N-V) center. The N-V center has a number of properties and characteristics that make it very promising to build a quantum computer (Awschalom et al., 2007). The N-V center electrons have a spin state that is extremely stable against environmental disturbances. One of the most exciting aspects of the N-V center is that it exhibits quantum behavior even at room temperature; this characteristic is very important because it makes this kind of systems easy to study and easy to turn into a practical technology. Other important aspects of N-V centers come from the fact that they have weak spin–orbit interaction and dipole–dipole interaction, which make the spin state very stable and it can be used to encode quantum information even at room temperatures. The decoherence time of the N-V center spin is about 1 ms and the operation time is about 10 ns, therefore $R = \tau_d/\tau_s = 100{,}000$ operations that can be performed in the millisecond lifetime of the spin quantum system. This rate of decay is well below the threshold and better than any other system of solid-state qubits to date.

Acknowledgments

We acknowledge support from NSF-ECCS 0725514, the DARPA/MTO Young Faculty Award HR0011-08-1-0059, NSF-ECCS 0901784, and AFOSR FA9550-09-1-0450. We would also like to thank Sergio Tafur for his comments on improving the manuscript.

References

Adamowski, J., S. Bednarek, and B. Szafram. Quantum computing with quantum dots. *Schedae Inform.*, 14:95–111, 2005.

Ahn, J., Weinacht, T.C., and P.H. Bucksbaum. Information storage and retrieval through quantum phase. *Science*, 287: 463–465, 2000.

Awschalom, D.D., R. Epstein, and R. Hanson. The diamond age of spintronics. *Sci. Am.*, 297:84–91, 2007.

Barenco, A., D. Deutsch, A. Ekert, and R. Josza. Conditional quantum dynamics and logic gates. *Phys. Rev. Lett.*, 74:4083, 1995.

Bennett, C.H., G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W.K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. *Phys. Rev. Lett.*, 70:1895–1899, 1993.

Bennett, C.H., D.P. DiVincenzo, J.A. Smolin, and W.K. Wootters. Mixed-state entanglement and quantum error correction. *Phys. Rev. A*, 54:3824–3851, 1996.

Berezovsky, J., M.H. Mikkelsen, O. Gywat, N.G. Stoltz, L.A. Coldren, and D.D. Awschalom. Nondestructive optical measurements of a single electron spin in a quantum dot. *Science*, 314:1916, 2006.

Bouwmeester, D., J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger. Experimental quantum teleportation. *Nature*, 390:575, 1997.

Cerf, N.J., C. Adami, and P.G. Kwiat. Optical simulation of quantum logic. *Phys. Rev. A*, 57:R1477, 1998.

Deutsch, D. Quantum theory, the Church–Turing principle, and the universal quantum computer. *Proc. R. Soc. Lond. A*, 400:97–117, 1985.

DiVincenzo, D.P. Two-bit quantum gates are universal for quantum computation. *Phys. Rev. A*, 51:1015–1022, 1995.

DiVincenzo, D.P. The physical implementation of quantum computation, in *Scalable Quantum Computers*, S.L. Braunstein and H.K. Lo (eds.), Wiley, Berlin, Germany, 2001.

Feynman, R.P. Simulating physics with computers. *Int. J. Theor. Phys.*, 21:467–488, 1982.

Fiorentino, M. and F.N.C. Wong. Deterministic controlled-not gate for single-photon two-qubit quantum logic. *Phys. Rev. Lett.*, 93:070502, 2004.

Friedman, J.R., M.P. Sarachik, J. Tejada, and R. Ziolo. Macroscopic measurement of resonant magnetization tunneling in high-spin molecules. *Phys. Rev. Lett.*, 76:3830–3833, 1996.

Gasparoni, S., J.-W. Pan, P. Walther, T. Rudolph, and A. Zeilinger. Realization of a photonic controlled-not gate sufficient for quantum computation. *Phys. Rev. Lett.*, 93:020504, 2004.

Gottesman, D. and I.L. Chuang. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature*, 402:390, 1999.

Greenberger, D.M., M.A. Horne, and A. Zeilinger. in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe*. Kluwer, Dordrecht, the Netherlands, p. 73, 1989.

Greenberger, D.M., M.A. Horne, A. Shimony, and A. Zeilinger. Bell theorem without inequalities. *Am. J. Phys.*, 58:1131, 1990.

Grover, L.K. Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.*, 79:325–328, 1997.

Gupta, J.A., R. Knobel, N. Samarth, and D.D. Awschalom. Ultrafast manipulation of electron spin coherence. *Science*, 292:2458, 2001.

Hu, C.Y., A. Young, J.L. O'Brien, W.J. Munro, and J.G. Rarity. Giant optical Faraday rotation induced by a single-electron spin in a quantum dot: Applications to entangling remote spins via a single photon. *Phys. Rev. B*, 78:085307, 2008.

Imamoglu, A., D.D. Awschalom, G. Burkard, D.P. DiVincenzo, D. Loss, M. Sherwin, and A. Small. Quantum information processing using quantum dot spins and cavity QED. *Phys. Rev. Lett.*, 83:4204, 1999.

Kasuya, T. A theory of metallic ferro- and antiferromagnetism on Zener's model, *Prog. Theor. Phys.* 16:45, 1956.

Kikkawa, J.M. and D.D. Awschalom. Resonant spin amplification in n-type gas. *Phys. Rev. Lett.*, 80:4313, 1998.

Kikkawa, J.M. and D.D. Awschalom. Lateral drag of spin coherence in gallium arsenide. *Nature*, 397:139, 1999.

Kikkawa, J.M., I.P. Smorchkova, N. Samarth, and D.D. Awschalom. Room-temperature spin memory in two-dimensional electron gases. *Science*, 277:1284, 1997.

Knill, E., R. Laflamme, and G.J. Milburn. A scheme for efficient quantum computation with linear optics. *Nature*, 409:46, 2001.

Leuenberger, M.N. Fault-tolerant quantum computing with coded spins using the conditional Faraday rotation in quantum dots. *Phys. Rev. B*, 73:075312–8, 2006.

Leuenberger, M.N. and D. Loss. Spin relaxation in Mn12-acetate. *EuroPhys. Lett.*, 46:692–698, 1999.

Leuenberger, M.N. and D. Loss. Spin tunneling and phonon-assisted relaxation in Mn12-acetate. *Phys. Rev. B*, 61:1286–1302, 2000a.

Leuenberger, M.N. and D. Loss. Incoherent Zener tunneling and its application to molecular magnets. *Phys. Rev. B*, 61:12200–12203, 2000b.

Leuenberger, M.N. and D. Loss. Quantum computing in molecular magnets. *Nature*, 410:789–793, 2001.

Leuenberger, M.N., D. Loss, M. Poggio, and D.D. Awschalom. Quantum information processing with large nuclear spins in gas semiconductors. *Phys. Rev. Lett.*, 89:207601–4, 2002.

Leuenberger, M.N., F. Meier, and D. Loss. Quantum spin dynamics in molecular magnets. *Monatshefte Chem.*, 134:217–233, 2003.

Leuenberger, M.N., M.E. Flatté, and D.D. Awschalom. Teleportation of electronic many-qubit states encoded in the electron spin of quantum dots via single photons. *Phys. Rev. Lett.*, 94:107401–4, 2005a.

Leuenberger, M.N., M.E. Flatté, and D.D. Awschalom. Teleportation of electronic many-qubit states encoded in the electron spin of quantum dots via single photons. *Phys. Rev. Lett.*, 94:107401–4, 2005b.

Loss, D. and D.P. DiVincenzo. Quantum computation with quantum dots. *Phys. Rev. A*, 57:120–126, 1998.

Nemoto, K. and W.J. Munro. Nearly deterministic linear optical controlled-not gate. *Phys. Rev. Lett.*, 93:250502, 2004.

O'Brien, J.L., G.J. Pryde, A.G. White, T.C. Ralph, and D. Branning. Demonstration of an all-optical quantum controlled-not gate. *Nature*, 426:264, 2003.

Piermarocchi, C., P. Chen, and L.J. Sham. Optical RKKY interaction between charged semiconductor quantum dots. *Phys. Rev. Lett.*, 89:167402, 2002.

Pittman, T.B., B.C. Jacobs, and J.D. Franson. Probabilistic quantum logic operations using polarizing beam splitters. *Phys. Rev. A*, 64:062311, 2001.

Pittman, T.B., B.C. Jacobs, and J.D. Franson. Demonstration of feed-forward control for linear optics quantum computation. *Phys. Rev. A*, 66:052305, 2002a.

Pittman, T.B., B.C. Jacobs, and J.D. Franson. Demonstration of nondeterministic quantum logic operations using linear optical elements. *Phys. Rev. Lett.*, 88:257902, 2002b.

Pittman, T.B., M.J. Fitch, B.C. Jacobs, and J.D. Franson. Experimental controlled-not logic gate for single photons in the coincidence basis. *Phys. Rev. A*, 68:032316, 2003.

Raussendorf, R. and H.J. Briegel. A one-way quantum computer. *Phys. Rev. Lett.*, 86:5188–5191, 2001.

Raussendorf, R., D.E. Browne, and H.J. Briegel. Measurement-based quantum computation on cluster states. *Phys. Rev. A*, 68:022312, 2003.

Ruderman, M.A. and C. Kittel. Indirect exchange coupling of nuclear magnetic moments by conduction electrons, *Phys. Rev*. 96:99, 1954.

Sangregorio, C., T. Ohm, C. Paulsen, R. Sessoli, and D. Gatteschi. Quantum tunneling of the magnetization in an iron cluster nanomagnet. *Phys. Rev. Lett.*, 78:4645–4648, 1997.

Seigneur, H.P., M.N. Leuenberger, and W.V. Schoenfeld. Single-photon Mach–Zehnder interferometer for quantum networks based on the single-photon Faraday effect. *J. Appl. Phys.*, 104:014307–13, 2008.

Shor, P. Scheme for reducing decoherence in quantum computer memory. *Phys. Rev. A*, 52:2493–2496, 1995.

Shor, P. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Comput.*, 26:1484–1509, 1997.

Silberfarb, A. and I.H. Deutsch. Entanglement generated between a single atom and a laser pulse. *Phys. Rev. A*, 69:042308–8, 2004.

Thiaville, A. and J. Miltat. Small is beautiful. *Science*, 284:1939–1940, 1999.

Thomas, L., F. Lionti, R. Ballou, D. Gatteschi, R. Sessoli, and B. Barbara. Macroscopic quantum tunnelling of magnetization in a single crystal of nanomagnets. *Nature*, 383: 145–147, 1996.

Wernsdorfer, W., R. Sessoli, A. Caneshi, D. Gatteschi, and A. Cornia. Nonadiabatic Landau–Zener tunneling in Fe8 molecular nanomagnets. *EuroPhys. Lett.*, 50:552–558, 2000.

Yoshie, T., A. Scherer, J. Hendrickson, G. Khitrova, H.M. Gibbs, G. Rupper, C. Ell, O.B. Shchekin, and D.G. Deppe. Vacuum Rabi splitting with a single quantum dot in a photonic crystal nanocavity. *Nature*, 432:200, 2004.

Yosida, K. Magnetic properties of Cu-Mn alloys, *Phys. Rev*. 106:893, 1957.

Zhao, Z., A.-N. Zhang, Y.-A. Chen, H. Zhang, J.-F. Du, T. Yang, and J.-W. Pan. Experimental demonstration of a nondestructive controlled-not quantum gate for two independent photon qubits. *Phys. Rev. Lett*., 94:030501, 2005.

2

Nanomemories Using Self-Organized Quantum Dots

Martin Geller
University of Duisburg-Essen

Andreas Marent
Technische Universität Berlin

Dieter Bimberg
Technische Universität Berlin

2.1 Introduction **2**-1
2.2 Conventional Semiconductor Memories **2**-2
Dynamic Random Access Memory • Nonvolatile Semiconductor Memories (Flash)
2.3 Nonconventional Semiconductor Memories **2**-4
Ferroelectric RAM • Magnetoresistive RAM • Phase-Change RAM
2.4 Semiconductor Nanomemories **2**-6
Charge Trap Memories • Single Electron Memories
2.5 A Nanomemory Based on III–V Semiconductor Quantum Dots **2**-7
III–V Semiconductor Materials • Self-Organized Quantum Dots • A Memory Cell Based on Self-Organized QDs
2.6 Capacitance Spectroscopy **2**-10
Depletion Region • Capacitance Transient Spectroscopy
2.7 Charge Carrier Storage in Quantum Dots **2**-16
Carrier Storage in InGaAs/GaAs Quantum Dots • Hole Storage in GaSb/GaAs Quantum Dots • InGaAs/GaAs Quantum Dots with Additional AlGaAs Barrier • Storage Time in Quantum Dots
2.8 Write Times in Quantum Dot Memories **2**-21
Hysteresis Measurements • Write Time Measurements
2.9 Summary and Outlook **2**-23
Acknowledgments **2**-23
References **2**-24

2.1 Introduction

One of the fundamental achievements of today's information society is providing storage and processing of ever-increasing amounts of data that is based on basic research in physics in combination with fundamental technology developments over the last four decades. As no low-cost universal memory that is suited for all kinds of applications exists, digital data storage is separated mainly into three different lines:

1. The digital versatile disk (DVD) and the compact disk (CD) are nonvolatile storage media based on an *optical* write and read-out process using laser light at wavelengths of 650 and 780 nm, respectively. The newest representative of optical data storage is the Blu-ray disk, using a laser wavelength at 405 nm with a maximum capacity of up to 50 GB disk^{-1}.
2. The hard-disk drive in a personal computer is a nonvolatile mass storage device with a capacity of up to 1000 GB in production year 2008 that uses the *magnetization* of rapidly rotating platters.
3. *Semiconductor* memories are the third line of digital storage media and will be the focus of this chapter with a perspective of future nanomemories.

Semiconductor memories can be divided into two groups: volatile memories, like the dynamic random access memory (DRAM), and nonvolatile memories, the so-called Flash. The DRAM presents the main working memory in a personal computer. Flash memories are found, for example, in mp3 players, cell phones, and memory sticks; in automobiles and microwave ovens; and have started to replace the hard-disk drive in notebooks as well as the DVD and CD as easily portable high-capacity nonvolatile memories. Driven by the increasing demand for such portable electronic applications, where nonvolatile data storage with low power consumption is needed, the market for Flash memories is growing rapidly and is replacing the DRAM as the market driver.

Up to now, the semiconductor memory industry improved the storage density and performance while simultaneously reducing the cost per information bit just by scaling down the feature size. This leads to an exponentially growing number of components on a memory chip as predicted by Moore in 1965 (Moore 1965) known nowadays as "Moore's Law." Contrary to all predictions, Moore's Law has held remarkably well over the last decades. The feature size for Flash memories has reached 45 nm in 2008. The International Roadmap for Semiconductors (ITRS 2008) predicts a further shrinkage down to 14 nm in 2020. The problems encountered during this shrinking process were mainly solved by developing new materials, for example, for isolation and interconnects, and for new types of cell structures. Upon further size reduction, quantum mechanics will dominate at least some of the physical properties. In addition, the amount of technological difficulties to realize such structures increases enormously. Therefore, a considerable effort is devoted to the search for alternative memory technologies that are even based on completely new non-semiconducting materials. Phase change access memories (PCRAMs), magnetic random access memories (MRAMs), and ferroelectric RAM (FeRAMs) are just three alternative memory concepts that are explored for having the potential to replace today's DRAM and/or Flash memory.

One completely different technology is the use of self-organized nanomaterials in future nanoelectronic devices. Especially self-organized quantum dots (QDs) based on III–V materials (e.g., GaAs, InAs, InP, etc.) provide a number of important advantages for new generations of nanomemories. Billions of self-organized QDs can be formed simultaneously and fast in a single technological step, allowing for massive parallel production in a bottom-up approach, and offering an elegant method to create huge ensembles of electronic traps without lithography. They can store just a few or even single charge carriers with a retention time depending on the material combination—potentially up to many years at room temperature. With an area density of up to 10^{11} cm^{-2}, an enormous storage density in the order of 1 TBit per square inch could be possible, if each single QD would represent one information bit and could be addressed individually. The carrier capture process into the QDs is of the order of pico to sub-picoseconds, an important prerequisite for a very fast write time in such memories. The use of self-organized QDs could thus lead to novel nonvolatile memories with high storage density combined with a fast read/write access time.

Using self-organized QDs for future semiconductor memory applications is discussed in this chapter. Section 2.2 gives a brief overview of the main semiconductor memories, DRAM and Flash, while the following Section 2.3 concentrates on three nonconventional memories that may replace Flash or DRAM in the future. Section 2.5 presents an alternative nanomemory concept that is based on III–V materials, especially self-organized QDs. By using capacitance spectroscopy, introduced in Section 2.6, the carrier storage time in different QD systems is studied in Section 2.7 and a storage time of seconds at room temperature is demonstrated. Finally, the chapter will show results on fast write times in QD-based nanomemories in Section 2.8 and will close with a summary and outlook in Section 2.9.

2.2 Conventional Semiconductor Memories

Presently the semiconductor memory industry focuses essentially on two memory types: the DRAM (Mandelman et al. 2002, Waser 2003) and the Flash (Geppert 2003). Both memory concepts have their advantages and disadvantages in speed, endurance, storage time, and cost. A memory concept that adds the advantages of a DRAM to a Flash would combine high storage density, fast read/write access, long data storage time, and good endurance with low production cost. In addition, for portable applications like mobile phones and mp3 players, low power consumption is demanded. This section describes the two conventional semiconductor memories, while Section 2.3 will focus on nonconventional alternatives that have the potential to replace DRAM and/or Flash.

2.2.1 Dynamic Random Access Memory

Since the invention of the DRAM in the late 1960s (Dennard 1968), its cell structure mainly stayed the same, consisting of a transistor and a capacitor (cf. Figure 2.1a). The capacitor stores the information by means of electric charge. The charge state is defined by the voltage level on the capacitor. The stored charge

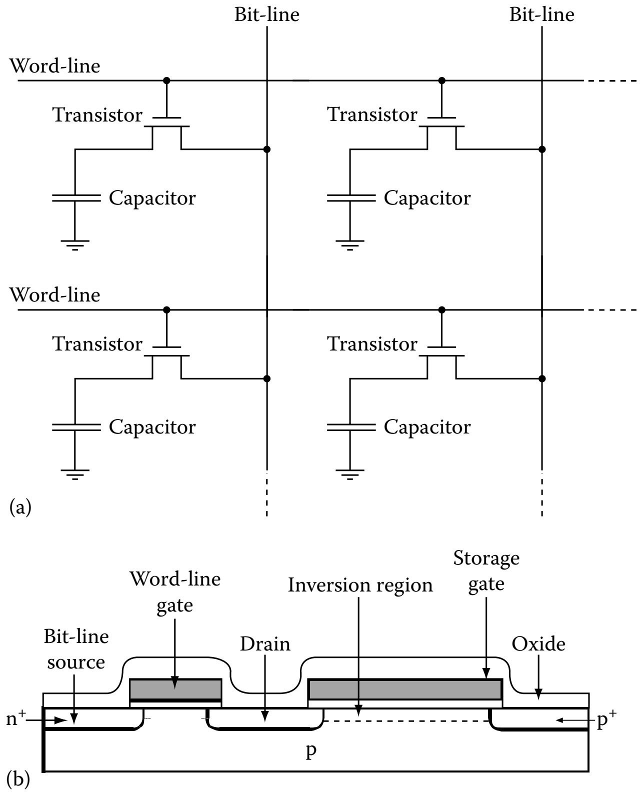

FIGURE 2.1 (a) Schematic picture of an array of DRAM cells, where the capacitors act as the storage units. Each cell can be addressed individually by a matrix of word- and bit-lines; it is the so-called random-access architecture. (b) Cell layout for a planar DRAM cell.

disappears typically in a few milliseconds mainly due to leakage and recombination currents. A DRAM is volatile and requires periodic "refreshing" of the stored charge. After any read process, the information has to be rewritten into the DRAM cell. A chip on the integrated circuit controls the refresh rewrite process automatically and this continuous read and write operation has led to the name "Dynamic."

Figure 2.1b shows schematically the layout of a simple planar DRAM cell (Sze 2002). The storage capacitor uses the inversion channel region as one plate, the storage gate as the other plate, and the gate oxide as the dielectric. The access transistor is a metal oxide semiconductor field effect transistor (MOSFET) with a source, drain, and a gate contact. The drain contact is connected to the storage capacitor. The source contact is connected to the bit-line and the gate contact to the word-line. Via word- and bit-line, a single cell can be addressed in an organized matrix of DRAM cells (Figure 2.1a). This so-called "random access" provides short access times independent of the location of the data (Waser 2003). For a write/read operation, a voltage pulse is passed through the selected bit- and word-line. Only at the crossing point of a given DRAM cell, the access transistor switches to "open" and the capacitor is charged or discharged for a write or read event, respectively.

DRAM cells provide fast read, write, and erase access times below 20 ns in combination with a very good endurance of more than 10^{15} write/erase cycles. The endurance is defined as the minimum number of possible write/erase operations until the memory cell is destroyed. DRAM memory cells draw power continuously due to the refresh process and the information is lost after switching off the computer. Another disadvantage is the relatively large number of electrons, presently in the order of 10^5, needed to store one information bit, leading additionally to permanent high power consumption. Both, the volatility and the power consumption make DRAMs unsuitable for mobile applications.

The goal to shrink the feature size, based on the assumption of scalability, of a DRAM (Kim et al. 1998) down to 14 nm by 2020 is a major challenge. It is speculated that the leakage currents of the capacitor might inhibit the scalability very soon. A fixed capacitance of the order of 50 pF is needed to maintain a sufficiently high voltage signal during the read-out process. Shrinking the area of the capacitor will decrease its capacitance, the number of stored electrons per DRAM cell, and the amplitude of the read-out signal. Downscaling increases leakage currents due to quantum-mechanical tunneling through thinner dielectrics (Frank et al. 2001). Therefore, there is a search for novel dielectrics.

2.2.2 Nonvolatile Semiconductor Memories (Flash)

Nonvolatile memories (NVMs) can retain their data for typically more than 10 years without power consumption. The most important nonvolatile memory is the Flash-EEPROM (electrically erasable and programmable read-only memory); in short just Flash (Geppert 2003, Lai 2008). A Flash offers the possibility of repeated electrical read, write, and erase processes. The Flash memory market is the fastest growing memory market today, since it is the ideal memory device for portable applications. In a mobile phone, it holds the instructions and data needed to send and receive calls, or stores phone numbers. But not only portable applications are ideal for Flash. In each computer, a Flash chip holds the data on how to boot up. Other electronic products of all types, from microwaves ovens to industrial machines, store their operating instructions in Flash memories.

Flash is based on a floating-gate structure (Pavan et al. 1997, Sze 1999), where the charge carriers are trapped inside a polysilicon floating gate embedded between two SiO_2 barriers (Figure 2.2a). The SiO_2 barriers having a height of ~3.2 eV and an average thickness of 10 nm guarantee a storage time of more than 10 years at room temperature. However, these barriers are also the origin of the two main disadvantages of a Flash cell: a slow write time (in the order of microseconds) and a poor endurance (in the order of 10^6 write/erase cycles).

The write process is realized by means of "hot-electron injection." Here a small voltage is applied between the source contact and the bit-line (drain contact) (cf. Figure 2.2b). The word-line (control gate) is set to a high positive bias of 10–20 V, the MOSFET is set to "open" and the electrons flow though the inversion channel from the source to drain and are, in addition, accelerated in the direction of the floating-gate due to the high electric field in the order of 10^7 V cm^{-1}. These "hot-electrons" have sufficient kinetic energy to reach the floating gate over one of the SiO_2 barriers, but destroy this barrier step by step by

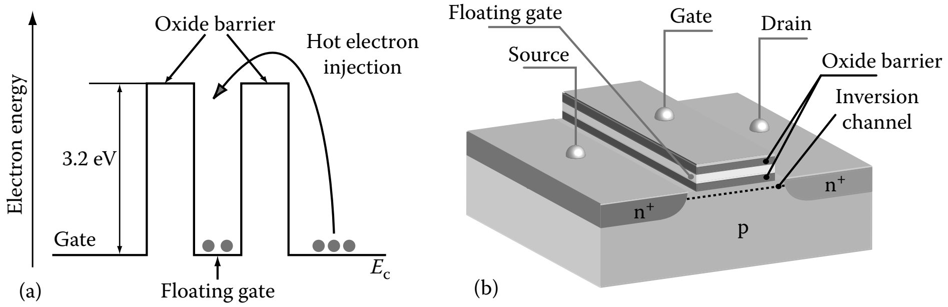

FIGURE 2.2 (a) Schematic band structure of a Flash memory based on Si as matrix material and SiO_2 for the oxide barriers. (b) Schematic cell layout of a floating-gate Flash memory.

Memory	Write	Read	Storage	Electrons	Endurance
DRAM	~10 ns	~10 ns	~ms	>10,000	$>10^{15}$
Flash	μs–ms	~20 ns	>10 years	~1,000	$\sim 10^{6}$

FIGURE 2.3 Comparison of DRAM and Flash.

creating defects leading eventually to leakage. This is the reason for the poor endurance of a Flash memory cell. The slow write time is due to the low probability for energy relaxation of these hot-electrons into the floating gate.

The read-out of the stored information is normally done by measuring the resistance of the two-dimensional electron gas (2DEG) of the MOSFET—the inversion channel. A bias is applied at the gate contact, forming inversion between the source and the drain. Stored electrons in the floating gate reduce the conductivity of the 2DEG and a higher resistance between the source and the drain is measured.

Figure 2.3 compares the main properties of the DRAM and Flash. The Flash memory has already replaced the DRAM as a technology driver for the semiconductor memory industry (Mikolajick et al. 2007) in the year 2003. It is leading in storage density and the main innovations are coming from the Flash industry. The feature size of the Flash has decreased from 1.5 μm in the year 1990 down to 45 nm in the year 2008. Accordingly, the number of electrons decreased from about 10^5 to less than 1000 per information bit (Atwood 2004) (cf. Figure 2.4). A further shrinkage to 14 nm in the year 2020, as predicted, would lead to only 100 electrons per information bit.

Flash memory scaling beyond 32 nm feature size will become more and more difficult, mainly due to the physical limitations (Atwood 2004). For instance, during the hot-electron injection a voltage of about 4.5 V is applied between the source and the drain. Shrinking the gate length while keeping the write voltage fixed at 4.5 V, leads to an increased electric field in the source–drain channel and in the end to a punch-through. Another limitation for future scaling is the capacitive coupling between different floating gates (cross talk) and the scalability of the floating gate. While the number of electrons decreases to about 200 per cell in year 2020, the main charge leakage mechanisms will remain. Dangling bonds and other defects at the Si/SiO$_2$ interface will dramatically decrease the storage time while scaling to smaller cell sizes.

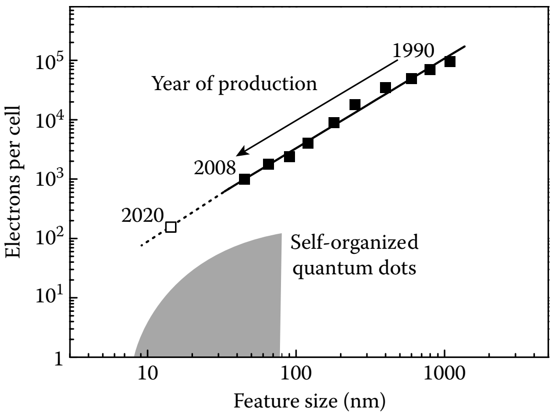

FIGURE 2.4 Electron number per Flash memory cell versus the minimum feature size. The feature size decreases from 1.5 μm in the production year 1990 down to 45 nm in the year 2008. Accordingly, the number of electrons decreased from about 10^5 to less than 1000 per information bit.

2.3 Nonconventional Semiconductor Memories

The increasing number of difficulties and challenges for further scaling of the DRAM and Flash has led to a continuous search for alternative memory concepts. A large variety of proposed concepts using different physical phenomena to store an information bit have been reported (Burr et al. 2008). This section concentrates on three nonconventional memories, which are the most advanced ones (Geppert 2003, Burr et al. 2008): the ferroelectric RAM (FeRAM), the magnetic RAM (MRAM), and the phase-change RAM (PCRAM).

2.3.1 Ferroelectric RAM

The ferroelectric RAM (FeRAM) (Jones et al. 1995, Sheikholeslami and Gulak 2000) has almost the same structure as a DRAM cell, except for the capacitor. The dielectric inside the capacitor of a DRAM is a non-ferroelectric material like silicon dioxide. When the charge is stored on the metal plates of the capacitor, it leaks away into the silicon substrate within a few milliseconds, causing the nonvolatility of a DRAM cell. In a FeRAM, a ferroelectric film such as zirconate titanate, also known as PZT, replays the dielectric of the capacitor. This ferroelectric material has a remnant polarization that occurs when an electric field has been applied. In PZT the center atom is zirconium or titanium which can be moved by an external electric field into two different stable states, representing a “0” or “1” in a ferroelectric memory. One state is near the top face of the PZT cube, the other one is near the bottom face. Even after removal of the external electric field, these stable states store the information for more than 10 years. When an opposite electric field is applied, the dipoles flip to the opposite direction, that is, the zirconium or titanium atoms are switching into the other stable state.

The write operations of a “0” and “1” state in a FeRAM are in principle the same as in a DRAM cell. An electric pulse via a word and a bit-line switches the ferroelectric state of the PZT. To read an information bit, an electrical pulse is applied via the access transistor to the ferroelectric capacitor and a sense amplifier can measure a current pulse. The amplitude depends on the position of the zirconium or titanium atoms in the PZT cube. For instance, if the external electric field from the pulse is pointing in the same direction of the PZT, that is, the atoms are already at the top of the cube and a smaller current pulse is detected than for opposite direction.

Reading the information in a FeRAM cell destroys the data stored in its ferroelectric capacitor. After a read operation, the information has to be rewritten into the FeRAM cell, which is a major disadvantage in comparison to the Flash memory based on a floating-gate structure. However, the FeRAM is nonvolatile having a random read/write access time below 50 ns and an almost unlimited endurance ($>10^{12}$ write/erase cycles).

2.3.2 Magnetoresistive RAM

The magnetoresistive RAM (Gallagher and Parkin 2006, Wolf et al. 2006, Burr et al. 2008) uses the giant magnetoresistance (GMR) effect that was discovered in the late 1980s by the groups of Albert Fert (Baibich et al. 1988) and Peter Grunberg (Binasch et al. 1989). Grunberg and Fert received the Nobel Prize in Physics 2007 for their discovery.

The MRAM is based on a magnetic tunnel junction (MTJ) (see Figure 2.5a), which consists of two magnetic layers separated by a thin insulating layer (like AlOx) with a thickness in the order of 1 nm. One of the ferromagnetic layers has a fixed magnetization, while the other layers can flip its magnetization by an external writing event. The read-out of the information in a MRAM cell works as follows: if the access transistor for a certain cell is turned on, a current is driven through the MTJ and the resistance of the MRAM cell is measured. If the two magnetic moments are parallel, the resistance is low, representing a "0" state. If the moments are antiparallel, the resistance is high, representing a "1" state. The difference between these two values can be up to 70%, therefore, GMR.

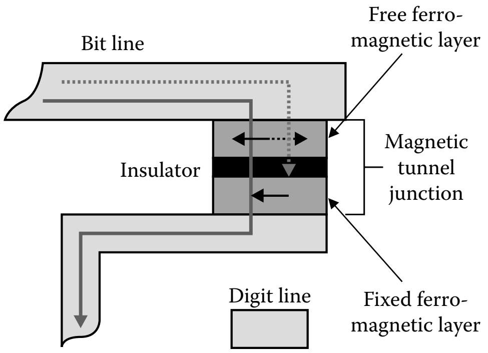

(a) Read-out process

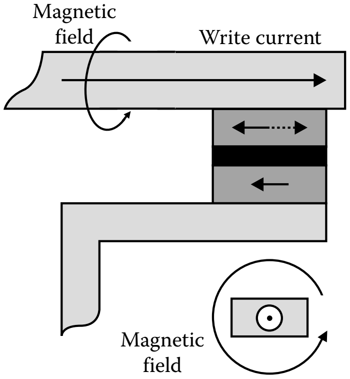

(b) Write process

FIGURE 2.5 Schematic picture of the read (a) and write (b) process of an MRAM cell. During read out a current passes through the MTJ. If the magnetization of the two magnetic layers is parallel to each other a low resistance is measured; if the magnetization is antiparallel the resistance is high. The write process is based on current flow through both the bit and digit line. The sum of both currents is strong enough to flip the magnetic domains inside the free magnetic layer to a "1" or "0" state, depending on the current direction inside the lines.

The physical effect that leads to this GMR effect is called tunnel magnetoresistance (TMR). The TMR effect can be understood in terms of spin polarized tunneling of the electrons. The electron spins are polarized inside one magnetic layer and if the spin is conserved during tunneling through the thin insulator layer, an initially spin up electron can only tunnel to a spin up final state. If the magnetic layers have parallel magnetic moments, a higher current is detected than for antiparallel directions.

The write mechanism can be understood by looking at Figure 2.5b. During the write mode, a current is passing through two wires: a bit-line that runs over the MTJ and a digit line that is below. The sum of both currents is strong enough to flip the magnetic domains inside the free magnetic layer to a "1" or "0" state, depending on the current direction. Both magnetic states are stable for more than 10 years; hence, an MRAM is a nonvolatile memory device with a random-access architecture and unlimited endurance. In addition, the write/read time has been demonstrated to be below 50 ns. However, scaling MRAM to smaller feature sizes is a big challenge as the write current has to remain very high (>1 mA), even if the feature size goes below 40 nm. Such large currents in very small devices might damage the structure.

2.3.3 Phase-Change RAM

The architecture of the phase-change RAM (PCRAM)—also called ovonic unified memory (OUM)—is again an array of access transistors that are connected to a word- and a bit-line. The capacitor is now replaced by a phase-change material (Geppert 2003, Burr et al. 2008). The PCRAM uses the large resistance change between a (poly) crystalline and an amorphous state in a chalcogenide glass, which is also used in rewritable CDs and DVDs. The chalcogenides used for this type of memory are alloys containing elements like selenium, tellurium, or antimony (GeSbTe, GeTe, Sb_2Te_3, etc.). The crystalline and amorphous states show a large difference of up to five orders of magnitude in the electrical resistance, representing the binary state "1" for low resistance and the state "0" for high resistance. The read-out can easily be done by a measurement of the resistance of the PCRAM cell.

Besides the chalcogenide, the PCRAM cell consists of a top and a bottom electrode that are connected to the word- and bit-line and a resistive heater below the phase-change material (see Figure 2.6). To switch the state of the cell to the crystalline

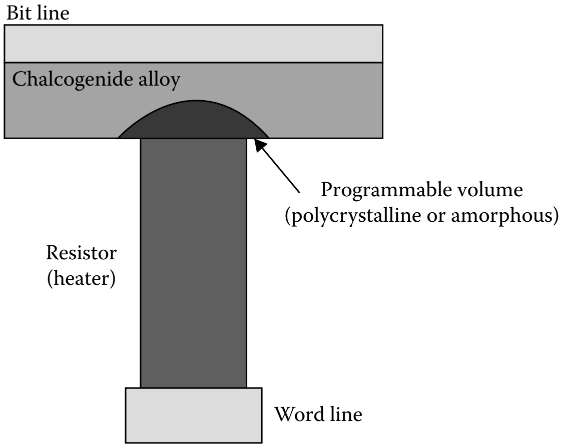

FIGURE 2.6 The PCRAM is based on a chalcogenide that can exist in two different states: in the (poly) crystalline state the resistance is low (representing a "1"), while the resistance is high in the amorphous state (representing a "0"). Switching to the amorphous state is done by heating the programmable volume up to its melting point with the resistive heater and cooling down rapidly. To make it crystalline, the heater is switched on for a short time period (~50 ns) to heat the volume above its crystallization temperature.

one, a short write pulse is applied that heats the programmable volume just below its melting point and holds it there for a certain time. This SET operation limits the write speed of the PCRAM because of the required duration to crystallize the phase-change material. The write speed is in the order of tens of nanoseconds (~50 ns), depending on the used material. In the RESET operation, the memory cell is switched to the amorphous "0" state by applying a larger electrical current (~mA) in the order of 100 ns that heats up the phase-change material just above the melting point.

The PCRAM access time is presently longer than for the MRAM. However, the write/erase time improves with scaling, as the active programmable volume is getting smaller and shorter write/erase pulses are sufficient to switch the state of the phase-change material. The read operation shows a very good scaling that gives PCRAM the highest market potential among all nonconventional memory concepts that could replace the Flash memory in the future. A big disadvantage is the endurance of a PCRAM cell (10^7–10^{12}), being above the conventional Flash cell but several orders of magnitude away from the DRAM ($>10^{15}$).

2.4 Semiconductor Nanomemories

The semiconductor memory industry has already entered the "nano-world" years ago and has introduced a feature size of 45 nm in the year 2008. To improve the device performance of a conventional Flash cell, nanomaterials have been used as replacements for the floating gate. For instance, replacing the Si floating gate with Si nanocrystals—Si particles with diameters in the order of 1–10 nm—leads to the advantage that charge leakage from any particular nanocrystal does not discharge the complete floating gate. The ultimate goal would be the usage of only one nanocrystal as one information bit to build a single electron memory.

This section will briefly review semiconductor memories based on charge traps and silicon nanocrystals in Section 2.4.1 and a single electron memory based on silicon QD in Section 2.4.2.

2.4.1 Charge Trap Memories

One of the major drawbacks is the limited number of write/erase cycles of a conventional Flash cell, the poor endurance. This will be a major problem during scaling down the feature size and simultaneously the thickness of the SiO_2 tunneling barrier. A single defect in the tunneling oxide will always destroy the entire Flash cell. The simplest way to improve the endurance of a conventional Flash memory and, hence, to extend the scalability is to replace the floating gate by a charge trapping material, schematically illustrated in Figure 2.7.

A charge trapping memory cell can be realized in different approaches. One possibility is to use a dielectric material that stores the charges in deep traps. Silicon nitride is the most established among such materials and the memory cell is often referred to as oxide–nitride–oxide, short just ONO (Lai 2008). A variation of this type of charge trapping device is the SONOS (silicon–oxide–nitride–oxide–silicon) (White et al. 2000). Figure 2.7a illustrates the floating-gate structure and the SONOS device principle. The thin silicon-nitrite (Si_3N_4) film is the storage unit where a single defect in the oxide layer will not discharge the complete memory cell. In addition, SONOS cells show a reduced cross talk and lower read/write voltages.

A second alternative to create a charge trapping memory is the usage of nanoparticles as the floating gate. The first attempts of such memory devices were made by Lambe and Jaklevic

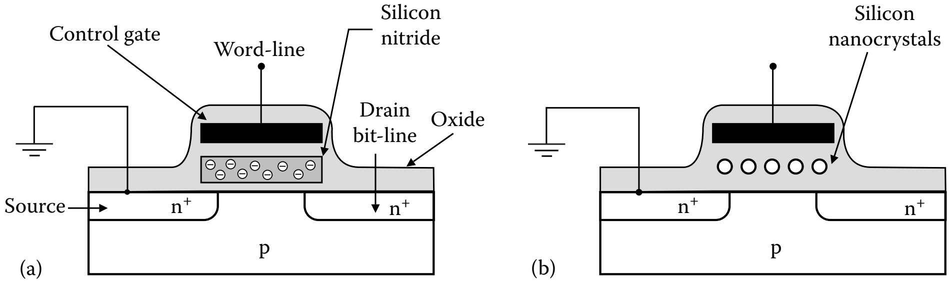

FIGURE 2.7 (a) Schematic picture of a SONOS nonvolatile charge trapping memory that consists instead of a Si floating gate of an oxide–nitride–oxide (ONO) sandwich. The nitride layer can store charges (electrons or holes) in deep traps. (b) A nanocrystal Flash based on silicon nanoparticles.

(1969) 40 years ago. They used an array of Al droplets—with a diameter in the order of 10 nm—embedded in thin oxide and observed a memory effect in capacitance measurements. Tiwari et al. (1996) embedded Si nanocrystals into a floating-gate structure to improve the Flash concept with nanomaterials. The principle of such memory devices is depicted in Figure 2.7b. The Si nanocrystals are fabricated through spontaneous decomposition during chemical vapor deposition onto the tunneling oxide barrier. These particles have a typical size in the order of about 1–10 nm. Replacing the floating gate with these Si nanocrystals has the advantage that charge leakage through defect inside the oxide layer will not completely discharge the whole memory cell. Thinner tunneling barriers can be used and an improved endurance is observed.

2.4.2 Single Electron Memories

One charge carrier inside a single nanocrystal or QD for one information bit is the ultimate goal for a memory cell. Such devices were realized with a single polysilicon QD embedded into an oxide matrix (Guo et al. 1997). This device showed a storage time of 5 s at room temperature and is schematically depicted in Figure 2.8. It was fabricated by electron beam lithography (EBL) and reactive ion etching (RIE). A source–drain channel with a width between 25 and 120 nm was created onto an oxide layer, followed by a second oxide layer and the polysilicon dot. The dot had a width of about 20 nm and is covered again by silicon oxide and polysilicon as a gate electrode. By an appropriate bias on the control gate, the single dot is charged and discharged at room temperature and the charge state—the read-out of the information—was possible via a resistance measurement of the source–drain channel.

This measurement demonstrates perfectly the possibility of single charge storage and read-out even at room temperature without conductance quantization inside the source–drain channel. However, these structures were fabricated via EBL, which is not suitable for mass production and the storage time is limited to 5 s at 300 K.

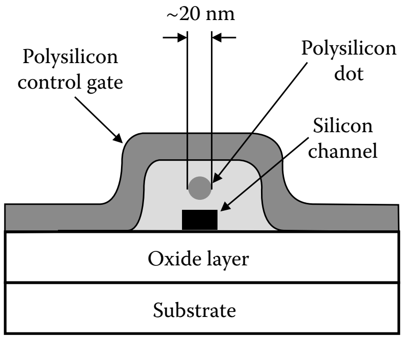

FIGURE 2.8 Schematic picture of single electron memory device based on a polysilicon QD above a silicon channel. The silicon channel is connected to a source and drain contact to measure the resistance of the channel. The charge state can be written and erased by the polysilicon control gate. (After Guo, L. et al., *Appl. Phys. Lett.*, 70(7), 850, 1997.)

2.5 A Nanomemory Based on III–V Semiconductor Quantum Dots

In this section, a memory concept based on III–V nanomaterials—self-organized QDs—that has the potential to overcome the drawbacks of the current conventional Flash and DRAM is presented (Geller et al. 2006a). Such a nanomemory cell should provide long storage times (>10 years) and good endurance (>10^{15} write/erase cycles) in combination with even better read/write access time than the DRAM (<10 ns). In addition, an ultimate memory cell should store one information bit by means of as few carriers as possible, let it be electrons or holes, to compete with the ongoing miniaturization in the semiconductor memory industry. With a lateral feature QD size of about 7–20 nm—depending on the material combination and growth conditions—to have a single QD as one information bit is a must for the future. Comparing this QD size with the predicted feature size of conventional Flash presented in Figure 2.4, one expects that the QD memory could be competitive for the next 10 years. A major challenge for such QD-based memory will be to address a single self-organized QD. However, a lot of work has been focused on QD alignment in two-dimensional arrays that would allow addressing a single self-organized QD in the future. An overview on QD alignment can be found in Kiravittaya et al. (2009).

This section concentrates on a brief introduction to III–V semiconductor hetereostructures and self-organized QDs in Sections 2.5.1 and 2.5.2, respectively. The last subsection will focus on a nanomemory concept based on these QDs.

2.5.1 III–V Semiconductor Materials

Metal organic chemical vapor deposition (MOCVD) or molecular beam epitaxy (MBE) are epitaxial growth techniques to deposit different defect-free semiconductor materials with different electronic properties (e.g., band offsets) on each other. Such a semiconductor crystal, made of more than one material, is referred to as a “heterostructure,” and the interface between the two materials as “heterojunction.” The electronic properties of the entire heterostructure are mainly determined by this heterojunction. Heterostructures pave the way to design devices with tailored optical or electronic properties.

The use of heterojunctions has led to the development of many device families like the double heterostructure (DHS) laser or the hetero-bipolar transistor, both invented in 1963 by Alferov and Kroemer (Alferov and Kazarinov 1963, Kroemer 1963) and honored with the Nobel prize in physics in the year 2000. Alferov built the first DHS semiconductor laser using gallium arsenide (GaAs) and aluminum arsenide (AlAs) in 1969.

By using epitaxial growth techniques, it is also easily possible to create ternary alloys, for instance $Al_xGa_{1-x}As$, which exhibit a bandgap smoothly varying between the binary alloys. Figure 2.9 shows the energy bandgap as a function of lattice constant for different III–V compound semiconductors (points) and their ternary alloys (lines) at zero temperature. The III–V compound

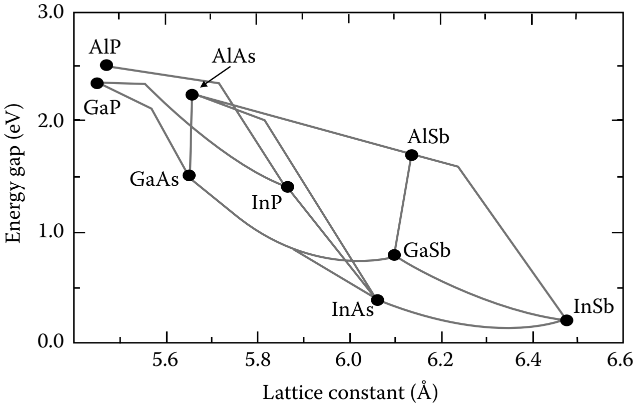

FIGURE 2.9 Energy bandgap as a function of lattice constant for different III–V compound semiconductors (points) and their ternary alloys (lines) at zero temperature. (After Vurgaftman, I. et al. *J. Appl. Phys.*, 89, 5815, 2001.)

semiconductors and their ternary alloys offer a large variety of material combinations in heterostructure devices with the possibility of a controlled band-structure engineering.

For a semiconductor heterostructure, besides the energy bandgap, the relative energetic position of the valence and conduction band at the interface is of central importance: the so-called band alignment. The band alignment and the bandgap control the band offsets in the conduction and valence band. The band offset is simply defined as the discontinuity in the band edges at the interface between the two semiconductors. Figure 2.10 shows the various types of band alignments that can arise for semiconductor heterojunctions: type I, type II staggered, and type II broken-gap. For a type I alignment, the bandgap of the narrower gap semiconductor E_g^2 lies completely within the bandgap of the wider gap E_g^1, shown in Figure 2.10a. This band alignment occurs in a large number of technically important heterojunctions, for example, InAs/GaAs, GaAs/AlGaAs, and GaSb/AlSb. For a type II staggered band alignment, the bandgaps of the two materials show only partial overlap (Figure 2.10b). An example for a type II staggered band alignment is GaSb/GaAs, which plays an important role for memories based on self-organized QDs. A type II broken-gap alignment occurs, when the bandgaps of the two materials do not overlap at all (Figure 2.10c).

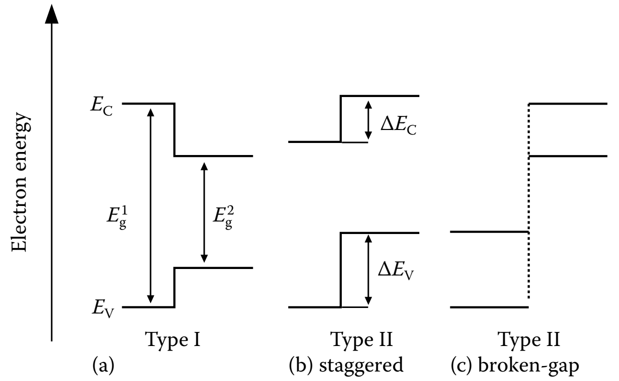

FIGURE 2.10 Types of band alignment at a heterojunction consisting of two semiconductors with different bandgap E_g^1 and E_g^2. The conduction and valence band edges have been labeled with E_C and E_V, while the conduction and valence band offsets have the label ΔE_C and ΔE_V, respectively.

2.5.2 Self-Organized Quantum Dots

Self-organized QDs are low-dimensional heterostructures (Bimberg et al. 1998) that confine electrons and/or holes in all three spatial directions. Such a zero-dimensional system exhibits the electronic and optical properties like "artificial" atoms and offers lots of possibilities for new applications, for example, in electronic devices such as field effect transistors (Kim et al. 2000, Koike et al. 2000) as well as in optoelectronic devices such as lasers (Kirstaedter et al. 1994, Heinrichsdorff et al. 2000, Kuntz et al. 2004, Hopfer et al. 2006), detectors (Campbell et al. 1997, Chu et al. 2001), amplifiers (Lämmlin et al. 2006), and single-photon sources (Santori et al. 2002, Lochmann et al. 2006) or memory cells as described in this chapter.

The fabrication of QDs is based on self-organization effects on crystal surfaces first reported by Stranski and Krastanow (1938). On a flat substrate surface, a layer of a material with a different lattice constant is deposited, leading to a strained wetting layer. The strain energy can be reduced by forming small islands from the material of the wetting layer. These small structures with dimensions in the nanometer range can be very regular in size and shape and can confine carriers in all three dimensions (Shchukin et al. 1995, 2003). The term Stranski–Krastanow growth mode (SK-growth) was reintroduced in heteroepitaxy for the formation of islands on an initially two-dimensional layer. The formation of coherent, that is, defect free islands as a result of SK-growth is today exploited to fabricate QDs (Eaglesham and Cerullo 1990, Mo et al. 1990). Self-organized QDs have often a pyramidal or truncated pyramidal shape (Heinrichsdorff et al. 1996, Eisele et al. 1999) with typical base widths and heights in the order of 7–30 nm and 2–5 nm, respectively, depending on the growth conditions and material combinations.

The reduction of the extension of the QDs to values below the de Broglie wavelength leads to charge carrier localization in all three directions and the disappearance of the continuous $E(k)$ dispersion, called band structure. As in a real atom, the resulting energy level structure is discrete. The QD provides a confining potential of finite depth for electrons and/or holes. Figure 2.11 shows schematically two types of QDs: type I QDs (like InAs/GaAs) confine electrons and holes, while a type II QD confines either electrons or holes. The depicted type II GaSb/GaAs QD system has a localization potential for holes only and provides a barrier for the electrons in the conduction band. The localization energy represents the energy difference of the confined electron or hole states with respect to the conduction or valence band edge, equivalent to the binding energy.

The single particle states in self-organized QDs depend on two structural properties: size and shape, and the chemical composition. The influence of the QD geometry on the electron and

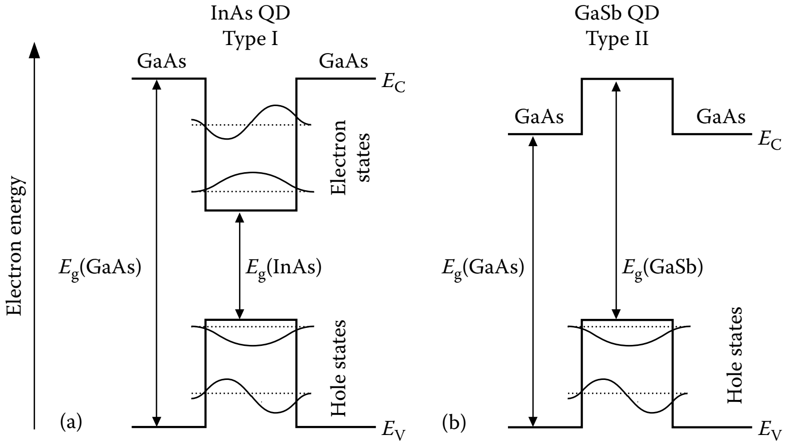

FIGURE 2.11 Schematic picture of the electron and hole states in a type I InAs/GaAs (a) and in a type II GaSb/GaAs (b) QD system. A type I QD confines electrons as well as holes, while a type II QD confines either electrons or holes. The depicted type II GaSb/GaAs QD confines only holes and has a potential barrier for electrons in the conduction band.

hole states can be understood in principle from a simple particle-in-a-box model (Grundmann et al. 1995, Stier et al. 1999). For an infinite potential well, the quantum-mechanical solution predicts energy separation and nonzero energy for the lowest level, depending on the size of the box. In contrast, the confining barrier in buried QDs is of finite height, given by the band offsets. Only a finite number of bound states exist, depending on the QD size and shape. As the QD becomes smaller, the separations between the electron and hole levels increase—quantum size effect—and all states are lifted in energy toward the valence/conduction band edge. Below a certain size, no bound electron and/or hole state exists.

Self-organized QDs can trap electrons or holes, storing these charge carriers for a certain time depending on the emission processes involved. Therefore, QDs may store information like a memory cell. Carrier emission limits the storage time in a QD-based memory. Confined charge carriers can escape by the following four mechanisms: thermal activation, tunneling, phonon-assisted tunneling, and optical activation. Phonon-assisted tunneling is a combination process of tunneling and thermal activation. The charge carrier is thermally activated by phonons to a higher energy state E_1 (a real or virtual state) and tunnels subsequent to the conduction or valence band through a barrier. The mechanisms are schematically depicted in Figure 2.12 for electron escape into the conduction band. Whether a charge state and, hence, the information bit is destroyed by one of the emission processes depends on the electric field and the temperature in comparison to the localization energy of the charge carriers E_{Loc}. If the electric field is weak and the thermal energy k_BT is large, the charge carriers will be emitted due to thermal activation. If the electric field is strong and the thermal energy is small in comparison to the localization energy, the charge carriers will be released due to tunneling. Phonon-assisted tunneling will occur for a situation, were the tunneling rate from the state E_1 is equal to the thermal activation rate from the ground state E_0 to the state E_1.

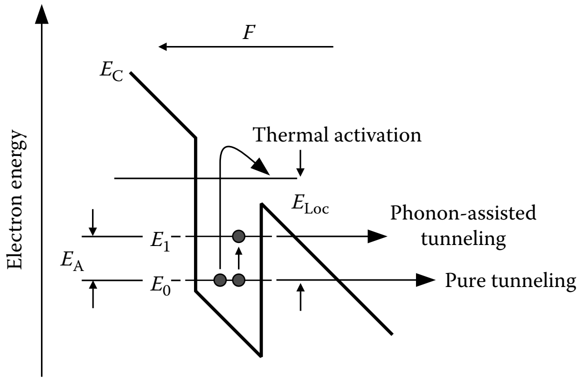

FIGURE 2.12 The possible carrier emission processes in an applied electric field: thermal activation, phonon-assisted tunneling and tunneling are shown for electron escape from a QD ground state into the conduction band. The term E_{Loc} represents the localization energy of the charge carriers in the ground state while E_A is the thermal activation energy for phonon-assisted tunneling.

To understand the limitations in storage time in a nanomemory based on self-organized QDs all above mentioned emission processes have to be taken into account. However, in the storage situation of a memory, the electric field should be sufficiently small such that tunneling and phonon-assisted tunneling can be ruled out and only pure thermal activation is of interest. The average time for a thermally activated emission process from the ground state of the QDs to the band edge (which presents an upper limit for the average storage time of the charge carriers) depends in this situation exponentially on the localization energy of the charge carriers; cf. Equation 2.27 in Section 2.6.2.1. Therefore, III–V material combinations with large band offsets that provide large localization energies are needed to build a nonvolatile nanomemory based on self-organized QDs. After a presentation of the QD-based memory concept, the next sections will concentrate on storage times and localization energies in different III–V self-organized QDs and their suitability for a nonvolatile memory.

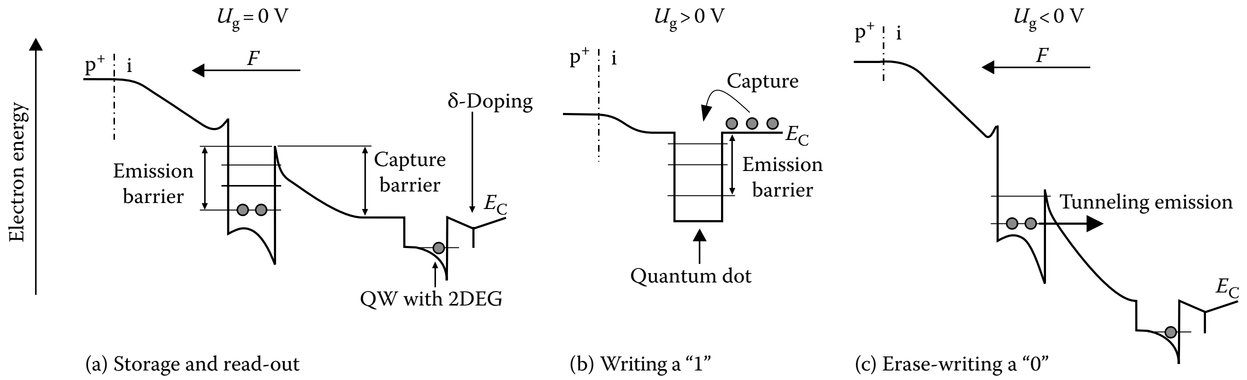

(a) Storage and read-out (b) Writing a "1" (c) Erase-writing a "0"

FIGURE 2.13 Schematic illustration of the (a) storage and read-out, (b) write, and (c) erase process in a possible future QD-based nanomemory.

2.5.3 A Memory Cell Based on Self-Organized QDs

Here, a memory concept is presented where the self-organized QDs are embedded in a p–n or p–i–n diode structure (Geller et al. 2006a). The QDs act as storage units, since they can be charged with electrons or holes representing the "0" (uncharged QDs) and "1" (charged QDs) of an information bit. An emission barrier is needed to store a "1" in such a memory concept. The barrier height—which is related to the localization energy of the charge carrier (Figure 2.13a)—can be varied by varying the material and size of the QDs and the material of the surrounding matrix (see Section 2.5.1). If the localization energy is increased, a longer storage time of the charge carriers is expected. Furthermore, a capture barrier is needed to store a "0" (Figure 2.13a). This barrier protects an empty QD cell from unwanted capture of charge carriers. In this concept, the capture barrier is realized by using the band bending of a p–i–n or p–n diode.

The major advantage in using a diode structure is the possibility to tune the height of the barriers by an external bias. Therefore, during the write process (Figure 2.13b), the capture barrier can be eliminated by a positive external bias between the gate and source contact ($V_g > 0$ V) and the charge carriers can directly relax into the QD states. This allows to benefit from another advantage of self-organized QDs: the charge carrier relaxation time into the QD states is in the range of picoseconds at room temperature (Müller et al. 2003, Geller et al. 2006b) (see Section 2.7), enabling very fast write times in a QD-based memory. In addition, a very good endurance of 10^{15} write operations should be feasible. To erase the information, the electric field is increased within the QD layer by a negative external bias between the gate and source ($V_g < 0$ V), such that tunneling of the charge carriers occurs (cf. Figure 2.13c).

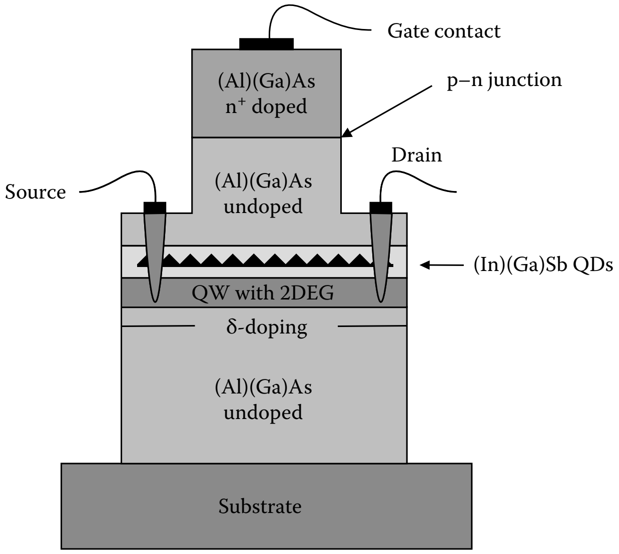

FIGURE 2.14 Schematic picture of the layer sequence for a possible QD-based memory. The QDs are located below the p–i–n junction and a 2DEG is placed below the QD layer to detect the charge state.

Figure 2.14 shows schematically the device structure of such a QD-based memory, where the QDs are charged with holes to represent an information bit. The doping sequence of the p- and n-regions would be inverted for electron storage. The distance to the junction is adjusted, such that the QDs are inside the depletion region for zero bias ($V_g = 0$ V in Figure 2.13a). The read-out of the stored information is done by a two-dimensional electron gas (2DEG) for an electron and a two-dimensional hole gas (2DHG) for a hole storage device. The 2DEG or 2DHG is situated 10–50 nm below the QD layer and is filled with charge carriers, provided by the additional n- or p-δ-doping, respectively. Stored charge carriers inside the QDs reduce the conductivity of the 2DEG/2DHG; hence, a higher resistance is measured between the source/drain contacts, in analogy to the read-out in a conventional Flash cell.

2.6 Capacitance Spectroscopy

To elucidate the potential of nanomaterials like self-organized QDs for future memory applications the localization energy and the carrier storage time at room temperature has to be determined for different III–V heterostructures. The first goal is to find a material combination that yields a minimum storage time of milliseconds at 300 K, the benchmark for present DRAM.

The capacitance spectroscopy is a powerful tool to study the energy levels and charge carrier emission times in QDs. It has been widely used for different device geometries, for instance in p–i–n diodes to investigate the tunneling dynamics and localization energies in self-organized InAs/GaAs QDs (Fricke et al. 1996, Luyken et al. 1999) or to probe single-electron levels in QDs that are laterally confined inside a 2DEG (Ashoori et al. 1992, Ashoori 1996).

In this section, capacitance spectroscopy is used to determine storage and emission times as well as localization energies in self-organized QDs. The QDs are embedded in a GaAs matrix nearby an abrupt p–n junction that forms a depletion region. This different capacitance spectroscopy method consists of measurements of the depletion capacitance of a p–n or Schottky diode, it is the "capacitance spectroscopy of a depletion capacitance" (Lang 1974).

2.6.1 Depletion Region

The work function or Fermi energy with respect to the vacuum level of a metal or doped semiconductor differs for different metals or semiconductors having a different doping concentration. If a metal and a semiconductor (or two differently doped semiconductors—a p–n diode) are in electric contact, the free charge carriers are exchanged between the different materials until a thermodynamic equilibrium is reached and the Fermi energy is equal throughout the entire structure. Sometimes only ionized donors and acceptors are present in the vicinity of the interface while all free charge carriers are moved away in the semiconductor (Figure 2.15). The layer depleted from free charge carriers is usually referred to as the "depletion region" (Sze 1985, Blood and Orton 1992). The width of the depletion region depends on the doping concentration and the potential difference between the materials, of which the latter can easily be modified by an externally applied bias.

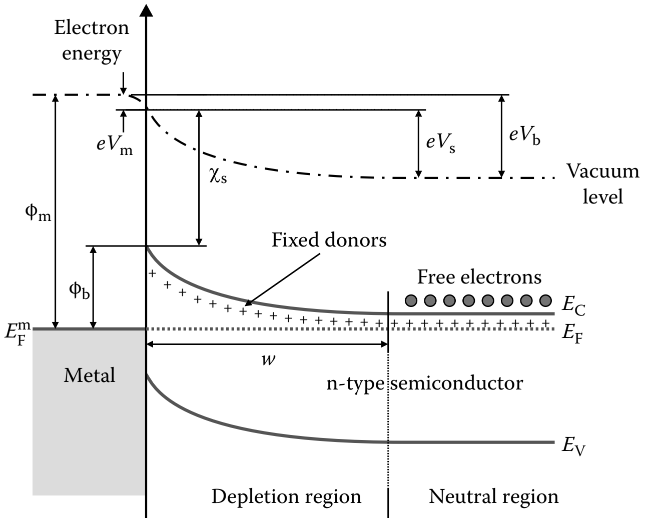

FIGURE 2.15 Energy band diagram of a depleted metal-semiconductor Schottky contact. The width of the depletion region w depends on the doping concentration in the semiconductor. A build-in voltage V_b is present without external bias due to the band bending of the depletion region.

A metal-semiconductor contact can be described by the Schottky model and is referred to as the "Schottky contact." A junction of a p-doped and n-doped semiconductor is a "p–n junction." Both types of contacts provide a depletion region and they are briefly described in the following.

2.6.1.1 Schottky Contact

A metal-semiconductor contact is usually described in the framework of the Schottky model, which is an acceptable approach to construct the band diagram of the contact (Figure 2.15). Within this model, the barrier height ϕ_b inside the metal is independent or just weakly dependent on the applied external bias.

According to the Schottky model, the energy band diagram is constructed by reference to the vacuum level, defined as the energy of a free electron to rest outside the material. The work function of the metal ϕ_m and the electron affinity of the semiconductor χ_s are defined as the energies required to remove an electron from the Fermi level of the metal or the semiconductor conduction band edge, respectively, to the vacuum level. These properties are assumed constant in a given material and it is further assumed that the vacuum level is continuous across the interface. The Fermi levels in the metal and semiconductor must be equal in thermal equilibrium. These conditions result in a band diagram for the interface as shown in Figure 2.15.

Since the vacuum level is the same at the interface of the metal and the semiconductor, a step between the Fermi level in the metal and the conduction band edge E_C of the semiconductor occurs. This is the barrier height ϕ_b given by*

$$\phi_b = \phi_m - \chi_s - eV_m. \qquad (2.1)$$

The band bending in the metal is very small due to the large density of electron states; hence, eV_m can be neglected. Therefore, the Schottky barrier height is usually written as

$$\phi_b = \phi_m - \chi_s. \qquad (2.2)$$

For increasing distance from the interface, the conduction band energy decreases. At the end of the depletion region, it has the same value as in the neutral semiconductor with respect to the Fermi level. The resulting band bending is an effect of the removed free charge carriers, leaving behind a distribution of fixed positive charges from ionized donors. The depletion region ends at that position, where the bands become flat and the associated electric field is zero. The width of the depletion region w is determined by the net ionized charge density according to Poisson's equation (see Section 2.6.1.3).

Since the density of states in the metal is much greater than the doping density in the semiconductor, the depletion width in the metal is much smaller. Therefore, it can be assumed that the potential difference across the metal near the contact (V_m) is negligibly small compared to that in the semiconductor (V_s).

---

* For a Schottky barrier, ϕ_m must exceed χ_s otherwise the bands bend in the opposite direction.

The total zero bias band bending of the Schottky contact, also referred to as "build-in potential" or "build-in voltage" V_b can be written as

$$eV_b \approx eV_s = \phi_m - \chi_s - (E_C - E_F) = \phi_b - (E_C - E_F). \qquad (2.3)$$

Experimental values for various metal Schottky contacts on GaAs can be found in Myburg et al. (1998).

2.6.1.2 p–n Junction

The band diagram of an abrupt p–n junction, in Figure 2.16, is considered in a similar manner. Again, two rules are used to construct the diagram: (1) the vacuum level is continuous and (2) the Fermi energy is constant across the junction in thermal equilibrium. For the same p- and n-doped semiconductor the electron affinity χ_s is the same on both sides of the junction. Therefore, the band bending is caused entirely by the difference in the Fermi level with respect to the conduction band of the two differently doped materials. From Figure 2.16, using subscripts to denote the n and p side, one obtains for zero external bias

$$(E_C^p - E_F) + \chi_s = eV_b + \chi_s + (E_C^n - E_F), \qquad (2.4)$$

hence, the build-in voltage is

$$eV_b = E_g - (E_F - E_V^p) - (E_C^n - E_F), \qquad (2.5)$$

where E_g is the energy gap of the semiconductor. The build-in voltage as the total electrostatic potential difference between the p-side and the n-side is temperature-dependent and a function of the fixed donor and acceptor charges of density N_d and N_a (Sze 1985):

$$V_b = \frac{k_B T}{e} \ln\left(\frac{N_a N_d}{n_i^2}\right), \qquad (2.6)$$

where n_i is the intrinsic carrier density

$$n_i = \sqrt{N_C N_V} \exp\left(-\frac{E_g}{2k_B T}\right), \qquad (2.7)$$

with the effective density of states in the valence N_V and conduction band N_C, respectively. A typical doping density of 10^{17} cm^{-3} yields for GaAs a build-in voltage of 1.3 V at 300 K.

In a p–n junction, depletion regions on each side of the contact exist, where the fixed donor and acceptor charges lead to the band bending. Since the total charge in the p–n diode is zero (charge neutrality), in the depletion approximation (Blood and Orton 1992) of an abrupt depletion layer edge

$$N_a w_p = N_d w_n \qquad (2.8)$$

must hold. For similar doping concentrations, the depletion width on the p- and n-side of the semiconductor will be comparable. However, for the purpose of material characterization (such as capacitance experiments) doping concentrations are often chosen such that the depletion region is situated almost entirely on one side of the junction. The depletion region of such an asymmetrical p–n junction resembles the depletion region of a Schottky contact. Such asymmetrical junctions are briefly denoted as p$^+$–n or n$^+$–p junction with $N_a \gg N_d$ or $N_d \gg N_a$, respectively.

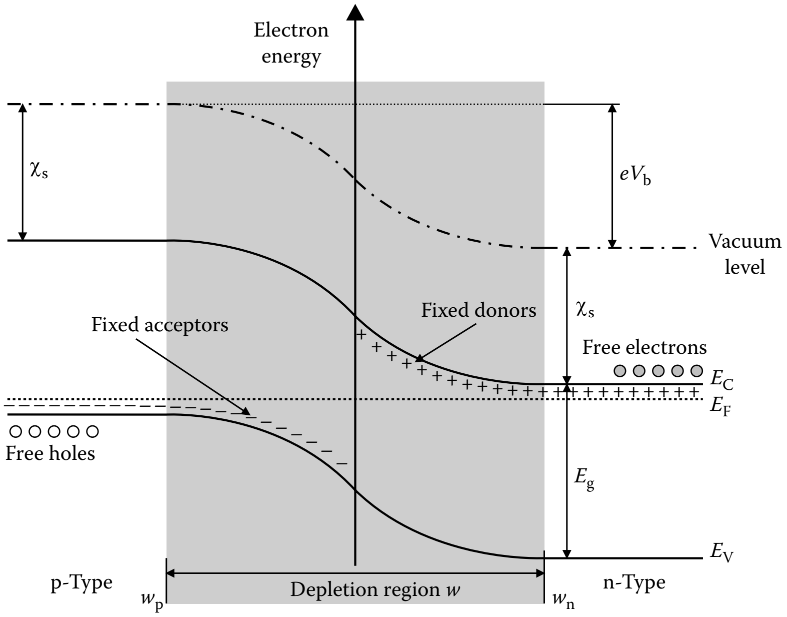

FIGURE 2.16 Energy band diagram of a p–n junction.

2.6.1.3 Width of the Depletion Region

The entire band bending across the depletion region is defined by the sum of the build-in voltage of the contact V_b (Equation 2.5 or 2.6), and the applied external bias V_a in reverse direction: $V = V_b + V_a$. The width of the depletion region and the electric field can be calculated using Poisson's equation. The electrostatic potential ψ at any point is given by

$$-\frac{\partial^2\psi}{\partial x^2} = \frac{\partial F}{\partial x} = \frac{\rho(x)}{\varepsilon\varepsilon_0}, \tag{2.9}$$

where

F is the electric field
ε_0 is the vacuum permittivity
ε is the dielectric constant of the semiconductor material

If the donors or acceptors are entirely ionized, the charge density ρ is eN_d or eN_a, respectively:

$$-\frac{\partial^2\psi}{\partial x^2} = \frac{eN_d}{\varepsilon\varepsilon_0} \quad \text{for } 0 \le x \le w_n. \tag{2.10}$$

For a Schottky contact or an abrupt asymmetric p⁺–n junction, integration of Equation 2.9 yields

$$F(x) = F_0 + \frac{eN_d x}{\varepsilon\varepsilon_0} \quad \text{for } 0 \le x \le w_n, \tag{2.11}$$

while $F(x) = 0$ for $x < 0$ and $x > w_n$. Hence, using the approximation $F(x) = 0$ at the edge of the depletion region, gives the boundary condition for the integration constant:

$$F_0 = -\frac{eN_d w_n}{\varepsilon\varepsilon_0}, \tag{2.12}$$

and represents the electric field at the interface $F(0)$, where it has its maximum. Therefore, the electric field F across the depletion region in an n-doped semiconductor with the donor concentration N_d is

$$F(x) = \frac{eN_d}{\varepsilon\varepsilon_0}(x - w_n) \quad \text{for } 0 \le x \le w_n. \tag{2.13}$$

Analogously, the electric field distribution across the depletion region on the p-doped side is

$$F(x) = -\frac{eN_a}{\varepsilon\varepsilon_0}(x + w_p) \quad \text{for } -w_p \le x \le 0. \tag{2.14}$$

The potential distribution across the depletion region for a p⁺–n junction inside the n-doped semiconductor is now obtained by integration of Equation 2.13:

$$\psi(x) = -\int_0^x \frac{eN_d}{\varepsilon\varepsilon_0}(x - w_n)\,dx = \frac{eN_d w_n}{\varepsilon\varepsilon_0}\left(x - \frac{x^2}{2w_n}\right) + \psi(0), \tag{2.15}$$

with $\psi(0) = 0$ as a reference for the potential distribution. For an asymmetric p⁺–n junction, one notices that the band bending in the present approximation only occurs in the n-doped region of the junction. The contact potential is equal to the total band bending, that means, equal to the build-in voltage plus the external bias $\psi(0) = -V = -(V_b + V_a)$. Since the potential on the edge of the depletion region is zero, $\psi(w_n) = 0$, one obtains

$$V = V_b + V_a = \frac{eN_d}{2\varepsilon\varepsilon_0} w_n^2, \tag{2.16}$$

or the expression for the depletion width

$$w_n = \sqrt{\frac{2\varepsilon\varepsilon_0}{eN_d} V}. \tag{2.17}$$

If the p–n junction is symmetric and the depletion region extends into the p- and n-doped semiconductor, Equation 2.16 is modified to

$$V = V_b + V_a = \frac{eN_d}{2\varepsilon\varepsilon_0} w_n^2 + \frac{eN_a}{2\varepsilon\varepsilon_0} w_p^2, \tag{2.18}$$

with the net acceptor doping concentration N_a and the width of the depletion region w_p in the p-doped side. The total depletion width w is given by

$$w = (w_n + w_p) = \sqrt{\frac{2\varepsilon\varepsilon_0}{e}\left(\frac{N_a + N_d}{N_a N_d}\right) V}. \tag{2.19}$$

The assumption of an abrupt depletion layer edge, is the so-called depletion approximation. A more general description of the depletion region can be found in reference Blood and Orton (1992).

2.6.1.4 Depletion Layer Capacitance

The capacitance that arises from the depletion region is the so-called "depletion capacitance." In comparison to a plate capacitor, here, the charge is accumulated inside the depletion region of the p–n diode. The number of ionized donors depends linearly on the width w_n, while again the width is a function of the square root of the voltage (Equations 2.17 and 2.19). Hence, the capacitance does not depend linearly on the voltage and has to be defined differentially for a small bias signal ΔV:

$$C = \lim_{\Delta V \to 0} \frac{\Delta Q}{\Delta V} = \frac{dQ}{dV}. \tag{2.20}$$

Using Equation 2.17, the entire stored charge inside the depletion region of a Schottky diode or an abrupt asymmetric p⁺–n junction with area A is

$$Q = eN_d A \cdot w_n(V) = A\sqrt{2e\varepsilon\varepsilon_0 N_d V}, \tag{2.21}$$

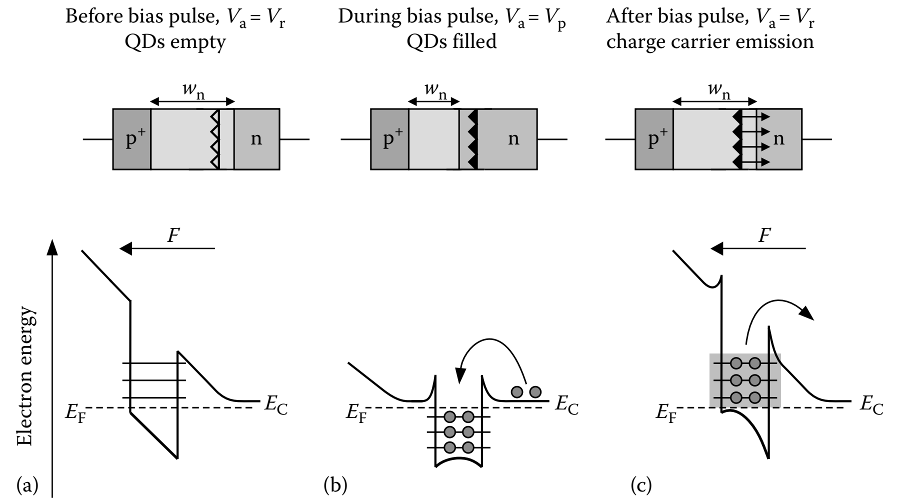
Before bias pulse, $V_a = V_r$
QDs empty
During bias pulse, $V_a = V_p$
QDs filled
After bias pulse, $V_a = V_r$
charge carrier emission
w_n
p^+
n
F
Electron energy
E_F
E_C
(a)
(b)
(c)

FIGURE 2.17 DLTS work cycle of a p⁺–n diode with embedded QDs. The upper row schematically shows sketches of the devices for three different bias situations: (a) before, (b) during, and (c) after the bias pulse, respectively. The lower row shows the corresponding band diagrams.

and the capacitance is

$$C(V) = \left|\frac{dQ}{dV}\right|_V = A\sqrt{\frac{\varepsilon\varepsilon_0 eN_d}{2V}} = \frac{\varepsilon\varepsilon_0 A}{w_n}. \quad (2.22)$$

The depletion capacitance resembles the capacitance of a plate capacitor with a distance of w_n between the plates and a dielectric with relative permittivity ε, although the charge is actually stored in the volume rather than on the edges of the space charge region.

2.6.2 Capacitance Transient Spectroscopy

The depletion width depends on the applied voltage and the doping concentration, that is, on the charge stored inside the space charge region. As a consequence, the measured capacitance is sensitive to the carrier population inside deep levels or QDs situated inside the depletion region (Kimerling 1974). Semiconductor QDs can be considered as deep levels (or traps), storing few or single charge carriers for a certain time. Time-resolved capacitance spectroscopy allows investigating the electronic structure of QDs, since the carrier dynamics can be studied in detail.

Historically, time-resolved capacitance spectroscopy was initially used to study and characterize deep levels caused by impurities or defects. In this context it is usually referred to as "deep level transient spectroscopy" (DLTS) or "capacitance transient spectroscopy" (Sah et al. 1970, Lang 1974, Miller et al. 1977, Rhoderick and Williams 1988, Grimmeiss and Ovrén 1981, Blood and Orton 1992). Besides the determination of activation energies and capture cross sections, DLTS also permits to obtain a depth profile of the trap concentration and to investigate the influence of the electric field on the emission process. As QDs act more or less as deep levels, the DLTS has been used successfully to study thermal emission processes, activation energies, and capture cross sections of various QD systems (Anand et al. 1995, 1998, Kapteyn et al. 1999, 2000a,b, Schulz et al. 2004).

2.6.2.1 Measurement Principle

In the following, the DLTS measurement principle will be described for QDs in a p–n or Schottky diode structure. A more detailed description can be found in Blood and Orton (1992).

First, a single layer of QDs with density N_{QD} per area in an n-doped material (with doping concentration N_d) shall be considered. The work cycle of a DLTS experiment of a p⁺–n structure with QDs embedded is depicted in Figure 2.17. The upper row displays schematically the p–n diode while the lower row depicts the corresponding potential distribution of the conduction band.

In an initial step (Figure 2.17a), a reverse bias V_r is chosen such that the depletion region extends well over the QDs. The QDs are completely depleted from charge carriers and the Fermi level is below the QD levels. During the pulse V_p, the depletion region is shorter than the distance of the QD layer from the p–n interface and the QDs are consequently filled with carriers—electrons in Figure 2.17b. The Fermi level is now above the QD states. After switching back to the reverse bias situation (Figure 2.17c), the QDs are again situated inside the depletion region but still filled with electrons. The depletion width is larger for QDs filled with charge carriers, hence, the capacitance after the pulse V_p is smaller and increases again when electrons/holes are emitted due to thermal activation or tunneling emission.* By recording the depletion capacitance as a function of time, transients are observed, which represent the carrier emission

---

* The emission of carriers from QD states is usually slow, whereas the free carriers in the matrix material are considered to follow the change in the external bias instantaneously at the timescale of the experiment.

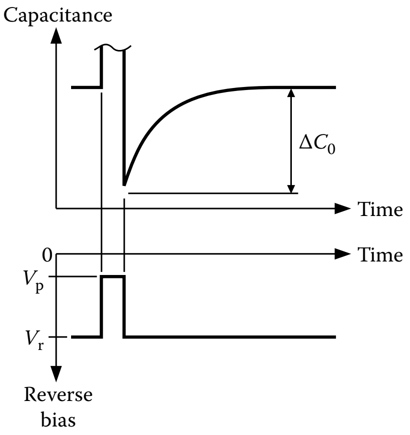

FIGURE 2.18 DLTS work cycle during a DLTS experiment. The lower part displays the external bias on the device as function of time, the upper part the corresponding capacitance.

processes from the QD states. The pulse sequence during the experiment and the measured capacitance transient is schematically depicted in Figure 2.18.

2.6.2.2 Rate Window and Double-Boxcar Method

The capacitance transient, recorded from a single emission process is usually mono-exponential having the time-constant τ. Hence, the capacitance transient is given by (Blood and Orton 1992)

$$C(t) = C(\infty) - \Delta C_0 \exp\left(-\frac{t}{\tau}\right), \tag{2.23}$$

where

$C(\infty)$ is the steady state capacitance at V_r

ΔC_0 is the entire change in the capacitance $C(t)$ for $t = \infty$ (see Figure 2.18)

The emission time constant can be determined by using Equation 2.23 and plotting the data on a semilogarithmic scale. However, in general, the capacitance transient for deep levels and QDs are multi-exponentials due to the ensemble broadening (Omling et al. 1983) and a linear fit is consequently impossible in most cases.

In order to obtain the emission time constant, the activation energies and capture cross section from multi-exponential transients, the rate window concept is commonly applied. One investigates the contribution to the observed emission process at a chosen reference time constant τ_{ref}. By plotting the boxcar amplitude $C(t_2) - C(t_2)$ for that reference time constant as function of temperature, the relation between temperature and emission rate can be evaluated for a thermal activated process.

In the rate window concept, the selection of the contribution for a chosen reference time constant is done by a simple technique: the DLTS signal at a certain temperature $S(T, t_1, t_2)$ is given by the difference of the capacitance at two times t_1 and t_2 by

$$S(T, t_1, t_2) = C(T, t_2) - C(T, t_1), \tag{2.24}$$

or with Equation 2.23

$$S(T, t_1, t_2) = \Delta C_0 \left[\exp\left(-\frac{t_2}{\tau(T)}\right) - \exp\left(-\frac{t_1}{\tau(T)}\right)\right]. \tag{2.25}$$

The two times t_1 and t_2 define the rate window, which has the reference time constant

$$\tau_{\text{ref}} = \frac{t_2 - t_1}{\ln(t_2/t_1)}. \tag{2.26}$$

Plotting $S(T, t_1, t_2)$ as function of temperature yields the DLTS spectrum, schematically depicted in Figure 2.19. A maximum appears at that temperature, where the emission time constant of the thermally activated process equals almost the applied reference time constant: $\tau(T) = \tau_{\text{ref}}$. A maximum appears only for a thermally activated process, a temperature independent tunneling process leads to a constant DLTS signal (Kapteyn

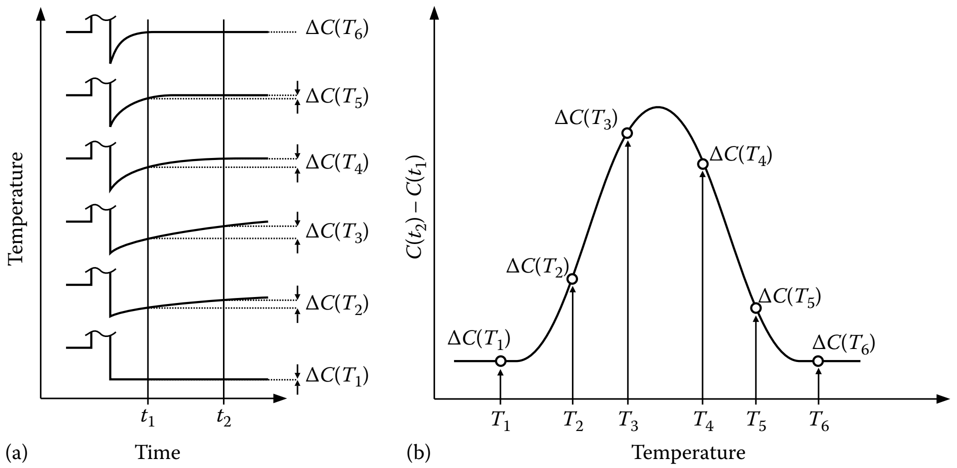

FIGURE 2.19 The evaluation of capacitance transients (a) for increasing temperature by a rate window, defined by t_1 and t_2, leads to a DLTS plot $[C(t_1) - C(t_2)]$ of a thermally activated emission process (b).

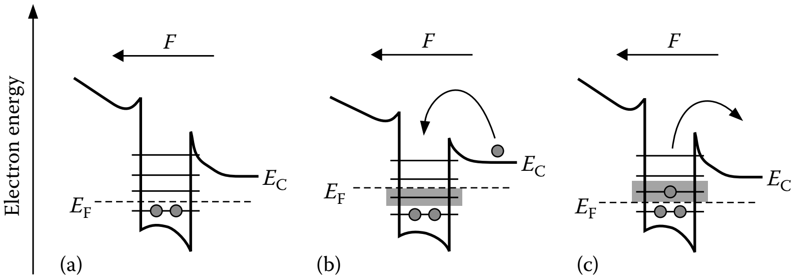

FIGURE 2.20 Work cycle during a charge-selective DLTS experiment (a) before, (b) during, and (c) after the bias pulse, respectively. The filling pulse V_p is chosen, that the emission or capture of approximately one charge carrier per QD is probed.

2001). This relation is also valid for inhomogeneous broadened DLTS spectra of self-organized QDs. The measured emission time constant is the average time constant at the maximum of the Gaussian ensemble distribution (Omling et al. 1983).

The thermal emission rate of the charge carriers, e_{th} is now given by (Lang 1974, Blood and Orton 1992)

$$e_{th} = \gamma T^2 \sigma_\infty \exp\left(-\frac{E_A}{kT}\right), \tag{2.27}$$

where

E_A is the thermal activation energy
σ_∞ is the apparent capture cross section for $T = \infty$
γ is a temperature-independent constant

Knowing the emission time constant $\tau = 1/e_{th}$ for different temperatures enables to derive the thermal activation energy and the capture cross section. For varying τ_{ref} different peak positions T_{max} are obtained. The plot of $\text{In}(T_{max}^2 \tau_{ref})$ as a function of T_{max}^{-1}, cf. Figure 2.22b, is called an Arrhenius plot and is a linearization of Equation 2.27. The slope yields the activation energy E_A and from the y-axis intersection, the apparent capture cross section σ_∞ at infinite temperature can be obtained.

In order to improve the signal-to-noise ratio (SNR), the capacitance transients near t_1 and t_2 can be averaged over an interval t_{av} (Day et al. 1979). The SNR in this case is found to scale as $\sim\sqrt{t_{av}}$ (Miller et al. 1977), and the reference time constant is in good approximation

$$\tau_{ref} = \frac{t_2 - t_1}{\ln\left((t_2 + (1/2)t_{av})/(t_1 + (1/2)t_{av})\right)}. \tag{2.28}$$

This method to improve the SNR is referred to as "Double-Boxcar" approach.

2.6.2.3 Charge-Selective Deep Level Transient Spectroscopy

As described in the previous section, usually the QDs are completely filled with charge carriers during the pulse bias V_p (Figure 2.17b). Consequently, after the pulse, the emission of many charge carriers from multiply charged QDs is probed. The activation energy of each emitted charge carrier depends on the actual charge state in such conventional DLTS experiments. The DLTS spectrum is broadened due to different emission processes from different QD states having different emission time constants, for instance, nicely observed in DLTS experiments on Ge/Si QDs (Kapteyn et al. 2000b).

In order to study the charge states in more detail, charge-selective DLTS probes the emission of approximately one charge carrier per QD. The principle of this method is schematically illustrated in Figure 2.20. Before the filling pulse, the QDs might be already filled with charge carriers up to the Fermi level. During the filling pulse (Figure 2.20b), the Fermi level is adjusted by the applied pulse bias, such that approximately one carrier per QD will be captured. To probe all QD states from the ground state up to the excited state the pulse bias with respect to the reverse bias is set to a constant height and the reverse bias is decreased step by step. After the bias pulse, the reverse bias is set to the initial condition. Now, the previously captured carrier is emitted again and the emission process of approximately one charge carrier per QD is observed. Narrow peaks appear now in the DLTS spectra, which are due to differently charged QDs (cf. Figure 2.23).

Moreover, even for an energy-broadened ensemble of QDs (normally measured in DLTS experiments) the charge-selective DLTS offers a simple method to probe the emission from many charge carriers in different QDs, all having the same activation energy. Decreasing the reverse bias and keeping the pulse bias height fixed (to ensure the emission of only one charge carrier per QDs) gives the activation energy starting at the ground states of the ensemble to the excited states and finishing at completely charged QDs.

2.7 Charge Carrier Storage in Quantum Dots

In this section, experimental results from capacitance spectroscopy measurements on different QD material systems are presented. This time-resolved method allows to derive thermal activation energies and capture cross sections of electron and hole states in QDs. In addition, the important storage time at room temperature can be quantified and connected to the localization energy of the charge carriers.

2.7.1 Carrier Storage in InGaAs/GaAs Quantum Dots

Emission of electrons and holes from InGaAs/GaAs QDs was observed using conventional DLTS measurements by Kapteyn et al. (1999, 2000a,b). The electron/hole emission from InGaAs/GaAs QDs was studied in more detail by using charge-selective DLTS (Geller et al. 2006d) and time-resolved tunneling capacitance measurements (TRTCM) (Geller et al. 2006c). Two samples H1 and E1 were investigated to study the hole and electron emission, respectively, where the QD layer was incorporated in the slightly p-doped/n-doped (~3×10^{16} cm^{-3}) region of a n$^+$–p or p$^+$–n diode structure. The QD layer was situated 500 nm/415 nm below the p–n junction for the hole/electron sample. Mesa structures with a diameter of 800 µm and ohmic contacts were formed by employing standard optical lithography. The results of these measurements are presented in Figure 2.21.

It turns out that phonon-assisted tunneling controls the charge-carrier emission process from self-organized QDs in an electric field. The influence of the tunneling part, however, depends strongly on the effective mass and the strength of the electric field. The observed hole ground state activation energy E_A^{H1} = 120 ± 10 meV, hence, underestimates the localization energy. It is the thermal activation part in a phonon-assisted tunneling process: thermal activation into an excited state and subsequent tunneling through the remaining triangular barrier, cf. schematic pictures in Figure 2.21. The true hole localization energy was determined by using the TRTCM method to E_{loc}^{H1} = 210 ± 20 meV.

For the electrons, two contributions to the DLTS signal were observed: a DLTS signal with an activation energy of E_{A1}^{E1} = 82 ± 10 meV that is in good agreement with the theoretically predicted value for the ground/excited state energy splitting (70 meV). In addition, a DLTS signal with a smaller activation energy of E_{A2}^{E1} = 44 ± 10 meV is observed. This value is attributed to the first/second-excited state energy splitting and in satisfactory agreement with the theoretically predicted value of 50 meV. The ground state localization energy was determined by the TRTCM method to E_{loc}^{E1} = 290 ± 30 meV (Geller et al. 2006c).

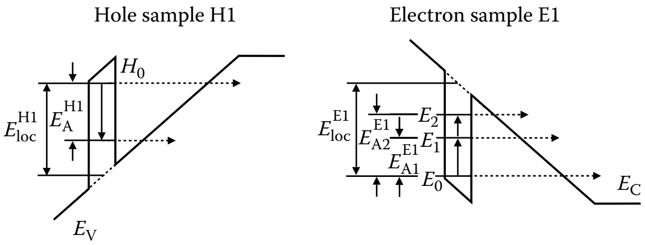

Thermal activation energy	$E_A^{H1} = 120$ meV	$E_{A1}^{E1} = 82$ meV
		$E_{A2}^{E1} = 44$ meV
Localization energy	$E_{loc}^{H1} = 210$ meV	$E_{loc}^{E1} = 290$ meV
Storage time (300 K)	$t_{Storage}^{H1} = 0.5$ ns	$t_{Storage}^{E1} = 200$ ns

FIGURE 2.21 Summary of the results from the charge-selective DLTS and tunneling emission experiments on self-organized InGaAs/GaAs QDs.

The electron/hole storage time at room temperature can be estimated by using Equation 2.27 and the capture cross section and the localization energy for sample E1/H1, respectively. An average storage time for electrons of about 200 ns and for holes of about 0.5 ns is obtained. This means, InAs QDs embedded in a GaAs matrix do not have a sufficiently long storage time to act as storage units in future nanomemories.

2.7.2 Hole Storage in GaSb/GaAs Quantum Dots

The storage time can be further increased by changing the material of the QDs and/or the surrounding matrix. A larger difference in the energy bandgap than in InGaAs/GaAs is more promising, for example, InAs/AlAs QDs. Moreover, large band discontinuities and, hence, strong hole localization is expected in type II QD heterostructures (Hatami et al. 1998), for example, GaSb/GaAs or InSb/GaAs. Only holes are confined in GaSb/GaAs QDs, while a repulsive potential barrier exists for the electrons in the conduction band. Type II material combinations are therefore very attractive for future memory applications.

Hole storage in and emission from $GaSb_{0.6}As_{0.4}$/GaAs QDs was investigated by DLTS and charge-selective DLTS (Geller et al. 2003). The sample was an n$^+$–p diode structure containing a single layer of GaSb QDs. An area density of about 3×10^{10} cm^{-2}, an average QD height of about 3.5 nm, and an average base width of about 26 nm were determined by structural characterization of uncapped samples grown under identical conditions (Müller-Kirsch et al. 2001). The QD layer was placed 500 nm below the n$^+$–p junction in a slightly p-doped ($p = 3 \times 10^{16}$ cm^{-3}) GaAs region. Mesa structures with a diameter of 800 µm and ohmic contacts were formed by employing standard optical lithography.

2.7.2.1 Multiple Hole Emission

In Figure 2.22, conventional DLTS measurements of the QD sample for a reverse bias V_r = 10 V and a pulse bias V_p = 4 V are displayed. The pulse length was 10 ms. For these conditions, the QDs are completely filled during the bias pulse and release all trapped holes after the pulse.

The DLTS spectrum of the QD sample (Figure 2.22a) shows three maxima, in the following denoted by "A," "B," and "C." From an Arrhenius plot, an activation energy E_A^C = 530 ± 20 meV and a capture cross section $\sigma_\infty^C \approx 6 \times 10^{-16}$ cm^2 is obtained for peak C. This signature can be identified as the bulk hole trap H3 in p-type GaAs (Stievenard et al. 1986).

Peaks A and B, observed at temperatures T = 140 and 230 K, are obviously related to QD formation. An activation energy of about 600 meV is found for peak B. This relatively high activation energy is comparable to the unstrained GaSb/GaAs valence band offset (Hatami et al. 1995, North et al. 1998) and could be explained by the existence of relaxed GaSb islands in the QD sample.

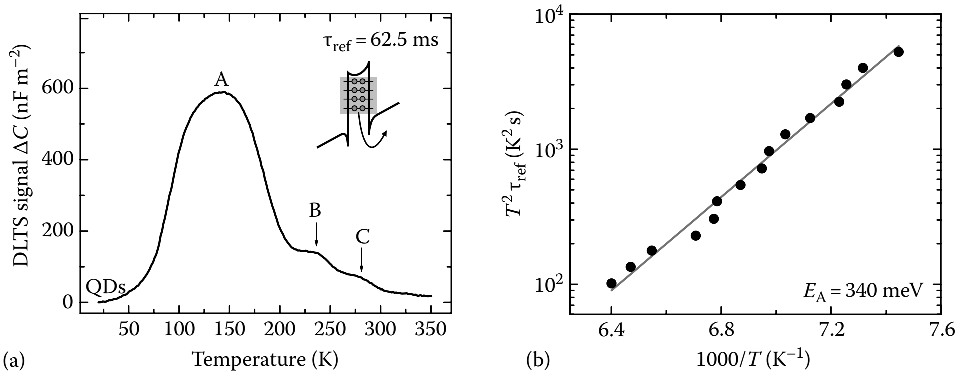

FIGURE 2.22 (a) DLTS signal of the $GaSb_{0.6}As_{0.4}$/GaAs QD sample, measured at a reverse bias V_r = 10 V and a pulse bias V_p = 4 V with a filling pulse width of 10 ms and a reference time constant τ_{ref} = 62.5 ms. (b) The corresponding Arrhenius plot for peak A, which is related to multiple hole emission from the GaSb QD states. An average activation energy of E_A = 340 meV is obtained.

Peak A in the DLTS spectrum of the QD sample in Figure 2.22a is attributed to the hole emission from the GaSb QD states, where the activation energy of each successively emitted hole depends on the actual charge state. The DLTS peak is broadened as previously observed for multiply charged Ge/Si QDs (Kapteyn et al. 2000b). From a standard Arrhenius plot in Figure 2.22b, an activation energy of E_A = 340 meV is obtained. Consequently, this activation energy of peak A represents only an average value for hole emission from completely charged QDs, where 15 holes from different states with different confinement are involved.

2.7.2.2 Charge-Selective DLTS

In order to study the charge states in more detail, the QD sample was studied using the charge-selective DLTS method. The pulse bias was always set to $V_p = V_r - 0.5$ V, while the reverse bias was increased from 4.5 to 9.5 V. Narrow peaks appear in the DLTS spectra in Figure 2.23. At V_r = 4.5 V the DLTS signal shows a maximum at about 80 K and for increasing reverse bias the DLTS peak shifts to higher temperature. The activation energy increases accordingly from 150 meV at 4.5 V to 450 meV for 9.5 V in Figure 2.24 obtained by standard Arrhenius plots.

The DLTS spectra for a reverse bias between 4.5 and 9.5 V in Figure 2.23 are attributed to the hole emission from differently charged QDs. All these spectra exhibit a maximum in the temperature range between 80 and 180 K, the range covered by peak A in Figure 2.22a. The maximum activation energy of 450 meV represents the average hole ground state energy of the QD ensemble, that is, the localization energy. The decrease in the activation energy from 450 meV down to 150 meV corresponds to an increase in the average occupation of the QDs, see the schematic insets in Figure 2.23. With increasing amount of charge in the QDs, state filling lowers the thermal activation barrier. The completely charged QDs are filled with 15 holes up to the Fermi level at the valence band edge, where Coulomb charging generates the barrier height.

In order to compare the storage time for $GaSb_{0.6}As_{0.4}$/GaAs with In(Ga)As/GaAs QDs, the observed emission rates were extrapolated to room temperature, using Equation 2.27. A storage time of about 1 μs for localized holes with the ground state energy of 450 meV is estimated, three orders of magnitude longer than the hole storage time in InGaAs/GaAs QDs.

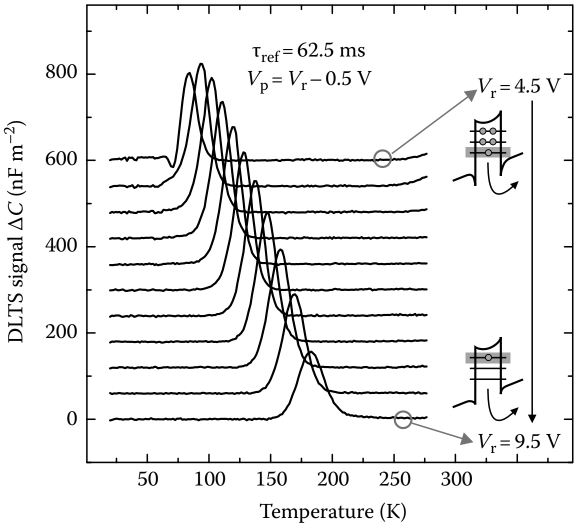

FIGURE 2.23 Charge-selective DLTS spectra of hole emission from $GaSb_{0.6}As_{0.4}$/GaAs QDs for a reference time constant of τ_{ref} = 62.5 ms. The reverse bias V_r is increased from 4.5 V up to 9.5 V while the pulse height is fixed at 0.5 V for all spectra. The pulse width is 10 ms and the data is displayed vertically shifted for clarity.

2.7.3 InGaAs/GaAs Quantum Dots with Additional AlGaAs Barrier

Hole storage in InGaAs/GaAs QDs with an additional AlGaAs barrier is presented in this section. The additional AlGaAs barrier increases the activation energies and a longer storage time at room temperature is observed. Two different samples having an AlGaAs barrier with different aluminum content were studied. The first contains an $Al_{0.6}Ga_{0.4}As$, the second an $Al_{0.9}Ga_{0.1}As$ barrier below the QD layer. The activation energy in the latter is increased sufficiently to reach a retention time of seconds at room temperature.

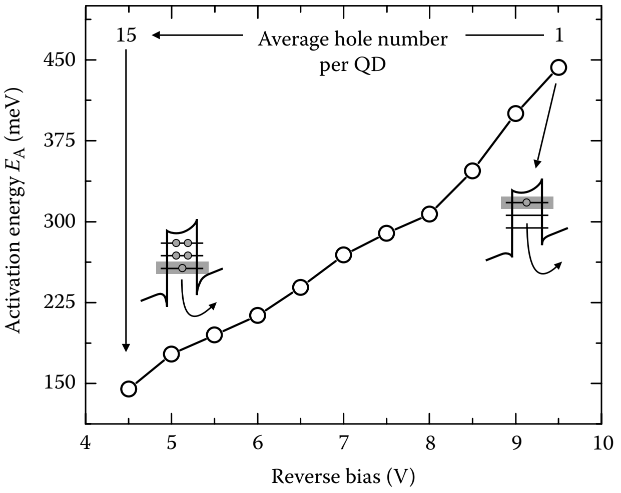

FIGURE 2.24 Dependence of the activation energy E_A on the reverse bias $GaSb_{0.6}As_{0.4}$/GaAs QDs. For a reverse bias of V_r = 9.5 V hole emission from the QD ground states is probed.

2.7.3.1 Storage Time: Milliseconds at Room Temperature

The first sample is an n^+–p diode structure, grown by MBE. It contains a single layer of InGaAs QDs embedded in slightly p-doped GaAs ($p = 2 \times 10^{15}$ cm^{-3}). The QDs are placed 1500 nm below the p–n junction and an additional undoped $Al_{0.6}Ga_{0.4}As$ barrier of 20 nm thickness is situated 7 nm below the QD layer to increase the hole storage time. Again, mesa structures with a diameter of 800 μm and ohmic contacts were formed.

Figure 2.25a shows the charge-selective DLTS spectra with a reference time constant of 5 ms (Marent et al. 2006). The pulse bias height was fixed to 0.2 V for all spectra ($V_p = V_r - 0.2$ V). For a reverse bias above V_r = 3.2 V no DLTS signal is visible, as the Fermi level is energetically above the QD states and no QD states are occupied. By decreasing the reverse bias the Fermi level reaches the QD ground state at V_r = 3.2 V and a peak in the DLTS spectrum appears at 300 K. This peak is related to thermally activated hole emission from the ground states of the QD ensemble across the AlGaAs barrier, see inset in the upper left corner of Figure 2.25b. A further decrease of the reverse bias leads to QD state-filling and emission from higher QD states is observed. A peak in the DLTS appears at that temperature, where the averaged time constant of the thermally activated emission equals the applied reference time constant, cf. Equation 2.26. Therefore, the peak at 300 K for τ_{ref} = 5 ms (V_r = 3.2 V in Figure 2.25a) represents an average emission time constant (storage time) of 5 ms for hole emission from the QD ground states across the $Al_{0.6}Ga_{0.4}As$ barrier.

The thermal activation energies are shown in Figure 2.25b. The highest value of 560 ± 60 meV at V_r = 3.2 V is related to thermal activation from the hole ground states of the QD ensemble across the AlGaAs barrier. The decrease in the activation energy corresponds to an increase in the average occupation of the QDs. At V_r = 1.4 V the energetic position of the Fermi level is at the valence band edge and no further QD state-filling is possible. As a consequence, between V_r = 1.4 and 0.2 V carrier emission from the valence band edge is probed. The activation energy remains roughly constant with a mean value of 340 meV. This energy represents the energetic height of the $Al_{0.6}Ga_{0.4}As$ barrier.

2.7.3.2 Storage Time: Seconds at Room Temperature

The second sample is also an n^+–p diode structure, grown by MOCVD. It contains a single layer of InGaAs QDs embedded in p-doped GaAs ($p = 3 \times 10^{16}$ cm^{-3}). The QDs are placed 400 nm below the p–n junction and an additional undoped $Al_{0.9}Ga_{0.1}As$ barrier of 20 nm thickness is situated 7 nm below the QD layer.

Figure 2.26a shows the charge-selective DLTS spectra (Marent et al. 2007). The pulse bias height was fixed here to 0.3 V for all spectra ($V_p = V_r - 0.3$ V). At V_r = 5.7 V a peak in the DLTS spectrum appears at 380 K, related to thermally activated hole emission from the ground states of the QD ensemble across the $Al_{0.9}Ga_{0.1}As$ barrier. As mentioned before, the peak at 380 K for τ_{ref} = 5 ms represents an average emission time constant (storage time) of 5 ms for hole emission from the ground states of the QD ensemble across the $Al_{0.9}Ga_{0.1}As$ barrier. To obtain the storage time at room temperature, the capacitance transient recorded at

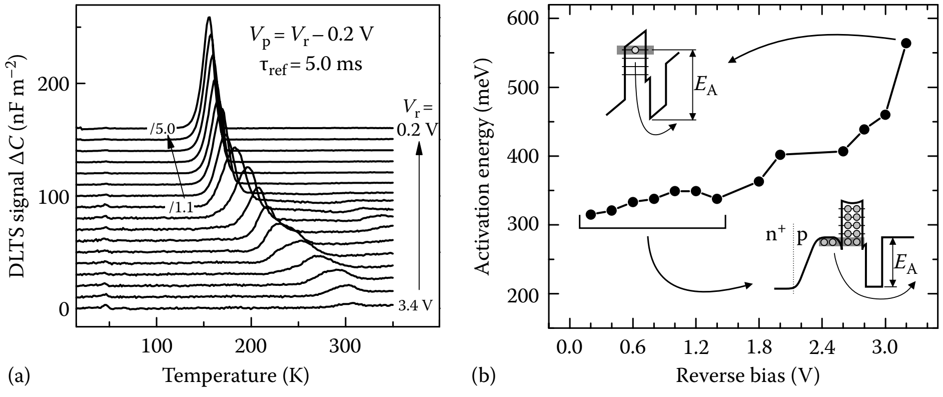

FIGURE 2.25 Charge-selective DLTS spectra of thermally activated hole emission from InGaAs/GaAs QDs with an additional $Al_{0.6}Ga_{0.4}As$ barrier below the QD layer (a). Spectra below V_r = 1.6 V are divided by a factor from 1.1 up to 5. (b) Dependence of the thermal activation energy on the reverse bias V_r. (Reprinted from Marent, A. et al., *Appl. Phys. Lett.*, 89, 072103, 2006. With permission.)

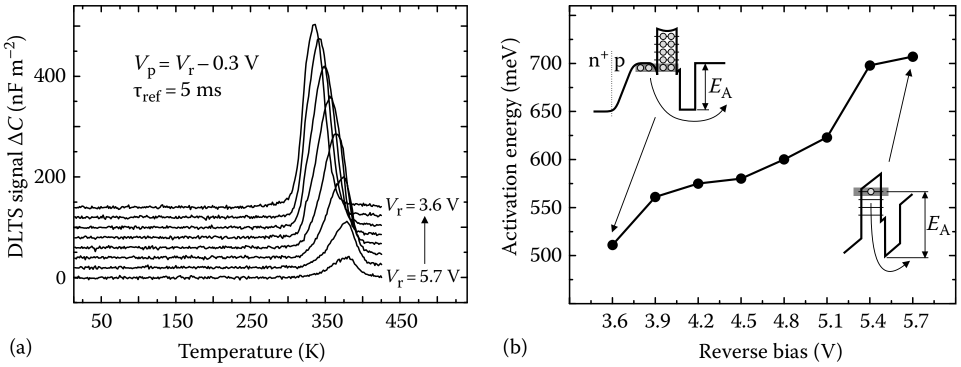

FIGURE 2.26 Charge-selective DLTS spectra of thermally activated hole emission from InGaAs/GaAs QDs with an additional $Al_{0.9}Ga_{0.1}As$ barrier below the QD layer (a). Dependence of the thermal activation energy on the reverse bias V_r.

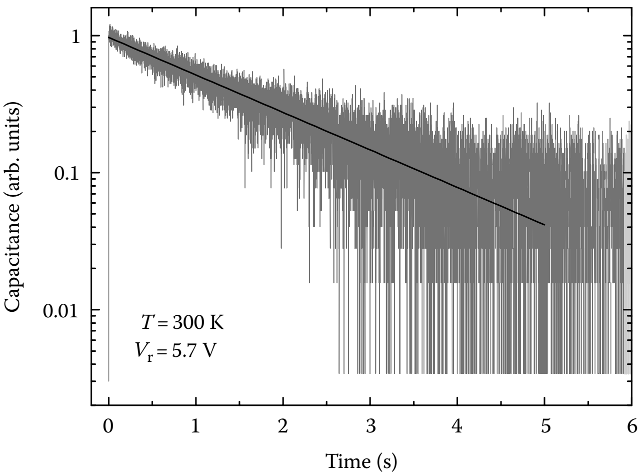

FIGURE 2.27 Capacitance transient of hole emission from InGaAs/GaAs QDs with an additional $Al_{0.9}Ga_{0.1}As$ barrier (for a reverse bias of $V_r = 5.7$ V). An average hole storage time of 1.6 s at $T = 300$ K is obtained from a linear fit of the transient on a semilogarithmic scale.

300 K is plotted on a semilogarithmic scale in Figure 2.27. From a linear fit an average hole storage time of 1.6 s at room temperature is determined.

Furthermore, the thermal activation energies are shown in Figure 2.26b. The highest value of 710 ± 40 meV at $V_r = 5.7$ V is related to thermal activation from the hole ground states of the QD ensemble across the $Al_{0.9}Ga_{0.1}As$ barrier. The decrease in the activation energy corresponds again to an increase in the average occupation of the QDs. At $V_r = 3.6$ V, carrier emission from the valence band edge is probed and the activation energy has a value of 520 meV. This energy now represents the energetic height of the $Al_{0.9}Ga_{0.1}As$ barrier.

2.7.4 Storage Time in Quantum Dots

Carrier emission from different QD systems has been studied in order to determine the carrier storage time at room temperature. In addition, the localization energy was obtained by using time-resolved capacitance spectroscopy (DLTS) and related to the hole/electron retention time. The results of the experiments are summarized in Figure 2.28.

An electron/hole localization energy of 290/210 meV, respectively, was obtained for InGaAs QDs embedded in a GaAs matrix. Based on these values the storage time at $T = 300$ K was estimated to ~200 ns for electrons and ~0.5 ns for holes. Furthermore, the more promising type II $GaSb_{0.6}As_{0.4}$/GaAs has been studied in detail. Ground state activation energy of 450 meV was determined, which accounts for a room temperature emission time in the order of 1 μs. The ground state localization is about twice as large and the retention time at room temperature is about three orders of magnitude longer than in InGaAs/GaAs QDs. Finally, the hole emission from InGaAs/GaAs QDs with an additional AlGaAs barrier was investigated. The activation energy for the QD ground states increases from 210 to 560 meV

Material System	Charge Carrier Type	Localization Energy	Storage Time at 300 K
InAs/GaAs	Hole Electron	210 meV 290 meV	~0.5 ns ~200 ns
Ge/Si	Hole	350 meV	~0.1 μs
$GaSb_{0.6}As_{0.4}$/GaAs	Hole	450 meV	~1 μs
InAs/GaAs with $Al_{0.6}Ga_{0.4}$ As barrier	Hole	560 meV	5 ms
InAs/GaAs with $Al_{0.9}Ga_{0.1}As$ barrier	Hole	710 meV	1.6 s

FIGURE 2.28 Summary of the measured electron/hole storage time (at 300 K) and localization energy for different QD systems.

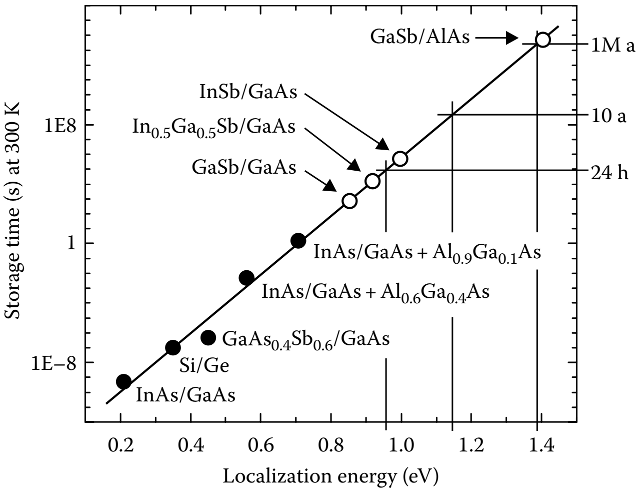

FIGURE 2.29 Dependence of the hole storage time on the localization energy for a variety of QD systems. The solid line is a fit to the experimentally obtained data (full circles). The open circles are estimated storage times for the labeled material systems according to the calculated localization energies and the fit. (Reprinted from Marent, A. et al., *Appl. Phys. Lett.*, 91, 42109, 2007. With permission.)

for an additional $Al_{0.6}Ga_{0.4}As$ barrier and the hole storage time is in the order of milliseconds. Using an $Al_{0.9}Ga_{0.1}As$ barrier increases the activation energy accordingly from 210 to 710 meV and the hole storage time at room temperature increases by nine orders of magnitude to 1.6 s. This value is already three orders of magnitude longer than today's DRAM refresh time, which is in the millisecond range.

From the experimental results, material combinations can be predicted to obtain a storage time of more than 10 years at room temperature. Figure 2.29 displays the hole storage time in dependence of the localization energy (full circles). The solid line is a fit to the data and corresponds to Equation 2.27. The storage time shows an exponential dependence on the localization energy as predicted by the common rate equation of thermally activated emission. The storage time increases by one order of magnitude for an increase of the localization energy of about 50 meV. From these results, hole localization energies can be estimated, which provide storage times of 24 h or 10 years. These storage times are reached for localization energies of 0.96 and 1.14 eV, respectively.

To find a material system with such large localization energies, the hole localization energies for $GaAs_xSb_{1-x}/GaAs$ and $In_xGa_{1-x}Sb/GaAs$ QDs have been calculated using eight-band k · p theory (Stier et al. 1999, Schliwa et al. 2007). Based on structural characterization of GaSb/GaAs QDs (Müller-Kirsch et al. 2001) the QDs are modeled as truncated pyramids with a base width of 21 nm and a height of 3.9 nm. The results are summarized in Figure 2.30 and plotted in Figure 2.29 as open circles. The localization energy in $GaAs_xSb_{1-x}/GaAs$ QDs increases from 350 meV up to 853 meV for an antimony content of 50% and 100%, respectively. Since the bandgap in III–V materials decreases with increasing lattice constant (Vurgaftman et al. 2001), InSb/GaAs QDs should offer a larger localization energy than GaSb/GaAs. Eight-band k·p calculations for $In_xGa_{1-x}Sb/GaAs$ QDs provide a localization energy of 919 and 996 meV for an indium content of 50% and 100%, respectively. A hole storage time in InSb/GaAs QDs of more than 24 h is predicted (see Figure 2.29). Using an AlAs matrix instead of a GaAs matrix a storage time of more than 10 years can be reached. Since the valence band offset between GaAs and AlAs is about 550 meV (Batey and Wright 1986) the entire hole localization energy in GaSb/AlAs QDs is about 1.4 eV. This value leads to the prediction of an average hole storage time of more than 1 million years at room temperature (see Figure 2.29), orders of magnitude longer than needed for a nonvolatile memory.

2.8 Write Times in Quantum Dot Memories

The carrier capture of electrons/holes from the valence/conduction band into the QD states limits physically the possible write time in a QD-based nanomemory. Normally, carrier capture into QDs is studied by interband absorption optical techniques, like time-resolved photoluminescence (PL) spectroscopy (Heitz et al. 1997, Giorgi et al. 2001). However, such experiments probe the exciton dynamics, that means, the electron–hole capture and relaxation into the QD states is measured simultaneously. A detailed knowledge of either electron or hole capture is of great importance, for future QD-based memory applications as only one sort of charge carriers is stored. Electron or hole capture has been investigated separately, for example, in interband pump and intraband probe experiments of Müller et al. (2003). In addition, the hole capture into $GaSb_{0.6}As_{0.4}/GaAs$ self-organized QDs has been studied (Geller et al. 2008) using DLTS experiments. The investigations demonstrate a fast

Material System	Charge Carrier Type	Localization Energy	Storage Time at 300 K
$GaSb_{0.5}As_{0.5}/GaAs$	Hole	350 meV	0.1 μs
GaSb/GaAs	Hole	853 meV	13 min
$In_{0.5}Ga_{0.5}Sb/GaAs$	Hole	919 meV	4 h
InSb/GaAs	Hole	996 meV	6 days
GaSb/AlAs	Hole	~1.4 eV	>10^6 years

FIGURE 2.30 Hole localization energies in different Sb-based QDs calculated by eight-band k.p theory. The storage time is estimated according to the fit of the experimental data in Figure 2.29.

capture and relaxation process in self-organized QDs in the range of picoseconds at room temperature, more than three orders of magnitude faster than the write time in a DRAM cell. This fast carrier capture should enable very fast write times in a QD-based nanomemory (<ns) which are independent of the storage time, possibly up to more than 10 years at room temperature. The QD-based memory concept should therefore allow a nonvolatile semiconductor memory with orders of magnitude faster write times than the present Flash concept.

To confirm the presumption of a fast write time that is independent of the storage time, this section shows the write time measurements for two different QD memory structures containing type I InAs/GaAs and type II $GaSb_{0.6}As_{0.4}$/GaAs QDs—both already described in Section 2.7—where the average storage time at 300 K ranges from nanoseconds (InAs/GaAs QDs) to microseconds ($GaSb_{0.6}As_{0.4}$/GaAs QDs).

2.8.1 Hysteresis Measurements

As already discussed in Section 2.7, where the charge carrier storage time was determined, the read-out of the charge state (which represents the stored information) inside the QDs is done by the measurement of the capacitance of the p–n diode. The capacitance of a diode structure with embedded QDs depends on the number of holes stored inside the depletion region. A larger capacitance corresponds to unoccupied QDs ("0") while a smaller capacitance represents a "1," where the QDs are filled with holes.

A fundamental property of storage devices is the appearance of hysteresis when switching between the two information states. Accordingly, a nanomemory device containing self-organized QDs shows a hysteresis in the capacitance measurement after writing (yielding occupied QDs, a "1" state) and erasing (yielding unoccupied QDs, a "0" state) process. Such a hysteresis measurement is shown in Figure 2.31 for InAs/GaAs QDs (a) and $GaSb_{0.6}As_{0.4}$/GaAs QDs (b). Both memory structures are n-p diodes where holes are stored inside the QDs. The sample description can be found in Section 2.7.1 for the InAs/GaAs QDs and in Section 2.7.2 for the $GaSb_{0.6}As_{0.4}$/GaAs QDs. One hysteresis sweep takes a few seconds; hence, the temperature was reduced down to 15 K for the InAs QDs where the hole storage time is in the order of minutes. Analogously, for the $GaSb_{0.6}As_{0.4}$/GaAs QDs—having hole localization energy twice as high as the InAs/GaAs QDs—a higher temperature of 100 K is already sufficient to obtain hole storage times of several minutes.

Figure 2.31 shows the switching between the two information states by a hysteresis curve of the capacitance for both samples. At the reverse bias of 14 V/16 V (point 1) the charge carriers tunnel out of the QDs (erasing the information). If the reverse bias is now swept from 14 V/16 V to the storage situation at 8.2 V (InAs QDs) or 7.2 V (GaSb QDs), respectively, the QDs are empty and a larger capacitance is observed (point 2). If the bias is swept to 0 V (point 3), the QDs are charged with holes and a smaller capacitance is observed upon sweeping back to the storage situation (point 4). The maximum hysteresis opening is now defined at the storage position as the difference between both capacitance values for a "0" and "1" state. The structures are not yet optimized since the carriers are not stored at zero bias. However, the memory structure can be easily adjusted in the future to yield a storage situation at zero bias, realized by changing the width between the QD layer and the p–n junction.

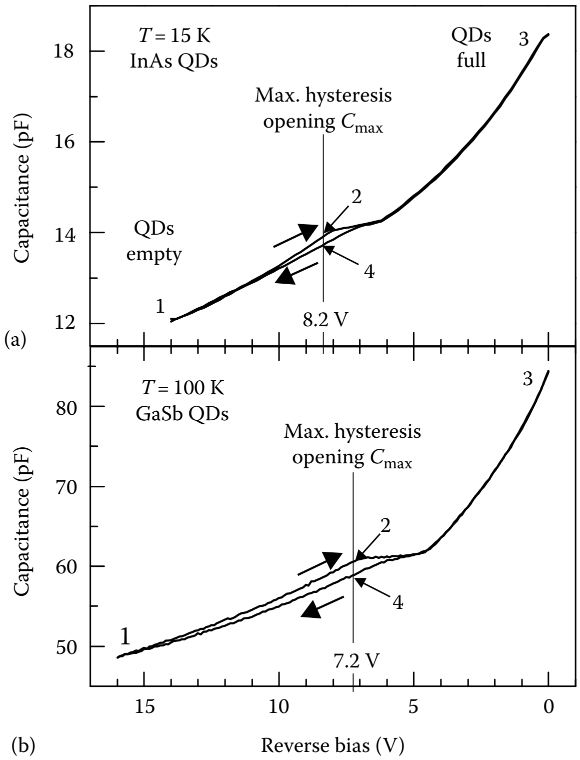

FIGURE 2.31 Capacitance sweeps of a p–n-diode containing InAs/GaAs (a) and $GaSb_{0.6}As_{0.4}$/GaAs (b) QDs. (Reprinted from Geller, M. et al., *Appl. Phys. Lett.*, 92, 092108, 2008. With permission.)

2.8.2 Write Time Measurements

To study the limit of the write-time, series of write pulses are applied with decreasing pulse widths down to 300 ps. The operation principle is schematically depicted in Figure 2.32. The cycle

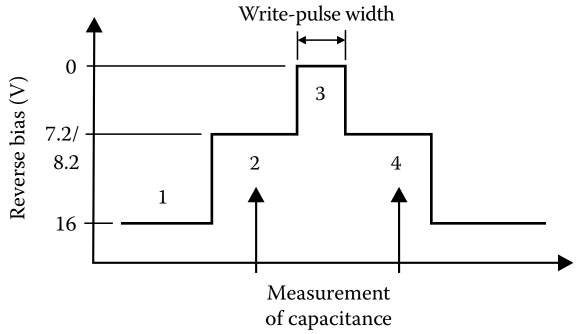

FIGURE 2.32 Operation principle for measuring the write time, while the labeled numbers refer to the numbers in Figure 2.31. (Reprinted from Geller, M. et al., *Appl. Phys. Lett.*, 92, 092108, 2008. With permission.)

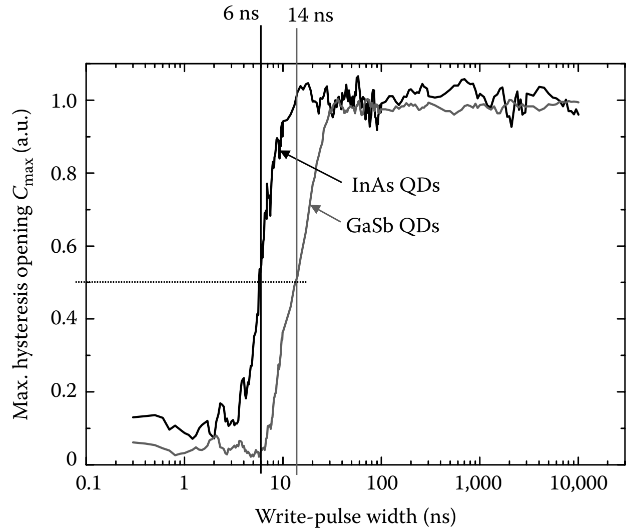

FIGURE 2.33 Maximum hysteresis opening $C_{\max}$ for varying write-pulse width. The InAs QD memory is successfully charged for write-pulses down to 6 ns. Analogously, the information in the GaSb QD memory structure can be written down to 14 ns.

started with a 10 s long erase pulse at $V_r = 14\,\text{V}/16\,\text{V}$ (point 1). Then the capacitance of the device structure is measured at the storage position (point 2) at 7.2 V (GaSb QDs) and 8.2 V (InAs QDs), respectively. A write pulse of $V_r = 0\,\text{V}$ was applied to the device (point 3) and the capacitance is again measured at the storage situation point 4.

The maximum hysteresis opening $C_{\max}$ was plotted versus the write-pulse width in Figure 2.33. The limit of the write time is reached when the QDs are not sufficiently charged anymore. The hysteresis opening $C_{\max}$ vanishes and the write time is defined as a drop in $C_{\max}$ to 50%. A minimum write time of 6 ns at $T = 15\,\text{K}$ for the InAs QD sample and 14 ns at $T = 100\,\text{K}$ for the GaSb QD sample is obtained, respectively. These values are in the order of the write time in a DRAM cell and are now only limited by the experimental set-up and the parasitic cut-off frequency of the RC low pass of the devices. Additional measurements for different temperatures on the same device showed a temperature-independent write time, for example, the difference between the InAs and GaSb structure is due to these different cut-off frequencies. This assumption is also confirmed by previous measurements (Geller et al. 2006b) of the same structure, where average hole capture and relaxation times were found in the order of picoseconds. Therefore, much faster write times below 1 ns are feasible for improved nanomemory structures having higher external cut-off frequencies.

2.9 Summary and Outlook

The semiconductor memory market is evenly split following two different layout approaches: the DRAM and the Flash. Both memories have their limitations in storage time, write/read/erase time, and endurance and will be facing fundamental physical problems in the near future when the feature size has to go down to 14 nm by the year 2020. Therefore, a large variety of novel concepts using different physical phenomena to store an information bit has been reported. One of the promising options is the use of self-organized nanostructures in future nanoelectronic devices. Especially self-organized QDs based on III–V materials (e.g., GaAs, InAs, InP, etc.) could realize a nonvolatile memory (like Flash) with fast read/write/erase access time (like DRAM). In addition, with an area density of up to 10^{11} cm^{-2}, a storage density in the order of 1 TBit in.$^{-2}$ could be possible, if each single QD—with a lateral feature size of about 10 nm—would represent one information bit and it could be addressed individually.

This chapter showed the potential and advantages of using self-organized QDs in a III–V semiconductor nanomemory concept. The charge carrier storage and emission from different QD systems has been studied in order to determine the carrier storage time at room temperature. Localization energies were determined by using time-resolved capacitance spectroscopy (DLTS) and were related to the hole/electron retention time. A maximum storage time of 1.6 s at room temperature has already been demonstrated in InAs/GaAs QDs with $Al_{0.9}Ga_{0.1}As$ barriers. A storage time of more than 10 years is predicted for hole storage in GaSb/AlAs QDs.

The use of III–V heterostructures enables tuning the band structure to a desired storage time at room temperature while simultaneously getting rid of one of the largest drawback of Si-based Flash memories: the fixed SiO_2 barrier that leads to the poor endurance and a bad write time. A very fast write time below 1 ns independent of the storage time (possibly beyond 10 years) is possible. The physical limitation of the write time in such a QD-based nanomemory is only given by the relaxation time of the charge carriers, known to be in the picosecond range. A write time of 14 ns in $GaSb_{0.6}As_{0.4}$/GaAs QDs and 6 ns for InAs/GaAs QDs has already been achieved. These numbers are now limited by the experimental set-up and the parasitic cut-off frequency of the RC low pass of the devices. Faster write times below 1 ns in combination with a storage time of more than 10 years at room temperature are expected in the future. A fast semiconductor Flash nanomemory—even faster than the present day DRAM—based on III–V self-organized QDs seems to be possible.

Acknowledgments

This work was partly funded by the SANDiE Network of Excellence of the European Commission, Contract No. NMP4-CT-2004-500101; the Nanomat project of the European Commission Growth Program, Contract No. G5RD-CT-00545; and SFB 296 of DFG. We also acknowledge financial support by the QD-Flash project in the framework of ProFIT of the Investitionsbank Berlin (IBB) and the DFG, BI 284/29-1. We want to acknowledge the support of and helpful discussions with Dr. C. Kapteyn, T. Nowozin, and E. Stock. Dr. L. Müller-Kirsch, K. Pötschke, D. Feise, A. P. Vasi'ev, E. S. Semenova, Prof. A. E. Zhukov, and Prof. V. M. Ustinov are acknowledged for providing us with the samples.

References

Alferov, Z. I. and Kazarinov, R. F. 1963. Semiconductor laser with electric pumping. Inventor's certificate no. 181737 [in Russian]. Application no. 950840.

Anand, S., Carlsson, N., Pistol, M.-E. et al. 1995. Deep level transient spectroscopy of InP quantum dots. *Appl. Phys. Lett.* **67**: 3016.

Anand, S., Carlsson, N., Pistol, M.-E. et al. 1998. Electrical characterization of InP/GaInP quantum dots by space charge spectroscopy. *J. Appl. Phys.* **84**: 3747.

Ashoori, R. C. 1996. Electrons in artificial atoms. *Nature* **379**: 413.

Ashoori, R. C., Störmer, H. L., Weiner, J. S. et al. 1992. Single-electron capacitance spectroscopy of discrete quantum levels. *Phys. Rev. Lett.* **68**: 3088.

Atwood, G. 2004. Future directions and challenges for Etox flash memory scaling. *IEEE Trans. Device Mater. Reliab.* **4**: 301.

Baibich, M. N., Broto, J. M., Fert, A. et al. 1988. Giant magnetoresistance of (001)Fe/(001)Cr magnetic superlattices. *Phys. Rev. Lett.* **61**: 2472.

Batey, J. and Wright, S. L. 1986. Energy band alignment in GaAs:(Al,Ga)As heterostructures: The dependence on alloy composition. *J. Appl. Phys.* **59**: 200.

Bimberg, D., Grundmann, M., and Ledentsov, N. N. 1998. *Quantum Dot Heterostructures*. Chichester, U.K.: John Wiley & Sons.

Binasch, G., Grunberg, P., Saurenbach, F. et al. 1989. Enhanced magnetoresistance in layered magnetic-structures with antiferromagnetic interlayer exchange. *Phys. Rev. B* **39**: 4828.

Blood, P. and Orton, J. W. 1992. *The Electrical Characterization of Semiconductors: Majority Carriers and Electron States*. London, U.K.: Academic Press.

Burr, G. W., Kurdi, B. N., Scott, J. C. et al. 2008. Overview of candidate device technologies for storage-class memory. *IBM J. Res. Dev.* **52**: 449.

Campbell, J. C., Huffaker, D. L., Deng, H. et al. 1997. Quantum dot resonant cavity photodiode with operation near 1.3 μm wavelength. *Electron. Lett.* **33**: 1337.

Chu, L., Zrenner, A., Bichler, M. et al. 2001. Quantum-dot infrared photodetector with lateral carrier transport. *Appl. Phys. Lett.* **79**: 2249.

Day, D. S., Tsai, M. Y., Streetman, B. G. et al. 1979. Deep-level-transient spectroscopy: System effects and data analysis. *J. Appl. Phys.* **50**: 5093.

Dennard, R. H. 1968. Field-effect transistor memory. U.S. patent no. 3,387,286.

Eaglesham, D. J. and Cerullo, M. 1990. Dislocation-free Stranski–Krastanow growth of Ge on Si (100). *Phys. Rev. Lett.* **64**: 1943.

Eisele, H., Flebbe, O., Kalka, T. et al. 1999. Cross-sectional scanning-tunneling microscopy of stacked InAs quantum dots. *Appl. Phys. Lett.* **75**: 106.

Frank, D. J., Dennard, R. H., Nowak, E. et al. 2001. Device scaling limits of Si MOSFETs and their application dependencies. *Proc. IEEE* **89**: 259.

Fricke, M., Lorke, A., Kotthaus, J. P. et al. 1996. Shell structure and electron–electron interaction in self-assembled InAs quantum dots. *Europhys. Lett.* **36**: 197.

Gallagher, W. J. and Parkin, S. S. P. 2006. Development of the magnetic tunnel junction MRAM at IBM: From first junctions to a 16-MB MRAM demonstrator chip. *IBM J. Res. Dev.* **50**: 333.

Geller, M., Kapteyn, C., Müller-Kirsch, L. et al. 2003. 450 meV hole localization energy in GaSb/GaAs quantum dots. *Appl. Phys. Lett.* **82**: 2706.

Geller, M., Marent, A., and Bimberg, D. 2006a. Speicherzelle and Verfahren zum Speichern von Daten. German patent application no. 10 2006 059 110.0.

Geller, M., Marent, A., Stock, E. et al. 2006b. Hole capture into self-organized InGaAs quantum dots. *Appl. Phys. Lett.* **89**: 232105.

Geller, M., Stock, E., Kapteyn, C. et al. 2006c. Tunneling emission from self-organized In(Ga)As/GaAs quantum dots observed via time-resolved capacitance measurements. *Phys. Rev. B* **73**: 205331.

Geller, M., Stock, E., Sellin, R. L. et al. 2006d. Direct observation of tunneling emission to determine localization energies in self-organized In(Ga)As quantum dots. *Physica E* **32**: 171.

Geller, M., Marent, A., Nowozin, T. et al. 2008. A write time of 6 ns for quantum dot-based memory structures. *Appl. Phys. Lett.* **92**: 092108.

Geppert, L. 2003. The new indelible memories—It's a three-way race in the multibillion-dollar memory sweepstakes. *IEEE Spectr.* **40**: 48.

Giorgi, M. D., Lingk, C., Plessen, G. V. et al. 2001. Capture and thermal re-emission of carriers in long-wavelength InGaAs/GaAs quantum dots. *Appl. Phys. Lett.* **79**: 3968.

Grimmeiss, H. G. and Ovrén, C. 1981. Fundamentals of junction measurements in the study of deep energy levels in semiconductors. *J. Phys. E: Sci. Instrum.* **14**: 1032.

Grundmann, M., Stier, O., and Bimberg, D. 1995. InAs/GaAs quantum pyramids: Strain distribution, optical phonons and electronic structure. *Phys. Rev. B* **52**: 11969.

Guo, L., Leobandung, E., and Chou, S. Y. 1997. A room-temperature silicon single-electron metal-oxide-semiconductor memory with nanoscale floating-gate and ultranarrow channel. *Appl. Phys. Lett.* **70**(7): 850.

Hatami, F., Grundmann, M., Ledentsov, N. N. et al. 1998. Carrier dynamics in type-II GaSb/GaAs quantum dots. *Phys. Rev. B* **57**: 4635.

Hatami, F., Ledentsov, N. N., Grundmann, M. et al. 1995. Radiative recombination in type-II GaSb/GaAs quantum dots. *Appl. Phys. Lett.* **67**: 656.

Heinrichsdorff, F., Krost, A., Grundmann, M. et al. 1996. Self-organization processes of InGaAs/GaAs quantum dots grown by metalorganic chemical vapor deposition. *Appl. Phys. Lett.* **68**: 3284.

Heinrichsdorff, F., Ribbat, C., Grundmann, M. et al. 2000. High-power quantum-dot lasers at 1100 nm. *Appl. Phys. Lett.* **76**: 556.

Heitz, R., Veit, M., Ledentsov, N. N. et al. 1997. Energy relaxation by multiphonon processes in InAs/GaAs quantum dots. *Phys. Rev. B* **56**: 10435.

Hopfer, F., Mutig, A., Kuntz, M. et al. 2006. Single-mode submonolayer quantum-dot vertical-cavity surface-emitting lasers with high modulation bandwidth. *Appl. Phys. Lett.* **89**: 141106.

ITRS. 2008. International Technology Roadmap for Semiconductors (ITRS) 2007 Edition (2008). Technical report.

Jones, R. E., Maniar, P. D., Moazzami, R. et al. 1995. Ferroelectric non-volatile memories for low-voltage, low-power applications. *Thin Solid Films* **270**: 584.

Kapteyn, C. 2001. *Carrier Emission and Electronic Properties of Self-Organized Semiconductor Quantum Dots*. Berlin, Germany: Mensch & Buch Verlag.

Kapteyn, C. M. A., Heinrichsdorff, F., Stier, O. et al. 1999. Electron escape from InAs quantum dots. *Phys. Rev. B* **60**: 14265.

Kapteyn, C. M. A., Lion, M., Heitz, R. et al. 2000a. Hole and electron emission from InAs quantum dots. *Appl. Phys. Lett.* **76**: 1573.

Kapteyn, C. M. A., Lion, M., Heitz, R. et al. 2000b. Many-particle effects in Ge quantum dots investigated by time-resolved capacitance spectroscopy. *Appl. Phys. Lett.* **77**: 4169.

Kim, K., Hwang, C. G., and Lee, J. G. 1998. DRAM technology perspective for gigabit era. *IEEE Trans. Electron Devices* **45**: 598.

Kim, H., Noda, T., Kawazu, T. et al. 2000. Control of current hysteresis effects in a GaAs/n-AlGaAs quantum trap field effect transistor with embedded InAs quantum dots. *Jpn. J. Appl. Phys.* **39**: 7100.

Kimerling, L. C. 1974. Influence of deep traps on the measurement of free-carrier distributions in semiconductors by junction capacitance techniques. *J. Appl. Phys.* **45**: 1839.

Kiravittaya, S., Rastelli, A., and Schmidt, O. G. 2009. Advanced quantum dot configurations. *Rep. Prog. Phys.* **72**: 046502.

Kirstaedter, N., Ledentsov, N. N., Grundmann, M. et al. 1994. Low threshold, large T_0 injection laser emission from (InGa)As quantum dots. *Electron. Lett.* **30**: 1416.

Koike, K., Saitoh, K., Li, S. et al. 2000. Room-temperature operation of a memory-effect AlGaAs/GaAs heterojunction field-effect transistor with self-assembled InAs nanodots. *Appl. Phys. Lett.* **76**: 1464.

Kroemer, H. 1963. Semiconductor laser with optical pumping. U.S. patent no. 3309553.

Kuntz, M., Fiol, G., Lämmlin, M. et al. 2004. 35 GHz mode-locking of 1.3 μm quantum dot lasers. *Appl. Phys. Lett.* **85**: 843.

Lai, S. K. 2008. Flash memories: Successes and challenges. *IBM J. Res. Dev.* **52**: 529.

Lambe, J. and Jaklevic, R. C. 1969. Charge-quantization studies using a tunnel capacitor. *Phys. Rev. Lett.* **22**: 1371.

Lämmlin, M., Fiol, G., Meuer, C. et al. 2006. Distortion-free optical amplification of 20–80 GHz modelocked laser pulses at 1.3 μm using quantum dots. *Electron. Lett.* **42**: 697.

Lang, D. V. 1974. Deep-level transient spectroscopy: A new method to characterize traps in semiconductors. *J. Appl. Phys.* **45**: 3023.

Lochmann, A., Stock, E., Schulz, O. et al. 2006. Electrically driven single quantum dot polarised single photon emitter. *Electron. Lett.* **42**: 774.

Luyken, R. J., Lorke, A., Govorov, A. O. et al. 1999. The dynamics of tunneling into self-assembled InAs dots. *Appl. Phys. Lett.* **74**: 2486.

Mandelman, J. A., Dennard, R. H., Bronner, G. B. et al. 2002. Challenges and future directions for the scaling of dynamic random-access memory (DRAM). *IBM J. Res. Dev.* **46**: 187.

Marent, A., Geller, M., Bimberg, D. et al. 2006. Carrier storage time of milliseconds at room temperature in self-organized quantum dots. *Appl. Phys. Lett.* **89**: 072103.

Marent, A., Geller, M., Schliwa, A. et al. 2007. 10^6 years extrapolated hole storage time in GaSb/AlAs quantum dots. *Appl. Phys. Lett.* **91**: 42109.

Mikolajick, T., Nagel, N., Riedel, S. et al. 2007. Scaling of non-volatile memories to nanoscale feature sizes. *Mater. Sci.-Pol.* **25**: 33.

Miller, G. L., Lang, D. V., and Kimerling, L. C. 1977. Capacitance transient spectroscopy. *Ann. Rev. Mater. Sci.* **7**: 377.

Mo, Y.-W., Savage, D. E., Swartzentruber, B. S. et al. 1990. Kinetic pathway in Stranski–Krastanov growth of Ge on Si(001). *Phys. Rev. Lett.* **65**: 1020.

Moore, G. E. 1965. Cramming more components onto integrated circuits. *Electronics* **38**: 82.

Müller, T., Schrey, F. F., Strasser, G. et al. 2003. Ultrafast intraband spectroscopy of electron capture and relaxation in InAs/GaAs quantum dots. *Appl. Phys. Lett.* **83**: 3572.

Müller-Kirsch, L., Heitz, R., Pohl, U. W. et al. 2001. Temporal evolution of GaSb/GaAs quantum dot formation. *Appl. Phys. Lett.* **79**: 1027.

Myburg, G., Auret, F. D., Meyer, W. E. et al. 1998. Summary of Schottky barrier height data on epitaxially grown n- and p-GaAs. *Thin Solid Films* **325**: 181.

North, S. M., Briddon, P. R., Cusack, M. A. et al. 1998. Electronic structure of GaSb/GaAs quantum dots. *Phys. Rev. B* **58**: 12601.

Omling, P., Samuelson, L., and Grimmeis, H. G. 1983. Deep level transient spectroscopy evaluation of nonexponential transients in semiconductor alloys. *J. Appl. Phys.* **54**: 5117.

Pavan, P., Bez, R., Olivo, P. et al. 1997. Flash memory cells—An overview. *Proc. IEEE* **85**: 1248.

Rhoderick, E. H. and Williams, R. H. 1988. *Metal Semiconductor Contacts*. Oxford, U.K.: Oxford Science Publications.

Sah, C. T., Forbes, L., Rosier, L. L. et al. 1970. Thermal and optical emission and capture rates and cross sections of electrons and holes at imperfection centers in semiconductors from photo and dark junction current and capacitance experiments. *Solid-State Electron.* **13**: 759.

Santori, C., Fattal, D., Vučković, J. et al. 2002. Indistinguishable photons from a single-photon device. *Nature* **419**: 594.

Schliwa, A., Winkelnkemper, M., and Bimberg, D. 2007. Impact of size, shape, and composition on piezoelectric effects and electronic properties of In(Ga)As/AaAs quantum dots. *Phys. Rev. B* **76**: 205324.

Schulz, S., Schnüll, S., Heyn, C. et al. 2004. Charge-state dependence of InAs quantum-dot emission energies. *Phys. Rev. B* **69**: 195317.

Shchukin, V. A., Ledentsov, N. N., Kopev, P. S. et al. 1995. Spontaneous ordering of arrays of coherent strained islands. *Phys. Rev. Lett.* **75**: 2968.

Shchukin, V. A., Ledentsov, N. N., and Bimberg, D. 2003. *Epitaxy of Nanostructures*. Berlin, Germany: Springer.

Sheikholeslami, A. and Gulak, P. G. 2000. A survey of circuit innovations in ferroelectric random-access memories. *Proc. IEEE* **88**: 667.

Stier, O., Grundmann, M., and Bimberg, D. 1999. Electronic and optical properties of strained quantum dots modeled by 8-band-k.p theory. *Phys. Rev. B* **59**: 5688.

Stievenard, D., Boddaert, X., and Bourgoin, J. 1986. Irradiation-induced defects in p-type GaAs. *Phys. Rev. B* **34**: 4048.

Stranski, I. N. and Krastanow, L. 1938. Zur Theorie der orientierten Ausscheidung von Ionenkristallen aufeinander. *Sitzungsber. Akad. Wiss. Wien, Math.-Naturwiss. K1, Abt. 2B* **146**: 797.

Sze, S. M. 1985. *Semiconductor Devices—Physics and Technology*. New York: John Wiley & Sons.

Sze, S. M. 1999. Evolution of nonvolatile semiconductor memory: From floating-gate concept to single-electron memory cell. In *Future Trends in Microelectronics*, S. Luryi, J. Xu, and A. Zaslavsky (Eds.), p. 291. New York: John Wiley & Sons.

Sze, S. M. 2002. *Semiconductor Devices*. New York: John Wiley & Sons.

Tiwari, S., Rana, F., Hanafi, H. et al. 1996. A silicon nanocrystal based memory. *Appl. Phys. Lett.* **68**: 1377.

Vurgaftman, I., Meyer, J. R., and Ram-Mohan, L. R. 2001. Band parameters for III–V compound semiconductors and their alloys. *J. Appl. Phys.* **89**: 5815.

Waser, R. 2003. *Microelectronics and Information Technology*. Berlin, Germany: Wiley-VCH.

White, M. H., Adams, D. A., and Bu, J. K. 2000. On the go with SONOS. *IEEE Circuits Devices* **16**: 22.

Wolf, S. A., Chtchelkanova, A. Y., and Treger, D. M. 2006. Spintronics—A retrospective and perspective. *IBM J. Res. Dev*. **50**: 101.

3

Carbon Nanotube Memory Elements

3.1 Introduction 3-1
3.2 CNFET-Based Memory Elements 3-2
General Description of a CNFET • FET-Based Memory Elements: Pioneering Works • FET-Based Memory Elements: Further Improvements • FET-Based Memory Elements: Controlling Storage Nodes • FET-Based Memory Elements: Optoelectronic Memory • Redox Active Molecules as Storage Nodes • Two-Terminal Memory Devices • Using Nanotubes as Storage Nodes • Conclusions
3.3 NEMS-Based Memory 3-9
NEMS: Generalities • Carbon Nanotube Crossbars for Nonvolatile Random Access Memory Applications • Nanorelays • Feedback-Controlled Nanocantilevers • Data Storage Based on Vertically Aligned Carbon Nanotubes • Linear Bearing Nanoswitch • Conclusions
3.4 Electromigration CNT-Based Data Storage 3-16
A "Bucky Shuttle" Memory Element • Memory Elements Based on the "Bucky Shuttle" Concept • Conclusions
3.5 General Conclusions 3-18
Acknowledgments 3-18
References 3-19

Vincent Meunier
Oak Ridge National Laboratory

Bobby G. Sumpter
Oak Ridge National Laboratory

3.1 Introduction

An incredibly rapid expansion in the demand of portable consumer electronics ranging from personal computers, cellular phones, MP3 players, digital cameras, PDAs, USB memory sticks to applications in the networking arena continues to drive the pursuit for the development of higher density, faster, more efficient, and economical electronic memory devices (Scott 2004). Currently, most commercially available memory elements consist of dynamic random access memory (DRAM), static random access memory (SRAM), or flash memory (nonvolatile memory). DRAM is considered to be very fast and cheap but its contents are lost when power is switched off (volatility). SRAM is much faster and needs less power but is far more expensive and also suffers from volatility. Flash memory is nonvolatile but currently operates at low write speed and thus has a slow random access. In addition, it is power-hungry and very expensive. In the past, incremental improvements in memory capacity and capability were primarily achieved by the simple scaling of the physical dimensions of the devices. However, the semiconductor industry is quickly reaching the fundamental limits on the miniaturization, encountering, for example, extreme difficulties due to short channel effects in the scaling of metal-oxide-semiconductor field-effect transistor (MOSFET). This difficulty has stimulated an intense growth in research and development in the area of new technologies and materials that can overcome these limitations and deliver unprecedented capabilities for memory storage and access. Some promising new types of memory-storage devices include magnetoresistive RAM (MRAM), ferroelectric RAM (FRAM), phase-change memory (PRAM), and novel high speed–density nonvolatile memory (these are primarily based on carbon nanotubes [CNTs]) (Waser 2003).

CNT-based data-storage devices or memory elements offer high potential for considerable improvements compared to current memory paradigms (Bichoutskaia et al. 2008, Zhang 2006). First, CNTs are considerably smaller than conventional cells, e.g., each nanotube is less than 1/10,000th the width of a human hair, thus allowing an impressive density to be achieved in a modest space. The intrinsic structural, thermal, and mechanical stability of CNTs are expected to confer to these memory devices a very long lifetime while the power requirements for their operation should be considerably less than DRAM. In addition, CNTs are known to be radiation-resistant and nanotube-based memory elements are therefore expected to be robust in numerous types of environments. As such, these novel types of memory elements have recently been touted as a "universal memory" (Bichoutskaia et al. 2008).

In this chapter, we review the developments in memory elements that exploit the properties of CNTs. These nanoscale systems fit into the realm of nanoelectronics and offer exceptional properties and well-characterized structures that have rapidly emerged as a viable route to satisfy the future requirements and needs for data-storage and memory devices. Given the difficulty in the integration of nanoscale systems, such as CNTs, into adequate arrangements for useful structures and devices, and the complexity in the assessment of the properties of the devices, a tremendous effort has been devoted in the past years toward the development of experimental and theoretical methods making it possible to creatively imagine, design, and test new systems for desired and tailored characteristics. In this chapter, we review the different types of CNT-based memory devices, which include memory elements based on CNT field-effect transistors (CNFET) (Section 3.2), nanoelectromechanical systems (NEMS) (Section 3.3), and electromigration (Section 3.4). The fundamental operational principles and characteristics of the different devices and concepts are examined in detail along with a discussion of the current status of experimental fabrication and practical realization.

3.2 CNFET-Based Memory Elements

A number of CNT memory elements have been developed from the functioning principle of CNFETs. The following section will be devoted to a description of a few of them. First, we will expose the basic working principle of the CNFET before showing how CNFETs can be implemented for memory applications.

3.2.1 General Description of a CNFET

Nanotube transistors are among the possible replacements for silicon field-effect transistors (FETs) as bulk-device scaling approaches quantum limits. A CNFET is a device consisting of a single nanotube and three contacts (Figure 3.1a). It has been demonstrated repeatedly in the laboratory to possess characteristics similar to those of conventional silicon FETs. Because nanotubes can be used as either n- or p-type CNFETs, existing complementary metal-oxide-semiconductor (CMOS) design techniques can be utilized for a nanotube-based FET technology. A CNT-based CMOS industry is very appealing, in particular because with their electrical and thermal properties, CNTs can sustain very large current in the billion amperes per square centimeter regime (about three orders of magnitude larger than conventional metallic wires).

The main hurdle facing nanotube-FET technology remains the difficulty of developing scalable manufacturing methods to compete with the well-developed silicon large-scale production.

CNTs can behave as either conductors or as semiconductors, depending on the arrangement of the atoms along the nanotube (Bernholc et al. 2002, Charlier et al. 2007). While about one-third of nanotubes are metals, the bandgap of the semiconducting tubes varies inversely with the tube diameter, making it possible to design a large range of devices and heterostructures. Conversely,

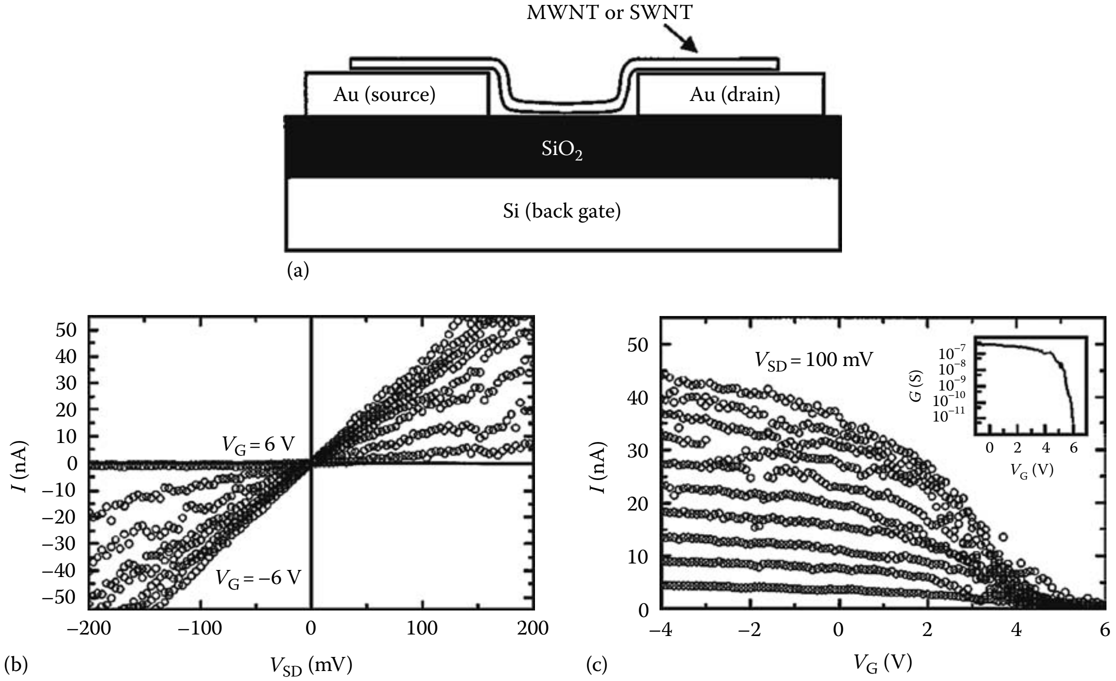

FIGURE 3.1 (a) Schematic cross section of a typical CNFET device. A single CNT of either multiwall- or single-wall-type bridges the gap between two gold electrodes. The silicon substrate is used as back gate. Output and transfer characteristics are shown in (b) and (c), respectively. The current vs. source–drain voltage curves in (b) were measured for gate voltage values of −6, 0, 1, 2, 3, 4, 5, and 6 V. In (c) the current vs. gate voltage are given for 10 different source–drain voltages between 10 and 100. The inset shows that the gate modulates the conductance by five orders of magnitude (for a source–drain voltage of 10 mV). (Reprinted from Martel, R. et al., *Appl. Phys. Lett.*, 73(17), 2447, 1998. With permission.)

conducting nanotubes possess a very low resistance for electron flow (ballistic conductors) because the decoherence length of the nanotube wavefunctions is typically larger than practical nanotube lengths. As long as the nanotube length is shorter than the decoherence length, the resistance will not depend upon the cylinder length. In that case, the only electrical resistance is at the contact points where the electrons enter and leave the tube. As shown below, it is precisely at the interface between the nanotube and the metallic electrodes that the physical processes responsible for the functioning of a nanotube transistor take place.

CNFETs can be fabricated with a structure very similar to silicon FETs. As shown in Figure 3.1a, such a device consists of a single-walled semiconducting nanotube whose ends are attached to two metallic contacts. In the terminology of transistors, the nanotube acts as the channel and the two contacts are the drain and the source, respectively. The device is completed by a gate electrode that is positioned somewhere between the source and the drain. There are a number of possible designs: the nanotube can be positioned above or under the metallic contacts and the gate can be placed as a back or front gate. For a transistor to work properly the semiconductor CNT must be doped. In air, at room temperature, the device behaves as a depletion mode p-FET (the *I*–*V* is typical of hole conduction, Figure 3.1b). The transistor action is due to the electric field around the gate modulating the Schottky barrier at the points where the source and drain terminals contact the nanotube. For reasonably short tubes (a few hundred nanometers or less), the effect of the gate voltage on channel resistance is minimal. Experimentally, the first-reported CNFETs were all p-type devices. This is because when exposed to air, nanotubes are p-doped due to the presence of oxygen that shifts the Fermi level at the metal–nanotube interface toward the valence band, leading to a hole-conduction mechanism. When the same device is heated in a vacuum and allowed to outgas, the Fermi level shifts toward the conduction band, which leads to an electron-conduction mechanism, i.e., acts as an n-type FET. Different types of FETs can be obtained depending on chemical doping, leading to p-type, n-type, or even p–n-type (ambipolar) systems.

Here we describe the general functioning of a p-type transistor. An understanding for the n-type systems can be easily obtained by considering electrons as majority carriers. As shown in Figure 3.2, the application of a negative gate voltage increases the number of majority carriers in the channel, thereby leading to an increased current density across the transistor. For positive gate voltage, the current is negligibly small, due to the depletion in the number of holes. At the same time there is no evidence of electron conduction, even at very large positive gate voltages.

While nanotubes show great promise in the mechanical and microelectronic realms, major problems still exist (Bernholc et al. 2002). Assuming the ability to produce large quantities of nanotubes with specific geometries, a separate difficulty lies in actually assembling large quantities of transistors out of them and placing those transistors, along with an interconnect, on a substrate to form a complete circuit. Solutions to these problems are only beginning to surface. With respect to electronics, the biggest difficulty is devising a production method that can provide the same facility for large-scale fabrication as silicon photolithography. This goal appears distant but it continues to be a topic of intense research. With the demonstrated potential payoff for a range of applications, it seems likely that a solution to this problem will eventually be found. Further details on the general aspects of CNFETs can be found in the vast literature published on the topic (e.g., Derycke et al. 2001, 2002, Martel et al. 2001, and Tans et al. 1998). Below we review how CNTs have been implemented for FET-based memory applications.

3.2.2 FET-Based Memory Elements: Pioneering Works

Nonvolatile and volatile memory elements can be designed from the functioning of a CNFET (Wunnicke 2006). The storage media is typically a charge that is stored in the vicinity of the nanotube

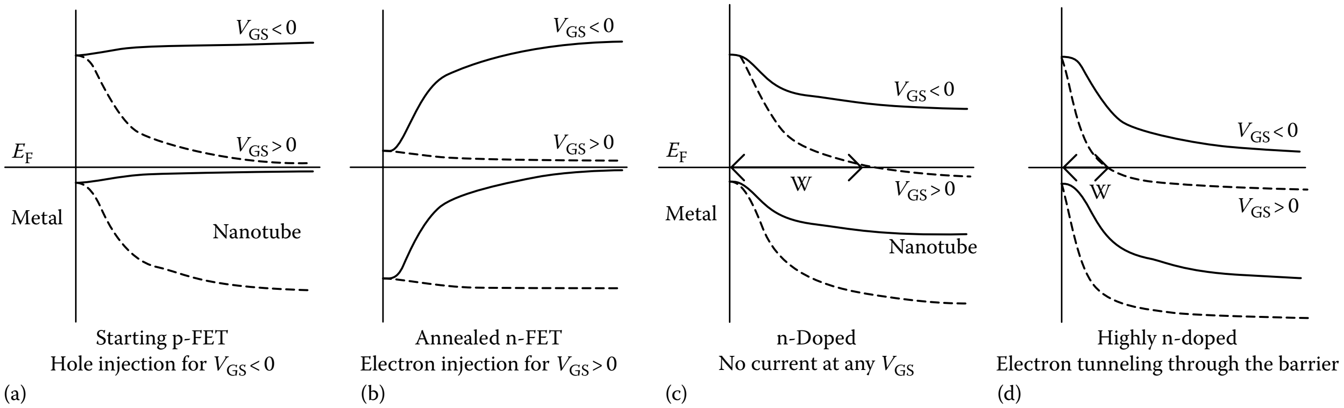

Starting p-FET
Hole injection for $V_{GS} < 0$
(a)

Annealed n-FET
Electron injection for $V_{GS} > 0$
(b)

n-Doped
No current at any V_{GS}
(c)

Highly n-doped
Electron tunneling through the barrier
(d)

FIGURE 3.2 Schematic of the bands in the vicinity of the contacts as a function of the electrostatic gate field for a p-type (a), ambipolar (b), and n-type (c) CNFETs. Annealing the device in vacuum results in a transition from the situation of (a) to that of (c) in case of low doping and (d) for high doping. The transition can be reversed by the introduction of air, clearly showing that the presence of oxygen p-doping of the nanotube. It is clear that the Schottky barrier at the nanotube–metal interface plays a key role. Note also the importance of Fermi-level pinning for the functioning of the device. (Reprinted from Derycke, V. et al., *Appl. Phys. Lett.*, 80(15), 2773, 2002. With permission.)

channel. The charge causes a threshold potential shift (the threshold potential is defined as the potential at which the majority carriers start flowing) of the nanotube FET. Because the nanotube has very high carrier mobility, the information can be stored in a few (as few as one) electrons configuration. The states can be reversibly written and erased. The reading mechanism is performed via a measurement of the source–drain current, while the writing mechanism is completed using a large bias voltage applied to the gate.

A number of groups have obtained convincing results on CNFET-based memory elements. How the information is stored and consequently under which conditions the memory is usable depends on the details of the manufacturing, particularly in the gating material. The first two examples of FET-memory elements were proposed independently by Fuhrer et al. (2002) and Radosavljevic et al. (2002). The functioning of the devices hinges on the operation of a single-electron memory. In this case, the capacitance of the storage node must be small enough so that its Coulomb-charging energy is significantly larger than the thermal energy at the operating temperature, and the readout device must be sensitive enough to detect a single nearby electronic charge. In Fuhrer's device, the charge is reversibly injected and removed from the dielectric (placed between the tube and the gate electrode) by applying a moderate (10 V) bias between the nanotube and the substrate. The nanotube is ideal as a charge-detecting device due to its high carrier mobility, large geometrical capacitance, and its one-dimensional (1D) nature ensuring that local changes in charge affects the global conductance (due to slow screening). In Fuhrer's pioneering work, for instance, discrete charge states corresponding to differences of a single, or at most a few, stored electrons are observed. The device is based on the characteristics of a p-type FET (i.e., it conducts at negative gate voltage and becomes insulating at positive gate voltage) and can be operated at temperatures up to 100 K. Properties relevant to memory operation are demonstrated by the large hysteresis I–V_g curve that is obtained by sweeping the gate voltage V_g between −10 and +10 V. As is shown in Figure 3.3, the threshold voltage is shifted by more than 6 V. The mechanism of charge storage is related to the rearrangement of charges in the dielectric or by the injection or removal of charges from the dielectric through the electrodes or the nanotube. Due to the geometry of the device, the electric field at the surface of the

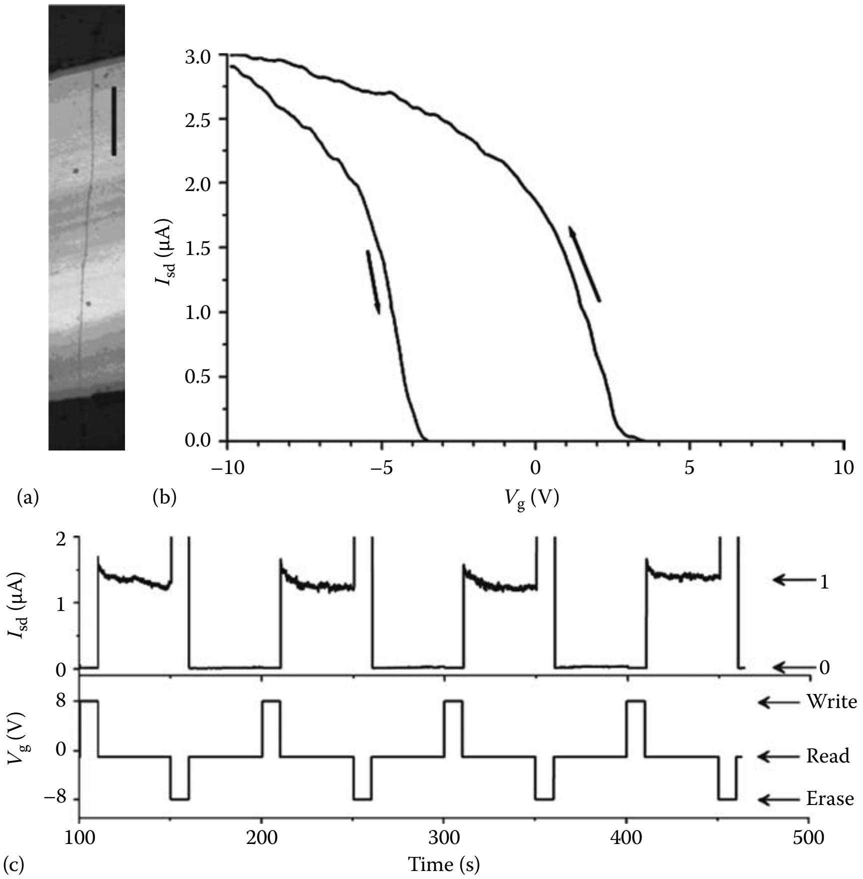

FIGURE 3.3 (a) Atomic force microscope topographic image of the nanotube device used in Fuhrer et al. (2002). The nanotube extends between the two dark blocks at the top and bottom of the image (i.e., the electrodes). The scale bar represents 1 μm. (b) Drain current as a function of gate voltage at room temperature and a source–drain bias of 500 mV. As the gate voltage is swept from positive to negative and back, a strong hysteresis is observed, as indicated by the arrows denoting the sweep direction. (c) Series of four read–write cycles of the nanotube memory at room temperature. The upper panel shows the drain current at a source–drain bias of 500 mV, while the lower panel shows the gate voltage. The memory state was read at −1 V, and written with pulses of ±8 V. (Reprinted from Fuhrer, M.S. et al., *Nano Lett.*, 2(7), 755, 2002. With permission.)

nanotube is very large, at least large enough to cause the movement of charge in the dielectric (Fuhrer et al. evaluated the field to be comparable to the breakdown field in SiO_2). At these high fields the electrons are easily injected into the dielectric from the nanotube and remain trapped in metastable states until the polarity is reversed.

In Radosavljevic's device, the FET used is based on an n-type transistor (obtained from annealing the nanotube in hydrogen gas). Again, it was shown that CNFETs are extremely sensitive to the presence of individual charges around the channel, largely because of the nanoscale capacitance of the CNTs. A technique known as scanning gate microscopy (SGM) (Freitag et al. 2002, Meunier et al. 2004) was used to study the local transport properties across the channel, revealing the key role of nanotube–metal contacts (Schottky barriers) and also the positions of the possible sites where the electrons are trapped (i.e., where the information is physically stored, the storage nodes). Figure 3.4 shows the reproducible hysteresis loop in the I–V_g curve that becomes larger as the range of V_g is increased, indicating that it originates from avalanche injection into bulk oxide traps. The hysteresis can become so large that the device varies from depletion mode (normally on at $V_g = 0$) and enhancement mode (normally off) behavior. The location and sign of the trapped charge can be determined by examining the directions of the hysteresis loop. After a sweep to positive V_g, the CNFET threshold voltage moves toward more positive gate values, indicating injection of negative charges into oxide traps. To read the memory, a 1 MΩ load resistor is added to create a voltage divider. Read ($V_{in} = 0$) and write ($V_{in} = +20$ or -20 V) are applied to the input terminal (back gate). A logical gate (1 or 0) is defined as $V_{out} = 1$ or 0. To write a "1" to the memory cell, V_{in} is switched rapidly to −20 or +20 V and back to "0," so the CNFET is in the on or off at the read voltage. This memory device was found to be nonvolatile at room temperature and with a bit-storage retention time of at least 16 h. The trap charging time limits the speed of the device and the bit was evaluated to be stored in no more than 2, 70, and 200 e.

In the pioneering works discussed above, it is noteworthy that the storage nodes are not precisely the nanotube itself, but consist of trapping sites in the dielectric layer of the gating material. The functional temperatures, storage density, operating speeds, etc., are very sensitive to the details of the trapping mechanisms. As summarized below, a number of studies have been reported following the seminal studies of Fuhrer and Radosavljevic, where the trapping mechanism was further analyzed and modified.

3.2.3 FET-Based Memory Elements: Further Improvements

Charge-storage stability of up to 12 days at room temperature was reported in a 150 nm long nanotube channel (Cui et al. 2002). The observed conductance decreases for increasing gate

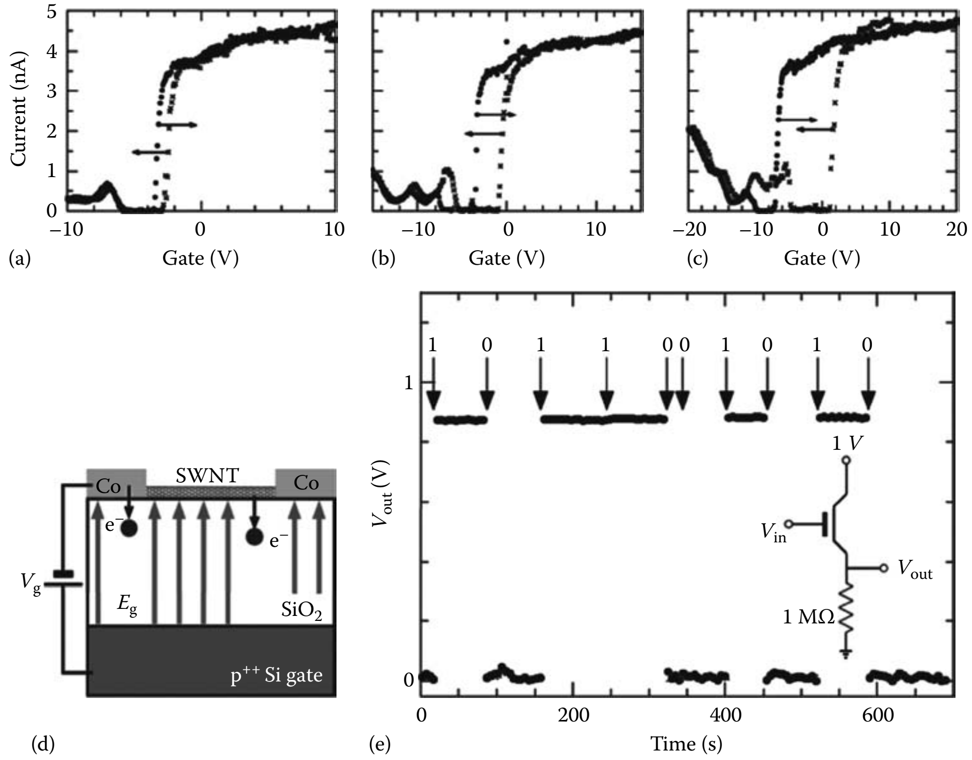

FIGURE 3.4 (a–c) Current vs. gate voltage data obtained in high vacuum for a source–drain voltage of 0.5 mV. The device hysteresis increases steadily with increasing gate voltage due to avalanche charge injection into bulk oxide traps, schematically shown in (d). (e) Demonstration of the CNFET-based nonvolatile molecular memory cell. A series of bits is written into the cell and the cell contents are continuously monitored as a voltage signal (V_{out}) in the circuit shown in the inset. (Reprinted from Radosavljevic, M. et al., *Nano Lett.*, 2(7), 761, 2002. With permission.)

potential, a normal feature of p-type, air-exposed single-wall carbon nanotubes (SWCNTs). The hysteresis loop is obtained by sweeping continuously the gate voltage from −3 to +3 V. This shows two conductance states at $V_g = 0$ differing by more than two orders of magnitude, associated with a threshold voltage of 1.25 V. The method used by Cui et al. is more technologically straightforward compared to the previously reported techniques, as it does not require separating SWCNT bundles into individual SWCNTs during sample preparation. The sample treatment is likely to be responsible for the modified type of trapping centers, compared to earlier similar devices. The heat treatment and exposure to oxygen plasma cause the metallic nanotubes present in the bundle to be preferentially oxidized, leading to increased gate dependence due to the remaining intact semiconducting tubes. In addition, oxidation-related defects are likely to be formed in the remaining amorphous carbon particles on the bundle surface or at the SiO_2 interface. These defects act as charge-storage traps and their close proximity to the surface of the channel accounts for the large threshold voltage shifts.

Another type of charge trapping was presented by Kim et al. (2003). In that case, water molecules were shown to be responsible for the hysteresis properties (i.e., responsible for the threshold potential shift) of the I–V_g curves. The water molecules contributing to the trapping could be located on the nanotube surface or on the SiO_2 close to the tube. Heating under dry conditions significantly reduces the hysteresis. A completely hysteresis-free CNFET was possible by passivation of the device using a polymer coating, clearly indicating that the storage nodes were removed by the treatment. This work also confirms the central role played by surface chemistry on the properties of the device and that truly robust passivation is needed in order to use CNT-based devices in practical electronic devices, unless it is simply used as a detection device (e.g., of humidity).

More recent work of Yang et al. confirmed the role of adsorbed (including water and alcohol) molecules for charge trapping, also showing charge retention of up to 7 days under ambient conditions (Yang et al. 2004).

3.2.4 FET-Based Memory Elements: Controlling Storage Nodes

Soon after the publication of the pioneering works of 2002, a number of groups confirmed that the as-prepared nanotube FETs possessed intrinsic charge-trapping centers that were responsible for the shift in the threshold voltage measured in the hysteresis loop of the I–V_g characteristics. As discussed briefly above, the trapping centers can be defects in the SiO_2, water, or alcohol molecules adsorbed on the tube or the dielectric, or oxidized amorphous carbon. It was soon understood that a better control of the storage medium was required in order to harness the full potential of CNFET-based memory elements.

In 2003, Choi et al. proposed a SWCNT nonvolatile memory device using SiO_2-Si_3N_4-SiO_2 (ONO) layers as the storage node (Choi et al. 2003). In that device, the top gate structure is placed above the ONO layer, which is positioned directly above a few-nanometers-long channel. Charges can tunnel from the CNT surface into the traps present in the ONO layers. The stored charges impose a threshold voltage shift of 60 mV, which is independent of charging time, suggesting that the ONO traps present a quasi-quantized state. The choice of SiO_2-Si_3N_4-SiO_2 is motivated by the fact that it presents a high breakdown voltage, low defect density, and high charge-retention capability (Bachhofer et al. 2001).

Another type of device was demonstrated experimentally in 2005 (Ganguly et al. 2005). In this case, the charge-storage nodes consist of gold nanocrystals placed on the top gate above the nanotube channel. The device was found to have a large memory window with low voltage operations and single-electron-controlled drain currents. The device is based on a Coulomb blockade behavior, in other words on the difficulty of adding supplementary electrons due to the large charging effect of the system with large capacitance and Coulomb repulsion. The Coulomb blockade in the nanocrystals, along with the single-charge sensitivity of the nanotube FET, is suggested to allow for multilevel operations. In this device, the reported retention is rather modest: 6200 s at 10 K and only 800 s at room temperature. Better dielectric and preprocessing should improve these low retention times.

More recently, Sakurai et al. proposed a modified design where the dielectric insulator was made up of a ferroelectric thin film (Sakurai et al. 2006). The experimental results show that the carriers in the CNTs are controlled by the spontaneous polarization of the ferroelectric films. Ferroelectric-gate FETs offer potential advantages as nonvolatile memory elements, such as low power consumption, the capability of high-density integration, and nondestructive readable operation. So far, the use of ferroelectric thin films in conventional Si-based FETs for memory applications has been complicated because of the extreme difficulty of fabricating the ferroelectric/Si structure with good interface properties due to the chemical reaction and interdiffusion of Si and ferroelectrics. It was suggested that this problem could be resolved by inserting an insulating layer between the ferroelectric and the semiconductor layers, but at the cost of increased voltage operation and depolarization field. Another method was to use an oxide semiconductor, such as indium tin oxide. However, even though this system also shows good memory operation, the carrier mobility is much lower than in Si, causing problems that are unacceptable for high-performance nonvolatile memory elements. The use of a semiconducting tube as the conducting channel and a ferroelectric as a gate insulator alleviates all of these problems: the new design allows the carriers in the channel to be controlled by the spontaneous polarization of the ferroelectric film, while keeping the high-conduction characteristics of CNTs. In the demonstrated prototypical device, a 400 nm thick PZT ($PbZr_{0.5}TiO_5O_3$) film was used. The threshold voltage was found to be shifted in the positive direction when V_g was swept from negative to positive bias and was shifted in the negative direction when V_g was swept from positive to negative. This gave a clear hysteresis loop and bistable current values at $V_g = 0$, due to the spontaneous polarization of the electric field (the effects of charge trapping in the dielectric was found to be negligible compared to the effects due to the ferroelectric, at least for small sweeping range of V_g).

3.2.5 FET-Based Memory Elements: Optoelectronic Memory

In all of the CNFET-based memory devices mentioned thus far in this chapter, the information is stored by applying a large enough gate voltage to store or remove the charge in, or from, the traps (for the ferroelectric-based FET, the gate is used to flip the polarization field). In Star's CNFET-memory (p-type) device, the information is still stored as electric charges (Star et al. 2004), however, it differs in that the writing mechanism relies on the use of optical illumination of a photosensitive polymer (which is part of the gate electrode) that converts photons into electric charge stored in the vicinity of the channel. The information is read by measuring the source–drain current, while it is erased (the charge is removed) using a large gate voltage. In other words, this optoelectronic memory is written optically and read and erased electrically. The threshold voltage change compared to the non-illuminated system was measured to be as large as +2 V after optical excitation, suggesting a charge transfer of about 300 electrons per micron in the tube.

3.2.6 Redox Active Molecules as Storage Nodes

In this experimentally built system, the overall configuration and operating principle consists of a nanotube or nanowire FET functionalized with redox active molecules, where an applied gate voltage or source–drain voltage pulse is used to inject net positive or negative charge into the molecular layer (Duan et al. 2002). The oxide layer, on the surface of the channel, serves as a barrier to reduce charge leakage between the molecules and the channel, and thus maintains the charge state of the redox molecules, thereby guaranteeing nonvolatility. The charged redox molecules gate the FET to a logic on state with higher channel conductance or the off state with lower channel conductance. Positive charges in the redox material act just like a positive gate, leading to an accumulation of electrons and an on state in n-type semiconducting nanowire-based FETs. For p-type FETs, the same effect is obtained with negative charges in the redox layer, which cause an increase in majority carrier density, in this case holes, in the channel (Figure 3.5). Mannick et al. also described a similar device, using SWCNTs as channels, where the conductance is switched on or off, guided by chemical reactivity of a CNT in H_2SO_4 (Goldsmith et al. 2007, Mannik et al. 2006).

3.2.7 Two-Terminal Memory Devices

Transistor devices consist of three electrodes: a source, a drain, and a gate terminal. The source and the drain are good metals (Au, Pt, Co, etc.) and the gate terminal is typically made up of a dielectric layer (e.g., SiO_2) sandwiched between the channel and a heavily doped Si-wafer, which is hooked up to the gate battery. The present authors have theoretically proposed a modified version of the FET memory device where the physical gate is replaced by a virtual electrode. The role of the

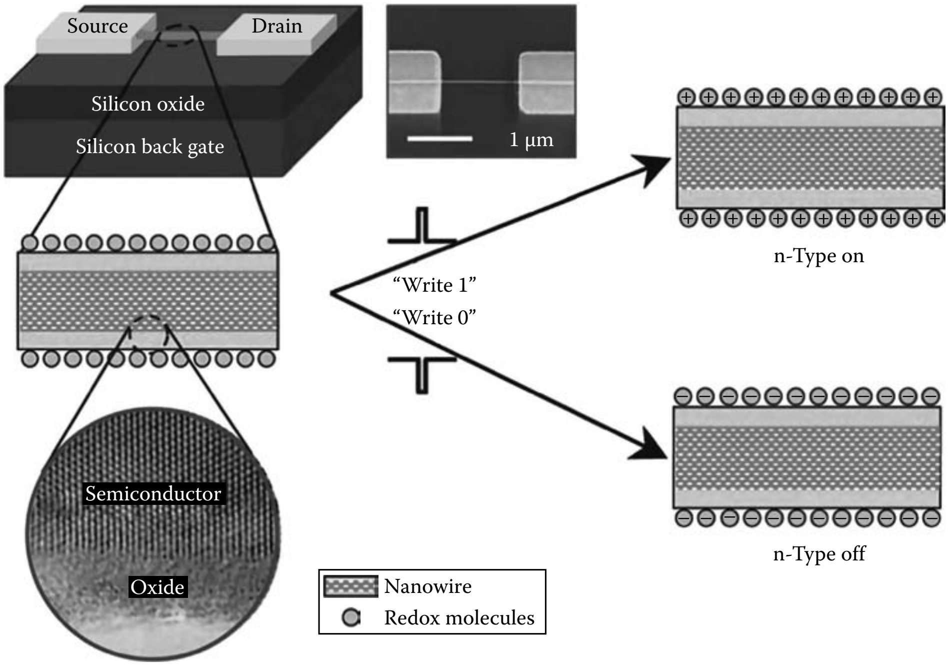

FIGURE 3.5 Nanowire-based nonvolatile devices. The devices consist of a semiconductor nanowire (NW) configured as a FET with the oxide surface functionalized with redox active molecules. The top-middle inset shows a scanning electron microscope (SEM) image of a device, and the lower circular inset shows a TEM image of an InP NW highlighting the crystalline core and surface oxide. Positive or negative charges are injected into, and stored in, the redox molecules with an applied gate or bias voltage pulse. In an n-type NW, positive charges create an on or logic "1" state, while negative charges produce an off or logic "0" state. (Reprinted from Duan, X.F. et al., *Nano Lett*., 2(5), 487, 2002. With permission.)

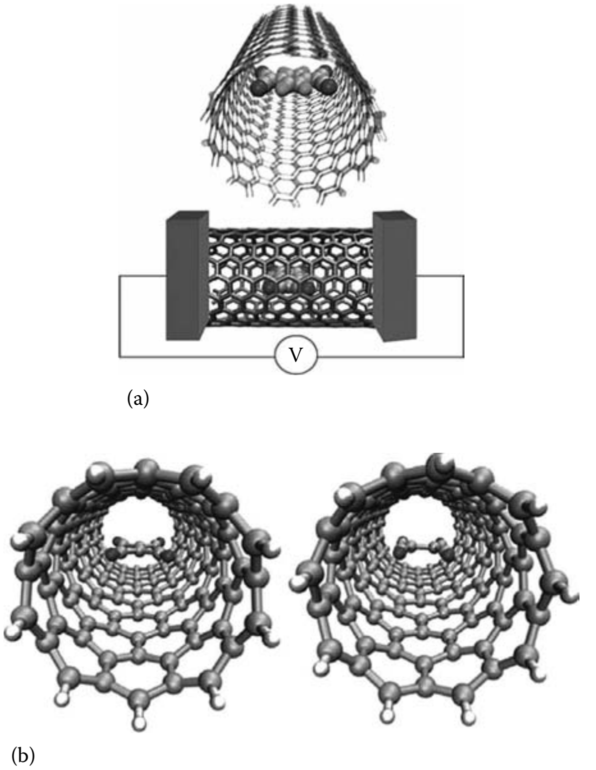

FIGURE 3.6 (a) Schematic representation of the typical setup used to probe the information stored in the nonvolatile memory element based on the encapsulation of a donor–acceptor molecule inside a metallic nanotube. In (a) the system is made up a tetrafluorotetracyano-p-quinodimethane molecule encapsulated inside a (10, 10) armchair nanotube. (b) Examples of bistable orientations of the C=C bond of tetracyanoethylene (TCNE) inside a (9, 0) zigzag CNT. (Reprinted from Meunier, V. and Sumpter, B.G., *Nanotechnology*, 18(42), 424032, 2007. With permission.)

gate is filled by a single donor or acceptor molecule embedded inside a metallic nanotube channel (Meunier and Sumpter 2007, Meunier et al. 2007) (Figure 3.6). The resulting device is a two-terminal system where the conduction properties can be turned on and off by modifying the position of the donor (or acceptor) molecule relative to the nanotube inner core. The intrinsic gating effect works for any nanotube-molecule couple, as long as the molecule possess two stable positions corresponding to a different interacting scheme with the nanotube host (bistability). The gating mechanism works as follows: In one orientation, charge transfer between the molecule and the nanotube imposes a suppression of the current. In another orientation, when the interaction is not only purely electrostatic (charge transfer), but also includes significant directional binding, the gating effect is suppressed and current flow is allowed. It was proposed that the molecule's orientation can be modified by imposing a lateral stress to the nanotube, changing the local atomic arrangement around the molecule, which in turn results in its flipping. Other optional writing mechanisms include the use of a magnetic field, optoelectronics, etc. This example underlines the importance of developing a memory device with a well-defined and well-characterized storage node (here the storage node is a single molecule). It is also a memory device that lies at the boundary of FET-type system and NEMS device, as it involves both transistor-like technology and "moving parts," on which NEMS are based, as shown in Section 3.3.

Tour and coworkers proposed a memory element, not based on FET operation, but relying on the conformation of a molecule as the storage media (He et al. 2006). The idea is based on the development of a metal-free silicon-molecule-nanotube test bed for exploring the electrical properties of single molecules that demonstrated a useful hysteresis *I*–*V* loop for memory storage.

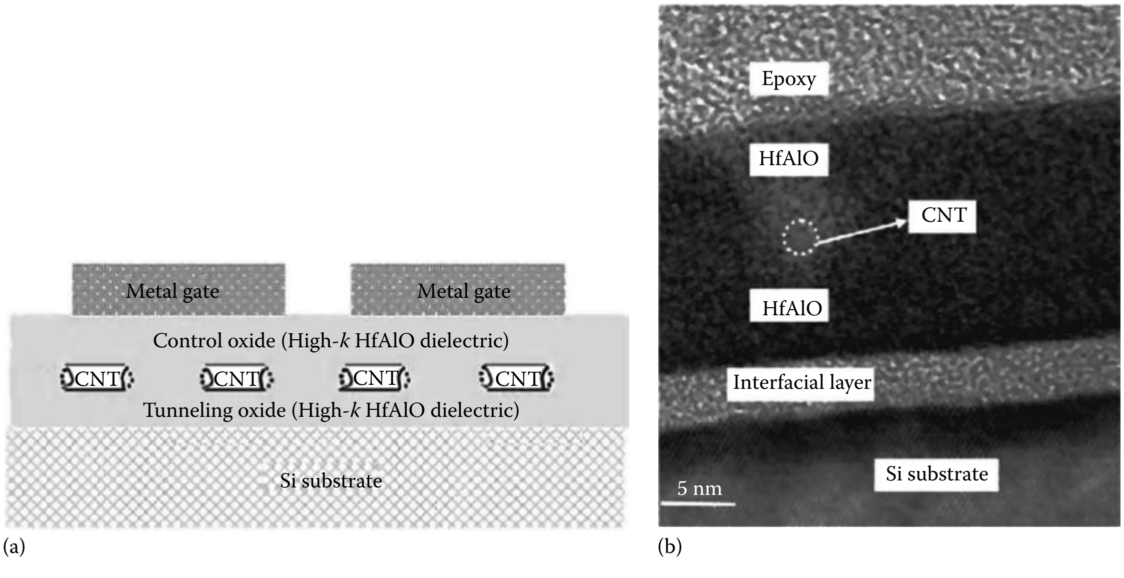

FIGURE 3.7 (a) Schematic structure of the HfAlO/CNTs/HfAlO/Si MOS memory structure of Lu and Dai where a nanotube is used as the storage node. (b) High-resolution transmission electron microscopy (HRTEM) image of CNTs embedded in HfAlO control and tunneling layers. (Reprinted from Lu, X.B. and Dai, J.Y., *Appl. Phys. Lett.*, 88(11), 113104-3, 2006. With permission.)

Nonvolatile memory for more than 3 days with cyclability of greater than 1000 read and write–erase operations was obtained for a molecular interface between the Si and SWCNTs consisting of π-conjugated organic molecules. In this case, the hysteresis in the *I*–*V* curve resulted from a conductivity change at high-voltage bias. The conduction state could be easily switched from high to low conductivity, by using a +5 V pulse, and makes it possible for the Si-molecule–SWCNT junction to be used as a nonvolatile memory element. The underlying mechanism for the switching was examined and it was concluded that it was not due to a thermoionic process but likely from a tunneling process across the molecule. Akdim and Pachter suggested that the switching in the device may be driven by conformational changes in the molecule upon the application of an electric field and that the nature of the contact at the interface of the SWCNT mat plays an important role in the switching (Akdim and Pachter 2008).

3.2.8 Using Nanotubes as Storage Nodes

Some authors have reported CNFETs, where nanotubes were used as the storage node. Yoneya et al. proposed a 40 K device with a crossed nanotube junction where one of the tubes is used as channel and the other one is used as floating gate (memory-storage node) (Yoneya et al. 2002). In this case it is the large capacitance property of the nanotube that is exploited. The floating gate is charged and discharged using a back gate. The main disadvantage of this approach is the requirement of a low operating temperature. Other reports have focused on room-temperature use of CNTs as storage nodes. Most notably, Lu et al. developed FET-based memory elements (Chakraborty et al. 2008, Lu and Dai 2006) in which the nanotubes are embedded inside the gate oxide in the metal oxide semiconductor structures. The memory structure is made up of an HfAlO/CNTs/HfAlO/Si structure. HfAlO was chosen as the tunneling and control oxides in the memory structures because of its promising performance for high-*k* gate dielectric applications and floating device applications. The schematic structure of a floating-gate memory device using CNTS as the floating gate is shown in Figure 3.7. The p-type substrate is designed as the current channel and the CNTs are designed to be embedded in the HfAlO film and act as the charge-storage nodes. Excellent long-term charge retention characteristics are expected for the memory structure using CNTs as a floating gate due to their hole-trapping characteristics, as is demonstrated in Lu and Dai's papers. While short-term charge retention was not found to be excellent in their prototype device, the memory window was found to remain at a reasonably large value over the long term.

3.2.9 Conclusions

CNFET-based devices for memory applications have received tremendous interest as is witnessed by the numerous published works outlined above. In all the examples presented, the working principle is that the channel of the transistor consists of a doped semiconducting CNT. For all the examples presented above, with the exception of those shown in Section 3.2.8, the storage node is usually located close to the channel but is not the channel itself, and depends dramatically on the surface chemistry of the dielectric, nanotube, and dielectric–nanotube interface. In those cases, it is the charge injection into the defects or charge traps in the dielectrics or interface responsible for the bistability and the memory properties.

3.3 NEMS-Based Memory

3.3.1 NEMS: Generalities

NEMS are made of electromechanical devices that have critical dimensions from hundreds to few nanometers (Ke and Espinosa 2006a,b). By exploiting nanoscale effects, NEMS offer a number of unique properties, which in some cases can differ significantly from those of the conventional microelectromechanical systems (MEMS). Those properties pave the way to applications such as force sensors, chemical sensors, and ultrahigh frequency resonators. For instance, NEMS operate in the microwave range and have a mechanical quality that allows low-energy dissipation, active mass in the femtogram range, unprecedented sensitivity (forces in the attonewton range, mass up to attograms, heat capacities below yoctocalories), power consumption on the order of 10 attowatts, and high integration levels (Ke and Espinosa 2006a,b). The most interesting properties of NEMS arise from the behavior of the active parts, which is typically in the form of cantilevers or doubly clamped beams with nanoscale dimensions. In NEMS, the charge controls the mechanical motion, and vice versa. The presence of mechanical motion demands that the moving element to possess high-quality mechanical properties, including high strengths, high Young's moduli, and low density. While limitations in strength and flexibility compromise the performance of Si-based NEMS actuators, CNTs are better candidates to realize the full potential of NEMS, in part due to their one-dimensional structure, high aspect ratio, perfect terminated surfaces, and exceptional electronic and mechanical properties. These properties, now complemented by significant advances in growth and manipulation techniques, make CNTs the most promising building blocks for next-generation NEMS (Bernholc et al. 2002, Chakraborty et al. 2007, and Yousif et al. 2008).

A majority of CNT-based NEMS devices exploit the specific propensity of CNTs to respond efficiently to capacitive forces. Capacitive forces can develop between a nanotube and a gate electrode when the presence of a gate potential induces the rearrangement of a net charge on the nanotube's surface, which in turn causes the appearance of a capacitance force between the nanotube and the gate. Because the nanotube is very flexible, the capacitance force bends the nanotube toward the gate electrode. This phenomenon hinges on two key properties of the nanotube: a large capacitance due to its shape and size and its exceptional mechanical flexibility. In addition to capacitance forces, other forces play a role in nanotube-based NEMS: elastic, van der Waals (vdW), and short-range forces. It is the delicate balance between those forces that makes the functioning of nanotube-based NEMS possible. Here we review a few examples of NEMS implementation for memory applications. We will discuss crossbar nonvolatile RAM, nano-relays, feedback-controlled nano-cantilevers, electro-actuated multiwalled nanotubes, linear-bearing nano-switches, and telescoping nanotube devices.

3.3.2 Carbon Nanotube Crossbars for Nonvolatile Random Access Memory Applications

Rueckes et al. proposed one of the most promising applications of CNTs for memory applications (Rueckes et al. 2000). The device exploits a suspended SWCNT crossbar array for both I/O and switchable bistable elements with well-defined on and off states. The crossbar consists of a set of parallel SWCNTs on a substrate and a set of parallel SWCNTs that are suspended on a periodic array of supports (Figure 3.8). The storage node is found at each place where two nanotubes cross. The bistability is obtained from a balance between the elastic energy (corresponding to the bending energy of the tube, having a minimum for a non-bent tube), and the attractive vdW energy (which creates a minimum corresponding to a situation where the upper nanotube is deflected downward into contact with the lower nanotube), as shown qualitatively in Figure 3.9. Since one minimum corresponds to a large vertical distance between the two mutually perpendicular arrays (defined roughly by the height of the supports), it corresponds to a situation where no current can flow between the two wires (this is the off state). At the second minimum (roughly 0.34 nm, the distance between two tubes in a bundle), tunneling current is possible as the tube–tube distance is small enough to allow wave-function overlap between the two tubes. The "bent" geometry therefore corresponds to the on state. The device can be switched between the on and off situation by transiently charging the nanotubes to produce attractive or repulsive electrostatic forces. In the integrated system, electrical contacts are made only at one end of each of the lower and upper sets of the nanoscale wires in the crossbar array, which makes it possible to address many device elements from a limited number of contacts. It was found that there is a wide range of parameters that yield a bistable potential for the proposed device configuration. The robustness of the two states suggests: this architecture is tolerant to variations in the structure. The other important feature of this device is the large difference in resistance between the two states (this is the key property of electron tunneling: the

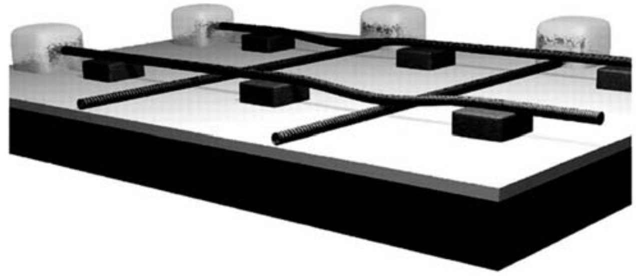

FIGURE 3.8 NRAM device made from a suspended nanotube device architecture with a three-dimensional view of a suspended crossbar array showing four junctions with two elements in the on state and two elements in the off state. The on state corresponds to nanotubes in contact and the off state to nanotubes separated. The substrate consists of a conducting layer (dark gray) that terminates in a thin dielectric layer (light gray). The lower nanotubes are supported directly on the dielectric film, whereas the upper nanotubes are suspended by periodic inorganic or organic supports (gray blocks). A metal electrode, represented by yellow blocks, contacts each nanotube. (Reprinted from Rueckes, T. et al., *Science*, 289(5476), 94, 2000. With permission.)

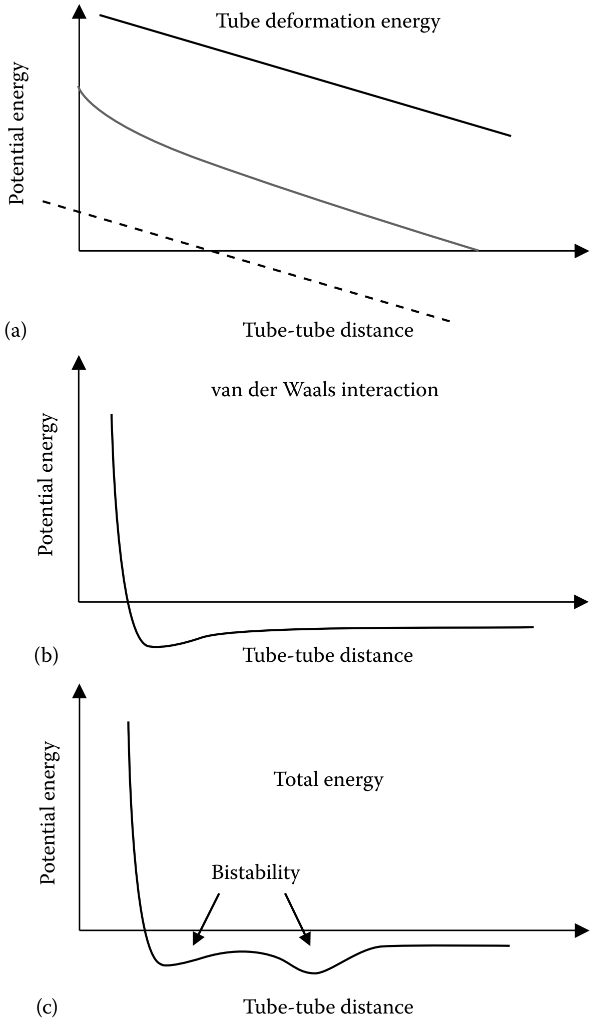

FIGURE 3.9 Interaction energies between two SWCNTs (one top and one bottom) in the device shown in Figure 3.8. (a) Mechanical strain in the top nanotube and (b) attractive van der Waals interaction, showing a minimum at the nanotube–nanotube distance of about 0.34 nm. Plot in (c) shows the resulting bi-stable potential well used for storing information. The first minimum is in the on state (contact) and the second minimum is the off state (physical separation). The existence of the energy barrier between the two minima ensures nonvolatility.

tunneling resistance decreases exponentially with the separation), which makes the device operation very reliable. The main drawback of the proposed geometry remains however, that at such small dimensions, the junction gap size imposes significant challenges in the nanofabrication of parallel device arrays.

The concept proposed by Rueckes et al. was further developed by the Nantero company, which has devoted a particular important effort toward integrating nanotube array-based memory devices for practical applications. In an IEEE communication in 2004, Ward and coworkers improved the initial device by proposing a novel technique to overcome the hurdle of manipulating individual nanotube structures at the molecular level (Ward et al. 2004). This technique allows CNT-based NEMS devices to be fabricated directly on existing production CMOS fabrication lines. The approach relies on the deposition and lithographic patterning of a 1–2 nm thick fabric of nanotubes, which retain their molecular-scale electromechanical characteristics, even when patterned with 80 nm feature sizes. Because the nonvolatile memory elements are created in an all-thin-film process, it can be monolithically integrated directly within existing CMOS circuitry to facilitate addressing and readout. The transfer of the NRAM fabrication process to a commercial CMOS foundry is ongoing, and, when commercialized, will be the first actual application of CNTs for their unique electronic properties.

A modified NEMS switch using a suspended CNT was studied by Cha et al. (2005). In that memory element, the device has a triode structure and is designed so that a suspended CNT is mechanically switched to one of two self-aligned electrodes by repulsive electrostatic forces between the tube and the other self-aligned nanotube electrodes. One of the self-aligned electrodes is set as the source electrode and the third is the gate electrode. The nanotube is suspended between the two other electrodes (acting as a drain electrode). As the gate bias increases, the force between the gate and the drain electrode deflects the suspended CNT toward the source electrode and establishes electric contact, which results in current flow between the source and the drain electrodes. The electrical measurements show well defined on and off states that can be changed with the application of the gate voltage.

Dujardin et al. exposed another type of memory device based on suspended nanotubes. In their NEMS, multiwall carbon nanotubes (MWCNTs) are suspended across metallic trenches at an adjustable height above the bottom electrode (Dujardin et al. 2005). When a voltage is applied between the nanotube and the bottom electrode, the nanotube switches between conducting and nonconducting states by physically getting closer or farther from the bottom electrode. The device acts as a very efficient electrical switch and can be improved by surface functionalizing the bottom electrode with a self-assembled monolayer. The nanotube bridge was also experimentally studied by Kang et al. who highlighted the importance of the interatomic interactions between the CNT bridge and the substrate and the damping rate on the operation of the NEM memory device as a nonvolatile memory (Kang et al. 2005).

3.3.3 Nanorelays

CNT nanorelays are three terminal devices made up of a conducting CNT positioned on a terrace on a silicon substrate (Figure 3.10). It is connected to a fixed-source electrode (single clamping) and a gate electrode is placed beneath the nanotube. When a bias is applied to it, charge is induced in the nanotube and the resulting capacitance force established between the gate and the tube triggers the bending of the tube, whose end is brought in closer contact to the drain electrode, thereby closing an electric circuit. This device was first proposed by Kinaret et al. (2003), and later demonstrated experimentally by Lee et al. (2004) and by Axelsson et al. (2005). As in other memory devices, the mechanism hinges on the existence of two stable positions (bistability), corresponding to the on and off states, respectively. The advantage of the device is that it is characterized by a sharp transition between the conducting and the nonconducting states. The large variation in the resistance is due to the exponential dependence of the tunneling resistance on the tube-end–drain distance. The transition occurs at fixed source–drain potential, when the gate voltage is varied. In this case, the tube is bent toward the drain and a large current is established, which subsequently disappears when the tube is moved far from it. Aside from possessing the property of a memory element, this device also allows for the amplification of weak signals superimposed on the gate voltage.

A number of investigations, theoretical and experimental, have been performed to further characterize and optimize the practical applications of the nanorelay for memory purposes. The fundamental property of the nanorelay is reached for a balance of the vdW, adhesive (close range), electrostatic (capacitance), and elastic forces. The so-called *pull-in* voltage (voltage required to bring the tube in contact with the drain) was first studied by Dequesnes et al. for a two-terminal device where the entire substrate acted as the gate electrode (Dequesnes et al. 2002). It was found that vdW forces reduce the pull-in voltage, but do not qualitatively modify the on–off transition. The importance of vdW forces decreases with decreasing tube length and increasing terrace height and tube diameter. The transition is, however, very sensitive to short-range attractive forces, which enhance the tendency of the CNT to remain in contact (*stiction* effect). Stiction makes

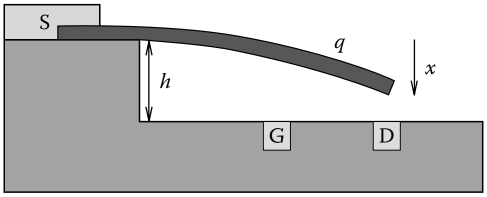

FIGURE 3.10 Nanorelay: schematic picture of the model system consisting of a conducting CNT placed on a terraced Si substrate. The terrace height is labeled h, and q denotes the excess charge on the tube. The CNT is connected to a source electrode (S), and the gate (G) and drain (D) electrodes are placed on the substrate beneath the CNT. The displacement x of the nanotube tip is measured toward the substrate. Typical practical lengths are $L \sim 50$–100 nm, $h \sim 5$ nm. (Reprinted from Kinaret, J.M. et al., *Appl. Phys. Lett.*, 82(8), 1287, 2003. With permission.)

the device unusable as it remains stuck in the on position (Jonsson et al. 2004a,b). The effect of the surface forces can be visualized by means of a stability diagram, which reveals the existence and positions of zero net force (local minima) on the cantilever as a function of gate voltage and tube deflection. A typical curve was computed by Jonsson et al. and is reproduced in Figure 3.11. A stability diagram with more than one local equilibrium for a specific voltage results in a hysteretic behavior in the I–V_g characteristics (more than one position of zero net force means that there actually are three such positions, one of which is not stable). The net force is positive to the right of the curve and negative to the left of the curve, so that, when lowering the voltage for a tube in contact with the drain electrode, it will not release until there is no stable position at the surface. The surface forces exacerbate this effect. The stiction problem corresponds to a stable nanotube position at the surface that can no longer be modified by the application of a gate voltage. The stiction problem is a general issue encountered in many nanotube-based devices. The problem of stiction hinders the use of SWNTs (because they are too flexible) as a bistable NEMS switch. MWNT and SWNT bundles are better candidates for bistable switching (Yousif et al. 2008). This problem can be alleviated, for example, by using stiffer (by avoiding total collapse of the tube, e.g., by using MWCNTs) or shorter tubes. It can also be avoided by applying a layer of adsorbates.

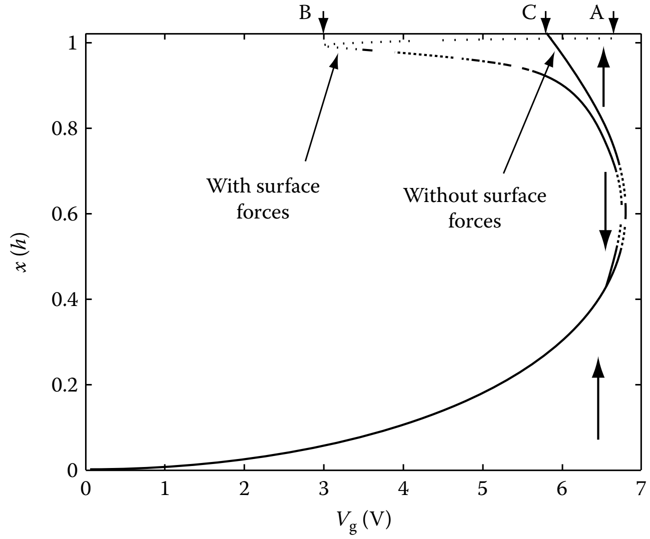

FIGURE 3.11 Nanorelay stability diagram with and without surface forces computed with parameters given in Jonsson et al. (2004b). The curve shows the positions of zero net force on the tube (or local minima) as functions of gate voltage (at constant source voltage = 0.01 V) and deflection x (in units of h, see Figure 3.10). The large arrows show the direction of the force on each side of the curves, indicating that one local equilibrium state is unstable. The required voltage for pulling the tube to the surface (*pull-in* voltage) is given by A (~6.73 V). A tube at the surface will not leave the surface until the voltage is lower than the "release voltage," B and C in the figure. Note that A > B, C, which indicates a hysteretic behavior in the current-gate voltage characteristics, a feature significantly enhanced by surface forces. (Reprinted from Jonsson, L.M. et al., *Nanotechnology*, 15(11), 1497, 2004a. With permission.)

Fujita et al. suggested a very different approach with a modified device design. They proposed using a floating gate that is built between a bundle of CNTs and electrically isolated from the input and output electrodes. The bundle of CNTs is, therefore, bent by the electrostatic force between the floating and control gate, thereby allowing the on–off threshold voltage to be controlled by changing the back-gate voltage, similar to silicon MOSFET (Fujita et al. 2007). A consequence of reducing the possibility of stiction is that a higher pull-in voltage is required, and field emission from the end of the tip can be important. Field emission will become important when the effective potential at the tube end is equal or larger to the work function of the nanotube (i.e., about 4.5 V). For a set of design parameters, field emission can be deliberately sought after (noncontact mode). In the contact mode described above, tunneling current passes from the tube to the drain electrode. In the noncontact mode, the device is designed (short tube) in such a way that the tube is never in physical contact with the drain electrodes. In that case, an electron flow is established via a field emission mechanism. The field emission current onsets with a sufficiently large source–drain voltage and then increases nonlinearly as the source–drain voltage (i.e., the applied field) is further increased. The nanotube can be switched very quickly between the on and off states, and the nanorelay operation was evaluated to work in the gigahertz regime (Jonsson et al. 2004a).

3.3.4 Feedback-Controlled Nanocantilevers

A variation of the nanorelay device presented in the previous section was proposed (Ke and Espinosa 2004), as shown in Figure 3.12a. It is made of an MWCNT placed as a cantilever over a micro-fabricated step. A bottom electrode, a resistor, and a power supply complete the device circuit. Compared to the nanorelay, this is a two-terminal device, providing more flexibility in terms of device realization and control. When the applied potential difference between the tube and the bottom electrode exceeds a certain potential, the tube becomes unstable and collapses onto the electrode. The potential that causes the tube to collapse is defined as the pull-in voltage, already encountered above (Dequesnes et al. 2002). Above the pull-in voltage, the electrostatic force becomes larger than the elastic forces and the CNT accelerates toward the electrode. At small nanotube-tip electrode distance (of the order of 0.3–1.0 nm), substantial tunneling current passes between the tip of the tube and the bottom electrode. The main difference with the nanorelay design is the presence of a resistance R placed in the circuit; with the increase in the current, the voltage drop at R increases, which causes a decrease in the effective tube-electrode potential and the tube moves away from the electrode. It follows that the current decreases and a new equilibrium is reached. Without damping in the system, the cantilever would keep on oscillating at high frequency. However, damping is usually present, and the kinetic energy of the CNT dissipates. The tip ends up at a position where the electrostatic force is equal to the

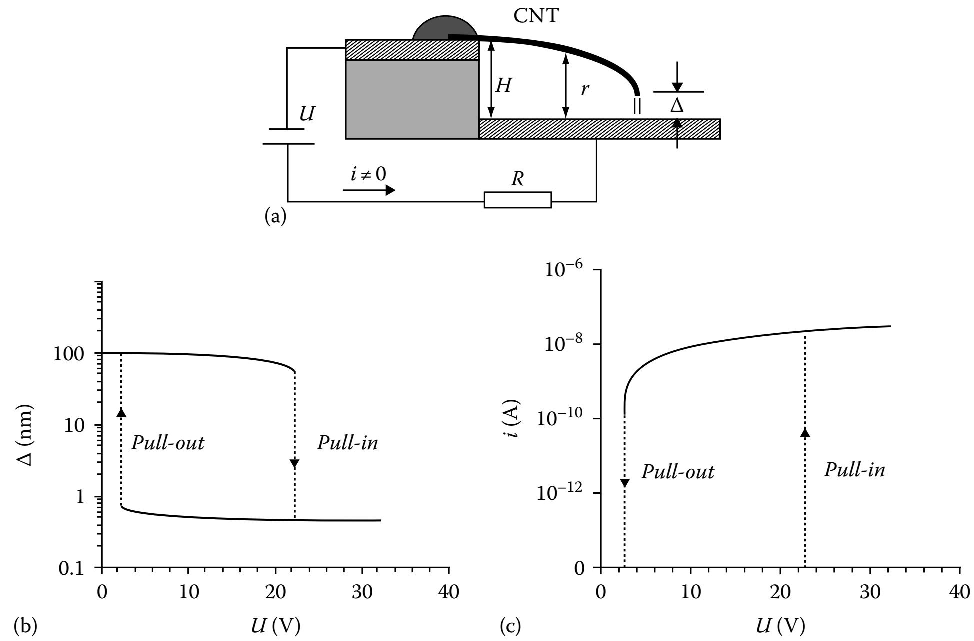

FIGURE 3.12 (a) Schematic of a feedback controlled nanocantilever device. H is the initial step height and Δ is the gap between the deflected tip and bottom conductive substrate. R is the feedback resistor, as explained in the text. (b,c) Characteristic of *pull-in* and *pull-out* processes for a device with $L = 500\,\text{nm}$, $H = 100\,\text{nm}$, and $R = 1\,\text{G}\Omega$. (b) The relation between the gap Δ and the applied voltage U. (c) The relation between the current I in the circuit and the applied voltage U. (Reprinted from Ke, C.H. and Espinosa, H.D., *Appl. Phys. Lett.*, 85(4), 681, 2004. With permission.)

elastic one and a stable tunneling current is established. This is the "lower" equilibrium position for the CNT cantilever. If the applied voltage decreases, the cantilever starts retracting. When the applied potential is lower than the so-called *pull-out* voltage, the CNT leaves the "lower" position and returns back to the upper equilibrium position where the tunneling current significantly decreases and becomes negligibly small. The pull-in and pull-out processes follow a hysteretic loop for the applied voltage and current in the device. The lower and upper positions correspond to the on and off states of the switch, respectively. The lower equilibrium state is very robust, thanks to the existence of the tunneling current and the feedback resistor. (The concept of robustness is related to the value of the ratio of current in on and off states. A ratio of 104 is usually needed for such a device to be considered robust.) Representative characteristics are reproduced in Figure 3.12b. This device has been demonstrated experimentally by Ke and Espinosa (2006a,b) and Ke et al. (2005), showing striking agreement between theoretical prediction and experimental realization. The presence of the resistor allows adjusting the electrostatic field to achieve the second stable equilibrium position. This is an advantage compared to the NRAM system developed by Rueckes et al., since it reduces the constraints in fabricating devices with nanometer gaps between the freestanding CNTs and the substrate, thereby increasing reliability and tolerance to variability in fabrication parameters. The main drawback of this device for memory applications is that the memory becomes volatile; when the applied potential is turned off, the tube retracts back to the upper position and the information stored in the position of the tube is erased.

3.3.5 Data Storage Based on Vertically Aligned Carbon Nanotubes

Another type of memory cell using a nanotube cantilever is based on the work of Kim et al. on nanotweezers (Kim and Lieber 1999). The nanotweezer consists of two vertically aligned MWCNTs that are brought together or separated by the application of a bias potential between them. Motivated by that realization, Jang et al. proposed a memory system that improves the integration density and provides a simple fabrication technique applicable to other types of NEMS (Jang et al. 2005). The device consists of three MWCNTs grown vertically from predefined positions on the electrodes, as shown in Figure 3.13. The first nanotube is electrically connected to the ground and acts as a negative electrode. Positive electrostatic charges build up in the second and the third tubes when they are connected to a positive voltage. The third tube pushes the second tube toward the first tube, because of the forces induced when the positive bias of the third tube increases, while maintaining a constant positive bias on the second one. Above a threshold bias, the second tube makes electrical contact to the first tube, establishing the on state. The balance of the electrostatic, elastic, and vdW forces involved in the device operation determines the threshold bias. If the attractive force between the first and second tube is larger than the repulsive electrostatic forces between them, they remain held together even after the driving bias is removed (otherwise they would return to the original position). Therefore, the system can act as either a volatile or nonvolatile memory element, depending on the design parameters. Note that a two-terminal memory device can be devised using the same principles. In that

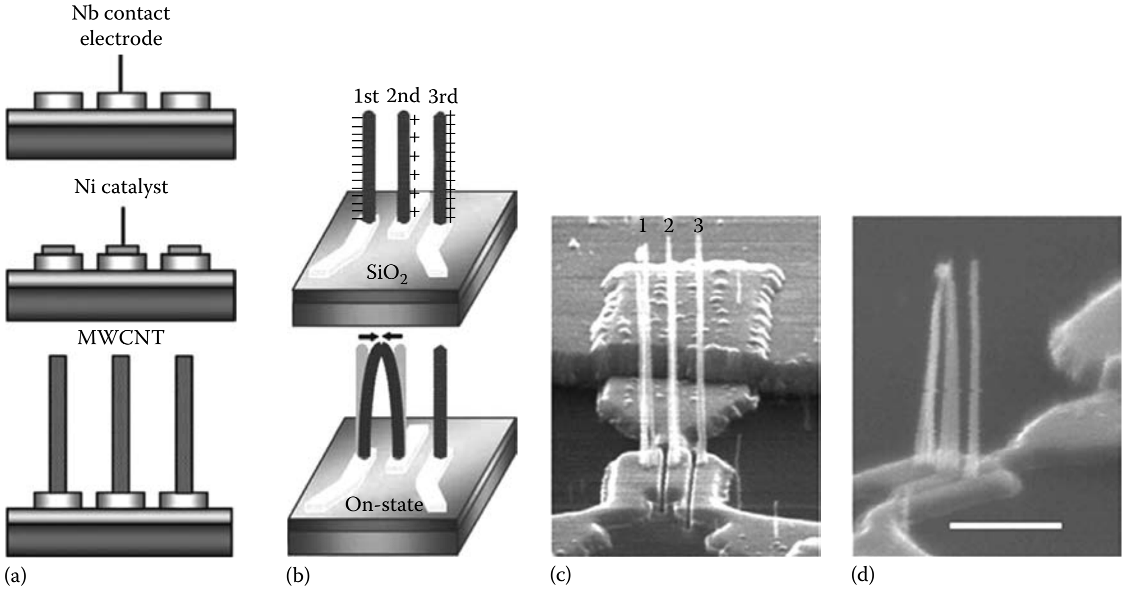

FIGURE 3.13 A schematic illustration of the CNT-based electromechanical switch device proposed by Jang et al. (a) Schematic of fabrication process: Three Nb electrodes are patterned by electron-beam lithography, followed by sputtering and lift off. Similarly, Ni catalyst dots were also formed on the predefined locations for the growth of MWCNTs. The MWCNTs were then vertically grown from the Ni catalyst dots. (b) Illustration of CNT-based electromechanical switch action. (c)–(d) SEM image of the actual device: The length and diameter of the MWCNTs are about 2 μm and 70 nm, respectively. The scale bar corresponds to 1 μm. (Reprinted from Jang, J.E. et al., *Appl. Phys. Lett.*, 87(16), 163114-3, 2005. With permission.)

case the "gate" electrode is no longer the third tube but a conventional gate that is used to separate the two tubes from their contact position.

Jang et al. also reported a nanoelectromechanical-switched capacitor structure based on vertically aligned MWCNTs in which the mechanical movement of a nanotube relative to a CNT-based capacitor defines on and off states. The CNTs are grown with controlled dimensions at predefined locations on a silicon substrate in a process that could be made compatible with existing silicon technology, and the vertical orientation allows for a significant decrease in cell area over conventional devices. It was predicted that it should be possible to read data with standard dynamic random-access memory-sensing circuitry. Jang et al.'s simulations suggest that the use of high-*k* dielectrics in the capacitors will increase the capacitance to the level needed for dynamic random-access memory applications (Jang et al. 2008).

3.3.6 Linear Bearing Nanoswitch

This type of memory device exploits the friction properties of MWCNTs. In MWCNTs, the intershell resistance force, or friction for the sliding of one shell inside another, is known to be negligibly small, as small as 10^{-14} N/atom (Cumings and Zettl 2000, Yu et al. 2000). The inner core of an open MWCNT can therefore be expected to move quasi-freely forward and backward along the tube axis, provided that at least one end of the tube is open. Deshpande et al. presented a remarkable application of the low-friction-bearing capabilities of MWCNTs to realize a NEM switch as shown in Figure 3.14 (Deshpande et al. 2006). The switch consists of two open-ended MWCNT segments separated by a nanometer-scale gap. The switching occurs through electrostatically actuated sliding of the inner nanotube shells to close the gap, producing

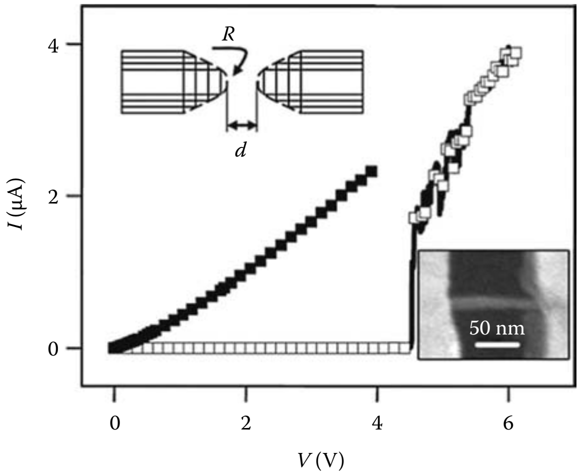

FIGURE 3.14 CNT linear bearing nanoswitch. Main panel: abrupt rise in conductance upon sweeping of source–drain voltage (open squares) and subsequent latching in the on state (filled squares). Lower-right inset: SEM image of the device after latching showing that the gap has closed. Upper-right inset: schematic cup and cone model of the system. (Reprinted from Deshpande, V.V. et al., *Nano Lett.*, 6(6), 1092, 2006. With permission.)

a conducting on state. For double-walled nanotubes, in particular, a gate voltage was found to restore the insulating off state. The device was shown to act as a nonvolatile memory element capable of several switching cycles. The authors indicate that the nanotubes are straightforward to implement, are self-aligned, and do not require complex fabrication, potentially allowing for scalability. The nanotube-bearing device is fabricated in high yield by using electric breakdown to create gaps in a single freestanding MWCNT device, producing an insulating off state. The device is actuated using electrostatic forces between the two ends of the tube that are connected to external electrodes. The force triggers a linear bearing motion that telescope the inner shells in the two multiwall or double-wall segments, so that they bridge the gap (Forro 2000). This restores electrical contact and produces the on state. Adhesion forces between the tube ends maintain the conductive state. The insulating state is controllably restored using a gate voltage, thereby completing the three-terminal nonvolatile memory devices. The gap is closed for a source–drain voltage around 9 V, leading to a conducting on state. At a large gate voltage (110 V) and small source–drain voltage (10 mV), the device snaps back to the zero conductance state. The explanation for the transition from the on state back to the off state is that the gate voltage imposes a bending stress on the nanotube and acts to break the connection. The use of elementary beam mechanics confirms that the force acting upon the two portions of the nanotube is significantly larger than the adhesion forces, allowing for the gap to be reopened.

The use of the exceptional low friction between individual walls in MWNTs is also the basic property exploited in a number of theoretically proposed structures for nanotube-based memory elements. In the following paragraphs, we will briefly review the work of Maslov (2006) and Kang and Jiang (2007).

Maslov proposed a concept that uses vertically aligned MWCNTs, that is first opened (or "peeled") layer by layer, so that the inner core is able to move along the vertical tube axis. By mounting another dielectric above the nanotube at a specific distance from the tube caps, two stable vdW states for the inner core are created and provide for nonvolatile data storage. The two stable states are (1) when the inner tube is away from the top electrode, stabilized by the interaction with the outer shell (off state, where the circuit is open) and (2) when the inner tube is in contact with the top electrode, stabilized by vdW interaction with the top electrode (on state, where the circuit is closed). The switching between the two states is realized by exploiting the electrostatic force resulting from the positive charging of the top (or bottom) electrode and negative charging of the tube inner core, pulling the core toward (away from) the top electrode. For nonvolatility purposes, it is important that the adhesion force is large enough to maintain the tube position, even when the battery potential is turned off.

Another interesting theoretical proposal using the friction properties of multiwalled systems was made by Kan and Jiang. The conceptual design and operation principle of the MWCNT-based storage are illustrated in Figure 3.15 (Kang and Jiang 2007). The main structural element of the devices is composed of a MWCNT deposited on a metallic electrode. The metallic electrode clamps the outer shell of the multiwall tube, while the inner core of the tube is allowed to freely slide back and forth (both ends of the tube outer-shell are opened). This telescoping device possesses three stable positions, resulting from the balance between the wall–wall interactions and the tube–metal electrode interactions. Two minima correspond to inner tube positions close to the electrode (allowing current to flow between that electrode and the bottom one, attached to the outer core of the tube). A third minimum corresponds to the tube positioned completely inside the outer shell. The movement of the inner core is realized using electrostatic forces induced by the voltage differences between the various electrodes. This leads to reversibility and nonvolatility, provided that the electrode material is carefully chosen (large enough vdW force to ensure contact when bias is turned off, while avoiding stiction where the nanotube cannot be separated

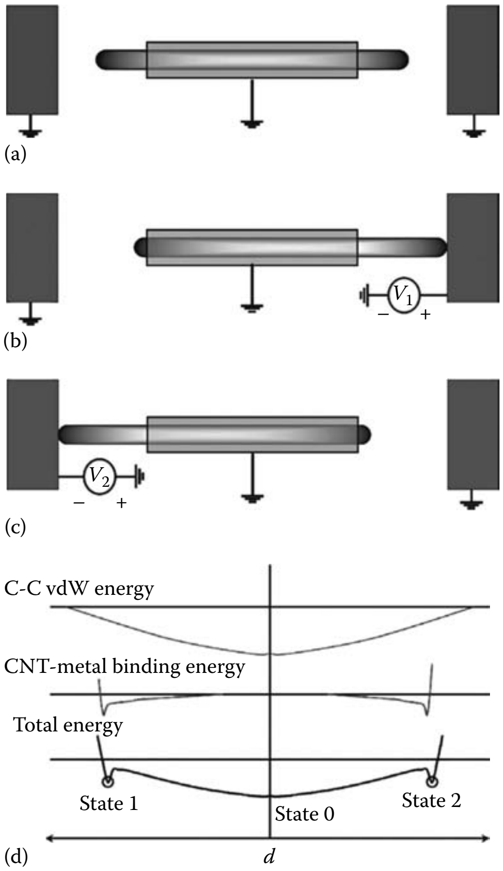

FIGURE 3.15 Electrostatically telescoping nanotube nonvolatile memory device. (a) The initial equilibrium position, (b) the core CNT contacts with the right electrode with V_1 and (c) the core CNT contacts with the left electrode with V_2. (d) Energetic schematics of the telescoping MWCNT-based memory, showing the presence of two local minima (bistability) and an energy barrier between them (nonvolatility). (Reprinted from Kang, J.W. and Jiang, Q., *Nanotechnology*, 18(9), 095705, 2007. With permission.)

from the electrode anymore due to small range forces, as illustrated in Figure 3.15b).

3.3.7 Conclusions

In this section, we have presented the state of the art of fundamental research of the use of CNTs NEMS for memory applications. The CNT-NEMS are particularly interesting because they exploit the two most remarkable properties of CNTs, namely, their electronic and mechanical properties. The viability of CNT-based NEMS switches and their comparison with their CMOS equivalents was recently presented by Yousif et al. (2008). A detailed analysis of performance metrics regarding threshold voltage control, static and dynamic power dissipation, and speed and integration density revealed that apart from packaging and reliability issues (which are the subject of intense current research), nanotube-based switches are competitive in low power, particularly low-standby power, logic, and memory applications.

3.4 Electromigration CNT-Based Data Storage

During the past decade there have been numerous promising concepts developed, based on utilizing ionized endohedral fullerenes and nanotubes as high-density nonvolatile memory elements. In this section the fundamental concepts underlying the operation and development of these types of systems are described and two specific examples are discussed in detail.

3.4.1 A "Bucky Shuttle" Memory Element

One of the earliest concepts for a CNT nonvolatile memory element was based on an ionized fullerene that could be electrically shuttled from one energy state to another inside a larger fullerene or CNT, as shown in Figure 3.16 (Kwon et al. 1999).

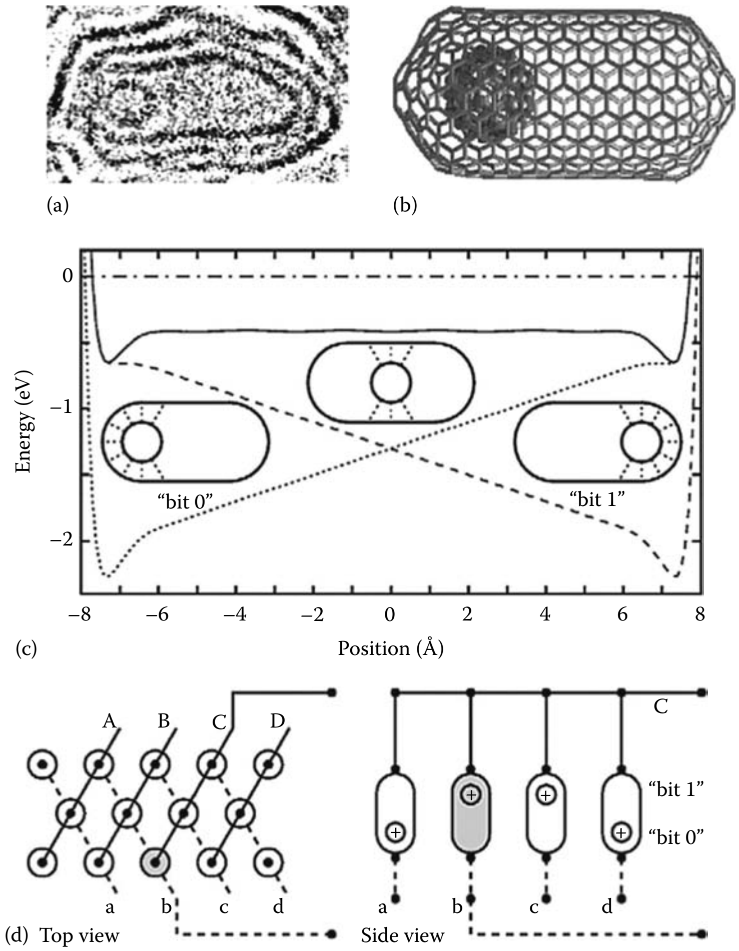

FIGURE 3.16 Illustration of the "bucky shuttle" concept as a memory element: (a) Transmission electron micrograph of a synthesized carbon structure showing a fullerene encapsulated inside a short nanotubule. (b) Molecular model depicting the ideal structure. (c) Position-dependent energy of an endohedral K@C_{60}^+ within the structure shown in (b). The results (total energy) were obtained from molecular dynamics simulations without an electric field (solid line) and following the application of a small electric field on the system (dashed line). A schematic of the corresponding C_{60} position and intrinsic state ("bit 0" or "bit 1") is shown on the figure. (d) A schematic of the overall "bucky shuttle" memory element concept that corresponds to a high-density memory board. (Reprinted from Kwon, Y.K. et al., *Phys. Rev. Lett.*, 82(7), 1470, 1999. With permission.)

One advantage of this concept was an appropriate material could be self-assembled from elemental carbon via thermal treatment of diamond power. A unique geometry of a nanotube encapsulating a fullerene, where a small fullerene structure could reside at one of the two ends of a capped nanotube or larger fullerene was experimentally identified (Figure 3.16a), and this structural arrangement became the basis for the original concept of a "bucky shuttle" memory element. In this system the on (bit 1) and off (bit 0) states correspond to the two ends of the nanotube that are local minima in the energy landscape of the endohedral fullerene (Figure 3.16c). The position of the ionized endohedral fullerene (obtained by doping with potassium, e.g., K^+@C60) can be manipulated by applying a bias voltage across the ends of the nanotube. The energetics of a switching field of 0.1 V/Å generated by applying a voltage of 1.5 V is shown in Figure 3.16c. With such a small electric field, integrity problems are not expected since graphitic structures are known not to degrade under fields of <3 V/Å. The overall thermal stability and nonvolatility of the system directly depends on the relative depth of the energy minima for the two states. Molecular dynamics simulations were used to explore the functional operation regime for the "bucky shuttle" and also to perform detailed performance analyses. Those calculations successfully demonstrated that information stored would be thermally stable well beyond room temperature, requiring temperatures above 3000 K to be destroyed. In addition, the results indicated that the "bucky shuttle" can be switched between the two states on a very fast timescale, ~10 ps, by applying a small electric field. This operation regime would allow for memory-switching access of close to 0.1 THz, giving a data throughput rate of 10 Gbyte/s in its simplest serial mode. Parallel arrays of the "bucky shuttles" could be envisioned from ordered close-packed nanotube arrays and such an assembly could potentially provide unprecedented memory storage (Figure 3.16d). This ideal system would offer a nonvolatile memory device with a combination of high switching speed and high density.

3.4.2 Memory Elements Based on the "Bucky Shuttle" Concept

While the "bucky shuttle" concept provides a relatively simple mechanism for storing information at the nanoscale, and it is strongly supported by molecular dynamics simulations that have solidified its theoretical operation, a practical measurement of the intrinsic on and off states is not clear. Also, the experimental production, although shown possible using self-assembly, is not able to exclusively produce the required "bucky shuttle" structures and would require novel separation and sorting techniques to facilitate their practical use. These shortcomings stimulated the development of a number of related concepts that have continued to evolve, such as a "nanotube shuttle" employing a transition from single-walled to double-walled nanotube for detection of the "shuttle" position (Kang and Hwang 2004c, 2005a), multi-endo-fullerenes- or peapod-based shuttles (Kang and Hwang 2004c, 2005b), and variations of these including boron nitride nanotubes (Choi et al. 2004, Hwang et al. 2005b), bipolar endo-fullerenes (Hwang et al. 2005a), metallofullerenes (Byun et al. 2004), and ionic fluidic media (Kang and Hwang 2004b). Again, the basic underlying concept relies on the electromigration of an encapsulated architecture(s) from one local energy minimum to another. The "nanotube shuttle," for example (see Figure 3.17), is based on the use of a capped peapod (nanotube with multiple endohedral C_{60}s, see Figure 3.17a), which is subjected to thermal and e-beam treatment (Figure 3.17b) that subsequently causes coalescence of the encapsulated C_{60}s to form an inner tubule. This is a well-known experimental process and one of the primary methods used for producing high-quality double-walled CNTs. The capped double-walled carbon nanotube (DWCNT) is then cut using lithography and etching technology in order to produce two aligned open ends with the inner tubule initially residing within one end (Figure 3.17c through e). The shuttle mechanism allows the inner tubule to move from one tube end across the gap into the other tube (Figure 3.17f). The separation of the two open ends is required in order to have two independent nanotube structures, which is critical for detecting the state that the overall system is in. The separation must be large enough to preempt strong interactions between the two nanotube structures (the two ends need to be independent in order for the electronic structure to allow unique detection) but small enough so as to prevent the shuttling inner tubule from completely escaping. For such an ideal system, the position of the inner tubule could in principle be measured by detecting the electronic structure of one of the ends since it will change from that corresponding to a single-walled to a double-walled nanotube (Lee et al. 2002).

3.4.3 Conclusions

The general "bucky shuttle" concept relies on utilizing the basic principle of electromigration in a two-level system defined by an ionized encapsulated fullerene-based structure. Its operation has been demonstrated to be very efficient (very short switching times) and is in general highly robust in regard to thermal and electrical stability. In addition, the possibility of assembling very high density architectures promises to allow construction of memory elements with high switching speed, high density, and nonvolatile storage of data. Unfortunately, despite the tremendous interest in the fundamentally simple and attractive systems underlying the use of CNT-based electromigration for memory elements, experimental realization has been difficult. This is because of the complexities in practical fabrication of the required structures and for making appropriate electrical contacts with them. As such, these types of systems have remained as conceptual ideas, to date, relying on the proof-of-principles results primarily obtained from classical molecular dynamics simulations.

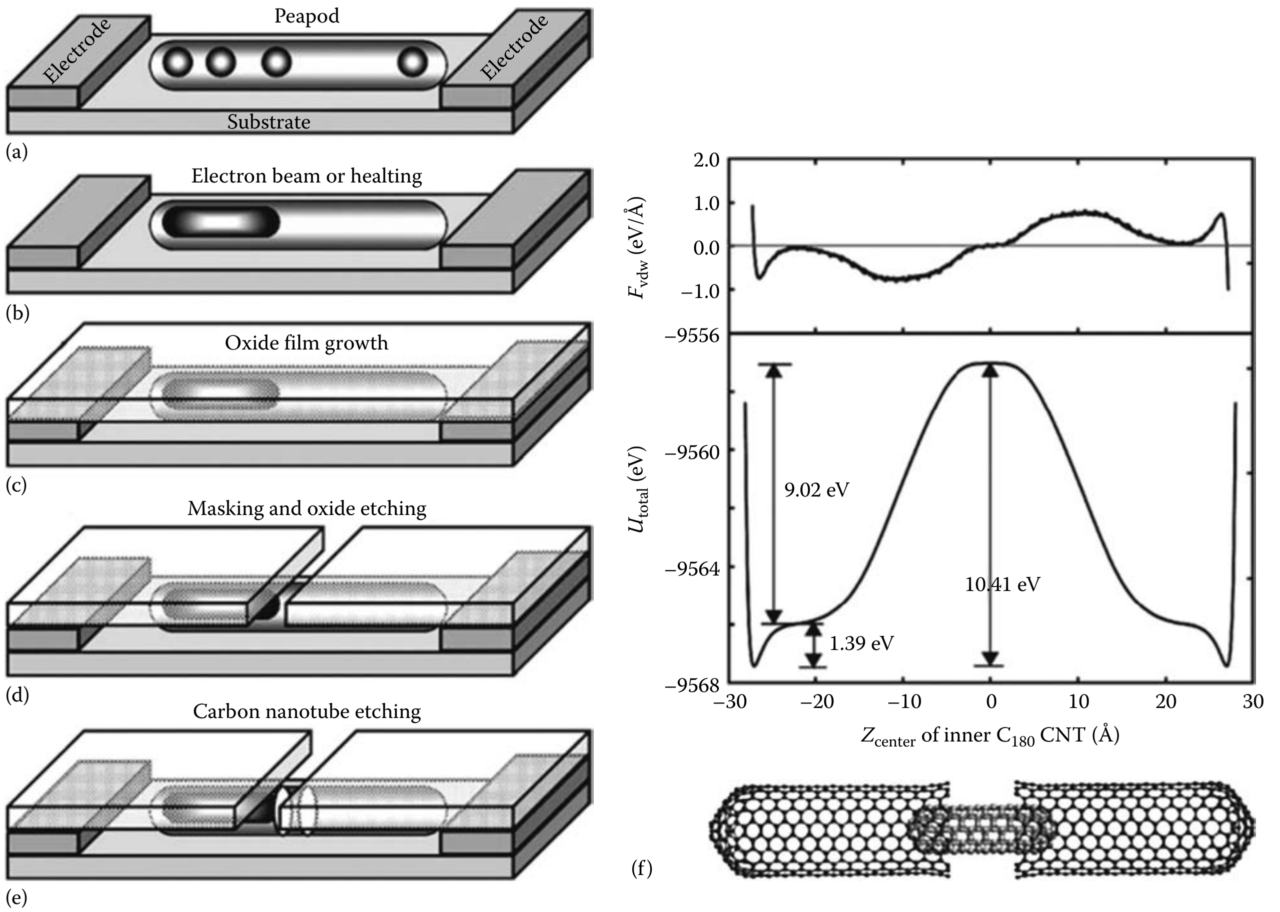

FIGURE 3.17 A schematic illustration of the design (a–e) and operation of a "nanotube shuttle" memory element (f): (a) a peapod CNT that can be coalesced by e-beams and thermal treatment. (b) the coalesced DWCNT; (c) oxide film growth; (d) Masking and oxide etching; (e) etching of the DWCNT; (f) The van der Waals force (F_{vdW}) and total energy (U_{total}) as a function of the center position of the inter-nanotubule. (Reprinted from Kang, J.W. and Hwang, H.J., *Carbon*, 42, 3018, 2004a. With permission.)

3.5 General Conclusions

At the nanoscale, where electronics is integrated on a lateral scale associated with a minimum feature size significantly smaller than tens of nanometers, a number of new phenomena stemming from quantum effects and confinement dominate the processes governing the electron flow across an active device (Waser 2003). It is when those quantum effects are understood, that new device concepts can be developed and novel mechanisms tuned to desired functionality. For practical applications, the role of fundamental research is to address the question of how a new material can be employed in useful devices. This is particularly important for nanoelectronics and information storage since any technology that has ambitions to compete with conventional approaches must display significant benefits and distinct advantages. As discussed in this chapter, CNTs are among the most cited materials that have been used for prototype nanoelectronics and information-storage devices, a dominant position that originates from their intrinsic structural and electronic properties. It is clear that numerous memory devices based on CNTs have been proposed and successfully demonstrated during the past decade. Many of these devices show tremendous promise for providing enhanced densities, lower power requirements, more efficient read–write processes, and nonvolatility of data. These advantages indicate that CNT-based memory could, in principle, compete with current memory devices such as DRAM. However, cycling stability and data-retention times that are competitive with flash memory has proven to be difficult to achieve. Other critical issues revolve around practical fabrication and the scalability for CNT-based memory devices that can provide the same facility for large-scale fabrication as silicon photolithography. Presently, this goal appears to be in the distant future but it continues to be a topic of intense research. With the demonstrated potential payoff for a range of useful applications, it seems likely that a solution to these problems will eventually be found.

Acknowledgments

This research was sponsored in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), by the Division of Materials Sciences and

Engineering, and by the Center of Nanophase Materials Science, U.S. Department of Energy under Contract No. DEAC05-00OR22725 with UT-Battelle, LLC at ORNL.

References

Akdim, B. and Pachter, R., Switching behavior in pi-conjugated molecules bridging nonmetallic electrodes: A density functional theory study, *Journal of Physical Chemistry C* 112(9), 3170–3174, 2008.

Axelsson, S., Campbell, E. E. B., Jonsson, L. M. et al., Theoretical and experimental investigations of three-terminal carbon nanotube relays, *New Journal of Physics* 7, 245, 2005.

Bachhofer, H., Reisinger, H., Bertagnolli, E., and von Philipsborn, H., Transient conduction in multidielectric silicon-oxide-nitride-oxide semiconductor structures, *Journal of Applied Physics* 89(5), 2791–2800, 2001.

Bernholc, J., Brenner, D., Nardelli, M. B., Meunier, V., and Roland, C., Mechanical and electrical properties of nanotubes, *Annual Review of Materials Research* 32, 347–375, 2002.

Bichoutskaia, E., Popov, A. M., and Lozovik, Y. E., Nanotube-based data storage devices, *Materials Today* 11(6), 38–43, 2008.

Byun, K. R., Kang, J. W., and Hwang, H. J., Nanoscale data storage device of metallofullerene peapods, *Journal of the Korean Physical Society* 45(2), 416–422, 2004.

Cha, S. N., Jang, J. E., Choi, Y. et al., Fabrication of a nanoelectromechanical switch using a suspended carbon nanotube, *Applied Physics Letters* 86(8), 083105-3, 2005.

Chakraborty, R. S., Narasimhan, S., and Bhunia, S., Hybridization of CMOS with CNT-based nano-electromechanical switch for low leakage and robust circuit design, *IEEE Transactions on Circuits and Systems I—Regular Papers* 54(11), 2480–2488, 2007.

Chakraborty, G., Sarkar, C. K., Lu, X. B., and Dai, J. Y., Study of the tunnelling initiated leakage current through the carbon nanotube embedded gate oxide in metal oxide semiconductor structures, *Nanotechnology* 19(25), 255401, 2008.

Charlier, J. C., Blase, X., and Roche, S., Electronic and transport properties of nanotubes, *Reviews of Modern Physics* 79(2), 677–732, 2007.

Choi, W. B., Chae, S., Bae, E. et al., Carbon-nanotube-based nonvolatile memory with oxide-nitride-oxide film and nanoscale channel, *Applied Physics Letters* 82(2), 275–277, 2003.

Choi, W. Y., Kang, J. W., and Hwang, H. J., Bucky shuttle memory system based on boron-nitride nanopeapod, *Physica E—Low-Dimensional Systems & Nanostructures* 23(1–2), 135–140, 2004.

Cui, J. B., Sordan, R., Burghard, M., and Kern, K., Carbon nanotube memory devices of high charge storage stability, *Applied Physics Letters* 81(17), 3260–3262, 2002.

Cumings, J. and Zettl, A., Low-friction nanoscale linear bearing realized from multiwall carbon nanotubes, *Science* 289(5479), 602–604, 2000.

Dequesnes, M., Rotkin, S. V., and Aluru, N. R., Calculation of pull-in voltages for carbon-nanotube-based nanoelectromechanical switches, *Nanotechnology* 13(1), 120–131, 2002.

Derycke, V., Martel, R., Appenzeller, J., and Avouris, P., Carbon nanotube inter- and intramolecular logic gates, *Nano Letters* 1(9), 453–456, 2001.

Derycke, V., Martel, R., Appenzeller, J., and Avouris, P., Controlling doping and carrier injection in carbon nanotube transistors, *Applied Physics Letters* 80(15), 2773–2775, 2002.

Deshpande, V. V., Chiu, H. Y., Postma, H. W. C. et al., Carbon nanotube linear bearing nanoswitches, *Nano Letters* 6(6), 1092–1095, 2006.

Duan, X. F., Huang, Y., and Lieber, C. M., Nonvolatile memory and programmable logic from molecule-gated nanowires, *Nano Letters* 2(5), 487–490, 2002.

Dujardin, E., Derycke, V., Goffman, M. F., Lefevre, R., and Bourgoin, J. P., Self-assembled switches based on electroactuated multiwalled nanotubes, *Applied Physics Letters* 87(19), 193107-3, 2005.

Forro, L., Nanotechnology: Beyond Gedanken experiments, *Science* 289(5479), 560–561, 2000.

Freitag, M., Johnson, A. T., Kalinin, S. V., and Bonnell, D. A., Role of single defects in electronic transport through carbon nanotube field-effect transistors, *Physical Review Letters* 89(21), 216801, 2002.

Fuhrer, M. S., Kim, B. M., Durkop, T., and Brintlinger, T., High-mobility nanotube transistor memory, *Nano Letters* 2(7), 755–759, 2002.

Fujita, S., Nomura, K., Abe, K., and Lee, T. H., 3-d nanoarchitectures with carbon nanotube mechanical switches for future on-chip network beyond CMOS architecture, *IEEE Transactions on Circuits and Systems I—Regular Papers* 54(11), 2472–2479, 2007.

Ganguly, U., Kan, E. C., and Zhang, Y., Carbon nanotube-based nonvolatile memory with charge storage in metal nanocrystals, *Applied Physics Letters* 87(4), 043108-3, 2005.

Goldsmith, B. R., Coroneus, J. G., Khalap, V. R. et al., Conductance-controlled point functionalization of single-walled carbon nanotubes, *Science* 315(5808), 77–81, 2007.

He, J. L., Chen, B., Flatt, A. K. et al., Metal-free silicon-molecule-nanotube testbed and memory device, *Nature Materials* 5(1), 63–68, 2006.

Hwang, H. J., Byun, K. R., Lee, J. Y., and Kang, J. W., A nanoscale field effect data storage of bipolar endo-fullerenes shuttle device, *Current Applied Physics* 5(6), 609–614, 2005a.

Hwang, H. J., Choi, W. Y., and Kang, J. W., Molecular dynamics simulations of nanomemory element based on boron-nitride nanotube-to-peapod transition, *Computational Materials Science* 33(1–3), 317–324, 2005b.

Jang, J. E., Cha, S. N., Choi, Y. et al., Nanoelectromechanical switches with vertically aligned carbon nanotubes, *Applied Physics Letters* 87(16), 163114-3, 2005.

Jang, J. E., Cha, S. N., Choi, Y. J. et al., Nanoscale memory cell based on a nanoelectromechanical switched capacitor, *Nature Nanotechnology* 3(1), 26–30, 2008.

Jonsson, L. M., Axelsson, S., Nord, T., Viefers, S., and Kinaret, J. M., High frequency properties of a CNT-based nanorelay, *Nanotechnology* 15(11), 1497–1502, 2004a.

Jonsson, L. M., Nord, T., Kinaret, J. M., and Viefers, S., Effects of surface forces and phonon dissipation in a three-terminal nanorelay, *Journal of Applied Physics* 96(1), 629–635, 2004b.

Kang, J. W. and Hwang, H. J., Carbon nanotube shuttle' memory device, *Carbon* 42, 3018–3021, 2004a.

Kang, J. W. and Hwang, H. J., Nano-memory-element applications of carbon nanocapsule encapsulating potassium ions: Molecular dynamics study, *Journal of the Korean Physical Society* 44(4), 879–883, 2004b.

Kang, J. W. and Hwang, H. J., Schematics and atomistic simulations of nanomemory element based on carbon tube-to-peapod transition, *Japanese Journal of Applied Physics Part 1—Regular Papers Short Notes & Review Papers* 43(7A), 4447–4452, 2004c.

Kang, J. W. and Hwang, H. J., Carbon nanotube shuttle memory device based on singlewall-to-doublewall carbon nanotube transition, *Computational Materials Science* 33(1–3), 338–345, 2005a.

Kang, J. W. and Hwang, H. J., Schematics and simulations of nanomemory device based on nanopeapods, *Materials Science & Engineering C—Biomimetic and Supramolecular Systems* 25(5–8), 843–847, 2005b.

Kang, J. W. and Jiang, Q., Electrostatically telescoping nanotube nonvolatile memory device, *Nanotechnology* 18(9), 095705, 2007.

Kang, J. W., Lee, J. H., Lee, H. J., and Hwang, H. J., A study on carbon nanotube bridge as a electromechanical memory device, *Physica E: Low-Dimensional Systems and Nanostructures* 27(3), 332–340, 2005.

Ke, C. H. and Espinosa, H. D., Feedback controlled nanocantilever device, *Applied Physics Letters* 85(4), 681–683, 2004.

Ke, C.-H. and Espinosa, H. D., Nanoelectromechanical systems and modeling, in *Handbook of Theoretical and Computational Nanotechnology* American Scientific Publishers, Valencia, CA, 2006a.

Ke, C. H. and Espinosa, H. D., In situ electron microscopy electromechanical characterization of a bistable NEMS device, *Small* 2(12), 1484–1489, 2006b.

Ke, C. H., Pugno, N., Peng, B., and Espinosa, H. D., Experiments and modeling of carbon nanotube-based NEMS devices, *Journal of the Mechanics and Physics of Solids* 53(6), 1314–1333, 2005.

Kim, P. and Lieber, C. M., Nanotube nanotweezers, *Science* 286(5447), 2148–2150, 1999.

Kim, W., Javey, A., Vermesh, O. et al., Hysteresis caused by water molecules in carbon nanotube field-effect transistors, *Nano Letters* 3(2), 193–198, 2003.

Kinaret, J. M., Nord, T., and Viefers, S., A carbon-nanotube-based nanorelay, *Applied Physics Letters* 82(8), 1287–1289, 2003.

Kwon, Y. K., Tomanek, D., and Iijima, S., "Bucky shuttle" memory device: Synthetic approach and molecular dynamics simulations, *Physical Review Letters* 82(7), 1470–1473, 1999.

Lee, J., Kim, H., Kahng, S.-J. et al., Bandgap modulation of carbon nanotubes by encapsulated metallofullerenes, *Nature* 415, 1005–1008, 2002.

Lee, S. W., Lee, D. S., Morjan, R. E. et al., A three-terminal carbon nanorelay, *Nano Letters* 4(10), 2027–2030, 2004.

Lu, X. B. and Dai, J. Y., Memory effects of carbon nanotubes as charge storage nodes for floating gate memory applications, *Applied Physics Letters* 88(11), 113104-3, 2006.

Mannik, J., Goldsmith, B. R., Kane, A., and Collins, P. G., Chemically induced conductance switching in carbon nanotube circuits, *Physical Review Letters* 97(1), 016601, 2006.

Martel, R., Schmidt, T., Shea, H. R., Hertel, T., and Avouris, P., Single- and multi-wall carbon nanotube field-effect transistors, *Applied Physics Letters* 73(17), 2447–2449, 1998.

Martel, R., Derycke, V., Lavoie, C. et al., Ambipolar electrical transport in semiconducting single-wall carbon nanotubes, *Physical Review Letters* 87(25), 256805, 2001.

Maslov, L., Concept of nonvolatile memory based on multiwall carbon nanotubes, *Nanotechnology* 17(10), 2475–2482, 2006.

Meunier, V. and Sumpter, B. G., Tuning the conductance of carbon nanotubes with encapsulated molecules, *Nanotechnology* 18(42), 424032, 2007.

Meunier, V., Kalinin, S. V., Shin, J., Baddorf, A. P., and Harrison, R. J., Quantitative analysis of electronic properties of carbon nanotubes by scanning probe microscopy: From atomic to mesoscopic length scales, *Physical Review Letters* 93(24), 246801, 2004.

Meunier, V., Kalinin, S. V., and Sumpter, B. G., Nonvolatile memory elements based on the intercalation of organic molecules inside carbon nanotubes, *Physical Review Letters* 98(5), 056401, 2007.

Radosavljevic, M., Freitag, M., Thadani, K. V., and Johnson, A. T., Nonvolatile molecular memory elements based on ambipolar nanotube field effect transistors, *Nano Letters* 2(7), 761–764, 2002.

Rueckes, T., Kim, K., Joselevich, E. et al., Carbon nanotube-based nonvolatile random access memory for molecular computing, *Science* 289(5476), 94–97, 2000.

Sakurai, T., Yoshimura, T., Akita, S., Fujimura, N., and Nakayama, Y., Single-wall carbon nanotube field effect transistors with non-volatile memory operation, *Japanese Journal of Applied Physics Part 2—Letters & Express Letters* 45(37–41), L1036–L1038, 2006.

Scott, J. C., Materials science: Is there an immortal memory? *Science* 304(5667), 62–63, 2004.

Star, A., Lu, Y., Bradley, K., and Gruner, G., Nanotube optoelectronic memory devices, *Nano Letters* 4(9), 1587–1591, 2004.

Tans, S. J., Verschueren, A. R. M., and Dekker, C., Room-temperature transistor based on a single carbon nanotube, *Nature* 393(6680), 49–52, 1998.

Ward, J. W., Meinhold, M., Segal, B. M. et al., A nonvolatile nanoelectromechanical memory element utilizing a fabric of carbon nanotubes, in *Non-Volatile Memory Technology Symposium*, Orlando, FL, 2004, pp. 34–38.

Waser, R., *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices*, Wiley-VCH, Weinheim, Germany, 2003.

Wunnicke, O., Gate capacitance of back-gated nanowire field-effect transistors, *Applied Physics Letters* 89(8), 083102, 2006.

Yang, D. J., Zhang, Q., Wang, S. G., and Zhong, G. F., Memory effects of carbon nanotube-based field effect transistors, *Diamond and Related Materials* 13(11–12), 1967–1970, 2004.

Yoneya, N., Tsukagoshi, K., and Aoyagi, Y., Charge transfer control by gate voltage in crossed nanotube junction, *Applied Physics Letters* 81(12), 2250–2252, 2002.

Yousif, M. Y. A., Lundgren, P., Ghavanini, F., Enoksson, P., and Bengtsson, S., CMOS considerations in nanoelectromechanical carbon nanotube-based switches, *Nanotechnology* 19(28), 285204, 2008.

Yu, M. F., Yakobson, B. I., and Ruoff, R. S., Controlled sliding and pullout of nested shells in individual multiwalled carbon nanotubes, *Journal of Physical Chemistry B* 104(37), 8764–8767, 2000.

Zhang, Y., Carbon nanotube based nonvolatile memory devices, *International Journal of High Speed Electronics and Systems* 16(4), 959–975, 2006.

4

Ferromagnetic Islands

Arndt Remhof
Swiss Federal Laboratories for Materials Testing and Research

Andreas Westphalen
Ruhr-Universität Bochum

Hartmut Zabel
Ruhr-Universität Bochum

4.1 Introduction **4**-1
4.2 Background **4**-1
Ferromagnetism • Ferromagnetic Domains • Hysteresis Loops • Micromagnetic Simulations • Experimental Techniques
4.3 State of the Art **4**-5
Sample Preparation • Rectangular and Elliptical Islands • Ferromagnetic Dots • Ferromagnetic Rings • Noncircular Rings • Ferromagnetic Spirals • Magnetostatic Interactions
4.4 Summary and Discussion **4**-14
References **4**-14

4.1 Introduction

A ferromagnetic island is a detached piece of ferromagnetic material, supported by a nonmagnetic* substrate. The term *ferromagnetic island* is usually used to describe a prolate object with a thickness of a few nanometers and lateral dimensions of up to a few micrometers. If the geometrical confinement becomes comparable or smaller than the intrinsic magnetic length scales, such as the exchange length, the domain size, or the domain wall thickness, novel magnetic properties and magnetization reversal mechanisms emerge, different from the properties of the respective bulk material. The domain configuration, the reversal mechanism, and the demagnetizing field of ferromagnetic islands can be tailored by their size, their shape, and their chemical composition.

Nanomagnetism is a rapidly growing area in solid-state research and development. There exist several excellent reviews covering the whole field, including preparation and characterization techniques (Dennis et al. 2002, Martín et al. 2003, Adeyeye and Singh 2008). We do not aim to give another comprehensive overview. This chapter is intended to give an introduction to ferromagnetic islands and to illustrate their properties and interactions on some of our own selected recent experimental results. First, we will present an overview of the fundamentals of ferromagnetism, particularly with regard to its impacts on nanostructured materials. Subsequent, examples of selected systems starting from single dots and rectangular-shaped islands to more complex geometries will be discussed. Preparation and characterization techniques are only covered if they are necessary for the understanding of the physical properties. This chapter concludes with an outlook on the future perspectives.

4.2 Background

4.2.1 Ferromagnetism

The discovery of magnetism is one of the oldest cultural achievements in human history. The attraction of iron to loadstone was observed before recorded history began. Magnetism arises from the magnetic dipole moments of atoms and molecules influenced by their symmetry and by the relative orientation of their electron orbits. Depending on the response of substances in an external magnetic field, they are classified as diamagnetic, paramagnetic, ferromagnetic, antiferromagnetic, or ferrimagnetic. The main characteristic of ferromagnetic materials is their permanent magnetization, caused by a natural tendency of the magnetic moments of its atoms or molecules to align under their mutual interactions. The ferromagnetism of the 3d transition metals Fe, Co, and Ni is essentially due to a quantum mechanical phenomenon which has its origin in electrostatics (Heisenberg 1928).

The permanent magnetic moment of a free atom of the 3d transition metals is caused by the uncompensated spin moments of partly filled 3d shells and the related orbital moments. The wave function of the 3d electrons can be separated into an orbital wave function and a spin wave function. Paulis exclusion principle forces the total wave function to be antisymmetric, so that no two electrons may have the same set of quantum numbers in an atom. Parallel alignment of the spins forces the occupation of higher energy levels, allowing for a larger separation of the electrons and a reduction of their Coulomb interaction. The win in Coulomb energy is confronted with an increasing kinetic energy, as some electrons have to occupy higher energy levels for the symmetric spin configuration. Thus, the developing of ferromagnetism depends on which energy term dominates. In the most simple case of a two-electron system, the energy difference between the parallel and the antiparallel spin alignment is expressed by the exchange integral J:

---

* Throughout the text the term "nonmagnetic" will be used to address nonferromagnetic materials.

$$J = 2\int \Big(\Psi_1^{*}(r_1)\Psi_2^{*}(r_2)V(r_1,r_2)\Psi_2(r_1)\Psi_2(r_2) \Big) dr_1 dr_2. \tag{4.1}$$

where

$V(r_1, r_2)$ denotes the interaction potential

$\Psi_1(r_1)$ and $\Psi_2(r_2)$ are the respective single-particle wave functions

A strong exchange interaction implies a large spatial overlap of the electron wave functions, which cannot be realized if these electrons are strictly localized. At least in the metals Fe, Co, and Ni, conduction electrons of the outer shell are not bound to their atoms, they are free to move and must interact in some way with the electrons at the lattice sites. In these metals the 3d band is overlapped in energy by a much wider 4s band. The bands are filled up to the Fermi level, so that the electrons which each atom contributes to the conduction band are not all from the 4s band, but partly from the 3d band. Therefore, the number of d electrons per atom contributing to the bulk magnetism is not an integral number (Aharoni 2000). In the bulk, the direction of the magnetization vector is defined by the crystalline symmetry. In materials with a large magnetic anisotropy there is a strong coupling between the spin and orbital angular moments within an atom. In addition, the atomic orbitals are generally nonspherical. Because of their shape, the orbits prefer to lie in certain crystallographic directions. The spin–orbit coupling then assures a preferred direction for the magnetization, called the easy direction. To rotate the magnetization away from the easy direction costs energy, the so-called anisotropy energy. The anisotropy energy depends on the lattice structure. In a perfect single crystal, the anisotropy axis is well defined macroscopically, while in a polycrystalline sample the anisotropy axis is still defined locally in each crystallite, but macroscopically it is averaged out due to the random orientation of the crystallites. In ferromagnetic islands the induced anisotropy of the geometrical shape plays a mayor role and will be discussed in more detail in the following sections. Anisotropies may also be induced by the interface to the substrate, the surface, or strains. A more detailed discussion of the relevant energy terms will be given in Section 4.2.4.

4.2.2 Ferromagnetic Domains

Most ferromagnetic materials show no resulting magnetic moment outside the material. The reason for this observation are domain structures (Hubert and Schäfer 1998). The equilibrium magnetization configuration is a result of the competition between the local effects of exchange and anisotropy with the nonlocal effect of the magnetostatic field (in the case without applied magnetic field). Exchange and anisotropy alone would favor a homogeneously magnetized sample with the magnetization along one of the system's easy axes. But this configuration creates a large, long-range dipolar stray field. The energy contained in the field can be reduced thoroughly by separation of the sample into so-called magnetic domains or Weiss regions, in which the magnetization is still orientated along one of the easy axes, and where the demagnetizing fields of the different domains compensate each other. Only at the boundary between the domains, a domain wall is created, where locally some exchange and anisotropy energy is invested. A domain structure is stable if the interplay between magnetic field energy and domain wall energy has reached an optimal value.

Inside a domain the homogeneous magnetization is parallel to an easy axis direction of the magnetic anisotropy. Thus, domain formation leads to a minimization of exchange and anisotropy energy. If an external magnetic field is applied to the sample, it will also bring energy into the system, the so-called Zeeman energy. The Zeeman energy favors domains parallelly aligned to the external field and leads to domain wall moving or rotation of the magnetization. In a soft magnetic material where no anisotropy exists, a continuous magnetic vector field is expected. But also in the absence of anisotropy, regular domain pattern develops due to the shape of the element. The reason for this is the requirement of a minimum demagnetizing field. The following conditions have to be fulfilled: the magnetic vector field has to be free of divergence, the field has to be aligned parallel to the edges and to the surface, and the vector field has to have a constant length. The conditions will be only achieved simultaneously, if line discontinuities are admitted (van den Berg 1986). These discontinuities are replaced by domain walls. Thus, the shape of the sample enforces a regular domain pattern in a magnetic film of no anisotropy. Van den Berg invented a geometrical procedure to construct this equilibrium domain pattern: the basic domain structure is the locus of centers of all circles inside the sample that touch the sample edge at least at two points. The most important result of van den Berg is that in most cases domains will form along the edges (closure domains) in order to minimize the demagnetizing field. Figure 4.1 schematically shows the reduction of the stray field by domain formation with the help of van den Berg's construction. Examples of the domain structure of a three-dimensional Fe island, discussed on the basis of van den Berg's construction, are provided by Hertel et al. (2005).

Inside a domain wall the magnetization changes its direction in numerous small steps (Bloch 1932). The cost of exchange energy for building up a domain wall decreases as the domain wall thickness increases. On the other hand a minimum value for the anisotropy energy is observed in thin domain walls as the magnetization direction inside the wall is different from the easy direction. In the equilibrium state, the domain wall thickness δ_W follows from the balance of the exchange energy J, and the anisotropy energy K_1:

$$\delta_W = \sqrt{\frac{JS^2\pi^2}{K_1 a}} = \pi\sqrt{\frac{A}{K_1}} \tag{4.2}$$

A is a material constant, the so-called exchange constant, and equals JS^2/a for a simple cubic lattice with lattice constant a.

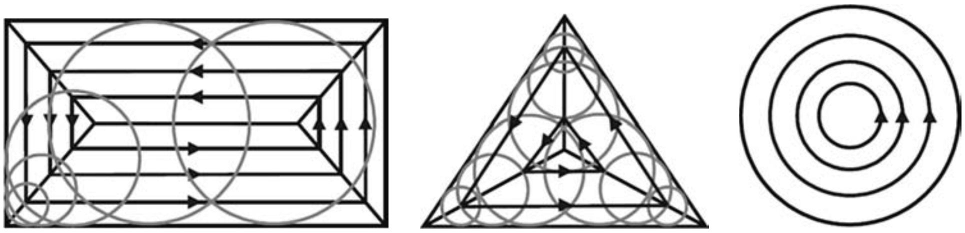

FIGURE 4.1 Van den Berg's construction on the example of ferromagnetic islands with simple geometric shapes.

The domain wall thickness can vary from a few nanometers in Co, to approximately 30 nm in Fe up to a few micrometers in extreme soft permalloy. If the domain wall thickness exceeds the lateral extension of the island, it will stay as a single domain. Larger islands will develop a domain structure if the gain in field energy overcompensates the domain wall energy. There is also a lower limit for the existence of a stable ferromagnetic single-domain island. As the anisotropy energy of an island scales with the volume of the island, there is a critical volume below which the anisotropy energy becomes comparable to the thermal energy k_BT. Below this *superparamagnetic limit*, the magnetization of the island is subject to thermal fluctuations. For magnetic data storage devices, the superparamagnetic limit marks the smallest possible bit size. A more detailed discussion of the relevant length scales in nanostructures is given by Martín et al. (2003) and Dennis et al. (2002).

4.2.3 Hysteresis Loops

The magnetic response of a ferromagnetic material to an external field is represented by a hysteresis loop in which the magnetization is plotted as a function of the magnetic field strength. The magnetization is a vector. Usually, the component parallel to the applied field (the x-component) is recorded. A hysteresis loop consists of two branches. They are called the ascending and descending branch, as it depicts the magnetization during field increase or field decrease. Generally, both branches do not overlap, i.e., the magnetization is not unambiguously defined on the field strength alone, it also depends on the magnetic history of the sample. The area enclosed by the two branches represents the magnetic energy stored in the system. A hysteresis loop shows three characteristic points: (1) The point where a further increase in magnetic field strength does not result in a further increase of the magnetization is called the saturation field. At this point all domains are aligned with the external field. (2) The persisting magnetization in zero field is called the remanent magnetization or the remanence. (3) The field needed to reduce the magnetization to zero is the coercive field or the coercivity. Coercivities may range from the mOe range for soft magnetic FeNi alloys to several 10 kOe for hard magnetic $SmCo_5$. Figure 4.2 shows two model hysteresis loops for a ferromagnetic island magnetized along the easy axis (left) and along the hard axis (right). The remagnetization along the easy axis usually is dominated by the rapid growth of oppositely oriented domains as soon as the applied field exceeds the nucleation field. In the case of elongated islands often domains nucleate at the two ends of the islands that grow by domain wall movement. The hysteresis loop is therefore characterized by a sudden change of the magnetization close to the coercive field and a high remanence, resulting in a square-shaped hysteresis loop. If on the other hand the field is applied along the hard axis, the remagnetization is realized by a coherent rotation of all the individual spins. The resulting hystereses loop therefore shows a continuous variation of the magnetization and a low remanence.

4.2.4 Micromagnetic Simulations

For the prediction of the expected domain configuration and the time-dependent development of the magnetization, micromagnetic computer simulations are used. A realistic approach for the description of the magnetic behavior of a magnetic material is to ignore the atomic nature of matter, to neglect quantum

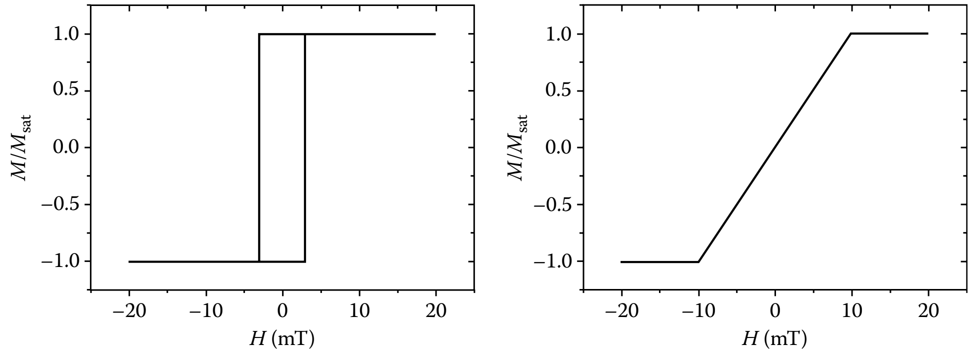

FIGURE 4.2 Model hysteresis loops for a single-domain island magnetized along the easy axis (left) and along the hard axis (right).

effects, and to use classical physics in a continuum description. Micromagnetism is the link between the atomistic, quantum mechanical description of the single atom on the one hand and the macroscopic view of a material characterized by the shape of its hysteresis curve on the other hand. Essentially, the magnetization is to be assumed as a continuous vector field $\vec{M}(\vec{r})$; thus, only the local average over the magnetic moments in a certain neighborhood instead of individual spins is taken into account. The history of micromagnetics starts from 1935 with the paper of L. Landau and E. Lifshitz on the structure of a wall between two antiparallel domains (Landau and Lifschitz 1935), and the work of W. F. Brown Jr. around 1940. A detailed treatment of micromagnetism is given by Brown (1963) and Prohl (2001). The energy will be formulated in terms of the continuous magnetization vector field, and this energy will be minimized in order to determine static magnetization structures. Thermodynamically, the system is described by the free energy $F = E - TS$ with the inner energy E of the system. There are several contributions to the inner energy:

- The exchange energy, responsible for the ferromagnetic coupling, is given by

  $$E_{\text{ex}} = -2\sum_{i<j} J_{ij}\vec{S}_i \cdot \vec{S}_j, \tag{4.3}$$

  where
    the coefficients J_{ij} are the exchange integrals
    the coefficients $\vec{S}_i$ are the atomic spins

  Positive J_{ij} lead to ferromagnetic order while negative J_{ij} cause antiferromagnetic order. The exchange energy is short-ranged and involves a summation over the nearest neighbors.
- The magnetostatic or demagnetization energy, which is the potential energy that the magnetic moments experience in their own magnetic field, can be written as

  $$E_{\text{demag}} = -\frac{\mu_0}{2}\int_V \vec{H}_{\text{d}} \cdot \vec{M} dV, \tag{4.4}$$

  where μ_0 is the magnetic constant. The magnetostatic or demagnetizing field, $\vec{H}_{\text{d}}$, is related to the magnetization via $\nabla \cdot \vec{H}_{\text{d}} = -\nabla \cdot \vec{M}$ as Maxwell's equation require that $\nabla \cdot \vec{B} = \mu_0 \nabla \cdot (\vec{H} + \vec{M}) = 0$. In the case of a homogenous magnetized ferromagnetic sample the demagnetizing energy can be expressed as

  $$E_{\text{demag}} = \frac{V}{2}\left(N_{11}M_x^2 + N_{22}M_y^2 + N_{33}M_z^2\right) \tag{4.5}$$

  where N_{ii} are the components of the so-called magnetometric demagnetizing tensor **N**. The components of the tensor are purely determined by the geometry.
- The magnetocrystalline anisotropy originates at the atomic level as discussed in Section 4.2.1. In a system with uniaxial anisotropy along the z direction, the leading energy terms are

  $$E_{\text{mca}} = VK_{u1}\sin^2\theta(1 + \sin^2\theta) \tag{4.6}$$

  where
    θ is the angle between the z direction and the magnetization
    K_{u1} is the anisotropy constant
    V is the volume of the sample

  A large positive anisotropy constant K_{u1} describes an easy axis, and a large negative constant K_{u1} an easy plane, perpendicular to the anisotropy axis. Uniaxial anisotropy occurs in hexagonal crystals such as Co, for example. Cubic anisotropy occurs for example in Fe and Ni, the leading energy terms are

  $$E_{\text{mca}} = VK_1\left(m_x^2m_y^2 + m_y^2m_z^2 + m_x^2m_z^2\right) + VK_2m_x^2m_y^2m_z^2 \tag{4.7}$$

  The anisotropy constants K_1 and K_2 and their ratio determine the easy axis of the system. Typical values for K_1 are in the order of 10^3–10^5 J/m^3. Most work on ferromagnetic islands has been either performed on polycrystalline samples, where the magnetocrystalline anisotropy is averaged out, or on materials with a negligible anisortopy. Therefore, contributions of magnetocrystalline anisotropy are not considered throughout this contribution.
- The surface anisotropy arises from the broken symmetry at the interface between the ferromagnetic and a nonmagnetic material. It can phenomenologically be described by the anisotropy constant K_{S}:

  $$E_{\text{S}} = \frac{1}{d}VK_{\text{S}}\sin^2\theta, \tag{4.8}$$

  where
    d denotes the thickness of the island
    θ denotes the angle between the magnetization and the surface normal

  K_{S} lies typically in the order of 10^{-4} to 10^{-3} J/m^2. The sign of K_{S} defines whether the easy axis is in-plane or out-of-plane. For the ferromagnetic islands discussed in this chapter, the surface anisotropy is negligible compared to the shape anisotropy, forcing in-plane magnetization.
- Finally the Zeeman energy is the energy of the magnetic moments in an external magnetic field H_{a}. It is given by

  $$E_{\text{Zeeman}} = -\mu_0 M_{\text{sat}} \int_V \vec{H}_{\text{a}} \cdot \vec{m} dV, \tag{4.9}$$

As experimentally the applied magnetic field $\vec{H}_\mathrm{a}$ is the control variable, it is more adapted to use Gibbs free energy or thermodynamic potential

$$G(T, \vec{H}_\mathrm{a}, \vec{M}) = F - \mu_0 \int \vec{H}_\mathrm{a} \cdot \vec{M}\, dV. \tag{4.10}$$

We assume that all experiments are performed at constant temperature, so that the entropy term can be neglected. Of course, it is necessary to take into account that the material constants are temperature-dependent. The system has the thermodynamic tendency to evolve toward the minimum of the potential G when approaching equilibrium. The metastable states of the ferromagnetic system are local minima of the potential G.

4.2.5 Experimental Techniques

To investigate the magnetic domain structure as well as the remagnetization process of micro- and nanostructured materials, a number of experimental techniques are available to the experimenter. On the one hand, magnetic domains can be imaged in real space by various techniques, an overview is given in Ref. Hopster and Oepen (2005): Kerr microscopy (Feldtkeller and Stein 1967), Lorentz microscopy (Chapman et al. 1994), scanning electron microscopy with polarization analysis (SEMPA) (Oepen and Kirschner 1991), scanning transmission x-ray microscopy (STXM) (Fischer et al. 1999), photo emission electron microscopy (PEEM) (Mundschau et al. 1996), spin-polarized low-energy electron microscopy (SPLEEM) (Bauer 1994), or magnetic force microscopy (MFM) (Martin and Wickramasinghe 1987). If real space images are obtained at different external magnetic field values, the hysteresis loop is derived by evaluating the total size of the magnetic domains having a particular direction of the magnetization vector with respect to the applied field. On the other hand, magnetic hysteresis loops can be measured with magneto-optical Kerr effect (MOKE) (Kerr 1877), micro MOKE (Allwood et al. 2003), superconducting quantum interference device (SQUID) magnetometry (Philo 1977), or vibrating sample magnetometry (VSM) (Foner 1959). Micro MOKE is an advanced magnetometry for the analysis of single magnetic nanostructures. In the case of the other techniques, a large number of identical elements are investigated in parallel in order to increase the resolution and accuracy. Therefore, the measured hysteresis loops yield information on the average remagnetization process of all elements and not on a single element. Thus, correlation effects between elements also become visible. The periodicity of an artificially structured array of ferromagnetic islands enables the use of diffraction techniques, such as Bragg-MOKE, where the Kerr effect measurements are performed at the diffraction spots from a lateral structure. Diffraction techniques suppress magnetic contributions of the substrate and contain information about the domain configuration inside the magnetic structures via the magnetic form factor (Remhof et al. 2007). Most experimental techniques can be extended to a vector-magnetometer: by applying a field perpendicular to the direction of measurement, the transverse (the y) component of the magnetization in the ferromagnetic structure can be investigated. From the measurements with the longitudinal and the transverse field orientation, the complete in-plane magnetization vector for the remagnetization can be derived (Westphalen et al. 2007). Using a photoelastic modulator and a digital lock-in amplifier for θ and 2θ response, all three components can be analyzed by measuring different polarizations of the incident beam (Vavassori 2000).

4.3 State of the Art

4.3.1 Sample Preparation

The preparation of sub-micrometer-sized ferromagnetic islands has benefited from the advancements in miniaturization techniques over the last decades. Usually optical lithography is used to define small patterns such as integrated circuits. Lithography, originally meaning "writing with stone," refers to a printing technique in which the imprint of an original pattern is reproduced on paper. In microtechnology it is used do describe a process, in which the desired pattern is first defined as a template on a mask, that can be transferred to the work piece. In optical lithography this is usually an opaque pattern on a transparent glass plate. An optical system is that used to downsize the pattern by projection onto a substrate coated with a photo-sensitive resist. The great advantage of this method is the high throughput. It is limited by the wavelength of the light and by the refractive power of the optical components. E-beam lithography (EBL) is a direct writing technique, in which the resist is exposed by a focused electron beam. In this case the mask only exists as a design file in the computer. The advantage is the high flexibility and the high resolution as compared to the optical methods. There are two kinds of resist: *positive* and *negative*. In the positive technique, first a resist template is produced by one of the aforementioned lithographic methods. Then the magnetic film is deposited onto the template. In a final chemical lift-off step only those magnetic islands remain, which have direct contact to the substrate or another metal layer. In the negative process a pattern is etched out of a continuous ferromagnetic film. Therefore, the homogeneous film is covered with an etch-resistant resist. After exposure, developing, and dry etching, only those magnetic islands that were exposed to the e-beam and covered with resist remain. Both methods are schematically illustrated in Figure 4.3. Apart from visible and UV light and electrons, other radiations such as x-rays and focused ion beams are well established in micro- and nanolithography. Also self-assembly and nanoimprint technologies are in use to prepare ferromagnetic islands. For a more detailed description of the various lithographic techniques we refer to Waser (2003), Bucknall (2005), and Suzuki and Smith (2005).

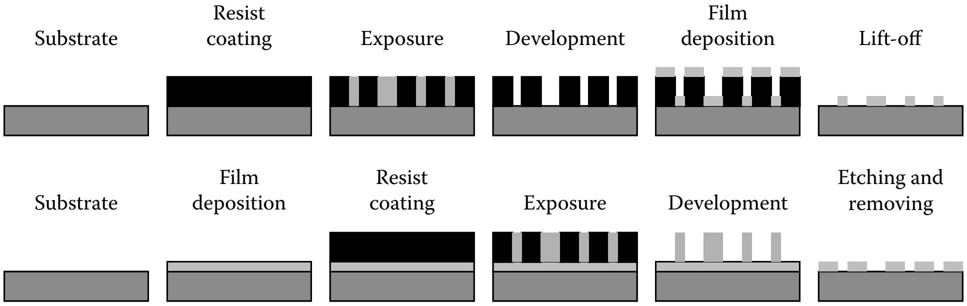

FIGURE 4.3 Patterning steps using positive resist (top panel) and negative resist (bottom panel).

4.3.2 Rectangular and Elliptical Islands

The simple geometry of rectangular and elliptical islands, featuring a well-defined and pronounced uniaxial shape anisotropy, is predestined as model systems. Especially in islands composed of materials with little or no crystalline anisotropy or in polycrystalline systems, the shape anisotropy becomes dominant. In elongated islands the shape anisotropy usually forces the easy axis along the major axis, while the hard magnetic axis lies along the minor axis. The strength of the anisotropy is thereby governed by the aspect ratio (Martín et al. 2003).

At remanence small rectangular or elliptic small islands are in a single-domain state with pronounced magnetic poles at the opposite ends of the island. MFM images these magnetic poles as areas of maximum contrast as depicted in the right panel of Figure 4.4. The left panel shows a secondary electron micrograph, recorded by scanning electron microscopy (SEM) of an array of 50 nm thick, 3 μm × 300 nm-sized permalloy islands placed on a square grid of 3 μm × 3 μm on a silicon wafer. While the SEM image shows the topological contrast, the MFM images the strength of the stay field (i.e., the density of the field lines) normal to the sample. Therefore, the poles at the end, where the stray field enters and leaves the island can, respectively, be identified as black and white spots.

Larger islands will decay into domains at remanence. The terms *small* and *larger* depend on the material's properties and the aspect ratio, determining the anisotropy. Last et al. (2004) have investigated polycrystalline permalloy (Py) micro- and nanostructures (thickness: 38 and 50 nm, length: 0.2–60 μm, width: 0.2–30 μm) in order to map the multi-domain to single-domain transition with MFM. They found a single-domain state, if the aspect ratio (length:width) is greater than 20:1; the multi-domain state is visible inside the rectangles with a aspect ratio between 1:1 and 3:1. Between these two cases they observed a transition state with a featureless inner region and fan-type closure domains at both ends of the rectangle. Figure 4.5 also shows that for very small particles, where the geometric extensions approach the domain wall width, smaller aspect ratios are sufficient to obtain a single-domain state. Extreme aspect ratios lead to *ferromagnetic nanowires*. In the case of thin wires the magnetization reversal process is initiated by the nucleation of oppositely oriented domains at the two ends of the wire and the formation of 180° domain walls. During the reversal process these domains grow by domain wall motion, until the walls meet in the middle of the film where they annihilate. Thicker wires show a more complex reversal mechanism, involving perpendicular in-plane domains that fill up the inner part of the wire and 180° as well as 90° domain walls are formed (Hausmanns et al. 2002). The domain walls can be manipulated (Tsoi et al. 2003) and used to perform logic operations (Allwood et al. 2002) and they can be employed for data storage (Parkin et al. 2008).

Returning to the original discussion of rectangular and elliptic islands and their dipolar behavior in remanence, the pronounced stray field of elongated micrometer-sized islands as discussed before are ideal to study the mutual magnetostatic interactions as a function of their separation.

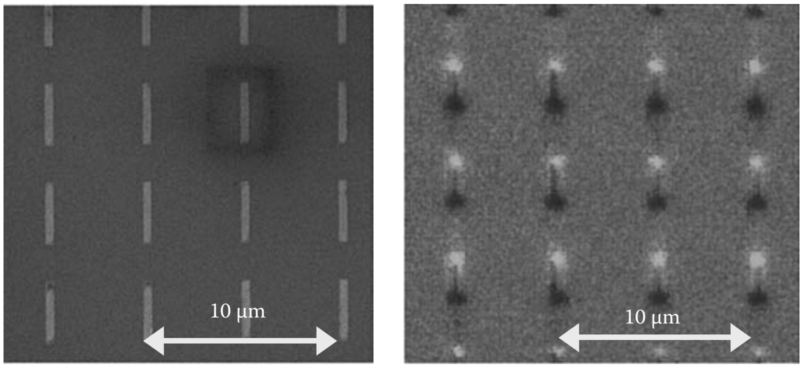

FIGURE 4.4 Micrographs of an array of 50 nm thick, 3 μm × 300 nm-sized permalloy islands placed on a square grid of 3 μm × 3 μm on a silicon wafer. The left panel shows a secondary electron micrograph, the right panel an magnetic force microscopy image.

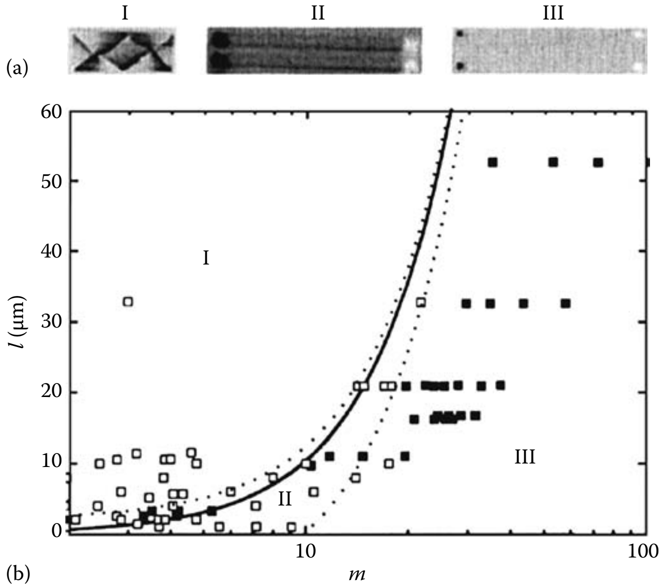

FIGURE 4.5 (a) Magnetic force microscopy images of Py structures in remanence with (I) low-aspect ratio (m = 2.5, with l = 10 μm, t = 38 nm), (II) medium-aspect ratio (m ≈ 14, l = 20.9 μm, t = 38 nm), and (III) high-aspect ratio (m = 57 (top wire), m = 43 (bottom wire), t = 50 nm). (b) Experimentally observed domain states of Py microstructures and nanostructures for varied length l vs. aspect ratios m. Open squares indicate multidomain states (I), gray squares denote high remanent states (II), and closed squares indicate a single-domain state (III). The black solid line depicts the transition line between multidomain and single-domain configuration from a phenomenological model. The dotted lines are guides to the eye indicating transitions between regions I, II, and III. (Reused from Last, T. et al., *J. Appl. Phys.*, 96, 6706, 2004. With permission.)

4.3.3 Ferromagnetic Dots

For technical applications such as the magnetic data recording, each island has to be addressed individually. Therefore, interactions between the islands have to be avoided. This can be done either by increasing the distances between the islands, or by avoiding strong and extended stray fields. Due to their geometric shape, circular dots and rings of certain aspect ratio exhibit a flux-closed state known as vortex state (Cowburn et al. 1999, Raabe et al. 2000, Shinjo et al. 2000, Fraerman et al. 2002). In this case a curling magnetic structure is established, in which the spin directions only change gradually, minimizing the loss in exchange energy. This configuration suppresses the occurrence of a net in-plane magnetization (zero remanence) without the need to form magnetic domains and to build domain walls. Close to the dot center, the increasing angle between adjacent spins forces them out-of-plane, so that they can align parallel to each other along the plane normal. Therefore, in the center of the vortex a spot of perpendicular magnetization exists, first observed by Shinjo et al. (2000) in circular permalloy dots with diameters ranging from 0.3 to 1 μm and a thickness of 50 nm. MFM images obtained by Shinjo et al. are represented in Figure 4.6.

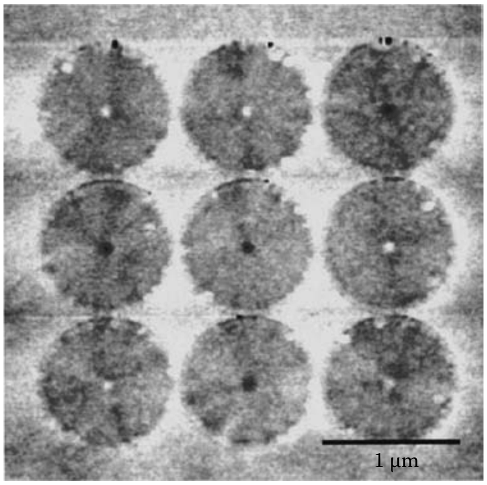

FIGURE 4.6 MFM image of an array of permalloy dots, 1 μm in diameter and 50 nm thick. (From Shinjo, T. et al., *Science*, 289, 930, 2000. With permission.)

Let us discuss the vortex state in a single permalloy dot of diameter 0.8 μm as an example for micromagnetic simulation, using the OOMMF 1.2a3 software (Donahue and Porter 1999) with a saturation magnetization M_s = 860 × 10^3 A/m, exchange stiffness A = 13 × 10^{-12} J/m, anisotropy constant K_1 = 0, and a damping coefficient α = 0.5. Figure 4.7 shows the result of the micromagnetic simulation.

As it is typical for an ideal vortex configuration, the y component of the magnetization is zero. Starting from negative saturation field (1) to point (2) the core of the vortex develops abruptly at the so-called *nucleation field*. In the remanent state (3) the core has moved into the center of the circular dot. This state lowers the system energy by reducing stray fields and hence lowering magnetostatic energy. Increasing the field then deforms the vortex, and its core moves away from the center of the dot to the upper edge (4). Finally it becomes unstable and again abruptly the core vanishes at the *annihilation field* (step in the hysteresis loop between (4) and (5) at approximately 90 Oe). Note that the nucleation and annihilation fields are different. The resulting minor loops are the main characteristics of the vortex state. Also note that the system is degenerate: the vortex may nucleate at the bottom and move to the top, or vice versa nucleate at the top and propagate to the bottom.

The geometrical conditions which favor the occurrence of the vortex state were studied by Cowburn at al. (1999) on the example of Supermalloy ($Ni_{80}Fe_{15}Mo_5$) dots with diameters ranging between 55 and 500 nm and a thickness range between 6 and 15 nm. The larger dots exhibit the remagnetization behavior of a vortex as described above, while the smaller dots show a more square-like hysteresis loop, a high remanence ($M_{rem}/M_{sat} \approx 0.8$) and a small coercive field, as depicted in Figure 4.8. This is characteristic of a uniform single-domain behavior. In agreement with micromagnetic simulations, the critical diameter drops with increasing thickness from 200 nm for 6 nm thick dots to 100 nm for 15 nm thick ones.

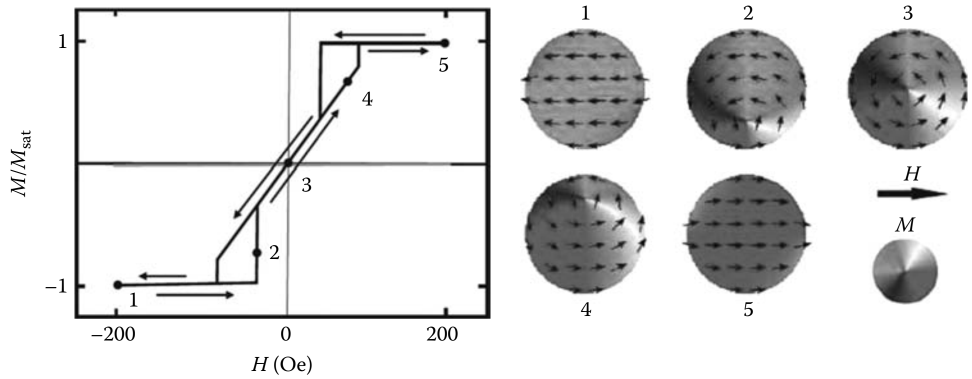

FIGURE 4.7 Hysteresis loop for the x (black line) and y component (gray line) of the magnetization, and some selected calculated magnetization profiles.

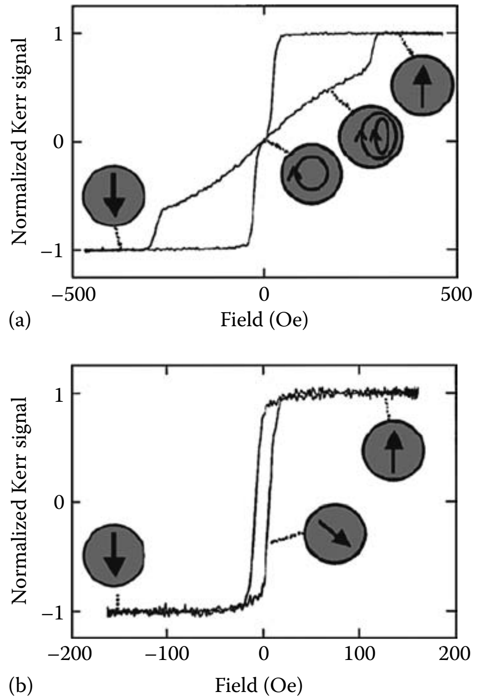

FIGURE 4.8 Hysteresis loops measured from nanomagnets of diameter d and thickness t: (a) $d = 300$ nm, $t = 10$ nm; (b) $d = 100$ nm, $t = 10$ nm. The schematic annotation shows the magnetization within a circular nanomagnet, assuming a field oriented up the page. (Reprinted from Cowburn, R.P. et al., *Phys. Rev. Lett.*, 83, 1042, 1999. With permission.)

The magnetic vortex is characterized by its rotational sense (chirality) which can be either clockwise or anticlockwise and the out-of-plane magnetization of the vortex core (polarity) pointing either upwards or downwards. Combining the chirality and the polarity, four different ground states can be realized. Often MOKE and Hall bar measurements are employed to determine the hysteresis of the magnetization reversal of circular magnetic disks (Cowburn et al. 1999, Rahm et al. 2003). The measured hysteresis has the features characteristic for the reversal process occurring through a vortex state, i.e., a vanishing remanence and closed lobes on either side of the remanence with the nucleation field and the annihilation field. However, these measurements are neither sensitive to the chirality nor to the polarity of magnetic vortices. Using MFM (Raabe et al. 2000, Shinjo et al. 2000), the polarity can be determined but not the chirality, whereas Lorentz microscopy operated in the Fresnel mode is sensitive to the chirality but not to the polarity (Raabe et al. 2000, Schneider et al. 2000).

Longitudinal Bragg MOKE on the other hand is sensitive to the average vortex chirality in a periodic array of magnetic islands. In case of longitudinal MOKE (L-MOKE) the measured variable, i.e., the Kerr rotation θ_x is proportional to the component of the magnetization vector $\vec{m} = (m_x, m_y)$ projected onto the intersection between the plane of incidence and the plane of the sample surface: $\theta_x \propto m_x$ (longitudinal component). For patterned samples such as an array of islands, the Kerr angle can be measured not only for the specularly reflected beam but also at the diffraction spots whose pattern represents the Fourier transform of the two-dimensional lattice structure of the real pattern. The Kerr rotation θ_x^{hk} measured for the diffraction orders h and k, corresponds to the magnetic form factor f_x^{hk}, i.e., the Fourier transform of the magnetization distribution within a single element of the array

$$f_x^{hk} \propto \int dS\,\vec{m}_x(\vec{r})\exp\left(i\vec{G}\,\vec{r}\right) \tag{4.11}$$

where

$\vec{G} = 2\pi(h\vec{e}_x + k\vec{e}_y)/a$ is the reciprocal lattice vector of the dot array
a is the periodicity of the dot array
the integral is taken over the area of a single dot

The relation between the magnetic form factor f_x^{hk} and the corresponding Kerr rotation θ_x^{hk} can be expressed as follows:

$$\theta_x^{hk} = c_{hk}(\phi_{\text{i}})\left(a_{hk}\Re\left(f_{\text{m}}\right) + b_{hk}\Im\left(f_{\text{m}}\right)\right) \tag{4.12}$$

where ϕ_i is the angle of incidence,

$$c_{hk}(\phi_i) = \left(h\frac{\lambda}{a} - \sin(\phi_i)\right) \Bigg/ \left(1 - \left(k\frac{\lambda}{a}\right)^2\right). \tag{4.13}$$

and a_{hk} and b_{hk} are coefficients describing the extent of the real and the imaginary contribution, respectively, of the magnetic form factor to the Bragg MOKE signal. They depend on a, ϕ_i, λ, the (magneto-) optical constants of the dots and the substrate, the shape, and the thickness of the dots. With the complete knowledge of all these parameters it is possible to determine the coefficients a_{hk} and b_{hk} (Lee et al. 2008). However, since the computation is rather complicated and cumbersome, they are usually treated here as adjustable parameters.

Grimsditch et al. concluded from Bragg MOKE measurements in the transverse geometry that the vortex chirality manifests itself in the sign of the imaginary part $\Im(f_m^{hk})$ (Grimsditch et al. 2002). This effect can also be applied in the longitudinal geometry. It can be shown that for a magnetization reversal process occurring through a vortex state, $\Im(f_m^{hk})$ always possesses the global extremum at the zero point of the external magnetic field and vanishes as the sample magnetization approaches the saturation level (Grimsditch et al. 2002, Grimsditch and Vavassori 2004). Hence, this imaginary contribution is expected to give rise to a splitting of the ascending branch and the descending branch of the hysteresis loop. For an array of perfectly circular disks a random distribution of left- and right-handed chiralities is expected. In this case the splitting of the hysteresis at remanence will vanish. Since the extent of the splitting corresponds to the degree of the imbalance between clockwise and anticlockwise vortices in magnetic islands, the Bragg MOKE technique provides a statistical means of measuring the chirality distribution in an array of magnetic islands. For a perfectly circular dot there are equal chances for the occurrence of either rotational senses. The vortex chirality can be controlled by breaking the circular symmetry (Lee et al. 2008). This can be most easily realized by shaping the dot with a flat edge at the top as depicted in Figure 4.9.

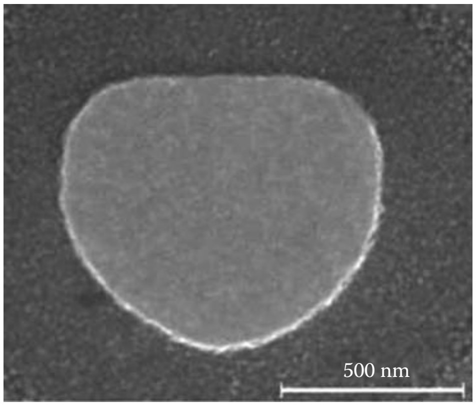

FIGURE 4.9 Scanning electron micrograph of a flat dot. (Reused from Lee, M.-S. et al., *J. Appl. Phys.*, 103, 093913, 2008. With permission.)

When a magnetic field is applied in the direction parallel to the flat edge (along the x-axis), the vortex nucleation always starts at the top close to the flat edge with a negative chirality (counterclockwise). With decreasing field strength the core moves down to the bottom perpendicularly to the field direction where it is annihilated. Micromagnetic simulation confirm that along the ascending branch the rotational sense is clockwise, whereas in the descending branch it is counterclockwise in agreement with the experimental observation (Lee et al. 2008).

4.3.4 Ferromagnetic Rings

Within a disk-shaped island, the vortex is only stable for diameters above approximately 100 nm, depending on the thickness of the disk. The vortex state can be stabilized by removing the vortex core in a ring-shaped element. Apart from the low remanent vortex state, which is the ground state of the system, ferromagnetic rings exhibit another, metastable magnetic configuration, the so-called *onion state* (Rothman et al. 2001). The onion state is characterized by the formation of two equally sized domains as shown in Figure 4.10. Due to the shape anisotropy, the magnetization in both domains follows the perimeter of the ring. The onion state allows the majority of the spins to be aligned with the applied magnetic field during the decrease of the magnetization from saturation to zero. Magnetization reversal is achieved by domain wall movement. The domain wall movement starts as soon as the applied reverse field exceeds the pinning strength of the walls. For perfectly symmetric rings, the pinning strength for both walls will be identical and both walls will move synchronous. Consequently, the ring will switch from the onion state to the reverse onion state. In a real system, however, shape irregularities and defects will give rise to differences in the pinning strength, which allows the wall that moves first to annihilate with the second wall, forming the vortex state. Therefore, specific shape irregularities in ferromagnetic rings allow to tailor their reversal behavior, to stabilize the vortex state, and to determine its chirality (Kläui et al. 2001, Vaz et al. 2007).

4.3.5 Noncircular Rings

Closed flux lines in the vortex state or the dipolar onion state are not restricted to circular-shaped objects. They can also be

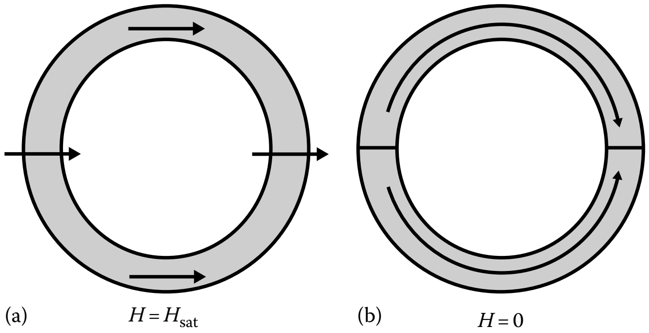

FIGURE 4.10 Schematic representation of the magnetization of a ring-shaped island. (a) In saturation by an applied field and (b) in remanence.

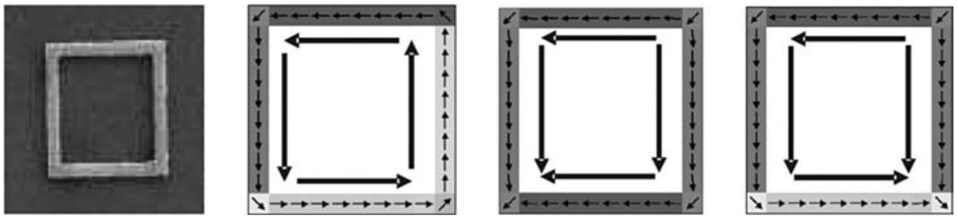

FIGURE 4.11 From left to right: Scanning electron micrograph of a square ring and sketches of the three different symmetrical ground states: vortex, onion, and horseshoe.

observed in square rings (or closed frames), as displayed in Figure 4.11. The shape anisotropy always forces the magnetization parallel to the sides of the frame. In the vortex state the magnetic moments form a closed ring; no stray field is expected to be observed outside the frame. The onion state consists of magnetically parallel-aligned sides of the frame, resulting in two magnetic poles in the opposite corners of the frames. The last panel of Figure 4.11 depicts a remanent state that is not observed in rings, the so-called *horseshoe state*. This state is formed by two magnetically parallel-aligned sides and by two antiparallel-aligned sites. Consequently two magnetic poles occur on one branch of the frame.

The remanent state is not only determined by the size and shape of the island, it can also be influenced by the direction of the external field during the magnetization reversal process. As long as the external field is applied in-plane, only the onion state is observed in remanence regardless of the azimuthal angle (Vavassori et al. 2003). However, if the field is applied perpendicular to the sample surface, the vortex state is the favorite remanent state. At intermediate angles (around 70°) the horseshoe state coexists with the onion state. MFM images of the different remanent states, together with an indication of the field direction, are presented in Figure 4.12.

Noncircular rings also allow the study of different symmetries. Triangular structures (Westphalen et al. 2006) have a sixfold magnetic shape anisotropy. In contrast to a square ring, where at the same time two sides are orientated in the hard axis direction and the other two sides point in the easy axis direction if the external magnetic field is orientated parallel to these sides, in a triangular ring structure two sides are always in an intermediate position if the third side is orientated parallel to the magnetic field, i.e., in the easy direction. This will lead to a more complex domain structure and to more possibilities for the remagnetization process according to the orientation of the structures to the applied magnetic field. By choosing different sizes for the same structure the shape anisotropy also affects the remagnetization process because it is more pronounced in the small triangular rings. Rings and islands have received most attention, because these elements show a vortex state in remanence which leads to a reduced stray field energy. Such elements can be stacked close together without disturbing each other, unlike elements with dipole character.

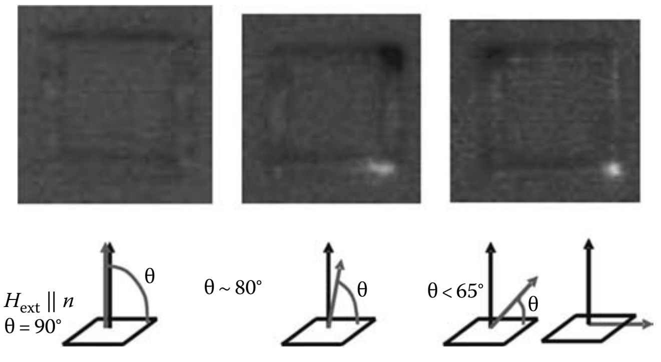

FIGURE 4.12 MFM images of square rings recorded in remanence after saturation in an external field. The field direction is indicated below each micrograph.

4.3.6 Ferromagnetic Spirals

Spirals are hybrids between circular islands and magnetic dipoles. In contrast to ring structures a closed magnetic flux configuration cannot be realized in the spirals during the remagnetization process. In fact they act as magnetic dipoles at the rim while the center is screened from the environment. On the other hand, a spiral is from a geometrical point of view a vortex. Therefore, spirals allow to investigate the interplay between a geometric and a magnetic vortex and to answer the question whether the geometry influences the magnetization: does the shape of a spiral lead to a vortex state in the remanent magnetization? In such a case the outer end piece of a long spiral would act as a magnetic monopole, because in the near region of the outer end piece the interaction with the second magnetic pole at the inner end piece could be neglected. The second question addressed by magnetic spirals is whether the core of the spiral is magnetically harder or softer than the outer parts and how the coercivity is affected by stray field interaction between the spirals. By choosing Fe as the material from which the spirals are fabricated a strong domain formation during the magnetization reversal is expected.

In order to address the aforementioned questions various quadratic arrays of polycrystalline Fe spirals (maximum radius = 2.8 μm, linewidth = 100 nm) were realized (Westphalen et al. 2008a). The polycrystallinity of the Fe suppresses the intrinsic magnetocrystalline anisotropy. In the first case (S1) only one spiral is set on the square grid. In a second sample (S2), it consists of two spirals facing each other without any connection while they are connected in the third sample (S3). Micrographs of the resulting pattern are shown in Figure 4.13.

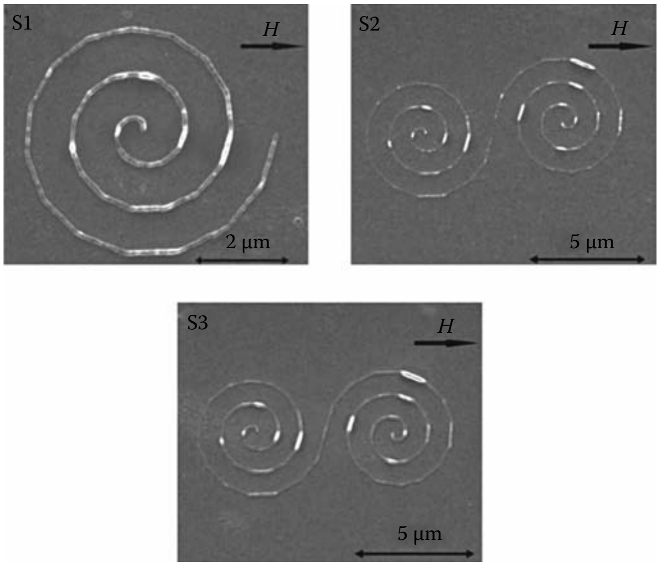

FIGURE 4.13 Scanning electron micrograph images of the four Fe spiral structures used in the experiments. The different spiral structures are named S1, S2, and S3. (Reused from Westphalen, A. et al., *J. Appl. Phys*., 104, 013906, 2008. With permission.)

The hysteresis loops are similar for all four patterns; however, strong variations occur in the coercive field values. S1 exhibits a coercive field of 240 Oe, while S2 and S3 exhibit coercive fields of 305 and 190 Oe, respectively. It is therefore possible to tailor the coercivity without disturbing the overall magnetization process.

The variations in coercivity are not only due to differences and inhomogeneities in the shape of the spiral structures. Micromagnetic simulations of ideal spiral structures, i.e., spirals without kinks and with constant linewidth, show the same behavior.

One important aspect which is obtained from the micromagnetic simulation is the observation that the remanent state of all spiral patterns is characterized by an onion-like magnetization configuration. The spirals are divided by a horizontal axis parallel to the applied field into an upper and lower part. In each part the magnetization follows the curvature of the spiral forming poles at the horizontal axis. Additionally, micromagnetic simulations of spiral structures with a diameter one order of magnitude smaller than for the ones considered so far still show a preference for the onion state. This behavior is different from closed ring structures, where a vortex state forms spontaneously during the magnetization reversal (Rothman et al. 2001, Castaño et al. 2003, Kläui et al. 2003). In order to generate a magnetic vortex in a spiral structure (geometric vortex), a staggered magnetic field is required that changes orientation after each half turn.

In general the inner parts of the spiral are more stable and switch later than the outer ones. In pattern S2 where the end pieces of two spirals face each other through a small gap, dipolar stray fields stabilize the magnetic configuration and the outer parts switch later than the inner parts. In contrast, joining two spirals by a magnetic bridge in pattern S3 results in a decrease of the coercive field, as this bridge allows domain walls to propagate across. Thus, the boundary condition for the outer pieces of the spiral controls the coercivity. Vice versa, using the same spiral parameters but arranging them in different ways the coercivity can be tailored. In fact the spiral structures have some similarity to the exchange spring heterostructures consisting of magnetically hard and soft materials, one providing the anisotropy and the other the magnetic moment (Fullerton et al. 1999). In comparison, the core of the spiral represents the hard part, whereas the outer tail of the spiral is either soft or can be hardened by interaction.

4.3.7 Magnetostatic Interactions

A key issue in fundamental physics and data storage technology is to understand the remagnetization behavior of submicron-sized magnetic elements and to functionalize their mutual interactions. Recently, the possibility to build logic devices such as AND and NOR gatters on the basis of elongated, dipolar-coupled permalloy (Py) has been demonstrated (Imre et al. 2006). Due to the demagnetizing field of elongated magnetic islands the remanent magnetic state and the reversal mechanism become sensitive to the shape of the elements. In well-ordered artificially structured nanomagnetic dipoles, the interactions arising from the stray field lead to highly correlated systems. In particular, geometric frustration leads to the spin ice state, in which the individual dipoles mimic the frustration of hydrogen ions in frozen water. In the ice structure, two hydrogen atoms are close to the oxygen atoms and two further away. This rule is usually quoted as the "two in–two out" ice rule. Figuratively speaking, the magnetic dipoles on the square lattice obey the spin ice rule, if two spins or magnetic dipoles point into a vertex and two point out. To realize this state, a two-dimensional square lattice of 25 nm thick Py islands (80 nm × 220 nm) was carefully demagnetized and visualized by means of MFM (Wang et al. 2006). The spin ice state is one of many possible magnetic superstructures which may appear in nanomagnetic arrays. Apart from the shape of the elements the symmetry of the superstructure depends on the relative distances between the particles. If the interparticle distance is comparable with or smaller than the length of an element, higher-order magnetostatic contributions become important (Vedmedenko et al. 2005). These contributions might induce additional anisotropy and/or select specific vertex configurations from the manifold of the spin ice symmetries. This, in turn, might influence the magnetization reversal in magnetic arrays.

The open frame system consisting of two perpendicularly crossed stripe arrays is a suitable prototype of interacting islands (Remhof et al. 2008). The open frame system is inherently frustrated. Each vertex where the two sublattices intersect has to accommodate four magnetic poles in close proximity. In the most unfavorable case, four equal magnetic poles meet in a vertex, i.e., the four magnetic moments point all inward or all outward. Energetically more favorable is a configuration in which two moments point inward and two point outward. A situation in which each magnetic pole is surrounded only by opposite poles is impossible. Analogous to the steric frustration of hydrogen ions in frozen water, the geometric frustration of spins in a magnetic material has been denoted as spin ice. The spin ice state is a demagnetized state. It is characterized by a lack of long-range order and by the so-called ice rules which require a minimization of the spin–spin interaction energy when two spins point inward and two spins point outward on each vertex. Model calculations based on the magnetostatic interactions including multipolar contributions clearly show that the spin ice state is indeed the demagnetized ground state of the system. At remanence, however, the samples exhibit a persistent net magnetization, indicative of the existence of a metastable, ordered state. The possible symmetric remanent states are shown schematically in Figure 4.14. The onion state, in which each of the two sublattices is parallel-aligned, has the highest symmetry and the highest magnetization. The resulting net magnetization of the whole array is tilted 45° against the long axes of the individual islands. At each vertex, the magnetization of two adjacent elements from different sublattices point inward, while the two others point outward. In the so-called horseshoe state only one of the two sublattices is parallel-aligned, while the other is antiparallel-ordered. Only the parallel-aligned sublattice contributes to the net magnetization, which consequently aligns with the anisotropy axis of the respective islands. There are two possible configurations in the horseshoe state. In the first one, in the following referred to as horseshoe 1, at each vertex the magnetization of two adjacent elements from different sublattices points inward, while the two others point outward. In the second one, in the following denoted as horseshoe 2, at each vertex the magnetization of three elements points inward (or outward), while only one points outward (inward). Experimentally, mainly the horseshoe 2 state is observed.

The micro-vortex state (Vedmedenko 2007) exhibits no net magnetic moment. In order to realize the micro-vortex state, the chirality of two adjacent vortices has to be antiparallel. Also in this state, the magnetization of two elements point inward and two outward. However, opposite to the aforementioned states, this time the magnetization of the two next nearest elements, i.e., from the same sublattice, points inward, while the two perpendicular elements from the other sublattice point outward, or vice versa. Consequently, the distance between two equal poles is larger than in the onion or the horseshoe state and the distance between two opposite poles is smaller. Thus, this configuration is energetically more favorable than the two arrangements discussed before. The micro-vortex state can be seen as an "ordered" spin ice state. It fulfills the condition of demagnetization and the ice rules.

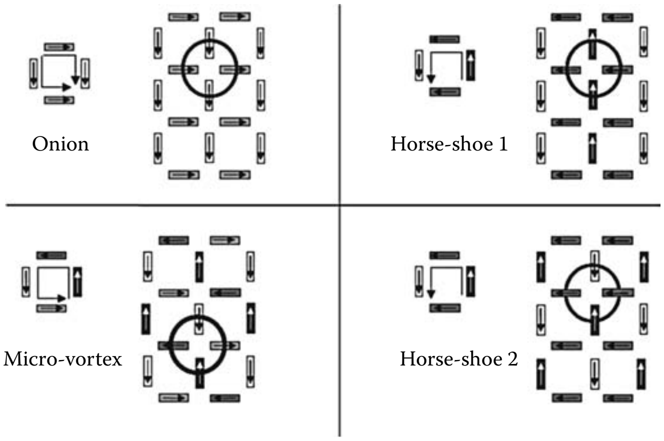

FIGURE 4.14 Schematic representation of the possible symmetrical (meta-) stable remanent states. On the right a representative square formed of four individual islands is depicted, on the left the long-range order of the respective structure is shown. The circles within the extended structure indicate the vertices.

The onion state can easily be realized experimentally. Applying an external field along the diagonal axis, i.e., under an angle of 45° with respect to the individual stripes, naturally leads to the onion state in remanence. In this case, the external field exhibits a strong component along the easy axis of each Py element, resulting in two magnetic poles in the opposite corners of the frames. For an applied magnetic field pointing along the horizontal stripes (0°), the situation is different. In this case the horizontal stripes are magnetized along the easy axis, while the vertical stripes are magnetized along their hard axis. In remanence, the shape anisotropy of the elongated islands supports the magnetization along the islands. While it is energetically favorable for the magnetization of the horizontal stripes to stay in the field direction, there is no preference for the vertical ones. If the effect of the stray field of the neighboring islands is negligible (i.e., at sufficiently large distances), 50% of the stripes are magnetically oriented north–south, while the other 50% are oriented south–north. For large distances (2 μm), this leads to the coexistence of the onion state together with the horseshoe 2 state (see Figure 4.15 right). For small distances (0.5 μm), the situation is different. Here, both sublattices are magnetically aligned parallel. Obviously, at small distances the stray field stabilizes the onion state irrespective of the direction of the field. At zero distance, i.e., in closed square rings, the onion state has also been observed in remanence. In this case the sample had to be carefully demagnetized to reach the vortex state.

The variation of distances between elongated Py elements without any variation of the particle shape changes the remanent configuration of the whole array. This means that with increasing interparticle distance the relative depth of energy minima corresponding to different configurations changes. For a deeper understanding of the observed phenomena, theoretical calculations are necessary. In the geometrically frustrated crossed bar configuration small distances lead to a quenching of the impressed fourfold anisotropy by the formation of the highly symmetric, long-range ordered onion state. At larger distances one of the sublattice starts to randomize and the onion state coexists with the horseshoe 2 state. The experimental results can be understood on the basis of the magnetostatic energy of the whole system, including higher-order terms. The one-dimensional energy curve displays three minima that can be identified as the micro-vortex state, the onion state, and the

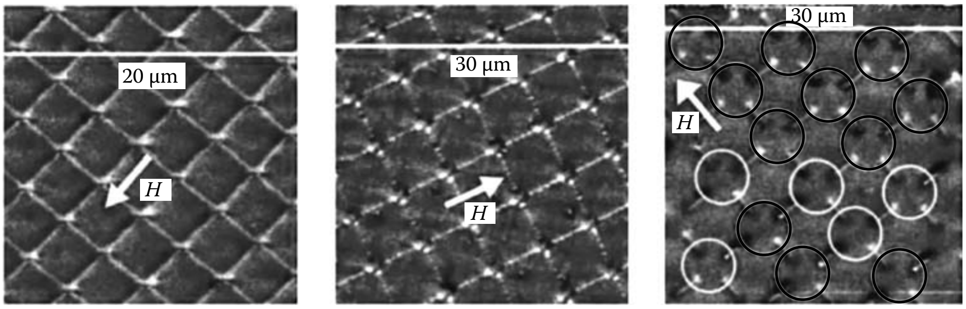

FIGURE 4.15 MFM images of Py dipole arrays on a square lattice, recorded in the remanent state ($H = 0$). The individual elements have separations of 0.42, 0.84, and 2.34 μm (from left to right). The field direction is indicated by the white arrow. In the panel to the right vertices displaying an onion state are highlighted with a black circle, vertices with a horseshoe state are highlighted with a white circle. (Reprinted from Remhof, A. et al., *Phys. Rev. B*, 77, 134409, 2008. With permission.)

horseshoe 2 state. Thereby, the formation of the micro-vortex state as the ground state is separated from the observed onion state by a substantial energy barrier.

Indeed, one of the initially longitudinally magnetized elements have to be completely reversed. Depending on the mechanism (coherent or incoherent magnetization rotation) the magnetization reversal requires an additional energy comparable with that of the magnetostatic coupling. The reversed element should be antiparallel to the applied magnetic field. This makes the magnetization reversal even more energetically expensive. The probability for climbing an energy barrier in the presence of available local minima is close to zero. Thus, it follows from the calculations that for initial 0° saturation of a Py array in the case of small separations, the most probable state is the onion magnetization configuration. However, for large interparticle distances a relevant number (≈30%) of horseshoe 2 states should appear. In remanence the energetically most favorable vortex state is inaccessible because of the large energy barrier of the magnetization reversal. A similar analysis has been performed for initially 45° saturated arrays. The main difference to the previous discussion is that the total energy of the 45° state is lower than that of the 0° configuration. Therefore, the energy minimum for the horseshoe 2 ordering disappears already for the more widely spaced arrays with dominating dipolar interactions. Hence, if an array has been initially magnetized under 45° to its axes the onion state should prevail in wider- as well as in denser-packed arrays. Energetically preferred micro-vortex configuration can be obtained only after a thorough demagnetization of an array. A detailed discussion of the theoretical calculations is given in Vedmedenko and Mikuszeit (2008).

Recently, magnetostatic interactions have also been studied in other frustrated systems. Especially the kagome lattice, a two-dimensional structure composed of corner-sharing triangles and the related honeycomb structure, gained considerable attraction. The kagome lattice and the honeycomb structure intrinsically exhibit a sixfold symmetry; both patterns are schematically represented in Figure 4.16. In the interconnected honeycomb net, three lines meet in each vertex. The ice rules change consequently from two-in–two-out to two-in–one-out or to one-in–two-out. In the magnetic ground state, prepared by careful demagnetization, Qi et al. observed that the magnetization along each connecting element of adopts a single domain, and that the domain walls are constrained within the vertices. Thereby, the ice rules are strictly obeyed. Compared to the square lattices discussed above, where the realization of the artificial spin ice is hindered by large energy barriers, the magnetic honeycomb structure turned out to be an ideal artificial spin ice system (Qi et al. 2008).

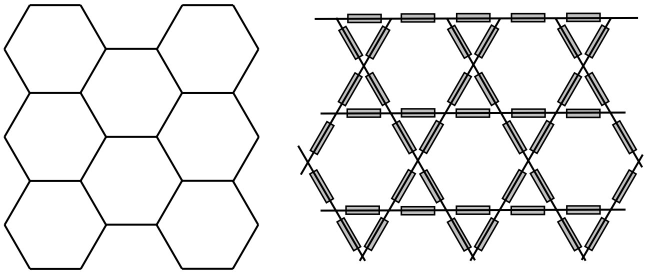

FIGURE 4.16 Schematic representation of the honeycomb structure (left) and of a periodic array of islands, placed between the vertices of a kagome lattice (right).

Rectangular islands, placed between the vertices of a kagome lattice, lead to a periodic structure in which the number of nearest neighbors of a lattice point is four as on a square lattice. However, the angles between the connecting lines to the nearest neighbors on the triangular lattice are 60° instead of 90°. Varying the size of the rectangular bars and the interparticle distance the effect of stray field interaction on the magnetization reversal was investigated by (Westphalen et al. 2008b). The islands on the kagome pattern can be subdivided into three sublattices with triangular symmetry. The magnetization reversal depends on the direction of the applied field. Two simple geometric cases may be distinguished: The field may be applied either parallel to one of the sublattices 0° orientation or perpendicular to one sublattice 30° orientation. For the 0° orientation, two reversal processes were observed. In case that the interaction between the net magnetic moments is weak, the particles inclined to the field direction by 60° flip first, followed by the sublattice parallel to the applied field. For stronger interaction at reduced particle distance, the parallel sublattice switches first, followed by the domain process in the inclined bars. Applying a magnetic field in the 30° orientation leads to a strong frustration of the particles with the hard axis orientation. This frustration causes domain formation preferentially in the perpendicular bars.

4.4 Summary and Discussion

The technological progress in the field of magnetoelectronics and the advancement in miniaturization techniques brought topological-ordered nano- and microstructured magnetic thin films in the focus of many research activities (Martín et al. 2003).

Understanding, modeling, and controlling the magnetization reversal of magnetic micro- and nanostructures as a function of shape, aspect ratio, and interparticle separation are of high current interest. Various shapes have been investigated in the past, including stripes, closed and open squares, and circular and elliptic dots.

Even though there exists a large variety of shapes, sizes, and materials, leading to a huge variety of properties, ferromagnetic islands have a few characteristics in common:

- In small ferromagnetic islands, the shape anisotropy dominates the remanent state and the reversal mechanism. Therefore, the geometrical extensions of magnetic structures determine their magnetic properties. Elongated structures tend to form magnetic dipoles, while small aspect ratios favor closed flux states.
- By lateral structuring of originally homogeneous, magnetically isotropic ferromagnetic thin films new anisotropy axes can be impressed in arbitrary directions.
- Micrometer and nanometer-sized ferromagnetic islands exhibit much larger coercive fields than bulk magnets composed of the same material.
- The magnetization reversal is dominated by coherent rotation in hard axis orientation and by domain wall motion in easy axis orientation.
- At small distances, magnetostatic interactions occur, leading to the collective behavior of an array of islands.

In future, the technological advancements of sample preparation will lead to even smaller individual structures and to larger arrays. Three major routes are currently followed. The first one tries to optimize the lithographic techniques by using electromagnetic radiation with smaller wavelength (Bakshi 2008) or by nanoimprint technologies (Chou et al. 1996). The second route involves self-assembly of clusters and surfaces to complex, periodic structures (Boncheva et al. 2005). The third route is the bottom-up technique. In this approach, the ferromagnetic island is built from its atoms. A tip from a scanning tunneling microscope can be used to move individual atoms on a surface, thereby reaching island sizes that will never be accessible with classical lithography (Stroscio and Eigler 1991). Further miniaturization will not only allow to study smaller objects, and bridge the gap between individual atoms, clusters and nano- to micrometer-sized islands, it also leads to a higher definition of the circumference of the island, reducing the often undesired influence of rough edges.

The research on ferromagnetic islands is nowadays embedded in the field of spintronics (Wolf et al. 2001). In the classical electronics the information is carried by the charge of the electrons, while its spin is ignored. In the future, the standard microelectronics will be combined with spin-dependent effects that arise from the interaction between spin of the carrier and the magnetic properties of the material. Therefore, the magnetic properties of ferromagnetic (FM) islands will be joined to other materials such as antiferromagnets (AF), semiconductors (SC), and superconductors (S). For extended films, the mutual interaction within F/AF and F/S heterostructures leading to the exchange bias effect and the superconducting proximity effect have been studied in great detail in the past (Zabel and Bader 2008). Patterning the ferromagnetic layer allows to control its coercivity and its anisotropy. It also allows to tune the size of the islands to the typical length scales of the AF, i.e., its domain size, and to the S, i.e., its coherence length. The exchange bias, for example, stabilizes the magnetization of ferromagnetic islands close to the superparamagnetic limit, fueling the hope of higher storage densities in magnetic recording devices (Skumryev et al. 1991). Ferromagnetic islands play an important role for the preparation of spin-polarized electric currents as the electrical conductivity of the majority spin electrons differs substantially from minority spin, resulting in a spin-polarized electric current. The spin-polarized current may then be transferred into a SC device ballistically via an ohmic contact. However, F/SC contacts involve doping of the semiconductor surface, leading to spin-flip scattering and loss of the spin polarization. Understanding and optimizing spin injection and spin transport are some of the key challenges for the development of truly spintronic devices.

Another degree of freedom are vertical structures, in which the islands are stacked on top of each other, separated by nonmagnetic material. In such a way, one-dimensional spin chains may be realized, in which the islands interact via their magnetostatic fields (Fraerman et al. 2008). It has already been shown, that interacting islands may be used as logic devices (Wang et al. 2006). Research in this area is still at a very early stage and far away from a technical application; the perspective, however, to functionalize the magnetostatic interactions will have stimulating effects on the research worldwide.

References

Adeyeye, A. O. and Singh, N. 2008. Large area patterned magnetic nanostructures. *J. Phys. D: Appl. Phys.* 41: 153001.

Aharoni, A. 2000. *Introduction to the Theory of Ferromagnetism*. Oxford, U.K.: Oxford Science Publications.

Allwood, D. A., Xiong, G., Cooke, M. D. et al. 2002. Submicrometer ferromagnetic NOT gate and shift register. *Science* 296: 2003–2006.

Allwood, D. A., Xiong, G., Cooke, M. D., and Cowburn, R. P. 2003. Magneto-optical Kerr effect analysis of magnetic nanostructures. *J. Phys. D: Appl. Phys.* 36(18): 2175–2182.

Bakshi, V. 2008. *EUV Lithography*, Bellingham, WA: SPIE Publications.

Bauer, E. 1994. Low energy electron microscopy. *Rep. Prog. Phys.* 57: 895–938.

Bloch, F. 1932. Zur Theorie des Austauschproblems und der Remanenzerscheinung der Ferromagnetika. *Z. Phys.* 74 (5–6): 295–335.

Boncheva, M., Andreev, S. A., Mahadevan, L. et al. 2005. Magnetic self-assembly of three-dimensional surfaces from planar sheets. *Proc. Natl. Acad. Sci. USA* 102(11): 3924–3929.

Brown, W. F. 1963. *Micromagnetics*. New York: John Wiley & Sons.

Bucknall, D. G. 2005. *Nanolithography and Patterning Techniques in Microelectronics*. Cambridge, U.K.: Woodhead Publishing in Materials.

Castaño, F. J., Ross, C. A., Frandsen, C. et al. 2003. Metastable states in magnetic nanorings. *Phys. Rev. B* 67(18): 184425.

Chapman, J. N., Johnston, A. B., and Heyderman, L. J. 1994. Coherent Foucault imaging: A method for imaging magnetic domain structures in thin Films. *J. Appl. Phys.* 76(9): 5349–5355.

Chou, S. Y., Krauss, P. R., and Renstrom, P. J. 1996. Imprint lithography with 25-nanometer resolution. *Science* 272: 85–87.

Cowburn, R. P., Koltsov, D. K., Adeyeye, A. O. et al. 1999. Single-domain circular nanomagnets. *Phys. Rev. Lett.* 83(5): 1042–1045.

Dennis, C. L., Borges, R. P., Buda, L. D. et al. 2002. The defining length scales of mesomagnetism: A review. *J. Phys. Condens. Matter* 14: R1175–R1262.

Donahue, M. J. and Porter, D. G. 1999. *OOMMF User's Guide, Version 1.0, Interagency Report NISTIR 6376*. Gaithersburg, MD: National Institute of Standards and Technology.

Feldtkeller, E. and Stein, K. U. 1967. Verbesserte Kerr-Technik zur Beobachtung magnetischer Domänen. *Z. Angew. Phys.* 23: 100–102.

Fischer, P., Eimüller, T., Schütz, G. et al. 1999. Magnetic domain imaging with a transmission X-ray microscope. *J. Magn. Magn. Mater.* 198–199: 624–627.

Foner, S. 1959. Versatile and sensitive vibrating-sample magnetometer. *Rev. Sci. Instrum.* 30(7): 548–557.

Fraerman, A. A., Gusev, S. A., Mazo, L. A. et al. 2002. Rectangular lattices of permalloy nanoparticles: Interplay of single-particle magnetization distribution and interparticle interaction. *Phys. Rev. B* 65(6): 064424.

Fraerman, A. A., Gribkov, B. A., Gusev, S. A. et al. 2008. Magnetic force microscopy of helical states in multilayer nanomagnets. *J. Appl. Phys.* 103(7): 073916.

Fullerton, E. E., Jiang, J. S., and Bader, S. D. 1999. Hard/soft magnetic heterostructures: Model exchange-spring magnets. *J. Magn. Magn. Mater.* 200(1–3): 392–404.

Grimsditch, M. and Vavassori, P. 2004. The diffracted magneto-optic Kerr effect: What does it tell you? *J. Phys. Condens. Matter* 16(9): R275–R294.

Grimsditch, M., Vavassori, P., Novosad, V. et al. 2002. Vortex chirality in an array of ferromagnetic dots. *Phys. Rev. B* 65(17): 172419.

Hausmanns, B., Krome, T. P., Dumpich, G. et al. 2002. Magnetization reversal process in thin Co nanowires. *J. Magn. Magn. Mater.* 240(1–3): 297–300.

Heisenberg, W. 1928. Zur Theorie des Ferromagnetismus. *Z. Physik* 49: 619–636.

Hertel, R., Fruchart, O., Cherifi, S. et al. 2005. Three-dimensional magnetic-flux-closure patterns in mesoscopic Fe islands. *Phys. Rev. B* 72: 214409.

Hopster, H and Oepen, H. P. 2005. *Magnetic Microscopy of Nanostructures*. Berlin Heidelberg: Springer Verlag.

Hubert, A. and Schäfer, R. 1998. *Magnetic Domains, The Analysis of Magnetic Microstructures*. Heidelberg, Germany: Springer Verlag.

Imre, A., Csaba, G., Ji, L. et al. 2006. Majority logic gate for magnetic quantum-dot cellular automata. *Science* 311(5758): 205–208.

Kerr, J. 1877. On the rotation of the plane of polarisation by reflection from the pole of a magnet. *Philos. Mag.* 3: 321–343.

Kläui, M., Rothman, J., Lopez-Diaz, L. et al. 2001. Vortex circulation control in mesoscopic ring magnets. *Appl. Phys. Lett.* 78(21): 3268–3270.

Kläui, M., Vaz, C. A. F., Bland, J. A. C. et al. 2003. Direct observation of spin configurations and classification of switching processes in mesoscopic ferromagnetic rings. *Phys. Rev. B* 68(13): 134426.

Landau, L. and Lifschitz, L. 1935. On the theory of dispersion of magnetic permeability in ferromagnetic bodies. *Physik. Z. Sowjetunion* 8: 153–169.

Last, T., Hacia, S., Wahle, M. et al. 2004. Optimization of nanopatterned permalloy electrodes for a lateral hybrid spin-valve structure. *J. Appl. Phys.* 96(11): 6706–6711.

Lee, M.-S., Westphalen, A., Remhof, A. et al. 2008. Extended longitudinal vector and Bragg magneto-optic Kerr effect for the determination of the chirality distribution in magnetic vortices. *J. Appl. Phys.* 103(9): 093913.

Martin, Y. and Wickramasinghe, H. K. 1987. Magnetic imaging by "force microscopy" with 1000 Å resolution. *Appl. Phys. Lett.* 50(20): 1455–1457.

Martín, J. I., Nogués, J., Liu, K. et al. 2003. Ordered magnetic nanostructures: Fabrication and properties. *J. Magn. Magn. Mater.* 256: 449–501.

Mundschau, M., Romanowicz, J., Wang, J. Y. et al. 1996. Imaging of ferromagnetic domains using photoelectrons: Photoelectron emission microscopy of neodymium-iron-boron ($Nd_2Fe_{14}B$). *J. Vac. Sci. Technol. B* 14(4): 3126–3130.

Oepen, H. and Kirschner, J. 1991. Imaging of magnetic microstructures at surfaces—The scanning electron microscope with spin polarization analysis. *Scanning Microsc.* 5(1): 1–16.

Parkin, S. S. P., Hayashi, M., and Thomas, L. 2008. Magnetic domain-wall racetrack memory. *Science* 320(5873): 190–194.

Philo, J. S. and Fairbank, W. M. 1977. High-sensitivity magnetic susceptometer employing superconducting technology *Rev. Sci. Instrum.* 48: 1529–1536.

Prohl, A. 2001. *Computational Micromagnetism*. Stuttgart, Germany: B. G. Teubner.

Qi, Y., Brintlinger, T., and Cumings, J. 2008. Direct observation of the ice rule in an artificial kagome spin ice. *Phys. Rev. B* 77: 094418.

Raabe, J., Pulwey, R., Sattler, R. et al. 2000. Magnetization pattern of ferromagnetic nanodisks. *J. Appl. Phys*. 88(7): 4437–4439.

Rahm, M., Biberger, J., Umansky, V. et al. 2003. Vortex pinning at individual defects in magnetic nanodisks. *J. Appl. Phys*. 93(10): 7429–7431.

Remhof, A., Westphalen, A., Theis-Bröhl, K. et al. 2007. Magnetization reversal studies of periodic magnetic arrays via scattering methods. In *Springer Series in Materials Science: Magnetic Nanostructures*, B. Aktas, L. Tagirov, and F. Mikailov (eds.), pp. 65–96. Heidelberg, Germany: Springer Verlag.

Remhof, A., Schumann, A., Westphalen, A. et al. 2008. Magnetostatic interactions on a square lattice. *Phys. Rev. B* 77: 134409.

Rothman, J., Kläui, M., Lopez-Diaz, L. et al. 2001. Observation of a Bi-domain state and nucleation free switching in mesoscopic ring magnets. *Phys. Rev. Lett*. 86(6): 1098–1101.

Schneider, M., Hoffmann, H., and Zweck, J. 2000. Lorentz microscopy of circular ferromagnetic permalloy nanodisks. *Appl. Phys. Lett*. 77(18): 2909–2911.

Shinjo, T., Okuno, T., Hassdorf, R. et al. 2000. Magnetic vortex core observation in circular dots of permalloy. *Science* 289(5481): 930–932.

Skumryev, V., Stoyanov, S., Zhang, Y. et al. 1991. Beating the superparamagnetic limit with exchange bias. *Nature (London)* 423: 850–853.

Stroscio, J. A. and Eigler, D. M. 1991. Atomic and molecular manipulation with the scanning tunneling microscope. *Science* 254(5036): 1319–1326.

Suzuki, K. and Smith, B. W. 2005. *Microlithography: Science and Technology*. Boca Raton, FL: CRC Press.

Tsoi, M., Fontana, R. E., and Parkin, S. S. P. 2003. Magnetic domain wall motion triggered by an electric current. *Appl. Phys. Lett*. 83(13): 2617–2619.

van den Berg, H. A. M. 1986. Self-consistent domain theory in soft-ferromagnetic media. II. Basic domain structures in thin-film objects. *J. Appl. Phys*. 60: 1104–1113.

Vavassori, P. 2000. Polarization modulation technique for magneto-optical quantitative vector magnetometry. *Appl. Phys. Lett*. 77(11): 1605–1607.

Vavassori, P., Grimsditch, M., Novosad, V., Metlushko, V., and Ilic, B. 2003. Metastable states during magnetization reversal in square permalloy rings. *Phys. Rev. B* 67(13): 134429.

Vaz, C. A. F., Hayward, T. J., Llandro, J. et al. 2007. Ferromagnetic nanorings. *J. Phys. Condens. Matter* 19(14): 255207.

Vedmedenko, E. Y. 2007. *Competing Interactions and Pattern Formation in Nanoworld*. Weinheim, Germany: Wiley-VHC.

Vedmedenko, E. Y. and Mikuszeit, N. 2008. Multipolar ordering in electro- and magnetostatic coupled nanosystems. *Chem. Phys. Chem*. 9(9): 1222–1240.

Vedmedenko, E. Y., Mikuszeit, N., Oepen, H. P., and Wiesendanger, R. 2005. Multipolar ordering and magnetization reversal in two-dimensional nanomagnet arrays. *Phys. Rev. Lett*. 95: 207202.

Wang, R. F., Nisoli, C., Freitas, R. S. et al. 2006. Artificial spin ice in a geometrically frustrated lattice of nanoscale ferromagnetic islands. *Nature (London)* 439: 303–306.

Waser, R. 2003. *Nanoelectronics and Information Technology*. Weinheim, Germany: Wiley-VCH.

Westphalen, A., Lee, M.-S., Remhof, A., and Zabel, H. 2007. Vector and Bragg magneto-optical Kerr effect for the analysis of nanostructured magnetic arrays. *Rev. Sci. Instrum*. 78(12): 121301.

Westphalen, A., Remhof, A., and Zabel, H. 2008a. Magnetization reversal in nanowires with a spiral shape. *J. Appl. Phys*. 104(1): 013906.

Westphalen, A., Schumann, A., Remhof, A., Zabel, H., Last, T., and Kunze, U. 2006. Magnetization reversal of equilateral Fe triangles. *Phys. Rev. B* 74(10): 104417.

Westphalen, A., Schumann, A., Remhof, A. et al. 2008b. Magnetization reversal of microstructured kagome lattices. *Phys. Rev. B* 77: 174407.

Wolf, S. A., Awschalom, D. D., Buhrman, R. A. et al. 2001. Spintronics: A spin-based electronics vision for the future. *Science* 294(5546): 1488–1495.

Zabel, H. and Bader, S. D. 2008. *Magnetic Heterostructures, Advances and perspectives in Spinstructures and Spintransport*. Heidelberg, Germany: Springer Verlag.

5

A Single Nano-Dot Embedded in a Plate Capacitor

5.1 Introduction 5-1
5.2 Studied Configuration: Geometry and Notations 5-3
5.3 Metallic Dot 5-3
Definitions of Capacitance and Coupling Coefficient • Influence of the Charge • Influence of Polarization • Numerical Implementation • Numerical Results and Comparison with a Classical Approximation • Conclusion
5.4 Semiconductor Dot: Theoretical Approach 5-9
General Considerations on Quantization Effects • Quantum Dot Containing a Single Electron: A Simplified Approach • Addition of a Second Electron in the Dot • Multiple Electrons in the Dot
5.5 Finite Element Modeling of a Silicon Quantum Dot 5-14
Studied Structure and Meshing Strategy • Numerical Results • Comparison with the Metallic Approximation
5.6 Conclusion 5-18
Appendix A: A Numerical Validation of the Semi-Analytical Approach 5-19
Convergence • Numerical Validation with FEM Approach
References 5-21

Gilles Micolau
Universités d'Aix Marseille

Damien Deleruyelle
Universités d'Aix Marseille

5.1 Introduction

Nowadays, nonvolatile memories (NVM) are used on a massive scale in different domains related to personal and industrial applications such as cellular phones, massive computer data storage on mobile support, digital photography, aeronautics, and embedded electronic in automotive applications.

The current NVM technology is based on electric charge storage in silicon-based devices. The scaling of these devices has been predicted by Gordon Moore. In his famous paper published in 1965, he mentions: "…The complexity for minimum component costs as increased at a rate of roughly a factor two per year […] Over the longer term […] there is no reason to believe it [the rate] will not remain nearly constant for at least 10 years" [1].

As a matter of fact, even after 40 years, this prediction is still valid [2]. However, the typical size of memory cells has been so much reduced that a further miniaturization becomes critical (recently, NVM cells have been proposed with 43 nm typical size [technology node]) [3]. Indeed, new phenomena such as quantum or finite size effects and single-electron effects [4–6] challenge the possibility of further reduction.

In the field of industrial and academic research, two major parallel axes are proposed to solve these issues. The first one is based on the use of new materials using singular physical properties (phase changed memory [7], resistive random access memory [8], etc.). These promising approaches are not yet completely mature to replace the well-known traditional devices, studied for at least a half century. The other approach consists in modifying, step by step, the current devices in order to push the scaling limits of silicon-based devices. This chapter is devoted to this second approach.

The essential element of the current NVM is the metal oxide semiconductor transistor (TMOS), which can be seen as a nonlinear valve of electronic current controlled by an external voltage. To simplify, the schematic behavior of the MOS transistor is presented in Figure 5.1. It can be described as follows: the current flowing between the two polarized pads—source and drain—is null when the voltage applied on the control gate is lower than a threshold voltage (V_{T0}). It becomes arbitrarily high when the voltage is equal or greater than V_{T0}. The value of V_{T0} is an intrinsic quantity of the TMOS. The main idea of the NVM is to modulate the value of V_{T0} by storing electric charges, spatially confined in a floating gate located between the control gate and the channel of electronic transport. Injecting negative (resp. positive) charges in the floating gate shifts the value of the threshold voltage to values greater (resp. lower) than V_{T0} as shown in Figure 5.2. Consequently, two different electrical states are possible: the "up" one defined by $V_{Tu} > V_{T0}$ and the "down"

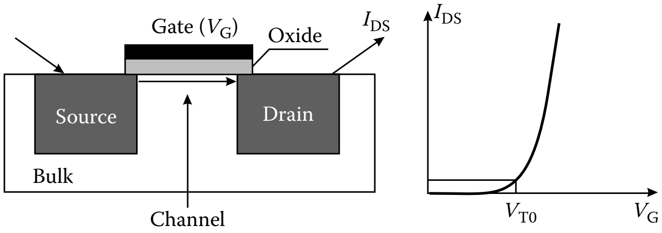

FIGURE 5.1 Schematic behavior of a MOS transistor.

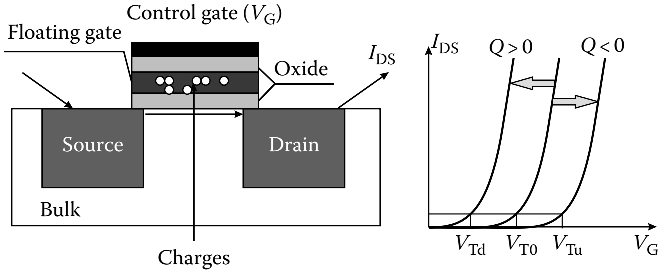

FIGURE 5.2 Schematic behavior of a TMOS with a floating gate.

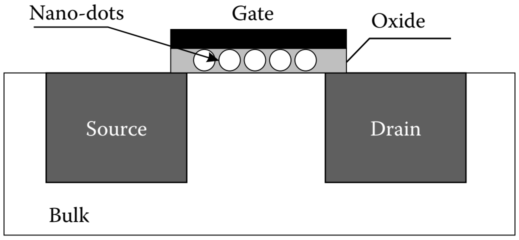

FIGURE 5.3 Floating nano-dots.

one with $V_{Td} < V_{T0}$. These two states define a level of programmation for the cell. Depending on the technology, the "up" (resp. "down") state is called programmed or erased (resp. erased or programmed) and the cell can store one fundamental memory unit, the well-known *bit*.

Classically, the floating gate is designed with degenerated polysilicon whose electrical behavior is quasi-metallic. It is electrically insulated by oxide layers (silicon oxide for the bottom and a stack of different oxides on its top) designed to maintain the electric charges as long as possible in the floating gate. One major issue in scaling is the reduction of the thickness of the oxides leading to an alteration of the retention ability of the device. Another critical issue is the size reduction of the floating gate. To work properly, the memory cell needs a certain quantity of electric charges—injected in its floating gate. Scaling the floating gate and the surrounding oxides implies using less charges for the programmation of the cell. The repartition of less charges in the volume of a small floating gate makes the discrimination between the two possible states of the cell more difficult (the classical term is "closure of the programming window"). Moreover, another issue is linked to the apparition of traps in the bottom oxide when cycling the memory device. This phenomenon is known as *stress-induced leakage current* (SILC) [9,10]. Because the charges are delocalized in the floating gate, the presence of the unique trap in the bottom dielectric can discharge the entire memory device and lead to the data loss. To tackle these issues, one solution could be to constrain charges in the vicinity of the channel. For accomplishing this goal, the use of nanoscale dots instead of a continuous floating gate has been experimentally and theoretically studied for many years [11–14]. The principle of the configuration is shown in Figure 5.3.

For modeling the electrical characteristics of these devices, such as programming (erasing) or retention, knowledge of the electrostatics and, more generally, the physics of this type of devices is needed. The physical properties of such nanoobjects have been studied for many years, with the help of different innovative techniques such as scanning tunneling microscopy (STM), atomic force microscopy (AFM), transmission electron microscopy (TEM), and so on. A detailed review about nanoclusters can be found in Ref. [15].

The modeling of dynamics phenomenon such as tunneling currents needs an accurate modeling of the electric field in the device. For metallic or quasi-metallic dots, a classical (non-quantum) approach based on equivalent capacitance is efficient and has been even used in the case of large semiconductor islands [16–19]. Theoretically, for small semiconductor dots, solving the electrostatics of the configuration must be achieved by a self-consistent resolution of Poisson and Schrödinger equations as shown in Refs. [20–22]. However, this huge complete resolution can be avoided. First of all, the number of dots involved in the problem has a crucial influence. The electrostatics of a device with numerous dots can be governed by the classical interaction between each dot, even if the typical size of the dot is lower than the typical quantum range. Second, the current typical size of dots does not necessarily imply that the quantum approach is justified. So, many models assume that the dots are metallic and interact as plane capacitors. The validity of this approach has been proved in Ref. [23]. However, it was shown in Ref. [24], in a 2D case and for a single dot arbitrarily shaped, that this approach is no longer valid in the range of aspect ratios encountered in nanodot arrays. To enclose this state of art, in the context of ultrascaled devices, the individual electrical behavior of a dot can become significant in front of the collective behavior of a matrix of dots [6,25].

We aim to model, academically and as precisely as possible, the electric behavior of a single dot embedded in a dielectric oxide. So, we choose a very canonical and academic physical configuration, particularly with the twin aims of

- Determining the critical size of the dot leading to quantum effects
- Comparing the classical approximations with models giving more physical insights

In this chapter, we focused on a single spherical island, electrically charged, embedded in a plane capacitor. In the first part, we propose a rigorous—by this word, we mean that there are no physical approximations—non-quantum model for the electrostatics of a metallic dot and give a semi-analytical expression for its capacitance. In the second part, we propose a quantum approach for a dielectric dot, based on solving the Poisson and Schrödinger equations, with a discrete number of charges in the dot. Finally, results obtained from both approaches are quantitatively confronted to determine in which extent the quantum effects have to be considered.

5.2 Studied Configuration: Geometry and Notations

A spherical dot, of radius a, is embedded in a dielectric medium of relative dielectric permittivity ε_r, located between two infinite conductor planes geometrically supported by two planes Π_1 and Π_2. Each is held at a constant electric potential (V_1 for Π_1, V_2 for Π_2). The spherical dot is initially charged with an electrical charge Q_0, with a process not studied here. We assume that $\Delta V = V_1 - V_2$ and Q_0 are independent. We denote by σ the surface charge density of the dot when metallic and ρ the volumic charge density when dielectric. h is the distance between the two plate electrodes. All distances and notations are defined in Figure 5.4. Any point M in the dielectric medium is characterized by its Cartesian coordinates (x, y, z) and also by the spherical ones (r, θ, ϕ). Because the physics of the configuration is invariant by any rotation along the longitudinal angle ϕ, only the two spherical coordinates $\left(r = OM, \theta = \widehat{(\mathbf{e}_z, \mathbf{OM})}\right)$ are required to locate M.

5.3 Metallic Dot

In this section, we assume that the dot is a perfect conductor medium. As a consequence, the potential is constant in the volume of the dot but also floating because it is electrically insulated. We solve the classic electrostatics (non-quantum approach) of the configuration by a rigorous semi-analytical expansion of the potential. By rigorous, we mean that there are no physical approximations in our model. There are only numerical truncations and numerical approximations arising from the implementation. The model is based on the work exposed in Ref. [26] where the author rigorously solves the electrostatics of this problem, with the help of the method of images and the Legendre polynomial expansions. The author gives the solution for the potential in the medium (and discuss its mathematical validity) and the electrostatic force acting on the charged dot. Nevertheless, the evaluation of floating potential is not explicit at all. Starting from this model, we derive a rigorous semi-analytical expression for the capacitance and a very simple expression for the polarization coupling coefficient. The floating potential is therefore directly determined without a recurring iterative process. Then, we validate the numerical implementation and compare results with classical capacitance approximations.

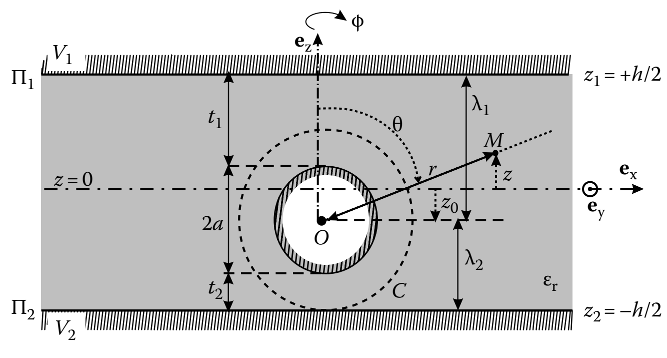

FIGURE 5.4 Configuration of study.

In this section, we first define the physical parameters required for the resolution of the electrostatics. Secondly, we find a semi-analytical expression for the capacitance and the coupling coefficient. Finally, we conclude with the numerical investigations of the model.

5.3.1 Definitions of Capacitance and Coupling Coefficient

Due to the linearity of the electrostatic equations and the unicity of the Poisson equation—with given boundary conditions—the potential, at any point M in the dielectric medium, can be described as the sum of two terms. The first one, denoted by $V^{Q}(M)$, is the potential created by the dot when charged (with electric charge Q_0) without polarization ($\Delta V = V_1 - V_2 = 0$). The second term, denoted by $V^{\mathrm{po}}(M)$, is the potential created by the polarization between the two electrodes ($\Delta V = V_1 - V_2$) when $Q_0 = 0$. Thus, the electric potential anywhere in the dielectric medium can be written as

$$V(M) = V^{Q}(M) + V^{\mathrm{po}}(M) \tag{5.1}$$

The computation of the function $V(M)$ involves the resolution of the Poisson equation with three Dirichlet boundary conditions: the potential is constant and imposed on the electrodes (Π_1 and Π_2), is constant, and floating on the dot. The floating potential, denoted by V_{f}, can be split into two parts: v_{f}^{Q}, due to the charge, and $v_{\mathrm{f}}^{\mathrm{po}}$, due to the polarization. It can now be written as

$$V_{\mathrm{f}} - V_2 = v_{\mathrm{f}}^{Q} + v_{\mathrm{f}}^{\mathrm{po}} - V_2 \tag{5.2}$$

Determining V_{f} requires another boundary condition. The charge density over the dot, σ, is linked to the normal derivative of the floating potential (normal component of the electric field):

$$\left(\frac{\partial V}{\partial r}\right)_{r=a} = -\frac{\sigma(\theta)}{\varepsilon_0 \varepsilon_r} \quad \text{and} \quad Q_0 = \int_{\theta=0}^{\pi} \sigma(\theta) 2\pi a^2 \sin\theta \, d\theta \tag{5.3}$$

Classically the capacitance is defined as follows: "…the capacitance of a conductor is the charge of the conductor when it is maintained at unit potential, all others conductors being held at zero potential" [27,28]. We define the capacitance of the dot, C_S, as

$$C_S = \left(\frac{\partial Q_0}{\partial v_{\mathrm{f}}^{Q}}\right)_{\Delta V} = \left(\frac{\partial Q_0}{\partial V_{\mathrm{f}}}\right)_{\Delta V} = \frac{Q_0}{v_{\mathrm{f}}^{Q}} \tag{5.4}$$

The floating potential linearly depends on the total difference of potential ΔV. We define the polarization coupling coefficient α as

$$\alpha = \left(\frac{\partial v_{\mathrm{f}}^{\mathrm{po}}}{\partial \Delta V}\right)_{Q_0} = \left(\frac{\partial V_{\mathrm{f}}}{\partial \Delta V}\right)_{Q_0} = \frac{\Delta V}{v_{\mathrm{f}}^{\mathrm{po}}} \tag{5.5}$$

We thus obtain the classical equation for the floating potential

$$V_{\mathrm{f}} - V_2 = \frac{Q_0}{C_S} + \alpha \Delta V \tag{5.6}$$

This equation still remains valid for any shape of the dot. Nevertheless, a rigorous semi-analytical expression of C_S can just be extracted for the canonical spherical shape. In other cases, computations can be carried out with the help of rigorous models that need numerical methods of discretization of the shape (method of moments, boundaries models, variational computations [24,29]) or by approximated analytical expansions (Rayleigh hypothesis on boundary conditions for instance [30]).

Let us denote by Q_1 (resp. Q_2) the charge supported by the top (resp. bottom) half-part of the dot. We write these quantities with an unique notation $Q_{(1,2)}$. Introducing here the angles $\theta_1 = 0$ (resp. $\theta_2 = \pi/2$), denoted by the unique notation $\theta_{(1,2)}$, we express the partial charges, as we did with Q_0, integrating σ on the top (resp. bottom) part of the dot:

$$Q_{(1,2)} = 2\pi a^2 \int_{\theta_{(1,2)}}^{\theta_{(1,2)}+\pi/2} \sigma(\theta) \sin\theta \, \mathrm{d}\theta \tag{5.7}$$

Deriving the relation (5.7) along v_{f}^{Q} defines the partial capacitances C_1 and C_2, formalized with the following unique notation $C_{(1,2)}$:

$$C_{(1,2)} = \left(\frac{\partial Q_{(1,2)}}{\partial v_{\mathrm{f}}^{Q}}\right)_{\Delta V} \tag{5.8}$$

The Chasles relation on the integral over θ leads obviously to the equality $C_1 + C_2 = C_S$.

We can define the ratio $\kappa_{(1,2)}$ between the partial capacitances and the total one:

$$\kappa_{(1,2)} = \frac{C_{(1,2)}}{C_S} = \left(\frac{\partial Q_{(1,2)}}{\partial Q_0}\right)_{\Delta V} \quad \text{and so} \quad \kappa_1 = 1 - \kappa_2 \tag{5.9}$$

As ΔV and Q_0 are independent, the coefficients $\kappa_{(1,2)}$ and α are not equal. Mathematically this property arises from the fact that the partial derivatives do not involve the same constant physical parameters (Q_0 for α and ΔV for κ_2).

The two following sections are dedicated respectively to the semi-analytical expansions of C_S and α.

5.3.2 Influence of the Charge

In this section, we assume that there is no polarization between the electrodes ($V_1 - V_2 = 0$). We focus on the term V^{Q} in the relation (5.1). The boundary conditions for the potential are such as

(Qa) The potential is constant on the dot. This floating potential is denoted by $v_{\mathrm{f}}^{Q} = V^{Q}(r = a) = \dfrac{Q_0}{C_S}$.

(Qb) The potential is null everywhere on each plate electrode.

(Qc) The relations (5.3) are still valid.

Because the geometry simultaneously involves plane and spherical structures (electrodes and dot), it is interesting to include the plate boundary conditions in the expansion of the potential. For this, we use the classical method of images. The expansion of the potential is based on the work in Ref. [26] where the author treats the electrostatics of a charged dot rigorously in a nonpolarized plate capacitor. He uses the method of images and the Legendre polynomials expansion. The author arrives at the solution for the potential in the medium (and discuss its mathematical validity) and the electrostatic force acting on the charged dot. Nevertheless, the evaluation of the floating potential is not explicit so that the capacitance cannot be known.

5.3.2.1 Method of Images

The main idea of the method is to find a virtual distribution of charges involving the same boundaries conditions than the initial distribution [28]. Because of the unicity of the solution of the Poisson equation for given boundary conditions, the virtual distribution and the real one generate the same potential in the domain investigated. Here, we build a distribution of charges ensuring that the two planes Π_1 and Π_2 are planes of antisymmetry of the charge distribution. This is done by taking recursively the images of the sphere by the symmetries along the planes Π_1 and Π_2, as done in Refs. [26,31], and described in Figure 5.5. The z-coordinate z_p of the center O_p of the pth image is: $z_p = ph + (-1)^p z_0$, p describing the relative natural integers.

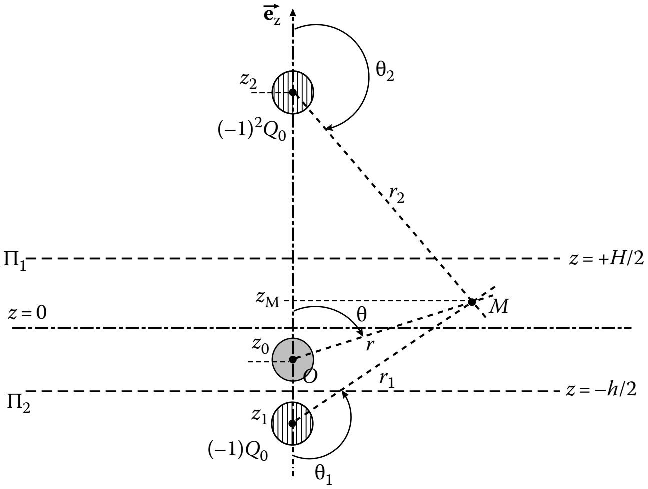

FIGURE 5.5 Coordinates of the images.

We denote by r_p the distance between the point of observation M and the center O_p, and θ_p the oriented angle between the line (O_pM) and vertical axis. To obtain a correctly oriented angle, it is necessary to alternate the sense of the axis $\mathbf{e}_z$ attached to the considerated image. We get so $\theta_p = \left(\widehat{(-1)^p\mathbf{e}_z, \mathbf{O_PM}}\right)$. These variables are shown in Figure 5.5. The classical Pythagore theorem and the definition of the geometric scalar product give the two coupled relations:

$$\begin{cases} r_p^2 = (z_0 - z_p)^2 + r^2 + 2r(z_0 - z_p)\cos\theta \\ \cos\theta_p = (-1)^p \dfrac{r\cos\theta + (z_0 - z_p)}{r_p} \end{cases} \tag{5.10}$$

To simplify the notations, we note for every point located on the surface of the dot $(r = a) : r_p^a(\theta) = (r_p)_{r=a}$ and $c_p^a(\theta) = (\cos\theta_p)_{r=a}$. Derivating the relations (5.10) along the variable r leads to

$$\begin{cases} dr_p^a(\theta) = \left(\dfrac{\partial r_p}{\partial r}\right)_{(r=a)} = \dfrac{a + (z_0 - z_p)\cos\theta}{r_p^a(\theta)} \\ dc_p^a(\theta) = \left(\dfrac{\partial \cos(\theta_p)}{\partial r}\right)_{(r=a)} = (-1)^p\left(\dfrac{\cos\theta}{r_p^a(\theta)} - \dfrac{a\cos\theta + (z_0 - z_p)}{r_p^a(\theta)^2}\, dr_p^a(\theta)\right) \end{cases} \tag{5.11}$$

5.3.2.2 Semi-Analytical Expression for the Capacitance

Each sphere centered in O_p generates a potential $v_p(M)$ expandable as an infinite set of Legendre polynomials [28]. Because the spheres get the same distribution of charges (or opposite), each sphere generates the same potential v_p (or opposite), vanishing far from its centre. The potential becomes [26,28]

$$v_p(M) = (-1)^p \sum_{l=0}^{+\infty} B_l \left(\frac{a}{r_p}\right)^{l+1} P_l(\cos\theta_p) \tag{5.12}$$

where P_l is the Legendre polynomial of order l and B_l are constant coefficients. The resolution of the electrostatics of the configuration is mathematically equivalent to determining B_l using boundary conditions.

The potential $V^Q(M)$, at every $M(r, \theta)$ in the dielectric medium, is the sum of the contributions of all spheres:

$$\begin{aligned} V^Q(r,\theta) = {} & \sum_{l=0}^{+\infty} B_l \left(\frac{a}{r}\right)^{l+1} P_l(\cos\theta) \\ & + \sum_{p*} (-1)^p \sum_{l=0}^{+\infty} B_l \left(\frac{a}{r_p(\theta)}\right)^{l+1} P_l(\cos\theta_p) \end{aligned} \tag{5.13}$$

where $p*$ denotes the summation over all the non-null relative integers.

The relation (5.13), combined with Equation 5.3, gives the charge of the sphere:

$$\begin{aligned} \frac{Q_0}{2\pi\varepsilon a} = {} & \sqrt{2}B_0 + a\sum_{p*}(-1)^p \sum_{l=0}^{+\infty} B_l \int_{\theta=0}^{\pi} \left[(l+1)\frac{dr_p^a(\theta)}{r_p^a(\theta)} P_l(c_p^a(\theta)) \right. \\ & \left. - \cdots dc_p^a(\theta) P_l^{(1)}(c_p^a(\theta)) \right] \left(\frac{a}{r_p^a(\theta)}\right)^{l+1} \sin\theta\, d\theta \end{aligned} \tag{5.14}$$

Because of the linearity between v_f^Q and Q_0, derivating Equation 5.14 along the floating potential is equivalent to dividing it by v_f^Q. So noting $\dot{B}_l = \left(B_l / v_f^Q\right) = \left(\partial B_l / \partial v_f^Q\right)$, it becomes

$$\begin{aligned} \frac{C_S}{2\pi\varepsilon a} = {} & \sqrt{2}\dot{B}_0 + a\sum_{p*}(-1)^p \sum_{l=0}^{+\infty} \dot{B}_l \int_{\theta=0}^{\pi} \left[(l+1)\frac{dr_p^a(\theta)}{r_p^a(\theta)} P_l(c_p^a(\theta)) \right. \\ & \left. - \cdots dc_p^a(\theta) P_l^{(1)}(c_p^a(\theta)) \right] \left(\frac{a}{r_p^a(\theta)}\right)^{l+1} \sin\theta\, d\theta \end{aligned} \tag{5.15}$$

Defining the row vector Γ, whom generic kth term Γ_k is

$$\begin{aligned} \Gamma_k = {} & \frac{\delta_{0,k-1}}{\sqrt{2}} + \frac{a}{2\sqrt{2}} \sum_{p*}(-1)^p \int_{\theta=0}^{\pi} \left[k\frac{dr_p^a(\theta)}{r_p^a(\theta)} P_{k-1}(c_p^a(\theta)) \right. \\ & \left. - \cdots dc_p^a(\theta) P_{k-1}^{(1)}(c_p^a(\theta)) \right] \left(\frac{a}{r_p^a(\theta)}\right)^{k} \sin\theta\, d\theta \end{aligned}$$

and defining the column vector $\dot{\mathbf{B}}$, the pth term of which is $\dot{B}_p$, we can write the relation (5.15) as a vector-product:

$$C_S = 4\pi a \boldsymbol{\Gamma}\dot{\mathbf{B}} \tag{5.16}$$

5.3.2.2.1 Expression of B_l Coefficients

The relation (5.13), written at $r = a$ and divided by v_f^Q, gives

$$\sum_{l=0}^{+\infty} \dot{B}_l P_l(\cos\theta) + \sum_{p*}(-1)^p \sum_{l=0}^{+\infty} \dot{B}_l \left(\frac{a}{r_p^a(\theta)}\right)^{l+1} P_l(c_p^a(\theta)) = 1, \quad \forall\theta \tag{5.17}$$

Projecting this relation on all the orders of the Legendre's polynomials (using their orthonormality and the fact that $P_0 = 1/\sqrt{2}$), we obtain the linear system:

$$\mathbb{M}\dot{\mathbf{B}} = \sqrt{2}\mathbf{1}_1$$

where

$\mathbf{1}_1$ is a column vector whose first coefficient is 1 and the others are zeros

$\mathbb{M}$ is a square matrix whose generic term M_{km} is

$$M_{km} = \delta_{k-1,m-1} + \sum_{p*} (-1)^p \int_{\theta=0}^{\pi} \left(\frac{a}{r_p^a(\theta)} \right)^m P_{m-1}(c_p^a(\theta)) P_{k-1}(\cos\theta) \sin\theta \, \mathrm{d}\theta \tag{5.18}$$

with $\delta_{k,m}$ the Kronecker symbol.

The coefficients $\dot{B}_l$ are so the unknown variables of a linear well-posed problem. The matrix $\mathbb{M}$ is numerically well-conditioned (its size and its numerical properties are explained in the numerical section of this part). The coefficients $\dot{B}_l$ are given by

$$\dot{\mathbf{B}} = \sqrt{2}\mathbb{M}^{-1}\mathbf{1}_1 \tag{5.19}$$

5.3.2.2.2 Matricial Expression for the Capacitance

Hence, the relation (5.16) becomes a simple matrix relation, numerically easy to manipulate:

$$C_S = 4\pi\varepsilon a \boldsymbol{\Gamma}\mathbb{M}^{-1}\mathbf{1}_1 \tag{5.20}$$

5.3.2.2.3 Corollary: Expression of Partial Capacitances

Using the relation (5.19) and the row vector $\Gamma^{(1,2)}$ (the superscript (1,2) indicates the part of the dot where we integrate σ) whose generic kth term $\Gamma_k^{(1,2)}$ is

$$\Gamma_k^{(1,2)} = \frac{1}{2}\delta_{0,k-1} + a\frac{1}{\sqrt{2}} \sum_{p*} (-1)^p \int_{\theta=\theta^{(1,2)}}^{\theta^{(1,2)}+\pi/2} \left[k \frac{dr_p^a(\theta)}{r_p^a(\theta)} P_{k-1}(c_p^a(\theta)) \right.$$
$$\left. - \cdots dc_p^a(\theta) P_{k-1}^{(1)}(c_p^a(\theta)) \right] \left(\frac{a}{r_p^a(\theta)} \right)^k \sin\theta \, \mathrm{d}\theta$$

the relation (5.8) becomes matricially

$$C_{(1,2)} = 4\pi\varepsilon a \boldsymbol{\Gamma}^{(\mathbf{1,2})}\mathbb{M}^{-1}\mathbf{1}_1 \tag{5.21}$$

5.3.2.3 Semi-Analytical Expression for the Potential

To finish this section, we rewrite the expression of the potential due to the charge as a matricial product. Using the relations (5.19) and (5.4) and the row vector $\mathbf{G}(r, \theta)$ whose generic term is

$$G_k(r,\theta) = \sqrt{2}\left(\frac{a}{r}\right)^k P_{k-1}(\cos\theta) + \sqrt{2}\sum_{p*}(-1)^p \left(\frac{a}{r_p}\right)^k P_{k-1}(\cos\theta_p)$$

the expression (5.13) for the potential $V^{\mathrm{Q}}(M)$ in the dielectric medium becomes

$$V^{\mathrm{Q}}(M) = \frac{Q_0}{C_S}\mathbf{G}(r,\theta)\mathbb{M}^{-1}\mathbf{1}_1 \tag{5.22}$$

5.3.3 Influence of Polarization

5.3.3.1 Analytical Expression for the Coupling Coefficient

In this section, we consider that $Q_0 = 0$ and that a polarization, $\Delta V = V_1 - V_2$, is applied between the two plate electrodes. As it is usually done in classical electrodynamics, the potential $V^{\mathrm{po}}(M)$—defined in relation (5.1)—is split into two parts. The first, denoted by v^i, is the potential due to the polarization ΔV in the dielectric medium considered as empty. It is often called "incident" or "primary" potential (classically used in scattering models [30,32]). It is the potential that should exist if the dot was not there. The second part of the potential, denoted by v^{d} is due to the presence of the dot in the polarized medium. Classically this term is called "scattered" potential or "secondary" potential. Then

$$V^{\mathrm{po}}(r,\theta) - V_2 = v^i(r,\theta) + v^{\mathrm{d}}(r,\theta) \tag{5.23}$$

v^i is the potential in a plane capacitor.

$$v^i(r,\theta) = (\lambda_2 + z - z_0)\frac{\Delta V}{h} = (\lambda_2 + r\cos\theta)\frac{\Delta V}{h} \tag{5.24}$$

The boundary conditions satisfied by $V^{\mathrm{po}}(M)$ are the following:

- (Pa) $V^{\mathrm{po}}(M) = V_1$ for $z = z_1$ and $V^{\mathrm{po}}(M) = V_2$ for $z = z_2$
- (Pb) The potential is floating and constant on the dot: $V^{\mathrm{po}}(a,\theta) - v_{\mathrm{f}}^{\mathrm{po}} \ \forall\theta$
- (Pc) Relations (5.3) with $Q_0 = 0$ are still valid

The scattered potential is expandable as a series of Legendre polynomials, for every M located in the same medium [28,30]. That is to say this expansion is valid only in the greatest sphere centered on O that does not cut the electrodes. This particular sphere, denoted by C in Figure 5.4, is tangent to the nearest electrode from the point O. Assuming that the nearest electrode is the bottom one (V_2), the radius of C is $\lambda_2 = a + t_2$. In this sphere ($r \le \lambda_2$), we have

$$v^{\mathrm{d}}(r,\theta) = \sum_{m=0}^{+\infty}\left[U_m\left(\frac{r}{a}\right)^m + W_m\left(\frac{a}{r}\right)^{m+1}\right]P_m(\cos\theta) \tag{5.25}$$

Using the boundary conditions (Pb) and (Pc) on the dot and the orthonormal properties of Legendre's polynomials, it follows that

$$\begin{cases} W_0 = 0 \\ U_0 = \left(v_{\mathrm{f}}^{\mathrm{po}} - V_2 - \lambda_2\dfrac{\Delta V}{h}\right)\sqrt{2} \\ U_m = -W_m - a\dfrac{\Delta V}{h}\sqrt{\dfrac{2}{3}}\delta_{m,1} \quad \forall m \ge 1 \end{cases} \tag{5.26}$$

The condition (Pa), for $\theta = \pi$, gives $U_m = -W_m$, $\forall m$. At zero-order it comes (with relations (5.26)):

$$v_{\mathrm{f}}^{\mathrm{po}} - V_2 = \frac{\lambda_2}{h}\Delta V \Rightarrow \alpha = \frac{\lambda_2}{h} \tag{5.27}$$

Remark: If the nearest electrode is the electrode 1, *the condition* (*Pa*) *gives* $v_{\mathrm{f}}^{\mathrm{po}} - V_2 = (1-(\lambda_1/h))\Delta V$ *which implies the same result for* α.

Finally, inside the sphere C, v^{d} can be written as

$$v^{\mathrm{d}}(r,\theta) = \frac{a}{h}\Delta V \frac{\lambda_2^3}{\lambda_2^3 - a^3}\left(\left(\frac{a}{\lambda_2}\right)^3 \frac{r}{a} - \left(\frac{a}{r}\right)^2\right)\cos\theta \tag{5.28}$$

5.3.3.2 Semi-Analytical Expression for the Potential in the Dielectric Medium

As a last corollary, the semi-analytical expansion for C_S and the simple expression of α allows us to express the total potential everywhere in the dielectric medium, when ΔV and Q_0 are non null.

To express the potential V^{po} everywhere in the dielectric, we use again the method of images for determining v^{d}. Indeed, v^i already contains the boundary conditions on the planes Π_1 and Π_2. The boundary conditions on v^{d} are $v^{\mathrm{d}}|_{\Pi_1} = v^{\mathrm{d}}|_{\Pi_2} = 0$. So we can expand the potential v^{d} as we did for V^Q in the previous section, with the same notations. Using the conditions (Pa), (Pb), (Pc), and the relation (5.27), we can finally write V^{po} as

$$V^{\mathrm{po}}(M) - V_2 = \frac{\lambda_2}{V}\Delta V \times \mathbf{G}\mathbb{M}^{-1}\mathbf{A}^h \tag{5.29}$$

where $\mathbf{A}^h$ is the column vector of generic term of order k, $A_k^h = -\sqrt{(1/3)}(a/h)\delta_{2,k}$.

Finally using the two relations (5.22) and (5.29), the total potential in the dielectric medium is

$$V(M) - V_2 = \mathbf{G}\mathbb{M}^{-1}\left(\frac{Q_0}{C_S}\mathbf{1}_1 + \frac{\lambda_2}{h}\mathbf{A}^h\right) \tag{5.30}$$

This expression allows one to estimate the distribution of charges on the electrodes, derivating this relation along z, at $z = 0$ and h. This is not the subject of this chapter but a parametric study of this distribution could be interesting to understand the classical approximations of C_S. Moreover, this should allow to obtain immediately the electric field everywhere in the medium. This physical quantity is necessary to model charge transport phenomenon through the dielectric layers.

To conclude this part, let us remind that the semi-analytical relations (5.20) and (5.27) solve the physical parameters of the configuration. As a final corollary, the matricial relation (5.30) gives the electrostatics of the dot.

We notice that this classical approach is invariant by scale change. So, these results can be directly used for larger configurations with the same aspect ratio. We can cite for instance chemical sciences where models of metallic bubbles embedded in dielectric solutions are often used.

5.3.4 Numerical Implementation

The numerical computation of the capacitance is numerically conditioned by the obtaining an invertible matrix $\mathbb{M}$. Numerically, the matrix $\mathbb{M}$ is a square matrix of the typical size 10 × 10. This is not a critical size for classical numeric inversion. Moreover, in practice, the conditioning (ratio between greatest and smallest values) of this matrix is about unity. Hence, the inversion of the matrix $\mathbb{M}$ is not the key point of this numerical implementation. Its quality is conditioned by the double truncation of the expansion of the potential (on the maximum order of polynomials in one hand and on the number of images in the other one). A discussion about the influence of this truncation is presented in Appendix A. Let us say here that the two truncations are independent one of each other, and numerically they are not critical at all for computations. Moreover, the semi-analytical expression was numerically validated by comparing results provided by finite element method (FEM) computations. This validation is also presented in Appendix A. To briefly sum up the conclusions, let us say that FEM may have some accuracy problems and it is clearly slower.

5.3.5 Numerical Results and Comparison with a Classical Approximation

Classically C_S and α are estimated by different physically based approaches using very simple expressions. Nevertheless, the validity of these latest approaches is not always well known. A very classical phenomenological approach is the equivalent capacitances approach (for details, see [24]) that we will briefly expose here.

It is assumed that a coupling capacitance can be defined between the dot and each electrode. Let us denote $C_{\mathrm{C}}^{\mathrm{b}}$ (resp. $C_{\mathrm{C}}^{\mathrm{t}}$) the capacitance between the dot and the bottom (resp. top) electrode. The sum of these two capacitances is equal to C_{C}. We further assume that $C_{\mathrm{C}}^{\mathrm{b}}$ and $C_{\mathrm{C}}^{\mathrm{t}}$ are classical capacitances, as defined for metallic bodies under mutual influence and that the polarization coupling coefficient α is given by the ratio $\beta = (C_{\mathrm{C}}^{\mathrm{t}}/C_{\mathrm{C}})$. The parallel plate approximation (PPA) consists in approximating the capacitances $C_{\mathrm{C}}^{\mathrm{b}}$ and $C_{\mathrm{C}}^{\mathrm{t}}$ by plate capacitances in which surface S is the projected surface of the dot on each electrode and whose thickness $t_{\mathrm{C}}^{\mathrm{b}}$ (resp. $t_{\mathrm{C}}^{\mathrm{u}}$) is the typical distance between the dot and each electrode. This rough approximation is even used to model single-electron effects such as Coulomb blockade [16]. Note that it is possible to use several surfaces and typical distances to calculate this plate capacitances. Here, we choose two different thicknesses for the same configuration ($t_{\mathrm{C}}^{\mathrm{t}} = t_1$ or $t_{\mathrm{C}}^{\mathrm{t}} = t_1 + a = \lambda_1$ to illustrate the capacitances.

1. The thicknesses of the capacitances are the distance between the border of the dot and the electrodes: $t_C^b = t_2$ and $t_C^t = t_1$. This approximation is named PPA(1).

$$C_C^b(1) = \varepsilon\frac{S}{t_2}, \quad C_C^t(1) = \varepsilon\frac{S}{t_1}, \quad C_C(1) = \varepsilon\left(\frac{S}{t_1} + \frac{S}{t_2}\right)$$

$$\text{and} \quad \beta(1) = \frac{t_2}{t_1 + t_2} \tag{5.31}$$

2. The thicknesses of the capacitances are the distance between the center of the dot and the electrodes: $t_C^b = \lambda_2$ and $t_C^t = \lambda_1$. This approximation is named PPA(2).

$$C_C^b(2) = \varepsilon\frac{S}{\lambda_2}, \quad C_C^t(2) = \varepsilon\frac{S}{\lambda_1}, \quad C_C(2) = \varepsilon\left(\frac{S}{\lambda_1} + \frac{S}{\lambda_2}\right)$$

$$\text{and} \quad \beta(2) = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{5.32}$$

To compare the validity of these approximations, we compare some results given by PPA and results given by the semi-analytical (SA) method. For a given radius, we compare the evolution of the capacitances and different ratios versus the value of t_2. The numerical values are the following: $a = 2.8$ nm, $t_1 = 12$ nm, $\varepsilon_r = 3.9$, and t_2 is varying from 0.5 to 10.5 nm. For the surface S, we have chosen the surface of the disk of radius a as $S = \pi a^2$.

In Figure 5.6, we plot the evolution of the three capacitances C_S (denoted by SA), $C_C(1)$ and $C_C(2)$, divided by the factor $4\pi\varepsilon a$, versus the ratio t_2/a. We also plot the ratios $C_C(1)/C_S$ and $C_C(2)/C_S$. The behavior given by the approximation (1) is qualitatively correct. Quantitatively, this is not very convincing. Indeed for $t_2/a < 1$, the ratio $C_C(1)/C_S$ is not quite constant. It varies from 0.7 to 0.2. There is no way to fit the values of C_S by $C_C(1)$ just by adjusting the numerical value of S. PPA(2) seems qualitatively and quantitatively more interesting to estimate C_S. For $t_2/a < 1$, the ratio $C_C(2)/C_S$ is varying from 0.12 to 0.14 (a variation about 20%).

For large values of t_2/a, the behavior of PPA(1) and PPA(2) are asymptotically equivalent. The asymptotic limit for $t_2 \gg a$ is $a/(4t_1) \sim 5.83 \times 10^{-2}$ for $C_C(1)/(4\pi\varepsilon a)$ and $a/(4(t_1 + a)) \sim 4.73 \times 10^{-2}$ for $C_C(2)/(4\pi\varepsilon a)$. In the figure, we see the beginning of this asymptotic behavior. Because the limiting value of t_2, in our simulation, is of the same order of t_1, the smallest values seen on the figure are approximatively twice the limiting value. For $t_2/a > 2$, the ratio $C_C(2)/C_S$ remains quite constant, varying from 0.11 to 0.09. The ratio $C_C(1)/C_S$ varies from 0.15 to 0.1.

Finally, in all the range of values of t_2/a presented here, only the PPA(2) allows to approximate C_S with finite relative error. The difficulty consists in finding a numerical value for the surface S that makes PPA match the capacitance.

To end this comparison, the dimensionless coefficients κ_1, $\beta(1)$, and $\beta(2) = \alpha$ are compared. For great values of t_2/a, $\beta(1)$ tends to 1 while $\beta(2)$ tends to 0.5. The limiting value of κ_1 is also 0.5 (because C_2 tends to the capacitance of a single half-sphere). On the other hand, for small t_2/a, the limiting value for $\beta(2)$ is $a/(t_1 + 2a) \sim 0.15$, while $\beta(1)$ tends to 0 and C_1 also to 0 (for small t_2, only C_2 contributes to the value of C_S).

In Figure 5.7, we plot the coefficients κ_1, $\beta(1)$, and $\beta(2)$ versus t_2/a for the same previous configuration. On this figure, it appears that the coefficient $\beta(2)$ underevaluates κ_1 for $t_2/a > 0.25$. The greatest relative difference between κ_1 and $\beta(2)$ is observed for $t_2/a \sim 1$ and is about 30%. For $t_2/a > 3$, the relative difference is lower than 10%. The coefficient $\beta(1)$ is far from κ_1, even in the range $t_2/a \ll 1$.

For values of t_2/a in the range of the unity, neither PPA(1) nor PPA(2) gives an approximation with an accuracy better than 10%. Only PPA(2) could give a good approximation, if we are able to evaluate an optimized value for the surface S.

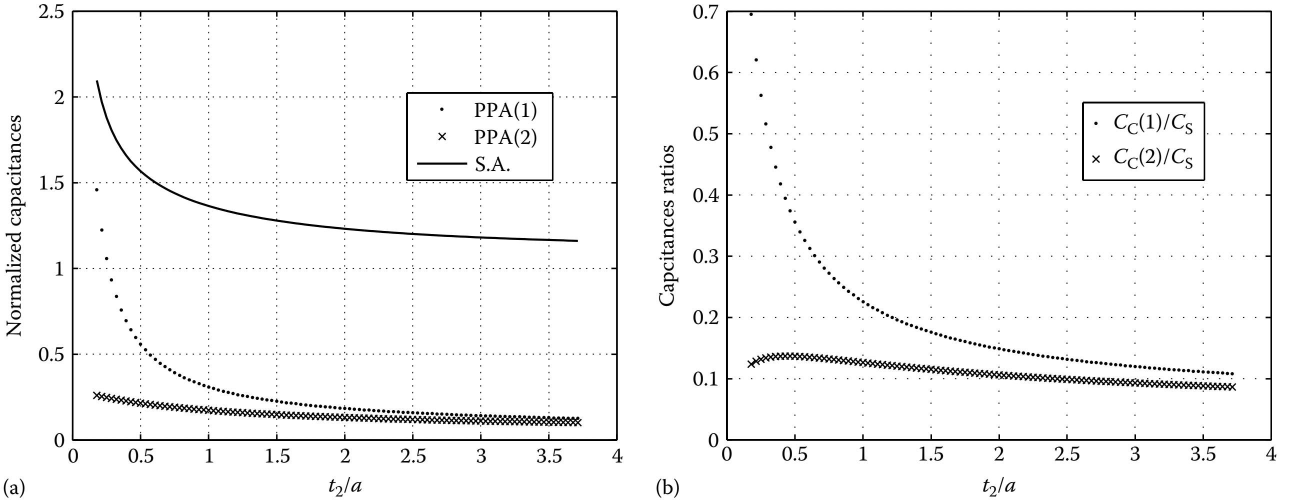

FIGURE 5.6 Comparison between normalized capacitances obtained by the approximations PPA(1) and PPA (2) and by SA. In the first figure, the capacitances are normalized by the factor $4\pi\varepsilon a$. The second figure represents the ratios C_C/C_S for the two approximations. The capacitances and the ratios are plotted versus the ratio t_2/a, for a fixed radius a and variable t_2.

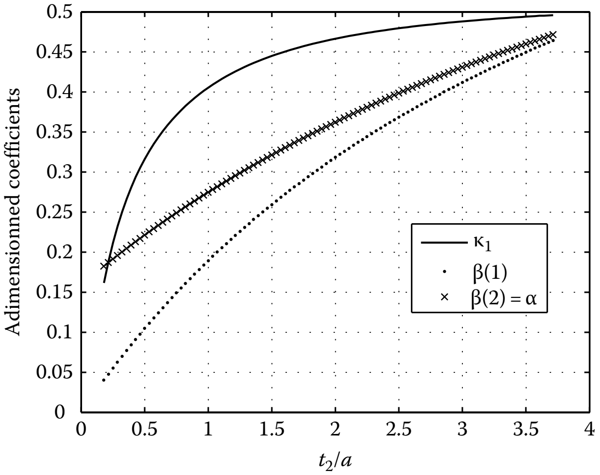

FIGURE 5.7 Undimensioned coefficients κ_1, $\beta(1)$, $\beta(2)$, and α versus the ratio t_2/a.

5.3.6 Conclusion

We devised a semi-analytical model for the potential and the capacitance of a spherical metallic dot located in a dielectric medium between two metallic planes at fixed potential. From this model, we establish a simple analytical expression for the polarization coupling coefficient. We have derived a semi-analytical formulation for the potential everywhere in the dielectric medium. The validity of our approach has been established by comparison with a variational numerical approach. Finally, we have discussed the validity of classical approximations for the capacitance, particularly when the distance between the border of the dot and the nearest electrode has the same order of magnitude as the radius of the dot.

Three major developments are under study. The first one, developed in the next part of this chapter, consists in using this approach in the quantum model of a dielectric dot. First, it allows to generate an initial evaluation of the floating potential used in iterative resolution of Schrödinger and Poisson systems. Second, the comparison between the classical and quantum approaches determines the critical dimensions leading to a quantum behavior. The second development is to adapt this model for dielectric dots. The method of images remains interesting but the spatial distribution of charges in the dot—even in a non-quantum approach—is a serious node of the resolution. The third point is to use the evaluation of the potential around the dot to model the electric field, first step to a modelization of an electronic transport through the dielectric.

5.4 Semiconductor Dot: Theoretical Approach

5.4.1 General Considerations on Quantization Effects

In the metallic case, the electric charge is distributed on the surface of the dot. On the contrary, in the case of a semiconductor dot, the electric charge is distributed in the volume of the nanocrystal and in the surrounding dielectrics. The resolution of such a configuration in a classical approach needs conditions of continuity of the electric field involving model of macroscopic polarization of the dielectric medium. This approach is classically exposed in different works (for instance, Refs. [28,30]). However, when the dot size falls into the nanometer range, classical electrostatics and semiclassical approaches of charge in semiconductor devices are no longer sufficient to describe the charge density and thus the electrostatic potential inside the nanocrystal. Two arguments favor a quantum approach to correctly describe the problem. First of all, let us consider the De Broglie wavelength, λ_B, that describes the wave–particle duality of any particle of mass *m*; it is equal to the ratio of Planck constant, *h*, on the kinetic momentum p. In the case of a free thermal electron of mass m_e, at temperature *T*, De Broglie wavelength is

$$\lambda_B = \frac{h}{p} = \frac{h}{\sqrt{2kTm_e}}$$

where *k* is the Boltzmann constant.

At ambient temperature ($T \sim 300\,K$), λ_B is around 8 nm; this values gives an order of magnitude of the extension of the electronic wavefunction.

In the frame of this work, we will consider semiconductor islands with typical radius around 10 nm, so it becomes a priori necessary to take into account quantum effects to solve the electrostatics of the dot.

Another way to convince ourselves of this latest point is to remind us of some results obtained in the case of an infinite quantum well. Indeed, electrons confined in quantum dots can be roughly compared to the problem of a single particle in an infinite quantum well.

Let us consider an infinite one-dimensional quantum well of width 2*a*. Due to the strong confinement of the wavefunction into the well, the only possible solution of Schrödinger equation inside the dot are standing waves with wave number, k_n, equal to

$$k_n = \frac{n\pi}{2a} \tag{5.33}$$

The non-normalized density of probability, $\tilde{\psi}_n(x)$ is given by

$$\tilde{\psi}_n(x) = \sin\left[k_n x - n\frac{\pi}{2}\right] \tag{5.34}$$

From now, the normalized density of probability is denoted by ψ. The normalization is such as the density is unitary.

The energy of the *n*th bound state, with respect to the bottom of the well, is proportional to the square of the wave number:

$$E_n = \frac{\hbar^2 k_n^2}{2m_e} = \frac{n^2 h^2}{32 m_e a^2} \tag{5.35}$$

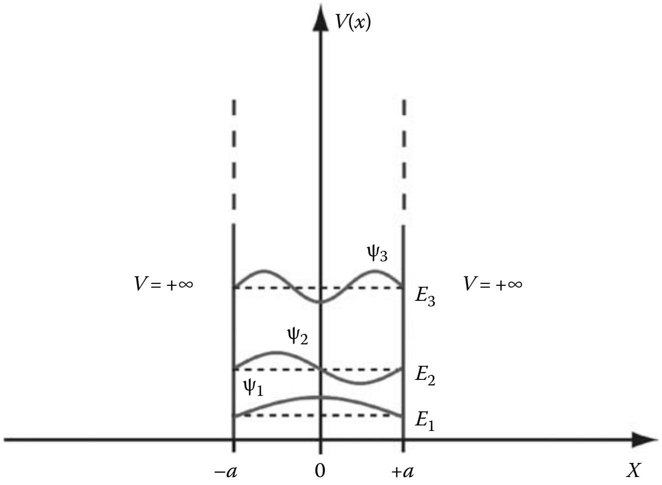

FIGURE 5.8 Schematic representation of an infinite quantum well of width $2a$ and of the three first bound states and their energy.

As an explicative pattern, the eigen wavefunction ψ_n and the level of energy E_n for the three first levels of quantification are plotted simultaneously in Figure 5.8. The evolution of the energy of the first bound state, E_1, is plotted as a function of the dimension of the quantum well in Figure 5.9. If we compare the energy level of the first bound state with the thermal energy, k_BT, we can see that when the quantum well's typical size, a, falls under 2–3 nm, E_1 becomes greater than the thermal energy at ambient temperature (~25.6 meV). From this consideration, we define R_Q the value of the quantum well radius at which the energy shift due to quantum confinement in the dot becomes equal to the thermal energy:

$$R_Q = \frac{nh}{4}\sqrt{\frac{1}{2m_e k_B T}} \tag{5.36}$$

As seen in Figure 5.10, for temperatures ranging from 5 to 300 K, R_Q values are in the 10 nm range, which is the typical size of the quantum dots that we will consider here.

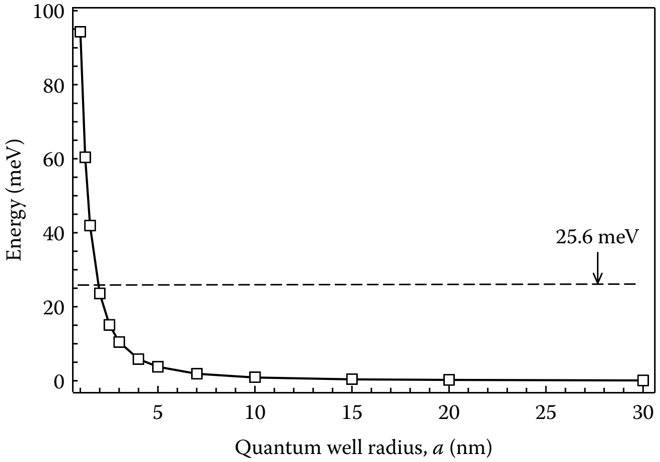

FIGURE 5.9 Evolution of the energy level of the first bound state, E_1, obtained in an infinite well of width $2a$ versus a.

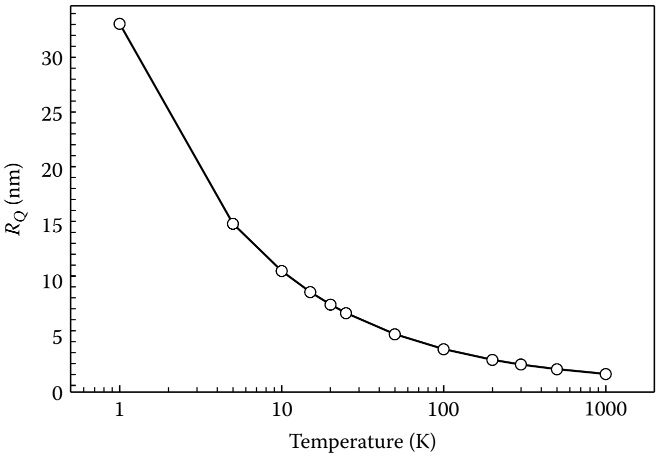

FIGURE 5.10 Evolution of quantum dot radius at which quantization energy becomes greater than thermal energy, R_Q, with temperature.

5.4.2 Quantum Dot Containing a Single Electron: A Simplified Approach

To start with, in this study of the electronic properties of semiconductor dots, we present the main equations describing the behavior of a single electron in the conduction band of a silicon nanocrystal. As explained in the introduction of this section, due to the fact that the De Broglie wavelength is of the same order as the dimension of the nanocrystal, a quantum approach has to be used. Due to the difference in electronic affinity between silicon and silicon oxide, a single silicon dot embedded in a SiO_2 matrix acts as a quantum well for electrons in the conduction band, as shown in Figure 5.11. This system is close to the well-known case of a finite potential well described in many graduate textbooks [33,34]. If analytical expression can be easily found for E_n and $\psi_n(\vec{r})$ in the case of an empty finite or infinite quantum well, it is not the case for a finite quantum well containing several electrons; as a consequence, solving Schrödinger equation is done numerically.

In this chapter, we consider an undoped semiconductor quantum dot embedded in a silicon dioxide matrix. Due to the extremely low concentration of free carriers in intrinsic semiconductor ($n_i \approx 10^{10}$ cm^{-3} for silicon, at 300 K), we assume that there is no free electron in the conduction band of the nanocrystal. Let us make a quick calculation to illustrate this assertion.

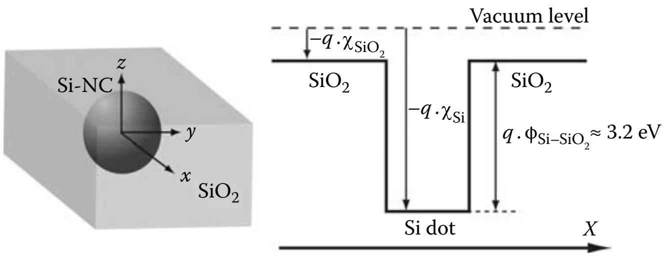

FIGURE 5.11 Energy band diagram representing a silicon nanocrystal embedded in a SiO_2 matrix along one direction.

Let us consider a silicon nanocrystal of radius $a = 10\,\text{nm}$, the intrinsic number of electrons N is

$$N = n_i \times \frac{4}{3}\pi a^3 \sim 4\times 10^{-8} \ll 1$$

For the sake of simplicity, we only consider the bound states existing in the conduction band. We neglect any perturbation arising from the valence band and spin degeneracy. Let us note $E_C(\vec{r})$ the spatial density of the potential energy in the conduction band. Considering the above finite quantum well with a single additional free electron in the conduction band, the spatial evolution of probability density, $\psi(\vec{r})$ is described by the Schrödinger equation in stationary regime. If we look at the first electronic states—at low energy—it is possible to consider that the electron behaves as a quasi-free particle in the dot with an isotropic effective mass m^*_{eff} [35] ($m^*_{\text{eff}} \approx 0.26 m_e$ in silicon):

$$\mathbb{H}\psi(\vec{r}) = E\psi(\vec{r}) \tag{5.37}$$

with

$$\mathbb{H} = \frac{\vec{p}^{\,2}}{2m^*_{\text{eff}}} + V_{\text{conf}}(\vec{r}) \tag{5.38}$$

where

$V_{\text{conf}}(\vec{r})$ is the potential confining the electron evaluated at the point $\vec{r}$ (see Figure 5.11)

$\vec{p}$ is the kinetic momentum operator (classically equivalent to a Laplacian [33,34])

$V_{\text{conf}}(\vec{r})$ must be evaluated with the help of hypothesis, described in the following, for each configuration.

In the case of a single electron in a quantum well, the potential energy $V_{\text{conf}}(\vec{r})$ is only equal to the conduction band energy value at position $(\vec{r})$, $E_C(\vec{r})$. For convenience, we take the bottom of the conduction band energy in the quantum dot as the energy reference.

As there is no other energy contribution in the Hamiltonian $\mathbb{H}$ due to the presence of other electrons in the conduction band, this problem is the same as the finite quantum well. The only available energy states for the electron, are the same (E_n; ψ_n) than those of the finite well. For the following, we suppose that the wavefunction ψ_n has been properly normalized.

The presence probability $\mathcal{P}(\vec{r})$, of the electron at position $\vec{r}$, is thus equal to

$$\mathcal{P}(\vec{r}) = \left|\psi_n(\vec{r})\right|^2 \tag{5.39}$$

Under these conditions, we can built the charge density at any position, $\rho(\vec{r})$, as the product of the electronic charge by the presence probability at $\vec{r}$ (which is also known as the charge density *observable*):

$$\rho(r) = -e \times \mathcal{P}(\vec{r}) = -e \times \left|\psi_n(\vec{r})\right|^2 \tag{5.40}$$

where e is the fundamental charge. As described above, charge density depends on the bound state occupied by the electron. Once the wavefunction of the bound state is known, a numerical resolution of the Poisson equation using the above charge density gives the electrostatic potential of the structure and the evolution of the conduction band energy. For our problem, we make the assumption that the electron, without any external perturbation, is on the first available quantum state, i.e., the fundamental bound state at energy E_1 with the normalized presence probability density $\psi_1(\vec{r})$ and solve the Poisson equation to derive the electrostatic potential, $V(\vec{r})$, in the whole structure:

$$\Delta V(\vec{r}) = -\frac{\rho(\vec{r})}{\varepsilon_0 \varepsilon_r} \tag{5.41}$$

where

ε_r is the relative dielectric permittivity of the material

Δ is the Laplacian operator

5.4.3 Addition of a Second Electron in the Dot

In this section, we consider the previous system made of a silicon nanocrystal embedded in a SiO_2 matrix containing a single electron on the first bound state. When a second electron is added in the conduction band, the probability density, ψ, rigorously becomes a function of two positions:

$$\psi_n = \psi_n(\vec{r}_1, \vec{r}_2) \tag{5.42}$$

Under these conditions, the Hamiltonian of the system is written as

$$\mathbb{H} = \frac{\vec{p}_1^{\,2}}{2m^*_{\text{eff}}} + \frac{\vec{p}_2^{\,2}}{2m^*_{\text{eff}}} + V_{\text{conf}}(\vec{r}_1) + V_{\text{conf}}(\vec{r}_2) + \frac{q^2}{4\pi\varepsilon_0\varepsilon_r \left\|\vec{r}_1 - \vec{r}_2\right\|^2} \tag{5.43}$$

where $\vec{p}_1$ and $\vec{p}_2$ are the kinetic momentum operators of each electron. At this stage, it is very important to notice that an energy term has been added in the Hamiltonian. This term is due to the *Coulombian interaction* of electrons with each other and corresponds to the *electrostatic energy*. Including the interaction term in the Hamiltonian and considering two position variables, solving the problem becomes very heavy, even numerically. Moreover, when we consider further more than two electrons in the nanocrystal, the problem quickly become a N-body problem. For this reason, some approximations are required to be developed further. In the following, we use an approximation that allows to write the wavefunction in a simpler way.

In the Hartree approximation [36], we make the assumption that each electron can be described by an individual wavefunction, respectively $\psi_1(\vec{r})$ and $\psi_2(\vec{r})$, as if they were *independent*. Therefore, the global wavefunction can be written as the product of these individual wavefunctions:

$$\psi(\vec{r}_1, \vec{r}_2) = \psi_1(\vec{r}_1) \times \psi_2(\vec{r}_2) \tag{5.44}$$

In this approximation, we calculate ψ_1 and ψ_2 in such a way that the *total energy* of the system is *minimal*. Moreover the separation of the wavefunction allows to use a *mean field* approximation in which we consider that the electrostatic interaction felt by one electron can be described as if it felt the electrostatic potential induced by the other one(s).

The problem now consists in calculating ψ_1 and ψ_2 by solving individual Schrödinger equations with Hamiltonians $\mathbb{H}_1$ and $\mathbb{H}_2$ with

$$\mathbb{H}_{1,2}\psi_{1,2}(\vec{r}_{1,2}) = E_{1,2}\psi_{1,2}(\vec{r}_{1,2}) \tag{5.45}$$

with

$$\mathbb{H}_{1,2} = \frac{\vec{p}_{1,2}^{\,2}}{2m_e} + V_{\text{conf}}(\vec{r}_{1,2}) + V_{\text{inter}_{1,2}}(\vec{r}_{1,2}) \tag{5.46}$$

$V_{\text{inter}_{1,2}}(\vec{r}_{1,2})$ is the term of Coulomb interaction defined in relation (5.43).

If we compare this latest equation with the one used for the single-electron case (5.37), V_{inter} has been added to take account of the Coulomb interaction between both electrons. As V_{inter} is the mean electrostatic potential induced by the antagonist electron, it is a solution of the Poisson equation:

$$\Delta V_{\text{inter}_{1,2}}(\vec{r}_{1,2}) = -\frac{\rho_{2,1}(\vec{r}_{1,2})}{\varepsilon_0\varepsilon_r} \tag{5.47}$$

where $\rho_{2,1}(\vec{r})$ is given by (5.40)

$$\rho_{2,1}(\vec{r}) = -e \times \left|\psi_{2,1}(\vec{r})\right|^2 \tag{5.48}$$

The total energy of the system is equal to the sum of the energy corresponding to the bound states of both electrons *plus* the electrostatic energy arising form the Coulomb interaction between both electrons, E_{inter}:

$$\mathbb{E}_{\text{tot}} = \mathbb{E}_1 + \mathbb{E}_2 + \mathbb{E}_{\text{inter}} \tag{5.49}$$

with

$$\begin{aligned} \mathbb{E}_{\text{inter}} &= \iiint_{\text{space}} -e\left|\psi_1(\vec{r})\right|^2 \times V_{\text{inter}_1}(\vec{r}) \cdot \mathrm{d}^3\vec{r} \\ &= \iiint_{\text{space}} -e\left|\psi_2(\vec{r})\right|^2 \times V_{\text{inter}_2}(\vec{r}) \cdot \mathrm{d}^3\vec{r} \end{aligned} \tag{5.50}$$

From Equations 5.46 to 5.48 it is obvious that this problem can be solved only *self-consistently*; ψ_1 is computed by solving (5.46) using an initial guess for ψ_2. Then, Poisson equation is solved using ρ_1 as charge density and the corresponding potential V_{inter_2} is used in (5.46) to compute ψ_2. This algorithm is iteratively repeated until the total energy $\mathbb{E}_{\text{tot}}$ calculated at the end of each loop converges.

5.4.4 Multiple Electrons in the Dot

In this section, we extend the results presented in the previous section to the case of multiple electrons in the dot with the same methodology. When the quantum dot contains N electrons, Hamiltonian can be derived in a similar way than in the two-electron problem (relation 5.43):

$$\mathbb{H} = \sum_{j=1}^{N} \frac{\vec{p}_j^{\,2}}{2m_{\text{eff}}^{*}} + \sum_{j=1}^{N} V_{\text{conf}}(\vec{r}_j) + \frac{1}{2} \times \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \frac{q^2}{4\pi\varepsilon_0\varepsilon_r \left\|\vec{r}_j - \vec{r}_i\right\|^2} \tag{5.51}$$

where

$\vec{p}_j$ is the kinetic momentum operator of electron j

V_{conf} is the confining potential

the last term of Equation 5.51 is the interaction potential coming from the electrostatic interaction of the N electrons with each other.

Note that the prefactor 1/2 has been added to avoid to count twice the same interaction.

Rigorously, the solution of the problem is given by solving Schrödinger equation with a N-variable wavefunction:

$$\psi = \psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N) \tag{5.52}$$

As mentioned in the previous section, solving this N-body problem needs an approximations on the wavefunction. In the Hartree approximation, the wavefunction can be written as the product of N individual *normalized* wavefunction:

$$\psi(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N) = \prod_{j=1}^{N} \psi_j(\vec{r}_j) \tag{5.53}$$

each wavefunction can now be associated to an Hamiltonian operator describing the behavior of the jth electron:

$$\mathbb{H}_j\psi_j(\vec{r}_j) = \mathbb{E}_j\psi_j(\vec{r}_j) \tag{5.54}$$

with

$$\mathbb{H}_j = \frac{\vec{p}_j^{\,2}}{2m_{\text{eff}}^{*}} + V_{\text{conf}}(\vec{r}_j) + V_{\text{inter}_j}(\vec{r}_j) \tag{5.55}$$

V_{inter_j} refers to the electrostatic interaction between electron j and the mean potential induced by the $(N - 1)$ other electrons.

Assuming that an equivalent charge density, $\rho_j(\vec{r})$ can be defined for each electron:

$$\rho_j(\vec{r}) = -e\times\left|\psi_j(\vec{r})\right|^2 \tag{5.56}$$

the electrostatic interaction potential V_{inter} is solution of the following Poisson equation:

$$\Delta V_{\text{inter}_j}(\vec{r}) = -\sum_{\substack{i=1\\ i\neq j}}^{N} \frac{\rho_i(\vec{r})}{\varepsilon_0\varepsilon_r} \tag{5.57}$$

The total energy of the system is the sum of the individual eigen energies of the electrons *plus* the electrostatic interaction energy:

$$\mathbb{E}_{\text{tot}} = \sum_{i=1}^{N} \mathbb{E}_i + \mathbb{E}_{\text{inter}} \tag{5.58}$$

where

$$\mathbb{E}_{\text{inter}} = -\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\iiint_{\text{space}} q\left|\psi_i(\vec{r})\right|^2 \times V_{\text{inter}_j}(\vec{r})\cdot \mathrm{d}^3\vec{r} \tag{5.59}$$

To solve the entire problem in the Hartree approximation, it is now necessary to self-consistently solve the set of coupled Equations 5.54 and 5.57 in order to minimize the total energy of the system (5.58).

In this study, electrical results presented in the next section are obtained by using the algorithm presented in Figure 5.12. Starting from an initial assumption for wavefunctions (assuming, for example, an homogenous dilution of the electrons in the quantum dot) and for the potential of the dot (metallic approach or empty dot with coupling polarization), the Poisson equation is solved to determine the interaction potential, V_{inter}. This potential is used in the system Hamiltonian (5.55) as well as the confining potential, V_{conf}, to determine the bound states, i.e., their energies, $\mathbb{E}_j$ and the normalized wavefunctions, ψ_j. This process is repeated for the N electrons and then, the total energy of the system, $\mathbb{E}_{\text{tot}}$ is calculated using Equations 5.58 and 5.59. If the difference of the total energy found at the nth and $(n-1)$th iteration is smaller than an user-defined value (ε), then the Poisson equation is solved for all electrons in the nanocrystal, otherwise the resolution restarts until this condition is satisfied.

We notice here that the influence of the initial guess on the final solution is very weak when the number of electrons is low. However, it can become critical in the case where the number of electrons becomes of the order of magnitude of 10.

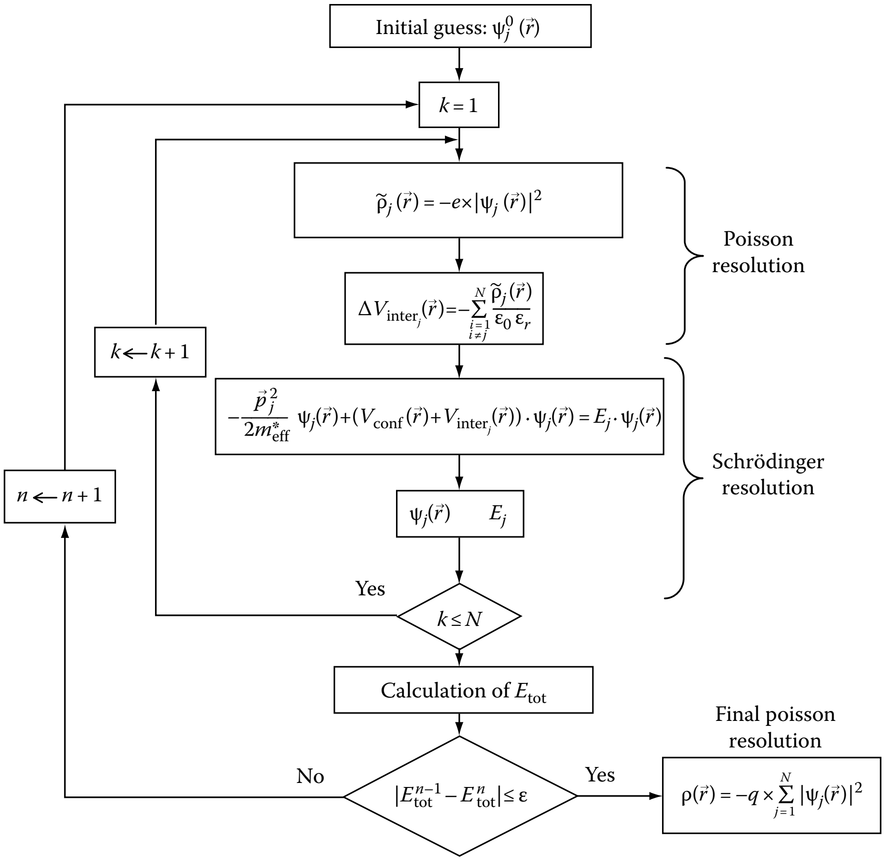
Initial guess: $\psi_j^0(\vec{r})$
$k=1$
$\tilde{\rho}_j(\vec{r}) = -e\times|\psi_j(\vec{r})|^2$
$\Delta V_{\text{inter}_j}(\vec{r}) = -\sum_{i=1, i\neq j}^{N}\frac{\tilde{\rho}_j(\vec{r})}{\varepsilon_0\varepsilon_r}$
Poisson resolution
$k \leftarrow k+1$
$-\frac{\vec{p}_j^{\,2}}{2m^*_{\text{eff}}}\psi_j(\vec{r}) + (V_{\text{conf}}(\vec{r}) + V_{\text{inter}_j}(\vec{r}))\cdot\psi_j(\vec{r}) = E_j\cdot\psi_j(\vec{r})$
$n \leftarrow n+1$
$\psi_j(\vec{r})$ E_j
Schrödinger resolution
Yes
$k \le N$
Calculation of E_{tot}
Final poisson resolution
No
$|E_{\text{tot}}^{n-1} - E_{\text{tot}}^{n}| \le \varepsilon$
Yes
$\rho(\vec{r}) = -q\times\sum_{j=1}^{N}|\psi_j(\vec{r})|^2$

FIGURE 5.12 Algorithm used for the self-consistent resolution of Poisson and Schrödinger equations for multiple electrons in a quantum dot.

5.5 Finite Element Modeling of a Silicon Quantum Dot

In this section, we will present some simulation results obtained on a silicon dot by finite element modeling (FEM). For this study, we used COMSOL Multiphysics© [37] to generate the geometry as well as to solve Poisson and Schrödinger equations [38,39]. Firstly, we will present the studied structure and focus on the methodology used to optimize meshing, which is a crucial point in FEM. Then, the simulation results will be presented, in particular, we will see the influence of the dot radius and the number of electrons on the evolution of the electrostatic potential into the nanocrystal. Finally, we will compare the electrical results obtained in the previous chapter on metallic dots with those obtained in a full quantum treatment of the charge in a semiconductor dot.

5.5.1 Studied Structure and Meshing Strategy

The system studied consists in a spherical silicon dot embedded into a dielectric matrix of SiO_2. The notations are recalled in Figure 5.13. In this work, the thickness of the top and bottom oxides (t_1 and t_2) are respectively fixed to 10 and 5 nm, which are typical values used in the semiconductor industry [17].

As shown in Figure 5.14, the device can be reduced to a quarter of structure for plane-symmetry reasons along the (YZ) and (XZ) planes. This optimization of simulated geometry considerably reduces the computation time that can be advantageously converted into a finer mesh of the structure.

The boundary conditions used for this work are the following:

- Dirichlet conditions on both top and bottom electrodes: electrostatic potential fixed, respectively, to V_1 and V_2
- Symmetry conditions on the symmetry planes of the structure (YZ) (XZ)
- Neumann conditions (i.e., null electric field) on the external planes of the structures parallel to (XZ) and (YZ)

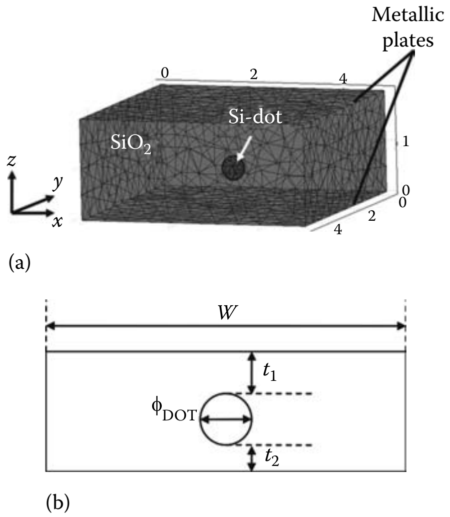

FIGURE 5.13 (a) Studied structure for the FEM simulation of the quantum dot containing multiple electrons. (b) 2D schematic view showing the dimensions of the structure.

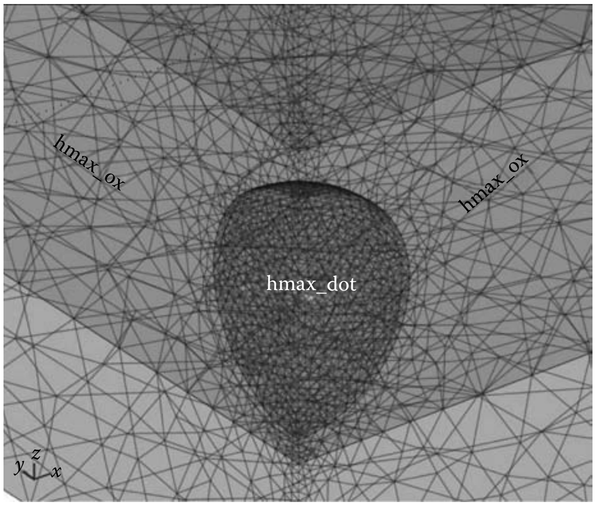

FIGURE 5.14 Detailed view of the quarter-structure studied in FEM showing some relevant meshing parameters considered during the meshing operation: maximum mesh size in the dot (hmax_dot) and maximum mesh size in the oxide (hmax_ox).

For these simulations, we paid a particular attention to meshing which is a crucial point in FEM simulations. From a practical point of view, each simulation result was validated only after having verified that its result remains unchanged after successive mesh refinements [40]. Meshing the structure described in Figure 5.13 is sensitive to many parameters related to the dot size, oxide thickness and the total width of the structure. In each of these domains (dot or silicon dioxide), many parameters such as maximum (resp. minimum) mesh size as well as maximum (resp. minimum) mesh growth rate can be defined. For the sake of simplicity, we will only focus on two main parameters to see the way to optimize them as well as their impact on the numerical results. The parameters studied here are (see Figure 5.14)

- The maximum mesh size in the oxide: hmax_ox
- The maximum mesh size in the dot: hmax_dot

From a general point of view, meshing optimization consists in finding the best compromise between numerical accuracy and computation time by tuning the relevant meshing parameters. Indeed, for a three-dimensional problem the number of mesh points is N^3 where N is the average number of mesh per dimension. So, reducing the number of mesh in regions that do not need much refinement can drastically increase computation time. The two structures depicted in Figure 5.15 show typical results obtained for a set of value for (hmax_ox; hmax_dot) giving a coarse mesh (Figure 5.15a) and a finer one (Figure 5.15b). The natural trend is to refine mesh as much as possible, like in Figure 5.15b, but sometimes it considerably increases computation time or dramatically leads to numerical instabilities.

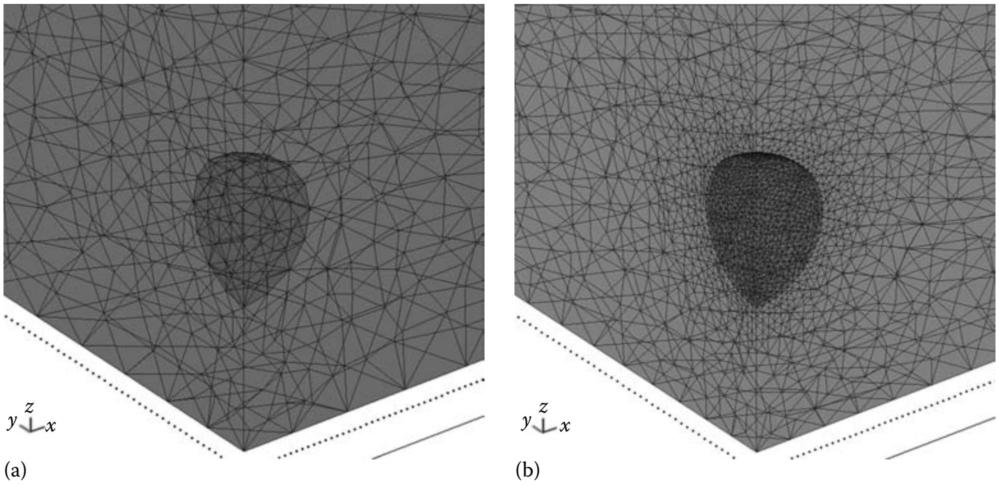

FIGURE 5.15 Comparison of the same structure meshed with different parameters: (a) coarse mesh giving fast computations (hmax_ox = 1 nm, hmax_ox = 3 nm) and (b) fine mesh giving a better numerical accuracy but more time consuming (hmax_dot = 0.25 nm, hmax_ox = 2 nm).

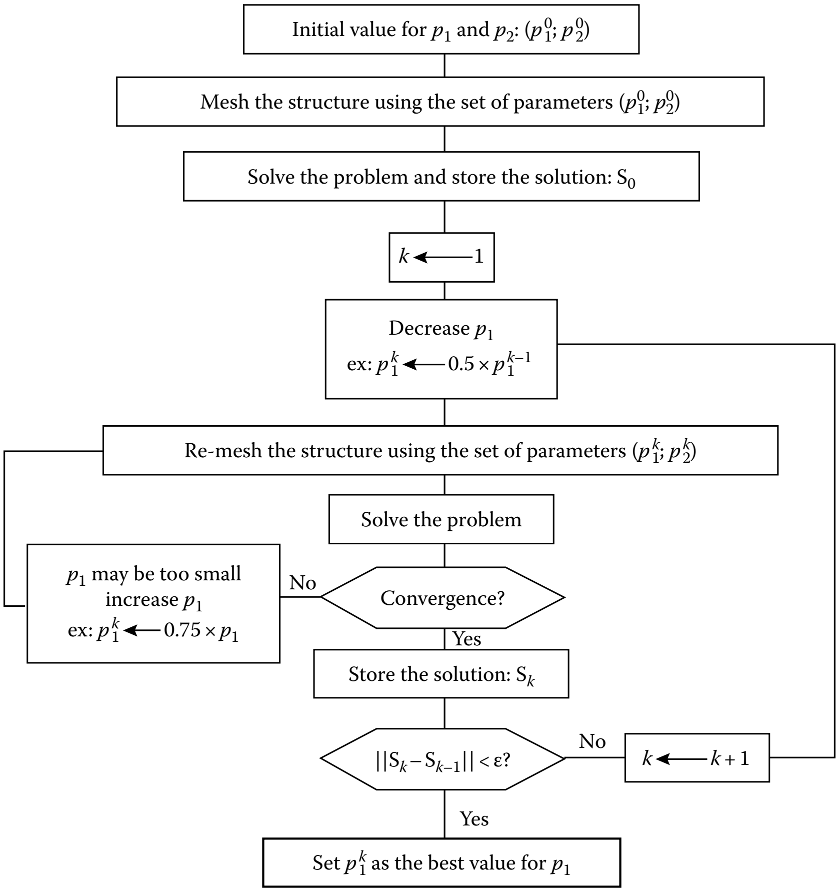
Initial value for p_1 and p_2: $(p_1^0; p_2^0)$
Mesh the structure using the set of parameters $(p_1^0; p_2^0)$
Solve the problem and store the solution: S_0
$k \leftarrow 1$
Decrease p_1
ex: $p_1^k \leftarrow 0.5 \times p_1^{k-1}$
Re-mesh the structure using the set of parameters $(p_1^k; p_2^k)$
Solve the problem
p_1 may be too small
increase p_1
ex: $p_1^k \leftarrow 0.75 \times p_1$
No
Convergence?
Yes
Store the solution: S_k
$||S_k - S_{k-1}|| < \varepsilon$?
No
$k \leftarrow k+1$
Yes
Set p_1^k as the best value for p_1

FIGURE 5.16 Algorithm used to mesh the structure.

As a consequence, for each of the two meshing parameters chosen in our study, a systematic study needs to be carried out to determine the best set of values for (hmax_ox; hmax_dot). To this aim, each parameter is separately adjusted, while the value of the other one is fixed, by an iterative study depicted in the algorithm of Figure 5.16. For each value of hmax_ox (resp. hmax_dot), the convergence is tested and the result is compared to the one obtained during the previous iteration. The value of the studied parameter is reduced until the difference between the simulation results obtained at iteration (*k*) and

$(k - 1)$ is lower than a user-fixed value. At this point, it means than reducing again the value of the studied parameter will not increase numerical accuracy anymore. The optimal set of meshing parameters values can be found by repeating this method as many times as the number of meshing parameters. To conclude with this method, let us mention that the factor used to reduce the studied parameter (i.e., 2) as well as the numerical accuracy criterion (i.e., ε) are very structure-dependent and need to be adjusted to the specificity of each problem.

In the typical case studied in this work, the meshing optimization has been performed on structure shown in Figure 5.14 with an applied bias of 5 V between the top and the bottom electrodes. To quantitatively evaluate the impact of the meshing parameters on the simulation results, the energy value of the first bound state in the dot without any electron inside (E_0) is followed when changing the values of hmax_ox and hmax_dot. Obviously, E_0 is obtained by solving Schrödinger equation in the dot after the meshing step. Note that the zero-energy reference has been taken at the bottom electrode. In Figure 5.17a and b, E_0 values are plotted for different values of hmax_ox and hmax_dot. As can be seen in Figure 5.17a and b, the maximum mesh size in the oxide does not seem to have any quantitative influence on E_0 (the total energy scale represented in the graph is below 1 meV). On the contrary, hmax_ox plays a crucial role in the numerical results; reducing its value from 1 × a (i.e., 2.5 nm) to 0.1 × a (i.e., 0.25 nm) changes the value found for E_0 from ≈5 meV. It is very interesting to notice that beyond 0.1 × a, reducing hmax_dot does not significantly change the numerical result obtained for E_0 meaning that an optimum value can be found around 0.1 × a. At this stage, it is useful to have a look at Figure 5.18 showing the exponential increase of the number of degree of freedom in the structure (i.e., number of mesh points) due to the reduction of the maximum mesh size in the nanocrystal; reducing hmax_dot from 0.1 × a to 0.05 × a makes the number of mesh points increases from ~20,000 to ~1 million units.

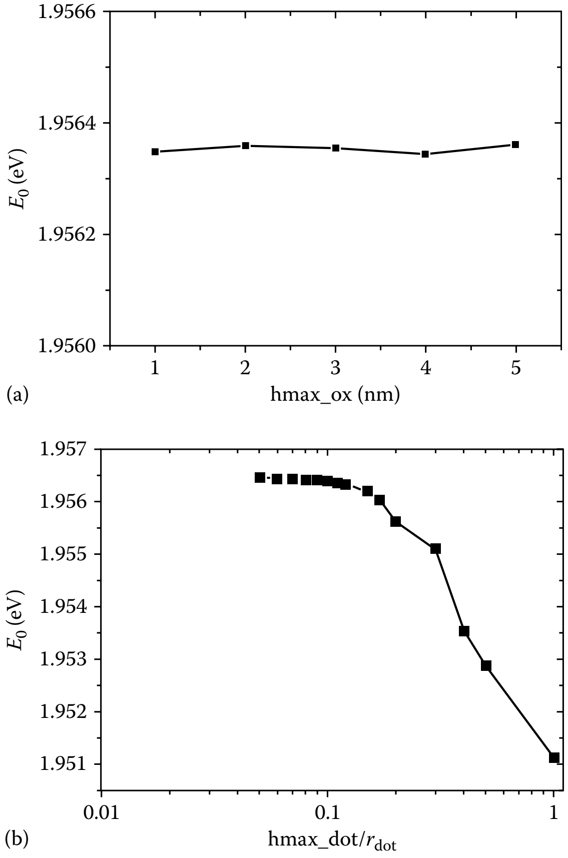

FIGURE 5.17 Meshing optimization. Evolution of the first self-energy found in the dot versus (a) hmax_ox and (b) hmax_dot. Note that values of are normalized with respect to the dot radius a.

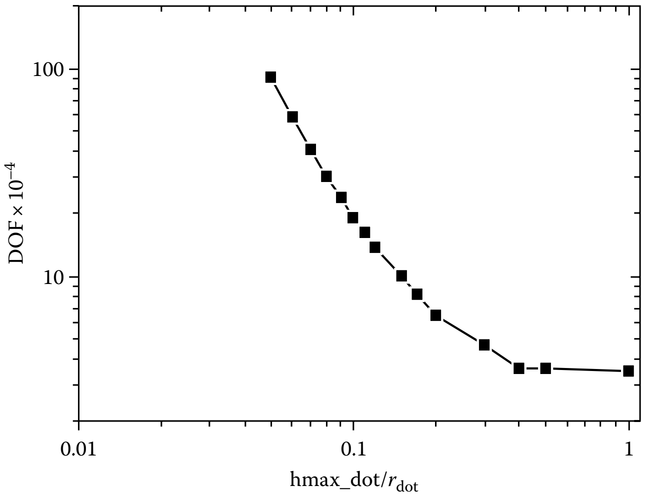

FIGURE 5.18 Number of degree of freedom (DOF) versus normalized maximum mesh size in the dot.

For this work, the criterion chosen to determine the best value for each meshing parameter is that the variation of E_0 should not exceed 0.1 meV between two successive meshing. This criterion corresponds to an arbitrary choice allowing a good accuracy in the energy levels obtained in our problem; in this way, the maximum allowed energy uncertainty remains smaller than the thermal energy at 300 K (~25.6 meV). Using this criterion the value retained for hmax_ox and hmax_dot are, respectively, 3 nm and 0.07 × a.

5.5.2 Numerical Results

Using the algorithm detailed in Figure 5.12 and the structure meshed with the methodology exposed in previous section, Poisson and Schrödinger equations have been solved for an increasing number of electrons in the dot. Figure 5.19a shows the evolution of the charge density in the nanocrystal along the high-symmetry axis. When no external bias is applied, the shape of the charge density is very similar to the wavefunction of the first state in an infinite quantum well. However, when increasing the number of charges in the dot, Coulomb interaction between

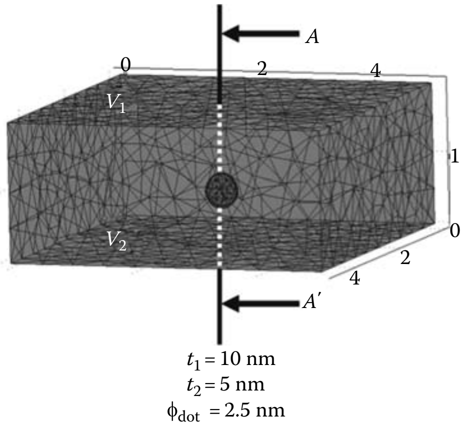

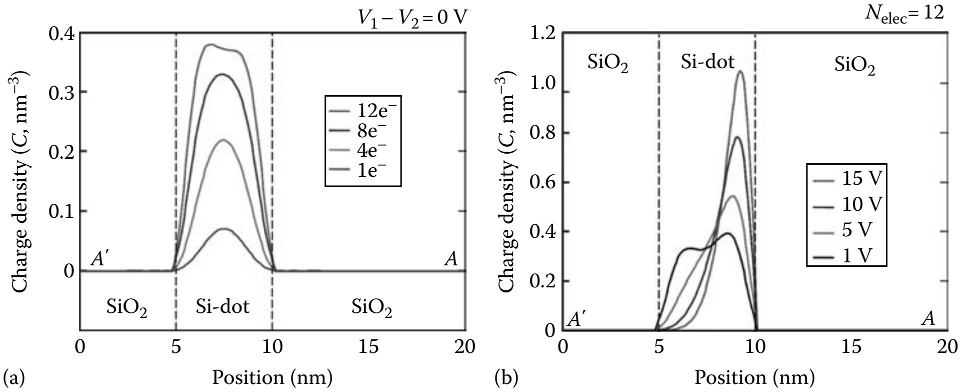

FIGURE 5.19 (a) Charge density in the nanocrystal without any external bias ($V_1 - V_2 = 0$ V) and (b) evolution of the charge density in the dot containing 12 electrons for various external biases.

electrons splits the charge density centroid into two peaks situated on each side of the symmetry plan of the dot (see Figure 5.19a for 12 electrons or Figure 5.19b for 1 V). When an external bias is applied on the electrodes, the charge density is moved in a classical way in the direction of the increasing potential, corresponding to expected electrostatic result. As shown in Figure 5.19b, the more the potential difference increases, the more the charge centroid moves toward the edge of the dot.

Then, using the relationship given in (5.58), the total energy of the system has been computed and reported in Figure 5.20. These results are very close to those obtained by other groups on similar structures [20,21]. As a reminder, this energy takes into account the confinement energy of each electron and the Coulomb interaction between all electrons. Let us remark that the increase in the total energy due to the addition of a single electron significantly increases as the dot radius decreases. As an example, adding a sixth electron in a dot already containing five electrons increases the total energy of 500 meV (~20k_BT at 300 K) if the dot radius is 6 nm. This value goes up to 3 eV when the radius is reduced to 1 nm (~120k_BT at 300 K). As explained in Figure 5.10, this result clearly underlines the role played by the quantum effects in nanostructures and the necessity to use a full quantum approach when the dimension of the system studied is in the nanometer range.

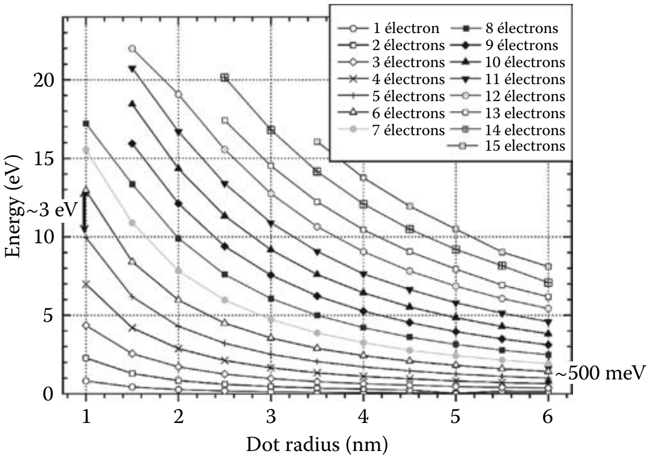

FIGURE 5.20 Evolution of the total energy of the system when varying the dot radius and the number of electrons stored in the nanocrystal.

5.5.3 Comparison with the Metallic Approximation

To conclude this study, let us compare the electrical results obtained in the case of a metallic dot with those obtained in a full quantum treatment of the charge density. Here, we will compare the electrostatic potential obtained in both methods. Let us note that, in the metallic case, the electrostatic potential is constant over the entire dot whereas it varies along the dot volume in the semiconductor cases. For ease of comparison, let us define the mean electrostatic potential, V_{mean}, as

$$V_{\text{mean}} = \frac{1}{Q_0}\iiint_{\text{space}} \rho(\vec{r})\times V(\vec{r})\,\mathrm{d}^3\vec{r} \qquad (5.60)$$

where Q_0 is the total electronic charge carried by the dot

$$Q_0 = -e\times N_{\text{elec}} \qquad (5.61)$$

As defined in Equation 5.60, V_{mean} is nothing else than the mean potential over the whole structure weighted by the charge density.

Figure 5.21 compares the potential of the metallic dot with the mean potential in the semiconductor dot, V_{mean}, for different values of the dot radius, a, and various number of electrons stored in the dot. As it can be seen, there is a small difference between both electrostatic potential values. In each case, the difference observed in the potential is smaller than 200 mV and this difference decreases as the dot radius increase. Even for small dot size, a good agreement can be found when the number of electrons is low; a difference of less than 50 mV is observed when the dot is charged by one or two electrons, whatever the radius is.

This comparison leads to an interesting conclusion regarding the different levels of approximations developed in each approaches. As described in the previous section, the metallic approach gives immediate results thanks to the potential equation once the coupling capacitances between the dot and the electrodes are extracted. On the contrary, the Poisson–Schrödinger solver is an iterative process much more time-consuming but giving a more realistic description of the physical processes occurring inside the nanocrystal. However, both approaches can successfully coexist; we should keep in mind that a Poisson–Schrödinger resolution always start with an initial guess for the Poisson or the Schrödinger problem. Giving the fact that the metallic approach gives electrostatic results quite close to the final solution (see Figure 5.21), the metallic solution can be used as an initial guess for the iterative Poisson–Schrödinger solver to improve the speed of convergence.

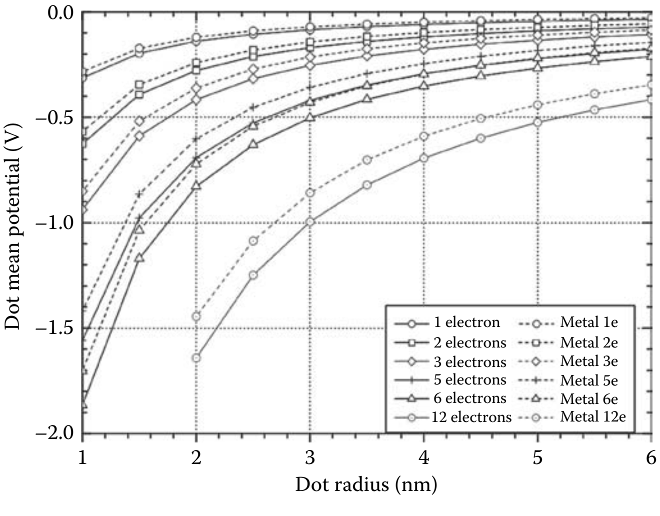

FIGURE 5.21 Comparison between the metallic and the quantum approach regarding the calculation of the electrostatic potential of the nanocrystal. No external bias is applied to the structure ($V_1 = V_2 = 0$ V).

5.6 Conclusion

The purpose of this chapter was to study the electrostatic properties of individual nanodots embedded in a dielectric medium in a 3D configuration. To this purpose, various approaches have been developed and compared. The first one, based on a metallic floating nanoconductor, consist in solving the classical electrostatics equation with a rigorous semi-analytical model. As shown in Section 5.1, this method is able to predict the electrostatic potential of the floating conductor from a simple equation involving the coupling capacitance and the total charge carried by the dot. These results were successfully compared with FEM results obtained on the same structure with an excellent agreement. As a corollary, it should also be stated that the results derived from this approach are not sensitive to the length scale of the system and can be applied to structures of greater (or smaller) dimensions unless they have the same aspect ratio.

In the second part of this chapter, the focus is put on semiconductor nanocrystals. In this case, the quantum effects cannot be neglected any more like in the metallic case. Indeed, it has been shown that when the quantum well dimension falls in the nanometer scale, quantum confinement of the carriers in the nanocrystal is expected, even at ambient temperature. In this case, the electrostatic problem has been solved by a self-consistent resolution of Poisson and Schrödinger equations using the Hartree approximation. The approach developed in this work on a 3D structure by FEM gives the evolution of the charge density (and thus the electrostatic potential) in the volume of the structure with more physical insight than the rough metallic approximation. Moreover, the energy distribution of the carrier can be derived from the energy levels found for each bound state. However, as shown in the last part of this chapter, if we only need the mean electrostatic potential of the nanocrystal, the first metallic approach surprisingly gives very accurate results, even for nanocrystal with dimensions as small as few nanometers, whatever the quantity of charges in the dot.

We should keep in mind these last results as a conclusion. Indeed, they demonstrate that even macroscopic approaches can be used efficiently to predict some physical properties in nanostructures. For the typical case of the electrostatics of a semiconductor dot, there is, a priori no need to handle a complex approach based on a Poisson–Schrödinger solver if the only unknown is the electrostatic potential of the nanodot; the simple metallic approach can give accurate and immediate results if the precision required does not exceed few hundreds of millivolts.

However, if quantum transport between the electrodes and the dot, or even inter-dot transport, has to be studied, the knowledge of the energy distribution of the electrons over their bound-states becomes primordial and will need the use of a quantum treatment of the charge. This is the first point of perspective of this work.

The second evolution of this work would consist in taking into account an irregular shape of the dot. Indeed, in the range of sizes of the current nanodot devices, it is expected that the shape of the dot could have a great influence on the electrostatics behavior. Such a study has been yet proposed in Ref. [24] in a 2D case, combining semi-analytical approach and FEM computations. The results suggest that, extended to a 3D case, a spherical approximation could be sufficient for some configurations.

Appendix A: A Numerical Validation of the Semi-Analytical Approach

5.A.1 Convergence

The tests of convergence of the numerical code are done by repeating a numerical simulation and refining precision until the convergence is satisfactory. For the numerical computation of C_S, there are two numerical truncations in relation (5.15). The first one concerns the required value of the maximum order, $L_{\max}$, ensuring the convergence of the series. The value of $L_{\max}$ is physically constrained by the ratio a/λ_2 (the relative size of the dot seen from the plane). In practice, a value of around a few units is enough to ensure convergence (to fix the idea let us say that $L_{\max} \sim 10 \times a/\lambda_2$ gives good results). $L_{\max}$ leads the discretization over θ that appears in the integration. This integration is assured here by a simple Riemann series with a finite step $\delta\theta$. In practice, $\delta\theta \sim \pi(10L_{\max})^{-1}$ is sufficient. The second truncation is about the number of images taken in the summation over p. We denote by P_0 the greatest natural integer taken for the images. The difficulty of evaluating correctly this summation is the fact that the generic term is alternating because of the factor $(-1)^p$. Doing the summation described in the relations (5.15) and (5.18) for $p = \{-P_0, \ldots, -1, 1, \ldots, +P_0\}$ shows that the convergence is very slow. However, a little numerical arrangement allows a reasonable convergence of the series. Doing the same computation with an odd number of images $p = \{-P_0 -1, -P_0, \ldots -1, 1, \ldots P_0\}$ allows a convergence almost twice faster. An instance of this behavior is shown in Figure 5.22. We compare the results for the capacitance C_S versus the greatest order P_0 in the case of a summation over an even number of images and the case of a summation over an odd one. From $P_0 \sim 30$, the "odd" summation gives result that varies on the fourth digit. The "even" summation gives result still varying on the second digit around $P_0 \sim 40$, and still oscillating on the third digit for $P_0 \sim 70$. This difference between the two types of summation can be clearly explained expressing the potential on the electrodes, highlighting the property of symmetry of each term. The required value of P_0 is constrained by the ratio λ_2/h (the relative distance between the border of the dot and the electrodes), but we have to verify *a posteriori* the validity of this truncation, and did not found heuristic law to fix it.

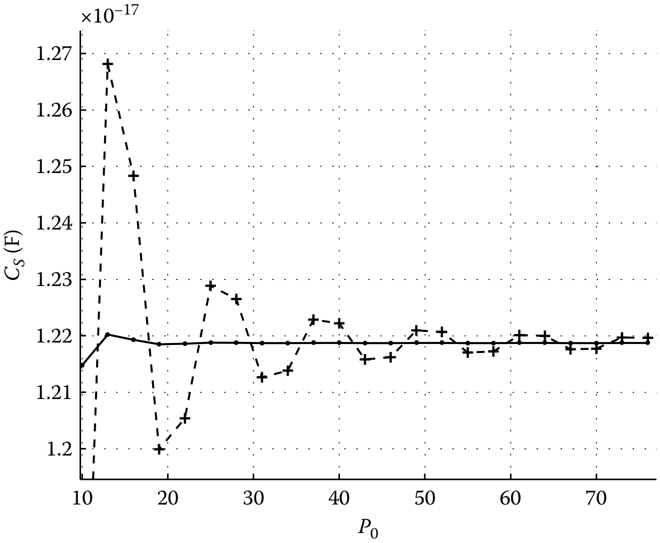

FIGURE 5.22 Behavior of the result for the capacitance versus the greatest order P_0. Slashed and stars, the summation is done for $p^* = \{-57\ldots57\}$. Full and dot, for $p^* = \{-58\ldots57\}$. The configuration is the following: $t_1 = 6\,\text{nm}$, $t_2 = 10\,\text{nm}$, $a = 15\,\text{nm}$. $L_{\max} = 7$.

In Figure 5.23, we plot the result for C_S versus $L_{\max}$ for three different values of P_0. The summation over p^* is made for $p^* = -P_0 -1, \ldots -1, 1, \ldots, P_0$. In the left figure, we see that the influence of P_0 is weak. That is to say that each truncation can be treated separately. In the zoom, in the right side of the figure, we can notice that from $L_{\max} \geq 8$, the result varies only on the fourth significant digit for the both three curves. The curve corresponding to $P_0 = 15$ differs from ~0.3% from the other one corresponding to $P_0 = 75$. The difference between the curve $P_0 = 35$ and $P_0 = 75$ is on the fifth significant digit. In this instance, the dot is particularly big and near from the electrode ($t_1 \sim a/15$ and $2a/h \sim 0.65$). This kind of configuration is the worst case for convergence.

5.A.2 Numerical Validation with FEM Approach

The numerical implementation has been finally validated comparing the results provided by a FEM computation over the same configuration. The capacitance C_S and the coefficient α have been computed for various radius a, keeping t_1 and t_2 constant. The methodology involved for the extraction of C_S and α by FEM calculation is exhaustively described in Ref. [24]. Let us explore this methodology. For each radius, V_f is computed for a given Q_0 and for a given ΔV, by an iterative process stopped on the condition of charge (relation (5.3)). This operation is repeated for each value of Q_0 and ΔV. The linear regressions between Q_0 and V_f (with ΔV constant) and between V_f and ΔV (with Q_0 constant) give, respectively C_S and α. The entire process is repeated for a new radius, giving so the results exposed in Figure 5.24.

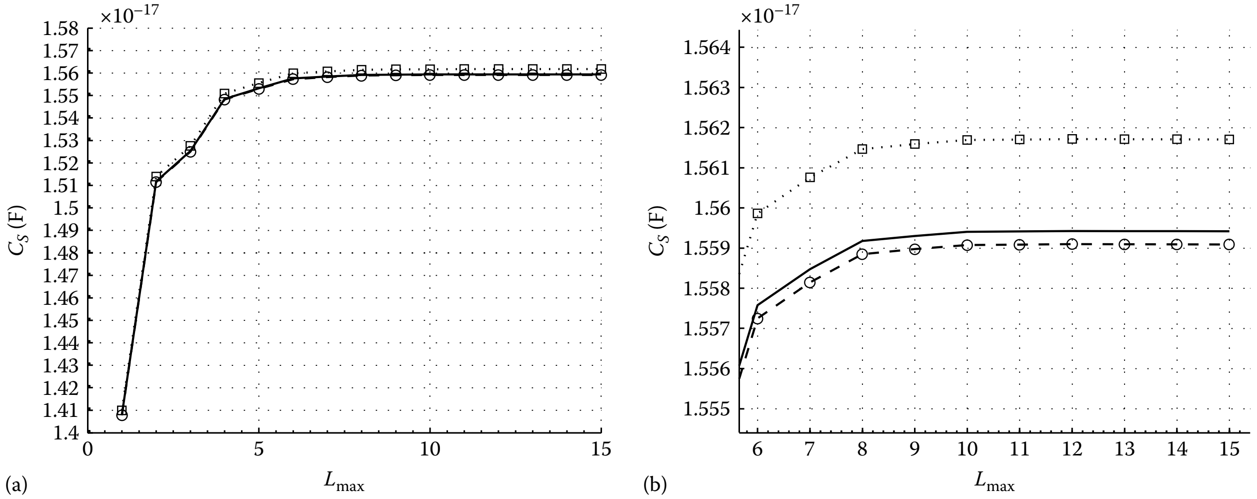

FIGURE 5.23 Behavior of the result for the capacitance versus the greatest order L_{max} for three values of P_0. Slashed and circle, the summation is done for $p^* = \{-76\ldots75\}$. Full, for $p^* = \{-36\ldots35\}$, dotted and square for $p^* = \{-16\ldots15\}$. The configuration is the following: $t_1 = 1$ nm, $t_2 = 10$ nm, $a = 15$ nm. Left-hand side: global representation; right-hand side: zoom for values of $L_{max} \geq 6$.

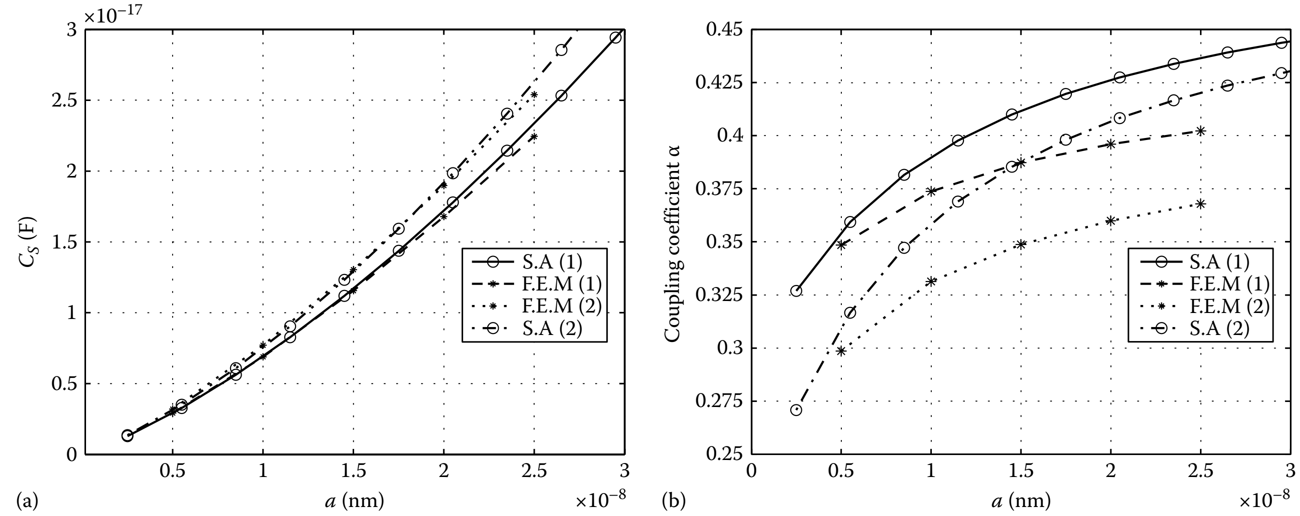

FIGURE 5.24 Comparison of FEM results with semi-analytical expression (denoted by SA). Left: Result for total capacitance versus the radius a for two sets of values for t_1 and t_2, right: Comparison of the coefficient α obtained by FEM and the value λ_2/h versus the radius a. For SA, $L_{max} = 9$ and $P_0 = 80$.

The two configurations are the following:

1. $t_2 = 6$ nm, $t_1 = 15$ nm, $h = t_1 + t_2 + 2a$. We plot C_S versus a, a varying from 0.5 to 30 nm.
2. $t_2 = 4$ nm, $t_1 = 15$ nm, $h = t_1 + t_2 + 2a$. We plot C_S versus a, a varying from 0.5 to 30 nm.

In Figure 5.24 are presented the results obtained by the two methods. Concerning C_S, the agreement between the two methods is better than 1% for values of a up to 20 nm and 2% for higher values. This difference for great values of a may be explained by the methodology involved for the extraction of C_S by FEM calculation. For large values of radius the meshing of the structure becomes a critical parameter. Indeed the space has to be meshed far from the dot and meshed very finely near the dot. The FEM computation is based on the conservation of Q_0. For big sphere, the computed floating potential may be not sufficiently constant, because of the meshing problem. As a conclusion for this analysis, let us insist on the point that the two methods are completely different, and such an agreement is not frequent in numerical simulations.

For the coupling coefficient α, the agreement is not as satisfying as the previous comparison. We observe that the greatest difference between λ_2/h and α obtained by FEM is around 10% for configuration (2) and around 5% for configuration (1). The FEM calculation pains to extract α especially for great values of a and small value of t_2. This fact could be enlighted by the asymptotic behavior of the coupling coefficient for great values of a. Whatever the configuration the coefficient $\alpha = \lambda_2/(t_1 + t_2 + 2a)$ tends to the value 0.5, for large radius. This behavior appears earlier more quickly with SA than with FEM.

This instances seems to show that for great values of a ($a > t_2$), FEM may have some problems of accuracy for computing those two parameters.

To finish this appendix, let us say that a typical calculation of C_S by SA takes few seconds, with a non-optimized.

MATLAB®-like code, while the same extraction with FEM could takes several minutes.

References

1. Moore G.E., Cramming more components onto integrated circuits, *Electronics*, 38 (8), 114–117, April 1965.
2. *2007 ITRS ORTC, Public Conference*, Makuhari, Japan, http://www.itrs.net.
3. Noguchi M., Yaegashi T., Koyama H., Morikado M., Ishibashi Y., Ishibashi S., Ino, K., et al., A high-performance multi-level NAND Flash Memory with 43 nm-node floating-gate technology, in: *Proceedings of the International Electron Device Meeting* (*IEDM*), pp. 445–448, Washington, DC, December 10–12, 2007, doi: 10.1109/iedm.2007.4418969.
4. Ishii T., Osabe T., Mine T., Murai F., Yano K., Engineering variations: Towards practical single-electron (few-electron) memory, in: *Technical Digest of IEDM'00*, San Francisco, CA, December 11–13, 2000.
5. Wang H., Takahashi N., Najima H., Inukai T., Saitoh M., Hiramoto T., Effects of dot size and its distribution on electron number control in metal-oxide-semiconductor-field-effect-transistor memories based on silicon nanocrystal floating dots, *Jpn. J. Appl. Phys*., 40, 2038–2040, 2001.
6. Molas G., Deleruyelle D., DeSalvo B., Ghibaudo G., Gely M., Perniola L., et al., Degradation of floating gate reliability by few electron phenomena, *IEEE. Trans. Electron Devices*, 53 (10), 2610–2619, 2006.
7. Redaelli A., Pirovano A., Benvenuti A., Lacaita, A., Threshold switching and phase transition numerical models for change memory simulations, *J. Appl. Phys*., 103, 111101-1–111101-18, 2008.
8. Sawa A., Resistive switching in transition metal oxides, *Mater. Today*, 11 (6), 28–36, June 2008.
9. Naruke K., et al., Stress induced leakage current limiting to scale down EEPROM tunnel oxide thickness, in: *Proceedings of International Electron Device Meeting* (*IEDM*), pp. 424–427, San Francisco, CA, December 11–14, 1988.
10. Bez R., et al., Introduction to flash memory, *Proc. IEEE*, 91 (4), 489–502, 2003.
11. Tiwari S., et al., Volatile and non-volatile memories in silicon with nano-crystal storage, in: *Proceedings of the International Electron Device Meeting* (*IEDM*), pp. 521–524, IEDM 1995, Washington, DC, December 10–13, 1995.
12. De Salvo B., Gerardi C., Lombardo S., Baron T., Perniola L., Mariolle D., Mur P., et al., How far will silicon nanocrystals push the scaling limits of NVMs technologies? in: *Technical Digest of IEDM'03*, Washington, DC, December 8–10, 2003.
13. Muralidhar R., Steimle R.F., Sadd M., Rao R., Swift C.T., Prinz E.J., Yater J., et al., A 6 V embedded 90 nm silicon nanocrystal nonvolatile memory, in: *Technical Digest of IEDM'03*, Washington, DC, December 8–10, 2003.
14. Choi S., Choi H., Kim T.-W., Yang H., Lee T., Jeon S., Kim C., Hwang H., High density silicon nanocrystal embedded in SiN prepared by low energy (<500 eV) SiH_4 plasma immersion ion implantation for non-volatile memory applications, in: *Technical Digest of IEDM'05*, Washington, DC, December 5–7, 2005.
15. Sattler K., *Handbook of Thin Films Materials*, Volume 5: *Nanomaterials and Magnetic Thin Film*, H.S. Nalwa (ed.), Academic Press, New York, Chapter 2, pp. 61–97, 2002, ISBN: 0-12-512913-0.
16. Averin D.V., Korotkov A.N., Likharev K.K., Theory of single-electron charging of quantum wells and dots, *Phys. Rev. B*, 44 (12), 6199–6211, 1991.
17. De Salvo B., Ghibaudo G., Luthereau P., Baron T., Guillaumot B., Reimbold G., Transport mechanisms and charge trapping in thin dielectric/Si nano-crystals structures, *Solid-State Electron*., 45 (8), 1513–1519, 2001.
18. Han K., Kim I., Shin H., Characteristics of P-channel Si nano-crystal memory, *IEEE Trans. Electron Devices*, 48 (5), 874–879, 2001.
19. Busseret C., Ferraton S., Montès L., Zimmermann J., Granular description of charging kinetics in silicon nanocrystals memories, *Solid-State Electron*., 50 (2), 134–141, 2006.
20. Sée J., Dollfus P., Galdin S., Hesto P., Electronic properties of semiconductor quantum dots for coulomb blockade applications, *Physica E*, 21 (2–4), 496–500, 2004.
21. Sée J., Dollfus P., Galdin S., Theoretical investigation of negative differential conductance regime of silicon nanocrystal single-electron devices, *IEEE Trans. Electron Devices*, 53 (5), 1268–1273, 2006.
22. Cordan A.S., Leroy Y., Leriche B., Electrostatic coupling between nanocrystals in a quantum flash memory, *Solid-State Electron*., 50 (2), 205–208, 2006.
23. Ferry D.K., Goodnick S.M., *Transport in Nanostructures*, Section 4.2, p. 226, Cambridge University Press, Cambridge, U.K., 1999, ISBN: 0-521-66365-2.
24. Deleruyelle D., Micolau G., On the electrostatic behavior of floating nanoconductors, *Solid State Electron*., 52 (1), 17–24, 2008.
25. Ganguly U., Narayanan V., Lee C., Hou T., Kan E., Three-dimensional analytical modeling of nanocrystal memory electrostatics, *J. Appl. Phys*., 99, 114516-1–114516-6, 2006.
26. Godin Yu.A., Electrostatic problem of a conducting sphere in the field of a plane capacitor, *Sov. Phys. Tech. Phys*., 33 (6), June 1988 [*Zh. Tech. Fiz*., 58, 1216–1219, June 1988].

27. Pan Y., Chew W.C., Wan, L.X., A fast multipole-method-based calculation of the capacitance matrix for multiple conductors above stratified dielectric media, *IEEE Trans. Microw. Theory Tech.*, 49 (3), March 2001.
28. Jackson J.D., *Classical Electrodynamics*, 3rd edn., pp. 95–144, Wiley, New York, 1998 (Chapter 3), ISBN: 047130932X.
29. Ngakosso E., Saillard M., Vincent P., Electromagnetic diffraction by homogeneous cylinders of the field radiated by a dipole: A rigorous computation, *J. Electromagn. Waves Appl.*, 9, 1189–1205, 1995.
30. Petit R., *Ondes Electromagnétiques*, Masson Publications, 1988.
31. Aqua J.N., *Physique statistique des fluides coulombiens classiques et quantiques au voisinnage d'une paroi*, Thèse Université Paris XI Orsay, décembre 20, 2000, pp. 23–100, partie B.
32. Chew W.C., *Waves and Fields in Inhomogeneous Media*, Van Nostrand Reinhold, New York, 1990.
33. Cohen-Tannoudji C., Diu B., Laloé F., *Mécanique Quantique*, I Editions, Hermann, Paris, 1998, ISBN: 2-7056-6074-7.
34. Messiah A., *Quantum Mechanics*, Dover Publications, Mineola, NY, 1999, ISBN: 978-0486409245.
35. Sze S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, NJ, 1981, ISBN: 0-471-87424-8.
36. Kittel C., *Quantum Theory of Solids*, John Wiley & Sons, New York, 1987, ISBN: 0-471-62412-8.
37. www.comsol.com.
38. Deleruyelle D., Guiraud A., Micolau G., Study of electronic properties of nanocrystals from different point of view: From a simple 2D electrostatic approach to 3D Poisson Schrödinger simulations, in: *Proceedings of 2007 European COMSOL Conference*, Volume 2, pp. 636–641, Grenoble, France, October 23–24, 2007.
39. Deleruyelle D., Guiraud A., Bassani F., Theoretical and experimental investigation of electronic properties of semiconductor quantum-dots for non-volatile memory applications, in: *Proceedings of Nanoscale VI*, p. 146, Berlin, Germany, July 9–11, 2008.
40. Nougier J.P., *Méthodes de calcul numérique, vol. 2: Fonctions, équations aux dérivées*, pp. 206–212, Paris Hermès Science Publications, Paris, 2001 (Chapter 7), ISBN: 2-7462-0279-4.

6

Nanometer-Sized Ferroelectric Capacitors

6.1 Introduction 6-1
6.2 Fundamentals of Ferroelectricity 6-2
6.3 Deposition and Patterning 6-4
General Aspects • Deposition Techniques • Patterning • Current Trends in Deposition and Patterning
6.4 Characterization of Ferroelectric Films and Capacitors 6-8
Rutherford Backscattering Spectrometry • X-Ray Diffraction for Thin-Film Analysis • Ferroelectric Capacitors: *P-E* Hysteresis Loop Measurements • Scanning Probe Techniques: Atomic Force Microscopy and Piezoresponse Force Microscopy
6.5 Physical Phenomena in Ferroelectric Capacitors 6-13
Strain Effect • Depolarizing-Field Effect • Intrinsic Size Effect in Ultrathin Films
6.6 Future Perspective 6-17
6.7 Summary and Outlook 6-19
References 6-19

Nikolay A. Pertsev
A.F. Ioffe Physico-Technical Institute

Adrian Petraru
Christian-Albrechts-Universität Zu Kiel

Hermann Kohlstedt
Christian-Albrechts-Universität Zu Kiel

6.1 Introduction

Ferroelectric materials possess unique dielectric, piezoelectric, pyroelectric, and electro-optic properties, which make them suitable for applications in various microelectronic and micromechanical devices (Lines and Glass 1977). Owing to the sustained trend toward the miniaturization of electronic devices, ferroelectrics are mostly used nowadays in thin-film form (Dawber et al. 2005, Setter et al. 2006). In particular, the switching of spontaneous polarization in a ferroelectric film is employed in low-power and fast nonvolatile random access memories (FeRAMs) produced for smart cards (Scott 2000, Ishiwara and Okuyama 2004, Kohlstedt et al. 2005a). Owing to their high piezoelectric and pyroelectric responses, ferroelectric thin films are promising for applications in various sensors and actuators and as the heart of microelectromechanical systems (MEMS) that combine Si-technology with the latest achievements in micromechanics (Scott 2007).

Numerous experimental and theoretical studies performed during the past decade have demonstrated that the physical properties of ferroelectric thin films may be very different from those of bulk ferroelectrics. When the film thickness or the lateral size of a ferroelectric capacitor (Figure 6.1) is reduced down to the nanoscale range, these properties generally become size-dependent. Therefore, the physics of nanoscale ferroelectrics represents an exciting field of research in solid-state physics with a close link to device applications. Impressive progress made in this field is mainly the result of the following three recent achievements: (1) new theoretical investigations based on the phenomenological Landau theory of phase transitions (Landau et al. 1984) and the abinitio calculations of ferroelectric crystals and thin films; (2) tremendous improvement in the thin-film technology of complex oxides by using sophisticated deposition techniques and pattering; and (3) the advent of advanced analytical tools. The corresponding “research triangle” is sketched in Figure 6.1.

The phenomenological theory of bulk ferroelectrics was founded in the middle of the twentieth century and proved to be very successful in describing the physical properties of these materials (Devonshire 1954, Lines and Glass 1977). However, this mean-field theory, which is usually termed the Landau–Ginzburg–Devonshire theory, cannot be directly applied to thin-film and nanoscale ferroelectrics. The presence of surfaces (interfaces) and considerable lattice strains in these material systems require significant modifications of this thermodynamic theory (Kretschmer and Binder 1979, Pertsev et al. 1998). In particular, Pertsev et al. (1998) have shown that the substrate-induced lattice strains may give rise to new polarization states and phase transitions in epitaxial ferroelectric thin films. Subsequent experimental studies of single-crystalline films grown on dissimilar substrates confirmed many of the theoretical predictions and demonstrated the importance of strain effects in ferroelectric films (Schlom et al. 2007).

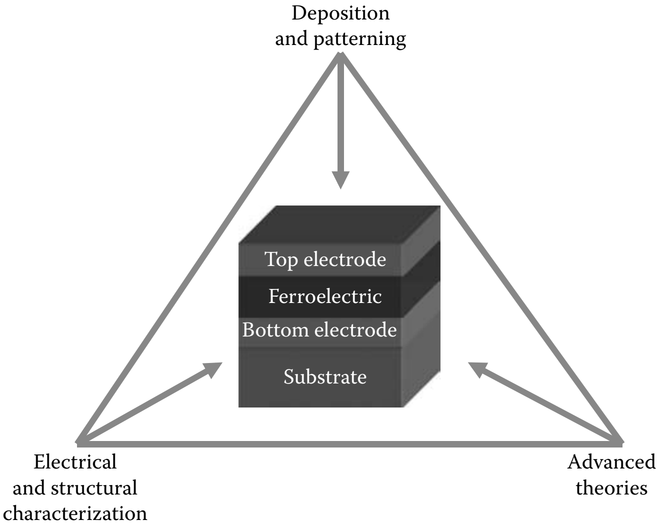

FIGURE 6.1 The "research triangle" summarizing the most important developments in the field of ferroelectric thin films and devices. Inside the triangle, a three-dimensional sketch of a ferroelectric capacitor is shown.

In addition to the advanced thermodynamic calculations, the past decade has been distinguished by the breakthrough in the first-principles density-functional theory investigations of ferroelectrics. Such theoretical studies, which do not involve any empirical input, were first performed for the two classical ferroelectrics, $BaTiO_3$ and $PbTiO_3$, and then were extended to other ferroelectric perovskites (Cohen 1992, King-Smith and Vanderbilt 1994, Waghmare and Rabe 1997). The treatment of ferroelectrics was greatly facilitated by the formulation of the correct definition of the electric polarization as a bulk quantity through the Berry-phase formalism of King-Smith and Vanderbilt (1993) and Resta (1994). This result and the related Wannier-function expression (King-Smith and Vanderbilt 1994) make it possible to calculate the spontaneous polarization and its derivatives after finishing the usual total-energy calculations (see the review paper by Dawber et al. (2005)).

Since the current computational limitations strongly restrict the size of material systems that can be directly calculated from first principles, the modeling based on first-principles results plays an important role in the theoretical studies of ferroelectrics as well. In particular, Monte Carlo simulations using an effective Hamiltonian approach make it possible to extend theoretical predictions to finite temperatures (Diéguez et al. 2004, Lai et al. 2005). It should be noted that the accuracy of first-principles calculations is still limited, which is clearly evidenced by the typical underestimation of the unit-cell size by about 1% and a substantial underestimation of the band gap of insulators. Despite these deficiencies and a considerable disparity between the size of systems studied theoretically and that of real devices, the ab initio calculations represent a very powerful tool for investigating bulk and thin-film ferroelectrics. Furthermore, by combining the first-principles results with the phenomenological Landau–Ginzburg–Devonshire theory, it is possible to develop a fruitful approach making use of the strength of both theories (Gerra et al. 2007, Tagantsev et al. 2008).

In parallel to important developments in the theoretical physics of ferroelectrics, we can notice essential improvements in thin-film growth and characterization techniques. Today, these improvements allow an atomic level of control of ferroelectric inorganic materials such as $BaTiO_3$ and $PbZr_xTi_{1-x}O_3$ (PZT) at the interfaces with lattice-matched substrates ($SrTiO_3$, $NdGaO_3$, $DyScO_3$, etc.) and electrodes made of metal oxides ($SrRuO_3$, $LaSr_xMn_{1-x}O_3$, etc.). Molecular beam epitaxy (MBE), a sophisticated method borrowed from semiconductor technology, pulsed laser deposition, and sputtering, is the work horse in the growth of thin-film ferroelectrics. Essential progress in the structural analysis by transmission electron microscopy and the advent of nanoscale scanning probe techniques of ferroelectric thin films led to a considerable progress in the thin-film deposition, as well as in the understanding of ferroelectrics at the nanometer and angstrom scale (Gruverman and Kholkin 2004, Schlom et al. 2007, Jia et al. 2008).

In this chapter, we first briefly discuss the basic physics of ferroelectric materials and then focus on nanoscale ferroelectric capacitors and future prospects in this research area. Because of the length limit, we restrict our discussion to the class of inorganic ferroelectrics and related devices. Readers interested in ferroelectric polymers, ferroelectric liquid crystals, and ferroelectric II–VI mixed compounds are referred to the relevant publications (Weil et al. 1989, Xu 1991).

This chapter is organized in the following way. In Section 6.2, the fundamentals of ferroelectricity are presented. Section 6.3 is devoted to the state-of-the-art deposition and pattering techniques, whereas Section 6.4 describes the various aspects of the structural and electrical characterization of ferroelectric thin films and capacitors. In Section 6.5, the most important physical phenomena occurring in ferroelectric capacitors are discussed. New developments and emerging trends in the field of nanoscale ferroelectric devices are briefly described in Section 6.6.

6.2 Fundamentals of Ferroelectricity

Ferroelectricity was discovered by J. Valasek in Rochelle salt in 1920 (Valasek 1921). The phenomenological definition of ferroelectricity in the textbook by Lines and Class reads: "A crystal is said to be ferroelectric material when it has two or more orientational states in the absence of an electric field. Any two of the orientation states are identical (or enantiomorphous) in crystal structure and differ only in electric polarization vector at null electric field" (Lines and Glass 1977). The crystal symmetry imposes restrictions on the existence of the spontaneous polarization $\mathbf{P}_s$. First, this property is not allowed in the presence of a center of symmetry, which excludes 11 centrosymmetric groups from the total set of 32 crystallographic point groups. Second, only 10 point groups out of the remaining 21 noncentrosymmetric ones posses a polar axis along which the spontaneous polarization may develop. This subset corresponds to *pyroelectric* crystals, where the build-in polarization manifests itself in temperature-induced changes of the total dipole moment of the unit cell. Ferroelectrics represents a subclass of pyroelectric crystals: they are distinguished from other pyroelectric materials by their ability to switch between two or more stable states with different spontaneous polarization under the action of sufficiently strong external electric field. Indeed, this switching field must be lower than the breakdown field of the material.

Currently, about 700 ferroelectric materials are known (Scott 2007), which can be divided into several groups in accordance with the microscopic origin of ferroelectricity and their atomic structure. The most important examples are ferroelectric oxides such as $BaTiO_3$ and $SrBi_2Ta_2O_9$, hydrogen-bonded ferroelectrics (e.g., KH_2PO_4), ferroelectric polymers like poly(vinylidenefluoride), and ferroelectric liquid crystals (Lines and Glass 1977, Xu 1991). There are many industrial applications of ferroelectric materials, ranging from $BaTiO_3$ capacitors to ferroelectric-liquid-crystal displays for flat TV screens.

In this chapter, we focus on ferroelectric oxides with the perovskite structure, i.e., with the crystal structure characteristic of the mineral $CaTiO_3$ found in the Ural Mountains and named "perovskite." The general chemical formula for perovskite oxides is ABO_3, where A and B represent two cations of very different sizes. In the cubic unit cell of a perovskite oxide, the A-type atoms are situated at the cube corners, the B-type atom sits at the cube center, and oxygen atoms are located at the face centers. In Figure 6.2, the unit cell of the perovskite oxide $BaTiO_3$ is shown schematically for the cubic and tetragonal phases. In the high-temperature cubic state, $BaTiO_3$ is indeed paraelectric with a zero dipole moment of the unit cell. Below about 120°C, stress-free bulk crystals of $BaTiO_3$ become ferroelectric owing to a small shift of the Ti^{4+} ion relative to the center of the surrounding oxygen cage, which leads to the appearance of dipole moment and lattice spontaneous polarization. Since the direction of this shift depends on temperature T, $BaTiO_3$ has three different ferroelectric phases: tetragonal ($10°C \le T \le 120°C$), orthorhombic ($-71°C \le T \le 10°C$), and rhombohedral ($T \le -71°C$) (Jona and Shirane 1962).

The build-in switchable polarization **P** represents the most important physical characteristic of a ferroelectric material. In the thermodynamic theory of ferroelectrics, the polarization vector **P** is used as an *order parameter* and the additional Gibbs free energy density $\Delta G(\mathbf{P})$ of the ferroelectric phase is expanded in terms of the polarization components P_i (i = 1, 2, 3). For a stress-free tetragonal ferroelectric crystal subjected to an electric field **E** parallel to the polar x_3 axis ($P_1 = P_2 = 0$, $P_3 \neq 0$), the Gibbs free energy G_{ferro} can be written as

$$G_{\text{ferro}} = G_{\text{para}} + a_1 P_3^2 + a_{11} P_3^4 + a_{111} P_3^6 + \cdots - P_3 E_3,$$

where a_1, a_{11}, and a_{111} are the dielectric stiffness and higher-order stiffness coefficients of the centrosymmetric paraelectric phase at constant stress. The dielectric stiffness a_1 is given the linear temperature dependence $a_1 = (T - \theta)/(2\varepsilon_0 C)$, where θ and C are the Curie–Weiss temperature and constant and ε_0 is the permittivity of the vacuum. The minimization of $\Delta G(P_3)$ makes it possible to calculate the polarization P_3 of a homogeneously polarized crystal as a function of temperature T and applied field E_3. When the coefficient a_{11} of the fourth-order polarization term is positive, the spontaneous polarization $P_s = P_3$ ($E_3 = 0$) remains zero down to the transition temperature $T_c = \theta$, below which it increases gradually as $P_s \sim \sqrt{\theta - T}$. The energetics of this *second-order* ferroelectric phase transition is described schematically in Figure 6.3b. In contrast, at $a_{11} < 0$ and $a_{111} > 0$, the spontaneous polarization displays a step-like increase up to a value of $P_c = \sqrt{|a_{11}|/2a_{111}}$ at $T_c = \theta + \varepsilon_0 C a_{11}^2/(2a_{111})$. Therefore, a *first-order* ferroelectric phase transition takes place in this situation, as illustrated in Figure 6.3a.

Under an external field directed against the spontaneous polarization, the magnitude of $P_3(E_3)$ first gradually decreases with increasing field intensity E_3. When it reduces down to a certain minimum value $P_{\min}$, the "antiparallel" polarization state becomes unstable. As a result, the polarization switches by 180° into the direction parallel to the applied field. In the case of ferroelectrics with a second-order transition, $P_{\min} = P_s/\sqrt{3}$, and the critical switching field equals $E_{\text{th}} = -(4/3)a_1 P_{\min} \sim (\theta - T)^{3/2}$. The field E_{th} represents the *thermodynamic coercive field* that corresponds to a homogeneous polarization reversal in the whole crystal. The polarization-field curve resulting from the thermodynamic calculations is hysteretic, as shown in Figure 6.4a. The theoretical hysteresis loop is qualitatively similar to the experimental ones, which differ mainly by a gradual polarization

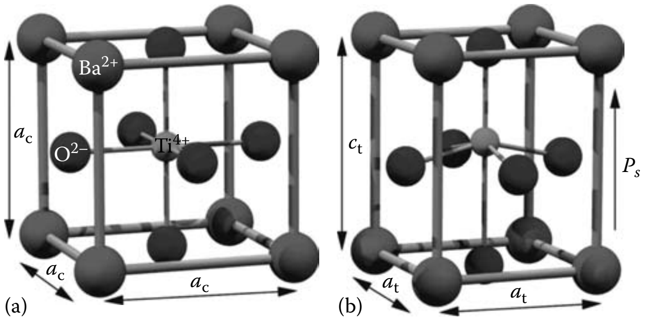

FIGURE 6.2 Schematic representation of the unit cell of $BaTiO_3$ in the paraelectric cubic (a) and ferroelectric tetragonal (b) phases.

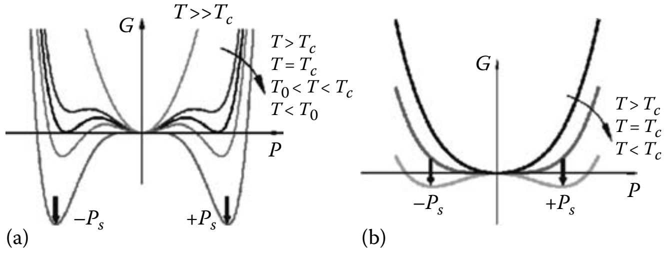

FIGURE 6.3 Temperature evolution of the free energy density as a function of polarization shown schematically for ferroelectrics with the first-order (a) and second-order (b) phase transition.

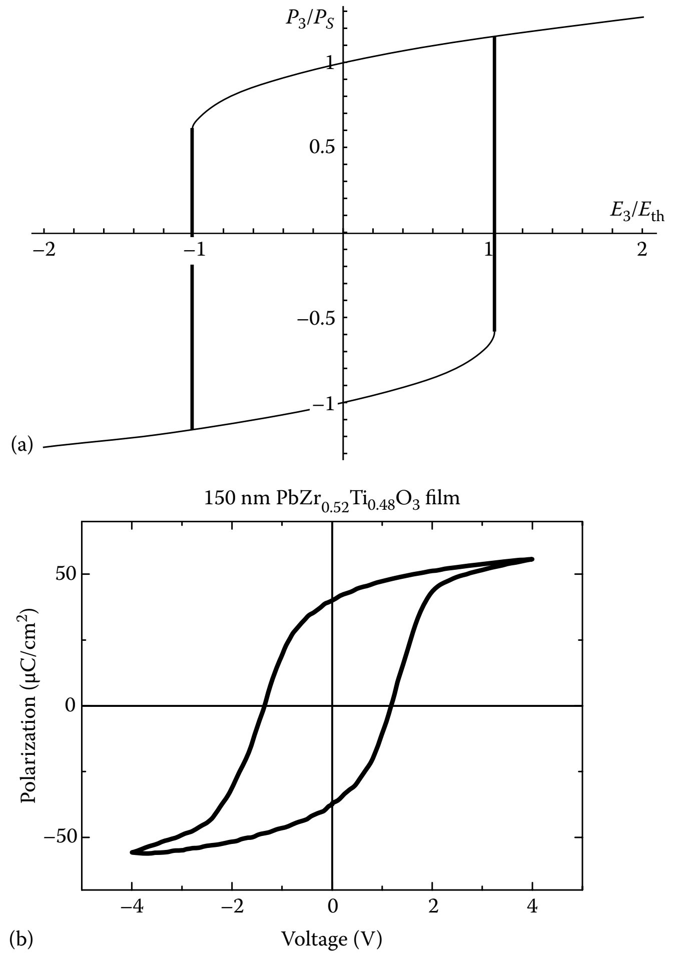

FIGURE 6.4 Theoretical (a) and measured (b) ferroelectric hysteresis loops. The theoretical dependence of the normalized polarization on the normalized applied field corresponds to a homogeneous polarization reversal in the whole crystal. The experimental polarization-voltage loop was measured for the 150 nm thick PZT film.

reversal seen in Figure 6.4b. The measured coercive fields E_c of bulk crystals, however, are typically several orders of magnitude lower than E_{th}, because the polarization switching develops in reality via the nucleation and growth of ferroelectric domains (Lines and Glass 1977).

When the size of a ferroelectric crystal is reduced down to a length scale comparable to the so-called *ferroelectric correlation length*, its physical properties generally become size-dependent. This feature is due to the fact that ferroelectricity is a collective phenomenon resulting from a delicate balance between long-range Coulomb forces (dipole–dipole interactions), which are responsible for the ferroelectric state, and a short-range repulsion favoring the paraelectric state (Lines and Glass 1977). The scaling of physical characteristics such as the remanent mean polarization $P_r = \langle P_3(E_3 = 0)\rangle$ and coercive field E_c is currently in the focus of experimental and theoretical studies in the field of nanometer-sized ferroelectric capacitors. The most important results of these studies will be discussed in Sections 6.4 and 6.5.

6.3 Deposition and Patterning

6.3.1 General Aspects

The deposition of thin films belongs to the heart of today's micro- and nano-electronics. In order to grow thin films with desired properties, several important issues have to be considered. Besides the choice of the deposition method, a number of process parameters are essential, such as the background pressure of the vacuum system, the deposition rate (measured in nm/s), the substrate material and temperature, and the composition of the material source. Furthermore, the process pressure is important, irrespective of the conditions in the chamber, i.e., the state of an ultra-high vacuum or the presence of a noble gas (typically argon) or reactive gases (oxygen or nitrogen). The basic principles of thin-film deposition are described in several textbooks (Chopra 1969, Maissel and Glang 1979, Bunshah 1994). Here we focus on the deposition of complex oxides by means of physical methods.

As already mentioned in the introduction, thin-film deposition techniques for the growth of heteroepitaxial oxides have made tremendous progress recently. The purpose of this section is to give a short overview of the current status of the MBE, pulsed laser deposition (PLD), and sputter deposition (SD). For chemical-based methods, such as metal-organic chemical vapor deposition (MOCVD), atomic layer deposition (ALD), and chemical solution deposition (CSD), we refer the reader to the relevant papers (Oikawa et al. 2004, Kato et al. 2007, Schneller and Waser 2007). We also note that ferroelectric polymers are deposited by either a spin-on technique or the Langmuir-Blodgett method (Ducharme et al. 2002).

Consider a planar ferroelectric capacitor sketched in Figure 6.1. The choice of materials for the substrate, electrodes, and ferroelectric layer strongly depends on the application or the research task. For the epitaxial growth of perovskite ferroelectrics, single-crystalline substrates having small lattice mismatches with these complex oxides are preferable. At present, the commercially available single crystals of $SrTiO_3$ are most popular, although other substrates such as MgO, $KTaO_3$, $GdScO_3$, and $DyScO_3$ have also been successfully employed. When the ferroelectric overlayer is commensurate with a dissimilar thick substrate, it appears to be strained to a certain extent defined by the mismatch in their in-plane lattice parameters. Above some critical thickness, however, these lattice strains start to relax due to the generation of misfit dislocations (see Section 6.4).

As for the electrodes, noble metals such as Pt, Ir (also IrO_2), and Ru are used in most device applications, e.g., in FeRAMs and MEMS (Kohlstedt et al. 2005a). On the contrary, conducting complex oxides were favored so far as electrode materials in the basic research studies of scaling effects. Prominent examples of such electrode materials are $SrRuO_3$, $LaSr_xMn_{1-x}O_3$, and $LaCa_xMn_{1-x}O_3$, which are routinely used in the complex-oxide heterostructures (Eom et al. 1992, Sun 1998). A comparison of the advantages and disadvantages of metal and oxide electrodes shows a delicate trade-off. Let us compare, for instance, Pt and $SrRuO_3$.

On the one hand, the resistivity of a sputtered thin-film Pt at room temperature (~10 μΩ cm) is much smaller than that of even a high-quality $SrRuO_3$ (≈300 μΩ cm). On the other hand, the screening-space-charge capacitance density, which is important for the stabilization of ferroelectricity in ultrathin films (Pertsev and Kohlstedt 2007), is equal to 0.9 F/m^2 for the $SrRuO_3$ electrode and only to 0.4 F/m^2 for the Pt electrode (Pertsev et al. 2007). This feature seems to make $SrRuO_3$ electrodes preferable for nanoscale ferroelectric capacitors. In addition, the electrode surface roughness, crystallographic orientation of the ferroelectric layer grown on a particular electrode, and the quality of the electrode–ferroelectric interface must be taken into account. Currently, conducting complex oxides are preferred for the fabrication of the bottom electrode, whereas the top electrode can be made of Pt or other noble metal as well. In the rest of this section, we focus on entirely complex-oxide heterostructures for ferroelectric capacitors.

6.3.2 Deposition Techniques

6.3.2.1 Molecular Beam Epitaxy

MBE has developed from a simple evaporation technique via the use of ultra-high vacuum (UHV) to avoid disturbances by residual gases and additional incorporation of various effusion (Knudson) cells as material sources. Figure 6.5 schematically shows an MBE system involving several material sources that allow controlled deposition of multi-element compounds. In contrast to the deposition of most semiconductor materials such as GaN, GaAs, and InP, the growth of oxides by MBE requires relatively high partial pressure of oxygen (~10^{-7} mbar) during the deposition. This is necessary to avoid the oxygen deficiency in the final film, which could seriously deteriorate the quality of a ferroelectric capacitor. Partial pumping, the use of reactive oxygen (e.g., ozone), and post-annealing of the films in a high-pressure oxygen atmosphere (several mbar) are used to supply the films with a sufficient amount of oxygen.

Owing to the UHV conditions in the MBE chamber, all UHV surface techniques can be employed. This feature, indeed, constitutes the strength of MBE (Haeni et al. 2000). This tool offers the highest degree of freedom to apply sophisticated in situ analytical techniques to study films during the growth and just after the deposition without breaking the vacuum. Complex MBE systems using low-energy electron microscopy (LEEM) and Auger electron spectroscopy (AES), for example, were developed (Habermeier 2007 and Clayhold et al. 2008). The standard technique currently is the reflection high-energy electron diffraction (RHEED), which allows control of the surface chemistry of the last layer during the deposition. This technique provides an opportunity to fabricate oxide films with a definite termination at the surface (e.g., the BaO or TiO_2 termination in $BaTiO_3$ films). Many groups successfully demonstrated this approach (Logvenov and Bozovic 2008) with similar oxide materials. An obvious research goal now is to find correlations between atomic terminations at ferroelectric–metal interfaces and the electrical properties of capacitors.

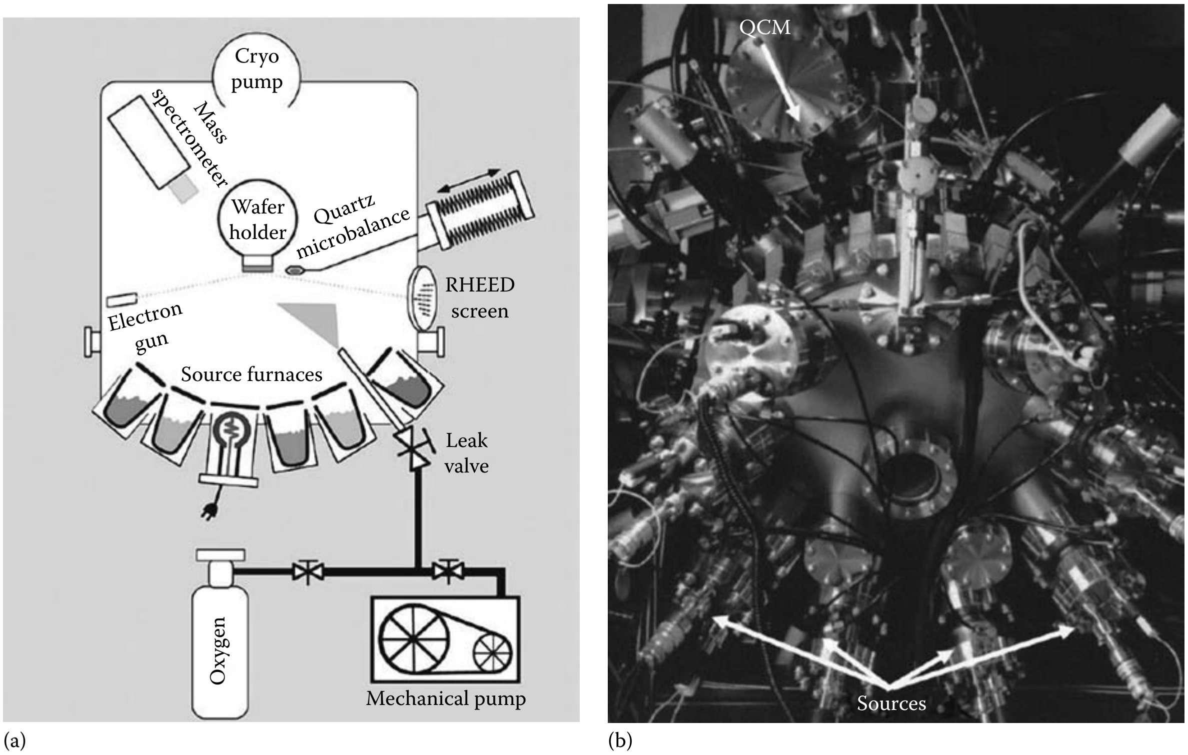

(a) (b)

FIGURE 6.5 Schematic view of MBE system for the growth of multi-element-compound thin films (a) and a photograph of MBE chamber (b). (From Lettieri, J. et al., *J. Vac. Sci. Technol. A*, 20, 1332, 2002. With permission.)

An additional important feature of MBE is the low energy of deposited species. Indeed, the temperature of the material source in effusion cells or electron-beam evaporators does not exceed 3500 K. Hence, the corresponding thermal energy of the deposited species is about 300 meV. This value is an order of magnitude lower than the energies characterizing pulsed laser deposition and typical sputter deposition. Among available deposition techniques, MBE is the most flexible one with respect to the incorporation of analytical tools and offers the highest degree of atomic layer control. On the other hand, MBE systems are difficult to handle, considerable time is necessary for their maintenance, and, last but not least, special methods are needed to supply complex oxides with a sufficient amount of oxygen. Due to the highly complex machinery, experienced researchers working with MBE systems for years translated the acronym MBE as "many boring evenings."

6.3.2.2 Pulsed Laser Deposition

PLD is a very useful and flexible tool for growing oxide materials (Hubler and Chrisey 1994). A sketch of a PLD system is shown in Figure 6.6. A pulsed laser beam, from a KrF (248 nm) or ArF (193 nm) excimer laser, for example, is focused on a rotating target made of an oxide material (e.g., $BaTiO_3$, PZT, or $SrRuO_3$). The oxygen gas pressure during ablation can be varied from 10^{-7} to 0.5 mbar. Owing to the intense laser beam, a plasma containing energetic ions, electrons, neutral atoms, and molecules is formed. The energy density is in the range of 2–5 J/cm^2 at the target surface. As a result, the energy of the ablated material may reach values well above 10 eV at the substrate surface. The wavelength of the used laser beam may be 248 or 193 nm (at this UV wavelength, the absorption in the oxide target materials is sufficiently large). A repetition rate of several Hz and a pulse length of 25 ns represent typical parameters. Initially, a serious problem of the method was the formation of droplets on the substrate, which can easily deteriorate the device properties. Currently, various methods to reduce this effect are known, e.g., the time-of-flight selection of ablated material. PLD systems are widely used for basic research studies of thin films. The main advantage of this method is that a certain film stoichiometry can be easily achieved by PLD. Many targets (six or more) can be placed on the target carrousel holder. In this way, numerous materials can be deposited without time-consuming rearrangements of the deposition chamber or complicated source exchange procedures, as in the case of MBE or MOCVD.

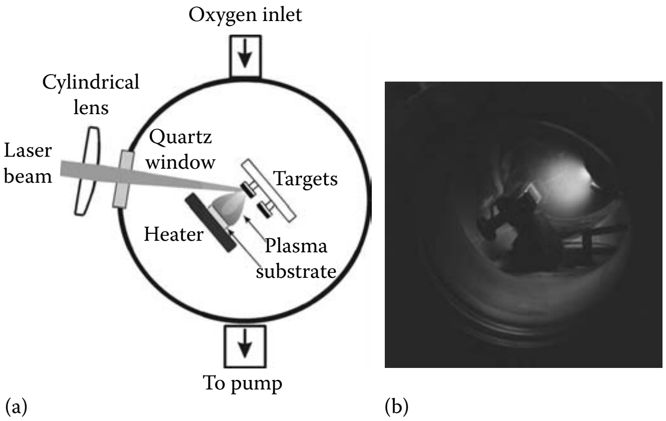

(a) (b)

FIGURE 6.6 Pulsed laser deposition system for the growth of oxide films: PLD setup scheme (a) and a photograph showing the ablation process and the plasma plume (b). (From Schlom, D., Cornell University, Ithaca, NY.)

6.3.2.3 Sputter Deposition

Plasma sputtering is a physical vapor deposition technique that has been known for 150 years since the time when W.R. Grove first observed the sputtering of surface atoms. Different sputtering techniques, such as dc- and rf-sputtering with or without a magnetron arrangement, have been used to grow a variety of materials. Figure 6.7a illustrates the principle of dc-sputtering. A potential of several hundred volts is applied between the target (cathode) and the heater (anode), accelerating positively charged ions toward the target. These accelerated particles sputter off the target material, which finally arrives at the substrate. The discharge is maintained because the accelerated electrons continuously collide with the gas circulating in the chamber and ionize new atoms.

For insulating targets such as ferroelectric ones, the dc-sputtering is not suitable. Insulating targets have to be sputtered using alternating electric fields to generate the plasma. Typically, an rf-frequency of 13.56 MHz is employed. This frequency is not a magic number, rather a frequency that is approved by the government for industrial purposes. A symmetrical arrangement of cathode and anode and the use of a low-frequency alternating field would result in similar sputtering and re-sputtering rates so that the film will not grow. In the case of a high-frequency alternating field, however, light electrons can respond to the field at this frequency, whereas heavy Ar^+ ions see only an average electric field (Kawamura et al. 1999). Moreover, the geometrical asymmetry between small cathodes (target side) and large anodes (heater and chamber) leads to a higher electron concentration at the former, resulting in a self-generated dc bias that accelerates Ar^+ ions toward the target.

The high-pressure sputtering technique of oxide materials was developed by Poppe et al. (1988) and served initially for the

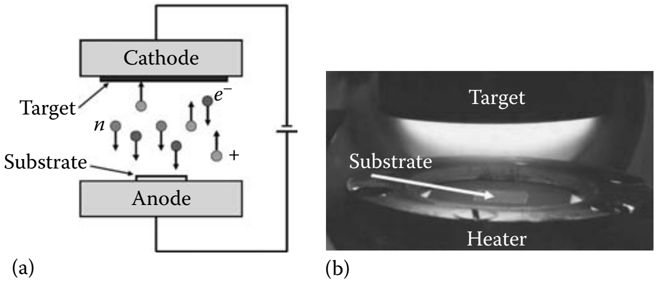

(a) (b)

FIGURE 6.7 Schematic representation of the dc-sputtering process showing the electrons, positive plasma ions (light gray), and neutral atoms (dark gray) moving toward cathode or anode (a) and a photograph of a high-pressure sputtering system (b). (From Rodríguez Contreras, J. et al., PhD thesis)

growth of oxide superconductors. A planar on-axis arrangement of the target and substrate is used, as shown in Figure 6.7b. A high sputtering pressure of 2.5–3.5 mbar, corresponding to a mean-free-path $\lambda_{\text{mean free}} = 6 \times 10^{-3}$ cm at 600°C and exceeding largely the pressure of 10^{-2} mbar used for conventional sputtering ($\lambda_{\text{mean free}} = 2$ cm at 600°C), leads to multiple scattering of the negatively charged oxygen ions accelerated toward the substrate. As a result of the thermalization of ions, the re-sputtering of the deposited films, which is caused by negatively charged ions, is negligible. This technique yields excellent thin films due to the low kinetic energy (as in the case of MBE) of sputtered particles.

A disadvantage of the high-pressure sputtering technique could be a low deposition rate of several nanometers per hour, which may lead to interdiffusion at heterogeneous interfaces. To enhance the deposition rate, a low ionization degree of less than 1% of the atoms in the plasma is increased by the use of magnetic fields forcing electrons onto helical paths close to the cathode, which leads to much higher ionization probability. This so-called magnetron sputtering can be employed for high-pressure sputtering (Poppe et al. 1988), as well as for conventional, low-pressure sputtering (Fisher et al. 1994). Sputtering is routinely used as a vapor deposition method for the growth of complex-oxide films.

6.3.3 Patterning

The patterning of oxide heterostructures represents an important step in the fabrication of ferroelectric capacitors with small lateral dimensions ranging from a few micrometers to tens of nanometers. As in many other areas of nanoelectronics, two different approaches exist for the device fabrication. A conventional approach relies on the well-established processes used in the modern semiconductor industry: deposition, lithography, and etching. Using these techniques sequentially, ferroelectric capacitors can be fabricated. It should be emphasized that this fabrication procedure employs the so-called *top-down* approach, where external tools are used to create a nanoscale device out of a larger structure. In contrast, the *bottom-up* approach is based on the self-organization of constituents or their positional assembly necessary for a desired nanodevice. Such techniques recently became very fashionable (Spatz et al. 2000) because they do not require advanced and expensive patterning tools. The bottom-up approach has had considerable success, but improvements are needed to achieve registered arrays of devices, such as those produced by the state-of-the-art complementary metal-oxide-semiconductor (CMOS) technology. In some works, mixed top-down and bottom-up methods were used to produce nanoscale ferroelectric dots and crystals (Kronholz et al. 2006, Szafraniak et al. 2008).

One of the simplest ways to fabricate ferroelectric capacitors is the lift-off technique. The main steps of this technique are shown in Figure 6.8a. First, the bottom electrode and the ferroelectric layer are deposited on a substrate. A subsequent photo-lithographic step defines the area of the capacitor. Next, the top electrode is deposited, for example, by the sputtering of Pt. After a lift-off in acetone, the metal with photoresist underneath is removed and the capacitor is ready for electrical characterization. Because the top interface is subjected to photoresist and chemical developer during this procedure, relatively poor electrical properties (e.g., large leakage) are observed here (Rodríguez et al. 2003a and Rodríguez Contreras 2004). The post-annealing of capacitors at high temperatures and in an oxygen atmosphere was successfully used to improve the electrical properties considerably (Schneller and Waser 2007).

The aforementioned drawback, however, can be avoided using another method, which involves the fabrication steps shown schematically in Figure 6.8b. Here the whole sandwich (bottom

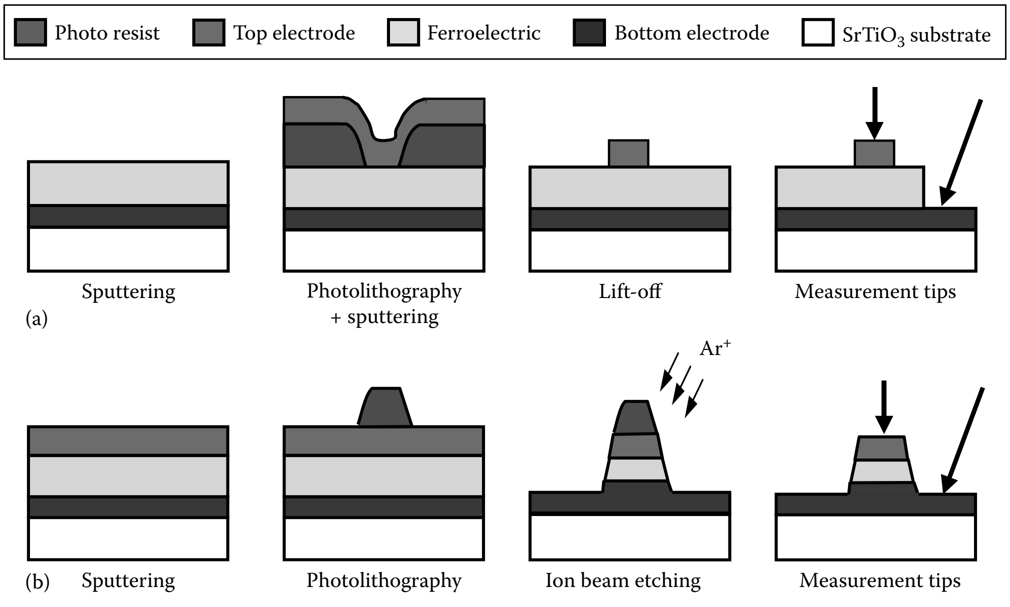

FIGURE 6.8 Patterning of capacitors by lift-off technique (a) and ion beam etching (b). (Reprinted from Rodríguez Contreras, J. et al., *Appl. Phys. Lett.*, 83, 126, 2003a. With permission. American Institute of Physics.)

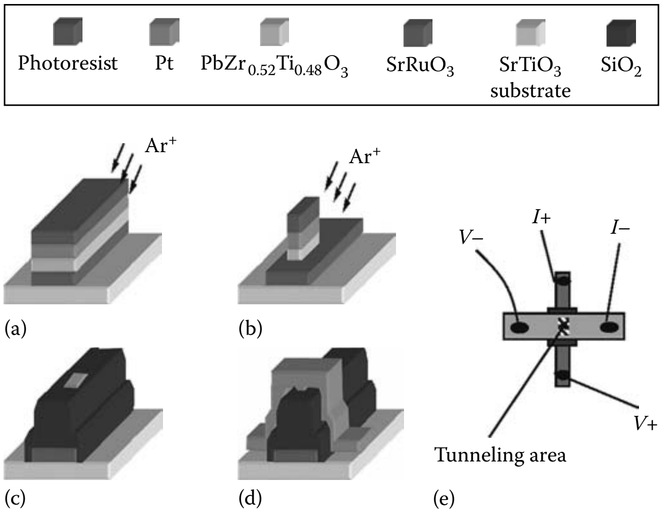

FIGURE 6.9 Patterning of tunnel junctions with the aid of photolithography and ion beam etching. (From Rodriquez Contreras, J. et al., *Mater. Res. Soc. Symp. Proc.*, 688, C8.10, 2002. With permission.)

electrode/ferroelectric/top electrode) is deposited without breaking the vacuum. Then the capacitor area is defined by a photolithographic process and dry-etching, typically using Ar ion beam milling. Again acetone is employed to remove the photoresist from the top electrode. Although this method delivers better electrical interface properties than the first one, etch residuals at the sidewalls of the mesa can cause short-circuiting along the sidewall from the top electrode to the bottom one. An etch stop right after reaching the top surface of the ferroelectric can avoid this problem in a simple way. Ion-beam etching is the preferred method for the studies of scaling effects, because it results in the smallest degradation at the top electrode–ferroelectric interface.

A more complex procedure should be used to pattern oxide tunnel junctions. As shown in Figure 6.9, three photo-mask steps in total allow the fabrication of tunnel junctions by conventional photolithography and ion-beam etching (Sun 1998). First, the whole layer sequence such as the $SrRuO_3$(20 nm)/$BaTiO_3$(2 nm)/$SrRuO_3$(20 nm) trilayer is deposited in situ. Using a photolithographic step and ion-beam milling, the shape of the bottom electrode is defined. Here the whole trilayer has to be etched down to the substrate. Next, the tunnel junction is defined by the second photo-mask step and an etching down to the bottom electrode. The subsequent deposition of SiO_2 is needed to isolate the surroundings of the mesa. A third photo-mask step and the deposition of the wiring layer complete the fabrication of a tunnel junction. Figure 6.9e shows the top view of the junction layout. The four-point probe arrangement, which avoids electrical artifacts during current-voltage measurements, is essential (Rodríguez et al. 2003b and Rodríguez Contreras 2004).

6.3.4 Current Trends in Deposition and Patterning

The physical vapor deposition methods described above make it possible to grow complex oxides with a high structural quality and a low surface roughness. Further developments in this field involve combinatorial PLD and MBE systems for the fabrication of epitaxial layers with composition gradients and wedge-like films with a thickness varying across the substrate (Nicolaou et al. 2002, Ohtani et al. 2005, Habermeier 2007). This interesting approach helps to reduce run-to-run uncertainties and allows the variation of certain film parameters in a single deposition run while keeping other conditions fixed.

New patterning methods can be developed as well, in particular, a combination of conventional top-down methods and bottom-up techniques. It should be noted that electron beam lithography has been used to produce capacitors with nanoscale dimensions (Szafraniak-Wiza et al. 2008). Focused ion-beam direct-writing techniques already showed their strength, but they suffer from sidewall contaminations at the mesa structure created by etching ions or atoms, e.g., Ga (Hambe et al. 2008).

6.4 Characterization of Ferroelectric Films and Capacitors

6.4.1 Rutherford Backscattering Spectrometry

Rutherford backscattering spectrometry (RBS) is an accurate nondestructive technique for measuring the stoichiometry, layer thickness, quality of interfaces, and crystalline perfection of thin films. RBS offers a quantitative determination of the absolute concentrations of different elements in multi-elemental thin films. A collimated mono-energetic beam of low-mass ions hits the specimen to be analyzed. Typically, He^+ ions with energy of 1.4 MeV are used in RBS experiments. A small fraction of the ions that impinge on the sample is scattered back elastically by the atomic nuclei and are then collected by a detector. The detector determines the energy of the backscattered ions, which provides an RBS energy spectrum. The RBS spectra describe the yield of backscattered particles as a function of their energy. The analysis of RBS spectra is done using modern software. A more detailed description of the technique is given by Chu et al. (1978).

In the so-called random experiments, the ion beam is not aligned with respect to the crystallographic directions of the specimen. The energy distribution of the collected ions provides information on the masses of atoms constituting the sample and on the thicknesses of deposited layers. Information on the sharpness of interfaces between these layers is given by the abruptness of the low-energy edge in the “random” spectrum.

Epitaxial films usually have the same major channeling axis as the substrate. The degree of epitaxy is determined from ion channeling experiments by a ratio of the elemental signals from the film for the channeled and random sample orientations. This ratio is called the “minimum yield,” χ_{min}, and its value provides information on the crystalline perfection of a film. Defects inside the film lead to higher values of χ_{min}.

6.4.2 X-Ray Diffraction for Thin-Film Analysis

X-ray diffraction (XRD) represents a powerful tool for the characterization of thin films. It can be used to determine whether the film grown on a crystalline substrate is amorphous, polycrystalline, or single-crystalline (epitaxial growth). Moreover, this technique makes it possible to determine the film thickness, lattice parameters, and the amount of strain in an epitaxially grown film with a high precision.

In particular, the 2θ scans performed at a fixed glancing incident angle of the incoming x-ray beam (in the range of 0.5°–2°) are suited for the investigations of polycrystalline films, since the spectrum contains only the peaks coming from the XRD of randomly oriented crystallites. (The single-crystal substrate does not contribute to the XRD because the Bragg condition is not satisfied for this angle of incidence.) For (001)-oriented epitaxial films, the normal θ–2θ scans reveal only the (00*l*) reflections. Thus, we can distinguish between an epitaxial film on a crystalline substrate and a polycrystalline one. Moreover, from the 2θ position of these reflections, one can precisely determine the out-of-plane lattice parameter. Once this parameter is measured, the in-plane lattice constants could be determined as well by finding the peak positions of the (*h*0*l*) reflections, for example. For ultrathin films, however, this becomes difficult because of the overlap with substrate peaks. Therefore, the grazing incidence diffraction, which is characterized by a low penetration depth of the incoming x-ray beam, has to be used to measure the in-plane lattice constants of ultrathin films.

The amount of strain in an epitaxial film can be nicely visualized by the reciprocal space maps measured, for example, around an asymmetric (103) reflection. These maps can indicate whether the film is fully strained by the substrate or is partially relaxed owing to the generation of misfit dislocations. Representative reciprocal space maps of strained and relaxed epitaxial $BaTiO_3$ films grown on $SrRuO_3$-covered $SrTiO_3$ substrates are given in Figure 6.10a and b. The out-of-plane and in-plane lattice parameters of $BaTiO_3$ films extracted from such maps are plotted in Figure 6.10c as a function of the film thickness t (Petraru et al. 2007). It can be seen that ultrathin films with $t < 30$ nm are commensurate with the substrate, which results in a compressive biaxial in-plane strain and an out-of-plane elongation of the unit cell.

The synchrotron x-ray scattering measurements give additional possibilities for the characterization of ultrathin films (Fong et al. 2005). In particular, it was demonstrated that electrode-free $PbTiO_3$ films grown on $SrTiO_3$ remain ferroelectric for thicknesses down to only 3 unit cells (Fong et al. 2004).

Finally, we note that the film thickness itself can be measured precisely using the x-ray specular reflectivity method based on interference fringes whose spacing is characteristic for this thickness (Fewster 1996). This method can be applied to films with any structure, crystalline or amorphous, but requires a flat surface over the region studied. It was demonstrated to work even

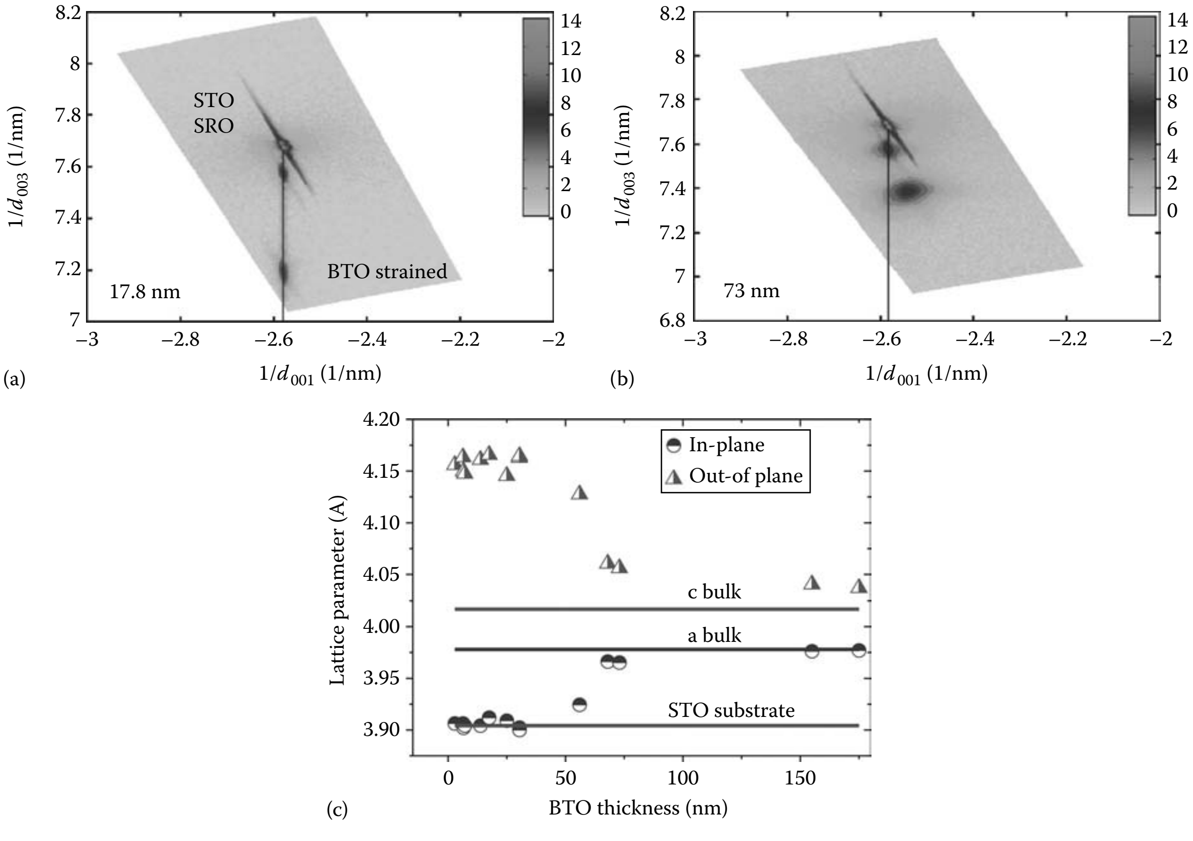

FIGURE 6.10 X-ray reciprocal space maps around the (103) Bragg reflection obtained for fully strained (a) and partially relaxed (b) $BaTiO_3$ films epitaxially grown on $SrRuO_3$-covered $SrTiO_3$. The film lattice parameters are plotted as a function of the film thickness in panel (c). (From Petraru, A. et al., *J. Appl. Phys.*, 101, 114106, 2007. With permission.)

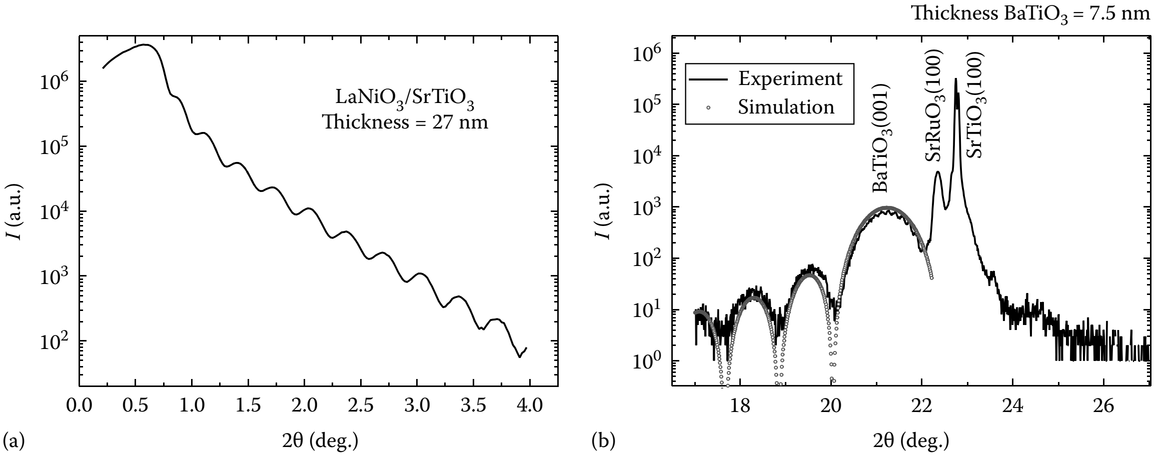

FIGURE 6.11 (a) Interference fringes appearing in an x-ray specular reflectivity scan for the 27 nm thick $LaNiO_3$ film deposited on $SrTiO_3$. The film thickness was calculated from the spacing of these fringes. (b) High-angle finite-size oscillations occurring in the θ–2θ scan around the (001) peak of the $BaTiO_3$ film (7.5 nm thick) grown on $SrRuO_3$-covered $SrTiO_3$. The solid line shows the measured signal, whereas the dots denote the results of simulations.

in the case of ultrathin films with thicknesses down to 24 Å. The amplitude of oscillations depends mainly on the density contrast between the layers, and the number of oscillations correlates with the roughness of the surface and interfaces involved. In the case of rough surfaces, the average intensity of reflectivity decreases rapidly with an increasing 2θ angle (Nevot and Croce 1980). For epitaxial films, high-angle finite-size oscillations occurring in the θ–2θ scans around the (001) peak allow determination of the number of planes involved in the diffraction, and, therefore, of the film thickness (Schuller 1980). Examples of the low- and high-angle finite-size oscillations are given in Figure 6.11.

6.4.3 Ferroelectric Capacitors: *P-E* Hysteresis Loop Measurements

A ferroelectric capacitor usually displays a polarization-field (*P-E*) hysteresis loop similar to that shown in Figure 6.4b. There are several techniques that are used to measure the *P-E* loops of ferroelectric capacitors. The simplest method employs a circuit proposed by Sawyer and Tower, which is shown schematically in Figure 6.12a. The circuit consists of a fixed capacitor with known capacitance, the test ferroelectric capacitor, an oscilloscope, and a function generator. The method relies on the fact that two capacitors in a series have the same charge. The ac voltage created by the generator and the potential across the standard capacitor are shown on the *x*- and *y*-axes of the oscilloscope. The capacitance of the standard capacitor is chosen to be large enough so that the voltage drop across this capacitor is much smaller than the potential difference across the tested ferroelectric capacitor.

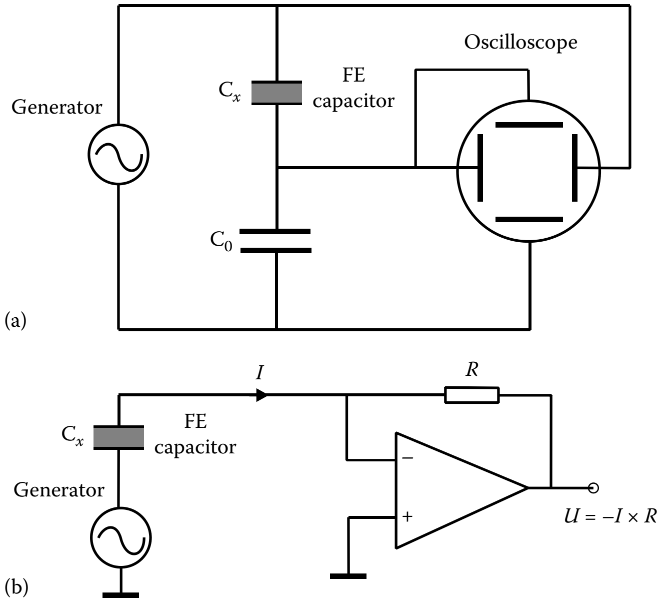

FIGURE 6.12 Experimental techniques for the measurement of polarization-voltage loops of ferroelectric capacitors: (a) Sawyer-Tower circuit allowing the visualization of ferroelectric hysteresis loops, and (b) an alternative method using a fast current-to-voltage converter, where the loop is obtained via the integration of measured electric current.

Another method uses a fast current-to-voltage converter connected in series with the ferroelectric capacitor (see the circuit shown in Figure 6.12b). In this case, the current-voltage curve is measured as a response of the ferroelectric capacitor to a triangular signal excitation. It is very useful to look at the switching current peaks that appear in this curve in order to distinguish the ferroelectric switching from artifacts, especially in the case of leaky ferroelectric samples. By numerical integration of the current over the time, the classical *P-E* hysteresis loop is obtained, from which the remanent polarization and the coercive field can be determined. The remanent polarization of nanoscale $SrRuO_3/BaTiO_3/SrRuO_3$ capacitors fabricated on the $SrTiO_3$ substrate is shown in Figure 6.13 as a function of the $BaTiO_3$ thickness *t* (Petraru et al. 2008). Remarkably, even at t = 3.5 nm, the strained $BaTiO_3$ film

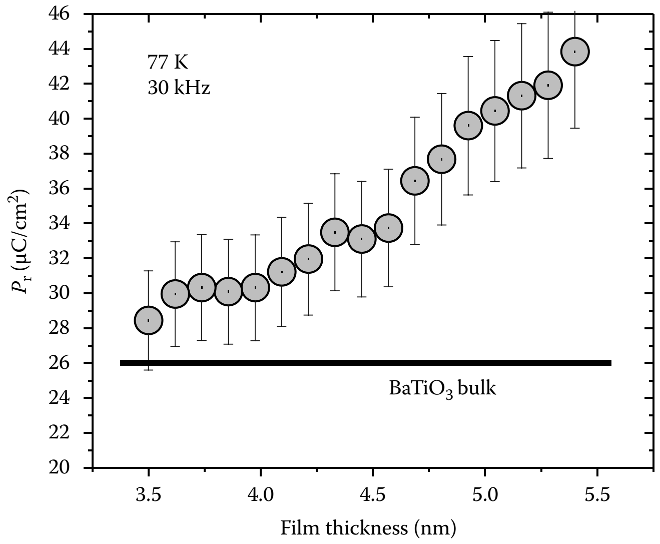

FIGURE 6.13 Thickness dependence of remanent polarization in the SrRuO$_3$/BaTiO$_3$/SrRuO$_3$ ferroelectric capacitors measured at 77 K. (From Petraru, A. et al., *Appl. Phys. Lett.*, 93, 072902, 2008. With permission.)

remains ferroelectric and has a remanent polarization larger than the spontaneous polarization $P_s = 26\,\mu C/cm^2$ of bulk BaTiO$_3$. The coercive field E_c of BaTiO$_3$ capacitors relatively weakly depends on the thickness t (Jo et al. 2006b), which contrasts with a strong increase of E_c (Figure 6.14) in ultrathin PZT capacitors with Pt top electrodes (Pertsev et al. 2003b).

Ferroelectric capacitors are often rather leaky, because thin films, especially at small thicknesses, are not perfect insulators. The conduction here results from the Schottky injection or Fowler–Nordheim tunneling through the interfacial barrier followed by the charge transport across the film via the Poole–Frenkel conduction mechanism, space-charge-limited conduction, or variable range hopping (Dawber et al. 2005). The leakage contribution to the total current can be singled out with the aid of the positive-up negative-down (PUND) pulsed method (Smolenskii et al. 1984). It involves the application of a series of voltage pulses from a function generator and the measurement of the transient current response of a ferroelectric device, which allows the separation of different contributions. As a representative example, we consider the sequence of five train pulses shown in Figure 6.15. The first one (0) is the pre-polarization pulse—it puts the sample into a definite polarization state. The pulse (1) switches the polarization of the sample, and its current response is the sum of the ferroelectric displacement current caused by the switching of spontaneous polarization, the dielectric displacement current, and the leakage current. The pulse (2) has the same polarity but comes after a certain delay time. Therefore, in case of a stable polarization, the current response contains only the components arising from the dielectric response and leakage current. In order to find the switchable polarization (the quantity of primary interest), the current response due to pulse (2) is subtracted from the current created by pulse (1), and the result is numerically integrated over the measuring time. Moreover, this method makes it possible to study the stability of ferroelectric polarization against back-switching. To that end, we can vary the delay time between pulses (1) and (2) and determine the relaxation time of the polarization. A similar analysis can be done for currents resulting from pulses (3) and (4) applied to the capacitor with opposite polarization.

6.4.4 Scanning Probe Techniques: Atomic Force Microscopy and Piezoresponse Force Microscopy

Atomic force microscopy (AFM) is one of the most widely used scanning probe microscopy (SPM) techniques (Garcia and Perez 2002). The primary purpose of an AFM instrument is to quantitatively measure the roughness of various surfaces. The lateral and vertical resolutions are typically about 5 and 0.01 nm, respectively. An atomically sharp tip is scanned over a surface with feedback mechanisms that enable the piezoelectric scanners to maintain the tip at a constant force (to obtain height information) or height (to obtain force information) above the sample surface.

Tips are typically made of Si$_3$N$_4$ or Si and extend down from the end of a cantilever. The AFM head employs an optical

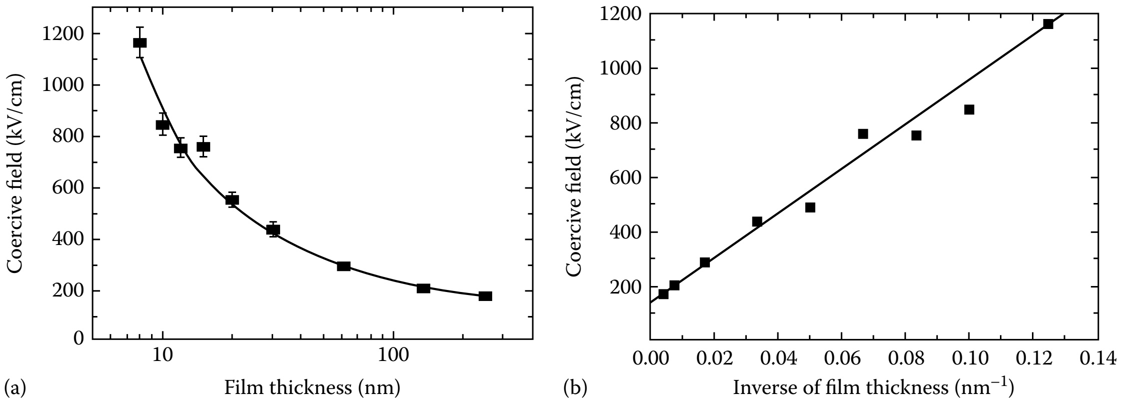

FIGURE 6.14 Coercive field of PZT 52/48 epitaxial films measured at 20 kHz and plotted versus the film thickness t (a) and the inverse of film thickness $1/t$ (b). The straight line in (b) shows a linear fit to the experimental data, whereas the curve in (a) is a guide to the eyes.

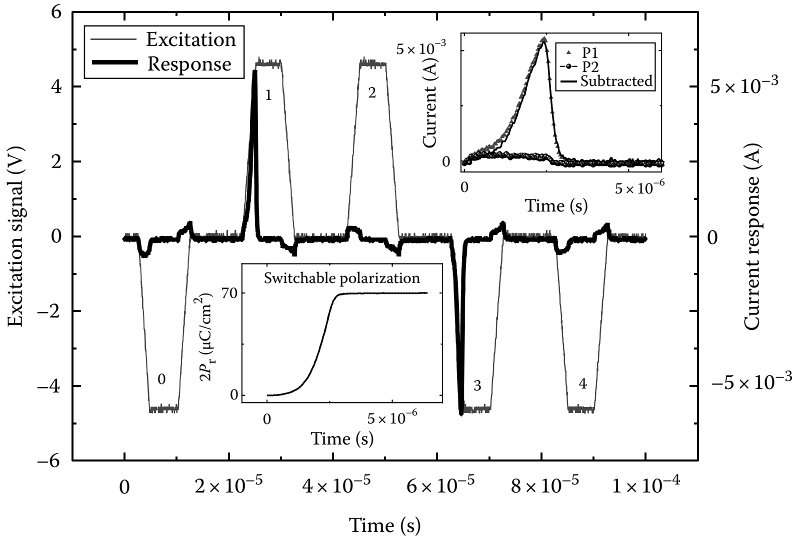

FIGURE 6.15 Measurements of switchable ferroelectric polarization by the PUND pulsed method. The excitation signal consists of five pulses denoted by thin lines, and the current response is shown by a thick line. The upper inset demonstrates the switching (1) and nonswitching (2) current responses. The integration of their difference gives the switchable polarization plotted in the lower inset.

detection system, in which the tip is attached to the bottom of a reflective cantilever. A laser diode is focused onto the back of this cantilever. As the tip scans the surface of a sample, the laser beam is deflected by the cantilever into a four-quadrant photodiode. In contact mode, feedback from the photodiode difference signal, through the software control from a computer, enables the tip to maintain either a constant force or a constant height above the sample. In the constant force mode, the piezoelectric transducer monitors real-time height variations. In the constant height mode, the deflection force acting on the tip is recorded. The instrument gives a topographical map of the sample surface by plotting the local sample height versus the horizontal probe tip position. For many soft materials like polymers and biological samples, the operation in contact mode often modifies or destroys the surface. These complications can be avoided using the tapping-mode AFM. In tapping mode, the AFM tip–cantilever assembly oscillates at the sample surface during the scanning. As a result, the tip lightly taps the surface while scanning and only touches the sample at the bottom of each oscillation. This prevents damage of soft specimens and avoids the "pushing" of specimens along the substrate. By using a constant oscillation amplitude, a constant tip–sample distance is maintained until the scan is complete. Tapping-mode AFM can be performed on both wet and dry surfaces.

Scanning probe microscopy techniques also offer several different possibilities for the investigation of domain patterns in crystals. For imaging domain structures in ferroelectrics, piezoresponse force microscopy (PFM) is most widely used nowadays (Figure 6.16). Introduced in 1992 by Güthner and Dransfeld (1992), the PFM method has been developed by several groups to visualize domain structures in ferroelectric thin films. It became a popular tool in the science and technology of ferroelectrics and is considered to be a main instrument for getting information on ferroelectric properties at the nanoscale. Several reviews on the SPM-based methods for the characterization of ferroelectric domains are available in the literature (Gruverman and Kholkin 2004, Kholkin et al. 2007).

Ideally, when a modulation voltage V is applied to a piezoelectric material, the vertical displacement of the probing tip, which is in mechanical contact with the sample, accurately follows the motion of the sample surface resulting from the converse piezoelectric effect. The applied voltage V generates an electric field $\mathbf{E}(\mathbf{r})$ in the ferroelectric film, which creates the lattice strain $\delta u_3 = d_{13}E_1 + d_{23}E_2 + d_{33}E_3$ in the film thickness direction. Here, d_{ij} are the local piezoelectric coefficients of the ferroelectric material, which depend on the polarization orientation. The strain field $\delta u_3(\mathbf{r})$ changes the film thickness at the tip position by an amount δt so that the local piezoresponse signal proportional to $\delta t(V)$ can be recorded. The amplitude of the tip vibration measured by the lock-in technique provides information on the effective piezoelectric coefficient $d_{33}^{\text{eff}} = \delta t / V$. The phase yields information on the polarization direction in a studied ferroelectric domain (Rodriguez et al. 2002). It should be noted that not only the surface electromechanical response but also the electrostatic forces could contribute to the measured signal in the PFM setup, which complicates the analysis of the PFM results. In particular, there exists a nonlocal contribution caused by the capacitive cantilever–sample interaction (Kalinin and Bonnell 2002).

Most PFM measurements are performed in a local-excitation configuration where the modulation voltage is applied between the bottom electrode and conductive SPM tip, which scans the bare surface of the film having no top electrode. In this case, the PFM image has a lateral resolution of about 10 nm (Gruverman et al. 1998). It should be noted that the electric field generated by the SPM tip in such film is highly inhomogeneous, which makes the *quantitative* analysis of the field-induced signal extremely difficult. In other words, PFM measurements on a sample

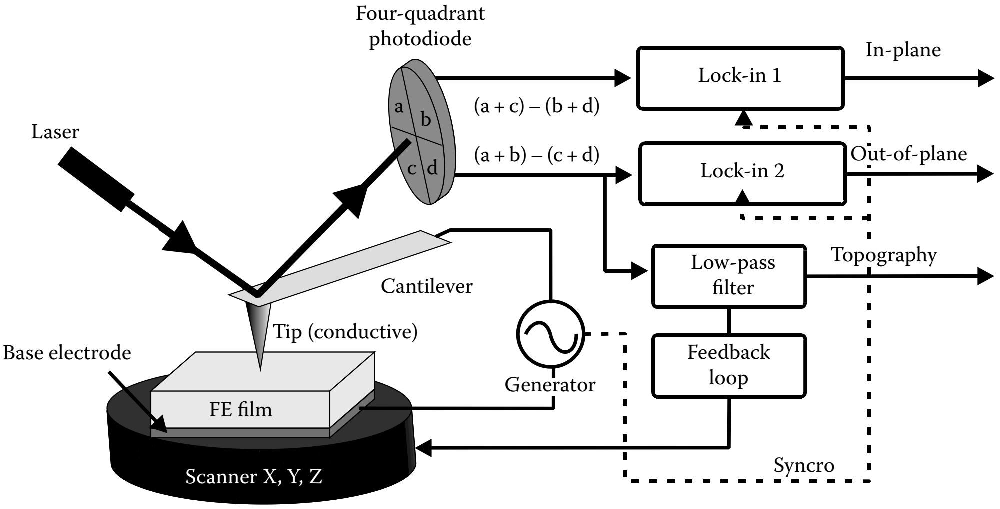

FIGURE 6.16 PFM setup for simultaneous measurements of the surface topography and the out-of-plane and in-plane piezoelectric responses of a ferroelectric sample. The cantilever deflection caused by the voltage-induced surface displacements is detected by a laser beam reflecting into a four-quadrant photodiode. The vertical and horizontal piezoresponses are determined with the aid of two lock-in amplifiers by demodulating the corresponding signals (a + c) – (b + d) and (a + b) – (c + d), respectively. The instrument is operated in a constant force mode.

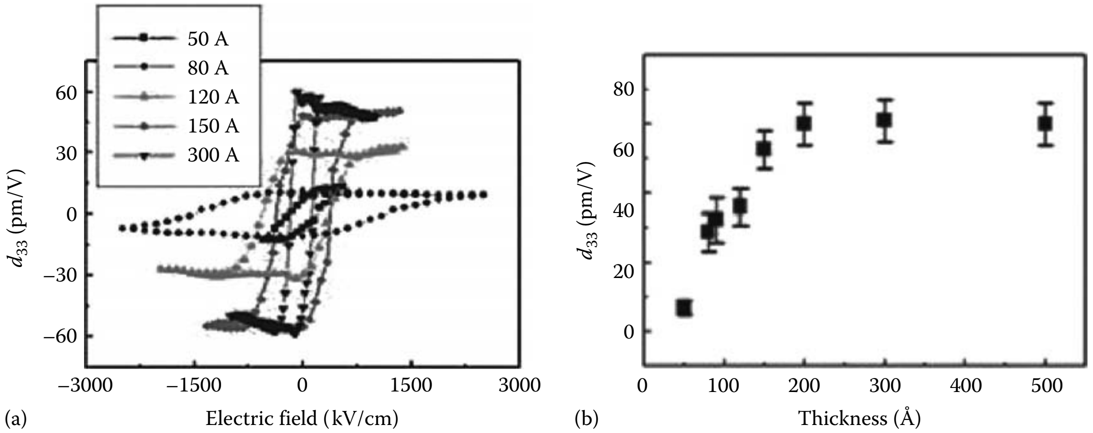

FIGURE 6.17 Out-of-plane piezoelectric response of $SrRuO_3/PbZr_{0.2}Ti_{0.8}O_3/SrRuO_3$ capacitors measured by the PFM technique: (a) piezoelectric hysteresis loops obtained for five different film thicknesses, and (b) the piezoelectric coefficient d_{33} as a function of the film thickness. (From Nagarajan, V. et al., *J. Appl. Phys.*, 100, 051609, 2006, With permission.)

without extended top electrodes collect signals from a subsurface layer of unknown thickness that is a function of dielectric permittivity and contact conditions (Gruverman et al. 1998).

Alternatively, a ferroelectric film with a deposited top electrode may be studied at the expense of a lower lateral resolution. By applying a voltage to the SPM tip contacting the top electrode, a submicron variation of piezoelectric properties in PZT capacitors was demonstrated (Christman et al. 2000, Setter et al. 2006). Under these conditions, a homogeneous electric field is generated in a ferroelectric layer, and the electrostatic tip–sample interaction is suppressed. This approach allows the investigations of domain-wall dynamics and polarization reversal mechanisms in ferroelectric capacitors and quantitative studies of the scaling of piezoelectric properties in ultrathin ferroelectric films. In particular, it was found (Nagarajan et al. 2006) that the piezoelectric coefficient d_{33} of epitaxial $PbZr_{0.2}Ti_{0.8}O_3$ films sandwiched between $SrRuO_3$ electrodes decreases rapidly as the thickness is reduced from 20 to 5 nm (see Figure 6.17).

6.5 Physical Phenomena in Ferroelectric Capacitors

There are several physical effects that make the phase states and electric properties of thin-film ferroelectric capacitors different from those of bulk ferroelectrics. First, the ferroelectric film is generally subjected to an in-plane straining and clamping due to the presence of a dissimilar thick substrate. Second, an internal electric field exists in the capacitor, which depends on the electrode material and the film thickness. Third, the scaling of physical properties may result from the short-range interatomic interactions at the film–electrode interfaces. The current status of the theoretical description of these effects in thin films of perovskite ferroelectrics is given below.

6.5.1 Strain Effect

Owing to the electrostrictive coupling between lattice strains and polarization, the mechanical film–substrate interaction may strongly affect the physical properties of ferroelectric thin films (Pertsev et al. 1998). In a film deposited on a dissimilar thick substrate, the in-plane strains u_1, u_2, and u_6 are totally governed by the substrate, whereas the stresses σ_3, σ_4, and σ_5 are usually equal to zero. (We use the Voigt matrix notation and the reference frame with the x_3 axis orthogonal to the film surfaces.) Under such "mixed" mechanical boundary conditions, the equilibrium polarization state corresponds to a minimum of the modified thermodynamic potential $\widetilde{G}$ (Pertsev et al. 1998), but not of the standard elastic Gibbs function G (Haun et al. 1987). In the most important case of a film grown in the (001)-oriented cubic paraelectric phase on a (001)-oriented cubic substrate ($u_1 = u_2 = u_m$, $u_6 = 0$), the stability ranges of different polarization states can be conveniently described with the aid of two-dimensional phase diagrams, where the misfit strain $u_m = (b - a_0)/a_0$ and temperature T are used as two independent parameters (a_0 is the equivalent cubic cell constant of the free standing film and b is the substrate lattice parameter). Such "misfit strain-temperature" diagrams were developed with the aid of thermodynamic calculations for single-domain $BaTiO_3$, $PbTiO_3$, and $Pb(Zr_{1-x}Ti_x)O_3$ (PZT) films (Pertsev et al. 1998, 2003a). Since the substrate-induced strains lower the symmetry of the paraelectric phase from cubic to tetragonal, the film polarization state may be very different from the ferroelectric phases observed in the corresponding bulk material (see Figure 6.18).

At large negative misfit strains, films of perovskite ferroelectrics stabilize in the tetragonal c phase with the spontaneous polarization $\mathbf{P}_s$ orthogonal to the film–substrate interface, whereas at large positive strains the orthorhombic aa phase forms, where $\mathbf{P}_s$ is directed along the in-plane face diagonal of the prototypic cubic cell. At low temperatures, the stability ranges of the c and aa phases are separated by a "monoclinic gap," where the monoclinic r phase with three nonzero polarization components P_i becomes the energetically most favorable state. These predictions of the thermodynamic theory were confirmed by the first-principles calculations (Bungaro and Rabe 2004, Diéguez et al. 2004).

It should be emphasized that the orthorhombic and monoclinic phases do not exist in the bulk $PbTiO_3$ crystals, where only the tetragonal ferroelectric state is stable (Haun et al. 1987). In the case of $BaTiO_3$, the aa phase may be compared with the orthorhombic phase forming in the bulk crystal in the low-temperature range between 10°C and −71°C, whereas the r phase can be regarded as a distorted modification of the rhombohedral phase that exists in a free crystal below −71°C (Jona and Shirane 1962).

The most remarkable manifestation of the strain effect appears in thin films of strontium titanate. In a mechanically free state, bulk $SrTiO_3$ crystals remain paraelectric down to zero absolute temperature despite a strong softening of the transverse optic polar mode near $T = 0\,K$ (Müller and Burkard 1979). The thermodynamic calculations show that this "incipient ferroelectricity" exists in epitaxial $SrTiO_3$ films grown on dissimilar cubic substrates only at small misfit strains ranging from -2×10^{-3} to -2×10^{-4} (Pertsev et al. 2000). Outside this "paraelectric gap," the ferroelectric phase transition takes place in the $SrTiO_3$ film at a finite temperature, which rises rapidly with the increase of the strain magnitude. The predicted phenomenon of strain-induced ferroelectricity was observed experimentally in $SrTiO_3$ films grown on (110)-oriented $DyScO_3$, which were found to display ferroelectric properties at room temperature (Haeni et al. 2004).

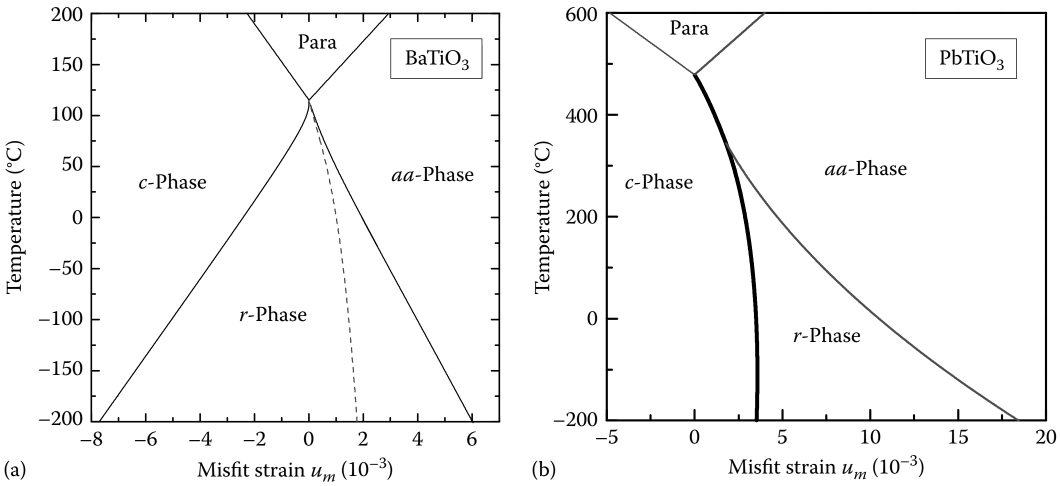

FIGURE 6.18 Misfit strain-temperature phase diagrams of single-domain $BaTiO_3$ (a) and $PbTiO_3$ (b) thin films epitaxially grown on (001)-oriented cubic substrates. The second- and first-order phase transitions are shown by thin and thick lines, respectively. (From Pertsev, N.A. et al., *Phys. Rev. Lett.*, 80, 1988, 1998. With permission.)

The strain-induced increase of the temperature T_c, at which the paraelectric to ferroelectric phase transition takes place, is characteristic of all studied perovskite ferroelectrics (Pertsev et al. 1998, 2003a). In addition, the two-dimensional clamping of the film by a thick substrate may change the order of this transition (Pertsev et al. 1998). Strong dependence of T_c on the misfit strain u_m explains the very high transition temperatures observed in epitaxial films grown on dissimilar substrates (Choi et al. 2004, He and Wells 2006).

The magnitude of the spontaneous polarization is also sensitive to the lattice strains. This effect is especially pronounced in ferroelectric films grown on "compressive" substrates ($u_m < 0$), where the polarization $\mathbf{P}_s$ is orthogonal to the film surfaces. For fully strained $BaTiO_3$ films grown on $SrTiO_3$ ($u_m = -2.6\%$), the thermodynamic theory predicts $P_s = 35\,\mu C/cm^2$, which is close to the experimental values of 43–$44\,\mu C/cm^2$ (Kim et al. 2005, Petraru et al. 2007). Remarkably, the film polarization exceeds the polarization $P_b = 26\,\mu C/cm^2$ of bulk $BaTiO_3$ significantly. At the same time, the strain sensitivity of polarization in highly polar Pb-based perovskites, where the ferroelectric ionic displacements are already large in the bulk, is relatively low (Lee et al. 2007).

The enhancement of polarization P_s and the decrease of the in-plane permittivity ε_{11} in the strained *c* phase (Koukhar et al. 2001) should lead to a considerable increase of the coercive field E_c in thin films (Pertsev et al. 2003b). From the Landauer model of domain nucleation (Landauer 1957) it follows that $E_c \sim \gamma^{6/5}/\left(\varepsilon_{11}^{1/5}P_s^{3/5}\right)$, where $\gamma \sim P_s^3$ is the domain-wall energy. Hence, the coercive field $E_c \sim P_s^3/\varepsilon_{11}^{1/5}$ of $BaTiO_3$ films grown on $SrTiO_3$ ($\varepsilon_{11} \approx 170$) may be about eight times larger than that of the bulk crystal ($\varepsilon_{11} \approx 3600$). Although it is certainly a very strong increase, the strain effect alone cannot explain the observed drastic difference between the measured coercive fields $E_c \sim$ 150–300 kV/cm of epitaxial $BaTiO_3$ films (Jo et al. 2006b, Petraru et al. 2007) and the bulk $E_c \sim 1$ kV/cm.

Finally, it should be noted that ferroelectric properties of epitaxial thin films may strongly depend on the orientation of the crystal lattice with respect to the substrate surface. In particular, the phase states and dielectric properties of single-domain $PbTiO_3$ films with the (111)-orientation of the paraelectric phase were found to be very different from those of the (001)-oriented films (Tagantsev et al. 2002).

6.5.2 Depolarizing-Field Effect

When the polarization charges $\rho = -\text{div}\,\mathbf{P}$ existing at the film surfaces are not perfectly compensated for by other charges, an internal electric field appears in the ferroelectric layer (Figure 6.19). This "depolarizing" field $\mathbf{E}_{\text{dep}}$ may be significant even in short-circuited ferroelectric capacitors with perfect interfaces because the electronic screening length in metals is finite (Guro et al. 1970, Mehta et al. 1973).

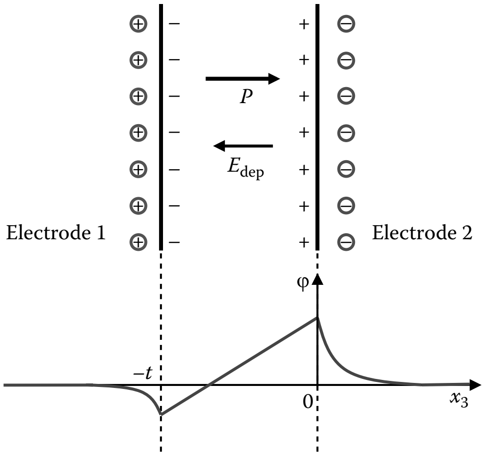

FIGURE 6.19 Imperfect screening of polarization charges in a ferroelectric capacitor. Distribution of the electrostatic potential φ is shown schematically for the case of dissimilar electrodes kept at a bias voltage compensating for the difference of their work functions. (From Pertsev, N.A. and Kohlstedt, H., *Phys. Rev. Lett.*, 98, 257603, 2007. With permission.)

For a homogeneously polarized film, the electrostatic calculation gives $E_{\text{dep}} = -P_3/(\varepsilon_0\varepsilon_b + c_i t)$, where P_3 is the equilibrium out-of-plane polarization in the film of thickness t, ε_0 is the permittivity of the vacuum, $\varepsilon_b \sim 10$ is the background dielectric constant of a ferroelectric material (Tagantsev and Gerra 2006), and c_i is the total capacitance of the screening space charge in the electrodes per unit area (Ku and Ullman 1964). When $P_1 = P_2 = 0$ (the *c* phase), the polarization $P_3(t)$ can be calculated from the nonlinear equation of state $\partial\tilde{G}/\partial P_3 = 0$ written for a strained film with an internal field E_{dep}. Since the capacitance c_i affects the polarization only via the product $c_i t$, the dependencies $P_3(t)$ corresponding to different electrode materials can be described by one universal curve $P_3(t_{\text{eff}})$. Here the effective film thickness t_{eff} may be defined as $t_{\text{eff}} = (c_i/c_1)t$, where $c_1 = 1\ F/m^2$.

Figure 6.20 shows the dependencies $P_3(t_{\text{eff}})$ calculated for fully strained PZT 50/50 and $BaTiO_3$ films grown on $SrTiO_3$ (Pertsev and Kohlstedt 2007). It can be seen that the spontaneous polarization decreases in thinner films and vanishes at a critical film thickness t_0. In the case of capacitors with $SrRuO_3$ electrodes ($c_i = 0.444\ F/m^2$), the thickness t_0 is about 2 nm for PZT 50/50 films and about 2.6 nm for $BaTiO_3$ films. However, the size-induced phase transition at t_0, also predicted by the first-principles calculations (Junquera and Ghosez 2003), cannot be observed in reality since at a slightly larger film thickness $t_c < 3$ nm the single-domain ferroelectric state becomes unstable and transforms into the 180° polydomain state (Pertsev and Kohlstedt 2007).

The experimental studies of ultrathin $BaTiO_3$ films (Kim et al. 2005, Petraru et al. 2008) showed that the remanent polarization decreases monotonically with decreasing thickness at $t < 30$ nm, where the misfit strain becomes constant. This behavior is similar to the dependence shown in Figure 6.20, but the thermodynamic theory does not provide a precise quantitative description of the experimental data (Kim et al. 2005, Petraru et al. 2008). The first-principles calculations performed by Junquera and Ghosez (2003) also cannot explain the measured dependence quantitatively (Kim et al. 2005).

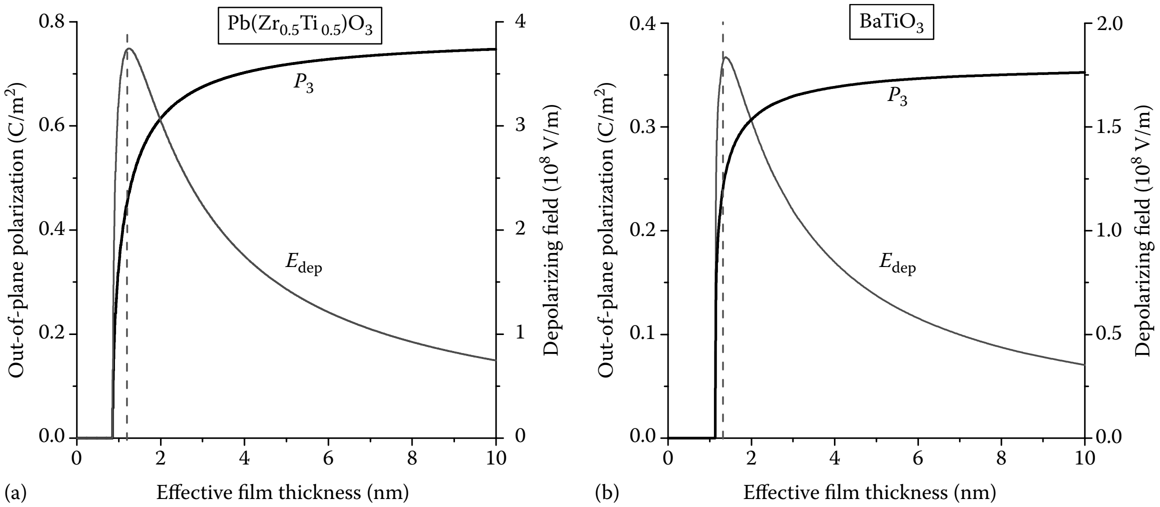

FIGURE 6.20 Thickness dependence of the out-of-plane polarization and depolarizing field in ultrathin Pb(Zr$_{0.5}$Ti$_{0.5}$)O$_3$ (a) and BaTiO$_3$ (b) films epitaxially grown on SrTiO$_3$. The misfit strain is taken to be -39×10^{-3} for Pb(Zr$_{0.5}$Ti$_{0.5}$)O$_3$ films and -26×10^{-3} for BaTiO$_3$ films; $T = 25°C$. The dashed line shows the film thickness below which the single-domain state becomes unstable.

For BaTiO$_3$ capacitors with SrRuO$_3$ electrodes, the depolarizing field E_{dep} has been evaluated experimentally as well and was found to increase dramatically with decreasing film thickness (Kim et al. 2005). Such a strong rise is consistent with the theoretical thickness dependence of E_{dep} shown in Figure 6.20, which was calculated with the account of the polarization variation $P_3(t_{eff})$. The predicted saturation of the depolarizing field and its decrease in ultrathin films, however, were not observed. In this thickness range, the single-domain state becomes unstable (Pertsev and Kohlstedt 2007), and the mean value of the depolarizing field vanishes due to the formation of 180° domain structure.

The presence of a depolarizing field in a capacitor is expected to affect the shape of polarization-voltage (*P-V*) hysteresis loops and the magnitude of coercive fields (Tagantsev and Gerra 2006, Jo et al. 2006a,b). Indeed, a simple calculation shows that imperfect compensation of polarization charges at the interfaces not only reduces the remanent polarization, but also leads to a tilt of the hysteresis loop (Tagantsev and Gerra 2006). The influence of E_{dep} on the magnitude of the coercive field seems to be less pronounced. An accurate electrostatic modeling of the polarization switching on the macroscopic level indicates that E_c reduces with decreasing film thickness, but only when the *P-V* loop is not saturated (Tagantsev and Gerra 2006). In the limit of high amplitudes of the driving field applied to a capacitor, the influence of the depolarizing field on E_c disappears. This feature is explained by the absence of voltage drop across the metal–ferroelectric interface at $E = E_c$, where the macroscopic polarization $P(E)$ goes to zero by definition.

At the same time, Jo et al. (2006a,b) proposed that the depolarizing field may strongly reduce the potential barriers hindering polarization switching in ultrathin films. Such reduction may happen at the early stage of switching, when the macroscopic polarization P is large, but the influence of E_{dep} becomes negligible in a film with $P = 0$. Therefore, it is not clear whether the coercive field changes. It would decrease if the switching develops as an "avalanche" process after the nucleation of a few reversed domains. On the contrary, the change of E_c should be small when the applied field must increase significantly in the course of switching to facilitate further domain nucleation or domain-wall motion.

It should be noted that the nucleation of reversed domains in a perfect bulk crystal is impossible since the activation energy is many orders of magnitude larger than the thermal energy (Landauer's paradox). Jo et al. (2006a) suggested that such "homogeneous" domain nucleation becomes possible in ultrathin ferroelectric films, where the field-dependent activation energy decreases dramatically due to strong depolarizing fields. This conclusion, however, is based on a rough estimate of the activation energy obtained in the approximation of a homogeneous depolarizing field, which is valid for a single-domain film only.

Domain nucleation is evidently easier near crystal surfaces, ferroelectric–electrode interfaces, and lattice defects. Gerra et al. (2005) proposed that the interface coupling between a ferroelectric layer and electrodes changes its sign after the polarization reversal and, therefore, stimulates the switching. According to their model, the surface-stimulated nucleation may reduce the coercive field of BaTiO$_3$ down to a value two orders of magnitude smaller than the thermodynamic coercive field. High local electric fields $E_{loc} \gg E$ created by spikes on the electrodes may also permit the domain formation at small applied fields E. Besides, residual domains probably exist in ferroelectric films, especially in polycrystalline ones. The growth of these domains may play an important role in the polarization switching, as it happens in ferroelectric polymers (Pertsev and Zembilgotov 1991).

6.5.3 Intrinsic Size Effect in Ultrathin Films

Since the unit cells adjacent to the film–electrode interfaces have an atomic environment different from that of the inner cells, the ferroelectric polarization may depend on the film thickness even in the absence of a depolarizing field. This "intrinsic" size effect can be described with the aid of a modified thermodynamic theory based on the concept of extrapolation length (Kretschmer and Binder 1979). In this theory, the total energy of the ferroelectric layer involves an additional surface contribution, and the polarization distribution across the film, in general, is taken to be inhomogeneous.

In the most important case of the (001)-oriented film grown on a compressive substrate ($P_1 = P_2 = 0$, $P_3 \neq 0$), the polarization profile $P_3(x_3)$ can be calculated from the Euler–Lagrange equation (Zembilgotov et. al. 2002). For a film having the same atomic terminations at both surfaces and sandwiched between identical electrodes, the boundary conditions can be written as $dP_3/dx_3 = P_b/\delta$ at $x_3 = 0$ and $dP_3/dx_3 = -P_b/\delta$ at $x_3 = t$, where P_b is the polarization value at the film boundaries and δ is the extrapolation length. The polarization suppression (enhancement) near the film surfaces is described by positive (negative) values of the extrapolation length. In a weakly conducting ferroelectric, such as $BaTiO_3$ or $PbTiO_3$, the inner polarization charges $\rho = -dP_3/dx_3$ are largely compensated by charge carriers so that the associated depolarizing field should be negligible. In this case, the spatial scale of polarization variations is determined by the ferroelectric correlation length $\xi^* = \sqrt{g_{11}/|a_3^*|}$ of a strained film, and the strength of the intrinsic size effect is governed by the ratio $\xi^*/|\delta|$. (Here g_{11} and $a_3^*(u_m)$ are the coefficients of the gradient term and the renormalized second-order polarization term in the free energy expansion, respectively.)

The numerical calculations demonstrate that the polarization suppression in the surface layers reduces the temperature T_c of ferroelectric transition at a given misfit strain u_m. This reduction, however, is significant only in ultrathin films with thicknesses about a few $\xi^*(T = 0)$. Below the transition temperature $T_c(u_m)$, the mean polarization reduces with decreasing film thickness and vanishes at a critical thickness $t_c \approx \xi^*$ (Zembilgotov et al. 2002). Thus, the intrinsic surface effect may lead to a size-induced ferroelectric to paraelectric transformation.

Since the discussed thermodynamic theory is a continuum theory, it is valid only when the characteristic length ξ^* of the polarization variations is larger than the interatomic distances. In the case of a negligible depolarizing field, this condition is satisfied at least near the transition temperature T_c because the coefficient a_3^* goes to zero at this temperature (Pertsev et al. 1998). The situation, however, changes dramatically in a perfectly insulating ferroelectric, where the uncompensated polarization charges $\rho = -\text{div}\,\mathbf{P}$ inside the film create a nonzero depolarizing field (Kretschmer and Binder 1979, Tagantsev et al. 2008). The characteristic length of polarization variations becomes $\xi_d^* = \sqrt{g_{11}/\left|a_3^* + (\varepsilon_0\varepsilon_b)^{-1}\right|}$, which, in contrast to ξ^*, does not increase significantly near T_c. Since $\xi_d^* \sim 0.1\text{nm}$ only, the continuum approach based on the concept of extrapolation length cannot be used to describe the surface effect in perfectly insulating perovskite ferroelectrics (Tagantsev et al. 2008).

In the latter case, however, the ferroelectric film may be assumed to be homogeneously polarized in the thickness direction, and the intrinsic size effect can be described with the aid of a phenomenological approach as well (Tagantsev et al. 2008). To that end, the film free energy is written as the sum of the "bulk" and "surface" contributions, each represented by a polynomial in terms of ferroelectric polarization. In general, the surface contribution should involve not only the even-power terms, but also the odd-power terms, because the surface breaks the inversion symmetry of the ferroelectric (Levanyuk and Sigov 1988, Bratkovsky and Levanyuk 2005). However, when the film–electrode interfaces are identical, the linear term vanishes and the surface energy can be approximated by a quadratic polarization term. The coefficient of this term may be evaluated from the comparison of the phenomenological theory with the results of first-principles calculations performed for ultrathin ferroelectric films. For $BaTiO_3$ capacitors with $SrRuO_3$ electrodes, this procedure reveals that the surface energy is positive (Tagantsev et al. 2008), which implies the polarization suppression at the interfaces.

In other metal/ferroelectric/metal heterostructures, however, polarization could be enhanced near the interfaces. Such enhancement was demonstrated by the first-principles-based calculations performed for ultrathin $PbTiO_3$ and $BaTiO_3$ films (Ghosez and Rabe 2000, Lai et al. 2005). The short-range interactions between the film and electrodes must be also taken into account to prove that this effect exists in ferroelectric capacitors as well. The importance of ionic displacements in the boundary layers of $SrRuO_3$ electrodes for the stabilization of ferroelectricity in ultrathin films was revealed by the first-principles investigations (Sai et al. 2005, Gerra et al. 2006).

6.6 Future Perspective

When the thickness of the ferroelectric layer in a biased capacitor becomes as small as a few nanometers, the quantum mechanical electron tunneling across the insulating barrier should become significant (Kohlstedt et al. 2005b, Zhuravlev et al. 2005a). Since the tunnel current exponentially depends on the barrier thickness, the crossover from a capacitor to a *tunnel junction* takes place near some threshold thickness. The presence of spontaneous polarization in a ferroelectric barrier and its piezoelectric properties are expected to make the current-voltage (*I-V*) characteristic of a ferroelectric tunnel junction (FTJ) very different from those of conventional tunnel junctions involving nonpolar dielectrics (Kohlstedt et al. 2005b and Rodríguez Contreras 2004). Theoretically, the junction conductance can change strongly after the polarization reversal in the barrier so that the FTJs are promising for the memory storage with nondestructive readout. For the memory applications, asymmetric FTJs with dissimilar electrodes seem to be preferable since such junctions should exhibit much larger conductance on/off ratios (Kohlstedt et al. 2005b, Zhuravlev et al. 2005a). This feature is due to the fact that here the mean barrier height changes after the polarization reversal by the amount $\Delta\phi \cong eP_3(t)(c_{m2}^{-1} - c_{m1}^{-1})$, where c_{m1} and c_{m2} are the capacitances of two electrodes and e is the electron charge.

The ferroelectric tunnel barrier may also be combined with ferromagnetic electrodes. For such a *multiferroic* tunnel junction (MFTJ), new functionalities may be expected since the tunneling probability becomes different for the spin-up and spin-down electrons owing to the exchange splitting of electronic bands in ferromagnetic electrodes. In particular, Zhuravlev et al. proposed a new spintronic device, where an electric current is injected from a diluted magnetic semiconductor through the ferroelectric barrier to a normal (nonmagnetic) semiconductor (Zhuravlev et al. 2005b). Their theoretical calculations indicated that the switching of ferroelectric polarization in the barrier may change the spin polarization of the injected current markedly, which provides a two-state electrical control of the device performance.

When both electrodes are ferromagnetic, the tunnel current becomes dependent on the mutual orientation of the electrode magnetizations. This phenomenon, which is termed *tunneling magnetoresistance* (TMR), is important for the applications in spin-electronic devices such as magnetic sensors and magnetic random-access memories. Since the TMR ratio depends not only on the properties of ferromagnetic electrodes, but also on the barrier characteristics (Slonczewski 1989), it may be sensitive to the orientation of the ferroelectric polarization in the MFTJ. This supposition was confirmed by the theoretical calculations performed for junctions involving two magnetic semiconductor electrodes (Zhuravlev et al. 2005b). It was shown that, under certain conditions, the MFTJ works as a device that allows the switching of TMR between positive and negative values.

The experimental realization of ferroelectric and multiferroic tunnel junctions, however, is a task with many obstacles because it requires the fabrication of ultrathin films retaining pronounced ferroelectric properties at a thickness of only a few unit cells. Moreover, the ferroelectric state with a nonzero net polarization must be stable at such a small thickness and switchable by a moderate external voltage. In our opinion, reliable FTJs showing resistive switching in the tunneling regime have not been fabricated yet, despite several attempts made in this direction (Rodríguez et al. 2003b, Gajek et al. 2007). The observation of a hysteretic *I-V* curve or resistance jumps after short-voltage pulses alone is not sufficient to prove the existence of an FTJ (Kohlstedt et al. 2008). Hysteretic *I-V* curves were also measured for *non*ferroelectric $LaSr_xMn_{1-x}O_3/SrTiO_3/LaSr_xMn_{1-x}O_3$ tunnel junctions and explained by other effects (Sun 2001). Further experiments are necessary to demonstrate the functioning of FTJs unambiguously.

In case the above challenges will be overcome in the future, a number of new exciting opportunities for technological applications and interesting physical phenomena are anticipated. Figure 6.21 summarizes the variety of novel functional oxide tunnel junctions. Here, the Josephson tunnel junctions with a ferroelectric or multiferroic barrier are included as well. The influence of a ferroic tunnel barrier on the Cooper pair and quasiparticle tunneling might lead to interesting new physics. Besides oxide materials, ferroelectric polymers such as PVDF and P(VDF-TrFE) (Xu 1991, Bune et al. 1998) can be incorporated in low-temperature superconducting Josephson junctions [e.g., Nb/Al-AlO$_x$-P(VDF-TrFE)/Nb] (Huggins and Gurvitch 1985) and in magnetic tunnel junctions [e.g., Co_xFe_{1-x}/AlO_x-P(VDF-TrFE)/$Ni_{20}Fe_{80}$] (Moodera et al. 1995). In principle, one would expect the appearance of physical effects similar to those

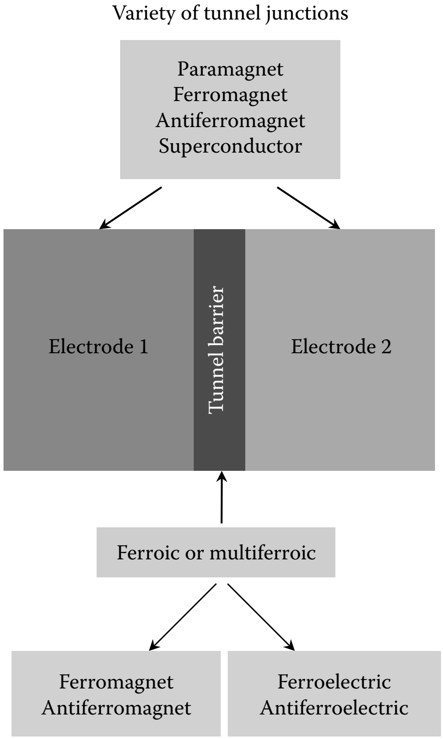

FIGURE 6.21 A "zoo" of novel tunnel junctions involving multifunctional tunnel barriers and electrodes of various types.

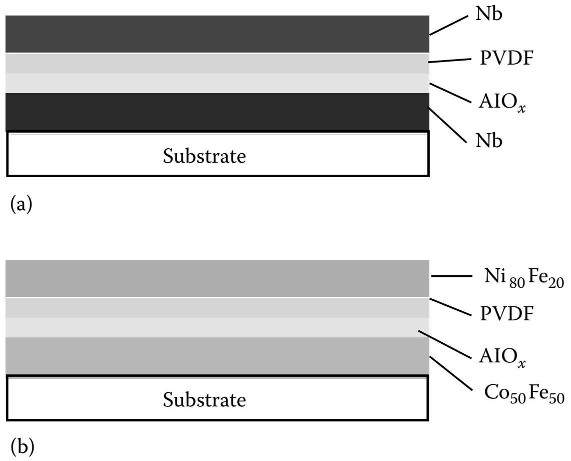

FIGURE 6.22 Low-temperature superconducting Josephson junctions (a) and magnetic tunnel junctions (b) with the AlO$_x$-ferroelectric polymer composite barriers.

described above for entirely oxide tunnel junctions. On the other hand, metallic (Josephson and magnetic) tunnel junctions are more reliable than oxide ones, and, in addition, ultrathin PVDF films are compatible with AlO_x so that composite ferroelectric-oxide barriers can be fabricated. The possible structures of superconducting and magnetic metallic tunnel junctions involving PVDF are shown schematically in Figure 6.22.

In conclusion of this section, it should be noted that in addition to the planar metal/ferroelectric/metal multilayers discussed above, heterostructures of other geometries may be useful for certain applications in nanoelectronics and may even display specific physical properties. Remarkably, one-dimensional structures in the form of ferroelectric nanowires (Urban et al. 2003) and nanotubes with inner and outer electrodes (Alexe et al. 2006) have been successfully fabricated. Ferroelectric quantum dots are of great interest as well, in particular, for electro-optical devices (Ye et al. 2000).

6.7 Summary and Outlook

The overview presented in this chapter demonstrates impressive achievements in the deposition, characterization, and theoretical description of nanoscale ferroelectric films and heterostructures. Remarkably, capacitors involving only a few nanometers of perovskite ferroelectrics were successfully fabricated, displaying a high remanent polarization. Several advanced analytical tools are now available, showing that high structural quality and sharp interfaces can be retained even in nanoscale capacitors. The basic physical effects in ferroelectric thin films, such as the strain and depolarizing-field effects, are already well understood theoretically, and the influence of short-range interactions at the ferroelectric–metal interfaces is intensively studied by first-principles calculations. Thus, all three constituents of the research triangle shown in Figure 6.1 are functioning effectively, which promises new advances in the physics of nanoscale ferroelectrics and their device applications in the near future.

References

Alexe, M., Hesse, D., Schmidt, V. et al. 2006. Ferroelectric nanotubes fabricated using nanowires as positive templates. *Applied Physics Letters* 89: 172907.

Bratkovsky, A. M. and Levanyuk, A. P. 2005. Smearing of phase transition due to a surface effect or a bulk inhomogeneity in ferroelectric nanostructures. *Physical Review Letters* 94: 107601.

Bune, A. V., Fridkin, V. M., Ducharme, S. et al. 1998. Two-dimensional ferroelectric films. *Nature* 391: 874–877.

Bungaro, C. and Rabe, K. M. 2004. Epitaxially strained [001]-$(PbTiO_3)_1(PbZrO_3)_1$ superlattice and $PbTiO_3$ from first principles. *Physical Review B* 69: 184101.

Bunshah, R. F. 1994. *Handbook of Deposition Technologies for Films and Coatings*, Park Ridge, NY: Noyes Publications.

Choi, K. J., Biegalski, M., Li, Y. L. et al. 2004. Enhancement of ferroelectricity in strained $BaTiO_3$ thin films. *Science* 306: 1005–1009.

Chopra, K. L. 1969. *Thin Film Phenomena*, New York: McGraw-Hill.

Christman, J. A., Kim, S. H., Maiwa, H. et al. 2000. Spatial variation of ferroelectric properties in $Pb(Zr_{0.3},Ti_{0.7})O_3$ thin films studied by atomic force microscopy. *Journal of Applied Physics* 87: 8031–8034.

Chu, W., Mayer, J., and Nicolet, M. 1978. *Backscattering Spectrometry*, New York: Academic Press.

Clayhold, J. A., Kerns, B. M., Schroer, M. D. et al. 2008. Combinatorial measurements of Hall effect and resistivity in oxide films. *Review of Scientific Instruments* 73: 033908.

Cohen, R. E. 1992. Origin of ferroelectricity in perovskite oxides. *Nature* 358: 136–138.

Dawber, M., Rabe, K. M., and Scott, J. F. 2005. Physics of thin-film ferroelectric oxides. *Reviews of Modern Physics* 77: 1083–1130.

Devonshire, A. F. 1954. Theory of ferroelectrics. *Advances in Physics* 3: 85–130.

Diéguez, O., Tinte, S., Antons, A. et al. 2004. Ab initio study of the phase diagram of epitaxial $BaTiO_3$. *Physical Review B* 69: 212101.

Ducharme, S., Palto, S. P., Fridkin, V. M., and Blinov, L. M. 2002. Ferroelectric Polymer Langmuir-Blodgett Films. In *Ferroelectric and Dielectric Thin Films*, Vol. 3 of *Handbook of Thin Films Materials*, ed. H. S. Nalwa, San Diego, CA: Academic Press.

Eom, C. B., Cava, R. J., Fleming, R. M. et al. 1992. Single-crystal epitaxial thin films of the isotropic metallic oxides $Sr_{1-x}Ca_xRuO_3$ $(0 < x < 1)$. *Science* 258: 1766–1769.

Fewster, P. F. 1996. X-ray analysis of thin films and multilayers. *Reports on Progress in Physics* 59: 1339–1407.

Fisher, O., Triscone, J. M., Fivat, P. et al. 1994. Investigations of coupled $DyBe_2Cu_3O_7$-$(Y_{1-x}Pr_x)$-$Be_2Cu_3O_7$ multilayer structures. *Superconducting Superlattices and Multilayers* 2157: 134–141.

Fong, D. D., Stephenson, G. B., Streiffer, S. K. et al. 2004. Ferroelectricity in ultrathin perovskite films. *Science* 304: 1650–1653.

Fong, D. D., Cionca, C., Yacoby, Y. et al. 2005. Direct structural determination in ultrathin ferroelectric films by analysis of synchrotron x-ray scattering measurements. *Physical Review B* 71: 144112.

Gajek, M., Bibes, M., Fusil, S. et al. 2007. Tunnel junctions with multiferroic barriers. *Nature Materials* 6: 296–302.

Garcia, R. and Perez, R. 2002. Dynamic atomic force microscopy methods. *Surface Science Reports* 47: 197–301.

Gerra, G., Tagantsev, A. K., and Setter, N. 2005. Surface-stimulated nucleation of reverse domains in ferroelectrics. *Physical Review Letters* 94: 107602.

Gerra, G., Tagantsev, A. K., Setter, N. et al. 2006. Ionic polarizability of conductive metal oxides and critical thickness for ferroelectricity in $BaTiO_3$. *Physical Review Letters* 96: 107603.

Gerra, G., Tagantsev, A. K., and Setter, N. 2007. Ferroelectricity in asymmetric metal-ferroelectric-metal heterostructures: A combined first-principles-phenomenological approach. *Physical Review Letters* 98: 207601.

Ghosez, P. and Rabe, K. M. 2000. Microscopic model of ferroelectricity in stress-free $PbTiO_3$ ultrathin films. *Applied Physics Letters* 76: 2767–2769.

Gruverman, A. and Kholkin, A. 2004. Ferroelectric nanodomains. In *Encyclopedia of Nanoscience and Nanotechnology*, ed. H. S. Nalwa, pp. 359–375. Los Angeles, CA: American Scientific Publishers.

Gruverman, A., Auciello, O., and Tokumoto, H. 1998. Imaging and control of domain structures in ferroelectric thin films via scanning force microscopy. *Annual Review of Materials Science* 28: 101–123.

Guro, G. M., Ivanchik, I. I., and Kovtonyuk, N. F. 1970. C-Domain barium titanate crystal in a short-circuited capacitor. *Soviet Physics Solid State* 11: 1574.

Güthner, P. and Dransfeld, K. 1992. Local poling of ferroelectric polymers by scanning force microscopy. *Applied Physics Letters* 61: 1137.

Habermeier, H. U. 2007. Thin films of perovskite-type complex oxides. *Materials Today* 10: 34–43.

Haeni, J. H., Theis, C. D., and Schlom, D. G. 2000. RHEED intensity oscillations for the stoichiometric growth of $SrTiO_3$ thin films by reactive molecular beam epitaxy. *Journal of Electroceramics* 4: 385–391.

Haeni, J. H., Irvin, P., Chang, W. et al. 2004. Room-temperature ferroelectricity in strained $SrTiO_3$. *Nature* 430: 758–761.

Hambe, M., Wicks, S., Gregg, J. M. et al. 2008. Creation of damage-free ferroelectric nanostructures via focused ion beam milling. *Nanotechnology* 19: 175302.

Haun, M. J., Furman, E., Jang, S. J. et al. 1987. Thermodynamic theory of $PbTiO_3$. *Journal of Applied Physics* 62: 3331–3338.

He, F. Z. and Wells, B. O. 2006. Lattice strain in epitaxial $BaTiO_3$ thin films. *Applied Physics Letters* 88: 152908.

Hubler, G. K. and Chrisey, D. B. 1994. *Pulsed Laser Deposition of Thin Films*, New York: John Wiley & Sons, Inc.

Huggins, H. A. and Gurvitch, M. 1985. Preparation and characterization of Nb/Al-oxide/Nb tunnel junctions. *Journal of Applied Physics* 57: 2103–2109.

Ishiwara, H. and Okuyama, M. 2004. *Ferroelectric Random Access Memories: Fundamentals and Applications*, Berlin, Germany: Springer.

Jia, C. L., Mi, S. B., Urban, K. et al. 2008. Atomic-scale study of electric dipoles near charged and uncharged domain walls in ferroelectric films. *Nature Materials* 7: 57–61.

Jo, J. Y., Kim, D. J., Kim, Y. S. et al. 2006a. Polarization switching dynamics governed by the thermodynamic nucleation process in ultrathin ferroelectric films. *Physical Review Letters* 97: 247602.

Jo, J. Y., Kim, Y. S., Noh, T. W. et al. 2006b. Coercive fields in ultrathin $BaTiO_3$ capacitors. *Applied Physics Letters* 89: 232909.

Jona, F. and Shirane, G. 1962. *Ferroelectric Crystals*, New York: MacMillan.

Junquera, J. and Ghosez, P. 2003. Critical thickness for ferroelectricity in perovskite ultrathin films. *Nature* 422: 506–509.

Kalinin, S. V. and Bonnell, D. A. 2002. Imaging mechanism of piezoresponse force microscopy of ferroelectric surfaces. *Physical Review B* 65: 125408.

Kato, Y., Kaneko, Y., Tanaka, H. et al. 2007. Overview and future challenge of ferroelectric random access memory technologies. *Japanese Journal of Applied Physics Part 1—Regular Papers Brief Communications & Review Papers* 46: 2157–2163.

Kawamura, E., Vahedi, V., Lieberman, M. A. et al. 1999. Ion energy distributions in rf sheaths; review, analysis and simulation. *Plasma Sources Science & Technology* 8: R45–R64.

Kholkin, A., Kalinin, S., Roelofs, A. et al. 2007. Review of ferroelectric domain imaging by piezoresponse force microscopy. In *Scanning Probe Microscopy: Electrical and Electromechanical Phenomena at the Nanoscale*, eds. S. Kalinin, and A. Gruverman, pp. 173–214. New York: Springer.

Kim, D. J., Jo, J. Y., Kim, Y. S. et al. 2005a. Polarization relaxation induced by a depolarization field in ultrathin ferroelectric $BaTiO_3$ capacitors. *Physical Review Letters* 95: 7602.

Kim, Y. S., Kim, D. H., Kim, J. D. et al. 2005b. Critical thickness of ultrathin ferroelectric $BaTiO_3$ films. *Applied Physics Letters* 86: 102907.

King-Smith, R. D. and Vanderbilt, D. 1993. Theory of polarization of crystalline solids. *Physical Review B* 47: 1651–1654.

King-Smith, R. D. and Vanderbilt, D. 1994. First-principles investigation of ferroelectricity in perovskite compounds. *Physical Review B* 49: 5828–5844.

Kohlstedt, H., Mustafa, Y., Gerber, A. et al. 2005a. Current status and challenges of ferroelectric memory devices. *Microelectronic Engineering* 80: 296–304.

Kohlstedt, H., Pertsev, N. A., Rodríguez Contreras, J. R. et al. 2005b. Theoretical current-voltage characteristics of ferroelectric tunnel junctions. *Physical Review B* 72: 125341.

Kohlstedt, H., Petraru, A., Szot, K. et al. 2008. Method to distinguish ferroelectric from nonferroelectric origin in case of resistive switching in ferroelectric capacitors. *Applied Physics Letters* 92: 062907.

Koukhar, V. G., Pertsev, N. A., and Waser, R. 2001. Thermodynamic theory of epitaxial ferroelectric thin films with dense domain structures. *Physical Review B* 64: 214103.

Kretschmer, R. and Binder, K. 1979. Surface effects on phase transitions in ferroelectrics and dipolar magnets. *Physical Review B* 20: 1065.

Kronholz, S., Rathgeber, S., Karthauser, S. et al. 2006. Self-assembly of diblock-copolymer micelles for template-based preparation of $PbTiO_3$ nanograins. *Advanced Functional Materials* 16: 2346–2354.

Ku, H. Y. and Ullman, F. G. 1964. Capacitance of thin dielectric structures. *Journal of Applied Physics* 35: 265–267.

Lai, B. K., Kornev, I. A., Bellaiche, L. et al. 2005. Phase diagrams of epitaxial $BaTiO_3$ ultrathin films from first principles. *Applied Physics Letters* 86: 132904.

Landau, L. D., Lifshitz, E. M., and Pitaevskii, L. P. 1984. *Electrodynamics of Continuous Media*, Oxford, NY: Pergamon.

Landauer, R. 1957. Electrostatic considerations in $BaTiO_3$ domain formation during polarization reversal. *Journal of Applied Physics* 28: 227–234.

Lee, H. N., Nakhmanson, S. M., Chisholm, M. F. et al. 2007. Suppressed dependence of polarization on epitaxial strain in highly polar ferroelectrics. *Physical Review Letters* 98: 217602.

Lettieri, J., Haeni, J. H., and Schlom, D. G. 2002. Critical issues in the heteroepitaxial growth of alkaline-earth oxides on silicon. *Journal of Vacuum Science and Technology A* 20: 1332.

Levanyuk, A. P. and Sigov, A. S. 1988. *Defects and Structural Phase Transitions*, Amsterdam, the Netherlands: Gordon and Breach.

Lines, M. E. and Glass, A. M. 1977. *Principles and Applications of Ferroelectrics and Related Materials*, Oxford, NY: Oxford University Press.

Logvenov, G. and Bozovic, I. 2008. Artificial superlattices grown by MBE: Could we design novel superconductors? *Physica C—Superconductivity and Its Applications* 468: 100–104.

Maissel, L. I. and Glang, R. 1979. *Handbook of Thin Film Technology*, New York: McGraw-Hill.

Mehta, R. R., Silverman, B. D., and Jacobs, J. T. 1973. Depolarization fields in thin ferroelectric films. *Journal of Applied Physics* 44: 3379–3385.

Moodera, J. S., Lisa R. K., Wong, T. M., and Meservey, R. 1995. Large magnetoresistance at room temperature in ferromagnetic thin film tunnel junctions. *Physical Review Letters* 74: 3273–3276.

Muller, K. A. and Burkard, H. 1979. $SrTiO_3$—Intrinsic quantum paraelectric below 4 K. *Physical Review B* 19: 3593–3602.

Nagarajan, V., Junquera, J., He, J. Q. et al. 2006. Scaling of structure and electrical properties in ultrathin epitaxial ferroelectric heterostructures. *Journal of Applied Physics* 100: 051609.

Nevot, L. and Croce, P. 1980. Characterization of surfaces by grazing x-ray reflection—Application to study of polishing of some silicate-glasses. *Revue De Physique Appliquee* 15: 761–779.

Nicolaou, K. C., Hanko, R., and Hartwig, W. 2002. *Handbook of Combinatorial Chemistry: Drugs, Catalysts, Materials*, Weinheim, Germany: Wiley-VCH Verlag GmbH.

Ohtani, M., Lippmaa, M., Ohnishi, T. et al. 2005. High throughput oxide lattice engineering by parallel laser molecular-beam epitaxy and concurrent x-ray diffraction. *Review of Scientific Instruments* 76: 062218.

Oikawa, T., Morioka, H., Nagai, A. et al. 2004. Thickness scaling of polycrystalline $Pb(Zr,Ti)O_3$ films down to 35 nm prepared by metalorganic chemical vapor deposition having good ferroelectric properties. *Applied Physics Letters* 85: 1754–1756.

Pertsev, N. A. and Kohlstedt, H. 2007. Elastic stabilization of a single-domain ferroelectric state in nanoscale capacitors and tunnel junctions. *Physical Review Letters* 98: 257603.

Pertsev, N. A. and Zembilgotov, A. G. 1991. Microscopic mechanism of polarization switching in polymer ferroelectrics. *Soviet Physics—Solid State* 33: 165–175.

Pertsev, N. A., Zembilgotov, A. G., and Tagantsev, A. K. 1998. Effect of mechanical boundary conditions on phase diagrams of epitaxial ferroelectric thin films. *Physical Review Letters* 80: 1988–1991.

Pertsev, N. A., Tagantsev, A. K., and Setter, N. 2000. Phase transitions and strain-induced ferroelectricity in $SrTiO_3$ epitaxial thin films. *Physical Review B* 61: R825–828; 2002. Erratum. *Physical Review B* 65: 219901.

Pertsev, N. A., Kukhar, V. G., Kohlstedt, H. et al. 2003a. Phase diagrams and physical properties of single-domain epitaxial $Pb(Zr_{1-x}Ti_x)O_3$ thin films. *Physical Review B* 67: 054107.

Pertsev, N. A., Rodríguez Contreras, J., Kukhar, V. G. et al. 2003b. Coercive field of ultrathin $Pb(Zr_{0.52}Ti_{0.48})O_3$ epitaxial films. *Applied Physics Letters* 83: 3356–3358.

Pertsev, N. A., Dittmann, R., Plonka, R. et al. 2007. Thickness dependence of intrinsic dielectric response and apparent interfacial capacitance in ferroelectric thin films. *Journal of Applied Physics* 101: 074102.

Petraru, A., Pertsev, N. A., Kohlstedt, H. et al. 2007. Polarization and lattice strains in epitaxial $BaTiO_3$ films grown by high-pressure sputtering. *Journal of Applied Physics* 101: 114106.

Petraru, A., Kohlstedt, H., Poppe, U. et al. 2008. Wedgelike ultrathin epitaxial $BaTiO_3$ films for studies of scaling effects in ferroelectrics. *Applied Physics Letters* 93: 072902.

Poppe, U., Schubert, J., Arons, R. R. et al. 1988. Direct production of crystalline superconducting thin-films of $YBa_2Cu_3O_7$ by high-pressure oxygen sputtering. *Solid State Communications* 66: 661–665.

Resta, R. 1994. Macroscopic polarization in crystalline dielectrics—The geometric phase approach. *Reviews of Modern Physics* 66: 899–915.

Rodríguez, B. J., Gruverman, A., Kingon, A. I. et al. 2002. Piezoresponse force microscopy for polarity imaging of GaN. *Applied Physics Letters* 80: 4166–4168.

Rodríquez Contreras, J., Schubert, J., Poppe, U. et al. 2002. Structural and ferroelectric properties of epitaxial $PbZr_{0.52}Ti_{0.48}O_3$ and $BaTiO_3$ thin films prepared on $SrRuO_3$/$SrTiO_3$(100) substrates. *Materials Research Society Symposium Proceedings* 688: C8.10.

Rodríguez Contreras, J. 2004. *Ferroelectric Tunnel Junctions*, Jülich, Germany: Forschungszentrum Jülich GmbH Verlag.

Rodríguez Contreras, J., Kohlstedt, H., Poppe, U. et al. 2003a. Surface treatment effects on the thickness dependence of the remanent polarization of $PbZr_{0.52}Ti_{0.48}O_3$ capacitors. *Applied Physics Letters* 83: 126–128.

Rodríguez Contreras, J., Kohlstedt, H., Poppe, U. et al. 2003b. Resistive switching in metal-ferroelectric-metal junctions. *Applied Physics Letters* 83: 4595–4597.

Sai, N., Kolpak, A. M., and Rappe, A. M. 2005. Ferroelectricity in ultrathin perovskite films. *Physical Review B* 72: 020101(R).

Schlom, D. G., Chen, L. Q., Eom, C. B. et al. 2007. Strain tuning of ferroelectric thin films. *Annual Review of Materials Research* 37: 589–626.

Schneller, T. and Waser, R. 2007. Chemical modifications of $Pb(Zr_{0.3},Ti_{0.7})O_3$ precursor solutions and their influence on the morphological and electrical properties of the resulting thin films. *Journal of Sol-Gel Science and Technology* 42: 337–352.

Schuller, I. K. 1980. New class of layered materials. *Physical Review Letters* 44: 1597–1600.

Scott, J. F. 2000. *Ferroelectric Memories*, Berlin, Germany: Springer.

Scott, J. F. 2007. Applications of modern ferroelectrics. *Science* 315: 954–959.

Setter, N., Damjanovic, D., Eng, L. et al. 2006. Ferroelectric thin films: Review of materials, properties, and applications. *Journal of Applied Physics* 100: 051606.

Slonczewski, J. C. 1989. Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier. *Physical Review B* 39: 6995–7002.

Smolenskii, G. A., Bokov, V. A., Isupov, V. A. et al. 1984. *Ferroelectrics and Related Materials*, New York: Gordon and Breach.

Spatz, J. P., Mossmer, S., Hartmann, C. et al. 2000. Ordered deposition of inorganic clusters from micellar block copolymer films. *Langmuir* 16: 407–415.

Sun, J. Z. 1998. Thin-film trilayer manganate junctions. *Philosophical Transactions of the Royal Society of London Series A—Mathematical Physical and Engineering Sciences* 356: 1693–1711.

Sun, J. Z. 2001. Spin-dependent transport in trilayer junctions of doped manganites. *Physica C—Superconductivity and Its Applications* 350: 215–226.

Szafraniak-Wiza, I., Hesse, D., and Alexe, M. 2008. *Nanosized Ferroelectric Crystals*, Cambridge, U.K.: Woodhead Publishing, Limited.

Tagantsev, A. K. and Gerra, G. 2006. Interface-induced phenomena in polarization response of ferroelectric thin films. *Journal of Applied Physics* 100: 051607.

Tagantsev, A. K., Pertsev, N. A., Muralt, P. et al. 2002. Strain-induced diffuse dielectric anomaly and critical point in perovskite ferroelectric thin films. *Physical Review B* 65: 012104.

Tagantsev, A. K., Gerra, G., and Setter, N. 2008. Short-range and long-range contributions to the size effect in metal-ferroelectric-metal heterostructures. *Physical Review B* 77: 174111.

Urban, J. J., Spanier, J. E., Lian, O. Y. et al. 2003. Single-crystalline barium titanate nanowires. *Advanced Materials* 15: 423–426.

Valasek, J. 1921. Piezo-electric and allied phenomena in Rochelle salt. *Physical Review* 17: 475–481.

Waghmare, U. V. and Rabe, K. M. 1997. Ab initio statistical mechanics of the ferroelectric phase transition in $PbTiO_3$. *Physical Review B* 55: 6161–6173.

Weil, R., Nkum, R., Muranevich, E. et al. 1989. Ferroelectricity in zinc-cadmium telluride. *Physical Review Letters* 62: 2744–2746.

Xu, Y. 1991. *Ferroelectric Materials and Their Applications*, Amsterdam, the Netherlands: North-Holland.

Ye, H., Xu, Y. H., and Mackenzie, J. D. 2000. Semiconducting ferroelectric SbSl quantum dots in organically modified TiO_2 matrix. *Proceedings of SPIE* 3943: 95–101.

Zembilgotov, A. G., Pertsev, N. A., Kohlstedt, H. et al. 2002. Ultrathin epitaxial ferroelectric films grown on compressive substrates: Competition between the surface and strain effects. *Journal of Applied Physics* 91: 2247–2254.

Zhuravlev, M. Y., Jaswal, S. S., Tsymbal, E. Y. et al. 2005b. Ferroelectric switch for spin injection. *Applied Physics Letters* 87: 222114.

Zhuravlev, M. Y., Sabirianov, R. F., Jaswal, S. S. et al. 2005a. Giant electroresistance in ferroelectric tunnel junctions. *Physical Review Letters* 94: 246802.

7

Superconducting Weak Links Made of Carbon Nanostructures

7.1 Introduction 7-1
7.2 Superconducting Transport through a Weak Link 7-2
Density of Electronic States in a Superconductor • Coherent Transport at a Normal/ Superconducting Interface • A Brief Introduction to the Josephson Effect • Normal Electron Transport through an Atomic or Molecular Conductor • Superconducting Transport through Andreev Bound States • Introduction to Low-Dimensional sp^2 Carbon Structures • Gate Control of the Weak Link Transparency
7.3 Superconducting Transport in a Carbon Nanotube Weak Link 7-6
Introduction to Carbon Nanotube Devices • Experimental Realizations of Carbon Nanotube Weak Links • Nanotube Quantum Dot Connected to Superconducting Electrodes • High- Frequency Irradiation of Nanotube Weak Links
7.4 Nanotube-Based Superconducting Quantum Interferometers 7-10
Quantum Interference with Single Nanotube Weak Links • Tunability of the Phase Shift across Nanotube Weak Links • Application of Nanotube Weak Links to Magnetometry • Application of Nanotube Weak Links to Quantum Information
7.5 Graphene-Based Superconducting Weak Links 7-15
Proximity Effect in Graphene Weak Links • Graphene Nanotube Superconducting Quantum Interferometers
7.6 Fullerene-Based Superconducting Weak Links 7-17
Molecular Transistors Made by Electromigration • Implementation of a C60 Superconducting Transistor
7.7 Concluding Remarks 7-18
Acknowledgment 7-18
References 7-18

Vincent Bouchiat
Centre National de la Recherche Scientifique

7.1 Introduction

It has been known since the 1960s (De Gennes and Guyon 1963) that charge carriers emitted from a superconductor can propagate in a non-superconducting medium, provided that the electron wave functions of the two electrons originating from the superconductor (the so-called Cooper pairs) keep their correlation during their diffusion outside the superconductor. This process requires the absence of defects in the propagating medium that break time-reversal symmetry, such as magnetic impurities. These correlated electrons form an evanescent state that "bleeds" from the superconductors into the non-superconducting electrode on mesoscopic distances. The physics of these evanescent states, known as the "superconducting proximity effect" depends on many parameters (the coherence length of the superconductor, the contact barrier at the interface, the temperature, the electron diffusion length, etc.).

It appeared then possible to sever a superconducting electrode with non-superconducting elements (a narrow constriction, a normal metal, a thin insulating barrier, known as "weak links") while keeping a supercurrent flow. As tunneling processes are known to depend exponentially on the thickness of oxide barriers, controlling the structure and the geometry of weak links down to the atomic scale appears to be of capital importance if one wants to get reproducible devices. Therefore, improving the miniaturization and the crystalline quality of these weak links was clearly identified as a critical issue. Among the possible candidates are sp^2-hybridized carbon nanostructures (i.e., graphenes, carbon nanotubes [CNTs], or fullerenes), which, among other advantages, offer good control of the crystallinity and provide ideal media to control a supercurrent flow. Quantum confinement of electronic states within the carbon nanostructure generates sharp variations of the electronic density of states. Tuning the weak link chemical potential with an electrostatic gate allows to adjust the weak link transparency.

The purpose of this chapter is to propose a basic overview of the principles that govern the physics of these carbon based devices and illustrate them by presenting experimental demonstrations and some applications.

7.2 Superconducting Transport through a Weak Link

7.2.1 Density of Electronic States in a Superconductor

At energies around the chemical potential μ and below the energy gap Δ, a superconductor (noted in the following by the symbol S) only has a single ground state, which depends on a macroscopic phase. In a bulk superconductor, all electrons are in the same quantum state characterized by a global phase φ. The charge carriers are condensed in a bound state of two electrons, the so-called Cooper pair, which has a charge of $-2e$ (Figure 7.1a). The ground state wave function depends on the phase φ and can be written as

$$\Delta \approx \sqrt{n_s}\exp(i\varphi), \tag{7.1}$$

where n_s is the total number of Cooper pairs.

Δ is the "superconductor order parameter" and specifies the choice for the quantum phase φ. There are no other states present in the interval $[\mu - \Delta, \mu + \Delta]$ because it takes an energy 2Δ to break a Cooper pair. An electron in this energy window cannot be injected in a superconductor alone, generating a "gap" in the transmission of charge carriers at a normal/superconductor metal interface.

7.2.2 Coherent Transport at a Normal/Superconducting Interface

At the interface between a normal metal (N) and a superconductor (S), an electron at the Fermi level can nevertheless be converted into a Cooper pair with zero momentum if a hole (charge $+e$) is reflected with energy $-\varepsilon$ and opposite momentum, conserving the charge, energy, and momentum. The reflected hole acquires an additional phase factor φ, the order parameter phase. This conversion of an electron–hole pair into a Cooper pair is a process known as the Andreev reflection (Figure 7.1b). It dominates the charge transport at low bias through the interface. This Andreev reflection can lead to a supercurrent transport if correlated charge carriers are collected by a second superconductor placed in a series. The process is still effective at the interface between superconducting electrodes and nanoscale constriction, which we will refer to as a "weak link."

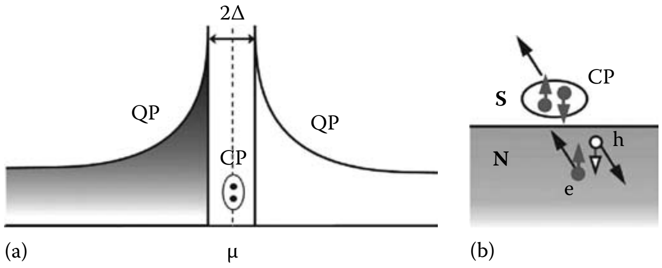

FIGURE 7.1 (a) Density of state of a superconducting electrode: all the electrons are in the ground state wave function at the chemical potential μ. (b) Schematics depicting the Andreev reflection. Describing coherent transport at a normal/superconducting interface below the gap. An electron (e) impinging the interface between a normal conductor (N) and a superconductor (S) produces a Cooper pair (CP, oval) in the top superconducting electrode and a retro-reflected hole (h) in the normal conductor. Vertical arrows indicate the spin band occupied by each particle.

7.2.3 A Brief Introduction to the Josephson Effect

The order parameter characterizing the superconducting state varies on a spatial scale equal to $\xi = hv_F/\Delta$, where v_F is the Fermi velocity. The quantity ξ, which is a measure of the size of Cooper pairs, is called the coherence length. At the interface with an insulator, a metal, or a nano-object, the superconducting order parameter $\Delta(\mathbf{r})$ (its wave function) weakens gradually close to the interface on a characteristic scale, which is on the order of ξ.

When two superconductors with phase φ_1 and φ_2 are separated by a weak link (see Figure 7.2), they interact through direct tunneling or by the Andreev reflection process presented in the previous paragraph. This sandwich structure is called a Josephson junction. It is dominated by the interaction energy between the two superconductors. In 1962, Brian Josephson predicted that due to the interaction between the two superconducting condensates in close proximity, a non-superconducting nanostructure can let flow a superconducting current I, the value of which depends on the phase twist between the two condensates (Figure 7.2) (Josephson 1962). For a thin insulating tunnel barrier, he has shown that the current has the following form:

$$I = I_c \sin(\varphi_2 - \varphi_1). \tag{7.2}$$

This surprising prediction (Josephson 1962) was rapidly experimentally confirmed, leading to a new field in superconducting electronics. Until the mid-1990s, such superconducting "weak links" were either made of insulating barriers (the so-called Josephson tunnel junctions, at the origin of all the recent advances in superconducting Qubits), narrow diffusive metallic

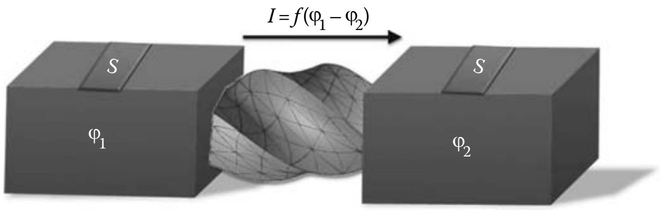

FIGURE 7.2 Schematics of a Josephson weak link. In this device, two superconducting electrodes, characterized by their phase φ_1 and φ_2 are separated by a weak link (symbolized by a twisted solid). The superconducting current I flowing through the weak link is a given function f of the "phase twist" i.e., the phase difference $\varphi_1 - \varphi_2$. This current–phase function f exhibits a 2π-periodicity and is highly non linear. It is the main characteristics of a Josephson weak link.

or semiconducting wires, or point contacts inserted between bulk electrodes (Likharev 1979). The first realization of a ballistic weak link with a limited number of conduction channels and controlled at the atomic scale came in 1997, with the measurement transport through gold atomic contacts in close vicinity of superconducting aluminum contacts (Scheer et al. 1997). In the last 5 years, it has been shown that carbon nanostructures can provide another class of weak links of good structural quality, for which strong quantum confinement offers gate tunability.

7.2.4 Normal Electron Transport through an Atomic or Molecular Conductor

At the atomic level, electron transport occurs by hopping through the atomic or molecular orbitals that are quantized in the case of a laterally confined nanostructure. They provide discrete conducting channels for electron transport (Imry and Landauer 1999). At the interface with a nanostructure, these contributions are related to the overlaps between molecular orbitals and the electronic wave functions in the metal. The physics of this quantum transport follows the description of Rolf Landauer from the 1970s, which refers to treating incoming charge carriers as plane waves traveling through "conduction channels" that are the eigen-modes for electronic propagation within the nanostructure. The electronic wave is partially transmitted and reflected at the interface. The theoretical work by Landauer and subsequently confirmed by numerous experiments have shown that the conductance of an electronic mesoscopic conductor is indeed divided into a discrete sum over all conduction channels *i* placed in parallel:

$$G = \sum_i G_i = \sum_i G_0 T_i = \frac{2e^2}{h} \sum_i T_i. \tag{7.3}$$

Each channel bears a quantized conductance G_0 called the quantum of conductance given by a ratio of two fundamental constants, $G_0 = 2e^2/h = (12.8\,\text{k}\Omega)^{-1}$. In practice, each conduction channel is partially coupled to the modes of the reservoirs. This induces a reduction of the conductance of a dimension-less prefactor T_i with $0 < T_i < 1$. The origin of these reduction factors is essentially of technologic origin and can be optimized by choosing the right interface materials as shown below. Nevertheless, even for a perfectly transmitting interface ($T = 1$), the normal transport through a nanoscale conductor is limited to a resistive transport of $12\,\text{k}\Omega$ per channel, the resistance being localized at the interface between the reservoir and the leads (see Figure 7.8a for an example).

In the case of coupling to superconducting electrodes, the Landauer formula 7.3 is no more valid: it is possible to cancel this contact resistance and obtain a zero-resistance electron flow through an atomic-sized contacts. This simple observation might have important technical consequences in the future of nanoelectronics, for example, to limit heat dissipation in the case of the ultra-high density integration of nanodevices involving few conduction channels.

7.2.5 Superconducting Transport through Andreev Bound States

As shown in Section 7.2.2, the microscopic transfer of charges at a superconducting interface arises by an electron–hole process known as the Andreev reflection. In a weak link, there are two normal/superconducting interfaces back-to-back (S/weak link and weak link/S). The two proximity effects on each side percolate and give rise to a superconducting current flow through the weak link. Furthermore, one has to consider the confinement in the weak link of the electron and holes created by the successive Andreev reflections at the interfaces: the electron and the holes remain spatially and energetically confined (Figure 7.3), thus defining a quantum well. The eigenstates of this well, the so-called Andreev bound states, require constructive interferences to propagate. The electron state, once converted into its hole conjugate, is converted back into the same electron state with the same quantum phase. This defines a quantization condition that depends on the phase difference between the superconductors: they shift the electron–hole Andreev wave function by their phase $\exp(i\varphi)$.

When no scattering occurs between the superconductors, there are two bound states per conductance channel with the energies $E_\pm = \pm\Delta\cos(\phi/2)$. More realistically, in the presence of a finite transmission T_i, their energies become $E_\pm = \pm\Delta\sqrt{1-\sin^2(\phi/2)}$ (Goffman et al. 2000). At low temperatures, only the Andreev-bound state lowest in energy is occupied. The Josephson superconducting currents carried by the different channels add their contribution to each other and for the superconducting current give

$$I = -\frac{\partial \sum_i E_-(\phi, T_i)}{\partial \phi} = \frac{e\Delta}{2\hbar} \sum_i T_i \frac{\sin\phi}{\sqrt{1 - T_i \sin^2(\phi/2)}}. \tag{7.4}$$

This "current-phase" relation, linking the supercurrent *I* to the phase shift ϕ across the ballistic weak link is a characteristic of Andreev processes and more generally will be the fingerprint of any superconducting weak link (Golubov et al. 2004). From that last formula, one can see that the "critical" current I_c, defined as the maximum supercurrent that can flow through the weak link with a length shorter than the superconducting coherence length, is quantized. Its maximum value is equal to

$$I_c = \frac{Ne\Delta}{\hbar}. \tag{7.5}$$

In the case of a nanodevice that can be modeled by *N* ballistic conduction channels placed in parallel, the quantum interference of the Andreev electrons bouncing back and forth between the two S/N and N/S serial interfaces at the junction boundaries is also at the origin of resonance currents called "Multiples Andreev reflections" (Klapwijk et al. 1982), which occur at voltages below the superconducting gap Δ (see examples in Figure 7.18). Due to resonating energy conditions they occur at bias

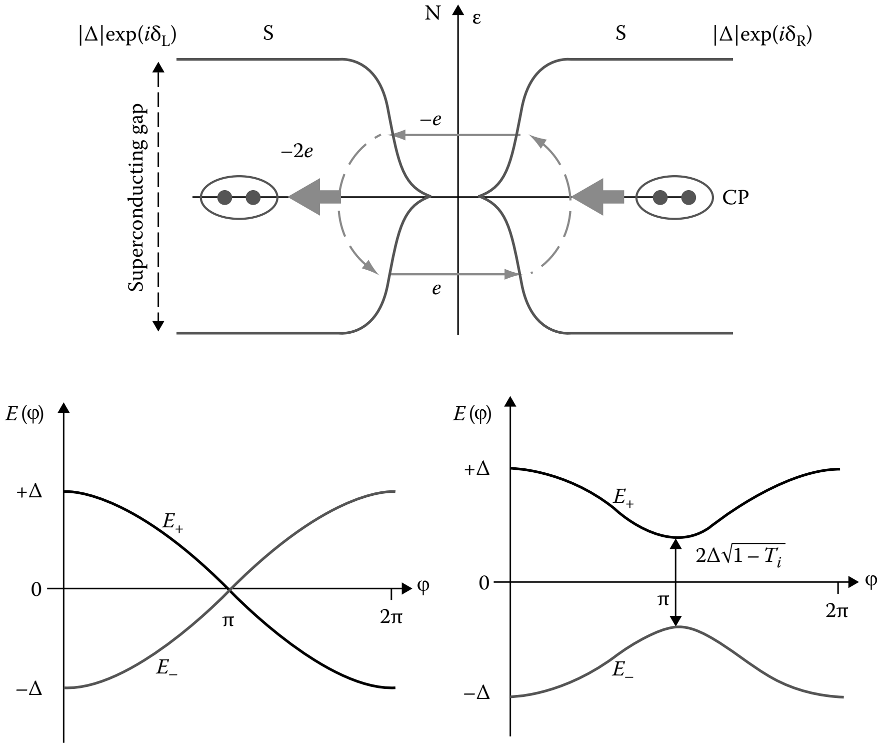

FIGURE 7.3 Top: Schematics depicting superconducting transport through a normal weak link N sandwiched between two superconducting electrodes. An electron is reflected as a hole at the interface with the left superconductor and the hole is reflected back as an electron on the right interface. The net result is a transfer of one Cooper pair through the S/N/S junction. Bottom: Phase dependence of the Andreev bound states energy levels ($\pi=\delta_r-\delta_l$). Left: For a perfectly transmitted channel without scattering potential ($T_i = 1$). Right: For a partially transmitted channel $T_i < 1$. The lower energy state is the only occupied state at low temperature.

voltages corresponding to the sub-harmonics of the superconducting gap ($1/n^*(2\Delta/e)$), where n is an integer. These multiple Andreev Reflections, which position are independent of the weak link density of state but are strongly affected by scattering, will then provide the most convincing experimental signatures of a ballistic weak link.

7.2.6 Introduction to Low-Dimensional sp^2 Carbon Structures

Low-dimensional allotropes based on carbon nanostructures with sp^2-hybridized bonds (Figure 7.4a), such as CNTs, graphenes, and fullerenes, have been extensively studied in the last decade (see appropriate chapters in the Handbook for immediate reference). Their exceptional electronic properties could provide alternative solutions in many fields of electronics including industrial applications for logic. They also behave as ideal media for transmitting superconducting charge carriers for several reasons.

First, they experimentally provide a quasi-perfect crystalline semiconducting lattice on a rather large scale length (typically a fraction of a micron, which corresponds to several thousands of C–C sp^2-hybridized bonds in every direction). Second, the inertness and absence of dangling bonds on the carbon surface solves the problem of surface termination (passivation), which is usually found in almost all other systems including all semiconductors candidates. This allow to make atomically clean edges and boundaries. This important feature enables the fabrication of reliable nanoscale metal/sp^2 carbon contacts with reproducible transparency.

7.2.7 Gate Control of the Weak Link Transparency

As we will detail in the following paragraph, the quantum confinement created by the low dimensionality of carbon nano-objects is at the origin of a sharp dependence of the electron density of states with the chemical potential of the weak link (Figure 7.4a). The resulting Andreev bound states will then be directly affected by this confinement and will create gate-tunable "filters" for the coupling between the two superconducting electrodes (Figure 7.4b). In the superconducting state, a supercurrent will flow if an Andreev bound state is aligned with respect to the Fermi level of the leads as depicted in paragraph 7.2.5. A voltage V_g applied on an electrostatic gate allows for a shift in the chemical potential of the weak link, which is equivalent to translating the confined energy levels (Figure 7.4b) with respect to the Fermi level of the leads. This will in turn tune the maximum superconducting current through the junction, thus implementing a "gate tunable Cooper pair filter." In the following section, we provide details of the typical cases for the different types of nanocarbon structures considered.

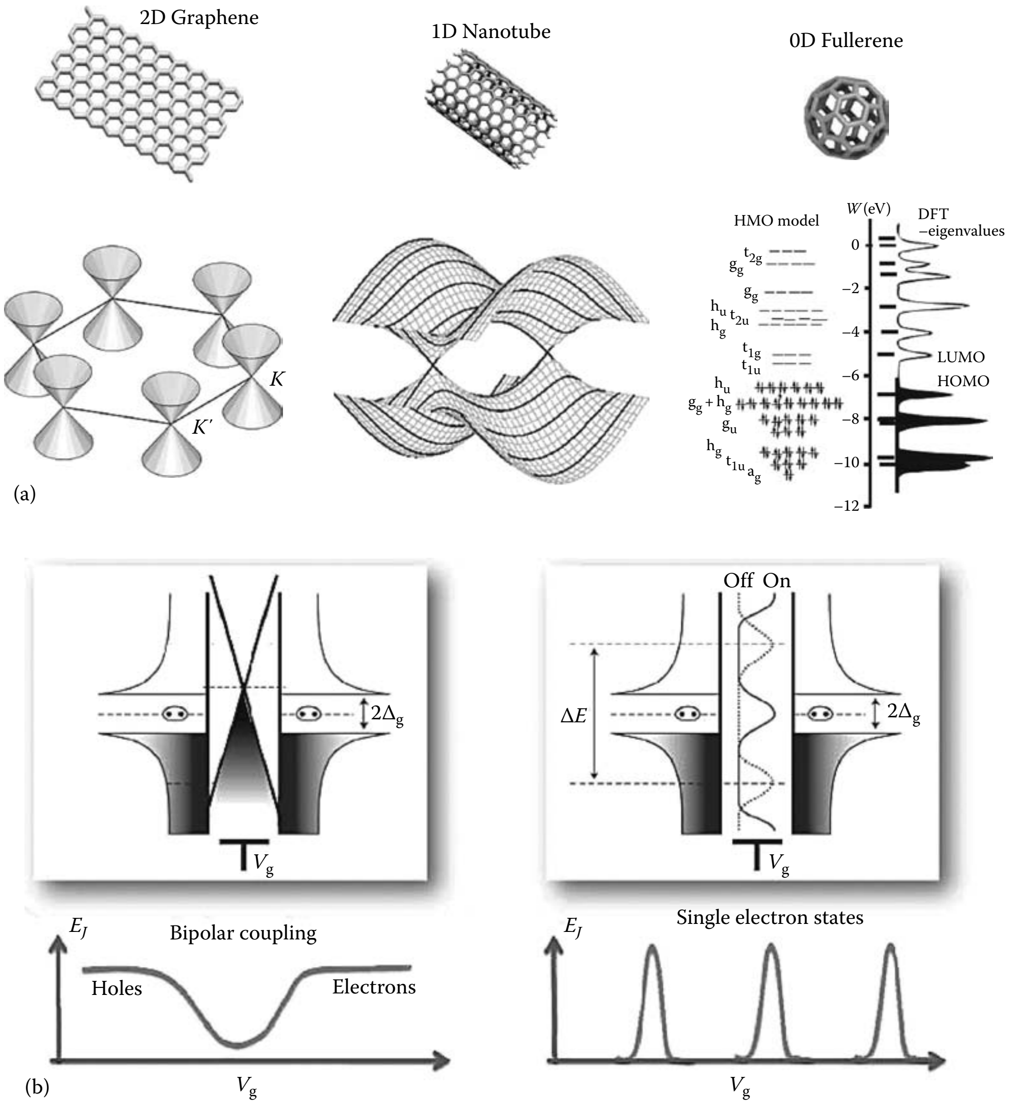

FIGURE 7.4 (a) Crystalline structure of the three low-dimensional phases of sp^2 carbon allotropes (top), and its associated densities of states (bottom). Left: in Graphene, the two dimensional density of states is described by a set of two "Dirac" cones that intersect the Brillouin zone at points K and K'. These cones provide a linear gap-less dispersion curve. Middle: Metallic single-walled carbon nanotube. The periodic boundary conditions generated by the rolled-up graphene sheet generate solutions symbolized by thick black lines. For selected chiral angles and diameter of the nanotube, these lines intersect the K and K' points of the Brillouin zone, leading to metallic nanotubes which behave as perfect 1D channels. Right: C60 and derivated fullerenes. The energy spectrum shows discrete states, whose positions could be derived through molecular orbitals (HMO, left side) or density functional theory (DFT, right side of axis). The molecule shows a gap between the occupied and empty states of about 1 eV. (b) "Semiconducting" representations of the energy levels of a Superconductor/Nano-carbon/Superconductor junctions for two dimensional (left) and zero dimensional (right) weak links. The chemical potential of the weak link can be translated by adjustment of a gate voltage V_g which allows to tune the transmission of the Cooper pairs through the weak link. Left: For a 2D junction (i.e., graphene-based weak links), the transport of Cooper pairs can either be transmitted by the valence or the conduction band, leading to an ambipolar semiconductor-like transmission function. At the charge neutrality point (position where the valence and conduction cones intersect, superconducting transport is minimal and is mediated through evanescent waves. Right: for a quantum dot junction (short nanotube junction or fullerene), superconducting transport occurs when a single electron or single hole energy levels is aligned with the Fermi level of the electrode. Such discrete energy spectrum of the weak link acts as a tunable filter for superconducting transport and leads to a series of peaks in the gate dependence of the maximum superconducting current.

7.3 Superconducting Transport in a Carbon Nanotube Weak Link

7.3.1 Introduction to Carbon Nanotube Devices

CNTs are cylinders of carbon of about one nanometer in diameter and up to several micron in length, which can be seen as a rolled-up sheet of sp^2 carbon "graphene." Their cylindrical shape and original 1D electronic structure (for a review, see Charlier et al. 2007) constitute the unique experimental realizations of a quasi-perfect 1D electronic system.

Depending on the chiral angle of the crystalline direction with the cylinder axis, CNTs can exhibit either metallic or semiconducting properties. For the sake of simplicity, one assumes in the following that the nanotube is single-walled and is of a metallic variety (Figure 7.4).

In practice, semiconducting nanotubes are so heavily doped by the metal contacting electrodes that they can also behave in a similar fashion.

A metallic single-walled CNT has two channels of conduction electrons available for electron conduction. According to the Landauer Formula (see Equation 7.3), when perfectly connected to a normal state electrode, it has a resistance that equals the fundamental constant called resistance quantum $h/4e^2$ = 6.5 kΩ. Due to the mismatch of the electron orbitals in the nanotube and in the metal electrodes, the transmission is generally not perfect. The result is equivalent to a tunnel barrier at the contact ($T < 1$). This induces a probability of transmission between the electrode and the nanotube that can be less than unity. These two semi-transparent electronic junctions placed in a series are in all means reminiscent to what is occurring for photon beams in a "Fabry–Perot" cavity (Liang et al. 2001) in which a photon is trapped between two semi-reflective mirrors. Such a cavity has internal resonating modes that correspond to stationary waves given by integer multiples of the half-wavelength and the resonating mode width (invert of the so-called finesse of the cavity) increases linearly with the mirror transparency. Similarly, the CNT junction behaves as a "quantum dot" that the gap between levels (normal modes) increases proportionally to $1/L$ where L is the nanotube length. The width of the levels is proportional to the transmission of contacts. For a short nanotube junction, we obtain discrete levels when the nanotube is completely decoupled from the electrodes like the energy level of an isolated molecule (see the case of C60 Figure 7.4a). These levels widen a bit when the quantum dot interacts with contacts. The position of the nanotube levels with respect to the energy level of the contacts can be adjusted by varying the electrostatic voltage applied to a gate electrode. The transmission of electrons is maximal when the energy level of the quantum dot coincides with the contacts. It thus realizes a molecular transistor (Figure 7.9, left) in which junction transparency can be periodically tuned with a voltage applied to the electrostatic gate (Figure 7.6).

7.3.2 Experimental Realizations of Carbon Nanotube Weak Links

Progress in the late 1990s in improving the contact of CNTs to superconducting electrodes showed that such junctions can exhibit superconducting fluctuations that can be gate controlled (Morpurgo et al. 1999) and that they can accommodate a superconducting current (Kasumov et al. 1999). In 2006, it was then shown that such junctions behave as gate-controlled Josephson junctions (Jarillo-Herrero et al. 2006, Jørgensen et al. 2006). During this period, fabrication techniques have been optimized to generate low-resistive contacts at the metal/sp^2 carbon interfaces. They required the use of specific metals for connection such as titanium (Jarillo-Herrero et al. 2006) or palladium (Cleuziou et al. 2006, Javey et al. 2003) that both provide good wetting properties onto the sp^2 carbon layer and minimize the Schottky barrier that could arise from the differences of the electronic work functions at the metal/carbon interface. More subtle effects linked to the hybridization of the metal orbitals to sp^2 carbon (Nemec et al. 2006) could explain the differences

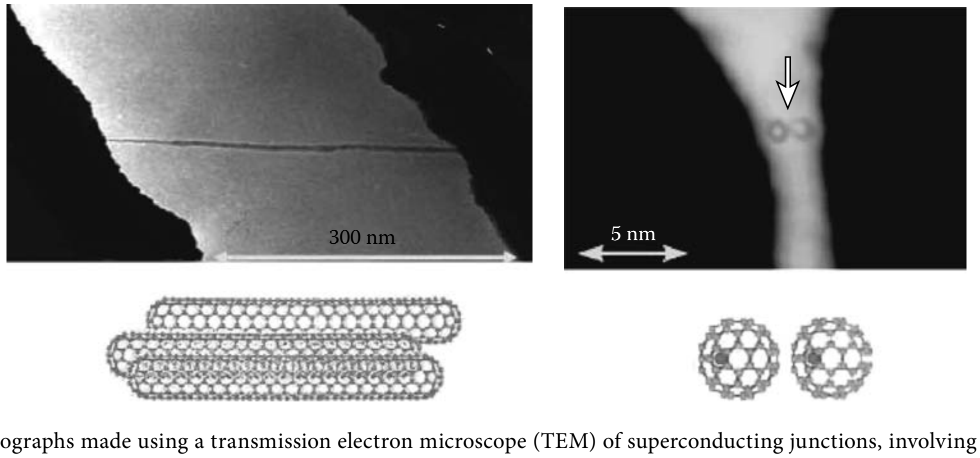

FIGURE 7.5 Micrographs made using a transmission electron microscope (TEM) of superconducting junctions, involving a carbon nanotube bundle, left and for a fullerene dimer formed by two C82 molecules in series, right. The weak links are free-standing over a slit and bridge superconducting electrodes deposited on a silicon membrane. This allows TEM imaging and electrical measurements on the same samples. (After Kasumov, A.Y. et al., *Science*, 284, 1508, 1999. With permission.)

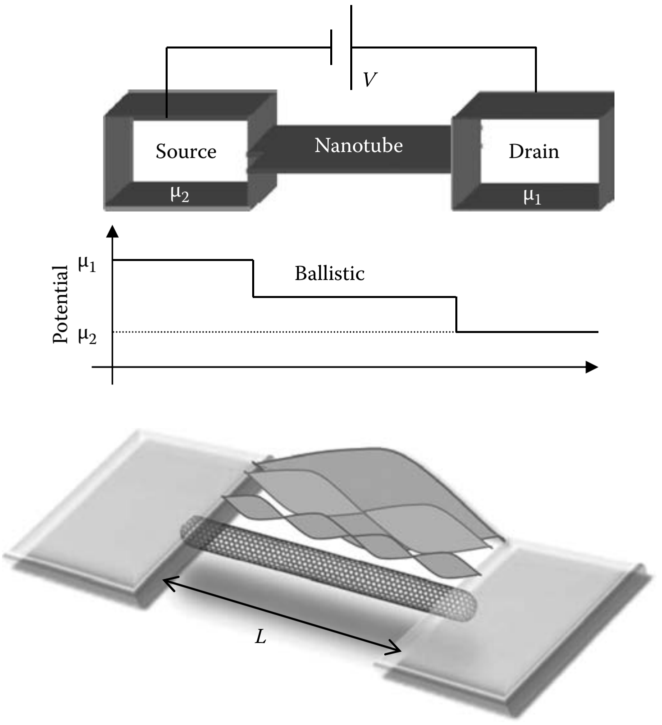

FIGURE 7.6 Top schematic of electron transport through a "defect-less" carbon nanotube contacted with two metallic electrodes. Electron transport is ballistic along the nanotube: the voltage drop $V = \mu_1 - \mu_2$ is occurring at the metal/nanotube contacts which act as tunnel barriers partially transmitting the electron. Bottom: For low resistance barrier contacts, the device act as a Fabry–Perot interferometer in which standing electron waves are separated by the energy difference $\Delta E = hv_F/2L$.

of interface resistances that have been reproducibly observed between these different metals (see Figure 7.7).

While not superconducting by itself, the Pd or Ti interfacing layer is made thin enough (a few nanometers) so that it lets the superconductivity induced by a covering top layer made of aluminum (Jarillo-Herrero et al. 2006) or niobium (Pallecchi et al. 2008) establish through the whole sandwich by the proximity effect. Using those materials, high transparency contacts with almost transparency T close to unity (Javey et al. 2003) can be reproducibly achieved by metal evaporation thus placing connected nanotube junctions in the high coupling limit (Liang et al. 2001). Advances in synthesis methods of nanotubes (in situ grown by CVD), which allow nanotubes to be connected during their growth onto platinum contacts (Cao et al. 2005) are another promising fabrication route. This last method makes clean suspended nanotube junctions, which in the near future could make possible the coupling of nanomechanical vibration modes to superconducting charge transport, as has been recently observed with normal charge carriers, by using non-superconducting metal electrodes as contacts (Lassagne et al. 2009, Steele et al. 2009).

7.3.3 Nanotube Quantum Dot Connected to Superconducting Electrodes

In a CNT weak link, the device considered is a portion of a nanotube with a length L connected by the two superconducting contacts (see Figure 7.8).

At one dimension, a defect along the channel cannot be overcome by the diffusion of the incoming electrons like it occurs in a 2D or 3D system. Therefore, the slightest defect along the nanotube generates back scattering which cancel the transmission of a supercurrent by Andreev bound states. It is then extremely important to realize the injection of Cooper pairs in a perfectly clean system. The nanotube portion length L is usually chosen not to exceed 300 nm in order to limit the occurrence of these structural defects. In such a case, a ballistic transport is reached (i.e., the voltage drop is localized at the contacts, see the top of Figure 7.6). The sub-micron length and nanometer diameter of the nanotube junction makes the intrinsic capacitance C of the junction very small, creating a Coulomb blockade effect (Beenakker and van Houten 1992). A charging energy $e^2/2C$ contributes to their energy and adds to the kinetic energy which leads to the Andreev bound states becoming

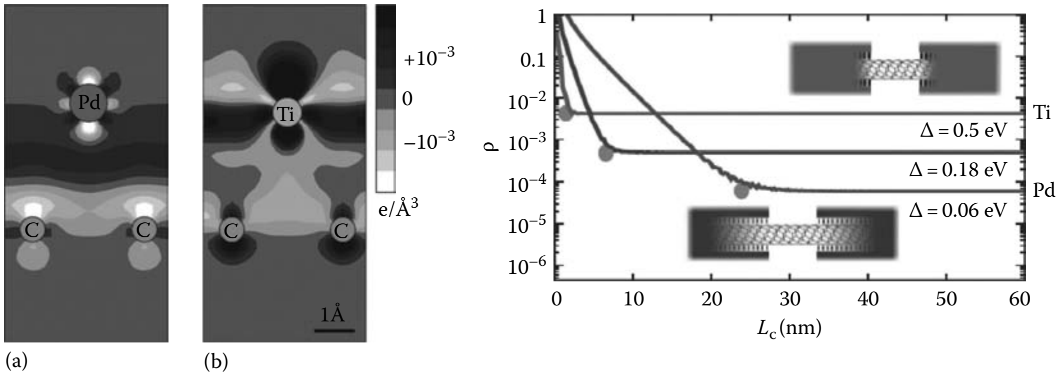

FIGURE 7.7 Left: Cross section showing the charge density redistribution at a metal/carbon nanotube interface for Palladium (a) and Titanium contacts (b). These results are obtained using an ab-initio density functional theory. The titanium contact show a much stronger hybridization with the carbon surface compared to Palladium. Right: Contact reflection coefficient $r = 1 - t$, (where t is contact transmission) as a function of the contact length L_c, the transparency first increases linearly with the contact length then saturates at a constant value independent on the contact length. Their results shows that the optimal metal/carbon interface transparency is obtained for a contact combining low hybridization with a large contact length. (Adapted from Nemec, N. et al., *Phys. Rev. Lett.*, 96, 076802, 2006. With permission.)

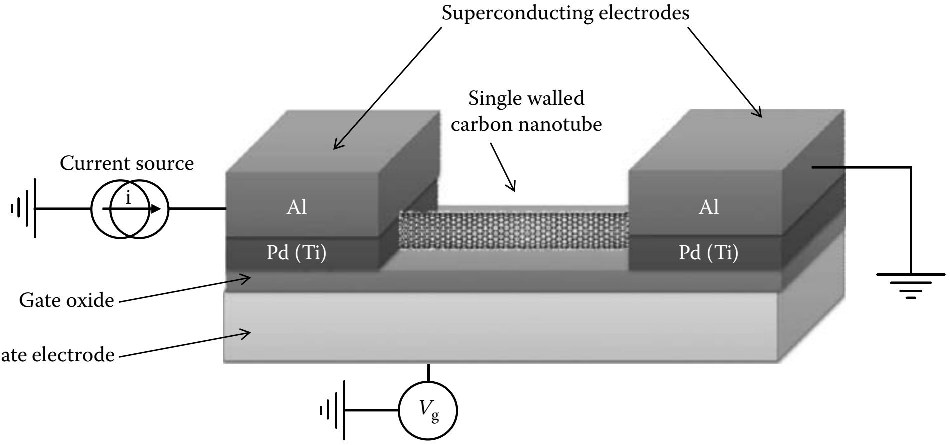

FIGURE 7.8 Schematic diagram of a carbon nanotube superconducting transistor. The contacting electrodes are made of a bilayer of aluminum (for the source of superconductivity) and underlayer of few nanometer thick film either made of palladium or titanium to increase contact transparency. These latter materials are known for their good contact abilities on sp^2 carbon layers (see Figure 7.7). The highly doped silicon substrate is connected to a voltage source V_g and is used as a "backgate" electrode.

$$E_{e,h} = \frac{e^2}{2C}(\pm 1 - n_g)^2 \pm \hbar v_F k, \tag{7.6}$$

where n_g is the gate-induced charge on the nanotube [$n_g = C_g V_g / e$]; electrostatic degeneracy between charge states occurs at zero gate bias.

The wavevector k is quantized by the Andreev boundary condition at the superconducting interfaces in multiples of $1/L$. In the presence of the Coulomb repulsion in the Quantum dot, the Andreev bound states described in Section 7.2.5 split their energy in a sum of the charging energy term and the confinement energy term, which for the electron states give

$$E_e = \frac{e^2}{2C}(1 - n_g)^2 + \frac{\hbar v_F T}{2L}(\phi + \psi_e(n_g, \phi)), \tag{7.7}$$

and for hole states

$$E_h = \frac{e^2}{2C}(1 + n_g)^2 + \frac{\hbar v_F T}{2L}(\phi + \psi_h(n_g, \phi)), \tag{7.8}$$

where $\psi_h(n_g,\phi)$ is a gate-dependent phase. In most experiments, the charging energy e^2/C and quantization energies are of the same order of magnitude (typically 1–10 meV). The eigenstates are then mixed states of charge and Andreev bound states. Close to a degeneracy point (n_g=1/2), the CNT Andreev bound states become resonant with the Fermi level of the leads, thus maximizing the supercurrent flow. As this energy separation can be tuned by applying a gate voltage V_g to the nanotube, we are able to tune the maximum supercurrent (referred to as "switching current") with the gate voltage.

The gate dependence of this last quantity is then periodic (see Figure 7.9, left) and directly reflects the conductance variations observed in the normal state (linked to the Fabry–Perot-like oscillations described in the previous paragraph). By varying the electrostatic voltage V_g applied to the nanotube, it is possible to shift the position of these energy levels relative to the energy levels of the electrons in superconducting contacts (Figure 7.4b, right). When levels coincide, the maximum supercurrent that can pass before obtaining a resistive state (called "current jump") reaches its maximum. On the opposite, if Andreev bound states spectral weight decrease due to levels misalignment, the current will then be minimal. It has been shown (Cleuziou 2006) that even in the case of abscence of Andreev bound states, a tiny supercurrent (of the order of tens of pico-amperes) can be maintained through the CNT probably due to higher order cotunneling processes. It has been experimentally shown by Jarillo-Herrero et al. (2006) that the maximum superconducting current can indeed be controlled by the gate voltage (Figure 7.9).

Therefore, such CNT junctions implement superconducting transistors for which the switching current can be tuned with the gate voltage over a large dynamic range (typically up to 2–3 orders of magnitude, depending on the coupling; see Figure 7.9 right). The current–voltage curves show some marked hysteresis especially in the off-state. The origin of this hysteresis (Gang et al. 2009), which is usually attributed to the existence of a capacitance in parallel with the junction for tunnel weak links (Barone and Paterno 1982), is still debated in the case of these carbon weak links in which the capacitance is extremely small. Hysteresis could rather be associated with the Joule dissipation that occurs in the weak link after switching to the normal state as it has been recently observed in metallic junctions (Courtois et al. 2008). Heat dissipation within the tiny volume of the junction could delay the re-trapping of the weak link in the superconducting state once the current bias is decreased. This occurs especially when the nanotube is in the high resistance state (Off State, see Figure 7.9).

The maximum superconducting current that can be carried by a single-walled CNT is given by Equation 7.5 and is equal to $2\Delta/\hbar \sim 30\,\text{nA}$ for the two conducting channels of a metallic

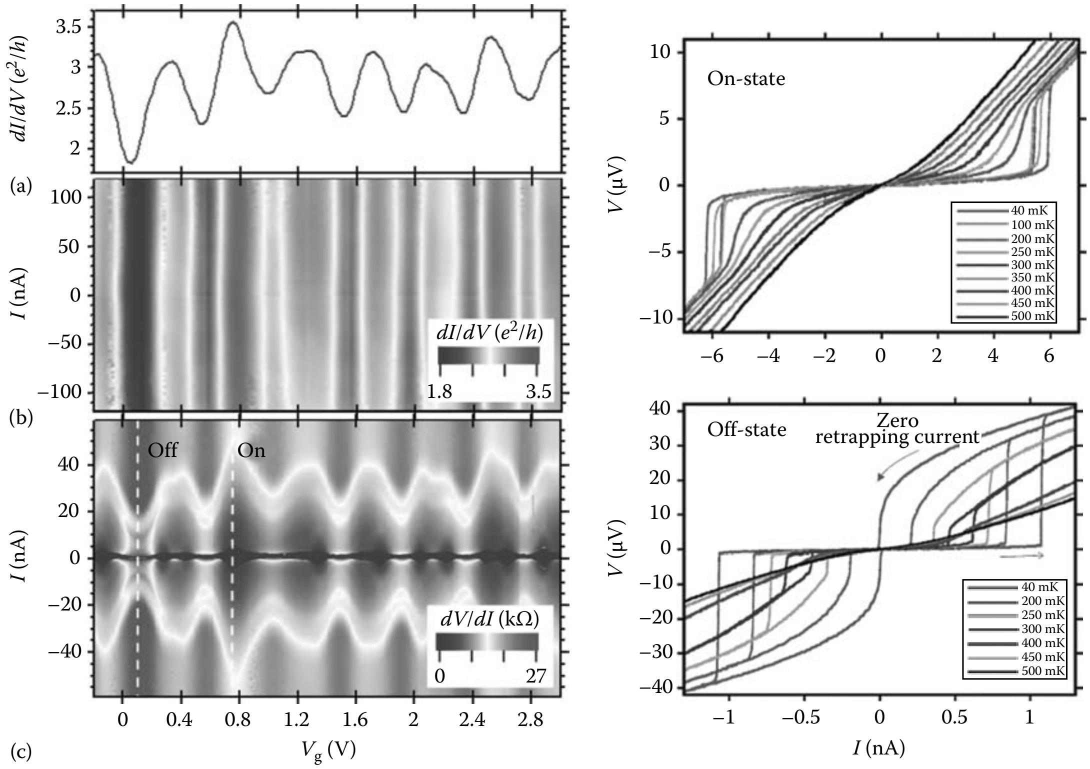

FIGURE 7.9 Electron transport through a transparent carbon nanotube junction as a function of the gate-voltage V_g. Top left: (a) Gate modulation of the nanotube conductivity dI/dV at zero bias in the normal state. (b) Map of the conductivity for current bias I. At resonance, the nanotube conductance is close to the theoretical maximum $4e^2/h$ thus corresponds to the high transparency limit ($t \sim 1$). Gate oscillations are broadly distributed in gate due to the strong hybridization of the nanotube levels with the contacting leads. (c) Differential resistivity dV/dI versus current I and V_g in the superconducting state. Darkest zones around zero bias corresponds to the supercurrent flow which are occurring to the same gate voltages as the high conducting state for non-superconducting electrodes. Right column: Voltage–Current characteristics of a carbon nanotube superconducting transistor measured at low temperature, showing a modulation of the superconducting current for gate voltage corresponding respectively to the on and off state (depicted by the white dotted lines on the resistance map shown in left). The off-state show a strong hysteresis which completely vanished in the on-state.

single-walled CNT connected to Aluminum electrode. In early experiments, measured switching currents were of a few nanoamperes, typically 10%–20% of the expected value. The discrepancy between the predicted and the observed values have been attributed to the sensitivity of the superconducting current to external perturbations, such as micro-wave photons emitted by thermal fluctuations and traveling along the lines. A better control of the electromagnetic environment by using LC and RC filtering on the measurement lines (Jarillo-Herrero et al. 2006) and by implementing locally dissipative elements, such as on-chip resistors close to the nanotube junctions (Jorgensen et al. 2007), has proven to be helpful to stabilize the superconducting current and increase the measured value of the switching current.

7.3.4 High-Frequency Irradiation of Nanotube Weak Links

According to what is known as the AC Josephson Effect, the phase twist across the weak links varies in time at the frequency given by the characteristic frequency of the coupling. Upon irradiation with radio frequency (RF) electromagnetic waves (Barone and Paterno 1982), superconducting weak links are known to develop a series of voltages steps that are so precisely positioned that they are used to define the voltage standards (Hamilton 2000). In the case of the irradiation of a single quantum channel, a peculiar behavior is expected for the RF-irradiated Andreev states (Averin and Bardas 1995).

Experimental measurements (Cleuziou et al. 2007) showed that the voltage steps across the nanotube weak link follow a completely different behavior upon the RF irradiation power depending on if the nanotube is placed either in the "on" or "off" state (see Figure 7.10). This behavior appears to be linked to the existence of the hysteresis found in the direct current (DC) measurements (Figure 7.9) and is discussed in the previous paragraph. They have been associated with the change of dissipation (Liu et al. 2009) that occurs when the nanotube levels are aligned or misaligned with the level of the leads. The RF properties of the superconducting carbon-based junction is expected to be very interesting principally because of their tiny size, which induces an extremely low intrinsic capacitance C, thus leading to extremely high cut-off frequencies $(RC)^{-1}$ that could exceed tens of GHz (Burke 2004). The coupling of these weak links to superconducting resonators will allow for the assessment of their additional RF properties in the near future.

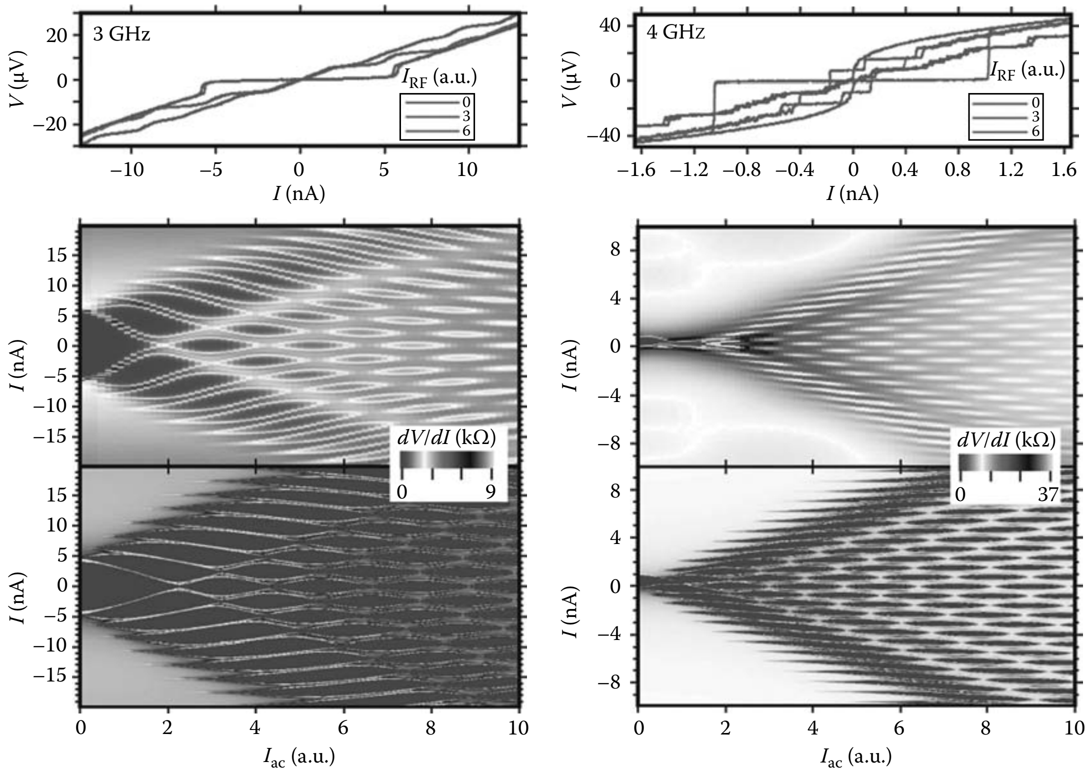

FIGURE 7.10 Carbon nanotube superconducting transistor behavior upon radio-frequency (RF) irradiation in the on and off state (respectively shown in the left and right columns). Top curves: Voltage current characteristics at several RF amplitudes showing the typical constant voltage steps, known as "Shapiro steps." Hysteresis in these steps is seen on the "on" state and vanishes in the "off" state. Middle panels: differential resistance *dV/dI* maps for increasing RF power showing a different behavior in the off and on state. Bottom panels: Numerical simulation of the data of the middle panels using the RSJ (resp. RCSJ) non linear model for the on (resp. off) states.

7.4 Nanotube-Based Superconducting Quantum Interferometers

7.4.1 Quantum Interference with Single Nanotube Weak Links

An interferometer is an apparatus that measures the phase differences between two wave components. Interferometers work by splitting a wave into two components, sending them off along different paths, and then recombining them to record an interference pattern. The pattern is dependent on the phase difference arising from changes in the conditions along their paths. A superconducting interferometer (referred to by its acronym "DC-SQUID" [Clarke and Braginski 2004], which stands for direct current superconducting quantum interferometer device) achieves a kind of "Young slits" experiment with Cooper pairs, the slit being implemented by 2 weak links placed in parallel.

The coherence of the photon beam is provided here by the coherence of the Cooper pair supercurrent. Such a device was first implemented in the 1960s by using oxide-based tunnel weak links (Jaklevic et al. 1964), which rapidly became ubiquitous in the field of superconducting electronics: SQUIDs have been successfully used as ultra-sensitive magnetometers as well as amplifiers of low-level signals and low-impedance sources (Clarke and Braginski 2004).

The main feature of this device is that the maximum superconducting current can be periodically modulated by a tiny magnetic flux threaded by the loop, with a period equal to the flux quantum ϕ_0 $h/2e$ (Figure 7.11, right). We briefly recall in the next paragraph the principle of operation of a DC-SQUID.

The two currents circulating in each weak link interfere with each other and each weak link induces a phase shift $\Delta\phi_i$. according to Equation 7.2.

The total supercurrent is therefore written as the sum of two currents flowing in each branch.

If one assumes a sinusoidal current-phase relation, one obtains

$$I = I_{c1}\sin\left(\Delta\phi_1\right) + I_{c2}\sin\left(\Delta\phi_2\right). \qquad (7.9)$$

Furthermore, an additional controlled phase (analogous to a geometric phase) can be introduced by applying a magnetic flux Φ_{ext} through the loop formed by the two weak link branches. Because of the uniqueness of the superconducting wave function, Φ_{ext} couples the phase shifts induced by each of these junctions in the two branches

$$2\pi\left(\frac{\Phi_{ext}}{\Phi_o}\right) = \Delta\phi_1 - \Delta\phi_2\,(\text{mod}.2\pi). \qquad (7.10)$$

In practice, the SQUID is current-biased and one measures the maximum superconducting current before the onset of a finite voltage

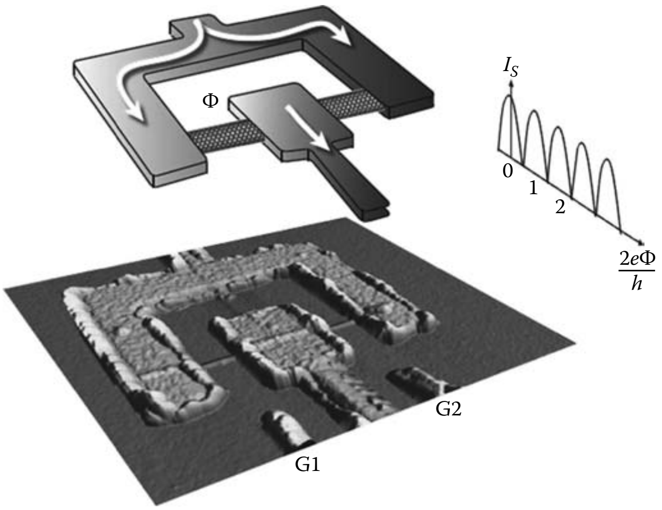

FIGURE 7.11 Schematic diagram of a superconducting quantum interferometer (SQUID) based on carbon nanotube weak links (top) aligned with an atomic force micrograph of a real device. (Adapted from Cleuziou, J.-P. et al., *Nat. Nanotechnol.*, 1, 53, 2006.) Thanks to a peculiar fork geometry design, the weak links in each branch can be made of two portions of the same carbon nanotube generating two independent quantum dots. The phase shift between the two currents flowing in each arm (white arrows) is adjusted by the magnetic flux *F* threaded by the loop. The two electrodes, G1 and G2 allow the application of locally addressed gate voltages which tune independently the transparency of each portion of the nanotube according to the principles of the carbon nanotube junctions shown above.

across the device. This switching current is directly related to the critical current of the SQUID, which can be obtained by computing the maximum value of the solutions of Equations 7.9 and 7.10.

$$I_c = \sqrt{\left(I_{c1} - I_{c2}\right)^2 + 4I_{c1}I_{c2}\cos^2\left(\frac{\pi\Phi_{\text{ext}}}{\Phi_o}\right)}. \tag{7.11}$$

Due to thermal or quantum fluctuations of the phase across the weak links, the measured switching current does not reach that critical value I_c but its magnetic flux dependence still follows the relation in Equation 7.11. The quantum interference, therefore, produces periodic oscillations of the current as a function of external magnetic flux, analogous to the bright fringes observed in an experiment of Young's slit and whose period is equal to the flux quantum $h/2e$.

At the nanometer scale, a single CNT can be used to implement the two weak links that are placed on both arms of the interferometer by using a forked geometry (Figure 7.11).

The two branches implement two nanotube quantum dots; their behavior has been detailed in Section 7.3.3. It is therefore possible to modulate the quantum interference by independently adjusting the intensity of coupling in each quantum dot (see the three limit cases in Figure 7.12). Such electrostatic control of a DC-SQUID is extremely useful, both to optimize its operation and its coupling to local magnetic fields to study the detailed physics of the molecular Josephson junction. More generally, it is an experimental model system for the study of the variation of the phase of charge carriers across a quantum dot. Many physical phenomena such as an equal number of electrons in the nanotube (including those related to interactions between the spins of electrons) have been predicted and may now be studied.

The integration of these nanotubes between superconducting electrodes opens the way for the realization of superconducting circuits operating at radio frequencies, particularly useful to follow the time evolution of magnetization of individual nano-objects. They could afford to explore how quantum states of magnetization of a molecule magnet can couple to those generated in a superconducting circuit (which are the basis of superconducting quantum bits).

7.4.2 Tunability of the Phase Shift across Nanotube Weak Links

For an odd electron number of charge sitting on the nanotube, a CNT weak link can behave as a quantum dot that has a nonzero spin state. It thus behaves as a "spin impurity" seen from the superconducting electrodes. Quantum dots populated with a odd number of electrons are prone to an additional electronic resonance called the Kondo Effect (Kouwenhoven and Glazman 2001). This resonance originates from the tendency of conduction electrons to screen the nonzero spin states of the quantum dot in a similar fashion as it occurs in a 3D metal, which contains a low level of magnetic impurities. This screening current adds to the usual conduction channel of the quantum dot and generates a conduction channel at zero bias, which bypasses the Coulomb blockade regime (Nygard et al. 2000). In the superconducting state, the nonzero spin tunneling induces higher order electron transfers events (Buitelaar et al. 2002) that allow the flipping of both spins of the Cooper pair during the tunneling through the nanotube. This amounts to changing the sign of the Cooper pair wave function. As can be seen in Equation 7.1, a change of sign in this wave function is equivalent to a phase shift of π in the superconducting phase.

Such an effect, in which the minimum energy state of one of the Josephson junctions (the so-called π-junction) is obtained for a phase difference of π instead of 0, has been the focus of intense studies. Reminiscent of ferromagnetic impurities in a Josephson junction, it was predicted some 20 years ago (Glazman and Matveev 1989) that a reverse Josephson current would take place in a junction involving tunneling through a quantum dot populated with an odd number of electrons. For a strong Kondo effect (Figure 7.2), the Josephson coupling is expected to be positive (0-junction) since the localized spin is screened due to the Kondo effect. On the other hand, for a weak Kondo effect, the large on-site interaction only allows the electrons in a Cooper pair to tunnel one-by-one via virtual processes in which the spin ordering of the Cooper pair is reversed, leading to a negative Josephson coupling (π-junction) and hence a reversed supercurrent. Controlling the phase-shift across a nanotube weak link can be achieved by tuning the charge induced on the weak link (see Figure 7.13).

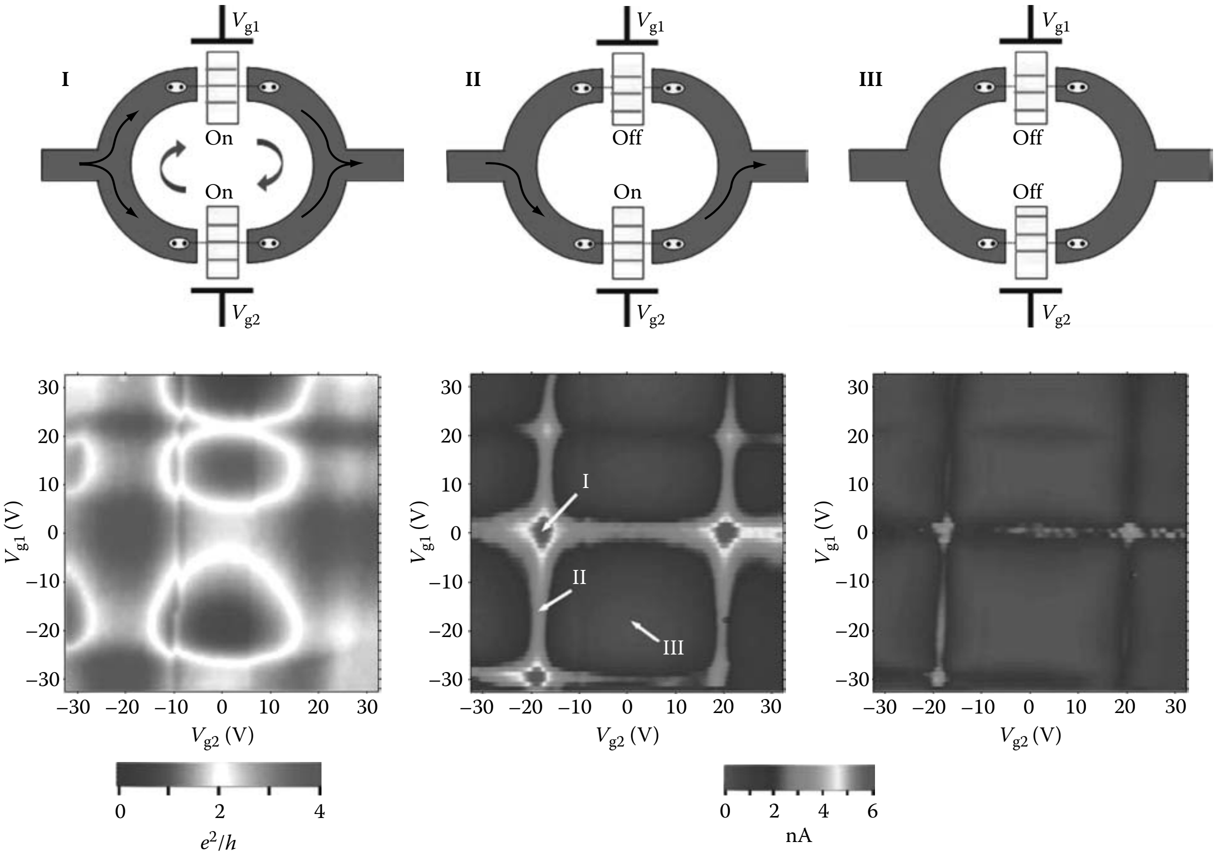

FIGURE 7.12 Top: Schematics of three typical working points of a carbon nanotube SQUID, in case **I**, both junctions have a quantum level adjusted to the Fermi energy of the leads (on-resonance) and supercurrent can flow through the device. In **II** and **III**, one and two junctions are tuned off-resonance, respectively. Bottom left: Modulation of the conductance of the device in the normal state. Center: Modulation of the maximum superconducting current of the same device. The number refers to the three cases described above. (Adapted from Cleuziou, J.-P. et al., *Nat. Nanotechnol.*, 1, 53, 2006. With permission.)

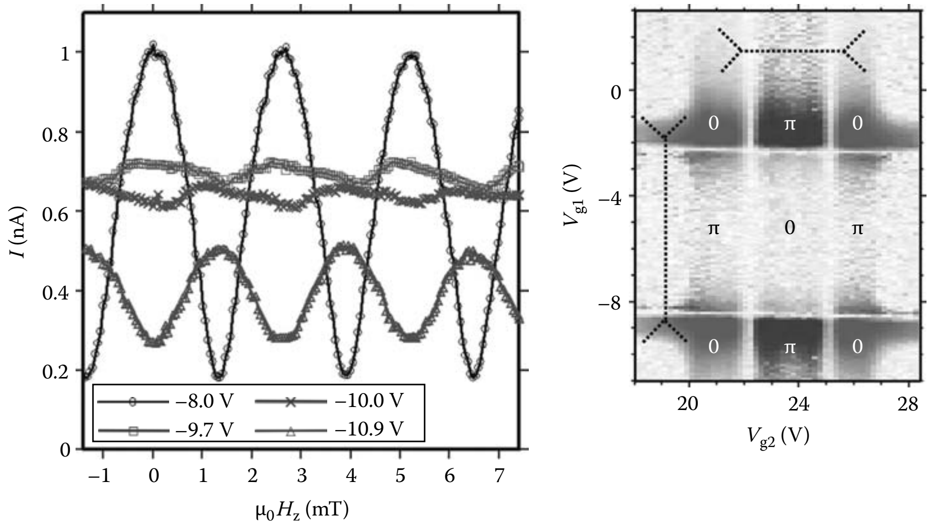

FIGURE 7.13 Gate control of the phase in a carbon nanotube SQUID. Left: magnetic field dependence of the maximum supercurrent in a nanotube SQUID as a function for different lateral gate voltages V_g, the modulation can be tuned from an even (0-junction SQUID) to an odd (π-junction SQUID) curve. The modulation is highly non-sinusoidal between the two states (Right), Color-scale representation of differential resistivity *dV/dI* as a function of the lateral gate voltages V_{g1} and V_{g2} showing the periodic switching between zero and π modulation. The dotted lines corresponds to voltage gate range for which an odd number of electrons is in the quantum dot. Note that when both SQUID junctions has π modulation, the SQUID recovers normal (0) modulation. (Adapted from Cleuziou, J.-P. et al., *Nat. Nanotechnol.*, 1, 53, 2006. With permission.)

7.4.3 Application of Nanotube Weak Links to Magnetometry

SQUID-based magnetometry in the near field regime (i.e., for the detection of fields emitted by source at the distance smaller than the SQUID loop radius) has been exploited for nearly 15 years in the field of molecular magnetism (Wernsdorfer 2001). It is based on the inductive coupling of a nano-magnetic object placed near a constriction of the loop of a SQUID. A local change in flux, for example, induced by the magnetization reversal of a magnetic dipole in the vicinity of the loop (Wernsdorfer et al. 1995) will be detected by a variation in the SQUID switching current. By recording these changes while sweeping the magnetic field in direction and intensity, one can accurately study the physical conditions of magnetization switching of a nanoparticle or a assembly of molecules. Therefore, it is possible to study the details of the dynamics of magnetization reversal on nano-magnets and highlight, in the case of molecular magnets, the existence of a tunnel effect of this reversal (Wernsdorfer et al. 1997).

The current system is limited in sensitivity and the threshold corresponds to the detection of a few thousands of elementary spins. The miniaturization of the most sensitive SQUID (superconducting weak seal, represented here by the CNT) seems to be a promising way of gaining sensitivity.

The strong 1D geometrical form factor of CNT junctions combined with their tiny cross-section open some interesting perspectives concerning the magnetometry at the nanometer scale. Indeed one can show that shrinking the cross-section of a weak link allows for the optimization of the inductive coupling between the loop of the SQUID magnetometer and a magnetic object sitting in close vicinity of the weak link (Bouchiat 2009). This inductive coupling factor is known to be proportional to the inverse of the weak link cross-section radius. Therefore, miniaturization of the weak link footprint is indeed necessary to match the size of a single molecule magnet (see Figure 7.14).

Noise measurements on practical devices (Cleuziou et al. 2006) have shown that the detection threshold in CNT magnetometers should allow measurements in the near field regime of magnetized nano-objects of typically 10 Bohr magnetons, for example, strays emitted by a single molecular magnet (see Figure 7.14). The control and measurement of the magnetization of a single molecular magnet is indeed a promising way to implement spin Qubits, which are readily interfaced to coherent superconducting electronic currents. The fact that nanotube SQUIDs can see their current gated within a large dynamic from 10 nA down to picoampere critical currents (Cleuziou et al. 2006) (see Figure 7.9) is also a promising feature. This could allow the limiting of the back-action of the measuring instrument to the evolving quantum spin, thus helping to preserve long coherence times. More generally, magnetometers based on these ultra-miniaturized weak links could provide new instrumentation set-ups, useful for the building of a spintronics at the molecular scale (Bogani and Wernsdorfer 2008).

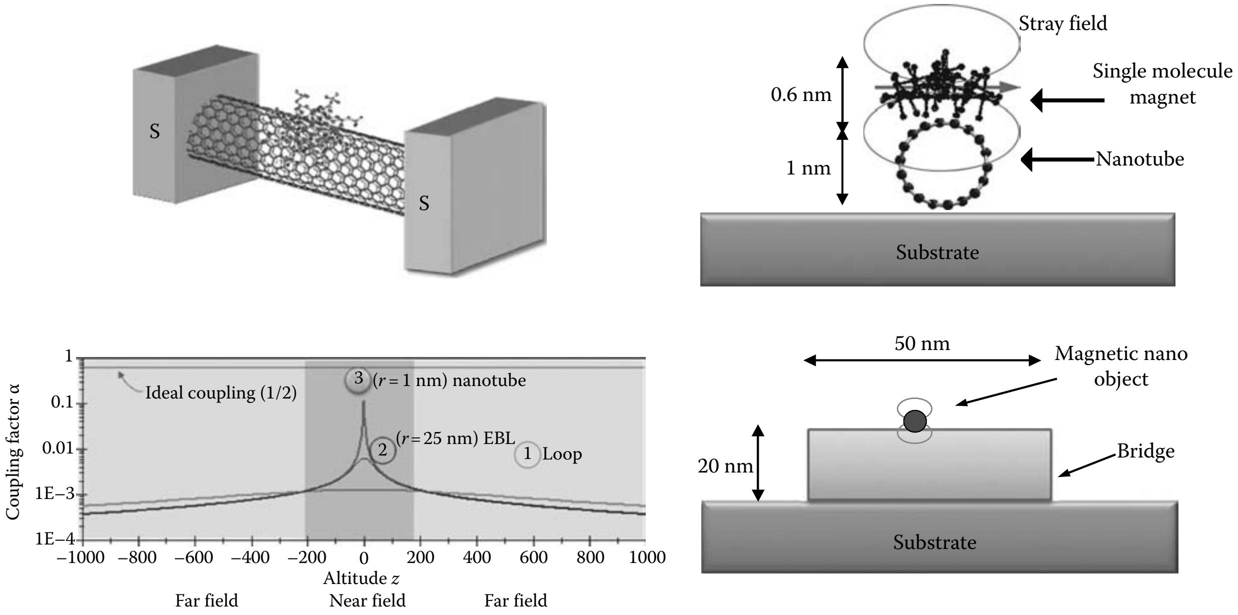

FIGURE 7.14 Optimization of the inductive coupling of single molecule magnet to the nanotube superconducting junction. Top left: artist view of the grafting of a single molecule magnet (here a "manganese-12" molecule) in close contact onto a carbon nanotube junction. Top right, cross section of the same assembly showing the stray field emitted from the magnetic molecule. Bottom right: cross section of a similar situation for the previous generation of weak links made of a nanowire with state-of-the art electron beam lithography. Bottom left: Simulation of the inductive coupling factor α for the two cases shown in right as a function of the distance z between the centers of the two coupled objects: the black curve stand for a lithography-made nanobridge ($r = 25$ nm) while the curve (labelled 3) depicts the situation for a single-walled carbon nanotube ($r = 1$ nm). The nanotube based weak link which diameter matches the molecule size, better grab the stray field and thus increase the inductive coupling by two order of magnitude with respect to the weak link made by top down lithography. The top straight line depicts the theoretical coupling limit ($\alpha = 1/2$) which is almost reached with a carbon nanotube junction.

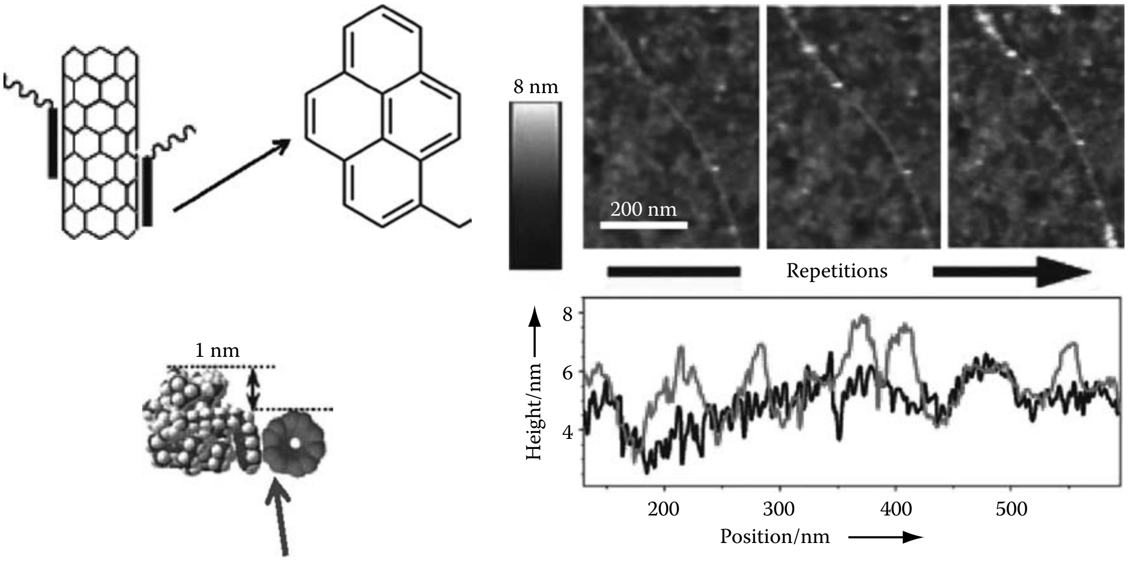

FIGURE 7.15 Chemical functionalization of carbon nanotube weak links for single molecule magnetism applications. Top left: Schematics showing the grafting of carbon nanotubes with organic molecules using pyrene group (inset) that specifically bind onto nanotube side walls by p-p interaction. (After Chen, R. et al., *J. Am. Chem. Soc.*, 123, 3838, 2001.) Right: AFM topographic analysis of molecular magnet/nanotube hybrids. Top right: Height images of the same CNT acquired on repeating the grafting process: left one time, center four times, right ten times. Bottom right: Section profile along the same CNT before (black line) and after (gray line) multiple graftings of 1. Bottom left: Heights of the grafted objects. (Adapted from Bogani, L. et al., *Angew. Chem.*, 48, 746, 2009.)

The full integration of magnetic molecules coupled with the device is still to be made. This requires control of the binding of single magnetic molecules to the nanotube while preserving its good conduction properties. Progress has been made in that direction by using chemically functionalized organic entities (Chen et al. 2001) that can mimic on a few cells, the graphitic honeycomb lattice (pyrene groups, see Figure 7.15), thus through π–π interaction, they provide specific grafting along the nanotube side walls. Pyrene functionalized molecular magnets have been shown (Bogani et al. 2009) to specifically bind in solution in a rather controlled way onto nanotube junctions.

7.4.4 Application of Nanotube Weak Links to Quantum Information

As shown above, CNT-based weak links offer the possibility of realizing the controlled injection of Cooper pairs that travel as

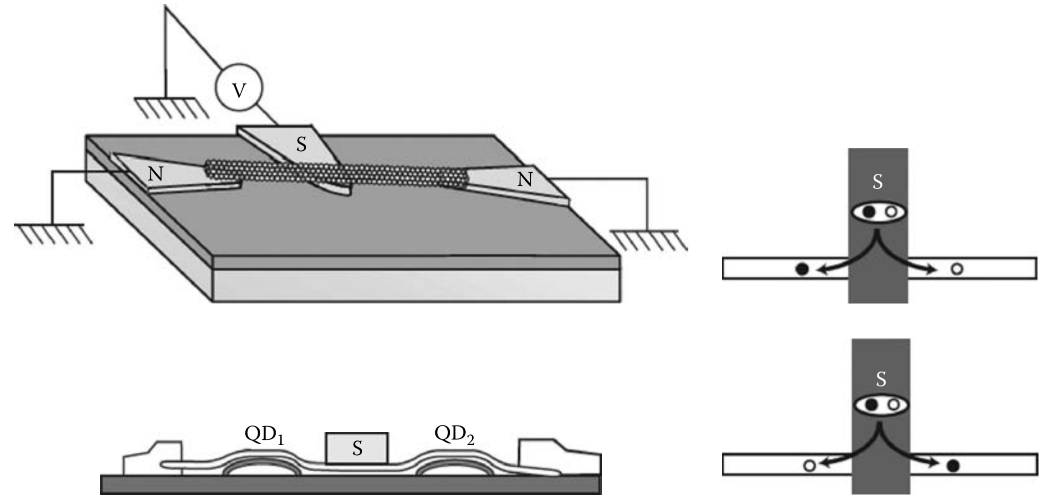

FIGURE 7.16 Top left: Schematics of a carbon nanotube based Cooper pair splitter: the nanotube is coupled in its middle part to a voltage biased superconducting electrode (noted S) which injects Cooper pairs in the nanotube. The nanotube splits the current into two components which are collected by symmetrically positioned normal metal electrodes. Bottom left: Cross section of the same device. The two branches act as nanotube quantum dots noted QD$_1$ and QD$_2$ in a similarly geometry as previously presented nanotube SQUID. Right: schematics showing the two possible paths for the two correlated electrons that forms the Cooper pair which splits in the nanotube and travel in opposite directions. (Adapted from Bouchiat, V. et al., *Nanotechnology*, 14, 77, 2003.)

correlated quasi-particles in a 1D or 0D conductor. Many experiments taking advantage of splitting the quantum correlated charge carriers emitted from a superconducting electrode can be devised. They are mostly inspired from those already realized in quantum optics and propose to realize their fully-integrated solid-state counterparts involving Cooper pairs instead of correlated photon pairs. In Cooper pairs splitting experiment, an additional quantum degree of freedom is also provided by the spin of the electron.

Among interesting basic devices is the nanotube-based Cooper pair beam splitter (Figure 7.16). It consists of a single-walled CNT connected at both ends to normal state electrodes and coupled in its middle part to a superconducting nanowire. Such a device acts as an electronic beam splitter for correlated electrons originating from the superconductor. Note that this geometry is very similar to the nanotube SQUID described in Section 7.4, but with the notable difference of having spatially separated normal-state "collecting electrodes" instead of a single superconducting one found in the SQUID geometry (Figure 7.11).

The Cooper pair splitter device was first discussed on a theoretical level in 2002–2003 (Bena et al. 2002, Recher and Loss 2002, Bouchiat et al. 2003) and its first preliminary experimental realizations just appeared recently using either a CNT (Herrmann et al. 2010) or a III–V semiconducting nanowire (Hofstetter et al. 2009).

In this device, the splitting of Cooper pairs in the superconductor provides two electrons with opposite spins. In the case of a full splitting, they can travel into opposite directions in the nanotube and are individually collected at both ends of the nanotube (Figure 7.16). Thanks to the Andreev process, the Cooper pair splitter generates electrons that are simultaneously entangled in energy and in spin. Correlation and noise measurements of the current emerging from the nanotube should allow the testing of Bell inequalities (Bouchiat et al. 2003) with electrons instead of photons.

7.5 Graphene-Based Superconducting Weak Links

7.5.1 Proximity Effect in Graphene Weak Links

Graphene, a single atomic layer of graphite, exhibits unique electrical and mechanical properties on account of its reduced dimensionality and "relativistic" band structure. The transport properties of the graphene sheets have recently received a lot of attention on both experimental (Geim and Novoselov 2007) and theoretical points of view. On one hand, this is due to its 2D character and on the other hand, to its remarkable band structure combining a linear dispersion relation of electronic wave functions similar to mass-less particles and perfect electron hole symmetry. The Fermi surface consists of two cones touching at one singular point, the so-called Dirac point, where the density of states is zero (see Figure 7.4a). Over the last 5 years, graphene has emerged as a unique platform for testing new electronic properties of condensed matter (Figure 7.17).

By tuning a gate voltage positively or negatively, one can respectively induce a finite density of electrons and holes with a linear energy dispersion curve. Furthermore, this 2D electron gas can be exposed to external physicochemical functionalization and be easily connected with superconducting electrodes. Soon after first electron transport measurements, graphene has been shown to allow Cooper pair transport (Heersche et al. 2007) with a well-defined Josephson effect exhibiting microamperes of critical current (Figure 7.18) and clearly defined (Figure 7.18) multiple Andreev reflections (Du et al. 2008a, Heersche et al. 2007).

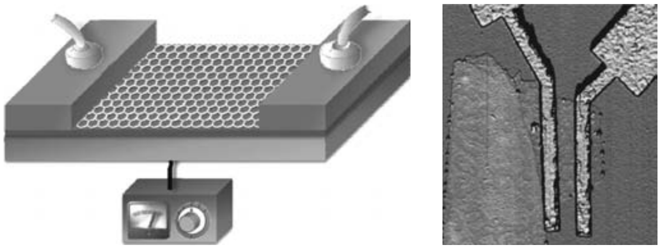

FIGURE 7.17 Left: Schematic diagram of a graphene-based superconducting weak link. The graphene carrier density is controlled by adjusting the backgate voltage on the silicon substrate, which is isolated from the device by a thermal oxide layer. Right: Atomic force micrograph of a first realized graphene superconducting junction. (Adapted from Heersche, H.B. et al., *Nature*, 446, 56, 2007.)

The supercurrent of such a junction can be modulated in intensity with the gate, which tunes the carrier type and density of the graphene sheet (see Figure 7.4b). Depending on the gate voltage applied to the graphene, a supercurrent can be mediated by either electrons or holes (Heersche et al. 2007), which makes graphene the first 2D ambipolar superconducting material. Unlike nanotube weak links, the modulation no more shows any gate periodicy in the modulation of the switching supercurrent because of the absence of lateral quantum confinement. The superconducting current is monotonically increased by increasing the gate and, as it is predicted by the number of conduction channels contributing in parallel to the superconducting current (Equations 7.4 and 7.5), the intensity of the maximum superconducting current is usually thousands of times the one measured for a CNT junction.

New features associated with the existence of Dirac Fermions have been predicted (Beenakker 2008). For example, due to the multiple conical valley band structure, the Andreev reflection, which usually is a retro-reflection (as seen in Figure 7.1), can be induced specularly with a high probability in undoped samples when the Fermi energy lies in the vicinity of the Dirac point (Beenakker 2006). Such effects, as well as the observation of peculiarities induced by the Andreev bound states (Titov and Beenakker 2006), require the realization of the ballistic graphene weak links, a challenging fabrication task since the mean free path of graphene samples usually made are below $0.1\,\mu$m. Promising routes are offered by making suspended graphene obtained by under-etching the supporting substrate. Such a process has been shown to greatly enhance the normal state electronic properties in the non-superconducting state by boosting mean free paths (Du et al. 2008b) and electronic mobilities (Bolotin et al. 2008).

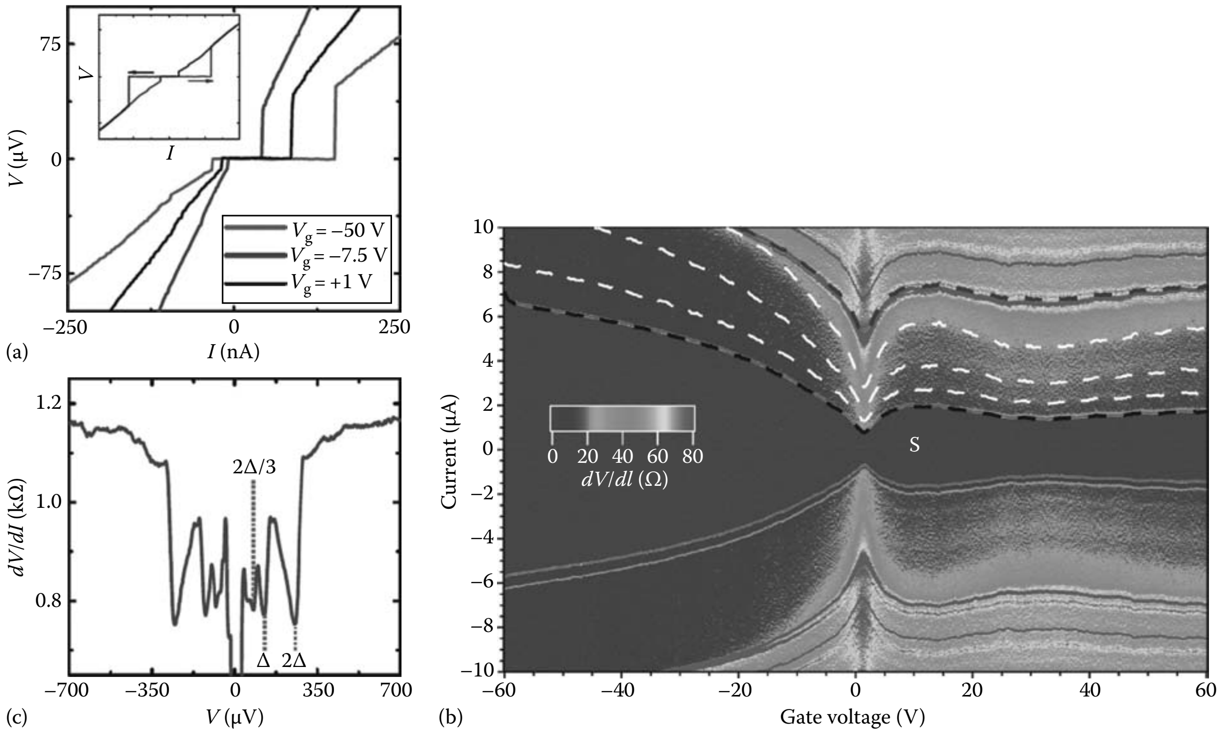

FIGURE 7.18 Characteristics of a superconducting/graphene/superconducting junction. (a) Voltage–current characteristics of a graphene-based superconducting weak link. (From Heersche, H.B. et al., *Nature*, 446, 56, 2007. With permission.) (b) Resistance map of a two-junction graphene device showing the ambipolar gate dependence of the superconducting state (central zone denoted by the letter S). Differential resistance of a device showing dependence of critical current on back gate voltage (dashed line) at 20 mK. Traces of multiple Andreev reflections at constant voltages $2\Delta g/ne$ for $n = 1, 2, 3$ are indicated by the white dashed lines (n) 1 at top, $\Delta g \sim 75\,\mu\text{eV}$. (Adapted from Girit, Ç. et al., *Nano Lett.*, 9, 198, 2009.) (c) Differential resistance as a function of the voltage bias showing the multiple Andreev reflection peaks below the Gap 2D. (From Heersche, H.B. et al., *Nature*, 446, 56, 2007.)

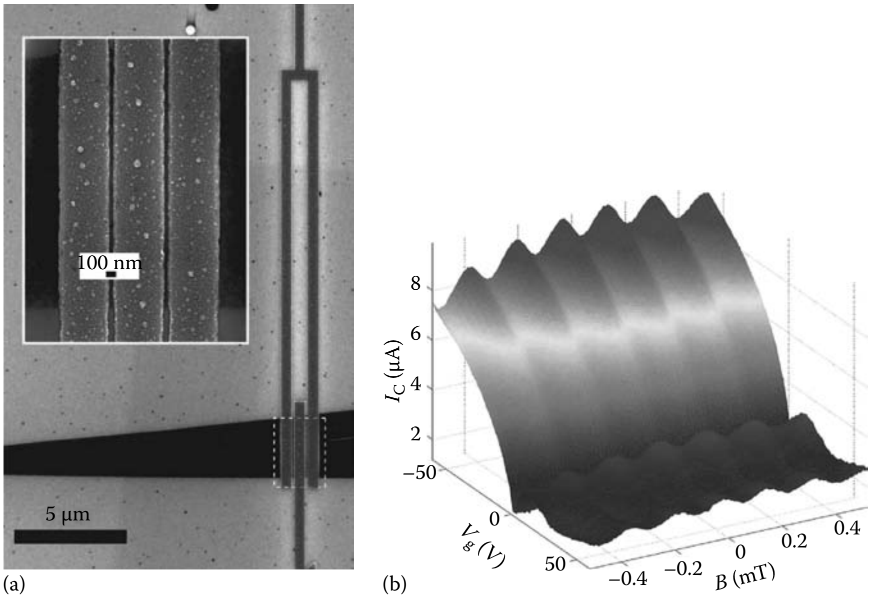

FIGURE 7.19 (a) Scanning electron microscope image of a superconducting quantum interference device with weak links made of portion of the same single layer graphene flake (black, graphene; light gray, silicon oxide substrate; dark gray, Pd/Al electrodes Inset: close-up of sub-100 nm wide graphene Josephson junctions. (b) Maximum switching current as a function of magnetic field B and gate voltage V_g. The sinusoidal oscillations indicate quantum interference between the two graphene Josephson junctions. For a fixed magnetic field, the magnitude of the modulation can be controlled by the back gate. (After Girit, Ç. et al., *Nano Lett.*, 9, 198, 2009.)

In the case of graphene, this is predicted to lead to an unusual bias dependence of the differential conductance of the weak link. In most experiments, superconducting charge carriers are injected by the measuring electrodes. However, it has been predicted (Feigel'man et al. 2010) and experimentally shown (Kessler et al. 2010) that the superconductivity could be induced globally by decorating the graphene with an array of superconducting electrodes. In this latter case, the proximity effect generates a supercurrent, which percolates through the entire graphene sheet.

7.5.2 Graphene Nanotube Superconducting Quantum Interferometers

Graphene-based SQUIDs devices can be obtained in a similar way as has been done for CNT weak links. Indeed, the fabrication and operation of a two junction SQUID formed by a single graphene sheet contacted with aluminum/palladium electrodes in the geometry of a loop (see Figure 7.19) has been demonstrated in 2009 (Girit et al. 2009).

As for the nanotube SQUID, graphene SQUIDs allow the tuning of the quantum interference with an electrostatic gate with a dependence (Figure 7.19b) that follows the transparency given by the graphene weak link (Figure 7.18). It also suggests a new modality for the ultrasensitive magnetometry of nanomagnets chemically or physically attached to the carbon surface.

7.6 Fullerene-Based Superconducting Weak Links

This last category considers the smallest sp^2 carbon weak links that involve fullerene molecules (Smalley 1997). We consider a device made of a fullerene molecule (see Figure 7.4a, right) placed in a transistor geometry (depicted in Figure 7.4b). Interaction of the discrete levels of a molecular entity with the superconducting electrode has triggered a lot a theoretical studies (Choi et al. 2004, Novotny et al. 2005). Unlike all the previous carbon-based weak links discussed above, for which the properties depend not only upon the material properties but also on the electrode geometries

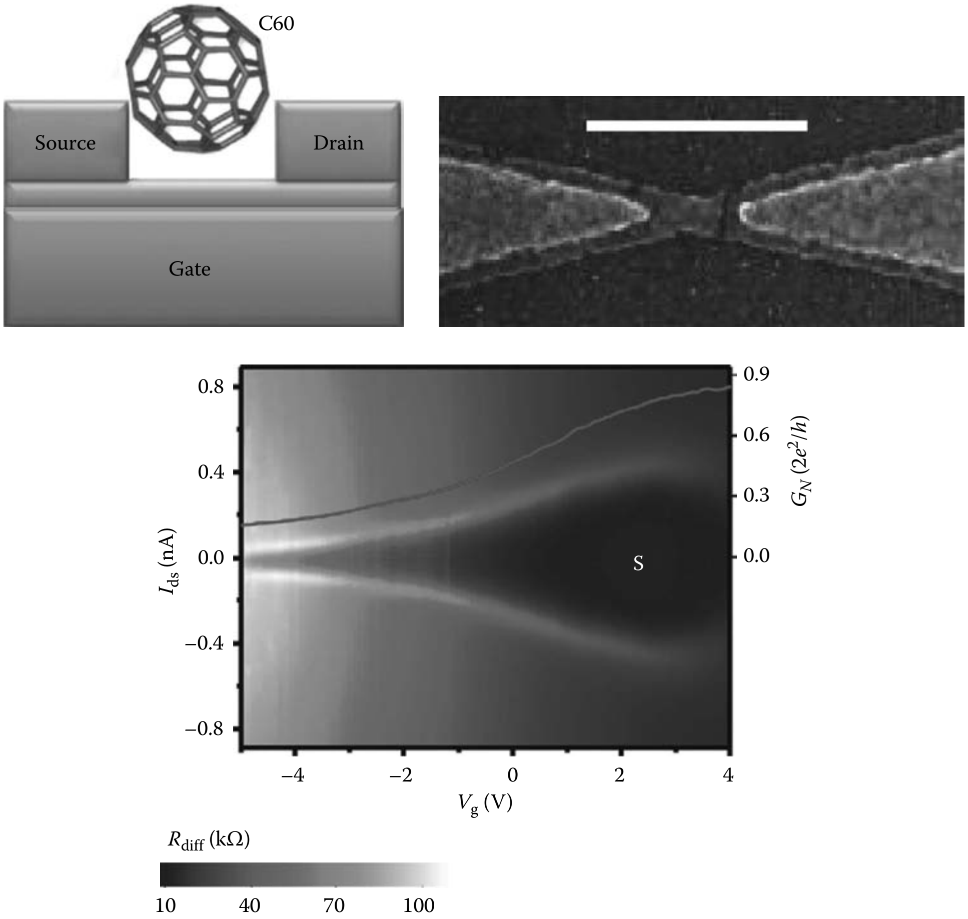

FIGURE 7.20 Fullerene based superconducting weak link. Top left: Schematics of the device C60 connected to superconducting source and drain. The device sits over an aluminum gate electrode whose thin native oxide forms the gate dielectrics. Top right: scanning electron micrograph of the device, showing the nanoscale gap created by electromigration. (scale bar = 300 nm). Bottom: Map of the differential resistance of the device as a function of the gate voltage (*x* axis) and the current bias (*y* axis). The black region corresponds to the superconducting state. The gate dependence of the superconducting current is in agreement with the expected dependence for a proximity coupled quantum dot. (After Winkelmann, C.B. et al., *Nat. Phys*., 5, 876, 2009.)

(for example, in a nanotube weak link, we have seen that the weak link gate modulation is inversely proportional to the length of the weak link, which is defined by the gap between the electrode), fullerene-based weak links involve the entire molecule. Therefore, electron transport through the junction is supposed to exhibit fingerprints (Zazunov et al. 2006) linked to the intrinsic molecule states, such as vibrational modes (Park et al. 2000) or molecular magnetism (Lee et al. 2008). Indeed a set of experiments (see Figure 7.5, right) involving magnetically active molecules such as endofullerenes (Grose et al. 2008, Kasumov et al. 2005), which consist of additional atoms inserted in the carbon cage formed by the fullerene shell, are a sensitive probe of molecular magnetism.

7.6.1 Molecular Transistors Made by Electromigration

Recent progress in the field of molecular electronics now allows for the direct integration of molecular objects into electrical on-chip circuits. It has been shown that by inserting molecules into a nanometer gap by sectioning a metallic wire, it is possible to create a molecular transistor which conducting states are linked to molecular orbitals available for conduction (Park et al. 2000). This possibility was provided by the electromigration process (Park et al. 1999), which allows the generation of nanogaps in metallic wires with good yield and an unprecedented reproducibility.

7.6.2 Implementation of a C60 Superconducting Transistor

Such electromigration processes have been recently adapted to the fabrication of superconducting transistors involving a single C60 molecule inserted between gold/aluminum electrodes that can be measured at milliKelvin temperatures (Winkelmann et al. 2009). Figure 7.20 shows the gate control of the superconducting state achieved in this molecular transistor, in the case of low dissipation in the environment of the linking conductor. The observed switching current is significantly lower than the critical current and has been shown to follow the law

$$\frac{I_S}{I_0}=\left(1-\sqrt{1-\frac{G_N h}{2e^2}}\right)^{3/2}.$$

The results of a C60 molecular transistor (Winkelmann et al. 2009) show good agreement between the measured critical current as a function of gate voltage and the prediction for $I_c(V_g)$ when taking into account the phase oscillations across the weak link.

7.7 Concluding Remarks

A new class of molecular weak links has now become available for research, in which superconductors can be coupled with reliable transparent interfaces to nanostructures composed of a crystalline arrangement of sp^2-hybridized carbon. The main original feature offered by this new type of device is the tunability of the superconducting transport being possible through gate control of the weak link chemical potential offered by quantum confinement. The integration control over these nanostructures, which has been made possible by advances in nanofabrication, allows for the fabrication of hybrid devices that mix bottom-up and top-down approaches. They provide building blocks that pave the way for the realization of more complex devices that should be useful for quantum information and nanomagnetism.

Acknowledgment

The author warmly acknowledges Wolfgang Wernsdorfer for stimulating discussions. The author would like to thank Sathyanarayanamoorthy Sridharan and Klaus Sattler for their kind help.

References

Averin, D. and Bardas, A. 1995. ac Josephson effect in a single quantum channel. *Physical Review Letters*, 75: 1831.

Barone, A. and Paterno, G. 1982. *Physics and Applications of the Josephson Effect*. New York: John Wiley & Sons.

Beenakker, C. W. J. 2006. Specular Andreev reflection in graphene. *Physical Review Letters*, 97: 067007.

Beenakker, C. W. J. 2008. Colloquium: Andreev reflection and Klein tunneling in graphene. *Reviews of Modern Physics*, 80: 1337.

Beenakker, C. W. J. and van Houten, H. 1992. *Single-Electron Tunneling and Mesoscopic Devices* (eds. Koch, H. and Lübbig, H.) http://xxx.lanl.gov/abs/condmat/0111505l. Berlin, Germany: Springer.

Bena, C., Vishveshwara, S., Balents, L., and Fisher, M. P. A. 2002. Quantum entanglement in carbon nanotubes. *Physical Review Letters*, 89: 037901.

Bogani, L. and Wernsdorfer, W. 2008. Molecular spintronics using single-molecule magnets. *Nature Materials*, 7: 179.

Bogani, L., Danieli, C., Biavardi, E. et al. 2009. Single-molecule-magnet carbon-nanotube hybrids. *Angewandte Chemie*, 48: 746–750.

Bolotin, K. I., Sikes, K. J., Jiang, Z. et al. 2008. Ultrahigh electron mobility in suspended graphene. *Solid State Communications*, 146: 351.

Bouchiat, V. 2009. Detection of magnetic moments using a nano-SQUID: Limits of resolution and sensitivity in near-field SQUID magnetometry. *Superconductor Science and Technology*, 22: 064002.

Bouchiat, V., Chtchelkatchev, N., Feinberg, D. et al. 2003. Single-walled carbon nanotube–superconductor entangler: Noise correlations and Einstein–Podolsky–Rosen states. *Nanotechnology*, 14: 77–85.

Buitelaar, M. R., Nussbaumer, T., and Schonenberger, C. 2002. Quantum dot in the Kondo regime coupled to superconductors. *Physical Reviews Letters*, 89: 256801.

Burke, P. J. 2004. Ac performance of nanoelectronics: Towards a ballistic THz nanotube transistor. *Solid State Electronics*, 40: 1981.

Cao, J., Wang, Q., and Dai, H. 2005. Electron transport in very clean, as-grown suspended carbon nanotubes. *Nature Materials*, 4: 745–749.

Charlier, J.-C., Blase, X., and Roche, S. 2007. Electronic and transport properties of nanotubes. *Reviews of Modern Physics*, 79: 677.

Chen, R., Zhang, Y., Wang, D., and Dai, H. 2001. Non-covalent sidewall functionalization of single-walled carbon nanotubes for protein immobilization. *Journal of the American Chemical Society*, 123: 3838–3839.

Choi, M. S., Lee, M., Kang, K., and Belzig, W. 2004. Kondo effect and Josephson current through a quantum dot between two superconductors. *Physical Review B*, 70: 020502.

Clarke, J. and Braginski, A. I. 2004. *The SQUID Handbook*. Weinheim, Germany: Wiley-VCH.

Cleuziou, J.-P., Wernsdorfer, W., Bouchiat, V., Ondarcuhu, T., and Monthioux, M. 2006. Carbon nanotube superconducting quantum interference device. *Nature Nanotechnology*, 1: 53.

Cleuziou, J. P., Wernsdorfer, W., Andergassen, S. et al. 2007. Gate-tuned high frequency response of carbon nanotube Josephson junctions. *Physical Review Letters*, 99: 117001.

Courtois, H., Meschke, M., Peltonen, J. T., and Pekola, J. P. 2008. Origin of hysteresis in a proximity Josephson junction. *Physical Review Letters*, 101: 067002.

De Gennes, P. G. and Guyon, E. 1963. Superconductivity in “normal” metals. *Physics Letters*, 3: 168–169.

Du, X., Skachko, I., and Andrei, E. Y. 2008a. Josephson current and multiple Andreev reflections in graphene SNS junctions. *Physical Review B*, 77: 184507.

Du, X., Skachko, I., Barker, A., and Andrei, E. Y. 2008b. Approaching ballistic transport in suspended graphene. *Nature Nanotechnology*, 3: 491–495.

Feigel’man, M. V., Skvortsov, M. A., and Tikhonov, K. S. 2008. Proximity-induced superconductivity in graphene. *JETP Letters*, 88: 862.

Gang, L., Yong, Z., and Chun Ning, L. 2009. Gate-tunable dissipation and “superconductor-insulator” transition in carbon nanotube Josephson junctions. *Physical Review Letters*, 102: 016803.

Geim, A. K. and Novoselov, K. S. 2007. The rise of graphene. *Nature Materials*, 6: 183.

Girit, Ç., Bouchiat, V., Naaman, O. et al. 2009. Tunable graphene superconducting quantum interference device. *Nano Letters*, 9: 198–199.

Glazman, L. I. and Matveev, K. A. 1989. Resonant Josephson current through Kondo impurities in a tunnel barrier. *JETP Letters*, 49: 659.

Goffman, M. F., Cron, R., Levy Yeyati, A. et al. 2000. Supercurrent in atomic point contacts and Andreev states. *Physical Review Letters*, 85: 170.

Golubov, A. A., Kupriyanov, M. Y., and Il’ichev, E. 2004. The current-phase relation in Josephson junctions. *Reviews of Modern Physics*, 76: 411–469.

Grose, J. E., Tam, E. S., Timm, C. et al. 2008. Tunnelling spectra of individual magnetic endofullerene molecules. *Nature Materials*, 7: 884.

Hamilton, C. A. 2000. Josephson voltage standards. *Review of Scientific Instruments*, 71: 3611.

Heersche, H. B., Jarillo-Herrero, P., Oostinga, J. B., Vandersypen, L. M., and Morpurgo, A. F. 2007. Graphene JJ. *Nature*, 446: 56.

Herrmann, L. G., Portier, F., Roche, P. et al. 2010. Carbon nanotubes as cooper pair beam splitters. *Physical Review Letters*, 104: 026801.

Hofstetter, L., Csonka, S., Nygard, J., and Schonenberger, C. 2009. Cooper pair splitter realized in a two-quantum-dot Y-junction. *Nature*, 461: 960–963.

Imry, Y. and Landauer, R. 1999. Conductance viewed as transmission. *Reviews of Modern Physics*, 71: S306.

Jaklevic, R. C., Lambe, J., Silver, A. H., and Mercereau, J. E. 1964. Quantum interference effects in Josephson tunneling. *Physical Review Letters*, 12: 159–160.

Jarillo-Herrero, P., van Dam, J. A., and Kouwenhoven, L. P. 2006. Quantum supercurrent transistors in carbon nanotubes. *Nature*, 439: 953–956.

Javey, A., Guo, J., Wang, Q. et al. 2003. Ballistic carbon nanotube field-effect transistors. *Nature*, 424: 654–657.

Jørgensen, H. I., Grove-Rasmussen, K., Novotny, T., Flensberg, K., and Lindelof, P. E. 2006. Electron transport in single-wall carbon nanotube weak links in the Fabry-Perot regime. *Physical Review Letters*, 96: 207003.

Jorgensen, H. I., Novotny, T., Grove-Rasmussen, K., Flensberg, K., and Lindelof, P. E. 2007. Critical current 0-pi transition in designed Josephson quantum dot junctions. *Nano Letters*, 7: 2441–2445.

Josephson, B. D. 1962. Possible new effects in superconductive tunnelling. *Physics Letters*, 1: 251–253.

Kasumov, A. Y., Deblock, R., Kociak, M. et al. 1999. Supercurrents through single-walled carbon nanotubes. *Science*, 284: 1508.

Kasumov, A. Y., Tsukagoshi, K., Kawamura, M. et al. 2005. Proximity effect in a superconductor-metallofullerene-superconductor molecular junction. *Physical Review B*, 72: 033414.

Kessler, B. M., Girit, C. O., Zettl, A., and Bouchiat, V. 2010. Tunable superconducting phase transition in metal-decorated graphene sheets. *Physical Review Letters*, 104: 047001.

Klapwijk, T. M., Blonder, G. E., and Tinkham, M. 1982. Explanation of subharmonic energy gap structure in superconducting contacts. *Physica B*, 109: 1657.

Kouwenhoven, L. and Glazman, L. 2001. Revival of the Kondo Effect. *Physics World*, 14: 33.

Lassagne, B., Tarakanov, Y., Kinaret, J., Garcia-Sanchez, D., and Bachtold, A. 2009. Coupling mechanics to charge transport in carbon nanotube mechanical resonators. *Science*, 325: 1107–1110.

Lee, M., Jonckheere, T., and Martin, T. 2008. Josephson effect through a magnetic metallofullerene molecule. *Physical Review Letters*, 101: 146804.

Liang, W., Bockrath, M., Bozovic, D. et al. 2001. Fabry - Perot interference in a nanotube electron waveguide. *Nature*, 411: 665.

Likharev, K. K. 1979. Superconducting weak links. *Reviews of Modern Physics*, 51: 101.

Liu, G., Zhang, Y., and Lau, C. N. 2009. Gate-tunable dissipation and "Superconductor-insulator" transition in carbon nanotube Josephson junctions. *Physical Review Letters*, 102: 016803.

Morpurgo, A. F., Kong, J., Marcus, C. M., and Dai, H. 1999. Gate-controlled superconducting proximity effect in carbon nanotubes. *Science*, 286: 263–265.

Nemec, N., Tomanek, D., and Cuniberti, G. 2006. Contact dependence of carrier injection in carbon nanotubes: An Ab initio study. *Physical Review Letters*, 96: 076802.

Novotny, T., Rossini, A., and Flensberg, K. 2005. Josephson current through a molecular transistor in a dissipative environment. *Physical Review B*, 72: 224502.

Nygard, J., Cobden, D. H., and Lindelof, P. E. 2000. Kondo physics in carbon nanotubes. *Nature*, 408: 342–346.

Pallecchi, E., Gaaß, M., Ryndyk, D., and Strunk, C. 2008. Carbon nanotube Josephson junctions with Nb contacts. *Applied Physics Letters*, 93: 072501.

Park, H., Lim, A. K. L., Alivisatos, A. P., Park, J., and McEuen, P. L. 1999. Fabrication of metallic electrodes with nanometer separation by electromigration. *Applied Physics Letters*, 75: 301.

Park, H., Park, J., Lim, A. K. L. et al. 2000. Nanomechanical oscillations in a single-C60 transistor. *Nature*, 407: 57–60.

Recher, P. and Loss, D. 2002. Superconductor coupled to two Luttinger liquids as an entangler for electron spins. *Physical Review B*, 65: 165327.

Scheer, E., Joyez, P., Esteve, D., Urbina, C., and Devoret, M. H. 1997. Conduction channel transmissions of atomic-size aluminum contacts. *Physical Review Letters*, 78: 3535.

Smalley, R. E. 1997. Discovering the fullerenes. *Reviews of Modern Physics*, 69: 723.

Steele, G. A., Huttel, A. K., Witkamp, B. et al. 2009. Strong coupling between single-electron tunneling and nanomechanical motion. *Science*, 325: 1103–1107.

Titov, M. and Beenakker, C. W. J. 2006. Josephson effect in ballistic graphene. *Physical Review B*, 74: 041401.

Wernsdorfer, W. 2001. Classical and quantum magnetization reversal studies in nanometer-sized particles and clusters. *Advances in Chemical Physics*, 188: 99–190.

Wernsdorfer, W., Hasselbach, K., Mailly, D. et al. 1995. Seminal microSQUID particle WW. *Journal of Magnetism and Magnetic Materials*, 145: 33.

Wernsdorfer, W., Bonet Orozco, E., Hasselbach, K. et al. 1997. Macroscopic quantum tunneling of magnetization of single ferrimagnetic nanoparticles of Barium Ferrite. *Physical Review Letters*, 79: 4014–4017.

Winkelmann, C. B., Roch, N., Wernsdorfer, W., Bouchiat, V., and Balestro, F. 2009. Superconductivity in a single-C60 transistor. *Nature Physics*, 5: 876–879.

Zazunov, A., Fierberg, D., and Martin, T. 2006. Phonon squeezing in a superconducting molecular transistor. *Physical Review Letters*, 97: 196801.

8

Micromagnetic Modeling of Nanoscale Spin Valves

Bruno Azzerboni
University of Messina

Giancarlo Consolo
University of Messina

Giovanni Finocchio
University of Messina

8.1 Introduction **8**-1
8.2 Background........ **8**-3
Exchange Interactions • Anisotropy • Magnetostatic Interactions • External Field • Energy Balance and Free Energy Functional Formulation • Equation of Motion • Equation of Motion Including Spin-Transfer Torque Effects
8.3 Computational Micromagnetics of Nanoscale Spin-Valves: State of the Art **8**-7
Numerical Solution of LLGS Equation Based on FD Methods • Spin-Transfer Torque and Giant Magnetoresistance: Dual Phenomena • Nanoscale Devices: Pillar vs. Point-Contact
8.4 Conclusions and Future Perspective........ **8**-15
References........ **8**-16

8.1 Introduction

Magnetism is an open field in which physicists, electrical engineers, material scientists, mathematicians, chemists, metallurgists, and others practice together.

Today information technologies ranging from personal computers to mainframes use magnetic materials to store information on tapes, floppy diskettes, and hard disks. Our seemingly insatiable appetite for more computer memory will probably be met by a variety of magnetic recording technologies based on nanocrystalline thin-film media and magneto-optic materials. Personal computers and many of our consumer and industrial electronics components are now powered largely by lightweight switch-mode power supplies using new magnetic materials technology that was unavailable 20 years ago. Magnetic materials touch many other aspects of our lives. Each automobile contains dozens of motors, actuators, sensors, inductors, and other electromagnetic and magneto-mechanical components using hard (permanent) as well as soft magnetic materials. Electric power generation, transformation, and distribution systems rely on hundreds of millions of transformers and generators that use the principles of electromagnetic fields as well. Finally, magnetism is present in telecommunication systems, such as mobile phones, electronic article surveillance, asset protection, and access control because of the presence of oscillating circuits and resonators working at microwave frequencies.

Magnetism, along with electricity, belongs to a larger phenomenon, electromagnetism, by which we describe the force generated by the passage of an electric current through matter. When two electric charges are at rest, it appears to the observer that the force between them is merely electric. If the charges are in motion, however, and in this instance motion or rest is referred to in relation to the observer, then it appears as though a different force, known as magnetism, exists between them. In fact, the difference between magnetism and electricity is purely artificial. Both are manifestations of a single fundamental force, with magnetism simply being an abstraction that people use for the changes in electromagnetic force created by the motion of electric charges.

To gain a deeper understanding of magnetism and the materials classification, it is necessary to first introduce some basic concepts. First, within the area of research of magnetism, one can identify at least four different scales that give correspondingly four different levels of investigation, as shown in Figure 8.1.

The top level, magnetic hysteresis models, treats the phenomenology of technical magnetization curves. Such an approach is encountered whenever connections between hysteresis and domain phenomena can be established. There are several models to treat the magnetization phenomena (curves) at this level. The first scalar model of hysteresis was proposed by Preisach (1929). From 1929, lots of efforts have been focused on the physical interpretation of its parameters as well as the inclusion of aftereffects, accommodation, and magnetostriction phenomena (Bertotti 1998).

At the bottom level, we find the atomic description, which deals with the spin structure of magnetically ordered material and the arrangements of spins on the crystal lattice sites. It also describes the origin, interactions, mutual arrangement, and statistical thermodynamics of elementary magnetic moments by means of quantum mechanic theories.

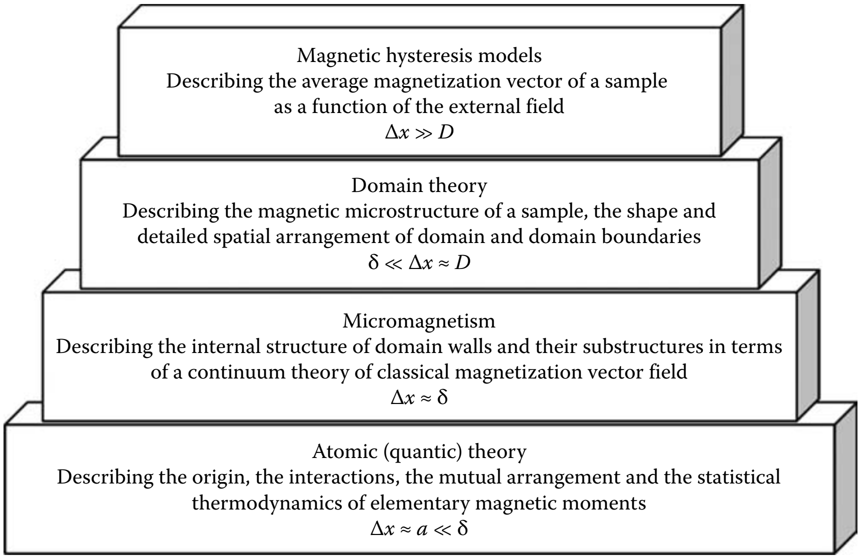
Magnetic hysteresis models
Describing the average magnetization vector of a sample as a function of the external field
$\Delta x \gg D$
Domain theory
Describing the magnetic microstructure of a sample, the shape and detailed spatial arrangement of domain and domain boundaries
$\delta \ll \Delta x \approx D$
Micromagnetism
Describing the internal structure of domain walls and their substructures in terms of a continuum theory of classical magnetization vector field
$\Delta x \approx \delta$
Atomic (quantic) theory
Describing the origin, the interactions, the mutual arrangement and the statistical thermodynamics of elementary magnetic moments
$\Delta x \approx a \ll \delta$

FIGURE 8.1 The hierarchy of descriptive levels of magnetically ordered materials. The values in parenthesis indicate the sample dimensions for which the different theories are applicable. Δx is the characteristic scale for each theory, D, δ, and a are the domain size, the wall width, and the interatomic distance, respectively.

There exist two models working at the intermediate mesoscopic scale: domain theory and micromagnetism. The former combines discrete, uniformly magnetized domains, domain walls, and their microstructure. The latter is the continuum theory of magnetic moments, underlying the description of the magnetic microstructure. The domain theory is located at a scale larger than the domain wall width, whereas the micromagnetic scale is comparable with that scale (see Figure 8.2).

The starting point of the domain theory is to suppose a given domain's distribution in a specimen according to both experimental observations and energy balance. The main assumption is that the wall that divides the adjacent domains has a zero width. In other words, this is a model where the magnetization vector presents discontinuities when passing from a domain to the adjacent one. As cited briefly above, this theory is valid under the assumption that the domain wall width is negligible with respect to the other dimensions involved in the problem. Without going into detail, it has to be noted that the domain theory offers the possibility of describing a lot of real systems. In spite of this, its limitations lead to the need for building a more accurate and general theory.

The micromagnetic theory was developed first by Landau and Lifshitz (1935) on a variational principle: it searches for magnetization distributions with the total smallest energy. This variational principle leads to a set of differential equations, the micromagnetic equations, which were initially developed by the same authors for one-dimensional problems. Stimulated again by experimental work and its analysis, W.F. Brown extended the equations to three dimensions (Brown 1940, 1963).

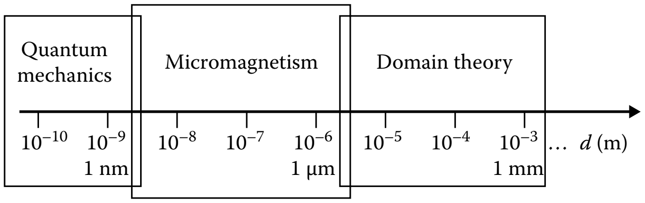
Quantum mechanics
Micromagnetism
Domain theory
10^{-10}
10^{-9}
1 nm
10^{-8}
10^{-7}
10^{-6}
1 μm
10^{-5}
10^{-4}
10^{-3}
1 mm
… d (m)

FIGURE 8.2 Characteristic length scales for mathematical models of magnetism.

The micromagnetic equations are complicated nonlinear and nonlocal equations; they are therefore difficult to solve analytically, except in cases in which a linearization is possible. However, a number of problems in research need micromagnetic methods for their adequate treatment, such as the following:

1. The analysis of the behavior of small magnetic particles that are, at the same time, too small to accommodate a regular domain structure and too large to be described as uniformly magnetized
2. The determination of the spatial distribution of magnetization inside a domain wall
3. The calculation of the magnetic stability limits
4. The analysis of rapid magnetization dynamics occurring in nanoscale devices

To treat such problems, works on numerical solutions of the micromagnetic equations are increasingly pursued. It appears utopian, however, to apply micromagnetic methods to large-scale domain structures because the gap between the size of samples and the size of computational grids is simply too large and unaffordable computational times and memory allocation would be consequently required. According to these discussions, micromagnetic and domain theories should not be considered as mutually exclusive at all, but as complementary tools.

In this chapter, we will address our discussion on the usage of micromagnetic frameworks to describe the behavior of a class of novel devices that are currently investigated for their attracting potential applications (mainly as magnetic memories and microwave oscillators).

In 1996, in fact, almost contemporarily, Slonczewski (1996) and Berger (1996) predicted that a direct electric current passing through a magnetic material, adequately thick, becomes spin-polarized along the direction of its local magnetization vector. The current moving outward such a polarizer is thus referred to as "spin-polarized current." To some extent, it is possible to consider this phenomenon in analogy to what occurs in optical systems where a unpolarized light traverses a slab of polarizing material and causes the outgoing electromagnetic waves to oscillate along the direction of the transmission axis of the polarizer. If a spin-polarized current presents a sufficiently large amplitude and now traverses a second magnetic layer, it can transfer its spin-angular momentum and destabilize the static equilibrium orientation of the magnetization of this latter layer. Such phenomenon, which is correspondingly called "spin-transfer torque effect," has been indeed applied, from the very beginning, in multilayer systems called "spin-valve" like the one shown in Figure 8.3.

A classical spin-valve system consists of: a thicker magnetic layer, called fixed or pinned layer (PL), which acts as the spin-polarizer; a thinner magnetic layer, called free layer (FL), which is the one generally subjected to the action of the spin-transfer torque; and a nonmagnetic spacer, used to decouple the exchange interactions between PL and FL. Depending on the electrical conductivity of the intermediate layer, it is possible to build-up either "classical" spin-valve (CSV) devices, if the spacer is a metal, or nonclassical magnetic tunnel-junctions (MTJ), if the spacer is an insulator. The mechanism of the torque and the resulting magnetic dynamics have also been studied experimentally in magnetic wires based on the capability of the spin-transfer torque to move domain walls.

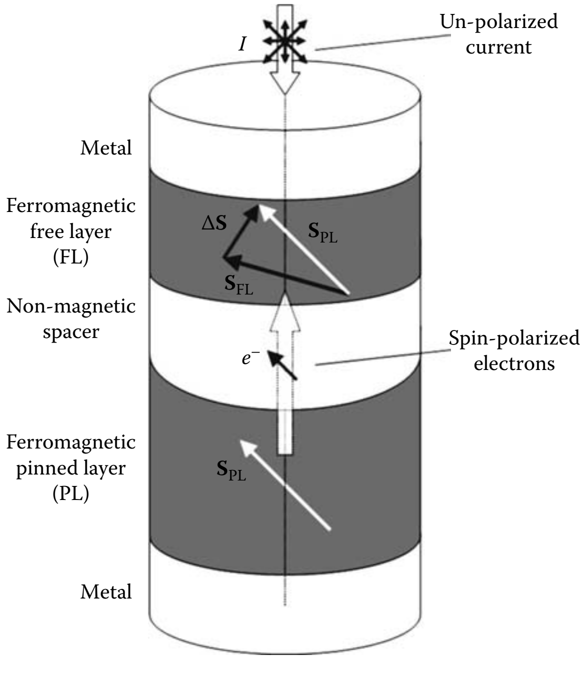

FIGURE 8.3 Schematic representation of the active region of a spin-valve when a perpendicular-to-plane electric current is injected through the structure. Please note that, because electrons bear negative charge, a downward current flow corresponds to an upward electron flow. A thicker ferromagnetic layer ("fixed" layer), having a magnetization vector $\mathbf{S}_{PL}$ fixed in space, and a thinner ferromagnetic layer ("free" layer), having a magnetization vector $\mathbf{S}_{FL}$ "free" to move, are separated by a nonmagnetic layer (spacer). The electrons of the current I exert a torque on the magnetization of the free layer. The current I is taken as positive when directed from the fixed to the free layer (electrons flow from the free into the fixed layer).

Moreover, depending on the geometry, properties of the layered magnetic structure, and magnitude and orientation of the external bias field, such a "spin-transfer torque effect" can lead either to the switching of the magnetization direction or to a stable magnetization precession of the magnetization of the thinner FL.

The discovery that spin-polarized current can alter a magnetic state opened several perspectives for a new class of nano-scale magnetic devices in which the additional degree of freedom constituted by the "spin" plays an important role. The corresponding new research area was consequently called *spintronics*, a short notation for "*spin-based electronics*". In general, spintronics refers to the study of the role played by electron (and more generally nuclear) spin in solid-state physics and in those devices that specifically exploit spin properties instead of, or in addition to, the degree of freedom given by the charge. The spin-transfer torque, by which spin-polarized currents can replace the role of a bias magnetic field in managing the spatial configuration of magnetization, offers the possibility of designing and building up a new class of devices and applications that includes, for example, magnetoresistive random access memory, logic gates, diodes and transistors, nano-oscillators, radiofrequency modulators, and detectors. To control device performance, it is critical to understand the nature of the spin-torque-driven magnetic excitations (Berkov and Miltat 2008, Ralph and Stiles 2008, Slavin and Tiberkevich 2008).

8.2 Background

We need, thus, a mathematical model to describe the magnetization dynamics driven by spin-polarized currents in nanoscale spin-valve systems.

To do that, as briefly introduced in Section 8.1, at a mesoscopic scale we need to introduce a continuum model, in terms of magnetic polarization per unit volume, and to characterize the state of a generic ferromagnetic body by means of its free energy.

To this aim, let us consider a region occupied by a magnetic body. Let us now focus on a "small" region dV_r within the body denoted by the position vector $r \in \Omega$. We refer to a "small" region to indicate that the volume dV_r is large enough to contain a huge number N of elementary magnetic moments $\boldsymbol{\mu}_j$, $j = 1, \ldots, N$, but small enough so that the average magnetic moment varies smoothly. In this respect, we define the magnetization vector field $\mathbf{M}(\mathbf{r})$, so that the product $\mathbf{M}(\mathbf{r})dV_r$ represents the net magnetic moment of the elementary volume dV_r:

$$\mathbf{M}(\mathbf{r}) = \frac{\sum_{j=1}^{N} \boldsymbol{\mu}_j}{\mathrm{d}V_r} \tag{8.1}$$

Moreover, we assume that the magnetization is also a function of time t:

$$\mathbf{M} = \mathbf{M}(\mathbf{r}, t) \tag{8.2}$$

The main hypotheses of *micromagnetism* consist of considering the magnetization vector $\mathbf{M}$ to be a continuous function of the position $\mathbf{r}$ within the body and whose modulus $|\mathbf{M}| = M_S$ is constant in time.

It is well known that the magnetization configuration of a ferromagnetic body should be such as to minimize its free energy $G(X, H, T) = F(X, T) - HX$. In this equation, $F(X, T)$ is the Helmholtz free energy of the system, H is the external field, and X is the state variable. The local minima obey the condition $\partial G/\partial X = 0$, by which all the metastable states are found. The great complication is that, in a space-dependent approach, X represents the full magnetization vector field $\mathbf{M}(\mathbf{r})$ defined over the entire body volume. Thus, the energy minimization has to be carried out in the infinite-dimensional functional space of all possible magnetization configurations, which, from the mathematical point of view, constitutes a nontrivial issue.

To characterize the energy landscape, it is necessary to first introduce the principal energy terms that control the behavior of a ferromagnetic body and show how they compete with each other.

It is well known that the most significant contributions to the energy balance arise from Maxwellian fields (magnetostatic and Zeeman) and non-Maxwellian ones (exchange and anisotropy). Let us briefly discuss each of them.

8.2.1 Exchange Interactions

The exchange interactions, which are the *source of the ferromagnetism,* should be vigorously analyzed by means of quantum theory, since this latter strongly concerns with spin–spin interactions. More specifically, on a scale in the order of the atomic scale, the exchange interaction tends to align neighboring magnetic moments. In other words, whenever $\mathbf{M}(\mathbf{r})$ changes orientation from point to point, we have some misalignment of the neighboring magnetic moments and this costs extra exchange energy. In view of a continuum average analysis in terms of a magnetization vector field, we expect that the exchange interactions tend to produce small uniformly magnetized regions, indeed observed experimentally and called magnetic domains.

The contribution of exchange interactions to the free energy of the whole magnetic body can be expressed as (Brown 1962)

$$F_{\mathrm{ex}} = \int_{\Omega} A\left[(\nabla m_x)^2 + (\nabla m_y)^2 + (\nabla m_z)^2\right] \mathrm{d}V \tag{8.3}$$

where

the parameter A is called "exchange constant"

the integration is extended over the whole volume region Ω

8.2.2 Anisotropy

Anisotropic effects in ferromagnetic bodies arise from the structure of the lattice and from particular symmetries that are produced in certain crystals. The experimental evidence of the existence of such energy-favored directions can be observed in certain ferromagnetic materials when, in the absence of external fields, the magnetization vector tends to be magnetized along precise directions, which in literature are referred to as *easy directions*. The fact that there is a "force" that tends to align magnetization along easy directions can be taken into account, in the micromagnetic framework, by means of an additional phenomenological term in the free energy functional

$$F_{\mathrm{an}}(\mathbf{m}) = \int_{\Omega} f_{\mathrm{an}}(\mathbf{m}) \,\mathrm{d}V \tag{8.4}$$

where $f_{\mathrm{an}}(\mathbf{m})$ is the anisotropy free energy density. In this phenomenological analysis, the easy directions correspond to the minima of the anisotropy energy density, whereas the saddle-points and maxima of $f_{\mathrm{an}}(\mathbf{m})$ determine the medium-hard axes and the hard axes, respectively.

8.2.3 Magnetostatic Interactions

Magnetostatic interactions represent the way the elementary magnetic moments interact over long distances within the body. In fact, the magnetostatic field at a given location within the body depends on the contributions from the whole magnetization vector field. We refer to this property as "nonlocal" character. Magnetostatic interactions can be taken into account by introducing the appropriate magnetostatic field $\mathbf{H}_m$ according to the Maxwell equations for magnetized media

$$\begin{cases} \nabla \cdot \mathbf{H}_m = -\nabla \cdot \mathbf{M} & \text{in } \Omega \\ \nabla \cdot \mathbf{H}_m = 0 & \text{in } \Omega^c \\ \nabla \times \mathbf{H}_m = 0 \end{cases} \tag{8.5}$$

where the quantity $-\nabla \cdot \mathbf{M} = \rho_M$ plays the role of "magnetic charge" density.

The system (8.5) has to be completed with the boundary conditions at the body discontinuity surface $\partial\Omega$

$$\begin{cases} \mathbf{n} \cdot [\mathbf{H}_m]_{\partial\Omega} = \mathbf{n} \cdot \mathbf{M} \\ \mathbf{n} \times [\mathbf{H}_m]_{\partial\Omega} = 0 \end{cases} \tag{8.6}$$

where

the quantity $\mathbf{n} \cdot \mathbf{M} = \sigma_M$ plays the role of "surface magnetic charge" density

$\mathbf{n}$ is the outward normal to the boundary $\partial\Omega$ of the magnetic body

$[\mathbf{H}_m]_{\partial\Omega}$ is the jump of the vector field $\mathbf{H}_m$ across $\partial\Omega$

The expression for the contribution of magnetostatic interactions to the free energy of the system is given by

$$F_m = -\int_{\Omega_\infty} \frac{1}{2}\mu_0 \mathbf{M}\cdot\mathbf{H}_m \mathrm{d}V \tag{8.7}$$

Equation (8.7) denotes the *nonlocal* character of the magnetostatic field, as it functionally depends, through the boundary value problem (8.6), on the spatial distribution of the magnetization vector field through the entire body volume.

8.2.4 External Field

When a magnetic body is subjected to an external bias field $\mathbf{H}_{\text{ext}}$, the contribution to the energy is simply

$$F_{\text{ext}} = -\int_{\Omega} \mu_0 \mathbf{M}\cdot\mathbf{H}_{\text{ext}}\, \mathrm{d}V \tag{8.8}$$

This energy term is referred to in literature as Zeeman energy.

8.2.5 Energy Balance and Free Energy Functional Formulation

It is now possible to summarize the role played by each of the previous energy terms described above. In fact, because each energy contribution favors different energy minima configurations, the metastable configuration arises from a sort of competition among them. For example, a magnetic material uniformly magnetized along its easy axis presents a low energy contribution arising from both exchange and magnetocrystalline anisotropy fields, but, at the same time, a higher contribution coming from both the magnetostatic and Zeeman fields (if the applied field direction is different from the easy axis). On the contrary, a configuration having a minimum magnetostatic energy is a closed-loop structure, but such a configuration presents higher contributions coming from all the other energy terms.

The expression for the free energy of the ferromagnetic body can hence be formulated by collecting the previous equations

$$\begin{aligned} G(\mathbf{M},\mathbf{H}_{\text{ext}}) &= F_{\text{ex}} + F_{\text{an}} + F_m + F_{\text{ext}} \\ &= \int_{\Omega}\Big\{A\Big[\left(\nabla m_x\right)^2 + \left(\nabla m_y\right)^2 + \left(\nabla m_z\right)^2\Big] \\ &\quad + f_{\text{an}} - \frac{1}{2}\mu_0\mathbf{M}\cdot\mathbf{H}_m - \mu_0\mathbf{M}\cdot\mathbf{H}_{\text{ext}}\Big\}\,\mathrm{d}V \end{aligned} \tag{8.9}$$

which can be put in the compact form by expressing the exchange interaction energy density as $A(\nabla\mathbf{m})^2$

$$G(\mathbf{M},\mathbf{H}_{\text{ext}}) = \int_{\Omega}\left\{A\left(\nabla\mathbf{m}\right)^2 + f_{\text{an}} - \frac{1}{2}\mu_0\mathbf{M}\cdot\mathbf{H}_m - \mu_0\mathbf{M}\cdot\mathbf{H}_{\text{ext}}\right\}\mathrm{d}V \tag{8.10}$$

By using a variational calculus, we are able to introduce now the so-called *effective field* as the variational derivative of the energy density of the system

$$\mathbf{H}_{\text{eff}} = -\frac{\delta G}{\delta \mathbf{M}} = \frac{2}{\mu_0 M_S}\nabla\cdot(A\nabla\mathbf{m}) - \frac{1}{\mu_0 M_S}\frac{\partial f_{\text{an}}}{\partial \mathbf{m}} + \mathbf{H}_m + \mathbf{H}_{\text{ext}} \tag{8.11}$$

where M_S expresses the saturation magnetization of the body.

8.2.6 Equation of Motion

Once the effective field is defined, it is possible to present the so-called Landau–Lifshitz (LL) equation (Landau and Lifshitz 1935)

$$\frac{\partial \mathbf{M}}{\partial t} = -\gamma\mathbf{M}\times\mathbf{H}_{\text{eff}} \tag{8.12}$$

which expresses the undamped precessional dynamics of the magnetization vector along the direction of the effective field (γ is the gyromagnetic ratio).

We observe that the Landau–Lifshitz equation (8.12) is a *conservative* (Hamiltonian) equation. Nevertheless, dissipative processes take place within the dynamic magnetization processes. The microscopic nature of this dissipation is still not clear and is currently the focus of considerable research. The approach generally followed to take into account dissipation precesses consists of introducing damping in a *phenomenological* way. Landau and Lifshitz introduced an additional torque term that pushes magnetization in the direction of the effective field. Then, the damped *Landau–Lifshitz equation* becomes, in the Gilbert formulation, (Gilbert 2004)

$$\frac{\partial \mathbf{M}}{\partial t} = -\gamma\mathbf{M}\times\mathbf{H}_{\text{eff}} + \frac{\alpha}{M_S}\left(\mathbf{M}\times\frac{\partial \mathbf{M}}{\partial t}\right) \tag{8.13}$$

where α is the Gilbert damping parameter. To better understand the meaning of undamped and damped precessional dynamics, in Figure 8.4 we sketched the trajectories described by Equations 8.12 and 8.13, respectively.

By means of simple algebraic steps, it could be easily demonstrated that Equation 8.13 can be rewritten in the equivalent form

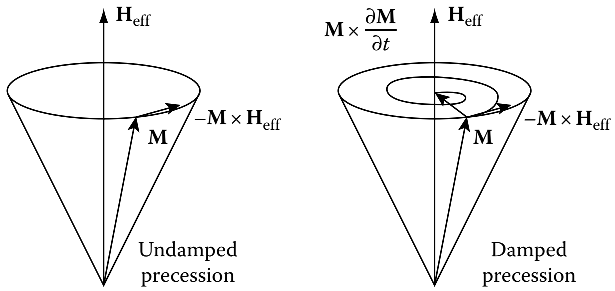
$\mathbf{H}_{\text{eff}}$
$-\mathbf{M}\times\mathbf{H}_{\text{eff}}$
$\mathbf{M}$
Undamped precession
$\mathbf{M}\times\frac{\partial\mathbf{M}}{\partial t}$
$\mathbf{H}_{\text{eff}}$
$-\mathbf{M}\times\mathbf{H}_{\text{eff}}$
$\mathbf{M}$
Damped precession

FIGURE 8.4 Comparison between undamped and damped precession.

$$\frac{\partial \mathbf{M}}{\partial t} = -\frac{\gamma}{1+\alpha^2}\mathbf{M}\times\mathbf{H}_{\text{eff}} - \frac{\gamma\alpha}{1+\alpha^2}\frac{1}{M_S}\mathbf{M}\times(\mathbf{M}\times\mathbf{H}_{\text{eff}}) \tag{8.14}$$

where the damping term is now expressed as a double vector product. Such formulation brings the advantage of being an explicit expression for the time derivative of the magnetization so that, as it will be more clear in the next paragraph, it is more suitable to be implemented in a numerical framework.

8.2.7 Equation of Motion Including Spin-Transfer Torque Effects

At this stage, spin-polarized current has not yet been considered.

It has been demonstrated that current-induced spin-transfer effects in a spin-valve layered structure, such as the one sketched in Figure 8.3, are taken into account by including another torque term (Slonczewski 1996) into the equation of motion. The dynamic equation for magnetization dynamics driven by spin-polarized currents is called the Landau–Lifshitz–Gilbert–Slonczewski (LLGS) equation, and it is in the form

$$\frac{\partial \mathbf{M}}{\partial t} = -\gamma(\mathbf{M}\times\mathbf{H}_{\text{eff}}) + \frac{\alpha}{M_S}\left(\mathbf{M}\times\frac{\partial \mathbf{M}}{\partial t}\right) + \frac{\sigma I}{M_S}f(r)\big[\mathbf{M}\times(\mathbf{M}\times\mathbf{p})\big] \tag{8.15}$$

In this case, **M** is the magnetization vector of the FL while **p** is the unit vector in the direction of spin-polarization (magnetization of the PL), *I* is the electric current traversing the spin-valve structure, the dimensionless function *f(r)* describes the spatial distribution of the current across the area of the FL, and the parameter σ is given by

$$\sigma = \frac{g\mu_B}{2eM_S LS}\varepsilon(\mathbf{M},\mathbf{p}) \tag{8.16}$$

where

g is the Landè factor
μ_B is the Bohr magneton
e is the modulus of the electron charge
L is the FL thickness
S is the area of the region involved by the current flow
$\varepsilon(\mathbf{M},\mathbf{p})$ is the dimensionless spin-polarization efficiency (Slonczewski 1996, 1999)

The latter parameter is, in general, a function of the relative alignment between PL and FL and of the material under investigation. In its original formulation, Slonczewski derived the expression

$$\varepsilon(\mathbf{M},\mathbf{p}) = \left[\frac{-4+\left(1+\eta^3\right)\left(3+\left(\mathbf{M}\cdot\mathbf{p}/M_{S_{\text{FL}}}M_{S_{\text{PL}}}\right)\right)}{4\eta^{3/2}}\right]^{-1} \tag{8.17}$$

with η being the polarization factor characteristic of the polarizing material. It has to be noticed, however, that in most cases, this dependence does not lead to any qualitative or significant quantitative effects and, especially when only a qualitative understanding of the problem is required, can it be approximated by a constant value (generally treated as a free parameter to fit experimental data).

By comparing the formulation of the dissipation torque in Equation 8.14 and the formulation of the spin-transfer torque in Equation 8.15, one observes that, for a proper direction of the applied current (and exactly for *positive* current, which conventionally means electrons flowing from the FL to the PL), these expressions present a similar vector structure but with opposite signs (Slonczewski 1999, Slavin and Kabos 2005). A way to understand how these effects work is to consider that, while dissipation pushes the magnetization vector *toward* the direction of $\mathbf{H}_{\text{eff}}$, the spin-transfer pushes the magnetization vector *away* from the direction of **p**. For such a reason, we might associate with the spin-transfer torque term the character of a "negative" damping, which opposes the natural "positive" Gilbert dissipation term (Slavin and Kabos 2005). The equality between these two terms represents the necessary condition under which stable persistent magnetization oscillations take place in a dissipative medium.

To give a complete picture on the mechanism underlying the process of spin-polarization and the subsequent spin-transfer torque effect, let us summarize in Figure 8.5 the FL magnetization dynamics occurring for both positive and negative currents.

The spin-polarization occurs via the absorption of the incident transverse spin current. The mechanism responsible for the transfer of the spin-angular momentum is the exchange interaction felt by electrons in the ferromagnet, which exerts a torque on the electron spins and, in turn, induces a reaction torque on the magnetization. In fact, let us consider the situation in which electrons incident from a nonmagnet are arranged

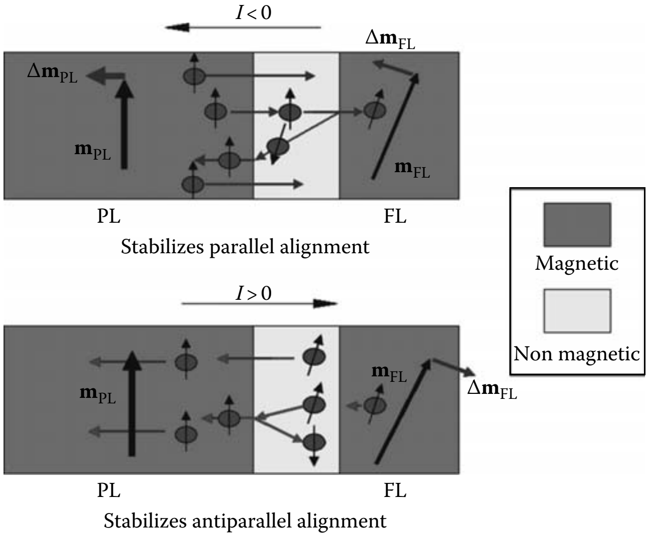

$I<0$
$\Delta\mathbf{m}_{\text{FL}}$
$\Delta\mathbf{m}_{\text{PL}}$
$\mathbf{m}_{\text{PL}}$
$\mathbf{m}_{\text{FL}}$
PL
FL
Stabilizes parallel alignment
Magnetic
Non magnetic
$I>0$
$\mathbf{m}_{\text{FL}}$
$\Delta\mathbf{m}_{\text{FL}}$
$\mathbf{m}_{\text{PL}}$
PL
FL
Stabilizes antiparallel alignment

FIGURE 8.5 Spin-transfer torque effects for both directions of the current.

over a distribution of states represented by three different incident directions. All of these electrons are in the same spin state, which is transverse to the ferromagnetic spin density (e.g., the magnetization of either FL or PL). The reflected electron spins have a predominantly minority character and undergo a random spin rotation whose precession axis is on average directed along the direction antiparallel to the one of the ferromagnet. The transmitted majority electron spins, on the other hand, precess along the direction parallel to the ferromagnetic spin density as they go into the ferromagnet because the wave vectors for the majority and minority components are different. It has to be pointed out that, although the proposed picture is qualitatively correct for most of the commonly used metal/ferromagnet interfaces (e.g., Cu/Co), additional phenomena (such as spin-dephasing and spin-accumulation) should be taken into account to get a quantitative understanding of the investigated dynamics (Stiles and Miltat 2006).

By looking at Figure 8.5 it turns out that, for negative current (top figure), the electrons crossing the interface metal/FL are on average polarized along the direction of the PL and the applied torque ($\Delta\mathbf{m}_{FL}$) favors the parallel alignment between the two magnetic layers. The back-torque operated by the reflected electrons on the PL ($\Delta\mathbf{m}_{PL}$) is generally negligible as the PL presents usually large values for saturation magnetization and thickness, which keep it fixed in its equilibrium state. On the contrary, for positive current (bottom figure), the electrons reflected from the PL hit the interface metal/FL with an antiparallel polarization with respect to that of the PL. In this case, the consequent torque causes a destabilization of the FL magnetization, which tries to favor an antiparallel configuration between the magnetization of the two layers.

These phenomena hold under two main assumptions: (1) the length of the nonmagnetic path that electrons experience between the two magnetic layers has to be smaller than a given characteristic length scale, named "spin-diffusion length" (according to the classical drift diffusion model of spin transport, the ensemble averaged spin of electrons drifting and diffusing in a solid decays exponentially with distance due to spin dephasing interactions) and (2) the thickness of the magnetic layers is assumed to be too great for appreciable tunneling of minority-spin electrons, therefore, the component antiparallel to the magnetization is totally reflected back into the nonmagnetic spacer, while the component parallel to the magnetization is totally transmitted into the magnetic layer.

Of course, it has to be recalled that the current-driven destabilization mechanism, apart from occurring only for a proper direction of the spin-polarized current, also exhibits a *threshold* character, i.e., there exists a given critical current above which the phenomena observable in such devices (reversal or persistent oscillations) take place. The simplest explanation lies in the energy that the applied current has to supply to the system in order to modify the energy landscape created by all the contributions of the effective field and to balance the natural positive dissipation (Slonczewski 1999, Slavin and Kabos 2005).

8.3 Computational Micromagnetics of Nanoscale Spin-Valves: State of the Art

Let us come back to the mathematical viewpoint by recalling that the influence of current-induced spin-transfer effects on the magnetization dynamics of nanoscale devices is expressed by the LLGS equation (Equation 8.15). This differential equation presents a complex nonlinear and nonlocal structure, which, in general, prevents the formulation of an exact analytical solution unless strong simplifying assumptions are imposed to the analyzed problem.

To gain a better physical understanding of the complicated magnetization dynamics, alternative strategies make use of the numerical integration of the equation of motion by means of different techniques, such as boundary element (BE), finite elements (FE), finite differences (FD), or hybrid procedures. Such approaches, to some extent, offer better-accuracy solutions, with respect to analytical approaches, simply because they generally use a smaller set of restrictions and simplifying hypotheses. Numerical tools can also be used to explore the range of validity of analytical theories, an analysis which sometimes could not be carried out otherwise.

On a practical level, when numerically solving partial differential equations, the primary challenge is to create an equation that approximates the equation to be studied. The main requirement is the numerical stability of the integration scheme, meaning that errors in the input data and intermediate calculations do not accumulate and cause the resulting output to deviate substantially from the exact solution or, in the worst case, to be meaningless.

Let us give a brief overview of the numerical methods cited above.

The BE method is a numerical computational method used to solve linear partial differential equations that have been formulated as integral equations, and is applicable to problems for which Green's functions can be calculated. The boundary element method attempts to use the given boundary conditions to fit boundary values into the integral equation, rather than values throughout the space defined by a partial differential equation. Once this is done, in the post-processing stage, the integral equation can then be used again to numerically calculate the solution directly at any desired point in the interior of the solution domain. The boundary element method is often more efficient than other methods, including finite elements, in terms of computational resources for problems where there is a small surface/volume ratio even though it generally gives rise to fully populated matrices, which in turn yield a substantial increase of the storage and computational time requirements. In these cases, compression techniques might be applied, even though additional restrictions to the problem have to be added.

The FE method is a numerical technique for finding the approximate solutions of partial differential equations as well as of integral equations. The solution approach is based either on eliminating the differential equation completely (steady-state problems) or rendering the partial differential equation into an approximating system of ordinary differential equations, which are then numerically integrated using standard techniques (e.g., Euler's method, Runge-Kutta, etc.). In other words, the basic idea is to replace the infinite dimensional linear problem with a finite dimensional version. The algorithm can be subdivided into two steps: (1) rephrase the original problem in its weak or variational form and (2) discretize the weak form in a finite dimensional space. In step (2), one has to choose both finite-dimensional subspace of the original space and the shape functions that constitute the basis of the searched solution in that subspace. Because the original space is subdivided into a mesh of polygonal cells having whatever geometrical shape (triangular, tetragonal, etc.) the FE method is a good choice for solving partial differential equations when the computational domain presents a complex geometrical structure (e.g., for nonflat boundaries problems), but also when the desired precision varies over the entire domain or when the solution lacks smoothness (see Figure 8.6).

FD methods are numerical methods for approximating the solutions to differential equations using finite difference equations to approximate derivatives. The FD method implies a discretization of the computational region by using prismatic cells and then each derivative in the equation of motion has to be approximated (by using a given integration scheme) by a certain discretization.

A comparison between the FE and FD methods brings out the following main considerations:

- The FD method is an approximation of the differential equation, while the FE method is an approximation of the solution.
- The most attractive feature of the FE method lies in its ability to handle complex geometries (and boundaries) by means of a proper choice of the discretization cells. The FD method, on the contrary, in its basic formulation, is restricted to handle rectangular shapes.
- The most attractive feature of the FD method lies in the fact that it might be very easy to be implemented.

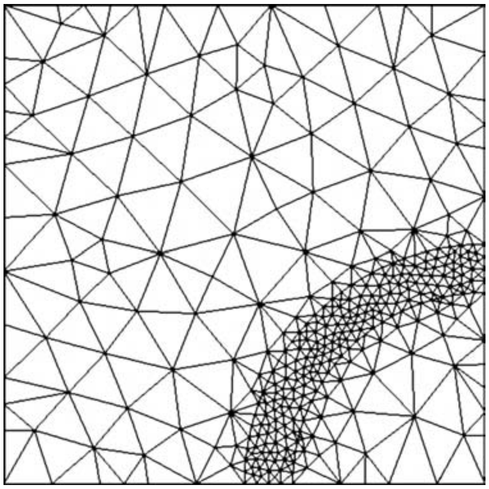

FIGURE 8.6 Example of mesh for FE methods with variable grid size to get a better resolution in correspondence of regions of interest.

Let us now focus on the usage of FD methods with the aim of showing how the numerical solution of the LLGS equation can be carried out and, at the same time, offer a brief overview of the state-of-the-art micromagnetic studies applied to nanoscale devices.

8.3.1 Numerical Solution of LLGS Equation Based on FD Methods

The first step of a numerical approach consists of subdividing the entire volume of the specimen by using a regular mesh of prismatic cells ($\Delta_x \times \Delta_y \times \Delta_z$), as in Figure 8.7.

To establish a proper dimension for the grid size, it is necessary to define before a fundamental quantity, named "exchange length," as follows:

$$l_{\text{ex}} = \sqrt{\frac{2A}{\mu_0 M_S^2}} \qquad (8.18)$$

The exchange length gives an estimation of the characteristic dimension on which exchange interactions are dominant. For typical magnetic materials (e.g., silicon-iron alloys, called "Permalloy") l_{ex} is in the order of 5 ÷ 10 nm. Therefore, one expects that on a spatial scale in the order of l_{ex}, the magnetization is spatially uniform. For such a reason, within a micromagnetic framework, the cell size has to be smaller than the exchange length value, i.e., typically a cell-size as large as a 2 ÷ 5 nm-side is used. At the same time, the cell size should not be considered too small for at least two reasons: (1) the computational time should be kept in an affordable range and (2) the cell dimensions should be consistent with the mesoscopic scale.

Once the spatial discretization is addressed, the next step is to choose a proper time-step $\Delta\tau$. This value changes according to the integration scheme adopted, and the main constrain it has to satisfy is the convergence of the solution. To check that convergence, the criterion generally used is to analyze the time-evolution of the total system energy. Because of dissipation, the energy can only diminish, so that a gradual increase of the total energy reflects a divergent algorithm.

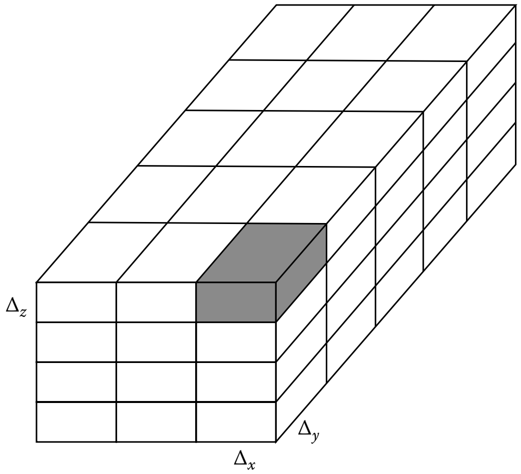

FIGURE 8.7 Example of mesh for FD methods.

Then, starting from an initial configuration of magnetization $\mathbf{m}(\mathbf{r},\tau)\ \forall\mathbf{r}\in V$, we compute the effective field $\mathbf{h}_{\text{eff}}(\mathbf{r},\tau) = \mathbf{h}_{\text{eff}}(\mathbf{m}(\mathbf{r},\tau),\tau)$ (see below for details) in each computational cell.

After that, it is necessary to apply a numerical integration scheme to solve the LLGS equation and to obtain the updated value of the magnetization distribution in the next time step $\tau+\Delta\tau$: $\mathbf{m}(\mathbf{r},\tau+\Delta\tau)\ \forall\mathbf{r}\in V$. Many integration schemes can be applied, either explicit or implicit, either with a fixed time-step or with an adaptive time-step (e.g., Euler, predictor-corrector, Runge-Kutta, mid-point, Adams-Bashforth-Moulton, Milne-Simpson, etc.). In most cases, one has to compute a finite differences representation for the derivate: that is the main reason why an explicit formulation for $\partial\mathbf{M}/\partial t$, like the one in Equation 8.14, is preferable with respect to the one shown in Equation 8.13.

Because our goal is to model the dynamic behavior of spintronic devices having a spin-valve structure as the one presented in Section 8.2, it is now necessary to focus on some aspects related to the implementation of the effective field for such devices. In the cases under investigation, the effective field has to be updated to take into account the additional contributions:

$$\mathbf{H}_{\text{eff}} = \mathbf{H}_{\text{ex}} + \mathbf{H}_{\text{an}} + \mathbf{H}_{\text{m}} + \mathbf{H}_{\text{ext}} + \mathbf{H}_{\text{mc}} + \mathbf{H}_{\text{Amp}} + \mathbf{H}_{\text{eddy}} + \mathbf{H}_{\text{th}} \tag{8.19}$$

Here $\mathbf{H}_{\text{ex}}$, $\mathbf{H}_{\text{an}}$, $\mathbf{H}_{\text{m}}$, and $\mathbf{H}_{\text{ext}}$ represent the exchange, anisotropy, magnetostatic, and external field contributions already introduced previously. The new contributions arise from $\mathbf{H}_{\text{mc}}$, which represents the magnetostatic coupling between PL and FL (it has to be distinguished from $\mathbf{H}_{\text{m}}$, which is the self-magnetostatic, or demagnetizing, field induced by the magnetization of each layer within its volume); $\mathbf{H}_{\text{Amp}}$, which represents the Ampère (or Oersted) magnetic field created by the current flow; $\mathbf{H}_{\text{eddy}}$, the field contribution arising from eddy currents; and $\mathbf{H}_{\text{th}}$, which represents the stochastic contribution arising from thermal noise.

Let us now give some details about the numerical implementation of each contribution within a FD micromagnetic framework. To do that, we will use a normalized notation, where each term is divided by the saturation magnetization value and written in lowercase letter, e.g., $\mathbf{m} = \mathbf{M}/M_S$.

The exchange field, $\mathbf{h}_{\text{exc}}$, is computed by considering the discrete expression for the exchange energy

$$f_{\text{ex}}(i,j,k) = A\left[\left(\nabla m_x(i,j,k)\right)^2 + \left(\nabla m_y(i,j,k)\right)^2 + \left(\nabla m_z(i,j,k)\right)^2\right] \tag{8.20}$$

where we used the notation for each Cartesian component $\alpha = x,y,z$

$$(\nabla m_\alpha)^2 = \left(\frac{\partial m_\alpha}{\partial x}\right)^2 + \left(\frac{\partial m_\alpha}{\partial y}\right)^2 + \left(\frac{\partial m_\alpha}{\partial z}\right)^2 \tag{8.21}$$

Under the approximation of finite differences, the derivative is substituted by the incremental differences (computed at the center of each cell), that is

$$(\nabla m_\alpha)^2 \approx \left(\frac{\Delta_x m_\alpha}{\Delta x}\right)^2 + \left(\frac{\Delta_y m_\alpha}{\Delta y}\right)^2 + \left(\frac{\Delta_z m_\alpha}{\Delta z}\right)^2 \tag{8.22}$$

where Δx, Δy, Δz are the sizes of the computational cells, whereas Δ_x, Δ_y, Δ_z represent the finite difference operators along the Cartesian axes x, y, z. Let us consider the magnetization vector $\mathbf{M}(i,j,k) = M_S\mathbf{m}(i,j,k)$ at the point (i,j,k) and $\mathbf{M}(i+1,j,k) = M_S\mathbf{m}(i+1,j,k)$ the corresponding vector at the neighboring cell $(i+1,j,k)$ in the x direction. We can rewrite the expression (8.22) as follows:

$$\begin{aligned}\left(\frac{\Delta_x m_\alpha}{\Delta x}\right)^2 &= \left(\frac{m_x(i+1,j,k) - m_x(i,j,k)}{\Delta x}\right)^2\\ &= \frac{m_x^2(i+1,j,k) - 2m_x(i+1,j,k)m_x(i,j,k) + m_x^2(i,j,k)}{\Delta x^2}\end{aligned} \tag{8.23}$$

By taking into account that $|\mathbf{m}(i,j,k)| = 1\ \forall\ (i,j,k)$, the expression (8.23) can be simplified as

$$\left(\frac{\Delta_x m_\alpha}{\Delta x}\right)^2 = \frac{1}{\Delta x^2}\left[2 - 2m_x(i+1,j,k)m_x(i,j,k)\right] \tag{8.24}$$

Also, by taking into account the neighboring cell in the negative directions, one ends up with

$$\begin{aligned}(\nabla m_x)^2 \approx{}& \frac{1}{\Delta x^2}\left[\left(2-2m_x(i+1,j,k)m_x(i,j,k)\right) + \left(2-2m_x(i,j,k)m_x(i-1,j,k)\right)\right]\\ &+ \frac{1}{\Delta y^2}\left[\left(2-2m_x(i,j+1,k)m_x(i,j,k)\right) + \left(2-2m_x(i,j,k)m_x(i,j-1,k)\right)\right]\\ &+ \frac{1}{\Delta z^2}\left[\left(2-2m_x(i,j,k+1)m_x(i,j,k)\right) + \left(2-2m_x(i,j,k)m_x(i,j,k-1)\right)\right]\end{aligned} \tag{8.25}$$

Under the hypothesis of cubic cells ($\Delta x = \Delta y = \Delta z$), we can generalize Equation 8.25 to the case of $N_x \times N_y \times N_z$ cells (along the x, y, and z axis, respectively) as follows:

$$\begin{aligned}f_{\text{ex}}(i,j,k) &= \frac{2A}{\Delta x^2}\sum_{i'=1}^{N_x}\sum_{j'=1}^{N_y}\sum_{z'=1}^{N_z}\left[1 - \mathbf{m}(i,j,k)\mathbf{m}(i',j'k')\right]\\ &= \frac{2A}{\Delta x^2}N_xN_yN_z - \frac{2A}{\Delta x^2}\mathbf{m}(i,j,k)\sum_{i'=1}^{N_x}\sum_{j'=1}^{N_y}\sum_{z'=1}^{N_z}\mathbf{m}(i',j'k')\end{aligned} \tag{8.26}$$

The exchange field is computed by using the functional derivative of the energy ($\delta/\delta\mathbf{m}$), which becomes the ordinary derivative ($\partial/\partial\mathbf{m}$) in the discrete domain:

$$h_{\text{exc}}(i,j,k) = -\frac{1}{\mu_0 M_S^2}\frac{\partial f_{\text{ex}}(i,j,k)}{\partial\mathbf{m}(i,j,k)} \tag{8.27}$$

By substituting Equation 8.26 into Equation 8.27, one obtains the final implementation of the six-neighbors exchange field:

$$h_{\mathrm{exc}}(i,j,k)=\frac{2A}{\mu_0 M_S^2}\left[\frac{\mathbf{m}(i+1,j,k)+\mathbf{m}(i-1,j,k)}{\Delta x^2}+\frac{\mathbf{m}(i,j+1,k)+\mathbf{m}(i,j-1,k)}{\Delta y^2}+\frac{\mathbf{m}(i,j,k+1)+\mathbf{m}(i,j,k-1)}{\Delta z^2}\right] \tag{8.28}$$

It should be pointed out that the partial derivative approximation for finite differences holds for small arguments, that is, when the angle formed between the magnetization vectors in two adjacent cells is small.

The anisotropy field, $\mathbf{h}_{\mathrm{an}}$, in the case of uniaxial anisotropy, is simply computed by

$$h_{\mathrm{ani}}(i,j,k)=\frac{2K_C}{\mu_0 M_S^2}\left(\mathbf{m}(i,j,k)\cdot\mathbf{u}_{\mathrm{K}}\right)\mathbf{u}_{\mathrm{K}} \tag{8.29}$$

where

K_C is the anisotropy constant

$\mathbf{u}_{\mathrm{K}}$ is the direction of uniaxial anisotropy

Both magnetostatic fields ($\mathbf{H}_{\mathrm{m}}$ and $\mathbf{H}_{\mathrm{mc}}$) are computed by using the generalization of the demagnetizing tensor given in (Newell et al. 1993)

$$h_{m,\alpha}(i,j,k)=\sum_{\beta}^{(x,y,z)}\sum_{i'=1'}^{N_x}\sum_{j'=1}^{N_y}\sum_{z'=1}^{N_z}N_{\alpha\beta}(i-i',j-j',k-k')m_{\beta}(i',j',k') \tag{8.30}$$

where

$h_{m,\alpha}$ represents the component of the magnetostatic field

$N_{\alpha,\beta}(i-i',j-j',k-k')$ are the components of the Newell magnetostatic tensor, which only depend on both the distance between the source and the cell position, and the specimen geometry

The expression (8.30) presents the form of a three-dimensional convolution in the "*coordinates space*" (x,y,z). By using the *Convolution Theorem*, it is possible to express the previous convolution as a product scalar in the "*Fourier space*," or phases space (k_x,k_y,k_z):

$$\begin{aligned}\tilde{h}_{m,\alpha}(k_x,k_y,k_z)&=\Im\left[h_{m,\alpha}(i,j,k)\right]\\&=\sum_{\beta}^{(x,y,z)}\Im\left[N_{\alpha\beta}(i,j,k)\right]\Im\left[m_{\beta}(i,j,k)\right]\\&=\sum_{\beta}^{(x,y,z)}\tilde{N}_{\alpha\beta}(k_x,k_y,k_z)\tilde{m}_{\beta}(k_x,k_y,k_z)\end{aligned} \tag{8.31}$$

To calculate the magnetostatic field, $\mathbf{h}_{\mathrm{m}}$, it is necessary to follow these steps: (1) calculate the Newell tensor coefficients $N_{\alpha,\beta}$; (2) perform the FFT of the previous tensor $\tilde{N}_{\alpha\beta}$; (3) perform the FFT of the magnetization distribution in each cell $\tilde{m}_{\beta}$; and (4) perform the inverse FFT of the inner product between $\tilde{N}_{\alpha\beta}$ and $\tilde{m}_{\beta}$:

$$\begin{aligned}h_{m,\alpha}(i,j,k)&=\Im^{-1}\left[\tilde{h}_{m,\alpha}(k_x,k_y,k_z)\right]\\&=\Im^{-1}\left[\sum_{\beta}^{(x,y,z)}\Im\left[N_{\alpha\beta}(i,j,k)\right]\Im\left[m_{\beta}(i,j,k)\right]\right]\end{aligned} \tag{8.32}$$

By denoting with N the total number of grid cells, such approach brings a computational complexity that grows as $N\log N$.

The external field $\mathbf{h}_{\mathrm{ext}}$ is simply computed, in the limit of uniform fields, by applying the corresponding components of the external field to each computational cell:

$$h_x(i,j,k)=\mathbf{h}_{\mathrm{ext}}\cdot\hat{\mathbf{e}}_{\mathbf{x}},\quad h_y(i,j,k)=\mathbf{h}_{\mathrm{ext}}\cdot\hat{\mathbf{e}}_{\mathrm{y}},\quad h_z(i,j,k)=\mathbf{h}_{\mathrm{ext}}\cdot\hat{\mathbf{e}}_{\mathbf{z}} \tag{8.33}$$

Moreover, once we inject an electric current density $\mathbf{J}=J_0\mathbf{e}_{\mathbf{z}}$ into a spin-valve stack composed by ferromagnetic and metallic materials, a new magnetic field is generated according to the Ampère (or Oersted) law. The values of the *Ampère field*, $\mathbf{H}_{\mathrm{Amp}}$, in each point of the volume can be calculated by using the Ampère law:

$$\nabla\times\mathbf{H}_{\mathrm{Amp}}(\mathbf{r},t)=\mathbf{J}(\mathbf{r},t) \tag{8.34}$$

The Ampère field is a solenoidal field and thus exhibits the following properties:

$$\begin{cases}\nabla\cdot\mathbf{H}_{\mathrm{Amp}}(\mathbf{r},t)=0\\\nabla\times\mathbf{H}_{\mathrm{Amp}}(\mathbf{r},t)=\mathbf{J}(\mathbf{r},t)\end{cases} \tag{8.35}$$

An additional contribution arises from *eddy currents*. An eddy current (also known as a Foucault current) is an electrical phenomenon discovered by Léon Foucault in 1851. It is caused when a moving (or changing) magnetic field intersects a conductor or vice-versa. The relative motion causes a circulating flow of electrons, or current, within the conductor. These circulating eddies of current create electromagnets with magnetic fields that oppose the effect of the applied magnetic field (see Lenz's law). The stronger the applied magnetic field, or the greater the electrical conductivity of the conductor, or the greater the relative velocity of motion, the greater the currents developed and the greater the opposing field. The Maxwell equations for eddy currents are given by

$$\begin{cases}\nabla\times\mathbf{H}_{\mathrm{eddy}}(\mathbf{r},t)=\sigma'\mathbf{E}(\mathbf{r},t)+\varepsilon'\dfrac{\partial\mathbf{E}(\mathbf{r},t)}{\partial t}\\\nabla\cdot\mathbf{H}_{\mathrm{eddy}}(\mathbf{r},t)=0\\\nabla\times\mathbf{E}(\mathbf{r},t)=-\dfrac{\partial\mathbf{B}(\mathbf{r},t)}{\partial t}\\\nabla\cdot\mathbf{E}(\mathbf{r},t)=0\end{cases} \tag{8.36}$$

where
- σ' is the electrical conductivity
- ε' is the electric permeability
- $\mathbf{E}$ is the electric field

To perform an analogy between $\mathbf{H}_{\text{Amp}}(\mathbf{r}, t)$ and $\mathbf{H}_{\text{eddy}}(\mathbf{r}, t)$, in the former the source of the field is given by the applied current density $\mathbf{J}(\mathbf{r}, t)$, whereas in the latter the sources of the field are given by both the time variation of the magnetic field $-\partial\mathbf{B}(\mathbf{r}, t)/\partial t$ and the conduction current $\mathbf{J}_C = \sigma'\mathbf{E}(\mathbf{r}, t) + \varepsilon'(\partial\mathbf{E}(\mathbf{r}, t)/\partial t)$. The effects of eddy currents become relevant with an increase in the device volume and, since we will deal with structures with relative small volumes, this contribution is generally disregarded (Martinez et al. 2004).

It has to be noted that both $\mathbf{H}_{\text{Amp}}$ and $\mathbf{H}_{\text{mc}}$ are nonuniform fields so that their effects cannot be correctly evaluated by using *macrospin* models (by "macrospin" we mean a particle that remains homogeneously magnetized independently of external conditions). Figures 8.8 and 8.9 show examples of the spatial configuration of these fields for a structure having circular and rectangular cross-sections, respectively.

The last contribution to the effective field arises from the effect of thermal fluctuations (or noise) on the magnetization dynamics of nanoscale devices. Considering that experimental data are well described by the current dependent activation barriers that agree with the prediction of the LLG-based models (Garcia-Palacios and Lazaro 1998, Krivorotov et al. 2004, Li and Zhang 2004), it is possible to include in a micromagnetic framework a thermal field as an additive random field to the deterministic effective field for each cell. It is done by starting from the formalism of Brownian motion (Gardiner 1985, Kloeden and Platen 1999).

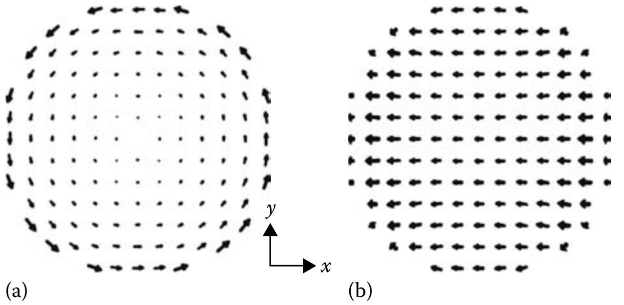

FIGURE 8.8 Spatial distribution of (a) Ampère field and (b) magnetostatic coupling field for a device having a circular cross-section.

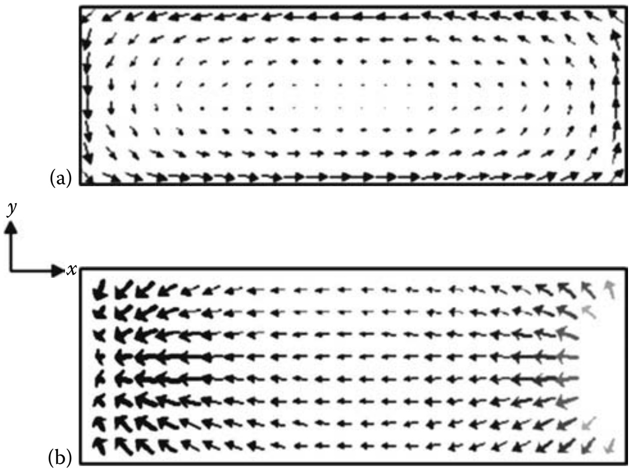

FIGURE 8.9 Spatial distribution of (a) Ampère field and (b) magnetostatic coupling field for a device having a rectangular cross-section.

In this case, the previous deterministic equation become a stochastic one, named the *Langevin equation*. The fluctuating term, which appears in the Langevin equation presents an integral structure that needs to be analyzed by using the theory of stochastic processes (Brown 1963, 1979, Garcia-Palacios and Lazaro 1998).

From the physical point of view (Garcia-Palacios and Lazaro 1998), the more appropriate interpretation of those stochastic integrals is called the *Stratonovich interpretation*. In this sense, Brown (1963, 1979) developed the *Fokker–Planck equation* to study the temporal evolution of nonequilibrium probability distributions. Several techniques have been developed to solve the Fokker–Planck equation. Some of those allow the direct computation of the probability distribution of magnetization, while others numerically compute the eigenvalues and amplitudes of the most relevant dynamics modes. Instead of solving the Fokker–Planck equation, an approximation is to build the solutions of the Langevin equation (Garcia-Palacios and Lazaro 1998). The Langevin dynamics gives origin to the stochastic trajectories of the system variables (time-evolution of magnetization) and, starting from this approach, we can compute the statistic behavior by averaging over many realizations (for "realizations" we mean a particular statistic trajectory obtained by starting from a given sequence of aleatory numbers for the statistic system variables).

From the computational point of view, we should take into account that thermal fluctuations in magnetic media occur at correlation times much shorter than the typical response time of magnetic systems (Brown 1963, 1979). In other words, thermal fluctuations arise from high-frequency perturbations. The thermal field that is generally used to simulate thermal effects can be therefore represented as a *Wiener stochastic process* (Van Kampen 1987). Moreover, because fluctuations arise from the interaction between magnetization and an enormous number of microscopic degrees of freedom (phonons, conduction electrons, nuclear spins, etc.) having equivalent stochastic properties, it is generally assumed that the above stochastic process can be treated as a *Gaussian white noise* (Brown 1963, Garcia-Palacios and Lazaro 1998). The same microscopic degrees of freedom are responsible for the relaxation of the magnetization precession, because fluctuations and dissipation are phenomena strictly related each other (*Fluctuation-Dissipation Theorem*). Because of those degrees of freedom, the thermal field $\mathbf{H}_{\text{th}}$, to be added to the deterministic effective field, has to satisfy the properties of a Gaussian stochastic vector process.

This leads to a definition of the stochastic Langevin–Landau–Lifshitz–Gilbert (LLLG) equation. In order to also take into account the spin-torque term within this new formulation, the main hypotheses are: (1) the spin torque does not contain a fluctuating field, (2) the fluctuating field does not depend on spin torque, and (3) the magnetization configuration of the PL does not depend on temperature.

The thermal field, $\mathbf{H}_{\text{th}}$, is hence included as a random fluctuating three-dimensional vector quantity given by

$$\mathbf{H}_{\text{th}} = \xi\sqrt{2\frac{\alpha}{1+\alpha^2}\frac{k_B T}{\mu_0 \gamma' \Delta V M_S \Delta t}} \tag{8.37}$$

where

k_B is the Boltzmann constant

ΔV is the volume of the computational cubic cell

Δt is the simulation time-step

T is the temperature of the sample

$\gamma' = \gamma/(1 + \alpha^2)$

ξ is a Gaussian stochastic process

The thermal field satisfies the following statistical properties:

$$\begin{cases} \langle H_{\text{th},k}(t) \rangle = 0 \\ \langle H_{\text{th},k}(t) H_{\text{th},l}(t') \rangle = D\delta_{kl}\delta(t-t') \end{cases} \tag{8.38}$$

where k and l represent the Cartesian coordinates x, y, and z. According to this, each component of $\mathbf{H}_{\text{th}}$ is a space and time independent random Gaussian distributed number (Wiener process) with zero mean value. The constant D measures the strength of thermal fluctuations and its value is obtained from the Fokker–Planck equation.

Finally, the micromagnetic framework has to be completed with proper boundary conditions, which describe the behavior of the magnetization vector at the computational boundaries. The formulation of boundary conditions is therefore dependent on the geometry under investigation. Some examples will be given in Section 8.3.3.

8.3.2 Spin-Transfer Torque and Giant Magnetoresistance: Dual Phenomena

In Section 8.3.1, we described the spin-transfer torque mechanism and showed the dynamic equation of motion for the magnetization which includes this effect. We described that, if a spin-polarized current presents proper sign and amplitude, it may destabilize the equilibrium configuration of the magnetization of the thinner FL leading it to either a reversal process or a persistent precession. Lots of questions could immediately follow such analysis: are these phenomena so interesting to deserve potential applications in the field of information technology? If yes, how do we observe and detect experimentally these magnetization dynamics?

The answer lies in the existence of another phenomenon, called giant magnetoresistance (GMR), which can be considered the dual mechanism of the spin-transfer torque (STT). Observing the phenomena from a more general viewpoint, we can describe the STT as the mechanism by which we account for the modifications of the magnetic state of a system induced by an electric current. For duality, the GMR effect (Baibich et al, 1988, Parkin et al. 1990, 1991, Schad et al. 1994) accounts for the variations of the electrical state of a system induced by the magnetic state (see Figure 8.10). For the discovery of the GMR effect, Fert and Gruenberg shared the Nobel Prize for Physics in 2007.

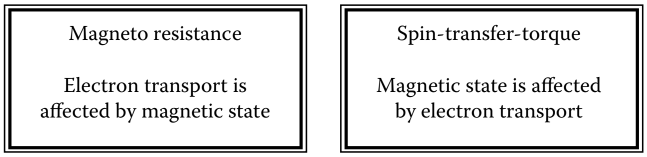
Magneto resistance
Electron transport is affected by magnetic state
Spin-transfer-torque
Magnetic state is affected by electron transport

FIGURE 8.10 STT and GMR effects as dual mechanisms.

The GMR is a quantum-mechanical effect observed in thin film structures composed of alternating ferromagnetic and nonmagnetic metal layers and arises from the spin-dependent scattering of conduction electrons by spins in ferromagnetic layers. When the spin of ferromagnetic layers is anti-parallel to the conduction electrons, then the scattering is increased. Therefore, the magnetoresistivity will be varied by the relative orientation of magnetization in neighboring ferromagnetic layers separated by a nonmagnetic metallic thin layer. In detail, the GMR effect manifests itself as a significant decrease in resistance from the zero-field state, when the magnetization of adjacent ferromagnetic layers are antiparallel due to a weak anti-ferromagnetic coupling between layers, to a lower level of resistance when the magnetization of the adjacent layers align due to an applied external field. The spin of the electrons of the nonmagnetic metal align parallel or antiparallel with an applied magnetic field in equal numbers, and therefore suffer less magnetic scattering when the magnetizations of the ferromagnetic layers are parallel.

In the analyzed case of the three-layers system, it behaves as a spin-valve for electron transport from one ferromagnetic layer to the other. The relative alignment of the magnetization vectors of the two ferromagnetic layers can be easily detected by measuring the system resistance (see Figure 8.11). At the same time, it is possible to associate two different logic states to the different electric configurations (e.g., bit 0 = low-resistance state, bit 1 = high-resistance state). This is the basic working principle of magnetoresistive random access memories (MRAM). On the other hand, if the magnetization undergoes a persistent precession, the output signal might manifest itself as a periodic variation of the resistance of the device, which could be used to build up nanoscale oscillators.

8.3.3 Nanoscale Devices: Pillar vs. Point-Contact

Within the class of nanoscale devices, the two most common experimental geometries are the so-called nano-pillar and nano-contact (or point-contact).

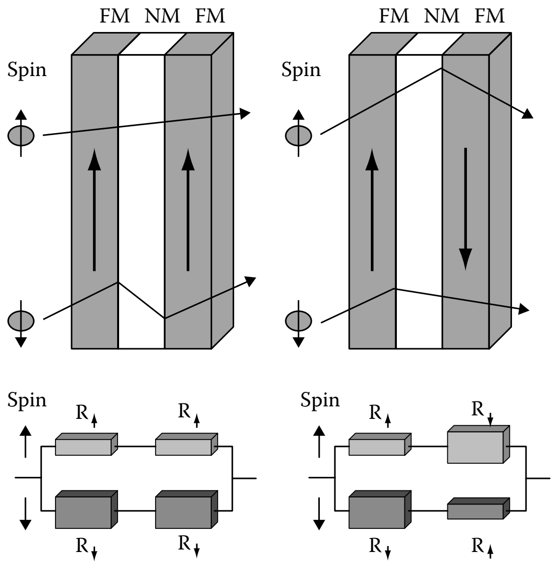

Low-resistance state High-resistance state

FIGURE 8.11 A schematic of a GMR-based spin-valve GMR showing a higher resistance state in the anti-parallel configuration with respect to the parallel one.

The main differences existing between them involve both the lateral sizes and the current density distribution. In fact, nano-contact devices generally present larger lateral dimensions (in the order of tenths of microns) with respect to those of nano-pillar (hundreds of nanometers). In addition, the perpendicular-to-plane applied current involves the whole physical (cross-section) area of pillar devices, whereas only a reduced region (the contact area, typically of the radius of 10–80 nm) is involved in nano-contact devices (see Figure 8.12).

The nano-pillar structure is widely used as an elementary memory cell where spin-polarized current drives the process of magnetization reversal (or switching), even in the absence of an external bias magnetic field.

On the other hand, the nano-contact geometry becomes particularly suitable to design microwave spin-transfer oscillators with very narrow linewidths because it combines the advantages of having a larger effective volume (with respect to pillar devices) (Kim et al. 2008) and a negligible influence of edge roughness and defects (Rippard et al. 2004). In fact, in the case of extended nano-contact geometry (also referred to as "nonconfined" systems), the FL is not bounded in the plane; therefore, the magnetic oscillations induced by spin-polarized current can excite spin-wave modes, which propagate through the FL material (in analogy to the radiation of electromagnetic waves via a transmitting antenna). On the contrary, the finite sizes of a nano-pillar device make it more similar to a magnetic resonator because of its reflecting boundaries. In fact, the spin-wave eigenmodes of such resonators, even though they could be spatially nonuniform, can have only a discrete frequency spectrum determined by the finite in-plane sizes of the device.

Nano-pillar

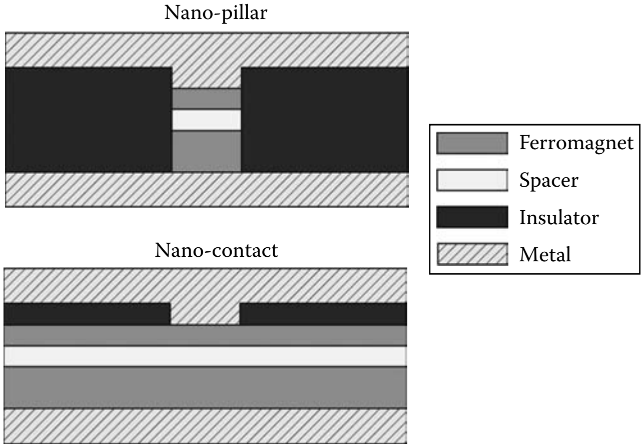

Nano-contact

FIGURE 8.12 Nano-pillar and nano-contact spin-valve geometries.

The structural differences between these two geometries imply not only different technological applications, but also require different approaches in the corresponding numerical investigations.

In fact, the micromagnetic study of the dynamics involved in a nano-pillar device can be carried out by directly applying the hints given in Sections 8.3.1 and 8.3.2 with the inclusion of normal boundary conditions at the computational edges (which also correspond to the physical edges, because of the finite dimensions):

$$\left.\frac{\partial \mathbf{m}}{\partial \mathbf{n}}\right|_{\partial\Omega} = 0 \tag{8.39}$$

($\mathbf{n}$ is the versor of the normal to the boundary $\partial\Omega$). We should notice that the normal derivative is different from zero only if surface anisotropy effects are taken into account.

On the other hand, the analysis of dynamics involved in nano-contact geometry requires a more sophisticated approach. In fact, one of the difficulties of FD methods lies in the fact that the investigated motion equation has to be solved in a discretized domain which, if applied to extended areas, would require prohibitive memory allocation and computational times. This is the case of nano-contact devices.

Nevertheless, theoretically boundless space extension problems have been solved by using special conditions at the boundaries of the reduced computational domain, in order to absorb the outgoing waves. Such a need for free-space simulation happens in a lot of physical problems and especially in wave-structure interactions.

Following the previous discussions, it is clear that FD techniques applied to extended geometries impose the numerical implementation of computational areas smaller than the physical ones. Therefore, it is necessary to develop a method for the waves absorption at the artificial boundaries to prevent the reflections of the propagating spin wave modes from these boundaries that, in a medium with relatively low dissipation (as Permalloy), might occur otherwise. If proper absorbing boundary conditions (ABC) are not considered, a spurious (purely computational) interference occurs which, in turn, can introduce a substantial distortion in the computed picture of the phenomenon and can seriously affect the properties of the excited spin-waves. As it will be shown below, due to the impossibility of implementing some sort of analytical operators within a micromagnetic framework for point-contact geometries, absorbing boundary conditions based on spatial-dependent damping functions have to be introduced.

The problem of finding the exact analytical formulation of the perfectly absorbing conditions for spin waves at the edges of the computational region has not been solved so far. The analytical formulation of ABC generally requires the knowledge of the exact wave solution (Enquist and Majda 1977, Renaut 1992, Alpert et al. 2002). In fact, the most general first-order analytical formulation for local ABC was introduced by Higdon (1987) as follows:

$$\left[\cos\theta\frac{\partial}{\partial t}-c\frac{\partial}{\partial x}\right]U=0 \tag{8.40}$$

where the Higdon operator is enclosed in brackets and it is used to completely absorb plane waves U propagating with speed c at the angles $\pm\theta$ with respect to the x axis. Unfortunately, from the numerical point of view, the lack of knowledge of the exact spin-wave velocity c does not allow the implementation of this formalism within a micromagnetic framework. Moreover, if we restrict the analysis to the case of perpendicular incidence (where $\cos\theta = 0$), the previous condition becomes the common boundary condition (see Equation 3.22) used in the modeling of the finite specimen $\partial U/\partial x = 0$, which cannot assure the waves absorption at each boundary.

Consequently, we need an alternative numerical strategy for the waves absorption at the computational boundaries. The attempts to find such conditions in numerical simulations have been undertaken by assuming that magnetic dissipation in the magnetic medium of the FL increases near the borders of the computational region according to a certain chosen empirical dependence on coordinates (Berkov and Gorn 2006, Consolo et al. 2007a,c). Following this idea, a micromagnetic approach could be generalized by including a site-dependent damping function in the model.

A possible scheme was first introduced by Berkov et al. in (Berkov and Gorn 2006) where the damping parameter was considered to vary between its physical value (α_O) and a higher artificial one ($\alpha_O + 2\Delta\alpha$) in a *smoothed* way, which starts at the distance from the center $r \approx (R_O - \sigma_\alpha)$ and involves a ring with a width of $\approx 2\sigma_\alpha$:

$$\alpha(r)=\alpha_O+\Delta\alpha\left(1+\tanh\frac{r-R_O}{\sigma_\alpha}\right) \tag{8.41}$$

The ABC introduced by Berkov are based on the following three main assumptions:

1. The dissipation within and nearby the contact area must be equal to its physical value.
2. Dissipation far from the contact area and close to the simulated boundaries must be large enough to ensure the wave energy absorption.
3. The spatial variation between these two values must be smooth enough to prevent wave reflection at the boundary.

Leaving the first two Berkov's assumptions unchanged, an alternative approach has been proposed in Consolo et al. (2007c) for perpendicular-to-plane magnetized films, and is based on the following requirements:

1. Creating a nonphysical absorbing medium surrounding the computational area
2. Considering a damping coefficient two orders of magnitude greater than the physical one just in the boundary cells
3. Entirely preserving the physical properties of the computational area

It has to be pointed out that, in spite of the formal equivalence between these empiric procedures and of the effectiveness of both procedures to absorb sufficiently well outgoing waves, the latter proposal, differently from the first one, allows the preservation of the physical properties in a larger computational area since it is based on the association of a localized *abrupt* change of the energy absorption at the computational boundaries

$$\begin{cases}\alpha(r)=\alpha_O & \text{for } r<R\\ \alpha(r)=m\cdot\alpha_O & \text{for } r=R\end{cases} \tag{8.42}$$

where

R is the radius of the computational area

m is the factor that amplifies the dissipation at the boundaries

Finally, as briefly introduced above, we need to mathematically formulate the difference existing in the spatial distribution of current for pillar and nano-contact devices, which is expressed by a proper choice of the dimensionless function $f(r)$ in Equation 8.15.

In particular, for pillar devices, by assuming a uniform current density distribution, it is quite straightforward to consider it as

$$\begin{cases}f(r)=1 & \text{for } r\le R\\ f(r)=0 & \text{for } r>R\end{cases} \tag{8.43}$$

which simply implies that all the computational cells are traversed by the same current value I.

On the other hand, for nanocontact geometry, it has to be taken into account that the current flow involves only the contact region of radius R_C so that, under the same condition of uniformity, we consider

$$\begin{cases}f(r)=1 & \text{for } r\le R_C\\ f(r)=0 & \text{for } r>R_C\end{cases} \tag{8.44}$$

which implies that only the computational cells within the contact area are traversed by the current I, whereas there exists an abrupt cut-off of the current outside the contact. While such an approximation is physically unrealistic, it has been demonstrated that it is able to capture most of the dynamics observed in laboratory experiments. However, more sophisticated shape functions can be derived by numerically solving the *Poisson equation* to get the corresponding current density distribution through the whole computational area.

8.4 Conclusions and Future Perspective

In this contribution, we presented a brief overview of a micromagnetic technique that is used to investigate the behavior of nanoscale spintronic devices. It is based on the numerical integration of the nonlinear LLGS equation. After developing a full-scale FD micromagnetic tool, a question might be asked. Is it possible to develop a framework simpler than the micromagnetic one but that allows for a gain in the equivalent understanding of the complex magnetization dynamics? Actually, magnetization dynamics might be most easily investigated using *macrospin* (less properly called "single-domain") approximation. The macrospin approximation assumes that the magnetization of a sample stays spatially uniform throughout its motion and can be treated as a single macroscopic spin. Since the spatial variation of the magnetization is frozen out, exploring the dynamics of magnetic systems is much more tractable using the macrospin approximation than it is using full micromagnetic simulations. The macrospin model makes it easy to explore the phase space of different torque models, and it has been a very useful tool for gaining a *zeroth-order* understanding of spin-torque physics. On the other hand, it obviously suffers from some intrinsic limitations (e.g., the impossibility of reproducing the nonuniform spatial distribution of magnetization and fields), which sometimes prevents the possibility of mimicking experimental data. It is possible to estimate qualitatively the critical size for which a macrospin approximation could be considered still sufficiently appropriate. It mainly depends on the competition between exchange and magnetostatic energy: the former favoring collinear (uniform) magnetic state, the latter favoring closed-flux configurations. As the exchange energy density of the closed-loop configuration obviously increases with decreasing particle size (because magnetization gradients become larger), it leads to the result that only a collinear magnetization state is energetically stable below a certain critical size. It could be demonstrated that such critical length-scale is about—four to eight times the exchange length (about 20–40 nm for usual soft magnetic materials).

On the other hand, micromagnetic frameworks offer a powerful tool for investigating magnetization phenomena occurring at the mesoscopic scale because of the accurate spatial resolution (few nanometers) and of the theoretically infinite bandwidth (restricted only by the integration time step). Because of that, it is also possible to gain additional information on the investigated problem, which could not be acquired from a laboratory experiment.

How can we summarize this issue between an accurate but time-consuming micromagnetic approach and a fast and often oversimplified macrospin model?

Let us recall that full-scale simulations are supposed to include more (and not fewer) known features of the investigated system than macrospin models and, at the same time, use fewer free (adjustable) parameters. Because of that, if a simplified macrospin approach were able to produce a better agreement with experiment than a micromagnetic approach, it would imply an incomplete comprehension of some crucial properties of the system under study. In other words, it does not necessarily mean that a wrong choice of the values of parameters has been considered, but rather it should recall for further studies to gain a better understanding of the problem, which in turn should bring some corrections to the model. Moreover, because full-scale simulations are quite time-consuming, it is necessary to get an accurate knowledge of the system parameters, together with their spatial dependence (especially near critical regions, such as close to the edges) from independent sources, such as a high-quality experiment. Such information is of crucial importance for the modeling of nano-scale devices due to their extremely small sizes.

What about present and future applications of micromagnetic framework?

So far, full-micromagnetic frameworks have been able to mimic the experimental behavior of most spintronic devices (pillar, nano-contact, MTJ, exchange-bias systems, phase-locked nano-contacts) as well as to perform predictions on new high-performance setups (Finocchio et al. 2006a,b, 2007, 2008, Choi et al. 2007, Consolo et al. 2007d, 2008, Hrkac 2008). At the same time, micromagnetic tools are successfully used to validate analytical theories as well. For example, the analysis of the disagreement reported between an analytical theory about the excitation of the nonlinear evanescent spin-wave "bullet" mode in in-plane magnetized nanocontact devices (Slavin and Tiberkevich 2005) and earlier results of micromagnetic simulations for the same geometry was intriguing and stimulating (Berkov and Gorn 2006). In fact, while the approximated theory was able to capture the underlying physics and consequently reproduce most of the experimental observations, micromagnetic simulations failed in that attempt. The problem was solved by using a new numerical strategy based on the application of decreasing currents starting from a large supercritical regime (Consolo et al. 2007b), which contains a lot of nonlinear "seeds." Apart from validating the theory, the micromagnetic approach has also been able to attribute the additional "subcritically-unstable" nature of these modes, enlarging the knowledge on the spin-wave modes supported by nano-contact devices.

From the experimental point of view, there is also an increasing interest in the dynamics (both switching and precession) involving magnetic tunnel junctions because of the discovery of very large TMR values (TMR is the difference in resistance between parallel and antiparallel orientation for the electrode magnetizations of a magnetic tunnel junction) at room temperature.

One reason for the interest in spin-torque effects in tunnel junctions is that these devices are better-suited than metallic magnetic multilayers for many types of applications. Tunnel junctions have higher resistances that can often be better impedance-matched to silicon-based electronics, and TMR values can now be made larger than the GMR values in metallic devices.

Devices based on those effects have already found very widespread applications as magnetic-field sensors in the read heads of magnetic hard disk drives as well as nonvolatile random access memory based on magnetic tunnel junctions.

Applications of spin transfer torques are so fascinating as well. Magnetic switching driven by the spin transfer effect can be much more efficient and more industrially intriguing than the usage of static magnetic fields or current-induced magnetic fields. This may enable the production of magnetic memory devices with much lower switching currents and hence greater energy efficiency as well as a larger integration density. The steady-state magnetic precession mode that can be excited by spin transfer is under investigation for a number of high-frequency applications, for example, nanometer-scale microwave sources (tunable by both magnetic field and current), detectors, mixers, modulators, phase shifters, and arrays (or matrixes) of phase-locked devices (whose coupling is used to increase the output power). One potential area of use is short-range chip-to-chip or even within-chip communications.

References

Alpert, B., Greengard, L., and Hagstrom T. 2002. Nonreflecting boundary conditions for the time-dependent wave equation. *J. Comput. Phys*. 180: 270–296.

Baibich, M.N. et al. 1988. Giant magnetoresistance of (001)Fe/(001) Cr magnetic superlattices. *Phys. Rev. Lett*. 61: 2472–2475.

Berger, L. 1996. Emission of spin waves by a magnetic multilayer traversed by a current. *Phys. Rev. B* 54: 9353–9358.

Berkov, D.V. and Gorn, N. 2006. Micromagnetic simulations of the magnetization precession induced by a spin-polarized current in a point-contact geometry. *J. Appl. Phys*. 99: 08Q701.

Berkov, D.V. and Miltat, J. 2008. Spin-torque driven magnetization dynamics: Micromagnetic modelling, *J. Magn. Magn. Mater*. 320: 1238–1259.

Bertotti, G. 1998. *Hysteresis in Magnetism*. Academic Press, Boston, MA.

Brown, W.F. 1940. Theory of the approach to magnetic saturation. *Phys. Rev*. 58: 736–743.

Brown, W.F. Jr. 1962. *Magnetostatic Principles in Ferromagnetism*, North-Holland Publishing Company, Amsterdam, the Netherlands.

Brown, W.F. 1963. *Micromagnetics*. Wiley, New York.

Brown, W.F. 1979. Thermal fluctuations in fine magnetic particles. *IEEE Trans. Magn*. MAG-15(5): 1196–1208.

Choi, S. et al. 2007. Double-contact spin-torque nano-oscillator with optimized spin-wave coupling: Micromagnetic modelling. *Appl. Phys. Lett*. 90: 083114.

Consolo, G. et al. 2007a. Boundary conditions for spin-wave absorption based on different site-dependent damping functions. *IEEE Trans. Magn*. 43: 2974–2976.

Consolo, G. et al. 2007b. Excitation of self-localized spin-wave bullets by spin-polarized current in in-plane magnetized magnetic nanocontacts: A micromagnetic study. *Phys. Rev. B* 76: 144410.

Consolo, G. et al. 2007c. Magnetization dynamics in nanocontact current controlled oscillators. *Phys. Rev. B* 75: 214428.

Consolo, G. et al. 2007d. Nanocontact spin-transfer oscillators based on perpendicular anisotropy in the free layer. *Appl. Phys. Lett*. 91: 162506.

Consolo, G. et al. 2008. Micromagnetic study of the above-threshold generation regime in a spin-torque oscillator based on a magnetic nano-contact magnetized at an arbitrary angle. *Phys. Rev. B* 78: 014420.

Enquist, B. and Majda, A. 1977. Absorbing boundary conditions for the numerical simulation of waves. *Math. Comput*. 31: 629–651.

Finocchio, G. et al. 2006a. Magnetization dynamics driven by the combined action of AC magnetic field and DC spin-polarized current. *J. Appl. Phys*. 99: 08G507.

Finocchio, G. et al. 2006b. Trends in spin-transfer driven magnetization dynamics of CoFe/AlO/Py and CoFe/MgO/Py magnetic tunnel junctions. *Appl. Phys. Lett*. 89: 262509.

Finocchio, G. et al. 2007. Magnetization reversal driven by spin-polarized current in exchange-biased nanoscale spin valves. *Phys. Rev. B* 76: 174408.

Finocchio, G. et al. 2008. Numerical study of the magnetization reversal driven by spin-polarized current in MgO based magnetic tunnel junctions. *Physica B* 403: 364–367.

Garcia-Palacios, J.L. and Lazaro, F.J. 1998. Langevin-dynamics study of the dynamical properties of small magnetic particles. *Phys. Rev. B* 58: 14937–14958.

Gardiner, C.W. 1985. *Handbook of Stochastic Methods*. Springer, Berlin, Germany.

Gilbert, T.L. 2004. A phenomenological theory of damping in ferromagnetic materials. *IEEE Trans. Magn*. 40: 3443–3449.

Higdon, R.L. 1987. Numerical absorbing boundary conditions for the wave equation. *Math. Comput*. 49: 65–90.

Hrkac, G. 2008. Mutual phase locking in high-frequency microwave nano-oscillators as a function of field angle. *J. Magn. Magn. Mater*. 320: L111–L115.

Kim, J.V., Tiberkevich, V., and Slavin, A. 2008. Generation linewidth of an auto-oscillator with a nonlinear frequency shift: Spin-torque nano-oscillator. *Phys. Rev. Lett*. 100: 017207.

Kloeden, P.E. and Platen E. 1999. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, Germany.

Krivorotov, I.N. et al. 2004. Temperature dependence of spin-transfer-induced switching of nanomagnets. *Phys. Rev. Lett*. 93: 166603.

Landau, L.D. and Lifshitz, E.M. 1935. On the theory of the dispersion of magnetic permeability in ferromagnetic bodies. *Phys. Z. Sowjetunion* 8: 153–169.

Li, Z. and Zhang, S. 2004. Thermally assisted magnetization reversal in the presence of a spin-transfer torque. *Phys. Rev. B* 69: 134416.

Martinez, E., Torres, L., and Lopez-Diaz, L. 2004. Computing solenoidal fields in micromagnetic simulations. *IEEE Trans. Magn*. 40: 3240–3243.

Newell, A.J., Williams, W., and Dunlop, D.J. 1993. A generalization of the demagnetizing tensor for nonuniform magnetization. *J. Geophys. Res*. 98: 9551–9555.

Parkin, S.S., More, N., and Roche, K.P. 1990. Oscillations in exchange coupling and magnetoresistance in metallic superlattice structures: Co/Ru, Co/Cr, and Fe/Cr. *Phys. Rev. Lett*. 64: 2304–2307.

Parkin, S.S., Bhadra, R., and Roche, K.P. 1991. Oscillatory magnetic exchange coupling through thin copper layers. *Phys. Rev. Lett*. 66: 2152–2155.

Preisach, F. 1929. Investigations on the Barkhausen effect. *Ann. Physik* 3: 737–799.

Ralph, D.C. and Stiles, M.D. 2008. Spin-transfer torques. *J. Magn. Magn. Mater*. 320: 1190–1216.

Renaut, R.A. 1992. Absorbing boundary conditions, difference operators and stability. *J. Comput. Phys*. 102: 236–251.

Rippard, W.H. et al. 2004. Direct-current induced dynamics in $Co_{90}Fe_{10}/Ni_{80}Fe_{20}$ point contacts. *Phys. Rev. Lett*. 92: 027201.

Schad, R. et al. 1994. Giant magnetoresistance in Fe/Cr superlattices with very thin Fe layers. *Appl. Phys. Lett*. 64: 3500–3502.

Slavin, A.N. and Kabos, P. 2005. Approximate theory of microwave generation in a current-driven magnetic nanocontact magnetized in an arbitrary direction. *IEEE Trans. Magn*. 41: 1264–1273.

Slavin, A. and Tiberkevich, V. 2005. Spin-wave mode excited by spin-polarized current in a magnetic nanocontact is a standing self-localized wave bullet. *Phys. Rev. Lett*. 95: 237201.

Slavin, A. and Tiberkevich, V. 2008. Excitation of spin waves by spin-polarized current in magnetic nano-structures. *IEEE Trans. Magn*. 44: 1916–1927.

Slonczewski, J.C. 1996. Current-driven excitation of magnetic multilayers. *J. Magn. Magn. Mater*. 159: L1–L7.

Slonczewski, J.C. 1999. Excitation of spin waves by an electric current. *J. Magn. Magn. Mater*. 195: L261–L268.

Stiles, M.D. and Miltat J. 2006. Spin-transfer torque and dynamics. In *Spin Dynamics in Confined Magnetic Structures III*. Eds. B. Hillebrands and A. Thiaville, pp. 225–308. Springer, Berlin/Heidelberg, Germany.

Van Kampen, N.G. 1987. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, the Netherlands.

9

Quantum Spin Tunneling in Molecular Nanomagnets

Gabriel González
University of Central Florida

Michael N. Leuenberger
University of Central Florida

9.1 Introduction **9**-1
9.2 Spin Tunneling in Molecular Nanomagnets **9**-2
9.3 Phonon-Assisted Spin Tunneling in Mn_{12} Acetate **9**-5
9.4 Interference between Spin Tunneling Paths in Molecular Nanomagnets **9**-5
Berry's Phase and Path Integrals
9.5 Incoherent Zener Tunneling in Fe_8 .... **9**-7
9.6 Coherent Néel Vector Tunneling in Antiferromagnetic Molecular Wheels **9**-8
9.7 Berry-Phase Blockade in Single-Molecule Magnet Transistors **9**-11
9.8 Concluding Remarks **9**-12
Acknowledgment **9**-13
References **9**-13

9.1 Introduction

Molecular magnets have attracted considerable interest recently among researchers because they are considered to be ideal systems to probe the interface between classical and quantum physics as well as to study decoherence in nanoscale systems. The advances in nanoscience during the last decade have made it possible to design and fabricate a wide variety of nanosize objects, ranging from a couple of micrometers all the way down to a few tens of nanometers, where quantum effects become important and give rise to new properties. One of the goals of miniaturization was the possibility of observing quantum tunneling effects in mesoscopic systems. One source for such a task is the so-called single-molecule magnets (SMM) and antiferromagnetic molecular wheels, in which the spin state of the molecule is known to behave quantum mechanically at low temperatures.

During the last decade, a tremendous progress in the experimental methods for contacting single molecules and measuring the electrical current through them has been achieved. The current through single magnetic molecules like Mn_{12} and Fe_8 has been measured and magnetic excited states have been identified (Jo et al., 2006; Henderson et al., 2007). In a three-terminal molecular single electron transistor, the current can flow between the source and drain leads via a sequential tunneling process through the molecular charge levels, thereby bringing the whole field of Coulomb-blockade physics to molecular systems.

Single-molecule magnets are mainly organic molecules containing multiple transition-metal ions bridged by organic ligands. These ions are strongly coupled by exchange interaction, yielding a large magnetic moment per molecule. The large spin combined with the large magnetic anisotropy provides an energy barrier for magnetization reversal. These systems provided for the first time evidence of quantum tunneling of the magnetization and interference effects as well as oscillations of the tunnel splitting.

Ferromagnetic molecular magnets such as Mn_{12} and Fe_8 show incoherent tunneling of the magnetization (Korenblit and Shender, 1978; Enz and Schilling, 1986; van Hemmen and Süto, 1986; Chudnovsky and Gunther, 1988) and allow one to study the interplay of thermally activated processes and quantum tunneling. The spin tunneling leads to two effects. First, the magnetization relaxation is accelerated whenever spin states of opposite direction become degenerate due to the variation of the external longitudinal magnetic field (Friedman et al., 1996; Thomas et al., 1996; Leuenberger and Loss, 1999, 2000a; Leuenberger and Loss). Second, the spin acquires a Berry phase during the tunneling process, which leads to oscillations of the tunnel splitting as a function of the external transverse magnetic field (Wernsdorfer and Sessoli, 1996; Wernsdorfer et al., 2000; Leuenberger and Loss, 2000b, 2001a).

Antiferromagnetic molecular wheels are another type of molecular magnets where an even number of transition metal ions with spin form a closed ring in which the exchange interaction between neighbors gives rise to a strong spin quantum dynamics. Antiferromagnetic molecular magnets such as ferric wheels belong to the most promising candidates for the observation of coherent quantum tunneling on the mesoscopic scale (Chiolero and Loss, 1998; Meier and Loss, 2000, 2001). In contrast to incoherent tunneling, in quantum coherent tunneling spins tunnel back and forth between energetically degenerate configurations at a tunneling rate which is *large* compared to the decoherence rate. The detection of quantum behavior is more challenging in antiferromagnetic molecular magnets than in ferromagnetic systems, but is feasible with present day experimental techniques.

Understanding the properties of molecular magnets is only a first step toward achieving technological applications in data storage, data processing, and quantum technologies. A possible next step will be the preparation and control of a well-defined single-spin quantum state of a molecular cluster. Although challenging, this task appears feasible with present-day experiments and would allow one to carry out quantum computing with molecular magnets (Leuenberger and Loss, 2001b).

9.2 Spin Tunneling in Molecular Nanomagnets

An effective spin Hamiltonian that captures the basic physics of a molecular nanomagnet is given by the following form:

$$\mathcal{H} = \mathcal{H}_0(\vec{S}) + \mathcal{H}_Z + \mathcal{H}_T + \mathcal{H}_{sb}. \tag{9.1}$$

We will now analyze each term in the Hamiltonian given in Equation 9.1 separately. The first term in Equation 9.1 is the dominant term of the Hamiltonian with spin quantum number $S = |\vec{S}| \gg 1$ and is called the uniaxial anisotropy, which results from the spin–orbit interactions and generates and easy magnetic axis for the magnetic moment of the nanomagnet. The assumption is that we have a single domain magnetic molecule in which the very strong exchange interactions align the microscopic electronic spins into a parallel or antiparallel configuration, resulting in a total spin $\vec{S} = \sum_j \vec{s}_j$ for the ferromagnetic moment or the Neél vector with $\vec{S} = \sum_j (-1)^j \vec{s}_j$, respectively.

For a nanomagnet like Fe$_8$, $\mathcal{H}_0(\vec{S}) = -AS_z^2$ where $A/k_B \approx 0.275\,\text{K}$ and generates an energy barrier that separates opposite spin projections (see Figure 9.1).

Quantum mechanics specifies that a particle in a symmetric double well potential undergoes a tunnel effect that makes the particle go back and forth from one well to the other with a frequency ω_T. This implies that the eigenfunctions of the Hamiltonian are delocalized, which means that there is a probability to find the particle either in one well or the other. In a similar fashion, there is a probability for the spin of the nanomagnet to change direction, i.e., if the spin of the nanomagnet points up there is a probability that at a later time the spin will be pointing down. At equilibrium and finite temperature, the probability that the spin is in state $|m\rangle$, where $S_z|m\rangle = m|m\rangle$ for $m = -s, -s+1, \ldots, s-1, s$, is given by

$$P_m = \frac{e^{-E_m/k_BT}}{Z} \tag{9.2}$$

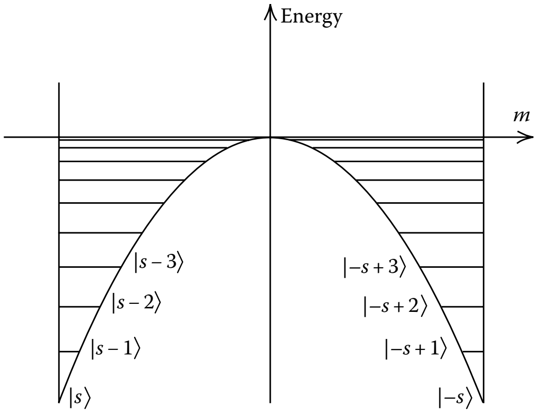

FIGURE 9.1 The graph shows $\mathcal{H}_0$ vs. S_z where the horizontal lines indicate the energies of the quantum states.

where
Z is the partition function
$E_m = -Dm^2$

For the case when $T \to 0$ the probability for the spin to go from one well to the other satisfies the following equation:

$$\frac{dP(t)}{dt} \approx -\Gamma P(t) < 0. \tag{9.3}$$

Therefore $P(t) \propto e^{-\Gamma t}$, where Γ is the rate of decay and Γ^{-1} is known as the lifetime or characteristic time of the particle.

The second term in Equation 9.1 is the Zeeman energy resulting from the interaction of the spin with an external magnetic field, i.e., $\mathcal{H}_Z = \mu_B g\vec{S}\cdot\vec{H}$. A magnetic field applied along the easy axis of the molecule will tilt the potential well and for certain values of the magnetic field ($\delta H_z = (m + m')|A|/g\mu_B$), the energy levels on both sides of the anisotropy barrier coincide. Therefore the Hamiltonian given by $\mathcal{H} = -AS_z^2 + g\mu_B\delta H_z S_z$ is still doubly degenerate (see Figure 9.2).

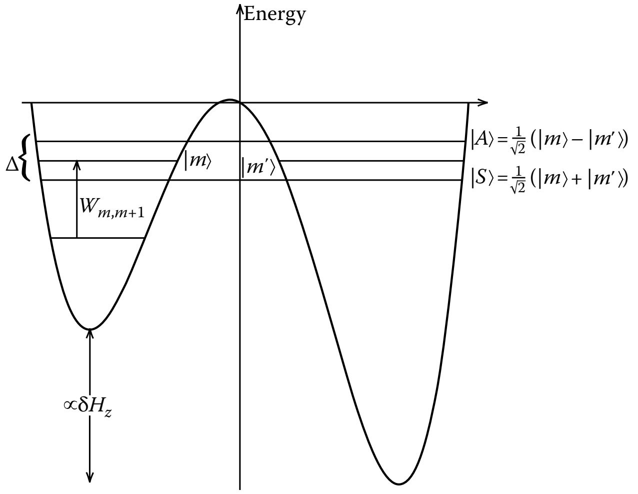
Energy
Δ
|m⟩
|m′⟩
$W_{m,m+1}$
$|A\rangle = \frac{1}{\sqrt{2}}(|m\rangle - |m'\rangle)$
$|S\rangle = \frac{1}{\sqrt{2}}(|m\rangle + |m'\rangle)$
$\propto\delta H_z$

FIGURE 9.2 The graph shows $(\mathcal{H}_0 + \mathcal{H}_Z)$ vs. S_z in a constant magnetic field parallel to the easy axis. Δ denotes the tunnel splitting between states $|m\rangle$ and $|m'\rangle$.

These degeneracies can be removed by the term $\mathcal{H}_T$, which corresponds to the transverse anisotropies. For the nanomagnet Fe_8, the transverse anisotropy term is given by $\mathcal{H}_T = E(S_x^2 - S_y^2)$, where $(E/k_B \approx 0.046\,\text{K})$. The anisotropy term lifts the degeneracy, thereby creating an energy gap between the spin states $|m\rangle$ and $|m'\rangle$, denoted by $\Delta_{mm'}$. If the Hamiltonian possesses transverse terms, the pairwise degenerate states get split by $\Delta_{mm'}$ amount of energy. If the coupling of our system to an external bath is assumed to be zero, each pair of the degenerate states can be described by means of a two-state Hamiltonian:

$$\mathcal{H} = \begin{bmatrix} hm & \dfrac{\Delta_{mm'}}{2} \\ \dfrac{\Delta_{mm'}}{2} & hm' \end{bmatrix} = \frac{h}{2}(m+m')\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{h}{2}(m-m')\begin{bmatrix} 1 & \tan\theta \\ \tan\theta & -1 \end{bmatrix}, \tag{9.4}$$

where $\tan\theta = \Delta_{mm'}/h(m-m')$, $0 \le \theta < \pi$ and h is the longitudinal magnetic field with the coupling constant g and Bohr magneton μ_B absorbed, which makes the double well asymmetric, and $\Delta_{mm'}$ is the tunnel splitting between the states $|m\rangle$ and $|m'\rangle$. The eigenstates of Equation 9.4 read

$$|\psi_+\rangle = \cos\frac{\theta}{2}|m\rangle + \sin\frac{\theta}{2}|m'\rangle, \tag{9.5}$$

$$|\psi_-\rangle = -\sin\frac{\theta}{2}|m\rangle + \cos\frac{\theta}{2}|m'\rangle, \tag{9.6}$$

with eigenvalues

$$E_\pm = \frac{h}{2}\left[(m+m') \pm \frac{(m-m')}{\cos\theta}\right] = \frac{1}{2}\left[h(m+m') \pm \sqrt{h^2(m-m')^2 + \Delta_{mm'}^2}\right]. \tag{9.7}$$

$|\psi_+\rangle$ and $|\psi_-\rangle$ are delocalized as long as $h(m - m') \le \Delta_{mm'}$. It is only in this regime that a spin state can tunnel from one side of the barrier to the other, which is also relevant for the Zener tunneling (see Section 9.5). In the limit $h \to 0$, the eigenstates $|\psi_+\rangle$ and $|\psi_-\rangle$ are the symmetrical and antisymmetrical combinations of $|m\rangle$ and $|m'\rangle$, i.e., $|\psi_+\rangle = |S\rangle = \frac{1}{\sqrt{2}}\left(|m\rangle + |m'\rangle\right)$ and $|\psi_-\rangle = |A\rangle = \frac{1}{\sqrt{2}}\left(|m\rangle - |m'\rangle\right)$. For $\Delta_{mm'} \to 0$ we are left with $|\psi_+\rangle = |m\rangle$ and $|\psi_-\rangle = |m'\rangle$ for $h > 0$ and $|\psi_-\rangle = |m\rangle$ and $|\psi_+\rangle = |m'\rangle$ for $h < 0$ (Figure 9.3).

Let us now assume that the system starts in the state

$$|\psi(t=0)\rangle = |m\rangle = \cos\frac{\theta}{2}|\psi_+\rangle - \sin\frac{\theta}{2}|\psi_-\rangle, \tag{9.8}$$

from which one obtains immediately the time evolution

$$|\psi(t)\rangle = \cos\frac{\theta}{2}e^{-iE_+t/\hbar}|\psi_+\rangle - \sin\frac{\theta}{2}e^{-iE_-t/\hbar}|\psi_-\rangle. \tag{9.9}$$

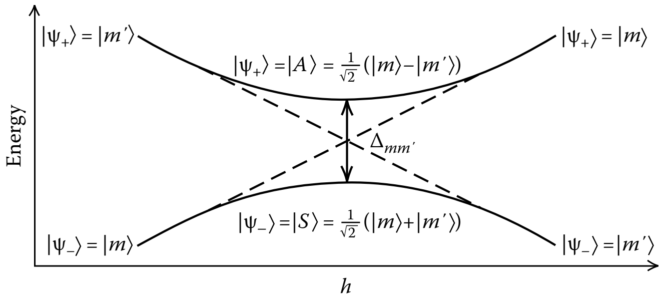
$|\psi_+\rangle = |m'\rangle$
$|\psi_+\rangle = |m\rangle$
$|\psi_+\rangle = |A\rangle = \frac{1}{\sqrt{2}}(|m\rangle - |m'\rangle)$
Energy
$\Delta_{mm'}$
$|\psi_-\rangle = |S\rangle = \frac{1}{\sqrt{2}}(|m\rangle + |m'\rangle)$
$|\psi_-\rangle = |m\rangle$
$|\psi_-\rangle = |m'\rangle$
h

FIGURE 9.3 Anticrossing of the adiabatic eigenvalues $E_\pm = \frac{1}{2}\left[h(m+m') \pm \sqrt{h^2(m-m') + \Delta_{mm'}^2}\right]$. The Zener tunneling transition of the state $|\psi_+\rangle$ is adiabatic (see Section 9.5).

Then one can calculate the probability that the state $|m\rangle$ has tunneled to the state $|m'\rangle$ after time t:

$$\begin{aligned} |\langle m' \mid \psi(t)\rangle|^2 &= \left|\sin\frac{\theta}{2}\langle\psi_+ \mid \psi(t)\rangle + \cos\frac{\theta}{2}\langle\psi_- \mid \psi(t)\rangle\right|^2 \\ &= \left|\sin\frac{\theta}{2}\cos\frac{\theta}{2}(e^{-iE_+t/\hbar} - e^{-iE_-t/\hbar})\right|^2 \\ &= \sin^2\theta\sin^2\left(\frac{\sqrt{h^2(m-m')^2 + \Delta_{mm'}^2}}{2\hbar}t\right) \\ &= \frac{\Delta_{mm'}^2}{h^2(m-m')^2 + \Delta_{mm'}^2}\sin^2\left(\frac{\sqrt{h^2(m-m')^2 + \Delta_{mm'}^2}}{2\hbar}t\right) \end{aligned} \tag{9.10}$$

This means that both the tunneling behavior and also the two-level spin resonance behavior are described by a Rabi oscillation.

In reality, the spin of the SMM is interacting with the environment and this interaction is described by the last term in Equation 9.1. The interaction of the molecular magnet with the environment is due to the hyperfine interaction, dipole interaction, spin–phonon interaction, the interaction between the molecule and two contact leads, etc. When the spin interacts with its environment, the wavefunction loses the memory of its phase. This phenomenon is known as decoherence. In this case, the wavefunction is not longer appropriate to describe the system. The problem of a system that interacts with the environment is formulated by means of the density matrix formalism. The spin has a certain probability to be in state $|m\rangle$ at time t, which is described by the density matrix $\rho_m(t)$. Due to the interaction with the environment, the spin undergoes incoherent transitions from a state $|m\rangle$ to another state $|m'\rangle$ at a rate $W_{m'm}$ until $\rho_m(t)$, $-s \le m \le s$ reach their equilibrium value. This process is known as relaxation. If these transitions are independent of each other, the density matrix evolves according to the Pauli equation or master equation

$$\frac{d\rho_m(t)}{dt} = \sum_{m'}\left[W_{mm'}\rho_{m'}(t) - W_{m'm}\rho_m(t)\right] \tag{9.11}$$

where $W_{m'm}$ is the transition probability from state $|m\rangle$ to state $|m'\rangle$. The master equation (Equation 9.11) can also be written as

$$\frac{d\rho_m(t)}{dt} = \sum_{m'} A_{mm'}\rho_{m'}(t) \tag{9.12}$$

where

$$A_{mm'} = \begin{cases} W_{mm'} & \text{if } m \neq m' \\ -\sum_{m''}{}' W_{m''m'} & \text{if } m' = m, \end{cases} \tag{9.13}$$

the prime indicates that the term $m'' = m$ is to be omitted in the summation. A formal solution to Equation 9.11 can be obtained in the following manner, let

$$\rho_m(t) = p_m^{(n)} e^{-\Gamma_n t} \tag{9.14}$$

where Γ_n^{-1} is the characteristic time labeled by the index $n = 0, 1, \ldots, 2s$. Substituting Equation 9.14 into Equation 9.11, one obtains

$$-\Gamma_n p_m^{(n)} = \sum_{m'}\left[W_{mm'} - \delta_{mm'}\sum_{m''} W_{mm''}\right] p_{m'}^{(\mathrm{n})} = \sum_{m'} A_{mm'} p_{m'}^{(\mathrm{n})} \tag{9.15}$$

which implies that $-\Gamma_n$ is an eigenvalue of the master matrix defined in Equation 9.13. The master matrix is a square matrix $(2s + 1) \times (2s + 1)$, which is not Hermitian. However, the eigenvalue problem can be written in terms of the Hermitian real matrix by making the following transformation

$$B_{mm'} = A_{mm'} \exp\left[\frac{\beta(E_m - E_{m'})}{2}\right] = B_{m'm}. \tag{9.16}$$

It follows that the general solution of Equation 9.11 is then

$$\rho_m(t) = \exp\left[\frac{-\beta E_m}{2}\right]\sum_n c_n \mathcal{P}_m^{(n)} e^{-\Gamma_n t} \tag{9.17}$$

where $(\mathcal{P}_m^{(n)}, -\Gamma_n)$ are the $(2s + 1)$ orthonormal set of eigenvectors with the corresponding eigenvalues of the matrix (9.16) and $\mathcal{P}_m^{(n)} = p_m^{(n)} \exp(\beta E_m/2)$. The c_n are the constants related to the initial distribution of ρ_m by

$$c_n = \sum_k \rho_k(0)\mathcal{P}_k^{(n)} \exp\left[\frac{\beta E_k}{2}\right]. \tag{9.18}$$

Equation 9.11 is only valid for times where the correlation time τ_c in the heat bath is much smaller than the relaxation time of the spin system, necessary for the establishment of irreversibility in a macroscopic system. In fact, in order to describe quantum tunneling, it is necessary to use the *generalized master equation*. The generalized master equation that describes the relaxation of the spin due to phonon-assisted transitions including resonances due to tunneling is given by

$$\dot\rho_{mm'} = \frac{i}{\hbar}[\rho, \mathcal{H}_0]_{mm'} + \delta_{mm'}\sum_{n \neq m}\rho_n W_{mn} - \gamma_{mm'}\rho_{mm'}. \tag{9.19}$$

The difference to the usual master equation is that Equation 9.19 takes also off-diagonal elements of the density matrix $\rho(t)$ into account. This is essential to describe tunneling of the magnetization, which is caused by the overlap of the S_z states.

Let us consider the two-state system $\{|m\rangle, |m'\rangle\}$, which yields the two-state Hamiltonian in the presence of a bias field given in Equation 9.4. Next we insert the two-state Hamiltonian into the generalized master equation (Equation 9.19), which yields for the diagonal elements of the density matrix

$$\dot\rho_m = \frac{i\Delta_{mm'}}{2\hbar}(\rho_{mm'} - \rho_{m'm}) - W_m\rho_m + \sum_{n \neq m, m'} W_{mn}\rho_n \tag{9.20}$$

and for the off-diagonal elements

$$\dot\rho_{mm'} = -\left(\frac{i}{\hbar}\xi_{mm'} + \gamma_{mm'}\right)\rho_{mm'} + \frac{i\Delta_{mm'}}{2\hbar}(\rho_m - \rho_{m'}), \tag{9.21}$$

with $\xi_m = -Am^2 + g\mu_B\delta H_z m$ and $\xi_{mm'} = \xi_m - \xi_{m'}$, similarly for $m \leftrightarrow m'$. Ultimately, we are interested in the overall relaxation time τ of the quantity $\rho_s - \rho_{-s}$ (see Section 9.7) due to phonon-induced transitions. This τ turns out to be much longer than $\tau_\mathrm{d} = 1/\gamma_{mm'}$, which is the decoherence time of the decay of the off-diagonal elements $\rho_{mm'} \propto e^{-t/\tau_\mathrm{d}}$ of the density matrix ρ. Thus, we can neglect the time dependence of the off-diagonal elements, i.e., $\dot\rho_{mm'} \approx 0$. Physically this means that we deal with incoherent tunneling for times $t > \tau_\mathrm{d}$. Inserting then the stationary solution of Equation 9.21 into Equation 9.20, which leads to the complete master equation including resonant as well as nonresonant levels, we get

$$\dot\rho_m = -W_m\rho_m + \sum_{n \neq m, m'} W_{mn}\rho_n + \Gamma_{mm'}(\rho_{m'} - \rho_m), \tag{9.22}$$

where

$$\Gamma_{mm'} = \Delta_{mm'}^2 \frac{W_m + W_{m'}}{4\xi_{mm'}^2 + \hbar^2(W_m + W_{m'})^2} \tag{9.23}$$

is the transition rate from m to m' (induced by tunneling) in the presence of phonon damping. Note that Equation 9.22 is now of the usual form of a master equation, i.e., only diagonal elements of the density matrix $\rho(t)$ occur. For levels $k \neq m, m'$, Equation 9.22 reduces to

$$\dot\rho_k = -W_k\rho_k + \sum_n W_{kn}\rho_n. \tag{9.24}$$

We note that $\Gamma_{mm'}$ has a Lorentzian shape with respect to the external magnetic field δH_z occurring in $\xi_{mm'}$. It is thus this $\Gamma_{mm'}$ that will determine the peak shape of the magnetization resonances.

9.3 Phonon-Assisted Spin Tunneling in Mn_{12} Acetate

The magnetization relaxation of crystals and powders made of molecular magnets Mn_{12} has attracted much recent interest since several experiments (Sessoli et al., 1993; Novak and Sessoli, 1995; Novak et al., 1995; Paulsen and Park, 1995; Paulsen et al., 1995) have indicated unusually long relaxation times as well as increased relaxation rates (Friedman et al., 1996; Hernández et al., 1996; Thomas et al., 1996) whenever two spin states become degenerate in response to a varying longitudinal magnetic field H_z. According to earlier suggestions (Barbara et al., 1995), this phenomenon has been interpreted as a manifestation of incoherent macroscopic quantum tunneling (MQT) of the spin.

As long as the external magnetic field H_z is much smaller than the internal exchange interactions between the Mn ions of the Mn_{12} cluster, the Mn_{12} cluster behaves like a large single spin $\vec{S}$ of length $|\vec{S}| = 10$. For temperatures $1 \le T \le 10\,\mathrm{K}$, its spin dynamics can be described by the spin Hamiltonian given in Equation 9.1, including the coupling between this large spin and the phonons in the crystal (Villain et al., 1994; Garanin and Chudnovsky, 1997; Hernández et al., 1997; Fort et al., 1998; Luis et al., 1998; Leuenberger and Loss, 1999, 2000a; Leuenberger and Loss).

The most general spin–phonon coupling reads

$$\begin{aligned}\mathcal{H}_{\mathrm{sp}} &= g_1(\epsilon_{xx} - \epsilon_{yy}) \otimes (S_x^2 - S_y^2) + \frac{1}{2} g_2 \epsilon_{xy} \otimes \{S_x, S_y\} \\ &\quad + \frac{1}{2} g_3(\epsilon_{xz} \otimes \{S_x, S_z\} + \epsilon_{yz} \otimes \{S_y, S_z\}) \\ &\quad + \frac{1}{2} g_4(\omega_{xz} \otimes \{S_x, S_z\} + \omega_{yz} \otimes \{S_y, S_z\}), \end{aligned} \tag{9.25}$$

where

g_i are the spin–phonon coupling constants

$\epsilon_{\alpha\beta}(\omega_{\alpha\beta})$ is the (anti-)symmetric part of the strain tensor

From the comparison between experimental data (Friedman et al., 1996; Thomas et al., 1996) and calculation, it turns out that the constants $g_i \approx A\ \forall i$ (Leuenberger and Loss, 1999, 2000a; Leuenberger and Loss).

9.4 Interference between Spin Tunneling Paths in Molecular Nanomagnets

Recent experiments have pointed out the importance of the interference between spin tunneling paths in molecules. For instance, measurements of the magnetization in bulk Fe_8 have observed oscillations in the tunnel splitting $\Delta_{m,-m}$ between states $S_z = m$ and $-m$ as a function of a transverse magnetic field at temperatures between 0.05 and 0.7 K (see Landau, 1932). This effect can be explained by the interference between Berry phases associated to spin tunneling paths of opposite windings (DiVincenzo et al., 1992; van Delft and Henley, 1992).

Usually the Berry phase effect arises in systems that undergo an adiabatic cyclic evolution, and which occurs in such diverse fields as atomic, condensed matter, nuclear and elementary particle physics, and optics. The basic assumption for the derivation of the Berry phase is that, for a slowly varying time-dependent Hamiltonian $H(\vec{R}(t))$ that depends on parameters $R_1(t), R_2(t), \ldots, R_N(t)$ components of a vector $\vec{R}$, the eigenstate evolves in time according to

$$\left|\psi_i(\vec{R}(0))\right\rangle \to e^{i\varphi_i(t)} \left|\psi_i(\vec{R}(t))\right\rangle \exp\left[-\frac{i}{\hbar}\int_0^t E_i(\vec{R})dt\right]. \tag{9.26}$$

Plugging Equation 9.26 into Schrödinger's equation we have

$$\frac{d\varphi}{dt} = i\left\langle \psi_i \,|\, \nabla_{\vec{R}} \,|\, \psi_i \right\rangle \cdot \frac{d\vec{R}}{dt}, \tag{9.27}$$

thus, for $t = 0$ to $t = \tau$ we end up with

$$\varphi(\tau) - \varphi(0) = \int_{\vec{R}(0)}^{\vec{R}(\tau)} \left\langle \psi_i \,|\, i\nabla_{\vec{R}} \,|\, \psi_i \right\rangle d\vec{R}. \tag{9.28}$$

The term of Equation 9.28 is known as Berry's phase. The overall phase of a quantum state is not observable but the relative phases for a quantum system, which undergoes a coherent evolution, can be detected experimentally. Consider for example two paths $\vec{R}$ and $\vec{R}'$ with the same end points $\vec{R}(0) = \vec{R}'(0)$ and $\vec{R}(\tau) = \vec{R}'(\tau)$, the net Berry phase change is given by

$$\Delta\varphi = \oint \left\langle \psi_i \,|\, i\nabla_{\vec{R}} \,|\, \psi_i \right\rangle \cdot d\vec{R}. \tag{9.29}$$

This is a line integral around a closed loop in parameter space, which is nonzero in general. The classic example of Berry's phase is an electron at rest subjected to a time-dependent magnetic field of constant magnitude but changing direction. The Hamiltonian for this system is given by

$$\mathcal{H} = -\mu_B \vec{\sigma} \cdot \vec{H} = -\mu_B \begin{bmatrix} H_z & H_x - iH_y \\ H_x + iH_y & -H_z \end{bmatrix}, \tag{9.30}$$

where $\vec{H}(t) = H(\sin\theta(t)\cos(\phi(t)), \sin(\theta(t))\sin(\phi(t)), \cos(\theta(t)))$ and μ_B is Bohr magneton. It is easy to show that the eigenstate representing spin up along $\vec{H}(t)$ for Equation 9.30 has the form

$$\left|\uparrow\right\rangle = \begin{pmatrix} e^{-i\phi/2}\cos(\theta/2) \\ e^{i\phi/2}\sin(\theta/2) \end{pmatrix}. \tag{9.31}$$

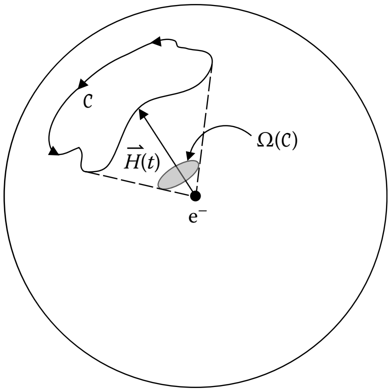

FIGURE 9.4 The figure shows an electron in a constant magnetic field which sweeps around a closed curve $\mathcal{C}$ and subtends a solid angle Ω.

For the case when the magnetic field sweeps out an arbitrary closed circuit $\mathcal{C}$ (see Figure 9.4) we can apply Equation 9.29 to get the Berry phase

$$\Delta\varphi = -\frac{1}{2}\Omega(\mathcal{C}), \tag{9.32}$$

where $\Omega(\mathcal{C})$ is the solid angle described by the field $\vec{H}$ along the circuit $\mathcal{C}$.

9.4.1 Berry's Phase and Path Integrals

An alternative approach for quantum mechanics and hence for the calculation of the tunneling frequency is by means of path integrals which were introduced by Richard Feynman (Feynman, 1948). In this section, we will briefly explain the path integral approach to calculate the Berry phase in single molecular magnets.

Feynman showed that the probability amplitude $\langle x' | \exp[i(t'-t)\mathcal{H}/\hbar] | x\rangle$ for a particle that is at point x' at time t' if it was at point x at a time t can be expressed as the path integral

$$U(x',t' \,|\, x,t) = \int_x^{x'} \mathcal{D}[x'']\exp\left[\frac{iS_E}{\hbar}\right], \tag{9.33}$$

where

$$S_E = \int_t^{t'} dt\, L[x(t),\dot{x}(t)], \tag{9.34}$$

is the *Euclidean action*, which involves the classical Lagrangian $L[x(t), \dot{x}(t)]$ and $\int\mathcal{D}[x'']$ is an integral over all paths from x to x' defined as

$$\int_x^{x'} \mathcal{D}[x''] \equiv \lim_{N\to\infty}\left(\frac{m}{2\pi i\hbar\Delta t}\right)^{N/2}\prod_{n=1}^{N-1}\int dx_n, \tag{9.35}$$

where $x_n = x(t_n)$ and $t_n \in [t, t_1, \ldots, t_{N-1}, t']$. For purposes of actual evaluation of quantities using path-integral methods, it is common to get rid of the factor i in the exponential of Equation 9.33 by analytic continuation to imaginary times, through $\beta = it/\hbar$. Then, the path integral becomes equivalent to the partition function Z by noting $\beta = 1/k_BT$.

Quantum tunneling of a spin is often described with the terminology of a path integral. Consider a general single-spin Hamiltonian $\mathcal{H}_{z,n} = -AS_z^2 + B_n(S_+^n + S_-^n)$, with easy-axis ($A$) and transverse ($B_n$) anisotropy constants satisfying $A \gg B_n > 0$. Here, n is an even integer, i.e., $n = 2, 4, 6, \ldots$, so that $\mathcal{H}_z$ is invariant under time reversal. Such Hamiltonians are relevant for molecular magnets such as Mn_{12} (Fort et al., 1998; Leuenberger and Loss, 1999, 2000a) and Fe_8 (Leuenberger and Loss, 2000b). The corresponding classical anisotropy energy $E_{z,n}(\theta, \phi) = -As^2\cos^2\theta + B_ns^n\sin^n\theta\cos(n\phi)$ has the shape of a double-well potential, with the easy axis pointing along the z direction. It is obvious that the anisotropy energy remains invariant under rotations around the z-axis by multiples of the angle $\eta = 2\pi/n$. For the following calculation, it proves favorable to choose the easy axis along the y direction where the contour path does not include the south pole. Then the spin Hamiltonian and the anisotropy energy are changed into

$$\mathcal{H}_{y,n} = -AS_y^2 + B_n(S_+^n + S_-^n), \tag{9.36}$$

where now $S_\pm = S_z \pm iS_x$, and

$$E_{y,n}(\theta,\phi) = -As^2\sin^2\theta\sin^2\phi + B_ns^n \times\left[(\cos\theta + i\sin\theta\cos\phi)^n + (\cos\theta - i\sin\theta\cos\phi)^n\right]. \tag{9.37}$$

We are interested in the tunneling between the eigenstates $|m\rangle$ and $|-m\rangle$ of $-AS_y^2$, corresponding to the global minimum points ($\theta = \pi/2$, $\phi = -\pi/2$) and ($\theta = \pi/2$, $\phi = +\pi/2$) of $E_{y,n}$. For this, we evaluate the imaginary time transition amplitude between these points. For the ground-state tunneling ($m = s \gg 1$), this can be done by means of the coherent spin-state path integral (DiVincenzo et al., 1992)

$$\left\langle -\frac{\pi}{2}\middle| e^{-\beta\mathcal{H}} \middle| +\frac{\pi}{2}\right\rangle = \int_{\pi/2,-\pi/2}^{\pi/2,+\pi/2} \mathcal{D}\Omega\, e^{-S_E}, \tag{9.38}$$

where $\beta = 1/k_BT$ is the inverse temperature, $\mathcal{D}\Omega = \Pi_\tau d\Omega_\tau$, $d\Omega_\tau = [4\pi/(2s+1)]d(\cos\theta_\tau)d\phi_\tau$ the Haar measure of the S^2 sphere, and $S_E = \int_0^\beta d\tau[is\dot{\phi}(1-\cos\theta) + E_{y,n}]$ the Euclidean action, where the first term in S_E defines the Wess–Zumino (or Berry phase) term, which gives rise to topological interference effects for spin tunneling (DiVincenzo et al., 1992; von Delft and Henley, 1992). We can divide the S^2 surface area of integration in Equation 9.38 into n equally shaped subareas $\mathcal{A}_n$, so that we can factor out the following sum over Berry phase terms without the need of

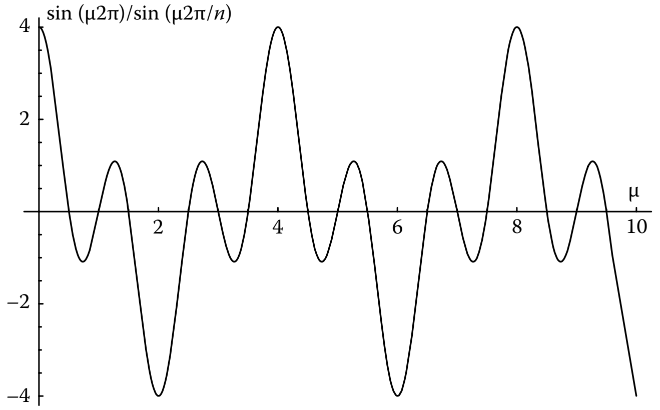

FIGURE 9.5 Sum over Berry phase terms for $s = 10$ and $n = 4$.

evaluating the dynamical part of the path integral (i.e., the following result is valid for all m),

$$\left\langle -\frac{\pi}{2}\right| e^{-\beta\mathcal{H}_{y,n}} \left|+\frac{\pi}{2}\right\rangle \propto \sum_{k=1}^{n} e^{i(2\pi/n)(2k-1)m} = \frac{\sin(2\pi m)}{\sin(2\pi m/n)}, \qquad (9.39)$$

which vanishes whenever n is not a divisor of $2m$. Thus, the tunnel splitting energy $\Delta_{m,-m}$ between the states $|m\rangle$ and $|-m\rangle$ vanishes if $2m/n \notin Z$. However, if $2m/n \in Z$, the variable m must be extended to real numbers, i.e., $m \rightarrow \mu \in R$, in order to calculate the limit $\mu \rightarrow m$ of the ratio of the sine functions, which is plotted in Figure 9.5 for a special case. In order to visualize the interference between the Berry phases in Equation 9.39, we select one representative path of each subarea $\mathcal{A}_n$. Then the vanishing of the amplitude in Equation 9.39 for the case $n = 6$ (see Figure 9.6) can be thought of as a destructive interference between six different paths. Also, it is important to note that from the semiclassical point of view, the total classical energy of the spin system E must be conserved during the tunneling process. Therefore, the semiclassical paths do not follow the local minima of E with respect to θ alone, except for the case of purely quadratic (i.e., $n = 2$) anisotropies (DiVincenzo et al., 1992; Garg, 1993).

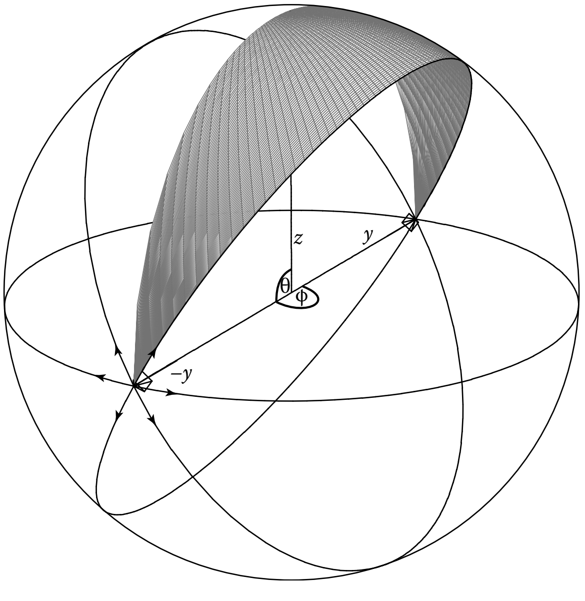

FIGURE 9.6 Berry phase interference between six different paths for $s = 10$ and $n = 6$.

From the above treatment, we conclude that tunneling between two degenerate spin states $|m\rangle$ and $|m'\rangle = |-m\rangle$ is topologically suppressed (i.e., the twofold degeneracy is not lifted) whenever n is *not* a divisor of $2m$. Since n is even this excludes immediately tunneling for all half-odd integer spins s (for all m and n), in accordance with Kramers degeneracy. For s integer, however, tunneling can be either allowed or suppressed, depending on the ratio $2m/n$. In the latter case, the twofold degeneracy of spin states is not lifted by the anisotropy, and we can view this result as a generalization of the Kramers theorem to integer spins.

9.5 Incoherent Zener Tunneling in Fe$_8$

Besides Mn$_{12}$, there have been several experiments on the molecular magnet Fe$_8$ that revealed macroscopic quantum tunneling of the spin (Barra et al., 1996; Wernsdorfer and Sessoli, 1996; Sangregorio et al., 1997; Ohm et al., 1998; Wernsdorfer et al., 2000). In particular, recent measurements on Fe$_8$ (Wernsdorfer and Sessoli, 1996; Wernsdorfer et al., 2000) lead to the development of the concept of the incoherent Zener tunneling (Leuenberger and Loss, 2000b; Leuenberger et al., 2003). The resulting Zener tunneling probability P_{inc} exhibits Berry phase oscillations as a function of the external transverse field H_x.

For many physical systems, the Landau–Zener model (Landau, 1932; Zener, 1932) has become an important tool for studying tunneling transitions (Crothers and Huges, 1977; Garga et al., 1985; Shimshoni and Gefen, 1991; Averin and Bardas, 1995). It must be noted that all quantum systems to which the Zener model (Landau, 1932) is applicable can be described by *pure* states and their *coherent* time evolution. In particular, the theory presented in Leuenberger and Loss (2000b) agrees well with recent measurements of $P_{\text{inc}}(H_x)$ for various temperatures in Fe$_8$ (Wernsdorfer and Sessoli, 1996; Wernsdorfer et al., 2000).

For the Zener transition, usually only the asymptotic limit is of interest. Therefore it is required that the range over which $\varepsilon_{mm'}(t) = \varepsilon_m - \varepsilon_{m'}$ is swept is much larger than the tunnel splitting $\Delta_{mm'}$ and the decoherence rate $\hbar\gamma_{mm'}$. In addition, the evolution of the spin system is restricted to times t that are much longer than the decoherence time $\tau_d = 1/\gamma_{mm'}$. In this case, tunneling transitions between pairs of degenerate excited states are incoherent. This tunneling is only observable if the temperature T is kept well below the activation energy of the potential barrier. Accordingly, one is interested only in times t that are larger than the relaxation times of the excited states. It was shown by Leuenberger and Loss (2000b) that the Zener tunneling can be described by Equation 9.22, where

$$\Gamma_{mm'}(t) = \frac{\Delta_{mm'}^2}{2}\frac{\gamma_{mm'}}{\varepsilon_{mm'}^2(t)+\hbar^2\gamma_{mm'}^2} \tag{9.40}$$

is time dependent, in contrast to Equation 9.23. As usual, the abbreviations $\gamma_{mm'} = (W_m + W_{m'})/2$ and $W_m = \Sigma_n W_{nm}$ are used, where W_{nm} denotes the approximately time-independent transition rate from $|m\rangle$ to $|n\rangle$, which can be obtained via Fermi's golden rule (Leuenberger and Loss, 1999, 2000a). The tunnel splitting (Leuenberger and Loss, 1999, 2000a) is given by

$$\Delta_{mm'} = 2\left|\sum_{\substack{m_1,\ldots,m_N\\ m_i\neq m,m'}} \frac{V_{m,m_1}}{\varepsilon_m-\varepsilon_{m_1}}\prod_{i=1}^{N-1}\frac{V_{m_i,m_{i+1}}}{\varepsilon_m-\varepsilon_{m_{i+1}}}V_{m_N,m'}\right|. \tag{9.41}$$

V_{m_i,m_j} denote off-diagonal matrix elements of the total Hamiltonian $\mathcal{H}_{\text{tot}}$.

Since all resonances n lead to similar results, Equation 9.22 is solved only in the unbiased case—corresponding to $n = 0$ (see below)—where the ground states $|s\rangle$, $|-s\rangle$ and the excited states $|m\rangle$, $|-m\rangle$, $m \in [-s + 1, s - 1]$ of the spin system with spin s are pairwise degenerate. In addition, it is assumed that the excited states are already in their stationary state, i.e., $\dot{\rho}_m = 0$ $\forall m \neq s, -s$. Equation 9.22 leads then to

$$1 - P_{\text{inc}} \equiv \Delta\rho(t) = \exp\left\{-\int_{t_0}^{t} dt'\,\Gamma_{\text{tot}}(t')\right\}, \tag{9.42}$$

where $\Delta\rho(t) = \rho_s - \rho_{-s}$, which satisfies the initial condition $\Delta\rho(t = t_0) = 1$, and thus $P_{\text{inc}}(t = t_0) = 0$. The total time-dependent relaxation rate is given by $\Gamma_{\text{tot}} = 2[\Gamma_{s,-s} + \Gamma_{\text{th}}]$, where the thermal rate Γ_{th}, which determines the incoherent relaxation via the excited states, is evaluated by means of relaxation diagrams (Leuenberger and Loss, 1999, 2000a).

Assuming linear time dependence, i.e., $\varepsilon_{mm'}(t) = \alpha_m^{m'} t$, in the transition region (Landau, 1932), and with $|\varepsilon_{mm'}^{<,>}| \gg \hbar\gamma_{mm'}$ one obtains from Equation 9.42

$$\begin{aligned}\Delta\rho &= \exp\left\{-\frac{2\Delta E_{s,-s}^2}{\hbar\alpha_s^{-s}}\arctan\left(\frac{\alpha_s^{-s}}{\hbar\gamma_{s,-s}}t\right) - \int_{-t}^{t} dt'\,\Gamma_{\text{th}}\right\}\\ &\approx \exp\left\{-\frac{\pi\Delta E_{s,-s}^2}{\hbar\alpha_s^{-s}} - \int_{-t}^{t} dt'\,\Gamma_{\text{th}}\right\},\end{aligned} \tag{9.43}$$

where $t_0 = -t$. In the low-temperature limit $T \rightarrow 0$ the excited states are not populated anymore and thus Γ_{th}, which consists of intermediate rates that are weighted by Boltzmann factors b_m (Leuenberger and Loss, 1999, 2000a), vanishes. Consequently, Equation 9.43 simplifies to

$$\Delta\rho = \exp\left\{-\frac{\pi E_{s,-s}^2}{\hbar\alpha_s^{-s}}\right\} = \exp\left\{-\frac{\pi E_{s,-s}^2}{\hbar\,|\dot{\varepsilon}_{s,-s}(0)|}\right\}. \tag{9.44}$$

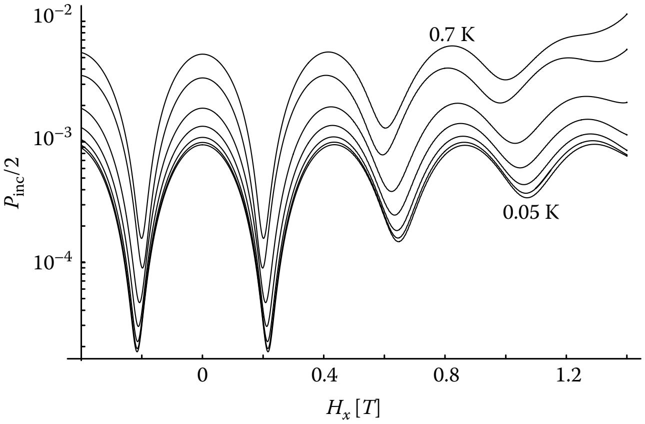

FIGURE 9.7 Zener transition probability $P_{\text{inc}}(H_x)$ for temperatures $T = 0.7$, 0.65, 0.6, 0.55, 0.5, 0.45, and 0.05 K. The fit agrees well with data. Note that P_{inc} is equal to $2P$. (From Wernsdorfer, W. et al., *Europhys. Lett.*, 50, 552, 2000.)

The exponent in Equation 9.44 differs by a factor of 2 from the Zener exponent (Landau, 1932). This is not surprising since Γ_{tot} is the relaxation rate of $\Delta\rho$, where both ρ_s and ρ_{-s} are changed in time by the same amount, and *not* an escape rate like in the case of coherent Zener transition, where only the population of the initial state is changed in time. Equation 9.44 implies $P_{\text{inc}} = 1$ for $|\dot{\varepsilon}_{s,-s}(0)| \rightarrow 0$ (adiabatic limit) and $P_{\text{inc}} = 0$ for $|\dot{\varepsilon}_{s,-s}(0)| \rightarrow \infty$ (sudden limit). After fitting the parameters the incoherent Zener theory is in excellent agreement with experiments (Wernsdorfer and Sessoli, 1996; Wernsdorfer et al., 2000) for the temperature range $0.05\,\text{K} \leq T \leq 0.7\,\text{K}$ if the states $|\pm 10\rangle$, $|\pm 9\rangle$, and $|\pm 8\rangle$ are taken into account. In particular, the path leading through $|\pm 8\rangle$ gives a non-negligible contribution for $T \geq 0.6\,\text{K}$. One obtains the following from Equation 9.43 for Fe_8 in the case $n = 0$:

$$\begin{aligned}\Gamma_{\text{tot}} &= 2\left(\Gamma_{10}^{-10} + \sum_{n=9}^{8}\frac{b_n}{\dfrac{2}{W_{10,n}} + \dfrac{1}{\Gamma_{n,-n}}}\right),\\ \Delta\rho &= \exp\left\{-\frac{\pi E_{10,-10}^2}{\hbar\alpha_{10}^{-10}} - \sum_{n=9}^{8}\frac{\pi E_{n,-n}^2 W_{10,n} b_n}{\alpha_n^{-n}\sqrt{E_{n,-n}^2 + \hbar^2 W_{10,n}^2}}\right\},\end{aligned} \tag{9.45}$$

where the approximation $\gamma_{n,-n} \approx W_{10,n}$ and $|\varepsilon_{mm'}^{<,>}| \gg E_{n,-n}, \gamma_{n,-n}$ is used. $P_{\text{inc}} = 1 - \Delta\rho$, which is plotted in Figure 9.7, is in excellent agreement with the measurements (Wernsdorfer et al., 2000).

9.6 Coherent Néel Vector Tunneling in Antiferromagnetic Molecular Wheels

Antiferromagnetic molecular clusters are the most promising candidates for the observation of coherent quantum tunneling on the mesoscopic scale currently available (Chiolero and Loss, 1998). Several systems in which an even number N of antiferromagnetically coupled ions is arranged on a ring have been synthesized to date (Papaefthymiou et al., 1994; Fabretti et al., 1996; Schromm et al., 2001; Gatteschi et al., 2002). These systems are well described by the spin Hamiltonian

$$\hat{\mathcal{H}} = J\sum_{i=1}^{N} \hat{\mathbf{s}}_i \cdot \hat{\mathbf{s}}_{i+1} + g\mu_B \mathbf{H}\cdot\sum_{i=1}^{N}\hat{\mathbf{s}}_i - k_z\sum_{i=1}^{N}\hat{s}_{i,z}^2, \qquad (9.46)$$

where

- $\hat{\mathbf{s}}_i$ is the spin operator at site i with spin quantum number s, $\hat{\mathbf{s}}_{N+1} \equiv \hat{\mathbf{s}}_1$
- J is the nearest neighbor exchange
- $\mathbf{H}$ is the magnetic field
- $k_z > 0$ the single ion anisotropy directed along the ring axis

The parameters J and k_z have been well established both for various ferric wheels (Papaefthymiou et al., 1994; Fabretti et al., 1996; Kelemen et al., 1998; Cornia et al., 1999; Jansen et al. 1999; Koch et al., 1999; Schromm et al., 2001; Zotos et al., 2001) with $N = 6, 8, 10$, and, more recently, also for a Cr wheel (Gatteschi et al., 2002). For $\mathbf{H} = 0$, the classical ground-state spin configuration of the wheel shows alternating (Néel) order with the spins pointing along $\pm\mathbf{e}_z$. The two states with the Néel vector $\mathbf{n}$ along $\pm\mathbf{e}_z$ (Figure 9.8), labeled $|\Uparrow\rangle$ and $|\Downarrow\rangle$, are energetically degenerate and separated by an energy barrier of height Nk_zs^2. Because antiferromagnetic exchange induces dynamics of Néel-ordered spins, the states $|\Uparrow\rangle$ and $|\Downarrow\rangle$ are not energy eigenstates. Rather, a molecule prepared in spin state $|\Uparrow\rangle$ would tunnel coherently between $|\Uparrow\rangle$ and $|\Downarrow\rangle$ at a rate Δ/h, where Δ is the tunnel splitting (Barbara and Chudnovsky, 1990; Krive and Zaslavskii, 1990). This tunneling of the Néel vector corresponds to a simultaneous tunneling of all N spins within the wheel through a potential barrier governed by the easy axis anisotropy. Within the framework of coherent state spin path integrals, an explicit expression for the tunnel splitting Δ as a function of magnetic field $\mathbf{H}$ has been derived (Chiolero and Loss, 1998). A magnetic field applied in the ring plane, H_x, gives rise to a Berry phase acquired by the spins during tunneling (DiVincenzo et al., 1992; van Delft and Henley, 1992; Garg, 1993; Leuenberger et al., 2003). The resulting interference of different tunneling paths leads to a sinusoidal dependence of Δ on H_x, which allows one to continuously tune the tunnel splitting from 0 to a maximum value which is of order of some Kelvin for the antiferromagnetic wheels synthesized to date.

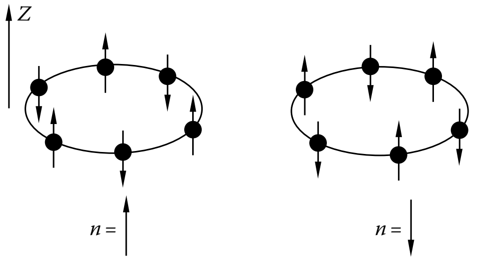

FIGURE 9.8 The two degenerate classical ground state spin configurations of an antiferromagnetic molecular wheel with easy axis anisotropy.

The tunnel splitting Δ also enters the energy spectrum of the antiferromagnetic wheel as level spacing between the ground and first excited state. Thus, Δ can be experimentally determined from various quantities such as magnetization, static susceptibility, and specific heat. Even more information on the physical properties of antiferromagnetic wheels (Equation 9.46) can be obtained from a theoretical and experimental investigation of dynamical quantities, such as the correlation functions of the total spin $\hat{\mathbf{S}} = \sum_{i=1}^{N} \hat{\mathbf{s}}_i$ or of single spins within the antiferromagnetic wheel (Meier and Loss, 2000, 2001). By symmetry arguments, it follows that the correlation function of total spin, $\langle \hat{S}_\alpha(t)\hat{S}_\alpha(0)\rangle$, which is experimentally accessible via measurement of the alternating current (AC) susceptibility does not contain a component, which oscillates with the tunnel frequency Δ/h (Meier and Loss, 2000, 2001). Hence, neither the tunnel splitting nor the decoherence rate of Néel vector tunneling can be obtained by experimental techniques which couple to the *total* spin of the wheel. In contrast, the correlation function of a single spin

$$\left\langle \hat{s}_{i,z}(t)\hat{s}_{i,z}(0)\right\rangle \simeq s^2\left(\frac{e^{-\beta\Delta/2}}{2\cosh(\beta\Delta/2)}e^{i\Delta t/\hbar} + \frac{e^{\beta\Delta/2}}{2\cosh(\beta\Delta/2)}e^{-i\Delta t/\hbar}\right) \qquad (9.47)$$

exhibits the time dependence characteristic of coherent tunneling of the quantity $\hat{S}_{i,z}$ with a tunneling rate Δ/h (Meier and Loss, 2001). We conclude that *local* spin probes are required for the observation of the Néel vector dynamics. Nuclear spins, which couple (predominantly) to a given single spin $\hat{\mathbf{s}}_i$ are ideal candidates for such probes (Meier and Loss, 2001) and have already been used to study spin cross-relaxation between electron and nuclear spins in ferric wheels (Lascialfari et al., 1999; Pini et al., 2000).

For simplicity, we consider a single nuclear spin $\hat{\mathbf{I}}$, $I = 1/2$, coupled to one-electron spin by a hyperfine contact interaction $\hat{\mathcal{H}}' = A\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{I}}$. According to Equation 9.47, the tunneling electron spin $\hat{\mathbf{s}}_1$ produces a rapidly oscillating hyperfine field $As\cos(\Delta t/\hbar)$ at the site of the nucleus.

Signatures of the coherent electron spin tunneling can thus also be found in the nuclear susceptibility. For a static magnetic field applied in the plane of the ring, H_x, it can be shown that the nuclear susceptibility

$$\chi''_{I,yy}(\omega) \simeq \frac{\pi}{4}\left[\tanh\left(\frac{\beta\gamma_I H_x}{2}\right)\delta(\omega - \gamma_I H_x/\hbar) + \left(\frac{As}{\Delta}\right)^2 \tanh\left(\frac{\beta\Delta}{2}\right)\delta(\omega - \Delta/\hbar)\right] - [\omega \to -\omega] \qquad (9.48)$$

exhibits a *satellite resonance at the tunnel splitting* Δ *of the electron spin system* (Meier and Loss, 2001). Here, $\gamma_I H_x$ is the Lamor frequency of the nuclear spin and the first term in Equation 9.48 corresponds to the transition between the Zeeman split energy levels of I. Because typically $As \simeq 1\,\text{mK}$ and $\Delta \leq 2\,\text{K}$ in Fe_{10}, the spectral weight of the satellite peak is small compared to the one of the first term in Equation 9.48 unless the magnetic field is tuned such that Δ is significantly reduced compared to its maximum value. The observation of the satellite peak in Equation 9.48 is challenging, but possible with current experimental techniques (Meier and Loss, 2001). The experiment must be conducted with single crystals of an antiferromagnetic molecular wheel with sufficiently large anisotropy $k_z > 2J/(Ns)^2$ at high, tunable fields (10 T) and low temperatures (2 K). Moreover, because the tunnel splitting $\Delta(\mathbf{H})$ depends sensitively on the relative orientation of $\mathbf{H}$ and the easy axis (Jansen et al., 1999; Zotos et al., 2001), careful field sweeps are necessary to ensure that the satellite peak in Equation 9.48 has a large spectral weight.

The need for local spin probes such as NMR or inelastic neutron scattering to detect coherent Néel vector tunneling can be traced back to the translation symmetry of the spin Hamiltonian $\hat{\mathcal{H}}$ (Meier and Loss, 2000). If this symmetry is broken, e.g., by doping of the wheel, ESR also provides an adequate technique for the detection of coherent Néel vector tunneling. If one of the original Fe or Cr ions of the wheel with spin $s = 5/2$ or $s = 3/2$, respectively, is replaced by an ion with different spin $s' \neq s$, this will in general also result in a different exchange constant J' and single ion anisotropy k_z' at the dopand site, i.e.,

$$\hat{\mathcal{H}} = J\sum_{i=2}^{N-1} \hat{\mathbf{s}}_i \cdot \hat{\mathbf{s}}_{i+1} + J'(\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}_2 + \hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}_N) + g\mu_B \mathbf{H} \cdot \sum_{i=1}^{N} \hat{\mathbf{s}}_i - \left(k_z' \hat{s}_{1,z}^2 + k_z \sum_{i=2}^{N} \hat{s}_{i,z}^2 \right). \tag{9.49}$$

Although thermodynamic quantities, such as magnetization, of the doped wheel may differ significantly from the ones of the undoped wheel, the picture of spin tunneling in antiferromagnetic molecular systems (Barbara and Chudnovsky, 1990; Krive and Zaslavskii, 1990) remains valid (Meier and Loss, 2000). However, due to unequal sublattice spins, a net total spin remains even in the Néel ordered state of the doped wheel (Figure 9.9). This allows one to distinguish the configurations sketched in Figure 9.8 according to their total spin. The dynamics of the total spin $\hat{s}$ is coupled to the one of the Néel vector (DiVincenzo et al., 1992), and coherent tunneling of the Néel vector results in a coherent oscillation of the total spin. Coherent Néel vector tunneling in doped wheels can hence also be probed by ESR. The AC susceptibility shows a resonance peak at the tunnel splitting Δ,

$$\chi''_{zz}(\omega \simeq \Delta/\hbar) = \pi(g\mu_B)^2 \left|\langle e|\, \hat{S}_z \,|g\rangle\right|^2 \tanh\left(\frac{\beta\Delta}{2}\right)\delta(\omega - \Delta/\hbar). \tag{9.50}$$

with a transition matrix element between the ground state $|g\rangle$ and first excited state $|e\rangle$,

$$\left|\langle e|\hat{S}_z|g\rangle\right| \simeq |s' - s| \frac{8Jk_z s^2}{(g\mu_B H_x)^2} \tag{9.51}$$

for $g\mu_B H_x \gg s\sqrt{8Jk_z}$. The matrix element in Equation 9.51 determines the spectral weight of the absorption peak in the ESR spectrum. The analytical dependence has been determined within a semiclassical framework and is in good agreement with numerical results obtained from exact diagonalization of small systems (Figure 9.9).

In conclusion, several antiferromagnetic molecular wheels synthesized recently are promising candidates for the observation of coherent Néel vector tunneling. Although the observation of this phenomenon is experimentally challenging, nuclear magnetic resonance, inelastic neutron scattering, and ESR on doped wheels are adequate experimental techniques for the observation of coherent Néel vector tunneling in antiferromagnetic molecular wheels.

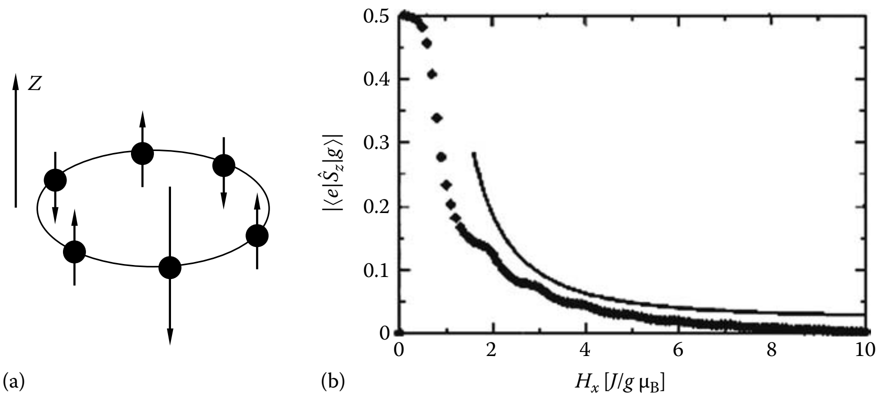

(a) (b)

FIGURE 9.9 The doped antiferromagnetic molecular wheel acquires a tracer spin which follows the Néel vector dynamics (a). Comparison of results obtained for the matrix element $|\langle e|\hat{S}_z|g\rangle|$ with a coherent state spin path integral formalism (solid line) and by numerical exact diagonalization (symbols) for $N = 4$, $s = 5/2$, $s' = 2$, $J' = J$, $k_z = k_z' = 0.0055J$ (b).

9.7 Berry-Phase Blockade in Single-Molecule Magnet Transistors

Single-electron devices show many promising characteristics such as ultimate low power consumption, down scalability to atomic dimensions, and high switching speed. In addition, these devices are supposed to be good candidates for applications in quantum computation and quantum information technologies. Recently there has been a huge interest in using magnetic molecules with large spin in single-electron devices to uncover the effects that a large spin has upon electron transport through single-molecule magnets (SMM) (Sangregorio et al., 1997). During the past few years, recent experiments demonstrated the possibility to place individual molecules between the source and drain leads allowing electron transport measurements (Jo et al., 2006; Henderson et al., 2007). In a three terminal molecular single-electron transistor (SET), the molecule is situated between the source and drain leads with an insulated gate electrode underneath. The insulating ligands on the periphery of the molecule act as isolating barriers and current can flow between the source and drain leads via a sequential tunneling process through the molecular energy levels, which are tuned by the gate electrode (see Figure 9.10). Several experimental and theoretical groups have been trying to predict and prove the effects on electron transport through an individual molecular nanomagnet SET. Heersche et al. reported Coulomb blockade and conduction excitation characteristics in a Mn_{12} individual molecular nanomagnet SET. Negative differential conductance and current suppression effects were explained with a model that combines spin properties of the molecule with standard sequential tunneling theory.

Recently, experiments have pointed out the importance of the interference between spin tunneling paths in molecules and its effects on electron transport scenarios involving SMMs. These results indicate that the Berry-phase interference plays an important role in the transport properties of SMMs in SET devices. Last year, the authors of this book chapter published an article (González and Leuenberger, 2007) in which an effect called the *Berry-phase blockade* was explained. The effect we described is a quantum interference effect that can be detected experimentally by measuring the current through a single molecular electron transistor with oppositely polarized leads and by applying a transverse magnetic field along specific angles. In the following we summarize our results.

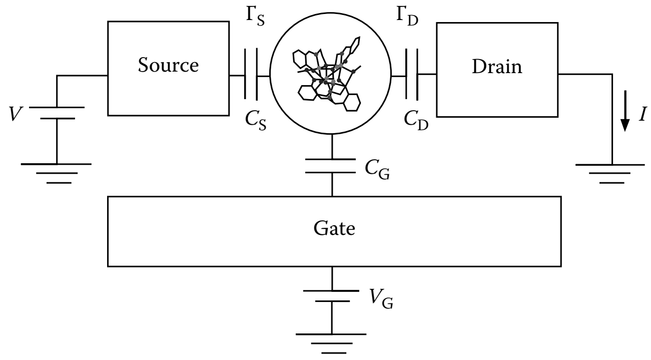

FIGURE 9.10 The sketch shows a three terminal molecular single-electron transistor. The ligands on the periphery of the molecule act as isolating barriers and electrons can flow between source and drain leads by sequential or cotunneling processes through the discrete energy levels of the molecule.

For weak coupling between the leads and the SMM we used the generalized master equation describing the electronic spin states of the SMM (see Section 9.2). The sequential tunneling rates for absorption of an electron in Equation 9.19 for ground states with spin s and s' in the case of low temperatures are given by

$$\begin{aligned} W_{s',s} &= \sum_l W_{s',s}^{(l)}, \quad W_{s',s}^{(l)} = w_\downarrow^{(l)} f_l(\Delta_{s',s}), \\ W_{-s',-s} &= \sum_l W_{-s',-s}^{(l)}, \quad W_{-s',-s}^{(l)} = w_\uparrow^{(l)} f_l(\Delta_{-s',-s}), \end{aligned} \tag{9.52}$$

and the tunneling rates for the emission of an electron are given by

$$\begin{aligned} W_{s,s'} &= \sum_l W_{s,s'}^{(l)}, \quad W_{s,s'}^{(l)} = w_\downarrow^{(l)}[1 - f_l(\Delta_{s,s'})], \\ W_{-s,-s'} &= \sum_l W_{-s,-s'}^{(l)}, \quad W_{-s,-s'}^{(l)} = w_\uparrow^{(l)}[1 - f_l(\Delta_{-s,-s'})], \end{aligned} \tag{9.53}$$

where $f_l(\Delta_{s',s}) = [1 + e^{(\Delta_{s',s} - \mu_l)/kT}]^{-1}$ is the Fermi function. $w_{\downarrow\uparrow}^{(l)}$ represents the spin-dependent transition rate from the $l \in$ [Left, Right] lead to the SMM and are defined in Fermi's golden rule approximation by $w_\downarrow^{(l)} = 2\pi D \nu_\downarrow^{(l)} |t_\downarrow^{(l)}|^2/\hbar$ and $w_\uparrow^{(l)} = 2\pi D \nu_\uparrow^{(l)} |t_\uparrow^{(l)}|^2/\hbar$, respectively, where D is the density of states and $\nu_\uparrow^{(l)}$ and $\nu_\downarrow^{(l)}$ are fractions of the number of spins polarized up and down of lead l such that $\nu_\downarrow^{(l)} + \nu_\uparrow^{(l)} = 1$. $t_\uparrow^{(l)}$ and $t_\downarrow^{(l)}$ are the tunneling amplitudes of lead l, respectively. Typical values for the tunneling rate of the electron range from around $w = 10^6\ s^{-1}$ to $w = 10^{10}\ s^{-1}$ (see González and Leuenberger, 2007, and references therein).

Solving the generalized master equation for the stationary limit, we obtained the coupled differential equations

$$\begin{aligned} \dot{\rho}_s = \left(\frac{\Delta_{s,-s}}{2\hbar}\right)^2 & \frac{2\gamma_{s,-s}}{(g\mu_B H_z(s-(-s)))^2/\hbar^2 + \gamma_{s,-s}^2} \\ & \times(\rho_{-s} - \rho_s) + W_{s,s'}\rho_{s'} - W_{s',s}\rho_s, \end{aligned} \tag{9.54}$$

$$\begin{aligned} \dot{\rho}_{-s} = \left(\frac{\Delta_{s,-s}}{2\hbar}\right)^2 & \frac{2\gamma_{s,-s}}{(g\mu_B H_z(s-(-s)))^2/\hbar^2 + \gamma_{s,-s}^2} \\ & \times(\rho_s - \rho_{-s}) + W_{-s,-s'}\rho_{-s'} - W_{-s',-s}\rho_{-s}. \end{aligned} \tag{9.55}$$

The other two differential equations are obtained by just replacing $s \leftrightarrow s'$ in the above equations. Solving the set of differential equations for ρ_s, ρ_{-s}, $\rho_{s'}$ and $\rho_{-s'}$ in the stationary case ($t \gg 1/W_{m,n}$), we obtain

$$\begin{aligned}
\rho_s &= (W_{s,s'}(W_{-s',-s}+\Gamma_{s,-s})\Gamma_{s',-s'}+W_{-s,-s'}(W_{s',s}+\Gamma_{s',-s'})\Gamma_{s,-s})/\eta,\\
\rho_{-s} &= (W_{s,s'}(W_{-s,-s'}+\Gamma_{s',-s'})\Gamma_{s,-s}+W_{-s,-s'}(W_{s',s}+\Gamma_{s,-s})\Gamma_{s',-s'})/\eta,\\
\rho_{s'} &= (W_{s',s}(W_{-s,-s'}+\Gamma_{s',-s'})\Gamma_{s,-s}+W_{-s',-s}(W_{s,s'}+\Gamma_{s,-s})\Gamma_{s',-s'})/\eta,\\
\rho_{-s'} &= (W_{s',s}(W_{-s',-s}+\Gamma_{s,-s})\Gamma_{s',-s'}+W_{-s',-s}(W_{s,s'}+\Gamma_{s',-s'})\Gamma_{s,-s})/\eta,
\end{aligned} \tag{9.56}$$

where η is a normalization factor such that $\sum_n \rho_n = 1$. The incoherent tunneling rate is

$$\Gamma_{s,-s} = \left(\frac{\Delta_{s,-s}}{2\hbar}\right)^2 \frac{2\gamma_{s,-s}}{\left(g\mu_B H_z\left(s-(-s)\right)\right)^2/\hbar^2+\gamma_{s,-s}^2}. \tag{9.57}$$

We now proceed to define the current through the SMM in terms of the density matrix for the case of the single molecule magnet Ni_4. In the case of Ni_4 we have $s = 4$ and $s' = 7/2$, therefore the current reads

$$I = e(W_{4,7/2}\rho_{7/2} + W_{-4,-7/2}\rho_{-7/2}). \tag{9.58}$$

In the case of leads that are fully polarized in opposite directions, i.e., $\nu_\uparrow^{(L)} = \nu_\downarrow^{(R)} = 1$ or $\nu_\downarrow^{(L)} = \nu_\uparrow^{(R)} = 1$, we get one of the two following conditions for the transition rates, respectively:

$$W_{-4,-7/2} = W_{7/2,4} = 0 \quad \text{or} \quad W_{4,7/2} = W_{-7/2,-4} = 0. \tag{9.59}$$

Choosing the case $\nu_\uparrow^{(L)} = \nu_\downarrow^{(R)} = 1$ and using the condition $W_{-4,-7/2} = W_{7/2,4} = 0$, we obtain from Equation 9.58

$$\frac{e}{I} = \frac{2}{W_{-7/2,-4}} + \frac{1}{\Gamma_{4,-4}} + \frac{2}{W_{4,7/2}} + \frac{1}{\Gamma_{7/2,-7/2}}. \tag{9.60}$$

Equation 9.60 reflects the fact that the current through the SMM depends on the tunnel splittings. The transitions that contribute to the current through the SMM in the case of fully polarized leads $\nu_\uparrow^{(L)} = \nu_\downarrow^{(R)} = 1$ are $4 \to 7/2 \to -7/2 \to -4$. Figure 9.11 shows the current as a function of $H_\perp$ for fully polarized leads. If the tunnel splitting $\Delta_{4,-4}$ or $\Delta_{7/2,-7/2}$ is topologically quenched (see Figure 9.11), $\Gamma_{4,-4}$ or $\Gamma_{7/2,-7/2}$ vanishes (see Equation 9.57), which leads to complete current suppression according to Equation 9.60. Since this current blockade is a consequence of the topologically quenched tunnel splitting, we call it *Berry-phase blockade*. Note that the current can also be suppressed by applying H_z, which follows immediately from Equations 9.57 and 9.60.

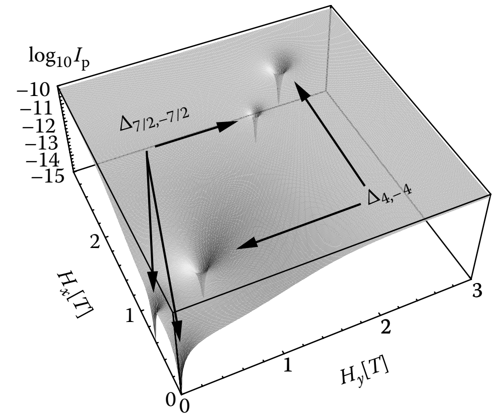

FIGURE 9.11 The graph shows the $\log_{10} I$ vs. $H_\perp$. The scale varies from $I = 0.1$ nA to 1 fA. From the figure we see that the current is completely suppressed at the zeros of the tunnel splitting, i.e., $\Delta_{4,-4} = \Delta_{7/2,-7/2} = 0$.

9.8 Concluding Remarks

We have seen how molecular nanomagnets offer an interesting platform to explore the quantum dynamics of mesoscopic systems. The theoretical advantage of using molecular nanomagnets is that molecules are all identical to each other, allowing the performance of experiments on large assemblies of identical particles and detect quantum effects. Molecules can be investigated in solutions or solids depending on the experiment you want to perform, allowing accurate measurements. We have paid special attention to the quantum tunneling of the magnetization of molecular nanomagnets and the different approaches that exist to describe this phenomenon. In particular we have placed emphasis on the interference effects of the quantum tunneling of the magnetization via Berry's phase and we have included some of our recent results in this field (González and Leuenberger, 2007).

Molecular nanomagnets are also the subject of intense research in molecular electronics. Single-electron devices have taken advantage of the small size of nanomagnets to create electronic structures of atomic dimensions. These devices are supposed to be good candidates for applications in quantum computation and quantum information technologies. There has been also a lot of research on single-electron transistors (SET) made of molecular nanomagnets to study quantum effects such as the Kondo effect in this kind of structures (Leuenberger and Mucciolo, 2006). The Kondo effect is a very well known and studied phenomenon in condensed matter physics. This effect arises when a magnetic impurity is placed into a conductor, which causes a dramatic increase in the resistivity of the metal at low temperatures. The Kondo effect can be seen also in a single-electron transistor where the molecular nanomagnet plays the role of the magnetic impurity and the leads of the SET mimics the bulk metal. It has been shown that the Hamiltonian for a molecular nanomagnet placed in a SET device can be mapped onto the Kondo Hamiltonian (González et al., 2008). This result is of great interest because one can apply the poor man's scaling theory to describe the Kondo physics at low temperatures and see how the Berry-phase oscillations become temperature dependent. This fact allows us to conclude that the scaling equations can be checked experimentally by measuring the renormalized zero points of the Berry phase.

The field of molecular nanomagnets is growing very fast and chemistry is playing a major role concerning the growth of these nanostructures. New nanomagnets with novel and interesting magnetic properties are being produced, which demand more sophisticated theories to explain and describe the dynamics of these magnetic mesoscopic systems. One direction which remains a major area of investigation lately is related to storing and decoding information on a single spin state to be able to create a quantum computer. The implementation of Grover's algorithm with molecular nanomagnets has already been proposed, however, it is required to perform simultaneous manipulation of many spin phases without losing quantum coherence, which remains still an experimental challenge.

Acknowledgment

We acknowledge support from NSF-ECCS 0725514, the DARPA/MTO Young Faculty Award HR0011-08-1-0059, 0901784, and AFOSR FA 9550-09-1-0450.

References

Averin, D. and A. Bardas. AC Josephson effect in a single quantum channel. *Phys. Rev. Lett.*, 75:1831–1834, 1995.

Barbara, B. and E.M. Chudnovsky. Macroscopic quantum tunneling in antiferromagnets. *Phys. Lett. A*, 145:205–208, 1990.

Barbara, B., W. Wernsdorfer, L.C. Sampaio, J.G. Park, C. Paulsen, M.A. Novak, R. Ferré, D. Mailly, R. Sessoli, A. Caneschi, K. Hasselback, A. Benoit, and L. Thomas. Mesoscopic quantum tunneling of the magnetization. *J. Magn. Magn. Mater*, 140–144:1825–1828, 1995.

Barra, A.L., P. Debrunner, D. Gatteschi, C.E. Schulz, and R. Sessoli. Superparamagnetic-like behavior in an octanuclear iron cluster. *Europhys. Lett.*, 35:133–138, 1996.

Chiolero, A. and D. Loss. Macroscopic quantum coherence in molecular magnets. *Phys. Rev. Lett.*, 80:169–172, 1998.

Chudnovsky, E.M. and L. Gunther. Quantum tunneling of magnetization in small ferromagnetic particles. *Phys. Rev. Lett.*, 60:661–664, 1988.

Cornia, A., M. Affronte, J.C. Lasjaunias, and A. Caneschi. Low-temperature specific heat of Fe6 and Fe10 molecular magnets. *Phys. Rev. B*, 60:1161–1166, 1999.

Crothers, D.S. and J.G. Huges. Stueckelberg close-curve-crossing phases. *J. Phys. B*, 10:L557–L560, 1977.

DiVincenzo, D.P., D. Loss, and G. Grinstein. Suppression of tunneling by interference in half-integer-spin particles. *Phys. Rev. Lett.*, 69:3232–3235, 1992.

Enz, M. and R. Schilling. Magnetic field dependence of the tunnelling splitting of quantum spins. *J. Phys. C: Solid State Phys.*, 19:L711–L715, 1986.

Fabretti, A.C., S. Foner, D. Gatteschi, R. Grandi, A. Caneshi, A. Cornia, and L. Schenetti. Synthesis, crystal structure, magnetism, and magnetic anisotropy of cyclic clusters comprising six iron(iii) ions and entrapping alkaline ions. *Chem. Eur. J.*, 2:1379–1387, 1996.

Feynman R.P. Space-Time Approach to Non-Relativistic Quantum Mechanics. *Rev. Mod. Phys.*, 20:367–387, 1948.

Fort, A., A. Rettori, J. Villain, D. Gatteschi, and R. Sessoli. Mixed quantum-thermal relaxation in Mn12 acetate molecules. *Phys. Rev. Lett.*, 80:612–615, 1998.

Friedman, J.R., M.P. Sarachik, J. Tejada, and R. Ziolo. Macroscopic measurement of resonant magnetization tunneling in high-spin molecules. *Phys. Rev. Lett*, 76:3830–3833, 1996.

Garanin, D.A. and E.M. Chudnovsky. Thermally activated resonant magnetization tunneling in molecular magnets: Mn12ac and others. *Phys. Rev. B*, 56:11102–11118, 1997.

Garg, A. Topologically quenched tunnel splitting in spin systems without kramers' degeneracy. *Europhys. Lett.*, 22:205–210, 1993.

Garga, A., N.J. Onuchi, and V. Ambegaokar. Effect of friction on electron transfer in bio-molecules. *J. Chem. Phys.*, 83:4491–4503, 1985.

Gatteschi, D., A.A. Smith, M. Helliwell, R.E.P. Winpenny, A. Cornia, A.L. Barra, A.G.M. Jansen, E. Rentschler, J. van Slageren, R. Sessoli, and G.A. Timco. Magnetic anisotropy of the antiferromagnetic ring [cr8f8piv16]. *Chem. Eur. J.*, 8:277–285, 2002.

González, G. and M.N. Leuenberger. Berry-phase blockade in single-molecule magnets. *Phys. Rev. Lett.*, 98:256804-4, 2007.

González, G., M.N. Leuenberger, and E.R. Mucciolo. Kondo effect in single-molecule magnet transistors. *Phys. Rev. B*, 78:054445-12, 2008.

Henderson, J.J., C.M. Ramsey, E. del Barco, A. Mishra, and G. Christou. Fabrication of nano-gapped single-electron transistors for transport studies of individual single-molecule magnets. *J. Appl. Phys.*, 101:09E102, 2007.

Hernández, J.M., X.X. Zhang, F. Luis, J. Bartolomé, J. Tejada, J.R. Friedman, M.P. Sarachik, and R. Ziolo. Field tuning of thermally activated magnetic quantum tunnelling in mn12—ac molecules. *Europhys. Lett.*, 35:301–306, 1996.

Hernández, J.M., X.X. Zhang, J. Tejada, F. Luis, and R. Ziolo. Evidence for resonant tunneling of magnetization in Mn12sacetate complex. *Phys. Rev. B*, 55:5858–5865, 1997.

Jansen, A.G.M., A. Cornia, and M. Affronte. Magnetic anisotropy of Fe6 and Fe10 molecular rings by cantilever torque magnetometry in high magnetic fields. *Phys. Rev. B*, 60:12177–12183, 1999.

Jo, M.-H., J.E. Grose, K. Baheti, M.M. Deshmukh, J.J. Sokol, E.M. Rumberger, D.N. Hendrickson, J.R. Long, H. Park, and D.C. Ralph. Signatures of molecular magnetism in single-molecule transport spectroscopy. *Nano Lett.*, 6:2014–2020, 2006.

Kelemen, M.T., M. Weickenmeier, B. Pilawa, R. Desquiotz, and A. Geisselmann. Magnetic properties of new Fe6(triethanolaminate(3-))6 spin-clusters. *J. Magn. Magn. Mat.*, 177–181:748–749, 1998.

Koch, R., P. Müller, I. Bernt, R.W. Saalfrank, H.P. Andres, H.U. Güdel, O. Waldmann, J. Schülein, and O. Allenspach. Magnetic anisotropy of two cyclic hexanuclear Fe(iii) clusters entrapping alkaline ions. *Inorg. Chem.*, 38:5879–5886, 1999.

Korenblit, I.Ya. and E.F. Shender. Low-temperature properties of amorphous magnetic materials with random axis of anisotropy. *Sov. Phys. JETP*, 48:937–942, 1978.

Krive, I.V. and O.B. Zaslavskii. Macroscopic quantum tunnelling in antiferromagnets. *J. Phys. Condens. Matter*, 2:9457–9462, 1990.

Landau, L.D. On the theory of transfer of energy at collisions. *Phys. Z. Sowjetunion*, 2:46, 1932.

Lascialfari, A., F. Borsa, M. Horvati, A. Caneschi, M.H. Julien, Z.H. Jang, and D. Gatteschi. Proton NMR for measuring quantum level crossing in the magnetic molecular ring Fe10. *Phys. Rev. Lett.*, 83:227–230, 1999.

Leuenberger, M.N. and D. Loss. Spin relaxation in Mn12-acetate. *Europhys. Lett.*, 46:692–698, 1999.

Leuenberger, M.N. and D. Loss. Spin tunneling and phonon-assisted relaxation in Mn12-acetate. *Phys. Rev. B*, 61:1286–1302, 2000a.

Leuenberger, M.N. and D. Loss. Incoherent zener tunneling and its application to molecular magnets. *Phys. Rev. B*, 61:12200–12203, 2000b.

Leuenberger, M.N. and D. Loss. Spin tunneling and topological selection rules for integer spins. *Phys. Rev. B*, 63:054414-4, 2001a.

Leuenberger, M.N. and D. Loss. Quantum computing in molecular magnets. *Nature*, 410:789–793, 2001b.

Leuenberger, M.N. and D. Loss. Reply to the comment of E. M. Chudnovsky and D. A. Garanin on Spin relaxation in Mn_{12}-acetate. *Europhys. Lett.*, 52: 247–248, 2000.

Leuenberger, M.N., F. Meier, and D. Loss. Quantum spin dynamics in molecular magnets. *Monatshefte fuer Chemie*, 134:217–233, 2003.

Leuenberger, M.N. and E.R. Mucciolo. Berry-phase oscillation of the Kondo effect in single-molecule magnets. *Phys. Rev. Lett.*, 97:126601, 2006.

Luis, F., J. Bartolomé, and F. Fernández. Resonant magnetic quantum tunneling through thermally activated states. *Phys. Rev. B*, 57:505–513, 1998.

Meier, F. and D. Loss. Thermodynamics and spin-tunneling dynamics in ferric wheels with excess spin. *Phys. Rev. B*, 64:224411–224414, 2000.

Meier, F. and D. Loss. Electron and nuclear spin dynamics in antiferromagnetic molecular rings. *Phys. Rev. Lett.*, 86:5373–5376, 2001.

Novak, M.A. and R. Sessoli. *Quantum Tunneling of Magnetization*. L. Gunther and B. Barbara (eds.), Kluwer, Dordrecht, the Netherlands, 1995.

Novak, M.A., R. Sessoli, A. Caneschi, and D. Gatteschi. Magnetic properties of a Mn cluster organic compound. *J. Magn. Magn. Mater*, 146:211–213, 1995.

Ohm, T., C. Sangregorio, and C. Paulsen. Local field dynamics in a resonant quantum tunneling system of magnetic molecules. *Europhys. J. B*, 6:195–199, 1998.

Papaefthymiou, G.C., S. Foner, D. Gatteschi, K.L. Taft, C.D. Delfs, and S.J. Lippard. [Fe(OMe)2(O2CCH2Cl)]10, a molecular ferric wheel. *J. Am. Chem. Soc.*, 116:823–832, 1994.

Paulsen, C. and J.G. Park. *Quantum Tunneling of Magnetization*. L. Gunther and B. Barbara (eds.), Kluwer, Dordrecht, the Netherlands, 1995.

Paulsen, C., J.G. Park, B. Barbara, R. Sessoli, and A. Caneschi. Novel features in the relaxation times of Mn12Ac. *J. Magn. Magn. Mater*, 140–144:379–380, 1995.

Pini, M.G., A. Cornia, A. Fort, and A. Rettori. Low-temperature theory of proton NMR in the molecular antiferromagnetic ring Fe10. *Europhys. Lett.*, 50:88–93, 2000.

Sangregorio, C., T. Ohm, C. Paulsen, R. Sessoli, and D. Gatteschi. Quantum tunneling of the magnetization in an iron cluster nanomagnet. *Phys. Rev. Lett.*, 78:4645–4648, 1997.

Schromm, S., J. Schülein, P. Müller, I. Bernt, R.W. Saalfrank, O. Waldmann, R. Koch, and F. Hampel. Magnetic anisotropy of a cyclic octanuclear Fe(iii) cluster and magneto-structural correlations in molecular ferric wheels. *Inorg. Chem.*, 40:2986–2995, 2001.

Sessoli, R., D. Gatteschi, A. Caneschi, and M.A. Novak. Magnetic bistability in a metal-ion cluster. *Nature*, 365:141–143, 1993.

Shimshoni, E. and Y. Gefen. Onset of dissipation in zener dynamics: Relaxation vs. dephasing. *Ann. Phys.*, 210:16–80, 1991.

Thomas, L., F. Lionti, R. Ballou, D. Gatteschi, R. Sessoli, and B. Barbara. Macroscopic quantum tunnelling of magnetization in a single crystal of nanomagnets. *Nature*, 383:145–147, 1996.

von Delft, J. and C.L. Henley. Destructive quantum interference in spin tunneling problems. *Phys. Rev. Lett.*, 69:3236–3239, 1992.

van Hemmen, J.L. and A. Süto. Tunnelling of quantum spins. *Europhys. Lett.*, 1:481–490, 1986.

Villain, J., F. Hartmann-Boutron, R. Sessoli, and A. Rettori. Magnetic relaxation in big magnetic molecules. *Europhys. Lett.*, 27:159–164, 1994.

Wernsdorfer, W. and R. Sessoli. Quantum phase interference and parity effects in magnetic molecular clusters. *Science*, 284:133–135, 1996.

Wernsdorfer, W., R. Sessoli, A. Caneshi, D. Gatteschi, and A. Cornia. Nonadiabatic Landau-Zener tunneling in Fe8 molecular nanomagnets. *Europhys. Lett.*, 50:552–558, 2000.

Zener, C. Non-adiabatic crossing of energy levels. *Proc. R. Soc. Lond. A*, 137:696, 1932.

Zotos, X., B. Normand, X. Wang, and D. Loss. Magnetization in molecular iron rings. *Phys. Rev. B*, 63:184409, 2001.

10

Inelastic Electron Transport through Molecular Junctions

10.1 Introduction **10**-1
10.2 Coherent Transport **10**-2
10.3 Buttiker Model for Inelastic Transport **10**-5
10.4 Vibration-Induced Inelastic Effects **10**-6
10.5 Dissipative Transport **10**-9
10.6 Polaron Effects: Hysteresis, Switching, and Negative Differential Resistance **10**-11
10.7 Molecular Junction Conductance and Long-Range Electron-Transfer Reactions **10**-13
10.8 Concluding Remarks **10**-15
References **10**-16

Natalya A. Zimbovskaya
University of Puerto Rico

10.1 Introduction

Molecular electronics is known to be one of the most promising developments in nanoelectronics, and the past decade has seen extraordinary progress in this field (Aviram et al. 2002, Cuniberti et al. 2005, Nitzan 2001). Present activities on molecular electronics reflect the convergence of two trends in the fabrication of nanodevices, namely, the top-down device miniaturization through the lithographic methods and bottom-up device manufacturing through the atom-engineering and the self-assembly approaches. The key element and the basic building block of molecular electronics is a junction including two electrodes (leads) linked by a molecule, as schematically shown in Figure 10.1. Usually, the electrodes are microscopic large but macroscopic small contacts that could be connected to a battery to provide the bias voltage across the junction. Such a junction may be treated as a quantum dot coupled to the charge reservoirs. The discrete character of energy levels on the dot (molecule) is combined with nearly continuous energy spectra on the reservoirs (leads) occurring due to their comparatively large size.

When the voltage is applied, an electric current flows through the junction. Successful transport experiments with molecular junctions (Ho 2002, Lortscher et al. 2007, Park et al. 2000, Poot et al. 2006, Reichert et al. 2002, Smit et al. 2002, Yu et al. 2004) confirm their significance as active elements in nanodevices. These include applications as rectifiers (molecular diodes), field-effect transistors (molecular triodes), switches, memory elements and sensors. Also, these experiments emphasize the importance of a thorough analysis of the physics underlying electron transport through molecular junctions. A detailed understanding of electron transport at the molecular scale is a key step to future device operations. A theory of electron transport in molecular junctions is being developed since the last two decades, and main transport mechanisms are currently elucidated in general terms (Datta 2005, Imry and Landauer 1999). However, a progress of the experimental capabilities in the field of molecular electronics brings new theoretical challenges causing a further development of the theory.

Speaking of transport mechanisms, it is useful to make a distinction between elastic electron transport when the electron energy remains the same as it travels through the junction, and inelastic transport processes when the electron undergoes energy changes due to its interactions with the environment. There are several kinds of processes bringing inelasticity in the electron transport in mesoscopic systems including molecular junctions. Chief among these are electron–electron and electron–phonon scattering processes. These processes may bring about significant inelastic effects modifying the transport properties of molecular devices and charging, desorption, and chemical reactions as well. To keep this chapter at a reasonable length, we concentrate on the inelastic effects originating from electron–phonon interactions.

In practical molecular junctions, the electron transport is always accompanied by nuclear motions in the environment. Therefore, the conduction process is influenced by the coupling between the electronic and the vibrational degrees of freedom. Nuclear motions underlie the interplay between the coherent electron tunneling through the junction and the inelastic thermally assisted hopping transport

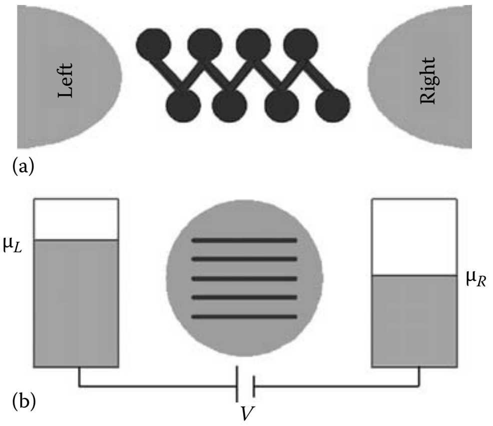

FIGURE 10.1 Schematic drawing of a junction including two electrodes and a molecule in between (a). When the voltage is applied across the junction, electrochemical potentials μ_L and μ_R differ, and the conduction window opens up (b).

(Nitzan 2001). Also, electron–phonon interactions may result in polaronic conduction (Galperin et al. 2005, Gutierrez et al. 2005, Kubatkin et al. 2003, Ryndyk et al. 2008), and they are directly related to the junction heating (Segal et al. 2003) and to some specific effects such as alterations in both shape of the molecule and its position with respect to the leads (Komeda et al. 2002, Mitra et al. 2004, Stipe et al. 1999). The effects of electron–phonon interactions may be manifested in the inelastic tunneling spectrum (IETS) that presents the second derivative of the current in the junction d^2I/dV versus the applied voltage V. The inelastic electron tunneling microscopy has proven to be a valuable method for the identification of molecular species within the conduction region, especially when employed in combination with scanning tunneling microscopy and/or spectroscopy (Galperin et al. 2004).

Inelasticity in the electron transport through molecular junctions is closely related to the dephasing effects. One may say that incoherent electron transport always includes an inelastic contribution with the possible exception of the low temperature range. The general approach to theoretically analyzing electron transport through molecular junctions in the presence of dissipative/phase-breaking processes in both electronic and nuclear degrees of freedom is based on advanced formalisms (Segal et al. 2000, Skourtis and Mukamel 1995, Wingreen et al. 1989, 1993). These microscopic computational approaches have the advantages of being capable of providing detailed dynamics information. However, while considering stationary electron transport through molecular junctions, one may turn to the less time-consuming approach based on scattering matrix formalism (Buttiker 1986, Li and Yan 2001a,b), as discussed below.

10.2 Coherent Transport

To better show the effects of dissipation/dephasing on the electron transport through molecular junctions, it seems reasonable to start from the case where these effects do not occur. So, we consider a molecule (presented as a set of energy levels) placed in between two electrodes with nearly continuous energy spectra. While there is no bias voltage applied across the junction, the latter remains in equilibrium characterized by the equilibrium Fermi energy E_F, and there is no current flowing through it. When the bias voltage is applied, it keeps the left and right electrodes at different electrochemical potentials μ_L and μ_R. Then the electric current appears in the junction, and the molecular energy levels located in between the electrochemical potentials μ_L and μ_R play a major part in maintaining this current. Electrons from occupied molecular states tunnel to the electrodes in accordance with the voltage polarity, and the electrons from one electrode travel to another one using unoccupied molecular levels as intermediate states for tunneling. Usually, the electron transport in molecular junctions occurs via the highest occupied (HOMO) and the lowest unoccupied (LUMO) molecular orbitals that work as channels for electron transmission. Obviously, the current through the junction depends on the quality of contacts between the leads and the molecule ends. However, there also exists the limit for the conductance in the channels. As was theoretically shown (Landauer 1970), the maximum conductance of a channel with a single-spin degenerate-energy level equals:

$$G_0 = \frac{e^2}{\pi\hbar} = (25.8\ \text{k}\Omega)^{-1} \tag{10.1}$$

where
 e is the electron charge
 $\hbar$ is Planck's constant

This is a truly remarkable result for it proves that the minimum resistance $R_0 = G_0^{-1}$ of a molecular junction cannot become zero. In another words, one never can short-circuit a device operating with quantum channels. Also, the expression (10.1) shows that the conductance is a quantized quantity.

Conductance g in practical quantum channels associated with molecular orbitals can take on values significantly smaller that G_0, depending on the delocalization in the molecular orbitals participating in the electron transport. In molecular junctions it also strongly depends on the molecule coupling to the leads (quality of contacts), as was remarked before. The total resistance $r = g^{-1}$ includes contributions from the contact and the molecular resistances, and could be written as (Wingreen et al. 1993)

$$r = \frac{1}{G_0}\left(1 + \frac{1-T}{T}\right). \tag{10.2}$$

Here, T is the electron transmission coefficient that generally takes on values less than unity.

The general expression for the electric current flowing through the molecular junction could be obtained if one calculates the total probability for an electron to travel between two electrodes at a certain tunnel energy E and then integrates the latter over the whole energy range (Datta 1995). This results in the well-known Landauer expression:

$$I = \frac{e}{\pi\hbar}\int T(E)\left[f_L(E) - f_R(E)\right]dE. \tag{10.3}$$

Here, $f_{L,R}(E)$ are Fermi distribution functions for the electrodes with chemical potentials $\mu_{L,R}$, respectively. The values of $\mu_{L,R}$ differ from the equilibrium energy E_F, and they are determined by the voltage distribution inside the system. Assuming that coherent tunneling predominates in electron transport, the electron transmission function is given by (Datta et al. 1997, Samanta et al. 1996):

$$T(E) = 2Tr\left\{\Delta_L G \Delta_R G^+\right\}, \tag{10.4}$$

where

the matrices $\Delta_{L,R}$ represent the imaginary parts of the self-energy terms $\Sigma_{L,R}$ describing the coupling of the molecule to the electrodes

G is the Green's function matrix for the molecule whose matrix elements between the molecular states $\langle i|$ and $|j\rangle$ have the form

$$G_{ij} = \left\langle i\middle|E - H\middle|j\right\rangle. \tag{10.5}$$

Here, H is the molecular Hamiltonian including the self-energy parts $\Sigma_{L,R}$.

When a molecule contacts the surface of electrodes, this results in a charge transfer between the molecule and the electrodes, and in a modification of the molecule energy states due to the redistribution of the electrostatic potential within the molecule. Besides, the external voltage applied across the junction brings additional changes to the electrostatic potential further modifying the molecular orbitals. The coupling of the molecule to the leads may also depend on the voltage distribution. So, generally, the electron transmission T inserted in Equation 10.3 and the electrochemical potentials $\mu_{L,R}$ depend on the electrostatic potential distribution in the system. To find the correct distribution of the electric field inside the junction one must simultaneously solve the Schrodinger equation for the molecule and the Poisson equation for the charge density, following a self-consistent converging procedure. This is a nontrivial and complicated task, and significant effort was applied to study the effect of electrostatic potential distribution on the electron transport through molecules (Damle et al. 2001, Di Ventra et al. 2000, Galperin et al. 2006, Lang and Avouris 2000a,b, Mujica et al. 2000, Xue and Ratner 2003a,b, 2004, Xue et al. 2001). Here, we put these detailed considerations aside, and we use the simplified expression for the electrochemical potentials:

$$\mu_L = E_F + \eta|e|V; \quad \mu_R = E_F - (1-\eta)|e|V, \tag{10.6}$$

where the parameter η indicates how the bias voltage is distributed between the electrodes. Also, we assume that inside the molecule, the external electrostatic field is screened due to the charge redistribution, and the electron transmission is not sensitive to the changes in the voltage V. Although very simple, this model allows one to analyze the main characteristics of the electron transport through molecular junctions. Within this model one may write down the following expression for the self-energy parts (D'Amato and Pastawski 1990):

$$(\Sigma_\beta)_{ij} = \sum_k \frac{\tau^*_{ik,\beta}\,\tau_{kj,\beta}}{E - \varepsilon_{k,\beta} + is}. \tag{10.7}$$

Here,

$\beta \in L, R$, $\tau_{ik,\beta}$ is the coupling strength between the ith molecular state and the kth state on the left/right lead

$\varepsilon_{k,\beta}$ are the energy levels on the electrodes

s is a positive infinitesimal parameter

Assuming that the molecule is reduced to a single orbital with the energy E_0 (a single-site bridge), Green's function accepts the form

$$G(E) = \frac{1}{E - E_0 - \Sigma_L - \Sigma_R}. \tag{10.8}$$

Accordingly, in this case, one may simplify the expression (10.4) for the electron transmission:

$$T(E) = \frac{4\Delta_L\Delta_R}{(E - E_0)^2 + (\Delta_L + \Delta_R)^2}. \tag{10.9}$$

To elucidate the main features of electron transport through molecular junctions we consider a few examples. In the first example, we mimic a molecule as a one-dimensional chain consisting of N identical hydrogen-like atoms with the nearest neighbors interaction. We assume that there is one state per isolated atom with the energy E_0, and that the coupling between the neighboring sites in the chain is characterized by the parameter b. Such a model was theoretically analyzed by D'Amato and Pastawski and in some other works (see e.g., Mujica et al. 1994). Based on Equations 10.4 and 10.5, it could be shown that for a single-site chain ($N = 1$) the electron transmission reveals a well-distinguished peak at $E = E_0$, shown in Figure 10.2. The height of this peak is determined by the coupling of the bridge site to the electrodes. The peak in the electron transmission arises because the molecular orbital $E = E_0$ works as the channel/bridge for electron transport between the leads. Similar peaks appear in the conductance $g = dI/dV$. Assuming the symmetric voltage distribution ($\eta = 1/2$), the peak in the conductance is located at $V = \pm 2E_0$. As for the current voltage characteristics, they display step-like shapes with the steps at $V = \pm 2E_0$. When the chain includes several sites, we obtain a set of states (orbitals) for our bridge instead of the single state $E = E_0$, and their number equals the number of

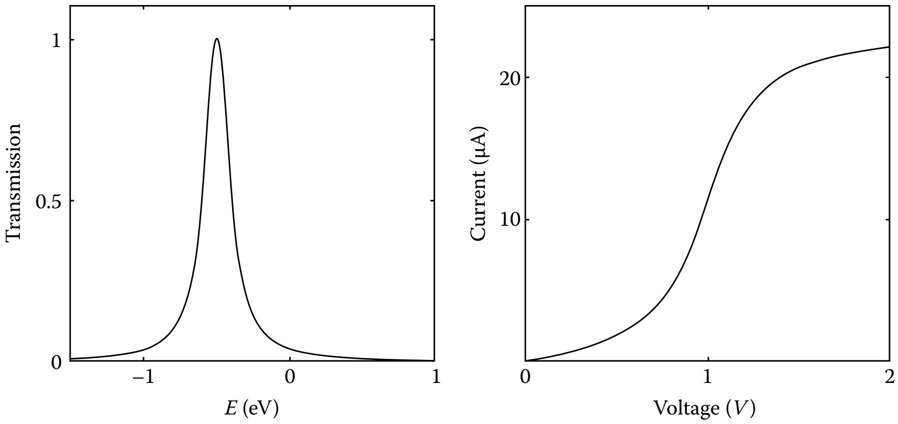

FIGURE 10.2 Coherent electron transmission (left panel) and current (right panel) versus bias voltage applied across a molecular junction where the molecule is simulated by a single electronic state. The curves are plotted assuming $\Delta_L = \Delta_R = 0.1\,\text{eV}$, $E_0 = -0.5\,\text{eV}$, $T = 30\,\text{K}$.

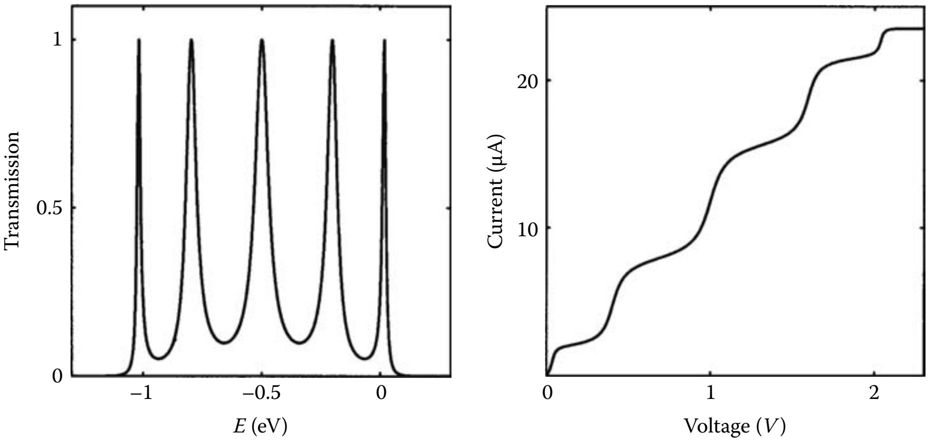

FIGURE 10.3 Coherent electron transmission (left panel) and current (right panel) through a junction with a five electronic states bridge. The curves are plotted for $\Delta_L = \Delta_R = 0.1\,\text{eV}$, $b = 0.3\,\text{eV}$, $E_0 = -0.5\,\text{eV}$, $T = 30\,\text{K}$.

sites in the chain. All these states are the channels for the electron transport. Correspondingly, the transmission reveals a set of peaks as presented in Figure 10.3. The peaks are located within the energy range with the width $4b$ around $E = E_0$. The coupling of the chain ends to the electrodes affects the transmission, especially near $E = E_0$. As the coupling strengthens, the transmission minimum values increase. Now, the current voltage curves exhibit a sequence of steps. The longer the chain, the more energy levels it possesses, and the more the number of steps in the I–V curves.

The second example concerns the electron transport through a carbon chain placed between copper electrodes. In this case, as well as for practical molecules, a preliminary step in transport calculations is to compute the relevant molecular energy levels and wave functions. Usually, these computations are carried out employing quantum chemistry software packages (e.g., Gaussian) or density functional-based software. Also, a proper treatment of the molecular coupling to the electrodes is necessary for it brings changes into the molecular energy states. For this purpose, one may use the concept of an "extended molecule" proposed by Xue et al. (2001). The point of this concept is that only a few atoms on the surface of the metallic electrode are significantly disturbed when the molecule is attached to the latter. These atoms are located in the immediate vicinity of the molecule end. Therefore, one may form a system consisting of the molecule itself and the atoms from the electrode surfaces perturbed by the molecular presence. This system is called the extended molecule and treated as such while computing the molecular orbitals. In the example considered, the extended molecule included four copper atoms on each side of the carbon chain. The results for electron transport are shown in Figure 10.4. Again, we observe a comb-like shape of the electron transmission corresponding to the set of transport channels provided by the molecular orbitals, and the stepwise I–V curve originating from the latter. Transport calculations similar to those described above were repeatedly carried out in the last two decades for various practical molecules (see e.g., Galperin et al. 2006, Xue and Ratner 2003a,b, Xue et al. 2001, Zimbovskaya 2003, 2008, Zimbovskaya and Gumbs 2002).

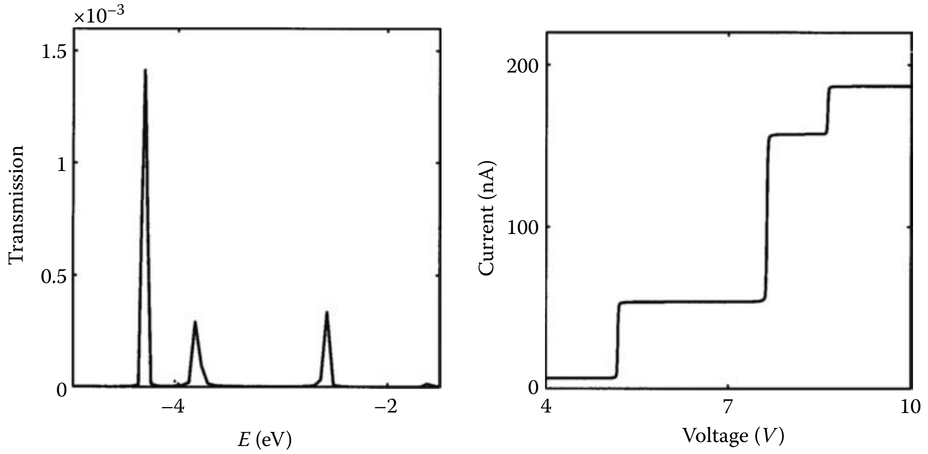

FIGURE 10.4 Coherent electron transmission (left panel) and current (right panel) through a carbon chain coupled to the copper leads at $T = 30$ K.

10.3 Buttiker Model for Inelastic Transport

An important advantage of the phenomenological model for the incoherent/inelastic quantum transport proposed by Buttiker (1986) is that this model could easily be adapted to analyze various inelastic effects in electron transport through molecules (and some other mesoscopic systems) avoiding complicated and time-consuming advanced methods, such as those based on the nonequilibrium Green's functions formalism (NEGF).

Here, we present the Buttiker model for a simple junction including two electrodes linked by a single-site molecular bridge. The bridge is attached to a phase-randomizing electron reservoir, as shown in Figure 10.5. Electrons tunnel from the electrodes to the bridge and vice versa via the channels 1 and 2. While on the bridge, an electron could be scattered into the channels 3 and 4 with a certain probability ε. Such an electron arrives at the reservoir where it undergoes inelastic scattering accompanied by phase-breaking and then the reservoir reemits it back to the channels 3 and 4 with the same probability. So, within the Buttiker model, the electron transport through the junction is treated as the combination of tunnelings through the barriers separating the molecule from the electrodes and the interaction with the phase-breaking electron reservoir coupled to the bridge site. The key parameter of the model is the probability ε that is closely related to the coupling strength between the bridge site and the reservoir. When $\varepsilon = 0$ the reservoir is detached from the bridge, and the electron transport is completely coherent and elastic. Within the opposite limit ($\varepsilon = 1$), electrons are certainly scattered into the reservoir that results in the overall phase randomization and inelastic transport.

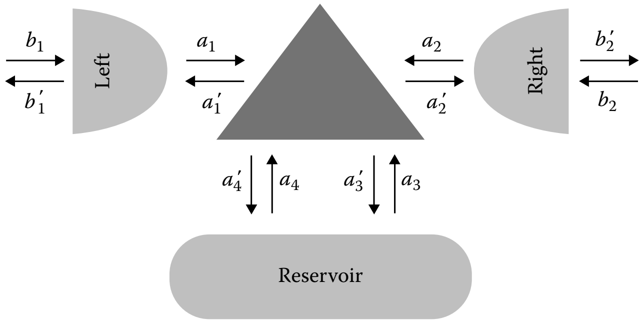

FIGURE 10.5 Schematic drawing illustrating inelastic electron transport through a molecular junction within the Buttiker model.

Within the Buttiker model, the particle fluxes outgoing from the junctions J_i' could be presented as the linear combinations of the incoming fluxes J_k where the indexes i, k label the channels for the transport: $1 \le i, k \le 4$.

$$J_i' = \sum_k T_{ik} J_k. \tag{10.10}$$

The coefficients T_{ik} in these linear combinations are matrix elements of the transmission matrix that are related to the elements of the scattering matrix S, namely, $T_{ik} = |S_{ik}|^2$. The matrix S expresses the outgoing wave amplitudes b_1', b_2', a_3', a_4' in terms of the incident ones b_1, b_2, a_3, a_4. To provide the charge conservation in the system, the net current in the channels 3 and 4 linking the system with the dephasing reservoir must be zero, so we may write

$$J_3 + J_4 - J_3' - J_4' = 0. \tag{10.11}$$

The transmission for quantum transport could be defined as the ratio of the particle flux outgoing from the system and that one incoming to the latter. Solving Equations 10.10 and 10.11 we obtain

$$T(E) = \frac{J_2'}{J_1} = T_{21} + \frac{K_1 \cdot K_2}{2R}, \tag{10.12}$$

where

$$K_1 = T_{31} + T_{41}; \quad K_2 = T_{23} + T_{24}; \qquad R = T_{33} + T_{44} + T_{43} + T_{34}. \tag{10.13}$$

For the junction including the single-site bridge the scattering matrix S has the form (Buttiker 1986)

$$S = \frac{1}{Z}\begin{pmatrix} r_1 + \alpha^2 r_2 & \alpha t_1 t_2 & \beta t_1 & \alpha\beta t_1 r_2 \\ \alpha t_1 t_2 & r_2 + \alpha^2 r_1 & \alpha\beta r_1 t_2 & \beta t_2 \\ \beta t_1 & -\beta r_1 t_2 & \beta^2 r_1 & \alpha r_1 r_2 - \alpha \\ \alpha\beta t_1 r_2 & \beta t_2 & \alpha r_1 r_2 - \alpha & \beta^2 r_2 \end{pmatrix}. \tag{10.14}$$

Here, $Z = 1 - \alpha^2 r_1 r_2$, $\alpha = \sqrt{1-\varepsilon}$, $\beta = \sqrt{\varepsilon}$, and $r_{1,2}$ and $t_{1,2}$ are the amplitude reflection and the transmission coefficients for the two barriers. Later, the expression for this matrix suitable for the case of multisite bridges including several inelastic scatterers was derived (Li and Yan 2001a,b).

Assuming for certainty the charge flow from the left to the right, we may write down the following expression (Zimbovskaya 2005):

$$T(E) = \frac{g(E)(1+\alpha^2)\left[g(E)(1+\alpha^2) + 1 - \alpha^2\right]}{\left[g(E)(1-\alpha^2) + 1 + \alpha^2\right]^2}, \tag{10.15}$$

where

$$g(E) = 2\sqrt{\frac{\Delta_L \Delta_R}{(E - E_0)^2 (\Delta_L + \Delta_R)^2}}. \tag{10.16}$$

Now, the electron transmission strongly depends on the dephasing strength ε. As shown in Figure 10.6, coherent transmission ($\varepsilon = 0$) exhibits a sharp peak at $E = E_0$ that gives a step-like shape to the volt–ampere curve, as was discussed in the previous section. In the presence of dephasing, the peak gets eroded. When the ε value approaches 1 the *I–V* curve becomes linear, corroborating the ohmic law for the inelastic transport.

Within Buttiker's model, ε is introduced as a phenomenological parameter whose relation to the microscopic characteristics of the dissipative processes affecting electron transport through molecular junctions remains uncertain. To further advance this model one should find out how to express ε in terms of the relevant microscopic characteristics for various transport mechanisms. This should open the way to a making of a link between the phenomenological Buttiker model and the NEGF. Such an attempt has been carried out in recent works (Zimbovskaya 2005, 2008) where the effect of stochastic nuclear motion on electron transport through molecules was analyzed.

10.4 Vibration-Induced Inelastic Effects

The interaction of electrons with molecular vibrations is known to be an important source of inelastic contribution to the electron transport through molecules. Theoretical studies of vibrationally inelastic electron transport through molecules and other similar nanosystems (e.g., carbon nanotubes) have been carried out over the past few years by a large number of researchers (Cornaglia et al. 2004, 2005, Donarini et al. 2006, Egger and Gogolin 2008, Galperin et al. 2007, Gutirrez et al. 2006, Kushmerick et al. 2004, Mii et al. 2003, Ryndyk and Cuniberti 2007, Siddiqui et al. 2007, Tikhodeev and Ueba 2004, Troisi and Ratner 2006, Zazunov and Martin 2007, Zazunov et al. 2006a,b, Zimmerman et al. 2008). Also, manifestations of the electron–vibron interactions were experimentally observed (Agrait et al. 2003, Djukic et al. 2005, Lorente et al. 2001, Qin et al. 2004, Repp et al. 2005a,b, Segal 2001, Smit et al. 2004, Tsutsui et al. 2006, Wang et al. 2004, Wu et al. 2004, Zhitenev et al. 2002). To analyze vibration induced effects on electron transport through molecular bridges, one must assume that molecular orbitals are coupled

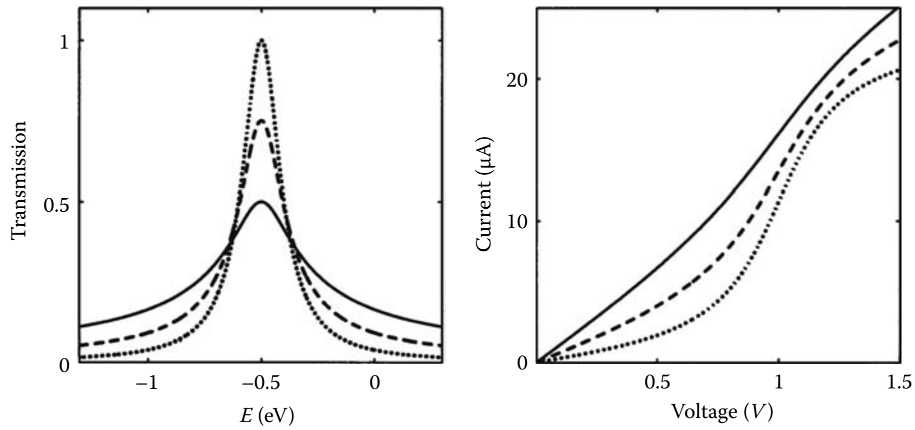

FIGURE 10.6 Electron transmission (left panel) and current (right panel) computed within the Buttiker model at various values of the dephasing parameter ε, namely, $\varepsilon = 0$ (dotted lines), $\varepsilon = 0.5$ (dashed lines), and $\varepsilon = 1$ (solid lines). The curves are plotted assuming that the molecule is simulated by a single orbital with $E_0 = -0.5$ eV at $\Delta_L = \Delta_R = 0.1$ eV, $T = 30$ K.

to the phonons describing vibrations. While on the bridge, electrons may participate in the events generated by their interactions with vibrational phonons. These events involve a virtual phonon emission and absorption. For rather strong electron–phonon interaction this leads to the appearance of metastable electron levels that could participate in the electron transport through the junctions, bringing an inelastic component to the current. As a result, vibration induced features occur in the differential molecular conductance dI/dV and in the IETS d^2I/dV^2. This was observed in experiments (see e.g., Qin et al. 2004, Zhitenev et al. 2002). Sometimes, even current voltage curves themselves exhibit an extra step originating from the electron–vibron interactions (Djukic et al. 2005).

Particular manifestations of electron–vibron interaction effects in the transport characteristics are determined by the relation of three relevant energies. These are the coupling strengths of the molecule to the electrodes $\Delta_{L,R}$, the electron–phonon coupling strength λ, and the thermal energy kT (k is the Boltzmann constant). When the molecule is weakly coupled to the electrodes ($\Delta_{L,R} \ll \lambda, \hbar\Omega$) and the temperature is low ($kT \ll \Delta_{L,R}$), the electron transfer through the junction may give rise to a strong vibrational excitation, and one may expect a pronounced vibrational resonance structure to appear in the electron transmission. Correspondingly, extra steps should occur in the I–V curves. A proper theoretical consideration of the electron transport in this regime could be carried out within the approach proposed by Wingreen et al. (1989). Here, we employ a very simple semi-quantitative approximation that, nevertheless, allows the qualitative description of this structure including the effect of higher phonon harmonics on the transport characteristics.

We consider a junction including a single-site bridge that is coupled to a single vibrational mode with the frequency Ω. An electron on the bridge may virtually absorb several phonons, which results in the creation of a set of metastable states with the energies $E_n = \tilde{E}_0 + n\,\hbar\Omega$ ($n = 0, 1, 2, \ldots$). Here, the energy $\tilde{E}_0$ is shifted with respect to E_0 due to the electron–phonon interaction. The difference in these energies E_p is called a polaronic shift and could be estimated as $E_p = \lambda^2/\hbar\Omega$ (Gutirrez et al. 2006, Ryndyk and Cuniberti 2007, Wingreen et al. 1989). At a weak coupling of the bridge state to the electrodes, the lifetime of these metastable states is long enough for them to serve as channels for the electron transmission. Therefore, one may roughly approximate the transmission as a sum of contributions from all these channels. The terms in the sum have a form similar to the well-known expression for coherent transmission (see Equation 10.9). However, every term includes the factor $P(n)$ that corresponds to the probability of the appearance of the metastable state. So, we obtain

$$T(E) = 2\Delta_L\Delta_R \sum_n \frac{P(n)}{(E - \tilde{E}_0 - n\hbar\Omega)^2 + (\Delta_L + \Delta_R)^2}. \tag{10.17}$$

Here (Cizek et al. 2004):

$$P(n) = \frac{1}{n!}\left(\frac{\lambda^2}{2\hbar^2\Omega^2}\right)^n \exp\left(-\frac{\lambda^2}{2\hbar^2\Omega^2}\right). \tag{10.18}$$

The phonon-induced peaks in the transmission are displayed in Figure 10.7 along with the transmission peak for coherent transport through a single-site bridge. As expected, the coupling of the electronic degrees of freedom to the vibrational motion splits the single peak in the coherent transmission into a set of smaller peaks associated with vibrational levels. The peaks could be resolved when $\Delta_{L,R} \ll \hbar\Omega$. This agrees with the results of the earlier theoretical works (Wingreen et al. 1989), as well as with the experiments (Qin et al. 2004, Zhitenev et al. 2002). Phonon-induced peaks in the transmission give rise to the steps in the I–V curves and rather sharp features (peaks and dips) in the IETS. The latter are shown in Figure 10.8 (left panel) and they resemble those obtained using proper NEGF-based calculations (Galperin et al. 2004).

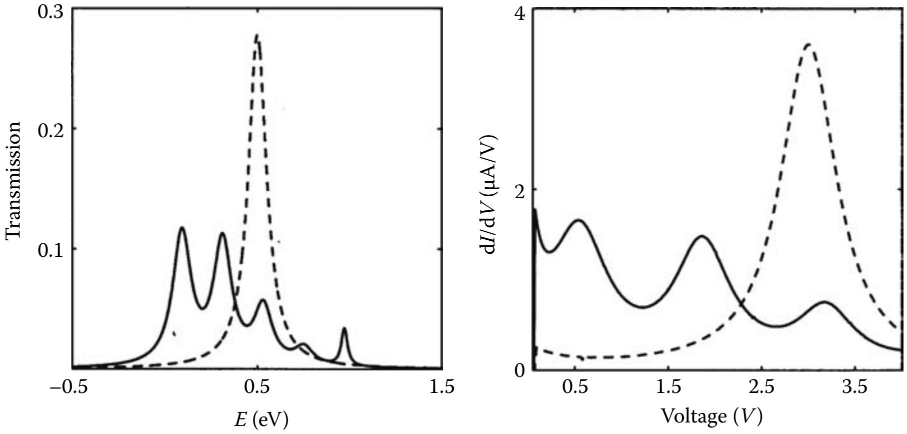

FIGURE 10.7 The electron transmission versus energy (left panel) and the conductance versus voltage (right panel) for a junction with the molecular bridge simulated by a single electronic state weakly coupled to the leads: $\Delta_L = \Delta_R = 0.01$ eV, $E_0 = 0.5$ eV. Solid lines are plotted assuming that the bridge is coupled to a single phonon mode ($\hbar\Omega = 0.22$ eV, $\lambda = 0.3$ eV). Dashed lines correspond to the coherent electron transport.

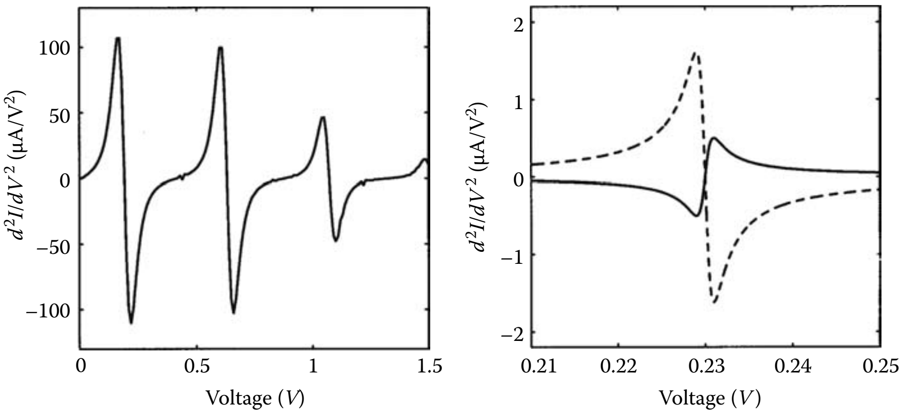

FIGURE 10.8 The inelastic electron tunneling spectrum plotted against the bias voltage at $\hbar\Omega = 0.22\,\text{eV}$, $\lambda = 0.3\,\text{eV}$. Left panel: $\Delta_L = \Delta_R = 0.01\,\text{eV}$, $E_0 = 0.5\,\text{eV}$. Right panel: $\Delta_L = \Delta_R = 0.5\,\text{eV}$, $E_0 = 0.5\,\text{eV}$ (solid line), $E_0 = -0.5\,\text{eV}$ (dashed line).

When the molecule is strongly coupled to the electrodes ($\Delta_{L,R} \gg \lambda$) and the temperature is still low ($kT \ll \lambda$, $\hbar\Omega$), electron–vibron interaction effects are less pronounced. Both the current and the conductance are weakly affected by the electron–phonon coupling (Galperin et al. 2007, Tikhodeev and Ueba 2004). However, the IETS features remain distinguishable. These features appear at the threshold $V = \hbar\Omega/|e|$ that corresponds to the opening of a channel for inelastic transport. To analyze IETS in the simplest way one may use the result for electron transmission derived within the Buttiker model where the dephasing parameter ε is expressed in terms of the relevant energies, namely,

$$\varepsilon = \frac{\Gamma_{ph}}{2\left(\Delta_L + \Delta_R\right) + \Gamma_{ph}}, \tag{10.19}$$

where $\Gamma_{ph} = 2\mathrm{Im}(\Sigma_{ph})$, and Σ_{ph} is the self-energy term originating from the electron–phonon interaction. Based on the nonequilibrium Green's function formalism, the expression for Γ_{ph} was derived in the form (Mii et al. 2003)

$$\begin{aligned}\Gamma_{ph}(E) = 2\pi\lambda^2 \int d\omega\rho(\omega)\big\{&N(\omega)[\rho_{el}(E-\hbar\omega)\\ &+\rho_{el}(E+\hbar\omega)] + \left(1+N(\omega)\right)\left(\left[1-n(E-\hbar\omega)\right]\rho_{el}(E-\hbar\omega)\right.\\ &\left.+n(E+\hbar\omega)\rho_{el}(E+\hbar\omega)\right) + N(\omega)\left(\left[1-n(E+\hbar\omega)\right]\rho_{el}(E+\hbar\omega)\right.\\ &\left.+n(E-\hbar\omega)\rho_{el}(E-\hbar\omega)\right)\big\}\,.\end{aligned} \tag{10.20}$$

Here, $\rho_{el}(E)$ and $\rho_{ph}(\omega)$ are the phonon and the electron densities of states, respectively; $N(\omega)$ is the Bose–Einstein distribution function at the temperature T, and

$$n(E) = \frac{1}{2}\left[f_L(E) + f_R(E)\right]. \tag{10.21}$$

While considering a junction with a single-site bridge state coupled to the single vibrational mode, we can write the following expressions for ρ_{ph} and ρ_{el}:

$$\rho_{ph}(\omega) = \frac{1}{\pi\hbar}\frac{\gamma}{(\omega-\Omega)^2 + \gamma^2}, \tag{10.22}$$

$$\rho_{el}(E) = \frac{1}{\pi}\frac{\Delta_L + \Delta_R}{(E-E_0)^2 + (\Delta_L + \Delta_R)^2}, \tag{10.23}$$

where the polaron shift is neglected, and the parameter γ characterizes the broadening of the maximum in ρ_{ph} at $\omega = \Omega$ due to the interaction of the vibronic mode with the environment. At low temperatures, we may significantly simplify the expression for Γ_{ph}. Within the conduction window $\mu_R < E < \mu_L$, we get

$$\Gamma_{ph}(E) \approx \pi\lambda^2 \left\{ \int_0^{(\mu_L - E)/\hbar} d\omega\rho_{ph}(\omega)\rho_{el}(E+\hbar\omega) + \int_0^{(E-\mu_R)/\hbar} d\omega\rho_{ph}(\omega)\rho_{el}(E-\hbar\omega) \right\}. \tag{10.24}$$

Omitting from consideration the coupling of the phonon mode to the environment ($\gamma \to 0$), we may easily carry out the integration over ω in Equation 10.24, and we arrive at the result

$$\begin{aligned}\Gamma_{ph}(E) \approx \pi\lambda^2\{&\rho_{el}(E+\hbar\Omega)\theta(\mu_L - \hbar\Omega - E)\\ &+\rho_{el}(E-\hbar\omega)\theta(E-\mu_R-\hbar\Omega)\},\end{aligned} \tag{10.25}$$

where $\theta(x)$ is the step function.

Substituting the approximation for Γ_{ph} given by Equations 10.24 and 10.25 into the expression (10.19) for the dephasing parameter $\int$ and employing the earlier result (10.15) for the electron transmission, we may calculate the transport characteristics. The adopted simplified approach, as well as the NEGF-based calculations, shows that the IETS signal appears at the threshold determined by the frequency of the vibrational mode (see Figure 10.8, right panel). At first glance, one may expect the net current through the junction to increase at the threshold. Indeed, the inelastic contribution to the current increases from zero to a certain nonzero value at this threshold for the channel for inelastic transport to open up. However, more thorough studies show that both elastic and nonelastic contributions to the net current undergo changes at the inelastic tunneling threshold, and the elastic current could decrease there, as was first shown by Persson and Baratoff (1987). Moreover, this decrease in the elastic current may overweigh the contribution coming from the inelastic channel. Depending on the relative value of the elastic and the inelastic contributions to the net current near threshold, the IETS reveals a peak or a dip at the corresponding voltage. Experiments corroborate the variety in the IETS taken for molecular junctions (Djukic et al. 2005, Hahn et al. 2000, Wang et al. 2004, Zhitenev et al. 2002). The shape of the signal is very sensitive to the characteristics of the junction, such as the position of the electronic state, the electron–phonon and the molecule-to-leads coupling strengths, and the vibron frequency. For instance, at a very strong coupling of the molecular bridge to the leads, it could so happen that the backscattering by the negatively biased electrode electrons whose energies belong to the conduction window between μ_L and μ_R, are locally depleted near the junction. In this situation, the opening of the inelastic channel may cause the increased reflection (otherwise forbidden by the Fermi exclusion) leading to the decrease in the conduction. Consequently, the described scenario should result in the dip in the IETS signal. Also, as concluded in the recent work of Ryndyk and Cuniberti (2007, the sideband phonon-induced features in the electron spectral density discussed above could give rise to the corresponding features in the differential conductance dI/dV and IETS assuming that the molecule coupling to the leads is not too strong. Contributions from these sideband features may be responsible for the shape of the IETS signal at the threshold of the inelastic tunneling channel. These contributions could produce an extra inelastic signal, as well. The latter appears as an additional peak or a dip in the differential conductance.

The question of the current decrease/increase at the phonon excitation threshold that corresponds to the peak/dip in the IETS is not completely answered so far, and the appropriate theory is being developed (Balseiro et al. 2006, Egger and Gogolin 2008). Nevertheless, it is presently understood that three relevant energies, namely, molecule–electrode couplings $\Delta_{L,R}$, electron–phonon coupling strength λ, and the phonon energy $\hbar\Omega$ play a very important part in determining the shape of the IETS signal. Varying these parameters one may convert the IETS signal from a peak to a dip and vice versa.

At finite temperatures, molecular vibrations always occur in the presence of stochastic nuclear motions. These motions could be described as a phonon thermal bath. Coupling of vibrational modes to this bath further affects the electron transport causing energy dissipation. The dissipative processes must be taken into account to properly analyze the effects of electron–phonon interactions in the electron transport. Also, the displacements of ions involved in the molecular vibrations are accompanied by the changes in the electrostatic field inside the molecule. This could give rise to polaronic effects in electron transport. We discuss these issues in the next sections.

10.5 Dissipative Transport

Electron transport through molecular junctions is always accompanied by stochastic nuclear/ion motions in the close environment. Interactions of traveling electrons with these environmental fluctuations cause energy dissipation. The importance of dissipative effects depends on several factors. Among these factors the temperature, the size, and the complexity of the molecular bridge predominate. The temperature determines the intensity of the nuclear motions, and the size of the molecule determines the so-called contact time, that is, the time for an electron to travel through the junction, and in consequence, to contact the environment. It was shown by Buttiker and Landauer (1985) that the contact time is proportional to the number of sites (subunits) in the molecule providing intermediate states for the electron tunneling. For small molecules at low temperatures, the contact time is shorter than the characteristic times for fluctuations in the environment, so, the effect of the latter on the electron transport is not very significant. One may expect a small broadening of the molecule energy states to occur that brings a moderate erosion of the peaks in the electron transmission and the steps in the *I–V* characteristics.

On the contrary, in large-sized molecules such as proteins and DNA, electron transport is accompanied by strong energy dissipation. The significance of the system-environment interactions in macromolecules was recognized long ago in studies of long-range electron-transfer reactions. In these reactions, electrons travel between distant sites on the molecule called a donor and an acceptor. It was established that when an electron initially localized on the donor site moves to the acceptor site with a lower energy, the energy difference must be dissipated to the environment to provide the irreversibility of the transfer (Garg et al. 1985).

A usual way to theoretically analyze dissipative effects in the electron transport through molecules is to introduce a phonon bath representing the random motions in the environment. In general, there is no one/several dominating modes in the bath. Instead, the bath is characterized by a continuous spectral function of all relevant phonon modes $\rho_{ph}(\omega)$. The electrons are supposed to be coupled to the phonon bath, and this coupling is specified by the spectral function. The particular form of $\rho_{ph}(\omega)$ may be found based on molecular dynamics simulations. However, to qualitatively study the effect of dissipation on electron transport one can employ the expression (Mahan 2000)

$$\rho_{ph}(\omega) = \lambda \frac{\omega}{\omega_c} \exp\left(-\frac{\omega}{\omega_c}\right), \qquad (10.26)$$

where

the parameter λ characterizes the electron–phonon coupling strength

ω_c is the cutoff frequency for the bath related to its thermal relaxation time $\tau_c = \omega_c^{-1}$

To illustrate the possible effects of dissipation on the electron transport we return to our simple model where the molecular bridge is represented by a single state. Now, we assume that this state is coupled to the phonon bath. This model is hardly appropriate to properly analyze the dissipative effects in the electron transport through practical molecules for the molecule length is very important for the dissipative effects to be pronounced. Nevertheless, it still could serve to basically outline the main features of the dissipative electron transport through molecular junctions. Also, the proposed model could be useful to analyze electron transport in doped polyacetylene/polyaniline–polyethylene oxide nanofibers (Zimbovskaya 2008). These conducting polymers could be treated as some kind of granular metals where metallic-like regions (grains) are embedded in the poorly conducting medium of the disorderly arranged polymer chains (MacDiarmid 2001). While in the metallic state, the intergrain electron transport in these nanofibers is mostly provided by electron tunneling through the intermediate states on the polymer chains between the grains (Prigodin and Epstein 2002). In this case, the contact time could be long enough for the effects of dissipation to be well manifested that justifies the adoption of the above model.

Again, one may carry out transport calculations using Equation 10.15 for the electron transmission, and by expressing ε in terms of the relevant energies. Substituting Equation 10.26 in the expression (10.20), we may calculate $\Gamma_{ph}(E)$. The energy dissipation effects are more distinctly pronounced at moderately high temperatures, so we assume $kT \gg \hbar\omega_c$. Then the main contribution to the integral over ω in Equation 10.20 originates from the low frequency region $\omega \ll \omega_c$; and we obtain the following approximation:

$$\Gamma_{ph}(E) = \frac{8kT\lambda\Gamma}{(E - E_0)^2 + \Gamma^2}, \qquad (10.27)$$

where $\Gamma = \Delta_L + \Delta_R + (1/2)\Gamma_{ph}(E)$. Solving the obtained equation for Γ_{ph} and using Equation 10.19, we obtain (Zimbovskaya 2008)

$$\varepsilon = \frac{1}{2} \frac{\rho^2\left(1 + \sqrt{1 + \rho^2}\right)}{4\left(\dfrac{E - E_0}{\Delta_L + \Delta_R}\right)^2 + \dfrac{1}{2}\left(1 + \sqrt{1 + \rho^2}\right)^3}, \qquad (10.28)$$

where $\rho^2 = 32kT\lambda/(\Delta_L + \Delta_R)^2$.

Voltage dependencies of the conductance computed based on Equations 10.3, 10.15, and 10.28 are presented in Figure 10.10 (left panel). One may see that at low values of the bias voltage, the electrons coupling to the phonon bath bring an enhancement in the conduction. The effect becomes reversed as the voltage grows above a certain value. This happens because the phonon-induced broadening of the molecular level (the bridge) assists the electron transport at a small bias voltage. When the voltage rises, this effect is surpassed by the scattering effect of phonons that resist electron transport. The significance of the electron–phonon interactions is determined by the ratio of the coupling constant λ and the self-energy terms describing the bridge coupling to the electrodes. The phonon bath makes a significant effect on the transport characteristics when $\lambda > \Delta_{L,R}$.

Dissipative electron transport through large DNA molecules was studied both theoretically and experimentally (Gutierrez et al. 2005, Xu et al. 2004). Theoretical studies were based on a model where the molecule was simulated by a tight-binding chain of sites linking the electrodes and the attached side chains. Electrons were allowed to travel along the bridge chain and to hop to the nearby side chains. These chains were coupled to the phonon bath providing the energy dissipation (see Figure 10.9). Although proposed for specific kinds of poly (*dG*)-poly (*dC*) molecules, this model seems to be quite generic and useful for a larger class of macromolecules.

Several coupling regimes to the bath may be analyzed. The most preferred regime for dissipative effects to appear is the strong-coupling limit defined by the condition $\lambda/\omega_c > 1$. Within this regime, the characteristic time for the electron bath interactions is much shorter than the typical electron time scales. Consequently, the bath makes a significant impact on the molecule electronic structure. New bath-induced states appear in the molecular spectrum inside the HOMO–LUMO gap. However, these states are strongly damped due to the dissipative action of the bath (Gutierrez et al. 2005). As a result, a small finite density of phonon-induced states appears inside the gap supporting electron transport at a low bias voltage. So again, the environment induces incoherent phonon-assisted transport through molecular bridges. For illustration, we show here the results of

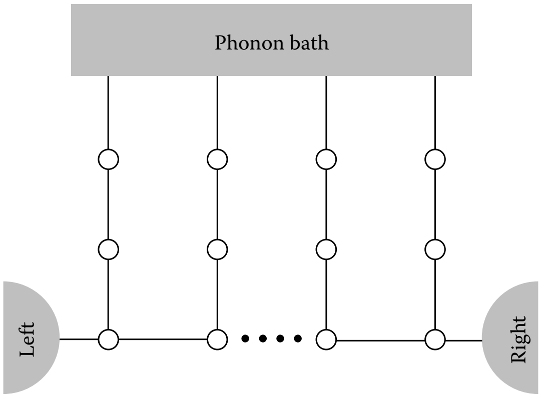

FIGURE 10.9 Schematic drawing of a molecular junction where the molecular bridge is coupled to the phonon bath via side chains.

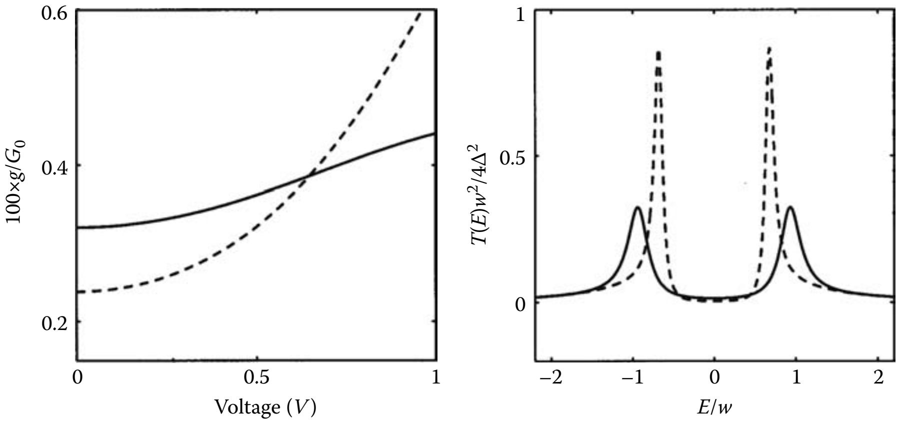

FIGURE 10.10 Left panel: The electron conductance versus voltage for a junction with a single electronic state bridge directly coupled to the phonon thermal bath. The curves are plotted assuming $E_0 = 0.4\,\mathrm{eV}$, $\Delta_L = \Delta_R = 0.1\mathrm{eV}$, $T = 30\,\mathrm{K}$, $\lambda = 0.3\,\mathrm{eV}$ (solid line), $\lambda = 0.05\,\mathrm{eV}$ (dashed line). Right panel: Electron transmission through the junction in the case when the bridge state interacts with the phonon bath via the side chain coupled to the bridge state with the coupling parameter w. The curves are plotted assuming $\Delta_L = \Delta_R = \Delta$, $w/\Delta = 20$, $E_0 = 0$, $\lambda = 0.3\,\mathrm{eV}$ (solid line), $\lambda = 0.05\,\mathrm{eV}$ (dashed line).

calculations carried out for a toy model with a single-site bridge with a side chain attached to the latter. The side chain is supposed to be coupled to the phonon bath. The results for the electron transmission are displayed in Figure 10.10 (right panel). We see that the original bridge state at $E = 0$ is completely damped but two new phonon-induced states emerge nearby that could support electron transport.

An important characteristics of the dissipative electron transport through molecular junctions is the power loss in the junction, that is, the energy flux from the electronic into the phononic system. Assuming the current flow from the left to the right, this quantity may be estimated as the sum of the energy fluxes $Q_{L,R}$ at the left and the right terminals (leads):

$$P = Q_L + Q_R. \tag{10.29}$$

One may express the energy fluxes in terms of renormalized currents at the electrodes $\tilde{I}_{L,R}(E)$ that are defined as follows (Datta 2005):

$$I_{L,R} = \frac{e}{\pi\hbar}\int \tilde{I}_{L,R}(E)dE. \tag{10.30}$$

Then $Q_{L,R}$ may be presented in the form

$$Q_{L,R} = \frac{1}{\pi\hbar}\int E\tilde{I}_{L,R}(E)dE. \tag{10.31}$$

We remark that the current $I_{L,R}$ in Equation 10.30 are related to the corresponding leads, and their signs are accordingly defined. An outgoing current is supposed to be positive whereas an incoming one is negative for each lead. To provide an electric charge conservation in the junction one must require that $I = I_L = -I_R$ for the chosen direction of the current flow. Therefore, the energy fluxes also have different signs. As for the $Q_{L,R}$ magnitudes, they may differ only if the renormalized currents $\tilde{I}_L(E)$ and $\tilde{I}_R(E)$ are distributed over energies in different ways. This cannot happen in the case of elastic transport, for in this case, $\tilde{I}_L(E) = -\tilde{I}_R(E)$. However, if the transport process is accompanied with the energy dissipation, the energy distributions for $\tilde{I}_L$ and $\tilde{I}_R$ may differ. In this case, electrons lose some energy while moving through the junction and this gives rise to the differences in the renormalized currents' energy distributions. For instance, in the case when the electrodes are linked with a single-state bridge, the energy distribution of $\tilde{I}_L(E)$ has a single maximum whose position is determined by the site energy E_0 and the applied bias voltage V. Assuming that the average energy loss due to dissipation could be estimated as ΔE, the maximum in the $\tilde{I}_R(E)$ distribution is shifted by this quantity, so the current $\tilde{I}_R(E)$ flows at lower energies compared to $\tilde{I}_L(E)$. This results in the power loss and the Joule heating in the junction (Segal et al. 2003).

10.6 Polaron Effects: Hysteresis, Switching, and Negative Differential Resistance

While studying electron transport through molecular junctions, hysteresis in the current–voltage characteristics was reported in some systems (Li et al. 2003). Multistability and stochastic switching were reported in single-molecule junctions (Lortscher et al. 2007) and in single metal atoms coupled to a metallic substrate through a thin ionic insulating film (Olsson et al. 2007, Repp et al. 2004).

The coupling of an electron belonging to a certain atomic energy level to the displacements of ions in the film brings a possibility of polaron formation in there. This leads to the polaron shift in the electron energy. It was noticed that multistability and hysteresis in molecular junctions mostly occurred when the molecular bridges included centers of long-living charged electronic states (redox centers). On these grounds, it was suggested that hysteresis in the I–V curves observed in molecular junctions appear due to the formation of polarons on the molecules (Galperin et al. 2005).

The presence of the polaron shift in the energy of a charged (occupied) electron state creates a difference between the latter and the energy of the same state while it remains unoccupied. Assuming, for simplicity, a single-state model for the molecular bridge coupled to a single optical phonon mode, we may write the following expression for the renormalized energy:

$$\tilde{E}_0(n_0) = E_0 - \frac{\lambda^2 n_0}{\hbar\Omega}, \tag{10.32}$$

where the electronic population on the bridge n_0 is given by

$$n_0 = \frac{1}{\pi}\int dE \frac{f_L(E)\Delta_L + f_R(E)\Delta_R}{\left[E - \tilde{E}_0(n_0)\right]^2 + \left(\Delta_L + \Delta_R\right)^2}. \tag{10.33}$$

So, as follows from Equation 10.32, the polaron shift depends on the bridge occupation n_0, and the latter is related to $\tilde{E}_0$ by Equation 10.33. Therefore, the derivation of an explicit expression for $\tilde{E}_0(n_0)$ is a nontrivial task even within the chosen simple model. Nevertheless, it could be shown that two local minima emerge in the dependence of the potential energy of the molecular junction including two electrodes linked by the molecular bridge, of the occupation number n_0. These minima are located near $n_0 = 0$ and 1, and they correspond to the neutral (unoccupied) and charged (occupied) states, respectively. This is illustrated in Figure 10.11 (left panel). These states are metastable, and their lifetime could be limited by the quantum switching (Mitra et al. 2005, Mozyrsky et al. 2006). When the switching time between the two states is longer than the characteristic time for the external voltage sweeping, one may expect the hysteresis to appear, for the states of interest live long enough to maintain it. Within the opposite limit, the average washes out the hysteresis. When the states are especially short-lived this could even result in a telegraph noise at a finite bias voltage that replaces the controlled switching.

Further, we consider long-lived metastable states and we concentrate on the I–V behavior. Let us for certainty assume that the bridge state at zero bias voltage is situated above the Fermi energy of the system and remains empty. As the voltage increases, one of the electrode's chemical potentials crosses the bridge level position, and the current starts to flow through the system. The I–V curve reveals a step at the bias voltage value corresponding to the crossing of the unoccupied bridge level with the energy E_0 by the chemical potential. However, while the current flows through the bridge, the level becomes occupied and, consequently, shifted due to the polaron formation. When the bias voltage is reversed, the current continues to flow through the shifted bridge state of the energy $\tilde{E}_0$, until the recrossing happens. Due to the difference in the energies of the neutral and the charged states, the step in the I–V curves appears at different values of the voltage, and this is the reason for the hysteresis loop to appear as shown on the right panel of Figure 10.11. One could also trace the hysteresis in the I–V characteristics starting from the filled (and shifted) bridge state.

Again, the hysteresis loop in the I–V curves may occur when both occupied and unoccupied states are rather stable, which means that the potential barrier separating the corresponding minima in the potential energy profile is high enough, so that quantum switching between the states is unlikely. This happens when the bridge is weakly coupled to the electrodes ($\Delta_{L,R} \ll \lambda^2/\hbar\Omega$), which is an obvious requirement for the involved states (neutral or charged) to the distinguishable. In other words, the broadening of the relevant levels due to the coupling to the electrodes must be much smaller that the polaron shift. As was recently shown (Ryndyk et al. 2008), the stronger the electron–phonon coupling, the less becomes the probability of the switching. At large values of λ, the tunneling between the charged and

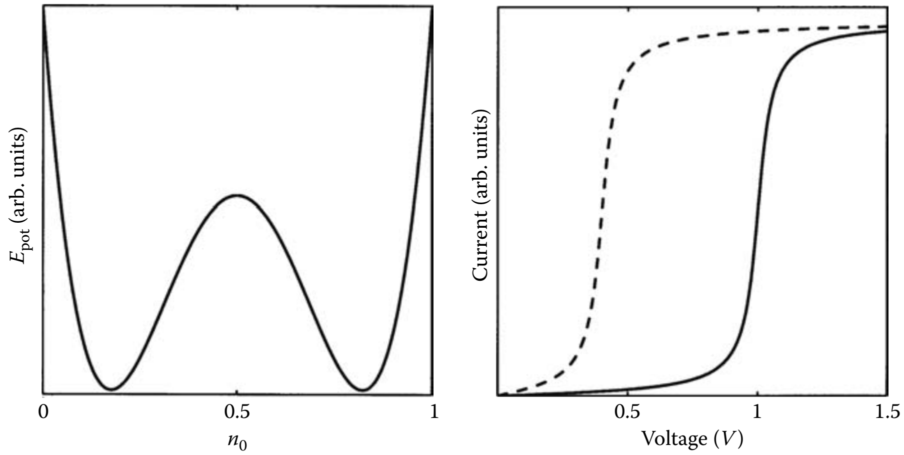

FIGURE 10.11 Left panel: Schematic of the potential energy of the molecular junction versus the occupation number n_0. Right panel: Hysteresis in the current–voltage characteristics. The solid line corresponds to the transport via the unoccupied bridge electronic state, and the dashed line corresponds to the transport via the occupied state shifted due to the polaron formation.

the neutral states is exponentially suppressed. Also, it was shown that the symmetry/asymmetry in the coupling to the electrodes could significantly affect the hysteresis behavior (D'Amico et al. 2008, Ryndyk et al. 2008). For asymmetric junctions ($\Delta_L \neq \Delta_R$), two nearly stable states exist at the zero bias voltage. When the asymmetry is very strong, both states could appear stable at one bias voltage polarity and unstable when the polarity is reversed. It was suggested that under an appropriate choice of parameters, one may create a situation where the instability regions for the involved states do not overlap. These properties give grounds to conjecture that such strongly asymmetric junctions could reveal memory functionalities, that make them potentially useful in the fabrication of nanodevices.

Among the various potentially important properties of the electron transport through metal–molecule junctions, one may separate the negative differential resistance (NDR), that is, the decrease in the current I while the bias voltage across the molecule increases. The NDR effect was originally observed in tunneling semiconducting diodes (Kastner 1992). Later, it was viewed in quantum dots (Grobis et al. 2005, Repp et al. 2005a,b). Several possible scenarios were proposed to explain the NDR occurrence in electron transport through molecules. The effect could originate from the alignment and the subsequent disalignment of the Fermi levels of the electrodes with the molecular orbitals that happens as the bias voltage varies (Xue et al. 1999). Also, the NDR could appear as a Coulomb blockade-induced effect (Simonian et al. 2007) and/or due to some other reasons (Zimbovskaya and Pederson 2008). It is likely that different mechanisms could play a major part in the NDR appearance in different molecular junctions where it was observed so far.

Here, we discuss the NDR features in the current–voltage characteristics that originate from the electron coupling with the molecular vibrational modes. As was shown (Galperin et al. 2005, Yeganeh et al. 2007a,b), the polaron formation could give rise to the NDR. This may happen if the polaron shift in the energy of the occupied state moves the energy level away from the conduction window. As the shift depends on the electronic population, and the latter changes as the bias voltage increases, the occupied state falls out from the window between the chemical potentials of the electrodes at a certain value of the voltage. If the transport is being conducted via this occupied state, then it stops when this voltage value is reached. Correspondingly, the current value drops to zero revealing a distinctive NDR feature. This is a realistic scenario based on the main features of the electron transport through molecules under polaron formation.

In conclusion, electron–vibrion coupling could lead to the formation of a polaron that results in the energy differences between the occupied and the unoccupied states on the molecular bridge. At a weak interaction between the bridge and the electrodes and the strong electron–phonon coupling these charged and neutral states are metastable and could serve for electron conduction. This results in such interesting and potentially useful effects as the hysteresis and the NDR features in the current–voltage characteristics.

10.7 Molecular Junction Conductance and Long-Range Electron-Transfer Reactions

Long-range electron-transfer reactions play an important part in many biological processes such as photosynthesis and enzyme catalyses (Kuznetsov and Ulstrup 1999). Theoretical and experimental studies of these processes have lasted for more than four decades but they still remain within a very active research area. In the intramolecular electron-transfer reactions, the electric charge moves from one section of the molecule to another section of the same molecule. Long-range transfer typically occurs in large molecules such as proteins and/or DNA, so that these two sections are situated far apart from each other. A common setup for the transfer reactions includes a donor, a bridge, and an acceptor. Due to the large distances between the donor (where an electron leaves) and the acceptor (where it arrives) typical for the long-range transfer reactions, a direct coupling between the two is negligible. Therefore, the electron participating in the transfer needs a molecular bridge providing a set of intermediate states for electron transport.

Essentially, intramolecular electron transfer is a combination of nuclear environment fluctuations and electron tunneling. Electron-transfer reactions result from the response of a molecule to environmental polarization fluctuations that accompany nuclear fluctuations. The molecule responds by redistribution of the electronic density thus establishing opportunities for the charge transfer to occur. The main characteristic of the electron-transfer processes is the transfer rate K_{et} that is the inversed time of the reaction. The transfer rate for the electron tunneling in the molecule interior from the donor to the acceptor depends both on the electronic transmission amplitudes and the vibrionic coupling. The latter provides the energy exchange between the electronic and the nuclear systems.

Viewing fast electronic motions in the background of the slowly moving nuclei (nonadiabatic electron transfer) and applying the Fermi golden rule of perturbation theory, it was shown that K_{et} could be written in the form first suggested by Marcus (1965)

$$K_{et} = \frac{2\pi}{\hbar}\left|H_{\mathrm{DA}}\right|^2 (FC). \tag{10.34}$$

Here, the first cofactor is the electron transmission coefficient, H_{DA}, which is the effective matrix element between the donor and the acceptor. The second term is the density of states weighted Franck–Condon factor that describes the effect of nuclear motions in the environment.

There exists a noticeable resemblance between the long-range electron transfer and the molecular conductance, and this resemblance was analyzed in several theoretical works (Dahnovsky 2006, Mujica et al. 1994, Nitzan 2001, Yeganeh et al. 2007a,b, Zimbovskaya 2003, Zimbovskaya and Gumbs 2002). The fundamental part in both molecular conduction and electron-transfer reactions is taken by the electron tunneling inside the molecule via the set of intermediate states. In large molecules, the electron involved in the transfer reaction or contributing to the conductance, with a high probability follows a few pathways that work as molecular bridges. The rest of the molecules do not significantly participate in the transfer/conduction and may be omitted from consideration in calculations of electron transmission (Jones et al. 2002; Beratan et al. 1992). On these grounds, it was suggested that the original molecule could be reduced to a simpler chain-like structure, and the latter may replace the former in calculations of molecular conductance and/or transfer rates (Daizadeh et al. 1997). Such simplification significantly eases computations of the molecular conductance and the electron-transfer rates in macromolecules.

The resemblance between the molecular conductance and the long-range electron transfer does not mean that these processes are nearly identical. Along with the similarities there are substantial differences between the former and the latter. For instance, the continuum of states causing the charge transport arises from the multitude of electronic states on the electrodes in the case of molecular conduction, and from the vibrations and fluctuations in the environment in the case of electron-transfer reactions. Also, the driving force that puts electrons in motion originates from the external bias voltage applied across the junction in the case of molecular conduction. In the electron-transfer situation, this force appears due to the electron–vibrionic interactions in the system, or the electron-transfer reaction starts as a result of a photoexcitation of the donor part of the molecule. The observables, namely, the molecular conductance g and/or current I, and the transfer rate K_{et} differ as well.

Notwithstanding these differences, the electron transport through molecular junctions and the electron transfer could be theoretically analyzed using the common formalism as recently proposed by Yeganeh et al. (2007b). Further, we follow their approach. The starting point is that one may simulate the donor–bridge–acceptor system as some kind of molecular junction, shown in Figure 10.12. Here, we represent the donor as a single state $|i\rangle$, and this state mimics the left lead (assuming the transport is from the left to the right). The donor is coupled to the bridge and the latter is coupled to the continuum of the final states $|f\rangle$ simulating the right lead. These couplings are described by the self-energy terms $\Delta_{i,f}$. For simplicity, we assume that there exists a single state on the bridge, and this state is coupled to a vibrionic mode of the frequency Ω. The latter mimics environment fluctuations. We assume that the "leads" are weakly coupled to the bridge ($\Delta_{L,R} \ll \lambda$), which is typical for electron-transfer situations. Also, the representation of the environment motions by the single vibrionic mode is justified only at low temperatures when the thermal energy kT is much smaller than the electron–vibrion coupling parameter λ.

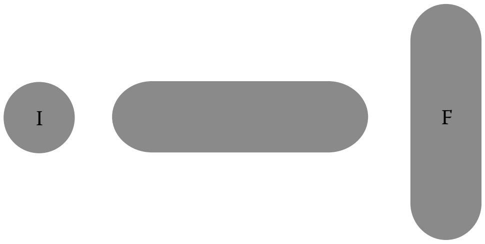

FIGURE 10.12 Schematic of the model system used in the transfer rate calculations. The initial/final reservoirs correspond to the left/right leads in the molecular junction.

Now, we can write the Landauer expression for the current flowing through the "junction" using Equations 10.3 and 10.4. Assuming that the bridge includes only one state, we simplify Equation 10.4:

$$T(E) = 2T_r\left\{\Delta_i G \Delta_f G^\dagger\right\} = -2\frac{\Delta_i \Delta_f}{\Delta_i + \Delta_f}\,\mathrm{Im}(G). \qquad (10.35)$$

Here, the subscripts i/f label the initial and the final states. Within the Fermi golden rule regime that allows the introduction of the transfer rate, the bridge must be coupled to the final states much more strongly than to the initial state ($\Delta_f \gg \Delta_i$). Therefore

$$\frac{\Delta_i \Delta_f}{\Delta_i + \Delta_f} \approx \Delta_i \qquad (10.36)$$

and the coupling to the final states falls out of the expressions for electron transmission and current. Within the chosen model, this is a physically reasonable result for the final states reservoir (the right "electrode" in our junction) was merely introduced to impose a continuum of states maintaining the transfer process at a steady-state rate. Also, considering the current flow, we may suppose that the initial state is always filled [$f_i(E) = 1$], and the final states are empty [$f_f(E) = 0$]. Therefore, the current flow through the "junction" accepts the form

$$I = \frac{2e}{\pi\hbar}\int dE\Delta_i\,\mathrm{Im}(G). \qquad (10.37)$$

Both the current and the transfer rate are fluxes closely related to each other, namely,

$K_{et} = I/e$. So, we may write

$$K_{et} = -\frac{2}{\pi\hbar}\int dE\Delta_i\,\mathrm{Im}(G). \qquad (10.38)$$

Now, Δ_i could be computed using the expression (10.7) for the corresponding self-energy term. Keeping in mind that the "left electrode" includes a single state with the certain energy $\int_i$ we obtain

$$\Delta_i = \mathrm{Im}\left(\frac{|\tau_i|^2}{E - \varepsilon_i + is}\right) = \pi|\tau_i|^2\,\delta(E - \varepsilon_i). \qquad (10.39)$$

Accordingly, the expression (10.38) for the transfer rate may be reduced to the form:

$$K_{et} = -\frac{2}{\hbar}|\tau_i|^2\,\mathrm{Im}\left[G(\varepsilon_i)\right], \qquad (10.40)$$

where τ_i represents the coupling between the donor and the molecule bridge. It must be stressed that within the chosen model, τ_i is the only term representing the relevant state coupling that may be identified with the electronic transmission coefficient H_{DA} in the general expression (10.34) for K_{et}. This leaves us with the following expression for the Franck–Condon factor:

$$(FC) = -\frac{1}{\pi}\mathrm{Im}\left[G(\varepsilon_i)\right]. \tag{10.41}$$

As discussed in Section 10.4, at a weak coupling of the bridge to the leads the electron–vibrionic interaction opens the set of metastable channels for the electron transport at the energies $E_n = \tilde{E}_0 + n\hbar\Omega$ ($n = 0, 1, 2\ldots$) where $\tilde{E}_0$ is the energy of the bridge state with the polaronic shift included. Green's function may be approximated as a weighted sum of contributions from these channels:

$$G(\varepsilon_i) = \sum_{n=0}^{\infty} P(n)\left[\varepsilon_i - \tilde{E}_0 - n\hbar\Omega + is\right]^{-1}, \tag{10.42}$$

where $s \to 0^+$, and the coefficients $P(n)$ are probabilities for the channels to appear given by Equation 10.18.

Substituting Equation 10.42 into Equation 10.41 we get

$$(FC) = \sum_{n=0}^{\infty} P(n)\delta(\Delta F - n\hbar\Omega). \tag{10.43}$$

Here, $\Delta F = \int_i - \tilde{E}_0 \equiv \int_i - E_0 + \lambda^2/(\hbar\Omega)$ is the exoergicity of the transfer reaction, that is, the free energy change originating from the nuclear displacements accompanied by polarization fluctuations. The effect of the latter is inserted via the reorganization term $\lambda^2/(\hbar\Omega)$ related to polaron formation. The exoergicity in the transfer reaction takes on a part similar to that of the bias voltage in the electron transport through molecular junctions. It gives rise to the electron motion through the molecules. In the particular case when the voltage drops between the initial state (left electrode) and the molecular bridge these two quantities are directly related by $|e|V = \Delta F$.

Usually, the long-range electron transfer is observed at moderately high (room) temperatures, so the low temperature approximation (10.43) for the Franck–Condon factor cannot be employed. However, the expression (10.41) remains valid at finite temperatures, only the expression for Green's function must be modified to include the thermal effects. It is shown (Yeganeh et al. 2007a,b) that within the high-temperature limit ($kT > \hbar\Omega$) the expression for the (FC) may be converted to the well-known form first proposed by Marcus:

$$(FC) = \frac{1}{\sqrt{4\pi E_p kT}}\exp\left[-\frac{(\Delta F - E)_p^2}{4E_p kT}\right], \tag{10.44}$$

where $E_p = \lambda^2/(\hbar\Omega)$ is the reorganization energy.

While studying the electron-transfer reactions in practical macromolecules, one keeps in mind that both the donor and the acceptor subsystems in the standard donor–bridge–acceptor triad are usually complex structures including multiple sites coupled to the bridge. Correspondingly, the bridge has a set of entrances and a set of exits that an electron can employ. At different values of the tunnel energy different sites of the donor and/or acceptor subsystems can give predominant contributions to the transfer. Consequently, an electron involved in the transfer arrives at the bridge and leaves from it via different entrances/exits, and it follows different pathways while on the bridge.

Also, nuclear vibrations in the environment could strongly affect the electron transmission destroying the pathways and providing a transition to a completely incoherent sequential hopping mechanism of the electron transfer. All this means that a proper computation of the electron transmission factor H_{DA} for practical macromolecules is a very complicated and nontrivial task. The strong resemblance between the electron-transfer reactions and the electron transport through molecules gives grounds to believe that studies of molecular conduction can provide important information concerning the quantum dynamics of electrons participating in the transfer reactions. One may expect that some intrinsic characteristics of the intramolecular electron transfer, such as the pathways of the tunneling electrons and the distinctive features of the donor/acceptor coupling to the bridge could be obtained in experiments on the electron transport through molecules. For instance, it was recently suggested to characterize electron pathways in molecules using the inelastic electron tunneling spectroscopy, and other advances in this area are to be expected.

10.8 Concluding Remarks

Presently, the electron transport through molecular-scale systems is being intensively studied both theoretically and experimentally. Largely, unceasing efforts of the research community to further advance these studies are motivated by important application potentials of single molecules as active elements of various nanodevices intended to compliment current silicon-based electronics. An elucidation of the physics underlying electron transport through molecules is necessary in designing and operating molecular-based nanodevices. Elastic mechanisms for the electron transport through metal–molecule–metal junctions are currently understood and described fairly well. However, while moving through the molecular bridge, an electron is usually affected by the environment, and it results in a change of its energy. So, inelastic effects appear, and they may bring noticeable changes in the electron transport characteristics. Here, we concentrated on the inelastic and the dissipative effects originating from the molecular bridge vibrations and thermally activated stochastic fluctuations. For simplicity and also to keep this chapter within a reasonable length, we avoided a detailed description of computational formalisms commonly used to theoretically analyze electron transport through molecular junctions. These formalisms are described elsewhere. We mostly focus on the physics of the inelastic effects in the electron transport. Therefore, we employ very simple models and techniques.

We remark that along with nonelasticity originating from electron–vibron interactions, which is the subject of the present review, there exist inelastic effects of different kinds. For instance, inelastic effects arise due to electron–photon interactions. Photoassisted transport through molecular junctions was demonstrated and theoretically addressed using several techniques. It is known that optical pumping could give rise to the charge flow in an unbiased metal–molecule–metal junction, and light emission in biased current-carrying junctions could occur. Also, there is the issue of molecular geometry. There are grounds to conjecture that in some cases the geometry of a molecule included in the junction may change as the current flows through the latter for bonds can break with enough amount of current. Obviously, this should bring a strong inelastic component to the transport, consequently affecting observables.

It is common knowledge that electron–electron interactions may significantly influence molecular conductance leading to a Coulomb blockade and a Kondo effect. To properly treat electron transport through molecular junctions one must take these interactions into consideration. Corresponding studies were carried out omitting electron–phonon interactions. However, a full treatment of the problem including both electron–electron and electron–phonon interactions has not been completed so far. There exist other theoretical challenges, such as the effect of bipolaron formation that originates from an effective electron–electron attraction via phonons.

Finally, practical molecular junctions are complex systems, and a significant effort is necessary to bring electron transport calculations to a result that could be successfully compared with the experimental data. For this purpose, one needs to compute molecular orbitals and the voltage distribution inside the junctions to get sufficient information on the vibrionic spectrum of the molecule, the electron–phonon coupling strengths, and electron–electron interactions. One needs a good quantitative computational scheme for transport calculations where all this information can be accounted for. Currently, there remain some challenges that have not been properly addressed by theory. Therefore, a comparison between the theoretical and the experimental results on the molecular conductance sometimes does not bring satisfactory results. However, there are firm grounds to believe that further efforts of the research community will result in a detailed understanding of all the important aspects of molecular conductance including inelastic and dissipative effects. Such an understanding is paramount to the conversion of molecular electronics into a viable technology.

References

Agrait, N.; Yeati, A. L., and van Ruitenbeek, L. M. 2003. Quantum properties of atomic-sized conductors. *Phys. Rep.* **377**: 81–279.

Aviram, A.; Ratner, M. A., and Mujica, V. (Eds). 2002. *Molecular Electronics II*. New York: Annals of the New York Academy of Sciences.

Balseiro, C.A.; Cornaglia, P. S., and Grempel, D. R. 2006. Electron–phonon correlation effects in molecular transistors. *Phys. Rev. B* **74**: 235409.

Beratan, D. N.; Onuchic, J. N.; Winkler, J. R. et al. 1992. Electron-tunneling pathways in proteins. *Science* **258**: 1740–1741.

Buttiker, M. 1986. Role of quantum coherence in series resistors. *Phys. Rev. B* **33**: 3020–3026.

Buttiker, M. and Landauer, R. 1985. Traversal time for tunneling. *Phys. Scri.* **32**: 429–434.

Cizek, M.; Thoss, M., and Domcke, W. 2004. Theory of vibrationally inelastic electron transport through molecular bridges. *Phys. Rev. B* **70**: 125406.

Cornaglia, P. S.; Ness, H., and Grempel, D. R. 2004. Many-body effects on the transport properties of single-molecule devices. *Phys. Rev. Lett.* **93**: 147201.

Cornaglia, P. S.; Grempel, D. R., and Ness, H. 2005. Quantum transport through a deformable molecular transistor. *Phys. Rev. B* **71**: 075320.

Cuniberti, G.; Fagas, G., and Richter, K. (Eds). 2005. *Introductory Molecular Electronics: A Brief Overview*, Lecture Notes in Physics, Volume **680**. Berlin, Germany: Springer.

Dahnovsky, Yu. 2006. Modulating electron dynamics: Modified spin-boson approach. *Phys. Rev. B* **73**: 144303.

Daizadeh, I.; Gehlen, J. N., and Stuchebrukhov, A. A. 1997. Calculation of electronic tunneling matrix element in proteins: Comparison of exact and approximate one-electron methods for Ru-modified azurin. *J. Chem. Phys.* **106**: 5658–5666.

D'Amato, J. L. and Pastawski, H. M. 1990. Conductance of a disordered linear chain including inelastic scattering events. *Phys. Rev. B* **41**: 7411–7420.

D'Amico, P.; Ryndyk, D. A.; Cuniberti, G. et al. 2008. Charge-memory effect in a polaron model: Equation-of-motion method for Green functions. *New J. Phys.* **10**: 085002.

Damle, P. S.; Ghosh, A. W., and Datta, S. 2001. Unified description of molecular conduction: From molecules to metallic wires. *Phys. Rev. B* **64**: 201403.

Datta, S. 1997. *Electron Transport in Mesoscopic Systems*. Cambridge, U. K.: Cambridge University Press.

Datta, S. 2005. *Quantum Transport: Atom to Transistor*. Cambridge, U. K.: Cambridge University Press.

Di Ventra, M.; Pantelides, S. T., and Lang, N. D. 2000. First-principles calculations of transport properties of a molecular device. *Phys. Rev. Lett.* **84**: 979–982.

Djukic, D.; Thygesen, K. S.; Untiedt, C. et al. 2005. Stretching dependence of the vibration modes of a single-molecule Pt–H_2–Pt bridge. *Phys. Rev. B* **71**: 161402(R).

Donarini, A.; Grifoni, M., and Richter, K. 2006. Dynamical symmetry breaking in transport through molecules. *Phys. Rev. Lett.* **97**: 166801.

Egger, R. and Gogolin, A. O. 2008. Vibration-induced correction to the current through a single molecule. *Phys. Rev. B* **77**: 1098.

Galperin, M.; Ratner, M. A., and Nitzan, A. 2004. Inelastic electron tunneling spectroscopy in molecular junctions: Peaks and dips. *J. Chem. Phys.* **121**: 11965–11979.

Galperin, M.; Ratner, M. A., and Nitzan, A. 2005. Hysteresis, switching, and negative differential resistance in molecular junctions: A polaron model. *Nano Lett.* **5**: 125–130.

Galperin, M.; Nitzan, A., and Ratner, M. A. 2006. Molecular transport junction: Current from electronic excitations in the leads. *Phys. Rev. Lett.* **96**: 166803.

Galperin, M.; Ratner, M. A., and Nitzan, A. 2007. Topical review: Molecular transport junctions: Vibrational effects. *J. Phys. Condens. Matter* **19**: 103201.

Garg, A.; Onuchic, J. N., and Ambegaokar, V. 1985. Effect of friction on electron transfer in biomolecules. *J. Chem. Phys.* **83**: 4491–4503.

Grobis, M.; Wachowiak, A.; Yamachika, M. F. et al. 2005. Tuning negative differential resistance in a molecular film. *Appl. Phys. Lett.* **86**: 204102.

Gutierrez, R.; Mandal, S., and Cuniberti, G. 2005. Dissipative effects in the electronic transport through DNA molecular wires. *Phys. Rev. B* **71**: 235116.

Gutirrez, R.; Mohapatra, S.; Cohen, H. et al. 2006. Inelastic quantum transport in a ladder model: Implications for DNA conduction and comparison to experiments on suspended DNA oligomers. *Phys. Rev. B* **74**: 235105.

Hahn, J. R.; Lee, H. J., and Ho, W. 2000. Electronic resonance and symmetry in single-molecule inelastic electron tunneling. *Phys. Rev. Lett.* **85**: 1914–1917.

Ho, J. W. 2002. Single molecule chemistry. *J. Chem. Phys.* **117**: 11033–11061.

Imry, Y. and Landauer, R. 1999. Conductance viewed as transmission. *Rev. Mod. Phys.* **71**: S306–S312.

Jones, M. L.; Kurnikov I. V., and Beratan, D. N. 2002. The nature of tunneling pathway and average packing density model for protein-mediated electron transfer. *J. Phys. Chem. A* **106**: 200206.

Kastner, M. A. 1992. The single-electron transistor. *Rev. Mod. Phys.* **64**: 849–858.

Komeda, T.; Kim, Y.; Kawai, M. et al. 2002. Lateral hopping of molecules induced by excitation of internal vibration mode. *Science* **295**: 2055–2058.

Kubatkin, S.; Danilov, A.; Hjort, M. et al. 2003. Single-electron transistor of a single organic molecule with access to several redox states. *Nature* **425**: 698–701.

Kushmerick, J. G.; Lazorcik, J.; Patterson, C. H. et al. 2004. Vibronic contributions to charge transport across molecular junctions. *Nano Lett.* **4**: 63942.

Kuznetsov, A. M. and Ulstrup, I. 1999. *Electron Transfer in Physics and Biology*. Chichester, U. K.: John Wiley.

Landauer, R. 1970. Electrical resistance of disordered one-dimensional lattices. *Philos. Mag.* **21**: 863–867.

Lang, N. D. and Avouris, Ph. 2000a. Carbon-atom wires: Charge-transfer doping, voltage drop and the effect of distortions. *Phys. Rev. Lett.* **84**: 358–361.

Lang, N. D. and Avouris, Ph. 2000b. Electrical conductance of parallel atomic wires. *Phys. Rev. B* **62**: 7325–7329.

Li, X.-Q. and Yan, Y.-J. 2001a. Electrical transport through individual DNA molecules. *Appl. Phys. Lett.* **79**: 2190–2192.

Li, X.-Q. and Yan, Y.-J. 2001b. Scattering matrix approach to electronic dephasing in longrange electron transfer. *J. Chem. Phys.* **115**: 4169–4174.

Li, C.; Zhang, D.; Liu, X. et al. 2003. In_2O_3 nanowires as chemical sensors. *Appl. Phys. Lett.* **82**: 1613.

Lorente, N.; Persson, M.; Lauhon, L. J. et al. 2001. Symmetry selection rules for vibrationally inelastic tunneling. *Phys. Rev. Lett.* **86**: 2593–2596.

Lortscher, E.; Weber, H. B., and Riel, H. 2007. Statistical approach to investigating transport through single molecules. *Phys. Rev. Lett.* **98**: 176807.

MacDiarmid, A. G. 2001. Nobel lecture: "Synthetic metals": A novel role for organic polymers. *Rev. Mod. Phys.* **73**: 701–712.

Mahan, G. D. 2000. *Many-Particle Physics*. New York: Plenum.

Marcus, R. A. 1965. On the theory of electron-transfer reactions. VI. Unified treatment for homogeneous and electrode reactions. *J. Chem. Phys.* **43**: 679–701.

Mii, T.; Tikhodeev, S. G., and Ueba, H. 2003. Spectral features of inelastic electron transport via a localized state. *Phys. Rev. B* **68**: 205406.

Mitra, A.; Aleiner I., and Millis, A. J. 2004. Phonon effects in molecular transistors: Quantal and classical treatment. *Phys. Rev. B* **69**: 245302.

Mitra, A.; Aleiner I., and Millis, A. J. 2005. Semiclassical analysis of the nonequilibrium local polaron. *Phys. Rev. Lett.* **94**: 076404.

Mozyrsky, D.; Hastings, M. B., and Martin, I. 2006. Intermittent polaron dynamics: Born–Oppenheimer approximation out of equilibrium. *Phys. Rev. B* **73**: 035104.

Mujica, V.; Kemp, M., and Ratner, M. A. 1994. Electron conduction in molecular wires. I. A scattering formalism. *J. Chem. Phys.* **101**: 6849–6855.

Mujica, V.; Roitberg, A. E., and Ratner, M. A. 2000. Molecular wire conductance: Electrostatic potential spatial profile. *J. Chem. Phys.* **112**: 6834–6839.

Nitzan, A. 2001. Electron transmission through molecules and molecular interfaces. *Ann. Rev. Phys. Chem.* **52**: 681–750.

Olsson, F. E.; Paavilainen, S.; Persson, M. et al. 2007. Multiple charge states of Ag atoms on ultrathin NaCl films. *Phys. Rev. Lett.* **17**: 176803.

Park, H.; Park, J.; Lim, A. K. L. et al. 2000. Nanomechanical oscillations in a single-C_{60} transistor. *Nature* **407**: 57–60.

Persson, B. N. J. and Baratoff, A. 1987. Inelastic electron tunneling from a metal tip: The contribution from resonant processes. *Phys. Rev. Lett.* **59**: 339–342.

Prigodin, V. N. and Epstein, A. J. 2002. Nature of insulator–metal transition and novel mechanism of charge transport in highly doped electronic polymers. *Synth. Met.* **125**: 43–53.

Poot, M.; Osorio, E.; O'Neil, K. et al. 2006. Temperature dependence of three-terminal molecular junctions with sulfur end-functionalized tercyclohexylidenes. *Nano Lett.* **6**: 1031–1035.

Qin, X. H.; Nazin G. V., and Ho, W. 2004. Vibronic states in single molecule electron transport. *Phys. Rev. Lett.* **92**: 206102.

Reichert, J.; Ochs, R.; Beckmann, D. et al. 2002. Driving current through single organic molecules. *Phys. Rev. Lett.* **88**: 176804.

Repp, J.; Meyer, G.; Olsson, F. E. et al. 2004. Controlling the charge state of individual gold adatoms. *Science* **305**: 493–495.

Repp, J.; Meyer, G.; Stojkovic, S. M. et al. 2005a. Molecules on insulating films: Scanning-tunneling microscopy imaging of individual molecular orbitals. *Phys. Rev. Lett.* **94**: 026803.

Repp, J.; Meyer, G.; Paavilainen, S. et al. 2005b. Scanning tunneling spectroscopy of Cl vacancies in NaCl films: Strong electron–phonon coupling in double-barrier tunneling junctions. *Phys. Rev. Lett.* **95**: 225503.

Ryndyk, D. A. and Cuniberti, G. 2007. Nonequilibrium resonant spectroscopy of molecular vibrons. *Phys. Rev. B* **76**: 155430.

Ryndyk, D. A.; D'Amico, P.; Cuniberty, G. et al. 2008. Charge-memory polaron effect in molecular junctions. *Phys. Rev. B* **78**: 085409.

Samanta, M. P.; Tian, W.; Datta, S. et al. 1996. Electronic conduction through organic molecules. *Phys. Rev. B* **53**: R7626–R7629.

Segal, D.; Nitzan, A.; Davis, W. B. et al. 2000. Electron transfer rates in bridged molecular systems. 2. A steady-state analysis of coherent tunneling and thermal transitions. *J. Phys. Chem. B* **104**: 3817–3829.

Segal, D.; Nitzan, A., and Hnggi, P. 2003. Thermal conductance through molecular wires. *J. Chem. Phys.* **119**: 6840–6855.

Siddiqui, L.; Ghosh, A. W., and Datta, S. 2007. Phonon runaway in nanotube quantum dots. *Phys. Rev. B* **76**: 085433.

Simonian, N.; Li, J., and Likharev, K. 2007. Negative differential resistance at sequential single-electron tunnelling through atoms and molecules. *Nanotechnology* **18**: 424006.

Skourtis, S. S. and Mukamel, S. 1995. Superexchange versus sequential long range electron transfer: Density matrix pathways in Liouville space. *Chem. Phys.* **197**: 367–388.

Smit, R. H. M.; Noat, Y.; Untiedt, C. et al. 2002. Measurement of the conductance of a hydrogen molecule. *Nature* **419**: 906–909.

Smit, R. H. M.; Untiedt, C., and van Ruitenbeek, J. M. 2004. The high-bias stability of monatomic chains. *Nanotechnology* **15**: S472–S478.

Stipe, B. C.; Rezaei, M. A., and Ho, W. 1999. Localization of inelastic tunneling and the determination of atomic-scale structure with chemical specificity. *Phys. Rev. Lett.* **82**: 1724–1727.

Tikhodeev, S. G. and Ueba, H. 2004. Relation between inelastic electron tunneling and vibrational excitation of single adsorbateson metal surfaces. *Phys. Rev. B* **70**: 125414.

Troisi, A. and Ratner, M. A. 2006. Molecular transport junctions: Propensity rules for inelastic tunneling. *Nano Lett.* **6**: 1784–1788.

Tsutsui, M.; Kurokawa, S., and Sakai, A. 2006. Bias-induced local heating in Au atom-sized contacts. *Nanotechnology* **17**: 5334–5338.

Wang, W.; Lee, T.; Kretzchmar, I. et al. 2004. Inelastic electron tunneling spectroscopy of an alkanedithiol self-assembled monolayer. *Nano Lett.* **4**: 643–646.

Wingreen, N. S.; Jacobsen, K. W., and Wilkins, J. W. 1989. Inelastic scattering in resonant tunneling. *Phys. Rev. B* **40**: 11834–11850.

Wingreen, N. S.; Jauho, A.-P., and Meir, Y. 1993. Time-dependent transport through a mesoscopic structure. *Phys. Rev. B* **48**: 8487–8490.

Wu, S. W.; Nazin, G. V.; Chen, X. et al. 2004. Control of relative tunneling rates in single molecule bipolar electron transport. *Phys. Rev. Lett.* **93**: 236802.

Xu, B.; Zhang, P.; Li, X. et al. 2004. Direct conductance measurement of single DNA molecules in aqueous solution. *Nano Lett.* **4**: 1105–1108.

Xue, Y. and Ratner, M. A. 2003a. Microscopic study of electrical transport through individual molecules with metallic contacts. I. Band lineup, voltage drop, and high-field transport. *Phys. Rev. B* **68**: 115406.

Xue, Y. and Ratner, M. A. 2003b. Microscopic study of electrical transport through individual molecules with molecular contacts. II. Effect of interface structure. *Phys. Rev. B* **68**: 115407.

Xue, Y. and Ratner, M. A. 2004. End group effect on electrical transport through individual molecules: A microscopic study. *Phys. Rev. B* **69**: 085403.

Xue, Y.; Datta, S.; Hong, S. et al, 1999. Negative differential resistance in the scanning-tunneling spectroscopy of organic molecules. *Phys. Rev. B* **59**: R7852–R7855.

Xue, Y.; Datta, S., and Ratner, M. A. 2001. Charge transfer and "band lineup" in molecular electronic devices: A chemical and numerical interpretation. *J. Chem. Phys.* **115**: 4292–4299.

Yeganeh, S.; Galperin, M., and Ratner M. A. 2007a. Switching in molecular transport junctions: Polarization response. *J. Am. Chem. Soc.* **129**: 13313–13320.

Yeganeh, S.; Ratner, M. A., and Mujica, V. 2007b. Dynamics of charge transfer: Rate processes formulated with nonequilibrium Green's functions. *J. Chem. Phys.* **126**: 161103.

Yu, L. H.; Keane, Z. K.; Ciszek, J. W. et al. 2004. Inelastic electron tunneling via molecular vibrations in single-molecule transistors. *Phys. Rev. Lett.* **93**: 266802.

Zazunov, A. and Martin, T. 2007. Transport through a molecular quantum dot in the polaron crossover regime. *Phys. Rev. B* **76**: 033417.

Zazunov, A.; Feinberg, D., and Martin, T. 2006a. Phonon squeezing in a superconducting molecular transistor. *Phys. Rev. Lett.* **97**: 196801.

Zazunov, A.; Feinberg, D., and Martin, T. 2006b. Phonon-mediated negative differential conductance in molecular quantum dots. *Phys. Rev. B* **73**: 115405.

Zhitenev, N. B.; Meng, H., and Bao, Z. 2002. Conductance of small molecular junctions. *Phys. Rev. Lett.* **88**: 226801.

Zimbovskaya, N. A. 2003. Low temperature electronic transport and electron transfer through organic macromolecules. *J. Chem. Phys.* **118**: 4–7.

Zimbovskaya, N. A. 2005. Low temperature electronic transport through macromolecules and characteristics of intramolecular electron transfer. *J. Chem. Phys.* **123**: 114708.

Zimbovskaya, N. A. 2008. Inelastic electron transport in polymer nanofibers. *J. Chem. Phys.* **123**: 114705.

Zimbovskaya, N. A. and Gumbs, G. 2002. Long-range electron transfer and electronic transport through macromolecules. *Appl. Phys. Lett.* **81**: 1518–1520.

Zimbovskaya, N. A. and Pederson, M. R. 2008. Negative differential resistance in molecular junction: The effect of electrodes electronic structure. *Phys. Rev. B* **78**: 153105.

Zimmerman, J.; Pavone, P., and Cuniberti, G. 2008. Vibrational modes and low-temperature thermal properties of graphene and carbon nanotubes: Minimal force constant model. *Phys. Rev. B* **78**: 045410.

11

Bridging Biomolecules with Nanoelectronics

11.1 Introduction and Background **11**-1
11.2 Preparation of Molecular Magnets **11**-2
Folding of Magnetic Protein Mn,Cd-MT • Protein: A Mesoscopic System • Quasi-Static Thermal Equilibrium Dialysis for Magnetic Protein Folding
11.3 Nanostructured Semiconductor Templates: Nanofabrication and Patterning **11**-6
Patterned Self-Assembly for Pattern Replication • Fabrication by Direct Inkjet and Mold Imprinting • Nanopatterned SAMs as 2D Templates for 3D Fabrication
11.4 Self-Assembling Growth of Molecules on the Patterned Templates **11**-8
11.5 Magnetic Properties of Molecular Nanostructures **11**-10
11.6 Conclusion and Future Perspectives **11**-10
References **11**-11

Kien Wen Sun
National Chiao Tung University

Chia-Ching Chang
National Chiao Tung University

11.1 Introduction and Background

The field of nanostructures has grown out of the lithographic technology developed for integrated circuits, but is now much more than simply making smaller transistors. In the early 1980s, microstructures became small enough to observe interesting quantum effects. These structures were smaller than the inelastic scattering length of an electron so that the electrons could remain coherent as they traversed them, giving rise to interference phenomena. Studies on the Aharonov–Bohm effect and universal conductance fluctuations led to the field of "mesoscale physics"—between macroscopic classical systems and fully quantized ones.

Now the size of the structures that can be produced is approaching the de Broglie wavelength of the electrons in the solids, leading to stronger quantum effects. In addition to interesting new physics, this drive toward smaller length scales has important practical consequences. When semiconductor devices reach about 100 nm, the essentially classical models of their behavior will no longer be valid. It is not yet clear how to make devices and circuits that will operate properly on these smaller scales. The replacement for the transistor, which must carry the technology to well below 100 nm, has not been identified. It is anticipated that the semiconductor industry will run up against this "wall" within about 10 years.

The current very large-scale integrated circuit paradigm based on complementary metal oxide semiconductor (CMOS) technology cannot be extended into a region with features smaller than 10 nm.[1] With a gate length well below 10 nm, the sensitivity of the silicon field-effect transistor parameters may grow exponentially due to the inevitable random variations in the device size. Therefore, an alternative nanodevice concept of molecular circuits was proposed, which was a radical paradigm shift from the pure CMOS technology to the hybrid semiconductor.[2] The concept combines the advantages of nanoscale components, such as the reliability of CMOS circuits, and the advantages of patterning techniques, which include the flexibility of traditional photolithography and the potentially low cost of nanoimprinting and chemically directed self-assembly. The major attraction of this concept is the incorporation of the richness of organic chemistry with the versatilities of semiconductor science and technology. However, before this, one needs to bring directed self-assembly from the present level of single-layer growth on smooth substrates to the reliable placement of three-terminal molecules on patterned semiconductor structures.

While physics and electrical engineering have been evolving from microstructures, to mesoscopic structures, and now to nanostructures, molecular biologists have always worked with objects of a few nanometers or less. A DNA molecule, for example, is very long (when stretched out), but it is about 2.5 nm wide with base pairs separated by 0.34 nm. Since the technology has not existed to directly fabricate and manipulate objects this small, various chemical techniques have been developed for cutting, tagging, and sorting large biological molecules. While enormously successful, these methods utilize batch processing of huge numbers of the molecules and rely on statistical interpretations. It is not possible, in principle, to sequence one particular DNA molecule with these techniques, for example. Ideally, one would like to stretch the DNA straight and simply read the sequence of the base pairs. While we are still far from this goal, nanofabrication techniques promise to bring us much closer.

During recent years, self-assembly has become one of the most important strategies used in biology for the development of complex, functional structures. Self-assembly on modified surfaces is one of the approaches to self-assemble structures that are particularly successful. By the coordination of molecules to surfaces, the molecular systems form ordered systems—self-assembled monolayers (SAMs). The SAMs are reasonably well understood and are increasingly useful technologically. A thin film of diblock copolymers can be self-assembled into ordered periodic structures at the molecular scale (~5–50 nm) and have been used as templates to fabricate quantum dots,[3,4] nanowires,[5–7] and magnetic storage media.[8] More recently, in epitaxial assembly of block-copolymer films, molecular level control over the precise size, shape, and spacing of the order domains was achieved with advanced lithographic techniques.[9] The development of methods for patterning and immobilizing biologically active molecules with micrometer and nanometer scale control has been proven integral to ranges of applications such as basic research, diagnostics, and drug discovery. Some of the most important advances have been in the development of biochip arrays that present either DNA,[10] protein,[11] or carbohydrates.[12] The use of patterned substrates for components of microfluidic systems for bioanalysis is also progressing rapidly.[13–15] Surface modification and patterning at the nanoscale to anchoring protein molecules is an important strategy on the way of obtaining the construction of new biocompatible materials with smart bioactive properties. In fact, surfaces patterned by protein molecules can act as active agents in a large number of important applications including biosensors capable of multifunctional biological recognition. In particular, physical adsorption of proteins onto semiconductor surfaces makes possible to combine the simplicity of the method with the versatility of chemical and physical properties of proteins.

In recent years, there has been substantial attention focused on the reactions of organic compounds with silicon surfaces. The major attraction is the incorporation of the richness of organic chemistry with the versatilities of semiconductor science and technology. More recently, it has been demonstrated that TEMPO,2,2,6,6-tetramathylpiperidinyloxy can bond with a single dangling bond on hydrogen-terminated Si(100) and Si(111) surfaces.[16] Functional organic molecular layers were found to self-assemble on metal[2] and semiconductor surfaces.[17] The technique of self-assembly is one of the few practical strategies available to arrive at one to three-dimensional ensembles of nanostructures. There are many different mechanisms by which self-assembly of molecules and nanoclusters can be accomplished, such as chemical reactions, electrostatic and surface forces, and hydrophobic and hydrophilic interactions.

In this chapter, we present most recent advances in the development of techniques in immobilizing a single nanostructure and producing arrays of 3D magnetic protein nanostructures with high throughput on surface modified semiconductor substrates. Well-characterized test nanostructures were prepared using current state-of-the-art nanofabrication techniques. Both nanoimaging and scanning probe microscopy studies (AFM, SEM, and TEM) on semiconductor nanostructures and these molecular self-assembly systems were performed. The combination of e-beam lithography, scanning probe microscope imaging, spectroscopy, and self-assembly approaches provide not only the high throughput of producing arrays of protein nanostructures but also with highest precision of positioning single nanostructure and/or single molecule.

11.2 Preparation of Molecular Magnets

11.2.1 Folding of Magnetic Protein Mn,Cd-MT

Metallothionein (MT) is a metal binding protein that binds seven divalent transition metals avidly via its twenty cysteines (Cys).[18] These Cys' form two metal binding clusters located at the carboxyl (α-domain) and amino (β-domain) terminals of MT.[19] The two clusters were identified as α-cluster $(M_4S_{11})^{3-}$ and the β-cluster $(M_3S_9)^{3-}$ (Figure 11.1),[20–22] where M denotes metal ions (Zn^{2+}, Cd^{2+}, or others), according to both x-ray crystallographic and NMR studies.[3] MT binds to metals ions via metal-thiol linkages.[19] As shown in Figure 11.1, the $(M_3S_9)^{3-}$ and $(M_4S_{11})^{3-}$ have the zinc-blende like structure that is similar to the "diluted magnetic semiconductor (DMS)" compounds.[23]

In general, semiconductors are not magnetic. However, a DMS exhibits magnetic properties by doping with Mn and Cd or other II–VI metal ions in certain ratio. The doped metal clusters among the semiconductors are in zinc-blende structures. Meanwhile, these semiconductors possess magnetic property only in low temperatures.[23] The magnetic properties may be the result of the d-sp^3 orbital hybridization and the alignment of the electron spins. The bridging sulfur atoms may also contribute to the alignment of the spins of the Mn^{2+} ions. Thus, by chelating the Mn^{2+} and Cd^{2+} with MT (i.e., Mn,Cd-MT), a "magnetic protein" may be obtained.

Recently, the single molecule magnets (SMMs) have attracted much attention.[24] However, the Curie temperature of these molecules has to be as low as 2–4 K[24–26] to avoid the thermal fluctuation among the electron spin within the molecules.[27] To be of practical utilization, it is highly desirable to create a room temperature molecular magnet.[28] With this intention in mind, one has to construct and investigate a new metal binding protein, MT, which sustained characteristic magnetic hysteresis loop from 10 to 330 K. The protein backbone may restrain the net spin moment of Mn^{2+} ions to overcome the minor thermal fluctuation. The magnetic-metallothionein (mMT) presented may reveal a possible approach to create high temperature molecular magnet. In order to prepare the Mn,Cd-MT magnetic proteins, the folding mechanism of protein should be introduced.

11.2.2 Protein: A Mesoscopic System

Protein is a complex biomolecule that contains a large number of basic residues—amino acids. Therefore, it is not possible to analyze it completely by macroscopic approaches. Meanwhile, it

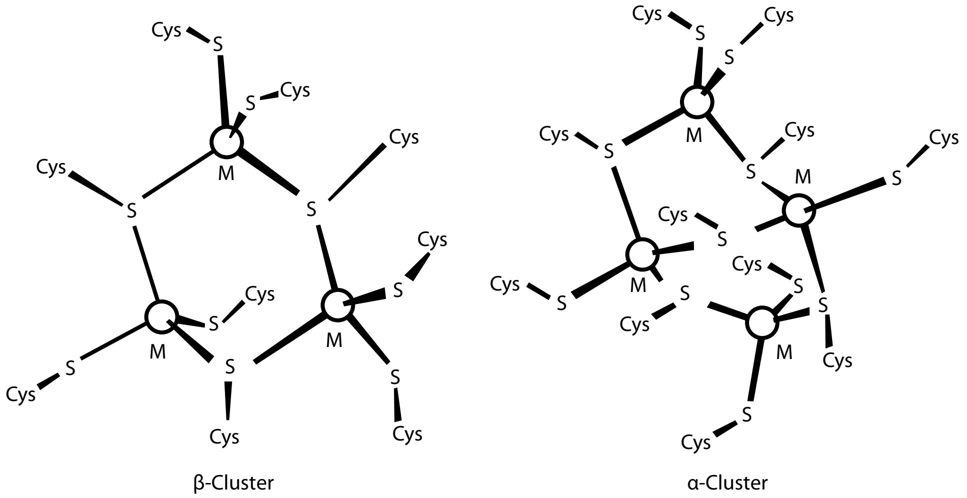

FIGURE 11.1 Metal binding clusters of MT that was modified from x-ray crystal structure (2). Where the circles denote metal ions Zn^{2+}, Cd^{2+}, or Mn^{2+}. Each metal ion was linked with protein via metal-thiol bonds.

is not feasible to describe the dynamics of its polypeptide chain behavior by using conventional statistical approaches either. Thus, a protein can be thought of as a mesoscopic system.[29] However, the conformational transition from unfolded state to the native state of a protein may be similar to the conventional phase transition model. In physics, a phase transition is the transformation of a thermodynamic system from one phase to another. In a folding process, proteins follow the thermodynamic theories and transform from the unfolded state to the folded state, where the state is defined as a region of configuration space with minimal potential.[30] Therefore, we named this conformational change as "state transition."[31] Due to the complexity of a protein folding system, a single experimental study may reveal only a part of the fact of protein folding. Therefore, we should examine the protein-folding problem with multidimensional approaches and integrate the findings to reveal the true mechanism of protein folding.

Protein folding may follow a spontaneous process[29] or a reaction-path directed process[30] *in vitro*. A choice between the two may be determined by the intrinsic properties of proteins, for example, the varying folding transition boundaries. However, a general model, named "first-order-like state transition model," in which the aggregated proteins exist within finite boundaries can encompass both processes without any conflictions.[31–36] According to this model, the folding path of the protein may not be unique. It can be folded, without being trapped in an aggregated state, via a carefully designed refolding path circumventing the transition boundary, that is, via an overcritical path.[31–36] The intermediates, following an overcritical path, are in a molten globular state,[37] and their behavior is consistent with both a sequential[38] and a collapse model.[39] However, both soluble (folded) and precipitated (unfolded) proteins can be observed in the direct folding reaction path *in vitro*. In terms of the "first-order-like state transition model" language, this can be described as stepping across the state transition line in the protein folding reaction phase diagram.[31,32] Since in protein refolding it is important to prevent protein aggregation *in vitro*, similarly, in biomedical applications, the revelation of the mechanism of the formation of the two states (unfolded and folded) becomes significant.

Previous studies have indicated that chemical environment,[40] temperature,[41] pH,[42] ionic strength,[43] dielectric constant,[44] and pressure[49]—considered as solvent effects collectively—could affect the fundamental structure, thermodynamics, and dynamics of polypeptides/proteins. The reaction ground state can be expressed as a dual-well potential according to the two-state transition model. The conformational energy, in general, of the unfolded state is relatively higher than that of the native state (Figure 11.1). When the system reaches thermal equilibrium, most protein molecules are found in their native state. No unfolded or intermediate states are observable. However, as described previously,[31–36] if a denaturant is added, the reaction potential may change accordingly as indicated in Figure 11.1. Then, an unfolded protein may be stable, as it is now at the lowest energy under the newly established equilibrium. The energy of the system can be expressed as follows:

$$H_T = H_p + \lambda H_s \tag{11.1}$$

where H_T, H_p, and H_s denote the potential energy of the interacting protein-solvent total system, the protein, and the solvent, respectively. The factor, λ, is a weighting factor of the solvent environment ($0 \le \lambda \le 1$). It approaches unity when a denaturant is present as pure solvent and decreases in value as the concentration of the denaturant is reduced.

When λ of the system is changed drastically, direct folding ensues and leads to the release of some of the bound denaturant. According to the Donnan effect in a macromolecule-counter ions interactive system, the diffusion of the bound denaturant can be expressed by Fick's first law:

$$\vec{J} = -D\nabla n \tag{11.2}$$

where

n denotes the concentration of the denaturant that is dissociated from the protein
D denotes the diffusive constant
vector J denotes the flux, respectively, of the solute

According to the Einstein relation

$$D = \frac{kT}{6\pi\eta R_H} \tag{11.3}$$

where

k is the Boltzmann constant
T is the temperature in Kelvin
η is the viscosity of the solvent
R_H is the hydration radius of the solutes

Due to the intrinsic diffusion process, the solute exchange processes are not synchronous for all protein molecules. Therefore, the folding rate of protein may not be measured directly by a simple spectral technique, that is, the stopped-flow CD,[45] continuous-flow CD,[46] or fluorescence.[47] However, the reaction interval of protein folding can be revealed by the autocorrelation of reaction time from these direct measurements. The detailed mechanism and an example will be discussed later.

If we look at the energy landscape funnel model of protein folding,[48] it appears that proteins can be trapped in a multitude of local minima of the potential well in a complicated protein system. The native state, though, is at the lowest energy level. When thermal equilibrium is reached, most of the protein molecules are located in the lowest energy state, with a population ratio as low as $e^{-\Delta E/kT}$, according to the Maxwell–Boltzmann distribution in thermodynamics. The ΔE denotes the energy difference between the native state and a local minimum; k and T denote the Boltzmann constant and temperature in Kelvin, respectively. At high concentration (>0.1 mg/mL), however, considerable amount of insoluble protein has been observed in protein folding,[31,33–37,48] indicating that insoluble proteins are at an even lower energy state than the native protein. Therefore, by considering the intermolecular interactions during the protein folding process, the reaction energy landscape may be expressed as a three-well model (Figure 11.2). As shown in Figure 11.2, the unfolded protein (U) is in the highest energy state; the native protein (N) is in the lower energy state. However, the intermediate (I) that may cause further protein aggregation/precipitation is in the lowest energy state. Although the energy state of the intermediate/aggresome is the lowest energy state, the conformational energy of the individual proteins composing the aggresome may not be lower than the native protein. Namely, in single molecular simulation, this extra potential well of intermediate (I) is nonexistent. Therefore, in the conventional energy landscape model (the single molecule simulation model), the lowest energy state "I" cannot be observed. According to the Zwanzig's definition of state, the protein molecules in the intermediate (I) belong to an unfolded state.[13] Hence, in a direct folding reaction, the soluble (N) and the insoluble parts (U) can coexist and they can be observed simultaneously, which is similar to the situation where the phase transition line is crossed in reactions congruent to the "first-order phase transition" model. Therefore, we named the protein folding reaction as "first-order like state transition model" (as shown in Figure 11.3).

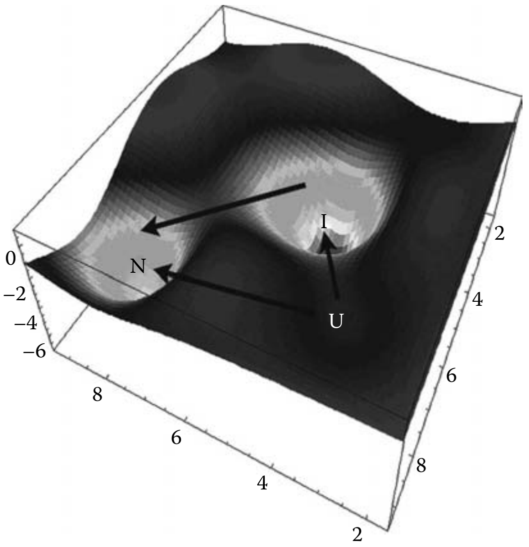

FIGURE 11.2 Three well model of multi-protein molecules folding reaction. The U denotes the unfolded state. N denotes the native state and I denotes the protein–protein complex (aggresome) intermediate.

The $\Phi(n_1, n_2, \ldots)$ in Figure 11.3 denotes the folding status of protein, where $n_1, n_2, \ldots$ represent the variables affecting the folding status, such as, temperature, concentration of denaturants, etc. The reaction curve indicates an overcritical reaction path of a quasi-static folding reaction. The gray area in Figure 11.3 indicates the state transition boundary of protein folding. The gray line and dash line indicate the reaction path of direct folding. By combining the three-well model (Figure 11.2) and the direct

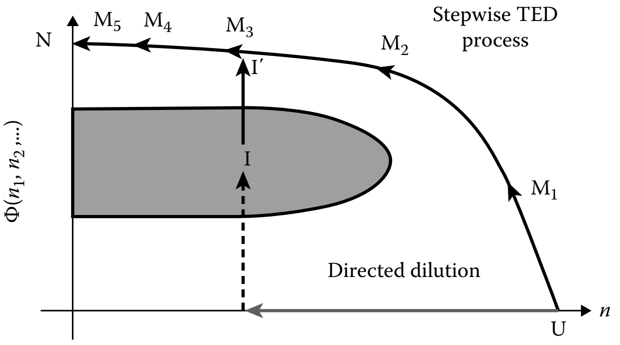

FIGURE 11.3 The protein folding phase diagram, where the $\Phi(n_1, n_2, \ldots)$ denotes the folding status (the order parameter) of protein. The $n, n_1, n_2, \ldots$ denote the variables that affect the folding status such as temperature, concentration of denaturants, etc.

folding reaction of the "first-order like state transition model" (Figure 11.3), we realized that those folded protein molecules along the direct folding path might fold spontaneously or form aggregates. Spontaneous folding may be driven by enthalpy–entropy compensation.

As indicated previously, the conformation of protein changed with changes of the solvent environment. It seems that the protein may fold spontaneously, such as in Anfinsen's experiment[44] and direct folding reactions. The protein folding reaction, similar to all chemical reactions, reaches its equilibrium by following the fundamental laws of thermodynamics. Although protein folding has been studied extensively in certain model systems for over 40 years, the driving force at the molecular level remained unclear until recently.

It is known that polymers and macromolecules may self-assemble/self-organize into a wide range of highly ordered phases/states at thermal equilibrium.[49–52] In a condensed solvent environment, large molecules may self-organize to reduce their effective volume. Meanwhile, the number of the allowed states (Ω) of small molecules, such as buffer salt and other counter-ions in solution, increases considerably. Therefore, the entropy of the system, $\Delta S = R\ \ln(\Omega_f/\Omega_i)$, becomes large, where i and f denote the initial and final states, respectively. Meanwhile, the enthalpy change (ΔH) between the unfolded and native protein is around hundreds kcal/mol.[53] Therefore, the Gibbs free energy of the system, $\Delta G = \Delta H - T\mathrm{d}S$, becomes more negative in this system when the large molecules self-organize.[54] A similar entropy–enthalpy compensation mechanism has been used to solve the reaction of colloidal crystals that self-assemble spontaneously.[55,56]

According to our studies,[31,33–37] the effective diameter of the unfolded protein is about 1.7–2.5 fold larger than the folded protein. Therefore, with the same mechanism, those macromolecules (proteins) may tend to reduce their effective volumes and increase the system entropy when thermal equilibrium is reached. The increase in entropy may compensate for the change of the enthalpy of the system and enable the reaction to take place spontaneously. This may be the reaction molecular mechanism of spontaneous protein folding reactions. Meanwhile, a similar mechanism can be adopted into the self-assembly process of magnetic protein in nanopore arrays.

11.2.3 Quasi-Static Thermal Equilibrium Dialysis for Magnetic Protein Folding

Due to the intrinsic diffusion process, the solvent exchange rate is slow and thus the variation of λ is slow and can be thought of as quasi-static. Therefore, we named this buffer exchanging process as a quasi-static process. We manipulated the reaction direction of the protein folding through this process. Meanwhile, we can obtain stable intermediates in each thermal equilibrium state. These intermediates may help us reveal the molecular folding mechanism of protein that is to be discussed in Section 11.3. The following is an example of the stepwise folding method,[31,33–37] and the buffers used were described in these studies.

Step 1: The unfolded protein (U) was obtained by treating the precipitate or inclusion body with denaturing/unfolding buffer to make it 10 mg/mL in concentration. This solution was left at room temperature for 1 h. This process was meant to relax the protein structure by urea and pH (acidic or basic) environments. The disulfide bridges were reduced to SH groups and the protein was unfolded completely.

Step 2: The unfolded protein (U) in the denature/unfolding buffer was dialyzed against the folding buffer 1 for 72 h to dilute the urea concentration to 2 M, producing intermediate 1, or M_1.

Step 3: M_2 was obtained by dialyzing M_1 against the folding buffer 2 for 24 h to dilute urea concentration to 1 M.

Step 4: M_3, an intermediate without denaturant (urea) in solution, was then obtained by dialyzing M_2 against the folding buffer 3 for 24 h.

Step 5: M_3 was further dialyzed against the folding buffer 4 for 24 h, and the pH changed from 11 to 8.8 to produce M_4.

Step 6: Finally, the chemical chaperonin mannitol was removed by dialyzing M_4 against the native buffer for 8 h to yield M_5.

It should be noted that all the equilibrium time of each step is longer than the conventional dialysis time. In general, for the free solvent case, the solute may exchange with the buffer completely within hours. However, it is known that the denaturant molecules interact with protein, similar to the Donna effect, and the solute exchange may be slow and needs more time for the system to reach thermal equilibrium, especially for the first refolding stage. The folding time of each process is relatively longer than the regular solvent exchange process. Therefore, we can obtain the magnetic protein that follows a similar process. Protein microenvironment protects the net electron spin of molecules from thermal fluctuation.

The bridging ligands (i.e., sulfur atom, S) between the magnetic ions may be responsible for aligning the electron spin of magnetic ions. As indicated in Figure 11.4, the valance bonding electrons of the bridging Cys may hop between the bonded metal ions, such as Mn^{2+} and Cd^{2+}; whereas the Cd^{2+} in the β metal cluster is rather important in restraining the orientation of the electron spins of the bridging sulfurs and in aligning the spins of Mn^{2+} in the metal binding clusters. Therefore, this electron hopping effect may turn the Mn,Cd-MT into a magnetic molecule. However, the protein backbone surrounding the β metal cluster may provide a strong restraining effect to overcome the thermal fluctuations from the environment. Therefore, the magnetization can be observed in room temperature. However, the geometrical symmetry of the spin arrangement in all Mn-MT may cause partial or complete cancellation of detectable magnetization. These results also indicated that the threshold temperature of the molecular magnet might rise to room temperature if the proper prosthetic environment, such as protein backbone, can be linked against the thermal fluctuation of the temperature.

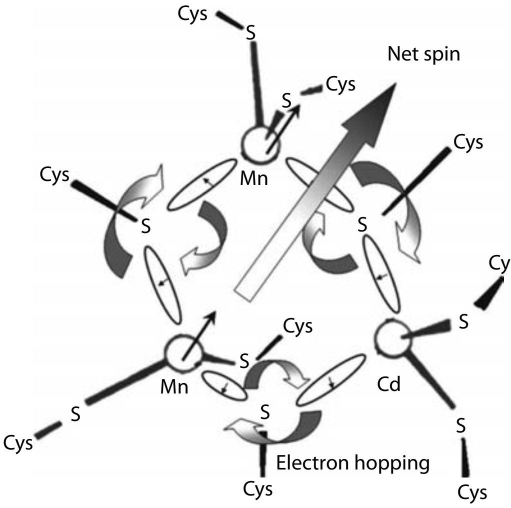

FIGURE 11.4 Proposed electron spin model of Mn^{2+} in β metal binding cluster of Mn,Cd-MT-2.

Therefore, we have successfully constructed a molecular magnet, Mn,Cd-MT, that is stable from 10 to 330 K. The observed magnetic moment can be explained by the highly ordered alignment of $(Mn_2CdS_9)^{3-}$ clusters embedded in the β-domain in which sulfur atoms serve as key bridging ligands. The discovery of mMT may allude to new schemes in constructing a completely different category of molecular magnets.

11.3 Nanostructured Semiconductor Templates: Nanofabrication and Patterning

The rapidly developing of interdisciplinary activity in nanostructuring is truly exciting. The intersections between the various disciplines are where much of the novel activity resides, and this activity is growing in importance. The basis of the field is any type of material (metal, ceramic, polymer, semiconductor, glass, and composite) created from nanoscale building blocks (clusters of nanoparticles, nanotube, nanolayers, etc.) that are themselves synthesized from atoms and molecules. Thus, the controlled synthesis of those building blocks and their subsequent assembly into nanostructures is one fundamental theme of this field. This theme draws upon all of the material-related disciplines from physics to chemistry to biology and to essentially all of the engineering disciplines as well.

The second and most fundamental important theme in this field is that the nanoscale building blocks, because of their size being below about 100 nm, impart to the nanostructures that are created from them new and improved properties and functionalities that are still unavailable in conventional materials and devices. The reason for this is that the materials in this size range can exhibit fundamentally new behavior when their sizes fall below the critical length scale associated with any given property. Thus, essentially any material property can be dramatically changed and engineered through the controlled size-selective synthesis and assembly of nanoscale building blocks. The present juncture is important in the fields of nanoscale solid-state physics, nanoelectronics, and molecular biology. The length scales and their associated physics and fabrication technology are all converging to the nanometer range.

The ability to fabricate structures with nanometer precision is of fundamental importance for any exploitation of nanotechnology. In particular, cost effective methods that are able to fabricate complex structures over large areas will be required. One of the main goals in the nanofabrication area is to develop general techniques for rapidly patterning large areas (a square centimeter, or more) with structures of nanometer sizes. Presently, electron-beam (e-beam) lithography is capable of defining patterns that are less than 10 nm. These patterns can then be transferred to a substrate using various ion milling/etching techniques. However, these are “heroic” experiments, and can only be made over a very limited area—typically a few thousand square microns, at most. While such areas are immediately useful for investigating the physics of nanostructures, the applications we would like to pursue will eventually require a faster writing scheme and much larger areas. The time and area constraints are determined by the direct e-beam writing. It is a “serial” process, defining single small regions at a time. Furthermore, the field of view for the e-beam system is typically less than 100 nm when defining the nanometer-scale structures. One simply cannot position the electron beam with nm precision over larger areas.

While e-beam lithographic methods are very general, in that essentially any shape can be written, we will also make use of “natural lithography” (tricks). Similarly, advances in the knowledge of the DNA structure has recently been applied to the fabrication of self-organized nm surface structures. There are many such “tricks” that could prove crucial to the success of the projects in allied fields. A list of current methods toward nanofabrication is given below.

11.3.1 Patterned Self-Assembly for Pattern Replication

By exploiting e-beam and focused ion beam lithography, self-assembled monolayers can be patterned into 10–20 nm features that can be functionalized with single molecules or small molecular groupings. These patterned areas will then be used as templates to direct the vertical assembly of stacks of molecules or to direct the growth of polymeric molecules. Schematics of the processes and the possible templates used are shown in Figure 11.5. The initial 2D pattern will thus be translated into 3D nanosized objects. With the capability of the full control over the interfacial properties, it will be possible to release the objects from the templates and transfer them to another substrate, after which the nanopatterned surface can be used again to provide an inexpensive replication technique.

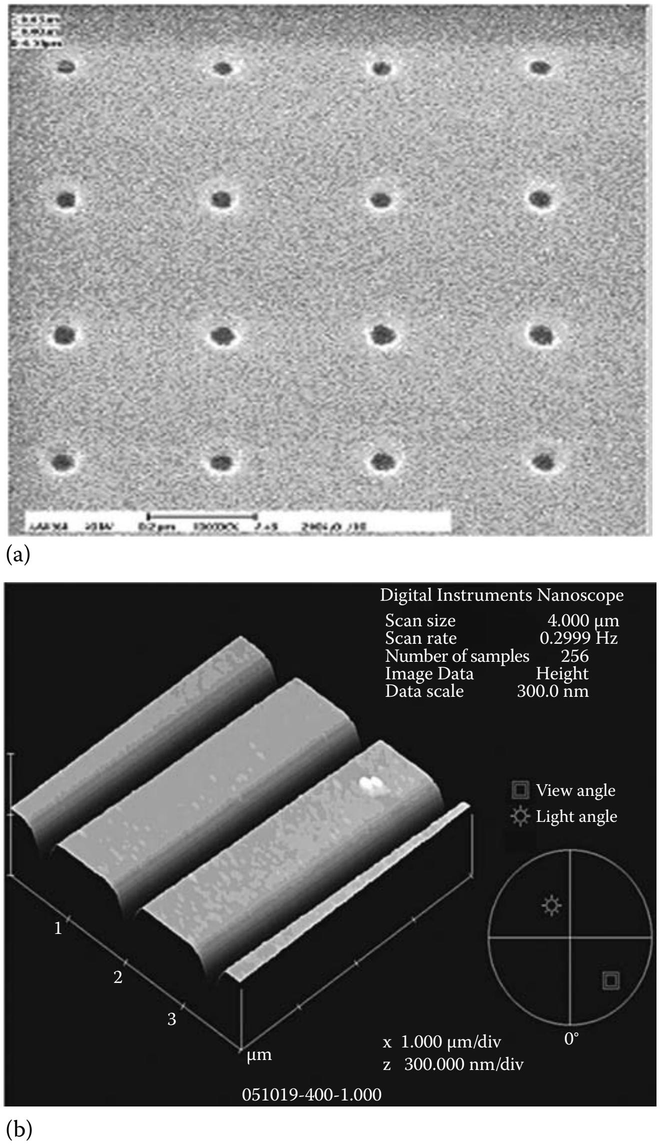

FIGURE 11.5 E-beam lithography defined (a) nanopores and (b) nanotrenches on silicon templates. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

11.3.2 Fabrication by Direct Inkjet and Mold Imprinting

This part of the technique will draw on recent advances in printing techniques for direct patterning of surfaces. This involves the use of inkjet printing to deliver a functional material (semiconductor or metal) to a substrate, which is then controlled by patterning in the surface free energy of the substrate, to allow very accurate patterning of the printed material. Currently, devices show channel lengths down to 5 µm, but indications are that geometries can be reduced to submicron dimensions. The method will be concerned with the limits to resolution that can be achieved by this process, and also the structure and the associated electronic structure at polymer–polymer interfaces, such as that between semiconductor and insulator layers in the field-effect device.

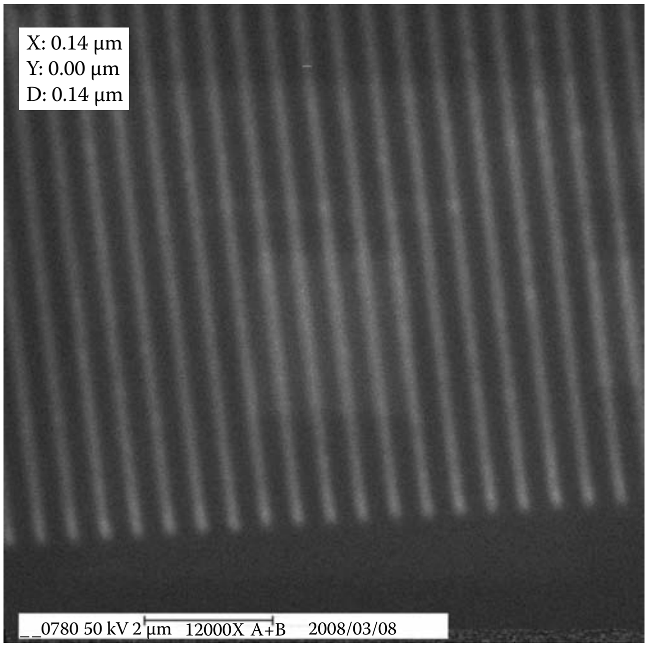

FIGURE 11.6 Nanogratings on silicon templates by using thermal nano-imprinting technique.

In 1995, Professor Stephen Y. Chou of Princeton University invented a new fabrication method in the field of semiconductor fabrication. It is called nano-imprint lithography (NIL).[57] Briefly speaking, this technique was demonstrated by pressing the patterned mold to contact with the polymer resist directly. The patterns on the mold will transfer to the polymer resist without any exposure source. Therefore, the diffraction effect of light can be ignored, and the limitation is dependent only on the pattern size of the mold rather than on the wavelength of the exposure light. In Figure 11.6 we show a nano-grating structure made by using the thermal imprinting technique.

NIL technology is a physical deformation process and is very different from conventional optical lithography. This technology provides a different way to fabricate nanostructures with easy processes, high throughput, and low cost. Currently, there are three main NIL techniques under investigation, namely, hot-embossing nano-imprint lithography (H-NIL[57]), ultraviolet nano-imprint lithography (UV-NIL[58]), and soft lithography.[59] Those NIL technologies can be applied to many different research fields, including nano-electric devices,[60] bio-chips,[61] micro-optic devices,[62] micro-fluidic channels,[63] etc.

11.3.3 Nanopatterned SAMs as 2D Templates for 3D Fabrication

Modern lithographic techniques (e-beam or focused ion beam (FIB), SNOM lithography) are able to generate topographic (relief) patterns in the 10–50 nm size ranges. As a further step toward more complicated and functional 3D structures, chemical functionalization at a similar size scale is necessary. By exploiting e-beam, FIB, and near-field optical lithography, it is possible to pattern SAMs directly. Using lithography, the SAMs can either be locally destroyed and refilled with other molecules, or the surface of the SAMs can be activated to allow further chemical reactions. The resulting patterned surfaces will be chemically patterned. Patterns can be introduced to incorporate H-bonding, pi–pi stacking, or chemical reactivity. These patterned areas will be used as templates to direct the vertical assembly of stacks of molecules or to direct the growth of polymeric molecules. Large, extended aromatic molecules prefer to stack on top of each other due to pi–pi stacking. These molecules have interesting electronic properties as molecular wires. When surfaces can be patterned to incorporate "seeds" for the large aromatic molecules, the stacking can be directed away from the surface. The initial 2D pattern will thus be translated into 3D nanosized objects. With the full control over the interfacial properties, it will be possible to release the objects from the templates and transfer them to another substrate, after which the nanopatterned surface can be used again to provide an inexpensive replication technique.

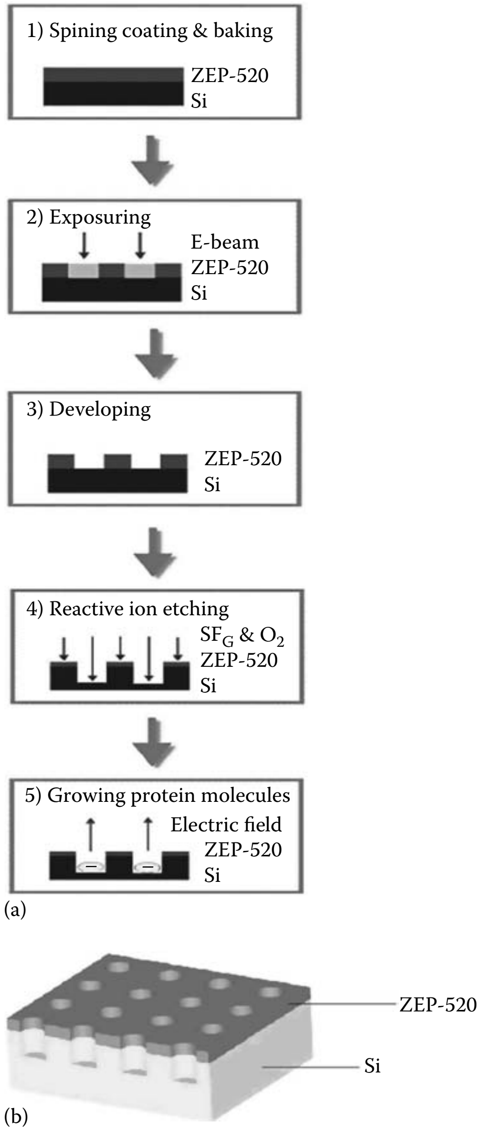

FIGURE 11.7 (a) Flowchart of the lithography, etching processes, and growth of protein molecules, (b) schematics of the patterned templates with nanopores. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

11.4 Self-Assembling Growth of Molecules on the Patterned Templates

The self-assembling growth of the MT-2 proteins is demonstrated as follows. One mg/mL magnetic MT in Tris. HCL buffer solution was placed onto the patterned surface, and an electric field with an intensity of 100 V/cm was then applied for 5 min to drive the MT molecules into the nanopores. The sample was then washed with DI water twice to remove the unbounded MT molecules and salts on the surface (the schematic of the process is also shown in Figure 11.7a. Figure 11.8 shows the atomic force microscopy (AFM) image of the template surface with 40 nm nanopores after they were filled by the MT-molecules. Keep in mind that most of the Si surface was still protected by photoresist after the etching processes, which has prevented the MT-molecules from forming strong OH bonds with the Si surface underneath. Therefore, the electrical field-driven MT molecules were all anchored on those areas that were not covered with photoresist. The molecules landing in each pore were then self-assembly grown vertically from the bottom of the pore into the shape of a rod (as shown in Figure 11.8). These molecular nanorods have an average height of ~120 nm above the template surface and a diameter equal to the size of the nanopore.

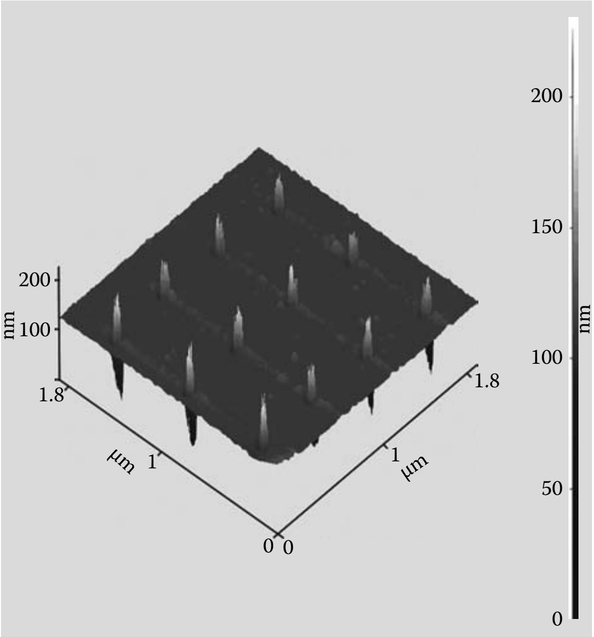

FIGURE 11.8 Three-dimensional AFM image of the patterned magnetic molecules. The molecules have self-assembled to grow into a rod shape. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

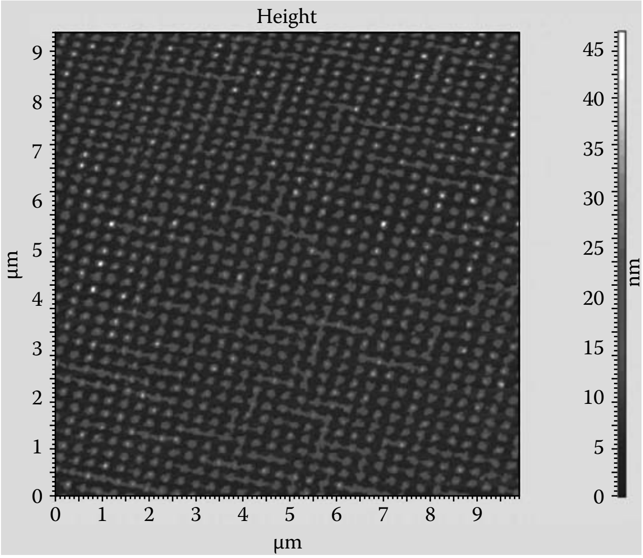

FIGURE 11.9 Two-dimensional AFM images of the patterned MT-molecules on the template with pore size of 130 nm and pitch size of 300 nm. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

However, experiments on the templates with pore sizes larger than 100 nm gave quite different results. Figure 11.9 shows the two-dimensional AFM image of the template surface with larger pores, where we can see that the molecules did not grow vertically above the template surface. Therefore, we were not able to generate 3D images of this type of template. However, judging from the AFM phase images, the MT molecules did form a more dense structure in the larger pores compared with the case of the smaller pores. On templates with thinner photoresist and smaller pitch sizes (less than 600 nm), we also found that the molecules anchored in the pore can grow laterally toward the neighboring pores (data not shown).

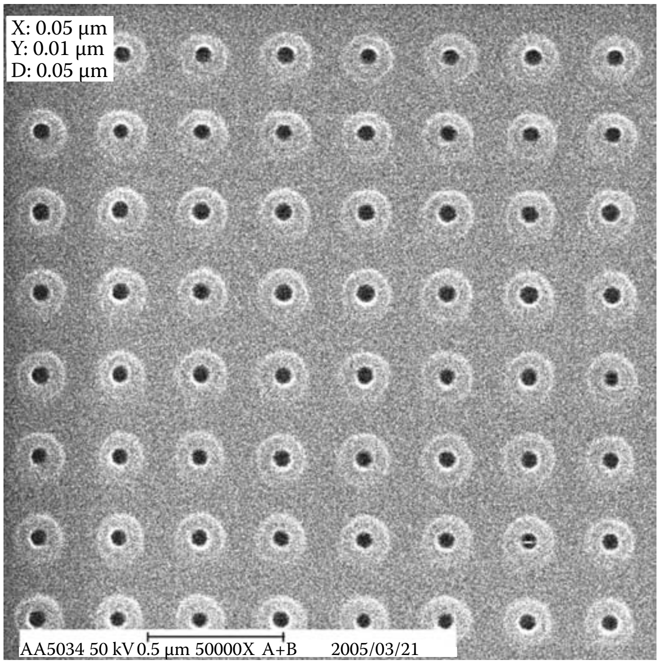

FIGURE 11.10 SEM image of the Si template shows a ring shape Si exposed area around the circumference of nanopores. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

By increasing both the e-beam exposure and dry etching time on the Si surface covered with a thin photoresist layer with a thickness of less than 150 nm, we were able to create a ring-type area with an exposed Si surface along the periphery of the nanopores. The SEM image of this type of template is shown in Figure 11.10. On this particular template, the molecules not only independently grew inside the pores, but they also grew along the circumference of the pores to form molecular rings on the template. Figure 11.11 shows the two-dimensional (2D) and three-dimensional (3D) AFM images of such molecular rings.

In order to gain better control of the formation of molecular nanostructures, it is important to uncover the underlying self-assembling growth mechanism. Molecular self-assembly

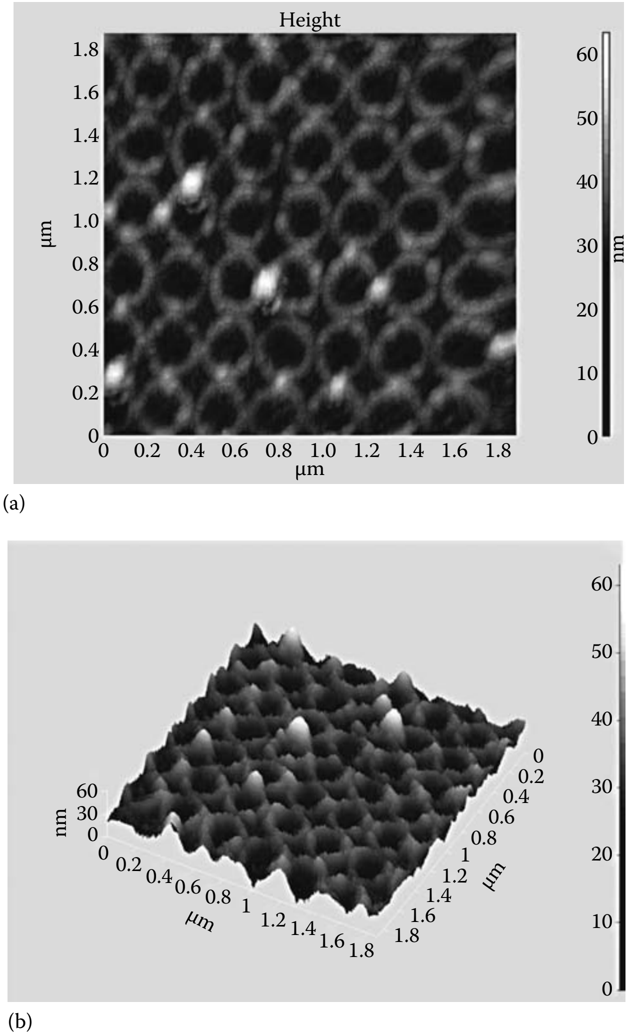

FIGURE 11.11 (a) 2D and (b) 3D AFM images of the patterned MT-molecules. (From Chang, C.-W. et al., *Appl. Phys. Lett.*, 88(26), 263104, 2006. With permission. AIP.)

can be mediated by weak, noncovalent bonds—notably hydrogen bonds, ionic bonds (electrostatic interactions), hydrophobic interactions, van der Waals interactions, and water-mediated hydrogen bonds. Although these bonds are relatively insignificant in isolation, when combined together as a whole, they govern the structural conformation of all biological macromolecules and influence their interaction with other molecules. The water-mediated hydrogen bond is especially important for living systems, as all biological materials interact with water. We believe that the first layer of proteins anchored inside the nanopores was bonded with the Si surface dangling bonds. They have provided building blocks for proteins that arrived later. With the assistance of spatial confinement from the patterned nanostructures, the rest of the proteins are able to self-assemble via the van der Waals interactions and perform molecular self-assembly.

11.5 Magnetic Properties of Molecular Nanostructures

The magnetic properties of the self-assembled molecular nanorods were investigated with magnetic force microscopy (MFM). We monitored the change of contour of a particular nanorod on the template when an external magnetic field was applied. Figure 11.12a shows the MFM image of the nanorod without the external magnetic field. In Figure 11.12b, a magnetic field of 500 Oe was applied during the measurement with a field direction from the right to left. The strength of the field was kept at a minimum so as not to perturb the magnetic tip on the instrument. In Figure 11.12 we can see clearly that the contour of the nanorod has changed in shape as compared to the case with no applied field. It indicates that the molecular self-assembly carries a magnetic dipole moment that interacts with the external magnetic field.

11.6 Conclusion and Future Perspectives

Success in the synthesis of the magnetic molecules produced from metallothionein (MT-2) by replacing the Zn atoms with Mn and Cd has been demonstrated in this chapter. Hysteresis behavior in the magnetic dipole momentum measurements was observed over a wide range of temperatures when an external magnetic field was scanned. These magnetic MT molecules were also found to self-assemble into nanostructures with various shapes depending on the nanostructures patterned on the Si templates. Data from the MFM measurements indicate that these molecular self-assemblies also carry magnetic dipole momentum. Since the pore size, spacing and shape can easily and precisely be controlled by lithography and etching techniques, this work should open up a new path toward an entire class of new biomaterials that can be easily designed and prepared. The techniques developed in this particular work promise to facilitate the creation of many bio-related nanodevices and spintronics. Magnetic molecular self-assembly may find its use in data storage or magnetic recording systems, as an example. They can also act as spin biosensors and be placed at the gate of the semiconductor spin valve to control the spin current from the source to the drain. More importantly, this work should not be limited to MT-2 molecules

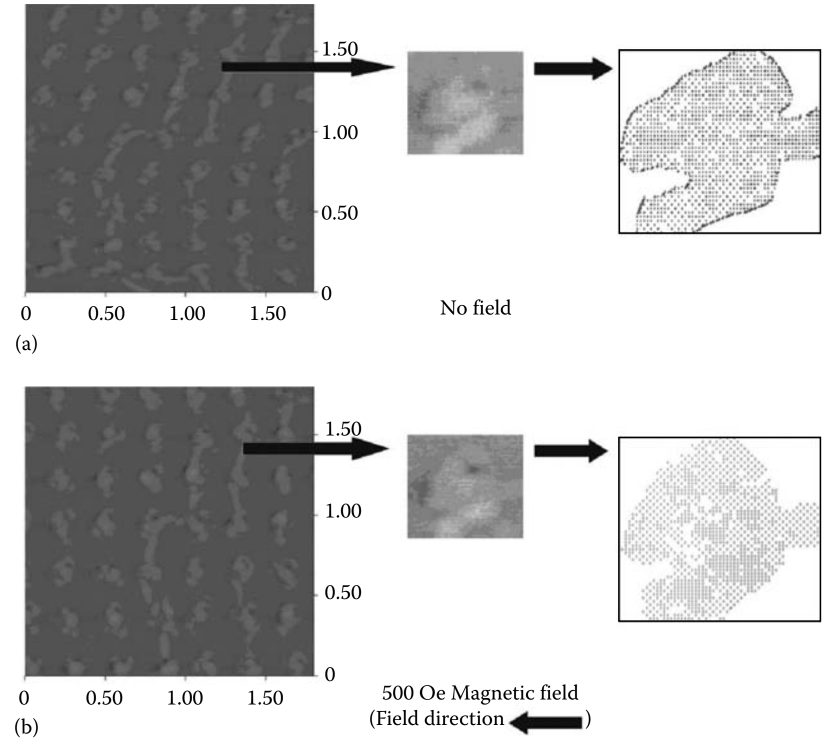

FIGURE 11.12 MFM images of nanorods (a) without the magnetic field (b) with a 500 Oe magnetic field applied with a field direction from right to left. (From Chang, C.-C. et al., *Biomaterials*, 28, 1941, 2007. With permission. Elsevier.)

and should be extended to other type of molecules and proteins as well. As mentioned in the beginning of this chapter, the various surface patterning techniques developed over the years to interface organic or biological materials with semiconductors have not only provided new tools for controlled 2D and 3D self-organized assemblies, but have also been essential to the creation and emergence of new semiconductor-molecular nanoelectronics as recently discussed by Likharev.[64]

In the future, it is important to develop techniques for growing and characterizing molecular self-assembly, single nanostructure, and molecule on semiconductor templates to bring a measure of control to the density, order, and size distribution of these molecular nanostructures. Using self-assembly techniques, one can routinely make molecular assembly with precise distances between them. Recent explorations of molecular self-assembly have sought to provide transverse dimensions on the mesoscopic nanometer scale. As a general—although not inviolate—rule, these attempts have led to very good local ordering (e.g., nearest neighbors).

We can anticipate, (1) the development of new (supra) molecular nanostructures via the self-assembling method (bottom up technique) and immobilization of single nanostructure or molecule; (2) the design of methods to functionalize molecular self-assembly and devices; (3) an integration of bottom-up and top-down procedures for the nano- and microfabrication of molecularly driven sensors, actuators, amplifiers, and switches; and (4) an increased understanding and appreciation of the science and engineering that lie behind nanoscale processes. All this and more is in the nature of the nanotechnology bonds as it impacts on biology and beyond. In the final analysis, however, the practice of biological synthesis that relies on molecular recognition and self-assembling processes within a very much more catholic framework than is currently being contemplated by most researchers that will dictate the pace of progress in synthesis.

The final goal is to understand this whole notion of what self-assembly is. One needs to really learn how to make use of the methods of organizing structures in more complicated ways than we can do now. On a molecular scale, the accurate and controlled application of intermolecular forces can lead to new and previously unachievable nanostructures. This is why molecular self-assembly (MSA) is a highly topical and promising field of research in nanotechnology today. MSA encompasses all structures formed by molecules selectively binding to a molecular site without external influence. With many complex examples all around us in nature (ourselves included), MSA is a widely observed phenomenon that has yet to be fully understood. Being more a physical principle than a single quantifiable property, it appears in engineering, physics, chemistry, and biochemistry, and is therefore truly interdisciplinary.

References

1. Sessoli R., Gatteschi D., Caneschi A., and Novak M.A. (1993), *Nature* (*London*), **365**, 141–143.
2. Fonticelli M., Azzaroni O., Benitez G., Martins M.E., Carro P., and Salvarezza R.C. (2004), *J. Phys. Chem. B*, **108**, 1898.
3. Park M., Harrison C., Chaikin P.M., Register R.A., and Adamson D.H. (1997), *Science* (*Washington*, *DC*), **276**, 1401–1404.
4. Li R.R., Dapkus P.D., Thompson M.E., Jeong W.G., Harrison C., Chaikin P.M., Register R.A., and Adamson D.H. (2000), *Appl. Phys. Lett.*, **76**, 1689–1691.
5. Thurn-Albrecht T., Schotter J., Kästle G.A., Emley N., Shibauchi T., Krusin-Elbaum L., Guarini K., Black C.T., Tuominen M.T., and Russell T.P. (2000), *Science* (*Washington, DC*), **290**, 2126–2129.
6. Kim H.C. et al. (2001), *Adv. Mater.*, **13**, 795–797.
7. Lopes W.A. and Jaeger H.M. (2001), *Nature*, **414**, 735–738.
8. Cheng J.Y. et al. (2001), *Adv. Mater.*, **13**, 1174–1178.
9. Kim S.O., Solak H.H., Stoykovich M.P., Ferrier N.J., de Pablo J.J., and Nealey P.F. (2003), *Nature*, **424**, 411–414.
10. Demers L.M., Ginger D.S., Park S.J., Li Z., Chung S.W., and Mirkin C.A. (2002), *Science* (*Washington*, *DC*), **296**, 1836–1838.
11. Hodneland C.D., Lee Y.S., Min D.H., and Mrksich M. (2002), *Proc. Natl. Acad. Sci.*, **99**, 5048–5052.
12. Houseman B.T. and Mrksich M. (2002), *Chem. Biol.*, **9**, 443–454.
13. Chen C.S., Mrksich M., Huang S., Whitesides G.M., and Ingber D.E. (1997), *Science* (*Washington*, *DC*), **276**, 1345–1347.
14. Mrksich M., Dike L.E., Tien J.Y., Ingber D.E., and Whitesides G.M. (1997), *Exp. Cell. Res.*, **235**, 305–313.
15. Chen C.S., Mrksich M., Huang S., Whitesides G.M., and Ingber D.E. (1998), *Biotech. Prog.*, **14**, 356–363.
16. Pitters J.L., Piva P.G., Tong X., and Wolkow R.A. (2003), *Nano Lett.*, **3**, 1431.
17. Wacaser B.A., Maughan M.J., Mowat I.A., Niederhauser T.L., Linford M.R., and Davis R.C. (2003), *Appl. Phys. Lett.*, **82**, 808.
18. Eaton D.L. (1985), *Toxicol. Appl. Pharmacol.*, **78**, 158–162.
19. Robbins A.H., McRee D.E., Williamson M., Collett S.A., Xuong N.H., Furey W.F., Wang B.C., and Stout C.D. (1991), *J. Mol. Biol.*, **221**, 1269–1293.
20. Messerle B.A., Schaffer A., Vasak M., Kagi J.H.R., and Wuthrich K. (1992), *J. Mol. Biol.* **225**, 433–443.
21. Boulanger Y., Goodman C.M., Forte C.P., Fesik S.W., and Armitage I.M. (1983), *Proc. Natl. Acad. Sci.*, **80**, 1501–1505.
22. Chang C.C. and Huang P.C. (1996), *Protein Eng.*, **9**, 1165–1172.
23. Wei S.H. and Zunger A. (1986), *Phys. Rev. Lett.*, **56**, 2391–2394.
24. Christou G., Gatteschi D., Hendrickson D.N., and Sessoli R. (2000), *MRS Bull.*, **25**, 66–71.
25. Sessoli R., Tsai H.L., Schake A.R., Wang S., Vincent J.B., Folting K., Gatteschi D., Christou G., and Hendrickson D.N. (1993), *J. Am. Chem. Soc.*, **115**, 1804–1816.
26. Aubin S.M.J., Dilley N.R., Pardi L., Kryzystek J., Wemple M.W., Brunel L.C., Maple M.B., Christou G., and Hendrickson D.N. (1998), *J. Am. Chem. Soc.*, **120**, 4991.
27. Friedman J.R., Sarachik M.P., Tejada J., and Ziolo R. (1996), *Phys. Rev. Lett.*, **76**, 3830–3833.

28. Bushby R.J. and Pailland J.-L. (1995), Molecular magnets, in *Introduction to Molecular Electronics*, M.C. Petty, M.R. Bryce, and D. Bloor (eds.), Edward Arnold, Hodder Headline, London, U.K.
29. Wolynes P.G. (1997), *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 6170–6175.
30. Zwanzig R. (1997), *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 148–150.
31. Chang C.-C., Yeh X.-C., Lee H.-T., Lin P.-Y., and Kan L.S. (2004), *Phys. Rev. E*, **70**, 011904.
32. Welker E., Wedemeyer W.J., Narayan M., and Scheraga H.A. (2001), *Biochemistry*, **40**, 9059–9064.
33. Chang C.-C., Su Y.-C., Cheng M.-S., and Kan L.S. (2002), *Phys. Rev. E*, **66**, 021903.
34. Chang C.-C., Tsai C.T., and Chang C.Y. (2002), *Protein Eng.*, **5**, 437–441.
35. Chang C.-C., Su Y.-C., Cheng M.-S., and Kan L.S. (2003), *J. Biomol. Struct. Dyn.*, **21**, 247–256.
36. Chang C.-C. and Kan L.-S. (2007), *Chin. J. Phys.*, **45**, 693–702.
37. Liu Y.-L., Lee H.-T., Chang C.-C., and Kan L.S. (2003), *Biochem. Biophys. Res. Commun.*, **306**, 59–63.
38. Levy Y., Jortner J., and Becker O. (2001), *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 2188–2193.
39. Creighton T.E. (1993), *Proteins*, W.H. Freeman, New York.
40. Elcock A.H. (1999), *J. Mol. Biol.*, **294**, 1051–1062.
41. Ulrih N.P., Anderluh G., Macek P., and Chalikian, T.V. (2004), *Biochemistry*, **43**, 9536–9545.
42. Despa F., Fernandez A., and Berry R.S. (2004), *Phys. Rev. Lett.*, **93**, 228104.
43. Seefeldt M.B., Ouyang J., Froland W.A., Carpenter J.F., and Randolph T.W. (2004), *Protein Sci.*, **13**, 2639–2650.
44. Anfinsen C.B., Haber E., Sela M., and White F.H. Jr. (1961), *Proc. Natl. Acad. Sci.*, **47**, 1309–1314.
45. Su Z.D., Arooz M.T., Chen H.M., Gross C.J., and Tsong T.Y. (1996), *Proc. Natl. Acad. Sci.*, **93**, 2539–2544.
46. Roder H., Maki K., Cheng H., and Shastry M.C. (2004), *Methods*, **34**, 15–27.
47. Royer C.A. (1995), *Methods Mol. Biol.*, **40**, 65–89.
48. Misawa S. and Kumagai I. (1999), *Biopolymers*, **51**, 297–307.
49. Dinsmore A.D., Crocker J.C., and Yodh A.G. (1998), *Curr. Opin. Colloid Interface Sci.*, **3**, 5–11.
50. Gast A.P. and Russel W.B. (1998), *Phys. Today*, **51**, 24–30.
51. Pusey P.N. and van Megan W. (1986), *Nature*, **320**, 340–341.
52. Ackerson B.J. and Pusey P.N. (1988), *Phys. Rev. Lett.*, **61**, 1033–1036.
53. Tsong T.Y., Hearn R.P., Warthal D.P., and Sturtevant J.M. (1970), *Biochemistry*, **9**, 2666–2677.
54. Tokuriki N., Kinjo M., Negi S., Hoshino M., Goto Y., Urabe I., and Yomo T. (2004), *Protein Sci.*, **13**, 125–133.
55. Lin K.-H., Crocker J.C., Prasad V., Schofield A., Weitz D.A., Lubensky T.C., and Yodh A.G. (2000), *Phys. Rev. Lett.*, **85**, 1770–1773.
56. Lau A.W.C., Lin K.-H., and Yodh A.G. (2002), *Phys. Rev. E*, **66**, 020401.
57. Stephen Y.C., Peter R.K., and Preston J. R. (1995), *Appl. Phys. Lett.*, **67**, 3114–3116.
58. Bender M., Otto M., Hadam B., Spangenberg B., and Kurz H. (2000), *Microelectron. Eng.*, **53**, 233–236.
59. Xia Y. and Whitesides G.M. (1998), *Angew. Chem. Int.*, **37**, 550–575.
60. Lingjie G., Peter R.K., and Stephen Y.C. (1997), *Appl. Phys. Lett.*, **71**, 1881–1883.
61. Pépin A., Youinou P., Studer V., Lebib A., and Chen Y. (2002), *Microelectron. Eng.*, **61–62**, 927–932.
62. Li M., Tan H., Chen L., Wang J., and Chou S.Y., (2003), *J. Vac. Sci. Technol. B*, **21**, 660–663.
63. Cho Y.H., Lee S.W., Kim B.J., and Fujii T. (2007), *Nanotechnology*, **18**, 465303.
64. Likharev K.K. (2003), Hybrid semiconductor–molecular nanoelectronics. *The Industrial Physicist*, June/July 2003, Forum: 20–23.
65. Chang C.-C., Sun K.W., Lee S.F., and Kan L.S. (2007), *Biomaterials*, **28**, 1941–1947.
66. Chang C.-C., Sun, K.W., Kan, L.-S., and Kuan, C.-H. (2006), *Appl. Phys. Lett.*, **88**, 263104.

II

Nanoscale Transistors

12 Transistor Structures for Nanoelectronics *Jean-Pierre Colinge and Jim Greer* **12**-1
Introduction • Evolution of Silicon Processing: Technology Boosters • Nonplanar Multi-Gate Transistors • The Rise of Quantum Effects • Carbon Nanotube Transistors • Nonclassical Transistor Structures • Graphene Ribbon Nanotransistors • Sub-kT/q Switch • Single-Electron Transistor • Spin Transistor • Molecular Tunnel Junctions • Point Contacts and Conductance Quantum • Atomic-Scale Technology Computer-Aided Design • Conclusion • References

13 Metal Nanolayer-Base Transistor *André Avelino Pasa* **13**-1
Introduction • Band Diagram • J–V Characteristic • Fabrication • References

14 ZnO Nanowire Field-Effect Transistors *Woong-Ki Hong, Gunho Jo, Sunghoon Song, Jongsun Maeng, and Takhee Lee* **14**-1
Introduction • Background • Results and Discussion • Summary • Acknowledgment • References

15 C_{60} Field Effect Transistors *Akihiro Hashimoto* **15**-1
Introduction • Fullerene (C_{60}) • Solid C_{60} • van der Waals Epitaxy of C_{60} • Principles of Organic FET Functions • Fabrication of Fullerene Field Effect Transistors (C_{60} FET) • Characterization of C_{60} FETs on SiO_2 or on AlN • Summary • References

16 The Cooper-Pair Transistor *José Aumentado* **16**-1
Introduction • Theory of Operation • Fabrication • Practical Operation and Performance • Present Status and Future Directions • Acknowledgment • References

12

Transistor Structures for Nanoelectronics

12.1 Introduction 12-1
12.2 Evolution of Silicon Processing: Technology Boosters 12-2
12.3 Nonplanar Multi-Gate Transistors 12-3
12.4 The Rise of Quantum Effects 12-5
12.5 Carbon Nanotube Transistors 12-6
12.6 Nonclassical Transistor Structures 12-7
12.7 Graphene Ribbon Nanotransistors 12-7
12.8 Sub-*kT/q* Switch 12-8
12.9 Single-Electron Transistor 12-8
12.10 Spin Transistor 12-9
12.11 Molecular Tunnel Junctions 12-9
12.12 Point Contacts and Conductance Quantum 12-11
12.13 Atomic-Scale Technology Computer-Aided Design 12-11
12.14 Conclusion 12-12
References 12-12

Jean-Pierre Colinge
University College Cork

Jim Greer
University College Cork

12.1 Introduction

In 1965, Gordon Moore predicted that the number of transistors that could be placed on a silicon chip would double every 18 months. This prediction became the official roadmap of the semiconductor industry and it still is the yardstick by which the progress of microelectronic devices and circuits is measured (Moore 1965). The MOS transistor, also called MOSFET (metal-oxide-semiconductor field-effect transistor) is the workhorse of the semiconductor industry and is at the heart of every digital circuit. Without the MOSFET, there would be no computer industry, no digital telecommunication systems, no video games, no pocket calculators, and no digital wristwatches.

Figure 12.1 shows the evolution of the number of MOS transistor per chip vs. calendar year, known as "Moore's law." Note that the vertical axis has a logarithmic scale. The effective length of the transistor, *L*, is the distance between the source and the drain in a region called the "channel region" where the electrostatics and the current flow are controlled by the gate. When the effective gate length becomes too small, the gate loses the control of that region and the so-called short-channel effects appear. These effects increase the OFF current of the device and render the current dependent of the drain voltage. In the most extreme cases, the transistor can no longer be turned off. The effective length of MOSFETs has shrunken from 10 µm in 1971 (Intel® 4004 processor) to 1 µm in the early 1980s, 0.1 µm in 2000, and 10 nm dimensions should be reached around 2015. Short-channel effects become so important in classical, planar MOSFETs with dimensions below 5 nm that new device structures that improve the electrostatic control of the channel by the gate are being explored. These devices are called multi-gate MOSFETs. They basically are silicon nanowires with a gate electrode wrapped around the channel region. This device architecture maximizes the control of the electrostatics in the channel region and allows one to reduce the effective channel length to sub-10 nm dimensions. At those dimensions, quantum transport effects start to kick in, and device simulators need to incorporate the Schrödinger equation to accurately model the transistor's electrical characteristics. Of course, the reduction of transistor size goes hand in hand with the increase of transistors on a chip. The first 32 Gb flash memory chips made using a 40 nm technology were announced in 2006. Each of these chips contains over 32 billion transistors. The number of transistor per chip is literally reaching astronomical proportions, considering that the 400 Gb mark will be reached around the year 2020. When this milestone is reached, there will be as many transistors on a single chip as there are stars in the Milky Way (Figure 12.1).

In 2008, the effective gate length of transistors that are used to fabricate microprocessors is 22 nm, and the thickness of the gate insulator is approximately 1 nm. Fully functional

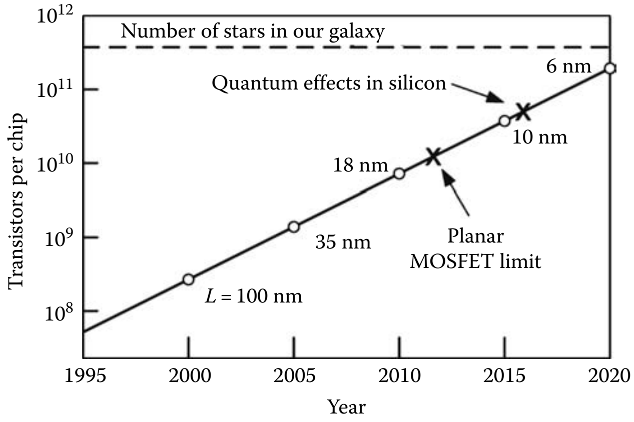

FIGURE 12.1 Moore's law: number of transistors per chip vs. calendar year.

"gate-all-around" transistors made in a silicon nanowire with a diameter of only 3 nm, that is, roughly the same diameter as a carbon nanotube, were reported in 2006 (Singh et al. 2006). Although these transistors are not formally called "nanoelectronic devices," they are clearly of nanometer dimensions. Strictly speaking, the nanoelectronics era has already begun.

Section 12.2 describes the evolution of the MOS transistor as its structure reaches nanometer dimensions: new materials are used to optimize electrical conductivity and dielectric constant, semiconductor alloying and the introduction of stress are used to increase carrier mobility, and novel transistor geometries are employed to increase the electrostatic control of the channel by the gate.

12.2 Evolution of Silicon Processing: Technology Boosters

In the 1980s, only a handful of elements were used in silicon chips: boron, phosphorus, arsenic, and antimony for doping the silicon, oxygen, and nitrogen for growing or depositing insulators, and aluminum for making interconnections. A few elements, such as hydrogen, argon, chlorine, and fluorine are used during processing in the form of etching plasmas or oxidation-enhancing agents. Sulfur is used in sulfuric acid to clean wafers. In the 1990s, a few more elements were added to the list, such as titanium, cobalt, and nickel, used to form low-resistivity metal silicides, tungsten, used to form vertical interconnects known as "plugs," and bromine, used in a plasma form to etch silicon. The 2000s saw an explosion in the number of elements used in silicon processing: lanthanide (rare earth) metals are being used to form oxides with high dielectric constants (high-k dielectrics), carbon and germanium are used to change the lattice parameter and induce mechanical stresses in silicon, fluorides of noble gases are used in excimer laser lithography, and a variety of metals are used to synthesize compounds that have desirable work functions or Schottky characteristics. Virtually all elements are used, with the notable exception of alkaline metals, which create mobile charges in oxides, and, of course, radioactive elements (Figure 12.2).

The use of new elements to obtain new desirable properties is a technology booster that has made it possible to extend the life of CMOS and reduce dimensions beyond barriers that were previously considered insurmountable. For instance, the reduction

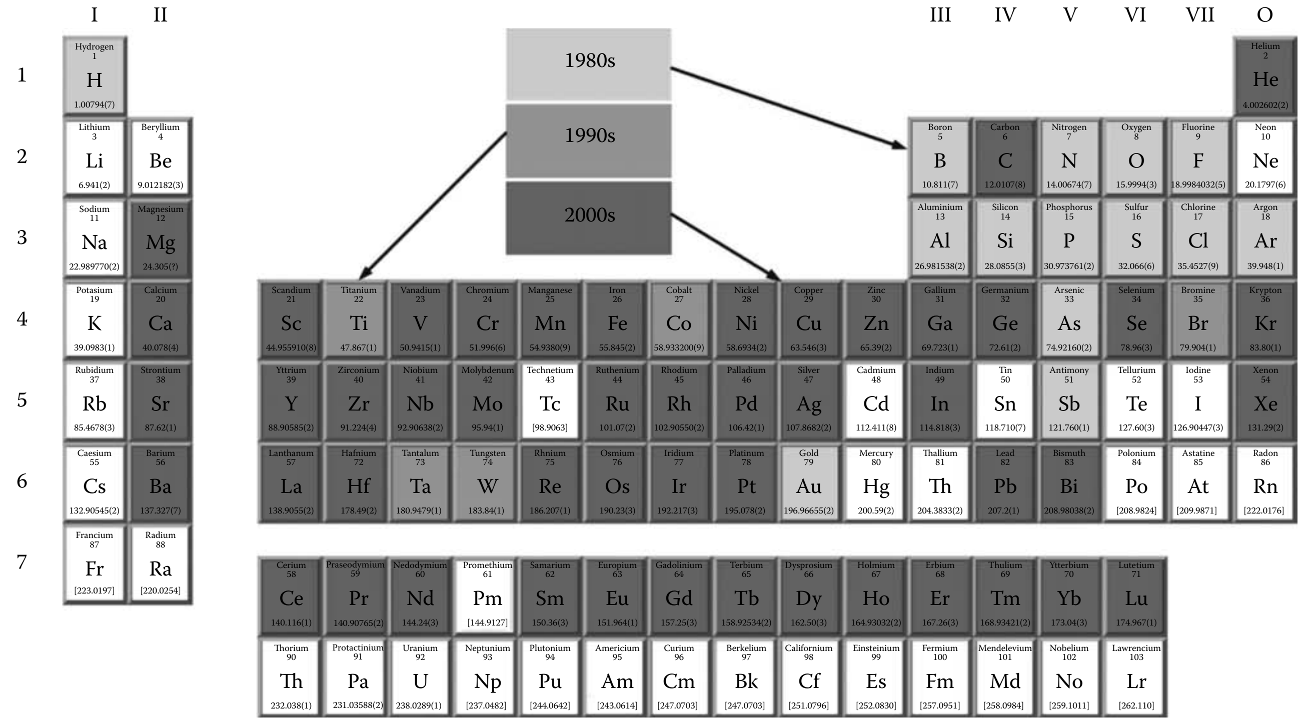
I
II
III
IV
V
VI
VII
O
1980s
1990s
2000s
1
2
3
4
5
6
7
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba La Hf Ta W Re Os Ir Pt Au Hg Th Pb Bi Po At Rn
Fr Ra
Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr

FIGURE 12.2 Elements used in silicon processing.

of gate oxide thickness below 1.5 nm leads to a gate tunnel current that quickly becomes prohibitively high. Replacing silicon dioxide by a high-k dielectrics such as hafnium oxide (HfO_2), which has a dielectric constant of 22 (vs. 3.9 for SiO_2), allows one to increase the thickness of the gate dielectric by a factor 22/3.9 = 5.5 without reducing the gate capacitance (i.e., the current drive of the transistor). The use of new gate dielectrics gave rise to the notion of equivalent oxide thickness, EOT, which is defined by the relationship $\text{EOT} = t_d(\varepsilon_{ox}/\varepsilon_d)$, where t_d is the thickness of the dielectric layer, and ε_{ox} and ε_d are the permittivity of silicon dioxide and the dielectric material, respectively. For example, a 4 nm thick layer of HfO_2 is electrically equivalent to an SiO_2 layer of 0.7 nm.

To improve the properties of transistors, another technology booster is commonly used: stress. Compressive stress increases hole mobility in (110) silicon (i.e., in the direction of current flow of most transistors), while tensile stress increases electron mobility. Mobility can also be modified by using Si:Ge or Si:Ge:C alloys. Compressive stress can be induced in the channel region of a transistor by introducing germanium in the source and drain. The resulting "swelling" of the silicon in the source and drain compresses the channel region situated between them. Tensile stress can readily be obtained by depositing a silicon nitride contact-etch stop layer on top of the device. Mobility (and thus speed) improvement in excess of 50% can be obtained using stress techniques.

The third technology booster deals with the physical geometry of the transistor and is worth being described in detail. As the dimensions of the transistors are shrunk, the close proximity between the source and the drain reduces the ability of the gate electrode to control the potential distribution and the flow of current in the channel region, and undesirable effects, called the "short-channel effects" render MOSFETs inoperable. For all practical purposes, it seems impossible to scale the dimensions of classical "bulk" MOSFETs below 15 nm. If that limitation cannot be overcome, Moore's law would reach an end around year 2012.

Short-channel effects arise when electric field lines from source and drain affect the control of the channel region by the gate. These reduce the threshold voltage V_{TH} according to the expression $V_{TH} = V_{TH\infty} - \Delta V_{TH} - \text{DIBL}$ where $V_{TH\infty}$ is the threshold voltage of a long-channel device, ΔV_{TH} is the "threshold voltage roll-off," due to the sharing of the space charge region underneath the gate between the gate, and the unbiased source and drain junctions, and DIBL is the "drain barrier lowering" which results from the increase of the share of the space charge region related to the drain junction when the drain voltage is increased. Short-channel effects can be minimized by reducing the junction depth and the gate oxide thickness. They can also be minimized by reducing the depletion depth through an increase in channel doping concentration. In modern devices, however, practical limits on the scaling of junction depth and gate oxide thickness lead to a significant increase of short channel effects and excessively large values of DIBL can quickly be reached.

12.3 Nonplanar Multi-Gate Transistors

The field lines in MOS transistors are illustrated graphically in Figure 12.3. In a bulk device (Figure 12.3A), the electric field lines propagate through the depletion regions associated with the junctions. Their influence on the channel can be reduced by increasing the doping concentration in the channel region. In very small devices, unfortunately, the doping concentration becomes very high ($>10^{19}$ cm^{-3}), which degrades carrier mobility. The situation can be improved by making the transistor in a thin film of silicon atop of an insulator. The silicon film is thin enough that it is fully depleted of majority carriers when a gate voltage is applied. Such a device is called fully depleted silicon-on-insulator (FDSOI) MOSFET. In FDSOI devices, most of the field lines propagate through the buried oxide (BOX) before reaching the channel region (Figure 12.3B). Short-channel effects can be further reduced by using a thin buried oxide and an underlying ground plane. In that case, most of the electric field lines from the source and drain terminate on the buried ground plane instead of the channel region (Figure 12.3C). This approach, however, has the inconvenience of increased junction capacitance and body effect (Xiong et al. 2002).

A much more efficient device configuration is obtained by using the double-gate transistor structure. This device structure was first proposed by Sekigawa and Hayashi in 1984 and was shown to reduce threshold voltage roll-off in short-channel devices (Sekigawa and Hayashi 1984). In a double-gate device, both gates are connected together. The electric field lines from source and drain underneath the device terminate

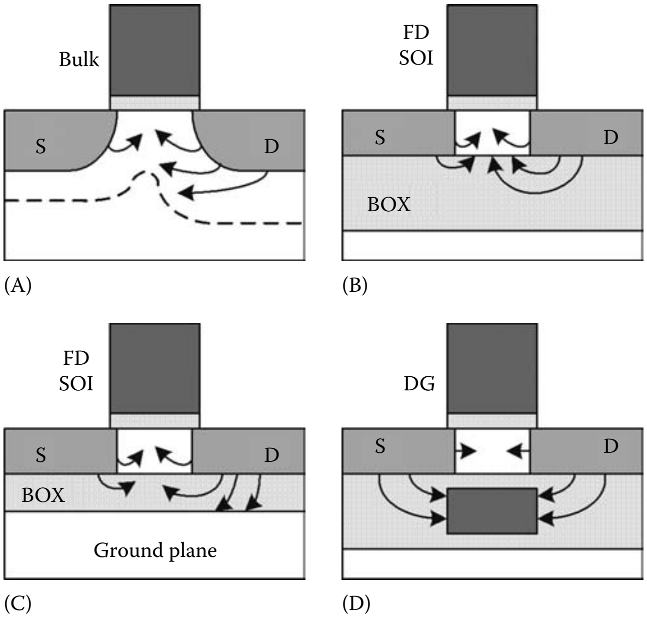

(A) (B)

(C) (D)

FIGURE 12.3 Encroachment of electric field lines from source and drain on the channel region in different types of MOSFETs: (A) Bulk MOSFET, (B) fully depleted SOI MOSFET, (C) fully depleted SOI MOSFET with thin buried oxide and ground plane, (D) double-gate MOSFET.

on the bottom gate electrode and, therefore, cannot reach the channel region (Figure 12.3D). Only the field lines that propagate through the silicon film itself can encroach on the channel region and degrade short-channel characteristics. This encroachment can be reduced by reducing the silicon film thickness.

Further improvement of short-channel effect control can be achieved by using a gate electrode that is wrapped around three sides of the channel region. The resulting device is known as a triple-gate FET, a term that includes the quantum-wire silicon-on-insulator (SOI) MOSFET (Baie et al. 1995) and the tri-gate MOSFET (Doyle et al. 2003, Kavalieros et al. 2006). The electrostatic integrity of triple-gate MOSFETs can be improved by extending the sidewall portions of the gate electrode to some depth in the buried oxide and underneath the channel region, which results in the formation of a Π-gate MOSFET (Park et al. 2001) or an Ω-gate FET (Yang et al. 2002, Ritzenthaler et al. 2006). The best possible electrostatic of the gate is obtained using "four" gates, or a gate that is wrapped around all sides of the channel region, thereby forming a gate-all-around silicon nanowire transistor (Singh et al. 2006). Using a combination of technology boosters, that is, using a tri-gate device in conjunction with strained silicon, a metal gate and/or high-k dielectric as gate insulator can be used to further enhance the performances of the device (Krivokapic et al. 2003, Andrieu et al. 2006, Kavalieros et al. 2006).

The quality of the electrostatic gate control over the channel can be expressed using a simple expression that can be derived from Poisson's equation (Colinge 2007a,b). It is found that short channel effects are basically nonexistent if the effective channel length is five times larger than that of a value called the "natural length" of the device, and noted λ. The natural length is given by

$$\lambda \cong \sqrt{\frac{\varepsilon_{si}}{n\varepsilon_{ox}} t_{si} t_{ox}}$$

where

t_{si} and t_{ox} are the thickness of the silicon film and the gate dielectric

ε_{si} and ε_{ox} is the permittivity of the silicon and the gate dielectric, respectively

n is the equivalent number of gates (ENG)

The ENG is basically equal to the physical number of sides of the devices covered by a gate and it expresses how efficiently the gate controls the electrostatics in the channel, assuming a square cross section ($W_{si} = t_{si}$). In a single gate device, n is equal to unity, $n = 2$ in a double-gate device, $n = 3$ in a tri-gate configuration, $n = 4$ in a gate-all-around (4 gates device), and $n \approx \pi$ in a π-gate device (Lee et al. 2007). The Ω-gate structure is a π-gate configuration with a larger lateral extension of the gate electrode underneath the channel region. The Ω-gate MOSFET has an effective number of gates larger than a π-gate, but smaller than a 4-gate device. The different configurations are shown in Figure 12.4.

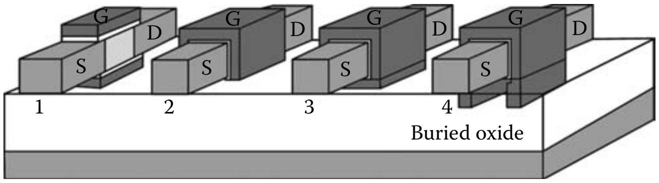

FIGURE 12.4 Different multi-gate MOSFET configurations. 1, double-gate; 2, triple-gate; 3, quadruple-gate or gate-all-around; 4, Π or Ω gate.

Short-channel effects are absent if the effective gate length is larger than seven to eight times the natural length, λ. This behavior is common to all gate configurations. Acceptable level of short-channel effects (DIBL < 50 mV and subthreshold slope <75 mV/decade) are obtained if the effective gate length, L, is smaller than 4λ (Figure 12.5) (Lee et al. 2007). Thus, for instance, to have the same short-channel behavior than a gate-all-around device, a single-gate SOI transistor needs to be made in a silicon film that is twice as thin.

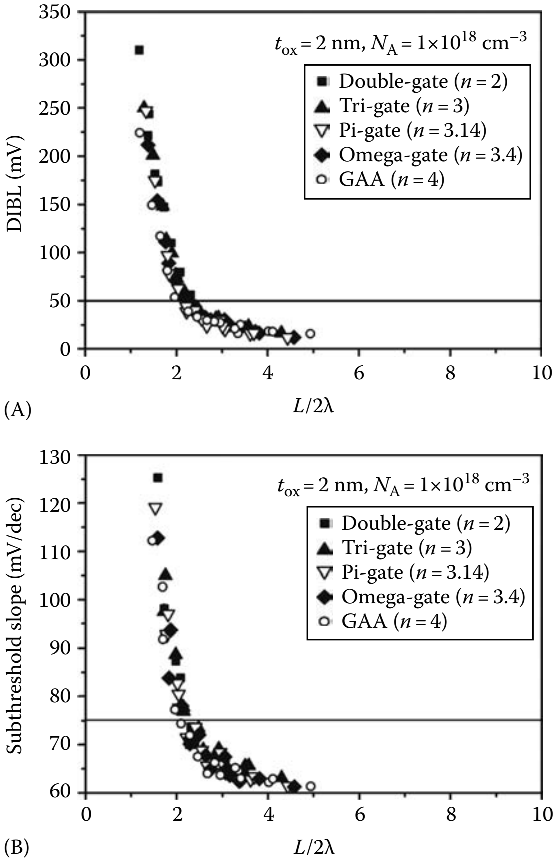

FIGURE 12.5 (A) Drain-induced barrier lowering (DIBL) and (B) subthreshold slope as a function of $L/2\lambda$ for different gate configurations. The gate oxide thickness is equal to 2 nm and the p-type channel doping concentration, N_A, is equal to 10^{18} cm^{-3}.

12.4 The Rise of Quantum Effects

As long as the cross section of the transistor stays above 10 nm classical models using the Poisson equation and drift-diffusion can be used. In smaller devices, however, quantum confinement of the electrons in two directions (perpendicular to the current flow) introduces some fundamental changes in the electronic properties of silicon. It is now necessary to solve the Schrödinger equation in conjunction with the Poisson equation. The density of states (DoS) for electrons in the conduction band in a three-dimensional crystal is a square root function of the electron energy: DoS $\approx \sqrt{E - E_{co}}$, where E_{co} is the minimum energy in the conduction band. In a two-dimensional crystal, the electrons are confined in one direction of space and free to move in the two other directions. As a result, energy subbands are formed within the conduction band. Each subband corresponds to a discrete energy level $E_{nz} = \hbar^2/2m(\pi n/t_{si})^2$ in the direction of confinement, z, and free motion in the two other directions of space (x and y). In this relationship, m is the mass of the electron, t_{si} is the crystal thickness in the z direction, and $n = 1, 2, 3, \ldots$. Within each subband the density of states is constant and thus independent of the electron energy: DoS = $(m/\hbar^2 \pi t_{si})$ (Ando et al. 1982). In a one-dimensional crystal, where the electrons are confined in two directions of space and free to move in the third direction, energy subbands are formed within the conduction band. Each subband corresponds to a discrete energy level $E_{mnyz} = \hbar^2/2m[(\pi m/W_{si})^2 + (\pi n/t_{si})^2]$ in the directions of confinement, y and z, and free motion in the third directions, x. In this relationship, t_{si} is the crystal thickness in the z direction, W_{si} is the crystal width in the y direction and m and $n = 1, 2, 3, \ldots$. Within each subband the density of states is shaped as a peak as a function of the electron energy: DoS $\approx (E - E_{nmyz})^{-1/2}$ (Davies 1998). This is illustrated in Figure 12.6, which shows the DoS as a function of energy in a tri-gate transistor with a thickness and width of 5 nm × 5 nm (Figure 12.6A) and a similar device with a thickness and width of 100 nm × 5 nm (Figure 12.6B). The density of states in 3D silicon is shown as well for comparison. In the device with the smaller section the DoS shows a series of peaks, which is characteristic of a one-dimensional system. In the taller device, the peaks gather to form steps of a staircase. Each step of the staircase corresponds to an energy subband of a 2D system. An ideal 2D device would be infinitely tall, while the one simulated here has a thickness of only 100 nm. One nonetheless can observe that the curve tends to a series of steps having each a constant DoS value, which is characteristic of an ideal 2D system (Colinge 2007a,b).

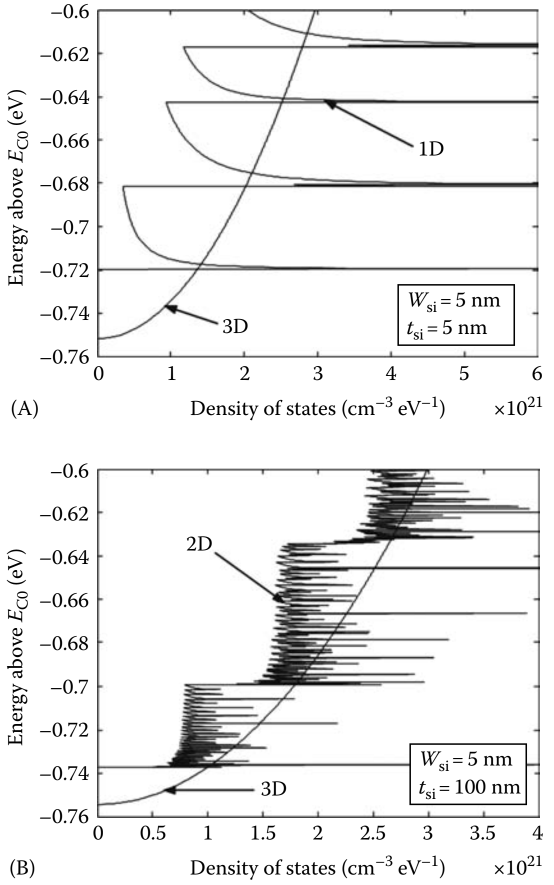

FIGURE 12.6 Density of states (DoS) vs. electron energy in a tri-gate MOSFET with (A) $W_{si} = t_{si} = 5$ nm and (B) $W_{si} = 100$ nm and $t_{si} = 5$ nm.

The confinement of carriers also modifies the spatial distribution of the electrons in the transistor. Let us take the example of a pi-gate device. Classical simulation tools predict that inversion channels form preferentially along the corner edges of the device (Figure 12.7A). A simulation of the same device

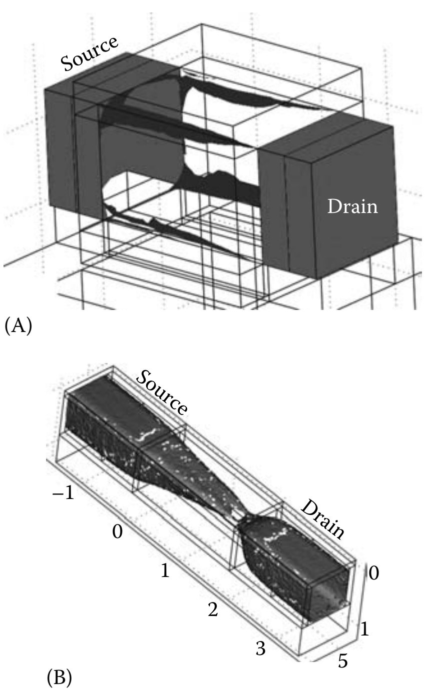

FIGURE 12.7 Electron isoconcentration contours in an n-channel pi-gate MOSFET with $t_{si} = W_{si} = 7$ nm and $L = 10$ nm. $V_{GS} = 0.7$ V and $V_{DS} = 0.5$ V. (A) Classical simulation using Poisson and drift-diffusion equations; (B) quantum simulation using Poisson and Schrödinger equations with nonequilibrium Green function formalism.

that includes Schrödinger's equation paints a very different figure: in this case, the electrons are confined to one or two channels located midway between the center of the device and the top corners (Figure 12.7B). Under the particular bias conditions used in Figure 12.7B the device is near saturation and the channel emanating from the source splits into two channels near the drain. Quite clearly, quantum confinement has profound influences on the electron distribution and, therefore, on scattering and transport mechanisms (Kim and Lundstrom 2008).

12.5 Carbon Nanotube Transistors

Carbon (C), silicon (Si), and germanium (Ge) are group IV elements, a description used to indicate that these elements share the same column in the periodic table of the elements. Another way of stating that all three are group IV elements is to say that they share the same configurations for their outer shell electrons. One consequence is that C, Si, and Ge can occur naturally in the diamond structure; see Figure 12.8A. Carbon is also commonly found in the graphite structure consisting of layers of carbon sheets. Each sheet is formed by a monolayer of carbon atoms arranged in a close packed hexagonal pattern. A single layer or carbon sheet in graphite is known as graphene. Carbon nanotubes (CNTs) are graphene strips rolled up into hollow columns or tubes. Carbon in three of its naturally occurring forms, diamond, graphene, and nanotube, is depicted within Figure 12.8. The CNT shown in Figure 12.8C is formed by wrapping a single graphene sheet into itself and onto a cylindrical shape. Nanotubes formed in this way are referred to as single walled carbon nanotubes (SWCNT). Although not difficult to produce, SWCNTs were not observed experimentally until 1991 (Iijima 1991, Bethune et al. 1993). Under typical growth conditions, many concentric tubes can form and this structure is referred to as multi-walled carbon nanotubes (MWCNT). Graphene can be wrapped onto a cylinder in different ways, referred to as the chirality or "handedness" of the nanotube. Specific chiralities give rise to different structures that are categorized as "armchair," "zig-zag," or simply "chiral." The specific structures give rise to different electronic properties for a specific CNT, which may be insulating, semiconducting, or metallic. Selecting the chirality, and hence controlling a CNT's electronic properties, remains a challenge to the growth and subsequent processing for these materials (Ding et al. 2008).

SWCNTs can have diameters of less than 1 nm but they are typically found within a range of 1–3 nm diameters, whereas their lengths can be on the order of centimeters. For a given diameter and chirality, the electronic, thermal, and mechanical properties between two SWCNT are well matched as the atomic structure of a tube is essentially defect free. This perfect atomic structure results in near ballistic transport, i.e., no scattering along the tube length, and is an attractive feature suggesting the use of CNTs for nanoelectronics. Also, the nanotubes are predicted to be able to carry current densities of up to three orders of magnitude larger (×1000) than typical conductors such as copper and aluminium making them attractive for applications

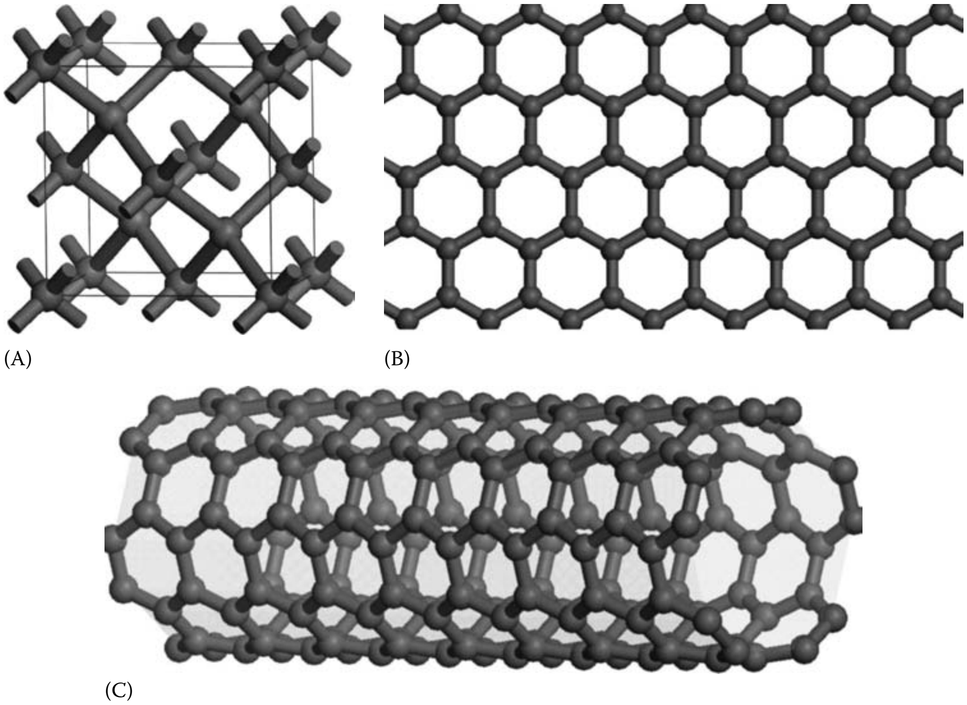

FIGURE 12.8 Carbon in three of its forms: (A) diamond structure, (B) graphene sheet, (C) an example of a carbon nanotube. Note that other forms of carbon exist including amorphous carbon and hollow spheres known as the fullerenes.

in nanoelectronic interconnects. Initially, use for CNTs in electronics focuses on two applications: as “vias” between different levels of metallization (the “wiring” in an integrated circuit) and as the channel material in field effect transistors.

The use of CNT’s as a material to fill vias is motivated by the need to have a highly conducting interconnection between two metallization layers, and the fact as circuit dimensions shrink, the cross sections of the vias also decreases. Extremely narrow vias are difficult to fill with conventional material deposition methods. One alternative is to deposit carbon in molecular form, for example, in the form of hollow carbon spheres C_{60} or “buckminsterfullerene.” The radius of the C_{60} molecule is on the order of 1 nm and can readily fill vias with cross sections of only a few square nanometers. As the deposited carbon is heated with an appropriate catalytic metal at the base of the via, CNTs grow. The resulting lengths of the CNTs are easily able to span between two metallization layers. However, as mentioned previously, it is presently not possible to control during growth whether a tube will be insulating, semiconducting, or metallic. Hence it is required to grow MWCNTs or bundles of SWCNTs to insure that an appropriate number of metallic tubes are contained within the via. This imposes a constraint on the minimum allowable cross section and introduces a statistical element to via resistances.

CNT field effect transistors (CNTFETs) are transistors fabricated with a CNT as the channel. In the first demonstrations of transistor with a CNT channel, a nanotube is placed from solution or other means onto an oxidized silicon substrate, and a metal layer is deposited and patterned as two metal electrodes. The resulting structure is a CNT spanning the gap between a metal source and drain (Martel et al. 1998, Tans et al. 1998). A cross section of the resulting structure is depicted in Figure 12.9. These original structures required the CNTs to be deposited and located, followed by the deposition of the electrodes. The silicon substrate is used as the gate electrode in a so-called back gate configuration, limiting the performance of the transistors fabricated in this manner. Typical metals used to form the source and drain contacts include but are not limited to Ti, Ni, Al, and Au. For semiconducting nanotubes, the metal/CNT junctions form a Schottky barrier and it is found that holes are injected into the nanotubes resulting in a p-channel MOS transistor. It has been determined that thermal treatments of the structures can be used to generate a shift in the source/drain Fermi levels, believed to be caused by driving oxygen out of the metal, with the result that n-channel CNTFET devices can be fabricated (Appenzeller et al. 2002).

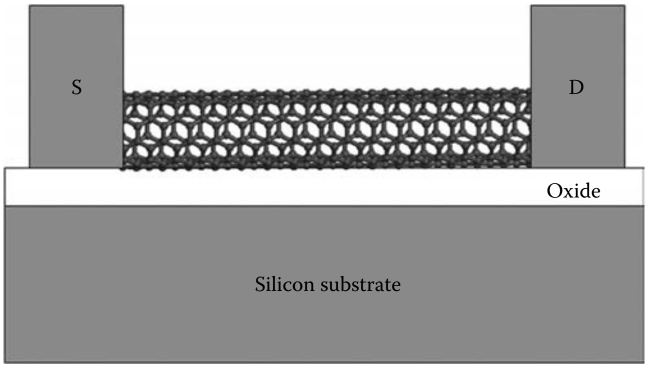

FIGURE 12.9 A CNT transistor in a back gated configuration. In this configuration, the CNT is connected between source and drain contacts and sits on an insulating oxide layer. The silicon substrate is used as a “back gate.”

There are now more sophisticated procedures for fabricating CNTFETs including deposited gate oxides and gate electrodes. CNTFETs have been fabricated and compared to Si MOSFETs and it is found that the CNTFETs can have lower switching delays for devices with similar on–off ratios (Javey et al. 2004, Seidel et al. 2005). CNTFET device layout and fabrication has not been fully optimized for high-frequency behavior, but the theoretical predictions for the high velocities achievable for the charge carriers in the CNTs and the measured results for high frequency performance on non-optimized structures indicate that ballistically limited CNTFETS should outperform ballistically limited Si FETs (Guo et al. 2005). However, in order to achieve the benefits of CNTs in large-scale integration schemes, progress is needed in the placement and controlled growth of nanotubes with preselected electronic character.

12.6 Nonclassical Transistor Structures

Classical transistors are based on the drift and diffusion of electrons and holes due to the presence of the electric fields and carrier concentration gradients. Novel transistor concepts based on quantum effects such as tunneling through a barrier, intraband tunneling, or spin polarization, open the possibility of improved performance and new functionality.

12.7 Graphene Ribbon Nanotransistors

As mentioned, one of the obstacles to the use of CNTs in nanoelectronics is the ability to grow nanotubes with a specific electronic character, that is, insulating, semiconducting, or conducting. Recently, the ability to produce single sheets of graphene with linear dimensions of micrometers (or “microns”) and with thicknesses of a single carbon atom has been achieved (Novoselov et al. 2004, 2005). In this sense, graphene is a truly “two-dimensional” material and displays unique electron properties—it is the only two-dimensional material with atomic thickness to be isolated to date. Graphene exhibits high conductivity and is being considered for use in high-speed electronic transistors. Due to the unique electronic structure of graphene, charge carriers move through a layer with zero mass and constant velocity. In common with CNTs, there is little scattering from defects. However, an ideal graphene sheet has a zero bandgap limiting its use in conventional electronic devices, although the unusual behavior of charge carriers in the material suggests it may be useful for nonclassical electronics. On the other hand, it is known that interactions of graphene with a semiconducting substrate can induce a bandgap of a fraction of an electron volt resulting in a narrow bandgap semiconductor.

Alternatively, for very narrow sheets or "nanoribbons" of graphene, the sheet edges can be chemically modified or "terminated" by different atoms or chemical groups. This results in a tailoring of the nanoribbon's electronic structure, and the resulting structure is a two-dimensional confined electron gas with lateral confinement lengths of only a few nanometers and a layer confinement length given by a single atom thickness. The material properties of graphene are only now beginning to be explored and transistor structures allowing for the introduction of local gate-tunable potential barriers within a graphene plane have been developed to explore charge transport mechanisms (Huard et al. 2007).

12.8 Sub-*kT/q* Switch

The power dissipated by a switching device is equal to $W = fC(V^2/2)$, where f is the switching frequency, C is the capacitance of the device, and V is the supply voltage. Quite clearly, reducing the supply voltage is the best way to reduce power consumption. Unfortunately, transistors, unlike mechanical switches and reed relays, cannot switch "instantly," and require a finite input voltage variation to switch from the off state to the on state. This behavior is fixed by the energy distribution of the electrons, and is a function of temperature T. The thermal energy is kT/q, where k is the Boltzmann constant, and q is the electron charge. At room temperature, the current in the best possible transistor needs an increase of input voltage of 60 mV to increase a decade (a 10-fold increase). This limit is referred to as the 60 mV/decade subthreshold slope. Thus, if one uses a supply voltage of 60 mV, the on/off current ratio of the transistor will only be equal to 10. In a chip that contains a billion of transistors, it is mandatory to reduce the off current to the subnanoampere level, which requires on/off current ratios in the order of 10^6 and thus a supply voltage of at very least 600 mV (in practice 0.8–1 V to get a decent current drive). Apart from cooling the devices in liquid nitrogen (where they require only 150 mV to achieve an on/off current ratio of 10^6), there is no real practical solution around the problem. Impact ionization in silicon-on-insulator (SOI) transistors has been long known to be capable of reducing the subthreshold slope below 60 mV/decade (Sundaresan and Chen 1990), but impact ionization requires large supply voltages, which defeats the purpose of realizing a low-voltage switch. More promising are devices based on band-to-band tunneling. Here, the variation of current is controlled by the variation of a tunnel barrier rather than by a change in electron concentration, which overcomes the kT/q limit. Such devices have been simulated and made in thin SOI films (Wang et al. 2004) and in carbon nanotubes (Appenzeller et al. 2004) but present technological challenges such as realizing extremely high doping concentrations in very thin films.

12.9 Single-Electron Transistor

The single-electron transistor (SET) comprises a source, a drain, a gate, and a channel region isolated from the source and drain by tunnel junctions. Since this region is usually very small, we will call it a "dot." When a small voltage is applied across a tunnel junction, it behaves like a capacitor, but electrons can tunnel through the barrier. If a constant voltage is applied to a tunnel junction, periodic current pulses, called "Coulomb oscillations," are produced. If the average current through the structure is I, the frequency of the Coulomb oscillations is equal to $f = I/q$, each pulse caused by the tunneling of a single electron through the barrier. Let us now consider two identical tunnel barriers in series (Figure 12.10A) with initial conditions $V_A = V_D = 0$ V. If voltage V_A is decreased, V_D is equal to $V_A/2$, since $C_1 = C_2 \equiv C$, but no current flows through the structure as long as the potential across C_1 is small enough. At the precise moment where the voltage across C_1 becomes large enough for an electron to tunnel through the junction, the voltage across C_1 is equal to $V_{At}-V_{At}/2 = V_{At}/2$. Note that the voltage across C_2 is also equal to $V_{At}/2$. When an electron tunnels into the dot, V_D jumps from $V_{At}/2$ to $V_{At}/2-qC_2$ (V_A and V_D are negative values). At that moment, the voltage across C_1 drops to $V_{At}/2 + qC$ that is smaller than $V_{At}/2$ in absolute value, which inhibits the tunneling of another electron from the source into the dot. The voltage across C_2, on the other hand, is equal to $V_{At}/2-qC$ that is larger than $V_{At}/2$ in absolute value, and the electron that was "stored" in the dot can now tunnel through C_2, giving rise to a drain current. In that process, the potential across C_1 is now increased above the critical value for tunneling, and the process repeats itself. The device is thus conducting if the applied voltage is larger, in absolute value, than a critical value, V_{At}. A more detailed analysis concludes that $V_{At} = -q/2C$ (Wasshuber et al. 1997). The region where no current flow is permitted ($V_{At} < V_A < -V_{At}$ with $V_{At} < 0$) is called the "Coulomb gap" (Figure 12.10B).

If we add a gate electrode that is capacitively coupled with the dot, we obtain a single-electron transistor. The dot potential, and thus the current flow, can now be modulated by the gate voltage. Depending on the gate voltage a discrete number of electrons

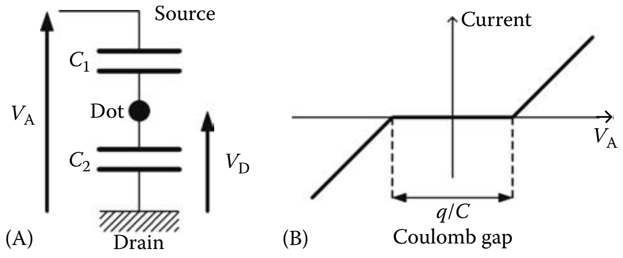

FIGURE 12.10 (A) Two tunnel barriers, C_1 and C_2, in series. (B) I–V characteristics of a double tunnel barrier structure.

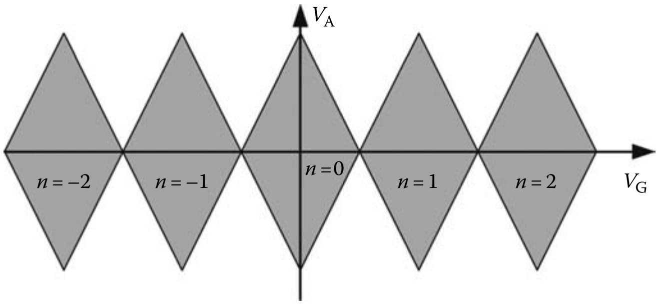

FIGURE 12.11 Current flow in an SET in the V_G–V_A plane. The shadowed areas indicate regions of the V_G–V_A plane where Coulomb blockade prevents current flow and $n = -2, -1, 0, 1, 2$ indicates the number of electrons stored in the dot. Outside these regions, the current increases with the gate voltage.

(or holes, $n = \ldots -2, -1, 0, 1, 2, \ldots$) can be stored in the dot, which produces a series of Coulomb gaps. These can be observed in the V_G–V_A plane as a series of rhombus-shaped regions where no current flows through the device (Figure 12.11).

12.10 Spin Transistor

Another nonclassical device is based on the manipulation of information not by a transfer of charge, but by a transfer or modification of electron spin. "Spintronics," in analogy with "electronics" groups a series of devices that use electron spin instead of charge flow to carry information. One advantage of spin over charge is that spin can be easily manipulated by externally applied magnetic fields, a property that has been utilized in magnetic storage technology for many years (Das Sarma 2001). Spin transistors involve the injection of spin-polarized electrons from an emitter made out of a ferroelectric material into a semiconductor. A ferroelectric gate is used to either conserve the spin or reverse it as the electrons flow through toward the collector. The collector is ferroelectric, with the same spin as the source. The electrons will reach the collector if their original spin has been conserved, but if they are no longer aligned with the direction of magnetization of the collector, no current can pass.

12.11 Molecular Tunnel Junctions

The idea using single molecules as circuit elements along with a proposal for building a molecular rectifier dates back over 30 years (Aviram and Ratner 1974). However, experimental developments to explore the use of single molecules in electronics did not significantly advance until the introduction of scanning probe microscopy (SPM) and the use of break junctions to measure current voltage characteristics for nanometer scale objects. Measurement of currents across single molecules has been investigated for just over a decade (Reed et al. 1997). State of the art photo- and e-beam lithography is not able to reliably produce metal lines with gap spacing of less than 1 nm as needed to contact single molecules. SPM and break junction methods allow for molecules to span gaps of a nanometer or less formed between two metal electrodes. The motivation for using SPM and break junctions to form molecular tunnel junctions is similar, and we briefly describe the formation of a break junction gap used in the investigation of molecular electronics.

In Figure 12.12, the idea behind a break junction is illustrated. Two triangular metal regions are patterned on a polymer substrate and they meet at the "apex" formed by each of the triangular regions. A piezoelectrically activated piston stretches the polymer layer leading to the "breaking" of the metal junction. Since piezoelectric actuation can be controlled on an Ångström-scale, a gap spacing of a few atomic lengths can separate the metallic contacts formed. Many ingenious methods for trapping molecules in the resulting gaps have been devised, but it remains true that the detailed structure of the resulting molecular junctions remains largely undetermined. Given the ability to trap single molecules in the gaps and to measure the current–voltage characteristics for simple tunnel junctions, many experiments and calculations of the currents across molecules have been performed (Cuniberti et al. 2005). A typical atomistic view of a single molecule bonded between two metal electrodes is given in Figure 12.13.

Although there remain discrepancies between different measurements, different theoretical treatments, and between theory and experiment, there is convergence and improvement in the

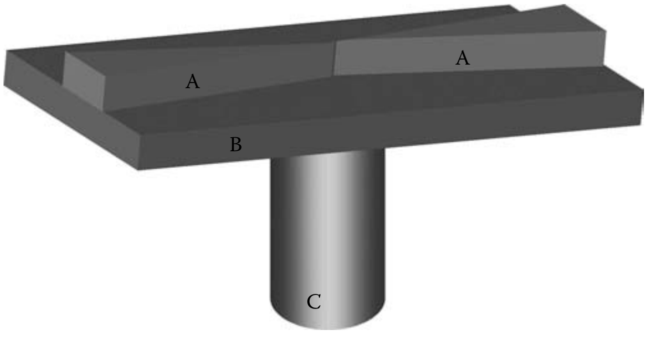

FIGURE 12.12 Schematic of a break junction experiment: Two triangular metal patterns A are deposited on a polymer substrate B and meet at a "point." The piston C under the substrate is piezoelectrically actuated causing the polymer layer to stretch resulting in the metal contacts to separate producing gaps of less than 1 nm.

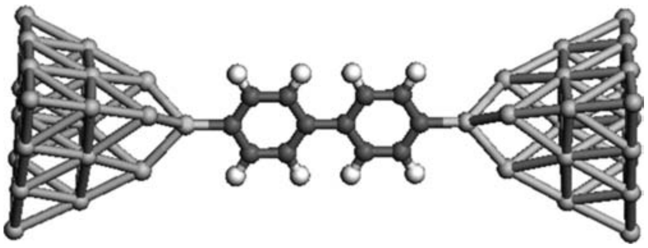

FIGURE 12.13 A "ball and stick" representation of the atomic positions and bonding structure of a 4,4′-biphenyldithiol (BPD) molecule bonded between two metal electrodes. Electrode atoms are metals, often gold is used. The "apex" of the electrode is a "linker", in this case a sulphur atom. The two molecular rings are formed by carbon atoms and "saturated" by hydrogen bonds.

reproducibility of the measurements of current–voltage characteristics in the experimental measurements on, for example, simple prototypical molecular junctions such as benzene dithiol (BDT) molecule, which is a benzene ring attached to the metal contacts through sulfur end groups (Reed et al. 1997, Dadosh et al. 2005, Lörtscher et al. 2007, Martin et al. 2008). Current experimental estimates for the resistance for a BDT molecule bonded between gold electrodes range between several MΩ to tens of MΩ. Early theoretical predictions often predicted currents orders of magnitude higher than those experimentally observed, and at least some of the reasons for these discrepancies have been identified and resolved (Delaney and Greer 2004, Lindsay and Ratner 2007). Theoretical predictions are now capable of predicting currents across molecular tunnel junctions within experimental uncertainties (Fagas and Greer 2007).

In Figure 12.14A, a representation of molecular tunnel junction is given in terms of an energy level diagram. On the left and right of the junction, the electrodes are represented as a set of dense energy levels characteristic of metals. Between the two electrodes, a few widely spaced energy levels represent the molecular region. For typical molecules being investigated in tunnel junctions, the molecular energy level spacings are on the order of an electron volt. If we take the Fermi level as the highest energy level in the electrodes, it is seen that in Figure 12.14A, one of the molecular levels lies above the Fermi energy and the other lies below. Hence, the highest energy level on the molecule is unoccupied and the lower level is filled. If a small voltage is applied across the junction, and if it is assumed that little current can flow through the lower occupied level (the electrons on the left and right "block" each other and current cannot flow), then it is seen that electrons leaving the electrodes near the Fermi level have no electronic states to flow across. The molecule in this case represents a barrier to the electrons, which must tunnel from the electrode at a lower voltage to that of a higher voltage. In this case, the current flow is described as nonresonant tunneling. As the voltage V across the tunnel is increased, the energy levels in the left and right electrodes are shifted up and down by $\varepsilon + eV/2$ and $\varepsilon - eV/2$, respectively, with the voltage converted to energy by multiplication of the electronic charge e. This leads to a situation where the unoccupied level on the molecule becomes "resonant" or aligned with an electrode energy level on one side of the molecular junction allowing current to flow freely between the electrodes as shown schematically in Figure 12.14B. In this case, the current across the molecule is due to resonant tunneling. These considerations can be quantified by what is commonly referred to as Landauer's formula (Datta 1997, Imry and Landauer 1999)

$$I = \frac{2e}{h}\int T(\varepsilon)[f_L(\varepsilon) - f_R(\varepsilon)]d\varepsilon,$$

with h Planck's constant. In this picture, states occupied in both the left and right electrodes "block" the current flow. This is described as the difference between the Fermi occupations $[f_L(\varepsilon) - f_R(\varepsilon)]$ for each state of energy ε in the left and right electrodes, respectively. For states that are not blocked, current can flow with transmission $T(\varepsilon)$. In Figure 12.14C, a typical current–voltage (IV) characteristic for a molecular junction is shown. At low voltages, the electrode states on one side of the junction become "unblocked" and current can flow, but there is no molecular state in the voltage window that allows electrons to traverse the molecule. The nonresonant tunneling behavior of the junction then implies a low value for the transmission coefficient and a corresponding small current. As a molecular state enters the voltage window as in Figure 12.14B, the transmission coefficient at energy ε increases substantially and results in a sudden increase in the current as the voltage increases. In Figure 12.14C, the onset of the resonant tunneling regime is clearly marked at the sudden increase in current above and below ±2 V.

Much of the recent work in molecular electronics is focused on describing the physics of current flow on a nanometer scale. There is effort needed to understand how to build switching or logic functions into a single molecule at a level that meets the stringent electronics requirements in terms of reproducibility, integration, and reliability, or indeed to pursue new functions for molecular electronics (Galperin et al. 2008). To explore the use of molecular electronics, efficient simulations of the behavior of molecules embedded into electronic circuits needs to be developed in a manner that allows the device models to be incorporated into circuit design methods (Fransson et al. 2006), but

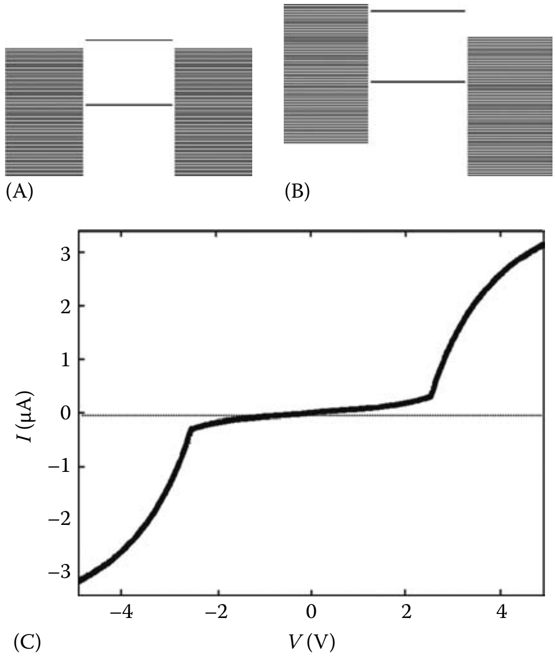

FIGURE 12.14 (A) An energy level diagram of a molecular tunnel junction. The dense set of electron states on the left and right represent the electrodes; the discrete set of widely separated states represents the molecule in the junction. (B) As a voltage is applied, the electrode states are shifted and the unoccupied molecular energy level enters the voltage window. (C) A representative current–voltage characteristic for a molecular tunnel junction. (After Delaney, P. and Greer, J.C., *Phys. Rev. Lett.*, 93, 036805-1, 2004.)

this work can be characterized as in its infancy. The goal in these approaches is to develop analytical models combining mesoscopic quantum transport equations with information about the molecular junctions determined from electronic structure calculations. Unknown physical parameters are then extracted by fitting to measurements of electron transport across molecular tunnel junctions, for example, for the prototypical case of benzene-1,4-dithiol bonded between gold electrodes. The coupling strength of the molecule attached to the metal electrodes is difficult to measure or calculate, so this parameter is extracted by minimizing the difference between the model predictions and measured data. The resulting model can be used to reproduce the *IV* characteristics for a molecule in a manner similar to that used for conventional compact models of transistors.

12.12 Point Contacts and Conductance Quantum

There is a remarkable consequence of Landauer's formula. In Figure 12.15, a pictorial representation of a break junction is given with a single atom bridging between two electrodes. It has been demonstrated using high-resolution transmission electron microscopy that a single atom can be reproducibly trapped in the break junction just prior to breaking of the junction (Rodrigues et al. 2000). For these junctions, a single atomic state can be in resonance with the electrodes and the transmission can be near ideal with $T \approx 1$ within the applied voltage window. Landauer's formula predicts the conductance in this case to be $g = \partial I/\partial V \approx 2e^2/h$, corresponding to a resistance of approximately 12.9 kΩ. This result indicates for a single quantum "channel" with perfectly ballistic transmission $T = 1$, there is an intrinsic resistance known as the "contact resistance." This resistance cannot be eliminated for a quantum conduction channel and in this sense, resistance is quantized per ballistic conducting channel. For metals with many parallel conducting channels, this intrinsically quantum effect is not noticeable. However for quantum scale conductors with a few or only one conducting channel, the contact resistance is significant and has been observed both in low-dimensional semiconductor devices (van Wees et al. 1988, Wharam et al. 1988) and for atomic point contacts (Scheer et al. 1998).

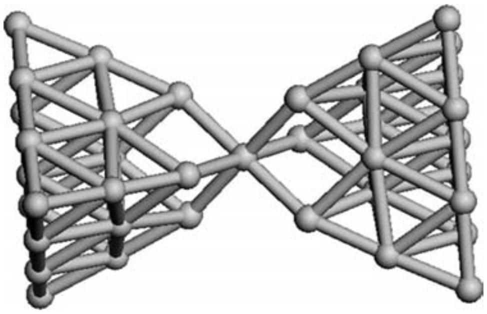

FIGURE 12.15 A "ball and stick" representation of quantum point contact. In this case, a single atom is located at the center of the junction.

12.13 Atomic-Scale Technology Computer-Aided Design

The continued down scaling of MOSFETs and the exploration of nonclassical transistor structures impose new requirements for technology computer-aided design (TCAD). For the "end-of-the-roadmap" silicon devices and for transistors relying on quantum mechanical effects for their operation, explicit atomic-scale structures of the devices and the ability to predict quantum behavior needs to be included in TCAD tools. Methods from computational chemistry and physics are being coupled to TCAD tools for design at the atomic and molecular length scales, and this represents a true change in simulation philosophy. Microelectronics device design was incremental for much of the period 1960–2000; few revolutionary technology changes were needed. Hence, parameterizations and simulations optimized for a production technology generation were adequate to predict properties for the design of the next technology generation. As the new technology generation was developed, re-parameterizations of the models were then made to refine and optimize transistors and circuits. In this way, technology design and manufacture leapfrogged one another forward. Now, the fields of molecular electronics, quantum computation, as well as classical alternatives are actively under investigation as means to overcome anticipated technology limitations. Hence, the nature of TCAD simulation is changing: instead of predicting incremental changes to existing technologies, new TCAD tools must be able to assess proposed candidates from a large set of replacement technologies in order to remain useful. And, useful has a rigorous definition in this field—TCAD simulation must be capable of reliably assessing new technologies, be cheaper and faster than design-by-trial strategies, with the constraints including design cycle time, cost, and accuracy.

The technology simulation community is not starting from scratch in terms of atomic scale and quantum computations. The development of algorithms and programs dates back to early days of quantum theory (Hartree 1928). However, the development of algorithms and programs were not written with nanoelectronic TCAD needs in mind and with means for treating open quantum systems, that is, systems that exchange charge or energy with their environment as needed to describe transistors and other circuit elements driven by voltage sources are currently under development.

In Figure 12.16, a simple graphical representation of Moore's law deduced from the minimum transistor feature size is compared to the system sizes accessible to atomistic and ab initio calculations. Clearly increased computing power and memory resources have allowed for computation and simulation at larger and larger sizes. In the diagram, we have differentiated between ab initio calculations, which directly solve the electronic Schrödinger equation, and much less accurate but computationally cheaper atomistic simulations. Atomistic simulations rely on simple empirical models of atom–atom interactions, but cannot deliver any information about the electronic structure of materials. They require, however, many orders of magnitude

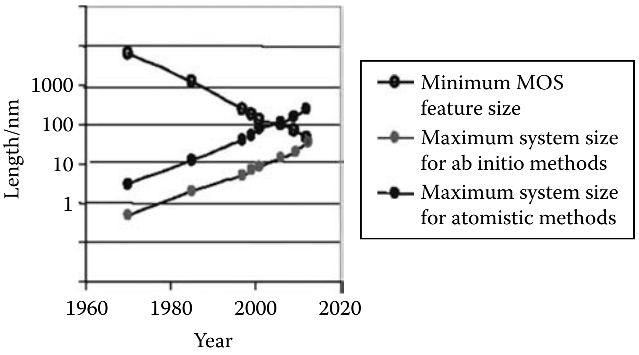

FIGURE 12.16 Convergence of atomistic and quantum mechanical simulations and nanoelectronics length scales.

less computational power for their description of matter and their use in the pharmaceutical and chemical industries is widespread. As a crude rule of thumb, accurate ab initio electronic structure codes can treat systems with a few hundred atoms to thousands of atoms, whereas atomistic methods can handle system sizes up to several tens of millions of atoms. Within a multiscale approach, atomistic simulations can be used to model aspects of the growth and processing of nanoscale structures. The electronic structure and charge transport in these systems can then be determined from ab initio techniques.

12.14 Conclusion

The size of transistors produced during the last 30 years by the microelectronics industry has shrunk by a factor of a hundred, from a few micrometers to a few tens of nanometers. As the limits of classical planar architecture are being reached, new transistor structures based on nanowire geometries are used to push the scaling down to nanometer dimensions. Quantum effects then begin to emerge, and the electrical characteristics of transistors no longer follow the predictions of classical simulation tools. The new behaviors imposed by quantum mechanics, on the other hand, pave the way for a new generation of devices with quantized current transport and nonconventional switching mechanisms.

References

Ando, T., Fowler, A.B., and Stern, F. 1982. Electronic properties of two-dimensional systems. *Review of Modern Physics* 54:437–672.

Andrieu, F., Dupré, C., Rochette, F., Faynot, O., Tosti, L., Buj, C., Rouchouze, E., Cassé, M., Ghyselen, B., Cayrefourcq, I., Brévard, L., Allain, F., Barbé, J.C., Cluzel, J., Vandooren, A., Denorme, S., Ernst, T., Fenouillet-Béranger, C., Jahan, C., Lafond, D., Dansas, H., Previtali, B., Colonna, J.P., Grampeix, H., Gaud, P., Mazuré, C., and Deleonibus, S. 2006. 25 nm short and narrow strained FDSOI with TiN/HfO$_2$ gate stack. *Symposium on VLSI Technology*, 16.4, Honolulu, HI.

Appenzeller, J., Knoch, J., Martel, R., Derycke, V., Wind, S.J., and Avouris, P. 2002. Carbon nanotube electronics. *IEEE Transactions on Nanotechnology* 1(4):84–89.

Appenzeller, J., Lin, Y.M., Knoch, J., and Avouris, P. 2004. Band-to-band tunneling in carbon nanotube field-effect transistors. *Physical Review Letters* 93(19):196805-1–196805-4.

Aviram, A. and Ratner, M.A. 1974. Molecular rectifiers. *Chemical Physics Letters* 29:277–283.

Baie, X., Colinge, J.P., Bayot, V., and Grivei, E. 1995. Quantum-wire effects in thin and narrow SOI MOSFETs. *Proceedings of the IEEE International SOI Conference*, Tucson, AZ, pp. 66–67.

Bethune, D.S., Klang, M.S., de Vries, M.S., Gorman, G., Savoy, R., Vazquez, J., and Beyers, R. 1993. Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. *Nature* 363:605–607.

Colinge, J.P. 2007a. Quantum-wire effects in trigate SOI MOSFETs. *Solid-State Electronics* 51(9):1153–1160.

Colinge, J.P. 2007b. *FinFETs and Other Multi-Gate Transistors*. Springer, Berlin, Germany.

Cuniberti, G., Fagas, G., and Richter, K. (eds.) 2005. *Introducing Molecular Electronics*, Lecture Notes in Physics. Springer, Berlin, Germany.

Dadosh, T., Gordin, Y., Krahne, R., Khivrich, I., Mahalu, D., Frydman, V., Sperling, J., Yacoby, A., and Bar-Joseph, I. 2005. Measurement of the conductance of single conjugated molecules. *Nature* 436:677–680.

Das Sarma, S. 2001. Spintronics. *American Scientist* 89:516–523.

Datta, S. 1997. *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, Cambridge, U.K.

Davies, J.H. 1998. *The Physics of Low-dimensional Devices*. Cambridge University Press, Cambridge, U.K.

Delaney, P. and Greer, J.C. 2004. Correlated electron transport for molecular electronics. *Physical Review Letters* 93:036805-1–036805-4.

Ding, F., Larsson, P., Larsson, J.A., Ahuja, R., Duan, H., Rosen, A., and Bolton, K. 2008. The importance of strong carbon-metal adhesion for catalytic nucleation of single-wall carbon nanotubes. *Nano Letters* 8:463–468.

Doyle, B.S., Datta, S., Doczy, M., Jin, B., Kavalieros, J., Linton, T., Murthy, A., Rios, R., and Chau, R. 2003. High performance fully-depleted tri-gate CMOS transistors. *IEEE Electron Device Letters* 24(4):263–265.

Fagas, G. and Greer, J.C. 2007. Tunnelling in alkanes anchored to gold electrodes via amine end groups. *Nanotechnology* 18:424010-1–424010-4.

Fransson, J., Bengone, O.M., Larsson, J.A., and Greer, J.C. 2006. A physical compact model for electron transport across single molecules. *IEEE Nanotechnology* 5(6):745–749.

Galperin, M., Ratner, M.A., Nitzan, A., and Troisi, A. 2008. Nuclear coupling and polarization in molecular transport: Beyond tunneling to function. *Science* 319(5866):1056–1060.

Guo, J., Hasan, S., Javey, A., Bosman, G., and Lundstrom, M. 2005. Assessment of high frequency performance of carbon nanotube transistors. *IEEE Transactions on Nanotechnology* 4(6):715–721.

Hartree, D.R. 1928. The wave mechanics of an atom with a non-Coulomb central field. *Proceedings of the Cambridge Philosophical Society* 24:89–132.

Huard, B., Sulpizio, J.A., Stander, N., Todd, K., Yang, B., and Goldhaber-Gordon, D. 2007. Transport measurements across a tunable potential barrier in graphene. *Physical Review Letters* 98:236803-1–236803-4.

Iijima, S. 1991. Helical microtubules of graphitic carbon. *Nature* 354(6348):56–58.

Imry, Y. and Landauer, R. 1999. Conductance viewed as transmission. *Review of Modern Physics* 71:S306–S312.

Javey, A., Guo, J., Farmer, D.B., Wang, Q., Wang, D., Gordon, R.G., Lundstrom, M., and Daih. 2004. Carbon nanotube field-effect transistors with integrated ohmic contacts and high-? gate dielectrics. *Nano Letters* 4(3):447–450.

Kavalieros, J., Doyle, B., Datta, S., Dewey, G., Doczy, M., Jin, B., Lionberger, D., Metz, M., Rachmady, W., Radosavljevic, M., Shah, U., Zelick, N., and Chau, R. 2006. Tri-gate transistor architecture with high-k gate dielectrics, metal gates and strain engineering. *Symposium on VLSI Technology* 7.1, Honolulu, HI.

Kim, R. and Lundstrom, M. 2008. Characteristic features of 1-D ballistic transport in nanowire MOSFETs. *IEEE Transactions on Nanotechnology* 7(6):787–794.

Krivokapic, Z., Tabery, C., Maszara, W., Xiang, Q., and Lin, M.R. 2003. High-.performance 45-nm CMOS technology with 20-nm multi-gate devices. *Extended Abstracts of the International Conference on Solid State Devices and Materials* (*SSDM*), Nagoya, Japan, pp. 760–761.

Lee, C.W., Yun, S.R.N., Yu, C.G., Park, J.T., and Colinge, J.P. 2007. Device design guidelines for nano-scale MuGFETs. *Solid-State Electronics* 51(3):505–510.

Lindsay, S.M. and Ratner, M.A. 2007. Molecular transport junctions: Clearing mists. *Advanced Materials* 19:23–31.

Lörtscher, E., Weber, H.B., and Riel, H. 2007. Statistical approach to investigating transport through single molecules. *Physical Review Letters* 98:176807-1–176807-4.

Martel, R., Schmidt, T., Shea, H.R., Hertel, T., and Avouris, P. 1998. Single- and multi-wall carbon nanotube field-effect transistors. *Applied Physics Letters* 73:2447–2449.

Martin, C.A., Ding, D., van der Zant, H.S.J., and van Ruitenbeek, J.M. 2008. Lithographic mechanical break junctions for single-molecule measurements in vacuum: Possibilities and limitations. *New Journal of Physics* 10:065008-1–065008-18.

Moore, G. 1965. Cramming more components onto integrated circuits. *Electronics* 38(8):114–117.

Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., and Firsov, A.A. 2004. Electric field effect in atomically thin carbon films. *Science* 306:666–669.

Novoselov, K.S., Jiang, D., Schedin, F., Booth, T.J., Khotkevich, V.V., Morozov, S.V., and Geim, A.K. 2005. Two-dimensional atomic crystals. *Proceedings of the National Academy of Sciences USA* 102:10451–10453.

Park, J.T., Colinge, J.P., and Diaz, C.H. 2001. Pi-gate SOI MOSFET. *IEEE Electron Device Letters* 22:405–406.

Reed, M.A., Zhou, C., Muller, C.J., Burgin, T.P., and Tour, J.M. 1997. Conductance of a molecular junction. *Science* 278:252–254.

Ritzenthaler, R., Dupré, C., Mescot, X., Faynot, O., Ernst, T., Barbé, J.C., Jahan, C., Brévard, L., Andrieu, F., Deleonibus, S., and Cristoloveanu, S. 2006. Mobility behavior in narrow Ω-gate FET devices. *Proceedings of the IEEE International SOI Conference*, Niagara falls, ON, pp. 77–78.

Rodrigues, V., Fuhrer, T., and Ugarte, D. 2000. Signature of atomic structure in the quantum conductance of gold nanowires. *Physical Review Letters* 85:4124–4127.

Scheer, E., Agraït, N., Cuevas, J.C., Yeyati, A.L., Ludoph, B., Martín-Rodero, A., Bollinger, G.R., van Ruitenbeek, J.M., and Urbina, C. 1998. The signature of chemical valence in the electrical conduction through a single atom contact. *Nature* 394:154–157.

Seidel, R.V., Gvaham, A.P., Kretz, J., Rajasekharan, B., Duesberg, G.S., Liebau, M., Unger, E., Kreupl, F., and Hoenlein, W. 2005. Sub-20nm short channel carbon nanotube transistors. *Nano Letters* 5(1):147–150.

Sekigawa, T. and Hayashi, Y. 1984. Calculated threshold-voltage characteristics of an XMOS transistor having an additional bottom gate. *Solid-State Electronics* 27:827–829.

Singh, N., Agarwal, A., Bera, L.K., Liow, T.Y., Yang, R., Rustagi, S.C., Tung, C.H., Kumar, R., Lo, G.Q., Balasubramanian, N., and Kwong, D.L. 2006. High-performance fully depleted silicon nanowire (diameter<5 nm) gate-all-around CMOS devices. *IEEE Electron Device Letters* 27(5):383–386.

Sundaresan, R. and Chen, C.E. 1990. SIMOX devices and circuits. *Electrochemical Society Proceedings* 90(6):437–454.

Tans, S.J., Verschueren, A.R.M., and Dekker, C. 1998. Room-temperature transistor based on a single carbon nanotube. *Nature* 393:49–52.

Van Wees, B.J., van Houten, H., Beenakker, C.W.J., Williamson, J.G., Kouwenhoven, L.P., van der Marel, D., and Foxon, C.T. 1988. Quantized conductance of point contacts in a two-dimensional electron gas. *Physical Review Letters* 60:848–850.

Wang, P.F., Hilsenbeck, K., Nirschl, T., Oswald, M., Stepper, C., Weis, M., Schmitt-Landsiedel, D., and Hansch, W. 2004. Complementary tunneling transistor for low power application. *Solid-State Electronics* 48:2281–2286.

Wasshuber, C., Kosina, H., and Selberherr, S. 1997. SIMON-A simulator for single-electron tunnel devices and circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 16(9):937–944.

Wharam, D.A., Thornton, T.J., Newbury, R., Pepper, M., Ahmed, H., Frost, J.E.F., Hasko, D.G., Peacock, D.C., Ritchie, D.A., and Jones, G.A.C. 1988. One-dimensional transport and the quantisation of the ballistic resistance. *Journal of Physics C* 21:L209–L214.

Xiong, W., Ramkumar, K., Jamg, S.J., Park, J.T., and Colinge, J.P. 2002. Self-aligned ground-plane FDSOI MOSFET. *Proceedings of the IEEE International SOI Conference*, San Jose, CA, pp. 23–24.

Yang, F.L., Chen, H.Y., Cheng, F.C., Huang, C.C., Chang, C.Y., Chiu, H.K., Lee, C.C., Chen, C.C., Huang, H.T., Chen, C.J., Tao, H.J., Yeo, Y.C., Liang, M.S., and Hu, C. 2002. 25 nm CMOS Omega FETs. *Technical Digest of IEDM*, New York, pp. 255–258.

13

Metal Nanolayer-Base Transistor

André Avelino Pasa
Universidade Federal
de Santa Catarina

13.1 Introduction **13**-1
13.2 Band Diagram **13**-1
13.3 *J–V* Characteristic **13**-3
13.4 Fabrication **13**-4
References **13**-7

13.1 Introduction

A metal nanolayer-base transistor is a transistor with a metallic layer with nanometric thickness sandwiched between two semiconductor layers. The metal forms the base and the semiconductors the emitter and the collector. This device was proposed and demonstrated at the beginning of the 1960s by Rose (1960), Geppert (1962), and Atalla and Kahng (1962). The main characteristic of the device for applied purposes was the expected high operation frequency when compared to bipolar transistors with low resistance and short transit times in the thin metallic nanolayer base.

13.2 Band Diagram

The mechanism of the metal-base transistor can be understood by drawing the energy band diagram of the sandwich structure with emphasis on the interfaces. The contact between a metal and a semiconductor forms a metal–semiconductor (MS) junction that has a special electric characteristic described independently by Schottky (1938) and Mott (1938). These junctions have a rectifying electrical behavior due to the presence of a potential barrier at the interface between the two materials. In Figure 13.1, we illustrate two different ways of performing MS contacts. The most common one at the time when metal-base transistors (MBT) emerged was by pressing a metallic tip against the surface of the semiconductor (a), known as the point contact, and the most reliable one used currently (b), obtained by depositing a metallic layer on top of a semiconductor substrate.

The band diagram of an MS contact is displayed in Figure 13.2a. The metal, represented by the energy of the Fermi level (E_{Fm}) and the semiconductor, represented by the bottom of the conduction band (E_{C}), and the top of the valence band (E_{V}), are in contact and under equilibrium conditions. The Fermi level in the semiconductor side (E_{FS}) is close to E_{C}, since we are considering an n-type material. In this case, the barrier at the interface ($q\phi_{\mathrm{b}}$), named the Schottky barrier, is due to the flow of electrons from the semiconductor to the metal to attain the equilibrium conditions, i.e., $E_{\mathrm{FM}} - E_{\mathrm{FS}}$ along the two materials corresponding in solid-state physics to the condition of the same chemical potential through the whole system. Electrons flow to the metal instead of from the metal to the semiconductor, since we are considering the work function of the metal ($q\phi_{\mathrm{m}}$) to be higher than the semiconductor affinity ($q\chi_{\mathrm{S}}$). The band banding in the semiconductor is then a consequence of the flow of electrons and leads to a depletion region and the appearance of an electric field close to the interface. The field is contrary to that flow and reaches its maximum intensity in the equilibrium condition. Figure 13.2b and c display the band banding conditions under direct and reversed bias, respectively. Electrons at E_{FM} have to overcome the Schottky barrier with a height of $q\phi_{\mathrm{b}}$, which is independent of the applied voltage, to cross the interface. On the other hand, electrons at the conduction band of the semiconductor are favored to flow to the metal by applying a forward bias as shown in Figure 13.2b. Due to the fact that the flow from the semiconductor to the metal is facilitated by the external voltage V, the MS interface acts as a diode rectifying the current, and the device is named Schottky diode. From the point of view of fundamental physics, this device is known as a hot-carrier diode. The carriers are injected with energy higher than the energy of the Fermi level. In the case of silicon, the most used semiconductor in the microelectronic industry, the height of the Schottky barrier varies in the range of 0.5 and 1.0 eV, this being the energy in excess of the carriers entering the metal. In Figure 13.2c, the diode is a reverse bias and no flux of charge is expected from both sides of the barrier. However, in real devices, a reverse current is always observed with a magnitude depending on the thermal ejection of carriers over the barrier, and most important, on the existence of defects that locally control the transport mechanism. Band diagrams similar to the ones shown in Figure 13.2 can also be drawn for p-type semiconductors.

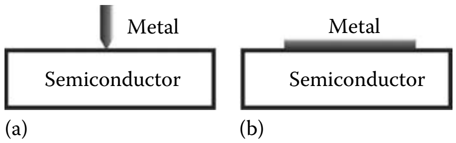

FIGURE 13.1 (a) Metal semiconductor contact obtained by pressing a metallic tip against the semiconductor surface (point contact). (b) MS contact formed by depositing a metallic layer on the surface of a semiconducting substrate.

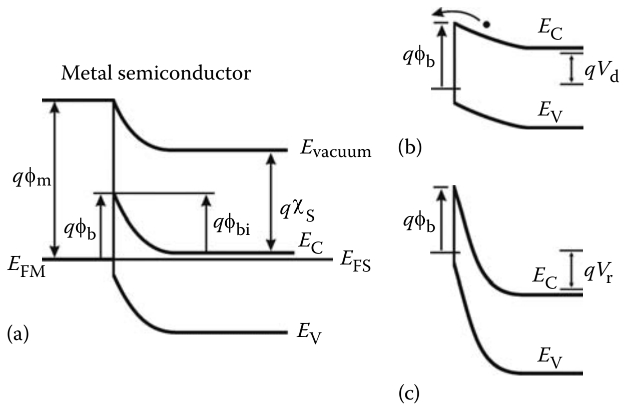

FIGURE 13.2 (a) Energy band diagram of the MS contact with a Schottky barrier at the interface. (b) and (c) Shift of conduction and valence bands of the semiconductor under influence of an external bias. The quantities indicated in the figure are: E_{FM}, Fermi level of the metal; $q\phi_b$, Schottky barrier; $q\chi_s$, electron affinity; $q\psi_{bi}$, built in potential; E_{FS}, Fermi level of the semiconductor; E_C, energy of the bottom of the conduction band; E_V, energy of the top of the valence band; E_{vacuum}, energy of the vacuum level; and q is just the electronic charge.

The current density J across an MS interface under the application of the external bias voltage V is described by the thermionic emission (TE) theory (Rhoderick and Williams 1988), which depicts the transport of carriers injected over the interface barrier $q\phi_b$, given by

$$J = J_o e^{\frac{qV}{nkT}}\left[1 - e^{\frac{-qV}{kT}}\right], \tag{13.1}$$

where

n is the ideality factor

k Boltzmann's constant

T the absolute temperature

J_o the saturation current density defined as

$$J_o = A^* T^2 e^{\frac{-q\phi_b}{kT}}, \tag{13.2}$$

where A^* is Richardson's constant, given by

$$A^* = \frac{4\pi q m^* k^2}{h^3}, \tag{13.3}$$

where

m^* is the effective mass

h is Planck's constant

For silicon, A^* is equal to 32 A/cm^2 K^2 for the p-type and 110 A/cm^2 K^2 for the n-type. The ideality factor n, equal to 1 corresponds to the case when the TE theory ideally describes the transport of charge over the potential barrier. For $n > 1$ are represented situations with deviations from the ideal case and are originated by the presence of, for example, inhomogeneities and interfacial layers.

Figure 13.3a shows the characteristic J–V curve of a Co/Si-n Schottky diode. The Co layer with a thickness of 80 nm was electrodeposited and the contact area was 0.5 cm^2. The curve follows the expected behavior of the TE theory and barrier heights and ideality factors of about 0.6 eV and 1.2, respectively, were obtained for this system (Zandonay et al. 2008). Figure 13.3b shows the characteristic of the diode when the logarithm function is applied to absolute values of current, i.e., the logarithm function was applied to the positive current values in the positive voltage range and to the negative current values, multiplied by −1, in the negative voltage range. This typical representation of the diode response is easily obtained through the observation of the magnitude of the reverse current and the linear interval of the forward current that follows the exponential dependence on the voltage predicted by the TE theory.

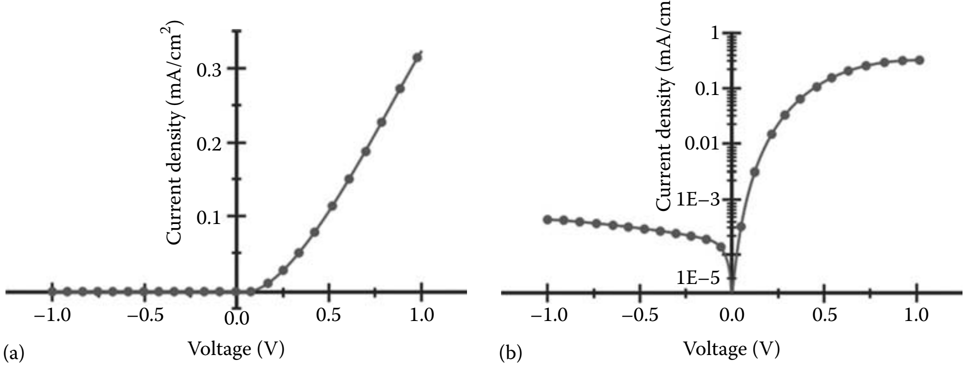

FIGURE 13.3 Linear (a) and logarithm (b) representation of the J–V electric characteristic of a Co/Si-n electrodeposited Schottky diode with a thickness of Co of about 80 nm and $q\phi_b$ = 0.60 eV and n = 1.2. (After Zandonay, R. et al., *ECS Trans.*, 14, 359, 2008.)

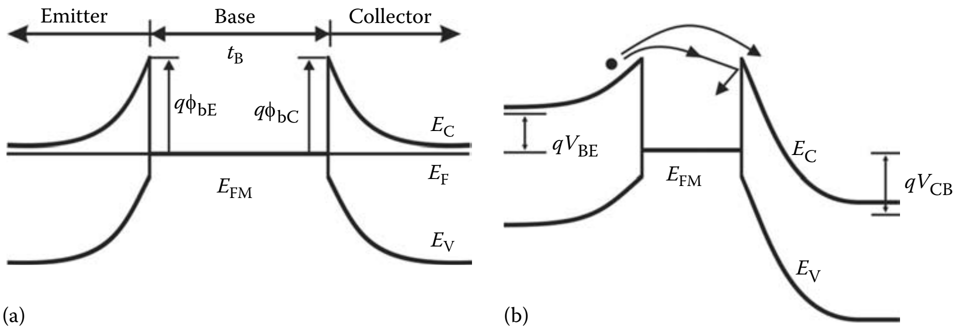

FIGURE 13.4 Band diagram of the MBT consisting of a semiconductor-metal-semiconductor planar trilayer structure under equilibrium conditions (a) and under external bias (b). Emitter-base diode forward biased and collector-base diode reverse biased. V_{CB} and V_{EB} are the applied voltages at the collector and the emitter relative to the base, $q\phi_{bE}$ and $q\phi_{bC}$ are the Schottky barrier heights at the emitter-base and collector-base interfaces, respectively, and t_B is the thickness of the base.

In an MBT, the metal layer of the base forms two Schottky barriers, as shown in Figure 13.4a, one with the emitter, and one with the collector. The barriers are placed face-to-face and can be polarized independently by applying voltages to the emitter and collector terminals relative to the base terminal. The forward operation mode of the transistor consists in applying a forward voltage at the emitter-base diode, to inject carriers into the base, and a reverse bias at the collector-base diode, in order to assist the collection of carriers crossing the base, as displayed in Figure 13.4b.

For the case of n-type semiconductors, as illustrated in Figure 13.4, the electrons ejected by the emitter have to have enough energy to cross the base and overcome the barrier at collector side. In order to have current gains close to 1, the collector barrier height should be lower than the barrier height at the emitter side and the base thickness should be relatively thin. In other words, the number of electrons that reaches the collector is reduced by processes occurring when traveling from the emitter to the collector through the base, which could be expressed in general as

$$J_C = \alpha_T J_E, \tag{13.4}$$

where α_T is the total transport factor of hot electrons that takes into account a series of individual factors, such as electron scattering in the metal base, quantum-mechanical reflection at the base-collector interface, and emission and collection efficiency due to interaction in the emitter and the collector interfaces (Sze 1969). All these processes contribute toward reducing the current that is measured in the collector-base circuit. Specifically, the base-transport factor, α_B, responsible for the scattering of the electrons in the base by collisions with electrons, phonons, impurities, and defects, can be written as

$$\alpha_B = e^{-\frac{t_B}{\lambda}}, \tag{13.5}$$

where λ is the mean free path of the injected hot electrons into the base. This dependency indicates that α_B, and consequently α_T, is strongly affected by the base thickness. Typical values for λ are a few tens of nanometers as $\lambda = 22$ nm observed for electrons with a 1 eV crossing Au layer sandwiched between Si-n and Ge-n at room temperature (Sze 1969).

13.3 *J–V* Characteristic

Similarly to bipolar transistors, the MBTs can be characterized electrically by measuring the performance of the device in the three different configurations: common-base, common-emitter, and common-collector. The common-base configuration is the most commonly used to obtain the electrical response of the transistor and to determine the common-base current gain α_o, defined by the ratio between collector– emitter current densities, i.e.,

$$\alpha_o = \frac{J_C}{J_E} = \alpha_T - \frac{J_{\text{reverse}}}{J_E}, \tag{13.6}$$

where J_{reverse} is the density of the current of the reverse biased collector-base diode. When the ratio between the reverse and the emitter current densities is considerable low, it can be neglected and $\alpha_o \approx \alpha_T$. In the common-base configuration, a fixed current is applied to the emitter and the collector-current density J_C is measured as a function of the collector-to-base voltage (V_{CB}). Figure 13.5 shows the vertical structure of the MBT, with emitter, base, and collector, in planar technology, and the circuit used for common-base measurements.

The transistor action for the common-base measurement corresponds to collector currents weakly dependent on the applied voltage between the collector and the base and strongly dependent on the emitter current. Figure 13.6 shows the typical response of a metal-base transistor for four different emitter currents densities, J_{E1} to J_{E4} with increasing values, showing a small increase of the collector current with the increase of the collector-to-base voltage, and the defined stages for the collector current that are determined by the intensity of the emitter current. The weak dependence of J_C on V_{CB} is attributed to a barrier height lowering with the reverse applied voltage at the collector Schottky diode

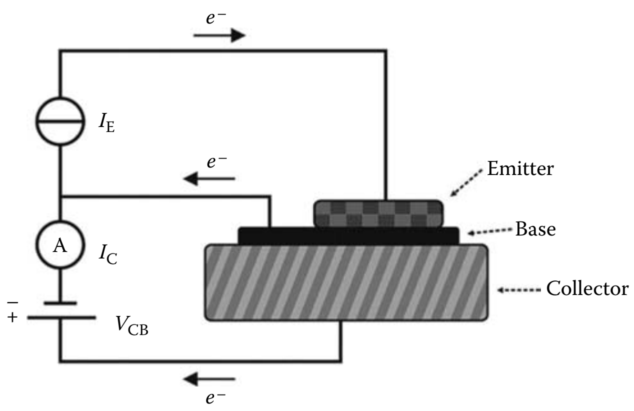

FIGURE 13.5 Schematic representation of a vertical structure with the emitter-base-collector of an MBT in planar technology. The bias circuit for common-base characterization is also shown. I_E is the current source connected in series to the emitter-base circuit and *A* is the current meter connected in series to the collector-base circuit that measures I_C, the collector current, as a function of voltage V_{CB} between the collector and the base.

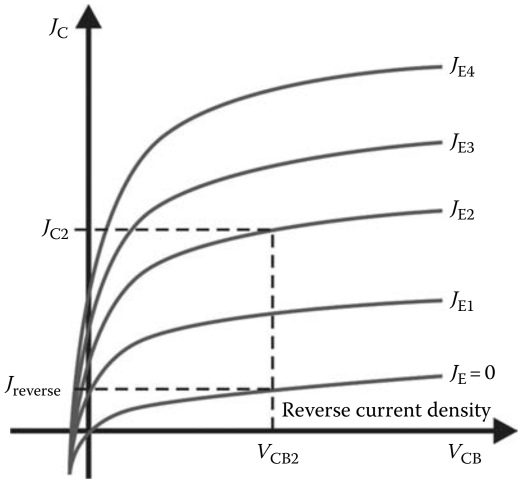

FIGURE 13.6 Typical common-base characteristic of an n-metal-n transistor with collector current levels determined by the intensity of the emitter current and weakly dependent on the collector-to-base reverse voltage.

(Sze 1969), and for this reason, the current gain has to be calculated for a fixed voltage as shown in the figure, where

$$\alpha_o = \frac{J_{C2}}{J_{E2}} - \frac{J_{reverse}}{J_{E2}}. \tag{13.7}$$

13.4 Fabrication

The main concern related to metal-base transistors is the fabrication process. It is necessary to prepare a structure with metal-semiconductor interfaces that have Schottky barriers and reduced losses for the electrons ejected by the emitter. Electronic devices are usually based on single crystal semiconductors and the growth of semiconductor epitaxial layers on top of metals is a difficult issue, requiring sophisticated methods. Epitaxial growth is a process whereby the atoms reaching the surface are deposited following the same atomic structure (order) of the crystalline substrate. Molecular beam epitaxy (MBE), liquid phase epitaxy (LPE), and chemical vapor deposition (CVD) are techniques that allow the growth of monolithic structures. This kind of process increases the production cost and reduces possibilities of large-scale integration (circuits with a large density of devices), requiring ultrahigh vacuum deposition or high temperature fabrication steps.

The pioneering work of Geppert (1962) and Atalla and Kahng (1962) were performed by depositing metal layers (base) on top of single crystal semiconductor substrates (collector). The emitter was a crystalline semiconductor piece with a tip or a blunt corner pressed against the metallic layer (point contact). Geppert electrodeposited Cu on n-Ge and contacted with an n-Ge probe. Atalla and Kahng deposited Au thin films with thicknesses of 10 nm on n-Ge and the point contact was n-Si. A few years later, Kanô (1964) used a vacuum-evaporated emitter, CdSe, an Au thin layer as base, and Ge as the collector to fabricate a structure with a Schottky barrier height higher at the emitter (0.75 eV) and lower at the collector (0.45 eV). However, in all these cases described above, the achieved current gains were not more than 0.5, and others factors, such as reproducibility and rough interfaces limited the performance of the device for applications.

To improve the quality of the interfaces, especially since one of the main issues to be worked out was the point contacts with a low yield of reproducibility, different approaches were tried. Brodsky and Deneuville (Deneuville and Brodsky 1978) grew amorphous semiconductor layers to surmount the need for monolithic structures. The glow discharge-deposition process was used for depositing hydrogenated silicon carbide (a-CSi:H) as the emitter, and hydrogenated silicon (a-Si:H) as the collector. The base was a Pt layer. The barrier heights at the emitter and the collector were 0.96 and 0.75 eV, respectively. Though the fact that the heights are adequate for the collection process, the common-base current gain measured was not higher than 0.1.

Moreover, no significant increase in the transfer ratio was observed by growing the semiconductor-metal-semiconductor (SMS) structure using the epitaxy techniques mentioned above. Rosencher et al. (1984) and Hensel et al. (1985) grew monolithic Si-$CoSi_2$-Si structures and obtained transistors with characteristic similar to the one shown in Figure 13.6. The monolithic atomic structure obtained by these authors is shown in Figure 13.7. $CoSi_2$ is a metallic disilicide with a resistivity comparable to metals and with a small lattice mismatch (~1.4%) with Si allowing the epitaxial growth.

Tung et al. (1986) observed high values of α_o, as high as 0.95, with MBTs grown by MBE. However, they also detected the existence of pinholes in the base that directly connected the emitter to the collector. The electron transfer was in this case mainly through the small openings in the base layer, where the barrier height is lower than the Schottky height. This mechanism of transport is the operation principle behind the permeable-base

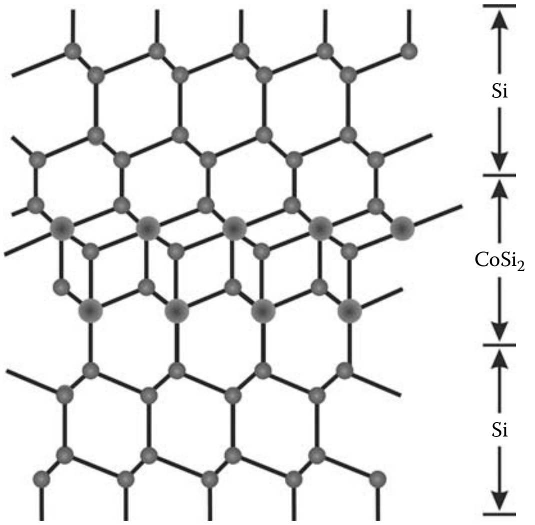

FIGURE 13.7 Atomic arrangement of the monocrystalline Si-$CoSi_2$-Si structure.

transistors (PMT), which incorporate a metal grid as base between the emitter and the collector (Bozler and Alley 1980). Even though the Si/$CoSi_2$/Si MBT showed a useful and a controllable current gain, its fabrication remained a challenge, requiring ultrahigh vacuum deposition and annealing.

The presence of pinholes in the base needs to be checked in all cases in order to see if the measured α_o is not a misleading interpretation of the direct contact between the emitter and the collector through the pinholes. A checking procedure was suggested by Rosencher et al. (1986) where the voltage between the emitter and the base is measured as a function of the collector-to-base voltage. The absence of pinholes leads to an independent voltage at the emitter-base junction, which is screened by the compact metal of the base.

A very significant innovation concerning metal nanolayer-base transistors was achieved with the advent of the spin transistor. Monsma et al. (1995) fabricated a metal-base transistor with a magnetic multilayer at the base. Magnetic multilayers are composed of ferromagnetic (F) and non-ferromagnetic (N) layers, with individual layer thicknesses of a few nanometers, and show electrical resistance depending on the applied magnetic field, a phenomenon known as giant magnetoresistance (GMR) (Baibich et al. 1988). Figure 13.8a shows the usual band structure of a metal-base transistor with the magnetic multilayer (MM) represented at the base. The difference is that in this case, the base-transport factor will also depend on the applied magnetic field, i.e., $\alpha_B = \alpha_B(H)$. For a ferromagnetic configuration of the multilayer, represented by (↑↑) and indicating the same magnetization direction (parallel) for the ferromagnetic layers, the magnetoresistance (MR) is low, and for anti-ferromagnetic, with the non-ferromagnetic layers aligned anti-parallel (↑↓), the MR is high. These two MR levels correspond to distinct α_B, $\alpha_B(\uparrow\uparrow) < \alpha_B(\uparrow\uparrow)$, since lower values of MR correspond to less scattering of the electrons when crossing the base, and consequently higher values of collector current. Figure 13.8b shows a common-base characteristic curve of a spin transistor for both configurations of the MM, with the excess current for the ferromagnetic configuration named magnetocurrent, quantified as the difference $\Delta J = J(\uparrow\uparrow) - J(\uparrow\downarrow)$. The magnetocurrent effect is calculated as $MC(\%) = 100\ [(J(\uparrow\uparrow) - J(\uparrow\downarrow)/J(\uparrow\downarrow))]$. The device demonstrated by Monsma et al. (1995) had initially a Co/Cu/Co multilayer at the base, with Co as ferromagnetic and Cu as non-ferromagnetic materials, and posteriorly a Pt layer was added to the emitter interface to tailor the barrier height relative to the collector interface (Monsma et al. 1998).

The magnetoresistance phenomenon is an effect observed at nanometer dimensions. It was first described in metallic nanolayers (Baibich et al. 1988) and subsequently in granular systems (Berkowitz et al. 1992, Xiao et al. 1992). It can be understood by assuming the model proposed by Mott (1964) that considers two channels for the conduction of electrons in ferromagnetic metals, one for spin up and one for spin down. The electrons transversally crossing a magnetic multilayer are assumed to suffer scattering in the layer with a magnetization contrary to their spin orientation. At low magnetic fields, when the magnetic multilayer is supposed to be in the anti-ferromagnetic alignment, the electron of both spin orientations will suffer scattering in alternate layers, and a higher MR will be measured. For high

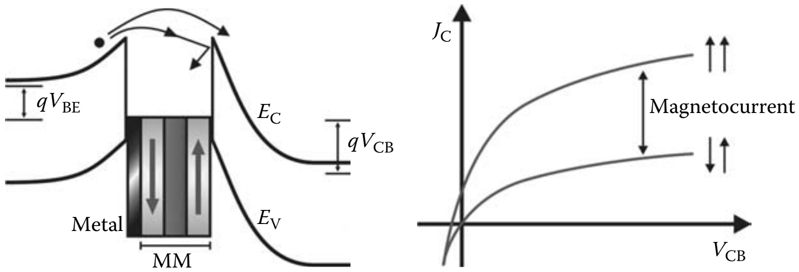

(a) (b)

FIGURE 13.8 (a) Band diagram of an MBT with an F/N/F magnetic multilayer (MM) in the base. A metal was added at the emitter interface to tailor the barrier height in the emitter relative to the collector. The arrows indicate the direction of the magnetization of the layers. (b) Common-base characteristic of the spin transistor collector current for both configurations of the multilayer, ferromagnetic (↑↑), and antiferromagnetic (↑↓). (After Monsma, D.J. et al., *Phys. Rev. Lett.*, 74, 5260, 1995; Monsma, D.J. et al., *Science*, 281, 407, 1998.)

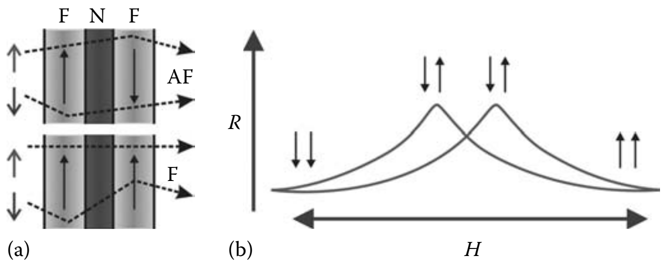

FIGURE 13.9 (a) Model of two channels and antiferromagnetic $(\uparrow\downarrow)$ and ferromagnetic $(\uparrow\uparrow)$ ordering in magnetic multilayers. (b) Typical curve for the magnetoresistance versus the applied magnetic field in magnetic multilayers.

magnetic fields, when the MM is ferromagnetic ordered, the electrons in the channel with the same orientation of the magnetization of the F layers will not suffer scattering by crossing the structure, and the electrons in the other channel will be scattered at each F layer. Since the channels are parallel, the total resistance will be reduced by the existence of the channel aligned with the magnetization of the ferromagnetic ordering of the layers. The alignment of the MM will have lower resistance, and the difference between both magnetic ordering is the magnetoresistive effect. Similarly to the magnetocurrent effect defined above, the MR effect can be written as $\mathrm{MR}(\%) = 100\,[(I(\uparrow\downarrow) - I(\uparrow\uparrow)/I(\uparrow\uparrow))]$. Figure 13.9a illustrates the scattering of electrons through the multilayered structure for both magnetic alignments, and in Figure 13.9b, the corresponding magnetoresistance curve when the field is varied from negative to positive values is shown.

The main disadvantages of the spin transistor proposed by Momsma et al. (1995) were related to the fabrication process that requires a vacuum bonding in order to sandwich the ferromagnetic multilayer between two-device quality silicon substrates, and the low transport factor observed ($<10^{-4}$). However, an easy method of fabricating MBTs with high current gains remained elusive.

A significant progress toward the fabrication process of TBMs was achieved by Meruvia et al. (2004) by using an organic semiconductor as the top layer in the vertical structure shown in Figure 13.5. A current gain as high as 1 was observed in the MBT with fullerene (C_{60}) as the emitter, Au thin layers as the base, and n-type silicon as the collector. The device is easy to fabricate since the organic layers can be prepared by regular methods such as electrodeposition, spin coating, and evaporation, and the approach opened the field of MBTs to hybrid structures with conducting polymers as the emitter and also as a pseudometallic base (Meruvia et al. 2005). The energy-band diagram of organic semiconductors is analogous to the one shown in Figure 13.4 for inorganic MBTs with barriers at interfaces and for the injection of hot carriers.

Most of the work expended on metal nanolayer-base transistors was in n-type devices. A p-type MBT was achieved by electrodepositing cuprous oxide (Cu_2O) (Delatorre et al. 2006). This semiconducting oxide is a p-type material with a 2.1 eV band-gap, a factor 2 higher than the band-gap of silicon of 1.1 eV. Figure 13.10 shows characteristic common-base curves obtained with the Cu_2O/Co/Si metal-base transistors with Co nanolayers of (a) 29 nm and (b) 33 nm, which display the transistor action in the third quadrant of the graphic as expected for p-metal-p devices.

The differences observed in Figure 13.10 by comparing the family of curves for the two base thicknesses used, 29 and 33 nm, indicate that the current gain is not only dependent on the thickness of the metallic layer but also that different transport mechanism are acting. For the 29 nm case, the collector current saturates at levels that are proportional to the emitter current, corresponding to a constant current gain. However, for 33 nm the I_C plateaus are not linear with I_E. When the test suggested in reference Rosencher et al. (1986) and described above to detect the presence of pinholes was applied to this p-metal-p transistor, a dependence of V_{EB} on V_{CB} was found. As can be seen in Figure 13.11a, for the thinner base layer, a significant dependence of V_{EB} on V_{CB} is observed. This dependency is a signature of the presence of pinholes in the base, which were also confirmed by the atomic force microscopy image shown in Figure 13.11b. The dark regions correspond to areas where no electrodeposition occurred, acting then as pinholes, allowing a direct contact of

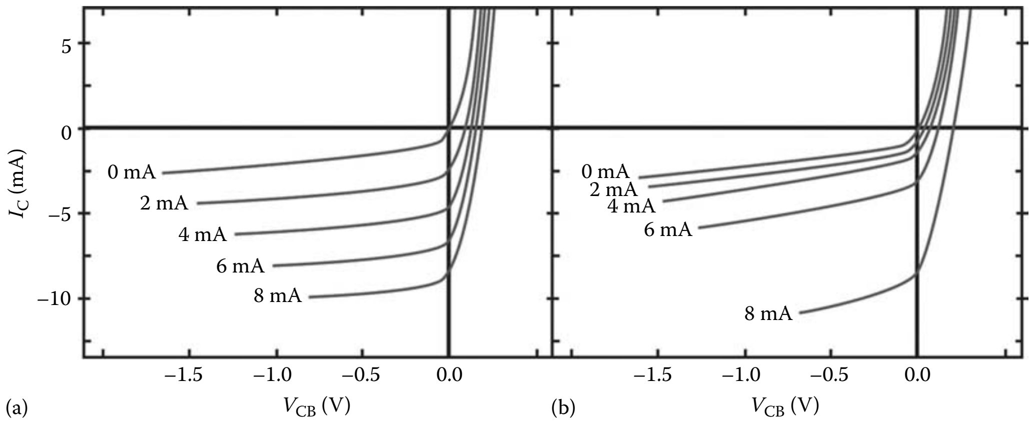

FIGURE 13.10 Common-base characteristics of a p-type metal transistor with Cu_2O as emitter, p-type silicon as collector, and Co with thicknesses of (a) 29 nm and (b) 33 nm as base. (After Delatorre, R.G. et al., *Appl. Phys. Lett.*, 88, 233504-1, 2006. With permission.)

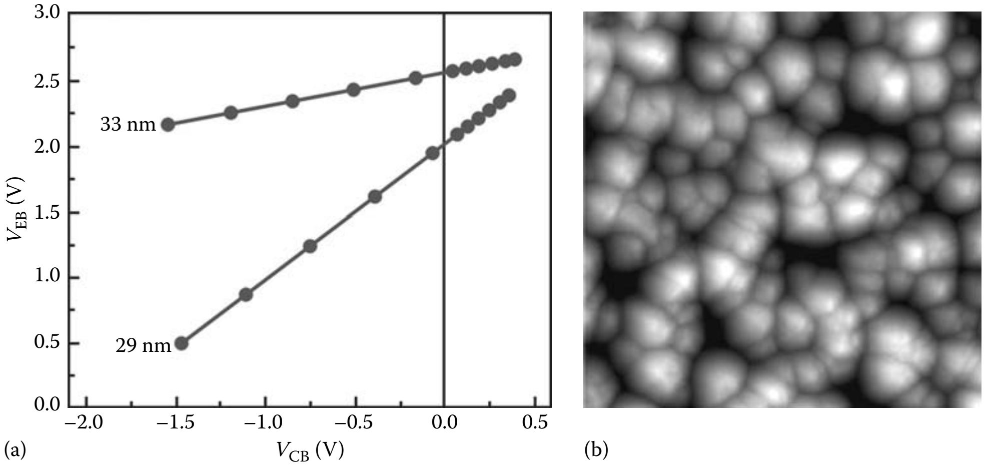

FIGURE 13.11 (a) Test of the dependency of V_{EB} on V_{CB} to check for the existence of pinholes in the base. (b) Atomic force microscopy of the Co metallic base with dark areas corresponding to regions not covered by the metal. (After Delatorre, R.G. et al., *Appl. Phys. Lett.*, 88, 233504-1, 2006. With permission.)

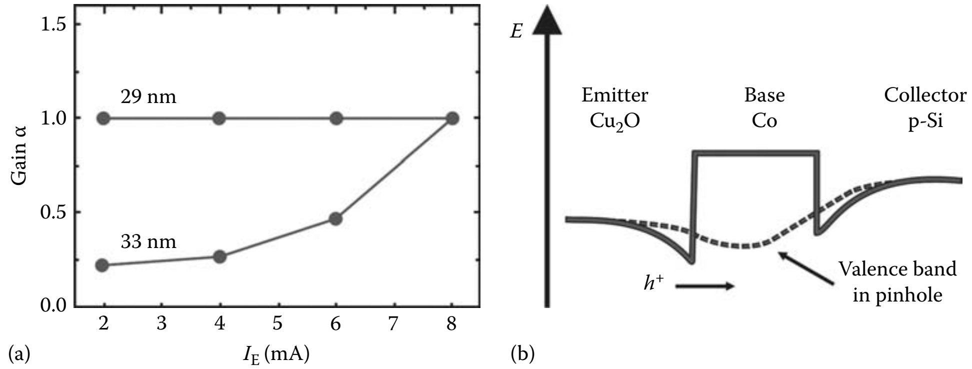

FIGURE 13.12 (a) Dependency of the current gain in the common-base configuration for two different Co thicknesses (29 and 33 nm). (b) Band diagram for the p-type metal base transistor with the two Schottky barriers and the expected barrier (dotted line) in the pinholes where the emitter and the collector are in contact, as expected for a permeable-base transistor. (After Delatorre, R.G. et al., *Appl. Phys. Lett.*, 88, 233504-1, 2006. With permission.)

the emitter with the collector. The bright areas are the Co grains. By increasing the base thickness, a considerable reduction in the pinholes occurred, since a lower dependence of V_{EB} on V_{CB} is observed for 33 nm in Figure 13.11a.

Figure 13.12a shows the dependency of the gain as a function of the emitter current for the two thicknesses of Co. The gain is equal to 1 and is independent of the emitter current for the case where the transistor has a higher density of pinholes. For a lower density of pinholes, the dependency of the current gain with the emitter current indicates that the gain increases with the energy of the ejected electron, i.e., with an increase of the emitter current, which is compatible with the view of electrons with higher energies being collected more easily at the collector interface. Figure 13.12b depicts the expected energy barriers for the p-type metal-base transistor in the case of two Schottky interfaces in the Cu_2O/Co/Si structure and the gradual variation of the barrier in the pinholes where the Cu_2O is in contact with p-Si. With an increasing base thicknesses, the MBT action dominates the PMT one, and on increasing the emitter current, higher gains should be observed since the energy of the electrons is higher, favoring the overcome of the collector to the base barrier.

References

Atalla, M. M. and Kahng, D. 1962. A new hot electron triode structure with semiconductor metal emitter. *IEEE Transactions on Electron Devices* ED-9: 507–508.

Baibich, M. N., Broto, J. M., Fert, A., Nguyen van Dau, F., Petroff, F., Eitenne, P., Creuset, G., Friederich, A., and Chazelas, J. 1988. Giant magnetoresistance of (001)Fe/(001)Cr magnetic superlattices. *Physical Review Letters* 61: 2472–2475.

Berkowitz, A. E., Mitchell, J. R., Carey, M. J., Young, A. P., Zhang, S., Spada, F. E., Parker, F. T., Hutten, A., and Thomas, G. 1992. Giant magnetoresistance in heterogeneous Cu-Co alloys. *Physical Review Letters* 68: 3745–3748.

Bozler, C. O. and Alley, G. D. 1980. Fabrication and numerical simulation of the permeable base transistor. *IEEE Transactions on Electron Devices* ED-27: 1128–1141.

Delatorre, R. G., Munford, M. L., Zandonay, R., Zoldan, V. C., Pasa, A. A., Schwarzacher, W., Meruvia, M. S., and Hümmelgen, I. A. 2006. p-Type metal-base transistor. *Applied Physics Letters* 88: 233504-1–233504-3.

Deneuville, A. and Brodsky, M. H. 1978. Thin-film metal base transistor structure with amorphous silicon. *Thin Solid Films* 55: 137–141.

Geppert, D. V. 1962. The metal base transistor. *Proceedings of the IRE* 50: 1527.

Hensel, J. C., Levi, A. F. J., Tung, R. T., and Gibson, J. M. 1985. Transistor action in Si/$CoSi_2$/Si heterostructures. *Applied Physics Letters* 47: 151–153.

Kanô, G. 1964. Schottky emitter type metal base transistor. *Japanese Journal of Applied Physics* 3: 363–364.

Meruvia, M. S., Hummelgen, I. A., Sartorelli, M. L., Pasa, A. A., and Schwarzacher, W. 2004. Organic-metal-semiconductor transistor with high gain. *Applied Physics Letters* 84: 3978–3980.

Meruvia, M. S., Benvenho, A. R. V., Hümmelgen, I. A., Pasa, A. A., and Schwarzacher, W. 2005. Pseudo-metal-base transistor with high gain. *Applied Physics Letters* 86: 263504-1–263504-3.

Monsma, D. J., Lodder, J. C., Popma, T. J. A., and Dieny, B. 1995. Perpendicular hot electron spin-valve effect in a new magnetic field sensor: The spin-valve transistor. *Physical Review Letters* 74: 5260–5263.

Monsma, D. J., Vlutters, R., and Lodder, J. C. 1998. Room temperature operating spin-valve transistors formed by vacuum bonding. *Science* 281: 407–409.

Mott, N. F. 1938. Note on the contact between a metal and an insulator on semiconductor. *Proceedings of the Cambridge Philosophical Society* 34: 568–572.

Mott, N. F. 1964. Electrons in transition metals. *Advances in Physics* 13: 325.

Rhoderick, E. H. and Williams, R. H. 1988. *Metal-Semiconductor Contacts*, 2nd edn. Oxford, U.K.: Clarendon.

Rose, A. 1960 (June). RCA Interim report 6A.

Rosencher, E., Delage, S., Campidelli, Y., and Arnaud D'Avitaya, F. 1984. Transistor effect in monolithic Si/$CoSi_2$/Si epitaxial structures. *Electronics Letters* 20: 762–764.

Rosencher, E., Badoz, P. A., Pfister, J. C., Arnaud d'Avitaya, F., Vincent, G., and Delage, S. 1986. Study of ballistic transport in Si-$CoSi_2$-Si metal base transistors. *Applied Physics Letters* 49: 271–273.

Schottky, W. 1938. Halbleitertheorie der sperrschicht. *Naturwissenschaften* 26: 843.

Sze, S. M. 1969. *Physics of Semiconductor Devices*. New York: Wiley.

Tung, R. T., Levi, A. F. J., and Gibson, J. M. 1986. Control of a natural permeable $CoSi_2$ base transistor. *Applied Physics Letters* 48: 635–637.

Xiao, J. Q., Jiang, J. S., and Chien, C. L. 1992. Giant magnetoresistance in nonmultilayer magnetic systems. *Physical Review Letters* 68: 3749–3752.

Zandonay, R., Delatorre, R. G., and Pasa, A. A. 2008. Electrical characterization of electrodeposited Co/p-Si Schottky diodes. *ECS Transactions* 14: 359–363.

14

ZnO Nanowire Field-Effect Transistors

Woong-Ki Hong
Gwangju Institute of Science and Technology

Gunho Jo
Gwangju Institute of Science and Technology

Sunghoon Song
Gwangju Institute of Science and Technology

Jongsun Maeng
Gwangju Institute of Science and Technology

Takhee Lee
Gwangju Institute of Science and Technology

14.1 Introduction **14**-1
14.2 Background.......... **14**-2
Basic Physical Parameters of ZnO • General Growth Methods of ZnO Nanowires • Field-Effect Transistor
14.3 Results and Discussion **14**-4
Growth by a Vapor Transport Method and Characterization of ZnO Nanowires • Fabrication and Characterization of ZnO Nanowire Field-Effect Transistors • N-Channel Depletion-Mode and Enhancement-Mode ZnO Nanowire FETs • Electronic Transport in Surface-Architecture-Controlled ZnO Nanowire FETs
14.4 Summary.......... **14**-14
Acknowledgment.......... **14**-14
References.......... **14**-14

14.1 Introduction

The top-down fabrication techniques, which involve photolithographic and subsequent etching techniques, have been used as conventional methods for device scaling in semiconductor industry for integrated electronic systems. However, due to the further downscaling of device size, the top-down approach faces the limitations of photolithographic/etching techniques as well as the increasing costs associated with lithography equipment (Lu and Lieber 2007). One promising method to overcome the limitations with the top-down methods is the bottom-up approach, which is a unique technique for fabricating functional devices, without the use of lithography techniques, through the chemical synthesis, self-assembly, and the manipulation of nanoscale elements such as small molecules, nanoparticles, nanotubes, and nanowires (Agarwal 2008). Recently, for the realization of integrated electronic systems through the bottom-up approach, these nanomaterials have attracted considerable interest due to their unique properties such as large surface-to-volume ratio, carrier and photon confinement, and high sensitivity. In particular, various semiconductor nanowires offer great potential as building blocks for nanoscale electronic and photonic device applications.

Among many semiconductor nanowires, ZnO with the preferentially hexagonal wurzite-type structure has a wide range of properties, from metallic to insulating conductivities (including n- and p-type conductivities), a direct wide band gap (~3.4 eV), large exciton-binding energy (~60 meV), radiation hardness, high transparency, piezoelectricity, room-temperature ferromagnetism, and chemical-sensing effects (Fan and Lu 2005, Klingshirn 2007, Schmidt-Mende and MacManus-Driscoll 2007). In particular, the advantage of large exciton-binding energy and a variety of nanostructures with relative cost-effectiveness provides excellent optical and electrical properties as a promising material for blue and ultraviolet optical devices, compared with GaN and GaN-based materials. Moreover, since ZnO has stronger radiation hardness than other common semiconductor materials such as Si, GaAs, CdS, and GaN, ZnO materials are promising for space applications (Look et al. 1999). In recent decades, due to these remarkable properties of ZnO materials and the demand for further miniaturization by semiconductor technology, tremendous efforts have been devoted to the growth and characterization of one-dimensional semiconductor ZnO nanostructures (or ZnO nanowires) as well as their versatile applications as a potential material for future nanoelectronics. For example, field-effect transistors (FETs) (Goldberger et al. 2005), sensors (Dorfman et al. 2006), light-emitting devices (Bao et al. 2006), solar cells (Law et al. 2005), as well as logic circuits (Park et al. 2005, Yeom et al. 2008) have been extensively investigated and demonstrated.

In particular, since nanowire-based FETs are the fundamental element for nanoelectronics among these versatile applications, ZnO nanowire-based FETs have been fabricated and intensively investigated in a typical back-gate configuration due to fabrication simplicity (Goldberger et al. 2005, Cha et al. 2006, Chang et al. 2006). The early studies of ZnO nanowire FETs have focused on their device performance (Cha et al. 2006, Chang et al. 2006), gas-sensing application (Fan and Lu 2006, Hsueh et al. 2007), photodetection (Kind et al. 2002, Fan et al. 2004a,b, Soci et al. 2007), and chemisorption/photodesorption (Li et al. 2005, Suehiro et al. 2006). However, in order to fabricate the functional devices from the ZnO nanowires, their fundamental properties must be carefully investigated because the presence of a large surface and interfacial area in nanowires can profoundly alter their performance (Hong et al. 2007, Xiong et al. 2007). Recently, there are a few studies on the role of geometric properties, surface states, and passivation on the transport properties of ZnO nanowire FETs, specifically on the influence of the presence of surface trap states at the interfaces associated with nanowire size and surface roughness (Hong et al. 2007, 2008a,b,c). The interfaces and interfacial states related to interface roughness play an important role not only in nanowire heterostructures, but also in the electronic transport properties of nanoscale transistors as fundamental elements for nanoelectronics (Hastas et al. 2002, Wang et al. 2005, Jung et al. 2007, Mieszawska et al. 2007). In addition, it is well known that metal-oxide semiconductors including ZnO nanostructures are strongly affected by the chemical adsorptions of ambient gases (Lagowski et al. 1977, Kolmakov and Moskovits 2004). Therefore, it is essential to understand the surface-roughness- and size-dependent effects on the electronic transport properties of nanowire transistors. Moreover, since ZnO nanowires are strongly influenced by chemical environments due to large surface-to-volume ratio (Goldberger et al. 2005, Li et al. 2005, Fan and Lu 2006, Suehiro et al. 2006, Hsueh et al. 2007), it is important to study the passivation effects of dielectric layer on the transport properties correlated to geometry and adsorbed species on the ZnO nanowires.

In this chapter, we review the basic electrical properties of FETs made from ZnO nanowires that were grown by a vapor transport method. We demonstrate that the transport properties of ZnO nanowire FETs are associated with the surface roughness, nanowire size, surface states, and/or defects, as well as the surface chemistry in surrounding environments. These explorations will give insights not only in understanding the transport properties but also in developing practically useful applications of nanowire-based devices.

14.2 Background

14.2.1 Basic Physical Parameters of ZnO

The ZnO in wurzite structure, which can be considered as a hexagonal close-packed lattice of zinc atoms tetrahedrally coordinated with four oxygen atoms, has the primitive translation vectors $\mathbf{a_0}$, $\mathbf{b_0}$, and $\mathbf{c_0}$ in which $\mathbf{a_0}$ and $\mathbf{b_0}$ lay in the x–y plane and are of equal length, whereas $\mathbf{c_0}$ is parallel to the z-axis (Figure 14.1). The values of the primitive translation vectors are $a_0 = b_0 = 0.3249$ nm and $c_0 = 0.5206$ nm at room temperature. The ratio c_0/a_0 of 1.602 deviates slightly from the value $c_0/a_0 = 1.633$ of the ideal hexagonal structure. The basic physical parameters for ZnO are summarized in Table 14.1 (Fan and Lu 2005, Klingshirn 2007, Schmidt-Mende and MacManus-Driscoll 2007). Based on these fundamental physical properties, intense research by many groups has been focused on the growth and characterization of ZnO nanostructures and their versatile device applications.

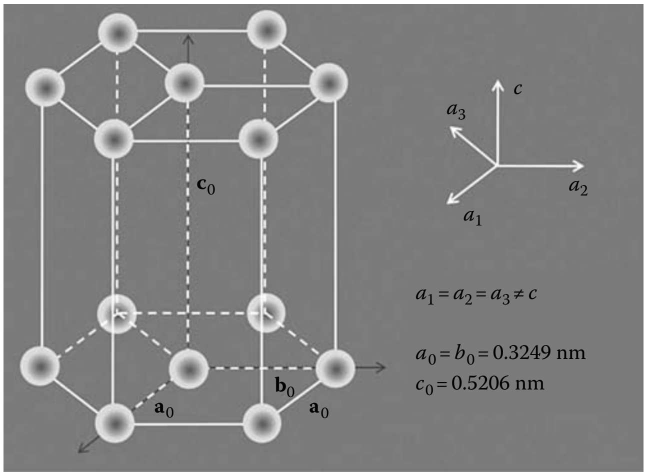

FIGURE 14.1 Crystal structure of ZnO.

TABLE 14.1 Basic Physical Parameters of ZnO

Property	Value
Lattice constants (at $T = 300$ K)	
a_0	0.3249 nm
c_0	0.5206 nm
a_0/c_0	1.602 (ideal structure: 1.633)
Density (ρ)	5.606 g/cm^3
Melting point	1975°C (2248 K)
Dielectric constant (ε)	8.66
Energy gap (E_g)	3.4 eV
Exciton-binding energy	60 meV
Electron effective mass (m_e^*)	0.24
Electron mobility (μ_e)	200 cm^2/V s
Hole effective mass (m_h^*)	0.59
Hole mobility (μ_h) (at $T = 300$ K)	5–50 cm^2/V s

In the following we explain briefly the growth by a vapor transport method and the characterization of ZnO nanowires and the detail fundamental properties of the FETs made from ZnO nanowires are explained in Section 14.3.

14.2.2 General Growth Methods of ZnO Nanowires

Semiconductor nanowires including ZnO nanowires have attracted considerable attention due to their unique properties, including their high anisotropic geometry and large surface-to-volume ratio. In order to utilize semiconductor nanowires for nanoelectronic devices, it is essential to rationally grow high-quality nanowires with tunable and modulated chemical composition, size, and morphology, and to precisely control doping concentration with both n- and p-type dopants. To this end, there are many efforts to control the nanowire properties, and numerous growth techniques have been developed (Dai et al. 2003, Law et al. 2004, Schmidt-Mende and MacManus-Driscoll 2007). In recent years, the interest in ZnO nanostructures has increased drastically and many researchers have intensively investigated on them. In general, ZnO nanostructures can be grown by either vapor-phase growth or aqueous solution growth methods. In particular, vapor-phase growth methods use gas-phase species as the initial starting reactants for the nanowire formation, and numerous growth techniques such as laser-assisted growth (Sun et al. 2006), chemical vapor deposition (Yang et al. 2002, Park et al. 2003), and thermal evaporation (Dai et al. 2003) have been developed to prepare gas-phase species. In these gas-phase production methods, nanowires are commonly grown via vapor–liquid–solid (VLS) or vapor–solid (VS) mechanisms. VLS process can be regarded as nanowire nucleation and growth via Zn accumulation in liquid catalyst particle (often Au), supersaturation, precipitation, and oxidation, whereas VS process can be described as one-dimensional nanostructure formation from vapor phase precursors in the absence of metal catalysts (Law et al. 2004).

14.2.3 Field-Effect Transistor

The FET, which is often called a unipolar transistor due to a majority carrier device, uses either electrons or holes for conduction. The FET is a three-terminal (source, drain, and gate) device in which the current through two terminals between source and drain is controlled by the gate (Sze 1981, Streetman and Banerjee 2000). The conductivity is varied by the electric field that is produced when the gate–source voltage (V_G) is increased; the drain–source current (I_{DS}) increases exponentially for V_G below threshold. The FETs are divided into several forms: a junction FET (JFET), a metal-semiconductor FET (MESFET), and a metal-insulator semiconductor FET (MISFET) or metal-oxide semiconductor FET (MOSFET) (Figure 14.2) (Sze 1981, Streetman and Banerjee 2000). In the JFET, the control (gate) voltage varies the depletion width of a reverse biased p–n junction with the channel between the source and drain (Figure 14.2a). Similarly, the MESFET is the JFET in which the reverse biased p–n junction is replaced by a metal-semiconductor Schottky junction (Figure 14.2b) The MOSFET is used to amplify or switch electronic signals and generally is consisting of three terminals (source, drain, and gate) and substrate (or body) (Figure 14.2c through f). In particular, since a gate electrode is separated from the semiconductor (substrate) by an oxide, there is effectively no dc gate current, and the channel is capacitively coupled to the gate via the electric field in the oxide. Furthermore, the FETs are

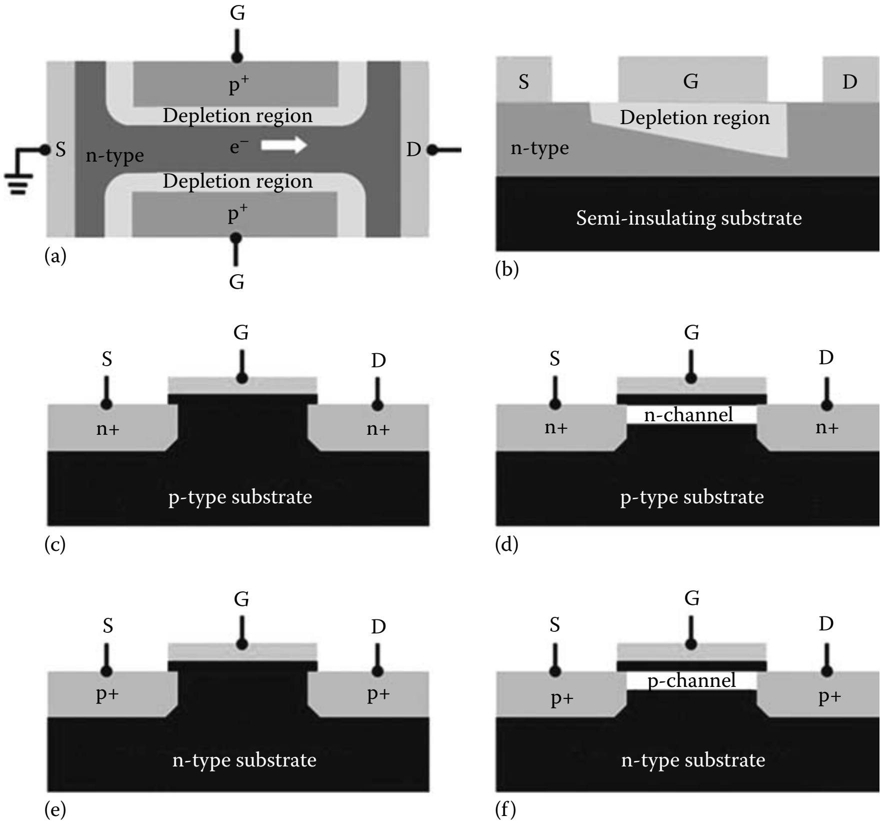

FIGURE 14.2 Basic classification of FETs. (a) JFET, (b) MESFET, and (c)–(f) four different types of MOSFET. (c) n-channel E-mode MOSFET and (d) n-channel D-mode MOSFET. (e) p-channel E-mode MOSFET and (f) p-channel D-mode MOSFET.

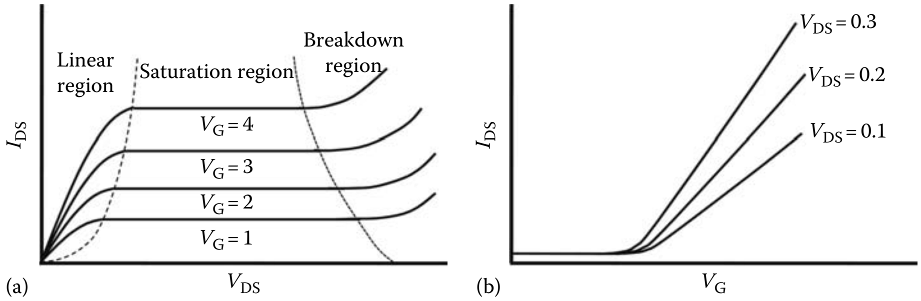

FIGURE 14.3 Basic current–voltage characteristics of FETs. (a) Typical transistor output characteristics (I_{DS}–V_{DS}). (b) Typical transistor transfer characteristics (I_{DS}–V_G).

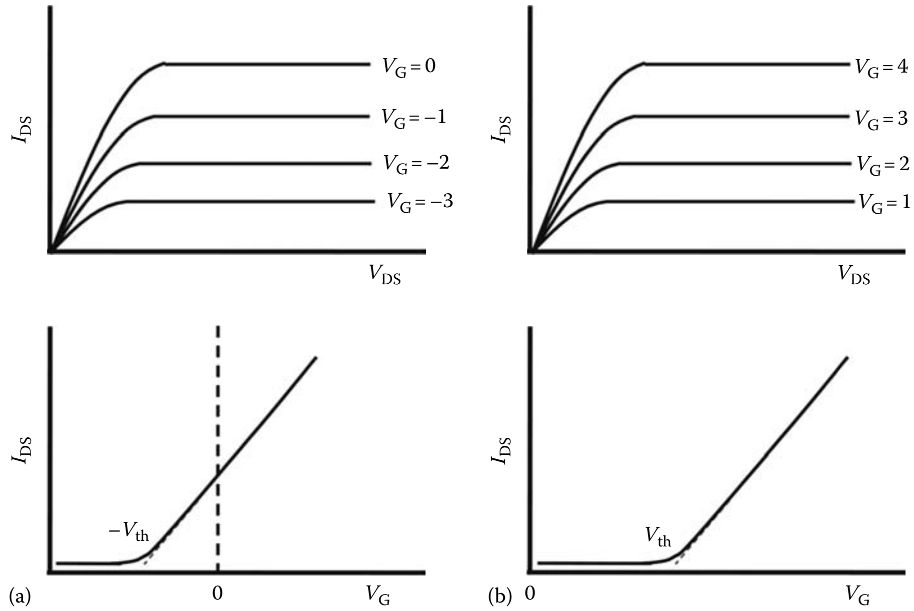

FIGURE 14.4 Comparison of basic current–voltage characteristics of (a) "normally on" FET and (b) "normally off" FET.

distinguished as depletion-mode (D-mode) and enhancement-mode (E-mode) types in terms of the turn-on/off position and the polarity of the threshold voltage. The D-mode FETs exhibit the turn-on of the channel at zero gate bias and negative threshold voltages (Figure 14.2d and f), whereas the E-mode FETs exhibit the turn-off at zero gate bias and positive threshold voltages (Figure 14.2c and e). Therefore, D- and E-mode types are often called "normally on" and "normally off," respectively.

The basic current–voltage (*I*–*V*) characteristics of FETs are shown in Figure 14.3, where the complete set of I_{DS}–V_{DS} (drain–source current versus voltage) curves are called "output characteristics" (Figure 14.3a) and I_{DS}–V_G (drain–source current versus gate voltage) curves are called "transfer characteristics" (Figure 14.3b). In particular, the output characteristics are plotted as a function of the drain–source voltage (V_{DS}) for various gate voltages (V_G) and can be divided into three regions: (1) the linear region, where I_{DS} increase linearly with V_{DS} for a given $V_G(>V_{th})$; (2) the saturation region, where I_{DS} no longer increases as V_{DS} increases, i.e., it remains essentially constant; and (3) the breakdown region, where I_{DS} suddenly increases with the further increase of V_{DS} beyond saturation. Figure 14.4 displays the comparison of *I*–*V* characteristics between "normally on" (D-mode) FET (Figure 4.4a) and "normally off" (E-mode) FET (Figure 14.4b). The main difference is the position of the threshold voltage (V_{th}) along the V_G.

14.3 Results and Discussion

14.3.1 Growth by a Vapor Transport Method and Characterization of ZnO Nanowires

ZnO nanostructures can be simply grown using a thermal furnace system consisting of quartz tube, gas flow controller, and vacuum system. Figure 14.5 shows a schematic of furnace system and a series of the field emission scanning electron microscopy (FESEM) images of various ZnO nanostructures with distinctive geometrical morphologies formed by the vapor transport method using a furnace system, which is the most common method to grow ZnO nanowires. The shape and size of these nanostructures are strongly dependent on the various growth

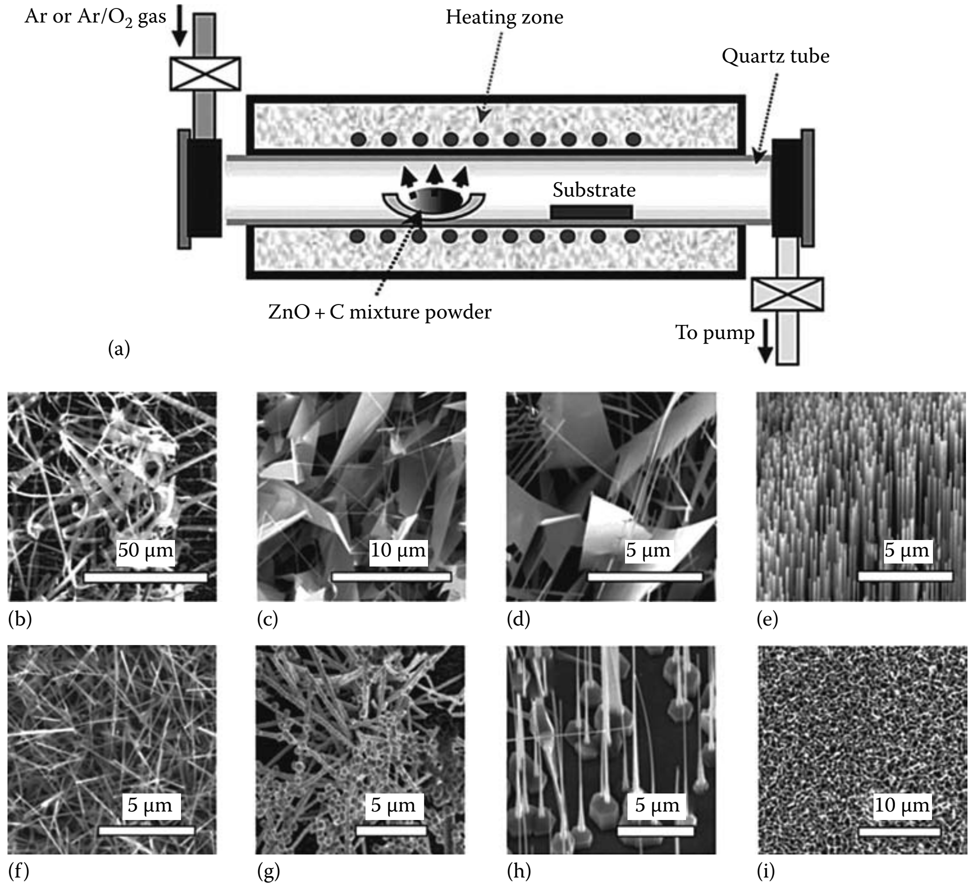

FIGURE 14.5 (a) Schematic diagram of experimental apparatus for growth of ZnO nanostructures. FESEM images of (b) nanoribbons or nanoblets, (c) nanosheets, (d) nanocombs, (e) vertically well-aligned nanowire/nanorods, (f) random-oriented nanowire/nanorods, (g) corrugated nanowires, (h) nanotips or nanoneedles, and (i) nanowalls.

conditions such as temperature, gas flow rate, substrate itself, and the position of substrate. In general, the size, surface morphology, and the crystal structure of ZnO nanowires are commonly characterized using FESEM and transmission electron microscopy (TEM). In order to study the optical properties of ZnO nanowires, the photoluminescence (PL) spectra are measured. In addition, to investigate the electrical properties of ZnO nanowires, the FET-based device structures are fabricated as shown in Figure 14.6e and f and characterized.

The vertically well-aligned ZnO nanowires were grown on the sapphire substrates or ZnO-film-coated sapphire substrates (Figure 14.6a). Figure 14.6b shows the micro-photoluminescence (μPL) spectra of ZnO nanowires grown on the two different substrates. A He–Cd laser (325 nm) was used as an excitation source in the PL measurements. The μPL spectra were measured on the ZnO nanowires after they were transferred from the growth substrates to a silicon wafer in order to eliminate the signals coming from the ZnO film substrate itself. The ZnO nanowires grown on an Au-catalyst-free ZnO film substrate show a stronger defect emission in the PL spectrum than those grown on an Au-coated sapphire substrate. Since it is generally agreed that the defect emission is a surface-related process (Wang et al. 2006), the PL spectra suggest that the ZnO nanowires grown on an Au-catalyst-free ZnO film have a significantly larger number of surface defect sites. Figure 14.6c and d shows typical high-resolution TEM (HRTEM) images of ZnO nanowires grown on an Au-coated sapphire substrate and on an Au-catalyst-free ZnO film, respectively. The upper insets of Figure 14.6c and d are computed fast Fourier transform patterns obtained from the lattice fringes of ZnO nanowires, indicating that the ZnO nanowires grown on both substrates are single crystalline with a preferred growth direction of [0001]. The lower insets of Figure 14.6c and d are the low-magnification TEM images of ZnO nanowires showing the overall surface roughness along the nanowires. Both the low-magnification TEM images and the HRTEM images allow comparison between the surface structures of the ZnO nanowires grown on the two different types of substrates. Compared to the ZnO nanowires grown on an Au-coated sapphire substrate (Figure 14.6c), the ZnO nanowires grown on an Au-catalyst-free ZnO film (Figure 14.6d) are seen to be significantly rougher across the surfaces parallel to the growth direction, as is apparent in the HRTEM images, and as is indicated by the circled regions in the low-magnification TEM images of the lower inset in Figure 14.6d. Figure 14.6e and f show the schematic of nanowire FET device structure and the SEM image of a single ZnO nanowire connecting the source and drain electrodes.

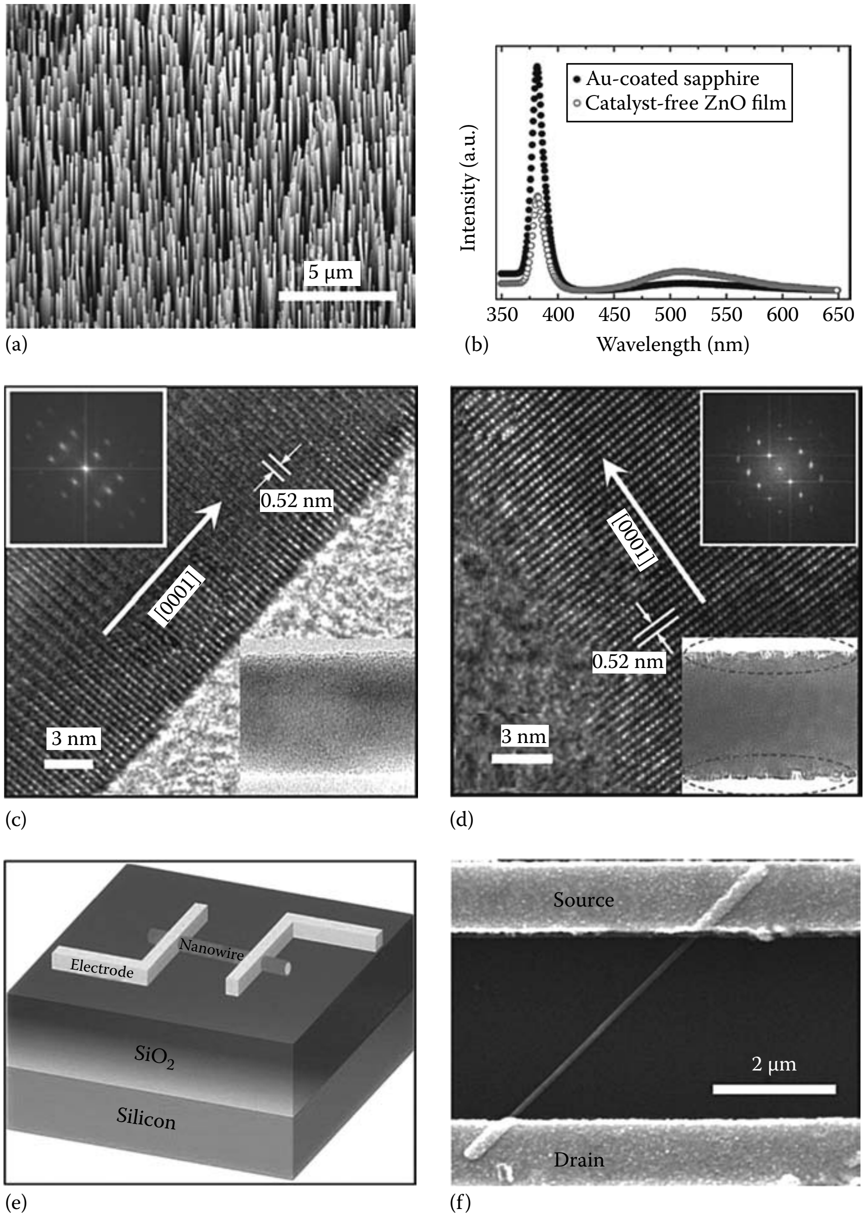

FIGURE 14.6 (a) FESEM images of vertically well-aligned ZnO nanowires grown on Au-coated sapphire substrates or ZnO buffer film-coated substrates. (b) The μPL spectra of ZnO nanowires on the two substrates showing emissions at approximately 378 nm. HRTEM images of ZnO NWs grown on (c) an Au-coated sapphire substrate and (d) an Au-catalyst-free ZnO film. Insets are computed fast Fourier transform patterns (upper) and low-magnification TEM images (lower). (e) A schematic of the FET device structures. (f) A FESEM image of a single ZnO nanowire connected between source and drain electrodes in FET. (Adapted from Hong, W.-K. et al., *Appl. Phys. Lett.*, 90, 243103, 2007. With permission.)

14.3.2 Fabrication and Characterization of ZnO Nanowire Field-Effect Transistors

ZnO nanowire FETs with different gate structures have been fabricated and extensively investigated in the past several years. For example, there are four different types of gate configurations: back gate (Goldberger et al. 2005, Hong et al. 2007), top gate (Huang et al. 2001, Yeom et al. 2008), side gate (Cha et al. 2006), and surrounding gate (Ng et al. 2004, Wade et al. 2007). In particular, the FETs with back-gate and top-gate configurations have been intensively explored due to the fabrication simplicity. To fabricate ZnO nanowire FET device, the grown ZnO nanowires are first transferred from the growth substrate to a silicon wafer with thermally grown oxide or a bare silicon wafer by dropping and drying a nanowire suspension (Figure 14.7). The nanowire suspension is made by briefly sonicating the growth substrate of ZnO nanowires in ethanol or isopropyl alcohol (IPA) for tens of seconds. Metal electrodes forming Ohmic contacts between metal and ZnO nanowires are deposited by an evaporator or a sputter and defined as source and drain electrodes by

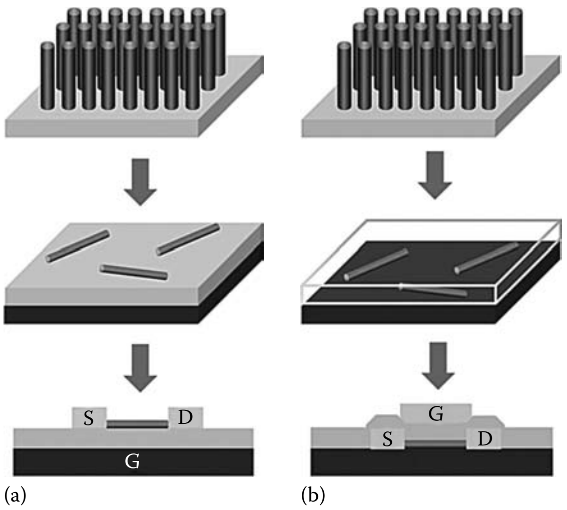

FIGURE 14.7 Fabrication process of ZnO nanowire FETs with (a) back gate and (b) top gate configurations.

lithography techniques and liftoff process (Figure 14.7). In back-gated FET configuration (Figure 14.7a), a pair of metal electrodes patterned on the ZnO nanowire channel serves as the source and drain and a degenerately doped silicon substrate serves the gate electrode. In top-gated FET configuration (Figure 14.7b), a metal gate electrode is formed above a gate dielectric layer on nanowires. Recently, tremendous efforts have been devoted to the fabrication and characterization of nanowire-based transistors. However, if functional FET devices with electrical contacts are to be made from nanowires, their fundamental properties must be carefully investigated due to the presence of the different interfaces and interfacial states in nanowire-based device. The different interfaces exist between the nanowire and passivation layer, nanowire and gate dielectrics, nanowire and electrodes, and nanowire and ambient gases.

In particular, since the semiconductor surfaces are strongly affected by the chemical adsorption of ambient gases (Wolkenstein 1991), the influence of chemisorptions of ambient gases in the case of semiconducting nanowires including ZnO nanowires are much more sensitive than their bulk counter part due to large surface-to-volume ratio (Lagowski et al. 1977, Kolmakov and Moskovits 2004). For example, surface defects such as oxygen vacancies act as the adsorption sites of O_2 molecules and form oxygen ions (O^-, O^{2-}, or O_2^-), and then the chemisorption of O_2 molecules depletes the surface electron and reduces the channel conductivity (Li et al. 2004, Zhang et al. 2004). Therefore, the passivation of ZnO nanowire FET devices is significantly important to remove the influence of water or gas molecules in ambient air on nanowire and to improve the FET performance by enhancing the gate-coupling effects (Hong et al. 2008a,b,c). In addition, other characterization conditions such as gate bias sweep rate can also influence the electronic transport in nanowire transistors due to the presence of surface states (Maeng et al. 2008).

In this section, we review the electrical properties of the surface-tailored smooth and rough ZnO nanowire FETs with and without surface passivation using poly(methyl metahacrylate) (PMMA), which is a polymeric resist commonly used in nanolithographic processes involving electron beam, deep UV, or x-ray radiation (Hong et al. 2008a,b,c). Furthermore, a detailed study of the passivation and gate sweep rate effects on the ZnO nanowire FETs under different oxygen environments is presented.

14.3.2.1 Passivation and Surface-Roughness Effects

The electrical characteristics before and after the passivation of the FET devices made from the surface-tailored smooth and rough ZnO nanowires are summarized in Figure 14.8. Figure 14.8a and b shows the output characteristics (source–drain current versus voltage, I_{DS}–V_{DS}) and transfer characteristics (source–drain current versus gate voltage, I_{DS}–V_G) for a FET made from smooth ZnO nanowires, respectively. The FETs made from smooth ZnO nanowires exhibit more well-defined saturation and pinch-off characteristics after passivation in comparison with the FET devices before passivation. In particular, the FET devices before passivation show decreasing separation in current between I_{DS} curves at larger currents in I_{DS}–V_{DS} curves (Figure 14.8a), which is attributed to either an electron injection barrier at the source electrode or to mobility degradation associated with the interface roughness scattering of channel electrons at the channel–insulator interface with increasing gate voltage (Dehuff et al. 2005). The I_{DS}–V_G curves in Figure 14.8b show the threshold voltages (V_{th}) of −4.16 and −2.25 V before and after passivation, respectively. The threshold voltage after passivation shifts toward the positive gate bias direction due to surface-depletion-induced channel-narrowing effect (Fan et al. 2004a,b, Ju et al. 2006, 2007). The current on/off ratios (I_{on}/I_{off}) both before and after passivation for this FET device of smooth ZnO nanowire exhibited 10^4–10^5 (insets of Figure 14.8b).

In contrast, Figure 14.8c and d shows the I_{DS}–V_{DS} and I_{DS}–V_G curves for a FET made from rough ZnO nanowires, respectively. The unpassivated FET devices exhibited the poor electrical characteristics, as the I_{DS}–V_{DS} curves shown in Figure 14.8c. The I_{DS} suddenly decreased with the further increase of V_{DS} and V_G beyond saturation and also the current between I_{DS} curves was not clearly separated. However, the passivated FET devices exhibited well-defined linear regions at low biases and saturation regions at high biases as typical transistors, indicating clear pinch-off behavior. Unlike the FET devices made from smooth ZnO nanowires, the threshold voltage for the FET devices made from rough ZnO nanowires exhibits a considerable shift toward the positive gate bias direction after passivation. The threshold voltage of the particular FET device shown in Figure 14.8d shifted from −12.5 to 8.9 V, indicating transition from D-mode to E-mode operations. Most of the passivated FET devices made from both smooth and rough ZnO nanowires exhibit superior electrical performance than the unpassivated FET devices made from rough ZnO nanowires. These results are in good agreement with the superior characteristics of the FETs passivated by SiO_2/Si_3N_4 or PMMA layer (Kim et al. 2003, Chang et al. 2006, Hong et al. 2008a,b,c). The current on/off ratios (I_{on}/I_{off}) both before

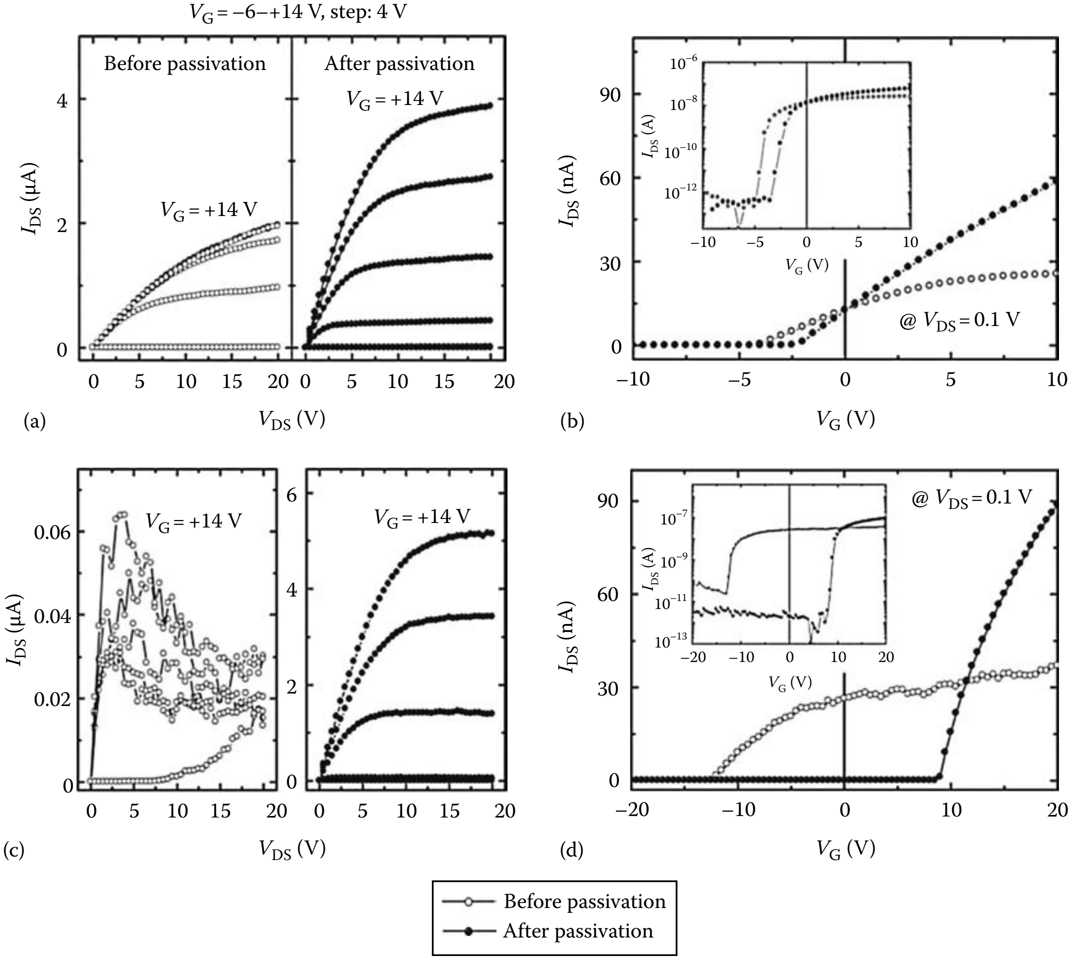

FIGURE 14.8 (a) I_{DS}–V_{DS} and (b) I_{DS}–V_G curves of a FET device made from rough ZnO nanowires before and after passivation. (c) I_{DS}–V_{DS} and (d) I_{DS}–V_G curves of a FET device made from rough ZnO nanowires before and after passivation. (Adapted from Hong, W.-K. et al., *Appl. Surf. Sci.*, 254, 7559, 2008c. With permission.)

and after passivation for the FET device of rough ZnO nanowire exhibited approximately 10^4–10^5 (insets of Figure 14.8d).

Furthermore, the electrical properties of the ZnO nanowire FETs have also been investigated and compared before and after the PMMA passivation under different environments of ambient air (20% O_2), dry O_2, and vacuum (10^{-3} Torr) environments (Song et al. 2008). Figure 14.9 shows a series of the I_{DS}–V_{DS} and the I_{DS}–V_G curves for a ZnO nanowire FET measured at different gate voltages under various oxygen environments before and after passivation, respectively. Figure 14.9a and b displays the I_{DS}–V_{DS} curves for the ZnO nanowire FET before and after passivation, respectively. As shown in Figure 14.9a, the current level is significantly affected by the environmental condition with O_2 molecules. The currents decrease significantly when the ZnO nanowire FET is exposed to dry O_2, as compared to the case in ambient air. When the oxygen is evacuated, the currents increase again. On the contrary, the I_{DS}–V_{DS} curves in Figure 14.9b show a negligible effect on the passivated ZnO nanowire FET by oxygen molecules in different oxygen environments. This is because the PMMA passivation layer prevents oxygen molecules from being adsorbed onto the nanowire. One can also notice that the ZnO nanowire FETs after passivation have a better electrical performance in terms of well-defined linear and saturation regions in I_{DS}–V_{DS} curves (Figure 14.9b). It has been reported that the passivation can improve the FET performance by enhancing the gate-coupling effect (Kim et al. 2006, Wunnicke 2006, Hong et al. 2008a,b,c). This passivation effect is more prominent in transfer characteristics, as shown in Figure 14.9c and d. A series of plots in Figure 14.9c and d shows the I_{DS}–V_G curves for the ZnO nanowire FET before and after passivation, respectively. The FET device was measured at a fixed source–drain voltage ($V_{DS} = 0.1$ V) under the same different environments of ambient air (20% O_2), dry O_2, and vacuum (~10^{-3} Torr). The I_{DS}–V_G plots in the semilogarithmic scale (insets of Figure 14.9c and d) show an on/off current ratio as large as 10^4–10^5 for both before and after passivation. Before passivation (Figure 14.9c), the threshold voltages of ZnO nanowire FET shifted from −1.73 (in ambient air) to 6.05 V (in dry O_2) in the positive gate bias direction. This is because more electrons are captured by the oxygen as the oxygen pressure is raised. Since the charge density is reduced after being exposed to dry O_2, more positive gate bias is needed to make currents flow in the nanowire channel, and thus the

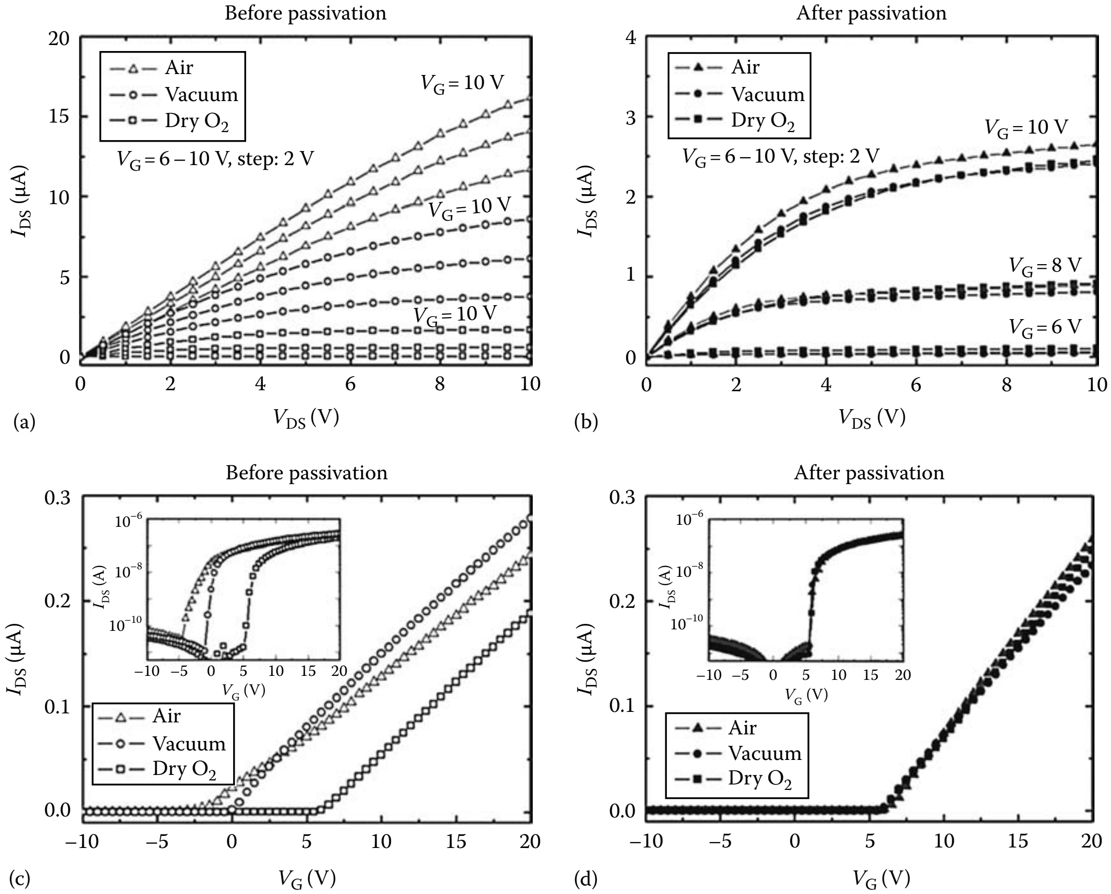

FIGURE 14.9 I_{DS}–V_{DS} curves for different gate biases (from 6 to 10 V, with a step of 2 V) for a ZnO nanowire FET, acquired under ambient air, dry O_2, and vacuum conditions (a) before and (b) after passivation. I_{DS}–V_G curves at V_{DS} = 0.1 V for the ZnO nanowire FET, acquired under ambient air, dry O_2, and vacuum conditions (c) before and (d) after passivation. The insets in (c and d) show the semilogarithmic plot of the I_{DS}–V_G curves. (Adapted from Song, S. et al., *Appl. Phys. Lett.*, 92, 263109, 2008. With permission.)

threshold voltage shifts to the positive gate bias direction. When oxygen gas was evacuated to a vacuum level of 10^{-3} Torr, the threshold voltages of the nanowire FET shifted back from 6.05 (in dry O_2) to −0.10 V (in a vacuum) in the negative gate bias direction, as compared when they were exposed to a dry O_2 environment. This is due to the reduction of the trapping effects of oxygen in a vacuum. On the contrary, when the FET was passivated with PMMA, the threshold voltages were ~6 V and shifted very little under different oxygen environments, as shown in Figure 14.9d. Note that the threshold voltage is defined as the gate voltage obtained by extrapolating the linear portion of the I_{DS}–V_G curve from the point of maximum slope to zero drain current, in which the point of maximum slope is the point where transconductance (dI_{DS}/dV_G) is at a maximum (Arora 1993).

As a result, these phenomena can be explained by oxygen effects. The ambient oxygen partial pressure has a considerable effect on the electrical properties of ZnO nanowire FETs (Fan et al. 2004a,b). Particularly, the adsorbed oxygen molecules deplete the electrons in the ZnO nanowire and form oxygen ions (O^-, O^{2-}, or O_2^-) (Li et al. 2004). Then, the electrons in the ZnO nanowires are trapped by the adsorbed oxygen molecules, and thus the surface depletion region of ZnO nanowire can be formed, leading to current reduction and threshold voltage shift to the positive gate bias direction (Figure 14.9c). As the oxygen pressure is raised, more electrons are captured by the oxygen molecules at the nanowire surface. Correspondingly, the depletion region is widened and the carrier density in the ZnO nanowire is decreased even more when the FET is exposed to dry O_2 environment.

14.3.2.2 Sweep Rate Effect of Gate Bias

As aforementioned, the corresponding characteristics of nanowire FETs are sensitively dependent on the surrounding environments to which the devices are exposed, and whether or not the devices are passivated (Park et al. 2004, Chang et al. 2006, Song et al. 2008). Other characterization conditions can also influence electrically the interactive absorption of oxygen molecules on the nanowire surface. Particularly, the effect of gate bias sweep rate on the electronic transport properties of ZnO nanowire FETs has been investigated (Maeng et al. 2008). The electrical properties of ZnO nanowire FETs were measured as a function of gate bias sweep rate in different measurement environments.

Figure 14.8 show a series of I_{DS}–V_G curves measured at different gate bias sweep rates (2500, 250, 130, 100, 12, 6, 1.2, 0.3, 0.2, 0.1 V/s) at $V_{DS} = 0.5$ V under various environments. The insets in Figure 14.10 show the I_{DS}–V_G plots on a semilogarithmic scale. A ZnO nanowire FET device was systematically measured in ambient air, in an N_2-filled glove box, and in ambient air after the same device was passivated with a PMMA layer. Figure 14.10a shows the case of a ZnO nanowire FET (unpassivated) measured in ambient air. As the gate bias sweep rate was decreased from 2500 to 0.1 V/s, the current decreased and the threshold voltage shifted to the positive gate bias direction. These phenomena can be explained by the depletion of electrons by oxygen adsorption on the surface of the ZnO nanowire. As previously mentioned, the increase of oxygen concentration causes the change of conductivity by surface depletion in the nanowire channel, i.e., the adsorbed oxygen molecules that bind the electrons localized at the nanowire surface become oxygen ions in the forms of O^-, O^{2-}, or O_2^-, resulting in depletion of electrons and thus lowering the conductivity (Takata et al. 1976, Li et al. 2004, Sadek et al. 2007). Oxygen adsorption is sustained at an equilibrium condition in ambient air at zero gate bias. However, the gate bias induces more adsorption of oxygen ions due to the applied positive gate bias. Therefore, when the nanowire FET is applied with a slower gate bias rate, more electrons on the ZnO nanowire surface are trapped by the adsorbed oxygen ions, which results in a reduction of nanowire conduction channel region by the more depleted region. Thus, the current decreases, and the threshold voltage shifts to the positive gate bias direction. The effect of the gate bias sweep rate on the ZnO nanowire FET was investigated under different oxygen environments. The same ZnO nanowire FET device measured in ambient air (Figure 14.10a) was placed into an N_2-filled glove box after holding at a vacuum (~10^{-5} Torr) for 24 h. The I_{DS}–V_G curves measured in this N_2 environment are shown in Figure 14.10b. The conductivity of the ZnO nanowire FET device in the N_2 environment increased by about an order of magnitude, compared with the case of ambient air. In the N_2 environment, the nanowire FET has fewer oxygen ions on the surface, thus less surface depletion of electrons, resulting in an increase of the current. Unlike the case of the nanowire measured in ambient air, the nanowire device (unpassivated) in the N_2 environment was not influenced by the gate bias sweep rate, as is shown in Figure 14.10b. The current did not change and the threshold voltage did not shift. This is because of the relative absence of oxygen in the N_2-filled glove box; therefore, there is a negligible effect of oxygen absorption even at different gate bias sweep rates. In addition, the effect of the gate bias sweep rate on the same ZnO nanowire FET device after the passivation of PMMA layer was explored. The resulting I_{DS}–V_G curves after the passivation, measured in ambient air, are shown in Figure 14.10c. Although the change and shift are not as significant, the I_{DS}–V_G curves exhibited somewhat similar trends of the current change and threshold voltage shift for different gate bias sweep rates as the case of Figure 14.10a. Therefore, the effect of gate bias sweep rate on the nanowire channel can be explained as the following. Slower gate bias sweep rates imply longer gate biasing time. Captured oxygen ions are strongly bonded by positive gate bias, so that oxygen ions would not be easily detached. Thus, slow gate bias sweep rates will cause more oxygen absorption than fast gate bias sweep rates. More adsorbed oxygen molecules extend the surface depletion region. Consequently, a larger depletion region in the nanowire occurs by slower gate bias sweep rates and longer gate-biasing time, which results in the reduction of current and carrier density and a shift of gate bias in the positive gate bias direction. On the contrary, faster gate bias sweep rates are applied in relatively shorter gate biasing time. Therefore, fast gate bias sweep rates lead to less oxygen absorption.

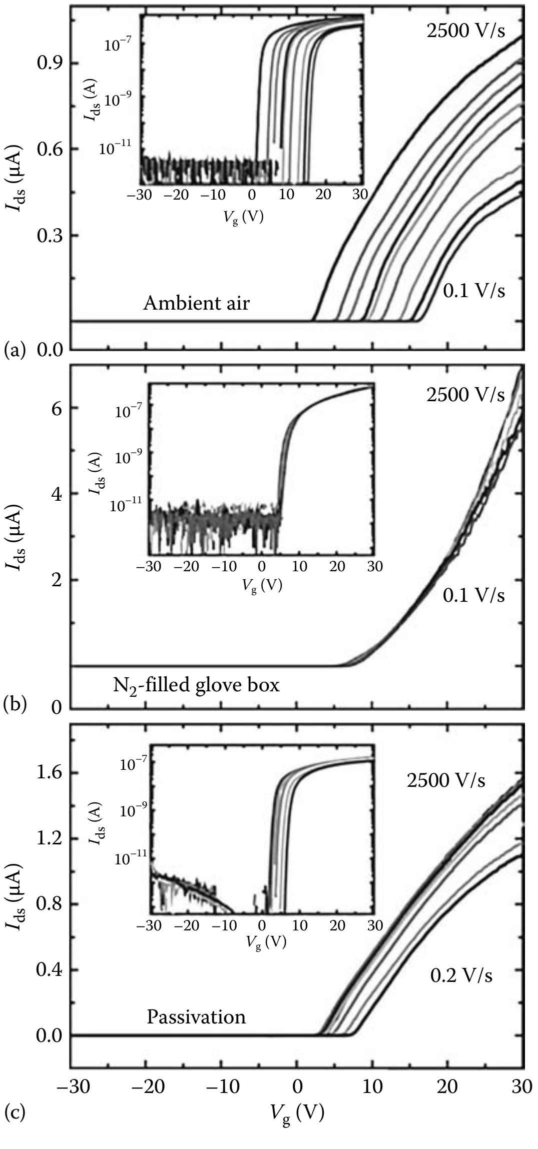

FIGURE 14.10 I_{DS}–V_G curves for a ZnO nanowire FET as a function of gate bias sweep rate (2500, 250, 130, 100, 12, 6, 1.2, 0.3, 0.2, and 0.1 V/s) measured at $V_{DS} = 0.5$ V (a) under ambient air, (b) N_2-filled glove box, and (c) ambient air after PMMA passivation. The semilogarithmic plots of I_{DS}–V_G curves are shown in the insets. (Adapted from Maeng, J. et al., *Appl. Phys. Lett.*, 92, 233120, 2008. With permission.)

The current, threshold voltage, and carrier density were sensitively dependent on the gate bias sweep rate. These phenomena became more pronounced under environments with increased oxygen. The results of this study may offer how to characterize the surface effects of ZnO nanowire devices or other materials.

14.3.3 N-Channel Depletion-Mode and Enhancement-Mode ZnO Nanowire FETs

To date, although FETs using ZnO nanowires as active channels have been extensively investigated, most of the fabricated ZnO nanowire FETs were typically normally on type, n-channel depletion-mode (D-mode) behavior, which exhibited nonzero current at zero gate bias and negative threshold voltages. However, for a wide number of the applications of nanowire FETs in logic circuits, both D-mode and enhancement-mode (E-mode) FETs are required (Sedra and Smith 1991, Fortunato et al. 2004, Lee et al. 2006, Ma et al. 2007).

The electrical characteristics of D-mode and E-mode ZnO nanowire FETs are summarized in Figure 14.11. All the fabricated ZnO nanowire FETs were passivated by PMMA layer to remove the influence of water or gas molecules in ambient air on nanowire and to improve the FET performance by enhancing the gate-coupling effects. Note that although the structure of ZnO nanowire FET device is different from the conventional metal-oxide-semiconductor transistors, the operation mode of ZnO nanowire FETs is distinguished as D-mode or E-mode in terms of the polarity of the threshold voltage (Hong et al 2007). ZnO nanowire FETs with negative threshold voltages are normally on type, n-channel D-mode transistors, indicating that more negative gate bias should be applied to deplete carriers in the channel to reduce channel conductance, since the n-channel for current flow already exists at zero gate bias. In contrast, ZnO nanowire FETs with positive threshold voltages are normally off type, n-channel E-mode transistors, indicating that more positive gate bias is needed to make the channel, since the channel current does not flow at zero gate bias.

Figure 14.11a and b shows the electrical characteristics for n-channel D-mode ZnO nanowire FETs, whereas Figure 14.11c and d shows those for n-channel E-mode FETs. The I_{DS}–V_{DS} curves of both D-mode and E-mode FETs have well-defined linear regimes at low biases and saturation regimes at high biases. Figure 14.11b shows that the threshold voltage V_{th} is −4.14 V, indicating n-channel D-mode behavior. The I_{DS}–V_G plot in the semilogarithmic scale displays an on/off current ratio as large as 10^5 (inset of Figure 14.11b). In contrast, in the I_{DS}–V_G curves of Figure 14.11d, the threshold voltage V_{th} is +10.85 V, indicating n-channel E-mode behavior. The I_{DS}–V_G plot in the semilogarithmic scale shows an on/off current ratio as large as 10^6 (inset of Figure 14.11d).

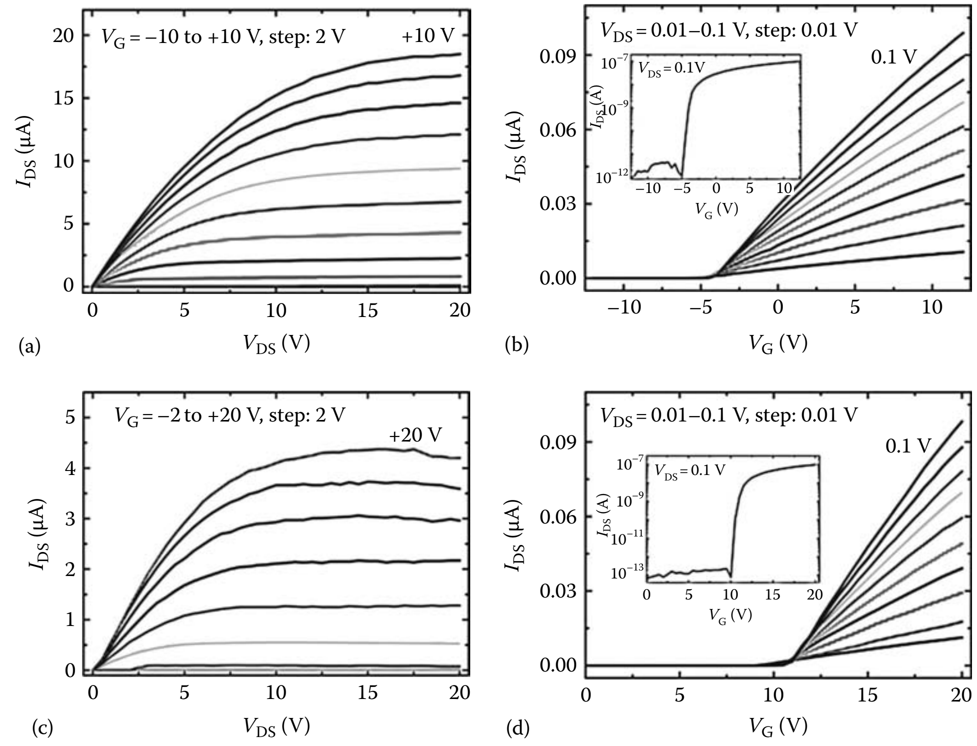

FIGURE 14.11 (a) I_{DS}–V_{DS} and (b) I_{DS}–V_G curves for n-channel D-mode FET of ZnO nanowires grown on an Au-coated sapphire substrate. The inset shows the semilogarithmic plot of the I_{DS}–V_G curve at V_{DS} = 0.1 V. (c) I_{DS}–V_{DS} and (d) I_{DS}–V_G curves for n-channel E-mode FET of ZnO nanowires grown on an Au-catalyst-free ZnO film. The inset shows the semilogarithmic plot of the I_{DS}–V_G curve at V_{DS} = 0.1 V. (Adapted from Hong, W.-K. et al., *Appl. Phys. Lett.*, 90, 243103, 2007. With permission.)

14.3.4 Electronic Transport in Surface-Architecture-Controlled ZnO Nanowire FETs

It is well-known that interface roughness plays an important role in electronic transport for transistor devices (Hastas et al. 2002, Wang et al. 2005). In particular, the electronic transport in nanowire transistors can be influenced by the surface morphology of nanowires associated with the interface roughness, including the presence of surface trap states at the interfaces between the nanowires and the dielectric layers of nanowire FETs (Wang et al. 2004, Dayeh et al. 2007a,b). In order to investigate the influence of surface states and/or defects (associated with surface roughness and nanowire size) on the electronic transport properties in nanowire transistors, the surface morphology- and size-controlled ZnO nanowires were grown on the three sets of sapphire substrates with ZnO buffer films: (1) an undoped ZnO film with Au catalyst (denoted as Au–ZnO film) or without Au catalyst (denoted as ZnO film), (2) a gallium-doped ZnO film with Au catalyst (denoted as Au–GZO film) or without Au catalyst (denoted as GZO film), (3) an aluminum-doped ZnO film with Au catalyst (denoted as Au–AZO film) or without Au catalyst (denoted as AZO film). In addition, an Au-coated sapphire (denoted as Au-sapphire) substrate was also used for comparison with the ZnO buffer film-coated sapphire substrates. ZnO buffer films with approximately 1 μm thickness were grown on the sapphire substrates by a radio frequency (rf) sputtering system using a commercially sintered ZnO target, a Ga_2O_3 (1 wt%)-doped ZnO target, and an Al_2O_3 (1 wt%)-doped ZnO target. The Au thin film (~3 nm) was deposited on some substrates using an e-beam evaporator to form Au catalysts for the growth of nanowires.

Figure 14.12 shows the I_{DS}–V_G curves at V_{DS} = 0.1 V of the surface-architecture-controlled ZnO nanowire FETs, which exhibit different electronic transport characteristics. Insets in Figure 14.10 show the low-magnification TEM images of the surface-architecture-controlled ZnO nanowires grown on different substrates. The ZnO nanowires with relatively smaller diameters were found to have rougher surfaces than those with relatively larger diameters. The rough ZnO nanowires can have a larger surface-area-to-volume ratio than smooth ZnO nanowires. The FET devices made from smooth ZnO nanowires (nanowires grown on GZO, Au–AZO films, and Au-sapphire) have negative threshold voltages, indicating n-channel D-mode behavior that exhibits nonzero current at the zero gate bias. In contrast, the FETs made from rough ZnO nanowires (nanowires grown on Au–ZnO, ZnO, Au–GZO, and AZO films) have positive threshold voltages, indicating n-channel E-mode behavior that exhibits off-current status at zero gate bias.

Figure 14.13 shows cross-sectional schematics (Figure 14.13a and c) across the electrodes, ZnO nanowire, and dielectric layers with the corresponding equilibrium energy band diagram (Figure 14.13b and d) of the ZnO nanowire FETs at V_G = 0 V. Figure 14.13a and b shows the case for smooth ZnO nanowires, and Figure 14.13c and d depicts the case for rough ZnO

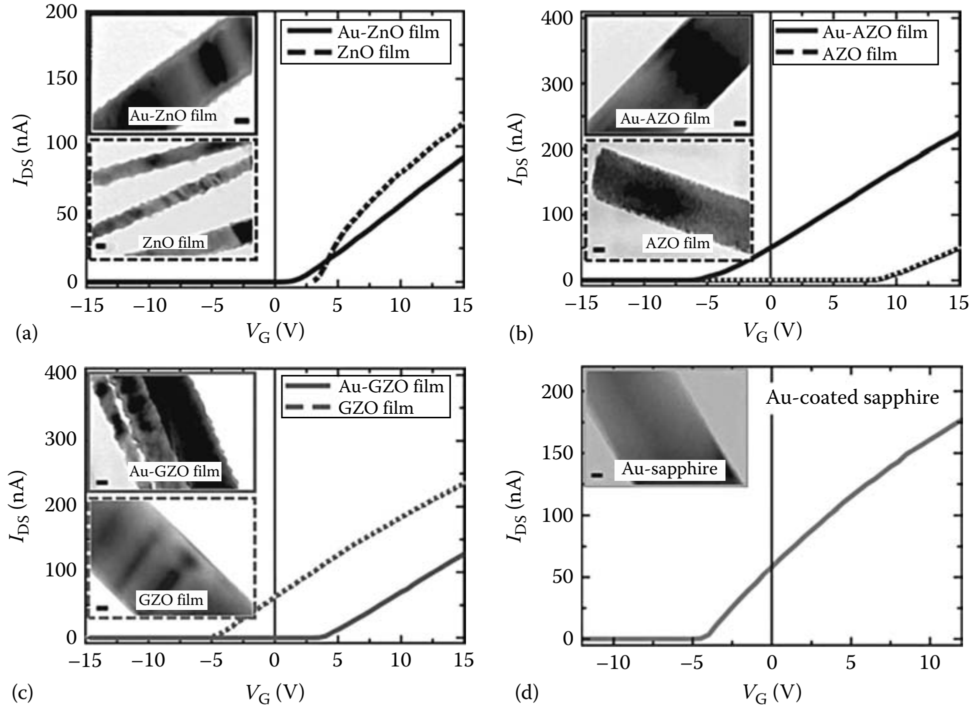

FIGURE 14.12 I_{DS}–V_G curves (measured at V_{DS} = 0.1 V) of FETs made from ZnO nanowires grown on (a) Au–ZnO and ZnO films, (b) Au–AZO and AZO films, (c) Au–GZO and GZO films, and (d) as Au-sapphire substrate. Labels are explained in the text. The insets show the low magnification TEM images of the surface-architecture-controlled ZnO nanowires. (Adapted from Hong, W.-K. et al., *Nano Lett.* 8, 950, 2008b. With permission.)

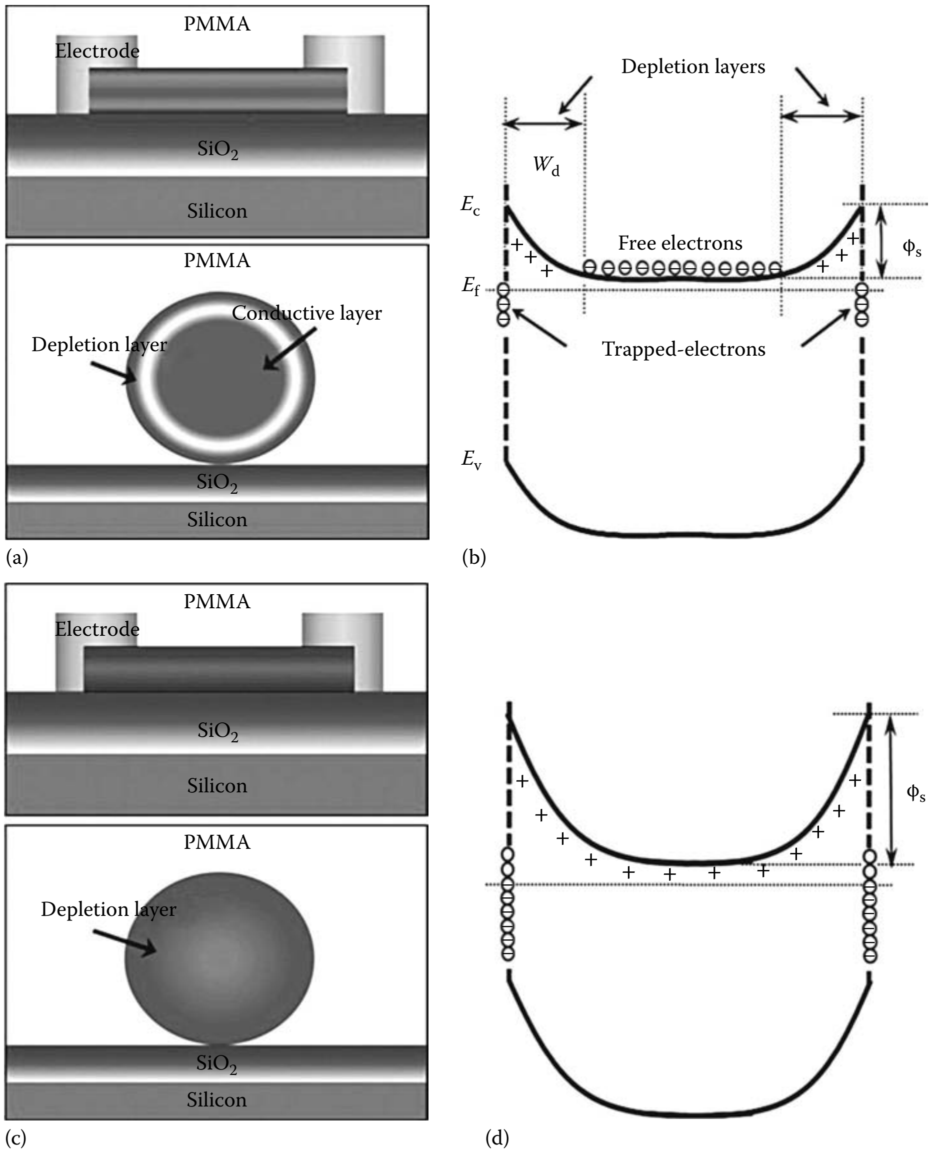

FIGURE 14.13 A cross-sectional view across the electrodes, ZnO nanowire, and dielectric layers (a and c) and the corresponding equilibrium energy band diagrams of nanowire FETs at $V_G = 0$ V (b and d) for smooth ZnO nanowires (a and b) and rough ZnO nanowires (c and d). (Adapted from Hong, W.-K. et al., *Nano Lett.*, 8, 950, 2008b. With permission.)

nanowires. For single crystalline semiconductor materials, band bending due to Fermi-level pinning can occur at surfaces and/or interfaces because the surface states and/or defects on single crystalline nanowires can induce trap energy levels at the interfaces (Wang et al. 2004, Calarco et al. 2005, Jones et al. 2007, Liao et al. 2007). The trapping of carrier electrons in the trap states can cause electron depletion in the channel, resulting in a threshold voltage shift and a conductance modulation (Ikeda 2002, Hossain et al. 2003, Fan et al. 2004a,b, Ju et al. 2007, Yoon et al. 2007). Correspondingly, the different electronic transport behavior between smooth and rough ZnO nanowires can be explained by considering the depletion of electron carriers due to surface band bending at the PMMA/ZnO nanowire and/or ZnO nanowire–SiO_2 interfaces as shown in Figure 14.13. Note that for the convenience of discussion and simplification of charge transport mechanism, it is assumed that the uniform charge and gate potential distribution for nanowires is with circular cross sections. However, the typical back-gate nanowire FET geometry can be significantly influenced by the cross-sectional shapes of nanowires and whether the nanowire is embedded or not (Wunnicke 2006). In addition, the back-gate nanowire FET geometry has the limitation in accuracy for calculating the gate-nanowire capacitance as well as the complication for the experimental extraction of transport parameters under the presence of depletion region and trap states (Dayeh et al. 2007a,b, Khanal and Wu 2007).

The depletion region by surface band bending and the effective conduction region (or effective diameter) can be estimated as (Taur and Ning 1998, Yu and Cardona 2001, Wang et al. 2004, Goldberger et al. 2005)

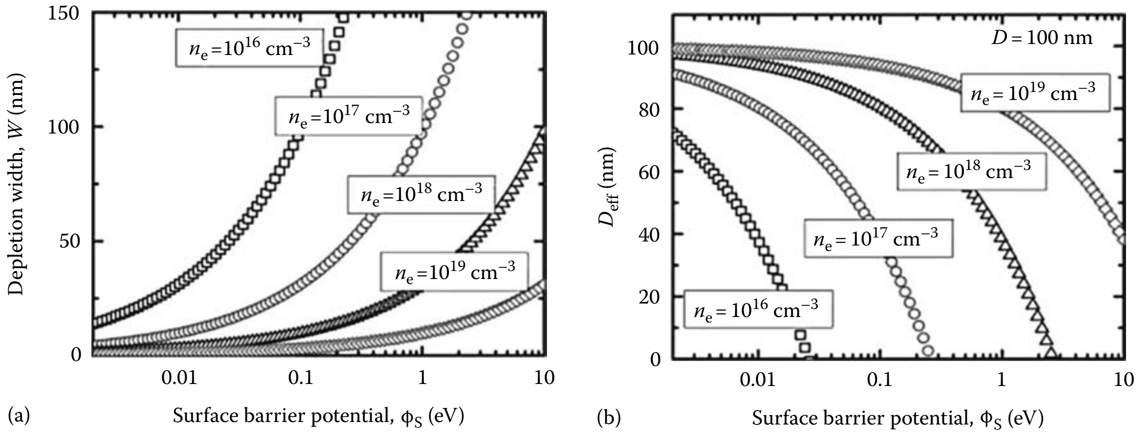

FIGURE 14.14 (a) Depletion width (W_d) and (b) effective diameter (D_{eff}) of ZnO nanowires as a function of surface barrier potential of ZnO nanowires at $N_D = 10^{16}$–10^{19} cm^{-3}.

$$W_d = \left(\frac{2\varepsilon_{ZnO}\,\phi_S}{e\,N_D} \right)^{1/2} \tag{14.1}$$

where

ϕ_S is the surface barrier potential
e is the electronic charge
N_D is the doping density
ε_{ZnO} is the dielectric constant of ZnO

Then by using the depletion approximation and charge neutrality condition (Taur and Ning 1998, Yu and Cardona 2001, Hossain et al. 2003), the surface trap density (N_t) can be estimated as $N_t = 2N_D W_d$. Thus, the effective diameter (D_{eff}) of the conduction layer in the nanowire channel can be expressed as

$$D_{eff} = D - \left(\frac{N_t}{N_D} \right) \tag{14.2}$$

where D is the geometrical diameter of nanowire. Therefore, by assuming that the $N_D = 10^{17}/\text{cm}^3$, $\varepsilon_{ZnO} = 8.66$ (Fan and Lu 2005), $\phi_S = 0.3$ eV, and the nanowire diameter (D) is 100 nm, we obtain $W_d = 54$ nm and $D < 2W_d$ from Equations 14.1 and 14.2 as shown in Figure 14.14.

14.4 Summary

In order to fabricate the functional devices from the ZnO nanowires, their fundamental properties must be carefully investigated because the presence of a large surface and interfacial area in nanowires can profoundly alter their performance. Therefore, it is essential to understand the surface-roughness- and size-dependent effects on the electronic transport properties of nanowire transistors. Furthermore, since ZnO nanowires are strongly influenced by chemical environments due to large surface-to-volume ratio, it is also important to study on the passivation effects of dielectric layer on the transport properties correlated to geometry and adsorbed species on the ZnO nanowires. In this chapter, we reviewed the electrical properties of FETs made from the surface-tailored smooth and rough ZnO nanowires that were grown by a vapor transport method. The transport properties of ZnO nanowire FETs are associated with the surface roughness, nanowire size, surface states, and/or defects as well as the surface chemistry in surrounding environments. The characteristics show that the electrical properties of ZnO nanowire FETs must be carefully investigated due to the various effects of nanowire surface and surrounding environments as well as interfaces. In the field of nanowire-based electronics, these explorations are important in understanding the transport properties and developing into practically useful device applications.

In future, in order to enable ZnO nanowire FETs to function as building blocks for the next-generation electronic circuitry, the ability to rationally grow nanowires with precisely the control of chemical composition, size, morphology, structure, and doping concentration will be required. Furthermore, the implementation of assembling any kind of functional nanodevices into highly integrated nanosystems must be intensively investigated.

Acknowledgment

This work was supported in part by the National Research Laboratory (NRL) Program, the National Core Research Center (NCRC) grant, and the World Clan University (WCU) program by the Korean Ministry of Education, Science and Technology (MEST).

References

Agarwal, R. 2008. Heterointerfaces in semiconductor nanowires. *Small* 4: 1872–1893.

Arora, N. 1993. *MOSFET Models for VLSI Circuit Simulation*. New York: Springer, Wien.

Bao, J., Zimmler, M. A., and Capasso, F. 2006. Broadband ZnO single-nanowire light-emitting diode. *Nano Lett.* 6: 1719–1722.

Calarco, R., Marso, M., Richter, T. et al. 2005. Size-dependent photoconductivity in MBE-grown GaN-nanowires. *Nano Lett*. 5: 981–984.

Cha, S. N, Jang, J. E., Choi, Y. et al. 2006. High performance ZnO nanowire field effect transistor using self-aligned nanogap gate electrodes. *Appl. Phys. Lett*. 89: 263102 (3 pp).

Chang, P.-C., Fan, Z., Chien, C.-J. et al. 2006. High-performance ZnO nanowire field effect transistors. *Appl. Phys. Lett*. 89: 133113 (3 pp).

Dai, Z. R., Pan, Z. W., and Wang, Z. L. 2003. Novel nanostructures of functional oxides synthesized by thermal evaporation. *Adv. Funct. Mater*. 13: 9–24.

Dayeh, S. A., Soci, C., Yu, P. K. L., Yu, E. T., and Wang, D. 2007a. Influence of surface states on the extraction of transport parameters from InAs nanowire field effect transistors. *Appl. Phys. Lett*. 90: 162112 (3 pp).

Dayeh, S. A., Soci, C., Yu, P. K. L., Yu, E. T., and Wang, D. 2007b. Transport properties of InAs nanowire field effect transistors: The effects of surface states. *J. Vac. Sic. Technol. B* 25: 1432–1436.

Dehuff, N. L., Kettenring, E. S., Hong, D. et al. 2005. Transparent thin-film transistors with zinc indium oxide channel layer. *J. Appl. Phys*. 97: 064505 (5 pp).

Dorfman, A., Kumar, N., and Hahm, J. 2006. Nanoscale ZnO-enhanced fluorescence detection of protein interactions. *Adv. Mater*. 19: 2685–2690.

Fan, Z. and Lu, J. G. 2005. Zinc oxide nanostructure: Synthesis and properties. *J. Nanosci. Nanotechnol*. 5: 1561–1573.

Fan, Z. and Lu, J. G. 2006. Chemical sensing with ZnO nanowire field-effect transistor. *IEEE Trans. Nanotechnol*. 5: 393–396.

Fan, Z., Chang, P.-C., Lu, J. G. et al. 2004a. Photoluminescence and polarized photodetection of single ZnO nanowires. *Appl. Phys. Lett*. 85: 6128–6130.

Fan, Z., Wang, D., Chang, P.-C., Tseng, W.-Y., and Lu, J. G. 2004b. ZnO nanowire field-effect transistor and oxygen sensing property. *Appl. Phys. Lett*. 85: 5923–5925.

Fortunato, E. M. C., Barquinha, P. M. C., Pimentel, C. C. M. B. G. et al. 2004. Wide-bandgap high-mobility ZnO thin-film transistors produced at room temperature. *Appl. Phys. Lett*. 85: 2541–2543.

Goldberger, J., Sirbuly, D. J., Law, M., and Yang, P. 2005. ZnO nanowire transistors. *J. Phys. Chem. B* 109: 9–14.

Hastas, N. A., Dimitriadis, C. A., and Kamarinos, G. 2002. Effect of interface roughness on gate bias instability of polycrystalline silicon thin-film transistors. *J. Appl. Phys*. 92: 4741–4745.

Hong, W.-K., Hwang, D.-K., Park, I.-K. et al. 2007. Realization of highly reproducible ZnO nanowire field effect transistors with n-channel depletion and enhancement modes. *Appl. Phys. Lett*. 90: 243103 (3 pp).

Hong, W.-K., Kim, B.-J., Kim, T.-W. et al. 2008a. Electrical properties of ZnO nanowire field effect transistors by surface passivation. *Colloids Surf. A* 313–314: 378–382.

Hong, W.-K., Sohn, J. I., Hwang, D.-K. et al. 2008b. Tunable electronic transport characteristics of surface-architecture-controlled ZnO nanowire field effect transistors. *Nano Lett*. 8: 950–956.

Hong, W.-K., Song, S., Hwang, D.-K. et al. 2008c. Effects of surface roughness on the electrical characteristics of ZnO nanowire field effect transistors. *Appl. Surf. Sci*. 254: 7559–7564.

Hossain, F. M., Nishii, J., Takagi, S. et al. 2003. Modeling and simulation of polycrystalline ZnO thin-film transistors. *J. Appl. Phys*. 94: 7768–7777.

Hsueh, T.-J., Chang, S.-J., Hsu, C.-L., Lin, Y.-R., and Chen, I.-C. 2007. Highly sensitive ZnO nanowire ethanol sensor with Pd adsorption. *Appl. Phys. Lett*. 91: 053111 (3 pp).

Huang, Y., Duan, X. F., Cui, Y. et al. 2001. Logic gates and computation from assembled nanowire building blocks. *Science* 294: 1313–1317.

Ikeda, H. 2002. Evaluation of grain boundary trap states in polycrystalline–silicon thin-film transistors by mobility and capacitance measurements. *J. Appl. Phys*. 91: 4637–4645.

Jones, F., Léonard, F., Talin, A. A., and Bell, N. S. 2007. Electrical conduction and photoluminescence properties of solution-grown ZnO nanowires. *J. Appl. Phys*. 102: 014305 (7 pp).

Ju, S., Lee, K., Janes, D. B. et al. 2006. ZnO nanowire field-effect transistors: Ozone-induced threshold voltage shift and multiple nanowire effects. In *Proceedings of the 6th IEEE Conference on Nanotechnology,* Cincinnati, OH, pp. 445–448.

Ju, S., Lee, K., Yoon, M.-H. et al. 2007. High performance ZnO nanowire field effect transistors with organic gate nanodielectrics: Effects of metal contacts and ozone treatment. *Nanotechnology* 18: 155201 (7 pp).

Jung, Y., Ko, D.-K., and Agarwal, R. 2007. Synthesis and structural characterization of single-crystalline branched nanowire heterostructures. *Nano Lett*. 7: 264–268.

Khanal, D. R. and Wu, J. 2007. Gate coupling and charge distribution in nanowire field effect transistors. *Nano Lett*. 7: 2778–2783.

Kim, W., Javey, A., Vermesh, O. et al. 2003. Hysteresis caused by water molecules in carbon nanotube field-effect transistors. *Nano Lett*. 3: 193–198.

Kim, H.-J., Lee, C.-H., Kim, D.-W., and Yi, G.-C. 2006. Fabrication and electrical characteristics of dual-gate ZnO nanorod metal-oxide semiconductor field-effect transistors. *Nanotechnology* 17: S327–S331.

Kind, H., Yan, H., Messer, B., Law, M., and Yang, P. 2002. Nanowire ultraviolet photodetectors and optical switches. *Adv. Mater*. 14: 158–160.

Klingshirn, C. 2007. ZnO: From basis towards applications. *Phys. Stat. Sol. B* 244: 3027–3073.

Kolmakov, A. and Moskovits, M. 2004. Chemical sensing and catalysis by one-dimensional metal-oxide nanostructures. *Annu. Rev. Mater. Res*. 34: 151–180.

Lagowski, J., Sproles, E. S. Jr., and Gatos, H. C. 1977. Quantitative study of the charge transfer in chemisorption; oxygen chemisorption on ZnO. *J. Appl. Phys*. 48: 3566–3575.

Law, M., Goldberger, J., and Yang, P. 2004. Semiconductor nanowires and nanotubes. *Annu. Rev. Mater. Res*. 34: 83–122.

Law, M., Greene, L. E., Johnson, J. C., Saykally, R., and Yang, P. 2005. Nanowire dye-sensitized solar cells. *Nat. Mater*. 4: 455–459.

Lee, C. A., Jin, S. H., Jung, K. D., Lee, J. D., and Park, B.-G. 2006. Full-swing pentacene organic inverter with enhancement-mode driver and depletion-mode load. *Solid State Electron*. 50: 1216–1218.

Li, Q. H., Liang, Y. X., Wan, Q. et al. 2004. Oxygen sensing characteristics of individual ZnO nanowire transistors. *Appl. Phys. Lett*. 85: 6389–6391.

Li, Q. H., Gao, T., Wang, Y. G., and Wang, T. H. 2005. Adsorption and desorption of oxygen probed from ZnO nanowire films by photocurrent measurements. *Appl. Phys. Lett*. 86: 123117 (3 pp).

Liao, Z.-M., Liu, K.-J., Zhang, J.-M., Xu, J., and Yu, D.-P. 2007. Effect of surface states on electron transport in individual ZnO nanowires. *Phys. Lett. A* 367: 207–210.

Look, D. C., Reynolds, D. C., Hemsky, J. W., Jones, R. L., and Sizelove, J. R. 1999. Production and annealing of electron irradiation damage in ZnO. *Appl. Phys. Lett*. 75: 811–813.

Lu, W. and Lieber, C. M. 2007. Nanoelectroncis from the bottom up. *Nat. Mater*. 6: 841–850.

Ma, R. M., Dai, L., and Qin, G. G. 2007. Enhancement-mode metal-semiconductor field-effect transistors based on single n-CdS nanowires. *Appl. Phys. Lett*. 90: 093109 (3 pp).

Maeng, J., Jo, G., Kwon, S.-S. et al. 2008. Effect of gate bias sweep rate on the electronic properties of ZnO nanowire field-effect transistors under different environments. *Appl. Phys. Lett*. 92: 233120 (3 pp).

Mieszawska, A. J., Jalilian, R., Sumanasekera, G. U., and Zamborini, F. P. 2007. The synthesis and fabrication of one-dimensional nanoscale heterojunctions. *Small* 3: 722–756.

Ng, H. T., Han, J., Yamada, T. et al. 2004. Single crystal nanowire vertical surrounded-gate field-effect transistor. *Nano Lett*. 4: 1247–1252.

Park, W. I., Yi, G.-C., Miyoung K., and Pennycook, S. J. 2003. Quantum confinement observed in ZnO/ZnMgO nanorod heterostructures. *Adv. Mater*. 15: 526–529.

Park, W. I, Kim, J. S., Yi, G.-C., Bae, M. H., and Lee, H.-J. 2004. Fabrication and electrical characteristics of high-performance ZnO nanorod field-effect transistors. *Appl. Phys. Lett*. 85: 5052–5054.

Park, W. I., Kim, J. S., Yi, G.-C., and Lee, H.-J. 2005. ZnO nanorod logic circuits. *Adv. Mater*. 17: 1393–1397.

Sadek, A. Z., Choopun, S., Wlodarski, W. et al. 2007. Characterization of ZnO nanoblet-based gas sensor for H_2, NO_2, and hydrocarbon sensing. *IEEE Sens. J*. 7: 919–924.

Schmidt-Mende, L. and MacManus-Driscoll, J. L. 2007. ZnO-nanostructures, defects, and devices. *Mater. Today* 10: 40–48.

Sedra, A. S. and Smith, K. C. 1991. *Microelectronic Circuits*, 3rd edn. Orlando, FL: Harcourt Brace College Publishers.

Soci, C., Zhang, A., Xiang, B. et al. 2007. ZnO nanowire UV photodetectors with high internal gain. *Nano Lett*. 7: 1003–1009.

Song, S., Hong, W.-K., Kwon, S.-S., and Lee, T. 2008. Passivation effects of ZnO nanowire field effect transistors under oxygen, ambient, and vacuum environments. *Appl. Phys. Lett*. 92: 263109 (3 pp).

Streetman, B. G. and Banerjee, S. 2000. *Solid State Electronic Devices*. Upper Saddle River, NJ: Prentice Hall.

Suehiro, J., Nakagawa, N., Hidaka, S.-I. et al. 2006. Dielectrophoretic fabrication and characterization of a ZnO nanowire-based UV photosensor. *Nanotechnology* 17: 2567–2573.

Sun, Y., Fuge, G. M., and Ashfold, M. N. R. 2006. Growth mechanisms for ZnO nanorods formed by pulsed laser deposition. *Superlattice. Microstruet*. 39: 33–40.

Sze, S. M. 1981. *Physics of Semiconductor Devices*. New York: John Wiley & Sons.

Takata, M., Tsubone, D., and Yanagida, H. 1976. Dependence of electrical conductivity of ZnO on degree of sintering. *J. Am. Ceram. Soc*. 59: 4–8.

Taur, Y. and Ning, T. H. 1998. *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge University Press.

Wade, T. L., Hoffer, X., Mohammed, A. D. et al. 2007. Nanoporous alumina wire templates for surrounding-gate nanowire transistors. *Nanotechnology* 18: 125201 (4 pp).

Wang, D., Chang, Y.-L., Wang, Q. et al. 2004. Surface chemistry and electrical properties of germanium nanowires. *J. Am. Chem. Soc*. 126: 11602–11611.

Wang, J., Polizzi, E., Ghosh, A., Datta, S., and Lundstrom, M. 2005. Theoretical investigation of surface roughness scattering in silicon nanowire transistors. *Appl. Phys. Lett*. 87: 043101 (3 pp).

Wang, D., Seo, H. W., Tin, C.-C. et al. 2006. Effects of postgrowth annealing treatment on the pthotoluminescence of zinc oxide nanorods. *J. Appl. Phys*. 99: 113509 (3 pp).

Wolkenstein, T. 1991. *Electronic Processes on Semiconductor Surfaces during Chemisorptions*, (ed.) R. Morrison. New York: Consultants Bureau.

Wunnicke, O. 2006. Gate capacitance of back-gated nanowire field-effect transistors. *Appl. Phys. Lett*. 89: 083102 (3 pp).

Xiong, H. D., Wang, W., Li, Q. et al. 2007. Random telegraph signals in n-type ZnO nanowire field effect transistors at low temperature. *Appl. Phys. Lett*. 91: 053107 (3 pp).

Yang, P., Yan, H., Mao, S. et al. 2002. Controlled growth of ZnO nanowires and their optical properties. *Adv. Funct. Mater*. 12: 323–331.

Yeom, D., Keem, K., Kang, J. et al. 2008. NOT and NAND logic circuits composed of top-gate ZnO nanowire field-effect transistors with high-k Al_2O_3 gate layers. *Nanotechnology* 19: 265202 (5 pp).

Yoon, Y., Lin, J., Pearton, S. J., and Guo, J. 2007. Role of grain boundaries in ZnO nanowire field-effect transistors. *J. Appl. Phys*. 101: 024301 (5 pp).

Yu, P. Y. and Cardona, M. 2001. *Fundamentals of Semiconductors: Physics and Materials Properties*. Berlin, Germany: Springer-Verlag.

Zhang, Y., Kolmakov, A., Chretien, S. et al. 2004. Control of catalytic reactions at the surface of a metal oxide nanowire by manipulating electron density inside it. *Nano Lett*. 4: 403–407.

15

C_{60} Field Effect Transistors

15.1 Introduction 15-1
Fullerene (C_{60}) • Solid C_{60} Crystals • C_{60} Field Effect Transistors (C_{60} FETs) • Bell Laboratory Misconduct • Interfaces between Solid C_{60} and Insulator Films • Field Effect Doping • Operation Principles of C_{60} FETs • Superconductivity • Guides for This Chapter
15.2 Fullerene (C_{60}) 15-4
Discovery of Fullerene Family • Synthesis of C_{60} Molecules • Properties of C_{60} Molecule
15.3 Solid C_{60} 15-7
Crystal Structure • van der Waals Interaction of C_{60} • Structural Phase Transition • Electronic Structures • Optical Properties
15.4 van der Waals Epitaxy of C_{60} 15-10
Molecular Beam Epitaxy • van der Waals Epitaxy • Initial Growth Surface
15.5 Principles of Organic FET Functions 15-16
Device Structures of OFETs • Metal/Organic Layer Interface • Field Effect Doping for Solid C_{60} Layers • Am-Bipolar OFET Operations
15.6 Fabrication of Fullerene Field Effect Transistors (C_{60} FET) 15-18
Substrates: SiO_2 or AlN • Device Process • C_{60} Epitaxy • Measurements of FET Performance
15.7 Characterization of C_{60} FETs on SiO_2 or on AlN 15-19
I–V Characteristics • Sub-Threshold Region
15.8 Summary 15-20
References 15-21

Akihiro Hashimoto
University of Fukui

15.1 Introduction

15.1.1 Fullerene (C_{60})

Fullerene is a new carbon material system closely related with carbon micro-clusters and/or carbon ultrafine particles. In general, *micro-clusters* are defined as atomic clusters consisting of less than 1000 atoms. On the other hand, *ultrafine particles* are usually defined as particles consisting of atoms above the upper limit of micro-clusters. The discovery of fullerene was achieved in the process of researching of carbon micro-clusters and/or carbon ultrafine particles. Fullerene is the name of a family of carbon clusters such as C_{60}, C_{70}, and so on. In this chapter, we will use the word "fullerene" mainly as the name of the C_{60} clusters.

C_{60} clusters were first physically observed as a magic number cluster in the mass spectrum of cluster molecular beams, although their theoretical prediction had already existed in the pioneering research of these clusters. Furthermore, several large-scale production methods for C_{60} clusters were found in a study on the formation process of carbon fine particles from the resistive heating for carbon rods. In a sense, the discovery of the fullerene was due to serendipity. Further information will be provided in Section 15.2.

15.1.2 Solid C_{60} Crystals

High-purity solid C_{60} molecular crystals can be obtained from high-purity C_{60} powders by using the sublimation method or the vacuum evaporation method in an ultrahigh vacuum (UHV) environment similar to the molecular beam epitaxy (MBE). This method is preferable to the solution growth from a solvent for the separation of the C_{60} molecules because of the reduction of residual impurities during growth. The solid C_{60} crystal has a face-centered cubic structure (fcc, $Fm\bar{3}m$ symmetry) above 260 K, and the lattice constant is 1.4154 nm at 270 K. It is well known that each of the C_{60} molecules in the solid C_{60} crystal rotates continuously very fast with a rotation correlation time on the order of 10 ps. The C_{60} molecules behave as a sphere with a diameter of 0.71 nm. Therefore, the solid C_{60} crystal is called "the plastic crystal" with "a rotational diffusion" of the C_{60} molecules. The rotational diffusion means that the degrees of freedom of the molecular rotation are softened before the ones of the translational displacements. The electronic structure of the solid C_{60} has been extensively studied by many groups, and its semiconductor properties such as an energy band gap of about 1.6 eV are commonly known. Moreover, one of the alkali-doped C_{60}s, $RbCs_2C_{60}$, shows superconductivity with a critical temperature of 33 K. In addition, it is well known that all alkali-doped C_{60} systems, in general, show superconductivity. The epitaxy of the solid C_{60} layer has also been extensively studied by many groups known as the van der Waals epitaxy. The concept of lattice mismatches of the epitaxial layers to the substrate is applicable as they are similar to the cases of the conventional semiconductor epitaxy. Therefore, it is a little difficult to obtain high-quality solid C_{60} layers without any dislocations and/or any boundary for high lattice-mismatch combinations even in the van der Waals epitaxy. Further information will be provided in Sections 15.3 and 15.4.

15.1.3 C_{60} Field Effect Transistors (C_{60} FETs)

Fullerene field effect transistors (C_{60} FETs) are typical of the family of the organic FETs (OFETs). It is well known that solid C_{60} FETs have the highest electron mobility in the organic FET family, although most OFETs show the *p*-type, as shown in Figure 15.1 [1]. The various types of the OFET have been extensively studied in recent years, because of their large area fabrication, flexibility, light weight, and low cost. The most essential reason for their extensive research may be due to a resource problem in conventional Si and compound semiconductors. Carbon materials exist almost infinitely on earth and the modification of carbon compounds is very easy to obtain, because of vast chemical knowledge known about carbon compounds. Essentially, organic materials used in the OFETs are high-resistance materials without doping and, moreover, rather unique electronic properties appear even in the doped organic materials due to a strong electron–lattice interaction compared with conventional covalent semiconductors such as Si and GaAs. The doping materials in organic semiconductors, in general, do not form impurity levels in the mid-gap of the band gap. However, the electrons of highest isolated occupied molecular orbital (HOMO) levels of the doping materials can transfer directly to the lowest unoccupied molecular orbital (LUMO) levels of the organic semiconductor. In general, such charge transfer process is often accompanied with strong lattice distortion. Consequently, the mid-gap level or the impurity band never forms in conductive organic materials. Almost all conductive organic materials have similar electronic properties to intrinsic semiconductor systems even in impurity doping. The conductivity in the organic materials essentially comes from the properties of the electrode metals for organic materials. In that sense, the conventional Shockley–Bardeen theory cannot be applied to OFET operations. The situation for general OFETs is also common for solid C_{60} FETs. Further information will be provided in Sections 15.5 and 15.6.

15.1.4 Bell Laboratory Misconduct

In 2001, a research group in the Bell Laboratory had reported about the high-temperature superconductivity in the solid C_{60} FETs, and consequently an intensive study of C_{60} FET superconductivity has been performed by many research groups. However, none of the groups could show similar high-temperature superconductivity such as presented in the Bell Laboratory's presentation. This story is the beginning of infamous misconduct by Bell Laboratory. The investigation team for misconduct was immediately formed and they concluded that all reports for the solid C_{60} FET superconductivity from the Bell Laboratory were false [2]. One of the causes of the misconduct may be due to the severe competitive circumstances in the U.S. research society based on the market-driven principle in the research field. From a scientific viewpoint, we have not succeeded in forming

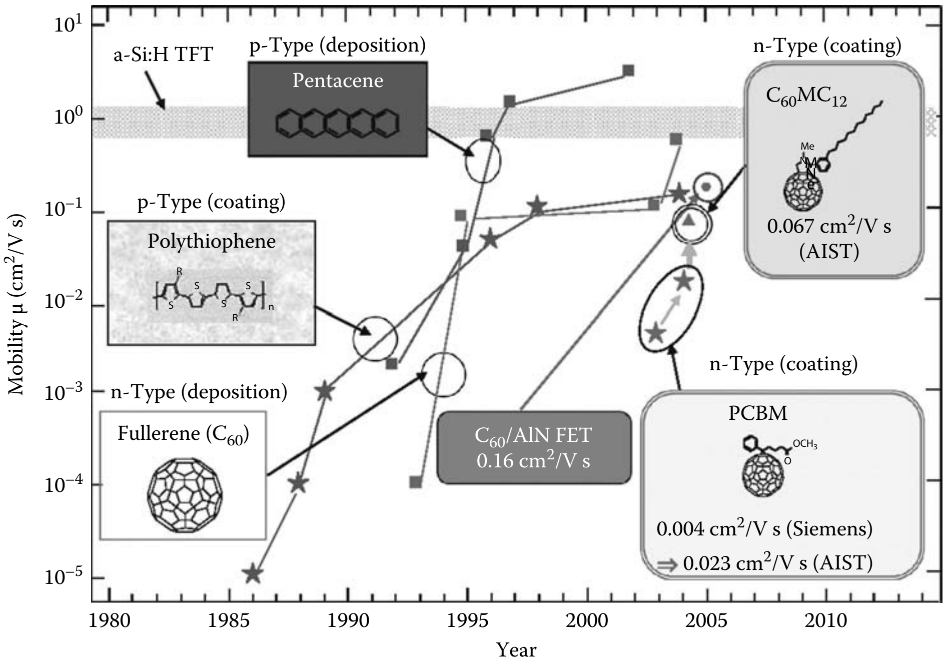

FIGURE 15.1 Improvement of mobility in OFET research field.

an ideal interface between the solid C_{60} and the insulator layers in the FET structure until now. In that sense, the superconductivity problem in the solid C_{60} FET remains until now an open question, in spite of the Bell Laboratory misconduct.

15.1.5 Interfaces between Solid C_{60} and Insulator Films

The interface between the solid C_{60} layer and the insulator layer such as SiO_2 is formed mainly by two different ways. The first way is a conventional FET fabrication process technique for C_{60} bulk. The second is an epitaxy of the solid C_{60} layer on a patterned insulator layer with electrode metals. In both ways, there are several problems in forming an ideal interface. Namely, there are damage problems accompanied with the formation process of the insulator layer on the C_{60} bulk surface in the first way, and lattice mismatch problems for the solid C_{60} epitaxy in the second way. The interface problem is very essential for electron transport; therefore, the characteristics of the C_{60} FETs such as I–V characteristics are strongly dependent on the interfacial states. It is well known that the I–V characteristics depend on adsorption molecules such as OH^- ions during the I–V measurements under the atmosphere. The conventional sputtering method to form the insulator layers on the solid C_{60} surface easily results in physical plasma damages from ions and/or the sputtered elements from the target during the deposition of the layers. Therefore, we can only obtain a rather disordered interface like the interface found between the solid C_{60} and the SiO_2 layers. On the other hand, even in the second way, we can also only obtain the amorphous C_{60} interface when the solid C_{60} layers are grown on the amorphous surfaces like SiO_2. The author's research group has proposed a new insulator layer made of epitaxial single crystalline AlN grown on the SiC and/or the Si substrates for single crystalline solid C_{60} growth. In principle, we can obtain a single crystal epitaxial layer of the solid C_{60} on the single crystalline AlN insulator layer. We will discuss about this new structure and the FET characteristics in the later sections. In these sections, we will see that we can obtain a higher performance C_{60} FET in comparison with the FET structures consisting of conventional amorphous solid C_{60}/SiO_2. Further information will be found in Sections 15.5 through 15.7.

15.1.6 Field Effect Doping

It is well known that chemical doping in solid C_{60} is not as controllable as with conventional semiconductors such as Si. This issue is the common problem for all the organic materials, including solid C_{60}. The key point of the problem is that almost all of organic materials behave as ideal intrinsic semiconductors with low conductivity and without mid-gap states. The carriers come from the source and the drain electrodes into the channel layer induced by an applied gate voltage. The carrier concentrations of the channel layer can be varied by controlling the supply of the carriers from the source and the drain electrodes through the modulation of the applied gate bias voltages. Certainly, in high-performance OFET, the am-bipolar transistor operation has been observed as direct proof of field effect doping, rather than by conventional chemical doping. Namely, this doping technique is completely different to chemical doping in conventional FET, in which the carriers come from bulk conduction bands due to the band bending at the interface, except for the similarity of the device structures. An important advantage of the field effect doping technique is am-bipolar carrier injections by controlling the polarity of the gate bias voltage. The operation polarity is determined by the combination of metals and organic materials. For example, when gold, Au, is used as an electrode, the channel layer of the C_{60} FET shows n-type conduction and it is said that the solid C_{60} is n-type material.

15.1.7 Operation Principles of C_{60} FETs

When the channel layer is formed by applying the gate bias voltage, the modulation of the gate bias voltage produces modulation of the channel conductance between the source and the drain. The typical characteristics of the drain current and the drain-source voltage (I_D–V_{SD} characteristics) with some gate voltage (as shown in Figure 15.2 as a schematic illustration) consist of three parts: the sub-threshold, the linear, and the saturation regions. In the sub-threshold region, the drain current gradually increases with the drain-source voltage due to the hopping transport mechanism. In the linear region, the uniform channel layer is formed by the applied gate voltages. Therefore, the drain current is proportional to the drain-source voltage just before the "pinch-off" voltage, according to the conventional FET theory. The channel layer just disappears due to a reverse bias voltage near the drain electrode at the pinch-off voltage. After the formation of depletion regions near the drain electrode at the pinch-off voltage, the external excess applied voltage of the drain-source electrodes is used as an extension of the depletion region in the channel layer. Therefore, in this region, the drain current saturates as a function of the applied drain-source voltage. These pictures are commonly observed for conventional semiconductor FET performances, except for the sub-threshold regions. However, in principle, the solid C_{60} FET, and in general, also high-performance OFETs are expected to result in the am-bipolar operation, which is completely different from the

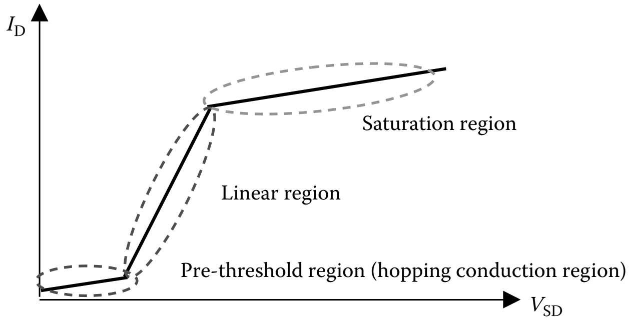

FIGURE 15.2 Schematic illustration of typical I–V characteristics of OFET.

conventional semiconductor FET, because of the difference in the origin of the induced channel carriers. In solid C_{60} FETs, we can inject both the electrons and the holes into the channel layers. In principle, we can control the *p*- and/or the *n*-channels of FETs by using the polarity of the external applied gate voltage.

15.1.8 Superconductivity

In the future, it is expected that the C_{60} FET will possibly have superconductivity with a considerably high critical temperature, in spite of the Bell Laboratory misconduct. If we can realize the ideal interface between the high-quality intrinsic solid C_{60} layer and the insulator layer, then the am-bipolar operation will be observed with very high mobility. This will be possible, though the present stage for C_{60} FETs is very far from achieving the above-ideal situation even with the highest performances. Therefore, the superconductivity of the C_{60} FET still now remains an open question. It is a very interesting issue, especially related to the holes' injection in the FET structure. We have already proposed a new type of C_{60}/single crystalline AlN FET structures on SiC and/or Si substrates for investigating the possibility of superconductivity with a considerably high critical temperature. However, so far, the preliminary C_{60} FET performance is not at all so high.

15.1.9 Guides for This Chapter

In this chapter, we discuss in detail the basic and the advanced topics about C_{60} FETs. In Section 15.2, we describe the fundamental physical properties of C_{60} molecules, including a brief historical review, a large-scale production synthesis method, an electronic structure of C_{60} molecules, and the significant aspects of C_{60} FETs. In Section 15.3, we describe the various physical properties of solid C_{60} crystals. In Section 15.4, we describe the van der Waals epitaxy of the solid C_{60} layer to be used in a realistic application of the FET structures. In Section 15.5, the basic concept of organic FETs, including the C_{60} FET, will be described in order to understand the C_{60} FET operation. In Section 15.6, the realistic fabrication process of the C_{60} FET is discussed in detail. In Section 15.7, the characterizations of the C_{60} FET performance are described and we discuss about a strategy for the realization of high-performance C_{60} FETs. Finally, we summarize this article in Section 15.8.

15.2 Fullerene (C_{60})

In this section, we describe the minimal knowledge about fullerene, in particular, the C_{60} molecule, to help in understanding the following sections. For those who want to know more details than presented in every section, look in the review and textbook references at the end of this chapter.

15.2.1 Discovery of Fullerene Family

The discovery of fullerene certainly gave a heavy shock to the scientific society, especially in carbon chemistry. In 1985, Curl and coworkers discovered fullerene, the C_{60} molecules, at the Sussex and the Rice Universities [3]. It is well known that the name of "fullerene" is named after Richard Buckminster Fuller who is one of the representative modern architects in the United States. The existence of the C_{60} molecule had already theoretically been predicted as a "corannulene" molecule by Eiji Osawa in 1970. He noticed that the structure of a corannulene molecule was a subset of a soccer-ball shape and, therefore, he predicted that a full ball shape could also exist. Unfortunately, he reported his idea only in a Japanese journal, but news of his discovery did not reach Europe and America. Before the discovery of the fullerene, the only known stable structures of the carbon system were diamond, graphite, and amorphous carbon. Therefore, most people thought that it would be difficult for a new structure of the carbon system to exist, because of the long research history of carbon systems by the chemistry. However, in 1985, Harold Kroto, who was studying the origin of interstellar molecules, was interested in interstellar molecules that have curious absorption and emission spectra. He thought there was a possibility that some carbon systems, produced in the supernova process, were strongly related with the origin of the emission spectra. Initially, Kroto did not think about the C_{60} molecular structure at the beginning of the 1980s, but he thought about some carbon molecules with strange chain structures. In 1984, Kroto visited the Richard Smalley Lab in Rice University to discuss a new proposal for "a laser-vaporization cluster beam mass spectroscopy" (see Figure 15.3) in which a similar condition to the supernova process for the carbon cluster formation seems to have been generated [4]. The experimental results were unexpected. The mass spectra for the laser-ablated carbon systems showed strong intensity in the C_{60} molecule [5–7]. However, Kroto, Smalley,

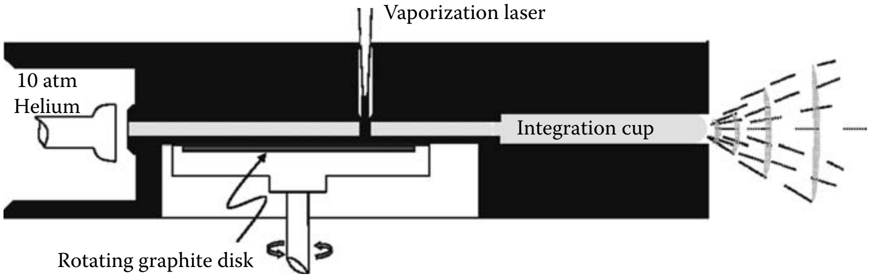

FIGURE 15.3 Laser-vaporization cluster beam mass spectroscopy.

and Robert Curl could not explain at all the stable structure of the C_{60} molecule when they first observed the strong mass spectra from it. It is said that they had a sudden inspiration about the stable structure of the C_{60} molecule from a geodesic dome designed by a famous modern American architect, Buckminster Fuller. The geodesic dome has a mechanically stable structure due to the proper dispersion of the internal force of the dome. They reported their ideas about the stable structure of the C_{60} in an academic paper with several other scientific data. However, many carbon researchers did not pay so much attention to the paper. The researchers needed direct proof such as optical spectroscopic data of the C_{60} for the discussions about the structure of C_{60} molecule to be taken seriously. However, the amounts of C_{60} molecules produced by the laser-vaporization cluster beam mass spectroscopy were not enough at all for an optical spectroscopic measurement in spite of the high generation rate of the C_{60} molecules. It was difficult to construct an apparatus large enough for the laser-vapor evaporation method of the large-scale production of the C_{60} molecules to occur. Therefore, the fullerene proposal by Kroto et al. had to wait until large-scale production methods for the fullerene had improved.

15.2.2 Synthesis of C_{60} Molecules

In 1990, the American and German cooperative research groups found a new "gas evaporation method" for the large-scale production of fullerene [8]. In the new method, a highly purified carbon rod was directly heated by an electric current under the He gas ambient of 100 Torr. The yield of C_{60} molecules by using this method was about 10%. Moreover, Smalley's group also developed another large-scale production method named the "contact arc method," as shown in Figure 15.4, in which the maximum yield of C_{60} molecules reached at about 13% [9]. Some other large-scale production methods such as the RF-heating method [10] and the gas burning method [11] have been developed by now. They also have been used for the commercial large-scale production of C_{60} molecules.

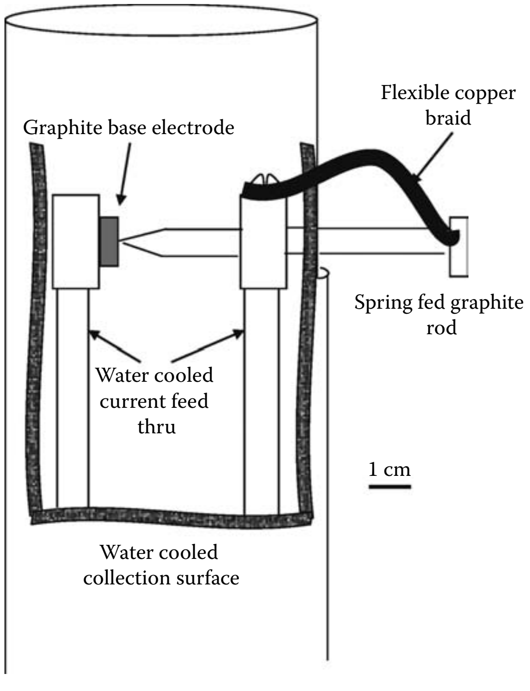

FIGURE 15.4 Contact arc method.

Typical carbon soot produced by the contact arc method, for example, usually contains 0%–15% fullerene, in which C_{60} molecules consist of 70%–85% of the weight, 10%–15% of the weight is C_{70}, and other higher-order residual fullerene molecules. Krätschmer et al. discovered abstraction methods for obtaining C_{60} molecules from carbon soot produced by large-scale production methods like the gas evaporation method. They found out that fullerene such as C_{60} molecules can be easily dissolved into an organic solvent, while the rest of the carbon cannot be dissolved. They also observed that the sublimation temperatures of C_{60} and C_{70} molecules in UHV are different from each other: 300°C and 350°C for the C_{60} and C_{70} molecules, respectively [8]. The high purity abstraction rate of more than 99.97% of the C_{60} molecules from the soot was achieved by this molecular distillation method using the sublimation apparatus with a constant temperature gradient [12]. Furthermore, very recently, it has been reported that a purity of more than 99.99999% classes has been achieved by a N_2 atmospheric train-sublimation method by Laudise et al. (see Figure 15.5) [13].

On the other hand, liquid chromatography such as open-column chromatography for C_{60} molecules [14] and high-performance liquid chromatography (HPLC) [15] are also very popular and useful methods for the abstraction of C_{60} and higher-order fullerene like the C_{70} molecules. However, in general, the purity of the fullerene obtained by these methods cannot be as high compared with the molecular distillation method because of residues of the organic solvent.

In any case, the C_{60} molecular powder with a purity of 99.98% can be easily bought commercially. For the van der Waals epitaxy of solid C_{60} for the FETs, in general, the C_{60} molecular powder with the purity more than 99.98% has been used.

15.2.3 Properties of C_{60} Molecule

The hypothesis for the structure of the Buckminster fullerene C_{60} molecule proposed by Kroto and Smalley in 1985 was researched for the large-scale production of C_{60} samples using various types of spectroscopy. In particular, there were several important theoretical predictions about the lattice-vibration spectrum, for example, for the infrared absorption (IR) spectrum [16–19]. Theoretical investigations predicted that only the first four-order vibration modes in the IR spectrum were active among 174 degrees of freedom in the lattice vibrations of the C_{60} molecules, when the structure was a soccer-ball shape. Indeed, Krätchmer and Huffman used the four IR modes in the IR spectra to identify the C_{60} molecules in the carbon soot produced by their gas evaporation method, as mentioned in their 1990 academic paper on the large-scale production of fullerene.

According to their theoretical predictions, the C_{60} molecules with I_h symmetry have 46 vibration modes, including many

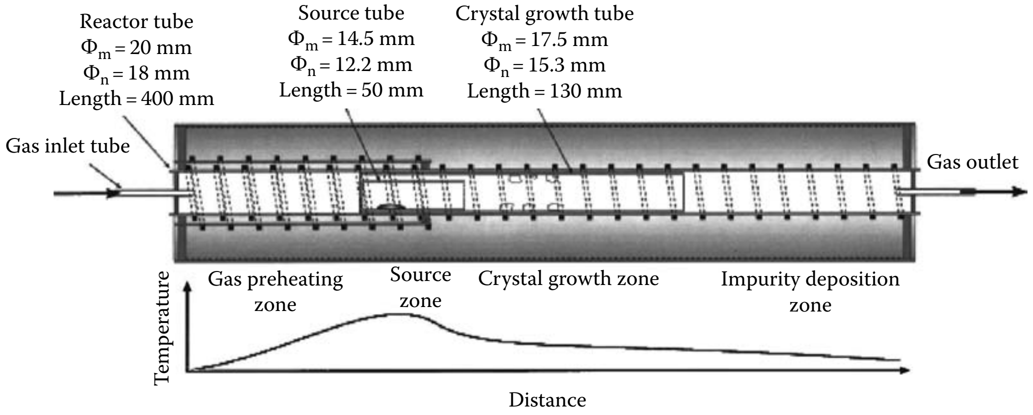

FIGURE 15.5 N_2 atmospheric train-sublimation method.

degenerated modes with 174 degrees of freedom in lattice vibrations. The 46 inner-molecule lattice-vibration modes of the C_{60} molecule were classified by the I_h symmetry as follows [20]:

$$\Gamma(C_{60}) = 2A_g + 3F_{1g} + 4F_{2g} + 6G_g + 8H_g + A_u + 4F_{1u} + 5F_{2u} + 6G_u + 7H_u \qquad (15.1)$$

where the $4F_{1u}$ modes are only allowed for the IR active mode. On the other hand, the 10 modes of the $2A_g$ and the $8H_g$ modes are also allowed as Raman active modes. The IR (A) and the Raman (B) spectra from the solid C_{60} layers deposited on the Si substrate are shown in Figure 15.6. The 4 IR and the 10 Raman peaks in the spectra are clearly observed [21].

Other strong experimental data from the soccer-ball shape of C_{60} molecules can be obtained by ^{13}C nuclear magnetic resonance (^{13}C-NMR) measurements. Indeed, the Kroto research group at Sussex University and the Bethune research group at IBM (Almaden) obtained only one NMR signal due to the high I_h symmetry in the ^{13}C-NMR measurements [14,22]. Moreover, the x-ray diffraction for the $C_{60}(OsO_4)$ single crystal [23] and the neutron diffraction for the low-temperature solid C_{60} crystal succeeded in obtaining more direct proof of the soccer-ball shape structure of C_{60} molecules after stopping of the high-speed rotation of the C_{60} molecules by using various scientific techniques [24].

Buckminster fullerene consists of 12 pentagons and 20 hexagons that is a geometrical stable combination due to Euler's theorem as follows:

$$V + F = E + 2 \qquad (15.2)$$

where V, F, and E are the numbers of the vertex, the face, and the side of a polyhedron, respectively. Euler's theorem is very helpful for the discussion about the geometrical stability of the higher-order fullerene structures. However, because the chemical stability is not equal to the geometrical one, we have to take into account the electronic structure of Buckminster fullerene. In the simplest calculation using the Hückel approximation, we can estimate the electron energy levels in the C_{60} molecule. When the electron pairs, in which each electron has the opposite spin from each other, occupy the lowest energy level and the electron pairs finally form a closed electron shell, the Buckminster fullerene becomes chemically stable, in addition to being geometrically stable. This is known as the "Hückel rule," and it is used for describing the chemical stability of the fullerene family.

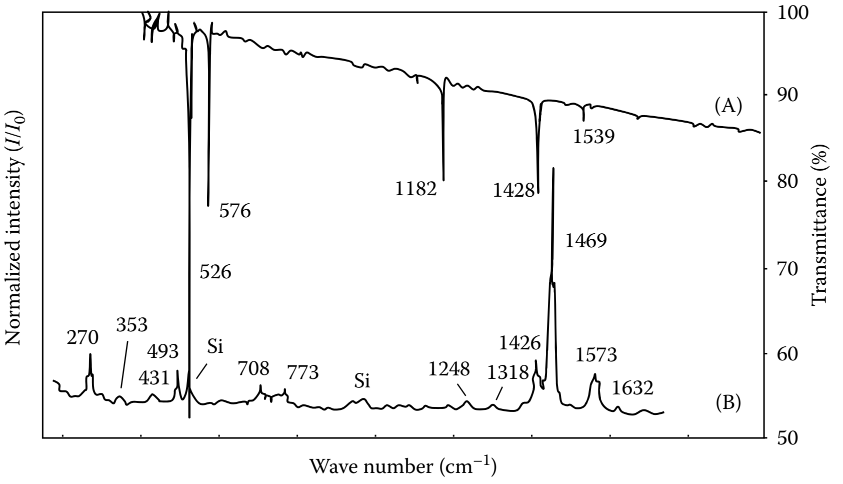

FIGURE 15.6 Infrared (A) and Raman (B) spectra from solid C_{60} layers deposited on Si substrate.

The electronic structure of the C_{60} molecule is very unique because it has many degenerate levels due to high I_h symmetry. In the Hückel approximation, it is estimated that the HOMO level, h_u, is five-folded degenerated and the LUMO level, t_{1u}, is three-folded, and that the energy gap between the HOMO and the LUMO levels is about 1.5 eV. Therefore, it is assumed that the solid C_{60} crystal becomes a semiconductor [25–28]. A scanning tunneling microprobe (STM) observation under the UHV showed that the C_{60} molecules on a well-defined solid surface like Cu (111) or Si (111) 7 × 7 do not spin due to the interaction between the surface and the C_{60} molecules induced by a charge transfer from surface to the C_{60} molecules. Many other interesting properties of the C_{60} molecules are referred to another chapter in this handbook series.

15.3 Solid C_{60}

A solid C_{60} single crystal without any residual organic solvent can be obtained by using the sublimation method or a vacuum evaporation method like MBE, while solid C_{60} crystals grown by solution growth can easily include residual solvent.

15.3.1 Crystal Structure

The structure of the solid C_{60} bulk crystal grown by using the sublimation method has two kinds of structural phases. They are a high-temperature and a low-temperature phase above and below 260 K, respectively [29]. In the high-temperature phase above 260 K, the structure of the solid C_{60} is a fcc lattice with ($Fm\bar{3}m$) symmetry in the space group and the lattice constant is 1.4145 nm. The four C_{60} molecules in the unit cell of the solid C_{60} crystal look like balls with a diameter of 0.71 nm due to the high-speed spin rotation in the x-ray diffraction measurements. Each C_{60} molecule in the solid C_{60} crystal rotates very fast continuously with a rotation correlation time on the order of 10 ps, estimated from the line width of the ^{13}C-NMR spectra [30]. The rotational diffusion property of the C_{60} molecules in the solid C_{60} bulk was shown estimating the neutron inelastic scattering and the rotation potential of the C_{60} molecules that have many local minima with the thermal activation energy of 35 meV [31]. Therefore, the solid C_{60} crystal is called "the plastic crystal" with a rotational diffusion of C_{60} molecules, in which the degrees of freedom of molecular rotation are softened before the translational displacement.

15.3.2 van der Waals Interaction of C_{60}

The lattice vibration of solid C_{60} consists of two kinds of vibration modes, an inner-molecular and an intermolecular vibration. The inner-molecular vibration mode is due to the vibration in the degree of freedom in the C_{60} inner molecule. Therefore, the spectrum of the inner-molecule vibration is very similar to the isolated C_{60} molecule, described in Section 15.2. On the other hand, the intermolecular vibration modes are the proper vibration modes of solid C_{60}, and there are three kinds of modes: acoustic, optical, and libration modes. The phonon frequency of all modes (0–50 cm^{-1}) is lower than the vibration modes in conventional covalent crystal modes such as Si (0–530 cm^{-1}), because of the large mass of the fullerene molecules and the weak van der Waals interaction between the C_{60} molecules. The innermolecular vibration modes have a considerable higher frequency than the intermolecular vibration modes [32].

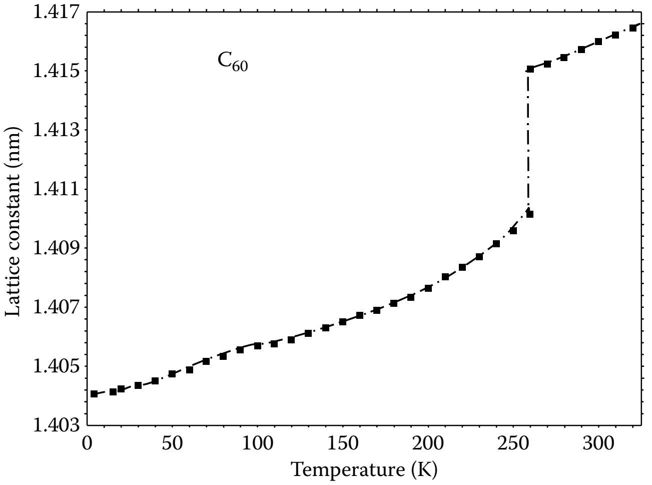

FIGURE 15.7 Lattice constant of solid C_{60} as a function of temperature.

15.3.3 Structural Phase Transition

The crystal structure of solid C_{60} suddenly changes from a fcc structure to a simple cubic (sc) structure when below 270 K in the first-order phase transition, because of the discontinuity in the variations of the lattice constant, as shown in Figure 15.7 [33].

15.3.4 Electronic Structures

The electronic structure of solid C_{60} has been extensively studied by many groups. They have already revealed certain aspects of its semiconductor nature, for example, the energy band gap of about 1.5 eV. Figure 15.8 shows the band structure of the fcc C_{60} crystal calculated by the density functional method, accompanied with π electron levels from an isolated C_{60} molecule calculated by the Hückel method, as presented by Saito and Oshiyama [28]. The electronic structure has a rather small dispersion and a narrow bandwidth (0.4 eV). They are just the origin of the band gap of fcc solid C_{60}. The solid C_{60} is a direct band-gap semiconductor in which the gap is the difference between the bottom of the LUMO and the top of the HOMO at *X* point.

The photo-emission spectroscopy (PES) and the inversion photo-emission spectroscopy (IPES) are very useful in directly observing the electronic structures of solid C_{60}. The HOMO–LUMO gap (3.7 eV) estimated from the PES and IPES spectra is larger than the gap estimated by the other techniques such as the absorption measurements. The middle energy points of the HOMO and the LUMO bands are above 2.25 eV and below 1.15 eV from the Fermi level, respectively. It is said that the large band gap in the PES and the IPES is due to the electron correlation effect [34].

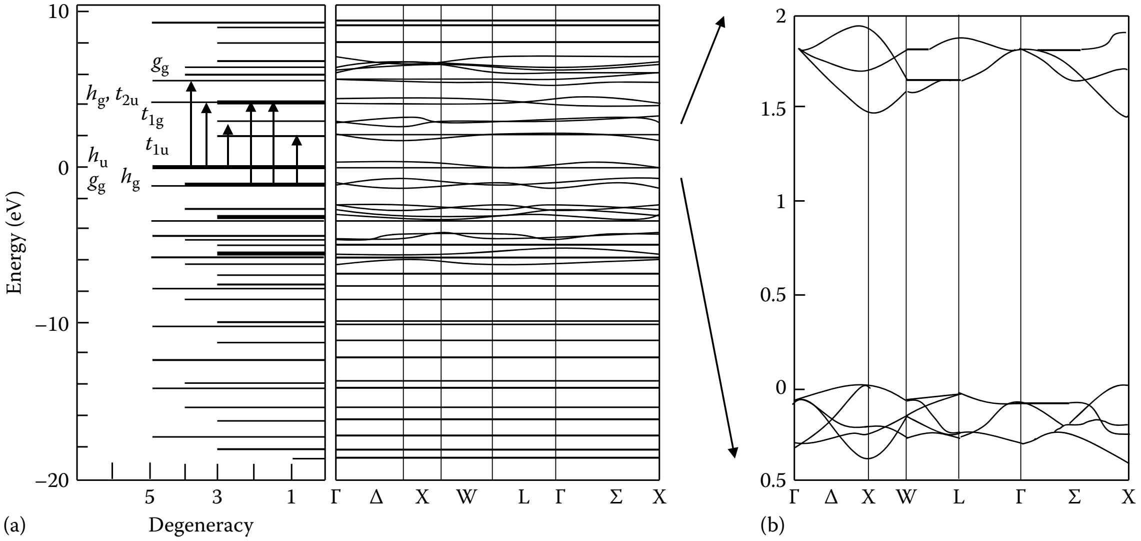

FIGURE 15.8 Band structure of fcc C_{60} crystal calculated by density functional theory.

The HOMO and the HOMO-1 bands, which come from the h_u and the h_g-g_g, respectively, are π electron-like states. These bands are 5- and the 9-degenerated electron states. On the other hand, the empty states consist of the LUMO and the LUMO+1 bands that originate from the t_u and the t_g. They form 3-degenerated states in the empty bands.

The electrical conductivity σ of solid C_{60} is extremely low, that is, from 10^{-8} to 10^{-14} $(\Omega\ \text{cm})^{-1}$. Therefore, solid C_{60} is like an insulator [35,36]. The fluctuation in electrical conductivity is due to impurities and defects. Furthermore, there are difficulties in measuring high-resistance materials. The relative dielectric constant of solid C_{60} is 4.4 ± 0.2 [37]. The relative deviation of the dielectric constant (Δε/ε) is due slightly to the temperature with a small discontinuity at 260 K. This variation is due to the phase transition from the fcc to the sc structures and a broad step-like dependence around 165 K in the low-temperature phase which occurs due to a disorder in the molecule's axis in the pentagon and hexagon directions [38].

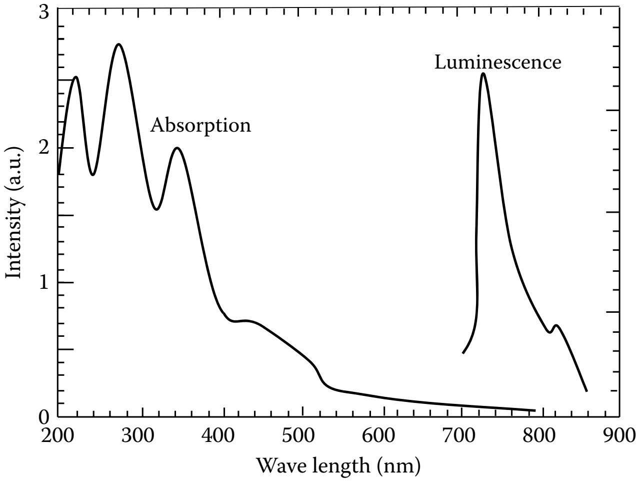

FIGURE 15.9 Absorption and luminescence spectra for solid C_{60} deposited on a quartz substrate.

15.3.5 Optical Properties

The absorption and luminescence spectra for solid C_{60} deposited on a quartz substrate are shown in Figure 15.9 [39]. Three strong absorption peaks are observed at 221, 271, and 347 nm. These peaks come from the π–π* transition of the C_{60} molecules, because the peaks are also observed from C_{60} molecules dispersed in an organic solvent. The broad and weak bands, observed at around 600 nm, come from the forbidden transition between the t_{1u} and h_u levels. In contrast to these absorption peaks, the rather strong band at 430–520 nm comes from the proper absorption of solid C_{60}. A broad photoluminescence spectrum with peak energy at 737 nm is observed above 10 K, as shown in Figure 15.9b. However, the photoluminescence spectrum at 1.2 K, as shown in Figure 15.10, shows two luminescence bands excited by

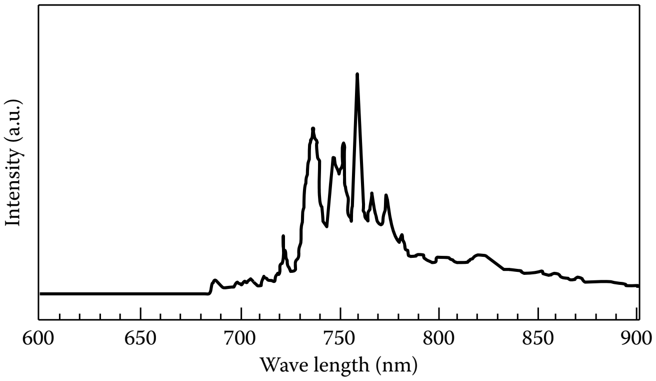

FIGURE 15.10 Photoluminescence spectrum from solid C_{60} at 1.2 K.

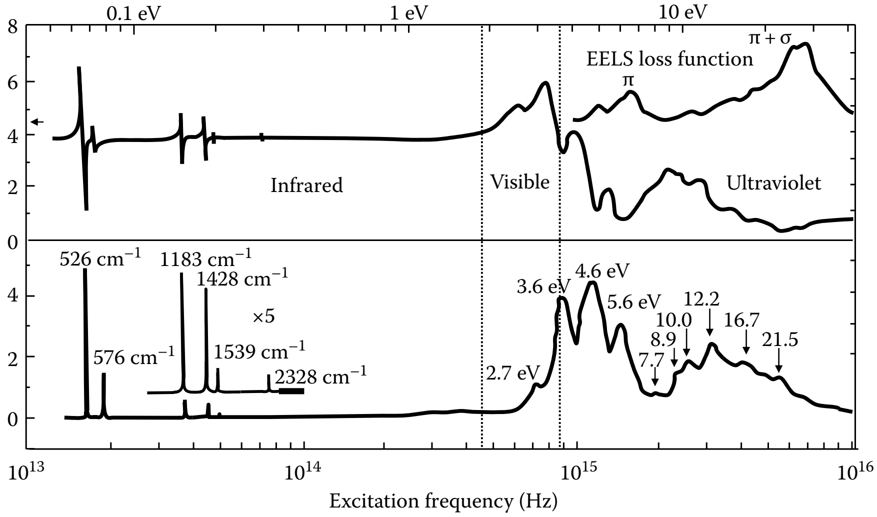

FIGURE 15.11 Optical measurements results from infrared to ultraviolet wavelengths for $\varepsilon_1(\omega)$ and/or the $\varepsilon_2(\omega)$.

734 and 684 nm. The 734 and the 684 nm excited bands consist of the electronic transition and a vibronic transition [40].

The dielectric function, $\varepsilon(\omega)$, consists of the real part $\varepsilon_1(\omega)$ and the imaginary part $\varepsilon_2(\omega)$, as follows:

$$\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega) \tag{15.3}$$

The complex refractive index, $n_c(\omega)$, is the root of the dielectric function:

$$n_c(\omega)^2 = \varepsilon(\omega) \tag{15.4}$$

in which $n_c(\omega)$ consists of the real part, the refractive index $n(\omega)$, and the imaginary part, the extinction coefficient $\kappa(\omega)$. The $\varepsilon_1(\omega)$ and the $\varepsilon_2(\omega)$ are connected with the $n(\omega)$ and the $\kappa(\omega)$, as follows:

$$\begin{aligned} \varepsilon_1(\omega) &= n(\omega)^2 - \kappa(\omega)^2 \\ \varepsilon_2(\omega) &= 2n(\omega)\kappa(\omega) \end{aligned} \tag{15.5}$$

These optical constants can be obtained from the absorption and the reflection spectra. The dielectric function is the most important and basic physical quantity concerning the optical properties of the solid C_{60} layers. The typical results of the optical measurements results from the infrared to the ultraviolet wavelengths for the $\varepsilon_1(\omega)$ and/or the $\varepsilon_2(\omega)$ are shown in Figure 15.11 [41]. In the measurements, the solid 100 nm thick C_{60} layers were deposited on proper substrates such as Si and glass by using the vacuum evaporation method. Many coupled modes were observed in the thick solid C_{60} spectra in addition to inner-molecular F_{1u} vibration modes (526, 576, 1183, and 1428 cm^{-1}) and their two coupled modes (1539 and 2328 cm^{-1}), roughly below 0.3 eV. A solid C_{60} film is also assumed to be a nonlinear optical system due to the π-electrons in the molecules. Third-harmonic generations were reported by Kajzar. The third nonlinear susceptibility as a function of the wavelength is shown in Figure 15.12 [42].

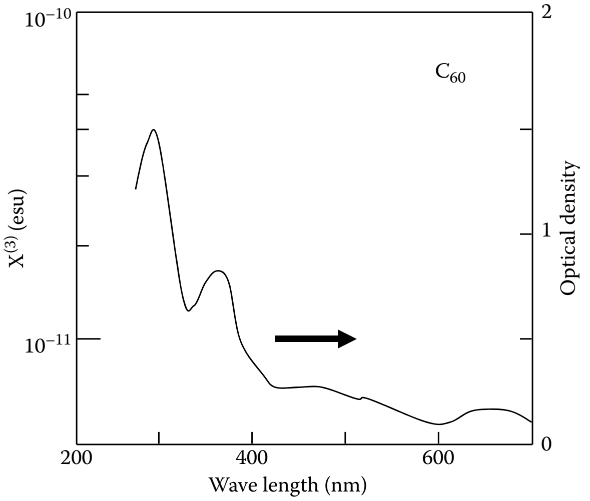

FIGURE 15.12 Third nonlinear susceptibility as a function of the wavelength.

The photo-conducting spectra of the solid C_{60} deposited on glass substrates are shown in Figure 15.13 [39]. The photo-conducting spectrum from the thin C_{60} layer just corresponds to the absorption spectrum. However, the behavior of the photoconductivity is out of phase with the absorption spectrum, as shown in Figure 15.13. The photo-conducting spectra show that the photo-active region of solid C_{60} is only limited to around the thin surface region and, therefore, the diffusion length of the photo-produced electrons and/or holes is very short. The electron-hole pairs can be produced within the thin small region, because the photons cannot penetrate deeply into the solid C_{60} layers. There is a decrease in one order of magnitude of the photoconductance due to oxygen exposure, because the oxygen molecules absorbed into the solid C_{60} layer work as recombination centers for the carriers, decreasing their lifetime and mobility. The absorption of the oxygen molecules into the solid C_{60} layers

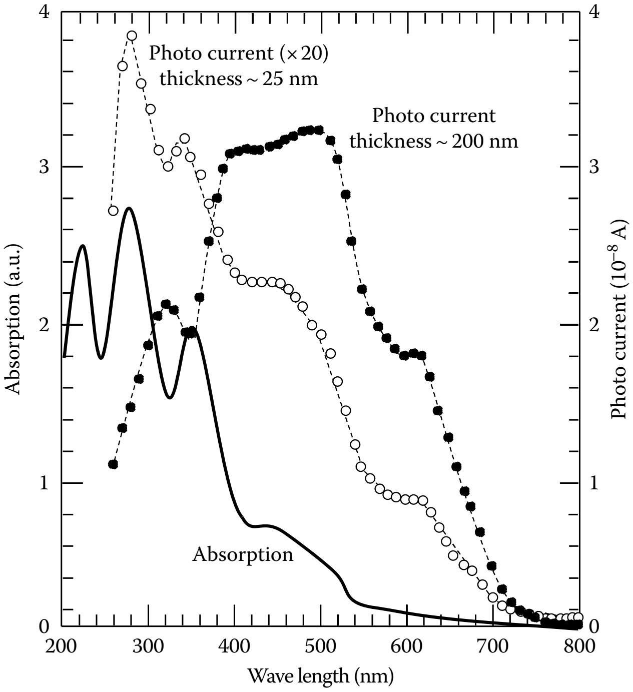

FIGURE 15.13 Photo-conducting spectra of solid C_{60} deposited on glass substrates.

is a reversible process. In addition, photo-conductance recovering accompanied with oxygen adsorption is observed by the 180°C temperature mark annealing under a vacuum. However, a large irreversible deterioration of photo-conductance is induced by white light irradiation exposure to oxygen in the solid C_{60} layers, because of a chemical reaction between the oxygen atoms and the C_{60} molecules induced by white light irradiation.

The polymerization of solid C_{60} is induced by ultraviolet strong light irradiation and/or a low-energy (3–1500 eV) electron beam. We can check the polymerization in several ways; for example, by using the Raman frequency shift. The polymerization by ultraviolet light and the low-energy electron beam is due to an excitation of the electrons in the double bonds of the hexagons and the following reconstruction of new bonds structures between the C_{60} molecules [21,43]. However, the oxygen contamination in solid C_{60} layers has a suppression effect in the polymerization process, because the oxygen atoms work as the killer centers of the excited electrons before the formation of new bonds. Furthermore, the transition from dimmer to monomers is also induced by a thermal annealing at 200°C under an UHV environment. The activation energy of this process is estimated at about 1.25 eV [44].

15.4 van der Waals Epitaxy of C_{60}

The epitaxial growth of solid C_{60} on several proper substrates offers very attractive potential due to its new electrical and optical properties, provided we can obtain the ideal interface between solid C_{60} and the insulator layers. The intermolecular distance on the fcc solid C_{60} film is about 1 nm. This is several times longer than the inter-atomic distance on the top surface of semiconductor substrates such as GaN. The atomic arrangement of the GaN (0001) surface has a sixfold symmetry similar to a fcc solid C_{60} (111) surface. The inter-atomic distance GaN (0001) is about 0.319 nm, which is nearly one-third of the intermolecular distance of the C_{60} molecules in a fcc solid C_{60} (111) surface. If we define the lattice-mismatch as the difference between the intermolecular distances on the C_{60} (111) surface [45] and the distance is almost equal to three times the inter-atomic one on the GaN (0001) surface, then the lattice-mismatch becomes about 4.5% which is not so serious of a difference in the lattice-mismatch of solid C_{60} epitaxial growth. Additionally, it is well known that nitride is rather stable surface for chemical reactions. Therefore, this surface is assumed to be a chemically inactive plane. This plane is necessary for the epitaxy of a van der Waals molecular crystal without any covalent bonds to substrates [46]. It is equally important in the formation of a van der Waals interface in layered materials such as Mica [47] and MoS_2 [48]. Consequently, the solid C_{60} epitaxial growth on the proper substrates can be expected to appear under proper growth conditions. We call the epitaxy like this (solid C_{60} epitaxial growth without any covalent bonds to the substrates, except for the potential periodicity), "the van der Waals epitaxy." In this section, the basic knowledge about the epitaxial growth of a solid C_{60} single crystal layer is described.

15.4.1 Molecular Beam Epitaxy

Conventional solid source MBE is named after the epitaxial growth using molecular beams such as the C_{60} beam produced by conventional Knudsen thermal cells (K-cells) under an UHV environment below a background pressure of lower than 10^{-7} Pa. The conventional MBE system usually consists of a main chamber for epitaxial growth, a preparation chamber for the surface cleaning of the substrate, and an introduction chamber for the substrate charge from the atmosphere, as shown in Figure 15.14. Each chamber has a pumping system with several vacuum

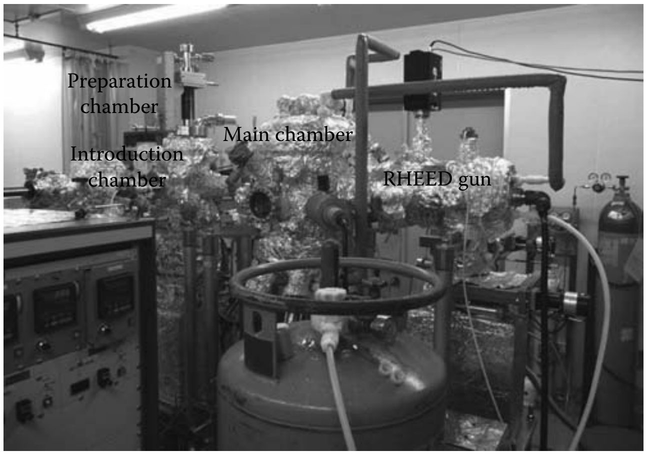

FIGURE 15.14 Conventional MBE system.

monitoring systems such as a nude ion gauge in order to keep the UHV environment, and all chambers are connected via gate valves with each other. The main chamber has a sample holder with a sample heating system and a monitoring system for epitaxy. For example, a reflection of high-energy electron diffraction (RHEED) system and a quadrupole mass spectroscopic (Q-Mass) system are used for monitoring the growth surface and the background ambient during epitaxy, respectively. The main chamber should maintain the UHV environment as much as it is possible in order to suppress contamination during growth, allowing a cooling bag using liquid nitrogen or a water circular system to settle inside the chamber. The preparation chamber has a thermal heating system for sample cleaning and, in many cases, other in-situ characterization systems such as the Auger electron spectroscopic (AES) system in order to characterize the sample surface before and/or after epitaxy. The substrate prepared by the proper chemical cleaning treatment is installed into the introduction chamber from the atmosphere to the high vacuum environment. Then, the substrate is transferred from the introduction chamber to the preparation chamber through the gate valve to clean the surface and, occasionally, to perform some in-situ characterization under a UHV environment before growth. After the surface cleaning, the substrate is transferred into the main chamber for growth. In the main chamber, the K-cell and/or the substrate heating systems should fully do the out-gas procedure before the installation of the substrate. The beam flux also should be estimated by a nude ion gauge monitor before and after growth, if possible. The details of the general information of MBE growth and systems can be obtained from the proper reference books.

15.4.2 van der Waals Epitaxy

van der Waals epitaxy is an important technique in overcoming the restrictions in large lattice-mismatches in the crystal growth. According to Koma, the van der Waals interface can drastically relax the lattice-matched growth conditions, because of a weak interaction between the films and the substrates [46]. There has been much interest in using single crystalline C_{60} thin films grown by van der Waals epitaxy on hydrogen-terminated (H-terminated) Si substrates. These newly functional films have a great potential in many areas due to the discovery of the large-scale production synthesis methods of C_{60} molecules. However, it has still been difficult to obtain single-domain C_{60} thin solid films with the fcc structure on Si substrates, though single-domain C_{60} films can be grown on layered material substrates such as MoS_2 [48], GaSe [49], and Mica [47]. The growth of C_{60} films on H-terminated Si (111) substrates with quasi van der Waals interfaces shows a three-dimensional island growth mode. In addition, the film's grown structure contains a fcc double domain. Therefore, a large fcc single-domain film has not been obtained yet to the best of our knowledge.

The most popular device structures of C_{60} FETs is a bottom contact structure with a SiO_2 insulator layer formed on a conductive Si substrate as a gate electrode, and the details will be described in the later sections. In the bottom contact structure with the SiO_2 insulator layer, the C_{60} layer becomes an amorphous or a polycrystalline structure due to the amorphous surface of the SiO_2 layer. If we want to use the single crystalline solid C_{60} surface, then we have to apply the top contact structure to the single crystalline solid C_{60} bulk. The solid C_{60} bulk can only be obtained on the inside of the furnace wall by methods such as train sublimation. However, the obtained single crystal is usually very small. Although it is not impossible to make C_{60} FETs using such tiny C_{60} bulks, interface problems such as the ion damage always occur in the formation process of the insulator layer when using such a sputtering method. Therefore, in the current state-of-the-art technology, it is very difficult to obtain an ideal interface between solid C_{60} and insulator layer in both FET structures. There is a lot of discussion about the use of the bottom contact structure in the C_{60} FET because of the easy fabrication process, even for amorphous-like interfaces.

15.4.3 Initial Growth Surface

Here, we will discuss the relationship between the van der Waals epitaxy and the initial growth surface. We have already pointed out that the interaction between the C_{60} molecules and the initial surface plays an important role in the epitaxial growth of the solid C_{60} layer. Therefore, we discuss several types of surfaces such as the H-terminated Si (111) surface and/or the single crystalline AlN surface.

15.4.3.1 H-Terminated Si (111) Surfaces

In Figure 15.15, typical atomic force microscope (AFM) image and RHEED pattern of C_{60} epitaxy on an H-terminated Si (111) surface are shown. Petal-like islands with considerable low density are formed on the H-terminated Si (111) surface. The results strongly indicate that the van der Waals interaction between C_{60} molecules and the surface is not so strong and, therefore, that the migration length of the C_{60} molecules on a quasi van der Waals surface like the H-terminated Si (111) surface is considerably longer than conventional epitaxy such as with Si layers. However, we can control the island density and size by using a two-step growth technique, as shown in Figure 15.15. The two-step growth technique is well known as one of the most important techniques in achieving a high-quality single crystalline for high lattice-mismatch hetero-epitaxial systems. According to the two-step growth procedure, the C_{60} layers are grown at 150°C to enhance the surface migration of C_{60} molecules, after the formation of low-temperature buffer layer growth at 30°C successively annealed at 150°C to form a high nucleus density, respectively. The typical beam equivalent pressure of the C_{60} molecular beam (purity; >99.99%) was 5.0×10^{-9} Torr during growth, and the beam pressure corresponds to a growth rate of 0.8 nm/min [50].

A RHEED pattern from the grown layer by using the two-step growth technique showed a fcc single-domain structure, as shown in Figure 15.16. However, the RHEED pattern from the directly grown layer showed a double-domain fcc structure in

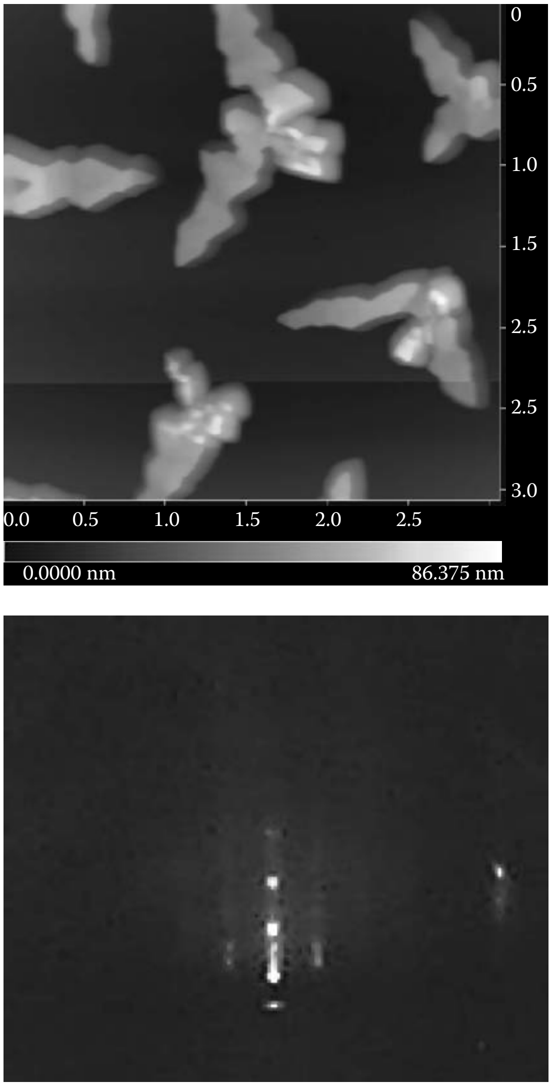

FIGURE 15.15 Typical AFM image and RHEED pattern of C_{60} epitaxy on an H-terminated Si (111) surface.

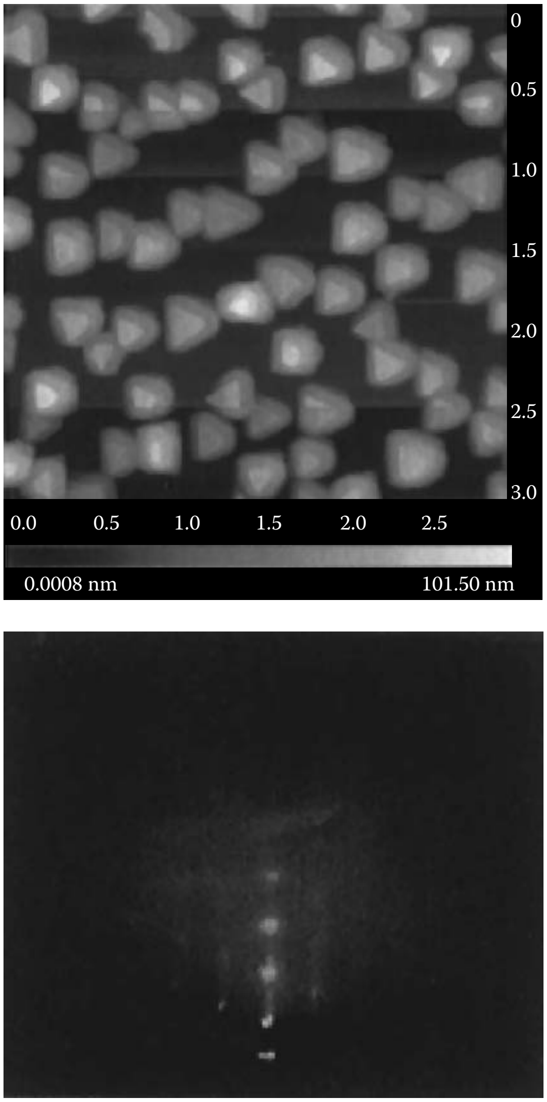

FIGURE 15.16 Typical AFM image and RHEED pattern of C_{60} grown layer by using two-step growth technique.

which the normal and the rotated domains on the Si substrate mixed with each other, as shown in Figure 15.16. According to Saito, in the case of C_{60} epitaxy on NaCl substrates, not only hexagonal (or truncated triangular) plates (which are single crystals) but also particles with irregular shapes which look like petals are observed. Such petal-like particles are rich in planar defects such as parallel and cyclic twining [47]. Therefore, the triangle islands formed in the two-step growth have a single-domain fcc structure. The petal-like islands formed in the direct growth have a double-domain fcc structure. These well-arranged dense islands on the surface obtained by the two-step growth may have a potential to form thick C_{60} films with the fcc single-domain structure.

15.4.3.2 Hexagonal GaN (0001) Surfaces with Ga-Polarity

The epitaxial growth of solid C_{60} on semiconductor substrates offers a very attractive potential due to its newly discovered electrical and optical properties. The atomic arrangement of the hexagonal GaN (0001) surface has a sixfold symmetry similar to the fcc solid C_{60} (111) surface. Its inter-atomic distance is about 0.319 nm, which is near one-third of the intermolecular distance of the C_{60} molecules in fcc solid C_{60} (111) surfaces. An extended lattice-mismatch in the van der Waals epitaxy can be defined as the difference between the intermolecular distance in the C_{60} (111) surface and the distance equal to three times the inter-atomic one in the h-GaN (0001) surface. The extended lattice-mismatch then becomes 4.5% which is the lowest among conventional semiconductor substrates in solid C_{60} epitaxial growth. Additionally, it is well known that the h-GaN surface is rather stable for chemical reactions. Therefore, the surface is expected to act as a chemically inactive plane which is necessary for van der Waals epitaxy as well as the van der Waals interface in layered materials such as Mica and MoS_2 [51].

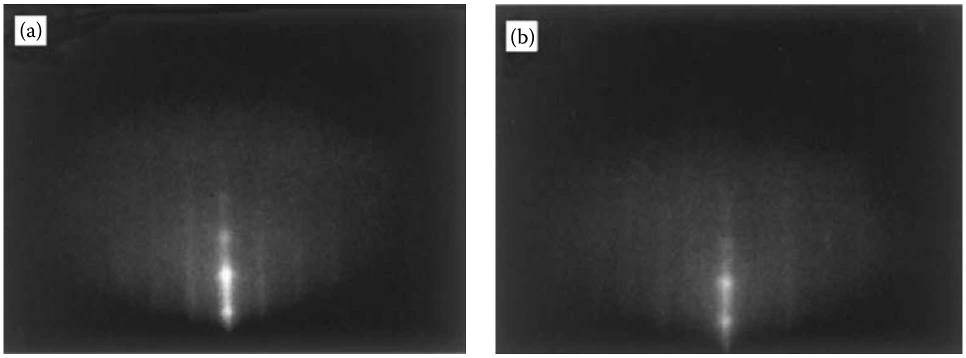

FIGURE 15.17 Typical RHEED patterns from solid C_{60} layer grown on flat h-GaN (0001) surface at 100°C for 60 min.

The GaN (0001) surfaces formed on the $\alpha\text{-}Al_2O_3$ substrates are prepared by a metal organic vapor phase epitaxy. Typical RHEED patterns from a solid C_{60} layer grown on a flat h-GaN (0001) surface at 100°C for 60 min are shown in Figure 15.17. The directions of the incident electron beam are parallel to (a) [1120] and (b) $[11\bar{1}0]$ crystal axes of the h-GaN (0001) surface. The in-situ RHEED observation revealed that the C_{60} film growth did not occur at 150°C because of the re-evaporation of the C_{60} molecules from the surface. On the other hand, sharp streaks from the fcc (111) surface can be observed in the RHEED patterns, as shown in Figure 15.17. The results imply that the lattice-structure of the grown layer is mainly the fcc structure, and the relationship of the crystal axes between the epitaxial grown layers and the h-GaN (0001) substrates is C_{60} (111)‖h-GaN (0001), C_{60} $[10\bar{1}]$‖h-GaN $[11\bar{2}0]$, C_{60} $[11\bar{2}]$‖h-GaN $[01\bar{1}0]$, as shown in Figure 15.18. Figure 15.19 shows a typical AFM image of the C_{60} layer on a flat h-GaN (0001) surface.

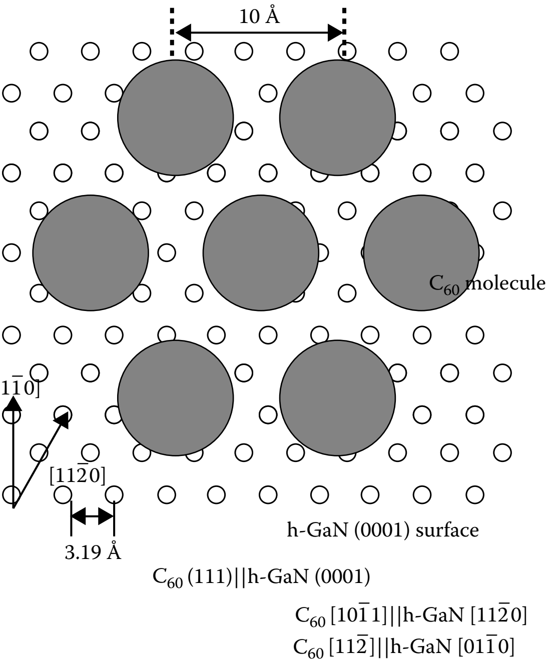
10 Å
C_{60} molecule
$[01\bar{1}0]$
$[11\bar{2}0]$
3.19 Å
h-GaN (0001) surface
C_{60} (111)||h-GaN (0001)
C_{60} $[10\bar{1}1]$||h-GaN $[11\bar{2}0]$
C_{60} $[11\bar{2}]$||h-GaN $[01\bar{1}0]$

FIGURE 15.18 Relationship of crystal axes between epitaxial grown layers and h-GaN (0001) substrates.

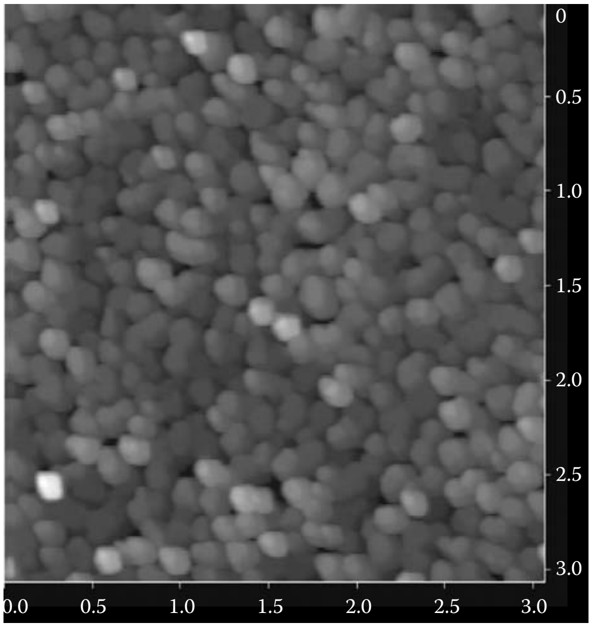

FIGURE 15.19 Typical AFM image of C_{60} layer on flat h-GaN (0001) surface.

The whole surface of the substrate was completely covered with highly dense C_{60} islands. The size of the islands was on the order of a sub-micron meter. Figure 15.20 shows the growth temperature dependency on grain size and density estimated from AFM images, as shown in Figure 15.19, in C_{60} growth on h-GaN and H-terminated Si. In the case of C_{60} growth on the h-GaN, the grain size slowly increases and the grain density rapidly decreases as a function of the growth temperature in comparison with the H-terminated Si substrate. The results indicate that the interaction between the C_{60} molecule and the h-GaN surface is extremely weak compared with C_{60} film growth on a H-terminated Si (111) surface. The evaporation of the C_{60} molecules on the GaN surface may very well easily occur and, consequently, it leads to a reduction of the diffusion length of C_{60} molecules during growth. Short diffusion and/or migration lengths may induce highly dense small grains, as shown in Figure 15.20.

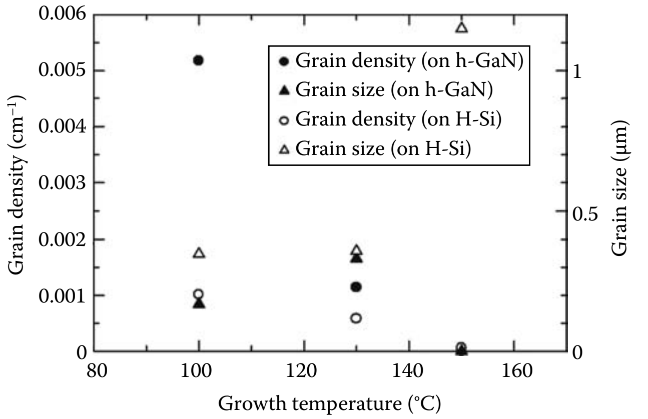

FIGURE 15.20 Growth temperature dependency on grain size and density estimated from AFM images, as shown in Figure 15.19.

The experimental results for C_{60} direct growth on rough h-GaN (0001) surfaces show that a polycrystalline C_{60} layer can only be obtained. Furthermore, it suggests that the surface migration of the C_{60} molecules may be suppressed by the existence of various planes on the rough surface. The fcc C_{60} layers can be grown epitaxially even on a rough h-GaN (0001) surface by using the two-step growth technique, though only the polycrystalline C_{60} layers are obtained by direct growth. All experimental results strongly indicate that the periodicity of the h-GaN surface potential for the van der Waals epitaxy is easily and strongly modified by surface roughness. Therefore, the flatness of GaN substrate is also a very important factor in obtaining a high-quality C_{60} epitaxial layer.

15.4.3.3 Hexagonal AlN (0001) Surface with N-Polarity

The AlN single crystal layer can be grown on 6H-SiC [53,54], Si [55,56], and so on. The single crystal AlN layer has large bandgap energy (6.2 eV) and enough high dielectric breakdown voltage for C_{60} FET applications. Lattice-mismatch between the C_{60} fcc structure (111) and the h-AlN (0001) is 7.3% in C_{60} $[10\bar{1}]\|$AlN $[11\bar{2}0]$ (the 0° orientation) and 7.0% in C_{60} $[11\bar{2}]\|$AlN $[11\bar{2}0]$ (the 30° orientation), as shown in Figure 15.21. These values are acceptable under C_{60} van der Waals epitaxy [45]. Therefore, the AlN layer is assumed to be the substrate for single crystal C_{60} layers. However, the lattice-mismatch in both orientations, shown in Figure 15.21, is almost the same. Because the C_{60} van der Waals epitaxy has a character that the growth orientation on the smooth surface strongly depends on the lattice mismatch [45], it is a little difficult to predict the C_{60} orientations on the AlN surface. If the single crystal C_{60} layers can be grown on a AlN surface, a more perfect interface can be obtained than by an interface formed by the conventional sputtering method.

Figure 15.22 shows typical RHEED patterns from a solid C_{60} layer grown on a smooth AlN surface (RMS value: 0.9 nm) for 2 h at 100°C–160°C. All the RHEED patterns in Figure 15.22 indicate a fcc (111) structure. The ring pattern appears on a spot pattern below a growth temperature of 100°C. Both the RHEED patterns with the azimuth of electron beam paralleled to AlN $[10\bar{1}0]$ and AlN $[11\bar{2}0]$ at 100°C–130°C show the inclusion of C_{60} $[10\bar{1}]$ and C_{60} $[11\bar{2}]$ patterns, as shown in Figure 15.22. Moreover, twin patterns are also observed. On the other hand, the C_{60} $[10\bar{1}]$ pattern is only observed with a AlN $[11\bar{2}0]$ azimuth of the electron beam at 160°C, and the C_{60} $[11\bar{2}]$ pattern is only observed with a AlN $[10\bar{1}0]$ azimuth of the electron beam at 160°C, respectively, although the twin patterns are still observed at 160°C.

From the RHEED results we can summarize that the C_{60} layers that are grown at 100°C–130°C contain four kinds of grains, which are the C_{60} $[10\bar{1}]\|$AlN $[11\bar{2}0]$ (the 0° orientation) and its twin grains and the C_{60} $[11\bar{2}]\|$AlN $[11\bar{2}0]$ (the 30° orientation) and its twin grains, as shown in Figure 15.23. However, the RHEED intensity from the 0° orientation is always stronger than that from the 30° orientation. Figure 15.24 shows the intensity ratio of the 0° orientation grain as a function of the growth temperature. In this figure, the 0° orientation grains are grown more than the 30° orientation grains, and its ratio is about 7:3. The C_{60} layers that are grown at 160°C contain only the 0° orientation grains, although the twin grains still remain. The results strongly

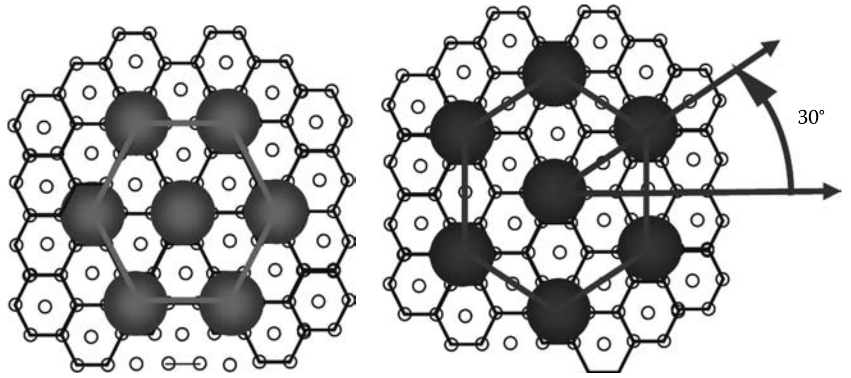

0° orientation
(C_{60} $[10\bar{1}]$||AlN $[11\bar{2}0]$)

30° orientation
(C_{60} $[11\bar{2}]$||AlN $[11\bar{2}0]$)

FIGURE 15.21 C_{60} $[10\bar{1}]\|$AlN $[11\bar{2}0]$ (the 0° orientation) and 7.0% in C_{60} $[11\bar{2}]\|$AlN $[11\bar{2}0]$ (the 30° orientation).

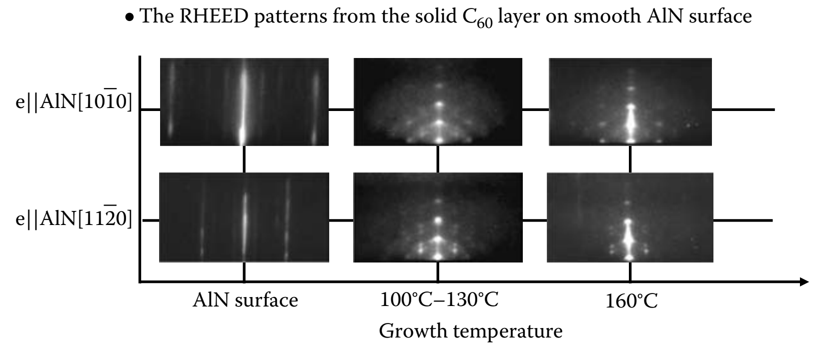
• The RHEED patterns from the solid C_{60} layer on smooth AlN surface
e||AlN[10$\bar{1}$0]
e||AlN[11$\bar{2}$0]
AlN surface
100°C–130°C
160°C
Growth temperature

FIGURE 15.22 Typical RHEED patterns from solid C_{60} layer grown on smooth AlN surface (RMS value: 0.9 nm) for 2 h at 100°C–160°C.

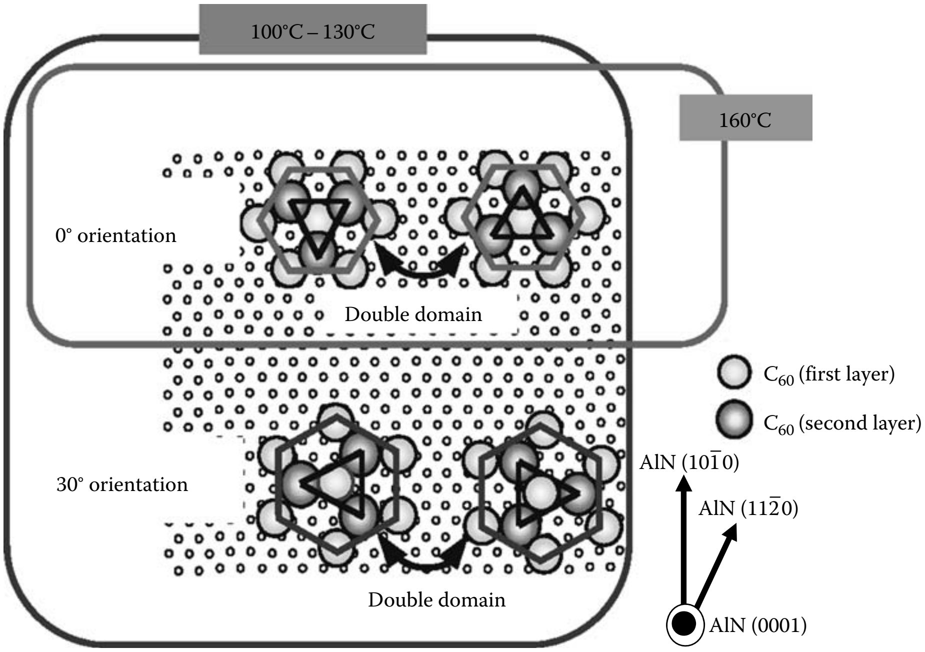
100°C – 130°C
160°C
0° orientation
Double domain
30° orientation
Double domain
C_{60} (first layer)
C_{60} (second layer)
AlN (10$\bar{1}$0)
AlN (11$\bar{2}$0)
AlN (0001)

FIGURE 15.23 Summarized illustration of epitaxial relationship between solid C_{60} and AlN (0001) surface.

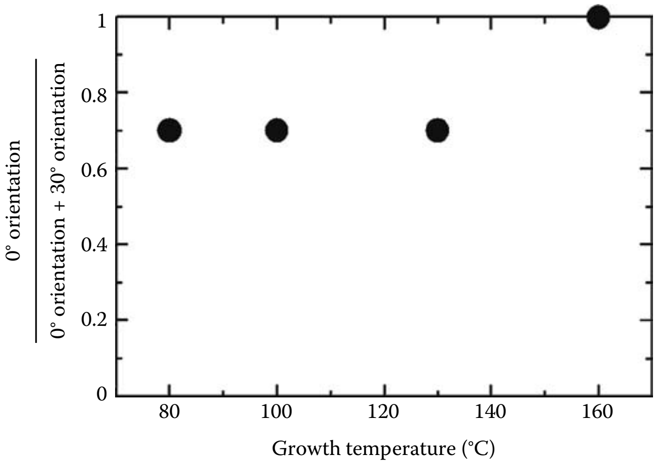

FIGURE 15.24 RHEED intensity ratio of 0° orientation grain as a function of growth temperature.

indicate that the C_{60} grains with the 0° orientations are stable on a smooth AlN surface in high-temperature growth, although the lattice-mismatch for each orientation is almost equal.

The solid C_{60} grains grown on the smooth AlN surface for 5 h at 100°C–160°C showed a three-dimensional island growth. Their grain density and grain size as a function of growth temperature show a trade-off relation which is strongly connected to the migration of C_{60} molecules and the covering of C_{60} layers on the AlN surface. The RHEED pattern for the solid C_{60} layer grown on the rough AlN (RMS value: 9.5 nm) surface at the 130°C shows a mixed pattern with a spot and a ring. The RHEED result indicates that the orientation of the grains is not controlled at all. Such random orientation can be improved by thermal annealing at 130°C after the low temperature growth at 30°C. However, the obtained C_{60} layers have multi-domain structures. The results indicate that the roughness of the AlN surface is also an important factor for C_{60} epitaxy. The RHEED pattern from the solid C_{60} layer grown on the polycrystalline AlN (RMS value: 0.7 nm) at 130°C shows a full ring pattern, which means the polycrystalline C_{60} layer is grown on the surface. It strongly indicates that the periodic atomic arrangement of the surface is also an important factor in C_{60} epitaxy as well as the surface roughness [52].

15.5 Principles of Organic FET Functions

The first report on the mobility measurement of an organic semiconductor material, merocyanine, using the field effect was published in the *Japanese Journal of Applied Physics* in 1984 by Kudo et al. [57]. Just after the first report on the field effect of the organic semiconductors, Koezuka et al. described the first OFET using polythiophene in 1987 [58]. After these innovations by Japanese researchers, the study of organic FETs has been extensively developed through the last two decades. The mobility of the carrier has achieved the same level as the ploy Si FETs in the late 1990s, as shown in Figure 15.1. Now, it has been reported that the highest mobility of organic FETs is 40 cm²/V s in a single crystalline rubrene FET [59]. According to Haddon et al., the solid C_{60} layers were utilized at the active element in thin film FETs [60]. However, the mobility of the FET was found to be rather low (10^{-4} cm²/V s) by Tang et al. [61]. Very recently, a group from the University of Tokyo has reported the highest electron mobility of 6.0 cm²/V s in the C_{60} FET structures at a low temperature [62]. The developments of the OFETs performance have been rapid in recent years like other organic devices such as organic electro-luminescence (OEL) [63]. In this section, we discuss about the general principles of OFETs, including C_{60} FETs for a deeper understanding of C_{60} FET operation in the later sections.

15.5.1 Device Structures of OFETs

Organic transistor materials are classified mainly into two categories: low- and high-molecular compound systems such as rubrene and polythiophene, respectively. In general, low-molecular compound systems have a rather high mobility. Consequently, the research on the OFETs using low-molecular systems has been intense in recent years. Therefore, in this section, we will discuss about FETs using low-molecular compound systems. The various types of FET structures such as vertical structure [64] and the Schottky gate structure [65] have already been reported. However, in this section we will discuss metal insulator semiconductor (MIS) structures, which are the most popular FET structures. In the OFET structure, mainly two types of structures are usually used in research: (a) bottom and (b) top contact structures, as shown in Figure 15.25a and b, respectively. The bottom contact structure is the most popular structure in OFETs, because a fine-patterning technique for the short channel is easily applicable just after the deposition of the electrode metal, as shown in Figure 15.25a. The organic active elements are deposited on the source and drain electrodes and the insulator layer after the fine-patterning of the electrodes. In general, in the bottom contact structure, the considerably large height difference between the electrodes and the insulator surface becomes an origin of high contact resistance between the active organic element and the metal electrodes. The discontinuity between the electrode and the organic materials at the edge of the electrodes becomes an origin of fluctuation in the FET performance such as in the *I–V* curves. In contrast, active organic elements are deposited on a flat insulator surface before the electrode formation in the top contact structure. However, achieving the fine-patterning of the organic semiconductor material systems is considerably difficult. Therefore, the metal electrodes formation is usually performed by using a metal-mask vacuum evaporation method. Moreover, a high contact resistance between the top metal electrodes and the interface of the insulator–organic layer in the active channel region cannot be essentially avoided. In practice, the contact resistance is lower than the case of the bottom contact FET, because the metal atoms of the electrodes may diffuse into the organic layer. In general, SiO_2 layers thermally grown on a Si substrate are used as the insulator layer and gold is used as source and drain electrode. The interface between metal and organic semiconductor becomes a Schottky junction because of the large difference in the carrier density between the metal and organic materials without doping. Therefore, the off-current is very low due to the large contact resistance in the OFET operations.

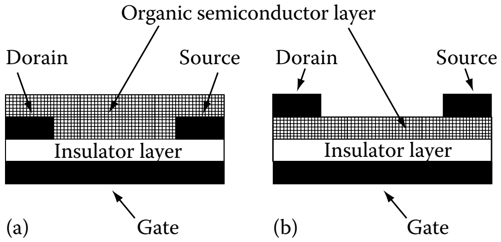

FIGURE 15.25 Schematics of (a) bottom and (b) top contact FET structures.

15.5.2 Metal/Organic Layer Interface

Organic semiconductors show intrinsic semiconductor nature and have very low impurity states in the band gap. Therefore, carriers excited by thermal energy play an essential role in the current conduction of OFETs. Both the electrons and the holes contribute to the carrier transport in the same manner as in am-bipolar operations. However, many experimental reports have been published on *p*-type materials such as rubrene and *n*-type materials such as solid C_{60} layers. Such a material dependency on the conducting types in spite of an essentially am-bipolar nature in organic semiconductors had been a large mystery in the OFET research for many years. However, very recently, Yasuda et al. have reported about the am-bipolar operations in OFETs such as pentacene or Cu-ftalocianina which had been considered a *p*-type OFET [66]. According to them, the am-bipolar operation can be realized by using a potassium electrode with a low work function, instead of the Au with a considerably high work function. Moreover, Nishikawa et al. have reported about the am-bipolar operation of OFETs using the work function as a control by using a modification technique that uses self-organized mono-layered molecules on the surface of Au electrodes [67]. These results strongly indicate that the formation of the carrier accumulation layer at the interface of OFETs is completely different from that of conventional semiconductor FETs. Namely, although the carrier accumulation layer for conventional semiconductor materials is formed by impurity doping into bulk, the carrier injection from the electrode plays an essential and important role for the formation of the carrier accumulation layer in OFET structures. According to Tamura et al., the formation of the accumulation layer in OFETs can be explained consistently using by the Maxwell–Wagner model for dielectric materials. In other words, the formation of the carrier accumulation layer can only be realized by carrier injection from the electrode with a low injection barrier height. However, the carrier accumulation layer cannot be formed for a high injection barrier [68]. The results of the analysis of the formation of accumulation layers indicate that the type of carriers of the accumulation layer can be determined by the relationship between the work function of the electrode metals and the HOMO–LUMO levels of the organic semiconductor materials, that is, the Schottky barrier between the electrode metals and organic semiconductor surfaces.

15.5.3 Field Effect Doping for Solid C_{60} Layers

A field effect doping technique has been proposed as a new type carrier doping that is applied to the interface between solid C_{60} and insulator layers by applying an external voltage to it in a similar manner to conventional field effect transistor (FET) [60]. In this doping technique, the solid C_{60} layer is treated as a semiconductor that was theoretically predicted by Saito and Oshiyama [28]. Injected career density and/or polarity through the interactions between the electrodes metals and the solid C_{60} layer can be controlled by the value and the polarity of external gate voltage in this technique. This point is a great advantage in the field effect doping for C_{60} device fabrication techniques in comparison with conventional chemical doping. However, several serious problems need to be resolved before there are realistic device applications using the field effect doping technique. The most important problem to be fixed is how to form a well-controlled interface between the insulator and the solid C_{60} layers. The channel region is formed in several mono layers from the interface. In the fabrication of the FET structure, an excellent C_{60} interface without any damage can be formed by the deposition of a C_{60} vacuum on a single crystal insulator layer such as the AlN layer (as shown in Figure 15.26) because the AlN single crystal layer can be grown on 6H-SiC [53,54], Si [55,56], and so on. In the conventional sputtering method for insulator film deposition, it is easy to damage the interface region. Therefore, it is a little difficult to form a reproducible and excellent interface. On the other hand, the single crystalline AlN layer has the large band-gap energy (6.2 eV). In addition, it also has enough high dielectric breakdown and dielectric constant for C_{60} FET applications. Consequently, it is expected that the interface between the single crystalline solid C_{60} layers grown on a single crystalline AlN insulator layer will become one of the most promising structures for field effect doping.

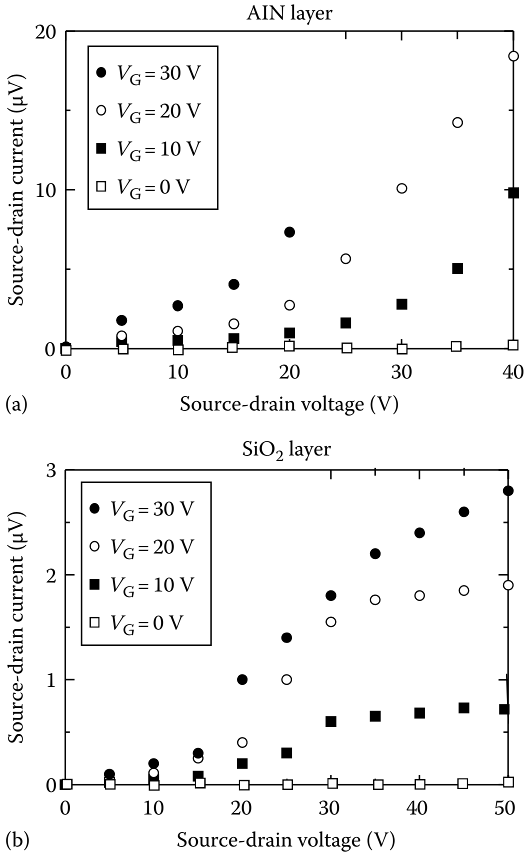

FIGURE 15.26 Typical I_D–V_{SD} characteristics of fabricated C_{60} FETs with (a) conventional SiO_2 layers and with (b) newly proposed AlN layers.

15.5.4 Am-Bipolar OFET Operations

In general, organic semiconductors such as solid C_{60} films are essentially intrinsic semiconductors, with the characteristics of low carrier density and an absence of impurity levels in the band gap. Therefore, if we can remove the extrinsic factors such as the absorption of gas and the carrier traps and can choose the proper electrode metals, an am-bipolar transistor can be easily realized in contrast to conventional semiconductors. There are many ways to realize a practical am-bipolar transistor such as fabricating the device under a vacuum or using inert gas ambient to prevent the absorption gas effect. Another typical idea is the usage of an asymmetric Schottky barrier source–drain electrode structure using different metals with large or small work functions from each other. Moreover, organic polymers such as polymethyl methacrylate (PMMA) can be applied as insulator layers to remove the electron traps on the SiO_2 surface. The electrons and the holes can be injected simultaneously under am-bipolar operation. Therefore, it is expected that a pseudo-PN junction will result in light emission due to a recombination of simultaneously injected electrons and holes in the channel. Am-bipolar light-emitting transistors have been reported in the carbon nanotubes, the organic polymers, and the tetracene channel FETs [69–71]. Unfortunately, however, am-bipolar operations have not been observed in solid C_{60} FETs until now for no essential reason.

15.6 Fabrication of Fullerene Field Effect Transistors (C_{60} FET)

The most important problem in the fabrication of the C_{60} FETs is how to form an effective interface between the insulator and the solid C_{60} layer. In this section, we discuss a new structure of the C_{60} FETs, which consists of an epitaxial AlN layer as the insulator and a solid C_{60} epitaxial layer grown on a AlN (0001) surface to realize the ideal interface, in comparison with the most popular structure of C_{60}/SiO_2 FETs. First of all, the solid C_{60} growth on the AlN (0001) surface by using the MBE technique is described again, although we have already described it in Section 15.4. Furthermore, how C_{60} islands with a fcc structure can be obtained on a smooth single crystalline AlN surface has been explained. The experimental results show that four types of C_{60} islands, in general, are obtained: those with 0° of orientation and its twin ones, those with 30° of orientation and its twin ones. However, only the islands with 0° orientation can be obtained by controlling of the solid C_{60} growth temperature. Such results strongly indicate that an effective interface between the AlN insulator and the solid C_{60} layer can be formed for device applications. The fabrication of polycrystalline solid C_{60}/AlN FET structures and the conductive characteristic of C_{60} FETs as a function of the C_{60} grain size have also been described. The C_{60} grain size grown on the polycrystalline AlN insulator layer is always larger than on SiO_2 insulator layers. Moreover, the electron mobility increases with the C_{60} grain size. It is also shown that the activation energy in the hopping conduction below the threshold voltage decreases along with the C_{60} grain size. These results all strongly indicate that the C_{60}/AlN FET structure has a great potential in achieving an outstanding performance in C_{60} FETs [72].

15.6.1 Substrates: SiO_2 or AlN

The most popular substrate is a conductive Si substrate with a thermal oxidation layer, SiO_2. The SiO_2 layer has some excellent properties as an insulator layer, and it is well known in the Si-MOSFET world. However, the fatal shortcoming in the SiO_2 layer as the insulator layer of the solid C_{60} FET is that the amorphous surface structure prevents single crystalline solid C_{60} epitaxy. On the other hand, AlN layers grown on conductive Si (or SiC) substrates have similar excellent properties as the insulator films of C_{60} FETs to SiO_2 films, except that they have a crystal surface. The crystal surface of the insulator film is necessary in forming an ideal interface between solid C_{60} layers and the insulator layers. For van der Waals epitaxy of solid C_{60} layer to form the ideal interface, however, there are also other necessary conditions such as lattice-matched conditions and the surface roughness of the insulator layer to be met. Therefore, the AlN layer grown on the Si (or SiC) substrates has a great potential for the insulator layer of solid C_{60} FETs.

15.6.2 Device Process

The drain and source electrodes of gold are usually fabricated on the insulator layer by conventional lithography techniques. The channel length (L) and width (W) of the fabricated FET described here are 7.5 and 175 μm, respectively. After the lithography of the electrodes, the C_{60} layers are grown at 30°C–160°C by a conventional solid source MBE system.

15.6.3 C_{60} Epitaxy

The lattice mismatches between the solid C_{60} (111) surface with the fcc structure and the AlN (0001) surface with the wurtzite structure (h-AlN) are 7.3% in C_{60} $[10\bar{1}]\|$AlN $[11\bar{2}0]$ (0° orientation) and 7.0% in C_{60} $[11\bar{2}]\|$AlN $[11\bar{2}0]$ (30° orientation), as shown in Figure 15.21. The lattice mismatch can be permitted in the van der Waals epitaxy of the solid C_{60} layer [45]. It has been reported that a fcc C_{60} layer can be grown on h-GaN (0001) which is chemically stable just like AlN [51]. Therefore, AlN can be expected to be a substrate for the single crystal C_{60} layers. However, the lattice-mismatch in both orientations (shown in Figure 15.21) is almost the same. Because the C_{60} van der Waals epitaxy growth orientation on the smooth surface strongly depends on the lattice mismatch [45], it is a little difficult to predict the C_{60} orientations on the AlN surface by using the above considerations. If single crystalline C_{60} layers can be grown on the AlN surface, a more effective interface can be obtained than from an interface formed by the conventional sputtering method. However, no one has succeeded in forming a single crystalline C_{60} layer grown on the AlN surface so far. Therefore, we have to investigate the characteristics of FETs with polycrystalline solid C_{60} layers on the AlN surface while at the same time studying the single C_{60} layer growth on single AlN layers. In particular, it is very interesting that the size of the C_{60} grains may be strongly related to conductive characteristic such as the mobility and/or the activation energy of hopping conduction in FET performance.

A smooth single crystalline AlN surface (root mean square: RMS value: 0. 9 nm), a rough single crystalline AlN surface (RMS value: 9.5 nm), and a smooth polycrystalline AlN surface (RMS value: 0. 7 nm) can be formed by the nitridation of Al_2O_3 (0001) substrate or by the AlN growth on them by using radio-frequency plasma-assisted molecular beam epitaxy (RF-MBE). The C_{60} layers described here can also be grown on a AlN surface for 2–5 h at 60°C–160°C by a conventional solid-source MBE system under a base pressure of $\sim 10^{-7}$ Pa. The monitoring of the growth process is usually performed by RHEED. The surface morphology of the solid C_{60} grown layer can be characterized by using the AFM. The AlN layers should be grown on a p-type

Si (111) substrate by using the RF-MBE system at about 500 nm (RMS value: 1.7 nm) in order to fabricate the bottom contact FET structure. For the C_{60}/SiO_2 FET structure, for example, the *p*-type Si (100) substrate is ordinarily used as a substrate covered with a SiO_2 insulator layer thermally formed at roughly 250 nm (RMS value: 0.2 nm).

Figure 15.22 shows typical RHEED patterns from solid C_{60} layers grown on a smooth AlN surface (RMS value: 0.9 nm) for 2 h at 100°C–160°C. All RHEED patterns in Figure 15.22 show ones from the fcc (111) structure. Ring patterns appear in spot patterns below a growth temperature of 100°C. Both the RHEED patterns, the one with azimuth electron beam paralleled to the AlN $[10\bar{1}0]$ and the AlN $[11\bar{2}0]$ crystal directions from samples grown at 100°C–130°C, show the inclusion of both C_{60} $[10\bar{1}]$ and C_{60} $[11\bar{2}]$ patterns, as shown in Figure 15.22. Moreover, twin patterns are also observed. On the other hand, the C_{60} $[10\bar{1}]$ and the C_{60} $[11\bar{2}]$ patterns are only observed with the AlN [11–20] and AlN $[10\bar{1}0]$ azimuths from samples grown at 160°C, respectively. However, the twin patterns are still observed even at 160°C.

The RHEED results can be summarized as follows: the C_{60} layers grown at 100°C–130°C contain four kinds of islands which are in the C_{60} $[10\bar{1}] \| $AlN [11] (the 0° orientation) epitaxial relation and its twin islands. They are also in the C_{60} $[11\bar{2}] \| $AlN $[11\bar{2}0]$ (the 30° orientation) epitaxial relation and its twin islands, as shown in Figure 15.23. However, the RHEED intensity from the islands with the 0° orientation is always stronger than from the islands with the 30° orientation. Figure 15.24 shows the RHEED intensity ratio for both islands in the epitaxial relation with the 0° and the 30° orientations as a function of the growth temperature. The islands with 0° orientation are grown much larger than ones with 30° orientation. The ratio is about 7:3. The C_{60} layer grown at 160°C contains only islands with 0° orientation. However, the twin islands still remain. The results strongly indicate that the C_{60} islands with the 0° orientation are stable on a smooth AlN surface in high-temperature growth, while the lattice-mismatch for each orientation is almost the same.

The C_{60} islands grown on the smooth AlN surface at 100°C–160°C show a three-dimensional island growth. The density and the size of the islands, as a function of growth temperature, show a trade-off relation which is strongly connected to the migration of the C_{60} molecules and the covering of the C_{60} layers on the AlN surface.

The RHEED pattern for a solid C_{60} layer grown on a single AlN with a rough surface (RMS value: 9.5 nm) at 130°C shows a mixed pattern with a spot and a ring. The RHEED results indicate that the orientation of the grains is not controlled at all. Such random orientation can be improved by the thermal annealing at 130°C after a low-temperature growth at 30°C. However, the obtained C_{60} layers have multi-domain structures. The results indicate that the roughness of a single crystalline AlN surface is also one of the important factors for solid C_{60} epitaxy.

The RHEED pattern from a solid C_{60} layer grown on a polycrystalline AlN (RMS value: 1.7 nm) at 130°C shows a complete ring pattern. This means that a polycrystalline C_{60} layer is grown on the surface. Furthermore, it strongly indicates that the periodic atomic arrangement on the AlN surface is also an important factor in solid crystalline C_{60} epitaxy as well as in surface roughness [52].

The RHEED observations show that polycrystalline C_{60} layers can be grown on both AlN and SiO_2 layers. By observing the AFM in surface morphology, it was revealed that the grain size of the polycrystalline C_{60} layers grown on polycrystalline AlN layers was larger than those grown on SiO_2 layers under the same conditions of C_{60} growth. The results indicate that a better interface due to the obtained larger C_{60} grains can be made on AlN layers rather than on SiO_2 layers.

15.6.4 Measurements of FET Performance

A drain current versus a source–drain voltage (I_D–V_{SD}) characteristic of FETs are usually measured by a conventional three probe current-voltage measuring system in an ambient N_2 atmosphere to avoid oxidation.

15.7 Characterization of C_{60} FETs on SiO_2 or on AlN

15.7.1 *I–V* Characteristics

Typical I_D–V_{SD} characteristics of fabricated C_{60} FETs with (a) conventional SiO_2 layers and with (b) newly proposed AlN layers are shown in Figure 15.26a and b, respectively. The typical I_D–V_{SD} characteristics consist of three regions: (1) the sub-threshold, (2) the linear, and (3) the saturation regions, similar to the I_D–V_{SD} characteristics of polycrystalline Si thin-film transistors (TFTs) [73]. According to the standard explanation of polycrystalline Si TFTs, those three regions can be explained by the differences in the conduction mechanisms. The linear and/or the saturation regions in the I_D–V_{SD} curves can be understood by normal band conduction. The electron mobility can be estimated from the I_D–V_{SD} curves. The electron mobility of such C_{60} FETs with (a) SiO_2 and with (b) AlN layers are estimated from the I_D–V_{SD} curves, as shown in Figure 15.26a and b and are 2.0×10^{-2} and 9.0×10^{-2} cm^2/V s, respectively. It is well known that the grain size is closely related to the mobility in conventional polycrystalline Si TFTs. The electron mobility estimated from I_D–V_{SD} curves as a function of the C_{60} grain size is shown in Figure 15.27. The electron mobility increases with grain size similar to conventional polycrystalline Si TFTs.

15.7.2 Sub-Threshold Region

The sub-threshold region at a low V_{SD} can be understood as hopping a conduction region [74]. According to many reports on organic FETs, a hopping conduction is commonly observed in them, and it strongly depends on the temperature and the electric-field-strength in the channel [75]. The activation energy of carrier hopping in C_{60} FETs, which is estimated from the temperature dependence of the I_D–V_{SD} curves in the range of 200–300 K by a gate voltage of 7.5 V, is shown in Figure 15.28. The

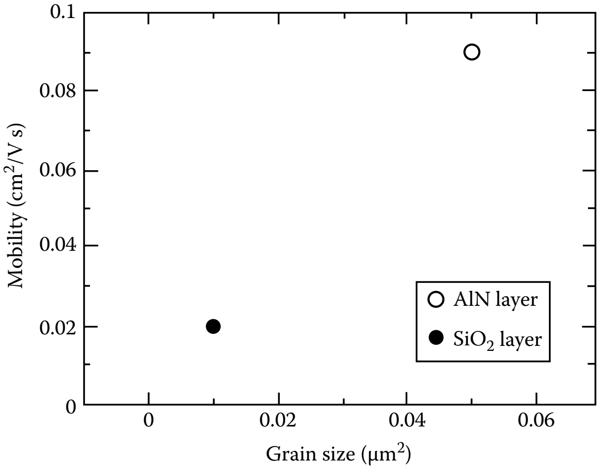

FIGURE 15.27 Electron mobility estimated from I_D–V_{SD} curves as a function of C_{60} grain size.

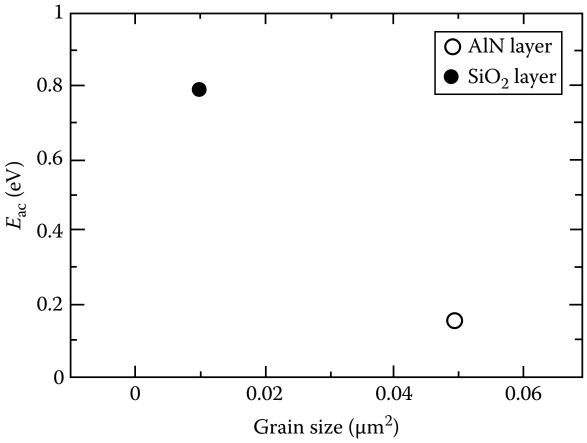

FIGURE 15.28 Activation energy of carrier hopping in C_{60} FETs with AlN and SiO_2 insulator layers.

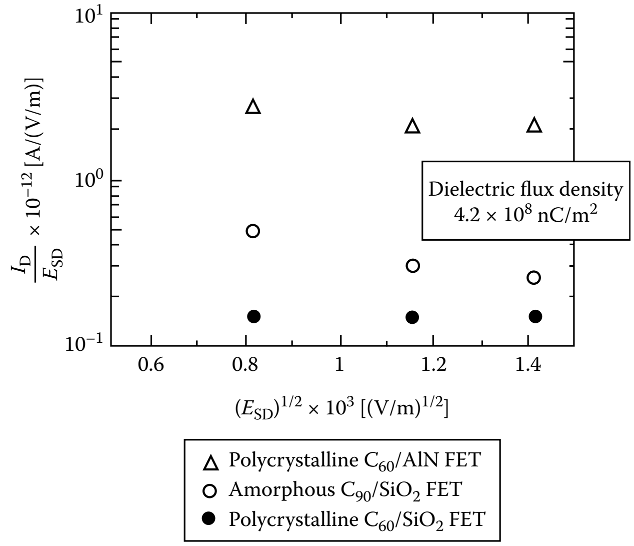

FIGURE 15.29 Electron mobility in hopping region as a function of electrical field.

results clearly indicate that the activation energy decreases with grain size in a manner similar to organic FET cases [76].

Moreover, a disorder model for a hopping conduction is given by the following equation:

$$\mu = \mu_0 \exp\left[-\left(\frac{2}{3}\hat{\sigma}\right)^2\right] \exp\left[C(\hat{\sigma}^2 - \Sigma^2)E^{1/2}\right] \quad \hat{\sigma} = \frac{\sigma}{kT} \tag{15.6}$$

where μ_0, σ, and Σ are the pre-exponential factors, the parameters of diagonal disorder, and the parameter of off-diagonal disorders, respectively [75]. The diagonal and the off-diagonal disorders mean the energetic and the positional disorders, respectively. When the Σ is larger than the σ, the electron mobility in a hopping conduction has a negative dependence on the electric-field-strength. The electron mobility in the hopping region is shown in Figure 15.29 as a function of the electrical field in the interface. The permittivity and the thickness of the AlN and the SiO_2 layers are different from each other. Consequently, the number of carriers induced by the gate voltage should be estimated as a function of the dielectric flux density. As shown in Figure 15.29, the electron mobility in a polycrystalline C_{60}/SiO_2 FET does not depend on the electric-field-strength. However, the electron mobility in the amorphous C_{60}/SiO_2 FET and the polycrystalline C_{60}/AlN FET decreases with the electric-field-strength. Therefore, such a negative slope in the electron mobility as a function of the electric-field-strength in the hopping region means that a positional disorder is dominant at the interface of those FETs. That is, a route for the conductive current without any relation to the electric-field-strength must be formed in polycrystalline C_{60}/AlN FETs and amorphous C_{60}/SiO_2 FETs. Such large positional disorders can be understood by looking at the geometrical interface roughness which is due to a rough surface morphology of the AlN layers in the case of C_{60}/AlN FETs and by the structural interface roughness due to the poor amorphous structure of the C_{60} layers in C_{60}/SiO_2 FETs.

15.8 Summary

We have described solid C_{60} FETs starting from a basic explanation of C_{60} molecules to FET applications, through an introductory discussion about van der Waals epitaxy and organic FET operations. The most essential issue in C_{60} FET research at the present stage is the formation of an ideal interface between solid C_{60} and the insulator layers. To accomplish this, we need a new combination of solid C_{60} and insulator layers. In this chapter, we discussed that one of the possible strong candidates is a combination of solid C_{60} and single-crystalline AlN grown on either SiC or Si substrates. However, this problem still remains open to debate.

In the near future, it is expected that rather high-temperature superconductivity will be obtained to form the ideal interface in some solid C_{60} FET structures. The reasons of this expectation are based on the potential of hole-injection in field effect doping and high-energy inner-cluster phonon (~1000 K) which will assist in the superconductivity in the bulk of solid C_{60} [77,78]. Furthermore, the various combinations of carbon materials such as nanotubes and graphene with solid C_{60} FET structures will bring new applications like new molecular devices to reality. The C_{60} FET structures must play essential and important roles in carbon nano-electronics in the research beyond CMOS in the future.

References

1. http://www.aist.go.jp/aist e/latest reserach/2004/20041118/20041118.html: Latest researches, "n-Type Organic Thin Film Transistor Prepared by Printing Method."
2. http://www.lucent.com/news: Bell Lab Research Review Report, Report of the investigation committee on the possibility of scientific misconduct in the work of Hendrik Schön and Coauthors, 2006/11/29.
3. H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smally, *Nature*, ***318***, 162 (1985).
4. E. A. Rohlfing, D. M. Cox, and A. Kaldon, *J. Chem. Phys.*, ***81***, 3322 (1984).
5. H. W. Kroto, *Angew. Chem. Int. Ed. Engl.*, ***31***, 111 (1992).
6. R. E. Smally, *The Science*, March/April, p. 22 (1991).
7. J. Baggott, *Perfect Symmetry*, Oxford University Press, Oxford, U.K. (1994).
8. W. Kratchmer, L. D. Lamb, K. Fostiropoulos, and D. R. Huffman, *Nature*, ***347***, 354 (1990).
9. R. E. Haufler et al., *J. Phys. Chem.*, ***94***, 8634 (9990).
10. G. Peters and M. Jansen, *Angew. Chem. Int. Ed. Engl.*, ***31***, 223 (1992).
11. J. B. Howard, J. T. McKinnon, Y. Markarousky, A. L. Lafleur, and M. E. Johnson, *Nature*, ***352***, 139 (1991).
12. R. D. Averitt, J. M. Alford, and N. J. Halas, *Appl. Phys. Lett.*, ***65***, 374 (1994).
13. R. A. Laudise, Ch. Kloc, P. G. Sinpkins, and T. Siegrist, *J. Cryst. Growth* ***187***, 449 (1998).
14. R. Taylor, J. P. Hare, A. K. Abdul-Sada, and H. W. Kroto, *J. Chem. Soc. Chem. Commun.*, ***20***, 1423–1425 (1990).
15. C. J. Welch and H. Pirkle, *J. Chromatogr.*, ***609***, 89 (1992).
16. S. Larsson, A. Volosov, and A. Rosen, *Chem. Phys. Lett.*, ***137***, 501 (1987).
17. D. E. Weeks and W. G. Harter, *J. Chem. Phys.*, ***90***, 4744 (1989).
18. R. E. Stanton and M. D. Newton, *J. Phys. Chem.*, ***92***, 2141 (1988).
19. S. J. Cyvin, E. Brendsdal, B. N. Cyvin, and J. Brunroll, *Chem. Phys. Lett.*, ***143***, 377 (1988).
20. M. S. Dresselhaus, G. Dresselhaus, and P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes*, Chap.11, Academic Press, New York (1996).
21. A. M. Rao et al., *Science*, ***259***, 955 (1993).
22. R. D. Johnson, G. Meijer, and D. S. Bethune, *J. Am. Chem. Soc.*, ***112***, 8983 (1990).
23. J. M. Hawkins, A. L. Meyer, A. Timothy, S. Loren, and F. J. Salam, *J. Phys. Chem.*, ***95***, 9 (9991).
24. W. I. F. David et al., *Nature*, ***353***, 147 (1991).
25. R. C. Haddon, *Acc. Chem. Res.*, ***25***, 127 (1992).
26. R. C. Haddon, L. E. Brus, and K. Raghavachari, *Chem. Phys. Lett.*, ***125***, 459 (1996).
27. A. D. J. Haymet, *Chem. Phys. Lett.*, ***122***, 421 (1985).
28. S. Saito and A. Oshiyama, *Phys. Rev. Lett.*, ***66***, 2637–2640 (1991).
29. W. I. F. David, R. M. Ibberson, and T. Matsuo, *Proc. R. Soc. Lond.*, ***A442***, 129 (1993).
30. R. Tycko, R. C. Haddon, G. Dabbagh, S. H. Giarum, D. C. Douglassc, and A. M. Muujsce, *J. Phys. Chem.*, ***95***, 518 (1991).
31. D. A. Neumann et al., *Phys. Rev. Lett.*, ***67***, 3808 (1991).
32. A. F. Hebard, *Phys. Today*, ***45***, 26 (November 1992).
33. K. Prassides, H. W. Kroto, R. Taylor, D. H. M. Walton, W. I. F. David, J. Tomkinson, R. C. Haddon, M. J. Rosseinsky, and D. W. Murphy, *Carbon*, ***30***, 1277 (1992).
34. J. H. Weaver, *J. Phys. Chem. Solids*, ***53***, 1433 (1993).
35. T. Arai, Y. Murakami, H. Suenatsu, K. Kikuchi, Y. Achiba, and I. Ikemoto, *Solid State Commun.*, ***84***, 827 (1992).
36. J. Mort et al., *Chem. Phys. Lett.*, ***186***, 284 (1992).
37. A. F. Hebard, R. C. Haddon, R. M. Fleming, and A. R. Kortan, *Appl. Phys. Lett.*, ***59***, 2109 (1991).
38. G. B. Alers, B. Golding, A. R. Kortan, R. C. Haddon, and F. A. Thiel, *Science*, ***257***, 511 (1992).
39. S. Kazaoui, R. Ross, and N. Minami, *Solid State Commun.*, ***90***, 623 (1994).
40. D. J. van den Heuvel et al., *Chem. Phys. Lett.*, ***233***, 284 (1995).
41. B. Pevzner, A. F. Hebard, R. C. Haddon, S. D. Senturia, and M. D. Dresselhaus, *Mater. Res. Soc. Symp. Proc.*, ***359***, 423 (1995).
42. F. Kajzar, *Synth. Met.*, ***54***, 21 (1993).
43. Y. B. Zhao, D. M. Poirier, R. J. Pachman, and J. H. Weaver, *Appl. Phys. Lett.*, ***64***, 577 (1994).
44. P. Zhou et al., *Appl. Phys. Lett.*, ***60***, 2871 (1992).
45. K. Tanigaki, S. Kuroshima, and T. W. Ebbesen, *Thin Solid Film*, ***257***, 154 (1995).
46. A. Koma, *J. Cryst. Growth*, ***201/202***, 236 (1999).
47. Y. Saito, *Mater. Sci. Eng. B*, ***18***, 229 (1993).
48. M. Sakurai, H. Tada, and A. Koma, *Jpn. J. Appl.*, ***30***, L1892 (1991).
49. G. Genstlnblum, L. M. Yu, J. J. Pireaux, P. A. Thiry, R. Caudano, J. M. Themlin, S. Bouzidi, F. Coletti, and J. M. Debever, *Appl. Phys.*, ***A56***, 175 (1993).
50. H. Takashima, M. Nakaya, A. Yamamoto, and A. Hashimoto, *J. Cryst. Growth*, ***227/228***, 825 (2001).
51. H. Takashima, M. Nakaya, A. Yamamoto, and A. Hashimoto, *J. Cryst. Growth*, ***227/228***, 829 (2001).
52. D. Yokoyama, H. Nojiri, A. Yamamoto, and A. Hashimoto, *Phys. Stat. Sol.* (a), ***195***, 3 (2003).
53. N. Teraguchi, A. Suzuki, Y. Saito, T. Yamaguchi, T. Araki, and Y. Nanishi, *J. Cryst. Growth*, ***230***, 392 (2001).

54. N. Onojima, J. Suda, and H. Matsunami, *J. Cryst. Growth*, ***237–239***, 1012 (2002).
55. V. Lebedev, B. Schroter, G. Kipshidze, and W. Richter, *J. Cryst. Growth* ***207***, 266 (1999).
56. U. Kaiser, P. D. Brown, I. Khodos, C. J. Humphreys, H. P. D. Schenk, and W. Richter, *J. Mater. Res*., ***14***, 2036 (1999).
57. K. Kudo, M. Yamashina, and T. Moriizumi, *Jpn. J. Appl. Phys*., ***23***, 130 (1984).
58. H. Koezuka, A. Tsumura, and T. Ando, *Synth. Met*., ***18***, 699 (1987).
59. J. Takeya, M. Yamagishi, Y. Tominari, R. Hirahara, Y. Nakazawa, T. Nishikawa, T. Kawase, T. Shimoda, and S. Ogawa, *Appl. Phys. Lett*., ***90***, 102120 (2007).
60. R. C. Haddon, A. S. Perei, R. C. Morris, T. T. M. Palsta, A. F. Hebard, and R. M. Fleming, *Appl. Phys. Lett*., ***67***, 121 (1995).
61. J. Tang, G. Xing, Y. Zhao, L. Jing, H. Yuan, F. Zhao, X. Gao, H. Qian, R. Su, K. Ibrahim, W. Chu, L. Zhang, and K. Tanigaki, *J. Phys. Chem*., ***B111***, 11929 (2007).
62. M. Kitamura, S. Aomori, J. Ho Na, and Y. Arakawa, *Appl. Phys. Lett*., ***93***, 033313 (2008).
63. J. Kido, M. Kimura, and K. Nagai, *Science*, ***267***, 1332 (1995).
64. D. X. Wang, Y. Tanaka, M. Iizuka, S. Kuniyoshi, K. Kudo, and K. Tanaka, *Jpn. J. Appl. Phys*., ***38***, 256 (1999).
65. Y. Ohmori, K. Muro, M. Onoda, and K. Yoshino, *Jpn. J. Appl. Phys*., ***31***, L646 (1992).
66. T. Yasuda, T. Goto, K. Fujita, and T. Tsutsui, *Appl. Phys. Lett*., ***85***, 2098 (2004).
67. T. Nishikawa, S. Kobayashi, T. Nakanowatari, T. Mitani, T. Shimoda, Y. Kubozono, G. Yamamoto, H. Ishii, M. Niwano, and Y. Iwasa, *J. Appl. Phys*., ***97***, 104509 (2005).
68. R. Tamura, E. Lim, T. Manaka, and M. Iwamoto, *J. Appl. Phys*., ***100***, 114515 (2006).
69. J. A. Misewich, R. Martel, Ph. Avouris, J. C. Tsang, S. Heinze, and J. Tersoff, *Science*, ***300***, 783 (2003).
70. J. Zaumseil, R. H. Friend, and H, Sirringhaus, *Nat. Mater*., ***5***, 69, (2006).
71. T. Takahashi, T. Takenobu, J. Takeya, and Y. Iwasa, *Adv. Funct. Mater*., ***17***, 1623 (2007).
72. H. Nojiri, D. Yokoyama, A. Yamamoto, and A. Hashimoto, *Diamond Relat. Mater*., ***14***, 518 (2005).
73. Y. G. Yoon, G. B. Kim, B. I. Lee, and S. K. Joo, *Thin Solid Films*, ***466***, 303–306 (2004).
74. C. P. Jarrett, K. Pichler, R. Newbould, and R. H. Friend, *Synth. Met*., ***77***, 35–38 (1996).
75. H. Bassler, *Phys. Stat. Sol*., (b) ***175***, 15–56 (1993).
76. K. Horiuchi, S. Uchino, K. Nakada, N. Aoki, M. Shimizu, and Y. Ochiai, *Physica B*, ***329–333***, 1538–1539 (2003).
77. T. W. Ebbesen, J. S. Tsai, K. Tanigaki, J. Tabuchi, Y. Shimakawa, Y. Kubo, I. Hirosawa, and J. Mizuki, *Nature*, ***355***, 620 (1992).
78. T. W. Ebbesen, J. S. Tsai, K. Tanigaki, H. Hiura, Y. Shimakawa, Y. Kubo, I. Hirosawa, and J. Mizuki, *Physica* C, **203**, 163 (1992).

16

The Cooper-Pair Transistor

16.1 Introduction **16**-1
16.2 Theory of Operation **16**-1
The Single Josephson Junction • Coulomb Blockade • Band Structure of the CPT • Quasiparticle Poisoning
16.3 Fabrication **16**-9
Electron-Beam Lithography and Two-Angle Deposition • Aluminum and Gap Engineering
16.4 Practical Operation and Performance **16**-10
Current Switching Electrometry • Zero-Biased rf Electrometry
16.5 Present Status and Future Directions **16**-16
Acknowledgment **16**-16
References **16**-16

José Aumentado
National Institute of Standards and Technology

16.1 Introduction

The Cooper-pair transistor (CPT) is a three-terminal superconducting device composed of a mesoscale superconducting island connected to drain and source leads via ultrasmall (~100 nm) Josephson junctions (see Figure 16.1). Due to its small size, the energy to add a single Cooper pair to the island can be large compared with the typical temperatures at which similar quantum circuits are operated, $T < 100\,\text{mK}$. Because of this, the transport properties of the device can be strongly dependent on the polarization charge presented by a capacitively coupled gate electrode and operated as a transistor with high input impedance. As an electrometer, it is useful as it can be operated with very little dissipation and should minimally influence the systems that it is measuring. As we will see, its similarity to another quantum circuit, the Cooper-pair box (CPB), marks it as a useful device to study decoherence in superconducting quantum bits (qubits) due to unpaired electrons.

While the CPT is topologically identical to the normal-state single-electron transistor (SET), it is operated in the superconducting state, and the superconducting nature of the device modifies the transport properties in several important ways. For our purposes, we also distinguish the CPT from the superconducting single-electron transistor (SSET) as a mode of operating this particular double-junction device. As a CPT, it is operated on the supercurrent branch near zero voltage in the coherent Cooper-pair transport regime, while as an SSET, it is operated at finite voltage bias and can include several incoherent hybrid quasiparticle/Cooper-pair tunneling processes.

16.2 Theory of Operation

As shown in Figure 16.1a, the device is composed of two Josephson junctions and can, at its core, be regarded as a single Josephson-like element with an added internal charge degree of freedom. This modification is important as it allows the CPT to act as a tuneable Josephson junction that is sensitive to charge. In this spirit, it is important that we begin with a conceptual understanding of a single Josephson junction.

16.2.1 The Single Josephson Junction

A superconducting Josephson junction is composed of a thin insulating barrier connecting two superconducting electrodes. In 1962, Brian Josephson made the remarkable statement that these junctions should support a supercurrent, that is, a dissipationless current at zero voltage (Josephson, 1962). This is a consequence of the finite quantum mechanical tunneling probability that arises from the spatial overlap in the superconducting order parameter on either side of a thin junction barrier. Josephson encapsulated this effect in the following equations that relate the current and the voltage to the dynamics of the phase difference across the junction*

$$I = I_0 \sin\delta \tag{16.1}$$

$$V = \varphi_0 \dot{\delta}. \tag{16.2}$$

---

* We use the notational shorthand $\varphi_0 \equiv \hbar/2e$. This is simply the superconducting flux quantum $\Phi_0 = h/2e$ divided by 2π.

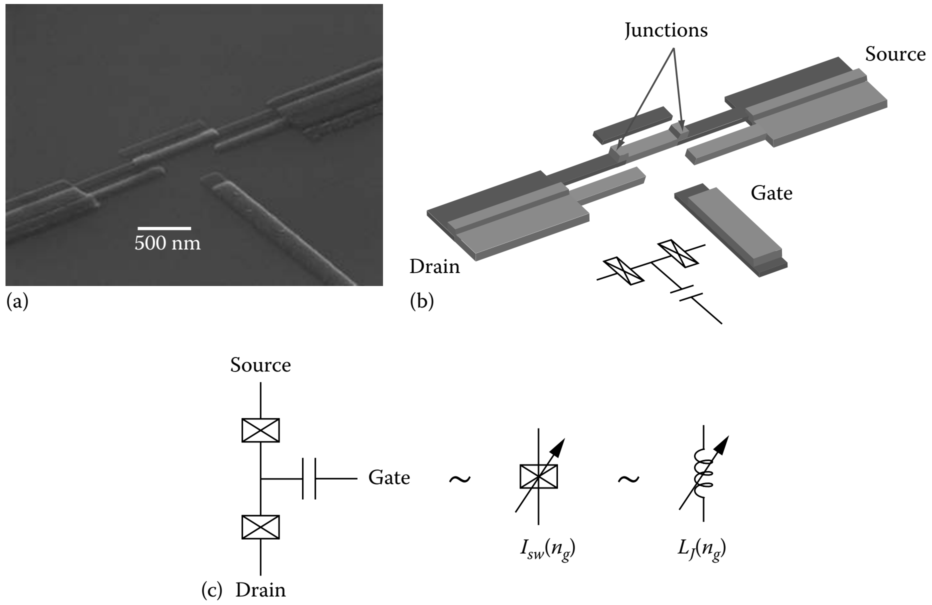

FIGURE 16.1 (a) Field emission micrograph of typical CPT. (b) 3D equivalent model showing two-angle/double image deposition (see Section 16.3) and location of junctions at the overlap. (c) The CPT is effectively a Josephson junction with a critical current and inductance that modulate with gate polarization charge/voltage $n_g \equiv C_g V_g / e$.

Taken at face value, these equations are the constitutive equations defining the Josephson junction as an inductor-like nonlinear circuit element. I_0 is a "critical current" beyond which the junction cannot support a supercurrent and the junction switches to a finite voltage branch in the current–voltage (I–V) characteristic. It can be computed from the low-temperature form of the Ambegaokar–Baratoff relation* (Ambegaokar and Baratoff, 1963).

$$I_0 = \frac{2}{eR_N}\frac{\Delta_1\Delta_2}{\Delta_1+\Delta_2}K\left(\frac{|\Delta_1-\Delta_2|}{\Delta_1+\Delta_2}\right), \tag{16.3}$$

where
- Δ_1 and Δ_2 are the superconducting gap energies on either side of the insulating barrier
- K is the complete elliptic integral of the first kind

If $\Delta_1 \sim \Delta_2$, $K \to \pi/2$ and

$$I_0 \simeq \frac{\pi}{eR_N}\frac{\Delta_1\Delta_2}{\Delta_1+\Delta_2}. \tag{16.4}$$

The most commonly used superconductor for CPTs (and SETs) is aluminum, Δ_{Al} = 180–250 μeV. Typical junction sizes are ~100 nm × 100 nm that yield junction resistances in the range of $10\,\mathrm{k}\Omega < R_N < 100\,\mathrm{k}\Omega$. With these parameters, the critical current ranges from 1 to 100 nA.

The presence of the phase δ across the junction in the equations marks it as an inductor-like element that can store magnetic flux. Following this reasoning, one can derive a Josephson inductance by finding the scaling coefficient between the voltage and the time derivative of the current,

$$L_J(I) = \frac{L_{J0}}{|\cos\delta|} = \frac{L_{J0}}{\sqrt{1-(I/I_0)^2}}, \quad \text{where } L_{J0} \equiv \frac{\varphi_0}{I_0}. \tag{16.5}$$

One can also define a Josephson energy that is a function of the phase difference δ,

$$E_J = \varphi_0 I_0 |\cos\delta| = E_{J0}\sqrt{1-\left(\frac{I}{I_0}\right)^2}, \quad \text{where } E_{J0} \equiv \varphi_0 I_0. \tag{16.6}$$

This can be interpreted as the magnetic energy that is stored in the element. This energy has a deeper significance as the coupling energy between Cooper-pair charge states on either side of the junction and will be important in calculating the energy bands of the CPT. For CPT Josephson junctions, the equivalent inductances are typically ~10–30 nH and the energies 20–80 μeV (0.2–1 K in temperature units).

16.2.1.1 The *I–V* Curve and Phase Dynamics

16.2.1.1.1 The I–V Curve

The CPT is operated on the so-called "supercurrent branch" of the *I–V* curve at (or near) zero voltage. The two measurement modes that we discuss in this chapter involve measuring the

---

* The form of the Ambegaokar–Baratoff relation shown here is uncommon in that it acknowledges the difference in gap energies that seems to be typical in evaporated aluminum junctions. This gap energy difference is usually ignored in the early SSET and CPT literature but has become important in more recent CPT experiments.

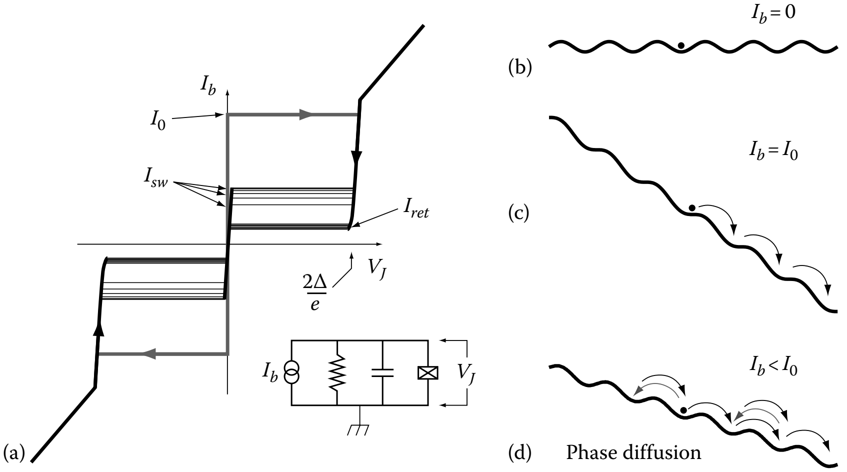

FIGURE 16.2 (a) Schematic I–V curve/characteristic for an ultrasmall Josephson junction. The ideal $T = 0$ I–V is shown in light gray, switching to the voltage state at $I_b = I_0$. For $T > 0$, thermal fluctuations can drive the system to the voltage state at much lower bias currents in ultrasmall ($A_{jn} <$ 0.01 μm^2) junctions, as shown in black. (b) $I_b < I_0$. The phase particle is trapped in a local minima of a washboard potential and $\langle\dot{\delta}\rangle = 0$ and no voltage is developed across the junction. (c) $I_b = I_0$. The washboard potential tilted at exactly the critical slope necessary to "release" the phase particle from its local minimum and enter a free-running state at finite voltage, V (= $\varphi_0\langle\dot{\delta}\rangle > 0$). (d) $I_b < I_0$. Phase diffusion. In ultrasmall junctions, thermal fluctuations can drive the phase to evolve in a 1D random walk of phase steps $\Delta\phi \simeq 2\pi$. Since the washboard is tilted, this results in a slow diffusion in one direction, V (= $\varphi_0\langle\dot{\delta}\rangle$) > 0, although it is still considered to be on the "supercurrent branch."

switching current I_{sw}, at which the CPT switches from the supercurrent branch to the finite-voltage branch as well as measuring the effective inductance L_J of the CPT when biased at zero current. While the small size of the junctions and internal charge degree of freedom will affect these parameters significantly, it is worthwhile to first look at the single junction case and understand how the dynamics of the phase determine the structure of the I–V curve (see Figure 16.2a).

16.2.1.1.2 The Resistively and Capacitively Shunted Junction (RCSJ) Model

A more realistic model of a single Josephson junction includes a parallel shunt resistance and capacitance (see inset in Figure 16.2a). The shunt resistance R encapsulates the resistance of the junction to normal/quasiparticle currents through the junction, while the capacitance C is simply the physical capacitance that arises from having two planar electrodes overlapping with an insulating barrier in between. In the junctions comprising a CPT, the intrinsic $R > 10\,\mathrm{k}\Omega$ and is considered effectively "unshunted" since R is much bigger than the impedance presented by the Josephson inductance. As noted previously, the Josephson tunnel element is a nonlinear inductance but, for small phase excursions, it looks like a linear inductor. The circuit model presented then looks very much like a parallel LCR circuit or damped simple harmonic oscillator with a natural frequency

$$\omega_p = \frac{1}{\sqrt{L_{J0}C}} = \sqrt{\frac{2eI_0}{\hbar C}}. \tag{16.7}$$

This is usually called the "plasma frequency" and has the significance of simply being the frequency of small oscillations for a Josephson junction.

Although this model can be used to describe the intrinsic Josephson junction dynamics, it is also useful in determining the dynamics when the junction is placed in an arbitrary measurement circuit. If, for instance, we want to current bias the junction, we might add a voltage source V_b in series with a resistor R_b and construct a Norton equivalent that is simply an ideal current bias $I_b = V_b/R_b$ in parallel with an output impedance R_b.* In this case, we can roll R_b into the shunt R in our RSJ model. With this simple circuit equivalent, we can derive a set of first-order differential equations defining the equations of motion for the phase by writing down Kirchoff's equations for the currents in each branch. We can write this system of equations as a single second-order differential equation

$$\ddot{\delta} + \frac{1}{RC}\dot{\delta} + \omega_p^2\left(\sin\delta - \left(\frac{I_b}{I_0}\right)\right) = 0. \tag{16.8}$$

For small δ at zero current bias, this describes the motion of a particle with coordinate δ oscillating in a quadratic potential with a friction term, viz., it's a damped simple harmonic oscillator (and is consistent with our earlier statement that L_J is linear for small oscillations). At finite current bias, however, this equation describes the motion of the particle in a tilted sinusoidal (or *washboard*) potential.† We know from Equation 16.2 that in order for the junction to support a finite dc voltage across it, the particle must have some net motion in the phase, that is, $\langle\dot{\delta}\rangle \neq 0$. Since the particle is trapped at a fixed phase when $I_b < I_0$, we have a well-defined phase and the junction remains in the supercurrent state (see Figure 16.2b). By applying a bias current, we tilt the washboard and, when $I_b = I_0$, the particle is tipped out of its well and begins to run away (Figure 16.2c). In this free-running state, the time derivative of the phase has both an ac and a dc component corresponding to ac and dc voltages across the junction. In a more realistic treatment, I_b includes a stochastic contribution from thermal fluctuations due to the dissipation in the environment and so the current at which the junction switches to the free-running state is lower than I_0 since thermal excitations can drive the phase particle over the potential barrier prematurely. Because of this, the measured current is distributed according to the statistics of a thermally excited (Kramers) escape process (Fulton and Dunkleberger, 1974). We will differentiate this switching current I_{sw} from the ideal Ambegaokar–Baratoff value I_0 and remember that it is actually a statistical variable.

---

* This vast oversimplification has its caveats and must be amended to account for stray capacitance in the bias lines and any other reactances relevant in the frequency range of our junction dynamics (typically $\omega < 50\,\mathrm{GHz}$ for small aluminum junctions).

† More in depth discussions of the RCSJ model can be found in Tinkham (2004) and Clarke and Braginski (2004).

If we reduce the current bias in the free-running, finite voltage state, the particle velocity begins to slow down due to the dissipation in the circuit. Once the dissipation rate can compensate for the ac power generated by the free-running voltage oscillations, the particle "retraps" into a minima in the washboard potential and the junction returns to the zero-voltage supercurrent branch. The current at which this happens, I_{ret}, the "return current" is an indication of the amount of damping or dissipation in the environment at the free-running oscillation frequency (Tinkham, 2004). Roughly, the closer to zero I_{ret} is, the less damping there is. Although the CPT is designed and operated in the unshunted, hysteretic regime, in many practical dc SQUID magnetometers, the junctions are intentionally shunted with damping resistors to eliminate hysteresis in the *I*–*V* measurement permitting stable finite voltage operation.

16.2.1.1.3 Concerns Specific to Small Junctions

While the picture above is the standard way to look at relatively large junctions (area $\gtrsim 1\,\mu\text{m}^2$), the junctions in CPTs are usually about a factor of 100 smaller and have much smaller capacitances. The main consequence of this is that the observed switching currents can easily be smaller than I_0 by an order of magnitude. One way to understand this is to return to the harmonic oscillator analogy. We can see that $\omega_p^2 \propto 1/C$, so C is an effective mass for our "phase particle." In CPT junctions, this means that the equivalent phase particle can be relatively light compared with its larger junction cousins and so thermal fluctuations, even at 100 mK, can have a more significant effect, kicking the particle into the free-running state much more easily. This has the end effect of reducing I_{sw} far below I_0 and also widening the distribution of I_{sw} (Figure 16.2a).

A more problematic effect of the light phase particle mass is the fact that the same thermal fluctuations can "kick" the phase particle from well to well in the washboard potential in a process called "phase diffusion" (Kautz and Martinis, 1990). At zero current bias, this results in a one-dimensional random walk in the phase with zero mean. At finite current, the tilt in the washboard biases this diffusive motion in one direction, such that the mean phase velocity $\langle\dot{\delta}\rangle$ can be positive (negative) at positive (negative) bias (see Figure 16.2d). The result of all of this is that at finite currents below the switching current, we often see a small finite voltage. This slope is fairly uniform near zero current and gives the supercurrent branch a slight resistive contribution that can correspond to 10 s or even 100 s of Ωs. The downside of this behavior is that the junction or CPT is not truly dissipationless.

For electrometer operation, we want I_{sw} to have as narrow a distribution as possible since it is the quantity that will vary with applied gate voltage/polarization charge. It has been demonstrated, however, that one can narrow this distribution and also increase I_{sw}, so that it is closer to I_0 by fabricating a larger capacitance (a few picofarad) in the leads shunting the junction to "weigh down" the phase particle (Joyez et al., 1994; Joyez, 1995).

16.2.2 Coulomb Blockade

While the above discussion of a single junction will eventually be relevant to the CPT measurements we are interested in, we have not yet seriously considered the role of the internal charge degree of freedom in determining the overall critical current and Josephson inductance in this device. This is, afterall, the knob that turns the CPT into a transistor.

We can qualitatively recognize the role of the island charge by first determining the energy required to charge the CPT island with a single electron charge *e*,

$$E_C = \frac{e^2}{2C_\Sigma}, \tag{16.9}$$

where

$$C_\Sigma \equiv C_{J1} + C_{J2} + C_g + C_{stray}. \tag{16.10}$$

Here, C_Σ is the total capacitance as seen by the island, and while it includes the junction and gate capacitances, it also includes a "catch-all" term, C_{stray}, that is meant to include any contributions from unintentional coupling to ground. In practice, $C_{stray} \ll C_\Sigma$ and can usually be ignored. The energy E_C is called the single-electron Coulomb blockade energy (Devoret and Grabert, 1992).

For typical junction capacitances $C_{J1,J2} \lesssim 1$ fF and gate capacitances of 0.01–1 fF, E_C can be on the order of ~100–400 μeV. In temperature units, this temperature is 1–5 K and is therefore a significant energy scale compared with typical operating temperatures. As noted previously, the Josephson energies of these junctions are roughly on par with the Coulomb blockade energies, so we can qualitatively assume that both energy scales will play a role in the transport.

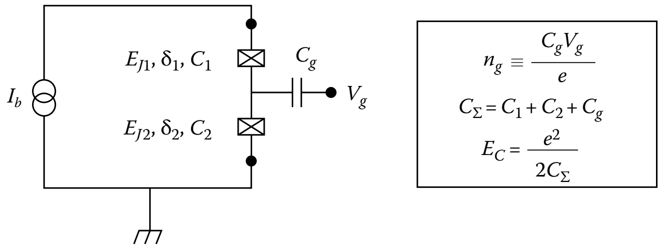
E_{J1}, δ_1, C_1
E_{J2}, δ_2, C_2
C_g
V_g
I_b
$n_g \equiv \frac{C_g V_g}{e}$
$C_\Sigma = C_1 + C_2 + C_g$
$E_C = \frac{e^2}{2C_\Sigma}$

FIGURE 16.3 Simplified CPT schematic. Each ultrasmall junction has an associated Josephson energy and phase along with an overall Coulomb blockade energy that defines the energy required to place discrete charge on the island. Voltage applied to the gate injects "gate polarization charge" into modulating the Coulomb or ionization energy. This modulates its Josephson-like behavior through charge-phase duality (see text).

16.2.3 Band Structure of the CPT

Since we have two junctions in series (Figure 16.3), we must define separate phases for each junction δ_1 and δ_2.* We can now be more precise about the role of the charging energy by looking at the Hamiltonian for the CPT and how the contributions couple to the phases and the island charge n,

$$H = H_{J1} + H_{J2} + H_{CB} + H_{QP}. \tag{16.11}$$

$H_{J1,J2}$ are the contributions due to the Josephson energy coupling the charge states on either side of the junctions, while H_{CB} is the Coulomb blockade contribution. Explicitly, we have

$$H_{J1,J2} = -E_J \cos\delta_{1,2} \tag{16.12}$$

$$H_{CB} = \frac{E_C}{4}(n - n_g)^2. \tag{16.13}$$

Here we have used the reduced gate voltage $n_g \equiv C_g V_g/e$ that is the polarization charge applied by the gate in electron units. H_{QP} is the contribution from quasiparticle excitations in the leads and on the island. Physically, they correspond to unpaired electrons/holes that can be introduced to the system via thermal excitations or some nonequilibrium process that can break Cooper pairs. While this is an important contribution, we ignore it now to facilitate the discussion.

We can write down a matrix equivalent of Equation 16.11 in the basis of island charge states $|n\rangle$, noting that the Josephson coupling terms that link Cooper-pair charge states on either state of the junction also link "adjacent" Cooper-pair charge states $|n\rangle \rightarrow |n \pm 2\rangle$.† Defining "external" and "internal" phases $\delta = \delta_1 + \delta_2$ and $\theta = \delta_1 - \delta_2$ and setting $\theta = 0$ (equivalent to assuming that the island has negligible inductance), we can write H in the charge basis. For example, the Hamiltonian spanning the Cooper-pair island charge states $|-2\rangle$, $|0\rangle$, and $|2\rangle$ is

$$H = \begin{bmatrix} E_C(-2-n_g)^2 & -\frac{1}{2}(E_{J1}e^{-i\delta/2} + E_{J2}e^{+i\delta/2}) & 0 \\ -\frac{1}{2}(E_{J1}e^{+i\delta/2} + E_{J2}e^{-i\delta/2}) & E_C(-n_g)^2 - \frac{1}{2} & (E_{J1}e^{-i\delta/2} + E_{J2}e^{+i\delta/2}) \\ 0 & -\frac{1}{2}(E_{J1}e^{+i\delta/2} + E_{J2}e^{-i\delta/2}) & E_C(2-n_g)^2 \end{bmatrix}. \tag{16.14}$$

The eigenvalues for this matrix represent the energy bands and are functions of n_g and external phase δ. The choice of basis size (how big a matrix one really needs to write down) is dictated by the ratio of $E_{J1,2}/E_C$. If this ratio is small, charge is easily localized and the corresponding external phase fluctuations are large, so the device has a weak effective Josephson energy. In this case, the charge basis may be described well by a small subspace of $|n\rangle$ spanning only a few charge states near $|0\rangle$. If $E_{J1,2}/E_C \gtrsim 1$, this approximation fails and one must write down a larger matrix spanning a bigger subspace of $|n\rangle$. In this limit, the charge on the island is no longer as well defined as its quantum mechanical state is now described by a linear superposition of several charge states. Likewise, fluctuations in the CPT external phase are then small and the CPT begins to look like a single Josephson junction.

16.2.3.1 The Uncertainty Principle at Work

We can numerically diagonalize Equation 16.14 to calculate the eigenvalues at various n_g and δ. Figure 16.4 shows the resulting eigenenergy surfaces corresponding to the ground and first excited state energies, ϵ_0 and ϵ_1. The quantities that we want to measure, $I_{sw}(n_g)$ and $L_J(n_g)$, are defined by the first and second derivatives of these surfaces in the phase. The size of this modulation is a direct consequence of how well the CPT island localizes Cooper pairs, $n2e$, through the uncertainty relation between charge number and phase, $\Delta n \Delta\delta \geq 1/2$. In this sense, the CPT is a strange charge-based device because it relies on the competition between number and phase uncertainty. This is in contrast to the SSET or SET where E_C is entirely dominant. In the CPT, we would like E_C to be big but would also like E_J to be comparable in magnitude, so that the switching current is not so small that it is difficult to measure well. Similarly, rf measurements of the Josephson inductance (described later in this review) become easier at bigger E_J as larger signal powers can be used while biased on the supercurrent branch.

16.2.3.2 Calculation of the Critical Current Modulation

The critical current of the CPT can now be determined from the eigenenergies calculated above. By restricting to the ground state band $\epsilon(n_g, \delta)$, we can compute the critical current (Joyez, 1995)

$$I_0(n_g) = \frac{1}{\varphi_0} \left. \frac{\partial \epsilon_0(n_g, \delta)}{\partial \delta} \right|_{\delta_{max}}, \tag{16.15}$$

---

[*] The phases δ_1 and δ_2 bear a strict conjugate correspondence to the number of Cooper pairs having "flown" through junctions 1 and 2. This discussion closely follows that given by Joyez (1995) and a more extensive discussion of choosing appropriate quantum variables appears there.

[†] For consistency with Joyez (1995), n in $|n\rangle$ refers to the single electron number and *not* Cooper pair number.

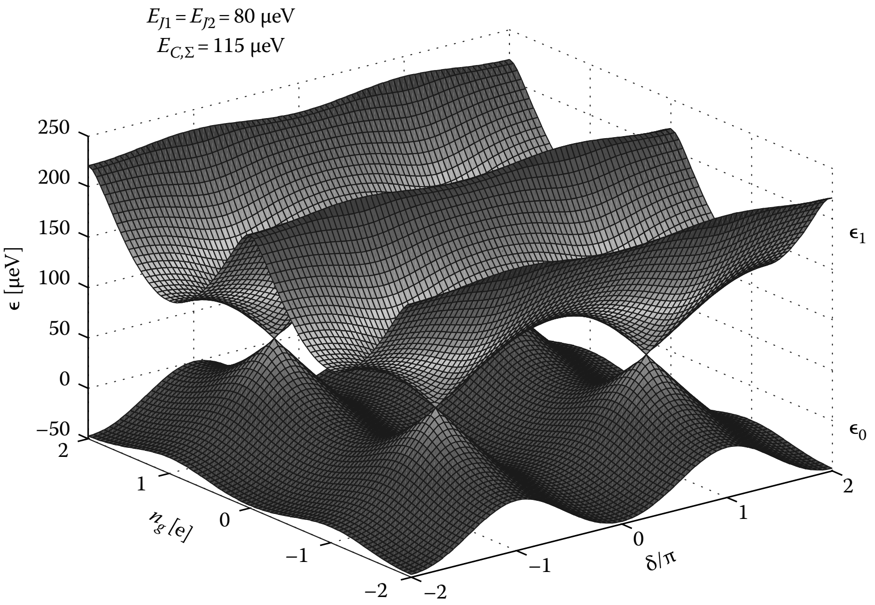

FIGURE 16.4 **(See color insert following page 20-14.)** The ground (ϵ_0) and first excited (ϵ_1) states calculated by numerically diagonalizing Equation 16.11 with the energies indicated in the figure.

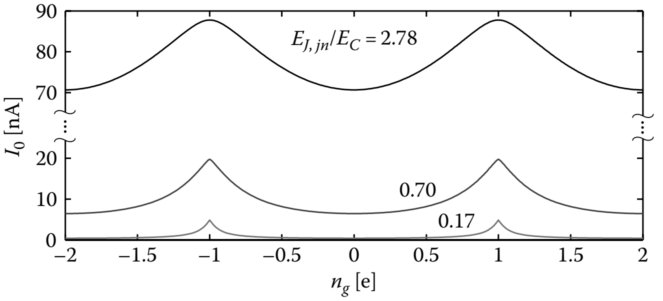

FIGURE 16.5 Numerically calculated CPT critical currents for various ratios of $E_{J,jn}/E_C$. In this example, E_C is fixed at 115 μeV, while $E_{J,jn}$ is varied.

where δ_{max} is the phase maximizing the derivative.* In Figure 16.5, we show the calculated critical currents for several ratios of E_J/E_C. Note that the calculated currents are 2*e* periodic in the applied gate charge which is the result of the Josephson coupling of Cooper-pair charge states. In practice, measured switching currents in CPTs are usually significantly smaller than the calculated value due to small-junction effects noted above. However, the total magnitude of the modulation can be on the order of 10 nA for practical aluminum device parameters and is easily measured.

16.2.3.3 Calculation of the Effective Inductance Modulation

The ground state energy surface $\epsilon_0(n_g, \delta)$ corresponds to an effective Josephson energy that is a function of charge as well as phase. When the ratio $E_J/E_C \lesssim 1$, the phase dependence, while still 2π periodic, gets "peaky" near odd-integer n_g. Despite this, one may still define a Josephson inductance in a similar manner (Sillanpää et al., 2004; Naaman and Aumentado, 2006b), that is,

$$L_J(n_g) = \frac{1}{\varphi_0^2} \frac{\partial^2 \epsilon_0(n_g, \delta)}{\partial \delta^2}\bigg|_{\delta=0}. \tag{16.16}$$

This is similar to Equation 16.5, but since the $\epsilon_0(n_g, \delta)$ is not strictly sinusoidal, we do not get the exactly the same result. For typical aluminum device parameters, the inductance modulation can be 10–100 nH and, like the switching current, is easily measured.

16.2.4 Quasiparticle Poisoning

Although the "quasiparticle" excitations that are talked about in superconductivity do not explicitly bear a well-defined charge, they do represent the screened excitations presented by an unpaired electron that might exist due to the breaking of a Cooper pair. A quasiparticle can tunnel onto a CPT island by "undressing" itself of its screening cloud and tunneling through a junction barrier alone. At this point, this additional charge presents a full electron charge offset to the CPT island. Since the CPT switching current and inductance are nominally 2*e* periodic in the gate charge, the addition of an extra electron can present a significant change in the way the CPT is operated, since it fluctuates the charge by *e*. In the literature, this problem was first discussed in terms of island "parity," but more recent work has favored the more colorful term "quasiparticle poisoning." The most important thing about quasiparticle poisoning for electrometry is that it is a stochastic process whose dynamics can happen on timescales comparable to our measurement time, so its effect on the resulting measurement must be well understood. That being said, quasiparticle poisoning has proven to be an interesting problem in itself, and the CPT has been very useful in its study.

---

* Incidentally, one can also take derivatives of the excited state eigenenergies and compute critical currents for those bands. Interested readers can see Flees et al. (1997) for further details.

To see how quasiparticles enter the picture, we now include the Hamiltonian H_{QP} in our description of the CPT,

$$H_{QP} = \sum_j \epsilon_j \gamma_j \gamma_j^\dagger. \tag{16.17}$$

The $\gamma_j, \gamma_j^\dagger$ are annihilation/creation operators for quasiparticle excitations in the superconductor (Tinkham, 2004), while ϵ_j is the energy of the excitation. In our case, this can be Δ_1 or Δ_2, remembering that the island and leads can have different gap energies. These quasiparticles are usually taken to be thermally generated and therefore have an exponentially small probability of existing at low temperatures. The expression for the quasiparticle density is (cf., Shaw et al., 2008),

$$n_{qp} = D(\epsilon_F)\sqrt{2\pi\Delta_j k_B T}\,\exp\left(\frac{-\Delta_j}{k_B T}\right). \tag{16.18}$$

$D(\epsilon_F)$ ($=2.3 \times 10^4\ \mu\text{m}^{-3}\ \text{J}^{-1}$) is the density of states at the Fermi energy and Δ_j is the superconducting gap energy in either film 1 or 2. Plugging in $\Delta \sim 200\,\mu\text{eV}$ (aluminum) and $T = 100\,\text{mK}$ gives us $2 \times 10^{-4}\ \mu\text{m}^{-3}$. This is a very low number considering the volume of a generic CPT island and leads; yet we know, experimentally, that the actual quasiparticle density is typically 10–$1000\,\mu\text{m}^{-3}$ (Mazin, 2004; Shaw et al., 2008). The disparity in the thermal prediction versus the experimentally measured numbers is the first indication that the source for the quasiparticles we see at low temperatures is distinctly nonthermal in nature. As yet, no one has determined the source of these *nonequilibrium* quasiparticles and we are stuck with the problem of understanding them.

16.2.4.1 Energetics of the Nonequilibrium Quasiparticle Poisoning Process

Since we are forced to work in an environment filled with nonequilibrium quasiparticles, we must figure out how to include them in the band picture that we constructed above. We have three states of interest (see Figure 16.6) (Aumentado et al., 2004):

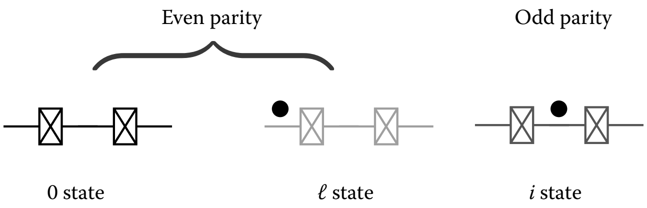

FIGURE 16.6 Three state nonequilibrium quasiparticle model. "0" state: no quasiparticles in vicinity of CPT, even parity. "ℓ" state: quasiparticle in leads even parity. "*i*" quasiparticle on island, odd parity.

State 0 No quasiparticles in or near the CPT island. *Even parity.*

State ℓ A quasiparticle is in the leads, in the vicinity of the CPT island. No quasiparticles on the island. *Even parity.*

State i A quasiparticle is on the CPT island. No quasiparticles in the leads near the island. *Odd parity.*

These states can all be described with the band structure derived in the previous section by offsetting the modulated energy surfaces by the superconducting gap energy, corresponding to where the quasiparticle lives, viz.,

$$\begin{aligned} E_0(n_g) &\equiv \epsilon_0(n_g, \delta = 0), \\ E_\ell(n_g) &\equiv \epsilon_0(n_g, \delta = 0) + \Delta_\ell, \\ E_i(n_g) &\equiv \epsilon_0(n_g + 1, \delta = 0) + \Delta_i. \end{aligned} \tag{16.19}$$

Here, we include the possibility of different gap energies in the leads and island, Δ_ℓ and Δ_i, respectively, and confine ourselves to $\delta = 0$. In principle, phase diffusion will smear these levels somewhat, but the average phase on the supercurrent branch will always be localized in the bottom of a potential well in the phase when biased near $I_b = 0$.

In Figure 16.7, we show the energy bands for two different pairs of island + gap energies using typical Coulomb, Josephson, and gap energies for an aluminum device. In the "type H" device (Figure 16.7a), the island gap is greater than the lead gap, and in the "type L" device, the reverse is true. Assuming that nonequilibrium quasiparticles are present, we can limit our discussion to E_ℓ and E_i. In the $T = 0$ limit, we expect the system to relax to the lowest energy state and the problem is reduced to whether the ℓ or i state energy is smallest (denoted in Figure 16.7a and d by the black dotted trace). For this purpose, we introduce the energy difference:[*]

$$\delta E_{\ell i}(n_g) \equiv E_i(n_g) - E_\ell(n_g). \tag{16.20}$$

At $T = 0$, the sign of this quantity determines what state we are in, that is,

$$\operatorname{sgn}[\delta E_{\ell i}] = \begin{cases} +1 & \text{even parity} \\ -1 & \text{odd parity} \end{cases} \tag{16.21}$$

---

[*] *A brief history detour.* In early treatments of quasiparticle poisoning, all quasiparticles were assumed to be thermal in nature, so the important energy in the problem is the free energy required for transition from 0 to *i* state (Tuominen et al., 1992, 1993;Amar et al., 1994; Joyez et al., 1994; Tinkham et al., 1995). This is basically the process of breaking a Cooper pair with thermal fluctuations. Experimentally, this model was more or less verified early on, but the situation was complicated by the anecdotal evidence that this picture failed to explain the numerous unpublished experiments that showed poisoning at low temperatures. In the absence of nonequilibrium quasiparticles, the early free energy theories are still valid and, in any case, when the system is heated sufficiently $T > 250\,\text{mK}$, the thermal quasiparticle generation rate dominates over the nonequilibrium rate. For a good review of the early theory, see Tinkham (2004).

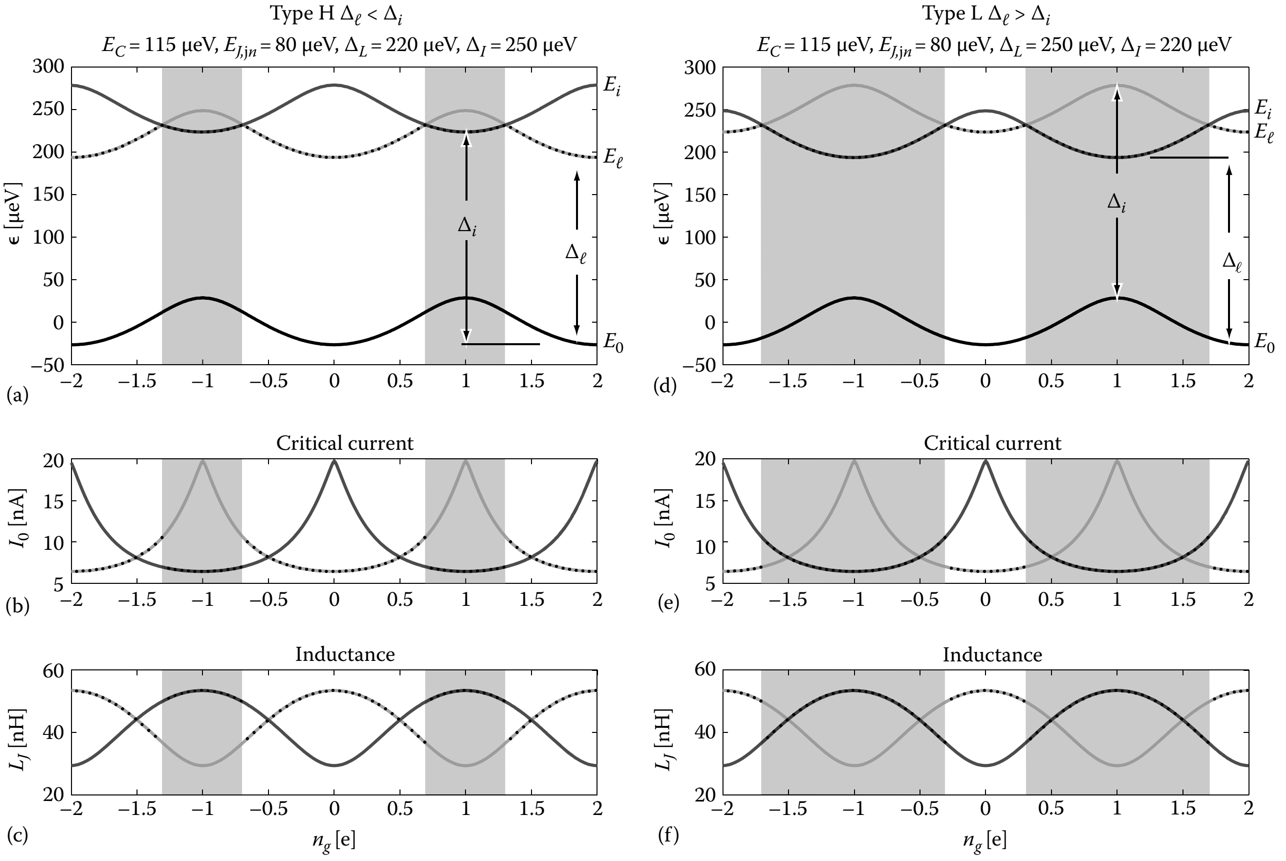

FIGURE 16.7 (a) Type H ($\Delta_\ell < \Delta_i$) energy bands. ℓ state/even parity (light gray), *i* state/odd parity (medium gray), 0 state/even parity (black solid). Minimum energy state is denoted as the black dotted trace. Corresponding even (light gray) and odd (medium gray) state (b) critical current and (c) effective inductance for a type H device. (d) Type L ($\Delta_\ell > \Delta_i$) energy bands. Corresponding even (light gray) and odd (medium gray) state (e) critical currents and (f) effective inductances for a type L device. Gray areas mark range in n_g where the CPT island is "trap-like," that is, a potential well for quasiparticles and the parity state is bimodal and can rapidly switch when coupled to thermal excitations.

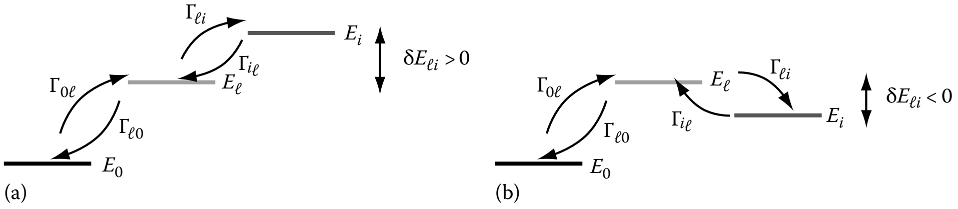

FIGURE 16.8 (a) Barrier-like configuration of levels. (b) Trap-like configuration of levels.

$\delta E_{\ell i}$ is an effective potential barrier height for quasiparticles. When $\delta E_{\ell i}$ is positive, the CPT island looks like a barrier, and when it is negative, it looks like a trap (Figure 16.8). In reality, the system is at finite temperature, so this qualitative picture must be amended to include the possibility of thermal excitations (phonons) coupling to the quasiparticles and exciting them out of the trap. Since this is a thermal escape process, we must characterize the system in terms of the average lifetimes of the poisoned and unpoisoned states, τ_o and τ_e.* Likewise, it is also useful to talk about these lifetimes as rates, $\Gamma_{eo} = \Gamma_{0\ell} + \Gamma_{\ell i} = 1/\tau_o$ and $\Gamma_{oe} = \Gamma_{i\ell} = 1/\tau_e$. Γ_{eo} is known as the "poisoning rate" and Γ_{oe} is the "ejection rate."

Experimentally, at $T \lesssim 250\,\text{mK}$, the poisoning rate Γ_{eo} has little temperature dependence and can be anywhere between 10^3 and $10^5\ \text{s}^{-1}$ (Aumentado et al., 2004; Ferguson et al., 2006; Naaman and Aumentado, 2006b; Court et al., 2008b). The fact that this rate is constant in this temperature range points to the notion that the source is nonthermal in nature and the rate for Cooper-pair breaking in the leads determines the poisoning rate, $\Gamma_{eo} \sim \Gamma_{0\ell} \gg \Gamma_{\ell i}$. The ejection rate Γ_{oe} does, however, have a temperature dependence at these low temperatures since the process of kicking a quasiparticle out of a potential well is a thermal escape process. One can derive the probability of the even parity state (no quasiparticle on the island) using detailed balance arguments (Aumentado et al., 2004),

---

* The "*e*" and "*o*" refer to "even" and "odd" parity states. Odd parity corresponds to a single excess quasiparticle on the CPT island.

$$P_e = P_0 + P_\ell = \frac{1}{1+\beta_{0\ell}e^{-\delta E_{\ell i}/k_BT}}. \tag{16.22}$$

The factor, $\beta_{0\ell} \equiv \Gamma_{0\ell}/(\Gamma_{0\ell} + \Gamma_{\ell 0})$, accounts for the generation and recombination rates for nonequilibrium quasiparticles in the leads. What this all means is that there is not really any strictly quasiparticle-free zone for typical energies that we would use in practice.* In Figure 16.7, we denote the places where the CPT looks trap-like and it is apparent that both the L and H devices are affected. Even with the higher island gap, the H device can trap quasiparticles, albeit in a shallower potential, $\delta E_{\ell i}$ than presented by the L device and also over a smaller range in n_g.

It is interesting to note that if there is no source for nonequilibrium quasiparticles, $\Gamma_{0\ell} = 0$ ($\beta_{0\ell} = 0$) and $P_e = 1$ whether or not we have a barrier-like or trap-like profile. In other words, there would be no poisoning regardless of the relative gap energies. The fact that one sees poisoning easily in trap-like devices at low temperatures is the strongest indication that the quasiparticles in these systems are generated by some nonequilibrium process. Surprisingly, the temperature dependence above also predicts that the even state probability can actually be enhanced since, even though the poisoning rate might be fixed, the ejection rate can be increased by heating the CPT and giving quasiparticles energy to escape (Aumentado et al., 2004; Palmer et al., 2007).

This model of nonequilibrium quasiparticle poisoning was first roughly outlined and tested in Aumentado et al. (2004) using CPTs with engineered gap energy profiles and measured with the ramped current technique (see below). It was subsequently confirmed in several later experiments in both CPTs and CPBs using rf and dc techniques (Gunnarsson et al., 2004; Yamamoto et al., 2006; Palmer et al., 2007; Savin et al., 2007; Court et al., 2008b; Shaw et al., 2008).

16.3 Fabrication

16.3.1 Electron-Beam Lithography and Two-Angle Deposition

The majority of CPTs (and single-charge tunneling devices in general) are fabricated from evaporated aluminum using conventional electron beam lithography and electron-gun (or thermal) deposition techniques.

The most conventional method of fabrication is to first define the island and lead pattern of the CPT in a special double-layer e-beam resist stack that is spun onto a silicon substrate. The double-layer is constructed such that the upper "image" layer defines the pattern outline, while the lower "ballast" layer is meant to provide a vast "undercut" region underneath the image. This is usually achieved using resist for the lower layer that is much more sensitive to the e-beam exposure than the upper layer. When the pattern is written, the dosage required to generate the image overexposes the lower layer resist such that the pattern is much wider in the lower layer resist, forming an "undercut" region when the sample is developed (cf., Cord et al., 2006).

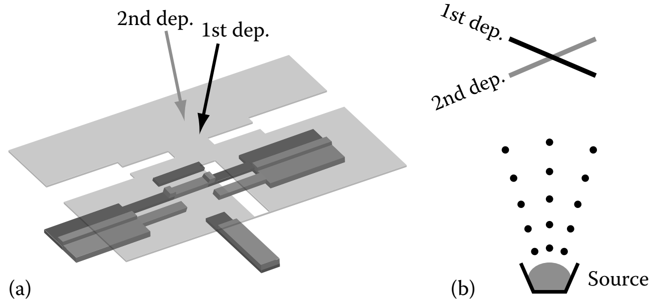

FIGURE 16.9 Angle deposition for the CPT. A mask is defined in electron-beam sensitive resist, usually in a bilayer configuration such that the underlayer is overdeveloped and the top (image) layer pattern can cast a shadow on the surface of the substrate from several angles. In this example, the leads for the device are deposited in the first deposition (dark gray) at an angle θ_1. After this deposition, oxygen is admitted into the deposition chamber forming an oxide on the surface of the metal. Finally, a second evaporation (light gray) is performed at θ_2 depositing the CPT island. In the process, Josephson junctions are formed at the overlap between the island and oxidized leads. (a) Shadow deposition through the top layer "image" resist. (b) Schematic of tilt configuration with respect to source.

The undercut region allows us to tilt the substrate so that the impinging aluminum atoms can deposit an image through the image resist mask at an angle (Figure 16.9). After the first aluminum deposition, oxygen is released into the chamber that grows a thin (<1 nm) insulating Al_xO_y layer on the surface of the aluminum. The substrate is then rotated to another angle and a second image is deposited onto the substrate such that it overlaps with the first oxidized image. In this manner, insulating junctions can be formed *in situ* without exposing the device to air. Subsequent processing removes the resist, so that it can be bonded up in a suitable measurement circuit. This process is called the "Dolan bridge" or "shadow deposition" technique (Dolan, 1977) and is used in almost all single-charge tunneling devices as well as most superconducting qubit designs. There are many variations on this technique but, surprisingly, the basic method has remained unchanged for more than 30 years.

16.3.2 Aluminum and Gap Engineering

Aluminum is regarded as the material of choice for these devices because of the general ease with which one can grow a reliable oxide layer at room temperature within the deposition chamber. Aluminum is easily deposited using both thermal and e-beam deposition techniques that are common to most clean rooms and requires very little in the way of special equipment. That being said, the CPT is a superconducting device whose operation is somewhat dependent on the quality of the superconducting gap.* In most general-purpose metal evaporators, the quality of the aluminum is completely beholden to the deposition hygiene of the community that uses the machine. For instance, the gap energy of aluminum is notoriously sensitive to the disorder and impurities that are present in the as-deposited film. For instance, two films deposited at identical rates but at different base pressures can yield gap energies that differ by 10 s of μeV. This was done intentionally by Aumentado et al. (2004) to fabricate type L and H devices, but the same effect can be achieved by simply changing the thickness of the aluminum. It turns out that the gap energy can be enhanced substantially by making the film thinner (Townsend et al., 1972), and subsequent CPT experiments demonstrated that this method can be used to make CPTs (Yamamoto et al., 2006; Court et al., 2008a).

---

* The reason for this is not obvious. Basically, the difference in aluminum gap energies that one can achieve practically is ~50 μeV. In order to see any level of charge localization at $T = 100$ mK, we need $E_C \sim 100$ μeV. If we want useful switching current modulation (~10 nA) and inductance modulation (~10 nH), then we are stuck with $E_{J,jn} \sim 50$ μeV. For these values, it is difficult to get the E_ℓ and E_i bands to separate.

For reference, typical cryopumped e-gun deposition systems can have base pressures in the low 10^{-8} Torr range, especially if they are dedicated to aluminum and a few other well-behaved metals and employ a load-lock system for exchanging samples. These systems usually surpass more general use machines in film quality and oxide reproducibility. The author prefers e-gun deposition out of a bare water-cooled copper hearth rather than using thermal deposition out of a tungsten boat. Tungsten has a tendency to alloy with aluminum and repeated depositions from the same source yield films with inconsistent quality (resistivity and gap quality) and must be replaced on every pumpdown.

We note in passing that CPTs have been successfully fabricated from niobium with aluminum oxide junctions, but in none of the published work has anyone been able to reduce quasiparticle poisoning in any significant way. While this is awful for making charge-based qubits, these devices can still be used as electrometers, but the fabrication techniques can prove to be much more difficult since niobium is a refractory metal and difficult to evaporate. At its evaporating temperature, most practical deposition systems outgas so much that the quality of the film can be compromised and the integrity of the resist can also become an issue (Dolata et al., 2002, 2005). Others have used a sputtered niobium/aluminum oxide/niobium trilayer (ubiquitous in SQUID fabrication) and patterned using focused ion beam (Watanabe et al., 2004).

16.4 Practical Operation and Performance

The Coulomb, Josephson, and superconducting gap energies are all 10–300 μeV, so practical measurements are performed well below 1 K in a dilution refrigerator, usually at $T \lesssim 100$ mK. There are currently two generic modes of operation of the CPT: current switching electrometry and rf electrometry. The main advantage of the current switching electrometry is its cost. rf electrometry can require relatively expensive microwave amplifiers, generators, and fast digitizers, while current switching electrometry can be achieved with simple function generators and voltage preamps operating with less than 1 MHz of bandwidth.

16.4.1 Current Switching Electrometry

One of the earliest methods for measuring the CPT was to measure the modulation of the switching current I_{sw} as a function of an applied gate voltage. This measurement is performed by cycling the current bias between the supercurrent and voltage branches in the *I*–*V* characteristic. We can track the current at which the CPT switches to the voltage state by tracking the voltage across the device and using a simple threshold trigger (see Figure 16.10). We can subdivide this measurement approach into ramped and

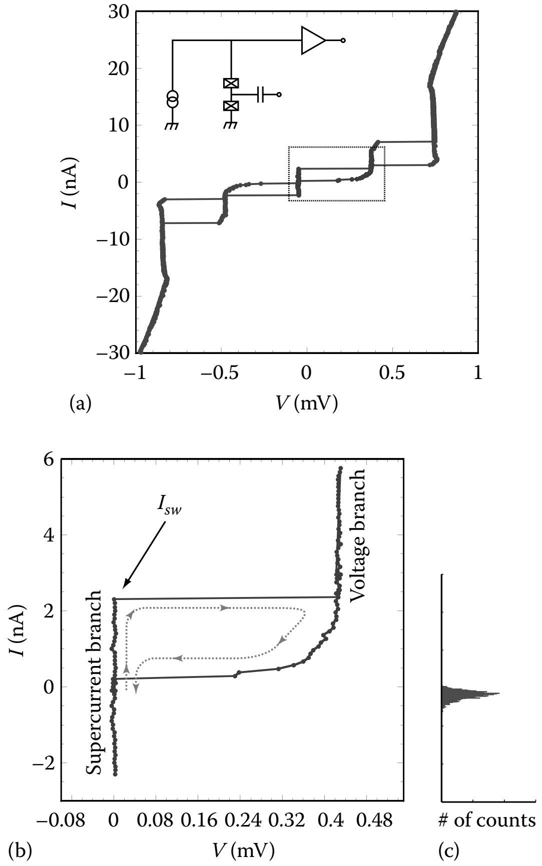

FIGURE 16.10 The switching current measurement. (a) Typical *I*–*V* curve for CPT device. The feature at ~0.4 mV ($\Delta_i + \Delta_\ell$) is the so-called Josephson quasiparticle peak or JQP and is due to more a complex Cooper-pair quasiparticle tunneling cycle. The device switches to the true voltage branch at $\gtrsim 2(\Delta_i + \Delta_\ell)$. Inset: Simplified measurement circuit schematic. (b) Expanded view of switching current cycle. The current is ramped along the supercurrent branch until the voltage across the CPT switches to the finite valued voltage branch. The current at which this happens is recorded, and the results of many ramp cycles are recorded in a switching current histogram as in (c).

---

* If the gap edge is not well defined and there is some finite density of states in the subgap, then the gap engineering arguments that we made previously lose their applicability. This is mostly only an issue in "dirtier" deposition systems.

pulsed current bias techniques. Each has its own merits, but the experimental literature is largely dominated by the ramped technique.

16.4.1.1 Ramped Switching Current Measurement

As the name suggests, the bias current through the CPT is ramped linearly in time such that the current exceeds the maximum switching current and is then returned to zero bias for the next cycle. The probability of switching at a particular current can be backed out from the histogram of switching currents accumulated over many cycles (see Figure 16.10a). This ramped bias current measurement was originally done in the late 1970s in relatively large junctions (Fulton and Dunkleberger, 1974), but the methodology is also suitable to small Josephson junction devices such as the CPT (Joyez, 1995).

This method can be used for electrometry since it yields a switching current histogram whose mean value changes with gate, but its power lies in the fact that one can obtain the switching or "escape" rate $\gamma_{sw}(I_p, n_g)$ through a straightforward transformation of the whole histogram as in the original work by Fulton and Dunkleberger (1974). This is useful since the escape rate contains information about the electron temperature and how well the system is isolated from external noise (Devoret et al., 1987).

16.4.1.2 Quasiparticle Poisoning in the Ramped Current Measurement

If quasiparticle poisoning is present, the effective Josephson energy of the CPT flickers between two different values corresponding to even and odd parities. Each of these energies has its own escape rate $\gamma_{sw,e}$ and $\gamma_{sw,o}$, and the observed switching current in any given ramp cycle is determined by the instantaneous state of the system as the current is ramped. If the ramp rate is fast enough such that it can span the distance separating the odd and even switching currents faster than the odd/even lifetimes, then the ramped measurement becomes a snapshot of the parity state and we see that the switching currents group around two distinct values as shown in type L device histograms in Figure 16.12.

In our example, the modulation of the switching current histograms is similar to that of the type H device, except that it shows a distinct 1*e* shifted image corresponding to the presence of the odd state, particularly where we predict the island potential to be trap-like. This bimodal behavior is really only evident if the measurement is faster than the state lifetimes. If the current bias ramp rate were slowed down significantly, then the system could flicker back and forth between parity states many times and we would only see switching distributions grouped around whichever parity state had the lowest switching current. That is, we would see a purely 1*e* modulation in the switching current. In the early literature, the available dc measurement techniques were unable to capture the dynamical nature of the quasiparticle parity states, and the periodicity (1*e* versus 2*e*) of the modulation was the only handle with which to gauge whether poisoning was present. Because of the dynamics of the poisoning process, this correlation can be misleading.

In the type L example given here, the device is definitely poisoned, particularly where we expect the island potential to look trap-like. However, if we examine the type H device's switching current modulation, it is distinctly 2*e* periodic (Figure 16.11). The early picture of poisoning would naively assume that the poisoning is not present when, in fact, it is happening much faster than a simple analysis would indicate. We know, for instance, that the island potential is trap-like over a narrower range and that the trap potential is shallow compared with the operating temperature (this is due to the gap engineering for the H device). Therefore, quasiparticles that get trapped on the island

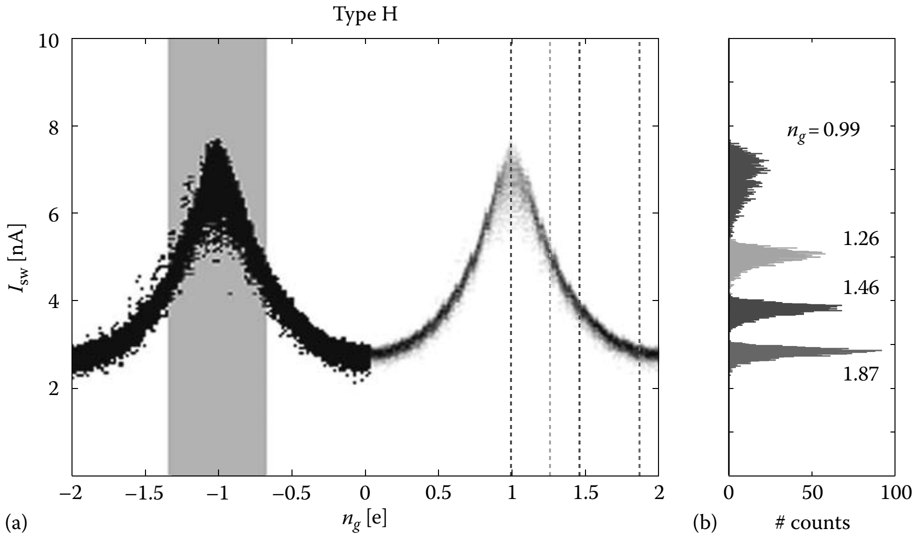

FIGURE 16.11 **(See color insert following page 20-14.)** (a) Type H I_{sw} histograms versus n_g. Histogram height is displayed in grayscale on the right-hand side, whereas all counts are displayed equally on the left-hand side. As in Figure 16.7, the gray box in (a) denotes regions where the island potential is trap like for quasiparticles. (b) Selected histograms corresponding to several gate voltages. Device parameters: $\Delta_i = 246\,\mu\text{eV}$, $\Delta_\ell = 205\,\mu\text{eV}$, $E_C \simeq 115\,\mu\text{eV}$, and $E_{J1} = E_{J2} \simeq 82\,\mu\text{eV}$.

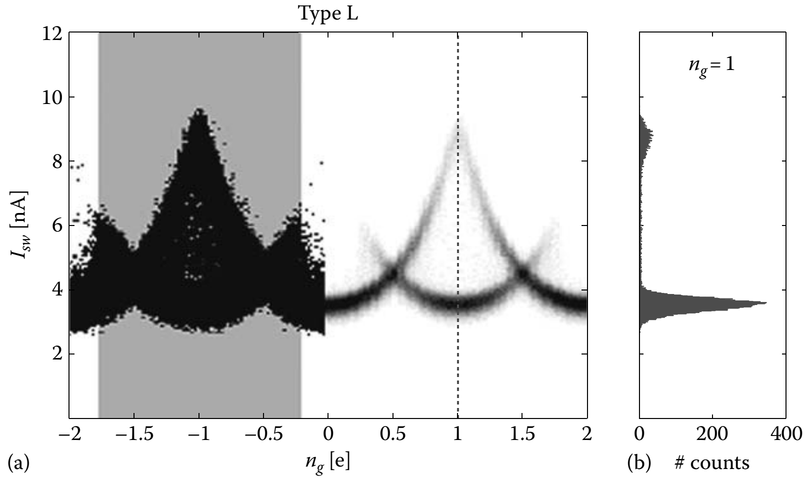

FIGURE 16.12 (a) Type L I_{sw} histograms versus n_g. Histogram height is displayed in grayscale on the right-hand side, whereas all counts are displayed equally on the left-hand side. As in Figure 16.7, the gray box in (a) denotes regions where the island potential is trap-like for quasiparticles. (b) Histogram at $n_g = 1$. Bimodal structure indicates relatively long-lived even and odd states. Device parameters: $\Delta_i = 205\,\mu\text{eV}$, $\Delta_\ell = 246\,\mu\text{eV}$, $E_C \simeq 115\,\mu\text{eV}$, and $E_{J1} = E_{J2} \simeq 78\,\mu\text{eV}$.

are rapidly ejected by thermal fluctuations. In this case, quasiparticles can jump in and out of the island before the CPT has time to latch into the free-running voltage state. In other words, the switching current measurement is bandwidth limited and is not able to see short-lived odd-parity events. While the average probability of being in the even state might be close to unity, quasiparticles may constantly be getting trapped and ejected at very rapid rates. In terms of the poisoning and ejection rates,

$$P_e = \frac{\Gamma_{oe}}{\Gamma_{eo} + \Gamma_{oe}}. \tag{16.23}$$

If the ejection rate Γ_{oe} is big compared with the poisoning rate Γ_{eo}, then P_e approaches one even though Γ_{eo} may be arbitrarily large itself. This is an insidious effect since it was previously common for many groups working in CPB qubits to assess their quasiparticle situation from what appeared to be clean 2*e* modulation characteristics. In fact, 2*e* modulation is observed in the presence of these fast quasiparticle dynamics when the measurement technique is slow in comparison. Since quasiparticles might have lifetimes far shorter than 1 μs, most available methods (dc and rf alike) can have problems with this.

In CPTs with deeper trap potentials, such as the type L devices we show here, the ejection rate can be slowed down significantly. Since the measurement shown in Figures 16.12 and 16.13c,d was performed with a ramp rate that was comparable to the quasiparticle ejection rate, it was possible to obtain a bimodal histogram of I_{sw}. In this case, it is easy to correlate the derived escape function with meaningful rates. The two slopes in Figure 16.13d are the even and odd state escape rates, $\gamma_{sw,e}$ and $\gamma_{sw,o}$, respectively. Thus, the bias current ramp is not infinitely fast, and the parity can flip from even to odd in the time it takes to ramp between the two current values corresponding to the odd and even state switching currents resulting in several switching events in the region between the peaks. These correspond to quasiparticles jumping into the CPT island during the bias ramp. Since the plateau that connects the even and odd switching rates is derived from the histogram counts between the even and odd peaks, it is a direct indication of the poisoning rate, Γ_{eo}. We can apply similar reasoning to the type H data in Figure 16.13a,b and notice that the escape rate for $n_g = 0.99$ is a little curvy, although the other escape rates shown look quite linear (on a semilogarithmic scale). While we do not develop a distinct plateau in this case, the deformation in the form of the escape rate is still an indication of very fast poisoning. This should not be a surprise since we expect the island to look trap-like at this gate voltage as we noted above. Thus, the lesson to be learned here is that despite appearances, the data shown in Figure 16.11 masks the fact that the type H device is poisoned.

16.4.1.3 Single Shot Measurement

In a single shot operation (Cottet et al., 2001), we pulse the current bias up from zero to some value I_p for a time τ_p. The probability of switching to the voltage state is given by

$$P_{sw}(n_g) = 1 - e^{-\gamma_{sw}(I_p, n_g)\tau}, \tag{16.24}$$

where $\gamma_{sw}(I_p, n_g)$ is the rate at which the CPT phase escapes to the free-running voltage state when instantaneously biased at I_p and n_g. If we are biased at n_g and wish to determine whether the polarization charge has shifted by δn_g, we need to figure out the likelihood of seeing a switching event that is actually due to the shift in switching probability versus just the probability of switching with no charge shift at all,

$$\Delta P_{sw}(n_g, \delta n_g) = e^{-\gamma_{sw}(I_p, n_g)\tau} - e^{-\gamma_{sw}(I_p, n_g + \delta n_g)\tau}. \tag{16.25}$$

If we want to measure this change in polarization charge in a single measurement, then we require that $\Delta P_{sw} = 1$. That is, if we pulse the bias current and see the CPT switch to the voltage state,

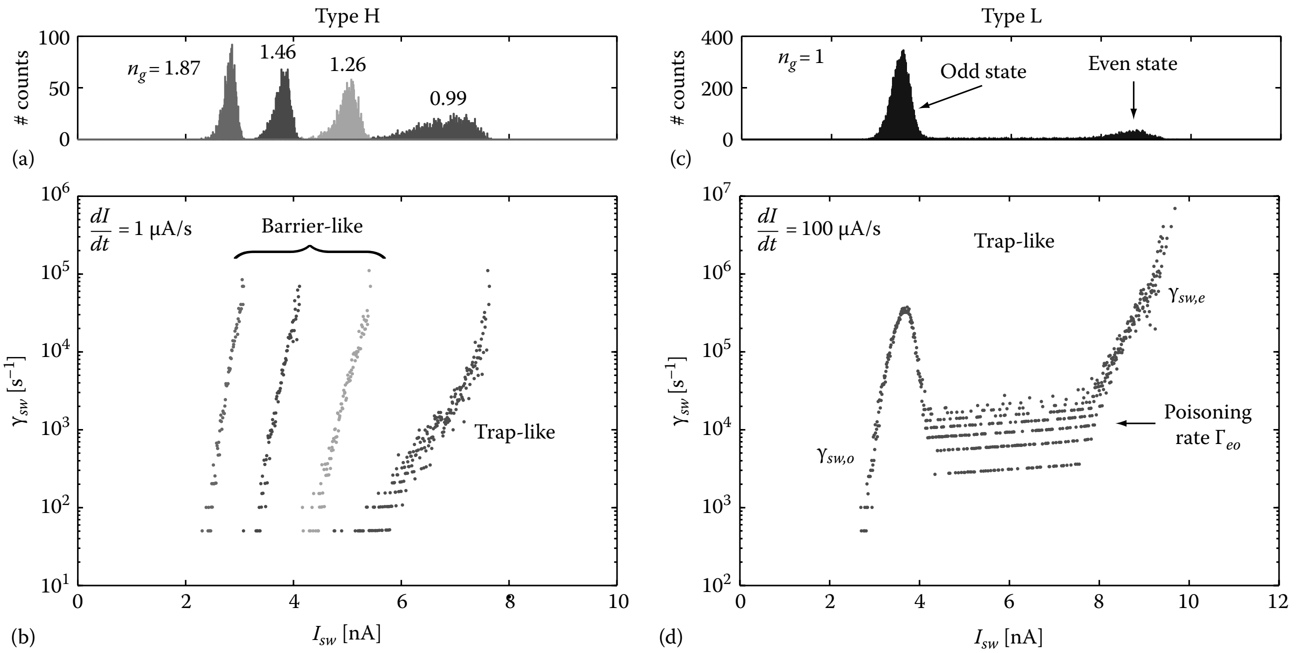

FIGURE 16.13 (See color insert following page 20-14.) (a,c) Switching current (I_{sw}) histograms and (b,d) derived switching/escape rates for a type H (barrier-like) and type L (trap-like) CPTs. For the type L device, the quasiparticle trapping behavior is evident in the bimodal I_{sw} distribution. In this case, the poisoning rate Γ_{eo} can be read directly from the derived escape rate in (d) as shown. Although the type H device is barrier-like for most n_g, it still looks like a trap near $n_g = 1$ (see Figure 16.7). This is apparent in the "curvy" structure of the escape curve for $n_g = 0.99$ as compared with the escape rates at other n_g in (b).

then the polarization charge has shifted by δn_g with certainty. This defines a minimum charge shift $\delta n_{g,ss}$ or voltage change $\delta V_{g,ss}$ ($= \delta n_{g,ss} e/C_g$) on the gate that can be detected with one measurement. If we want to detect a smaller change in gate voltage, then we would have to take multiple measurements until the uncertainty in the shift of $\langle I_{sw} \rangle$ is satisfactory. In many experiments, this is a completely reasonable approach. It is possible, however, to attempt to reduce the width of the switching current distribution by increasing the shunt capacitance across the CPT (increasing the effective mass of our imaginary phase particle) (Joyez, 1995).

In the end, pulsed single shot measurements have never been very popular. This may be because their chief use would have been in charge-based qubits, and other significant measurement schemes were shown to outperform it (Vion et al., 2002). However, it might still be a viable measurement method outside of superconducting quantum computing.

16.4.2 Zero-Biased rf Electrometry

The initial aim of the rf inductance measurements was to demonstrate a dissipationless (or near dissipationless) electrometer that could be used in charge-based quantum circuits such as the CPB, but the most useful application in recent years seems to have been to study the dynamics of quasiparticle tunneling.

Initial measurements of the Josephson (or "quantum") inductance of a CPT were performed by Sillanpää et al. (2004) and soon thereafter by Naaman and Aumentado (2006b). rf measurements of CPTs were also performed by Court and coworkers (Ferguson et al., 2006) as well, but these measurements focused on a modulation of the dissipation and were not strictly confined to the supercurrent branch.*

16.4.2.1 The rf Measurement Setup

Figure 16.14a shows a typical microwave circuit used in rf-CPT electrometry. In this scheme, the CPT is embedded in a tank circuit composed of the parasitic capacitance provided by the leads, the Josephson inductance of the CPT, and an extra surface mount inductor soldered to a printed circuit board near the chip to lower the tank resonance to the range of the microwave amplifier.†

Following the circuit path in Figure 16.14a, we reflect an incoming microwave signal (the *carrier*) off of the CPT resonator circuit and measure the amplitude and phase of the outgoing signal, that is, we measure the scattering parameter S_{11}. As shown

---

* It's important to note that measurement of the Josephson inductance (the second derivative of the ground state energy in phase) is conceptually linked to the "quantum capacitance" (the second derivative in charge). rf measurements of the capacitance have recently been shown to be an extremely useful measurement of charge-based qubits such as the CPB (Wallraff et al., 2004; Duty et al., 2005) and the transmon (Schuster, 2007).

† The word "parasitic" should be a tip-off to the reader that this scheme might be a little hit-or-miss and, indeed, it can be frustrating to target the operating frequency within 10% (this is the typical bandwidth for sub-1 GHz low-noise cryogenic amplifiers). While this setup is very similar to that used in typical rf-SET operation (Schoelkopf et al., 1998), it is not necessarily trivial to implement.

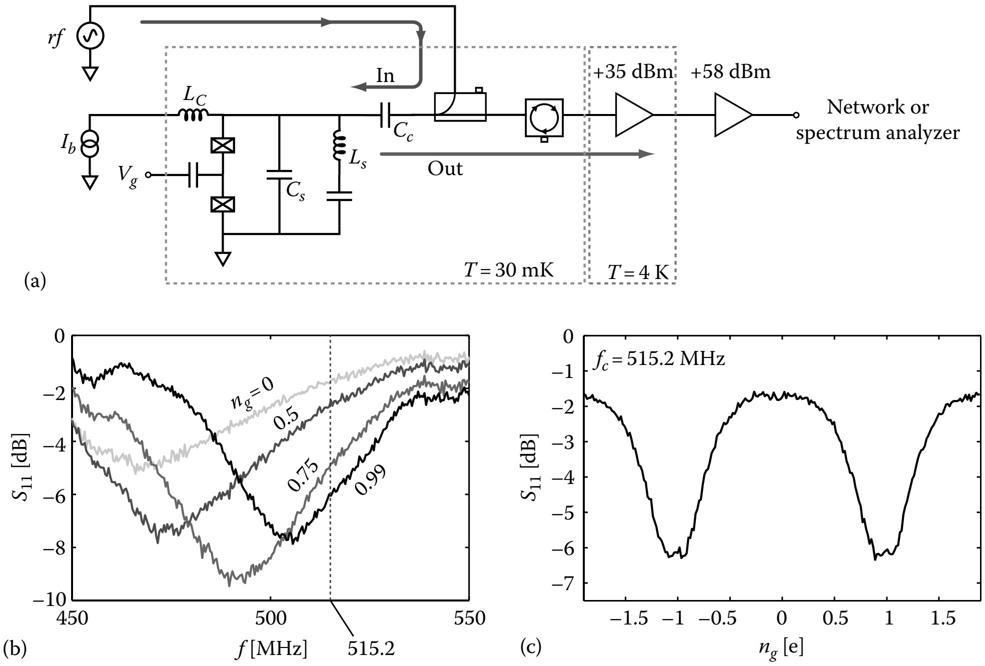

FIGURE 16.14 (a) Simplified rf-CPT measurement schematic. Incoming microwave power is directed toward CPT+ resonator through a directional coupler. The reflected wave is then amplified and measured at room temperature. In this schematic, several bandpass filters and attenuators have been omitted for simplicity. (b) Reflected power versus frequency at various gate voltages. (c) Reflected power modulation as a function of gate voltage for a fixed carrier frequency, f_c = 515.2 MHz. (Adapted from Naaman, O. and Aumentado, J., *Phys. Rev. B*, 73, 172504, 2006b.)

earlier, the gate-dependent Josephson inductance can be many 10s of nano-Henries and, since this inductance provides much of the total inductance available to the resonator, the frequency shift can be on the order of the bandwidth (typical Qs ~ 20–30). While a purely dissipationless circuit will reflect all of the signal ($|S_{11}|^2 = 1$) (Pozar, 2004), there are, in fact, a number of sources of dissipation, including the lossy traces on the pc board and leads on chip, the losses in the surface mount inductor and wire bonds, and, finally, the intrinsic CPT losses from any phase-diffusion resistance present. The end result is that there is a visible resonance dip in the reflected power amplitude $|S_{11}|^2$ (Figure 16.14b,c). For our purposes then, the frequency shift due to the modulation of the CPT Josephson inductance can be inferred from the reflected amplitude modulation.* That is, since the CPT inductance is a function of island charge, the final output microwave power amplitude and phase modulate with the gate-induced polarization charge.

Like the rf-SET, one can characterize the ultimate noise performance of this device as an electrometer in terms of an effective charge resolution in a 1 Hz bandwidth. In the rf measurement of the CPT inductance, the best number that has been reported is $\delta Q \sim 50\,\mu e/\sqrt{\text{Hz}}$ (Naaman and Aumentado, 2006b). That is, if we integrate the reflected power at the end of our measurement chain for 1 s, we can resolve a change in polarization charge at the gate electrode C_gV_g of 5.2×10^{-5} electrons. If we compare this with the best rf-SSET numbers $\delta Q < 5\,\mu e/\sqrt{\text{Hz}}$ (Brenning et al., 2006), the rf-CPT is more than 100 times slower at resolving the polarization charge. At first blush, this seems to put the rf-CPT at a disadvantage, but the rf-SSET charge resolution comes at the expense of using a relatively complex charge transport cycle that involves *both* quasiparticles and Cooper pairs. The back-acting voltage fluctuations of this cycle presented at the device they are measuring are equally nontrivial and have even been used to *cool* an electromechanical system that it was intended to measure (Naik et al., 2006). While some would consider this a feature of the rf-SSET, it seems to get away from the notion of simply wanting to use the device as a noninvasive electrometer. In fact, quasiparticle poisoning the rf-CPT has been used as a method to measure the effect of nearby SSETs, demonstrating that the latter emits nontrivial microwave power into its environment when voltage-biased (as they would be when operated as electrometers) (Naaman and Aumentado, 2007).

16.4.2.2 Operation Beyond 1 GHz

Although all of the published CPT experiments were operated with carrier frequencies below 1 GHz, this is not a fundamental requirement. In fact, this limitation seems to be determined by parasitic self-resonances in the surface-mount inductors, often used in these experiments. In principle, the operating frequency can be raised to many gigahertz using coplanar resonator techniques. The advantage of moving to higher frequencies is that wideband, low-noise HEMT amplifiers are now readily available in the 4–8 GHz range and all of the associated passive components (directional couplers and isolators) are smaller and much more common (less expensive).

---

* In principle, this information is also accessible through the phase of the outgoing signal and can be measured using an IQ mixer. Since loss was present in many of these experiments and the frequency shift was usually very obvious, measurement of the amplitude modulation has ended up being the simplest method.

16.4.2.3 Quasiparticle Poisoning in the rf-CPT

As in the switching current measurement, quasiparticle poisoning also has a dramatic effect on rf measurement. Unlike the switching current measurement, a qualitative understanding of the measurement response is very straightforward. Since the inductance of the CPT can switch instantaneously between two different parity states when the gate is biased into a trap-like regime, the reflected power also switches between two different amplitudes. This kind of response is more commonly known as "telegraph noise." A typical example is shown in Figure 16.15a for a type L device. We note that this is the same bimodal behavior that we observe in the switching current measurements except that we can sit at zero bias and watch quasiparticles jumping in and out of the CPT island. In Figure 16.15c, we histogram the telegraph time traces as a function of gate voltage, we get data that, unsurprisingly, are very similar to our bimodal switching current histograms in Figure 16.12.

The rates Γ_{eo} and Γ_{oe} can be derived from an analysis of these time-domain traces but requires a careful characterization of the system measurement bandwidth (Naaman and Aumentado, 2006a). Although it is possible to extract the poisoning rate Γ_{eo} from a switching current measurement, it is far more difficult to pull out an ejection rate. This is a direct consequence of the time-ordered nature of the current bias ramp. In contrast, the rf measurement is not burdened by this kind of time ordering in any obvious way, and both the poisoning and ejection rates are available.

The availability of the poisoning and ejection rates has provided further verification of the model presented in Aumentado et al. (2004), while confirming that nonequilibrium quasiparticles are generated in the leads at a constant rate below ~250 mK. In addition, the ejection rate Γ_{oe} has been used to validate the notion that biased SSETs emit nontrivial levels of microwave power into their environment (Naaman and Aumentado, 2007). This is important as for a long time rf-SSET electrometry had been considered a viable method of measuring CPB qubit charge states.

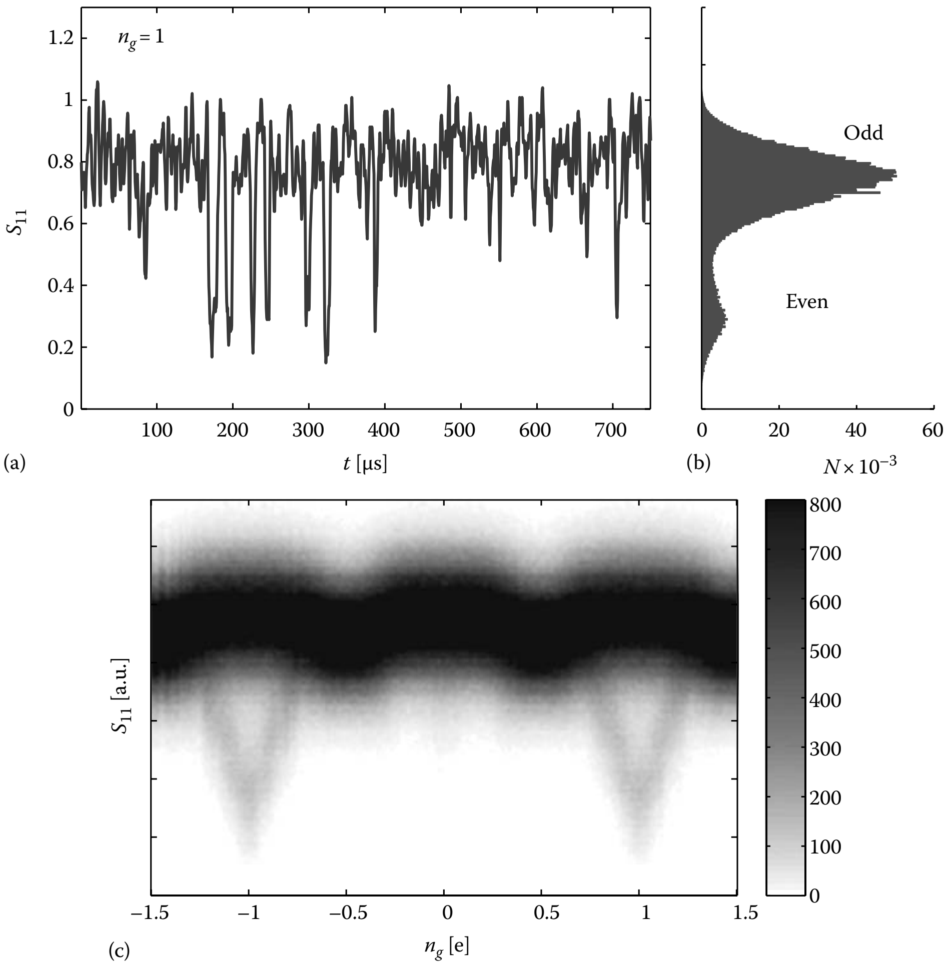

FIGURE 16.15 (a) Reflected power (linear units) of type L CPT at $n_g = 1$ in the time domain. The even (upper) and odd (lower) state levels are evident in telegraph-noise time traces. (b) Histogram of the full time trace. (c) Histograms versus gate voltage. (Adapted from Naaman, O. and Aumentado, J., *Phys. Rev. B*, 73, 172504, 2006b.)

16.4.2.4 Relation to the Cooper-Pair Box

Earlier, we alluded to the fact that the CPT was related to the CPB. In fact, when the CPT is biased at zero, it is identical to the CPB if the biasing circuit series impedance is small at the relevant parallel-junction plasma frequency. Since the problem of quasiparticle poisoning is also important to charge-based qubits such as the CPB (Nakamura et al., 1999; Wallraff et al., 2004), quantronium (Vion et al., 2002), and transmon (Koch et al., 2007), the fact that we can study it with relatively high instantaneous bandwidth in the CPT motivated several experiments aimed at studying the dynamics of the poisoning process in detail (Aumentado et al., 2004; Ferguson et al., 2006; Naaman and Aumentado, 2006b; Court et al., 2008b). Ultimately, it was realized that quantum capacitance measurements could achieve the same objective measuring CPBs directly and have yielded the most detailed quasiparticle poisoning studies to date (Lutchyn and Glazman, 2007; Shaw et al., 2008).

16.5 Present Status and Future Directions

The CPT is the simplest device in which dc transport characteristics can be correlated to the duality between charge and phase. Although the most prominent recent CPT experiments have focused on the problem of nonequilibrium quasiparticle poisoning, the CPT is also a promising general-purpose low-temperature, low-backaction electrometer. Since the most reliable techniques for fabricating these devices involve aluminum fabrication with gap, Josephson, and Coulomb blockade energies in the ~100–300 μeV range, operation is restricted to dilution refrigeration, so that the systems that one attaches it to must also be cold. While this seems restrictive at present, there are already several mesoscopic condensed matter systems that might benefit from minimally invasive fast electrometry.

Acknowledgment

The author wishes to acknowledge several important conceptual discussions with Ofer Naaman, Michel H. Devoret, and John M. Martinis.

References

Amar, A., Song, D., Lobb, C. J., and Wellstood, F. C. (1994). 2e and e periodic pair currents in superconducting coulomb-blockade electrometers. *Physical Review Letters*, 72(20):3234.

Ambegaokar, V. and Baratoff, A. (1963). Tunneling between superconductors. *Physical Review Letters*, 10(11):486.

Aumentado, J., Keller, M. W., Martinis, J. M., and Devoret, M. H. (2004). Nonequilibrium quasiparticles and 2e periodicity in single-cooper-pair transistors. *Physical Review Letters*, 92(6):066802.

Brenning, H., Kafanov, S., Duty, T., Kubatkin, S., and Delsing, P. (2006). An ultrasensitive radio-frequency single-electron transistor working up to 4.2 k. *Journal of Applied Physics*, 100:114321.

Clarke, J. and Braginski, A. I. (eds.). (2004). *The SQUID Handbook: Volume 1: Fundamentals and Technology of SQUIDs and SQUID Systems*. Wiley-VCH, Weinheim, Germany.

Cord, B., Dames, C., and Bergren, K. K. (2006). Robust shadow-mask evaporation via lithographically defined undercut. *Journal of Vacuum Science and Technology B*, 24(6):3139.

Cottet, A., Steinbach, A., Joyez, P., Vion, D., and Pothier, H. (2001). *Macroscopic Quantum Coherence and Quantum Computing*, p. 111. Kluwer Academic, Plenum Publishers, New York.

Court, N. A., Ferguson, A. J., and Clark, R. G. (2008a). Energy gap measurement of nanostructured aluminium thin films for single cooper-pair devices. *Superconductor Science and Technology*, 21(1):015013.

Court, N. A., Ferguson, A. J., Lutchyn, R., and Clark, R. G. (2008b). Quantitative study of quasiparticle traps using the single-cooper-pair transistor. *Physical Review B*, 77(10):100501.

Devoret, M. H. and Grabert, H. (1992). Introduction to single-charge tunneling. *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures (NATO Science Series: B)*, p. 1. Springer, New York.

Devoret, M. H., Esteve, D., Martinis, J. M., Cleland, A., and Clarke, J. (1987). Resonant activation of a brownian particle out of a potential well: Microwave-enhanced escape from the zero-voltage state of a Josephson junction. *Physical Review B*, 36(1):58.

Dolan, G. (1977). Offset masks for lift-off photoprocessing. *Applied Physics Letters*, 31(5):333.

Dolata, R., Scherer, H., Zorin, A., and Niemeyer, J. (2002). Single electron transistors with high-quality superconducting niobium islands. *Applied Physics Letters*, 80(15):2776.

Dolata, R., Scherer, H., Zorin, A., and Niemeyer, J. (2005). Single-charge devices with ultrasmall nb/alo/nb trilayer Josephson junctions. *Journal of Applied Physics*, 97:054501.

Duty, T., Johansson, G., Bladh, K., Gunnarsson, D., Wilson, C., and Delsing, P. (2005). Observation of quantum capacitance in the cooper-pair transistor. *Physical Review Letters*, 95(20):206807.

Ferguson, A. J., Court, N. A., Hudson, F. E., and Clark, R. G. (2006). Microsecond resolution of quasiparticle tunneling in the single-cooper-pair transistor. *Physical Review Letters*, 97(10):106603.

Flees, D., Han, S., and Lukens, J. (1997). Interband transitions and band gap measurements in bloch transistors. *Physical Review Letters*, 78(25):4817.

Fulton, T. and Dunkleberger, L. (1974). Lifetime of the zero-voltage state in Josephson tunnel junctions. *Physical Review B*, 9(11):4760.

Gunnarsson, D., Duty, T., Bladh, K., and Delsing, P. (2004). Tunability of a 2e periodic single cooper pair box. *Physical Review B*, 70(22):224523.

Josephson, B. (1962). Possible new effects in superconductive tunnelling. *Physics Letters*, 1(7):251.

Joyez, P. (1995). Le Transistor a une Paire de Cooper: un Systeme Quantique Macro-scopique. PhD thesis, L'Universite Paris, Paris, France.

Joyez, P., Lafarge, P., Filipe, A., Esteve, D., and Devoret, M. H. (1994). Observation of parity-induced suppression of Josephson tunneling in the superconducting single…. *Physical Review Letters*, 72(15):2458.

Kautz, R. and Martinis, J. (1990). Noise-affected iv curves in small hysteretic Josephson junctions. *Physical Review B*, 42(16):9903.

Koch, J., Yu, T. M., Gambetta, J. M., Houck, A. A., Schuster, D. I., Majer, J., Blais, A., Devoret, M. H., Girvin, S. M., and Schoelkopf, R. J. (2007). Charge insensitive qubit design from optimizing the cooper-pair box. *Physical Review A*, 74(4):042319.

Lutchyn, R. M. and Glazman, L. I. (2007). Kinetics of quasiparticle trapping in a cooper-pair box. *Physical Review B*, 75(18):184520.

Mazin, B. A. (2004). Microwave kinetic inductance detectors. PhD thesis, California Institute of Technology, Pasadena, CA.

Naaman, O. and Aumentado, J. (2006a). Poisson transition rates from time-domain measurements with a finite bandwidth. *Physical Review Letters*, 96(10):100201.

Naaman, O. and Aumentado, J. (2006b). Time-domain measurements of quasiparticle tunneling rates in a single-cooper-pair transistor. *Physical Review B*, 73(17):172504.

Naaman, O. and Aumentado, J. (2007). Narrow-band microwave radiation from a biased single-cooper-pair transistor. *Physical Review Letters*, 98(22):227001.

Naik, A., Buu, O., LaHaye, M. D., Armour, A. D., Clerk, A. A., Blencowe, M. P., and Schwab, K. C. (2006). Cooling a nanomechanical resonator with quantum back-action. *Nature*, 443:193.

Nakamura, Y., Pashkin, Y. A., and Tsai, J. S. (1999). Coherent control of macroscopic quantum states in a single-cooper-pair box. *Nature*, 398:786.

Palmer, B. S., Sanchez, C. A., Naik, A., Manheimer, M. A., Schneiderman, J. F., Echternach, P. M., and Wellstood, F. C. (2007). Steady-state thermodynamics of nonequilibrium quasiparticles in a cooper-pair box. *Physical Review B*, 76(5):054501.

Pozar, D. M. (2004). *Microwave Engineering*, 3rd edn. Wiley, New York.

Savin, A. M., Meschke, M., Pekola, J. P., Pashkin, Y. A., Li, T. F., Im, H., and Tsai, J. S. (2007). Parity effect in Al and Nb single electron transistors in a tunable environment. *Applied Physics Letters*, 91:063512.

Schoelkopf, R. J., Wahlgren, P., Kozhevnikov, A. A., Delsing, P., and Prober, D. E. (1998). The radio-frequency single-electron transistor (rf-set): A fast and ultrasensitive electrometer. *Science*, 280:1238.

Schuster, D. I. (2007). Circuit quantum electrodynamics. PhD thesis, Yale University, New Haven, CT.

Shaw, M. D., Lutchyn, R. M., Delsing, P., and Echternach, P. M. (2008). Kinetics of nonequilibrium quasiparticle tunneling in superconducting charge qubits. *Physical Review B*, 78(2):024503.

Sillanpää, M., Roschier, L., and Hakonen, P. (2004). Inductive single-electron transistor. *Physical Review Letters*, 93(6):066805.

Tinkham, M. (2004). *Introduction to Superconductivity*, 2nd edn. Dover, New York.

Tinkham, M., Hergenrother, J. M., and Lu, J. G. (1995). Temperature dependence of even-odd electron-number effects in the single-electron transistor. *Physical Review B*, 51(18):12649.

Townsend, P., Taylor, R., and Gregory, S. (1972). Superconducting behavior of thin-films and small particles of aluminum. *Physical Review B*, 5(1):54.

Tuominen, M. T., Hergenrother, J. M., Tighe, T. S., and Tinkham, M. (1992). Experimental evidence for parity-based 2e periodicity in a superconducting single-electron tunneling transistor. *Physical Review Letters*, 69(13):1997.

Tuominen, M. T., Hergenrother, J. M., Tighe, T. S., and Tinkham, M. (1993). Even-odd electron number effects in a small superconducting island: Magnetic-field dependence. *Physical Review B*, 47(17):11599.

Vion, D., Aassime, A., Cottet, A., Joyez, P., Pothier, H., Urbina, C., Esteve, D., and Devoret, M. H. (2002). Manipulating the quantum state of an electrical circuit. *Science*, 296:886.

Wallraff, A., Schuster, D. I., Blais, A., Frunzio, L., Huang, R.-S., Majer, J., Kumar, S., Girvin, S. M., and Schoelkopf, R. J. (2004). Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature*, 431:162.

Watanabe, M., Nakamura, Y., and Tsai, J. (2004). Circuit with small-capacitance high-quality Nb Josephson junctions. *Applied Physics Letters*, 84(3):410.

Yamamoto, T., Nakamura, Y., Pashkin, Y. A., Astafiev, O., and Tsai, J. S. (2006). Parity effect in superconducting aluminum single electron transistors with spatial gap profile controlled by film thickness. *Applied Physics Letters*, 88(21):212509.

III

Nanolithography

17 **Multispacer Patterning: A Technology for the Nano Era** *Gianfranco Cerofolini, Elisabetta Romano, and Paolo Amato* **17**-1
Introduction • The Crossbar Process • Nonlithographic Preparation of Nanowires • Multispacer Patterning Techniques • Influence of Technology on Architecture • Fractal Nanotechnology • Appendixes • Abstract Technology • Concrete Technology • References

18 **Patterning and Ordering with Nanoimprint Lithography** *Zhijun Hu and Alain M. Jonas* **18**-1
Introduction • Nanoimprint Lithography • Nanoshaping Functional Materials by Nanoimprint Lithography • Controlling Ordering and Assembly Processes by Nanoimprint Lithography • Conclusions and Perspectives • Acknowledgments • References

19 **Nanoelectronics Lithography** *Stephen Knight, Vivek M. Prabhu, John H. Burnett, James Alexander Liddle, Christopher L. Soles, and Alain C. Diebold* **19**-1
Introduction • Photoresist Technology • Deep Ultraviolet Lithography • Electron-Beam Lithography • Nanoimprint Lithography • Metrology for Nanolithography • Abbreviations • References

20 **Extreme Ultraviolet Lithography** *Obert R. Wood II* **20**-1
Introduction • Historical Background • Current State of the Art • EUVL Infrastructure Status • Future Perspectives • Abbreviations • References

17

Multispacer Patterning: A Technology for the Nano Era

17.1 Introduction **17**-1
17.2 The Crossbar Process **17**-2
17.3 Nonlithographic Preparation of Nanowires.. **17**-3
Imprint Lithography • Spacer Patterning Technique
17.4 Multispacer Patterning Techniques.. **17**-5
Additive Route—S^nPT_+ • Multiplicative Route—$S^nPT_\times$ • Three-Terminal Molecules
17.5 Influence of Technology on Architecture **17**-10
Addressing • Comparing Crossbars Prepared with Additive or Multiplicative Routes • Applications—Not Only Nanoelectronics
17.6 Fractal Nanotechnology **17**-12
Fractals in Nature • Producing Nanoscale Fractals via $S^nPT_\times$
Appendixes.. **17**-14
17.A Abstract Technology **17**-14
Bodies and Surfaces • Conformal Deposition and Isotropic Etching • Directional Processes • Selective Processes
17.B Concrete Technology **17**-18
References.. **17**-19

Gianfranco Cerofolini
University of Milano-Bicocca

Elisabetta Romano
University of Milano-Bicocca

Paolo Amato
Numonyx and University of Milano-Bicacca

17.1 Introduction

The evolution of integrated circuits (ICs) has been dominated by the idea of scaling down its basic constituent—the metal-oxide-semiconductor (MOS) field-effect transistor (FET). In turn, this has required the development of suitable lithographic techniques for its definition on smaller and smaller length scales. There are several generations of lithographic techniques, usually classified according to the technology required for the definition of the wanted features on photo- or electro-sensitive materials (resists): standard photolithography (436 nm, Hg *g*-line; refractive optics), deep ultraviolet (DUV) photolithography (193 nm, ArF excimer laser; refractive optics), immersion DUV photolithography (refractive optics), extreme ultraviolet (EUV) photolithography (13.5 nm, plasma-light source; reflective optics), and electron beam (EB) lithography (electron wavelength controlled by the energy, typically in the interval 10^{-3} to 10^{-2} nm).

The industrial system has succeeded in that, but the cost of ownership has in the meanwhile dramatically increased, because of either the required investment per machine,

$$\text{DUV} \lll \text{immersion DUV} \lll \text{EUV},$$

or the throughput,

$$\text{EB} \lll \text{DUV}.$$

It is just the dramatic cost escalation that is necessary for the reduction of the feature size that casts doubts on the possibility of continuing the current increase of IC density beyond the next 10 years.

Entirely new revolutionary technological device platforms, overcoming the complementary MOS (CMOS) paradigm, must likely be developed to enable the economical feasibility and scalability of electronic circuits to the tera scale intergation (TSI).

On another side, the preparation of ICs with bit density as high as 10^{11} cm^{-2} seems now possible, with modest changes in the current production process and marginal investment for the fabrication facility, within a different paradigm. The new paradigm is based on a structure, where a crossbar embodies in each of its cross-points a functional material able to perform by itself the functions of a memory cell [1].

The crossbar is indeed producible (via nonconventional lithography or even without any lithographic method) with geometry on the 10 nm length scale. Although the crossbar is not yet a circuit, it may nonetheless become a circuit if each cross-point contains a memory cell and each of them can be addressed, written, and read—that requires an external circuitry for addressing, power supply, and sensing. The best architecture for satisfying those functions is manifestly achieved embedding the crossbar in a CMOS circuitry [2]:

$$\begin{aligned}\text{TSI IC} = {} & \text{submicro CMOS IC} \cup \\ & \text{nano crossbar} \cup \\ & \text{nanoscopic cells.}\end{aligned}$$

This architecture reduces the problem of preparing TSI ICs to that of producing nanoscopic memory cells and inserting them into the cross-points of a crossbar structure. It would be a mere declaration of will were it not for the fact that molecules by themselves able to behave as memory cells have been not only designed [3–5] and synthesized [6,7], but also inserted in hybrid devices [8–12]. This fact opens immediately the possibility of a *hybrid route* to TSI ICs:

$$\begin{aligned}\text{Hybrid TSI IC} = {} & \text{submicro CMOS IC} \cup \\ & \text{nano crossbar} \cup \\ & \text{grafted functional molecules.}\end{aligned} \tag{17.1}$$

In this approach, the transport properties of programmable molecules are exploited for the preparation of externally accessible circuits. Because of this, it is usually referred to as molecular electronics.

The hypothesized TSI IC has thus a hybrid structure, formed by a nanometer-sized kernel (the functionalized crossbar), linked to a conventional submicrometer-sized CMOS control circuitry (producible with currently achievable technologies) and hosting molecular devices (whose production is left to chemistry).

17.2 The Crossbar Process

That self-assembled monolayers on preformed gold contacts may behave as nanoscale memory elements was demonstrated for thiol-terminated π-conjugated molecules containing amino or nitro groups [9]. The first demonstration of nonvolatile molecular crossbar memories employed self-assembled monolayers of thiol-terminated rotaxanes as reprogrammable cells [10]. The molecules were embedded between the metal layers forming the crossbar via a process that can be summarized as follows:

XB^1, deposition and definition of the first-level ("bottom") wire array

XB^2, deposition of the active reconfigurable molecules, working also as vertical spacer separating lower and upper arrays

XB^3, deposition and definition of the second-level ("top") wire array

Figure 17.1 sketches the XB process.

Although potentially revolutionary, the XB approach, with double metal strips, has been found to have serious limits:

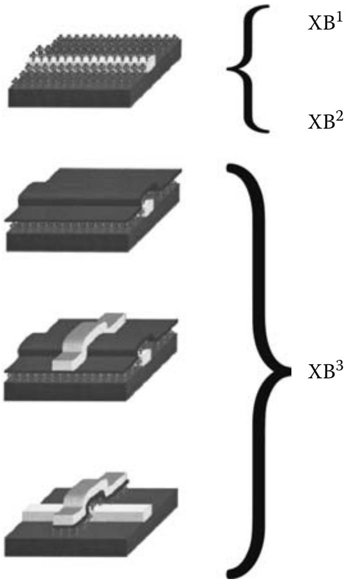

FIGURE 17.1 The first proposed crossbar-architecture fabrication steps. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

- The organic active element is incompatible with high-temperature processing, so that the top layer must be deposited in XB^3 at room or slightly higher temperature. This need implies a preparation based on physical vapor deposition, where the metallic electrode results from the condensation of metal *atoms* on the outer surface of the deposited organic films. This process, however, poses severe problems of compatibility, because isolated metal atoms, quite irrespective of their chemical nature, are mobile and decorate the molecule, rather than being held as a film at its outer extremity [13–15].
- A safe determination of the conductance state of bistable molecules requires the application of a voltage V appreciably larger than k_BT/e (with k_B being the Boltzmann constant, T the absolute temperature, and $-e$ the electron charge), say $V = 0.1–0.2$ V. Applied to molecules with typical length around 3 nm, this potential sustains an electric field, of the order of 5×10^5 V cm^{-1}, sufficiently high to produce metal electromigration along the molecules [16].
- The energy barrier for metal-to-molecule electron transfer is controlled by the polarity of the contact, in turn increasing with the electronegativity difference along the bond linking metal and molecule [17]. The use of thiol terminations for the molecule, as implicit in the XB approach, is expected to be responsible for high-energy barriers because of the relatively high electronegativity of sulfur.

Even though the first difficulty can in principle be removed by slight sophistication of the process (for instance, as follows: spin-coating the organic monolayer with a dispersion of metal nanoparticles in a volatile solvent, evaporating the solvent, forming a relatively compact layer via coalescence of the metal particles, and compacting the resulting film by means of an additional amount of PVD metal), other difficulties are more fundamental in nature and require different materials.

A solution to the electromigration problem can be achieved preparing the bottom electrodes in the form of silicon wires

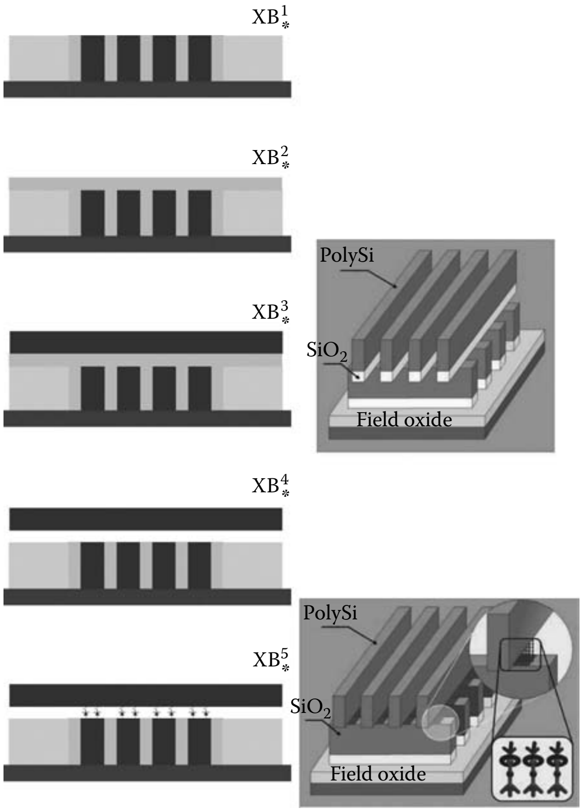

FIGURE 17.2 The basic idea of XB^*: preparing the crossbar before its functionalization.

(as done in [12]), and the top electrodes in the form conducting π-conjugated polymers (as suggested in [18]):

- XB_+^1, deposition and definition of the bottom array of poly-silicon wires
- XB_+^2, deposition of the active (reconfigurable) element, working also as vertical spacer separating lower and upper arrays
- XB_+^3, deposition and definition of the top array of conducting π-conjugated polymers

The use of poly-silicon as material for the top array too seems impossible because it is prepared almost uniquely via chemical vapor deposition at incompatible temperatures with organic molecules.* The only way to overcome this difficulty consists thus in a process, $XB_\star$, where the two poly-silicon arrays defining the crossbar matrix are prepared *before* the insertion of the organic element [11]. Preserving a constant separation on the nanometer length scale is possible only via the growth of a sacrificial thin film on the first array before the deposition of the second one [21]:

- $XB_\star^1$, preparation of a bottom array of poly-silicon wires
- $XB_\star^2$, deposition of a sacrificial layer as vertical spacer separating lower and upper arrays
- $XB_\star^3$, preparation of a top array of poly-silicon wires crossing the first-floor array
- $XB_\star^4$, selective chemical etching of the spacer
- $XB_\star^5$, insertion of the reprogrammable molecules in a way to link upper and lower wires in each cross-point

The basic idea of process $XB_\star$, of inserting the functional molecules after the preparation of the crossbar, is sketched in Figure 17.2.

Of the three considered processes (XB, XB_+, and $XB_\star$), the one based on double-silicon strips is certainly the most conservative one and is thus expected to be of easiest integration in IC processing. For this reason (and for the possibility of using three-terminal molecules, see Section 17.4.3), our attention will be concentrated on the $XB_\star$ route.

17.3 Nonlithographic Preparation of Nanowires

The preparation of a crossbar requires the use of simple geometries—essentially arrays of dielectrically insulated conductive wires. What is especially interesting is that wire arrays with pitch on the nanometer length scale are producible via nonlithographic techniques (NLTs). Not only is this preparation possible, but also the wire linear density already achieved with the NLTs described in the following is smaller than the one achievable via the most advanced EUV or EB lithographies. Such NLTs exploit the following features:

- (V), the "vertical" control of film thickness, possible down to the subnanometer length scale provided that the film is sufficiently homogeneous.
- (V-to-H), the transformation of films with "vertical" thickness t into patterns with "horizontal" width w:

$$t \xrightarrow{\text{NLT}} w.$$

These techniques are *imprint lithography* and two variants of the *multispacer patterning technique*.

17.3.1 Imprint Lithography

Imprint lithography (IL) is a contact lithography where properties (V) and (V-to-H) are exploited for the preparation of the mask. The process is essentially based on the sequential alternate deposition of two films, A and B, characterized by the existence of a preferential etching for one (say A) of them. After cutting at 90°, polishing, and controlled etching of A, one eventually gets a mask formed by nanometer-sized trenches running parallel to one another at a distance fixed by the thickness of B [22,23]. For instance, a contact mask for imprint lithography with pitch of 16 nm was prepared by growing on a substrate a quantum well via molecular beam epitaxy, cutting the sample perpendicularly to the surface, polishing the newly exposed surface, and etching selectively the different strata of the well [23]. The potentials of the preparation method based on superlattice nanowire pattern transfer (SNAP) are reviewed in Ref. [24].

---

[*] It is instead possible if the spacer is an inorganic film. This case, although not for molecular electronics, is interesting for memories based on phase-change materials [19] or mimicking the memristor [20].

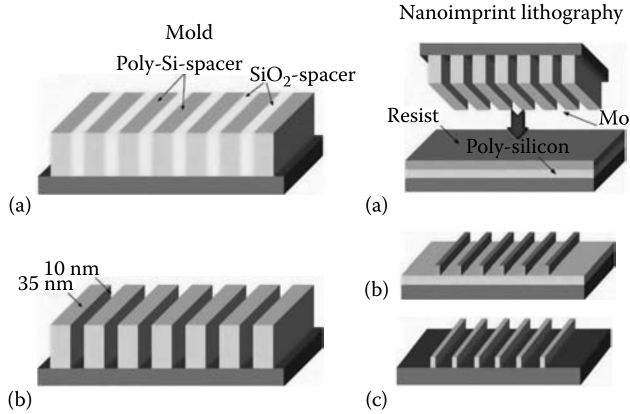

FIGURE 17.3 Preparation of mold for imprint lithography (left) and its use as contact mask (right); the multilayer has been supposed to be produced with cycles of sequential depositions of silicon and SiO_2. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

Actually a number of variants for transferring the pattern to the surface have been developed: molding, embossing, and stamping are the ones most frequently considered [25]. In one of them (molding), after filling the trenches with a suitable polymer, the mask is used as a stamp, pressing it onto the surface; if the polymer has a higher affinity for the surface than for the mask, the pattern is transferred to the surface when the mask is eventually removed [25]. The transfer of the polymer to the surface is possible without loss of geometry only if the trench is sufficiently shallow; this implies that the polymer must sustain a subsequent process where it is used as mask for the definition (via directional etching) of the underlying structure with a high aspect ratio. Another method (embossing), sketched in the right-hand side of Figure 17.3, involves the pressure-induced transfer of the pattern from the mask to a plastic film and its subsequent polymerization.

Imprint lithography is generally believed to have potential advantages over conventional lithography because it can be carried out in ordinary laboratory with no special equipments [26]. This situation is expected to make it easy to run along the learning curve to a mature technology. However, very little is known about the overall yield, eventually resulting in production cost, of this process (preparation of mask and stamp, imprint, etching) when the geometries are on the length scale of tens of nanometers.

Imprint lithography has been the matter of extended investigation (see for instance, Refs. [25,26]) and will not be discussed here. Rather, this chapter is devoted to describe the multispacer patterning technique and to compare its two variants.

17.3.2 Spacer Patterning Technique

The multispacer patterning technique (S^nPT) is essentially based on the repetition of the spacer patterning technique (SPT). In turn, the SPT is an age-old technology originally developed for the dielectric insulation of metal electrodes contacting source and drain from the gate of metal-oxide-semiconductor (MOS) transistors.

The SPT involves the following steps:

SPT^0, the *lithographic definition* of a seed with sharp edge and high aspect ratio
SPT^1, the *conformal deposition* on this feature of a film of uniform thickness
SPT^2, the *directional etching* of the film until the original seed surface is exposed

If the process is stopped at this stage, it results in the formation of sidewalls of the original seed; otherwise, if

SPT^3, the original seed is removed via a *selective etching*

what remains is constituted only by the walls of the seed edges. Figure 17.4 sketches the various stages of SPT.

This technique has been demonstrated to be suitable for the preparation of features with minimum size of 7 nm [27,28] and has already achieved a high level of maturity, succeeding in the definition, with yield very close to unity, of nanoscopic bars with high aspect ratio.

The SPT may be sophisticated via the deposition of a multilayered film; Figure 17.5 shows the sidewalls resulting from the deposition of a multilayer and compares it with what is really done in the original application of SPT—the insulation of the source and drain electrodes from the gate [29].

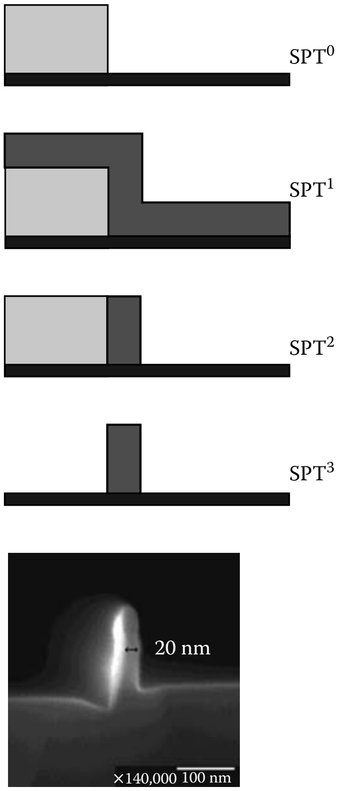

FIGURE 17.4 Up: The spacer patterning technique: SPT^0, definition of a pattern with sharp edges; SPT^1, conformal deposition of a uniform film; SPT^2, directional etching of the deposited film up to the appearance of the original seed; and SPT^3, selective etching of the original feature. Down: Cross-section of a wire produced via SPT. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

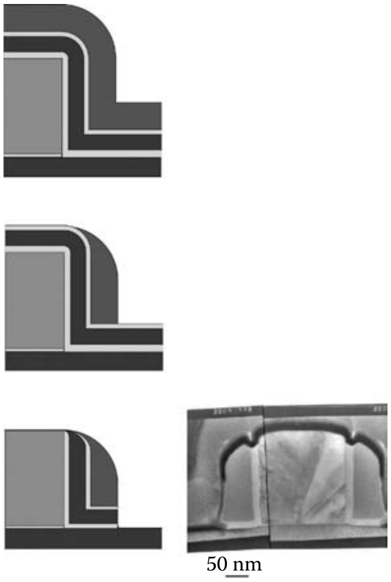

FIGURE 17.5 The original application of the SPT in microelectronics—dielectric insulation of the gate from source and drain electrodes. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

17.4 Multispacer Patterning Techniques

Two S^nPT routes have been considered: the additive (S^nPT_+) and multiplicative ($S^nPT_\times$) routes.

The S^nPT_+ is recent and was proposed having in mind the preparation of crossbars for molecular electronics [30–32]. The $S^nPT_\times$ is instead much older: the first demonstrators were developed for the generation of gratings with sub-lithographic period [33]; recently, however, this technique has been used for the preparation of wire arrays in biochips also [28]. Since no detailed comparison of the limits and relative advantages of these techniques is known, the following part will try to provide an understanding about them on the basis of fundamental considerations.

17.4.1 Additive Route—S^nPT_+

The S^nPT_+ is substantially based on n STP repetitions where *the original seed is not removed and each free wall of newly grown bars is used as a seed for the subsequent STP*. Each SPT_+ cycle starts from an assigned seed and proceeds with the following steps:

$S^nPT_+^1$, conformal deposition of a conductive material
$S^nPT_+^2$, directional etching of this material up to the exposure of the original seed
$S^nPT_+^3$, conformal deposition of an insulating material
$S^nPT_+^4$, directional etching of this material up to the exposure of the original seed

The basic idea of the S^nPT_+ is shown in Figure 17.6: the upper part sketches the process; the lower part shows instead how poly-silicon arrays separated by SiO_2 dielectrics with sub-lithographic pitch (35 nm) can indeed be produced [30–32].

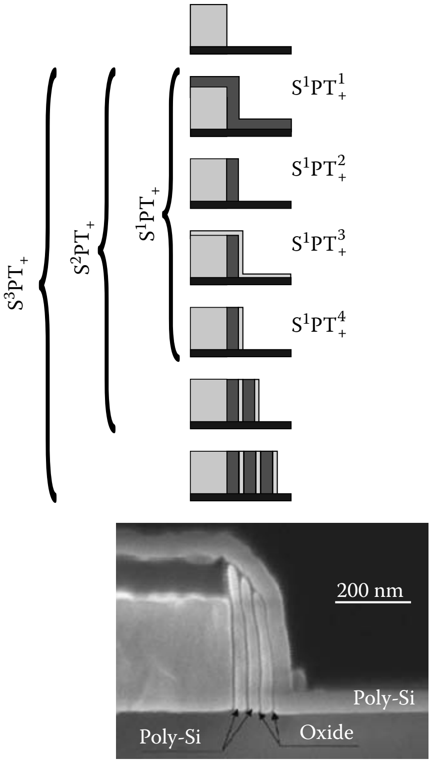

FIGURE 17.6 Up: the additive multispacer patterning technique. Down an example of S^3PT_+ multispacer (with pitch of 35 nm and formed by a double layer poly-Si|SiO_2) resulting after three repetitions of the SPT_+. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

The sketch in Figure 17.6 shows a process in which lines are additively generated onto a progressively growing seed, preserving the original lithographic feature along the repetitions of the unit process. The unit process is based on two conformal depositions of uniform layers (poly-silicon and SiO_2) each followed by a directional etching.

Figure 17.7 shows, however, that a similar structure could be obtained by a cycle formed by

S^nPT_+', conformal deposition of a bilayer film (formed by an insulating layer deposited before the conductive one—the order of deposition is fundamental)
S^nPT_+'', directional etching of this film up to the exposure of the original seed

Consider an array with pitch P of lithographic seeds, each with width W (and thus separated from one another by a distance P–W). Denote with the same symbols in lower case, p and w, the corresponding sub-lithographic quantities. Starting from the said array of lithographic seeds, after n repetitions of SPT_+ any seed is surrounded by $2n$ lines (an example with $n = 4$ is shown in Figure 17.8), so that the corresponding effective linear density K_n of spacer bars is given by

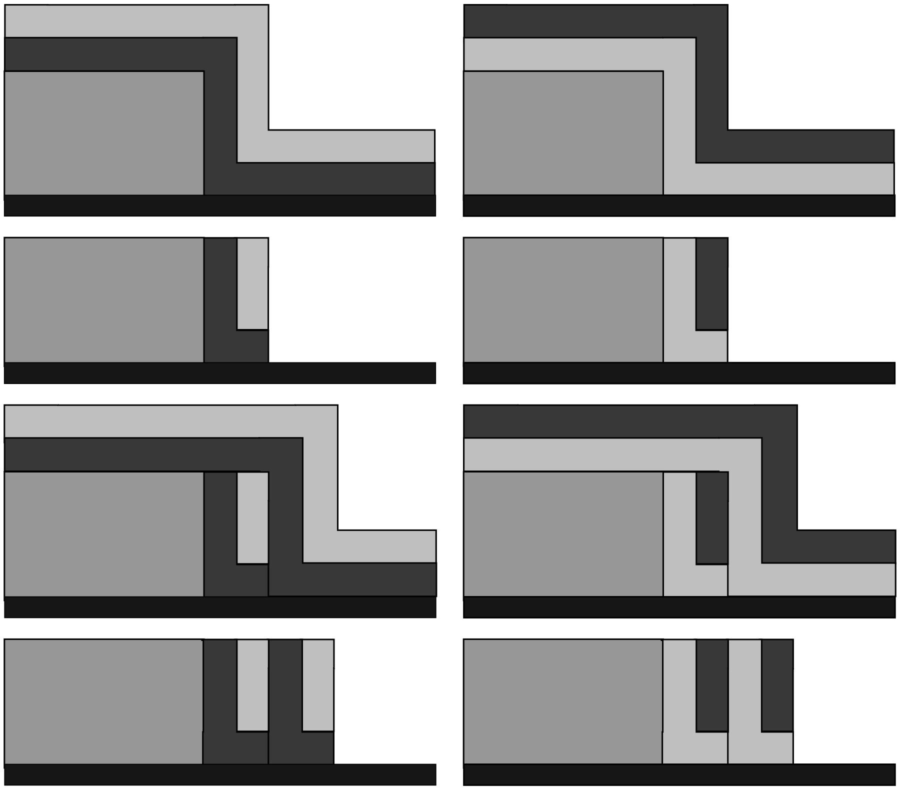

FIGURE 17.7 A variant of the additive multispacer patterning technique able to reduce the number of directional etching by a factor of 2 via sequential deposition of a bilayered film. Two possibilities are considered, consisting in the deposition first of either poly-silicon and then of SiO_2 (left), or the same layers in reverse order (right). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

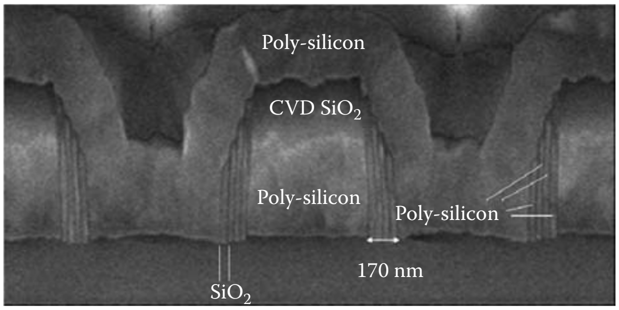

FIGURE 17.8 An example of S^4PT_+, showing the construction of four silicon bars per side of the seed. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

$$K_n = \frac{2n}{P}.$$

The example of Figure 17.8 shows also that the process can be tuned to preserve the constancy with n of wire width w_n, $w_n = w$, and pitch p_n, $p_n = p$. On the contrary, the spacer heights s_n decrease almost linearly with n,

$$s_n \simeq s_0 - \tau n \tag{17.2}$$

(with τ the spacer height loss per SPT cycle), at least for n lower than a characteristic value $n^{\max}$.

This unavoidable decrease is ultimately due to the fact that the conformal coverage of a feature with high aspect ratio results necessarily in a rounding off of the edge shape with curvature radius equal the film thickness and that the subsequent directional etching produces a nonplanar surface. Figure 17.9 explains the reasons for the shape of the resulting bars: even assuming a perfect directional attack, the resulting spacer is not flat and there is a loss of height τ not smaller than t: $\tau \geq t$.

Figure 17.6 shows that the process can actually be controlled to have τ coinciding, within error, with its minimum theoretical value, $\tau = 1.0t$. Figure 17.8 shows however that τ depends on the process, and this can be tuned to have $\tau = 3.2t$.

Although the loss of height may seem a disadvantage, in Section 17.5.1 it will be shown how the controlled decrease of s_n with n may be usefully exploited.

Assuming the validity of Equation 17.2 until s_n vanishes, the maximum number $n^{\max}$ of SPT_+ repetitions is given by $n^{\max} = s_0/\tau$; after $n^{\max}$ SPT_+ repetitions, the seed is lost and the process cannot continue further. In view of the availability of techniques for the production of deep trenches with very high aspect ratios, in this analysis s_0 (and hence $n^{\max}$) may be regarded as almost a free parameter.

The optimum distance allowing the complete filling of the void regions separating the original lithographic seeds is therefore given by $P - W = 2n^{\max}p$. Hence, the maximum number of cross-points that can be arranged in any square of side P is given

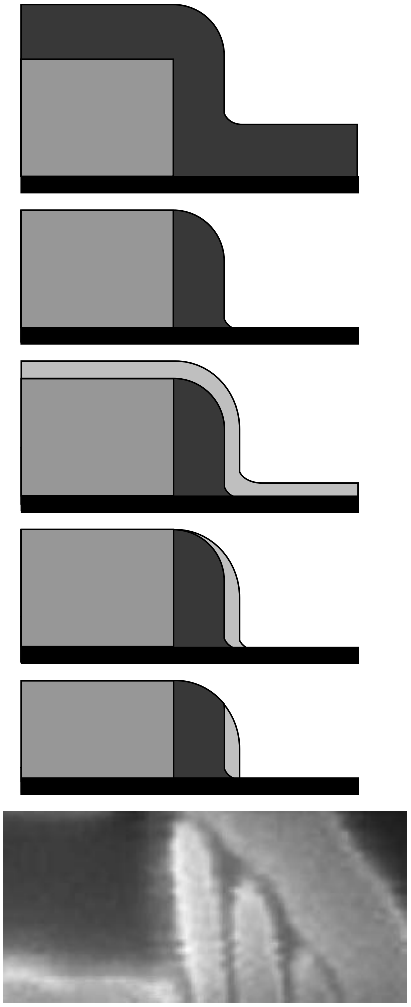

FIGURE 17.9 Shape of a sidewall resulting after ideal conformal deposition and directional etching (top; five sketches) and an image of how it results in practice (bottom; magnification of the spacers shown in Fig. 6). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

by $(2n^{\max})^2$, and the maximum effective cross-point density $\delta_+^{\max}$ achievable with the S^nPT_+ is given by

$$\delta_+^{\max} = \left(\frac{2n^{\max}}{P}\right)^2$$
$$= \frac{1}{p^2}\left(\frac{1}{1+W/2n^{\max}p}\right)^2. \qquad (17.3)$$

Equation 17.3 shows that $\delta_+^{\max}$ depends on the lithography (through W) and on the sub-lithographic technique S^nPT_+ (through $n^{\max}$ and p).

Just to give an idea of the maximum obtainable density, Figure 17.6 shows that $p = 35\,\text{nm}$ has already been achieved and $n^{\max} \simeq 10$ is at the reach of the S^nPT_+; for $W = 0.1\,\mu\text{m}$ (characteristic value for IC high volume production) and $p = 30\,\text{nm}$, Equation 17.3 gives $\delta_+^{\max} \simeq 8\times10^{10}\ \text{cm}^{-2}$. The comparison of this prediction with the lithographically achievable cross-point density (currently of about $2 \times 10^9\ \text{cm}^{-2}$), shows that S^nPT_+ allows the cross-point density to be magnified by a factor of about 40. This is however achieved only with the construction of 10 consecutive spacers per (bottom and top) layer. The spacer technology is a mature technology with yields close to unity also when employed in more complex geometries than single lines. The increase of processing cost implied by its repeated application in IC processing has been discussed in Ref. [34]. Although this increase is moderate, the integration of so many SPT_+ cycles may, however, be not trivial and passes through the development of dedicated cluster tools.

It is however noted that the basic idea sketched in Figure 17.7 can be extended to remove this difficulty at least partially: the conformal deposition of a slab with n poly-silicon|insulator bilayers (the insulator being SiO_2, Al_2O_3,…) followed by its directional etching would indeed result in the formation of $2n$ dielectrically insulated poly-silicon wires. Figure 17.10 sketches the process.

The figure however shows that this dramatic simplification of the process can be done only for the deposition of the bottom array; its use for the top array would result in a distance of the top wire from the lower one varying with the order of poly-silicon layer.

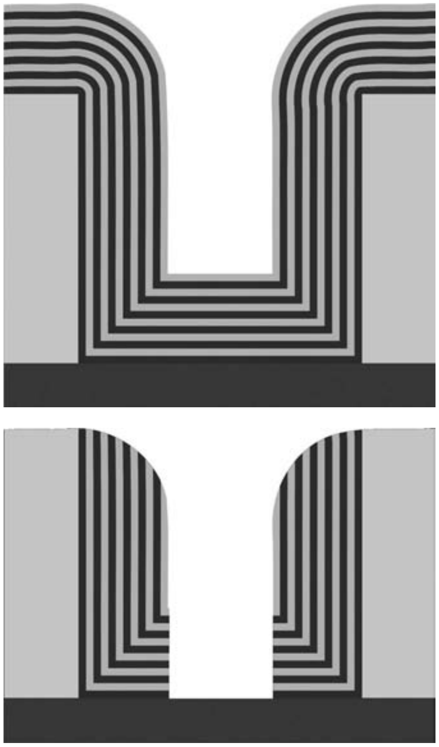

FIGURE 17.10 Structure resulting after the deposition of a slab of six poly-silicon|SiO_2 bilayers (top) in one shot followed by the conformal attack stopped with the exposure of the original lithographic seed, resulting in an array of 12 sub-lithographic wires (bottom). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

17.4.2 Multiplicative Route—$\mathbf{S}^n\mathbf{PT}_\times$

The SPT allows, starting from one seed, the preparation of *two* spacers [27]; in principle, this fact allows another, multiplicative, growth technique—$S^nPT_\times$. The multiplicative generation requires that both sides of each newly grown spacer are used as seeds for the subsequent growth—that is possible only if the original seed is etched away at the end of any cycle. In $S^nPT_\times$ each multiplicative $SPT_\times$ cycle involves therefore the following steps:

$S^nPT_\times^1$, conformal deposition of a film on the seed
$S^nPT_\times^2$, directional etching of the newly deposited film up to the exposure of the seed
$S^nPT_\times^3$, selective etching of the original seed

Figure 17.11 sketches two $S^nPT_\times$ repetitions.

Assume that the process starts from a seed formed by an array with pitch P of lithographically defined seeds (lines) each of width W, the linear density K_0 of lines being thus given by $K_0 = P^{-1}$. If lower and upper arrays have the same linear density K_0, the lithographic cross-point density is K_0^2 and the repetition of n (bottom) plus n (top) $S^nPT_\times$ results in a sub-lithographic cross-point density $\delta_\times$ given by

$$\delta_\times = 2^{2n} K_0. \tag{17.4}$$

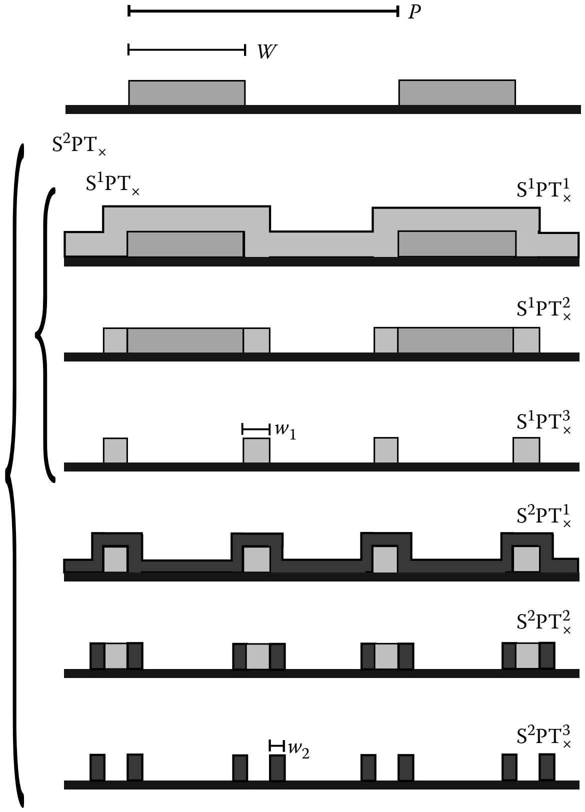

FIGURE 17.11 Two $STP_\times$ steps for the formation of a sub-lithographic wire array starting from a lithographic seed array. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

For any assigned n, this density is optimized maximizing K_0, i.e., minimizing P. The minimum value of P is determined by the considered lithography, while W is adjusted to the wanted value controlling exposure, etching, etc. The appropriate ratio W/P is obtained with the following considerations.

Let the seeds be formed by a given material A (to be concrete we shall think of it as SiO_2) and the $SPT_\times$ be carried out depositing another material B (again for concreteness, we shall think of B as poly-silicon) forming a conformal layer of thickness $t_1 = \varrho_1 W$, with $\varrho_1 < 1$. After completion of the multiplicative SPT cycle, the surface will thus be covered by an array of $2K_0$ wires per unit length each of width $w_1 = t_1 = \varrho_1 W$.

Let the process proceed with the deposition of a film of A with thickness $t_2 = \varrho_2 w_1 = \varrho_1 \varrho_2 W$, with $\varrho_2 < 1$ (in the considered example, this process could be the oxidation of poly-silicon to an SiO_2 thickness t_2). After completion of the second $SPT_\times$ cycle, the surface will be covered by a spacer array of linear density $2^2 K_0$ each of width $w_2 = t_2 = \varrho_1 \varrho_2 W$. It is noted that in $S^nPT_\times$ the seed material at the end of each $SPT_\times$ is inverted from A to B or vice versa, so that the material of the original seed must be chosen in the relation to the parity (even or odd) of n.

In the following, the focus is on the search of the mask geometry that maximizes the spacer density. After n reiterations of the $SPT_\times$, the spacers will extend both beyond and beneath the original lithographic feature. The zone containing the spacers extends from the edge of the original lithographic feature both into the region separating them and into the region beneath the original feature by amounts l_{out}^n and l_{in}^n given by

$$\begin{aligned} l_{\text{out}}^n &= w_1 + w_2 + \cdots + w_n \\ &= W \sum_{k=1}^{n} \prod_{j=1}^{k} \varrho_j, \end{aligned} \tag{17.5}$$

$$\begin{aligned} l_{\text{in}}^n &= w_2 + \cdots + w_n \\ &= W \sum_{k=2}^{n} \prod_{j=1}^{k} \varrho_j. \end{aligned} \tag{17.6}$$

The estimate of l_{out}^n and l_{in}^n requires knowledge of various ϱ_k values; at this stage, it is impossible to state anything about them. Without pretending to describe the actual technology, but simply to have quantitative (although presumably correct at the order of magnitude) estimates, we assume ϱ_k independent of k, $\forall k(\varrho_k = \varrho)$. With this assumption Equations 17.5 and 17.6 become

$$\begin{aligned} l_{\text{out}}^n &= W \sum_{k=1}^{n} \varrho^k \\ &= W \frac{\varrho}{1-\varrho}(1-\varrho^n), \end{aligned} \tag{17.7}$$

$$l_{\mathrm{in}}^{n} = W\sum_{k=2}^{n}\varrho^{k}$$

$$= W\frac{\varrho^{2}}{1-\varrho}(1-\varrho^{n-1}). \tag{17.8}$$

The least upper bounds of l_{out} and l_{in} are obtained taking the limit for $n \to +\infty$ in Equations 17.7 and 17.8: $l_{\mathrm{out}} = \varrho/(1-\varrho)$ and $l_{\mathrm{in}} = \varrho^2/(1-\varrho)$; moreover, already for relatively low values of n both ϱ^n and ϱ^{n-1} are negligible with respect to 1 so that we can reasonably assume

$$l_{\mathrm{out}}^{n} \simeq W\varrho/(1-\varrho), \tag{17.9}$$

$$l_{\mathrm{in}}^{n} \simeq W\varrho^{2}/(1-\varrho). \tag{17.10}$$

For any W, the optimum ϱ is obtained imposing the condition that all the region beneath the original lithographic feature is filled with nonoverlapping spacers: $2l_{\mathrm{in}} = W$. Inserting this condition into Equation 17.10 gives

$$\varrho \simeq \frac{1}{2}. \tag{17.11}$$

Similarly, the optimum size of the outer region is given by the following condition: $2l_{\mathrm{out}} = P - W$. Inserting this condition into Equation 17.9 gives

$$P \simeq 3W. \tag{17.12}$$

Choi et al. [28] have demonstrated that three $\mathrm{SPT}_{\times}$ repetitions on a lithographically defined seed result in nanowire arrays of device quality and suggest that very long wires can indeed be produced with a high yield. However, even accepting that the process have a yield so high as to allow the preparation of noninterrupted wires over a length (on the centimeter length scale) comparable with the chip size, if the lines are used as conductive wires of the crossbar its length is so high to have a series resistance larger than the resistance of the molecules forming the memory cell. This problem was considered in Ref. [2], where it was shown that the crossbar memory can conveniently be organized in modules each hosting a sub-memory of size 1–4 kbits. This implies that each module must be framed in a region sufficiently large to allow the addressing of the memory cells. In this way the density calculated with Equation 17.4 is an upper value to the exploitable density.

17.4.3 Three-Terminal Molecules

The use of molecules in molecular electronics is essentially due to the fact that they embody in themselves the electrical characteristics of existing devices. The characteristics of nonlinear resistors, diodes, and Schmitt triggers have been reported for two-terminal molecules; their use as nonvolatile memory cells is possible thanks to the stabilization of a metastable state excited by the application of a high voltage (thus behaving as a kind of virtual third terminal). Three-terminal molecules offer more application perspectives not only because they can mimic transistors but also because they could exploit genuine quantum phenomena like the Aharonov–Bohm effect [35].

The application potentials of three-terminal molecules can however be really exploited only if all terminals can be contacted singularly. This is manifestly impossible using the XB or XB_{+} routes, but is possible in the $\mathrm{XB}_{\star}$ framework. The major advantage of poly-silicon in the $\mathrm{XB}_{\star}$ route is that it does not pose the problem of metal electromigration. However, the multispacer technology can also be adapted for the preparation of nanowire arrays of poly-silicon and metals in arrangements that

- Allow the use of three-terminal molecules
- Facilitate the self-assembly of functional molecules
- Avoid the problem of metal electromigration

Assume that the top array defining the crossbar is formed by poly-silicon nanowires whereas the bottom array has a more complicate structure. Assume, as sketched in Figure 17.12a, that each conformal deposition is formed by the following multilayer:

$$\mathrm{SiO_2}\,|\,\text{polysilicon}\,|\,\mathrm{SiO_2}\,|\,\text{metal}\,|\,\mathrm{SiO_2}\,|\,\text{polysilicon},$$

where the metal might be, for instance, platinum obtained via CVD from $\mathrm{PtF_6}$ precursor.

After an SPT, one gets the structure sketched in Figure 17.12b; a subsequent time-controlled selective etch of the metal electrode will result in the formation of a recessed region, as sketched in Figure 17.12c.

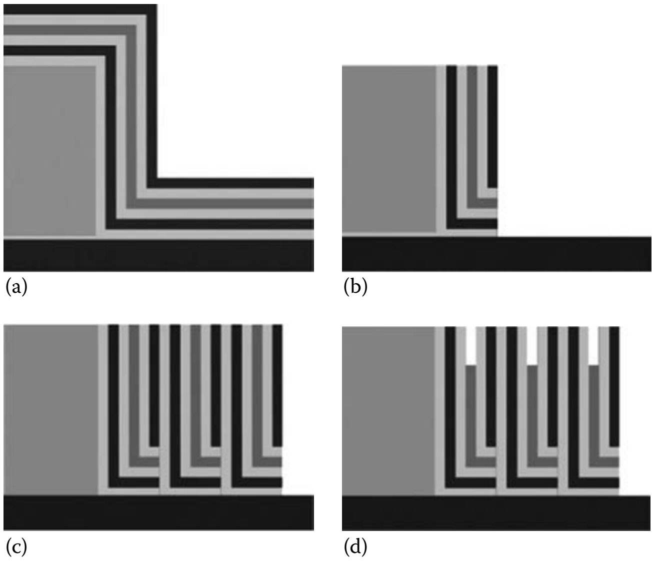

FIGURE 17.12 The structure resulting after (a) conformal deposition of a multilayer, (b) its directional etching, (c) three repetitions of the above processes, and (d) the time-limited preferential attack of the metal. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

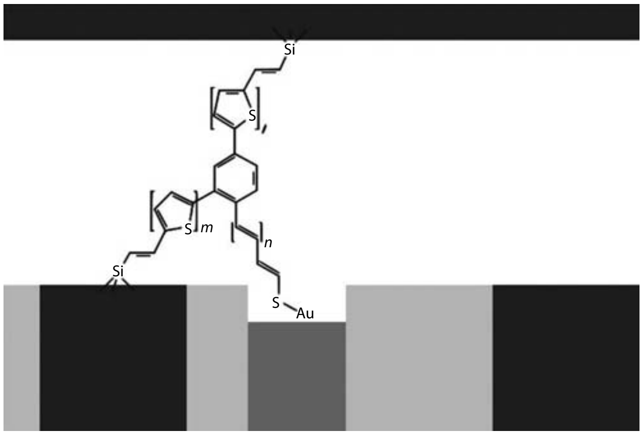

FIGURE 17.13 The three-terminal molecule after the formation of a self-assembled monolayer on the metal and the grafting to the silicon. The larger separation between metal and top poly-silicon electrode than between top and bottom poly-silicon electrodes suggests that the metal surface is subjected to an electric stress much lower than that at the bottom silicon surface.

Now, observe that thiol-terminated molecules self-assemble spontaneously on many metals (like platinum, gold, etc.) forming closely packed monolayers. Therefore, if the considered three-terminal molecules contain two alkyne and one thiol terminations, they can arrange in the cross-points in an ordered way, allowing their covalent grafting by simple heat treatment, as sketched in Figure 17.13. This figure also explains why the electric field at the metal–sulfur interface may be significantly lower than at the silicon–carbon interface.

17.5 Influence of Technology on Architecture

Device architecture and preparation procedure are strongly interlocked. This will become especially clear considering that crossbars obtained via S^nPT_+ may be linked to the external world via methods that cannot be extended to $S^nPT_\times$.

17.5.1 Addressing

If the availability of nanofabrication techniques is fundamental in establishing a nanotechnology, not less vital is the integration of the nanostructures with higher-level structures: once the crossbar structure is formed, it is necessary to link it to the conventional silicon circuitry. This is especially difficult because the nanoworld is not directly accessible by means of standard lithographic methods—"the difficulties in communication between the nanoworld and the macroworld represent a central issue in the development of nanotechnology" [36].

The importance of addressing nanoscale elements in arrays goes beyond the area of memories and will be critical to the realization of other integrated nanosystems such as chemical or biological sensors, electrically driven nanophotonics, or even quantum computers.

In the following the attention will however be limited to the problem of addressing cross-points in a nanoscopic crossbar structure by means of externally accessible lithographic contacts.

Several strategies have been adopted to attack this problem: many of them involve materials and methods quite far from, if not orthogonal to, those of the planar technology [37–43].

The consistent strategies with the planar technology are discussed in Ref. [2]. Of them, one can be applied to all crossbars irrespective of their preparation methods. According to this strategy, each line defining the crossbar extends beyond the crossing region and in this zone it is used for addressing. This region is then covered with a protecting cap, which is etched away along a narrow (sub-lithographic) line misoriented with respect to the array by a small angle α. In this way the zones where the bars are not covered are separated by a distance that diverges for $\alpha \to 0$; thus, if α is sufficiently small, the separation between the zones no longer protected makes them accessible to conventional lithography and suitable for contacting the CMOS circuitry. In this method, each line is linked separately from the others to the external circuitry—addressing n^2 cross-points requires therefore $2n$ contacts.

The multispacer technique (in particular, the S^nPT_+) permits, however, novel strategies for the nano-to-litho link in addition to the ones suitable for crossbar prepared with other techniques.

In the first of such strategies, the original mask defining the seed is shaped with n indentations with size so scaled that (1) the first indentation is filled, with the fusion of the wires, after the first deposition; (2) the second indentation is filled, with the fusion of the wires, after the second deposition; (3)…; and (4) n, the nth indentation is filled, with the fusion of the wires, after the nth deposition.

Taking into account that the minimum distance between the centers of two adjacent fused layers is $W + 3p$ (say 150 nm) and that each contact requires the definition of a hole in a region with side $2p$ (say 70 nm), this technique allows the nano-to-litho link. Figure 17.14 shows the cross section demonstrating

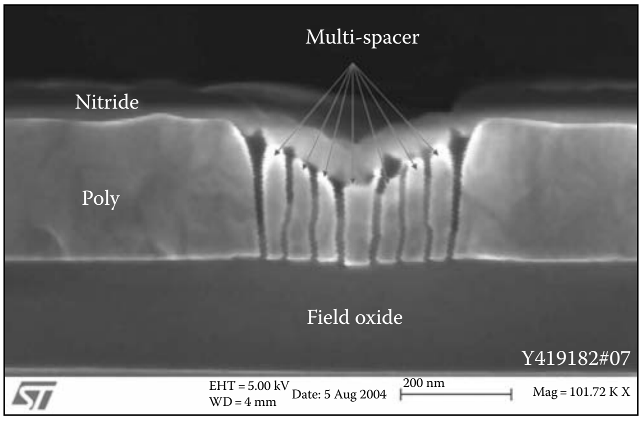

FIGURE 17.14 Cross-section showing the fusion of the arms of the fourth wire grown on two sides of the indentation. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

how filling indentation with the fusion of the central wires may render them accessible to lithography. Similarly to the strategy considered above, in this method, each line is linked separately from the others to the external circuitry, so that addressing n^2 cross-points requires therefore $2n$ contacts.

The S^{n}PT$_+$ technique results in wires with different heights. This fact can be exploited to inhibit or enable them by means of dielectrically insulated lithographically defined electrodes in the geometry described in Ref. [2]. A crossbar with $n \times n$ cross-points can therefore be addressed by controlling the conduction along the wires from one Ohmic contact to the other by means of 2 + 2 electrodes. Therefore, for the addressing of n^2 cross-points, this method needs 3 + 3 electrodes only; however, as discussed in Ref. [2], this architecture requires a complex elaboration of the information involving analog-to-digital conversion, with subsequent analysis and elaboration of data.

17.5.2 Comparing Crossbars Prepared with Additive or Multiplicative Routes

While the repetition in additive way of n SPTs per (bottom and top) layers magnifies the lithographically achievable cross-point density K_0^2 by a factor of $(2n)^2$, the repetition in multiplicative way gives a magnification of 2^{2n}. The magnification factor increases quadratically for the additive way and exponentially with the multiplicative way.

To estimate numerically the process simplifications offered by S^{n}PT$_\times$ over SPT$_+$, consider for instance the case of the 3 SPT$_\times$ repetitions per layer. This would produce a magnification of the lithographic cross-point density by a factor of $2^3 \times 2^3$. Taking $W = 0.1\,\mu\text{m}$, after 3 SPT$_\times$ repetitions, the spacer width should be of 12.5 nm, with minimum separation of 25 nm. Taking into account Equation 17.12, the cross-point density achievable with the repetition of 2 × 3 SPT$_\times$ would thus be almost the same as that obtainable with the repetition of 2 × 10 SPT$_+$ (7×10^{10} cm^{-2} vs. 8×10^{10} cm^{-2}).

Figure 17.15 shows in plan view a comparison between the following crossbars:

(a) A 2 × 2 crossbar obtained by crossing lithographically defined lines.
(b) A 16 × 16 crossbar obtained via S^{8}PT$_+$ starting from lithographically defined seeds separated by a distance allowing the optimal arrangement of the wire arrays.
(c) An 16 × 16 crossbar obtained via S^{3}PT$_\times$ starting from lithographically defined seeds separated by a distance satisfying Equation 17.12.

The figure has been drawn in the following hypotheses:

- The lithographic lines in (a) and (b) have width at the current limit for large-volume production, say $W = 65$ nm.
- The height loss τ is such that the maximum number of repetitions in the additive route is 8, and the sub-lithographic pitch is the same as shown in Figure 17.6.

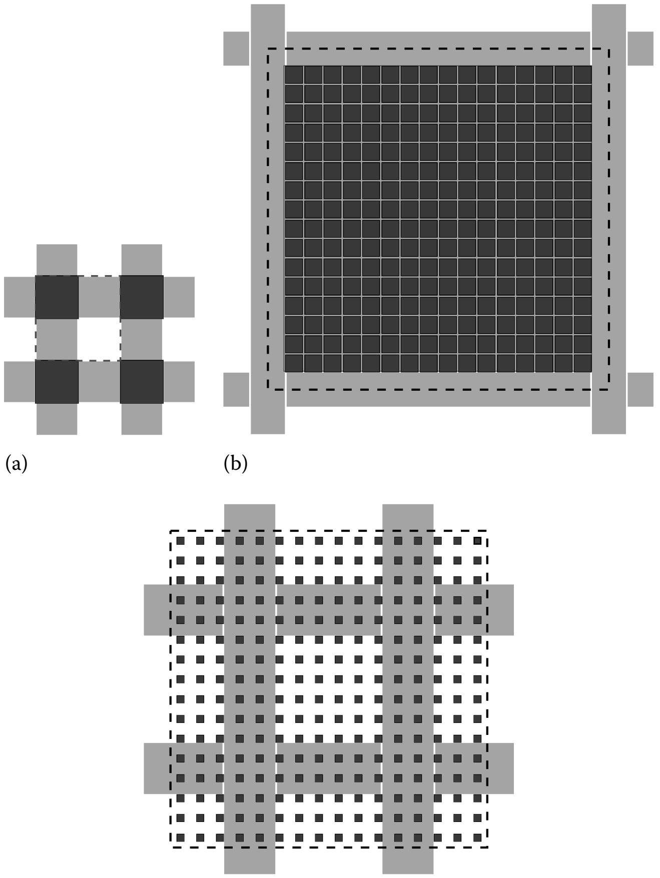

FIGURE 17.15 Plan-view comparison of the crossbars obtained (a) crossing lithographically defined lines, (b) using the lithographically defined lines above as seeds for S^{8}PT$_+$, and (c) using the lithographically defined lines above as seeds for S^{3}PT$_\times$. In each structure the square with dashed sides denotes a unit cell suitable for the complete surface tiling. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

- The lithographic width of (c) is chosen to allow the minimum pitch to be consistent with the one obtained with the additive route ($W = 100$ nm); in this way, producing sub-lithographic wires with width (12.5 nm) has been proved to be producible [28].

For $n \geq 3$, this comparison is so favorable to S^{n}PT$_\times$ to suggest its practical application. The following factors, however, would make S^{n}PT$_+$ preferable to S^{n}PT$_\times$:

1. If addressing the wires defining the crossbar are used also as addressing lines, they cannot run along the entire plane; their interruption for addressing reduces the available area.
2. In S^{n}PT$_\times$ all wires are produced collectively and have the same height and material characteristics. On the contrary, in S^{n}PT$_+$ the wires are produced sequentially, each SPT$_+$ repetition produces wires of decreasing height, and the characteristics of the material (or even the materials themselves) may vary in a controlled way from one cycle to another.

Whereas at the first glance all the S^nPT_+ features described the second item may seem detrimental, the analysis of Section 17.5.1 has clarified that they may be usefully exploited for cross-point addressing.

Deciding which route, between the additive and multiplicative ones, is actually more convenient for the preparation of hybrid devices depends on the particular application and circuit architecture. In fact, the multiplicative route (certainly less demanding for what concerns the preparation of the crossbar but more expensive for what concerns the nano-to-litho link) is presumably suitable for random access memories; on the contrary, the additive route (more demanding for the crossbar preparation, but much heavier for the elaboration of the signal) seems consistent with nonvolatile memories.

17.5.3 Applications—Not Only Nanoelectronics

The proposal of the S^nPT is quite recent [30,31], so that it has had only short time to be tested.

17.5.3.1 Electronics

Concerning the preparation of crossbars, the attention has mainly been concentrated on one side on the verification of the possibility of scaling the S^nPT_+ to large values of n (the productions of arrays with $n = 3$ [30,31], 4 [44], and 6 [45] have been reported) and on the other on the architectural impact of this technology [46–48]. In the middle, a demonstrator has been prepared showing the feasibility, without stressing the technology, of crossbars with cross-point density of 10^{10} cm^{-2} [45].

The electrical characterization of silicon nanowires has been the subject of only a few studies: Ref. [24], addressed to the nanonowires produced via the SNAP technique and Ref. [45], devoted to nanowires built with the S^nPT_+. The latter paper demonstrates that the S^nPT_+ is already suitable for the preparation of ultrahigh density memory.

The interest of a technology for the cheap production of nanowires is not limited to nanoelectronics but might extend to energetics.

17.5.3.2 Energetics

The large-scale availability of devices able to transform low-enthalpy heat (which would otherwise be dispersed into the environment) into electrical energy without complicate mechanical systems might impact the energy problem on a global scale. In principle the Seebeck effect provides a way for that. The ability of a material to operate as a Seebeck generator is contained in a parameter, ZT—a function of the Seebeck coefficient and of the electrical and thermal conductivities. The possibility of practical application is related to the occurrence of $ZT \gtrsim 1$.

From the technical point of view, Seebeck devices are already known, but materials with high Seebeck coefficient (multicomponent nanostructured thermoelectrics, such as Bi_2Te_3/Sb_2Te_3 thin-film superlattices, or embedded PbSeTe quantum dot superlattices) are expensive. This fact has allowed the application of direct thermoelectric generators only to situations (e.g., space) where cost is not important or due to other factors (e.g., weight in space application).

This state of affairs would significantly change only in the presence of highly efficient, low-cost materials. A new avenue to progress in such a direction has been the claim, by two independent collaborations [49,50], that silicon nanowires with width of 20–30 nm and rough surfaces having $ZT \simeq 1$. Although silicon is certainly a cheap material and a lot of technologies are known for its controlled deposition, this result is of potential practical interest only if the production of nanowires does not involve electron beam or extreme ultraviolet lithographies—hence the relevance of S^nPT.

It is also noted that the researchers of Refs. [49,50] employed wires with comparable height and width, thus characterized with a very high resistance. The S^nPT allows the preparation of wires with much higher aspect ratio (same width and much higher height—actually nanosheets). If it were possible to impart a sufficient roughness to the sidewalls of these sheets, it would result in much more efficient Seebeck generators.

17.6 Fractal Nanotechnology

There is an application where the multiplicative route is manifestly superior: the preparation of fractal structures on the sub-lithographic length scale.

17.6.1 Fractals in Nature

The volume V of any body with regular shape (spheric, cubic, …) varies with its area A as

$$V = g_3 A^{3/2}, \tag{17.13}$$

with g_3 being a coefficient related to the shape (for instance, $g_3 = (1/6)^{3/2}$ for cubes, $g_3 = 1/6\sqrt{\pi}$ for spheres, etc.; a well-known variational properties of the sphere guarantees that for all bodies $g_3 \leq 1/6\sqrt{\pi}$). For bodies with regular shape the ratio A/V diverges for small V and vanishes for large V.

Life can be preserved against the second law of thermodynamics only in the presence of a production of negative entropy in proportion to the total mass of the organism. This is achieved via the establishment of a diffusion field sustained by the metabolism inside the organism [51]. Since the consumption of energy and neg-entropy of living systems increases in proportion to V whereas the exchange of matter increases A, small unicellular organisms (like bacteria) have no metabolic problem due to their shape and can grow satisfying Equation 17.13 preserving highly symmetric shapes (the smallest living bacterium, the pleuropneumonia like organism, with diameter 0.1–0.2 μm, is spherical). On the contrary, larger organisms (even procaryotic, like the amoeba) may survive only adapting their shapes to have an energy uptake coinciding with what is required to preserve living functions [52,53]:

$$V = g_2 A, \tag{17.14}$$

with g_2 being a coefficient characteristic of two-dimensional growth. Moreover, the need to adapt itself to the variable environmental conditions is satisfied only thanks to the existence of an inner organellum, the endoplasmatic reticulum [54], which can fuse to the external membrane, thus allowing an efficient change of area at constant volume [52,53]. This mechanism, however, can sustain the metabolism of unicellular organisms (like the amoeba) only for diameter of at most a few tens of micrometers.

Larger organisms require the organization of the constituting cells in tissues specialized to single functions. This specialization, however, requires the formation of a vascular network (to transport catabolites to, and anabolites from, each constituting cell) whose space-filling nature implies a fractal character [55]. Nature prefers to manifest itself with fractal shapes in other situations too, like for the mammalian lung (to allow a better O_2–CO_2 exchange) [56], the dendritic links of the neuron (to allow a high interconnection degree [57]).

That smoothness is not a mandatory feature of the way how nature expresses itself, but rather fractality is of ubiquitous occurrence in a large class of phenomena even in the mineral kingdom has become clear with Mandelbrot's question about the length of the Great Britain coast [58].

The fractality of nature is manifestly approximate, the lowest length scale being ultimately limited by the atomic nature of matter. That surfaces may continue to have a scale invariance down to the atomic size became however clear only after Avnir, Farin, and Pfeifer's study of the adsorption behavior of porous adsorbents [59–61].

Surprisingly enough, the artificial production of self-similar structures has remained largely unexplored. This is somewhat disappointing because the use of arrays or matrices with self-similar structures is potentially interesting even for applications. For instance, a matrix formed by the Cartesian product of two Cantor sets would provide a way for sensing with infinite probes a surface, leaving it almost completely uncovered. Although the minimum length scale is actually larger than the atomic one, that ideal case suggests the usefulness of the idea.

The lowest length scale of biological fractals is determined by cell size, say 10^4 nm. The preparation of self-similar structures on length scales between 10^2 and 10^4 nm is relatively easy via photolithography, but this scale seems inadequate for interesting applications like the highly parallel probing of single cells (as required, for instance, by the needs of systems biology [62]). Rather, for that application the appropriate length scale seems the one characteristic of nanotechnology, 1–10^2 nm.

17.6.2 Producing Nanoscale Fractals via $S^nPT_\times$

Imagine for a moment that, in spite of the atomistic nature of matter and of the inherent technological difficulties, the multiplicative route can be repeated indefinitely. Remembering that the (n + 1)th step generates a set $\mathcal{S}_{n+1^-}$, which is, nothing but the one at the nth, $\mathcal{S}_n$, at a lower scale, the sequence $\{\mathcal{S}_0,\ldots,\mathcal{S}_n,\ldots\}$ defines a fractal. This fractal, referred to as multispacer fractal set, is self-similar only if the height of each spacer varies with n as 2^{-n}, otherwise, the fractal is self-affine [63]. As mentioned above, the "spontaneous" decrease of height with n, Equation 17.2 renders the fractal self-affine. A self-similar fractal can be obtained at the end of process planarizing the whole structure with a resist and sputter-etching in a nonselective way the composite film until the thickness is reduced to $s_0/2^n$.

It is however noted that even assuming our ability to scale down the fabrication technology, the atomistic structure of matter limits anyway the above considerations to an interval of 1–2 orders of magnitude, ranging from few atomic layers to the lower limits of standard lithography.

Having clarified in which limits the set $\mathcal{S}_n$ may be considered a fractal, it is interesting to compare it with other fractal sets. The prototype of such sets, and certainly the most interesting from the speculative point of view, is the Cantor middle-excluded set. Figure 17.16 compares sequences of three process steps eventually leading to the multispacer fractal set $\mathcal{S}$ and to the Cantor set $\mathcal{C}$. The comparison shows interesting analogies: consider a multiplicative multispacer with $P = 2W$; if $w_n = (1/3)w_{n-1}$, the measure at each step of multifractal set coincides with that of the Cantor set. This implies that the multispacer fractal set has null measure. Similarly, it can be argued that the multispacer set, considered as a subset of the unit interval, has the same fractal dimension as the Cantor middle-excluded set—ln(2)/ln(3) [63]. At each step, the multispacer fractal set is characterized by a more uniform distribution of single intervals than the Cantor set; this makes the former more interesting for potential applications than the latter.

Once one has one-dimensional fractal sets, two-dimensional fractal sets can be constructed taking their Cartesian products. Although the mixed product $\mathcal{C}\times\mathcal{S}$ is possible, in the following, we concentrate on a comparison of the Cantor and multispacer fractal crossbars. Figure 17.17 compares their plan views showing

(A) A 16 × 16 crossbar obtained via $S^3PT_\times$ starting from lithographically defined seeds separated by a distance satisfying Equation 17.12
(B) A 16 × 16 crossbar obtained via $S^3PT_\times$ starting from lithographically defined seeds and arranging the process to generate the Cantor middle-excluded set

Although the potential applications of the Cantor set are quite far, trying to reproduce it on the nanometer length scale seems of a certain interest. This is compatible with existing technologies; a possible process would involve

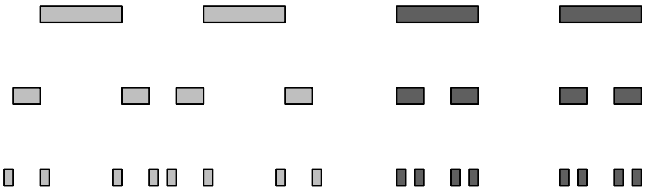

FIGURE 17.16 Generation of the multispacer set (left) and of the Cantor middle-excluded set (right). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

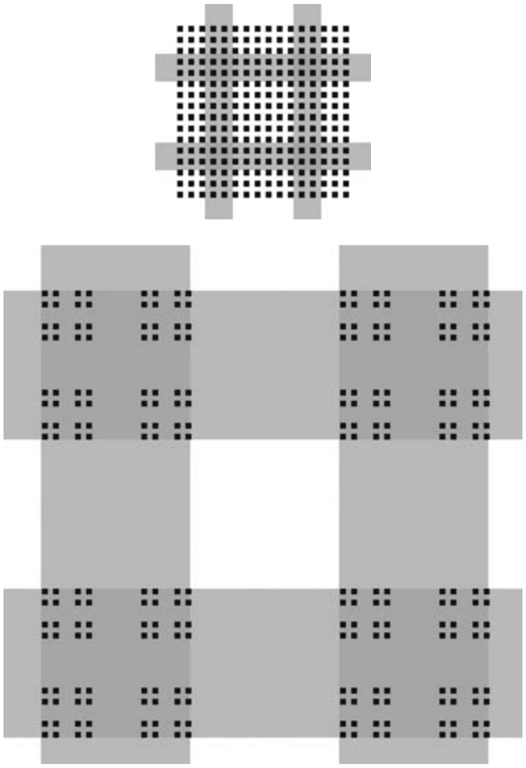

FIGURE 17.17 Plan-view comparison of the crossbars obtained via S^3PT_x, and a corresponding Cantor middle-excluded set; the width of the lithographic seed in S^3PT_x has been assumed to be at the technology forefront, whereas in the Cantor set it has been assumed sufficiently large to allow three reiterations of the process. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

(C1) the lithographic definition of the seed (formed, for instance, by poly-silicon) generating the Cantor set
(C2) its planarization (for instance, via the deposition of a low viscosity glass and its reflow upon heating)
(C3) the etching of this film to a thickness controlled by the exposure of the original seed
(C4) the selective etching of the original film
(C5) the conformal deposition of a film of the same material as the original seed (poly-silicon, in the considered example) and of thickness equal to 1/3 of its width
(C6) its directional etching
(C7) the selective etching of the space seed (glass, in the considered example)

Figure 17.18 sketches the overall process [64].

The preparation of fractal structures may appear at a first sight nothing but a mere exercise of technology stressing. The following examples suggest however the usefulness of a fractal technology:

1. The highly parallel and real-time sensing of single cells (as required, for instance, by the needs of systems biology [62] of by label-free immunodetection [65]) could be done with a minimum of perturbation (i.e., leaving almost completely uncovered the cell surface) by a matrix formed by the Cartesian product of two Cantor sets [63].

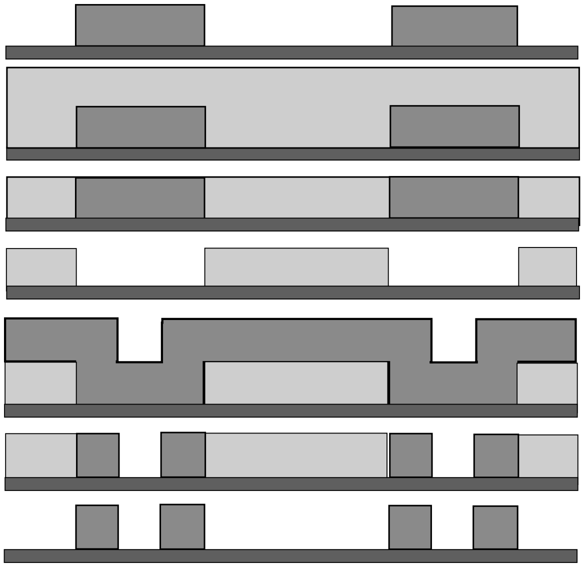

FIGURE 17.18 A process for the generation of Cantor's middle-excluded set. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

2. Superhydrophobic surfaces may be prepared controlling roughness and surface tension of nonwetting surfaces [66]. Whereas surface tension is a material property, roughness can be controlled by the preparation. For instance, roughness may be achieved by imparting a suitable (fractal) relief on the surface.
3. If the S^nPT is used for the preparation of crossbar structures for molecular electronics, the functionalization with organic molecules of the cross-points can only be done after the preparation of the hosting structure. According to the analysis of Ref. [44], this requires an accurate control of the rheological and diffusion properties in a medium embedded in a domain of complex geometry. Understanding how such properties change when the size is scaled and clarifying to which extent the domain can indeed be viewed as a fractal (so allowing the analysis on fractals [67] to be used for their description) may be a key point for the actual exploitation of already producible nanometer-sized wire arrays in molecular electronics.

Appendixes

17.A Abstract Technology

The processes considered in this chapter (conformal deposition, directional etching, and selective etching) have been introduced without specifying exactly what they are and which materials they involve. Although the reader certainly has a built-in idea of them, we however believe useful to provide their rigorous definition.

Once defined in rigorous mathematical terms, the processes of interest for this work are the building blocks of a nanotechnology that may be seen as the equivalent of Euclid's geometry, where the construction with straightedge and compass is replaced by the construction via conformal, directional or selective, deposition or etchings. Of course, the properties of a figure built with any real straightedge and compass are not exactly the same as those of the same figure constructed with the ideal straightedge and compass. Similarly, the use of real materials and processes produces structures that differ from those achievable via the ideal materials and processes. Nonetheless, it is convenient to develop the theory in abstract terms.

17.A.1 Bodies and Surfaces

The minimal characterization of a body is in terms of its geometry and constituting materials.

Definition 17.1 (Simple body) *A simple body $\mathcal{B}_X$ is a three-dimensional closed connected subset B of the Euclidean space $\mathbb{E}^3$ filled with a material X:*

$$\mathcal{B}_X = (B, X).$$

What is a material is considered here a primitive concept. It may be a substance, a mixture, a solution, or an alloy. The emphasis is on homogeneity—whichever region, however small, of B is considered, the constituting material is the same. When not necessary, the index denoting the constituting material will not be specified.

In view of the granular nature of matter, considering the limit for vanishing size is not relevant. Here and in the following when dealing with the concept of vanishing length (as it implicitly happens when considering the frontier $B^\star$ of B) we mean that the property holds true on the *ultimate length scale*. Although the ultimate length scale is an operative concept related to the probe used for observing the body, we however have in mind the molecular one. Analogously, we consider a body as indefinitely extended when its size is much bigger than the size variations induced by the considered processes. In this sense, a *wafer* is a simple body indefinitely extended over two dimensions.

Definition 17.2 (Composite body) *Consider the N-tuple of simple body $(\mathcal{B}_{X^1}, \mathcal{B}_{X^2}, \ldots, \mathcal{B}_{X^N})$ formed by nonoverlapping sets $(B_1, B_2, \ldots, B_N)$, with*

$$\forall I, J\left(B_i \cap B_j = B_i^\star \cap B_j^\star\right),$$

and let

$$B = \bigcup_{i=1}^{N} B_i.$$

If B is connected and the simple bodies $(\mathcal{B}_{X^1}, \mathcal{B}_{X^2}, \ldots, \mathcal{B}_{X^N})$ contain at least two different materials X^I and X^J, then the N-tuple is a composite body.

Definition 17.3 (Interface) *Let a composite body $\mathcal{B}$ be formed by simple bodies $\mathcal{B}_{X^1}$ and $\mathcal{B}_{X^2}$ with $X^1 \neq X^2$; the region*

$$F_{1|2} = B_1^\star \cap B_2^\star$$

defines the interface between $\mathcal{B}_{X^1}$ and $\mathcal{B}_{X^2}$.

For any $\mathbf{x}$ in $F_{1|2}$ all neighborhoods, however small, contain materials X^1 and X^2.

Definition 17.4 (Total surface) *The set*

$$S_{\text{tot}} = B^\star$$

is the total surface of $\mathcal{B}$.

Proposition 17.1 *Let a composite body $\mathcal{B}$ be formed by the N-tuple of simple body $(\mathcal{B}_{X^1}, \mathcal{B}_{X^2}, \ldots, \mathcal{B}_{X^N})$. Then*

$$S_{\text{tot}} = \bigcup_i B_i^\star \Big\backslash \bigcup_{j,k} F_{j/k}.$$

Definition 17.5 (Surface) *Let $\mathcal{B}$ be a composite body, and $B_\bullet$ be the smallest simply-connected (i.e., without holes) set containing B. Then*

$$S = B_\bullet^\star$$

is the outer surface (or simply surface) of $\mathcal{B}$.

Definition 17.6 (Inner surface) *Let $\mathcal{B}$ be a composite body. Then*

$$S_{\text{in}} = S_{\text{tot}} / S$$

is the inner surface of $\mathcal{B}$.

17.A.2 Conformal Deposition and Isotropic Etching

Definition 17.7 (Delta coverage) *For any body $\mathcal{B} \leftarrow$ with surface S, the additive delta coverage $\overline{D}_{+\delta}$ of thickness δ is the set*

$$\overline{D}_{+\delta} = \left\{\mathbf{x} : |\mathbf{x} - \mathbf{y}| \leq \delta \wedge \mathbf{y} \in S \wedge \mathbf{x} \notin B_\bullet \backslash S\right\}.$$

The delta coverage can be imagined to result from the application of an operator a_δ to B:

$$\overline{D}_{+\delta} = a_\delta(B).$$

The delta coverage $\overline{D}_{+\delta}$ is the "pod" of thickness δ covering the set B. The set obtained as union of B with its delta coverage is referred to a $B_{+\delta}$:

$$B_{+\delta} = B \cup a_{\delta}(B).$$

This operation can be reiterated, and the final set obtained after n iteration is referred to as $B^{n}_{+\delta}$:

$$B^{n}_{+\delta} = B^{n-1}_{+\delta} \cup a_{\delta}\left(B^{n-1}_{+\delta}\right),$$

where $B^{1}_{+\delta} = B_{+\delta}$ and $B^{0}_{+\delta} = B$.

The reiteration of a_{δ} is the base for the following definition:

Definition 17.8 (Conformal coverage) *A conformal coverage* $C_{+\delta}$ *of thickness* δ *is the following limit*[*]:

$$C_{+\delta} = \lim_{n \to +\infty} \bigcup_{i=0}^{n-1} a_{\delta}\left(B^{i}_{+\delta/n}\right).$$

In other words, the conformal coverage process can be thought of as obtained by applying infinitely many delta coverages; one after another, and each one of infinitesimal thickness. As shown in Figure 17.19, in general conformal coverage and delta coverage of the same thickness can lead to different bodies. Moreover from the same figure it is clear that these transformations are not topological invariants.

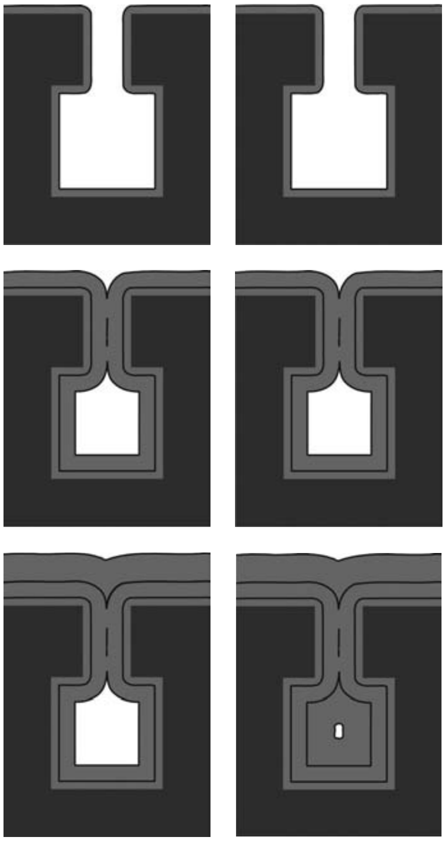

FIGURE 17.19 The different behavior of conformal coverage (left) and delta coverage (right). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

Definition 17.9 (Conformal deposition) *A conformal deposition of a film of material Z of thickness* δ *is the pair* ($C_{+\delta}$, Z).

Note that the material Z may be different from the materials X^i forming the original body.

The following theorem is trivial:

Theorem 17.1 (Planarization) *Let* $d_{\mathcal{S}}$ *the diameter of* $\mathcal{B}$. *Irrespective of the shape of* $\mathcal{B} \leftarrow$ *the conformal deposition of a layer for which* $\delta \gg d_{\mathcal{S}}$ *produces a body whose shape becomes progressively closer and closer to the spherical one as* δ increases. *If* $\mathcal{B} \leftarrow$ *is indefinitely extended in two directions, it undergoes a progressive planarization.*

Isotropic etching can be defined in a way similar to that used for conformal coverage by introducing an operator s_{δ}, characterized by the following definitions:

$$s_{\delta}(B) = a_{\delta}(\mathbb{E}^3 \backslash B_{\bullet}),$$

$$B^{n}_{-\delta} = B^{n-1}_{-\delta} \backslash s_{\delta}\left(B^{n-1}_{-\delta}\right).$$

In general, the operator s_{δ} is not the inverse of a_{δ}. In fact, $s_{\delta}(B) = a_{-\delta}(B)$ only if B is a simply connected set. We refer to s_{δ} as *delta depletion*.

Definition 17.10 (Conformal depletion) *A conformal depletion* $E_{-\delta}$ *of thickness* δ *is the following limit*:

$$E_{-\delta} = \lim_{n \to +\infty} \bigcup_{i=0}^{n-1} s_{\delta}\left(B^{i}_{-\delta/n}\right).$$

Also in this case, it is true that in general the conformal depletion and delta depletion s_{δ} of the same thickness can lead to different bodies (see Figure 17.20).

Definition 17.11 (Isotropic etching) *An isotropic* (*nonselective*) *etching is a conformal depletion of thickness* δ *applied to a body* $\mathcal{B} \leftarrow$ *regardless of its constituting materials.*

The etching process just defined is nonselective. However the most part of etching are selective (see Section 17.A.4).

---

[*] Such a limit is naively clear, but its rigorous specification should require to elaborate many technical details that we prefer to skip in this appendix. The same considerations hold true anytime we introduce limits of this kind.

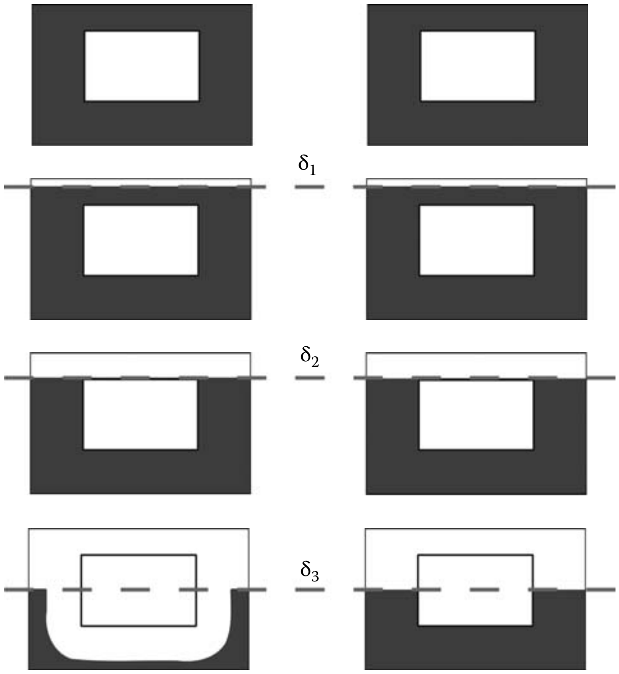

FIGURE 17.20 The different behavior of conformal depletion (left) and delta depletion (right). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

17.A.3 Directional Processes

Definition 17.12 (Shadowed surface along d) *Let* **d** *be a unit vector in* $\mathbb{E}^3$. *For any point s belonging to the surface S consider the straight line* l_s *starting from s and oriented as* **d**. *If* l_s *intersects S, then s is said to be shadowed along* **d**. *The set all shadowed points of S is referred to as the shadowed surface* $S_{\mathbf{d}}$ *along* **d** *of the body* $\mathcal{B}$.

Definition 17.13 (Directional delta coverage along d) *The directional delta coverage along* **d** *is the set of points lying on the straight lines directed along* **d** *and going from the exposed surface to the same surface shifted along* **d** *by an amount* δ:

$$\overline{D}_{+\delta,+\mathbf{d}} = \{\mathbf{x} : \mathbf{x} = \mathbf{y} + a\mathbf{d} \wedge a \le \delta \wedge \mathbf{y} \in S \setminus S_{\mathbf{d}}\}.$$

Figure 17.21 shows the results of applying to a given set the delta coverage and the directional delta coverage along **d**.

By using the directional delta coverage along δ instead of the delta coverage, it is possible to define the directional deposition.

Definition 17.14 (Directional deposition) *A directional deposition* $C_{+\delta,+\mathbf{d}}$ *of thickness* δ *along* **d** *is the following limit*:

$$C_{+\delta,+\mathbf{d}} = \lim_{n\to+\infty} \bigcup_{i=0}^{n-1} a_{\delta,+\mathbf{d}}\left(B^{i}_{+\delta/n,+\mathbf{d}}\right),$$

where $a_{\delta,+\mathbf{d}}$ *is the operator associated with the directional delta coverage and* $B^{n}_{+\delta,+\mathbf{d}} = B^{n-1}_{+\delta,+\mathbf{d}} \cup a_{\delta,+\mathbf{d}}(B^{n-1}_{+\delta,+\mathbf{d}})$.

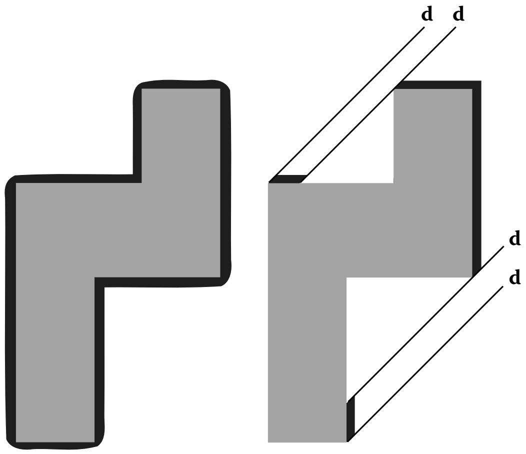

FIGURE 17.21 Comparison between delta coverage (left) and directional delta coverage along **d** (right) of the same thickness δ. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

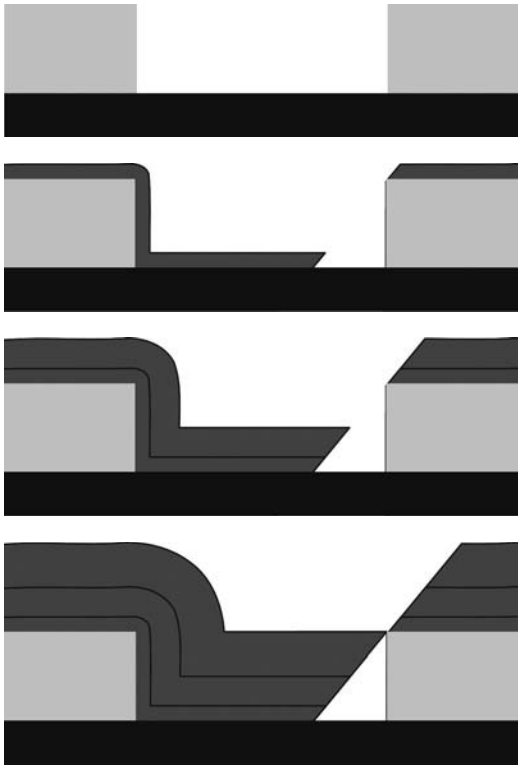

FIGURE 17.22 An example of directional deposition. (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

Figure 17.22 shows an example of directional deposition. This process (as shown in the figure) can create holes inside the bodies.

Analogously, to define the directional depletion we introduce the directional delta depletion:

Definition 17.15 (Directional delta depletion along d)

$$\overline{D}_{+\delta,-\mathbf{d}} = \left\{\mathbf{x} : \mathbf{x} = \mathbf{y} - a\mathbf{d} \wedge a \le \delta \wedge \mathbf{y} \in S \setminus S_{\mathbf{d}}\right\}.$$

Definition 17.16 (Directional etching along d) *A directional etching $E_{-\delta,-\mathbf{d}}$ of thickness δ is the following limit*:

$$E_{-\delta,-\mathbf{d}} = \lim_{n\to+\infty} \bigcup_{i=0}^{n-1} s_{\delta,-\mathbf{d}}\left(B^{i}_{-\delta/n,-\mathbf{d}}\right),$$

where $s_{\delta,-\mathbf{d}}$ is the operator associated to the directional delta depletion and $B^{n}_{-\delta,+\mathbf{d}} = B^{n-1}_{-\delta,-\mathbf{d}} \backslash s_{\delta,-\mathbf{d}}\left(B^{n-1}_{-\delta,-\mathbf{d}}\right)$.

17.A.4 Selective Processes

No indication has hitherto been given about the dependence of the same unit process on the material to which it is applied. In general the effect depends on the material, and this difference is referred to in terms of selectivity.

Before introducing selectivity, we note that the surface S of a body $\mathcal{B} \leftarrow$ is, in general, composed by different materials. In particular, by using Proposition 1 it is possible to define $S_{\mathrm{tot},X}$, the subset of S_{tot} containing material X:

$$S_{\mathrm{tot},X} = \bigcup_{X^i = X} B_i^{\star} \Big\backslash \bigcup_{j,k} F_{j|k},$$

and then the (outer) surface S_X composed by material X:

$$S_X = S_{\mathrm{tot},X} \cap S.$$

Definition 17.17 (Selective delta depletion) *For any body $\mathcal{B} \leftarrow$ with surface S, the selective delta depletion $s_{\delta,X}(B)$ of thickness δ is the set*

$$s_{\delta,X}(B) = \left\{ \mathbf{x} : |\mathbf{x} - \mathbf{y}| \le \delta \wedge \mathbf{y} \in S_X \wedge \mathbf{x} \notin B_{\bullet} \backslash S_X \right\}.$$

In other words, the selective delta depletion is a delta depletion that affects S_X only, not the whole surface S. We remark that, in general, $s_{\delta,X}(B)$ is not a connected set.

Definition 17.18 (Selectivity) *Any process is said to be selective with respect to materials X when, applied to a composite body $\mathcal{B}$ with frontier S, affects S_X only.*

Definition 17.19 (Selective etching) *A selective etching $E_{-\delta,X}$ of thickness δ is the following limit (in a sense to be defined)*:

$$E_{-\delta,X} = \lim_{n\to+\infty} \bigcup_{i=0}^{n-1} s_{\delta,X}\left(B^{i}_{-\delta/n,X}\right),$$

where

$$B_{-\delta/n,X} = \left(B \setminus s_{\delta,X}(B)\right) \cup \left(\bigcup_{X^j \neq X} B_j \right),$$

and the iterative process is defined as in the previous cases.

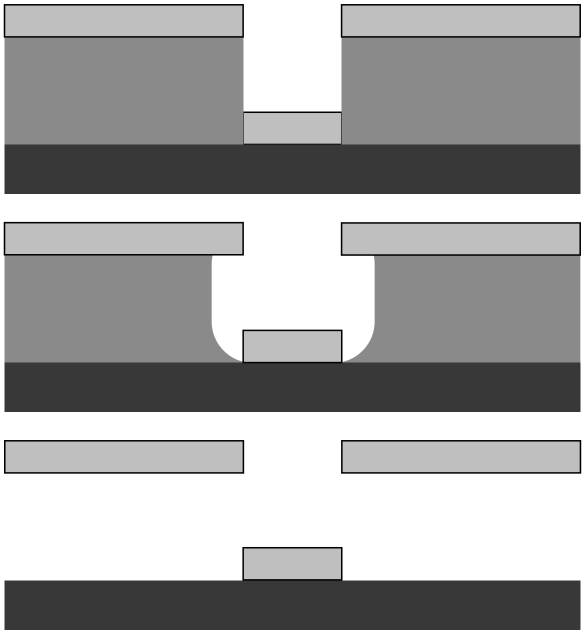

FIGURE 17.23 An example of selective etching ("liftoff"). (Reprinted from Cerofolini, G., *Nanoscale devices*, Springer, Berlin, Germany, 2009. With permission.)

As example of selective etching is given in Figure 17.23.

17.B Concrete Technology

Together with the above operations, one could certainly define other, more complicate (or perhaps even formally more interesting), operations. What is of uppermost importance here is that *there already exist combinations of materials and processes of the silicon technology mimicking the ideal behaviors described above.*

The materials considered in this chapter are well known in the silicon technology: a practical model for wafer is any body extending in two directions exceedingly more than the change of thickness resulting from the considered processes and such that only the processes are actually carried out on only one of its major surfaces. Before undergoing any operations, such wafers have generally homogeneous chemical composition and high flatness and are referred to as substrates. In the typical situations considered herein the substrates are slices of single crystalline silicon. The other materials are polycrystalline silicon (poly-Si), SiO_2, Si_3N_4, and various metals (Al, Ti, Pt, Au,…).

The typical processes involved in the silicon technology are lithography (for the definition of geometries), wet or

TABLE 17.1 Shape Resulting after Etching or Growth Processes

	Shape	
Process	Conformal	Directional
Attack	Wet etching	Sputter etching
	← Plasma etching	
		Reactive ion etching →
Growth	CVD	PVD

gas-phase etchings, chemical or physical vapor deposition, planarization, doping (typically via ion implantation), and diffusion. A special class of materials is formed by resists, i.e., photoactive materials undergoing polymerization (or depolymerization) under illumination. The processes of interest here are the following:

Lithography requires a preliminary planarization with a resist, its patterning (i.e., the definition of a geometry) via the exposure through a mask to light, the selective etching of the exposed (or unexposed) resist, the selective etching first of the region not protected by the patterned resist, and eventually of this material.

Wet etchings are usually isotropic and are used for their selectivity: HF_{aq} etches isotropically SiO_2 leaving unchanged silicon and Si_3N_4; H_3PO_4 etches isotropically Si_3N_4 leaving unchanged silicon and SiO_2; $HF_{aq} + HNO_{3\ aq}$ etches Si leaving unchanged Si_3N_4 (but has poor selectivity with respect to SiO_2).

Sputter etching is produced by momentum transfer from a beam to a target and results typically in nonselective directional etching. Selectivity can be imparted exploiting reactive ion etching.

Plasma etching can be tuned to the situation: via a suitable choice of the atmosphere it can be used for the isotropic selective etching of Si, SiO_2 or Si_3N_4; it becomes progressively more directional and less selective applying a bias to the body ("target") with respect to the plasma.

Chemical vapor deposition is the typical way for the conformal deposition that occurs when the growth is controlled by reactions occurring at the growing surface. Poly-Si grows well on SiO_2 but requires an SiO_2 buffer layer for the growth on Si_3N_4; conversely Si_3N_4 is easily deposited on SiO_2. Silicon can be conformally covered by extremely uniform layers of SiO_2 with controlled thickness on the subnanometer range via *thermal oxidation*.

Physical vapor deposition is the typical way for the deposition of metal films. The corresponding growth mode is the positive counterpart of directional etching. The conformal deposition of a metal layer can only be done via CVD from volatile precursors (like metal carbonyls or metalorganic monomers).

Table 17.1 summarizes the shape (conformal or directional) resulting from the above processes.

References

1. Heath J. R., Kuekes P. J., Snider G. S., Williams R. S., A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science, 280* (1998) 1716–1721.
2. Cerofolini G. F., Realistic limits to computation. II. The technological side. *Appl. Phys. A, 86* (2007) 31–42.
3. Aviram A., Ratner M., Molecular rectifiers. *Chem. Phys. Lett., 29* (1974) 277–283.
4. Joachim C., Gimzewski J. K., Aviram A., Electronics using hybrid-molecular and mono-molecular devices. *Nature, 408* (2000) 541–548.
5. Joachim C., Ratner M. A., Molecular electronics: Some views on transport junctions and beyond. *Proc. Natl. Acad. Sci. USA, 102* (2005) 8801–8808.
6. Tour J. M., Rawlett A. M., Kozaki M., Yao Y., Jagessar R. C., Dirk S. M., Price D. W., et al., Synthesis and preliminary testing of molecular wires and devices. *Chem. Eur. J., 7* (2001) 5118–5134.
7. Mendes P. M., Flood A. H., Stoddart J. F., Nanoelectronic devices from self-organized molecular switches. *Appl. Phys. A, 80* (2005) 1197–1209.
8. Chen J., Reed M. A., Rawlett A. M., Tour J. M., Large on-off ratios and negative differential resistance in a molecular electronic device. *Science, 286* (1999) 1550–1552.
9. Reed M. A., Chen J., Rawlett A. M., Price D. W., Tour J. M., Molecular random access memory cell. *Appl. Phys. Lett., 78* (2001) 3735–3737.
10. Luo Y., Collier C. P., Jeppesen J. O., Nielsen K. A., Delonno E., Ho G., Perkins J., et al, Two-dimensional molecular electronics circuits. *Chem. Phys. Chem., 3* (2002) 519–525.
11. Cerofolini G. F., Ferla G., Toward a hybrid micro-nanoelectronics. *J. Nanoparticle Res., 4* (2002) 185–191.
12. Green J. E., Choi J. W., Boukai A., Bunimovich Y., Johnston-Halperin E., Delonno E., Luo Y., et al., A 160-kilobit molecular electronic memory patterned at 10^{11} bits per square centimetre. *Nature, 445* (2007) 414–417.
13. Service R. F., Next-generation technology hits an early midlife crisis. *Science, 302* (2003) 556–559.
14. Stewart D. R., Ohlberg D. A. A., Beck P., Chen Y., Williams R. S., Jeppesen J. O., Nielsen K. A., Stoddart J. F., Molecule-independent electrical switching in Pt/organic monolayer/Ti devices. *Nano Lett., 4* (2004) 133–136.
15. Lau C. N., Stewart D. R., Williams R. S., Bockrath D., Direct observation of nanoscale switching centers in metal/molecule/metal structures. *Nano Lett., 4* (2004) 569–572.
16. Zhitenev N. B., Jiang W., Erbe A., Bao Z., Garfunkel E., Tennant D. M., Cirelli R. A., Control of topography, stress and diffusion at molecule metal interfaces. *Nanotechnology, 17* (2006) 1272–1277.
17. Stewart M. P., Maya F., Kosynkin D. V., Dirk S. M., Stapleton J. J., McGuiness C. L., Allara D. L., Tour J. M., Direct covalent grafting of conjugated molecules onto Si, GaAs and Pd surfaces from aryldiazonium salts. *J. Am. Chem. Soc., 126* (2004) 370–378.

18. Akkerman H. B., Blom P. W. M., de Leeuw D. M., de Boer B., Towards molecular electronics with large-area molecular junctions. *Nature, 441* (2006) 69–71.
19. Lankhorst M. H. R., Ketelaars B. W. S. M. M., Wolters R. A. M., Low-cost and nanoscale non-volatile memory concept for future silicon chips. *Nature Mater. 4* (2005) 347–352.
20. Strukov D. B., Snider G. S., Stewart D. R., Williams R. S., The missing memristor found. *Nature, 453* (2008) 80–83.
21. Cerofolini G. F., Romano E., Molecular electronics in silico. *Appl. Phys. A, 91* (2008) 181–210.
22. Natelson D., Willett R. L., West K. W., Pfeiffer L. N., Fabrication of extremely narrow metal wires. *Appl. Phys. Lett., 77* (2000) 1991–1993.
23. Melosh N. A., Boukai A., Diana F., Gerardot B., Badolato A., Heath J. R., Ultrahigh-density nanowire lattices and circuits. *Science, 300* (2003) 112–115.
24. Wang D., Sheriff B. A., McAlpine M., Heath J. R., Development of ultra-high density silicon nanowire arrays for electronics applications. *Nano Res., 1* (2008) 9–21.
25. Gates D. B., Xu Q. B., Stewart M., Ryan D., Willson C. G., Whitesides G. M., New approaches to nanofabrication. *Chem. Rev., 105* (2005) 1171–1196.
26. Whitesides G. M., Love J. C., The art of building small. *Sci. Am. Rep., 17* (2007) 12–21.
27. Choi Y.-K., Zhu J., Grunes J., Bokor J., Somorjai G. A., Fabrication of sub-10-nm silicon nanowire arrays by size reduction lithography. *J. Phys. Chem. B, 107* (2003) 3340–3343.
28. Choi Y.-K., Lee J. S., Zhu J., Somorjai G. A., Lee L. P., Bokor J., Sub-lithographic nanofabrication technology for nanocatalysts and DNA chips. *J. Vac. Sci. Technol. B, 21* (2003) 2951–2955.
29. Augendre E., Rooyackers R., de Potter de ten Broeck M., Kunnen E., Beckx S., Mannaert G., Vrancken C., et al., Thin L-shaped spacers for CMOS devices. *Eur. Solid-State Device Res., 2003. ESSDERC'03,* 2003, pp. 219–222.
30. Cerofolini G. F., Arena G., Camalleri M., Galati C., Reina S., Renna L., Mascolo D., Nosik V., Strategies for nanoelectronics. *Microelectron. Eng., 81* (2005) 405–419.
31. Cerofolini G. F., Arena G., Camalleri M., Galati C., Reina S., Renna L., Mascolo D., A hybrid approach to nanoelectronics. *Nanotechnology, 16* (2005) 1040–1047.
32. Cerofolini G. F., An extension of microelectronic technology to nanoelectronics. *Nanotechnol. E-Newslett., 7* (2005) 5–6.
33. Flanders D. C., Efremow N. N., Generation of <50 nm period gratings using edge defined techniques. *J. Vac. Sci. Technol. B, 1* (1983) 1105.
34. Cerofolini G. F., Mascolo D., A hybrid micro-nano-molecular route for nonvolatile memories. *Semicond. Sci. Technol., 21* (2006) 1315–1325.
35. Cardamone D. M., Stafford C. A., Mazumdar S., Controlling quantum transport through a single molecule. *Nano Lett., 6* (2006) 2422–2426.
36. Roukes M., Plenty of room indeed. *Sci. Am. Rep., 17*, 3 (2007) 4–11.
37. Huang Y., Duan X., Cui Y., Lauhon L. J., Kim K.-H., Lieber C. M., Logic gates and computation from assembled nanowire building blocks. *Science, 294* (2001) 1313–1317.
38. Zhong Z., Wang D., Cui Y., Bockrath M. W., Lieber C. M., Nanowire crossbar arrays as address decoders for integrated nanosystems. *Science, 302* (2003) 1377–1379.
39. DeHon A., Lincoln P., Savage J. E., Stochastic assembly of sublithographic nanoscale interfaces. *IEEE Trans. Nanotechnol., 2* (2003) 165–174.
40. Likharev K. K., Strukov D. B., CMOL: Devices, circuits, and architectures. In: *Introducing Molecular Electronics*, Cuniberti G., Fagas G., Richter K. Berlin, Germany: Springer, 2005, pp. 447–477.
41. Strukov D. B., Likharev K. K., Prospects for terabit-scale nanoelectronic memories. *Nanotechnology, 16* (2005) 137–148.
42. Beckman R., Johnston-Halperin E., Luo Y., Green J. E., Heath J. R., Bridging dimensions: Demultiplexing ultrahigh-density nanowire circuits. *Science, 310* (2005) 465–468.
43. Aswal D. K., Lenfant S., Guerin D., Yakhmi J. V., Vuillaume D., Self assembled monolayers on silicon for molecular electronics. *Anal. Chim. Acta, 568* (2006) 84–108.
44. Cerofolini G. F., Casuscelli V., Cimmino A., Di Matteo A., Di Palma V., Mascolo D., Romanelli E., Volpe M. V., Romano E., Steps farther toward micro-nano-mole integration via the multispacer patterning technique. *Semicond. Sci. Technol., 22* (2007) 1053–1060.
45. Ben Jamaa M. H., Cerofolini G., Leblebici Y., De Micheli G., Cost-efficient and CMOS-compatible fabrication of high-density poly-silicon nanowire crossbars with the multi-spacer patterning technique. In: *2009 VLSI Technology Symposium*, Kyoto, Japan, 2009.
46. Ben Jamaa M. H., Moselund K. E., Atienza D., Bouvet D., Ionescu A. M, Leblebici Y., De Micheli G., Fault-tolerant multi-level logic decoder for nanoscale crossbar memory arrays. In: *Proceedings of the 2007 IEEE/ACM International Conference on Computer-Aided Design*, IEEE Press, Piscataway, NJ, 2007, pp. 765–772.
47. Ben Jamaa M. H., Moselund K. E., Atienza D., Bouvet D., Ionescu A. M, Leblebici Y., De Micheli G., Variability-aware design of multilevel logic decoders for nanoscale crossbar memories. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 27* (2008) 2053–2067.
48. Ben Jamaa M. H., Atienza D., Leblebici Y., De Micheli G., A stochastic perturbative approach to design a defect-aware thresholder in the sense amplifier of crossbar memories. In: *14th Asia and South Pacific Design Automation Conference*, Yokohama, Japan, 2009.
49. Hochbaum A. I., Chen R. K., Delgado R. D., Liang W., Garnett E. C., Najarian M., Majumdar A., Yang P., Enhanced thermoelectric performance of rough silicon nanowires. *Nature, 451* (2008) 163–167.

50. Boukai A. I., Bunimovich Y., Tahir-Kheli J., Yu J.-K., Goddard III, W. A., Heath J. R., Silicon nanowires as efficient thermoelectric materials. *Nature, 451* (2008) 168–171.
51. Rashevsky N., *Mathematical Biophysics*, 3rd edn. New York: Dover, 1960.
52. Cerofolini G. F., Size, shape, growth and reproduction—Towards a physical morphology. *Solid Films, 79* (1981) 277–299.
53. Cerofolini G. F., The biomedium. Adsorbed water as a model for the aqueous medium, supporting life functions. *Adv. Colloid Interface Sci., 19* (1983) 103–136.
54. Ploegh H. L., A lipid-based model for the creation of an escape hatch from the endoplasmic reticulum. *Nature, 448* (2007) 435–438.
55. Gazit Y., Berk D. A., Leunig M., Baxter L. T., Jain R. K., Scale-invariant behavior and vascular network formation in normal and tumor tissue. *Phys. Rev. Lett., 75* (1995) 2428–2431.
56. Shlesinger M. F., West B. J., Complex fractal dimension of the bronchial tree. *Phys. Rev. Lett., 67* (1991) 2106–2108.
57. Simons M. J., Pellionisz A. J., Genomics, morphogenesis and biophysics: Triangulation of Purkinje cell development. *Cerebellum, 4* (2005) 1–9.
58. Mandelbrot B. B., *The Fractal Geometry of Nature*. New York: Freeman, 1982.
59. Pfeifer P., Avnir D., Chemistry in noninteger dimensions between two and three. I. Fractal theory of heterogeneous surfaces. *J. Chem. Phys., 79* (1983) 3558–3565; Erratum *80* (1984) 4573.
60. Avnir D., Farin D., Pfeifer P., Chemistry in noninteger dimensions between two and three. II. Fractal surfaces of adsorbents. *J. Chem. Phys., 79* (1983) 3566–3571.
61. Avnir D., Farin D., Pfeifer P., Molecular fractal surfaces. *Nature, 308* (1984) 261–263.
62. Hood L., Heath J. R., Phelps M. E., Lin B., Systems biology and new technologies enable predictive and preventative medicine. *Science, 306* (2004) 640–643.
63. Falconer K., *Fractal Geometry: Mathematical Foundations and Applications*, 2nd edn. New York: Wiley, 2003.
64. Cerofolini G. F., Narducci D., Amato P., Romano E., Fractal nanotechnology. *Nanoscale Res. Lett., 3* (2008) 381–383.
65. Stern E., Klemic J. F., Routenberg D. A., Wyrembak P. N., Turner-Evans D. B., Hamilton A. D., LaVan D. A., Fahmy T. M., Reed M. A., Label-free immunodetection with CMOS-compatible semiconducting nanowires. *Nature, 445* (2007) 519–522.
66. Tuteja A., Choi W., Ma M., Mabry J. M., Mazzella S. A., Rutledge G. C., McKinley G. H., Cohen R. E., Designing superoleophobic surfaces. *Science, 318* (2007) 1618–1622.
67. Kigami J., *Analysis on Fractals*. Cambridge, U.K.: Cambridge University Press, 2001.
68. Cerofolini, G., *Nanoscale Devices*. Berlin, Germany: Springer, 2009.

18

Patterning and Ordering with Nanoimprint Lithography

Zhijun Hu
Soochow University

Alain M. Jonas
Catholic University of Louvain

18.1 Introduction **18**-1
18.2 Nanoimprint Lithography **18**-3
18.3 Nanoshaping Functional Materials by Nanoimprint Lithography **18**-5
18.4 Controlling Ordering and Assembly Processes by Nanoimprint Lithography **18**-7
Ordering under Partial Confinement • Ordering under Full Confinement
18.5 Conclusions and Perspectives **18**-14
Acknowledgments **18**-14
References **18**-14

18.1 Introduction

Photolithography, which consists of etching locally a photosensitive polymer film spin-cast on a flat substrate (Figure 18.1, left column), is central to top-down micro- and nanofabrication methodologies. The image of a drawing is first projected on the film through a mask and possibly a set of lenses; subsequent development of the exposed regions, that is, removal of the exposed regions when a positive tone resist is used, results in the etching of the drawing in the polymer film (Rai-Choudhury 1997). This topographically patterned film then serves for further operations such as localized deposition of metals, etching of the underlying substrate, or ion implantation, which are basic processes for the fabrication of integrated circuits (IC) (Campbell 2001).

The need to increase the density of integration has resulted in a continuously decreasing size of the features to be lithographied. It is well known that a wave of wavelength λ is diffracted when passing through an opening of size a, and that the width of the central diffraction lobe is proportional to λ/a in the far-field limit (Born and Wolf 1980). Therefore, it does not come as a surprise that the resolution limit of photolithography is directly linked to the wavelength of the radiation used as well. Other factors also come into play (Campbell 2001) but need not interest us here. As a result, progress in downsizing photolithography led to a continuous decrease of λ, resulting simultaneously in rapidly rising costs (Resnick 2007). Starting from visible light, the trend thus went to deep UV, extreme UV (Wurm and Gwyn 2007), and soft x-ray lithography (Ueno 2007). In parallel, electron beam lithography (EBL) was developed to benefit from the short wavelength of accelerated electrons (Owen 1985, Suzuki 2007). However, EBL consists of scanning a beam of electrons over a film of an electrosensitive resist, which is essentially a serial process much slower than the all-at-once projection of a drawing. Therefore, EBL is restricted to niche uses such as research, low-volume production, or the fabrication of masks for photolithography.

The photolithographic process is actually more related to the etching technique used by artists to reproduce drawings than to the original lithographic technique. In etching, a film of wax is deposited on the surface of a metal plate. The artist removes locally the wax using a sharp stylus, and the metal plate is subsequently etched with acid where the wax is missing. The resulting metal plate is then used to reprint the drawing (bottom-left of Figure 18.1). Historically, other methods have been devised by mankind to reproduce drawings. Probably the oldest method is stamping, which can be traced back to prehistoric times when men and women of all continents inked their hands to impress the walls of caves. Traditional ink seals such as used in East Asia, known as 印章 in China, are remote descendants of this method. The stamping technique was also translated to micro- and nanotechnologies, in which it is currently known as microcontact printing (μCP) (Kumar and Whitesides 1993), the most prominent of the soft lithographic methods (Xia and Whitesides 1998, Xia et al. 1999, Gates et al. 2005). In μCP, a soft elastomer stamp is inked with a solution of a molecule capable to chemisorb on a substrate (Figure 18.1, central column). By pressing the stamp on the substrate, local grafting of the molecules ensues. The stamp softness is important to ensure conformal contact with the surface; however, it also results in undesirable sagging and deformation that usually limit μCP to the micrometer scale, even though higher resolutions were reported in favorable circumstances (Biebyuck et al. 1997, Odom et al. 2002). But more importantly, stamping methods only transfer a monolayer of ink on the surface of the substrate. The resulting drawing is thus

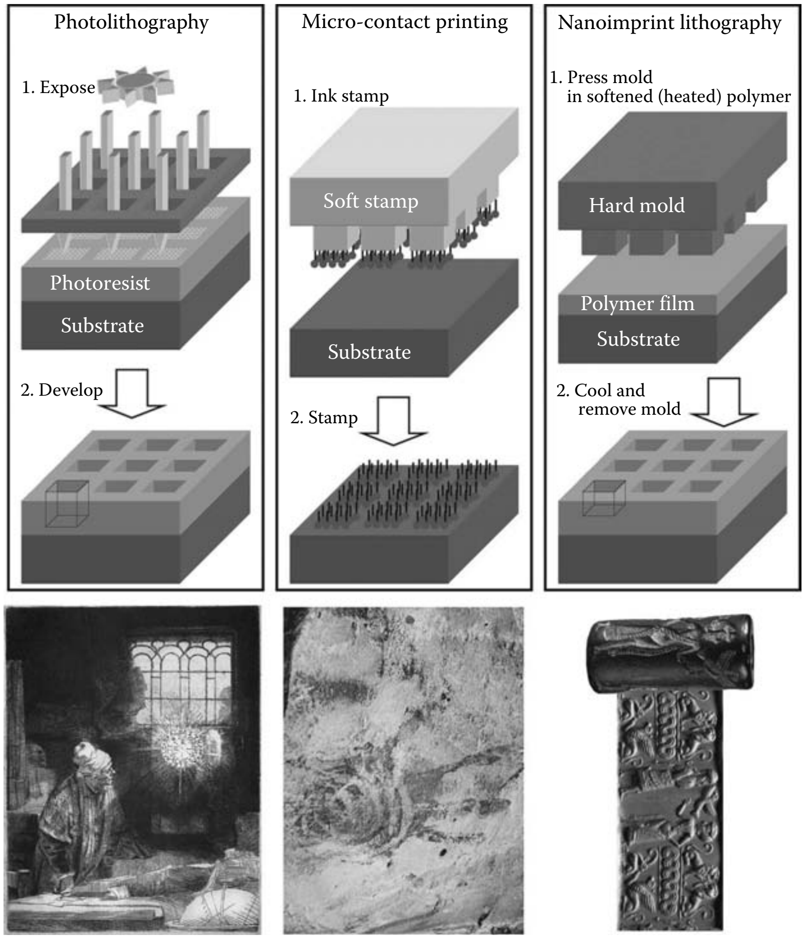

FIGURE 18.1 Schematic presentation of photolithography, microcontact printing, and nanoimprint lithography. The art drawings obtained by the corresponding historical drawing method are shown at the bottom. (Left) Doctor Faust, etching from Rembrandt, 1652; (center) a paleo-Indian handprint from Pedra Pintada, Brazil; (right) a Mesopotamian seal and its printout, *ca*. 1720–1650 BC, Syria.

essentially two-dimensional, as opposed to the three-dimensional nature of the topographically patterned polymer films used for integration technology. Therefore, μCP is more suited to create micrometer-scale chemical patterns that are used as platforms for bottom-up assembly processes (Mahalingam et al. 2004) or for the fabrication of arrays of biosensors (Kane et al. 1999), rather than for nanolithography.

Another ancient method of reproducing drawings is the embossing of soft materials such as paper, clay, or wax by hard seals. Historical records trace back to the kingdoms of Ancient Egypt and Mesopotamia, and there is evidence that seals were used as signatures in official Chinese documents before 220 BC. The method of embossing found its way in nanotechnology 5000 years after its initial discovery, where it is currently known as nanoimprint lithography (NIL) (Chou et al. 1995), hot embossing (Heyderman et al. 2000), or nanoembossing (Studer et al. 2002). NIL is a newcomer in the lithography field, since it was initially proposed in the mid-1990s only by Chou and coworkers (Chou et al. 1995, 1996); however, the method has rapidly attracted strong interest and is currently considered as one of the possible industrial options for reaching feature sizes below 30 nm (ITRS 2007). An overview of the method that consists of transferring a drawing in a polymer film by pressing a topographically patterned hard mold into the film (Figure 18.1, right column) is given in Section 18.2. Because NIL as such has already been reviewed extensively (Torres 2003, Guo 2004, Stewart and Willson 2005, Guo 2007, Resnick 2007, Schift and Kristensen 2007, Schift 2008), this section concentrates only on the most important features of the method required to understand the rest of the chapter. Note that NIL, like all other techniques mentioned above, leaves aside the issue of the fabrication of the master mold (or stamp for μCP, or mask for photolithography). Actually, the idea behind most advanced lithographic techniques is to fabricate masters by a slow, expensive but efficient technique such as EBL, and to replicate them in large numbers by a less expensive, more rapid method. This is not basically different from printing or stencil techniques, which also create one template from which many copies are produced. Obviously, the wear of the master is a potentially serious problem that is certainly more significant for methods involving a direct contact between the master and the film to be patterned, as in NIL or μCP.

The main advantages of NIL as a lithographic technique are the relatively limited cost of the method, its high throughput, and its capacity to reach very small feature sizes. But there is intrinsically more to NIL than simply the fabrication of three-dimensionally patterned polymer films for integration technology. NIL is basically a molding technique acting at the nanometer scale, which can thus be used to shape functional materials into nanowires, nanopillars, or other useful nanoobjects. This new trend is reviewed in Section 18.3. Even more interestingly, NIL can also be used as a way to control ordering processes (such as crystallization) at the nanometer scale. In this case, not only is the shape of the nanoobjects controlled during embossing but also their internal structure, which may give rise to enhanced performance compared with macroscopic materials. Recent examples of this promising emerging trend are given in Section 18.4. NIL is thus progressively moving from a purely lithographic method to a real polymer nanoprocessing tool, which might well be the niche where NIL will thrive in the future.

Before closing this section, it is useful to mention an interesting parallelism between the historical development of microscopy and lithography. Microscopy began with optical microscopes that are limited by diffraction to a spatial resolution of about λ/2 (Bergmann and Schaefer 1999), except when special setups are used in combination with fluorescent molecules (Klar et al. 2000). Higher resolutions are attained by decreasing the wavelength, leading to electron microscopes and soft x-ray microscopes (Chao et al. 2005). This clearly parallels the evolution of photolithography. However, the diffraction limit can also be beaten by changing entirely the principle of operation of the microscope. This is the case of scanning probe microscopes, in which a sharp tip is placed in contact or close to contact with a surface. In atomic force microscopy (AFM), for instance, the local force experienced by the tip is translated into an image in a way that depends on the specific mode of the operation selected (Kalinin and Gruverman 2007). The main point in the present context is that images of very high resolution can be obtained by switching from a wave-based imaging process to a contact-mode imaging process. This is not unlike the switch from a photolithography-based process to an embossing-based process. The comparison is even stronger when one realizes that AFM tips can be used as mechanical tools to orient polymer molecules (Leung and Goh 1992), just as NIL is shown in Section 18.4.

18.2 Nanoimprint Lithography

Contrarily to photolithography and EBL, NIL replicates a relief pattern in a heated thermoplastic or in a thermosetting polymer by mechanical deformation, rather than by a local photo- or electrochemical modification of the resist. The pattern is then frozen by cooling down or cross-linking the polymer. A schematic drawing of the original NIL process that was proposed for thermoplastic polymers is shown in Figure 18.2a (Chou et al. 1997). A thin film of an amorphous thermoplastic polymer is cast onto a rigid surface. The polymer film is softened by heating above its glass transition temperature. A rigid mold bearing nanoscale features is then pressed into the polymer film until the polymer flows into the recesses of the mold. The polymer film is finally hardened by decreasing the temperature below the glass transition, before the mold is removed. A thickness contrast is thus created in the polymer film after imprinting. A thin residual layer of polymer is most often intentionally left between the protrusions of the mold and the substrate in order to prevent direct impact of the hard mold on the substrate and thereby protect the fragile protrusions of the mold (Guo 2007). This residual layer is then removed by an anisotropic O_2 plasma etching process, which completes the pattern replication. However, Schulz et al. (2003) have concluded from a study of local mass transport and its effects on pattern replication that for optimal performance, the initial film thickness should be selected so as to have less than 100% filling of the mold cavities, which corresponds to nearly zero residual layer thickness.

As can be realized, a very attractive aspect of NIL is its intrinsic simplicity: the basic elements needed for NIL are limited to (1) a mold with a relief comprising a large number of nanofeatures, (2) a suitable resist such as a thermoplastic polymer that can be softened during imprinting and hardened after imprinting, and (3) a printing equipment setup allowing some control of process parameters such as temperature and

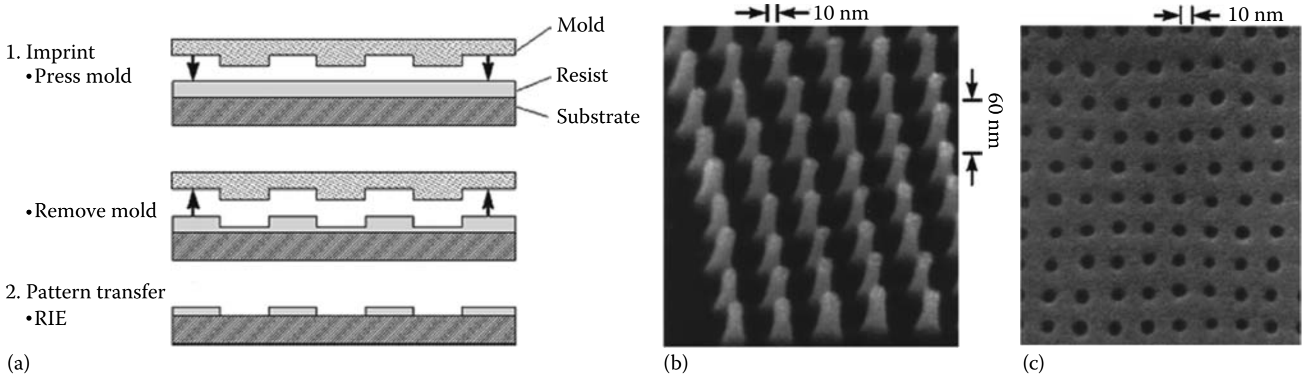

FIGURE 18.2 (a) Schematic drawing of the originally proposed NIL method. (b) SEM image of a mold with pillars of 10 nm lateral size and 60 nm height. (c) SEM image of the corresponding array of holes imprinted in a PMMA film. (Reprinted from Chou, S.Y. et al., *J. Vac. Sci. Technol. B*, 15, 2897, 1997. With permission.)

pressure. NIL is thus cost-efficient compared with other lithographic methods, while at the same time offering a good performance with respect to resolution and throughput. Unlike the development of photolithography and EBL, the lack of expensive optical components and light sources makes it much easier to construct an imprint setup. To meet the requirements of an NIL process, one needs only a well-controlled heating and cooling system and a setup for applying pressures. There are now many different companies selling NIL equipment, such as Obducat, EVGroup, or Süss MicroTec in the European Union, and Nanonex in the United States.

The mold used in NIL can essentially be any type of solid material that has a good mechanical strength and reasonable durability, and is usually made of silicon, dielectric materials such as silicon dioxide or silicon nitride, metals such as nickel, or even polymeric materials that are harder than the resist (Guo 2007). The desired nano- or microscale features defined on the molds are produced by any other lithographic processing such as photolithography, EBL, or even NIL itself, followed by localized metal deposition, reactive ion etching of metal-uncovered regions, and metal removal. In order to facilitate the separation of the mold from the substrate after imprinting, an anti-adhesion layer is normally coated on the mold surface. The requirements for this layer are that its interfacial tension with the imprinted material be large, whereas its surface tension in air be low. This favors the separation of the mold and imprinted material for thermodynamic reasons. Therefore, the most widely adopted approach is to form a self-assembled monolayer of a partially perfluorinated silane (e.g., 1*H*,1*H*,2*H*,2*H*-perfluorodecyltrichlorosilane) on the mold surface, preferably by a vapor-phase reaction (Jung et al. 2005). Mixtures of mono- and trichloro-perfluorinated silanes were also proposed (Schift et al. 2005). Perfluorinated organic materials are interesting because of their low surface tension in air and known incompatibility with hydrocarbons that favors a high interfacial tension with most polymers.

The imprinting resists used in NIL are typically thermoplastic amorphous homopolymers such as poly(methyl methacrylate) (PMMA), polycarbonate (PC), and polystyrene (PS). One common feature of these thermoplastic materials is that their viscosity and elastic modulus are significantly reduced by increasing the temperature above their glass transition temperature, T_g. Hence, the polymer viscoelastic fluid can be forced to flow into the recesses of the mold under a pressure of a few tens of bar and to conform exactly to the surface relief of the mold. As a rule of thumb, the pressure ranges usually from 10 to 100 bar and the imprint temperature is most often ~70°C–90°C above the T_g of the imprinted polymers. The polymer fluid can be easily hardened by cooling back to below T_g to freeze the imprinted nanostructures.

One key parameter for the selected polymer is its molar mass M that controls physical properties such as the glass transition temperature, the stiffness, the strength, and the viscosity (Gedde 1995). For instance, the viscosity scales as $M^{3.4}$ above some critical M_e, below which entanglements between chains do not exist in the melt (Rubinstein and Colby 2003). The pressure, imprinting times, and/or imprinting temperatures have therefore to be significantly increased for polymers of higher molar mass. Both printing temperature and pressure should be kept as low as possible in view of the time needed for temperature and pressure cycling. It is thus favorable to work with polymers of lower M, hence lower viscosity. However, because amorphous polymers of molar mass lower than M_e are brittle, they are not suitable for nanoimprinting. Typically, M_e is in the 10,000 g mol^{-1} range for styrenic and methacrylate amorphous polymers (Rubinstein and Colby 2003). Note also that the distribution of molar mass and the topology of the polymer (e.g., presence of long chain branches or cyclic chains) are important parameters with respect to rheological properties; these may thus be tuned for optimal throughput.

One unique advantage of NIL is that its resolution is not limited by factors such as light diffraction, scattering, and interference in the resist. The resolution limit of NIL can be smaller than the radius of gyration of polymer chains that is on the level of a few or a few tens of nanometers depending on molar mass. Figure 18.2b and c shows scanning electron microscopy (SEM) images of a mold with 10 nm diameter pillars and the replicated 10 nm hole array in PMMA (Chou et al. 1997), showing that indeed NIL can be used as a high-resolution nanolithographic technique.

Since the first report on NIL, the method has attracted strong interest from both industry and academia due to its merits such as high-resolution, low-cost, and inherent high-throughput resulting from parallelized processing. For lithography purposes, however, one of the drawbacks of the originally proposed NIL process is the thermal cycle that limits the throughput for semiconductor IC manufacturing. A variation of the NIL technique that uses a transparent mold and low-viscosity UV-curable monomers that cross-link upon UV exposure (step and flash imprint lithography, SFIL) was developed soon after (Ruchhoeft et al. 1999), allowing the process to be carried out at lower pressures and room temperature. Since the thermal cycle is avoided, SFIL is much faster than thermal NIL which makes it very attractive for the manufacturing of semiconductor IC devices (Resnick et al. 2003). Other variations include solvent-assisted NIL, in which the polymer is softened by a solvent vapor, and imprinting is performed at room temperature (Khang et al. 2001, Voicu et al. 2007).

In order to optimize the NIL process or widen its applications, understanding the cavity-filling process during imprinting is essential. At first sight, the initial step in the process of NIL is very similar to the technique of hot embossing, in which a polymer layer is thermally deformed using a rigid mold. However, with ever smaller feature sizes and increasingly complex mold geometries, the behavior of the thin polymer film during both molding and demolding becomes more complicated. The cavity fill process can be understood in first approximation as a squeeze flow, as shown in Figure 18.3a (Heyderman et al. 2000). Initially, the polymer flows on the cavity walls and wets them. The material in the center of the cavity is under compressive stress, causing a convex profile to form. The polymer surface thus raises and eventually touches the top of the cavity. Because squeeze flow, electrostatic interactions between the mold and polymer, capillary forces, van der Waals forces between the mold and the substrate

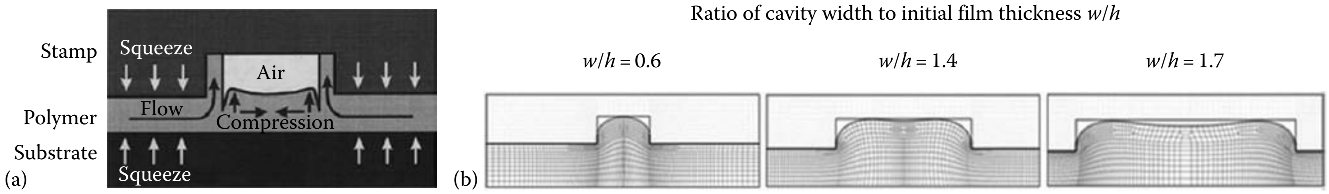

FIGURE 18.3 (a) Scheme of the forces acting on a polymer film during the filling of a cavity. (Reprinted from Heyderman, L.J. et al., *Microelectron. Eng.*, 54, 229, 2000. With permission.) (b) Cavity fill process depending on the ratio of cavity width to film thickness. (Reprinted from Rowland, H.D. et al., *J. Micromech. Microeng.*, 15, 2414, 2005. With permission.)

across the polymer fluid, and surface tensions appear in the balance of forces (Schift et al. 2001), a variety of self-organization phenomena can be triggered during imprinting, resulting sometimes in the formation of different undesirable patterns such as pillars, fractal-like patterns, and viscous fingers. Rowland et al. (2005) have investigated by numerical simulation the impact of polymer properties, mold geometry, and process conditions on the cavity-filling process. When the polymer flows vertically into an open cavity during imprinting, the polymer can deform either as a single peak centered in the cavity or as a dual peak where each peak remains close to the vertical sidewalls, depending upon geometry, as shown in Figure 18.3b. The ratio of cavity width to film thickness determines single versus dual-peak cavity filling, regardless of the absolute size of the features and also of the pressure or temperature applied during embossing.

18.3 Nanoshaping Functional Materials by Nanoimprint Lithography

It is important to note that the cavity fill process of NIL only moves the macromolecules of the imprinted resist from one place to another. The chemical properties of the imprinted resist are thus kept after imprinting, contrary to conventional lithography where a pattern is generally created by the irradiation-induced reaction of organic materials. Therefore, NIL can be adapted directly to pattern or shape of a wide range of relatively fragile materials such as functional polymers and biological macromolecules instead of photo- or electroresists. In this respect, NIL is attracting increasing attention for its ability to shape directly functional materials into nanostructures. NIL can therefore be used as an efficient nanoprocessing tool to shape a material for applications, requiring a synergy between its topography and its functional properties. NIL was therefore used

1. *To generate nanopores in materials*: The demand for low dielectric constant (low k) materials in the microelectronics industry has recently led to considerable interest in porous materials, especially nanoporous materials (Maier 2001). Imprinting can be used to create pores into a thin film of a dielectric material and therefore decreases its dielectric constant. For instance, Soles et al. demonstrated the direct patterning by NIL of sub-100 nm features into porous spin-on organosilicate glass materials (Ro et al. 2007, 2008).
2. *To shape the interface between two functional materials*: In organic heterojunction solar cells, the interface between donor and acceptor materials is crucial. This interface could be tailored by NIL. A thin film of conjugated polymer (serving as acceptor) was first imprinted in a controlled atmosphere (low vacuum or inert gas). Well-defined nanoscale structures were thus created in the acceptor layer. The other component (donor) of the system was then poured in the created trenches (Kim et al. 2007, Cheyns et al. 2008 and Aryal et al. 2008). The fill factor and power conversion efficiency were correspondingly increased.
3. *To create optical gratings in light-emitting materials*: Within recent years, organic dye-doped or conjugated polymer laser devices have received much attention as convenient and compact light sources. One of the important issues to make polymer lasers is the fabrication of a periodic surface relief in the conjugated or dye-doped polymers, with a period close to the wavelength of light. The periodic relief can modify the generation and propagation of light in these materials by interference effects. Chou et al. were the first to apply NIL to directly shape organic light-emitting materials, such as Alq3 (8-hydroxyquinoline aluminum) doped with DSMII dye molecules, or their blends with PMMA (Wang et al. 1999). Figure 18.4a shows AFM images of the resulting Alq3/DSMII gratings of 300 nm period. It was demonstrated that no degradation of the light emission efficiency is caused by the NIL process (Figure 18.4b). In addition, because the small dye molecules embedded in the PMMA matrix are slightly aligned by the imprinting process (see Section 18.4), an anisotropic photoluminescence behavior was observed (Wang et al. 2000).

Subsequently, Guo et al. (2002) successfully imprinted two-dimensional photonic crystal nanostructures in nonlinear optical dye-doped polymers. Pisignano et al. (2003, 2004) and Mele et al. (2005, 2006) showed that conjugated polymers and oligomers can be directly imprinted at room temperature under a very high pressure, based on the plastic deformation of the polymer. Figure 18.5a is an AFM image of a room temperature-imprinted organic semiconductor thin film. A study of the optical properties before and after imprinting (absorbance, luminescence, and quantum yield) again ruled out the possible degradation of the active materials even when imprinting was performed in air (Figure 18.5b). The imprinting of

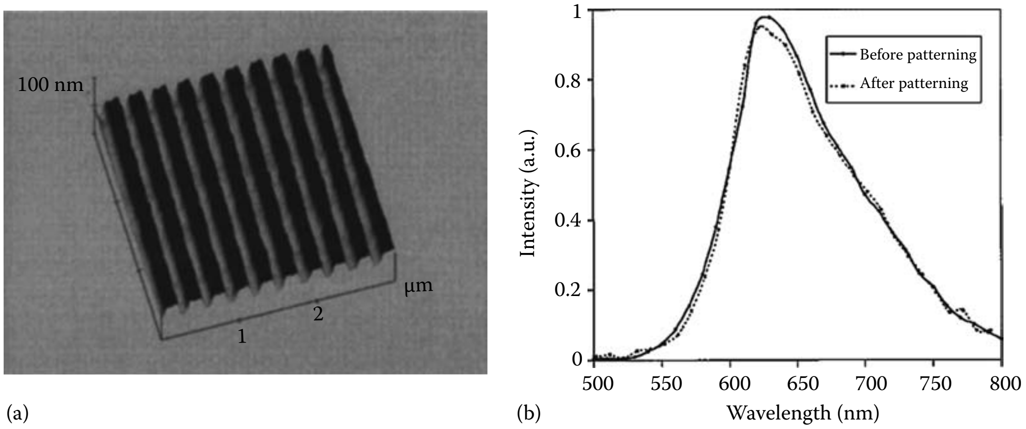

FIGURE 18.4 Optical grating imprinted by NIL in a light-emitting material. (a) AFM height image of Alq3/DMSII gratings with a period of 300 nm. (b) The measured luminescence spectral intensity before and after imprinting, demonstrating the absence of sample degradation. (Reprinted from Wang, J. et al., *Appl. Phys. Lett.*, 75, 2767, 1999. With permission.)

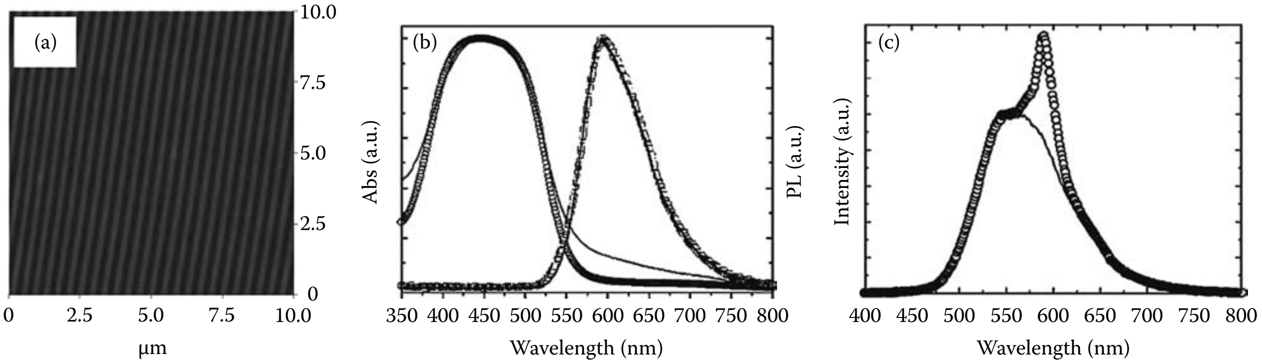

FIGURE 18.5 Optical grating imprinted by NIL at room temperature in a light-emitting organic semiconductor. (a) AFM height image of the surface grating with a period of 400 nm. Image side: 10 µm. (b) Absorbance and integrated photoluminescence spectra before (line) and after (circles) imprinting, showing the absence of degradation. (c) Photoluminescence spectra at a selected emission angle, showing the presence of an interference peak after imprinting (circles) compared with the uniform film (line). The interference peak is due to the surface grating. (Reprinted from Pisignano, D. et al., *Adv. Mater.*, 16, 525, 2004. With permission.)

a surface periodic grating with wavelength-scale periodicity enhances the light emitted in specific directions due to interference effects (Figure 18.5c), and reduces the effective length covered by the photons in the material and thus the self-absorption inside the organic slab. These interference effects were subsequently used to build a distributed feedback (DFB) laser, wherein the surface grating serves to produce a narrow band of wavelengths by interference. Figure 18.6 displays the photoluminescence spectrum of such a DFB laser obtained by NIL using the conjugated light-emitting polymer (poly[2-methoxy-5-(2-ethylhexyloxy)-1,4-phenylenevinylene]) (MEH-PPV), showing a narrow emission at about 630 nm wavelength.

4. *To create antireflective structures on the surface of optical materials*: Two-dimensional antireflection subwavelength structures were also created by NIL on the surface of thin films in order to suppress Fresnel reflection. Fresnel reflection is a drawback for many optical systems, since it reduces the fraction of transmitted light, deteriorates the contrast of displays, and generates ghost images in imaging systems (Ting et al. 2008a,b). Surface relief gratings with periods smaller than the wavelength of light, named subwavelength gratings, may behave as antireflection surfaces. Figure 18.7a is an SEM image of antireflective PMMA nanostructures fabricated by NIL. The optical properties of a plain PMMA sheet with the ones of a patterned PMMA sheet are shown in Figure 18.7b, showing a significant reduction of glare after imprinting.
5. *To shape biomaterials*: Ohtake et al. (2004) reported that DNA can be directly patterned by NIL; interestingly, the patterned DNA retains its activity, showing that these biomacromolecules can withstand the relatively harsh processing conditions of NIL. Park et al. (2007) showed that chitosan, a derivative of a natural polysaccharide, can also be directly patterned by NIL at low temperature

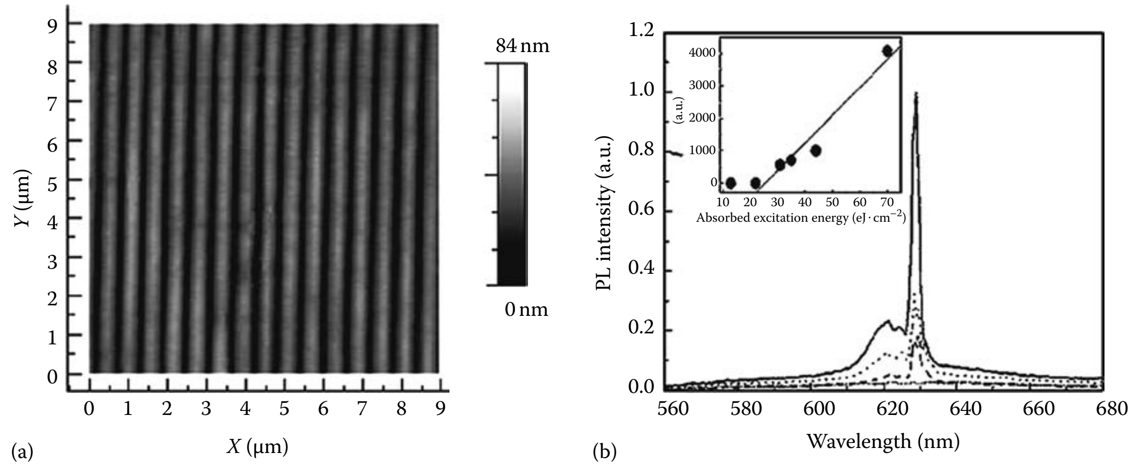

FIGURE 18.6 (a) AFM height image of MEH-PPV gratings with a period of 600 nm patterned directly by NIL. (b) Distributed feedback emission spectra under pulsed excitations of increasing dose. The narrow band of wavelengths emitted at about 630 nm results from interference from the surface grating. Inset: the influence of excitation dose on the DFB emission intensity. (Reprinted from Mele, E. et al., *Appl. Phys. Lett.*, 89, 131109, 2006. With permission.)

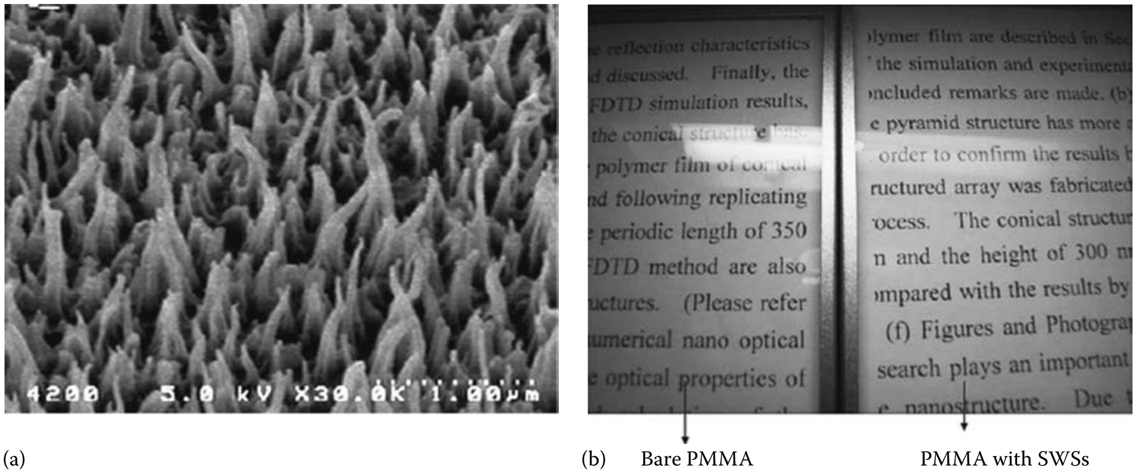

e reflection characteristics
d discussed. Finally, the
FDTD simulation results,
the conical structure has
polymer film of conical
nd following replicating
periodic length of 350
DTD method are also
uctures. (Please refer
umerical nano optical
optical properties of

lymer film are described in Sec
the simulation and experiment
ncluded remarks are made. (b
e pyramid structure has more
order to confirm the results b
ructured array was fabricated
ocess. The conical structur
n and the height of 300 n
mpared with the results by
(f) Figures and Photogra
search plays an important
e nanostructure. Due

FIGURE 18.7 (a) SEM image of PMMA films with imprinted antireflective nanostructures. (b) Comparison between a bare PMMA sheet and the imprinted PMMA sheet, showing an obvious damping of backlight glare. (Reprinted from Ting, C.-J. et al., *Nanotechnology*, 19, 205301, 2008a. With permission.)

and low pressure. The nanoimprinted chitosan patterns contain a large amount of primary amino groups for further chemical modification, and can be used, for instance, to graft proteins and DNA strands for biosensing purposes.

Interestingly, nanoshaping of functional *inorganic* materials is also possible by NIL. For instance, Grigoropoulos et al. (Ko et al. 2007, Park et al. 2008) demonstrated the direct nanoimprinting of thin films of gold nanoparticles. Because of their small size (2–3 nm), the nanoparticles start to melt at ~130°C–140°C, well below the melting point of bulk gold. A variety of gold nanostructures were fabricated, such as nanodots, nanowires, and serpentine nanowire arrays, showing the potential of the method for the direct fabrication of electrodes for nanoscopic devices. Imprinting of soft precursors of inorganic crystals and glasses was also reported, showing the possibility to use NIL to fabricate ceramic or inorganic glass nanostructures (Harnagea et al. 2003, Okinaka et al. 2006, Peroz et al. 2007). Finally, direct imprinting of silicon wafers was also achieved via laser-induced transient liquefaction of the imprinted material (Chou et al. 2002).

18.4 Controlling Ordering and Assembly Processes by Nanoimprint Lithography

As mentioned in Section 18.1, the application of NIL to organic materials such as semicrystalline polymers or liquid crystalline oligomers is capable to control the internal structure of the resulting nanoobjects. Consider semicrystalline polymers as an example. These materials crystallize in ~10 nm thick lamellar-shaped crystals, in which the chains are perpendicular or slightly inclined to the lamellar surface, and when emerging from the crystal surface either fold back in the same lamella or reenter a neighboring quasi-parallel lamella (Figure 18.8) (Gedde 1995, Reiter and Strobl 2006). The lateral sizes of polymer lamellae are in the micrometer range, and the lamellae are elongated in the preferred growth direction of the crystal. Successive lamellae are separated by amorphous interlayers, which accommodate the folds, loose loops, tie segments, and dangling ends of the chains (Mandelkern 1979, Gautam et al. 2000). The crystallization process occurs through nucleation from a central seed, most often an impurity (Wunderlich 1976), from which a complex assembly of lamellae forms by crystal growth, branching, and splaying (Bassett 1984). The resulting so-called spherulite has a radius in the micrometer range, and the preferred growth direction of the lamellae is essentially radial within the spherulites. Hence, at the supraspherulitic scale, semicrystalline polymers are isotropic, and consist of an intricate mixture of crystallized and amorphous nanoregions. Therefore, their properties are averages of different phases over all possible orientations, and include components resulting from the numerous crystal–amorphous interfaces present in a spherulite. This significantly reduces the performance of such materials.

Consider now the possible effects of NIL on such a complex morphology. Embossing is performed in the molten state, but the polymer crystallizes in the mold during cooling. First, the nucleation process could be affected, because NIL splits the melt into a very large number of small cavities. Since the probability to find a heterogeneous nucleating seed in a cavity depends on the product between the volume of each nanocavity and the number density of nucleating seeds in the melt, this probability may be well below 1. If so, crystallization in the nanocavity will either require the spontaneous formation of a nucleating germ in the melt, a process called homogeneous nucleation rarely encountered in practice when crystallizing bulk polymers, or the propagation of a crystal from the outside into the nanocavity (if the cavity is not totally isolated), which will only be possible when a nearby crystal grows in the proper orientation. The situation is similar to the one encountered when crystallizing polymers in microdroplets (Cormia et al. 1962, Massa and Dalnoki-Veress 2004), in microdomains of phase-separated block copolymers (Reiter et al. 2001, Loo et al. 2002, Muller et al. 2002), or in the nanopores of membranes (Steinhart et al. 2006, Woo et al. 2007), cases for which retarded crystallization and preferred orientation are often reported. One can thus expect identical effects to occur in NIL as well.

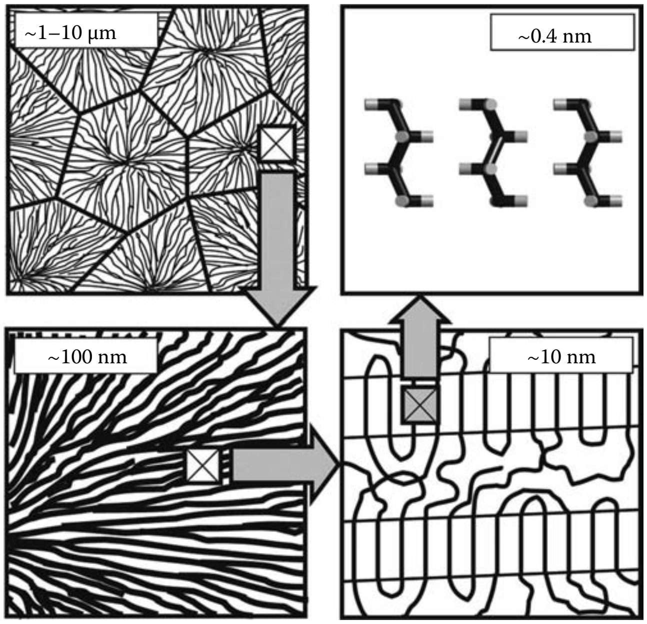

FIGURE 18.8 Scheme of the morphology of semicrystalline polymers seen at different length scales, showing spherulites (top-left image), stacks of lamellae (bottom-left image), lamellar crystals and amorphous interlayers (bottom-right image), and crystalline stems (top-right image).

Second, the crystal growth process might also be affected. It was demonstrated for the first time in Russia in 1972 (Sheftal and Bouzynin 1972), and rediscovered later in the United States (Smith and Flanders 1978), that topographically patterned amorphous surfaces may induce preferential orientation of growing crystals, provided the symmetry of the topography be adapted to the habit of the growing crystals. This phenomenon is known as "graphoepitaxy" or "artificial epitaxy" (Givargizov 1994, 2008), with reference to the epitaxy of crystals on crystalline substrates, where matching between the lattice planes of the growing crystal and of the substrate provides for preferential orientation. In graphoepitaxy, however, the matching occurs at a higher level and does not require a crystalline substrate. Key thermodynamic parameters are the shape of the growing entity and the symmetry of the substrate topography, as well as the surface tensions and surface tension anisotropies of the crystal–substrate and melt–substrate interfaces, which dictate the energy balance of the growing system (Mouthuy et al. 2007). On the kinetic side, the growth rates of the crystal along specific crystalline directions are also important parameters because they control the crystal shape, hence the graphoepitaxial matching. In the present context, the topography of the NIL mold may thus control graphoepitaxially the crystallization process of the polymer, resulting in a preferential alignment of specific crystallographic directions with respect to the nanofeatures of the mold.

A third factor that could affect the crystallization process in a NIL experiment is the degree of chain alignment that results from the melt embossing process. Polymers are viscoelastic fluids, which may keep for a long time a partial memory of the considerable squeeze flow experienced during the imprinting. In addition, polymer chains tend to align along flow lines during processing. This favors oriented crystallization (Miller 1979, Kumaraswamy et al. 1999, Somani et al. 2000) and the emergence of specific morphologies such as "shish-kebabs," in which long crystalline nanofibers made of aligned chains (the "shishes") serve as nuclei for oriented lamellar crystals (the "kebabs") (Eder et al. 1989, Kimita et al. 2007). In NIL, the residual degree of chain alignment will depend on the molar mass of the polymer, on its previous processing history, and on its annealing time in the melt before crystallization. Note that chain orientation under flow is not easy to predict at the nanometer scale, since molecular confinement of the chains in a thin film leads to a substantially perturbed rheological behavior (Rowland et al. 2008).

A fourth factor is confinement itself. Most macromolecules have radii of gyration in the 10–50 nm range, close to the thickness of films used in NIL and to mold feature sizes, resulting in strong confinement of the chains. There is currently an ongoing discussion in the literature regarding the way confinement affects polymer properties such as glass transition and segmental mobility (Ellison and Torkelson 2003, Alcoutlabi and McKenna 2005) or whole chain motion (Rowland et al. 2008). Conflicting results have been published, and it is not possible to review them fairly in the present chapter. Instead, one should keep in mind that the conformation and dynamic properties of confined polymers may be substantially different in a nanocavity than in the bulk, which is likely to modify their crystallization behavior as well. Think only of the need that chains fold back and forth to form lamellar crystals (Figure 18.8). If this occurs in a space only slightly larger than the intrinsic lamellar thickness, significant perturbations of the crystallization process are clearly to be expected.

What is true for polymer crystallization should also be valid for other supramolecular assembly processes, such as the formation of liquid crystalline domains or the ordering of block copolymer mesophases. In the sequel, we will review the results obtained in this emerging field, for a range of systems such as amorphous rigid polymers, liquid crystalline polymers, semicrystalline polymers, and phase-separated block copolymers. NIL will be shown to be unique in its ability to orient crystals and chains in specific directions, resulting in favorable cases in a significant improvement of the performance of nanodevices. The results will be divided in two classes: NIL under either full or partial confinement. In the first case, the starting film thickness and the height of the protrusions on the mold are such that the mold can touch the surface of the substrate if pressed strongly enough (Figure 18.9). For a mold having a protruding surface fraction σ, each protrusion being of height h, full confinement will only be possible when the starting film thickness $d < (1 - \sigma)h$. Under this circumstance, each recess of the mold will be effectively isolated from other recesses. In contrast, when $d > (1 - \sigma)h$, a continuous residual film always remains between the substrate and the protruding features of the mold, no matter how hard the mold is pressed in the film (Schift and Heyderman 2003). In this case, information on the ordering process may propagate from recess to recess, which will have significant consequences on the global ordering of the sample.

18.4.1 Ordering under Partial Confinement

The first experimental demonstration of molecular alignment by NIL was obtained for chromophores dissolved in a standard PMMA resist, using line gratings as molds (Wang et al. 2000). Polarized photoluminescence and absorption measurements indicated that the dye molecules are aligned after imprint with their long axis along the grating lines, most probably due to the effect of flow during processing. The fluorescence anisotropy, $R = (I_{//} - I_{\perp})/(I_{//} + 2I_{\perp})$, quantifies the degree of alignment by comparing the fluorescence emission polarized parallel to the grating lines ($I_{//}$) and perpendicular to them ($I_{\perp}$). R values of 1 correspond to complete alignment, whereas $R = 0$ for isotropic systems. From the reported dichroic ratios $DR = I_{//}/I_{\perp}$, R values of 0.1–0.2 can be computed for this specific example, which shows that the polymer flow is able to align the dye molecules only moderately.

For semicrystalline polymers crystallized in partial confinement, no or limited preferential orientation of polymer crystals occurs when line gratings are used as molds (Hu et al. 2005, Okerberg et al. 2007). Thin films of poly(vinylidene fluoride) (PVDF, α-phase) crystallize in a normal spherulitic morphology and appear to be insensitive to the presence of the grooves of the NIL mold (Hu et al. 2005). The crystalline lamellae propagate from groove to groove, and the electron diffraction patterns do not show preferred orientation relative to the line grating direction (Figure 18.10). This can be ascribed to fast crystal growth in the thin residual film below the protrusions of the mold, from which crystallization propagates into the linear grooves of the mold. The final morphology is thus not dictated by the mold, but mainly by the processes occurring in the residual film. Similar results were obtained for poly(ethylene oxide) (PEO) (Okerberg et al. 2007), although partial elongation of the crystalline lamellae along the grooves of the mold was sometimes observed. As is readily apparent in Figure 18.11a, the global spherulitic shape is not affected by

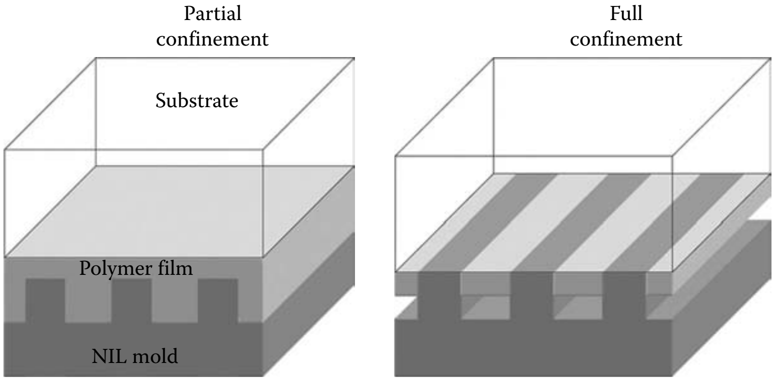

FIGURE 18.9 Two configurations of NIL leading to different microstructures after imprint. For clarity, the mold was placed at the bottom of the drawing.

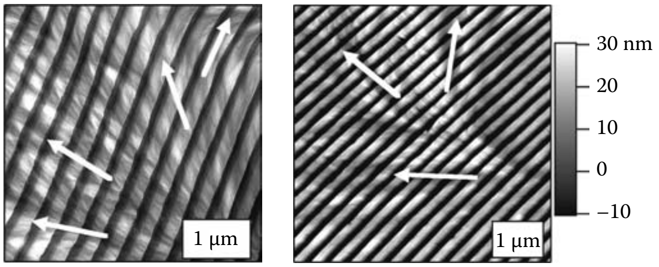

FIGURE 18.10 Microstructure of PVDF (α-phase) crystallized in NIL line grating molds under partial confinement (left: transmission electron micrograph; right: atomic force microscopy topograph). No preferential alignment is obtained, as shown by the arrows that highlight the direction of a few lamellar crystals. (Adapted from Hu, Z. et al., *Nano Lett*., 5, 1738, 2005.)

the NIL mold, and lamellae are found growing either perpendicular to the grooves (zone 1 in Figure 18.11a and b) or parallel to them (zone 2 in Figure 18.11a and c), depending on the orientation of the radius of the spherulite with respect to the line grating direction. These results clearly indicate that chain alignment by the flow is not sufficient to induce oriented crystal growth and that graphoepitaxy of semicrystalline polymers does not occur in partial confinement.

The situation is somewhat different when imprinting a low molar mass liquid crystalline conjugated polymer, poly(9,9-dioctylfluorene-*co*-benzothiadiazole) (F8BT), in its liquid crystalline nematic mesophase (Zheng et al. 2007, Schmid et al. 2008). This green-light-emitting polymer displayed after imprinting *R* values as large as 0.97, indicating that most polymer chains are aligned parallel to the line grating direction. The order parameter *R* was found to increase for thinner films, suggesting that the ordering starts from the surface of the mold and progressively vanishes when going down into the residual film. Polymer-light-emitting diodes (PLEDs) were successfully built based on nanoimprinted F8BT, giving rise to the emission of preferentially linearly polarized light (Figure 18.12). A field-effect transistor was also realized, and the hole mobility was 10–15 times larger parallel to the grating lines than perpendicular to them. These results directly illustrate the beneficial impact of aligning molecules, as far as charge mobility or emission polarization is concerned. It should,

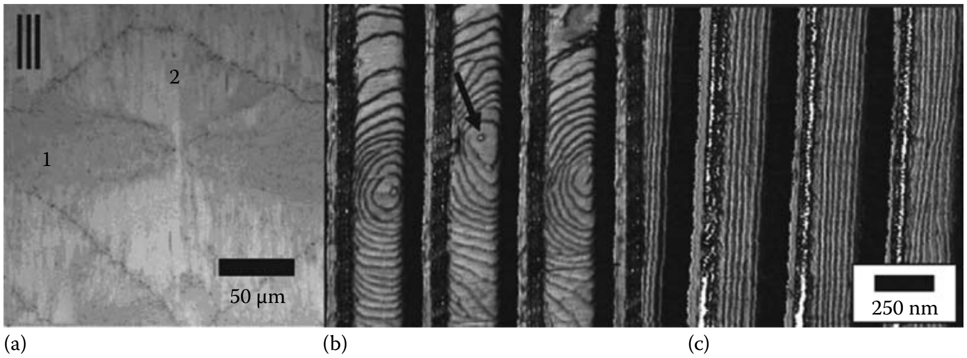

FIGURE 18.11 Microstructure of PEO crystallized in NIL grating molds under partial confinement. Panel (a) is an optical micrograph displaying a spherulite; the vertical stripes indicate the direction of the lines of the grating used to imprint the sample. Panels (b) and (c) are AFM topographic images corresponding to the regions 1 and 2 of panel (a), respectively, and show different relative orientations of the lamellae compared with the line grating direction. (Reprinted from Okerberg, B.C. et al., *Macromolecules*, 40, 2968, 2007. With permission.)

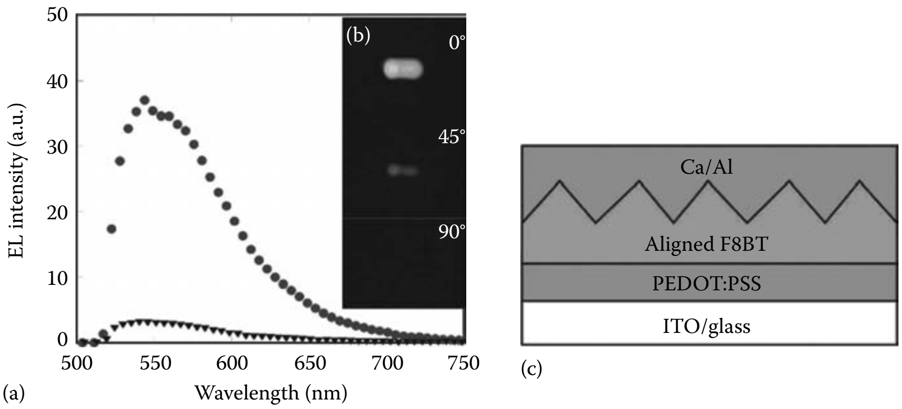

FIGURE 18.12 (a) Electroluminescence of a PLED made of F8BT imprinted in partial confinement with a line grating. The luminescence is measured through a polarizer parallel (circles) or perpendicular (triangles) to the grating lines. (b) Optical images of the electroluminescent device seen through a polarizer oriented with respect to the grating lines as indicated in the figure. (c) Schematic drawing of the PLED structure. (Reprinted from Zheng, Z. et al., *Nano Lett*., 7, 987, 2007. With permission.)

however, be noted that NIL failed to align a similar light-emitting conjugated polymer, poly(9,9-dioctylfluorene) (F8) also displaying a liquid crystalline nematic mesophase (Song et al. 2008). This was tentatively ascribed to the higher molar mass (hence larger viscosity) of this polymer compared with F8BT. Clearly, NIL under partial confinement is able to align liquid crystalline polymers in favorable circumstances; however, further work is required to fully understand the factors which control this alignment that are probably similar to the ones driving the alignment of liquid crystals on buffing layers (Toney et al. 1995).

18.4.2 Ordering under Full Confinement

When the starting film thickness is small enough for the NIL mold to enter in contact with the substrate, full confinement arises (Figure 18.9b). In this case, the propagation of information from groove to groove is suppressed, and complete graphoepitaxial alignment happens as is demonstrated below. It is critical to realize that graphoepitaxial alignment occurs at some stage during the formation of a structure, depending on the details of the processing history. It may thus orient a growing crystal, or a liquid crystalline domain, or even a single molecule, depending on the basic structural element that interacts with the mold. The notion of basic structural element is thus key to understand and reconcile the results published in the literature, and will serve us as guide in the sequel.

When a *semicrystalline* polymer crystallizes from the isotropic melt in a NIL line grating mold, the basic structural element to consider is the *nascent crystal*. Because polymer crystals are usually elongated along their fast growth axis direction, alignment of this crystal axis along the grating lines is favored for geometrical reasons. This is, for instance, the case for PVDF crystallized in complete confinement in its orthorhombic α-form (Hu et al. 2005), whose fast b-axis is found by electron diffraction to align parallel to the grooves of the NIL mold (Figure 18.13a). Interestingly, when NIL molds containing curved grooves are used, the b-axis follows these curved tracks (Hu and Jonas, unpublished). This can only be possible if the resulting curved nanowires are polycrystalline. Note that graphoepitaxy in linear channels only aligns one crystal axis, leaving open the issue of the rotation of the crystal about this axis. It was found for NIL-imprinted α-PVDF that the chain c-axis is perpendicular to the substrate for films below about 100 nm (flat-on lamellae), whereas in thicker films, the crystals have their chain c-axis parallel to the substrate (edge-on lamellae) (Hu and Jonas, unpublished). In both cases, however, the b-axis remains parallel to the grooves. The variation from a flat-on to an edge-on orientation upon increasing film thickness is frequent with polymers and was tentatively rationalized recently (Wang et al. 2008).

For polymers imprinted in a *liquid crystalline phase*, then cooled into their crystalline phase, the basic structural element to consider differs depending on the detailed microstructure of the liquid crystalline phase. Poly(9,9-dioctylfluorene) (F8) was imprinted in its nematic liquid crystalline phase, in which chain axes tend to orient parallel to a director with no other form of long-range ordering (Hu et al. 2007). Thus, the basic structural element is simply a *rod-like chain*, which for entropic and geometrical reasons aligns parallel to the grooves. After cooling and crystallization, the preferential ordering is maintained, and the chain axis (which is the c-axis in the low-temperature crystal phase) remains parallel to the grooves, hence to the axis of the resulting nanowire (Figure 18.13b). The preferential alignment translates into polarized light emission, with an R factor of 0.65–0.7 testifying for a high degree of ordering. This is in stark contrast with the absence of preferential ordering reported when F8 was imprinted in partial confinement (Song et al. 2008), and shows the importance of full confinement for NIL-induced graphoepitaxy.

The semiconducting poly(3,3‴-didodecyl-quaterthiophene) (PQT) was also imprinted in its liquid crystalline phase under full confinement (Hu et al. 2007). PQT chains form in the liquid crystalline phase supramolecular rod-like nanostructures, in which extended chains pack with their π-stacking direction along the rod axis. The basic structural element is thus the *rod-like nanostructure*, whose long axis aligns parallel to the grooves of the mold. As a consequence, the π-stacking direction is parallel to the axis of the imprinted nanowire, and this orientation is kept when the polymer crystallizes on cooling. Therefore, the b-axis of the crystal, which is the π-stacking direction, is similarly aligned at the end of the process (Figure 18.13c), whereas the chain c-axis direction is perpendicular to the nanowire axis. The situation is therefore entirely different from F8, where the chain axis was aligned parallel to the nanowire axis. In the present case, this specific crystal setting is interesting for applications, because the PQT b-axis is the axis of high carrier mobility. A nanowire-based field-effect transistor was thus realized by nanoimprint, and it was demonstrated that the hole mobility is increased by a factor of 1.7 along the nanowire axis compared with an isotropic PQT film. Essentially identical results were reported later for nanoimprinted poly (3-hexylthiophene) (P3HT), which adopts a similar morphology as PQT (Aryal et al. 2009).

For rigid conjugated *amorphous* polymers, which cannot crystallize and do not adopt a liquid crystalline phase, the basic structural element to consider during imprinting is simply their *Kuhn segment*. The flexibility of a polymer is quantified by its persistence length L_p (Grosberg and Khokhlov 1994), which is the average distance over which the local orientation of the chain persists. Real chains can be modeled as equivalent freely jointed chains having rigid (Kuhn) segments of length $b_K = 2L_p$ connected by freely rotating junctions (Rubinstein and Colby 2003). Conjugated polymers have large persistence lengths, and therefore long Kuhn segments. These will tend to align parallel to the groove direction during imprinting. The resulting nanowire thus contains chains whose axes are parallel to the nanowire axis, which is favorable for conductivity. For polypyrrole (PPy), for instance, the conductivity was improved by a factor of 1.7 in the imprinted nanowires compared with the isotropic film (Hu et al. 2007), and this was correlated to the preferential chain alignment observed by birefringence measurements (Figure 18.13d).

As a final example of graphoepitaxial alignment by NIL in full confinement, the supramolecular assembly of a phase-separated

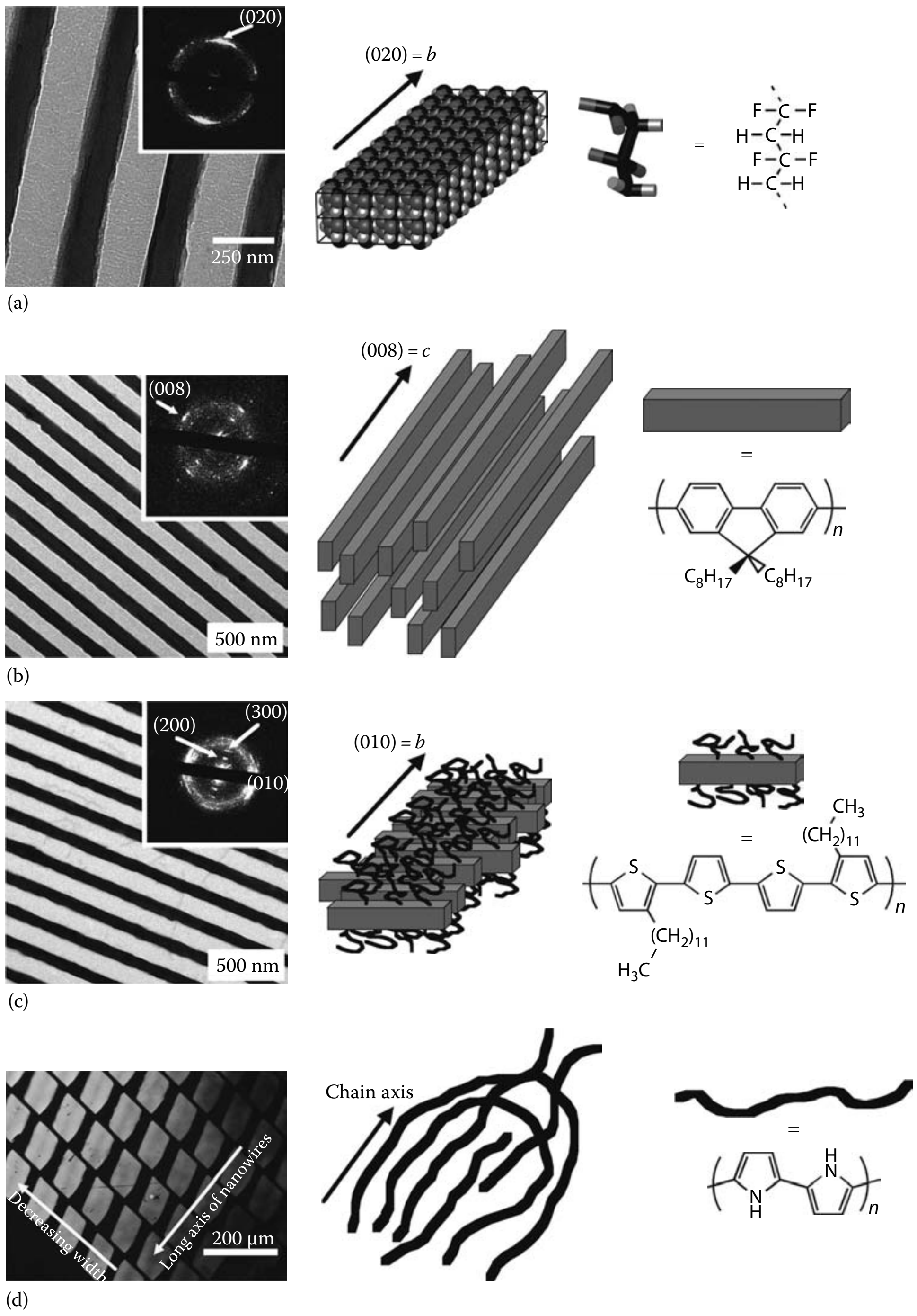
(020)
250 nm
(020) = b
=
F–C–F
H–C–H
F–C–F
H–C–H
(a)
(008)
500 nm
(008) = c
=
n
C8H17
C8H17
(b)
(300)
(200)
(010)
500 nm
(010) = b
=
CH3
(CH2)11
S
S
S
S
n
(CH2)11
H3C
(c)
Chain axis
=
H
N
N
H
n
Decreasing width
Long axis of nanowires
200 µm
(d)

FIGURE 18.13 Selected examples of functional polymers aligned preferentially by NIL in complete confinement. The left column contains transmission electron microscopy images of (a) nanowires of PVDF, (b) electroluminescent F8, (c) semiconducting PQT, and (d) a polarized microscopy image of birefringent arrays of nanowires of conducting PPy (the nanowire axes are at 45° with respect to the polarizer and analyzer directions). Insets in (a)–(c) are the corresponding electron diffraction, showing the crystallographic axis aligned parallel to the line grating direction of the mold. The right column presents schematic drawings of the basic structural elements aligned during imprinting (a nascent crystal for semicrystalline PVDF, rod-like chains of the nematic liquid crystalline phase of F8, a supramolecular rod of π-stacked chains in the liquid crystalline phase of PQT, and chain segments of the amorphous PPy). (Adapted from Hu, Z. et al., *Nano Lett*., 5, 1738, 2005; Hu, Z. et al., *Nano Lett*., 7, 3639, 2007.)

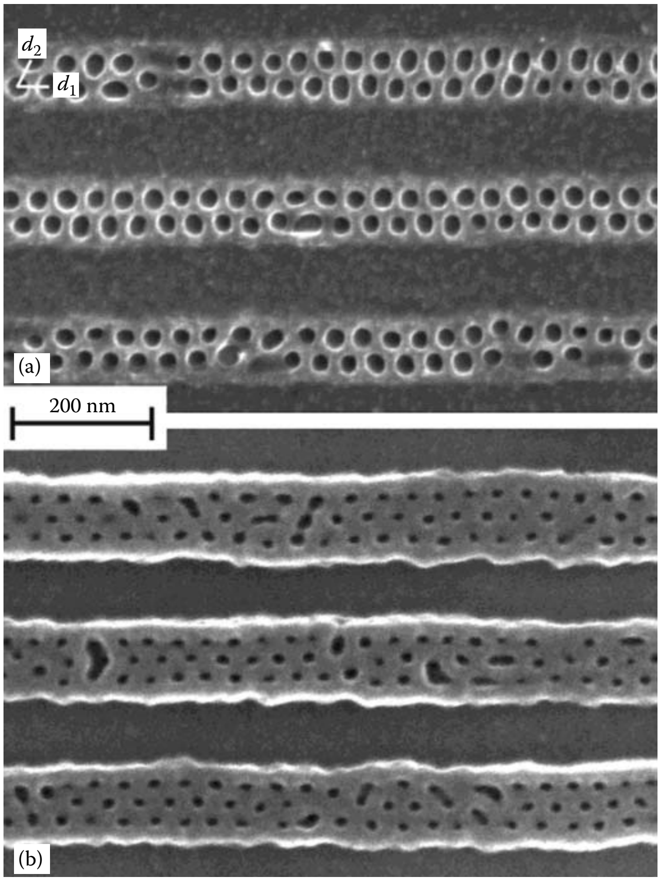

FIGURE 18.14 SEM images of a (PS-*b*-PMMA) diblock copolymer imprinted in full confinement in its hexagonal mesophase. The PMMA cylinders were etched away after mold removal, showing the graphoepitaxial alignment of the mesophase. (Reprinted from Li, H.-W. and Huck, W.T.S., *Nano Lett*., 4, 1633, 2004. With permission.)

block copolymer is presented (Li and Huck 2004). The selected poly(styrene)-*block*-poly(methyl methacrylate) (PS-*b*-PMMA) copolymer consists of flexible amorphous blocks, which do not align by themselves. However, due to the microphase separation of the two incompatible blocks, the PMMA regions form nanocylinders located on the nodes of a hexagonal lattice, within a continuous PS matrix (Hamley 1998). Such microphase-separated copolymers are known to form graphoepitaxially oriented supramolecular crystals (Cheng et al. 2006, Bita et al. 2008). They should thus also align in NIL. Here, the proper structural element is the unit cell of the supramolecular crystal, whose unit cell parameters are in the 10 nm range and above. The axes of the block copolymer unit cell were indeed found to be parallel to the direction of the grooves of the mold (Figure 18.14), although which axis is parallel to the grooves depends on the thickness of the starting block copolymer film (Li and Huck 2004). However, the regularity of the packing was limited, most probably because the dimensions of the grooves were not exactly adapted to the natural period of the block copolymer (Li and Huck 2004). Nevertheless, this example illustrates nicely how one can combine two lithographic techniques together, namely, NIL and block copolymer lithography, since the PMMA can be etched away with UV radiation after imprinting, giving rise to a second level of patterning.

So far, all examples given above referred to line gratings as molds. Other feature shapes can be selected, but the success or failure of graphoepitaxy will depend on the match between the intrinsic size of the basic structural element and the shape of the nanofeatures of the mold (Givargizov 1994). When square cavities are selected instead of grooves, it is obvious that one cannot expect anymore preferential alignment in the plane of the film. However, this does not preclude ordering effects from appearing. This was reported for a statistical copolymer poly(vinylidene fluoride-*stat*-trifluoroethylene) [P(VDF-TrFE)], which was imprinted in its paraelectric liquid crystalline phase, then crystallized by cooling into the ferroelectric pseudohexagonal β-phase (Hu et al. 2009). The mold consisted of square nanocavities about 100 nm in lateral size, allowing to shape the polymer film into a dense array of nanosquares (Figure 18.15a). As expected, there is no preferential orientation in the plane of the film after imprint; however, the polar *b*-axis is aligned almost vertically, and the crystalline perfection is improved. This results in a much lower voltage than usually required to switch the direction of the electrical dipole moment. Figure 18.15b is a map of local polarization acquired by piezoresponse force microscopy (PFM), after writing the word “FeRAM” with a positive or negative bias (5 V) in the array of P(VDF-TrFE) imprinted nanostructures. Interestingly,

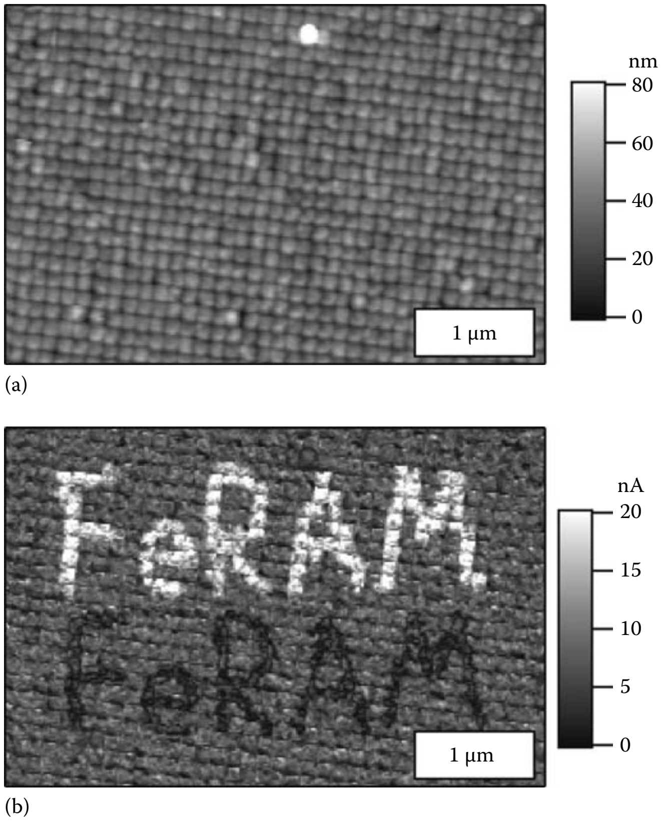

FIGURE 18.15 Array of ferroelectric P(VDF-TrFE) nanostructures nanoimprinted in complete confinement. The orientation of the polar *b*-axis is close to vertical. (a) AFM topography. (b) Piezoresponse amplitude of the same region after local poling with a positive (bright letters) or negative (dark letters) 5 V voltage. The piezoresponse amplitude is proportional to the local electric dipole moment, and was measured by PFM. (Adapted from Hu, Z. et al., *Nat. Mater.*, 8, 62, 2009.)

when nanoimprinting is performed with micro- instead of nanometer-sized cavities, no preferential orientation of the *b*-axis is noted, and the performance of the device is not improved compared with the macroscopic film (Zhang et al. 2007). This underlines the importance of confining crystallization to dimensions comparable to the intrinsic crystal size for optimal results. This example also illustrates that the combination of shaping and crystalline improvement afforded by NIL is extremely useful for device development. Indeed, high-density arrays of ferroelectric polymer crystals exhibiting a low switching voltage are attractive for the development of cheap organic random-access memories. Such systems are considered to be superior to electrically erasable and programmable read-only memories and to Flash memories in terms of write-access time and power consumption (Sheikholeslami and Gulak 2000).

18.5 Conclusions and Perspectives

NIL is by now a well-established lithographic technique, expected to play a significant role in the development of semiconductor technology. Attractive features are the parallelism afforded by the methodology, its potential low-cost, its high resolution, its versatility toward a large range of materials, the possibility to mix it with photoresist technology, and its compatibility with integration technology. However, the potential of NIL as a method to control the shape and internal order of functional materials is only being realized currently. NIL was shown in this chapter to be able to shape functional materials into gratings, arrays, nanowires, and other interesting nanoobjects for applications in optics or microelectronics. Far from being limited to polymeric materials, NIL is further opening its application range to biomacromolecules, sol–gel precursors of inorganic ceramics and glasses, and even hard crystalline solids such as silicon. When applied to materials capable to self-organize or crystallize, NIL is also capable to orient graphoepitaxially the basic structural elements of the material, provided it be performed under full confinement. In much rarer cases such as for liquid crystals of low molar mass, partial confinement may also lead to preferential orientation. The control over the internal structure of nanomolded materials translates into improved device performance. Thus, it was shown that NIL-induced graphoepitaxy results in polarized light emission for electroluminescent polymers, increased mobility of charge carriers in semiconducting polymers, increased conductivity for conjugated conducting polymers, or easier switching of electrical dipole moments for ferroelectric materials. It was also demonstrated in this chapter that the principles governing graphoepitaxial orientation can be rationalized based on the knowledge of the structure of the material and on the notion of basic structural element. This should help device designers to predict the effects of NIL on the microstructure of materials, and therefore to use fully the power of this promising nanoprocessing method.

Acknowledgments

The authors acknowledge the generous financial support from the Wallonia Region (Nanolitho, Nanosens, and Nanotic projects), the French Community of Belgium (ARC Nanorg and Dynanomove), the Belgian Federal Science Policy (IUAP SC^2 and FS^2), the Belgian National Fund for Scientific Research, the Solvay company through the Fondation Louvain, and the European Commission (STREP Metamos, NoE FAME), which permitted part of the research mentioned here to be performed. Access to the WinFab clean rooms is also gratefully acknowledged, as well as stimulating discussions and collaborations with colleagues and students in UCLouvain and elsewhere, too numerous to be all cited here.

References

Alcoutlabi, M. and McKenna, G. B. 2005. Effects of confinement on material behaviour at the nanometre size scale. *J. Phys. Condens. Matter* 17: R461–R524.

Aryal, M., Buyukserin, F., Mielczarek, K., Zhao, X.-M., Gao, J., Zhakidov, A., and Hu, W. W. 2008. Imprinted larger-scale high density polymer nanopillars for organic solar cells. *J. Vac. Sci. Technol. B* 26: 2562–2566.

Aryal, M., Trivedi, K., Hu, W. W. 2009. Nano-confinement induced chain alignment in ordered P3HT nonostructures defined by nanoimprint lithography. ACS Nano 3: 3085–3090.

Bassett, D. C. 1984. Electron microscopy and spherulitic organization in polymers. *CRC Crit. Rev. Solid State Mater. Sci*. 12: 97–163.

Bergmann, L. and Schaefer, C. 1999. *Optics of Waves and Particles*. Berlin, Germany: Walter de Gruyter.

Biebyuck, H. A., Larsen, N. B., Delamarche, E., and Michel, B. 1997. Lithography beyond light: Microcontact printing with monolayer resists. *IBM J. Res. Dev*. 41: 159–170.

Bita, I., Yang, J. K. W., Jung, Y. S., Ross, C. A., Thomas, E. L., and Berggren, K. K. 2008. Graphoepitaxy of self-assembled block copolymers on two-dimensional periodic patterned templates. *Science* 321: 939–943.

Born, M. and Wolf, E. 1980. *Principles of Optics*, 6th edition. Oxford, NY: Pergamon Press.

Campbell, S. A. 2001. *The Science and Engineering of Microelectronic Fabrication*. New York: Oxford University Press.

Chao, W. L., Harteneck, B. D., Liddle, J. A., Anderson, E. H., and Attwood, D. T. 2005. Soft X-ray microscopy at a spatial resolution better than 15 nm. *Nature* 435: 1210–1213.

Cheng, J. Y., Zhang, F., Smith, H. I., Vancso, G. J., and Ross, C. A. 2006. Pattern registration between spherical block-copolymer domains and topographical templates. *Adv. Mater*. 18: 597–601.

Cheyns, D., Vasseur, K., Rolin, C., Genoe, J., Poortmans, J., and Heremans, P. 2008. Nanoimprinted semiconducting polymer films with 50 nm features and their application to organic heterojunction solar cells. *Nanotechnology* 19: 424016.

Chou, S. Y., Krauss, P. R., and Renstrom, P. J. 1995. Imprint of sub-25 nm vias and trenches in polymers. *Appl. Phys. Lett*. 67: 3114–3116.

Chou, S. Y., Krauss, P. R., and Renstrom, P. J. 1996. Imprint lithography with 25 nm resolution. *Science* 252: 85–87.

Chou, S. Y., Krauss, P. R., Zhang, W., Guo, L. J., and Zhuang, L. 1997. Sub-10 nm imprint lithography and applications. *J. Vac. Sci. Technol. B* 15: 2897–2904.

Chou, S. Y., Keimel, C., and Gu, J. 2002. Ultrafast and direct imprint of nanostructures in silicon. *Nature* 417: 835–837.

Cormia, R. L., Price, F. P., and Turnbull, D. 1962. Kinetics of crystal nucleation in polyethylene. *J. Chem. Phys*. 37: 1333–1340.

Eder, G., Janeschitz-Kriegel, H., and Krobath, G. 1989. Shear induced crystallization, a relaxation phenomenon in polymer melts. *Prog. Coll. Polym. Sci*. 80: 1–7.

Ellison, C. J. and Torkelson, J. M. 2003. The distribution of glass-transition temperatures in nanoscopically confined glass formers. *Nat. Mater*. 2: 695–700.

Gates, B. D., Xu, Q. B., Stewart, M., Ryan, D., Wilson, C. G., and Whitesides, G. M. 2005. New approaches to nanofabrication: Molding, printing, and other techniques. *Chem. Rev*. 105: 1171–1196.

Gautam, S., Balijepalli, S., and Rutledge, G. C. 2000. Molecular simulations of the interlamellar phase in polymers: Effect of chain tilt. *Macromolecules* 33: 9136–9145.

Gedde, U. W. 1995. *Polymer Physics*. London, U.K.: Chapman & Hall.

Givargizov, E. I. 1994. Artificial epitaxy (Graphoepitaxy). In *Handbook of Crystal Growth*, Vol. 3, D. T. J. Hurle (Ed.), pp. 941–995. Amsterdam, the Netherlands: Elsevier.

Givargizov, E. I. 2008. Graphoepitaxy as an approach to oriented crystallization on amorphous substrates. *J. Cryst. Growth* 310: 1686–1690.

Grosberg, A. Y. and Khokhlov, A. R. 1994. *Statistical Physics of Macromolecules*. New York: American Institute of Physics.

Guo, L. J. 2004. Recent progress in nanoimprint technology and its applications. *J. Phys. D Appl. Phys*. 37: R123–R141.

Guo, L. J. 2007. Nanoimprint lithography: Methods and material requirements. *Adv. Mater*. 19: 495–513.

Guo, L. J., Cheng, X., and Chao, C. Y. 2002. Fabrication of photonic nanostructures in nonlinear optical polymers. *J. Mod. Opt*. 49: 663–673.

Hamley, I. W. 1998. *The Physics of Block Copolymers*. Oxford, NY: Oxford University Press.

Harnagea, C., Alexe, M., Schilling, J., Choi, J., Wehrspohn, R. B., Hesse, D., and Gosele, U. 2003. Mesoscopic ferroelectric cell arrays prepared by imprint lithography. *Appl. Phys. Lett*. 83: 1827–1829.

Heyderman, L. J., Schift, H., David, C., Gobrecht, J., and Schweizer, T. 2000. Flow behavior of thin polymer films used for hot embossing lithography. *Microelectron. Eng*. 54: 229–245.

Hu, Z., Baralia, G., Bayot, V., Gohy, J.-F., and Jonas, A. M. 2005. Nanoscale control of polymer crystallization by nanoimprint lithography. *Nano Lett*. 5: 1738–1743.

Hu, Z., Muls, B., Gence, L., Serban, D. A., Hofkens, J., Melinte, S., Nysten, B., Demoustier-Champagne, S., and Jonas, A. M. 2007. High-throughput fabrication of organic nanowire devices with preferential internal alignment and improved performance. *Nano Lett*. 7: 3639–3644.

Hu, Z., Tian, M., Nysten, B., and Jonas, A. M. 2009. Regular arrays of highly ordered ferroelectric polymer nanostructures for non-volatile low-voltage memories. *Nat. Mater*. 8: 62–67.

ITRS (International Technology Roadmap for Semiconductors), 2007 edition (http://www.itrs.net).

Jung, G.-Y., Li, Z., Wu, W., Chen, Y., Olynick, D. L., Wang, S.-Y., Tong, W. M., and Williams, R. S. 2005. Vapor-phase self-assembled monolayer for improved mold release in nanoimprint lithography. *Langmuir* 21: 1158–1161.

Kalinin, S. and Gruverman, A. 2007. *Scanning Probe Microscopy: Electrical and Electromechanical Phenomena at the Nanoscale*. New York: Springer.

Kane, R. S., Takayama, S., Ostuni, E., Ingber, D. E., and Whitesides, G. M. 1999. Patterning proteins and cells using soft lithography. *Biomaterials* 20: 2363–2376.

Khang, D.-Y., Yoon, H., and Lee, H. H. 2001. Room-temperature imprint lithography. *Adv. Mater*. 13: 749–752.

Kim, M.-S., Kim, J.-S., Cho, J. C., Shtein, M., Guo, L. J., and Kim, J. 2007. Flexible conjugated polymer photovoltaic cells with controlled heterojunctions fabricated using nanoimprint lithography. *Appl. Phys. Lett.* 90: 123113.

Kimita, S., Sakurai, T., Nozue, Y., Kasahara, T., Yamaguchi, N., Karino, T., Shibayama, M., and Kornfiled, J. A. 2007. Molecular basis of the shish-kebab morphology in polymer crystallization. *Science* 316: 1014–1017.

Klar, T. A., Jakobs, S., Dyba, M., Egner, A., and Hell, S. W. 2000. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proc. Natl. Acad. Sci. U.S.A.* 97: 8206–8210.

Ko, S. H., Park, I., Pan, H., Grigoropoulos, C. P., Pisano, A. P., Luscombe, C. K., and Fréchet, J. M. J. 2007. Direct Nanoimprinting of metal nanoparticles for nanoscale electronics fabrication. *Nano Lett.* 7: 1869–1877.

Kumar, A. and Whitesides, G. M. 1993. Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol ink followed by chemical etching. *Appl. Phys. Lett.* 63: 2002–2004.

Kumaraswamy, G., Issaian, A. M., and Kornfield, J. A. 1999. Shear-enhanced crystallization in isotactic polypropylene. 1. Correspondence between in situ rheo-optics and ex situ structure determination. *Macromolecules* 32: 7537–7547.

Leung, O. M. and Goh, M. C. 1992. Orientational ordering of polymers by atomic force microscope tip-surface interaction. *Science* 255: 64–66.

Li, H.-W. and Huck, W. T. S. 2004. Ordered block-copolymer assembly using nanoimprint lithography. *Nano Lett.* 4: 1633–1636.

Loo, Y. L., Register, R. A., and Ryan, A. J. 2002. Modes of crystallization in block copolymer microdomains: Breakout, templated, and confined. *Macromolecules* 35: 2365–2374.

Mahalingam, V., Onclin, S., Peter, M., Ravoo, B. J., Huskens, J., and Reinhoudt, D. N. 2004. Directed self-assembly of functionalized silica nanoparticles on molecular printboards through multivalent supramolecular interactions. *Langmuir* 20: 11756–11762.

Maier, G. 2001. Low dielectric constant polymers for microelectronics. *Prog. Polym. Sci.* 26: 3–65.

Mandelkern, L. 1979. Relation between properties and molecular morphology of semicrystalline polymers. *Faraday Discuss.* 68: 310–319.

Massa, M. V. and Dalnoki-Veress, K. 2004. Homogeneous crystallization of poly(ethylene oxide) confined to droplets: The dependence of the crystal nucleation rate on length scale and temperature. *Phys. Rev. Lett.* 92: 255509.

Mele, E., Di Benedetto, F., Persano, L., Cingolani, R., and Pisignano, D. 2005. Multilevel, room temperature nanoimprint lithography for conjugated polymer-based photonics. *Nano Lett.* 5: 1915–1919.

Mele, E., Camposeo, A., Stabile, R., Del Carro, P., Di Benedetto, F., Persano, L., Cingolani, R., and Pisignano, D. 2006. Polymeric distributed feedback lasers by room-temperature nanoimprint lithography. *Appl. Phys. Lett.* 89: 131109.

Miller, R. L. (Ed.). 1979. *Flow-Induced Crystallization in Polymer Systems*. New York: Gordon & Breach.

Mouthuy, P.-O., Melinte, S., Geerts, Y. H., and Jonas, A. M. 2007. Uniaxial alignment of nanoconfined columnar mesophases. *Nano Lett.* 7: 2627–2632.

Muller, A. J., Balsamo, V., Arnal, M. L., Jakob, T., Schmalz, H., and Abetz, V. 2002. Homogeneous nucleation and fractionated crystallization in block copolymers. *Macromolecules* 35: 3048–3058.

Odom, T. W., Love, J. C., Wolfe, D. B., Paul, K. E., and Whitesides, G. M. 2002. Improved pattern transfer in soft lithography using composite stamps. *Langmuir* 18: 5314–5320.

Ohtake, T., Nakamatsu, K., Matsui, S., Tabata, H., and Kawai, T. 2004. DNA nanopatterning with self-organization by using nanoimprint. *J. Vac. Sci. Technol. B* 22: 3275–3278.

Okerberg, B. C., Soles, C. L., Douglas, J. F., Ro, H. W., Karim, A., and Hines, D. R. 2007. Crystallization of poly(ethylene oxide) patterned by nanoimprint lithography. *Macromolecules* 40: 2968–2970.

Okinaka, M., Tsukagoshi, K., and Aoyagi, Y. 2006. Direct nanoimprint of inorganic-organic hybrid glass. *J. Vac. Sci. Technol. B* 24: 1402–1404.

Owen G. 1985. Electron lithography for the fabrication of microelectronic devices. *Rep. Prog. Phys.* 48: 795–851.

Park, I., Cheng, J., Pusano, A. P., Lee, E.-S., and Jeong, J.-H. 2007. Low temperature, low pressure nanoimprinting of chitosan as a biomaterial for bionanotechnology applications. *Appl. Phys. Lett.* 90: 093902.

Park, I., Ko, S. H., Pan, H., Grigoropoulos, C. P., Pisano, A. P., Fréchet, J. M. J., Lee, E.-S., and Jeong, J.-H. 2008. Nanoscale patterning and electronics on flexible substrate by direct nanoimprinting of metallic nanoparticles. *Adv. Mater.* 20: 489–496.

Peroz, C., Heitz, C., Barthel, E., Sondergard, E., and Goletto, V. 2007. Glass nanostructures fabricated by soft thermal nanoimprint. *J. Vac. Sci. Technol. B* 25: L27–L30.

Pisignano, D., Persano, L., Visconti, P., Cingolani, R., Gigli, G., Barbarella, G., and Favaretto, L. 2003. Oligomer-based organic distributed feedback lasers by room-temperature nanoimprint lithography. *Appl. Phys. Lett.* 83: 2545–2547.

Pisignano, D., Persano, L., Raganato M. F., Visconti, P., Cingolani, R., Barbarella, G., Favaretto, L., and Gigli, G. 2004. Room temperature nanoimprint lithography of non-thermoplastic organic film. *Adv. Mater.* 16: 525–529.

Rai-Choudhury, P. 1997. *Handbook of Microlithography, Micromachining, and Microfabrication: Vol. 1: Microlithography*. Bellingham, WA: SPIE Optical Engineering Press.

Reiter, G. and Strobl, G. R. (Eds.). 2006. *Progress in Understanding of Polymer Crystallization*. Lecture Notes in Physics, Vol. 714. Berlin, Germany: Springer.

Reiter, G., Castelein, G., Sommer, J. U., Rottele, A., and Thurn-Albrecht, T. 2001. Direct visualization of random crystallization and melting in arrays of nanometer-size polymer crystals. *Phys. Rev. Lett.* 87: 226101.

Resnick, D. J. 2007. Imprint lithography. In *Microlithography Science and Technology*, K. Suzuki and B. W. Smith (Eds.), pp. 465–499. Boca Raton, FL: Taylor & Francis.

Resnick, D. J., Dauksher, W. J., Mancini, D., Nordquist, K. J., Bailey, T. C., Johnson, S., Stacey, N., Ekerdt, J. G., Willson, C. G., Sreenivasan, S. V., and Schumaker, N. 2003. Imprint lithography for integrated circuit fabrication. *J. Vac. Sci. Technol. B* 21: 2624–2631.

Ro, H. W., Jones, R. L., Peng, H., Hines, D. R., Lee, H.-J., Lin, E. K., Karim, A., Yoon, D. Y., Gidley, D. W., and Soles, C. L. 2007. The direct patterning of nanoporous interlayer dielectric insulator films by nanoimprint lithography. *Adv. Mater.* 19: 2919–2924.

Ro, H. W., Peng, H., Niihara, K., Lee, H.-J., Lin, E. K., Karim, A., Gidley, D. W., Jinnai, H., Yoon, D. Y., Gidley, D. W., and Soles, C. L. 2008. Self-sealing of nanoporous low dielectric constant patterns fabricated by nanoimprint lithography. *Adv. Mater.* 20: 1934–1939.

Rowland, H. D., Sun, A. C., Schunk, P. R., and King, W. P. 2005. Impact of polymer film thickness and cavity size on polymer flow during embossing: Toward process design rules for nanoimprint lithography. *J. Micromech. Microeng.* 15: 2414–2425.

Rowland, H. D., King, W. P., Pethica, J. B., and Cross, G. L. W. 2008. Molecular confinement accelerates deformation of entangled polymers during squeeze flow. *Science* 322: 720–724.

Rubinstein, M. and Colby, R. H. 2003. *Polymer Physics*. Oxford, NY: Oxford University Press.

Ruchhoeft, P., Colburn, M., Choi, B., Nounu, H., Johnson, S., Bailey, T., Damle, S., Stewart, M., Ekerdt, J., Sreenivasan, S. V., Wolfe, J. C., and Willson, C. G. 1999. Patterning curved surfaces: Template generation by ion beam proximity lithography and relief transfer by step and flash imprint lithography. *J. Vac. Sci. Technol. B* 17: 2965–2969.

Schift, H. 2008. Nanoimprint lithography: An old story in modern times? A review. *J. Vac. Sci. Technol. B* 26: 458–480.

Schift, H. and Heyderman, L. J. 2003. Nanorheology. Squeeze flow in hot embossing of thin films. In *Alternative Lithography. Unleashing the Potentials of Nanotechnology*, C. M. Sotomayor Torres (Ed.), pp. 47–76. New York: Kluwer Academic/Plenum.

Schift, H. and Kristensen, A. 2007. Nanoimprint lithography. In *Springer Handbook of Nanotechnology*, 2nd edition, B. Bushan (Ed.), pp. 239–278. Berlin, Germany: Springer.

Schift, H., Heyderman, L. J., Auf der Maur, M., and Gobrecht, J. 2001. Pattern formation in hot embossing of thin polymer films. *Nanotechnology* 12: 173–177.

Schift, H., Saxer, S., Park, S., Padeste, C., Peiles, U., and Gobrecht, J. 2005. Controlled co-evaporation of silanes for nanoimprint stamps. *Nanotechnology* 16: S171–S175.

Schmid, S. A., Yim, K. H., Chang, M. H., Zheng, Z., Huck, W. T. S., Friend, R. H., Kim, J. S., and Herz, L. M. 2008. Polarization anisotropy dynamics for thin films of a conjugated polymer aligned by nanoimprinting. *Phys. Rev. B* 77: 115338.

Schulz, H., Wissen, M., and Scheer, H.-C. 2003. Local mass transport and its effect on global pattern replication during hot embossing. *Microelectron. Eng.* 67–68: 657–663.

Sheftal, N. N. and Bouzynin, N. A. 1972. Preferred orientation of crystallites on the substrate and effect of scratches. *Vestn. Mosk. Univ. Ser. Geol.* 27: 102–104.

Sheikholeslami, A. and Gulak, P. G. 2000. A survey of circuit innovations in ferroelectric random-access memories. *Proc. IEEE* 88: 667–688.

Smith, H. I. and Flanders, D. C. 1978. Oriented crystal growth on amorphous substrates using artificial surface-relief gratings. *Appl. Phys. Lett.* 32: 349–350.

Somani, R. H., Hsiao, B. S., Nogales, A., Srinivas, S., Tsou, A. H., Sics, I., Balta-Calleja, F. J., and Ezquerra, T. A. 2000. Structure development during shear flow-induced crystallization of i-PP: In-situ small-angle x-ray scattering study. *Macromolecules* 33: 9385–9394.

Song, M. H., Wenger, B., and Friend, R. H. 2008. Tuning the wavelength of lasing emission in organic semiconducting laser by the orientation of liquid crystalline conjugated polymer. *J. Appl. Phys.* 104: 033107.

Steinhart, M., Göring, P., Dernaika, H., Prabhukaran, M., Gösele, U., Hempel, E., and Thurn-Albrecht, T. 2006. Coherent kinetic control over crystal orientation in macroscopic ensembles of polymer nanorods and nanotubes. *Phys. Rev. Lett.* 97: 027801.

Stewart, M. D. and Willson, C. G. 2005. Imprint materials for nanoscale devices. *MRS Bull.* 30: 947–951.

Studer, V., Pepin, A., and Chen, Y. 2002. Nanoembossing of thermoplastic polymers for microfluidic applications. *Appl. Phys. Lett.* 80: 3614–3616.

Suzuki, K. 2007. Electron beam lithography systems. In *Microlithography Science and Technology*, K. Suzuki and B. W. Smith (Eds.), pp. 465–499. Boca Raton, FL: Taylor & Francis.

Ting, C.-J., Huang, M.-C., Tsai, H.-Y., Chou, C.-P., and Fu, C.-C. 2008a. Low cost fabrication of the large-area anti-reflection films from polymer by nanoimprint/hot-embossing technology. *Nanotechnology* 19: 205301.

Ting, C.-J., Chang, F.-Y., Chen, C.-F., and Chou, C.-P. 2008b. Fabrication of an antireflective polymer optical film with subwavelength structures using a roll-to-roll micro-replication process. *J. Micromech. Microeng.* 18: 075001.

Toney, M. F., Russell, T. P., Logan, J. A., Kikuchi, H., Sands, J. M., and Kumar, S. K. 1995. Near-surface alignment of polymers in rubbed films. *Nature* 374: 709–711.

Torres, C. M. (Ed.). 2003. *Alternative Lithography. Unleashing the Potentials of Nanotechnology*. New York: Kluwer Academic/Plenum.

Ueno, T. 2007. X-ray lithography. In *Microlithography Science and Technology*, K. Suzuki and B. W. Smith (Eds.), pp. 361–382. Boca Raton, FL: Taylor & Francis.

Voicu, N. E., Ludwigs, S., Crossland, E. J. W., Andrew, P., and Steiner, U. 2007. Solvent-vapor-assisted imprint lithography. *Adv. Mater.* 19: 757–767.

Wang, J., Sun, X., Chen, L., and Chou, S. Y. 1999. Direct nanoimprint of submicron organic light-emitting structures. *Appl. Phys. Lett*. 75: 2767–2769.

Wang, J., Sun, X., Chen, L., Zhuang, L., and Chou, S. Y. 2000. Molecular alignment in submicron patterned polymer matrix using nanoimprint lithography. *Appl. Phys. Lett*. 77: 166–168.

Wang, Y., Chan, C. M., Ng, K. M., and Li, L. 2008. What controls the lamellar orientation at the surface of polymer films during crystallization? *Macromolecules* 41: 2548–2553.

Woo, E., Huh, J., Jeong, Y. G., and Shin, K. 2007. From homogeneous to heterogeneous nucleation of chain molecules under nanoscopic cylindrical confinement. *Phys. Rev. Lett*. 98: 136103.

Wunderlich, B. 1976. *Macromolecular Physics. Vol. 2. Crystal Nucleation, Growth, Annealing*. New York: Academic Press.

Wurm, S. and Gwyn, C. 2007. EUV lithography. In *Microlithography Science and Technology*, K. Suzuki and B. W. Smith (Eds.), pp. 383–464. Boca Raton, FL: Taylor & Francis.

Xia, Y. N. and Whitesides, G. M. 1998. Soft lithography. *Annu. Rev. Mater. Sci*. 28: 153–184.

Xia, Y. N., Rogers, J. A., Paul, K. E., and Whitesides, G. M. 1999. Unconventional methods for fabricating and patterning nanostructures. *Chem. Rev*. 99: 1823–1848.

Zhang, L., Ducharme, S., and Li, J. 2007. Microimprinting and ferroelectric properties of poly(vinylidene fluoride-trifluoroethylene) copolymer films. *Appl. Phys. Lett*. 91: 172906.

Zheng, Z., Yim, K.-H., Saifullah, M. S. M., Welland, M. E., Friend, R. H., Kim, J.-S., and Huck, W. T. S. 2007. Uniaxial alignment of liquid-crystalline conjugated polymers by nanoconfinement. *Nano Lett*. 7: 987–992.

19

Nanoelectronics Lithography

Stephen Knight
National Institute of Standards and Technology

Vivek M. Prabhu
National Institute of Standards and Technology

John H. Burnett
National Institute of Standards and Technology

James Alexander Liddle
National Institute of Standards and Technology

Christopher L. Soles
National Institute of Standards and Technology

Alain C. Diebold
University at Albany

19.1 Introduction **19**-1
19.2 Photoresist Technology **19**-3
Fundamentals • Advances by Material Structure • Progress in Resists for EUV • Progress in Resists for 193 nm Immersion Lithography • Concluding Remarks
19.3 Deep Ultraviolet Lithography **19**-8
DUV Lithography Steppers • 193 nm Immersion Lithography • Double Patterning • Ultimate Resolution Limits of DUV Lithography
19.4 Electron-Beam Lithography **19**-14
Resolution • Throughput • Overlay • Cost of Ownership
19.5 Nanoimprint Lithography **19**-17
Variations of NIL • Fundamental Issues • Overlay Accuracy and Control • Technology Examples • Future Perspectives
19.6 Metrology for Nanolithography **19**-24
Advanced Lithographic Processes • Metrology for Advanced Lithography Processes • Proposed New CD Metrology Methods
Abbreviations **19**-27
References **19**-28

19.1 Introduction

The modern integrated circuit (IC), comprising memory, logic processors and analog function devices, are multicomponent and multilevel nanostructures prepared by a series of patterning and pattern-transfer steps. Figure 19.1 shows the cross-sectional hierarchical structure that starts from the smallest feature, the transistor, to dielectrics and metal contacts that are each well defined and must precisely overlay the previous layer. This three-dimensional nanoelectronics structure is manufactured by a rapid patterning process called lithography.

Since the invention of the transistor in 1947 by Bell Labs and Intel's first microprocessor in 1971, modern lithography has enabled the semiconductor industry to shrink device dimensions. The early progress was first quantified by Gordon Moore in 1965 and is now known as Moore's law.[1] While progress in all process steps are required to achieve the continual shrinking of circuit elements, the advancements in lithography have been the overwhelming driving force. Figure 19.2 shows how Moore's law has guided the industry's systematic increase in the number of transistors per chip since the introduction of the IC. Indeed, now and in the future, the industry will achieve productivity improvement primarily by feature reduction.[2]

While a wide variety of lithography technologies have been developed, optical step and repeat lithography technologies are the predominant methods of printing features on the semiconductor surfaces and overlying the interconnect structures. A number of competing technologies have been explored, such as x-ray[3,4] and flood electron beam (SCALPEL),[5,6] but optical lithography has dominated through vigorous optical technology and tool developments that have left it the lowest cost solution with highest throughput. Direct-write electron-beam technology provides the highest resolution and remains the leading technique for manufacturing the masks used by optical lithography. A novel way of comparing different lithography strategies is by a plot of resolution versus throughput, also known as "Tennant's Law,"[7,8] as plotted in Figure 19.3. In this plot, direct-write approaches, such as single-atom placement by scanning tunneling microscopy (STM), atomic-force microscopy (AFM) tip-induced oxidation, and electron-beam

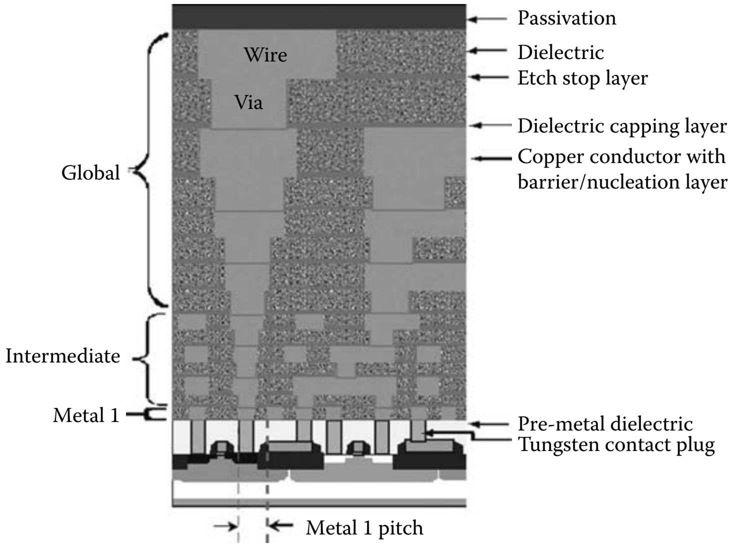

FIGURE 19.1 Cross section of an integrated circuit, showing the active semiconductor and multilayer interconnect levels. (Reproduced from *International Technology Roadmap for Semiconductors*, 2007 edition, SEMATECH, Austin, TX, Figure INTC2, 2007. With permission.)

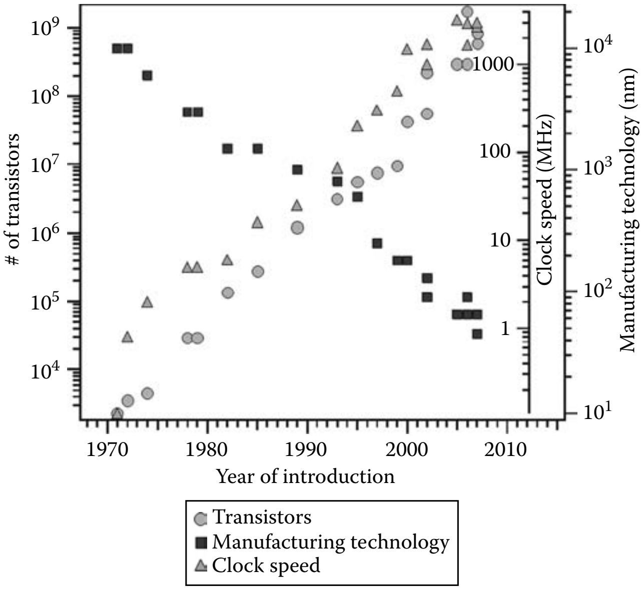

FIGURE 19.2 A composite plot of the scaling of the number of transistors, clock speed, and manufacturing technology versus the year of introduction of Intel Processors. (Data from www.intel.com/technology/timeline.pdf.)

lithography are compared with optical step and repeat lithography that replicate the features of a mask. The continued advancements in optical lithography have pushed the resolution to smaller features at higher pixel throughput, as analyzed by Brunner.[9]

The fundamental guide for optical step and repeat lithography is the Rayleigh equation, which provides the scaling criteria for predicting the smallest optically definable image[2]

$$R = k_1 \frac{\lambda}{\mathrm{NA}} = k_1 \frac{\lambda}{n \sin\theta}; \tag{19.1}$$

where

λ is the source wavelength
n is the medium refractive index
$n \sin\theta$ is the numerical aperture (NA) with the incidence angle (θ)
k_1 is a process dependent factor

While the Rayleigh equation defines the resolution, the ability of the photoresist to replicate the mask features is of critical importance. All approaches and manufacturing tools have taken advantage of the Rayleigh equation to predict transitions from each manufacturing technology node (smallest feature size) up to the diffraction limit. Optical lithography started with the near *ultra*violet (UV) g-line (436 nm) and i-lines (365 nm) of mercury-arc lamps and made way for laser sources in the deep *ultra*violet (DUV) from KrF (248 nm) and ArF (193 nm) excimers. A significant extension, immersion lithography, decreases the 193 nm wavelength by using water as the immersion fluid and is on track to produce features down to 32 nm. The next-generation photoresist materials, described in Section 19.2, for sub 22 nm features expect to image 13.4 nm radiation in extreme *ultra*violet (EUV) with all reflective lithography imaging tools.

While most of this chapter focuses on the leading-edge and next-generation technologies used for defining the finest features in nanoelectronics, it is important to recognize that for most applications it is necessary to connect the nano-world (transistor) to the macro-world (a computer motherboard). Figure 19.1 shows the cross section of an IC. Note that the interconnect sizes increase at subsequent higher levels. Therefore, the previous-generation tools continue to play crucial roles by migrating to the higher levels of interconnect.

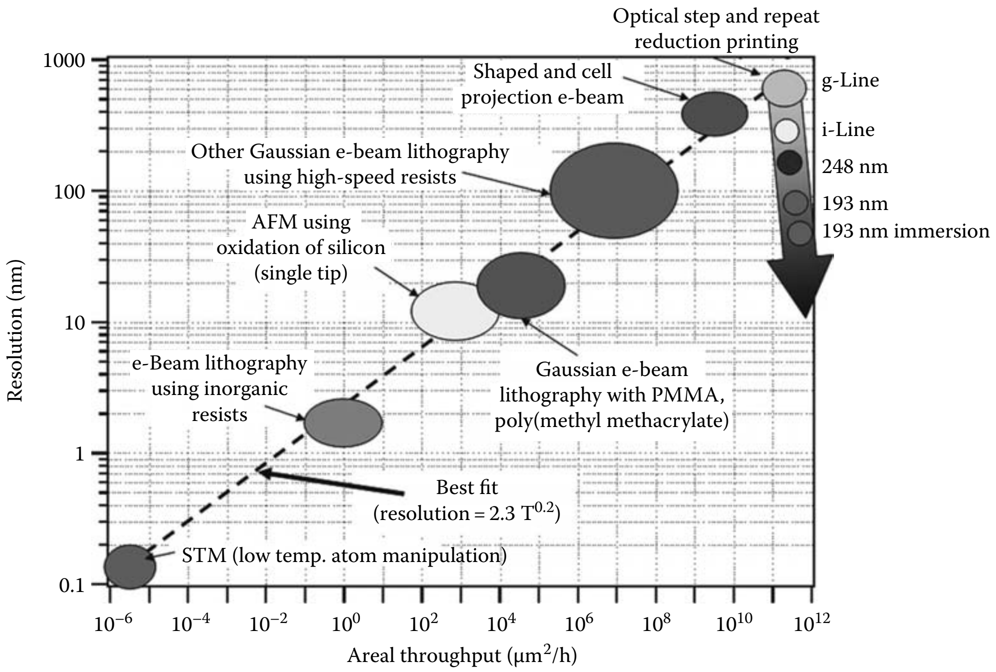

FIGURE 19.3 The progress of optical step and repeat lithography has driven to higher resolution and throughput as described by "Tennant's Law," whereby the resolution (R) versus areal throughput (T) scales as $R = 2.3\,T^{0.2}$ (Adapted from Tennant, D., Limits of conventional lithography, in Timp, G.M. (ed.), *Nanotechnology*, AIP Press, New York, Chapter 4, p. 161, 1999. With permission.)

Microelectronics and now nanoelectronics technology have become a collaborative and competitive worldwide effort, as exemplified by the *International Technology Roadmap for Semiconductors* (ITRS), available to the public at http://www.itrs.net. The ITRS is updated every two years by subject matter experts from the semiconductor manufacturing industry, the tool and materials supplier industries, the factory automation infrastructure, academia, and government agencies. An important guiding aspect of the ITRS roadmap is the identification of the status of the technology nodes and guidance of the phases of research, development, and pre-production. In particular, the 2007 edition identifies potential lithographic solutions out to 2022, with a predicted dynamic random access memory half pitch of 11 nm and FLASH memory half pitch of 9 nm as shown in Figure 19.4.

In this roadmap, several emerging lithography approaches are highlighted. Double patterning with 193 nm DUV water immersion is expected to extend to the 32 nm half pitch era with high-volume production in 2013. Alternatively, contending technologies for 22 nm half pitch are EUV Lithography, 193 nm DUV immersion with higher index fluid and lens materials, maskless lithography (ML2), and nanoimprint lithography (NIL).[10,11] NIL is rapidly emerging as a low-cost, high-resolution, and versatile alternative to optical lithography. Below 22 nm half-pitch, the likely technology solutions are less clear. All the technologies mentioned above, along with new contenders, such as directed self assembly, have credible paths. However, each would have to surmount numerous technical difficulties while limiting exorbitant cost and loss in throughput.

In the rest of this chapter, we describe some of the challenges facing the photoresist materials (Section 19.2) used in optical lithography including some crucial aspects facing these materials as the feature dimensions are reduced to the length scale of the basic photoresist polymers. In Section 19.3, DUV lithography, the basic optics, advancements in steppers, and approaches to extend to higher-resolution, denser features are described. In Section 19.4, electron-beam lithography is covered with respect to the metrics of resolution, throughput, overlay requirements, and cost. Section 19.5 highlights a non-optical lithography approach, NIL. In this section, several variations of NIL for transferring mask features into a photoresist are described. Finally, in Section 19.6, we end with an overview of the metrology requirements for nanolithography. Figure 19.1 is a reminder that one must print and measure a wide-range of feature dimensions, from the nanoscale to the macroscale. Lastly, a separate chapter in this handbook is dedicated to EUV lithography.

19.2 Photoresist Technology

The optical image projected from a mask upon a thin film at the semiconductor wafer plane is the first step of photolithography.[10,12] The thin films that replicate the mask features are called photoresists with basic process as shown in Figure 19.5. The high sensitivity of photoresists to radiation have consistently met the challenges of smaller, high-fidelity features with increased throughput driven by memory and processor chip performance to feature size gains (Moore's law). The etch resistance of the

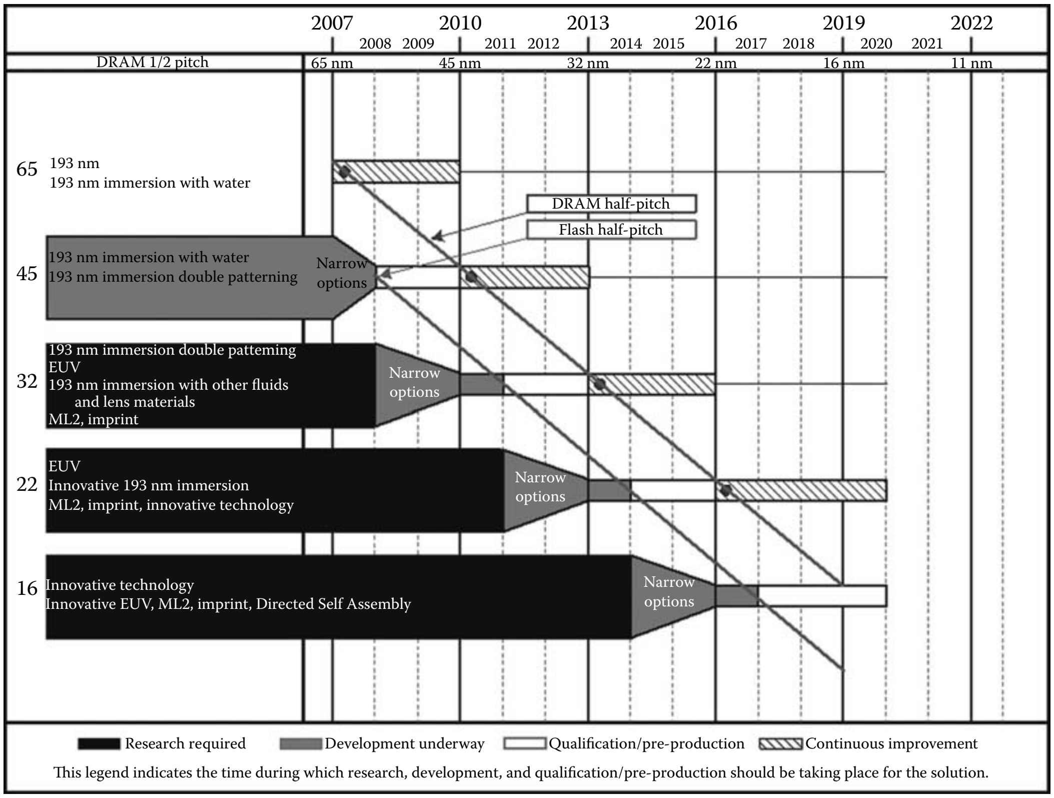

FIGURE 19.4 Roadmap for half pitch scaling. (Reproduced from *International Technology Roadmap for Semiconductors*, 2007 edition, SEMATECH, Austin, TX, Figure LITH5, 2007. With permission.)

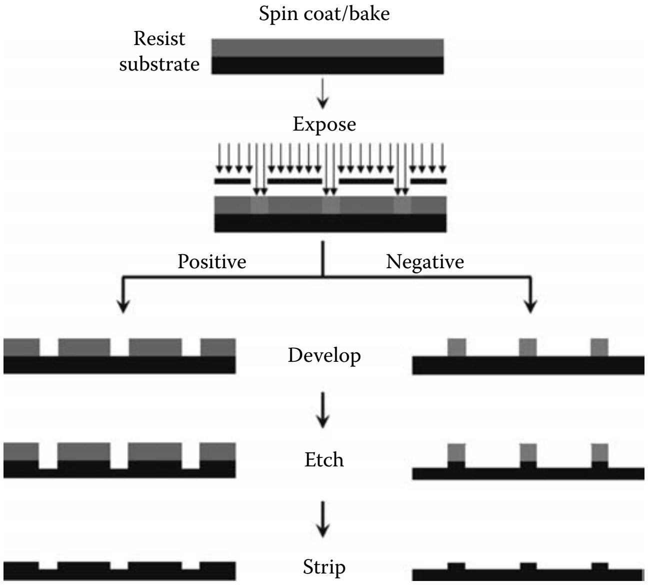

FIGURE 19.5 Schematic of the lithographic process for positive and negative tone resist.

photoresist allows pattern transfer into the underlying semiconductor wafer.

Test structures, such as a line and space pattern as shown in Figure 19.6, must meet criteria as defined by the ITRS roadmap for critical dimension (CD). The CD is the feature size, for example, in a 32 nm half-pitch 1:1 dense line, a line with a width of 32 nm is followed by a space of equal size. The Rayleigh equation defines the resolution; however, the ability of the photoresist to perfectly replicate the mask features is of critical importance. The line-width variations called line-width roughness (LWR) and line-edge roughness (LER) must be reduced to less than 2 nm for 32 nm half pitch (HP) as they impact device performance.[13,14] Therefore, the photoresist materials chemistry plays a substantial role for both resolution and LER. Methods to extend resist resolution by double patterning methods and new photoresist architectures must be leveraged against meeting the LER requirements for sub-22 nm lithography.

19.2.1 Fundamentals

The requirements for advanced photoresists are discussed in the 2007 ITRS roadmap.[14] Specifically, the "Lithography" chapter highlights difficult challenges:

> Resist materials at <32 nm indicates three issues (1) Resist and antireflective coating materials composed of alternatives to PFAS [perfluoroalkysulfonate] compounds, (2) Limits of chemically amplified resist sensitivity for <32 nm half-pitch due to acid diffusion length, and (3) materials with improved dimensional and LWR control.

The details of the photoresist materials chemistry are at the heart of turning the optically defined patterns into three-dimensional

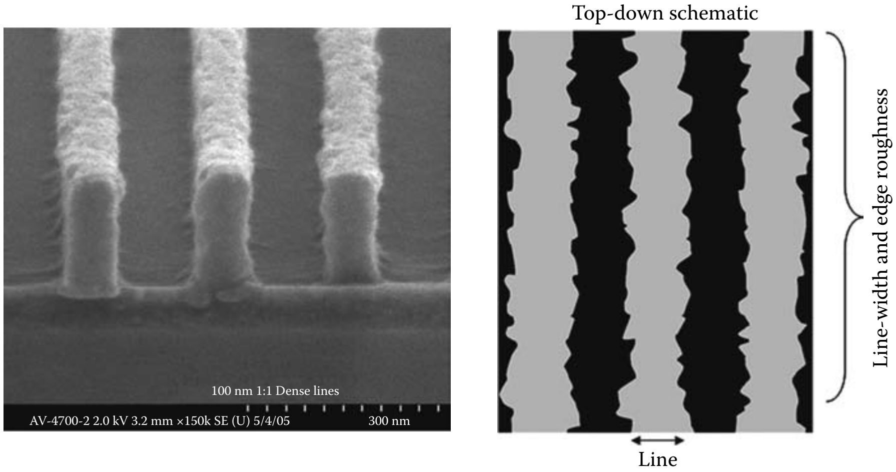
100 nm 1:1 Dense lines
AV-4700-2 2.0 kV 3.2 mm ×150k SE (U) 5/4/05
300 nm
Top-down schematic
Line-width and edge roughness
Line

FIGURE 19.6 Example of a cross-sectional SEM image of a 100 nm 1:1 line-space lithographic line. The fluctuations in the feature critical dimension along the line are quantified by line-edge roughness and line-width roughness metrics as indicated in the schematic.

nanoscale features. Chemically amplified photoresists are formulations of an acid-sensitive polymer film mixed with photoacid generators (PAGs) and other additives, such as base quenchers. Upon exposure through a mask, strong acids are formed by photolysis of the molecularly mixed PAG within the thin film. A post-exposure bake is then applied and the acidic protons (photoacids) diffuse along with the counter-anions and catalyze a deprotection reaction on the acid-sensitive polymer to change the local solubility for development in an aqueous hydroxide solution as shown in Figure 19.7. The photoacids are true catalysts as they are regenerated by each deprotection reaction;[15] hence, the term chemical amplification refers to the cascade of reactions that occur within the photoresist induced by a single photon. Therefore, these photoresists may be used at low exposure doses.

The chemical reaction-diffusion that forms the latent image and development must be understood and controlled at the nanometer length scale. Chemically amplified photoresists are also deposited onto bottom anti-reflection coatings (BARC) to eliminate the effects of standing waves. Interactions and component transport between the BARC and resist layer can lead to loss of profile control or pattern collapse and therefore must also be understood. Detailed studies of these interactions and transport mechanisms are needed to design materials for the successful fabrication of sub-32 nm structures.

As feature sizes are reduced to below 32 nm, a general problem of simultaneously minimizing dose sensitivity, CD, and LER was observed experimentally and theoretically. These observations suggest only two of these metrics may be met at the sacrifice of

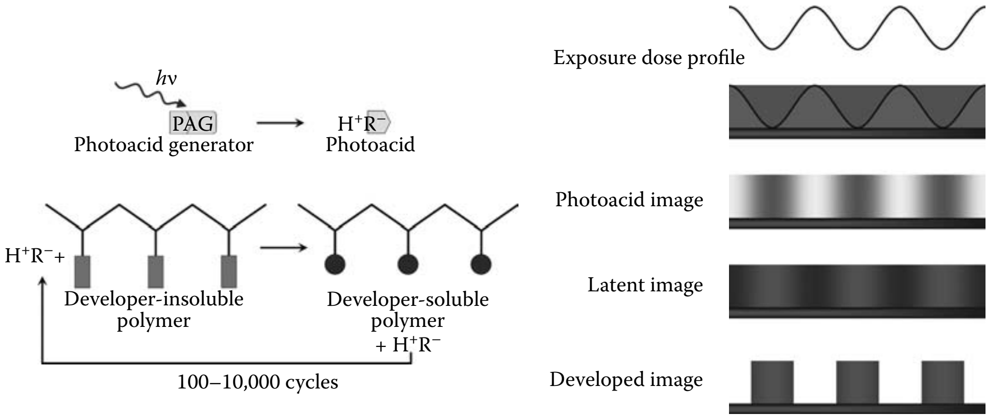
hν
PAG
Photoacid generator
H^+R^-
Photoacid
H^+R^-+
Developer-insoluble polymer
Developer-soluble polymer + H^+R^-
100–10,000 cycles
Exposure dose profile
Photoacid image
Latent image
Developed image

FIGURE 19.7 Schematic of the photolysis process of a photoacid generator that forms a strong acid (H^+R^-) and the subsequent acid catalytic reaction that changes the solubility of the polymer in a developer solution. Schematic of the mask-defined dose profile, subsequent profile in the photoresist film that leads to a photoacid image and chemical latent image formed during the post-exposure bake. The final developed image is formed after selective dissolution in a developer solution as determined by the extent of chemical reaction in the nominally exposed regions.

the third. Several theoretical models[16–18] have been proposed, such as those by Bristol (2007)

$$\mathrm{LER} \approx \frac{1}{\sqrt{q\alpha D_e}}\sqrt{1+\frac{1}{\varepsilon}}\,\frac{1}{\mathrm{LILS}\cdot\delta}, \tag{19.2}$$

where

- q is the number of photons/nm^2
- D_e is the dose at the line-edge (threshold for development)
- α is the fraction of light absorbed
- ε is the quantum efficiency defined by the (molecules of acids produced)/(number of absorbed photons)
- δ is the effective photoacid diffusion length
- LILS is the latent-image log slope

As shown in the schematic of Figure 19.7, the initial photoacid image is followed by a post-exposure bake step whereby the photoacid diffuses and reacts. The extent of the reaction will define a chemical latent image in the photoresist. The local slope of the chemical transformation, or latent-image log slope (LILS), at the point of development is critical to maximize. Qualitatively, a large exposure dose leads to a lower LER at fixed LILS and δ. However, notice that once smaller features are desired (δ < CD), then the photoacid diffusion length should be reduced; hence at fixed D_e, a smaller δ may increase LER. The full problem is a nonlinear reaction-diffusion equation and typically requires additional parameterization to include the effects of photoacid loss, photoacid trapping, and amine quenchers that react and diffuse. However, approaches such as Equation 19.2 provide qualitative insights into the resist problem. In cases when the material specifics were not known, the form of material constant = (half-pitch)^3 × (LER)^2 × (Sensitivity) has been used.[19]

Such resolution limits are linked to the fundamentals of photoacid generation (via D_e, α, ε) and chemical reaction-diffusion during the post-exposure bake that defines the LILS, choice of photoresist chemistry, PAG size, and processing time and the temperature that ultimately determines the photoacid diffusion length and development step that resolves the final feature. Experimental methods have been developed to quantitatively measure many of these parameters. More recently, parameters such as the photoacid diffusion length were estimated by modeling of the lithographic feature power spectrum.[16]

The efficient generation of photoacids is also crucial to photoresist technology. Analytical methods have been developed to measure photoacid generation and quantum efficiency.[20–23] Many of these techniques rely on a well-defined dose upon photoresist thin films followed by analytical approaches to determine the acid concentration via titration with a standard base. More recently, quantifying the quantum efficiency and acid generation mechanisms in EUV photoresists has taken a more important role as increasing the acid concentration or efficiency (Equation 19.2) is a method for reducing LER. For 248 and 193 nm lithography, photoresist polymers were designed for low absorbance, while the PAG absorbs strongly for photolysis. Since organic photoresist polymers absorb strongly in the EUV, the photolysis induced by the direct exposure of the 92 eV photon and secondary electrons are equally important. In fact, the ε for EUV exposure is greater than 1 due to the non-negligible contribution from lower energy secondary electrons. In both DUV and EUV lithography, the photoacid image plays the important role of changing the local solubility of the photoresist polymer in the nominally exposed regions (Figure 19.7).

During the post-exposure bake, the latent image (Figure 19.7) formed is a composition profile of the photoresist within the line:space feature. The true shape of this profile is due to the photoacid catalyzed reaction-diffusion process that cleaves (deprotection) a nonpolar side-group of the photoresist polymer. The influence of the optical image quality on the final developed feature may be probed by varying the aerial image contrast. Hinsberg et al. elegantly showed that the initial photoacid distribution, controlled through the exposure quality by interferometric lithography, significantly affects the printed feature quality.[24] Furthermore, using an image-fading technique, Pawloski et al. identified an apparent resolution limit in the final feature of 193 nm resists as quantified by LER versus the image-log slope (ILS).[25] These changes in the feature quality are also controlled by the reaction-diffusion of the photoacid[12] into the unexposed regions that leads to image spreading or blurring.[26] However, even with an ideal step-exposure condition, mimicked by forming a bilayer stack, the reaction-diffusion process induces an image blur. Controlling the photoacid reaction-diffusion (via LILS and δ) remains an important strategy for improving feature quality.

The transport properties of the photoacid are also influenced by the changes in the local chemical composition that occur during the deprotection reaction.[27] Houle et al. demonstrated that the evolving resist polymer chemistry plays a crucial role in lithographic imaging.[26] An increased photoacid size decreased the apparent diffusion length,[28,29] but also quantitatively increased the LILS.[30,31] However, the catalytic efficiency of the photoacid proton can dominate, such that by increasing the size of the photoacid counter-anion, image blur occurs primarily due to the local proton mobility, not diffusion of the acid-counter-anion pair.[32] Neutron reflectivity methods were developed to characterize the shape of the reaction-diffusion front with nanometer resolution.[14,30,31,33–37] In these studies, the influence of the evolving chemical composition on the spatial-extent of the reaction profile was directly measured.

Base quencher additives are also used in photoresist formulations to limit the reaction and diffusion of the photoacid catalyst into unexposed regions.[12,38] The influence of these neutralizing species on the reaction-diffusion process is complex.[39] The simplistic view of the quenchers solely acting to neutralize photoacid, thus decreasing the acid concentration, is not always correct.[40] The quencher appears to partially neutralize the photoacid less than stoichiometrically, influence the dissolution either as promoters or inhibitors, and increase the development induction time.[39] The partial neutralization initially reduces the amount of acid available after photolysis proportional to the base concentration.[40–42] The influence of quenchers on the chemical gradients in the film were suspected to control LER.[43] Direct measurements of the effect of amine quenchers on the reaction-diffusion profile shape (LILS) and spatial extent (δ) were determined by neutron reflectivity.[35]

A central assumption in these resolution limit models is a direct transfer of the chemical deprotection reaction-diffusion heterogeneity on the feature quality; the details of the development process are not considered. However, the bulk of photoresist thin films dissolve via reactive dissolution kinetics involving a well-defined steady-state swollen layer.[44–46] This swollen layer must approach the nominally unreacted and unexposed zone as bulk development ceases.[47] This crucial transition zone results from the initial deprotection latent image that can be controlled by aerial images, polymer chemistry, photoacid generators and base additives, and post-exposure bake conditions.[30,35] The mechanism of how the advancing swelling dissolution front faces the transition of soluble to insoluble species (solubility switch) is crucial for the understanding of resist resolution limits. This was addressed by neutron reflectivity techniques that directly measured the developer penetration and extent of line-edge swelling.[48–50] The residual swelling fraction at the feature edge remains diffuse over length scales (>10 nm) far exceeding the polymer chain dimensions during hydroxide development and water rinse step; the swelling layer eventually collapses upon drying. Alternative development approaches to control this residual swelling fraction may be needed to smooth and reduce LER.

19.2.2 Advances by Material Structure

In a typical polymer photoresist, the PAG is dissolved along with the photoresist polymer and spin cast to form a thin film mixture, or binary blend. Polymers provide flexible platforms to change functional groups in order to meet etch resistance, optical transparency, refractive index requirements, and a variety of acid-sensitive protecting groups. The high glass transition temperatures (typically, T_g > 140°C) provide dimensional stability and wide latitude in post-exposure bake temperature that increases the rates of reaction and photoacid diffusivity. Lastly, polymers have a large degree of lipophilicity in an aqueous base developer that provides a high-development contrast.[12]

Driven by the CD requirements, it was considered that reducing the photoacid diffusion length would enable smaller CD. The PAG and polymer blend approach may not be the most effective route, since the photoacid could diffuse to lengths longer than the CD. In order to address this viewpoint, an alternative resist structure was devised that covalently bonds the photoacid generator to the polymer as shown in Figure 19.8a.[51–54] With this approach, after exposure, the photoacid counter-anion remains covalently bound to the polymer thereby restricting the acidic

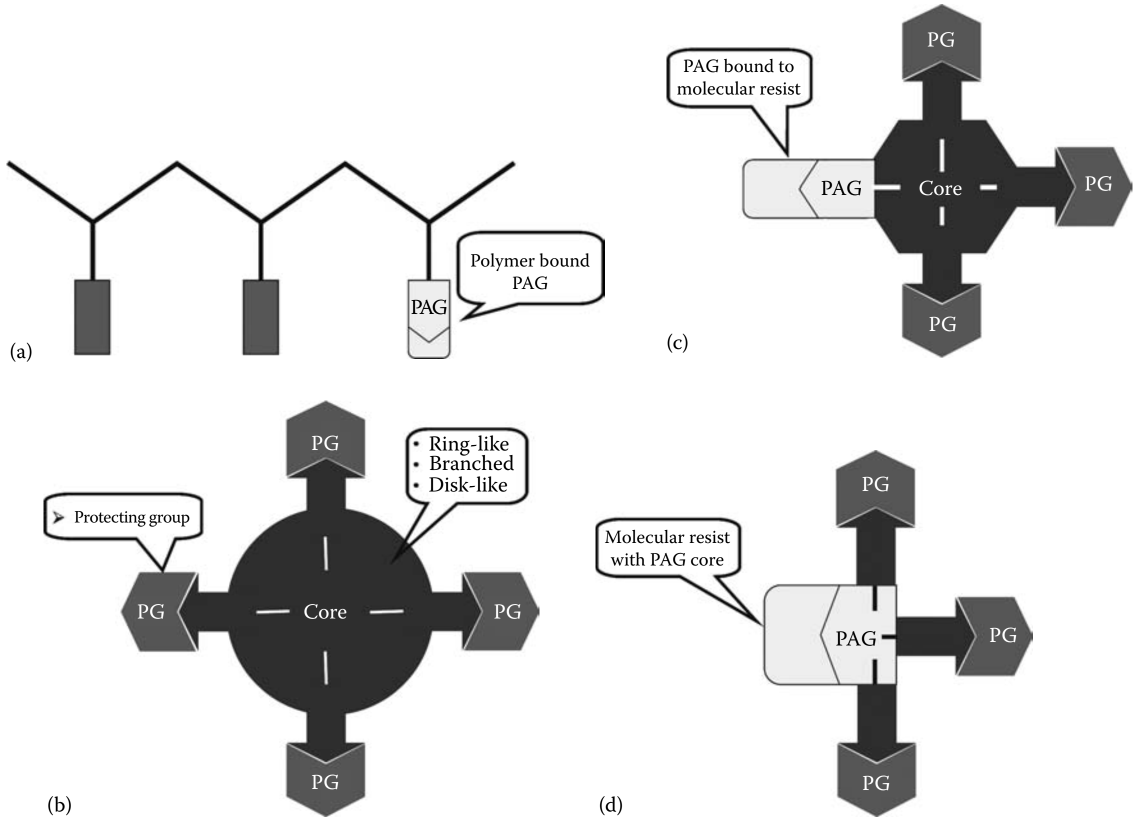

FIGURE 19.8 Cartoons of photoresist architecture alternative to polymer and PAG blends: (a) Photoresist polymer bound to the PAG, (b) molecular glass resist with variable core structure, (c) molecular glass resist bound to the PAG, and (d) PAG-core molecular resist.

proton diffusion length. Such approaches remain promising especially to increase the dose sensitivity.

Since the CD and LER metrics are approaching the characteristic dimensions of the photoresist polymers, alternative architectures were considered to extend photolithography by using lower molar mass molecules.[55] These molecular glass (MG) resists, while smaller, may also improve the uniformity of blends with PAG and other additives since miscibility of polymer blends decreases with increasing molar mass.[56] In general, the molecular glass resist has a well-defined small-molecule core that bears protected base-soluble groups (such as hydroxyls and carboxyls) as shown in Figure 19.8b. With this approach, the core chemistry can vary from calix[4]resorcinarenes (ring-like),[57–59] branched phenolic groups,[60,61] and hexaphenolic groups (disk-like).[62,63] Early approaches with MG led to low glass transition temperatures, however, such problems were resolved by increased hydrogen bonding functionality and the design of the core structure. The MG resists may also benefit from a more uniform development due to the lack of chain entanglements and reduced swelling, when compared with polymers; these are active areas of research. Experimental data using the quartz crystal microbalance method demonstrate that swelling appears during development even with molecular glass resists.[64] Most of these alternative resist structures adhere to the chemical amplification strategy. However, nonchemically amplified photoresists are also being considered as they do not contain photoacid generators and hence do not suffer from photoacid diffusion length constraints.[65]

Two other novel MG variants are a PAG covalently bound to the molecular resist (Figure 19.8c) and the core of the molecule serving as the PAG[66] (Figure 19.8d). As designed, there would be no need for the blending of PAG with such resist systems. In the case of Figure 19.8d, photolysis produces a photoacid, which then would deprotect the unexposed acid-sensitive protecting groups of the PAG-core molecular glass. These two approaches (Figure 19.8c and d) are smaller pixel sizes and true one-component systems that in principle eliminate surface segregation and phase separation in cast films. These alternate resist structures, however, typically require additives such as amine base quenchers to limit the diffusion of the photoacid catalyst into unexposed regions.

19.2.3 Progress in Resists for EUV

There has been progress in designing photoresists for EUV lithography in anticipation of the 22 nm nodes. The testing and development of new materials typically relies on direct lithographic testing to screen formulations. However, due to the lack of widely available EUV exposure tools, micro-field exposure tools (MET), such as the 0.3 NA SEMATECH Berkeley MET have an important role for resist testing. Substantial progress[67] in reaching CD challenges with commercial chemically amplified photoresist were reported with 20 nm half-pitch with an EUV dose sensitivity of (12.7–15.2) mJ/cm^2, which is near the theoretical Rayleigh resolution limit of the 0.3-NA system with a k_1 of 0.45. While progress in resolution and sensitivity were observed, LER remains a challenge. In fact, the fidelity of the resist feature may have non-negligible LER contributions from the EUV mask and optical trains as ascertained by modeling. After subtracting an estimated mask contribution, resist LER values appear to approach the 2 nm level. Additionally, a pattern collapse of sub-20 nm dense features suggests the intrinsic resolution could be better than expected. In fact, unoptimized model EUV polymers and MG photoresists clearly show 20 nm features as quantified by the latent image and developed image roughness.[68] Alternative development approaches[69] may provide new directions to break the resolution limits (Equation 19.2), which currently do not directly consider development effects, such as swelling and swelling layer collapse.

19.2.4 Progress in Resists for 193 nm Immersion Lithography

By inspection, the Rayleigh equation (Equation 19.1) directs that the smallest feature may be achieved by employing higher refractive index media (photoresist and immersion fluids). Currently, highly purified water with $n = 1.44$ is used as the immersion fluid and typical 193 nm polymers have $n \approx 1.7$. The development of photoresists and immersion fluids with higher n are needed. A strategy to increase the refractive index is by incorporating more polarizable heteroatoms, such as sulfur,[70–73] into the polymer structure. Halides such as Cl, Br, and I would increase the refractive index at the expense of absorption and possible photo-induced side reactions. Developer-soluble top-coat barriers[74–77] or engineered surface-segregating barrier-layer additives[78–83] are typically used to reduce or mitigate the leaching of critical components (PAG and quenchers) into the immersion fluid.

19.2.5 Concluding Remarks

Sub-32 nm critical dimensions come with many resist challenges. The challenges of reducing feature size and decreasing LER and sensitivity are inter-related. Current attempts to surpass resolution challenges include novel resist process strategies such as double patterning and double exposure as well as novel architectures. However, it is clear that the quality of the final patterned structures is inter-dependent upon the spatial distribution of the photoacid (optical image quality), the spatial extent of the reaction-diffusion process (latent image), and the development mechanisms. The guidance of the resist resolution metrics in cooperation with improved materials fundamentals will help photoresist materials achieve ever smaller features as suggested by the Rayleigh equation.

19.3 Deep Ultraviolet Lithography

Deep ultraviolet (DUV) lithography, using KrF or ArF excimer lasers with illumination wavelengths of 248 or 193 nm, respectively, is now the predominant technology for producing critical

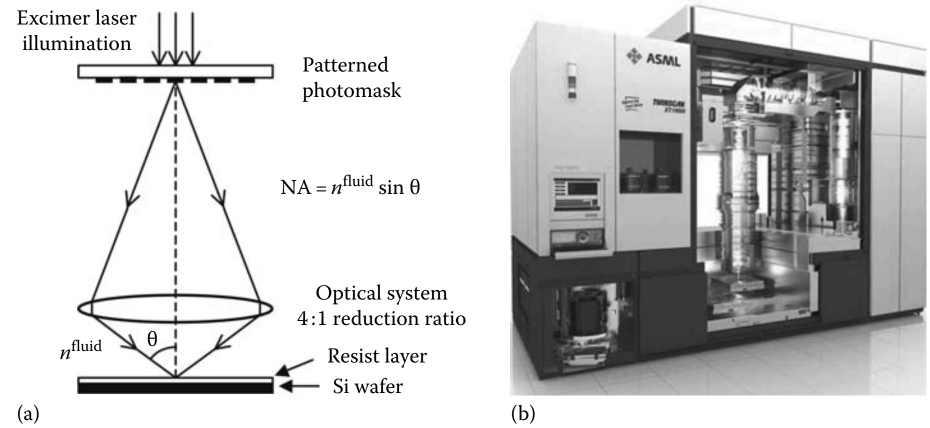
Excimer laser illumination
Patterned photomask
NA = n^{fluid} sin θ
Optical system 4:1 reduction ratio
n^{fluid}
θ
Resist layer
Si wafer
(a)
ASML
TWINSCAN
(b)

FIGURE 19.9 (a) Schematic of the DUV photolithography exposure process. The numerical aperture is defined by NA = n^{fluid} sin θ, where θ is the half angle of the marginal rays and n^{fluid} is the index of the fluid (gas or liquid) between the final lens element and the resist. (b) Illustration of a commercial 193 nm immersion lithography system (ASML TWINSCAN XT:1950i), showing the cylindrical optics barrel in the center, the mask above the optics system, and an illuminated wafer below. (Courtesy of ASML, Veldhoven, the Netherlands. This illustration is based on an artist impression. No rights can be derived from it.)

layers of leading-edge, mass-market semiconductor logic and memory ICs. The technology is based on illuminating a patterned fused silica mask with diffused excimer laser radiation and imaging the transmitted light patterns with 4:1 image reduction onto a photosensitive-resist-coated Si wafer, using a diffraction-limited optical system. The process is indicated schematically in Figure 19.9a. The illuminated wafer is removed from the optical system for chemical processing steps to convert the exposed pattern (latent image) in the resist to a patterned structure on the wafer and put back into the optical system multiple times for further exposure/processing steps. State-of-the-art ICs are built up from about 20–30 layers, a significant fraction of which involve DUV exposure steps. An illustration of a leading-edge 193 nm immersion lithography exposure tool for volume production is shown in Figure 19.9b.

DUV lithography is an evolution of optical-projection lithography developed in the late 1960s and the 1970s using the strong spectral lines from mercury discharge lamps to illuminate first at 436 nm (g-line) and later at 365 nm (i-line). In these early systems, the pattern on the mask for the entire circuit was illuminated and imaged onto the wafer for a specified exposure time, and then the wafer was "stepped" to an adjacent unexposed position for another imaging of the circuit pattern. This was repeated until the wafer was filled with as many identical exposures as the wafer size allowed. These wafer "step-and-repeat" systems were termed "steppers."

In modern leading-edge DUV lithography systems, the full circuit illumination of the mask is replaced by a "scanned" illumination. A fraction of the circuit pattern on the mask is illuminated in a narrow strip across its width and the mask is scanned under this strip, while its image is projected with a 4:1 reduction ratio onto a wafer scanning in the opposite direction, at a speed relative to the mask reduced by the factor of 4. When the full circuit pattern has been exposed, the wafer is stepped to the next position and the process repeated. The exposure image is stepped-and-scanned across the wafer, producing typically over 50 full chip exposures on a 300 mm diameter wafer at a rate giving about 100 wafers per hour. These wafer "step-and-scan" systems are referred to as "scanners," though they are often referred to as "steppers" as well.

The first commercially available production optical-projection lithography system, the DSW4800, was introduced by GCA in 1978. It achieved a minimum feature resolution of a little over 1 μm and a slightly larger depth of focus.[10] At the time, it was fully understood that the resolution of the optical-projection lithography approach was fundamentally restricted by the diffraction limit given roughly by the Rayleigh resolution criterion in Equation 19.1. In DUV lithography, the process-dependent factor k_1 is of order 1 Ref. [10,84]. Actually, Equation 19.1 is intended to capture both the limiting effects of diffraction and the impact of the resist processing. Considering just the diffraction effects, Equation 19.1 would refer to the aerial image at the image plane, and k_1 would be determined only by the profile of the light intensity. However, the actual size of a feature ultimately created by the lithography process also depends on the chemical processing of the exposed resist as described in Section 19.2. For an isolated feature, this size can be any value, i.e., k_1 has no fundamental restrictions. On the other hand, for the dense structures of real circuits, e.g., modeled by a periodic structure, the separation between printed features (the pitch) is fundamentally limited. With the minimum feature size in Equation 19.1 defined as half the pitch for a single exposure, k_1 can be rigorously shown to have a minimum value of $k_1 = 0.25$ Ref. [10,84]. This limit holds for any single-exposure process, as long as the process has a linear dose response.

Because a functional circuit must have some topology and because imaging control is imperfect, the resist exposure process requires some finite depth of focus (DOF). This also has a diffraction and geometric limitation, which is characterized by Equation 19.3

$$\text{Depth of focus} = \frac{\lambda}{2n\left(1-\sqrt{1-\dfrac{\mathrm{NA}^2}{n^2}}\right)} \xrightarrow{\text{Small NA}} k_2\frac{n\lambda}{\mathrm{NA}^2}$$

$$= k_2\frac{\lambda}{n\sin^2\theta}, \tag{19.3}$$

where, for small NAs, a process-dependent prefactor k_2 is conventionally used.[10,84] For large NAs, the exact form of the equation conventionally incorporates a different process-dependent prefactor $k_3 \times 4$.[85]

In the early 1980s, reasonable considerations about the prospect of significantly altering the parameters in Equations 19.1 and 19.3 to improve the resolution and DOF led to the conclusion that optical-projection lithography had an ultimate dense feature size resolution of about 0.5 μm. Consequently, it was assumed then that progress in high-volume-production ICs would require switching within a decade or so to alternative lithography techniques with more extendibility potential. These included electron-beam direct-write lithography (with multiple parallel beams), electron-projection lithography, ion-projection lithography, x-ray-proximity lithography, extreme-ultraviolet lithography, and nanoimprint lithography.[10] Few people, if anyone, publicly predicted then that the factors in Equations 19.1 and 19.3 would be relentlessly driven to enable optical-projection lithography to out-complete all the alternative technologies mentioned above, in cost and performance capabilities for volume production, at least into the second decade of this century.

19.3.1 DUV Lithography Steppers

A primary driver for this progress has been the wavelength factor in Equation 19.1. This can be seen in Figure 19.10, which shows a log-linear plot of critical feature sizes of leading-edge ICs as a function of year introduced into large-scale production, along with the illumination wavelength used. Wavelength reduction has been aggressively pursued in part because, as is clear from Equations 19.1 and 19.3, feature size reduction by decreasing wavelength reduces DOF less than that by increasing the NA. However, the switch from Hg-arc lamp illuminators with wavelengths at 436 and 365 nm to the DUV excimer laser sources with wavelengths at 248 nm (KrF) around 1995 and 193 nm (ArF) around 2000 was a very challenging one. Hg-arc lamp sources are compact, relatively inexpensive, reliable, and operate continuously. KrF and ArF excimer lasers are much more complex, are substantially more expensive to purchase and operate reliably, have undesirable laser coherence properties, and the illumination is pulsed. This last characteristic has been particularly troublesome because obtaining the required exposure doses fast enough for acceptable wafer throughputs requires very high pulse peak intensities. This puts very stringent requirements on the durability of lens and window materials in the optical system. Further, the shorter UV wavelengths substantially limit material options for these optical elements. In particular, at the time of the introduction of 248 nm lithography, UV-fused silica glass was the only UV-transmitting material that could meet the tight optical properties specifications for the large lenses required. For 193 nm systems, in addition to UV-fused silica, a second high-quality material, crystalline calcium fluoride (CaF_2), had to be developed for lenses to correct for chromatic aberrations in the optical system due to the wavelength dispersion of the index of fused silica (see Figure 19.10).

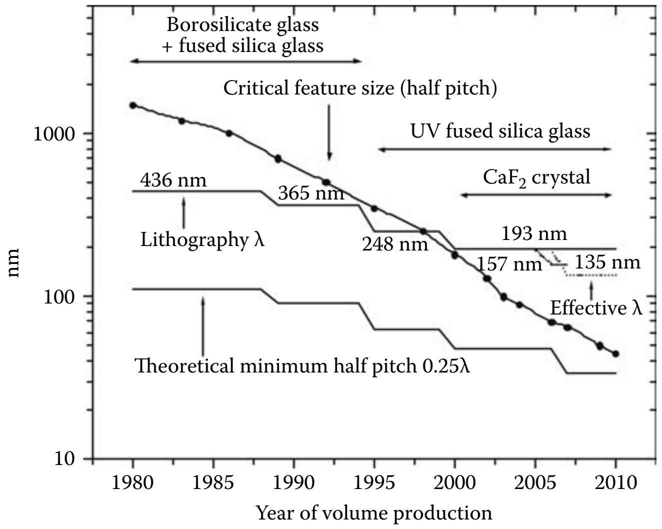

FIGURE 19.10 Plot of critical feature size (half pitch) of leading-edge semiconductor integrated circuits versus year of volume production, on a log-linear plot. Also included is the lithography wavelength used and the theoretical minimum half pitch for single exposures (0.25λ) for this wavelength. 157 nm technology has not been brought into production. Effective λ is λ/n for 193 nm immersion lithography. Optical materials used for lenses at each λ are also indicated.

Fused silica glass is an amorphous form of SiO_2, which minimizes the large birefringence effects of the uniaxial crystal structure of crystalline SiO_2 (quartz). However, the amorphous structure makes fused silica a thermodynamically metastable material and it suffers structural changes under exposure to high 193 nm laser intensities. These changes result in both volume compaction and rarefaction effects with different intensity dependencies.[86] The resulting refractive index and lens geometry changes can substantially degrade lens performance. To minimize these effects, maximum intensities have to be limited throughout the stepper optics, and the more durable CaF_2 lens elements generally must be used in the positions with highest intensities, for example, at the final lens position.

For wavelengths much below 193 nm, UV-fused silica glass has poor transmission, leaving only the cubic structure Group II fluorides as practical lens materials. Of these, only CaF_2 has been made with lithography-grade optical quality. Hence, wavelength extension to 157 nm with F_2 excimer lasers would require the lenses to be made only from CaF_2, or possibly from other related fluorides, such as BaF_2, if the optical quality could be significantly improved.[87] The availability of lithography-quality CaF_2 for lens material has been a key issue for the development of both 193 and 157 nm lithography technologies, and building a CaF_2 manufacturing infrastructure to support the needs of both of these technologies has been a major challenge. A principal difficulty with CaF_2 production is that its relatively low thermal conductivity and high thermal expansion coefficient requires it to be cooled very slowly from the melt in order to obtain low-strain, single-crystal material of the sufficiently large sizes needed for lenses. Even many weeks of controlled cooling in elaborate temperature-controlled furnaces gives a low yield of lens blanks meeting specifications. One result is that 193 nm stepper systems have been designed to use the minimum number of CaF_2 lens elements possible, though CaF_2 has not been designed out entirely.

A further unexpected complication from the use of CaF_2 lens elements in 193 and 157 nm lithography systems resulted from a faulty assumption that the cubic crystalline structure of CaF_2 would ensure isotropic- and polarization-independent optical properties (for high-quality crystals), as could be "demonstrated" by naive symmetry arguments and measurements at longer wavelengths. In fact, CaF_2 turned out to have substantial index anisotropy and birefringence at the short wavelengths of 193 and 157 nm.[88] This is due to the symmetry-breaking effects of the finite photon momentum at these wavelengths, giving rise to a "spatial-dispersion-induced" or "intrinsic" birefringence.[89] Fortunately, the symmetric nature of this effect has enabled it to be minimized by judiciously orienting the crystal axes of several lens elements to substantially cancel the effects.

Beyond 157 nm, a few shorter wavelengths have been considered for further resolution extension: 126 nm from Ar_2 excimer lasers[90] and 121.6 nm from hydrogen Lyman-α discharge sources.[91] At these wavelengths, the only transmissive optical materials known are LiF, which has high extrinsic absorption and poor exposure durability, and MgF_2, which has high natural birefringence due to its noncubic crystal structure. These shorter wavelength technologies have not been developed beyond feasibility studies.[90,91] Below about 100 nm, no practical material is transparent and refractive optics are not possible. For reasons primarily associated with the development of immersion lithography discussed later, 157 nm lithography, though demonstrated to be technically feasible,[87] has been dropped off of technology roadmaps.[14] It now appears likely that the shortest wavelength that will be used for production lithography with refractive optics is 193 nm.

As of 2008, 193 nm steppers are the primary tools for producing critical layers of high-volume, leading-edge ICs. Their ArF excimer laser sources are pulsed (≈6 kHz, 10 mJ/pulse) and line narrowed (≤0.25 pm spectral bandwidth).[92] The transmission photomasks are made from fused silica with Cr absorbing features and have thin-membrane pellicles to keep uncontrolled particles from collecting on the mask and imaging onto the wafer. The stepper optics are made of fused silica spherical and aspheric lenses, calcium fluoride lenses with clocked crystal axes, mirror elements with aspheric surfaces, and immersion fluids between the last lens element and the wafer, as discussed below.

While the wavelength was progressing down to 193 nm, the other two factors in Equation 19.1, k_1 and NA, were also being pushed towards their limits. A number of resolution enhancement techniques have been used to drive the k_1 factor down to achieve IC structures with half pitches corresponding to a value of k_1 approaching the limiting value of $k_1 = 0.25$.[93] Nearly all of these have been taken over from established optical techniques in other fields such as microscopy. These have included (1) illumination methods (off-axis illumination and partial-coherence control), (2) mask modifications to engineer desired wavefronts (phase shift masks, sub-resolution mask structures, and other optical proximity corrections), and (3) resist contrast improvements. For the 45 nm technology node, logic manufacturers are expected to operate with k_1 factors down to about 0.31,[94] and further extension with k_1 factors down to 0.29 is considered feasible.[95] Unfortunately, accompanying these gains, process latitudes have been shrinking to marginally tolerable levels. Clearly, k_1 factor improvements for single exposures are nearly tapped out.

The last factor left in Equation 19.1 is the numerical aperture, $NA = n \sin \theta$. For air between the final lens element and the wafer, the liming value is NA = 1. The GCA 4800 DSW stepper in 1978 had an NA of 0.28.[96] The NA has been increased steadily in newer designs by increasing the size and complexity, and consequently the cost of the lens systems. To contain the number and size of lens elements as the NA was increased, the lens system designs had to incorporate aspheric lenses and off-axis mirror elements (catadioptric designs). Further, as the NAs approached 0.9, the polarization effects could no longer be ignored and illumination polarization control became an essential part of the lithography process, further complicating the lithography tools and processes.[84,97] State-of-the-art dry 193 nm stepper optics have an NA of about 0.93, contain about 30 lenses, with a path length through lens material of about 1 m, and weigh 500 kg or more.[10] With dry NA limit of 1.0, there is very little possible gain left to justify the vastly increased cost and complexity needed for improvement—an exponential increase in cost for an asymptotic gain. The $\sin \theta$ factor is now also essentially tapped out.

Figure 19.10 shows, along with the critical feature size (half pitch), the limiting value of 0.25λ for each wavelength. Clearly, by 2006, the critical feature size 65 nm had approached its limit for 193 or 157 nm. In the 1990s, this projection was the reason why it was nearly universally assumed that alternative technologies, such as EUV lithography, would have to take over below this feature size. This has turned out to be incorrect primarily for two reasons: the NA is not limited to 1 if immersion fluids are included and the effective k_1 factor can be decreased below 0.25 with multiple exposures and processing on the same layer.

19.3.2 193 nm Immersion Lithography

Equation 19.1 is valid at the image plane, which of course must be in the resist. In principle, the $NA = n \sin \theta$ should be evaluated in the resist, which typically has a 193 nm index in the range

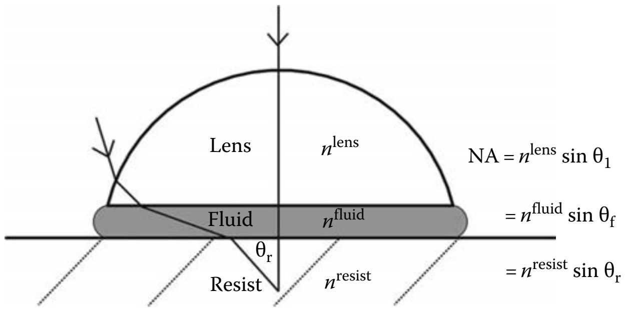
Lens
n^{lens}
NA = n^{lens} sin θ_1
Fluid
n^{fluid}
= n^{fluid} sin θ_f
θ_r
Resist
n^{resist}
= n^{resist} sin θ_r

FIGURE 19.11 Schematic of an on-axis ray and a marginal ray through the last elements of an immersion lithography optical system. The angles θ_l, θ_f, and θ_r are the angles of the marginal ray with respect to the surface normal in the lens, fluid, and resist, respectively. For the case shown here, the refractive index of the fluid is less than that of the lens and resist, and n^{fluid} is then the maximum possible NA of the system.

$n = 1.6$–1.7. However, by Snell's law ($n_1 \sin \theta_1 = n_2 \sin \theta_2$), NA = $n \sin \theta$ can equally well be evaluated above the resist. Until the recent introduction of the immersion stepper, the space between the final lens element and the wafer was filled with air or N_2 gas, with an index near $n = 1.0$. A consequence is that in spite of the much higher index of the resist, the NA has a maximum value of 1.0. However, as microscopists have known and exploited for centuries, adding a fluid between the image plane and the final lens element allows the maximum NA to be increased.[98–100] For a planer final lens element, the NA can theoretically be increased up to the lowest index of the resist, the fluid, or the lens, as indicated in Equation 19.4 and Figure 19.11:

$$\text{NA}_{\text{max}} \leq (n^{\text{resist}}, n^{\text{fluid}}, n^{\text{lens}}) \tag{19.4}$$

A nearly ideal immersion fluid for 193 nm lithography turned out to be purified water. It has sufficient transparency at this wavelength, is relatively innocuous to the resist and lens materials, is compatible with resist processing, has low enough viscosity to enable rapid wafer scanning, and is inexpensive. These properties enabled remarkably rapid development and implementation of immersion technology.[101] From the inception of substantial 193 nm immersion lithography efforts in late 2002,[102–105] it took only about 4 years for production-worthy water immersion systems to be built and the process brought into production. Immersion imaging does introduce some new issues, however, including bubble formation in the fluid, evaporation residue defects, resist-immersion fluid interactions, and fluid thermal effects. Polarization effects, already issues for dry systems, have been exacerbated at the extreme NAs of immersion lithography, requiring careful polarization control for differently oriented structures, which puts some restrictions on IC design.[106]

With water as the immersion fluid, having a 193 nm index of $n^{\text{water}} = 1.437$, the theoretical minimum half pitch (HP^{min}) for a 193 nm exposure tool decreases to $\text{HP}^{\text{min}} = 0.25 \times 193.4\,\text{nm}/1.437 = 33.6\,\text{nm}$. Furthermore, at HPs achievable for dry systems, Equations 19.1 and 19.3 show that for the same HP, the DOF is larger for the immersion approach. Note that the minimum HP for 157 nm dry systems is $\text{HP}^{\text{min}} = 0.25 \times 157.6\,\text{nm}/1.0 = 39.4\,\text{nm}$, higher than that for 193 nm immersion systems. Attempts were made to identify high-index 157 nm immersion fluids, but only relatively low-index (n [157 nm] $\approx$ 1.35) fluorocarbon liquids were found to have any practical transparency at 157 nm,[107] and consequently 157 nm lithography technology was dropped from technology roadmaps. As of 2008, water-based 193 nm immersion lithography systems are being operated at an NA in the range of 1.30–1.35 to satisfy the 45 nm technology node with acceptable process latitudes.[14]

Approaches to extend 193 nm immersion lithography technology further with higher-index fluids have been pursued. Practical 193 nm transmitting organic fluids with 193 nm indices near $n = 1.65$ nm (second generation fluids) have been developed.[108,109] However, the last stepper lens element is made of calcium fluoride with a 193 nm index of $n = 1.50$, or fused silica, with an index of $n = 1.57$. By Equation 19.4, these materials are the bottlenecks for resolution extension. Thus, significant NA gain from higher-index fluids requires higher-index last lens materials as well. Higher-index UV lens materials have been explored for this purpose. Key practical materials identified include $\text{Lu}_3\text{A}_5\text{O}_{12}$ (LuAG) [$n = 2.21$] and polycrystalline MgAl_2O_4 (ceramic spinel) [$n = 1.93$].[110] LuAG has the highest index, but efforts have not yet succeeded in improving the 193 nm transmission to the specifications, which are stringent due to its high thermo-optic coefficient, dn/dT. Also, the large value of the intrinsic birefringence for this material is difficult to compensate for. Ceramic spinel solves this problem with its polycrystalline structure, but the polycrystalline nature also results in a high degree of scattering. A further possible high-index lens material being considered is $\alpha\text{-Al}_2\text{O}_3$ (crystalline sapphire) [$n = 1.92$]. Its large natural birefringence due to its uniaxial crystal structure has limited its use in precision optics. However, with its crystal optic axis oriented along the optical axis of the system and with the polarization of all rays arranged to be oriented perpendicular to this axis, a sapphire last lens element could be manageable. As of this writing, LuAG is considered the most promising high-index lens material candidate,[111] and the industry is still considering whether the practical NA gain to about 1.50 is worth the development effort.

If a third-generation immersion fluid with an index near or above $n = 1.80$ could be developed, then an NA increase to near NA $\approx$ 1.7 could provide enough incentive to justify the substantial development effort required to implement it, because this would enable the 32 nm half-pitch technology node. Pure organic fluids with indices this high cannot have practical 193 nm transmissions. However, the indices of the fluids could in principle be increased to this level by loading the liquids with approximately (2–10) nanometer-size suspended particles of high-index oxide crystals, e.g., HfO_2 or LuAG, while maintaining acceptable transmission, scattering levels, and viscosities.[112–114] This approach is being explored, but it is regarded at best as complementary to multiple patterning approaches to resolution extension discussed next.

19.3.3 Double Patterning

The process-dependent k_1 factor has been driven down to near the hard limit of $k_1 = 0.25$ for single exposures. However, for multiple exposures and processing steps on the same layer, structures with half pitches corresponding to lower effective k_1 values are possible. A number of these multiple patterning approaches are now being pursued and these are expected to satisfy the requirements for the 32 nm half-pitch technology node and possibly the 22 nm node and below. These approaches differ by the number and sequence of exposure and processing steps used to achieve the desired features. Four basic types are illustrated schematically in Figure 19.12.[14] They all use the fact that while the structure pitch created by a single exposure is limited by diffraction, as characterized by Equation 19.1, processing can be used to tailor the line/space ratios in each period, and subsequent exposure and/or processing steps can interleave further structures to increase the structure density. In the Litho-Etch-Litho-Etch process, Figure 19.12a, two Litho-Etch processes are done in sequence with the second process shifted by half the period to give a pitch doubling. This general process can be used to create arbitrary structures with the half pitch shrunk by a factor of 2 below the diffraction limit, though this scale shrink requires tighter tolerances on line edge control and overlay, among other issues.

A serious difficulty with this process is that it requires that the wafer be taken out of the stepper track for etching and then realigned before the second exposure, creating overlay challenges and decreasing the throughput for the layer by a factor of about 2. A preferred approach would have the two exposures done in sequence with just one etch step at the end: Litho-Litho-Etch, shown in Figure 19.12b. Unfortunately, it can be shown that for a resist system with a linear dose response, as has been universal in DUV lithography, two exposures cannot generate patterns with pitches below the diffraction limit. However, with a nonlinear dose response system, this is possible. A number of nonlinear response approaches are being pursued, including contrast enhancement layers, 2-photon resists, and positive-and negative-tone threshold response resists.[115] One version, Litho-Freeze-Litho-Etch, shown in Figure 19.12c, is, particularly promising.[116] In this approach, the latent image captured in the resist from the first exposure is "frozen" by chemical treatment. The first frozen image region is protected from any photo response to a second exposure, then a single etch step can create the pitch-doubled structures.

A different approach, known as side wall (or spacer assist) double patterning, requires only one exposure step to create pitch doubling.[117] As illustrated in Figure 19.12d, the first exposure and development step is only to create sidewalls at the desired positions. A subsequent thin-film deposition on the sidewalls, chemical-mechanical polishing (CMP) to split the deposited structure,

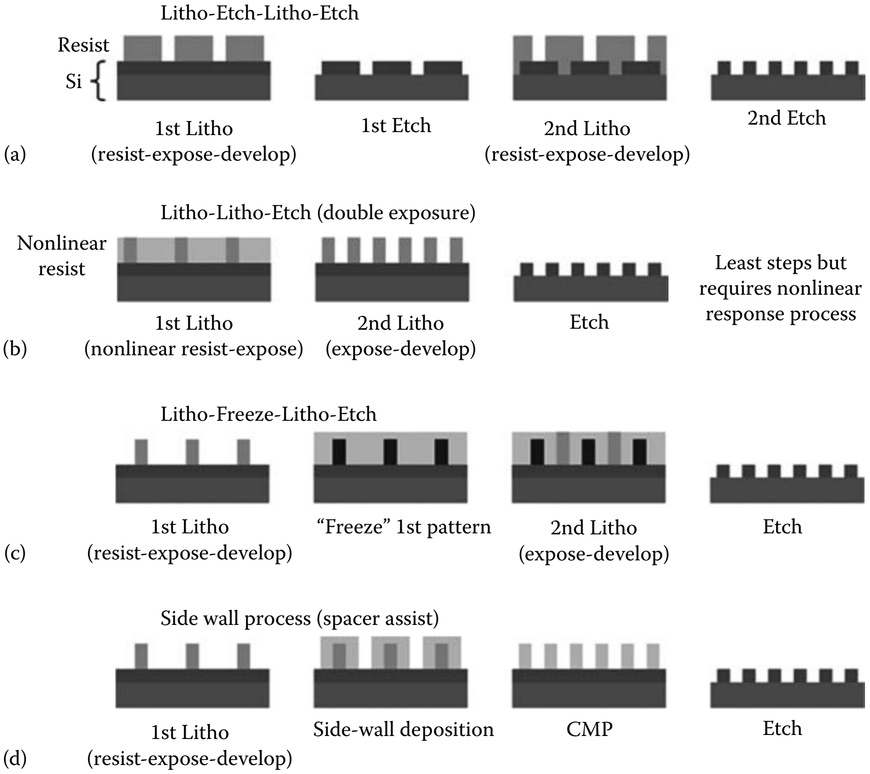
Litho-Etch-Litho-Etch
Resist
Si
1st Litho (resist-expose-develop)
1st Etch
2nd Litho (resist-expose-develop)
2nd Etch
(a)
Litho-Litho-Etch (double exposure)
Nonlinear resist
1st Litho (nonlinear resist-expose)
2nd Litho (expose-develop)
Etch
Least steps but requires nonlinear response process
(b)
Litho-Freeze-Litho-Etch
1st Litho (resist-expose-develop)
"Freeze" 1st pattern
2nd Litho (expose-develop)
Etch
(c)
Side wall process (spacer assist)
1st Litho (resist-expose-develop)
Side-wall deposition
CMP
Etch
(d)

FIGURE 19.12 Double patterning approaches: (a) Litho-etch-litho-etch requires 2 exposures and 2 etch steps. (b) Litho-litho-etch (or double exposure) requires 2 exposures and only one etch step, but needs a resist process with a nonlinear dose response. (c) Litho-freeze-litho-etch has intermediate complexity, requiring 2 exposures, a chemical freezing process after the first exposure, and one etch step. (d) Side wall process (spacer assist) requires one exposure and one etch. It uses side wall deposition and chemical-mechanical polishing (CMP) to achieve doubled pitch.

and elimination of the original pattern enables the desired pitch doubled structures. This approach takes advantage of the well-understood and highly controllable thin-film-deposition technology to obtain the desired linewidths, independent of diffraction limits. Furthermore, the sidewall-derived structures are automatically self-aligned by the original structure. A difficulty with this approach is that restriction to patterning along sidewalls makes design layout much more difficult, and not all structures are topologically possible with single exposures. At least a further exposure/processing step is generally required for arbitrary structures. Still, this method is inherently extendable, and pitch quadrupling, etc. is possible in principle.

19.3.4 Ultimate Resolution Limits of DUV Lithography

Double patterning methods are already being used in IC production, and they are the declared technology solutions, using 193 nm water immersion tools, for several volume semiconductor manufacturers for the 32 nm half-pitch technology node projected for about 2013.[14] These approaches appear feasible for further resolution shrinking as well, e.g., the 22 nm half-pitch technology node. Redesign of 193 nm immersion steppers with high-index fluids and lens materials, in combination with double patterning techniques, offer the potential for further resolution extension and relaxation of the k_1 factor for a given resolution. When double patterning becomes routine, consideration of triple patterning, quadruple patterning, etc. may become tempting. However, this would likely have to come at the cost of much tighter process control requirements, such as CD and overlay control, than is presently attainable. Improvements to meet the requirements, along with the increased complexity (more exposures and processing steps per layer) will surely drive up costs substantially. There is little doubt that these DUV lithography extension approaches can be made to work technically. It is just a matter of whether they can operate at commercially viable costs and whether any reliable alternatives can be implemented at less cost. For example, rapidly increasing mask costs have made the numerous maskless technologies, such as multi electron–beam direct write schemes, very attractive, at least for low-volume production.

For high-volume leading-edge ICs, it is generally agreed that for the 22 nm half-pitch technology node and below, a reliable single-exposure EUV solution, with its larger depth of focus and more natural extendibility, would be preferred. However, formidable challenges, repeated introduction delays, and especially rapidly rising projected costs, have made this solution not so inevitable as it once seemed at the beginning of the decade. Cost is the ultimate driver. If or when DUV lithography is displaced, and what its ultimate practical resolution limits are, is now less clear than ever. The technology has thrived on remarkable innovativeness and resourcefulness and its development history does not suggest that these will cease. Previous predictions of its limits and of its imminent replacement have always been wrong.

19.4 Electron-Beam Lithography

Electron-beam lithography has demonstrated its utility over several decades as both a primary pattern generation tool for the semiconductor industry and as the premiere means of patterning small structures for advanced device development and research in myriad fields. Like any lithographic technology, its suitability for a given application depends on its performance with respect to the four metrics of resolution, throughput, pattern placement, and cost-of-ownership. As we shall see in the following sections, electron-beam systems excel with respect to resolution and pattern placement, but suffer from fundamental limits in terms of throughput.

19.4.1 Resolution

Resolution in a lithographic process is a rather ill-defined quantity since, by virtue of the fact that it is a process involving many steps, a large number of variables are involved. These can frequently be manipulated to produce isolated features far smaller than might, at first sight, be judged possible. The true test of the process is the minimum pitch of the features that can be fabricated. In what follows, we will consider primarily those factors that are unique to electron-beam exposure and neglect those that are common across the various exposure techniques—those are covered elsewhere in this chapter.

The ultimate resolution that can be obtained using electron-beam exposure is determined by the nature of the interactions of the electrons with the material being exposed. There are two parts to be considered: first, the trajectories that the incident, or primary, electrons follow within the material and second, the processes by which the energy deposited by the primary electrons is translated into developable chemistry in the exposed material.

The electron trajectories are determined by the combination of incident electron energy and resist/substrate atomic number.[118,119] At low energies/high atomic numbers, the electrons scatter strongly within the solid, which leads to substantial broadening of the beam as it penetrates the material. Since most resist materials are organic, their average atomic numbers are not dissimilar from that of carbon, so the beam broadening, or forward scattering in the resist, is determined principally by the electron energy. For this reason, most high-resolution systems operate at energies of 50–100 keV. In addition, because the beam broadening increases progressively as the electrons undergo additional elastic scattering events on their way through the resist, thin resists yield higher resolution images. Very low (<500 eV) incident energies have also been proposed as a means of generating high-resolution images.[120–122] In this case, the beam broadening is on the order of and constrained by the very short total range of the electrons within the resist. The resist thickness is therefore limited to that range, which may be insufficient for subsequent processing needs, and the resolution can only be improved by reducing the energy (and the resist thickness) to impractically small values.

The incident electron energy, E_0, has a dramatic impact on the dose needed to expose resist, which, for energies above about 5 keV, varies linearly with E_0. This is a reflection of how the inelastic scattering cross section varies with electron energy. The effect is captured by the Bethe continuous slowing-down approximation[123] that describes the average rate at which fast electrons lose energy along their trajectories. At high energies and resist thicknesses that are small relative to the electron range, the feature width is determined not by the deviations in the trajectories of the primary electrons due to elastic scattering, but by the volume in which they deposit energy. Although the primary electrons can cause chemical changes in the resist directly, it is normally assumed that the so-called fast secondary electrons generated by inelastic scattering of the primary beam are responsible for the majority of broken bonds in the resist.[124] This is because the low-energy secondary electrons have a much larger interaction cross section than those of the primary beam.[118] Conceptually, each primary electron trajectory can be thought of as surrounded by a cylinder of material 10–20 nm in diameter exposed by the secondaries that it generates. At very low incident electron energies, the stopping power increases significantly and the electrons deposit all their energy efficiently in the resist film, rather than carrying on through to the substrate.[125]

The resolution that can be achieved in a resist can also be estimated by considering the impact parameter for energy-loss events for the primary electrons. An electron traveling with velocity v, interacting with a stationary electron at a distance b delivers an impulse having a duration of approximately b/v and contains frequency components, ω, up to v/b. These correspond to a maximum energy transfer, ΔE, of $h\omega/2\pi$.[126,127] For 100 kV electrons and ΔEs of 5, 50, and 500 eV, b is 10, 1, and 0.1 nm, respectively. Normally, the energy scale relevant to resist exposure is on the order of a few electron volts (3.6 eV for the C–C bond), leading to an interaction distance of a few nanometers but in a few cases the energy needed is much higher, corresponding to a core-shell excitation, for example. In systems such as alkali metal halides, in which the exposure mechanism depends on the dissociation of the halide ions from the lattice, resolutions of 1–2 nm are achievable.[128–130] Unfortunately, the doses required are in the C/cm^2 range, which is not practical for manufacturing. Currently, the highest resolution material that is suitable for nanofabrication is hydrogen silsesquioxane (HSQ), which has demonstrated feature sizes as small as 7 nm half-pitch.[131] This is an inorganic, small-molecule resist that essentially forms silica upon exposure and development. Its increased mechanical strength relative to organic resist materials is one reason for its ability to produce small features, while its higher average atomic number may help reduce the range of the fast secondaries, contributing to its high resolution.

Electron–solid interactions also affect the resolution in electron-beam systems through the proximity effect. This is the additional exposure of the resist by electrons backscattered from the substrate and it leads to a loss in contrast and thus to a loss in resolution. At higher voltages and on low-atomic number substrates (e.g., Si or SiO$_2$), the backscattered electrons have a relatively large range ($\propto V^{5/3}$) and the contribution of the forward and backscattered electron doses can be captured by the two-Gaussian model[132]

$$D = D_0\left[e^{x^2/\alpha^2} + \eta e^{x^2/\beta^2}\right] \qquad (19.5)$$

where

D_0 is the incident dose
α is the width of the forward-scattering distribution
β is the width of the backscattered distribution
η is the effective dose contributed by the backscattered electrons

As we have discussed above, α is typically on the order of 10 nm, while β is on the order of micrometers. η can be quite large, depending on the substrate's atomic number and is approximately 0.5 for Si at 100 kV.[133] This model is applicable if the forward and backscattered dose distributions can be readily separated. In this case, small, isolated features receive a dose of D_0, the centers of large features receive a dose of $D_0(1 + \eta)$, and all other types of features receive a pattern-density dependent dose that varies between those values. As long as the feature size is small compared with β, this dose variation can be corrected for either by adjusting the dose delivered to individual features,[134] applying the inverse of the background dose to produce a flat background,[135,136] or adjusting the size of the features[137,138] by using a pattern density averaged over a length scale on the order of β. However, if either α or β approach the feature size, then much more computationally intensive techniques must be used to perform an iterative correction procedure, as the correction applied to each feature affects its neighbor and vice versa.[139] At low energies and in high atomic number substrates such as GaAs or InP, the deposited energy distribution is more complex[140,141] and the proximity effect becomes much more difficult to correct. It is, however, important to point out that these phenomena are analogous to the flare and optical proximity effects encountered in advanced optical lithography, for which correction techniques are at an advanced stage of development.

So far in our discussion, we have implicitly assumed an ideal point source of electrons incident upon the resist. A close approximation can be achieved in practice with low-current, focused probes in electron microscopes. However, for practical applications, we need high beam currents, and this has a dramatic impact on the design and operation of electron-beam lithography tools and leads to systems in which the intrinsic resolution of the resist and process is often better than that of the electron optics.

19.4.2 Throughput

In the design of most optical systems, including electron microscopes, there is a trade-off between blur due to diffraction, which scales as 1/NA, where NA is the numerical aperture of the system, and aberrations that generally scale as some power of NA (a spherical aberration is often a dominant term and varies as NA^3). In electron-beam lithography systems, throughput is critical to their cost of ownership and ultimate utility. Unlike photons in a light-optical system, the electrons in an electron-optical system affect one another strongly through unscreened Coulomb repulsion. The interactions give rise to three principal effects: (1) the forces between electrons along the electron-optical axis lead to the speeding up or slowing down of individual electrons, which is equivalent to an energy spread in the beam, known as the Boersch effect; (2) the average force due to all the other electrons in the beam leads to deterministic changes in electron trajectories and leads, to first order, to an overall defocusing of the beam, known as the global space-charge effect; and (3) the interactions between individual electrons lead to random changes in individual electron trajectories, known as the stochastic space-charge effect. The severity of the Boersch effect depends on the chromatic aberration in the system. The global space-charge effect can be thought of as creating a negative lens in the system whose effects can be compensated for by refocusing the beam. However, if the current density is not uniform across the beam, that lens becomes highly aberrated and its effects cannot be compensated. The stochastic space-charge effect is not correctable. As might be expected, the space-charge effects are made worse when the electrons are forced closer together or can interact for longer periods and increase with beam current, I, and column length, L, and are mitigated when the electrons are spread farther apart or interact for shorter times and decrease with increasing numerical aperture α, the demagnification factor, M, and the accelerating voltage, V. The detailed dependence of beam blur on I, L, V, M, and α is particular to a given system and has been studied in depth for both single-column, probe-forming,[142] and projection systems.[143,144] Note that improvements realized by increasing the voltage are offset by the need to increase the beam current to compensate for the linear decrease in resist sensitivity with voltage.

The net result of these effects is that rather than balancing diffraction and aberration blurs, the system optimization must be performed by balancing space-charge and aberration blurs. This means that for any given beam current in a system, there is an optimum numerical aperture that must be used and a corresponding minimum blur that can be achieved. Any attempt to increase the beam current leads to an increase in beam blur. Electron-beam lithography systems for nanofabrication research are generally designed to provide the optimum resolution, with throughput as an important, but secondary goal. However, the semiconductor industry's continuing demand for ever-smaller features and higher throughputs has resulted in the disappearance of electron-beam lithography from the direct IC fabrication process because of the conflict between the need for increased beam currents to increase throughput and the need for the decreased beam currents necessary for improving resolution cannot be resolved at the feature sizes now required.

In the future, the throughput of electron-beam lithography, even for mask making, is likely to become worse. To understand why this is the case, it is necessary to examine how resist sensitivity scales with feature size.

If we consider a feature of CD L and allow it to vary by no more than $L/10$, then we can define an image pixel as $L/10 \times L/10$. In order for each pixel to be faithfully reproduced, we need the resist to be able to clearly distinguish between exposed and unexposed pixels. Even if the dose over a large area is fixed at a precise value, the arrival of the electrons at any pixel is a statistical process, with the number of electrons delivered following a Poisson distribution. This effect is frequently referred to as shot noise and has been extensively studied.[145–149] We can, therefore, define a minimum number, n, of electrons per pixel that is necessary to ensure that the error rate of improperly exposed pixels is suitably small. The total dose needed per unit area therefore becomes $100n/L^2$, i.e., it scales as $1/L^2$, and the beam current necessary to maintain a given real exposure rate must therefore scale in the same fashion.

19.4.3 Overlay

The fact that magnetic and electric fields can deflect electrons precisely and at high speed is one of the great strengths of electron-beam systems. In optical lithography systems, accurate and precise registration and overlay can only be achieved by ensuring nanometer-scale mechanical alignment between mask and wafer stages—physically massive systems that can only be controlled with bandwidths of a few tens of Hz. In electron-beam tools, it is not necessary to control the stage precisely (offsets of several micrometers between the actual and planned stage coordinates are common), because the position of the electron beam can be adjusted, at speeds of up to 100 kHz, to make up the difference, provided a suitable signal from the stage interferometers is available. There are generally differences in the coordinate system that are generated by the stage and the substrate. Normally, these are accounted for by periodically acquiring alignment mark information from the substrate. This procedure is susceptible to errors arising from drift occurring between mark measurements and schemes have been developed that enable registration data to be generated from the substrate continuously.[150] The ability to adjust the position of the beam, as well as characteristics such as focus and stigmation, in real time, makes it possible to correct for a large variety of errors, such as those caused by variations in substrate height. Along with the benefits it brings, the sensitivity to fields is also a source of difficulty associated with electron-beam systems: interference from dynamic fields must be reduced as far as possible, while that from static ones, such as those due to the small magnetic inhomogeneities of the stage, must be carefully mapped and calibrated out.

Another source of error peculiar to electron-beam tools, particularly at high voltages, is substrate heating. The total amount of energy deposited in the substrate scales as the square of the accelerating voltage because of the combination of the increasing energy per electron and the increasing resist dose required. The concomitant substrate heating can affect both pattern placement/overlay, as a result of thermal expansion,[151] and CD control, as resist sensitivity may vary with temperature.[152] Temperature increases are more severe in SiO_2 substrates because of silica's poor thermal conductivity, while Si substrates, with their higher coefficient of thermal expansion, are liable to greater placement errors.

19.4.4 Cost of Ownership

As we have seen, although electron-beam systems provided exceptional resolution, and are capable of precise pattern placement,[153] they are fundamentally limited in throughput. Historically, for relatively large feature sizes, the limiting factor in throughput was the serial nature of the writing process in probe-forming, or Gaussian-beam, systems and this led to the development of schemes such as shaped-beam,[154] cell-projection,[155] large-area,[6,156] and full-field projection[157] that sought to deliver more than one pixel's worth of information at a time. Unfortunately, as feature sizes decreased and the resolution requirements became more demanding, they all ran into space-charge limits. This has relegated electron-beam systems for production purposes to the fabrication of small numbers of high-value items such as photomasks, CD/DVD masters, and nanoimprint templates. There are two routes out of this dilemma: the first is to accept the limited throughput, but use the essential addressability of electron beams to create flexible systems—maskless lithography tools—that can create ICs without the need for masks and their associated costs;[158,159] the second is to avoid the space-charge limits of single columns by dividing and conquering—using multiple small columns to keep the electrons far enough apart so that their interactions are insignificant.[160,161] In the first case, the cost of ownership is competitive with photolithography for short-production run devices where only a few wafers worth of devices are produced for each mask set. In the second case, providing that issues associated with source uniformity, control, and brightness can be overcome, then it may be possible to have the best of both worlds—flexibility in patterning together with high throughput—making it directly competitive with optical lithography. Research and development are active in both areas.

19.5 Nanoimprint Lithography

Nanoimprint lithography (NIL) is rapidly emerging as a low-cost, high-resolution, and versatile alternative to optical lithography as a patterning technology for semiconductor fabrication and other applications that incorporate nanoelectronic devices.[162,163] Optical lithographic techniques, including state-of-the-art tools based on 193 nm radiation[98] or next-generation potential solutions, including extreme ultraviolet (EUV, 13.5 nm radiation) lithography,[164–166] rely upon selectively exposing a photoactive film to radiation via a photomask that defines the lithographic pattern of interest. A chemical reaction occurs in the exposed regions of the film that renders them soluble in an aqueous base, creating a physical pattern defined by the unexposed regions in the film. It is becoming increasingly difficult to control the resolution of this patterning process.

NIL offers a potentially simplified alternative to chemically amplified resists. All NIL techniques rely on a simple squeeze-flow mechanism whereby a resist flows into a well-defined physical cavity of nanoscale dimensions. It is a molding or replication technique where the ultimate resolution is defined by the physical dimensions of the imprint mold cavities.[167–170] If high-resolution masters can be created, high-resolution copies can be generated. It is also a repetitive stamping or replication technique. The well-defined patterns in the imprint master can be used to create thousands of replicas, analogous to how the metallic master for an LP record were used to stamp thousands of inexpensive vinyl copies for consumer use. Compared with optical lithography for high-volume nanomanufacturing, there is an important distinction worth highlighting. With optical lithography, every printing or copy of the nanostructured surface requires extremely tight control of the environmental conditions to maintain the sensitive balance of the complicated reaction, diffusion, and dissolution process for the chemically amplified resists.[171] This is precisely why state-of-the-art optical lithography tools cost upwards of $30 M apiece, providing a 65 nm patterning resolution, and semiconductor fabrication lines cost billions of dollars to construct.

The resolution of NIL is largely controlled by a simple squeeze-flow process.[167–170] As long as high-quality, high-resolution imprint masters can be fabricated, it is relatively easy to stamp out copies of the mold. In optical lithography, the ultimate patterning resolution is intrinsically coupled with the high volume manufacturing process making process control measures extremely important. Every single copy manufactured requires stringent control over the reaction, diffusion, and dissolution processes. With NIL, the resolution is less coupled with the high-volume manufacturing process and more dependent on the mold or imprint master fabrication. For these reasons, NIL tools tend to be significantly less expensive. Research and development grade imprint tools can be purchased for about $100 K, while manufacturing grade tools cost approximately $2 M, both with a sub 10 nm resolution.

Both optical and nanoimprint lithography are dependent on high-resolution masks or masters. In optical lithography, the mask sets required to fabricate a semiconductor device can cost several millions of dollars to produce, comprising a significant fraction of the manufacturing costs. These are phase-shifting optical masks in which blanket UV radiation is exposed to one side, generating selectively exposed regions in the resist film on the other side. This exposure is also done through a series of lenses where the features in the mask are approximately four times larger than the image at the wafer plane. In NIL, these mask costs are likely to increase. The squeeze-flow is a direct write process where the features in the mold are the same size as the features in the wafer plane; there is no longer this 4x reduction. This means that high-resolution and costly electron-beam lithography is needed to fabricate the imprint masks, compared with optical lithography where the mask resolution requirements are somewhat reduced because of the lens system. There are no lenses in NIL and the imprint masks may very well be more costly to produce. Furthermore, these expensive NIL mask sets must be mechanically pressed into the resist film. In optical lithography, the masks are isolated from the resist behind a transparent and protective pellicle. There are no protective pellicles in NIL and one must make direct mechanical contact of the expensive mask with the resist film.

In this section, we provide a general overview of the NIL technology. Our discussion will be largely limited to the field of nanoelectronics, including complementary metal-oxide semiconductor (CMOS) devices, although NIL is also widely considered to be a manufacturing alternative for several other forms of nanotechnology. We will elaborate on the differences between NIL and existing optical lithography techniques and discuss some of the unique materials and metrology challenges that NIL processes introduce. We will also provide a brief overview of some of the recent developments in applying NIL to nanoelectronics technologies.

19.5.1 Variations of NIL

19.5.1.1 Thermal NIL

NIL generically refers to pressing a hard mold or template decorated with nanoscale patterns into a softer material to lithographically define well-controlled nanostructures.[167–170] In most cases, pattern formation is accomplished by a squeeze-flow mechanism whereby the material flows into the mold in a liquid-like state (Figure 19.13). Of course, this requires "setting" the pattern in the liquid material. In the thermal embossing form of NIL, the resist is typically a glassy polymer.[167,168] During the imprint process, the film and the imprint mold are heated to above the glass transition of the polymer while applying an external pressure. Under elevated pressure and temperature, the polymer flows into the cavities of the mold. The pattern is set by cooling the system to below the glass transition of the polymer, where the material vitrifies into a rigid glass. When the mold is separated from the substrate, the surface is left with nanoscale patterns that are relief images of the mold cavities. A cavity in the mold becomes a polymeric pattern on the surface.

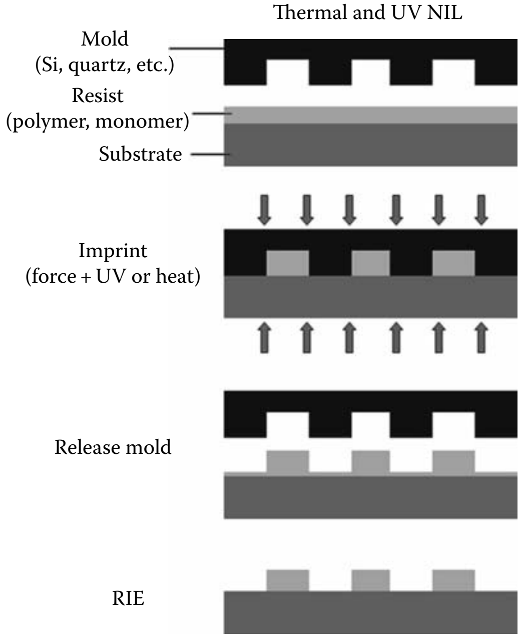

FIGURE 19.13 Schematic of the NIL process, showing both the thermal and UV forms.

The thermal embossing form of NIL is similar in concept to many of the bulk molding or forming processes used to manufacture large scale plastic components. High molecular mass polymers, even at temperatures far above their glass transition, tend to be high viscosity fluids. Flow does not always readily occur under quiescent conditions meaning that pressures must be applied to squeeze the material into the mold. Most NIL tools designed for thermal embossing are capable of applying pressure as high as 5 MPa. This is combined with maximum temperatures of typically 200°C–300°C, which means that most tools are capable of exceeding the rubbery plateau modulus for most polymer melts and able to induce material flow.

There are several different options when it comes to applying the heat and pressure for thermal NIL. One class of tools is largely based on the photomask aligner and wafer bonding systems widely used in the semiconductor industry. In many instances, an imprint module can be purchased as an add-on for an existing tool. In these systems, mechanical platens with heating elements are used to apply the temperature and pressure to the imprint mold and substrate. Their advantages are that they provide uniform heating through the large thermal mass of the platens and they often have a built-in capability for optical overlay alignment between the imprint mold and the substrate. The primary disadvantage is that maintaining parallel contact between the platens can be difficult. Misalignment of just a few μm can prevent conformal contact, which makes it very difficult to imprint nanoscale features over large areas. To overcome the conformal contact issues, some imprint tools utilize hydrostatic gas pressure to squeeze the imprint mold into the resist film.[172] This pressure is applied through what is essentially a deformable gas bladder. Since the NIL molds or masters are usually etched into thin Si or quartz wafers and imprinted onto similar thin Si substrates, the molds and wafers are able to bend or flex to a certain extent and increase the degree of conformal contact. In some tools, the pressure is applied through the top through the gas bladder while the substrate rests on a rigid metallic platen that provides the heating. These systems provide an optimal balance between uniform heating and conformal contact. There are also tools that utilize an entirely soft press technology, where pressure is applied through both the top and bottom through deformable gas bladders. In these cases, the bladders are transparent and the heating of the substrate is provided through infrared heat lamps. This soft press technology maximizes the pressure uniformity across the mold and substrate while the reduced thermal mass (no platens) allows for rapid heating and cooling rates. However, the noncontact optical heating means that temperature uniformity and stability are more difficult to achieve.

19.5.1.2 UV NIL

One disadvantage of thermal NIL is that it requires heating and applying pressure to the substrate. For certain applications, the patterned substrate may or may not be able to withstand the thermal budget associated with this process. Likewise, large temperature changes can induce thermal distortions, which make overlay and alignment difficult. UV NIL offers a low temperature alternative to thermal NIL. In UV NIL, the resist is a liquid-like (or low viscosity) monomer that readily flows into the mold cavities at room temperature with no heating and very little pressure.[169,170] In many instances, the liquid monomer is actually wicked into the mold via capillary forces, minimizing distortions of the substrate or the template (Figure 19.13). In the mold, the monomer undergoes a cross-linking reaction upon exposure to UV radiation, creating a hard glassy pattern that retains its shape after separating the mold from the substrate. This means that either the NIL mold or the supporting substrate must be transparent to the appropriate UV radiation. For this reason, the molds are typically fabricated in quartz wafers, whereas silicon and nickel are more common materials in thermal NIL. An additional requirement is that UV NIL is limited to patterning materials that cross-link under UV radiation. Thermal NIL can, in principle, be applied to any material in which flow can be induced.

19.5.1.3 Transfer Printing

Both the thermal and UV forms of NIL described thus far are similar to optical lithography in that a pattern is created in a resist film that has been previously deposited on the substrate of interest. It is also possible to employ the NIL technologies in a material transfer mode that is similar to most industrial forms of printing. In transfer printing, the material that is to be patterned is first applied to a patterned stamp or mold and then transferred via mechanical contact to the target substrate. This requires tuning of the differential adhesion between the pattern mold and the substrate such that the material adheres to the substrate more strongly than the patterned mold. There are two primary modes in which transfer printing can occur[173] (Figure 19.14). The first is a whole layer transfer where a continuous film with a nanoscale texture defined by the mold topology is transferred to the substrate. The second is a discrete transfer or inking mode where material is selectively transferred to the substrate from the protrusions of the mold; any material in the cavities of the mold is not transferred over. Whole layer transfer is used primarily to create topological patterns of a given material on the target substrate. Examples of this technology are being used to create nano- or microfluidic channels or devices. The discrete transfer mode can also be used to create physical patterns, but it also has the ability to create chemical patterns of a given surface energy or chemical functionality by essentially inking monolayers of a chemical compound. In this last case, the pattern may, to a first approximation, lack topology and not alter the smooth surface of the substrate. Chemical patterns like these are commonly considered in sensor-type applications.

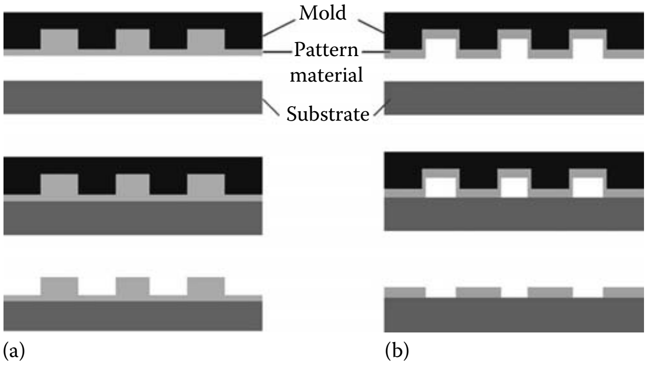

FIGURE 19.14 Schematic of two modes of transfer printing: (a) whole-layer transfer and (b) inking.

19.5.1.4 Functional Materials

One of the biggest distinctions between NIL and competing optical lithography methods is rooted in the generality of the patterning process. In nearly all forms, optical lithography focuses on creating high-resolution patterns in sacrificial photoresist formulations. As mentioned previously, photoresists are highly complicated materials systems that balance chemical reactivity, acid diffusivity, and their dissolution behavior to optimize their patterning resolution. Once the pattern is formed, the resist becomes a sacrificial component for subsequent additive or subtractive processes to transfer the pattern into the functional materials of interest. On the contrary, the mechanical stamping and squeeze-flow processes used in NIL are generic mechanisms applicable to a wide range of materials. This means that a wide range of materials can be patterned directly by NIL, including functional materials. The resolution of NIL processes is controlled by the mold or mask fabrication process more so than optical lithography. The ability to directly stamp nanoscale structures into functional materials is a highly attractive way to simplify the number of parameters impacting patterning resolution, reduce the number of nanofabrication steps, and thereby reduce manufacturing costs.[174]

19.5.1.5 Tool Types

NIL is a low-cost nanopatterning alternative to optical lithography with the potential for high-volume manufacturability. Mechanical stamping processes have a long history in manufacturing environments and the same appears to be bearing true for the emerging field of nanomanufacturing. Stamping processes are intrinsically compatible with manufacturing assembly lines. However, NIL is still in its infancy period and most of the tools that are commercially available today are geared for R&D or the technology development. There are examples of high-throughput tools in the market, but most of the activity is still in the lab as opposed to pre-production. In this section, we review the different types of tools with respect to patterning throughput.

The most versatile tools on the market today also have the lowest throughput and least applicability for high volume nanofabrication. These account for most of the imprint tools sold today from companies like Nanonex, Obducat, or Jenoptik that utilize gas pressure through some type of deformable bladder to apply a very uniform pressure to the mold and the substrate. Typically, these tools are very flexible and provide the user with a wide variety of process control, at the sacrifice of throughput. The user has the freedom to place an arbitrary mold onto an arbitrary substrate, sandwiched between one or two sheets of deformable rubber or foil. This stack is then placed inside a movable gas chamber. The gas chamber closes down on the deformable bladders, creating pressure reservoirs on the top and/or bottom of the sample. A vacuum pumping system is also usually involved to evacuate the sample region between the bladders. Nitrogen or some other inert gas is then pumped into the chamber, generating pressures as high as approximately 5 MPa to press the mold into the substrate. Heat is then applied through either the optical heat lamps or thermal sample block to enable flow of the material into the mold cavity. After a prescribed heating cycle of a few minutes, the temperature is cooled back to room temperature and gas chambers are vented to release the pressure. The sample and the mold are then manually removed from the chamber and manually separated to reveal the patterned substrate. This process obviously describes the thermal form of NIL. The same general process applies for UV NIL with these tools with the exception that lower pressure is required and UV radiation is applied instead of heat. In either case, UV or NIL, these single stamp tools are capable of generating one patterned substrate approximately every 10–15 min with manual input from the user. For most applications, this throughput is not suitable for manufacturing. Rather they are designed for flexibility and ease of use. The sample chambers typically vary from approximately 5–20 cm in diameter and allow arbitrary form factors or piece parts to be imprinted. This is primarily useful for technology development. One notable exception is a large area single imprint tool being developed at the Korean Institute of Machinery and Materials.[175] This is a low pressure UV imprint tool capable of imprinting an entire 45 cm (diagonal) flat panel display uniformly in a single imprint. Cycle times of 4–5 min here would probably be more acceptable given the limited demand for such high-end displays. These cycle times would be totally unacceptable for the high volume demand of bit patterned magnetic media for data storage hard drives.

There are efforts to increase the throughput of imprinting tools. The next class of tools generically utilizes a stamp and repeat technology to generate multiple copies of a single pattern across a large substrate. This form of stamp or step and repeat technology is naturally compatible with the form factor of multiple dies (or devices) per wafer that is currently used in semiconductor fabrication.[170] So these types of tools tend to be targeted for CMOS devices. Molecular Imprints, Inc. is the pioneer in developing step and repeat tools, although other companies including EV Group and Nanonex now offer tools with similar capabilities. Throughput or manufacturability tends to be the primary driver for going from a single imprint to a step and repeat tool. For this reason, most step and repeat tools are based on the UV NIL process, which is usually faster because it lacks the heating and cooling cycles associated with thermal NIL. Furthermore, locally heating a selected region for imprinting on a larger wafer or substrate is difficult. It is much easier and faster to locally expose the imprinted region to UV radiation. Lastly, these step and repeat tools are usually optimized for throughput that comes at the price of the flexibility in the patterning processes or conditions. The general use single imprint tools typically offer greater flexibility. For example, single imprint tools are usually capable of performing either UV or thermal NIL on a variety of materials whereas step and repeat tools tend to be limited to UV NIL patterning of a given resist formulation that has been optimized for the system.

The NIL tools discussed thus far have largely been designed as drop-in replacement technologies for optical lithography, with the target applications in CMOS. However, there have been recent advances in building roll-to-roll NIL tools that are capable of patterning continuous sheets or strips of a nanopatterned substrate.[176,177] Roll-to-roll processes are capable of generating patterned media at a rate at least several orders of magnitude faster than traditional optical lithography, enabling high volume nanofabrication. However, roll-to-roll processes tend to provide very poor overlay capabilities, practically limiting the advantages to single-layer devices. For these reasons, the target applications migrate away from CMOS to emerging markets where nanopatterning has yet to be realized. Examples might include antifouling biological films, polarizing or antireflective optical coatings,[178] brightness enhancing films for large area displays, or flexible electronic devices like radio frequency ID tags.[179] In the roll-to-roll form of NIL, the mold cavities are fabricated onto either a circular drum or a tractor tread-like device. A continuous film is fed in between a pair of rollers that rotate in opposite directions, squeezing their textured surface(s) into the film. Roll-to-roll processes can be implemented in many different forms. The thermal form of NIL can be used to emboss nanoscale topology on the film or substrate by using heated rollers. Likewise, there are examples of UV NIL being implemented to create textured surfaces with roll-to-roll devices. The traditional forms of gravure or flexography can also be adapted to create roll-to-roll NIL transfer printing processes.

19.5.2 Fundamental Issues

There are a few commercial products being made with imprint today, proof that the technology is inherently manufacturable. Some of the companies operating in this space include Nano Opto, Heptagon, Omron, OVD Kingram, LG Electronics, Reflexite, MacDermid, Wavefront Technologies, and Spectratek.

Dr. Michael P.C. Watts is the founder and president of Impattern Solutions, an independent consulting company in the field of NIL. The company's website tracks the progress and commercialization of imprint-related technologies and provides updates at www.impattern.com. Most of the commercial success stories in the imprint field have thus far been related to optics or light. These include, in order of decreasing critical dimensions, refractive, diffractive, and sub-wavelength optical devices. However, even with the sub-wavelength devices, the minimum dimensions of the structures being fabricated commercially are typically on the order of 100 nm or larger, which by many definitions, is not within the nanoscale domain. The community has yet to see the realization of devices fabricated by NIL with dimensions strongly into the sub-100 nm regime. There are many examples of nanoelectronic technologies in the R&D phase that appear to have strong potential with respect to NIL, but several fundamental issues need to be resolved before these ideas can successfully make the transition from the lab to manufacturing. In this section, we briefly summarize some of the remaining technical roadblocks to the commercialization of nanoelectronics via NIL techniques.

19.5.2.1 Nanoimprint Materials Development

There are several examples of high resolution patterning being achieved with NIL. For years, Professor Chou's fabrication of 10 nm vias was the ultimate demonstration of high resolution pattering with NIL.[168] More recently, Professor John Roger's group demonstrated that the fabrication of 2 nm replicas of carbon nanotubes randomly distributed across a rigid substrate could be created within a polydimethylsiloxane (PDMS) resin using NIL processes.[180] Professor Chou's group has gone on to report the patterning periodic line-space patterns with a 6 nm half-pitch,[181] indicating that NIL has the potential resolution and patterning control that is commensurate with nanoelectronics (Figure 19.15).

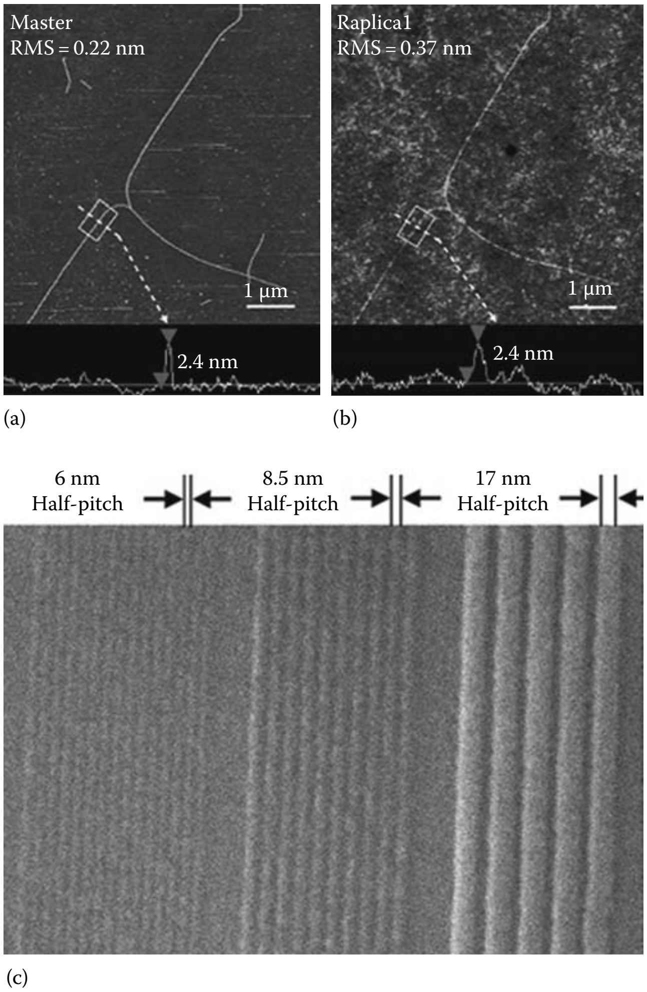

FIGURE 19.15 Examples of high resolution patterning with NIL. Panels (a) shows a few carbon nanotubes with a diameter of 2 nm on a flat surface that served as the model imprint mold while panel (b) shows the resulting patterns that were replicated in PDMS. (Reprinted from Hua, F. et al., *Nano Lett.*, 4(12), 2467, 2004. With permission. © 2004, American Chemical Society.) Panel (c) shows well regular line-space patterns that have been imprinted with critical dimension to a 6 nm half-pitch. (Reprinted from Austin, M.D. et al., *Nanotechnology*, 16(8), 1058, 2005. With permission.)

However, there is still a need for improving NIL materials or resists. Resolution is only part of the picture and a NIL resist must also satisfy a number of different properties. First, the material must be able to readily flow into the imprint mold. If harsh conditions that plastically deform the patterned material are required, the NIL processes may induce residual stress into the patterns.[182–184] It is desirable that the patterned media readily flows into the imprint mold. Likewise, the imprinted patterns must have the mechanical strength to resist pattern collapse. There are several examples of the critical aspect ratio for pattern collapse with decreasing feature size.[185] NIL is capable of generating very small patterns that make pattern collapse a significant concern. Furthermore, it is critical that the mold separates easily from the imprinted structure without destroying the sample. This requires both low surface energy mold treatment strategies that minimize adhesion between the imprint and the substrate as well as imprint materials with the cohesive strength to withstand the shear stress generated during mold separation.[186] Since external forces are required to pry the mold from the substrate, a mechanical load is placed on the patterns. It is a significant materials science challenge to maximize these properties simultaneously especially with functional nanoimprint materials.

19.5.2.2 Template Fabrication and Availability

The fabrication and availability of high quality templates or molds are crucial for NIL patterning technologies. With an ultimate resolution approaching just a few nm, even the slightest flaw in the master has the potential to be replicated into hundreds or thousands of imprinted copies. This means that significant efforts must be made to produce high quality imprint masters. Electron-beam (e-beam) lithography is the industry standard for producing high-resolution patterns with a resolution on the order of 10–20 nm. However, e-beam lithography is not practical for manufacturing since it is a slow, serial writing process and is very costly. One of the greatest attributes of NIL is that it can be considered as an e-beam replication technique. Once a costly pattern is fabricated by e-beam lithography into a mold or master material, thousands of high-resolution copies of the pattern can be replicated via NIL processes at a fraction of the initial cost.[187] This is a very attractive proposition, but it also means that a high quality imprint master is critical. Any mistake will be easily propagated through all of the copies.

19.5.2.3 Defect Inspection and Dimensional Metrology

The need to have high quality imprint templates has been articulated in the preceding section. Hand-in-hand with this requirement is the need to evaluate, quantify, and certify the pattern quality in both the imprint master and the resulting imprints. High-resolution patterning demands high-resolution pattern metrology. We have already mentioned the excessive costs for fabricating a NIL master or mold with large areas patterned with nanoscale features. Quantitative measurements to qualify that the imprint master meets the design specifications in terms of pattern shape are absolutely critical for making this investment. Likewise, fast and efficient high-resolution pattern metrology methods are also needed to monitor the high volume manufacturing processes. It is critical to have quantitative measures in place to determine if the manufacturing processes drift out of tolerance. For high-resolution patterning, this means high-resolution metrology. Even at the current 65 nm technology node, the “Metrology” chapter of the ITRS Roadmap contains roadblocks in terms of high-resolution CD metrologies.[188] The leap from 65 nm optical lithography to NIL processes that offer better than a 5 nm resolution raise significant concerns regarding our ability to quantitatively evaluate the patterns that we fabricate. This is not that big of an issue in the R&D phase of technology development, where one usually deals with highly optimized prototypes. However, high volume manufacturing processes demand quantitative process monitoring capabilities. With respect to pattern shape metrology, the ability to pattern with NIL greatly exceeds the limits of the current inspection tools.

NIL is essentially a direct pattern transfer process whereby the patterns in the imprint have the potential to be a mirror image of the patterns in the mold. This brings up the fidelity of the pattern transfer concept—how closely do the imprinted features resemble the mold. To achieve the greatest fidelity of pattern transfer, one must be able to minimize the shrinkage of the imprinted material in the mold (either up cross-linking or cooling) or any distortions of the pattern after imprinting. The fidelity of the pattern transfer concept is fundamentally unique to NIL. However, in NIL one can directly compare the dimensions of the mold cavity to the pattern to quantify the fidelity of the pattern transfer. This means that dimensional metrologies must be able to quantify both the mold and the imprinted patterns.[189]

19.5.2.4 Residual Layer Control

The objective in most imprint processes is typically to create isolated patterns on the substrate. However, there is always a continuous residual layer of resist material between the substrate and the patterns that needs to be removed. Simple volume filling arguments dictate the thickness of the residual layer, comparing the volume of material per unit area in the smooth film to the volume of the cavities in the mold. Even if there is less material in the film than there is in the mold cavities, the protrusions of the mold cannot be completely pressed into direct contact with the substrate; there is always a finite residual layer on the order of a few nanometers thick. While this residual layer is easily removed through reactive ion etching (RIE) processes, it is critical to achieve both a minimized and uniform residual layer. The minimized residual layer is important to maintain CD control. RIE processes are highly anisotropic, etching primarily vertically, however, there is a small amount of lateral trimming that occurs (Figure 19.16). If one has to etch through a thick residual layer, then the extent of lateral trimming can be significant; minimizing the residual layer minimizes the trimming. Furthermore, it is important to achieve a uniform residual layer across the entire field of the imprint. If there are thick and thin regions of residual layer, the RIE will break through to the substrate first in the thinner regions. When this happens, a micro loading effect occurs in the etch process where the concentration of plasma increases upon breakthrough (the plasma etch chemistries are biased to consume resist over the substrate). The result of micro-loading is an increased lateral etch rate, i.e., a localized enhancement in the lateral trimming. This nonuniform lateral trimming leads to a distribution of the pattern dimensions (loss of CD control).

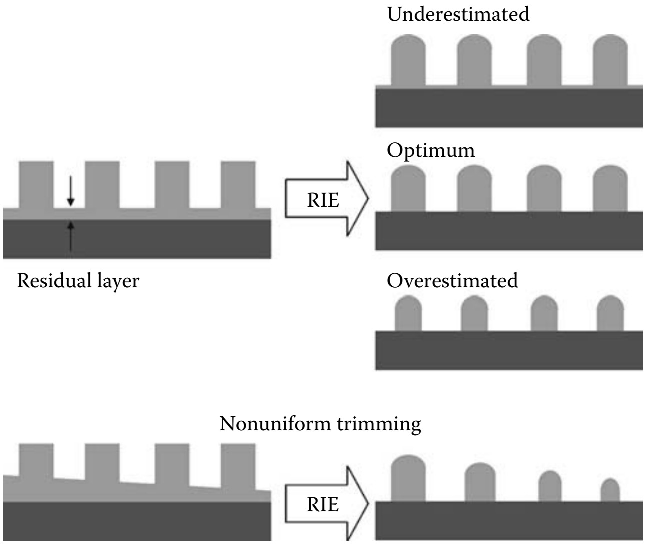

FIGURE 19.16 A schematic illustrating how an accurate knowledge and control of the residual layer thickness is critical to controlling lateral trimming and critical dimension control. Overestimating the thickness of the residual layer can lead to excessive lateral trimming. A non-uniform residual layer can lead to a distribution of linewidths and a loss of CD control through a micro-loading effect.

Controlling the residual layer thickness is straight forward for imprint molds with uniform pattern fields. Simple volume filling calculations can be used to adjust the thickness of the film to minimize the residual layer, assuming that a uniform pressure is applied.[190] However, residual layer control becomes difficult when the pattern field contains a distribution of different patterns with different sizes and density. In this case, the volume filling calculations are different for the different regions of the mold.[191] Resist films are usually applied via spin coating techniques, which leads to a film of uniform thickness. In these situations, it is difficult to control the residual layer. The Molecular Imprints series of tools use a GDS-based drop on demand technology to deposit an optimized amount of resist material in each location to properly fill the mold. This technique can be very effective for minimizing the residual layer.[192,193]

19.5.3 Overlay Accuracy and Control

Depending on the application, it may or may not be important to align the imprint within the lateral directions with some features or fiduca markers on the substrate. There are many applications, such as anti-reflective optical coatings, where pattern overlay is not required. However, for nanoelectronic applications, one could imagine it being very important to align the imprint with circuit elements in the substrate. Most imprint tools offer some type of alignment capability. The alignment is usually achieved by fabricating a series of vernier gratings in both the imprint mold and the patterning substrate and use interferometers to sense very small changes in the in-plane dimensions.[194] Piezoelectric actuators can be incorporated into the imprint head to generate very small lateral position corrections that help precisely align the mold with the substrate. Most general use imprint tools are designed for molds and substrates that are not transparent, meaning that the comparison of the interference pattern from the two verniers is achieved through a series of mirrors and reflective light. With these types of tools, one is able to achieve overlay accuracies on the order of ±200 nm in the in-plane directions. Given the 5 nm or better patterning resolution of NIL, the accuracy with which a feature can be placed within the plane of the film is significantly worse. There are tools with higher overlay accuracy. Molecular Imprints produces highly specialized tools that are optimized for patterning with UV radiation only, using quartz molds that are optically transparent.[195] When it comes to overlay alignment, systems that are dedicated to patterning with transparent molds have a significant advantage because through wafer comparison of the optical verniers is possible. Eliminating the series of mirrors and reflective optics reduces the error in the overlay alignment, enabling sub-15 nm 3σ overlay accuracy. This is more commensurate with the ultimate patterning resolution of a few nm.

19.5.4 Technology Examples

There are a growing number of examples in the literature of nanoelectronic devices being fabricated by NIL processes, proving that the technique has potential (Figure 19.17). The most common examples are the cross-bar memory structures where a semiconducting material is sandwiched between orthogonally situated parallel line-space electrodes to be used as either data storage or logic devices.[196] Here, NIL processes are used to fabricate the parallel line-space electrode arrays. These devices can have half-pitch values as small as 17 nm for the electrode arrays and memory densities as high as 100 Gbit/cm^2. On the CMOS front, a recent joint effort from Toshiba and Molecular Imprints have demonstrated the fabrication of a functional sub-32 nm logic circuit using NIL.[197] Likewise, IBM and SEMATECH have reported the fabrication of 27 nm functional FINFET structures using NIL technologies.[198] Another sector where NIL processes are being widely adapted is for magnetic data storage media. The equivalent Moore's law for the hard disc industry is even more aggressive than that of the semiconductor industry, doubling in data storage capacity ever 12 months. To keep pace with this trend, the industry is moving towards bit patterned media. NIL processes are being developed to fabricate 25 nm posts of magnetic media on a pitch of approximately 40 nm.[199] These are just a few examples related to nanoelectronics where NIL technologies are being actively pursued. There are also a number of other applications in the optical and biotechnology fields where NIL methods are poised to have an impact. Two of the most prominent examples include wire grid polarizer elements for screen and display technologies[200,201] and photonic structures to facilitate light extraction in high brightness LEDs.[187]

19.5.5 Future Perspectives

In some respects, NIL is just one of a long list of next-generation lithography techniques that has been developed as an alternative to deep UV optical lithography for the semiconductor industry. However, none of these alternate technologies have even come close to supplanting optical lithography as the high volume lithography of choice for the electronics industry. While a debate over the merits of NIL to replace optical lithography for CMOS fabrication is beyond the scope of this review, it is important to realize that NIL is the first next generation lithography candidate to have a viable life outside of CMOS, and there are already success stories for the patterning technology in non-CMOS technologies. This means that the success or failure of NIL is not tied to the semiconductor industry; it has already succeeded. As microelectronics moves beyond CMOS into nanoelectronics with novel device architectures, dramatically different principals of operation, and the need for nanoscale patterning, NIL is on-pace to be a well-established generic nanoscale patterning technology of relevance. Next generation lithography techniques tied directly to the CMOS community (193 nm immersion, double patterning, EUV) are solely dependent on this community for success or failure. These patterning technologies will likely go away if they are not adapted by the semiconductor community. The same is not true for NIL.

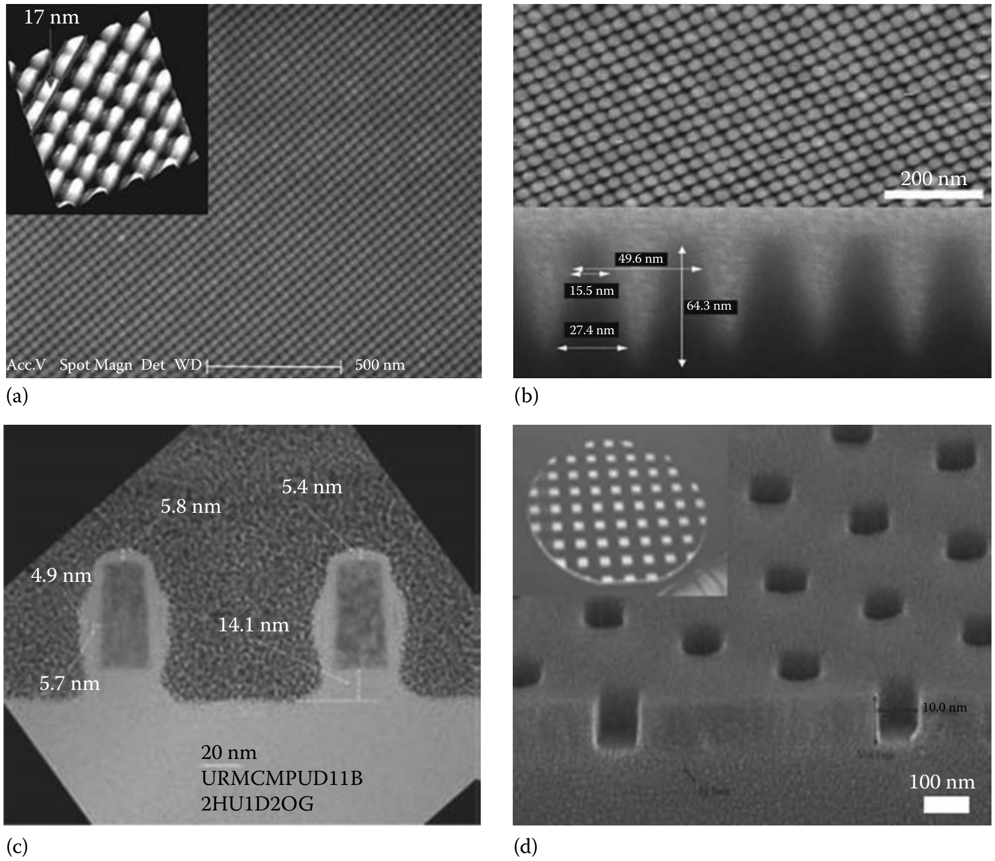

FIGURE 19.17 Examples of several different technologies fabricated with NIL techniques: (a) a high-resolution image taken from the 17 nm half-pitch cross-bar nanowire circuit arrays made by SNAP. (Reprinted from Jung, G.Y. et al., *Nano Lett.*, 6(3), 351, 2006. With permission. © 2006, American Chemical Society.) (b) Bit patterned magnetic media for high density data storage showing 25 nm magnetic posts fabricated by NIL. (Reprinted from Dobisz, E.A. et al., *Proc. IEEE*, 96(11), 1836, 2008. With permission. © 2008, IEEE.) (c) A cross-sectional view of a 27 nm FinFET structure whose critical dimension was patterned by Step-and-Flash Imprint Lithography at Molecular Imprints, Austin, TX. Nominally identical structures led to functional FinFET devices. (From IBM Almaden Research Center, San Jose, CA. With permission.) (d) An array of photonic devices patterned uniformly over an entire 6 nm wafer for high brightness LED devices. (Reprinted from Miller, M. et al., *Proc. SPIE*, 6883, 68830D, 2008. With permission.)

19.6 Metrology for Nanolithography

19.6.1 Advanced Lithographic Processes

As optical lithography reaches its limits, new processes such as double patterning and double exposure are being explored as means of extending 193 nm immersion-based lithography. The goal of these patterning methods is patterning at dimensions that are less than those that can be printed in a single lithographic exposure step. Each of these process-based methods poses a unique challenge to all aspects of process control including metrology. These processes are summarized in Figures 19.18 through 19.20.[188]

Double exposure uses two reticles that must be precisely aligned so that the pattern placement is one pitch across the printed area. One photoresist layer is exposed twice with specially designed reticles resulting in densely spaced lines. This pattern can then be etched into the underlying films patterning the wafer with features having a ½ pitch less than can be printed using one reticle. This process is shown in Figure 19.18.

Double patterning is also done using two reticles that must be aligned with accuracy, so that the pattern placement has one pitch and close to one linewidth across the printed area. The first photoresist layer is exposed and then the film stack is etched leaving the pattern in the hard mask layer. Then the wafer is coated with BARC and resist and the second set of lines is patterned into the second BARC/resist layer. The resultant double pattern structure is etched into the poly silicon. Figure 19.19 provides great detail for double patterning processes for lines. Similar process steps are used for contacts and trenches.

Spacer double patterning is believed to be already in use in manufacturing. A slightly different film stack is used for spacer double patterning. First, a sacrificial poly silicon layer is patterned, spacer oxide is deposited, and the sacrificial poly silicon layer is removed resulting in a spacer pattern. This pattern is etched into the poly silicon layer. In Figure 19.20, this single reticle process is more fully described. The origin of the two sets of line shapes is due to the different shapes of each side of a spacer. The two sets of critical dimensions are due to the nature of the spacer deposition process.

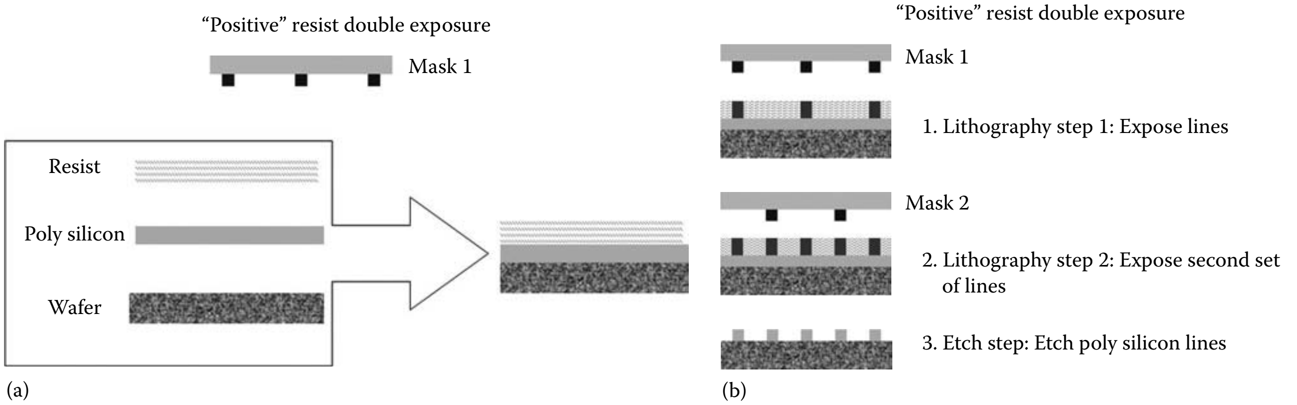
"Positive" resist double exposure
Mask 1
Resist
Poly silicon
Wafer
(a)
"Positive" resist double exposure
Mask 1
1. Lithography step 1: Expose lines
Mask 2
2. Lithography step 2: Expose second set of lines
3. Etch step: Etch poly silicon lines
(b)

FIGURE 19.18 The double exposure method of patterning lines is shown. Lines are patterned using positive resist. (a) shows the film stack, and (b) shows the patterning process.

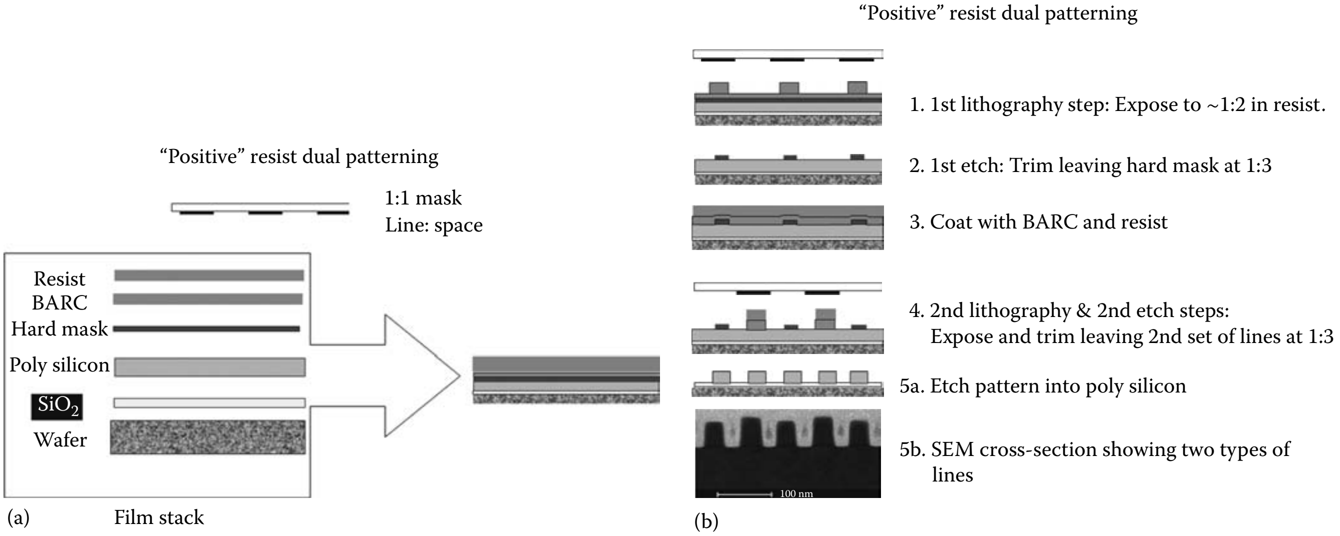
"Positive" resist dual patterning
1:1 mask
Line: space
Resist
BARC
Hard mask
Poly silicon
SiO_2
Wafer
(a)
Film stack
"Positive" resist dual patterning
1. 1st lithography step: Expose to ~1:2 in resist.
2. 1st etch: Trim leaving hard mask at 1:3
3. Coat with BARC and resist
4. 2nd lithography & 2nd etch steps: Expose and trim leaving 2nd set of lines at 1:3
5a. Etch pattern into poly silicon
5b. SEM cross-section showing two types of lines
100 nm
(b)

FIGURE 19.19 The double patterning method of patterning lines is shown. Lines are patterned using positive resist. (a) shows the film stack, and (b) shows the patterning process.

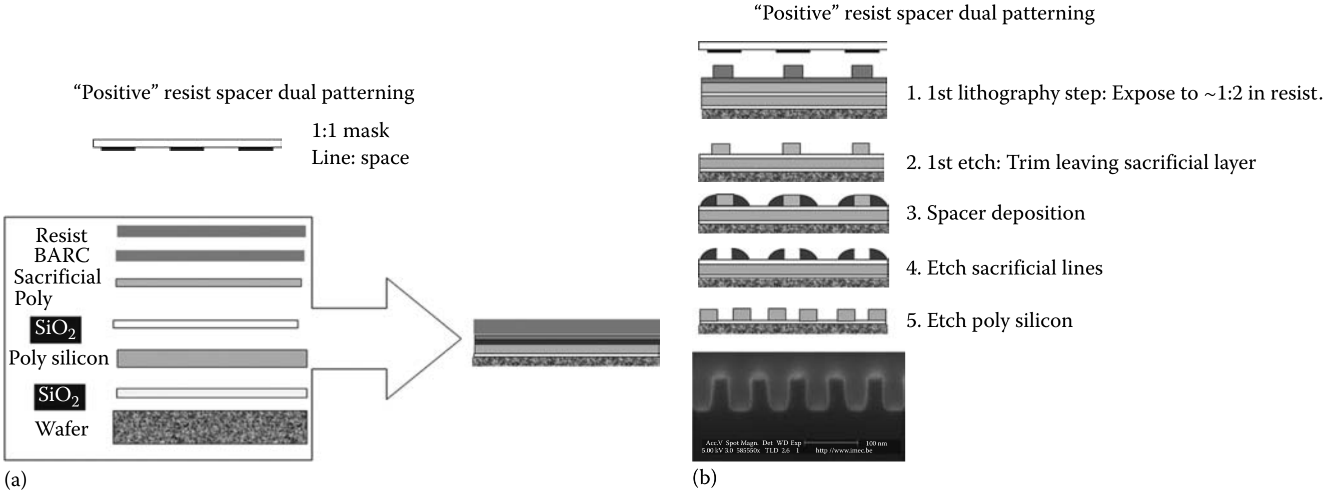
"Positive" resist spacer dual patterning
1:1 mask
Line: space
Resist
BARC
Sacrificial Poly
SiO_2
Poly silicon
SiO_2
Wafer
(a)
"Positive" resist spacer dual patterning
1. 1st lithography step: Expose to ~1:2 in resist.
2. 1st etch: Trim leaving sacrificial layer
3. Spacer deposition
4. Etch sacrificial lines
5. Etch poly silicon
100 nm
http://www.imec.be
(b)

FIGURE 19.20 The spacer double patterning method is shown. Lines are patterned using positive resist. (a) shows the film stack, and (b) shows the patterning process.

Overlay plays an important role in double pattering, spacer patterning, and double exposure pattering of lines. Overlay must be near perfect to achieve a uniform pitch. If the second exposure is shifted, every other space will be different. This impacts further processing, the measurement of CD and lineshape, and device performance. Etch is known to be sensitive to lineshape and the distance between lines. Etch uniformity can be reduced when the resist pattern is etched into the layer below.

19.6.2 Metrology for Advanced Lithography Processes

Measurement needs start with control of feature shape and placement on the mask (reticle), and follow through until the features are etched into the active layer of the device or interconnect structure.[188] Since optical masks are glass with chrome features, scanning electron microscope (SEM)-based measurements must overcome the charging inherent in the measurement of an insulator. The chrome features must have the correct linewidth and shape and must be placed on the mask in the correct position relative to each other. The recent introduction of "environmental" SEM allows for charge neutralization during the SEM-based measurement of CD and line shape on the mask.

The advanced lithographic process challenge metrology mainly through the two distributions of critical dimensions and line shapes especially the side wall angles of the lines. Two methods of CD measurement are used during manufacturing, namely CD-scanning electron microscopy (CD-SEM) and scatterometry. The term scatterometry typically refers to one of two optical methods of measuring critical dimensions using a grating test structure. One example of a grating test structure is an array of lines with a constant pitch and linewidth. Others include an array of contact holes. Scatterometry measurements can be done in a spectroscopic ellipsometer; the main challenge is analyzing the changes in the optical response that the grating structure imparts versus the response of an un-patterned film stack. The other approach to scatterometry is to measure the diffraction pattern from a single wavelength across a range of angles. Both methods of scatterometry require that an extensive database of simulated ellipsometric responses be placed into a library. Some libraries contain tens to hundreds of thousands of simulated structures for a single process step. Data from a sample is matched to this set of simulated responses. Accuracy requires that ellipsometry be sensitive to the small differences in optical response.

Ellipsometry measures the change in polarization of light after reflection from the sample surface. Spectroscopic ellipsometry-based scatterometry relies on the changes in polarization due to changes in the grating test structure as a function of wavelength. Advanced lithographic processes pose two challenges to scatterometry. One is shrinking feature size and the second is the two sets of distributions of CD and feature shape. Sensitivity to small changes in dimension is becoming increasingly difficult as feature size decreases below 50 nm for both CD-SEM and scatterometry. The previous method of single exposure and etch patterning required that scatterometry measure the average CD and line shape of a uniform grating structure. Double patterning and double exposure requires that scatterometry produce the average CD and line shape of two distributions of line shape. In situations where overlay errors result in a nonuniform pitch, the number of simulated data sets is greatly increased. Small differences in optical response make it very difficult to distinguish changes in CD and line shape. Recent commercial instrumentation extended the wavelength range of scatterometry further into the UV down to 150 nm. In general, UV wavelengths have improved the ability of scatterometry to measure CD and line shape for features with sub 50 nm dimensions.

The specifics of ellipsometry-based scatterometry are useful in understanding how critical dimensions are determined from a method associated with the determination of film thickness for unpatterned layers. The premise of ellipsometry is that light polarized in-plane of reflection (P) reflects differently than light polarized perpendicular to the plane of reflection (S). Ellipsometry determines the change in polarization of light in terms of Δ, which is the difference in phase change = $\Delta_P - \Delta_S$ and Ψ, which is related to the ratio of the change in intensity upon reflection through $\tan^{-1}(|R_P|/|R_S|)$, where R_P and R_S are the complex reflectivity's parallel and perpendicular to the reflectance plane. The reflectivity is determined by the dielectric function of the film stack. Reflection from a single film on a substrate depends on the complex dielectric function (refractive index) of the materials and the Fresnel reflection coefficients r_{12} between (1) air and (2) the film and r_{23} between (2) film and (3) the substrate. The reflection from an unpatterned film is determined by well-understood reflection coefficients[202]

$$R_P = \frac{(r_{12}^P + r_{23}^P e^{-i2\beta})}{(1 + r_{12}^P r_{23}^P e^{-i2\beta})} \tag{19.6}$$

$$R_S = \frac{(r_{12}^S + r_{23}^S e^{-i2\beta})}{(1 + r_{12}^S r_{23}^S e^{-i2\beta})} \tag{19.7}$$

where

$$\beta = 2\pi\left(\frac{d}{\lambda}\right) N_2 \cos\phi \tag{19.8}$$

The film thickness dependence comes from the exponential term β where d is the film thickness, λ is the wavelength of the light, and N_2 is the complex refractive index of the film. The measurement of a patterned film stack requires that Maxwell's equations be solved. Typically, this is done using a rigorous coupled wave approximation (RCWA), and the RCWA equations must be solved at each wavelength for each structure. Changes in Ψ and Δ vs wavelength require solving the RCWA equations at enough wavelengths to observe subtle changes due to CD and line shape.[202]

The CD-SEM measurement of critical dimensions is also challenged by double patterning and double exposure. Measuring patterned photoresist requires the calibration of the rate of shrinkage of the resist vs. electron beam exposure time. This calibration requires the collection of enough data to provide statistically significant shrinkage rates. This must be done for each resist material. It is important to note that different resists are used for the fabrication of line and vias. Many IC manufacturers hold the resist formulas to be proprietary meaning that each resist formula will have different responses to electron beam exposure. Recent unpublished studies show that resist lines with a CD less than 25 nm shrink more rapidly than lines with a greater CD. Scatterometry measurements may also impact resist lines especially when measurements are done in UV.

There are important differences between CD-SEM and scatterometry in terms of the nature of information each provides. Advanced CD-SEM is capable of measuring the CD of two or more lines simultaneously. Future CD-SEMs may be able to measure several lines in a test area providing the average CD as well as the range of the CD in the test area. Scatterometry provides the average CD in the test area.

The measurement of line edge roughness in resist is critical for the control of device performance including leakage current. Industry standards organizations such as SEMI have driven consensus documents describing measurement methods for quantifying line edge roughness.[203] Line edge roughness on a nm scale produces gate CD width roughness that results in an increase in leakage current. Line edge roughness at longer scales produces an increase in carrier surface scattering that increases line resistivity. At this time, line edge roughness measurements can be done using CD-SEM or CD-AFM.

Overlay metrology is considerably more challenging for advanced lithography processes. Overlay measurements are typically done using specially designed optical microscopes and target structures that are patterned into both the upper and lower layers. Double patterning of trenches for interconnect lines have additional complications for metrology. If the overlay is shifted, then the trench CD also changes. Overlay measurement precision for all double patterning methods becomes at best a factor of 0.7 of the overlay required for a single exposure patterning of all the trenches. This has driven the development of new overlay test structures that include scattering bars that diffract light. SEM may provide an alternate means of measuring overlay.

19.6.3 Proposed New CD Metrology Methods

Many have speculated about the demise of CD-SEM and scatterometry. Even optical microscopy-based CD measurement is still done for features have larger CDs such as microelectromechanical systems (MEMS). It is possible that CD-SEM and scatterometry can not extend beyond the 22 nm half pitch. Two methods are being considered as potential replacements for scatterometry and CD-SEM. These are small angle x-ray scattering, which is known as CD-SAXS,[204] and He ion microscopy.[205] In CD-SAXS, a focused, monochromatic x-ray beam is transmitted through a grating test structure, as shown in Figure 19.21, that is similar to the ones used for scatterometry measurement. It is important to note that CD-SAXS can measure line edge roughness, average CD, individual CD, and line shape for the lines in the test structure.[206–209] At this time (2009), evaluation of the capabilities of He ion microscopy are in the initial stages.

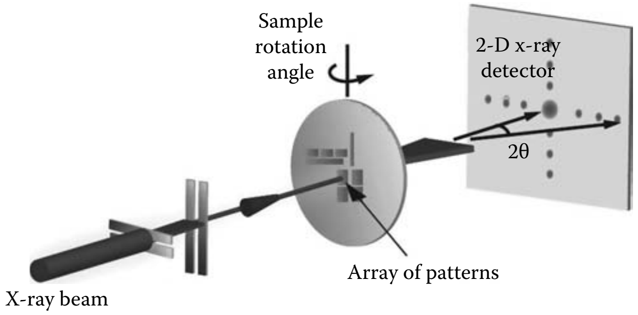

FIGURE 19.21 CD SAXS schematic of using collimated x-ray beams with energy that are transmissive through a patterned silicon wafer resulting in scattering from an array of patterns.

Abbreviations

α	Fraction of light absorbed, when referring to resist technology
α	Width of the forward-scattering distribution, when referring to electron-beam lithography
β	Width of the backscattered distribution, when referring to electron-beam lithography
δ	Photoacid diffusion length, when referring to resist technology
ε	Photoacid generation efficiency, when referring to resist technology
η	The effective dose contributed by the backscattered electrons, when referring to electron-beam lithography
BARC	Bottom anti-reflective coating
CD	Critical dimension
CD-SEM	Critical dimension scanning electron microscope
CD-SAXS	Critical dimension small-angle x-ray scattering
CD-AFM	Critical dimension atomic force microscopy
CMOS	Complementary metal-oxide semiconductor
CMP	Chemical mechanical polishing
De	Dose at line-edge, when referring to resist technology
D_0	Incident dose, when referring to electron-beam lithography
DOF	Depth of focus
DUV	Deep-ultraviolet
E_0	Incident electron energy, when referring to electron-beam lithography
EUV	Extreme-ultraviolet
FINFET	Nonplanar, double-gate field effect transistor, “fin” refers to shape of semiconductor
HP	Half-pitch

HSQ	Hydrogen silsesquioxane
IC	Integrated circuit
ITRS	*International Technology Roadmap for Semiconductors*
LILS	Latent-image log slope
LED	Light emitting diode
LER	Line-edge roughness
LWR	Line-width roughness
MEMS	Microelectromechanical systems
MG	Molecular glass
ML2	Maskless lithography
NA	Numerical aperture
NIL	Nanoimprint lithography
PAG	Photoacid generator
PDMS	Polydimethylsiloxane
q	Number of photons/nm^2, when referring to resist technology
RIE	Reactive ion etch
SCALPEL	Scattering with angular limitation in projection electron-beam lithography
STM	Scanning tunneling microscopy
UV	Ultraviolet

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

Certain commercial equipment and materials are identified in this paper in order to specify adequately the experimental procedure. In no case does such identification imply recommendations by the National Institute of Standards and Technology nor does it imply that the material or equipment identified is necessarily the best available for this purpose.

References

1. Moore, G. E., Cramming more components onto integrated circuits, *Electronics*, **38**, 114–117 (1965).
2. Huff, H., *Into the Nano Era, Springer Series in Materials Science*, Springer-Verlag, Berlin, Germany, **106**, (2009).
3. Maydan, D., Coquin, G. A., Maldonado, J. R. et al., High-speed replication of submicron features on large areas by x-ray lithography, *IEEE Transactions on Electron Devices*, **ED22**(7), 429 (1975).
4. Silverman, J. P., X-ray lithography: Status, challenges, and outlook for 0.13 mu m, *Journal of Vacuum Science & Technology B*, **15**(6), 2117 (1997).
5. Harriott, L. R., Berger, S. D., Biddick, C. et al., The SCALPEL proof of concept system, *Microelectronic Engineering*, **35**(1–4), 477 (1997).
6. Liddle, J. A., Berger, S. D., Biddick, C. J. et al., The scattering with angular limitation in projection electron-beam lithography (SCALPEL) system, *Japanese Journal of Applied Physics Part 1—Regular Papers Short Notes & Review Papers*, **34**(12B), 6663 (1995).
7. Tennant, D., Limits of conventional lithography, in G. M. Timp (ed.), *Nanotechnology*, AIP/Springer, New York, Chap. 4, p. 161 (1999).
8. Marrian, C. R. K. and Tennant, D. M., Nanofabrication, *Journal of Vacuum Science & Technology A*, **21**(5), S207–S215 (2003).
9. Brunner, T. A., Why optical lithography will live forever, *Journal of Vacuum Science & Technology B*, **21**(6), 2632 (2003).
10. Levinson, H. J., *Principles of Lithography*, 2nd edn., SPIE Press, Bellingham, WA, (2005).
11. Brillouet, M., An introduction to ultimate lithography, *Comptes Rendus Physique*, **7**(8), 837 (2006).
12. Ito, H., Chemical amplification resists for microlithography, *Advances in Polymer Science*, **172**, 37 (2005).
13. Chandhok, M., Datta, S., Lionberger, D., and Vesecky, S., Impact of line-width roughness on Intel's 65-nm process devices, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65191A (2007).
14. *International Technology Roadmap for Semiconductors*, 2007 edition, http://www.itrs.net, Lithography, p. 10. SEMATECH, Austin, TX.
15. Wallraff, G. M. and Hinsberg, W. D., Lithographic imaging techniques for the formation of nanoscopic features, *Chemical Reviews*, **99**(7), 1801 (1999).
16. Gallatin, G. M., Naulleau, P., Niakoula, D. et al., Resolution, LER, and sensitivity limitations of photoresists, *Emerging Lithographic Technologies XII, SPIE*, **6921**, 69211E (2008).
17. Bristol, R. L., The tri-lateral challenge of resolution, photospeed, and LER: Scaling below 50 nm? *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65190W (2007).
18. Van Steenwinckel, D., Gronheid, R., Lammers, J. H. et al., A novel method for characterizing resist performance, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65190V (2007).
19. Wallow, T., Higgins, C., Brainard, R. et al., Evaluation of EUV resist materials for use at the 32 nm half-pitch node, *Emerging Lithographic Technologies XII, SPIE*, **6921**, 69211F (2008).
20. Fedynyshyn, T. H., Goodman, R. B., and Roberts, J., Polymer matrix effects on acid generation, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692319 (2008).
21. Pawloski, A. R., Christian, and Nealey, P. F., A standard addition technique to quantify photoacid generation in chemically amplified photoresist, *Chemistry of Materials*, **13**(11), 4154 (2001).
22. Szmanda, C. R., Brainard, R. L., Mackevich, J. F. et al., Measuring acid generation efficiency in chemically amplified resists with all three beams, *Journal of Vacuum Science & Technology B*, **17**(6), 3356 (1999).
23. Brainard, R., Hassanein, E., Li, J. et al., Photons, electrons, and acid yields in EUV photoresists: A progress report, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692325 (2008).

24. Hinsberg, W., Houle, F. A., Hoffnagle, J. et al., Deep-ultraviolet interferometric lithography as a tool for assessment of chemically amplified photoresist performance, *Journal of Vacuum Science & Technology B*, **16**(6), 3689 (1998).
25. Pawloski, A., Acheta, A., Lalovic, I., LaFontaine, B., and Levinson, H., Characterization of line edge roughness in photoresist using an image fading technique, *Proceedings of the SPIE, Advances in Resist Technology and Processing XXI*, **5376**, 414 (2004).
26. Houle, F. A., Hinsberg, W. D., Sanchez, M. I., and Hoffnagle, J. A., Influence of resist components on image blur in a patterned positive-tone chemically amplified photoresist, *Journal of Vacuum Science & Technology B*, **20**(3), 924 (2002).
27. Houle, F. A., Hinsberg, W. D., Morrison, M. et al., Determination of coupled acid catalysis-diffusion processes in a positive-tone chemically amplified photoresist, *Journal of Vacuum Science &Technology B*, **18**(4), 1874 (2000).
28. Croffie, E., Yuan, L., Cheng, M. S. et al., Modeling influence of structural changes in photoacid generators an 193 nm single layer resist imaging, *Journal of Vacuum Science & Technology B*, **18**(6), 3340 (2000).
29. Itani, T., Yoshino, H., Fujimoto, M., and Kasama, K., Photoacid bulkiness on dissolution kinetics in chemically amplified deep ultraviolet resists, *Journal of Vacuum Science & Technology B*, **13**(6), 3026 (1995).
30. Vogt, B. D., Kang, S., Prabhu, V. M. et al., Measurements of the reaction-diffusion front of model chemically amplified photoresists with varying photoacid size, *Macromolecules*, **39**(24), 8311 (2006).
31. Vogt, B. D., Kang, S., Prabhu, V. M. et al., The deprotection reaction front profile in model 193 nm methacrylate-based chemically amplified photoresists, *Advances in Resist Technology and Processing XXIII, Proceedings of SPIE*, **6153**, 615316 (2006).
32. Shi, X. L., Effect of Coulomb interaction and pKa on acid diffusion in chemically amplified resists, *Journal of Vacuum Science & Technology B*, **17**(2), 350 (1999).
33. Goldfarb, D. L., Angelopoulos, M., Lin, E. K. et al., Confinement effects on the spatial extent of the reaction front in ultrathin chemically amplified photoresists, *Journal of Vacuum Science & Technology B*, **19**(6), 2699 (2001).
34. Lin, E. K., Soles, C. L., Goldfarb, D. L. et al., Direct measurement of the reaction front in chemically amplified photoresists, *Science*, **297**(5580), 372 (2002).
35. Vogt, B. D., Kang, S., Prabhu, V. M. et al., Influence of base additives on the reaction diffusion front of model chemically amplified photoresists, *Journal of Vacuum Science & Technology B*, **25**, 175 (2006).
36. Wu, W. L., Prabhu, V. M., and Lin, E. K., Identifying materials limits of chemically amplified photoresists, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 651902 (2007).
37. Lavery, K. A., Choi, K. W., Vogt, B. D. et al., Fundamentals of the reaction-diffusion process in model EUV photoresists, *Advances in Resist Technology and Processing XXIII, SPIE*, **6153**, 615313 (2006).
38. Houlihan, F., Person, D., Nalamasu, O. et al., Study of base additives for use in a single layer 193 nm resist based upon poly(norbornene/maleic anhydride/acrylic acid/tert-butyl acrylate), *Proceedings of the SPIE—The International Society for Optical Engineering*, **4345**, 67 (2001).
39. Pawloski, A. R., Christian, and Nealey, P. F., The multifunctional role of base quenchers in chemically amplified photoresists, *Chemistry of Materials*, **14**(10), 4192 (2002).
40. Houle, F. A., Hinsberg, W. D., and Sanchez, M. I., Acid-base reactions in a positive tone chemically amplified photoresist and their effect on imaging, *Journal of Vacuum Science & Technology B*, **22**(2), 747 (2004).
41. Hinsberg, W. D., Houle, F. A., Sanchez, M. I., and Wallraff, G. M., Chemical and physical aspects of the post-exposure baking process used for positive-tone chemically amplified resists, *IBM Journal of Research and Development*, **45**(5), 667 (2001).
42. Hinsberg, W., Houle, F., Sanchez, M. et al., Effect of resist components on image spreading during postexposure bake of chemically amplified resists, *Proceedings of the SPIE—The International Society for Optical Engineering*, **3999**, 148 (2000).
43. Michaelson, T. B., Pawloski, A. R., Acheta, A., Nishimura, Y., and Willson, C. G., The effects of chemical gradients and photoresist composition on lithographically generated line edge roughness, *Proceedings of SPIE*, **5753**, 368 (2005).
44. Hinsberg, W., Houle, F. A., Lee, S. W., Ito, H., and Kanazawa, K., Characterization of reactive dissolution and swelling of polymer films using a quartz crystal microbalance and visible and infrared reflectance spectroscopy, *Macromolecules*, **38**(5), 1882 (2005).
45. Hinsberg, W. D., Houle, F. A., Ito, H., Kanazawa, K., and Lee, S. W., Kinetics of reactive dissolution of lithographic polymers, *Proceedings of the 13th International Conference on Photopolymers*, *Advances in Imaging Materials and Processes*, *RETEC* 2003, 193 (2003).
46. Hinsberg, W. D., Houle, F. A., and Ito, H., Reactive dissolution of lithographic copolymers, *Proceedings of the SPIE, Advances in Resist Technology and Processing XXI*, **5376**, 352 (2004).
47. Houle, F. A., Hinsberg, W. D., and Sanchez, M. I., Kinetic model for positive tone resist dissolution and roughening, *Macromolecules*, **35**(22), 8591 (2002).
48. Prabhu, V. M., Vogt, B. D., Kang, S. et al., Direct measurement of the spatial extent of the in situ developed latent image by neutron reflectivity, *Journal of Vacuum Science & Technology B*, **25**(6), 2514 (2007).
49. Prabhu, V. M., Vogt, B. D., Kang, S. et al., Direct measurement of the in situ developed latent image: The residual swelling fraction, *Proceedings of SPIE*, **6519**, 651910 (2007).
50. Prabhu, V. M., Rao, A., Kang, S., Lin, E. K., and Satija, S. K., Manipulation of the asymmetric swelling fronts of photoresist polyelectrolyte gradient thin films, *The Journal of Physical Chemistry B*, **112**(49), 15628 (2008).

51. Wang, M. X., Gonsalves, K. E., Rabinovich, M., Yueh, W., and Roberts, J. M., Novel anionic photoacid generators (PAGs) and corresponding PAG bound polymers for sub-50 nm EUV lithography, *Journal of Materials Chemistry*, **17**(17), 1699 (2007).
52. Wang, M. X., Lee, C. T., Henderson, C. L. et al., Incorporation of ionic photoacid generator (PAG) and base quencher into the resist polymer main chain for sub-50 nm resolution patterning, *Journal of Materials Chemistry*, **18**(23), 2704 (2008).
53. Fukushima, Y., Watanabe, T., Ohnishi, R. et al., Optimization of photoacid generator in photoacid generation-bonded resist, *Japanese Journal of Applied Physics*, **47**(8), 6293 (2008).
54. Wu, H. P. and Gonsalves, K. E., Preparation of a photoacid generating monomer and its application in lithography, *Advanced Functional Materials*, **11**(4), 271 (2001).
55. Hirayama, T., Shiono, D., Hada, H., and Onodera, J., New photoresist based on amorphous low molecular weight polyphenols, *Journal of Photopolymer Science and Technology*, **17**(3), 435 (2004).
56. VanderHart, D. L., De Silva, A., Felix, N., Prabhu, V. M., and Ober, C. K., The effect of EUV molecular glass architecture on the bulk dispersion of a photo-acid generator, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69231M (2008).
57. Young-Gil, K., Kim, J. B., Fujigaya, T., Shibasaki, Y., and Ueda, M., A positive-working alkaline developable photoresist based on partially tert-Boc-protected calix[4]resorcinarene and a photoacid generator, *Journal of Materials Chemistry*, **12**(1), 53 (2002).
58. Chang, S. W., Ayothi, R., Bratton, D. et al., Sub-50 nm feature sizes using positive tone molecular glass resists for EUV lithography, *Journal of Materials Chemistry*, **16**(15), 1470 (2006).
59. Yang, D., Chang, S. W., and Ober, C. K., Molecular glass photoresists for advanced lithography, *Journal of Materials Chemistry*, **16**(18), 1693 (2006).
60. De Silva, A., Lee, J. K., Andre, X. et al., Study of the structure-properties relationship of phenolic molecular glass resists for next generation photolithography, *Chemistry of Materials*, **20**(4), 1606 (2008).
61. Felix, N. M., De Silva, A., Luk, C. M. Y., and Ober, C. K., Dissolution phenomena of phenolic molecular glass photoresist films in supercritical CO2, *Journal of Materials Chemistry*, **17**(43), 4598 (2007).
62. De Silva, A., Felix, N., Sha, J., Lee, J. K., and Ober, C. K., Molecular glass resists for next generation lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69231L (2008).
63. De Silva, A., Felix, N., Forman, D., Sha, J., and Ober, C. K., New architectures for high resolution patterning, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69230O (2008).
64. Toriumi, M., Santillan, J., Itani, T., Kozawa, T., and Tagawa, S., Dissolution characteristics and reaction kinetics of molecular resists for extreme-ultraviolet lithography, *Journal of Vacuum Science & Technology B*, **25**(6), 2486 (2007).
65. Nishimura, I., Heath, W. H., Matsumoto, K. et al., Non-chemically amplified resists for 193 nm lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69231C (2008).
66. Lawson, R. A., Lee, C. T., Yueh, W., Tolbert, L., and Henderson, C. L., Single molecule chemically amplified resists based on ionic and non-ionic PAGs, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69230K (2008).
67. Naulleau, P. P., Anderson, C. N., Chiu, J. et al., Latest results from the SEMATECH Berkeley extreme ultraviolet microfield exposure tool, *Journal of Vacuum Science & Technology B* **27**, 66 (2009).
68. Woodward, J. T., Choi, K. W., Prabhu, V. M. et al., Characterization of the latent image to developed image in model EUV photoresists, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69232B (2008).
69. Choi, K. W., Prabhu, V. M., Lavery, K. A. et al., Effect of photo-acid generator concentration and developer strength on the patterning capabilities of a model EUV photoresist, *Proceedings of SPIE*, **6519**, 651943 (2007).
70. Zimmerman, P. A., Byers, J., Piscani, E. et al., Development of an operational high refractive index resist for 193 nm immersion lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692306 (2008).
71. Matsumoto, K., Costner, E., Nishimura, I., Ueda, M., and Willson, C. G., High-index resist for 193-nm immersion lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692305 (2008).
72. Gonsalves, K. E., Wang, M., and Pujari, N. S., High refractive-index resists composed of anionic photoacid generator (PAG) bound polymers for 193 nm immersion lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69231P (2008).
73. Liu, H. P., Blakey, I., Conley, W. E. et al., Application of quantitative structure property relationship to the design of high refractive index 193i resist, *Journal of Micro-Nanolithography MEMS and MOEMS*, **7**(2), 023001 (2008).
74. Allen, R. D., Brock, P. J., Sundberg, L. et al., Design of protective topcoats for immersion lithography, *Journal of Photopolymer Science and Technology*, **18**(5), 615 (2005).
75. Terai, M., Kumada, T., Ishibashi, T. et al., Mechanism of immersion specific defects with high receding-angle topcoat, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65191S (2007).
76. Takebe, Y., Shirota, N., Sasaki, T., and Yokokoji, O., Development of top coat materials for ArF immersion lithography, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65191Y (2007).
77. Nakagawa, H., Goto, K., Shima, M. et al., Process optimization for developer soluble immersion topcoat material, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 651923 (2007).

78. Sundberg, L. K., Sanders, D. P., Sooriyakumaran, R., Brock, P. J., and Allen, R. D., Contact angles and structure/surface property relationships of immersion materials, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 65191Q (2007).
79. Shirota, N., Takebe, Y., Wang, S. Z., Sasaki, T., and Yokokoji, O., Development of non-topcoat resist polymers for 193-nm immersion lithography, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 651905 (2007).
80. Wada, K., Kanna, S., and Kanda, H., Novel materials design for immersion lithography, *Advances in Resist Materials and Processing Technology XXIV, SPIE*, **6519**, 651908 (2007).
81. Wu, S., Tseng, A., Lin, B. et al., Non-topcoat resist design for immersion process at 32-nm node, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692307 (2008).
82. Sanders, D. P., Sundberg, L. K., Brock, P. J. et al., Self-segregating materials for immersion lithography, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 692309 (2008).
83. Wang, D., Caporale, S., Andes, C. et al., Design consideration for immersion 193: Embedded barrier layer and pattern collapse margin, *Journal of Photopolymer Science and Technology*, **20**(5), 687 (2007).
84. Smith, B. W., Optics for photolithography, in K. Suzuki and B. W. Smith (eds.), *Microlithography: Science and Technology*, 2nd edn., CRC Press, Boca Raton, FL, (2007).
85. Lin, B. J., The k_3 coefficient in nonparaxial λ/NA scaling equations for resolution, depth of focus, and immersion lithography, *Journal of Microlithography Microfabrication and Microsystems*, **1**(1), 7 (2002).
86. Oldham, W. G. and Schenker, R. E., 193-nm lithographic system lifetimes as limited by UV compaction, *Solid State Technology*, **40**(4), 95–102 (1997).
87. Sewell, H., McClay, J., Jenkins, P. et al., 157 nm lithography—Window of opportunity, *Journal of Photopolymer Science and Technology*, **15**(4), 569 (2002).
88. Burnett, J. H., Levine, Z. H., and Shirley, E. L., Intrinsic birefringence in calcium fluoride and barium fluoride, *Physical Review B*, **64**(24), 241102 (2001).
89. Burnett, J. H., Levine, Z. H., Shirley, E. L., and Bruning, J. H., Symmetry of spatial-dispersion-induced birefringence and its implications for CaF_2 ultraviolet optics, *Journal of Microlithography Microfabrication and Microsystems*, **1**(3), 213 (2002).
90. Kang, H., Bourov, A., and Smith, B. W., Optical lithography at a 126-nm wavelength, *Emerging Lithographic Technologies V, SPIE*, **4343**, 797 (2001).
91. Liberman, V., Rothschild, M., Murphy, P. G., and Palmacci, S. T., Prospects for photolithography at 121 nm, *Journal of Vacuum Science & Technology B*, **20**(6), 2567 (2002).
92. Trintchouk, F., Ishihara, T., Gillespie, W. et al., XLA-300: The fourth-generation ArF MOPA light source for immersion lithography, *Optical Microlithography XIX, SPIE*, **6154**, 615423 (2006).
93. Wong, A. K. K., *Resolution Enhancement Techniques in Optical Lithography*, SPIE Press, Bellingham, WA, (2001).
94. Smayling, M. C., Liu, H. Y., and Cai, L., Low k_1 logic design using gridded design rules, *Design for Manufacturability through Design-Process Integration II, SPIE*, **6925**, 69250B (2008).
95. Borodovsky, Y., Marching to the beat of Moore's Law, *Advances in Resist Technology and Processing XXIII, SPIE*, **6153**, 615301 (2006).
96. Bruning, J. H., Optical lithography: 40 years and holding, *Optical Microlithography XX, SPIE*, **6520**, 652004 (2007).
97. Matsuyama, T., Ohmura, Y., and Williamson, D. M., The lithographic lens: Its history and evolution, *Optical Microlithography XIX, SPIE*, **6154**, 615403 (2006).
98. Kawata, H., Carter, J., Yen, A., and Smith, H. I., Optical projection lithography using lenses with numerical apertures greater than unity, *Microelectronic Engineering*, **9**, 31 (1989).
99. Mulkens, J., Flagello, D., Streefkerk, B., and Graeupner, P., Benefits and limitations of immersion lithography, *Journal of Microlithography Microfabrication and Microsystems*, **3**(1), 104 (2004).
100. Owen, G., Pease, R. F. W., Markle, D. A. et al., 1/8 μm optical lithography, *Journal of Vacuum Science & Technology B*, **10**(6), 3032 (1992).
101. de Klerk, J., Wagner, C., Droste, R. et al., Performance of a 1.35NA ArF immersion lithography system for 40-nm applications, *Optical Microlithography XX, SPIE*, **6520**, 65201Y (2007).
102. Lin, B., Drivers, prospects, and challenges for immersion lithography, *Third International Symposium on* 157 *nm Lithography*, Antwerp, Belgium (2002).
103. Owa, S., Shiraishi, N., Tanaka, I. et al., Nikon F2 exposure tool, *Third International Symposium on* 157 *nm Lithography*, Antwerp, Belgium (2002).
104. Smith, B. W., Kang, H., Bourov, A., and Cropanese, F., Extreme-NA water immersion lithography for 35–65 nm technology, *Third International Symposium on* 157 *nm Lithography*, Antwerp, Belgium (2002).
105. Switkes, M. and Rothschild, M., Resolution enhancement of 157-nm lithography by liquid immersion, *Optical Microlithography XV, SPIE*, **4691**, 459 (2002).
106. Adam, K. and Maurer, W., Polarization effects in immersion lithography, *Journal of Microlithography Microfabrication and Microsystems*, **4**(3), 031106 (2005).
107. Kunz, R. R., Switkes, M., Sinta, R. et al., Transparent fluids for 157-nm immersion lithography, *Journal of Microlithography Microfabrication and Microsystems*, **3**(1), 73 (2004).

108. French, R. H., Qiu, W., Yang, M. K. et al., Second generation fluids for 193 nm immersion lithography, *Optical Microlithography XIX, SPIE*, **6154**, 615415 (2006).
109. Furukawa, T., Hieda, K., Wang, Y. et al., High refractive index fluid for next generation ArF immersion lithography, *Journal of Photopolymer Science and Technology*, **19**(5), 641 (2006).
110. Burnett, J. H., Kaplan, S. G., Shirley, E. L. et al., High-index optical materials for 193 nm immersion lithography, *Optical Microlithography XIX, SPIE*, **6154**, 615418 (2006).
111. Parthier, L., Wehrhan, G., Seifert, F. et al., Development update of high index lens material LuAG for ArF Hyper NA immersion systems, *Fifth International Symposium on Immersion Lithography Extensions*, The Hague, the Netherlands, 22–25 September 2008.
112. Jahromi, S., Bremer, L., Tuinier, R., and Liebregts, S., Development of third generation immersion fluids based on dispersion of nanoparticles, *Fifth International Symposium on Immersion Lithography Extensions*, The Hague, the Netherlands, 22–25 September 2008.
113. Zimmerman, P., Rice, B., Rodriguez, R. et al., High index fluids to enable 1.55 and higher NA 193 nm immersion imaging, *Fifth International Symposium on Immersion Lithography Extensions*, The Hague, the Netherlands, 22–25 September 2008.
114. Zimmerman, P. A., Rice, B., Rodriguez, R. et al., The use of nanocomposite materials for high refractive index immersion lithography, *Journal of Photopolymer Science and Technology*, **21**(5), 621 (2008).
115. Byers, J., Lee, S., Jeri, K. et al., Double exposure materials: Simulation study of feasibility, *Journal of Photopolymer Science and Technology*, **20**(5), 707 (2007).
116. Hori, M., Nagai, T., Nakamura, A. et al., Sub-40-nm half-pitch double patterning with resist freezing process, *Advances in Resist Materials and Processing Technology XXV, SPIE*, **6923**, 69230H (2008).
117. Jung, W. Y., Kim, C. D., Eom, J. D. et al., Patterning with spacer for expanding the resolution limit of current lithography tool, *Design and Process Integration for Microelectronic Manufacturing IV, SPIE*, **6156**, 61561J (2006).
118. Reimer, L., *Transmission Electron Microscopy: Physics of Image Formation and Microanalysis*, 4th edn., Springer-Verlag, Berlin, Germany, (1985).
119. Bolorizadeh, M. and Joy, D. C., Effects of fast secondary electrons to low-voltage electron beam lithography, *Journal of Micro-Nanolithography MEMS and MOEMS*, **6**(2), 023004 (2007).
120. Baylor, L. R., Lowndes, D. H., Simpson, M. L. et al., Digital electrostatic electron-beam array lithography, *Journal of Vacuum Science & Technology B*, **20**(6), 2646 (2002).
121. Utsumi, T., Low-energy e-beam proximity lithography (LEEPL): Is the simplest the best? *Japanese Journal of Applied Physics Part 1-Regular Papers Short Notes & Review Papers*, **38**(12B), 7046 (1999).
122. Utsumi, T., Present status and future prospects of LEEPL, *Microelectronic Engineering*, **83**(4–9), 738 (2006).
123. Bethe, H., The theory of the passage of rapid neutron radiation through matter, *Annalen Der Physik*, **5**(3), 325 (1930).
124. Broers, A. N., Resolution limits of PMMA resist for exposure with 50 kV electrons, *Journal of the Electrochemical Society*, **128**(1), 166 (1981).
125. Tanuma, S., Powell, C. J., and Penn, D. R., Calculations of stopping powers of 100 eV to 30 keV electrons in 10 elemental solids, *Surface and Interface Analysis*, **37**(11), 978 (2005).
126. Han, G., Khan, M., Fang, Y. H., and Cerrina, F., Comprehensive model of electron energy deposition, *Journal of Vacuum Science & Technology B*, **20**(6), 2666 (2002).
127. Jackson, J. D., *Classical Electrodynamics*, 3rd edn., John Wiley & Sons Inc., New York, (1999).
128. Salisbury, I. G., Timsit, R. S., Berger, S. D., and Humphreys, C. J., Nanometer scale electron-beam lithography in inorganic materials, *Applied Physics Letters*, **45**(12), 1289 (1984).
129. Langheinrich, W. and Beneking, H., The resolution of the inorganic electron-beam resist LiF(AlF(3)), *Microelectronic Engineering*, **23**(1–4), 287 (1994).
130. Muray, A., Isaacson, M., and Adesida, I., AiF3—A new very high-resolution electron-beam resist, *Applied Physics Letters*, **45**(5), 589 (1984).
131. Yang, J. K. W. and Berggren, K. K., Using high-contrast salty development of hydrogen silsesquioxane for sub-10-nm half-pitch lithography, *Journal of Vacuum Science & Technology B*, **25**(6), 2025 (2007).
132. Parikh, M. and Kyser, D. F., Energy deposition functions in electron resist films on substrates, *Journal of Applied Physics*, **50**(2), 1104 (1979).
133. Watson, G. P., Berger, S. D., Liddle, J. A. et al., Precise measurement of the effective backscatter coefficient for 100-keV electron-beam lithography on Si, *Journal of Vacuum Science & Technology B*, **13**(6), 2535 (1995).
134. Watson, G. P., Fetter, L. A., and Liddle, J. A., Dose modification proximity effect correction scheme with inherent forward scattering corrections, *Journal of Vacuum Science & Technology B*, **15**(6), 2309 (1997).
135. Owen, G. and Rissman, P., Proximity effect correction for electron-beam lithography by equalization of background dose, *Journal of Applied Physics*, **54**(6), 3573 (1983).
136. Watson, G. P., Berger, S. D., Liddle, J. A., and Waskiewicz, W. K., A background dose proximity effect correction technique for scattering with angular limitation projection electron lithography implemented in hardware, *Journal of Vacuum Science & Technology B*, **13**(6), 2504 (1995).
137. Groves, T. R., Efficiency of electron-beam proximity effect correction, *Journal of Vacuum Science & Technology B*, **11**(6), 2746 (1993).
138. Seo, E. and Kim, O., Dose and shape modification proximity effect correction for forward-scattering range scale features in electron beam lithography, *Japanese Journal of Applied Physics Part 1-Regular Papers Short Notes & Review Papers*, **39**(12B), 6827 (2000).

139. Lee, S. Y., Jacob, J. C., Chen, C. M., Mcmillan, J. A., and Macdonald, N. C., Proximity effect correction in electron-beam lithography—A hierarchical rule-based scheme-pyramid, *Journal of Vacuum Science & Technology B*, **9**(6), 3048 (1991).
140. Patrick, W. and Vettiger, P., Optimization of the proximity parameters for the electron-beam exposure of nanometer gate-length GaAs metal-semiconductor field-effect transistors, *Journal of Vacuum Science & Technology B*, **6**(6), 2037 (1988).
141. Wuest, R., Strasser, P., Jungo, M. et al., An efficient proximity-effect correction method for electron-beam patterning of photonic-crystal devices, *Microelectronic Engineering*, **67–8**, 182 (2003).
142. Jansen, G. H., *Coulomb Interactions in Particle Beams*, Academic Press, New York, (1990).
143. Mkrtchyan, M. M., Liddle, J. A., Berger, S. D. et al., Stochastic scattering in charged particle projection systems: A nearest neighbor approach, *Journal of Applied Physics*, **78**(12), 6888 (1995).
144. Stickel, W., Simulation of Coulomb interactions in electron beam lithography systems—A comparison of theoretical models, *Journal of Vacuum Science & Technology B*, **16**(6), 3211 (1998).
145. Gallatin, G. M. and Liddle, J. A., Analytical model of the "Shot Noise" effect in photoresist, *Microelectronic Engineering*, **46**(1–4), 365 (1999).
146. Gallatin, G. M., Continuum model of shot noise and line edge roughness, *Lithography for Semiconductor Manufacturing II, SPIE*, **4404**, 123 (2001).
147. Kruit, P., Steenbrink, S., and Wieland, M., Predicted effect of shot noise on contact hole dimension in e-beam lithography, *Journal of Vacuum Science & Technology B*, **24**(6), 2931 (2006).
148. Neureuther, A. R., Pease, R. F. W., Yuan, L. et al., Shot noise models for sequential processes and the role of lateral mixing, *Journal of Vacuum Science & Technology B*, **24**(4), 1902 (2006).
149. Smith, H. I., A statistical-analysis of ultraviolet, x-ray, and charged-particle lithographies, *Journal of Vacuum Science & Technology B*, **4**(1), 148 (1986).
150. Ferrera, J., Wong, V. V., Rishton, S. et al., Spatial-phase-locked electron-beam lithography—Initial test-results, *Journal of Vacuum Science & Technology B*, **11**(6), 2342 (1993).
151. Fares, N., Stanton, S., Liddle, J., and Gallatin, G., Analytical-based solutions for SCALPEL wafer heating, *Journal of Vacuum Science & Technology B*, **18**(6), 3115 (2000).
152. Babin, S. and Kuzmin, I. Y., Experimental verification of the TEMPTATION (temperature simulation) software tool, *Journal of Vacuum Science & Technology B*, **16**(6), 3241 (1998).
153. Chao, W. L., Harteneck, B. D., Liddle, J. A., Anderson, E. H., and Attwood, D. T., Soft x-ray microscopy at a spatial resolution better than 15 nm, *Nature*, **435**(7046), 1210 (2005).
154. Sturans, M. A., Hartley, J. G., Pfeiffer, H. C. et al., EL5: One tool for advanced x-ray and chrome on glass mask making, *Journal of Vacuum Science & Technology B*, **16**(6), 3164 (1998).
155. Nakayama, Y., Okazaki, S., Saitou, N., and Wakabayashi, H., Electron-beam cell projection lithography—A new high-throughput electron-beam direct-writing technology using a specially tailored Si aperture, *Journal of Vacuum Science & Technology B*, **8**(6), 1836 (1990).
156. Pfeiffer, H. C., Projection reduction exposure with variable axis immersion lenses (PREVAIL)-A high throughput e-beam projection approach for next generation lithography, *Japanese Journal of Applied Physics Part 1-Regular Papers Short Notes & Review Papers*, **38**(12B), 7022 (1999).
157. Heritage, M. B., Electron-projection microfabrication system, *Journal of Vacuum Science & Technology*, **12**(6), 1135 (1975).
158. http://www.darpa.mil/MTO/Programs/nanowriter/index.html, 2009.
159. Kruit, P., High throughput electron lithography with the multiple aperture pixel by pixel enhancement of resolution concept, *Journal of Vacuum Science & Technology B*, **16**(6), 3177 (1998).
160. Chang, T. H. P., Thomson, M. G. R., Kratschmer, E. et al., Electron-beam microcolumns for lithography and related applications, *Journal of Vacuum Science & Technology B*, **14**(6), 3774 (1996).
161. Groves, T. R. and Kendall, R. A., Distributed, multiple variable shaped electron beam column for high throughput maskless lithography, *Journal of Vacuum Science & Technology B*, **16**(6), 3168 (1998).
162. Pease, R. F. and Chou, S. Y., Lithography and other patterning techniques for future electronics, *Proceedings of IEEE*, **96**(2), 248 (2008).
163. Sreenivasan, S. V., Nanoscale manufacturing enabled by imprint lithography, *MRS Bulletin*, **33**(9), 854 (2008).
164. Barty, A. and Goldberg, K. A., The effects of radiation induced carbon contamination on the performance of an EUV lithographic optic, *Proceedings of SPIE*, **5037**, 450 (2003).
165. Brainard, R. L., Henderson, C., Cobb, J. et al., Comparison of the lithographic properties of positive resists upon exposure to deep- and extreme-ultraviolet radiation, *Journal of Vacuum Science & Technology B*, **17**, 3384 (1999).
166. Brainard, R. L., Barclay, G. G., Anderson, E. H., and Ocola, L. E., Resists for next generation lithography, *Microelectronic Engineering*, **62**(707), 715 (2002).
167. Chou, S. Y., Krauss, P. R., and Renstrom, P. J., Imprint of sub-25 Nm vias and trenches in polymers, *Applied Physics Letters*, **67**(21), 3114 (1995).
168. Chou, S. Y., Krauss, P. R., and Renstrom, P. J., Imprint lithography with 25-nanometer resolution, *Science*, **272**(5258), 85 (1996).

169. Haisma, J., Verheijen, M., vandenHeuvel, K., and van den Berg, J., Mold-assisted nanolithography: A process for reliable pattern replication, *Journal of Vacuum Science & Technology B*, **14**(6), 4124 (1996).
170. Colburn, M., Johnson, S., Stewart, M. et al., Step and flash imprint lithography: A new approach to high-resolution patterning, *Proceedings of SPIE*, **3676**, 379 (1999).
171. Bratton, D., Yang, D., Dai, J. Y., and Ober, C. K., Recent progress in high resolution lithography, *Polymers for Advanced Technologies*, **17**(2), 94 (2006).
172. Gao, H., Tan, H., Zhang, W., Morton, K., and Chou, S. Y., Air cushion press for excellent uniformity, high yield, and fast nanoimprint across a 100 mm field, *Nano Letters*, **6**(11), 2438 (2006).
173. Huang, X. D., Bao, L. R., Cheng, X. et al., Reversal imprinting by transferring polymer from mold to substrate, *Journal of Vacuum Science & Technology B*, **20**(6), 2872 (2002).
174. Schmid, G. M., Stewart, M. D., Wetzel, J. et al., Implementation of an imprint damascene process for interconnect fabrication, *Journal of Vacuum Science & Technology B*, **24**(3), 1283 (2006).
175. Kim, K., Jeong, J., Park, S. et al., Development of a very large-area ultraviolet imprint lithography process, *Microelectronic Engineering*, **86**, 1983 (2009).
176. Ahn, S. H. and Guo, L. J., High-speed roll-to-roll nanoimprint lithography on flexible plastic substrates, *Advanced Materials*, **20**(11), 2044 (2008).
177. Tan, H., Gilbertson, A., and Chou, S. Y., Roller nanoimprint lithography, *Journal of Vacuum Science & Technology B*, **16**(6), 3926 (1998).
178. Ahn, S. H., Kim, J. S., and Guo, L. J., Bilayer metal wire-grid polarizer fabricated by roll-to-roll nanoimprint lithography on flexible plastic substrate, *Journal of Vacuum Science & Technology B*, **25**(6), 2388 (2007).
179. PolyIC. http://www.polyic.com, 2009.
180. Hua, F., Sun, Y. G., Gaur, A. et al., Polymer imprint lithography with molecular-scale resolution, *Nano Letters*, **4**(12), 2467 (2004).
181. Austin, M. D., Zhang, W., Ge, H. X. et al., 6 nm half-pitch lines and 0.04 mu m(2) static random access memory patterns by nanoimprint lithography, *Nanotechnology*, **16**(8), 1058 (2005).
182. Ding, Y. F., Ro, H. W., Germer, T. A. et al., Relaxation behavior of polymer structures fabricated by nanoimprint lithography, *ACS Nano*, **1**(2), 84 (2007).
183. Ding, Y. F., Ro, H. W., Alvine, K. J. et al., Nanoimprint lithography and the role of viscoelasticity in the generation of residual stress in model polystyrene patterns, *Advanced Functional Materials*, **18**(12), 1854 (2008).
184. Jones, R. L., Hu, T. J., Soles, C. L. et al., Real-time shape evolution of nanoimprinted polymer structures during thermal annealing, *Nano Letters*, **6**(8), 1723 (2006).
185. Hirai, Y., Yoshida, S., and Takagi, N., Defect analysis in thermal nanoimprint lithography, *Journal of Vacuum Science & Technology B*, **21**(6), 2765 (2003).
186. Jung, G. Y., Li, Z. Y., Wu, W. et al., Vapor-phase self-assembled monolayer for improved mold release in nanoimprint lithography, *Langmuir*, **21**(4), 1158 (2005).
187. Miller, M., Brooks, C., Lentz, D. et al., Step and flash imprint process integration techniques for photonic crystal patterning: Template replication through wafer patterning irrespective of tone, *Proceedings of SPIE*, **6883**, 68830D (2008).
188. *International Technology Roadmap for Semiconductors*, 2007 edition, http://www.itrs.net, Metrology. SEMATECH, Austin, TX
189. Li, M. T., Chen, L., Zhang, W., and Chou, S. Y., Pattern transfer fidelity of nanoimprint lithography on six-inch wafers, *Nanotechnology*, **14**(1), 33 (2003).
190. Lee, H. J., Ro, H. W., Soles, C. L. et al., Effect of initial resist thickness on residual layer thickness of nanoimprinted structures, *Journal of Vacuum Science & Technology B*, **23**, 3023 (2005).
191. Schulz, H., Wissen, M., and Scheer, H. C., Local mass transport and its effect on global pattern replication during hot embossing, *Microelectronic Engineering*, **67–68**, 657 (2003).
192. Kim, K. D., Jeong, J. H., Sim, Y. S., and Lee, E. S., Minimization of residual layer thickness by using the optimized dispensing method in S-FIL (TM) process, *Microelectronic Engineering*, **83**, 847 (2006).
193. Melliar-Smith, M., Lithography beyond 32 nm—A role for imprint? *Proceedings of SPIE*, **6520**, 652001 (2007).
194. Zhang, W. and Chou, S. Y., Multilevel nanoimprint lithography with submicron alignment over 4 in. Si wafers, *Applied Physics Letters*, **79**(6), 845 (2001).
195. Resnick, D. J., Dauksher, W. J., Mancini, D. et al., Imprint lithography for integrated circuit fabrication, *Journal of Vacuum Science & Technology B*, **21**(6), 2624 (2003).
196. Jung, G. Y., Johnston-Halperin, E., Wu, W. et al., Circuit fabrication at 17 nm half-pitch by nanoimprint lithography, *Nano Letters*, **6**(3), 351 (2006).
197. Hand, A., *Toshiba Validates Imprint Lithography for < 32 nm*, Semiconductor International, October, (2007).
198. Hart, M. W. IBM Almaden Research Center, DARPA Presentation. 2007.
199. Dobisz, E. A., Bandic, Z. Z., Tsai-Wei, W., and Albrecht, T., Patterned media: Nanofabrication challenges of future disk drives, *Proceedings of IEEE*, **96**(11), 1836 (2008).
200. Ahn, S. W., Lee, K. D., Kim, J. S. et al., Fabrication of a 50 nm half-pitch wire grid polarizer using nanoimprint lithography, *Nanotechnology*, **16**(9), 1874 (2005).
201. Yu, Z. N., Deshpande, P., Wu, W., Wang, J., and Chou, S. Y., Reflective polarizer based on a stacked double-layer subwavelength metal grating structure fabricated using nanoimprint lithography, *Applied Physics Letters*, **77**(7), 927 (2000).
202. Raymond, C., Scatterometry for semiconductor metrology, in A. C. Diebold (ed.), *Handbook of Silicon Semiconductor Metrology*, Dekker, New York, p. 477 (2001).
203. SEMI Standard. Test method for evaluation of line edge roughness and line width roughness, SEMI P47-0307.

204. Jones, R. L., Lin, E. K., Lin, Q. et al., Cross-section and critical dimension metrology in dense high aspect ratio patterns with CD-SAXS, *AIP Conference Proceedings*, **788**, 403 (2005).
205. Notte, J., Ward, B., Economou, N. et al., An introduction to the helium ion microscope, *AIP Conference Proceedings*, **931**, 489 (2007).
206. Wang, C. Q., Jones, R. L., Lin, E. K. et al., Characterization of correlated line edge roughness of nanoscale line gratings using small angle x-ray scattering, *Journal of Applied Physics*, **102**(2), 024901 (2007).
207. Wang, C. Q., Jones, R. L., Lin, E. K., Wu, W. L., and Leu, J., Small angle x-ray scattering measurements of lithographic patterns with sidewall roughness from vertical standing waves, *Applied Physics Letters*, **90**(19), 193122-3 (2007).
208. Hu, T. J., Jones, R. L., Wu, W. L. et al., Small angle x-ray scattering metrology for sidewall angle and cross section of nanometer scale line gratings, *Journal of Applied Physics*, **96**(4), 1983 (2004).
209. Jones, R. L., Hu, T., Lin, E. K. et al., Small angle x-ray scattering for sub-100 nm pattern characterization, *Applied Physics Letters*, **83**(19), 4059 (2003).

20

Extreme Ultraviolet Lithography

20.1 Introduction 20-1
20.2 Historical Background..20-2
EUV Imaging System Development • EUVL-Enabling Technologies
20.3 Current State of the Art 20-5
EUVL Exposure Tools • EUV Device Integration
20.4 EUVL Infrastructure Status..20-9
EUV Sources • EUV Masks • EUV Resists • EUV Mask Handling • EUV Optics and Mask Contamination
20.5 Future Perspectives...20-14
Abbreviations 20-16
References..20-16

Obert R. Wood II
GLOBALFOUNDRIES

20.1 Introduction

Extreme ultraviolet (EUV) lithography extends conventional optical lithography to higher resolution because it utilizes an exceedingly short imaging wavelength (λ = 13.5 nm) and provides a larger depth of focus because it employs a small numerical aperture (NA = 0.25) imaging system. Although it has a number of similarities with visible, deep ultraviolet (DUV), and 193 nm lithography, EUV lithography (EUVL) presents several unique technical challenges. For example, because EUV radiation is not transmitted through ambient air, an EUVL tool must be operated in a vacuum environment. Since all solid materials strongly absorb EUV radiation, conventional refractive optics is not an option and reflective reticles and optics must be employed. Since no naturally occurring material provides more than about 1% reflectivity at normal incidence in the EUV spectral region, EUV mirrors must be coated with special multilayer (ML) reflective coatings. Since EUV sources are inefficient, a high-temperature plasma source requiring extremely high input power and careful management of waste heat must be used. Furthermore, because hydrocarbons and water vapor are cracked by EUV radiation, depositing carbon on mirror surfaces and oxidizing their reflective coatings, hydrocarbon and water vapor levels in an EUV exposure tool must be carefully controlled. While the development of EUVL has been and continues to be challenging, the successful insertion of EUVL into semiconductor manufacturing will bring many benefits, among which are a much wider process window and an ability to print linewidths at 22 nm and smaller.

The key components of a system for EUVL include (1) an illumination system consisting of a plasma EUV radiation source, a condenser to collect the light, and an illuminator; (2) a patterned reflective mask; (3) a 4×-reduction reflective imaging system comprised of six or more highly precise ML-coated aspheric mirrors; and (4) an EUV-sensitive resist. A diagram illustrating the key components in an EUVL exposure tool is shown in Figure 20.1.

Although a variety of sources can generate EUV radiation, the source illustrated in Figure 20.1 is a high-temperature plasma of the kind that can be produced by focusing a high-power pulsed laser on a target of lithium, tin, or xenon. EUV radiation reflected from a patterned reticle located on the mask stage is imaged by 4×-reduction projection optics onto a resist-coated silicon wafer located on the wafer stage. EUVL exposure tools use the same step and scan architecture employed in today's DUV and 193 nm exposure tools and will ultimately operate at similar throughputs, approximately 80–100 wafers/h.

The development of EUV imaging systems from the early 1980s to the present is summarized in Section 20.2, which also describes the key technology developments that made EUVL possible: ML reflective coatings, highly precise aspheric mirrors, and phase-shifting point-diffraction interferometers. The current state of the art in EUVL is summarized in Section 20.3. The current status of EUVL infrastructure development in sources, masks, resists, mask handling, and optics and mask contamination is described in Section 20.4. The elements of EUV technology at highest risk of not being ready when needed and the timing of EUV insertion into manufacturing are discussed in Section 20.5, and an extensive list of references is given at the end.

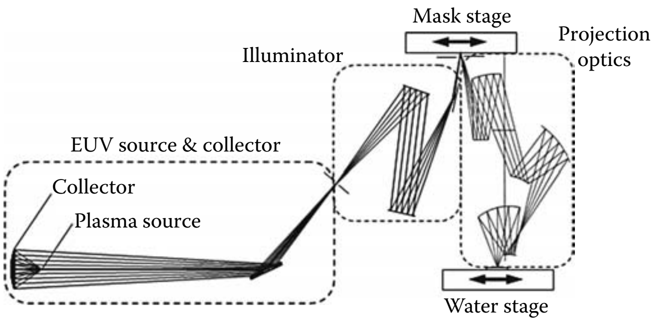
Mask stage
Projection optics
Illuminator
EUV source & collector
Collector
Plasma source
Water stage

FIGURE 20.1 Sketch showing the major subsystems in an exposure tool for extreme ultraviolet lithography.

20.2 Historical Background

20.2.1 EUV Imaging System Development

The most important developments in EUV imaging systems that took place in Europe, Japan, and the United States from 1985 to the present are illustrated in Figure 20.2.

The surface of each optical element in an EUV imaging system must be provided with a ML coating that reflects EUV radiation at near-normal angles of incidence. The use of normal-incidence reflective optics for x-ray microscopes and x-ray telescopes was proposed first in the early 1980s,[1–3] but efficient ML reflective coatings were not sufficiently advanced at that time to be of much use and the results of the early soft x-ray imaging experiments were not very encouraging. The earliest EUV imaging experiments were carried out using projection optics consisting of two spherical mirrors, a convex primary, and a concave secondary, arranged in a Schwarzschild form. In fact, projection optics of this type were used by Hiroo Kinoshita of NTT in 1985 in the very first EUVL imaging experiments. Initially, the imaging system mirrors consisted of tungsten–carbon MLs deposited on SiC substrates.[4] Later, an imaging system with reflective coatings designed to work near the absorption edge of Si was assembled at the High Energy Physics Laboratory in Tsukuba, Japan, but the alignment accuracy of the optics was poor and all of the replicated patterns were severely distorted. In 1984, an image of a 4 μm line and space pattern was finally produced and in 1986, at the Autumn Meeting of the Japan Society of Applied Physics, the results of the early imaging experiments were first made public.[5] Other early experiments that employed two-spherical-mirror projection optics were the demonstration by Bjorkholm et al. of AT&T in 1990 of diffraction-limited imaging near 14 nm wavelength using 20× reduction 0.08 NA Mo/Si ML-coated Schwarzschild optics to print features as small as 0.05 μm in poly (methyl metacrylate) resist,[6] and the development by Tichenor et al. of Sandia National Laboratory (SNL) in 1995 of the first EUVL laboratory tool capable of precise overlay combining 10× reduction, 0.088 NA Schwarzschild optics with accurate mask and wafer stages, and an integrated through-the-lens alignment system.[7]

The second stage in the evolution of EUVL imaging systems involved the development of two-aspherical-mirror projection optics. In 1993, Kinoshita et al. of NTT fabricated a 5× reduction, 0.07 NA two-aspherical-mirror imaging system with a 20 × 0.4 mm exposure area[8] based on a design first described by Kurihara et al.[9] and demonstrated the imaging of features as small as 0.25 μm across a 10 × 0.6 mm area and imaging of 0.15 μm features over a portion of a 2 × 0.6 mm ring-shaped field. In 1996, Haga et al. succeeded in expanding the exposure area of the two-aspherical-mirror imaging system to 20 × 25 mm by

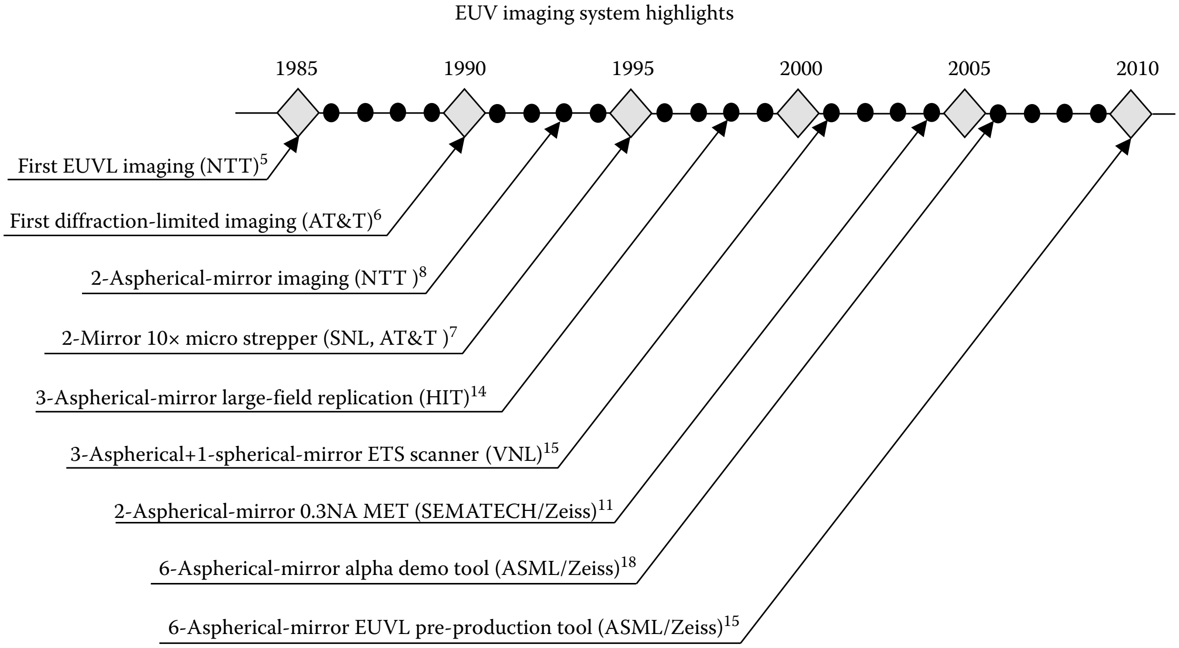
EUV imaging system highlights
1985
1990
1995
2000
2005
2010
First EUVL imaging (NTT)[5]
First diffraction-limited imaging (AT&T)[6]
2-Aspherical-mirror imaging (NTT)[8]
2-Mirror 10× micro strepper (SNL, AT&T)[7]
3-Aspherical-mirror large-field replication (HIT)[14]
3-Aspherical+1-spherical-mirror ETS scanner (VNL)[15]
2-Aspherical-mirror 0.3NA MET (SEMATECH/Zeiss)[11]
6-Aspherical-mirror alpha demo tool (ASML/Zeiss)[18]
6-Aspherical-mirror EUVL pre-production tool (ASML/Zeiss)[15]

FIGURE 20.2 Major developments in EUVL imaging systems from 1985 to present.

utilizing new critical-illumination optics to illuminate a 20 × 0.6 mm ring-shaped area of the mask and by scanning the mask and wafer stages in synchronism.[10] In 2004, Naulleau et al. of Lawrence Berkeley National Laboratory (LBNL) demonstrated the printing capabilities of a 0.3 NA EUV micro-exposure tool (MET) at the Advanced Light Source in Berkeley.[11] The MET's 0.3 NA 5×-reduction two-aspherical-mirror imaging system (with maximum aspheric departure of 3.82 μm on the primary and 5.61 μm on the secondary), manufactured by Carl Zeiss in Germany, had an annular pupil with a central obscuration radius equal to 30% of the full pupil, a 1 × 3 mm field of view at the reticle plane (200 × 600 μm at the wafer plane), and provided ~22.5 nm resolution with conventional illumination (Rayleigh $k_1 = 0.5$) and smaller than 16 nm resolution with dipole illumination (Rayleigh $k_1 = 0.35$).[12]

The third stage in the evolution of EUV imaging systems involved the development of projection optics, which could support the larger exposure fields needed for device demonstrations. In 1996, Murakami et al. of Nikon Corporation described the design of a three-aspherical-mirror ring-field imaging system.[13] In 1998, Kinoshita et al. of Himeji Institute of Technology (Hyogo University today) described a three-aspherical mirror system for EUVL that included illumination optics, a scanning and alignment mechanism, a 0.1 NA 5× reduction imaging system that utilized three aspherical mirrors and one plane mirror, which supported 0.1 μm resolution printing by scanning a 1 mm wide by 30 mm long ring-shaped image field across a wafer, and a load-lock chamber for bringing wafers into and out of vacuum.[14] In 2001, Tichenor et al. of the Virtual National Laboratory, with personnel from LBNL, Lawrence Livermore National Laboratory (LLNL), and SNL, described system integration and performance of the EUV Engineering Test Stand (ETS) developed for the EUVLLC,[15] an Intel-led consortium of semiconductor manufacturers formed in 1997 to accelerate the commercialization of EUVL. The ETS's 0.1 NA 4× reduction projection optics, consisting of three aspherical mirrors (with max aspheric departures less than 10 μm) and one spherical mirror, provided a ±0.5 μm depth of focus (DOF), a scanned field size of 24 × 32.8 mm, and 100 nm resolution imaging. The extensive experimental work on the ETS and the involvement of personnel from six semiconductor companies in the EUVLLC project resulted in a high level of confidence that EUVL was ready for commercial development.

The most recent stage in the evolution of EUV imaging systems includes the development of six-aspherical-mirror ring-field projection optics by each of the three major suppliers of commercial lithographic exposure tools. The 0.25 NA 4× reduction imaging system in ASML's full-field scanning EUV Alpha Demo Tool (ADT), two of which were delivered to customers in 2007, is designed to print 40 nm dense lines and spaces over a 26 × 33 mm field and has recently been used to print 28 nm dense lines and spaces in a state-of-the-art resist.[16] The 0.25 NA 4× reduction six-aspherical-mirror imaging system in Nikon's EUV1 process development tool is designed to print 45 nm dense lines over a 26 × 23 mm field, and a set of EUV1 projection optics has recently demonstrated an extraordinarily low 0.4 nm root mean square (rms) wavefront error over a 26 × 2 mm static exposure field.[17] ASML's Model NXE: 3100 EUV preproduction tool, which is expected to be delivered to customers in 2010, will include a 0.25 NA six-aspherical-mirror imaging system with even lower wavefront error and with reduced imaging system flare.[16]

20.2.2 EUVL-Enabling Technologies

Four scientific and technical breakthroughs critical to the development of EUVL occurred between 1981 and 1996. These four major advances were the first demonstrations of normal-incidence imaging in the soft x-ray spectral region, the development of efficient ML reflective coatings, the fabrication and characterization of precision aspheres, and the improvement of phase-measuring interferometry (PMI).

20.2.2.1 Multilayer Reflective Coatings

Below 30 nm wavelength, no naturally occurring material provides more than about 1% normal incidence reflectance because the index of refraction of all materials in this region is nearly unity (actually slightly less than one). In 1972, Eberhard Spiller of IBM showed that ML films composed of two materials that have nearly identical index of refraction, but that have widely different values of EUV absorption, can lead to constructive interference and, hence, enhanced reflection in a narrow wavelength band peaked at the Bragg wavelength, $\lambda = 2\Lambda \sin(\theta)$, where Λ is the ML period and θ is the angle of incidence.[18] In 1985, Barbee et al.[19] demonstrated that the particular material combination of molybdenum and silicon results in an exceptionally high normal-incidence reflectivity at wavelengths just longer than the Si L absorption edge at 12.4 nm. Today, ML coatings can provide a usable normal incidence reflectance in narrow bands of EUV wavelength from 4.5 to 30 nm.[20] The normal incidence reflectance of the best modern Mo/Si ML, an example of which is shown in Figure 20.3, now approaches 70% in a narrow band of wavelengths near 13.5 nm.[21]

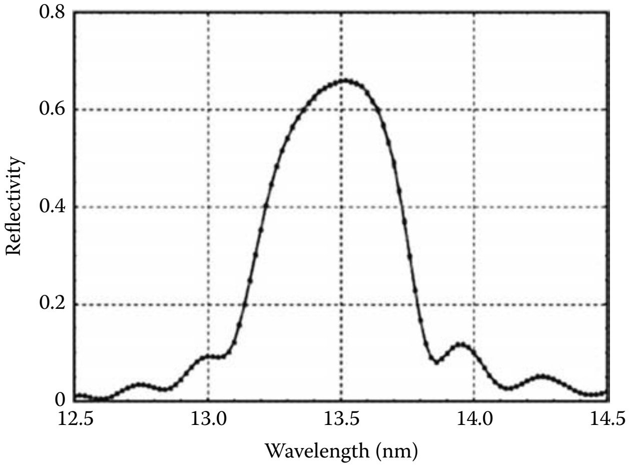

FIGURE 20.3 Reflectance of a Mo/Si ML measured with a commercial EUV reflectometer. (Reprinted from Wood, O., *AIP Conf. Proc.*, 931, 375, 2007. With permission.)

To maximize tool throughput, each of the mirrors in EUVL projection optics must have the highest possible reflectivity at 13.5 nm. To maximize imaging performance, the central wavelength of the reflectivity must be controlled to better than 0.01 nm. To support diffraction-limited printing, the mirror surfaces must have surface figure errors of the order of about 0.1 nm rms, and the ML coatings must not contribute more than approximately one-quarter of the total error, a requirement that leads to strict requirements on the thickness uniformity. The processing conditions for the deposition of Mo/Si ML coatings have now been optimized to obtain high reflectivity, low internal stress, and graded thickness simultaneously, and ML coatings are now available commercially at a number of institutions around the globe. Nikon Corporation recently succeeded in fabricating an aspheric mirror for an EUV imaging system with as little as 60 pm rms surface figure change before and after ML coating (small enough so that its effect on wavefront error can be ignored).[17]

20.2.2.2 Fabrication and Characterization of Aspheric Mirrors

The progress that has taken place in the fabrication and characterization of aspheric mirrors since the early 1990s is impressive. In the beginning, only low-departure aspheres could be fabricated and these were found to be only marginally usable in two-mirror EUV imaging systems. Polishing processes for aspheric mirrors have since been developed that can simultaneously reduce low spatial frequency roughness (LSFR), mid-spatial frequency roughness (MSFR), and high spatial frequency roughness (HSFR). Today, aspheres that meet the stringent requirements of a high-NA six-mirror EUV lithographic imaging system, less than 0.1 nm rms for LSRF, MSFR, and HSFR, are available commercially. Figure 20.4 shows some recent data from Nikon Corporation from a polished aspheric mirror for its EUV1 projection optics on which Nikon achieved an LSFR of 38 pm rms, an MSFR of 80 nm rms, and a HSFR of 68 nm rms.[17]

Surface figure data at low spatial frequencies, obtained from visible light interferometry, are needed to ensure diffraction-limited imaging. Surface finish data at high- and mid-spatial frequencies, obtained with an atomic force microscope and a white-light interferometric microscope, are needed to predict the distribution of scattered radiation in the imaging system. Surface roughness at high-spatial frequencies results in a loss of EUV reflectivity and, hence, system throughput. Surface roughness at mid-spatial frequencies can scatter light into the image plane and result in a loss of image contrast. For this reason, MSFR and the imaging system flare it causes is a serious concern for EUVL even today. A complete set of power spectral density (PSD) surface height data from an asphere for Nikon's EUV1 projection optics is also shown in Figure 20.4.[17] The smooth PSD curve shows that a high degree of precision has been successfully achieved in each of the important spatial wavelength regions.

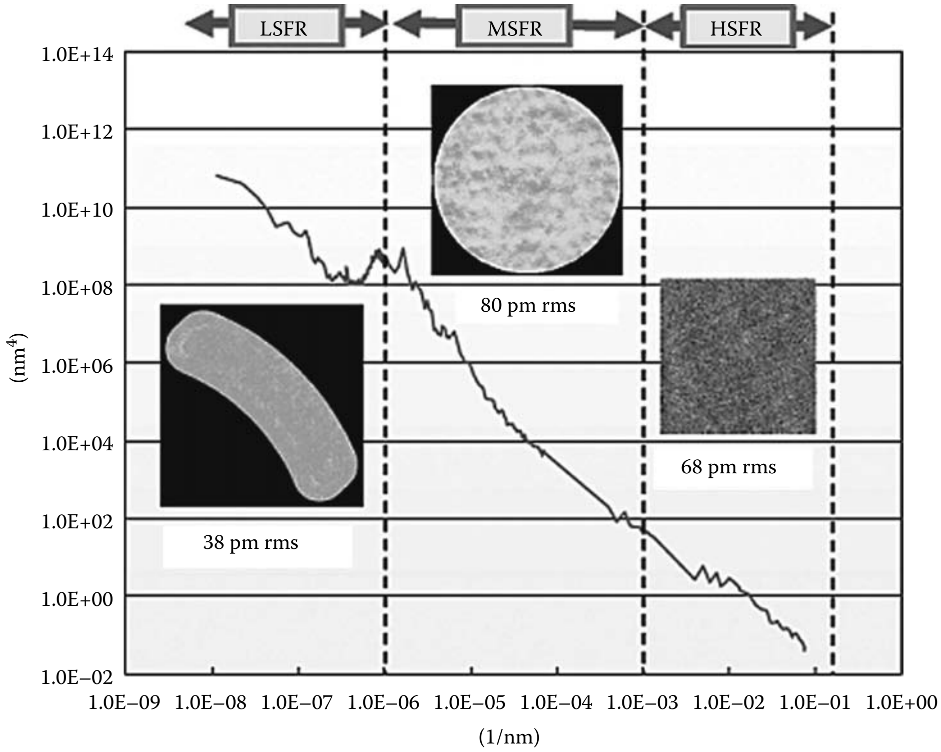

FIGURE 20.4 Plot of surface height data from a polished aspheric mirror for the projection optics of Nikon's EUV1 projection optics expressed as power spectral density (nm^4) versus spatial frequency (nm^{-1}). (Reprinted from Murakami, K. et al., *Proc. SPIE*, 6921, 69210Q, 2008. With permission.)

20.2.2.3 The Phase-Shifting Point- Diffraction Interferometer

To achieve the highest possible resolution (diffraction-limited printing), the wavefront error of an imaging system must be less than λ/14, where λ is the operating wavelength (less than 1 nm when λ = 13.5 nm). For a six-mirror imaging system, this means that the error contributions from each individual mirror can be no larger than about 0.1 nm rms (assuming that the errors are uncorrelated and taking into account the doubling of the error that occurs on reflection). Visible-light phase-measuring interferometry (PMI) is the metrology most commonly used to provide the feedback needed for figuring a mirror surface. Since PMI is a comparative technique, its accuracy is limited by the quality of the reference. Recognizing this, Gary Sommargren (LLNL) in 1996 developed a new type of PMI, called phase-shifting diffraction interferometry (PSDI) that utilized a reference based on diffraction.[22] An experimental arrangement for PSDI configured to measure the surface figure of a concave off-axis aspheric mirror is illustrated in Figure 20.5.

Light leaving the end of a single-mode optical fiber diffracts, forming a spherical wavefront. Part of this wavefront is incident on the mirror being tested and is reflected back toward the fiber. This aberrated wavefront reflects from a semitransparent metal film on the face of the fiber and interferes with part of the original spherical wavefront to produce an interference pattern on a charge-coupled-device camera. Subsequent work has shown that PSDI can result in measurement accuracies approaching 0.10 nm rms.[23]

The total wavefront error of an EUV imaging system depends not only on the surface figure and alignment of each mirror but also on the spatially varying properties of their ML coatings. EUV interferometry, with phase-shifting point-diffraction interferometry (PS/PDI) and various forms of lateral-shearing interferometry, has been used to test and align a number of EUV imaging systems.[24] In fact, the 0.3 NA SEMATECH Berkeley MET was interferometrically aligned at EUV wavelengths using the PS/PDI technique and its wavefront error was shown to be about 0.8 nm rms in a 37-term annular Zernike polynomial dominated by higher order spherical aberration.[25] Because the wavefront error measured with visible light is not exactly the same as that measured with EUV radiation due to the phase change in reflection that occurs at a ML, to ensure that the six-mirror projection optics in future EUV exposure tools have wavefront errors below 1 nm, EUV interferometry has been or is being developed by each of the major suppliers of commercial lithography tools.

20.3 Current State of the Art

EUVL development has accelerated significantly in the last few years. Alpha-class EUV exposure tools are now in operation in Europe, Japan, and the United States. Significant improvements in both line/space and contact-hole performance of EUV-resist materials have been made due in large part to the availability of EUV MET tools that provide high-quality and high-contrast aerial images. The first EUV device integration exercises have been carried out using full-field EUVL exposure tools in Albany, New York and in Leuven, Belgium. This section presents the current state of the art in EUV exposure tools and a brief summary of the results of the device integrations exercise.

20.3.1 EUVL Exposure Tools

Two full-field EUVL research and development exposure tools were delivered by ASML in 2006, one to the College of Nanoscale Science and Engineering in Albany, New York and the other to the IMEC microelectronics research laboratory in Leuven, Belgium.[26] Another full-field EUVL R&D tool was delivered by Nikon Corporation to the Semiconductor Leading

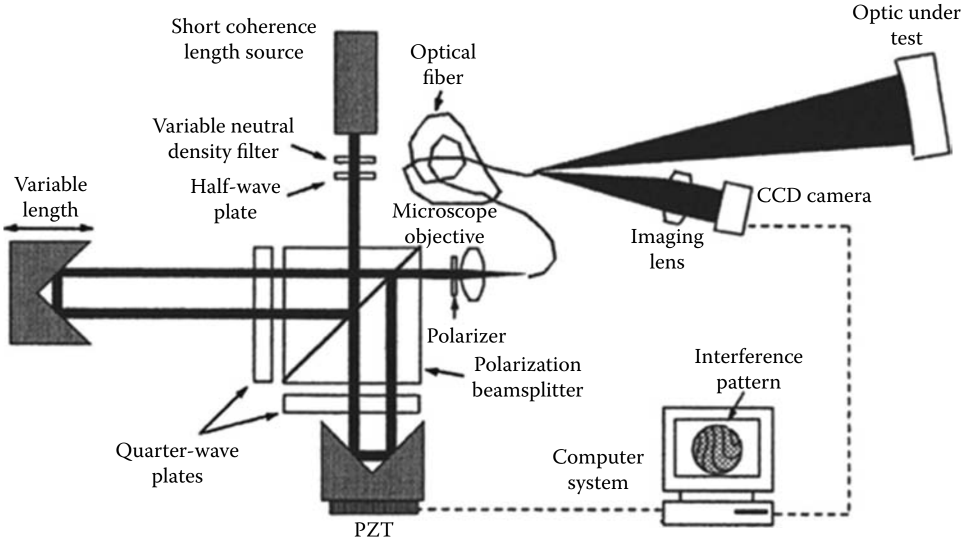

FIGURE 20.5 Experimental arrangement for a phase-shifting diffraction interferometer configured to measure the surface figure of a concave off-axis aspheric mirror. (Reprinted from Sommargren, G., *OSA Trends Opt. Photon.*, 4, 108, 1996. With permission.)

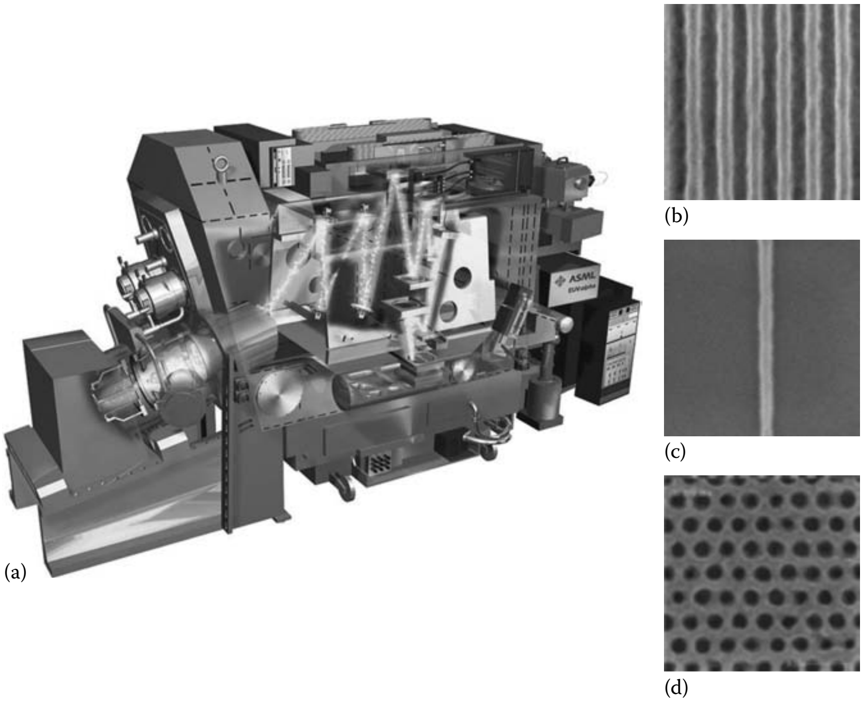

FIGURE 20.6 **(See color insert following page 20-14.)** (a) Artist illustration of ASML's EUV ADT showing a DPP EUV source on the left, illumination optics in the middle, and mask, projection optics, and wafer on the right. (Courtesy of ASML, Veldhoven, the Netherlands.) (b) Image of 28 nm lines and spaces printed in a 50 nm thick layer of a chemically amplified resist using an ASML ADT. (Reprinted from Meiling, H. et al., *Proc. SPIE*, **7271**, 727102, 2009. With permission.) (c) Image of 27 nm isolated line in 80 nm thick EUV-65 resist printed with an ASML ADT. (Reprinted from Pierson, B. et al., EUV resist performance on the ASML ADT and LBNL MET, *International EUVL Symposium*, Lake Tahoe, CA, 2008. With permission.) (d) Image of 28 nm diameter contact holes in 80 nm thick film of a chemically amplified resist printed at 45 mJ/cm^2 dose using an ASML ADT. (Reprinted from Koh, C. et al., *Proc. SPIE*, 7271, 727124, 2009. With permission.)

Edge Technologies (Selete) consortium in Tsukuba, Japan in 2007.[27] An artist's rendering of an ASML EUV ADT is shown in Figure 20.6a.

The EUV source for the ADT is a gas discharge–produced plasma fueled by Sn vapor. The ADT utilizes a reflective illuminator to supply partially coherent illumination ($\sigma = 0.5$) to a 6″ × 6″ × 0.25″ patterned reflective mask. A 4× reduction 0.25 NA reflective imaging system, comprised of six highly precise ML-coated aspheric mirrors, produces an image of the mask on resist-coated 300 mm diameter wafers. The ADT utilizes the same step and scan architecture employed in today's DUV and 193 nm exposure tools, although its throughput is limited to a few wafers per hour by the available source power. The specifications for the ADT include 40 nm resolution for dense lines and spaces, 30 nm for isolated lines, and 55 nm for contact holes, a field size of 26 × 33 mm, an overlay error of 12 nm mean $+3\sigma$, and an imaging system flare of 16%.[28] Nikon's EUV1 exposure tool includes a 4× reduction 0.25 NA six-mirror imaging system with about 10% flare and an illumination system that provides $\sigma = 0.8$ conventional illumination, and was designed for use in process development at the 45 nm half-pitch node. The specifications for the EUV1 include 45 nm resolution for dense lines and spaces and a field size of 26 × 33 mm.[29] Using 0.25 NA EUV ADTs, ASML recently demonstrated 28 nm resolution for dense lines and spaces in a 50 nm thick layer of a state-of-the-art chemically amplified resist (CAR) (illustrated in Figure 20.6b).[15] ASML in joint work with AMD, IBM, and Toshiba demonstrated 27 nm resolution for isolated lines at a DOF of 200 nm (illustrated in Figure 20.6c),[30] and SEMATECH demonstrated 28 nm resolution for dense contact holes in 80 nm thick films of resist at a DOF of 150 nm (illustrated in Figure 20.6d).[31]

Micro-field exposure tools have played and continue to play a key role in the development of EUV resists. This is because the relative simplicity of these tools enables them to provide a higher aerial image contrast than the early alpha-class EUVL exposure tools, since they utilize higher NA projection optics (0.3 NA instead of 0.25 NA) and support more sophisticated illumination modes, and their owners have been more willing to risk contaminating the projection optics by agreeing to print with resist materials that have high outgassing rates. Two MET tools, the Canon 0.3 NA small-field exposure tool at Selete in Tsukuba, Japan[32] and the SEMATECH Berkeley MET at the Advanced Light Source in California[11] use synchrotron radiation as their EUV source. The other two, an Exitech MS-13 at Intel Corporation in Oregon and another at the SEMATECH Resist Test Center in Albany, New York use a Xe discharge–produced plasma source. A CAD model of the SEMATECH Berkeley MET illustrating the major components of the exposure system as well as the EUV beam path[33] is shown in Figure 20.7a.

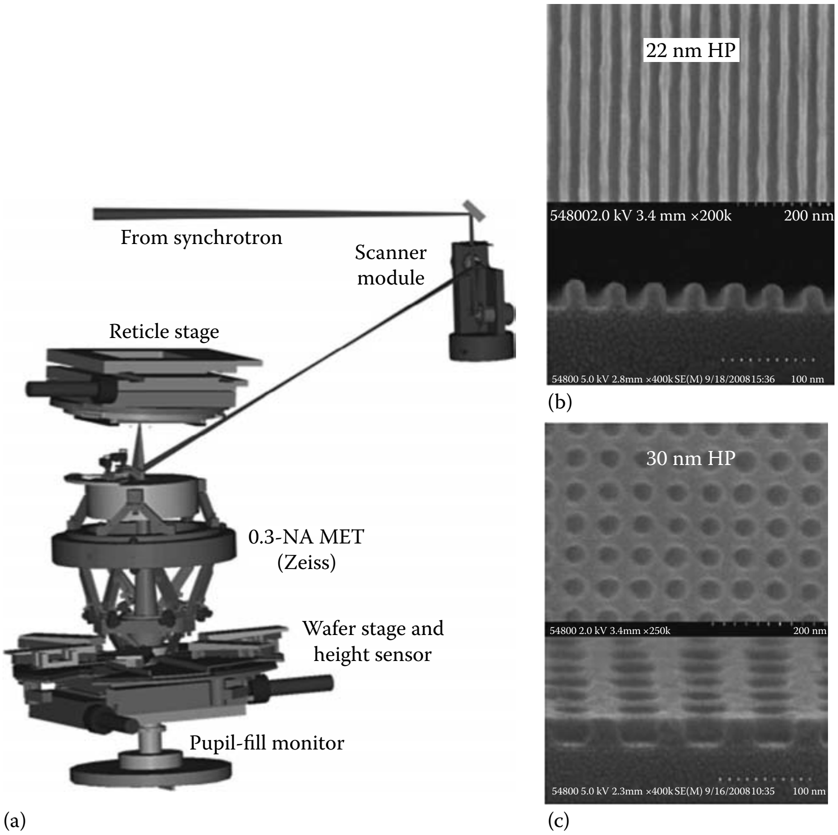
22 nm HP
548002.0 kV 3.4 mm ×200k 200 nm
54800 5.0 kV 2.8mm ×400k SE(M) 9/18/2008 15:36 100 nm
(b)
From synchrotron
Scanner module
Reticle stage
0.3-NA MET (Zeiss)
Wafer stage and height sensor
Pupil-fill monitor
(a)
30 nm HP
54800 2.0 kV 3.4mm ×250k 200 nm
54800 5.0 kV 2.3mm ×400k SE(M) 9/16/2008 10:35 100 nm
(c)

FIGURE 20.7 (a) CAD model of major exposure station components and beam path of the SEMATECH Berkeley MET. (b) Patterning of 1:1 dense lines and spaces in a 50 nm thick film of a chemically amplified resist at a sizing dose of 15.2 mJ/cm^2 using the SEMATECH Berkeley MET using rotated dipole illumination. (c) Patterning of 1:1 dense contact holes in a chemically amplified resist using the SEMATECH Berkeley MET using annular illumination. (Reprinted from Naulleau, P. et al., *Proc. SPIE*, 7271, 72710W, 2009. With permission.)

As mentioned in Section 20.2.1, the 0.3 NA MET optics support a resolution of 22.5 nm at a Rayleigh k_1 factor of 0.5 and a resolution of 16 nm at a Rayleigh k_1 factor of 0.35. Although the SEMATECH Berkeley MET utilizes the same projection optics design as the commercial METs, its unique lossless programmable-coherence illuminator allows it to print at considerably lower Rayleigh k_1 factor (that is, it supports better resolution) than the other MET tools. Another significant benefit of the SEMATECH Berkeley MET is that it is the only MET to have been interferometry-aligned at EUV wavelengths after integration into the tool at the operational location. Impressive resist performance improvements have been demonstrated for both line/space and contact-hole printing during the last year using the SEMATECH Berkeley MET. Figure 20.7b shows the printing of 22 nm half-pitch lines/spaces in a 50 nm thick resist film at about 20 mJ/cm^2 dose using 45° rotated dipole illumination[33]; Figure 20.7c shows the printing of 30 nm diameter contact holes at 48.6 mJ/cm^2 dose using 0.35/0.55 annular illumination.[33]

EUV interference lithography (IL) tools, like the ones at Paul Scherrer Institute (PSI) in Switzerland and at the University of Wisconsin in Madison, are exploring the ultimate resolution limits of resist materials and screening new candidate resist materials for lithographic performance before subjecting the expensive projection optics in the full-field EUV exposure tools to the risk of contamination from resist outgas products. The EUV-IL tool at PSI uses a spatially coherent EUV beam from the XIL beamline at the Swiss Light Source, with a central wavelength of 13.4 nm and a spectral width of about 2%–3%, to illuminate a transmission-diffraction-grating-based interferometer like the one illustrated in Figure 20.8a. The period of the interference fringes formed by this interferometer is equal to half the period of the diffraction gratings, which are patterned on thin silicon nitride membranes using electron beam lithography. Although the image contrast of EUV-IL can be influenced by a number of factors (transverse coherence length of the source, a difference in intensity between the two interfering EUV beams, mechanical instabilities, etc.), EUV-IL offers the advantage of nearly perfect aerial image contrast, performance that is difficult to achieve using projection or proximity lithography methods. EUV-IL at PSI has demonstrated resolution below 12.5 nm half-pitch in several electron beam resist materials. For example, a cross-sectional scanning electron microscope (SEM) image of 16.25 nm half-pitch lines and spaces in a 25 nm thick film of calixarene resist (*p*-chloromethyl-methoxy-calix[4] arene) is shown in Figure 20.8b[34] and a cross-sectional SEM image of 20 nm half-pitch lines and spaces in a 35 nm thick film of HSQ resist (hydrogen silsesquioxane) is shown in Figure 20.8c.[35]

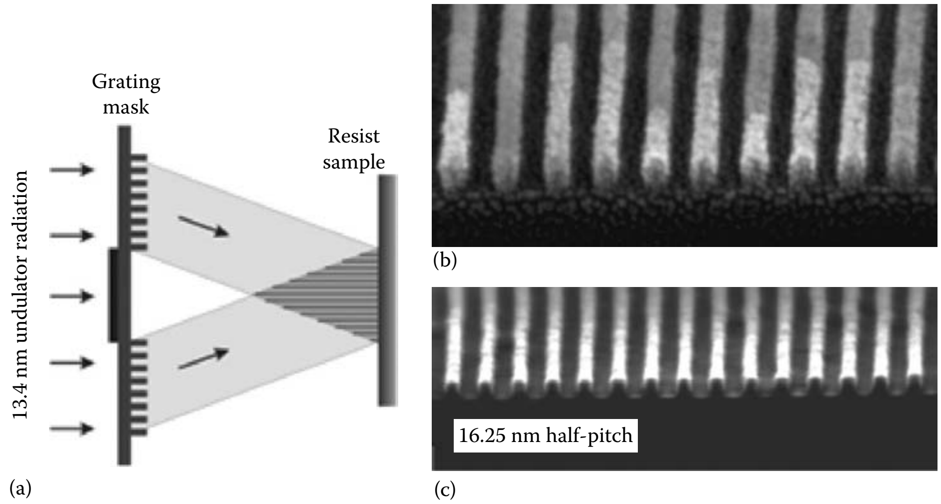

FIGURE 20.8 (a) Sketch of system for EUV interference lithography produced by EUV radiation from two transmission gratings illuminated with EUV radiation from an undulator beamline at a synchrotron light source. (b) Cross-sectional SEM image of HSQ resist patterns with half-pitch of 20 nm produced by the EUV interference lithography at the PSI. (Reprinted from Ekinci, Y. et al., *Microelectron. Eng.*, 84, 700, 2007. With permission from Elsevier.) (c) Cross-sectional SEM image of calixarene patterns with half-pitch of 16.25 nm produced by the EUV interference lithography at PSI. (Reprinted from Solak, H. et al., *J. Vac. Sci. Technol. B*, 25, 91, 2007. With permission.)

20.3.2 EUV Device Integration

Now that the first generation of alpha-class EUVL exposure tools has been deployed and major exposure tool suppliers have established detailed roadmaps for the ramp-up of EUVL technology, one of the main challenges facing the semiconductor industry is to begin integrating EUVL into the semiconductor fabrication process so the technology will be ready to support high-volume manufacturing (HVM) when needed. EUV device integration exercises not only provide a realistic test of technology readiness but also highlight any remaining technical issues. Two EUV device integration exercises were carried out in 2008 and a number of others are currently underway. In 2008, the ASML EUV ADT in Albany was integrated into a standard semiconductor manufacturing flow by the AMD/IBM/Toshiba Alliance and used to pattern the first metal interconnect level of AMD test chips fabricated using 45 nm logic design rules,[36] and the ADT at IMEC in Belgium was used to pattern the contact hole level of a 32 nm

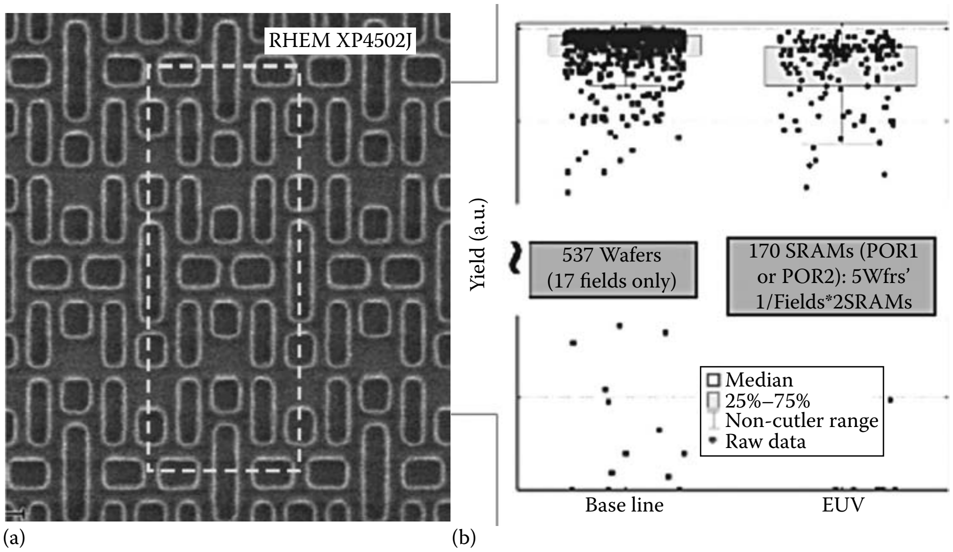

FIGURE 20.9 (a) EUV resist image of an area of the mask containing SRAM patterns. The unit cell is defined by the dashed rectangle. Rohm and Haas resist XP4502J with a film thickness of 150 nm was used. (Reprinted from La Fontaine, B. et al., *Proc. SPIE*, 6921, 69210P, 2008. With permission from SPIE.) (b) Comparison of yield from a baseline 45 nm process and a process using EUVL to print the M1 level. Each point in the baseline statistics represents a full wafer, whereas each point in the EUV statistics represents a 256 kbit SRAM area. (Reprinted from La Fontaine, B. et al., The future of EUV lithography, in Yamamoto, N. (ed), *Proceedings of SEMI Technology Symposium 2008*, SEMI Japan, Chiba, Japan, 2008. With permission.)

six-transistor SRAM cell based on FinFET technology.[37] More recently, EUV Alpha Demo Tools were used by the IBM/AMD/Toshiba Alliance in Albany to pattern two levels of a 22 nm test chip[38] by Samsung to pattern a contract hole level at the 4× DRAM node,[39] and by Hynix to compare the patterning fidelity of 193 nm immersion and EUVL at the 4×, 3×, and 2× DRAM nodes.[40]

Some results from the AMD/IBM/Toshiba Alliance EUV device integration exercise at the 45 nm logic node are shown in Figure 20.9. Figure 20.9a shows an EUVL image of memory structure (SRAM) patterns in a 150 nm thick film of Rohm and Haas XP4502J resist.[41] The SRAM unit cell is defined by the dashed rectangle. Once EUVL of the M1 layer was completed, the patterns were transferred to the underlying dielectric layer using reactive ion etching. The resulting trenches were filled with Cu using a damascene process and then polished. Electrical tests at the M1 level showed that individual transistors connected in this way had electrical characteristics very similar to those patterned using standard 193 nm immersion lithography. After further processing of the 45 nm node test chips to the fifth interconnect level, 256 kbit SRAMs were analyzed electrically for yield. The results in Figure 20.9b demonstrate the potential for EUVL to produce yield in extended devices when used to pattern one level. The yield data shown for the baseline process was calculated on full wafers, whereas the yield data for the EUV case was based on individual 256 kbit SRAM areas on five wafers.[41] Some of the best yielding 26 Mbit SRAM blocks had nearly 100% yield but did contain two failing cells. The location of the failing cells coincided with the location of a hard defect on the mask discovered during subsequent wafer and reticle inspection.

20.4 EUVL Infrastructure Status

Even after more than two decades of development, a number of technical challenges still need to be met before EUVL can be used in HVM. The steering committee of the International EUVL Symposium, with representatives from leading semiconductor manufacturers, EUVL Consortia in Europe, Japan, and the United States, and commercial EUVL exposure tool suppliers, has identified and ranked the list of critical technical issues on a yearly basis since 2002. The current list of critical issues, which was last updated and re-ranked in October 2008, is shown in Table 20.1. Even though this year's list is shorter than ever, the top three issues (focus areas, sources, masks, and resists) have remained at or near the top of each year's critical issue list and must be completely resolved before EUVL can be used in HVM.

TABLE 20.1 2008 EUV Focus Areas

Rank	Key Focus Areas
1	Long-term source operation with 100 W at IF and delivery of 5 MJ/day
2	Defect-free masks through lifecycle and inspection/review infrastructure
3	Resist resolution, sensitivity, and LEF met simultaneously
Areas of Significant Concern	
•	Reticle protection during storage, handling, and use
•	Projection/illuminator optics and mask lifetime

20.4.1 EUV Sources

The wavelength and spectral bandwidth of the source for an EUVL exposure tool, a 2% wavelength band around 13.5 nm, are set by the ML reflective coatings on the imaging system mirrors. The source power is set by the required exposure tool throughput, which to first order is determined by the transmission of the optical train (collector efficiency and illuminator, reticle, and projection optic reflectivities) and the resist sensitivity. Lithography exposure tool manufacturers have agreed on the following top-level specifications for an HVM EUV source: 115 W of collected source power at a plane behind the first source collector optic called the intermediate focus (IF), an etendue limit of 3 mm^2 sr, and a repetition rate of 10 kHz or higher.[42] The cost of ownership (COO) for an EUV source is likely to be determined by the lifetime of the source collector, which in turn will be limited by the effectiveness of the source debris mitigation system. The current EUV source requirements are very severe and significant progress needs to be made to achieve the performance required for HVM while still maintaining an acceptable COO. For the later stages of HVM, EUV sources with up to 400 W of IF power are likely to be needed. All sources under consideration for EUVL are based on hot plasmas created by an electrical discharge or by laser irradiation. The current status of the two leading EUV source technologies, discharge-produced plasma (DPP) and laser-produced plasma (LPP) are described in the next two sections.

20.4.1.1 Discharge-Produced Plasma EUV Sources

The major components of a DPP EUV source-collector module, the DPP source, the debris mitigation system, and the collector optic, are illustrated in Figure 20.10a.

A number of discharge configurations have been evaluated for use in a DPP EUV source. What these discharge configurations have in common is the azimuthal magnetic field generated by the large axial discharge current between the discharge electrodes, which compresses the source fuel (typically Xe gas or Sn vapor) to the hot plasma (about 30 eV electron temperature) required for efficient generation of EUV radiation at 13.5 nm. The debris mitigation system for a DPP EUV source must protect the collector against (1) sputtering by high-energy ions and atoms, (2) contamination by tin fuel, and (3) contamination by electrode material. A reduction in the collector sputter rate by a factor of approximately 50,000 has been demonstrated through improvements in debris mitigation, and the lifetime of the collector in a Sn DPP source today is limited only by contamination. When the limits of active protection of the collector against contamination are reached, *in situ* cleaning will be needed to further prolong the collector lifetime. The collector used with DPP sources usually consists of a number of nested shells (in a so-called Wolter design) that are fabricated using a nickel electroforming process and coated with a thick layer of ruthenium

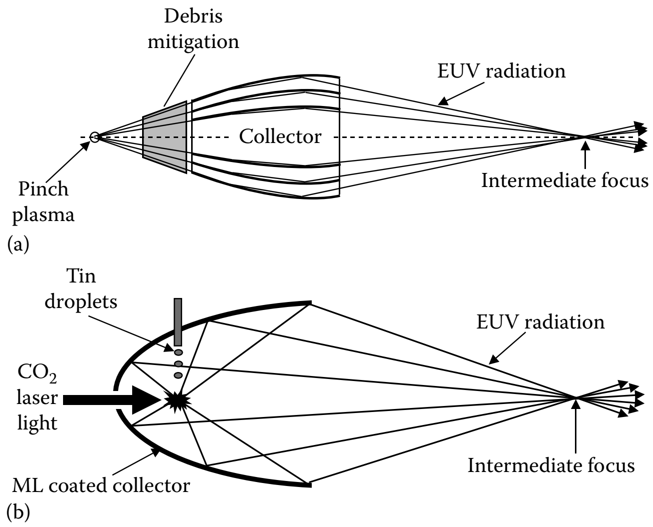

FIGURE 20.10 (a) Schematic diagram of an EUV source-collector module containing an EUV Sn DPP source. (b) Schematic diagram of an EUV source-collector module for a Sn droplet source powered by a carbon dioxide laser.

to provide high grazing-incidence reflectance of EUV radiation. At SPIE Advanced Lithography 2009, Philips Extreme UV in a joint Philips/Xtreme solution reported achieving 8 W of in-band power at IF using 170 W/2πsr Sn DPP EUV source, a foil-trap-based debris mitigation system, and a six-shell Ru-coated grazing-incidence collector.[43]

The advantages of a DPP EUV source are its lower cost of manufacturing, its lower level of complexity, and its use of a lower cost grazing-incidence collector optic. DPP EUV source technology is more mature than LPP EUV source technology and DPP EUV sources are commercially available today. The disadvantages of DPP EUV sources are the large thermal load and short lifetime of the debris mitigation system and collector optic and the difficulty of scaling to higher powers without sacrificing component lifetime and reliability.

20.4.1.2 Laser-Produced Plasma EUV Sources

The major components of an LPP EUV source-collector module, an LPP source consisting of high-power pulsed laser, a droplet target generator, and a ML-coated normal-incidence collector, are shown in Figure 20.10b. In the early 1990s, a number of experimental and theoretical studies showed that LPP EUV sources could achieve conversion efficiencies from laser energy to EUV energy in a 2% bandwidth near a 13 nm wavelength of about 1%–2% depending on target material, laser intensity, laser pulse width, and laser wavelength. It was apparent from the very earliest experiments that debris ejected from the plasma posed a major threat to the lifetime of the collector due to (1) deposition of debris (particulates) from the source hardware, (2) deposition of atomic vapor from the source plasma, and (3) erosion by high-energy ions. Since that time, a great deal of effort has been put into designing and testing a variety of debris-mitigation schemes (gas curtains, electrostatic repeller fields, etc.) for LPP EUV sources. The only alternative to debris mitigation is to minimize debris production using mass-limited targets that supply only the amount of target material needed for efficient production of EUV radiation. One of the most promising methods of producing a mass-limited target is to introduce a stream of tiny droplets of target material into the laser focus, as shown in Figure 20.10b. At SPIE Advanced Lithography 2009, Cymer Inc. reported achieving about 45 W of EUV power at 50 kHz repetition rate in 400 ms-long bursts from a LPP EUV source operating at 50 kHz repetition rate using an approximately 12 kW CO_2 laser to ionize a succession of tiny Sn droplet targets.[44] Cymer has recently been able to accumulate nearly 1 MJ of EUV dose by operating a LPP Sn droplet source over a 3 day period for about 8 h per day.[44]

The advantages of an LPP EUV source are a lower rate of debris production due to the use of a mass-limited target and a larger standoff distance between the hot plasma and the collector optic, a higher radiation-collection efficiency and better thermal management due to the use of a normal-incidence collector, and the possibility of power scaling through spatial, temporal, and spectral multiplexing. The disadvantages of an LPP EUV source are its potentially higher capital cost and greater complexity (a high-power pulsed laser system, a droplet generator, and an ML-coated collector are required) and the relative immaturity of LPP technology (even after many years of development, the measured power at IF from LPP EUV sources has only recently surpassed that from DPP EUV sources).

20.4.2 EUV Masks

The second most critical issue for EUVL is the inability to fabricate a defect-free mask blank and the lack of the infrastructure needed for defect inspection and review. A reflective mask for EUVL consists of buffer and absorber (e.g., Cr or TaN) layers deposited on a ruthenium-capped Mo/Si ML-coated blank. The performance requirements for an EUV mask blank can be found in the SEMI P38-1103 Specification for Absorber Film Stacks and Multilayers on Extreme Ultraviolet Mask Blanks.[45] EUV masks are similar to masks for conventional optical lithography, except they are coated with 40 layer pairs of Mo (2.8 nm thick) and Si (4.1 nm thick) instead of a single layer of chrome, and the 6″ × 6″ × 0.25″ mask substrate is a low thermal expansion material instead of quartz, as shown in Figure 20.11.[46] The performance specifications for EUV mask substrates can be found in the SEMI P27-1102 Specifications for Extreme Ultraviolet Lithography Mask Substrates.[47] A patterned EUV mask, like the one shown in Figure 20.11, was fabricated from an ML-coated blank by selectively etching the absorber and buffer layers.

20.4.2.1 EUV Mask Blanks

Fabrication of a defect-free EUV mask blank is exceedingly difficult. First, it requires an almost perfectly flat (front-side flatness no more than 32 nm peak to valley for the 32 nm technology node) and defect-free substrate with an extremely low coefficient of thermal expansion (0 ± 12 ppb/K in the temperature

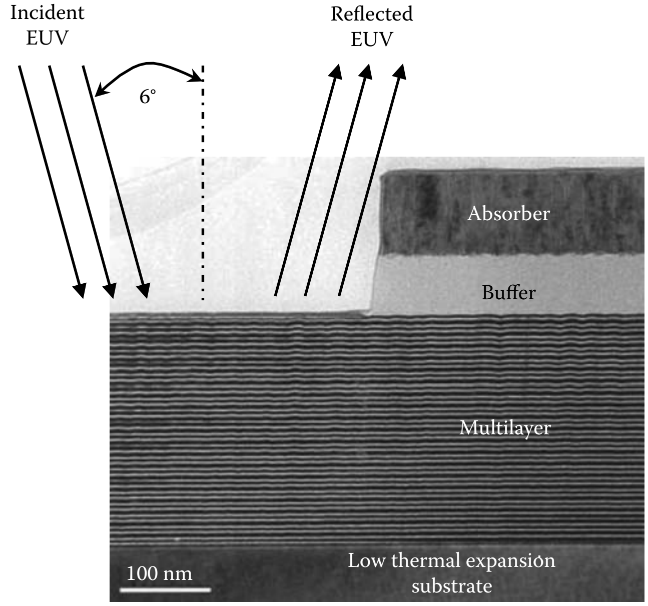

FIGURE 20.11 Cross-sectional TEM image showing EUV reflective mask architecture. (Reprinted from Levinson, H.J., *Principles of Lithography* 2nd edn., SPIE Press, Bellingham, WA, 2005. With permission.)

range from 12°C to 25°C). Then, it requires a coating technology that can deposit Mo/Si MLs with high reflectivity (65%–70%), good thickness uniformity, and nearly zero defect density (the defectivity specification for the 32 nm technology node is a defect density less than 0.003 defects/cm^2 at defect sizes greater than 25 nm). EUV mask blanks meeting the SEMI P38-1103 Class A production specifications for reflectance and uniformity are available today; deposition technologies and processes for zero-added–defect Mo/Si ML deposition are not.

In an attempt to accelerate the development of the zero-defect-adder Mo/Si deposition process needed to produce defect-free mask blanks, SEMATECH established the Mask Blank Development Center (MBDC) in Albany, New York in 2003. During the MBDC ramp-up, SEMATECH, in partnership with Veeco Instruments, successfully transferred technology developed at LLNL specifically for depositing low-defect ML coating on EUV blanks. A record of the MBDC's progress in reducing the defectivity on EUV blanks is shown in Figure 20.12.[48]

The champion plate from a recent deposition run exhibited only eight defects in the quality area of the plate (132 × 132 mm), corresponding to a defect density of 0.04 defects/cm^2 at defect sizes greater than 53 nm (polystyrene latex sphere equivalent size). At the present time, the defect density on the best blanks are dominated by defects that originate from substrate pits and bumps. The rate of progress in blank defect reduction is currently limited by the performance of available defect inspection tools. According to recent data from Intel Corporation, the most sensitive defect inspection tools have a 98% capture rate for 45 nm (SiO_2 sphere equivalent) defects on quartz and for 50 nm defects on MLs.[49] Further progress in defect-free EUVL mask development will require defect inspection tools with sensitivities down to approximately 25 nm. Whether this level of performance can be achieved by incremental improvements to existing tools or will require the development of an entirely new defect inspection tool platform is still an open question.

20.4.2.2 EUV Mask Defectivity

As with chrome-on-quartz masks for optical lithography today, the absorbing pattern on an EUV mask can be inspected using a commercially available DUV mask inspection tool and repaired as needed using focused-ion or electron-beam techniques. While the repair of defects that occur during the absorber patterning process is reasonably straightforward, repair of defects

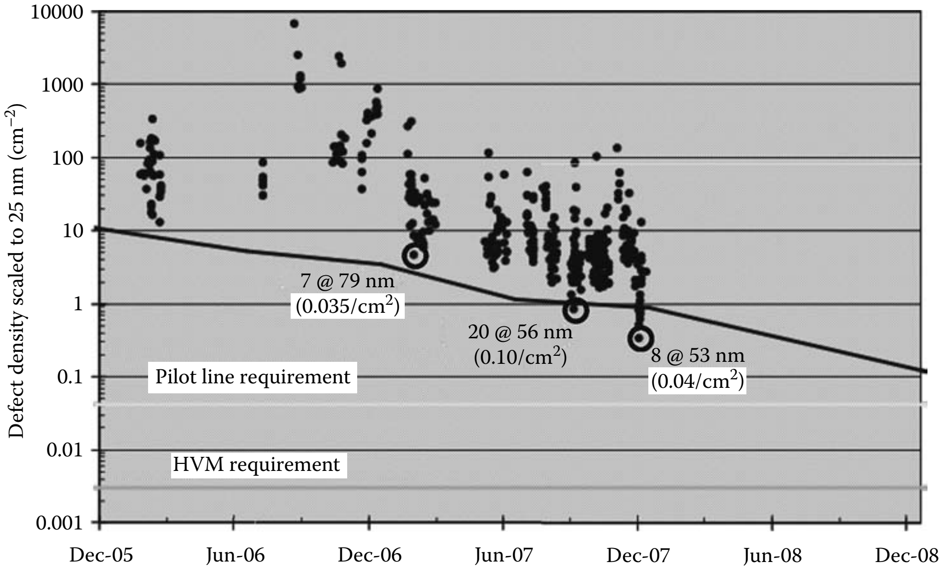

FIGURE 20.12 Record of the defect reduction progress at the SEMATECH Mask Blank Development Center in Albany, New York. (Reprinted from Kearney, P. et al., *Proc. SPIE*, 6921, 69211X, 2008. With permission.)

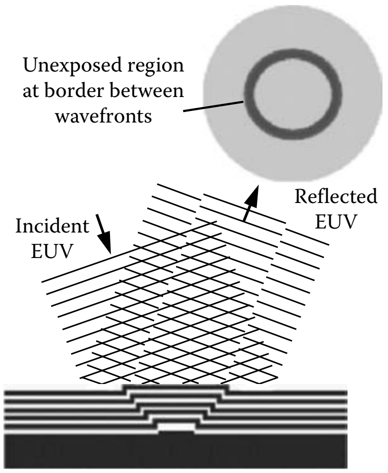

FIGURE 20.13 Illustration of a multilayer phase defect that could result from a growth discontinuity or replicated substrate imperfection, which leads to an abrupt phase shift and, if resolved, a printable defect.

in the ML coating or on the substrate below the reflective coating is problematic. Examples of these "phase defects" were noticed during the course of EUV printing experiments as early as 1992 and were studied systematically by Nguyen et al. in 1994.[50] The illustration in Figure 20.13 shows how a small height difference in a ML coating can lead to cancellation of the reflective waves and an unexpected, high-contrast feature in the aerial image.

Recent results on EUV mask defectivity, determined by printing an EUV mask on resist-coated wafers and inspecting those wafers in a KLA-Tencor 2800 wafer inspection system operating in cell-to-cell mode, showed that only a few defects originated from the mask.[46] The resulting mask blank defect density inferred from these measurements (about 1 defect/cm^2) is approximately 30 times smaller than expected based on mask blank inspection statistics (30–60 defects/cm^2) and masks fabricated from state-of-the-art mask blanks should provide even better defectivity performance.

20.4.3 EUV Resists

The development of an EUV resist that can simultaneously meet the resolution, line-edge roughness (LER), and sensitivity requirements for insertion of EUVL into manufacturing is now ranked by the International EUVL Symposium Technical Steering Committee as the third highest risk to the commercialization of EUVL.

20.4.3.1 Current State of the Art

Because of the limited power available from current EUV sources, state-of-the-art EUV resists must be chemically amplified. The important properties of a chemically amplified resist (CAR) are due mainly to its photo acid generator and its polymer type, which in most cases is the same as those used in resists for 248 and 193 nm lithographies. The chief advantages of a CAR is its high sensitivity and resolution; its drawbacks are high LER, tendency towards pattern collapse, and high EUV light absorption. Pattern collapse depends on the aspect ratio of the resist pattern and, for features less than 30 nm in width, the critical aspect ratio is approximately 2. Assuming aspect ratio scaling continues to hold true, patterning for the 22 nm half-pitch node will require film thicknesses near 40 nm, although the situation can be improved somewhat using surfactant rinses, bilayer resist technologies, and supercritical CO_2 drying. The mechanism responsible for EUV light absorption in EUV resists differs from that in optical resists because the absorption results from the atomic nature of the excitation rather than from the making or breaking of molecular bonds. The high light absorption of all resist material in the EUV spectral region leads to resist patterns with sloping (non-vertical) sidewall profiles that, like pattern collapse, can be improved by using an ultrathin imaging layer over a hard mask in a bilayer or trilayer approach. The resolution performance of state-of-the-art EUV CAR resists printed with the SEMATECH Berkeley MET with 45° rotated dipole illumination is shown in Figure 20.7b, and with 0.35/0.55 annular illumination in Figure 20.7c.[33]

20.4.3.2 The RLS Challenge

The key lithographic performance parameters of CARs, resolution (R), line-edge roughness (L), and sensitivity (S), are highly interdependent—a fact that is often called the RLS challenge or the RLS trade-off. The correlation between resist sensitivity and resolution for a collection of EUV resists evaluated with the SEMATECH Berkeley MET in 2007[51] is shown in Figure 20.14a. Clearly, the materials that exhibit the highest resolution tend to be the least sensitive. The correlation between line edge roughness and sensitivity for a large selection of EUV resists, also tested on the SEMATECH Berkeley MET,[52] is shown in Figure 20.14b. Several attempts to model the interdependences among the performance parameters R, L, and S have been made. Among these, the lithographic uncertainty principle (K_{LUP}) described by van Steenwinckel et al.[53] and the Z-constant proposed by Wallow et al.[51] where [$Z \sim (\text{half-pitch})^3 \cdot \text{LER}^2 \cdot \text{sensitivity}$] have proven to be the most useful. While no single material simultaneously meets resolution, LER, and sensitivity targets for the 22 nm half-pitch node, a number of materials have been developed that meet both the resolution and sensitivity targets. LER remains the single most challenging performance parameter and, at this point, satisfactory values can only be obtained with a high aerial image contrast imaging system and at an unacceptably high exposure dose. The curve in Figure 20.14b is the prediction of a simple model for the shot noise-limited LER. Because of its short exposure wavelength, it is possible that EUVL could have an additional contribution to LER due to shot noise; for the same exposure energy, a EUV image will have a smaller number of photons per pixel. For the most sensitive resist formulations shown in Figure 20.14b, LER tends to be dominated by shot noise, whereas the LER in slower resists appears to be dominated by other effects.

20.4.4 EUV Mask Handling

Because EUV radiation is highly absorbed by all materials, a membrane pellicle for an EUVL mask is not generally considered to be practical and the current strategy for maintaining a defect-free EUV mask is to utilize a removable pellicle to protect the mask

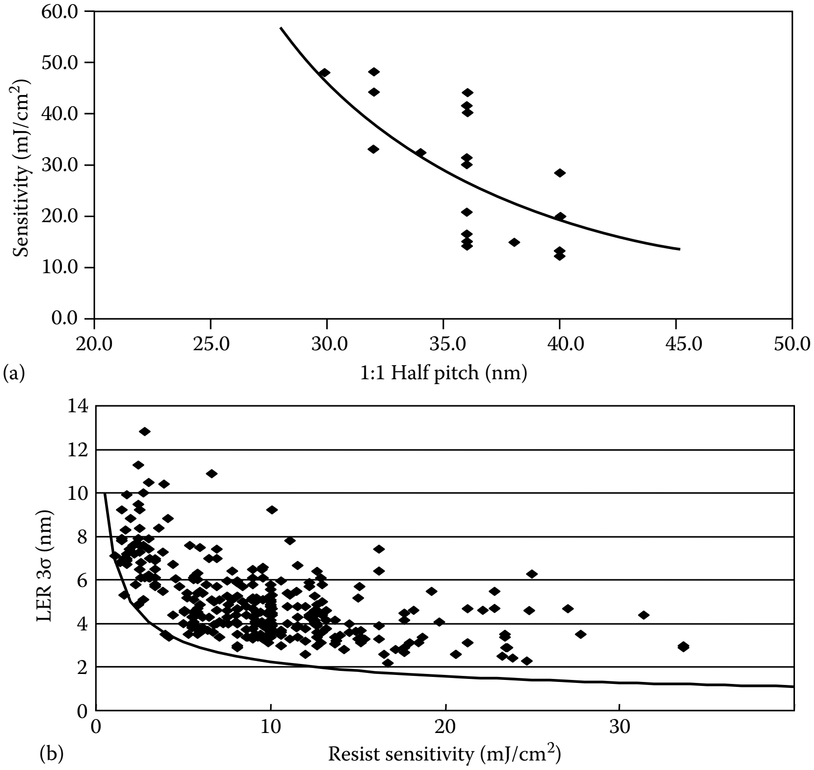

FIGURE 20.14 (a) Observed correlation between resist sensitivity (dose to size at 40 nm half-pitch) and limiting resolution for a collection of EUV resist evaluated using the SEMATECH Berkeley MET. (Reprinted from Wallow, T. et al., *Proc. SPIE*, 6921, 69211F, 2008. With permission from SPIE.) (b) LER and sensitivity for a large selection of EUV resists evaluated using the SEMATECH Berkeley MET. The curve represents a simple model estimate for short-noise-limited LER. (Reprinted from La Fontaine, B. et al., *J. Vac. Sci. Technol. B*, 25, 2089, 2007. With permission.)

during shipping, handling, and storage and to remove the pellicle only when the mask is being inspected or when it is in the vacuum environment of an EUV exposure tool and is ready for exposure. An EUV-removable pellicle can be as simple as a protective, flat face plate that shields the patterned side of the mask from particulate contamination. Variants of the simple face plate are the Fala/Intel Bracket, which adds an additional level of protection by shielding the sides of the mask, the ASML frame, and storage box, which is used as the reticle carrier in their ADTs, and the Nikon, Canon, and Entegris inner pod, which includes a backside lid to form the enclosure shown in the inset of Figure 20.15. The results of four recent tests of a SEMI (draft) 4466-compliant inner pod fabricated by SEMATECH (a so-called S-pod) are shown in Figure 20.15. The data shows, on average, about 0.1 particle added per cycle of use, where cycle is defined as a round-trip shipment, vacuum pump vent, and accelerated storage test at 53 nm PSL equivalent inspection capability.[54]

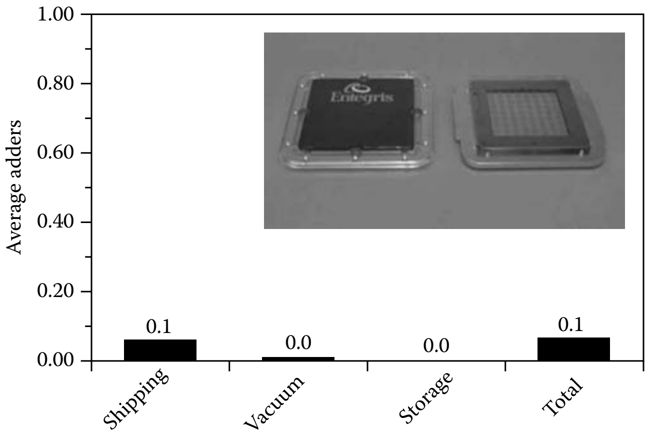

FIGURE 20.15 s-Pod EUV reticle carrier test results for round-trip shipping, vacuum pumping and venting, and accelerated storage. Inset: photograph of reticle pod. (Reprinted from Gomei, Y. et al., *Proc. SPIE*, 6517, 65170W, 2008. With permission from SPIE; He, L. et al., *Proc. SPIE*, 6921, 69211Z, 2008. With permission.)

20.4.5 EUV Optics and Mask Contamination

Another area of significant concern is the shortening of the lifetime of the projection and illuminator optics and the reflective mask due to surface contamination. Water vapor and hydrocarbons are the two major contaminants present in the vacuum environment of an EUVL exposure tool that are known to reduce the lifetime of the reflecting surfaces. Of the two contaminants, carbon contamination is generally considered to be less damaging to optics' lifetimes because carbon films can be removed by hydrogen or oxygen cleaning. A particularly egregious example of carbon contamination is depicted in Figure 20.16,

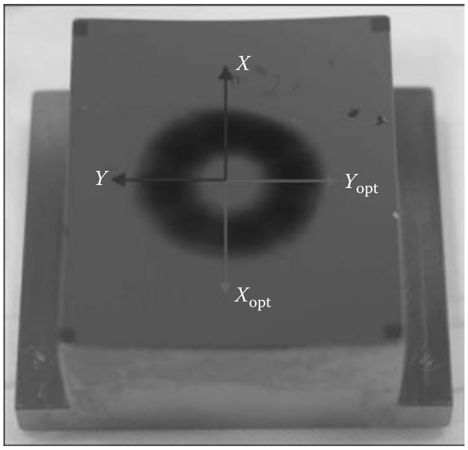

FIGURE 20.16 EUV illuminator mirror showing strong signs of contamination. (Reprinted from La Fontaine, B. et al., *J. Vac. Sci. Technol. B*, 25, 2089, 2007. With permission.)

which shows a photograph of the first illuminator mirror from the SEMATECH EUV MET in Albany with an annular-shaped zone of darkening where the EUV radiation impinged on the mirror surface.

Because an oxide layer strongly absorbs EUV radiation (a 0.3 nm increase in oxide thickness leads to a more than 1% EUV reflectance loss) and is more difficult to remove, oxidation is viewed as a more serious threat to optics' lifetimes. The vacuum environment in an EUVL exposure tool can be modified (by introducing the appropriate gases) to strike a balance between oxidation and carbon growth. This method of contamination control was utilized in 2001 in the EUV Engineering Test Stand[55] (discussed in Section 20.2.1). Oxidation control involving the use of long-life anti-oxidative capping layers comprised of the metal oxides of V, Cr, or W together with carbon-film suppression and removal using EUV + O_2 *in situ* cleaning is a contamination-control strategy currently being explored in Japan.[56] At the 2008 EUVL Symposium, ASML presented data showing that the lifetime of the projection optics in both of its ADTs is at least 4.5×10^9 source pulses (about 1.5 years).[57] For an acceptable COO, the operational lifetime of the imaging system in an EUV exposure tool for HVM must be greater than 5 years (or 30,000 light-on hours). Since the models that have been developed to predict how contamination scales with EUV power density and environmental conditions have not yet reached the level of sophistication needed to accurately predict what the lifetime will be in an HVM tool, the integration of a moderately high-power EUV source with a full-field EUV scanner has become a critical need.

The masks in the EUV METs currently in operation around the world have now received a sufficiently large cumulative dose of EUV radiation that carbon contamination on their surfaces is becoming an issue. Attempts are being made to characterize mask contamination so the lifetime of an EUV mask can be predicted and a safe and effective mask-cleaning process can be developed. Preliminary results from Auger emission spectroscopy of the mask surface show clear signals of carbon and oxygen in the contaminated areas. Preliminary results from a print-based study of a contaminated mask showed that the primary effect of the contamination is an increase in dose to size with increasing contamination.[41] An analysis of wafers printed with the same mask after cleaning show that the original dose to size was recovered with only a slight reduction in DOF. The results also suggest that the increase in dose to size could be used as a metric to trigger a mask clean.

20.5 Future Perspectives

The two EUVL technology elements that are at highest risk of not being ready when needed are source power at intermediate focus (IF) and mask blank defectivity. Source power at IF, which is critically important to exposure tool throughput, currently falls far short of what will be needed for an EUVL pilot production tool. However, according to Cymer Inc.'s LPP source roadmap, a LPP EUV source with beta-level performance will be delivered to ASML in 1Q09 and a LPP EUV source with 100 W at IF will be available before the end of the 2009 calendar year.[57] If Cymer's LPP EUV source delivery schedule slips, ASML's EUV preproduction tool schedule will also slip and—because of this—source power at IF is generally regarded as the technology element that will determine the date of EUVL manufacturing insertion. Mask blank defectivity, which is critically important to the yield of electrically functional devices, also falls far short of what will be needed for HVM, and given the current rate of progress in blank defectivity reduction, it now seems likely that completely defect-free blanks will not be available in the foreseeable future. However, several of the blank-defect workarounds that are now being actively pursued, such as defect compensation and fiducials for blank disposition, could allow functional devices to be manufactured even at current blank defect levels. Most of the other EUVL technology elements are already able to satisfy 15 nm pilot line requirements. Mask handling, optics quality, resist resolution, and sensitivity are already in specification. The only other performance metric that is of concern is resist LER. The current best LER, about 3.0 nm (3σ) at 32 nm HP and 15 mJ/cm^2 dose, is much higher than the International Technology Roadmap for Semiconductors specification for the 15 nm hp node, 1.1 nm (3σ)[31]; however, the use of smoothing under-layers, special rinse liquids, and/or smoothing during etch transfer is expected to help close this gap.

EUVL provides a wider manufacturing process window, has a potentially higher throughput, and is the most easily extendible of all the lithographic technologies that have been evaluated as possible replacements for optical projection lithography. The wider process window and easier extendibility for EUVL can be seen most clearly by rearranging the Rayleigh equation for resolution (RES = $k_1\lambda$/NA, where RES is resolution, λ is wavelength, NA is numerical aperture, and k_1 is a process-dependent constant) to give an expression for k_1 in terms of resolution:

$$k_1 = \left[\frac{\text{NA}}{\lambda}\right]\text{RES} = \left[\frac{\text{NA}}{\lambda}\right]\text{HP} \tag{20.1}$$

where resolution (RES) has been replaced with the period, or half-pitch, of dense line and space patterns. Plots of k_1 versus HP (nm) for a 193 nm immersion exposure tool with a 1.30 NA imaging system and for EUV exposure tools with 0.25, 0.35, and 0.45 NA imaging systems are shown in Figure 20.17.

Also shown in Figure 20.17 are some dates from the roadmaps of the most aggressive semiconductor chip makers for the production ramps at the 32, 22, and 15 nm technology nodes (i.e., at which the logic HP is assumed to undergo an approximate 0.7 shrink). In the not-too-distant past, k_1 values of 0.6 and larger were needed for HVM. More recently, it has been found possible to work at lower values of k_1, at the cost of tighter process control, by employing a variety of resolution enhancement techniques such as off-axis illumination, phase-shifting masks, and application of optical proximity correction to the patterns on the mask. The limiting value for k_1 is 0.25, below which the image contrast for periodic line and space patterns is zero. The calculated values plotted in Figure 20.17 show that EUVL, even with the smallest NA imaging system, will allow a return to manufacturing process windows that seems impossibly large by today's standards. By utilizing EUV imaging systems with larger and larger NAs, the route to higher resolution that optical lithography took in the past, EUVL should be able to print dense features at HP well below 20 nm. In contrast, the process window for single-exposure 193 nm immersion lithography, with water as the immersion fluid, will shrink to zero at an HP approximately equal to 39 nm (solid line at k_1 = 0.25 in Figure 20.17). Even if major advances in high-index immersion fluids, high-index lens materials, and high-index resists are made, SE 193 nm immersion lithography is unlikely to be extended much below 32 nm HP, and double-exposure double-etch (DE^2) 193 nm immersion lithography is unlikely to be extended much below 22 nm HP (dashed line at k_1 = 0.18 in Figure 20.17).

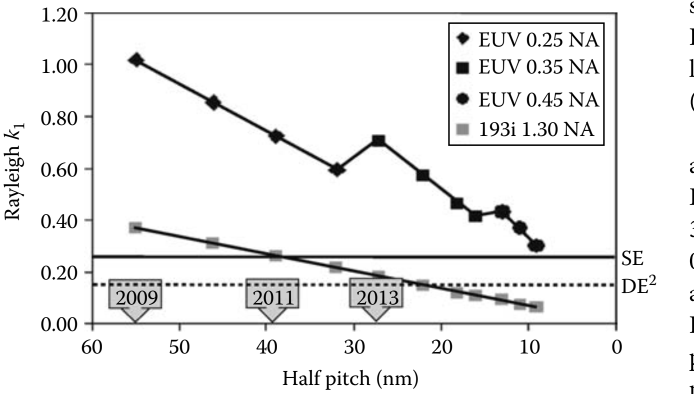

FIGURE 20.17 A plot of the Rayleigh k_1 parameter vs. the half-pitch of dense 1:1 lines and spaces for a 193 nm immersion imaging system with 1.30 NA and 13.5 nm EUV imaging systems with 0.25, 0.35, and 0.45 NA.

Several full-field EUVL pilot production tools will be available in 2011. According to ASML's EUVL tool roadmap,[15] two EUVL exposure tool models will be available in 2011: the NXE 3100 preproduction tool (illustrated in Figure 20.18), which has a 0.25 NA imaging system, will support 4 nm overlay, and provide a throughput of about 60 wafers per hour; and a first-generation HVM tool, which will have a 0.32 NA imaging system, will support 3 nm overlay, and provide a throughput of about 150 wafers per hour. According to Canon's EUV tool development plan,[58] a six-mirror 0.3 NA imaging system will be available in 2011 for in-house testing purposes, but the first of Canon's VS2 HVM EUV exposure tools will not be available until late in 2015. According to Nikon's EUVL tool development scenario,[56] a 0.25 NA EUV2

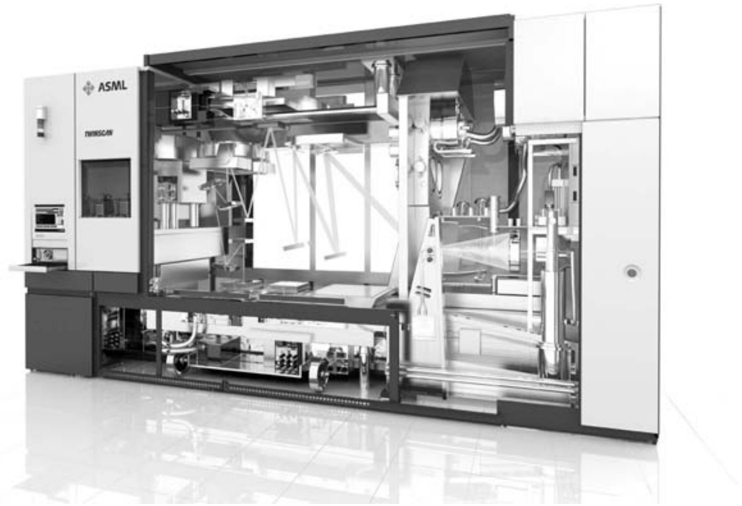

FIGURE 20.18 **(See color insert following page 20-14.)** Artist illustration of ASML NXE: 3100 preproduction EUV exposure tool showing the LPP EUV source on the right, the illuminator optics in the center, and the mask and the projection optics on the right. (Courtesy of ASML, Veldhoven, the Netherlands.)

process development tool with lower wavefront error and lower flare imaging optics will be available at its factory early in 2011 and a greater than 0.3 NA EUV3 HVM EUVL exposure tools will be available early in 2012.

Abbreviations

ADT	Alpha Demo Tool
CAR	chemically amplified resist
COO	cost of ownership
DOF	depth of focus
DPP	discharge-produced plasma
DRAM	dynamic random access memory
DUV	deep ultraviolet
ETS	Engineering Test Stand
EUV	extreme ultraviolet
EUVL	extreme ultraviolet lithography
FET	field-effect transistor
HP	half-pitch
HSFR	high-spatial frequency roughness
HVM	high-volume manufacturing
IF	intermediate focus
IL	interference lithography
LBNL	Lawrence Berkeley National Laboratory
LER	line-edge roughness
LLNL	Lawrence Livermore National Laboratory
LPP	laser-produced plasma
LSRF	low spatial frequency roughness
LSI	lateral-shearing interferometry
MBDC	Mask Blank Development Center
MET	micro-exposure tool
ML	multilayer
MSRF	mid-spatial frequency roughness
NA	numerical aperture
PMI	phase-measuring interferometry
PSD	power spectral density
PSDI	phase-shifting diffraction interferometry
PSI	Paul Scherrer Institute
PSL	polystyrene latex (sphere)
PS/PDI	phase-shifting point-diffraction interferometry
SEM	scanning electron microscope
SNL	Sandia National Laboratory
SRAM	static random access memory
TEM	transmission electron microscope

References

1. Henry, J., Spiller, E., and Weisskopf, M., Imaging performance of a normal incidence x-ray telescope measured at 0.18 KeV, *Proc. SPIE* **316**, 166–168 (1981).
2. Lovas, I., Stanty, W., Spiller, E. et al., Design and assembly of a high resolution Schwarzschild microscope for soft-x-rays, *Proc. SPIE* **316**, 90–97 (1981).
3. Underwood, J.H. and Barbee, T.W. Jr., Soft x-ray imaging with a normal incidence mirror, *Nature* **294**, 429–431 (1981).
4. Takei, K. and Maeda, Y., Preparation of multi-layered tungsten-carbon films by ion beam sputtering, *Jpn. J. Appl. Phys.* **24**, 1366–1367 (1985).
5. Kinoshita, H., Kaneko, T., Takei, H. et al., Study on x-ray reduction projection lithography, *Presentation at the 47th Autumn Meeting of the Japan Society of Applied Physics*, Sapporo, Japan, Paper No. 28-ZF-15 (1986).
6. Bjorkholm, J., Bokor, J., Eichner, L. et al., Reduction imaging using multilayer coated optics: Printing of features smaller than 0.1 microns, *J. Vac. Sci. Technol. B* **8**, 1509–1513 (1990).
7. Tichenor, D.A., Kubiak, G.D., Haney, S.J. et al., Recent results in the development of an integrated EUVL laboratory tool, *Proc. SPIE* **2437**, 292–307 (1995).
8. Kinoshita, H., Kurihara, K., Mizota, T. et al., Large-area, high-resolution pattern replication by the use of a two-aspherical-mirror system, *Appl. Opt.* **32**, 7079–7083 (1993).
9. Kurihara, K., Kinoshita, H., Mizota, T. et al., Two-mirror telecentric optics for soft x-ray reduction lithography, *J. Vac. Sci. Technol. B* **9**, 3189–3192 (1991).
10. Haga, T., Tinone, M., Takenaka, H., and Kinoshita, H., Large-field (>20 × 25 mm^2) replication by EUV lithography, *Microelectron. Eng.* **30**, 179–182 (1996).
11. Naulleau, P., Goldberg, K.A., Anderson, E. et al., Status of EUV microexposure capabilities at the ALS using the 0.3-NA MET optic, *Proc. SPIE* **5374**, 881– 891 (2004).
12. Naulleau, P., Anderson, C., Chiu, J. et al., Advanced extreme ultraviolet resist testing using the SEMATECH Berkeley 0.3-NA micro field exposure tool, *Proc. SPIE* **6921**, 69213N1–69213N11 (2008).
13. Murakami, K., Oshino, T., Shimizu, S. et al., Basic technologies for extreme ultraviolet lithography, *OSA Trends Opt. Photon.* **4**, 16–20 (1996).
14. Kinoshita, H., Watanabe, T., Niibe, M. et al., Three-aspherical mirror system for EUV lithography, *Proc. SPIE* **3331**, 20–31 (1998).
15. Tichenor, D.A., Ray-Chaudhuri, A.K., Replogle, W.C. et al., System integration and performance of the EUV engineering test stand, *Proc. SPIE* **4343**, 19–37 (2001).
16. Meiling, H., Buzing, N., Cummings, K. et al., EUV system—Moving towards production, *Proc. SPIE* **7271**, 7271021–72710215 (2009).
17. Murakami, K., Oshino, T., Kondo, H. et al., Development status of projection optics and illumination optics for EUV1, *Proc. SPIE* **6921**, 69210Q1–69210Q7 (2008).
18. Spiller, E., Multilayer optics for x-rays, in P. Dhez and C. Weisbuch (eds.), *Physics, Fabrication and Applications of Multilayer Structures*, Plenum, New York (1987), pp. 271–309.
19. Barbee, T.W. Jr., Mrowka, S., and Hettrick, M., Molybedenum-silicon multilayer mirrors for the extreme ultraviolet, *Appl. Opt.* **24**, 883–886 (1985).

20. Stearns, D.G., Rosen, R.S., and Vernon, S.P., Multilayer mirror technology for soft-x-ray projection lithography, *Appl. Opt.* **32**, 6952–6960 (1993).
21. Perera, R.C.C. and Underwood, J.H., Results from our recently delivered automated reflectometer for measurement of reflectivity of EUV lithographic masks, *Poster Presentation at the 2006 International EUVL Symposium*, Barcelona, Spain, 15–18 October 2006.
22. Sommargren, G., Phase shifting diffraction interferometry for measuring extreme ultraviolet optics, *OSA Trends Opt. Photon.* **4**, 108–112 (1996).
23. Sommargren, G., Phillion, D., Johnson, M. et al., 100-picometer interferometry for EUVL, *Proc. SPIE* **4688**, 316–328 (2002).
24. Goldberg, K., Naulleau, P., Denham, P. et al., EUV interferometry of the 0.3 NA MET optics, *Proc. SPIE* **5037**, 69–74 (2003).
25. Goldberg, K., Naulleau, P., Denham, P. et al., EUV interferometric testing and alignment of the 0.3 NA MET optic, *Proc. SPIE* **5373**, 64–73 (2004).
26. Harned, N., Goethals, M., Groeneveld, R. et al., EUV lithography with the Alpha Demo Tools: Status and challenges, *Proc. SPIE* **6517**, 6517061–65170612 (2007).
27. Tawarayama, K., Aoyama, H., Magoshi, S. et al., Recent progress of EUV full-field exposure tools in Selete, *Proc. SPIE* **7271**, 7271181–7271188 (2009).
28. Meiling, H., Meyer, H., Banine, V. et al., First performance results of the ASML alpha demo tool, *Proc SPIE* **6151**, 6151081–61510812 (2006).
29. Miura, T., Murakami, K. Suzuki, K. et al., Nikon EUVL development progress update, *Proc. SPIE* **6517**, 6517071–65170710 (2007).
30. Pierson, B., Wallow, T., Mizuno, H. et al., EUV resist performance on the ASML ADT and LBNL MET, *Oral Presentation at the 2008 International EUVL Symposium*, Lake Tahoe, CA, 28 September 2008.
31. Koh, C., Ren, L., Georger, J. et al., Assessment of EUV resist readiness for 32 nm hp manufacturing, and extendibility study of EUV ADT using state-of-the-art resist, *Proc. SPIE* **7271**, 7271241–72712411 (2009).
32. Uzawa, S., Kubo, H., Miwa, Y. et al., Path to the HVM in EUVL through the development and evaluation of the SFET, *Proc. SPIE* **6517**, 6517081–65170811 (2007).
33. Naulleau, P., Anderson, E., Baclea-an, L.-M. et al., The SEMATECH Berkeley microfield exposure tool: Learning at the 22 nm node and beyond, *Proc. SPIE* **7271**, 72710W1–72710W11 (2009).
34. Solak, H.H., Ekinci, Y., Kaser, P., and Park, S., Photon-beam lithography reaches 12.5 nm half-pitch resolution, *J. Vac. Sci. Technol. B* **25**, 91–95 (2007).
35. Ekinci, Y., Solak, H.H., Padeste, C. et al., 20 nm line/space patterns in HSQ fabricated by EUV interference lithography, *Microelectron. Eng.* **84**, 700–704 (2007).
36. La Fontaine, B., Deng, Y., and Kim, R.-H., The use of EUV lithography to produced demonstration devices, *Proc. SPIE* **6921**, 69210P1–69210P10 (2008).
37. Lorusso, G.F., Hermans, J., Goethals, A.M. et al., Imaging performance of the EUV alpha demo tool at IMEC, *Proc. SPIE* **6921**, 69210O1–69210O11 (2008).
38. Wood, O., Koay, C.-S., Petrillo, K. et al., Integration of EUV lithography in the fabrication of 22-nm node devices, *Proc. SPIE* **7271**, 7271041–72710411 (2008).
39. Park, J.-O., Koh, C., Goo, D. et al., The application of EUV lithography for 40 nm node DRAM device and beyond, *Proc. SPIE* **7271**, 7271141–7271149 (2009).
40. Eom, T.-S., Park, S., and Park, J.-T., Comparative study of DRAM cell patterning between ArF immersion and EUV lithography, *Proc. SPIE* **7271**, 7271151–72711511 (2009).
41. La Fontaine, B., Wood, O., and Medeiros, D., The future of EUV lithography, N. Yamamoto (ed.), *Proceedings of SEMI Technology Symposium*, SEMI Japan, Chiba, Japan, 2008.
42. Ota., K., Watanabe, Y., Banine, V., and Frankin, H., EUV source requirements for EUV lithography, in Vivek Bakshi (ed.), *EUV Sources for Lithography*, SPIE Press, Bellingham, WA (2006), pp. 27–43.
43. Corthout, M., Apetz, R., Bruederman, J. et al., Sn DPP source-collector modules: Status of Alpha sources, Beta developments, and HVM experiments, *Proc. SPIE* **7271**, 72710A1–72710A10 (2009).
44. Brandt, D.C., Fomenkov, I.V., Ershov, A.I. et al., LPP source system development for HVM, *Proc. SPIE* **7271**, 727031–72710311 (2009).
45. SEMI P38-1103, Specifications for absorbing film stacks and multilayers on extreme ultraviolet lithography blanks, Semiconductor Equipment and Materials International, San Jose, CA (2003).
46. Levinson, H.J., *Principles of Lithography*, 2nd edn, SPIE Press, Bellingham, WA (2005).
47. SEMI P37-1102, Specifications for extreme ultraviolet lithography mask substrates, Semiconductor Equipment and Materials International, San Jose, CA (2003).
48. Kearney, P., Lin, C.C., Sugiyama, T. et al., Ion beam deposition for defect-free EUVL mask blanks, *Proc. SPIE* **6921**, 69211X1–69211X7 (2008).
49. Ma, A., Liang, T., Park, S.-J. et al., EUVL blank defect inspection capability at Intel, *2008 International EUVL Symposium*, Lake Tahoe, CA, September 29, 2008.
50. Nguyen, K., Attwood, D., Mizota, T. et al., Imaging of EUV lithography masks with programmed defects, *OSA Proc. Extreme Ultraviolet Lithography* **23**, 193–203 (1994).
51. Wallow, T., Higgins, C., Brainard, R. et al., Evaluation of EUV resist materials for use at the 32 nm half-pitch node, *Proc. SPIE* **6921**, 69211F1–69211F11 (2008).
52. La Fontaine, B., Deng, Y., Kim, R.-H. et al., Extreme ultraviolet lithography: From research to manufacturing, *J. Vac. Sci. Technol. B* **25**, 2089–2093 (2007).

53. Van Steenwinckel, D., Gronheid, T., Lammers, J.H. et al., A novel method for characterizing resist performance, *Proc. SPIE* **6519**, 65190V1–65190V12 (2007).
54. He, L., Wurm, S., Seidel, P. et al., Status of EUV reticle handling solutions in meeting 32 nm HP EUV lithography, *Proc. SPIE* **6921**, 69211Z1–69211Z10 (2008).
55. Malinowski, M., Grunow, P., Steinhaus, C. et al., Use of molecular oxygen to reduce EUV-induced carbon contamination of optics, *Proc. SPIE* **4343**, 347–356 (2001).
56. Miura, T., Murakami, K., Kawai, H. et al., Nikon EUVL development progress update, *Proc. SPIE* **7271**, 72711X1–72711X11 (2009).
57. Meiling, H., Lok, B., Hultermans, B. et al., EUV alpha demo tools—Stepping stones towards volume production, *2008 International EUVL Symposium*, Lake Tahoe, CA, 28 September 2008.
58. Hasegawa, T., Uzawa, S., Honda, T. et al., Development status of Canon's full-field EUVL tool, *Proc. SPIE* **7271**, 72711Y1–72711Y11 (2009).

IV

Optics of Nanomaterials

21 **Cathodoluminescence of Nanomaterials** *Naoki Yamamoto* **21**-1
Introduction • Fundamentals of Cathodoluminescence • Optical Properties of Semiconductor Quantum Structures • Quantum Structures • Application of TEM-CL • Acknowledgment • References

22 **Optical Spectroscopy of Nanomaterials** *Yoshihiko Kanemitsu* **22**-1
Introduction • Carbon Nanotubes • Nanoparticle Quantum Dots • Multiexciton Generation • Summary • Acknowledgments • References

23 **Nanoscale Excitons and Semiconductor Quantum Dots** *Vanessa M. Huxter, Jun He, and Gregory D. Scholes* **23**-1
Introduction to Nanoscale Excitons • Nonlinear Optical Properties of Semiconductor Quantum Dots • Two-Photon Absorption in Semiconductor Quantum Dots • Conclusion • References

24 **Optical Properties of Metal Clusters and Nanoparticles** *Emmanuel Cottancin, Michel Broyer, Jean Lermé, and Michel Pellarin* **24**-1
Introduction • Theoretical Description • Experimental Results: State of the Art • Conclusion and Outlooks • Acknowledgments • References

25 **Photoluminescence from Silicon Nanostructures** *Amir Sa'ar* **25**-1
Introduction • Synthesis of Silicon Nanostructures • Luminescence Properties of Silicon Nanostructures • The Vibron Model: The Relationship between Surface Polar Vibrations and Nonradiative Processes • Concluding Remarks • Acknowledgments • References

26 **Polarization-Sensitive Nanowire and Nanorod Optics** *Harry E. Ruda and Alexander Shik* **26**-1
Introduction • Single Nanowires and Nanorods • Arrays of Nanorods • Conclusion • Acknowledgments • References

27 **Nonlinear Optics with Clusters** *Sabyasachi Sen and Swapan Chakrabarti* **27**-1
General Introduction: Fundamentals of Nonlinear Optics and Different Nonlinear Optical Effects • Quantum Formulation of Nonlinear Optics • Nonlinear Optics with Materials • Why Are Clusters So Important? • Review of Nonlinear Optics with Clusters • Conclusions • References

28 **Second-Harmonic Generation in Metal Nanostructures** *Marco Finazzi, Giulio Cerullo, and Lamberto Duò* **28**-1
Introduction • Nonlinear Optics and Second Harmonic Generation • Second Harmonic Generation in Nanosystems • Leading Order Contributions to the Second Harmonic Generation Process in Nanoparticles • Emission Patterns and Light Polarization • Allowed and Forbidden Second Harmonic Emission Modes: Examples • Second Harmonic Generation in Single Gold Nanoparticles • Perspectives • Conclusions • Acknowledgments • References

29 **Nonlinear Optics in Semiconductor Nanostructures** *Mikhail Erementchouk and Michael N. Leuenberger* **29**-1
Introduction • Semiconductor-Field Hamiltonian in $\mathbf{k}\cdot\mathbf{p}$-Approximation • Quantum Equations of Motion • Perturbation Theory: Four-Wave Mixing Response • Nonperturbative Methods: Rabi Oscillations • Conclusion • Acknowledgment • References

30 **Light Scattering from Nanofibers** *Vladimir G. Bordo* **30**-1
Introduction • Background • Basic Theory • Some Numerical Results • Experimental Implementation • Discussion • Summary and Outlook • Acknowledgments • References

31 **Biomimetics: Photonic Nanostructures** *Andrew R. Parker* **31**-1
Introduction • Engineering of Antireflectors • Engineering of Iridescent Devices • Cell Culture • Diatoms and Coccolithophores • Iridoviruses • The Mechanisms of Natural Engineering and Future Research • Acknowledgments • References

21

Cathodoluminescence of Nanomaterials

21.1 Introduction **21**-1
21.2 Fundamentals of Cathodoluminescence **21**-2
Radiative Recombination • Exciton • TEM-Cathodoluminescence Technique
21.3 Optical Properties of Semiconductor Quantum Structures **21**-5
Band Structure of Semiconductor • Theory of Light Emission • Polarization
21.4 Quantum Structures **21**-8
Quantum Well • Quantum Wire
21.5 Application of TEM-CL **21**-15
InP Nanowires • GaAs Nanowire • ZnO Nanowires
Acknowledgment **21**-24
References **21**-24

Naoki Yamamoto
Tokyo Institute of Technology

21.1 Introduction

Dimensionality and size are two factors that introduce new properties into semiconductor materials and enable us to provide new functionality in semiconductor nanoelectrics and photonics. Semiconductor quantum wires have been intensively studied, on one hand, for the theoretical interest in physical properties due to quasi-one-dimensional quantum confinement of an electronic system (Sakaki, 1980; Ogawa and Takagahara, 1991; Tanatar et al., 1998) and, on the other hand, for practical application to future optoelectronic devices such as low-threshold lasers (Arakawa and Sakaki, 1982) and polarization-sensitive devices (Wang et al. 2001). One-dimensional nanostructures have been primarily synthesized by lithography and an epitaxial technique on semiconductor substrates (Arakawa and Sakaki, 1982; Ils et al., 1994; Someya et al., 1995). V- (Kapon et al., 1989) and T-shaped quantum wires (Pfeiffer et al., 1990) were fabricated using epitaxial growth techniques (Wang and Voliotis, 2006), namely, molecular beam epitaxy (MBE) and metalorganic vapor phase epitaxy (MOVPE). However, excitons in the quantum wires were observed to be localized in monolayer step-induced islands at low temperatures (Guilet et al., 2003). The fluctuation of confinement potential along a quantum wire is a serious issue for realizing the characteristic property of a one-dimensional structure.

The free-standing nanowires of many semiconductors are recently produced by the growth technique using nanosized liquid droplets of the metal solvent (Hiruma et al., 1995; Gudiksen et al., 2002; Wu et al., 2002). This technique has an advantage to produce heterogeneous structure along the wire such as p–n junctions (Gudiksen et al., 2002) and superlattices (Wu et al., 2002). The optical properties of the nanowires have been studied by photoluminescence (PL) and absorption spectroscopy, which showed the blue shift of the peak energy due to the quantum confinement effect and giant polarization anisotropy in PL and optical absorption. Those properties will provide a potentiality for new application of the semiconductor nanowires. In spite of many works for synthesizing nanowires, there have been very few experimental works for studying the optical properties of self-standing nanowires, especially for a single nanowire, because of the limitation in spatial resolution. An isolated single nanowire of InP was studied by PL (Wang et al., 2001), which showed giant polarization anisotropy in PL and photoconductivity measurement. The giant polarization anisotropy is partly due to the excitation process by light, because the electromagnetic field in the nanowire, which generates carriers, depends on the polarization direction of incident light. Polarization anisotropy can also be expected in the light emission process, as in the case of electroluminescence (EL) and cathodoluminescence (CL).

We studied the optical properties of semiconductor nanowires, observing cathodoluminescence (CL) spectra and polarized monochromatic CL images of nanowires by using a transmission electron microscope (TEM) combined with a CL detection system (TEM-CL). The TEM-CL technique has an advantage in high spatial resolution and ability to characterize nanoscale structures (Yamamoto, 1984; Mitsui et al., 1996; Yamamoto et al., 2003, 2008; Ino and Yamamoto, 2008). In Section 21.2, a principle of TEM-CL technique is described based on the fundamental knowledge of radiative recombination processes and property of exciton. Luminescence in semiconductor nanowires shows characteristic properties in emission energy and polarization, because the optical transition between electronic states is affected by the quantum confinement effect. For understanding these properties, the theory of the band structure in semiconductors and optical transitions are described in Section 21.3. Quantum confinement effect on the optical properties of semiconductor nanowires and dielectric effect on polarization properties are described in Section 21.4. Finally, the experimental results of the TEM-CL technique for the InP and GaAs nanowires, which show the quantum confinement effect, and for the ZnO nanowires, which show the dielectric effect, are described in Section 21.5.

21.2 Fundamentals of Cathodoluminescence

21.2.1 Radiative Recombination

Luminescence in semiconductors is light emission phenomenon due to the recombination of excited electrons in the conduction band with holes in the valence band (Voos et al., 1980). It is called photoluminescence (PL) when those electrons and holes, the carriers, are excited by light, and called cathodoluminescence (CL) when they are excited by fast electrons. The band structures are classified into two types, direct gap type and indirect gap type. In the former type, the conduction band minimum is located at the Γ point ($k = 0$) in the k-space, whereas it is shifted from the Γ point in the latter type. The valence band maximum is located at the Γ point. Then in the case of the direct gap band structure, the radiative interband transition from the electron states in the conduction band to the hole states in the valence band occur at the Γ point accompanying with emission of photon at high recombination probability. GaAs and InP have band structure of this type. Whereas in the indirect gap type, the transition occurs with the association of phonon, and the recombination rate is rather small. Si and GaP have band structure of this type.

Typical recombination processes in semiconductors are depicted in Figure 21.1 (Pankove, 1971). First group is radiative recombination processes, i.e., (a) interband transition from conduction band to valence band, (b)–(d) annihilation of free exciton (FX) composed of a pair of electron and hole, and related processes of bound excitons (BX), (e)–(g) transitions of carriers trapped by impurity states such as donor and acceptor states. Second group is nonradiative recombinations, i.e., (h) multiphonon scattering, and (i) Auger process.

Photon energies in the recombination processes in Figure 21.1 are slightly lower than the band gap energy E_g, though their values are different from each other. The emission intensity by the interband transition from conduction band to valence band is generally small compared with the other processes. Photon energies in the impurity-associated recombination processes (Figure 21.1e and f) are given as follows:

$$(D,h): E = E_g - E_D \tag{21.1}$$

$$(e,A): E = E_g - E_A, \tag{21.2}$$

where E_D and E_A are binding energies of the donor level and acceptor level, respectively. Photon energy associated with the donor–acceptor pair transition (D,A) is given by

$$(D,A): E = E_g - E_D - E_A + \frac{e^2}{4\pi\varepsilon R}, \tag{21.3}$$

where

R is a distance between donor and acceptor
ε is the dielectric constant of the material
e is the elemental charge

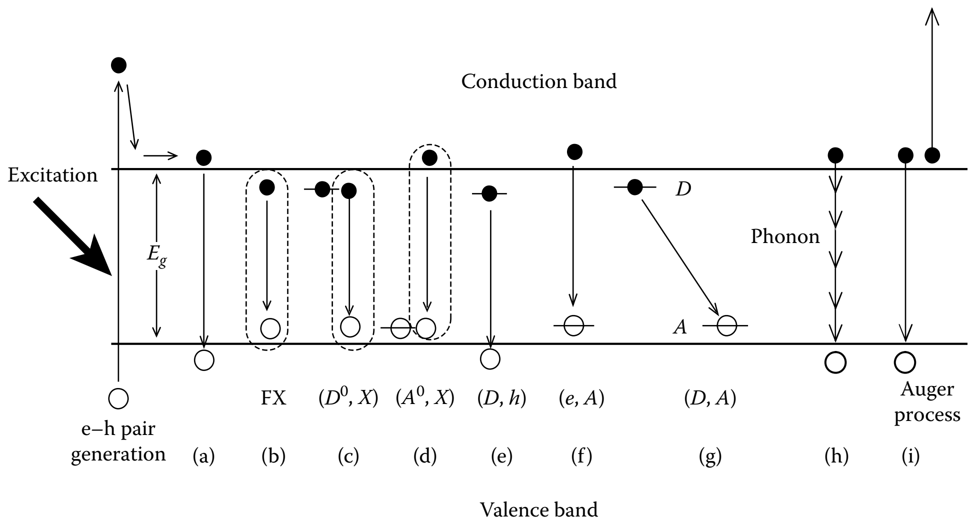

FIGURE 21.1 Recombination processes in semiconductors.

TABLE 21.1 Basic Parameters on Exciton in Typical Semiconductors

			Band Gap (eV)				
	Structure	Band Type	E_g (4 K)	E_g (300 K)	Dielectric Constant	E_x (meV)	FX Radius (nm)
Si	Diamond	Indirect	1.1698	1.11	11.4	14.7	403
Ge	Diamond	Indirect	0.7454	0.669	15.36	4.15	11
GaAs	Zinc-blend	Direct	1.519	1.43	13.1	4.21	13
InP	Zinc-blend	Direct	1.4236	1.34	12.61	5.12	11
GaN	Wurtzite	Direct	3.507	3.39	10.4 ($\|\|_c$) 9.5 ($\perp_c$)	23	3
ZnO	Wurtzite	Direct	3.436	3.37	8.84 ($\|\|_c$) 8.47 ($\perp_c$)	59	1.4

Note: The data were gathered from many literatures. The values of FX radius are calculated from the data in the other columns.

The last term indicates the Coulomb energy of the electron and hole pair. The impurities locate at lattice sites and R takes many different values. Then the emission peak due to (D, A) is composed of the multiple peaks to become a single broad peak, and dependence of the peak position on temperature and excitation rate is different from that of the other luminescence peaks of a single recombination process.

21.2.2 Exciton

An electron and a hole in semiconductors attract each other to form a FX. FX emission dominates in the radiative recombination process at low temperatures. The binding energy of exciton can be obtained by solving Schrödinger equation for electron and hole system attracted by Coulomb potential (Basu, 1997; Yu and Cardona, 1999). The binding energy of the lowest energy level and Bohr radius of FX in a bulk crystal are expressed by the hydrogen atom model as

$$E_X = 13.6\left(\frac{\mu}{m_0}\right)\frac{1}{\varepsilon^2}\ [\text{eV}] \tag{21.4}$$

$$a_X = \varepsilon\left(\frac{m_0}{\mu}\right)a_B, \tag{21.5}$$

where

ε is a relative dielectric constant
m_0 is the electron mass
μ is the reduced effective mass $((1/\mu = (1/m_e^*) + (1/m_h^*))$
a_B is Bohr radius (0.0529 nm), respectively

Photon energy of the FX emission is given by

$$E = E_g - E_X. \tag{21.6}$$

Some of excitons are trapped by impurities to form BX. The recombinations of excitons bounded by a neutral donor and ionized donor are expressed by (D^0, X) and (D, X), respectively. Similarly those of excitons bounded by a neutral acceptor and ionized acceptor are expressed by (A^0, X) and (A, X), respectively. The photon energy of the recombination is given by

$$E = E_g - E_X - E_b, \tag{21.7}$$

where E_b is a binding energy for trapping exciton at impurities. The magnitude of E_b is about 1/3 of E_D and E_A or less.

Basic parameters are listed in Table 21.1 for typical semiconductors. The values of the band gap energy and exciton binding energy are taken after several literatures (Pankove, 1971; Vugaftman et al., 2001). Exciton radius in the last column is calculated by (21.4) and (21.5) using the relative dielectric constants and exciton binding energies listed in the table. The values give only a measure of the exciton radius, because the differences among the reported values for the listed parameters are fairly large.

21.2.3 TEM-Cathodoluminescence Technique

TEM-cathodoluminescence (TEM-CL) is a technique to observe luminescence spectra and scanning images using a converged electron beam having a probe size of nm order (Yamamoto, 2002, 2008). The characteristics of TEM-CL are as follows: (1) Inner structures can be observed by TEM using thin samples and comparison between TEM and CL images is possible and (2) high spatial resolution of about 100 nm can be achieved. In CL, incident electrons are scattered in a specimen from the incident beam direction, as shown in Figure 21.2. The scattered electrons produce secondary electrons and plasmons, which generate electron-hole pairs in the hatched region (generation region). The excited carriers diffuse into surroundings and some of them recombine radiatively in a surrounding region (diffusion region). The diffusion of the excess minority carriers is important, because majority carriers exist everywhere. The resolution of CL is given by the diameter of the diffusion region, which is determined by the following factors: (1) the electron beam diameter, (2) size of the generation region, and (3) the diffusion length of the minority carrier. The diameter of the electron beam is less than 10 nm, and is negligible compared with the other factors.

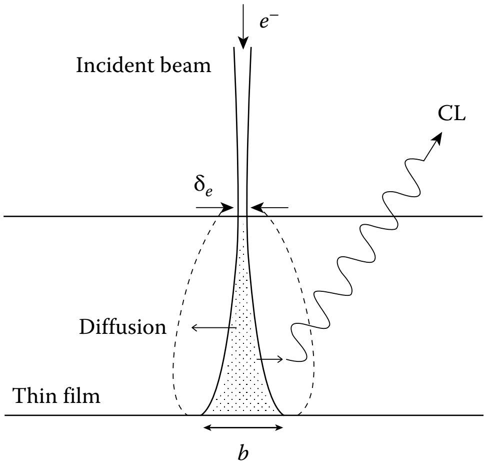

FIGURE 21.2 Carrier generation and light emission process in TEM-CL.

The diameter of the beam spread due to elastic scattering in a thin specimen of thickness t (μm) is given by

$$b = 6.25\left(\frac{Z}{E}\right)\left(\frac{\rho}{A}\right)^{1/2} t^{3/2}\ [\mu\text{m}], \tag{21.8}$$

where

Z is the atomic number
A is the atomic mass
E (keV) is an incident electron energy
ρ (g/cm^3) is the density (Goldstein et al., 1977)

For example, assuming that for GaAs, $Z_{\text{av}} = 32$, $A_{\text{av}} = 72.3$, $\rho = 5.3\,\text{g/cm}^3$, and a specimen thickness is 0.5 μm, we obtain $b = 0.24\,\mu\text{m}$. However, Monte Carlo simulation of the electron scattering showed that the electron density is concentrated on the center of the beam, so the effective beam spread is smaller than that given by Equation 21.8 (see Figure 21.3).

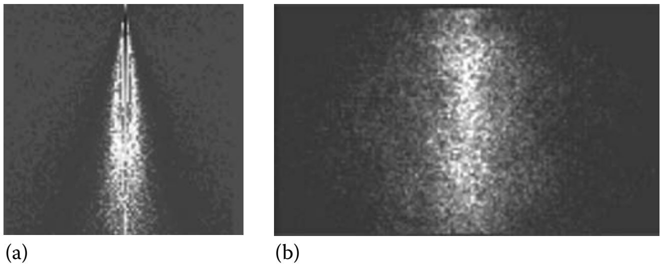

FIGURE 21.3 Monte Carlo simulation of (a) carrier generation by incident electrons and (b) carrier diffusion.

Equation 21.8 indicates that the higher accelerating voltage and thinner specimen thickness are advantageous for realizing high spatial resolution. However, when the accelerating voltage exceeds 100 kV, CL intensity starts to decrease because of the formation of specimen damage such as vacancies. TEM is typically operated at an accelerating voltage of 80 kV for CL measurements.

Diffusion length is an effective parameter for determining spatial resolution in CL of semiconductors. In insulators we can ignore carrier diffusion, but in semiconductors, carrier diffusion occurs with diffusion length of about 100 nm for GaN and 1 μm for GaAs. The diffusion length depends on dopant density and temperature, and reduces by surface recombination effect in thin specimens (Nakaji et al., 2005; Ino and Yamamoto, 2008). In quantum structures of semiconductors, carriers are confined in the localized structures smaller than the diffusion length, and individual nanostructures can be observed in CL images with high resolution as in the case of self-standing nanowires.

TEM-CL system based on a 200 kV transmission electron microscope (JEM-2000FX) is illustrated in Figure 21.4. Samples

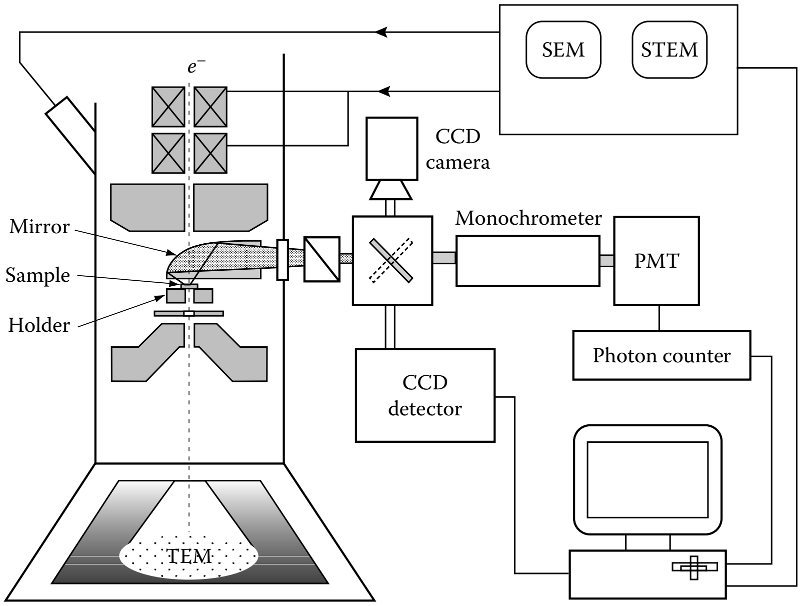

FIGURE 21.4 TEM-CL system.

are examined in the TEM, both in the stationary beam illumination mode and the scanning beam illumination mode. The scanning electron microscopy (SEM) imaging using a secondary electron detector and the scanning transmission electron microscopy (STEM) imaging using a transmitted electron beam detector are both possible in the TEM. A beam size of 10 nm in diameter and a probe current of 1 nA was conventionally used. The temperature of a sample can be varied from 20 K to room temperature using a liquid-He cooling holder and liquid-N_2 cooling holder.

Light emitted from a sample is collected by an ellipsoidal mirror or parabolic mirror, being guided through a polarizer and a monochromator, and is detected by a photomultiplier tube (PMT) with a GaAs photocathode for visible light and an InGaAs photocathode for infrared light. A CCD detector is also used for measuring spectra with high sensitivity. Monochromatic CL images can be obtained by scanning the electron beam in the SEM mode. This imaging technique is similar to SEM or EDX mapping, because we use the light intensity collected by the mirror, instead of secondary electron yields or x-ray intensities, as in the case of the SEM. An image is composed of, for example, 100 × 100 pixels, so that the focused electron beam is scanned across the specimen and stays 0.1 s for each pixel. Therefore, it takes 1,000 s to obtain one complete image. The CL images in this paper are shown with the absolute intensity.

In the CL detection system, a linear polarizer is located between the ellipsoidal mirror and monochrometer to select linearly polarized components of the emitted light; here *p*-polarized (*s*-polarized) light, which is detected with the polarization direction parallel (perpendicular) to the monochrometer slit, mainly involves linearly polarized light component parallel (perpendicular) to the longer axis of the ellipsoidal mirror or the axis of the parabola. A mask with a small hole is set in front of the CCD detector for the angular resolved measurement using the parabolic mirror. The mask is movable using an *X-Y* stage in a plane normal to the light path and selects an emission angle.

21.3 Optical Properties of Semiconductor Quantum Structures

21.3.1 Band Structure of Semiconductor

The wavefunction of a carrier in a semiconductor crystal of volume V is of a form of Bloch function,

$$\psi_{nk} = \frac{1}{\sqrt{V}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} u_{nk}(\boldsymbol{r}), \tag{21.9}$$

which is a product of a periodic function $u_{nk}(\boldsymbol{r})$ of the crystal lattice and a slowly varying function $\exp(i\boldsymbol{k}\cdot\boldsymbol{r})$. Subscripts n and $\boldsymbol{k}$ indicate a number of the Bloch band and wavevector, respectively. In many semiconductors of diamond and zinc-blend structures, the carrier wavefunction at the bottom of the conduction band has the *s*-like orbital and that at the top of the valence band has the *p*-like orbital (Basu, 1997; Yu and Cardona, 1999). The function $u_{n0}(\boldsymbol{r})$ has a form of the atomic orbital in the vicinity of the atoms. The band structure of semiconductor has been treated by the $\boldsymbol{k}\cdot\boldsymbol{p}$ perturbation theory (Luttinger and Kohn, 1955). The wavefunctions of electron states at the Γ point in the conduction band are given by the product of the atomic *s*-orbital function and spin orbital function as

$$u^c \frac{1}{2} = |s\uparrow\rangle \quad \text{and} \quad u^c\left(-\frac{1}{2}\right) = |s\downarrow\rangle \tag{21.10}$$

Here, $|\uparrow\rangle$ and $|\downarrow\rangle$ are the spin orbital functions with up and down spins, respectively. The wavefunctions of the hole states at Γ point in the valence band are given by the $\boldsymbol{k}\cdot\boldsymbol{p}$ theory using the atomic *p*-orbital functions $|X\rangle$, $|Y\rangle$, and $|Z\rangle$. Assuming that $\boldsymbol{k}$ is parallel to the z direction, the following functions are used as the valence band wavefunctions,

$$\begin{aligned}
u_{\frac{3}{2},\frac{3}{2}} &= -\frac{1}{\sqrt{2}}\,|(X+iY)\uparrow\rangle \\
u_{\frac{3}{2},-\frac{3}{2}} &= \frac{1}{\sqrt{2}}\,|(X-iY)\downarrow\rangle \\
u_{\frac{3}{2},\frac{1}{2}} &= -\frac{1}{\sqrt{6}}\big[|(X+iY)\downarrow\rangle - |2Z\uparrow\rangle\big] \\
u_{\frac{3}{2},-\frac{1}{2}} &= \frac{1}{\sqrt{6}}\big[|(X-iY)\uparrow\rangle + |2Z\downarrow\rangle\big] \\
u_{\frac{1}{2},\frac{1}{2}} &= -\frac{1}{\sqrt{3}}\big[|(X+iY)\downarrow\rangle + |Z\uparrow\rangle\big] \\
u_{\frac{1}{2},-\frac{1}{2}} &= -\frac{1}{\sqrt{3}}\big[|(X-iY)\uparrow\rangle - |Z\downarrow\rangle\big]
\end{aligned} \tag{21.11}$$

where subscripts j and m_j in u_{j,m_j} are quantum number of total angular momentum $\boldsymbol{J}$ and the z component J_z, respectively. The functions $u_{\frac{3}{2},\frac{3}{2}}$ and $u_{\frac{3}{2},-\frac{3}{2}}$ express the heavy hole (hh) wavefunctions, $u_{\frac{3}{2},\frac{1}{2}}$ and $u_{\frac{3}{2},-\frac{1}{2}}$ the light hole (lh) wavefunctions, and $u_{\frac{1}{2},\frac{1}{2}}$ and $u_{\frac{1}{2},-\frac{1}{2}}$ are those of the split-off band due to the spin–orbit interaction. Band structure, energy level, and dispersion can be calculated by orthogonalizing the Luttinger–Kohn Hamiltonian containing 2nd order $\boldsymbol{k}\cdot\boldsymbol{p}$ perturbation terms (Basu, 1997). The band structure of the direct gap type is schematically illustrated in Figure 21.5a. In the valence band of the zinc-blend structure, the hh and lh bands are degenerated at the Γ point, as seen in the figure, and the split-off band locates below them. Whereas in the wurtzite structure this degeneracy is dissolved, and the A, B, C bands exist separately (Figure 21.5b). The effective mass of hole is determined by the curvature of the dispersion curve at $\boldsymbol{k} = 0$, i.e., the small curvature gives a large effective mass.

If the wavevector of hole $\boldsymbol{k}_\text{h}$ is inclined from the z-axis, the hole wavefunctions of the valence band change from (21.11). Here we

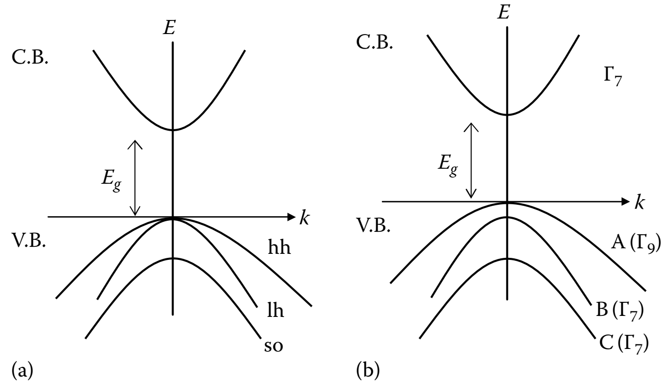

FIGURE 21.5 Band structures near the direct gap of (a) zinc-blend structure and (b) wurtzite structure.

take a z' axis parallel to $\boldsymbol{k}_h$, and x' and y' axes to be normal to z', as shown in Figure 21.6. The Cartesian coordinates (x, y, z) are fixed in space, and the coordinates (x', y', z') are related to (x, y, z) by the rotation of angle θ around the y-axis and successive rotation of angle ϕ around the z-axis. The conversion matrix between the wavefunctions in the two coordinates is expressed as

$$\boldsymbol{T} = \begin{pmatrix} \cos\theta\cos\phi & -\sin\phi & \sin\theta\cos\phi \\ \cos\theta\sin\phi & \cos\phi & \sin\theta\sin\phi \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}. \tag{21.12}$$

Namely, the wavefunction of the heavy hole is expressed in each coordinates as

$$-\frac{1}{\sqrt{2}}\left|(X'+iY')\uparrow\right\rangle = -\frac{1}{\sqrt{2}}\{(\cos\theta\cos\phi - i\sin\phi)\,|X\rangle + (\cos\theta\sin\phi + i\cos\phi)\,|Y\rangle - \sin\theta\,|Z\rangle\}\,|\!\uparrow\rangle \tag{21.13}$$

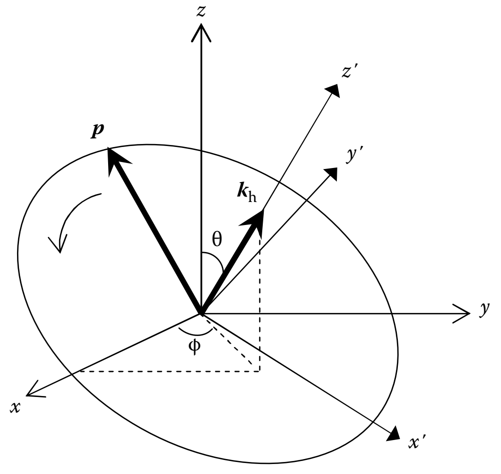

FIGURE 21.6 Relation between the coordinates fixed in space and those associated with the wavevector of hole $\boldsymbol{k}_\mathrm{h}$.

and that of the light hole is expressed as

$$\frac{1}{\sqrt{6}}\,|(X'-iY')\uparrow\rangle + \sqrt{\frac{2}{3}}\,|Z'\downarrow\rangle = \frac{1}{\sqrt{6}}\{(\cos\theta\cos\phi + i\sin\phi)\,|X\rangle + (\cos\theta\sin\phi - i\cos\phi)\,|Y\rangle - \sin\theta\,|Z\rangle\}\,|\!\uparrow\rangle + \sqrt{\frac{2}{3}}\{\sin\theta\cos\phi\,|X\rangle + \sin\theta\sin\phi\,|Y\rangle + \cos\theta\,|Z\rangle\}\,|\!\downarrow\rangle \tag{21.14}$$

21.3.2 Theory of Light Emission

Photon is emitted when an electron in the higher energy state transfers to the lower state. The transition probability can be given by the Fermi Golden rule,

$$W = \frac{2\pi}{\hbar}\sum_{\lambda}\sum_{f,i}|\langle f\,|H'\,|i\rangle|^2\,\delta(E_{f\lambda} - E_{i\lambda}). \tag{21.15}$$

Here, H' is the perturbation term due to the interaction between electron and photon appearing in the Hamiltonian, and is expressed as $H' = -(e/m)\boldsymbol{A}\cdot\boldsymbol{p}$ using a vector potential $\boldsymbol{A}$ and momentum $\boldsymbol{p}$. The wavefunctions of the initial and final states are denoted by $|i\rangle$ and $|f\rangle$, which are given by a product of wavefunctions of electron and photon. The summation over λ in (21.15) is taken over modes of photons. Substituting the above expression for the perturbation term, (21.15) can be rewritten as

$$W = \frac{2\pi}{\hbar}\sum_{\lambda}\sum_{f,i}\left(\frac{eA_0}{m}\right)^2 |\langle\psi_f|(\boldsymbol{e}\cdot\boldsymbol{p})e^{-i\boldsymbol{K}\cdot\boldsymbol{r}}|\psi_i\rangle|^2 \times (n_\lambda + 1)\delta(E_i - E_f - \hbar\omega). \tag{21.16}$$

where $\boldsymbol{K}$ is a wavevector of photon. The first term in $(n_\lambda + 1)$ expresses the stimulated emission and the second term expresses the spontaneous emission. In the following, we are only concerned with the spontaneous emission.

The rate of the spontaneous emission at a frequency ω in the transition from electron state in the conduction band to the hole state in the valence band is given by

$$R(\omega) = \frac{2\pi}{\hbar}\left(\frac{eA_0}{m}\right)^2 \sum_{c,v}|M|^2 G(\hbar\omega) f_c(E_c)\left[1 - f_v(E_v)\right]\delta(E_c - E_v - \hbar\omega). \tag{21.17}$$

The summation over the photon modes is replaced by the optical density of states G, which is derived from the van Roosbroeck–Shockly relation as

$$G(\hbar\omega) = \frac{n^3}{\pi^2 c^3 \hbar^3}(\hbar\omega)^2, \tag{21.18}$$

where n is the refractive index of the material. $f_c(E_c)$ and $f_v(E_v)$ are Fermi distribution functions. It is noted that we should use the pseudo-Fermi energy instead of the equilibrium one, when carriers are excited by outer sources. The product of $f_c(1-f_v)$ in (21.17) produces a sharp peak near the band gap energy $(E_c - E_v)$ in luminescence spectra. This is clear difference from the absorption spectrum which shows multiple peaks due to $|M(\omega)|^2$ above the band gap energy.

The term M is an optical transition matrix element, giving emission intensity of the transition, and is written as

$$M = \langle \psi_v | e^{-i\boldsymbol{K}\cdot r} (\boldsymbol{e}\cdot\boldsymbol{p}) | \psi_c \rangle \tag{21.19}$$

where $\boldsymbol{e}$ is a unit vector parallel to the electric field of light, i.e., the polarization direction. Using the momentum operator $\boldsymbol{p} = -i\hbar\nabla$ and wavefunctions of the conduction and valence bands, (21.19) can be rewritten as

$$M = -i\frac{\hbar}{V}\left[\int e^{i(\boldsymbol{k}_c - \boldsymbol{K} - \boldsymbol{k}_v)\cdot \mathbf{r}} \left\{ u_v^*(\boldsymbol{e}\cdot\nabla u_c) + i\,\boldsymbol{e}\cdot\boldsymbol{k}_c\, u_v^*\, u_c \right\} \mathrm{d}\mathbf{r}\right] \tag{21.20}$$

The second term is zero because of orthogonality of the wavefunctions. The following equation is valid for any periodic function of the lattice translation, $f(\boldsymbol{r}) = f(\boldsymbol{r} + \boldsymbol{R}_n)$,

$$\int f(\boldsymbol{r}) e^{i(\boldsymbol{k}_c - \boldsymbol{K} - \boldsymbol{k}_v)\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{r} = \sum_n e^{i(\boldsymbol{k}_c - \boldsymbol{K} - \boldsymbol{k}_v)\cdot\boldsymbol{R}_n} \times \int_{Vc} f(\boldsymbol{r}) e^{i(\boldsymbol{k}_c - \boldsymbol{K} - \boldsymbol{k}_v)\cdot\boldsymbol{r}}\, \mathrm{d}\boldsymbol{r} \propto \delta(\boldsymbol{k}_c - \boldsymbol{K} - \boldsymbol{k}_v) \tag{21.21}$$

Using this, the integral in (21.20) is found to be non-zero only when $\boldsymbol{k}_c = \boldsymbol{k}_v + \boldsymbol{K}$, indicating the conservation of wavevector in the transition. The wavelength of light with band gap energy is relatively long, i.e., $\boldsymbol{K} \ll \boldsymbol{k}_c, \boldsymbol{k}_v$ so we may approximate as $\boldsymbol{k}_c \approx \boldsymbol{k}_v$. As a result we obtain

$$M = \langle u_v | \boldsymbol{e}\cdot\boldsymbol{p} | u_c \rangle \delta_{\boldsymbol{k}_v, \boldsymbol{k}_c} \tag{21.22}$$

Here the condition $\boldsymbol{k}_c \approx \boldsymbol{k}_v$ indicates the "vertical" interband transition.

21.3.3 Polarization

The vector $\boldsymbol{e}$ in the optical transition matrix element indicates polarization direction of the electric field of light. Thus, the polarization in luminescence derives from the property of the matrix element. Here we consider

$$\langle u_v | \boldsymbol{p} | u_c \rangle$$

as a momentum vector. The vector $\boldsymbol{p}$ reflects the polarization of the emitted light; when $\boldsymbol{e}$ is perpendicular to $\boldsymbol{p}$, M is zero, which means that the polarization component perpendicular to $\boldsymbol{p}$ is zero. In other words, the x component of the electric field, E_x, is proportional to $\langle u_v | p_x | u_c \rangle$. From symmetry, only the following matrix elements of

$$\langle X | p_x | s \rangle = \langle Y | p_y | s \rangle = \langle Z | p_z | s \rangle = -i\frac{\hbar}{m}P = p_{cv} \tag{21.23}$$

are non-zero, where P is the factor introduced by Kane (1957).

For example, in the transition from the electron state of $|s\uparrow\rangle$ in the conduction band to the heavy hole state, the matrix elements can be written as

$$\begin{aligned} \left\langle \frac{3}{2}, \frac{3}{2} \middle| p_x \middle| s\uparrow \right\rangle &= -\frac{1}{\sqrt{2}}\langle X | p_x | s \rangle = -\frac{1}{\sqrt{2}} p_{cv} \\ \left\langle \frac{3}{2}, \frac{3}{2} \middle| p_y \middle| s\uparrow \right\rangle &= -\frac{i}{\sqrt{2}}\langle Y | p_y | s \rangle = -\frac{i}{\sqrt{2}} p_{cv} \\ \left\langle \frac{3}{2}, \frac{3}{2} \middle| p_z \middle| s\uparrow \right\rangle &= 0 \end{aligned} \tag{21.24}$$

From these, the relation among the electric field components can be written as

$$\frac{E_y}{E_x} = e^{i\frac{\pi}{2}}, \quad E_z = 0 \tag{21.25}$$

This means that the light emitted toward the z direction has a right-handed circular polarization. This situation is illustrated in Figure 21.7, where we express $|X\rangle$ and $|Y\rangle$ in terms of pure angular momentum states. The wavevector $\boldsymbol{k}_h$ of the heavy hole is pointing to the z direction. The wavefunction of the heavy hole is extended parallel to the x-y plane, and its momentum vector $\boldsymbol{p}$ rotates on the same plane. The wavevector $\boldsymbol{k}_e$ of the electron in the conduction band is parallel to the z-axis, and the wavefunction has a spherical symmetry. In this recombination process of the electron and heavy hole, the angular momentum of the heavy hole is preserved as that of the circular polarized light.

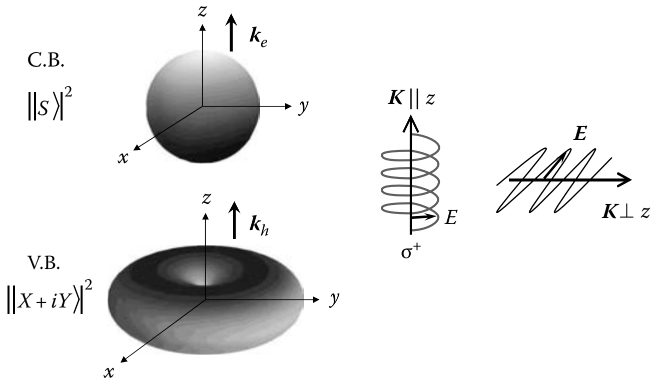
z
k_e
C.B.
$\left\| s \rangle \right|^2$
y
x
$K \| z$
E
E
σ^+
$K \perp z$
z
k_h
V.B.
$\left\| X + iY \rangle \right|^2$
y
x

FIGURE 21.7 Wavefunctions of the electron in the conduction band and the heavy hole in the valence band. Circularly polarized and linearly polarized light are emitted by the optical transition between them.

Similarly in the transition from the electron state $|s\uparrow\rangle$ in the conduction band to the light hole state $\left|\frac{3}{2},-\frac{1}{2}\right\rangle$, the matrix elements can be written as

$$\begin{aligned}
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_x \left|s\uparrow\right\rangle &= \frac{1}{\sqrt{6}}\left\langle X\right| p_x \left|s\right\rangle = \frac{1}{\sqrt{6}}p_{cv}\\
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_y \left|s\uparrow\right\rangle &= -\frac{i}{\sqrt{6}}\left\langle Y\right| p_y \left|s\right\rangle = -\frac{i}{\sqrt{6}}p_{cv} \qquad (21.26)\\
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_z \left|s\uparrow\right\rangle &= 0
\end{aligned}$$

Then, light with a left-handed circular polarization is emitted. Whereas in the transition from the electron state $|s\downarrow\rangle$ to the light hole state $\left|\frac{3}{2},-\frac{1}{2}\right\rangle$, the matrix elements are written as

$$\begin{aligned}
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_x \left|s\downarrow\right\rangle &= 0\\
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_y \left|s\downarrow\right\rangle &= 0 \qquad (21.27)\\
\left\langle\frac{3}{2},-\frac{1}{2}\right| p_z \left|s\downarrow\right\rangle &= \sqrt{\frac{2}{3}}\left\langle Z\right| p_z \left|s\right\rangle = \sqrt{\frac{2}{3}}p_{cv}
\end{aligned}$$

In this case, light is linearly polarized parallel to the z-axis.

The emission intensity of CL is proportional to a mean square of the electric field vector. Therefore, in the transition from $|s\uparrow\rangle$ in the conduction band to the heavy hole state in the valence band, the emission intensity of light polarized in each direction is expressed as

$$I_x \propto \frac{1}{2}p_{cv}^2, \quad I_y \propto \frac{1}{2}p_{cv}^2, \quad I_z \propto 0 \qquad (21.28)$$

Similarly in the transition from $|s\uparrow\rangle$ in conduction band to the light hole state $\left|\frac{3}{2},-\frac{1}{2}\right\rangle$ in the valence band,

$$I_x \propto \frac{1}{6}p_{cv}^2, \quad I_y \propto \frac{1}{6}p_{cv}^2, \quad I_z \propto 0 \qquad (21.29)$$

And in the transition from $|s\downarrow\rangle$ in the conduction band to the light hole state $\left|\frac{3}{2},-\frac{1}{2}\right\rangle$ in the valence band,

$$I_x \propto 0, \quad I_y \propto 0, \quad I_z \propto \frac{2}{3}p_{cv}^2 \qquad (21.30)$$

The relative intensity and polarization of the emission associated with the transition between the conduction band and the valence band states are schematically depicted in Figure 21.8.

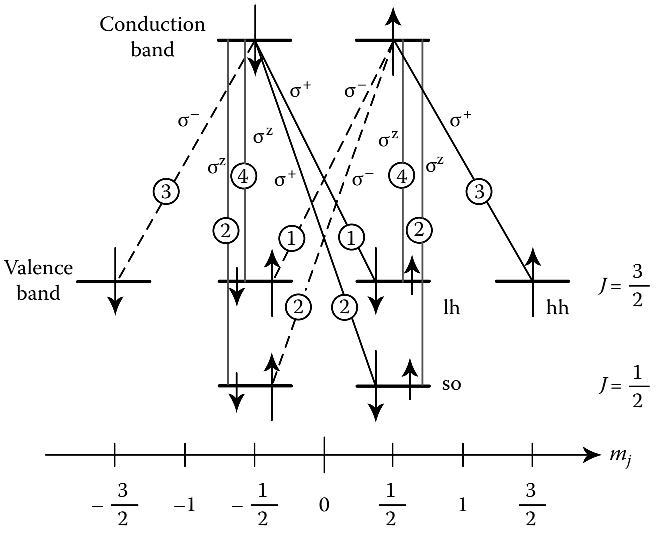

FIGURE 21.8 Relative intensity and polarization of the emission associated with the transition between the conduction band and the valence band states.

The semiconductors of zinc-blend structure are crystallographically isotropic, and any direction can be taken as the z-axis. Thus, the luminescence from the bulk crystal is not polarized, because the polarization of each recombination process cancels out as a whole. Typical cases in which polarized emission occurs are luminescence from quantum structures and that from semiconductors of wurtzite structure. In the case of quantum well structures, a unique axis appears normal to the well, and the heavy and light hole states have anisotropic distribution which causes dissociation of the band degeneration (Yu and Cardona, 1999). Similarly in semiconductors of wurtzite structure, the c-axis is a unique axis, and the bands of the heavy and light hole states are dissolved into the A and B bands. If the highest energy state in the valence band is heavy hole as in the case of GaN, the optical transition to the heavy hole state mainly occurs, and the luminescence is polarized perpendicular to the c-axis. In addition, the polarization property of luminescence changes by a strain field in semiconductor crystals, because a lattice strain causes an anisotropic deformation potential and reconstructs the band structure (Hensel and Feher, 1963; Pollak and Cardona, 1968).

21.4 Quantum Structures

When carriers are confined in small-scaled semiconductor structures, carrier wavefunctions form standing waves and start to show quantum character. The wavelength and wave number of a carrier in a one-dimensional potential well of an infinite height with a width of L are given as

$$\lambda = \frac{2L}{n}, \quad k = \frac{n\pi}{L} \qquad (21.31)$$

where n is an integer number. The wave number k can only take discrete values in this direction corresponding to $n = 1, 2, 3, \ldots$, and has continuous value in the other direction. Kinetic energy in this direction takes discrete values given by

$$\Delta E = \frac{\hbar^2}{2m}k^2 = \frac{\hbar^2}{2m}\left(\frac{n\pi}{L}\right)^2 \tag{21.32}$$

This anisotropy in k gives rise to anisotropy in optical transition and polarization of luminescence. In the following section, we start with a quantum well structure because of simplicity in theoretical treatment and next move to a quantum wire structure which shows strong polarization property.

21.4.1 Quantum Well

We consider carriers confined in a thin semiconductor plate of a width L_z in the z direction. The wavefunction of electrons in the conduction band and holes in the valence band are expressed using the Bloch function $u_{c,v}(\boldsymbol{r})$ at $\boldsymbol{k} = 0$ under the effective mass approximation as

$$\Psi^e(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\sin\left(\frac{n_c\pi}{L_z}z\right)e^{i(k_x x + k_y y)}\,u_c(\boldsymbol{r}) = \phi_e(z)e^{i(k_x x + k_y y)}\,u_c(\boldsymbol{r}) \tag{21.33}$$

$$\Psi^v(\boldsymbol{r}) = \frac{1}{\sqrt{V}}\sin\left(\frac{n_v\pi}{L_z}z\right)e^{i(k_x x + k_y y)}\,u_v(\boldsymbol{r}) = \phi_h(z)e^{i(k_x x + k_y y)}\,u_v(\boldsymbol{r}) \tag{21.34}$$

where ϕ_e and ϕ_h are envelop functions of the electron and hole, respectively, showing standing waves in the confined direction. The energy of the electron and hole measured from the top of the valence band are given by

$$E_e(n) = \frac{\hbar^2}{2m_e}\left(\frac{n_c\pi}{L_z}\right)^2 + \frac{\hbar^2}{2m_e}(k_x^2 + k_y^2) + E_g \tag{21.35}$$

$$E_h(n) = \frac{\hbar^2}{2m_h}\left(\frac{n_v\pi}{L_z}\right)^2 + \frac{\hbar^2}{2m_h}(k_x^2 + k_y^2) \tag{21.36}$$

where

m_e and m_h are the effective mass of electron and hole, respectively

n_c and n_v are their quantum numbers

Since wavevector components, k_x and k_y, take continuous values, mini bands are formed for each quantum number, as shown in Figure 21.9. It should be noted that (21.36) is a rough approximation because the band mixing should be considered for the valence band in the quantum structure, as shown in Figure 21.9. However, this gives a simple picture for the band splitting in the valence band. In the quantum well structure, the degenerated bands at the Γ point are dissolved, and the highest energy state is the heavy hole. Consequently, the transition from the electron state to the heavy hole state becomes dominant. Since the energy shift of the valence band is relatively small compared to that of the conduction band, the emission energy is roughly estimated from (21.35) and (21.36).

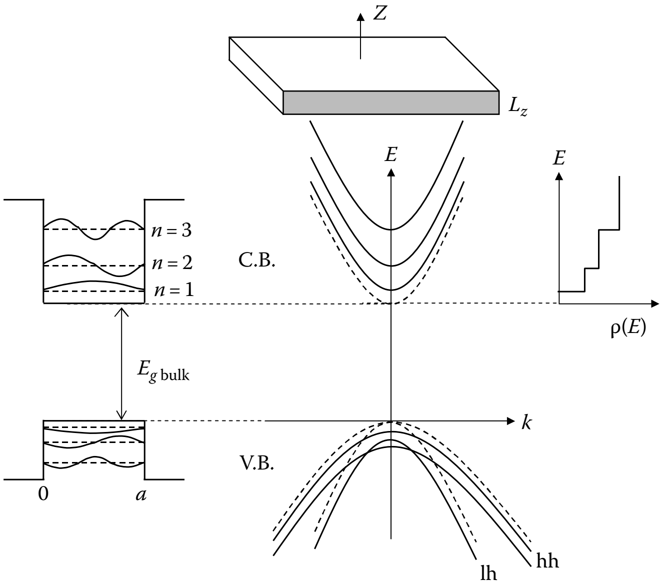

FIGURE 21.9 Wave functions, band structure, and density of states in a quantum well.

The probability of interband transition from the conduction band to the heavy hole state at the band edge (BE) is given by the following matrix element,

$$M = \langle\Psi^v|\boldsymbol{e}\cdot\boldsymbol{p}|\Psi^c\rangle = \langle\phi_h(z)|\phi_e(z)\rangle\left\langle u_{\frac{3}{2},\frac{3}{2}}\middle|\boldsymbol{e}\cdot\boldsymbol{p}\middle|u_c\right\rangle \tag{21.37}$$

The product of the electron and hole envelop functions is non-zero only when

$$n_c = n_v \tag{21.38}$$

If M is zero, the corresponding transition is inhibited. This gives the selection rule of the transition, and means that the optical interband transition only occurs between the electron and hole states of the same quantum number. The emission energy is given by

$$E(n) = E_e(n) - E_h(n) = \frac{\hbar^2}{2\mu}\left(\frac{n_c\pi}{L_z}\right)^2 + \frac{\hbar^2}{2\mu}(k_x^2 + k_y^2) + E_g \tag{21.39}$$

where μ is the reduced effective mass. The lowest energy transition occurs for $n_c = n_v = n = 1$. However, if the barrier height is finite, other transitions become permitted.

The polarization of emission from the quantum well is deduced from (21.37). Here we take the z-axis normal to the well. In the transition between the lowest mini bands at the BE, the wavevector of carrier is $\boldsymbol{k}_c = \boldsymbol{k}_v = (0, 0, \pi/L_z)$, which points to the z direction. Therefore, the result of (21.24) can be applied for this case, so the emission toward the z direction is a right-handed circular polarized light. Whereas light emitted along the well is linearly polarized parallel to the well.

In the interband transition between the electronic states with a finite $k_\perp$ component normal to the z-axis, the wavevector of the carriers is inclined from the z-axis, and the wavefunction of the valence band changes according to (21.13). The emission energy increases according to (21.39). In order to find the polarization of emitted light with this energy, the transition matrix element should be integrated over equivalent transitions of the same energy. Firstly we consider the transition from the $|s\uparrow\rangle$ state in the conduction band to the heavy hole state in the valence band. The wavevector component parallel to the z-axis is π/L_z, and then $\boldsymbol{k} = (k\sin\theta\cos\phi,\ k\sin\theta\sin\phi,\ \pi/L_z)$. The wavefunction of the heavy hole is expressed by (21.13). The x and y directions are equivalent, so here we consider the emission propagating in the x direction with the electric field parallel to the y direction (TE wave in Figure 21.10). After integrating over the azimuth angle ϕ, the averaged transition matrix element is given as

$$\begin{aligned}\langle |M|^2\rangle_{hh,TE} &= \frac{1}{2\pi}\int_0^{2\pi}\left|\left\langle \frac{3}{2},\frac{3}{2}\middle|p_y\middle|s\right\rangle\right|^2 d\phi\\ &= \frac{1}{2}\cdot\frac{1}{2\pi}\int_0^{2\pi}(\cos^2\theta\sin^2\phi+\cos^2\phi)\,d\phi = \frac{1}{4}p_{cv}^2\,(1+\cos^2\theta)\end{aligned} \tag{21.40}$$

Similarly for the electric field parallel to the z direction (TM wave), the averaged transition matrix element is given as

$$\begin{aligned}\langle |M|^2\rangle_{hh,TM} &= \frac{1}{2\pi}\int_0^{2\pi}\left|\left\langle \frac{3}{2},\frac{3}{2}\middle|p_z\middle|s\right\rangle\right|^2 d\phi\\ &= \frac{1}{2}\cdot\frac{1}{2\pi}\int_0^{2\pi}\sin^2\theta\,d\phi = \frac{1}{2}p_{cv}{}^2\sin^2\theta\end{aligned} \tag{21.41}$$

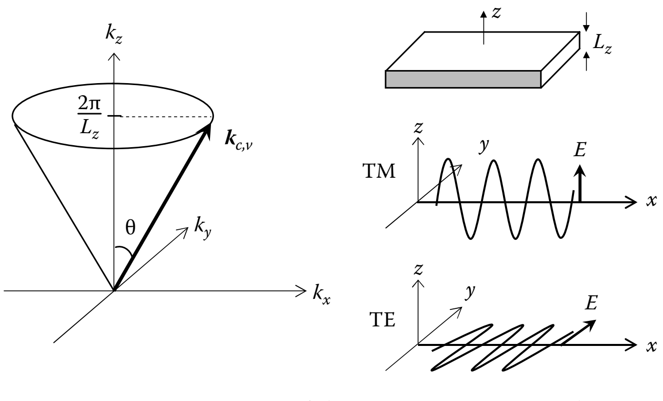

FIGURE 21.10 Wavevectors of the same energy states in the miniband. Light emission of both the TM mode and TE mode occurs from a quantum well.

If the $k_\perp$ component is zero ($k_x = k_y = 0$), i.e., $\theta = 0$ at the mini-BE, the TM component is zero and the polarization of the emission is parallel to the well. As the tilted angle θ from the z-axis increases, the emission energy increases with increasing $\boldsymbol{k}$ and the polarization direction tilts from the well plane.

21.4.2 Quantum Wire

21.4.2.1 Energy Levels in Quantum Wire

A semiconductor quantum wire of an infinite length along the z-axis with a rectangular cross section of the widths of L_x and L_y confines a carrier in an infinite barrier height. Electron wavefunction in the quantum nanowire is expressed in the effective mass approximation by a product of the Bloch function at the BE and the envelop function in the x-y plane normal to the wire axis. For simplicity we consider the electron state at the BE of the conduction band in semiconductors with a cubic symmetry and a direct band structure. The electron wavefunction is expressed as

$$\Psi^e(\boldsymbol{r}) = \phi^c(x,y)e^{ik_z z}\,u_c(\boldsymbol{r}) \tag{21.42}$$

The envelop function of the electron is written as

$$\phi^c(x,y) = \frac{1}{\sqrt{L_xL_y}}\sin\left(\frac{\pi n_c}{L_x}x\right)\,\sin\left(\frac{\pi m_c}{L_y}y\right) \tag{21.43}$$

where n_c and m_c are integer. The energy of the electron in the (n_c, m_c) band is given by

$$E_e(n_c,m_c) = \frac{\hbar^2}{2m_e}\left\{\left(\frac{\pi n_c}{L_x}\right)^2+\left(\frac{\pi m_c}{L_y}\right)^2\right\}+\frac{\hbar^2}{2m_e}k_z^2+E_g \tag{21.44}$$

where m_e is the electron effective mass. The lowest energy band corresponds to the quantum numbers of $n_c = m_c = 1$. Mini bands are formed from (21.44) in the conduction band, as shown in Figure 21.11. The density of states as a function of energy has sharp peaks at the mini-BE given by the first term, and shows a sawtooth shape.

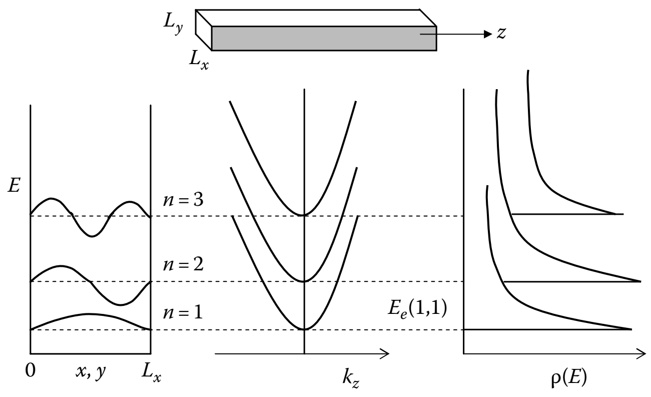

FIGURE 21.11 Wave functions, band structure, and density of states in a quantum wire with a square cross section.

On the other hand, electrons in a quantum wire with a circular cross section of radius R has the following envelop functions expressed using the cylindrical coordinates (ρ, ϕ) as

$$\phi_e(\rho,\phi) \propto J_\ell\left(\frac{j_{n\ell}\rho}{R}\right)\cos(\ell\phi) \quad \text{or} \quad J_\ell\left(\frac{j_{n\ell}\rho}{R}\right)\sin(\ell\phi) \tag{21.45}$$

where

J_l is the Bessel function of the ℓth order

$j_{l,n}$ is the 0 point value of nth order given by the condition that $J_l(j_{l,n}) = 0$

The lowest values are given as $j_{0,1} = 2.405$, $j_{1,1} = 3.832$, and $j_{0,2} = 5.520$. A set of energy levels are derived as

$$E_e(\ell,n) = \frac{\hbar^2 j_{\ell,n}{}^2}{2m_e R^2} + \frac{\hbar^2}{2m_e}k_z{}^2 + E_g \tag{21.46}$$

All the energy states indicated by the quantum number (ℓ, n) are doubly degenerated, except for $\ell = 0$.

Change in electron energy levels in the conduction band due to the quantum confinement effect is large, whereas those of holes in the valence band are relatively small because of small curvature of the bands or large effective mass of hole. However, the band mixing occurs, and then the hole wavefunctions are formed by a mixture of the bulk wavefunctions in (21.11). This affects the property of the optical transitions. The hole wavefunctions are expressed by a linear combination of the bulk wavefunctions as

$$\Psi^v(\boldsymbol{r}) = e^{ik_z z}\sum_j \phi_j^v(x,y)u_j(\boldsymbol{r}) \quad \left(j = \frac{3}{2},\frac{1}{2},-\frac{1}{2},-\frac{3}{2}\right) \tag{21.47}$$

The envelop functions are written as

$$\phi_j^v(x,y) = \frac{1}{\sqrt{L_x L_y}}\sum_{n_v,m_v} c_j(n_v,m_v)\sin\left(\frac{\pi n_v}{L_x}x\right)\sin\left(\frac{\pi m_v}{L_y}y\right) \tag{21.48}$$

Here, we ignore the contribution of the split-off band to the mixing. The wavefunctions and energies in the valence band of the quantum wire can be derived from the theoretical calculation using the Luttinger–Kohn Hamiltonian in the effective mass approximation (Sercel and Vahala, 1990; Bockelmann and Bastard, 1992; McIntyre and Sham, 1992).

Theoretical treatments were firstly applied to the quantum wires manufactured from the quantum well structures, in which the wire has an axis parallel to the growth surface and rectangular shape in cross section (Tanaka and Sakaki, 1989; Tsuchiya et al., 1989). Because of the technological development for the fabrication of self-standing nanowire, energy band calculations have been done for self-standing wires with circular and rectangular cross sections of Si (Zhao et al., 2004), GaAs (Persson and Xu, 2004, Redlinski and Peeters, 2008), and InP (Pellegrini et al., 2005). The band dispersion of camel back structure appears in some of the high energy bands of the valence band. The character of the band structure such as the sequence of the levels does not change with the diameter of the quantum wire. The highest energy band in the valence band has a character of the light hole, and the second highest energy band has a character of the heavy hole. The difference in band structure between the nanowires grown in the [100] and [111] directions was shown in the GaAs nanowires (Redlinski and Peeters, 2008).

The luminescence due to exciton is generally dominant at low temperatures. The emission energy of the FX is given by the band gap energy minus the exciton binding energy as in (21.6). If the carrier confinement in quantum wires is perfect, the exciton binding energy becomes infinite in the limit of small wire size (Basu, 1997). A real nanowire has a finite size in diameter, the divergence can be avoided using a following potential for the Coulomb interaction between the electron-hole pair (Ando et al., 1991),

$$V(z,z_0) = -\frac{e^2}{4\pi\varepsilon}\frac{1}{\left(|z| + z_0\right)} \tag{21.49}$$

where z_0 is a parameter for the size effect of nanowire. Quantum confinement effect is large enough to be detected, when the diameter of the wire becomes comparable to or less than the exciton radius. The size dependence of the emission energy is approximately proportional to R^{-2}, as in (21.46). The theoretical calculations, including other effects, have been carried out for the size dependence of the emission energy; for example, the dielectric mismatch effect results in the $R^{-1.8}$ dependence (Redlinski and Peeters, 2008).

The barrier height of the carrier confinement is finite in real when a quantum wire is surrounded by another semiconductor or even by air. In such cases, the energy levels decrease from those given by (21.44) and (21.46), and the distances between the energy levels decrease, because they must exist below the barrier height level. When the diameter of wire decreases with the barrier height fixed, the energy levels move upward to gather near the barrier height level, and their total number decreases. However, at least one level is left at the barrier height in the symmetrical potential well in the limit of small size.

21.4.2.2 Luminescence from Quantum Wire

Luminescence from bulk crystals with cubic symmetry such as Si and GaAs is generally unpolarized, whereas that with hexagonal symmetry such as GaN and ZnO shows anisotropy in polarization because the c-axis is a unique axis. The luminescence from nanowires has been frequently observed to be polarized even for semiconductors with cubic symmetry. There are two reasons for the appearance of the polarization anisotropy in luminescence from nanowire; one is the anisotropy originated from the quantum effect, and another is the anisotropic response of nanowire to the electromagnetic fields parallel or perpendicular to the wire axis (Ruda and Shik, 2005, 2006).

The former effect appears when a diameter of nanowire becomes smaller than the exciton radius, which gives rise to the polarization anisotropy in the luminescence of excitons in the nanowire. The latter effect appears even in nanowires with diameter larger than the exciton radius.

21.4.2.2.1 Quantum Confinement Effect

The quantum confinement of carriers in thin nanowires affect on the polarization of luminescence. Here we treat a nanowire of a square cross section with a dimension of a and an infinite length. The wire is isotropic in the cross-sectional plane and the confinement potential has an infinite barrier height for carriers in the wire. At first we ignore the band mixing in the valence band, and use the wavefunctions of the one-band model expressed as follows:

$$\Psi^{v}(\boldsymbol{r}) = e^{ik_z z}\phi_j^{v}(x, y)u_j(\boldsymbol{r}) \tag{21.50}$$

When we take the heavy hole state for j, the transition matrix element is written as

$$M = \langle\psi_v|e^{-i\boldsymbol{K}\cdot\boldsymbol{r}}(\boldsymbol{e}\cdot\boldsymbol{p})|\psi_c\rangle = \sum_{k_i}\langle\phi_j^{v}|\phi_e\rangle\langle u_j(\boldsymbol{k}_i)|\boldsymbol{e}\cdot\boldsymbol{p}|u_c\rangle \tag{21.51}$$

Electron wavefunctions of the lowest energy states in the conduction band are expressed by (21.42) and (21.43) with $n_c = m_c = \pm 1$. The transition occurs under the selection rule of $n_c = n_v$ and $m_c = m_v$ in (21.43) and (21.48). There are four corresponding heavy hole states of the same energy at the BE ($k_z = 0$), having the following wavevectors:

$$\boldsymbol{k}_i = \left(\pm\frac{\pi}{a}, \pm\frac{\pi}{a}, 0\right) : i = 1 \sim 4 \tag{21.52}$$

The wavefunctions of the heavy hole states with these wavevectors are given by substituting $\theta = 90°$ in (21.13) as

$$\left|\frac{3}{2}, \frac{3}{2}\right\rangle_{k_i} = -\frac{1}{\sqrt{2}}\left\{-i\sin\phi_i\left|X\right\rangle + i\cos\phi_i\left|Y\right\rangle - \left|Z\right\rangle\right\}\left|\uparrow\right\rangle \tag{21.53}$$

Four azimuth angles of

$$\phi_1 = \frac{\pi}{4}, \quad \phi_2 = \frac{3\pi}{4}, \quad \phi_3 = \frac{5\pi}{4}, \quad \phi_4 = \frac{7\pi}{4} \tag{21.54}$$

correspond to the four wavevectors. Similarly, the wavefunctions of the light hole states with the same wavevectors are written as

$$\left|\frac{3}{2}, -\frac{1}{2}\right\rangle = \frac{1}{\sqrt{6}}\left\{i\sin\phi_i\left|X\right\rangle - i\cos\phi_i\left|Y\right\rangle - \left|Z\right\rangle\right\}\left|\uparrow\right\rangle + \sqrt{\frac{2}{3}}\left\{\cos\phi_i\left|X\right\rangle + \sin\phi_i\left|Y\right\rangle\right\}\left|\downarrow\right\rangle \tag{21.55}$$

If the transitions from the conduction band to those valence band states are assumed to occur independently, mean square of the transition matrix element is given as

$$|M|^2_{e-hh} = \frac{1}{4}\sum_{k_i}\left|\left\langle u_{\frac{3}{2},\frac{3}{2}}(\boldsymbol{k}_i)\middle|\boldsymbol{e}\cdot\boldsymbol{p}\middle|s\uparrow\right\rangle\right|^2 \tag{21.56}$$

Here, the term from $|s\downarrow\rangle$ to $\left|u_{\frac{3}{2},-\frac{3}{2}}(\boldsymbol{k}_i)\right\rangle$ is ignored, because it gives the same value as above. If we take the z direction along the wire axis, it follows that $\boldsymbol{e}\cdot\boldsymbol{p} = p_z$ for the polarization of emitted light parallel to the wire axis and $\boldsymbol{e}\cdot\boldsymbol{p} = p_x$ for that normal to the wire axis. Then the square of the transition matrix element is calculated using (21.52), and the two components parallel and perpendicular to the z-axis are obtained as

$$|M|^2_{e-hh\parallel} = \frac{1}{2}p_{ev}^2, \quad |M|^2_{e-hh\perp} = \frac{1}{4}p_{ev}^2 \tag{21.57}$$

The component parallel to the wire axis is larger than the perpendicular one because the transformed wavefunction in (21.52) involves a large $|Z\rangle$ component. When the polarization of light is inclined from the wire axis by an angle θ (Figure 21.12), the square of the transition matrix element can be written as

$$|M|^2_{e-hh} = \frac{1}{2}p_{ev}^2\cos^2\theta + \frac{1}{4}p_{ev}^2\sin^2\theta \tag{21.58}$$

The transition matrix element from the conduction band to the light hole state is calculated in the same way using (21.54),

$$|M|^2_{e-\ell h} = \frac{1}{6}p_{ev}^2\cos^2\theta + \frac{5}{12}p_{ev}^2\sin^2\theta \tag{21.59}$$

In this transition, the component perpendicular to the wire axis is larger than the parallel one. The two hole states locate at different energy levels, so the emission energy of this transition is

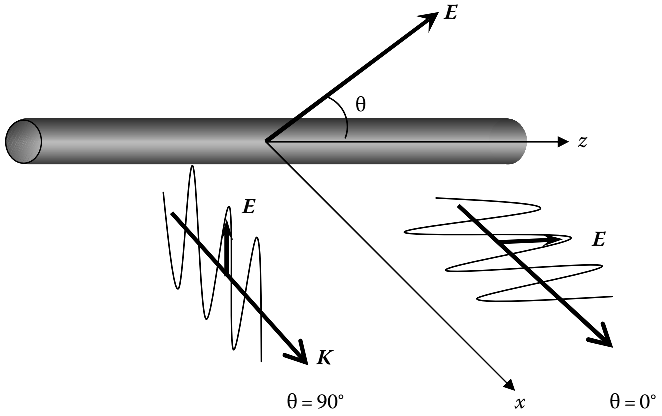

FIGURE 21.12 Light emissions from a nanowire polarized in directions inclined from the wire axis by angles of θ = 0° and 90°.

also different. If we use the relation $p_{cv}^2 = 3M^2$ in which M is the averaged value in the isotropic bulk crystal, the ratio between these transition matrix elements can be written as

$$\frac{|M|^2_{e-\ell h}}{|M|^2_{e-hh}} = \frac{\frac{1}{2}\cos^2\theta + \frac{5}{4}\sin^2\theta}{\frac{3}{2}\cos^2\theta + \frac{3}{4}\sin^2\theta} \tag{21.60}$$

This result was derived by several authors on the basis of the one-band model (Tanaka and Sakaki, 1989; Tsuchiya et al., 1989), though this treatment was claimed to be too simple, and later more reliable one was proposed on the basis of band mixing in the valence band.

In the photoluminescence excitation (PLE) spectrum of the quantum wires formed in the GaAs quantum well structure, two peaks were observed to appear near the BE energy, as shown in Figure 21.13 (Tsuchiya et al., 1989). On the other hand in the PL spectrum, only the lower energy peak appeared. The lower energy peak was assigned to be due to the transition to the heavy hole state, and the higher peak due to the transition to the light hole state, respectively. It appeared to be consistent with the result of (21.57). However, it was suggested that the one-band model cannot explain the polarization property correctly, and the theoretical treatment taking account of the band mixing in the valence band should be necessary. Sercel and Vahala (1990) showed that if the coherent combination of the hole states with the four wavevectors is considered in the one-band model, Equation 21.57 has a constant value (=1/3) independent of θ. They calculated the polarization dependence of the optical transition using the mixed states of the heavy and light holes in a cylindrical wire. As a result, the transition matrix elements associated with the two highest hole states were obtained as

$$|M|^2_{C-V\left(\frac{1}{2}\right)} = \left(\frac{2}{3}\cos^2\theta + \frac{1}{6}\sin^2\theta\right) p_{ev}^2 I^2_{C-V\left(\frac{1}{2}\right)} \tag{21.61}$$

$$|M|^2_{C-V\left(\frac{3}{2}\right)} = \frac{1}{2}\sin^2\theta \cdot p_{ev}^2\, I^2_{C-V\left(\frac{3}{2}\right)} \tag{21.62}$$

where I is an overlap integral. The optical transition of the lowest energy is C–V(1/2). The V(1/2) state involves only 41% of the heavy hole state. The single peak in the PL spectrum is due to this transition. The polarization component parallel to the wire axis is dominant in the emission, and the polarization ratio deduced from (21.60) is 60%. The polarization ratio becomes smaller as the barrier height decreases. The optical transition of the next lowest energy is C–V(3/2), and the V(3/2) state involves 67% of the heavy hole state. The polarization direction of the corresponding emission is perpendicular to the wire axis. The property of the polarization in these two transitions is well consistent with those of the two emission peaks in the PLE spectrum in Figure 21.13.

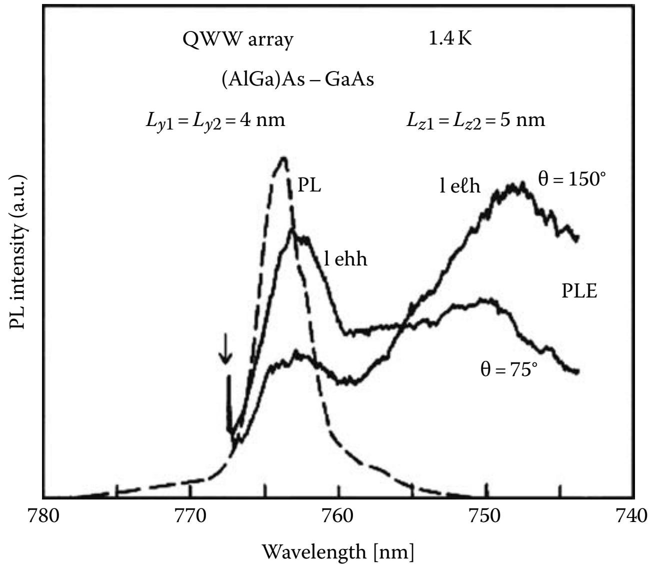

FIGURE 21.13 Photoluminescence (PL) and photoluminescence excitation (PLE) spectra of the quantum wires formed in the GaAs quantum well structure. (From Tsuchiya, M. et al., *Phys. Rev.*, 62, 455, 1989. With permission.)

Similar results were obtained in the calculation of optical absorption for self-standing GaAs nanowires (Redlinski and Peeters, 2008). The highest energy state in the valence band has a character of the light hole. The optical transition matrix element associated with this hole state has both polarization components parallel and perpendicular to the wire axis, and the former is much larger than the latter. The fine splitting of the lowest energy peak in the absorption spectrum appears in the [111] wire because of the anisotropy in the cross-sectional plane. The next highest energy state in the valence band has a character of the heavy hole, and the absorption occurs only for the polarization direction perpendicular to the wire axis. Emission energies in these two transitions are slightly different, so in PL and CL, the transition to the band having the light hole character is predominant.

21.4.2.2.2 Dielectric Effect

The second factor to affect the polarization of luminescence from semiconductor nanowires is the dielectric effect (Ruda and Shik, 2005). When a spatially homogeneous electric field is applied to a nanowire of cylindrical shape with the electric field vector parallel to the wire axis, magnitude and direction are the same both in and outside of the wire. Whereas the electric field is applied perpendicular to the wire axis, its magnitude is reduced by a factor of $(2\varepsilon_0/\varepsilon + \varepsilon_0)$, where ε and ε_0 are relative dielectric constants of semiconductor and surrounding material, respectively (Landau et al., 1984), i.e.,

$$\begin{aligned} E_{\parallel} &= E_{0\parallel} \\ E_{\perp} &= \frac{2\varepsilon_0}{\varepsilon + \varepsilon_0} E_{0\perp} \end{aligned} \tag{21.63}$$

This is true for the oscillating electric field of light if the diameter of the wire is sufficiently smaller than the wavelength of light. From this, the polarization dependence of the optical absorption is expressed by

$$\frac{k_{\parallel}}{k_{\perp}} = \left|\frac{\varepsilon + \varepsilon_0}{2\varepsilon_0}\right|^2 \tag{21.64}$$

This anisotropy appears in PL and light-induced current excited by polarized light.

As for the dielectric effect on the polarization of luminescence, we consider an oscillating electric dipole with a frequency of ω and a dipole moment of $\boldsymbol{d}_0 = (d_{0x}, d_{0y}, d_{0z})$ in a cylindrical nanowire. We take the z-axis parallel to the wire axis and the x-axis parallel to the propagating direction of emitted light which is normal to the wire axis. Here we consider an effective dipole moment $\boldsymbol{d} = (d_x, d_y, d_z)$ of which the electric potential in a free space is equivalent to that produced by $\boldsymbol{d}_0$ outside the nanowire. The relation between $\boldsymbol{d}$ and $\boldsymbol{d}_0$ can be written as

$$d_{x,y} = \frac{2\varepsilon_0}{\varepsilon + \varepsilon_0} d_{0x,y}, \quad d_z = d_{0z} \tag{21.65}$$

The magnitude of the pointing vector of light at the position far from the center of wire by x_0 is given by

$$\begin{aligned} S_{\parallel} &= \frac{a\omega^4}{6\pi\varepsilon_0^2 c^3 x_0}(d_x^2 + 2d_z^2) \\ S_{\perp} &= \frac{a\omega^4}{2\pi\varepsilon_0^2 c^3 x_0} d_y^2 \end{aligned} \tag{21.66}$$

for the components parallel and perpendicular to the wire axis, respectively. Here, c is the speed of light and a is a radius of the wire. Therefore, when the semiconductor is isotropic and $d_{0x} = d_{0y} = d_{0z}$, the ratio of intensity between the emissions polarized parallel and perpendicular to the wire axis is given as

$$\frac{I_{\parallel}}{I_{\perp}} = \frac{S_{\parallel}}{S_{\perp}} = \frac{d_x^2 + 2d_z^2}{3d_x^2} = \frac{(\varepsilon + \varepsilon_0)^2 + 2\varepsilon_0^2}{6{\varepsilon_0}^2} \tag{21.67}$$

The polarization ratio is given by

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} = \frac{(\varepsilon + \varepsilon_0)^2 - 4\varepsilon_0^2}{(\varepsilon + \varepsilon_0)^2 + 8\varepsilon_0^2} \tag{21.68}$$

This indicates that P does not depend on the wire size. Since the dielectric constants of semiconductors are much larger than 1, the dielectric effect affects seriously on the polarization anisotropy. In case that the internal emission is not isotropic such as in the case of the wurtzite-type semiconductors and quantum wires, we have to consider that $d_{0x} = d_{0y} \neq d_{0z}$. In such cases, (21.68) changes to

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} = \frac{(\varepsilon + \varepsilon_0)^2 p - 4\varepsilon_0^2}{(\varepsilon + \varepsilon_0)^2 p + 8\varepsilon_0^2} \tag{21.69}$$

where we define

$$p = \frac{d_{0z}^2}{d_{0x}^2} = \frac{|M_{\parallel}|^2}{|M_{\perp}|^2} \tag{21.70}$$

as a measure of the anisotropic internal emission.

For large nanowires, the above treatment becomes invalid when the diameter of the wire is comparable with the wavelength of emitted light, and the polarization depends on wire size. In this case, the retardation effect should be involved, and we have to solve the Helmholtz equation to find the electromagnetic field in the nanowires instead of the Laplace equation (Ruda and Shik, 2006). Consequently, the following equations are derived corresponding to (21.65)

$$\begin{aligned} d_x &= \sqrt{\varepsilon}\,\frac{J_1'(ka)H_1^{(1)}(ka) - J_1(ka)H_1^{(1)\prime}(ka)}{\sqrt{\varepsilon_0}\,J_1'(ka)H_1^{(1)}(k_0a) - \sqrt{\varepsilon}\,J_1(ka)H_1^{(1)\prime}(k_0a)} d_{0x} \equiv A_x d_{0x} \\ d_z &= \sqrt{\varepsilon}\,\frac{J_1(ka)H_0^{(1)}(ka) - J_0(ka)H_1^{(1)}(ka)}{\sqrt{\varepsilon}\,J_1(ka)H_0^{(1)}(k_0a) - \sqrt{\varepsilon_0}\,J_0(ka)H_1^{(1)}(k_0a)} d_{0z} \equiv A_z\, d_{0z} \end{aligned} \tag{21.71}$$

where J and H are the Bessel and first-kind Hankel functions

$$k = \sqrt{\varepsilon}\,\frac{\omega}{c}$$

$$k_0 = \sqrt{\varepsilon_0}\,\frac{\omega}{c}$$

The coefficient A expresses the effect of image force. It is noted that ε_0 and ε are the relative dielectric constants at the frequency ω, and different from the static values listed in Table 21.1. The relation of the wave numbers of light in the nanowire and surrounding material is written as $k = k_0\sqrt{\varepsilon/\varepsilon_0} = nk_0$, using the refractive index n of the material of the wire. The polarization ratio is given as

$$P = \frac{A_z^2 p - A_x^2}{A_z^2 p + 2A_x^2} \tag{21.72}$$

Figure 21.14 shows polarization ratio P as a function of $k_0a(=(\omega/c)a)$, calculated for various values of ε and p. The polarization is seen to strongly oscillate with k_0a. In the case that $\varepsilon = 4$ and $p = 1$, the points where $P = 0$ are given by $k_0a = 1.21, 1.91, 2.72, 3.5, \ldots$. It is found that the polarization ratio has a value calculated from (21.68), when the radius a is sufficiently smaller than the wavelength of light, λ_0. Then, the polarization ratio P decreases

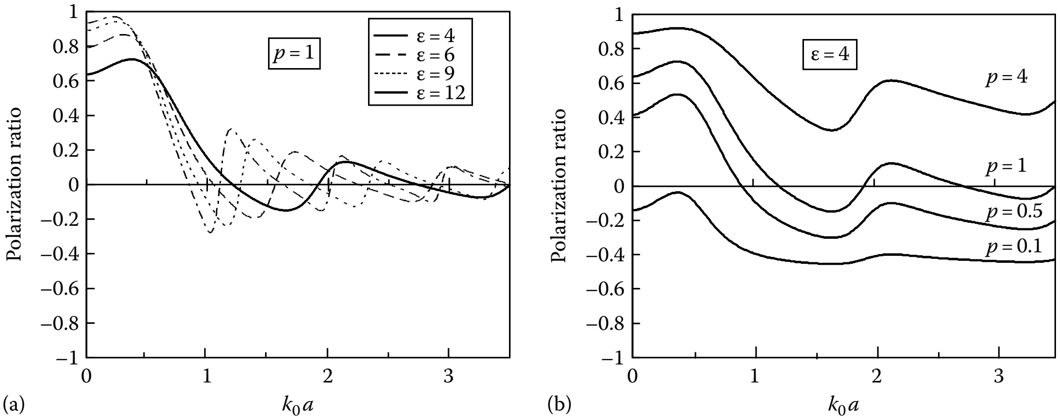

FIGURE 21.14 Polarization ratio as a function of $k_0a(=(\omega/c)a)$ calculated by (21.72) for (a) various values of ε with $p = 1$, and (b) various values of p with $\varepsilon = 4$.

with increasing a, and is equal to zero at $a_1 = 0.193\lambda_0$. With further increasing a, P gradually approaches zero with oscillation. If p is not equal to 1, the asymptotic value of P is to be given by (21.69) for $k_0a = 0$ and $(p - 1)/(p + 2)$ in the limit of $k_0a \to \infty$.

21.5 Application of TEM-CL

21.5.1 InP Nanowires

InP has a direct band gap structure with a gap energy of 1.42 eV at 4 K. Exciton Bohr radius is about 11 nm, so the quantum confinement effects, energy shift of the near-edge emission, and polarization characteristics are expected to appear for nanowires with a diameter smaller than this radius. Exciton binding energy is 5 meV. The emission intensity of exciton in the bulk crystal is weak at higher temperatures, but it can be observed even at room temperature if the quantum effect is strong enough. The size dependence of emission energy by quantum confinement effect has been measured in PL and absorption spectroscopy (Yu et al., 2003), and theoretically treated for cylindrical nanowires (Pellegrini et al., 2005). The PL study by Wang et al. (2001) using a polarized light excitation showed extremely large degree of polarization ($P = 0.96$) along the wire axis in the luminescence from single InP nanowires. The peak shift in PL was small, and the anisotropic polarization was considered to be due to the dielectric effect. Mishra et al. (2007) showed that the PL emission from the InP nanowires of 80 nm in diameter having zinc-blend and wurtzite structures are polarized parallel and perpendicular to the wire axis, respectively.

InP nanowires were grown on InP substrates by the metal-organic vapor phase epitaxy (MOCVPE) technique with gold nanoparticles as the catalyst (Figure 21.15). From the TEM observation, the InP nanowire has the zinc-blend structure of cubic lattice. The wire axis is parallel to the [111] axis, and the nanowire contains many twin boundaries normal to the wire axis (Bhunia et al., 2003). The diameter of the nanowire is distributed in the range from 5 to 30 nm, and the most frequently observed size is 20 nm. The length can be controlled by the growth conditions such as group-III source gas flow rate and total growth time, and that of the present sample is 700 nm. Cross-sectional samples were made by cleaving the sample crystal normal to the (111) substrate surface, and were examined by TEM-CL (Yamamoto et al., 2006).

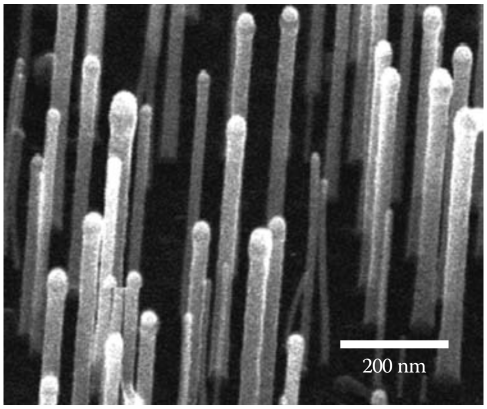

FIGURE 21.15 A SEM image of InP nanowires grown on InP.

Figure 21.15a shows a SEM image of InP nanowires standing straight on the InP substrate with various diameters. The image was taken at a tilt angle of 40°. Figure 21.16a shows a SEM image of a cross-sectional sample where the InP nanowires are standing perpendicular to the incident electron beam. The CL spectra from the nanowires and a substrate region are shown in Figure 21.16b with a common vertical scale, which were taken at room temperature with an electron beam scanning over an 0.5 μm × 0.5 μm area in each region. A broad peak appears in the CL spectrum from the nanowire region at peak energy of 1.50 eV. The band gap energy of the bulk InP is 1.34 eV at room temperature, and this peak energy is much higher than that expected from the quantum confinement effect for the nanowire of 20 nm in diameter. The intensity of the nanowire emission is one order stronger than that of the InP substrate. A monochromatic CL image in Figure 21.16c clearly shows that an intense emission comes from the

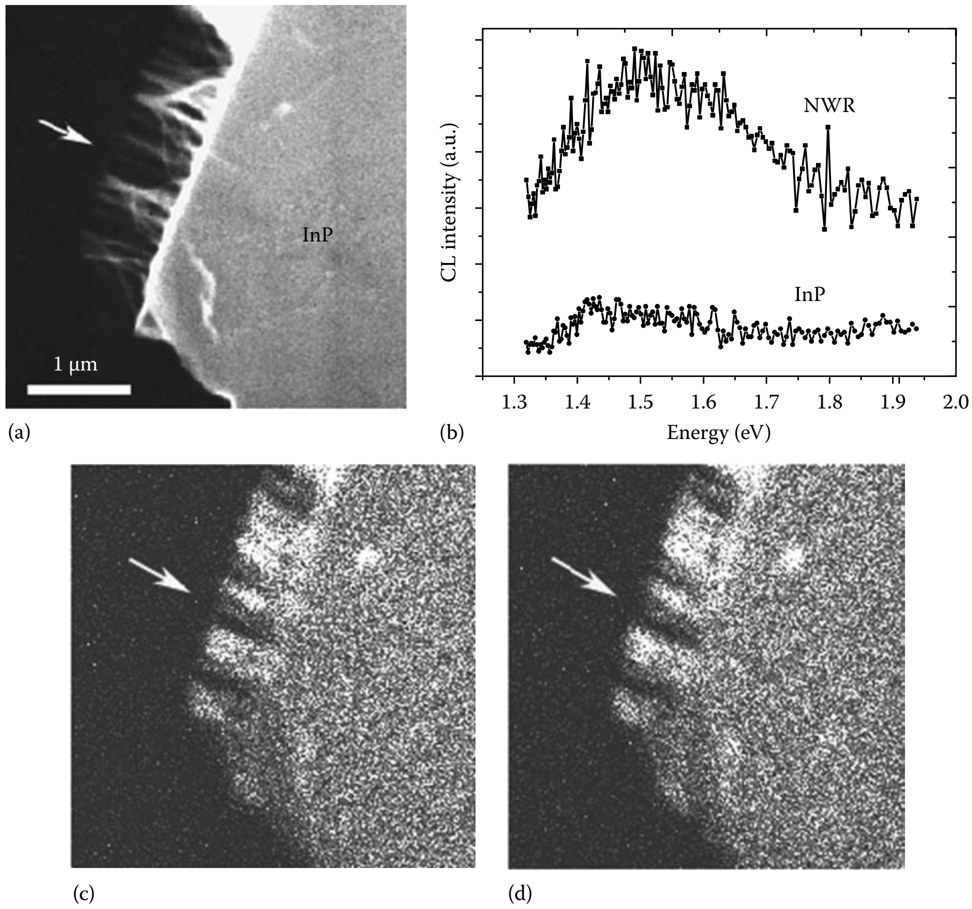

FIGURE 21.16 (a) A SEM image of a cross-sectional sample, (b) room temperature CL spectra from the nanowires (NWRs) and substrate region with a common vertical scale. (c) and (d) are monochromatic CL images taken at photon energies of 1.419 and 1.473 eV, respectively.

nanowires rather than from the substrate at room temperature. Individual nanowire can be resolved in the CL image, as indicated by an arrow. The broad emission peak from the nanowires is considered to be due to the size distribution in wire diameter because the electron confinement energy of nanowire changes with wire diameter. Then the emissions of different energies are expected to come from the nanowires with different diameters. Figure 21.16c and d shows monochromatic CL images taken at photon energies of 1.419 and 1.473 eV, respectively. The images should reveal the distribution of nanowires which produce the emissions of these energies. However, the intensity distribution is almost the same in those images, and the nanowires with different diameters cannot be distinguished. This fact indicates that each nanowire gives the broad peak in the emission spectrum.

The temperature dependence of the emission spectrum from the nanowire is shown in Figure 21.17. The emission intensity of the broad peak does not increase with decreasing temperature from room temperature to 40 K, while that from the InP substrate strongly increases with cooling. The broad peak blueshifts with decreasing temperature and the behavior of the shift follows the Varshni's law. The deviation of the broad peak from the exciton peak in the bulk InP is about 170 meV. For the

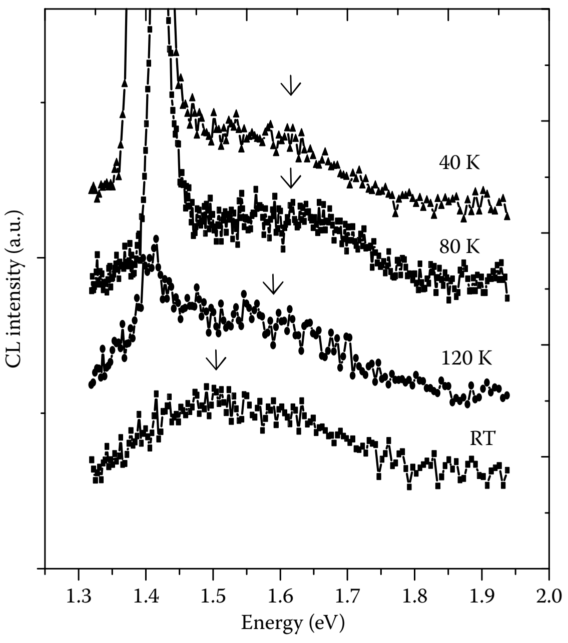

FIGURE 21.17 Emission spectra from the nanowires taken at various temperatures.

recombination of carriers in the ground states ($l = 0$ and $n = 1$), the emission energy shift from the bulk one is roughly estimated from (21.46) to be $\Delta E_{0,1}$ = 31 meV for the InP wire of R = 10 nm with $\mu = 0.071$ m and $j_{0,1} = 2.405$. From the comparison, the observed value is much larger than the value expected from the average diameter of 20 nm. The diameter of the InP quantum wire with the energy shift of 170 meV is estimated to be about 5 nm from the previous data by (Yu et al., 2003). It can be considered that there exists a depletion region near the surface or an oxide layer on the nanowire sidewalls (Dionizio et al., 2008), which decreases a practical size of carrier confinement region.

Figure 21.18a shows CL spectra from the nanowire region and substrate region taken at 80 K, with an electron beam scanning over an 0.5 μm × 0.5 μm area. The emission peak at 1.412 eV comes from the InP substrate with a strong intensity, and the emission from the nanowires shows a strong peak at 1.417 eV. The deviation from the InP peak is 5 meV. This peak shift is much smaller than the value estimated from the quantum confinement effect. It is more likely to be considered that the emission comes from the strained InP substrate region near the surface, because a part of the carriers generated in the nanowires can drift into the substrate.

Figure 21.18b shows CL spectra in the wide energy range extended to more than 1.5 eV. In this spectral range, the CL intensity from the nanowire region is much stronger than that from the InP substrate. A broad peak appears in the spectrum from the nanowire region at 1.60 eV with a full-width half-maximum (FWHM) of 150 meV. Figure 21.18c is a SEM image of the nanowire region, and a corresponding panchromatic CL image taken at 80 K is shown in Figure 21.18d. The InP substrate shows a brighter contrast than the nanowires in this image. On the other hand, in monochromatic CL images taken at the photon energy of 1.55 and 1.632 eV (Figure 21.18e and f), the nanowires show much brighter contrast than the InP substrate. It is noticed that the intensity distributions in Figure 21.18e and f are similar to each other in spite of the existence of the size distribution in the nanowires. This means that the emission peak from each nanowire is sufficiently broad, even at low temperatures. If the broadening of the emission peak due to the phonon scattering process is dominant, the FWHM of the peak should become narrow with decreasing sample temperature, and split peaks from the nanowires with different diameters are expected to be observed. However, any split from the broad peak has not been observed, even for the measurement under a converged electron beam illumination at the nanowire region. One possible

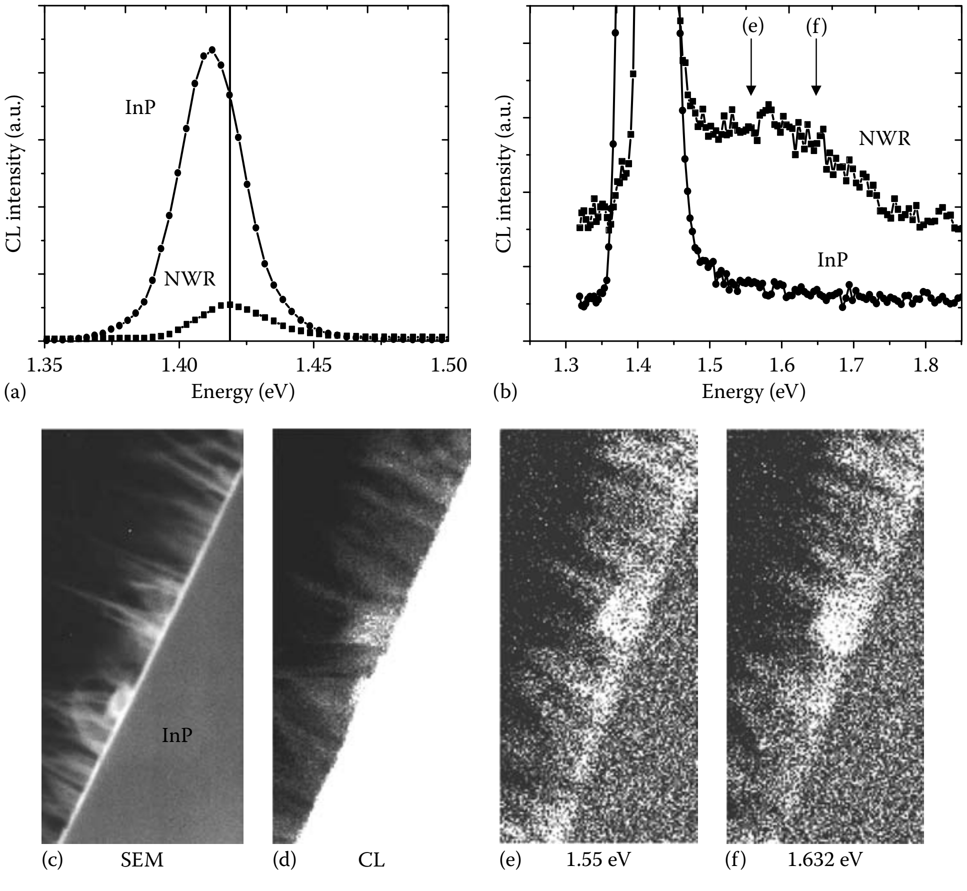

FIGURE 21.18 (a) A CL spectrum from the nanowire region and substrate region taken at 80 K, and (b) a magnified spectrum of (a). (c) A SEM image of the nanowire region and (d) a corresponding panchromatic CL image. (e) and (f) are monochromatic CL images taken at photon energy of 1.55 and 1.632 eV, respectively.

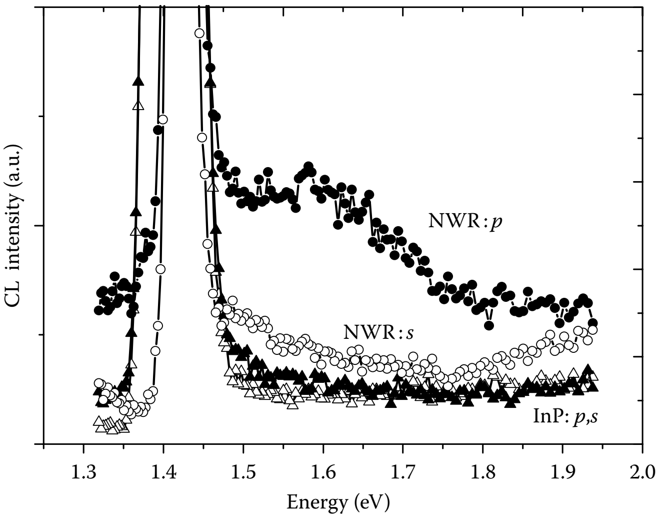

FIGURE 21.19 CL spectra from the nanowire region and substrate region taken for the *p* (solid spots) and *s* (open spots) polarization directions.

explanation is that the broadening of the luminescence peak occurs due to carrier scattering by the impurities or defects such as twin boundaries involved in the nanowire during the growth. In such case, the FWHM is independent of temperature. However, the origin of the broadening is not yet clear.

The polarization dependence of the CL spectrum was measured, as shown in Figure 21.19, where *p*(*s*) polarization corresponds to the electric field oscillation of light parallel (normal) to the nanowire direction. It is found that the luminescent light emitted from the nanowires is highly polarized parallel to the wire direction, while that from the substrate is not polarized. The polarization ratio at the peak energy of 1.6 eV is 0.50. The value expected from the dielectric effect is $P = 0.93$ calculated in (21.68) with $\varepsilon = 12.0$, where the size dependence of the dielectric effect is neglected. In addition, the quantum confinement effect should be involved, i.e., (21.69) is used with $p(>1)$, the expected value must be larger than 0.93. The observed value is much smaller than the expected one. This is partly because the polarization mixing occurs due to reflection by the ellipsoidal mirror which corrects light emitted to a large solid angle. Wang et al. (2001) reported a large degree of polarization ($P = 0.96$) in the photoluminescence from InP nanowires. The PL was excited by a polarized light, so the polarization anisotropy is enhanced in the absorption process. Whereas in the CL experiment, the excitation is performed by the incident electrons, there is no anisotropy in excitation of carriers. Thus, the degree of polarization in the CL measurement is smaller than that in the polarized PL. The CL detection is recently improved by introducing the angler-resolved measurement system, as shown in the later section of ZnO nanowires.

21.5.2 GaAs Nanowire

Gallium arsenide (GaAs) has a zinc-blend structure, with a direct band gap of 1.519 eV at 0 K. Exciton radius is rather large, about 15 nm, so the quantum confinement effect on photon energy and polarization of the FX emission is expected for thin nanowires. The binding energy of FX in the bulk is about 4 meV, and the emission intensity of FX becomes weak at higher temperatures above 200 K, even in thin nanowires. Recently, self-standing nanowires of many semiconductors are produced by a growth technique using nanosized liquid droplets of a metal solvent (Hiruma et al., 1995; Gudiksen et al., 2002; Wu et al., 2002), and optical properties were theoretically studied (Persson and Xu, 2004; Redlinski and Peeters, 2008).

Figure 21.20a shows a SEM image of a GaAs nanowire sample covered by an $Al_{0.4}Ga_{0.6}As$ layer, self-standing on a Si(111) substrate. The average outer dimensions of the nanowires are 80 nm in diameter and 700 nm in height, and the average density of the nanowires is about 10 μm^{-2}. The sample was fabricated by a metalorganic chemical vapor phase epitaxy (MOCVPE) technique using Au nanoparticles as the catalyst (Bhunia et al., 2003; Tateno et al., 2004). First, an $Al_{0.3}Ga_{0.7}As$ nanowire was grown to 100 nm in height on the Si substrate, and then GaAs nanowire and $Al_{0.3}Ga_{0.7}As$ nanowires were successively grown to 100 and 180 nm, respectively. Finally, $Al_{0.4}Ga_{0.6}As$ was deposited as a capping layer, with a thickness of 100 nm to reduce the nonradiative recombination on the nanowire surface. Au dots with a radius of about 20 nm can be observed on top of the nanowires in Figure. 21.20a. Figure 21.20b schematically represents a diagram of the sample structure. The materials are direct-gap

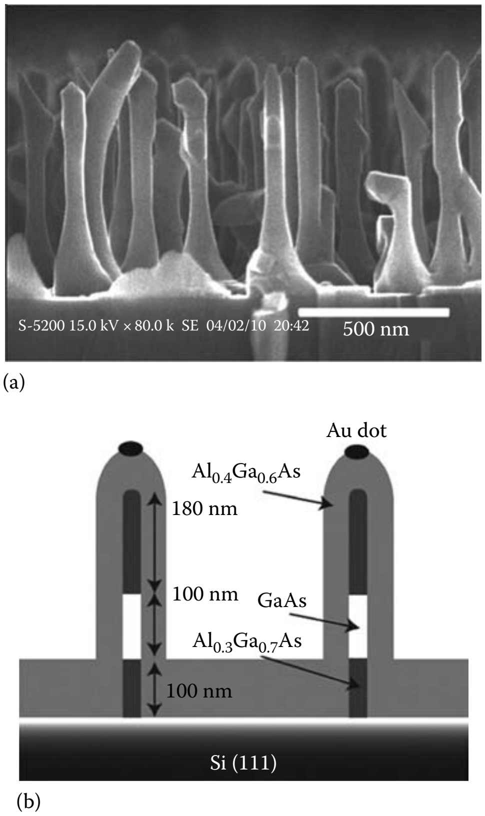

FIGURE 21.20 (a) A SEM image of GaAs/AlGaAs nanowires with a core-shell structure, and (b) schematic diagram of the sample structure.

semiconductors, with energy gaps of 1.52 eV for GaAs, 1.96 eV for $Al_{0.3}Ga_{0.7}As$, and 2.11 eV for $Al_{0.4}Ga_{0.6}As$ at 0 K (Pavesi and Guzzi, 1994). The $Al_{0.4}Ga_{0.6}As$ capping layer acts as a potential barrier, so a stronger (weaker) confinement structure is formed in the GaAs ($Al_{0.3}Ga_{0.7}As$) wire region. Carriers excited in the wire by an electron beam are expected to flow into the GaAs wire region, and recombine at the bottom of the band. Therefore, CL only from GaAs nanowires is expected to have sharp emission peaks in the CL spectrum corresponding to their diameters.

Figure 21.21a shows a SEM image of the sample observed from the direction normal to the surface. The image was observed in the SEM mode using the TEM operated at an accelerating voltage of 80 kV with a scanning area of 1 μm^2 (Ishikawa et al., 2008). Figure 21.21b shows a panchromatic CL image of the same area as that shown in Figure 21.20a taken at 60 K. The localized CL intensity is distributed at the nanowires, and we can specify the emission of each nanowire from the comparison between the CL image and the SEM image. A spectrum from this area is shown in Figure 21.21c, where two broad peaks with a complex shape appear, reflecting the size distribution of the nanowires existing in this area. The emission spectra of similar shapes were frequently observed when the scanning area is of the order of 1 μm^2, although the peak energies are different. The spectrum taken from the wider area becomes a single broad peak extending over the energy range from 1.7 to 2.0 eV. The broadening of the CL peak is due to the superposition of many CL peaks from the nanowires of different diameters. To determine the effect of size distribution, we used a focused electron beam illuminating individual nanowires, A, B, and C indicated by circles in Figure 21.21a. As shown in Figure 21.21d, sharp peaks were observed to appear at different energies because of the difference in diameter of the wire. It is noted that some nanowires such as A and C show double emission peaks.

As for the diameters of the nanowires, it is difficult to determine them from the SEM image since the wires are covered by the AlGaAs layer with outer diameters of about 100 nm. The

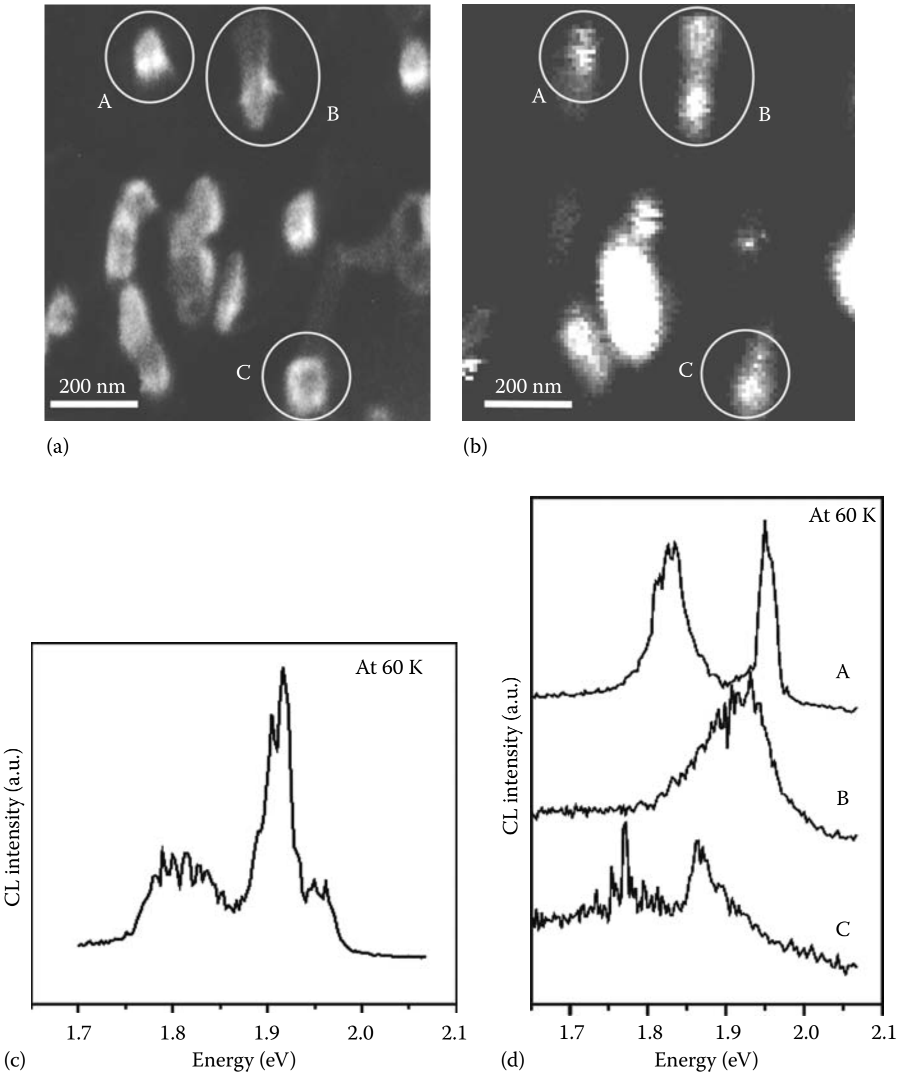

FIGURE 21.21 (a) A SEM image of the GaAs/AlGaAs nanowires, and (b) a panchromatic CL image taken at 130 K. A scan area is 1 μm × 1 μm. (c) A CL spectrum taken from the 1 μm × 1 μm area in (b). (d) CL spectra taken from the individual nanowires marked A, B, and C in (a) and (b).

energy levels were previously theoretically studied by Persson and Xu (2004) for GaAs nanowhisker grown in the [111] direction, on the basis of a tight binding approach. According to the calculation of the transition at the Γ point, the peak energies from 1.7 to 2.0 eV in the CL spectrum from the wide area correspond to the diameters from 5 to 2 nm. These values are much smaller than that expected from a Au particle size of 20 nm. This has been similarly observed in InP self-standing nanowires (Yamamoto et al., 2006), where the peak shift is much larger than that expected from the outer dimensions of the wire. This means that a practical size of carrier confinement region is much smaller than the apparent diameter. Such narrowing of the confinement region may result from the inner stress near the GaAs/AlGaAs interface and interdiffusion of Al into the wire.

The temperature dependence of the CL spectra from the single nanowire A shown in Figure 21.21a is shown in Figure 21.22a. The lower energy peak (P1) appears below 140 K and the higher energy peak (P2) appears below 100 K. Their intensities increase with decreasing temperature. The intensity of the P2 peak increases rapidly at lower temperatures, and exceeds the P1 peak intensity at about 60 K. As shown in Figure 21.22b, the P1 peak energy shows a blue shift as temperature decreases, following the band gap shift described by Varshni's law (solid lines). The P1 emission can be attributed to the radiative recombination of excitons at the ground state of the GaAs wire because the emission peak predominates at low temperatures. The luminescence of excitons remains at rather high temperatures compared with that in bulk GaAs because the binding energy of excitons becomes large in nanowires owing to the quantum confinement effect (Basu, 1997). The appearance of the two peaks can be explained, assuming that the two parts having different energy states coexist along a single nanowire. The inhomogeneous intensity distribution was frequently observed along the tilted wires in the CL images and the polarization of the emitted light was randomly distributed with respect to the wire axis (Ishikawa et al., 2008). These facts indicate that the quantum-dot-like potential structure is formed along the wire, which does not emit a polarized light parallel to the wire axis.

Figure 21.23a shows a SEM image of a GaAs/AlGaAs nanowire with a core-shell structure, as illustrated in Figure 21.20, and Figure 21.23b shows a corresponding panchromatic CL image taken at 130 K. In the CL image, discontinuous contrast is seen to appear along the wire, which suggests the existence of a quantum dot-like structure in the GaAs core. Figure 21.23c is a spectrum image taken with the electron beam scanning along the wire axis. A peak line at a wavelength of 830 nm is attributed to the bulk luminescence of GaAs. This indicates that the carriers excited in the wire diffuse into the GaAs substrate because this peak vanishes for the nanowire separated from the substrate. It is noted that the emissions from those dots have a common spectral shape in the wavelength range from 700 to 800 nm, and the anisotropy in polarization is very small. We still need more consideration about the property of this luminescence.

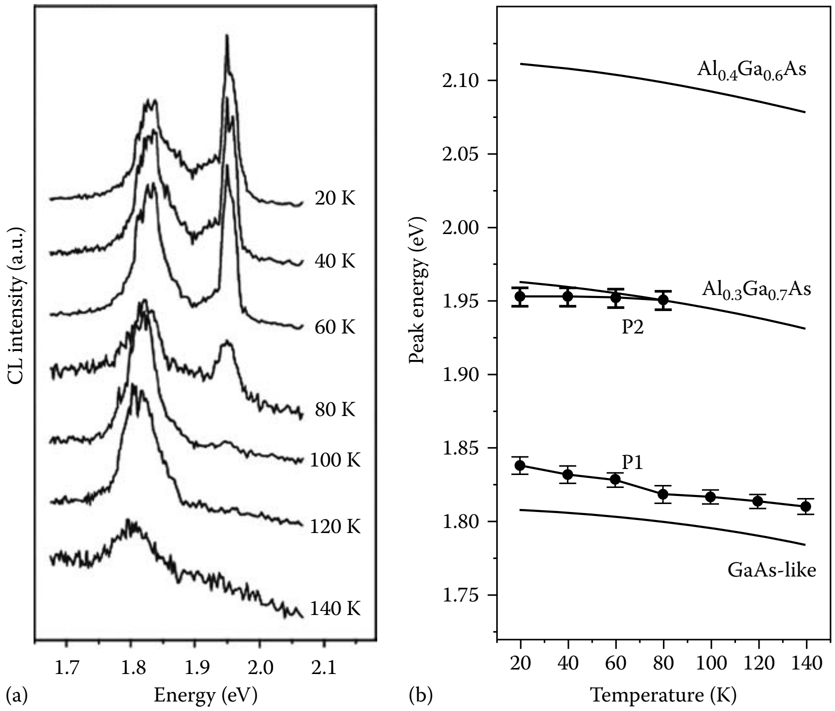

FIGURE 21.22 (a) Temperature dependence of the CL spectrum from the nanowire A in Figure 21.21. (b) Temperature dependence of the peak energies of the two peaks. Solid lines show the temperature dependence of the band gaps of GaAs, $Al_{0.3}Ga_{0.7}As$, and $Al_{0.4}Ga_{0.6}As$, respectively, given by the Varshni's law. The solid curve of GaAs is shifted for easy comparison with the observed one.

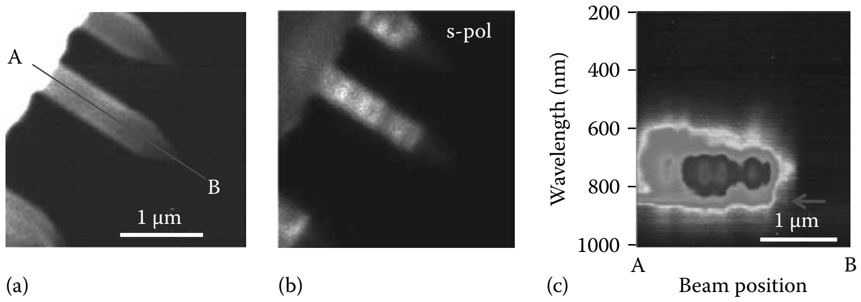

FIGURE 21.23 (a) A SEM image of GaAs/AlGaAs nanowires and (b) a panchromatic CL image of the same area (3 μm × 3 μm) taken at 130 K. (c) A spectrum image taken with the electron beam scanning along a line A-B through the central nanowire in (a).

21.5.3 ZnO Nanowires

Zinc oxide (ZnO) is a wide gap semiconductor of wurtzite structure, with a gap energy of 3.44 eV at 4 K. The valence band structure of bulk ZnO near the BE is composed of three bands (A, B, and C band), and the character of the band has been controversy for many years. The highest energy band (the A band) of many wurtzite-type semiconductors such as GaN and CdSe has a character of heavy hole (the Γ_9 symmetry, as shown in Figure 21.5b), whereas many authors assume the Γ_7 symmetry character for ZnO (Meyer et al., 2004). The polarization studies showed that the inner emission due to the transition from the conduction band to the A band contain both polarization components parallel and perpendicular to the *c*-axis. The energy difference between the A and B band was reported to be very small, 4.9 meV. The exciton binding energy is about 60 meV which is much larger than the thermal energy at room temperature ($k_BT = 25$ meV), so the band-edge emission can be observed even at room temperature. However, the exciton radius is very small ($a_X = 1.3$ nm), and it is difficult to fabricate a ZnO quantum wire of such a small diameter. The quantum effect of ZnO nanowires has rarely been reported so far.

Recently, the optical properties of ZnO nanowires have been studied using self-standing nanowires grown parallel to the *c*-axis by MOCVD technique. Hsu et al. (2004) measured PL from narrow nanowires at room temperature and observed BE emission and green emission originated from defects. They showed that the BE emission is polarized parallel to the *c*-axis with the intensity ratio of $I_{\parallel} : I_{\perp} = 1 : 0.7$, whereas the green emission is polarized perpendicular to the *c*-axis with the ratio of $I_{\parallel} : I_{\perp} = 0.85{:}1$. The polarization dependence of photoconductivity was studied by Fan et al. (2004) using nanowires of 30–150 nm in diameter. The photoconductivity had a maximum value for polarization direction parallel to the *c*-axis. Yu et al. (2003) measured temperature dependence of PL spectrum from ZnO nanowires, showing peaks of BX emission, the *P* emission, and their phonon replica. The nanowires of wurtzite-type semiconductors such as GaN (Zhang and Xia, 2006; Chen et al., 2008) and CdSe (Lan et al., 2008) have also been studied. Chen et al. (2008) measured the size dependence in polarization of the BE emission from GaN nanowires, and explained the results from the dielectric effect and the anisotropy in the inner emission. The polarization ratio changed with temperature; the BE emission is polarized parallel to the *c*-axis at room temperature, and is unpolarized at 20 K. They explained this from the change in anisotropy of the internal emission.

In the TEM-CL study, ZnO smoke particles were made by burning a Zn wire in air, and were collected by a copper mesh coated by a carbon film with micrometer-size holes. The smoke particle has a tetrapod-like shape, with four legs growing along the *c*-axis. Figure 21.24 shows CL spectra taken from a thick leg of the ZnO smoke particle at various temperatures. Two peaks at wavelengths of 368 nm (3.370 eV) and 375 nm (3.307 eV) appear in the spectrum at 20 K. The peak at 368 nm can be attributed to the ionized donor BX emission (Meyer et al., 2004), and the peak

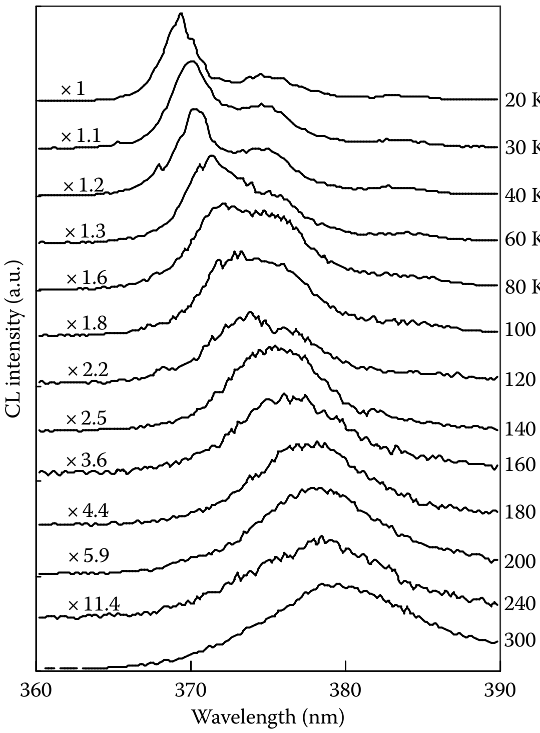

FIGURE 21.24 CL spectra taken from a thick leg of a ZnO smoke particle at various temperatures.

at 375 nm is due to the P emission, which is generated by Auger-like process associated with two excitons (Hvam, 1973; Li et al., 2008). A small peak at 393 nm (3.155 eV) is a phonon replica of the P emission. The photon energies of those peaks shift to lower energies with temperature, following the Varshni's law. A broad single peak (a BE emission) at room temperature can be recognized as a mixed peak composed of the BX emission, P emission, and their phonon replica.

Figure 21.25a shows a SEM image of a ZnO smoke particle with thick legs, and Figure 21.25b and c are monochromatic CL images taken at the peak wavelength at room temperature for the p and s polarizations, respectively. The BE emission from a leg marked A extending along the y direction is strong in the p polarization, whereas that from a leg marked B extending along the x direction is strong in the s polarization. These facts indicate that the BE emission is polarized perpendicular to the c-axis. The central part of the particle shows a dark contrast because the emission peak from this part is slightly shifted to the longer wavelength. The central part showed a bright contrast in the monochromatic CL image taken at 395 nm (Figure 21.25d). This peak shift may be caused by the band gap narrowing due to strain in the multiple twin structure at the central part.

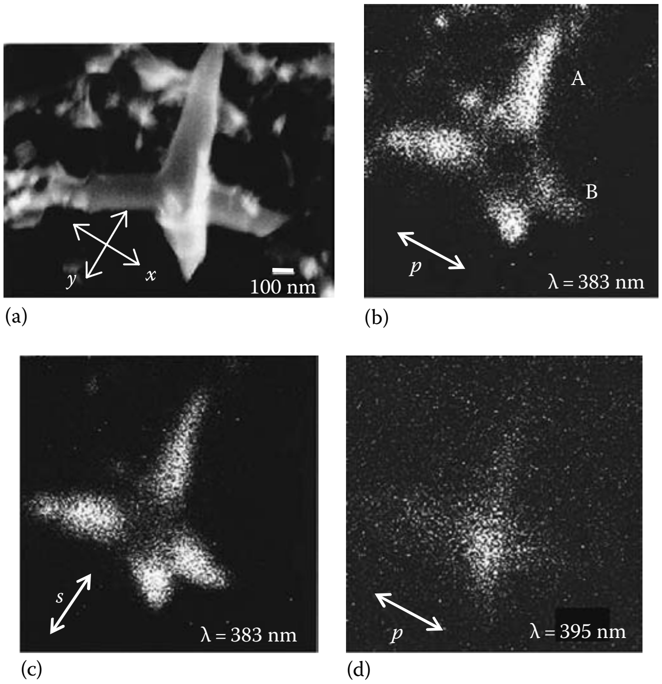

(a) (b)

(c) (d)

FIGURE 21.25 (a) A SEM image of a ZnO smoke particle with thick legs, and (b) and (c) are monochromatic CL images taken at the peak wavelength (383 nm) at room temperature for the p and s polarizations, respectively. (d) A monochromatic CL image taken at 395 nm.

The polarization of the BE emission from ZnO nanowires depends on the diameter of the wire. CL light emitted from a single wire with a large diameter (300 nm) elongating perpendicular to the axis of the parabolic mirror was measured with changing polarization angle θ defined in Figure 21.12. The CL intensity as a function of θ is shown in Figure 21.26a. The light was emitted in a direction perpendicular to the thick wire axis. The solid angle of the detection is about 0.05 str. The intensity of the BE emission at 383 nm has a minimum at $\theta = 0°$, and shows $\cos^2\theta$-dependence. Polarization ratio is estimated to be −0.43 from the intensity ratio of $I_{\|c}/I_{\perp c} = 0.40$, which is deduced using the fitting function of

$$I(\theta) = (I_{\|} - I_{\perp})\cos^2\theta + I_{\perp} \tag{21.73}$$

Whereas the BE emission from a thin nanowire with a diameter of 60 nm presented an opposite behavior in polarization, as shown in Figure 21.26b; the intensity has a maximum at $\theta = 0°$. The polarization ratio is estimated to be 0.31 from the intensity ratio of $I_{\|c}/I_{\perp c} = 1.9$.

The quantum confinement effect cannot affect the polarization property of the CL emission from the ZnO nanowires because the diameters of those wires are much larger than the exciton radius. Therefore, the parameter p related to the anisotropic internal emission is fixed regardless of the diameter of the nanowires, and should be given by the character of the optical transition in the bulk crystal. If the intensity ratio $I_{\|c}/I_{\perp c} = 0.40$ is considered as a bulk one, the parameter p can be calculated from (21.67) and (21.70) as follows:

$$p = \frac{{d_{0z}}^2}{{d_{0x}}^2} = \frac{1}{2}\left\{3\left(\frac{I_{\|}}{I_{\perp}}\right) - 1\right\} = 0.1 \tag{21.74}$$

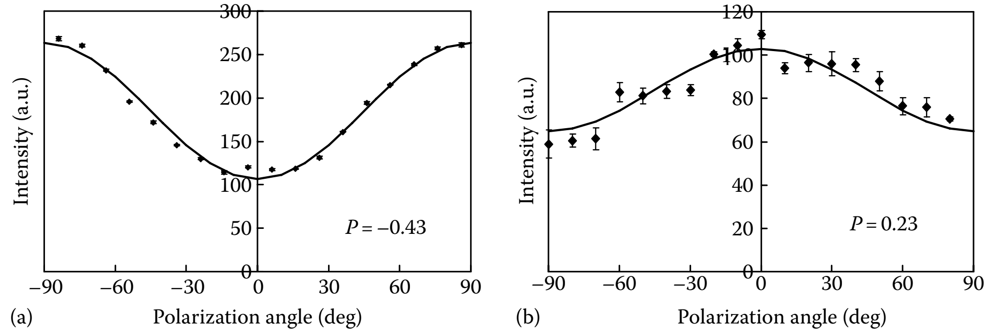

(a) (b)

FIGURE 21.26 CL intensity from (a) a thick wire (diameter of 300 nm) and (b) a thin wire (60 nm) of ZnO measured by changing an angle θ.

If the highest energy state in the valence band has a character of the light hole, $p = 4$ is expected from (21.58) for the transition from the conduction band to valence band. The above value is too small compared to the expected one. In other words, the emission due to the transition to the light hole state should be mainly polarized parallel to the *c*-axis, but the observed one is rather polarized perpendicular to the *c*-axis. This fact indicates that the transition to the heavy hole state component in the A band considerably contributes to the CL emission. It was also observed that the polarization ratio does not change with temperature from 20 to 300 K.

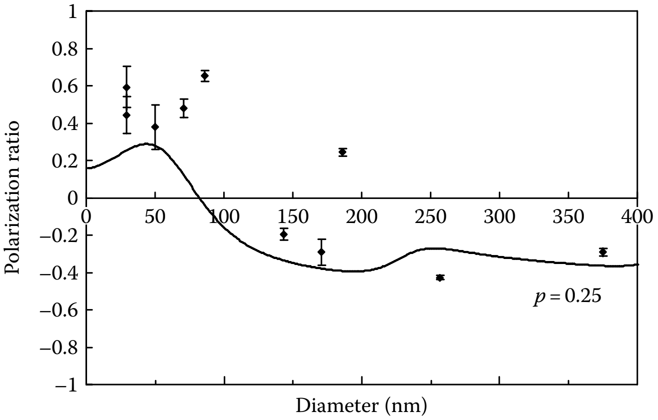

FIGURE 21.27 Dependence of the polarization ratio on nanowire diameter measured from several nanowires with various diameters. Solid line is a calculated curve using the dielectric constant at the peak wavelength of 4.0 and $p = 0.25$.

The CL emission from the thick nanowire is polarized perpendicular to the wire axis (the *c*-axis), whereas that from the thin nanowire is polarized parallel to the wire axis. We should consider the dielectric effect for explaining this size dependence of the polarization. Figure 21.27 shows the dependence of the polarization ratio on nanowire diameter calculated from (21.71) and (21.72) with the dielectric constant at the peak wavelength to be 4.0 and with $p = 0.25$. This result well explains the behavior that the polarization ratio P is positive for thin nanowires, and turns to be negative for thick nanowires. The experimental values measured from ZnO nanowires with different diameters are plotted in Figure 21.27. The polarization ratio changes sign around a diameter of 100 nm, being fitted with the calculated curve.

The spatial variation of polarization can be measured from spectral images taken with scanning electron beam. Figure 21.28a and b shows a SEM image of a ZnO smoke particle with thick legs and a panchromatic CL image, respectively. Change in CL intensity appears along the leg. Figure 21.28c and d are *p*- and *s*-polarized spectral images taken with scanning electron beam along the line A-B in Figure 21.28a. The peak wavelength is 383 nm at room temperature. Polarization ratio calculated from the division of the peak intensities in the images of Figure 21.28c and d is shown in Figure 21.28e. The polarization ratio is seen to be −0.45 at a thinner part (D = 150 nm) and −0.25 at a thicker part along the line. One of the legs is standing vertically near the position B, and the polarization is vanished there.

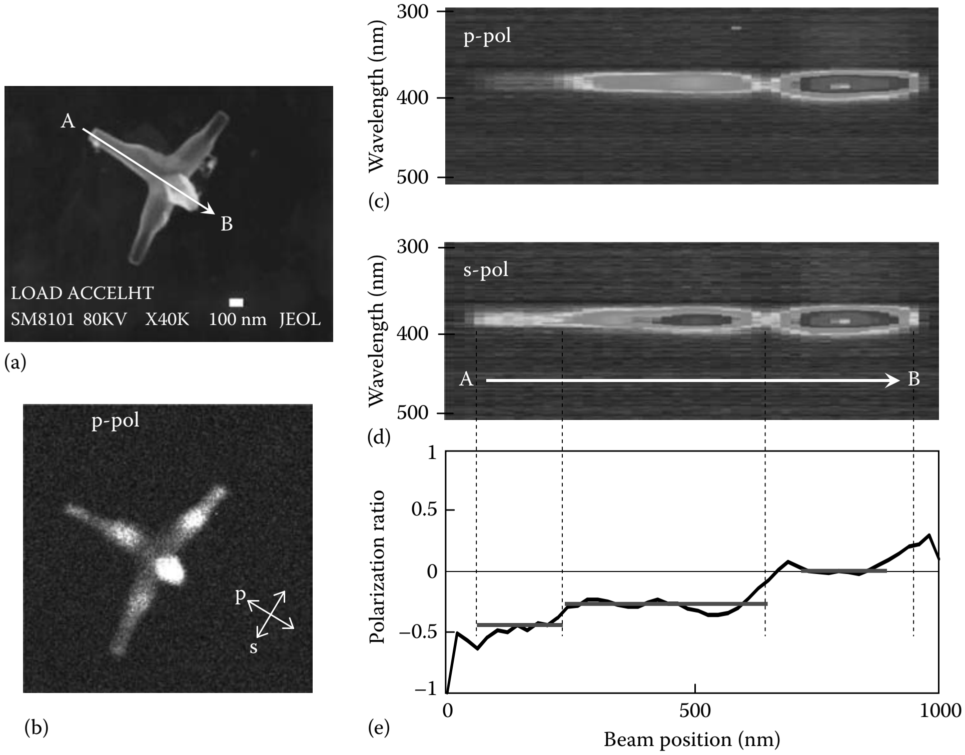

FIGURE 21.28 (a) A SEM image of a ZnO smoke particle and (b) a panchromatic CL image, respectively. (c) and (d) are *p*- and *s*-polarized spectral images taken with scanning electron beam along the line A-B in (a). (e) Spatial variation of the polarization ratio along the line.

This behavior is qualitatively explained by the theoretical curve in Figure 21.27.

Other types of emission appear under a high excitation condition, when a thin nanowire is illuminated by a strong electron beam. The *P* emission is one of such types. Emission peak due to the exciton molecules appears near the BE peak, and shifts to the lower energies with increasing electron beam density. The emission peak due to the electron-hole plasma also appears at the lower energy side of the BE peak, and becomes predominant under the high excitation rate. The effect of shape and size of nanowires on the property of these emissions such as threshold condition of lasing has been studied recently (Johnson et al., 2003).

Acknowledgment

The author acknowledges Y. Watanabe, K. Tateno, G. Salviati, and L. Lazzarini for the supply of nanowire samples. This work has been supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Ando, H, Oohashi, H, and Kanbe, H, 1991, *J. Appl. Phys.* 70, 7024.

Arakawa, Y and Sakaki, H, 1982, *Appl. Phys. Lett.* **40**, 939.

Basu, PK, 1997, *Theory of Optical Processes in Semiconductors*, Oxford University Press, New York.

Bockelmann, U and Bastard, G, 1992, *Phys. Rev. B* **45**, 1688.

Bhunia, S, Kawamura, T, Watanabe, Y, Fujikawa, S, and Tokushima, K, 2003, *Appl. Phys. Lett.* **83**, 3371.

Chen, H-Y, Yang, Y-C, Lin, H-W, Chang, S-C, and Gwo, S, 2008, *Opt. Express* 16(17), 13465.

Dionizio, M, Venezuela, P, and Schmidt, TM, 2008, *Nanotechnology* 19, 065203.

Fan, Z, Chang, P-C, Lu, JG, Walter, EC, Penner, RM, Lin, C-H, and Lee, HP, 2004, *Appl. Phys. Lett.* 85(25), 6128.

Goldstein, JI, Costley, JL, Lorimer, GW, and Reed, SJB, 1977, *Scanning Electron Microscopy*, Vol. 1, pp. 315, IITRI, Chicago, IL.

Gudiksen, MS, Lauhon, LJ, Wang, J, Smith, DC, and Lieber, CM, 2002, *Nature* **415**, 617.

Guilet, T, Grousson, R, Voliotis, V, Wang, X-L, and Ogura, M, 2003, *Phys. Rev. B* **68**, 045319.

Hiruma, K, Yazawa, M, Katsuyama, T, Ogawa, K, Haraguchi, K, Koguchi, M, and Kakibayashi, H, 1995, *J. Appl. Phys.* **77**, 447.

Hsu, NE, Hung, WK, and Chen, YF, 2004, *J. Appl. Phys.* **96**(8), 4671.

Hvan, JM, 1973, Solid State Com. **12**, 95.

Hensel, JC and Feher, G, 1963, *Phys. Rev.* **129**, 1041.

Ils, P, Michel, M, Forchel, A, Gyuro, I, Klenk, M, and Zielinski, E, 1994, *Appl. Phys. Lett.* **64**, 496.

Ino, N and Yamamoto, N, 2008, *Appl. Phys. Lett.* **93**, 232103.

Ishikawa, K, Yamamoto, N, Tateno, K, and Watanabe, Y, 2008, *Jpn. J. Appl. Phys.* **47**(8), 6596.

Johnson, JC, Yan, H, Yang, P, and Saykally, RJ, 2003, *J. Phys. Chem. B* **107**, 8816.

Kane, EO, 1957, *J. Phys. Chem. Solids.* **1**, 249.

Kapon, E, Hwang, DM, and Bhat, R, 1989, *Phys. Rev. Lett.* **63**, 430.

Lan, A, Giblin, J, Protasenko, V, and Kuno, M, 2008, *Appl. Phys. Lett.* **92**, 183110.

Landau, LD, Lifshitz, EM, and Pitaevskii, LP, 1984, *Electrodynamics of Continuous Media*, Pergamon, Oxford, U.K.

Li, W, Gao, M, Chen, R, Zhang, X, Xie, S, and Peng, L-M, 2008, *Appl. Phys. Lett.* **93**, 023117.

Luttinger, JM and Kohn, W, 1955, *Phys. Rev.* **97**, 869.

McIntyre, CR and Sham, LJ, 1992, *Phys. Rev.* **45**, 9443.

Meyer, BK, Alves, H, Hofmann, DM et al., 2004, *Phys. Stat. Sol. (b)* **241**(2), 231.

Mishra, A, Titova, LV, Hoang, TB et al., 2007, *Appl. Phys. Lett.* **91**, 263104.

Mitsui, T, Yamamoto, N, Tadokoro, T, and Ohta, S, 1996, *J. Appl. Phys.* **80**, 6972.

Nakaji, D, Grillo, V, Yananoto, N, and Mukai, T, 2005, *J. Elect. Microscopy.* **54**, 223.

Ogawa, T and Takagahara, T, 1991, *Phys. Rev. B* **43**, 14325.

Pankove, JI, 1971, *Optical Processes in Semiconductors*, Dover Pub. Inc., New York.

Pavesi, L and Guzzi, M, 1994, *J. Appl. Phys.* **75**, 4779.

Pellegrini, G, Mattei, G, and Mazzoldi, P, 2005, *J. Appl. Phys.* **97**, 073706.

Persson, MP and Xu, HQ, 2004, *Nano Lett.* **4**, 2409.

Pfeiffer, L, West, KW, Stormer, HL, Eisenstein, JP, Baldwin, KW, Gershoni, D, and Spector, J, 1990, *Appl. Phys. Lett.* **56** 1697.

Pollak, FH and Cardona, M, 1968, *Phys. Rev.* **172**, 816.

Redlinski, P and Peeters, FM, 2008, *Phys. Rev. B* **77**, 075329.

Ruda, HE and Shik, A, 2005, *Phys. Rev. B* **72**, 115308.

Ruda, HE and Shik, A, 2006, *J. Appl. Phys.* **100**, 024314.

Sakaki, H, 1980, *Jpn. J. Appl. Phys.* **19**, L735.

Sercel, PC and Vahala, KJ, 1990, *Appl. Phys. Lett.* **57**, 545.

Someya, T, Akiyama, H, and Sakaki, H, 1995, *Phys. Rev. Lett.* **74**, 3664.

Tanaka, M and Sakaki, H, 1989, *Appl. Phys. Lett.* **54**, 1326.

Tanatar, B, Al-Hayek, I, and Tomak, M, 1998, *Phys. Rev. B* **58**, 9886.

Tateno, K, Gotoh, H, and Watanabe, Y, 2004, *Appl. Phys. Lett.* **85**, 1808.

Tsuchiya, M, Gaines, JM, Yan, RH, Simes, RJ, Holtz, PO, Coldren, LA, and Petroff, PM, 1989, *Phys. Rev.* **62**, 455.

Voos, M, Leheny, RF, and Shah, J 1980, Radiative recombination, in *Handbook on Semiconductors: Vol. 2 Optical Properties of Solids*, M. Balkanski (ed.), North-Holland, Amsterdam, the Netherlands.

Vugaftman, I, Meyer, JR, and Ram-Mohan, LR, 2001, *Appl. Phys. Rev.* **89**, 5815.

Wang, J, Gudiksen, MS, Duan, X, Cui, Y, and Lieber, CM, 2001, *Science* **293**, 1455.

Wang, X-L and Voliotis, V, 2006, *J. Appl. Phys.* **99**, 121301.

Wu, Y, Fan, R, and Yang, P, 2002, *Nano Lett.* **2**, 83.

Yananoto, N, Spence, JCH, and Fatty, D, 1984, *Phil. Mag. A* **49**, 609.

Yamamoto, N 2002, Development of CL for semiconductor research I: EM-CL study of microstructures and defects in semiconductor epilayers, in *Nanoscale Spectroscopy and Its Applications to Semiconductor Research*, Watanabe Y et al. (eds.), *Lecture Notes in Physics*, Vol. 588, Springer Verlag, Berlin, Germany.

Yamamoto, N, 2008, TEM-Cathodoluminescence study of semiconductor quantum dots and quantum wires, in *Beam Injection Based Nanocharacterization of Advanced Materials*, G. Salviati et al. (eds.), Research Signpost, Kerala, India.

Yamamoto, N, Itoh, H, Grillo, V, et al., 2003, *J. Appl. Phys.* **94**, 4315.

Yamamoto, N, Bhunia, S, and Watanabe, Y, 2006, *Appl. Phys. Lett.* **88**, 153106.

Yamamoto, N, Ishikawa, K, Akiba, K, Bhunia, S, Tateno, K, and Watanabe, Y, 2008, TEM-Cathodoluminescence study of semiconductor quantum dots and quantum wires, in *Beam Injection Based Nanocharacterization of Advanced Materials*, G. Salviati et al. (eds.), Research Signpost, Kerala, India.

Yu, H, Li, J, Loomis, RA, Wang L-W, and Buhro, WE, 2003, *Nat. Mater.* **2**, 517.

Yu, PY and Cardona, M 1999, *Fundamentals of Semiconductors*, Springer Verlag, Berlin, Germany.

Zhang, XW and Xia, JB, 2006, *J. Phys. Condens. Matter* **18**, 3107.

Zhao, X, Wei, CM, Yang, L, and Chou, MY, 2004, *Phys. Rev. Lett.* **92**(23), 236805.

22

Optical Spectroscopy of Nanomaterials

22.1 Introduction 22-1
22.2 Carbon Nanotubes 22-2
22.3 Nanoparticle Quantum Dots 22-4
22.4 Multiexciton Generation 22-6
22.5 Summary 22-7
Acknowledgments 22-7
References 22-7

Yoshihiko Kanemitsu
Kyoto University

22.1 Introduction

Over the past two decades, there have been extensive studies on the optical properties of semiconductor nanomaterials from the fundamental physics viewpoint and from the interest in the application to functional devices, because they exhibit unique size-dependent quantum properties [1–11]. In this chapter, we discuss optical properties of semiconductor nanomaterials of zero-dimensional (0D) nanoparticle quantum dots and one-dimensional (1D) carbon nanotubes. In optical studies of nanoparticle quantum dots and carbon nanotubes, we would like to point out two important reports opening new active fields: the discovery of room-temperature-visible luminescence from porous silicon in 1990 [12] and the discovery of efficient luminescence from isolated carbon nanotubes in 2002 [13]. These observations of efficient luminescence clearly show that nanoparticles and carbon nanotubes are high-quality crystalline semiconductors. Many different fabrication methods have been developed to obtain stable and efficient luminescence from nanoparticles and carbon nanotubes, e.g., core/shell nanoparticles, suspended isolated nanotubes, and so on [14–20]. These nanomaterials become new materials for optoelectronic devices such as wavelength-tunable light-emitting diodes and lasers, quantum light sources, and solar cell applications.

When semiconductor nanoparticles of sizes are comparable to or smaller than the exciton Bohr radius in bulk crystals, the excited state energies and optical properties are very sensitive to their sizes [1,21,22]. Usually, nanoparticle samples are an inhomogeneous system in the sense that they have a distribution of size and shape, and variations of surface structures and surrounding environments [2]. A large nanoparticle has small band-gap energy and a small nanoparticle has large band-gap energy. Furthermore, the nanoparticles have large surface-to-volume ratios, and then the optical properties of nanoparticles are also sensitive to surrounding environments. The exciton band-gap energy of semiconducting carbon nanotubes is also sensitive to the nanotube diameter and the chiral index. Then, we need to study the intrinsic optical processes in nanoparticle quantum dots and carbon nanotubes hidden by sample inhomogeneity using sophisticated optical spectroscopy.

If the laser light excites all nanoparticles or all nanotubes in the sample, the sample shows broad luminescence, reflecting size or diameter distributions. This "global" photoluminescence (PL) or nonresonantly excited luminescence contains contributions from all nanoparticles or all nanotubes in the sample, and the PL spectrum is inhomogeneously broadened, as shown in Figure 22.1a. These sample inhomogeneities are the origin of nonexponential PL decay. In inhomogeneously broadened systems, resonant excitation spectroscopy is a powerful method to obtain intrinsic information from broadened optical spectra. Under resonant excitation at energies within the global PL band, we can observe fine structures in PL spectra at low temperatures, as shown in Figure 22.1b. Resonant excitation at energies within the luminescence band results in a single zero-phonon PL line or a well-resolved phonon progression in PL spectra at low temperatures. In this case, we suppress the inhomogeneous broadening of the luminescence by selectively exciting a narrow subset of nanoparticles. Resonant excitation results in fluorescence line narrowing (FLN) in nanoparticle samples [23–27]. The resonantly excited PL spectra are sensitive to the nature of the band-edge structure and the surface structure [27–29]. Moreover, luminescence hole-burning (LHB) spectroscopy is another resonant excitation spectroscopy. In the LHB experiments, the sample is excited by intense laser at the energy within the PL band [30–32]. After prolonged laser irradiation (burning laser excitation), a spectral hole is formed near the burning laser energy in the luminescence

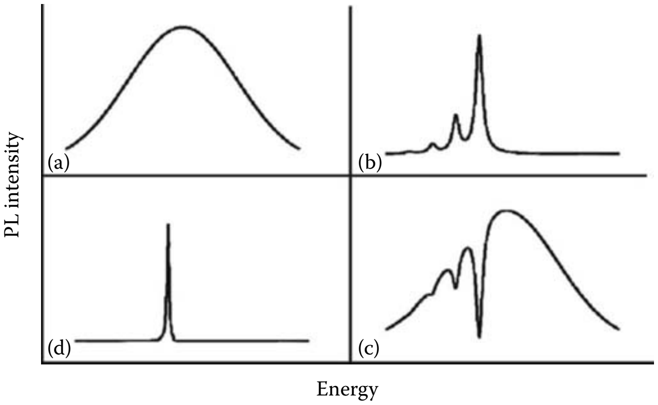

FIGURE 22.1 Luminescence spectra of semiconductor nanoparticles: (a) global luminescence spectrum, (b) resonantly excited luminescence spectrum, (c) LHB spectrum, and (d) single nanoparticle luminescence spectrum.

band. We obtain similar optical information from the FLN and LHB experiments. However, in the resonantly excited PL or FLN spectra, it is difficult to obtain the detailed spectrum near the excitation energy because of the scattering of the excitation laser light. In the LHB experiment, it is comparatively easy to observe the spectral change just at the excitation energy [32].

Single molecule (or nanomaterial) spectroscopy is the most powerful tool for understanding intrinsic optical properties of isolated molecules and nanomaterials [5,33]. Single nanoparticle spectroscopy makes it possible to probe inherent and novel optical properties in nondoped and impurity-doped semiconductor nanoparticles hidden by inhomogeneity, such as size distributions and surrounding environment variations [34–37]. As an example of unique optical phenomena, PL blinking (or PL intermittency) is revealed by single nanoparticle spectroscopy [35]. We have developed many different types of luminescence imaging microscopes. Our systems cover a wide spectral region and wide response times. The spatial resolution is typically about 1 μm in our confocal microscopes and about 50–100 nm in home-built near-field optical microscopes. In these space-resolved optical measurements, the most important point is the fabrication of stable and isolated nanomaterial samples. The nanomaterials should be isolated from each other, and the number density is dilute (about 1/μm^2 or less). The sample should be stable against oxidation and strong light illumination. Then, chemically synthesized materials are good samples for our optical studies. Here, we discuss the luminescence properties and the mechanism of spectral diffusion and PL blinking of single carbon nanotubes and nanoparticle quantum dots.

22.2 Carbon Nanotubes

A single-walled carbon nanotube (SWNT) with about 1 nm diameter and a length greater than several hundred nanometers is a prototype of 1D structures. The recent discovery of efficient PL from semiconducting SWNTs [13,38] has stimulated considerable efforts in understanding optical properties of SWNTs. The semiconducting SWNTs are 1D direct-gap band structures [8]. Because of the extremely strong electron–hole interactions (excitonic effects) in 1D materials, unique optical properties of SWNTs are determined by the dynamics of 1D excitons [8,39]. In addition, the electronic structure and the PL energy of SWNTs strongly depend on the diameter and the chiral index [38]. The SWNT samples are also inhomogeneous systems, similar to the nanoparticle samples, because many different species of nanotubes exist in the sample. The inhomogeneous broadening and the spectral overlapping of PL spectra cause the complicated PL dynamics of SWNTs. Single nanotube spectroscopy reveals the intrinsic excitonic properties of SWNTs [40–46], such as exciton energy, bright and dark exciton structures, exciton–phonon interaction, and so on.

For single nanotube spectroscopy, we synthesized spatially isolated carbon nanotubes on Si substrates using an alcohol catalytic chemical vapor deposition method [45,46]. In our experiments, the Si or SiO$_2$ substrates were patterned with parallel grooves, typically 500 nm wide and 500 nm deep using an electron-beam lithography technique. The isolated SWNTs grow from one side toward the opposite side of the groove. We used these SWNT samples without matrix and surfactant around the nanotubes to reduce the local environmental fluctuation effect. We show a typical PL spectrum of a single carbon nanotube suspended on the groove [assigned chiral index: (7,6)] at about 40 K in Figure 22.2a [46]. Very broad PL bands are observed in the ensemble-averaged spectrum of micelle-wrapped SWNTs dispersed in gelatin. The PL spectral shape of a single carbon

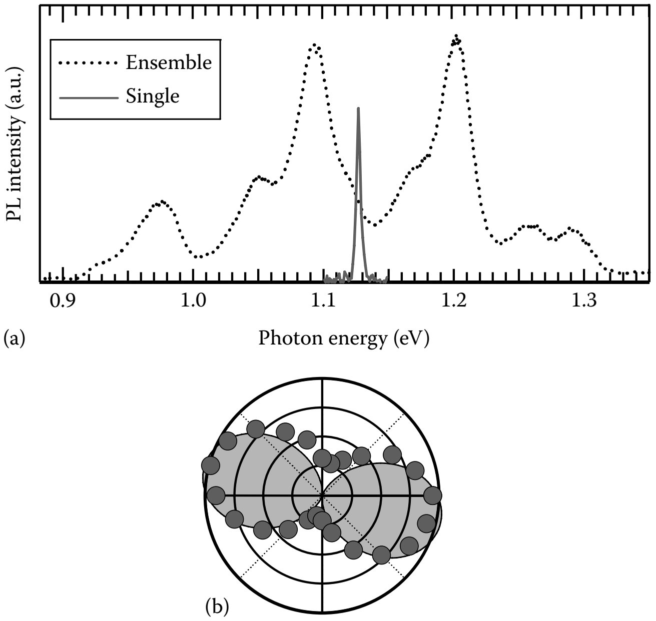

FIGURE 22.2 (a) PL spectrum of a typical suspended single SWNT [assigned chiral index (7,6)] in comparison with the ensemble-averaged spectrum of micelle-wrapped SWNTs dispersed in gelatin. (b) Polar plot of the PL intensity of a typical single SWNT versus the polarized direction of the excitation laser. The PL data (circle) were fitted using $\cos^2\theta$ (solid line). (Reprinted from Matsunaga, R. et al., *Phys. Rev. Lett.*, 101, 147404, 2008. With permission.)

nanotube is given by a Lorentzian function, and its linewidth of a few meV reflects homogeneous broadening. Figure 22.2b shows a polar plot of the PL intensity of a typical single carbon nanotube versus the polarization direction of the excitation laser light [46]. Since a 1D dipole moment exists, strong optical absorption occurs when the polarization of the excitation light parallels the nanotube axis. This PL anisotropy is useful for determining the direction of the observed nanotube axis for single nanotube spectroscopy and modulation spectroscopy.

The sharp luminescence spectra provide detailed information on the exciton fine structures. We studied the PL fine structure of single SWNTs under magnetic fields at low temperatures. A single sharp PL spectrum arising from bright exciton recombination is observed at zero magnetic field, as shown in Figure 22.2a. When the magnetic field is parallel to the nanotube axis, a new peak appears below the bright exciton peak. Figure 22.3a shows PL spectra of a single carbon nanotube under magnetic fields [46]. These PL spectra are fit well by two Lorentzian functions. With the magnetic field, the lower energy peak shows a redshift and the lower-energy peak intensity increases. We cannot observe these changes when the magnetic field is perpendicular to a single nanotube axis, as shown in Figure 22.3b [46]. The splitting of the PL peak occurs due to the magnetic flux parallel to the nanotube axis. These splitting and magnetic field dependence can be explained by the Aharonov–Bohm splitting of excitons based on the Ajiki and Ando model [8,47]. The singlet exciton states split into the bonding and antibonding exciton states, and this due to the short-range Coulomb interaction. The bonding state is odd parity (bright) and the antibonding is even (dark). The energy difference between the bright and dark exciton states also depends on the diameter of SWNTs. These experimental observations are consistent with the theoretical calculation. The dark exciton state exists about several meV below the bright exciton state. Studies of 1D dark excitons influencing optical responses of carbon nanotubes [48–52] are very important for optical device applications.

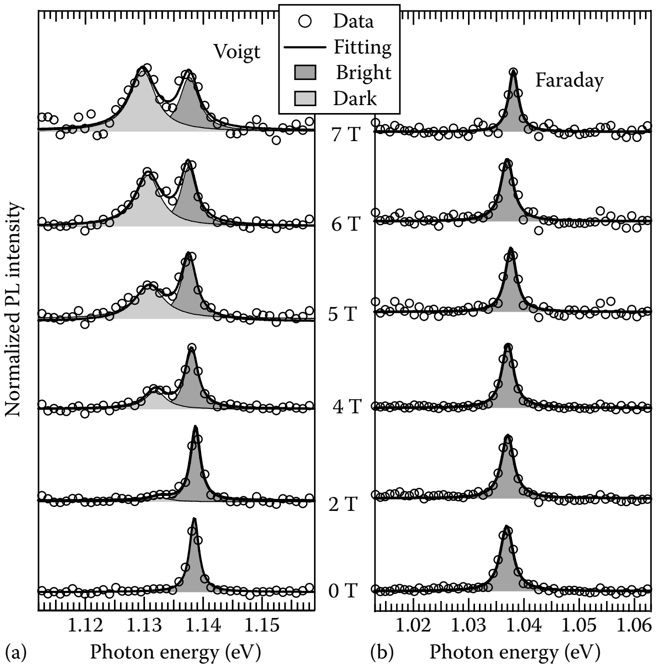

FIGURE 22.3 (a) Normalized magneto-PL spectra of a single (9,4) carbon nanotube at 20 K in the Voigt geometry. The split PL spectra are fit by two Lorentzian functions. (b) The normalized magneto-PL spectra of a single (9,5) carbon nanotube at 20 K in the Faraday geometry. The PL spectra are fit by a Lorentzian function. (Reprinted from Matsunaga, R. et al., *Phys. Rev. Lett.*, 101, 147404, 2008. With permission.)

The diameter dependence of the exciton energy in single carbon nanotubes is also revealed by single nanotube spectroscopy. At room temperature, the experimentally obtained PL spectra can be approximately reproduced by single Lorentzian functions. The observed PL peaks correspond to the zero-phonon lines of free excitons, and the spectral linewidth of the PL spectra is determined by the homogeneous broadening. We obtained PL spectra from many different isolated SWNTs with a variety of chiral indices. Figure 22.4a shows a distribution of the PL peak energies for the single SWNTs, indicated by diamonds [45]. In Figure 22.4b, we show some of the PL spectra from isolated SWNTs with various emission energies [45]. Only a single sharp peak can be seen in each spectrum. The PL linewidth clearly becomes broader as the diameter decreases. This shows that the exciton–phonon interaction is stronger in smaller diameter tubes.

The lowest exciton has fine structures (bright and dark excitons), and the fine structures will determine optical responses and cause unique phenomena. At low temperatures, we observe an interesting phenomenon, spectral diffusion. A few ten percents

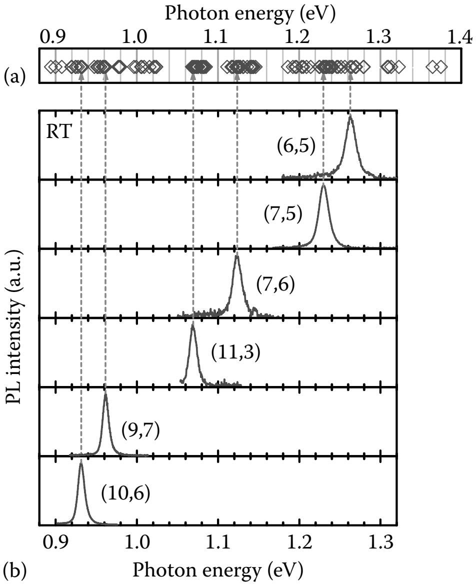

FIGURE 22.4 (a) PL peak energy distribution of obtained PL spectra from about 180 different isolated SWNTs. (b) PL spectra for several species of single SWNTs at room temperature. SWNTs with higher PL emission energy tend to have a larger spectral linewidth. (Reprinted from Inoue, T. et al., *Phys. Rev. B*, 73, 233401, 2006. With permission.)

of the nanotubes show spectral diffusion. We consider that spectral diffusion is related to the exciton fine structure, bright and dark excitons. The PL fluctuation due to spectral diffusion is clearly observed at low temperatures. Spectral fluctuation occurs very slowly, the order of several seconds. Figure 22.5a shows a typical temporal evolution of the PL spectrum of SWNT showing spectral fluctuations at 40 K [53]. During spectral diffusion, the PL spectra clearly show two peaks. The lower energy is fitted by Gaussian and the higher energy is Lorentzian. From spectral fitting, we can determine the peak positions and linewidth of both peaks of the PL spectra. Temporal changes in two PL peak energies and linewidths of the lower energy PL band are shown in Figure 22.5b [53]. The higher energy peak is almost constant, but the lower energy peak fluctuates. We find a good correlation between the PL peak energy and the PL linewidth of the lower energy band. When the PL peak shows a low energy, the linewidth becomes wider. To clarify the origin of the spectral diffusion in the lower energy peak, the PL linewidth of the lower energy peaks is plotted as a function of the emission energy. We obtain the square root dependence of the linewidth on the emission energy [53]. This dependence suggests the quantum-confined Stark effect [54]. The spectral diffusion can be explained by the fluctuation of local electric field. The Stark effect causes a redshift in the exciton energy. A small, fast, local electric field fluctuation results from surface charge oscillations. These observations show that the energy splitting between the bright and the dark exciton states is estimated to be about a few meV [53]. This conclusion is well consistent with the magneto-optic results as mentioned before. Detailed understanding of fine structures of the lowest excitons is important for the optoelectronic applications of carbon nanotubes.

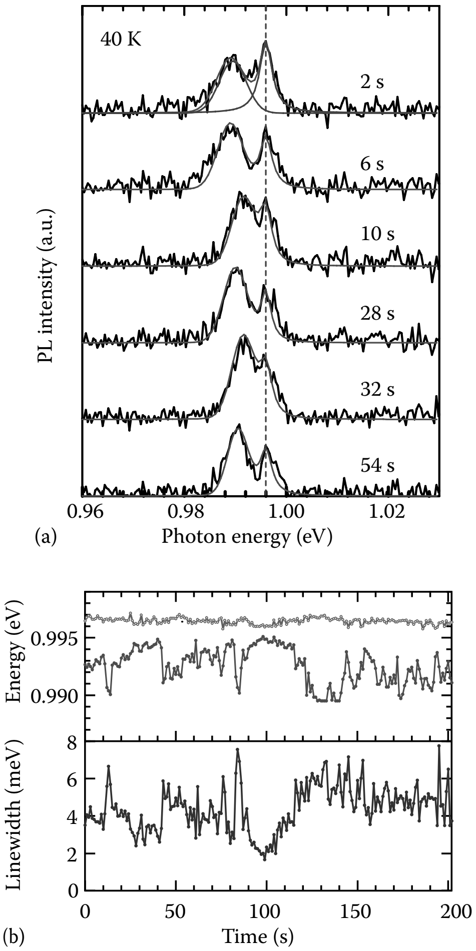

FIGURE 22.5 (a) Temporal evolutions of the PL spectrum of a single SWNT showing spectral fluctuations at 40 K. The solid curves indicate the results of fitting analysis assuming Gaussian and Lorentzian functions. (b)Temporal trace of the higher and lower energy peaks of the PL spectra of a single SWNT at 40 K. The temporal trace of the linewidth (FWHM) of the lower energy peaks. (Reprinted from Matsuda, K. et al., *Phys. Rev. B*, 77, 193405, 2008. With permission.)

22.3 Nanoparticle Quantum Dots

Similar spectral diffusion and PL blinking phenomena are also observed in nanoparticle quantum dots. PL blinking phenomenon is quite enhanced in 0D nanoparticles rather than 1D carbon nanotubes. PL blinking in single nanoparticles is caused by a random switching between light-emitting "on" and non-light-emitting "off" states under continuous-wave (cw) laser excitation. Since the first blinking observation in nanoparticles [35], the mechanism of nanoparticles PL blinking has been extensively discussed [55–70]. Ionization of nanoparticles due to nonradiative Auger recombination plays an essential role in PL blinking of single nanoparticles [55]. It is well accepted in this field that PL blinking originates from the photoionization and neutralization of nanocrystals under cw light illumination. In CdSe nanocrystals, for example, the non-light-emitting off-time state is due to positively charged nanoparticles, and the light-emitting on-time state is due to the neutral nanoparticle [55,70]. Spectral fluctuations during on-time suggest that both electrons and holes trapped on the nanoparticle surface cause transient and local electric field fluctuations in the light-emitting neutral nanoparticle. In neutral nanoparticles, excitons recombine radiatively. In ionized (or charged) nanoparticles, the photogenerated excitons and excess holes recombine nonradiatively through fast three-body Auger recombination. The PL blinking means that a nanoparticle repeats the neutralization and ionization under cw laser excitation.

PL blinking behavior is very sensitive to the local surrounding environments. Figure 22.6 shows the PL intensity time traces of a single CdSe nanoparticle on different substrates: glass, rough, and flat Au surfaces [66]. The samples are excited by cw laser at room temperature. The on-off PL blinking behavior is clearly observed on the glass substrate. The time distribution of the on and off states can be characterized by power law functions [57]. The power law distributions suggest that the PL blinking is caused by a very complicated process. On the metal surfaces, on the other hand, the PL off-time is drastically suppressed. The PL blinking suppression indicates that the very rapid neutralization of ionized nanoparticles occurs through the fast energy transfer. However, the enhancement

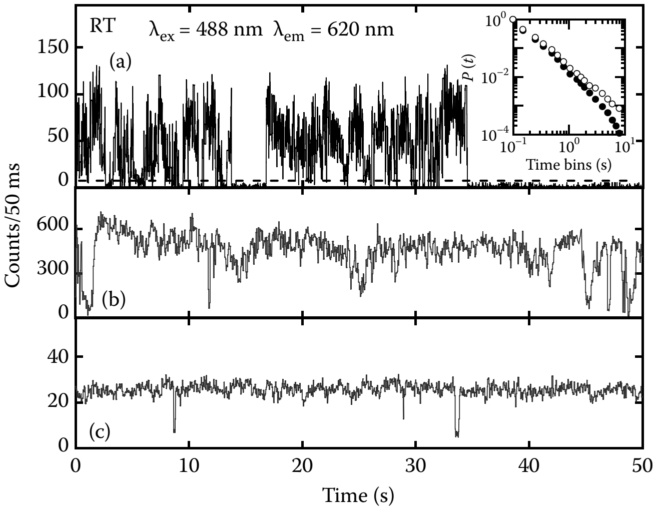

FIGURE 22.6 PL intensity time-traces of single CdSe/ZnS nanoparticles on the glass (a), the rough Au surface (b), and the flat Au surface (c). The inset of (a) is the histogram of the on-time (solid circles) and off-time (open circles) duration of the PL intensity time trace on the glass. (Reprinted from Matsunaga, R. et al., *Phys. Rev. Lett.*, 101, 147404, 2008. With permission.)

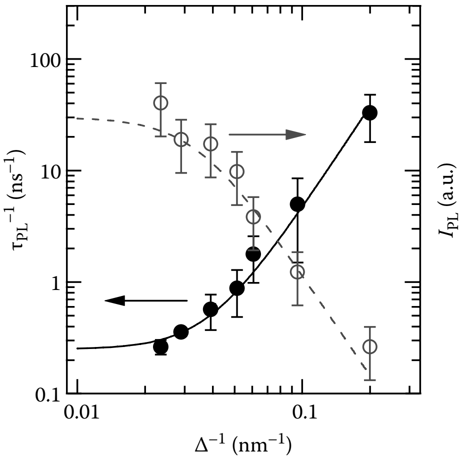

FIGURE 22.7 PL lifetime and PL intensity as a function of the distance between the CdSe nanoparticles monolayer and the Au film. (Reprinted from Ueda, A. et al., *Appl. Phys. Lett.*, 92, 133118, 2008. With permission.)

and quenching of the PL intensity depend on the roughness of the metal surface. On the rough surface, the PL intensity increases. In this experiment, rough Au surfaces are composed of an assembly of hemispherical particles with lateral sizes of 20–50 nm and peaks and valleys of roughly 15 nm. On the flat surface, however, the PL intensity decreases. The enhancement of the PL intensity depends on the excitation laser wavelength. The enhancement spectrum agrees well with the absorption spectrum for localized plasmon resonance [66].

Even on rough Au surfaces, we can observe both enhancement and quenching phenomena of the PL intensity. PL intensity on a rough Au surface depends on the polarization angle of the excitation laser [71]. Electric field enhancement depends on the polarization direction. The PL intensity of a single nanoparticle on glass does not change with the polarization angle. We obtain the microscopic structure of semiconductor nanoparticles and rough surfaces from the polarization and wavelength dependences. Thus, we conclude that the PL intensity enhancement is related to the electric field enhancement due to localized plasmon excitation. Studying PL blinking behaviors is a way to understand energy transfer processes between nanoparticles and surrounding environments.

Close-packed nanoparticle films or nanoparticle arrays show unique exciton energy transfer and charge carrier transport beyond isolated nanoparticles [72–79]. Many different types of closely packed nanoparticle films, arrayed nanoparticle solids, and nanoparticle suprasolids have been prepared [72–79]. In order to control energy transfer between excitons in semiconductors and plasmons in metals, we fabricate metal-semiconductor hetero-nanostructures and their PL spectrum and dynamics. We fabricated two types of semiconductor-metal nanoparticle heterostructures using the Langmuir–Blodgett technique: close-packed CdSe nanoparticle monolayers on Au substrates [80] and mixed CdSe and Au nanoparticle monolayers [81]. In close-packed CdSe nanoparticle monolayers on Au surfaces, the inert polymer thin film was inserted between the nanoparticle monolayer and the Au substrate. The distance between the excitons and plasmons, Δ, is controlled by the polymer thickness. Figure 22.7 summarizes the distance dependence of the PL lifetime and the time-integrated PL intensity [80]. There is a good correlation between the PL decay rate and the PL intensity. PL quenching only occurs when the distance between excitons and plasmons is less than 30 nm. In large distance samples, the PL decay rates in close-packed monolayers are much larger than those in isolated nanoparticles in solutions. The PL lifetime in the CdSe monolayers on the glass is governed by the nonradiative recombination of excitons in nanoparticles and the energy transfer from small to large CdSe nanoparticles, which have lower exciton energies. Furthermore, the PL decay increases with a decrease of the distance between the Au surface and the CdSe nanoparticle monolayer. This can be attributed to energy transfer from nanoparticles to surface plasmons of the Au surfaces. The reduction of both the PL lifetime and the PL intensity simultaneously occurs. PL quenching only occurs in the CdSe nanoparticle monolayer in close proximity to the Au films [80].

Close-packed monolayer films composed of CdSe and Au nanoparticles have simple two-dimensional hexagonal lattices [81]. The PL and optical density of the sample films depend on the Au nanoparticle concentration in the film, because of the spectral overlap between the exciton luminescence of CdSe nanoparticles and the plasmon absorption in Au nanoparticle. In the CdSe and Au nanoparticle mixed monolayer samples, the PL decay curves can be reproduced successfully using three exponential decay components. Three kinds of decay channel of excited states exist in the close-packed CdSe and Au nanoparticle monolayer. In mixed monolayer samples, the decay times are classified into three components: 0.2, 1, and 10 ns, as summarized in Figure 22.8a [81]. These decay times are almost independent of

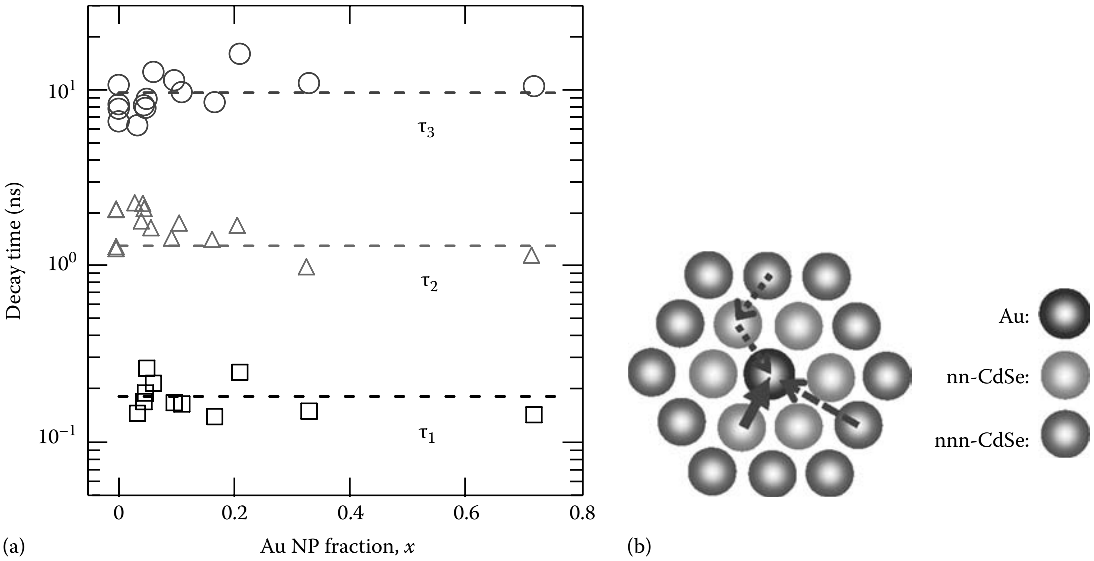

FIGURE 22.8 (a) Decay time obtained by fitting using the three exponential decays as a function of the Au nanoparticle fraction. The broken lines are guides. (b) Schematic illustration of the configurations of the metal and semiconductor nanoparticles. Direct energy transfer from a nearest-neighbor (nn) CdSe nanoparticle (solid arrow) and from a next-nearest-neighbor (nnn) CdSe nanoparticle [broken arrow]. Stepwise CdSe → CdSe → Au nanoparticle energy transfer [dotted arrow]. (Reprinted from Hosoki, K. et al., *Phys. Rev. Lett.*, 100, 207404, 2008. With permission.)

the Au nanoparticle mixing ratio in the film. The 10-ns decay time originates from radiative recombination within CdSe nanoparticles in the films. As the Au nanoparticle fraction increases, the amplitudes of two 1- and 10-ns components decrease. The amplitude of the fast 0.2-ns decay component becomes dominant. These results indicate that the fast PL quenching is caused by the energy transfer to Au nanoparticles for the CdSe nanoparticles in contact with Au nanoparticles. Here, note that the decay component of about 1 ns is a unique characteristic of close-packed mixed nanoparticles solids. Energy transfer between the nearest-neighbor CdSe nanoparticles takes part in the slow PL quenching process of 1 ns in the mixed film. The 1-ns PL quenching process is the stepwise energy transfer from a CdSe nanoparticle to a CdSe nanoparticle to a Au nanoparticle, as illustrated in Figure 22.8b [81]. Therefore, we conclude that the PL dynamics are explained by three kinds of decay channel. The energy relaxation rate in semiconductor nanoparticles is controlled by changing local surrounding environments. These close-packed nanoparticle heterostructures will show unique exciton energy transfer, and charge carrier transport beyond isolated nanoparticles opens new application fields.

22.4 Multiexciton Generation

Finally, we discuss unique exciton–exciton interactions in nanoparticles and nanotubes. Nanoparticle quantum dots and carbon nanotubes provide an excellent stage for experimental studies of many-body effects of excitons or electrons on optical processes in semiconductors [82–84]. The reduced dielectric screening and the relaxation of the energy–momentum conservation rule in nanostructures enhance the Coulomb carrier–carrier interactions, leading to multi-carrier processes such as the quantized Auger recombination, multiple exciton generation (MEG), carrier multiplication (CM), and so on [7,85]. The achievement of efficient CM in semiconductors makes it possible to produce highly efficient solar cells with conversion efficiencies that exceed the Shockley–Queisser limit of 32% [86]. Strongly confined electrons and excitons in nanomaterials show unique nonlinear optical properties, compared to semiconductor bulk crystals.

Strong interactions between carriers or between excitons cause fast nonradiative Auger recombination of multiple excitons or carriers [82]. Intense interest in Auger recombination in nanoparticles has been stimulated by investigations and searches for new laser and solar cell materials [87–89]. In laser and solar cell applications, nonradiative Auger recombination dominates both the carrier density and the carrier lifetime, determining the device performance. Moreover, in transient absorption, PL, and terahertz conductivity experiments, fast Auger recombination has also been used as a probe in MEG and CM processes [89–96]. The CM efficiencies of nanoparticles are not clear, and the CM mechanism is under discussion. In SWNTs, for example, strong Coulomb interactions enhance the many-body effects of excitons. The fast Auger recombination of excitons has been observed by means of exciton homogeneous linewidth [97] and pump-probe measurements [98–100]. We studied MEG processes in SWNTs at room temperature by temporal change in the carrier density [101]. The fast-decay component grows at increasing excitation intensity. When the photon energy is three times larger than the band-gap energy, Auger recombination occurs efficiently even in the weak intensity region. In our experiment, CM is estimated to be about 1.3 under 4.65 eV excitation [101]. We pointed out that a possible mechanism of CM in carbon nanotubes is the impact ionization. Strong interactions between carriers or between excitons cause unique optical processes in semiconductor nanomaterials. Highly excited semiconductor nanomaterials show new optical functionalities.

22.5 Summary

We briefly discussed luminescence properties of carbon nanotubes and nanoparticle quantum dots by means of single molecular spectroscopy and time-resolved optical spectroscopy. These semiconductor nanomaterials show unique luminescence properties such as spectral diffusion and luminescence blinking. Energy transfer between nanomaterials and surrounding environments affects the PL spectra and dynamics of nanomaterials. Although this chapter is written as a review-type survey of our recent studies, we hope that discussions and many references cited are useful for the readers.

Acknowledgments

The author would like to thank many colleagues and graduate students for their contributions and discussions. In particular, Prof. K. Matsuda deserves mention here. Part of this work was supported by a Grant-in-Aid for Scientific Research on Innovative Area "Optical Science of Dynamically Correlated Electrons (DYCE)" (No. 20104006) from MEXT, Japan, and a Grant-in-Aid for Scientific Research (No. 21340084) from Japan Society for Promotion of Science (JSPS).

References

1. L. Brus: *J. Phys. Chem.* 90, 2555 (1986).
2. Y. Kanemitsu: *Phys. Rep.* 263, 1 (1995).
3. A. P. Alivisatos: *J. Phys. Chem.* 100, 13226 (1996).
4. A. G. Cullis, L. T. Canham, and P. D. J. Calcott: *J. Appl. Phys.* 82, 909 (1997).
5. S. A. Empedocles, R. Neuhauser, K. Shimizu, and M. G. Bawendi: *Adv. Mater.* 11, 1243 (1999).
6. A. P. Alivisatos: *Nat. Biotechnol.* 22, 47 (2004).
7. V. I. Klimov: *J. Phys. Chem. B* 110, 16827 (2006).
8. T. Ando: *J. Phys. Soc. Jpn.* 75, 024707 (2006).
9. D. J. Norris, Al. L. Efros, and S. C. Erwin: *Science*, 319, 776 (2008).
10. Ph. Avouris, M. Freitag, and V. Perebeinos: *Nat. Photon.* 2, 341 (2008).
11. R. Beaulac, P. I. Archer, S. T. Ochsenbein, and D. R. Gamelin: *Adv. Funct. Mater.* 18, 3873 (2008).
12. L. T. Canham: *Appl. Phys. Lett.* 57, 1046 (1990).
13. M. J. O'Connell, S. M. Bachilo, X. B. Huffman, V. C. Moore, M. S. Strano, E. H. Haroz, K. L. Rialon, P. J. Boul, W. H. Noon, C. Kittrell, J. Ma, R. H. Hauge, R. B. Weisman, and R. E. Smalley: *Science* 297, 593 (2002).
14. Y. Kanemitsu, T. Ogawa, K. Shiraishi, and K. Takeda: *Phys. Rev. B* 48, 4883 (1993).
15. C. B. Murray, D. J. Norris, and M. G. Bawendi: *J. Am. Chem. Soc.* 115, 8706 (1993).
16. M. A. Hines and P. Guyot-Sionnest: *J. Phys. Chem.* 100, 468 (1996).
17. X. Peng, M. C. Schlamp, A. V. Kadavanich, and A. P. Alivisatos: *J. Am. Chem. Soc.* 119, 7019 (1997).
18. Z. A. Peng and X. G. Peng: *J. Am. Chem. Soc.* 123, 183 (2001).
19. J. Lefebvre, Y. Homma, and P. Finnie: *Phys. Rev. Lett.* 90, 217401 (2003).
20. A. Nish, J.-Y Hwang, J. Doig, and R. J. Nicholas, *Nat Nanotechnol.* 2, 640 (2007).
21. Y. Kayanuma: *Phys. Rev. B* 38, 9797 (1988).
22. S. V. Gaponenko: *Optical Properties of Semiconductor Nanoparticles* (Cambridge University, Cambridge, U.K., 1998).
23. P. D. J. Calcott, K. J. Nash, L. T. Canham, M. J. Kane, and D. Brumhead: *J. Phys. Condens. Matter* 5, L91 (1993); *J. Lumin.* 57, 257 (1993).
24. T. Itoh, M. Nishijima, A. I. Ekimov, C. Gourdon, Al. L. Efros, and M. Rosen: *Phys. Rev. Lett.* 74, 1645 (1995).
25. N. Nirmal, D. J. Norris, M. Kuno, M. G. Bawendi, Al. L. Efros, and M. Rosen: *Phys. Rev. Lett.* 75, 3728 (1995).
26. Al. L. Efros, M. Rosen, M. Kuno, M. Nirmal, D. J. Norris, and M. Bawendi: *Phys. Rev. B* 54, 4843 (1995).
27. Y. Kanemitsu: *J. Lumin.* 100, 209 (2002).
28. Y. Kanemitsu, S. Okamoto, M. Otobe, and S. Oda: *Phys. Rev. B* 55, R7375 (1997).
29. Y. Kanemitsu and S. Okamoto: *Phys. Rev. B* 58, 9652 (1998).
30. T. Kawazoe and Y. Masumoto: *Phys. Rev. Lett.* 77, 4942 (1996).
31. D. Kovalev, H. Heckler, B. Averboukh, M. Ben-Chorin, M. Schwartzkopff, and F. Koch: *Phys. Rev. B* 57, 3741 (1998).
32. Y. Kanemitsu, H. Tanaka, Y. Fukunishi, T. Kushida, K. S. Min, and H. A. Atwater: *Phys. Rev. B* 62, 5100 (2000).
33. W. E. Moerner: *J. Phys. Chem. B* 106, 910 (2002).
34. S. A. Empedocles, D. J. Norris, and M. G. Bawendi: *Phys. Rev. Lett.* 77, 3873 (1996).
35. N. Nirmal, B. O. Dabbousi, M. G. Bawendi, J. I. Macklin, J. K. Trautman, T. D. Harris, and L. E. Brus: *Nature* 383, 802 (1996).
36. A. Ishizumi, K. Matsuda, T. Saiki, C. W. White, and Y. Kanemitsu: *Appl. Phys. Lett.* 87, 133104 (2005).
37. A. Ishizumi, C. W. White, and Y. Kanemitsu: *Appl. Phys. Lett.* 84, 2397 (2004).
38. S. M. Bachilo, M. S. Strano, C. Kittrell, R. H. Hauge, R. E. Smalley, and R. B. Weisman: *Science* 298, 2361 (2002).
39. F. Wang, G. Dukovic, L. E. Brus, and T. F. Heinz: *Science* 308, 838 (2005).
40. A. Hartschuh, H. N. Pedrosa, L. Novotny, and T. D. Krauss: *Science* 301, 1354 (2003).
41. J. Lefebvre, J. M. Fraser, P. Finnie, and Y. Homma: *Phys. Rev. B* 69, 075403 (2004).
42. K. Matsuda, Y. Kanemitsu, K. Irie, T. Saiki, T. Someya, Y. Miyauchi, and S. Maruyama: *Appl. Phys. Lett.* 86, 123116 (2005).

43. J. Lefebvre, D. G. Austing, J. Bond, and P. Finnie: *Nano Lett*. 6, 1603 (2006).
44. H. Htoon, M. J. O'Connell, P. J. Cox, S. K. Doorn, and V. I. Klimov: *Phys. Rev. Lett*. 93, 027401 (2004).
45. T. Inoue, K. Matsuda, Y. Murakami, S. Maruyama, and Y. Kanemitsu: *Phys. Rev. B* 73, 233401 (2006).
46. R. Matsunaga, K. Matsuda, and Y. Kanemitsu: *Phys. Rev. Lett*. 101, 147404 (2008).
47. H. Ajiki and T. Ando: *J. Phys. Soc. Jpn*. 62, 1255 (1993).
48. S. Zaric, G. N. Ostojic, J. Kono, J. Shaver, V. C. Moore, M. S. Strano, R. H. Hauge, R. E. Smalley, and X. Wei: *Science* 304, 1129 (2004).
49. S. Zaric, G. N. Ostojic, J. Shaver, J. Kono, O. Portugall, P. H. Frings, G. L. J. A. Rikken, M. Furis, S. A. Crooker, X. Wei, V. C. Moore, R. H. Hauge, and R. E. Smalley: *Phys. Rev. Lett*. 96, 016406 (2006).
50. H. Hirori, K. Matsuda, Y. Miyauchi, S. Maruyama, and Y. Kanemitsu: *Phys. Rev. Lett*. 97, 257401 (2006).
51. I. B. Mortimer and R. J. Nicholas: *Phys. Rev. Lett*. 98, 027404 (2007).
52. A. Srivastava, H. Htoon, V. I. Klimov, and J. Kono: *Phys. Rev. Lett*. 101, 087402 (2008).
53. K. Matsuda, T. Inoue, Y. Murakami, S. Maruyama, and Y. Kanemitsu: *Phys. Rev. B* 77, 193405 (2008).
54. S. A. Empedocles and M. G. Bawendi: *Science* 278, 2114 (1997); *J. Phys. Chem. B* 103, 1826 (1999).
55. Al. L. Efros and M. Rosen: *Phys. Rev. Lett*. 78, 1110 (1997).
56. R. G. Neuhauser, K. T. Shimizu, W. K. Woo, S. A. Empedocles, and M. G. Bawendi: *Phys. Rev. Lett*. 85, 3301 (2000).
57. M. Kuno, D. P. Fromm, H. F. Hamann, A. Gallagher, and D. J. Nesbitt: *J. Chem. Phys*. 115, 1028 (2001).
58. F. Koberling, A. Mews, and T. Basché: *Adv. Mater*. 13, 672 (2001).
59. K. T. Shimizu, W. K. Woo, B. R. Fisher, H. J. Eisler, and M. G. Bawendi: *Phys. Rev. Lett*. 89, 117401 (2002).
60. W. G. J. H. M. van Sark, P. L. T. M. Frederix, A. A. Bol, H. C. Gerritsen, and A. Meijerink: *Chem. Phys. Chem*. 3, 871 (2002).
61. M. Kuno, D. P. Fromm, S. T. Johnson, A. Gallagher, and D. J. Nesbitt: *Phys. Rev. B* 67, 125304 (2003).
62. J. Muller, J. M. Lupton, A. L. Rogach, J. Feldmann, D. V. Talapin, and H. Weller: *Appl. Phys. Lett*. 85, 381 (2004).
63. A. Issac, C. von Borczyskowski, and F. Cichos: *Phys. Rev. B* 71, 161302(R) (2005).
64. D. E. Gómez, J. van Embden, and P. Mulvaney: *Appl. Phys. Lett*. 88, 154106 (2006).
65. K. Zhang, H. Chang, A. Fu, A. P. Alivisatos, and H. Yang: *Nano Lett*. 6, 843 (2006).
66. Y. Ito, K. Matsuda and Y. Kanemitsu: *Phys. Rev. B* 75, 033309 (2007).
67. Y. Ito, K. Matsuda, and Y. Kanemitsu: *J. Phys. Soc. Jpn*. 77, 103713 (2008).
68. P. Frantsuzov, M. Kuno, B. Janko, and R. A. Marcus: *Nat. Phys*. 4, 519 (2008).
69. F. Stefani, J. Hoogenboom, and E. Barkal: *Phys. Today* 62(2), 34 (2009).
70. T. D. Krauss and L. E. Brus: *Phys. Rev. Lett*. 83, 4840 (1999).
71. K. Matsuda, Y. Ito, and Y. Kanemitsu: *Appl. Phys. Lett*. 92, 211911 (2008).
72. C. M. Murray, C. R. Kagan, and M. G. Bawendi: *Science* 270, 1966 (1995).
73. C. A. Leatherdale, C. R. Kagan, N. Y. Morgan, S. A. Empedocles, M. A. Kastner, and M. G. Bawendi: *Phys. Rev. B* 62, 2669 (2000).
74. M. P. Pileni: *J. Phys. Chem. B* 105, 3358 (2001).
75. C. R. Kagan, C. B. Murray, M. Nirmal, and M. G. Bawendi: *Phys. Rev. Lett*. 76, 1517 (1996).
76. C. R. Kagan, C. B. Murray, and M. G. Bawendi: *Phys. Rev. B* 54, 8633 (1996).
77. S. A. Crooker, J. A. Hollingsworth, S. Tretiak, and V. I. Klimov: *Phys. Rev. Lett*. 89, 186802 (2002).
78. M. Achermann, M. A. Petruska, S. A. Crooker, and V. I. Klimov: *J. Phys. Chem. B* 107, 13782 (2003).
79. O. I. Mićić, K. M. Jones, A. Cahill, and A. J. Nozik: *J. Phys. Chem. B* 102, 9791 (1998).
80. A. Ueda, T. Tayagaki, and Y. Kanemitsu: *Appl. Phys. Lett*. 92, 133118 (2008).
81. K. Hosoki, S. Yamamoto, T. Tayagaki, K. Matsuda, and Y. Kanemitsu: *Phys. Rev. Lett*. 100, 207404 (2008).
82. V. I. Klimov, A. A. Mikhailovsky, D. W. McBranch, C. A. Leatherdale, and M. G. Bawendi: *Science* 287, 1011 (2000).
83. V. I. Klimov, A. A. Mikhailovsky, S. Xu, A. Malko, J. A. Hollingsworth, C. A. Leatherdale, H. J. Eisler, and M. G. Bawendi: *Science* 290, 314 (2000).
84. Y. Kanemitsu, T. J. Inagaki, M. Ando, K. Matsuda, T. Saiki, and C. W. White: *Appl. Phys. Lett*. 81, 141 (2002).
85. A. J. Nozik: *Chem. Phys. Lett*. 457, 3 (2008).
86. W. Shockley and H. J. Queisser: *J. Appl. Phys*. 32, 510 (1961).
87. V. I. Klimov, S. A. Ivanov, J. Nanda, M. Achermann, I. Bezel, J. A. McGuire, and A. Piryatinski: *Nature* 447, 441 (2007).
88. T. Tayagaki, S. Fukatsu, and Y. Kanemitsu: *Phys. Rev. B* 79, 041301(R) (2009).
89. R. D. Schaller and V. I. Klimov: *Phys. Rev. Lett*. 92, 186601 (2004).
90. R. Ellingson, M. C. Beard, J. C. Johnson, P. Yu, O. I. Micic, A. J. Nozik, A. Shabaev, and A. L. Efros: *Nano Lett*. 5, 865 (2005).
91. M. C. Beard, K. P. Knutsen, P. Yu, J. M. Luther, Q. Song, W. K. Metzger, R. J. Ellingson, and A. J. Nozik: *Nano Lett*. 7, 2506 (2007).
92. G. Nair and M. G. Bawendi: *Phys. Rev. B* 76, 081304(R) (2007).
93. V. I. Klimov, J. A. McGuire, R. D. Schaller, and V. I. Rupasov: *Phys. Rev. B* 77, 195324 (2008).
94. G. Nair, S. M. Geyer, L. Y. Chang, and M. G. Bawendi: *Phys. Rev. B* 78, 125325 (2008).

95. J. J. H. Pijpers, E. Hendry, M. T. W. Milder, R. Fanciulli, J. Savolainen, J. L. Herek, D. Vanmaekelbergh, S. Ruhman, D. Mocatta, D. Oron, A. Aharoni, U. Banin, and M. Bonn: *J. Phys. Chem. C* 111, 4146 (2007); 112, 4783 (2008).
96. M. T. Trinh, A. J. Houtepen, J. M. Schins, T. Hanrath, J. Piris, W. Knulst, A. P. L. M. Goossens, and L. D. A. Siebbeles: *Nano Lett*. 8, 1713 (2008).
97. K. Matsuda, T. Inoue, Y. Murakami, S. Maruyama, and Y. Kanemitsu: *Phys. Rev. B* 77, 033406 (2008).
98. F. Wang, G. Dukovic, E. Knoesel, L. E. Brus, and T. F. Heinz: *Phys. Rev. B* 70, 241403(R) (2004).
99. F. Wang, Y. Wu, M. S. Hybertsen, and T. F. Heinz: *Phys. Rev. B* 73, 245424 (2006).
100. Y.-Z. Ma, L. Valkunas, S. L. Dexheimer, S. M. Bachilo, and G. R. Fleming: *Phys. Rev. Lett*. 94, 157402 (2005).
101. A. Ueda, K. Matsuda, T. Tayagaki, and Y. Kanemitsu: *Appl. Phys. Lett*. 92, 233105 (2008).

23

Nanoscale Excitons and Semiconductor Quantum Dots

Vanessa M. Huxter
University of Toronto

Jun He
University of Toronto

Gregory D. Scholes
University of Toronto

23.1 Introduction to Nanoscale Excitons **23**-1
23.2 Nonlinear Optical Properties of Semiconductor Quantum Dots **23**-5
23.3 Two-Photon Absorption in Semiconductor Quantum Dots **23**-6
Biological Imaging • Amplified Stimulated Emission and Lasing • Optical Power Limiting • Two-Photon Sensitizer
23.4 Conclusion **23**-8
References **23**-8

23.1 Introduction to Nanoscale Excitons

The optical properties of nanoscale excitons are of interest to researchers because their size tunability makes them of practical use; however, understanding the nature of these delocalized electronic states is a challenge (Scholes and Rumbles 2006, 2008). Nanoscale excitons are the electronic excited states formed by the absorption of light by nanoscale systems. A nanoscale system is any chemical or physical system that is in the nanometer-size regime. The kinds of systems that are the subject of this chapter tend to have properties lying between those of molecules and those of bulk semiconductors. Unlike excitons in bulk semiconductors, the energies of nanoscale excitons can often be changed by the size of the system. That property is envisioned, for example, to allow the color of solid-state lasers to be tuned.

Well-studied examples of nanoscale systems include semiconductor nanocrystals, carbon nanotubes (CNTs), organic conjugated polymers, and molecular aggregates, shown in Figure 23.1. The excitons in quantum dots and CNTs are closely related to molecular excited states; however, the size of the nanoscale systems has forced researchers to make severe approximations in their quantum-mechanical descriptions. Nonetheless, the excited states of nanoscale systems tend to be more amenable to approximate descriptions than molecules because the wavefunction delocalization reduces the importance of the electron correlation for an accurate calculation of the energies of electronic states. Such electronic excited states are often described as Wannier–Mott excitons (Banyai and Koch 1993; Basu 1997; Gaponenko 1998; Jorio et al. 2008). Other systems, like molecular aggregates, crystals, and certain proteins—namely, those involved in photosynthetic energy transduction—have optical properties that are better described with reference to the molecular building blocks of the aggregate. The lowest electronic excited states of these aggregates are called Frenkel excitons (Kasha 1976). Conjugated polymers, now used in organic light-emitting diodes and displays (OLEDs), have excited states somewhere between these limits (Sariciftci 1997; Hadziioannou and Malliaras 2006). This relationship is illustrated in Figure 23.2.

In the limiting case of Wannier–Mott excitons, it is assumed that the electronic interaction between the building blocks of the system—the atoms or unit cells—is large, akin to a chemical bond. In this case, orbitals that are delocalized over an entire system are a good starting point for describing one-electron states. Photoexcitation introduces an electron into the conduction orbitals, leaving a "hole" in the valence orbitals. The Wannier–Mott description assumes that there are many electrons in the system, so the atomic centers are highly screened from the outermost electrons, that is, the dielectric constant is high. In that case, it is considered reasonable to assume that the electron and the hole move freely in the background dielectric continuum and are weakly bound. The strength of this electron–hole attraction determines the "binding" of the lower energy, optically allowed states compared to the dense manifold of charge carrier states that lie higher in energy. In this model, the electron and the hole move under their mutual attraction in a dielectric continuum, and then the exciton energy levels are found as a series analogous to the states of the hydrogen atom. Note that this model can be useful, but it must be realized that it is highly approximate (Scholes 2008b), and it is only able to give limited insights. The conceptual advantage of this model is that it converges to the free carrier limit where the electron-hole attraction is negligible compared to the thermal energies. It can thereby be seen how the photoexcitation of a bulk semiconductor exciton efficiently produces charge carriers, which is how a typical semiconductor solar cell works.

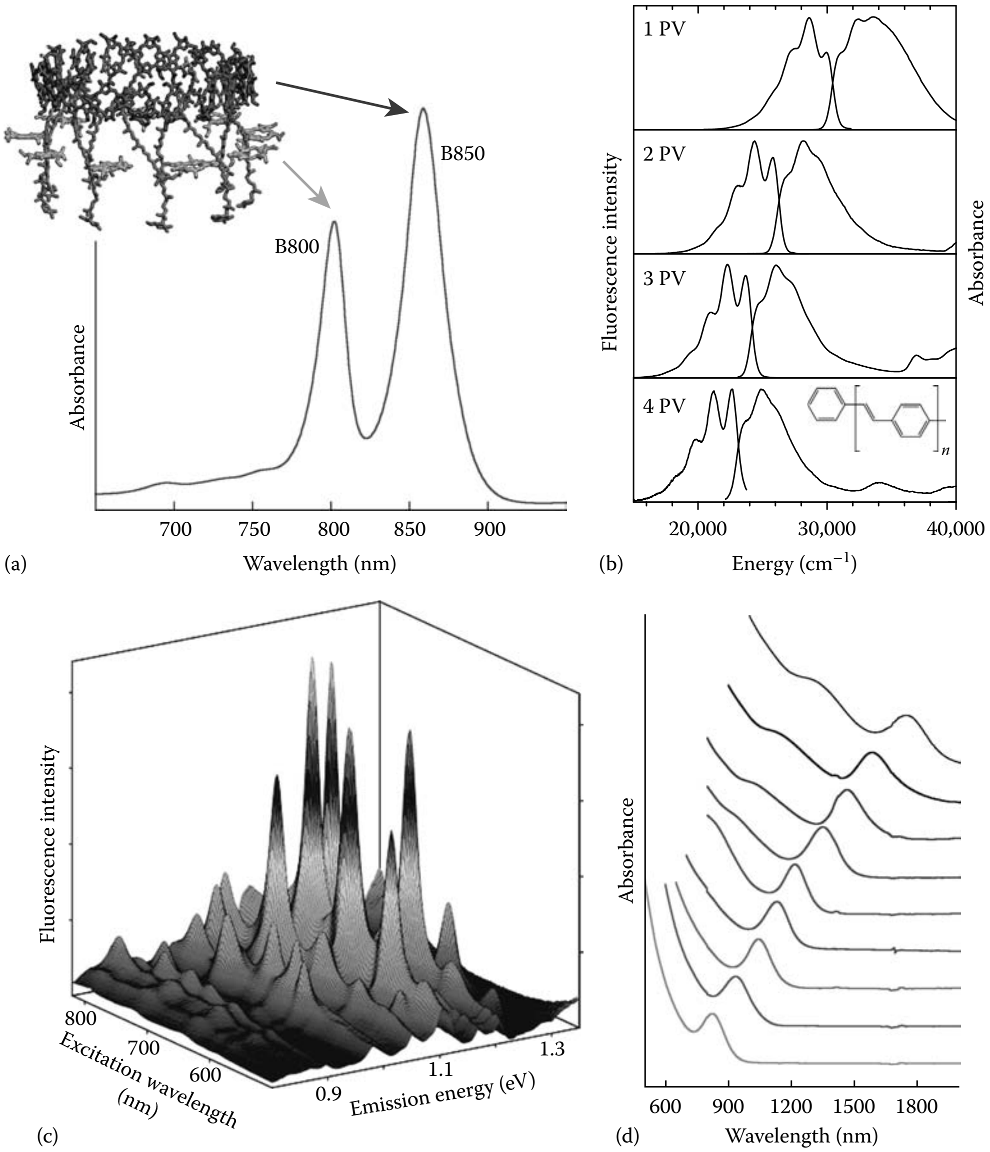

FIGURE 23.1 **(See color insert following page 20-14.)** Excitons and structural size variations on the nanometer length scale. (a) The photosynthetic antenna of purple bacteria, LH2, is an example of a molecular exciton. The absorption spectrum clearly shows the dramatic distinction between the B800 absorption band, arising from essentially "monomeric" bacteriochlorophyll-a (Bchl) molecules, and the redshifted B850 band that is attributed to the optically bright lower exciton states of the 18 electronically coupled Bchl molecules. (b) The size-scaling of polyene properties, for example, oligophenylenevinylene oligomers, derives from the size-limited delocalization of the molecular orbitals. However, as the length of the chains increases, disorder in the chain conformation impacts the picture for exciton dynamics. Absorption and fluorescence spectra are shown as a function of the number of repeat units. (c) SWCNT size and "wrapping" determine the exciton energies. Samples contain many different kinds of tubes, therefore optical spectra are markedly inhomogeneously broadened. By scanning excitation wavelengths and recording a map of fluorescence spectra, the emission bands from various different CNTs can be discerned, as shown. (Courtesy of Dr. M. Jones). (d) Rather than thinking in terms of delocalizing the wavefunction of a semiconductor through interactions between the unit cells, the small size of the nanocrystal confines the exciton relative to the bulk. Size-dependent absorption spectra of PbS quantum dots are shown. (Adapted from Scholes, G. D. and Rumbles, G., *Nat. Mater.*, 5, 683, 2006.)

In order to further understand the formation of excitons in nanoscale systems, we will examine the case of semiconductors as a model system. The optical properties of semiconductor nanostructures lie in an intermediate regime between a molecular and a bulk description. In a bulk semiconductor, the dimensions of the system are practically infinite compared to the dimensions of the carriers (electrons and holes). In this case, the wavefunctions, which are standing waves in the material, are spread over an infinite number of unit cells (the repeat unit in a crystal), and the carriers (electrons and holes) are free particles. The density of states for the bulk material is continuous, as shown in Figure 23.3a. Note that in the case of a bulk semiconductor, the density of states for both the hole and the electron are continuous. These two continuous regions are separated by

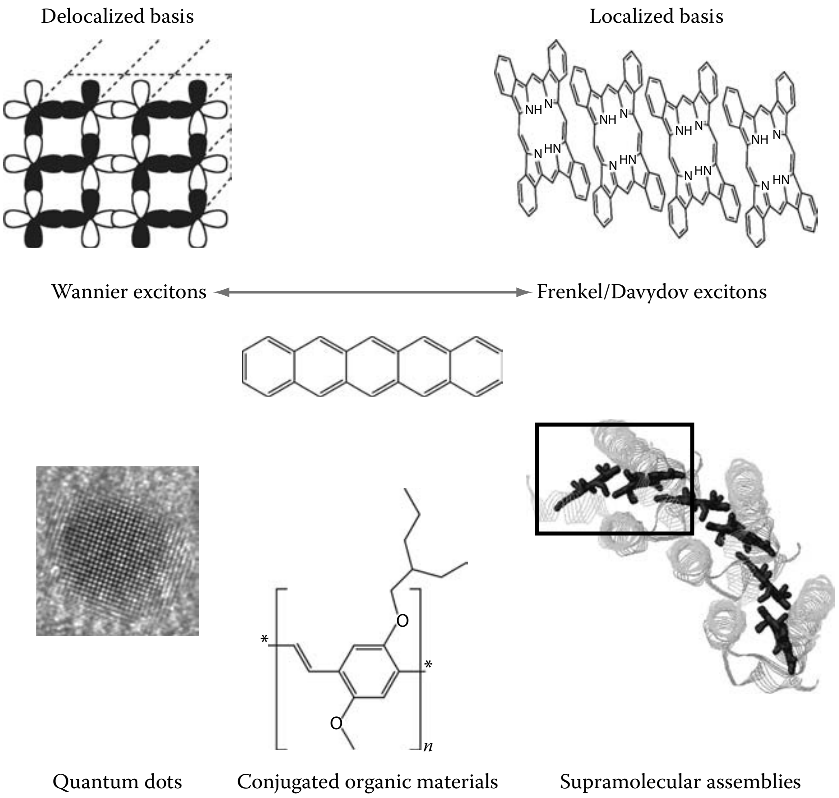
Delocalized basis
Localized basis
NH
N
HN
Wannier excitons
Frenkel/Davydov excitons
O
*
n
Quantum dots
Conjugated organic materials
Supramolecular assemblies

FIGURE 23.2 Collective properties of nanoscale materials modify the optical properties, such as the wavelength and the dipole strength for light absorption. The elementary excitations are known as excitons. Excitons are formed through the collective absorption of light by two or more repeat units in a crystal, a molecular assembly, or a macromolecule. Wannier excitons are typical of atomic crystals, semiconductor quantum dots, aromatic molecules, CNTs, conjugated polymers, and so on. Supramolecular assemblies, including J-aggregates and photosynthetic light-harvesting antennae, typify Frenkel excitons—excitons in which the repeat units retain their identity to a significant degree.

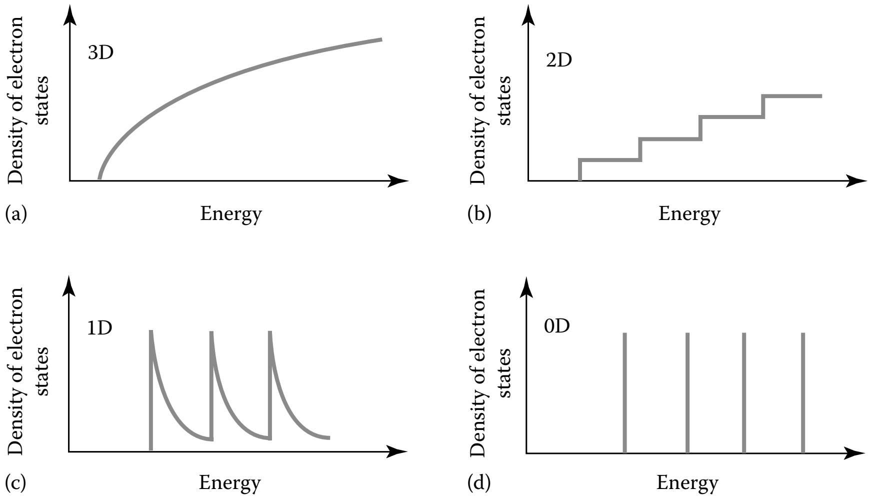

FIGURE 23.3 Idealized representation of the density of electron states for (a) a bulk semiconductor system (3D) and a semiconductor system confined in (b) one dimension (a 2D system), (c) confined in two dimensions (a 1D system) and (d) confined in all three dimensions (a 0D system). In the bulk case, the energy levels form a continuous band. With increasing confinement, the density of states becomes more discrete, resulting in the delta functional form of the density of states in the 0D case.

the bandgap in which the density of states goes to zero for an ideal system. The bulk system is infinite in all three dimensions. The effect of spatially confining the electrons and holes in one, two, or three of those dimensions is to change the density of states. The spatial confinement of the electrons and holes in one dimension means that the wavefunctions of those particles are quantized in that same one dimension. However, the electrons and holes are not confined and can move freely in the other two dimensions. This is similar to the concept of a plane as a two-dimensional object in a three-dimensional space. Examples of these two-dimensional systems with a spatial confinement in one dimension are nanometer-height thin films and quantum wells. The quantization of the electrons and holes in one dimension changes the density of states from the bulk continuum to a step-type function, as shown in Figure 23.3b. Similarly, spatial confinement of the electrons and holes in two dimensions results in the quantization of the particles in those same two dimensions. In these systems, called quantum wires, the electrons and the holes can only move freely in one dimension. The quantization in two of three dimensions results in a further change in the density of states, which becomes more discrete with individual peaks as shown in Figure 23.3c. Finally, the spatial confinement of the electron and the hole in all three dimensions results in quantization in all three dimensions. In this case, the density of states is discrete and takes the form of delta functions, as shown in Figure 23.3d. These systems are called quantum dots and are discussed in more detail later in this chapter.

The physical confinement in one, two, or three dimensions changes the boundary conditions imposed on the wavefunctions and quantizes the behavior of the electrons and holes in one, two, or three dimensions. The boundary conditions associated with the spatial confinement of the wavefunctions, in turn, modify the density of states. The resulting change in the density of states from a continuous band in the bulk to discrete levels in a quantum dot is analogous to the particle in a box model in quantum mechanics. The physical dimensions required to spatially confine the electron or the hole depend on the size of the electron and hole wavefunctions. These are determined by the physical properties of the specific material and are usually on the order of one to tens of nanometers.

The quantum confinement of the electron and the hole wavefunctions modifies their interaction. The spatially overlapped electron and hole wavefunctions can associate through an energy-lowering Coulomb interaction due to their relative negative and positive charges, forming an exciton. This association can be described in analogy to a hydrogen atom and, as such, the size of the exciton is described using a Bohr radius. Just like the hydrogen atom, the lowest energy set of states is comprised of closely bound electron-hole pair configurations. These are the bound exciton states, which are the optically active states of the system (those that absorb and emit light). They tend to be clearly distinguished from a band of many higher energy states, which do not absorb light, but are the nanoscale-free carrier states (Scholes 2008b).

The energy separation between the lowest energy-exciton states and the onset of the carrier states is called the exciton-binding energy. We can think of the exciton-binding energy as the energy required to ionize an exciton. Notably, this ionization energy is significantly reduced compared to small, molecular systems because of the many different ways that the electron and the hole can be separated. The exciton-binding energy, meditated by an attractive Coulomb interaction, can also be thought of as the energy reduction associated with forming an exciton as compared to the energies of the free electron and hole confined in the three-dimensional potential. This exciton-binding energy shifts the positions of the energy levels. The exciton in a three-dimensional potential and the associated exciton levels are shown in Figure 23.4.

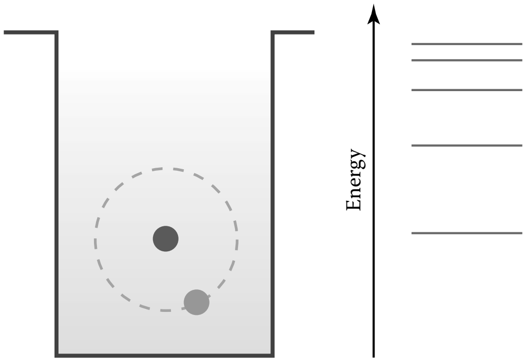

FIGURE 23.4 Confining the electron and the hole in a potential results in the formation of a hydrogen-like exciton and bound exciton states. The binding of the exciton is mediated through the Coulomb interaction.

The binding energy of excitons is a topic of widespread interest, especially because of its relevance to photovoltaic science. In high dielectric constant bulk-semiconductor materials the exciton-binding energy is typically small: 27 meV for CdS, 15 meV for CdSe, 5.1 meV for InP, and 4.9 meV for GaAs. The small exciton-binding energies of these materials make them well-suited for photovoltaic applications because the optically active exciton states absorb light, then the ambient thermal energy (~25 meV) is sufficient to convert the exciton to free carriers. Thus, the potential energy of the absorbed photon is converted to an electrical potential. On the other hand, in molecular materials, the electron-hole Coulomb interaction is substantial—usually a few eV. In nanoscale materials, we find a middle ground where exciton-binding energies are significant in magnitude—that is, excitons are important.

As an example, the valence and the conduction orbitals of a single-wall carbon nanotube (SWCNT) are shown in Figure 23.5. These are typical 1D densities-of-states as were introduced in Figure 23.3. If the electrons were non-interacting, then the lowest excitation energy of a SWCNT would be the same as the energy gap between the highest valence orbital and the lowest conduction orbital. In fact, this is the energy corresponding to the onset of carrier formation. Suitable quantum-chemical calculations introduce interactions between the electrons in these various orbitals, which captures the electron-hole attraction and thus leads to a significant stabilization of the excited states relative to

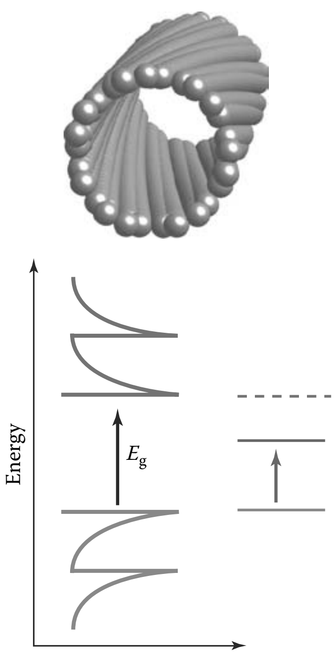

FIGURE 23.5 The attractive interaction of electrons and holes in a CNT (represented by the model at the top of the figure) results in the formation of excitons. This stabilizes the excited states relative to the orbital energy difference, lowering the energy of the optical gap.

those estimated with the more primitive model of orbital energy differences. That gives the correct energy of the optical absorption bands. The energy difference between the orbital energy difference and the optical gap is the exciton-binding energy. It has been predicted for SWCNTs to be ~0.3–0.5 eV (Zhao and Mazumdar 2004), which was later confirmed by an experiment (Ma et al. 2005; Wang et al. 2005).

As a result of the nature of excitons, the optical properties of nanoscale systems provide an interesting link between the properties of extended "bulk" systems and those of molecules. The study of excitons provides the opportunity for new insights into the behavior of nanoscale systems. For the rest of this chapter, we will focus on the properties of a particular nanoscale system: semiconductor quantum dots.

23.2 Nonlinear Optical Properties of Semiconductor Quantum Dots

Quantum-confined semiconductor nanocrystals, or quantum dots, have been the focus of intense study over the past three decades due to their size-tunable optical properties and unique physical characteristics. This area of research was founded in the early 1980s when researchers at Bell Laboratories (Rossetti et al. 1983) in the United States and the Yoffe Institute (Ekimov et al. 1980, Ekimov and Onushchenko 1982; Efros and Efros 1982) in Russia (then the U.S.S.R) independently described the properties of nanometer-sized semiconductor quantum dots. This work was quickly followed by studies on colloidal samples (Spanhel et al. 1987), leading to a further understanding of the optical and physical properties of quantum dots. Within less than a decade, a basic theoretical framework to describe the observed properties had been proposed and work was underway to explore the fundamental physics of these materials.

Much of the early work on quantum dots focused on semiconductor-doped glasses. These materials are characterized by broad size distributions and offer no possibility of control over the shape or the interface characteristics of the particles. The limitations of these doped glasses, particularly the broad size distribution that results in a large static inhomogeneity, masked many of the fundamental physical processes that were occurring in the quantum dots. However, the study of quantum dot systems underwent a revolution in 1993 when it was discovered that nucleation and a controlled growth of colloidal semiconductors could be achieved by injecting highly reactive organometallic precursors into a solvent system that coordinates to the colloid surface (Murray et al. 1993). This coordinating solvent, trioctylphosphine oxide, was an important discovery because it serves to arrest growth and stabilize the nanocrystals. This method allowed a simple and reproducible synthesis of high quality, nearly monodisperse cadmium chalcogenide nanocrystal samples. This development led to an explosion in the field and opened new avenues for research and technology including increased processibility, the possibility of mass production, and chemical manipulation for tailored shape control. In particular, this reliable method of making high-quality samples allowed the exploration of phenomena that had been previously unobservable due to static inhomogeneity associated with broad size distributions.

In addition to doped glasses and colloidal samples, there are epitaxially grown quantum dots. While colloidal nanocrystals tend to be sized in the range of 1–10 nm in diameter, epitaxial dots, which are grown on solid substrates, may be a couple of nanometers in height but with lateral dimensions of tens of nanometers. The physics of these materials differ from colloidal samples in some fundamental ways; however, the basic properties associated with quantum-confined nanocrystals apply to all three types of quantum dots. Different aspects of these properties have been explored in many comprehensive reviews. For instance, a review of the electronic properties was presented by Yoffe (1993, 2001). The optical nonlinearities of semiconductor nanocrystals were reviewed by Banfi's group (Banfi et al. 1998). Research on semiconductor quantum dots has evolved from fundamental science (Alivisatos 1996; Empedocles and Bawendi 1997; Klimov 2000; Klimov et al. 2000a,b) to lasing and amplification (Klimov 2006; Klimov et al. 2007), optical power limiting (He et al. 2007a,b), biological imaging (Bruchez et al. 1998; Dubertret et al. 2002; Larson et al. 2003; Michalet et al. 2005) and labeling (Seydack 2005), sensitization (Dayal and Burda 2008), and optical switching (Etienne et al. 2005; He et al. 2005a).

One of the defining features of a semiconductor is the bandgap, which separates the conduction band and the valence band. When a semiconductor material absorbs light, an electron is promoted from the valence to the conduction band. The wavelength of light absorbed and emitted from a semiconductor material is determined by the width of the bandgap. In semiconductors, when an electron moves from the valence to the conduction band, the resulting gap left in the valence band is called a hole. An exciton can be formed through an energy-lowering Coulomb interaction between the negative electron and the positive hole. In analogy with the hydrogen atom, the spatial extent of the exciton wavefunction is described by a quantity called the exciton Bohr radius. As a result of the three-dimensional spatial confinement of the exciton wavefunction, the density of states becomes discrete, as described earlier in the chapter. The position of these states, and therefore the energy of the gap, depends on the spatial confinement of the exciton, which is determined by the physical size of the nanocrystal. To a first approximation, this quantum-confinement effect can be described using the particle in a box or a simple quantum box model (Efros and Efros 1982; Brus 1984), in which the electron motion is restricted in all three dimensions by impenetrable walls. For a spherical nanocrystal with radius R, the quantum box model predicts that the size-dependence of the energy gap is proportional to $1/R^2$, indicating that the energy of the lowest transition increases as the nanocrystal size decreases. In addition, as described above, quantum confinement changes the continuous energy bands of a bulk semiconductor into discrete exciton energy levels. The exciton energy levels produce peaks in the absorption spectrum of quantum dots, which is in contrast to the continuous absorption spectrum of a bulk semiconductor (Alivisatos 1996).

The colloidal quantum dots discussed here are composed of a semiconductor core surrounded by a shell of organic ligand molecules (Murray et al. 1993). The organic capping prevents the uncontrolled growth and agglomeration of the nanoparticles. It also allows quantum dots to be chemically manipulated as if they were large molecules, with solubility and chemical reactivity determined by the identity of the organic molecules. The capping also provides an electronic passivation of the nanocrystals by terminating the dangling bonds on the surface. The unterminated dangling bonds can affect the emission efficiency of the quantum dots because they lead to a loss mechanism where electrons are rapidly trapped at the semiconductor surface before they have a chance to emit photons. Using colloidal chemical synthesis, one can prepare nanocrystals with nearly atomic precision with diameters ranging from nanometers to tens of nanometers and a size dispersion as narrow as 5% standard deviation. Because of the quantum-confinement effect, the ability to tune the size or shape of the nanocrystals translates into a means of controlling their optical properties, such as the absorption and emission wavelengths (Scholes and Rumbles 2006; Scholes 2008a).

In a quantum dot system, the three-dimensional confinement modifies the Hamiltonian, adding a potential that restricts how far apart the electron and the hole can be. This forced spatial overlap changes the density of states as described earlier in the chapter. The interaction between the confined electron and the hole, mediated by a Coulomb potential, leads to the formation of an exciton, which can be described in analogy to a hydrogen atom where an electron interacts with a nucleus. As a result of the discrete character of the density of states in quantum dots, as shown in Figure 23.3d, the oscillator strength is concentrated into those few transitions instead of being spread over a continuum of states. This means that the oscillator strength of the states in quantum dots is significantly enhanced. The concentration of the oscillator strength into a few transitions also enhances the nonlinear optical properties of quantum dots as compared to the bulk (Shalaev et al. 1996). As a result of their enhanced nonlinear optical properties, synthetically controllable size tunability and photostability, quantum dots continue to be of great interest both for fundamental research and device applications. One of these areas of research involves nonlinear two-photon absorption (2PA).

23.3 Two-Photon Absorption in Semiconductor Quantum Dots

Two-photon absorption in semiconductors is the simultaneous absorption of two photons of identical or different frequencies required to move an electron from the valence to the conduction band. As opposed to the linear intensity dependence of one-photon absorption (1PA), 2PA depends on the square of the light intensity. Therefore, 2PA is a third-order nonlinear optical process and many orders of magnitude weaker than 1PA. The 2PA coefficient, β, is directly related to the imaginary part of the third-order nonlinear susceptibility, $\chi^{(3)}$, by $\beta = 3\pi \mathrm{Im}\,\chi^{(3)}/(\lambda n_0^2 c \varepsilon_0)$, where n_0 is the linear refractive index, ε_0 is the dielectric constant in vacuum, λ the laser wavelength, and c the speed of light in vacuum (Sutherland 2003). Compared to 1PA, 2PA is associated with different selection rules for dipole transitions. For example, in CdSe quantum dots, one-photon transitions satisfy selection rules $\Delta L = 0, \pm 2$ and $\Delta F = 0, \pm 1$ while two-photon transitions satisfy $\Delta L = \pm 1, \pm 3$ and $\Delta F = 0, \pm 1, \pm 2$, where L is the orbital angular momentum of the envelope wavefunction and F is the total angular momentum (Schmidt et al. 1996). Therefore, the comparison between 1PA and 2PA spectra allows a more detailed optical analysis of electronic states in semiconductor quantum dots (Schmidt et al. 1996). For direct bandgap semiconductor quantum dots, a simple theory based on the effective mass approximation was proposed for both 1PA and 2PA (Fedorov et al. 1996). Although this simple model does not consider the mixing between the heavy and light holes bands, it works quite well for describing degenerate 2PA. Recently, this theory has been extended to obtain the analytical expressions for both degenerate and nondegenerate 2PA spectra of semiconductor quantum dots with the parabolic band approximation and $\vec{k}\cdot\vec{p}$ theory (Padilha et al. 2007).

As the incident laser intensity increases, the 2PA in semiconductor quantum dots will be saturated due to the limited density of states of the quantized energy levels (He et al. 2005b). If the allowed 2PA transitions are assumed to occur between the one atomic-like energy level in the conduction band and the other in the valence band, the ensemble of quantum dots can be approximately treated as a two-level system. Since the semiconductor quantum dots are not uniform in size, the saturation of 2PA for such an inhomogeneous system may be derived as $\beta(I) = \beta / \sqrt{1 + I_0^2/I_{\text{s,2PA}}^2}$ (Sutherland 2003). The saturation intensity, $I_{\text{s,2PA}}$, in quantum dots can be described quantitatively by an inhomogeneously broadened, saturated 2PA model as (He et al. 2005b)

$$I_{\text{s,2PA}}^2 = \frac{\hbar\omega\pi\Delta\omega p(2\omega)N_0}{{}_p\sqrt{\pi/2}\left(1 + \frac{g_k}{g_n}\right)\beta}\frac{g_k}{g_n} \tag{23.1}$$

where

- $g_k(g_n)$ is the electronic degeneracy of the upper (lower) state
- τ_p is the half width at the maximum of the femtosecond laser pulse
- $\hbar\omega$ is the photon energy
- N_0 is the number density of the quantum dots in the sample material
- $p(2\omega)$ is the probability of a homogeneous class of absorbers with a central frequency of 2ω

The quantity $\Delta\omega$ is related to the dephasing time (T_2) of the excitation, i.e., the width of the homogeneous line shape.

Two-photon transition does not involve a real intermediate state. The single photon energy is less than that of the quantum dots bandgap. So 2PA is a non-resonant optical nonlinearity. The 2PA coefficients, β, can be determined by two-photon excited fluorescence, nonlinear transmission, Z scan, and pump-probe techniques (Sutherland 2003). For example, the differential equation describing the intensity change in a one-photon transparent but two-photon absorbing material is given by $dI_0/dz = -\beta I_0^2$, where I_0 is the peak intensity of the input laser beam inside the sample. The nonlinear transmission in such a material, excited by a focused continuous-wave (cw) Gaussian beam, can be expressed as follows (Sutherland 2003):

$$T(I_0) = \frac{\left[\ln(1 + \beta l I_0)\right]}{(\beta l I_0)} \tag{23.2}$$

where l is the thickness of the quantum dot sample. A comparison of the expressions for the energy transmittance of a cw top-hat, pulsed Gaussian and sech^2 beam are available in Sutherland (2003). From Equation 23.2 one can see that at a given input intensity (I_0) level, if the nonlinear transmission value is measured, the β value can be readily determined. The 2PA coefficient β is a macroscopic parameter characterizing the quantum dot composite material. The intrinsic 2PA coefficient of quantum dots, β_{QD}, can be derived as $\beta_{\text{QD}} = \beta_{\text{composite}} n_{0,\text{composite}}^2 / \left(n_{0,\text{QD}}^2 f_\text{v} |f|^4\right)$, where f_v is the volume fraction of the quantum dots in the matrix and $f = 3n_{0,\text{matrix}}^2 / (n_{0,\text{QD}}^2 + 2n_{0,\text{matrix}}^2)$ is the local field correction that depends on the dielectric constant of the quantum dots and the matrix material. In addition, the 2PA cross section can be calculated by the use of the definition: $\sigma_2 = \beta\hbar\omega/N_0$.

Two-photon absorption in semiconductor quantum dots has many important potential applications, some of which are discussed in more detail below.

23.3.1 Biological Imaging

Using quantum dot labeling, 2PA provides a possible way of performing biological imaging that is not possible by traditional one-photon methods as visible wavelengths cannot penetrate human tissue. As tagging materials, semiconductor quantum dots have advantages over fluorescent dyes, such as broad excitation and narrow emission bands, emission wavelength tunability, photostability, and enhanced brightness. The 2PA in robust water-soluble CdSe/ZnS core-shell quantum dots was found to be well-suited for use as fluorescent labels in multiphoton microscopy for biological imaging (Larson et al. 2003). The near-infrared two-photon excitation of such quantum dots in the "tissue optical window" (0.7–1.1 μm), in which water and hemoglobin absorb very little light, allows the extensive imaging of living systems, making further developments in medical diagnostics possible.

23.3.2 Amplified Stimulated Emission and Lasing

Quantum dot lasers have potential advantages, such as a temperature-insensitive lasing threshold and wide-range color tunability. Due to the quantum-size confinement, the 2PA cross-sections in semiconductor quantum dots are enhanced compared to the corresponding bulk material. This allows for a lower lasing threshold with a 2PA pumping mechanism. In addition, quantum dot lasers with 2PA excitation in the "tissue optical window" have important application prospects on laser-assisted biological-medical diagnostics and therapy. Recently, upconverted laser emission from a solution-processed CdSe/CdS/ZnS quantum dot waveguide-resonant cavity has been successfully demonstrated with femtosecond excitation at a wavelength of 800 nm (Zhang et al. 2008).

23.3.3 Optical Power Limiting

Optical power limiting and stabilization can be used to protect sensitive equipment or to control noise in laser beams (He et al. 2008). In this case, the output laser intensity approaches a constant value when the input intensity increases beyond a certain threshold, limiting the amount of optical power entering a system. Compared to optical-limiting materials based on organic chromophores, crystals, and polymers, semiconductor quantum dots have a large 2PA cross section and better photostability.

23.3.4 Two-Photon Sensitizer

Semiconductor quantum dots are excellent sensitizers for near-infrared 2PA due to their large and size-tunable 2PA cross sections. For example, quantum dots linked to phthalocyanines

(Pcs) can be excited using the 2PA of the quantum dots without any significant direct excitation of Pc molecules. In this case, Pcs are electron acceptor molecules, and a near-infrared excitation of the Pc molecules within the spectral therapeutic window (0.7–1.2 μm) can be realized via a two-photon sensitization of the quantum dot followed by an energy transfer to the Pc (Dayal and Burda 2008).

23.4 Conclusion

Excitons in nanoscale systems, particularly in semiconductor quantum dots, provide a link between the bulk and the molecular regimes. In quantum dots, the discrete density of states associated with the exciton results in an increased oscillator strength in the optically active levels and an increased nonlinear cross section. The increased nonlinear optical properties of quantum dots provide an opportunity for novel devices and experimental applications including those associated with 2PA. The unique properties of nanoscale excitons and the materials associated with them will provide inspiration and direction to future research.

References

Alivisatos, A. P. 1996. *Science* 271: 933.

Banfi, G. P., Degiorgio, V., and Ricard, D. 1998. *Adv. Phys*. 47: 447.

Banyai, L. and Koch, S. W. 1993. *Semiconductor Quantum Dots*. River Edge, NJ: World Scientific.

Basu, P. K. 1997. *Theory of Optical Processes in Semiconductors: Bulk and Microstructures*. New York: Oxford University Press.

Bruchez, M., Moronne, M., Gin, P., Weiss, S., and Alivisatos, A. P. 1998. *Science* 281: 2013.

Brus, L. E. 1984. *J. Chem. Phys*. 80: 4403.

Dayal, S. and Burda, C. 2008. *J. Am. Chem. Soc*. 130: 2890.

Dubertret, B., Skourides, P., Norris, D. J. et al. 2002. *Science* 298: 1759.

Efros, Al. L. and Efros, A. L. 1982. *Semiconductors* 16: 1209–1214.

Ekimov, A. I. and Onushchenko, A. A. 1982. *Semiconductors* 16: 1215–1219.

Ekimov, A. I., Onushchenko, A. A., and Tsekhomskii, V. A. 1980. *Sov. Glass Phys. Chem*. 6: 511–512.

Empedocles, S. A. and Bawendi, M. G. 1997. *Science* 278: 2114.

Etienne, M., Biney, A., Walser, A. D. et al. 2005. *Appl. Phys. Lett*. 87: 181913.

Fedorov, A. V., Baranv, A. V., and Inoue, K. 1996. *Phys. Rev. B* 54: 8627.

Gaponenko, S. V. 1998. *Optical Properties of Semiconductor Nanocrystals*. New York: Cambridge University Press.

Hadziioannou, G. and Malliaras, G. G. eds. 2006. *Semiconducting Polymers: Chemistry, Physics and Engineering*, Weinheim, Germany: Wiley-VCH.

He, J., Ji, W., Ma, G. H. et al. 2005a. *J. Phys. Chem. B* 109: 4373.

He, J., Mi, J., Li, H. P., and Ji, W. 2005b. *J. Phys. Chem. B* 109: 19184.

He, G. S., Yong, K. T., Zheng, Q. et al. 2007a. *Opt. Express* 15: 12818.

He, G. S., Zheng, Q., Yong, K. T. et al. 2007b. *Appl. Phys. Lett*. 90: 181108.

He, G. S., Tan, L., Zheng, Q., and Prasad, P. N. 2008. *Chem. Rev*. 108: 1245.

Jorio, A., Dresselhaus, M. S., and Dresselhaus, G. 2008. *Carbon Nanotubes: Advanced Topics in Synthesis, Structure, Properties and Applications*. New York: Springer.

Kasha, M. 1976. Molecular excitons in small aggregates. In *Spectroscopy of the Excited State*, ed. B. DiBartolo. New York: Plenum Press.

Klimov, V. I. 2000. *J. Phys. Chem. B* 104: 6112.

Klimov, V. I. 2006. *J. Phys. Chem. B* 110: 16827.

Klimov, V. I., Mikhailovsky, A. A., McBranch, D. W., Leatherdale, C. A., and Bawendi, M. G. 2000a. *Science* 287: 1011.

Klimov, V. I., Mikhailovsky, A. A., Xu, S. et al. 2000b. *Science* 290: 314.

Klimov, V. I., Ivanov, S. A., Nanda, J. et al. 2007. *Nature* 447: 441.

Larson, D. R., Zipfel, W. R., Williams, R. M. et al. 2003. *Science* 300: 1434.

Ma, Y. Z., Valkunas, L., Bachilo, S. M., and Fleming, G. R. 2005. *J. Phys. Chem. B* 109: 15671–15674.

Michalet, X., Pinaud, F. F., Bentolila, L. A. et al. 2005. *Science* 307: 538.

Murray, C. B., Norris, D. J., and Bawendi, M. G. 1993. *J. Am. Chem. Soc*. 115: 8706–8715.

Padilha, L. A., Fu, J., Hagan, D. J. et al. 2007. *Phys. Rev. B* 75: 075325.

Rossetti, R., Nakahara, S., and Brus, L. E. 1983. *J. Chem. Phys*. 79: 1086–1088.

Sariciftci, N. S. 1997. *Primary Photoexcitations in Conjugated Polymers: Molecular Exciton versus Semiconductor Band Model*. Singapore: World Scientific.

Schmidt, M. E., Blanton, S. A., Hines, M. A., and Guyot-Sionnest, P. 1996. *Phys. Rev. B* 53: 12629.

Scholes, G. D. 2008a. *Adv. Funct. Mater*. 18: 1157.

Scholes, G. D. 2008b. *ACS Nano* 2: 523–537.

Scholes, G. D. and Rumbles, G. 2006. *Nat. Mater*. 5: 683.

Scholes, G. D. and Rumbles, G. 2008. Excitons in nanoscale systems: Fundamentals and applications. In *Annual Review of Nano Research*, eds. G. Cao and C. J. Brinker. Hackensack, NJ: World Scientific.

Seydack, M. 2005. *Biosens. Bioelectron*. 20: 2454.

Shalaev, V. M., Poliakov, E. Y., and Markel, V. A. 1996. *Phys. Rev. B* 53: 2437.

Spanhel, L., Haase, M., Weller, H., and Henglein, A. 1987. *J. Am. Chem. Soc*. 109: 5649–5655.

Sutherland, R. L. 2003. *Handbook of Nonlinear Optics*. New York: Marcel Dekker.

Wang, F., Dukovic, G., Brus, L. E., and Heinz, T. F. 2005. *Science* 308: 838–841.

Yoffe, A. D. 1993. *Adv. Phys*. 42: 173.

Yoffe, A. D. 2001. *Adv. Phys*. 50: 208.

Zhang, C. F., Zhang, F., Zhu, T. et al. 2008. *Opt. Lett*. 33: 2437.

Zhao, H. B. and Mazumdar, S. 2004. *Phys. Rev. Lett*. 93: 157402.

24

Optical Properties of Metal Clusters and Nanoparticles

Emmanuel Cottancin
Université de Lyon

Michel Broyer
Université de Lyon

Jean Lermé
Université de Lyon

Michel Pellarin
Université de Lyon

24.1 Introduction **24**-1
24.2 Theoretical Description **24**-2
General Considerations on Bulk Metals • Optical Properties of Confined Metallic Systems
24.3 Experimental Results: State of the Art **24**-12
Nanoparticle Synthesis • Techniques of Optical Spectroscopy • Experimental Results
24.4 Conclusion and Outlooks **24**-21
Acknowledgments **24**-21
References **24**-22

24.1 Introduction

The discovery of fullerenes (Kroto et al. 1985) and the experimental evidence of the so-called magic cluster sizes (Knight et al. 1984, de Heer 1993), more than 25 years ago, were at the origin of unceasing investigations on small clusters and nanoparticles and have contributed to the emergence of nanosciences, at the crossing point of several branches such as physics, chemistry, or even biology. In view of their high surface-to-volume ratio, clusters possess properties, different from those of bulk matter, that are very sensitive to their size and shape, rendering them very attractive from both the fundamental and technological points of view. In particular, the metallic species of a few nanometers in diameter disclose electronic properties intermediate between those of molecular systems and those of bulk matter. Whereas the sparse energy levels are quantized in atoms and molecules and continuously distributed in the energy bands of the crystal (Ashcroft and Mermin 1976), they tend to bunch together in metallic clusters and thus pattern the so-called electronic shells. This shell structure was evidenced experimentally (magic sizes) in the early 1980s in several metals (de Heer 1993) and nicely interpreted in the frame of the *jellium* model (Brack 1993). However, it was shown that magic sizes can also be correlated with atomic shell closures (Martin 1996) and even that both electronic and geometric structures may compete with each other, depending on the temperature (Martin et al. 1990).

The optical properties of metallic clusters being directly underlain by their electronic structure, their optical study is, in this respect, of fundamental interest. For the rest, the fascinating colors of glasses doped with metallic powders have been known for ages even if their origin long remained mysterious. The synthesis of such materials is utilized since antiquity in the art of making jewels, ornamental glassware, or stained glass in the Middle Ages. One can quote the famous cup of Licurgus (Barber and Freestone 1990) from the fourth century AC (appearing red in transmission and green in reflection) or the stained-glass windows of the Chartres Cathedral in France. The colorful shade of these materials (such as glass ruby) was discovered empirically and various colors were obtained during the seventeenth century by alchemists, metallurgists, or glassworkers. For instance, Glauber mentions that *purple can be obtained by precipitating gold from its solution in aqua regia with a solution of tin compound* (Hunt 1976). It was however only in the nineteenth century that their optical properties started to become more systematically studied by Faraday who succeeded in producing gold colloids by reducing gold salts (Faraday 1857b) and showed that color effects are intimately correlated to the size and morphology of colloids. In the Bakerian lecture (Faraday 1857a), Faraday spoke about the experimental relations of gold to light in the following terms: "Light has a relation to the matter which it meets with in its course, and is affected by it, being reflected, deflected, transmitted, refracted, absorbed, &c. by particles very minute in their dimensions." In the lineage of Faraday, Zsigmondy worked on colloidal suspensions and set up the ultra-microscope based on the scattering of particles observed in dark field by illuminating the material to be viewed with a light source placed at right angle to the plane of the objective (Zsigmondy 1926). In 1925, Zsigmondy was awarded the Nobel Prize for Chemistry for his work on the heterogeneous nature of colloidal solutions.

On a theoretical point of view, the pioneering experiments by Faraday have been interpreted later by Mie (1908) who solved Maxwell's equations of a metallic sphere in a homogeneous medium submitted to an external electromagnetic field (plane wave). Mie showed that the original properties of metallic nanoparticles are a consequence of the "dielectric confinement" (limited volume of the material) of particles whose sizes are smaller or of the same order as the wavelength of the excitation field. At the same time, Maxwell-Garnett developed an effective field model (Maxwell-Garnett 1904) in order to describe the optical properties of a medium containing "*minute metal spheres*." The main feature in the optical extinction spectra of small metallic clusters is the emergence of a giant resonance in the near UV–visible range, called surface plasmon resonance (SPR) that is related to the collective motion of the conduction electrons induced by the applied field. This resonance is clearly noticeable only in the case of simple metal (alkali, trivalent) and noble metal (gold, silver, and copper) clusters. Its spectral position and width depend on the morphology of the particles (size, shape, and internal structure for alloyed systems), but also on their dielectric environment (medium in which the particles are embedded, local neighborhood) (Kreibig and Vollmer 1995). The development of numerous cluster sources (Sattler et al. 1980, Smalley 1983, Milani and de Heer 1990, Siekmann et al. 1991) in the 1980s enabled to probe the intrinsic properties of the clusters of very small size for which quantum size effects were expected. From a fundamental point of view, alkali clusters constitute a perfect model owing to their simple electronic structure and have been widely studied in the gas phase (Pedersen et al. 1991, Blanc et al. 1992, Bréchignac et al. 1992, de Heer 1993). However, they are immediately oxidized in contact with air once deposited on a surface, and thus are not suitable for applications. In spite of their more complex electronic structure, noble metal clusters in solution or embedded in a transparent matrix are more promising for potential applications because they are more robust toward oxidization. By varying the morphology, structure, or environment of these clusters, their optical response may be more or less controlled, making them attractive in several areas (linear and nonlinear optics [Kreibig and Genzel 1985], nanomaterials, nano-photonics, plasmonics, biosensors [Raschke et al. 2003]). Conversely, as the optical response is closely linked with the electronic structure, the SPR can also be used as a probe of the structure of metallic clusters, especially bimetallic systems or nanoalloys. Furthermore, the exaltation of the electromagnetic field in the vicinity of the particle can be exploited to increase the coupling of molecules with light, for developing biological markers for instance (McFarland and Duyne 2003).

Until the past few years, most of the experiments in this field were performed on cluster assemblies, and except the investigation of size-selected clusters in the gas phase, the results are blurred by averaging effects due to the unavoidable cluster size and shape dispersions in samples (Kreibig and Vollmer 1995, Cottancin et al. 2006). To overcome this drawback, new methods of spectroscopy have been developed within the past 10 years, in order to study a single nanoparticle (Tamaru et al. 2002, Raschke et al. 2003, Arbouet et al. 2004, Dijk et al. 2006, Billaud et al. 2007). This has opened up a new field of research, allowing for instance to investigate reliably more complex nano-objects or to study in detail shape effects on the optical response best than ever.

The aim of this chapter is to give some keys to understand the linear optical properties (absorption, scattering, and extinction) of metallic clusters and nanoparticles,* and to present the state of the art of the research in this field. Section 24.2 deals with the theoretical description of the optical response of a metallic particle submitted to an external field. After a brief focus on the optical response of a bulk metal, particular attention is paid to the dipolar approximation that is appropriate for clusters sizes much smaller than the wavelength of excitation. Size, shape, and structural effects that can be expected are then fully discussed. The case of small clusters, for which quantum finite size effects are expected, is concisely sketched, disregarding the case of very small clusters (below 100 atoms per cluster) for which ab initio calculations are necessary (Rubio et al. 1997, Bonacic-Koutecky et al. 2001, Harb et al. 2008). Finally, the broad outlines of the Mie theory required to describe the optical response of larger nanoparticles are given.

Section 24.3 is divided into three subsections. The first one briefly sets out various methods for producing clusters together with spectroscopic techniques for probing their optical properties. The second one gives an overview of the major results obtained on simple and noble metal clusters concerning size, shape, and multipolar effects in clusters and nanoparticles. Illustrations are taken equally from results obtained for cluster assemblies or single nanoparticles. Finally, the case of bimetallic clusters is described to illustrate the possibility of using the SPR as a structural probe.

24.2 Theoretical Description

24.2.1 General Considerations on Bulk Metals

In the bulk phase, the close packing of atoms involves an overlap of their outer atomic orbitals that strongly interact. This leads to a broadening of the discrete levels of the free atoms into bands (a continuum of electronic states), of very high density of states in metals. In this last case, the highest occupied energy band is called the conduction band and is filled with electrons originating from the outermost atomic orbitals. These electrons weakly interact with the ionic cores, and can be considered as quasi-free particles. They are delocalized in the metal and responsible for most of electrical and thermal transport properties. The highest occupied energy level in the conduction band (not completely filled in metals) is the Fermi level ε_F for which the electron velocity is the Fermi velocity v_F. Lower-energy atomic orbitals will give rise to deeper energy bands (valence bands) in the solid. In noble metals for instance, the highest of these bands originates from the *nd* atomic orbitals ($n = 3, 4, 5$ for Cu, Ag, and Au, respectively). It is narrow owing to its weak hybridization with the conduction band (Figure 24.1). When a metal interacts with light, it may absorb a photon of energy $h\nu$ promoting an electron from an occupied state to an unoccupied state, the latter being located in the incompletely filled conduction band. As for the occupied state, it belongs either to the conduction band or to the valence band. In the first case, the corresponding transition is called "intraband" transition and occurs in the infrared (IR)–visible range. The second case corresponds to "interband" transitions that can occur only if the photon energy is larger than the limiting value $\hbar\Omega_{ib}$ corresponding to the energy threshold required for reaching the Fermi level from the top of the valence band (see Figure 24.1). In simple metals like alkali or trivalents, this threshold is out of the optical domain (near UV–visible–near IR). Thus, only optical transitions within the conduction band may occur. For noble metals, it lies in the UV–visible range and interband transitions may happen in the visible-UV range (see Figure 24.1). The energy threshold occurs in the UV range for silver ($\hbar\Omega_{ib} \sim 4\,\text{eV}$, $\lambda_{ib} \sim 310\,\text{nm}$) and in the visible range for gold and copper ($\hbar\Omega_{ib} \sim 1.9\,\text{eV}$, $\lambda_{ib} \sim 650\,\text{nm}$).

---

* The term "cluster" is generally reserved for sizes lower than a few nanometers in diameter, whereas the term "nanoparticle" is used in the range between a few nanometers to a few hundreds of nanometers.

The optical response of the metal can be described entirely by its dielectric function $\varepsilon(\omega)$, which reflects its electronic structure. If a metal is submitted to an external electromagnetic (EM) field $\vec{E} = \vec{E}_o \cos(\omega t) = \Re(\vec{E}_o e^{-i\omega t})$, it is polarized such that its polarization (defined as the dipolar moment per volume unit) is $\vec{P} = \varepsilon_o \chi \vec{E}$, where ε_o is the vacuum permittivity and χ the dielectric susceptibility. The displacement vector can then be written as $\vec{D} = \varepsilon_o \vec{E} + \vec{P} = \varepsilon_o \varepsilon(\omega) \vec{E}$, where $\varepsilon(\omega) = 1 + \chi(\omega)$ denotes the relative dielectric function which is characteristic of the metal. A part of the response may be in phase with the exciting field ($\alpha \cos(\omega t)$), whereas the absorption effects induce a response in quadrature phase ($\alpha \sin(\omega t)$). Therefore, the dielectric function can be decomposed into a complex form $\varepsilon = \varepsilon_1 + i\varepsilon_2$. Let us recall that it is directly correlated to the optical index of the medium, $n_{\text{opt}} = n + i\kappa = \sqrt{\varepsilon}$, in which n is the refractive index and κ the extinction coefficient (Fox 2001). n_{opt} is real if the medium is non-absorbing.

In the case of simple metals, the Drude model (Drude 1900a,b, Ashcroft and Mermin 1976), developed originally to explain why metals are good conductors of heat and electricity, remains successful to interpret their optical properties (such as the fact that metals are good reflectors for frequencies lower than a threshold frequency, called plasma frequency [Fox 2001]). This model makes the basic assumption that most of the metal properties can be explained in first approximation by those of the conduction electrons if they are considered as independent and quasi-free. In this frame, when an oscillating electric field is applied, the free electrons oscillate and undergo collisions with other particles (electrons, ions, defects) with a characteristic scattering time $\tau = (1/\gamma_o)$, where γ_o is the average collision rate of electrons. Since the electrons are independent, their global response is the sum of all individual responses. By applying the principle of dynamics for an electron of effective mass m_e and charge $-e$, one obtains:

$$m_e \frac{d^2\vec{r}}{dt^2} = -\gamma_o m_e \frac{d\vec{r}}{dt} - e\vec{E}_o e^{-i\omega t}, \tag{24.1}$$

where $\vec{r}$ is the complex position vector of the electron. One can easily solve this equation and deduce the polarization in the metal:

$$\vec{P} = -\rho e\vec{r} = \frac{\varepsilon_o \omega_P^2}{m_e \omega(\omega + i\gamma_o)} \tag{24.2}$$

where ρ is the number of electrons per volume unit (electronic density) and

$$\omega_P = \sqrt{\frac{\rho e^2}{m_e \varepsilon_o}} \tag{24.3}$$

the plasma angular frequency that is introduced to interpret the high reflectivity of metals (Fox 2001) for frequencies lower than ω_P.

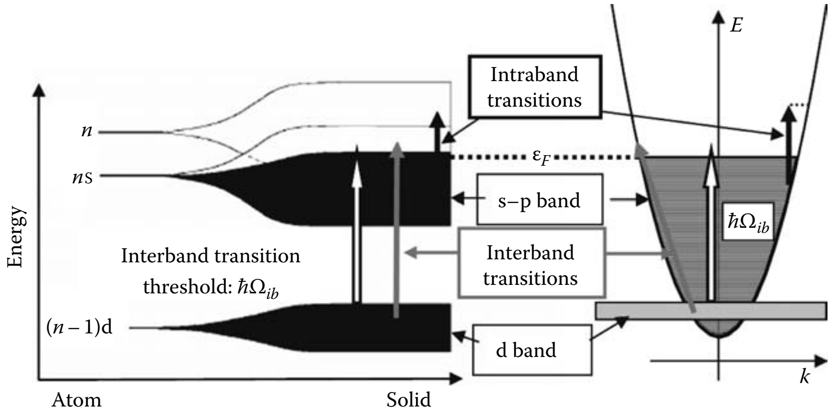

FIGURE 24.1 (Left) Schematic diagram of the transition from discrete electronic levels in atoms to electronic bands in the solid for noble metals. (Right) Schematic density of states for the $(n-1)$d and ns–p bands (left) of a noble metal. The ns–p band corresponding to quasi-free electrons in the metal is almost parabolic. As the atoms are brought closer together, their outer orbitals begin to overlap with each other and their strong interaction involves the formation of energy bands. Optical transitions inside the s–p band take place in the IR–visible range whereas interband transitions between the d-band and the conduction band may occur in the visible–near UV range.

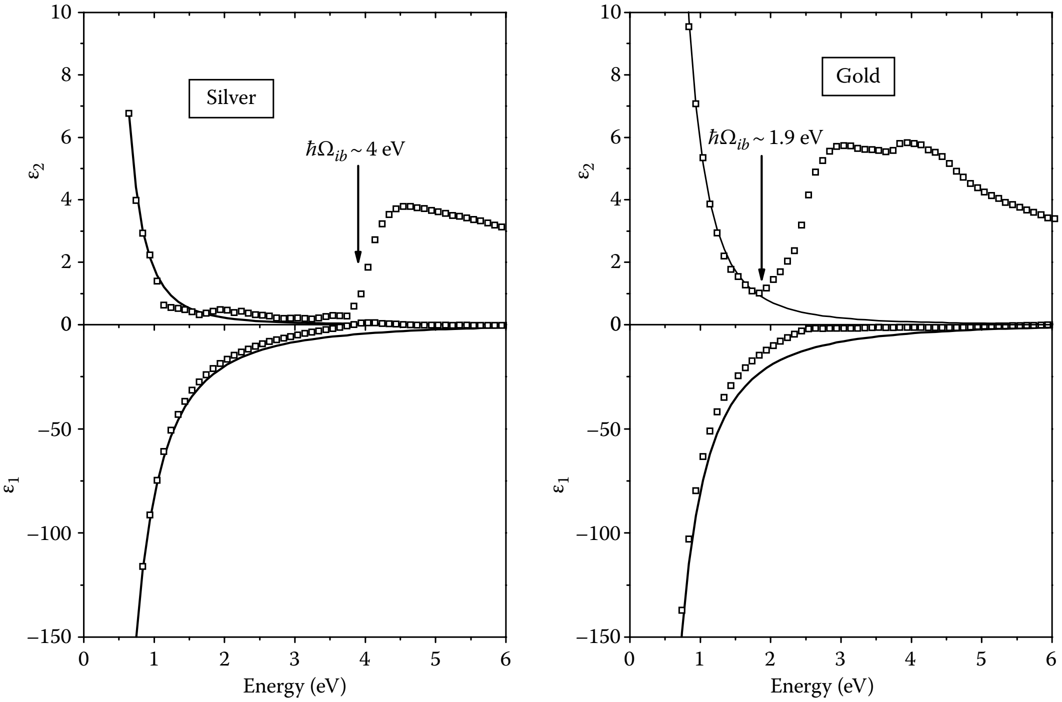

FIGURE 24.2 Real and imaginary parts ε_1 and ε_2 of the dielectric function of silver (left) and gold (right). Comparison between the Drude model (gray lines) and the experimental measurements of Johnson and Christy (1972) (squares). The influence of d-core electrons is patent in the imaginary part of ε where an increase of ε_2 appears around 1.9 eV for gold and 4 eV for silver corresponding to the interband transition threshold $\hbar\Omega_{ib}$.

The so-called Drude dielectric function can then be written as (see also Figure 24.2)

$$\varepsilon^D(\omega) = 1 - \frac{\omega_P^2}{\omega(\omega + i\gamma_o)} \quad \text{or}$$

$$\begin{cases} \varepsilon_1^D(\omega) = 1 - \dfrac{\omega_P^2}{\omega^2 + \gamma_o^2} \approx 1 - \dfrac{\omega_P^2}{\omega^2} & (\text{if } \gamma_o \ll \omega) \\ \varepsilon_2^D(\omega) = \dfrac{\omega_P^2 \gamma_o}{\omega(\omega^2 + \gamma_o^2)} \approx \dfrac{\omega_P^2}{\omega^3}\gamma_o & (\text{if } \gamma_o \ll \omega) \end{cases} \tag{24.4}$$

This model is well suited to reproduce the dielectric function of simple metals, but in the case of noble metals, another term has to be added to the dielectric function in order to account for the polarization of the core electrons. The dielectric function then takes the following form:

$$\varepsilon = 1 + \chi^D + \chi^{IB} = \varepsilon^D + \varepsilon^{IB} - 1, \tag{24.5}$$

where ε^D denotes the component of the dielectric function associated with the conduction electrons and well described by the Drude model (D). It is related to the electron excitations inside the conduction band. ε^{IB} is related to interband (IB) transitions and depicts the contribution of the core electrons (d-electrons mainly) to the metal polarizability (Ashcroft and Mermin 1976). Figure 24.2 displays a comparison between the dielectric function calculated with the Drude model and the measured one, for gold and silver. If an almost clear agreement is observed in the IR range (below 1–1.5 eV), the experimental values distinctly deviate from the calculated ones in the visible-UV range, because of the interband transition contribution. The steep increase in ε_2 corresponds to the interband transition threshold $\hbar\Omega_{ib}$. In the case of copper (not shown here) the increase is steeper than in gold. It occurs at slightly higher wavelengths and is at the origin of the difference of the color of these two metals. As $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$ are connected to each other with the Kramers–Kronig relationship (Ashcroft and Mermin 1976), the influence of interband transitions also explains the disagreement observed in the real part of $\varepsilon(\omega)$ between the Drude model and the experiment, the difference being larger in gold because of the proximity of the IB transitions.

24.2.2 Optical Properties of Confined Metallic Systems

24.2.2.1 Scattering, Absorption, and Extinction

In general, light is absorbed and scattered by a particle (see Figure 24.3). Absorption corresponds to the energy dissipation (heating) in the particle, whereas scattering is an elastic interaction of the wave with the particle. The EM field is reradiated by the particle in the whole space without a change of wavelength (Rayleigh scattering). If one assumes a particle in a homogeneous transparent medium (of dielectric function ε_m) submitted to an incoming plane wave $\vec{E}_i = \vec{E}_o e^{i(\vec{k}\vec{r} - \omega t)}$ of average intensity $I_o = (1/2)c\varepsilon_o\sqrt{\varepsilon_m}E_o^2$ (units: $\text{W} \cdot \text{m}^{-2}$), the absorption and scattering cross sections (units: m^2) are defined as the ratio between the absorbed power W_{abs} (unit: W) (respectively the scattered power W_{sca}) and the intensity I_o: $\sigma_{\text{abs}} = W_{\text{abs}}/I_o$ and $\sigma_{\text{sca}} = W_{\text{sca}}/I_o$.

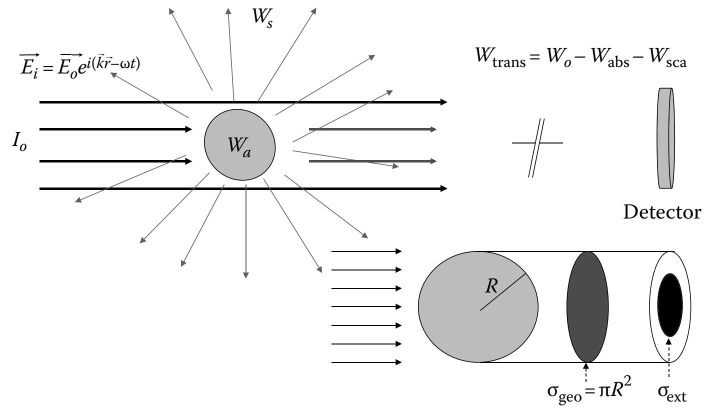
W_s
$\vec{E_i} = \vec{E_o}e^{i(\vec{k}\vec{r}-\omega t)}$
$W_{\text{trans}} = W_o - W_{\text{abs}} - W_{\text{sca}}$
I_o
W_a
Detector
R
$\sigma_{\text{geo}} = \pi R^2$
σ_{ext}

FIGURE 24.3 Schematic view of light absorption, scattering, and extinction of a particle. (Top) A particle excited by an incident plane wave reemits a part of the incident power by scattering (W_{sca}) and absorbs the power W_a. An in-axis sensor placed far away from the particle detects the incident power minus the lost power related to extinction. The detector is far enough for neglecting a direct collection of the forward scattered light. (Bottom) Naive picture of the extinction cross section σ_{ext} corresponding to the loss induced by the presence of a spherical particle in the course of the plane wave. σ_{ext} may be smaller or larger than the apparent surface of the particle πR^2.

The extinction cross section, related to the total power lost in the propagation direction of the incident plane wave far away from the particle, is defined as the sum of both absorption and scattering cross sections: $\sigma_{\text{ext}} = \sigma_{\text{abs}} + \sigma_{\text{sca}}$. This can be viewed naively in terms of perfectly opaque transverse areas shadowing the incoming plane wave (see Figure 24.3). It should be emphasized that, depending on the wavelength, σ_{ext} may be different from the mere geometrical apparent area of the particle (see Figure 24.3).

In the case of metals, in particular simple and noble metals, the dielectric confinement gives rise to a giant resonance (called SPR or Mie resonance) in the UV–visible range due to collective electronic excitations. It will be shown that the dielectric confinement refers to as classical size effects.

24.2.2.2 Dipolar or Quasi-Static Approximation

Let us consider a spherical metallic particle of radius R embedded in a transparent medium of real dielectric function ε_m submitted to an incident EM field $\vec{E}_i = \vec{E}_o e^{i(\vec{k}\vec{r}-\omega t)}$, whose diameter is very small as compared to the wavelength of excitation (diameter $\phi \ll \lambda$). In this approximation, called quasi-static approximation, the spatial variations of the field inside the particle can be ignored and the field can be written as $\vec{E}_i(t) = \vec{E}_o e^{-i\omega t}$ (see the smaller particle in Figure 24.4). Moreover, as will be shown below, the scattering is negligible as compared to absorption for diameters below 20–40 nm (depending on the element) and $\sigma_{\text{ext}} \simeq \sigma_{\text{abs}}$.

In the presence of the EM field, the small metallic sphere is polarized (see Figure 24.4). By solving the Poisson equation in the absence of free charges ($\Delta V(\vec{r}) = 0$) inside and outside the particle, and considering the boundary conditions at the interface, the internal field is shown to be as follows (Bohren and Huffman 1983, Berthier 1993): $\vec{E}_{\text{int}} = 3\varepsilon_m/(\varepsilon + 2\varepsilon_m)\vec{E}_i(t)$, and the external field can be viewed as the sum of the incident field and the one created by a fictitious point dipole $\vec{p}(\omega,t) = 4\pi\varepsilon_o\varepsilon_m R^3((\varepsilon-\varepsilon_m)/(\varepsilon+2\varepsilon_m))\vec{E}_i(t) = \alpha(\omega)\vec{E}_i(t)$, where $\alpha(\omega)$ is the dynamical polarizability (it is the reason why the quasi-static approximation is also called dipolar approximation). The mean power dissipated inside the particle can be expressed in terms of the imaginary part of the dynamic polarizability (component in quadrature phase with the incident field):

$$P_{\text{mean}} = \iiint \left\langle \vec{j}(t)\vec{E}_{\text{int}}(t) \right\rangle d\tau = \frac{E_o^2\omega}{2}\Im m(\alpha)$$

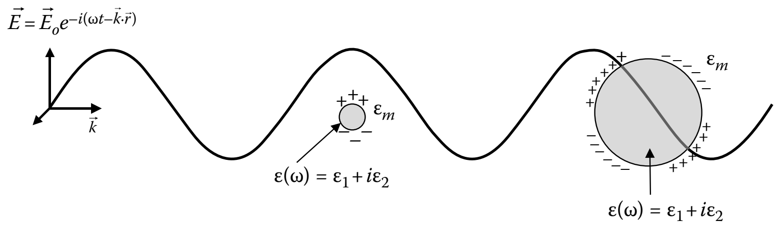
$\vec{E} = \vec{E}_o e^{-i(\omega t - \vec{k}\cdot\vec{r})}$
$\vec{k}$
ε_m
$\varepsilon(\omega) = \varepsilon_1 + i\varepsilon_2$
ε_m
$\varepsilon(\omega) = \varepsilon_1 + i\varepsilon_2$

FIGURE 24.4 Metallic sphere of dielectric function $\varepsilon(\omega)$ in a transparent medium of real dielectric function ε_m submitted to an external field $\vec{E}_i = \vec{E}_o e^{i(\vec{k}\vec{r}-\omega t)}$. Case of a small particle of diameter $\phi \ll \lambda$: the particle does not feel the spatial variations of the electric field: this is the quasi-static or dipolar approximation. The metallic sphere is polarized and the surface charge distribution creates a dipole. Case of a large particle of diameter $\phi \sim \lambda$: the spatial variations of the field inside the particle induces a more complex surface charge distribution, leading to multipolar modes like the displayed quadrupolar mode (see Mie theory Section 24.2.2.5).

where $\vec{j}(t) = -\rho e\vec{v} = \frac{\partial \vec{P}}{\partial t}$ and finally the absorption cross section takes the following form:

$$\sigma_{\text{abs}}(\omega) = \frac{9\omega\varepsilon_m^{3/2}}{c}\left(\frac{4}{3}\pi R^3\right)\frac{\varepsilon_2(\omega)}{(\varepsilon_1(\omega)+2\varepsilon_m)^2+\varepsilon_2(\omega)^2}, \tag{24.6}$$

where c denotes the velocity of light.

In general $\varepsilon_2(\omega)$ is relatively small in the case of simple and silver metals in the UV–visible range and $\sigma_{\text{abs}}(\omega)$ is strongly enhanced when

$$\varepsilon_1(\omega) + 2\varepsilon_m = 0 \tag{24.7}$$

This condition named resonance condition may be fulfilled for metals for which the real part of the dielectric function $\varepsilon_1(\omega)$ is negative (see Figure 24.2). The correlative resonance phenomenon is known as the SPR or Mie resonance and corresponds to the collective oscillation of the conduction electrons relative to the ionic background. By using Equations 24.4 and 24.5, the spectral position of the SPR can be approximately deduced from the resonance condition Equation 24.7. One obtains

$$\omega_{SPR} \simeq \frac{\omega_P}{\sqrt{\varepsilon_1^{IB}(\omega_{RPS}) + 2\varepsilon_m}}. \tag{24.8}$$

This equation indicates that the frequency resonance depends both on the material through ω_P and ε_1^{IB}, and on the medium in which the particle is embedded through ε_m. In alkali metals, as the IB transitions are negligible, one gets in vacuum ($\varepsilon_m = 1$) the simple relation $\omega_{SPR} = \omega_P/\sqrt{3}$, which generally holds for a perfectly conducting metallic sphere.

It can be underlined that, in the frame of the quasi-static approximation, the extinction and scattering cross sections can be deduced directly from the Mie theory in the small-size range (Bohren and Huffman 1983). They take the following form at the lowest order:

$$\sigma_{\text{ext}}(\omega) = \frac{9\omega\varepsilon_m^{3/2}}{c}\left(\frac{4}{3}\pi R^3\right)\frac{\varepsilon_2(\omega)}{(\varepsilon_1(\omega)+2\varepsilon_m)^2+\varepsilon_2(\omega)^2} \tag{24.9}$$

$$\sigma_{\text{sca}}(\omega) = \frac{3\omega^4\varepsilon_m^2}{2\pi c^4}\left(\frac{4}{3}\pi R^3\right)^2\frac{(\varepsilon_1(\omega)-\varepsilon_m)^2+\varepsilon_2(\omega)^2}{(\varepsilon_1(\omega)+2\varepsilon_m)^2+\varepsilon_2(\omega)^2} \tag{24.10}$$

σ_{sca} varies with λ as $1/\lambda^4$ (Rayleigh-scattering) ($\lambda = (2\pi c/\omega)$ being the wavelength in vacuum) and exhibits the same resonance condition as the absorption cross section. For a given nanoparticle of diameter R, $(\sigma_{\text{sca}}/\sigma_{\text{ext}}) \propto (R/\lambda)^3$. In the dipolar approximation ($\phi \ll \lambda$) the extinction is clearly dominated by absorption ($\sigma_{\text{sca}} \ll \sigma_{\text{abs}} = \sigma_{\text{ext}}$ [comparison of Equations 24.6 and 24.9]).

24.2.2.3 Intrinsic Size Effects

24.2.2.3.1 Dielectric Function of a Confined Metal

Except a mere volume factor, it is clear from Equation 24.6 that no size effects are predicted in the dipolar approximation. However, in this description, the dielectric function of the particle was assumed to be the one of the bulk. Actually, some modifications of the metallic dielectric function are expected because of confinement. Indeed, in a confined system, the collision rate of electrons is modified because of surface scattering that reduces the mean free path of electrons (Kreibig and Fragstein 1969, Kreibig and Genzel 1985, Kreibig and Vollmer 1995). The extra-scattering rate scales at v_F/ℓ, where v_F is the Fermi electron velocity and ℓ the limited mean free path for electron motion (of the order of the particle radius). For a spherical particle of radius R, the collision rate can be written as

$$\gamma(R) = \gamma_o + g\frac{v_F}{R}, \tag{24.11}$$

where g is the surface-scattering coefficient whose value is about unity (Kreibig and Vollmer 1995). The modification of the collision rate with size induces a modification of $\varepsilon^D(\omega, R)$. The dielectric function is then $\varepsilon(\omega, R) = \varepsilon^D(\omega, R) + \varepsilon^{IB}(\omega, \infty) - 1$, where $\varepsilon^{IB}(\omega, \infty)$ refers to as the interband dielectric function.

In the dipolar approximation, this size dependence of the dielectric function induces a size variation of the absorption cross section, as illustrated Figure 24.5 in the case of copper, silver, and gold clusters embedded in alumina. The main features in the spectra are a damping and broadening of the SPR with decreasing size. For gold and copper, a slight red-shift of the resonance is expected within the frame of this model, which is not observed experimentally due to quantum size effects (Cottancin et al. 2006) as will be shown in the following section.

This phenomenological size evolution of $\varepsilon(\omega, R)$ may be recovered from a simple quantum approach of confined electrons in an infinite potential well (Kubo model [Kawabata and Kubo 1966]). Nevertheless this simple model is too crude to give an interpretation of the size effects observed in small alkali and noble metal clusters (except silver) because an infinite potential well cannot lead to the well-known spillout of electrons beyond the classical radius of clusters.

24.2.2.3.2 Quantum Size Effects

In the classical Mie theory, the electronic density $\rho_e(r)$ perfectly matches the positive ionic charge, but quantum calculations (Brack 1993) predict a significant overflow of electrons beyond the classical radius* of the cluster $R_C = r_S N_e^{1/3}$. Because of the finite depth of the potential well, the electronic density does not vanish beyond the cluster radius R_C (exponential decay). This spillout corresponds to an increase of a few angstroms of the effective cluster radius (conduction electron cloud) that becomes significant in small clusters (ϕ < 5 nm). In particular, it has a great influence on the optical properties of metallic clusters, especially on the size evolution of the SPR spectral position.

---

* Keeping in mind that the Wigner–Seitz radius r_S is correlated with the volume occupied by a conduction electron $(V_e = (4/3)\pi r_s^3)$ in the bulk, one can easily deduce that the classical radius of a cluster containing N_e conduction electrons is: $R_C = r_S N_e^{1/3}$.

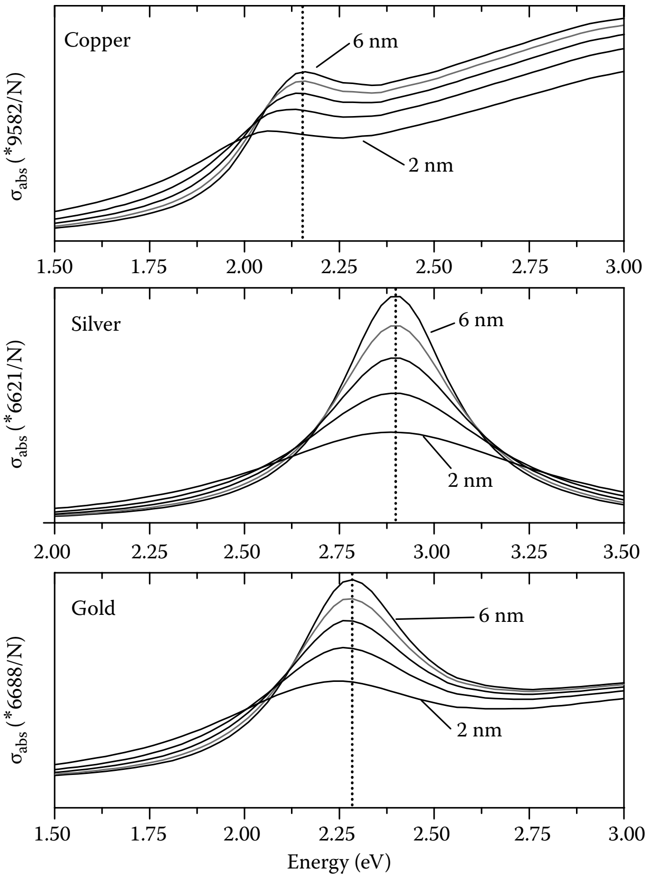

FIGURE 24.5 Size evolution of the theoretical absorption cross section σ_{abs} versus energy within the classical Mie theory in the dipolar approximation, taking into account the reduction of the mean free path in the particle, for copper, silver, and gold clusters embedded in porous alumina ($\varepsilon_m \simeq 2.7$). (From Cottancin, E. et al., *Theor. Chem. Acc.*, 116, 514, 2006.)

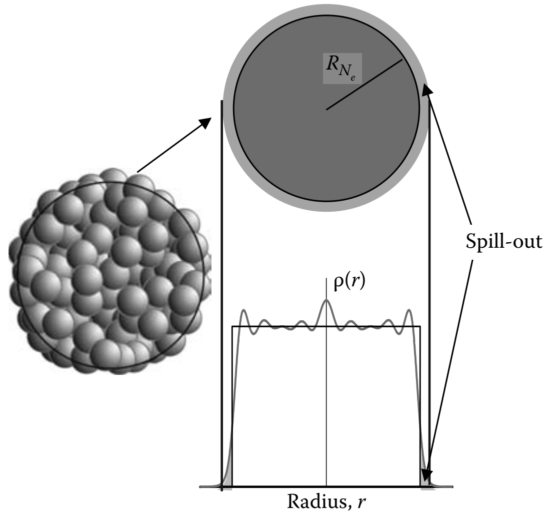

FIGURE 24.6 Schematic view of the jellium model: the ionic background is replaced by a homogeneous positive charge distribution with which the conduction electrons interact. (Right, top) Geometrical aspect of the jellium (dark gray) and the electron spillout (light gray). (Bottom) Corresponding evolution of both ionic (black line) and electronic (gray line) densities versus *r*. The electronic density overflows beyond the classical cluster radius: this is the so-called spillout.

On a quantum basis, the metallic clusters of a few nanometers in diameter are usually described by the standard jellium model (Brack 1993) in which the conduction electrons of a spherical cluster in vacuum are quantum mechanically treated, whereas the granular ionic structure is replaced by a spherical, homogeneous, positive-charge distribution: the jellium (of total positive charge *Q*). The problem to be solved is then the one of N_e electrons interacting together and with the potential created by the positively charged jellium. For $r < R_{N_e}$, the interaction potential energy between an electron and the jellium is equal to

$$V_{\text{jel}}(r) = \frac{1}{2}m(\omega_{RPS}^{cl})^2 r^2 - \frac{3}{8\pi\varepsilon_o}\frac{qQ}{R_C} \quad \text{where } \omega_{RPS}^{cl} = \frac{\omega_P}{\sqrt{3}}$$

The first term corresponds to the potential of a harmonic oscillator of angular frequency ω_{RPS}^{cl} (the SPR frequency in vacuum) in the case of simple metals for which $\varepsilon^{IB} = 0$ (see Equation 24.8).

The resolution of the problem may be achieved in the frame of the density functional theory (DFT) by solving self-consistently the Kohn–Sham equations (Hohenberg and Kohn 1964, Kohn and Sham 1965). When convergence is reached, the system can be considered formally as equivalent to a system of N_e-independent electrons moving in an effective central potential $V_{\text{eff}}(r)$ of finite depth. The most relevant results obtained in this framework are the electronic shell and supershell structures explaining the famous magic sizes, the beating patterns (Pedersen et al. 1991, Pellarin et al. 1995) and the rise of the spillout (see Figure 24.6) whose profile is almost independent of cluster size (Brack 1993). The interaction of a cluster with light may be investigated in the frame of the random phase approximation (Yannouleas et al. 1989) (RPA) or in the frame of the time-dependent local-density approximation (Eckardt 1985) (TD LDA), both using the jellium model. These formalisms account for the quantum size effects ignored in the classical approach. Figure 24.7 shows for instance results obtained with the TD LDA for the magic sizes of sodium clusters in vacuum. For small clusters ($N < 100$), the absorption spectrum is broad and highly structured. This fragmentation of the plasmon band is due to a phenomenon known as Landau damping and can be explained in a simple manner in terms of a discrete state (the plasmon) coupled to a continuum (Cohen-Tannoudji et al. 1996) (one-electron excitations). When the cluster size increases, the fragmentation persists, but the oscillator strength is concentrated in an ever-narrower spectral range around the plasmon frequency, adopting a quasi-Lorentzian distribution. A red-shift of the SPR with decreasing size is also observed, in qualitative agreement with experimental spectra obtained for alkali clusters as will be shown in the following section.

This red-shift of the resonance obtained through quantum calculations is intimately linked to the increasing spillout

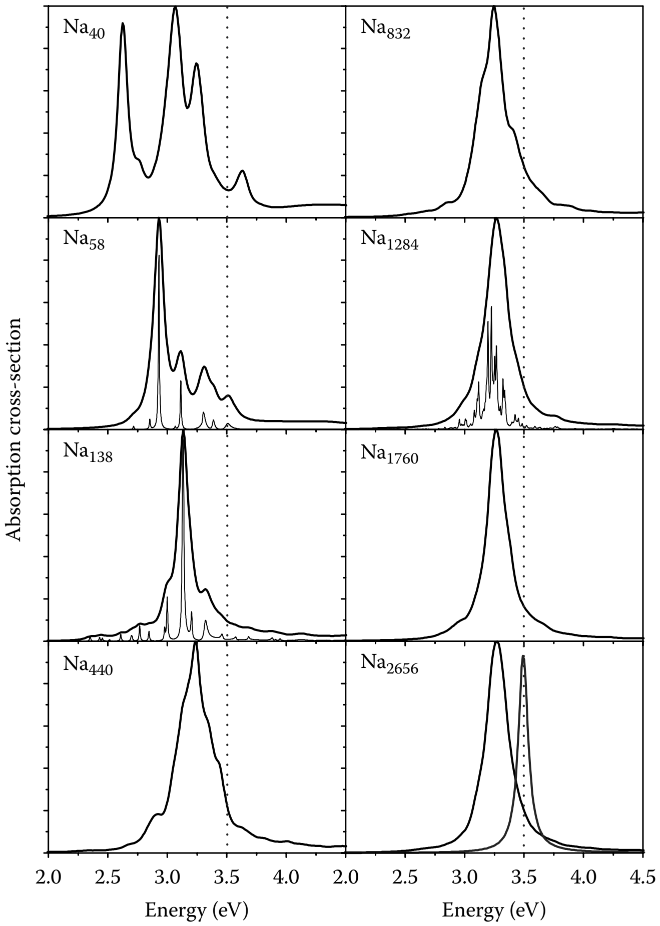

FIGURE 24.7 Absorption cross sections of sodium clusters calculated using the quantum TDLDA formalism within the framework of the jellium model. The gray line curve is the classical result obtained with the sodium dielectric function calculated with the Drude model (r_s = 0.208 nm). Spectra with thin lines for N = 58,138,1284 were calculated using a smoothing parameter 10 times smaller (Perez et al. 2007). The spectra are highly structured at small sizes and red-shifted as compared to classical predictions. They gradually converge toward the classical predictions (gray dotted line) as the size increases. (From Perez, A. et al., Nanoscience, *Clusters and Colloids*, Springer Verlag, Berlin/Heidelberg, Germany, 2007.)

influence. This correlation may also be viewed through a simple classical argument by considering the expression of the Plasmon angular frequency in vacuum $\omega_{RPS}^{cl} = (\omega_P/\sqrt{3})$.

Recall that $\omega_P = \sqrt{\rho e^2/(m_e\varepsilon_o)}$ and let δ_{so} be the spillout parameter. The effective electronic cluster radius can be written as $R_e = R_C + \delta_{so} = r_s N_e^{1/3} + \delta_{so}$. As the electronic density is the ratio between the number of electrons and their occupied volume in the cluster $((4/3)\pi R_e^3)$, the angular frequency is approximately

$$\omega_{SPR} \simeq \frac{\omega_P}{\sqrt{3}}\left(1 - \frac{3\delta_{so}}{2R_{N_e}}\right). \qquad (24.12)$$

The profile of the spillout being almost constant with the size, this relation shows that the SPR frequency shifts toward lower energies with decreasing size. The increasing influence of the spillout with decreasing size is pertinent to explain the red-shift of the SPR in alkali clusters. In the case of noble metal clusters, the screening of the core electrons (especially the d-electrons) and its modification near the cluster surface plays also a key role (see Section 24.3) that induces a blue-shift in copper and gold and a quenching of the resonance shift in silver (Lermé et al. 1998b, Palpant et al. 1998, Celep et al. 2004).

Let us recall that the case of very small clusters (a few tens of atoms per cluster) requires ab initio calculations (Bonacic-Koutecky et al. 1991, Rubio et al. 1997, Bonacic-Koutecky et al. 2001, Harb et al. 2008) and would demand a large development that is not the subject of this chapter.

24.2.2.4 Shape and Structure Effects

24.2.2.4.1 Influence of the Shape on the Optical Properties of Clusters

The optical response of clusters is noticeably modified by their shape when they deviate from a perfect sphere. For instance, in the case of ellipsoids, the degenerate SPR splits into as many SPRs as the different principal axes of deformation (three in general and two for spheroids). Although quantum calculations account well for these shape effects (Clemenger), they can be more easily understood in a classical approach assuming a quasi-static approximation. If the system exhibits a sufficient symmetry, like an ellipsoid, the problem can be solved analytically. Consider an ellipsoid of a metal of complex dielectric function $\varepsilon(\omega)$ in a medium of dielectric constant ε_m. In the frame of the quasi-static approximation, the Maxwell equations can be solved in ellipsoidal coordinates (Bohren and Huffman 1983). Taking into account the boundary conditions at the metal–external-medium interface, the absorption cross section for a light polarization along the principal axes $i = x, y, z$ (correlated to the semi-axes a, b, c of the ellipsoid [see Figure 24.7]) can be expressed as (Bohren and Huffman 1983)

$$\sigma_{\text{abs}}^{i} \simeq \sigma_{\text{ext}}^{i} = \frac{\omega\varepsilon_m^{3/2}}{c}\frac{V_{\text{ell}}}{L_i^2}\frac{\varepsilon_2(\omega)}{\left(\varepsilon_1(\omega)+\dfrac{1-L_i}{L_i}\varepsilon_m\right)^2+\varepsilon_2(\omega)^2}, \qquad (24.13)$$

where $V_{\text{ell}} = (4/3)\pi abc$ is the volume of the ellipsoid. The relation is similar to the one obtained for a sphere except for the additional geometrical factor:

$$L_i = \frac{abc}{2}\int_0^{\infty}\frac{du}{(u+l_i^2)\sqrt{(u+a^2)(u+b^2)(u+c^2)}} \quad \text{with } l_x = a, l_y = b, l_z = c.$$

The sum $L_x + L_y + L_z = 1$ and for a sphere $L_i = (1/3)$.

The optical response exhibits three resonances associated with the three principal axes of the ellipsoid if the excited light is not polarized. For a light polarization directed along the axis i, the resonance condition is in a first approximation: $L_i\varepsilon_1(\omega_{SPR}^i) + (1 - L_i)\varepsilon_m = 0$ and the resonance occurs at

$$\omega_{SPR}^{i} \simeq \frac{\omega_P}{\sqrt{\varepsilon_1^{IB}(\omega_{RPS}) + \left(\dfrac{1-L_i}{L_i}\right)\varepsilon_m}}.$$

In the case of a spheroid for which two semi-axes are equal ($b = c < a$: prolate or $a < b = c$: oblate), the geometrical factors are analytically correlated to the aspect ratio (see Figure 24.8) of the spheroid (Bohren and Huffman 1983).

The influence of the shape on the optical extinction cross section is displayed in Figure 24.8 for silver prolate and oblate nanospheroids. In both cases, two resonance modes appear on each side of the resonance of the sphere of equivalent volume, the difference between both modes being all the more important as the aspect ratio is small. The resonances, red- or blue-shifted as compared to the resonance of the sphere, are correlated to the electronic motion along the large axis or the small axis of the particle. Experimentally, if the measurements are performed on non-oriented cluster assemblies all the resonance modes are observed and the resulting optical response is the sum of the responses along each axis. Actually, the shape influence is easier to investigate experimentally on single nanoparticles.

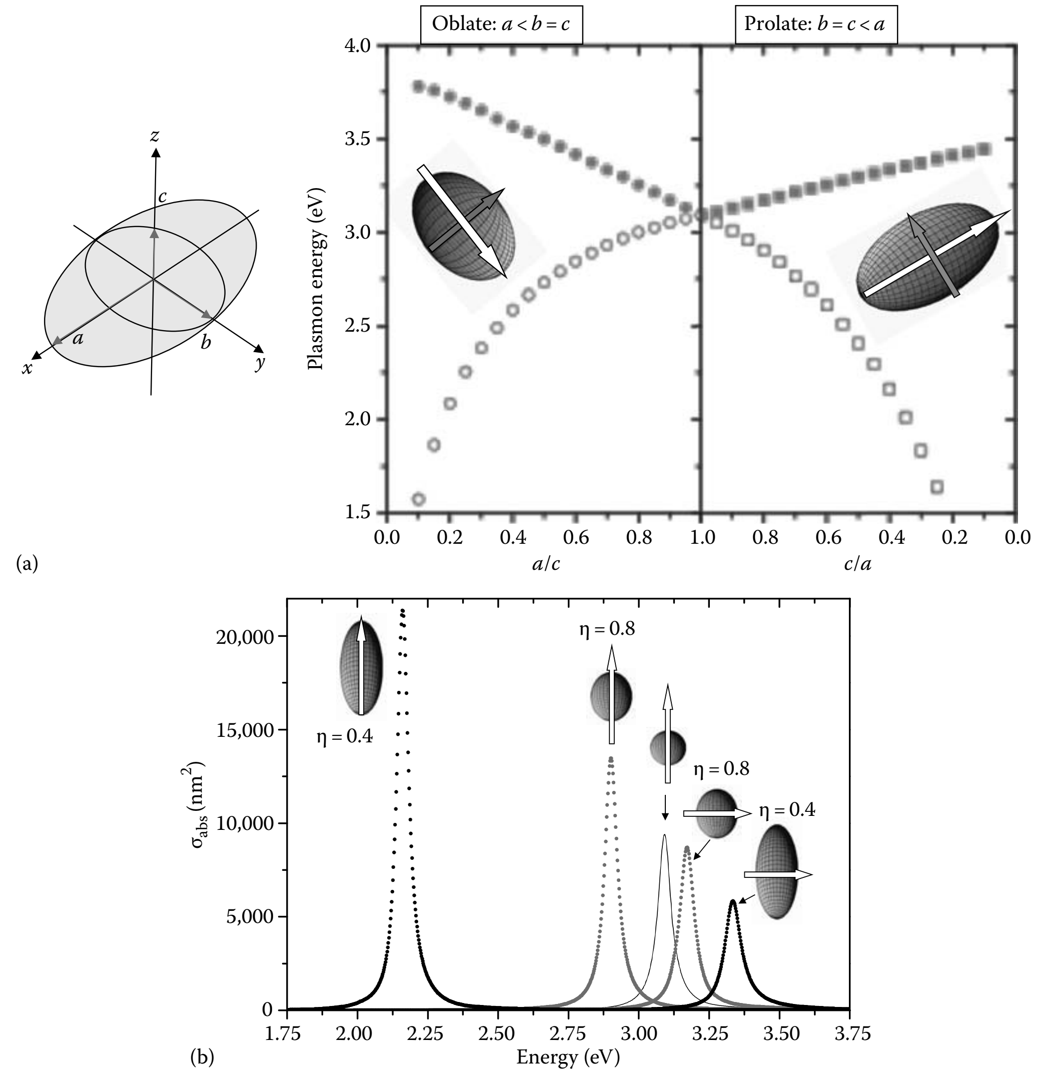

FIGURE 24.8 Shape evolution of the SPR for spheroidal silver clusters embedded in silica ($n = \sqrt{\varepsilon_m} = 1.5$). In each case, as one axis is degenerate the optical response gives rise to two resonances. The volume of the particle is constant and corresponds to the volume of a spherical particle of 20 nm in diameter. (a) Evolution of the maximum of the SPR versus the ratio between the small axis and the large axis (prolate and oblate cases). The gray and white points correspond to the white and gray arrows indicating the polarization direction of the incident light, respectively. (b) SPR for different aspect ratios $\eta = (c/a)$ in silver clusters embedded in silica (prolate case). Depending on the light polarization direction (see arrows), two resonances shifted as compared to the resonance obtained in the spherical case (black line) may be excited (gray and black circles). When the light polarization direction is along the large axis of the particle, the resonance is red-shifted, whereas it is blue-shifted if the particle is excited along its small axis, the smaller the aspect ratio, the larger the shift.

24.2.2.4.2 Influence of Chemical Structure

The optical response of metallic nanoparticles not only depends on their size and shape but also on their internal structure that can be heterogeneous. This is, for example, the case of mixed clusters made of several species (two for simplicity) in various proportions. Consider two chemical structures: a more or less homogeneous bimetallic alloying or a full spatial segregation of both metals (core/shell structure for instance). In the case of alloyed structures, the main difficulty is the knowledge of an effective dielectric function to use as an input for calculating the optical response. If the dielectric function has not been previously determined on corresponding alloys in the bulk phase (by optical measurements for instance), it is usually taken as the weighted average of the dielectric function of each constituent. But this hypothesis is crude because it assumes that the constituents are randomly segregated in nanodomains, each of them being described by a dielectric function taken as the one of the bulk phase. For alloyed structures at an atomic level, this approach may fail because the effective dielectric function of the alloy may be very different from those of each component.

The only structure for which calculations are currently feasible is the segregated core/shell structure that may develop in some particular systems. The problem can be solved exactly in the dipolar approximation for spherical or ellipsoidal shapes, and may be generalized to multi-shell structures. By solving Maxwell's equations and applying the successive boundary conditions on each interface, one can deduce the extinction and scattering cross sections. For a spherical core/shell cluster of core radius R_{core} and total radius R composed of two materials (with core and shell complex dielectric functions ε_c and ε_{sh}, respectively), the cross sections take the following form (Bohren and Huffman 1983):

$$\begin{cases} \sigma_{\text{ext}}(\omega) \simeq \sigma_{\text{abs}}(\omega) = \dfrac{3\omega\varepsilon_m^{1/2}}{c}\left(\dfrac{4}{3}\pi R^3\right)\Im\left(\dfrac{(\varepsilon_{sh}-\varepsilon_m)(\varepsilon_c+2\varepsilon_{sh})+f_v(\varepsilon_c-\varepsilon_{sh})(\varepsilon_m+2\varepsilon_{sh})}{(\varepsilon_{sh}+2\varepsilon_m)(\varepsilon_c+2\varepsilon_{sh})+2f_v(\varepsilon_{sh}-\varepsilon_m)(\varepsilon_c-\varepsilon_{sh})}\right) \\ \sigma_{\text{sca}}(\omega) = \dfrac{3\omega^4\varepsilon_m^2}{2\pi c^4}\left(\dfrac{4}{3}\pi R^3\right)^2 \Im\left|\dfrac{(\varepsilon_{sh}-\varepsilon_m)(\varepsilon_c+2\varepsilon_{sh})+f_v(\varepsilon_c-\varepsilon_{sh})(\varepsilon_m+2\varepsilon_{sh})}{(\varepsilon_{sh}+2\varepsilon_m)(\varepsilon_c+2\varepsilon_{sh})+2f_v(\varepsilon_{sh}-\varepsilon_m)(\varepsilon_c-\varepsilon_{sh})}\right|^2 \end{cases}$$

where $f_v = (R_{\text{core}}/R)^3$ is the volume ratio between the core and the cluster.

Such calculations can be performed on various bimetallic clusters or in the case of nanohybrid systems (dielectric–metal or semiconductor–metal). The reduction of the mean free path in the core and in the shell may be also taken into account (Granqvist and Hunderi 1978). As an illustration the optical response of a nanoshell containing a water core and a silver or copper shell embedded in water is displayed in Figure 24.9 for various shell thicknesses. This calculation, made in the frame of the dipolar approximation, is valid only for relatively small particles (Bohren and Huffman 1983). The most important feature in the spectra is a large red-shift of the resonance when the thickness of the metal decreases. In the case of copper, the SPR, damped and broadened in the full metal cluster, is progressively uncoupled from interband transitions (and thus enhanced) when the ratio between the thickness and the total radius decreases. A very similar evolution is expected for a gold nanoshell (Jain and El-Sayed 2007). One can also notice that for small shell thicknesses the resonance occurs at the same spectral range independently of the material (silver, copper, or gold). Such systems are very interesting because the position of the plasmon resonance can be easily tuned in the visible range by adjusting the shell thickness, which is very convenient for potential applications in medicine (Weissleder 2001).

24.2.2.5 Beyond the Dipolar Approximation: The Mie Theory

As emphasized before, when the size of the particle is very small compared to the wavelength of the exciting field, light scattering is negligible and the absorption cross section can be appropriately calculated by assuming the quasi-static (or dipolar) approximation. On the contrary, for large sizes as compared to the wavelength, the spatial distribution of the field amplitude in the particle volume cannot be disregarded (see the larger particle in Figure 24.4). One has to resort to the Mie theory for calculating the optical properties of particles subject to an incoming plane wave of angular frequency ω. This theory allows to determine the absorption and scattering cross sections, as well as the angular scattering pattern, of homogeneous spherical particles, of any radius, embedded in a homogeneous non-absorbing medium of dielectric function ε_m.

The theory (Bohren and Huffman 1983, Lermé et al. 2008) may be shortly sketched as follows. In a homogeneous medium the electric and magnetic field, $\vec{E}$ and $\vec{H}$ respectively, both satisfy the equations $\nabla\vec{V} = 0$ (zero divergence) and $\nabla^2\vec{V} + k^2\vec{V} = 0$ (vector Helmholtz wave equation), where $k = [\omega\varepsilon(\omega)\mu_0]^{1/2}$ is the wave vector in the medium and $\vec{V}$ stands for $\vec{E}$ or $\vec{H}$ ($\varepsilon(\omega) = \varepsilon_{\text{metal}}(\omega)$ in the particle and $\varepsilon(\omega) = \varepsilon_m$ outside the particle). In the Mie theory the incident, internal and scattered fields are expanded on the two sets of appropriate vector functions $\vec{M}_{nm}$ and $\vec{N}_{nm}$ (referred to as "the vector spherical harmonics") satisfying both previous Maxwell equations in the relevant homogeneous medium. These vector functions can be expressed in spherical coordinates (the origin being the particle center), and written as follows:

$$\vec{M}_{nm}(\vec{r},k) = K_{nm}\nabla\times\left[\vec{r}z_n(kr)P_n^m[\cos(\theta)]\exp(im\varphi)\right]$$

$$\vec{N}_{nm}(\vec{r},k) = \frac{1}{k}\nabla\times\vec{M}_{nm}(\vec{r}),$$

where

the $P_n^m(\theta)$'s are associated Legendre polynomials ($n = 1, 2, 3, \ldots$; $-n \le m \le n$)

$z_n(kr)$ are either spherical Bessel functions of the first kind $j_n(kr)$ (involved in the incident and internal electromagnetic field expansions), or spherical Hankel functions of the first kind $h_n^+(kr)$ (involved in the scattered field expansion, for ensuring the outgoing wave radiation condition, $\vec{M}_{nm}$ and $\vec{N}_{nm}$ are then denoted $\vec{M}_{nm}^+$ and $\vec{N}_{nm}^+$)

K_{nm} is a numerical factor

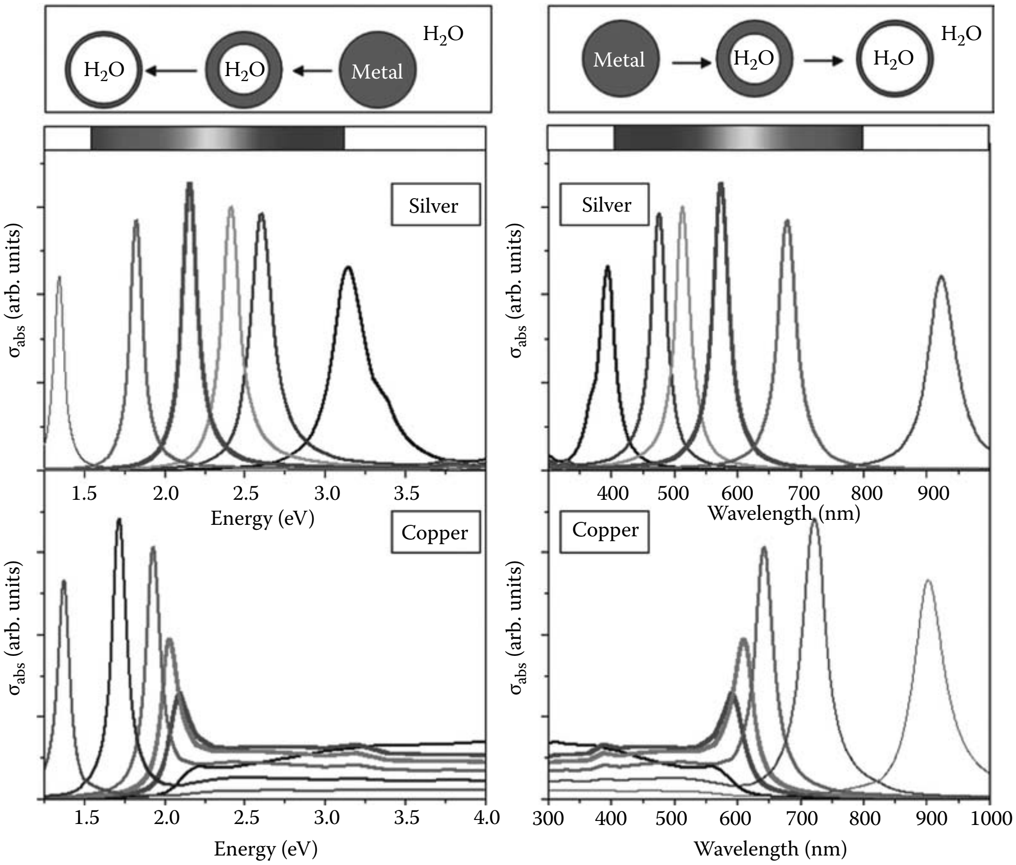

FIGURE 24.9 **(See color insert following page 20-14.)** Evolution of the absorption cross section of a nanoshell of silver (top) or copper (bottom) in the dipolar approximation versus energy (left) or versus wavelength (right) for various thicknesses of the shell. The core is filled with water and the external medium is also water ($n = 1.33$). The dielectric functions of copper and silver have been extracted from Palik (1985–1991). The correspondence between the energy in eV and the wavelength in nm is E (eV) = 1239.85/λ (nm). The total radius of the cluster is always 15 nm and the thickness takes the following values: $e = R - R_c = 5; 4; 3; 2; 1$ nm corresponding to ratios between the shell thickness and the total cluster radius: $(e/R) = 0.33; 0.27; 0.2; 0.13; 0.07$. The spectra in black correspond to the fully homogeneous cluster.

The three involved fields (incident, scattering, and internal fields) can then be expressed as follows:

$$\begin{cases} \vec{E}_i(\vec{r}) = \displaystyle\sum_{n=1}^{\infty}\sum_{m=-n}^{n}[p_{nm}\vec{M}_{nm} + q_{nm}\vec{N}_{nm}] \\ \vec{E}_{\text{sca}}(\vec{r}) = \displaystyle\sum_{n=1}^{\infty}\sum_{m=-n}^{n}[a_{nm}\vec{M}_{nm}^{+} + b_{nm}\vec{N}_{nm}^{+}] \\ \vec{E}_{\text{int}}(\vec{r}) = \displaystyle\sum_{n=1}^{\infty}\sum_{m=-n}^{n}[c_{nm}\vec{M}_{nm} + d_{nm}\vec{N}_{nm}] \end{cases}$$

Owing to the orthogonality properties of the vector spherical harmonics, the four boundary conditions expressing the tangential continuity of the electric and magnetic fields at the surface of the sphere lead to a one-to-one relationship between the six various expansion coefficients corresponding to a given (n, m)-index set. For a given index set, they result in four linear algebraic equations allowing the expansion coefficients of the internal and scattered fields to be determined as a function of those characterizing the incident field. A mere proportionality relationship is obtained first between p_{nm}, a_{nm}, and c_{nm}, and second between q_{nm}, b_{nm}, and d_{nm}, the coefficients depending on $\varepsilon(\omega)$, ε_m, R, $j_n(x)$, and $h_n^{+}(x)$ (and their derivatives) for $x = kR$ or $x = k_mR$. Finally, the integration of the flux of the Poynting vector over a sphere enclosing the particle allows one to obtain the energy dissipated per second by scattering and absorption and thus extinction, and consequently the cross sections. All these observables can be expressed in terms of the expansion coefficients of the three involved fields:

$$\begin{cases} \sigma_{\text{sca}} = \dfrac{4\pi}{k^2}\displaystyle\sum_{n,m}\left\{|a_{nm}|^2 + |b_{nm}|^2\right\}\dfrac{1}{E_o^2} \\ \sigma_{\text{abs}} = -\dfrac{4\pi}{k^2}\displaystyle\sum_{n,m}\text{Re}\left\{a_{nm}p_{nm}^{\star} + b_{nm}q_{nm}^{\star}\right\}\dfrac{1}{E_o^2} - \sigma_{\text{sca}} \\ \sigma_{\text{ext}} = \sigma_{\text{abs}} + \sigma_{\text{sca}} \end{cases}$$

The vector spherical harmonics are defined according to Appendix A in reference Lermé et al. (2008).

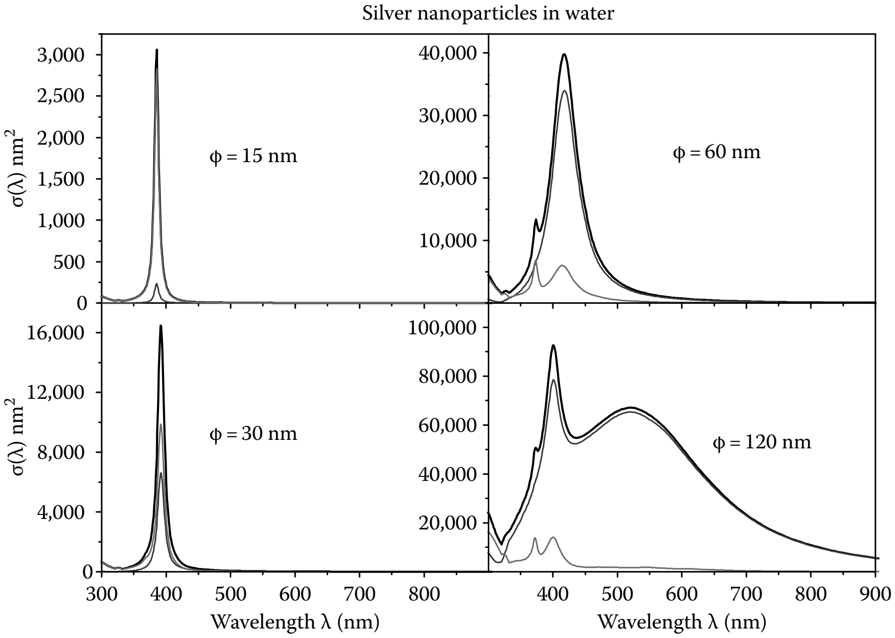

FIGURE 24.10 Extinction (bold black line), absorption (gray line), and scattering (thin black line) cross sections calculated for silver clusters of diameter ϕ in water (n = 1.33) in the frame of the Mie theory. The dielectric function of silver is extracted from Johnson and Christy (1972). For small sizes extinction is dominated by absorption (for ϕ = 15 nm absorption and extinction are identical), whereas scattering becomes predominant for larger sizes.

Figure 24.10 gives the results of the Mie theory for silver nanoparticles of various sizes embedded in water (n = 1.33). A red-shift of the resonance with increasing size, which is correlated to multipolar effects, is clearly evidenced. Such size effects are sometimes referred to as extrinsic size effects (Kreibig and Vollmer 1995). The resonance is very different from the one expected in the quasi-static approximation and scattering is seen to be fastly dominant in the optical response for larger sizes. Notice that the spectra for smaller sizes do not involve surface effects and are thinner than the ones obtained previously since the modification of the collision rate at the cluster surface has not been included here (Section 24.2.2.3). For larger sizes, a shoulder appears at about 375 nm, which corresponds to the quadrupolar mode. This quadrupolar mode only appears in the absorption cross section for a diameter of about 60 nm. Experimentally, it has been observed in silver nanoparticles embedded in silica glass (p. 289 from reference Kreibig and Vollmer (1995)), but it is easier to observe on single nanoparticles (Billaud et al. 2007).

The Mie theory can be generalized to spherical core/shell structures and extended to several interacting spherical nanoparticles excited by an incident plane wave. In this latter case, the field felt by the particle is then the sum of the applied field and the fields scattered by the other particles. The optical response can be auto-coherently resolved. It should be emphasized that in addition to the analytical approaches, numerous numerical methods exist to deal with the particles of any shape or size (Kahnert 2003) or with pairs of nanoparticles in interaction (Romero, 2006). This subject arouses a large interest but is out of the range of this chapter.

24.3 Experimental Results: State of the Art

This section gives a brief outline of the main results obtained in this field of investigation through some illustrations on size, shape, and multipolar effects observed in simple and noble metal clusters. But first, the most common methods for producing clusters and some techniques for performing their optical spectroscopy are described.

24.3.1 Nanoparticle Synthesis

Clusters and nanoparticles can be synthesized either by chemical or physical processes. The former have been known for a long time and allow producing clusters in solution, in matrices or deposited on various substrates. The latter have been developed for almost 30 years.

24.3.1.1 Chemical Methods

Chemical methods can be classified into two main routes: the formation of clusters in glasses or the colloidal synthesis. The former has been worked out from antiquity and consists in including metallic powders and some reducing agent in a melting glass, which is then fastly cooled down in order to obtain a homogeneous glass doped with metal in an oxidized or ionic state (Doremus et al. 1992, Kreibig and Vollmer 1995, Shiomi and Umehara 2000). Various thermal or irradiation treatments are further applied to induce metallic reduction and the growth of metallic nanoparticles embedded in a transparent glass. This technique is nevertheless limited to noble metals with a low concentration (10^{-6} to 10^{-3}), but this is the only one giving volumic nanostructured glasses. The solgel technique (Brinker and Scherer 1990) is an alternative to elaborate such composite materials but only in the form of thin films. It consists in transforming a solution of precursors (solid particles in suspension: the sol) into a polymerized gel containing metallic ions on various substrates (by dip-coating, spin-coating, or pulverization). After drying, the films are thermally treated (annealed under reducing atmosphere, UV, or laser irradiation) allowing the reduction in metallic nanoparticles. This technique enables to prepare noble and transition metal clusters (Nogami et al. 1995, De et al. 1996, Canut et al. 2007) of sizes between a few nanometers to a few tens of nanometers embedded in various matrices with concentration ranging from 10^{-3} to 10^{-1}.

The second route is the colloidal technique (Link and El-Sayed 1999), which has known a large development within the past decades for elaborating self-organized nanostructures. The standard way consists in reducing metallic salts in aqueous solution at intermediate temperatures (300–400 K) thanks to reducing agents. The addition of organic compounds adsorbing on cluster surfaces will stabilize and control the growth reaction. The final cluster diameter is controlled by the nature and the concentration of the reagents (organic compounds and reducing agents): it can vary from a few nanometers to a few hundreds of nanometers. By this way, bimetallic clusters can also be prepared in the form of alloyed or segregated structures (Papavassiliou 1976, Torigoe et al. 1993, Liz-Marzan and Philipse 1995, Link et al. 1999). The optical properties of colloids are investigated either in solution or once they have been deposited on a substrate. However, they can be also transferred into the gas phase, with an electrospray source for instance, and then analyzed through mass spectrometry. They can even be deposited after an accurate selection of their size (Rauschenbach et al. 2006, Böttger et al. 2007). More recently, new methods have emerged allowing a reduction of the size dispersion that is a prerequisite for cluster self-assembling. By those ways clusters are generally passivated with ligands or surface-active molecules having an hydrophilic head and an hydrophobic chain (surfactants). Depending on their shape, related to the ratio between the head size and chain length, these molecules may form direct or reverse dynamic micelles (Pileni 1989) that can be viewed as molds in which nanoparticles can be chemically synthesized. For instance, the reverse micelles form aqueous droplets in nonaqueous solution that serve as nanoreactors for the cluster growth (Lisiecki and Pileni 1995, Lisiecki et al. 1996). By adding anions that adsorb on clusters, the final cluster size may be accurately controlled. Moreover, the shape may be also monitored by taking advantage of a selective adsorption on the various faces of the cluster during the growth. The mean size of the so-produced clusters is of a few nanometers in diameter, and nanorods, nanodisks, or nanowires may be obtained as well. This process reduces considerably the size dispersion and is thus a suitable way to generate 2D or 3D superlattices (Schmitt et al. 1997, Courty et al. 2001) that are model systems for investigating the influence of the cluster interactions on their optical or magnetic properties.

24.3.1.2 Physical Methods

An alternative way to generate clusters is the physical route. Two main approaches are usually considered. The first starts with the production under vacuum of an atomic vapor (thermal heating, laser ablation, etc.) that is further deposited on a suited substrate (de Heer 1993, Perez et al. 1997, 2001). Taking advantage of their diffusion on the surface, atoms can meet or be trapped on defects so as to form nucleation seeds that further grow under the form of clusters by the subsequent capture of other mobile atoms. This gives rise to 2D assemblies of cluster islands with size and shape dispersions that are generally difficult to control accurately. The atoms can also be deposited simultaneously with a dielectric matrix in which they are trapped and isolated. A further thermal treatment may then promote diffusion in the matrix volume so as to induce their grouping and the formation of embedded clusters. In this way, thin nanostructured films can be obtained (3D arrangement), but the cluster size and concentration are hardly controlled independently. It must be stressed that such samples can also be obtained, before annealing, by implanting high-energy metallic ions in a preformed thin film. A second method allows an independent and easier control of the cluster size and volumic concentration. It consists in directly producing clusters in the gas phase and depositing them on a substrate simultaneously with an embedding matrix. The available sources generating cluster beams work generally on the following principle (de Heer 1993): a metallic vapor obtained for instance by heating a metal in an oven is mixed with an inert gas to induce the cluster nucleation and growth. Then, the mixture undergoes an expansion through a nozzle giving birth to a continuous cluster beam. Depending on the temperature and pressure conditions in the source, the size range may strongly vary from a source to another one (seeded supersonic beam sources [Larsen et al. 1974], Sattler sources [Sattler et al. 1980], etc.). Instead of a metallic vapor, a plasma can be produced by lasers (pulsed laser vaporization sources [Smalley 1983, Milani and de Heer 1990]), by an electric discharge (PACIS sources [Siekmann et al. 1991]), or by magnetron sputtering (Hahn and Averback 1990). Cluster growth takes place during the high cooling of the plasma with an inert gas and the so-formed clusters may be charged or neutral. The main advantage is that clusters of refractory metals or alloys may be produced. Moreover, the cluster size distribution can be varied by controlling the inert gas pressure in the source (Richtsmeier et al. 1985).

The co-deposition technique of preformed clusters by the so-called low-energy cluster beam deposition (LECBD) has proven to be one of the most efficient technique for elaborating nanostructured thin films for optical as well as for other investigations. Here we mainly focus on the optical studies of supported nanoparticles, but it must be emphasized that many experiments have been carried out on free cluster beams produced by sources such as those mentioned before. This will be outlined in the following.

24.3.2 Techniques of Optical Spectroscopy

Metallic clusters may be optically investigated either in the gas phase or in nanostructured films. In the former, several methods are available to probe their optical response such as photodissociation, multiphoton ionization (MPI) or photo-depletion spectroscopy (Kreibig and Vollmer 1995). They are of particular importance as they allow to get information about their electronic structure. The principle of such methods is the following: a cluster absorbs a given number of photons of known energy, and its desexcitation leads to the appearance of at least two reaction products (electrons, ions, or neutral fragments) that are detected and counted by mass spectrometry. The nature of these products and the wavelength dependence of their counting rates give insight into the electronic structure of the primary cluster. MPI is rather devoted to very small clusters (of a few atoms), whereas beam depletion spectroscopy can be extended to larger clusters. By using size-selected cluster ion beams and detecting the fragment-size distribution, it is also possible to deduce photo-absorption cross sections (Bréchignac et al. 1992, 1993).

Concerning embedded clusters (samples of nanostructured films or of colloids in solution), the general setup to investigate their optical properties is composed of a light source directed through the sample and detectors for the measurements of the incident, transmitted, reflected, and scattered light that may be analyzed with a grating monochromator (Fox 2001) (see Figure 24.11). The transmittivity T yields the extinction of light correlated to absorption and scattering losses.

For small clusters, the scattering being negligible with respect to absorption, the transmittivity directly reflects absorption (if the thin film reflectivity R can be neglected) through the Beer–Lambert law: $T = (I_t/I_o) = e^{-K_{\text{obs}}e}$, where $K_{\text{abs}} = (4\pi/\lambda)$ $\kappa(\omega)$, κ being the imaginary part of the effective optical index of the composite sample. This effective optical index may be estimated through effective medium theories, for instance the Maxwell–Garnett theory developed in the frame of the dipolar approximation for small embedded spheres and valid for low-volumic particle concentrations (<10%). Note that the particles can be considered almost isolated from each other if the volumic metal proportions are lower than a few percents. The optical response can then be compared to the expected response of a single cluster in a transparent medium. Of course, this requires an accurate characterization of the matrix in which the clusters are embedded. On the other hand, in the case of weakly absorbing thin films, in particular when the resonance is strongly damped as is observed in small clusters), the reflectivity can no longer be neglected and multilayer effects (Fabry–Pérot fringes) may blur the intrinsic response of the embedded clusters. To get rid of this spurious effect, a simple trick consists in using p-polarized light under Brewster incidence when measuring transmission (Cottancin et al. 2003).

All these measurements are carried out on cluster assemblies and the optical response corresponds to the average response of all the probed clusters. The unavoidable geometry fluctuations in the samples make delicate a precise determination of their specific properties. To overcome this difficulty the clusters may be size selected (Issendorff and Palmer 1999, Alayan et al. 2004) before co-deposition, but shape and environment fluctuations will persist. Another approach developed within the past 10 years aims at investigating a single deposited or embedded metallic nanoparticle. It requires highly sensitive spectroscopic methods since the optical density of the sample is considerably lower than in conventional nanostrutured films. Such investigations may be achieved through near-field or far-field microscopy. Indeed, when a particle is excited by an incident field, the diffracted field gives rise to evanescent waves in the vicinity of the particle and propagating waves far from the particle. In near-field microscopy an optical fiber tip is used as a probe of the field near the particle. By scanning the position of the tip on a sample, a map of the surface covered with deposited nanoparticles may be recorded, but the resolution is still limited by the tip size. In this way, gold nanoparticles of 20 nm in diameter have been detected (Klar et al. 1998). However, the interaction between the tip and the nanoparticles prevents a straightforward deduction of the

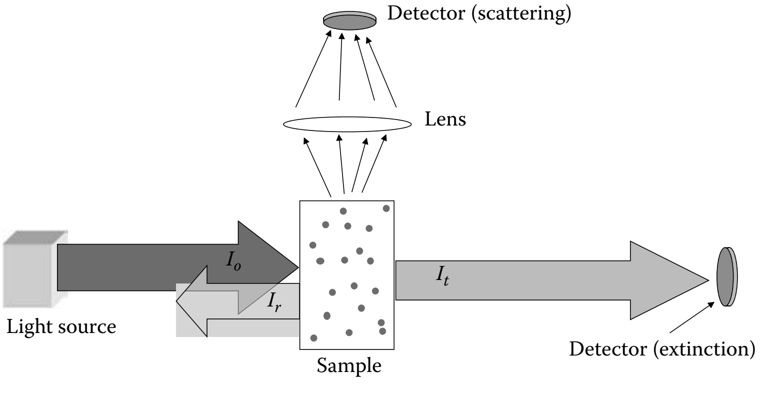

FIGURE 24.11 Schematic experimental setup for conventional scattering and extinction spectroscopy. Light sources and detectors may be combined with monochromators.

particle intrinsic properties. Far-field is more convenient as the obtained response is easier to compare with theory. To “bypass” the Rayleigh criterion, the basic idea is to isolate nanoparticles from each other by depositing them with very low concentration (less than one particle per square micrometer) to be sure that only one particle will be shined by a highly focused light beam. The most widespread technique is the dark-field confocal microscopy allowing the measurement of single-nanoparticle scattering and which can be viewed as an improvement of the ultramicroscope of Zsigmondy. The high sensitivity is obtained by using high numerical aperture refractive objectives with a mask (Dijk et al. 2006) or reflective objectives, both generating a shadow cone in which the scattered light can be collected with a negligible background signal. Nevertheless, as the scattering varies as V^2, such investigations remain restricted to particles larger than 20 nm in diameter (Tamaru et al. 2002, Raschke et al. 2003). On the other hand, interferometric methods have been also developed to detect smaller nanoparticles (Lindfors et al. 2004, Berciaud et al. 2005).

Besides scattering experiments, it is also possible to detect the overall extinction of light by a nanoparticle in a transmission measurement configuration, just as in conventional spectrophotometers. Using tightly focused light beams in a confocal geometry, an increased sensitivity may be reached thanks to a spatial modulation of the sample and a lock-in detection (Arbouet et al. 2004, Muskens et al. 2006, Billaud et al. 2007). The main advantage of this technique is that it enables the absolute measurement of the extinction cross-section if the incident field distribution in the focal plane is accurately characterized. Finally, it should be noticed that scattering or extinction measurements on single nanoparticles may be correlated to their electron microscopy images by using substrates compatible with optical measurements and electron microscopy (Tamaru et al. 2002, Billaud et al. 2008).

24.3.3 Experimental Results

This last section draws up some of the major experimental results relative to size and shape effects on the optical properties of metal clusters, emphasizing the role of the SPR as a probe of the internal structure of the cluster material. Because of the wealth of publications in this field for 20 years, the relevant state of the art will not be exhaustive. A more complete bibliography may be found in the appendix of Kreibig and Vollmer (1995) as well as in the references list cited below.

24.3.3.1 Size and Shape Effects in the Optical Response of Metal Clusters

24.3.3.1.1 Alkali Clusters

Let us first examine the experimental results concerning the SPR in alkali clusters. As mentioned before, alkalis are the simplest metals and their clusters represent ideal model systems that were widely studied during the 1990s. Because of their high sensitivity toward oxidization, most of the corresponding experiments were usually performed on cluster beams in the gas phase through photo-evaporation or depletion spectroscopy (Bréchignac et al. 1992, 1993, Borggreen et al. 1993, Pedersen et al. 1993). In all cases (lithium, sodium, potassium), the optical spectra are dominated from very small sizes (~8 atoms) by a broad absorption peak (the SPR) qualitatively similar to the classical predictions of the Drude model. Nevertheless, a red-shift of the resonance with decreasing size, correlated to the increasing spillout influence, is patent (see Figure 24.12a) whatever the element. The signature of the spillout effect has been also observed in the measurements of the static polarizability of sodium and lithium clusters (Knight et al. 1985, Bénichou et al. 1999). Indeed we have already underlined that because of the electronic spillout, the sphere containing the conduction electrons (of effective radius R_e) is larger than the geometrical volume enclosing the ionic cores (sphere of radius R_C). The main consequence is a correlated increase of the static cluster polarizability since it is proportional to the effective volume $(4/3)\pi R_e^3$. The calculations in the frame of the time-dependent local-density-approximation formalism (TDLDA) (within the simple jellium model) that have been shown in Figure 24.7 reproduce qualitatively the experimental results, in particular the red-shift of the resonance with decreasing size. Nevertheless, it should be emphasized that these calculations slightly overestimate the absolute value of the SPR in sodium clusters (the SPR occurs generally around 2.5–2.8 eV for sizes smaller than 100 atoms per cluster, whereas the calculations give values around 3 eV). In lithium the disagreement is larger. These discrepancies show that various ingredients, not included in the standard jellium model, have a noticeable influence. For instance the ionic-core polarization, which is disregarded in this framework, may induce a shift of the resonance toward smaller energies. This shift can be viewed classically through Equation 24.8 in which ε_1^{IB} would be chosen different from zero. Moreover, the profile of the effective potential confining the conduction s-electrons and hence also the electron density are known to be softened when electron–ion interaction pseudo-potentials (non-coulombic) are taken into consideration (Rubio et al. 1993, Lermé et al. 1995). The surface softness increases the electron spillout relative to the standard jellium model in which the electron–ion interaction is assumed to be coulombic. At last, the nonlocal character of the electron–ion pseudo-potential results in an increase of the effective conduction electron mass (for instance, in lithium the effective mass is $m_{\text{eff}} = 1.4\, m_e$). Both these effects lead to a decrease of the bulk plasma angular frequency ω_P, and hence of ω_{SPR}. On the other hand, the Drude parametrization of the dielectric function also leads to a larger value of the SPR frequency (~3.5 eV) than that deduced from experimental measurements on thin alkali films (~3.1 eV), showing similarly the crudeness of the model for the bulk phase. This shows that the conduction electrons do not really constitute a free electron gas in the bulk. If the details of the cluster electronic structure are obviously of great importance to obtain an accurate quantitative description of the optical properties of alkali clusters, the possible deviation of their shape from the ideal case of a sphere relative to the ideal case of a sphere deserves to be considered with care. The one-peak resonance observed for Na_{21}^+ and Na_{41}^+ (see Figure 24.12b) indicates that these clusters adopt an almost spherical shape since these magic sizes correspond to exact electronic closures in

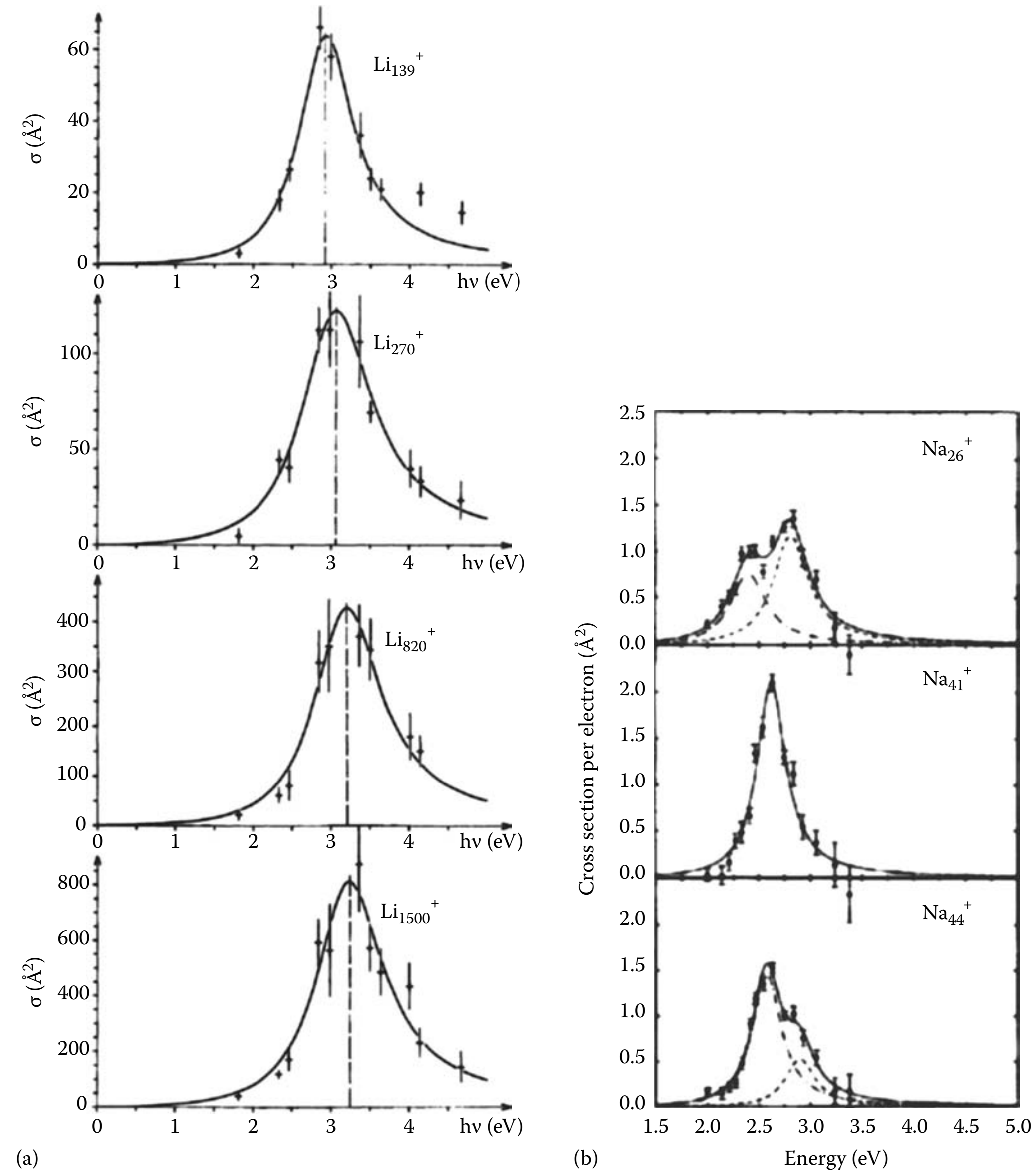

FIGURE 24.12 (a) Photoabsorption cross sections for large lithium clusters ions Li_n^+ versus the photon energy. (From Bréchignac, C. et al., *Phys. Rev. Lett.*, 70, 2036, 1993. With permission.) (b) Photoabsorption cross sections per electron versus the photon energy, for various sizes of sodium clusters. (From Pedersen et al., *Z. Phys. D*, 26, 281, 1993. With permission.) In both figures, the lines are Lorentzian fits of the data.

the spherical jellium model. For non-magic sizes, the resonance is asymmetric, reflecting the existence of at least two resonance modes corresponding to plasmon excitations along the different principal axes of deformation of the cluster (see Figure 24.12b). One can thus say that the profile of resonance is an efficient probe of the cluster shape. Nevertheless, for very small sizes the multi-peak absorption spectra also reflect the electronic-level quantization.

At last, potassium and sodium clusters have been also investigated once deposited under ultrahigh vacuum (UHV) and exposed to various gases (Iline et al. 1999). Since the gases modify their surface tension after being adsorbed, clusters may experience a change in shape that can be detected through variations in absorption spectra. Nevertheless, the influence of the gas is not so easy to interpret as the size was not well controlled.

24.3.3.1.2 Noble Metal Clusters

In noble metal clusters, as the valence d-band and the conduction s–p band are energetically close to each other, the SPR emerges more or less easily depending on the element and is strongly influenced by the d-electrons. The simplest metal is silver because the threshold for interband transitions in the bulk is quite high (around 4 eV). The SPR emerges for very small sizes, in the range of less than 20 atoms, and it was measured both for ionic clusters (Tiggesbäumker et al. 1993) and for size-selected neutral clusters deposited in a rare gas matrix (Fedrigo et al. 1993). Nevertheless, the core electrons (d electrons mainly) play an important role. The screening they induce is responsible for a global red-shift of the SPR frequency as compared to the frequency expected from the standard jellium model (Pollack et al. 1991, de Heer 1993, Liebsch 1993a,b, Rubio and Serra 1993, Kreibig and Vollmer 1995, Lermé et al. 1998a,b, Palpant et al. 1998). For gold, the threshold for interband transitions is much lower (of the order of 2 eV). This results in a stronger red-shift of the SPR frequency and, in addition, a strong damping and broadening of the resonance band, especially for small particles (Kreibig and Vollmer 1995, Alvarez et al. 1997, Palpant et al. 1998, Gaudry et al. 2001). For copper, the interband threshold is similar to that of gold, of the order of 2 eV. The SPR occurs in the same frequency range as gold, but experiments show that the resonance band is much more damped and even disappears at low sizes for particles of typically 2–3 nm in diameter (few hundred of atoms) depending of the observation conditions (Lisiecki and Pileni 1995, Celep et al. 2004), e.g., on the dielectric constant of the matrix in which the particles are embedded. This specific behavior of copper clusters will be explained below.

Figure 24.13 illustrates the size evolution of the SPR in noble metal clusters in a size range below 10 nm in diameter for which quantum size effects are expected. The upper part of the figure displays experimental results on the size evolution of the SPR on the samples of noble metal clusters (copper, silver, and gold) embedded in alumina matrix (Lermé et al. 1998b, Palpant et al. 1998, Lermé 2000, Celep et al. 2004, Cottancin et al. 2006) at low concentrations (the matrix and the cluster size distributions have been carefully characterized). As the samples were elaborated with the same protocol and in the same matrix, a comparative study can be achieved. In gold clusters, the SPR band occurs around 2.3–2.4 eV in the same spectral region than the threshold of the interband transitions (2 eV), which continue, as for them, to induce adsorption in the UV range. Moreover, a blue shift and a damping of the resonance are observed when the cluster size decreases. In silver clusters, the SPR band peaks at about 2.9–3.0 eV, well below 4 eV the threshold of the interband transitions. Only a very small blue-shift of the resonance is measured when the cluster size decreases. For copper, the SPR band emerges hardly at about 2.15 eV from the steep increase of the interband transitions. Actually in copper the interband threshold is close to 2 eV (as in gold), but the corresponding absorption increases much more rapidly above 2 eV. The damping and the broadening of the SPR band, with decreasing sizes, are the most important features of the experimental observations, rather than a clear blue-shift as in gold. At very small sizes, the SPR

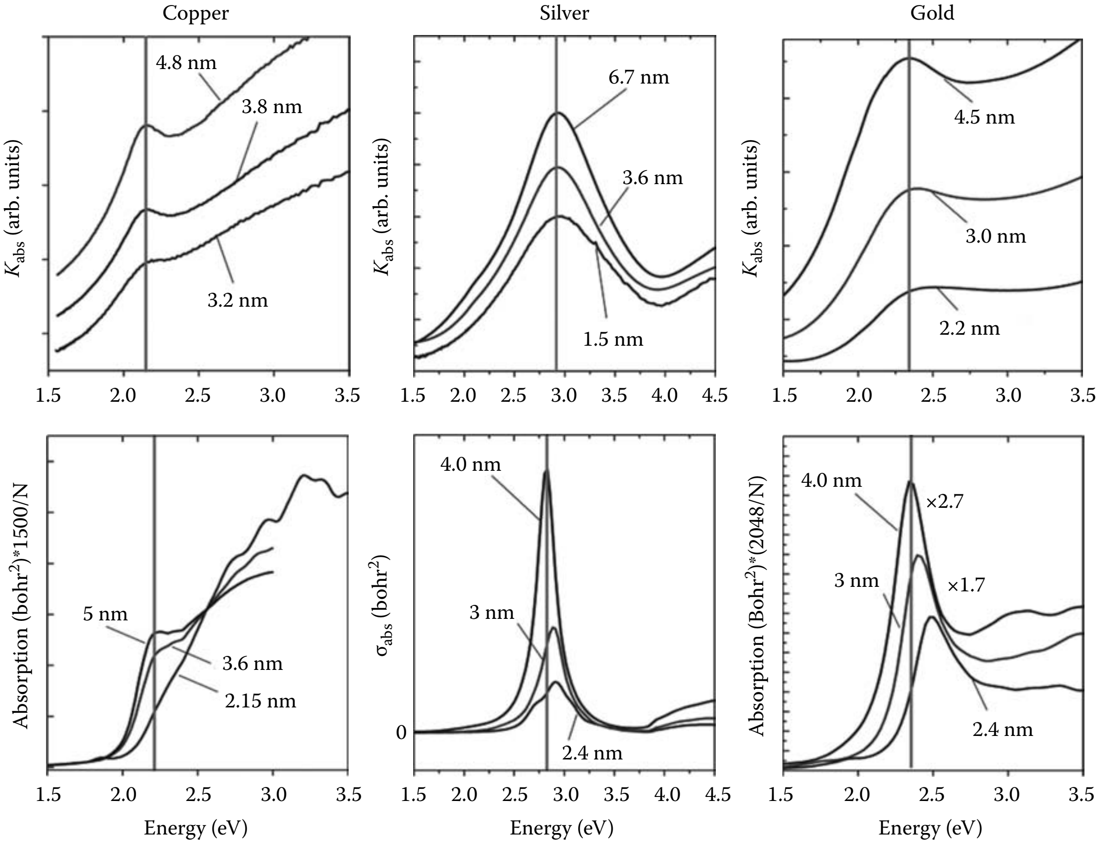

FIGURE 24.13 Size evolution of the experimental (top) and theoretical (bottom) absorption spectra obtained in the semi-quantal model for noble metal clusters of copper, silver and gold embedded in alumina (Cottancin et al. 2006). The volumic concentrations of metallic clusters embedded in the transparent matrix are in all cases rather low (<5%–6%) allowing to consider these systems as isolated clusters randomly dispersed. (From Cottancin, E. et al., *Theor. Chem. Acc.*, 116, 514, 2006.)

band tends to vanish in the steep rising edge of the interband transitions, but in the present experiments, it was not possible to produce clusters as small as 2 nm.

In the present size range (2 nm < ϕ < 5 nm), the Mie theory is not suited to explain the size evolution of the SPR, even if a size dependence of the scattering rate is assumed (see Equation 24.11). Indeed, the size evolution obtained from such calculations (Figure 24.5) shows a red-shift of the resonance in gold and copper in contradiction with experimental results. For a better interpretation of the experimental results, and as long as large particles are involved, a structureless jellium-type model approach seems quite appropriate and reasonable for investigating mean-size trends over broad-size domains. The calculations are based on the TDLDA within a jellium model including phenomenologically the absorption and screening properties of both the ionic-core background and the surrounding matrix through their dielectric functions $\varepsilon^{IB}(\omega)$ and ε_m (Lermé 2000). A first ingredient added in the model, well established in surface physics and correlated to the spatial localization of d-electrons wavefunctions, is a skin of a vanishing polarizability of thickness d of a few angstroms in which the polarization of the d-core electrons is ineffective ($\varepsilon^{IB} = 1$ in this shell). This skin of ineffective polarizability was firstly introduced by Liebsch for silver surfaces (Liebsch 1993a) and by Kresin for clusters (Kresin 1995). Moreover, a vacuum ring surrounding the particle, of thickness d_m, is introduced in the model in order to mimic the local matrix porosity estimated experimentally (see Figure 24.14).

Figure 24.13 shows the comparison between experiment and this semi-quantal theory. The theoretical calculations well reproduce the experimental trends when the cluster size decreases, namely the strong blue-shift of the SPR for gold, the very small blue-shift for silver, and the gradual disappearance of the SPR in the rising edge of the interband transitions for copper. This blue-shift is mainly due to the skin of ineffective screening by the d-electrons, which results in an all the more important blue-shift that the sizes are small. This blue-shift competes with the red-shift associated with the spillout of the s-electrons, that is a pure quantum effect. In gold, the blue-shift is predominant as compared to the red-shift. In silver, both are close to be compensated, inducing a rough quenching of the size effects. This different behavior of gold and silver clusters is easy to understand because the real part of ε^{IB} (which quantifies the screening) is much smaller in silver than in gold (of the order of 4 and 11, respectively, in the spectral range of the SPR [Cottancin et al. 2006]). The difference between gold and copper may be understood with regard to their respective dielectric functions. The main difference lies in the steeper increase of the imaginary part of ε^{IB} for copper, just above the interband threshold, leading to a stronger damping of the resonance. Actually, the vanishing of the resonance in copper is a signature of a blue-shift of the SPR, resulting in an increasing coupling with IB transitions.

Finally, the semi-quantal calculations appear to be the best compromise between completely ab initio methods limited to relatively small sizes and completely classical calculations that disregard the quantum size effects. These semi-quantal calculations, with a very limited number of semiempirical parameters, are able to predict the optical properties of the three noble metals in a size range from 1.5 to 10 nm in diameter. Above 10 nm, the quantum size effects are negligible and the Mie theory is well adapted.

As for shape effects, experiments have been performed on small silver cluster beams showing similar results as in alkalis (Tiggesbaümker et al. 1992, Tiggesbäumker et al. 1993). However the shape influence on the optical response of noble metal clusters and nanoparticles (essentially gold and silver) has been mostly investigated within the past 10 years on single nanoparticles. Indeed, the tremendous advances in the synthesis of various nano-objects (such as rods, prisms, chains, core/shell structures, pairs of nanoparticles, etc.) opened up a new field of investigations on their optical properties, in a size range where the multipolar effects are of major importance. For instance, silver nanoparticles of various shapes have been studied using dark-field optical microscopy and spectroscopy, in correlation with their electron microscopy images (Mock et al.

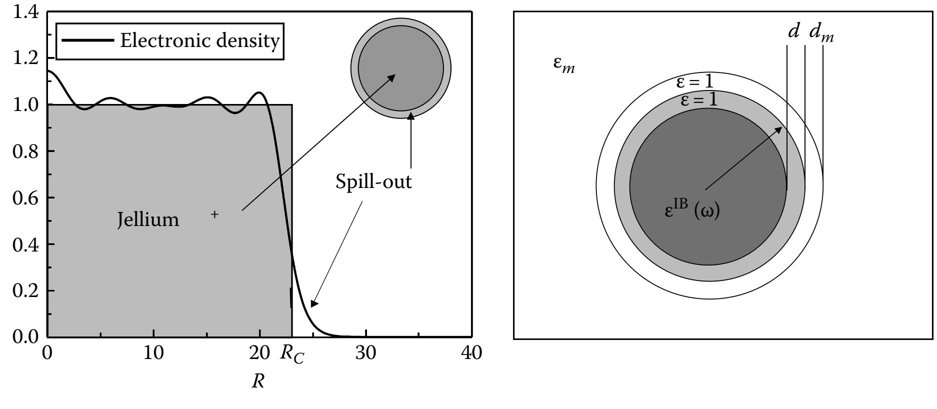

FIGURE 24.14 (Left) *Jellium* model: ionic (gray) and electronic (black line) densities. (Right) Semi-quantal model: cluster of radius $R_C = r_S N_e^{1/3}$ in matrix of dielectric function ε_m (d et d_m denote respectively the thicknesses of the shell of reduced polarizability and the vacuum ring surrounding the particle).

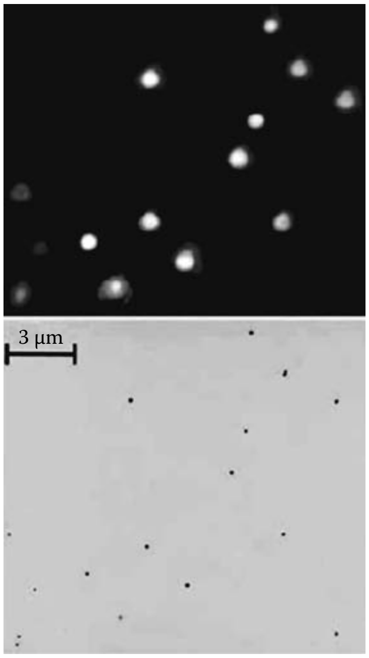

FIGURE 24.15 **(See color insert following page 20-14.)** (Top) Color image of a typical sample of silver nanoparticles as viewed under the dark-field microscope. The brightness of the particles increases from blue to red due to both the intrinsic optical scattering cross section and the spectral output of the light source (the red particle is overexposed). This image is correlated to its electron microscopy image (Bottom). (Reprinted from Mock, J.J. et al., *J. Chem. Phys.*, 116, 6755, 2002. With permission.)

2002). The spectral position of the resonance is so much shape dependent that spheres appear blue, pentagons green, and triangles mostly red (see Figure 24.15). A significant dispersion of the SPR peak position for triangular particles is observed. This phenomenon is probably due to the high sensitivity of the resonance to the height, the edge length, and the apex shapes (sharp or truncated) of the triangles as has been evidenced through numerical calculations based on the discrete dipole approximation (Sherry et al. 2006).

The main conclusion concerning all the numerous studies for which the state of the art is out of the topic of this chapter is that the SPR may be tailored by varying the shape or the structure of the nano-objects. Consequently, the attachment of specific molecules (such as proteins or antibodies) on the nanoparticle surface can be detected by an induced shift of the SPR (biosensors) (McFarland and Duyne 2003).

24.3.3.2 The SPR as a Probe of the Internal Structure of Bimetallic Clusters

Bimetallic particles are of double interest since they disclose size- and shape-dependent properties and mirror the original features of bulk alloys (a complete bibliography about nanoalloys can be found in Ferrando et al. (2008)). Depending on the thermodynamic properties of both components, mixed particles can form either homogeneous alloys or segregated core/shell structures. An insight into the bulk properties of both constituent materials (Wigner–Seitz radius (r_S), surface energies) and their corresponding binary phase diagrams brings indications about the possible structure of the composite clusters. In general, one expects the metal component with the lower surface energy to accumulate preferentially at the surface. However, in the case of systems in solution like colloids, one has to be careful with such simple rules, because the kinetics of reduction as well as the surface-adsorbed ions and the surrounding medium also come into play. Thermodynamical equilibrium conditions are not necessary reached and such particle architectures may be only metastable in this respect.

The dielectric function of mixed clusters being governed by its overall electronic structure, the optical response of bimetallic systems will be modified by varying the relative proportion of both constituents (Kelly et al. 2003, Hubenthal et al. 2005). For instance, in the case of gold–silver nanoparticles, by far the most-studied binary system (Morriss and Collins 1964, Papavassiliou 1976, Teo et al. 1987, Mulvaney et al. 1993, Liz-Marzan and Philipse 1995, Link et al. 1999), the SPR falls between those of gold and silver for alloyed systems. On the contrary, in core/shell geometries (that can be essentially obtained by chemical ways), the optical response displays two resonance peaks for sufficiently large particles.

As an illustration, Figure 24.16 shows the evolution of the optical response with varying compositions of aqueous solutions of alloyed gold–silver nanoparticles of diameters ranging from 15 to 50 nm (Russier-Antoine et al. 2008). The spectra exhibit a single SPR whose wavelength is found to move regularly, with increasing gold content, from the silver SPR to the gold one. This observation thus rules out the possibility of a core/shell structure for which two resonances would be expected.

However, the experimental spectra cannot be correctly reproduced by the Mie theory if the dielectric function of the alloy is taken as the volume-weighted average of the pure material dielectric functions. Such an hypothesis is physically relevant if the particle can be considered as built by the stacking of pure nanodomains that individually retain the electronic structure (dielectric function) of their corresponding bulk phase. Its failure indicates that the effective dielectric function is closely related to an alloying at a much lower scale (atomic level). It cannot be simply inferred from the dielectric function of each component but must be either obtained from available experimental data on similar bulk alloys or calculated from solid-state theory assuming an alloying at an atomic scale. In this system, measurements on annealed alloys (Nilsson 1970) have been performed and the corresponding dielectric functions exhibit a regular shift of the interband transition threshold with the gold composition that is not reproduced in the average dielectric function evoked just before. If this IB threshold is taken into account, i.e., if the dielectric function reproduces correctly the one measured experimentally, the optical spectra are in agreement with the theory, confirming thus that the nanoparticles adopt an alloyed structure at an atomic level. Similar works on smaller bimetallic gold–silver clusters produced by laser vaporization and embedded in alumina had lead to the

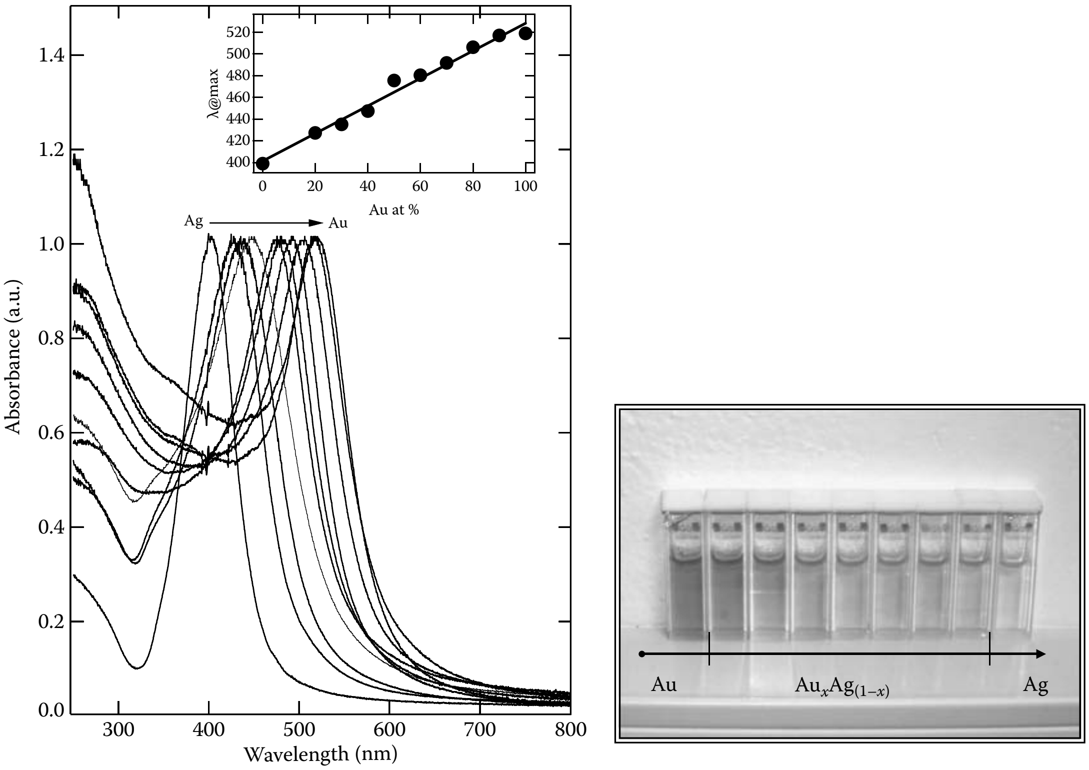

FIGURE 24.16 **(See color insert following page 20-14.)** (Left) Normalized UV–visible spectra of Au–Ag alloy nanoparticles with varying composition. (Insert) Location of the SPR maximum as a function of the gold content. (Right) Corresponding solutions whose colors vary from the red (pure gold nanoparticles) to the yellow (pure silver nanoparticle). (From Russier-Antoine, I. et al., *Phys. Rev. B*, 78, 35436, 2008. With permission.)

same conclusions about the internal structure of clusters, near from the one of annealed alloys (Gaudry 2001). Such a homogeneous alloyed structure is the likely one (for clusters produced by physical ways) in view of both their relative surface energies and Wigner–Seitz radii and regarding the bulk phase diagram of the Au–Ag system in which homogeneous solutions are thermodynamically possible for any relative composition (Massalski et al. 1986).

In the same way, the internal structure of mixed Ni–Ag and Pt–Ag clusters of a few nanometers in diameter, produced by laser vaporization and embedded in an alumina matrix (Cottancin et al. 2003), has been investigated. Concerning the bulk properties of these materials, the phase diagram of Ni–Ag shows that Ni and Ag are immiscible over the entire composition range (Massalski et al. 1986). One can suppose that the silver atoms will move onto the cluster surface for both systems since the surface energy of silver is the lowest one and that the resulting structure will be a core/shell one. As for the Pt–Ag phase diagram, several alloys are present (Massalski et al. 1986), so the assumption of a core/shell structure has to be taken with caution. Moreover, the atomic Wigner–Seitz radii of Ag and Ni are noticeably different (r_s(Ag) = 3.02 a.u. and r_s(Ni) = 2.6 a.u.) and a pronounced lattice mismatch is expected, whereas they are closer for Pt and Ag (r_s(Pt) = 2.9 a.u.).

The experimental spectra of both systems are displayed in Figure 24.17. In the case of the Ni–Ag system, the optical properties are found to be intermediate between those of pure silver and pure nickel clusters, with an SPR enlarged and blueshifted compared to pure silver clusters. This feature can be understood within the classical Mie theory assuming a segregation between both metals in a core/shell geometry and including the reduction of the mean free path in the silver shell. Moreover, low energy ion scattering (LEIS) measurements that consist in probing the cluster surface provide the evidence that the surface of the clusters is mainly covered with silver atoms. Such results permit to conclude that the core/shell geometry with silver on the surface is the most likely structure and that the dielectric function of the bimetallic system remains closely connected with the ones of its constituents. As for Pt–Ag clusters, LEIS measurements lead also to the conclusion that the cluster surface is mainly composed of silver; nevertheless, the absorption spectra do not exhibit a marked SPR, in contradiction with the calculated spectra for which a clear resonance is expected as in Ni–Ag clusters. Calculations assuming a dielectric function equal to the volumic-weighted average of the constituent ones predict also a large resonance band. Those results suggest that the Pt–Ag particles do not form a segregated core/shell structure, but rather an alloyed core with a surface enriched with silver atoms. Monte Carlo simulations (Calvo et al. 2008) performed on these systems show the same trends as from small sizes. In particular at zero temperature, silver and platinum seem to be segregated, whereas an increase in the temperature

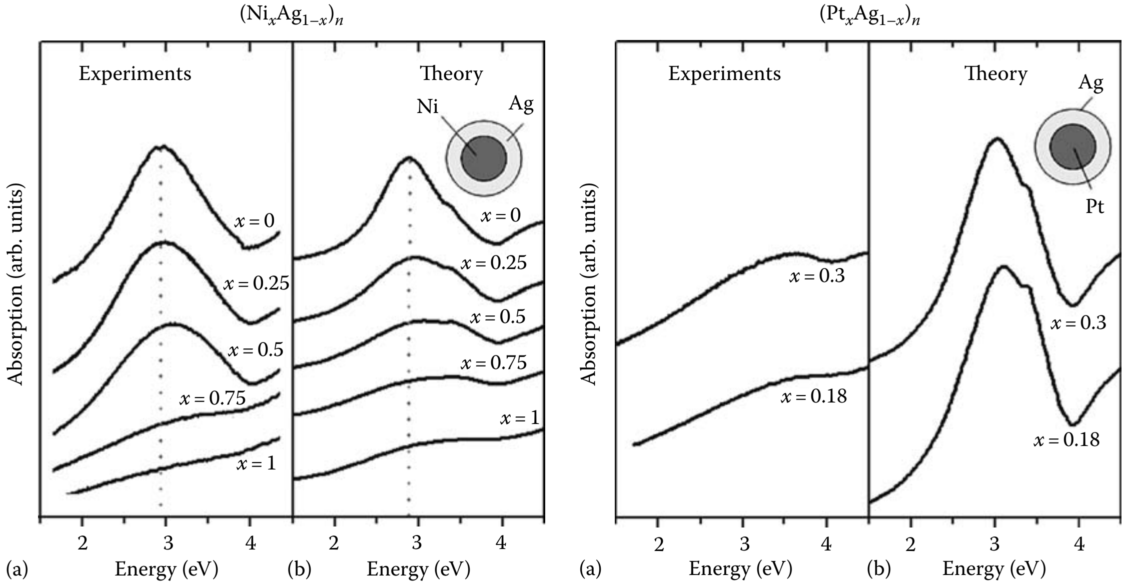

FIGURE 24.17 (Left) (a) Evolution of the optical absorption spectra of Ni–Ag clusters (optical diameter of about 2.6–2.7 nm) with the nickel proportion. (b) Theoretical absorption K_{abs} within the classical core/shell model when the reduction of the mean free path in the silver shell is taken into account. (Right) (a) Evolution of the optical absorption spectra of Pt–Ag clusters (optical diameter of about 4 nm) with the platinum proportion. (b) Theoretical absorption K_{abs} within the classical core/shell model.

induces a better alloying with the surface remaining rich in silver. Finally, one can conclude that the lack of the SPR in experimental spectra shows in this case that the dielectric function of the bimetallic system is not simply correlated to those of its constituents as it is the case for Ni–Ag.

24.4 Conclusion and Outlooks

This chapter intended to shed light on the main ingredients involved in the linear optical properties of metallic clusters and nanoparticles such as size, shape, environment, and the chemical structure of particles. The optical studies of nanoparticles are naturally a wide-ranging field including dynamical processes or nonlinear optics that have not been developed in this chapter.

It has been shown that the colors of samples containing clusters or nanoparticles originate from the strong absorption and scattering (SPR) of the metal nanoparticles when the frequency of the exciting electromagnetic field is coherent with the collective conduction electron motion.

In small particles ($\phi < 20$–30 nm), the optical response can be in a first approximation described in the frame of the dipolar approximation. In addition, quantum size effects have to be taken into account for cluster sizes below 5–10 nm in diameter, involving a red-shift of the SPR in alkali clusters because of the increasing effect of the electronic spillout. In noble metals, the polarization of the core electrons (*d* electrons mainly) and their ineffective influence near the cluster surface result in a blue-shift that competes with the red-shift associated with the spillout effect. The size evolution of the SPR depends then on the element (Cu, Ag, or Au) constituting the clusters.

For larger sizes ($\phi > 20$–30 nm), the Mie theory accounts well for the optical response of spherical nanoparticles in a homogeneous medium. When the size increases, higher resonance modes become dominant and the SPR is red-shifted and spectrally broadened.

The synthesis of new particles of various sizes and shapes in the past decades and the development of highly sensitive methods for probing single nano-objects have opened a new field of investigations showing that the SPR may be tailored by tuning the shape or the composition of the metal particles. The understanding of their behavior with the help of computational methods and experiments on single particles gives high hopes for potential applications. For instance, the high sensitivity of the SPR toward the local neighborhood of the particle or the chemical nature of its surface can be used for making biosensors, giving information about surface changes of the particles, as for instance the red-shift observed in silver nanoparticles after several molecules have adsorbed (McFarland and Duyne 2003). Moreover the large local field enhancement that occurs near the particle surface is being used in Surface-Enhanced Raman Spectroscopy (Campion and Kambhampati 1998) or Plasmon Enhanced Fluorescence (Lee et al. 2004). The strong absorption of metal nanoparticles can even be exploited in plasmonic photothermal therapy (Dickerson et al. 2008, El-Sayed et al. 2006).

Acknowledgments

The authors are grateful to F. Calvo for the reading of the chapter and for his fruitful remarks.

References

Alayan, R., L. Arnaud, M. Broyer et al. (2004), Application of a static quadrupole deviator to the deposition of size-selected cluster ions from a laser vaporization source, *Rev. Sci. Instrum.*, **75**, 2461.

Alvarez, M. M., J. T. Khoury, T. G. Schaaff et al. (1997), Optical absorption spectra of nanocrystal gold molecules, *J. Phys. Chem. B*, **101**, 3706.

Arbouet, A., D. Christofilos, N. Del Fatti et al. (2004), Direct measurement of the single-metal-cluster optical absorption, *Phys. Rev. Lett.*, **93**, 127401.

Ashcroft,N. W. and N. D. Mermin (1976), *Solid State Physics*, International Edition, Saunders College, Philadelphia, PA.

Barber, D. J. and I. C. Freestone (1990), An investigation of the origin of the colour of the Licurgus cup by analytical Transmission Electron Microscopy, *Archaeometry*, **32**, 33–45.

Bénichou, E., R. Antoine, D. Rayane et al. (1999), Measurement of static electric dipole polarizabilities of lithium clusters: Consistency with measured dynamic polarizabilities, *Phys. Rev. A*, **59**, R1–R4.

Berciaud, S., L. Cognet and B. Lounis (2005), Photothermal absorption spectroscopy of individual semiconductor nanocrystals, *Nanoletters*, **5**, 2160.

Berthier, S. (1993), *Optique des Milieux Composites*, Polytechnica, Paris, France.

Billaud, P., J.-R. Huntzinger, E. Cottancin et al. (2007), Optical extinction spectroscopy of single silver nanoparticles, *Eur. Phys. J. D*, **43**, 271–274.

Billaud, P., S. Marhaba, E. Cottancin et al. (2008), Correlation between the extinction spectrum of a single metal nanoparticle ant its electron microscopy image, *J. Phys. Chem. C*, **112**, 978–982.

Blanc, J., V. Bonacic-Koutecky, M. Broyer et al. (1992), Evolution of the electronic structure of lithium clusters between four and eight atoms, *J. Chem. Phys.*, **96**, 1793.

Bohren, C. F. and D. P. Huffman (1983), *Absorption and Scattering of Light by Small Particles*, Wiley, New York.

Bonacic-Koutecky, V., P. Fantucci, and J. Koutecky (1991), Quantum chemistry of small clusters of elements of groups Ia, Ib, and IIa: Fundamental concepts, predictions, and interpretation of experiments, *Chem. Rev.*, **91**, 1035–1108.

Bonacic-Koutecky, V., V. Veyret, and R. Miltric (2001), Ab initio study of the absorption spectra of Ag_n (n = 5–8) clusters, *J. Chem. Phys.*, **115**, 10450.

Borggreen, J., P. Chowdhury, N. Kebaïli et al. (1993), Plasma excitations in charged sodium clusters, *Phys. Rev. B*, **48**, 17507.

Böttger, P. H. M., Z. Bi, D. Adolph et al. (2007), Electrospraying of colloidal nanoparticles for seeding of nanostructure growth, *Nanotechnology*, **18**, 105304.

Brack, M. (1993), The physics of simple metal clusters: Self-consistent jellium model and semiclassical approaches, *Rev. Mod. Phys.*, **65**, 677–731.

Bréchignac, C., P. Cahuzac, N. Kebaïli, J. Leygnier, and A. Sarfati (1992), Collective resonance in large free potassium cluster ions, *Phys. Rev. Lett.*, **68**, 3916.

Bréchignac, C., P. Cahuzac, J. Leygnier, and A. Sarfati (1993), Optical response of large lithium clusters: Evolution toward the bulk, *Phys. Rev. Lett.*, **70**, 2036–2039.

Brinker, C. J. and G. W. Scherer (1990), *Sol-Gel Science*, Academic Press, San Diego, CA.

Calvo, F., E. Cottancin, and M. Broyer (2008), Segregation, core alloying, and shape transitions in bimetallic nanoclusters: Monte Carlo simulations, *Phys. Rev. B*, **77**, 121406(R).

Campion, A. and P. Kambhampati (1998), Surface-enhanced Raman scattering, *Chem. Soc. Rev.*, **27**, 241–250.

Canut, B., M. G. Blanchin, V. Teodorescu, and A. Traverse (2007), Structure of Ni/SiO_2 films prepared by sol-gel dip coating, *J. Non Cryst. Solids*, **353**, 2646–2653.

Celep, G., E. Cottancin, J. Lermé et al. (2004), Optical properties of copper clusters embedded in alumina: An experimental and theoretical study of size dependence, *Phys. Rev. B*, **70**, 165409.

Clemenger, K. (1985), Ellipsoidal shell structure in free-electron metal clusters, *Phys. Rev. B*, **32**, 1359–1362.

Cohen-Tannoudji, C., B. Diu, and F. Laloë (1996) *Mécanique Quantique, Tome II*, Hermann, Paris, France.

Cottancin, E., G. Celep, J. Lermé et al. (2006), Optical properties of noble metal clusters as a function of the size: Comparison between experiments and a semi-quantal theory, *Theor. Chem. Acc.*, **116**, 514.

Cottancin, E., M. Gaudry, M. Pellarin, et al. (2003), Optical properties of mixed clusters: Comparative study of Ni/Ag and Pt/Ag clusters, *Eur. Phys. J. D*, **24**, 111–114.

Courty, A., C. Fermon, and M.-P. Pileni (2001), "Supra Crystals" made of nanocrystals, *Adv. Mater.*, **13**, 254.

De, G., M. Gusso, L. Tapfer et al. (1996), Annealing behavior of silver, copper and silver-copper nanoclusters in a silica matrix synthesized by the sol-gel technique, *J. Appl. Phys.*, **80**, 6734.

de Heer, W. A. (1993), The physics of simple metal clusters: Experimental aspects and simple models, *Rev. Mod. Phys.*, **65**, 611–675.

Dickerson, E. B., E. C. Dreaden, X. Huang et al. (2008), Gold nanorod assisted near-infrared plasmonic photothermal therapy (PPTT) of squamous cell carcinoma in mice, *Cancer Lett.*, **269**, 57–66.

Dijk, M. A. V., A. L. Tchebotareva, M. Orrit et al. (2006), Absorption and scattering microscopy of single nanoparticles, *Phys. Chem. Chem. Phys.*, **8**, 3486.

Doremus, R., S. C. Kao, and R. Garcia (1992), Optical absorption of small copper particles and the optical properties of copper, *Appl. Opt.*, **31**, 5773.

Drude, P. (1900a), Zur Elektronentheorie der Metalle, *Ann. Phys.*, **306**, 566–613.

Drude, P. (1900b), Zur Elektronentheorie der Metalle; II. Teil. Galvanomagnetische und thermomagnetische Effecte, *Ann. Phys.*, **308**, 369–402.

Eckardt, W. (1985), Size-dependent photoabsorption and photoemission of small metal particles, *Phys. Rev. B*, **31**, 6360.

El-Sayed, I. H., X. Hunang, and M. A. El-Sayed (2006), Selective laser photo-thermal therapy of epithelial carcinoma using anti-EGFR antibody conjugated gold nanoparticles, *Cancer Lett.*, **239**, 129–135.

Faraday, M. (1857a), The Bakerian lecture: Experimental relations of gold (and other metals) to light, *Philos. Trans. R. Soc. Lond.*, **147**, 145–181.

Faraday, M. (1857b), Experimental relations of gold (and other metals) to light, *Proc. R. Soc. Lond.*, 145–181.

Fedrigo, S., W. Harbich, and J. Buttet (1993), Collective dipole oscillations in small silver clusters embedded in rare-gas matrices, *Phys. Rev. B*, **47**, 10706.

Ferrando, R., J. Jellinek, and R. L. Johnston (2008), Nanoalloys: From theory to applications of alloys clusters and nanoparticles, *Chem. Rev.*, **108**, 845.

Fox, M. (2001), *Optical Properties of Solids*, Oxford University Press, New York.

Gaudry, M., J. Lermé, E. Cottancin et al. (2001), Optical properties of $(Au_xAg_{1-x})_n$ clusters embedded in alumina: Evolution with size and stoichiometry, *Phys. Rev. B*, **64**, 085407.

Granqvist, C. and O. Hunderi (1978), Optical absorption of ultrafine metal spheres with dielectric cores, *Z. Phys. B: Condens. Matter*, **30**, 47–51.

Hahn, H. and R. S. Averback (1990), The production of nanocrystalline powders by magnetron sputtering, *J. Appl. Phys.*, **67**, 1113.

Harb, M., F. Rabilloud, D. Simon et al. (2008), Optical absorption of small silver clusters: Ag_n, (n = 4–22), *J. Chem. Phys.*, **129**, 194108.

Hohenberg, P. and W. Kohn (1964), Inhomogeneous electron gas, *Phys. Rev.*, **136**, B864.

Hubenthal, F., T. Ziegler, C. Hendrich, M. Alschinger, and F. Träger (2005), Tuning the surface plasmon resonance by preparation of gold-core/silver-shell and alloy nanoparticles, *Eur. Phys. J. D*, **34**, 165–168.

Hunt, L. B. (1976), The true story of purple of cassius. The birth of gold-based glass and enamel colours, *Gold Bull.*, **9**, 134–139.

Iline, A., M. Simon, F. Stietz, and F. Träger (1999), Adsorption of molecules on the surface of small metal particles studied by optical spectroscopy, *Surf. Sci.*, **436**, 51–62.

Issendorff, B. V. and R. E. Palmer (1999), A new high transmission infinite range mass selector for cluster and nanoparticle beams, *Rev. Sci. Instrum.*, **70**, 4497.

Jain, P. K. and M. A. El-Sayed (2007), Universal scaling of plasmon coupling in metal nanostructures: Extension from particle pairs to nanoshells, *Nano Lett.*, **7**, 2854–2858.

Johnson, P. B. and R. W. Christy (1972), Optical constants of the noble metals, *Phys. Rev. B*, **6**, 4370–4379.

Kahnert, N. M. (2003), Numerical methods in electromagnetic scattering theory, *J. Quant. Spectrosc. Radiative Trans.*, **79–80**, 775–824.

Kawabata, A. and R. Kubo (1966), Electronic properties of fine metallic particles, *J. Phys. Soc. Jpn.*, **21**, 1767.

Kelly, K. L., E. Coronado, L. L. Zhao, and G. C. Schatz (2003), The optical properties of metal nanoparticles: The influence of size, shape and dielectric environment, *J. Phys. Chem. B*, **107**, 668–677.

Klar, T., M. Perner, S. Grosse et al. (1998), Surface plasmon resonance in single metallic nanoparticles, *Phys. Rev. Lett.*, **80**, 4249.

Knight, W. D., K. Clemenger, W. A. de Heer et al. (1984), Electronic shell structure and abundances of sodium clusters, *Phys. Rev. Lett.*, **52**, 2141 LP–2143.

Knight, W. D., K. Clemenger, W. A. de Heer, and W. A. Saunders (1985), Polarizability of alkali clusters, *Phys. Rev. B*, **31**, 2539–2540.

Kohn, W. and L. J. Sham (1965), Self-consistent equations including exchange and correlation effects, *Phys. Rev.*, **140**, A1133.

Kreibig, U. and C. V. Fragstein (1969), The limitation of electron mean free path in small silver particles, *Z. Phys.*, **224**, 307.

Kreibig, U. and L. Genzel (1985), Optical absorption of small metallic particles, *Surf. Sci.*, **156**, 678–700.

Kreibig, U. and M. Vollmer (1995) *Optical Properties of Metal Clusters*, Springer, Berlin, Germany.

Kresin, V. V. (1995), Collective resonances in silver clusters: Role of d electrons and the polarization-free surface layer, *Phys. Rev. B*, **51**, 1844.

Kroto, H. W., J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley (1985), C-60—Buckminsterfullerene, *Nature*, **318**, 162.

Larsen, R. A., S. K. Neoh, and D. R. Herschbach (1974), Seeded supersonic alkali atom beams, *Rev. Sci. Instrum.*, **45**, 1511.

Lee, J., A. O. Govorov, J. Dulka, and N. A. Kotov (2004), Bioconjugates of CdTe nanowires and Au nanoparticles: Plasmon-exciton interactions, luminescence enhancement, and collective effects, *Nano Lett.*, **4**, 2323.

Lermé, J. (2000), Introduction of quantum finite-size effects in the Mie's theory for a multilayered metal sphere in the dipolar approximation: Application to free and matrix-embedded noble metal clusters, *Eur. Phys. J. D*, **10**, 265.

Lermé, J., G. Bachelier, P. Billaud et al. (2008), Optical response of a single spherical particle in a tightly focused light beam: Application to the spatial modulation spectroscopy technique, *J. Opt. Soc. Am. A*, **25**, 493.

Lermé, J., B. Palpant, B. Prével et al. (1998a), Optical properties of gold metal clusters: A time dependent local density approximation investigation, *Eur. Phys. J. D*, **4**, 95–108.

Lermé, J., B. Palpant, B. Prével et al. (1998b), Quenching of the size effects in free and matrix-embedded silver clusters, *Phys. Rev. Lett.*, **80**, 5105.

Lermé, J., M. Pellarin, J. L. Vialle, and M. Broyer (1995), Effects of nonlocal ion pseudopotential on the electronic shell structure of metal clusters, *Phys. Rev. B*, **52**, 2868–2877.

Liebsch, A. (1993a), Surface plasmon dispersion and size dependence of Mie resonance: Silver versus simple metals, *Phys. Rev. B*, **48**, 11317.

Liebsch, A. (1993b), Surface plasmon dispersion of Ag, *Phys. Rev. Lett.*, **71**, 145.

Lindfors, K., T. Kalkbrenner, P. Stoller, and V. Sandoghdar (2004), Detection and spectroscopy of gold nanoparticles using supercontinuum white light confocal microscopy, *Phys. Rev. Lett.*, **93**, 037401.

Link, S. and M. A. El-Sayed (1999), Size and temperature dependance of the plasmon absorption of colloidal gold nanoparticles, *J. Phys. Chem. B*, **103**, 4212–4217.

Link, S., Z. L. Wang, and M. A. El-Sayed (1999), Alloy formation of gold-silver nanoparticles and the dependence on their absorption, *J. Phys. Chem. B*, **103**, 3529–3533.

Lisiecki, I. and M. P. Pileni (1995), Copper metallic particles synthesized "in situ" in reverse micelles: Influence of various parameters on the size of the particles, *J. Phys. Chem.*, **99**, 5077.

Lisiecki, I., F. Billoudet, and M. P. Pileni (1996), Control of the shape and the size of copper metallic clusters, *J. Phys. Chem.*, **100**, 4160–4166.

Liz-Marzan, L. M. and A. P. Philipse (1995), Stable hydrosols of bimetallic nanoparticles immobilized on imogolite fibers, *J. Phys. Chem.*, **99**, 15120–15128.

Martin, T. P. (1996), Shells of atoms, *Phys. Rep.*, **273**, 199–241.

Martin, T. P., T. Bergmann, H. Göhlich, and T. Lange (1990), Observation of electronic shells and shells of atoms in large Na clusters, *Chem. Phys. Lett.*, **172**, 209.

Massalski, T. B., J. L. Murray, L. H. Bernett, and H. Baker (1986), *Binary Alloy Phase Diagrams* (Vol. 1), American Society for Metals, Metals Park, OH.

Maxwell-Garnett, J. C. (1904), Colours in metal glasses and in metallic films, *Philos. Trans. R. Soc. Lond.*, **203**, 385.

McFarland, A. D. and R. P. V. Duyne (2003), Single silver nanoparticles as real-time optical sensors with zeptomole sensitivity, *Nano Lett.*, **3**, 1057–1062.

Mie, G. (1908), Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen, *Ann. Phys.*, **25**, 377.

Milani, P. and W. A. de Heer (1990), Improved pulsed laser vaporization source for production of intense beams of neutral and ionized clusters, *Rev. Sci. Instrum*, **61**, 1835.

Mock, J. J., M. Barbic, D. R. Smith, D. A. Schultz, and S. Schultz (2002), Shape effects in plasmon resonance of individual colloidal silver nanoparticles, *J. Chem. Phys.*, **116**, 6755.

Morriss, R. H. and L. F. Collins (1964), Optical properties of multilayer colloids, *J. Chem. Phys.*, **41**, 3357–3363.

Mulvaney, P., M. Giersig, and A. Henglein (1993), Electrochemistry of multilayer colloids: Preparation and absorption spectrum of gold-coated silver particles, *J. Chem. Phys.*, **97**, 7061–7064.

Muskens, O. L., N. Del Fatti, F. Vallée et al. (2006), Single metal nanoparticle absorption spectroscopy and optical characterization, *Appl. Phys. Lett.*, **88**, 063109.

Nilsson, P. O. (1970), Electronic structure of disordered alloys: Optical and photoemission measurements on Ag-Au and Cu-Au alloys, *Phys. Kondens. Mater.*, **11**, 1–18.

Nogami, M., Y. Abe, and A. Nakamura (1995), Cu microcrystals in sol-gel derived glass, *J. Mater. Res.*, **10**, 2648.

Palik, E. D. (1985–1991), *Handbook of Optical Constants of Solids*, Academic Press, New York.

Palpant, B., B. Prével, J. Lermé et al. (1998), Optical properties of gold clusters in the size range 2–4 nm, *Phys. Rev. B.*, **57**, 1963.

Papavassiliou, G. C. (1976), Surface plasmons in small Au-Ag alloy particles, *J. Phys. F*, **6**, L103–L105.

Pedersen, J., S. Bjørnholm, J. Borggreen et al. (1991), Observation of quantum supershells in clusters of sodium atoms, *Nature*, **353**, 733–735.

Pedersen, J., J. Borggreen, P. Chowdhury et al. (1993), Plasmon profiles and shapes of sodium cluster ions, *Z. Phys. D*, **26**, 281–283.

Pellarin, M., E. Cottancin, B. Baguenard et al. (1995), Observation of two successive quantum supershells in a 15000-electron fermionic system, *Phys. Rev. B*, **52**, 16807.

Perez, A., P. Mélinon, V. Dupuis et al. (1997), Cluster assembled materials: A novel class of nanostructured solids with original structures and properties, *J. Phys. D*, **30**, 709.

Perez, A., P. Mélinon, V. Dupuis et al. (2001), Nanostructured materials from clusters: synthesis and properties, *Mater. Trans.*, **42**, 1460.

Perez, A., P. Mélinon, J. Lermé, and P.-F. Brevet (2007), Nanoscience (Chapter 7), *Clusters and Colloids*, Springer Verlag, Berlin/Heidelberg, Germany.

Pileni, M. P. (1989), *Structure and Reactivity in Reverse Micelles*, Elsevier Science, Amsterdam, the Netherlands.

Pollack, S. P., C. R. C. Wang, and M. M. Kappes (1991), On the optical response of Na_{20} and its relation to computational prediction, *J. Chem. Phys.*, **94**, 2496.

Raschke, G., S. Kowarik, T. Franzl et al. (2003), Biomolecular recognition based on single gold nanoparticle light scattering, *Nanoletters*, **3**, 935.

Rauschenbach, S., F. L. Stadler, E. Lunedei et al. (2006), Electrospray ion beam deposition of Clusters and Biomolecules, *Small*, **2**, 540–547.

Richtsmeier, S. C., E. K. Parks, K. Liu, L. G. Pobo, and S. J. Riley (1985), Gas phase reactions of iron clusters with hydrogen. I. Kinetics, *J. Chem. Phys.*, **82**, 3659.

Romero, I., J. Aizpurua, G. W. Bryant, and J. G. de Abajo (2006), Plasmons in nearly touching metallic nanoparticles: Singular response in the limit of touching dimers, *Opt. Express*, **14**, 9988.

Rubio, A. and L. Serra (1993), Dielectric screening effects on the photoabsorption cross section of embedded metallic clusters, *Phys. Rev. B*, **48**, 18222.

Rubio, A., J. A. Alonso, X. Blase, and S. G. Louie (1997), Theoretical models for the optical properties of clusters and nanostructures, *Int. J. Mod. Phys. B*, **11**, 2727–2776.

Rubio, A., L. C. Balbás, and J. A. Alonso (1993), Photoabsorption cross sections of sodium clusters: Electronic and geometrical effects, *Z. Phys. D*, **26**, 284–286.

Russier-Antoine, I., G. Bachelier, V. Sablonière et al. (2008), Surface heterogeneity in Au-Ag nanoparticles probed by hyper-Rayleigh scattering, *Phys. Rev. B*, **78**, 35436.

Sattler, K., J. Mühlbach, and E. Recknagel (1980), Generation of metal clusters containing from 2 to 500 atoms, *Phys. Rev. Lett.*, **45**, 821.

Schmitt, J., G. Decher, W. J. Dressick et al. (1997), Metal nanoparticle/polymer superlattice films: Fabrication and control of layer structure, *Adv. Mater.*, **9**, 61–65.

Sherry, L. J., R. Jin, C. A. Mirkin, G. C. Schatz, and R. P. V. Duyne (2006), Localized surface plasmon resonance spectroscopy of single silver triangular nanoprisms, *Nano Lett.*, **6**, 2060–2065.

Shiomi, H. and K. Umehara (2000), Preparation of SnO_2-Glass composite containing Cu particles reduced from copper ions in glass matrix—Effect of glass particle size on microstructure and electrical property, *Mater. Sci. Res. Int.*, **6**, 144–149.

Siekmann, H. R., C. Lüder, J. Faehrmann, H. O. Lutz, and K. H. Meiwes-Broer (1991), The pulsed arc cluster ion source (PACIS), *Z. Phys. D*, **20**, 417.

Smalley, R. E. (1983), Laser studies of metal cluster beams, *Laser Chem.*, **2**, 167.

Tamaru, H., H. Kuwata, H. Miyazaki, and K. Miyano (2002), Resonant light scattering from individual Ag nanoparticle and particle pairs, *Appl. Phys. Lett.*, **80**, 1826–1828.

Teo, B. K., K. Keating, and Y. H. Kao (1987), Observation of plasmon frequency in the optical spectrum of $Au_{18}Ag_{20}$ clusters: The beginning of the collective phenomenon characteristic of the bulk?, *J. Am. Chem. Soc.*, **109**, 3494–3495.

Tiggesbaümker, J., L. Köller, H. O. Lutz, and K. H. Meiwes-Broer (1992), Giant resonances in silver-cluster photofragmentation, *Chem. Phys. Lett.*, **190**, 42.

Tiggesbäumker, J., L. Köller, K. H. Meiwes-Broer, and A. Liebsch (1993), Blue shift of the Mie plasma frequency in Ag clusters and particles, *Phys. Rev. A*, **48**, R1749.

Torigoe, K., Y. Nakajima, and K. Esumi (1993), Preparation and characterization of colloidal silver platinum alloys, *J. Phys. Chem.*, **97**, 8304–8309.

Weissleder, R. (2001), A clearer vision for in vivo imaging, *Nat. Biotechnol.*, **19**, 316–317.

Yannouleas, C., R. A. Broglia, M. Brack, and P. F. Bortignon (1989), Fragmentation of the photoabsorption strength in neutral and charged metal microclusters, *Phys. Rev. Lett.*, **63**, 255–259.

Zsigmondy, R. A. (1926), Properties of colloids, *Nobel Lecture*.

25

Photoluminescence from Silicon Nanostructures

25.1 Introduction **25**-1
25.2 Synthesis of Silicon Nanostructures **25**-3
25.3 Luminescence Properties of Silicon Nanostructures **25**-7
Optical Characterizations and Luminescence Bands • CW Photoluminescence Experiments • Time-Resolved Photoluminescence Experiments
25.4 The Vibron Model: The Relationship between Surface Polar Vibrations and Nonradiative Processes **25**-15
The Silicon–Silicon Dioxide Interface • The Vibron Model • Experiments Supporting the Vibron Model
25.5 Concluding Remarks **25**-22
Acknowledgments **25**-22
References **25**-22

Amir Sa′ar
The Hebrew University of Jerusalem

25.1 Introduction

Silicon, the second-most abundant element on earth (after oxygen) having superior mechanical and electronic properties, has become the principal material of the semiconductors industry from the dawn of microelectronics, and apparently will remain dominant in the foreseen future. Silicon is a semiconductor whose electrical conductivity can be controlled over a wide range, either dynamically or permanently. Its oxidized state (SiO_2) is one of the best and most stable electrical insulator and its superior chemical and mechanical properties make silicon the ideal material for advanced materials processing. For all these reasons, silicon became essentially the sole player in electronic integrated circuits, being the basic building block of most electronic devices, from transistors and diodes to microprocessors, solar cells, wireless communication devices, and more.[1] Yet, silicon is not a good choice for photonic applications where optically active elements are required due to its indirect energy bandgap characteristics where energy's minima of the conduction and the valence bands do not fall at the same wavevector (i.e., the crystal Bloch momentum normalized to $\hbar$). The situation is schematically illustrated in Figure 25.1 where the energy-band diagram of silicon is shown and compared to that of GaAs, which is a direct bandgap semiconductor.

In a steady state, electrons (open circles) occupy the lowest energy states of the conduction band while holes (filled circles) occupy the upper states of the valence band. The emission of a photon takes place once the electron and the hole recombine radiatively, namely, conduction electrons drop down to empty states of holes in the valence band and, releasing their energy to photons of energy, $\hbar\omega = E_C - E_V \cong E_g$, where E_g is the bandgap energy of the semiconductor and $\hbar\omega$ is the photon energy. Besides energy, momentum should also be conserved during the process. However, as the photon wavelength ($\lambda = 2\pi/k \sim 1\,\mu m$ for silicon) is about three orders of magnitude larger than the de Broglie wavelength of the electrons (which is of the order of the lattice constant of the semiconductors ~5 Å), the photon momentum can be neglected relative to that of the electrons and the holes. In direct bandgap semiconductors like GaAs, both the electrons and the holes have the same momentum at the center of the Brillouin zone (Γ-point in Figure 25.1) and vertical radiative recombination can take place. In silicon, however, the large momentum mismatch between electrons and holes does not permit direct radiative recombination unless another entity, a phonon, for example, is involved in the recombination process. As a result, silicon is a poor emitter of light and cannot be utilized for applications where active light sources are required. In many papers and even textbooks this property of silicon is considered to be a disadvantage. However, one should remember that radiative recombination in direct bandgap semiconductors is a fast process (usually of the order of a few nanoseconds in direct bandgap semiconductors with $E_g \sim 1$–$2\,eV$) that limits the lifetime of the carriers (mainly that of the minority carriers in the semiconductor). In silicon, the slow radiative lifetime (of the order of few milliseconds in pure silicon[1]) allows the minority carriers to diffuse over macroscopically large distances (a few hundreds of micrometers and more), a highly favorable property for electronic applications.

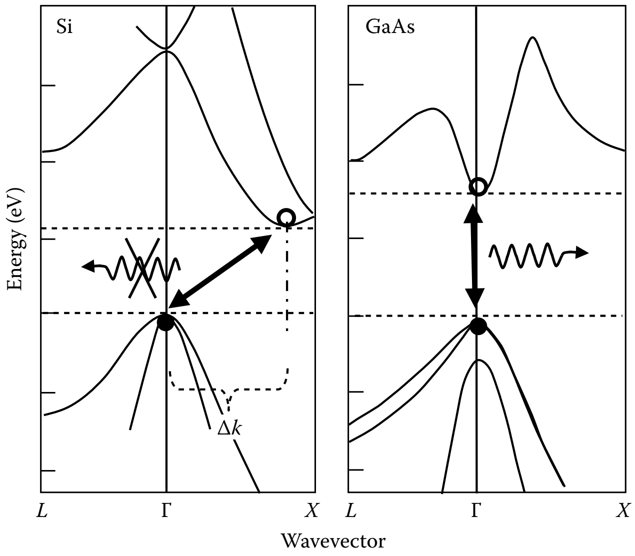

FIGURE 25.1 Energy-band diagrams for silicon and GaAs where the dispersion curves are along the [100] direction (*X*-point) and the [111] direction (*L*-point). Conduction electrons (open circles) and valence holes (filled circles) occupy the band's minima with Δk (momentum/$\hbar$) being the wavevector's difference for silicon. Photon emission is not allowed for silicon.

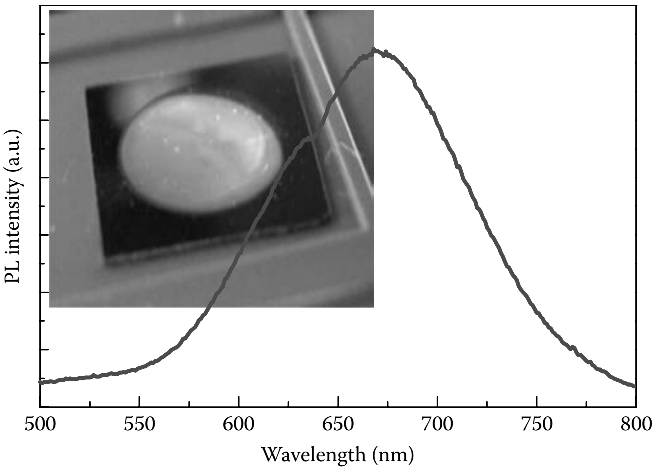

FIGURE 25.2 **(See color insert following page 20-14.)** The PL spectrum from porous silicon with a maximum PL at a wavelength of about 675 nm. The photograph at the inset demonstrates the red color of the emitted PL from a circular layer of porous silicon (the sample has been illuminated with a UV lamp).

The situation is substantially different when dealing with silicon nanostructures. Generally speaking, semiconductor nanostructures[2,3] are artificially made semiconductor objects whose dimensions have been shrunk down to the nanometer length scales. They are classified according to their dimensionality, e.g., two-dimensional (2D) quantum wells, one-dimensional (1D) nanowires (or quantum wires), and zero-dimensional (0D) nanocrystals (or quantum dots[4]), where the classification is according to the number of unconfined dimensions along which carriers (electrons and holes) are still free to move. At the nanometric length scales, quantum confinement (QC) phenomena become important. For example, one should anticipate the momentum conservation law to relax with the decreasing size of the object. In case of silicon nanostructures, for example, a 0D silicon nanocrystal of radius R, we may use the Heisenberg uncertainty principle to estimate the wavevector relaxation: $\Delta k \sim 1/R$. As the relaxation of the wavevector conservation law increases with the decreasing size of the nanostructure, one may expect the radiative recombination rate to considerably increase once the wavevector mismatch, shown in Figure 25.1, becomes comparable to Δk, raising the question: Can we generate active photonic elements[5,6] from silicon nanostructures?

The above question has turned to be a practical subject in 1990 when Canham[7] reported on efficient red light emission from porous silicon (PS) under UV light illumination. PS is a nanometric random network of pores and silicon prepared by electrochemical etching of a silicon substrate.

Figure 25.2 shows a typical spectrum of the photoluminescence (PL) and a photograph of red light coming out from a PS layer. As the efficiency of the PL is comparable to that obtained from direct bandgap semiconductors, Canham in his pioneering work has suggested that quantum size effects might be responsible for efficient PL. Two complementary results seem to support this conclusion. The first is the presence of fairly small crystalline silicon nanostructures, in the form of nanocrystals and undulating nanowires, in the PS medium.[8] Secondly, the considerable blueshift of the maximum PL energy, and accordingly the energy bandgap, from about 1.12 eV (of bulk silicon crystals) to ~1.8 eV in PS, is another manifestation of QC in small semiconductor nanostructures.[2] Following Canham's discovery, a very extensive investigation has been conducted by many groups, aimed at verifying the QC model. Surprisingly, while many investigations provided additional support to the QC model,[9] a considerable number of works have reported results and properties of the PL that cannot be explained by the QC picture.[10] This puzzle has led numerous researchers to propose alternative models and theories that do not rely on QC, the more notable ones suggesting that surface phenomena are responsible for the PL.[11,12] In this picture, radiative transitions take place on the surface of the nanostructures either due to surface bonds, surface defects, imperfections, or even surface molecular species. In particular, the fact that certain properties of the PL depend on "surface chemistry," for example, the specific way that silicon surface bonds are terminated, have led certain researchers to suggest that surface phenomena are responsible for the luminescence from silicon nanostructures.[13]

Let us emphasize that surface phenomena are expected to play a major role in small nanostructures. To follow this, let us estimate the surface-to-volume (STV) ratio for few nanostructures of different dimensionality but having a simple geometrical form such as spheres (0D), cylinders (1D), and slabs (2D), as schematically shown in Figure 25.3.

In order to define a "volume" of the surface, we may estimate the surface's thickness of a given nanostructure, d, to be about 1–3 monolayers thick. For silicon (with a lattice constant of ~0.54 nm),

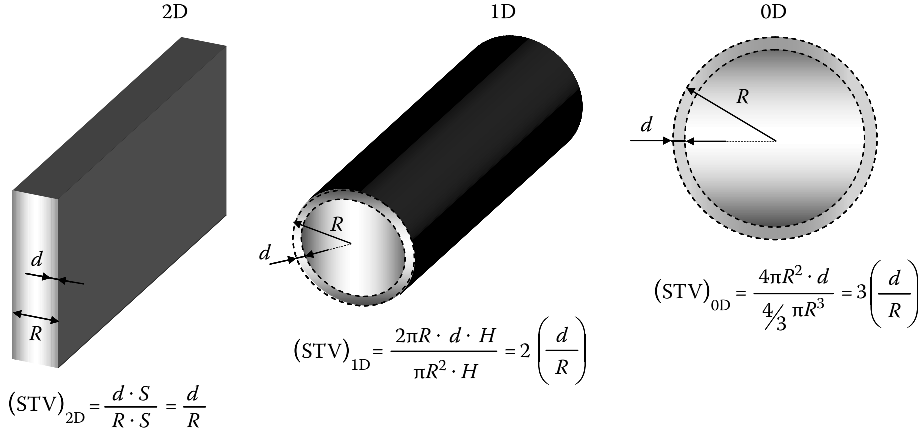
2D
1D
0D
d
R
$(\mathrm{STV})_{2D} = \frac{d \cdot S}{R \cdot S} = \frac{d}{R}$
d
R
$(\mathrm{STV})_{1D} = \frac{2\pi R \cdot d \cdot H}{\pi R^2 \cdot H} = 2\left(\frac{d}{R}\right)$
R
d
$(\mathrm{STV})_{0D} = \frac{4\pi R^2 \cdot d}{4/3\, \pi R^3} = 3\left(\frac{d}{R}\right)$

FIGURE 25.3 The surface-to-volume (STV) ratio for a 2D slab, 1D cylinder, and 0D sphere where R is the thickness of the slab, radius of the cylinder, and radius of the sphere, respectively. In all three cases, d represents the thickness of the surface.

we may take, $d \sim 1\,\mathrm{nm}$, to be an estimate of the surface's thickness. Hence, for a 0D spherical nanocrystal we find, STV = $3(d/R)$, where R is the radius of the sphere, so that for $R = 6\,\mathrm{nm}$ nanocrystal we get, STV ~ 0.5, meaning that 50% of the silicon atoms belong to the surface while for $R = 10\,\mathrm{nm}$, 33% of the silicon atoms sit on the surface of the nanocrystal. We conclude that surface phenomena become more and more appreciable with the decreasing size of the nanostructures and, on the nanoscales not only quantum size effects can affect the electronic properties of the nanostructures but also surface phenomena should carefully be considered as a possible source of interactions that may affect electronic properties of the nanostructures. This is particularly true for silicon, which is known to be material sensitive to surface termination and usually requires a special treatment for passivation.

Unfortunately, PS is not the ideal medium for the study of quantum size effects and surface phenomena at the nanoscale. PS is a random network of electrochemically etched pores and silicon that is characterized by a broad size and shapes distribution, it is chemically and mechanically unstable and it tends to change its properties with time (aging effects).[11] Hence, to a great extent, the puzzles and discrepancies between the QC and the surface chemistry (SC) models could not be resolved despite of the extensive investigation on PS. Yet, over the last decade, alternative techniques to fabricate silicon nanostructures, with better capabilities to control their size, shape, and the host matrix in which they are embedded, have emerged.[14] With these improved techniques, it is now possible to investigate the evolution of optical and electrical properties of silicon nanostructures versus size and dimensionality. Recent experimental results together with refined theories indicate that none of the above models alone, for example, the QC and the SC models, can explain the entire optical properties of silicon nanostructures.[15–17] Instead, a refined comprehensive model, which takes into account both quantum size effects and surface phenomena has to be developed.[18,19] The purpose of this chapter is to describe these recent developments in the field of silicon nanostructures, particularly those experiments and models that provide a support and verification to the mutual role played by QC and SC in shaping the optical properties of silicon nanostructures.[19] The interested readers are referred to earlier reviews and the rich literature in the field discussing each individual model (QC and SC) and the supporting/contradicting experimental results.[5–12]

25.2 Synthesis of Silicon Nanostructures

In this section, we briefly review some of the most popular techniques to synthesize light-emitting silicon nanostructures. The simplest and the less expensive method is based on PS, which is fabricated by electrochemical etching of crystalline silicon substrates using hydrofluoric acid (HF)–based solutions. The specific conditions to obtain light-emitting PS (like the PS sample shown in Figure 25.2)[20,21] are reviewed in detail in Refs. [22,23]. Light-emitting PS has a random, nanometric sponge-like structure with a fairly large surface area that can easily be accessed and chemically modified due to its porous characteristics. While this property of PS is of great advantage for certain applications such as chemical[24,25] and biochemical sensing,[26,27] it represents a major limitation for a consistent study of quantum size effects as the size, shape, and even the dimensionality of the nanocrystalline silicon objects are not well defined and can vary between samples and preparation techniques. Furthermore, as the PS matrix may contain various species such as SiO_x, amorphous silicon (a:Si), and other amorphous derivatives of silicon, the exact surface termination of the nanocrystalline silicon objects cannot be determined accurately. Aging effects can be quite significant in PS and may affect both optical properties and transport phenomena. For all these reasons and despite the extensive literature on luminescence from PS, the author of this chapter is in the opinion that in practice, it is impossible to draw consistent conclusions about nanoscale phenomena from PS alone.

The first experimental report on room-temperature PL from silicon nanostructures other than PS is related to silicon nanocrystals (SiNCs) fabricated by ion implantation of silicon into SiO_2 matrix.[28] Recently after, other methods to produce SiNCs embedded in SiO_2 matrices emerged,[14] including chemical vapor deposition (CVD)[29] of sub-stoichiometric oxide (SiO_x), RF magnetron sputtering,[30] reactive evaporation,[31] and plasma-enhanced CVD (PECVD).[32–34] In principle, all these methods require high-temperature annealing of the deposited films to produce phase separation of the excess silicon from the SiO_2 matrix followed by crystallization of the silicon into nanocrystalline particles. The ion implantation method[35–37] is quite popular due to its compatibility with the standard silicon CMOS technology (where it is routinely used to create doped silicon regions). In this technique, silicon ions are extracted from the plasma, accelerated toward the SiO_2 substrate by an electric field and losing their energy after traveling a given depth in the substrate. The thickness and the profile of the implanted layer depend mainly on the implantation dose and the ion energy. In principle, the lower the ion energy the narrower the implanted zone, however, implantation dose also decreases at lower ion energies, setting up a practical limit on the ion energy, which is typically in the range of 1–10 keV.

A similar result of forming oxide layers with excess silicon concentration has been obtained by RF magnetron sputtering. The sputtering process (sometime referred as physical vapor deposition ≡ PVD) involves a bombardment of silicon and SiO_2 solid targets by energetic ions (usually argon ions), removing atoms and molecules from the targets that are deposited on the substrate. As silicon dioxide is an electric insulator, an RF electric field is used to create the plasma between the substrate and the targets. A popular method to produce SiNCs is the co-sputtering technique[15,38] in which, the two targets (usually pure silicon and quartz) are simultaneously exposed to the ions, producing a mixed layer of silicon and oxide with the excess silicon being used to create SiNCs after high-temperature annealing. The thickness and the amount of excess silicon concentration are determined by the RF power as well as by the geometry of the sputtering chamber. Usually, in order to improve uniformity, the substrate is rotated and placed far enough from the targets. In certain cases however, particularly when size-dependent phenomena are to be investigated, it is desirable to put the substrate near the targets without rotation. In this case, the excess silicon concentration along the deposited film continuously varies from one edge of the substrate to the other edge,[15] as schematically illustrated in Figure 25.4.

The amount of excess silicon concentration can be determined by measuring the volume ratio of silicon and SiO_2 obtained from reference depositions of the single components (Si and SiO_2) under the same deposition conditions and is measured in unit of % excess silicon content in a given volume (x). In the experiments reported in Refs. [15,16], x varied from about 10% (at the edge

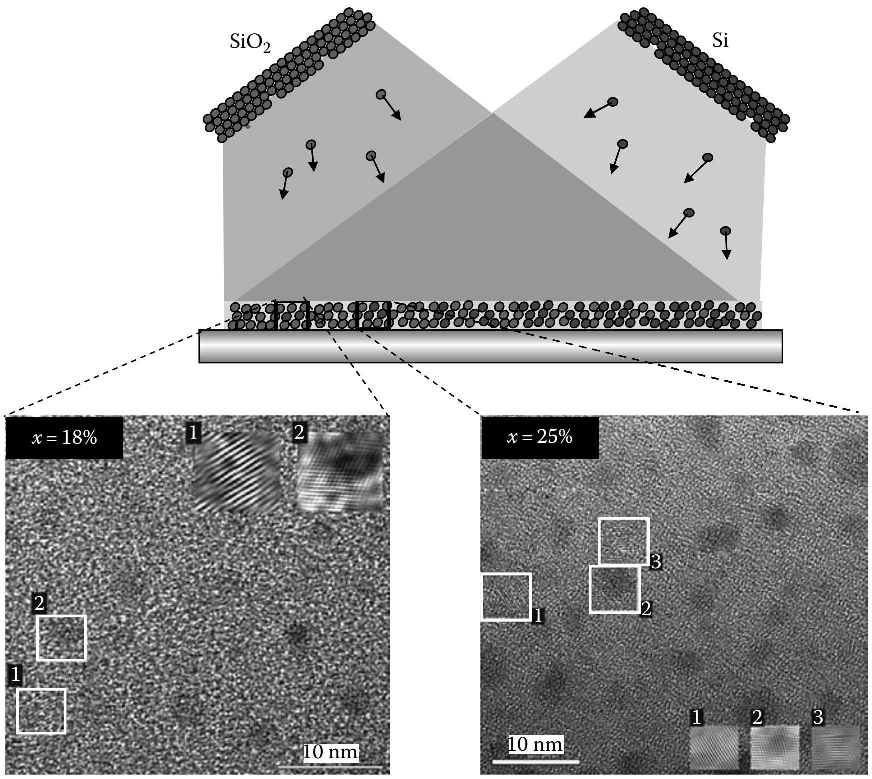

FIGURE 25.4 Schematic view of the co-sputtering method with the two targets of silicon and SiO_2 simultaneously exposed to energetic ions for producing a mixed layer of silicon and SiO_2. The amount of excess silicon decreases from the silicon target edge toward the SiO_2 target edge. Isolated silicon nanocrystals (and PL) have been observed in the range of $10\% \leq x \leq 35\%$. The two cross-section TEM images were taken from two different locations along the substrate with $x = 18\%$ and $x = 25\%$. A few nanocrystals are marked by white squares where the insets are the corresponding images obtained after spatial filtering.

closer to the SiO_2 target) up to about 80% (at the edge closer to the Si target). Luminescent SiNCs are obtained after high-temperature annealing from a limited range of $10\% \le x \le 35\%$ where isolated nanocrystals can be observed. At higher excess silicon content, a connected network of silicon clusters is formed, the PL diminishes, and both dark- and photoconductivities from the deposited layers appear, leading Balberg et al.[39] to propose that a formation of percolating connected network of silicon clusters gives rise to mutual exclusion of the PL and the transport.[40] Annealing of the as-deposited films at high temperatures of about 1150°C–1200°C in a controlled environment (Ar or N gas) leads to phase separation and crystallization of the silicon monomers. In Figure 25.4, we present a few examples of cross-section transmission electron microscope (TEM) micrographs obtained after thinning the samples by ion milling to allow passing of the electron beam through the specimen. The images shown in Figure 25.4 are related to two different values of excess silicon content, $x \approx 18\%$ and 25%. Nearly spherical SiNCs can clearly be seen in both images, revealing typical silicon lattice fringes of mainly the {111} silicon crystal planes (and sometime also the {220} crystal planes). The insets to each figure show the silicon crystallographic planes obtained after spatial filtering of the images in the frequency domain for selected SiNCs marked by solid white squares. The diameter of the nanocrystals is not constant but rather varies in the range of 2–10 nm. The histograms shown in Figure 25.5 were obtained after statistical processing of numerous TEM images, identifying SiNCs and measuring their diameters for each value of the excess silicon volume content. The profile of the histograms can be fitted to a Gaussian shape size distribution with a full width at half maximum (FWHM) of approximately 2 nm at the lower content ($x \approx 10\%$) and 3 nm at the higher silicon volume content ($x \approx 30\%$), reflecting the inhomogeneous characteristics of the samples, which are typical to almost all fabrication techniques. The average diameter of the nanocrystals increases with increasing silicon content and can be fitted to the following power law (solid line in Figure 25.5b):

$$d\,(\text{nm}) \cong 2(x - x_0)^{1/3} \tag{25.1}$$

where

d is the average diameter of the SiNCs

$x_0 = (8 \pm 2)\%$

The power exponent of 1/3 stems from the relationship between the diameter and the volume of the nanocrystals, reflecting the spherical characteristics of the SiNCs. Let us emphasize that SiNCs with a diameter smaller than 2 nm have not been observed so far, independent of the fabrication method that has been utilized, and Equation 25.1 is rather limited to the range of $10\% \le x \le 35\%$.

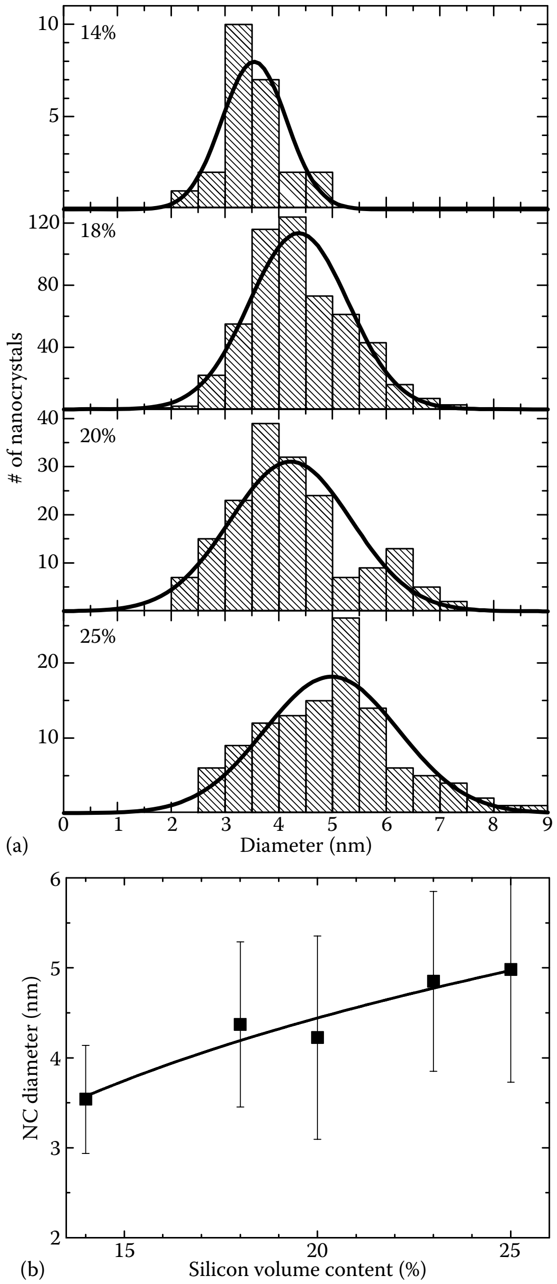

FIGURE 25.5 (a) Histograms showing the size distribution for selected samples having different excess silicon volume contents. The histograms have been obtained after statistical analysis of numerous TEM images, similar to those presented in Figure 25.4, identifying the nanocrystals and measuring their diameters. The solid lines represent the best fit to a Gaussian distribution function. (b) The variation of the average diameter of the nanocrystals versus the excess silicon volume content. The solid line represents the best fit to Equation 25.1.

Another elegant method to produce continuous size variation of SiNCs in a single deposition run is the laser pyrolysis of silane in a gas-flow reactor, which has been reported by Ledoux et al.[41] In this technique, a pulsed CO_2 laser produces a molecular beam of silicon nanoparticles having a size distribution of 3–7 nm with the larger nanoparticles moving slower than the smaller ones. Using a rotating mechanical chopper synchronized with the laser pulses, size-selected nanoparticles are transmitted and deposited at different locations across the substrate. The PL image shown in Figure 25.6 reveals a color variation of the PL

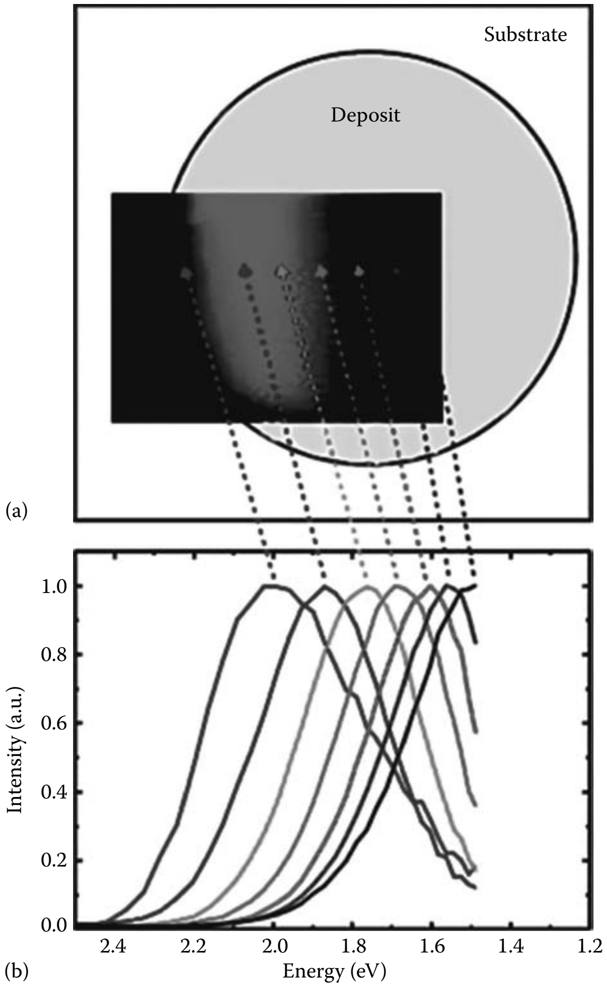

FIGURE 25.6 **(See color insert following page 20-14.)** (a) A photograph showing the PL variation along the substrate from silicon nanocrystals deposited by the laser pyrolysis technique. (b) The normalized PL spectra from different positions along the substrate. (Reprinted from Ledoux, G. et al., *Appl. Phys. Lett.*, 80, 4834, 2002. With permission.)

across the substrate that is correlated with the average size of the silicon nanoparticles.

A slightly narrower size distribution but a much better control of the position and the density of SiNCs has been achieved by synthesizing Si/SiO_2 superlattices. This method has been introduced by Lockwood et al.[42] who have used silicon molecular beam epitaxy (MBE) system combined with ex situ UV ozone oxidation for growing alternating nanolayers of amorphous silicon (a:Si) and SiO_2. Later, this approach has been adopted to create alternating layers of SiNCs and SiO_2 using PECVD,[43] low-pressure CVD (LPCVD)[44,45] and RF magnetron sputtering in the serial sputtering mode.[46] A key factor in all these methods is a good control of the layer's thicknesses down to the nanometric length scales and good uniformity, particularly of the interfaces between layers. A similar approach, based on depositing alternating SiO_x/SiO_2 layers with $1 \leq x < 2$, has also been introduced where phase separation during annealing gives rise to a formation of a:Si layers that later on crystallize to create SiNCs.[47] Results obtained by serial sputtering of silicon/SiO_2 superlattices are shown in Figure 25.7.[48]

In this method, individual layers are deposited one on top of the other at deposition rates of about 2–4 nm/min to achieve a good thickness control. Cross-section HRTEM images, shown in Figure 25.7, reveal the presence of 2D alternating layers of SiNCs separated by amorphous SiO_2 layers. The order and the high quality of the interfaces can be appreciated from the low-magnification TEM image at the inset to Figure 25.7(a) while the higher magnification HRTEM micrograph of Figure 25.7(b) reveals 2D layers of SiNCs with average diameter of about 4.3 nm and a statistical size distribution similar to those presented in Figure 25.5. Figure 25.7(a) presents a STEM (scanning TEM) micrograph obtained after introducing high angle annular dark field (HAADF) detector into the microscope for collecting

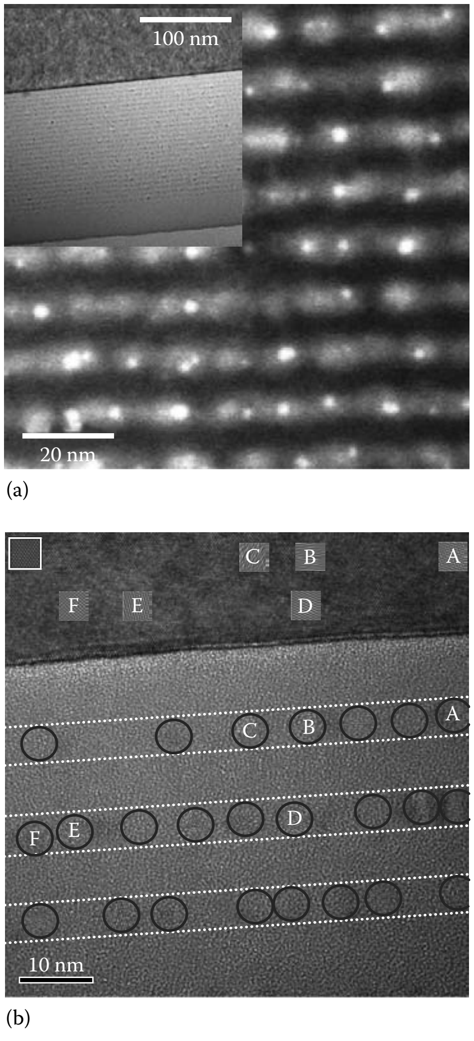

FIGURE 25.7 High-resolution, cross-section TEM micrographs of serially sputtered SiNCs embedded in SiO_2 matrices. (a) STEM image (high z-contrast) revealing nearly spherical SiNCs (bright spots) with average diameter of 3 nm, separated by 10 nm of SiO_2 layers. The inset shows a lower magnification image of the superlattice. (b) HRTEM image of selected layers of SiNCs (marked by circles) having average diameter of 4.3 nm.

electrons that undergo high angle inelastic scattering to yield high *z*-contrast (where *z* is the atomic number of the element). In this image, the average diameter of the SiNCs is about 3 nm.

25.3 Luminescence Properties of Silicon Nanostructures

25.3.1 Optical Characterizations and Luminescence Bands

In general, several luminescence bands can be excited from silicon nanostructures, but most of them are not directly related to the nanostructures but rather to defects and imperfections in the host matrix. A relatively strong but misleading band is the so-called "F"-band (Fast-band) which has been observed for oxidized PS and for most silicon nanostructures embedded in SiO_2 matrices.[11,49,50] The emission spectrum of this band extends over the blue-green (400–550 nm) range of the visible spectrum and is characterized by a fast PL decay time of several nanoseconds. This band has been observed from fully oxidized PS and from silicon nanostructures, such as silicon nanowires,[51,52] with fairly large diameters and therefore, cannot be associated with the nanostructures themselves but with luminescence oxide and interface defects. Other luminescence bands that have been reported but do not relate to quantum size effects are UV bands (~350 nm) and near-IR bands of bulk silicon, unsaturated silicon bonds, and rare-earth heavy ions in the silicon matrix.[11,12]

The size-dependent luminescence band, which shows a clear correlation with quantum size effects in silicon nanostructures, is called the "S"-band (Slow-band) and has been the subject of extensive investigation over the past two decades.[5–13] The emission spectrum from this band can be tuned over the green–red–near IR spectral ranges (500–900 nm) and is characterized by significantly slower PL decay times (compared to the F-band) in the range of few microseconds. In a typical PL experiment, carriers (electrons and holes) are optically excited via absorption of photons of energies exceeding the energy bandgap of the nanostructures, followed by nonradiative relaxation of the photocarriers into the lowest energy levels of the nanostructures and finally, radiative recombination of the carriers that generates PL photons (see Figure 25.8).

Several complementary methods including direct optical absorption,[53] ellipsometry,[54–56] and PLE[11,57] (PL excitation) have been utilized for revealing the absorption spectra from silicon nanostructures. In general, the absorption is substantially blueshifted relative to the PL spectrum with a weak tail below 3 eV and increasing absorption above 3 eV that resembles some of the crystalline and the amorphous silicon characteristics. Therefore, it is not simple to correlate absorption data with the nanostructures themselves as contributions from silicon species, defects, and other absorbing centers in the host matrix cannot be distinguished. The photoexcited carriers undergo a fast relaxation, on timescales of several picoseconds, releasing their energy via nonradiative processes (such as phonons emission) and are trapped in the lower energy levels of the nanostructures. Hence, the luminescence spectrum, particularly the luminescence from the S-band, can directly be correlated with the silicon nanostructures to provide a direct tool for studying the size effects in these nanostructures. Two kinds of PL experiments can be performed. In continuous wave (cw) PL experiments, the system approaches a steady state where all relaxation and transient phenomena have already disappeared and a steady state PL spectrum is measured. Time-resolved PL experiments allow studying dynamical aspects of the nanocrystals where, the investigated dynamics depends on the temporal resolution of the measurement system, particularly the time resolution of the optical detection system that follows the decay of the PL signal shortly after the excitation pulse is switched off. In silicon nanostructures, the dynamics associated with interband (e.g., conduction to valence) radiative and nonradiative relaxation processes is on timescales of few microseconds and therefore, the excitation source can be either a pulsed laser or a beam of a cw laser modulated by an external light modulator (such as acousto-optical modulator).

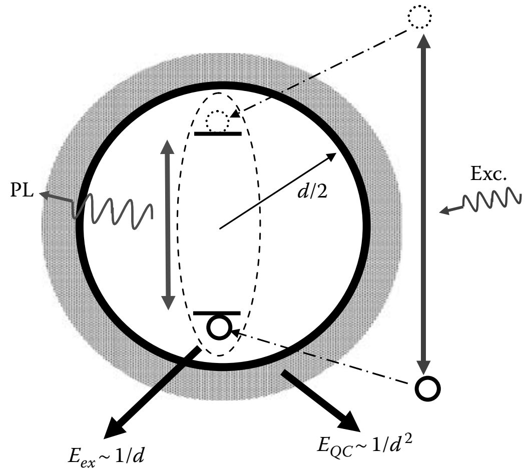

FIGURE 25.8 Schematic of the optical excitation process (solid line) generating electron (dotted circle)–hole (solid circle) pairs, followed by nonradiative relaxation into the lowest energy levels of the nanocrystal (dashed-dotted lines) and finally, radiative recombination of the electron–hole pair (exciton) to generate a photon (gray arrow). The scaling laws for the exciton energy and the bandgap energy (according to the QC model) are shown at the bottom of the figure.

25.3.2 CW Photoluminescence Experiments

One of the most convincing evidences for quantum size effects in silicon nanostructures comes from cw PL experiments on series of SiNCs having different diameters. The room temperature, normalized PL spectra shown in Figure 25.6 (SiNCs fabricated by laser pyrolysis of silane[41]) and Figure 25.9 (SiNCs fabricated by co-sputtering[15]) demonstrate this phenomenon where all PL spectra are blueshifted to higher energies for smaller nanocrystals. This behavior of the PL, e.g., a blueshift of the PL peak energy with the decreasing size of the nanocrystals, has also been observed for PS[9–11] and for SiNCs fabricated by other methods.[14] The relatively large bandwidth of all PL spectra (FWHM ≈ 200 meV) has been

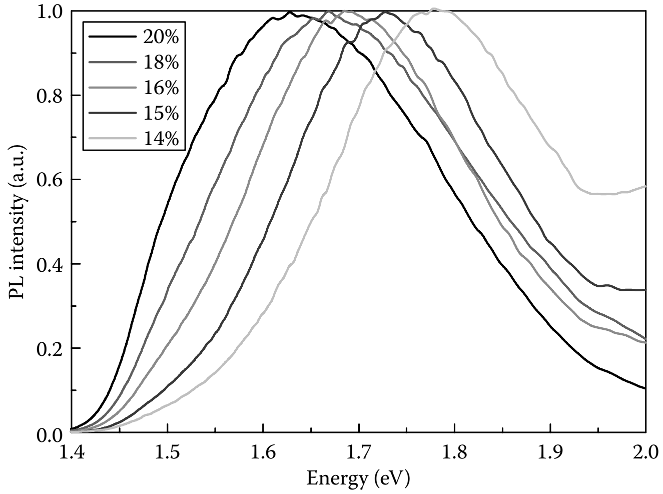

FIGURE 25.9 The normalized PL spectra for co-sputtered SiNCs of different excess silicon volume content.

assigned to the inhomogeneous characteristics of the samples, for example, the broad size distributions shown in Figure 25.5 for co-sputtered SiNCs.

Figure 25.10 (top) presents the variation of the PL peak energy with the increasing excess silicon volume content, x, in the range of $10\% \leq x \leq 35\%$ (bottom axis of Figure 25.10) and accordingly, with the decreasing average diameter of the co-sputtered SiNCs (upper axis of Figure 25.10). The solid line in this figure represents the best fit of the experimental results to an inverse power law as follows:

$$E - E_G = \frac{b}{(x - x_0)^{\delta/3}} \tag{25.2}$$

where

- $\delta/3 \cong 0.45 \pm 0.1$ is the power exponent
- $E_G = 1.12\,\text{eV}$ is the energy bandgap of bulk silicon
- $x_0 \cong (9 \pm 2)\%$ (very close to the experimental value of x_0 obtained for the variation of the average diameter with x; see Equation 25.1)

Using Equations 25.1 and 25.2, we find a relationship between the PL energy shift (relative to bulk silicon) and the diameter of the SiNCs:

$$\Delta E = E - E_G = \frac{A}{d^{\delta}} \tag{25.3}$$

where

- $\delta \cong 1.35 \pm 0.3$
- ΔE is called the "confinement energy" of the nanocrystals

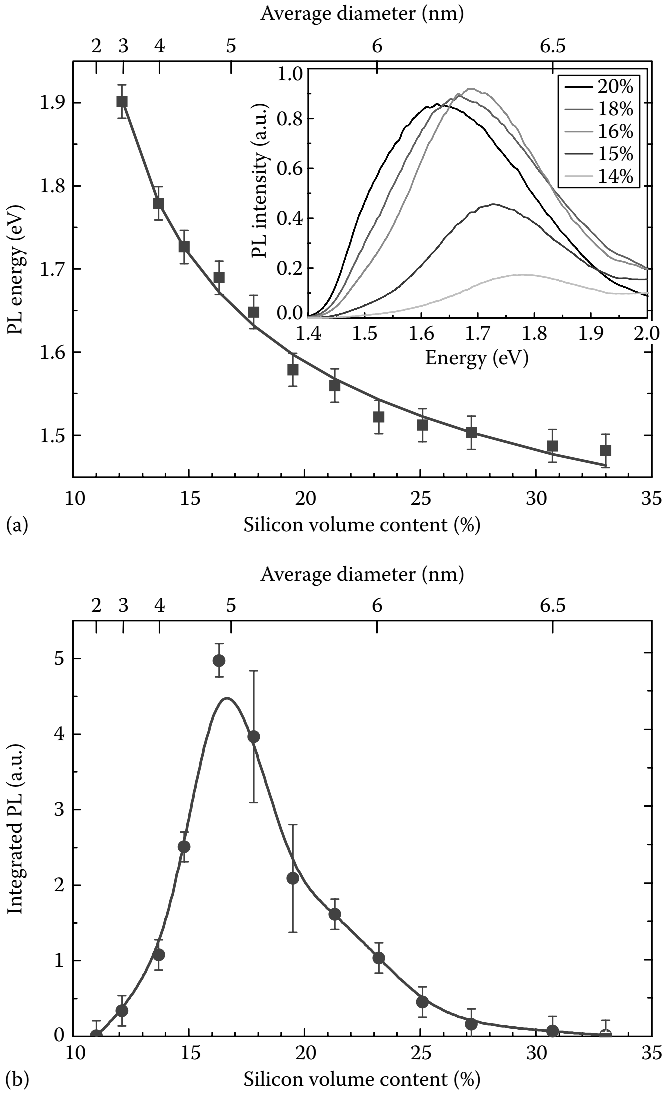

FIGURE 25.10 The variation of (a) the PL peak energy and (b) the integrated PL versus the excess silicon volume content (bottom axis) and the average diameter of the nanocrystals (top axis). The inset shows the non-normalized PL spectra at room temperature.

A very similar inverse power law with essentially the same power exponent has been found for SiNCs fabricated by laser pyrolysis of silane,[41] by PECVD,[34] and by other techniques. However, this power exponent deviates from the exponent predicted from a simple QC model, e.g., a particle confined into a spherical quantum dot having infinite potential barrier. In the latter case, the quantized energy levels of the dot are given by[2] $E_{n,l} = 2\hbar^2 z_{n,l}^2/(m^* d^2) \sim 1/d^2$, where $z_{n,l}$ are the lth root of the nth order spherical Bessel function and m^* is the effective mass of the carriers in the bulk medium. Therefore, in such a simple effective mass model, one would expect the confinement energy to vary as $\Delta E \sim 1/d^2$. Several factors may account for the deviation of δ from this simple model. A major factor, which has been ignored so far, is the Coulomb interaction between electrons and holes to create excitons. Excitons are bound states of electron–hole pairs that are created by the Coulomb attraction between the charged carriers. In a bulk silicon crystal, the Bohr radius of the exciton is given by $a_{ex} = \hbar^2 \varepsilon_{Si}/\mu e^2 \approx 5-7\,\text{nm}$, where ε_{Si} is the dielectric constant of silicon and μ is the reduced mass of the electron–hole pairs. QC modifies this picture, particularly when the size of the nanocrystals becomes comparable to the

Bohr radius of the exciton. In this case ($d/2 \leq a_{ex}$), which is called the strong confinement regime, the electrostatic Coulomb energy is scaled as $E_{ex} \approx e^2/(\varepsilon_{Si}d/2) \sim 1/d$ as opposed to $\sim 1/d^2$ of the effective mass model; see Figure 25.8. Thus, we can expect the power exponent to vary in the range of $1 \leq \delta \leq 2$ with $\delta \approx 2$ being the weak confinement limit for big nanocrystals. Indeed, recent first-principle calculations (meaning that the many-body problem of having a finite number of silicon atoms in a dot, including the specific form of surface termination bonds, has numerically been solved) such as the LCAO (linear combination of atomic orbitals)[58] and the pseudo-potential[59] methods, have found a power exponent of $\delta \cong 1.4$, in a very good agreement with the experimental results, thus providing a solid support to the QC model.

Yet, cw PL experiments provide additional information about SiNCs that cannot be easily interpreted by the QC model. Figure 25.10b shows the variation of the integrated PL intensity, which is obtained by integrating the area below the non-normalized PL spectra[15] (inset to Figure 25.10a) versus the silicon volume content (and the average diameter of the nanocrystals). In order to obtain the actual PL yield, one should take into account the varying density of SiNCs as lower x means lower density of SiNCs. The corrected PL yield, obtained after dividing the integrated PL by x, is shown in Figure 25.11. In this presentation, the precise number of photons absorbed per each unit area could only be estimated. While this quantity has not been measured directly, in several other experiments[41,60–63] aimed at measuring the absolute value of the PL yield, it has been found that the PL yield for SiNCs having a diameter of 4–5 nm is fairly high in the range of 60%–80%. Similar values have also been found for PS.[64] The maximum yield of 30%–50% shown in Figure 25.11 (for nanocrystals emitting their light in the 650–750 nm, red-orange spectral range) should be considered as a lower limit. However, the general trend of the PL yield, e.g., sharp increase with the decreasing diameter up to a maximum for 4–5 nm in diameter nanocrystals and a decrease of the yield for smaller nanocrystals, has been confirmed by numerous groups.[41,63] This finding cannot simply be explained by the QC model. According to this model, the smaller the nanocrystals the more pronounced is the relaxation of the momentum conservation rule (as confirmed experimentally[65]) and accordingly, a larger PL yield should be measured.[66] Further implications of the high PL yield and its relation to the rates of radiative and nonradiative processes in SiNCs is discussed in Section 25.3.3.

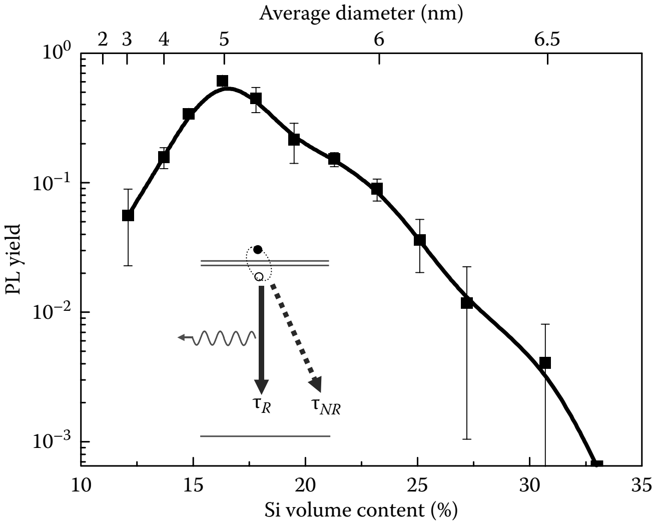

FIGURE 25.11 The PL yield versus the silicon volume content (bottom axis) and the average diameter of the nanocrystals (top axis). The inset illustrates schematically the two relaxation channels of radiative recombination (followed by photon emission) and nonradiative relaxation of the exciton.

Among the first experiments indicating the mutual role of QC and SC is the one reported by Wolkin et al.[17] In this experiment, a series of PS samples having different porosities have been kept under controlled environment to avoid oxidation and to ensure that the surface of the silicon nanostructures is passivated by Si–H chemical bonds. The state of passivation has been monitored by FTIR infrared spectroscopy to exclude the presence of Si–O and Si–OH bonds. A second group of similar PS samples were exposed to air for 24 h, allowing surface oxidation and the presence of Si–O bonds. Results of room-temperature PL measurements from both series of samples are presented in Figure 25.12. Despite that the actual average size of the silicon nanostructures cannot accurately be measured for PS, it is well known that smaller silicon features appear for larger porosities of the samples.[67] The PL spectra, shown in Figure 25.12a, show indeed a consistent blueshift with the increasing porosity for hydrogen-terminated PS samples. On the other hand, oxygen-terminated samples (Figure 25.12b) show, at first, a blueshift of the PL spectra up to the yellow (~600 nm) and then, no blueshift and even a redshift for the blue-green emitting PS samples. Clearly, this phenomenon cannot be explained by the QC model alone and the role of SC, particularly the passivation of the PS surface must be taken into account. The model proposed by Wolkin et al.[17] assumes that some of the dangling bonds, on the silicon oxide interface, are passivated by the Si=O double bond, which is known to be more stable than other forms of surface passivation.[68] Based on this assumption, the authors simulated the electronic band structure of SiNCs having Si=O double bonds at the silicon oxide interface. The results, shown in Figure 25.13, suggest that the QC model is valid for hydrogen-terminated SiNCs as radiative recombinations are via free excitons (e.g., excitons of the bulk nanocrystalline silicon core) for all sizes of the nanocrystals. However, for Si=O passivated SiNCs, both electrons and holes can be localized (or trapped) by the Si=O bonds where, in zone I of Figure 25.13 none of the carriers are trapped, in zone II only electrons are trapped, and in zone III both electrons and holes are trapped to form surface-trapped excitons. This model provides a nice and elegant explanation to the experimental results shown in Figure 25.12, particularly for kinetic properties associated with the joint contribution of SC and QC to the shift of the energy bandgap. Yet, as pointed out by these authors, the intensity of the PL (which is proportional to the integrated PL) increases by few orders of magnitude up to the orange emitting samples and then, diminishes for higher porosity samples emitting in the blue-green; quite similar to the PL yield presented in Figure 25.11 for co-sputtered

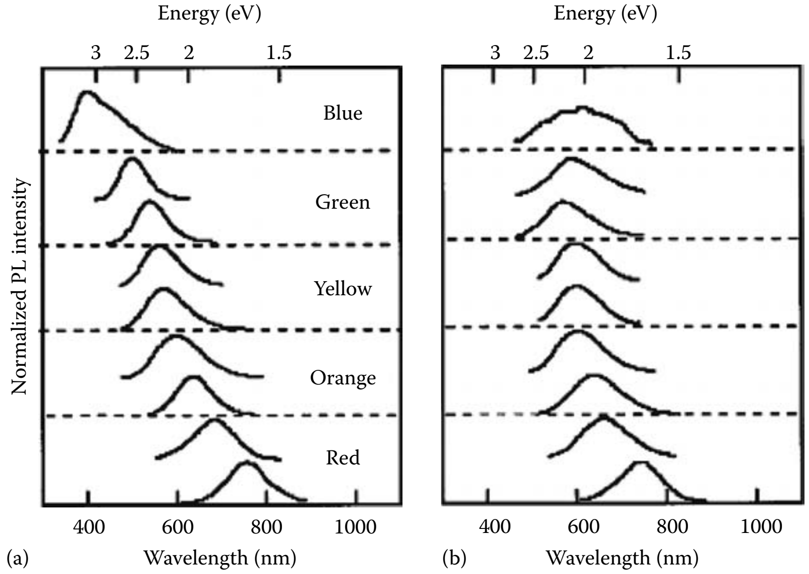

FIGURE 25.12 Room-temperature PL spectra from PS with different porosities (a) kept in Ar atmosphere, (b) after exposure to air. (Reprinted from Wolkin, M.V. et al., *Phys. Rev. Lett.*, 82, 197, 1999. With permission.)

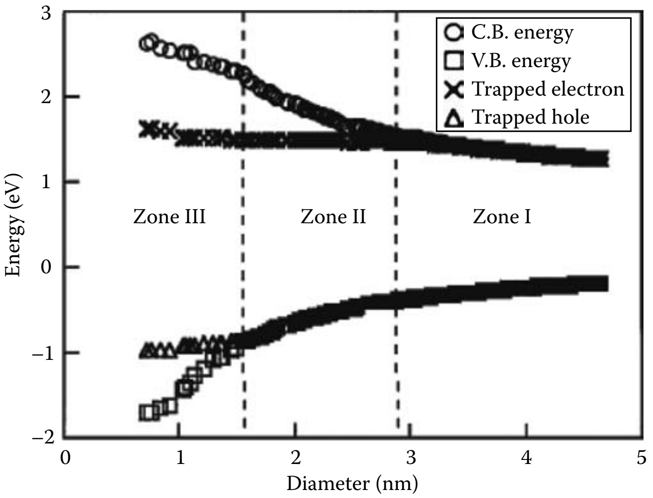

FIGURE 25.13 Electronic states of SiNCs as a function of the cluster size and surface passivation. (Reprinted from Wolkin, M.V. et al., *Phys. Rev. Lett.*, 82, 197, 1999. With permission.)

SiNCs. Hence, dynamical characteristics of the nanocrystals still require sufficient explanations, particularly as the PL decay time from the blue-green PS samples have been reported[17] to be much faster (0.07 μs in the blue) than those of the red samples (~2 μs) and therefore, according to the QC model, should exhibit higher PL yield. Further discussion of this topic will be given in Section 25.4.2 after presenting the vibron model.

25.3.3 Time-Resolved Photoluminescence Experiments

The PL decay process from silicon nanostructures has been intensively investigated over the recent years, both from PS[11] and from SiNCs embedded in SiO_2 matrices.[69] In a typical setup, a train of short laser pulses excites the carriers to their lowest excitonic state from which they radiatively recombine to generate PL photons. The PL decay time is measured shortly after the pump laser is switched off using, for example, a gated photon-counting system that is mostly suitable for SiNCs since the relevant dynamics occurs on timescales of sub-microseconds up to few milliseconds.

A typical room-temperature, time-resolved PL decay curve from co-sputtered SiNCs with x = 18% (average diameter of ~4.5 nm) and at a PL energy of 1.65 eV (PL wavelength of ~750 nm), is shown in Figure 25.14. The temporal behavior of the

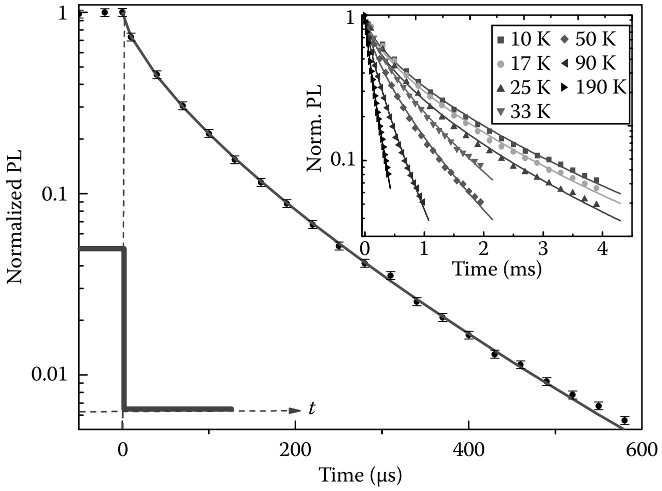

FIGURE 25.14 Room temperature, time-resolved PL decay normalized to the initial PL intensity taken at PL photon energy of 1.65 eV for the x ~ 18% sample (average diameter of 4.5 nm). The solid line represents the best fit of the experimental data (symbols) to the stretched-exponential decay function (Equation 25.4). The inset shows the PL decay curves at various low temperatures.

PL does not follow a simple exponential decay function [e.g., $I/I_0 = \exp(-t/\tau)$; notice the semilogarithmic scale of Figure 25.14], but rather a stretched-exponential decay function of the form:

$$I(t) = I_0 \exp\left[-\left(\frac{t}{\tau}\right)^{\beta}\right] \tag{25.4}$$

where

τ is the characteristic PL lifetime

$0 < \beta \leq 1$ is the dispersion exponent of the PL decay function

This form of the PL decay function has been observed for many classes of silicon nanostructures[37,70] and is frequently assigned to disorder characteristics of systems where dispersive transport take place, particularly migration[71,72] and trapping–detrapping (or release) of photoexcited carriers.[73] This topic is directly related to transport phenomena in PS and SiNCs where dispersive transport has been observed and discussed,[21,74] but will not be reviewed here. The solid line in Figure 25.14 represents the best fit of the experimental data to Equation 25.4, yielding a PL lifetime of about ~40 μs and $\beta \sim 0.85$. As opposed to the dispersion exponent that is approximately independent on the PL energy, the PL decay time decreases with the increasing PL energy, from about 100 μs at 1.4 eV to few μs at 2 eV. This is rather important result as it indicates, again, about the inhomogeneous nature of the PL spectrum, in accordance with the inhomogeneous size distributions of the nanocrystals shown in Figure 25.5. According to the QC model, the smaller the SiNCs (and therefore, the breakdown of the k-conservation rule), the larger the radiative transitions rate and the faster the radiative lifetime. We will follow this picture, assuming that each wavelength in the PL spectrum probes a different size of the nanocrystals, as suggested by the QC model. In addition, let us denote the substantial dependence of the PL decay time (at a given PL energy) on temperature, shown at the inset to Figure 25.14. The lower is the sample temperature the longer is the PL decay time, approaching fairly slow PL lifetimes of about few milliseconds at low temperatures (below 50 K). The variation of the PL lifetime with temperature, for several PL energies, is presented in the Arrhenius plot (semilogarithmic scale versus the inverse temperature) of Figure 25.15.

We can identify two distinct temperature regimes in Figure 25.15: a low-temperature regime (below 50–60 K) where the PL lifetime is essentially independent of temperature and a high-temperature regime (above 60 K) where the PL lifetime gets shorter with the increasing temperature. To follow these characteristics of the PL lifetime let us briefly describe the exchange-splitting model that has originally been proposed by Calcott et al.[75] for PS, and later on adopted for other classes of silicon nanostructures.[69] In typical semiconductors the lowest excitonic state is composed of conduction electrons having a total angular momentum of $J = 1/2$ (the sum of $L = 0$ orbital angular momentum of the conduction band and spin 1/2 of the electrons) and heavy holes states having $J = 3/2$ ($L = 1$ of the valence band and spin 1/2). Hence, the total angular momentum of the exciton can be either 1 or 2. Of course, QC makes this picture much more complicated as different bands can be mixed and the spherical symmetry of the ideal dot can be distorted (in fact, the spherical symmetry of the dot is always slightly distorted when constructed from a cubic crystal). Therefore, Calcott et al.[75] treated the exciton as composed of two spin 1/2 particles assuming that only the spin components of the total angular momentum are conserved. In this case, the total spin of the exciton can be either $S = 1$ (triplet) or $S = 0$ (singlet). The Coulomb exchange interaction between electrons and holes lifts the singlet–triplet degeneracy. To follow the role of the exchange interaction, let us point out that, according to the Pauli principle, a singlet state (which is antisymmetric with respect to exchange of the electron–hole spins) has to be symmetric with respect to exchange of the electron–hole orbital states, while the spin-triplet state has to be orbitally antisymmetric with respect to this exchange. Hence, the Coulomb energy associated with the exchange interaction is always larger for the spin-singlet state ($S = 0$) relative to the spin-triplet state ($S = 1$). The situation is schematically illustrated at the inset to Figure 25.15 where Δ is the singlet–triplet exchange splitting energy. In a bulk silicon crystal, the splitting energy is fairly small, of the order of ~150 μeV; however, since the exchange interaction is proportional to the overlap between the electron and the hole states, it can significantly be enhanced by confining the exciton into small nanostructures. This picture explains very well the behavior of the PL lifetime. Denoting the lifetimes of the upper singlet and the lower triplet states by τ_U and τ_L, respectively, and taking the population of the two states to be in thermal equilibrium (which is a reasonable assumption as relaxation processes between these neighboring states should very fast, of the order of few picoseconds), we find the following expression for the PL radiative lifetime:

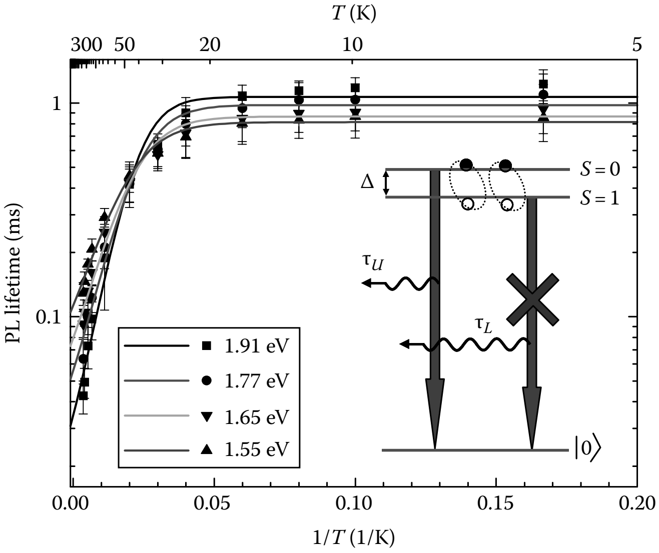

FIGURE 25.15 Arrhenius plot (semilog scale versus the inverse temperature) of the PL lifetime, from the $x \sim 18\%$ sample, for several photon energies. The solid lines represent the best fit of the experimental data to the exchange-splitting model shown schematically at the inset with τ_U, τ_L, and Δ being the upper state (singlet) lifetime, the lower state (triplet) lifetime, and the splitting energy, respectively.

$$\frac{1}{\tau_R} = \frac{g/\tau_L + (1/\tau_U)e^{-\Delta/kT}}{g + e^{-\Delta/kT}} \tag{25.5}$$

where

$1/\tau_L$ is the rate of radiative transitions from the lower triplet state

$1/\tau_U$ is the rate of radiative transitions from the upper singlet state

$g = g_L/g_U = 3$ is the level's degeneracy ratio* and the Bolzmann factor, $\exp(-\Delta/kT)$, takes into account the relative (thermal) population of the states

The solid lines in Figure 25.15 represent the best fit of the PL lifetimes to Equation 25.5. The appearance of two temperature regimes can be understood now as follows. At low temperatures (where $kT < \Delta$), only the lower triplet state of the exciton is populated and the PL decay time is dominated by the long lifetime of the triplet state. At higher temperatures, the upper singlet state becomes populated. If the radiative lifetime of this state is much faster than that of the lower triplet state, it will dominate the PL decay time at high temperatures, giving rise to exponentially faster lifetimes (versus the inverse temperature), which is in a very good agreement with the results shown in Figure 25.15. The upper state lifetime is expected to be (relatively) short since the spin singlet is an optically active state (e.g., radiative recombination of the exciton into the ground state of no exciton is allowed) as opposed to the lower triplet state, which is optically forbidden state according to the spin selection rules. The spin–orbit interaction[11] can mix the singlet–triplet states, making the triplet state weakly allowed. However, this interaction is fairly weak in silicon so that the lifetime of the triplet state is expected to be quite long.

Before discussing the application of this analysis to the various SiNCs samples (of variable silicon volume content), let us denote the contribution of nonradiative relaxation processes to the measured PL decay time. In principle, the measured (inverse) PL decay time is the sum of radiative and nonradiative relaxation rates, $\tau^{-1} = \tau_R^{-1} + \tau_{NR}^{-1}$, where τ_{NR} is the nonradiative relaxation time. Yet, as will be further explained hereafter, at high temperatures the measured PL decay time is dominated by radiative processes and we may take, $\tau_U \cong \tau_S$ (where τ_S is the singlet lifetime). At low temperatures, the contribution of nonradiative processes cannot be neglected; thus we have $\tau_L^{-1} = \tau_T^{-1} + \tau_{NR}^{-1}$, with τ_T being the triplet (radiative) lifetime.

The above analysis, which yields the upper level lifetime, the lower level lifetime, and the exchange-splitting energy for each of the PL energies, has been exploited to the entire set of SiNC samples having different average diameters.[15,16] The results are presented in Figure 25.16 for selected samples. Let us emphasize again that each of these samples is characterized by inhomogeneous, approximately Gaussian, size distribution of the nanocrystals (see Figure 25.5). The PL energy probes a given size (or diameter) of the nanocrystals within this distribution. This is schematically illustrated at the bottom of Figure 25.16, where a smaller PL energy is linked to a larger nanocrystal with a scale of diameters (based on Equation 25.3—the relationship between the PL energy and the diameter of the nanocrystals). For a given size distribution, the PL intensity varies across the PL spectrum yielding more photons near the average size of the

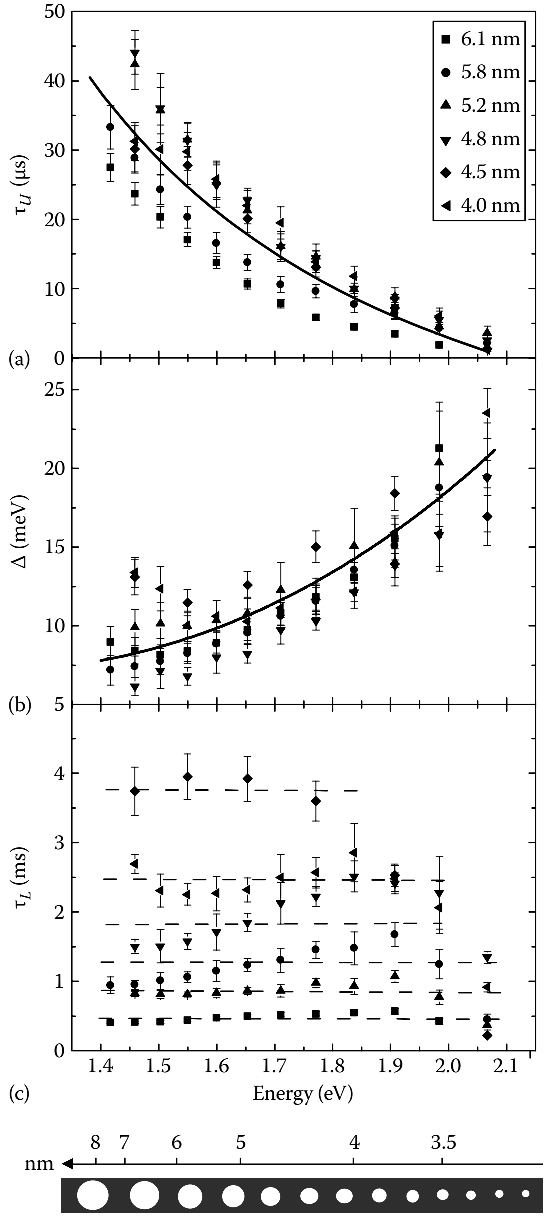

FIGURE 25.16 The variation of (a) the upper level lifetime, (b) the triplet–singlet exchange-splitting energy and (c) the lower level lifetime, as a function of the PL energy for various samples having different average diameters. The lower diameter scale correlates the PL energy to the diameter of the nanocrystals according to Equation 25.3. The solid lines in (a and b) represent a universal dependence of the upper level lifetime and the splitting energy on the diameter of the nanocrystals, while the dashed lines in (c) show the dependence of the lower level lifetime on the average size of the samples.

---

* The degeneracy ratio depends on the assumption regarding the total angular momentum of the states. However, the lifetime analysis is almost insensitive to the specific value of this ratio.

SiNCs. Nevertheless, according to the QC model the PL lifetime should be independent of the size distribution but rather characteristic of a given nanocrystal's diameter. In other words, according to the QC model, the PL lifetime should vary with the PL energy but should be *independent of the size distribution of the nanocrystals*. The results shown in Figure 25.16a for the upper state (singlet) lifetime and in Figure 25.16b for the exchange-splitting energy, remarkably follow the QC model. The upper state lifetime, τ_U, decreases by an order of magnitude (from about 40–50 µs down to 2–3 µs) with increasing PL energy, in agreement with the QC model where larger relaxation of the *k*-conservation rule is expected for smaller nanocrystals. Furthermore, all the results for τ_U from different samples having different average diameters, collapse into a single line (the blue line in Figure 25.16a) independent of the size distributions of the samples. The PL energy probes a lifetime of nanocrystals having the same diameter, which should be independent of how many nanocrystals of that size appear at a given distribution. A similar conclusion holds for the exchange-splitting energy. The smaller the nanocrystal, the larger the exciton confinement and the bigger the splitting energy. Here again, as the splitting energy is a size-dependent property of the nanocrystals, this energy should be independent of the size distribution and the results for all samples should collapse into a single line.

So far, we have found the upper state (singlet) lifetime and the exchange-splitting energy to provide a solid support to the QC model. However, a completely different behavior appears for the lower state (triplet) lifetime, τ_L. The lower state lifetime is essentially independent of the PL energy and therefore, is not a size-dependent property of the nanocrystals. In contrast, τ_L depends on the excess silicon volume content (x), and accordingly, on the average diameter of the SiNCs. Thus, the lower state lifetime cannot be assigned to radiative relaxation from the triplet state but rather to nonradiative relaxation processes that depend on the environment of the SiNCs, e.g., the amount of excess silicon in the host matrix and, as will be explained later, on the SC (or passivation) of the SiNCs. These findings suggest that *QC is responsible for the relatively fast radiative transitions rates from SiNCs, while SC is responsible for the fairly slow nonradiative relaxation processes in these nanocrystals*. Yet, both processes are responsible for the PL yield, η (see Section 25.3.2), which is derived from the competition between radiative and nonradiative recombination rates, as schematically illustrated at the inset to Figure 25.11, and is given by the following expression:

$$\eta = \frac{\tau_R^{-1}}{\tau_R^{-1} + \tau_{NR}^{-1}} = \frac{\tau_{NR}}{\tau_R + \tau_{NR}} = \frac{\tau}{\tau_R} \qquad (25.6)$$

where $\tau^{-1} = \tau_R^{-1} + \tau_{NR}^{-1}$ is the total (measured) PL decay time. For high-yield samples (average diameter in the range of 4.5–5.5 nm), the PL decay time varies by more than two orders of magnitude over the 50–300 K temperature range (see Figure 25.15), while the integrated PL is approximately constant over the same temperature range (it varies by less than a factor of 2; see Figure 25.17).

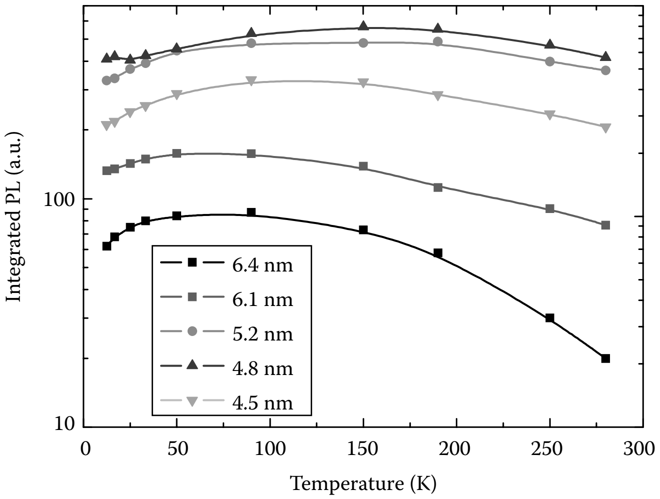

FIGURE 25.17 The variation of the integrated PL as a function of temperature for various samples having different average diameters.

Both experiments (Figures 25.16 and 25.17) suggest that over the 50–300 K temperature range, the PL decay time is dominated by *radiative transitions* of the exciton where, $\tau_R \ll \tau_{NR}$. In this range, we find $\eta \cong$ constant ~1, in a good agreement with the high yield measured for these SiNCs. At this temperature range, radiative transitions are dominated by the upper-state lifetimes, $\tau \cong \tau_U \sim 2$–$50\,\mu$s, which should be compared to the nonradiative lifetimes, $\tau_{NR} \sim 1$–3 ms. Notice that nonradiative rates have been measured at low temperatures and, in general, one should expect faster nonradiative lifetimes at higher temperatures. Nevertheless, all experimental results indicate that $\tau_R \ll \tau_{NR}$ over the entire temperature range, up to room temperature.

Based on the assignment of radiative lifetime to the upper (singlet) state, we can estimate now the oscillator strength for optical transitions from this state. The oscillator strength, $f_{i\to j}$, is a dimensionless parameter that measures the strength of a dipole-allowed optical transition from a given *i*th state to another *j*th state, relative to all other dipole-allowed transitions from the same (*i*th) state. The oscillator strength is normalized so that the sum over all oscillator strengths originated from the same level, is equal to 1, a property which is known as the *f*-sum rule[76] (or the Thomas–Reiche–Kuhn sum rule). The relationship between the oscillator strength and the (spontaneous) radiative lifetime of the singlet state is given by[18]

$$f_{OS} = \frac{2\pi m c^3 \hbar^2 \varepsilon_0}{E^2 e^2 n} \frac{1}{\tau_U} \qquad (25.7)$$

where
- m is the electron mass
- n is the refractive index of the medium
- E is the photon energy

In Figure 25.18, we have plotted the oscillator strength for optical transitions, from the upper singlet state, versus the confinement energy, ΔE. Here again, all results from samples having

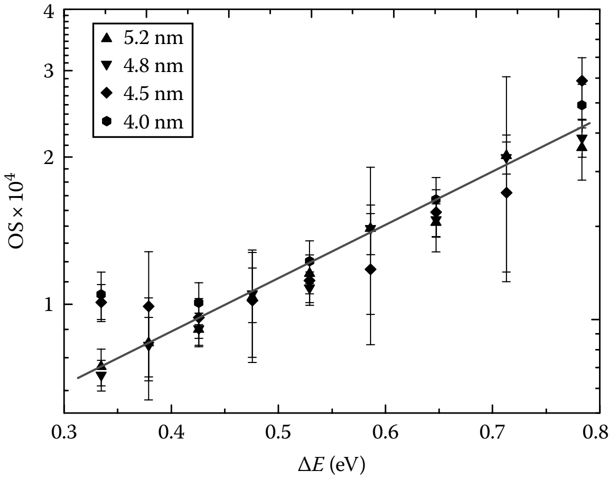

FIGURE 25.18 Semilogarithmic plot of the upper level oscillator strength versus the confinement energy, for various samples having different average diameters. The red line represents the best fit to Equation 25.8.

different average diameters collapse into a single function that, taking into account the limited range of variation, can be fitted into exponential form:*

$$f_{OS} = f_0 \exp\left(\frac{\Delta E}{E_{OS}}\right) \tag{25.8}$$

where

$E_{OS} \cong 0.4 \pm 0.1\,\text{eV}$ is the characteristic oscillator strength energy

$f_0 \cong (3.5 \pm 1) \times 10^{-5}$ is the oscillator strength of the bulk crystal (no confinement)

The above analysis of the PL characteristics seems to adequately explain most of the experimental findings but, at the same time, open up other fundamental questions concerning the origin of the PL. Let us briefly summarize these findings and the most relevant questions[19]:

1. *Radiative Processes*: QC plays a major role here. For small silicon nanostructures, QC gives rise to a blueshift of the bandgap energy, up to the visible range of the spectrum, according to the power law of Equation 25.3. At the same time, the increasing bandgap energy gives rise to shorter radiative lifetimes since $\tau_U \sim 1/E^2$ (without taking into account the effect of confinement on the oscillator strength, which will be discussed hereafter). A second contribution of QC to radiative processes comes from the breakdown of the *k*- (or momentum) conservation rule. However, the effect of confinement on the rate of radiative transitions has been overestimated by many researchers over the years. Our estimate of the oscillator strength for the case of no confinement is, $f_0 \sim 10^{-5}$, in a fairly good agreement with the poor optical emission and the fairly slow radiative lifetimes (≥10 ms) from bulk silicon. QC enhances the oscillator strength for small nanocrystals, but only to the order of ~(2–3) × 10^{-4} and, apparently up to 10^{-3} for blue-emitting nanocrystals. For a comparison, in direct bandgap semiconductors the oscillator strength for interband transitions is ~1, about 3–4 orders of magnitude larger than the oscillator strength of our silicon nanostructures. A direct consequence of this result is the fairly slow radiative lifetimes in silicon nanostructures relative to direct bandgap semiconductors. For example, the radiative lifetime in GaAs (with energy bandgap of ~1.42 eV) is about few nanoseconds compared to few microseconds in silicon nanostructures; a result which is a direct manifestation of the 3–4 orders of magnitude differences in the oscillator strength. Some researchers might assign this conclusion to the so-called quasi-direct bandgap of silicon nanostructures, meaning that the oscillator strength for dipole allowed optical transitions is substantially smaller compared to direct bandgap semiconductors. However, the author of this chapter prefers not to talk about "bands" in small silicon nanostructures as we actually deal with a discrete set of levels rather than "bands."
2. *Nonradiative Processes*: As opposed to radiative processes, QC does not play a major role here. The efficient light emission from SiNCs is due to exceptionally slow nonradiative relaxation processes in SiNCs, of the order of few milliseconds. While this result explains very well the high quantum yield from SiNCs, it raises another fundamental question. The measured nonradiative lifetimes of few milliseconds are typical for the purest silicon wafers, with impurity levels of about 10^{13}–10^{14} cm^{-3} (e.g., one impurity atom per more than a billion of silicon atoms).[1] It is well known that higher density of impurities enhances nonradiative relaxation rates, for example, in p-type silicon crystals with boron concentration of about 10^{16} cm^{-3} the minority carriers' lifetime is about 0.1–1 μs; adding 10^{17} cm^{-3} Au impurities into the crystal gives rise to nonradiative lifetimes as short as 10^{-10} s. The high PL yield and the slow nonradiative relaxation times in SiNCs seem to be a universal property of SiNCs, essentially independent of the method of fabrication (see Section 25.2). Yet, these fabrication methods are not necessarily "clean" and free of imperfections, defects, and impurities. In fact, the quality of interfaces between SiNCs and the host matrices cannot be compared to those obtained by epitaxial growth methods; their shape and symmetry are less regular than nanocrystals synthesized by colloidal chemistry and the level of impurities and defects is far from those obtained by advanced silicon crystal growth techniques. Considering all these facts, one would expect to find much faster nonradiative relaxation times in SiNCs, at least a considerable variation of the PL yield among the various preparation techniques. It seems that *SC plays a major role in excluding nonradiative relaxation channels*

---

* Over the limited range where the oscillator strength has been measured, other functions such as power laws could also be fitted. The current exponential form should be considered as a guideline only.

in SiNCs. Understanding what possible mechanism could give rise to remarkably slower nonradiative relaxation rates is a major issue in the field and is expected not only to explain the origin of the efficient PL, but also to explore novel aspects of SC at the nanoscales. A possible mechanism that could prevent nonradiative relaxation processes at these length scales, and is related to SC, is presented and discussed in the following section (the vibron model). In principle, understanding this mechanism may open up new horizons concerning surface engineering of nonradiative processes at nanoscales.

25.4 The Vibron Model: The Relationship between Surface Polar Vibrations and Nonradiative Processes

The purpose of this section is to discuss basic properties of SiNCs surfaces, particularly the silicon–SiO_2 interface (Section 25.4.1) as most experimental data reported so far are related to SiNCs embedded in silicon dioxide matrices. The "vibron" model[19] (Section 25.4.2), which correlates surface vibrations at the silicon–SiO_2 interfaces to nonradiative processes, is proposed as a specific mechanism that passivates the nanocrystals against nonradiative relaxation channels. Such a mechanism can explain the role of SC and its influence on nonradiative relaxation processes in SiNCs. Finally, in Section 25.4.3, we describe recent experimental results that seem to support the vibron model.

25.4.1 The Silicon–Silicon Dioxide Interface

The silicon–SiO_2 interface is one of the most studied interfaces due to the vital role it plays in microelectronics, particularly in the MOS (metal-oxide-semiconductor) technology. In general, it is well known that thermally grown oxide on top of a crystalline silicon substrate has amorphous structure down to distances of about 1–3 nm away from the interface. The common picture of the interface is that of 1–2 monolayers of non-stoichiometric SiO_x (with $1 < x < 2$), followed by 1–2 nm of strained SiO_2 and a remaining layer of stoichiometric, strain-free amorphous SiO_2.[1] The Si–SiO_2 interface plays a major role in the MOS technology as it can permanently trap charges (interface charges) acting as a source of voltage that shift the electrical characteristics of MOS devices. The amount of surface traps can be as high as 10^{15} cm^{-2} (meaning that essentially all surface atoms can trap charges); however, after hydrogen annealing that passivates most of surface dangling bonds, the amount of surface traps can be reduced to a level of ~10^{10} cm^{-2} or less. The electronic properties of bulk silicon crystals are almost insensitive to these imperfections due to the very small STV ratio of bulk materials (see, Figure 25.3). However, for small SiNCs, the circumstances are substantially different as the number of atoms (or molecules) belonging to the surface of the nanocrystal is similar to the number of silicon atoms of the "bulk" nanocrystal.

Several theoretical and experimental investigations have been conducted for exploring the nature of the interface transition region, of about 1–3 nm, where the structure changes from a perfect order of crystalline silicon to a disordered structure of amorphous SiO_2. Numerous reports[77–80] have shown that the interface may include few monolayers of compressed crystalline, epitaxial-oxide phase in the form of either cristoblite,[81] quartz, or even tridymite.[82] After few monolayers of strained crystalline oxide, the stress is released and amorphous phase of SiO_2 appears. A quite remarkable demonstration to the presence of crystalline SiO_2 phase at the Si–SiO_2 interface has been reported by Cho et al.[83] using thermally oxidized SOI (silicon-on-insulator) substrates to create a thin crystalline silicon layer embedded in between relatively thick amorphous SiO_2 (a 2D silicon quantum well*). The cross-section HRTEM images of these structures, which can be considered as 2D analogous[84] of 0D SiNCs, are shown in Figure 25.19a and b. Despite that the exact crystalline phase of SiO_2 could not be identified in this experiment, these images demonstrate the possibility of creating crystalline SiO_2 at the Si–SiO_2 transition region. Notice also that the model of surface passivation by Si=O double bonds (see Section 25.3.2) proposed by Wolkin et al.[17] has been shown to be consistent with the presence of cristoblite phase at the transition region of the interface.[85] The remarkable point about the TEM images shown in Figure 25.19a and b is the fairly thick crystalline-oxide layer, of ~2–3 nm, which may contain more than 10 monolayers of crystalline SiO_2. The situation is much more complicated when dealing with nonplanar surfaces such as the spherical surface of SiNCs. The EFTEM (energy filtered TEM) image of a single SiNC, shown in Figure 25.19c, was taken from Ref. [86] and is one of the best images of a single nanocrystal ever been reported. Daldosso et al.[86] provided clear evidences to the existence of a Si–SiO_2 spherical transition region, of about 1 nm in thickness, which is marked by the dashed lines in Figure 25.19c. While the signature of crystalline SiO_2 is absent in this image, it is worth noting that such a signature from few monolayers of a crystalline phase, thinner than 2 nm, has not been observed even for planar surfaces and definitely not for the much more challenging spherical surfaces discussed here. The presence of a spherical transition region made of crystalline SiO_2 is not a prerequisite for the vibron model to be discussed in Section 25.4.2. Yet, the existence of such a surface would provide a supplementary support to the model as coherent vibrations across the SiNCs surface, are expected to enhance the vibron effect (see Section 25.4.2) as compared to noncoherent vibrations. The presence of such a surface would also help to understand the analogy between the "classical" polaron problem and the vibron model as "phonons"

---

* Some readers may ask if such 2D silicon nanostructures could show similar quantum size effects (such as PL). However, as far as electronic properties are concerned, these 2D structures behave as 0D SiNCs. The reason is local fluctuations, on length scales of 1–2 monolayers, of the oxide layer that cannot be avoided during thermal oxidation. These fluctuations induce strong localization of the electronic wavefunctions that resembles the 0D characteristics of SiNCs (with lesser control of the additional confinement scales[84]).

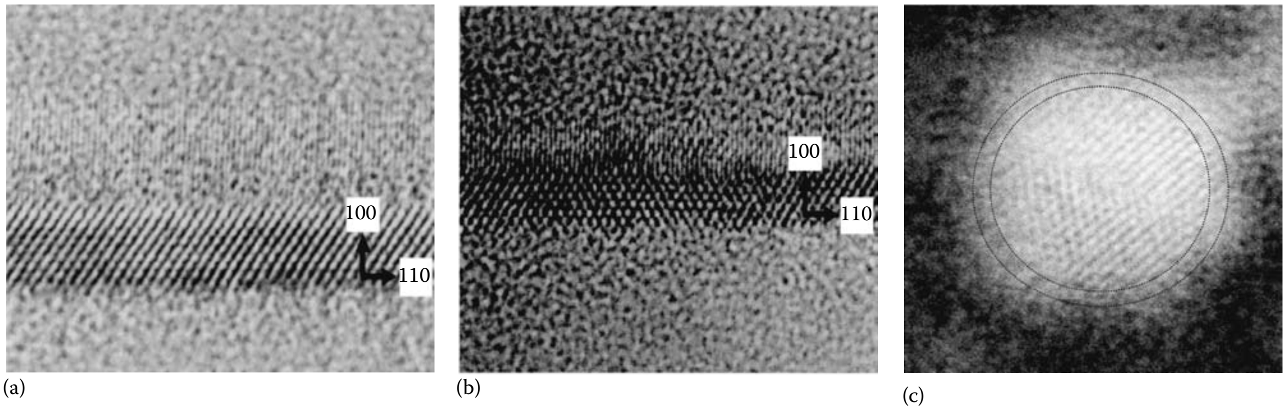

FIGURE 25.19 HRTEM images of crystalline SiO_2 on ultrathin silicon (100) layers with thicknesses of (a) 1.9–2.2 nm and (b) 1.4–1.6 nm. (c) High-resolution energy filtered TEM of a SiNC. The presence of silicon (111) reticular planes is clearly visible in the core region. The dotted lines mark the interface transition region. (Reprinted from Daldosso, N. et al., *Phys. Rev. B.*, 68, 085327, 2003, Cho, E.-C. et al., *J. Appl. Phys.*, 96, 3211, 2004. With permission.)

are natural vibrations of a crystalline SiO_2 phase. However, this is essentially a technical point and the vibron model can be justified even in the absence of a crystalline SiO_2 shell.

25.4.2 The Vibron Model

Following the conclusions of Section 25.3, particularly the conclusion that the high PL efficiency is due to the inhibition of nonradiative relaxation in SiNCs, we should look for a specific mechanism, associated with SC, which can "passivate" the nanocrystals against nonradiative relaxation processes. Such a mechanism has been proposed in Refs. [18,19,87,88] and is called the "vibron" model. Let us briefly explain this model that assigns the inhibition of nonradiative channels to a resonant coupling between electronic states of the nanocrystals and surface vibrations of the silicon–oxygen bonds. These vibrations can be either noncoherent, e.g., vibrations associated with the Si–O and/or the Si=O bonds on the surface of the SiNCs, or coherent vibrations of a spherical crystalline SiO_2 shell that wraps the inner crystalline silicon core of the nanocrystals.

In general, a major source of scattering and relaxation in any semiconductor is the electron scattering from lattice vibrations of its own crystal, known as phonons.[89] The phonon dispersion relation (e.g., the relation between the frequency and the wavevector of the phonons) consists of two branches: an acoustic branch (where ω goes to zero as $\vec{k}$ approaches zero) and an optical branch of phonons that can interact with light (with nonzero value of ω at $\vec{k} \to 0$). In addition, each of these branches is divided into longitudinal and transversal phonon modes where, for longitudinal modes (LO and LA), the atoms vibrate parallel to the propagation direction while for transversal modes (TO and TA), the atoms vibrate perpendicular to the propagation direction of the phonons. In polar semiconductors (such as GaAs), where ions constructing a unit cell of the crystal carry positive and negatives (ionic) charges, longitudinal vibrations of the crystal generate a long-range polarization field. The polarization field is an electric field produced by the electric dipole moment of the vibrating ions, which is responsible for the strong interaction between LO phonons and charged carriers (electrons and holes) in polar semiconductors, and is a major source of scattering and energy dissipation in these semiconductors.[90] In silicon, however, the situation is quite different as silicon is a covalent semiconductor having the same atoms in a unit cell of the crystal that do not carry a net charge. Therefore, neither longitudinal nor transversal phonon vibrations in silicon generate a polarization field, giving rise to significantly slower phonon scattering rates compared to polar semiconductors. The presence of charged impurities, vibrating bonds, or any other source of oscillating electric dipoles (or charges) on the surface of silicon could generate electric field that interacts with the charged carriers. However, these are small surface effects that have minor impact on the electronic properties of bulk silicon crystals.

The situation is substantially different when dealing with low-dimensional silicon nanostructures, particularly small SiNCs. In these nanocrystals, the surface can no longer be treated as a small perturbation since a significant portion of the atoms sit on the surface of the nanocrystals (see Figure 25.3). While vibrations of the core silicon crystal cannot produce a polarization field, polar vibrating bonds on the surface of the nanocrystals can generate a polarization field, in the vicinity of the nanocrystals, that interacts with the charged carriers. This is the basic mechanism that stands behind the vibron model[19]. Some readers may argue that this picture seems to contradict our goal, e.g., finding a mechanism that will allow to passivate the nanocrystals against nonradiative processes. Indeed, the smaller are the nanocrystals the bigger is the surface area giving rise to larger amount of nonradiative relaxation channels. This is also the case for SiNCs with a diameter larger than 10 nm. However, for smaller SiNCs with a diameter less than 10 nm, a new mechanism of resonant coupling between surface vibrations and electronic states appears and allows passivating the nanocrystals against nonradiative channels.

To follow the origin of this mechanism, let us consider the SiNC schematically illustrated in Figure 25.20, which consists

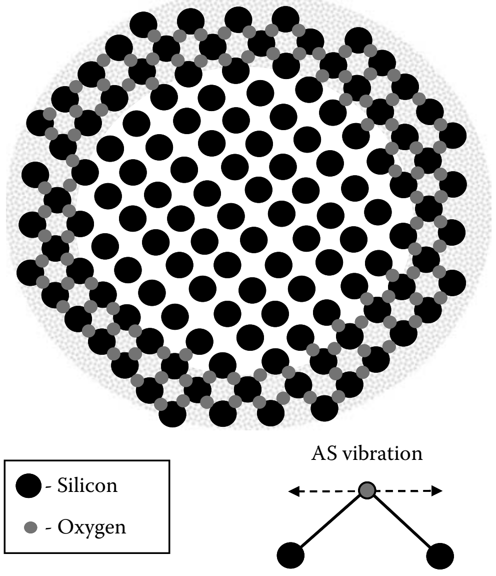

FIGURE 25.20 Schematic illustration of a spherical silicon nanocrystal terminated with silicon–oxygen bonds. The dashed area represents a surface region where a crystalline SiO_2 shell could be formed. The inset illustrates the Si–O–Si asymmetric stretching mode of vibration.

of (approximately) spherical nanocrystalline silicon core terminated with silicon–oxygen bonds on its surface. Traditionally, finding the electronic levels and the vibrational modes of a bulk semiconductor have been treated as two independent problems (the so-called adiabatic approximation) as the electronic energies are of the order of few eV while vibrational (or phonon) energies are an order of magnitude smaller. At present, we will follow this picture, but later on we will reexamine this assumption. The SiNCs spectrum of electronic levels has been calculated by several groups using various techniques such as tight-binding,[58,91] pseudo-potential,[59] and effective mass.[92,93] Most reports have focused on the lowest unoccupied conduction level (called the LUMO state) and the highest occupied valence level (the HOMO state) as these levels are involved in interband optical transitions contributing to the PL. However, the vibron model requires some knowledge about higher excited states of the system. In direct bandgap semiconductor nanocrystals, it is convenient to classify the electronic levels according to their orbital symmetry using the standard atomic orbital notations, e.g., *S*, *P*, *D*, etc.[2] Several factors make the classification for SiNCs more complex. First, the anisotropic effective mass of conduction electrons (with a transverse mass of $0.19m_0$ and longitudinal mass of $0.92m_0$) may gives rise to a splitting of the levels with lower energy levels associated with the heavier mass. Secondly, the six minima degeneracy of the conduction band in bulk silicon (near the *X*-point, along the family of (100) directions; see Figure 25.1), which should be preserved for an ideal spherical dot, will be lifted by any deviation from a perfect spherical symmetry. In addition, a unique splitting into two groups of levels that are symmetric/antisymmetric under inversion (e.g., $\vec{k} \rightarrow -\vec{k}$) are expected for small SiNCs due to intervalley couplings.[91] Finally, mixing between different bands is expected in small SiNCs.

For the purpose of our model, we will ignore most of these fine splitting and mixing features, referring to the lowest two sublevels of each band as $1S_e$ and $1P_e$ for conduction electrons ($1S_h$ and $1P_h$ for holes), as schematically illustrated on the left side of Figure 25.21. E_g is the effective energy bandgap (the energy difference between the $1S_e$ and the $1S_h$ states) and ΔE_C and ΔE_V are the energy differences between lowest conduction and valence sublevels, respectively. On the right side of Figure 25.21, we schematically illustrate the configuration space diagram for the Si=O surface bonds (e.g., the variation of the potential energy versus the length of the Si=O bond) with a ground vibrational state and a first excited state having a vibrational energy of $\hbar\omega_{VB}$ (a common approximation here is taking the potential energy to be parabolic and the quantized vibrational levels and energies being the levels of a simple harmonic oscillator). The vibrational energies are measured upward for conduction electrons and downward for valence holes. Notice that vibrational energies are size independent (e.g., independent on the diameter of the nanocrystal), depending only on the type of bonding used to terminate the surface of the nanocrystal. On the other hand, the quantized electronic levels of the nanocrystal are size dependent and should increase with the decreasing diameter of the SiNCs. Hence, for (relatively) big SiNCs, one expects the conduction/valence energy differences to be fairly small compared to vibrational energies. With the decreasing size of the

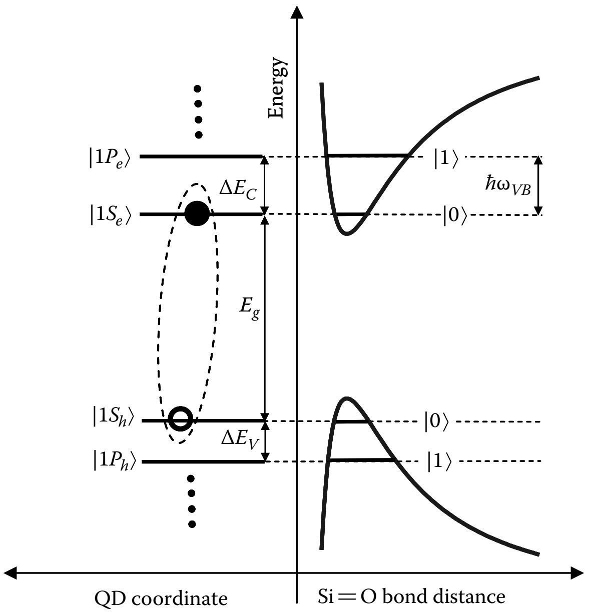

FIGURE 25.21 Schematic illustration of the condition for resonant coupling between electronic sublevels and surface vibrations. On the left, the lowest electronic sublevels of the conduction ($1S_e$ and $1P_e$) and the valence ($1S_h$ and $1P_h$) bands are shown. ΔE_C and ΔE_V are the energy differences between these sublevels. On the right, the ground and the first excited vibrational states are shown (with $\hbar\omega_{VB}$ being the vibrational energy). The solid lines represent the potential energy for surface vibrations (in the configuration space). Resonant coupling occurs once the vibrational energy is equal to ΔE_C or ΔE_V.

nanocrystals, ΔE_C (and ΔE_V) increases until, for a given diameter of the nanocrystals, the two energies (electronic and vibrational) coincide. This is the condition for *a resonant coupling between surface vibrations and the electronic states of the nanocrystals via a polarization field of the vibrations.*

The above model should slightly be modified when dealing with surface phonons of a crystalline SiO_2 shell, but the main conclusion concerning resonant coupling between electronic states and surface phonons, remains valid.* In this case, we will focus our attention on LO phonons as these phonons generate a considerably large polarization field. In general, the dispersion relation of the phonons should be taken into account (e.g., the dependence of the phonon energy on the phonon wavevector, $\vec{q}$), but, to a first order approximation we may take the LO phonon energy to be dispersionless and equal to $\hbar\omega_{LO}$. When the condition for a strong coupling is fulfilled, $\Delta E \approx \hbar\omega_{LO}$, the electrons and the phonons are not anymore in the weak coupling regime where scattering and energy dissipation occur. Instead, the strong coupling between electrons and phonons gives rise to a creation of virtually everlasting mixed electron–phonon modes, called polarons. In polar semiconductor quantum dots, these polarons are due to strong coupling between the electrons and the LO phonons of the same crystalline core while in the vibron case, a similar coupling occurs between surface phonons of the crystalline shell and the core electrons. These surface polarons "passivate" the surface against nonradiative relaxation processes. Let us emphasize that this process of creating surface polarons (vibrons) is expected to be particularly important in silicon where, as opposed to polar semiconductors, only surface phonons (or vibrations) can produce a polarization field explaining the crucial role of SC in SiNCs.

The polaron state of nanocrystals can be viewed as a "dressed" state of the electron–phonon system. A simple illustration of a polaron state is depicted in Figure 25.22 where the negatively charged electron produces a local disturbance to the polarization field of the lattice (which is created by longitudinal lattice vibrations). In the vibron case, the surface polaron is a "dressed" state of the surface, meaning that the electron and/or the hole are coupled to the surface of the SiNC. This provides a direct explanation and a specific mechanism to the model of Wolkin et al.[17] of surface-trapped electrons/holes/excitons discussed in Section 25.3.2. Here again, resonant coupling to surface vibrations can produce a localized (or trapped) electron-surface vibrations, hole-surface vibrations, or a localization of the entire exciton as suggested by Verzelen et al.[94] Another important characteristic of the polaron problem, which can be utilized for experimental verification of the model, is the polaron energy splitting due to confinement. To follow this phenomenon, one needs to solve the entire polaron Hamiltonian including the electron-polar phonon interaction, finding the new eigen-energies of the polaron that becomes a mixed state of the electrons (or the holes) and the phonons. While the exact algebra is quite complicated and tedious, a simplified model has been proposed by Mahdouani et al.[95] by limiting the interaction into the subspace of the $|1S_{e(h)}, 1\vec{q}\rangle$ and $|1P_{e(h)}, 0\vec{q}\rangle$ uncoupled states, which are composed of the electronic ground state ($1S_{e(h)}$) + 1 phonon state ($1\vec{q}$), and the first excited electronic state ($1P_{e(h)}$) + 0 phonon state ($0\vec{q}$), respectively. In this subspace, the electron–phonon interaction Hamiltonian (known as the Fröhlich Hamiltonian[96]) takes the following form:

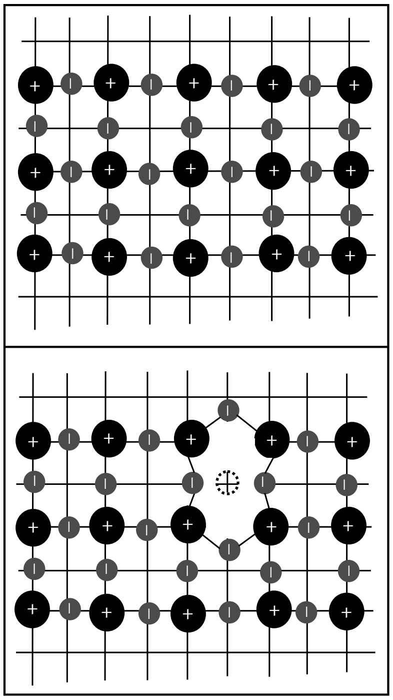

FIGURE 25.22 Schematic description of the polaron formation mechanism. On the left, the polarization field from the unperturbed lattice of charged ions is shown while on the right, the perturbed field (e.g., the polaron) created by a negatively charged electron (dotted circle) is presented.

$$H_{POL} = \begin{pmatrix} E_S + \hbar\omega_{LO} & W \\ W^* & E_P \end{pmatrix} \tag{25.9}$$

where

E_S, E_P are the electronic energies of the $1S$ and the $1P$ sublevels, respectively

$\hbar\omega_{LO}$ is the LO phonon energy

The off-diagonal electron–phonon interaction term is given by $W = \langle 1S_{e(h)}, 1\vec{q} | H_{el-ph} | 1P_{e(h)}, 0\vec{q} \rangle$

The solution to this problem (e.g., diagonalization of the Fröhlich Hamiltonian) yields the mixed electron (hole)–phonon states of the vibron and the energies of the vibron, which are schematically illustrated in Figure 25.23.

The lower energy branch (E_-) has a phonon-like behavior ($E_- \approx \hbar\omega_{LO}$) for small SiNCs and behaves as an electronic state ($1P_e$) for large SiNCs (with $E_- \approx \Delta E_C \sim 1/d^{\eta}$; see Equation 25.3). The upper branch (E_+) is electronic-like for small SiNCs and phonon-like for

---

* For the sake of simplicity, we have ignored the effect of confinement on the spectrum of phonons and their dispersion relation. However, this topic could lead to additional novel phenomena that have not been studied so far.

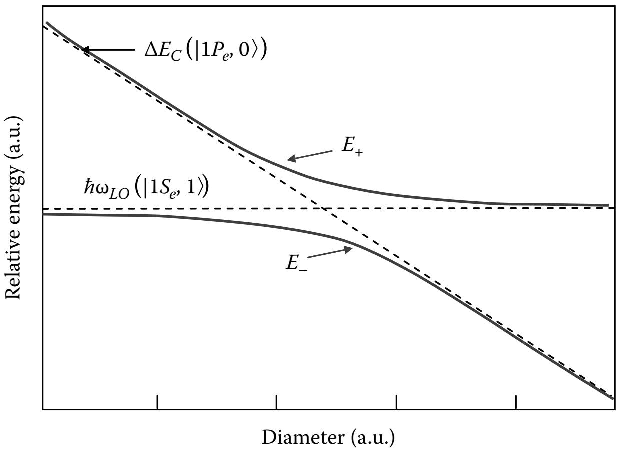

FIGURE 25.23 Schematic presentation of the vibron energies (solid lines) as a function of the nanocrystal diameter. The dashed lines represent the vibrational energy, which is not affected by the confinement (horizontal line), and the electronic energy (ΔE_C), which depends on the diameter of the nanocrystal.

large SiNCs. However, the most interesting behavior occurs for $\Delta E_C \approx \hbar\omega_{LO}$ (the resonant coupling condition), where the polaron states are mixed states of the form $\left|\psi_{POL}^{\pm}\right\rangle = a_{\pm}\left|1S_e, 1\vec{q}\right\rangle + b_{\pm}\left|1P_e, 0\vec{q}\right\rangle$ (with $\left|a_{\pm}\right|^2 + \left|b_{\pm}\right|^2 = 1$). Similar results can be obtained for the hole–phonon coupling[95] and for the exciton–phonon coupling.[94] In principle, the same formalism can be applied to noncoherent surface vibrations with $\hbar\omega_{VB}$ replacing $\hbar\omega_{LO}$. Yet, this problem requires further investigation as the application of the Fröhlich Hamiltonian for coupling to surface vibrations and/or amorphous phases has not been investigated so far. Finally, let us point out that measuring the vibron energies can provide a direct verification to the model. This is the subject of Section 25.4.3.

25.4.3 Experiments Supporting the Vibron Model

At first, let us mention several reports about polar optical phonons of the silicon–SiO_2 planar interfaces, which have been theoretically predicated[97] and experimentally observed.[98] In these planar Si–SiO_2 interfaces, polar optical phonons of the oxide layer give rise to enhanced scattering and energy dissipation, from charged carriers moving in the adjacent silicon channel of MOS devices and bipolar transistors having SiO_2 spacer layers.[98] In this case, the interaction with polar optical phonons causes a lower mobility of the carriers and a degradation of the current–voltage characteristics. As discussed in the previous section, this is the regime of weakly coupled electrons–phonons that is reflected by energy dissipation and enhanced nonradiative relaxation rates.

Among the first experiments providing direct evidence for vibrons in SiNCs is the one reported in Refs. [87,88]. In this experiment, the aim was to measure inter-sub-level (ISL) optical transitions between the quantized sublevels of SiNCs,[93] for example, the $|1S_e\rangle \rightarrow |1P_e\rangle$ and the $|1S_h\rangle \rightarrow |1P_h\rangle$ ISL transitions between the quantized states of the conduction and the valence bands, respectively (see Figure 25.21, left). Ignoring band mixing, let us indicate that these are optical transitions between sublevels coming from the same band and therefore, having the same Bloch wavefunction (for example, the same conduction Bloch state for both the $|1S_e\rangle$ and the $|1P_e\rangle$ sublevels). Hence, dipole matrix elements for optical transitions between these sublevels are reduced to dipole matrix elements between envelope states of these sublevels.[99,100] Thus, the $|1S_e\rangle \rightarrow |1P_e\rangle$ transition is a dipole allowed optical transition (with the oscillator strength for this transition being close to 1[91]) that can be measured by infrared absorption spectroscopy. Let us emphasize that ISL optical transitions are a direct manifestation of QC as these sublevels do not exist in bulk semiconductors.[101] In addition, according to the vibron model, these sublevels are expected to interact with surface polar vibrations yielding modified energy spectra similar to those presented in Figure 25.23. Therefore, this experiment should be sensitive to both QC and SC in SiNCs.

The experimental setup for measuring ISL transitions[101] in SiNCs includes a visible pump laser exciting electron–hole pairs* (excitons) and a weak, tunable IR probe beam measuring the photoinduced absorption (PIA) spectra. Both the photoinduced transmittance, ΔT, and the linear IR transmittance, T, are measured (by modulating the pump laser beam in one experiment and the probe beam in a second experiment using a lock-in detection method[87]) to obtain PIA: PIA = $\Delta T/T$.† The same set of co-sputtered SiNC samples, with varying average diameter (and silicon volume content; see Figure 25.5), which were used for the PL experiments (see Figures 25.9 through 25.11), have also been exploited for PIA measurements to allow investigation of possible correlation between these experiments.

Figure 25.24 presents the linear IR absorption (dashed line) and the PIA (solid line) spectra for SiNCs having an average diameter

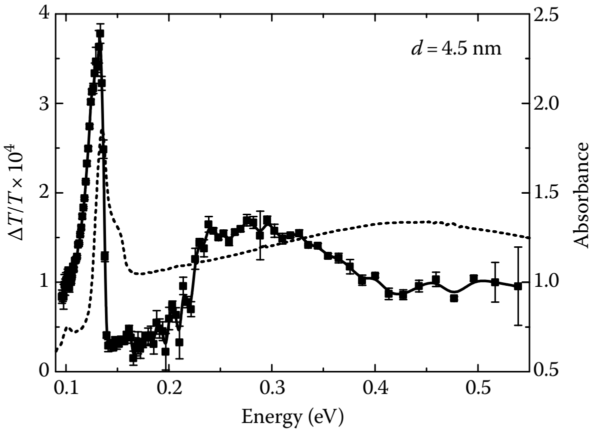

FIGURE 25.24 Room temperature PIA (symbols and a solid line) and linear IR absorption (dashed line) spectra for the sample with average diameter of 4.5 nm.

---

* Since our SiNcs are essentially undoped, the carriers must optically be induced. Once a technology to introduce dopants into SiNCs will emerge, it would be interesting to investigate direct ISL transitions from SiNCs having a single type of carriers.

† The linear and the photoinduced transmittance are given by $T = T_0 e^{-\alpha \cdot d}$ and $\Delta T = T_0[e^{-\alpha \cdot d} - e^{-(\alpha+\Delta\alpha)\cdot d}]$ where α is the linear absorption coefficient and $\Delta\alpha$ is the photoinduced absorption coefficient. Hence, we find $\mathrm{PIA} = \Delta T/T = 1 - e^{-\Delta\alpha \cdot d} \cong \Delta\alpha \cdot d$ where we have assumed that $\Delta\alpha \cdot d \ll 1$.

of 4.5 nm (x = 18%). The linear IR absorption spectrum includes a single absorption band, at about 130–150 meV (1050–1210 cm^{-1} in terms of wavenumbers), of the Si–O asymmetric stretching mode of vibration,[102,103] schematically illustrated in Figure 25.20. Similar linear IR absorption spectra have been obtained for all SiNCs embedded in SiO_2 matrices. The PIA spectrum reveals two photoinduced absorption bands; a relatively strong and narrow band at ~135 meV and a fairly broad, high-energy band at about 280 meV. The broadening of the PIA bands is expected here as absorption measurements average over all sizes of nanocrystals according to the size distribution of each sample (see Figure 25.5). In fact, the fairly narrow linewidth of the low-energy band (FWHM of ~20 meV) is quite surprising taking into account the size distribution. As the PIA technique is not sensitive to the polarity of the carriers, we may use the effective mass approximation for assigning the low-energy PIA band to valence-ISL (VISL) transitions and the high-energy band to conduction-ISL (CISL) optical transitions.[91–93] However, another explanation of assigning both bands to CISL transitions is possible, where the $1P_e$ state of the conduction splits into two sublevels, mainly due to differences between longitudinal and transversal conduction effective masses.[91] In the following, we follow the first scenario but the main conclusions of this experiment remain valid for the second scenario as well.

Next, in Figure 25.25, we show the PIA spectra for a series of SiNCs having variable average diameter, from 4 up to 6 nm. This is also the range of diameters where the two PIA bands could be resolved from the noise. Surprisingly, while the high-energy CISL absorption band shows a remarkable redshift with the increasing size of the nanocrystals, the low-energy VISL band presents a significantly weaker shift to the red, with the energy of the VISL transition (120–140 meV) being fairly close to the energy of the Si–O asymmetric stretching mode of vibration (~135 meV).

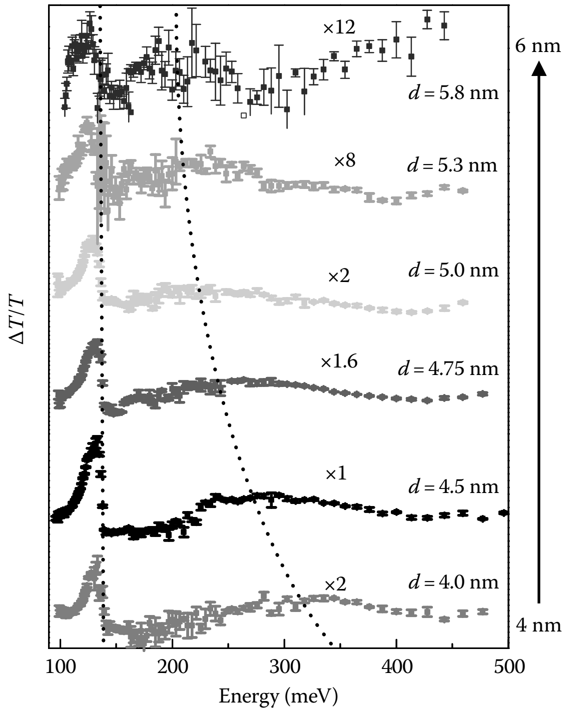

FIGURE 25.25 PIA spectra for selected samples of different average diameters from 4 up to 6 nm. The dashed lines indicate the remarkable redshift of the CISL band and the much weaker shift of the VISL band.

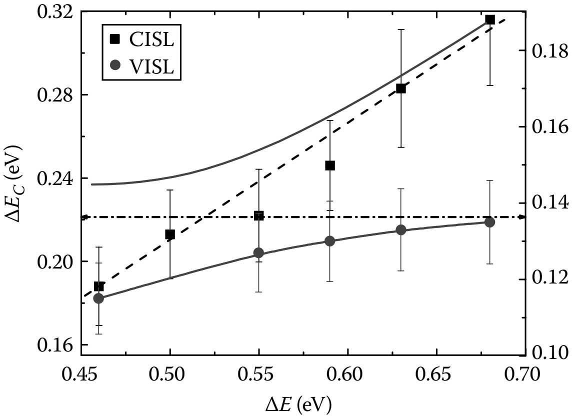

FIGURE 25.26 The measured ISL energies (left axis-CISL, right axis-VISL) versus the confinement energy. The horizontal dashed line indicates the energy of the Si–O asymmetric stretching mode of vibration.

To correlate the results with the vibron model, we present in Figure 25.26 the PIA peak energies for both ISL bands (ΔE_C and ΔE_V) versus the confinement energy, ΔE (defined here as $\Delta E = E_{PL} - E_G$ with E_{PL} being the peak PL energy of the sample). The CISL energies display a linear dependence on ΔE as expected from the QC model (where both ΔE and $\Delta E_C \sim 1/d^{\delta}$ according to Equation 25.3). On the other hand, the VISL energies show a sublinear dependence on ΔE, approaching the vibrational energy of the Si–O stretching mode of vibration. The dashed line in Figure 25.26 represents the expected dependence of the VISL energies on ΔE according to the QC model (i.e., $\Delta E_V \sim \Delta E$). This line crosses the vibrational energy of the Si–O vibration, which is independent of the size of the nanocrystals, and is represented by the horizontal dashed-dotted line in Figure 25.26. Hence, the anti-crossing of the VISL energies and the vibration energies, which resembles the anti-crossing shown in Figure 25.23 (the vibron model), suggests that electronic states of the valence band are resonantly coupled to Si–O surface vibrations to create vibrons.

Notice also that the amplitude (or the strength) of the PIA bands increases with the decreasing size of the nanocrystals, up to a diameter of ~4.5 nm and then, decreases until the PIA spectra could not be resolved for SiNCs smaller then 2.5 nm. These results provide additional support to the vibron model. The correlation between the integrated PIA (e.g., the area below the VISL absorption bands shown in Figure 25.25) and the integrated PL, which is presented in Figure 25.27, suggests that the same mechanism, e.g., the formation of vibrons, is also responsible for the high PL efficiency. For SiNCs having a diameter in the range of 4–4.5 nm, the energy difference between the lowest valence sublevels becomes comparable to those of the Si–O surface vibrations. As a result, long-lived vibrons are created giving rise to slow nonradiative relaxation rates that, in turn, enhance

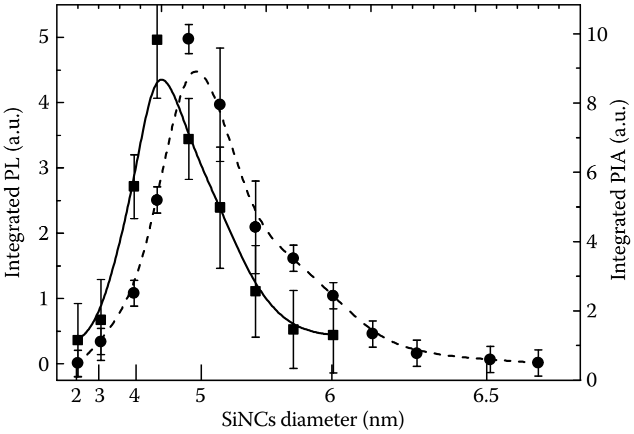

FIGURE 25.27 The correlation between the integrated PL (dashed line and circles) and the integrated PIA (solid line and squares) where both the PL and the PIA have a maximum in the range of 4.5–5 nm.

the PL efficiency. This mechanism, which takes into account the mutual role of QC and SC, also explains the correlation between the PIA and the PL shown in Figure 25.27. Finally, let us point out that the Si–O asymmetric vibration is related to a "longitudinal" mode of vibration where the oxygen atom vibrates back and forth along the line joining the axis of the two silicon atoms (see Figure 25.20). This vibrational mode is expected to generate a strong polarization field relative to other Si–O vibrations[103] such as the symmetric stretching mode (at about 810 cm^{-1}), which can be seen in the linear IR absorption spectrum but without any correlation to the PIA spectra, and the Si–O rocking mode of vibration at ~450 cm^{-1}. Notice that, it has been suggested that the asymmetric lineshape of the Si–O stretching mode at ~1100 cm^{-1} (see Figure 25.24) is due to the two "quasi-phonon" modes of the amorphous SiO_2 medium,[104] with the AS_1 (first asymmetric stretching) mode at 1076 cm^{-1} being related to "in-phase" motion of adjacent oxygen atoms while the AS_2 mode at ~1250 cm^{-1} is related to "out-of-phase" motion of adjacent oxygen atoms.[103] Clearly, only the AS_1 mode is expected to generate a long-range polarization field, a result that can explain the strong asymmetry of the VISL band (see Figure 25.24), once approaching the high-energy shoulder of the linear Si–O stretching vibration.

Additional experimental results supporting the vibron model come from a single quantum dot spectroscopy of SiNCs fabricated by the laser pyrolysis technique (see Section 25.2). Martin et al.[105] have succeeded to dissolve SiNCs (surrounded by SiO_2), disperse them in polymer matrices and to spin-cast the nanocrystals on glass substrates for a single silicon nanoparticle spectroscopy. The PL spectra, shown in Figure 25.28 for two different sizes of nanocrystals, reveal vibronic features of low-energy PL satellites, at about 130–160 meV below the high-energy peaks. These satellite peaks have been interpreted as (surface) phonon-assisted optical transitions where the high-energy PL band is related to a zero-phonon transition (having homogeneous linewidth of ~100 meV) and a low-energy PL shoulder that is related to a phonon-assisted transition. Notice that phonon-assisted optical transitions (both interband[106] and ISL[58]) have been predicted for SiNCs; however, in most cases the bulk phonons of silicon have been considered rather than polar surface vibrations (which could be either Si–O surface vibrations, or AS phonons of amorphous-SiO_2 or LO phonons of a crystalline SiO_2 shell). A similar phenomenon of vibronic satellites in the PL spectrum has also been reported for PS particles.[107] In another work, vibronic features due to surface-oxide absorption band using ellipsometric measurements[108] have been reported. Seraphin et al.[109] have shown that modifying the surface passivation of SiNCs does not alter the PL spectrum and only the PL intensity is changed, apparently due to modification of the surface bonds with a higher PL intensity associated with the more polar surface bonds.

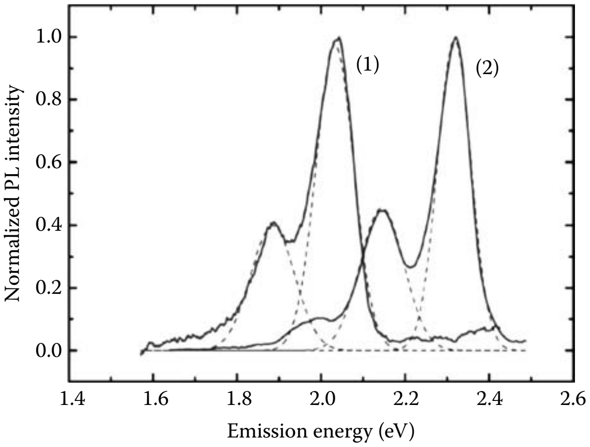

FIGURE 25.28 PL spectra of two, single SiNCs prepared by the laser pyrolysis techniques with vibronic features (low-energy shoulders). The dotted lines represent fitting to a double Gaussian lineshape. (Reprinted from Martin, J. et al., *Nano Lett.*, 8, 656, 2008. With permission.)

Finally, let us discuss the recent magneto-PL experiment reported by Godefroo et al.[110] In this experiment, pulsed magnetic fields up to 50 T have been applied to SiNCs and the corresponding energy shift of the PL has been measured. The high magnetic field acts to further confine the exciton with $\ell_B = \sqrt{\hbar/eB}$ being the magnetic confinement length. Hence, free excitons of bulk nanocrystals are expected to show a blueshift in response to high enough B-fields, while localized or trapped excitons are not expected to show a similar shift. Indeed, no blueshift of the PL has been measured for SiNCs embedded in SiO_2. The authors have interpreted these observations as a PL originated from defect centers in the oxide layer that wraps the nanocrystals. However, according to the vibron model, surface-localized vibrons should not present a blueshift either. In another experiment, the authors annealed the samples in hydrogen, at temperatures low enough to avoid the creation of hydrogenated a:Si. In this experiment a blueshift of the PL has been observed, thus confirming the experimental results of Wolkin et al.[17] where H-terminated SiNCs have shown monotonic blueshift of the PL with the decreasing size of the nanocrystals (or with the increasing B-field in the case of the Godefroo experiment[110]). The topic of hydrogen-terminated SiNCs remains a puzzle that has not been addressed in this chapter for the very simple reason that no reliable methods to produce stable, H-terminated SiNCs have been reported so far. Yet, the Si–H bonds are expected to be considerably more polar than the Si–O bonds. Hydrogen atoms are highly mobile ions that can easily penetrate over few nanometers into the nanocrystalline silicon core. Even small amount of hydrogen in the silicon core is expected to produce considerable polarization field of the "core" nanocrystal. This topic calls for further investigation as the current knowledge about H-terminated SiNCs is insufficient for drawing any reliable conclusions.

25.5 Concluding Remarks

This chapter has covered one specific topic from the much broader, rapidly expanding field of silicon nanostructures and their applications, focusing on the question about the origin of the PL and the specific role of QC and SC in the eluding game played by the luminescence. Yet, this specific question is an excellent example to a situation where *small is really different*. Namely, not a property of matter that scales down in proportion to the size of the material, but rather a truly new phenomenon that appears at, and only at the nanometric length scales. The PL from silicon nanostructures appears to be such a property. The role of QC and SC is another fascinating example to the case where *physics meets chemistry* and both disciplines of science merge together into one, multidisciplinary field of nanoscience.

The questions regarding the origin of the PL from silicon nanostructures are far from being fully resolved. Many open questions still exist, some have been discussed in this chapter, some did not, and most likely, some new fundamental questions and puzzles will emerge soon. Yet, the basic direction in attempting to merge QC and SC, rather than choosing one as a sole winner, seems to be the proper approach. It is just a matter of time to see if this subjective opinion of the author of this chpater is indeed correct or not.

Acknowledgments

The author would like to thank many of his (former and current) graduate students at the Hebrew University of Jerusalem for invaluable contributions to this work, particularly, M. Dovrat, D. Krapf, Y. Shalibo, N. Arad, Y. Oppenheim-Goshen, and Y. Raichman. I wish to acknowledge the fruitful collaboration with my colleagues at the Hebrew University, including J. Jedrzejewski, J. Shappir, and especially Prof. Isaac Balberg for stimulating and helpful discussions and for critical reading of the manuscript. I wish to acknowledge the technical staff of the unit for nano-characterization (UNC) at the Hebrew University of Jerusalem, particularly Dr. Inna Popov for her assistance with the TEM measurements. I have enjoyed a fruitful and productive collaboration with X.-H. Zhang from the Chinese Academy of Sciences in Beijing and Prof. S.-T. Lee from the City University of Honk Kong during a joint project on silicon nanowires. This work has been supported by the Israel Science Foundation (ISF), the Israeli Ministry of Science, the binational Chinese–Israeli research grant provided by the ministries of science of both countries.

References

1. S. M. Sze, *Physics of Semiconductors*, 2nd edn., Wiley-Interscience, New York (1981).
2. Al. L. Efros and M. Rosen, The electronic structure of semiconductor nanocrystals, *Annu. Rev. Mater. Sci.* **30**, 475–521 (2000).
3. P. Moriarty, Nanostructured materials, *Rep. Prog. Phys.* **64**, 297–381 (2001).
4. C. J. Murphy and J. L. Coffer, Quantum dots: A primer, *Appl. Spectrosc.* **56**, 16A–27A (2002).
5. L. Pavesi, Will silicon be the photonic material of the third millennium? *J. Phys.: Condens. Matter* **15**, R1169–R1196 (2003).
6. A. G. Nassiopoulou, Silicon nanocrystals in SiO_2 thin layers, in *Encyclopedia of Nanoscience and Nanotechnology*, Vol. 9, H. S. Nalwa, Ed., pp. 793–813, American Scientific, Stevenson Ranch, CA (2004).
7. L.T. Canham, Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers, *Appl. Phys. Lett.* **57**, 1046–1048 (1990).
8. A. G. Cullis and L. T. Canham, Visible light emission due to quantum size effects in highly porous crystalline silicon, *Nature* **353**, 335–338 (1991).
9. D. Kovalev, H. Heckler, G. Polisski, and F. Koch, Optical properties of Si nanocrystals, *Phys. Stat. Sol. (b)* **215**, 871–932 (1999).
10. G. Amato, C. Delerue, and H.-J. von Bardeleben, Eds., Structural and optical properties of porous silicon nanostructures, *Optoelectronic Properties of Semiconductor and Superlattices*, Vol. **5**, Gordon & Breach, Amsterdam, the Netherlands (1997) and references therein.
11. A. G. Cullis, L. T. Canham, and P. D. J. Calcott, The structural and luminescence properties of porous silicon, *J. Appl. Phys.* **82**, 909–965 (1997).
12. O. Bisi, S. Ossicini, and L. Pavesi, Porous silicon: A quantum sponge structure for silicon based optoelectronics, *Surf. Sci. Rep.* **38**, 1–126 (2000).
13. J. Gole and D. A. Dixon, Potential role of silanones in the photoluminescence-excitation, visible-photoluminescence-emission, and infrared spectra of porous silicon, *Phys. Rev. B.* **57**, 12002–12016 (1998).
14. J. J. Heitmann, F. Müller, M. Zacharias, and U. Gösele, Silicon nanocrystals: Size matters, *Adv. Mater.* **17**, 795–803 (2005).
15. M. Dovrat, Y. Oppenheim, J. Jedrzejewski, I. Balberg, and A. Sa'ar, Radiative versus nonradiative decay processes in silicon nanocrystals probed by time-resolved photoluminescence spectroscopy, *Phys. Rev. B.* **69**, 155311 (2004).
16. M. Dovrat, Y. Goshen, I. Popov, J. Jedrzejewski, I. Balberg, and A. Sa'ar, The role of radiative and nonradiative relaxation processes in the generation of light from silicon nanocrystals, *Phys. Stat. Sol. (c)* **2**, 3440–3444 (2005).
17. M. V. Wolkin, J. Jorne, P. M. Fauchet, G. Allan, and C. Delerue, Electronic states and luminescence in porous silicon quantum dots: The role of oxygen, *Phys. Rev. Lett.* **82**, 197–200 (1999).

18. A. Sa'ar, M. Dovrat, J. Jedrzejewski, and I. Balberg, Optical inter- and intra-band transitions in silicon nanocrystals: The role of surface vibrations, *Phys. E: Low-Dimensi. Syst. Nanostruct.* **38**, 122–127 (2007).
19. A. Sa'ar, Photoluminescence from silicon nanostructures: The mutual role of quantum confinement and surface chemistry, *J. Nanophoton* **3**, 032501 (2009).
20. A. Givant, J. Shappir, and A. Sa'ar, Photoluminescence anisotropy from laterally anodized porous silicon, *Appl. Phys. Lett.* **73**, 3150–3152 (1998).
21. B. Urbach, E. Axelrod, and A. Sa'ar, Correlation between transport, dielectric, and optical properties of oxidized and nonoxidized porous silicon, *Phys. Rev. B* **75**, 205330 (2007).
22. V. Lehmann, *Electrochemistry of Silicon*, Wiley-VCH, Weinheim, Germany (2002).
23. W. Theiβ, Optical properties of porous silicon, *Surf. Sci. Rep.* **29**, 91–192 (1997).
24. E. Sabatani, Y. Kalisky, A. Berman, Y. Golan, N. Gutman, B. Urbach, and A. Sa'ar, Photoluminescence of polydiacetylene membranes on porous silicon utilized for chemical sensors, *Opt. Mater.* **38**, 1766–1774 (2008).
25. B. Urbach, N. Korbakov, Y. Bar-David, S. Yitzchaik, and A. Sa'ar, Composite structures of polyaniline and mesoporous silicon: Electrochemistry, optical and transport properties, *J. Phys. Chem. C* **111**, 16586–16592 (2007).
26. S. Ben-Tabou de-Leon, R. Oren, M. E. Spira, S. Yitzchaik, and A. Sa'ar, Neurons culturing and biophotonic sensing using porous silicon, *Appl. Phys. Lett.* **84**, 4361–4363 (2004).
27. H. Lin, T. Gao, J. Fantini, and M. J. Sailor, A porous silicon–palladium composite film for optical interferometric sensing of Hydrogen, *Langmuir* **20**, 5104–5108 (2004).
28. S. Furukawa and T. Miyasato, Three-dimensional quantum well effects in ultrafine silicon particles, *Jpn. J. Appl. Phys.* **27**, L2207–L2209 (1988).
29. W. S. Cheong, N. M. Hwang, and D. Y. Yoon, Observation of nanometer silicon clusters in the hot-filament CVD process, *J. Cryst. Growth* **204**, 52–61 (1999).
30. S. Charvet, R. Madelon, R. Rizk, B. Garrido, O. Gonzàlez-Varona, M. López, A. Pérez-Rodríguez, and J. R. Morante, Substrate temperature dependence of the photoluminescence efficiency of co-sputtered Si/SiO_2 layers, *J. Lumin.* **80**, 241–245 (1988).
31. U. Kahler and H. Hofmaeister, Visible light emission from Si nanocrystalline composites via reactive evaporation of SiO, *Opt. Mater.* **17**, 83–86 (2001).
32. A. J. Kenyon, P. F. Trwoga, C. W. Pitt, and G. Rehm, Luminescence efficiency measurements of silicon nanoclusters, *Appl. Phys. Lett.* **73**, 523–525 (1998).
33. Z.-X. Ma, X.-B. Liao, J. He, W. C. Cheng, G. Z. Yue, Y. Q. Wang, and G. L. Kong, Annealing behaviors of photoluminescence from SiO_x:H, *J. Appl. Phys.* **83**, 7934–7939 (1998).
34. L. Ferraioli, M. Wang, G. Pucker, D. Navarro-Urrios, N. Dadosso, C. Kompocholis, and L. Pavesi, Photoluminescence of silicon nanocrystals in silicon oxide, *J. Nanomater.* **2007**, 43491 (2007).
35. T. Shimizu-Iwayama, S. Nakao, and K. Saitoh, Optical and structural characterization of implanted nanocrystalline semiconductors, *Nucl. Instrum. Methods B* **121**, 450–454 (1997).
36. K. S. Min, K. V. Scheglov, C. M. Yang, H. A. Atwater, M. L. Brongersma, and A. Polman, Defect-related versus excitonic visible light emission from ion beam synthesized Si nanocrystals in SiO_2, *Appl. Phys. Lett.* **69**, 2033–2035 (1996).
37. J. Linnros, N. Lalic, A. Galeckas, and V. Grivickas, Analysis of the stretched exponential photoluminescence decay from nanometer-sized silicon crystals in SiO_2, *J. Appl. Phys.* **86**, 6128–6134 (1999).
38. Y. Posada, I. Balberg, L. F. Fonseca, O. Resto, and S. Z. Wiesz, Diffusion length measurements of minority carriers in Si-SiO_2 using the photo-grating technique, in *Microcrystalline and Nanocrystalline Semiconductors*—2000, Boston, MA, Materials Research Society Symposia Proceedings, No. 638, F.14.44, pp. 1–6, M. Fauchet, J. M. Buriak, L.T. Canham, N. Koshida, and B. E. White, Jr. Eds., MRS, Warrendale, PA (2001).
39. I. Balberg, E. Savir, J. Jedrzeijewski, A. G. Nassiopoulou, and S. Gardelis, Fundamental transport processes in ensembles of silicon quantum dots, *Phys. Rev. B* **75**, 235329 (2007).
40. I. Balberg, E. Savir, and J. Jedrzeijewski, The mutual exclusion of luminescence and transport in nanocrystalline silicon networks, *J. Non-Cryst. Solids* **338–340**, 102–105 (2004).
41. G. Ledoux, J. Gong, F. Huisken, O. Guillois, and C. Reynaud, Photoluminescence of size-separated silicon nanocrystals: Confirmation of quantum confinement, *Appl. Phys. Lett.* **80**, 4834–4836 (2002).
42. D. J. Lockwood, Z. H. Lu, and J.-M. Baribeau, Quantum confined luminescence in Si/SiO_2 superlattices, *Phys. Rev. Lett.* **76**, 539–541 (1996).
43. V. Vinciguerra, G. Franzò, F. Priolo, F. Iacona, and C. Spinella, Quantum confinement and recombination dynamics in silicon nanocrystals embedded in Si/SiO_2 superlattices, *J. Appl. Phys.* **87**, 8165–8173 (2000).
44. Z. Ma, L. Wang, K. Chen, W. Li, L. Zhang, Y. Bao, X. Wang, J. Xu, X. Huang, and D. Feng, Blue light emission in nc-Si/SiO_2 multilayers fabricated using layer by layer plasma oxidation, *J. Non-Cryst. Solids* **299–302**, 648–652 (2002).
45. P. Photopooulos, A. G. Nassiopoulou, D. N. Kouvatsos, and A. Travlos, Photo- and electroluminescence from nanocrystalline silicon single and multilayer structures, *Mater. Sci. Eng. B* **69–70**, 345–349 (2000).
46. L. Tsybeskov, K. D. Hirschman, S. P. Duttagupta, M. Zacharias, P. M. Fauchet, J. P. McCaffrey, and D. J. Lockwood, Nanocrystalline-silicon superlattice produced by controlled recrystallization, *Appl. Phys. Lett.* **72**, 43–45 (1998).
47. M. Zacharias, J. Heitmann, R. Scholz, U. Kahler, M. Schmidt, and J. Bläsing, Size-controlled highly luminescent silicon nanocrystals: A SiO/SiO_2 superlattice approach, *Appl. Phys. Lett.* **80**, 661–663 (2002).

48. M. Dovrat, Y. Shalibo, N. Arad, S.-T. Lee, and A. Sa'ar, Fine structure and selection rules for excitonic transitions in silicon nanostructures, *Phys. Rev. B* **79**, 125306 (2009).
49. N. G. Shang, U. Vetter, I. Gerhards, H. Hofsäss, C. Ronning, and M. Seibt, Luminescence centres in silica nanowires, *Nanotechnology* **17**, 3215–3218 (2006).
50. K. Sato and K. Hirakuri, Influence of paramagnetic defects on multicolored luminescence from nanocrystalline silicon, *J. Appl. Phys.* **100**, 114303 (2006).
51. M. Dovrat, N. Arad, X.-H. Zhang, S.-T. Lee, and A. Sa'ar, Optical properties of silicon nanowires from cathodoluminescence imaging and time-resolved photoluminescence spectroscopy, *Phys. Rev. B* **75**, 205343 (2007).
52. M. Dovrat, N, Arad, S. T. Lee, and A. Sa'ar, Cathodoluminescence and photoluminescence of individual silicon nanowires, *Phys. Stat. Sol. (a)* **204**, 1512–1517 (2007).
53. C. Meier, A. Gondorf, S. Lüttjohann, A. Lorke, and H. Wiggers, Silicon nanoparticles: Absorption, emission, and the nature of the electronic bandgap, *J. Appl. Phys.* **101**, 103112 (2007).
54. J. A. Moreno, B. Garrido, P. Pellegrini, C. Garcia, J. Arbiol, J. R. Morante, P. Marie, F. Gourbilleau, and R. Rizk, Size dependence of refractive index of Si nanoclusters embedded in SiO_2, *J. Appl. Phys.* **98**, 013523 (2005).
55. M. Mansour, A. En Naciri, L. Johann, J. J. Grob, and M. Stchakovsky, Dielectric function and optical transitions of silicon nanocrystals between 0.6 eV and 6.5 eV, *Phys. Stat. Sol. (a)* **205**, 845–848 (2008).
56. A. En Naciri, M. Mansour, L. Johann, J. J. Grob, and H. Rinnert, Influence of the implantation profiles of Si^+ on the dielectric function and optical transitions in silicon nanocrystals, *J. Chem. Phys.* **129**, 184701 (2008).
57. C. Cimpean, V. Groenewegen, V. Kuntermann, A. Sommer, and C. Kryschi, Ultrafast exciton relaxation dynamics in silicon quantum dots, *Laser Photonics Rev.* **3**, 138–145 (2009).
58. C. Delerue, G. Allen, and M. Lannoo, Theoretical aspects of the luminescence of porous silicon, *Phys. Rev. B* **48**, 11024–11036 (1993).
59. A. Zunger and L. W. Wang, Theory of silicon nanostructures, *Appl. Surf. Sci.* **102**, 350–359 (1996).
60. J. Valenta, R. Juhasz, and J. Linnros, Photoluminescence spectroscopy of single silicon quantum dots, *Appl. Phys. Lett.* **80**, 1070–1072 (2002).
61. J. D. Holmes, K. J. Ziegler, C. Doty, L. E. Pell, K. P. Johnston, and B. A. Korgel, Highly luminescent silicon nanocrystals with discrete optical transitions, *J. Am. Chem. Soc.* **123**, 3743–3748 (2001).
62. R. M. Sankaran, D. Holunga, R. C. Flagan, and K. P. Giapis, Synthesis of Blue luminescent Si nanoparticles using atmospheric-pressure microdischarges, *Nano Lett.* **5**, 537–541 (2005).
63. R. J. Walters, J. Kalkman, A. Polman, H. A. Atwater, and M. J. A. de Dood, Photoluminescence quantum efficiency of dense silicon nanocrystal ensembles in SiO_2, *Phys. Rev. B* **73**, 132302 (2006).
64. T. Suemoto, K. Tanaka, and A. Nakajima, Interpretation of the temperature dependence of the luminescence intensity, lifetime, and decay profiles in porous Si, *Phys. Rev. B* **49**, 11005–1109 (1994).
65. D. Kovalev, H. Heckler, M. Ben-Chorin, G. Polisski, M. Schwartzkopff, and F. Koch, Breakdown of the *k*-conservation rule in Si nanocrystals, *Phys. Rev. Lett.* **81**, 2803–2806 (1998).
66. T. Takagahara and K. Takeda, Theory of the quantum confinement effect on excitons in quantum dots of indirect-gap materials, *Phys. Rev. B* **46**, 15578–15581 (1992).
67. H. Koyama and N. Koshida, Photo-assisted tuning of luminescence from porous silicon, *J. Appl. Phys.* **74**, 6365–6367 (1993).
68. F. Herman and R. V. Kasowski, Electronic structure of defects at Si/SiO_2 interfaces, *J. Vac. Sci. Technol.* **19**, 395–401 (1981).
69. M. L. Brongersma, P. G. Kik, A. Polman, K. S. Min, and H. A. Atwater, Size-dependent electron-hole exchange interaction in Si nanocrystals, *Appl. Phys. Lett.* **76**, 351–353 (2000).
70. X. Wen, L.V. Dao, P. Hannaford, E.-C. Cho, Y. H. Cho, and M. A. Green, Excitation dependence of photoluminescence in silicon quantum dots, *New J. Phys.* **9**, 337 (2007).
71. L. Pavesi, Influence of dispersive exciton motion on the recombination dynamics in porous silicon, *J. Appl. Phys.* **80**, 216–225 (1996).
72. J. Heitmann, F. Müller, L. Yi, M. Zacharias, D. Kovalev, and F. Eichhorn, Excitons in Si nanocrystals: Confinement and migration effects, *Phys. Rev. B* **69**, 195309 (2004).
73. I. Balberg, Electrical transport mechanisms in ensembles of silicon quantum dots, *Phys. Stat. Sol. (c)* **5**, 3771–3775 (2008).
74. M. Ben-Chorin, F. Müller, F. Koch, W. Schirmacher, and M. Eberhard, Hopping transport on a fractal: ac conductivity of porous silicon, *Phys. Rev. B* **51**, 2199–2213 (1995).
75. P. D. Calcott, K. J. Nash, L. T. Canham, M. J. Kane, and D. Brumhead, Identification of radiative transitions in highly porous silicon, *J. Phys. Condens. Matter* **5**, L91–L98 (1993).
76. J. M. Ziman, *Principles of the Theory of Solids*, 2nd edn., Cambridge University Press, Cambridge, U.K. (1972).
77. A. Ourmazd, D. W. Taylor, A. Rentschler, and J. Bevk, $Si \rightarrow SiO_2$ transformation: Interfacial structure and mechanism, *Phys. Rev. Lett.* **59**, 213–216 (1987).
78. H. Kageshima, M. Uematsu, K. Akagi, S. Tsuneyuki, T. Akiyama, and K. Shiraishi, Theoretical study on atomic structures of thermally grown silicon oxide/silicon interfaces, *e-J. Surf. Sci. Nanotechnol.* **4**, 584–587 (2006).
79. V. V. Afanas'ev, A. Stesmans, and M. E. Twigg, Epitaxial growth of SiO_2 produced in silicon by oxygen ion implantation, *Phys. Rev. Lett.* **77**, 4206–4309 (1996).
80. C. Kaneta, T. Yamasaki, T. Uchiyama, T. Uda, and K. Kiyoyuki, Structure and electronic property of Si(100)/SiO_2 interface, *Microelectron. Eng.* **48**, 117–120 (1999).
81. T. Shimura, H. Misaki, M. Umeno, I. Takahashi, and J. Harada, X-ray diffraction evidence for the existence of epitaxial microcrystallites in thermally oxidized SiO_2 thin films on Si(111) surfaces, *J. Cryst. Growth* **166**, 786–791 (1995).

82. W. A. Dollase, The crystal structure at 220°C of orthorhombic high tridymite from the Steinbach meteorite, *Acta Crystallogr.* **23**, 617–623 (1967).
83. E.-C. Cho, M. A. Green, J. Xia, R. Corkish, and A. Nikulin, Atomistic structure of SiO_2/Si/SiO_2 quantum wells with an apparently crystalline silicon oxide, *J. Appl. Phys.* **96**, 3211–3216 (2004).
84. N. Pauc, V. Calvo, J. Eymery, F. Fournel, and N. Magnea, Electronic and optical properties of Si/SiO_2 nanostructures. II. Electron-hole recombination at the Si/SiO_2 quantum-well–quantum-dot transition, *Phys. Rev. B* **72**, 205325 (2005).
85. H. Kageshima and K. Shiraishi, Microscopic mechanism for SiO_2/Si interface passivation: Si=O double bond formation, *Surf. Sci.* **380**, 61–65 (1997).
86. N. Daldosso, M. Luppi, S. Ossicini, E. Degoli, R. Magri, G. Dalba, P. Fornasini, R. Grisenti, F. Rocca, L. Pavesi, S. Boninelli, F. Priolo, C. Spinella, and F. Iacona, Role of the interface region on the optoelectronic properties of silicon nanocrystals embedded in SiO_2, *Phys. Rev. B.* **68**, 085327 (2003).
87. A A. Sa'ar, Y. Reichman, M. Dovrat, D. Krapf, J. Jedrzejewski, and I. Balberg, Resonant coupling between surface vibrations and electronic states in silicon nanocrystals at the strong confinement regime, *Nano Lett.* **5**, 2443–2447 (2005).
88. A. Sa'ar, M. Dovrat, J. Jedrzejewski, I. Popov, and I. Balberg, The role of quantum confinement and surface chemistry in silicon nanocrystals at the strong confinement regime, *Phys. Stat. Sol. (a)* **204**, 1491–1496 (2007).
89. N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders, Philadelphia, PA (1976).
90. J. Singh, *Physics of Semiconductors and Their Heterostructures*, McGraw-Hill, New York (1993).
91. G. Allen and C. Delerue, Efficient intraband optical transitions in Si nanocrystals, *Phys. Rev. B* **66**, 233303 (2002).
92. A. S. Moskalenko, J. Berakdar, A. A. Prokofiev, and I. N. Yassievich, Single-particle states in spherical Si/SiO_2 quantum dots, *Phys. Rev. B* **76**, 085427 (2007).
93. J. S. de Sousa, J.-P. Leburton, V. N. Freire, and E. F. da Silva, Intraband absorption in silicon nanocrystals: The combined effect of shape and crystal orientation, *Appl. Phys. Lett.* **87**, 031913 (2005).
94. O. O. Verzelen, R. Ferreira, and G. Bastard, Excitonic Polarons in semiconductor quantum dots, *Phys. Rev. Lett.* **88**, 146803 (2002).
95. M. Mahdouani, R. Bourguiga, and S. Jaziri, Polaronic states in Si nanocrystals embedded in SiO_2 matrix, *Physica E* **41**, 228–234 (2008).
96. G. D. Mahan, *Many-Particle Physics*, 3rd edn., Kluwer, New York (2000).
97. K. Hess and P. Vogl, Remote polar phonon scattering in silicon inversion layers, *Solid State Commun.* **30**, 797–799 (1979).
98. J.-Q. Lü and F. Koch, Polar phonon scattering at the Si-SiO_2 interface, *Microelectron. Eng.* **48**, 95–99 (1999).
99. For general reviews see: NATO ASI Series on *Quantum Well Intersubband Transitions: Physics and Devices*, E. Rosencher, B. Vinter, and B. Levine, Eds., *Series B: Physics*, Vol. **288**, Plenum Press, New York (1992); (b) *Series E: Applied Sciences*, Vol. **270**, Kluwer, Dordrecht, the Netherlands (1994); (c) *Intersubband Transitions in Quantum Wells: Physics and Devices*, S. S. Li and Yan-Kuin Su, Eds., Kluwer Academic, Boston, MA (1998).
100. A. Sa'ar, S. Calderon, A. Givant, O. Ben-Shalom, E. Kapon, and C. Caneau, Energy subbands, envelope states, and intersubband optical transitions in one-dimensional quantum wires: The local-envelope-states approach, *Phys. Rev. B* **54**, 2675–2684 (1996).
101. D. Krapf, S. H. Kan, U. Banin, O. Millo, and A. Sa'ar, Intersublevel optical transitions in InAs nanocrystals probed by photoinduced absorption spectroscopy: The role of thermal activation, *Phys. Rev. B* **69**, 073301 (2004).
102. H. J. Hrostowski and R. H. Kaiser, Infrared absorption of oxygen in silicon, *Phys. Rev.* **107**, 966–972 (1957).
103. C. T. Kirk, Quantitative analysis of the effect of disorder-induced mode coupling on infrared absorption in silica, *Phys. Rev. B* **38**, 1255–1273 (1988).
104. F. L. Galeener and G. Lucovsky, Longitudinal optical vibrations in glasses: GeO_2 and SiO_2, *Phys. Rev. Lett.* **37**, 1474–1478 (1976).
105. J. Martin, F. Cichos, F. Huisken, and C. von Borczyskowski, Electron–phonon coupling and localization of excitons in single silicon nanocrystals, *Nano Lett.* **8**, 656–660 (2008).
106. C. Delerue, G. Allan, and M. Lannoo, Electron-phonon coupling and optical transitions for indirect-gap semiconductor nanocrystals, *Phys. Rev. B* **64**, 193402 (2001).
107. M. D. Mason, G. M. Credo, K. D. Weston, and S. K. Buratto, Luminescence of individual porous Si chromophores, *Phys. Rev. Lett.* **80**, 5405–5408 (1998).
108. E. Lioudakis, A. Antoniou, A. Othonos, C. Christofides, A.G. Nassiopoulou, Ch.B. Lioutas, and N. Frangis, The role of surface vibrations and quantum confinement effect to the optical properties of very thin nanocrystalline silicon films, *J. Appl. Phys.* **102**, 083534 (2007).
109. A. A. Seraphin, S.-T. Ngiam, and K. D. Kolenbrander, Surface control of luminescence in silicon nanoparticles, *J. Appl. Phys.* **80**, 6429–6433 (1996).
110. S. Godefroo, M. Hayne, M. Jivanescu, A. Stesmans, M. Zacharias, O. I. Lebedev, G. Van Tendeloo, and V. V. Moshchalkov, Classification and control of the origin of photoluminescence from Si nanocrystals, *Nat. Nanotechnol.* **3**, 174–178 (2008).

26

Polarization-Sensitive Nanowire and Nanorod Optics

26.1 Introduction **26**-1
26.2 Single Nanowires and Nanorods **26**-1
Polarization of Absorption: General Concepts • Metal Nanostructures: Plasmon Phenomena • Semiconductor Nanostructures: Polarization of Luminescence • Nonlinear Phenomena in Metal and Semiconductor Nanowires • Core–Shell Nanowires
26.3 Arrays of Nanorods **26**-13
Polarization Memory in Random Nanorod Arrays • Plasmon Spectra in Self-Assembled Nanorod Structures
26.4 Conclusion **26**-17
Acknowledgments **26**-17
References **26**-18

Harry E. Ruda
University of Toronto

Alexander Shik
University of Toronto

26.1 Introduction

Our goal is to present a large group of optical phenomena observed in anisotropic nanostructures such as nanowires (NWs) and nanorods (NRs) placed in an environment with a dielectric constant ε_0 different from that of the nanostructure ε. Actually, this situation occurs for most experimental work on NWs and NRs. For instance, in recent years, a large number of works have emerged devoted to the fabrication and the investigation of NWs grown in a free-standing fashion from catalyst particle nano-islands (Duan and Lieber 2000), or in a dielectric template with cylindrical pores (Mei et al. 2003), as well as for solution-grown NRs (Peng et al. 2000). Since the dielectric constant of the semiconductor ε is usually of the order of 10 or larger, this situation is realized for NWs and NRs free-standing in air, dissolved in liquids, or embedded in most dielectric matrices. Such effects of dielectric mismatch are observed not only for the semiconductor but also for metal nanostructures, where ε also differs noticeably from ε_0 and, besides, has a complex character and a strong frequency dispersion.

Due to the dielectric mismatch, the optical electric field in the vicinity of nanostructures is dramatically distorted by polarization effects (image forces), and the field intensity inside nanostructures, and hence all light-induced phenomena (optical absorption, luminescence, photoconductivity, etc.), strongly depend on the orientation of NWs and NRs related to the light polarization. Similarly, light emitted by such nanostructures is also strongly polarized.

We emphasize that the effects considered have a pure electrodynamic nature. They are not related to the presence or absence of size quantization and other possible size phenomena in nanostructures. The effects have a very general character and, as will be shown later, the only condition for their presence, besides the dielectric mismatch with the environment, is the smallness of the NW and the NR diameter in comparison with the light wavelength, which is well satisfied in most semiconductor and metal nanostructures.

26.2 Single Nanowires and Nanorods

26.2.1 Polarization of Absorption: General Concepts

26.2.1.1 Thin Nanowires and Nanorods

It is well known that at the interface of two media with different dielectric constants, the parallel component of electric field **E**, and the normal component of electric induction $\varepsilon\mathbf{E}$ must be continuous. If we have a cylindrical NW in the external field $\mathbf{E_0}$, this means that its parallel component, $E_\parallel$, remains the same inside an NW: $E_\parallel = E_{0\parallel}$. On the contrary, the field normal to the NW axis, $E_\perp$, is different inside and outside the NW. For a planar interface, we would have $E_\perp = \varepsilon_0 E_{0\perp}/\varepsilon$ but for a cylinder where $E_\perp$ contains components both parallel and perpendicular to the interface, the expression is more complicated (Landau and Lifshitz 1984):

$$E_\perp = \frac{2\varepsilon_0}{\varepsilon + \varepsilon_0} E_{0\perp}. \tag{26.1}$$

If, instead of a long (in theory—infinitely long) NW, we have an NR, which can be considered as a prolate ellipsoid of revolution

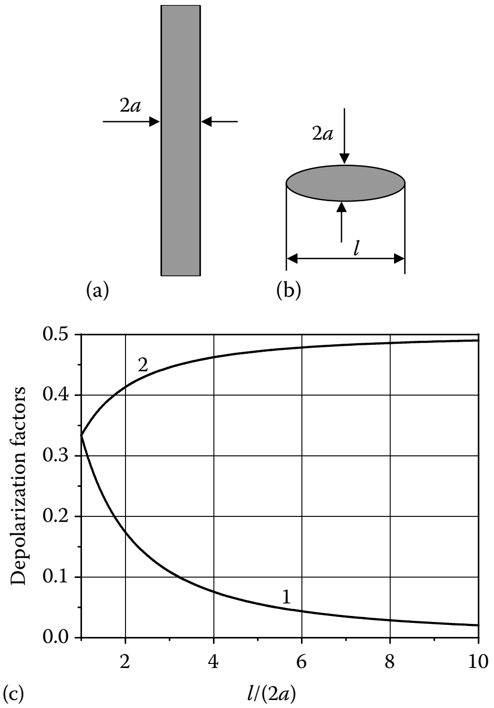

FIGURE 26.1 Schematic view of an NW (a) and an NR (b). Depolarization factors $n_{\|}$ (curve 1) and $n_{\perp}$ (curve 2) for an NR (c).

with semi-axes a and $l/2$ ($a < l/2$) (Figure 26.1b), the internal field is expressed in terms of the so-called depolarization factors $n_{\|,\perp}$:

$$E_{\|,\perp} = \frac{\varepsilon_0}{\varepsilon_0 + (\varepsilon - \varepsilon_0)n_{\|,\perp}} E_{0\|,\perp}. \tag{26.2}$$

These factors depend only on the aspect ratio a/l (see Figure 26.1c) but not on the absolute dimensions and material of NRs and satisfy the conditions $n_{\|} + 2n_{\perp} = 1$ and $n_{\|} < n_{\perp}$, so that for $\varepsilon > \varepsilon_0$ the internal field, as in NWs, is maximal for the NR axis parallel to the electric field of light. For infinitely long NWs, $n_{\|} = 0$, $n_{\perp} = 1/2$, which gives us $E_{\|} = E_{0\|}$ and Equation 26.1. For simplicity, most of the results below are presented for the case of NWs but can be easily generalized to NRs by repeating the same calculations starting from Equation 26.2, rather than Equation 26.1.

For $\varepsilon > \varepsilon_0$, which is always the case for semiconductor nanostructures, Equations 26.1 and 26.2 have two important consequences:

1. The amplitude of the high-frequency electric field and, hence, the probability of optical transitions in a nanostructure depend dramatically on the light polarization, acquiring maximal values for light with a polarization parallel to its axis.
2. Illumination with unpolarized light creates a high-frequency electric field in the nanostructure with the predominant direction of the polarization parallel to its axis.

The practical consequence of the first effect is a strong dependence of the absorption coefficient α (both inter- and intraband) on the light polarization. The ratio of α in NWs for the two light polarizations is

$$\frac{\alpha_{\|}}{\alpha_{\perp}} = \left| \frac{\varepsilon + \varepsilon_0}{2\varepsilon_0} \right|^2. \tag{26.3}$$

For most semiconductors, this factor exceeds 30 for free-standing NWs ($\varepsilon_0 = 1$) and 4.7 for NWs in the Al_2O_3 matrix ($\varepsilon_0 \cong 3$ in the optical spectral region). Since photoconductivity and photoluminescence are proportional to the number of optically generated carriers, their anisotropy is given by the same Equation 26.3.

Note that, we consider the classical electrodynamic effects caused exclusively by the difference of the refractive indices for nanostructures and the environment. In quantum nanostructures with a noticeable size quantization, the optical matrix element becomes anisotropic (Sercel and Vahala 1990, McIntyre and Sham 1992), which may quantitatively modify the described polarization effects so that Equation 26.3 acquires an additional factor equal to the square of the ratio of the matrix elements for the two perpendicular directions.

Though the given expressions are derived for static electric fields, they remain valid for high-frequency fields as well (with ε and ε_0 corresponding to the field frequency ω), as long as the NW radius a or both the NR dimensions a, l remain much less than the light wavelength λ. The situation when this inequality is violated will be referred to as the case of "thick" NWs and considered separately in Sections 26.2.1.2 and 26.2.3.2.

The direct measurement of optical absorption in a single NW is difficult but can be replaced by measurements of photoluminescence excitation or photoconductivity (for NWs provided with contacts). All the phenomena discussed should be polarization-sensitive acquiring their maximum for the exciting light polarization parallel to the NWs. The effect in its pure form was first observed and explained in (Wang et al. 2001). The authors observed photo-luminescence and photoconductivity in InP NWs increasing up to 49 times when the polarization of the exciting light changed from perpendicular to parallel (Figure 26.2). This number exactly corresponds to Equation 26.3 for InP in air. Similar effects with a slightly lesser degree of polarization (maybe, owing to the misorientation of the NWs, the role of that is discussed elsewhere (Ruda and Shik 2005a)) were also observed in the photoluminescence of Si NWs (Qi et al. 2003) and the photoconductivity of ZnO and GaN NWs (Fan et al. 2004, Han et al. 2004).

26.2.1.2 Thick Nanowires

The quasi-static approach of the previous section is adequate only for thin nanostructures with $a \ll \lambda$. At the same time, a number of optical experiments (Fan et al. 2004, Hsu et al. 2004, Philipose et al. 2005) are performed in semiconductor NWs with a diameter as large as 100 nm and more. This is comparable with

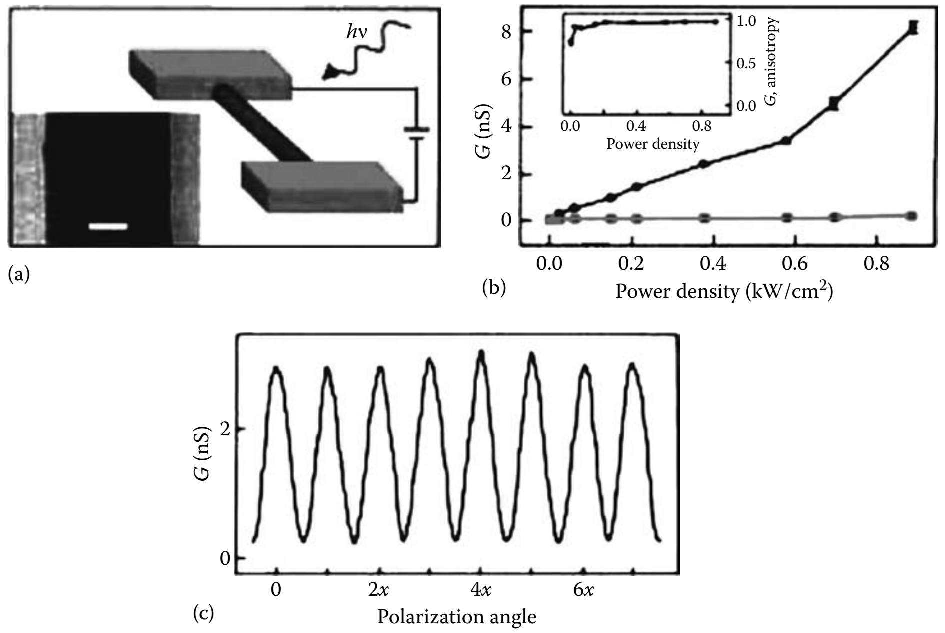

FIGURE 26.2 Photoconductivity versus light-polarization angle in an InP NW with $a = 10\,\text{nm}$. (a) Schematic depicting the use of an NW as a photodetector: inset, field-emission scanning electron microscopy image of an NW and contact electrodes, scale bar $2\,\mu\text{m}$. (b) Conductance, G, versus excitation power density for the illumination polarized parallel (upper curve) and perpendicular (lower curve) to the NW; inset, photoconductivity anisotropy versus excitation power, equal to 0.96 for the shown device. (c) Conductance versus polarization angle. The measurements were performed at room temperature for the excitation wavelength 514.5 nm. (From Wang, J. et al., *Science*, 293, 1455, 2001. With permission.)

the interband light wavelength, which makes the formulae of Section 26.2.1.1 inapplicable for this case.

Now we theoretically consider the anisotropy of light absorption and the emission caused by image forces in semiconducting NWs of an arbitrary radius a. For a comparable to the light wavelength, the high-frequency electric field inside NWs is no longer uniform, and the polarization characteristics acquire a noticeable, non-monotonic spectral dependence allowing one to observe a number of new effects in the optical properties of NWs with submicron thickness.

We consider a cylindrical NW placed in an external ac electric field $\mathbf{E_0}$ and find the real field distribution $\mathbf{E}(\mathbf{r})$ created by $\mathbf{E_0}$ together with the NW-induced image forces. Then we calculate the total absorbed power proportional to $|\mathbf{E}(\mathbf{r})|^2$ integrated over the NW volume and study the dependence of this characteristic on the angle between $\mathbf{E_0}$ and the NW axis, which gives us the polarization dependence of our interest.

The main distinction to Section 26.2.1.1 consists in the need to find $\mathbf{E}(\mathbf{r})$ not from the Laplace but from the Helmholtz wave equation. Using cylindrical coordinates (ρ, z, φ), it can be written inside the NW ($\rho < a$) as

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial E_i}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2 E_i}{\partial\varphi^2}+\frac{\partial^2 E_i}{\partial z^2}+\varepsilon\frac{\omega^2}{c^2}E_i=0 \qquad (26.4)$$

where c is the light velocity. Outside the NW ($\rho > a$) the equation differs from Equation 26.4 only by the replacement $\varepsilon \rightarrow \varepsilon_0$. The character of its solution depends on the mutual orientation of $\mathbf{E_0}$, the light wave vector $\mathbf{k}$, and the NW axis $\mathbf{z}$. Below, we will give only the final results of calculations referring the reader to the original work (Ruda and Shik 2006).

For $\mathbf{k} \perp \mathbf{z}$ (say, $k = k_x$) and $\mathbf{E_0} \parallel \mathbf{z}$, when $\mathbf{E}_0(\mathbf{r}) = E_0\,\mathbf{z}\exp[ik_0\rho\cos\varphi]$, the solution for $\mathbf{E}$ contains only a single component E_z given by the expression (Batygin and Toptygin 1978)

$$E_z(\rho,\varphi)=E_0\sum_{m=-\infty}^{\infty} i^m \frac{J'_m(ka)H_m^{(1)}(k_0a)-J_m(ka)H_m^{(1)\prime}(k_0a)}{\sqrt{\varepsilon}J'_m(ka)H_m^{(1)}(k_0a)-\sqrt{\varepsilon_0}J_m(ka)H_m^{(1)\prime}(k_0a)} \times J_m(k\rho)\exp(im\varphi) \qquad (26.5)$$

where J_m and $H_m^{(1)}$ are the Bessel and the first-kind Hankel functions, $k=\sqrt{\varepsilon}\omega/c$, $k_0=\sqrt{\varepsilon_0}\omega/c$.

For the same direction of $\mathbf{k}$ but perpendicular polarization, $\mathbf{E_0} \parallel \mathbf{x}$, Equation 26.4 gives the following general solution inside the NW:

$$\mathbf{E}(\rho,\alpha)=\frac{2E_0a\varepsilon_0\nabla\left[J_1\left(\dfrac{\sqrt{\varepsilon-\varepsilon_0}\,\omega\rho}{c}\right)\cos\varphi\right]}{\varepsilon_0 J_1\left(\dfrac{\sqrt{\varepsilon-\varepsilon_0}\,\omega a}{c}\right)+\varepsilon\dfrac{\sqrt{\varepsilon-\varepsilon_0}\,\omega a}{c}J'_1\left(\dfrac{\sqrt{\varepsilon-\varepsilon_0}\,\omega a}{c}\right)}. \qquad (26.6)$$

At small a Equation 26.6 transforms into the static expression (Equation 26.1).

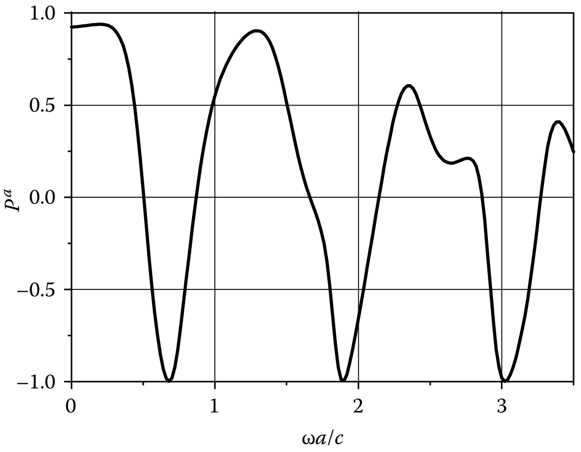

FIGURE 26.3 Spectral dependence of the polarization ratio $P^a = \left(I_{\|}^a - I_{\perp}^a\right)/\left(I_{\|}^a + I_{\perp}^a\right)$ in an NW for ε = 9, $\varepsilon_0 = 1$. (From Ruda, H.E. and Shik, A., *J. Appl. Phys.*, 100, 024314, 2006. With permission.)

Equations 26.5 and 26.6 show that, contrary to the case of a thin NW, the field distribution inside it is essentially nonuniform. However, real experimentally measured characteristics, such as the absorption, the photoconductivity, and the photoluminescence excitation spectra are determined not by details of this field distribution but by the total power absorbed, proportional to the integral of the light intensity over the NW cross section $I^a = \iint |E(\varphi,\alpha)|^2 \, d\varphi\rho d\rho$ (superscript *a* means absorption). Figure 26.3 shows the spectral dependence of the ratio $P^a = \left(I_{\|}^a - I_{\perp}^a / I_{\|}^a + I_{\perp}^a\right)$ characterizing the polarization dependence of the absorbed light intensity.

For $k_0 a \to 0$, this ratio tends to the value $(\varepsilon^2 + 2\varepsilon\varepsilon_0 - 3\varepsilon_0^2)/(\varepsilon^2 + 2\varepsilon\varepsilon_0 + 5\varepsilon_0^2)$ corresponding to Equation 26.3. However, at higher light frequencies (or for thicker NWs) we can no longer claim that light with a parallel polarization has a much higher absorption coefficient since $I_{\|}^a$ and $I_{\perp}^a$, as well as their ratio, demonstrate a strong frequency dependence. This dependence is especially dramatic for $I_{\perp}^a$, which tends to infinity in a series of critical points corresponding to roots of the denominator Equation 26.6 and $P^a = -1$ in Figure 26.3. In our example of ε = 9, $\varepsilon_0 = 1$, the first of these points are $k_0 a$ = 0.68, 1.89, 3.02,. … Considering a cylindrical NW as an optical fiber, we can express each of critical points as the cut-off of a fiber mode LP_{1m} (*m* = 1, 2, 3…) with the same angular field distribution as Equation 26.6. At the cut-off, these modes become purely transverse and excited by an incident wave. The spectrum of $I_{\|}^a(\omega)$ has no singularities due to the absence of modes with a purely longitudinal polarization (**E** || **z**).

Since an infinite value $I_{\perp}^a$ at some fixed ω is definitely nonphysical, we must discuss what initial theoretical assumptions are inadequate and what phenomena are responsible for the real height of the absorption peaks. In our calculations, we have considered one single NW in the electric field of an infinite plane wave. In this case, the field in an NW can, in theory, acquire arbitrary high values remaining equal to $\mathbf{E_0}$ at very large distances. For an array of NWs with an area density *N*, the electric field will be redistributed between the NWs and their environment so that the total wave energy remains the same as that provided with external excitation in the absence of NWs and equal to $\varepsilon_0 E_0^2/4\pi$ per unit volume and the absorption coefficient is finite. There are also some other mechanisms of smoothening the mentioned singularities in $I_{\perp}^a$, e.g., non-rectilinearity and thickness fluctuations of NWs, which may become dominant at a low NW density.

The described frequency dependence of polarization-sensitive absorption has not yet been studied experimentally, except for the observation (Fan et al. 2004) that two different frequencies of excited light caused different degrees of photoconductivity anisotropy.

26.2.2 Metal Nanostructures: Plasmon Phenomena

26.2.2.1 Longitudinal and Transverse Plasmons

The conclusions of Section 26.2.1.1 and formulae presented there can be applied to both metal and semiconductor NWs and NRs but result in the different spectral dependences of optical anisotropy. The degree of polarization in thin NWs, determined exclusively by $\varepsilon/\varepsilon_0$, in semiconductor nanostructures does not explicitly depend on the light frequency ω (if we ignore the frequency dependence of ε, which in semiconductors is rather weak, even near the interband absorption edge). In metals, on the contrary, ε has a strong frequency dispersion caused by the electron plasma, which makes the polarization characteristics frequency-dependent even in the quasistatic limit of small nanostructures.

For an analytical description of these effects, we will use for ε the simplest Drude expression: $\varepsilon = 1 - \left(\omega_p^2 / \omega(\omega - i\nu)\right)$ where ω_p is the plasma frequency in a bulk metal and ν is the scattering rate. Substituting this expression into Equation 26.2, we see that at a small ν (weak electron scattering), the optical electric field inside NRs increases dramatically at the frequency

$$\omega_{\|,\perp} = \frac{\omega_p \sqrt{n_{\|,\perp}}}{\varepsilon_0 (1 - n_{\|,\perp}) + n_{\|,\perp}} \qquad (26.7)$$

called the plasmon frequency,* which is smaller than ω_p and typically belongs to the visible spectral region. The optical absorption spectrum has a strong maximum near the plasmon frequency. In it seen that light polarized along the NR axis generates plasmons with the frequency $\omega_{\|}$ called longitudinal plasmons, while light with the perpendicular polarization generates transverse plasmons with the frequency $\omega_{\perp}$, larger than $\omega_{\|}$. In a random system of NRs with arbitrary orientations related to the light polarization (or for a nonpolarized optical excitation), the absorption spectrum has two absorption maxima: at $\omega_{\|}$ and $\omega_{\perp}$ (Kreibig and Vollmer 1995).

---

* $\omega_{\|,\perp}$ are often called the surface plasmon frequencies, to distinguish them from the plasmon frequency of bulk material ω_p.

We note that for long NWs $\omega_{\|} \rightarrow 0$, in other words, longitudinal plasmons cease to exist. Formally, it is due to the fact that in infinitely long cylinders, the parallel component of the electric field must be continuous at the interface and hence is uniform, without any structure characterizing plasma oscillations.

26.2.2.2 Transverse Plasmons in Nanowires

If we substitute the Drude expression for $\varepsilon(\omega)$ into Equation 26.3, we will see that the anisotropy of the optical absorption in metal NWs has an interesting spectral dependence. Far below the plasmon frequency, $|\varepsilon(\omega)|$ is very large, and the situation is similar to that in semiconductor NWs where the absorption is much higher for the light with parallel polarization. However, near the plasmon frequency, the absorption of the perpendicularly polarized light increases drastically, exceeding that for the parallel polarization and hence changing the sign of the anisotropy, as illustrated by Figure 26.4. At $\omega \rightarrow \infty$ the anisotropy tends to a constant value determined by ε_0 and equal to zero at $\varepsilon_0 = 1$.

Thus, the situation near the plasmon resonance is, to some extent, opposite to that of semiconductor NWs where, due to the large ε, the electric field of light is polarized mostly along the NW axis. In metal NWs, for light with the frequency in the transverse plasmon region polarized perpendicular to the NW axes, ε is close to zero. For this reason, the electric field in an NW has essentially normal polarization and very high amplitude—the effect often called the plasmon amplification (see also Section 26.2.5.3).

Though the described spectral dependences of absorption are very specific and interesting, their experimental observation is more difficult than for the effects in semiconductor NWs. In metals, we can measure neither the photoconductivity nor the luminescence excitation and have to restrict ourselves to direct measurements of optical absorption. The most straightforward method to employ is to investigate a large number of NWs suspended in some liquid optically inactive in the spectral region of interest. In this case, the orientation of NWs is random, and no polarization dependence should be observed. A strong anisotropy of absorbing objects will be revealed in a specific character of spectral dependence of the absorption coefficient (Figure 26.5). In this spectrum, strong low-frequency absorption is caused by NWs almost parallel to the light polarization while the plasmon maximum is formed by those with an orientation close to the perpendicular one.

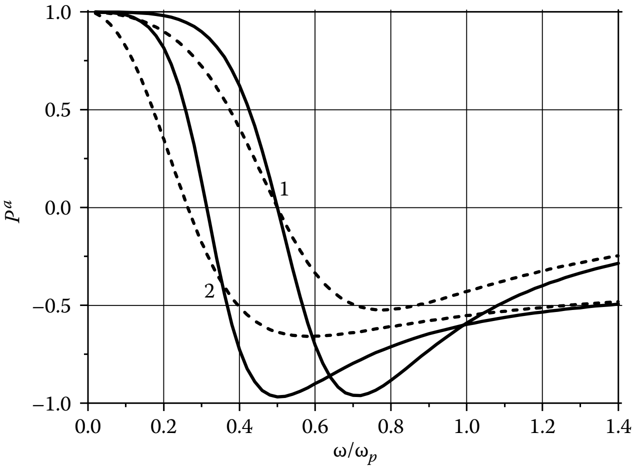

FIGURE 26.4 Absorption anisotropy $P^a = (k_{\|} - k_{\perp})/(k_{\|} + k_{\perp})$ in metal NWs for $\varepsilon_0 = 1$ (curves 1) and $\varepsilon_0 = 3$ (curves 2). Solid curves correspond to $\nu = 0.1\omega_p$, dashed curves to $\nu = 0.5\omega_p$. (From Ruda, H.E. and Shik, A., *Phys. Rev. B*, 72, 115308, 2005a. With permission.)

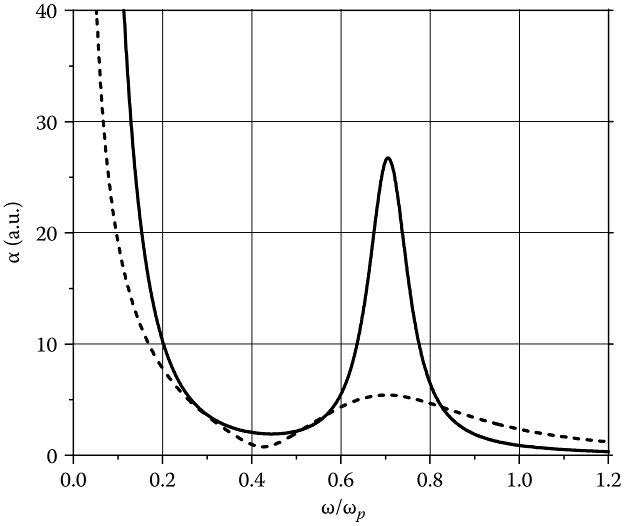

FIGURE 26.5 Absorption spectrum of randomly oriented metal NWs. Solid curve corresponds to $\nu = 0.1\omega_p$, dashed curves to $\nu = 0.5\omega_p$. (From Ruda, H.E. and Shik, A., *Phys. Rev. B*, 72, 115308, 2005a. With permission.)

26.2.3 Semiconductor Nanostructures: Polarization of Luminescence

26.2.3.1 Thin Nanowires and Nanorods

In the previous sections 26.2.1 through 26.2.2, we considered the physical phenomena related to the fact that external nonpolarized light becomes strongly polarized inside a nanostructure with a high refractive index. Now, we discuss the related problem of polarization of light emitted by an NW. As a first step, we solve the auxiliary problem of finding the electric field created by an electric dipole placed at the axis of a cylinder with a radius a and a dielectric constant ε, in an environment with ε_0. Details of these calculations can be found in Ruda and Shik (2005a), and here we present only the final results. For the dipole moment $\mathbf{d}_0$ parallel to the cylinder axis (z-axis), the electric field of the radiation far from the NW, has the same amplitude and configuration as would be created by a dipole with the moment $\mathbf{d}_0$ in free space. For the normal orientation of $\mathbf{d}_0$ (along x-axis), the corresponding effective moment is given by $\mathbf{d} = 2\varepsilon_0\mathbf{d}_0/(\varepsilon + \varepsilon_0)$.

As a result, even for NWs with isotropic interband matrix elements (which is the case for cubic semiconductors and a moderate a when size quantization is negligible), where luminescence would otherwise be nonpolarized, a strong polarization under the influence of image forces is acquired.

As a second step, we consider the net polarization of luminescence from a line x, $y = 0$ containing a mixture of different emitting dipoles. We take a point z along the wire axis containing three mutually perpendicular dipoles d_x, d_y, d_z with a frequency ω and calculate the resulting electric field far from the NW at some point $(x_0, 0, 0)$ using the standard formula for the electric field of an emitting dipole and adding contributions of all line segments by integrating over z. Since radiation from different segments is incoherent, we do not add electric fields, but Poynting vectors **S** containing no phase factors. The resulting **S** is directed along the x-axis and for the two different light polarizations is

$$(S_x)_{\|} = \frac{\eta\omega^4}{6\pi\varepsilon_0^2 c^3 x_0}\left(d_x^2 + 2d_z^2\right), \tag{26.8}$$

$$(S_x)_{\perp} = \frac{\eta\omega^4}{2\pi\varepsilon_0^2 c^3 x_0} d_y^2 \tag{26.9}$$

where η is the density of dipoles per unit line length.

Combining these two problems and substituting $d_{x,y} = 2\varepsilon_0 d_{0x,y}/(\varepsilon + \varepsilon_0)$; $d_z = d_{0z}$ into Equations 26.8 and 26.9, provides the resulting polarization of light emitted by an NW. The intensity ratio for different light polarizations in the case of isotropic matrix elements is

$$\frac{I_{\|}^e}{I_{\perp}^e} = \frac{d_x^2 + 2d_z^2}{3d_z^2} \tag{26.10}$$

where superscript e means emission. If the bare dipole moment components $d_{0x,y,z}$ do not coincide due to size quantization or/and crystal anisotropy, this modifies the polarization character introducing anisotropy even in the absence of image forces, similar to the case of absorption.

It is seen from Equation 26.10 that light emitted by an NW with a dielectric constant higher than that of its environment, will be strongly polarized in the wire direction, which was clearly demonstrated experimentally (Wang et al. 2001, Qi et al. 2003). The effect is qualitatively similar to the property of absorption considered in Section 26.2.1.1 but the magnitude of polarization is different. By comparing Equation 26.3 with Equation 26.10 one can see that the degree of polarization for luminescence is less than for absorption. It is caused by the fact that, according to Equation 26.8, light with parallel polarization is generated not only by d_z dipole moments but also by d_x that are weakened by image forces. For example, for InP NWs ($\varepsilon = 12.7$) where, as was already mentioned in Section 26.2.1.1, $\alpha_{\|}/\alpha_{\perp} \cong 49$, Equation 26.10 gives $I_{\|}^e / I_{\perp}^e \cong 30$, in good agreement with the experimentally observed values of $\alpha_{\|}/\alpha_{\perp} \cong 48$ and $I_{\|}^e / I_{\perp}^e \cong 24$ (Wang et al. 2001).

All the results obtained above for semiconductor NWs remain valid also for NRs, except for the numerical values of polarization coefficients, which for NRs depend not only on $\varepsilon/\varepsilon_0$ but also on the depolarization coefficients, or, in other words, on the NR aspect ratio.

26.2.3.2 Thick Nanowires

Now we consider the polarization properties of luminescence from thick NWs where, according to Section 26.2.1.2, the electric field cannot be assumed uniform. Though in fact luminescence occurs throughout the whole NW volume, we will model it by an effective emitting dipole at the axis ($\rho = 0$). (Precise results can be obtained, if necessary, by using general formulae for a field created in a system of coaxial cylindrical layers by a dipole with an arbitrary position $\mathbf{r_0}$ followed by integration over $\mathbf{r_0}$, to be performed only numerically.) The general approach to similar problems is described in (Chew 1990), which allows us to present only the final expression for the resulting electric field outside the NW.

For the dipole moment $\mathbf{d_0}$ oriented along the NW axis ($d_0 = d_{0z}$), the electric field E_z far from the NW (at $\rho \gg k_0^{-1}$) is given by

$$E_z \cong i\frac{d_{0z}\omega^2}{\varepsilon c^2 r}(\varepsilon - \varepsilon_0 \cos^2\theta)\exp(ik_0 r - i\pi/4) \times \frac{J_1(k_\rho^0 a)H_0^{(1)}(k_\rho^0 a) - J_0(k_\rho^0 a)H_1^{(1)}(k_\rho^0 a)}{J_1(k_\rho^0 a)H_0^{(1)}(k_{\rho 0}^0 a) - \dfrac{k_{\rho 0}^0}{k_\rho^0} J_0(k_\rho^0 a)H_1^{(1)}(k_{\rho 0}^0 a)} \tag{26.11}$$

where

$k_\rho^0 = k_0\sqrt{\varepsilon/\varepsilon_0 - \cos^2\theta} = k\sqrt{1 - \varepsilon_0\cos^2\theta/\varepsilon}$
$k_{\rho 0}^0 = k_0 \sin\theta$
$r = \sqrt{\rho^2 + z^2}$
k_ρ^0 is the distance to the dipole
$\theta = \arccos(z/r)$ is the angle from the NW axis

If we consider the light emission normal to the dipole (along the y-axis) where $\theta = \pi/2$, the corresponding electric field will have the same coordinate dependence as that of a free dipole $\sim(d_{0z}\omega^2/c^2 r)\exp(ik_0 r)$ with an additional factor given by the fraction in Equation 26.11. Thus, the influence of image forces is equivalent to the replacement of d_0z with some effective dipole (Ruda and Shik 2006):

$$d_z = d_{0z}\sqrt{\varepsilon}\frac{J_1(ka)H_0^{(1)}(ka) - J_0(ka)H_1^{(1)}(ka)}{\sqrt{\varepsilon}J_1(ka)H_0^{(1)}(ka) - \sqrt{\varepsilon_0}J_0(ka)H_1^{(1)}(ka)} \equiv A_z d_{0z}. \tag{26.12}$$

For the perpendicular dipole orientation ($d_0 = d_{0x}$), similar calculations give

$$d_x = d_{0x}\sqrt{\varepsilon}\frac{J_1'(ka)H_0^{(1)}(ka) - J_0(ka)H_1^{(1)\prime}(ka)}{\sqrt{\varepsilon_0}J_1'(ka)H_0^{(1)}(ka) - \sqrt{\varepsilon}J_0(ka)H_1^{(1)\prime}(ka)} \equiv A_x d_{0x}. \tag{26.13}$$

Figure 26.6a shows the frequency dependence of the effective dipoles d_x and d_z. At $\omega \to 0$ they acquire the static values $d_z = d_{0z}$, $d_x = 2\varepsilon_0 d_{0x}/(\varepsilon + \varepsilon_0)$ given in Section 26.2.3.1, and then demonstrate strong oscillations. At a larger ω (or a larger a), the phases

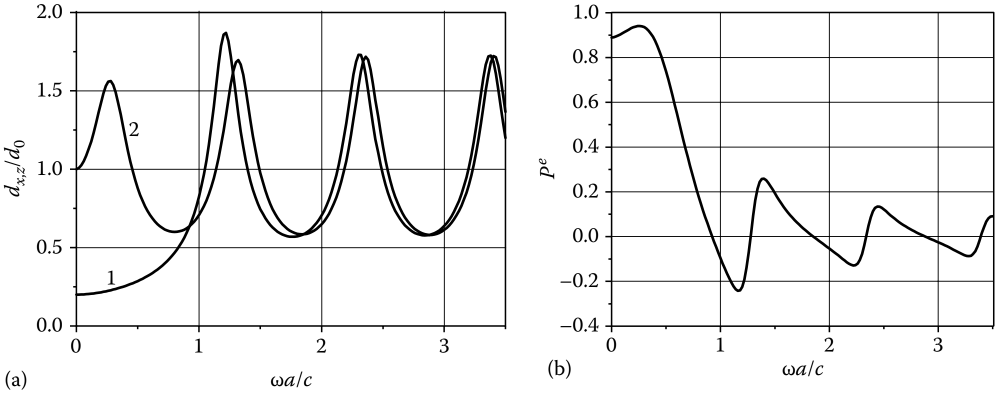

FIGURE 26.6 (a) Spectral dependence of the effective dipoles d_x (1) and d_z (2) for an NW with $\varepsilon = 9$, $\varepsilon_0 = 1$; (b) spectral dependence of the polarization ratio P^e for the same NW. (From Ruda, H.E. and Shik, A., *J. Appl. Phys.*, 100, 024314, 2006. With permission.)

of oscillations close in, d_z approaches d_x, so that the radiation becomes almost unpolarized.

So far, we have analyzed the emission characteristics of one single effective dipole in an NW. To obtain the radiation characteristics of the whole NW, we must perform an integration over its length assuming the dipoles distributed uniformly along z, exactly as it was done in Section 26.2.3.2. Substituting Equations 26.12 and 26.13 into the first equality of Equation 26.10 and assuming an isotropic internal emission ($d_{0z} = d_{0x}$), we obtain the spectral dependence of the resulting polarization ratio $P^e = \left(I^e_{\parallel} - I^e_{\perp}/I^e_{\parallel} + I^e_{\perp}\right)$ presented at Figure 26.6b. The latter is seen to have a strong oscillatory character changing sign at some critical frequencies differing from those describing the polarization dependence of absorption (Figure 26.3).

It is evident from Figure 26.6b that, if the luminescence spectrum of an NW material contains several lines with noticeably different frequencies, the polarizations of these lines can be essentially different and even have an opposite sign. It was confirmed by experiments (Hsu et al. 2004) in ZnO NWs. This material contains two strong luminescence lines with frequencies differed 1.3 times (Figure 26.7a). As is seen from Figure 26.7b, these lines have opposite angular dependences, indicating at opposite signs of luminescence anisotropy.

Note that the same polarization phenomena must be observed not only for photo- but for any other mechanisms of luminescence. In the case of photoluminescence, a change in the exciting light polarization (for a fixed ω_{ex}) does not change the described spectra $I^e_{\parallel}/I^e_{\perp}(\omega_{em})$, but modifies the amplitudes $I^e_{\parallel}$ and $I^e_{\perp}$ synchronously according to the polarization dependence of the absorbed light intensity described in Section 26.2.1.2.

The luminescence excitation spectrum in most physical systems (including thick NWs with microscopic electronic properties coinciding with those of bulk semiconductors) actually reproduces the optical absorption spectrum and hence is given by the formulae of Section 26.2.1.2. Thus, we may claim that the polarization dependence of the luminescence excitation spectrum for any emission frequency and polarization is given by Figure 26.3.

26.2.4 Nonlinear Phenomena in Metal and Semiconductor Nanowires

26.2.4.1 Theory

It is evident that the polarization dependence of the internal optical field in NWs and NRs must result in anisotropy not only of linear but also of nonlinear optical phenomena, such as, e.g., the second harmonic generation (SHG). Moreover, one may

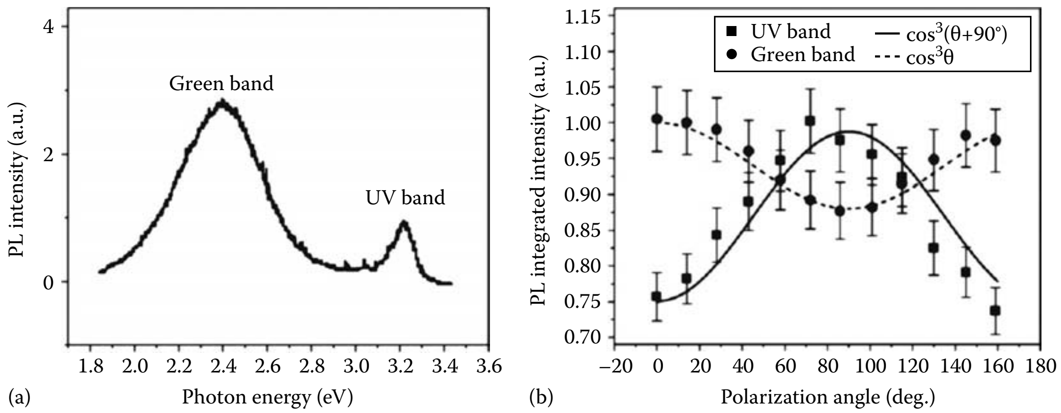

FIGURE 26.7 Luminescence spectrum of ZnO NW (a) with two emission lines and an angular dependence of the luminescence intensity (b) in these lines. (From Hsu, N.E. et al., *J. Appl. Phys.*, 96, 4671, 2004. With permission.)

expect the anisotropy of nonlinear effects to be larger than the linear ones, since the former are proportional to a higher degree of electric field.

The analysis of the SHG in NWs and NRs is less straightforward than that of linear effects where we took standard formulae for light absorption or emission in bulk materials and simply substituted the specific optical field distribution of nanostructures into them. On the contrary, in uniform bulk materials characterized by inversion symmetry, the SHG is completely absent. Hence, the theory of this effect in nanostructures should be developed from scratch.

We start from the classical expression connecting nonlinear polarization at the frequency 2ω with the spatial distribution of the electric field at the frequency ω (Blombergen et al. 1968):

$$\mathbf{P}^{2\omega} = \gamma\nabla(\mathbf{E}^{\omega})^2 + \beta\mathbf{E}^{\omega}(\nabla\mathbf{E}^{\omega}). \tag{26.14}$$

For nonconducting media, $\beta = -2\gamma$ and is determined by the square of linear susceptibility, while for metals $\beta = e/(8\pi m^*\omega^2\}$ and $\gamma = \beta[1 - \varepsilon(2\omega)]$ (for simplicity, $\varepsilon_0 = 1$ is assumed). Our task is to find $\mathbf{P}^{2\omega}$ for the $\mathbf{E}^{\omega}(\mathbf{r})$ determined for nanostructures in Section 26.2.1 and to use it for the calculation of the SHG intensity.

For the case of spherical nanocrystals, it was shown (Agarwal and Jha 1982) that the two terms on the right-hand side of Equation 26.14 correspond, respectively, to the volume component $\mathbf{P}_v^{2\omega}$ caused by the coordinate-dependent $\mathbf{E}^{\omega}$ inside the nanostructure and having a dipole symmetry, and the surface component $\mathbf{P}_s^{2\omega}$ caused exclusively by the nanostructure surface and with a quadrupole symmetry. Since the second term in Equation 26.14 for small nanocrystals essentially exceeds the first one, $\mathbf{P}_s^{2\omega}$ has a larger amplitude than $\mathbf{P}_v^{2\omega}$ but, at the same time, with a quadrupole, rather than a dipole symmetry, which emits less efficiently. For this reason, the relative role of these components is *a priori* not evident and depends on the direction and the polarization of light.

We restrict ourselves to the case of thin cylindrical NWs where $\omega a/c \ll 1$ can be used as a perturbation parameter. The particular expressions for $\mathbf{E}^{\omega}$ depend on the directions of the wave vector $\mathbf{k}$ and the electric field $\mathbf{E}_0^{\omega}$ of light related to the NW axis (z-axis) and can be obtained by expansion of the field distributions Equations 26.5 and 26.6 in terms of ω. There exist three possible optical geometries.

a. $\mathbf{k} \| \mathbf{z};\ \mathbf{E}_0^{\omega} \perp \mathbf{z}$ (longitudinal propagation, transverse polarization)

   In this case

$$E^{\omega} = E_x^{\omega} = \frac{2E_0^{\omega}}{[\varepsilon(\omega)+1]}(1+ikz) \tag{26.15}$$

($k = \omega/c$). Substitution of Equation 26.15 into Equation 26.14 shows that $\mathbf{P}_v^{2\omega}$ is determined only by the coordinate-dependent component of $\mathbf{E}^{\omega}(\mathbf{r})$, directed along the NW axis (z-axis) and has the value

$$P_v^{2\omega} = \frac{8i\gamma k[E_0^{\omega}]^2}{[\varepsilon(\omega)+1]^2} = \text{const}(\mathbf{r}). \tag{26.16}$$

Thus, each elementary volume of the NW dV can be considered as an elementary dipole $\mathbf{P}_v^{2\omega}dV$ parallel to the NW axis. For such dipole orientation, its radiation is not disturbed by the image forces and is the same as for free dipoles (Ruda and Shik 2005a), not depending on the NW dielectric constant at the emission frequency, ε(2ω). The resulting radiation electric field at the distance $r \gg a$ from NW can be found as a sum of elementary dipoles along the whole NW. Calculations, shown in more detail in Ruda and Shik (2007a), show that at a large distance r from an NW ($r \gg a$; $kr \gg 1$) the SHG intensity per unit NW length is

$$I_v^{2\omega} = \frac{\pi^3\omega^3 a^4\left[P_v^{2\omega}\right]^2}{c^2} = \frac{64\pi^3\gamma^2\omega^5 a^4\left[E_0^{\omega}\right]^4}{c^4\left[\varepsilon(\omega)+1\right]^4}. \tag{26.17}$$

The radiation is emitted isotropically (independent of a particular direction of light polarization).

The surface component of polarization $\mathbf{P}_v^{2\omega}$ is determined by an $\mathbf{E}^{\omega}$ discontinuity at the NW interface caused by a dielectric mismatch, and for its determination we can retain only the first, uniform term in Equation 26.15. After simple transformations, we have

$$P_s^{2\omega} = \frac{4\beta[\varepsilon(\omega)-1]\left[E_0^{\omega}\right]^2}{[\varepsilon(\omega)+1]^2}\delta(\rho - a)\cos\alpha. \tag{26.18}$$

(ρ,α) are polar coordinates in the xy-plane. The direction of $\mathbf{P}_v^{2\omega}$ coincides with that of the $\mathbf{E}_0^{\omega}$ (x-axis). This distribution of polarization has a quadrupole symmetry. Due to the coordinate dependence of $P_s^{2\omega}$, the determination of the electric field of emitted light must contain integration not only over z, but over ρ and α as well. Besides, in this case, we deal with elementary dipoles directed not parallel (as for Equation 26.16) but perpendicular to the NW axis. Due to the image forces, the ac electric field of these dipoles is suppressed by the factor 2/[ε(2ω) + 1] (see Section 26.2.3.1). Under the same conditions as before, we get the intensity of this surface-related SHG component:

$$I_s^{2\omega} = \frac{4\pi^3\beta^2\omega^5 a^4\left[\varepsilon(\omega)-1\right]^2\left[E_0^{\omega}\right]^4}{c^4\left[\varepsilon(\omega)+1\right]^4\left[\varepsilon(2\omega)+1\right]^2}\sin^2(2\varphi) \tag{26.19}$$

where φ is the angle between the exciting-light wave vector $\mathbf{k}$ and the direction to the observation point. As for any quadrupole emitter, the intensity $I_s^{2\omega}$ has four directional lobes.

b. $\mathbf{k} \perp \mathbf{z}$; $\mathbf{E}_0^\omega \parallel \mathbf{z}$ (transverse propagation, longitudinal polarization)

In this case

$$E^\omega = E_z^\omega = E_0^\omega\left(1 + ikx\frac{1+\sqrt{\varepsilon(\omega)}}{2}\right) \tag{26.20}$$

and $\mathbf{P}_v^{2\omega}$, which is always colinear with the light wave vector, is directed along the x-axis normal to the NW and is equal to

$$P_v^{2\omega} = i\gamma k\left[E_0^\omega\right]^2\left[1+\sqrt{\varepsilon(\omega)}\right] = \text{const}(\mathbf{r}). \tag{26.21}$$

For such an orientation of elementary dipoles, the SHG is anisotropic with the radiation pattern directed perpendicularly to $\mathbf{k}$, and the ac electric field of these dipoles contains the factor $2/[\varepsilon(2\omega)+1]$. Adding the fields of elementary dipoles in the same manner as before, we get

$$\begin{aligned} I_v^{2\omega} &= \frac{4\pi^3\omega^3a^4\left[P_v^{2\omega}\right]^2}{c^2\left[\varepsilon(2\omega)+1\right]^2}\sin^2\varphi \\ &= \frac{4\pi^3\gamma^2\omega^5a^4\left[E_0^\omega\right]^4\left|1+\sqrt{\varepsilon(\omega)}\right|^2}{c^4\left[\varepsilon(2\omega)+1\right]^2}\sin^2\varphi. \end{aligned} \tag{26.22}$$

Since $\mathbf{P}_s^{2\omega}$ is caused only by the normal component of $\mathbf{E}_0^\omega$ suffering discontinuity at the NW interface, in the case considered it vanishes, and the SHG has only a dipole component given by Equation 26.22 and no quadrupole component.

c. $\mathbf{k} \perp \mathbf{z}$; $\mathbf{E}_0^\omega \perp \mathbf{z}$ (transverse propagation, transverse polarization)

$$E^\omega = E_x^\omega = \frac{2E_0^\omega}{\left[\varepsilon(\omega)+1\right]}(1+iky). \tag{26.23}$$

This geometry combines characteristic features of the two previous cases. As in case (a), the SHG contains both dipole and quadrupole components. The dipole component of the polarization has the same value as Equation 26.16 but a perpendicular, rather than a longitudinal, orientation. For this reason, its radiation has an anisotropic pattern and is affected by the image force (as in case (b)) resulting in the additional dependence of the SHG on $\varepsilon(2\omega)$:

$$I_v^{2\omega} = \frac{4\pi^3\omega^3a^4\left[P_v^{2\omega}\right]^2}{c^2\left[\varepsilon(2\omega)+1\right]^2}\sin^2\phi = \frac{256\pi^3\gamma^2\omega^5a^4\left[E_0^\omega\right]^4}{c^4\left[\varepsilon(\omega)+1\right]^4\left[\varepsilon(2\omega)+1\right]^2}\sin^2\phi. \tag{26.24}$$

The surface polarization component caused by the first term in Equations 26.15 and 26.23 is exactly the same as in the case (a) (it results from the fact that this component is determined exclusively by light polarization being independent of the value and the direction of **k**) and hence is described by Equation 26.19.

Let us summarize the results obtained thus far. The SHG was shown to consist of volume and surface components. They have, respectively, dipole and quadrupole symmetry and a different spatial distribution of emitted radiation at 2ω. It can be isotropic (Equation 26.17), bidirectional (Equations 26.22 and 26.24), or of a four-lobed shape (Equation 26.19). However, the integral radiation intensity, as in the case of spherical nanodots (Agarwal and Jha 1982), is given by qualitatively similar expressions differing by numerical coefficients and factors depending only on the NW dielectric constant. For this reason, SHG anisotropy is independent of the NW radius a (until the condition $a\omega/c \ll 1$ used in our calculations, remains valid). The explicit frequency dependence $I^{2\omega} \sim \omega^5$ is also the same in all formulae, and the possible spectral dependence of SHG anisotropy may occur only due to the frequency dispersion of the NW dielectric constant ε. The latter is of minor importance in semiconductors and dielectrics but may be very strong in metal NWs. That is why the characteristics of SHG anisotropy of these two classes of NW materials are very different and will be considered separately.

Semiconductor materials are characterized by high values of the dielectric constant (~10 or even more) with relatively weak frequency dependence. For this reason, in our estimates we do not distinguish between ε(ω) and ε(2ω) denoting both of them as ε. The ratio of the integral SHG for different NW orientations is independent not only of the NW material and the diameter but also of the light frequency ω, being a function of one single parameter ε. In the limit $\varepsilon \gg 1$, the ratio of the total SHG intensities $\overline{I_v^{2\omega}} + \overline{I_s^{2\omega}}$ (symbol $\bar{I}$ means averaging over ϕ) for all three cases considered is $36:\varepsilon^3:4$. It means that the SHG is maximal for the case (b) when the exciting light is polarized along the NW axis. The result is not unexpected since only in this geometry the optical electric field inside the NW is not weakened by NW polarization and is equal to the external field. It correlates with the linear optical phenomena (absorption, photoconductivity, photoluminescence) which are also maximal for this light polarization. Much less trivial is the dependence of the SHG in perpendicularly polarized light in the direction of the light wave vector. The SHG is essentially larger for light propagating along the NW since, for this geometry, the effective dipoles forming $\mathbf{P}_v^{2\omega}$ are oriented parallel to the NW and their emission is not weakened by NW polarization.

It is worth noting that in all three cases, the spatial distribution of the SHG has different characteristics. In case (a), the second harmonic is emitted in all directions perpendicular to the NW (the contribution of the quadrupole component is around 10%), while in case (b), the emission has a dipole and in case (c), a quadrupole symmetry. This fact should be taken into account while interpreting experimental results (see Section 26.2.4.2).

The main distinguishing feature of metal systems is the strong frequency dispersion of $\varepsilon(\omega)$, which, as in Section 26.2.2, will be approximated by the Drude formula. It is seen that, due to the factors $\varepsilon(\omega) + 1$ and $\varepsilon(2\omega) + 1$ in the denominators of Equations 26.17, 26.19, 26.22, and 26.24, the SHG is amplified dramatically near the frequencies $\omega_p/\sqrt{2}$ and $\omega_p/(2\sqrt{2})$ when the excitation frequency or its second harmonic coincides with the transverse plasmon frequency in a cylinder. In case (b), the resonance exists only for 2ω since the exciting light is polarized along the NW and does not interact with plasmons.

It is interesting to consider SHG anisotropy near these two resonance frequencies when the effect itself is anomalously large. As has been mentioned, at $\omega \cong \omega_p/\sqrt{2}$ the SHG increases dramatically for cases (a) and (c) remaining constant in case (b). In other words, for light with parallel polarization, the SHG is negligibly small compared to the case of perpendicular polarization (the situation opposite to semiconductor NWs). The surface quadrupole component (Equation 26.19) is in this case almost two orders of magnitude less than the volume dipole component. The formulae, Equations 26.17 and 26.24, describing the latter for two different geometries, differ by only 12%. This means that for perpendicular polarization, contrary to the semiconductor case, the total SHG intensity is practically independent of the direction of illumination, though in case (a), emission is isotropic and is polarized parallel to the NWs, while in case (c), it has a dipole symmetry and is polarized perpendicular to them.

At the plasmon resonance for emitted light, $\omega \cong \omega_p/(2\sqrt{2})$, all contributions to the SHG, except Equation 26.17, contain the diverging factor $[\varepsilon(2\omega) + 1]^2$, so that the SHG for all light directions and polarization increases at $\omega \to \omega_p/(2\sqrt{2})$ in a similar manner with the anisotropy remaining constant. Taking into account that $\varepsilon(\omega_p/(2\sqrt{2})) \cong -7$, it can be seen that in the vicinity of this low-frequency resonance, light with the parallel polarization (case (b)) causes a SHG three orders of magnitude larger than that with perpendicular polarization, qualitatively similar to the situation in semiconductor NWs.

26.2.4.2 Experiment

First experimental investigations of SHG in NWs (Barzda et al. 2008) were performed on ZnSe NWs with a length of 8–10 microns and a diameter of 80–100 nm, excited by a laser at a 1029 nm wavelength, that satisfies the quasi-static condition $\omega a/c \ll 1$. The angular dependence of the SHG efficiency $I_{SH}(\theta)$ measured by changing the angle θ between the sample axis and the excitation beam polarization is presented by experimental points at Figure 26.8. It is seen that this dependence has a very strong character with $I_{SH}(0)$ (parallel polarization) exceeding $I_{SH}(90°)$ (perpendicular polarization) by more than an order of magnitude.

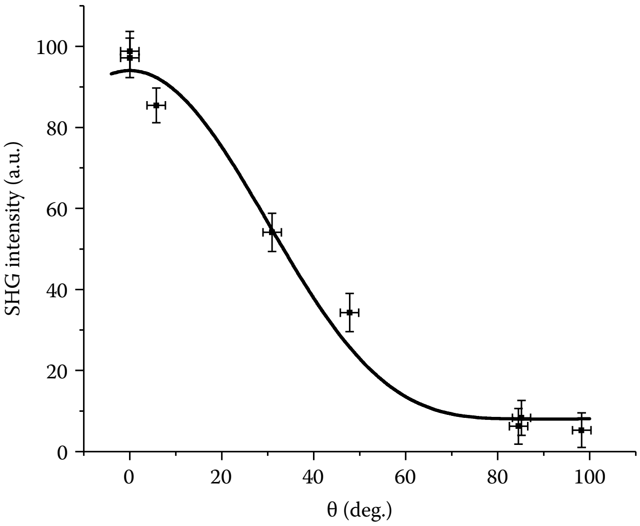

FIGURE 26.8 Polarization dependence of the SHG intensity (Barzda et al. 2008). The dots represent experimentally measured values, the solid line corresponds to the theoretical formulae (Equations 26.25 and 26.26). (From Barzda, V. et al., *Appl. Phys. Lett.*, 92, 113111, 2008. With permission.)

An adequate quantitative interpretation must be based on the theoretical results of Section 26.2.4.1. In the experiment, the light wave vector **k** was always perpendicular to the NW, so that only conclusions for the cases (b) and (c) are relevant. It is claimed that the SHG caused by the optical field with a longitudinal polarization $E_{\parallel}$ can be written as the radiation of an effective dipole parallel to the light wave vector with the intensity distribution Equation 26.22 proportional to $\sin^2 \varphi$, while in the SHG caused by $E_{\perp}$, the dominating component has a quadrupole character (Equation 26.19) with the intensity proportional to $\sin^2(2\varphi)$. In the experiments, the detector collected radiation from a wide angle $-49° < \varphi < 49°$ (see Figure 26.9), so that Equations 26.19 and 26.22 should be integrated in these limits. This results in an additional form-factor equal to 0.36 for parallel and to 0.93 for perpendicular light polarization. By taking into account that $E_{\parallel} = E_0 \cos\theta$ and $E_{\perp} = 2E_0 \sin\theta/(\varepsilon + 1)$, we obtain the final result for the detected intensity of the SHG:

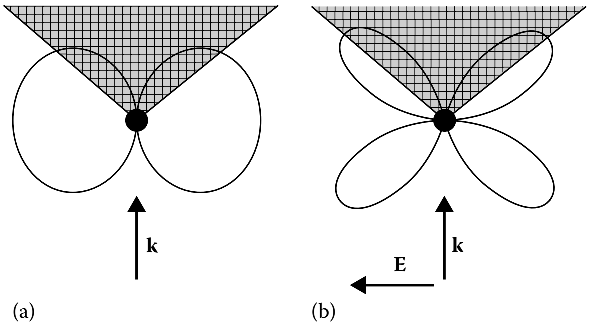

FIGURE 26.9 Schematic angular distribution of the SHG intensity for: (a) $\theta = 0$ (dipole mode) and (b) $\theta = 90°$ (quadrupole mode). **k** depicts the wave vector of the excitation light. The NW axis is perpendicular to the page. The orientation of the optical electric field $\mathbf{E}_0^{\omega}$ is perpendicular to the page in (a) and lies in the page plane in (b). The photodetector collects SHG radiation emitted in the forward direction within the shaded area. (From Barzda, V. et al., *Appl. Phys. Lett.*, 92, 113112, 2008. With permission.)

$$I_{SH} \propto E_0^4\left[\cos^4\theta + \frac{10.3(\varepsilon-1)^2\sin^4\theta}{(\varepsilon+1)^4(1+\sqrt{\varepsilon})^2}\right]. \tag{26.25}$$

For semiconductor NWs with $\varepsilon \gg 1$, Equation 26.25 predicts giant angular oscillations of the SHG with the maximum/minimum ratio of the order $\varepsilon^3/10$. For ZnSe, the value of ε reported by different authors varies from 5.9 to 7.2 (Grigoriev and Meilikhov 1997) but even for the smallest of these numbers, the theoretical ratio exceeds the corresponding experimental value from Figure 26.8, which is close to 20. The most evident explanation is attributed to deviations from an ideal cylindrical NW geometry used in deriving Equations 26.19 and 26.22. Free-standing NWs are inevitably bent, so that the local value of θ changes along the NW. After replacing θ by $\theta + \delta\theta$ and averaging over small $\delta\theta$ with Gaussian properties, Equation 26.25 acquires an additional term

$$\delta I_{SH} \propto 3E_0^4\overline{(\delta\theta)^2}\sin^2\theta\left[2\cos^2\theta + \overline{(\delta\theta)^2}\sin^2\theta\right]. \tag{26.26}$$

As a result, for $\overline{(\delta\theta)^2}\varepsilon^{3/2} > 2$ the maximum/minimum ratio becomes equal to $3\overline{(\delta\theta)^2}$, independent of the NW dielectric constant. The solid line in Figure 26.8 shows the theoretical angular dependence $I_{SH}(\theta)$ given by Equations 26.25 and 26.26 demonstrating a good agreement between theory and experiment. The best fitting of results was obtained for the value of the average NW corrugation $\overline{(\delta\theta)^2} \cong 0.2$.

26.2.5 Core–Shell Nanowires

26.2.5.1 Potential Distribution

So far, we have separately analyzed the optical properties of semiconductor and metal nanostructures and demonstrated their essential distinctions. In this connection, it is very interesting to investigate the properties of core–shell structures containing both types of layers, where luminescence in the semiconductor part can be amplified due to the plasmon effects in the metal part. Fabrication of such structures has been demonstrated experimentally (Lauhon et al. 2002, Choi et al. 2003).

As a starting point, we must find the formula for the intensity of the optical electric field inside such a structure, which would generalize Equation 26.1 to the case of an NW consisting of a central core with r_1 and a dielectric constant ε_1, and a shell with radius r_2 and a dielectric constant ε_2. Calculations show (Ruda and Shik 2005a) that an external electric field $\mathbf{E_0}$, perpendicular to the NW axis, creates in the core a uniform field

$$\mathbf{E_1} = \frac{4\varepsilon_0\varepsilon_2\mathbf{E_0}}{2\varepsilon_2(\varepsilon_1+\varepsilon_0)+p(\varepsilon_2-\varepsilon_1)(\varepsilon_2-\varepsilon_0)} \equiv A\mathbf{E_0} \tag{26.27}$$

where $p = 1-(r_1/r_2)^2$ is the relative shell volume. The field in the shell is a sum of a uniform field and that of a linear dipole:

$$\mathbf{E_2} = \frac{2\varepsilon_0}{2\varepsilon_2(\varepsilon_1+\varepsilon_0)+p(\varepsilon_2-\varepsilon_1)(\varepsilon_2-\varepsilon_0)} \times\left\{(\varepsilon_1+\varepsilon_2)\mathbf{E_0} + 2(\varepsilon_1-\varepsilon_2)(1-p)r_1^2\left[\frac{\mathbf{E_0}}{r^2} - \frac{2\mathbf{r}(\mathbf{r}\mathbf{E_0})}{r^4}\right]\right\}. \tag{26.28}$$

For $\mathbf{E_0}$ parallel to the NW axis, $\mathbf{E_1} = \mathbf{E_2} = \mathbf{E_0}$. Formulae (Equations 26.27 and 26.28) can be used to determine the spectra and the polarization dependence of optical absorption in core–shell NWs. We restrict our analysis to the already mentioned case of a metal shell and a semiconductor core, promising luminescence amplification.

The specific features of the optical response of metal shells arise from the fact that the plasmon-related absorption maximum for the light with perpendicular polarization is split into two maxima with positions depending on the relationship between r_1 and r_2. The frequency regions near the plasmon resonances are characterized by an anomalously high electric field strength (light amplification) both in the core and in the shell. As it has been already pointed out for core–shell nanodots (Neeves and Birnboim 1989), this can be used for the enhancement of optical effects in a core (in NWs—only for light with perpendicular polarization). There is also a reciprocal effect: an oscillating dipole $\mathbf{d_0}$ in the core is created far from the wire radiation field that may be considerably larger than that in a uniform medium. This effect of light amplification will be considered later in Section 26.2.5.3. In the next section, we investigate the optical absorption near the plasmon resonances caused by the increased electric field $\mathbf{E_2}$ in the shell.

26.2.5.2 Spectra and Polarization Dependence of Absorption

To calculate the optical absorption in a core–shell NW caused by light with a perpendicular polarization, we must integrate the local absorption $\mathrm{Im}\,\varepsilon_2|\mathbf{E_2}(\mathbf{r})|^2$ over the whole shell. Using the Drude formula for ε_2 and Equation 26.28 for $\mathbf{E_2}(\mathbf{r})$, we obtain the absorption spectra shown by solid lines in Figure 26.10. These contain two maxima corresponding to plasmon resonances where the real part of the denominator in Equations 26.27 and 26.28 vanishes and the electric field in the NW becomes anomalously large. With an increasing p, the maxima approach each other and the low-frequency one diminishes so that at $p \to 1$ we get a single peak at $\omega = \omega_p/\sqrt{2}$ corresponding to the plasma resonance in a metal cylinder.

Similar phenomena in core–shell nanodots have been already observed experimentally (Zhou et al. 1994, Oldenburg et al. 1998) and have a good theoretical description (Neeves and Birnboim 1989, Ruda and Shik 2005b). It is important to emphasize that, contrary to nanodots, the described two-peak absorption spectra in NWs are observed only for light polarized perpendicularly to the NWs. For parallel polarization, when the electric field in the NWs is equal to the external one, we have simple monotonically decreasing spectra reflecting the low-frequency absorption given by $\mathrm{Im}\,\varepsilon_2$.

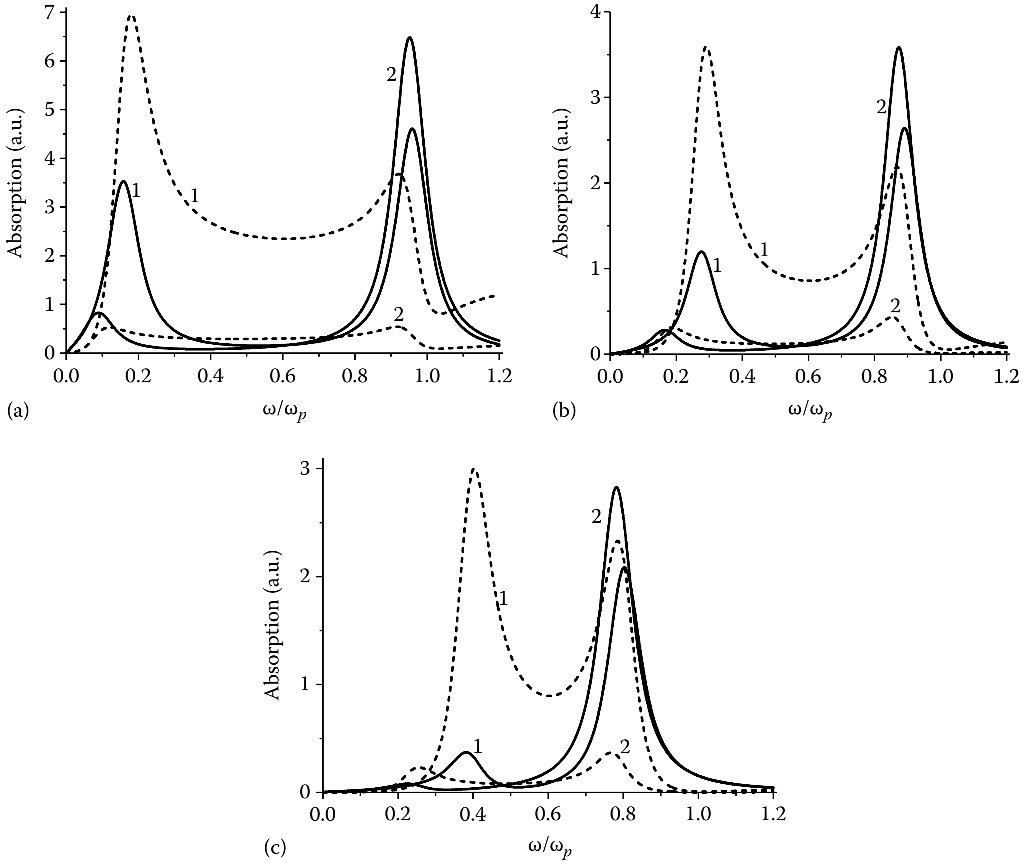

FIGURE 26.10 Absorption spectra of NW with a metal shell for $p = 0.2$ (a), 0.5 (b), 0.8 (c) and $\nu = 0.1\omega_p$ and a semiconductor core with $\varepsilon_1 = 3$ (curves 1) and $\varepsilon_1 = 10$ (curves 2) for light polarized perpendicularly to NWs. Solid curves correspond to the absorption in a metal shell, dashed curves—in a semiconductor core. (From Ruda, H.E. and Shik, A., *Phys. Rev. B*, 72, 115308, 2005a. With permission.)

If the core is a semiconductor with the bandgap $E_g < \hbar\omega$ then, besides the absorption described above, there is also interband absorption in the core proportional to $|\mathbf{E}_1(\mathbf{r})|^2$. The spectrum of this absorption can be easily calculated from Equation 26.27 and is shown by dashed lines in Figure 26.10. The spectrum also has a double maximum structure qualitatively similar to a metal shell one, but contrary to that, suffers a dramatic drop with an increasing ε_1 due to the extrusion of the electric field from the high-ε core. We emphasize that absorptions in a shell and a core depend on different parameters (i.e., the plasma frequency and the scattering rate in metals, and the bandgap width and the interband matrix element in semiconductors) and thus we can not compare the amplitudes of solid and dashed curves and say *a priori* what type of absorption is dominating. Moreover, it is also difficult to distinguish experimentally these two absorption components since the absorption spectra in a shell and a core have qualitatively similar characteristics. The problem can be solved by measuring the excitation spectrum for core luminescence, which is proportional only to the partial absorption in the core shown by the dashed lines in the figure.

As mentioned above, the dashed lines in Figure 26.10 actually show the spectral dependence of the electric field in a core $|\mathbf{E}_1(\mathbf{r})|^2$. As we have shown in Section 26.2.1, the semiconductor and the metal NWs have opposite characteristics of electrical anisotropy: the perpendicular component of the electric field is suppressed as compared with the parallel one, in a semiconductor NW, and enhanced in a metal NW in the vicinity of a plasmon frequency. In composite core–shell structures, both cases can be realized, depending on the shell thickness and the light frequency. The corresponding "phase diagram" using p-ω axes, is shown in Figure 26.11. The areas above the line(s) correspond to the "metal" case $E_\perp > E_{0\perp}$ while those below the line(s)—to the "semiconductor" case $E_\perp < E_{0\perp}$. In other words, these lines correspond to $|A| = 1$ where the amplification factor $A \equiv E_\perp/E_{0\perp}$ is given by Equation 26.27. At a small p, we naturally deal with the "semiconductor" regime. One might be surprised that at $p \to 1$, the regime is not "metal" but should be pointed so that the figure describes the field not in a metal shell, but in a semiconductor core, which in the limit of a pure metal nanowire $p \to 1$ simply disappears.

The condition $|A| > 1$ is satisfied in the vicinity of plasmon resonances, which means that for light-inducing interband absorption in the core, the intensity of the electron-hole pair generation, and hence, the intensity of the luminescence caused by these pairs is larger than in the absence of a metal shell. That is why, for A, we use the term "amplification factor."

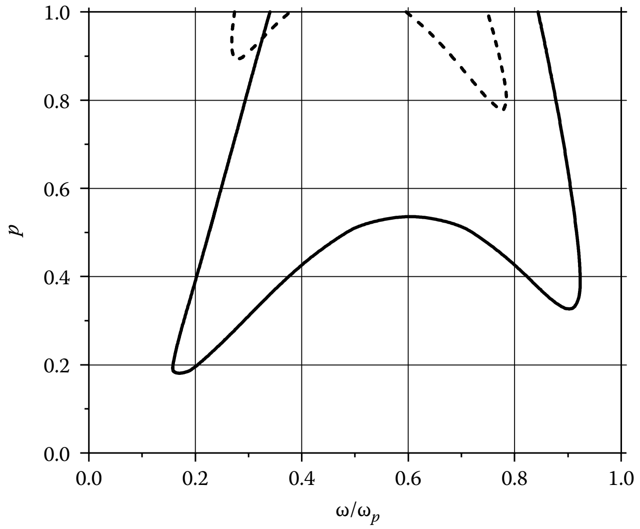

FIGURE 26.11 Values of p and ω, for which $E_{\perp} = E_{0\perp}$ at $\varepsilon_1 = 3$ (solid line) and $\varepsilon_1 = 10$ (dashed lines). The area above the line(s) corresponds to $E_{\perp} > E_{0\perp}$. (From Ruda, H.E. and Shik, A., *Phys. Rev. B*, 72, 115308, 2005a. With permission.)

26.2.5.3 Luminescence Amplification

Metal shells may cause amplification not only of the ac electric field of exciting light but also of the light emitted by a semiconductor core. The ac electric field emitted by a core, which can be considered as an effective dipole $\mathbf{d}_0$, is disturbed by a metal shell so that the field far from the nanostructure looks like one created by some other effective dipole $\mathbf{d}$. By analogy with the previous section, we may expect that in the vicinity of a plasmon resonance, the condition $|d/d_0| \gg 1$ can be realized, which means an effective amplification of the emitted radiation as well.

Calculations show that the luminescence of core–shell structures is characterized by exactly the same rules as absorption. The field of dipoles parallel to the NW axis is not disturbed by image forces, while the field of perpendicular dipoles acquires the same amplification factor A as given by Equation 26.27. As a result, for an isotropic distribution of effective dipoles, such as in a cubic semiconductor core in the absence of size quantization, the emission from an NW acquires a partial polarization parallel to its axis at $|A| < 1$ (as in Section 26.2.3) or perpendicular at $|A| > 1$. Since A has a strong frequency dependence, the components of the emitted light with a different polarization may have different spectra depending on the metal shell parameters.

From the applied point of view, the most important conclusion is as follows: by a proper choice of the metal-shell parameters, we can make one of the plasmon peaks coincide with the luminescence line in the core, which will result in an increase in the luminescence intensity (Ruda and Shik 2005b). It would be even more attractive to enhance the photoluminescence dramatically by a simultaneous amplification of both the exciting and the emitting light. Such a situation often occurs in surface-enhanced Raman scattering (Moskovits 1985) but in luminescence, its realization is more difficult since the difference in frequencies of exciting ω_{ex} and emitted ω_{em} light is usually quite large and cannot be covered by a single plasmon peak. However, in core–shell structures, such double amplification can be reached by using both plasmon peaks as shown in Figure 26.10. In this case, the shell parameters should be chosen in such a way that the low-frequency peak corresponds to ω_{em} in a given NW core while the source of excitation is tuned for ω_{ex} to coincide with the second high-frequency peak (Ruda and Shik 2005a).

26.3 Arrays of Nanorods

26.3.1 Polarization Memory in Random Nanorod Arrays

26.3.1.1 Theory

The whole sections 26.2 was devoted to the optical properties of individual NWs or NRs and now we will discuss similar effects in NW arrays. All the polarization-dependent effects remain the same in such arrays as a system of parallel NWs (say, in porous dielectric matrices) with a relatively large distance between the NWs, but in the case of randomly oriented semiconductor NWs or NRs (say, in a polymer matrix or solution), the situation changes. Such a system has macroscopically isotropic optical properties but simultaneously possesses a very interesting property of polarization memory (Lavallard and Suris 1995, Ruda and Shik 2005a). If we excite photoluminescence in the system using polarized light, in accordance with Section 26.2.1, nonequilibrium carriers will be generated mostly in the NWs that are oriented close to and parallel to the light polarization. According to Section 26.2.3, light emitted by these NWs will have a preferable polarization parallel to their axes, or in other words, parallel to the polarization of the exciting radiation. In other words, an anisotropic random array of nanostructures "remembers" the polarization of the exciting radiation and emits luminescence in the same direction of polarization.

To calculate this effect in a random system of thin NWs, we assume that the NW array occupies the semi-space $x < 0$ and study the polarization of the net luminescence excited by z-polarized light and measured at some point $(x_0, 0, 0)$ outside the array (that is, $x_0 > 0$). The principle of calculation is as follows. We take some NWs with an orientation characterized by the spherical angles (θ, α) and consider it as a line of emitting dipoles with components d_0 along the NW axis and $2d_0\varepsilon_0/(\varepsilon + \varepsilon_0)$ in two other directions. Then, as in Section 26.2.3.1, we find the components $(S_x)_{\parallel}$ and $(S_x)_{\perp}$ of the Poynting vector created by these dipoles at the point $(x_0, 0, 0)$. Since, according to Equation 26.3, the intensity of the absorbed exciting light and hence the intensity of the luminescence depends on θ being proportional to $(\varepsilon + \varepsilon_0)^2\cos^2\theta + 4\varepsilon_0\sin^2\theta$, we integrate $(S_x)_{\parallel}$ and $(S_x)_{\perp}$ over θ and α with this weighting factor obtaining the total intensities $I_{\parallel,\perp}$ of light emission with polarizations parallel and perpendicular to the polarization of the exciting light.

Referring the reader to Ruda and Shik (2005a) for the details of calculations, we present only the final result:

$$\frac{I_\parallel}{I_\perp} = \frac{9[(\varepsilon+\varepsilon_0)^2+2\varepsilon_0^2](\varepsilon+\varepsilon_0)^2+20[(\varepsilon+\varepsilon_0)^2+2\varepsilon_0^2]+21(\varepsilon+\varepsilon_0)^2+336\varepsilon_0^2}{3[(\varepsilon+\varepsilon_0)^2+2\varepsilon_0^2](\varepsilon+\varepsilon_0)^2+44[(\varepsilon+\varepsilon_0)^2+2\varepsilon_0^2]+63(\varepsilon+\varepsilon_0)^2+168\varepsilon_0^2}. \tag{26.29}$$

Equation 26.29 shows that maximal possible polarization, reached at $|\varepsilon/\varepsilon_0| \gg 1$, is equal to $I_\parallel/I_\perp = 3$. In terms of the polarization ratio $P = (I_\parallel - I_\perp)/(I_\parallel + I_\perp)$ it corresponds to $P = 0.5$. Figure 26.12 shows this polarization ratio for an arbitrary relationship between ε and ε_0. For InP NWs in air ($\varepsilon/\varepsilon_0 = 12.7$), we get $P = 0.44$. It is worth noting that at $\varepsilon < \varepsilon_0$ (e.g., in metal NWs near the plasmon resonance), when the electric field in the NWs has a preferably perpendicular orientation, the system also has a polarization memory. This effect is weaker than at a large ε, with the maximal $P = 6/68 = 0.088$.

As we have shown in Sections 26.2.1.2 and 26.2.3.2, for thick NWs, the polarization characteristics for absorption and emission vary dramatically with the light frequency. As a result, the polarization memory in these objects must demonstrate a strong dependence on the frequency of both the exciting (ω_{ex}) and the emitted (ω_{em}) light and even change its sign at some intervals of ω_{ex} and ω_{em}. The quantitative description of this effect (Ruda and Shik 2006) is based on the same approach as that for thin NWs resulting in Equation 26.29 but with two distinctions. First, the values of the longitudinal and the transverse components of the effective dipole moments in the NW are no longer constant but given by Equations 26.12 and 26.13 with $k = \omega_{em}/c$. Second, the intensity of the absorbed exciting light and hence the intensity of the luminescence (the weighting factor in the angular averaging) is now equal to $I_\parallel^a(\omega_{ex})\cos^2\theta + I_\perp^a(\omega_{ex})\sin^2\theta$ where the intensities of the absorbed light $I_\parallel^a$ and $I_\perp^a$ are determined from Equations 26.5 and 26.6 according to the procedure described in Section 26.2.1.2. The final answer for the ratio of luminescence intensities with polarization parallel and perpendicular to that of the exciting light is

$$\frac{I_\parallel^e}{I_\perp^e} = \frac{2I_\perp^a(\omega_{ex})[47d_x^2(\omega_{em})+10d_z^2(\omega_{em})]+3I_\parallel^a(\omega_{ex})[13d_x^2(\omega_{em})+12d_z^2(\omega_{em})]}{4I_\perp^a(\omega_{ex})[16d_x^2(\omega_{em})+11d_z^2(\omega_{em})]+3I_\parallel^a(\omega_{ex})[23d_x^2(\omega_{em})+4d_z^2(\omega_{em})]}. \tag{26.30}$$

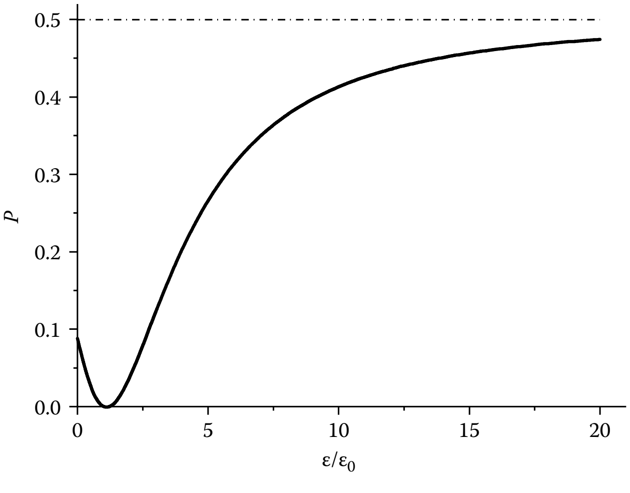

FIGURE 26.12 Luminescence polarization ratio in a random system of nanowires exited by polarized light. The dot-and-dash line corresponds to the limiting value $P = 0.5$. (From Ruda, H.E. and Shik, A., *Phys. Rev. B*, 72, 115308, 2005a. With permission.)

To clarify, we emphasize the difference in notations between Equations 26.29, 26.30 and 26.10 where $I_\parallel^e$ and $I_\perp^e$ referred to the light polarized parallel and perpendicular to the NW axis. Figure 26.13 shows the dependences of the system polarization ratio $P^e = (I_\parallel^e - I_\perp^e)/(I_\parallel^e + I_\perp^e)$ on the frequencies of the exciting

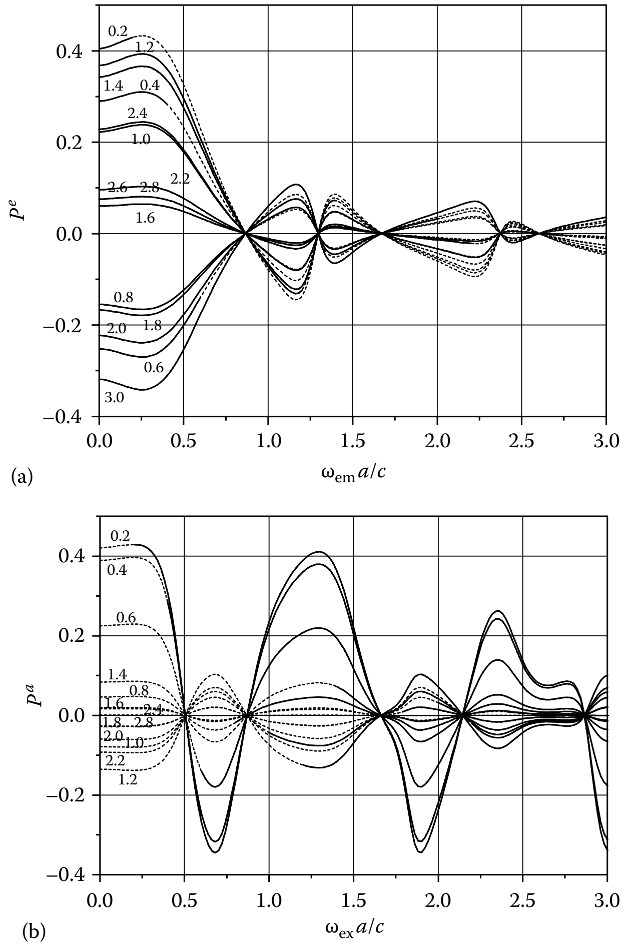

FIGURE 26.13 Polarization ratio of photoluminescence for a random NW system with $\varepsilon = 9$, $\varepsilon_0 = 1$. (a) Luminescence spectra $P^e(\omega_{em}a/c)$ for different ω_{ex}. Numbers at the curves indicate the values ω_{ex}/c. The curves corresponding to $\omega_{ex}a/c = 2.6$ and 2.8 practically coincide. (b) Excitation spectra $P^a(\omega_{ex}a/c)$ for different ω_{ems}. Numbers at the curves indicate the values of $\omega_{em}a/c$. The curves corresponding to $\omega_{em}a/c = 1.8$ and 2.8 practically coincide. For $\omega_{em}a/c = 2.6$ polarization is very close to zero and is not presented in the figure.

and the emitted light, given by Equation 26.30. Since in standard photoluminescence measurements ω_{ex} always exceeds ω_{em}, parts of the curves in Figure 26.13 where this condition is violated are given by dashed lines. At ω_{ex}, $\omega_{em} \to 0$, polarization tends to its low-frequency limit shown in Figure 26.12 and for our choice $\varepsilon/\varepsilon_0 = 9$ equal to $P^e \cong 0.40$.

It is important to mention that if the NW material is characterized by several luminescence lines with essentially different frequencies, then these lines may have different degrees of polarization, in accordance with Figure 26.6. Since the curve at the figure crosses the level $I^e_{\parallel} = I^e_{\perp}$, there exists even the possibility when one of the luminescent lines may have. The latter case could be called “polarization anti-memory.”

26.3.1.2 Experiment

The effect of polarization memory was first observed in porous Si (Kovalev et al. 1995) that has only a distant resemblance to a system of NWs or NRs. Hence, for demonstration of its main experimental features, we have chosen subsequently more detailed measurements in CdSe/ZnS NRs (Kravtsova et al. 2007) with the aspect ratio $l/2a \cong 2.4$ dissolved in a liquid and hence having a randomly distributed orientation. Figure 26.14 shows the luminescence spectrum consisting of two peaks corresponding to interband transitions between different size-quantized states. Polarization properties of the emitted light are presented in the form of luminescence spectra for different angular positions of the analyzer related to the direction of the exciting light polarization $\mathbf{E_0}$. The luminescence is seen to be polarized mostly parallel to $\mathbf{E_0}$, in qualitative agreement with theoretical predictions.

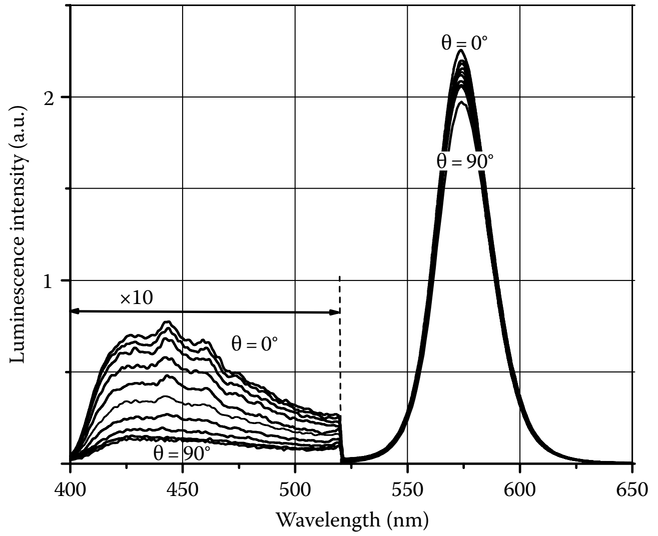

FIGURE 26.14 Photoluminescence spectrum of the NR solution for excitation by linearly polarized light with $\lambda = 366$ nm. Different curves correspond to the luminescence components with a polarization in the direction forming the angle θ with the polarization of excitation. θ increases from the upper to the lower curve with the step 10°. The amplitude of the luminescence in the spectral region $\lambda < 520$ nm is shown with the magnification factor 10. (From Kravtsova, Y. et al., *Appl. Phys. Lett.*, 90, 083120, 2007. With permission.)

The quantitative treatment of the results meets with some difficulties. According to Equation 26.29, the amplitude of polarization memory in small nanostructures has some universal value depending only on $\varepsilon/\varepsilon_0$. However, in the experiment, this amplitude differs noticeably for two different spectral lines in the same NRs. Besides, for the short wavelength peak, the amplitude of polarization memory exceeded the theoretically predicted one, while for the long wavelength peak, it was always less. These results could be explained if the matrix element responsible for the ground state 573 nm-long wavelength peak is larger for perpendicular light polarization (it is really the case for NRs with a small aspect ratio (Hu et al. 2001)) while that at 470 nm is either more isotropic or larger for parallel polarization. This is in agreement with the theoretical (Li and Wang 2003) and the experimental (Thomas et al. 2005) conclusions of essentially different polarization properties of different optical transitions in NRs.

26.3.2 Plasmon Spectra in Self-Assembled Nanorod Structures

26.3.2.1 Self-Assembling of Metal Nanorods

In the recent years, an impressive success has been achieved in fabricating various arrays of metal NRs by their self-assembling in solution. The most comprehensive results in this direction were obtained by using Au NRs with the lateral surface covered by a double layer of hydrophilic molecules and the ends terminated with polystyrene molecular chains (Nie et al. 2007, 2008). Such building blocks in solution can form different ordered arrays, such as chains, bundles, raft-like structures, rings, etc. (Figure 26.15), depending on the chemical composition of the solvent. In these systems, the distances between the NRs are very small, compared to their diameter, so that the electric fields of plasmons in neighboring NRs overlap and the plasmon spectrum of the system becomes different from that of individual NRs and depends on the NR array topology. Figure 26.16 shows the plasmon absorption spectra in some of these structures formed from identical NRs. They consist of two peaks (longitudinal and transverse plasmons) where the long wavelength one, corresponding to longitudinal plasmons (see Section 26.2.2.1), demonstrates a strong dependence on the array geometry.

Some general qualitative regularities were already formulated in literature and found their experimental confirmation in systems of NRs and nanodots. It was demonstrated (Rechenberger et al. 2003, Su et al. 2003) that by reducing the inter-nanoparticle separation in the direction of light polarization, a red shift in the plasmon resonance occurs, in contrast with a blue shift for the perpendicular direction of polarization. It has a simple qualitative explanation. The end-to-end alignment of NRs increases the effective length of an NR, which results in a decrease of the longitudinal depolarization factor $n_{\parallel}$ and, according to Equation 26.7, a decrease in the longitudinal plasmon frequency $\omega_{\parallel}$, while the side-to-side arrangement increases the effective $n_{\parallel}$ and $\omega_{\parallel}$.

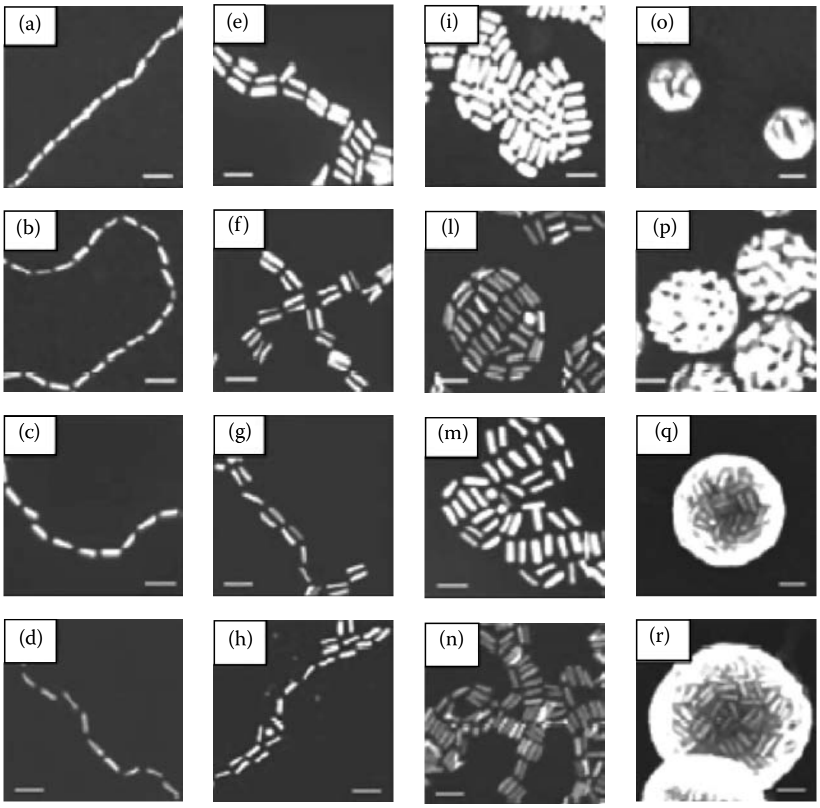

FIGURE 26.15 NR arrays obtained by self-assembling in different solvents. (From Fava, D. et al., *Adv. Mater.*, 20, 4318, 2008. With permission.)

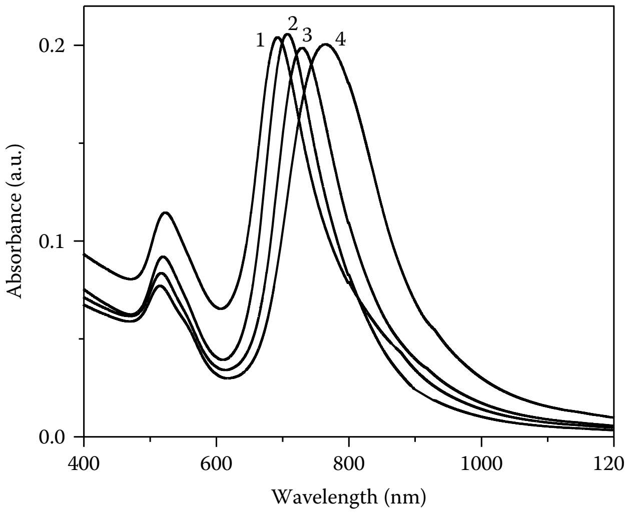

FIGURE 26.16 Optical absorption spectra for various NR arrays. (From Nie, Z.H. et al., *J. Am. Chem. Soc.*, 130, 3683, 2008. With permission.)

In the next sections, we present theoretical models allowing one to make some quantitative calculations of plasmon characteristics and their comparison with the experiment.

26.3.2.2 Longitudinal Plasmon Shift

Let us consider some arbitrary system of metal NWs or strongly anisotropic NRs (further we will mention them simply as NWs) and find the polarization of each NW in an external uniform electric field $\mathbf{E}_0$. We assume that the NW radius a is many times smaller than both their length l, so that for calculations of longitudinal plasmons, the radial distribution of the induced charges is inessential and we may characterize each NW by a linear charge density $q(z)$ depending on one single coordinate z measured along the NW axis. In the absence of a direct electrical contact between the NWs, their total charges remain zero: $\int_{-l/2}^{l/2} q(z)\,dz = 0$. For each metal NW, the charge distribution $q(z)$ is found from the condition of equipotentiality (Ruda and Shik 2007b). It means that the external field potential $-E_{0z}z$ is exactly compensated by the joint potential created by the charges of all NWs including the given one. To describe the potential created by $q(z)$ of the same NW in a medium with the dielectric constant ε, we use the approximate expression $2q(z)\ln(l/a)/\varepsilon$ valid for $l \gg a$ (Averkiev and Shik 1996). Potentials of all other NWs can be obtained by a simple Coulomb expression neglecting the finite thickness of NWs.

We demonstrate the described procedure for the case of two in-line NWs separated by the interval b and occupying the segments $(-l - b/2, -b/2)$ and $(b/2, l + b/2)$ of the z-axis. Noting the charge densities in the NWs as $q_-(z)$ and $q_+(z)$, respectively, we obtain the system of two equations:

$$\frac{2q_{\pm}(z)}{\varepsilon}\ln\left(\frac{l}{a}\right) - E_{0z}z + \int_{\mp b/2}^{\mp b/2 \mp l} \frac{q_{\mp}(z')}{\varepsilon(z'-z)}\,dz' = C_{\pm}. \qquad (26.31)$$

Instead of $q_{\pm}$, we introduce two dimensionless functions $\eta_{\pm}(\xi)$ determined from

$$q_{\pm}(z) = \frac{\varepsilon E_{0z}}{2\ln(l/a)}\left[z + l\eta_{\pm}(\xi)\right] \tag{26.32}$$

where ξ is the dimensionless coordinate measured from the center of the corresponding NWs: $\xi = [z \mp (b/2 + l/2)]/l$. It can be easily shown that $\eta_{-}(-\xi) = -\eta_{+}(\xi)$. Finding $C_{\pm}$ from the neutrality conditions: $\int_{-1/2}^{1/2}\eta_{\pm}(\xi)\,d\xi = 0$, we reduce Equation 26.31 to one single equation for $\eta_{+}(\xi)$:

$$\begin{aligned} &2\ln\left(\frac{l}{a}\right)\eta_{+}(\xi) - \int\limits_{-1/2}^{1/2}\frac{\eta_{+}(\xi')d\xi'}{\xi+\xi'+1+\beta} + \int\limits_{-1/2}^{1/2}\eta_{+}(\xi')\ln\left(\frac{2\xi'+3+2\beta}{2\xi'+1+2\beta}\right)d\xi' \\ &= \frac{1}{2} + \ln\left(\frac{2+\beta}{1+\beta}\right) - (\xi+1+\beta)\ln\left(\frac{2\xi+3+2\beta}{2\xi+1+2\beta}\right) \\ &+ \frac{\beta}{2}\ln\left[\frac{\beta(2+\beta)^3}{(1+\beta)^4}\right] + \frac{\beta^2}{2}\ln\left[\frac{\beta(2+\beta)}{(1+\beta)^2}\right]. \end{aligned} \tag{26.33}$$

This equation was solved numerically for $l/a = 10$, which corresponds to the experimental results (Nie et al. 2007) to be discussed later. After $q(z)$ is found, the total dipole moment P of an NW is easily found: $P = \int zq(z)\,dz$, which, in turn, determines the depolarization factor $n_{\parallel} = E_{0z}V/(4\pi P)$ (V is the NR volume) (Landau and Lifshitz 1984), connected with the longitudinal plasmon frequency $\omega_{\parallel}$ by Equation 26.7.

Using a similar approach, the frequency $\omega_{\parallel}$ was found (Ruda and Shik 2007b) for two other systems: the infinite periodic NR chain and a pair of two side-by-side NRs with the distance $d > 2a$ between the axes. The resulting dependences of the longitudinal plasmon frequency $\omega_{\parallel}$ on inter-NR distances b or d for all three cases are presented in Figure 26.17. It is seen that, in accordance with the above-mentioned qualitative considerations, $\omega_{\parallel}$ demonstrates a red shift for chains and a blue shift for bunches of NRs. Figure 26.17 also contains experimental points (Nie et al. 2007) partially taken from the curves of Figure 26.16. It is seen that the results for in-line structures (curves 1 and 2) demonstrate a good qualitative agreement but the experimental dependences are steeper than the theoretical. This effect can be caused by the fact that the polystyrene molecules connecting the NR tips are also polarized in the external electric field, so that the effective local dielectric constant of this gap exceeds that of the solvent. This distorts the electric field pattern and increases the intensity of inter-NR interaction and hence, the value of P, as compared to the ideal case considered above. For the side-by-side NR assemblies, practically all the lines of the electric field responsible for the inter-NR interaction are concentrated in the region filled by hydrophilic molecules, so that the dielectric inhomogeneity of the system is irrelevant, and curve 3 demonstrates not only qualitative but also quantitative agreement.

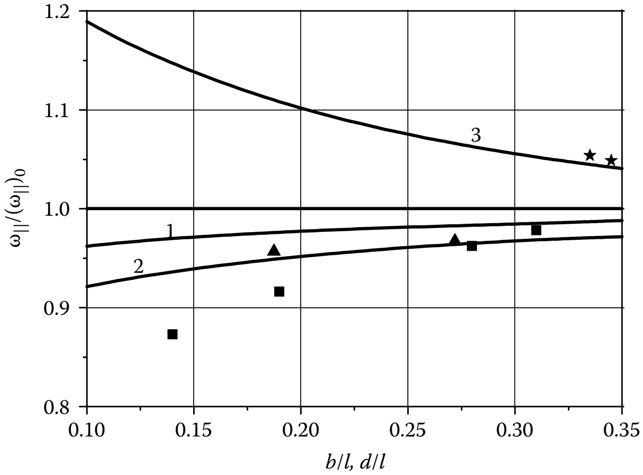

FIGURE 26.17 Dependence of the longitudinal plasmon frequency $\omega_{\parallel}$ (in units of $\omega_{\parallel}$ for individual NRs) on the distance between NRs with $l/a = 10$. 1, two in-line NRs (experimental points—triangles); 2, chain of NRs (experimental points—squares); 3, two parallel NRs (experimental points—stars).

26.4 Conclusion

We have described a large variety of polarization-sensitive optical phenomena in anisotropic nanostructures, such as NWs and NRs. The phenomena have a universal character being exclusively due to the difference in the dielectric constants between nanostructures and their environment. They include a strong dependence of linear (absorption, photoluminescence, photoconductivity, etc.) and nonlinear optical properties on polarization of the exciting light, as well as a high degree of polarization of luminescence emitted by these systems. The joint action of absorption and emission anisotropy results in a polarization memory in random arrays of NWs and NRs. In metal nanostructures, all the mentioned phenomena exhibit a dramatic dependence (including sometimes the change of sign) on the light frequency, related to plasmon effects. In composed metal-semiconductor nanostructures, these effects can be used for luminescence amplification. In ordered self-assembled arrays of metal NRs, the plasmon-related optical spectra are sensitive to the array geometry. All the described effects are provided with proper theoretical description and illustrated by experimental observations performed in recent years.

Acknowledgments

The authors would like to thank their collaborators, V. Barzda, R. Cisek, D. Fava, E. Kumacheva, Y. Kravtsova, U. Krull, L. Levina, S.F. Musikhin, Z.H. Nee, U. Philipose, and T.L. Spencer from the University of Toronto, whose experimental results are presented in this work.

References

Agarwal, G.S. and Jha, S.S. 1982. Theory of second harmonic generation at a metal surface with surface plasmon excitation. *Solid State Communications* 41: 499–501.

Averkiev, N.S. and Shik, A. 1996. Contact phenomena in quantum wires and porous silicon. *Semiconductors* 30: 112–116.

Barzda, V., Cisek, R., Spencer, T.L., Philipose, U., Ruda, H.E., and Shik, A. 2008. Giant anisotropy of second harmonic generation for a single ZnSe nanowire. *Applied Physics Letters* 92: 113111.

Batygin, V.V. and Toptygin, I.N. 1978. *Problems in Electrodynamics*. New York: Academic Press.

Blombergen, N., Chang, R.K., Jha, S.S., and Lee, C.H. 1968. Optical second-harmonic generation in reflection from media with inversion symmetry. *Physical Review* 174: 813–822.

Chew, W.C. 1990. *Waves and Fields in Inhomogeneous Media*. New York: Van Nostrand Reinhold.

Choi, H.-J., Johnson, J.C., He, R. et al. 2003. Self-organized GaN quantum wire UV lasers. *Journal of Physical Chemistry B* 107: 8721–8725.

Duan, X. and Lieber, C.M. 2000, General synthesis of compound semiconductor nanowires. *Advanced Materials* 12: 298–302.

Fan, Z., Chang, P., Lu, J.G. et al. 2004. Photoluminescence and polarized photodetection of single ZnO nanowires. *Applied Physics Letters* 85: 6128–6130.

Fava, D., Nie, Z.H., Winnik, M.A., and Kumacheva, E. 2008. Evolution of self-assembled structures of polymer-terminated gold nanorods in selective solvents. *Advanced Materials* 20: 4318–4322.

Grigoriev, I.S. and Meilikhov, E.Z. 1997. *Handbook of Physical Quantities*. Boca Raton, FL: CRC Press.

Han, S., Jin, W., Zhang, D. et al. 2004. Photoconduction studies on GaN nanowire transistors under UV and polarized UV illumination. *Chemical Physics Letters* 389: 176–180.

Hsu, N.E., Hung, W.K., and Chen, Y.F. 2004. Origin of defect emission identified by polarized luminescence from aligned ZnO nanorods. *Journal of Applied Physics* 96: 4671–4673.

Hu, J., Li, L., Yang, W., Manna, L., Wang, L., and Alivisatos, A.P. 2001. Linearly polarized emission from colloidal semiconductor quantum rods. *Science* 292: 2060–2063.

Kovalev, D., Ben Chorin, M., Diener, J. et al. 1995. Porous Si anisotropy from luminescence polarization. *Applied Physics Letters* 67: 1585–1587.

Kravtsova, Y., Krull, U., Musikhin, S.F., Levina, L., Ruda, H.E., and Shik, A. 2007. Polarization memory in a system of CdSe nanorods. *Applied Physics Letters* 90: 083120.

Kreibig, U. and Vollmer, M. 1995. *Optical Properties of Metal Clusters*. Berlin, Germany: Springer.

Landau, L.D. and Lifshitz, E.M. 1984. *Electrodynamics of Continuous Media*. New York: Pergamon Press.

Lauhon, L.J., Gudiksen, M.S., Wang, D., and Lieber, C.M. 2002. Epitaxial core-shell and core-multishell nanowire heterostructures. *Nature* 420: 57–61.

Lavallard, P. and Suris, R.A. 1995. Polarized photoluminescence of an assembly of non cubic microcrystals in a dielectric matrix. *Solid State Communications* 95: 267–269.

Li, J. and Wang, L. 2003. High energy excitations in CdSe quantum rods. *Nano Let*ters 3: 101–105.

McIntyre, C.R. and Sham, L.J. 1992. Theory of luminescence polarization anisotropy in quantum wires. *Physical Review B* 45: 9443–9446.

Mei, X., Blumin, M., Sun, M. et al. 2003. Highly-ordered GaAs/AlGaAs quantum dot arrays on GaAs(100) substrate grown by molecular beam epitaxy using nanochannels alumina masks. *Applied Physics Letters* 82: 967–969.

Moskovits, M. 1985. Surface-enhanced spectroscopy. *Reviews of Modern Physics* 57: 783–826.

Neeves, A.E. and Birnboim, M.H. 1989. Composite structures for the enhancement of nonlinear optical susceptibility. *Journal of the Optical Society of America B* 6: 787–796.

Nie, Z.H., Fava, D., Kumacheva, E. et al. 2007. Self-assembly of metal-polymer analogues of amphiphilic triblock copolymers. *Nature Materials* 6: 609–614.

Nie, Z.H., Fava, D., Rubinstein, M., and Kumacheva, E. 2008. Supramolecular assembly of gold nanorods end-terminated with polymer "pom-poms". *Journal of the American Chemical Society* 130: 3683–3689.

Oldenburg, S.J., Averitt, R.D., Westcott, S.L., and Halas, N.J. 1998. Nanoengineering of optical resonances. *Chemical Physics Letters* 288: 243–247.

Peng, X., Manna, L., Yang, W. et al. 2000. Shape control of CdSe nanocrystals. *Nature* 404: 59–61.

Philipose, U., Ruda, H.E., Shik, A., de Souza, C.F., and Sun, P. 2005. Light emission from ZnSe nanowires. *Proceedings of SPIE* 5971: 597116–597128.

Qi, J., Belcher, A.M., and White, J.M. 2003. Spectroscopy of individual silicon nanowires. *Applied Physics Letters* 82: 2616–2618.

Rechenberger, W., Hohenau, A., Leitner, A. et al. 2003. Optical properties of two interacting gold nanoparticles. *Optics Communications* 220: 137–141.

Ruda, H.E. and Shik, A. 2005a. Polarization-sensitive optical phenomena in semiconducting and metallic nanowires. *Physical Review B* 72: 115308.

Ruda, H.E. and Shik, A. 2005b. Plasmon phenomena and luminescence amplification in nanocomposite structures. *Physical Review B* 71: 245328.

Ruda, H.E. and Shik, A. 2006. Polarization-sensitive optical phenomena in thick semi-conducting nanowires. *Journal of Applied Physics* 100: 024314.

Ruda, H.E. and Shik, A. 2007a. Nonlinear optical phenomena in nanowires. *Journal of Applied Physics* 101: 034312.

Ruda, H.E. and Shik, A. 2007b. Polarization and plasmon effects in nanowire arrays. *Applied Physics Letters* 90: 223106.

Sercel, P.C. and Vahala, K.J. 1990. Analytical technique for determining the polarization dependence of optical matrix elements in quantum wires with band-coupling effects. *Applied Physics Letters* 57: 545–547.

Su, K.-H., Wei, Q.-H., Zhang, X. et al. 2003. Interparticle coupling effects on plasmon resonances in nanogold particles. *Nano Letters* 3: 1087–1090.

Thomas, N.L., Herz, E., Schöps, O., Woggon, U., and Artemyev, M.V. 2005. Exciton fine structure in single CdSe nanorods. *Physical Review Letters* 94: 016803.

Wang, J., Gudiksen, M.S., Duan, X., Cui, Y., and Lieber, C.M. 2001. Highly polarized luminescence and photodetection from single InP nanowires. *Science* 293: 1455–1457.

Zhou, H.S., Honma, I., Komiyama, H., and Haus, J.W. 1994. Controlled synthesis and quantum-size effect in gold-coated nanoparticles. *Physical Review B* 50: 12052–12056.

27

Nonlinear Optics with Clusters

Sabyasachi Sen
JIS College of Engineering

Swapan Chakrabarti
University of Calcutta

27.1 General Introduction: Fundamentals of Nonlinear Optics and Different Nonlinear Optical Effects **27**-1
27.2 Quantum Formulation of Nonlinear Optics **27**-2
27.3 Nonlinear Optics with Materials **27**-5
27.4 Why Are Clusters So Important? **27**-5
27.5 Review of Nonlinear Optics with Clusters **27**-5
27.6 Conclusions **27**-12
References **27**-12

27.1 General Introduction: Fundamentals of Nonlinear Optics and Different Nonlinear Optical Effects

In recent years, nonlinear optical (NLO) materials have attracted substantial research interests due to their potential applications in information processing and telecommunications [1–4]. NLO processes are particularly useful in photonic devices where photons are used to transmit and process information. Modification in the optical properties of a material by the strong oscillating electric field of a laser beam is in general discussed in the study of NLO. NLO processes have been observed in a wide variety of materials starting from molecules and clusters to bulk materials [5–14]. In molecular systems, the optical nonlinearity depends on the geometrical arrangement of molecules, whereas in case of bulk systems, the electronic characteristics of the system determine the optical nonlinearity. Various experimental as well as theoretical results have reported that NLO properties of small atomic clusters depend on the geometry and the electronic structure property of these materials [15,16].

In general, the electric dipole moment μ_a of a molecular system is expressed as

$$\begin{aligned}\mu_a &= \mu_a(E^b(\vec{r},t) = E^c(\vec{r},t) = E^d(\vec{r},t)\ldots. = 0)\\ &\quad + \sum_b \alpha_{ab} E^b(\vec{r},t) + \frac{1}{2!}\sum_{bc} \beta_{abc} E^b(\vec{r},t) E^c(\vec{r},t)\\ &\quad + \frac{1}{3!}\sum_{bcd} \gamma_{abcd} E^b(\vec{r},t) E^c(\vec{r},t) E^d(\vec{r},t) + \cdots, \end{aligned} \tag{27.1}$$

In the above equation, $E^a(\vec{r},t)$, $E^b(\vec{r},t)$, …, are the external electric fields consisting of a monochromatic and a static part, and a, b, …. are equal to the Cartesian directions x, y, or z. The first term in Equation 27.1 denotes the dipole moment of the system. α_{ab}, β_{abc}, and γ_{abcd} represent the elements of the linear polarizability tensor, the first-order hyperpolarizability tensor, and the second-order hyperpolarizability tensor, respectively. The time dependence of the dipole moment leads to various frequency-dependent polarizabilities and hyperpolarizability tensors. The first-order hyperpolarizability tensor is described in terms of static first-order hyperpolarizability [β (0; 0, 0)], second-harmonic generation (SHG) [β (−2ω; ω, ω)], electro-optical Pockels effect (EOPE) [β (−ω; ω, 0)], optical rectification (OR) [β (0; ω, −ω)], sum-frequency generation (SFG), and difference-frequency generation (DFG). Similarly, in the case of the second-order hyperpolarizability tensor, apart from the static part [γ (0; 0, 0, 0)], there exists frequency-dependent components, namely, the third-harmonic generation (THG) [γ (−3ω; ω, ω, ω)], the electro-optical Kerr effect (EOKE) [γ (−ω; ω, 0, 0)], the dc-induced second-harmonic generation (dc-SHG or electric field-induced second-harmonic generation (EFISH)) [γ (−2ω; ω, ω, 0)], and the intensity-dependent refractive index (IDRI) [γ (−ω; ω, −ω, ω)] or the degenerate four-wave mixing (DFWM). Among the different NLO properties, the first-order hyperpolarizability is important because it is the simple third-order molecular property and responsible for phenomena such as the SHG and the EOPE. On the other hand, the second-order hyperpolarizability tensor is related to physical properties like the EOKE. It is also significant in the perspective of quantum optical effects like self-focusing and self-defocusing effects in a material.

In the presence of an electric field of field strength

$$E(t) = E_1 e^{-i\omega_1 t} + E_2 e^{-i\omega_2 t} + c.c, \tag{27.2}$$

the second-order nonlinear polarization induced in a crystal is expressed as

$$P^{(2)}(t) = \chi^{(2)} E(t)^2, \tag{27.3}$$

$$\begin{aligned} P^{(2)}(t) = \chi^{(2)}[E_1^2 e^{-2i\omega_1 t} + E_2^2 e^{-2i\omega_2 t} + 2E_1E_2 e^{-i(\omega_1+\omega_2)t} \\ + 2E_1E_2^{\star} e^{-i(\omega_1-\omega_2)t} + c.c] + 2\chi^{(2)}[E_1E_1^{\star} + E_2E_2^{\star}], \end{aligned} \tag{27.4}$$

The first and second terms in Equation 27.4 represent the SHG, which is useful in the perspective of optoelectronic phenomenon such as tuning the frequency of the laser beam. The expression $2E_1E_2e^{-i(\omega_1+\omega_2)t}$ and $2E_1E_2^{\star}e^{-i(\omega_1-\omega_2)t}$ are associated with the sum-frequency generation (SFG) and the difference-frequency generation (DFG) respectively. The SFG is used to produce tunable radiation in the ultraviolet spectral region and the DFG produces tunable radiation in the infrared region. The process of the DFG is also used in the design of the optical parametric oscillator. The last term of Equation 27.4 does not show any frequency dependence and leads to a process known as optical rectification (OR).

The third order contribution to the nonlinear polarization is given by

$$P^{(3)}(t) = \chi^{(3)} E(t)^3, \tag{27.5}$$

In the case of the monochromatic applied electric field ($E(t) = \xi \cos \omega t$), the nonlinear polarization is expressed as

$$P^{(3)}(t) = \frac{1}{4}\chi^{(3)}\xi^3 \cos 3\omega t + \frac{3}{4}\chi^{(3)}\xi^3 \cos \omega t. \tag{27.6}$$

The first term of Equation 27.6 illustrates the third-harmonic generation (THG), which describes the response of the system at the frequency of 3ω. The second term in Equation 27.6 leads to a nonlinear contribution to the refractive index of the electromagnetic wave of frequency ω. This is recognized as the intensity-dependent refractive index (IDRI). In the presence of such nonlinearity, the refractive index becomes

$$n = n_0 + n_2 I, \tag{27.7}$$

where n_0 is the linear refractive index and

$$n_2 = \frac{12\pi^2}{n_0^2 c}\chi^{(3)}, \tag{27.8}$$

n_2 is an optical constant that characterizes the strength of the optical nonlinearity. I is the intensity of the incident wave expressed as $I = (n_0 c/8\pi)\xi^2$. The intensity-dependent refractive index (IDRI) is related to optical phenomenon, such as the self-focusing of the electromagnetic wave. Self-focusing occurs in a material in which n_2 is positive and the electromagnetic waves, with a nonuniform transverse intensity distribution, curve towards each other after passing through such material.

27.2 Quantum Formulation of Nonlinear Optics

A large number of quantum chemical techniques have been employed for the evaluation of the NLO properties of materials. Jensen et al. have developed a dipole interaction model. This model successfully describes the response properties of large aggregates of molecular clusters [17]. The density-matrix renormalization group (DMRG)-based formalism was proposed by Pati and coworkers and successfully applied in the description of the frequency-dependent NLO properties of π-conjugated systems [18]. Standard techniques like the time-density functional theory, the time-dependent Hartee Fock technique, as well as the finite-field methods [19] based on DFT are also very successful in describing the NLO responses of several classes of materials. The quantum chemical approach, like the density functional theory, had its limitations in the case of long-range dispersion interaction and has been improved by the introduction of the dispersion-corrected DFT DFD-D [20,21] technique. Higher-order wavefunction-based methods like the coupled cluster approach and the MP2 technique are also efficient but they are excessively expensive. In addition to these, the response properties are also evaluated through semi-empirical methods like ZINDO/MRDCI (multi reference doubles-configuration interaction) formalism with a correction vector. [22]. Another well-celebrated quantum chemical approach is the use of the two-state model proposed by Oudar and Chemla [23]. This model relates the nonlinear optical coefficients to the excited-state energy, the oscillator strength and the dipole moment of the molecule under consideration. The theoretical results of Zyss [24] and Andrews et al. [25] successfully established the two-state model of Oudar and Chemla. Datta and Pati [26] employed the two-state model to discuss the role of dipolar interaction and H-bonding in tuning the response properties in molecular aggregates. This model is simple and particularly useful in the case of charge-transfer systems [27]. In this two-state model, the expression of β is given by

$$\beta_{\text{two-level}} = \frac{3e^2}{2\hbar}\frac{\omega_{01} f \Delta\mu_{01}}{(\omega_{01}^2 - \omega^2)(\omega_{01}^2 - 4\omega^2)}$$

where

f is the oscillator strength
ω is the applied frequency
ω_{01} is the frequency of the optical transition between the ground state and the first dipole-allowed state
$\Delta\mu_{01}$ is the difference in moments between the ground state and the first dipole-allowed state

In the absence of an applied frequency, the above equation results in a first-order static hyperpolarizability tensor

$$\langle \beta(0;0,0) \rangle = 3\pi \frac{f \Delta\mu_{01}}{\omega_{01}^3}.$$

In two-state formalism, the parameters like f, ω_{01} and $\Delta\mu_{01}$ are estimated through standard excited approaches like the configuration interaction scheme (CIS) or the time-dependent density-functional theory (TDDFT).

Apart from all those techniques described in the above paragraph, another widely accepted quantum chemical technique of response calculation is the analytic approach. In this context, it will be useful to point out some of the basic differences between the finite-difference technique and the analytic approach. In the finite-difference technique, the energy is calculated at different values of the electric field; there after, the finite differentiation is performed to obtain the derivative of energies. The finite-difference method is suitable for any program capable of energy calculation in a perturbed system. However, the main limitation of this method is the prolonged computational time and it is not good enough when higher order derivatives of the energy are involved. On the contrary, in the analytical approach, physical properties are obtained by analytically evaluating the derivatives of the energy and are often preferred in response calculations. The main advantage of this technique is that it gives access to frequency-dependent properties and the data obtained are more accurate. This technique has been successfully applied by van Gisbergen et al. [28] in case of a set of small molecules. Sen et al. employed the same technique in case of CdSe clusters and Al_4M_4 clusters [11,13].

In the density-functional theoretical formulation of the analytic approach, the starting equation is a variation on the time-dependent Kohn–Sham (TDKS) equation. In order to obtain the starting equation of the analytic approach in the time-dependent density-functional theory (TDDFT), the action integral of the TDDFT is considered.

$$A=\int_{t_0}^{t_1}\mathrm{d}t\left\langle\psi(t)\right|i\frac{\partial}{\partial \mathrm{t}}-\hat{H}(t)\left|\psi(t)\right\rangle,\tag{27.9}$$

where ψ is the wave function of the system. In terms of time-dependent single-particle orbitals $\{\varphi_j(\vec{r},t)\}$, the action integral becomes

$$\begin{aligned}A[\{\varphi_j\}]=\sum_{n}^{j}\int_{-\infty}^{t_1}\mathrm{d}t\int\mathrm{d}^3\vec{r}\varphi_j^*(\vec{r},t)\left(i\frac{\partial}{\partial t}+\frac{\nabla^2}{2}\right)\varphi_j(\vec{r},t)\\ -\int_{-\infty}^{t_1}\mathrm{d}t\int\mathrm{d}^3\vec{r}\rho(\vec{r},t)v_{\text{ext}}(\vec{r},t)\\ -\frac{1}{2}\int_{-\infty}^{t_1}\mathrm{d}t\int\mathrm{d}^3\vec{r}\int\mathrm{d}^3\vec{r}'\frac{\rho(\vec{r},t)\rho(\vec{r}',t)}{|\vec{r}-\vec{r}'|}-A_{xc}[\{\varphi_j\}],\end{aligned}\tag{27.10}$$

In Equation 27.10, the term $v_{\text{ext}}(\vec{r},t)$ is the time-dependent exchange-correlation potential and A_{xc} is the exchange-correlation part of the action functional. On considering the orthonormal Kohn–Sham orbital, the action integral produces the time-dependent Kohn–Sham equation as

$$\sum_j\varepsilon_{ij}(t)\varphi_j(\vec{r},t)+i\frac{\partial}{\partial t}\varphi_i(\vec{r},t)=\left[-\frac{\nabla^2}{2}+\nu_s(\vec{r},t)\right]\varphi_i(\vec{r},t)=F_s\varphi_i(\vec{r},t),\tag{27.11}$$

In Equation 27.11, $\nu_s(\vec{r},t)$ is the time-dependent Kohn–Sham potential and $\varepsilon_{ij}(\vec{r},t)$ is the Lagrangian multiplier. Different choices of Lagrangian multipliers are allowed. With $\varepsilon_{ij}(\vec{r},t)=0$ we attain the canonical form of the Kohn–Sham equation. In the case of orbitals varying rapidly with time, $\varepsilon_{ij}(\vec{r},t)$ is suitably chosen to avoid an unphysical divergence in the Kohn–Sham equation. The time-dependent density $\rho(\vec{r},t)$ is obtained from the relation

$$\rho(\vec{r},t)=\sum_i^{\text{occupied}}\left|\varphi_i(\vec{r},t)\right|^2,\tag{27.12}$$

When $\varphi_i(\vec{r},t)$ is expanded in a fixed time-dependent basis set of atomic orbitals $\{\chi_\mu(\vec{r})\}$, the resulting expression becomes

$$\varphi_i(\vec{r},t)=\sum_\mu\chi_\mu(\vec{r})C_\mu(t),\tag{27.13}$$

In Equation 27.13, the coefficient $C_\mu(t)$ manifests the time dependence of $\varphi_i(\vec{r},t)$. Finally, Equations 27.11 through 27.13 produces the time-dependent Kohn–Sham equation in the matrix form given by

$$F_sC-i\frac{\partial}{\partial t}SC=SC\varepsilon,\tag{27.14}$$

where S is the overlap matrix of the atomic orbitals with

$$S_{\mu\nu}=\int\mathrm{d}\vec{r}\chi_\mu^*(\vec{r})\chi_\nu(\vec{r}),\tag{27.15}$$

The density matrix (D) is expressed in terms of the coefficient matrix C and the occupation number matrix n and is expressed as

$$D=CnC^\dagger.\tag{27.16}$$

The Kohn–Sham equation expressed in Equation 27.14 is the starting point of the perturbative expansion. Equation 27.14 is similar to the starting equation for the NLO calculation in the time-dependent Hartree–Fock technique (TDHF). In Equation 27.14, the term F_s represents the Kohn–Sham matrix or the Fock matrix. This is expressed as

$$F_s=h+D\mathrm{X}\,2J+v_{xc}.\tag{27.17}$$

In the above equation, h is the one-electron integral matrix comprising kinetic energy, the Coulomb field of the nuclei, and the applied external electric field. J is the four-index Coulomb supermatrices.

When all matrices are expanded in different orders of external perturbation and ε is chosen as the time-dependent zero-order matrix, the canonical Kohn–Sham equation for the ground-state DFT is obtained and given by the following expression:

$$F_s^{(0)}C^{(0)}=S^{(0)}C^{(0)}\varepsilon^{(0)}.\tag{27.18}$$

The Lagrangian multiplier matrix $[\varepsilon^{(0)}]$ can be chosen at each order of perturbation. Now, if an external electric field of $E^a(\vec{r},t)$ is applied, the mathematical form of $E^a(\vec{r},t)$ being

$$E^a(\vec{r},t) = E^a(\vec{r})\mathrm{X}[1 + e^{i\omega_a t} + e^{-i\omega_a t}], \tag{27.19}$$

then in the dipolar approximation, the external perturbation term (H) to the Kohn–Sham Hamiltonian becomes

$$H = \mu \cdot E(\vec{r},t) \tag{27.20}$$

where μ is the dipole-moment operator of the electrons.

The time dependence of the dipole moment produces various frequency-dependent hyperpolarizablity tensors. Frequency-dependent polarizability and hyperpolarizability tensors are obtained from the trace of the dipole moment matrix $[H^a]$ and the nth-order density matrix $[D^{(n)}]$ (where $n = 1$ for the linear polarizability α, $n = 2$ for the first-order hyperpolarizability tensor β, $n = 3$ for the second-order hyperpolarizability γ and so on). If the inducing electric field of frequency ω_a, ω_b, ω_c … acts in the direction a, b, c… the tensors can be represented as

$$\alpha_{ab}(-\omega_\sigma; \omega_b) = -\mathrm{Tr}\left[H^a D^b(\omega_b)\right], \tag{27.21}$$

$$\beta_{abc}(-\omega_\sigma; \omega_b, \omega_c) = -\mathrm{Tr}\left[H^a D^{bc}(\omega_b, \omega_c)\right], \tag{27.22}$$

$$\gamma_{abcd}(-\omega_\sigma; \omega_b, \omega_c, \omega_d) = -\mathrm{Tr}\left[H^a D^{bcd}(\omega_b, \omega_c, \omega_d)\right], \tag{27.23}$$

where $\omega_\sigma = \omega_\mathrm{b} + \omega_\mathrm{c} + \cdots$ These expressions are a set of generalized equations in terms of frequencies. The polarizability and the hyperpolarizability tensors defined above are determined through the iterative solution of the time-dependent Kohn–Sham (TDKS) equations given in Equation 27.14.

In ordinary TDDFT, the zeroth-order Kohn–Sham matrix is used, which is given by

$$F_s^{(0)} = h^{(0)} + D^{(0)}\mathrm{X}(2J) + \mathrm{v}_{xc}^{(0)}, \tag{27.24}$$

where $h^{(0)}$ contains the external potential terms that are of zero order in the external field, kinetic energy and the nuclear Coulomb field. The Coulomb super matrix J is independent of the external electric field and results in a Coulomb term of the form $D^{ab\ldots n}(\omega_a, \omega_b, \ldots, \omega_n)\mathrm{X}(2J)$ in the nth-order Kohn–Sham matrix $F_s^{ab\ldots n}(\omega_a, \omega_b, \ldots, \omega_n)$. The external perturbation appears only in the first-order Kohn–Sham matrices. Higher-order Kohn–Sham matrices consist of a Coulomb and an exchange-correlation part. Therefore, the general formula for the higher-order Kohn–Sham matrices becomes

$$\begin{aligned} F_s^{ab\ldots n}(\omega_a, \omega_b, \ldots, \omega_n) = {} & D^{ab\ldots n}(\omega_a, \omega_b, \ldots, \omega_n)\mathrm{X}(2J) \\ & + v_{xc}^{ab\ldots n}(\omega_a, \omega_b, \ldots, \omega_n), \end{aligned} \tag{27.25}$$

The Taylor expansions of F_s, C, ε, and D are inserted in the time-dependent Kohn–Sham equation (Equation 27.14), the normalization condition, and the density matrix (Equation 27.16). This results in a first-order time-dependent Kohn–Sham equation as

$$F_s^a(\omega)C^{(0)} + F_s^0 C^a(\omega) + \omega S^0 C^a(\omega) = S^{(0)}C^a(\omega)\varepsilon^{(0)} + S^{(0)}C^{(0)}\varepsilon^a(\omega). \tag{27.26}$$

Higher-order coupled equations are obtained by equating the left and the right hand sides of the TDKS equation, the normalization condition and the density matrix. The NLO properties are estimated through the iterative solution of the time-dependent Kohn–Sham equations up to a certain order n. Primarily, the static Kohn–Sham equations are solved that result in matrices $F_s^{(0)}$, $C^{(0)}$, $\varepsilon^{(0)}$, $D^{(0)}$, and the converged SCF density $\rho^{(0)}$. These matrices are needed for the solution of the first-order Kohn–Sham equation, which produces the first-order density matrix. The first-order density matrix yields the frequency-dependent polarizability component $[\alpha_\mathrm{ab}\ (-\omega_\sigma;\ \omega_\mathrm{b})]$ through Equation 27.21. The solution of the first-order equation provides matrices required for an iterative solution of the second-order equation. In either case, the technique adopted is called the $(2n + 1)$ theorem. The second-order equation is solved to get the second-order density matrix elements from which the frequency-dependent first-order hyperpolarizability tensor is obtained through Equation 27.23. If the external fields of the frequency zero and a common frequency ω are considered, a number of very important NLO properties become accessible. These are SHG $[\beta\ (-2\omega;\ \omega,\ \omega)]$, EOPE $[\beta\ (-\omega;\ \omega,\ 0)]$, OR $[\beta\ (0;\ \omega,\ -\omega)]$, and the static hyperpolarizability $[\beta\ (0;\ 0,\ 0)]$. Similar to the two previous cases, the $(2n + 1)$ theorem is also used to compute the higher-order hyperpolarizability tensors like γ, δ (third-order hyperpolarizability tensor), and the rest. When γ is calculated only at frequencies 0 and ω, THG $[\gamma\ (-3\omega;\ \omega,\ \omega,\ \omega)]$, EOKE $[\gamma\ (-\omega;\ \omega,\ 0,\ 0)]$, dc-SHG $[\gamma\ (-2\omega;\ \omega,\ \omega,\ 0)]$, and IDRI $[\gamma\ (-\omega;\ \omega,\ -\omega,\ \omega)]$ become accessible.

In practice, the average polarizability tensor is defined in terms of Cartesian components such as

$$\langle \alpha \rangle = \frac{\alpha_{xx} + \alpha_{yy} + \alpha_{zz}}{3}, \tag{27.27}$$

where, α_{xx}, α_{yy}, and α_{zz} are the diagonal elements of the polarizability-tensor matrix. The first-order hyperpolarizability tensor is defined as the third derivative of the energy with respect to the electric field components, and hence, involves one additional field differentiation compared to polarizabilities. The average first-order hyperpolarizability is defined as

$$\langle \beta \rangle = \left(\sum_i \beta_i \beta_i^* \right)^{1/2}; \quad \beta_i = \tfrac{1}{3} \sum_j (\beta_{ijj} + \beta_{jij} + \beta_{jji}), \tag{27.28}$$

where the sums are over the coordinates x, y, z ($i, j = x, y, z$), and β_i^* refers to the conjugate of the vector β_i. The second-order hyperpolarizability tensor involves one additional field differentiation compared to the first-order analog. The average second-order hyperpolarizability is defined as

$$\langle\gamma\rangle = \frac{1}{15}\sum_{ij}(2\gamma_{iijj} + \gamma_{ijji}); \quad (i,j = x,y,z). \tag{27.29}$$

In some of the theoretical approaches second-order hyperpolarizability tensors are determined from a combination of analytical and finite-difference techniques [28]. All components of the γ tensor of interest (dc-SHG/EOKE/static second-order hyperpolarizability) are obtained from the analytical time-dependent calculation of the SHG/EOPE/static first-order hyperpolarizability in the presence of small electric fields. E.g., the relation used in the evaluation of EOKE is

$$\gamma_{abcd}(-\omega;\omega,0,0) = \lim_{E^d \to 0} \frac{\beta_{abc}(-\omega;\omega,0)\big|_{E=E^d}}{E^d}, \tag{27.30}$$

Frequency-dependent response properties can also be calculated from the time-averaged quasienergy [29]. In this formalism, the response properties are obtained as the derivative of the quasi energy and the same is estimated by using the variational criterion for the quasienergy and the time-averaged time-dependent Hellmann–Feynman theorem. Molecular properties are obtained by using the variational Lagrangian technique in accordance with the $2n + 1$ and the $2n + 2$ rules. Within this approach, the different frequency-dependent response properties are obtained from the simple extension of the variational perturbation theory to the Fourier component variational perturbation theory.

27.3 Nonlinear Optics with Materials

Much of the present research work in the field of nonlinear optics is motivated by certain nonlinear optical phenomena in suitable materials [14]. Some of these potentially useful phenomena include the ability to alter the frequency or color of light, to amplify one source of light with another, and to alter its transmission features through a medium. The nonlinear optical materials require unusual stability with respect to ambient conditions and high-intensity light sources. These materials can broadly be categorized into two major classes. The first is the class of inorganic materials, which includes semiconducting and metallic clusters, inorganic crystals, bulk materials, etc. The second belongs to the organic or, in general, molecular materials that include mainly organic crystals and polymers. For these systems, the optical nonlinearities are usually derived from their structures. In the present chapter, we will focus only on the first type, i.e., semiconducting and metallic clusters and their NLO behavior.

27.4 Why Are Clusters So Important?

Nonlinear optical processes in cluster materials provide the necessary information about the exact understanding of the quantum confinements and the surface effects in these systems [11]. In general, the properties at the nanoscale are usually non-monotonic and oscillating in nature due to the quantum size effects, and as such, cluster materials can exhibit properties, which are quite different from those of the bulk [5,11,30]. More specifically, the energy band structure and the phonon distribution of cluster materials may differ from that of the bulk systems. Numerous theoretical as well as experimental investigations have been performed that show that the cluster-assembled materials have novel mechanical, electrical, and optical properties [5–16]. As mentioned earlier in this chapter, these NLO materials find huge applications in high-speed electro-optic devices for information processing and telecommunications [1–4]. In Section 27.5, we present a review on nonlinear optics with various cluster materials.

27.5 Review of Nonlinear Optics with Clusters

In this section, semiconducting clusters will be discussed first followed by metallic clusters. An example of a class of materials with a manifestation of unusual physicochemical and optical properties is cadmium selenide clusters [$(CdSe)_n$]. These $(CdSe)_n$ clusters are found to be the precursors of a wide range of low-dimensional materials such as quantum dots [7,10], tetrapods [31], and nanowires [32] that exhibit a variety of optical and electronic properties. The electronic properties of cadmium selenide clusters $(CdSe)_n$ and quantum dots are semiconducting in nature and these materials find tremendous application in the design and the development of novel nonlinear optical (NLO) materials. There have been a lot of experimental investigations [15,33] on the exploration of the NLO properties in CdSe clusters. The Hyper-Rayleigh scattering technique provides the experimental basis for NLO properties in CdSe clusters and nanoparticles [15]. Aktsipetrov et al. [33] have shown the size dependence of SHG from the surface of a composite material consisting of CdSe nanoparticles embedded in a glass matrix. The first-order hyperpolarizability in CdSe and CdS nanoparticles is reported to be very high [34].

Apart from the experimental works, theoretical *ab initio* investigations [8,11,12] have also been performed in order to calculate the NLO coefficients in CdSe clusters. Troparevsky and Chelikowsky [35], in their work, reported on the structural and the electronic properties (the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), binding energies, and polarizabilities) of small $(CdS)_n$ and $(CdSe)_n$ clusters, where n ranges from 2 up to 8 using the pseudopotential method in real space. Karamanis et al. [36] have used DFT for the computation of static polarizability and anisotropy in the static polarizability of small CdSe clusters up to tetramer units (see Figure 27.1).

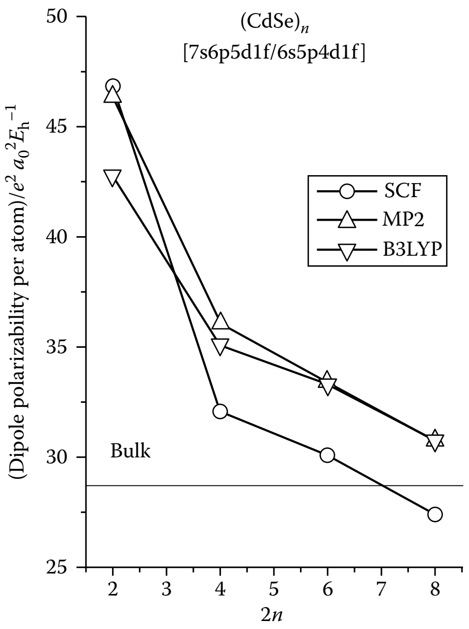

FIGURE 27.1 Evolution of the mean dipole polarizability with cluster size in a small CdSe cluster, (CdSe)$_n$. (From Karamanis, P. et al., *J. Chem. Phys.*, 124, 071101, 2006. With permission.)

Their investigations suggest that the optical properties of the bulk system can also be predicted from a plot of dipole polarizability per atom.

Later on, the same group of Karamanis extended their work on CdSe clusters, highlighting the importance of basis sets and electron correlation in these systems [12]. However, the calculations of the frequency-dependent optical polarizability, as well as that of hyperpolarizability tensors using theoretical methods remained an unresolved issue for CdSe clusters.

Recently, Sen and Chakrabarti [11], for the first time, investigated the frequency-dependent nonlinear optical properties (first- and second-order hyperpolarizability tensors) of (CdSe)$_n$ clusters up to the tetramer using the time-dependent density functional theory (TDDFT). Within the TDDFT, both the local-density approximation (LDA) and the generalized-gradient approximation (GGA) technique have been used. The more accurate LB94 (van Leeuwen and Baerends 94) [37] functional has been used under the GGA scheme. Response calculations have been implemented in the Amsterdam density functional package (ADF 2004.01) [38]. The analytical approach has been used for the evaluation of all the NLO coefficients. The calculated polarizability and hyperpolarizabilities depend on the choice of the basis set and the exchange-correlation potential. While the results obtained within the DFT are overestimated under the normal LDA and GGA functional, the coupled Hartree–Fock procedure exhibits exactly the opposite trend [39,40]. The static values of the average first-order hyperpolarizability (⟨β (0; 0, 0)⟩) and the average second-order hyperpolarizability (⟨γ (0; 0, 0, 0)⟩) of CdSe, Cd_2Se_2, Cd_3Se_3, and Cd_4Se_4 are depicted in Table 27.1.

The frequency dependence of the SHG [β (−2ω; ω, ω)], the EOPE [β (−ω; ω, 0)], the EFISH [γ (−2ω; ω, ω, 0)], and the EOKE [γ (−ω; ω, 0, 0)] was paid premier attention in the investigation of Sen and Chakrabarti [11]. For a complete evaluation of the frequency dependence of NLO properties in different CdSe clusters, a wide range of frequencies (0 to 0.45 a.u.) was considered.

TABLE 27.1 Average Static First and Second Order Static Hyperpolarizability in CdSe, Cd_2Se_2, Cd_3Se_3, and Cd_4Se_4

Sample	Average Static First-Order Hyperpolarizability, ⟨β (0;0,0)⟩ (a.u.)	Average Static Second-Order Hyperpolarizability, ⟨γ (0;0,0,0)⟩ (a.u.)
CdSe	−540.07 (LDA)	−556,680 (LDA)
Cd_2Se_2	Absent	7,306.5 (LDA)
		4,398.6 (LB94)
Cd_3Se_3	−0.12258 (LDA)	12,200 (LDA)
	−0.059125 (LB94)	8,929.3 (LB94)
Cd_4Se_4	Absent	89,565 (LDA)
		50,124 (LB94)

Source: Sen, S. and Chakrabarti, S., *Phys. Rev. B*, 74, 205435, 2006. With permission.

Figure 27.2 manifests the frequency behavior of ⟨β (−2ω; ω, ω)⟩ and ⟨β (−ω; ω, 0)⟩ in the case of CdSe and Cd_3Se_3 respectively. ⟨β (−2ω; ω, ω)⟩, being a frequency-dependent property, is of paramount interest and has been investigated extensively in numerous previous investigations [15,41,42]. High degrees of fluctuations in ⟨β (−2ω; ω, ω)⟩ are observed in all the cases. Experimentally, ⟨β (−2ω; ω, ω)⟩ in CdSe nanocrystals and quantum dots has been observed by the HRS technique and its size dependence was also verified [15]. It is evident from Figure 27.2 that ⟨β (−2ω; ω, ω)⟩ is negative over a wide range of frequencies. This makes CdSe and Cd_3Se_3 more significant in the perspective of quantum optics. The frequency variation in ⟨β (−2ω; ω, ω)⟩ divulges one more significant information, i.e., the abrupt increase in ⟨β (−2ω; ω, ω)⟩ at specific frequencies. At specific frequencies the magnitude of ⟨β (−2ω; ω, ω)⟩ becomes very high. It is a common notion that larger values of ⟨β (−2ω; ω, ω)⟩ are obtained at a near resonance of the input energy [5]. The presence of the near resonance indicates that linear absorption can occur at such frequencies.

Similar to the components of the first-order hyperpolarizability tensors, the components of the second-order hyperpolarizability tensors (⟨γ (−2ω; ω, ω, 0)⟩ and ⟨γ (−ω; ω, 0, 0)⟩) were reported to be highly sensitive to frequency variation. Figure 27.3 demonstrates the frequency dependence of different γ tensors. It was concluded from both LDA and LB94 results that both ⟨γ (−2ω; ω, ω, 0)⟩ and ⟨γ (−ω; ω, 0, 0)⟩ are negative over a wide range of frequencies. The change in the sign of ⟨γ (−2ω; ω, ω, 0)⟩ and ⟨γ (−ω; ω, 0, 0)⟩, due to the change in frequency is a significant observation. This led to a frequency selection in assigning the optical activity of the particular cluster.

In a separate study, Chakrabarti and coworkers [43] investigated the evolution of electric polarizability and the anisotropy of the polarizability at static as well as dynamic (Nd:YAG laser) frequencies with an increase in the cluster size for small, as well

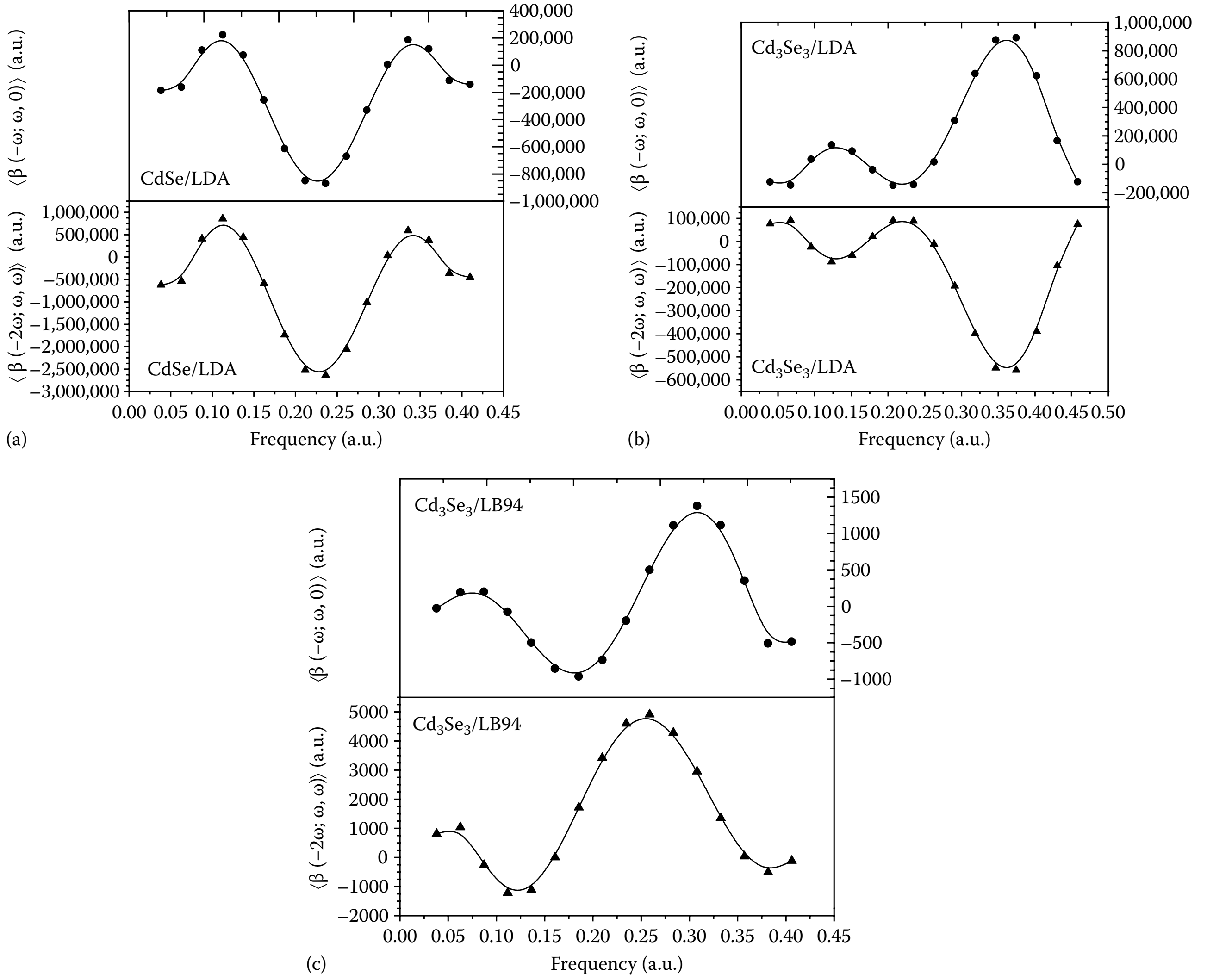

FIGURE 27.2 Variation of ⟨β (−2ω; ω, ω)⟩ (SHG) and ⟨β (−ω; ω, 0)⟩ (EOPE) with a frequency at (a) LDA scheme for the CdSe cluster, (b) LDA, and (c) GGA (LB94) scheme for the Cd_3Se_3 cluster. (From Sen, S. and Chakrabarti, S., *Phys. Rev. B*, 74, 205435, 2006. With permission.)

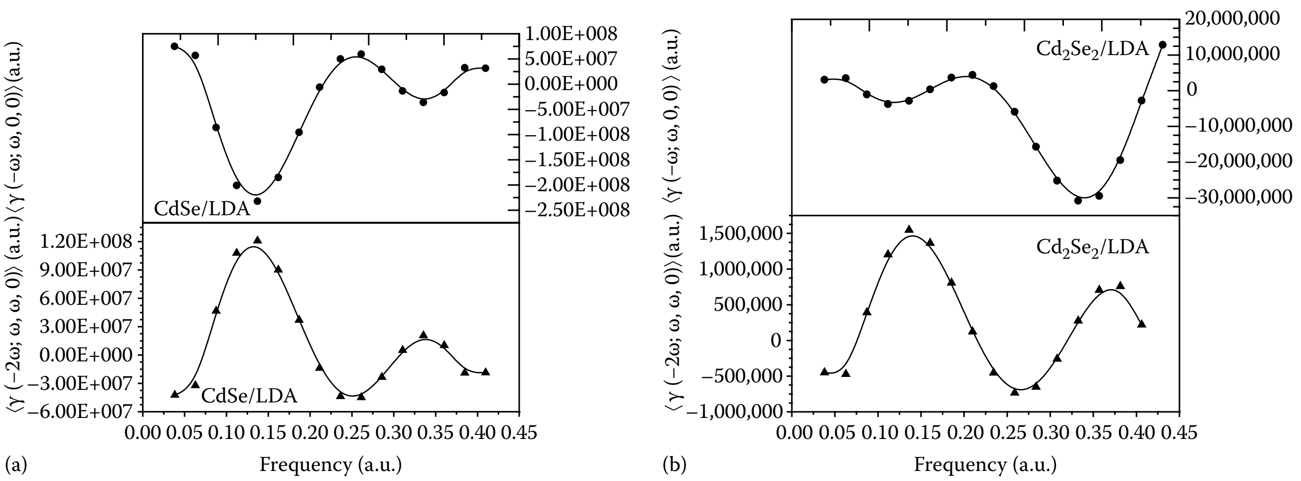

FIGURE 27.3 Variation of ⟨γ (−2ω; ω, ω, 0)⟩ and ⟨γ (−ω; ω, 0, 0)⟩ with a frequency at LDA. (a) Scheme for the CdSe cluster, (b) LDA and

(*continued*)

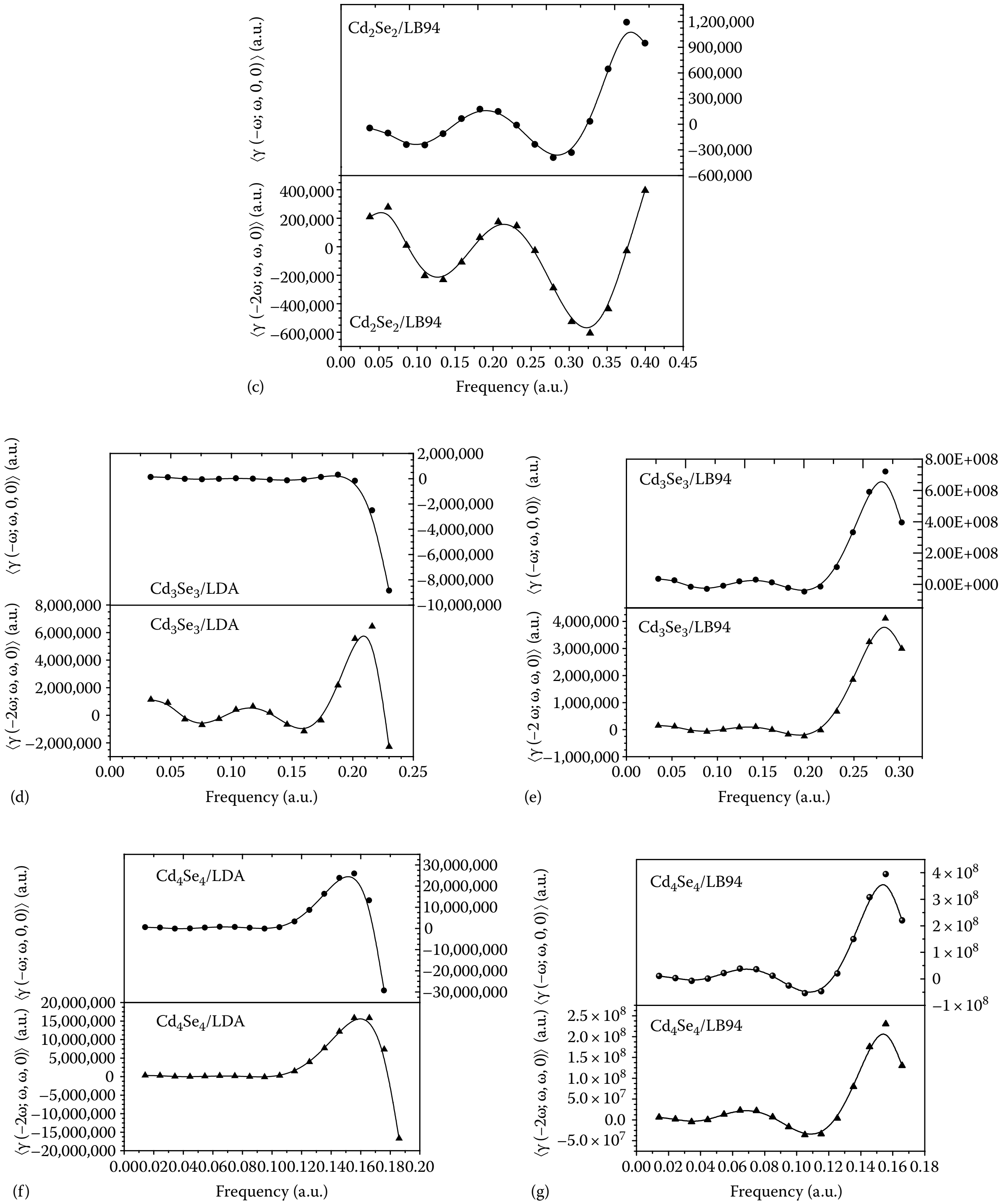

FIGURE 27.3 (continued) (c) GGA (LB94) scheme for the Cd_2Se_2 cluster, (d) LDA and (e) GGA (LB94) scheme for the Cd_3Se_3 cluster, (f) LDA and (g) GGA (LB94) scheme for the Cd_4Se_4 cluster. (From Sen, S. and Chakrabarti, S., *Phys. Rev. B*, 74, 205435, 2006. With permission.)

as medium-sized (CdSe)$_n$ (n = 1–16) clusters within the density functional theory (DFT). The main motivation of that investigation was

1. None of the theoretical studies reported had gone beyond a cluster size of eight for the polarizability calculations for the CdSe clusters.
2. Except the study by Sen and Chakrabarti [11], no one was aware of any investigations where the dynamic hyperpolarizability for any system size was reported for the CdSe clusters.

Knowing the strong dependence of polarizability on the cluster diameter (the cluster size), they performed polarizability calculations beyond the cluster size of eight and calculated both static and dynamic polarizabilities.

Table 27.2 describes computed polarizability values at static and Nd:YAG laser frequencies. An even-odd oscillating behavior is observed in the anisotropy values between the dimer and the heptamer CdSe clusters.

Just like the CdSe clusters, GaAs is one such semiconducting cluster, which has gained a lot of importance over the past few years [5]. These cluster materials display long absorption tails in the low energy region [44] owing to the existence of free surfaces in it. Moreover, the static polarizabilities of these systems exceed the bulk value and follow a decreasing tendency with the increase in the cluster size [45]. The variable nature in the polarizabilities also indicates its "metallic-like" behavior. In the experimental investigation [46], it has been observed that the polarizabilities for small and medium-sized Ga$_n$As$_m$ clusters reside above and below the bulk limit. Albeit, there have been a lot of experimental and theoretical investigations [47–49] based on the NLO properties of GaAs bulk materials which suggest that GaAs should possess very large static second- and third-order susceptibilities, only a few of them deal with the theoretical evaluation of the hyperpolarizabilities of these cluster materials [50,51]. A recent work by Lan et al. [5] provides a systematic insight into the NLO properties of several small and medium Ga$_n$As$_m$ clusters, where $m + n$ runs up to 10. They employed the TDDFT method in combination with the sum-over-states (SOS) formalism to calculate the first- and second-order hyperpolarizability tensors. The results indicate the presence of large second- and third-order nonlinear susceptibilities in GaAs cluster materials similar to that of the bulk. These results also suggest that resonance absorption can be avoided in the frequency-dependent NLO experiments. Furthermore, the frequency dependence of second- and third-harmonic generation reveals that these clusters are good candidates for future NLO materials.

Besides these two clusters, investigations of the NLO properties have also been carried out on various other III-V atomic nanoclusters. Pineda and Karna [52] estimated the linear and the nonlinear polarizabilities of isolated GaN nanoclusters employing the *ab initio* time-dependent Hartree–Fock (TDHF) method. Their results suggest a strong dependence of the NLO properties on the size and the shape of the clusters. The linear and the nonlinear optical properties of (CdS)$_n$ clusters were analyzed and the merits of the DFT level calculations have been compared with different *ab initio* results, as well as the basis set dependence of the optical properties of these clusters has been evaluated by Maroulis and Pouchan [53].

Figure 27.4 depicts the variation in the mean polarizability and the mean second-order hyperpolarizability values in small (CdS)$_n$ clusters. The values suggest a reduction in the mean

TABLE 27.2 Average Values of Electric Polarizability per Atom ($\langle\alpha\rangle/2n$) and Anisotropy in Polarizability ($\Delta\alpha$) against the Cluster Size, $2n$ at Both Static (0.0 a.u.) and Nd:YAG Laser (0.04283 a.u.) Frequencies

			At 0.0 a.u. Frequency		At Nd:YAG Laser (0.04283 a.u.) Frequency	
Samples	n	$2n$	Polarizability per Atom ($\langle\alpha\rangle/2n$) (a.u.)	Anisotropy in Polarizability ($\Delta\alpha$) (a.u.)	Polarizability per Atom ($\langle\alpha\rangle/2n$) (a.u.)	Anisotropy in Polarizability ($\Delta\alpha$) (a.u.)
CdSe	1	2	34.0349	36.5416	31.3777	49.7528
Cd_2Se_2	2	4	30.9670	60.5853	32.5664	59.0703
Cd_3Se_3	3	6	29.8098	99.7527	30.5716	102.1434
Cd_4Se_4	4	8	29.5301	0.0000	30.8635	0.0000
Cd_5Se_5	5	10	32.4484	159.8706	33.3488	166.2891
Cd_6Se_6	6	12	29.1341	43.8215	30.0033	43.5338
Cd_7Se_7	7	14	29.0036	134.4511	29.9893	143.1402
Cd_8Se_8	8	16	29.5565	114.3660	30.4782	122.1772
Cd_9Se_9	9	18	29.4061	49.9355	30.2877	54.6105
$Cd_{10}Se_{10}$	10	20	29.8171	124.1952	30.7628	132.7159
$Cd_{11}Se_{11}$	11	22	29.8559	150.1088	30.7738	160.6296
$Cd_{12}Se_{12}$	12	24	28.4830	0.0165	29.3089	0.0169
$Cd_{13}Se_{13}$	13	26	30.9808	271.1190	32.0344	287.1187
$Cd_{14}Se_{14}$	14	28	30.3060	192.7016	31.2651	207.4891
$Cd_{15}Se_{15}$	15	30	30.2442	158.5999	31.1661	170.6002
$Cd_{16}Se_{16}$	16	32	33.5302	428.7594	34.9350	470.1254

Source: Jha, P.C. et al., *Comput. Mater. Sci.*, 44, 728, 2008. With permission.

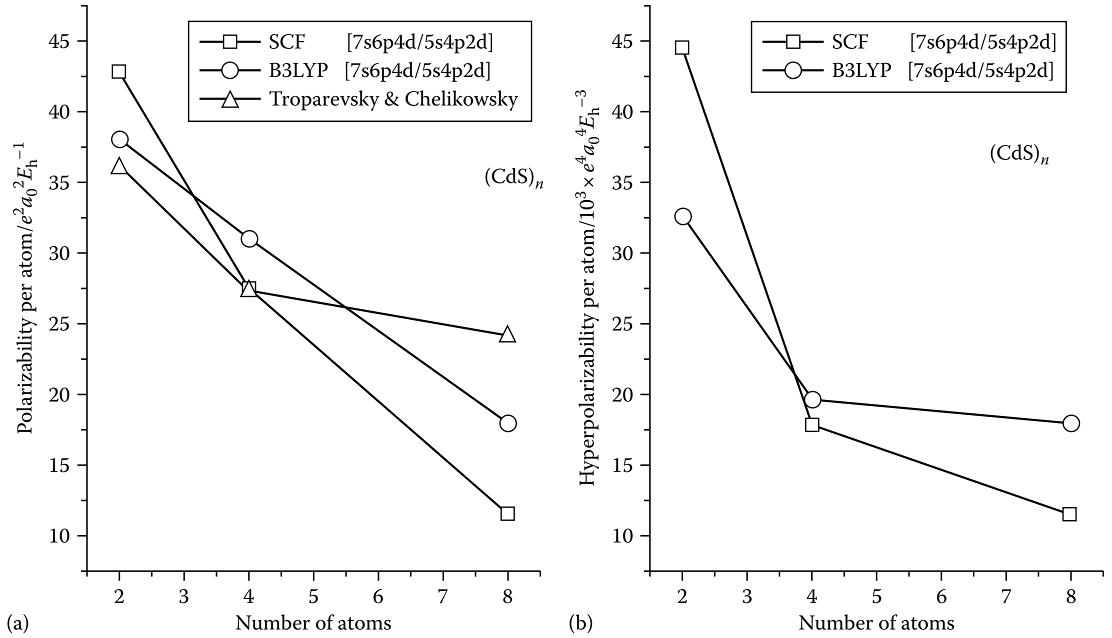

FIGURE 27.4 Mean (a) dipole polarizability and (b) second-order hyperpolarizability in (CdS)$_n$ clusters. (From Maroulis, G. and Pouchan, C., *J. Phys. Chem. B*, 107, 10683, 2003. With permission.)

dipole polarizability, as well as in the second-order hyperpolarizability values with an increase in the cluster size. Moreover, the results obtained from B3LYP and other conventional *ab initio* methods are well in accordance with each other and also with the earlier theoretical results [35]. However, in the case of the monomer geometry, the difference in the hyperpolarizability values between these methods is quite large. Champagne et al. estimated [54] average second-order hyperpolarizabilities of Si$_n$ (n = 3–38) clusters using MP2, MP3, MP4, CCSD, and CCSD(T) levels of approximations and demonstrated the variation of polarizability and hyperpolarizability against the cluster size.

Apart from these semiconducting materials, cluster science has also made rapid progress in determining the electric properties of various metal clusters [55] including that of mixed metal. At first, a review of the some of the works based on the metal clusters containing only one type of atom is discussed, which is followed by some detailed analysis on mixed metal clusters. There are investigations on lithium [18,56,57], sodium [58], copper [59,60], nickel [61], niobium [62], and zinc [63] clusters. Maroulis and Xenides [57] reported highly accurate *ab initio* calculations on the polarizability and the hyperpolarizabilities of lithium tetramer, Li$_4$ with specially designed basis sets for Li. The molecule, Li$_4$ has got a very high anisotropic dipole polarizability and quite large second-order hyperpolarizability values. The main feature of this work is the extensive analysis of the basis set and the correlation effects in calculating the electric properties. The results indicate a discrepancy in dipole polarizabilities between theory and experiment, which might be resolved by measuring the anisotropy values experimentally. The same group also presented a thorough analysis of the performance of several density functional theoretical (DFT) calculations of the (hyper) polarizability of Li$_4$. Their investigations elucidated that of the various DFT methods employed; only the mPW1PW91 and the O3LYP methods produce reliable results. Papadopoulos et al. [63] studied the (hyper) polarizabilities of small and medium-sized zinc clusters employing a hierarchy of basis sets and computational methods. The relativistic effect on the electric polarizabilities has also been investigated by employing the Douglas–Kroll approximation. It has been observed that the relativity contribution is significant in these cluster materials; however, the correlation effect plays a major part in influencing the electro-optic properties.

The NLO properties of Ag nanoparticles embedded in a Si$_3$N$_4$ matrix was investigated by Traverse et al. [64]. They performed a two-color sum-frequency generation (2c-SFG) spectroscopy experiment, which exhibited a surface plasmon resonance at a 421 nm wavelength in the visible region and absorption in the infrared region. The third-order optical nonlinearity of dielectrics in the presence of nanoparticles has been experimentally investigated by Flytzanis [65]. Chen and coworkers performed [66] high-level *ab-initio* calculations [HF, MP2, the fourth-order perturbation theory using single, double, quadrupole substitutions, CCSD, and QCISD] on (HCN)$_n$...Li and Li...(HCN)$_m$ clusters with electride characteristics. A high value of static first hyperpolarizability is reported in both samples. The static second-order hyperpolarizability of a series of tri-nuclear metal cluster MS4(M′PPh$_3$)$_2$(M′PPh$_3$) (M = Mo, W; M′ = Cu, Ag, Au) have been estimated by [67] Chen and coworkers. They employed a finite-field approach using the hybrid density functional theory (B3LYP) and reported a very high value of second-order hyperpolarizability. The NLO properties of Au$_n$ clusters

TABLE 27.3 Average Hirshfeld Charge on the Al_4 Ring, Ground State Dipole Moment, Magnitude of Average Static First- and Second-Order Hyperpolarizability in Al_4M_4 (M = Li, Na, and K)

Sample	Average Hirshfeld Charge on the Al_4 Ring	Ground-State Dipole Moment (Debye)	Average Static First-Order Hyperpolarizability, $\langle\beta(0;0,0)\rangle$ (a.u.)	Average Static Second-Order Hyperpolarizability, $\langle\gamma(0;0,0,0)\rangle$ (a.u.)
Al_4Li_4 (2D)	−0.247	0.049	145.922	1.299×10^6
Al_4Li_4 (3D)	−0.202	0.018	33.079	2.406×10^6
Al_4Na_4 (2D)	−0.258	0.003	21.537	1.275×10^6
Al_4Na_4 (3D)	−0.229	0.054	187.843	1.961×10^6
Al_4K_4 (2D)	−0.303	0.080	316.781	3.380×10^6
Al_4K_4 (3D)	−0.297	6.513	8863.074	1.149×10^7

Source: Sen, S. et al., *Phys. Rev. B*, 76, 115414, 2007. With permission.

and their alloyed clusters, $Au_{n-m}M_m$ (M = Ag, Cu; m = 1, 2) have been theoretically investigated by Xu and coworkers [68]. They used the usual TD-DFT as well as the B3LYP hybrid density functional employing the LANL2DZ basis set and measured the first-order hyperpolarizabilites of such clusters. They suggested that the high degree of NLO properties of these clusters are due to the lack of a centro-symmetric feature in these clusters and an enhancement in the transition electric dipole moment in alloyed clusters.

The optical properties of several mixed metal clusters such as Li_nCu_m, Na_4Pb_4, Na_6Pb, $Al_{11}Fe$, and $Al_{18}Fe$, Al_4M_4 (M = alkali metals) have also been investigated extensively [69,70]. In this context, we particularly mention the NLO properties of aluminum metal clusters [Al_4M_4 (M = Li, Na, and K)]. Various investigations have been performed on these Al_4M_4 systems and their anions due to their unique characteristic features and their structural similarity with that of cyclobutadiene, C_4H_4 [6]. These cluster materials are also important since they are better polarized than their organic counterparts. Earlier, Datta and Pati [6] investigated the static linear and nonlinear polarizabilities of Al_4M_4 (M = Li, Na, and K) systems. They predicted that structural changes through the broken inversion symmetry of Al_4M_4 could lead to a unique polarization response and large optical coefficients. In their work, both the static linear and nonlinear optical properties of these Al_4M_4 were estimated through ZINDO-multireference doubles configuration interaction (MRDCI) calculations. They emphasized that the large NLO properties of Al_4M_4 are due to the charge transfer from alkali metals to the Al_4 ring. In spite of the fact that their work exhibits an exceptionally high magnitude of static linear and nonlinear optical coefficients, it lacks the description of its dynamic counterparts. It is to be noted that the importance of a material in nonlinear optics and its applications in device fabrication demands a complete description of the frequency dependence. Hence, it was highly instructive to explore the dynamic (frequency-dependent) NLO properties of these clusters for their proper assessment in optical device fabrication and to explain whether the large NLO properties are due to a charge transfer interaction or due to other excitations.

Recently Sen et al. [13] reported various components of first- (SHG, EOPE, static first-order hyperpolarizability) and second- (EFISH, EOKE, and static second-order hyperpolarizability) order hyperpolarizability tensors at several near resonance frequencies. All these dynamic NLO properties of Al_4M_4 (M = Li, Na, and K) clusters have been explored using the TDDFT as a prime investigating tool. Within the TDDFT, the generalized gradient approximation (GGA) technique has been implemented with the use of a more accurate LB94 (van Leeuwen and Baerends 94) functional.

The magnitude of static $\langle\beta\rangle$ and $\langle\gamma\rangle$ along with an average Hirshfeld charge on the Al_4 ring and the ground-state dipole moments are shown in Table 27.3. It is quite clear from Table 27.3 that the magnitude of $\langle\beta(0;0,0)\rangle$ increases significantly from 2D to 3D in the case of Al_4Na_4 and Al_4K_4, on the other hand, a reverse trend is observed for Al_4Li_4.

Table 27.3 also suggests that the magnitudes of the average static $\langle\gamma(0;0,0,0)\rangle$ are very high (of the order of 10^6 and 10^7 a.u.) for all the clusters under investigation.

However, the central issue of the investigation of Sen and coworkers [13] was to explicate the frequency dependence of $\langle\beta(-2\omega;\omega,\omega)\rangle$, $\langle\beta(-\omega;\omega,0)\rangle$, $\langle\gamma(-2\omega;\omega,\omega,0)\rangle$, and $\langle\gamma(-\omega;\omega,0,0)\rangle$ in aluminum metal clusters. These components of the first- and second-order hyperpolarizability tensors were estimated in the vicinity of near-excitation frequencies with a substantial oscillator strength and presented in Table 27.4. A closer inspection of Table 27.4 reveals that $\langle\beta(-2\omega;\omega,\omega)\rangle$ and $\langle\beta(-\omega;\omega,0)\rangle$ attain negative values at certain frequencies in all these cluster materials. Exceptionally high values [of the order of 10^6–10^8 a.u.] of $\langle\gamma(-2\omega;\omega,\omega,0)\rangle$ and $\langle\gamma(-\omega;\omega,0,0)\rangle$ were also noticed. A significant observation was in the change in sign of $\langle\gamma(-2\omega;\omega,\omega,0)\rangle$ and $\langle\gamma(-\omega;\omega,0,0)\rangle$ due to the change in frequency, which lead to a frequency selection in assigning the optical activity of a particular cluster.

In order to explain the observed high NLO coefficients in these aluminum metal clusters, an analysis of the magnitude of gross populations of molecular orbitals with the most significant SFO (Symmetrized Fragment Orbitals) was carried out. The SFO study divulged the role of charge transfer interactions. Molecular orbitals with the most significant SFO gross population indicated that except for Al_4K_4 (2D and 3D) and Al_4Li_4 (3D) there exists excitations other than a charge transfer interaction at near resonance frequencies.

Moreover, it is worth mentioning that the dynamic first-order hyperpolarizability tensors are functions of the excitation

TABLE 27.4 ⟨β(−2ω; ω, ω)⟩, ⟨β(− ω; ω, 0)⟩, ⟨γ(−2ω; ω, ω, 0)⟩, and ⟨γ (−ω; ω, 0, 0)⟩ Measured at Frequencies Nearer to Excitation Frequencies in Al_4M_4 (M = Li, Na and K)

Sample	Excitation Frequencies (a.u.)	Oscillator Strength	Frequency at which Data Are Taken (a.u.)	⟨β(−2ω; ω, ω)⟩ (a.u.)	⟨β(−ω; ω, 0)⟩ (a.u.)	⟨γ(−2ω; ω, ω, 0)⟩ (a.u.)	⟨γ(−ω; ω, 0, 0)⟩ (a.u.)
Al_4Li_4 (2D)	0.0205	0.0042	0.0200	−6081.100	12,741	8.988×10^6	-2.744×10^6
	0.0522	0.0034	0.0500	224.900	46.143	-5.239×10^5	1.155×10^6
	0.0558	0.0094	0.0540	−371.630	361.960	1.226×10^7	-7.558×10^5
Al_4Li_4 (3D)	0.0377	0.0441	0.0360	4057.900	−15,252	6.049×10^6	1.663×10^7
	0.0650	0.0863	0.0640	968.580	−902.360	2.729×10^7	-8.458×10^4
	0.0709	0.3736	0.0690	−638.610	236.140	-1.199×10^8	-1.866×10^7
Al_4Na_4 (2D)	0.0238	0.0068	0.0220	−429.590	2697.100	1.266×10^7	-7.639×10^7
	0.0403	0.0012	0.0390	51.085	−96.614	4.735×10^4	1.245×10^6
	0.0570	0.0852	0.0590	4641.500	−93.815	1.540×10^7	-6.290×10^6
Al_4Na_4 (3D)	0.0271	0.0301	0.0260	2193.700	−13333	1.675×10^7	-6.581×10^6
	0.0446	0.0060	0.0430	−452.340	−178.82	4.459×10^6	5.213×10^5
	0.0543	0.0099	0.0560	−7067.500	−168.550	1.242×10^7	1.401×10^6
Al_4K_4 (2D)	0.0244	0.0066	0.0230	305.930	864.090	1.286×10^6	6.640×10^6
	0.0427	0.0910	0.0410	−1465.600	3477.900	-3.240×10^7	3.698×10^7
	0.0431	0.0197	0.0440	−64908	−5213.200	-3.143×10^8	2.548×10^7
Al_4K_4 (3D)	0.0215	0.0010	0.0200	−231620	−16,582	1.842×10^7	-1.926×10^7
	0.0391	0.0105	0.0380	29403	150,940	9.758×10^6	2.791×10^7
	0.0416	0.0067	0.0400	−120770	380,220	-7.515×10^7	4.444×10^7

Source: Sen, S. et al., *Phys. Rev. B*, 76, 115414, 2007. With permission.

energy, the dipolar matrix, and the difference between the dipole moment of the ground and the excited states. On the other hand, components of the second-order hyperpolarizability tensor are associated with four dipolar matrices in the numerator and three excitation frequencies in the denominator. Hence, it is very difficult to find all the factors that play a dominant role in the high β and γ values in particular clusters.

27.6 Conclusions

Therefore, it is well established from both the theoretical as well as the experimental standpoint that cluster materials can be considered as promising candidates for future NLO devices due to their high nonlinear optical coefficient. However, the exploration of the fundamentals of the unique physicochemical behavior of these clusters is a challenge to modern physical and chemical research. A theoretical determination of a higher-order nonlinear optical coefficient is still a matter of extensive computation and there is need for further developments to provide a proper physical background for some of the higher-order optical coefficients. There is enough scope for future research work in the field of cluster science and the design of new cluster materials with interesting physical behavior.

References

1. M. Lee, E. H. Katz, C. Erben, D. M. Gill, P. Gopalan, J. D. Heber, and D. J. McGee, *Science* **298**, 1401 (2002).
2. Y. Shi, C. Zhang, H. Zhang, J. H. Bechtel, L. R. Dalton, B. H. Robinson, and W. H. Steier, *Science* **288**, 119 (2000).
3. J. Luo, M. Haller, H. Li, H.-Z. Tang, A. K.-Y. Jen, K. Jakka, C.-H. Chou, and C.-F. Shu, *Macromolecules* **37**, 248 (2004).
4. C. Zhang, L. R. Dalton, M. C. Oh, H. Zhang, and W. H. Steier, *Chem. Mater*. **13**, 3043 (2001).
5. Y.-Z. Lan, W.-D. Cheng, D.-S. Wu, J. Shen, S.-P. Huang, H. Zhang, Y.-J. Gong, and F.-F. Li, *J. Chem. Phys*. **124**, 094302 (2006).
6. A. Datta and S. K. Pati, *J. Phys. Chem. A* **108**, 9527 (2004).
7. A. Puzder, A. J. Williamson, F. Gygi, and G. Galli, *Phys. Rev. Lett*. **92**, 217401 (2004).
8. M. C. Troparevsky, L. Kronik, and J. R. Chelikowsky, *J. Chem. Phys*. **119**, 2284 (2003).
9. R. Jose, N. U. Zhanpeisov, H. Fukumura, Y. Baba, and M. Ishikawa, *J. Am. Chem. Soc*. **128**, 629 (2006).
10. J. Zhao, J. A. Bardecker, A. M. Munro, M. S. Liu, Y. Niu, I.-K. Ding, J. Luo, B. Chen, A. K.-Y. Jen, and D. S. Ginger, *Nano Lett*. **6**, 463 (2006).
11. S. Sen and S. Chakrabarti, *Phys. Rev. B* **74**, 205435 (2006).
12. P. Karamanis, G. Maroulis, and C. Pouchan, *Chem. Phys*. **331**, 19 (2006).
13. S. Sen, P. Seal, and S. Chakrabarti, *Phys. Rev. B* **76**, 115414 (2007).
14. P. N. Prasad and D. J. Williams in *Introduction to Nonlinear Optical Effects in Molecules and Polymers*. Wiley, New York (1991); S. P. Karna and A. T. Yeates in *Nonlinear Optical Materials*, Eds.; ACS Symposium Series 628, American Chemical Society, Washington, DC (1996).
15. M. Jacobsohn and U. Banin, *J. Phys. Chem. B* **104**, 1 (2000).
16. S.-i. Inoue and Y. Aoyagi, *Phys. Rev. Lett*. **94**, 103904 (2005).

17. L. Jensen, P.-O. Astrand, A. Osted, J. Kongsted, and K. V. Mikkelsen, *J. Chem. Phys.* **116**, 4001 (2002).
18. S. K. Pati, S. Ramasheshа, Z. Shuai, and J. L. Bredas, *Phys. Rev. B: Condens. Matter* **59**, 14827 (1999).
19. K. B. Sophy and S. Pal, *J. Chem. Phys.* **118**, 10861 (2003); L. D. Freo, F. Terenziani, and A. Painelli, *J. Chem. Phys.* **116**, 755 (2002).
20. S. Grimme, *J. Comput. Chem.* **25**, 1463 (2004); Q. Wu and W. Yang, *J. Chem. Phys.* **116**, 515 (2002).
21. M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, and E. Kaxiras, *J. Chem. Phys.* **114**, 5149 (2001).
22. S. K. Pati, T. J. Marks, and M. A. Ratner, *J. Am. Chem. Soc.* **123**, 7287 (2001); D. Beljonne, Z. Shuai, J. Cornil, D. dos Santos, and J. L. Bredas, *J. Chem. Phys.* **111**, 2829 (1999).
23. J. L. Oudar and D. S. Chemla, *J. Chem. Phys.* **66**, 2664 (1977); J. L. Oudar, *J. Chem. Phys.* **67**, 446 (1977).
24. J. J. Zyss, *Chem. Phys.* **71**, 909 (1979).
25. D. L. Andrews, L. C. Dávila Romero, and W. J. Meath, *J. Phys. B: At. Mol. Opt. Phys.* **32**, 1 (1999).
26. A. Datta and S. K. Pati, *Chem. Soc. Rev.* **35**, 1305 (2006).
27. P. Seal, P. C. Jha, and S. Chakrabarti, *J. Mol. Struct.: Theochem*, **855**, 64 (2008); D. R. Kanis, M. A. Ratner, and T. J. Marks, *Chem. Rev.* **94**, 195 (1994).
28. S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends, *J. Chem. Phys.* **109**, 10644 (1998); S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends, *J. Chem. Phys.* **109**, 10657 (1998).
29. O. Christiansen, P. JØrgensen, and C. HÄttig, *Int. J. Quantum Chem.* **88**, 1 (1998).
30. K. R. S. Chandrakumar, T. K. Ghanty, and S. K. Ghosh, *Int. J. Quantum Chem.* **105**, 166 (2005).
31. L. Manna, E. C. Scher, and A. P. Alivisatos, *J. Am. Chem. Soc.* **122**, 12700 (2000); J. Li and L. W. Wang, *Nano Lett.* **3**, 1357 (2003).
32. W. U. Huynh, J. J. Dittmer, and A. P. Alivisatos, *Science* **295**, 2425 (2002).
33. O. A. Aktsipetrov, P. V. Elyutin, A. A. Nikulin, and E. A. Ostrovskaya, *Phys. Rev. B* **51**, 17591 (1995).
34. B. S. Santos, G. A. L. Pereira, D. V. Petrov, and C. de Mello Donegá, *Opt. Commun.* **178**, 187 (2000).
35. M. C. Troparevsky and J. R. Chelikowsky, *J. Chem. Phys.* **114**, 943 (2001).
36. P. Karamanis, G. Maroulis, and C. Pouchan, *J. Chem. Phys.* **124**, 071101 (2006).
37. R. van Leeuwen and E. J. Baerends, *Phys. Rev. A* **49**, 2421 (1994).
38. http://www.scm.com
39. A. J. Cohen, N. C. Handy, and D. J. Tozer, *Chem. Phys. Lett.* **303**, 391 (1999).
40. R. M. Dickson and A. D. Becke, *J. Phys. Chem.* **100**, 16105 (1996).
41. D. V. Petrov, B. S. Santos, G. A. L. Pereira, and C. deMello-Donegá, *J. Phys. Chem. B* **106**, 5325 (2002).
42. N. Song, L. Men, J. P. Gao, Y. Bai, A. M. R. Beaudin, G. Yu, and Z. Y. Wang, *Chem. Mater.* **16**, 3708 (2004).
43. P. C. Jha et al., *Comput. Mater. Sci.* **44**, 728 (2008).
44. I. Vasiliev, S. Öğüt, and J. R. Chelikowsky, *Phys. Rev. B* **60**, R8477 (1999).
45. I. Vasiliev, S. Öğüt, and J. R. Chelikowsky, *Phys. Rev. Lett.* **78**, 4805 (1997).
46. R. Schäfer, S. Schlecht, J. Woenckhaus, and J. A. Becker, *Phys. Rev. Lett.* **76**, 471 (1996).
47. M. Zh. Huang and W. Y. Ching, *Phys. Rev. B* **47**, 9464 (1993).
48. Z. H. Levine and D. C. Allan, *Phys. Rev. Lett.* **66**, 41 (1991).
49. D. J. Moss, E. Ghahramani, J. E. Sipe, and H. M. van Driel, *Phys. Rev. B* **41**, 1542 (1990).
50. J. L. Wang, M. L. Yang, G. H. Wang, and J. J. Zhao, *Chem. Phys. Lett.* **367**, 448 (2003).
51. P. P. Korambath and S. P. Karna, *J. Phys. Chem. A* **104**, 4801 (2000).
52. A. C. Pineda and S. P. Karna, *Chem. Phys. Lett.* **429**, 169 (2006).
53. G. Maroulis and C. Pouchan, *J. Phys. Chem. B* **107**, 10683 (2003).
54. B. Champagne, M. Guillaume, D. Bégué, and C. Pouchan, *J. Comput. Methods Sci. Eng.* **7**, 297 (2007).
55. K. D. Bonin and V. V. Kresin, in *Electric Dipole Polarizabilities of Atoms, Molecules and Clusters*, World Scientific, Singapore (1997).
56. E. Benichou, R. Antoine, D. Rayane, B. Vezin, F. W. Dalby, Ph. Dugourd, M. Broyer et al., *Phys. Rev. A* **59**, R1 (1999).
57. G. Maroulis and D. Xenides, *J. Phys. Chem. A* **103**, 4590 (1999).
58. G. Tikhonov, V. Kasperovich, K. Wong, and V. V. Kresin, *Phys. Rev. A* **64**, 063202 (2001).
59. G. Maroulis, *J. Phys. Chem. A* **107**, 6495 (2003).
60. M. B. Knickelbein, *J. Chem. Phys.* **120**, 10450 (2004).
61. M. B. Knickelbein, *J. Chem. Phys.* **115**, 5957 (2001).
62. R. Moro, X. Xu, S. Yin, and W. A. de Heer, *Science* **300**, 1265 (2003).
63. M. G. Papadopoulos, H. Reis, A. Avramopoulos, S. Erkoç, and L. Amirouche, *J. Phys. Chem. B* **109**, 18822 (2005).
64. A. Traverse(a), C. Humbert, C. Six, A. Gayral, and B. Busson, E. *Phys. Lett.* **83**, 64004 (2008).
65. C. Flytzanis, *J. Phys. B* **38**, S661 (2005).
66. W. Chen, Z.-Ru Li, D. Wu, R. Li, and C. Sun, *J. Phys. Chem. B* **109**, 601 (2005).
67. X. Chen, K. Wu, J. G. Snijders, and C. Lin, *Chinese Chem. Lett.* **13**, 893 (2002).
68. Y. Xu, C. Xu, T. Zhou, and C. Cheng, *J. Mol. Struct. Theochem* **893**, 88 (2009).
69. P. Fuentealba, L. P.-Campos, and O. Reyes, *J. Comput. Methods Sci. Eng.* **4**, 589 (2004).
70. M. B. Torres and L. C. Balbás, *J. Comput. Methods Sci. Eng.* **4**, 517 (2004) and references therein.

28

Second-Harmonic Generation in Metal Nanostructures

28.1 Introduction 28-1
28.2 Nonlinear Optics and Second Harmonic Generation 28-2
28.3 Second Harmonic Generation in Nanosystems....................28-3
Spherical Isolated Particles and Plane-Wave Illumination • Low-Symmetry Particles • Nonuniform Illumination • Second Harmonic Generation in Resonant Metal Particles • Random Metal Nanostructures and Rough Metal Surfaces
28.4 Leading Order Contributions to the Second Harmonic Generation Process in Nanoparticles....................28-5
Conservation of Parity and Angular Momentum • Long-Wavelength Limit
28.5 Emission Patterns and Light Polarization 28-8
Irradiated Intensity • Second Harmonic Polarization
28.6 Allowed and Forbidden Second Harmonic Emission Modes: Examples 28-9
Plane-Wave Illumination • Illumination with a Laterally Limited Light Beam
28.7 Second Harmonic Generation in Single Gold Nanoparticles....................28-11
Near-Field Microscopy Setup • Fundamental Wavelength and Second Harmonic Maps • Second Harmonic Polarization Analysis
28.8 Perspectives....................28-15
28.9 Conclusions....................28-15
Acknowledgments....................28-16
References....................28-16

Marco Finazzi
Politecnico di Milano

Giulio Cerullo
Politecnico di Milano

Lamberto Duò
Politecnico di Milano

28.1 Introduction

The explosive growth of nanoscience and nanotechnology during the last decade has led to great interest in the investigation of nanoscale optical fields and to the development of experimental tools for their study (Novotny and Hecht 2006). One of the most remarkable effects in light interaction with metal nanostructures is the strong and spatially localized field amplitude enhancement, due to lightning rod effects induced by the sharp curvatures (Novotny and Hecht 2006) and/or resonant excitation of collective electron oscillations (plasma oscillations) in single or coupled particles (Grand et al. 2003, Mühlschlegel et al. 2005). The resonance frequencies associated with such oscillations and thus the optical properties can be tuned in a broad spectral range according to the material and the shape of the nanostructure. Particularly, gold and silver may sustain plasma oscillations at optical frequencies.

Field enhancements are best observed by exploiting nonlinear optical effects such as second-harmonic generation (SHG) (Smolyaninov et al. 1997, 2000, Jakubczyk et al. 1999, Zayats et al. 2000, Shen et al. 2001, Biagioni et al. 2007, Breit et al. 2007) or two-photon photo-luminescence (TPPL) (Jakubczyk et al. 1999, Bouhelier et al. 2003, 2005, Imura et al. 2005, 2006), which depend on the square of the light intensity, thus amplifying local field variations. To observe a significant nonlinear response, the samples are typically illuminated by high peak intensities, such as those associated with ultrashort laser pulses. Nonlinear optical microscopies, either in the far field or in the near field, are thus becoming powerful tools for the study of nanoscale field enhancements.

SHG was first experimentally demonstrated in 1961 (Franken et al. 1961) and has since then found many applications, especially in converting laser light to a different color. Another application more relevant to the content of this chapter is surface SHG. Although the sensitivity of the SHG process to boundaries was soon recognized (Bloembergen and Pershan 1962), it was only in the early 1980s that nonlinear optical spectroscopy of surfaces and interfaces started to develop as a well-established analysis tool in a wide area of fields (Shen 1989, Mc Gilp 1996, Bloembergen 1999, Lüpke 1999, Downer et al. 2001). The SHG ability to selectively probe surfaces and interfaces derives from a selection rule (see Section 28.2) that does not allow for dipole emission from the bulk of centrosymmetric materials. At the surface or interface, however, the symmetry of the bulk is broken and dipole generation is no longer forbidden. This peculiar property has determined the enormous spread of SHG-based techniques in the community of surface and interface physics. This technique has the advantages of being highly surface-sensitive, capable of remote sensing and in situ measurement, and applicable to any interface between two condensed media accessible by light, even if buried. Among the applications, surface SHG can be used to monitor the growth of monolayers on a surface or to probe the alignment of absorbed molecules.

This chapter presents theoretical and experimental results on SHG from metal nanostructures. It is organized as follows: Section 28.2 contains a brief description of the principles of nonlinear optics and SHG; Section 28.3 presents a survey of the recent literature on SHG from nanostructures; Section 28.4 introduces a theoretical analysis of SHG from nanostructures, allowing to derive selection rules for the process; Section 28.5 describes the light emission patterns and polarization, while Section 28.6 presents experimentally relevant illumination conditions. Section 28.7 reports experimental results on SHG from gold nanoparticles, while Sections 28.8 and 28.9 contain a discussion of possible future developments and conclusions, respectively.

28.2 Nonlinear Optics and Second Harmonic Generation

Nonlinear optics (Shen 1984, Boyd 2003) is the branch of physics that describes the phenomena that occur when the matter is illuminated by very intense light fields that have the capability of modifying its optical properties, thus generating an optical response that is a nonlinear function of the light intensity. Since electric fields strong enough to induce a nonlinear optical response can only be obtained by using the spatially and temporally coherent radiations provided by lasers, the birth of nonlinear optics closely coincides with the invention of the laser.

The physical origins of the nonlinear optical response of a material can be understood by recalling the Lorentz model, in which the atoms/molecules are modeled as harmonic oscillators that, driven by the external light field, become oscillating dipoles emitting light at the same frequencies as the driving field. In a harmonic oscillator, the restoring force is a linear function of displacement; for large displacements, however, nonlinear terms in the restoring force start to become significant, thus leading to an anharmonic oscillation and emission of new frequency components not present in the driving field.

Let us consider an electromagnetic wave crossing a material: the time-varying electric field $E(t)$ induces a time-dependent polarization (dipole moment per unit volume) $P(t)$, which is responsible for light emission. At low light intensities, the polarization is a linear function of the driving field:

$$P(t) = \varepsilon_0 \chi^{(1)} E(t), \tag{28.1}$$

where

ε_0 is the vacuum permittivity
$\chi^{(1)}$ is known as the linear susceptibility

As the intensity increases, further nonlinear terms in which the polarization is proportional to higher powers of the electric field must be considered:

$$P(t) = \varepsilon_0 [\chi^{(1)} E(t) + \chi^{(2)} E^2(t) + \chi^{(3)} E^3(t) + \cdots], \tag{28.2}$$

where the quantities $\chi^{(2)}$ and $\chi^{(3)}$ are known as second- and third-order nonlinear susceptibilities, respectively. By considering the vectorial nature of E and P, $\chi^{(1)}$ becomes a second-rank tensor, $\chi^{(2)}$ a third-rank tensor, and so on.

The $\chi^{(2)}$ term is responsible for second-order nonlinear optical effects, such as sum frequency generation, difference frequency generation, and SHG. The following text briefly describes the SHG process. Let us consider the geometry shown in Figure 28.1a, in which a monochromatic wave at the fundamental frequency ω, $E(t) = E_0 \cos(\omega t)$, impinges on a crystal with nonzero second-order susceptibility. The second-order polarization can then be written as

$$P^{(2)}(t) = \varepsilon_0 \chi^{(2)} E^2(t) = \varepsilon_0 \chi^{(2)} E_0^2 \cos^2(\omega t) = \frac{1}{2} \varepsilon_0 \chi^{(2)} E_0^2 [1 + \cos(2\omega t)]. \tag{28.3}$$

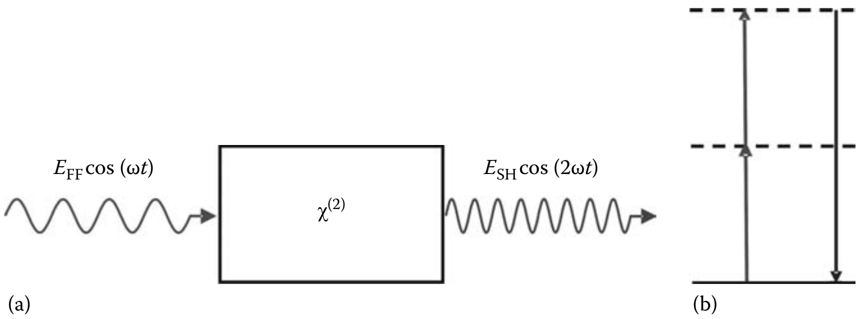

FIGURE 28.1 Scheme of SHG in a nanoparticle: two incoming FW photons are absorbed by the particle and their energy converted in a single photon with double frequency.

The nonlinear polarization thus contains, in addition to a constant term (optical rectification), a component at the second-harmonic (SH) frequency 2ω, which irradiates a field at the corresponding frequency. Note that, since the second-order nonlinear polarization depends on the square of the driving field amplitude E_0, the SHG efficiency is enhanced by high-peak power, pulsed illumination. The SHG process can also have a simple corpuscular interpretation, according to the energy-level scheme reported in Figure 28.1b: two fundamental wavelength (FW) photons at energy $\hbar\omega$ are absorbed to a virtual level of the material, and an SH photon with energy $2\hbar\omega$ is emitted, thus satisfying energy conservation.

In a centrosymmetric medium (i.e., a medium with a center of inversion), the $\chi^{(2)}$ term vanishes identically. In such a medium, in fact, by changing the sign of the driving field, also the sign of the nonlinear polarization must change. By substituting in Equation 28.3, one obtains

$$-P^2(t) = \varepsilon_0\chi^2[-E(t)]^2 = \varepsilon_0\chi^2E^2(t), \tag{28.4}$$

which can only be satisfied if $\chi^{(2)} \equiv 0$. Since gases, liquids, amorphous solids, and also several crystals possess an inversion symmetry, the second-order nonlinear susceptibility in such materials is zero, and thus, SHG is forbidden.

At the microscopic level, one can consider the single emitter (atom, molecule) as a nonlinear dipole radiating at the SH frequency 2ω. In order to generate a macroscopic SH signal, the fields emitted by the individual dipoles must interfere constructively, adding their contributions in phase. Since the phase of the nonlinear dipoles depends on the phase of the FW driving field, to achieve constructive interference between the nonlinear emissions at two different positions in the nonlinear medium, one must match the phase velocities of the FW field $v_{\mathrm{pFW}} = \omega/k_{\mathrm{FW}}$ and the SH field $v_{\mathrm{pSH}} = 2\omega/k_{\mathrm{SH}}$. From $v_{\mathrm{pFW}} = v_{\mathrm{pSH}}$, one obtains the condition $k_{\mathrm{SH}} = 2k_{\mathrm{FW}}$, also known as "phase-matching" condition, which implies that $n_{\mathrm{SH}} = n_{\mathrm{FW}}$, where n is the refractive index of the medium. If the phase-matching condition is satisfied, all the microscopic emissions add in phase and the SH field grows linearly (and the intensity quadratically) with the propagation distance. On the other hand, if $\Delta k = k_{\mathrm{SH}} - 2k_{\mathrm{FW}} \neq 0$, the SH field depends on the propagation coordinate x as $E_{\mathrm{SH}} \propto \chi^{(2)} \sin(\Delta k\, x/2)$. It thus increases for propagation up to a "coherence length" L_c such that $\Delta k\, L_c = \pi$, and then starts to decrease due to destructive interference.

Achievement of the phase matching condition is made difficult by the fact that most materials display, in the optical frequency range, a refractive index increasing monotonously with frequency (normal dispersion), so that $n_{\mathrm{SH}} > n_{\mathrm{FW}}$. One can obtain phase-matching condition exploiting the birefringence of noncentrosymmetric crystals, that is, the dependence of the refractive index on the polarization and propagation direction. By polarizing the SH wave along the direction giving the lower refractive index, one can satisfy the condition $n_{\mathrm{SH}} = n_{\mathrm{FW}}$ and thus achieve phase matching. Under such conditions, the SHG process can become so efficient that nearly all the FW power is converted to the SH (Parameswaran et al. 2002).

28.3 Second Harmonic Generation in Nanosystems

As discussed in Section 28.2, SHG is forbidden, within the dipole approximation, from the bulk of a centrosymmetric material but is allowed at its surface, where the inversion symmetry is broken. For this reason, SHG has been extensively applied since its discovery to the investigation of the surface properties of centrosymmetric media. In more recent years, thanks to its surface selectivity, SHG has attracted a considerable interest also as a tool to investigate the properties of nonplanar surfaces that characterize nanoscale systems. The following text presents a survey of the recent literature concerning SHG in nanostructures.

28.3.1 Spherical Isolated Particles and Plane-Wave Illumination

Perfectly spherical small particles made of isotropic material under FW plane-wave illumination represent the simplest geometry to study SHG in nanostructured systems. Experiments have been conducted on spherical dielectric (Wang et al. 1996, Yang et al. 2001, Shan et al. 2006) or metallic (Vance et al. 1998, Nappa et al. 2005) particles dispersed in dilute suspensions, or on colloidal particles, ordered in a centrosymmetric crystalline lattice (Martorell et al. 1997). The influence of the particle diameter (Yang et al. 2001, Nappa et al. 2005, Shan et al. 2006), concentration (Wang et al. 1996, Vance et al. 1998), and the presence of adsorbates (Wang et al. 1996) on SHG have been investigated, by performing angle- and polarization-resolved measurements (Martorell et al. 1997, Yang et al. 2001, Nappa et al. 2005, Shan et al. 2006).

The theory is in good agreement with the experiments, confirming that, when the sphere is small compared with the wavelength of light (Rayleigh limit), the leading-order SH radiation is emitted by a locally excited electric quadrupole and by an electric dipole directed along the direction of the propagation of the incident FW light beam that requires a nonlocal excitation mechanism, in which the phase variation of the pump beam across the sphere needs to be considered (Dadap et al. 1999, 2004). The locally excited electric dipole term, analogous to the source for linear Rayleigh scattering, is absent for the nonlinear case because the FW field induces mutually canceling polarizations at opposite sides of a centrosymmetric spherical particle (Dadap et al. 1999, 2004). The theory predicts the absence of any SHG signal in the forward direction, with both dipole and quadrupole contributions producing SH radiation with an intensity of leading order $(ka)^6$, k being the wave vector and a the particle diameter (Dadap et al. 1999, 2004). Theoretical models have been applied to describe SHG in particles larger than the Rayleigh limit (Dewitz et al. 1996, Pavlyukh and Hübner 2004) and to explicitly account for the electromagnetic response of conducting nanoparticles (Hua and Gersten 1986, Dewitz et al. 1996, Panasyuk et al. 2008).

28.3.2 Low-Symmetry Particles

The fact that SHG is forbidden in the bulk of centrosymmetric materials makes SHG very sensitive to the particle symmetry (Sandrock et al. 1999). For instance, experimental results suggest that SHG can be used to characterize the symmetry and chirality of carbon nanotubes (Su et al. 2008). It has been theoretically (Bachelier et al. 2008) and experimentally demonstrated (Martorell et al. 1997, Nappa et al. 2006) that slight structural deviations from the spherical shape may lead to SH radiation and polarization properties of nanoparticles differing significantly from those of a sphere. For gold particles with a diameter smaller than 50 nm, SHG is dominated by a dipole contribution that is not due to defects in the crystalline structure of the particle but to the deviation of the particle shape from that of a perfect sphere (Nappa et al. 2006, Bachelier et al. 2008). For larger sizes, retardation effects in the interaction of the electromagnetic fields with the particles cannot be neglected any longer, and the response exhibits a strong quadrupolar contribution (Martorell et al. 1997, Nappa et al. 2006).

The importance of local defects in breaking the symmetry of the particles and, consequently, in influencing the SHG process has been verified also in noncentrosymmetric gold nanostructures lithographed on a glass substrate (Canfield et al. 2004, Canfield et al. 2006). SHG from such particles show a high degree of polarization sensitivity and reveals that responses forbidden to ideal, symmetric particles are not only present but are relatively large compared with the allowed response (Canfield et al. 2004). Indeed, with respect to spherical or ellipsoidal nanoparticles, very different polarization selection rules and SH emission directions are obtained when the symmetry of the particle is lowered (Neacsu et al. 2005, Finazzi et al. 2007).

Despite the nonlinear optical properties of a particle are very sensitive to local field enhancement, the presence of very strong fields alone may not be sufficient for efficient SHG from centrosymmetric nanoparticles, as a consequence of the high symmetry selectivity of SHG. An example is SHG from arrays of noncentrosymmetric T-shaped gold nanodimers with a nanogap (Canfield et al. 2007). In this case, SHG arises from asymmetry in the local fundamental field distribution and is not strictly related to the nanogap size, which determines the intensity of the electric field between the dimers. Calculations show that the local field contains polarization components that are not present in the exciting field, which yield the dominant SHG response. The strongest SHG responses occur through the local surface susceptibility of the particles for a fundamental field distributed asymmetrically at the particle perimeters. Weak responses result from more symmetric distributions despite high field enhancement in the nanogap. Nearly constant field enhancement persists for relatively large nanogap sizes (Canfield et al. 2007).

28.3.3 Nonuniform Illumination

The shape of particles participating in the SHG process is not the only geometrical parameter that might affect the SH emission modes and efficiency. Actually, also the spatial distribution of the exciting field can have a great importance to determine the nonlinear response of particles. For instance, for a nonuniform polarizing field, the cancellation of the SH fields generated at opposite sides of a centrosymmetric spherical particle (see Section 28.3.1) is no longer exact and lower order SH emission becomes possible (Brudny et al. 2000). This issue is particularly important since, in practical experimental geometries, the polarizing field is in general neither of a pure longitudinal character nor a simple plane wave, as in the case illumination is performed by focusing the light beam with a microscope objective.

In this context, it is worth mentioning an important application of SHG, namely, second-harmonic imaging microscopy (SHIM), which is a technique used for imaging living cells or tissue by detecting the SHG signal generated by noncentrosymmetric molecules (Campagnola and Loew 2003). By exploiting the quadratic intensity dependence of the SHG signal that confines it to the focal volume, SHIM has the capability of three-dimensional imaging deep within the tissue. SHIM offers several advantages: since it does not rely on absorption from a real electronic state, as in fluorescence microscopy, it does not give rise to photobleaching or phototoxicity. In addition, as it exploits the intrinsic hyperpolarizability of noncentrosymmetric molecules, it is a label-free technique that does not require the use of exogenous probes.

A general theory for the quadratic nonlinear response of a single small nonmagnetic centrosymmetric sphere illuminated by a nonhomogeneous electromagnetic field was developed by Mochán et al. (2003), while general symmetry-based selection rules for SHG with arbitrary illumination are given by Finazzi et al. (2007) and discussed in Section 28.4. The importance of the nature of FW illumination has also been demonstrated experimentally. In fact, it has been shown that the SH radiation generated in Si spherical nanocrystallites displays a peak along the forward direction with a very narrow angular aperture (Jiang et al. 2001, 2002). As explained in Section 28.3.1, this result contradicts the behavior expected for the SH radiation produced by a single sphere illuminated by a plane wave, which should identically vanish along the forward direction (Dadap et al. 1999, 2004). Nevertheless, this observation can be explained (Brudny et al. 2003) by considering the nonuniform illumination FW field distribution, consisting of a strongly focused Gaussian beam. Similarly, in Si nanocrystals embedded uniformly in an SiO_2 matrix, a configuration consisting of two noncollinear, orthogonally polarized FW beams is found to greatly enhance the SHG yield with respect to single-beam illumination (Figliozzi et al. 2005). This effect has been attributed to the strong inhomogeneities in the FW field under double-beam illumination (Figliozzi et al. 2005).

28.3.4 Second Harmonic Generation in Resonant Metal Particles

Although SHG is subject to the same geometry-based selection rules in both metal and dielectric nanoparticles (Finazzi et al. 2007), a relevant difference is represented by the fact that, in metal particles, the linear and nonlinear optical properties are governed by collective electronic excitations (plasma oscillations), which can be tuned in a broad spectral range by choosing the proper particle geometry. In nanosystems, the quasiparticles associated with the collective electron excitations are often addressed to as localized surface plasmons (LSPs), or as localized plasmon polaritons, to indicate the associated electric field. In particular, depending on the size and the shape, metal particles made of gold and silver may display LSP resonance in the visible region. Thanks to this tunability, the nonlinear optical properties of noble-metal nanoparticles have found applications for in vitro and in vivo imaging (Nagesha et al. 2007), high-resolution analysis of tumor tissues (Durr et al. 2007, Bickford et al. 2008, Park et al. 2008), and enhanced photothermal therapies (O'Neal et al. 2004, El-Sayed et al. 2006).

A relevant enhancement of the SHG efficiency is observed when either the FW (Hubert et al. 2007) or the SH (Antoine et al. 1997, 1998, Hao et al. 2002, Johnson et al. 2002, Abid et al. 2004, Russier-Antoine et al. 2004) frequency is set at the LSP resonance of the nanoparticles. When the SH wavelength is tuned in the vicinity of the surface plasmon resonance, the wavelength analysis indicates that the interband transitions do not play a significant role in the total SH response (Antoine et al. 1998). Rather, the nonlinear optical behavior of the clusters is dominated by the free-electron gas-like response of the conduction band, the surface plasmon resonance being much sharper than the interband transition contribution (Antoine et al. 1998, Russier-Antoine et al. 2004). Similarly, the SHG efficiency as a function of the FW frequency shows a curve that reproduces the LSP resonance (Hubert et al. 2007). Both electric dipole and electric quadrupole contributions to the SH-radiated power are observed (Hao et al. 2002).

28.3.5 Random Metal Nanostructures and Rough Metal Surfaces

The strong local field enhancements in random metal nanoparticles and rough metal surfaces can be exploited to improve the efficiency of the interaction of light with molecules approached to such nanostructures (Hartschuh et al. 2003), a phenomenon that can be exploited in surface-enhanced spectroscopies to obtain single-molecule sensitivity. Among these, surface-enhanced Raman scattering (SERS) makes use of the excitation provided by LSP and high local fields in such structures to enhance the Raman emission of adsorbed molecules by several orders of magnitudes (generally 10^6, up to 10^{12}). Surface enhanced spectroscopies such as SERS are, therefore, intimately correlated with the plasmon properties of the nanostructures, offering, at the same time, unique potential for ultrasensitive molecular identification (Nie and Emery 1997). LSP-based nanosensors of biomolecules, based on the shift of the plasmon resonance frequency of thin metallic films, have been developed and are already commercially available.

Since SHG is extremely sensitive to both local fields and LSP resonances, both theoretical (Agarwal and Jha 1982, Stockman et al. 2004, Beermann et al. 2006, Singh and Tripathi 2007) and experimental studies performed with either far-field (Bozhevolnyi et al. 2003, Beermann and Bozhevolnyi 2004) or near-field techniques (Smolyaninov et al. 1997, Zayats et al. 2000) have addressed the SHG process in such disordered systems characterized by LSP interacting with local defects. The latter act as plasmon-scattering centers and might also provide highly localized electric fields due to the lightning rod effect. LSP resonances emerge as a consequence of multiple interparticle light scattering (Stockman et al. 2004, Beermann et al. 2006) and exhibit very different strength, phase, polarization, and localization characteristics (Bozhevolnyi et al. 2003, Beermann and Bozhevolnyi 2004). SH emission is experimentally observed from small and very bright spots, whose locations depend on the light wavelength (Bozhevolnyi et al. 2003) and polarization (Zayats et al. 2000, Bozhevolnyi et al. 2003). According to simulations (Stockman et al. 2004), the spatial distributions of the fundamental frequency and SH local fields are very different, with highly enhanced SH hot spots corresponding to areas where FW and SH eigenmodes overlap (Beermann et al. 2006). Another feature of SHG in rough metal surfaces is that SH fields show a very rapid spatial decay and are strongly depolarized and incoherent (Stockman et al. 2004).

28.4 Leading Order Contributions to the Second Harmonic Generation Process in Nanoparticles

As illustrated in Section 28.3, the shape of the nanoparticle plays a significant role in determining the conditions for efficient SHG, which can also be significantly influenced by the illumination geometry. The latter represents a particularly important point that needs to be considered when microscopic techniques are employed to locally study field-enhancement processes. In this case, the FW illumination is not a plane wave anymore, and strong field gradients (and even a nonvanishing longitudinal field component when scanning near-field optical microscopy (SNOM) is used) can be present in the illumination area. Although nanoscale SHG has already been discussed in the literature for few particle geometries excited by either far (Zhu 2007, Dadap 2008) or near (Bozhevolnyi and Lozovski 2000, 2002) FW fields, a set of general validity selection rules for SHG might be helpful in addressing more complicated particle geometries and FW field distributions. The following text addresses the concepts that allow one to derive selection rules for SHG in isolated nanoparticles from angular momentum and parity conservation laws, as discussed by Finazzi et al. (2007).

28.4.1 Conservation of Parity and Angular Momentum

In the rest reference frame of the particle, the SHG process consists of two photons, belonging to the FW field interacting with the particle, that transfer their energy to a photon in the SH field oscillating at double frequency (see Figure 28.1b). Therefore, the energy of the electromagnetic field is conserved and the particle is left in the ground state, that is, the particle initial and final states in the SHG process coincide. This fact has relevant consequences if the particle is symmetric. In this case, the particle ground state is an eigenstate of parity and, possibly, angular momentum operators. In other words, the particle has a well-defined parity (or angular momentum). Since the particle initial and final states are the same, the electromagnetic field cannot transfer parity or angular momentum quanta to the particle, and the total parity or angular momentum of the field will be conserved. These concepts can be expressed more formally by considering the explicit expression of the SHG process cross section σ. By using perturbation theory, σ can be expressed as a third-order term of the form

$$\sigma \propto \left| \sum_{u,w} \frac{\langle \varphi_0 | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | u \rangle \langle u | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | w \rangle \langle w | \hat{\mathbf{p}} \cdot \mathbf{A}_{2\omega} | \varphi_0 \rangle}{[(E_u - E_0) - \omega + i\Gamma_u][(E_w - E_u) - \omega + i\Gamma_w]} \right.$$
$$+ \frac{\langle \varphi_0 | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | u \rangle \langle u | \hat{\mathbf{p}} \cdot \mathbf{A}_{2\omega} | w \rangle \langle w | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | \varphi_0 \rangle}{[(E_u - E_0) - \omega + i\Gamma_u][(E_w - E_u) + 2\omega]}$$
$$\left. + \frac{\langle \varphi_0 | \hat{\mathbf{p}} \cdot \mathbf{A}_{2\omega} | u \rangle \langle u | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | w \rangle \langle w | \hat{\mathbf{p}} \cdot \mathbf{A}_\omega | \varphi_0 \rangle}{[(E_u - E_0) + 2\omega][(E_w - E_u) - \omega + i\Gamma_w]} \right|^2, \quad (28.5)$$

where

- φ_0 is the particle ground state
- u and w are eigenfunctions corresponding to excited states of the unperturbed particle
- E_0, E_u, and E_w are the corresponding energies

Equation 28.5 consists of the sum of three terms, each corresponding to one of the Feynman diagrams shown in Figure 28.2. In the expression for σ, $\hat{\mathbf{p}}$ is the particle momentum operator, $\mathbf{A}_{2\omega}$ the SH vector potential, and $\mathbf{A}_\omega$ the vector potential of the *externally applied* FW field. Note that the field generated by the particle should not be added to $\mathbf{A}_\omega$ in Equation 28.5. In fact, the particle self-interactions, that is, the interactions between the particle and the electromagnetic field the particle produces are accounted for by u and w, which describe the electronic as well as the electromagnetic excitations of the particle. Retardation effects that play an important role in determining SHG in nanoparticles (Dadap et al. 1999, 2004) are therefore implicitly accounted for in Equation 28.5.

As far as parity and angular momentum are concerned, it is convenient to expand the FW and SH fields appearing in Equation 28.5 in terms of electric (E) and magnetic (M) multipoles, which represent the photon angular momentum eigenstates. Let L'_ω and L''_ω be the L^2 quantum numbers of the two FW absorbed photons, M'_ω and M''_ω their L_z quantum numbers, z being an arbitrary quantization axis, and $L_{2\omega}$ and $M_{2\omega}$ the respective quantum numbers for the emitted SH photon (see Figure 28.3). The Feynman diagram in Figure 28.2a gives a contribution to Equation 28.5 that can be expressed as the sum of T terms defined as follows:

$$T^{u,w}_{L'_\omega, M'_\omega, L''_\omega, M''_\omega, L_{2\omega}, M_{2\omega}}$$
$$\propto \langle \varphi_0 | \hat{\mathbf{p}} \cdot \mathbf{A}(L'_\omega, M'_\omega) | u \rangle \langle u | \hat{\mathbf{p}} \cdot \mathbf{A}(L''_\omega, M''_\omega) | w \rangle \langle w | \hat{\mathbf{p}} \cdot \mathbf{A}(L_{2\omega}, -M_{2\omega}) | \varphi_0 \rangle, \quad (28.6)$$

where $\hat{\mathbf{p}} \cdot \mathbf{A}(L, M)$ represents the electric or magnetic multipole operator of order (L, M). $M_{2\omega}$ is taken with a negative sign since the emitted SH photon carries the corresponding L_z angular momentum component *away* from the particle. Similar terms can be given for the other Feynman diagrams in Figure 28.2. Each T term corresponds to an SHG interaction involving FW and SH photons characterized by a well-defined total angular momentum.

The selection rules that determine the conditions for which a given T term is identically zero depend on how the multipoles in Equation 28.6 transform under the particle symmetry group operations. When the particle has spherical symmetry, its ground as well as excited states are eigenstates of both L^2 and L_z. Therefore, the SH photon must carry away a total angular momentum equal to the sum of the angular momenta of the two FW incoming photons. In cylindrical symmetry, only the projection of the photon angular momentum along the symmetry axis (taken as the quantization axis z) needs to be conserved. The photon parity must be conserved in the case the particle displays inversion symmetry. If the particle is invariant under inversion symmetry with respect to a point (S_2 symmetry in Schoenflies notation, see Tinkham (1964)), then electric and magnetic multipoles of

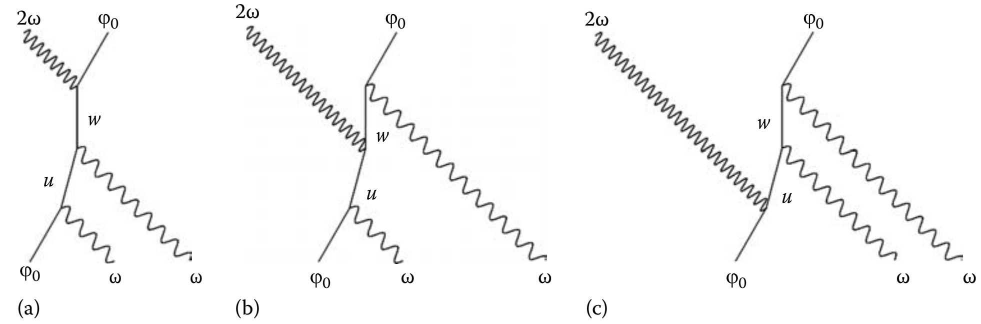

FIGURE 28.2 The three Feynman diagrams corresponding to SHG. (Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.)

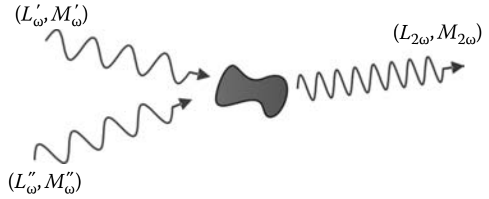

FIGURE 28.3 Scheme of SHG reporting the L and M quantum numbers that define the angular momentum and parity of the photons involved in the process.

order (L, M) have opposite contributions to the total parity of the electromagnetic field, equal to $(-1)^L$ and $(-1)^{L+1}$, respectively (Jackson 1975). In the case the particle displays inversion symmetry (C_2 symmetry) with respect to the quantization axis, electric and magnetic multipoles have the same parity, given by $(-1)^M$ (Jackson 1975). Finally, considering inversion with respect to a plane perpendicular to the quantization axis (C_{1h} symmetry), the parity of electric and magnetic multipoles of order (L, M) is given by $(-1)^{L+M}$ and $(-1)^{L+M+1}$, respectively (Jackson 1975).

By considering the electromagnetic field conservation laws imposed by the particle symmetry and contributions to angular momentum and parity associated with each multipole, one can obtain the selection rules listed in Table 28.1, which allow to determine the nonvanishing T terms that contribute to the SGH process for any combination of FW and SH multipoles (Finazzi et al. 2007). Thus, to apply the selection rules listed in Table 28.1, one has to determine the multipole expansion of the impinging FW electromagnetic field in order to understand which SH multipoles can contribute to the generated SH radiation.

TABLE 28.1 SHG Selection Rules for the Different Particle Symmetries Illustrated in Figure 28.4

Symmetry and Point Group	Selection Rule
Spherical	$\begin{pmatrix} L'_\omega & L''_\omega & L_{2\omega} \\ M'_\omega & M''_\omega & -M_{2\omega} \end{pmatrix} \neq 0$ and $(-1)^{L'_\omega + L''_\omega + L_{2\omega} + m} = 1$
Cylindrical	$M'_\omega + M''_\omega - M_{2\omega} = 0$
Central, S_2	$(-1)^{L'_\omega + L''_\omega + L_{2\omega} + m} = 1$
Axial, C_2	$(-1)^{M'_\omega + M''_\omega - M_{2\omega}} = 1$
Reflection, C_{1h}	$(-1)^{L'_\omega + M'_\omega + L''_\omega + M''_\omega + L_{2\omega} - M_{2\omega} + m} = 1$

Source: Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.

Note: Conservation of angular momentum for spherical symmetry corresponds to a condition involving a 3-j symbol (Tinkham 1964). The value m indicates the number of *magnetic* multipole transitions involved in the T term.

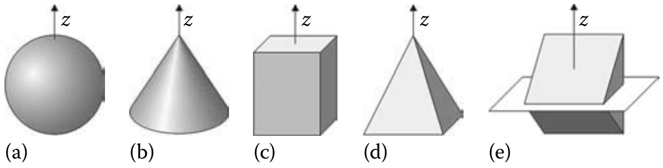

FIGURE 28.4 The particle symmetry groups discussed in the text and referred to the quantization axis z: (a) spherical, (b) cylindrical, (c) central, (d) axial, and (e) reflection symmetries across a mirror plane perpendicular to z. (Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.)

If the symmetry of the particle results from a combination of symmetry groups, the corresponding conditions listed in Table 28.1 must be satisfied at the same time. For instance, SHG from particles with C_{2h} symmetry must obey both the selection rules for axial (C_2) and reflection (C_{1h}) symmetries with respect to the same quantization axis. For the same reason, when the particle exhibits two or three orthogonal symmetry axes, the corresponding selection rules must be simultaneously satisfied by the field multipole expansions expressed by using each symmetry axis as the quantization axis.

Note that, strictly speaking, the spherical and cylindrical symmetry groups would have to be excluded since they are not compatible with any of the symmetry properties of the various Bravais lattices. Moreover, crystallographic defects (such as grain boundaries, dislocations, and atomic vacancies) would disrupt the internal symmetry of the particles. Thus, it would seem that symmetry-based selection rules would not be respected by SHG in real nanostructures. However, such limitations are not expected to play a significant role in the metal structures that are usually investigated in nano-optics (Finazzi et al. 2007), which are typically realized in aluminum, silver, or gold. The electronic and electromagnetic properties of these materials are in fact well described by the free-electron model, so that the particle symmetry properties are independent from the particle crystallography but are just determined by the boundary conditions imposed by the particle surface.

28.4.2 Long-Wavelength Limit

In general, selection rules allow one to discriminate the transitions that, for symmetry reasons, cannot contribute to the cross section of a particular physical phenomenon, but they do not give any information about the magnitude of the nonvanishing contributions. However, by considering the multipole expansion of the electromagnetic FW and SH fields in the SHG process, it is possible not only to identify the emission channels that are ineffective to SHG but also to estimate a hierarchy among them, identifying the ones that mostly contribute to the SHG process. This is possible in the long-wavelength limit (Rayleigh limit) characterized by $ka \ll 1$, with k being the FW wave vector and a the particle lateral size. In fact, matrix elements of the form $\langle u \mid \hat{\mathbf{p}} \cdot \mathbf{A}(L, M) \mid w \rangle$ rapidly decrease when the multipole order L is increased (Jackson 1975). The ratio $R(L)$ between matrix elements for successive orders, L and (L + 1), of either electric or magnetic multipoles of the same frequency is

$$R(L) \sim \frac{ka}{2L}, \tag{28.7}$$

while the magnetic multipoles of order L have cross sections of the same order in ka as electric multipoles of order $(L + 1)$ (Jackson 1975, Dadap et al. 2004). Since the expression of the SHG cross section σ reported in Equation 28.5 depends on the square value of the sum of products involving three matrix elements, we need to consider only the contributions associated with the lowest order multipoles.

By restricting the analysis to the lowest order terms, five distinct SHG emission modes can be indicated. Following the notation in Dadap et al. (2004), these are E1 + E1 → E1, E1 + E2 → E1, E1 + M1 → E1, E1 + E1 → E2, and E1 + E1 → M1. In this notation, the first two symbols refer to the nature of the interaction with the FW field, and the third symbol describes the SH emission. For example, the E1 + E2 → E1 interaction represents electric dipole (E1) SH emission that arises through combined FW electric dipole and electric quadrupole (E2) excitations. M1 indicates a magnetic dipole transition. The E1 + E1 → E1 transition corresponds to the leading order contribution to the SHG process, generating an SH field with a magnitude of a factor $(ka)^{-1}$ higher than the field generated by the other terms, resulting in a $(ka)^{-2}$ higher irradiated power. The E1 + E1 → E1 interaction violates parity conservation for any incident FW field distribution on centrosymmetric particles since $(-1)^{L'_{\omega}+L''_{\omega}+L_{2\omega}} = (-1)^{1+1+1} = -1$, thus violating the selection rule for S_2 symmetry particles reported in Table 28.1. This channel may, however, be allowed for lower symmetry particles.

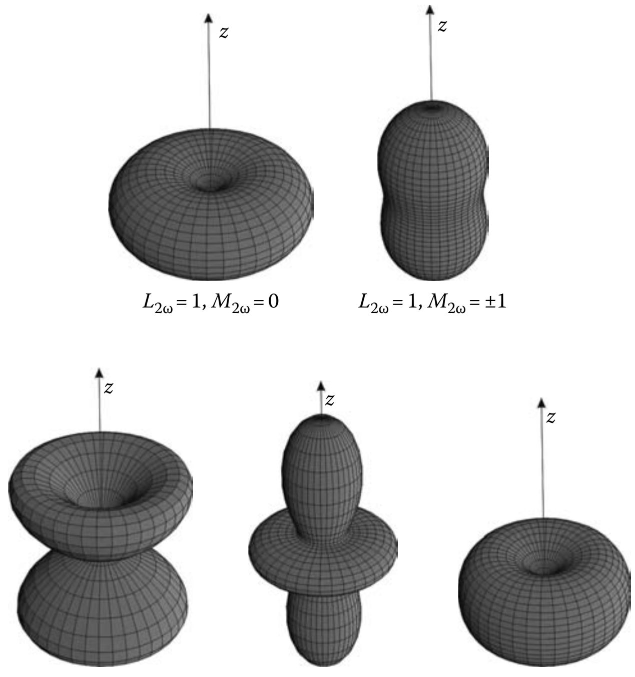

$L_{2\omega} = 1, M_{2\omega} = 0$ $L_{2\omega} = 1, M_{2\omega} = \pm 1$

$L_{2\omega} = 2, M_{2\omega} = 0$ $L_{2\omega} = 2, M_{2\omega} = \pm 1$ $L_{2\omega} = 2, M_{2\omega} = \pm 2$

FIGURE 28.5 Graphical representation of emission patterns $dP/d\Omega(L_{2\omega}, M_{2\omega})$ for dipoles (upper row) and quadrupoles (lower row). The time-averaged radiated power per unit solid angle displays cylindrical symmetry around the quantization axis z. (Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.)

28.5 Emission Patterns and Light Polarization

The emission modes that participate in the SHG process can be experimentally identified by considering SH emission pattern and light polarization. In fact, each multipole contributing to the SHG process is characterized by a well-defined angular dependence of the irradiated intensity and polarization.

28.5.1 Irradiated Intensity

Let $\mathbf{k}_{2\omega}$ be the SH wave vector and $\mathbf{r}$ the position with respect to the particle. According to Jackson (1975), in the far-field limit characterized by $k_{2\omega} \gg r^{-1}$, the SH electric fields $\mathbf{E}^{(E)}_{L_{2\omega}M_{2\omega}}(\mathbf{r})$ and $\mathbf{E}^{(M)}_{L_{2\omega}M_{2\omega}}(\mathbf{r})$, produced by either electric or magnetic multipoles of order $(L_{2\omega}, M_{2\omega})$, are given by

$$\mathbf{E}^{(E)}_{L_{2\omega},M_{2\omega}}(\mathbf{r}) = -(-i)^{(L_{2\omega}+1)} \frac{e^{ik_{2\omega}r}}{k_{2\omega}r} (\hat{\mathbf{u}}_r \times \mathbf{r} \times \nabla Y_{L_{2\omega},M_{2\omega}}), \qquad (28.8a)$$

$$\mathbf{E}^{(M)}_{L_{2\omega},M_{2\omega}}(\mathbf{r}) = -(-i)^{(L_{2\omega}+1)} \frac{e^{ik_{2\omega}r}}{k_{2\omega}r} (\mathbf{r} \times \nabla Y_{L_{2\omega},M_{2\omega}}). \qquad (28.8b)$$

In Equations 28.8 $\hat{\mathbf{u}}_r$ is a unit vector in the radial direction, and $Y_{L_{2\omega},M_{2\omega}}$ is the spherical harmonic of order $(L_{2\omega}, M_{2\omega})$. Pure electric and magnetic SH multipoles have the same angular distribution of the time-averaged radiated power per solid angle $dP/d\Omega(L_{2\omega}, M_{2\omega})$. A graphical representation of the radiated power is given in Figure 28.5 for dipoles ($L_{2\omega} = 1$) and quadrupoles ($L_{2\omega} = 2$) (Jackson 1975). If more than a single multiplet is responsible for the generation of SH radiation, interference among the SH radiation emitted by the contributing multipoles occurs and the emission pattern can be significantly different from those in Figure 28.5. In such cases, identifying the multipoles that are responsible for the SH emission might be quite complicated. Further information, however, can be extracted by analyzing the polarization dependence of the SH radiation as a function of the polarization of the impinging FW field.

28.5.2 Second Harmonic Polarization

From Equations 28.8, one can also obtain the polarization state of the emitted SH light. Table 28.2 reports the direction of the electric far field for pure electric and magnetic multipoles with $L_{2\omega} \leq 2$. The direction is indicated by a combination of the mutually perpendicular unit vectors $\hat{\mathbf{u}}_\theta$ and $\hat{\mathbf{u}}_\phi$ defined in a spherical coordinate system centered on the particle, in which θ and ϕ are the polar and azimuthal angles, respectively, the former being referred to the z quantization axis. An imaginary (real) ratio between the coefficients of $\hat{\mathbf{u}}_\theta$ and $\hat{\mathbf{u}}_\phi$ corresponds to a $\pi/2$ (zero) phase difference between the two field components and hence indicates elliptical (linear) polarization (Jackson 1975).

TABLE 28.2 SH Electric Field Polarization for Pure Electric or Magnetic Multipoles with $L_{2\omega} \leq 2$

	$M_{2\omega} = 0$	$M_{2\omega} = \pm 1$	$M_{2\omega} = \pm 2$
$L_{2\omega} = 1$ (E)	$\sin\theta\hat{\mathbf{u}}_\theta$	$e^{\pm i\phi}(\pm\cos\theta\hat{\mathbf{u}}_\theta \mp i\hat{\mathbf{u}}_\phi)$	
$L_{2\omega} = 1$ (M)	$\sin\theta\hat{\mathbf{u}}_\phi$	$-e^{\pm i\phi}(i\hat{\mathbf{u}}_\theta \mp \cos\theta\hat{\mathbf{u}}_\phi)$	
$L_{2\omega} = 2$ (E)	$-i\sin\theta\cos\theta\hat{\mathbf{u}}_\theta$	$e^{\pm i\phi}(\mp i\cos 2\theta\hat{\mathbf{u}}_\theta + \cos\theta\hat{\mathbf{u}}_\phi)$	$e^{\pm 2i\phi}\sin\theta(i\cos\theta\hat{\mathbf{u}}_\theta \pm \hat{\mathbf{u}}_\phi)$
$L_{2\omega} = 2$ (M)	$-i\sin\theta\cos\theta\hat{\mathbf{u}}_\phi$	$-e^{\pm i\phi}(\cos\theta\hat{\mathbf{u}}_\theta + i\cos 2\theta\hat{\mathbf{u}}_\phi)$	$e^{\pm 2i\phi}\sin\theta\,(\pm\hat{\mathbf{u}}_\theta + i\cos\theta\hat{\mathbf{u}}_\phi)$

Source: Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.

Note: For each point in space, the unit vectors $\hat{\mathbf{u}}_\theta$ and $\hat{\mathbf{u}}_\phi$ are, respectively, parallel and perpendicular to the plane containing the z (quantization) axis, and are both perpendicular to the radial unit vector $\hat{\mathbf{u}}_r$.

28.6 Allowed and Forbidden Second Harmonic Emission Modes: Examples

This section discusses some examples of applications of the arguments discussed above. Two relevant cases of FW illumination are addressed, namely, illumination with a plane wave and illumination with a strongly laterally limited beam, for example, through a high numerical aperture or via a near-field tip.

28.6.1 Plane-Wave Illumination

In an arbitrarily polarized FW plane wave propagating along the z direction, the field vector $\mathbf{A}(z)$ can be expressed as $\mathbf{A}(0)e^{ikz}$, with $\mathbf{A}(0)$ being a vector perpendicular to the propagation direction, $\mathbf{A}(0)\cdot\hat{\mathbf{u}}_z = 0$. By taking the origin of the z axis coincident with the particle position, and truncating the multipole expansion around $z = 0$ of the matrix element $\langle u \mid \hat{\mathbf{p}} \cdot \mathbf{A} \mid w\rangle$ to second order in ka, one obtains

$$\langle u|\hat{\mathbf{p}}\cdot\mathbf{A}|w\rangle \frac{E_w - E_u}{\hbar\omega}\langle u|\mathbf{r}\cdot\mathbf{E}(0)|w\rangle + \langle u|\hat{\mathbf{L}}\cdot\mathbf{B}(0)|w\rangle \\ + ik\langle u|\hat{p}_z[xA_x(0) + yA_y(0)]|w\rangle. \tag{28.9}$$

In this expression, $\mathbf{E}(0)$ and $\mathbf{B}(0)$ are complex vectors representing the amplitude and phase of the electric and magnetic fields, respectively, at the particle position, while $A_x(0)$ and $A_y(0)$ are the complex components of $\mathbf{A}(0)$, and $\hat{\mathbf{L}} = \mathbf{r} \times \hat{\mathbf{p}}$. The three terms on the right-hand side of Equation 28.9 correspond to E1, M1, and E2 transitions, respectively.

Let us first discuss the case of circularly polarized light. In this case, it can be shown (Finazzi et al. 2007) that all the three FW absorption transitions described by the matrix elements in Equation 28.9 involve FW photons characterized by an L_z quantum number M_ω either equal to +1 or −1 according to the sign of the circular polarization (right or left) of the FW light impinging on the particle. By recalling that dipole terms are characterized by $L^2 = 1$ and quadrupole terms by $L^2 = 2$, the lowest allowed emission multipoles participating in the SHG process can be obtained from Table 28.1 in a straightforward manner. The allowed SH emission modes are listed in Table 28.3. As anticipated above, the lowest order transition E1 + E1 → E1 is forbidden under circularly polarized FW plane-wave illumination for particles displaying central symmetry, but it is allowed for noncentrosymmetric, noncylindrical particles. In this case, SH emission is given by an E1 term with $M_{2\omega} = 0$ or $M_{2\omega} = \pm 1$ for C_2 or C_{1h} symmetry, respectively. This emission mode corresponds to SHG from an electric dipole oriented parallel (for C_2 symmetry) or perpendicular (in the case of C_{1h} symmetry) to z. The lowest allowed SHG channel for spherical or cylindrical particles is E1 + E1 → E2 (Dadap et al. 1999, 2004), with $M_{2\omega} = \pm 2$. This SH emission mode corresponds to an irradiated SH power angular distribution displaying cylindrical symmetry around the propagation axis of the FW field, with a maximum in the plane perpendicular to z and a null in the forward and backward directions (Dadap et al. 1999, 2004), as displayed in Figure 28.5.

In the case of linearly polarized FW illumination, it is convenient to consider the selection rules that can be obtained by considering both the propagation direction of the FW plane wave (z axis) and the direction of the FW electric field vector (x axis) as the quantization axis. Therefore, one should calculate two multipole expansions of the FW wave, one for each choice of the quantization axis, and consider the possible values of both the L_z and L_x photon quantum numbers. Moreover, both axes should be taken into account to define the particle symmetry group. Let M_ω and $M_{2\omega}$ be the FW and SH photon L_z quantum numbers, and N_ω and $N_{2\omega}$ be the FW and SH photon L_x quantum numbers, respectively. It can be shown (Finazzi et al. 2007) that the multipole expansion

TABLE 28.3 SHG Selection Rules for the Different Particle Symmetries Illustrated in Figure 28.4, under Circularly Polarized Far-Field Plane-Wave Illumination

Symmetry and Point Group	E1 + E1 → E1	E1 + E1 → E2	E1 + E2 → E1	E1 + M1 → E1	E1 + E1 → M1
Spherical	Forbidden	Allowed for $M_{2\omega} = +2$ or -2	Forbidden	Forbidden	Forbidden
Cylindrical	Forbidden	Allowed for $M_{2\omega} = +2$ or -2	Forbidden	Forbidden	Forbidden
Central, S_2	Forbidden	Allowed	Allowed	Allowed	Allowed
Axial, C_2	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$
Reflection, C_{1h}	Allowed for $M_{2\omega} = \pm 1$	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$

Source: Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.

Note: Every transition multipole is considered separately. The propagation direction (z axis) is chosen as the quantization axis.

TABLE 28.4 SHG Selection Rules for the Different Particle Symmetries Illustrated in Figure 28.4, under Linearly Polarized Far-Field Plane-Wave Illumination

Symmetry and Point Group	E1 + E1 → E1	E1 + E1 → E2	E1 + E2 → E1	E1 + M1 → E1	E1 + E1 → M1
Spherical	Forbidden	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Forbidden
Cylindrical	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Forbidden
Central, S_2	Forbidden	Allowed	Allowed	Allowed	Allowed
Axial, C_2	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$
Reflection, C_{1h}	Allowed for $M_{2\omega} = \pm 1$	Allowed for $M_{2\omega} = 0, \pm 2$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$	Allowed for $M_{2\omega} = 0$

Source: Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.
Note: The propagation direction (z axis) is chosen as the quantization axis.

TABLE 28.5 SHG Selection Rules for the Different Particle Symmetries Illustrated in Figure 28.4, under Linearly Polarized Far-Field Plane-Wave Illumination

Symmetry and Point Group	E1 + E1 → E1	E1 + E1 → E2	E1 + E2 → E1	E1 + M1 → E1	E1 + E1 → M1
Spherical	Forbidden	Allowed for $N_{2\omega} = 0$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = \pm 1$	Forbidden
Cylindrical	Allowed for $N_{2\omega} = 0$	Allowed for $N_{2\omega} = 0, \pm 2$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = 0$
Central, S_2	Forbidden	Allowed	Allowed	Allowed	Allowed
Axial, C_2	Allowed for $N_{2\omega} = 0$	Allowed for $N_{2\omega} = 0, \pm 2$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = 0$
Reflection, C_{1h}	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = 0, \pm 2$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = \pm 1$	Allowed for $N_{2\omega} = 0$

Source: Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007.
Note: The electric field polarization direction (x axis) is chosen as the quantization axis.

of the FW plane wave corresponds to E1, M1, and E2 terms in Equation 28.9 that involve FW photons with an M_ω quantum number that can be either +1 or −1 but not 0. On the other hand, the expansion obtained by choosing the x axis as the quantization axis shows that the E1, M1, and E2 terms correspond to transitions involving FW photons with L^2 and L_x quantum numbers restricted to ($L_\omega = 1$, $N_\omega = 0$), ($L_\omega = 1$, $N_\omega = \pm 1$), and ($L_\omega = 2$, $N_\omega = \pm 1$), respectively. With these figures in mind and the general rules summarized in Table 28.1, one can find the selection rules listed in Tables 28.4 and 28.5. From these, one can see that the E1 + E1 → E1 SHG channel can be excited by linearly polarized FW plane-wave illumination in all noncentrosymmetric particles (Finazzi et al. 2007). In particular, if the particle exhibits C_2 symmetry either around the z or the x axis, SH emission will be given by an oscillating dipole parallel to the C_2 symmetry axis (Finazzi et al. 2007). In centrosymmetric particles illuminated by linearly polarized light, electric dipole SH emission is always given by a dipole parallel to z, while the quadrupole SH emission pattern is symmetric around x (see Figure 28.4) (Finazzi et al. 2007). M1 emission is forbidden (Dadap et al. 1999, 2004) because the selection rules would lead to $M_{2\omega} = N_{2\omega} = 0$, which correspond to a monopole term that cannot radiate (Finazzi et al. 2007).

28.6.2 Illumination with a Laterally Limited Light Beam

When the particle is illuminated by a laterally limited FW beam, which might be obtained by focusing with a high numerical-aperture objective or even with near-field illumination, and the beam waist is comparable with the particle size, the beam cannot be considered as a plane wave any longer. In this case, other emission channels than those listed in Tables 28.3 through 28.5 become available. In these conditions, in fact, the particle experiences large longitudinal components of the FW field, which are absent in the case of illumination with a plane wave. The presence of such longitudinal components can be easily understood for a converging beam focused by an objective, where the FW light wave vector **k** has a broad angular distribution determined by the objective numerical aperture. Intense longitudinal components should also be expected in the proximity of near-field probes (Novotny and Hecht 2006). The consequence is that SHG can be induced by a combined FW photon absorption from perpendicular field components acting on the particle (Figliozzi et al. 2005), which is not possible for plane-wave illumination. For instance, illuminating with a laterally confined beam, absorption of FW photons with L_z quantum number $M_\omega = 0$ from a linearly polarized FW field becomes possible, which is excluded for plane-wave illumination (Finazzi et al. 2007). For the same reason, FW photons with $M_\omega = 0, \pm 1, \pm 2$ are available for E2 absorption (Finazzi et al. 2007).

Another important difference with respect to in-plane wave illumination is represented by the fact that the rapid spatial variations of the field intensity can increase the relative weight of the higher-order multipoles in the SHG process. This is readily understood in far-field illumination when the spot size, which is limited by diffraction to about $1/k$, is comparable with the particle size a, resulting in a violation of the long-wavelength approximation. Similarly, strong field gradients are expected at the apex of a near field probe (Novotny and Hecht 2006), again invalidating the long-wavelength approximation and implying that high-order multipoles cannot be a priori neglected.

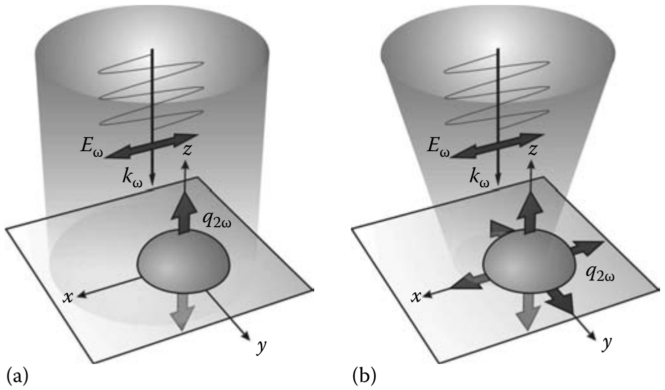

FIGURE 28.6 Allowed E1 second-harmonic emission channels for a particle with C_2 symmetry (axial particle over a substrate) excited by (a) a plane wave or (b) a laterally limited beam. The allowed directions of the particle electric dipole $q_{2\omega}$ generating the E1 second-harmonic radiation are indicated by the double-headed arrows on the particle. (Reprinted from Finazzi, M. et al., *Phys. Rev. B*, 76, 125414, 2007. With permission.)

As a relevant example, let us address SHG in lithographed particles on a substrate (see Section 28.7), such as those employed in surface-enhanced spectroscopy techniques. The presence of the substrate breaks the inversion symmetry with respect to the particle center. For ellipsoidal-shaped dots as those described by Grand et al. (2003) and Zavelani-Rossi et al. (2008), the particle symmetry group reduces to C_{2v} (see Figure 28.6). In this case, the selection rules governing SHG in C_2 symmetry around the z axis must apply, z being the substrate normal, which is taken parallel to the optical axis. The particles also belong to the C_{1h} symmetry group defined with respect to either the x or y axis, oriented parallel to the in-plane principal axes of the ellipsoids. However, this further symmetry does not provide more restrictive selection rules.

When such particles are excited with an FW plane wave with $\mathbf{k} \parallel z$ and considering only the lowest-order E1 + E1 → E1 transitions, M_ω can only assume the values ±1 both for linearly and circularly polarized light, so parity is conserved only for $M_{2\omega} = 0$, corresponding to SH emission from an electric dipole parallel to both $\mathbf{k}$ and z (see Figure 28.6) (Finazzi et al. 2007). When the particle is illuminated by a focused FW beam or by the near field of a tip, photons with $M_\omega = 0$ can be absorbed. This results in new SHG channels, namely the ones that correspond to SH radiation from electric dipoles lying in the plane of the substrate, becoming available (see Figure 28.6) (Finazzi et al. 2007). In this case, the details of the SH-radiated intensity and polarization will depend on the particle fine structure, which defines the relative strength of each channel.

28.7 Second Harmonic Generation in Single Gold Nanoparticles

As discussed in Section 28.3, the nonlinear optical response of nanostructured systems has been experimentally studied, so far, mainly by far-field techniques. These, however, are limited by diffraction and do not allow a direct mapping of the nanoscale field enhancement. These limitations can be overcome by SNOM, which can reach sub-100 nm lateral resolution. This technique consists of bringing the sample in interaction with the near field of a source (a tip or a backilluminated aperture in a metal-coated probe). In this case, the bandwidth of spatial frequencies associated with the evanescent waves from the probe is unlimited and the resolution can in principle be arbitrarily optimized. The response of the probe–sample interaction is recorded with standard far-field collection optics. State-of-the-art lateral resolution for aperture probes is around 50 nm. Near-field nonlinear optical microscopy is a powerful tool to characterize local field enhancements in metal nanostructures since it combines the great sensitivity of nonlinear optical response with the spatial superresolution of near-field microscopy. However, the technical challenges of combining scanning probe and ultrafast technologies and the typical low peak power available at the output of optical near field probes have so far limited the number of studies of SHG from single nanoparticles in closely packed arrangements (Biagioni et al. 2007, Breit et al. 2007, Celebrano et al. 2007, 2008a,b, Zavelani-Rossi et al. 2008).

This section describes SHG from single gold nanoparticles obtained with a SNOM setup. As illustrated in Section 28.6.2, the high lateral confinement typical of near-field illumination enables the observation and exploitation of unusual and peculiar SHG modes. Moreover, the acquisition of SH maps provides complementary information to FW images, which typically results from a complex interplay between scattering, absorption, and reflection, all contributing to light extinction (Novotny and Hecht 2006). The comparison between FW and SH maps therefore allows for a better interpretation of the optical response of metal nanoparticles and discriminating among different light extinction particle behaviors that would not be possible to address by just collecting the FW signals (Biagioni et al. 2007, Celebrano et al. 2007, 2008a,b, Zavelani-Rossi et al. 2008).

28.7.1 Near-Field Microscopy Setup

To observe SHG at the nanoscale, we combine high peak power and high spatial resolution by coupling femtosecond pulses to a hollow-pyramid aperture SNOM (see Figure 28.7). The pulses

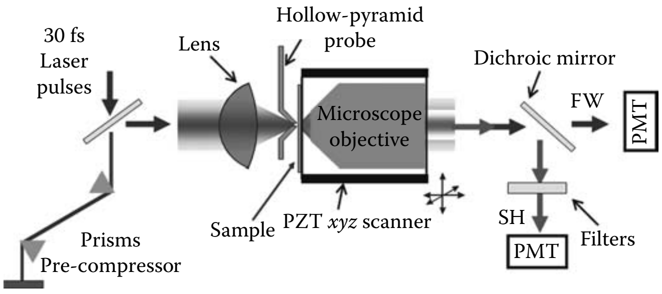

FIGURE 28.7 Schematics of our SNOM, with the ultrashort laser beam coupled to the measurement head: The collection optics, filters, and detector are also indicated. PMT indicates a photomultiplier tube.

are generated by a long-cavity mode-locked Ti:sapphire oscillator (26 MHz repetition rate), producing 20 nJ, 27 fs pulses at 800 nm. The SNOM probe consists of a silicon nitride cantilever with a hollow pyramid tip. The tip is aluminum-coated with a circular aperture at the apex, with diameter ranging between 100 and 200 nm. This probe offers several advantages compared with metal-coated tapered optical fibers, such as larger taper angle, lower absorption, preservation of light polarization (Biagioni et al. 2005), and pulse duration (Labardi et al. 2005) at the output. These improvements enable achieving peak powers more than two orders of magnitude higher in the near field, resulting in greatly enhanced nonlinear optical response of the sample. Typical tip throughputs at 800 nm range between ~10^{-4} for the 100 nm tips and 5×10^{-3} for the 200 nm ones. The tips with larger aperture were selected for the experiments presented in this section. Tip-sample distance is controlled by an optical lever that is sensitive enough to allow nondestructive contact mode stabilization on the samples.

The FW/SH light is collected in the far field by a long working distance microscope objective (numerical aperture = 0.75) and split by a dichroic filter. The FW and the SH are simultaneously detected by photomultiplier tubes. Bandpass filters are inserted in the SH beam to further reject the FW and TPPL from the sample. A 30 nm bandpass interference filter centered at 405 nm is inserted in the SH beam path to reject TPPL from the sample, which is negligible for wavelength shorter than 450 nm (Bouhelier et al. 2005, Imura et al. 2005, 2006). The quadratic dependence of SH intensity on the excitation beam power was verified. A mechanical chopper with a 1:6 duty cycle allows increasing the peak power for a given level of average power, typically 1 mW incident on the tip, and lock-in detection improves the signal-to-noise ratio, allowing for faster scans (integration time 30–100 ms per point).

Hereafter, we present two different types of gold nanoparticles. The first ones are triangles (height 15–25 nm) on a glass substrate obtained from a projection pattern (“Fischer pattern” (Fischer and Zingsheim 1981)), in a hexagonal array with 453 nm periodicity. Such a sample is chosen as an example of a SNOM study of SHG in a dense network of particles. The second type of particles are well-separated isolated ellipsoids (height about 60 nm) produced by electron beam lithography on a quartz substrate (Grand et al. 2003) in square arrays.

28.7.2 Fundamental Wavelength and Second Harmonic Maps

Figure 28.8a shows the topography of the Fischer pattern together with its SH optical image (Figure 28.8b). The topography shows the regular array of gold triangles: most of them are well separated, some are in contact with each other, and big defects cover portions of the sample. In the FW image (not shown), triangles appear dark, although the resolution is poor due to the quite large aperture diameter. Nevertheless, contrasted and well-resolved SHG maps from the gold triangles are detected. The background signal is attributed to SHG from both the glass substrate and the tip edges. The inset of Figure 28.8b shows a line profile of the SH image, demonstrating very good signal-to-noise ratio, high contrast (3:1), and good spatial resolution (better than 100 nm, see Figure 28.8c). Figure 28.8 highlights the unique capability of nonlinear SNOM to image SHG from closely packed metal nanostructures. It is interesting to note that not all the triangles observed in topography emit SH radiation with the same efficiency: some of them display an intense SH emission, while others are nearly dark. This supports the absence of topographical artifacts (Hecht et al. 1997) in the image, proved by the fact that in further measurements (not shown here), the topographic resolution was missing, yet a clear and well-resolved SHG image could still be collected. The high variability of the SHG signal is addressed further below when SHG in isolated nanoparticles is discussed.

The lithographed nanorods are about 60 nm high, with a short axis of about 70 nm and long axes equal to 100, 150, or 400 nm. The array period is 1 μm. According to the Mie theory, the plasmon resonances in elongated structures depend on their aspect ratio. Indeed, far-field extinction spectra (see Figure 28.9), with exciting white light beam polarization parallel to the major axis, display a peak around 690 nm for the 100 nm ellipsoids, around 800 nm (i.e., resonant with our FW) for the 150 nm ones and more than 1000 nm for the 400 nm ones (not shown). These three distinct particle geometries thus provide different resonating regimes that correspond to different linear and nonlinear optical behaviors and near-field properties. These are highlighted in Figure 28.10, showing the particle topography, FW near-field extinction, and SH emission obtained with the FW illumination linear polarization parallel to the nanorod long axis. The optical images at both

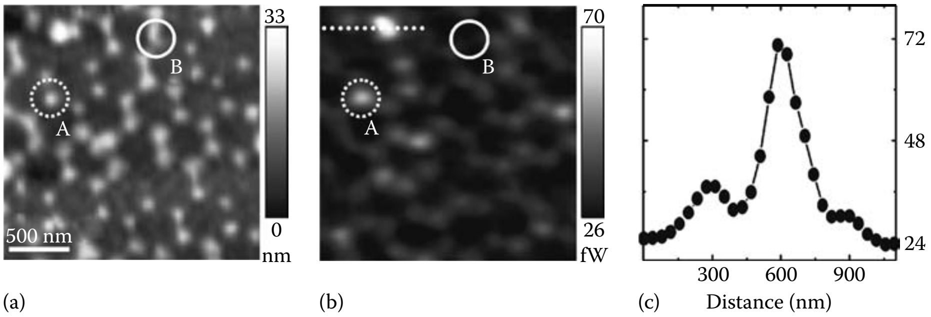

(a) (b) (c)

FIGURE 28.8 Projection pattern: (a) topography and (b) SH SNOM image (size: $5 \times 2.9\,\mu m^2$). The dashed circle (A) shows an SH emitting triangle and solid circle (B) shows a dark one. (c) Cross section from the raw data of the SH image, along the dotted line.

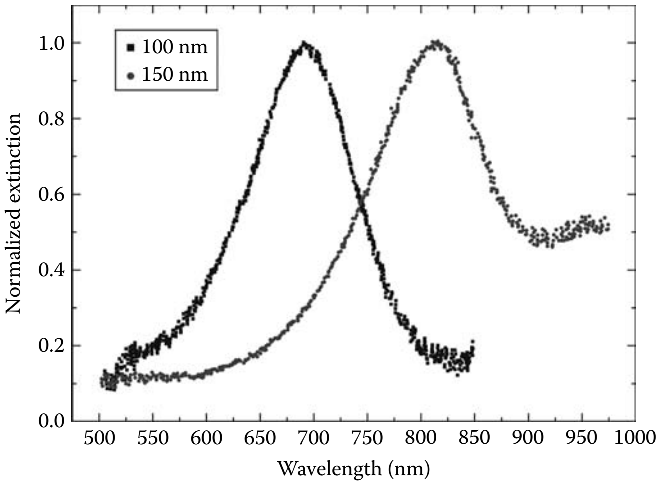

FIGURE 28.9 Extinction spectra of gold nanorods with a major axis length either equal to 100 (squares) or 150 nm (dots). The spectra have been normalized to each other at the maximum extinction. The light polarization is parallel to the particle long axis. The extinction spectra of the 400 nm-long nanorods also, discussed in the text, are not reported since the longitudinal plasmon resonance falls off the range of the analyzer. (Courtesy of J. Grand and P.-M. Adam.)

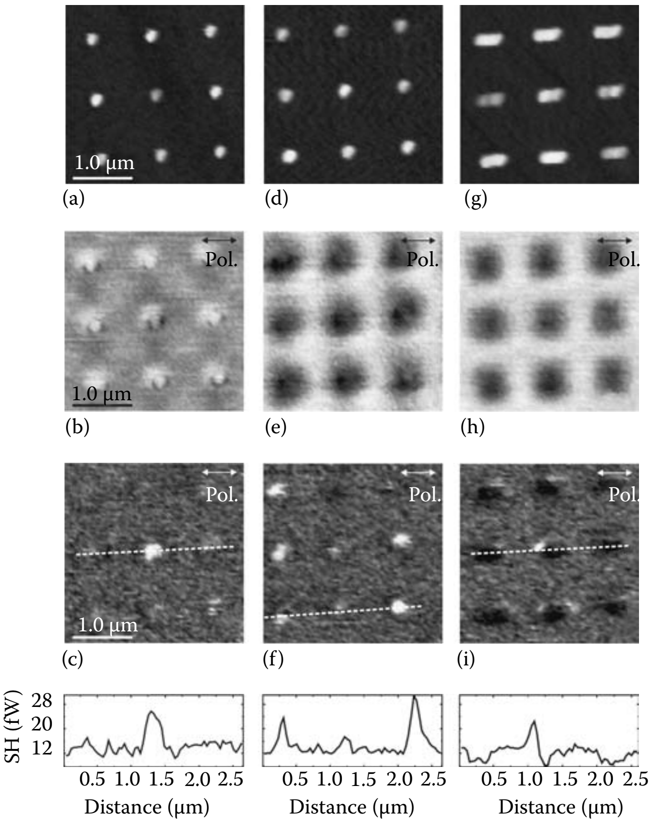

FIGURE 28.10 **(See color insert following page 20-14.)** Nanoparticles: (a), (d), and (g) topography; (b), (e), and (h) FW transmission; and (c), (f), and (i) SH emission SNOM images with corresponding cross sections along the dashed lines, from the raw data. Incident light is polarized parallel to the major axis. The particle major axis lengths are 100 nm (a)–(c), 150 nm (d)–(f), and 400 nm (g)–(i). Image size: $3 \times 3\,\mu\text{m}^2$. (Reprinted from Zavelani-Rossi, M. et al., *Appl. Phys. Lett.*, 92, 093119, 2008. With permission.)

the FW and the SH wavelengths strongly depend on the particle size. In particular, in FW maps, all the 100 nm ellipsoids appear bright, while all the 150 and 400 nm ellipsoids appear dark. In SH images, (1) most of the 100 nm-long particles do not emit, (2) most of the resonant 150 nm-long particles emit uniformly with high contrast (see line profiles in Figure 28.10), and (3) the 400 nm-long particles appear darker than the substrate. A general picture for these results can be obtained by a combined analysis of the FW and SH maps of the metal nanostructures. As apparent from Figure 28.10, the SH maps provide complementary information with respect to the FW images. For instance, only by comparing the SH emission properties of the 150 and 400 nm-long particles, it is possible to recognize that these fall in different optical regimes, due to the FW excitation being on or off resonance, respectively, with respect to the particle LSP excitations.

The near-field FW properties of the particles can be understood by recalling that the particle-to-background contrast in the near-field extinction maps results from the interference between an FW wave propagating into the far-field and a nonpropagating near-field distribution (Mikhailowsky et al. 2003). As a general rule, a metal particle should appear brighter than the background when the frequency of the excitation is higher than that of the particle LSP resonance, while it should be darker when the frequency of the FW excitation is set below the resonance (Mikhailowsky et al. 2003). At the LSP resonance, the particle should show no contrast with respect to the background. The presence of these interference effects explains the trend observed in the FW maps displayed in Figure 28.10, where the particles have a well-controlled aspect ratio, but it may hinder the interpretation of FW near-field images in other contexts, especially when the frequency of the LSP resonances of the particles is not known. However, the resonating behavior of the 150 nm-long nanorods becomes apparent from the SH maps, since these particles efficiently emit SH radiation as a result of the presence of strongly enhanced and localized electric field associated with the particle LSP oscillations, which are effectively excited only for these particles. Instead, most of the 100 nm-long particles do not emit SH radiation, and the 400 nm-long particles, being larger than the tip aperture, strongly absorb/scatter the SH light generated at the tip and appear darker than the background.

The nonlinear optical behavior of off-resonance particles is further highlighted in Figure 28.11, where FW and SH maps are collected by exciting the nanorods with an FW field parallel to the rod minor axis. With this type of excitation, the LSP resonance frequencies are shifted toward the blue region of the spectrum and cannot be excited by the FW light. In this case, SHG is observed at correspondence with the high-curvature regions of the particle, as a consequence of highly enhanced and localized electric fields by the lightning rod effect, in agreement with the theoretical predictions of Beermann and Bozhevolnyi (2004) and Stockman et al. (2004). It is interesting to note that, in Figure 28.11, the bright spots in the SH maps seem to correspond to

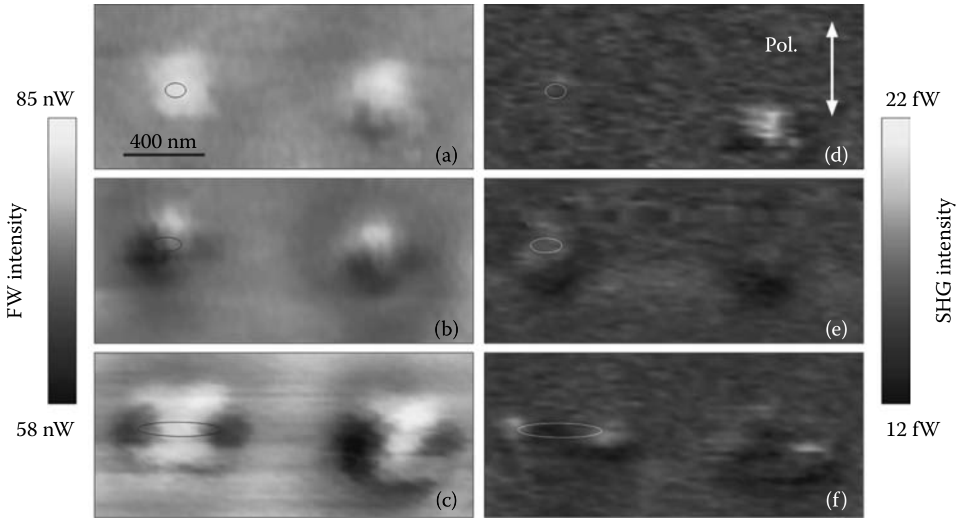

FIGURE 28.11 (a)–(c) FW transmission and (d)–(f) SH emission excited by incident FW light polarized perpendicular to the particle major axis. The particle major axis lengths are 100 nm (a) and (d), 150 nm (b) and (e), and 400 nm (c) and (f). (Reprinted from Zavelani-Rossi, M. et al., *Appl. Phys. Lett.*, 92, 093119, 2008. With permission.)

dark spots in the FW images, that is, the high-curvature regions where strong fields are localized also correspond to areas where larger FW scattering/absorption occurs, giving rise to a larger local linear extinction cross section. We believe that intense localized defects are also responsible for the high variability of the SHG efficiency displayed by nominally identical particles displaying quite similar FW (see Figures 28.8 and 28.10). In this case, local field enhancement would be associated with high-curvature defects in the particles.

To summarize, the SHG efficiency depends on the particle fine structure that determines (1) local FW field enhancements, induced either by LSP excitation or by lightning rod effects in areas with high-curvature or local imperfections and (2) the modes contributing to the emission process, which are described in the next section.

28.7.3 Second Harmonic Polarization Analysis

This section shows how the analysis of the polarization of the SH-emitted light might allow one to recognize which emission modes are contributing to the SHG process for a given particle geometry and illumination conditions. Such analysis will be applied to the same lithographed isolated gold particles, already described in Section 28.7.2.

The polarization state of the SH light emitted by the gold nanoellipsoids is analyzed as illustrated schematically in Figure 28.12. The FW light is polarized parallel to the major axes of the 150 nm-long particles to excite their strong longitudinal SPT resonance. The SH radiation polarization is analyzed by a polarizer on detection. Figure 28.13 shows two typical SH maps that are obtained with the analyzer parallel (Figure 28.13b) and perpendicular (Figure 28.13c) to the excitation polarization. From these maps, one can see that the SH light emitted by the nanoparticle is polarized parallel to its short axis (see Figure 28.13b,c,d). This particular polarization pattern of the emitted SH light is incompatible with the lowest allowed emission modes one should expect for such particles for an FW plane-wave illumination. In this case, in fact, SH radiation would be emitted by a dipole oscillating perpendicularly to the substrate sustaining the particle, as illustrated in Figure 28.6a. Such an emission mode would give the same SH maps independently from the direction of the analyzer axis. This is obviously in contradiction with the experimental results shown in Figure 28.13 and is a demonstration that a novel particle emission mode, which is forbidden for plane-wave illumination, becomes accessible. In fact, the presence of a symmetry-breaking substrate together with the strong longitudinal FW field component, typical of near-field illumination, allow for efficient excitation of an SH-emitting electric dipole perpendicular to both the incident FW light polarization and the detection axes,

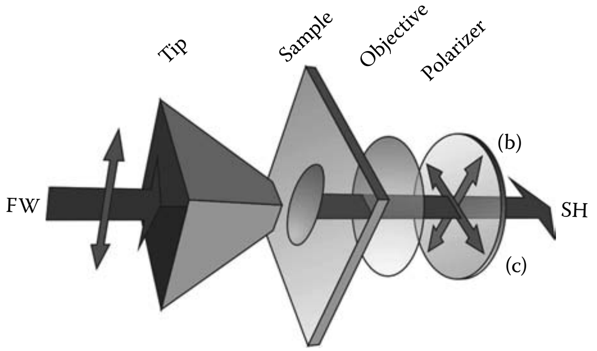

FIGURE 28.12 Experimental geometry of the polarization analysis setup in the SNOM microscope. The FW light is polarized parallel to the particle major axis. SH maps have been collected for the two analyzer directions (b) and (c), corresponding to SH linear polarizations parallel and perpendicular to the FW polarization, respectively. The corresponding SH maps are shown in Figure 28.13b and c, respectively.

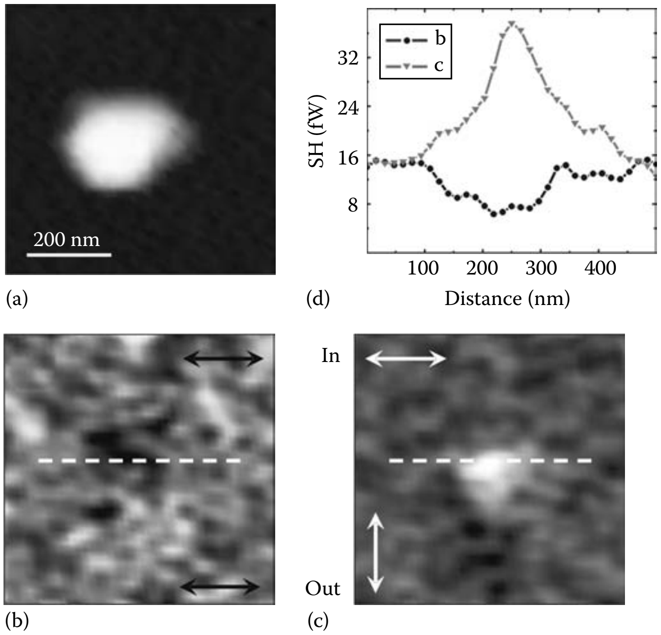

FIGURE 28.13 Nanoellipsoids with 150 nm major axis: topography (a), SH emission with a polarizer on detection parallel (b), and perpendicular (c) to the major axis (incident light is polarized parallel to the major axis). Image size: $0.6 \times 0.6\,\mu\text{m}^2$. (d) Intensity profiles from the raw data of images (b) (circles) and (c) (triangles).

as illustrated in Figure 28.6b. From the analysis of the SH polarization pattern in Figure 28.13, one can conclude that only the SH-emitting dipole perpendicular to the particle minor axis is significantly excited by the FW field. For this particle geometry, the SHG selection rules discussed in Section 28.4 do not exclude the possibility that SHG could be due to a dipole oscillating in the direction of the long axis of the particles, so there is no simple symmetry-based justification to the fact that such emission is not observed. This behavior has to be ascribed to the particular shape and resonance conditions of the particles. Such characteristics can in effect influence in a complex way the particle SH emission properties. To give a tentative explanation of this behavior, we remark that a much larger portion of gold–air interface is available along the long sides of the ellipsoid, so that a major contribution of SH light oriented parallel to the short axis might actually be expected as a surface contribution to SHG. Finally, we would like to stress that the spatial uniformity of the polarized SHG pattern measured on several particles is a hint of the negligible influence of particle defects, which can also lower the symmetry of the particles and contribute to SHG with modes that would otherwise be forbidden by parity and angular momentum conservation in a particle with a perfectly controlled shape.

28.8 Perspectives

A very promising perspective for future nano-optics is the possibility to simultaneously control of the spatial and temporal properties of the optical near field in the vicinity of a nanostructure by illumination with broadband optimally polarization- and time-shaped femtosecond light pulses. Adaptive shaping of the phase and amplitude of femtosecond laser pulses has been developed into an efficient tool for the directed manipulation of interference phenomena, thus providing coherent control over various quantum-mechanical systems. The interest to extend coherent control methods to nanostructures consists in overcoming the spatial limitation due to diffraction. Coherent control of the spatial and temporal evolution of optical near-fields by amplitude and polarization pulse shaping in plasmonic nanostructures has recently been theoretically proposed (Stockman et al. 2002, Brixner et al. 2005, 2006, Sukharev and Seideman 2007) and experimentally demonstrated (Aeschlimann et al. 2007) by applying an adaptative method based on measuring the local fields with two-photon photoemission electron microscopy. Progress in the coherent control of electric fields in nanostructures thus might open the access to simultaneously ultrafast and nanoscale new physics.

Nonlinear near-field optical microscopy might thus play an extremely important role in the development of nanoscale coherent control, since it can be employed as a probe of local field enhancement. In this frame, the temporal dynamics of the nonlinear phenomenon is, however, extremely important. SHG represents an ideal candidate to this task since the conversion of two FW photons into an SH photon is itself a coherent process (Imura et al. 2005), at variance from, for example, TPPL, which is incoherent in noble metal nanostructures (Imura et al. 2005). Indeed, we have recently demonstrated (Biagioni et al. 2009) that the TPPL yield in gold nanowires becomes independent of the pulse duration for laser pulses shorter than about 1 ps, while an inverse proportionality with the pulse duration would be expected for a fixed pulse energy. The origin of this behavior consists of the fact that TPPL in gold nanostructures is obtained after a 3*d* hole is produced by two sequential absorption steps involving a single photon (Imura et al. 2005), and is governed by the relaxation time of the *sp* conduction band hole generated after the first photon absorption. This temporal limitation would advise against the utilization of TPPL in probing near-field enhancement for coherent control applications, leaving SHG as a better adapted tool.

28.9 Conclusions

Thanks to the combination of surface sensitivity and symmetry selectivity, SHG is a very powerful tool to investigate the nanoscale properties of matter. The application of SHG techniques for investigating the structure and various other properties of small particles is a subject of considerable current interest and has stimulated a conspicuous number of both theoretical and experimental works in a broad range of subjects. In particular, being SHG a nonlinear optical process whose yield depends on the square power of the FW field intensity, SHG has found many applications in the study of the electric field enhancement mechanisms in metal nanostructures, where the lightning rod effect associated with high-curvature particles or resonantly excited plasma oscillations can lead to extremely intense and highly localized electric fields. Since nanoscience is attracting an exponentially growing attention, SHG in nanostructured systems will be a very hot topic for many years to come.

Acknowledgments

We would like to particularly thank Prof. Klaus Sattler for his kind invitation to write this chapter. We are also indebted to many colleagues and friends, whom we had the opportunity to work and discuss with during their activity concerning this work. Among these, we would like to mention M. Allegrini, P.-M. Adam P. Biagioni, M. Celebrano, G. Grancini, J. Grand, M. Labardi, D. Polli, M. Savoini, and M. Zavelani-Rossi.

M. F. and L. D. are also affiliated to L-NESS (Laboratory for Nanometric Epitaxial Structures on Silicon and for Spintronics), which is an Interuniversity Centre between Politecnico di Milano and Università di Milano Bicocca. G. C. is also affiliated to ULTRAS (National Laboratory for Ultrafast and Ultraintense Optical Science) research center, which is founded by the National Institute of the Physics of Matter (INFM) of the National Research Council (CNR).

References

Abid, J.-P., Nappa, J., Girault, H. H., and Brevet, P.-F. 2004. Pure surface plasmon resonance enhancement of the first hyperpolarizability of gold core–silver shell nanoparticles. *J. Chem. Phys.* **121**: 12577.

Aeschlimann, M., Bauer, M., Bayer, D. et al. 2007. Adaptive subwavelength control of nano-optical fields, *Nature* **446**: 301–304.

Agarwal, G. S. and Jha, S. S. 1982. Theory of second harmonic generation at a metal surface with surface plasmon excitation. *Solid State Commun.* **41**: 499–501.

Antoine, R., Brevet, P. F., Girault, H. H., Bethell, D., and Schiffrin, D. J. 1997. Surface plasmon enhanced non-linear optical response of gold nanoparticles at the air/toluene interface. *Chem. Commun.* **1997**: 1901–1902.

Antoine, R., Pellarin, M., Palpant, B. et al. 1998. Surface plasmon enhanced second harmonic response from gold clusters embedded in an alumina matrix. *J. Appl. Phys.* **84**: 4532.

Bachelier, G., Russier-Antoine, I., Benichou, E., Jonin, C., and Brevet, P. 2008. Multipolar second-harmonic generation in noble metal nanoparticles. *J. Opt. Soc. Am. B* **25**: 955–960.

Beermann, J. and Bozhevolnyi, S. I. 2004. Microscopy of localized second-harmonic enhancement in random metal nanostructures. *Phys. Rev. B* **69**: 155429.

Beermann, J., Bozhevolnyi, S. I., and Coello, V. 2006. Modeling of nonlinear microscopy of localized field enhancements in random metal nanostructures. *Phys. Rev. B* **73**: 115408.

Biagioni, P., Polli, D., Labardi, M. et al. 2005. Unexpected polarization behavior at the aperture of hollow-pyramid near-field probes. *Appl. Phys. Lett.* **87**: 223112.

Biagioni, P., Celebrano, M., Polli, D. et al. 2007. Nonlinear optics and spectroscopy at the nanoscale with a hollow-pyramid aperture SNOM. *J. Phys. Conf. Ser.* **61**: 125–129.

Biagioni, P., Celebrano, M., Savoini, M., Grancini, G., Brida, D., Mátéfi-Tempfli, S., Mátéfi-Tempfli, M., Duò, L., Hecht, B., Cerullo, G., and Finazzi, M. 2009. Dependence of the two-photon photoluminescence yield of gold nanostructures on the laser pulse duration. *Phys. Rev. B* **80**: 045411.

Bickford, L., Sun, J., Fu, K. et al. 2008. Enhanced multi-spectral imaging of live breast cancer cells using immunotargeted gold nanoshells and two-photon excitation microscopy. *Nanotechnology* **19**: 315102.

Bloembergen, N. 1999. Surface nonlinear optics: A historical overview. *Appl. Phys. B: Lasers Opt.* **68**: 289–293.

Bloembergen, N. and Pershan, P. S. 1962. Light waves at the boundary of nonlinear media. *Phys. Rev.* **128**: 606.

Bouhelier, A., Beversluis, M. R., and Novotny, L. 2003. Characterization of nanoplasmonic structures by locally excited photoluminescence. *Appl. Phys. Lett.* **83**: 5041–5043.

Bouhelier, A., Bachelot, R., Lerondel, G., Kostcheev, S., Royer P., and Wiederrecht, G. P. 2005. Surface plasmon characteristics of tunable photoluminescence in single gold nanorods. *Phys. Rev. Lett.* **95**: 267405.

Boyd, R. W. 2003. *Nolinear Optics*. Academic Press, San Diego, CA.

Bozhevolnyi, S. I. and Lozovski, V. Z. 2000. Self-consistent model for second-harmonic near-field microscopy. *Phys. Rev. B* **61**: 11139.

Bozhevolnyi, S. I. and Lozovski, V. Z. 2002. Second-harmonic scanning optical microscopy of individual nanostructures. *Phys. Rev. B* **65**: 235420.

Bozhevolnyi, S. I., Beermann, J., and Coello, V. 2003. Direct observation of localized second-harmonic enhancement in random metal nanostructures. *Phys. Rev. Lett.* **90**: 197403.

Breit, M., Malkmus, S., Feldmann, J., and Danzebrink, H. U. 2007. Near-field second harmonic generation by using uncoated silicon tips. *Appl. Phys. Lett.* **90**: 093114.

Brixner, T., Garcia de Abajo, F. J., Schneider, J., and Pfeiffer, W. 2005. Nanoscopic ultrafast space-time-resolved spectroscopy. *Phys. Rev. Lett.* **95**: 093901.

Brixner, T., García de Abajo, F. J., Schneider, J., Spindler, C., and Pfeiffer, W. 2006. Ultrafast adaptive optical near-field control. *Phys. Rev. B* **73**: 125437.

Brudny, V. L., Mendoza, B. S., and Mochán, W. L. 2000. Second-harmonic generation from spherical particles. *Phys. Rev. B* **62**: 11152.

Brudny, V. L., Mochán, W. L., Maytorena, J. A., and Mendoza, B. S. 2003. Second harmonic generation from a collection of nanoparticles. *Phys. Stat. Sol. B* **240**: 518–526.

Campagnola, P. J. and Loew, L. M. 2003. Second-harmonic imaging microscopy for visualizing biomolecular arrays in cells, tissues and organisms. *Nat. Biotechnol.* **11**: 1356–1360.

Canfield, B. K., Kujala, S., Jefimovs, K., Turunen, J., and Kauranen, M. 2004. Linear and nonlinear optical responses influenced by broken symmetry in an array of gold nanoparticles. *Opt. Express* **12**: 5418–5423.

Canfield, B. K., Kujala, S., Laiho, K., Jefimovs, K., Turunen, J., and Kauranen, M. 2006. Chirality arising from small defects in gold nanoparticle arrays. *Opt. Express* **14**: 950–955.

Canfield, B. K., Husu, H., Laukkanen, J., Bai, B. et al. 2007. Local field asymmetry drives second-harmonic generation in noncentrosymmetric nanodimers. *Nano Lett.* **7**: 1251–1255.

Celebrano, M., Zavelani-Rossi, M., Biagioni, P. et al. 2007. Mapping local field distribution at metal nanostructures by near-field second-harmonic generation. *Proc. SPIE* **6641**: 66411E.

Celebrano, M., Zavelani-Rossi, M., Polli, D., et al. 2008a. Mapping local field enhancements at nanostructured metal surfaces by second-harmonic generation induced in the near field. *J. Microsc.* **229**: 233–239.

Celebrano, M., Biagioni, P., Finazzi, M., et al. 2008b. Near-field second-harmonic generation from gold nanoellipsoids. *Phys. Stat. Sol. C* **5**: 2657–2661.

Dadap, J. I. 2008. Optical second-harmonic scattering from cylindrical particles. *Phys. Rev. B* **78**: 205322.

Dadap, J. I., Shan, J., Eisenthal, K. B., and Heinz, T. F. 1999. Second-harmonic Rayleigh scattering from a sphere of centrosymmetric material. *Phys. Rev. Lett.* **83**: 4045.

Dadap, J. I., Shan, J., and Heinz, T. F. 2004. Theory of optical second-harmonic generation from a sphere of centrosymmetric material: Small-particle limit. *J. Opt. Soc. Am. B* **21**: 1328–1347.

Dewitz, J. P., Hübner, W., and Bennemann, K. H. 1996. Theory for nonlinear Mie-scattering from spherical metal clusters. *Z. Phys. D: At. Mol. Clusters* **37**: 75–84.

Downer, M. C., Mendoza, B. S., and Gavrilenko, V. I. 2001. Optical second harmonic spectroscopy of semiconductor surfaces: Advances in microscopic understanding. *Surf. Interface Anal.* **31**: 966–986.

Durr, N. J., Larson, T., Smith, D. K., Korgel, B. A., Sokolov, K., and Ben-Yakar, A. 2007. Two-photon luminescence imaging of cancer cells using molecularly targeted gold nanorods. *Nano Lett.* **7**: 941–945.

El-Sayed, I. H., Huang, X., and El-Sayed, M. A. 2006. Selective laser photo-thermal therapy of epithelial carcinoma using anti-EGFR antibody conjugated gold nanoparticles. *Cancer Lett.* **239**: 129–135.

Figliozzi, P., Sun, L., Jiang, Y. et al. 2005. Single-beam and enhanced two-beam second-harmonic generation from silicon nanocrystals by use of spatially inhomogeneous femtosecond pulses. *Phys. Rev. Lett.* **94**: 047401.

Finazzi, M., Biagioni, P., Celebrano, M., and Duò, L. 2007. Selection rules for second-harmonic generation in nanoparticles. *Phys. Rev. B* **76**: 125414.

Fischer, U. Ch. and Zingsheim, H. P. 1981. Submicroscopic pattern replication with visible light. *J. Vac. Sci. Technol.* **19**: 881–885.

Franken, P. A., Hill, A. E., Peters, C. W., and Weinreich, G. 1961. Generation of optical harmonics. *Phys. Rev. Lett.* **7**: 118.

Grand, J., Kostcheev, S., Bijoeon, J.-L. et al. 2003. Optimization of SERS-active substrates for near-field Raman spectroscopy. *Synth. Met.* **139**: 621–624.

Hao, E. C., Schatz, G. C., Johnson, R. C., and Hupp, J. T. 2002. Hyper-Rayleigh scattering from silver nanoparticles. *J. Chem. Phys.* **117**: 5963.

Hartschuh, A., Pedrosa, H. N., Novotny, L., and Krauss, T. D. 2003. Simultaneous fluorescence and Raman scattering from single carbon nanotubes. *Science* **301**: 1354–1356.

Hecht, B., Bielefeldt, H., Inouye, Y., and Pohl, D. W. 1997. Facts and artifacts in near-field optical microscopy. *J. Appl. Phys.* **81**: 2492.

Hua, X. M. and Gersten, J. I. 1986. Theory of second-harmonic generation by small metal spheres. *Phys. Rev. B* **33**: 3756.

Hubert, C., Billot, L., Adam, P.-M. et al. 2007. Role of surface plasmon in second harmonic generation from gold nanorods. *Appl. Phys. Lett.* **90**: 181105.

Imura, K., Nagahara, T., and Okamoto, H. 2005. Near-field two-photon-induced photoluminescence from single gold nanorods and imaging of plasmon modes. *J. Phys. Chem. B* **109**: 13214–13220.

Imura, K., Nagahara, T., and Okamoto, H. 2006. Photoluminescence from gold nanoplates induced by near-field two-photon absorption. *Appl. Phys. Lett.* **88**: 023104.

Jackson, J. D. 1975. *Classical Electrodynamics*. Wiley, New York.

Jakubczyk, D., Shen, Y., Lal, M., Kim, K.S., Świątkewicz, J., and Prasad, P.N. 1999. Near-field probing of nanoscale nonlinear optical processes. *Opt. Lett.* **24**: 1151–1153.

Jiang, Y., Wilson, P. T., Downer, M. C., White, C. W., and Withrow, S. P. 2001. Second-harmonic generation from silicon nanocrystals embedded in SiO_2. *Appl. Phys. Lett.* **78**: 766.

Jiang, Y., Sun, L., and Downer, M. C. 2002. Second-harmonic spectroscopy of two-dimensional Si nanocrystal layers embedded in SiO_2 films. *Appl. Phys. Lett.* **81**: 3034.

Johnson, R. C., Li, J., Hupp, J. T., and Schatz, G. C. 2002. Hyper-Rayleigh scattering studies of silver, copper, and platinum nanoparticle suspensions. *Chem. Phys. Lett.* **356**: 534–540.

Labardi, M., Zavelani-Rossi, M., Polli, D. et al. 2005. Characterization of femtosecond light pulses coupled to hollow-pyramid near-field probes: Localization in space and time. *Appl. Phys. Lett.* **86**: 031105.

Lüpke, G. 1999. Characterization of semiconductor interfaces by second-harmonic generation. *Surf. Sci. Rep.* **35**: 75–161.

Martorell, J., Vilaseca, R., and Corbalán, R. 1997. Scattering of second-harmonic light from small spherical particles ordered in a crystalline lattice. *Phys. Rev. A* **55**: 4520.

Mc Gilp, J. F. 1996. A review of optical second-harmonic and sum-frequency generation at surfaces and interfaces. *J. Phys. D* **29**: 1812–1821.

Mikhailowsky, A. A., Petruska, M. A., Stockman, M. I., and Klimov, V. I. 2003. Broadband near-field interference spectroscopy of metal nanoparticles using a femtosecond white-light continuum. *Opt. Lett.* **28**: 1686–1688.

Mochán, W. L., Maytorena, J. A., Mendoza, B. S., and Brudny, V. L. 2003. Second-harmonic generation in arrays of spherical particles. *Phys. Rev. B* **68**: 085318.

Mühlschlegel, P., Eisler, H.-J., Martin, O. J. F., Hecht, B., and Pohl, D. W. 2005. Resonant optical antennas. *Science* **308**: 1607–1609.

Nagesha, D., Laevsky, G. S., Lampton, P. et al. 2007. In vitro imaging of embryonic stem cells using multiphoton luminescence of gold nanoparticles. *Int. J. Nanomed.* **2**: 813.

Nappa, J., Revillod, G., Russier-Antoine, I., Benichou, E., Jonin, C., and Brevet P. F. 2005. Electric dipole origin of the second harmonic generation of small metallic particles. *Phys. Rev. B* **71**: 165407.

Nappa, J., Russier-Antoine, I., Benichou, E., Jonin, Ch., and Brevet, P. F. 2006. Second harmonic generation from small gold metallic particles: From the dipolar to the quadrupolar response. *J. Chem. Phys.* **125**: 184712.

Neacsu, C. C., Reider, G. A., and Raschke, M. B. 2005. Second-harmonic generation from nanoscopic metal tips: Symmetry selection rules for single asymmetric nanostructures. *Phys. Rev. B* **71**: 201402(R).

Nie, S. M. and Emery, S. R. 1997. Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. *Science* **275**: 1102–1106.

Novotny, L. and Hecht, B. 2006. *Principles of Nano-Optics*. Cambridge University Press, Cambridge, NY.

O'Neal, D. P., Hirsch, L. R., Halas, N. J., Payne, J. D., and West, J. L. 2004. *Cancer Lett.* **209**: 181.

Panasyuk, G. Y., Schotland, J. C., and Marke, V. A. 2008. Classical theory of optical nonlinearity in conducting nanoparticles. *Phys. Rev. Lett.* **100**: 047402.

Parameswaran, K. R., Kurz, J. R., Roussev, M. M., and Fejer, M. 2002. Observation of 99% pump depletion in single-pass second-harmonic generation in a periodically poled lithium niobate waveguide. *Opt. Lett.* **27**: 43–45.

Park, J., Estrada, A., Sharp, K. et al. 2008. Two-photon-induced photoluminescence imaging of tumors using near-infrared excited gold nanoshells. *Opt. Express* **16**: 1590–1599.

Pavlyukh, Y. and Hübner, W. 2004. Nonlinear Mie scattering from spherical particles. *Phys. Rev. B* **70**: 245434.

Russier-Antoine, I., Jonin, Ch., Nappa, J., Bénichou, E., and Brevet, P.-F. 2004. Wavelength dependence of the hyper Rayleigh scattering response from gold nanoparticles. *J. Chem. Phys.* **120**: 10748.

Sandrock, M. L., Pibel, C. D., Geiger, F. M., and Foss, Jr. C. A. 1999. Synthesis and second-harmonic generation studies of noncentrosymmetric gold nanostructures. *J. Phys. Chem. B* **103**: 2668–2673.

Shan, J., Dadap, J. I., Stiopkin, I., Reider, G. A., and Heinz, T. F. 2006. Experimental study of optical second-harmonic scattering from spherical nanoparticles. *Phys. Rev. A* **73**: 023819.

Shen, Y. R. 1984. *The Principles of Nonlinear Optics*. Wiley, New York.

Shen, Y. R. 1989. Optical second harmonic generation at interfaces. *Annu. Rev. Phys. Chem.* **40**: 327–350.

Shen, Y., Markowicz, P., Winiarz, J., Swiatkiewicz, J., and Prasad, P. N. 2001. Nanoscopic study of second-harmonic generation in organic crystals with collection-mode near-field scanning optical microscopy. *Opt. Lett.* **26**: 725–727.

Singh, D. B. and Tripathi, V. K. 2007. Surface plasmon excitation at second harmonic over a rippled surface. *J. Appl. Phys.* **102**: 083301.

Smolyaninov, I. I., Zayats, A. V., and Davis, C. C. 1997. Near-field second harmonic generation from a rough metal surface. *Phys. Rev. B* **56**: 9290.

Smolyaninov, I. I., Liang, H. Y., Lee, C. H., Davis, C. C., Aggarwal, S., and Ramesh, R. 2000. Near-field second-harmonic microscopy of thin ferroelectric films. *Opt. Lett.* **25**: 835–837.

Stockman, M. I., Faleev, S. V., and Bergman, D. J. 2002. Coherent control of femtosecond energy localization in nanosystems. *Phys. Rev. Lett.* **88**: 067402.

Stockman, M. I., Bergman, D. J., Anceau, C., Brasselet, S., and Zyss, J. 2004. Enhanced second-harmonic generation by metal surfaces with nanoscale roughness: Nanoscale dephasing, depolarization, and correlations. *Phys. Rev. Lett.* **92**: 057402.

Su, H. M., Ye, J. T., Tang, Z. K., and Wong, K. S. 2008. Resonant second-harmonic generation in monosized and aligned single-walled carbon nanotubes. *Phys. Rev. B* **77**: 125428.

Sukharev, M. and Seideman, T. 2007. Coherent control of light propagation via nanoparticle arrays. *J. Phys. B: At. Mol. Opt. Phys.* **40**: S283–S298.

Tinkham, M. 1964. *Group Theory and Quantum Mechanics*. McGraw-Hill, New York.

Vance, F. W., Lemon, B. I., and Hupp, J. T. 1998. Enormous hyper-Rayleigh scattering from nanocrystalline gold particle suspensions. *J. Phys. Chem. B* **102**: 10091–10093.

Wang, H., Yan, E. C. Y., Borguet, E., and Eisenthal, K. B. 1996. Second harmonic generation from the surface of centrosymmetric particles in bulk solution. *Chem. Phys. Lett.* **259**: 15–20.

Yang, N., Angerer, W. E., and Yodh, A. G. 2001. Angle-resolved second-harmonic light scattering from colloidal particles. *Phys. Rev. Lett.* **87**: 103902.

Zavelani-Rossi, M., Celebrano, M., Biagioni, P. et al. 2008. Near-field second-harmonic generation in single gold nanoparticles. *Appl. Phys. Lett.* **92**: 093119.

Zayats, A.V., Kalkbrenner, T., Sandoghdar, V., and Mlynek, J. 2000. Second-harmonic generation from individual surface defects under local excitation. *Phys. Rev. B* **61**: 4545.

Zhu, J. 2007. Theoretical study of the tunable second-harmonic generation (SHG) enhancement factor of gold nanotubes. *Nanotechnology* **18**: 225702.

29

Nonlinear Optics in Semiconductor Nanostructures

29.1 Introduction 29-1
29.2 Semiconductor-Field Hamiltonian in $\mathbf{k}\cdot\mathbf{p}$-Approximation 29-2
29.3 Quantum Equations of Motion 29-6
Electromagnetic Field • Semiconductor Dynamics • Rotating Wave Approximation
29.4 Perturbation Theory: Four-Wave Mixing Response 29-11
29.5 Nonperturbative Methods: Rabi Oscillations 29-16
29.6 Conclusion 29-19
Acknowledgment 29-19
References 29-19

Mikhail Erementchouk
University of Central Florida

Michael N. Leuenberger
University of Central Florida

29.1 Introduction

The physics of semiconductor nonlinear response unites under the same cover a fascinating variety of different physical phenomena, theoretical ideas, and experimental approaches. It provides promising opportunities for technologies, for example, in the form of principally new light sources, and presents a convenient testing field for studying long-standing fundamental questions while raising new ones in fields ranging decoherence to low-energy quantum field theory.

It is not therefore, surprising, that the interest of researchers remains steadily persistent for already many decades. There are many books and reviews devoted to the details of different topics. However, as investigations move forward, new types of problems surface and the question of the basic principles of the theory of nonlinear response is raised again and again.

Perhaps the most clear demonstration of the recurring process of testing the background is the drastic change of the main theoretical methods, which took place in the 1990s. In the pre-1990s era, Green's functions method was the main technique used in the limit of zero, finite temperature, or nonequilibrium (the Keldysh technique). The main questions were formulated in terms of the spectral characteristics, and using the adiabatical approximation was the usual practice. With the development of new experimental methods and the improvement of the quality of the samples, the attention gradually moved toward the temporal behavior and the transient features of the excitations. As a result, in the post-1990s era, time dependence turned out to be the major question, and the typical theoretical approach became the method of the dynamical equations of motion. In a sense the whole topic was reinvented, making the question of fundamentals of the theory especially important.

In the following sections, we present the derivation of the main dynamical equations for the electromagnetic field and the semiconductor excitations in the general context of quantum field theory. Such an approach is particularly useful because the resonant interaction of light with the semiconductor results in constantly changing the number of elementary excitations in the semiconductor bands. Therefore, the respective description must admit the creation and annihilation of particles on the fundamental level. Also, in order to cover at least partially the various physical situations relevant for the problem of nonlinear response, we avoid using specific approximations such as, for example, the spherical symmetry of the hole's bands or the translational symmetry of the structures. The latter circumstance compels to formulate the major part of considerations in the coordinate representation. Finally, the main equations are derived for operators that makes them applicable for studying quantum statistics and other complex questions going beyond the single-particle correlation functions (Green's functions).

The structure of the presentation is as follows. In Section 29.2 we derive the quantum-field Hamiltonian in the $\mathbf{k}\cdot\mathbf{p}$-approximation and briefly review the basic properties of the semiconductor bands. Section 29.3 is devoted to the derivation of the quantum equations of motion for the electromagnetic field and the semiconductor excitations. In Section 29.4, the perturbative approach of analyzing the quantum equations of motion is reviewed. Finally, in Section 29.5 we consider the semiconductor Bloch equation.

29.2 Semiconductor-Field Hamiltonian in $\mathbf{k}\cdot\mathbf{p}$-Approximation

The quantum dynamics of the electrons moving in the periodic lattice field in the semiconductor under the action of the external field is described in SI units by the Hamiltonian

$$\mathcal{H}_{mat} = \sum_{s}\int d\mathbf{x}\,\psi_s^\dagger(\mathbf{x})\left\{\frac{1}{2m}[\mathbf{p}-e\mathbf{A}(\mathbf{x})]^2 + U_l(\mathbf{x}) + U_{ext}(\mathbf{x})\right\}\psi_s(\mathbf{x})$$
$$+\frac{1}{2}\sum_{s,s'}\int d\mathbf{x}_1\,d\mathbf{x}_2\,\psi_s^\dagger(\mathbf{x}_1)\psi_{s'}^\dagger(\mathbf{x}_2)V(\mathbf{x}_1-\mathbf{x}_2)\psi_{s'}(\mathbf{x}_2)\psi_s(\mathbf{x}_1), \tag{29.1}$$

where

$\psi_s(\mathbf{x})$ is the electron field operator
s and s' are the electron spin indices
$\mathbf{A}$ is the vector potential
$U_l(\mathbf{x})$ is the periodic lattice potential
$U_{ext}(\mathbf{x})$ is an external potential caused, for example, by the external static electric field
$V(\mathbf{x}) = e^2/4\pi\epsilon_0\epsilon_b|\mathbf{x}|$, with ϵ_b being the background dielectric constant, is the potential of the Coulomb interaction between the electrons

In the following paragraphs we will be interested in establishing the relation with the standard $\mathbf{k}\cdot\mathbf{p}$ approach, therefore, we will omit both the external potential and the Coulomb interaction and will consider their effect later.

The main idea of the $\mathbf{k}\cdot\mathbf{p}$ approach is to employ the fact that we are interested in the weak and spatially smooth perturbations of the semiconductor near its ground state. Thus, we want to account exactly the effect of the periodic lattice potential forming the semiconductor ground state and to consider the deviation from it in the spirit of the perturbation theory.

In order to account for the effect of the periodic potential on the electron dynamics we need to develop an appropriate description of the respective electron single-particle states. The field operator $\psi^\dagger(\mathbf{x})$ can always be expanded in terms of the Bloch modes as

$$\psi_s^\dagger(\mathbf{x}) = \sum_{n}\int d\mathbf{k}\,\phi_{n,\mathbf{k}}^\star(\mathbf{x})a_{n,\mathbf{k}}^\dagger, \tag{29.2}$$

where

n enumerates the bands and includes the spin index
the Bloch wave vector $\mathbf{k}$ lies in the first Brillouin zone
the wave functions $\phi_{n,\mathbf{k}}(\mathbf{x})$ give the solutions of the respective Schrödinger equation in the periodic potential
the operators $a_{n,\mathbf{k}}^\dagger$ are the operators, which create electrons in the respective Bloch states

According to the Bloch theorem the electron wave function $\phi_{n,\mathbf{k}}(\mathbf{x})$ can be presented as

$$\phi_{n,\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}u_{n,\mathbf{k}}(\mathbf{x}), \tag{29.3}$$

where $u_{n,\mathbf{k}}(\mathbf{x})$ is a periodic function solving the equation

$$\frac{\hbar^2}{2m}(-i\nabla+\mathbf{k})^2u_{n,\mathbf{k}}(\mathbf{x}) + U_l(\mathbf{x})u_{n\cdot\mathbf{k}}(\mathbf{x}) = \epsilon_n(\mathbf{k})u_{n,\mathbf{k}}(\mathbf{x}). \tag{29.4}$$

This equation is Hermitian and, hence, functions $u_{n,\mathbf{k}}(\mathbf{x})$ for a fixed $\mathbf{k} = \mathbf{k}_0$ constitute a complete set on the space of periodic functions. We can, therefore, represent functions corresponding to other $\mathbf{k}\neq\mathbf{k}_0$ as the respective Fourier series

$$u_{n,\mathbf{k}}(\mathbf{x}) = \sum_{m}\alpha_{n,m}(\mathbf{k};\mathbf{k}_0)u_{m,\mathbf{k}_0}(\mathbf{x}) \tag{29.5}$$

with unitary $\alpha_{n,m}(\mathbf{k};\,\mathbf{k}_0)$. The latter follows from orthonormality of the amplitudes $u_{n,\mathbf{k}}(\mathbf{x})$

$$\frac{1}{\mathcal{V}}\int d\mathbf{x}\,u_{n',\mathbf{k}}^\star(\mathbf{x})u_{n',\mathbf{k}}(\mathbf{x}) = \frac{1}{(2\pi)^3}\delta_{nn'}, \tag{29.6}$$

where $\mathcal{V}$ is the volume of the elementary cell. Indeed, substituting Equation 29.5 into Equation 29.6 we obtain

$$\sum_{m}\alpha_{n'm}^\star(\mathbf{k};\mathbf{k}_0)\alpha_{nm}(\mathbf{k};\mathbf{k}_0) = \delta_{nn'}, \tag{29.7}$$

which explicitly shows the unitarity of $\alpha_{n,m}(\mathbf{k};\,\mathbf{k}_0)$. This motivates the introduction of new (Luttinger–Kohn) function describing the electron states

$$\chi_{n,\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}}u_{n,\mathbf{k}_0}(\mathbf{x}). \tag{29.8}$$

These functions constitute a complete orthonormal set similarly to the electron wave functions $\phi_{n,\mathbf{k}}(\mathbf{x})$. The proof demonstrates the typical line of arguments and, therefore, it is useful to consider it in detail. The completeness of the Luttinger–Kohn functions

$$\sum_{n}\int d\mathbf{k}\,\chi_{n,\mathbf{k}}^\star(\mathbf{x})\chi_{n,\mathbf{k}}(\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}'). \tag{29.9}$$

follows from the respective relation for the electron wave functions

$$\sum_{n}\int d\mathbf{k}\,\phi_{n,\mathbf{k}}^\star(\mathbf{x})\phi_{n,\mathbf{k}}(\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}') \tag{29.10}$$

if we use Equations 29.5 and 29.7. For orthonormality we need to consider

$$(n,\mathbf{k}|n',\mathbf{k}') \equiv \int d\mathbf{x}\,\chi_{n,\mathbf{k}}^\star(\mathbf{x})\chi_{n',\mathbf{k}'}(\mathbf{x}) = \int d\mathbf{x}\,e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}}u_{n,\mathbf{k}_0}^\star(\mathbf{x})u_{n',\mathbf{k}_0}(\mathbf{x}). \tag{29.11}$$

The product $u^{*}_{n,\mathbf{k}_0}(\mathbf{x})u_{n',\mathbf{k}_0}(\mathbf{x})$ is a periodic function and therefore can be expanded in the Fourier series

$$u^{*}_{n,\mathbf{k}_0}(\mathbf{x})u_{n',\mathbf{k}_0}(\mathbf{x}) = \sum_{\mathbf{Q}} B_{nn'}(\mathbf{Q})e^{i\mathbf{Q}\cdot\mathbf{x}}, \tag{29.12}$$

where the sum runs over the vectors of the reciprocal lattice and

$$B_{nn'}(\mathbf{Q}) = \frac{1}{\mathcal{V}}\int d\mathbf{x}\, e^{-i\mathbf{Q}\cdot\mathbf{x}} u^{*}_{n,\mathbf{k}}(\mathbf{x})u_{n',\mathbf{k}}(\mathbf{x}). \tag{29.13}$$

Using this representation in Equation 29.11 we obtain

$$(n,\mathbf{k}|n',\mathbf{k}') = (2\pi)^3 \sum_{\mathbf{Q}} B_{nn'}(\mathbf{Q})\delta(\mathbf{k}-\mathbf{k}'+\mathbf{Q}). \tag{29.14}$$

The Bloch-wave-vectors $\mathbf{k}$ and $\mathbf{k}'$ are inside the first Brillouin zone and, therefore, the only possibility to reach the singularity point of the δ-function is when $\mathbf{Q} = 0$ leading to $(n,\mathbf{k}|n',\mathbf{k}') = (2\pi)^3 B_{nn'}(0)\delta(\mathbf{k}-\mathbf{k}')$. Taking into account Equation 29.6 we arrive at

$$(n,\mathbf{k}|n',\mathbf{k}') = \delta_{nn'}\delta(\mathbf{k}-\mathbf{k}'). \tag{29.15}$$

Such introduced functions $\chi_{n',\mathbf{k}'}(\mathbf{x})$ provide a general background and are used in different contexts in solid-state physics (Luttinger and Kohn 1955, Johnson and Hui 1993). In what follows we limit ourselves to the case $\mathbf{k}_0 = 0$ and will omit the index corresponding to the Bloch-wave-vector, simplifying notations.

In terms of the functions $\chi_{n,\mathbf{k}}(\mathbf{x})$ the electron field operator is presented as

$$\psi^{\dagger}_{s}(\mathbf{x}) = \sum_{n}\int d\boldsymbol{k}\,\chi^{*}_{n,\mathbf{k}}(\mathbf{x})b^{\dagger}_{n,\mathbf{k}}. \tag{29.16}$$

Substituting this representation into Equation 29.1, adopting the Coulomb gauge for the vector potential, $\nabla\cdot\mathbf{A}(\mathbf{x}) \equiv 0$, taking into account Equation 29.4 and using the same trick as for Equation 29.15 we obtain

$$\begin{aligned}\mathcal{H}_{mat} &= \sum_{n,n'}\int d\mathbf{k}\left[\left(\epsilon_n \frac{\hbar^2 k^2}{2m}\right)\delta_{n,n'} + \frac{\hbar}{m}\mathbf{k}\cdot\mathbf{p}_{n,n'}\right]b^{\dagger}_{n,\mathbf{k}}b_{n',\mathbf{k}} \\ &\quad -\frac{e}{m}\sum_{n,n'}\int d\mathbf{k}\,d\mathbf{q}\left[\mathbf{A}(\mathbf{q})(\mathbf{p}_{n,n'} + \hbar\mathbf{k}\delta_{n,n'}) - \frac{e}{2}A_2(\mathbf{q})\delta_{n,n'}\right] \\ &\quad \times b^{\dagger}_{n,\mathrm{k}+\mathrm{q}}b_{n',\mathrm{k}},\end{aligned} \tag{29.17}$$

where $A_2(\mathbf{q}) = (2\pi)^{-3/2}\int d\mathbf{x}\, A^2(\mathbf{x})e^{-i\mathbf{q}\cdot\mathbf{x}}$. The parameters of great importance introduced in Equation 29.17 are the matrix elements of the momentum operator between different bands

$$\mathbf{p}_{nn'} = -\frac{i\hbar}{\mathcal{V}}\int d\mathbf{x}\, u^{*}_{n}(\mathbf{x})\nabla u_{n'}(\mathbf{x}), \tag{29.18}$$

which are responsible for the *interband* coupling and are the main parameters quantifying the strength of the light–matter interaction.

The first term in the r.h.s. of Equation 29.17 describes the dynamics of the semiconductor excitations, and the second term accounts for the interaction with the external field. The specific feature of the "eigen-part" of the Hamiltonian (that is the part which is independent of the external field) is that it is not diagonal with respect to bands. The reason for this is the special status given to the states corresponding to $\mathbf{k} = \mathbf{k}_0 = 0$. These states do not exhaust the eigenstates of the semiconductor and, therefore, the effective coupling between different bands should not be surprising. This coupling, however, is proportional to $\mathbf{k}$. Thus, if the relevant excitations are relatively weak and smooth, then the main contribution into the semiconductor dynamics would come from small $\mathbf{k}$ and we can treat the respective band coupling as a perturbation. As the first step we could just drop this term. This, however, gives obviously wrong results for the dynamics near the extreme points of the bands. Indeed, the resulting Hamiltonian would predict that the mass of the excitations in the conductance band is the mass of electron in empty space and that the energy of states in the valence band increases with $\mathbf{k}$. Both of these predictions contradict the experiment. Thus, it is necessary to develop a consistent procedure that allows taking the band coupling into account, similarly to the perturbation theory in quantum mechanics. Unfortunately, a straightforward implementation of this approach is significantly complicated by the complexity of the semiconductor band structure. It poses not principal, but technical difficulties, the resolution of which requires developing a special approach employing the symmetry properties of the semiconductor lattice. These questions are beyond the scope of the present consideration (see, e.g., Chapter III in Bir and Pikus (1974)), therefore we limit ourselves to a semiqualitative analysis, which clearly illustrates the main ideas and can be readily adopted for a particular situation, for example for studying the effect of the static magnetic field and so on.

In order to eliminate the off-diagonal elements of the eigen-part of the Hamiltonian we perform a unitary Schrieffer-Wolff transformation $T = e^S$

$$b_{n,\mathbf{k}} = \sum_{m}\int d\mathrm{q}(n,\mathbf{k}|T|m,\mathbf{q})c_{m,\mathbf{q}}, \quad b^{\dagger}_{n,\mathbf{k}} = \sum_{m}\int d\mathbf{q}(m,\mathbf{q}|T^{\dagger}|n,\mathbf{k})c^{\dagger}_{m,\mathbf{q}} \tag{29.19}$$

such that the eigen-part of the transformed Hamiltonian $\tilde{\mathcal{H}}_{mat} = T^{\dagger}\mathcal{H}_{mat}T$ is *approximately* diagonal. In Bir and Pikus (1974) a general procedure is discussed, which allows developing the respective perturbation theory up to any order. Here, however, we restrict ourselves to the first relevant order, as it is usually done in the $\mathbf{k}\cdot\mathbf{p}$-approximation. In terms of the "generator" S the transformed Hamiltonian is written as

$$\tilde{\mathcal{H}}_{mat} = \mathcal{H}_{mat} + [\mathcal{H}_{mat}, S] + \frac{1}{2}[[\mathcal{H}_{mat}, S], S] + \cdots \tag{29.20}$$

where $[A,B] = AB - BA$ denotes the usual matrix commutator. We want to treat different contributions on the different basis and therefore we present the Hamiltonian as $\mathcal{H}_{mat} = \mathcal{H}_0 + \mathcal{H}_p + \mathcal{H}_A$, where $\mathcal{H}_0$ stands for the diagonal part of the first term in the r.h.s. of Equation 29.17, $\mathcal{H}_p$ is the $\mathbf{k}\cdot\mathbf{p}$ part and $\mathcal{H}_A$ is the term related to the external field. Using this presentation in Equation 29.20 we obtain

$$\begin{aligned}\tilde{\mathcal{H}}_{mat} = {} & \mathcal{H}_0 + \mathcal{H}_p + \mathcal{H}_A + [\mathcal{H}_0, S] + [\mathcal{H}_p, S] + [\mathcal{H}_A, S] \\ & + \frac{1}{2}[[\mathcal{H}_0, S], S] + \frac{1}{2}[[\mathcal{H}_p, S], S] + \frac{1}{2}[[\mathcal{H}_A, S], S] + \cdots \end{aligned} \tag{29.21}$$

Since the $\mathbf{k}\cdot\mathbf{p}$-coupling between the bands is small, the transformation diagonalizing the Hamiltonian is close to identical and, hence, the generator S is in some sense small $\sim\mathbf{k}\cdot\mathbf{p}$. Therefore, in the lowest order no contributions come from the terms not shown in Equation 29.21 and, moreover, we need to drop the term $[[\mathcal{H}_p, S], S]$.

We choose S in such a way that the off-diagonal part of $\mathcal{H}_p$ is canceled in the leading order of $\mathbf{k}\cdot\mathbf{p}$. Or, since the diagonal part of $\mathcal{H}_p$ is 0, we find such S that

$$\mathcal{H}_p + [\mathcal{H}_0, S] = 0. \tag{29.22}$$

Resolving this equation with respect to the matrix elements of S yields

$$(n,\mathbf{k}|S|n',\mathbf{k}') = -\frac{\hbar\mathbf{k}\cdot\mathbf{p}_{nn'}}{m\omega_{nn'}}\delta(\mathbf{k}-\mathbf{k}'), \tag{29.23}$$

where $\omega_{nn'} = \epsilon_n - \epsilon_{n'}$.

Using found S for the field related terms in Equation 29.21, we will have among the others the terms of the form $\mathbf{A}(\mathbf{q})\otimes\mathbf{q}$, where $\otimes$ denotes the tensor product, so that in a Cartesian basis $(\mathbf{k}\otimes\mathbf{q})_{ij} = k_i q_j$. Generally the tensor $\mathbf{A}(\mathbf{q})\otimes\mathbf{q}$ is not zero. Its trace vanishes because of the Coulomb gauge imposed on the vector potential, but there are no other general restrictions on its form. Such terms, however, can still be neglected because of the smoothness of the external field. Indeed, because of the slow spatial variation of $\mathbf{A}(\mathbf{x})$ the main contribution into $\mathbf{A}(\mathbf{q})$ arises from small $\mathbf{q}$, which leads to the essential reduction of the tensor $\mathbf{A}(\mathbf{q})\otimes\mathbf{q}$ comparing to the field terms contained in $\mathcal{H}_A$ noncommuted with S in Equation 29.21. Thus, we obtain the matrix elements of the transformed Hamiltonian

$$\begin{aligned}&\left(n,\mathbf{k}\,|\,\tilde{\mathcal{H}}_{mat}\,|\,n',\mathbf{k}'\right) \\ &\quad = \left(\epsilon_n + \frac{\hbar^2 k^2}{2m}\right)\delta_{nn'}\delta(\mathbf{k}-\mathbf{k}') \\ &\quad + \frac{\hbar^2}{2m^2}\sum_{n''}\mathbf{k}\cdot\mathbf{p}_{nn''}\mathbf{k}\cdot\mathbf{p}_{n''n'}\left(\frac{1}{\omega_{nn''}} + \frac{1}{\omega_{n'n''}}\right) \\ &\quad - \frac{e}{m}\int d\mathbf{q}\left[\mathbf{A}(\mathbf{q})\cdot(\mathbf{p}_{nn'} + \hbar\mathbf{k}'\delta_{nn'}) - \frac{e}{2}A_2(\mathbf{q})\delta_{nn'}\right]\delta(\mathbf{k}-\mathbf{k}'-\mathbf{q}) \\ &\quad - \frac{\hbar e}{m^2}\sum_{n''}\int d\mathbf{q}\,\delta(\mathbf{k}-\mathbf{k}'-\mathbf{q})\left[\frac{\mathbf{k}\cdot\mathbf{p}_{n''n'}\mathbf{A}(\mathbf{q})\cdot\mathbf{p}_{nn''}}{\omega_{n'n''}} + \frac{\mathbf{k}\cdot\mathbf{p}_{nn''}\mathbf{A}(\mathbf{q})\cdot\mathbf{p}_{n''n'}}{\omega_{nn''}}\right].\end{aligned} \tag{29.24}$$

It should be noted that because of the smoothness property discussed above the field terms are symmetric with respect to the change $\mathbf{k}\leftrightarrow\mathbf{k}'$.

The Hamiltonian with the matrix elements given by Equation 29.24 looks quite complex and barely simpler than the initial Hamiltonian. In order to discuss the physical meaning of different contributions into Equation 29.24, we introduce the spatial field operators $c_n(\mathbf{x})$ and $c_n^\dagger(\mathbf{x})$ according to

$$c_{n,\mathbf{k}} = \frac{1}{(2\pi)^{3/2}}\int d\mathbf{x}\,c_n(\mathbf{x})e^{-i\mathbf{k}\cdot\mathbf{x}} \tag{29.25}$$

and the tensors $\mathsf{M}_{nn'}$ of effective masses

$$\frac{1}{\mathsf{M}_{nn'}} = \frac{1}{m}1\delta_{nn'} + \frac{1}{m^2}\sum_{n''}\left[\frac{\mathbf{p}_{n''n'}\otimes\mathbf{p}_{nn''}}{\omega_{nn''}} + \frac{\mathbf{p}_{nn''}\otimes\mathbf{p}_{n''n'}}{\omega_{n'n''}}\right]. \tag{29.26}$$

Additionally we need to analyze the different types of interband, $n \neq n'$, and intraband, $n = n'$, couplings induced by the external field. It can be seen that the main term of interest from the perspective of the semiconductor nonlinear response is $\propto\mathbf{A}(\mathbf{q})\cdot\mathbf{p}_{nn'}$, while all others can be neglected. Indeed, the terms $\propto\mathbf{A}\otimes\mathbf{k}$ describing the interband coupling vanish in the limit $k\to 0$, and thus their contribution is negligible in the region of validity of the $\mathbf{k}\cdot\mathbf{p}$-approximation. In turn, the effect of the terms yielding the intraband coupling is small provided the gap Δ between the valence and the conduction bands is relatively wide. As will be shown below (see Equation 29.56), the strong interaction of light with the semiconductor occurs when the frequency of the electromagnetic field is tuned into the resonance with the gap, $\hbar\omega\approx\Delta$. Thus, the intraband coupling is $\propto\hbar(\Delta T)^{-1}$, where T is the duration of the external excitation field, and even at (reasonably) short time scales it can be neglected.

Taking into account these considerations the semiconductor Hamiltonian takes a compact form

$$\tilde{\mathcal{H}}_{mat} = \sum_{n,n'}\int d\mathbf{x}\,c_n^\dagger(\mathbf{x})\left[\epsilon_n\delta_{nn'} - \nabla\cdot\frac{\hbar^2}{2\mathsf{M}_{nn'}}\cdot\nabla - \mathbf{A}(\mathbf{x})\cdot\mathbf{d}_{nn'}\right]c_{n'}(\mathbf{x}), \tag{29.27}$$

where $\mathbf{d}_{nn'} = \mathbf{p}_{nn'}e/m$. This representation clarifies the physical meaning of the operators $c_n^\dagger(\mathbf{x})$. These are the field operators of quasiparticles corresponding to different bands and characterized by specific tensors of the effective mass. The logic of perturbative treatment of the band coupling implies neglecting the off-diagonal elements $\mathsf{M}_{nn'}$ for $n \neq n'$ as the respective correction are of higher order in $\mathbf{k} \cdot \mathbf{p}$ than those kept for derivation of Equation 29.27. This naturally leads to the definition of the effective masses M_{nn} of the respective bands. The explicit expression for the tensors of the effective mass given by $n = n'$ elements of Equation 29.26 gives the correct prediction of significant reduction of the masses compared to the electron mass in the empty space and even the negative sign of the mass of the upper valence bands, it also describes correctly the dependence on the width of the gap. However, it lacks more subtle but nevertheless important features, such as the mixing of heavy and light holes.

In order to discuss these features, we inspect closely the semiconductor band structure (see, e.g., Chapter 2 in Yu and Cardona (2004)). If the electrons were spinless, the lowest energy bands were the valence and conduction bands. In semiconductors with diamond- or zinc-blende structures (e.g., Ge, GaAs, InP) the valence band would be threefold degenerate. It turns out that under rotations which map the lattice into itself the states belonging to this band transform similarly to the states with the orbital momentum $\ell = 1$. Because of this, the band is called* $\Gamma_{4\upsilon}$. The states in the conduction band transform trivially, and in group-theoretical terms the band is called Γ_{1c}. In addition to this one needs to take into account the twofold degeneracy of each state due to the electron spin. This makes the conduction band twofold and the valence band sixfold degenerate. Directly it does not imply divergence of the terms in Equations 29.23 and 29.26 corresponding to the degenerate bands because the respective matrix elements of the momentum operator are zero. This, however, does imply certain sensitivity of these bands to the external perturbations. The interaction, which takes advantage of the degeneracy, is the spin–orbit interaction originating from the dependence of the electron energy on the orientation of the spin in spatially inhomogeneous potential (see, e.g., Chapter XII in Schiff (1949)). The spin–orbit interaction leads to partial lifting of the degeneracy forming the semiconductor band structure schematically shown in Figure 29.1.

Thus, in order to be able to describe the complex picture of the band coupling in semiconductors, we need to consider the tensors of the effective masses as phenomenological parameters and, in particular, to keep the interband tensors $\mathsf{M}_{n,n'}$ with $n \neq n'$. The great simplification comes from the consideration of restrictions imposed on the form of these tensors due to the lattice symmetry. Since the semiconductor lattice turns into itself under the action of the point symmetry group so does Hamiltonian (29.27). For a general set of tensors of the effective masses $\mathsf{M}_{n,n'}$ it might not be the case, thus the symmetry implies that these tensors must have a special structure. This line of arguments was implemented and clearly explained in the series of papers by Luttinger and Kohn (1955, Luttinger 1956).

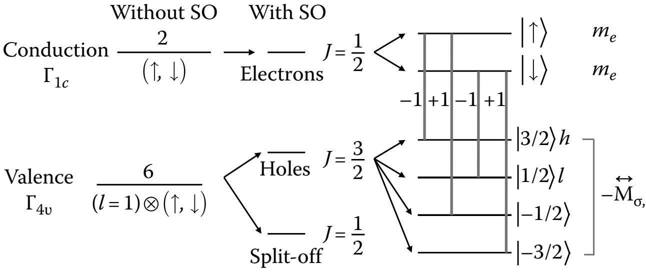

FIGURE 29.1 Effects of the spin–orbit interaction on the semiconductor band structure. The vertical lines indicate the subbands in the valence and conduction bands, between which the matrix elements of the momentum operator are not zero, together with the helicity of the respective transition. The convention for masses is shown.

Studying these restrictions shows that in the conduction band the tensors of the effective masses are decoupled from all bands and are trivial $\mathsf{M}_{s,s'} = 1 m_e \delta_{s,s'}$, where 1 is the unit tensor, and $\mathsf{M}_{s,\sigma} = \mathsf{M}_{\sigma,s} \equiv 0$. The tensors of the effective masses of the heavy and light holes are coupled $\mathsf{M}_{\sigma,\sigma'} \neq 0$ leading to the heavy–light hole hybridization. Additionally, the principal values of the tensors are not equal to each other leading to the anisotropy of the effective masses. It should be noted, however, that the anisotropy usually is not very strong and the mixing of the heavy and light holes is $\propto k^2$ and for small k the effect of hybridization is also small. This supports the popular approximation of spherical, decoupled heavy-hole and light-hole bands widely used for description of the semiconductor weak excitations. We also will use it for the qualitative discussions of some questions while keeping the main equations as general as possible.

So far we considered the case of a homogeneous semiconductor. In order to complete the derivation of the effective Hamiltonian we need to consider the spatial variation of the energies of the conduction and the valence bands due to the variation of composition and to include the terms responsible for the effect of the external static field and the Coulomb interaction between the electrons. To this end we take into consideration the last two terms in Hamiltonian (29.1) and substitute $\epsilon_n(\mathbf{x})$ for the variation of the bands edges. Generally, strong coordinate dependence of $\epsilon_n(\mathbf{x})$ would break the whole line of arguments, which allowed us to derive Equation 29.27 and to introduce the effective masses. It turns out, however, that as small as 10 elementary cells are sufficient for the approximation of effective mass to work. In particular, it can be safely applied for the description of the semiconductor quantum dots. Thus, the assumption of smoothness is fulfilled for composition-induced variations of the band edges (quantum heterostructures (Ivchenko 2005) or random interface fluctuations (Langbein et al. 2004)). It also usually holds for the external electric field (e.g., in the case of studying the quantum-confined Stark effect (Schmitt-Rink et al. 1989) on the optical semiconductor response). This assumption is employed in the standard manner. Taking into account that the main contribution into the dynamics comes from the small **k** part of Hamiltonian (29.24), one can show the validity of the important relation

---

* Strictly speaking, the correct formulation should sound like "the states transform according to irreducible representation Γ_4 of the point symmetry group of the lattice." Detailed and practical introduction into the group theory can be found in Yu and Cardona (2004) and Bir and Pikus (1974).

$$\int d\mathbf{x}\, f(\mathbf{x})\chi^{*}_{n,\mathbf{k}}(\mathbf{x})\chi_{n',\mathbf{k}'}(\mathbf{x}) = \delta_{n,n'}\int \frac{d\mathbf{x}}{(2\pi)^3} f(\mathbf{x})e^{-i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{x}} \qquad (29.28)$$

for sufficiently smooth $f(\mathbf{x})$. The key element is to observe that if the spatial Fourier spectrum $f(\mathbf{q})$ falls off fast enough, then $\delta(\mathbf{k}' - \mathbf{k} + \mathbf{q} = \mathbf{Q})$, with $\mathbf{Q}$ being the vector of the reciprocal lattice, reaches its singularity only when $\mathbf{Q} = 0$, or, in solid-state theory terms that the Umklapp processes can be neglected. Using relation (29.28) makes the transformation to the field operators $c_n(\mathbf{x})$, following Equations 29.16 and 29.19, trivial.

The Coulomb potential $V(\mathbf{x})$ is singular at the origin, therefore, generally its variation over an elementary cell is not small. However, the effect of the motion of electrons when they are extremely close to each other is not significant for the problem of the optical response. The strongest effect of the Coulomb interaction can be expected when it leads to the formation of bound states (excitons). The typical distance between the particles in such states significantly exceeds the lattice constants and the details of the behavior of the interaction potential at the origin are not important. Therefore, with good accuracy the Coulomb potential can be considered as slowly varying and one can apply relation (29.28).

Collecting these considerations, we can write down the full Hamiltonian in terms of the operators $c_n(\mathbf{x})$ and $c_n^{\dagger}(\mathbf{x})$

$$\begin{aligned}\mathcal{H}_{mat} = \sum_{n,m}\int d\mathbf{x}\, c_n^{\dagger}(\mathbf{x})\left[\epsilon_n(\mathbf{x})\delta_{nm} - \nabla\cdot\frac{\hbar^2}{2\mathsf{M}_{nm}}\cdot\nabla - \mathbf{A}(\mathbf{x})\cdot\mathbf{d}_{nm}\right]c_m(\mathbf{x})\\ +\frac{1}{2}\sum_{n,m}\int d\mathbf{x}_1\, d\mathbf{x}_2\, c_n^{\dagger}(\mathbf{x}_1)c_m^{\dagger}(\mathbf{x}_2)V(\mathbf{x}_1-\mathbf{x}_2)c_m(\mathbf{x}_2)c_n(\mathbf{x}_1).\end{aligned} \qquad (29.29)$$

29.3 Quantum Equations of Motion

The main object of interest while studying the nonlinear response is the time dependence of observables such as the macroscopic electromagnetic field $\mathbf{A}(\mathbf{x}) = \langle\hat{\mathbf{A}}(\mathbf{x})\rangle$. These are determined by the state of the whole electromagnetic field–matter system. For example, if the system is in the coherent regime, that is, if the system can be described by the vector of state $|\Phi(t)\rangle$, then

$$\mathbf{A}(\mathbf{x}) = \langle\Phi(t)|\hat{\mathbf{A}}(\mathbf{x})|\Phi(t)\rangle = \langle\Phi(0)|U^{\dagger}(t;t_0)\hat{\mathbf{A}}(\mathbf{x})U(\mathrm{t};\mathrm{t}_0)|\Phi(0)\rangle, \qquad (29.30)$$

where $|\Phi(0)\rangle$ is the initial state of the system and the time dependence of the system state $|\Phi(t)\rangle = U(t;t_0)|\Phi(0)\rangle$ is written in terms of the evolution operator

$$U(t;t_0) = \mathcal{T}_{+}\exp\left[-\frac{i}{\hbar}\int_{t_0}^{t} dt'\mathcal{H}(t')\right] \qquad (29.31)$$

with $\mathcal{T}_{+}$ being the standard time-ordering operator. Here we have taken into account that the Hamiltonian may explicitly depend on time because of the presence of the external driving electromagnetic field.

The same ideology works also if the initial state is a mixture (e.g., thermal equilibrium) and therefore it should be described by the density matrix $\hat{\rho}(0)$. If there are no decohering processes, then one has

$$\mathbf{A}(\mathbf{x}) = \mathbf{Tr}\left[\hat{\rho}(t)\hat{\mathbf{A}}(\mathbf{x})\right] = \mathrm{Tr}\left[\hat{\rho}(0)U^{\dagger}(t;t_0)\hat{\mathbf{A}}(\mathbf{x})U(t;t_0)\right]\hat{\rho}(0). \qquad (29.32)$$

Here we have explicitly spelled out the time dependence of the density matrix in the absence of the incoherent processes $\hat{\rho}(t) = U(t;t_0)\hat{\rho}(t)U^{\dagger}(t;t_0)$. The situation, however, becomes more complex when the decohering processes, for example, interaction with bath's phonons, is present. This problem can be treated rigorously using special methods (see, e.g., Chapter 3 in Mukamel (1995)), which are beyond the scope of the present consideration because they are somewhat excessive for the problem of finding the macroscopic electromagnetic field when a simple introduction of a phenomenological decay of the macroscopic polarization usually suffices.

The typical structure

$$\hat{O}(t) = U^{\dagger}(t;t_0)\hat{O}U(t;t_0) \qquad (29.33)$$

appearing in Equations 29.30 and 29.32 defines the *Heisenberg representation* of the field operator $\hat{O}$. Since this representation plays the same role whether the initial state is coherent or not, it is often more convenient to study its time dependence rather than directly the time dependence of the state of the system or its density matrix. It can be easily checked that the time evolution of $\hat{O}(t)$ is governed by the Heisenberg equations of motion.

$$i\hbar\frac{\partial\hat{O}}{\partial t} = \left[\hat{O},\mathcal{H}\right]. \qquad (29.34)$$

Since the explicit time dependence of the Hamiltonian does not change the derivation of the main equations of motion (all operators entering Equation 29.34 are taken at the same instant), we will omit the time argument in what follows.

29.3.1 Electromagnetic Field

We begin the general analysis of the semiconductor optical response from the derivation of the equations of motion of the electromagnetic field. The commutator in this equation is found using the commutation relation for the electromagnetic field in the Coulomb gauge (Cohen-Tannoudji et al. 1992)

$$\left[\hat{A}_j(\mathbf{x}_1),\ \epsilon_0\frac{\partial}{\partial t}\hat{A}_l(\mathbf{x}_2)\right] = \frac{i\hbar\delta_{jl}}{n^2(\mathbf{x}_1)}\delta^{\perp}(\mathbf{x}_1-\mathbf{x}_2), \qquad (29.35)$$

where

$n(\mathbf{x})$ is the refractive index, which is allowed to be spatially inhomogeneous

j and l denote the Cartesian coordinates

$\hat{\delta}^{\perp}(\mathbf{x})$ is the transverse δ-function (Cohen-Tannoudji et al. 1992)

As Helmholtz's theorem (Arfken and Weber 2005) states any vector field $\mathbf{V}(\mathbf{x})$ can be presented as the sum of irrotational and solenoidal parts, $\mathbf{U}(\mathbf{x}) = \mathbf{U}_l(\mathbf{x}) + \mathbf{U}_t(\mathbf{x})$ with $\nabla \times \mathbf{U}_l(\mathbf{x}) \equiv 0$ and $\nabla \cdot \mathbf{U}_t(\mathbf{x}) \equiv 0$. The convolution of the vector field with the transverse δ-function returns the value of the solenoidal component $\mathbf{V}_r(\mathbf{x})$ at the point of singularity

$$\int d\mathbf{x}'\, \hat{\delta}^{\perp}(\mathbf{x} - \mathbf{x}')\mathbf{U}(\mathbf{x}') = \mathbf{U}_t(\mathbf{x}). \tag{29.36}$$

The Hamiltonian of the electromagnetic field has the standard form

$$\mathcal{H}_F = \frac{\epsilon_0}{2}\int d\mathbf{x}\left[n^2(\mathbf{x})E^2(\mathbf{x},t) + c^2B^2(\mathbf{x},t)\right], \tag{29.37}$$

where the transverse electric and magnetic fields are defined in terms of the vector potential as $\mathbf{E}(\mathbf{x}) = -\partial\mathbf{A}(\mathbf{x})/\partial t$ and $\mathbf{B}(\mathbf{x}) = \nabla \times \mathbf{A}(\mathbf{x})$. Substituting into Equation 29.34 $\mathcal{H} = \mathcal{H}_F + \mathcal{H}_{mat}$, where $\mathcal{H}_{mat}$ is given by Equation 29.29 one obtains the *quantum* equations of motion for the electromagnetic field

$$\frac{n^2(\mathbf{x})}{c^2}\ddot{\hat{\mathbf{A}}} = \nabla^2\hat{\mathbf{A}} + \mu_0\sum_{n,m}\int d^3\mathbf{x}'\,\hat{\delta}^{\perp}(\mathbf{x} - \mathbf{x}')\mathbf{d}_{nm}c_n^{\dagger}(\mathbf{x}')c_m(\mathbf{x}'). \tag{29.38}$$

This equation has the standard form of the inhomogeneous wave equation for the vector potential with the integral in the r.h.s. representing the current source.

Writing Equation 29.38 we have taken into account that $\nabla \cdot \mathbf{A}$ vanishes in the Coulomb gauge used for derivation of Equation 29.27. Since, generally, the vector field $\mathbf{U}(\mathbf{x}) = \mu_0\sum_{n,m}\mathbf{d}_{nm}c_n^{\dagger}(\mathbf{x})c_m(\mathbf{x})$ is not necessarily solenoidal, the presence of the transverse δ-function in the r.h.s. of Equation 29.38 is important to ensure that at any instant $\hat{\mathbf{A}}(\mathbf{x})$ does not violate the Coulomb gauge. Clearly, if one replaces the transverse δ-function by the usual (local) δ-function it implies effective adding to the source the irrotational component of $\mathbf{U}(\mathbf{x})$. According to Equation 29.36, the convolution of $\mathbf{U}(\mathbf{x})$ with the transverse δ-function is $\int d\mathbf{x}'\,\tilde{\delta}^{\perp}(\mathbf{x} - \mathbf{x}')\mathbf{U}(\mathbf{x}') = \mathbf{U}(\mathbf{x}) - \mathbf{U}_l(\mathbf{x})$. Equation 29.38 is linear, therefore $\hat{\mathbf{A}}(\mathbf{x})$ can be presented as a superposition of two components, $\hat{\mathbf{A}}(\mathbf{x}) = \tilde{\hat{\mathbf{A}}}(\mathbf{x}) + \hat{\mathbf{A}}_l(\mathbf{x})$, which satisfy the wave equation of the same form as Equation 29.38, with $\mathbf{U}(\mathbf{x})$ and $-\mathbf{U}_l(\mathbf{x})$ serving as sources. Since $\mathbf{U}_l(\mathbf{x})$ is irrotational it can be presented as a gradient of a scalar function $\mathbf{U}_l(\mathbf{x}) = -\nabla\varphi(\mathbf{x})$. Furthermore, following the standard line of arguments as in electrostatics we obtain

$$\varphi(\mathbf{x}) = \frac{1}{4\pi}\int d\mathbf{x}'\,\frac{\nabla \cdot \mathbf{U}(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}. \tag{29.39}$$

Thus $\varphi(\mathbf{x})$ has the same form as the potential created by the charge spatially distributed with the density $\epsilon_0\nabla \cdot \mathbf{U}(\mathbf{x})$. If $\tilde{\hat{\mathbf{A}}}(\mathbf{x})$ satisfies the same initial conditions as $\hat{\mathbf{A}}(\mathbf{x})$ the component $\hat{\mathbf{A}}_l(\mathbf{x})$ is early shown to be irrotational and, hence, it does not contribute to the outgoing radiation field but leads to a modification of the electron–electron interaction. This modification identically vanishes in many important physical situations, such as the excitation by the normally incident wave at a frequency tuned to the response with the heavy-hole excitons. While a rigorous study of the effect of $\hat{\mathbf{A}}_l(\mathbf{x})$ on the semiconductor dynamics is yet to be conducted, it can be conjectured that for the problems which will be studied below the respective correction of the electron–electron interaction is rather small and can be neglected; in other words one can approximate $\hat{\mathbf{A}}(\mathbf{x}) \approx \tilde{\hat{\mathbf{A}}}(\mathbf{x})$. This significantly simplifies Equation 29.38 since we can use the conventional δ-function instead of $\hat{\delta}^{\perp}(\mathbf{x} - \mathbf{x}')$ and perform the integration over $\mathbf{x}'$, obtaining

$$\frac{n^2(\mathbf{x})}{c^2}\ddot{\hat{\mathbf{A}}}(\mathbf{x}) = \nabla^2\hat{\mathbf{A}}(\mathbf{x}) + \mu_0\sum_{n,m}\mathbf{d}_{nm}c_n^{\dagger}(\mathbf{x})c_m(\mathbf{x}). \tag{29.40}$$

This equation completely describes the electromagnetic field driven by the semiconductor excitations. The macroscopic field, as has been discussed above, is found by averaging $\hat{\mathbf{A}}(\mathbf{x})$ over the state of the combined semiconductor-field system. It should be noted, however, that Equation 29.40 can also be used in a more general context to find higher-order moments of the field, for example, single- or two-photon density matrix and so on.

29.3.2 Semiconductor Dynamics

The current, which enters the equation of motion of the electromagnetic field as a source, is determined by the operator of the interband polarization $c_n^{\dagger}(\mathbf{x}_1)c_m(\mathbf{x}_2)$ taken at the diagonal $\mathbf{x}_1 = \mathbf{x}_2$. Its time evolution is governed by the Heisenberg equation of motion (29.34). The electron operators obey the canonical anticommutation relation.

$$\left\{c_n^{\dagger}(\mathbf{x}_1), c_m(\mathbf{x}_2)\right\} = \hbar\delta_{n,m}\delta(\mathbf{x}_1 - \mathbf{x}_2), \tag{29.41}$$

where $\{A, B\} = AB + BA$. The commutator $\left[c_n^{\dagger}(\mathbf{x}_1)c_m(\mathbf{x}_2), \mathcal{H}\right]$ is expressed in terms of the anticommutators using the relation, which holds for any four operators

$$[AB, CD] = A\{B, C\}D - \{A, C\}BD + CA\{B, D\} - C\{A, D\}B. \tag{29.42}$$

Substituting $\mathcal{H} = \mathcal{H}_F + \mathcal{H}_{mat}$ we obtain

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}c_n^\dagger(\mathbf{x}_1)c_m(\mathbf{x}_2) = {} & \left[\epsilon_m(\mathbf{x}_2)-\epsilon_n(\mathbf{x}_1)\right]c_n^\dagger(\mathbf{x}_1)c_m(\mathbf{x}_2) \\ & +\sum_{j,k}\left[\delta_{km}\nabla_1\cdot\frac{\hbar^2}{2\mathsf{M}_{jn}}\cdot\nabla_1-\delta_{jn}\nabla_2\cdot\frac{\hbar^2}{2\mathsf{M}_{mk}}\cdot\nabla_2\right. \\ & \left.+\,\delta_{mk}\hat{\mathbf{A}}(\mathbf{x}_1)\cdot\mathbf{d}_{jn}-\delta_{jn}\hat{\mathbf{A}}(\mathbf{x}_2)\cdot\mathbf{d}_{mk}\right]c_j^\dagger(\mathbf{x}_1)c_k(\mathbf{x}_2) \\ & +c_n^\dagger(\mathbf{x}_1)\sum_j\int d\mathbf{x}'\left[V(\mathbf{x}_2-\mathbf{x}')-V(\mathbf{x}_1-\mathbf{x}')\right] \\ & \times c_j^\dagger(\mathbf{x}')c_j(\mathbf{x}')c_m(\mathbf{x}_2) \end{aligned} \tag{29.43}$$

where indices at ∇'s denote the number of the spatial variables they act on. The form of Equation 29.43 shows why it is not enough to consider directly the time dependence of the operator $c_n^\dagger(\mathbf{x}_1)c_m(\mathbf{x}_2)$ at the diagonal $\mathbf{x}_1 = \mathbf{x}_2$. Because of the presence of the spatial derivatives the rate of change of the operator at the diagonal is determined by the value the operator takes at the neighboring points requiring, thereby, studying the general case $\mathbf{x}_1 \neq \mathbf{x}_2$.

Equation 29.43 is presented in a general form unifying the conduction and the valence bands. This makes a constructive analysis difficult because the excitations belonging to different bands have distinctively different properties, for example, the mass of excitations in the conduction band (electrons) is positive while in the valence band (holes) it is negative. In order to account for these differences it is convenient to introduce special field operators for holes. It is done following the simple rule—creation of an electron in the valence band corresponds to annihilation of a hole and vice versa. As has been discussed above, different subbands in the valence and the conduction bands are enumerated by the projection of the angular momentum ($j = 1/2$ and $j = 3/2$ for the conduction and the valence bands, respectively). Thus, we introduce the hole operators according to $\upsilon_\sigma^\dagger(\mathbf{x}) \equiv c_{\upsilon,\sigma}(\mathbf{x})$ and $\upsilon_\sigma(\mathbf{x}) \equiv c_{\upsilon,\sigma}^\dagger(\mathbf{x})$. For the electrons in the conduction band we leave the same notation but will keep only the index corresponding to the spin of the electron so that $c_s^\dagger(\mathbf{x}) \equiv c_{c,s}^\dagger(\mathbf{x})$.

Rewriting Equation 29.43 in terms of the electron and hole operators is straightforward but is not free from the formal difficulties related to the fact that the holes exist on the background of all electrons filling the valence band. In a sense, in order to create a hole, one first needs to fill completely the valence band and then remove one electron. This naturally leads to divergences of the Coulomb energy. In order to demonstrate one kind of divergence let us consider the Coulomb term, which appears at the r.h.s. of Equation 29.43 for $n = \{c, s_1\}$, $m = \{c, s_s\}$ and $j = \{\upsilon, \sigma\}$

$$c_{s_1}^\dagger(\mathbf{x}_1)\int \mathrm{d}\mathbf{x}'\left[V(\mathbf{x}_2-\mathbf{x}')-V(\mathbf{x}_1-\mathbf{x}')\right]\upsilon_\sigma(\mathbf{x}')\upsilon_\sigma^\dagger(\mathbf{x}')c_{s_1}(\mathbf{x}_2). \tag{29.44}$$

If one tries to rewrite this expression in the normal form, that is, with all the creation operators collected on the left, one gets an additional term $\propto\{\upsilon_\sigma(\mathbf{x}'), \upsilon_\sigma^\dagger(\mathbf{x}')\} = \hbar\delta(\mathbf{x}'-\mathbf{x}')$, which is infinite if not meaningless. There are different ways to work around this divergence. One of them is to consider Equation 29.44 as a continuous limit of the respective model formulated in a discrete space, where the field operators taken at different cells of the space anticommute to Kronecker's δ so that $\left\{\upsilon_{\sigma,\mathbf{x}}, \upsilon_{\sigma,\mathbf{x}'}^\dagger\right\} = N\delta_{\mathbf{x},\mathbf{x}'}$ with some constant N. This gives for the term proportional to the anticommutator

$$c_{s_1,\mathbf{x}_1}^\dagger c_{s_1,\mathbf{x}_2} N\sum_{\mathbf{x}'}\left[V_{\mathbf{x}_2-\mathbf{x}'}-V_{\mathbf{x}_1-\mathbf{x}'}\right]. \tag{29.45}$$

Truncating the Coulomb potential $V_\mathbf{x}$ at infinity we see that this term vanishes regardless of the parameters of the truncation and the discretization. Thus, we can conclude that these divergencies do not contribute to the equations of motion.

A more different obstacle is the divergent band-gap renormalization, which enters the equations of motion in the form $\epsilon_\sigma \rightarrow \epsilon_\sigma + V(0)$. This divergence is of the same origin as the previous one but it cannot be discarded as easily. The reason is that we have introduced the band edges as the solutions of a single-electron problem, while the hole sees the band edges in the presence of all the electrons filling the valence band minus one. Therefore, the renormalization is a real physical effect. We incorporate it in the theory in a similar way as the elementary-charge renormalization is dealt with in quantum field theory. We will consider from now on that the hole energy offset $\epsilon_\sigma(\mathbf{x})$ in the equations of motion is the genuine one. This can be done because the actual band gap between the conduction and valence bands is directly observable (see, e.g., Equation 29.56).

Taking into account these considerations we obtain

$$\begin{aligned} & i\hbar\frac{\partial}{\partial t}\upsilon_\sigma^\dagger(\mathbf{x}_1)c_s^\dagger(\mathbf{x}_2) \\ & \quad = \upsilon_\sigma^\dagger(\mathbf{x}_1)c_s^\dagger(\mathbf{x}_2)\left[\epsilon_\sigma(\mathbf{x}_1)+\hat{U}_\Delta(\mathbf{x}_1)-\epsilon_s(\mathbf{x}_2)-\hat{U}_\Delta(\mathbf{x}_2)+V(\mathbf{x}_1-\mathbf{x}_2)\right] \\ & \qquad +\frac{\hbar^2}{2}\left(\frac{1}{m_e}\nabla_2^2+\sum_{\sigma'}\nabla_1\cdot\frac{1}{\mathsf{M}_{\sigma,\sigma'}}\cdot\nabla_1\right)\times\upsilon_{\sigma'}^\dagger(\mathbf{x}_1)c_s^\dagger(\mathbf{x}_2) \\ & \qquad -\hat{\mathbf{A}}(\mathbf{x}_1)\cdot\mathbf{d}_{\sigma,s}\delta(\mathbf{x}_1-\mathbf{x}_2)\;+\sum_{s'}\hat{\mathbf{A}}(\mathbf{x}_1)\cdot\mathbf{d}_{\sigma,s'}c_s^\dagger(\mathbf{x}_2)c_{s'}(\mathbf{x}_1) \\ & \qquad +\sum_{\sigma'}\hat{\mathbf{A}}(\mathbf{x}_2)\cdot\mathbf{d}_{\sigma',s}\upsilon_{\sigma'}^\dagger(\mathbf{x}_1)\upsilon_\sigma(\mathbf{x}_2), \\[2ex] & i\hbar\frac{\partial}{\partial t}c_{s_1}^\dagger(\mathbf{x}_1)c_{s_2}(\mathbf{x}_2) \\ & \quad = c_{s_1}^\dagger(\mathbf{x}_1)\left[\epsilon_{s_2}(\mathbf{x}_2)+\hat{U}_\Delta(\mathbf{x}_2)-\epsilon_{s_1}(\mathbf{x}_1)-\hat{U}_\Delta(\mathbf{x}_1)\right]c_{s_2}(\mathbf{x}_2) \\ & \qquad +\frac{\hbar^2}{2m_e}(\nabla_1^2-\nabla_2^2)c_{s_1}^\dagger(\mathbf{x}_1)c_{s_2}(\mathbf{x}_2) \\ & \qquad +\sum_{\sigma'}\left[\hat{\mathbf{A}}(\mathbf{x}_2)\cdot\mathbf{d}_{s_2,\sigma'}\upsilon_{\sigma'}^\dagger(\mathbf{x}_2)c_{s_1}^\dagger(\mathbf{x}_1)-\hat{\mathbf{A}}(\mathbf{x}_1)\cdot\mathbf{d}_{\sigma',s_1}c_{s_2}(\mathbf{x}_2)\upsilon_{\sigma'}(\mathbf{x}_1)\right], \\[2ex] & i\hbar\frac{\partial}{\partial t}\upsilon_{\sigma_2}^\dagger(\mathbf{x}_2)\upsilon_{\sigma_1}(\mathbf{x}_1) \\ & \quad = \upsilon_{\sigma_2}^\dagger(\mathbf{x}_2)\left[\epsilon_{\sigma_2}(\mathbf{x}_2)+\hat{U}_\Delta(\mathbf{x}_2)-\epsilon_{\sigma_1}(\mathbf{x}_1)-\hat{U}_\Delta(\mathbf{x}_1)\right]\upsilon_{\sigma_1}(\mathbf{x}_1) \\ & \qquad +\frac{\hbar^2}{2}\sum_{\sigma',\sigma''}\left(\nabla_1\cdot\frac{\delta_{\sigma'',\sigma_2}}{\mathsf{M}_{\sigma',\sigma_1}}\cdot\nabla_1-\nabla_2\cdot\frac{\delta_{\sigma',\sigma_1}}{\mathsf{M}_{\sigma_2,\sigma''}}\cdot\nabla_2\right)\upsilon_{\sigma''}^\dagger(\mathbf{x}_2)\upsilon_{\sigma'}(\mathbf{x}_1) \\ & \qquad +\sum_{s'}\left[\hat{\mathbf{A}}(\mathbf{x}_1)\cdot\mathbf{d}_{s',\sigma_1}\upsilon_{\sigma_2}^\dagger(\mathbf{x}_2)c_{s'}^\dagger(\mathbf{x}_1)-\hat{\mathbf{A}}(\mathbf{x}_2)\cdot\mathbf{d}_{\sigma_2,s'}c_{s'}(\mathbf{x}_2)\upsilon_{\sigma_1}(\mathbf{x}_1)\right], \end{aligned} \tag{29.46}$$

where

$$\hat{U}_\Delta(\mathbf{x}) = \int d\mathbf{x}' V(\mathbf{x}-\mathbf{x}')\left[\sum_{s'} c^\dagger_{s'}(\mathbf{x}')c_{s'}(\mathbf{x}') - \sum_{\sigma'} \upsilon^\dagger_{\sigma'}(\mathbf{x}')\upsilon_{\sigma'}(\mathbf{x}')\right]. \tag{29.47}$$

Equations 29.40 and 29.46 describe the coupled dynamics of the electromagnetic field and the semiconductor excitations in the endless variety of physical situations. They can be used for studying the modification of the radiative decay of the semiconductor excitations in photonic crystals (Sakoda 2005) or the speckle pattern caused by electron scattering on surface inhomogeneities of quantum wells and so on. In what follows we will be mostly interested in the effect of the many-body correlations on the nonlinear optical response and, therefore, we make the proper simplifications of the problem. We assume that the life time of a photon in the structure is not too long so that the polaritons, which are the coupled states of the photons and the semiconductor excitations, are not formed. In this case one can neglect the reabsorption of the photons emitted in the course of the radiative decay. This significantly simplifies the analysis of Equations 29.40 and 29.46 because now we can consider **A**(**x**) in Equation 29.46 as the external *classical* excitation field, solve Equation 29.46 for $c^\dagger_n(\mathbf{x}_1)c_m(\mathbf{x}_2)$, and then find the field produced by the semiconductor excitation using Equation 29.40.

29.3.3 Rotating Wave Approximation

The interaction of light with the semiconductor has a strong resonant character when the frequency of the excitation field, Ω, sweeps the vicinity of the gap, $\hbar\Omega \sim \Delta$. In order to see how the resonances show up in the equations of motion we consider Equations 29.46 in the linear approximation when we need only the equation with respect to the operator of the interband polarization

$$i\hbar\frac{\partial}{\partial t}\upsilon^\dagger_\sigma(\mathbf{x}_2)c^\dagger_s(\mathbf{x}_1) = -\hat{L}_{\sigma,s}\upsilon^\dagger_\sigma(\mathbf{x}_2)c^\dagger_s(\mathbf{x}_1) - \mathbf{A}(\mathbf{x}_1)\cdot\mathbf{d}_{\sigma,s}\delta(\mathbf{x}_1-\mathbf{x}_2), \tag{29.48}$$

where the operator $\hat{L}_{\sigma,s}$ is defined as

$$\hat{L}_{\sigma,s} = \epsilon_s(\mathbf{x}_1) - \epsilon_\sigma(\mathbf{x}_2) - V(\mathbf{x}_1-\mathbf{x}_2) - \frac{\hbar^2}{2}\left(\frac{1}{m_e}\nabla_1^2 + \nabla_2\cdot\frac{1}{\mathsf{M}_\sigma}\cdot\nabla^2\right). \tag{29.49}$$

We have also neglected the heavy-hole–light-hole mixing setting $\mathsf{M}_{\sigma',\sigma} = \delta_{\sigma',\sigma}\mathsf{M}_\sigma$. This does not make principle changes in the character of the following consideration, but greatly simplifies the notation.

Next we employ the spectral representation of the operator $\hat{L}_{\sigma,s}$

$$\hat{L}_{\sigma,s}f(\mathbf{x}_2,\mathbf{x}_1) = \sum_\mu E_\mu\phi^*_\mu(\mathbf{x}_2,\mathbf{x}_1)\int d\mathbf{x}_1'\,d\mathbf{x}_2'\,\phi_\mu(\mathbf{x}_2',\mathbf{x}_1')f(\mathbf{x}_2',\mathbf{x}_1'), \tag{29.50}$$

where the formal summation over μ implies summing over the discrete part and integrating over the continuous part of the spectrum of the operator $\hat{L}_{\sigma,s}$, and E_μ and $\phi_\mu(\mathbf{x}_2,\mathbf{x}_1)$ are, respectively, the eigenvalues and the eigenfunctions, $\hat{L}_{\sigma,s}\phi_\mu = E_\mu\phi_\mu$. The operator $\hat{L}_{\sigma,s}$ is (formally) self-adjoint and, hence, its eigenvalues are real and the eigenfunctions form a complete orthonormal set.

Using Equation 29.50 in Equation 29.48 and convoluting both sides of Equation 29.48 with ϕ^*_μ we obtain the ordinary differential equation

$$-i\hbar\frac{\partial}{\partial t}\hat{P}^\dagger_\mu(t) = E_\mu\hat{P}^\dagger_\mu(t) + \mathcal{A}_\mu(t), \tag{29.51}$$

where we have introduced

$$\hat{P}^\dagger_\mu(t) = \int d\mathbf{x}_1\,d\mathbf{x}_2\,\phi_\mu(\mathbf{x}_2,\mathbf{x}_1)\upsilon^\dagger_\sigma(\mathbf{x}_2)c^\dagger_s(\mathbf{x}_1) \tag{29.52}$$

and

$$\mathcal{A}_\mu(t) = \int d\mathbf{x}\,\phi_\mu(\mathbf{x},\mathbf{x})\mathbf{A}(\mathbf{x},t)\cdot\mathbf{d}_{\sigma,s}. \tag{29.53}$$

The solution of Equation 29.51 is written as the sum of the general solution of the homogeneous equation (with $\mathcal{A}_\mu(t) \equiv 0$) and the partial solution of inhomogeneous

$$\hat{P}^\dagger_\mu(t) = \hat{P}^\dagger_\mu(0)e^{iE_\mu t/\hbar} + \frac{i}{\hbar}\int_0^t dt'\,e^{iE_\mu(t-t')/\hbar}\mathcal{A}_\mu(t'). \tag{29.54}$$

The first term is determined by the initial value of the operators $\upsilon^\dagger_\sigma(\mathbf{x}_2)c^\dagger_s(\mathbf{x}_1)$ or, according to the definition of the Heisenberg representation (see Equation 29.33), by their bare values. If, for example, initially the semiconductor is in the vacuum state (empty conduction band and filled valence band) the average value of the first term vanishes (see Equation 29.30) and, hence, it does not contribute to the average interband polarization and, accordingly, to the macroscopic electromagnetic field, which, as follows from Equation 29.40, is driven by $\left\langle\upsilon^\dagger_\sigma(\mathbf{x})c^\dagger_s(\mathbf{x})\right\rangle$.

In order to discuss the effect of the external field, we first assume that the external field changes harmonically with time

$$\mathbf{A}(\mathbf{x},t) = \mathbf{A}(\mathbf{x},\Omega)e^{-i\Omega t} + \mathbf{A}^*(\mathbf{x},\Omega)e^{i\Omega t}, \tag{29.55}$$

which yields

$$\left\langle 0\middle|\hat{P}^{(\mu)}_{\sigma,s}(t)\middle|0\right\rangle = \mathcal{A}^{(\mu)}_{\sigma,s}(\Omega)\frac{e^{iE_\mu t/\hbar} - e^{-i\Omega t}}{E_\mu + \hbar\Omega} + \bar{\mathcal{A}}^{(\mu)}_{\sigma,s}(\Omega)\frac{e^{iE_\mu t/\hbar} - e^{i\Omega t}}{E_\mu - \hbar\Omega}, \tag{29.56}$$

where $\mathcal{A}^{(\mu)}_{\sigma,s}(\Omega)$ and $\bar{\mathcal{A}}^{(\mu)}_{\sigma,s}(\Omega)$ are found using in Equation 29.53 **A**(**x**, Ω) and **A***(**x**, Ω), respectively. If $\Omega \approx E_\mu/\hbar$, then at the time scale $(\Omega + E_\mu/\hbar)^{-1} \ll t \ll |\Omega - E_\mu/\hbar|^{-1}$ the second term in Equation 29.56 increases linearly with time reflecting the resonant semiconductor response and eventually prevails over the first term, so that one can neglect the nonresonant contribution. This constitutes the *rotating wave approximation* widely used in studying the semiconductor optical response. For estimates we note that the smallest value E_μ is determined by the width of the gap, $\Delta = \epsilon_s - \epsilon_\sigma$, minus binding energy of a possible bound state. The value of the binding energy cannot be expected to be significant compared to Δ and thus we can take $E_\mu \approx \Delta$. In GaAs $\Delta \approx 1.5\,\text{eV}$ yielding the characteristic time scale

$$\mathrm{T}_R = \hbar\pi/\Delta \approx 1\,\text{fs}. \tag{29.57}$$

Thus, if the excitation field is approximately tuned into the resonance with the semiconductor gap and the duration of the excitation pulse significantly exceeds 1 fs, one can use the resonant approximation and neglect the nonresonant contribution. The reverse, however, should also be mentioned. The experiments with pulses with duration of few femtoseconds are readily available now and one should be careful in applying the resonant approximation for the description of such experiments.

Next we consider a more general case wherein the external field changes harmonically with slowly varying amplitude, for example, when the excitation field is a pulse of finite duration much longer than T_R, so that the temporal Fourier transform is localized in a vicinity of Δ with both detuning $|\Omega - \Delta/\hbar|$ and the width of the distribution much smaller than Δ. Obviously, in this case we can again separate the resonant and nonresonant contributions. Indeed, this is true for each major harmonic and, hence, for the whole excitation pulse.

This motivates taking advantage of the resonance condition on the level of the equations of motion. This can be easily done. Let the external field be of the form

$$\mathbf{A}(\mathbf{x},t) = \mathbf{A}_\Omega(\mathbf{x},t)e^{-i\Omega t} + \text{c.c.}, \tag{29.58}$$

where $\mathbf{A}_\Omega(\mathbf{x}, t)$ is a function slowly changing with time. Following Equation 29.54, we use in Equation 29.46 the ansatz

$$\upsilon_\sigma(\mathbf{x}) \to e^{-i\Omega t/2}\upsilon_\sigma(\mathbf{x}), \quad c_s(\mathbf{x}) \to e^{-i\Omega t/2}c_s(\mathbf{x}). \tag{29.59}$$

After canceling the factor $e^{i\Omega t}$ in the equation with respect to the operator of the interband polarization there will be two type of terms explicitly depending on time in the r.h.s. of Equation 29.46. The terms of the first type oscillate much faster than any typical time scale for the operators defined in Equation 29.59 and their contribution is small on the time scale $t > T_R$. Neglecting such terms corresponds to the rotating wave approximation and defining the operators according to Equation 29.59 is called the transition to the *rotating frame*.

We do not provide the form of the semiconductor equations of motion in the rotating frame but we would like to note that the width of the gap between the conduction and the valence bands in the equation for the operator of the interband polarization is renormalized by the frequency of the external field $\epsilon_s - \epsilon_\sigma \to \epsilon_s - \epsilon_\sigma - \hbar\Omega$. These equations can be derived as the Heisenberg equations of motion from the Hamiltonian $\mathcal{H} = \mathcal{H}_{e-h} + \mathcal{H}_{exc}$, where $\mathcal{H}_{e-h}$ is the electron–hole Hamiltonian written in the rotating frame

$$\begin{aligned}\mathcal{H}_{e-h} = &\sum_s \int d\mathbf{x}\, c_s^\dagger(\mathbf{x})\left[\epsilon_s(\mathbf{x}) - \hbar\Omega/2 - \frac{\hbar^2}{2m_e}\nabla^2\right]c_s(\mathbf{x}) \\ &- \sum_{\sigma,\sigma'} \int d\mathbf{x}\, \upsilon_\sigma^\dagger(\mathbf{x})\left[\delta_{\sigma,\sigma'}\epsilon_\sigma(\mathbf{x}) - \delta_{\sigma,\sigma'}\hbar\Omega/2 - \nabla\cdot\frac{\hbar^2}{2\mathrm{M}_{\sigma,\sigma'}}\cdot\nabla\right]\upsilon_{\sigma'}(\mathbf{x}) \\ &+ \frac{1}{2}\sum_{s,s'} \int d\mathbf{x}_1\, d\mathbf{x}_2\, c_s^\dagger(\mathbf{x}_1)c_{s'}^\dagger(\mathbf{x}_2)V(\mathbf{x}_1 - \mathbf{x}_2)c_{s'}(\mathbf{x}_2)c_s(\mathbf{x}_1) \\ &+ \frac{1}{2}\sum_{\sigma,\sigma'} \int d\mathbf{x}_1 d\mathbf{x}_2 \upsilon_\sigma^\dagger(\mathbf{x}_1)\upsilon_{\sigma'}^\dagger(\mathbf{x}_2)V(\mathbf{x}_1 - \mathbf{x}_2)\upsilon_{\sigma'}(\mathbf{x}_2)\upsilon_\sigma(\mathbf{x}_1) \\ &- \sum_{s,\sigma} \int d\mathbf{x}_1 d\mathbf{x}_2 \upsilon_\sigma^\dagger(\mathbf{x}_1)c_s^\dagger(\mathbf{x}_2)V(\mathbf{x}_1 - \mathbf{x}_2)c_s(\mathbf{x}_2)\upsilon_\sigma(\mathbf{x}_1)\end{aligned} \tag{29.60}$$

and $\mathcal{H}_{exc}$ is the Hamiltonian of interaction with the electromagnetic field in the rotating approximation

$$\mathcal{H}_{exc} = \sum_{s,\sigma} \int d\mathbf{x}[\mathbf{A}_\Omega(\mathbf{x},t)\cdot\mathbf{d}_{s,\sigma}\upsilon_\sigma^\dagger(\mathbf{x})c_s^\dagger(\mathbf{x}) + \mathbf{A}_\Omega^\star(\mathbf{x},t)\cdot\mathbf{d}_{\sigma,s}c_s(\mathbf{x})\upsilon_\sigma(\mathbf{x})], \tag{29.61}$$

where $\mathbf{A}_\Omega(\mathrm{x}, t)$ is the amplitude slowly changing with time. It is worth noting that Hamiltonian (29.60) reflects that holes and electrons have opposite charges—the Coulomb interaction between electrons and between holes increases the energy and, hence, leads to repulsion while the Coulomb interaction between the electrons and holes decreases the energy and corresponds to attraction.

While we derived Equation 29.61, taking into account only external classical excitation field, it can be generalized to describe the interaction of the semiconductor with the quantized electromagnetic field in the rotating wave approximation. For this one should represent $\hat{\mathbf{A}}(\mathbf{x})$ in terms of the photon creation and annihilation operators and use in Equation 29.61 instead of $\mathbf{A}_\Omega(\mathbf{x}, t)$ the part containing only the annihilation operators and instead of $\mathbf{A}_\Omega^\star(\mathbf{x}, t)$ the part with the photon creation operators. This illustrates an evident but important feature of the light–semiconductor interaction. The process of absorption of a photon is accompanied with the creation of the electron–hole pair and the emission results in destruction of a pair, so that the *operator* of the difference between the total numbers of the electrons and holes is a constant of motion

$$\int d\mathbf{x}\left[\sum_s c_s^\dagger(\mathbf{x})c_s(\mathbf{x}) - \sum_s \upsilon_\sigma^\dagger(\mathbf{x})\upsilon_\sigma(\mathbf{x}), \mathcal{H}\right] = 0. \tag{29.62}$$

29.4 Perturbation Theory: Four-Wave Mixing Response

The best understood way of analyzing the semiconductor nonlinear response is the perturbation theory with respect to the external excitation field because it allows for a general model independent formulation. For example, in Ostreich et al. (1998) the powerful formalism of Hubbard operators is used, which, as has been demonstrated in Ostreich (2001), allows a relatively straightforward extension to higher orders. Here we choose a more traditional approach developed in the standard $T = 0$ quantum field theory. The derivation of the main equations, however, significantly benefits from the idea of describing the semiconductor excitations in terms of genuine eigenstates borrowed from Ostreich et al. (1998).

We limit our consideration to the case when there is no external static field, i.e., $U_{\text{ext}}(\mathbf{x}) \equiv 0$, and the spatial variations of the band edges correspond to a single quantum well grown along z direction, $\epsilon_n(\mathbf{x}) = \epsilon_n + \Delta\epsilon_n(z)$. As has been discussed above the states reached in the course of the resonance excitation of the semiconductor are characterized by the definite numbers of the electron–hole pairs. The simplest states contain one such pair and can be created by acting on vacuum with the operator

$$\hat{B}^\dagger_\mu = \sum_{\sigma_\mu, s_\mu} \int d\mathbf{x}_1 d\mathbf{x}_2 \phi_\mu(\mathbf{x}_1, \mathbf{x}_2) \upsilon^\dagger_{\sigma_\mu}(\mathbf{x}_1) c^\dagger_{s_\mu}(\mathbf{x}_2), \qquad (29.63)$$

where the index μ enumerates different states of the single electron–hole pair. The function $\phi_\mu(\mathbf{x}_1, \mathbf{x}_2)$ is chosen in such a way that the state $|\mu\rangle = \hat{B}^\dagger_\mu |0\rangle$ is the eigenstate of the semiconductor Hamiltonian without the excitation field

$$\mathcal{H}_{e-h} |\mu\rangle = E_\mu |\mu\rangle. \qquad (29.64)$$

Substituting Equations 29.60 and 29.63 into this equation, we find that $\phi_\mu(\mathbf{x}_1, \mathbf{x}_2)$ satisfies the differential equation, which has the familiar structure (compare with Equation 29.48)

$$E_\mu \phi_\mu(\mathbf{x}_1, \mathbf{x}_2) = \left[\epsilon_{s_\mu}(\mathbf{x}_2) - \epsilon_{\sigma_\mu}(\mathbf{x}_1) - \hbar\Omega \right.$$
$$\left. - \frac{\hbar}{2m_c} \nabla_2^2 - \nabla_1 \cdot \frac{\hbar^2}{2M_{\sigma', \sigma_\mu}} \cdot \nabla_1 - V(\mathbf{x}_2 - \mathbf{x}_1) \right] \phi_\mu(\mathbf{x}_1, \mathbf{x}_2). \qquad (29.65)$$

This equation has the meaning of the stationary Schrödinger equation for two particles with opposite charges (the hole at point $\mathbf{x}_1$ and the electron at the point $\mathbf{x}_2$) moving in potentials $\epsilon_{\sigma_\mu}(\mathbf{x}_1)$ and $\epsilon_{s_\mu}(\mathbf{x}_2)$. Thus, $\phi_\mu(\mathbf{x}_1, \mathbf{x}_2)$ has the physical meaning of the electron–hole wave function corresponding to the particular stationary state. For a general form of potentials $\epsilon_{\sigma_\mu}(\mathbf{x}_1)$ and $\epsilon_{s_\mu}(\mathbf{x}_2)$. Equation 29.65 cannot be solved exactly and different approximate methods are developed for investigating the specific features of the solutions. We discuss the general properties of solutions of Equation 29.65 adopting the δ-functional approximation for the quantum well. In the framework of this approximation, the support of electron–hole wave function, that is, the region where the function is not identically zero along z direction is assumed to be limited by the position of the quantum well. Additionally, we neglect the off-diagonal interband elements and the anisotropy of the tensor of the holes effective mass, so we use $\mathsf{M}_{\sigma', \sigma_\mu} = 1\delta_{\sigma', \sigma_\mu} \mathsf{M}_{\sigma_\mu}$, where 1 is the unit tensor. This corresponds to neglecting the heavy-hole–light-hole mixing, which is valid if the hole momentum in plane of the quantum well is not too high (Zhu and Huang 1987). Under these approximations Equation 29.65 turns into the well-studied (and yet nontrivial (Parfitt and Portnoi 2002)) problem of the two-dimensional hydrogen atom.

The spectrum consists of the discrete and the continuous part. The discrete part corresponds to the bound electron–hole states, which are called *excitons*. Often this term is reserved for the bound states, however, often it is convenient to treat bound and unbound states on the equal ground and, therefore, the term "exciton" is also used as the short synonym of "electron–hole pair" regardless whether they are bound or not. From the perspective of Equation 29.64 operators $B^\dagger_\mu$ with the respective kernels $\phi_\mu(\mathbf{x}_1, \mathbf{x}_2)$ acting on vacuum create the exciton and thus $B^\dagger_\mu$ have the meaning of the exciton creation operators.

Because of the translational invariance of the r.h.s. of Equation 29.65 in the plane of the quantum well, the wave function admits the presentation

$$\phi_\mu(\mathbf{x}_1, \mathbf{x}_2) = e^{i\mathbf{K}_\mu \cdot \mathbf{R}} \tilde{\phi}_\mu(\mathbf{x}_1 - \mathbf{x}_2), \qquad (29.66)$$

where

$\mathbf{K}_\mu$ is the momentum

$\mathbf{R} = (M_{\sigma_\mu} \mathbf{x}_1 + m_e \mathbf{x}_2)/(M_{\sigma_\mu} + m_e)$ is the position of the center of mass of the pair

Similarly to the three-dimensional case, the eigenstates of the two-dimensional hydrogen atom are enumerated by the main quantum number n and the angular momentum ℓ. Thus, in this case, the index μ in Equation 29.63 combines all the quantum numbers characterizing the specific exciton state $\mu = \{\mathbf{K}_\mu, s_\mu, \sigma_\mu, n_\mu, \ell_\mu\}$. The important property of the eigenstates of the two-dimensional hydrogen atom is that the wave function corresponding to the states with $\ell \neq 0$ vanishes at $\mathbf{x}_1 = \mathbf{x}_2$. Thus, as follows from Equation 29.61, these states do not contribute to the coupling with the electromagnetic field. Among the states with $\ell = 0$ the most important one is the state with the lowest energy, whose wave function has the form

$$\tilde{\phi}_\mu(\mathbf{x}) = \sqrt{\frac{2}{\pi r_\mu^2}} e^{-|x|/r_\mu}, \qquad (29.67)$$

where the Bohr radius of the two-dimensional exciton is given by $r_\mu = \epsilon_b / 2m_{\sigma_\mu} e^2$ (notice that it is two times smaller than in three dimensions) with m_{σ_μ} being the electron–hole-reduced mass, $1/m_{\sigma_\mu} = 1/m_e + 1/M_{\sigma_\mu}$.

Using solutions of Equation 29.65 we can incorporate the single-pair states into the theory. The full set of solutions is complete

$$\sum_\mu \phi_\mu^*(\mathbf{x}_1,\mathbf{x}_2)\phi_\mu(\mathbf{x}_1',\mathbf{x}_2') = \delta(\mathbf{x}_1 - \mathbf{x}_1')\delta(\mathbf{x}_2 - \mathbf{x}_2'), \tag{29.68}$$

where summation over μ implies summation over the discrete quantum numbers (e.g., ℓ and n for hydrogen-like bound excitons) and integration over continuous ones (such as the center of mass momentum $\mathbf{K}_\mu$). Relation (29.68) allows us to express any electron–hole operator in terms of the exciton operators, for example

$$\upsilon_\sigma^\dagger(\mathbf{x})c_s^\dagger(\mathbf{y}) = \sum_{\mu | s_\mu = s, \sigma_\mu = \sigma} \phi_\mu^*(\mathbf{x},\mathbf{y})\hat{B}_\mu^\dagger \tag{29.69}$$

and represent in terms of the exciton operators the Hamiltonian of the interaction of the semiconductor, Equation 29.61, with the electromagnetic field

$$\mathcal{H}_{exc} = \sum_\mu [\mathcal{A}_\mu^*(t)\hat{B}_\mu + \mathcal{A}_\mu(t)\hat{B}_\mu^\dagger], \tag{29.70}$$

where

$$\mathcal{A}_\mu(t) = \int d\mathbf{x}\, \mathbf{A}_\Omega(\mathbf{x},t) \cdot \mathbf{d}_{s_\mu,\sigma_\mu} \phi_\mu^*(\mathbf{x},\mathbf{x}). \tag{29.71}$$

The similarity between Equations 29.63 and 29.71 and Equations 29.52 and 29.53 is not accidental. If we act by the first equation in Equations 29.46 on vacuum $|0\rangle$ in the case of absent external field, we obtain the nonstationary Schrödinger equation for a single-pair state, so the operator $\hat{L}_{\sigma,s}$ defined in Equation 29.49 has the meaning of the single-pair Hamiltonian. The spectral representation of $\hat{L}_{\sigma,s}$ allows projecting on the eigenstates of the single-pair Hamiltonian, which are analogous to the projections on the genuine single-particle states defined by Equations 29.63 and 29.71. This demonstrates that the procedure shown above can be generalized to go beyond the adopted approximations.

The interaction Hamiltonian $\mathcal{H}_{exc}$ is considered as a perturbation and the theory is then built following the standard lines. First we account for the nonperturbed dynamics exactly introducing

$$\left|\tilde{\Phi}(t)\right\rangle = e^{i\mathcal{H}_{e-h}t/\hbar}\,|\,\Phi(t)\rangle, \tag{29.72}$$

so that if there is no perturbation, then $|\tilde{\Phi}(t)\rangle$ does not change with time. Using the fact that the system state satisfies the equation $i\hbar\partial\,|\,\Phi(t)\rangle\,/\,\partial t = \mathcal{H}|\Phi\rangle$, we find that the time evolution is governed by

$$i\frac{\partial}{\partial t}\left|\tilde{\Phi}(t)\right\rangle = \tilde{\mathcal{H}}_{exc}(t)\left|\tilde{\Phi}(t)\right\rangle, \tag{29.73}$$

where

$$\tilde{\mathcal{H}}_{exc}(t) = e^{i\mathcal{H}_{e-h}t/\hbar}\mathcal{H}_{exc}(t)e^{-i\mathcal{H}_{e-h}t/\hbar}. \tag{29.74}$$

The solution of Equation 29.73 is written in terms of the S-matrix

$$\left|\tilde{\Phi}(t)\right\rangle = S(t)\left|\tilde{\Phi}(0)\right\rangle, \tag{29.75}$$

where

$$S(t) = 1 + \sum_{n=1}^{\infty}\left(-\frac{i}{\hbar}\right)^n \int\limits_{0 \le t_1 \le \dots} \dots \int\limits_{\le t_n \le t} dt_1 \dots dt_n \tilde{\mathcal{H}}_{exc}(t_1)\dots\tilde{\mathcal{H}}_{exc}(t_n). \tag{29.76}$$

Often this expansion is presented in the compact form $S(t) = \mathcal{T}_+ \exp\left\{-i\hbar^{-1}\int_0^t dt'\tilde{\mathcal{H}}_{exc}(t')\right\}$ similarly to Equation 29.31; we, however, will need in what follows the explicit form of expansion (29.76).

Next, we express the observables

$$O(t) = \left\langle\hat{O}\right\rangle = \left\langle\Phi(t)\,|\hat{O}|\,\Phi(t)\right\rangle = \left\langle\tilde{\Phi}(t)\,|\,e^{i\mathcal{H}_{e-h}t/\hbar}\hat{O}\,e^{-i\mathcal{H}_{e-h}t/\hbar}\,|\,\tilde{\Phi}(t)\right\rangle \tag{29.77}$$

and notice that in this equation and in Equation 29.74 appears the same structure

$$\hat{O}(t) = e^{i\mathcal{H}_{e-h}t/\hbar}\hat{O}e^{-i\mathcal{H}_{e-h}t/\hbar}, \tag{29.78}$$

which is called the *interaction representation* of the operator $\hat{O}$. Taking into account Equation 29.75 we obtain

$$O(t) = \left\langle S^\dagger(t)\tilde{O}(t)S(t)\right\rangle_0, \tag{29.79}$$

where $\langle\dots\rangle_0 \equiv \langle\Phi(0)|\dots|\Phi(0)\rangle$. Finally, the perturbation series is obtained substituting expansion (29.76) into Equation 29.79. The order of the perturbation is determined by the highest order of $\tilde{\mathcal{H}}(t)$ kept in the resulting expression.

We illustrate the perturbation theory in action finding the macroscopic exciton polarization, $P_\mu = \left\langle\hat{B}_\mu\right\rangle$ or

$$P_\mu = \left\langle S^\dagger(t)\tilde{B}_\mu(t)S(t)\right\rangle_0. \tag{29.80}$$

Let the initial state of the system be vacuum $|\Phi(0)\rangle = |0\rangle$. Below we will consider only such initial state, therefore, in order to shorten notations we will omit the index 0 and write averages over vacuum simply $\langle\ldots\rangle$. Since the semiconductor Hamiltonian $\mathcal{H}_{e-h}$ preserves both the number of electrons in the conduction band and the number of holes in the valence band, the exciton operator in the interaction representation acts in a similar way as in the Schrödinger or Heisenberg representations, namely, $\tilde{B}_\mu^\dagger(t)$ creates and $\tilde{B}_\mu(t)$ destroys an electron–hole pair, respectively. Hence, the terms in the expansion of the r.h.s. of Equation 29.80, which give nonzero contribution, must have balanced numbers of the exciton creation and annihilation operators. In particular, one can easily check that there are no contributions of even orders (zeroth, second, and so on).

There is only one term in the first order

$$P_\mu^{(1)}(t) = -\frac{i}{\hbar}\sum_\nu \int_0^t dt_1 \left\langle \tilde{B}_\mu(t)\tilde{B}_\nu^\dagger(t_1)\right\rangle \mathcal{A}_\nu(t_1)$$

$$= -\frac{i}{\hbar}\int_0^t dt_1 e^{-iE_\mu(t-t_1)/\hbar}\mathcal{A}_\mu(t_1), \tag{29.81}$$

where in the second equality we have taken into account Equation 29.64 and the fact that $\hat{B}_\mu\hat{B}_\nu^\dagger\left|0\right\rangle = \delta_{\mu,\nu}\left|0\right\rangle$, which follows from the commutation relation for the exciton operators

$$[\hat{B}_\mu \hat{B}_\nu^\dagger] = \hbar(\delta_{\mu,\nu} - \hat{C}_{\mu,\nu}), \tag{29.82}$$

where

$$\hat{C}_{\mu,\nu} = \int d\mathbf{x}'\, d\mathbf{y}'\, d\mathbf{x}\, d\mathbf{y}\, \phi_\nu(\mathbf{x}', \mathbf{y}')\phi_\mu^\star(\mathbf{x}, \mathbf{y})[\upsilon_{\sigma_\nu}^\dagger(\mathbf{x}')\upsilon_{\sigma_\mu}(\mathbf{x})\delta(\mathbf{y}-\mathbf{y}')\delta_{s_\nu,s_\mu}$$

$$+ c_{s_\nu}^\dagger(\mathbf{y}')c_{s_\mu}(\mathbf{y})\delta(\mathbf{x}-\mathbf{x}')\delta_{\sigma_\nu,\sigma_\mu}]. \tag{29.83}$$

These operators describe the deviation of excitons from bosons. It should be noted, however, that the deviation depends on the number of the electrons and holes. Thus, in the limit of weak excitations, the excitons can be approximately considered as bosons. However, as the population of the valence and conduction bands increases, one has to take into account the full form of the commutation relation.

Differentiating Equation 29.81 with respect to time we find

$$i\hbar\frac{\partial}{\partial t}P_\mu^{(1)}(t) = E_\mu P_\mu^{(1)}(t) + \mathcal{A}_\mu(t). \tag{29.84}$$

Comparing this with Equation 29.51 we can see that Equation 29.84 is nothing else but the exciton polarization of the linear response. This shows that the perturbation theory built with the help of the interaction representation correctly accounts for the nonperturbed dynamics of the semiconductor. With this regard it should be noted that this presents a convenient opportunity to account for the incoherent processes, which take place in real systems. As has been noted before a rigorous introduction of the decoherence requires developing a complex technique. For many purposes, however, it suffices to introduce the imaginary part of the single-exciton energies, $E_\mu \to E_\mu - i\hbar\Gamma_\mu$.

Before we move further, we would like to discuss some basic features of the amplitudes $\mathcal{A}_\mu(t)$. Substituting Equation 29.66 within Equation 29.71, we obtain

$$\mathcal{A}_\mu(t) = \tilde{\phi}_\mu^\star(0)\int d\mathbf{x}\, \mathrm{e}^{-i\mathbf{K}_\mu\cdot\mathbf{x}}\mathbf{A}_\Omega(\mathbf{x},t)\cdot\mathbf{d}_{s_\mu,\sigma_\mu}, \tag{29.85}$$

where the integration is taken over the plane of the quantum well. Thus, if the excitation field has the form of a plain wave $\mathbf{A}_\Omega(\mathbf{x}) \propto \exp(i\mathbf{K}\cdot\mathbf{x})$, then $\mathcal{A}_\mu \propto \delta(\mathbf{K}_\parallel - \mathbf{K}_\mu)$, that is, the plane wave only excites the exciton states with the same center of mass momentum as the projection of $\mathbf{K}$ on the plane of the quantum well, $\mathbf{K}_\parallel$. Moreover, as follows from Equation 29.40 the same in-plane momentum will be passed to the outgoing electromagnetic field. Such conservation of the in-plane momentum is not a specific feature of the linear response. If the excitation field is a plane wave, the translations $\Delta\mathbf{x}$ in the plane of the quantum well lead to the amplitudes $\mathbf{A}_\Omega(\mathbf{x})$ acquiring a phase factor, $\mathbf{A}_\Omega(\mathbf{x} + \Delta\mathbf{x}) = \exp(i\mathbf{K}_\parallel \cdot \Delta\mathbf{x})\mathbf{A}_\Omega(\mathbf{x})$; in other words, the excitation field possesses the translational symmetry. If this is the case, then, as can be easily checked, the full system of the equations of motion, Equations 29.46, is also translationally invariant (more precisely it has the same translational symmetry as the external field) with the following transformation rules for the solutions

$$\begin{aligned}
\upsilon_\sigma^\dagger(\mathbf{x}_2 + \Delta\mathbf{x})c_s^\dagger(\mathbf{x}_1 + \Delta\mathbf{x}) &= \upsilon_\sigma^\dagger(\mathbf{x}_2)c_s^\dagger(\mathbf{x}_1)e^{i\mathbf{k}_\parallel\cdot\Delta\mathbf{x}},\\
c_{s_1}^\dagger(\mathbf{x}_1 + \Delta\mathbf{x})c_{s_2}(\mathbf{x}_2 + \Delta\mathbf{x}) &= c_{s_1}^\dagger(\mathbf{x}_1)c_{s_2}(\mathbf{x}_2),\\
\upsilon_{\sigma_2}^\dagger(\mathbf{x}_2 + \Delta\mathbf{x})\upsilon_{\sigma_1}(\mathbf{x}_1 + \Delta\mathbf{x}) &= \upsilon_{\sigma_2}^\dagger(\mathbf{x}_2)\upsilon_{\sigma_1}(\mathbf{x}_1).
\end{aligned} \tag{29.86}$$

This also results in the transfer of the incoming in-plane momentum to the outgoing wave.

The picture becomes more complex if the external field is a superposition of two-plane waves $\mathbf{A}_\Omega(\mathbf{x}) = \mathbf{A}_1\exp(i\mathbf{K}^{(1)}\cdot\mathbf{x}) + \mathbf{A}_2\exp(i\mathbf{K}^{(2)}\cdot\mathbf{x})$ with noncollinear in-plane components of the momenta, $\mathbf{K}_\parallel^{(1)} \neq a\mathbf{K}_\parallel^{(2)}$ for any number a. In this case, the external field is invariant with respect to infinitesimal translations in the direction perpendicular to $\Delta\mathbf{K} = \mathbf{K}_\parallel^{(2)} - a\mathbf{K}_\parallel^{(1)}$, while in the direction along $\Delta\mathbf{K}$ it is invariant only with respect to *finite* shifts $\Delta R = 2\pi/\Delta K$. For the polarization of linear response it brings nothing new, as one could expect only the exciton polarizations with the center of mass momenta $\mathbf{K}_\parallel^{(1)}$ and $\mathbf{K}_\parallel^{(2)}$ are produced. However, this is not necessarily the case for the nonlinear system of equations. Generally, it is invariant only with respect to finite translations. As we know from the basic theory of crystals whenever one has such symmetry, the momentum conserves only up to the vectors of the reciprocal lattice, which in our case are $m\Delta\mathbf{K}$ with integer m. Thus, in addition to the initial momenta presented in the excitation field, one might expect to see additional momenta appearing in the nonlinear response. This is, indeed, the case and the exciton polarizations corresponding to the "new" momenta are called the polarizations of *multiwave mixing* response for reasons which will become evident later.

In order to see the multiwave mixing response one needs to consider next orders of the perturbation theory. In the third order we have four terms, which we show schematically as

$$P_\mu^{(3)} \sim \left\langle B_\mu B B^\dagger B^\dagger \right\rangle + \left\langle B_\mu B^\dagger B B^\dagger \right\rangle + \left\langle B B^\dagger B_\mu B^\dagger \right\rangle + \left\langle B B_\mu B^\dagger B^\dagger \right\rangle \tag{29.87}$$

The operators that appear to the right from B_μ originate from the expansion of $S(t)$ in Equation 29.80, and those to the left from B_μ come from $S^\dagger(t)$. After some algebra we obtain

$$\begin{aligned} \hbar\frac{\partial}{\partial t}P_\mu^{(3)}(t) &= -(iE_\mu + \hbar\Gamma_\mu)P_\mu^{(3)}(t) \\ &\quad + i\hbar\sum_{\nu_1\nu_2\nu_3}\left\langle \hat{B}_{\nu_2}\hat{C}_{\mu,\nu_3}\hat{B}_{\nu_1}^\dagger\right\rangle \mathcal{A}_{\nu_3}(t)P_{\nu_2}^{(1)\star}(t)P_{\nu_1}^{(1)}(t) \\ &\quad + \sum_{\nu_1,\nu_2,\nu_3} P_{\nu_2}^{(1)^*}(t)\int_0^t dt'\left\langle \hat{D}_{\nu_2,\mu}e^{-i\mathcal{H}_{e-h}(t-t')/\hbar}\hat{B}_{\nu_3}^\dagger\hat{B}_{\nu_1}^\dagger\right\rangle \mathcal{A}_{\nu_3}(t')P_{\nu_1}^{(1)}(t'), \end{aligned} \tag{29.88}$$

where we have introduced the operator

$$\hat{D}_{\nu,\mu} = \left[\hat{B}_\nu, \left[\hat{B}_\mu, \mathcal{H}_{e-h}\right]\right], \tag{29.89}$$

which appears in calculations as $\langle 0|\hat{B}_\nu^\dagger\hat{B}_\mu^\dagger(\mathcal{H}_{e-h} - E_\nu - E_\mu) = \langle 0|\hat{D}_{\nu,\mu}$. In terms of the electron and hole operators $\hat{D}_{\nu,\,\mu}$ is written as

$$\begin{aligned} \hat{D}_{\nu,\mu} &= \int d\mathbf{x}_1 d\mathbf{y}_1 d\mathbf{x}_2 d\mathbf{y}_2 \phi_\mu^\star(\mathbf{x}_1,\mathbf{y}_1)\phi_\nu^\star(\mathbf{x}_2,\mathbf{y}_2)V(\mathbf{x}_1,\mathbf{y}_1 : \mathbf{x}_2,\mathbf{y}_2) \\ &\quad \times c_{s_\mu}(\mathbf{y}_1)\upsilon_{\sigma_\mu}(\mathbf{x}_1)c_{s_\nu}(\mathbf{y}_2)\upsilon_{\sigma\nu}(\mathbf{x}_2), \end{aligned} \tag{29.90}$$

where $V(\mathbf{x}_1, \mathbf{y}_2; \mathbf{x}_2, \mathbf{y}_2) = V(\mathbf{x}_1 - \mathbf{x}_2) + V(\mathbf{y}_1 - \mathbf{y}_2) - V(\mathbf{x}_1 - \mathbf{y}_2) - V(\mathbf{x}_2 - \mathbf{y}_1)$ is the energy of the Coulomb interaction between two excitons.

We rewrite the last term in Equation 29.88 in order to express it solely in terms of the polarizations of the linear response. For this we notice that for a symmetric matrix $G_{\kappa,\lambda} = G_{\lambda,\kappa}$ one has

$$\begin{aligned} &\frac{i\hbar}{2}\sum_{\kappa,\lambda} G_{\kappa,\lambda}e^{-i(E_\kappa+E_\lambda)t/\hbar}\frac{\partial}{\partial t}\left[P_\kappa^{(1)}(t)P_\lambda^{(1)}(t)e^{i(E_\kappa+E_\lambda)t/\hbar}\right] \\ &\quad = \sum_{\kappa,\lambda} G_{\kappa,\lambda}P_\kappa^{(1)}(t)\mathcal{A}_\lambda(t). \end{aligned} \tag{29.91}$$

Using this relation for the terms under the time integral in Equation 29.88 (the symmetry is ensured by $[\hat{B}_\mu^\dagger, \hat{B}_\nu^\dagger] = 0$) and integrating by part we obtain

$$\begin{aligned} \frac{\partial}{\partial t}P_\mu^{(3)}(t) &= -\left(i\frac{E_\mu}{\hbar} + \Gamma_\mu\right)P_\mu^{(3)}(t) + i\sum_{\kappa,\lambda,\nu} C_{\lambda,\mu}^{\nu,\kappa}\mathcal{A}_\nu(t)P_\lambda^{(1)\star}(t)P_\kappa^{(1)}(t) \\ &\quad + \frac{1}{2}\sum_{\kappa,\lambda,\nu} P_\lambda^{(1)^*}(t)\Big[i\beta_{\lambda,\mu}^{\nu,\kappa}P_\kappa^{(1)}(t)P_\nu^{(1)}(t) \\ &\quad + \frac{1}{\hbar}\int_0^t dt'\,e^{-(\Gamma_\kappa+\Gamma_\nu)(t-t')}F_{\lambda,\mu}^{\nu,\kappa}(t-t')P_\kappa^{(1)}(t')P_\nu^{(1)}(t')\Big], \end{aligned} \tag{29.92}$$

where we have introduced the incoherent decay into the integral term assuming that the decay of the pair of excitons is not affected by the interaction between them. We also have introduced $C_{\lambda,\mu}^{\nu,\kappa} = \left\langle \hat{B}_\lambda\hat{C}_{\mu,\nu}\hat{B}_\kappa^\dagger\right\rangle$, $\beta_{\lambda,\mu}^{\nu,\kappa} = \left\langle \hat{D}_{\lambda,\mu}\hat{B}_\nu^\dagger\hat{B}_\kappa^\dagger\right\rangle$, and $F_{\lambda,\mu}^{\nu,\kappa}(\tau) = \left\langle \hat{D}_{\lambda,\mu}e^{-i\mathcal{H}_{e-h}\tau/\hbar}\hat{D}_{\nu,\kappa}^\dagger\right\rangle$.

The term $\propto C_{\lambda,\mu}^{\nu,\kappa}$ in Equation 29.92 accounts for the effect of Pauli blocking of electrons. The parameters $\beta_{\lambda,\mu}^{\nu,\kappa}$ and the memory function $F_{\lambda,\mu}^{\nu,\kappa}(\tau)$ describe the effect of the exciton–exciton correlations. It should be emphasized that since we have taken into account exactly the dynamics of the semiconductor nonperturbed by the external field these parameters exactly account for the Coulomb interaction between the excitons. More complex many-body states, for example, triples of excitons will contribute only to the higher orders of the perturbation theory.

The important feature of $C_{\lambda,\mu}^{\nu,\kappa}$, $\beta_{\lambda,\mu}^{\nu,\kappa}$, and $F_{\lambda,\mu}^{\nu,\kappa}$ is that all of them are proportional to $\delta(\mathbf{K}_\mu + \mathbf{K}_\lambda - \mathbf{K}_\nu - \mathbf{K}_\kappa)$. The proof consists of two steps. First, we notice that after we expand the definitions of the operators entering the respective mean values, we will have at hand the vacuum averages of a set of the electrons and holes creation and annihilation operators taken at different coordinates multiplied by some scalar functions. The vacuum averages, however, do not change if we uniformly shift in the plane of the quantum well all the coordinates of the creation and annihilation operators. Indeed, when we transform the operator under-averaging to the normal form, that is, with all creation operators collected on the left side we perform transpositions of neighboring operators, which are functions of, say, $\mathbf{x}_1$ and $\mathbf{x}_2$, taking into account their anticommutation relations. This results at most in the appearance of $\delta(\mathbf{x}_1 - \mathbf{x}_2)$; thus, the vacuum average can depend only on the differences between different coordinates and, therefore, it does not change after the uniform translation.

Next, we notice that the scalar functions, which are obtained after the expansion of all definitions, acquire the phase factor only due to the translational properties of the exciton wave functions (see Equation 29.66), all other functions including $\tilde{\Phi}$ and the Coulomb potential in the Hamiltonian $\mathcal{H}_{e-h}$ are translationally invariant, so eventually we end up with an integral of the form

$$I(\mathbf{K}_1,\ldots,\mathbf{K}_n)=\int d\mathbf{x}_1\ldots d\mathbf{x}_n \exp\left(i\sum_{j=1}^{n}\mathbf{K}_j\cdot\mathbf{x}_j\right)f(\mathbf{x}_1,\ldots,\mathbf{x}_n), \tag{29.93}$$

where $f(\mathbf{x}_1+\Delta\mathbf{x},\ldots,\mathbf{x}_n+\Delta\mathbf{x})=f(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. We want to show that $I(\mathbf{K}_1,\ldots,\mathbf{K}_n)\propto\delta(\mathbf{K}_1+\cdots+\mathbf{K}_n)$. To this end we present the integral in Equation 29.93 as

$$I(\mathbf{K}_1,\ldots,\mathbf{K}_n)=\int d\mathbf{x}_1 e^{i\mathbf{K}_1\cdot\mathbf{x}_1}\int d\mathbf{x}_2\ldots d\mathbf{x}_n \exp\left(i\sum_{j=2}^{n}\mathbf{K}_j\cdot\mathbf{x}_j\right) \times f(0+\mathbf{x}_1,\mathbf{x}_2-\mathbf{x}_1+\mathbf{x}_1\ldots,\mathbf{x}_n-\mathbf{x}_1+\mathbf{x}_1), \tag{29.94}$$

using the translational invariance of the function $f(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ and change the variables of integration in the multiple integral. This yields

$$I(\mathbf{K}_1,\ldots,\mathbf{K}_n)=\int d\mathbf{x}_1 \exp\left(i\sum_{j=1}^{n}\mathbf{K}_j\cdot\mathbf{x}_1\right)\int d\mathbf{x}_2\ldots d\mathbf{x}_n \exp\left(i\sum_{j=2}^{n}\mathbf{K}_j\cdot\mathbf{x}_j\right) \times f(0,\mathbf{x}_2\ldots,\mathbf{x}_n). \tag{29.95}$$

The multiple integral does not depend on $\mathbf{x}_1$, so we can perform integration over $\mathbf{x}_1$ obtaining $I(\mathbf{K}_1,\ldots,\mathbf{K}_n)\propto\delta(\mathbf{K}_1+\cdots+\mathbf{K}_n)$.

Thus, we prove that $P_\mu^{(3)}\propto\delta(\mathbf{K}_\mu+\mathbf{K}_\lambda-\mathbf{K}_\nu-\mathbf{K}_\kappa)$. Because four momenta have to add up to zero in order to have nonvanishing $P_\mu^{(3)}$, this type of response is called *four-wave mixing* response (see Figure 29.2). The possible values that $\mathbf{K}_\mu$ can take are obtained looking at different combinations of the momenta of the polarizations of the linear response, which are determined by the momenta of the incoming excitation, $\mathbf{K}_\parallel^{(1)}$ and $\mathbf{K}_\parallel^{(2)}$. We can see that in addition to $\mathbf{K}_\parallel^{(1)}$ and $\mathbf{K}_\parallel^{(2)}$ we also have $2\mathbf{K}_\parallel^{(2)}-\mathbf{K}_\parallel^{(1)}=\mathbf{K}_\parallel^{(2)}+\Delta\mathbf{K}$ and $\mathbf{K}_\parallel^{(1)}-\Delta\mathbf{K}$ in agreement with expected conservation of the momentum only up to multiples of $\Delta\mathbf{K}$.

It is important that the response in the four-wave mixing direction is created only owing to the many-body correlations in the semiconductor. As a result, the signal in the respective directions is not blurred by the nonabsorbed excitation field and the linear response. Therefore, four-wave mixing spectroscopy is a popular tool for experimentally studying the many-body properties of solids. Higher order of the perturbation theory will produce higher orders of wave mixing originating from the respective correlations, which recently started attracting significant attention.

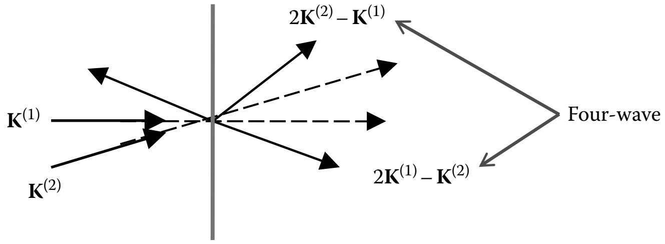
$2\mathbf{K}^{(2)}-\mathbf{K}^{(1)}$
$\mathbf{K}^{(1)}$
Four-wave
$\mathbf{K}^{(2)}$
$2\mathbf{K}^{(1)}-\mathbf{K}^{(2)}$

FIGURE 29.2 Linear and four-wave mixing responses. Bold solid arrows show the momenta of the external excitation field. Dashed lines show the linear. Notice that the horizontal component of momentum does not conserve and the outgoing wave exists in both half-spaces. The solid arrows show the outgoing waves produced by the exciton polarization of the four-wave mixing response.

We conclude the survey of the basics of the perturbation approach for developing a qualitative theory of the two-dimensional Fourier spectroscopy (Khalil et al. 2003, Jonas 2003), which recently became a powerful tool for studying the dynamics of the semiconductor excitations (Zhang et al. 2005, Li et al. 2006). The general scheme is the same as that of the standard four-wave mixing experiments with three pulses propagating along $\mathbf{K}^{(1)}$, $\mathbf{K}^{(2)}$, and $\mathbf{K}^{(3)}$ launched at $t=t_1$, t_2, and t_3, respectively. The difference from the standard technique is that measurements of the four-wave mixing signal are performed not at a fixed or at just a few values of the delay time $\tau=\min(t_2,t_3)-t_1$, but rather for a dense series of values lying in some interval. Subsequently, the Fourier transforms are done with respect to the delay time as well as with respect to the signal time. These two Fourier transforms constitute the two-dimensional Fourier spectrum.

Equation 29.92 can be directly applied for finding the exciton polarization produced by the excitation pulses, one only needs to take into account that the pulses with different directions of the wave vectors strike the quantum well at different times (Erementchouk et al. 2007). As the simplest approximation we neglect the contribution of the memory function $F(\tau)$ and assume that its effect can be accounted for by a renormalization of $\beta_{\lambda,\mu}^{\nu,\kappa}$. Let Ω and ω correspond to the Fourier transform with respect to the signal and delay times, respectively. As one would expect the two-dimensional spectrum $P_\mu^{(3)}(\Omega,\omega)$ has resonances at the frequencies corresponding to the light-hole and the heavy-hole excitons (see Figure 29.3). It turns out, however, that the form of the resonances depends on whether they are created due to the Pauli blocking $\propto C_{\lambda,\mu}^{\nu,\kappa}$ or due to the Coulomb interaction $\beta_{\lambda,\mu}^{\nu,\kappa}$. We denote the respective contributions to the spectrum by $P_p(\Omega,\omega)$ and $P_C(\Omega,\omega)$ respectively. One can show that near particular resonance

$$\begin{aligned}
P_p(\Omega,\omega)&\propto\frac{1}{(\Omega-\hbar^{-1}E_\kappa+i\Gamma_\kappa)(\omega+\hbar^{-1}E_\lambda+i\Gamma_\lambda)},\\
P_C(\Omega,\omega)&\propto\frac{1}{(\Omega-\hbar^{-1}E_\kappa+i\Gamma_\kappa)(\Omega-\hbar^{-1}E_\kappa+i\Gamma_\kappa+2i\Gamma_\lambda)(\omega+\hbar^{-1}E_\lambda+i\Gamma_\lambda)},
\end{aligned} \tag{29.96}$$

where κ and λ stand for 1s state of either heavy- or light-hole excitons.

The contribution to the spectrum from the Pauli blocking has the Lorentz form along both the vertical and horizontal axes. The reason is that, as can be readily seen from Equation 29.92, the dependence on both the signal time $[P_\mu^{(3)}(t)]$ and delay

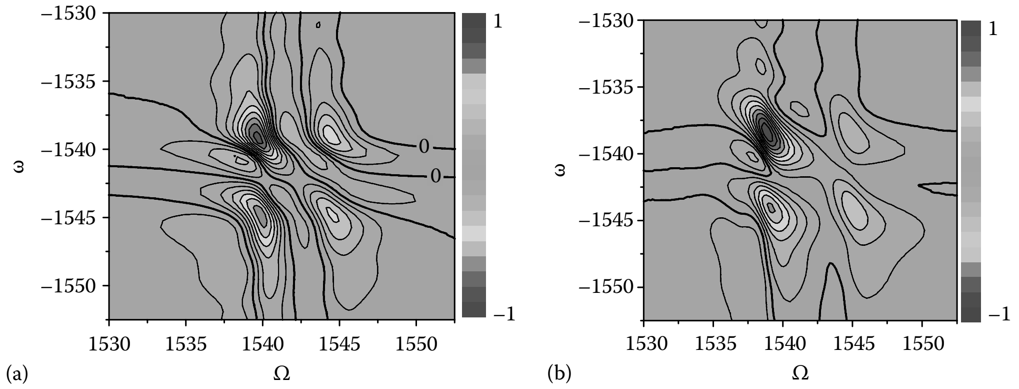

FIGURE 29.3 **(See color insert following page 20-14.)** The imaginary part of the two-dimensional Fourier spectrum, $P(\Omega, \omega)$. (a) The spectrum calculated using Equation 29.92. (b) The experimental results of Zhang et al. (2005).

time (through $P_{\lambda}^{(1)^*}(t)$ in the r.h.s. of Equation 29.92) is essentially the free evolution of the polarization directly created by the short excitation pulse. This evolution has the form of oscillations, which produce a simple pole after the Fourier transform. At the same time the resonances created by the Coulomb interaction, $P_C(\Omega, \omega)$, fall off asymptotically as $\propto 1/\omega$ and $\propto 1/\Omega^2$, along the ω- and Ω-axes, respectively. This is a consequence of the fact that the polarization of the four-mixing response $P_{\mu}^{(3)}(t)$ is continuously excited by the polarizations of the linear response. The dependence on the signal time is found as a convolution of the respective Green function with the source, similarly to how it was done in Equation 29.54. After the Fourier transform with respect to signal time it yields the product of Fourier images of the Green function and the source. This results in the asymptotic form $\propto 1/\Omega^2$ because of the harmonic time dependence of these functions.

This simple consideration shows that the two-dimensional Fourier spectroscopy allows one to distinguish between different types of the many-body correlations, which is almost impossible to do with the standard approach. For example, in experiments reported in Li et al. (2006), the form of the two-dimensional spectra is clearly asymmetric along different axes (see Figure 29.3b) indicating that due to relatively low intensity of excitations the contribution of the Pauli blocking into the nonlinear response was small compared to the effect of the Coulomb interaction.

29.5 Nonperturbative Methods: Rabi Oscillations

The drawback of the perturbative treatment of interaction of the electromagnetic field with the semiconductor is overestimating the resonant response. As follows from Equation 29.81 if the excitation field is tuned exactly to the resonance with an exciton state, the amplitude of the respective exciton polarization will indefinitely increase linearly with time. This, however, is impossible since with increase of the population of the conduction and the valence bands it becomes more difficult to add additional particles due to the Pauli blocking. Therefore, one would expect that the rate of change of the exciton polarization should change with time.

It turns out, however, that the effect of the environment is even stronger. The rate not only decreases, it reaches zero and then changes its sign, so the polarization starts to decrease until it completely vanishes. As a result the actual time dependence of the polarization has the form of oscillations, which are called *Rabi oscillations* by analogy with the effect of rotation of the magnetic moment in an external oscillating magnetic field.

In order to describe this effect we need to take into account the interaction with the electromagnetic field nonperturbatively. The problem of developing such nonperturbative description is that on the one hand there is no general way to study nonlinear operator equations other than using perturbation theory, while on the other hand there is no general way to map such equations onto the respective problem for regular functions. For example, averaging the first equation from Equations. 29.46 over vacuum (see e.g., Equation 29.30) one gets the dynamical equation for the interband polarization $p^{\star}_{\sigma,s}(\mathbf{x}_1,\mathbf{x}_2) = \left\langle \upsilon^{\dagger}_{\sigma}(\mathbf{x}_1) c^{\dagger}_{s}(\mathbf{x}_2) \right\rangle$, which contains in the r.h.s. the correlation functions of higher order, for example, $\langle \upsilon^{\dagger} c^{\dagger} c^{\dagger} c \rangle$, that is, the system will not be closed. An attempt to close the system by writing down the dynamical equation for new quantities will bring to attention correlation functions of even higher order and so on to infinity. Thus, one has to make the special approximation to truncate this hierarchy of the dynamical equations.

The simplest approximation of such kind is the Hartree–Fock approximation. It forces the closure at the level of the equation with respect to the interband polarization by applying the special decoupling scheme, which is illustrated by

$$\begin{aligned}\left\langle \upsilon^{\dagger}_{\sigma}(\mathbf{x}_1) c^{\dagger}_{s_1}(\mathbf{x}_2) c^{\dagger}_{s_2}(\mathbf{x}_3) c_{s_3}(\mathbf{x}_4) \right\rangle &\approx \left\langle \upsilon^{\dagger}_{\sigma}(\mathbf{x}_1) c^{\dagger}_{s_1}(\mathbf{x}_2) \right\rangle \left\langle c^{\dagger}_{s_2}(\mathbf{x}_3) c_{s_3}(\mathbf{x}_4) \right\rangle \\ &\quad - \left\langle \upsilon^{\dagger}_{\sigma}(\mathbf{x}_1) c^{\dagger}_{s_2}(\mathbf{x}_3) \right\rangle \left\langle c^{\dagger}_{s_1}(\mathbf{x}_2) c_{s_3}(\mathbf{x}_4) \right\rangle\end{aligned} \tag{29.97}$$

Thus, we can express the four-point correlation functions in terms of the products of the two-point ones.

In the framework of the Hartree–Fock approximation we obtain from Equations 29.46 in the rotating wave approximation (here and below the summation over dashed spin indices is implied)

$$\begin{aligned}
-i\hbar\frac{\partial}{\partial t}p^{\star}_{\sigma,s}(\mathbf{x}_1,\mathbf{x}_2) &= \hat{K}_{\sigma,s}[p]+\mathcal{A}^{\star}_{\sigma,s}(\mathbf{x}_1)\delta(\mathbf{x}_1-\mathbf{x}_2)\\
&\quad-\int d\mathbf{x}'[\tilde{\mathcal{A}}^{\star}_{\sigma',s}(\mathbf{x}',\mathbf{x}_2)h_{\sigma,\sigma'}(\mathbf{x}_1,\mathbf{x}')\\
&\quad-\tilde{\mathcal{A}}^{\star}_{\sigma,s'}(\mathbf{x}',\mathbf{x}_1)e_{s,s'}(\mathbf{x}_2,\mathbf{x}')],\\
-i\hbar\frac{\partial}{\partial t}e_{s_1,s_2}(\mathbf{x}_1,\mathbf{x}_2) &= \hat{K}_{s_1,s_2}[e]\\
&\quad+\int d\mathbf{x}'[\tilde{\mathcal{A}}^{\star}_{\sigma',s_1}(\mathbf{x}',\mathbf{x}_1)p_{\sigma',s_2}(\mathbf{x}',\mathbf{x}_2)\\
&\quad-\tilde{\mathcal{A}}^{\star}_{\sigma',s_2}(\mathbf{x}',\mathbf{x}_2)p^{\star}_{\sigma',s_1}(\mathbf{x}',\mathbf{x}_1)],\\
-i\hbar\frac{\partial}{\partial t}h_{\sigma_1,\sigma_2}(\mathbf{x}_1,\mathbf{x}_2) &= \hat{K}_{\sigma_1,\sigma_2}[h]\\
&\quad+\int d\mathbf{x}'[\tilde{\mathcal{A}}^{\star}_{\sigma_1,s'}(\mathbf{x}_1,\mathbf{x}')p_{\sigma_2,s'}(\mathbf{x}_2,\mathbf{x}')\\
&\quad-\tilde{\mathcal{A}}^{\star}_{\sigma_2,s'}(\mathbf{x}_2,\mathbf{x}')p^{\star}_{\sigma_1,s'}(\mathbf{x}_1,\mathbf{x}')],
\end{aligned}\qquad(29.98)$$

where $\mathcal{A}_{\sigma,s}(\mathbf{x})=\mathbf{A}_{\Omega}(\mathbf{x},t)\cdot d_{\sigma,s}$, the electron and hole correlation functions are $e_{s_1,s_2}(\mathbf{x}_1,\mathbf{x}_2)=\langle c^{\dagger}_{s_1}(\mathbf{x}_1)c_{s_2}(\mathbf{x}_2)\rangle$ and $h_{\sigma_1,\sigma_2}(\mathbf{x}_1,\mathbf{x}_2)=\left\langle \upsilon^{\dagger}_{\sigma_1}(\mathbf{x}_1)\upsilon_{\sigma_2}(\mathbf{x}_2)\right\rangle$, and we have introduced the modified coupling between the interband polarization and the charge densities

$$\tilde{\mathcal{A}}_{\sigma,s}(\mathbf{x},\mathbf{y})=\mathcal{A}_{\sigma,s}(\mathbf{x})\delta(\mathbf{x}-\mathbf{y})-V(\mathbf{x}-\mathbf{y})p_{\sigma,s}(\mathbf{x},\mathbf{y}).\qquad(29.99)$$

The operators

$$\begin{aligned}
\hat{K}_{\sigma,s}[p^{\star}] &= \hat{H}_{\sigma,\sigma'}p^{\star}_{\sigma',s}+p^{\star}_{\sigma,s'}\hat{H}_{s',s}-V(\mathbf{x}_1-\mathbf{x}_2)p^{\star}_{\sigma,s}(\mathbf{x}_1,\mathbf{x}_2)\\
&\quad-\hbar\Omega p^{\star}_{\sigma,s}(\mathbf{x}_1,\mathbf{x}_2),\\
\hat{K}_{s_1,s_2}[e] &= \hat{H}_{s_1,s'}e_{s',s_2}-e_{s_1,s'}\hat{H}_{s',s_2},\\
\hat{K}_{\sigma_1,\sigma_2}[h] &= \hat{H}_{\sigma_1,\sigma'}h_{\sigma',\sigma_2}-h_{\sigma_1,\sigma'}\hat{H}_{\sigma',\sigma_2}
\end{aligned}\qquad(29.100)$$

are expressed in terms of the integro-differential operators $\hat{H}$. The action of these operators is defined by

$$\begin{aligned}
\hat{H}_{s_1,s'}f_{\sigma'} &= \left[-\frac{\hbar^2}{2m_e}\nabla_1^2+\epsilon_{s_1}(\mathbf{x}_1)+U_{\Delta}(\mathbf{x}_1)\right]f_{s_1}(\mathbf{x}_1,\mathbf{x}_2)\\
&\quad-\int d\mathbf{x}'V(\mathbf{x}_1-\mathbf{x}')e_{s_1,s'}(\mathbf{x}_1,\mathbf{x}')f_{s'}(\mathbf{x}',\mathbf{x}_2),\\
\hat{H}_{\sigma_1,\sigma'}f_{\sigma'} &= -\nabla_1\cdot\frac{\hbar^2}{2M_{\sigma_1,\sigma'}}\cdot\nabla_1 f_{\sigma'}(\mathbf{x}_1,\mathbf{x}_2)-\left[\epsilon_{\sigma_1}(\mathbf{x}_1)+U_{\Delta}(\mathbf{x}_1)\right]f_{\sigma_1}(\mathbf{x}_1,\mathbf{x}_2)\\
&\quad-\int d\mathbf{x}'V(\mathbf{x}_1-\mathbf{x}')h_{\sigma_1,\sigma'}(\mathbf{x}_1,\mathbf{x}')f_{\sigma'}(\mathbf{x}',\mathbf{x}_2),
\end{aligned}\qquad(29.101)$$

where $U_{\Delta}(\mathbf{x})=\langle\hat{U}_{\Delta}(\mathbf{x})\rangle$ (see Equation 29.47). It should be noted that for a semiconductor excited by a single-plane wave this potential vanishes in an overall neutral system due to the translational invariance. In the two-wave excitation setup the effect of U_{Δ} is small for bound exciton states if the order of the multiwave mixing response is not too high.

Equations 29.100 constitute the semiconductor Bloch equations, which are widely used for studying nonperturbative effects in the dynamics of optically excited semiconductors. It should be emphasized that the quantity of main interest is the exciton polarization P_{μ} since these are the excitons that constitute true single-particle semiconductor states. As follows from Equations 29.63 and 29.72, P_{μ} is related to the interband polarization $p_{\sigma,s}$ through

$$P_{\mu}=\int d\mathbf{x}\,d\mathbf{y}\,\phi^{\star}_{\mu}(\mathbf{x},\mathbf{y})p_{\sigma_{\mu},s_{\mu}}(\mathbf{x},\mathbf{y}).\qquad(29.102)$$

For bound exciton states this naturally introduces a characteristic spatial scale, the exciton Bohr radius r_B (see Equation 29.67). At this point we assume that we deal with a deep and narrow quantum well, $\epsilon(\mathbf{x})=\epsilon+\Delta\epsilon(z)$, and that the system is invariant with respect to translations along the quantum well, that is, there are no random fluctuations of the width of the well.

The existence of the characteristic spatial scale together with the assumption that the typical spatial variation of the external excitation in the plane of the quantum well is small compared to the Bohr radius $K^{(1,2)}r_B \ll 1$ allows one to simplify the semiconductor Bloch equation using the parametric approximation. For this we introduce new spatial variables $\mathbf{R}=(\mathbf{x}+\mathbf{y})/2$ and $\mathbf{r}=\mathbf{x}-\mathbf{y}$. Having in mind the consecutive convolution of the solutions with the exciton wave function we can neglect the terms $\propto\Delta K r_B$ and leave only the dependence on $\mathbf{R}$ running along the quantum well as a parameter. Thus, assuming for simplicity that $\mathbf{K}^{(1)}=0$ we approximate Equations 29.98 by

$$\begin{aligned}
i\hbar\dot{p}_{\sigma,s}(\mathbf{r},\mathbf{R}) &= \tilde{K}_{\sigma,s}[p]+\mathcal{A}_{\sigma,s}(\mathbf{R})\delta(\mathbf{r})\\
&\quad-\int d\mathbf{r}'\left[\tilde{\mathcal{A}}_{\sigma',s}(\mathbf{r}',\mathbf{R})h_{\sigma',\sigma}(\mathbf{r}'-\mathbf{r},\mathbf{R})-\tilde{\mathcal{A}}_{\sigma,s'}(\mathbf{r}',\mathbf{R})e_{s',s}(\mathbf{r}'-\mathbf{r},\mathbf{R})\right],\\
-i\hbar\dot{e}_{s_1,s_2}(\mathbf{r},\mathbf{R}) &= \hat{K}_{s_1,s_2}[e]\\
&\quad+\int d\mathbf{r}'\left[\tilde{\mathcal{A}}^{\star}_{\sigma',s_1}(\mathbf{r}',\mathbf{R})p_{\sigma',s_2}(\mathbf{r}'+\mathbf{r},\mathbf{R})-\tilde{\mathcal{A}}_{\sigma',s_2}(\mathbf{r}',\mathbf{R})p^{\star}_{\sigma',s_1}(\mathbf{r}'-\mathbf{r},\mathbf{R})\right],\\
-i\hbar\dot{h}_{\sigma_1,\sigma_2}(\mathbf{r},\mathbf{R}) &= \hat{K}_{\sigma_1,\sigma_2}[h]\\
&\quad+\int d\mathbf{r}'\left[\tilde{\mathcal{A}}^{\star}_{\sigma_1,s'}(\mathbf{r}',\mathbf{R})p_{\sigma_2,s'}(-\mathbf{r}-\mathbf{r}',\mathbf{R})-\tilde{\mathcal{A}}_{\sigma_2,s'}(\mathbf{r}',\mathbf{R})p^{\star}_{\sigma_1,s'}(\mathbf{r}+\mathbf{r}',\mathbf{R})\right].
\end{aligned}\qquad(29.103)$$

Here $\tilde{\mathcal{A}}_{\sigma,s}(\mathbf{r},\mathbf{R})=\mathcal{A}_{\sigma,s}(\mathbf{R})\delta(\mathbf{r})-V(\mathbf{r})p_{\sigma,s}(\mathbf{r},\mathbf{R})$ and the integro-differential operators in Equations 29.100 are substituted by

$$\begin{aligned}
\hat{H}_{s_1,s'}f_{s'} &= \left[\frac{\hbar^2}{2m_e}\nabla^2_{\mathbf{r}}+\epsilon_{s_1}(z)\right]f_{s_1}(\mathbf{r},\mathbf{R})-\int d\mathbf{r}'V(\mathbf{r}')e_{s_1,s'}(\mathbf{r}',\mathbf{R})f_{s'}(\mathbf{r}-\mathbf{r}',\mathbf{R}),\\
\hat{H}_{\sigma_1,\sigma'}f_{\sigma'} &= -\nabla_{\mathbf{r}}\cdot\frac{\hbar^2}{2\mathrm{M}_{\sigma_1,\sigma'}}\cdot\nabla_{\mathbf{r}}f_{\sigma'}(\mathbf{r},\mathbf{R})-\epsilon_{\sigma_1}(z)f_{\sigma_1}(\mathbf{r},\mathbf{R})\\
&\quad-\int d\mathbf{r}'V(\mathbf{r}')h_{\sigma_1,\sigma'}(\mathbf{r}',\mathbf{R})f_{\sigma'}(\mathbf{r}-\mathbf{r}',\mathbf{R}).
\end{aligned}\qquad(29.104)$$

Equations 29.103 are sufficiently general for studying various nonlinear effects. Before we turn to further approximations in order to extract the Rabi oscillations, we would like to sketch the procedure to find the polarizations of multiwave mixing response. As has been discussed earlier in the case of two-wave excitation the translational invariance of the semiconductor equations along the quantum well is broken. Instead one has the symmetry with respect to finite translations $\Delta\mathbf{R} = 2\pi\Delta\mathbf{K}/\Delta K^2$. The same holds for the semiconductor Bloch equations, which is especially evident in the parametric approximation. Indeed, the parametric dependence on $\mathbf{R}$ originates from $\mathbf{A}(\mathbf{R})$, which is a periodic function of $\Delta\mathbf{R}$ and, hence, so are the solutions of Equations 29.103. Thus, the interband polarization $p_{\sigma,s}(\mathbf{r} = \mathbf{0}, \mathbf{R})$, which serves as the source of the outgoing re-emitted electromagnetic wave (see Equation 29.40), can be expanded in the Fourier series

$$p_{\sigma,s}(0,\mathbf{R}) = \sum_{m=-\infty}^{\infty} p_{\sigma,s}^{(m)} e^{im\Delta\mathbf{K}\cdot\mathbf{R}}. \tag{29.105}$$

The terms $m = 0$ and $m = 1$ correspond to the momenta brought by the excitation field, while the remaining terms are the polarizations of multiwave mixing response, for example, $m = -1, 2$ give the four-wave mixing polarizations, $m = -2, 3$ are the six-wave mixing components, and so on. In order to find the amplitudes of the multiwave mixing responses we merely use the formula of the inverse Fourier transform

$$p_{\sigma,s}^{(m)} = \int_0^1 d\alpha\, p_{\sigma,s}(0, \alpha\Delta\mathbf{R}) e^{-2\pi i\alpha m}. \tag{29.106}$$

Thus, once the solution of Equation 29.103 is known, the polarizations of the multiwave mixing response can be found immediately.

In order to see the oscillatory behavior of the interband polarization driven by the external field we need to make drastic simplifications of Equation 29.103, which correspond to considering the case of very short excitation pulses of high intensity. Such simplification is possible owing to the existence of the characteristic spatial scale and the radius of exciton states responsible for the resonant response. We consider all the terms in the r.h.s. of Equation 29.103 besides the kinetic energy as sources. The response, therefore, is governed by the free-particle propagators for the correlation functions and the Coulomb propagator for the interband polarization. The fundamental property of these propagators is that in the short time limit $t \to 0$ they turn to spatial δ-functions owing to fast spatial oscillations of the kernel

$$K(\mathbf{r}_1, \mathbf{r}_2; t) = \sqrt{\frac{m}{2\pi i t}} \exp\left[im\left|\mathbf{r}_1 - \mathbf{r}_2\right|^2 / 2t \right]. \tag{29.107}$$

For example, for the interband polarization in the short time limit, the Coulomb propagator can be approximated by the free-particle propagator (Blinder 1984, Kunikeev 2000) with $m \approx m_h = (m_e^{-1} + M_{hh})^{-1}$ and $m \approx m_l = (m_e^{-1} + M_{lh})^{-1}$, where m_σ are the excitons' reduced masses. The effect of the dispersion of the electrons and holes on the *exciton* polarization is expressed as the distortion of the exciton wave function under the action of the free-particle propagators. The time scale, at which the distortion becomes essential, can be estimated as the time required for the initial δ-shape of the kernel to acquire the width of the order of the exciton Bohr radius. This leads to the estimate

$$T_c \sim \frac{mr_B^2}{4\pi}. \tag{29.108}$$

The numerical value of the typical time scale for GaAs is determined by $m_{hh} = 0.45m_0$, $m_{lh} = 0.082m_0$, $m = m_\Delta$, and $r_B = r_{xh} = \epsilon_b/2m_{xh}e^2$, where m_0 is the electron mass in empty space and $\epsilon_b = 13$ is the background dielectric function. Using these values in Equation 29.108 we find $T_c \sim 20$ fs. We would like to draw attention to the fact that $T_c \gg T_R \approx 1$ fs (see Equation 29.57), so that the rotating wave approximation is applicable at such time scales. Thus, for pulses with duration shorter than 20 fs and sufficiently high intensity, we can treat the semiconductor excitations as dispersionless. It can be shown that for the states in the discrete spectrum the condition is even less strict (Erementchouk and Leuenberger 2008).

In the dispersionless limit, Equations 29.103 take the simple form

$$\begin{aligned}
i\hbar\dot{p}_{\sigma,s}(\mathbf{r},\mathbf{R}) &= \mathcal{A}_{\sigma,s}(\mathbf{R})\delta(\mathbf{r}) - \mathcal{A}_{\sigma',s}(\mathbf{R})h_{\sigma',\sigma}(-\mathbf{r},\mathbf{R}) - \mathcal{A}_{\sigma,s'}(\mathbf{R})e_{s',s}(\mathbf{r},\mathbf{R}),\\
-i\hbar\dot{e}_{s_1,s_2}(\mathbf{r},\mathbf{R}) &= \mathcal{A}^{\star}_{\sigma',s_1}(\mathbf{R})\rho_{\sigma',s_2}(\mathbf{r},\mathbf{R}) - \mathcal{A}_{\sigma',s_2}(\mathbf{R})\rho^{\star}_{\sigma',s_1}(-\mathbf{r},\mathbf{R}), \qquad (29.109)\\
-i\hbar\dot{h}_{\sigma_1,\sigma_2}(\mathbf{r},\mathbf{R}) &= \mathcal{A}^{\star}_{\sigma_1,s'}(\mathbf{R})\rho_{\sigma_2,s'}(-\mathbf{r},\mathbf{R}) - \mathcal{A}_{\sigma_2,s'}(\mathbf{R})\rho^{\star}_{\sigma_1,s'}(\mathbf{r},\mathbf{R}),
\end{aligned}$$

where we assumed the strong resonance regime and neglected the small contribution $(\epsilon_s - \epsilon_\sigma - \hbar\Omega)p_{\sigma,s}$. The solutions of Equations 29.109 are singular, $\propto\delta(\mathbf{r})$, but by virtue of Equation 29.102 it does not pose any problem.

We limit ourselves to the consideration of the simplest case when the external field is circularly polarized with positive helicity (right polarized) and has the amplitude constant in the time, which corresponds to the continuous wave excitation tuned to resonance with the gap. The condition of circular polarization leaves only two nonvanishing interband polarizations corresponding to p_{h+} and p_{l+}, that is $\{\sigma = -3/2, s = -1/2\}$ and $\{\sigma = -1/2, s = 1/2\}$, respectively. Thus we find for the respective interband polarizations

$$p_{\sigma,s}(\mathbf{r},\mathbf{R};t) = -i\delta(\mathbf{r})\frac{1}{\hbar\Omega_{\sigma,s}(\mathbf{R})}\mathcal{A}_{\sigma,s}(\mathbf{R})\sin\left[\Omega_{\sigma,s}(\mathbf{R})t\right], \tag{29.110}$$

where $\hbar\Omega_{\sigma,s}(\mathbf{R}) = \left|\mathcal{A}_{\sigma,s}(\mathbf{R})\right|$. It should be noted that only when $\Omega_{\sigma,s}$ is uniform along the quantum well, that is, in the case of the single-wave excitation, it defines the frequency of the Rabi oscillations. When the semiconductor is excited by two incoming waves, the frequency of temporal oscillations of the polarizations of the multiwave mixing response at sufficiently long

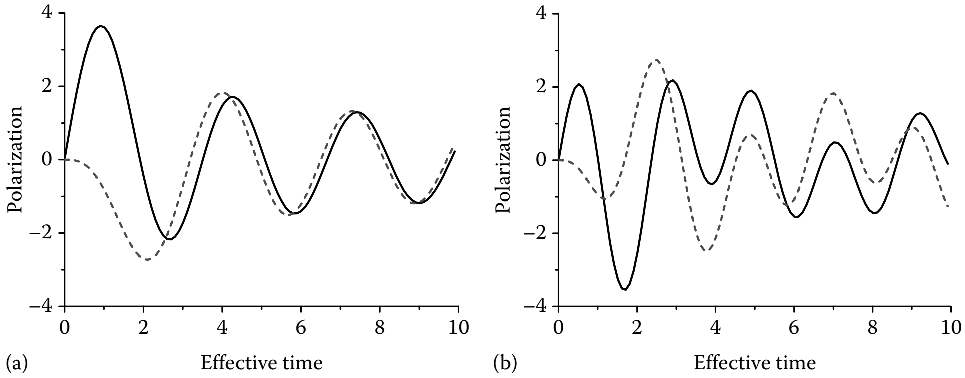

FIGURE 29.4 The exciton polarization in the case when the excitation consists of two plane waves with noncollinear wave vectors. The solid and dashed lines show the response in the forward ($m = 1$) and the four-wave mixing ($m = 2$) directions for (a) $\Omega^{(1)}_{\sigma,s}/\Omega^{(2)}_{\sigma,s} = 1$, (b) $\Omega^{(1)}_{\sigma,s}/\Omega^{(2)}_{\sigma,s} = 2$. The effective time is defined as $\tilde{t} = \Omega^{(2)}_{\sigma,s} t$.

time scales have two characteristic frequencies obtained as superpositions of the Rabi frequencies of each pulse $\left|\Omega^{(1)}_{\sigma,s} \pm \Omega^{(2)}_{\sigma,s}\right|$ (see Figure 29.4).

29.6 Conclusion

In order to keep the introduction into the theory of the semiconductor nonlinear response concise, we have concentrated on the problem of low energy excitation, when the dynamics of the semiconductor is determined by the bound states of the electron–hole pairs (excitons). In this regime the most natural description is provided by the effective quantum field theory, which accounts for the nonconstant number of particles in the semiconductor bands. The operator equations of motion, in turn, give access to all the information about the semiconductor and the electromagnetic fields. Using these equations one can either study the time evolution of the observables such as the macroscopic electromagnetic field $\langle \mathbf{A}(\mathbf{x},t)\rangle$ or the exciton polarization $\left\langle \hat{B}_\mu \right\rangle$; or one can turn, for example, to the single-particle exciton density matrix $\left\langle \hat{B}^\dagger_\lambda \hat{B}_\mu \right\rangle$ in order to incorporate into the picture the decoherence effect of the interaction with the phonon bath; or one can study the many-body properties of the semiconductor excitations looking at, say, the two-exciton density matrix $\left\langle \hat{B}^\dagger_\kappa \hat{B}^\dagger_\lambda \hat{B}_\mu \hat{B}_\nu \right\rangle$.

Unfortunately, the analysis of the equations of motion is tremendously hardened by their nonlinearity. The complexity is not of merely technical origin but reflects the concealed variety of physical phenomena. This should be taken into account when the equations of motion are simplified. As demonstrated in the main text, one can relatively easily develop the perturbational description of the semiconductor dynamics. However, one should be careful in applying the results of the perturbational approach to the situation, when the intensity of the excitation is significantly high, so that the properly defined pulse area approaches π. The perturbational series diverges in this case signifying the appearance of a new class of phenomena related to the reconstruction of the spectrum. The effect of the reconstruction on the semiconductor dynamics requires a non-perturbative treatment and has been illustrated above for the example of the Rabi oscillations.

We have covered the basics of the theory to the degree where it can be straightforwardly applied to various physical situations. However, we have left many details mentioned only on the formal level, such as the symmetries of the semiconductor bands, the specific form of the tensors of effective masses, and the matrix elements of the momentum operator between different bands. These details, while being easy to incorporate into the theory developed in the main text, may significantly simplify the analysis of the particular case. As a source of the fundamental insight into the physics of semiconductors we would like to refer the reader to the great book by Yu and Cardona (2004), the details of the symmetrical properties of the solids are provided in Bir and Pikus (1974), and the optical semiconductor response is considered in Haug and Koch (2004).

Acknowledgment

We would like to thank Prof. Xiaoqin Li for providing us with the data obtained in the experiments reported in Zhang et al. (2005). We also thank Sergio Tafur for useful comments. We acknowledge support from NSF-ECCS 0725514, the DARPA/MTO Young Faculty Award HR0011-08-1-0059, NSF-ECCS 0901784, and AFOSR FA9550-09-1-0450.

References

Arfken, G. B. and H. J. Weber. *Mathematical Methods for Physicists*. Academic Press, New York, 6th edn., 2005.

Bir, G. L. and G. E. Pikus. *Symmetry and Strain-Induced Effects in Semiconductors*. Wiley, New York, 1974.

Blinder, S. M. Semiclassical approximation for the nonrelativistic Coulomb propagator. *Phys. Rev. Lett.*, 52(20):1771–1773, 1984.

Cohen-Tannoudji, C., J. Dupont-Roc, and G. Grynberg. *Atom–Photon Interactions. Basic Processes and Applications*. Wiley, New York, 1992.

Erementchouk, M. V. and M. N. Leuenberger. Rabi oscillations in semiconductor multiwave mixing response. *Phys. Rev. B*, 78:075206, 2008.

Erementchouk, M. V., M. N. Leuenberger, and L. J. Sham. Many-body interaction in semiconductors probed with two-dimensional Fourier spectroscopy. *Phys. Rev. B*, 76:115307, 2007.

Haug, H. and S. W. Koch. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*. World Scientific, Singapore, 2004.

Ivchenko, E. L. *Optical Spectroscopy of Semiconductor Nanostructures*. Alpha Science International, Oxford, U.K., 2005.

Johnson, N. F. and P. M. Hui. Theory of propagation of scalar waves in periodic and disordered composite structures. *Phys. Rev. B*, 48(14):10118–10123, 1993.

Jonas, D. M. Two-dimensional femtosecond spectroscopy. *Annu. Rev. Phys. Chem.*, 54:425–463, 2003.

Khalil, M., N. Demirdoven, and A. Tokmakoff. Coherent 2d IR spectroscopy: Molecular structure and dynamics in solution. *J. Phys. Chem. A*, 107:5258–5279, 2003.

Kunikeev, S. D. Coulomb propagator in the WKB approximation. *J. Phys. A Math. Gen.*, 33:5405–5428, 2000.

Langbein, W., G. Kocherscheidt, and R. Zimmermann. Probing localized excitons by speckle analysis of resonant light scattering. In H. Kalt and M. Hetterich (eds.), *Optics of Semiconductors and Their Nanostructures*, vol. 146 of *Springer Series in Solid-State Sciences*, pp. 47–72. Springer, Berlin, Germany, 2004.

Li, X., T. Zhang, C. N. Borca, and S. T. Cundiff. Many-body interactions in semiconductors probed by optical two-dimensional Fourier transform spectroscopy. *Phys. Rev. Lett.*, 96:057406, 2006.

Luttinger, J. M. Quantum theory of cyclotron resonance in semiconductors: General theory. *Phys. Rev.*, 102(4):1030–1041, 1956.

Luttinger, J. M. and W. Kohn. Motion of electrons and holes in perturbed periodic fields. *Phys. Rev.*, 97(4):869–883, 1955.

Mukamel, S. *Principles of Nonlinear Optical Spectroscopy*. Oxford University Press, New York, 1995.

Ostreich, T. Higher-order Coulomb correlation in the nonlinear optical response. *Phys. Rev. B*, 64:245203, 2001.

Ostreich, T., K. Schonhammer, and L. J. Sham. Theory of exciton–exciton correlation in nonlinear optical response. *Phys. Rev. B*, 58(19):12920–12936, 1998.

Parfitt, D. G. W. and M. E. Portnoi. The two-dimensional hydrogen atom revisited. *J. Math. Phys.*, 43(10):4681–4691, 2002.

Sakoda, K. *Optical Properties of Photonic Crystals*, vol. 80 of *Optical Sciences*. Springer, Berlin, Germany, 2nd edn., 2005.

Schiff, L. I. *Quantum Mechanics*. McGraw Hill, New York, 1949.

Schmitt-Rink, S., D. S. Chemla, and D. A. B. Miller. Linear and nonlinear optical properties of semiconductor quantum wells. *Adv. Phys.*, 38(2):89–188, 1989.

Yu, P. Y. and M. Cardona. *Fundamentals of Semiconductors: Physics and Materials Properties*. Springer, Berlin, Germany, 2004.

Zhang, T., C. N. Borca, X. Li, and S.T. Cundiff. Optical two-dimensional Fourier transform spectroscopy with active interferometric stabilization. *Opt. Express*, 13(19):7432–7431, 2005.

Zhu, B. and K. Huang. Effect of valence-band hybridization on the exciton spectra in GaAs–$Ga_{1-}xAl_x$As quantum wells. *Phys. Rev. B*, 36(15):8102–8108, 1987.

30

Light Scattering from Nanofibers

30.1 Introduction 30-1
30.2 Background....................30-1
30.3 Basic Theory....................30-3
Diffraction by a Cylindrical Nanofiber • Nanofiber Normal Modes • Calculation of the Scattered Field • Semicylinder on an Ideally Reflecting Surface
30.4 Some Numerical Results....................30-9
Angular Distribution of Scattered Light • Total Scattered Intensity versus Incidence Angle • Total Scattered Intensity versus Frequency • Excitation near Exciton Resonance
30.5 Experimental Implementation 30-11
30.6 Discussion....................30-12
30.7 Summary and Outlook....................30-12
Acknowledgments....................30-13
References....................30-13

Vladimir G. Bordo
A.M. Prokhorov General Physics Institute

30.1 Introduction

The progress in nanotechnology has led to a possibility of fabricating needlelike crystals that have transverse dimensions in the submicron range. They can be grown from different inorganic semiconductor materials such as InP (Duan et al., 2001; Wang et al., 2001), ZnO (Huang et al., 2001), GaN (Johnson et al., 2002), CdS (Duan et al., 2003), GaSb (Chin et al., 2006), as well as from *p*-phenylene oligomers (Balzer and Rubahn, 2001, 2005; Schiek et al., 2005). Alternatively, one can also obtain fibers with a diameter of a few tens of nanometers by means of drawing from silica (Tong et al., 2003). Such structures are called nanofibers or nanowires. The morphology of individual nanofibers and their mutual alignment can be controlled by the use of appropriate growth conditions and substrate surfaces (Balzer and Rubahn, 2005). The nanofibers can be detached from the substrate, thus allowing one to study their properties under a variety of different conditions, including free-floating nanoaggregates (Brewer et al., 2005). They act as optical waveguides (Balzer et al., 2003; Johnson et al., 2003; Tong et al., 2003; Barrelet et al., 2004) and demonstrate remarkable photoluminescence properties (Wang et al., 2001; Thilsing-Hansen et al., 2005). Under pumping conditions, they can operate as nanoscale lasers (“nanolasers”) (Huang et al., 2001; Johnson et al., 2002, 2003; Duan et al., 2003; Agarwal et al., 2005; Quochi et al., 2005, 2006; Chin et al., 2006; Zimmler et al., 2008).

Light scattering by nanofibers plays an important role in optical processes at the nanoscale. Being grown on a substrate, nanofibers form an irregular array. The optical properties of such an ensemble are determined, on the one hand, by the optical response of individual aggregates and, on the other hand, by light scattering from nanofibers, which results in interaction between them. Light incident at a sample can launch the nanofiber optical modes or create excitons. Both processes manifest themselves in the intensity distribution of scattered light, thus providing valuable information on different nanofiber properties.

30.2 Background

In the simplest theoretical approach, a nanofiber is represented by an infinitely long dielectric circular cylinder. Light scattering by such an object was the first diffraction problem to be solved rigorously. Its solution dates back to Lord Rayleigh (Lord Rayleigh, 1881), who found the electromagnetic field diffracted from a cylinder of arbitrary radius and refractive index illuminated by a plane wave incident perpendicularly to its axis. Much later, that result was generalized to the case of oblique incidence (Wait, 1955). If α is the incidence angle relatively normal to the cylinder axis, then the scattered radiation is propagated along the surface of the cone with apical angle $(\pi - 2\alpha)$ (Kerker, 1969). The field amplitudes inside the cylinder and outside it are expressed in terms of a series of the Bessel functions of the first kind and Hankel functions, respectively. The result depends on the polarization of the incident wave. One can decompose it into two independent polarizations: (1) the transverse magnetic (TM) mode for which there is no magnetic vector component along the cylinder axis and (2) the transverse electric (TE) mode for which there is no electric field component along that axis. When the incident wave propagates perpendicularly to the cylinder axis, both the field inside the cylinder and the scattered field are of the same polarization as the incident wave. For oblique incidence, however, the incident TM waves are partially converted into scattered TE waves. Similarly, the incident TE waves are partially converted into scattered TM waves.

Light scattering in the far zone is characterized by the *scattering efficiency*, Q_{sca}, which is defined as the scattering cross section per unit length of the cylinder integrated over all the values of the scattering angle, θ, divided by the cylinder diameter, $2a$. If the polarization of the scattered radiation is not specified, light scattering is completely described by the two scattering efficiencies, $Q_{\text{sca}}^{\text{TM}}$ and $Q_{\text{sca}}^{\text{TE}}$, where the superscript denotes the polarization of the incident wave. In the limit where the diameter of the cylinder is sufficiently small relative to the wavelength, all the results are considerably simplified. If the refractive index of the cylinder relative to that of the surrounding medium is n and the wave number of the incident wave is k, the scattering efficiencies are reduced to the following form (Kerker, 1969):

$$Q_{\text{sca}}^{\text{TM}} = \frac{\pi^2 (ka)^3}{8}(n^2-1)^2; \quad Q_{\text{sca}}^{\text{TE}} = \frac{\pi^2 (ka)^3}{4}\left(\frac{n^2-1}{n^2+1}\right)^2. \tag{30.1}$$

If the angle θ is not close to 0 or π,* the ratio of the intensities of the two components is

$$\rho = \frac{I^{\text{TE}}}{I^{\text{TM}}} = \frac{4\sin^2\theta}{(n^2+1)^2}. \tag{30.2}$$

This leads to the conclusion that the radiation scattered in both forward and backward directions will be partially polarized even if the incident radiation is completely unpolarized.

The abandonment of the assumption of an infinite cylinder length leads to the necessity to use numerical methods. For some ranges of the relevant parameters, simplified treatments can provide accurate results. If the cylinder radius a is much smaller than the wavelength λ, the integral equation for the electric field amplitude can be solved by the method of moments (Uzunoglu et al., 1978). In the case where the parameter $ka(n-1)$ is comparable with unity or smaller, an iterative method can be applied, which is quite effective for both long (Cohen et al., 1983) and short (Haracz et al., 1984) cylinders. For the general case of a finite cylinder length l, the approach based on solving the electric field integral equation by using the method of moments is not very efficient. A more appropriate way is to apply a Fourier transform, which leads to a Fredholm integral equation of the first kind. This method has been applied to the cylinders with aspect ratios $p = l/(2a)$ as large as 10 and the lengths parameters up to $kl \sim 7$ (Shepherd and Holt, 1983). An alternative approach exploits the extended boundary condition method (EBCM), also known as the T-matrix method, in which both the incident and scattered fields are expanded in spherical vector wavefunctions. The truncation of the expansions at N terms leads to the system of $2N$ linear equations for the scattered field coefficients. The EBCM calculations carried out for the backscattering from a dielectric rod in the microwave region have demonstrated a satisfactory agreement with the existing experimental data (Ruppin, 1990). However, the EBCM method is tractable only when the size parameters ka and p are less than certain limits; e.g., for $n = 1.31$, $ka < 2.75$ and $p < 5.0$ (Kuik et al., 1994). Its convergence also becomes more difficult with increasing the dielectric constant of the cylinder.

A remarkable feature of light scattering from a dielectric cylinder at a fixed angle of incidence is the existence of resonant peaks in the scattering cross section plotted versus the size parameter (Van Bladel, 1977). Such peaks do not occur for a perfectly conducting cylinder and their sharpness increases with the dielectric constant of the cylinder. The resonant frequencies are determined by the poles of the expansion coefficients in the complex plane of the size parameter (Owen et al., 1981). The linewidths of the peaks are very sensitive to the circularity of the cylinder and to optical loss within or on the cylinder surface. These findings were confirmed experimentally in the scattering of either broadband radiation or light from a tunable laser by optical-communication fibers and used for the accurate determination of their diameters and shapes (Ashkin et al., 1981; Owen et al., 1981). The peaks of another kind are displayed in the dependence of the differential scattering efficiency on the angle between the incident light beam and the cylinder axis (Birkhoff et al., 1977). Fitting the angular positions of observed maxima in the scattered intensity provides a sensitive method for determining fiber radii, reproducible within ±3%–5%.

An essential feature of light scattering from nanofibers is reflection of the scattered radiation from the substrate surface. Therefore, it is desirable to take this effect into account when modeling nanofibers. If the cylinder axis is oriented parallelly to a flat reflecting surface, the diffraction problem is reduced to a two-dimensional one. The influence of the substrate can be calculated by introducing the Fresnel reflection coefficients. The corresponding solution can be found in the form of a series of the Bessel and Hankel functions. The procedure involves the truncation of the expansions for the field amplitudes at some index N, which is related to the size parameter by the rule $N \simeq 3ka$ (Borghi et al., 1997). Light scattering from nanofibers grown on a substrate involves the multiple-scattering effect. In some special cases, the nanofibers form an array of mutually parallel nanoaggregates oriented either parallelly or perpendicularly to the substrate surface. As it has been shown for an arbitrary configuration of parallel cylinders, the light scattering in such a system can be expressed as an infinite sum of the orders of scattering (Twersky, 1952). The first-order scattering originates from a diffraction of the incident wave on each cylinder. The second order results from a diffraction of the first-order-scattered radiation on each cylinder, and so on. A common approach to calculating the scattered field is to express the multiple-scattering linear equations in matrix form and then to find the corresponding scattering coefficients either by iterative method, or by direct matrix inversion, or by using a quadratic-programming algorithm. By this means, it has been found, in particular, that in an array of abutting cylinders, the resonant character of the cylinders with respect to the size parameter, ka, is damped (Bever and Allebach, 1992).

---

* We reckon the angle θ from the plane perpendicular to the plane of incidence (see Figure 30.1).

30.3 Basic Theory

In this section, we shall give the basic equations that describe light scattering from nanofibers (Bordo, 2006, 2007). We shall consider two models of a nanofiber: (1) an infinite circular dielectric cylinder (Figure 30.1) and (2) an infinite circular dielectric semicylinder placed on a perfectly reflecting surface (Figure 30.2). Although it might be possible that these models do not reproduce some details of light scattering, they nevertheless provide a fundamental understanding of this phenomenon. To match the conditions of many experiments with nanofibers, we assume that the incident radiation is represented by a pulsed Gaussian light beam.

30.3.1 Diffraction by a Cylindrical Nanofiber

Let us assume first that a nanofiber has a circular cross-section of radius a, both the nanofiber and surrounding medium are isotropic and are characterized by the dielectric functions ε_2 and ε_1, respectively. We choose the z axis along the nanofiber axis and the y axis in the plane containing both the z axis and the beam axis. Then the incident wave electric field has the form

$$\mathbf{E}_i(x,y,z,t) = \mathbf{E}_{i0}(x,y,z,t)\exp(-iq_{10}y + i\beta_0 z - i\omega_0 t), \quad (30.3)$$

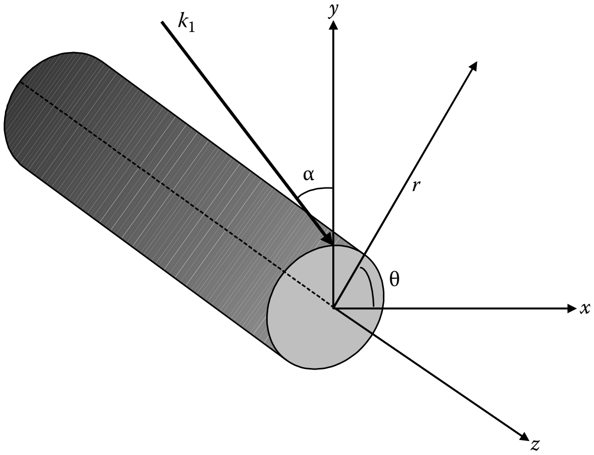

FIGURE 30.1 Model of a nanofiber: dielectric cylinder.

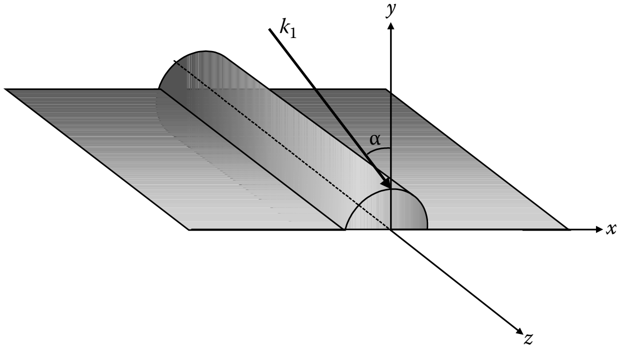

FIGURE 30.2 Model of a nanofiber: dielectric semicylinder on a perfectly reflecting surface.

where the field amplitude is given by

$$\mathbf{E}_{i0}(x,y,z,t) = \mathbf{E}_0 \exp\left[-\frac{x^2 + (y\sin\alpha + z\cos\alpha)^2}{2\sigma^2}\right]\exp\left(-\frac{t^2}{2\tau^2}\right), \quad (30.4)$$

where

- $-q_{10}$ and β_0 are the projections of the wave vector onto the axes y and z, respectively
- ω_0 is the frequency of the incident wave
- α is the angle of incidence with respect to the y axis
- σ and τ are the radius and pulse duration of the light beam, respectively

A similar expression is valid for the magnetic field of the incident beam, $\mathbf{H}_i(x, y, z, t)$.

In experiments, the beam radius is much larger than the nanofiber radius, i.e., $\sigma \gg a$. In such a case, in the vicinity of the nanofiber, Equation 30.4 is reduced to the following* form:

$$\mathbf{E}_{i0}(z,t) \approx \mathbf{E}_0 \exp\left(-\frac{z^2\cos^2\alpha}{2\sigma^2}\right)\exp\left(-\frac{t^2}{2\tau^2}\right). \quad (30.5)$$

On the other hand, to simplify the consideration, we shall assume that the beam radius is much less than the nanofiber length. Then one can neglect the edge effects and consider the nanofiber as infinitely long.

The total electric field has the form

$$\mathbf{E} = \begin{cases} \mathbf{E}_i + \mathbf{E}_1 & \text{if } r > a; \\ \mathbf{E}_2 & \text{if } r < a, \end{cases} \quad (30.6)$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ are the electric field vectors of the scattered field and the field inside the nanofiber, respectively. The total magnetic field vector, $\mathbf{H}$, is represented in a similar form. The fields $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{H}_1$, and $\mathbf{H}_2$ in turn can be found in terms of the Hertz vectors, Π_1 and Π_2, whose rectangular components satisfy the scalar wave equation. The Hertz vector for an infinite cylinder has the only nonzero component, $\Pi_z \equiv \psi$, which can be decomposed into two contributions, ψ^{TM} and ψ^{TE}, associated with TM and TE polarizations, respectively. Then, the electromagnetic field components in the cylindrical coordinates (r, θ, z) are found as (Stratton, 1941)

$$E_{jr} = \frac{\partial^2\psi_j^{\mathrm{TM}}}{\partial z\partial r} - \frac{1}{cr}\frac{\partial^2\psi_j^{\mathrm{TE}}}{\partial t\partial\theta}, \quad (30.7)$$

$$E_{j\theta} = \frac{1}{r}\frac{\partial^2\psi_j^{\mathrm{TM}}}{\partial z\partial\theta} + \frac{1}{c}\frac{\partial^2\psi_j^{\mathrm{TE}}}{\partial t\partial r}, \quad (30.8)$$

---

* This approximation is valid outside the nanofiber region $|z| \leq a \tan\alpha$.

$$E_{jz} = \frac{\partial^2 \psi_j^{\mathrm{TM}}}{\partial z^2} - \frac{\epsilon_j}{c^2}\frac{\partial^2 \psi_j^{\mathrm{TM}}}{\partial t^2}, \tag{30.9}$$

$$H_{jr} = \frac{\epsilon_j}{r}\frac{\partial^2 \psi_j^{\mathrm{TM}}}{\partial t \partial \theta} + \frac{\partial^2 \psi_j^{\mathrm{TE}}}{\partial z \partial r}, \tag{30.10}$$

$$H_{j\theta} = -\epsilon_j \frac{\partial^2 \psi_j^{\mathrm{TM}}}{\partial t \partial r} + \frac{1}{r}\frac{\partial^2 \psi_j^{\mathrm{TE}}}{\partial z \partial \theta}, \tag{30.11}$$

$$H_{jz} = \frac{\partial^2 \psi_j^{\mathrm{TE}}}{\partial z^2} - \frac{\epsilon_j}{c^2}\frac{\partial^2 \psi_j^{\mathrm{TE}}}{\partial t^2}, \tag{30.12}$$

where

- c is the speed of light in vacuum
- the subscript j labels different media: $j = 1$ corresponds to the surrounding medium, whereas $j = 2$ denotes the nanofiber interior

We shall seek the functions ψ_j^{μ} (μ = TM or TE) in the form of the Fourier integral

$$\psi_j^{\mu}(x,y,z,t) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{\psi}_j^{\mu}(x,y;\beta,\omega)e^{i\beta z - i\omega t}d\beta d\omega. \tag{30.13}$$

The representation (30.13) can be considered as a superposition of an infinite number of the elementary waves $\tilde{\psi}_j^{\mu}e^{i\beta z - i\omega t}$, each of which should satisfy the wave equation. The Fourier-transformed quantities $\tilde{\psi}_j^{\mu}$, therefore, can be expanded in the series of the cylindrical functions as follows:

$$\tilde{\psi}_j^{\mathrm{TM}}(r,\theta;\beta,\omega) = \frac{1}{q_j^2}\sum_{n=0}^{\infty} Z_n(q_j r)\left[a_{jn}(\beta,\omega)\sin(n\theta) + b_{jn}(\beta,\omega)\cos(n\theta)\right], \tag{30.14}$$

$$\tilde{\psi}_j^{\mathrm{TE}}(r,\theta;\beta,\omega) = \frac{1}{q_j^2}\sum_{n=0}^{\infty} Z_n(q_j r)\left[c_{jn}(\beta,\omega)\sin(n\theta) + d_{jn}(\beta,\omega)\cos(n\theta)\right], \tag{30.15}$$

where the functions $Z_n(X)$ are determined as

$$Z_n(X) = \begin{cases} H_n^{(1)}(q_1 r) & \text{if } r > a; \\ J_n(q_2 r) & \text{if } r < a, \end{cases} \tag{30.16}$$

with

$$q_j = \sqrt{\frac{\omega^2}{c^2}\epsilon_j - \beta^2}, \tag{30.17}$$

J_n and $H_n^{(1)}$ are the Bessel functions of the first kind and Hankel functions, respectively. The choice of the Bessel functions for the cylinder interior ensures finiteness of the solution at its center, whereas the Hankel functions have proper behavior at infinite distance from the cylinder.

The continuity of the tangential components of the total fields, $E_{j\theta}$, E_{jz}, $H_{j\theta}$, and H_{jz}, across the boundary $r = a$ leads to the equations for the coefficients a_{jn}, b_{jn}, c_{jn}, and d_{jn}. Taking into account that $y = r\sin\theta$, one can use the expansion

$$e^{-iq_{10}r\sin\theta} = \sum_{n=-\infty}^{\infty} J_n(q_{10}r)e^{-in\theta} \tag{30.18}$$

in Equation 30.3. Then, combining the terms with the same θ-dependence in the boundary conditions and performing the inverse Fourier transform, one obtains the following equations:

$$\hat{M}_n\vec{A}_n = \vec{F}_n, \tag{30.19}$$

$$\hat{M}_n\vec{B}_n = \vec{G}_n, \tag{30.20}$$

where

$$\hat{M}_n = \begin{pmatrix} J_n(q_2a) & 0 & -H_n^{(1)}(q_1a) & 0 \\ \frac{i\beta n}{q_2^2 a}J_n(q_2a) & -\frac{i\omega}{cq_2}J_n'(q_2a) & -\frac{i\beta n}{q_1^2 a}H_n^{(1)}(q_1a) & \frac{i\omega}{cq_1}H_n^{(1)\varepsilon'}(q_1a) \\ \frac{i\omega\epsilon_2}{cq_2}J_n'(q_2a) & -\frac{i\beta n}{q_2^2 a}J_n(q_2a) & -\frac{i\omega\epsilon_1}{cq_1}H_n^{(1)'}(q_1a) & \frac{i\beta n}{q_1^2 a}H_n^{(1)}(q_1a) \\ 0 & J_n(q_2a) & 0 & -H_n^{(1)}(q_1a) \end{pmatrix} \tag{30.21}$$

and

$$\vec{A}_n = \begin{pmatrix} a_{2n} \\ d_{2n} \\ a_{1n} \\ d_{1n} \end{pmatrix}, \quad \vec{B}_n = \begin{pmatrix} -b_{2n} \\ c_{2n} \\ -b_{1n} \\ c_{1n} \end{pmatrix}. \tag{30.22}$$

Here, the prime above the Bessel and Hankel functions denotes differentiation with respect to their argument. The vector functions $\vec{F}_n$ and $\vec{G}_n$ depend on the polarization of the incident beam and are determined as

$$\vec{F}_n^{\mathrm{TM}} = \begin{pmatrix} -2i\sigma_n\tilde{E}_{i0}\cos\alpha J_n(q_{10}a) \\ 2\sigma_n\tilde{E}_{i0}\sin\alpha\left[nJ_n(q_{10}a)/q_{10}a\right] \\ 2\sigma_n\tilde{H}_{i0}J_n'(q_{10}a) \\ 0 \end{pmatrix}, \tag{30.23}$$

$$\vec{G}_n^{\mathrm{TM}} = \begin{pmatrix} -2\tau_n(1-\sigma_n)\tilde{E}_{i0}\cos\alpha J_n(q_{10}a) \\ -2i\tau_n(1-\sigma_n)\tilde{E}_{i0}\sin\alpha\left[nJ_n(q_{10}a)/q_{10}a\right] \\ -2i\tau_n(1-\sigma_n)\tilde{H}_{i0}J_n'(q_{10}a) \\ 0 \end{pmatrix}, \tag{30.24}$$

and

$$\vec{F}_n^{\mathrm{TE}} = \begin{pmatrix} 0 \\ -2i\tau_n(1-\sigma_n)\tilde{E}_{i0}J_n'(q_{10}a) \\ -2i\tau_n(1-\sigma_n)\tilde{H}_{i0}\sin\alpha\left[nJ_n(q_{10}a)/q_{10}a\right] \\ 2\tau_n(1-\sigma_n)\tilde{H}_{i0}\cos\alpha J_n(q_{10}a) \end{pmatrix}, \tag{30.25}$$

$$\vec{G}_n^{\mathrm{TE}} = \begin{pmatrix} 0 \\ -2\sigma_n\tilde{E}_{i0}J_n'(q_{10}a) \\ -2\sigma_n\tilde{H}_{i0}\sin\alpha\left[nJ_n(q_{10}a)/q_{10}a\right] \\ -2i\sigma_n\tilde{H}_{i0}\cos\alpha J_n(q_{10}a) \end{pmatrix}, \tag{30.26}$$

with

$$\sigma_n = \begin{cases} 0 & \text{if } n \text{ is even;} \\ 1 & \text{if } n \text{ is odd,} \end{cases} \tag{30.27}$$

and

$$\tau_n = \begin{cases} 1/2 & \text{if } n = 0; \\ 1 & \text{if } n \neq 0. \end{cases} \tag{30.28}$$

Here, we have introduced the Fourier-transformed amplitudes of the incident beam

$$\tilde{E}_{i0}(\beta,\omega) = E_0\frac{2\pi\sigma\tau}{\cos\alpha}\exp\left[-\frac{(\beta-\beta_0)^2\sigma^2}{2\cos^2\alpha}\right]\exp\left[-\frac{(\omega-\omega_0)^2\tau^2}{2}\right], \tag{30.29}$$

and a similar expression for $\tilde{H}_{i0}(\beta,\omega)$.

Solving Equations 30.19 and 30.20, one obtains the Fourier-transformed electromagnetic field components both inside the nanofiber and outside it. We shall be interested, however, in the field scattered by the nanofiber, which is described by the coefficients a_{1n}, b_{1n}, c_{1n}, and d_{1n}. They are found as

$$a_{1n}(\beta,\omega) = \frac{D_{1n}(\beta,\omega)}{D_n(\beta,\omega)}, \tag{30.30}$$

$$b_{1n}(\beta,\omega) = -\frac{D_{2n}(\beta,\omega)}{D_n(\beta,\omega)}, \tag{30.31}$$

$$c_{1n}(\beta,\omega) = \frac{D_{3n}(\beta,\omega)}{D_n(\beta,\omega)}, \tag{30.32}$$

$$d_{1n}(\beta,\omega) = \frac{D_{4n}(\beta,\omega)}{D_n(\beta,\omega)}, \tag{30.33}$$

where

D_n is the determinant of the matrix $\hat{M}_n$

D_{1n} and D_{4n} are the determinants of the matrices obtained from $\hat{M}_n$ by replacing the third and fourth columns with the column given by $\vec{F}_n$, respectively

D_{2n} and D_{3n} are the determinants of the matrices obtained from $\hat{M}_n$ by replacing the third and fourth columns with the column given by $\vec{G}_n$, respectively

The corresponding Fourier-transformed electric and magnetic field amplitudes can be written in the form

$$\tilde{E}_{1\mu}(r,\theta;\beta,\omega) = \tilde{E}_{i0}(\beta,\omega) \times \sum_{n=0}^{\infty}\left[A_{\mu n}(r;\beta,\omega)\sin(n\theta) + B_{\mu n}(r;\beta,\omega)\cos(n\theta)\right], \tag{30.34}$$

$$\tilde{H}_{1\mu}(r,\theta;\beta,\omega) = \tilde{H}_{i0}(\beta,\omega) \times \sum_{n=0}^{\infty}\left[C_{\mu n}(r;\beta,\omega)\sin(n\theta) + D_{\mu n}(r;\beta,\omega)\cos(n\theta)\right], \tag{30.35}$$

where the subscript μ runs over the components r, θ, and z, and the functions $A_{\mu n}$, $B_{\mu n}$, $C_{\mu n}$, and $D_{\mu n}$ are given by the equations

$$A_{rn}(r;\beta,\omega) = \frac{i\beta}{q_1}H_n^{(1)\prime}(q_1r)a_{1n}(\beta,\omega) - \frac{i\omega n}{cq_1^2r}H_n^{(1)}(q_1r)d_{1n}(\beta,\omega), \tag{30.36}$$

$$B_{rn}(r;\beta,\omega) = \frac{i\beta}{q_1}H_n^{(1)\prime}(q_1r)b_{1n}(\beta,\omega) + \frac{i\omega n}{cq_1^2r}H_n^{(1)}(q_1r)c_{1n}(\beta,\omega), \tag{30.37}$$

$$A_{\theta n}(r;\beta,\omega) = -\frac{i\beta n}{q_1^2r}H_n^{(1)}(q_1r)b_{1n}(\beta,\omega) - \frac{i\omega}{cq_1}H_n^{(1)\prime}(q_1r)c_{1n}(\beta,\omega), \tag{30.38}$$

$$B_{\theta n}(r;\beta,\omega) = \frac{i\beta n}{q_1^2r}H_n^{(1)}(q_1r)a_{1n}(\beta,\omega) - \frac{i\omega}{cq_1}H_n^{(1)\prime}(q_1r)d_{1n}(\beta,\omega), \tag{30.39}$$

$$A_{zn}(r;\beta,\omega) = H_n^{(1)}(q_1r)a_{1n}(\beta,\omega), \tag{30.40}$$

$$B_{zn}(r;\beta,\omega) = H_n^{(1)}(q_1r)b_{1n}(\beta,\omega). \tag{30.41}$$

$$C_{rn}(r;\beta,\omega) = \frac{i\omega\epsilon_1 n}{cq_1^2r}H_n^{(1)}(q_1r)b_{1n}(\beta,\omega) + \frac{i\beta}{q_1}H_n^{(1)\prime}(q_1r)c_{1n}(\beta,\omega), \tag{30.42}$$

$$D_{rn}(r;\beta,\omega) = -\frac{i\omega\epsilon_1 n}{cq_1^2 r} H_n^{(1)}(q_1 r)a_{1n}(\beta,\omega) + \frac{i\beta}{q_1} H_n^{(1)\prime}(q_1 r)d_{1n}(\beta,\omega), \tag{30.43}$$

$$C_{\theta n}(r;\beta,\omega) = \frac{i\omega\epsilon_1}{cq_1} H_n^{(1)\prime}(q_1 r)a_{1n}(\beta,\omega) - \frac{i\beta n}{q_1^2 r} H_n^{(1)}(q_1 r)d_{1n}(\beta,\omega), \tag{30.44}$$

$$D_{\theta n}(r;\beta,\omega) = \frac{i\omega\epsilon_1}{cq_1} H_n^{(1)\prime}(q_1 r)b_{1n}(\beta,\omega) + \frac{i\beta n}{q_1^2 r} H_n^{(1)}(q_1 r)c_{1n}(\beta,\omega), \tag{30.45}$$

$$C_{zn}(r;\beta,\omega) = H_n^{(1)}(q_1 r)c_{1n}(\beta,\omega), \tag{30.46}$$

$$D_{zn}(r;\beta,\omega) = H_n^{(1)}(q_1 r)d_{1n}(\beta,\omega). \tag{30.47}$$

Taking the Fourier transforms of Equations 30.34 and 30.35,

$$E_{1\mu}(r,\theta,z,t) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{E}_{1\mu}(r,\theta;\beta,\omega)e^{i\beta z - i\omega t}d\beta d\omega, \tag{30.48}$$

$$H_{1\mu}(r,\theta,z,t) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{H}_{1\mu}(r,\theta;\beta,\omega)e^{i\beta z - i\omega t}d\beta d\omega, \tag{30.49}$$

one obtains the solution of the problem under consideration.

The integrands in Equations 30.48 and 30.49 have the poles given by the zeros of the denominators $D_n(\beta,\omega)$ in Equations 30.30 through 30.33. On the other hand, these zeros determine the allowed values of pairs (β, ω) for the electromagnetic field in a free nanofiber, i.e., its normal modes (Stratton, 1941).

30.3.2 Nanofiber Normal Modes

Among the solutions of the equations

$$D_n(\beta,\omega) = 0, \tag{30.50}$$

there are those which are represented by pairs of real quantities (β_{nm}, ω_{nm}). They give the propagation constants and frequencies of the waveguide (bound) modes. Such modes provide a complete description of the light propagation along the fiber in the steady-state regime far from the light source (Snyder and Love, 1983). Besides that, Equation 30.50 has solutions for which either β or ω, or both, have imaginary parts, i.e.,

$$\beta = \beta^r + i\beta^i, \tag{30.51}$$

$$\omega = \omega^r - i\omega^i. \tag{30.52}$$

The contribution of such poles to the integrals (30.48) and (30.49) leads to the transients along the fiber length or in time, or in both length and time, respectively. If β^i is small, the modes of the first type can propagate over long distances, and their portions localized within or near the fiber core are known as leaky modes (Snyder and Love, 1983). In the following, to distinguish between different modes, we shall call the first-type modes as *space-decaying* (*SD*) modes, whereas the second-type ones as *time-decaying* (*TD*) modes. Figure 30.3 shows the dispersion curves $\beta^r(\omega)$ calculated numerically for the SD modes, SD_{nm}, with n = 0–3 and represented in dimensionless variables. It has been assumed that $\varepsilon_1 = 1$ and $\varepsilon_2 = 2.89$, which correspond to an isolated nanofiber obtained from an isotropic *para*-hexaphenyl material immersed in vacuum. The dispersion curves for the bound modes are also shown in Figure 30.3. They are disposed between the light lines $\omega = c\beta/\sqrt{\epsilon_1}$ (LL_1) and $\omega = c\beta/\sqrt{\epsilon_2}$ (LL_2), and are denoted as it is accepted in the optical waveguide theory (Snyder and Love, 1983). The dispersion curves for the quantities $\beta^i(\omega)$ are shown in Figure 30.4. Figure 30.5 represents the dispersion relations $\omega^r(\beta)$ for n = 0–3 calculated for the TD modes, TD_{nm}. The corresponding imaginary parts, $\omega^i(\beta)$, for the modes that are not strongly decaying are plotted in Figure 30.6. Here, the subscript m numerates the modes with a given n according to the order they approach the axis β^r for the SD-modes or the axis ω^r for the TD-modes. One can observe that some of these modes, both SD and TD, can be considered as a continuation of waveguide modes beyond cutoff, which coincides with the light line LL_1. The imaginary parts of the propagation constants or frequencies for such modes increase with the deviation of the corresponding real parts from cutoff.

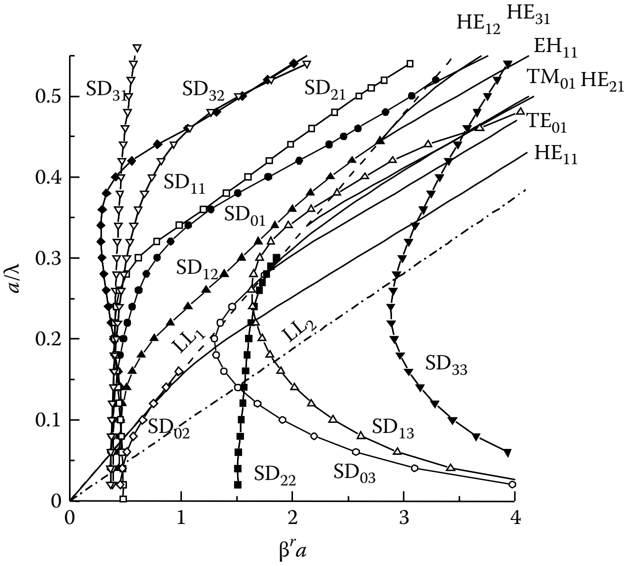

FIGURE 30.3 Dispersion curves of the SD modes. Real parts of the mode propagation constants. The modes with very large values of β^i are not shown in the figure. $\varepsilon_1 = 1$, $\varepsilon_2 = 2.89$. (Reprinted from Bordo, V., *J. Phys.: Condens. Matter*, 19, 236220-1, 2007. With permission.)

30.3.3 Calculation of the Scattered Field

When carrying out the integration in Equations 30.48 and 30.49, one can take into account that the functions $\tilde{E}_{i0}(\beta,\omega)$ and $\tilde{H}_{i0}(\beta,\omega)$ in the integrands are essentially different from

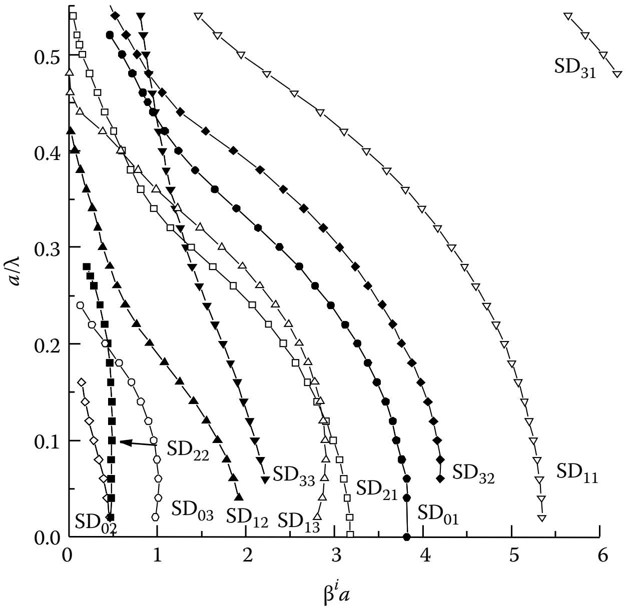

FIGURE 30.4 Same as Figure 30.3 but for the imaginary parts of the mode propagation constants. (Reprinted from Bordo, V., *J. Phys.: Condens. Matter*, 19, 236220-1, 2007. With permission.)

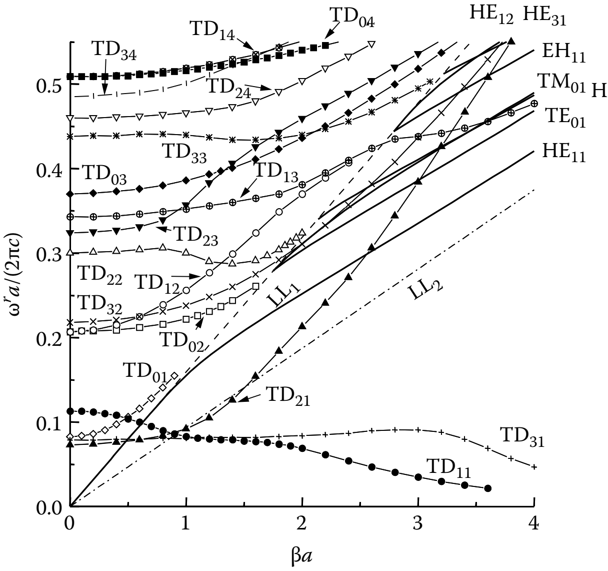

FIGURE 30.5 Dispersion curves of the TD modes. Real parts of the mode frequencies. $\varepsilon_1 = 1$, $\varepsilon_2 = 2.89$. (Reprinted from Bordo, V., *J. Phys.: Condens. Matter*, 19, 236220-1, 2007. With permission.)

zero around their maxima at $\beta = \beta_0$ and $\omega = \omega_0$. For typical experimental parameters, the widths of the corresponding peaks, $\Delta\beta \ll 1/a$ and $\Delta\omega \ll c/a$, and the slowly varying functions $A_{\mu n}(\beta, \omega)$, $B_{\mu n}(\beta, \omega)$, $C_{\mu n}(\beta, \omega)$, and $D_{\mu n}(\beta, \omega)$ can be taken off the integral at the point (β_0, ω_0).* The remaining integration leads to the result

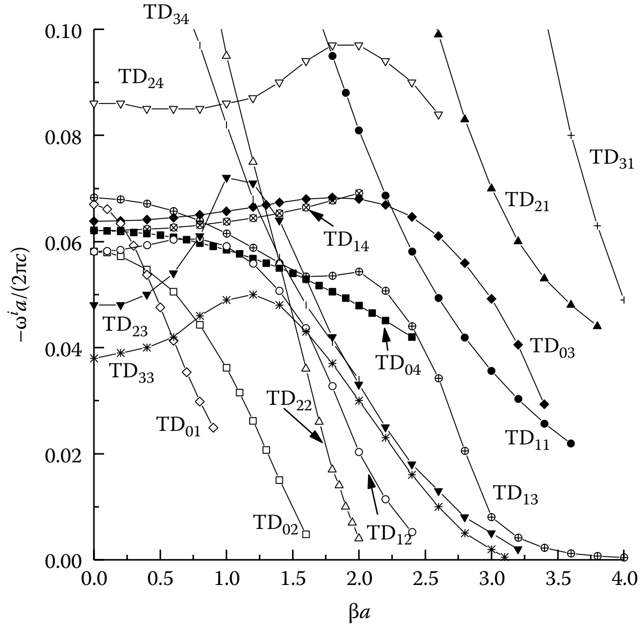

FIGURE 30.6 Same as Figure 30.5 but for the imaginary parts of the mode frequencies, which fall into the range $|\omega^i a/(2\pi c)| \le 0.1$. (Reprinted from Bordo, V., *J. Phys.: Condens. Matter*, 19, 236220-1, 2007. With permission.)

$$E_{1\mu}(r,\theta,z,t) = E_{i0}(z,t)\exp(i\beta_0 z - i\omega_0 t) \times \sum_{n=0}^{\infty}\left[A_{\mu n}(r;\beta_0,\omega_0)\sin(n\theta) + B_{\mu n}(r;\beta_0,\omega_0)\cos(n\theta)\right], \tag{30.53}$$

$$H_{1\mu}(r,\theta,z,t) = H_{i0}(z,t)\exp(i\beta_0 z - i\omega_0 t) \times \sum_{n=0}^{\infty}\left[C_{\mu n}(r;\beta_0,\omega_0)\sin(n\theta) + D_{\mu n}(r;\beta_0,\omega_0)\cos(n\theta)\right], \tag{30.54}$$

where $E_{i0}(z, t)$ and $H_{i0}(z, t)$ are the amplitudes of the incident wave electric and magnetic fields, respectively. In other words, under typical experimental conditions, the illumination by a pulsed light beam can be treated as the case of an incident plane monochromatic wave if instead of its amplitudes, E_0 and H_0, one substitutes the pulsed beam envelope functions, $E_{i0}(z, t)$ and $H_{i0}(z, t)$, respectively.

We shall be interested in the intensity of light scattered in the *xy* plane. This quantity is determined by the time-averaged radial component of the Poynting vector

$$S_r = \frac{c}{8\pi}\mathrm{Re}(E_{1\theta}H_{1z}^{\star} - E_{1z}H_{1\theta}^{\star}), \tag{30.55}$$

where the asterisk denotes complex conjugation. The substitution of the field components, as shown in Equations 30.53 and 30.54, into Equation 30.55 gives the intensity of the scattered electromagnetic field at any distance from the nanofiber, i.e., in both near-field and far-field regions. In the far zone where $r \gg \lambda$, the

---

* See Bordo (2007) for detail.

Hankel functions and their derivatives can be replaced by their asymptotical expansions. Then in the leading order one obtains

$$S_r = S_r^{\mathrm{TM}} + S_r^{\mathrm{TE}}, \tag{30.56}$$

where the TM- and TE-polarized components are given by

$$S_r^{\mathrm{TM}}(\theta) = \frac{\omega\epsilon_1}{4\pi^2 q_{10}^2 r}\left|\sum_{n=0}^{\infty}(-i)^n[a_{1n}(\beta_0,\omega_0)\sin(n\theta) + b_{1n}(\beta_0,\omega_0)\cos(n\theta)]\right|^2 \tag{30.57}$$

and

$$S_r^{\mathrm{TE}}(\theta) = \frac{\omega}{4\pi^2 q_{10}^2 r}\left|\sum_{n=0}^{\infty}(-i)^n[c_{1n}(\beta_0,\omega_0)\sin(n\theta) + d_{1n}(\beta_0,\omega_0)\cos(n\theta)]\right|^2, \tag{30.58}$$

respectively. The total intensity scattered in all directions per unit length of the nanofiber is determined by the quantity

$$S_{\mathrm{tot}} \equiv \int_0^{2\pi} S_r(\theta)d\theta = S_{\mathrm{tot}}^{\mathrm{TM}} + S_{\mathrm{tot}}^{\mathrm{TE}} \tag{30.59}$$

with

$$S_{\mathrm{tot}}^{\mathrm{TM}} = \frac{\omega\epsilon_1}{4\pi q_{10}^2 r}\left\{|b_{10}(\beta_0,\omega_0)|^2 + \sum_{n=1}^{\infty}\left[|a_{1n}(\beta_0,\omega_0)|^2 + |b_{1n}(\beta_0,\omega_0)|^2\right]\right\} \tag{30.60}$$

and

$$S_{\mathrm{tot}}^{\mathrm{TE}} = \frac{\omega}{4\pi q_{10}^2 r}\left\{|d_{10}(\beta_0,\omega_0)|^2 + \sum_{n=1}^{\infty}\left[|c_{1n}(\beta_0,\omega_0)|^2 + |d_{1n}(\beta_0,\omega_0)|^2\right]\right\}. \tag{30.61}$$

If in an experiment, one varies either the angle of incidence of the light beam at a fixed frequency, or tunes the light frequency at a fixed angle of incidence, the scattered intensity will also change. In the former case, the intensity (30.56) has maxima at the values $\beta_0 = \beta_{nm}^r$ corresponding to the SD modes. In the latter case, the maxima in intensity occur at $\omega_0 = \omega_{nm}^r$ and correspond to the TD modes. The widths of those peaks are determined by the quantities β_{nm}^i and ω_{nm}^i, respectively.[*]

30.3.4 Semicylinder on an Ideally Reflecting Surface

The approach developed above can be equally applied to the nanofiber model represented by a circular semicylinder placed on an ideally reflecting plane. When the reflecting plane at $y = 0$ is present, the total electric field in the half-space $y > 0$ can be written as

$$\mathbf{E} = \mathbf{E}_i + \mathbf{E}_i^{\mathrm{ref}} + \mathbf{E}_1 + \mathbf{E}_1^{\mathrm{ref}}, \tag{30.62}$$

where $\mathbf{E}_i^{\mathrm{ref}}$ and $\mathbf{E}_1^{\mathrm{ref}}$ are the electric fields specularly reflected from the plane $y = 0$ originating from the incident wave and the field scattered by the nanofiber, respectively. A similar expression is valid for the magnetic field vector. The condition of an ideally reflecting surface implies that the tangential components of the total electric field at it are equal to zero. This requirement will be fulfilled if to introduce, instead of the reflecting plane, an image semicylinder illuminated by the image of the incident wave below the plane $y = 0$ so that $E_r = E_z = 0$ at both $\theta = 0$ and $\theta = \pi$ (Rao and Barakat, 1994). The same is true for the magnetic field components. The fields scattered by the image semicylinder, $\mathbf{E}_1^{\mathrm{im}}$ and $\mathbf{H}_1^{\mathrm{im}}$, are determined by Equations 30.34 through 30.47 with the substitutions

$$a_{1n}^{\mathrm{im}} = -a_{1n}, \quad b_{1n}^{\mathrm{im}} = -b_{1n}, \quad c_{1n}^{\mathrm{im}} = -c_{1n}, \quad d_{1n}^{\mathrm{im}} = -d_{1n}. \tag{30.63}$$

Then the components of the field scattered by the semicylinder in the presence of the reflecting plane are found as $\mathbf{E}_1(\theta) + \mathbf{E}_1^{\mathrm{im}}(-\theta)$ and $\mathbf{H}_1(\theta) + \mathbf{H}_1^{\mathrm{im}}(-\theta)$. Formally, the field components can be obtained from Equations 30.53 and 30.54 with the following substitutions:

$$A_{\mu n} \to 2A_{\mu n}, \quad B_{\mu n} \to 0, \quad C_{\mu n} \to 2C_{\mu n}, \quad D_{\mu n} \to 0. \tag{30.64}$$

Let us assume that the angle of incidence, α, differs from zero. Then the fields $\mathbf{E}_i^{\mathrm{ref}}$ and $\mathbf{H}_i^{\mathrm{ref}}$ do not contribute to the scattered intensity determined by S_r. In the far zone, its TM- and TE-polarized components are given by

$$S_r^{\mathrm{TM}}(\theta) = \frac{\omega\epsilon_1}{\pi^2 q_{10}^2 r}\left|\sum_{n=1}^{\infty}(-i)^n a_{1n}(\beta_0,\omega_0)\sin(n\theta)\right|^2 \tag{30.65}$$

and

$$S_r^{\mathrm{TE}}(\theta) = \frac{\omega}{\pi^2 q_{10}^2 r}\left|\sum_{n=1}^{\infty}(-i)^n c_{1n}(\beta_0,\omega_0)\sin(n\theta)\right|^2, \tag{30.66}$$

respectively. The corresponding total intensity scattered in all directions per unit length of the nanofiber is determined by the quantity

$$S_{\mathrm{tot}} = \int_0^{\pi} S_r(\theta)d\theta = \frac{\omega}{2\pi q_{10}^2 r}\sum_{n=1}^{\infty}\left[\epsilon_1|a_{1n}(\beta_0,\omega_0)|^2 + |c_{1n}(\beta_0,\omega_0)|^2\right]. \tag{30.67}$$

---

[*] Besides the maxima relatively the light frequency discussed here, there may also be maxima originating from resonances in $\varepsilon_2(\omega)$.

Note that here the sum over the modes starts from $n = 1$, indicating that the modes with $n = 0$ are not excited in the course of light scattering. The suppression of such modes is conditioned by the presence of the reflecting plane.

Let us consider the special case of normal incidence with respect to the substrate surface ($\alpha = 0$). In such a case, $\beta_0 = 0$ and the equations for TM and TE polarizations are completely separated. This means that for TM incident wave polarization, one can set $c_{1n} = d_{1n} = 0$. Besides that, the vector $\vec{F}_n$, Equation 30.23, is nonzero only for odd indices n. Then the coefficients a_{1n} are nonzero for odd n and only such modes can be observed in scattering. For TE incident wave polarization, $a_{1n} = b_{1n} = 0$ and the vector $\vec{G}_n$, Equation 30.26, is nonzero for odd n. As a result, one obtains the same selection rule as for TM polarization.

30.4 Some Numerical Results

We shall illustrate the general theory considered above with some numerical examples. All the calculations have been carried out for $\varepsilon_2 = 2.89$ that corresponds to an isotropic *para*-hexaphenyl film. We shall compare the results obtained for two different models of a nanofiber.

30.4.1 Angular Distribution of Scattered Light

Let us consider first the dependence of the Poynting vector component S_r, Equation 30.55, calculated in the far zone on the detection angle, θ. This quantity is determined by the coefficients a_{1n}, b_{1n}, c_{1n}, and d_{1n}, as shown in Equations 30.30 through 30.33. Their variation with the nanofiber radius, a, is more rapid for larger values of $q_{10} = (\omega_0/c)\sqrt{\epsilon_1}\cos\alpha$ and hence for smaller angles of incidence, α. Figures 30.7 and 30.8 show the angular distributions of light scattered by a cylindrical nanofiber for the TM and TE incident light polarizations, respectively, and for $\alpha = 45°$. One can see that in the case of the TM incident wave polarization, there are two lobes directed along the normal to the plane of incidence (at $\theta = 0°$ and $\theta = 180°$). For relatively large nanofiber radius, additional two lobes appear. In the case of the TE incident wave polarization, these lobes are more pronounced. Besides that, for small nanofiber radius, the angular distribution of scattered light is more isotropic as compared to the TM polarization.

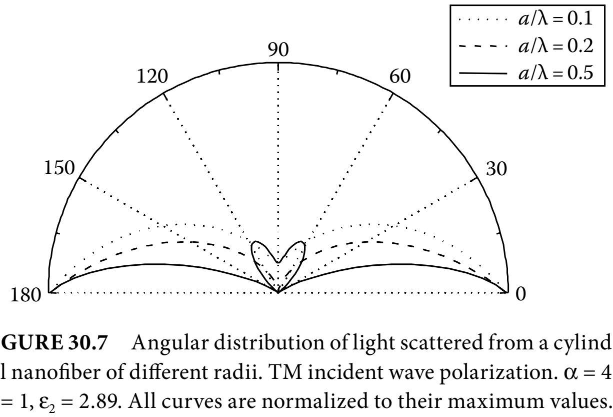

FIGURE 30.7 Angular distribution of light scattered from a cylindrical nanofiber of different radii. TM incident wave polarization. $\alpha = 45°$, $\varepsilon_1 = 1$, $\varepsilon_2 = 2.89$. All curves are normalized to their maximum values.

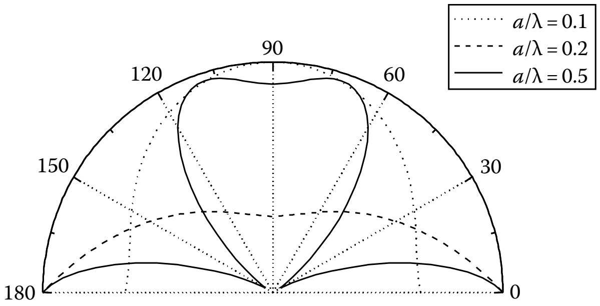

FIGURE 30.8 Same as Figure 30.7 but for TE incident wave polarization.

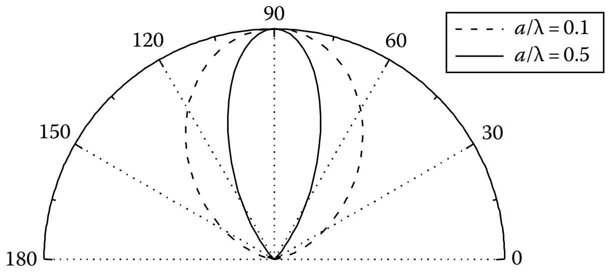

FIGURE 30.9 Same as Figure 30.7 but for a semicylindrical nanofiber on a perfectly reflecting surface.

The same quantity but calculated for a semicylindrical nanofiber placed on a perfectly reflecting substrate and for the TM incident wave polarization is depicted in Figure 30.9. The scattering diagram for the TE incident wave polarization has a similar form. In this case, the lobes directed along the substrate plane are suppressed and light is scattered mainly around the plane of incidence.

30.4.2 Total Scattered Intensity versus Incidence Angle

Let us consider the total scattered intensity, S_{tot}, as a function of the incidence angle, α, or, equivalently, the quantity $\beta_0 = (\omega_0/c)\sqrt{\epsilon_1}\sin\alpha$. Figure 30.10 shows this dependence calculated for a cylinder and for the TM incident wave polarization with an account of the terms $n = 0$–10 in Equations 30.60 and 30.61. The individual contributions from the terms with $n = 0$, 1, and 2 are also shown in Figure 30.10, whereas the other contributions have much smaller amplitude. The intensity drop at $\beta_0 a \approx 2.2$ originates from the approach to the cutoff at which $q_{10} = 0$. The curve corresponding to $n = 2$ exhibits a shoulder around $\beta_0 a \approx 1.75$, which is also seen in the total intensity curve. Turning to Figure 30.3, one can identify it as being originating from the mode SD_{21}. The contributions of the other modes are not pronounced because of the large values of the propagation constant imaginary parts, β^i, corresponding to them. This plot can be compared with Figure 30.11, which represents the results of calculations for a semicylindrical nanofiber placed on an ideally reflecting surface. In this case, the zeroth term does not contribute to S_{tot} and the contribution of the term with $n = 2$ is equal

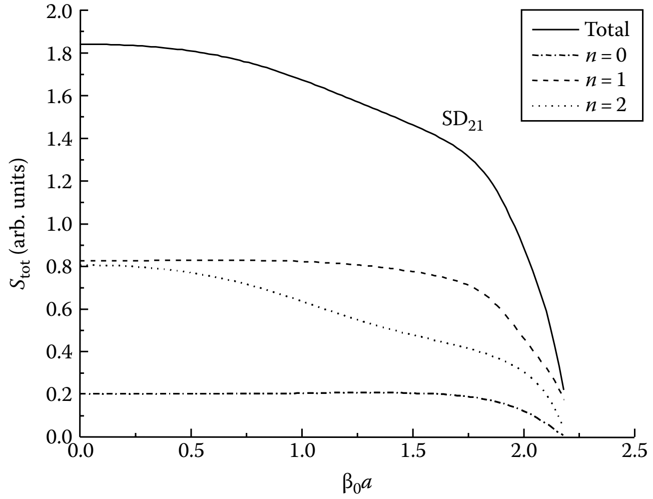

FIGURE 30.10 Light intensity scattered from a cylindrical nanofiber as a function of β_0 calculated for $\varepsilon_1 = 1$, $\varepsilon_2 = 2.89$, and $\omega_0 a/(2\pi c) = 0.35$. TM incident wave polarization.

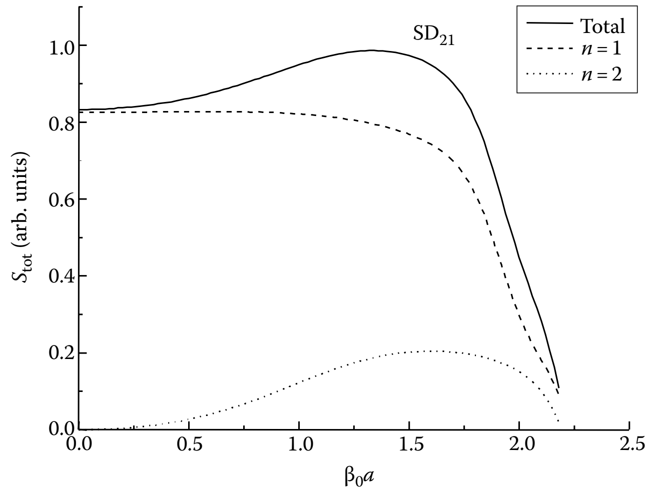

FIGURE 30.11 Same as Figure 30.10 but for a semicylindrical nanofiber on a perfectly reflecting surface.

to zero at $\beta_0 = 0$ (normal incidence) in accordance with the selection rule discussed in Section 30.3.4. As a result, the shoulder in the total intensity originating from the mode SD_{21} becomes more distinct.

30.4.3 Total Scattered Intensity versus Frequency

Another situation when the light beam frequency is scanned at a fixed angle of incidence ($\alpha = 0$) is illustrated in Figure 30.12 for the case of scattering of a TM-polarized beam by a cylindrical nanofiber. This time the features in the total scattered intensity originate from the TD modes. They can be identified when comparing the positions of maxima, which occur for the curves with different n with the mode frequencies at $\beta_0 = 0$ in Figure 30.5.

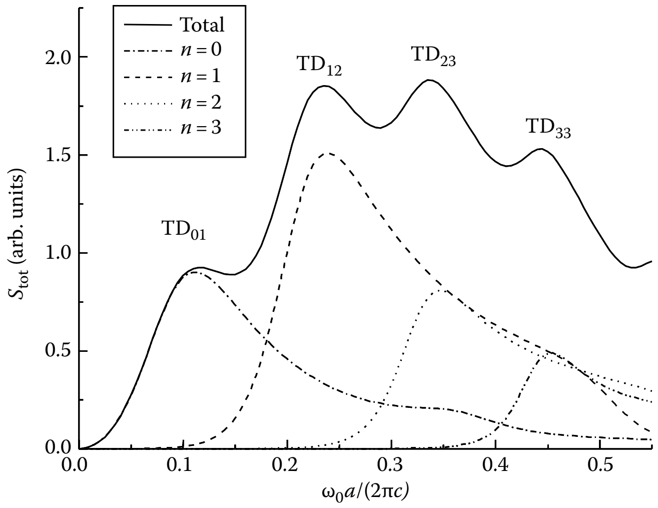

FIGURE 30.12 Light intensity scattered from a cylindrical nanofiber as a function of ω_0 calculated for $\varepsilon_1 = 1$, $\varepsilon_2 = 2.89$. Normal incidence. TM incident wave polarization.

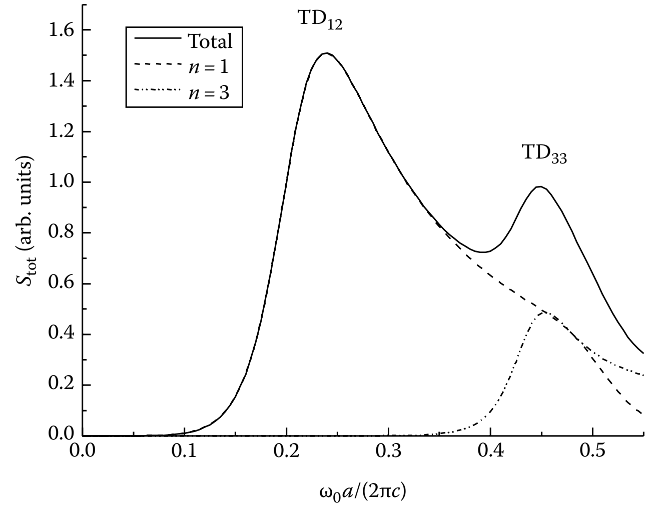

FIGURE 30.13 Same as Figure 30.12 but for a semicylindrical nanofiber on a perfectly reflecting surface. (Reprinted from Bordo, V., *J. Phys.: Condens. Matter*, 19, 236220-1, 2007. With permission.)

In the case of a semicylinder placed on an ideally reflecting plane (Figure 30.13), only modes with odd n can be excited, as it is expected from the selection rule (see Section 30.3.4). It is worthwhile to note that the modes distinct from those for TM polarization can be observed in the scattering of a TE-polarized light beam (not shown).

30.4.4 Excitation near Exciton Resonance

So far we have implied that the frequency of the incident beam is far from any resonance in the dielectric function of the nanofiber and its frequency dispersion can be neglected. Let us assume

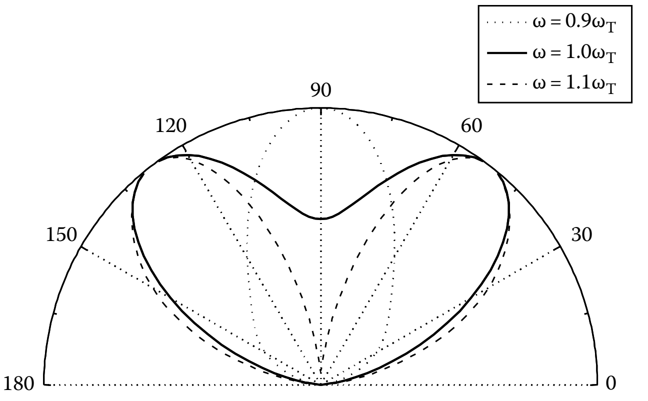

FIGURE 30.14 Angular distribution of scattered light in the vicinity of the exciton resonance. TM incident wave polarization. $\varepsilon_1 = 1$, $\varepsilon_\infty = 2.89$, $a = 4c/\omega_T$, $\omega_L = 1.2\omega_T$, $\Gamma = 10^{-3}\omega_T$, and $\alpha = 30°$.

now that the frequency ω_0 is close to an exciton resonance of the nanofiber. Then its dielectric function can be written as follows:

$$\epsilon_2(\omega) = \frac{\omega_L^2 - \omega^2}{\omega_T^2 - \omega^2 + i\omega\Gamma}\epsilon_\infty, \qquad (30.68)$$

where

ω_T and ω_L are the frequencies of the transverse and longitudinal excitons, respectively
Γ is the relaxation constant
ε_∞ is the high-frequency dielectric constant, and the spatial dispersion has been neglected

Figure 30.14 shows the angular distribution of the scattered light in this case calculated for a semicylindrical nanofiber on an ideally reflecting surface. It is seen that the polar diagram of scattering changes dramatically in a narrow spectral interval when the frequency of the incident beam scans across the exciton resonance.

30.5 Experimental Implementation

Experimentally, one can observe light scattering from nanofibers in the following setup (Figure 30.15) (Fiutowski et al., 2008). A sample with deposited organic nanofibers is attached to a half-sphere made of fused silica ($n_s = 1.48$ at 325 nm) with a radius of 10 mm. The sample is oriented so that the nanofiber axes would be parallel to the plane of light incidence. A flat domain of parallelly oriented organic nanofibers is illuminated with a linearly polarized beam of a He-Cd laser ($\lambda = 325$ nm) from the half-sphere side, and the scattered light intensity is detected along the normal to the sample surface from the backside. The extinction ratio of the polarization system is 10^5. The light beam passes a pinhole of diameter 400 μm and is collimated before entering the half-sphere, which results in a beam divergence of less than 0.7° after leaving the half-sphere. The angles of incidence can be varied between 20° and 65° with respect to the normal axis of the plane of the half-sphere. A photomultiplier mounted on a goniometric table allows one to detect scattered light intensity distributions from a spot on the sample of a diameter of about 500 μm. The acceptance angle of the detection system is 44°. The absolute and relative accuracies of the incident and the detection angles are 1° and 0.5°, respectively. To observe only the wavelengths of interest, one can use band pass or interferometric filters in front of the photomultiplier.

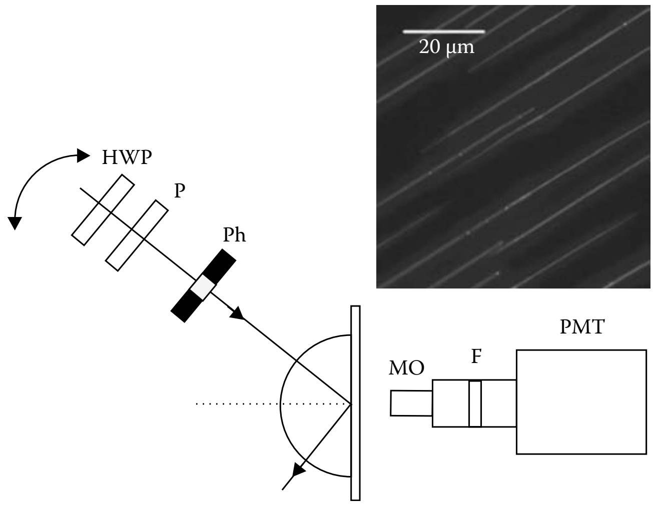

FIGURE 30.15 Experimental setup. HWP, half-wave plate; P, polarizer; Ph, pinhole; MO, microscope objective; F, filter; and PMT, photomultiplier. Inset: Fluorescence microscopy image of a sample domain. (Reprinted from Fiutowski, J. et al., *Appl. Phys. Lett.*, 92, 073302-1, 2008. With permission.)

The intensity of light scattered perpendicularly to the sample surface as a function of the incidence angle, α, is shown in Figure 30.16. Clear peaks at $\alpha = 40°$ in both TE and TM incident light polarizations are observed. Those peaks originate from the

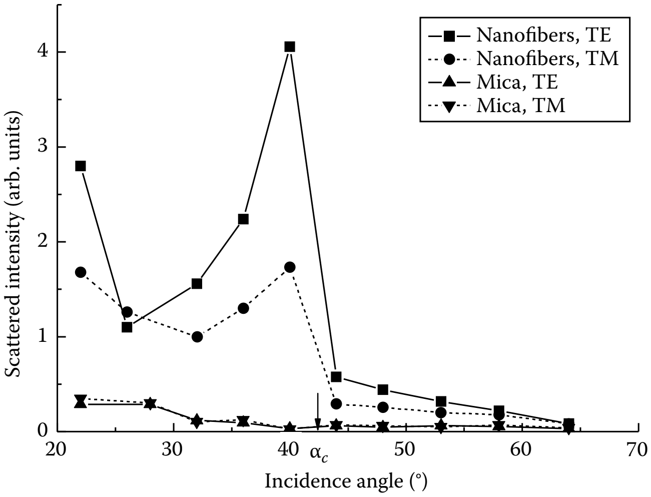

FIGURE 30.16 Intensity of light scattered perpendicularly to the sample surface as a function of the angle of incidence. The incident wave polarization is indicated in the inset. (Reprinted from Fiutowski, J. et al., *Appl. Phys. Lett.*, 92, 073302-1, 2008. With permission.)

nanofiber array exclusively, as seen from a comparison with the data obtained from a nanofiber-free mica surface. The scattered intensity decreases when the incidence angle exceeds the critical angle for total internal reflection at the quartz semisphere–air interface, $\alpha_c = 42.5^\circ$. To describe the optical properties of a nanofiber grown on a substrate, we use the nanofiber model represented by a semicylinder placed on an ideally reflecting surface and considered above. Figure 30.17 shows the dispersion curves of the nanofiber normal modes plotted in the range of the size parameter, a/λ, corresponding to the experimentally observed typical nanofiber widths ($2a$) as well as the wavelength of the He-Cd laser used for excitation. The condition of the phase matching between the incident light and the nanofiber normal mode can be written as $\beta = (2\pi/\lambda)n_s \sin\alpha$. The corresponding dispersion line for the angle $\alpha = 40^\circ$, at which the resonance in scattering is observed, is also shown in Figure 30.17.

The analysis of the polarization properties of the excited normal modes allows one to specify them. The sharp peak seen in the scattered light intensity at $\alpha = 40^\circ$ has comparable amplitudes in both TE and TM incident wave polarizations. It can therefore be associated with the excitation of the hybrid SD_{12} mode. The increase in intensity at $\alpha < 26^\circ$ originates probably from a resonance with the hybrid SD_{21} mode. The mode SD_{01} is suppressed in scattering due to the presence of the reflecting substrate (see Section 30.3.4).

The positions of the maxima in Figure 30.16 correspond to the most probable size parameter within the illuminated spot. The broadenings of the peaks observed in light scattering allow one to estimate the distribution of the nanofibers in the array over their widths. The angular broadening of about 7° corresponds to a scattering of 0.12λ in nanofiber diameters, which gives approximately 40 nm for the relevant wavelength. This value agrees qualitatively with that measured via atomic force microscope (AFM).

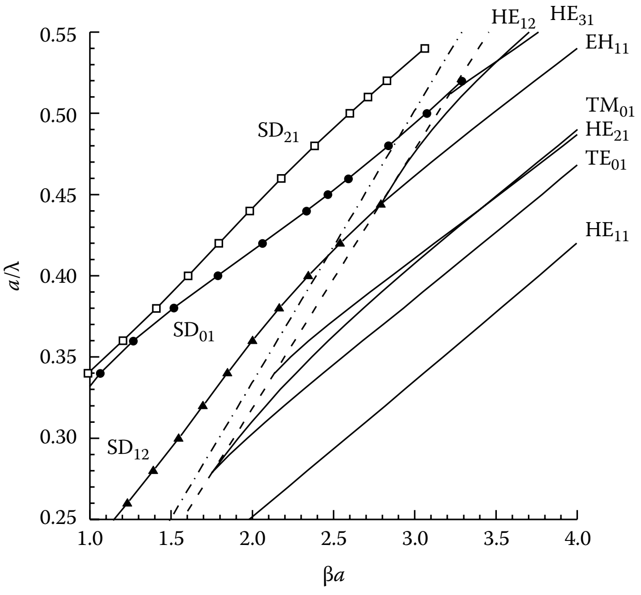

FIGURE 30.17 Dispersion curves of the SD modes plotted in the range of size parameters corresponding to measured nanofiber widths. The inclined dot-and-dash line indicates the dispersion for light incident at the angle $\alpha = 40^\circ$.

30.6 Discussion

We have considered here the simplest theoretical models of nanofibers and applied them for the interpretation of the experimental results on light scattering from an array of organic nanoaggregates deposited on a mica substrate. Although the morphology of nanoaggregates is far from a cylindrical shape, light scattering from them can be similarly described in terms of excitation of their radiative normal modes. This conclusion is also supported by the measurements of photoluminescence excited in the course of light scattering (Fiutowski et al., 2008). In that case, the peaks in the photoluminescence intensity versus the angle of incidence originate from matching with the nanoaggregates' normal modes, both radiative and waveguiding. Such measurements being carried out at a varying frequency of incident light should allow one to determine the dispersion curves of the actual normal modes of nanoaggregates. Then, using an appropriate model, one could extract information on the morphology of nanofibers. However, it is necessary to keep in mind that the actual normal modes may differ from those of an isotropic cylindrical nanofiber. In particular, the optical anisotropy of nanoaggregates can cause distortion of the dispersion curves as well as their splitting.

Another point which has to be taken into account is the different morphology of nanoaggregates within the illuminated spot. As we have already mentioned, this leads to a broadening of the peaks in the intensity of scattered light. On the other hand, the data obtained from the scattering of a focused light beam reflect the local morphology of a sample and, thus, can be used for its local characterization.

30.7 Summary and Outlook

In this chapter, we have considered the theory which describes the scattering of a pulsed Gaussian light beam at an infinitely long dielectric cylinder. The results obtained for this model have been generalized to the model represented by a semicylinder placed on a perfectly reflecting plane. Although being rather simple, these approaches reproduce the main features of light scattering from nanofibers. In particular, this process can be understood in terms of the excitation of nanofiber radiative modes. As a result, the dependence of the scattered light intensity on the angle of incidence exhibits maxima corresponding to matching with the nanofiber SD modes.

These conclusions are in agreement with the measurements of light scattering from organic nanoaggregates grown on mica, which have demonstrated experimental evidence of mode launching in this system. Those modes show up as pronounced peaks in light scattering from the nanofibers. Using the model, one is able to identify specific modes that have been excited in the course of light scattering from the sample. The obtained results provide detailed information about possible electromagnetic mode propagation in nanosized, needle-shaped aggregates and also form the basis for a new way of optically characterizing the morphology of sub-wavelength-sized nanostructures via far-field scattering. The further development of this technique by using a tunable light source should allow one to determine the actual dispersion curves of nanofibers. On the other hand, one can use a broadband light source to illuminate the sample instead of a laser. In such a case, one should expect to observe peaks in the spectrum of scattered light, which originates from the excitation of TD nanofiber modes. The implementation of this technique with the use of a scanning near-field optical microscope for the registration of scattered field would provide an additional advantage. The data obtained in such a setup are not influenced by an averaging over the illuminated spot and, thus, are related to individual nanofibers. Another possible application of light scattering from nanofibers stems from correlation between the positions of peaks in scattered intensity and the dispersion of nanofiber normal modes. Any change in the optical properties of nanofibers caused by either heating, or mechanical tension, or adsorption of molecules from ambient medium would result in a shift of the peaks in light scattering. This opens up new opportunities for the creation of nanosensors of various kinds. Due to its generic nature, the method discussed in this chapter is not limited to organic nanofibers but can be equally well applied to other kind of light-emitting nanoaggregates.

Acknowledgments

The author is grateful to Prof. H.-G. Rubahn and Dr. L. Jozefowski for the discussions of the experimental aspects of light scattering from nanofibers.

References

Agarwal, R., Barrelet, C., and Lieber, C. (2005). Lasing in single cadmium sulfide nanowire optical cavity. *Nano Lett.*, 5:917–920.

Ashkin, A., Dziedzic, J., and Stolen, R. (1981). Outer diameter measurement of low birefringence optical fibers by a new resonant backscatter technique. *Appl. Opt.*, 20:2299–2303.

Balzer, F., Bordo, V., Simonsen, A., and Rubahn, H.-G. (2003). Optical waveguiding in individual nanometer-scale organic fibers. *Phys. Rev. B*, 67:115408–1–8.

Balzer, F. and Rubahn, H.-G. (2001). Dipole-assisted self-assembly of light-emitting *p*-nP needles on mica. *Appl. Phys. Lett.*, 79:3860–3862.

Balzer, F. and Rubahn, H.-G. (2005). Growth control and optics of organic nanoaggregates. *Adv. Funct. Mater.*, 15:17–24.

Barrelet, C., Greytak, A., and Lieber, C. (2004). Nanowire photonic circuit elements. *Nano Lett.*, 4:1981–1985.

Bever, S. and Allebach, J. (1992). Multiple scattering by a planar array of parallel dielectric cylinders. *Appl. Opt.*, 31:3524–3532.

Birkhoff, R., Ashley, J., Hubbel, H. Jr., and Emerson, L. (1977). Light scattering from micron-size fibers. *J. Opt. Soc. Am.*, 67:564–569.

Bordo, V. (2006). Light scattering from a nanofiber: Exact numerical solution of a model system. *Phys. Rev. B*, 73:205117-1–205117-7.

Bordo, V. (2007). Theory of nanofibre excitation by a pulsed light beam. *J. Phys. Condens. Matter*, 19:236220-1–236220-12.

Borghi, R., Santarsiero, M., Frezza, F., and Schettini, G. (1997). Plane-wave scattering by a dielectric circular cylinder parallel to a general reflecting flat surface. *J. Opt. Soc. Am. A*, 14:1500–1504.

Brewer, J., Maibohm, C., Jozefowski, L., Bagatolli, L., and Rubahn, H.-G. (2005). A 3D view on free-floating, space-fixed and surface-bound *para*-phenylene nanofibers. *Nanotechnology*, 16:2396–2401.

Chin, A., Vaddiraju, S., Maslov, A., Ning, C., Sunkara, M., and Meyyappan, M. (2006). Near-infrared semiconductor subwavelength-wire lasers. *Appl. Phys. Lett.*, 88:163115-1–163115-3.

Cohen, L., Haracz, R., Cohen, A., and Acquista, C. (1983). Scattering of light from arbitrary oriented finite cylinders. *Appl. Opt.*, 22:742–748.

Duan, X., Huang, Y., Agarwal, R., and Lieber, C. (2003). Single-nanowire electrically driven lasers. *Nature*, 421:241–245.

Duan, X., Huang, Y., Cui, Y., Wang, J., and Lieber, C. (2001). Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices. *Nature*, 409:66–69.

Fiutowski, J., Bordo, V., Jozefowski, L., Madsen, M., and Rubahn, H.-G. (2008). Light scattering from an ordered array of needle-shaped organic nanoaggregates: Evidence for optical mode launching. *Appl. Phys. Lett.*, 92:073302-1–073302-3.

Haracz, R., Cohen, L., and Cohen, A. (1984). Perturbation-theory for scattering from dielectric spheroids and short cylinders. *Appl. Opt.*, 23:436–441.

Huang, M., Mao, S., Feick, H. et al. (2001). Room-temperature ultraviolet nanowire nanolasers. *Science*, 292:1897–1899.

Johnson, J., Choi, H.-J., Knutsen, K., Schaller, R., Yang, P., and Saykally, R. (2002). Single gallium nitride nanowire lasers. *Nat. Mater.*, 1:106–110.

Johnson, J., Yan, H., Yang, P., and Saykally, R. (2003). Optical cavity effects in ZnO nanowire lasers and waveguides. *J. Phys. Chem.*, 107:8816–8828.

Kerker, M. (1969). *The Scattering of Light and Other Electromagnetic Radiation*. Academic Press, New York.

Kuik, F., de Haan, J., and Hovenier, J. (1994). Single scattering of light by circular cylinders. *Appl. Opt.*, 33:4906–4918.

Lord Rayleigh (1881). On the electromagnetic theory of light. *Philos. Mag.*, 12:81–101.

Owen, J., Barber, P., Messinger, B., and Chang, R. (1981). Determination of optical-fiber diameter from resonances in the elastic scattering spectrum. *Opt. Lett.*, 6:272–274.

Quochi, F., Cordella, F., Mura, A., Bongiovanni, G., Balzer, F., and Rubahn, H.-G. (2005). One-dimensional random lasing in a single organic nanofiber. *J. Phys. Chem. B*, 109:21690–21693.

Quochi, F., Cordella, F., Mura, A., Bongiovanni, G., Balzer, F., and Rubahn, H.-G. (2006). Gain amplification and lasing properties of individual organic nanofibers. *Appl. Phys. Lett.*, 88:041106-1–041106-3.

Rao, T. and Barakat, R. (1994). Near-field scattering by a conducting cylinder partially buried in a conducting plane. *Opt. Commun.*, 111:18–25.

Ruppin, R. (1990). Electromagnetic scattering from finite dielectric cylinders. *J. Phys. D Appl. Phys.*, 23:757–763.

Schiek, M., Lützen, A., Koch, R. et al. (2005). Nanofibers from functionalized para-phenylene molecules. *Appl. Phys. Lett.*, 86:153107-1–153107-3.

Shepherd, J. and Holt, A. (1983). The scattering of electromagnetic radiation from finite dielectric circular cylinders. *J. Phys. A Math. Gen.*, 16:651–662.

Snyder, A. and Love, J. (1983). *Optical Waveguide Theory*. Chapman and Hall, London, U.K.

Stratton, J. (1941). *Electromagnetic Theory*. McGraw-Hill, New York.

Thilsing-Hansen, K., Neves-Petersen, M., Petersen, S., Neuendorf, R., Al-Shamery, K., and Rubahn, H.-G. (2005). Luminescence decay of oriented phenylene nanofibers. *Phys. Rev. B*, 72:115213-1–115213-7.

Tong, L., Gattass, R., Ashcom, J. et al. (2003). Subwavelength-diameter silica wires for low-loss optical wave guiding. *Nature*, 426:816–819.

Twersky, V. (1952). Multiple scattering of radiation by an arbitrary planar configuration of parallel cylinders. *J. Acoust. Soc. Am.*, 24:42–46.

Uzunoglu, N., Alexopoulos, N., and Fikioris, J. (1978). Scattering from thin and finite dielectric fibers. *J. Opt. Soc. Am.*, 68:194–197.

Van Bladel, J. (1977). Resonant scattering by dielectric cylinders. *IEEE J. Microwave Opt. Acoust.*, 1:41–50.

Wait, J. (1955). Scattering of a plane wave from a circular dielectric cylinder at oblique incidence. *Can. J. Phys.*, 33:189–195.

Wang, J., Gudiksen, M., Duan, X., Cui, Y., and Lieber, C. (2001). Highly polarized photoluminescence and photodetection from single indium phosphide nanowires. *Science*, 293:1455–1457.

Zimmler, M., Bao, J., Capasso, F., Müller, S., and Ronning, C. (2008). Laser action in nanowires: Observation of the transition from amplified spontaneous emission to laser oscillation. *Appl. Phys. Lett.*, 93:051101-1–051101-3.

31

Biomimetics: Photonic Nanostructures

31.1 Introduction 31-1
31.2 Engineering of Antireflectors 31-1
31.3 Engineering of Iridescent Devices 31-1
31.4 Cell Culture 31-3
31.5 Diatoms and Coccolithophores 31-5
31.6 Iridoviruses 31-7
31.7 The Mechanisms of Natural Engineering and Future Research 31-7
Acknowledgments 31-8
References 31-8

Andrew R. Parker
The Natural History Museum

31.1 Introduction

Three centuries of research, beginning with Hooke and Newton, have revealed a diversity of optical devices at the nanoscale (or at least the submicron scale) in nature.[1] These include structures that cause random scattering, 2D diffraction gratings, 1D multilayer reflectors, and 3D liquid crystals (Figure 31.1a through d). In 2001, the first photonic crystal was identified as such in animals,[2] and since then the scientific effort in this subject has accelerated. Now we know of a variety of 2D- and 3D-photonic crystals in nature (e.g., Figure 31.1e and f), including some designs not encountered previously in physics.

Biomimetics is the extraction of good design from nature. Some optical biomimetic successes have resulted from the use of conventional (and constantly advancing) engineering methods to make direct analogues of the reflectors and antireflectors found in nature. However, recent collaborations between biologists, physicists, engineers, chemists, and material scientists have ventured beyond merely mimicking in the laboratory what happens in nature, leading to a thriving new area of research involving biomimetics via cell culture. Here, the nano-engineering efficiency of living cells is harnessed, and nanostructures such as diatom "shells" can be made for commercial applications via culturing the cells themselves.

31.2 Engineering of Antireflectors

Some insects benefit from antireflective surfaces, either on their eyes to see under low-light conditions, or on their wings to reduce surface reflections in transparent (camouflaged) areas. Antireflective surfaces, therefore, occur on the corneas of moth and butterfly eyes[3] and on the transparent wings of hawkmoths.[4] These consist of nodules, with rounded tips, arranged in a hexagonal array with a periodicity of around 240 nm (Figure 31.2b). Effectively they introduce a gradual refractive index profile at an interface between chitin (a polysaccharide, often embedded in a proteinaceous matrix; r.i. 1.54) and air, and hence reduce reflectivity by a factor of 10.

This "moth-eye structure" was first reproduced at its correct scale by crossing three gratings at 120° using lithographic techniques, and employed as antireflective surfaces on glass windows in Scandinavia.[5] Here, plastic sheets bearing the antireflector were attached to each interior surface of triple-glazed windows using refractive-index-matching glue to provide a significant difference in reflectivity. Today the moth-eye structure can be made extremely accurately using e-beam etching,[6] and is employed commercially on solid plastic and other lenses.

A different form of antireflective device, in the form of a sinusoidal grating of 250 nm periodicity, was discovered on the cornea of a 45-million-year-old fly preserved in amber[7] (Figure 31.2a). This is particularly useful where light is incident at a range of angles (within a single plane, perpendicular to the grating grooves), as demonstrated by a model made in photoresist using lithographic methods.[7] Consequently it has been employed on the surfaces of solar panels, providing a 10% increase in energy capture through reducing the reflected portion of sunlight.[8] Again, this device is embossed onto plastic sheets using holographic techniques.

31.3 Engineering of Iridescent Devices

Many birds, insects (particularly butterflies and beetles), fishes, and lesser-known marine animals display iridescent (changing color with angle) and/or "metallic" colored effects resulting from

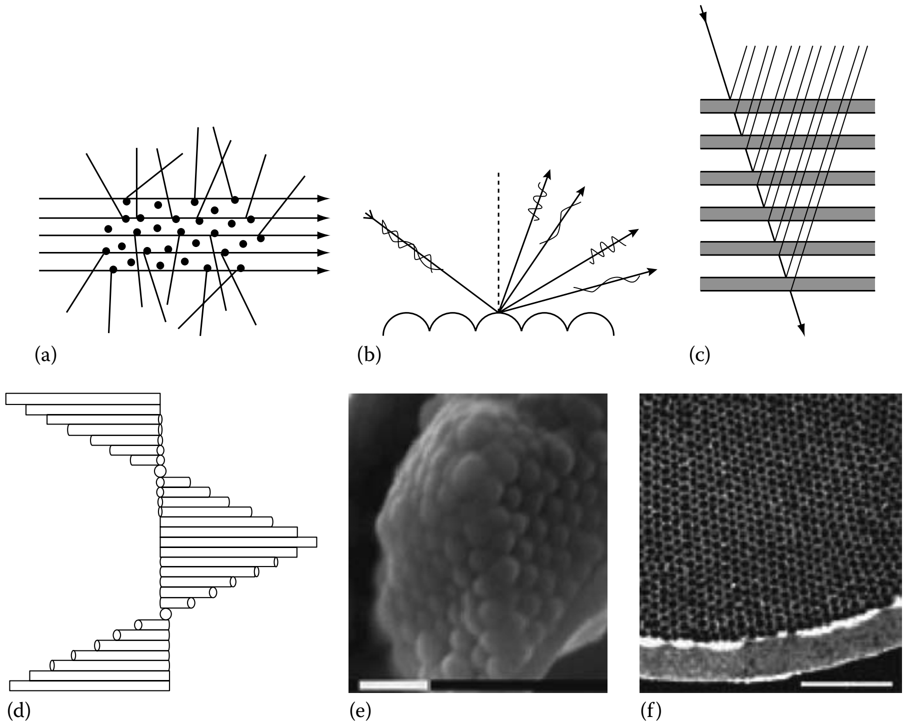

FIGURE 31.1 Summary of the main types of optical reflectors found in nature; a–d where a light ray is (generally) reflected only once within the system (i.e., they adhere to the single scattering, or first Born, approximation), and e and f where each light ray is (generally) reflected multiple times within the system. (a) An irregular array of elements that scatter incident light into random directions. The scattered (or reflected) rays do not superimpose. (b) A diffraction grating, a surface structure, from where light is diffracted into a spectrum or multiple spectra. Each corrugation is about 500 nm wide. Diffracted rays superimpose either constructively or destructively. (c) A multilayer reflector, composed of thin (ca. 100 nm thick) layers of alternating refractive index, where light rays reflected from each interface in the system superimpose either constructively or destructively. Some degree of refraction occurs. (d) A "liquid crystal" composed of nano-fibers arranged in layers, where the nano-fibers of one layer lie parallel to each other yet are orientated slightly differently to those of adjacent layers. Hence spiral patterns can be distinguished within the structure. The height of the section shown here—one "period" of the system—is around 200 nm. (e) Scanning electron micrograph of the "opal" structure—a close-packed array of submicron spheres (a "3D photonic crystal")—found within a single scale of the weevil *Metapocyrtus* sp.; scale bar = 1 μm. (f) Transmission electron micrograph of a section through a hair (neuroseta) of the sea mouse *Aphrodita* sp. (Polychaeta), showing a cross section through a stack of submicron tubes (a "2D photonic crystal"); scale bar = 5 μm.

photonic nanostructures. These appear comparatively brighter than the effects of pigments and often function in animals to attract the attention of a potential mate or to startle a predator. An obvious application for such visually attractive and optically sophisticated devices is within the anticounterfeiting industry. For secrecy reasons, work in this area cannot be described, although devices are sought at different levels of sophistication, from effects that are discernable by the eye to fine-scale optical characteristics (polarization and angular properties, for example) that can be read only by specialized detectors. However, new research aims to exploit these devices in the cosmetics, paint, printing/ink, and clothing industries. They are even being tested in art to provide a sophisticated color-change effect.

Original work on exploiting nature's reflectors involved copying the design but not the size, where reflectors were scaled up to target longer wavelengths. For example, rapid prototyping was employed to manufacture a microwave analogue of a *Morpho* butterfly scale that is suitable for reflection in the 10–30 GHz region. Here the layer thicknesses would be in the order of 1 mm rather than 100 nm as in the butterfly, but the device could be employed as an antenna with broad radiation characteristics, or as an antireflection coating for radar. However, today techniques are available to manufacture nature's reflectors at their true size.

Nanostructures causing iridescence include photonic crystal fibers, opal and inverse opal, and unusually sculpted 3D architectures. Photonic crystals are ordered, often complex, sub-wavelength (nano) lattices that can control the propagation of light at the single-wave scale in the manner that atomic crystals control electrons.[9] Examples include opal (a hexagonal or square array of 250 nm spheres) and inverse opal (a hexagonal array of similar-sized holes in a solid matrix). Hummingbird feather barbs contain variation ultrathin layers with variations in porosity that cause their iridescent effects, due to the alternating nanoporous/fully dense ultrastructure.[10] Such layers have been mimicked using aqueous-based layering techniques.[10] The greatest diversity of 3D architectures can be found in butterfly scales, which can include micro-ribs with nano-ridges, concave multilayered pits, blazed gratings, and randomly punctate nano-layers.[11,12] The cuticle of many beetles contain

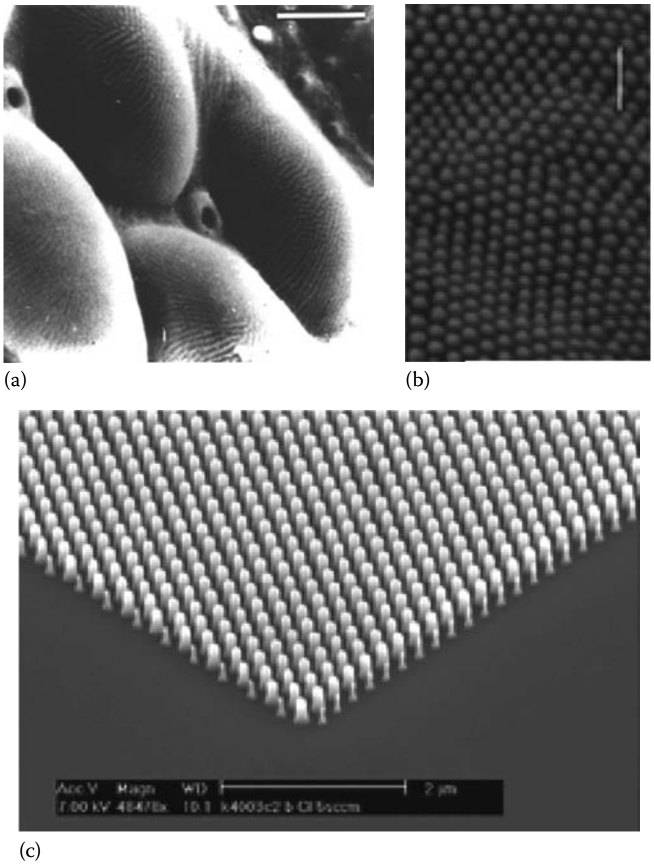

FIGURE 31.2 Scanning electron micrographs of antireflective surfaces. (a) Fly-eye antireflector (ridges on four facets) on a 45-million-year-old dolichopodid fly's eye. (Micrograph from Mierzejewski, P. With permission.) (b) Moth-eye antireflective surfaces. (c) Moth-eye mimic fabricated using ion-beam etching. Scale bars = 3 μm (a), 1 μm (b), 2 μm (c). (Micrograph by Boden, S.A. and Bagnall, D.M., Biomimetic subwavelength surfaces for near-zero reflection sunrise to sunset, *Proceedings 4th World Conference on Photovoltaic Energy, Conversion*, Waikoloa, HI, 2006. With permission.)

structurally chiral films that produce iridescent effects with circular or elliptical polarization properties.[13] These have been replicated in titania for specialized coatings,[13] where a mimetic sample can be compared with the model beetle and an accurate variation in spectra with angle is observed (Figure 31.3). The titania mimic can be nanoengineered for a wide range of resonant wavelengths; the lowest so far is a pitch of 60 nm for a circular Bragg resonance at 220 nm in a Sc_2O_3 film (Ian Hodgkinson, pers. com.).

Biomimetic work on the photonic crystal fibers of the *Aphrodita* sea mouse is underway. The sea mouse contains spines (tubes) with walls packed with smaller tubes of 500 nm, with varying internal diameters (50–400 nm). These provide a bandgap in the red region, and are to be manufactured via an extrusion technique. Larger glass tubes packed together in the proportion of the spine's nanotubes will be heated and pulled through a drawing tower until they reach the correct dimensions. The sea mouse fibre mimics will be tested for standard PCF applications (e.g., in telecommunications) but also for anticounterfeiting structures readable by a detector.

The analogues of the famous blue *Morpho* butterfly (Figure 31.4a) scales have been manufactured.[14,15] Originally, corners were cut. Where the *Morpho* wing contained two layers of scales—one to generate color (a quarter-wave stack) and another above it to scatter the light—the model copied only the principle.[14] The substrate was roughened at the nanoscale, and coated with 80 nm thick layers alternating in refractive index.[14] Therefore the device retained a quarter-wave stack centered in the blue region, but incorporated a degree of randomness to generate scattering. The engineered device closely matched the butterfly wing—the color observed changed only slightly with changing angle over 180°, an effect difficult to achieve and useful for a broad-angle optical filter without dyes.

A new approach to making the 2D "Christmas tree" structure (a vertical, elongated ridge with several layers of 70 nm thick side branches; Figure 31.4b) has been achieved using focused-ion-beam chemical-vapor deposition (FIB-CVD).[15] By combining the lateral growth mode with ion-beam scanning, the Christmas tree structures were made accurately (Figure 31.4c). However, this method is not ideal for low-cost mass production of 2D and 3D nanostructures, and therefore the ion-beam-etched Christmas trees are currently limited to high-cost items including nano- or micron-sized filters (such as "pixels" in a display screen or a filter). Recently further corners have been cut in manufacturing the complex nanostructures found in many butterfly scales, involving the replication of the scales in ZnO, using the scales themselves as templates[16] (Figure 31.4d and e).

31.4 Cell Culture

Sometimes nature's optical nanostructures have such an elaborate architecture at such a small (nano) scale that we simply cannot copy them using current engineering techniques. Additionally, sometimes they can be made as individual reflectors (as for the *Morpho* structure), but the effort is so great that commercial-scale manufacture would never be cost-effective.

An alternative approach to making nature's reflectors is to exploit an aspect other than design—that the animals or plants can make them efficiently. Therefore we can let nature manufacture the devices for us via cell-culture techniques. Animal cells are in the order of 10 μm in size and plant cells up to about 100 μm, and hence suitable for nanostructure production. The success of cell culture depends on the species and on type of cell from that species. Insect cells, for instance, can be cultured at room temperature, whereas an incubator is required for mammalian cells. Cell culture is not a straightforward method, however, since a culture medium must be established to which the cells adhere, before they can be induced to develop to the stage where they make their photonic devices.

The current work in this area centres on butterfly scales. The cells that make the scales are identified in chrysalises, dissected, and plated out. Then the individual cells are separated, kept alive in culture, and prompted to manufacture scales through the addition of growth hormones. Currently we have cultured blue *Morpho* butterfly scales in the lab that have identical optical and structural characteristics to natural scales. The cultured scales

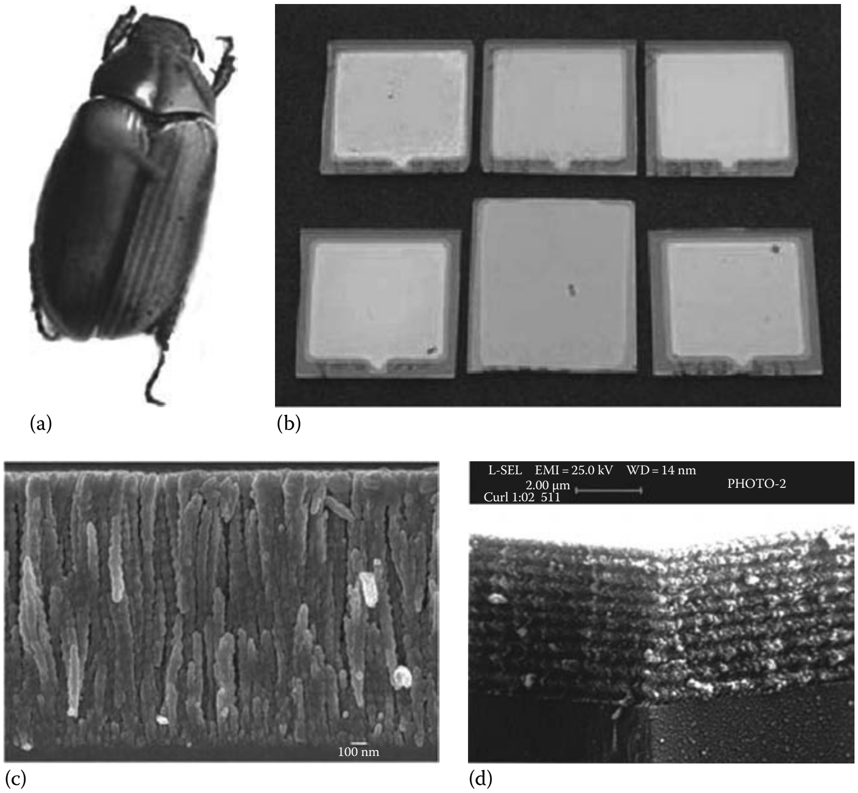

FIGURE 31.3 (a) A Manuka (scarab) beetle with (b) titania mimetic films of slightly different pitches. (c) Scanning electron micrograph of the chiral reflector in the beetle's cuticle. (d) Scanning electron micrograph of the titania mimetic film. (Reproduced from DeSilva, L. et al., *Electromagnetics*, 25, 391, 2005. With permission.)

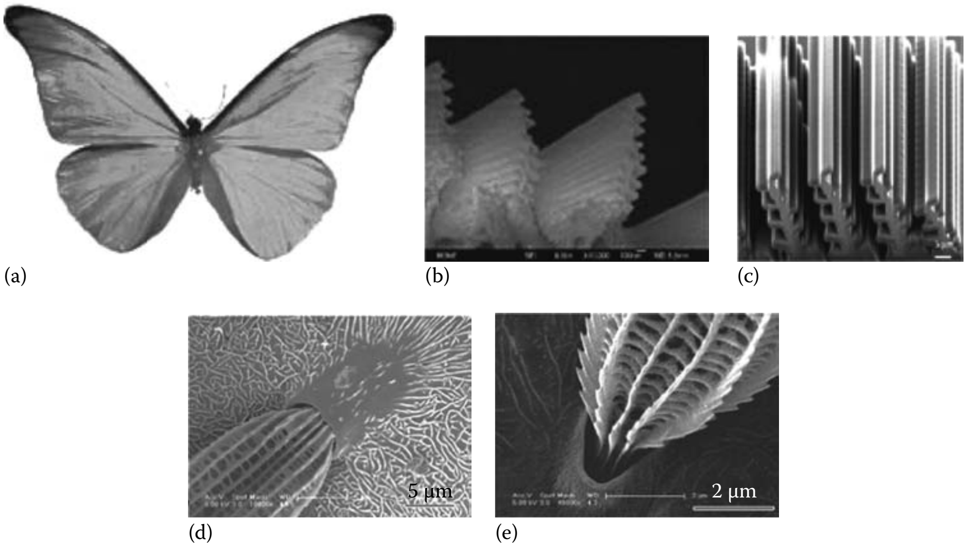

FIGURE 31.4 (a) A *Morpho* butterfly with (b) a scanning electron micrograph of the structure causing the blue reflector in its scales. (c) A scanning electron micrograph of the FIB-CVD-fabricated mimic. A Ga^+ ion beam (beam diameter 7 nm at 0.4 pA; 30 kV), held perpendicular to the surface, was used to etch a precursor of phenanthrene ($C_{14}H_{10}$). Both give a wavelength peak at around 440 nm and at the same angle (30°). (From Watanabe, K. et al., *Jpn. J. Appl. Phys.*, 44, L48, 2005. With permission.) (d) Scanning electron micrograph of the base of a scale of the butterfly *Ideopsis similes*. (e) Scanning electron micrograph of a ZnO replica of the same part of the scale in (d). (Reproduced from Zhang, W. et al., *Bioinspir. Biomim.*, 1, 89, 2006. With permission.)

could be embedded in a polymer or mixed into a paint, where they may float to the surface and self-align. Further work, however, is required to increase the level of scale production and to harvest the scales from laboratory equipment in appropriate ways. A far simpler task emerges where the iridescent organism is single-celled.

31.5 Diatoms and Coccolithophores

Diatoms are unicellular photosynthetic microorganisms. The cell wall is called the frustule and is made of the polysaccharide pectin impregnated with silica. The frustule contains pores (Figure 31.5a through c) and slits that give the protoplasm access to the external environment. There are more than 100,000 different of species of diatoms, generally 20–200 µm in diameter or length, but some can be up to 2 mm long. Diatoms have been proposed to build photonic devices directly in 3D.[17] The biological function of the optical property (Figure 31.5d) is at present unknown, but may affect light collection by the diatom. This type of photonic device can be made in silicon using a deep photochemical-etching technique (initially developed by Lehmann[18]) (e.g., Figure 31.5e). However, there is a new potential here since diatoms carry the added advantage of exponential growth in numbers—each individual can give rise to 100 million descendents in a month.

Unlike most manufacturing processes, diatoms achieve a high degree of complexity and hierarchical structure under mild physiological conditions. Importantly, the size of the pores does not scale with the size of the cell, thus maintaining the pattern. Fuhrmann et al.[17] showed that the presence of these pores in the silica cell wall of the diatom *Coscinodiscus granii* means that the frustule can be regarded as a photonic crystal slab waveguide. Furthermore, they present models to show that light may be coupled into the waveguide and give photonic resonances in the visible spectral range.

The silica surface of the diatom is amenable to simple chemical functionalization (e.g., Figure 31.6a through c). An interesting example of this uses a DNA-modified diatom template for the control of nanoparticle assembly.[19] Gold particles were coated with DNA complementary to that bound to the surface of the diatom. Subsequently, the gold particles were bound to the diatom surface *via* the sequence-specific DNA interaction. Using this method up to seven layers were added showing how a hierarchical structure could be built onto the template.

Porous silicon is known to luminesce in the visible region of the spectrum when irradiated with ultraviolet light.[20] This photoluminescence (PL) emission from the silica skeleton of diatoms was exploited by DeStafano[21] in the production of an optical gas sensor. It was shown that the PL of *Thalassiosira rotula* is strongly dependent on the surrounding environment. Both the optical intensity and peaks are affected by gases and organic vapors. Depending on the electronegativity and polarizing ability, some substances quench the luminescence, while others effectively enhance it. In the presence of the gaseous substances NO_2, acetone, and ethanol, the photoluminescence was quenched. This was because these substances attract electrons from the silica skeleton of the diatoms and hence quench the PL.

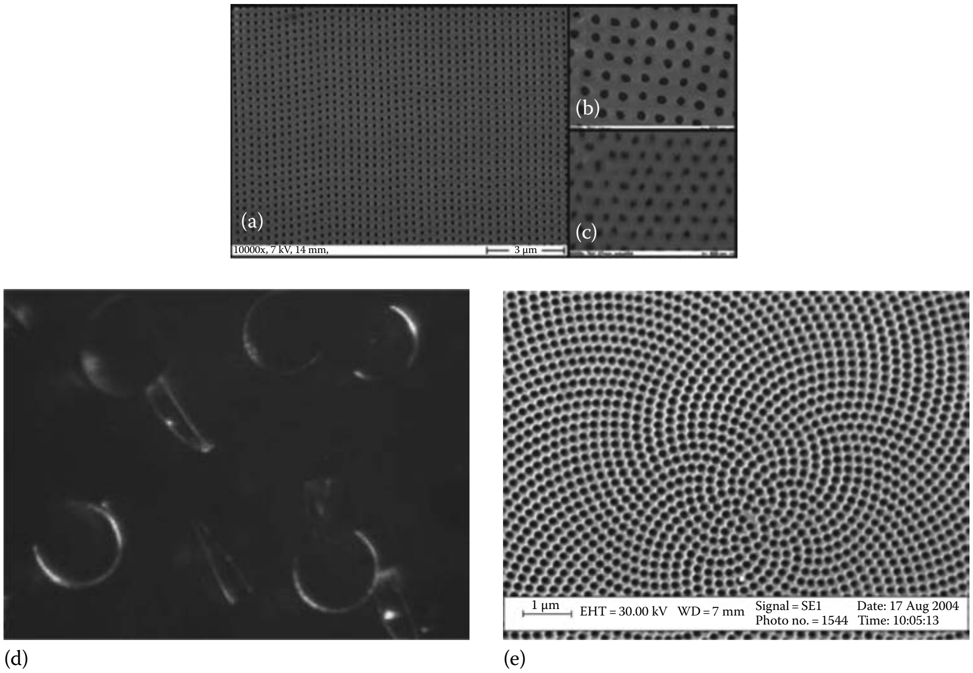

FIGURE 31.5 (a–c) Scanning electron micrographs of the intercalary band of the frustule from two species of diatoms, showing the square array of pores from *C. granii* ((a) and (b)) and the hexagonal arrays of pores from *C. wailesii* (c). These periodic arrays are proposed to act as photonic crystal waveguides. (d) Iridescence of the *C. granii* girdle bands. (e) Southampton University mimic of a diatom frustule (patented for photonic crystal applications); scanning electron micrograph. (Micrograph by Parker, G. With permission.)

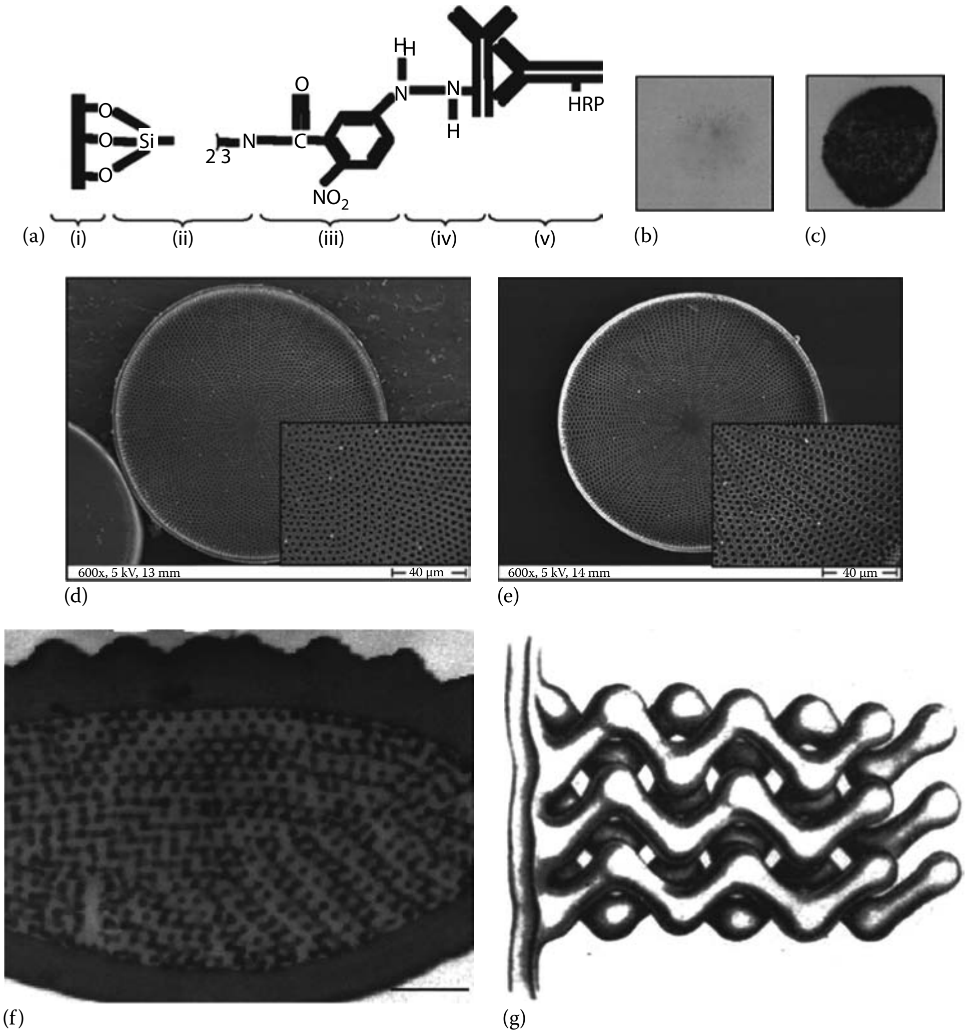

FIGURE 31.6 Modification of natural photonic devices. (a)–(c) Diatom surface modification. The surface of the diatom was silanized, then treated with a heterobifunctional cross-linker, followed by attachment of an antibody via a primary amine group. (a) (i) Diatom exterior surface (ii) APS (iii) ANB-NOS (iv) primary antibody (v) secondary antibody with HRP conjugate. Diatoms treated with primary and secondary antibody with (b) no surface modification (c) after surface modification. (d and e) Scanning electron micrographs showing the pore pattern of the diatom *C. wailesii* (d) and after growth in the presence of nickel sulphate (e). Note the enlargement of pores, and hence change in optical properties, in (e). (f) "Photonic crystal" of the weevil *Metapocyrtus* sp., section through a scale, transmission electron micrograph; scale bar: 1 µm (see Parker[33]). (g) A comparatively enlarged diagrammatic example of cell membrane architecture: tubular christae in mitochondria from the chloride cell of sardine larvae. Evidence suggests that preexisting internal cell structures play a role in the manufacture of natural nanostructures; if these can be altered then so will the nanostructure made by the cell. (From Threadgold, L.T., *The Ultrastructure of the Animal Cell*, Pergamon Press, Oxford, U.K., 1967.)

Nucleophiles, such as xylene and pyridine, which donate electrons, had the opposite effect, and increased PL intensity almost ten times. Both quenching and enhancements were reversible as soon as the atmosphere was replaced by air.

The silica inherent to diatoms does not provide the optimum chemistry/refractive index for many applications. Sandhage et al.[22] have devised an inorganic molecular conversion reaction that preserves the size, shape, and morphology of the diatom while changing its composition. They perfected a gas/silica displacement reaction to convert biologically derived silica structures such as frustules into new compositions. Magnesium was shown to convert SiO_2 diatoms by a vapor phase reaction at 900°C to MgO of identical shape and structure, with a liquid Mg_2Si byproduct. Similarly, when diatoms were exposed to titanium fluoride gas, the titanium displaced the silicon, yielding a diatom structure made up entirely of titanium dioxide; a material used in some commercial solar cells.

An alternative route to silica replacement hijacks that native route for silica deposition in vivo. Rorrer et al.[23] sought to incorporate elements such as germanium into the frustule—a semiconductor material that has interesting properties that could be of value in optoelectronics, photonics, thin film displays, solar cells, and a wide range of electronic devices. Using a two-stage cultivation process, the photosynthetic marine diatom *Nitzschia* *frustulum* was shown to assimilate soluble germanium and fabricate Si-Ge-oxide nanostructured composite materials.

Porous glasses impregnated with organic dye molecules are promising solid media for tunable lasers and nonlinear optical devices, luminescent solar concentrators, gas sensors, and active waveguides. Biogenic porous silica has an open sponge-like structure and its surface is naturally OH-terminated. Hildebrand and Palenik[24] have shown that rhodamine B and 6G are able to stain diatom silica in vivo, and determined that the dye treatment could survive the harsh acid treatment needed to remove the surface organic layer from the silica frustule.

Now attention is beginning to turn additionally to coccolithophores—single-celled marine algae, also abundant in marine environments. Here, the cell secretes calcitic photonic crystal frustules, which, like diatoms, can take a diversity of forms, including complex 3D architectures at the nano- and microscales.

31.6 Iridoviruses

Viruses are infectious particles made up of the viral genome packaged inside a protein capsid. The iridovirus family comprises a diverse array of large (120–300 nm in diameter) viruses with icosahedral symmetry. The viruses replicate in the cytoplasm of insect cells. Within the infected cell the virus particles produce a paracrystalline array that causes Bragg refraction of light. This property has largely been considered esthetic to date, but the research group of Vernon Ward (New Zealand), in collaboration with the Biomaterials laboratory at Wright–Patterson Air Force base, is using iridoviruses to create biophotonic crystals. These can be used for the control of light, with this laboratory undertaking large-scale virus production and purification as well as targeting the manipulation of the surface of iridoviruses for altered crystal properties. These can provide a structural platform for a broad range of optical technologies, ranging from sensors to waveguides.

Virus nanoparticles, specifically *Chilo* and *Wiseana* Invertebrate Iridovirus, have been used as building blocks for iridescent nanoparticle assemblies. Here, virus particles were assembled in vitro, yielding films and monoliths with optical iridescence arising from multiple Bragg scattering from closely packed crystalline structures of the iridovirus. Bulk viral assemblies were prepared by centrifugation followed by the addition of glutaraldehyde, a cross-linking agent. Long-range assemblies were prepared by employing a cell design that forced virus assembly within a confined geometry followed by cross-linking. In addition, virus particles were used as core substrates in the fabrication of metallodielectric nanostructures. These comprise a dielectric core surrounded by a metallic shell. More specifically, a gold shell was assembled around the viral core by attaching small gold nanoparticles to the virus surface using inherent chemical functionality of the protein capsid.[25] These gold nanoparticles then acted as nucleation sites for electroless deposition of gold ions from solution. Such nano-shells could be manufactured in large quantities, and provide cores with a narrower size distribution and smaller diameters (below 80 nm) than currently used for silica. These investigations demonstrated that direct harvesting of biological structures, rather than biochemical modification of protein sequences, is a viable route to create unique, optically active materials.

31.7 The Mechanisms of Natural Engineering and Future Research

Where cell culture is concerned it is enough to know that cells *do* make optical nanostructures, which can be farmed appropriately. However, in the future an alternative may be to emulate the natural engineering processes ourselves, by reacting to the same concentrations of chemicals under the same environmental conditions, and possibly substituting analogous nano- or macro-machinery.

To date, the process best studied is the silica cell wall formation in diatoms. The valves are formed by the controlled precipitation of silica within a specialized membrane vesicle called the silica deposition vesicle (SDV). Once inside the SDV, silicic acid is converted into silica particles, each measuring approximately 50 nm in diameter. These then aggregate to form larger blocks of material. Silica deposition is molded into a pattern by the presence of organelles such as mitochondria spaced at regular intervals along the cytoplasmic side of the SDV.[26] These organelles are thought to physically restrict the targeting of silica from the cytoplasm, to ensure laying down of a correctly patterned structure. This process is very fast, presumably due to optimal reaction conditions for the synthesis of amorphous solid silica. Tight structural control results in the final species-specific, intricate exoskeleton morphology.

The mechanism whereby diatoms use intracellular components to dictate the final pattern of the frustule may provide a route for directed evolution. Alterations in the cytoplasmic morphology of *Skeletonema costatum* have been observed in cells grown in sublethal concentrations of Mercury and Zinc,[27] resulting in swollen organelles, dilated membranes, and vacuolated cytoplasm. Frustule abnormalities have also been reported in *Nitzschia liebethrutti* grown in the presence of mercury and tin.[28] Both metals resulted in a reduction in the length to width ratios of the diatoms, fusion of pores, and a reduction in the number of pores per frustule. These abnormalities were thought to arise from enzyme disruption either at the silica deposition site or at the nuclear level. We grew *Coscinodiscus wailesii* in sublethal concentrations of nickel and observed an increase in the size of the pores (Figure 31.6d and e), and a change in the phospholuminescent properties of the frustule. Here, the diatom can be "made to measure" for distinct applications such as stimuli-specific sensors.

Further, *trans*-Golgi-derived vesicles are known to manufacture the coccolithophore 3D "photonic crystals."[29] So the organelles within the cell appear to have exact control of (photonic) crystal growth ($CaCO_2$ in the coccolithophores) and packing (SiO_3 in the diatoms).[30,31] Indeed, Ghiradella[9] suggested that the employment of preexisting, intracellular structures lay behind the development of some butterfly scales, and Overton[32] reported the action of microtubules and microfibrils during butterfly-scale morphogenesis. Further evidence has been found to suggest that these mechanisms, involving the use of molds and nano-machinery (e.g., Figure 31.6f and g), reoccur with unrelated species, indicating that the basic “eukaryote” (containing a nucleus) cell can make complex photonic nanostructures with minimal genetic mutation.[33] The ultimate goal in the field of optical biomimetics, therefore, could be to replicate such machinery and provide conditions under which, if the correct ingredients are supplied, the optical nanostructures will self-assemble with precision.

For further information on the evolution of optical devices in nature, including those found in fossils, or when they first appeared on earth, see references.[34,35]

Acknowledgments

This work was funded by The Royal Society (University Research Fellowship), The Australian Research Council, European Union Framework 6 grant, and an RCUK Basic Technology grant.

References

1. Parker, A.R. 515 million years of structural colour. *J. Opt. A* **2**, R15–R28 (2000).
2. Parker, A.R., McPhedran, R.C., McKenzie, D.R., Botten, L.C., and Nicorovici, N.-A.P. Aphrodite’s iridescence. *Nature* **409**, 36–37 (2001).
3. Miller, W.H., Moller, A.R., and Bernhard, C.G. The corneal nipple array. In C.G. Bernhard (ed.), *The Functional Organisation of the Compound Eye* (Pergamon Press, Oxford, U.K., 1966), pp. 21–33.
4. Yoshida, A., Motoyama, M., Kosaku, A., and Miyamoto, K. Antireflective nanoprotuberance array in the transparent wing of a hawkmoth *Cephanodes hylas*. *Zool. Sci.* **14**, 737–741 (1997).
5. Gale, M. Diffraction, beauty and commerce. *Phys. World* **2**, 24–28 (1989).
6. Boden, S.A. and Bagnall, D.M. Biomimetic subwavelength surfaces for near-zero reflection sunrise to sunset. *Proceedings of the 4th World Conference on Photovoltaic Energy, Conversion*, Waikoloa, HI (2006).
7. Parker, A.R., Hegedus, Z., and Watts, R.A. Solar-absorber type antireflector on the eye of an Eocene fly (45Ma). *Proc. R. Soc. Lond. B* **265**, 811–815 (1998).
8. Beale, B. Fly eye on the prize. *The Bulletin*, 46–48 (May 25, 1999).
9. Yablonovitch, E. Liquid versus photonic crystals. *Nature* **401**, 539–541 (1999).
10. Cohen, R.E., Zhai, L., Nolte, A., and Rubner, M.F. pH gated porosity transitions of polyelectrolyte multilayers in confined geometries and their applications as tunable Bragg reflectors. *Macromolecules* **37**, 6113 (2004).
11. Ghiradella, H. Structure and development of iridescent butterfly scales: Lattices and laminae. *J. Morph.* **202**, 69–88 (1989).
12. Berthier, S. *Les coulers des papillons ou l’imperative beauté. Proprietes optiques des ailes de papillons* (Springer, Paris, France, 2005), 142 pp.
13. DeSilva, L., Hodgkinson, I., Murray, P., Wu, Q., Arnold, M., Leader, J., and Mcnaughton, A. Natural and nanoengineered chiral reflectors: Structural colour of manuka beetles and titania coatings. *Electromagnetics* **25**, 391–408 (2005).
14. Kinoshita, S., Yoshioka, S., Fujii, Y., and Okamoto, N. Photophysics of structural color in the Morpho butterfly. *Forma* **17**, 103 (2002).
15. Watanabe, K., Hoshino, T., Kanda, K., Haruyama, Y., and Matsui, S. Brilliant blue observation from a *Morpho*-butterfly-scale quasi-structure. *Jpn. J. Appl. Phys.* **44**, L48–L50 (2005).
16. Zhang, W., Zhang, D., Fan, T., Ding, J., Gu, J., Guo, Q., and Ogawa, H. Bio-mimetic zinc oxide replica with structural color using butterfly (*Ideopsis similis*) wings as templates. *Bioinspir. Biomim.* 1, 89 (2006).
17. Fuhrmann, T., Lanwehr, S., El Rharbi-Kucki, M., and Sumper, M. Diatoms as living photonic crystals. *Appl. Phys. B* **78**, 257–260 (2004).
18. Lehmann, V. On the origin of electrochemical oscillations at silicon electrodes. *J. Electrochem. Soc.* **143**, 1313 (1993).
19. Rosi, N.L., Thaxton, C.S., and Mirkin, C.A. Control of nanoparticle assembly by using DNA-modified diatom templates. *Agnew Chem. Int. Ed.* **43**, 5500–5503 (2004).
20. Cullis, A.G., Canham, L.T., and Calcott, P.D.J. The structural and luminescence properties of porous silicon. *J. Appl. Phys.* **82**, 909–965 (1997).
21. De Stefano, L., Rendina, I., De Stefano, M., Bismuto, A., and Maddalena, P. Marine diatoms as optical chemical sensors. *Appl. Phys. Lett.* **87**, 233902 (2005).
22. Sandhage, K.H., Dickerson, M.B., Huseman, P.M., Caranna, M.A., Clifton, J.D., Bull, T.A., Heibel, T.J., Overton, W.R., and Schoenwaelder, M.E.A. Novel, bioclastic route to self-assembled, 3D, chemically tailored meso/nanostructures: Shape-preserving reactive conversion of biosilica (diatom) microshells. *Adv. Mater.* **14**, 429–433 (2002).
23. Rorrer, G.L., Chang, C.H., Liu, S.H., Jeffryes, C., Jiao, J., and Hedberg, J.A. Biosynthesis of silicon-germanium oxide nanocomposites by the marine diatom Nitzschia frustulum. *J. Nanosci. Nanotechnol.* **5**, 41–49 (2004).
24. Hildebrand, M. and Palenik, B. Grant report *Investigation into the Optical Properties of Nanostructured Silica from Diatoms*, La Jolla, CA (2003).
25. Radloff, C., Vaia, R.A., Brunton, J., Bouwer, G.T., and Ward, V.K. Metal nanoshell assembly on a virus bioscaffold. *Nano Lett.* **5**, 1187–1191 (2005).
26. Schmid, A.M.M. Aspects of morphogenesis and function of diatom cell walls with implications for taxonomy. *Protoplasma* **181**, 43–60 (1994).
27. Smith, M.A. The effect of heavy metals on the cytoplasmic fine structure of Skeletonema costatum (Bacillariophyta). *Protoplasma* **116**, 14–23 (1983).

28. Saboski, E. Effects of mercury and tin on frustular ultrastructure of the marine diatom *Nitzschia liebethrutti. Water, Air Soil Pollut.* **8**, 461–466 (1977).
29. Corstjens, P.L.A.M. and Gonzales, E.L. Effects of nitrogen and phosphorus availability on the expression of the coccolith-vesicle v-ATPase (subunit C) of Pleurochrysis (Haptophyta). *J. Phycol.* **40**, 82–87 (2004).
30. Klaveness, D. and Paasche, E. Physiology of coccolithophorids. In: *Biochemistry and Physiology of Protozoa*, 2nd edn., vol. 1 (Academic Press, New York, 1979), pp. 191–213.
31. Klaveness, D. and Guillard, R.R.L. The requirement for silicon in *Synura petersenii* (Chrysophyceae). *J. Phycol.* **11**, 349–355 (1975).
32. Overton, J. Microtubules and microfibrils in morphogenesis of the scale cells of *Ephestia kuhniella. J. Cell Biol.* **29**, 293–305 (1966).
33. Parker, A.R. Conservative photonic crystals imply indirect transcription from genotype to phenotype. *Recent Res. Develop. Entomol.* **5**, 1–10 (2006).
34. Parker, A.R. *In the Blink of an Eye* (Simon & Schuster, London, U.K., 2003), 316 pp.
35. Parker, A.R. A geological history of reflecting optics. *J. R. Soc. Lond. Interface* **2**, 1–17 (2005).

V

Nanophotonic Devices

32 **Photon Localization at the Nanoscale** *Kiyoshi Kobayashi* **32**-1
Introduction • Background • Dressing Mechanism and Spatial Localization of Photons • Summary and Future Perspective • Acknowledgments • References

33 **Operations in Nanophotonics** *Suguru Sangu and Kiyoshi Kobayashi* **33**-1
Introduction • Dissipation-Controlled Nanophotonic Devices • Nanophotonic Devices Using Spatial Symmetries • Summary • Acknowledgments • References

34 **System Architectures for Nanophotonics** *Makoto Naruse* **34**-1
Introduction • System Architectures Based on Optical Excitation Transfer • Hierarchical Architectures in Nanophotonics • Summary • References

35 **Nanophotonics for Device Operation and Fabrication** *Tadashi Kawazoe and Motoichi Ohtsu* **35**-1
Introduction • Excitation Energy Transfer in Nanophotonic Devices • Device Operation • Nanophotonics Fabrication • Summary • References

36 **Nanophotonic Device Materials** *Takashi Yatsui and Wataru Nomura* **36**-1
Introduction • Nanophotonic Devices Based on Quantum Dots • Nanophotonic AND-Gate Device Using ZnO Nanorod Double-Quantum-Well Structures • References

37 **Waveguides for Nanophotonics** *Jan Valenta, Tomáš Ostatnický, and Ivan Pelant* **37**-1
Introduction • Fabrication of Planar and Rib Waveguides • Experimental Techniques • Main Experimental Observations in Active Si-nc Waveguides • Theoretical Description of Active Lossy Waveguides • Application of Active Nanocrystalline Waveguides • Conclusions • Acknowledgments • References

38 **Biomolecular Neuronet Devices** *Grigory E. Adamov and Evgeny P. Grebennikov* **38**-1
Introduction • Some Techniques of Neuro-Molecular and Molecular Information Processing Using Bacteriorhodopsin (Basic Processes, Constructions, Technology) • Nanostructuring of Bacteriorhodopsin-Containing Molecular Media • Summary • References

32

Photon Localization at the Nanoscale

32.1 Introduction 32-1
32.2 Background 32-2
Difficulty in Defining a Wave Function of a Photon in Real Space • Effective Spatial Wave Function of a Single Photon • Canonical Field Quantization: Mode Functions, Field Operators, and Quantum States
32.3 Dressing Mechanism and Spatial Localization of Photons 32-5
Virtual Photon Cloud Surrounding a Neutral Source (in Ground State or Excited State) in QED • Electromagnetic Field Correlations and Intermolecular Interactions between Molecules in Either Ground or Excited States • Effective Near-Field Optical Interaction between Nanomaterials Disconnected but Closely Separated • Localization of a Photon Dressed by Matter Excitation in Nanomaterials at the Nanoscale
32.4 Summary and Future Perspective 32-17
Acknowledgments 32-17
References 32-17

Kiyoshi Kobayashi
University of Tokyo
Japan Science and Technology
University of Yamanashi

32.1 Introduction

You might think that the great success of quantum electrodynamics (QED) would settle the debate on the nature of light to provide a clear view of its behavior, where the photon is regarded as the unit of excitation associated with a quantized mode of the electromagnetic (radiation) field. However, Heisenberg's uncertainty principle tells us that a state of definite momentum, energy, and polarization associated with a plane wave used as a basis function of quantization must be completely indefinite in space and time. It suggests the difficulty of spatial localization of a photon as a particle. In fact, Newton and Wigner showed that a free photon, as a massless particle with spin 1, has no localized states on the basis of natural invariance requirements that localized states for which operators of the Lorentz group apply should be orthogonal to the undisplaced localized states, after a translation (Newton and Wigner 1949). According to them, one can obtain a general expression for a position operator for massive particles and for massless particles of spin 0 or 1/2, not for massless particles with finite spin, which indicates that there is no probability density for the position of the photon, and thus a position-representation wave function cannot be consistently introduced. It has also been shown that photons are not localizable, on the basis of imprimitive representations of the Euclidean group (Wightman 1962). It is now believed that photons are only weakly localizable, although single-photon states with arbitrarily fast asymptotic falloff of energy density exist, and that a lack of strict localizability is directly related to the absence of a position operator for a photon in free space and a position-representation photon wave function (Hawton 1999).

On the other hand, several authors have claimed that a minimum modification of the naive route leads to a wave-function description of a photon, even though the probability density for the position of a photon and a position-representation wave function cannot be consistently introduced. For example, it has been shown to be possible to introduce a position-representation wave function $\vec{\psi}(\vec{r}, t)$ for a photon, which is the expectation value of the photon energy in a region $d\vec{r}$ about $\vec{r}$ (Sipe 1995, Scully and Zubairy 1997, Hawton 1999, Roychoudhuri and Roy 2003). Mandel et al., from another viewpoint, have found that a photon wave function as a probability amplitude is possible in a coarse-grained volume whose linear dimensions are larger than the photon wavelength (Mandel 1966, Mandel and Wolf 1995, Inagaki 1998). The localization of the photon energy density and photodetection rates having an exponential or arbitrarily fast asymptotic falloff have been discussed, as well as causality (Hegerfeldt 1974, Pike and Sarkar 1987, Hellwarth and Nouchi 1996, Adlard et al. 1997, Bialynicki-Birula 1998, Keller 2005).

When the interactions with matter were considered, different views opened up. Dressed states and operators, dressed and half-dressed sources in nonrelativistic QED, electromagnetic field correlations, and intermolecular interactions between molecules in either ground or excited states have been discussed (Compagno et al. 1988, 1995, Cohen-Tannoudji et al. 1989, 1992, Power and Thirunamachandran 1993), focusing on the fact that a bare source interacting with a quantum field is surrounded by a cloud of virtual particles. It has been shown that the dressing of the source, or virtual cloud effects, can be detected by a test body (detector) located close to the source. Carniglia and Mandel (1971) proposed a complete basis for electromagnetic fields interacting with a material of refractive index n filled in a half space separated by vacuum in order to quantize the source fields, while Inoue and Hori (2001) discussed the detector modes and the behavior of a photon–atom interacting system near the material surface. Kobayashi et al. (2001) focused on the environmental effects on a nanomaterial interacting with photons and obtained a near-field optical interaction as an effective interaction between nanomaterials electronically disconnected, but closely located, in order to detect the cloud of virtual photons. They have also applied it to a discussion of nanophotonic devices (Sangu et al. 2004). Focusing on the photon degrees of freedom, on the other hand, photon hopping has been employed to discuss a photon–material interacting system (John and Quang 1995, Suzuura et al. 1996, Shojiguchi et al. 2003), and a photon dressing by material excitations has been recently discussed by using the photon-hopping model in real space and a quasiparticle model (Kobayashi et al. 2008).

Focusing on light–matter interactions at the nanoscale, we discuss a near-field optical interaction between nanomaterials surrounded by a macroscopic system, a dressing mechanism, and spatial localization of photons in this article, which is organized as follows. Section 32.2 is devoted to background issues for photon localization, in particular, the difficulty of the definition, effective methods, and free-field quantization. In Section 32.3 we discuss a dressing mechanism of photons and their localization in space, including virtual clouds of photons, electromagnetic field correlations and intermolecular interactions, effective near-field optical interactions, and phonons' effects on photon localization. Finally, a summary and a future outlook are presented.

32.2 Background

32.2.1 Difficulty in Defining a Wave Function of a Photon in Real Space

Since there is no probability density for the photon, and thus a position-representation wave function in free space cannot be consistently defined, one has to follow quantum electrodynamics (QED), that is, to redefine one- and a few-photon wave functions in a physically meaningful way in order to obtain fruitful insights into the photon–matter interacting system at the nanoscale. In the following subsections, we will give an overview of both approaches, after pointing out the difficulties involved with a position-representation wave function of a photon in free space.

We begin with Maxwell's equations in vacuum:

$$\nabla \times \vec{E}(\vec{r},t) = -\frac{1}{c}\frac{\partial \vec{B}(\vec{r},t)}{\partial t}, \tag{32.1a}$$

$$\nabla \times \vec{B}(\vec{r},t) = \frac{1}{c}\frac{\partial \vec{E}(\vec{r},t)}{\partial t}, \tag{32.1b}$$

where
c is the speed of light
$\vec{E}(\vec{r},t)$ and $\vec{B}(\vec{r},t)$ are electric and magnetic fields, respectively

Let us now define $\vec{\Phi}_\pm(\vec{r},t)$ as $\vec{\Phi}_\pm(\vec{r},t) = \vec{E}(\vec{r},t) \pm i\vec{B}(\vec{r},t)$; then, it follows from Maxwell's equations that $\vec{\Phi}_\pm(\vec{r},t)$ should satisfy

$$i\hbar\frac{\partial \vec{\Phi}_\pm(\vec{r},t)}{\partial t} = \pm c\hbar\nabla \times \vec{\Phi}_\pm(\vec{r},t). \tag{32.2}$$

Using the Fourier transform of these functions, we introduce the vectors $\vec{\gamma}_\pm(\vec{p},t)$ as

$$\vec{\Phi}_\pm(\vec{r},t) = \int \frac{d^3p}{(2\pi\hbar)^{3/2}}\vec{\gamma}_\pm(\vec{p},t)e^{i\vec{p}\cdot\vec{r}/\hbar}, \tag{32.3}$$

which is associated with $\gamma_\pm(\vec{p},t)$ as $\vec{\gamma}_\pm(\vec{p},t) = \gamma_\pm(\vec{p},t)\vec{e}_\pm(p)$ with

$$\vec{e}_\pm(\hat{p}) = \mp\frac{1}{\sqrt{2}}[\vec{e}_1(\hat{p}) \pm i\vec{e}_2(\hat{p})], \tag{32.4a}$$

where two unit vectors $\vec{e}_1(\hat{p})$ and $\vec{e}_2(\hat{p})$ are defined such that the unit vectors

$$\left[\vec{e}_1(\hat{p}), \vec{e}_2(\hat{p}), \hat{p} = \vec{p}/|\vec{p}|\right] \tag{32.4b}$$

form a right-handed triad. Then, it is easy to verify that $\gamma_\pm(\vec{p},t)$ satisfy a Schrödinger-like equation in the momentum representation

$$i\hbar\frac{\partial \gamma_\pm(\vec{p},t)}{\partial t} = cp\gamma_\pm(\vec{p},t), \tag{32.5}$$

which indicates that $\gamma_\pm(\vec{p},t)$ are probability amplitudes for photons of momentum $\vec{p}$, energy $E = cp$ with $p = |\vec{p}|$, and positive/negative helicity. Here note that

$$\vec{\gamma}^*_+(\vec{p},t)\cdot\vec{\gamma}_+(\vec{p},t)d\vec{p} = \vec{\gamma}^*_-(\vec{p}\cdot t)\cdot\vec{\gamma}_-(\vec{p},t)d\vec{p} \tag{32.6a}$$

and

$$\vec{\gamma}^*_+(\vec{p},t)\cdot\vec{\gamma}_-(\vec{p},t) = 0 \tag{32.6b}$$

show the probability of detecting a photon of positive helicity and momentum $\vec{p}$ between $\vec{p}$ and $\vec{p} + d\vec{p}$, and likewise for a photon of negative helicity. Then the normalization condition is

$$\int\left[\vec{\gamma}_+^*(\vec{p},t)\cdot\vec{\gamma}_+(\vec{p},t)+\vec{\gamma}_-^*(\vec{p},t)\right]d^3p=1, \qquad (32.7)$$

which leads to

$$\int\left[\vec{\Phi}_+^*(\vec{r},t)\cdot\vec{\Phi}_+(\vec{r},t)+\vec{\Phi}_-^*(\vec{r},t)\cdot\vec{\Phi}_-(\vec{r},t)\right]d^3r=1, \qquad (32.8)$$

and it follows from the dynamical equations that

$$\vec{\Phi}_\pm(\vec{r},t)=\int\frac{d^3p}{(2\pi\hbar)^{3/2}}\vec{\gamma}_\pm(\vec{p},0)e^{-icpt/\hbar}e^{i\vec{p}\cdot\vec{r}/\hbar}. \qquad (32.9)$$

Here we note that the sum of $\vec{\Phi}_\pm(\vec{r},t)$ might be regarded as the position-representation wave function of a photon, but it cannot be regarded as such because Newton and Wigner and also Wightman have shown that the photon, being a massless particle, is not localizable in free space, and that there does not exist a probability amplitude and density for the position of the photon in the usual sense (see also (32.15a) and (32.15b)) (Newton and Wigner 1949, Wightman 1962).

32.2.2 Effective Spatial Wave Function of a Single Photon

In order to avoid the difficulties mentioned above, Mandel defined an operator representing the number of photons in a volume *V* as the integral over *V* of a so-called "detection operator," which led a simple formula for the probability that *n* photons are present in *V*, when the linear dimensions of *V* are larger than the wavelength of light used (Mandel 1966).

Sipe took another approach to this issue, seeking a probability amplitude $\vec{\Psi}(\vec{r},t)$ for the photon energy to be detected about $d\vec{r}$ of $\vec{r}$ (Sipe 1995). Assuming that the integral of $\vec{\Psi}^*(\vec{r},t)\cdot\vec{\Psi}(\vec{r},t)d\vec{r}$ over all space is proportional to the photon energy, we normalize it as follows:

$$\int\vec{\Psi}^*(\vec{r},t)\cdot\vec{\Psi}(\vec{r},t)d^3r=\int cp\Big[\vec{\gamma}_+^*(\vec{p},t)\cdot\vec{\gamma}_+^*(\vec{p},t)$$
$$+\vec{\gamma}_-^*(\vec{p},t)\cdot\vec{\gamma}_-^*(\vec{p},t)\Big]d^3p. \qquad (32.10)$$

It is easily shown that if $\vec{\Psi}(\vec{r},t)$ is set as

$$\vec{\Psi}(\vec{r},t)=\vec{\Psi}_+(\vec{r},t)+\vec{\Psi}_-(\vec{r},t), \qquad (32.11a)$$

$$\vec{\Psi}_\pm(\vec{r},t)=\int\frac{\sqrt{cp}}{(2\pi\hbar)^{3/2}}\vec{\gamma}_\pm(\vec{p},t)e^{i\vec{p}\cdot\vec{r}/\hbar}d^3p, \qquad (32.11b)$$

the normalization condition is satisfied. Here note that

$$\vec{\gamma}_+^*(\vec{p},t)\cdot\vec{\gamma}_-(\vec{p},t)=0=\vec{\gamma}_-^*(\vec{p},t)\cdot\vec{\gamma}_+(\vec{p},t) \qquad (32.12)$$

and

$$i\hbar\frac{\partial\vec{\Psi}_\pm(\vec{r},t)}{\partial t}=\pm c\hbar\nabla\times\vec{\Psi}_\pm(\vec{r},t). \qquad (32.13)$$

At the same time, $\vec{\Psi}_\pm(\vec{r},t)$ should be chosen to satisfy an initial condition given by

$$\vec{\Psi}_\pm(\vec{r},t)=\int\frac{\sqrt{cp}}{(2\pi\hbar)^{3/2}}\vec{\gamma}_\pm(\vec{p},0)e^{-icpt/\hbar}e^{i\vec{p}\cdot\vec{r}/\hbar}d^3p. \qquad (32.14)$$

Since $\vec{\gamma}_\pm(\vec{p},t)$ and $\vec{\Psi}_\pm(\vec{p},\vec{r})$ are not the Fourier transform pairs, the arguments of the photon momentum $\vec{p}$ and the position $\vec{r}$ associated with the photon energy are not conjugate variables. When $\vec{\Phi}_\pm(\vec{r},t)$ and $\vec{\Psi}_\pm(\vec{r},t)$ are related by a local kernel, it can usually be regarded as a particle. However, $\vec{\Phi}_\pm(\vec{r},t)$, the Fourier transform of $\vec{\gamma}_\pm(\vec{p},t)$, is not a reasonable candidate for the position-representation wave function of the photon because of the following relation between $\vec{\Phi}_\pm(\vec{r},t)$ and $\vec{\Psi}_\pm(\vec{r},t)$

$$\vec{\Phi}_\pm(\vec{r},t)=\int w(\vec{r}-\vec{r}\,')\vec{\Psi}_\pm(\vec{r}\,',t)d^3r' \qquad (32.15a)$$

with the nonlocal kernel

$$w(\vec{r}-\vec{r}\,')=\int\frac{1}{(2\pi)^3}\frac{e^{i\vec{k}\cdot(\vec{r}-\vec{r}\,')}}{\sqrt{\hbar ck}}d^3k. \qquad (32.15b)$$

Nevertheless, $\vec{\Psi}(\vec{r},t)$ might be meaningful to describe the dynamics of a photon such as a spontaneous emission from an atom and the inverse process, or at least we can detect a photon, within the range of a detector's precision by placing a detector like an atom close to the source. In other words, it indicates that light–matter interactions near the source play an important role and should be treated consistently.

32.2.3 Canonical Field Quantization: Mode Functions, Field Operators, and Quantum States

It is natural to follow the canonical quantization of the electromagnetic field as a starting point for a discussion of light–matter interactions at the nanoscale. Since a lot of famous textbooks on

QED or quantum field theory have been published, we follow the essence of the theory and restrict ourselves to the free field that is free from charges and currents and whose scalar potential can be set to zero (Sakurai 1967, Roychoudhuri and Roy 2003). In the Coulomb gauge with $\nabla \cdot \vec{A}(\vec{r}, t) = 0$ for the vector potential $\vec{A}(\vec{r}, t)$, three basic equations we work with for the free-field case are

$$\vec{B}(\vec{r}, t) = \nabla \times \vec{A}(\vec{r}, t), \tag{32.16a}$$

$$\vec{E}(\vec{r}, t) = -\frac{1}{c}\frac{\partial \vec{A}(\vec{r}, t)}{\partial t}, \tag{32.16b}$$

$$\nabla^2 \vec{A}(\vec{r}, t) - \frac{1}{c^2}\frac{\partial^2 \vec{A}(\vec{r}, t)}{\partial t^2} = 0. \tag{32.16c}$$

We expand $\vec{A}(\vec{r}, t)$ into a complete set of mode functions $\vec{u}_{\vec{k},a}(\vec{r})$ defined by the Helmholtz equation

$$(\nabla^2 + \vec{k}^2)\vec{u}_{\vec{k},a}(\vec{r}) = 0, \tag{32.17}$$

and the boundary conditions set by the shape of a virtual cavity, for example, a box taken to be a cube of side $L = V^{1/3}$. It follows from the Coulomb gauge that the direction of $\vec{u}_{\vec{k},a}(\vec{r})$ has to be orthogonal to the wave vector $\vec{k}$, and thus there are two polarization degrees of freedom indicated by the index α. Note that the mode functions become more complicated, or even impossible for more sophisticated cavity shapes. The mode functions satisfy the orthonormality condition

$$\frac{1}{V}\int d^3r \vec{u}_\ell^{\star}(\vec{r}) \cdot \vec{u}_{\ell'}(\vec{r}) = \delta_{\ell\ell'}, \tag{32.18}$$

where we have for simplicity combined the three components of the wave vector $\vec{k}$ and the polarization index α to one index ℓ. Using the normalization constant A_t and the time-dependent amplitude $q_t(t)$, we have

$$\vec{A}(\vec{r}, t) = \sum_\ell \left[A_\ell q_\ell(t)\vec{u}_\ell(\vec{r}) + A_\ell^{\star} q_\ell^{\star}(t)\vec{u}_\ell^{\star}(\vec{r}) \right], \tag{32.19}$$

where $q_t(t)$ follows from the differential equation of a harmonic oscillator of frequency $\Omega_\ell \equiv c|\vec{k}_\ell|$:

$$\frac{d^2 q_\ell(t)}{dt^2} + \Omega_\ell^2 q_\ell(t) = 0. \tag{32.20}$$

The Hamiltonian of the field is

$$H = \frac{1}{8\pi}\int d^3r(\vec{E}^2 + \vec{B}^2) = \frac{1}{8\pi}\int d^3r \left[\left(\frac{1}{c}\frac{\partial \vec{A}}{\partial t} \right)^2 + (\nabla \times \vec{A})^2 \right]. \tag{32.21}$$

Noticing a typical term for the $\vec{E}^2$ integration

$$\int \left(\frac{1}{c}\frac{dq_\ell}{dt} \right)\left(\frac{1}{c}\frac{dq_{\ell'}^{\star}}{dt} \right)\vec{u}_\ell(\vec{r}) \cdot \vec{u}_{\ell'}^{\star}(\vec{r})d^3r = \left(\frac{\Omega_\ell}{c} \right) V|q_\ell|^2 \delta_{\ell\ell'} \tag{32.22}$$

and for the $\vec{B}^2$ integration

$$\begin{aligned} \int (\nabla \times \vec{u}_\ell) \cdot (\nabla \times \vec{u}_{\ell'}^{\star})d^3r &= \int \nabla \cdot (\vec{u}_\ell \times \nabla \times \vec{u}_{\ell'}^{\star})d^3r + \int \vec{u}_\ell \cdot \left[\nabla \times \nabla \times \vec{u}_{\ell'}^{\star} \right] d^3r \\ &= \int \vec{u}_\ell \cdot \left[\nabla(\nabla \cdot \vec{u}_{\ell'}^{\star}) - \nabla^2 \vec{u}_{\ell'}^{\star} \right] d^3r = -\int \vec{u}_\ell \cdot \nabla^2 \vec{u}_{\ell'}^{\star} d^3r \\ &= \left(\frac{\Omega_\ell}{c} \right)^2 V\delta_{\ell\ell'} \end{aligned} \tag{32.23}$$

we obtain

$$H = \frac{V}{2\pi}\sum_\ell \left(\frac{\Omega_\ell}{c} \right)^2 (A_\ell q_\ell)^{\star}(A_\ell q_\ell). \tag{32.24}$$

If we define

$$Q_\ell = \frac{1}{c}(q_\ell + q_\ell^{\star}), \tag{32.25a}$$

$$P_\ell = -i\frac{\Omega_\ell}{c}(q_\ell - q_\ell^{\star}), \quad A_\ell = \sqrt{\frac{4\pi}{V}}, \tag{32.25b}$$

the Hamiltonian can be expressed in terms of a collection of independent and uncoupled harmonic oscillators as

$$H = \sum_\ell \frac{1}{2}(P_\ell^2 + \Omega_\ell^2 Q_\ell^2). \tag{32.26}$$

Here, P_ℓ and Q_ℓ are seen to be canonical variables:

$$\frac{dP_\ell}{dt} = -\frac{\partial H}{\partial Q_\ell}, \tag{32.27a}$$

$$\frac{dQ_\ell}{dt} = \frac{\partial H}{\partial P_\ell}. \tag{32.27b}$$

The natural method to quantize the field is to replace the variable q_ℓ and its conjugate momentum $p_\ell \equiv dq_\ell/dt$ by operators $\hat{q}_\ell$ and $\hat{p}_\ell$ that satisfy the commutation relations $[\hat{q}_\ell, \hat{p}_\ell] = i\hbar\delta_{\ell\ell'}$, or $[Q_\ell, P_{\ell'}] = i\hbar\delta_{\ell\ell'}$. Then, the Hamiltonian operator for the quantized field can be written as follows:

$$\hat{H} = \sum_{\ell} \frac{1}{2}(\hat{P}_{\ell}^2 + \Omega_{\ell}^2 \hat{Q}_{\ell}^2). \tag{32.28}$$

Next we consider linear combinations of $\hat{P}_{\ell}$ and $\hat{Q}_{\ell}$, given by

$$\hat{a}_{\ell} = \sqrt{\frac{\Omega_{\ell}}{2h}}\left(\hat{Q}_{\ell} + \frac{i}{\Omega_{\ell}}\hat{P}_{\ell}\right), \tag{32.29a}$$

$$\hat{a}_{\ell}^{\dagger} = \sqrt{\frac{\Omega_{\ell}}{2h}}\left(\hat{Q}_{\ell} + \frac{i}{\Omega_{\ell}}\hat{P}_{\ell}\right), \tag{32.29b}$$

where we insert a factor to make $\hat{a}_{\ell}$ and $\hat{a}_{\ell}^{\dagger}$ dimensionless and thus satisfy the following boson commutation relations:

$$\left[\hat{a}_{\ell}, \hat{a}_{\ell'}\right] = -\frac{i}{2h}[\hat{Q}_{\ell}, \hat{P}_{\ell'}] + \frac{i}{2h}[P_{\ell}, Q_{\ell'}] = \delta_{\ell\ell'}. \tag{32.30}$$

Taking care of the order of $\hat{a}_{\ell}$ and $\hat{a}_{\ell'}^{\dagger}$, we obtain the Hamiltonian operator of the electromagnetic field as

$$\hat{H} = \sum_{\ell} \hbar\Omega_{\ell}\left(\hat{n}_{\ell} + \frac{1}{2}\right), \tag{32.31}$$

where $\hat{n}_{\ell} \equiv \hat{a}_{\ell}^{\dagger}\hat{a}_{\ell}$ denotes the number operator. The contribution 1/2 arises from the commutation relations and results in the familiar zero-point energy. Since there are infinitely many modes, the zero-point energy becomes infinite, but in general we drop this contribution by shifting the vacuum energy, which does not influence the dynamics. The quantization procedure is completed by writing the vector potential in terms of $\hat{a}_{\ell}$ and $\hat{a}_{\ell}^{\dagger}$ as a field operator

$$\hat{\vec{A}}(\vec{r}, t) = \sum_{\ell} \sqrt{\frac{2\pi\hbar c^2}{V\Omega_{\ell}}}\left(\hat{a}_{\ell}\vec{u}_{\ell} + \hat{a}_{\ell}^{\dagger}\vec{u}_{\ell}^{*}\right). \tag{32.32}$$

Thus, $\hat{a}_{\ell}^{\dagger}$ is called the creation operator for a photon specified in ℓ as corresponding to the quantum-mechanical excitations of the electromagnetic field, while $\hat{a}_{\ell}$ is interpreted as the annihilation operator for a photon in state ℓ. It is important to note that the $\vec{r}$ and t that appear in the quantized field $\hat{\vec{A}}(\vec{r}, t)$ are not quantum-mechanical variables but just parameters on which the field operator depends, and, in particular, $\vec{r}$ and t should not be regarded as the space–time coordinates of the photon.

When we adopt a linearly polarized plane wave as the mode function

$$\vec{u}_{\ell}(\vec{r}) = \vec{\varepsilon}_{\vec{k}\alpha} e^{i\vec{k}\cdot\vec{r}} \tag{32.33}$$

with the polarization vector $\vec{\varepsilon}_{\vec{k}\alpha}$, the energy and momentum of the photon are $\hbar\Omega_{\vec{k}} = \hbar|\vec{k}|c$ and $\hbar\vec{k}$, respectively. Therefore, the mass of the photon is zero. In addition, since $\vec{\varepsilon}_{\vec{k}\alpha}$ transforms like a vector, the general theory of angular momentum encourages us to associate with it one unit of angular momentum, which means that the photon has one unit of spin angular momentum.

The field operators described above operate on quantum state vectors, and quantum states $|\Psi\rangle$ of the electromagnetic field are, in general, multimode states that involve quantum states $|\psi_{\ell}\rangle$ for each mode ℓ. One of the useful quantum states is photon number states denoted by $|n_{\ell}\rangle$, which are eigenstates of the number operator $\hat{n}_{\ell}$

$$\hat{n}_{\ell}|n_{\ell}\rangle = n_{\ell}|n_{\ell}\rangle \tag{32.34}$$

with integer eigenvalues n_{ℓ}. At the same time, $|n_{\ell}\rangle$ are eigenstates of the Hamiltonian with eigenenergy $n_{\ell}\hbar\Omega_{\ell}$, that is, n_{ℓ} times the fundamental unit $\hbar\Omega_{\ell}$. It should be noted that n_{ℓ} quanta of energy $\hbar\Omega_{\ell}$ are in the mode, but the energy is distributed over the entire space, that is, not localized. Other useful quantum states used later are coherent states $|\alpha\rangle$, which are eigenstates of the annihilation operator $\hat{a}_{\ell}$ with eigenvalues α

$$\hat{a}_{\ell}|\alpha\rangle = \alpha|\alpha\rangle. \tag{32.35}$$

The phase of the coherent states is completely determined, while the number of photons is completely undetermined.

In the subsequent sections, we will employ a quantum electromagnetic field discussed above in order to discuss the nature of light–matter interactions apparently exhibited at the nanoscale.

32.3 Dressing Mechanism and Spatial Localization of Photons

32.3.1 Virtual Photon Cloud Surrounding a Neutral Source (in Ground State or Excited State) in QED

A quantum source material system interacting with a quantum field is influenced by virtual processes such as emission and absorption of virtual quanta of the field, and the source can be described as a dressed source, that is, the "bare" source surrounded by a cloud of virtual particles (Compagno et al. 1988, 1995). It is true for a detector. The virtual-cloud effects are responsible for the modification of the values of fundamental constants. For example, in nonrelativistic quantum electrodynamics, the presence of a virtual cloud around a hydrogen atom in its ground state contributes to the Lamb shift. Dressed-source effects can also be seen in different physical systems, such as a nucleon coupled to the meson field, and an electron coupled to the optical phonon modes of a semiconductor (polaron). The virtual cloud around the source also modifies the energy density distribution of the electromagnetic field, and the detailed properties have direct physical significance. The energy density of the virtual photon cloud at a given point is in fact related to the van der Waals interaction experienced by a suitable test object as a detector at that point. The presence of a virtual cloud around a source can influence not only its energy levels but also its dynamics.

Let us roughly estimate the linear dimensions of the virtual cloud surrounding the source or the detector. Even when the source–field system is in the ground state, a fluctuation of the field leads to the possibility of absorption or the emission of photons, not necessarily to the conservation of energy. Such energy imbalance δE is constrained by the Heisenberg uncertainty principle $\delta E \sim \hbar/\tau$, where τ is the duration of the fluctuation. Since these fluctuations take place continuously, a steady-state cloud of virtual photons is continuously emitted and reabsorbed. The virtual photon can only attain a finite distance from the source or the detector given by

$$r \sim c\tau \sim \frac{\hbar c}{\delta E}, \tag{32.36}$$

where c is the speed of light. For a transition corresponding to one of typical visible light, we set $\delta E \sim 2\,\text{eV}$ and obtain a typical linear dimension of 100 nm. This indicates that dressing effects might be prominent at the nanoscale.

We have discussed the virtual clouds of the source or detector in its ground state. From now on, we discuss the virtual cloud of the source or the detector in excited states, which can decay by emission of real photons. The above discussion inclines us to use the perturbation theory, but it fails due to the near degeneracy of states that gives rise to vanishing energy denominators at all the orders of perturbation theory. Another difficulty is the description of decay processes in a consistent way. One of such attempts is based on an extension of the eigenvalue problem to the complex E-plane. The underlying theory, unfortunately, has not yet been established, but it seems to imply time symmetry breaking and irreversibility in the dynamics of the system Section 32.3.2 will be devoted to an approach focusing on field correlations and intermolecular interactions due to the virtual clouds.

32.3.2 Electromagnetic Field Correlations and Intermolecular Interactions between Molecules in Either Ground or Excited States

The London–van der Waals interaction between two molecules in their ground states located in free space is attractive with an R^{-6} power law, where R is the intermolecular separation (Power and Thirunamachandran 1993). When both molecules are excited, the potential energy gives a repulsive force. If one of the pair is excited, the sign of the potential depends on the relative magnitudes of the relevant transition energies of the two molecules. In both cases the power law is R^{-6} in the near zone. On the other hand, the power law in the far zone tends to R^{-7} for large intermolecular separations, which is called the Casimir–Polder potential, since the finite speed of propagation (retardation effect) plays an important role in the far zone.

In these kinds of studies, the multipolar quantum dynamics in Coulomb gauge is employed because all the interactions, except for the Coulomb binding within each molecule, are mediated by transverse photons, and at the same time the retarded effects are automatically satisfied (Power and Thirunamachandran 1993). For example, for upward transitions from the ground state $|0\rangle$ to an excited state $|m\rangle$, the electric–electric spatial correlation expectation value after spatial averaging of the molecular orientation is given by using the second-order perturbation method to include virtual-cloud effects:

$$\left\langle D_i(\vec{r})D_j(\vec{r})\right\rangle_{m\leftarrow 0} \sim \begin{cases} \dfrac{\left|\vec{\mu}^{m0}\right|^2}{6\pi k_0 r^7}(13\delta_{ij}+7\hat{r}_i\hat{r}_j), & \text{for far zone } (k_0 r \gg 1), \\ \dfrac{\left|\vec{\mu}^{m0}\right|^2}{3r^6}(\delta_{ij}+3\hat{r}_i\hat{r}_j), & \text{for near zone } (k_0 r \ll 1) \end{cases} \tag{32.37}$$

where $D_i(\vec{r})$ is the i-component of the transverse displacement vector field $\vec{D}(\vec{r})$, which satisfies $\nabla\cdot\vec{D}(\vec{r})=0$, $\vec{\mu}^{m0}$ is the electric dipole transition moment for the molecular states $|m\rangle$ and $|0\rangle$, and $k_0 \equiv k_{m0} \equiv (E_m - E_0)/(\hbar c)$ denotes the wave number associated with the $m \leftarrow 0$ transition of the molecule. The i-component of the unit vector $\hat{r}$ is designated by $\hat{r}_i$, while the absolute value of the position vector $\vec{r}$ is expressed by r. The Kronecker delta is denoted by δ_{ij}.

Similarly, for downward transitions from an excited state $|p\rangle$ to the ground state $|0\rangle$, the electric–electric correlation function is obtained:

$$\left\langle D_i(\vec{r})D_j(\vec{r})\right\rangle_{0\leftarrow p} \sim \begin{cases} \dfrac{2p_0^4\left|\vec{\mu}^{0p}\right|^2}{3r^2}(\delta_{ij}-\hat{r}_i\hat{r}_j), & \text{for far zone } (p_0 r \gg 1), \\ \dfrac{\left|\vec{\mu}^{0p}\right|^2}{3r^6}(\delta_{ij}+3\hat{r}_i\hat{r}_j), & \text{for near zone } (p_0 r \ll 1), \end{cases} \tag{32.38}$$

where $p_0 \equiv p_{0p} \equiv (E_p - E_0)/(\hbar c)$ denotes the wave number associated with the $0 \leftarrow p$ transition of the molecule. It should be noted that the r^{-2} dependence arises from the real photon emission, while the r^{-6} dependence and the r^{-7} dependence are due to the virtual photon exchange. The far-zone behavior for the magnetic–magnetic correlation functions due to an electric-dipole source is also described by the same r^{-2} or r^{-7} dependence, while the near-zone result is different from its electric analog and the power law is r^{-5} instead of r^{-6}.

We have discussed the electric and magnetic correlation functions leading to the electric and magnetic energy densities associated with electric-dipole transitions in a source molecule, which can be detected by their effect on polarizable test bodies placed in the neighborhood of the source. This situation is analogous to an optical near-field system in which a nanometric material source connected to a macroscopic material and light source interacts with a nanometric detector connected to a macroscopic detector system. The difference is that the nanometric detector serving as a test body can disturb the field formed by the nanometric source in the case of the optical near-field system. We will move on to this topic in Section 32.3.3.

32.3.3 Effective Near-Field Optical Interaction between Nanomaterials Disconnected but Closely Separated

Several theoretical approaches to optical near-field problems, different from each other in viewpoints, have been proposed in the last two decades (Pohl and Courjon 1993, Ohtsu and Hori 1999). The optical near-field problems, including the application to nanophotonics, are ultimately regarded as how one should formulate a separated (more than two elements) composite system, each of which consists of a photon–electron–phonon interacting system on a nanometer scale as a source or a detector system and, which, at the same time, is connected with a macroscopic light–matter system. These questions must be clearly answered to achieve practical realization of nanophotonics. In order to provide a base for a variety of discussions in this research field, a new formulation has been developed within a quantum theoretical framework, putting matter excitations (electronic and vibrational) on an equal footing with photons (Kobayashi et al. 2001).

As discussed in Section 32.2, a "photon," whose concept has been established as a result of quantization of a free electromagnetic field, corresponds to a discrete excitation of electromagnetic modes in a virtual cavity. Unlike an electron, a photon is massless, and it is difficult to construct a wave function in the position representation that gives a picture of the photon as a spatially localized point particle like an electron. However, if there is a detector near the optical source, such as an atom, to absorb a photon in an area whose linear dimension is much smaller than the wavelength of light, it would be possible to detect a photon with the same precision as the detector size. In optical near-field problems, it is required to consider the interactions between light and nanomaterials surrounded by a macroscopic material system and detection of light by other nanomaterials on a nanometer scale. Then, a more serious question for the quantization of the field is how to define a virtual cavity, or which normal modes are to be used, since there exist more than two systems (nano-source and nano-detector) with arbitrary shape, size, and material in the nanometer region, which are still connected with a macroscopic material system, such as a source or a detector system. In this section, we describe a model and a theoretical approach to address the issue, which is essential to understand the operating principles nanophotonic devices, as well as nanofabrication using optical near fields (Ohtsu et al. 2008).

Let us consider a nanomaterial system surrounded by an incident light and a macroscopic material system, which electromagnetically interacts with one another in a complicated way, as schematically shown in Figure 32.1. Using the projection operator method, we can derive an effective interaction between the relevant nanomaterials in which we are interested (nano-source and nano-detector—either one is in the excited state), as a result of renormalizing the other effects. It corresponds to an approach to describe "photons localized around nanomaterials," as if each nanomaterial would work as a detector and light source in a self-consistent way. The effective interaction related to optical near-fields is hereafter called a near-field optical interaction. As will be discussed in detail in this section, the near-field optical interaction between nanomaterials separated by R is as follows:

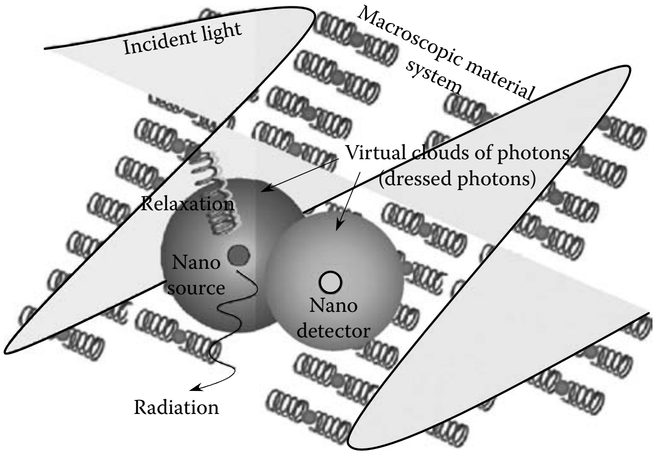
Incident light
Macroscopic material system
Virtual clouds of photons (dressed photons)
Relaxation
Nano source
Nano detector
Radiation

FIGURE 32.1 Nanometric source and detector system induced by incident light and a macroscopic material system.

$$V_{\text{eff}} = \frac{\exp(-aR)}{R}, \tag{32.39}$$

where a^{-1} is the interaction range that represents the characteristic size of the nanomaterials and does not depend on the wavelength of the light used. It indicates that photons are localized around the nanomaterials (either of which is in the excited state) as a result of the interaction with matter fields, from which a photon, in turn, can acquire a finite mass. Therefore, we might consider that the near-field optical interaction is produced via localized photon hopping between nanomaterials. On the basis of the projection operator method, we will investigate formulation of an optical near-field system that was briefly mentioned above. Moreover, the explicit functional form of the near-field optical interaction will be given by using the effective interaction $m\hat{V}_{\text{eff}}$ in a perturbative way.

32.3.3.1 Relevant Nanometric Subsystem and Irrelevant Macroscopic Subsystem

As schematically illustrated in Figure 32.1, the optical near-field system consists of two subsystems: One is a macroscopic subsystem including the incident light, whose typical dimension is much larger than the wavelength of the incident light. The other is a nanometric subsystem (nano-source and nano-detector), whose constituents are, for example, a nanometric aperture or a protrusion at the apex of a near-field optical probe, and a nanometric sample. We call such an aperture or a protrusion a probe tip.

As a nanometric sample we mainly suppose a single atom/molecule, or quantum dot (QD). Two subsystems are interacting with each other, and it is very important to formulate the interaction consistently and systematically. Here let us call the nanometric subsystem a relevant subsystem *n*, and the macroscopic subsystem an irrelevant subsystem *M*. We are interested in the subsystem *n*; in particular, the interaction induced in the subsystem *n*. Therefore, it is the key to renormalize the effects originating from the subsystem *M* in a consistent and systematic way. Now we show a formulation based on the projection operator method, described below.

32.3.3.2 P Space and Q Space

It is preferable, for a variety of discussions, to express exact eigenstates $|\psi\rangle$ for the total system described by the total Hamiltonian $\hat{H}$ in terms of a small number of bases of as small a number of degrees of freedom as possible, which span P space. In the following, let us assume two states as the P-space components: $|\phi_1\rangle = |s^*\rangle|p\rangle \otimes |0_{(M)}\rangle$ and $|\phi_2\rangle = |s\rangle|p^*\rangle \otimes |0_{(M)}\rangle$, both of which are eigenstates of the unperturbed Hamiltonian $\hat{H}_0$. Here $|s\rangle$ and $|s^*\rangle$ are eigenstates of the sample that is isolated from the others, while $|p\rangle$ and $|p^*\rangle$ are eigenstates of the probe tip, which is also isolated. In addition, exciton polariton states, which are a mixture of photons and electron–hole pairs, are used as bases to describe the macroscopic subsystem *M*, and thus $|0_{(M)}\rangle$ represents the vacuum for exciton polaritons. Note that there exist photons and electronic matter excitations even in the vacuum state $|0_{(M)}\rangle$. The direct product is denoted by the symbol $\otimes$. The complementary space to the P space is called Q space, which is spanned by a huge number of bases of a large number of degrees of freedom not included in the P space, as schematically shown in Figure 32.2.

32.3.3.3 Effective Interaction Exerted in the Nanometric Subsystem

Noticing the relation between a bare interaction Hamiltonian $\hat{V} = \hat{H} - \hat{H}_0$ and an effective interaction Hamiltonian operator $\hat{V}_{\text{eff}}$, namely, $\langle\psi|\hat{V}|\psi\rangle = \langle\psi|P\hat{V}_{\text{eff}}P|\psi\rangle$, we obtain the effective interaction Hamiltonian operator in the P space, given by

$$\hat{V}_{\text{eff}} = (P\hat{J}^\dagger\hat{J}P)^{-1/2}(P\hat{J}^\dagger\hat{V}\hat{J}P)(P\hat{J}^\dagger\hat{J}P)^{-1/2}, \tag{32.40}$$

and tracing out the other degrees of freedom gives an effective interaction Hamiltonian of the nanometric subsystem *n* after renormalizing the effects from the macroscopic subsystem *M*. Here, *P* is the projection operator, and *Q* is the complimentary operator defined by $Q = 1 - P$, both of which satisfy the following relations:

$$P = P^\dagger, \quad P^2 = P, \quad [P, \hat{H}_0] = 0, \tag{32.41a}$$

$$Q = Q^\dagger, \quad Q^2 = Q, \quad [Q, \hat{H}_0] = 0, \tag{32.41b}$$

$$PQ = QP = 0. \tag{32.41c}$$

The operator $\hat{J}$ is defined by

$$\hat{J} = [1 - (E - \hat{H}_0)^{-1}Q\hat{V}]^{-1}, \tag{32.42}$$

where *E* are the eigenvalues of the total Hamiltonian $\hat{H}$. Using the effective interaction Hamiltonian, one can forget the subsystem *M* as if the subsystem *n* were isolated and separated from the subsystem *M*.

To obtain an explicit expression of the effective interaction Hamiltonian, let us employ the bare interaction $\hat{V}$ between the two subsystems, which in the multipolar formalism (Craig and Thirunamachandran 1998) is given by

$$\hat{V} = -\left\{\hat{\vec{\mu}}_s \cdot \hat{\vec{D}}^\perp\left(\vec{r}_s\right) + \hat{\vec{u}}_p \cdot \hat{\vec{D}}^\perp\left(\vec{r}_p\right)\right\}, \tag{32.43}$$

where the canonical momentum of the vector potential operator $\hat{\vec{A}}(\vec{r})$ is proportional to the transverse displacement vector field operator* $\hat{\vec{D}}^\perp(\vec{r})$, while the electric dipole operator is denoted as $\hat{\vec{\mu}}(\vec{r})$. It should be noted that there are no interactions, i.e., $\hat{V} = 0$, without incident photons in the macroscopic subsystem *M*. The subscripts *s* and *p* represent physical quantities related to the sample and the probe tip, respectively. Representative positions of the sample and the probe tip are chosen, for simplicity, by the vectors $\vec{r}_s$ and $\vec{r}_p$, respectively, but may be composed of several positions. In that case the quantities inside curly brackets in (32.43) should be read as a summation. The operator $\hat{\vec{\Pi}}(\vec{r})$ conjugate to $\hat{\vec{A}}(\vec{r})$ is expressed in terms of $\hat{\vec{D}}^\perp(\vec{r})$ as follows:

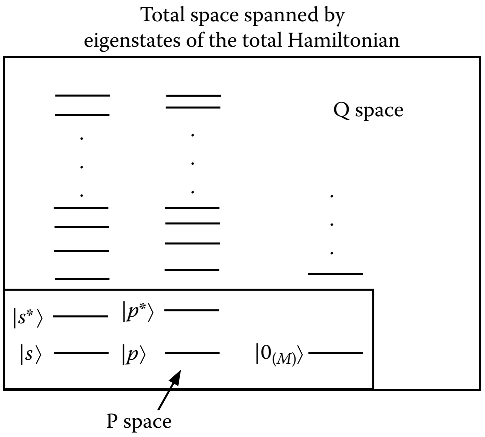

FIGURE 32.2 Subdivision of the space spanned by eigenstates of the total Hamiltonian of the system.

---

* The transverse component is defined by $\nabla \cdot \vec{F}^\perp = 0$, while the longitudinal component is defined by $\nabla \times \vec{F}^\parallel = 0$, for an arbitrary vector field $\vec{F}(\vec{r})$.

$$\hat{\vec{\Pi}}(\vec{r}) = -\frac{1}{4\pi c}\hat{\vec{E}}^{\perp}(\vec{r}) - \frac{1}{c}\hat{\vec{P}}^{\perp}(\vec{r}) = -\frac{1}{4\pi c}\hat{\vec{D}}^{\perp}(\vec{r}), \tag{32.44}$$

where $\hat{\vec{E}}^{\perp}(\vec{r})$ and $\hat{\vec{P}}^{\perp}(\vec{r})$ are the transverse electric field and the induced polarization field, respectively. With the help of the mode expansion of $\hat{\vec{A}}(\vec{r})$ and $\hat{\vec{\Pi}}(\vec{r})$, that is,

$$\hat{\vec{A}}(\vec{r}) = \sum_{k}\sum_{\lambda=1}\left(\frac{2\pi\hbar c^2}{V\omega_{\vec{k}}}\right)^{1/2}\vec{e}_\lambda(\vec{k})\left\{\hat{a}_\lambda(\vec{k})e^{i\vec{k}\cdot\vec{r}} + \hat{a}_\lambda^{\dagger}(\vec{k})e^{-i\vec{k}\cdot\vec{r}}\right\}, \tag{32.45}$$

and

$$\hat{\vec{\Pi}}(\vec{r}) = -\frac{i}{4\pi c}\sum_{\vec{k}}\sum_{\lambda=1}^{2}\left(\frac{2\pi\hbar\omega_{\vec{k}}}{V}\right)^{1/2}\vec{e}_\lambda(\vec{k})\left\{\hat{a}_\lambda(\vec{k})e^{i\vec{k}\cdot\vec{r}} - \hat{a}_\lambda^{\dagger}(\vec{k})e^{-i\vec{k}\cdot\vec{r}}\right\} \tag{32.46}$$

we can rewrite the transverse component of the electric displacement operator as

$$\hat{\vec{D}}^{\perp}(\vec{r}) = i\sum_{\vec{k}}\sum_{\lambda=1}^{2}\left(\frac{2\pi\hbar\omega_{\vec{k}}}{V}\right)^{1/2}\vec{e}_\lambda(\vec{k})\left\{\hat{a}_\lambda(\vec{k})e^{i\vec{k}\cdot\vec{r}} - \hat{a}_\lambda^{\dagger}(\vec{k})e^{-i\vec{k}\cdot\vec{r}}\right\}, \tag{32.47}$$

where the plane waves are used for the mode functions, and the creation and annihilation operators of a photon with wave vector $\vec{k}$, angular frequency $\omega_{\vec{k}}$, and polarization component λ are designated by $\hat{a}_\lambda^{\dagger}(\vec{k})$ and $\hat{a}_\lambda(\vec{k})$, respectively. The quantization volume is V, and the unit vector related to the polarization direction is shown by $\vec{e}_\lambda(\vec{k})$. Note that the electric field outside the material corresponds to $\hat{\vec{D}}^{\perp}(\vec{r})$.

Since exciton polariton states are employed as bases to describe the macroscopic subsystem M, the creation and annihilation operators of a photon in (32.47) are rewritten using the exciton polariton operators $\hat{\xi}^{\dagger}(\vec{k})$ and $\hat{\xi}(\vec{k})$, and then they are substituted into (32.43). Using the electric dipole operator defined by

$$\hat{\vec{\mu}}_\alpha = \left(\hat{B}(\vec{r}_\alpha) + \hat{B}^{\dagger}(\vec{r}_\alpha)\right)\vec{\mu}_\alpha, \tag{32.48}$$

with the creation and annihilation operators of excitation in subsystem n, $\hat{B}^{\dagger}(\vec{r}_\alpha)$ and $\hat{B}(\vec{r}_\alpha)$, and the transition dipole moments $\vec{\mu}_\alpha$ ($\alpha = s, p$), we obtain the bare interaction in the exciton polariton picture:

$$\hat{V} = -i\sum_{\alpha=s}^{p}\sum_{\vec{k}}\left(\frac{2\pi\hbar}{V}\right)^{1/2}\left(\hat{B}(\vec{r}_\alpha) + \hat{B}^{\dagger}(\vec{r}_\alpha)\right)\left(K_\alpha(\vec{k})\hat{\xi}(\vec{k}) - K_\alpha^{\star}(\vec{k})\hat{\xi}^{\dagger}(\vec{k})\right). \tag{32.49}$$

Here $K_\alpha(\vec{k})$ is the coupling coefficient between the exciton polariton and the nanometric subsystem n, given by

$$K_\alpha(\vec{k}) = \sum_{\lambda=1}^{2}\left(\vec{\mu}_\varepsilon\cdot\vec{e}_\lambda(\vec{k})\right)f(k)e^{i\vec{k}\cdot\vec{r}_\alpha} \tag{32.50}$$

with

$$f(k) = \frac{ck}{\sqrt{\Omega(k)}}\sqrt{\frac{\Omega^2(k) - \Omega^2}{2\Omega^2(k) - \Omega^2 - (ck)^2}}. \tag{32.51}$$

The complex conjugate of $K_\alpha(\vec{k})$ is denoted by $K_\alpha^{\star}(\vec{k})$, while c, $\Omega(k)$, and Ω are light speed in vacuum and the eigenfrequencies of both exciton polariton and electronic polarization of the macroscopic subsystem M, respectively. The dispersion relation for a free photon, $\omega_{\vec{k}} = ck$, is used in (32.51). Note that the wave-number dependence of $f(k)$ characterizes a typical interaction range of exciton polaritons coupled to the nanometric subsystem n.

The next step is to evaluate the amplitude of the effective interaction exerted on the nanometric subsystem, for example, the effective sample–probe-tip interaction in the P space after tracing out the polariton degrees of freedom:

$$V_{\text{eff}}(2,1) \equiv \langle\phi_2|\hat{V}_{\text{eff}}|\phi_1\rangle. \tag{32.52}$$

With the first-order approximation $\hat{J}^{(1)}$ of $\hat{J}$ in (32.42), we can explicitly write (32.52) in the following form:

$$\begin{aligned} V_{\text{eff}}(2,1) &= \langle\phi_2|P\hat{V}Q\hat{V}(E_P^0 - E_Q^0)^{-1}P\,|\,\phi_1\rangle + \langle\phi_2|P(E_P^0 - E_Q^0)^{-1}\hat{V}Q\hat{V}P|\phi_1\rangle \\ &= \sum_m \langle\phi_2|P\hat{V}Q|m\rangle\langle m|Q\hat{V}P|\phi_1\rangle\left(\frac{1}{E_{P1}^0 - E_{Qm}^0} + \frac{1}{E_{P2}^0 - E_{Qm}^0}\right). \end{aligned} \tag{32.53}$$

The second line shows that a virtual transition from the initial state $|\phi_1\rangle$ in the P space to an intermediate state $|m\rangle$ in the Q space is followed by a subsequent virtual transition from the intermediate state $|m\rangle$ to the final state $|\phi_2\rangle$ in the P space. Here $E_{P1}^0(E_{P2}^0)$ and E_{Qm}^0 denote eigenenergies of $|\phi_1\rangle(|\phi_2\rangle)$ in the P space and that of $|\phi_m\rangle$ in the Q space, respectively. Now we can proceed to the next step by substituting the explicit bare interaction $\hat{V}$ in (32.49) with (32.50) and (32.51) into (32.53). First of all, note that the one-exciton polariton state among arbitrary intermediate states $|m\rangle$ can only contribute to nonzero matrix elements. Therefore, (32.53) can be transformed into

$$V_{\text{eff}}(2,1) = -\frac{\hbar}{(2\pi)^2}\int d^3k\left[\frac{K_p(\vec{k})K_s^{\star}(\vec{k})}{\Omega(k) - \Omega_0(s)} + \frac{K_s(\vec{k})K_p^{\star}(\vec{k})}{\Omega(k) + \Omega_0(s)}\right], \tag{32.54}$$

where the summation over $\vec{k}$ is replaced by $\vec{k}$-integration, i.e., $\frac{V}{(2\pi)^3}\int d^3k$, in the usual manner. Excitation energies of the sample (between $|s^{*}\rangle$ and $|s\rangle$) and the probe tip (between $|p^{*}\rangle$ and $|p\rangle$) are denoted as $E_s = \hbar\Omega_0(s)$ and $E_p = \hbar\Omega_0(p)$, respectively.

Exchanging the arguments 1 and 2, or the role of the sample and probe tip, we can similarly calculate $V_{\text{eff}}(1,2) \equiv \langle\phi_1|\hat{V}_{\text{eff}}|\phi_2\rangle$:

$$V_{\text{eff}}(1,2) = -\frac{\hbar}{(2\pi)^2}\int d^3k\left[\frac{K_s(\vec{k})K_p^{\star}(\vec{k})}{\Omega(k)-\Omega_0(p)} + \frac{K_p(\vec{k})K_s^{\star}(\vec{k})}{\Omega(k)+\Omega_0(p)}\right]. \tag{32.55}$$

Therefore, the total amplitude of the effective sample–probe-tip interaction is given by the sum of (32.54) and (32.55), which includes the effects from the macroscopic subsystem M. We write this effective interaction for the nanometric subsystem n as $V_{\text{eff}}(\vec{r})$ in the following way:

$$V_{\text{eff}}\left(\vec{r}\right) = -\frac{1}{4\pi^2}\sum_{\lambda=1}^{2}\sum_{\alpha=p,s}\int d^3k\left[\left(\vec{\mu}_p\cdot\vec{e}_\lambda(\vec{k})\right)\left(\vec{\mu}_s\cdot\vec{e}_\lambda(\vec{k})\right)\right]$$
$$\times\hbar f^2(k)\left(\frac{e^{i\vec{k}\cdot\vec{r}}}{E(k)+E_\alpha} + \frac{e^{-i\vec{k}\cdot\vec{r}}}{E(k)-E_\alpha}\right) \tag{32.56}$$

where we have set $E(k) = \hbar\Omega(k)$, and $E_\alpha = \hbar\Omega_0(a^{*}) - \hbar\Omega_0(\alpha)$ for $\alpha = p$ and $\alpha = s$. The summation over polarization λ is performed as

$$\sum_{\lambda=1}^{2}\vec{e}_{\lambda i}(\vec{k})\vec{e}_{\lambda i}(\vec{k}) = \delta_{ij} - k_i k_j, \tag{32.57}$$

and thus the summation of $\left(\vec{\mu}_p\cdot\vec{e}_\lambda(\vec{k})\right)\left(\vec{\mu}_s\cdot\vec{e}_\lambda(\vec{k})\right)$ over λ can be reduced as follows:

$$\sum_{\lambda=1}^{2}\left(\vec{\mu}_p\cdot\vec{e}_\lambda(\vec{k})\right)\left(\vec{\mu}_s\cdot\vec{e}_\lambda(\vec{k})\right) = \sum_{\lambda=1}^{2}\sum_{i,j}\left(\vec{\mu}_{pi}\vec{e}_{\lambda i}(\vec{k})\right)\left(\vec{\mu}_{sj}\vec{e}_{\lambda j}(\vec{k})\right)$$
$$= \sum_{i,j}\mu_{pi}\mu_{sj}\left(\delta_{ij} - \hat{k}_i\hat{k}_j\right) \tag{32.58}$$

with the unit vector $\hat{k} \equiv \vec{k}/k$. Noticing that $d^3k = k^2dkd\Omega = k^2dk \sin\theta d\theta d\varphi$ and

$$\delta_{ij}\int e^{\pm i\vec{k}\cdot\vec{r}}d\Omega = \delta_{ij}\int_0^{2\pi}\int_{-1}^{1} e^{\pm ikr\cos\theta}d(\cos\theta)d\varphi = \delta_{ij}\frac{2\pi}{ikr}(e^{ikr} - e^{-ikr}), \tag{32.59a}$$

$$-\int\hat{k}_i\hat{k}_j e^{\pm i\vec{k}\cdot\vec{r}}d\Omega = \frac{1}{k^2}\nabla_i\nabla_j\int e^{\pm i\vec{k}\cdot\vec{r}}d\Omega = \frac{2\pi}{ik^3}\nabla_i\nabla_j\left(\frac{e^{ikr} - e^{-ikr}}{r}\right), \tag{32.59b}$$

we find

$$\int(\delta_{ij} - \hat{k}_i\hat{k}_j)e^{\pm i\vec{k}\cdot\vec{r}}d\Omega = \delta_{ij}\frac{2\pi}{ik}\left(\frac{e^{ikr} - e^{-ikr}}{r}\right) + \frac{2\pi}{ik^3}\nabla_i\nabla_j\left(\frac{e^{ikr} - e^{-ikr}}{r}\right)$$
$$= 2\pi\left[\delta_{ij}\frac{(e^{ikr} - e^{-ikr})}{ikr} + (\delta_{ij} - 3\hat{r}_i\hat{r}_j)\left\{\frac{(e^{ikr} + e^{-ikr})}{k^2r^2} - \frac{(e^{ikr} - e^{-ikr})}{ikr^3r^3}\right\}\right.$$
$$\left. - \frac{(e^{ikr} - e^{-ikr})}{ikr}\hat{r}_i\hat{r}_j\right] \tag{32.60}$$

where $\hat{r}$ is the unit vector defined by $\hat{r} \equiv \vec{r}/r$, and the jth component is denoted by $\hat{r}_j$. Hence, the effective interaction can be rewritten as

$$V_{\text{eff}}(\vec{r}) = -\frac{1}{2\pi}\int_{-\infty}^{\infty}k^2dk\hbar f^2(k)\sum_{a=s,p}\left(\frac{1}{E(k)+E_a} + \frac{1}{E(k)-E_a}\right)$$
$$\times\left\{(\vec{\mu}_s\cdot\vec{\mu}_p)e^{i\vec{k}\cdot\vec{r}}\left(\frac{1}{ikr} + \frac{1}{k^2r^2} - \frac{1}{ik^3r^3}\right) - (\vec{\mu}_s\cdot\hat{r})(\vec{\mu}_p\cdot\hat{r})e^{i\vec{k}\cdot\vec{r}}\right.$$
$$\left.\times\left(\frac{1}{ikr} + \frac{3}{k^2r^2} - \frac{3}{ik^3r^3}\right)\right\} \tag{32.61}$$

where the integration range is extended from (0, ∞) to (−∞, ∞). When the dispersion relation of exciton polaritons, which have been chosen as a basis for describing the macroscopic subsystem M, is approximated as

$$E(k) = \hbar\Omega + \frac{(\hbar k)^2}{2m_{\text{pol}}} = E_m + \frac{(\hbar ck)^2}{2E_{\text{pl}}}, \tag{32.62}$$

in terms of the effective mass of exciton polaritons, m_{pol}, or $E_{\text{pl}} = m_{\text{pol}}c^2$ and the electronic excitation energy of the macroscopic subsystem M, $E_m = \hbar\Omega$, (32.61) is further simplified:

$$V_{\text{eff}}(\vec{r}) = -\frac{1}{2\pi}\int_{-\infty}^{\infty}k^2dk\hbar f^2(k)\sum_{\alpha=s,p}\frac{2E_{\text{pl}}}{(\hbar c)^2}$$
$$\times\left\{\frac{1}{(k+i\Delta_{\alpha+})(k-i\Delta_{\alpha+})} + \frac{1}{(k+i\Delta_{\alpha-})(k-i\Delta_{\alpha-})}\right\}$$
$$\times\left\{(\vec{\mu}_s\cdot\vec{\mu}_p)e^{i\vec{k}\cdot\vec{r}}\left(\frac{1}{ikr} + \frac{1}{k^2r^2} - \frac{1}{ik^3r^3}\right) - \left(\vec{\mu}_s\cdot\hat{r}\right)\left(\vec{\mu}_p\cdot\hat{r}\right)e^{i\vec{k}\cdot\vec{r}}\right.$$
$$\left.\times\left(\frac{1}{ikr} + \frac{3}{k^2r^2} - \frac{3}{ik^3r^3}\right)\right\}$$
$$\equiv \sum_{\alpha=s,p}\left[V_{\text{eff},\alpha+}(\vec{r}) + V_{\text{eff},\alpha-}(\vec{r})\right] \tag{32.63a}$$

with

$$\Delta_{\alpha\pm} \equiv \frac{1}{\hbar c}\sqrt{2E_{\text{pl}}(E_m \pm E_\alpha)}, \quad (E_m > E_\alpha). \tag{32.63b}$$

The k-integration can be performed with the residues evaluated at $k = i\Delta_{\alpha\pm}$, and we have

$$\begin{aligned} V_{\text{eff},\alpha\pm}(\vec{r}) = \mp\frac{1}{2}\Bigg[&(\vec{\mu}_s\cdot\vec{\mu}_p)\left\{\frac{(\Delta_{\alpha\pm})^2}{r}+\frac{\Delta_{\alpha\pm}}{r^2}+\frac{1}{r^3}\right\}W_{\alpha\pm}e^{-\Delta_{\alpha\pm}r} \\ &-(\vec{\mu}_s\cdot\hat{r})(\vec{\mu}_p\cdot\hat{r})\left\{\frac{(\Delta_{\alpha\pm})^2}{r}+\frac{3\Delta_{\alpha\pm}}{r^2}+\frac{3}{r^3}\right\}W_{\alpha\pm}e^{-\Delta_{\alpha\pm}r}\Bigg], \end{aligned} \tag{32.64a}$$

where the constants $W_{\alpha\pm}$ are defined by

$$W_{\alpha\pm} \equiv \frac{E_{\text{pl}}}{E_\alpha}\frac{E_m^2 - E_\alpha^2}{(E_m \pm E_\alpha)(E_m - E_{\text{pl}} \mp E_\alpha) - E_m^2/2}. \tag{32.64b}$$

If the angular average of $(\vec{\mu}_s\cdot\hat{r})(\vec{\mu}_p\cdot\hat{r})$ is taken, the expression

$$V_{\text{eff}}(r) = -\frac{(\vec{\mu}_s\cdot\vec{\mu}_p)}{3}\sum_{\alpha=s,p}\left\{W_{\alpha+}(\Delta_{\alpha+})^2\frac{e^{-\Delta_{\alpha+}r}}{r} - W_{\alpha-}(\Delta_{\alpha-})^2\frac{e^{-\Delta_{\alpha-}r}}{r}\right\} \tag{32.65}$$

is obtained for the effective interaction, or the near-field optical interaction $V_{\text{eff}}(r)$, which consists of the sum of the Yukawa functions $Y(\Delta_{\alpha\pm}r) \equiv e^{-\Delta_{\alpha\pm}r}/r$ with a shorter interaction range $\Delta_{\alpha+}$(heavier effective mass) and a longer interaction range $\Delta_{\alpha-}$(lighter effective mass).

To sum up, we find that the major part of the effective interaction exerted in the nanometric subsystem n is the Yukawa potential after renormalizing the effects from the macroscopic subsystem. This interaction comes from the mediation of massive virtual photons corresponding to the counter-rotating term, or polaritons, where exciton polaritons have been employed in an explicit formulation, but in principle other types of polaritons would be applicable. In this section we have mainly focused on the effective interaction of the nanometric subsystem n, after tracing out the other degrees of freedom. It is certainly possible to have a formulation with a projection onto the P space that is spanned in terms of the degrees of freedom of the massive virtual photons, although this is left as the subject of future work. This kind of formulation emphasizes a "dressed photon" picture, in which photons are not massless but massive, due to light–matter interactions at the nanoscale.

32.3.4 Localization of a Photon Dressed by Matter Excitation in Nanomaterials at the Nanoscale

In this section, we consider a simple model system, for example, a pseudo one-dimensional near-field optical probe system, to discuss the mechanism of photon localization in space as well as the phonon's role. In order to focus on the photon–phonon interaction, the interacting part between photon and electronic excitation is first expressed in terms of a polariton, and is called a photon in the model. Then the model Hamiltonian that describes the photon and phonon interacting system is presented. Using the Davydov transformation, we rewrite the Hamiltonian in terms of quasiparticles. On the basis of the Hamiltonian, we present numerical results on the spatial distribution of photons and discuss the mechanism of photon localization due to phonons.

32.3.4.1 Model Hamiltonian

We consider a near-field optical probe, schematically shown in Figure 32.3, as an example system where light interacts with both phonons and electrons in the probe on a nanometer scale. Here, the interaction of a photon and an electronic excitation is assumed to be expressed in terms of a polariton basis, as discussed above, and is hereafter called a photon so that special attention is paid to the photon–phonon interaction. The system is simply modeled as a one-dimensional atomic or molecular chain coupled with photon and phonon fields. The chain consists of finite N molecules (representatively called), each of which is located at a discrete point (called a molecular site) whose separation represents a characteristic scale of the near-field system. Photons are expressed in the site representation and can hop to the nearest neighbor sites due to the short-range interaction nature of the optical near fields.

The Hamiltonian for the above model is given by

$$\begin{aligned} \hat{H} = \sum_{i=1}^{N}\hbar\omega\hat{a}_i^\dagger\hat{a}_i + \left\{\sum_{i=1}^{N}\frac{\hat{p}_i^2}{2m_i} + \sum_{i=1}^{N-1}\frac{k}{2}(\hat{x}_{i+1}-\hat{x}_i)^2 + \sum_{i=1,N}\frac{k}{2}\hat{x}_i^2\right\} \\ + \sum_{i=1}^{N}\hbar\chi\hat{a}_i^\dagger\hat{a}_i\hat{x}_i + \sum_{i=1}^{N-1}\hbar J\left(\hat{a}_i^\dagger\hat{a}_{i+1} + \hat{a}_{i+1}^\dagger\hat{a}_i\right), \end{aligned} \tag{32.66}$$

where

$\hat{a}_i^\dagger$ and $\hat{a}_i$ correspondingly denote the creation and annihilation operators of a photon with energy $\hbar\omega$ at site i in the chain

$\hat{x}_i$ and $\hat{p}_i$ represent the displacement and conjugate momentum operators of the vibration, respectively

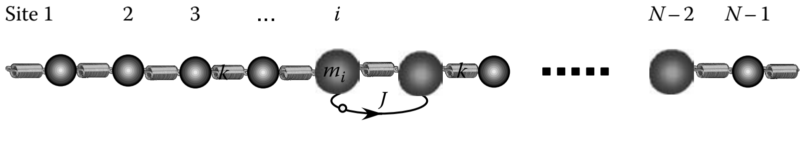

FIGURE 32.3 Simple one-dimensional model for a light–matter interacting system on a nanometer scale.

The mass of a molecule at site i is designated by m_i, and each molecule is assumed to be connected by springs with spring constant k. The third and fourth terms in (32.66) stand for the photon–vibration interaction with coupling constant χ and the photon hopping with hopping constant J, respectively. After the vibration field is quantized in terms of phonon operators of mode p and frequency Ω_p, $\hat{b}_p^\dagger$, and $\hat{b}_p$, the Hamiltonian (32.66) can be rewritten as

$$\hat{H} = \sum_{i=1}^{N} \hbar\omega \hat{a}_i^\dagger \hat{a}_i + \sum_{p=1}^{N} \hbar\Omega_p \hat{b}_p^\dagger \hat{b}_p + \sum_{i=1}^{N}\sum_{p=1}^{N} \hbar\chi_{ip}\hat{a}_i^\dagger \hat{a}_i(\hat{b}_p^\dagger + \hat{b}_p) + \sum_{i=1}^{N-1} \hbar J(\hat{a}_i^\dagger \hat{a}_{i+1} + \hat{a}_{i+1}^\dagger \hat{a}_i), \tag{32.67}$$

with the coupling constant χ_{ip} of a photon at site i and a phonon of mode p. It should be noted that the index p designates not the momentum but the mode number, because the translational invariance of the system is broken and the momentum is not a good quantum number. The site-dependent coupling constant χ_{ip} is related to the original coupling constant χ in terms of the transformation matrix as P_{ip} as

$$\chi_{ip} = \chi \mathrm{P}_{ip}\sqrt{\frac{\hbar}{2m_i\Omega_p}}, \tag{32.68}$$

and the creation and annihilation operators of a photon and a phonon satisfy the boson commutation relation as $\left[\hat{a}_i, \hat{a}_j^\dagger\right] = \delta_{ij}$ and $\left[\hat{b}_p, \hat{b}_q^\dagger\right] = \delta_{pq}$. The Hamiltonian (32.67), which describes the model system, is not easily handled because of the third order of the operators in the interaction term. To avoid this difficulty, this direct photon–phonon interaction term in (32.67) is eliminated by the Davydov transformation in the following subsection.

32.3.4.2 Davydov Transformation

Before going into the explicit expression, we discuss a unitary transformation $\hat{U}$ generated by an anti-Hermitian operator $\hat{S}$, defined as

$$\hat{U} \equiv \exp(\hat{S}), \quad \text{with } \hat{S}^\dagger = -\hat{S} \tag{32.69a}$$

and

$$\hat{U}^\dagger = \hat{U}^{-1}. \tag{32.69b}$$

Assume a Hamiltonian $\hat{H}$ that consists of a diagonalized part $\hat{H}_0$ and a non-diagonal interaction part $\hat{V}$:

$$\hat{H} = \hat{H}_0 + \hat{V}. \tag{32.70}$$

Transforming the Hamiltonian in (32.70) as

$$\tilde{H} \equiv \hat{U}\hat{H}\hat{U}^\dagger = \hat{U}\hat{H}\hat{U}^{-1}, \tag{32.71}$$

we have

$$\begin{aligned}\tilde{H} &= \hat{H} + [\hat{S}, \hat{H}] + \frac{1}{2}\left[\hat{S}, [\hat{S}, \hat{H}]\right] + \cdots \\ &= \hat{H}_0 + \hat{V} + [\hat{S}, \hat{H}_0] + [\hat{S}, \hat{V}] + \frac{1}{2}\left[\hat{S}, [\hat{S}, \hat{H}_0]\right] + \cdots.\end{aligned} \tag{32.72}$$

If the interaction $\hat{V}$ can be perturbative, and if the operator $\hat{S}$ is chosen so that the second and third terms in (32.72) are canceled out, then

$$\hat{V} = -[\hat{S}, \hat{H}_0], \tag{32.73}$$

the Hamiltonian (32.70) is rewritten as

$$\tilde{H} = \hat{H}_0 - \frac{1}{2}\left[\hat{S}, [\hat{S}, \hat{H}_0]\right] + \cdots, \tag{32.74}$$

and can be diagonalized within the first order of $\hat{V}$.

Now we apply the above discussion to the model Hamiltonian (32.67),

$$\hat{H}_0 = \sum_{i=1}^{N} \hbar\omega \hat{a}_i^\dagger \hat{a}_i + \sum_{p=1}^{N} \hbar\Omega_p \hat{b}_p^\dagger \hat{b}_p, \tag{32.75a}$$

$$\hat{V} = \sum_{i=1}^{N}\sum_{p=1}^{N} \hbar\chi_{ip}\hat{a}_i^\dagger \hat{a}_i(\hat{b}_p^\dagger + \hat{b}_p), \tag{32.75b}$$

tentatively neglecting the hopping term. Assuming the anti-Hermitian operator $\hat{S}$

$$\hat{S} = \sum_{i}\sum_{p} f_{ip}\hat{a}_i^\dagger \hat{a}_i(\hat{b}_p^\dagger - \hat{b}_p), \tag{32.76a}$$

we can determine f_{ip} from (32.73) as follows:

$$f_{ip} = \frac{\chi_{ip}}{\Omega_p}. \tag{32.76b}$$

This operator form of $\hat{S}$ leads us to not the perturbative but the exact transformation of the photon and phonon operators:

$$\hat{\alpha}_i^\dagger \equiv \hat{U}^\dagger \hat{a}_i^\dagger \hat{U} = \hat{a}_i^\dagger \exp\left\{-\sum_{p=1}^{N} \frac{\chi_{ip}}{\Omega_p}\left(\hat{b}_p^\dagger - \hat{b}_p\right)\right\}, \tag{32.77a}$$

$$\hat{\alpha}_i \equiv \hat{U}^\dagger \hat{a}_i \hat{U} = \hat{a}_i \exp\left\{ \sum_{p=1}^{N} \frac{\chi_{ip}}{\Omega_p} (\hat{b}_p^\dagger - \hat{b}_p) \right\}, \qquad (32.77b)$$

$$\hat{\beta}_p^\dagger \equiv \hat{U}^\dagger \hat{b}_p^\dagger \hat{U} = \hat{b}_p^\dagger + \sum_{i=1}^{N} \frac{\chi_{ip}}{\Omega_p} \hat{a}_i^\dagger \hat{a}_i, \qquad (32.77c)$$

$$\hat{\beta}_p \equiv \hat{U} \hat{b}_p \hat{U}^\dagger = \hat{b}_p + \sum_{i=1}^{N} \frac{\chi_{ip}}{\Omega_p} \hat{a}_i^\dagger \hat{a}_i. \qquad (32.77d)$$

These transformed operators can be regarded as the creation and annihilation operators of quasiparticles—dressed photons and phonons—that satisfy the same boson commutation relations as those of photons and phonons before the transformation, namely, $\left[\hat{\alpha}_i, \hat{\alpha}_j^\dagger\right] = \hat{U}^\dagger \left[\hat{a}_i, \hat{a}_j\right] \hat{U} = \delta_{ij}$ and $\left[\hat{\beta}_p, \hat{\beta}_q^\dagger\right] = \hat{U}^\dagger \left[\hat{b}_p, \hat{b}_q^\dagger\right] \hat{U} = \delta_{pq}$.

Using the quasiparticle operators, we can rewrite the Hamiltonian (32.67) as

$$\hat{H} = \sum_{i=1}^{N} \hbar\omega \hat{\alpha}_i^\dagger \hat{\alpha}_i + \sum_{p=1}^{N} \hbar\Omega_p \hat{\beta}_p^\dagger \hat{\beta}_p - \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N} \frac{\hbar \chi_{ip} \chi_{jp}}{\Omega_p} \hat{\alpha}_i^\dagger \hat{\alpha}_i \hat{\alpha}_j^\dagger \hat{\alpha}_j$$

$$+ \sum_{i=1}^{N-1} \hbar \left(\hat{J}_i \hat{\alpha}_i^\dagger \hat{\alpha}_{i+1} + \hat{J}_i^\dagger \hat{\alpha}_{i+1}^\dagger \hat{\alpha}_i \right), \qquad (32.78a)$$

with

$$\hat{J}_i = J \exp\left\{ \sum_{p=1}^{N} \frac{(\chi_{ip} - \chi_{i+1p})}{\Omega_p} (\hat{\beta}_p^\dagger - \hat{\beta}_p) \right\}, \qquad (32.78b)$$

where it is noted that the direct photon–phonon coupling term has been eliminated, while the quadratic form $\hat{N}_i \hat{N}_j$ with the number operator of $\hat{N}_i = \hat{\alpha}_i^\dagger \hat{\alpha}_i$ has emerged as well as the site-dependent hopping operator $\hat{J}_i$ in (32.78b). The number states of quasiparticles are thus eigenstates of each terms of the Hamiltonian (32.78a), except the last term that represents the higher-order effect of photon–phonon coupling through the dressed photon hopping. Therefore, it is a more appropriate form to discuss the phonon's effect on photon's behavior as localization.

32.3.4.3 Quasiparticle and Coherent State

In the previous section, we have transformed the original Hamiltonian with the Davydov transformation. In order to grasp the physical meanings of the quasiparticles introduced above, the creation operator $\hat{\alpha}_i^\dagger$ is applied to the vacuum state $|0\rangle$. Then, it follows from (32.77a) that

$$\hat{\alpha}_i^\dagger |0\rangle = \hat{a}_i^\dagger \exp\left\{ -\sum_{p=1}^{N} \frac{\chi_{ip}}{\Omega_p} \left(\hat{b}_p^\dagger - \hat{b}_p \right) \right\} |0\rangle,$$

$$= \hat{a}_i^\dagger \exp\left\{ -\sum_{p=1}^{N} \frac{1}{2} \left(\frac{\chi_{ip}}{\Omega_p} \right)^2 \right\} \exp\left\{ -\frac{\chi_{ip}}{\Omega_p} \hat{b}_p^\dagger \right\} |0\rangle, \qquad (32.79)$$

where a photon at site i is associated with phonons in a coherent state, i.e., a photon is dressed by an infinite number of phonons. This corresponds to the fact that an optical near field is generated from a result of interactions between the photon and matter fields.

When $\hat{\beta}_p^\dagger$ is applied to the vacuum state $|0\rangle$, we have

$$\hat{\beta}_p^\dagger |0\rangle = \hat{b}_p^\dagger |0\rangle, \qquad (32.80)$$

and it is expressed by only the bare phonon operator (before the transformation) in the same p mode. Therefore, we mainly focus on the quasiparticle expressed by $(\hat{\alpha}_i^\dagger, \hat{\alpha}_i)$ in the following subsection. Note that it is valid only if the bare photon number (the expectation value of $\hat{\alpha}_i^\dagger \hat{\alpha}_i$ is not so large that the fluctuation is more important than the bare photon number. In other words, the model we are considering is suitable for discussing the quantum nature of a few photons in an optically excited probe system.

In the coherent state of phonons, the number of phonons, as well as energy, is fluctuating. This fluctuation allows incident photons into the probe system to excite phonon fields. When all the phonons are in the vacuum state at time $t = 0$, the excitation probability $P(t)$ that a photon incident on site i in the model system excites the phonon mode p at time t is given by

$$P(t) = 1 - \exp\left\{ 2 \left(\frac{\chi_{ip}}{\Omega_p} \right)^2 \left[\cos(\Omega_p t) - 1 \right] \right\}, \qquad (32.81)$$

where the photon-hopping term is neglected for simplicity. The excitation probability oscillates at a frequency of $2\pi/\Omega_p$ and has the maximum value at $t = \pi/\Omega_p$. The frequencies of the localized phonon modes are higher than those of the delocalized ones, and the localized modes at the earlier time are excited by the incident photons.

Figure 32.4 shows the temporal evolution of the excitation probability $P_{p_0}(t)$ calculated from

$$P_{p_0}(t) = \left[1 - \exp\left\{ 2 \left(\frac{\chi_{ip_0}}{\Omega_{p_0}} \right)^2 \left[\cos(\Omega_{p_0} t) - 1 \right] \right\} \right]$$

$$\times \exp\left\{ \sum_{p \neq p_0} 2 \left(\frac{\chi_{ip}}{\Omega_p} \right)^2 \left[\cos(\Omega_p t) - 1 \right] \right\}, \qquad (32.82)$$

where a specific phonon mode p_0 is excited, while other modes are in the vacuum state. In Figure 32.4, the solid curve represents the probability that a localized phonon mode is excited as the p_0 mode, while the dashed curve illustrates how the lowest phonon mode is excited as the p_0 mode. It follows from the figure that the localized phonon mode is dominantly excited at an earlier time.

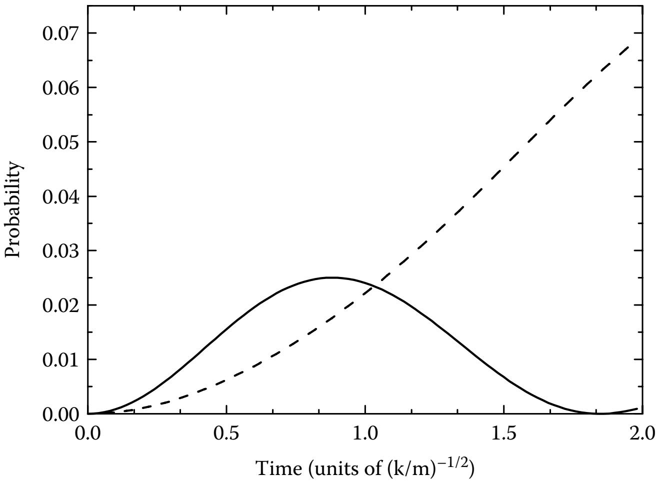

FIGURE 32.4 Temporal evolution of a specific phonon mode. The solid curve represents the probability for a localized phonon mode, while the dashed curve shows that of the lowest delocalized phonon mode.

32.3.4.4 Localization Mechanism of Dressed Photons

In this section, we discuss how phonons contribute to the spatial distribution of photons in the pseudo one-dimensional system under consideration. When there are no interactions between photons and phonons, the frequency and hopping constant are equal at all sites, and thus the spatial distribution of photons is symmetric. It means that no photon localization occurs at any specific site. However, if there are any photon–phonon interactions, spatial inhomogeneity or localization of phonons affects the spatial distribution of photons. On the basis of the Hamiltonian (32.78a), we analyze the contribution from the diagonal (the third term) and off-diagonal (the fourth term) parts of the Hamiltonian (32.78a) in order to investigate the localization mechanism of photons.

32.3.4.4.1 Contribution from the Diagonal Part

Let us rewrite the third term of the Hamiltonian (32.78a) with the mean field approximation as

$$-\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{p=1}^{N}\frac{\hbar\chi_{ip}\chi_{jp}}{\Omega_p}\hat{\alpha}_i^{\dagger}\hat{\alpha}_i\left\langle\hat{\alpha}_j^{\dagger}\hat{\alpha}_j\right\rangle \equiv -\sum_{i=1}^{N}\hbar\omega_i\hat{\alpha}_i^{\dagger}\hat{\alpha}_i, \qquad (32.83a)$$

with

$$\omega_i = \sum_{j=1}^{N}\sum_{p=1}^{N}\frac{\chi_{ip}\chi_{jp}}{\Omega_p}\left\langle\hat{\alpha}_j^{\dagger}\hat{\alpha}_j\right\rangle. \qquad (32.83b)$$

In addition, for the moment, we neglect the site dependence of the hopping operator $\hat{J}_i$, which is approximated as J. Then the Hamiltonian regarding the quasiparticles ($\hat{\alpha}$ and $\hat{\alpha}^{\dagger}$) can be expressed as

$$\hat{H} = \sum_{i=1}^{N}\hbar(\omega-\omega_i)\hat{\alpha}_i^{\dagger}\hat{\alpha}_i + \sum_{i=1}^{N-1}\hbar J\left(\hat{\alpha}_i^{\dagger}\hat{\alpha}_{i+1}+\hat{\alpha}_{i+1}^{\dagger}\hat{\alpha}_i\right), \qquad (32.84)$$

or in matrix form as

$$\hat{H} = \hbar\hat{\vec{\alpha}}^{\dagger}\begin{pmatrix}\omega-\omega_1 & J & \cdots & 0\\ J & \omega-\omega_2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & J\\ 0 & \cdots & J & \omega-\omega_N\end{pmatrix}\hat{\vec{\alpha}} \qquad (32.85a)$$

with

$$\hat{\vec{\alpha}} \equiv \left(\hat{\alpha}_1^{\dagger},\hat{\alpha}_2^{\dagger},\ldots,\hat{\alpha}_N^{\dagger}\right), \qquad (32.85b)$$

where the effect from the phonon fields is involved in the diagonal elements ω_i. Denoting an orthonormal matrix to diagonalize the Hamiltonian (32.85a) as Q and the rth eigenvalue as E_r, we have

$$\hat{H} = \sum_{r=1}^{N}\hbar E_r\hat{A}_r^{\dagger}\hat{A}_r \qquad (32.86a)$$

with

$$\hat{A}_r\sum_{i=1}^{N}(\mathrm{Q}^{-1})_{ri}\hat{\alpha}_i = \sum_{i=1}^{N}\mathrm{Q}_{ir}\hat{\alpha}_i, \qquad (32.86b)$$

and

$$\left[\hat{A}_r,\hat{A}_s^{\dagger}\right] \equiv \hat{A}_r\hat{A}_s^{\dagger}-\hat{A}_s^{\dagger}\hat{A}_r = \delta_{rs}. \qquad (32.86c)$$

Using the above relations (32.86a) through (32.86c), we can write down the time evolution of the photon number operator $\hat{N}_i$ at site i as follows:

$$\begin{aligned}\hat{N}_i(t) &= \exp\left(i\frac{\hat{H}}{\hbar}t\right)\hat{N}_i\exp\left(-i\frac{\hat{H}}{\hbar}t\right)\\ &= \sum_{r=1}^{N}\sum_{s=1}^{N}\mathrm{Q}_{ir}\mathrm{Q}_{is}\hat{A}_r^{\dagger}\hat{A}_s\exp\left\{i(E_r-E_s)t\right\}. \qquad (32.87)\end{aligned}$$

The expectation value of the photon number operator $\hat{N}_i(t)$ is then given by

$$\left\langle N_i(t)\right\rangle_j \equiv \left\langle\psi_j\left|\hat{N}_i(t)\right|\psi_j\right\rangle = \sum_{r=1}^{N}\sum_{s=1}^{N}\mathrm{Q}_{ir}\mathrm{Q}_{jr}\mathrm{Q}_{is}\mathrm{Q}_{js}\cos\left\{(E_r-E_s)t\right\}, \qquad (32.88)$$

in terms of one photon state at site j defined by

$$|\psi_j\rangle = \hat{\alpha}_j^\dagger |0\rangle = \sum_{r=1}^{N} Q_{jr} \hat{A}_r^\dagger |0\rangle. \tag{32.89}$$

Since the photon number operator $\hat{N}_i$ commutes with the Hamiltonian (32.84), the total photon number is conserved, which means that a polariton, called a photon in this section, conserves the total particle number within the lifetime. Moreover, $\langle N_i(t)\rangle_j$ can be regarded as the observation probability of a photon at an arbitrary site i and time t, initially populated at site j. This function is analytically expressed in terms of the Bessel functions $J_{j\pm i}$:

$$\langle N_i(t)\rangle_j = \left\{ J_{j-i}(2Jt) - (-1)^i J_{j+i}(2Jt) \right\}^2, \tag{32.90}$$

when there are no photon–phonon interactions ($\omega_i = 0$) and the total site number N becomes infinite. Here the argument J is the photon hopping constant, and (32.90) shows that a photon initially populated at site j delocalizes to the whole system.

Focusing on the localized phonon modes, we take the summation in (32.83b) over the localized modes only, which means that an earlier stage is considered after the incident photon excites the phonon modes, or that the duration of the localized phonon modes that are dominant over the delocalized modes is focused on (see Figure 32.4). This kind of analysis provides us with an interesting insight into the photon–phonon coupling constant and the photon hopping constant, which is necessary for understanding the mechanism of photon localization.

The temporal evolution of the observation probability of a photon at each site is shown in Figure 32.5. Without the photon–phonon coupling ($\chi = 0$), a photon spreads over the whole system as a result of the photon hopping, as shown in Figure 32.5a. Here the photon energy $\hbar\omega = 1.81$ eV and the hopping constant $\hbar J = 0.5$ eV are used in the calculation. Impurities are assumed to be doped at sites 3, 7, 11, 15, and 19, while the total number of sites N is 20, and the mass ratio of the host molecules to the impurities is 5. Figure 32.5b shows a result with $\chi = 1.4 \times 10^3$ fs^{-1} nm^{-1}; the other parameters used are the same as those in Figure 32.5a. It follows from the figure that a photon moves from one impurity site to another impurity site instead of delocalizing to the whole system. As the photon–phonon coupling constant becomes much larger than $\chi = 1.4 \times 10^3$ fs^{-1} nm^{-1}, a photon cannot move from the initial impurity site to others, but stays there.

The effect due to the photon–phonon coupling constant χ is expressed by the diagonal component in the Hamiltonian, while the off-diagonal component involves the photon hopping effect due to the hopping constant J. The above results indicate that a photon's spatial distribution depends on the competition between the diagonal and off-diagonal components in the Hamiltonian, i.e., χ and J, and that a photon can move among impurity sites and localize at those sites when both components are comparable under the condition

$$\chi \sim N\sqrt{\frac{kJ}{\hbar}}, \tag{32.91}$$

where the localization width seems very narrow.

32.3.4.4.2 Contribution from the Off-Diagonal Part

In the previous subsection, we have approximated J as a constant independent of the sites, in order to examine the photon's spatial distribution as well as the mechanism of the photon localization. Now let us treat the photon hopping operator $\hat{J}_i$ more rigorously, and investigate the site dependence of the off-diagonal contribution, which includes the inhomogeneity of the phonon fields. Noticing that a quasiparticle transformed from a photon

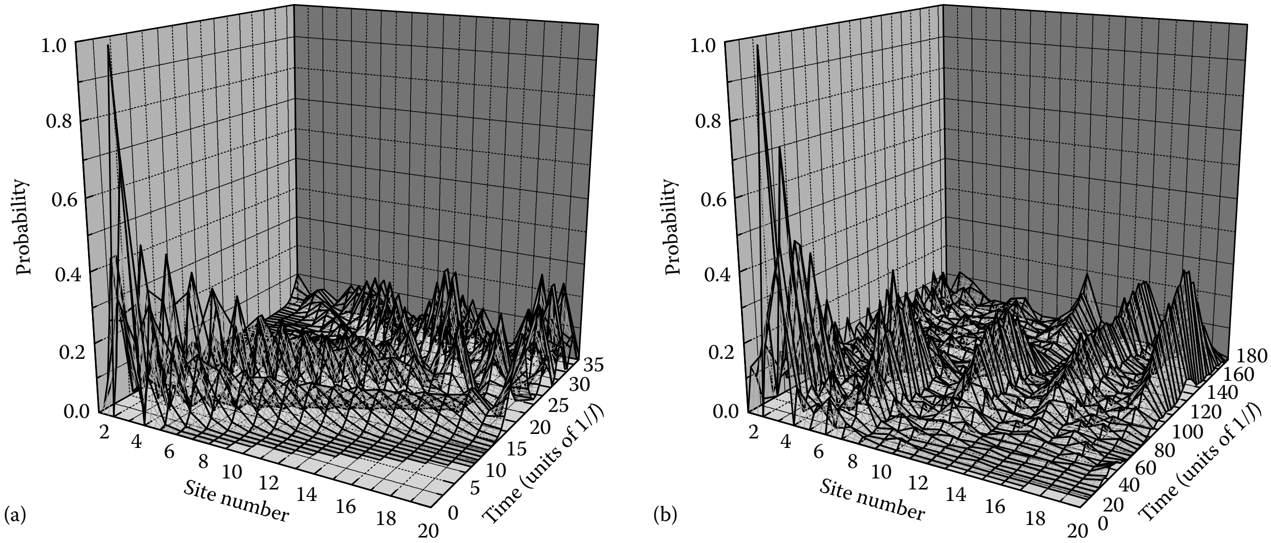

FIGURE 32.5 **(See color insert following page 20-14.)** Probability that a photon is found at each site as a function of time (a) without the photon–phonon coupling, and (b) with the photon–phonon coupling comparable to the photon hopping constant.

operator by the Davydov transformation is associated with phonons in the coherent state (see (32.77a) and (32.77b)), we take expectation values of $\hat{J}_i$ in terms of the coherent state of phonons $|\gamma\rangle$ as

$$J_i \equiv \langle\gamma|\hat{J}_i|\gamma\rangle. \tag{32.92}$$

Here, the coherent state $|\gamma\rangle$ is an eigenstate of the annihilation operator $\hat{b}_p$ with eigenvalue γ_p and satisfies

$$\hat{b}_p|\gamma\rangle = \gamma_p|\gamma\rangle. \tag{32.93}$$

Since the difference between the creation and annihilation operators of a phonon is invariant under the Davydov transformation, the following relation holds:

$$\hat{\beta}_p^\dagger - \hat{\beta}_p = \hat{b}_p^\dagger - \hat{b}_p. \tag{32.94}$$

Using (32.78b), (32.93), and (32.94), we can rewrite the site-dependent hopping constant J_i in (32.92) as

$$\begin{aligned} J_i &= J\langle\gamma|\exp\left\{\sum_{p=1}^{N} C_{ip}\left(b_p^\dagger - b_p\right)\right\}|\gamma\rangle \\ &= J\exp\left(-\frac{1}{2}\sum_{p=1}^{N} C_{ip}^2\right)\langle\gamma|\exp\left(\sum_{p'=1}^{N} C_{ip'}\hat{b}_{p'}^\dagger\right)\exp\left(-\sum_{p''=1}^{N} C_{ip''}\hat{b}_{p''}\right)|\gamma\rangle \\ &= J\exp\left(-\frac{1}{2}\sum_{p=1}^{N} C_{ip}^2\right), \end{aligned} \tag{32.95a}$$

where C_{ip} is denoted by

$$C_{ip} \equiv \frac{\chi_{ip} - \chi_{i+1p}}{\Omega_p}. \tag{32.95b}$$

Figure 32.6 shows the site dependence of J_i in the case of $N = 20$. Impurities are doped at sites 4, 6, 13, and 19. The mass ratio of the host molecules to the impurities is 5, while $\hbar J = 0.5$ eV and $\chi = 14.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$ are used. It follows from the figure that the hopping constants are highly modified around the impurity sites and the edge sites. The result implies that photons are strongly affected by localized phonons and hop to the impurity sites to localize. Here we have not considered the temperature dependence of J_i, which is important for phenomena dominated by incoherent phonons. This is because coherent phonons weakly depend on the temperature of the system. However, there remains room to discuss a more fundamental issue, i.e., whether the probe system is in a thermal equilibrium state or not.

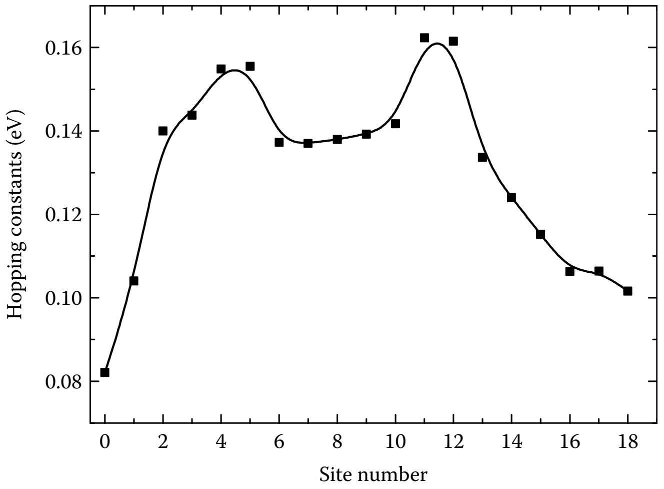

FIGURE 32.6 Site dependence of the hopping constant J_i in the case of $N = 20$. Impurities are doped at sites 4, 6, 13, and 19. The mass ratio of the host molecules to the impurities is 1–0.2, whereas $\hbar J = 0.5$ eV and $\chi = 40.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$ are used.

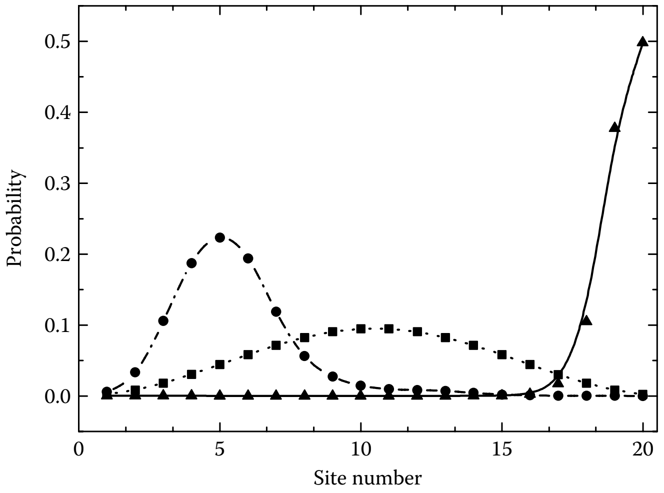

FIGURE 32.7 Probability of photons observed at each site. The filled squares, circles, and triangles represent the results for $\chi = 0$, 40.0, and 54.0 $\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$, respectively. Other parameters are the same as Figure 32.6.

In Figure 32.7, we present a typical result that photons localize around the impurity sites in the system as the photon–phonon coupling constants χ vary from 0 to $40.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$ or $54.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$, while keeping $\hbar J = 0.5$ eV. As depicted by the filled squares in the figure, photons delocalize and spread over the system without the photon–phonon couplings. When the photon–phonon couplings are comparable to the hopping constants, $\chi = 40.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$, photons can localize around the impurity site with a finite width and two sites at half width and half maximum (HWHM), as shown by the filled circles. This finite width of photon localization comes from the site-dependent hopping constants. As the photon–phonon couplings are larger than $\chi = 40.0\,\mathrm{fs}^{-1}\,\mathrm{nm}^{-1}$, photons can localize at the edge sites with a finite width, as well as the impurity sites. In Figure 32.7, the photon localization at the edge site, shown by the filled triangles, originates from the finite size effect of the molecular chain. This kind of localization of photons dressed by the coherent state of phonons leads us to a simple understanding of phonon-assisted photodissociation using an optical near field: molecules in the electronic ground state approach the probe tip within the localization range of the dressed photons, and can be vibrationally excited by the dressed-photon transfer to the molecules, via a multi-phonon component of the dressed photons, which might be followed by electronic excitation. Thus, it leads to the dissociation of the molecules even if an incident photon energy less than the dissociation energy is used.

As a natural extension of the localized photon model, we have discussed the inclusion of the phonon's effects into the model. The study was initially motivated by experiments on the photodissociation of molecules by optical near fields, whose results show unique features different from the conventional one with far fields (Ohtsu et al. 2008). After clarifying whether the vibration modes in a pseudo one-dimensional system are delocalized or localized, we focused on the interaction between dressed photons and phonons by using the Davydov transformation. We have theoretically shown that photons are dressed by the coherent state of phonons, and found that the competition between the photon–phonon coupling constant and the photon hopping constant governs the photon localization or delocalization in space. The obtained results lead us to a simple understanding of an optical near field itself as an interacting system of photon, electronic excitation (induced polarization), and phonon fields in a nanometer space, which are surrounded by macroscopic environments, as well as phonon-assisted photodissociation using an optical near field.

32.4 Summary and Future Perspective

We have briefly outlined the difficulties in defining the position-representation wave function of a photon, followed by several trials to overcome these issues. On the basis of canonical quantization of the electromagnetic fields in free space, we have discussed a dressing mechanism and spatial localization of photons, which is a natural viewpoint from light–matter interactions, or from virtual photon clouds and field correlations. Finally, with the projection operator method, we have shown an effective interaction between nanomaterials electronically disconnected, but closely separated, which are also surrounded by a macroscopic system, in order to detect the virtual clouds. We have discussed the photon dressing by material excitation and pointed out the importance of the phonon's role for spatial localization of photons at the nanoscale.

The pace of development in photonics has accelerated, but most of the underlying science of the field is still Maxwell's classical electromagnetism, not field-quantized photons. In the near future, however, we are anticipating new breakthroughs in nano- and atom photonics, where the localized and quantized nature of photons, as well as an exact quantization formulation for an optical near-field system, including relaxation processes at the nanoscale, will play a critical role.

Acknowledgments

The author is grateful to M. Ohtsu (University of Tokyo) and H. Hori (University of Yamanashi) for stimulating discussions and valuable comments from the early stage of this study. He is also thankful to S. Sangu (Ricoh Co. Ltd.), A. Shojiguchi (NEC Co.), Y. Tanaka (Tokyo Institute of Technology), and A. Sato (Tokyo Institute of Technology) for discussions and collaborations. He greatly acknowledges the valuable guidance given by M. Tsukada (University of Tokyo, emeritus, Tohoku University), K. Kitahara (Tokyo Institute of Technology, emeritus, International Christian University), M. Kitano (Kyoto University), Y. Masumoto (University of Tsukuba), K. Cho (Osaka University, emeritus), and T. Yabuzaki (Kyoto University, emeritus). Finally but not the least, he expresses his gratitude to H. Ito (Tokyo Inst. Technology). T. Kawazoe (University of Tokyo), T. Yatsui (University of Tokyo), T. Saiki (Keio University), K. Matsuda (Kyoto University), H. Nejo (National Institute of Materials Science), M. Naruse (National Institute of Information and Communication Technology), M. Ikezawa (University of Tsukuba), I. Banno (University of Yamanashi), and H. Ishihara (Osaka Prefecture University).

References

Adlard, C., Pike, E. R., and Sarkar, S. 1997. Localization of one-photon states. *Phys. Rev. Lett.* 79: 1585–1587.

Bialynicki-Birula, I. 1998. Exponential localization of photons. *Phys. Rev. Lett.* 80: 5247–5250.

Carniglia, C. K. and Mandel, L. 1971. Quantization of evanescent electromagnetic waves. *Phys. Rev. D* 3: 280–296.

Cohen-Tannoudji, C., Dupont-Roc, I., and Grynberg, G. 1989. *Photons and Atoms*. New York: Wiley.

Cohen-Tannoudji, C., Dupont-Roc, I., and Grynberg, G. 1992. *Atom-Photon Interactions: Basic Processes and Applications*. New York: Wiley.

Compagno, G., Passante, R., and Persico, F. 1988. Dressed and half-dressed neutral sources in nonrelativistic QED. *Phys. Scr*. T21: 33–39.

Compagno, G., Passante, R., and Persico, F. 1995. *Atom-Field Interactions and Dressed Atoms*. Cambridge, NY: Cambridge University Press.

Craig, D. P. and Thirunamachandran, T. 1998. *Molecular Quantum Electrodynamics*. New York: Dover Pub.

Hawton, M. 1999. Photon wave functions in a localized coordinate space basis. *Phys. Rev. A* 59: 3223–3227.

Hegerfeldt, G. 1974. Remarks on causality and particle localization. *Phys. Rev. D* 10: 3320–3321.

Hellwarth, R. W. and Nouchi, P. 1996. Focused one-cycle electromagnetic pulses. *Phys. Rev. E* 54: 889–895.

Inagaki, T. 1998. Physical meaning of the photon wave function. *Phys. Rev. A* 57: 2204–2207.

Inoue, T. and Hori, H. 2001. Quantization of evanescent electromagnetic waves based on detector modes. *Phys. Rev. A* 63: 063805(1)–063805(16).

John, S. and Quang, T. 1995. Photon-hopping conduction and collectively induced transparency in a photonic band gap. *Phys. Rev. A* 52: 4083–4088.

Keller, O. 2005. On the theory of spatial localization of photons. *Phys. Rep*. 411: 1–232.

Kobayashi, K., Sangu, S., Ito, H. et al. 2001. Near-field optical potential for a neutral atom. *Phys. Rev. A* 63: 013806(1)–063806(9).

Kobayashi, K., Tanaka, Y., and Kawazoe, T. et al. 2008. Localized photon model including phonon's degrees of freedom. In *Progress in Nano-Electro-Optics*, vol. VI, ed. M. Ohtsu, pp. 41–66. Berlin, Germany: Springer.

Mandel, L. 1966. Configuration-space photon number operators in quantum optics. *Phys. Rev*. 144: 1071–1077.

Mandel, L. and Wolf, E. 1995. *Optical Coherence and Quantum Optics*. Cambridge, NY: Cambridge University Press.

Newton, T. D. and Wigner, E. P. 1949. Localized states for elementary systems. *Rev. Mod. Phys*. 21: 400–406.

Ohtsu, M. and Hori, H. 1999. *Near-Field Nano-Optics*. New York: Kluwer Academic/Plenum.

Ohtsu, M., Kobayashi, K., Kawazoe, T. et al. 2008. *Principles of Nanophotonics*. Boca Raton, FL: CRC Press.

Pike, E. R. and Sarkar, S. 1987. Spatial dependence of weakly localized single-photon wave packets. *Phys. Rev. A* 35: 926–928.

Pohl, D. W. and Courjon, D. 1993. *Near Field Optics*. Dordrecht, the Netherlands: Kluwer Academic.

Power, E. A. and Thirunamachandran, T. 1993. Quantum electrodynamics with nonrelativistic sources. V. Electromagnetic field correlations and intermolecular interactions between molecules in either ground or excited states. *Phys. Rev. A* 47: 2539–2551.

Roychoudhuri, C. and Roy, R. eds. 2003. The nature of light: What is a photon? *Opn. Trends* 3: S1–S35.

Sakurai, J. J. 1967. *Advanced Quantum Mechanics*. Reading, MA: Addison-Wesley.

Sangu, S., Kobayashi, K., Shojiguchi, A. et al. 2004. Logic and functional operations using a near-field optically coupled quantum-dot system. *Phys. Rev. B* 47: 115334(1)–115334(13).

Scully, M. O. and Zubairy, M. S. 1997. *Quantum Optics*. Cambridge, U.K.: Cambridge University Press.

Shojiguchi, A., Kobayashi, K., Sangu, S. et al. 2003. Superradiance and dipole ordering of an N two-level system interacting with optical near fields. *J. Phys. Soc. Jpn*. 72: 2984–3001.

Sipe, J. E. 1995. Photon wave functions. *Phys. Rev. A* 52: 1875–1883.

Suzuura, H., Tsujikawa, T., and Tokihiro, T. 1996. Quantum theory for exciton polaritons in a real-space representation. *Phys. Rev. B* 53: 1294–1301.

Wightman, A. S. 1962. On the localizability of quantum mechanical systems. *Rev. Mod. Phys*. 34: 845–872.

33

Operations in Nanophotonics

33.1 Introduction 33-1
33.2 Dissipation-Controlled Nanophotonic Devices 33-2
Basic Principles • Nanophotonic Switch • Carrier Manipulation (Up-Converter)
33.3 Nanophotonic Devices Using Spatial Symmetries 33-8
Basic Principles • AND- and XOR-Logic Gates • Nanophotonic Buffer Memory • Manipulation of Quantum-Entangled States
33.4 Summary 33-13
Acknowledgments 33-13
References 33-13

Suguru Sangu
Ricoh Company, Ltd.

Kiyoshi Kobayashi
University of Yamanashi
Japan Science and Technology

33.1 Introduction

Recently, a vast number of studies on nanophotonics have been published in the field of nanoscale optical measurement and bioimaging with super-resolution (Ohtsu 1998, Maheswari et al. 1999, Hosaka and Saiki 2001, Matsuda et al. 2003) and information processing technologies (Biolatti et al. 2000, Rinaldis et al. 2002, Troiani et al. 2002), owing to the unique characteristics of optical near field, which far exceed the technical limitations of conventional optics (Ohtsu et al. 2008). Especially, nanophotonics has shown promise and is expected to be the technology for next-generation nanoscale devices dealing with large amounts of information resources and low energy consumption (Ohtsu et al. 2002). In such devices, optical near field plays an important role in nanophotonic device operations, since the optical near field is a mixed state between photon and matter excitation, not only breaking the diffraction limit that is dependent on the wave nature of light, but also utilizing interesting characteristics inherent in nanophotonics, such as unidirectional energy transfer, optical forbidden transition (Kawazoe et al. 2002), and operations based on coherence of nanometric matters (Sangu et al. 2004), which have not been used in conventional optical devices.

Nanophotonic device is interpreted as a system that consists of several matter systems, localized photon field (optical near field) for driving carriers in the systems, and free photon (radiation) field for extracting some information, as illustrated in Figure 33.1. The important point of nanophotonic system is that the spatial distribution of photons is localized in a nanometric space rather than the matter itself being nanometer-sized. From this point, valuable device operations inconsiderable in conventional optical devices are expected. Relating to a theoretical viewpoint, some restrictions under the long wavelength approximation is allowed in a nanophotonic device system (Cho 2004). Moreover, quantum nature appears in carrier dynamics.

In order to control the carriers, quantum systems with discrete energies, such as quantum dots and molecules, are adapted as matter systems because the processes with matter coherence and with energy dissipation are clearly distinguishable. And also, excitation energies and dissipation dynamics are readily determined by adjusting the sizes of quantum objects. In addition, optical allowed and forbidden transitions for external far-field light are usable to prepare initial excitations. In this chapter, a quantum-dot system is assumed as a matter system of nanophotonic device, and thus the signal carriers correspond to excitons, that is, electron–hole pairs, in the quantum dots. Although such systems resemble a quantum computation device, let us note that a nanophotonic device need not keep quantum coherence in the entire system, and is divided into several quantum coherent parts via dissipative process.

In this chapter, operation principles of typical nanophotonic operations are covered by describing exciton population dynamics theoretically and numerically on the basis of the density matrix formalism (Walls and Milburn 1994, Carmichael 1999, Breuer and Petruccione 2002). Basic formulations for energy transfer, dissipation, and exciton excitation are explained, and then, nanophotonic switch and carrier up-converter are numerically demonstrated in Section 33.2. In Section 33.3, nanophotonic device operations, such as logic gates and memory, are described, in which system coherence and spatial symmetries play important roles. As a summary, design concepts for nanophotonic device systems are discussed in Section 33.4.

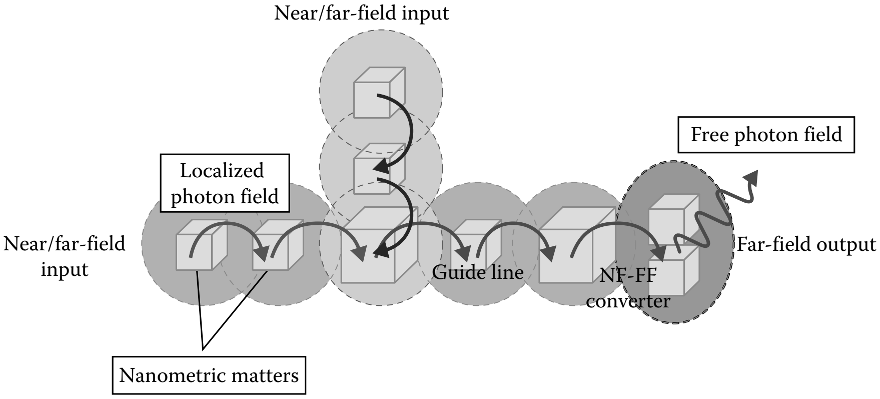

FIGURE 33.1 Schematic representation of a nanophotonic device system, in which carriers are transferred via localized photon field with interaction length in the order of a matter size.

33.2 Dissipation-Controlled Nanophotonic Devices

In this section, nanophotonic devices, where signal carrier in a single input terminal transfers to that in an output one, are formulated and evaluated numerically. Dynamics of signal transfer and control in a nanophotonic device can be described by density matrix formalism. Since quantum mechanical theory expresses an energy-conservative system, dissipation process should be introduced by some sort of approximation because a functional device operation needs to guarantee unidirectional signal transfer. After explanation of basic principles of exciton energy transfer and creation of excitation in a quantum-dot system with dissipation process, some concrete functional operations are explained in the later part of this section.

33.2.1 Basic Principles

33.2.1.1 Energy Transfer between Two Quantum Dots

As the simplest example of carrier dynamics, exciton energy transfer between two quantum dots with radiative and nonradiative relaxations is described by using density matrix formalism (Carmichael 1999, Sangu et al. 2003). Figure 33.2 denotes energy diagram for a two-quantum-dot system, where exciton ground state in QD-A and first exciton excited state in QD-B resonantly couple with each other. This resonant condition can be realized by adjusting the size of quantum dots. For example, in the case of a cubic quantum dot, such as CuCl quantum dot (Masumoto 2002, Kawazoe et al. 2003), the energy with the quantum numbers (n_x, n_y, n_z) is given as

$$E_{(n_x,n_y,n_z)} = \frac{\pi^2\hbar^2}{2Ma^2}\left(n_x^2 + n_y^2 + n_z^2\right) + E_g, \quad (n_x, n_y, n_z = 1, 2, 3, \ldots), \tag{33.1}$$

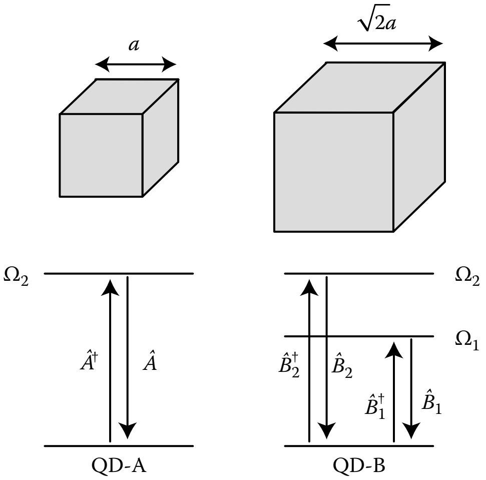

FIGURE 33.2 Energy diagram of a two-quantum-dot system, where the sizes of quantum dots are determined to satisfy the resonant condition, and are set as a (QD-A) and $\sqrt{2}a$ (QD-B). There are exciton sublevels in QD-B because of unidirectional energy transfer. Intra-sublevel relaxation originates from coupling between excitons and phonon reservoir.

where M and E_g denote the exciton mass and bulk exciton energy, respectively. Obviously, when the sizes of QD-A and QD-B are set as a and $\sqrt{2}a$, respectively, the (1, 1, 1)-level in QD-A and (2, 1, 1), (1, 2, 1), and (1, 1, 2)-levels in QD-B are the same energy and, thus, they couple resonantly. In the following, the energy levels for three quantum numbers are dealt with collectively and labeled as $\hbar\Omega_1$- and $\hbar\Omega_2$-levels, since degeneracy of several energy levels is not essential. Energy state of the system is described by the following non-perturbative and interaction Hamiltonians:

$$\begin{aligned} \hat{H}_0 &= \hbar\Omega_2\left(\hat{A}^\dagger\hat{A} + \hat{B}_2^\dagger\hat{B}_2\right) + \hbar\Omega_1\hat{B}_1^\dagger\hat{B}_1, \\ \hat{H}_{\text{int}} &= \hbar U\left(\hat{A}^\dagger\hat{B}_2 + \hat{A}\hat{B}_2^\dagger\right), \end{aligned} \tag{33.2}$$

where

Ω_1 and Ω_2 are eigenfrequencies of QD-A and QD-B

U represents the optical near-field coupling strength, and creation and annihilation operators for excitons are depicted in Figure 33.2

The interaction Hamiltonian $\hat{H}_{\text{int}}$ is well known as the Förster-type interaction (Förster 1965), which is often used by describing intermolecular and inter-quantum-dot interaction. Although the actual interaction between QD-A and QD-B occurs via intermediate virtual states of exciton-polaritons (Kobayashi et al. 2000), which are the coupled state of excitons and photons, such exciton-polariton degree of freedom is neglected in Equation 33.2 as is traced out and renormalized in the coupling strength U.

Exciton dynamics can be expressed by using density operator $\hat{\rho}(t)$, which is a projection operator expanding a certain energy state to appropriate bases depending on possible exciton states in a quantum-dot system. There are a three-exciton state, three two-exciton states, three one-exciton states, and a vacuum state in a two-quantum-dot system, as illustrated in Figure 33.3. Equation of motion of the density operator is given by the Liouville equation in the case without dissipation (Breuer and Petruccione 2002). Dynamics with dissipation is often expressed by so-called Lindblad type notation on the bases of the first-order Born approximation (Carmicheal 1999), in which free photon and phonon fields are considered as energy reservoirs as follows:

$$\frac{\mathrm{d}\hat{\rho}(t)}{\mathrm{d}t} = \frac{1}{i\hbar}\left[\hat{H}_0 + \hat{H}_{\text{int}}, \hat{\rho}(t)\right] + \mathcal{D}_{\text{photon}}\hat{\rho}(t) + \mathcal{D}_{\text{phonon}}\hat{\rho}(t), \tag{33.3}$$

where $\mathcal{D}_{\text{photon (phonon)}}$ is a superoperator defined as

$$\begin{aligned}
\mathcal{D}_{\text{photon}}\hat{\rho}(t) &= \sum_{\alpha=A,B_1} \frac{\gamma_\alpha}{2}\left[2\hat{\alpha}\hat{\rho}(t)\hat{\alpha}^\dagger - \hat{\alpha}^\dagger\hat{\alpha}\hat{\rho}(t) - \hat{\rho}(t)\hat{\alpha}^\dagger\hat{\alpha}\right], \\
\mathcal{D}_{\text{phonon}}\hat{\rho}(t) &= \frac{\Gamma}{2}(n+1) \\
&\quad\times\left[2\hat{B}_2\hat{B}_1^\dagger\hat{\rho}(t)\hat{B}_2^\dagger\hat{B}_1 - \hat{B}_2^\dagger\hat{B}_1\hat{B}_2\hat{B}_1^\dagger\hat{\rho}(t) - \hat{\rho}(t)\hat{B}_2^\dagger\hat{B}_1\hat{B}_2\hat{B}_1^\dagger\right] \\
&\quad+\frac{\Gamma}{2}n\left[2\hat{B}_2^\dagger\hat{B}_1\hat{\rho}(t)\hat{B}_2\hat{B}_1^\dagger - \hat{B}_2\hat{B}_1^\dagger\hat{B}_2^\dagger\hat{B}_1\hat{\rho}(t) - \hat{\rho}(t)\hat{B}_2\hat{B}_1^\dagger\hat{B}_2^\dagger\hat{B}_1\right],
\end{aligned} \tag{33.4}$$

n is the number of phonons in the reservoir, and the number of photons is assumed as zero because of external field being a vacuum state. As you readily understand from the bottom of Equation 33.4, energy transfer between the excited and ground states of an exciton occurs by mediating the intra-sublevel transition with non-radiative relaxation constant Γ, which guarantees unidirectional energy transfer in the two-quantum-dot system.

Since analytical solution of Equation 33.3 is complex and the physical meaning is unclear, the following is discussed on the basis of numerical calculations. For far-field light excitation, excitons in a two-quantum-dot system are created via mixed states of several bases, which are shown in Figure 33.3, because the two quantum dots cannot distinguish each other due to the diffraction limit of light. On the other hand, high spatial resolution of optical near field permits direct access for individual quantum dots. For example, an optical fiber probe possesses sub-100 nm spatial resolution (Hosaka and Saiki 2001,

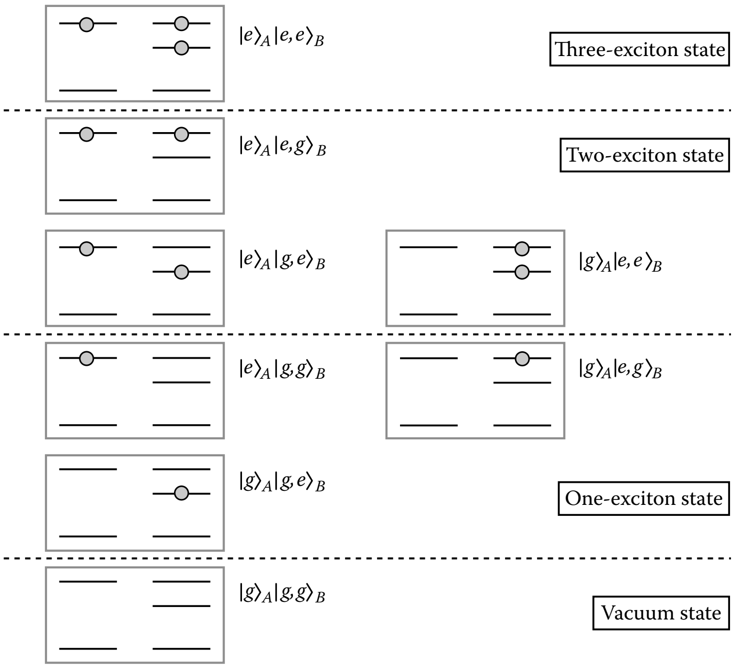
$|e\rangle_A|e,e\rangle_B$
Three-exciton state
$|e\rangle_A|e,g\rangle_B$
Two-exciton state
$|e\rangle_A|g,e\rangle_B$
$|g\rangle_A|e,e\rangle_B$
$|e\rangle_A|g,g\rangle_B$
$|g\rangle_A|e,g\rangle_B$
$|g\rangle_A|g,e\rangle_B$
One-exciton state
$|g\rangle_A|g,g\rangle_B$
Vacuum state

FIGURE 33.3 Appropriate base states in which excitation and ground states for an isolated quantum dot are used. There are eight bases, which include a three-exciton state, three two-exciton states, three one-exciton states, and a vacuum state.

Matsuda et al. 2003) and plasmonic excitation in metallic nanostructures are utilized as an interface in a nanophotonic device (Yatsui et al. 2001, Nomura et al. 2005). As an initial condition to demonstrate energy transfer between the two-quantum-dot systems, an exciton in QD-A is locally prepared that is expressed by density matrix elements of $\rho_{\text{in}}(t) = {}_A\langle e|_B\langle g,g|\hat{\rho}|e\rangle_A|g,g\rangle_B = 1$ and otherwise being zero. Output population is given by the density matrix element of $\rho_{\text{out}}(t) = {}_A\langle g|_B\langle g,e|\hat{\rho}|g\rangle_A|g,e\rangle_B$, which is the lower-level state excited in QD-B, and $(|g, e\rangle_B)^\dagger = {}_B\langle g, e|$. In this case, two- and three-exciton bases decouple from the output population dynamics, and it is enough to consider only four base states for calculating the dynamics.

In the following, numerical solutions in Equation 33.3 are explained, in which two types of typical dynamics appear depending on the optical near-field coupling strength. Figures 33.4 and 33.5 show the exciton population in the input (QD-A) and output (QD-B) energy levels, respectively, when the optical near-field coupling strength are applied as U^{-1} = 10 ps (Figures 33.4a and 33.5a) and 50 ps (Figures 33.4b and 33.5b). For Figure 33.4a, oscillating behavior still remains because the sublevel relaxation constant Γ is the same order of the coupling strength U. While in the case of $\Gamma > U$, coherence in the upper energy levels quickly reduces, and the exciton energy that is stored in the lowest energy level in the system is released as radiative relaxation. The exciton population given in Figure 33.5 is proportional to photoluminescence intensity from the coupled quantum-dot system.

Key point to design of a nanophotonic device is to determine energy transfer paths by using resonant coupling among plural energy states, and internal and external dissipation processes. Density matrix formalism discussed above is a useful tool to solve signal dynamics in an exciton–photon coupled or polariton-mediated system. In addition, multi-exciton process is important for functional operations, such as nanophotonic switch, which are given in Sections 33.2.2 and 33.2.3.

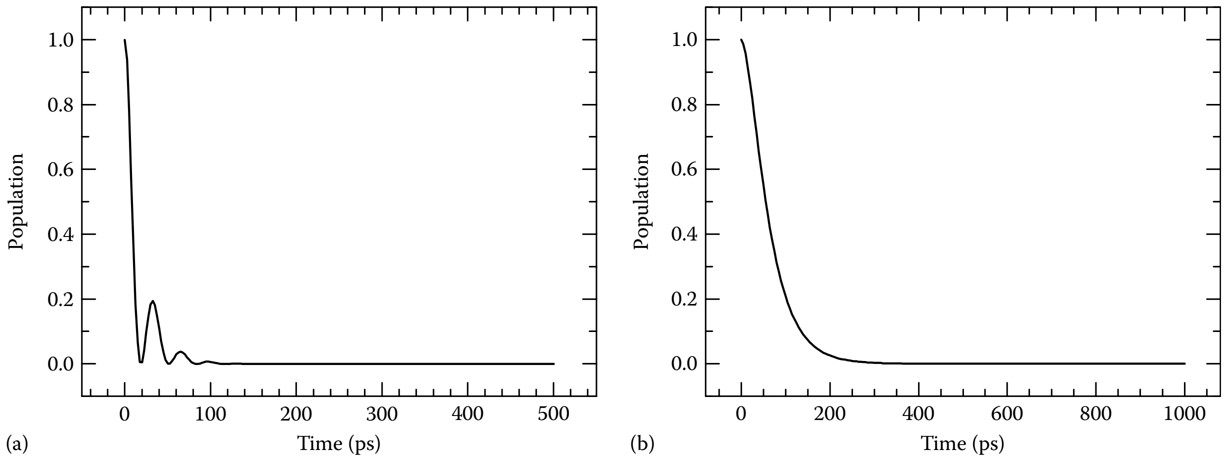

FIGURE 33.4 Temporal evolution of exciton population in the $\hbar\Omega_2$-level in QD-A (input). The optical near-field coupling strength are (a) U^{-1} = 10 ps and (b) 50 ps, and the other parameters are set as $\hbar\Omega_2$ = 3.22 eV ($M = 2.3m_e$, E_g = 3.20 eV, a = 5 nm or 7.07 nm), $\hbar\Omega_1$ = 3.21 eV (a = 7.07 nm), γ_A^{-1} = 1 ns, γ_B^{-1} = 0.5 ns, and Γ^{-1} = 10 ps, where a CuCl quantum cube is assumed as a coupled two-quantum-dot system. For simplicity, zero temperature (n = 0) is assumed.

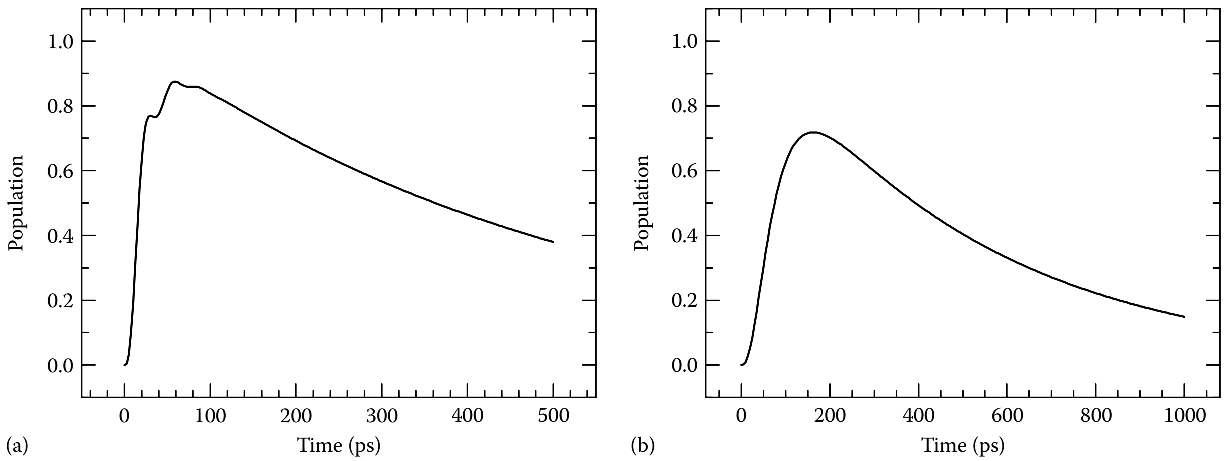

FIGURE 33.5 Temporal evolution of exciton population in the $\hbar\Omega_1$-level in QD-B (output). The optical near-field coupling strength are (a) U^{-1} = 10 ps and (b) 50 ps, and the other parameters are the same as in Figure 33.4.

33.2.1.2 Creation of Excitation by an External Field

Before explanation of multi-exciton process, a way to describe the creation of signal carriers in a quantum-dot system is discussed by means of the density matrix formalism. There are two types of carrier excitations: resonant and nonresonant. In the resonant excitation, energy of external photon field (laser pulse) corresponds to that of the exciton state (Carmicheal 1999), and the equation of motion, which is given in Equation 33.3, is modified with additional contribution of $\hat{H}_r$ as

$$\frac{d\hat{\rho}(t)}{dt} = \frac{1}{i\hbar}\left[\hat{H}_0 + \hat{H}_{\text{int}} + \hat{H}_r, \hat{\rho}(t)\right] + \mathcal{D}_{\text{photon}}\hat{\rho}(t) + \mathcal{D}_{\text{phonon}}\hat{\rho}(t),$$

$$\hat{H}_r = -\hbar g\left(\hat{B}_1^\dagger\langle\hat{a}\rangle + \langle\hat{a}^\dagger\rangle\hat{B}_1\right) \approx -\hbar V(t)\left(\hat{B}_1^\dagger + \hat{B}_1\right), \tag{33.5}$$

where $\hat{a}$ and $\hat{a}^\dagger$ represent annihilation and creation operators of external photon field, and on the basis of the semi-classical approximation, $\langle\hat{a}\rangle = \langle\hat{a}^\dagger\rangle \equiv V(t)/g$. It means that the annihilation and creation operators are renormalized in the parameter (Rabi frequency) $V(t)$. Equation 33.5 expresses the coherent process, which is well known as π-pulse excitation.

While in the case of the nonresonant excitation, coherence between an exciton and a photon is already lost because of interaction among a large number of energy states. This situation appears in such a case that an exciton creates mediated by continuous energy levels and surrounding barrier levels. The formulation of exciton excitation is equivalent to exciton creation from photon reservoir with finite temperature. According to Equation 33.4, the following term is added to Equation 33.3 for nonresonant excitation:

$$\mathcal{P}_{\text{photon}}\hat{\rho}(t) = \sum_{\alpha=A,B_1} \frac{V_\alpha(t)}{2}\left[2\hat{\alpha}^\dagger\hat{\rho}(t)\hat{\alpha} - \hat{\alpha}\hat{\alpha}^\dagger\hat{\rho}(t) - \hat{\rho}(t)\hat{\alpha}\hat{\alpha}^\dagger\right], \tag{33.6}$$

where $V_\alpha(t)$ is pumping rate in the optically allowed state in QD-α.

The above formulations are useful for making initial excited states in a quantum-dot system as well as for applying control signal to select energy transfer paths. In the following, exciton population dynamics for a two-quantum-dot system (Figure 33.2), which is driven by the external photon field, is explained by using numerical results of Equations 33.5 and 33.6. When the bases are restricted less than the one-exciton state, there are four states in this system (See Figure 33.3). Temporal evolution of exciton population in the $\hbar\Omega_2$-levels in QD-A and $\hbar\Omega_1$-level in QD-B are plotted in Figure 33.6a and b, respectively, which is in the case of resonant excitation. The results for nonresonant excitation are also given in Figure 33.7a and b. In both cases, the time width of pumping is set as 50 ps, which is denoted by the vertical dashed lines in Figures 33.6a and 33.7a. Temporal evolutions for resonant and nonresonant excitation are similar with each other except for those in the period when an external field is applied. Since the resonant oscillation keeps matter coherence, the population once increases and then decreases beyond π-pulse excitation. Figures 33.7b and 33.8b represent the population in QD-B, and the temporal profiles are similar in both cases, where the decay profiles are determined by the radiative relaxation constant in the lower energy level in QD-B, but total populations (the area of population curves) reflect the population for the early stage in QD-A.

Resonant and nonresonant exciton creations are important for projecting input signal to a quantum-dot system as well as for controlling signal carrier transiently in a nanophotonics device. Due to the above formulations for energy transfer, dissipation, and creation of excitation, some functional device operations can be demonstrated numerically, which are discussed in Sections 33.2.2 and 33.2.3.

33.2.2 Nanophotonic Switch

General switching device consists of three terminals, which correspond to input, output, and control. Similarly, a three-quantum-dot system can operate as a nanophotonics switch,

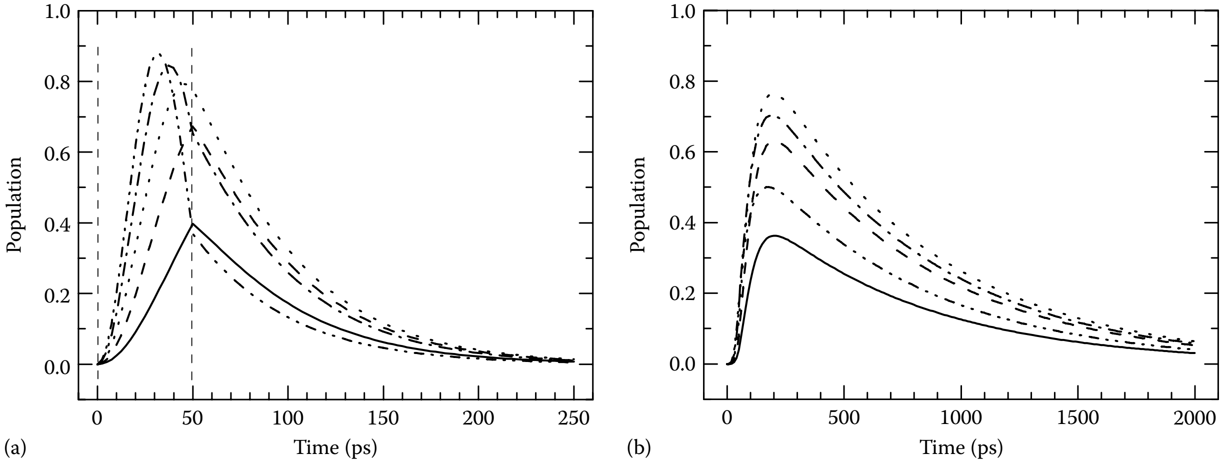

FIGURE 33.6 Temporal evolution of exciton population in (a) $\hbar\Omega_2$-level in QD-A (input terminal) and (b) $\hbar\Omega_1$-level in QD-B (output terminal), where $\hbar\Omega_2$-level in QD-A is resonantly excited. The optical near-field coupling strength is U^{-1} = 50 ps, and the other parameters are the same as in Figure 33.4. The solid, dashed, dotted, and dot-dash, dot-dot-dash curves represent the pulse area of 0.5π, 0.75π, 1.0π, 1.25π, and 1.5π, respectively.

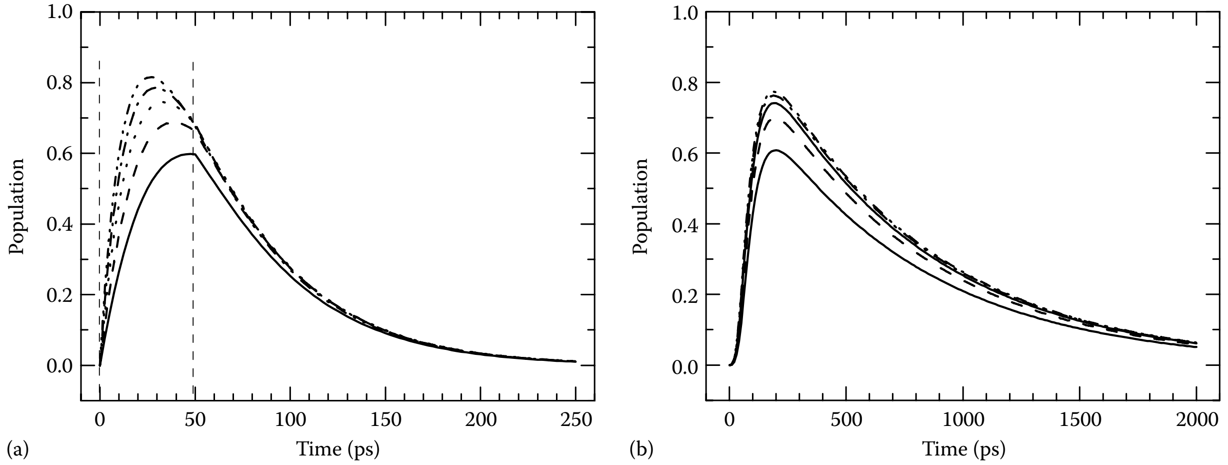

FIGURE 33.7 Temporal evolution of exciton population in (a) $\hbar\Omega_2$-level in QD-A (input port) and (b) $\hbar\Omega_1$-level in QD-B (output port), where $\hbar\Omega_2$-level in QD-A is nonresonantly excited. The parameters are the same as in Figure 33.6. The solid, dashed, dotted, dot-dash, and dot-dot-dash curves represent the pulse area of 0.5π, 0.75π, 1.0π, 1.25π, and 1.5π, respectively.

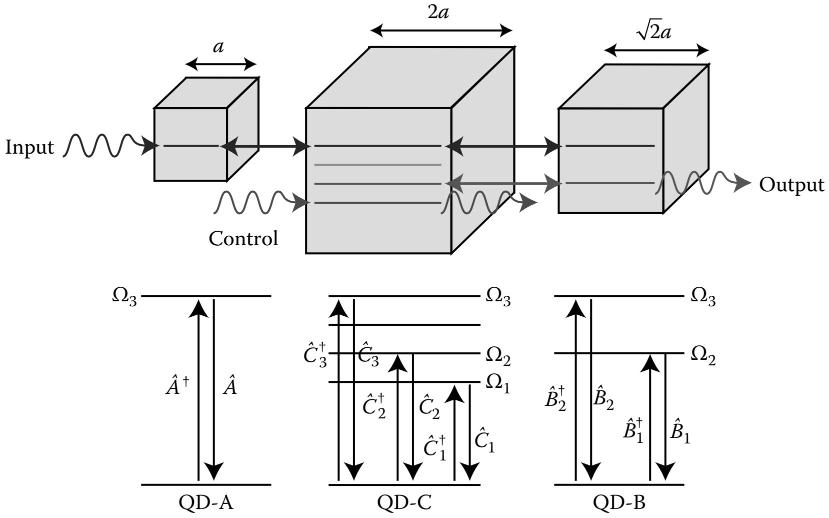

FIGURE 33.8 Energy diagram of a nanophotonic switch that consists of three quantum dots with the size ratio $A:B:C = 1:\sqrt{2}:2$. Input and control signals are injected in the $\hbar\Omega_3$-level in QD-A and the $\hbar\Omega_1$-level in QD-C, respectively, and output signal is detected from the $\hbar\Omega_2$-level in QD-B. By mediating intra-sublevel relaxations, unidirectional energy transfer is guaranteed in this system.

where three quantum dots correspond to input, output, and control terminals (Kawazoe et al. 2003). In Figure 33.8, energy diagram of the three-quantum-dot system is illustrated where the two resonant paths are labeled $\hbar\Omega_2$ and $\hbar\Omega_3$. The resonance conditions can be realized by choosing the quantum-dot sizes as $A:B:C = 1:\sqrt{2}:2$ (see Equation 33.1), where the interaction between QD-A and QD-B is not considered because interdot distance between them is assumed large enough for neglecting the interaction. The principle of the switching operation is as follows: when the control signal is injected in the $\hbar\Omega_1$-level in QD-C, the dissipation path toward the $\hbar\Omega_1$-level is blocked because of Fermionic feature of excitons in a system with discrete energies, an initial exciton in QD-A transfers to the second stable state of $\hbar\Omega_2$-levels in QD-B, and annihilates an exciton with a radiative photon. What is important for this type of functional operation, in which dissipation paths are selectively determined, is to deal with a multi-exciton dynamics in a system.

Equation of motion of the nanophotonic switch can be written in accordance with the manner presented in Section 33.2.1, where the considered bases are extended in two-exciton states, and resonant exciton excitations are applied as an input signal in QD-A and a control signal in QD-C. Since output luminescent intensity is proportional to the $\hbar\Omega_2$-level population in QD-B, the population dynamics in QD-B is discussed in the following.

Figure 33.9 represents the numerical result of one and two-exciton population dynamics. The dashed curve shows the output population in QD-B, when rectangular optical pulse (π-pulse) with the duration of 10 ps is injected in the input terminal of QD-A. The exciton created in QD-A transfers to QD-C

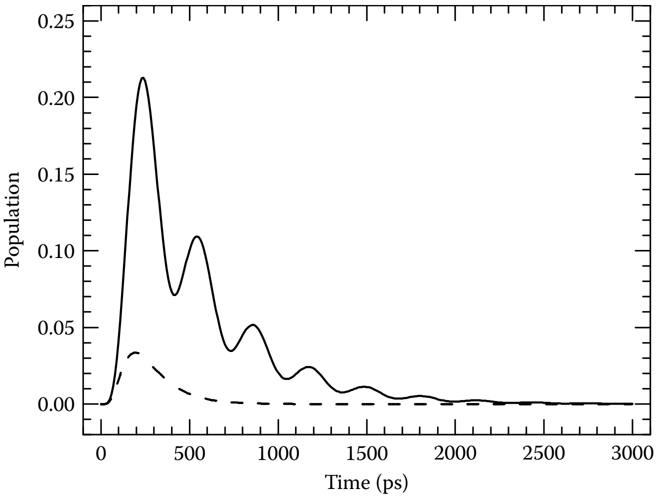

FIGURE 33.9 Temporal evolution of exciton population in the $\hbar\Omega_2$-level in QD-B (output terminal). The dashed and solid curves represent the cases where a single rectangular optical pulse is injected in QD-A (input terminal) and rectangular pulses are injected in the $\hbar\Omega_3$-level in QD-A (input terminal) and the $\hbar\Omega_1$-level in QD-C (control terminal). The parameters are set as follows: optical near-field coupling strength $U_{AC}^{-1} = U_{BC}^{-1} = 50$ ps, pulse duration $\Delta t = 10$ ps (π-pulse amplitude), radiative relaxation time $2^{-3}\gamma_A^{-1} = 2^{-3/2}\gamma_B^{-1} = \gamma_C^{-1} = 1000$ ps, intra-sublevel relaxation time $\Gamma_{B,32}^{-1} = 10$ ps, and $\Gamma_{B,21}^{-1} = \Gamma_{C,21}^{-1} = 20$ ps.

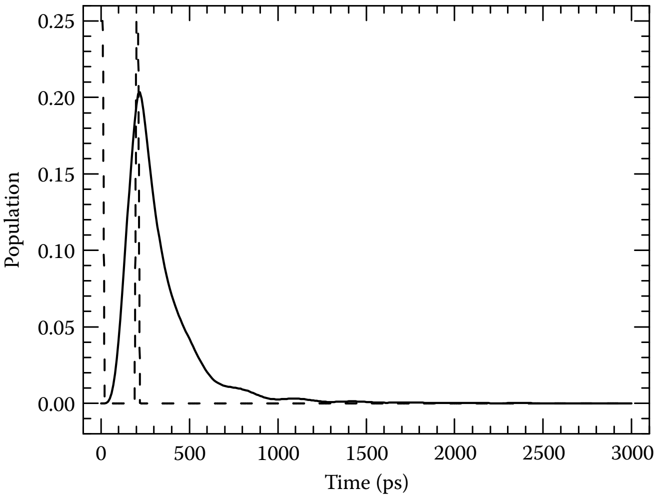

FIGURE 33.10 Temporal evolution of exciton population in the $\hbar\Omega_1$-level in QD-C (output port). The dashed curve represents the profile of the third rectangular optical pulse (π-pulse). The pumping pulse is injected with a delay of 200 ps and pulse duration of 10 ps. The other parameter values set are the same as in Figure 33.9.

via optical near-field coupling U_{AC} and then relaxes to the lowest energy level in QD-C because of fast intra-sublevel relaxation. Therefore, output population in QD-B is detected as very small value (OFF-state). While, the solid curve represents the case that two optical pulses with 10 ps-pulse duration are injected simultaneously in the lowest energy levels in QD-A and QD-C. In the early stage of the dynamics, a somewhat large signal with oscillation appears (ON-state). This is the result of occupation in the $\hbar\Omega_1$-level in QD-C, which is the optical nutation between lower resonant energy levels in QD-B and QD-C. The rise time or switching time is estimated less than 100 ps, which depends on the optical near-field coupling strength (Sangu et al. 2003).

The readers may see that the signal falling time from ON- to OFF-state is quite slower than the rising time from OFF- to ON-state. This is surely constrained by the radiative relaxation time in the control port of QD-C. However, the falling time is controllable by injection of the third control pulse, because π-pulse excitation can remove the exciton occupied in the $\hbar\Omega_1$-level in QD-C. Figure 33.10 represents the same result of exciton population dynamics shown in Figure 33.9, but additional rectangular optical pulse is injected with the delay of 200 ps in the control port of QD-C. The dashed curve represents timing of the control pulse. By injecting the third pulse, the optical nutation completely disappears and the dissipation path toward the lowest energy level in QD-C becomes effective.

Nanophotonic switches and nanophotonic devices, in which dissipation paths are controlled by the state-filling property of excitons, can realize quite low energy operations. Ideally, the energy loss is determined by intra-sublevel relaxation, which is in the order of a few tens of meV. This is a big advantage not only for miniaturizing photonic devices, but also for heat dissipation and wiring in nanometric devices.

33.2.3 Carrier Manipulation (Up-Converter)

Unidirectional energy transfer and switching operation are guaranteed by considering the phonon reservoir system which is coupled with excitons. On the other hand, intense pulse excitation can create the finite phonons in the phonon reservoir system. This leads to blow back an exciton from lower to upper energy levels, instantaneously. Therefore, direction of energy transfer is also controllable by external optical pulse, which has been demonstrated experimentally in the recent publication (Yatsui et al. 2009). This up-conversion dynamics is formulated in a two-quantum-dot system, as shown in Figure 33.2. The intense optical pulse should be dealt with coherent interaction between excitons in the $\hbar\Omega_1$-level in QD-B and external photons. Therefore, the equation of motion described in Equation 33.5 is adopted. Since the atomic lattice in the reservoir system is transiently vibrated by the optical pulse, the phonon number of $n(\omega_0)$ in Equation 33.4 is given as the following function:

$$n(\omega_0) = \frac{1}{\exp\left[\hbar\omega_0 / k_B T(I_{\text{in}})\right] - 1}, \tag{33.7}$$

where ω_0, T, and k_B are the angular frequency of incident light, temperature of the system, and the Boltzmann constant, respectively. Equation 33.7 is the Bose–Einstein distribution function. For simplicity, the temperature is assumed to be proportional to the time integration of the optical pulse, I_{in}. The calculated results of the exciton population dynamics are

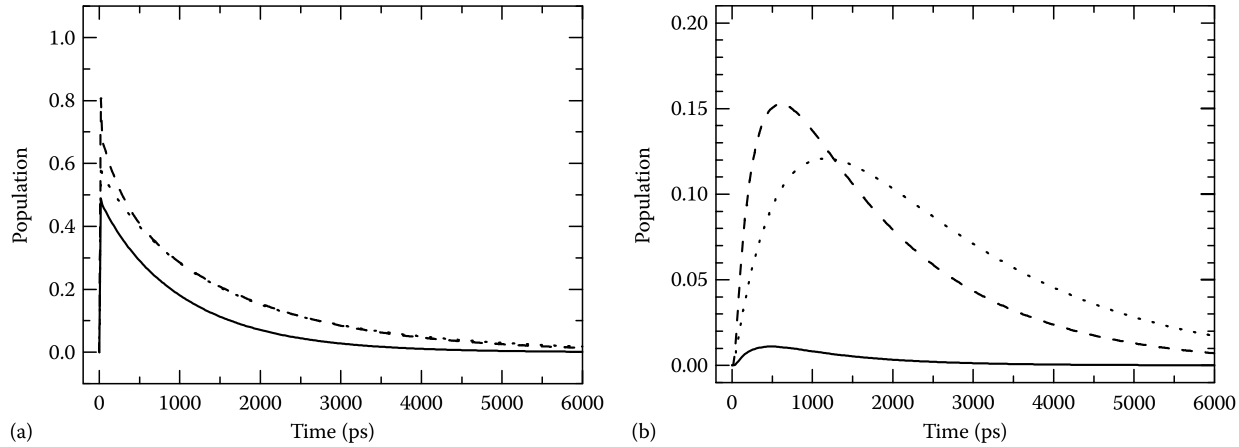

FIGURE 33.11 Temporal evolution of exciton population in (a) the $\hbar\Omega_1$-level in QD-B and (b) the $\hbar\Omega_2$-level in QD-A, when the optical pulse is injected at the time 10 ps. The solid, dashed, and dotted curves represent the pulse area of 0.5π, 1.0π, and 1.5π, respectively. The parameters are set as follows: optical near-field coupling strength $U^{-1} = 50$ ps, pulse duration $\Delta t = 10$ ps (π-pulse amplitude), radiative relaxation time $2^{-3/2}\gamma_A^{-1} = \gamma_B^{-1} = 1000$ ps, and intra-sublevel relaxation time $\Gamma = 10$ ps in QD-B. The phonon number in given by $[\exp(8.0/A^2) - 1]^{-1}$ where A denotes the pulse area.

plotted in Figure 33.11. Figure 33.11a and b represent the populations in the $\hbar\Omega_1$-level in QD-B and in the $\hbar\Omega_2$-level in QD-A which is the up-converted energy level. The solid, dashed, and dotted curves correspond to π/2-pulse, π-pulse, and 3π/2-pulse, respectively. Owing to coherent excitation process by optical pulse injection, the population in the lower energy level in QD-B maximally increases for π-pulse excitation, and decreases for more intense excitation. Temporal evolution in this energy level obeys solely relaxation time of the lower energy level. The population dynamics in QD-A is more complex. The peak values depend on the population in the $\hbar\Omega_1$-level in QD-B for each optical pulse area while the number of phonons in the phonon reservoir monotonically increases depending on the power of the optical pulse. Therefore, population lifetime depicted in the dotted curve in Figure 33.11b becomes longer than the dashed one because of longer pulse injection period. Such an up-conversion signal cannot observe in an isolated quantum dot, since dipole inactive transition from QD-B to QD-A never occurs without optical near-field interaction.

Carrier dynamics driven by both exciton–photon and exciton–phonon interactions has large possibilities for some applications, such as novel light sources, optical near–far field interface devices, efficient photodetectors, and efficient photoelectric conversion devices (Yatsui et al. 2009).

33.3 Nanophotonic Devices Using Spatial Symmetries

In Section 33.2, nanophotonic devices, in which dissipation paths are selected by a state-filling condition in multi-exciton excitation states, were discussed. While another type of device using a coherently-coupled state shows interesting features. For such devices, the number of excitons in a system as well as the spatial symmetry of the system determine the device operations. In this section, some logical operations, memory operations, and operations mediating quantum entangled states are explained with basic theoretical formulations.

33.3.1 Basic Principles

33.3.1.1 Symmetric and Antisymmetric States

When two identical quantum dots sharing an exciton are coupled with each other, as shown in Figure 33.12, appropriate bases for coupled states instead of those for individual quantum dots give a clear view for system dynamics. There are two coupled states via optical near-field interaction that are described by using mixed states between vacuum and exciton states for individual quantum dots, are as follows:

$$\begin{aligned} |S\rangle &= \left(|e\rangle_A|g\rangle_B + |g\rangle_A|e\rangle_B\right)\big/\sqrt{2}, \\ |A\rangle &= \left(|e\rangle_A|g\rangle_B - |g\rangle_A|e\rangle_B\right)\big/\sqrt{2}, \end{aligned} \tag{33.8}$$

where the bases for the coupled states, $|S\rangle$ and $|A\rangle$ are called as symmetric and antisymmetric states as explained later in this section. At the beginning, energies in the symmetric and antisymmetric

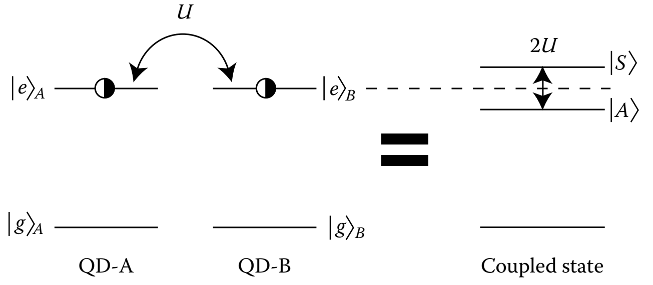

FIGURE 33.12 Energy diagram for two identical quantum dots (left). The system is expressed as a superposition of symmetric and antisymmetric states (right).

states are evaluated. Since the system Hamiltonian in a two-quantum-dot system with optical near-field coupling is given as

$$\begin{aligned}\hat{H} &= \hat{H}_0 + \hat{H}_{\text{int}},\\ \hat{H}_0 &= \hbar\Omega\hat{A}^{\dagger}\hat{A} + \hbar\Omega\hat{B}^{\dagger}\hat{B},\\ \hat{H}_{\text{int}} &= \hbar U\left(\hat{A}^{\dagger}\hat{B} + \hat{A}\hat{B}^{\dagger}\right),\end{aligned} \tag{33.9}$$

where radiative relaxation terms are ignored for simplicity and assuming that the relaxation lifetime is longer than that of optical near-field interaction. The energies for these states are easily derived as

$$\begin{aligned}\langle S|\hat{H}|S\rangle &= \hbar(\Omega + U),\\ \langle A|\hat{H}|A\rangle &= \hbar(\Omega - U).\end{aligned} \tag{33.10}$$

Equation 33.10 explains that the symmetric and antisymmetric states have positive and negative energy shift U from the energy in each quantum dot, respectively. In order to extract carrier excitation from the two identical quantum-dot systems, the third (output) quantum dot is set with resonance energy to the symmetric state or antisymmetric state. Detailed explanation of energy transfer properties in the three-quantum-dot system is given in Section 33.3.1.2.

In this stage, to examine total dipole moment in such a system is quite instructive. The total dipole moment is expressed by an expectation value of a summation of creation and annihilation operators, such as

$$\begin{aligned}\langle S|\mu\sum_{\alpha=A,B}\left(\alpha^{\dagger}+\alpha\right)|g\rangle_A|g\rangle_B &= \sqrt{2}\mu,\\ \langle A|\mu\sum_{\alpha=A,B}\left(\alpha^{\dagger}+\alpha\right)|g\rangle_A|g\rangle_B &= 0,\end{aligned} \tag{33.11}$$

where μ denotes the dipole moment for QD-A and QD-B. Apparently, transition from vacuum state to the antisymmetric state is inactive and, thus, it cannot couple to the external far-field light; this is the reason this state is known as the antisymmetric state. In contrast, optical near field can excite the antisymmetric state because it has a spatial localization character beyond the diffraction limit of light, and can create an exciton in a one-side quantum dot. The state with an exciton is equal to a superposition between symmetric and antisymmetric states.

As described above, if the symmetric and antisymmetric states can be manipulated freely, coupling from near- to far-field light becomes controllable, which is useful for some functional operations, such as nanophotonic buffer memory. Furthermore, optical near-field coupling between the two identical quantum dots and the third output dot creates characteristic behaviors depending on spatial symmetry, where the symmetric and antisymmetric states play important roles, as explained in Section 33.3.1.2.

33.3.1.2 Selective Energy Transfer

For far-field light, excitation in each quantum dot is not distinguishable and radiation is proportional to total dipole moment, which is discussed in Section 33.3.1.1. However, optical near field can resolve the spatial distribution of exciton occupation in a nanometer scale. In order to access the excited states in a two-identical quantum-dot system by using optical near field, third quantum dot, QD-C, is considered as an output port. Figure 33.13 represents the energy transfer from two-identical quantum-dot system to the third quantum dot. The two identical quantum dots make symmetric and antisymmetric states via optical near-field interaction, and these states also couple to the third quantum dot with optical near-field interaction. Here, to evaluate transition moment between two identical quantum dots to the third quantum dot is useful for understanding the effect of spatial alignment of quantum dots. The interaction Hamiltonian is given as

$$\begin{aligned}\hat{H}_{\text{int},AC} &= \hbar U_{AC}\left(\hat{A}^{\dagger}\hat{C} + \hat{C}^{\dagger}\hat{A}\right),\\ \hat{H}_{\text{int},BC} &= \hbar U_{BC}\left(\hat{B}^{\dagger}\hat{C} + \hat{C}^{\dagger}\hat{B}\right),\end{aligned} \tag{33.12}$$

where U_{AC} and U_{BC} represent optical near-field coupling strength between QD-A and QD-C, and QD-B and QD-C, respectively. Using these interaction Hamiltonians, the transition matrix element from symmetric state to excited state in QD-C reads

$${}_A\langle g|_B\langle g|_C\langle e|\left(\hat{H}_{\text{int},AC} + \hat{H}_{\text{int},BC}\right)|S\rangle|g\rangle_C = \frac{\hbar(U_{AC}+U_{BC})}{\sqrt{2}}, \tag{33.13}$$

and that from antisymmetric state to excited state in QD-C is

$${}_A\langle g|_B\langle g|_C\langle e|\left(\hat{H}_{\text{int},AC} + \hat{H}_{\text{int},BC}\right)|A\rangle|g\rangle_C = \frac{\hbar(U_{AC}-U_{BC})}{\sqrt{2}}. \tag{33.14}$$

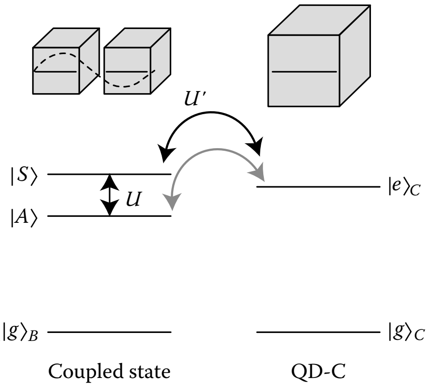

FIGURE 33.13 Schematic representation of energy transfer between two-identical quantum-dot system (input) and a larger quantum dot (output) than the other two. Antisymmetric state becomes dipole-allowed and dipole-forbidden states, depending on spatial symmetry of the output quantum dot.

From Equation 33.14, the transition matrix element can take a zero value when the initial state is an antisymmetric state and the coupling strength toward QD-C are set as $U_{AC} = U_{BC} \neq 0$. This condition can be realized by spatial symmetric alignment of quantum dots. Therefore, energy transfer mediating antisymmetric state is strongly dependent on spatial symmetry in a quantum-dot system.

The feature of selective energy transfer is useful for functional operations of a nanophotonic device, such as logic and memory operations. In the following sections, some functional operations using coherently coupled states, which are symmetric and antisymmetric states, are explained with numerical calculation results.

33.3.2 AND- and XOR-Logic Gates

Similar to Section 33.3.1.2, let us consider a three-quantum-dot system that consists of two identical input quantum dots (QD-A and QD-B) and the third output quantum dot (QD-C) with energy sublevels. QD-C is located in symmetrical position from QD-A and QD-B. Important points for this system are energy transfer reflecting spatial symmetry and difference of resonance conditions in one- and two-exciton states. By using these points, the principle of logic operations has been proposed (Sangu et al. 2004).

First, the case QD-C negatively detuned with energy shift $\hbar U$ is considered. Figure 33.14 represents energy diagram from initial state to final state in this system. The left (initial state) has no excitation in QD-C, and the right (final state) has an exciton in QD-C. Although negative energy shift creates resonance condition between the states of $|A\rangle|g\rangle_C$ and $|g\rangle_A|g\rangle_B|e\rangle_C$ for the one-exciton state, the energy transfer mediated by the antisymmetric state is not permitted in the quantum-dot system that is symmetrically aligned, as explained in Section 33.3.1.2. While, for the two-exciton state, the excited state $|e\rangle_A|e\rangle_B|g\rangle_C$ couples resonantly to the symmetric state $|S\rangle|e\rangle_C$, which is allowed transition. Therefore, this quantum-dot system permits the energy transfer only in the case of two-exciton state. This behavior represents just an AND-logic operation.

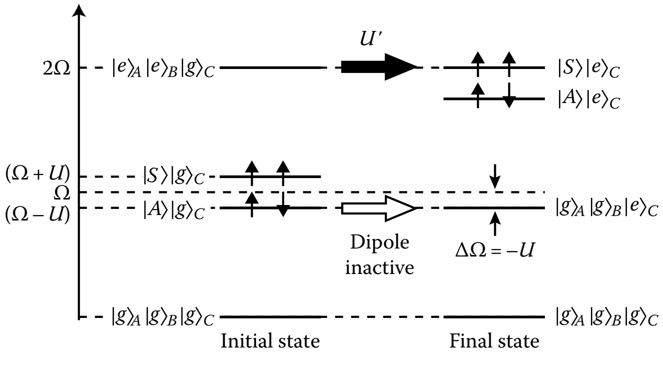
U′
2Ω
$|e\rangle_A|e\rangle_B|g\rangle_C$
$|S\rangle|e\rangle_C$
$|A\rangle|e\rangle_C$
(Ω + U)
Ω
(Ω − U)
$|S\rangle|g\rangle_C$
$|A\rangle|g\rangle_C$
$|g\rangle_A|g\rangle_B|e\rangle_C$
Dipole inactive
ΔΩ = −U
$|g\rangle_A|g\rangle_B|g\rangle_C$
$|g\rangle_A|g\rangle_B|g\rangle_C$
Initial state
Final state

FIGURE 33.14 Energy diagram in three-quantum-dot system, which consists of two identical quantum dots (input port) and a third quantum dot (output port). The third quantum dot is negatively detuned against each input quantum dot. The energy transfer occurs only in the case of two-exciton state.

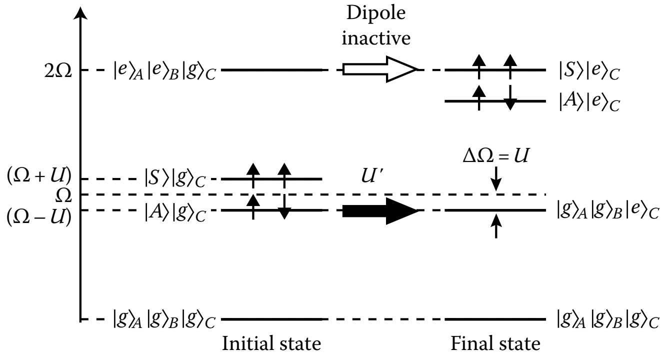
Dipole inactive
2Ω
$|e\rangle_A|e\rangle_B|g\rangle_C$
$|S\rangle|e\rangle_C$
$|A\rangle|e\rangle_C$
ΔΩ = U
(Ω + U)
Ω
(Ω − U)
$|S\rangle|g\rangle_C$
U′
$|A\rangle|g\rangle_C$
$|g\rangle_A|g\rangle_B|e\rangle_C$
$|g\rangle_A|g\rangle_B|g\rangle_C$
$|g\rangle_A|g\rangle_B|g\rangle_C$
Initial state
Final state

FIGURE 33.15 Energy diagram in three-quantum-dot system, which consists of two identical quantum dots (input port) and a third quantum dot (output port). The third quantum dot is positively detuned against each input quantum dot. The energy transfer occurs only in the case of one-exciton state.

Second, the energy level in QD-C is set as positive energy shift $\hbar U$. The energy diagram is illustrated in Figure 33.15. In this case, two-exciton state is inactive because of antisymmetric state, while symmetric one-exciton state $|S\rangle|g\rangle_C$ can resonantly couple to the final state of $|g\rangle_A|g\rangle_B|e\rangle_C$. This means that the energy transfer occurs when either QD-A or QD-B is initially excited, where excitation in a quantum dot in two-identical quantum-dot system is equivalent to excite symmetric and antisymmetric states, simultaneously (see Equation 33.8). This corresponds to the XOR-logic operation.

Temporal evaluation of the above two logic operations can be derived analytically on the basis of the density matrix formalism. The system Hamiltonian is the same as Equation 33.9 except for additional third quantum dot terms, which include non-perturbed energy and non-radiative intra-sublevel relaxation, and radiative relaxation is ignored because optical near-field coupling and sublevel relaxation process are assumed to be fast enough. Final solution for one-exciton state, which is detected in lower energy level in QD-C, is written as follows (Sangu et al. 2004):

$$\rho_{gge,gge}(t) = \frac{1}{2} + \frac{4U'^2}{\omega_+^2 - \omega_-^2} e^{-\frac{\Gamma}{2}t} \times \left[\cos\phi_+ \cos(\omega_+ t + \phi_+) - \cos\phi_- \cos(\omega_- t + \phi_-)\right], \tag{33.15}$$

where $\rho_{gge,gge}(t) \equiv {}_C\langle e|_B\langle g|_A\langle g|\rho(t)|g\rangle_A|g\rangle_B|e\rangle_C$, and the following abbreviations have been used:

$$\phi_\pm = \tan^{-1}\left(\frac{2\omega_\pm}{\Gamma}\right),$$

$$\omega_\pm^2 = \frac{1}{2}\left\{(\Delta\Omega - U)^2 + W_+W_- \pm \sqrt{\left[(\Delta\Omega - U)^2 + W_+^2\right]\left[(\Delta\Omega - U)^2 + W_-^2\right]}\right\},$$

$$W_\pm = 2\sqrt{2}U' \pm \frac{\Gamma}{2}, \tag{33.16}$$

where $\hbar\Delta\Omega$ is the energy shift in QD-C against that in QD-A or QD-B. In Equation 33.15, the second term denotes the energy transfer or nutation between two-identical quantum-dot system QD-A and QD-B, and the third of QD-C. Apparently, this term becomes the maximum value in the case of positive detuning $\Delta\Omega = U$, because the denominator $\omega_+^2 - \omega_-^2$ becomes minimum. This corresponds to dynamics in the XOR-logic operation, as is mentioned in Figure 33.15. The exciton population can reach maximum value of 0.5 not 1. This is because one-side quantum dot in QD-A or QD-B is excited as an initial excitation, which is same as coupled state between symmetric and antisymmetric states, and the antisymmetric state cannot couple to QD-C due to the symmetrically aligned system.

On the other hand, analytical solution for two-exciton state is given in a similar form as

$$\rho_{Se,Se}(t) + \rho_{gg2e,gg2e}(t)$$
$$= 2\left\{\frac{1}{2} + \frac{4U'^2}{\omega_+'^2 - \omega_-'^2} e^{-\frac{\Gamma}{2}t}\left[\cos\phi_+ \cos\left(\omega_+' t + \phi_+'\right) - \cos\phi_- \cos\left(\omega_-' t + \phi_-'\right)\right]\right\}, \tag{33.17}$$

where $\rho_{Se,Se}(t) + \rho_{gg2e,gg2e}(t) \equiv {}_C\langle e|\langle S|\hat{\rho}\,(t)|S\rangle|e\rangle_C + {}_C\langle 2e|_B\langle g|_A\langle g|\rho(t)|g\rangle_A|g\rangle_B|2e\rangle_C$ is defined as the population in the lower level in QD-C and

$$\phi_\pm' = \tan^{-1}\left(\frac{2\omega_\pm'}{\Gamma}\right),$$
$$\omega_\pm'^2 = \frac{1}{2}\left\{\left(\Delta\Omega + U\right)^2 + W_+W_- \pm \sqrt{\left[\left(\Delta\Omega+U\right)^2 + W_+^2\right]\left[\left(\Delta\Omega+U\right)^2 + W_-^2\right]}\right\}. \tag{33.18}$$

From Equations 33.17 and 33.18, negative detuning $\Delta\Omega = -U$ makes resonance condition for two-exciton state, while the negative detuning for one-exciton state is off-resonant to symmetric state (see Equations 33.15 and 33.16). Therefore, the energy transfer rate is quite small. This is the AND-logic operation.

Population dynamics in these devices are plotted in Figure 33.16, where exciton population in QD-C is analytically derived by using Equations 33.15 and 33.17. The solid and dashed curves represent one- and two-exciton states, respectively. Figure 33.16a is the result of the AND-logic gate with negative detuning $\Delta\Omega = -U$, where the strength of optical near-field interaction between QD-A and QD-B, and that between QD-A(B) and QD-C are set as $U^{-1} = 10\,\text{ps}$ and $U'^{-1} = 50\,\text{ps}$. In this operation, the two-exciton state in QD-A and QD-B is resonantly coupled with QD-C, while the one-exciton state is not, because of off-resonant condition. On the other hand, the XOR-logic operation is shown in Figure 33.16b. In this case, the one-exciton state in QD-A and QD-B is resonant to the energy level in QD-C, where initial population is prepared in QD-A, which is same as the coupled state between symmetric and antisymmetric states, and thus, the population reaches only the value of 0.5 because the antisymmetric state is a dark state in a spatially symmetric system. Although ON/OFF ratio is not so high, the XOR-logic gate operation is surely observed.

33.3.3 Nanophotonic Buffer Memory

As mentioned in Section 33.3.1.2, using energy transfer via a dipole inactive state, which can be controlled by energy detuning and quantum dot alignment or spatial symmetry, various functional operations inherent in nanophotonic devices are expected. In this section, one of interesting operations of a nanophotonic buffer memory is introduced. Quantum-dot alignment for realizing nanophotonic buffer memory is illustrated in Figure 33.17, in which two identical quantum dots (QD-A and QD-B) couple

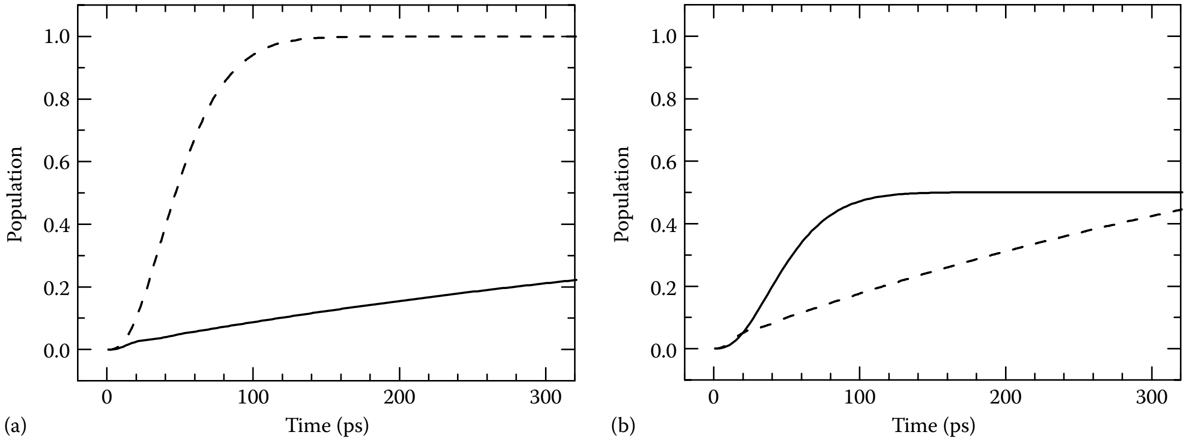

FIGURE 33.16 Temporal evolution of exciton population in three-quantum-dot systems, which consist of input-side two identical quantum dots (QD-A and QD-B) and an output-side quantum dot (QD-C) with intra-sublevel relaxation. (a) and (b) represent the cases where upper energy level is negatively (AND-logic gate) and positively (XOR-logic gate) detuned. The solid and dashed curves correspond to the one-exciton and two-exciton states, respectively. The calculation parameters in Equations 33.15 through 33.18 are set as $U^{-1} = 10\,\text{ps}$, $U'^{-1} = 50\,\text{ps}$, and $\Gamma^{-1} = 10\,\text{ps}$.

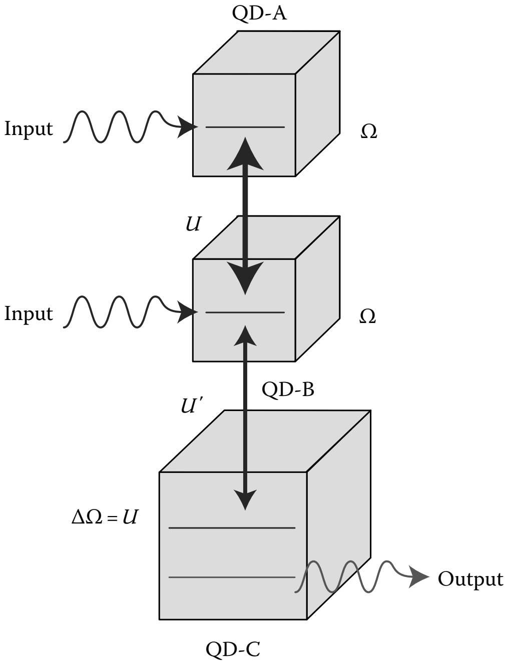

FIGURE 33.17 Schematic representation of nanophotonic buffer memory, which consists of two identical quantum dots, and the larger quantum dot with positive detuning $\Delta\Omega = U$. The three quantum dots are aligned maximally asymmetric and, thus, the coupling strength between QD-A and QD-C can be neglected.

to the third quantum dot (QD-C) with maximal asymmetric position. The third quantum dot is detuned positively with the energy $\hbar(\Omega + U)$, that is, $\Delta\Omega = U$, which has the same energy diagram as in Figure 33.15, but energy transfer via antisymmetric state is allowed. When both QD-A and QD-B are excited as an initial condition, the two-exciton state in the input side resonantly couples to antisymmetric state with excitation in QD-C and then relaxes to the one-exciton antisymmetric state. This antisymmetric state decouples with outer far-field light because of dipole inactive (dark) state. This is just buffer memory operation.

Dynamics of buffer memory operation is numerically evaluated by using density matrix formalism, as explained above. The calculation results are plotted in Figure 33.18, where the vertical axis represents population relating to the two-exciton coupled states, and the coupling strengths via optical near field are set as $U^{-1} = 10\,\text{ps}$ for QD-A and QD-B, and $U'^{-1} = 50\,\text{ps}$ for QD-B and QD-C (QD-A and QD-C are decoupled). The solid and dashed curves represent antisymmetric and symmetric states, respectively. In this calculation, radiative relaxation is ignored because the radiative relaxation time is longer than exciton population dynamics via optical near field and, thus, two-exciton antisymmetric state becomes the final state. Actually, the two-exciton state decays into the one-exciton antisymmetric state with a photon radiation from QD-C, and an exciton is retained in the system.

In Section 33.3.2, the dark state is used for suppressing energy transfer, leading to logical operations, while it is used to create excitation in a dark state in this section, which enables us to control far- and near-field conversion. It is one of the characteristic features of nanophotonic devices.

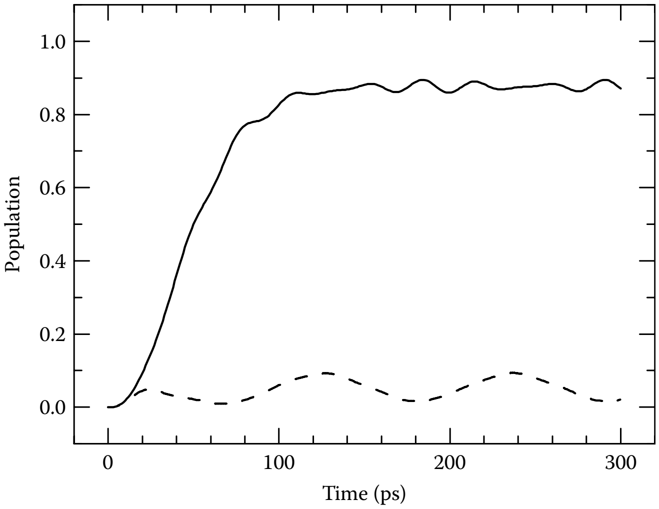

FIGURE 33.18 Temporal evolution of exciton population in three-quantum-dot systems, which have maximal asymmetric configuration (see Figure 33.17). The solid and dashed curves represent the populations for the antisymmetric two-exciton state and the symmetric two-exciton state, respectively. The antisymmetric state has a dipole inactive nature and, thus, the excitation is maintained in the three-quantum-dot system. The calculation parameters are set as $U^{-1} = 10\,\text{ps}$, $U'^{-1} = 50\,\text{ps}$, and $\Gamma^{-1} = 10\,\text{ps}$.

33.3.4 Manipulation of Quantum-Entangled States

As mentioned in Sections 33.3.1 through 33.3.3, a nanophotonic device system consists of coherent energy transfer process and dissipation process. Since the coherent energy transfer process maintains quantum coherence, this system can be used for quantum information processing as well as interface between quantum and classical computations. As an instructive example, a typical device that identifies quantum mixed states is numerically demonstrated in this section. Figure 33.19 schematically shows spatial alignment of quantum dots, where two identical quantum dots couple resonantly, and the excitations are led to two output quantum dots with positive (QD-C) and negative (QD-D) detuning. Both QD-C and QD-D are set at symmetric and asymmetric positions, respectively. Apparently, QD-C resonantly couples to the symmetric state in the input quantum dots, and QD-D couples to the antisymmetric state. Therefore, this device operates that input signal with arbitral quantum state divides two quantum bases of symmetric and antisymmetric states into two different output signals with different optical frequencies as well as different positions in a nanometric space. Temporal evolution of an input quantum mixed state is numerically calculated in Figure 33.20. When an input state $c_1|S\rangle + c_2|A\rangle$ is injected, where one-exciton state is assumed, the output exciton population in QD-C and QD-D

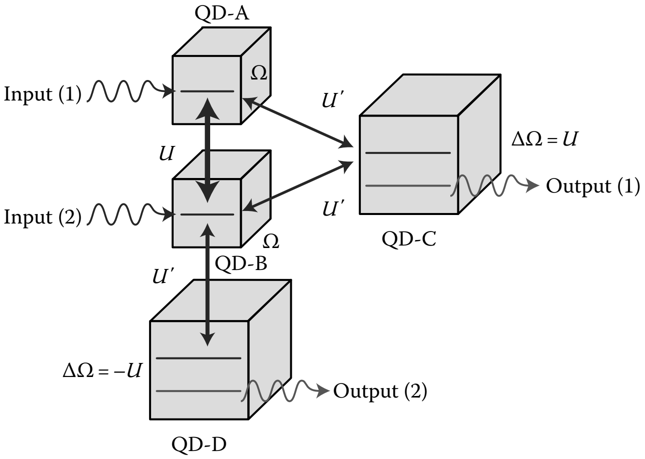

FIGURE 33.19 Spatial arrangement of a four-quantum-dot system, which has symmetric and asymmetric configurations simultaneously. QD-C resonantly couples with symmetric state in QD-A and QD-B, and QD-D couples with antisymmetric state.

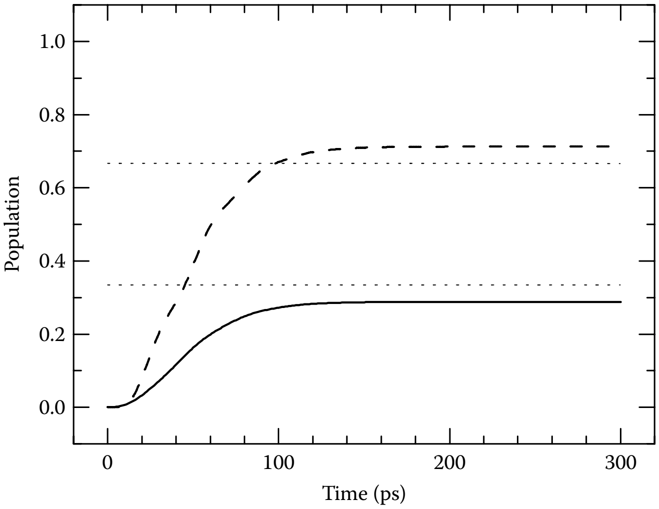

FIGURE 33.20 Population dynamics in four-quantum-dot systems that have two identical quantum dots (QD-A and QD-B) for input ports of an arbitral quantum state, and output quantum dots (QD-C and QD-D) for identifying the input quantum states. The solid and dashed curves represent the output population of QD-C and QD-D, respectively, and the dotted lines correspond to the coefficients of an initial quantum mixed state. The calculation parameters in Equations 33.15 through 33.18 are set as $U^{-1} = 10\,\text{ps}$, $U'^{-1} = 50\,\text{ps}$, and $\Gamma^{-1} = 10\,\text{ps}$.

reflects the coefficients of c_1 and c_2 of an input state. In Figure 33.20, the coefficients of initial quantum mixed state are set as $c_1 = 1/3$ and $c_2 = 2/3$, and thus, the output populations approach these values asymptotically.

Although general quantum information processing devices are built upon quantum coherent states in whole device systems, nanophotonic devices are characterized by the control of dissipation processes, but keeping the quantum coherence locally. Such a mixed operation of classical and quantum information processing may open up future device technologies.

33.4 Summary

In this chapter, operation of a nanophotonic device, in which characteristics of the light localized among nanometric objects, that is, optical near fields, are positively used, has been demonstrated theoretically and numerically. Resonant energy transfer between quantum systems mediated by optical near-field interaction and fast energy relaxation in intra-energy levels via exciton–phonon interaction are fundamental for a switching device and excitation carrier manipulations, as discussed in Section 33.2. By using coupled states, which are made from resonant optical near-field interaction between quantum systems, nanophotonic logic-gate, memory, and quantum-classical interface devices can be realized with the help of selective energy transfer depending on energies of the coupled state and a selection rule for optical near and far fields. These nanophotonic devices are taken as extensions of classical devices, in which simple functions are tandemly arranged, to novel-type devices locally utilized the quantum nature. Here, the optical near fields play important roles for controlling quantum coherence and/or decoherence.

As a summary, key fundamental design concepts for nanophotonic inherent operations are enumerated in the following:

- Externally switchable dissipation paths
- Dependence of the number of excitation carriers
- Selection rule for optical near field based on spatial arrangement of quantum systems
- Selection rule for optical far field based on total dipole moment

The above may not explain all concepts for designing nanophotonic devices—additional unique natures may be hidden. Accomplished device systems might be realized, which have possible advantages such as low energy consumption, low heat liberation, large parallelism, miniaturization, and environmental tolerance, by comprehending physical phenomena inherent in nanophotonics correctly and designing nanophotonic devices from a standpoint throughout from nano- to macro-scales.

Acknowledgments

The authors are grateful to M. Ohtsu, T. Kawazoe, and T. Yatsui from the University of Tokyo for fruitful discussions. This work was mainly carried out at the project of ERATO, Japan Science and Technology Agency, from 1998 to 2003. The authors would like to thank the persons concerned.

References

Biolatti, E., Iotti, R. C., Zanardi, P., and Rossi, F. 2000. Quantum information processing with semiconductor macroatoms. *Phys. Rev. Lett.* 85: 5647–5650.

Breuer, H.-P. and Petruccione, F. 2002. *The Theory of Open Quantum Systems*. New York: Oxford University Press.

Carmichael, H. J. 1999. *Statistical Methods in Quantum Optics 1*. Berlin/Heidelberg, Germany: Springer-Verlag.

Cho, K. 2004. *Optical Response of Nanostructures: Microscopic Nonlocal Theory (Springer Series in Solid-State Sciences)*. Tokyo, Japan: Springer-Verlag.

Förster, T. 1965. Delocalized excitation and excitation transfer. In *Modern Quantum Chemistry*. O. Sinanoglu (Ed.), pp. 93–137. London, U.K.: Academic Press.

Hosaka, N. and Saiki, T. 2001. Near-field fluorescence imaging of single molecules with a resolution in the range of 10 nm. *J. Microsc*. 202: 362–364.

Kawazoe, T., Kobayashi, K., Lim, J., Narita, Y., and Ohtsu, M. 2002. Direct observation of optically forbidden energy transfer between CuCl quantum cubes via near-field optical spectroscopy. *Phys. Rev. Lett*. 88: 067404-1–067404-4.

Kawazoe, T., Kobayashi, K., Sangu, S., and Ohtsu, M. 2003. Demonstration of a nanophotonic switching operation by optical near-field energy transfer. *Appl. Phys. Lett*. 82: 2957–2959.

Kobayashi, K., Sangu, S., Ito H., and Ohtsu, M. 2000. Near-field optical potential for a neutral atom. *Phys. Rev. A* 63: 013806-1–013806-9.

Maheswari, R. U., Mononobe, S., Yoshida, K., Yoshimoto, M., and Ohtsu, M. 1999. Nanometer level resolving near field optical microscope under optical feedback in the observation of a single-string deoxyribo nucleic acid. *Jpn. J. Appl. Phys*. 38: 6713–6720.

Masumoto, Y. 2002. Persistent spectral hole burning in semiconductor quantum dots. In *Semiconductor Quantum Dots: Physics, Spectroscopy and Applications*. Y. Masumoto and T. Takagahara (Eds.), pp. 209–244. Berlin/Heidelberg, Germany: Springer-Verlag.

Matsuda, K., Saiki, T., Nomura, S., Mihara, M., Aoyagi, Y., Nair, S., and Takagahara, T. 2003. Near-field optical mapping of exciton wave functions in a GaAs quantum dot. *Phys. Rev. Lett*. 91: 177401–177404.

Nomura, W., Ohtsu, M., and Yatsui, T. 2005. Nanodot coupler with a surface plasmon polariton condenser for optical far/near-field conversion. *Appl. Phys. Lett*. 86: 181108-1–181108-3.

Ohtsu, M. 1998. *Near-Field Nano/Atom Optics and Technology*. Tokyo, Japan: Springer-Verlag.

Ohtsu, M., Kobayashi, K., Kawazoe, T., Sangu, S., and Yatsui, T. 2002. Nanophotonics: Design, fabrication, and operation of nanometric devices using optical near fields. *IEEE J. Sel. Top. Quantum Electron*. 8: 839–862.

Ohtsu, M., Kobayashi, K., Kawazoe, T., Yatsui, T., and Naruse, M. 2008. *Principles of Nanophotonics*. London, U.K.: Taylor & Francis.

Rinaldis, S. D., D'Amico, I., and Rossi, F. 2002. Exciton–exciton interaction engineering in coupled GaN quantum dots. *Appl. Phys. Lett*. 81: 4236–4238.

Sangu, S., Kobayashi, K., Kawazoe, T., Shojiguchi, A., and Ohtsu, M. 2003. Excitation energy transfer and population dynamics in a quantum dot system induced by optical near-field interaction. *J. Appl. Phys*. 93: 2937–2945.

Sangu, S., Kobayashi, K., Shojiguchi, A., and Ohtsu. M., 2004. Logic and functional operations using a near-field optically coupled quantum-dot system. *Phys. Rev. B* 69, 115334-1–115334-13.

Troiani, F., Hohenester, U., and Molinari, E. 2002. Electron–hole localization in coupled quantum dots. *Phys. Rev. B* 65: 161301-1–161301-4.

Walls, D. F. and Milburn, G. J. 1994. *Quantum Optics*. Berlin/Heidelberg, Germany: Springer-Verlag.

Yatsui, T., Kourogi, M., and Ohtsu, M. 2001. Plasmon waveguide for optical far/near-field conversion. *Appl. Phys. Lett*. 79: 4583–4585.

Yatsui, T., Sangu, S., Kobayashi, K., Kawazoe, T. et al. 2009. Nanophotonic energy up-conversion using ZnO nanorod double-quantum-well structures. *Appl. Phys. Lett*. 94: 083113-1–083113-3.

34

System Architectures for Nanophotonics

34.1 Introduction **34**-1
34.2 System Architectures Based on Optical Excitation Transfer **34**-1
Optical Excitation Transfer via Optical Near-Field Interactions and Its Functional Features • Parallel Architecture Using Optical Excitation Transfer • Secure Signal Transfer in Nanophotonics
34.3 Hierarchical Architectures in Nanophotonics **34**-6
Physical Hierarchy in Nanophotonics and Functional Hierarchy • Hierarchical Memory Retrieval • Design of *Unscalable* Hierarchical Response • Versatile Functionalities Based on Hierarchy in Optical Near-Fields
34.4 Summary **34**-12
References **34**-12

Makoto Naruse
National Institute of Information and Communications Technology
The University of Tokyo

34.1 Introduction

To accommodate the continuously growing amount of digital data and qualitatively new requirements demanded by industry and people in society, optics is expected to be highly integrated and to play a wider role in enhancing system performance. However, many technological difficulties remain to be in overcome in adopting optical technologies in critical information and communication systems; one problem is the poor integrability of optical hardware due to the diffraction limit of light (Pohl and Courjon 1993, Ohtsu and Hori 1999).

Nanophotonics, on the other hand, which is based on local interactions between nanometer-scale matters via optical near fields, offers ultrahigh-density integration since it is not constrained by the diffraction limit. Fundamental nanophotonic processes, such as optical excitation transfer via optical near fields between nanometer-scale matters, have been studied in detail (Ohtsu et al. 2002, 2008, Maier et al. 2003). Moreover, this higher integration density is not the only advantage that optical near fields have over conventional optics and electronics. From a system architectural perspective, nanophotonics drastically changes the fundamental design rules of functional optical systems, and suitable architectures may be built to exploit this. As a result, it also gives qualitatively strong impacts on information and communication systems.

This chapter discusses system architecture for nanophotonics considering the unique physical principles of optical near-field interactions as well as their experimental verification based on technological vehicles, such as quantum dots and engineered metal nanostructures. In particular, two unique physical processes in light–matter interactions in the nanometer scale are exploited. One is optical excitation transfer via optical near-field interactions and the other is the hierarchical property in optical near-field interactions, which are explained in Sections 34.2 and 34.3, respectively.

The overall structure of this chapter is outlined in Figure 34.1. Through these architectural and physical insights, nanophotonic information and communication systems are demonstrated, which overcome the integration-density limit imposed by the diffraction of light with ultralow-power operation, as well as provide unique functionalities that are only achievable using optical near-field interactions.

34.2 System Architectures Based on Optical Excitation Transfer

34.2.1 Optical Excitation Transfer via Optical Near-Field Interactions and Its Functional Features

In this section, optical excitation transfer processes involving optical near-field interactions are reviewed from a system perspective. Here, their fundamental principles are first briefly reviewed and their functional features are introduced for later discussion.

The interaction Hamiltonian between an electron and an electric field is given by

$$\hat{H}_{\text{int}} = -\int \hat{\psi}^{\dagger}(\vec{r})\vec{\mu}\hat{\psi}(\vec{r}) \cdot \hat{\vec{D}}(\vec{r})d\vec{r}, \tag{34.1}$$

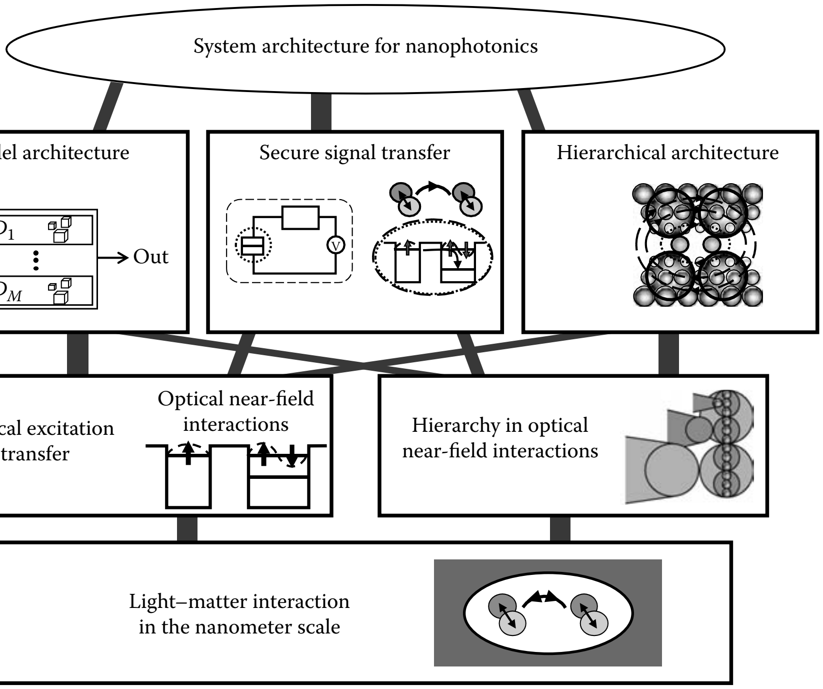
System architecture for nanophotonics
Parallel architecture
In
D_1
D_M
Out
Secure signal transfer
V
Hierarchical architecture
Optical excitation transfer
Optical near-field interactions
Hierarchy in optical near-field interactions
Light–matter interaction in the nanometer scale

FIGURE 34.1 Overview of this section: from light–matter interactions on the nanometer scale to system architectures for nanophotonics.

where

$\vec{\mu}$ is a dipole moment

$\hat{\psi}^{\dagger}(\vec{r})$ and $\hat{\psi}(\vec{r})$ are, respectively, creation and annihilation operators of an electron at $\vec{r}$

$\hat{\vec{D}}(\vec{r})$ is the operator of electric flux density

In usual light–matter interactions, the operator $\hat{\vec{D}}(\vec{r})$ is a constant since the electric field of propagating light is considered to be constant on the nanometer scale. Therefore, as is well known, one can derive optical selection rules by calculating a transfer matrix of an electrical dipole. As a consequence, in the case of cubic quantum dots, for instance, transitions to states described by quantum numbers containing an even number are prohibited. In the case of optical near-field interactions, on the other hand, due to the steep electric field of optical near fields in the vicinity of nanoscale matter, an optical transition that violates conventional optical selection rules is allowed.

Optical excitations in nanostructures, such as quantum dots, can be transferred to neighboring ones via optical near-field interactions (Ohtsu et al. 2002, 2008). For instance, assume two cubic quantum dots, QD_A and QD_B, whose side lengths L are a and $\sqrt{2}a$, respectively (see Figure 34.2a). Suppose that the energy eigenvalues for the quantized exciton energy level specified by quantum numbers (n_x, n_y, n_z) in a QD with side length L are given by

$$E_{(n_x,n_y,n_z)} = E_B + \frac{\hbar^2\pi^2}{2ML^2}(n_x^2 + n_y^2 + n_z^2), \qquad (34.2)$$

where

E_B is the energy of the bulk exciton

M is the effective mass of the exciton

According to Equation 34.2, there exists a resonance between the level of quantum number (1, 1, 1) for QD_A and that of quantum number (2, 1, 1) for QD_B. There is an optical near-field interaction, which is denoted by U, due to the steep electric field in the vicinity of QD_A. Therefore, excitons in QD_A can move to the (2, 1, 1)-level in QD_B. Note that such a transfer is prohibited for propagating light since the (2, 1, 1)-level in QD_B contains an even number (Tang et al. 1993). In QD_B, the exciton sees a sublevel energy relaxation, denoted by Γ, which is faster than the near-field interaction, and so the exciton goes to the (1, 1, 1)-level of QD_B. It should be emphasized that the sublevel relaxation determines the unidirectional exciton transfer from QD_A to QD_B.

Now, several unique functional aspects should be noted in the above excitation transfer processes. First, as already mentioned, the transition from the (1, 1, 1)-level in QD_A to the (2, 1, 1)-level in QD_B is usually a dipole-forbidden transfer. In contrast, the optical near field allows such processes. Second, in the resonant energy levels of those quantum dots, optical excitation can go back and forth between QD_A and QD_B, which is called optical nutation. The direction of excitations is determined by the energy dissipation processes. Therefore, based on the above mechanisms, the flow of optical excitations can be controlled in quantum-dot systems via optical near-field interactions.

From an architectural standpoint, such a flow of excitations directly leads to digital processing systems and computational

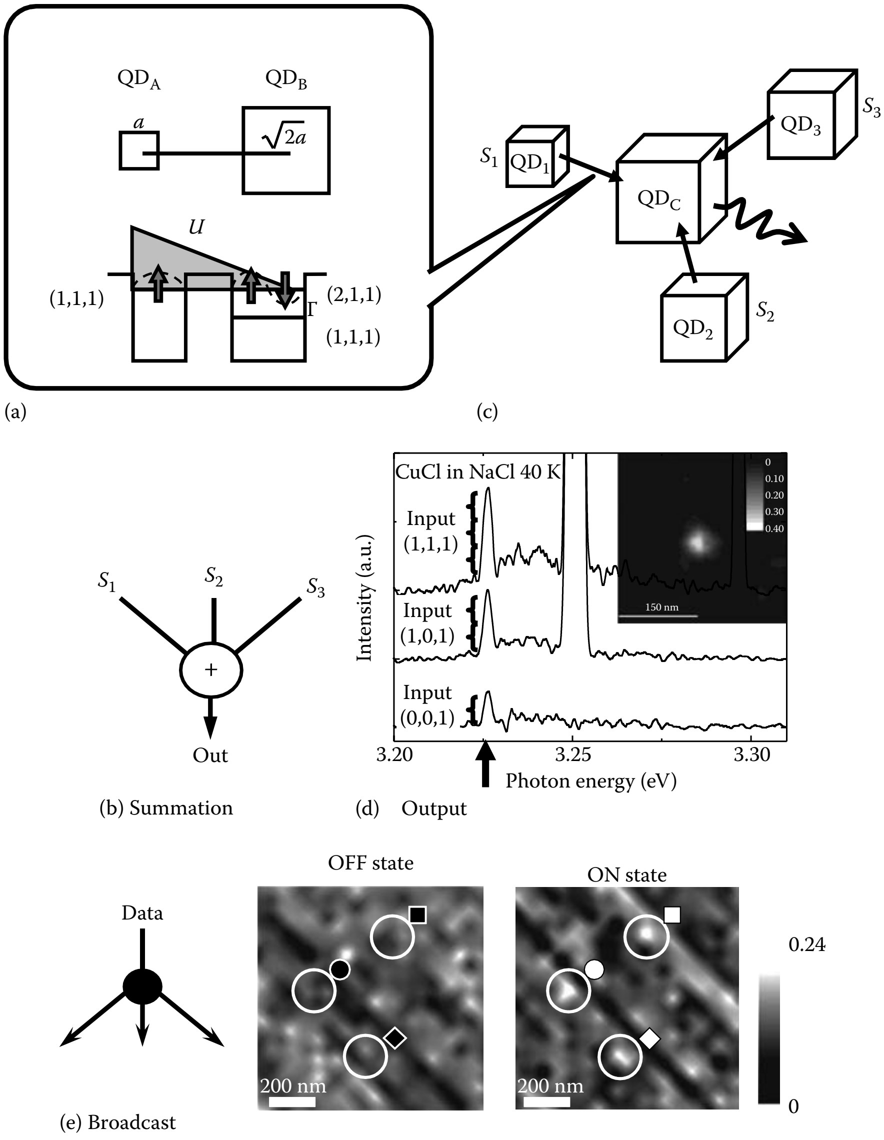
QD_A
QD_B
a
√2a
U
(1,1,1)
(2,1,1)
Γ
(1,1,1)
(a)
S_1
QD_1
QD_C
QD_3
S_3
QD_2
S_2
(c)
S_1
S_2
S_3
+
Out
(b) Summation
CuCl in NaCl 40 K
Input (1,1,1)
Input (1,0,1)
Input (0,0,1)
Intensity (a.u.)
3.20
3.25
3.30
Photon energy (eV)
150 nm
0
0.10
0.20
0.30
0.40
(d) Output
Data
(e) Broadcast
OFF state
ON state
200 nm
200 nm
0.24
0

FIGURE 34.2 (a) Optical excitation transfer between quantum dots via optical near-field interactions. (b) Global summation: a basic function for memory-based architectures. (c) Quantum-dot arrangement for summation via an optical near field. (d) Intensity for three different input combinations and the spatial intensity distribution of the output photon energy. (e) Broadcast interconnects for parallel processing. Spatial intensity distribution of the output of 3-dot AND gates.

architectures. First of all, two different physical states appear by controlling the dissipation processes in the larger dot; this is the principle of the nanophotonic switch (Kawazoe et al. 2003). Also, such a flow control itself allows an architecture known as a binary decision diagram, where an arbitrary combinatorial logic operation is determined by the destination of a signal flowing from a root (Akers 1978).

Such optical excitation transfer processes also lead to unique system architectures. In this regard, Section 34.2.2 discusses a massively parallel architecture and its nanophotonic implementations. Also, Section 34.2.3 demonstrates that optical excitation transfer provides higher tamper resistance against attacks than conventional electrically wired devices, by focusing on environmental factors for signal transfer.

34.2.2 Parallel Architecture Using Optical Excitation Transfer

34.2.2.1 Memory-Based Architecture

This section discusses a memory-based architecture where computations are regarded as a table lookup or database search problem, which is also called content addressable memory (CAM) (Liu 2002). The inherent parallelism of this architecture is well matched with the physics of optical excitation transfer, and provides performance benefits in high-density, low-power operations.

In this architecture, input signal (content) serves as a query to a lookup table, and the output is the address of the data matching the input. This architecture plays a critical role in various systems for example in a data router where the output port for an incoming packet is determined based on the lookup tables.

All optical means for implementing such functions have been proposed, for instance, by using planar lightwave circuits (Grunnet-Jepsen et al. 1999). However, since separate optical hardware for each table entry is needed if based on today's known methods, if the number of entries in the routing table is on the order of 10,000 or more, the overall physical size of the system becomes impractically large. On the other hand, by using diffraction-limit-free nanophotonic principles, huge lookup tables can be configured compactly.

Then, it is important to note that the table lookup problem is equivalent to an inner product operation. Assume an N-bit input signal $\mathbf{S} = (s_1,\ldots,s_N)$ and reference data $\mathbf{D} = (d_1,\ldots,d_N)$. Here, the inner product $\mathbf{S}\cdot\mathbf{D} = \sum_{i=1}^{N} s_i \cdot d_i$ will provide a maximum value when the input perfectly matches the reference data with an appropriate modulation format (Naruse et al. 2004). Then, the function of a CAM is to derive j, which maximizes the $\mathbf{S} \cdot \mathbf{D}_j$.

34.2.2.2 Global Summation Using Near-Field Interactions

As discussed in Section 34.2.2.1, the inner product operations are the key functionalities of the memory-based architecture. The multiplication of two bits, namely, $x_i = s_i \cdot d_i$, has already been demonstrated by a combination of three quantum dots (Kawazoe et al. 2003). Therefore, one of the key operations remaining is the summation, or data gathering scheme, denoted by $\sum x_i$, where all data bits should be taken into account, as schematically shown in Figure 34.2b.

In known optical methods, wave propagation in free-space or in waveguides, using focusing lenses or fiber couplers, for example, well matches such a data gathering scheme because the physical nature of propagating light is inherently suitable for the collection or distribution of information, such as global summation. However, the level of the integration of these methods is restricted due to the diffraction limit of light. In nanophotonics, on the other hand, the near-field interaction is inherently physically local, although functional global behavior is required.

The global data gathering mechanism, or summation, is realized based on the unidirectional energy flow via an optical near field, as schematically shown in Figure 34.2c, where surrounding excitations are transferred toward a quantum dot $\mathrm{QD_C}$ located at the center (Kawazoe et al. 2005, Naruse et al. 2005b). This is based on the excitation transfer processes presented in Section 34.2.1 and in Figure 34.2a, where an optical excitation is transferred from a smaller dot ($\mathrm{QD_A}$) to a larger one ($\mathrm{QD_B}$) through a resonant energy sublevel and a sublevel relaxation process occurring at a larger dot. In the system shown in Figure 34.2c, similar energy transfers may take place among the resonant energy levels in the dots surrounding $\mathrm{QD_C}$ so that excitation transfer can occur. The lowest energy level in each quantum dot is coupled to a free photon bath to sweep out the excitation radiatively. The output signal is proportional to the (1, 1, 1)-level population in $\mathrm{QD_B}$.

A proof-of-principle experiment was performed to verify the nanoscale summation using CuCl quantum dots in a NaCl matrix, which has also been employed for demonstrating nanophotonic switches (Kawazoe et al. 2003) and optical nano-fountains (Kawazoe et al. 2005). A quantum-dot arrangement where small QDs ($\mathrm{QD_1}$–$\mathrm{QD_3}$) surrounded a large QD at the center ($\mathrm{QD_C}$) was chosen. Here, at most three light beams with different wavelengths, 325, 376, and 381.3 nm, are irradiated, which excite the quantum dots $\mathrm{QD_1}$–$\mathrm{QD_3}$ having sizes of 1, 3.1, and 4.1 nm, respectively. The excited excitons are transferred to $\mathrm{QD_C}$, and their radiation is observed by a near-field fiber probe tip. Notice the output signal intensity at a photon energy level of 3.225 eV in Figure 34.2d, which corresponds to a wavelength of 384 nm, or a $\mathrm{QD_C}$ size of 5.9 nm. The intensity varies approximately as 1:2:3 depending on the number of excited QDs in the vicinity, as observed in Figure 34.2d. The spatial intensity distribution was measured by scanning the fiber probe, as shown in the inset of Figure 34.2d, where the energy is converged at the center. Hence, this architecture works as a summation mechanism, counting the number of input channels, based on exciton energy transfer via optical near-field interactions.

Such a quantum-dot-based data-gathering mechanism is also extremely energy efficient compared to other optical methods such as focusing lenses or optical couplers. For example, the transmittance between two materials with refractive indexes n_1 and n_2 is given by $4n_1n_2/(n_1 + n_2)^2$; this gives a 4% loss if n_1 and n_2 are 1 and 1.5, respectively. The transmittance of an N-channel-guided wave coupler is $1/N$ from the input to the output if the coupling loss at each coupler is 3 dB. In nanophotonic summation, the loss is attributed to the dissipation between energy sublevels, which is significantly smaller. Incidentally, it is energy- and space-efficient compared to electrical CAM VLSI chips (Lin and Kuo 2001, Arsovski et al. 2003, Naruse et al. 2005a).

34.2.2.3 Broadcast Interconnects

For the parallel architecture shown in Section 34.2.2.2, it should also be noted that the input data should be commonly applied to all lookup table entries. In other words, broadcast interconnect is another important requirement for parallel architectures. Broadcast is also important in applications such as matrix-vector products (Goodman et al. 1978, Guilfoyle and McCallum 1996) and switching operations, for example, broadcast-and-select architectures (Li et al. 2001). Optics is in fact well suited to such broadcast operations in the form of simple imaging optics (Goodman et al. 1978, Guilfoyle and McCallum 1996) or in optical waveguide couplers; thanks to the nature of wave propagation. However, the integration density of this approach is physically limited by the diffraction limit, which leads to bulky system configurations.

The overall physical operation principle of broadcast using optical near fields is as follows. Suppose that the arrays of nanophotonic circuit blocks are distributed within an area whose size is comparable to the wavelength. For broadcasting, multiple input QDs simultaneously accept identical input data carried by diffraction-limited far-field light by tuning their optical frequency so that the light is coupled to dipole-allowed energy sublevels.

The far- and near-field coupling mentioned above is explained based on a model assuming cubic quantum dots, which was introduced in Section 34.2.1. According to Equation 34.2, there exists a resonance between the quantized exciton energy sublevel of quantum number (1, 1, 1) for the QD with effective side length a and that of quantum number (2, 1, 1) for the QD with effective side length $\sqrt{2}a$. Energy transfer from the smaller QD to the larger one occurs via optical near fields, which is forbidden for far-field light (Kawazoe et al. 2003).

The input energy level for the QDs, that is, the (1, 1, 1)-level, can also couple to the far-field excitation. This fact can be utilized for data broadcasting. One of the design restrictions is that energy sublevels for input channels do not overlap with those for output channels. Also, if there are QDs internally used for near-field coupling, dipole-allowed energy sublevels for those QDs cannot be used for input channels since the inputs are provided by far-field light, which may lead to the misbehavior of internal near-field interactions if resonant levels exist. Therefore, frequency partitioning among the input, internal, and output channels is important. The frequencies used for broadcasting, denoted by $\Omega_i = \{\omega_{i,1}, \omega_{i,2}, \ldots, \omega_{i,A}\}$, should be distinct values and should not overlap with the output channel frequencies $\Omega_o = \{\omega_{o,1}, \omega_{o,2}, \ldots, \omega_{o,B}\}$. A and B indicate the number of frequencies used for input and output channels, respectively. Also, there will be frequencies needed for internal device operations, which are not used for either input or output, denoted by $\Omega_n = \{\omega_{n,1}, \omega_{n,2}, \ldots, \omega_{n,C}\}$, where C is the number of those frequencies. Therefore, the design criteria for global data broadcasting is to exclusively assign input, output, and internal frequencies, Ω_i, Ω_o, and Ω_n, respectively.

In a frequency multiplexing sense, this interconnection method is similar to multi-wavelength chip-scale interconnection (De Souza et al. 1995). Known methods, however, require a physical space comparable to the number of diffraction-limited input channels due to wavelength demultiplexing, whereas in the nanophotonic scheme, the device arrays are integrated on the sub-wavelength scale, and multiple frequencies are multiplexed in the far-field light supplied to the device.

To verify the broadcasting method, the following experiments were performed using CuCl QDs inhomogeneously distributed in a NaCl matrix at a temperature of 22 K (Naruse et al. 2006). To operate a 3-dot nanophotonic switch (2-input AND gate) in the device, at most two input light beams (IN1 and IN2) were irradiated. When both inputs exist, an output signal is obtained from the positions where the switches exist, as described above. In the experiment, IN1 and IN2 were assigned to 325 and 384.7 nm, respectively. They were irradiated over the entire sample (global irradiation) via far-field light. The spatial intensity distribution of the output, at 382.6 nm, was measured by scanning a near-field fiber probe within an area of approximately 1 μm × 1 μm. When only IN1 was applied to the sample the output of the AND gate was ZERO (OFF state). When both inputs were irradiated the output was ONE (ON state). Note the regions marked by ■, ●, and ◆ in Figure 34.2e. In those regions, the output signal levels were, respectively, low and high, which indicate that multiple AND gates were integrated at densities beyond the scale of the globally irradiated input beam area. That is to say, broadcast interconnects to nanophotonic switch arrays are accomplished by diffraction-limited far-field light.

Combining this broadcasting mechanism with the summation mechanism discussed in Section 34.2.2.2 will allow the development of the nanoscale integration of massively parallel architectures, which have conventionally resulted in bulky configurations.

34.2.3 Secure Signal Transfer in Nanophotonics

In addition to breaking through the diffraction limit of light, such local interactions of optical near fields also have important functional aspects, such as in security applications, which particularly tamper resistance against attacks (Naruse et al. 2007a). One of the most critical security issues in present electronic devices is so-called side-channel attacks, by which information is tampered either invasively or noninvasively. This may be achieved, for instance, merely by monitoring their power consumption (Kocher et al. 1998).

In this section, it is shown that devices based on optical excitation transfer via near-field interactions are physically more tamper-resistant than their conventional electronic counterparts. The key is that the flow of information in nanoscale devices cannot be completed unless they are appropriately coupled with their environment (Hori 2001), which could possibly be the weakest link in terms of their tamper resistance. A theoretical approach is presented to investigate the tamper resistance of optical excitation transfer, including a comparison with electrical devices.

Here, the tampering of information is defined as involving simple signal transfer processes, since the primary focus is on their fundamental physical properties.

In order to compare the tamper resistance, an electronic system based on single-charge tunneling is introduced here, in which a tunnel junction with capacitance C and tunneling resistance R_T is coupled to a voltage source V via an external impedance $Z(\omega)$, as shown in Figure 34.3a. In order to achieve

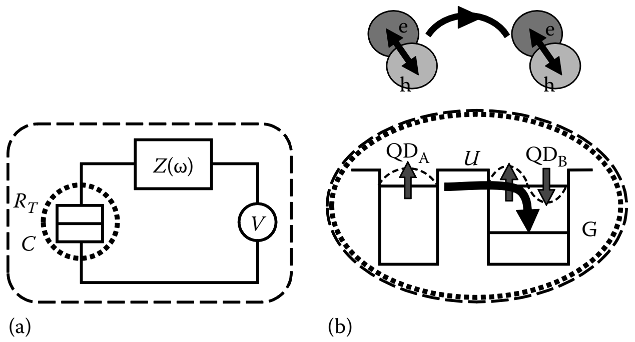

FIGURE 34.3 Model of tamper resistance in devices based on (a) single charge tunneling and (b) optical excitation transfer. Dotted curves show the scale of a key device and dashed curves show the scale of the environment required for the system to work.

single charge tunneling, besides the condition that the electrostatic energy $E_C = e^2/2C$ of a single excess electron be greater than the thermal energy k_BT, the environment must have appropriate conditions, as discussed in detail in Ingold and Nazarov (1992). For instance, with an inductance L in the external impedance, the fluctuation of the charge is given by

$$\langle \delta Q^2 \rangle = \frac{e^2}{4\rho}\coth\left(\frac{\beta\hbar\omega_S}{2}\right), \qquad (34.3)$$

where

$\rho = E_C/\hbar\omega_S$
$\omega_S = (LC)^{-1/2}$
$\beta = 1/k_BT$

Therefore, charge fluctuations cannot be small even at zero temperature unless $\rho \gg 1$. This means that a high-impedance environment is necessary, which makes tampering technically easy, for instance by adding another impedance circuit.

Here, let us define two scales to illustrate tamper resistance: (1) the scale associated with the key device size and (2) the scale associated with the environment required for operating the system, which are, respectively, indicated by dotted and dashed curves in Figure 34.3. In the case of Figure 34.3a, scale I is the scale of a tunneling device, whereas scale II covers all of the components. It turns out that the low tamper resistance of such wired devices is due to the fact that scale II is typically the macroscale, even though scale I is the nanometer scale.

In contrast, in the case of the optical excitation transfer shown in Figure 34.3b, the two quantum dots and their surrounding environment are governed by scale I. It is also important to note that scale II is the same as scale I. More specifically, the transfer of an exciton from QD_A to QD_B is completed due to the non-radiative relaxation process occurring at QD_B, which is usually difficult to tamper with. Theoretically, the sublevel relaxation constant is given by

$$\Gamma = 2\pi |g(\omega)|^2 D(\omega), \qquad (34.4)$$

where

$\hbar g(\omega)$ is the exciton–phonon coupling energy at frequency ω
$\hbar$ is Planck's constant divided by 2π
$D(\omega)$ is the phonon density of states (Carmichael 1999)

Therefore, tampering with the relaxation process requires somehow "stealing" the exciton–phonon coupling, which would be extremely difficult technically.

It should also be noted that the energy dissipation occurring in the optical excitation transfer, derived theoretically as $E_{(2,1,1)} - E_{(1,1,1)}$ in QD_B based on Equation 34.2, should be larger than the exciton-phonon coupling energy of $\hbar\Gamma$, otherwise, the two levels in QD_B cannot be resolved. This is similar to the fact that the condition $\rho \gg 1$ is necessary in the electron-tunneling example, which means that the mode energy $\hbar\omega_S$ is smaller than the required charging energy E_C. By regarding $\hbar\Gamma$ as a kind of mode energy in the optical excitation transfer, the difference between the optical excitation transfer and a conventional wired device is the physical scale at which this mode energy is realized: nanoscale for the optical excitation transfer and macroscale for electric circuits.

34.3 Hierarchical Architectures in Nanophotonics

34.3.1 Physical Hierarchy in Nanophotonics and Functional Hierarchy

In this section, another feature of nanophotonics, the inherent hierarchy in optical near-field interactions, is exploited. As schematically shown in Figure 34.4a, there are multiple layers associated with the physical scale between the macroscale world and the atomic-scale world, which are primarily governed by propagating light and electron interactions, respectively. Between these two extremes, typically in scales ranging from a few nanometers to wavelength size, optical near-field interactions play a crucial role. In this section, such hierarchical properties in this mesoscopic or sub-wavelength regime are exploited.

Such physical hierarchy in optical near-field interactions will be analyzed by a simple dipole–dipole interaction model and an angular spectrum representation of optical near fields, as shown in Section 34.3.2. Before going into the details of the physical processes, functionalities required for system applications are briefly reviewed in terms of hierarchy.

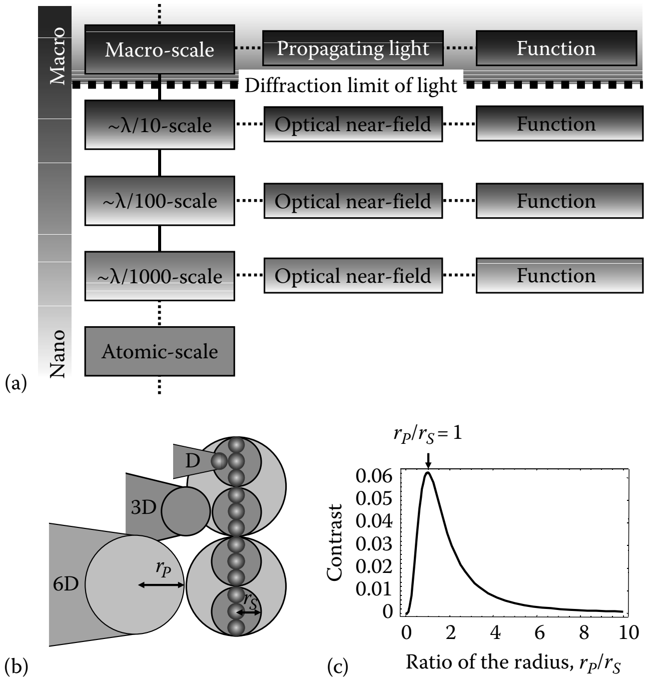

FIGURE 34.4 (a) Hierarchy in optical near-field interactions. (b) Dipole–dipole interaction. (c) Signal contrast as a function of the ratio of the radius of the sample and the probe.

One of the problems for ultrahigh-density nanophotonic systems is interconnection bottlenecks, which have been addressed previously in Section 34.2.3 regarding broadcast interconnects. In fact, a hierarchical structure can be found in these broadcast interconnects by relating far-field effects at a coarser scale and near-field effects at a finer scale.

In this regard, it should also be mentioned that such physical differences in optical near-field and far-field effects can be used for a wide range of applications. The behavior of usual optical elements, such as diffractive optical elements, holograms, or glass components, is associated with their optical responses in optical far fields. In other words, nanostructures can exist in such optical elements as long as they do not affect the optical responses in far fields. Designing nanostructures accessible only via optical near fields provides additional, or hidden, information recorded in those optical elements, while maintaining the original optical responses in far fields. In fact, a "*hierarchical hologram*" or "*hierarchical diffraction grating*" has been experimentally demonstrated, as schematically shown in Figure 34.5 (Tate et al. 2008).

Since there is more hierarchy in the optical near-field regime, further applications should be possible, for example, it should be possible for nanometer-scale high-density systems to be gradually hierarchically connected to coarse layer systems.

Hierarchical functionalities are also important for several aspects of memory systems. One is related to recent high-density, huge-capacity memory systems, in which data retrieval or searching from entire memory archives is made even more difficult. Hierarchy is one approach for solving such a problem by making systems hierarchical, that is, by recording abstract data, metadata, or tag data in addition to the original raw data.

Hierarchy in nanophotonics provides a physical solution to achieve such functional hierarchy. As will be introduced below, low-density, rough information is read out at a coarser scale, whereas high-density, detailed information is read out at a finer scale. Sections 34.3.2 and 34.3.3 will show physical mechanisms for such hierarchical information retrieval.

Another issue in hierarchical functionalities will be security. High-security information is recorded at a finer scale, whereas less-critical security information is associated with a coarse layer. Also, in addition to associating different types of information with different physical scales, another kind of information can also be related to one or more layers of the physical hierarchy, for instance, traceability, history, or aging of information. Section 34.3.3 will demonstrate a *traceable* memory as an example.

34.3.2 Hierarchical Memory Retrieval

This section describes a physical model of optical near-field interactions based on dipole–dipole interactions (Ohtsu and Kobayashi 2004). Suppose that a probe, which is modeled by a sphere of radius r_P, is placed close to a sample to be observed, which is modeled as a sphere of radius r_S. Figure 34.4b shows three different sizes for the probe and the sample. When they are illuminated by incident light, whose electric field is $\mathbf{E}_0$, electric dipole moments are induced in both the probe and the sample; these moments are denoted by $\mathbf{p}_P = \alpha_P\mathbf{E}_0$ and $\mathbf{p}_S = \alpha_S\mathbf{E}_0$, respectively. The electric dipole moment induced in the sample $\mathbf{p}_S$, then generates an electric field, which changes the electric dipole moment in the probe by an amount $\Delta\mathbf{p}_P = \Delta\alpha_P\mathbf{E}_0$. Similarly, $\mathbf{p}_P$ changes the electric dipole moment in the sample by $\Delta\mathbf{p}_S = \Delta\alpha_S\mathbf{E}_0$. These electromagnetic interactions are called

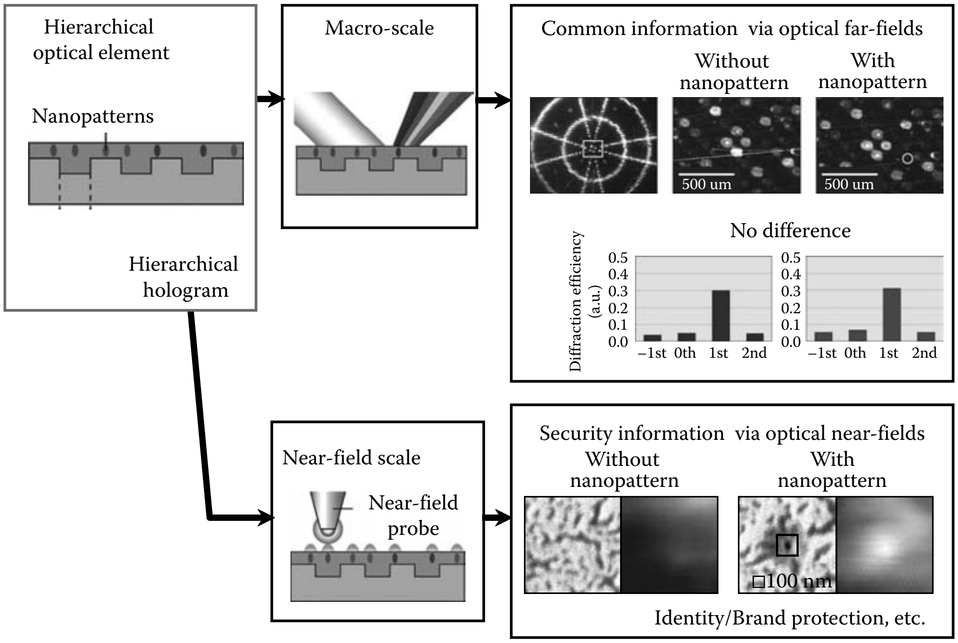

FIGURE 34.5 Hierarchical optical elements, such as *hierarchical holograms*, based on different optical responses obtained in optical far and near fields.

dipole–dipole interactions. The scattering intensity induced by these electric dipole moments is given by

$$\begin{aligned} I &= \left|\mathbf{p}_P + \Delta\mathbf{p}_P + \mathbf{p}_S + \Delta\mathbf{p}_S\right|^2 \\ &\approx (\alpha_P + \alpha_S)^2\,|E_0|^2 + 4\Delta\alpha(\alpha_P + \alpha_S)\,|E_0|^2 \end{aligned} \tag{34.5}$$

where $\Delta\alpha = \Delta\alpha_S = \Delta\alpha_P$ (Ohtsu and Kobayashi 2004). The second term in Equation 34.5 shows the intensity of the scattered light generated by the dipole–dipole interactions, containing the information of interest, which is the relative difference between the probe and the sample. The first term in Equation 34.5 is the background signal for the measurement. Therefore, the ratio of the second term to the first term of Equation 34.5 corresponds to a signal contrast, which will be maximized when the sizes of the probe and the sample are the same ($r_P = r_S$), as shown in Figure 34.4c. (A detailed derivation is found in Ohtsu and Kobabashi (2004).) Thus, one can see a scale-dependent physical hierarchy in this framework, where a small probe, say $r_P = D/2$, can nicely resolve objects with a comparable resolution, whereas a large probe, say $r_P = 3D/2$, cannot resolve a detailed structure but it can resolve a structure with a resolution comparable to the probe size. Therefore, although a large diameter probe cannot detect smaller-scale structure, it could detect certain features associated with its scale.

Based on the above simple hierarchical mechanism, a hierarchical memory system is constructed. Consider, for example, a maximum of N nanoparticles distributed in a region of a subwavelength scale. These nanoparticles can be nicely resolved by a scanning near-field microscope if the size of its fiber probe tip is comparable to the size of individual nanoparticles; in this way, the *first-layer* information associated with each distribution of nanoparticles is retrievable, corresponding to 2^N-different codes. By using a larger-diameter fiber probe tip instead, the distribution of the particles cannot be resolved, but a mean-field feature with a resolution comparable to the size of the probe can be extracted, namely, the number of particles within an area comparable to the size of the fiber probe tip. Thus, the *second-layer* information associated with the number of particles, corresponding to $(N + 1)$-different level of signals, is retrievable. Therefore, one can access different set of signals, 2^N or $N + 1$, depending on the scale of observation. This leads to hierarchical memory retrieval by associating this information hierarchy with the distribution and the number of nanoparticles using an appropriate coding strategy, as schematically shown in Figure 34.6a.

For example, in encoding N-bit information, $(N - 1)$-bit signals can be encoded by the distributions of nanoparticles while associating the remaining 1-bit with the number of nanoparticles. The details of encoding/decoding strategies will be found in Naruse et al. (2005c).

Simulations were performed assuming ideal isotropic metal particles to see how the second-layer signal varies depending on the number of particles using a finite-difference time-domain simulator (*Poynting for Optics*, a product of Fujitsu, Japan). Here, 80 nm-diameter particles are distributed over a 200 nm-radius circular grid at constant intervals. The solid circles in Figure 34.6d show calculated scattering cross sections as a function of the number of particles. A linear correspondence to the number of particles was observed. This result supports the simple physical model described above.

In order to experimentally demonstrate such principles, an array of Au particles, each with a diameter of around 80 nm, was distributed over a SiO_2 substrate in a 200 nm-radius circle. These particles were fabricated by a liftoff technique using electron-beam (EB) lithography with a Cr buffer layer. Each group of Au particles was spaced by 2 μm. A scanning electron microscope (SEM) image is shown in Figure 34.6b in which the values indicate the number of particles within each group. In order to illuminate all Au particles in each group and collect the scattered light from them, a near-field optical microscope (NOM) with a large-diameter-aperture (500 nm) metallized fiber probe was used in an illumination collection setup. The light source used was a laser diode with an operating wavelength of 680 nm. The distance between the substrate and the probe was maintained at 750 nm. Figure 34.6c shows an intensity profile captured by the probe, from which the second-layer information is retrieved. The solid squares in Figure 34.6d indicate the peak intensity of each section, which increased linearly. These results show the validity of hierarchical memory retrieval from nanostructures.

34.3.3 Design of *Unscalable* Hierarchical Response

The coarse graining process is usually an averaging process, meaning that the signal in the coarser layer is obtained in terms of a mean-field approximation of the fine-grained, lower-layer signals, which is also the case shown in Section 34.3.2. However, it should be noted that the optical near-field amplitude can be distributed independently at different scales of observation. In other words, the coarse graining of the optical near fields in fact provides an optical property independent of the lower-layer feature; such an *unscalable* hierarchical property of optical near fields will be analyzed below.

Here, the angular spectrum representation of the electromagnetic field is used for discussing the hierarchy in optical near fields (Wolf and Nieto-Vesperinas 1985, Inoue and Hori 2005, Naruse et al. 2007b). This allows an analytical treatment and gives an intuitive picture of the localization of optical near fields and represents relevance/irrelevance in optical near-field interactions at different scales of observation since it describes electromagnetic fields as a superposition of evanescent waves with different decay length and corresponding spatial frequency.

Suppose, for example, that there is an oscillating electric dipole, $\mathbf{d}^{(k)} = d^{(k)}(\cos\phi^{(k)}, 0)$, on the xz plane, which is oriented parallel to the x axis. Now, consider the electric field of radiation observed at a position displaced from the dipole by $\mathbf{R} = (r_{\parallel}^{(k)}\cos\varphi^{(k)}, z^{(k)})$. The angular spectrum representation of the z-component of the optical near field is given by

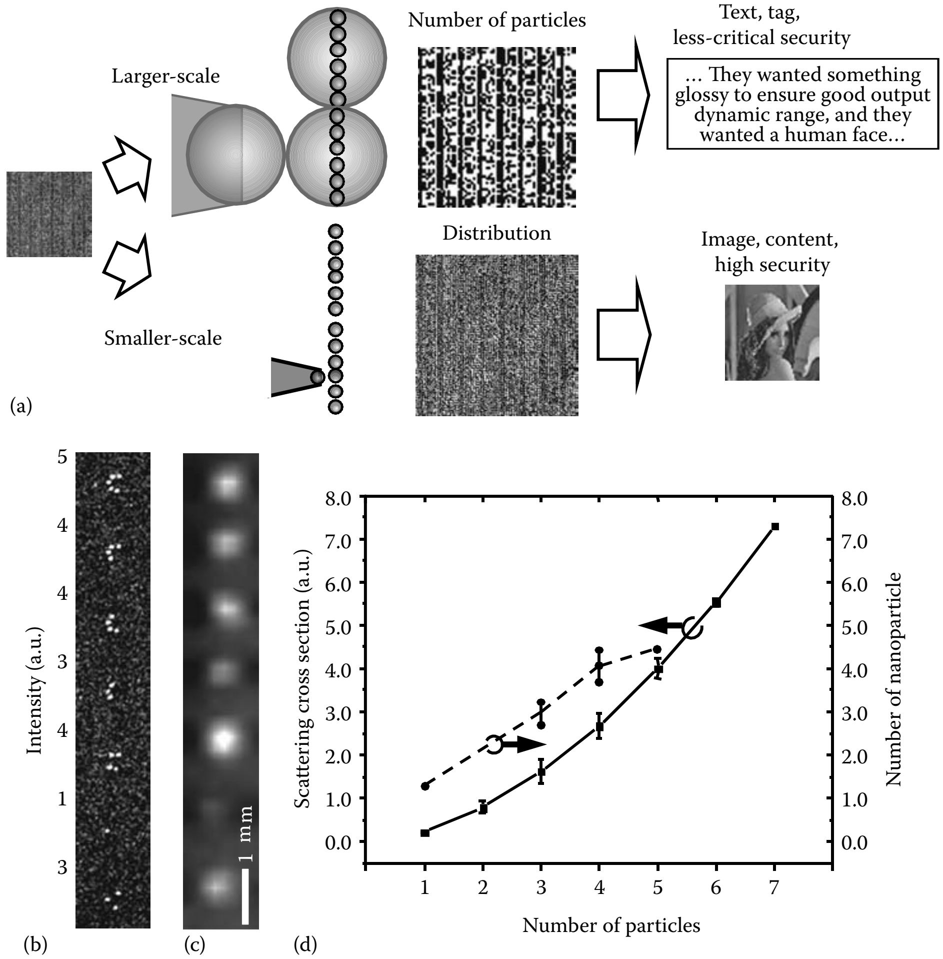

FIGURE 34.6 (a) Hierarchical memory retrieval from nanostructures. (b) SEM picture of an array of Au nanoparticles. Each section consists of up to seven nanoparticles. (c) Intensity pattern captured by a fiber probe tip whose diameter is comparable to the size of each section of nanoparticles. (d) Square marks: calculated scattering cross sections in each section, circular marks: peak intensity of each section in the intensity profile shown in (c).

$$E_z(\mathbf{R}) = \left(\frac{iK^3}{4\pi\varepsilon_0}\right)\int_1^{\infty} ds_{\parallel}\frac{s_{\parallel}}{s_z} f_z(s_{\parallel}, \mathbf{d}^{(1)},\ldots, \mathbf{d}^{(N)}) \tag{34.6}$$

where

$$f_z(s_{\parallel},\mathbf{d}^{(1)},\ldots,\mathbf{d}^{(N)}) = \sum_{k=1}^{N} d^{(k)} s_{\parallel}\sqrt{s_{\parallel}^2-1}\cos(\phi^{(k)}-\varphi^{(k)})J_1\left(Kr_{\parallel}^{(k)}s_{\parallel}\right) \times \exp\left(-Kz^{(k)}\sqrt{s_{\parallel}^2-1}\right). \tag{34.7}$$

Here, $s_{\parallel}$ is the spatial frequency of an evanescent wave propagating parallel to the x axis, and $J_n(x)$ represents Bessel functions of the first kind. Here, the term $f_z(s_{\parallel}, \mathbf{d}^{(1)},\ldots, \mathbf{d}^{(N)})$ is called the angular spectrum of the electric field.

In the following, a two-layer system is introduced where (1) by observing very close to the dipoles, two items of first layer information are retrieved, and (2) by observing relatively far from the dipoles, one item of second-layer information is retrieved.

Suppose that there are two closely spaced dipole pairs (so there are four dipoles in total). The dipoles $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ are oriented in the same direction, namely, $\varphi^{(1)} = \varphi^{(2)} = 0$, and another dipole pair, $\mathbf{d}^{(3)}$ and $\mathbf{d}^{(4)}$, are both oriented in the opposite direction to $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$, namely, $\varphi^{(3)} = \varphi^{(4)} = \pi$. These four dipoles are located at positions shown in Figure 34.7a. Here, at a position close to the x axis equidistant from $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$, such as at the position A_1 in Figure 34.7a, the electric field is weak (logical ZERO) since (1) the angular spectrum originating from $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ vanishes, and (2) the electric field originating from $\mathbf{d}^{(3)}$ and $\mathbf{d}^{(4)}$ is small because they are far from the position A_1. In fact, as shown by the dashed curve in Figure 34.7b, since the angular spectrum at position A_1 oscillates, the integral of the angular spectrum, which is correlated to the field intensity at that point, will be low.

For the second layer retrieval, consider the observation at an intermediate position between the dipole pairs, such as the position B in Figure 34.7a. From this position, the four dipoles effectively appear to be two dipoles that are oriented in opposite directions to each other. As shown by the solid curve in Figure 34.7b, the angular spectrum involving relatively low

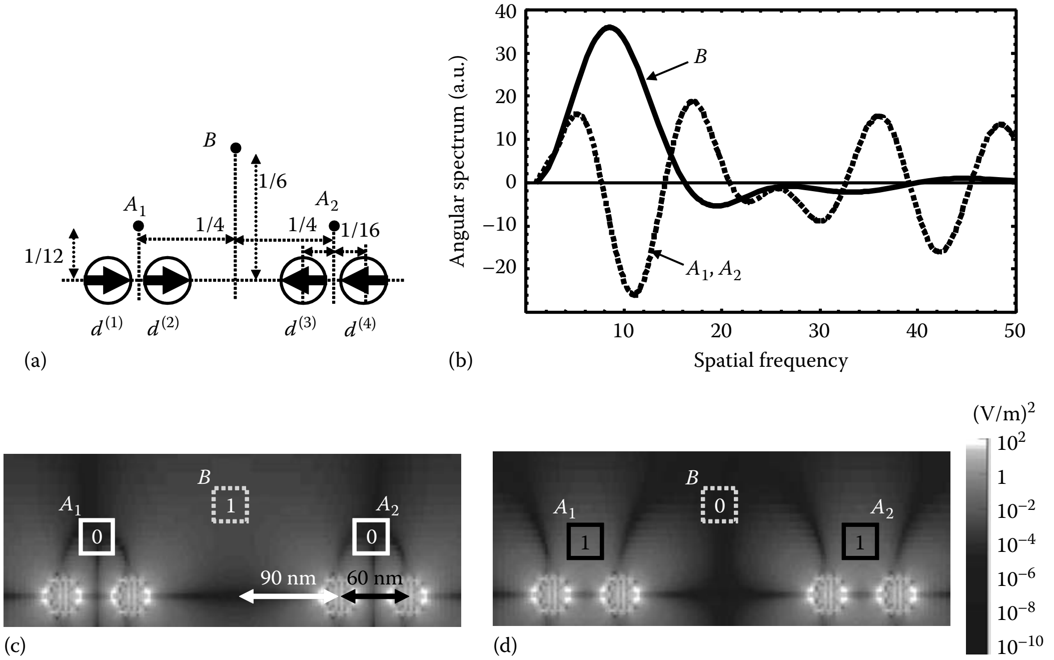

FIGURE 34.7 Examples of *unscalable* hierarchy in optical near fields. (a) The positions and orientations of four electric dipole systems. (b) And the corresponding angular spectrum observed at positions A_1, A_2, and B, indicating that the first-layer signals (obtained at A_1 and A_2) are logical ZEROs, while the second-layer signal (obtained at B) is logical ONE. (c) Calculated electric field intensity distribution agreed with theoretical predictions shown in (b). (d) Another electric field distribution where logical ONEs are retrieved at the first layer (A_1 and A_2) and logical ZERO is retrieved at the second layer (B).

spatial-frequency components shows a single peak, indicating that the electric field in the xy-plane is localized to the degree determined by its spectral width so that a logical ONE is retrievable at position B. Meanwhile, the angular spectrum observed at position A_2, shown in Figure 34.7a, is identical to that observed at position A_1, meaning that the electric field at A_2 is also at a low level.

To summarize the above mechanisms, a logical level of ONE in the second layer can be retrieved even though the two items of information retrieved in the first layer are both ZEROs; therefore, an unscalable hierarchy is achieved.

As described above, one of the two first-layer signals, the electric field at A_1, primarily depends on the dipole pair $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$, and the other, the electric field at A_2, depends primarily on the dipole pair $\mathbf{d}^{(3)}$ and $\mathbf{d}^{(4)}$. The second-layer signal is determined by all of those dipoles. Concerning such a hierarchical mechanism, it was shown that a total of eight different signal combinations were achieved by appropriately orienting the four dipoles (Naruse et al. 2007b).

Numerical simulations were performed based on finite-difference time-domain methods to see how they agree with the theoretical analysis based on the angular spectrum. Four silver nanoparticles (of radius 15 nm) containing a virtual oscillating light source were assumed in order to simulate dipole arrays. Their positions are shown in Figure 34.7c. The first and the second layers were located 40 and 80 nm away from the dipole plane, respectively. The operating wavelength was 488 nm. The electric fields obtained at A_1, A_2, and B agree with the combinations of the first- and second-layer signals to be retrieved, as shown in Figure 34.7c. As another unscalable hierarchy example, Figure 34.7d represents a situation where logical ONEs are obtained at the first layer, while logical ZERO is obtained at the second layer. These agree with the theoretical analysis based on the angular spectrum.

34.3.4 Versatile Functionalities Based on Hierarchy in Optical Near-Fields

The hierarchical nature has been further exploited by combining other physical mechanisms in nanophotonics. For example, we can associate one of the hierarchical layers with energy dissipation processes. Specifically, a two-layer system is demonstrated where (1) at smaller scale, called *Scale* 1, the system should exhibit a unique response, and (2) at a larger scale, called *Scale* 2, the system should output two different signals. Such a hierarchical response can be applied to functions like the *traceability* of optical memory in combination with a localized energy dissipation process at Scale 1 (Naruse et al. 2008c). Optical access to this memory will be automatically recorded due to energy dissipation occurring locally in Scale 1, while at the same time, information will be read out based on Scale 2 behavior. Therefore, such hierarchy enables the traceability of optical memory, which will be important for the security (confidentiality is ensured) and management of digital content.

Shape-engineered metal nanostructures can achieve the hierarchy required for traceable memory (Naruse et al. 2008e). Here, two types of shapes are assumed. The first one (*Shape* I) has two triangular metal plates aligned in the same direction, and the other one (*Shape* II) has them facing each other, as shown in Figure 34.8a. The metal is gold, the gap between the two apexes is 50 nm, the horizontal length of one triangular plate is 173 nm, the angle at the apex is 30°, and the thickness is 30 nm. An incident uniform plane wave with a wavelength of 680 nm is assumed for input light. The polarization is parallel to the x axis in Figure 34.8b. Now, Scale 1 is associated with the scale around the gap of the triangles, and Scale 2 is associated with the scale covering both of the triangles, as shown in Figure 34.8a.

Regarding the optical response at Scale 1, as shown in Figure 34.8b, the electric field near the surface (1 nm away from the metal surface) shows an intensity nearly five orders of magnitude higher than the surrounding area. It should also be noted that nearly comparable electric-field enhancements are observed near the apexes of Shapes I and II, which are, respectively, denoted by the squares and circles in Figure 34.8b.

On the other hand, Shapes I and II exhibit different responses at Scale 2. As shown in Figure 34.8c, Shape I exhibits larger scattering cross section compared to Shape II. This indicates that a digital output is retrievable by observing the scattering from the entire structure (Scale 2), where, for example, digital 1 and 0 are associated with Shape I and Shape II, respectively.

In order to experimentally demonstrate the principle, Shapes I and II were fabricated in gold metal plates on a glass substrate by a liftoff technique using EB lithography. A NOM in an illumination collection setup was used with an apertured fiber probe having a diameter of 500 nm, as shown in Figure 34.8d. The light

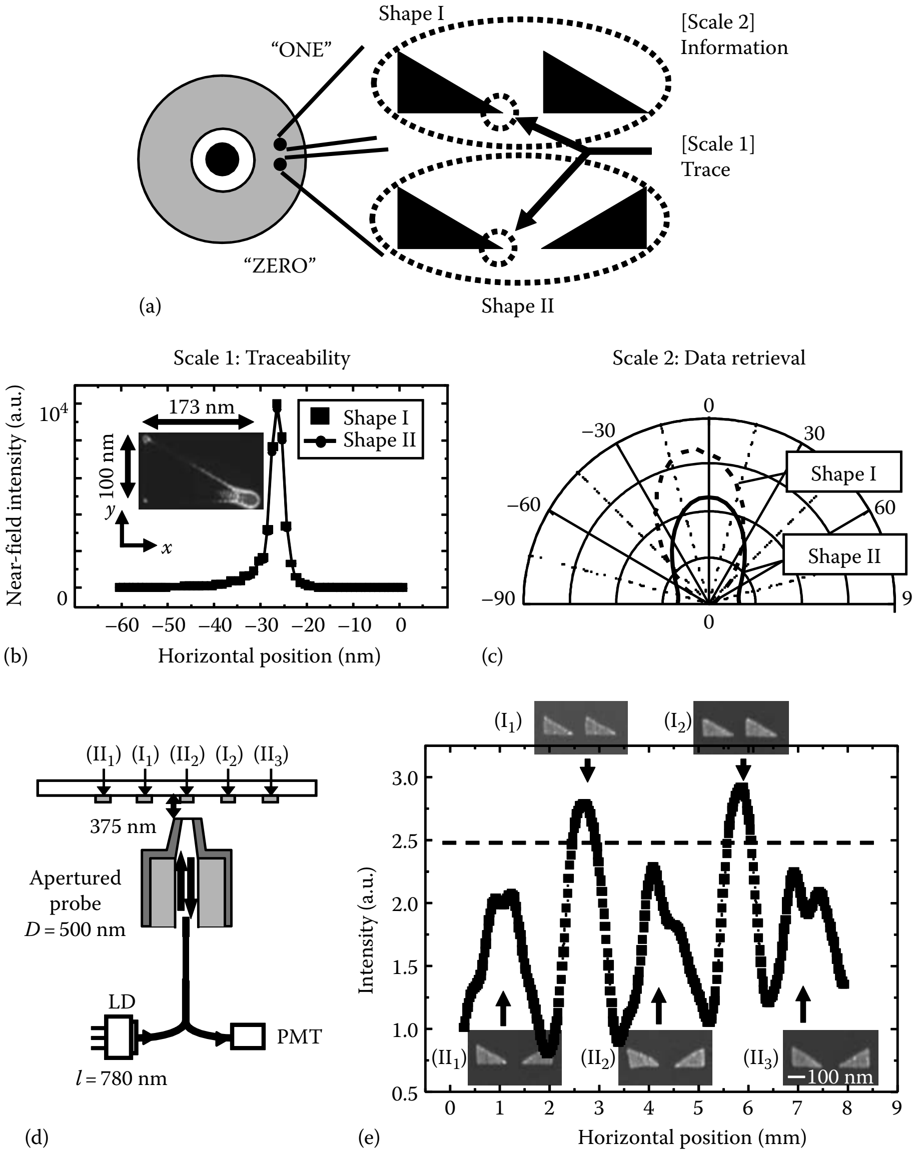

FIGURE 34.8 (a) Hierarchy in optical near fields by engineering the shape of metal plates at nanometer scale. (b) In Scale 1, both shapes exhibit comparable electric-field enhancement. (c) In Scale 2, they exhibit different system responses. (d) Experimental setup for Scale 2 signal retrieval. (e) Electric field intensity for Scale 2 signals for both Shapes I and II.

source used was a laser diode with an operating wavelength of 780 nm. The distance between the substrate and the probe was maintained at 375 nm. Figure 34.8e shows the electric field intensity depending on the shape of the metal plates, where the Shape I series exhibited larger values compared to the Shape II series, as expected.

Hierarchical nature in optical near-field interactions provides other unique functions. For example, the near-field photoluminescence of semiconductor quantum dots exhibits a hierarchical nature and its spectra diversity is maximized at an optimal scale of the optical near fields. This leads to a novel *non-pixelated memory architecture*, which can simultaneously retrieve a sequence of bits, as opposed to conventional bit-sequential pixelated architecture. The principles have been experimentally verified by InAs quantum dots (Naruse et al. 2008a). Also, a nanophotonic *lock-and-key* system has been demonstrated for the authentication of applications based on shape-engineered nanostructures and their associated optical near fields (Naruse et al. 2008b,d).

34.4 Summary

In this chapter, fundamental nanophotonic system architectures were presented along with two principal physical features of nanophotonics: one is optical excitation transfer and the other is hierarchy in optical near-field interactions. Both of these physical features originate from light–matter interactions on the nanometer scale. It should be emphasized that those basic features provide versatile applications and functionalities besides the example demonstrations shown in the above sections. Also, there are many other degrees-of-freedom in the nanometer scale that need to be deeply understood for systems. Further exploration and attempts to exploit nanophotonics for future devices and systems will certainly be exciting.

References

Akers, S. B. 1978. Binary decision diagrams. *IEEE Trans. Comput.* C-27:509–516.

Arsovski, I., Chandler, T., and Sheikholeslami, A. 2003. A ternary content-addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme. *IEEE J. Solid-State Circuits* 38:155–158.

Carmichael, H. J. 1999. *Statistical Methods in Quantum Optics I*. Berlin, Germany: Springer-Verlag.

De Souza, E. A., Nuss, M. C., Knox, W. H. et al. 1995. Wavelength-division multiplexing with femtosecond pulses. *Opt. Lett.* 20:1166–1168.

Goodman, J. W., Dias, A. R., and Woody, L. M. 1978. Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Lett.* 2:1–3.

Grunnet-Jepsen, A., Johnson, A. E., Maniloff, E. S. et al. 1999. Fibre Bragg grating based spectral encoder/decoder for lightwave CDMA. *Electron. Lett.* 35:1096–1097.

Guilfoyle, P. S. and McCallum, D. S. 1996. High-speed low-energy digital optical processors. *Opt. Eng.* 35:436–442.

Hori, H. 2001. Electronic and electromagnetic properties in nanometer scales. In *Optical and Electronic Process of Nano-Matters*, ed. M. Ohtsu, pp. 1–55. Tokyo: KTK Scientific/Dordrecht, the Netherlands: Kluwer Academic.

Ingold, G.-L. and Nazarov, Y. V. 1992. Charge tunneling rates in ultrasmall junctions. In *Single Charge Tunneling*, eds. H. Grabert and M. H. Devoret, pp. 21–107. New York: Plenum Press.

Inoue, T. and Hori, H., 2005. Quantum theory of radiation in optical near field based on quantization of evanescent electromagnetic waves using detector Mode. In *Progress in Nano-Electro-Optics IV*, ed. M. Ohtsu, pp. 127–199. Berlin, Germany: Springer-Verlag.

Kawazoe, T., Kobayashi, K., Sangu, S. et al. 2003. Demonstration of a nanophotonic switching operation by optical near-field energy transfer. *Appl. Phys. Lett.* 82:2957–2959.

Kawazoe, T., Kobayashi, K., and Ohtsu, M. 2005. Optical nanofountain: A biomimetic device that concentrates optical energy in a nanometric region. *Appl. Phys. Lett.* 86:103102 1–3.

Kocher, P., Jaffe, J., and Jun, B. 1998. Introduction to differential power analysis and related attacks. Cryptography Research. http://www.cryptography.com/resources/whitepapers/DPATechInfo.pdf

Li, B., Qin, Y., Cao, X. et al. 2001. Photonic packet switching: Architecture and performance. *Opt. Netw. Mag.* 2:27–39.

Lin, P.-F. and Kuo, J. B. 2001. A 1-V 128-kb four-way set-associative CMOS cache memory using wordline-oriented tag-compare (WLOTC) structure with the content-addressable-memory (CAM) 10-transistor tag cell. *IEEE J. Solid-State Circuits* 36:666–675.

Liu, H. 2002. Routing table compaction in ternary CAM, *IEEE Micro* 22:58–64.

Maier, S. A., Kik, P. G., Atwater, H. A. et al. 2003. Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides, *Nat. Mater*. 2:229–232.

Naruse, M., Mitsu, H., Furuki, M. et al. 2004. Terabit all-optical logic based on ultrafast two-dimensional transmission gating. *Opt. Lett.* 29:608–610.

Naruse, M., Miyazaki, T., Kawazoe, T. et al. 2005a. Nanophotonic computing based on optical near-field interactions between quantum dots. *IEICE Trans. Electron*. E88-C:1817–1823.

Naruse, M., Miyazaki, T., Kubota, F. et al. 2005b. Nanometric summation architecture using optical near-field interaction between quantum dots. *Opt. Lett.* 30:201–203.

Naruse, M., Yatsui, T., Nomura, W. et al. 2005c. Hierarchy in optical near-fields and its application to memory retrieval *Opt. Express* 13:9265–9271.

Naruse, M., Kawazoe, T., Sangu, S. et al. 2006. Optical interconnects based on optical far- and near-field interactions for high-density data broadcasting. *Opt. Express* 14:306–313.

Naruse, M., Hori, H., Kobayashi, K. et al. 2007a. Tamper resistance in optical excitation transfer based on optical near-field interactions. *Opt. Lett.* 32:1761–1763.

Naruse, M., Inoue, T., Hori, H. 2007b. Analysis and synthesis of hierarchy in optical near-field interactions at the nanoscale based on Angular Spectrum. *Jpn. J. Appl. Phys.* 46:6095–6103.

Naruse, M., Nishibayashi, K., Kawazoe, T. et al. 2008a. Scale-dependent optical near-fields in InAs quantum dots and their application to non-pixelated memory retrieval. *Appl. Phys. Express* 1:072101 1–3.

Naruse, M., Yatsui, T., Hori, H. et al. 2008b. Polarization in optical near- and far-field and its relation to shape and layout of nanostructures. *J. App. Phys.* 103:113525 1–8.

Naruse, M., Yatsui, T., Kawazoe, T. et al. 2008c. Design and simulation of a nanophotonic traceable memory using localized energy dissipation and hierarchy of optical near-field interactions *IEEE Trans. Nanotechnol.* 7:14–19.

Naruse, M., Yatsui, T., Kawazoe, T., Tate, N. et al. 2008d. Nanophotonic matching by optical near-fields between shape-engineered nanostructures. *Appl. Phys. Express* 1:112101 1–3.

Naruse, M., Yatsui, T., Kim, J. H. et al. 2008e. Hierarchy in optical near-fields by nano-scale shape engineering and its application to traceable memory *Appl. Phys. Express* 1:062004 1–3.

Ohtsu, M. and Hori, H. 1999. *Near-Field Nano-Optics*. New York: Kluwer Academic/Plenum Publishers.

Ohtsu, M. and Kobayashi, K. 2004. *Optical Near Fields*. Berlin, Germany: Springer-Verlag.

Ohtsu, M., Kobayashi, K., Kawazoe, T. et al. 2002. Nanophotonics: Design, fabrication, and operation of nanometric devices using optical near fields. *IEEE J. Sel. Top. Quantum Electron.* 8:839–862.

Ohtsu, M., Kobayashi, K., Kawazoe, T. et al. 2008. *Principles of Nanophotonics*. Boca Raton, FL: Taylor & Francis.

Pohl, D. W. and Courjon, D. eds. 1993. *Near Field Optics*. Dordrecht, the Netherlands: Kluwer Academic.

Tang, Z. K., Yanase, A., Yasui, T. et al. 1993. Optical selection rule and oscillator strength of confined exciton system in CuCl thin films. *Phys. Rev. Lett.* 71:1431–1434.

Tate, N., Nomura, W., Yatsui, T. et al. 2008. Hierarchical hologram based on optical near- and far-field responses. *Opt. Express* 16:607–612.

Wolf, E. and Nieto-Vesperinas, M. 1985. Analyticity of the angular spectrum amplitude of scattered fields and some of its consequences. *J. Opt. Soc. Am. A.* 2:886–890.

35

Nanophotonics for Device Operation and Fabrication

35.1 Introduction 35-1
35.2 Excitation Energy Transfer in Nanophotonic Devices 35-1
35.3 Device Operation 35-3
Nanophotonic AND Gate • Nanophotonic NOT Gate • Interconnection with Photonic Devices
35.4 Nanophotonics Fabrication 35-7
Nonadiabatic Near-Field Optical CVD • Nonadiabatic Near-Field Photolithography
35.5 Summary 35-11
References 35-12

Tadashi Kawazoe
The University of Tokyo

Motoichi Ohtsu
The University of Tokyo

35.1 Introduction

The optical near field is an electromagnetic field that mediates the interaction between nanometric particles located in close proximity to each other. Nanophotonics utilizes this field to realize novel devices, fabrications, and systems. That is, a photonic device with a novel function can be operated by transferring the optical near-field energy between nanometric particles and subsequent dissipation. In such a device, the optical near field transfers a signal and carries the information. Novel photonic systems become possible by using these novel photonic devices. Furthermore, if the magnitude of the transferred optical near-field energy is sufficiently large, structures or conformations of nanometric particles can be modified, which suggests the feasibility of novel photonic fabrications.

Note that the true nature of nanophotonics is to realize "qualitative innovation" in photonic devices, fabrications, and systems by utilizing novel functions and phenomena caused by optical near-field interactions, which are impossible as long as conventional propagating light is used. On reading this note, one may understand that the advantage of going beyond the diffraction limit, that is, "quantitative innovation," is no longer essential, but only a secondary nature of nanophotonics. In this sense, one should also note that optical near-field microscopy, that is, the methodology used for image acquisition and interpretation in a nondestructive manner, is not an appropriate application of nanophotonics because the magnitude of the optical near-field energy transferred between the probe and sample must be extrapolated to zero to avoid destroying the sample.

Quantitative innovation has already been realized by breaking the diffraction limit. Examples include the optical–magnetic hybrid disk storage systems, nanophotonic devices and systems, and photochemical vapor deposition and photolithography. However, it is important to note that these examples also realize qualitative innovation. Details of these examples are described in this chapter.

35.2 Excitation Energy Transfer in Nanophotonic Devices

Kagan et al. observed the energy transfer among CdSe quantum dots (QDs) coupled via a dipole–dipole inter-dot interaction [1]. Crooker et al. also studied the dynamics of the exciton energy transfer in close-packed assemblies of monodisperse and mixed-size CdSe nanocrystal QDs and reported the energy-dependent transfer rate of excitons from smaller to larger dots [2]. These examples are based on the optical near-field interaction. The physical model for the unidirectional resonant energy transfer between QDs via the optical near-field interaction has been presented, and the optically forbidden energy transfer among randomly dispersed CuCl QDs has been demonstrated experimentally using optical near-field spectroscopy [3]. The theoretical analysis and temporal evolution of the energy transfer via the optical near-field interaction were discussed in Ref. [4]. This section reviews the principles of nanophotonic devices and experimental works involving the direct observation of energy transfer from the exciton state in a CuCl QD to the optically forbidden exciton state in another CuCl QD using optical near-field spectroscopy.

Cubic CuCl QDs embedded in an NaCl matrix have the potential to be an optical near-field coupling system that exhibits the optically forbidden energy transfer. This is made possible because for this system, other forms of energy transfer, such as carrier tunneling and Coulomb coupling, can be neglected as the carrier wave function is localized in the QDs; this occurs because the potential depth exceeds 4 eV and the Coulomb interaction is weak due to the small exciton Bohr radius of 0.68 nm in CuCl [3]. The energy transfer via a propagating light is also negligible, since the optically forbidden transition in nearly perfect cubic CuCl QDs is used; that is, the transition to the confined exciton energy levels has an even principal quantum number [5].

So far, this type of energy transfer has not been observed directly because such a nanometric system is usually extremely complex. However, CuCl QDs in a NaCl matrix is a very simple system. The translational motion of the exciton center of mass is quantized due to the small exciton Bohr radius for CuCl QDs, and CuCl QDs become cubic in a NaCl matrix [6–8]. The potential barrier of CuCl QDs in a NaCl crystal can be regarded as infinitely high, and the energy eigenvalues E_{n_x,n_y,n_z} for the quantized Z_3 exciton energy level (n_x, n_y, n_z) in a CuCl QD with side length L depends on the values of quantum numbers n_x, n_y, n_z, and the effective side length $d = (L - a_B)$ found after considering the dead layer correction [6], where a_B is an exciton Bohr radius. The exciton energy levels with even quantum numbers are dipole-forbidden states, which are optically forbidden [5]. However, the optical near-field interaction is finite for such coupling to the forbidden energy state [9].

Figure 35.1 shows schematic drawings of different-sized cubic CuCl QDs (A and B) and the confined exciton Z_3 energy levels. Here, d and $\sqrt{2}d$ are the effective side lengths of cubic QDs A and B, respectively. The quantized exciton energy levels of (1,1,1) in QD A and (2,1,1) in QD B resonate with each other. Under this resonant condition, the coupling energy of the optical near-field interaction is given by the following Yukawa function [9,10]:

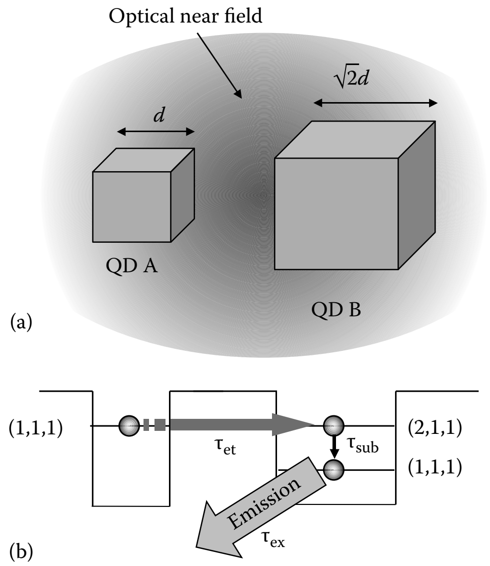

FIGURE 35.1 (a) Schematic drawings of closely located cubic CuCl QDs A and B with effective side lengths ($L - a_B$) of d and $\sqrt{2}d$, respectively, where L and a_B are the side lengths of the cubic quantum dots and the exciton Bohr radius, respectively. (b) Their exciton energy levels. n_x, n_y, and n_z represent quantum numbers of an exciton. E_B is the exciton energy level in a bulk crystal.

$$V(r) = \frac{A}{r}\exp\left(-\frac{\pi\sqrt{3m_p}}{a\sqrt{m_e}}r\right),$$

where

A is a proportional constant
r is the separation between the two QDs
a is the size of the QD
m_p is the effective mass of an exciton–polariton
m_e is the effective mass of an electron

Assuming that the two CuCl QDs in the NaCl matrix have side lengths 5 and 7 nm (a size ratio of $1:\sqrt{2}$) and the inter-dot distance is 6.1 nm, then the coupling energy $V(r)$ is 5.05 μeV. This corresponds to an energy transfer time of 130 ps due to the optical near-field coupling, which is much shorter than the exciton lifetime of a few nanoseconds. In addition, the inter-sublevel transition τ_{sub}, from higher exciton energy levels to the lowest, as shown in Figure 35.1, is generally less than a few picoseconds [11] and is much shorter than the transfer time τ_{et}. Therefore, most of the energy of the excitation in a cubic CuCl QD with a side length of d is transferred to the lowest exciton energy level in a neighboring QD with a side length of $\sqrt{2}d$ and recombines radiatively in the lowest level.

The CuCl QDs embedded in NaCl matrix used experimentally were fabricated using the Bridgman method and successive annealing, and the average size of the QDs was found to be 4.3 nm. The sample was cleaved just before the near-field optical spectroscopy experiment to keep the sample surface clean. The cleaved surface of a 100 μm-thick sample was sufficiently flat for the experiment; that is, its roughness was less than 50 nm, at least within a few microns squared. A 325 nm He–Cd laser was used as the light source. A double-tapered fiber probe was fabricated using chemical etching and a 150 nm gold coating was applied [12]. A 50 nm aperture was fabricated using the pounding method [13].

The spatial distributions of the luminescence intensity, that is, near-field optical microscope images, clearly show anti-correlation features in their intensity distributions. This anti-correlation feature can be clarified by noting that these spatial distributions in luminescence intensity represent not only the spatial distributions of the QDs, but also some kind of resonant interaction between the QDs. This interaction induces energy transfer from QDs A (L = 4.6 nm) to QDs B (L = 6.3 nm) because most of the 4.6 nm QDs located close to 6.3 nm QDs cannot emit light, but instead transfer the energy to the 6.3 nm QDs. As a result, in the region containing embedded 6.3 nm QDs, the luminescence intensity from 4.6 nm QDs is low, while the intensity from the 6.3 nm QDs is high at the corresponding position. This anti-correlation feature originates from the near-field energy transfer, which appears for every pair of QDs with different sizes to satisfy the resonant conditions of the confinement exciton energy levels. This is the first spatially resolved observation of energy transfer between QDs via an optical near field. This evidence of the near-field energy transfer between QDs can give rise to a variety of applications, as shown in the following sections.

35.3 Device Operation

Optical fiber transmission systems require increased integration of photonic devices for higher data transmission rates. Since conventional photonic devices, for example, diode lasers and optical waveguides, have to confine light waves within their cavities and core layers, respectively, their minimum sizes are limited by the diffraction of light [14]. Therefore, they cannot meet the size requirement, which is beyond this diffraction limit. An optical near field is free from the diffraction of light and enables the operation and integration of nanometric optical devices. That is, by using a localized optical near field as the carrier, which is transmitted from one nanometric element to another, the above requirements can be met. Based on this idea, nanometer-sized photonic devices have been proposed, which are called *nanophotonic devices* [15]. A nanometric all-optical AND gate (i.e., a nanophotonic switch) is one of the most important devices for realizing nanophotonic integrated circuits, and the operation of a nanophotonic AND gate has been already demonstrated using a coupled QD system.

A logic gate, for example, an AND gate and a NOT gate, is a block in a digital system. Logic gates have some inputs and some outputs, and every terminal is under one of two binary conditions, low (0) or high (1), given by different optical intensities for the optical device. The logic state of the input terminal is controlled by the optical input signal, and the logic state of the output terminal changes depending on the logic state of the input terminals. An intensity of approximately zero and a much higher intensity are preferable in the low and high logic states, respectively, such that the ratio of high to low intensity exceeds 30 db. For nanophotonic devices, the high state (1), which is called "true," is defined simply as the higher intensity state and the low state (0), which is called "false," is defined as the lower intensity state.

Section 35.3.1 presents the operation of nanophotonic AND gates using three CuCl QDs [16,17], while a nanophotonic NOT gate using CuCl QDs is outlined in Section 35.3.2. The optically forbidden energy transfer between neighboring nanostructures via the optical near-field interaction, which was reviewed in Section 35.2, is a key phenomenon for these operations.

35.3.1 Nanophotonic AND Gate

Operation of a nanophotonic AND gate using cubic CuCl QDs embedded in a NaCl matrix has been demonstrated [16,17]. When closely spaced QDs with quantized energy levels resonate with each other, near-field energy is transferred between them, even if the transfer is optically forbidden, as noted in Section 35.2. The output is "true" when both inputs are "true"; otherwise, the output is "false." Figure 35.2a and b explains the "false" and "true" states of the proposed nanophotonic AND gate. Three cubic QDs, QD$_{\text{input A}}$, QD$_{\text{input B}}$, and QD$_{\text{output}}$, are used as the two inputs and output ports of the AND gate, respectively. Assuming an effective size ratio of $1:\sqrt{2}:2$, the quantized energy levels (1,1,1) in QD$_{\text{input A}}$, (2,1,1) in QD$_{\text{output}}$, and (2,2,2) in QD$_{\text{input B}}$ resonate with each other. Furthermore, energy levels (1,1,1) in QD$_{\text{output}}$ and (2,1,1) in QD$_{\text{input B}}$ also resonate. In the "false" state operation (Figure 35.2a), for example, input A is "true" and input B is "false," almost all of the exciton energy in QD$_{\text{input A}}$ is transferred to the (1,1,1) level in the neighboring QD$_{\text{output}}$, and then to the (1,1,1) level in QD$_{\text{input B}}$. Therefore, the input energy escapes to QD$_{\text{input B}}$, and consequently no optical output signals are generated from QD$_{\text{output}}$. This means that the output is "false." In the "true" state (Figure 35.2a) when inputs A and B are both "true," the escape route to QD$_{\text{input B}}$ is blocked by the excitation of QD$_{\text{inputB}}$ due to state filling in QD$_{\text{input B}}$ on applying the input B signal. Therefore, the input energy is transferred to QD$_{\text{output}}$ and an optical output signal is generated. This means that the output is "true." These operating principles are realized with the condition $\tau_{ex} > \tau_{et} > \tau_{sub}$, where τ_{ex}, τ_{et}, and τ_{sub} are the exciton lifetime, energy transfer time between QDs, and inter-sublevel transition time, respectively. Since τ_{ex}, τ_{et}, and τ_{sub} are a few nanoseconds, 100 ps, and a few picoseconds, respectively, for the CuCl QDs used in a NaCl matrix, the condition of operation described in Section 35.2 is satisfied.

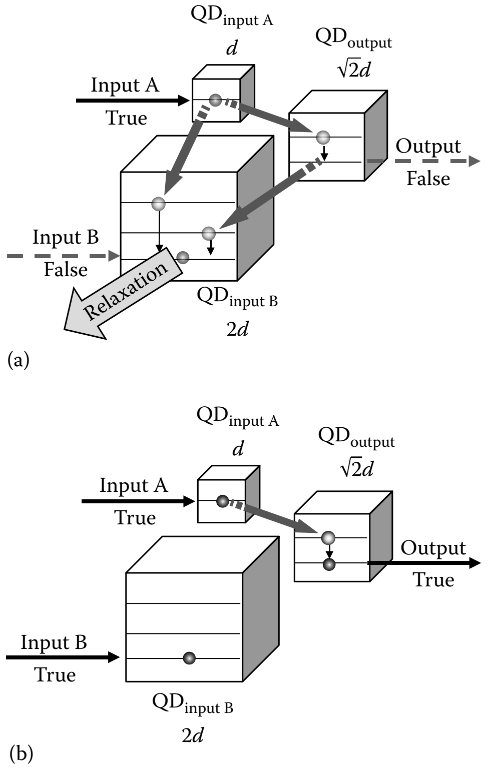

FIGURE 35.2 Principle of AND-gate operation. (a) and (b) The "false" and "true" states of the nanophotonic AND gate, respectively.

In an experiment using CuCl QDs embedded in a NaCl matrix, a double-tapered UV fiber probe was fabricated using chemical etching and coated with 150 nm-thick aluminum (Al) film. An aperture less than 50 nm in diameter was formed by the pounding method [13]. To confirm the AND-gate operation, the fiber probe was used to search for a trio of QDs that had an effective size ratio of $1:\sqrt{2}:2$. Since the homogeneous linewidth of a CuCl QD increases with the sample temperature [18,19], the allowance in the resonatable size ratio is 10% at 15 K. The separation of the QDs must be less than 30 nm for operation of the proposed AND gate because the energy transfer time increases with the separation; moreover, it must be shorter than the exciton lifetime. It is estimated that at least one trio of QDs that satisfies these conditions exists in a $2 \times 2\,\mu$m scan area. To demonstrate AND-gate operation, a QD trio had to be found in the sample, as shown in Figure 35.2.

Near-field photoluminescence (PL) pump-probe spectroscopy was used to find the QD trio. Figure 35.3 shows the PL spectrum obtained at the position where the QD trio exists. In this figure, peaks appear at the positions of the (1,1,1) levels in the 4.6 and 3.5 nm QDs. The appearance of the satellite peaks means that the AND-gate system proposed in Figure 35.2 was present in the area under the probe. In other words, a trio of cubic QDs with sizes of 3.5, 4.6, and 6.3 nm existed. Since their effective respective sizes $L - a_B$ were 2.8, 3.9, and 5.6 nm (a_B: an exciton Bohr radius of 0.7 nm in CuCl), the size ratio was close to $1:\sqrt{2}:2$ and they could be used as QD_{inputA}, QD_{output}, and QD_{inputB}, respectively. The pumping to the 6.3 nm QD blocks the energy transfer from the 3.5 and 4.6 nm QDs to the 6.3 nm QD due to state filling of the 6.3 nm QD, and the 3.5 and 4.6 nm QDs emit light that results in the peaks in Figure 35.3. Therefore, a QD trio for a nanophotonic AND gate was found. The PL peak from the 4.6 nm QD corresponds to the output signal in Figure 35.2b.

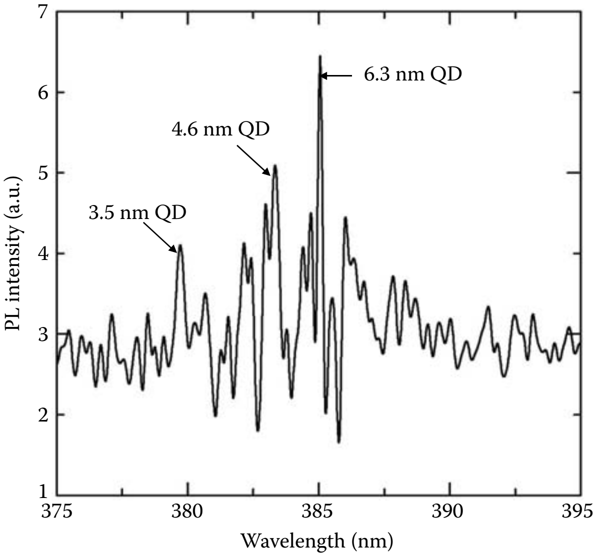

FIGURE 35.3 Near-field differential PL spectra measured at the position of a QD trio acting as a nanophotonic AND gate.

The PL intensity from the 4.6 nm QD was 0.05–0.02 times the PL intensity from the 6.3 nm QD, which was obtained with the probe laser only. This value is quite reasonable considering the pumping pulse energy density of $10\,\mu W/cm^2$ because the probability density of excitons in a 6.3 nm QD is 0.1–0.05 [18], which is close to the PL intensity from the 4.6 nm QD. This result indicates that the internal quantum efficiency of the AND-gate system is close to unity.

In the experiment examining the AND-gate operation, the second harmonics of Ti:sapphire lasers (wavelengths 379.5 and 385 nm), which were tuned to the (1,1,1) exciton energy levels of QD_{inputA} and QD_{inputB}, respectively, were used as the signal light sources for inputs A and B. The output signal was collected by the fiber probe, and its intensity was measured using a cooled microchannel plate after passing through three interference filters of 1 nm bandwidth tuned to the (1,1,1) exciton energy level in QD_{output} at 383 nm. Figure 35.4a and b shows the spatial distribution of the output signal intensity in the "false" state (i.e., with one input signal only) and in the "true" state (i.e., with both input A and input B signals) using near-field spectroscopy at 15 K. The insets in this figure are schematic drawings of the existing QD trio used for the AND gate, which was confirmed by the near-field PL spectra. Here, separation of the QDs by less than 20 nm was estimated theoretically from time-resolved PL measurements, as explained in the next paragraph (see the illustrations in Figure 35.4). In the "false" state, no output signal was observed because the energy of the input signal was transferred to QD_{inputB} and swept out as PL at 385 nm. To quench the output signal in the "false" state, which was generated by accumulating excitons in QD_{inputB}, the input signal density to QD_{inputA} was regulated to less than 0.1 excitons in QD_{inputB}. In the "true" state, a clear output signal was obtained in the dashed circle of Figure 35.4b. The output signal was proportional to the intensity of the control signal, which had a density of 0.01–0.1 excitons in QD_{inputB}.

Next, the dynamic properties of the nanophotonic AND gate were evaluated using the time-correlation single-photon counting method. As a pulse-input B signal source, the 385 nm second harmonic of a mode-locked Ti:sapphire laser was used. The repetition rate of the laser was 80 MHz. To avoid cross talk originating from spectral broadening of the pulse duration between input signals

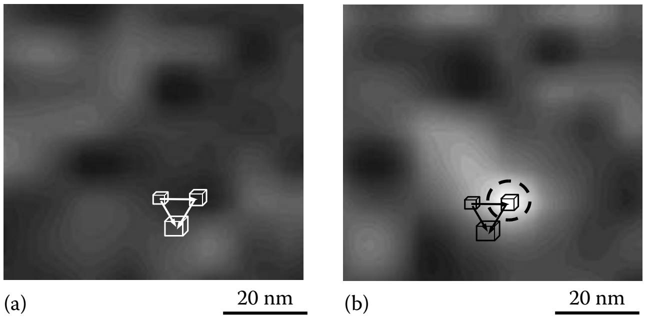

FIGURE 35.4 Spatial distribution of the output signal from the nanophotonic AND gate in the "false" (a) and "true" (b) states measured using near-field microscopy.

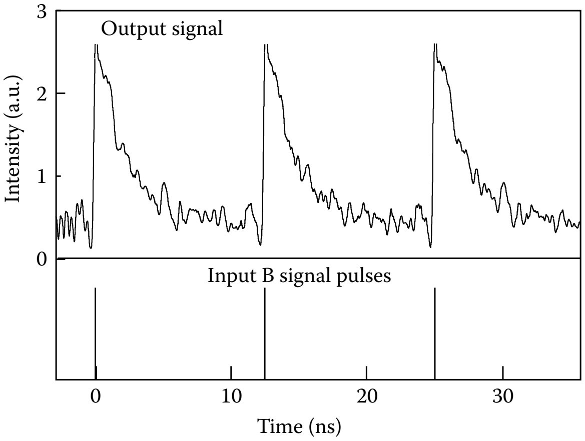

FIGURE 35.5 Temporal evolution of the output signal from the nanophotonic AND gate located in the dashed circle in Figure 35.4b. The duration and repetition rate of the control pulse were 10 ps and 80 MHz, respectively.

A and B, the pulse duration of the mode-locked laser was set to 10 ps. The time resolution of the experiment was 15 ps. Figure 35.5 shows the temporal evolution of the input B pulse signal (lower part) applicable to QD_{inputB} and the output signal (upper part) from QD_{output}. The output signal increases synchronously with the input B pulse within less than 100 ps, which agrees with the theoretically expected result based on the Yukawa model [9]. As this signal rise time is determined by the energy transfer time between the QDs, the separation between the QDs can be estimated from the rise time as being less than 20 nm; the rise time can be shortened to a few picoseconds by decreasing the separation of the QDs. Since the decay time of the output signal is limited by the exciton lifetime, this nanophotonic AND gate can be operated at a few hundred megahertz, and the operating frequency can be increased to several gigahertz by exciton quenching using plasmon coupling [20]. The output signal ratio between "true" and "false" was about 10, which is sufficient for use as an all-optical AND gate, and can be increased using a saturable absorber and electric field enhancement of the surface plasmon [21].

The advantages of this nanophotonic device are its small size and high-density integration capability based on the locality of the optical near field. The figure of merit (FOM) of an optical AND gate should be more important than the switching speed. Here, the FOM is defined as $F = C/Vt_{sw}P_{sw}$, where C, V, t_{sw}, and P_{sw} are the "true"–"false" (ON–OFF) ratio, volume of the device, switching time, and switching energy, respectively. The FOM of the nanophotonic AND gate is 10–100 times higher than that of conventional photonic gates because its volume and switching energy are 10^{-5} times and 10^{-3} times those of conventional photonic gates, respectively.

35.3.2 Nanophotonic NOT Gate

A nanophotonic NOT gate is a key device for realizing a functionally complete set of logic gates for nanophotonic systems, and its operation is demonstrated in this section using CuCl QDs [22]. Figure 35.6 shows a schematic explanation of the nanophotonic NOT gate. QD_{IN} and QD_{OUT} correspond to the input and output terminals of the NOT gate, respectively. Assuming a pair of QDs with a size ratio of $1+\alpha:\sqrt{2}\,(\alpha \ll 1)$, the quantized energy levels with the set of quantum numbers (2,1,1) in QD_{IN} and (1,1,1) in QD_{OUT} are slightly nonresonant with each other. The energy from the optical power supply generates an exciton in QD_{OUT}. Without the input signal (i.e., the input is "false"), the exciton in QD_{OUT} disappears and emits a photon, which is observed as an output signal, as shown in Figure 35.6a. That is, input = "false" and output = "true." Conversely, by applying the input signal (i.e., the input is "true"), the energy level (2,1,1) in QD_{IN} becomes resonant to (1,1,1) in QD_{OUT} due to broadening of its line width. This broadening was confirmed experimentally for CuCl QDs; anti-holes appear on both sides of the spectral hole in the absorption spectrum of CuCl QDs observed in a far-field hole-burning experiment at 5 K. This experimental result shows that the excitation broadens the homogeneous line width of QDs, which might arise from shortening of the phase relaxation time of the excitons in QDs due to carrier–carrier scattering. Consequently, the exciton energy in QD_{OUT} is transferred to QD_{IN} via an optical near-field interaction [3], which suppresses output signal generation (Figure 35.6b, i.e., input = "true" and output = "false"). As a result, the temporal behavior of the output signal is the inverse of that of the input signal. By selecting a suitable threshold to distinguish "true" and "false," these behaviors are used for a NOT gate.

CuCl QDs embedded in a NaCl matrix were used to verify operation of the NOT gate, as CuCl QDs offer discrete energy levels similar to the exciton described in Figure 35.6 [6]. The

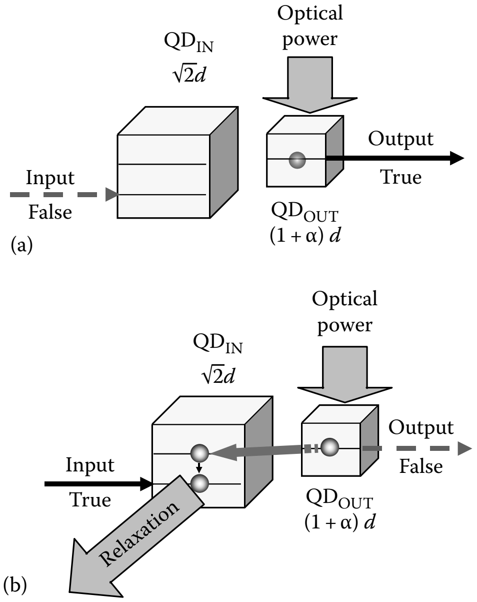

FIGURE 35.6 A nanophotonic NOT gate. (a) and (b) Schematic explanation of the "true" and "false" states using cubic QDs.

mean size of the QDs was 4.1 nm and the mean distance between the QDs was 25 nm. In the experiment, the second harmonics of a continuous-wave (CW) Ti:sapphire laser ($\hbar\omega$ = 3.2704 eV) and a mode-locked Ti:sapphire laser ($\hbar\omega$ = 3.2195 eV) were used as the optical power supply and input signal pulse, respectively. The respective power densities were 1 and 2 W/cm^2 at the sample surface. Under the excitation condition, fewer than 0.1 excitons occurred in a QD. These lasers excited the sample from its back and the output signal was observed using a near-field spectrometer in collection mode. The sample temperature was controlled at 15 K. To find the QD pair acting as a nanophotonic NOT gate, the QD positions were mapped by measuring the luminescence distribution on the sample, which was obtained using a near-field spectrometer using He–Cd laser excitation ($\hbar\omega$ = 3.81 eV). The signal was observed after several QD pairs were selected as candidates for nanophotonic NOT-gate operation. Finally, a nanophotonic NOT gate was found, with the probability of about one device per 1 µm^2 of scanning area. Figure 35.7a shows the spatial distribution of the optical near-field output-signal intensity when an input signal is "false" (i.e., with the optical power supply only). Figure 35.7b shows the distribution when the input signal is "true." The insets in this figure are schematic drawings of an existing QD pair that function as a NOT gate, which was confirmed from the PL spectra. The sizes of the two QDs estimated from the wavelengths of their luminescence were 5.0 and 6.3 nm, which satisfy the NOT-gate operation condition, as shown in Figure 35.6. Note that the photon energy of the optical power supply ($\hbar\omega$ = 3.2704 eV) is maintained nonresonant to the (1,1,1) exciton level in the 5.0 nm QD ($\hbar\omega$ = 3.2304 eV) to decrease the artifact originating from the laser by using narrowband optical filters and observing a clear output signal. Sufficiently low optical power was supplied to the 5.0 nm QD from neighboring QDs [3,23]. The NOT-gated signal appears at the center of the dashed circle in Figure 35.7a, from which the size of the device was estimated to be 20 nm. Comparison of Figure 35.7a and b clearly demonstrates the operation of a NOT gate.

The dynamic behavior of the NOT gate was observed using the time-correlation single-photon counting method. Figure 35.8 shows the temporal evolution of the output signal. The horizontal dashed line indicates the output signal level without the input signal pulses (i.e., the input is "false"). Without the input pulses, the signal level is constant, since a CW laser was used as the optical power supply. This signal level is defined as "true." With the input pulses, the output signal increases within a time period shorter than the time resolution of 20 ps due to the artifact of the input pulses, and it decreases to a level lower than the initial level. Here, the signal level lower than the dotted line is defined as "false." Then, the output signal level becomes "false" 50 ps after the "true" input pulse. The fall time of the output signal to the minimum level is about 100 ps, which corresponds to the energy transfer time between QDs. The "false" output signal level recovers to the "true" level within 10 ns. The recovery time is longer than the exciton lifetime in 6.3 nm CuCl QDs (τ_{ex} ~ 1 ns) because the energy transfer from the optical power supply affects the recovery. That is, the recovery time depends on the competition between exciton annihilation via recombination and exciton creation via energy transfer.

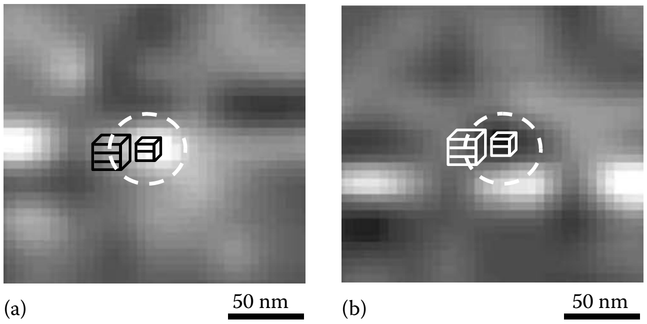

FIGURE 35.7 The spatial distribution of the output signal from a nanophotonic NOT gate measured using near-field microscopy at Input = "false" (a) and Input = "true" (b).

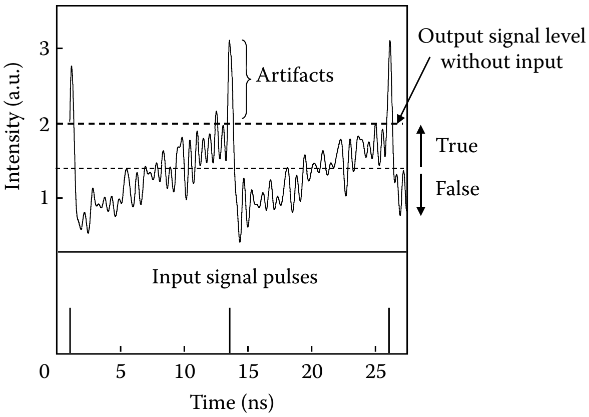

FIGURE 35.8 The temporal evolution of the output signal from the nanophotonic NOT gate circled in Figure 35.7.

35.3.3 Interconnection with Photonic Devices

An interconnection device needs to be developed to collect the incident propagating light and drive the nanophotonic device for efficient operation of the system [15,24,25]. Conventional far-field optical devices, such as convex lenses and concave mirrors, cannot be used for this interconnection because of their diffraction-limited operation. This section demonstrates a novel optical device, the *optical nanofountain*, which concentrates optical energy in a nanometric region using optical near-field energy transfer among QDs. This nanometric optical device enables highly efficient interconnection to nanophotonic devices. The optical nanofountain is operated using the energy transfer between QDs via an optical near-field interaction, as shown in Figure 35.1 [9–11]. The unidirectional energy transfer from smaller to larger QDs concentrates the optical energy in a nanometric region. When different sized QDs with resonant energy sublevels are distributed, as shown in Figure 35.9a, energy transfer occurs via the optical near field, as illustrated by the arrows. Light incident to the QD array is ultimately concentrated in

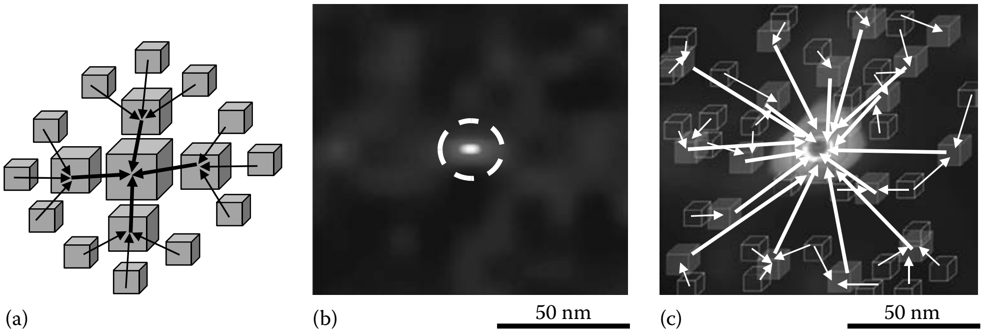

FIGURE 35.9 Optical nanofountain. (a) Schematic explanation of the optical nanofountain and unidirectional energy transfer. (b) Spatial distribution of the PL intensity in an optical nanofountain. The bright spot surrounded by a dashed circle is the focal spot. (c) Spatial distribution of the PL intensity of CuCl QDs for 2.5 nm ≤ L ≤ 10 nm (3.215 eV ≤ E_p ≤ 3.350 eV) for the same area as in (b). The cubes represent the positions estimated from the PL intensity distribution.

the largest QD. The area of optical energy concentration corresponds to the size of the QD.

The name optical nanofountain was proposed because light spurts from the largest QD after it is concentrated by stepwise energy transfer from smaller neighboring QDs, so that this device looks like a fountain in a basin. From the experimental tests of nanophotonic AND-gate operation in Section 35.3.1, the concentration efficiency of this device should be close to unity because no other possible relaxation paths exist in the system. The operation of an optical nanofountain was demonstrated using CuCl cubic QDs embedded in a NaCl matrix. The average QD size was 4.2 nm and the average separation was less than 20 nm. Although the QDs have an inhomogeneous size distribution and are arranged in the matrix randomly, the operation can be confirmed if an appropriate QD group is found using a nanometric-resolution near-field spectrometer. For the operation, the optimum sample temperature T was 40 K. At $T < 40$ K, the resonant condition becomes tight due to narrowing of the homogeneous line width of the quantized energy sublevels, while at $T > 40$ K, the unidirectional energy transfer is obstructed by the thermal activation of excitons in the QDs. A 325 nm He–Cd laser was used as the excitation light source. A double-tapered UV fiber probe with an aperture 20 nm in diameter was fabricated using chemical etching and coated with a 150 nm-thick Al film to ensure sufficiently high detection sensitivity [12].

Figure 35.9b shows the typical spatial distribution of the PL from QDs operating as an optical nanofountain. The bright spot inside the dashed circle corresponds to a spurt from an optical nanofountain, i.e., the focal spot of the nanometric optical condensing device. The diameter of the focal spot was less than 20 nm, which was limited by the spatial resolution of the near-field spectrometer. From the Rayleigh criterion (i.e., resolution = $0.61 \cdot \lambda/\mathrm{NA}$) [26], its numerical aperture (NA) was estimated to be 12 for $\lambda = 385$ nm. To demonstrate the detailed operating mechanism of the optical nanofountain, we show the size-selective PL intensity distribution; that is, the photon energy is shown in Figure 35.9c. Here, the collected PL photon energy, E_p, was 3.215 eV ≤ E_p ≤ 3.350 eV, which corresponds to the PL from QDs of size 2.5 nm ≤ L ≤ 10 nm. The area scanned by the probe are equivalent to those in Figure 35.9b. The PL intensity distribution is shown using a gray scale.

This device can also be used as a frequency selector based on the resonant frequency of the QDs, which can be applied, for example, to frequency domain measurements, multiple optical memories, multiple optical signal processing, and frequency division multiplexing.

Practical nanophotonic devices for room-temperature operation are under development using III–V compound semiconductor QDs [27] and ZnO nanorods [28].

35.4 Nanophotonics Fabrication

This section presents the nonadiabatic processes involved in optical chemical vapor deposition (CVD) and photolithography. These methods have realized qualitative innovation in nanofabrications by utilizing the spatially localized nature of optical near fields.

35.4.1 Nonadiabatic Near-Field Optical CVD

Conventional optical CVD utilizes a two-step process: photodissociation and adsorption. For photodissociation, a propagating light must resonate the reacting molecular gases to excite molecules from the ground state to an excited electronic state. The Franck–Condon principle holds that this resonance is essential for excitation. The excited molecules then relax to the dissociation channel, and the dissociated atoms adsorb to the substrate surface. However, a nonadiabatic photodissociation process is observed in near-field optical chemical vapor deposition (NFO-CVD) under the nonresonant condition of the electronic transition, which violates the Franck–Condon principle. This section discusses the nonadiabatic NFO-CVD of nanometric Zn dots and presents experimental results based on the exciton–phonon–polariton (EPP) model.

Figure 35.10 shows the cross-sectional profiles of the shear-force topographical images after NFO-CVD for photon energies of 5.08 eV ($\lambda = 244$ nm; broken curve) and 2.54 eV ($\lambda = 488$ nm; solid curve) of Zn dots deposited on a sapphire substrate in

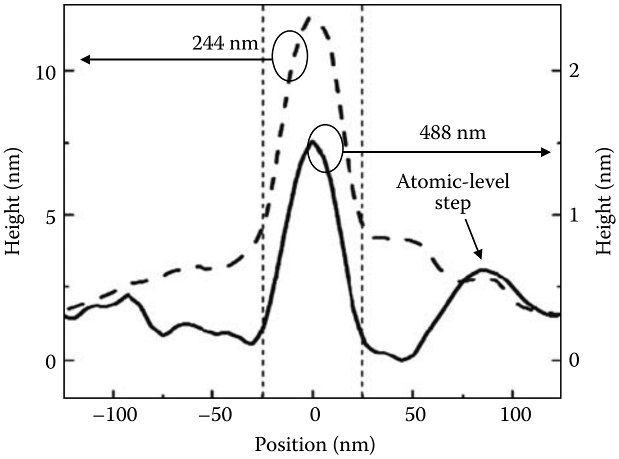

FIGURE 35.10 Cross-sectional profiles of the shear-force topographical images of the deposited Zn patterns.

atomic-level steps [29]. In the experiment, diethylzinc (DEZn) was used as the CVD gas source. For the broken curve, the laser power was 1.6 μW and the irradiation time was 60 s. Before carrying out NFO-CVD, 0.4 nm-high atomic-level step structures were clearly observed on the sapphire substrate. After NFO-CVD, they disappeared and a deposited Zn dot less than 50 nm in diameter was seen at the center of the image. This occurred because the optical near field deposited the Zn dot directly under the apex of the fiber probe. Furthermore, since high-intensity propagating light leaks from a bare fiber probe, that is, one without a metallic coating, and is absorbed by the DEZn, a Zn layer was deposited on top of the atomic-step structures.

For the solid curve, the laser power was 150 μW and the irradiation time was 75 s. The photon energy (2.54 eV) was higher than the dissociation energy of DEZn, but it was still lower than the absorption edge of DEZn [30]. Therefore, it was not absorbed by DEZn. However, a Zn dot less than 50 nm in diameter appears at the center. While using conventional CVD with propagating light, a Zn film cannot be grown using a light source with a photon energy lower than the absorption edge (4.13 eV: λ = 300 nm) [31]. Deposited Zn dots were observed on the substrate just below the apex of the fiber probe using NFO-CVD. The atomic-level steps in this figure are still observed, despite the leakage of the propagating light from the bare fiber probe. These curves confirm that Zn dots with a full-width at half-maximum (FWHM) of 30 nm were deposited in the region where the optical near field is dominant.

The dashed curve has 4 nm-high tails on both sides of the dot, which represent the deposition caused by the leaked propagating light. This deposition process is based on the conventional adiabatic photochemical process. In contrast, the solid curve has no tails; therefore, it is clear that the leaked 488 nm propagating light did not deposit a Zn layer. Note that a Zn dot 30 nm in diameter without tails was deposited under nonabsorbed conditions (λ = 488 nm).

Figure 35.11 shows shear-force topographical images of the sapphire substrate after NFO-CVD using an optical near field with photon energies of 3.81 eV (λ = 325 nm) (a), 2.54 eV (λ = 488 nm) (b), and 1.81 eV (λ = 684 nm) (c). The respective laser power and irradiation time were (a) 2.3 μW and 60 s, (b) 360 μW and 180 s, and (c) 1 mW and 180 s. The high quality of the deposited Zn was confirmed by x-ray photoelectron spectroscopy. Furthermore, PL was observed from ZnO dots, which were fabricated by oxidizing the Zn dots deposited by NFO-CVD [8]. In Figure 35.11a, the photon energy ($\hbar\omega$) exceeds the dissociation energy (E_d) of DEZn, and is close to the absorption band edge (E_{abs}) of DEZn, i.e., $\hbar\omega > E_d$ and $\hbar\omega \cong E_{abs}$ [30]. The diameter (FWHM) and height of the topographical image were 45 and 26 nm, respectively. The small tail (shown by the dotted curve) represents a Zn layer less than 2 nm thick, and was deposited by the propagating light leaking from the bare fiber probe. This deposition is possible because the DEZn absorbs some of the propagating light at $\hbar\omega = 3.18$ eV. The very high peak suggests that the optical near field enhances the photodissociation rate at this photon energy because its intensity increases rapidly at the apex of the fiber probe. In Figure 35.11b, the photon energy still exceeds the dissociation energy of DEZn, but is lower than the absorption band edge of DEZn, i.e., $E_{abs} > \hbar\omega > E_d$ [30]. The diameter and height of the image were 50 and 24 nm, respectively. This image has no tail because Zn was not deposited by the high-intensity propagating light leaking from the bare fiber probe. This confirmed that the photodissociation of DEZn and Zn deposition occurred only with an optical near field of $\hbar\omega = 2.54$ eV. Figure 35.11c represents the cases $\hbar\omega < E_d$ and $\hbar\omega < E_{abs}$. Zn dots were deposited successfully at these low photon energies.

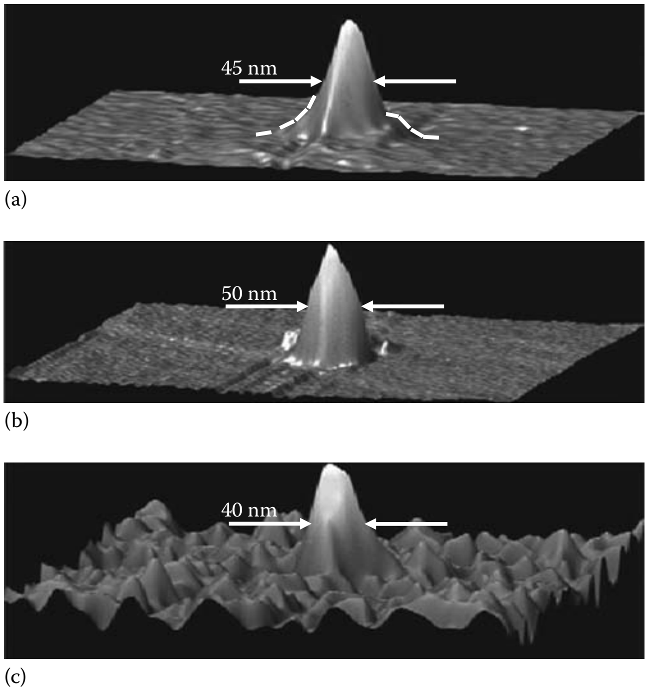

FIGURE 35.11 Shear-force topographical images after NFO-CVD at wavelengths of λ = 325 nm (a), 488 nm (b), and 684 nm (c). The laser output power and irradiation time for deposition were 2.3 μW and 60 s (a), 360 μW and 180 s (b), and 1 mW and 180 s (c), respectively.

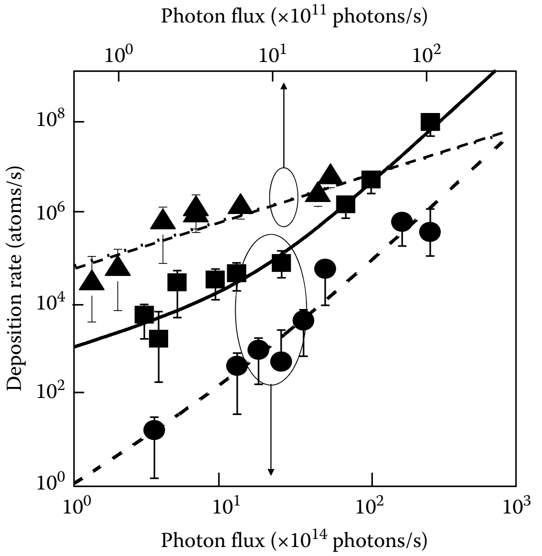

FIGURE 35.12 The relationship between the photon flux and the rate of Zn deposition. The dotted, solid, and dashed curves represent the calculated values fitted to the experimental results.

The topographical image showed dots with a diameter of 40 nm and a height of 2.5 nm. The experimental results in Figure 35.11 demonstrate dissociation based on a nonadiabatic photochemical process that violates the Franck–Condon principle.

To discuss this novel dissociation process quantitatively, Figure 35.12 shows the relationship between the photon flux (I) and the deposition rate of Zn (R). For $\hbar\omega = 3.81$ eV (▲), R is proportional to I. For $\hbar\omega = 2.54$ eV (■) and 1.81 eV (●), higher order dependencies appear and R is fitted by the third-order function $R = a_{\hbar\omega}I + b_{\hbar\omega}I^2 + c_{\hbar\omega}I^3$. The respective values of $a_{\hbar\omega}$, $b_{\hbar\omega}$, and $c_{\hbar\omega}$ are $a_{3.81} = 5.0 \times 10^{-6}$, $b_{3.81} = 0$, and $c_{3.81} = 0$ for $\hbar\omega = 3.81$ eV, $a_{2.54} = 4.1 \times 10^{-12}$, $b_{2.54} = 2.1 \times 10^{-27}$, and $c_{2.54} = 1.5 \times 10^{-42}$ for $\hbar\omega = 2.54$ eV, and $a_{1.81} = 0$, $b_{1.81} = 4.2 \times 10^{-29}$, and $c_{1.81} = 3.0 \times 10^{-44}$ for $\hbar\omega = 1.81$ eV. The results of fitting are shown with the solid, dashed, and dotted curves in Figure 35.12. Since no conventional photochemical processes, for example, the Raman process and two-photon absorption, can explain these experimental results, the discussion below uses a unique theoretical model based on the discussion in Ref. [32].

Figure 35.13 shows the potential curves of an electron in a DEZn molecular orbital drawn as a function of the internuclear distance of the C–Zn bond, which is involved in photodissociation [30]. The relevant energy levels of the molecular vibration mode are indicated by the horizontal broken lines in each potential curve. When a propagating light is used, photo-absorption (indicated by the white arrow) triggers the dissociation of DEZn [33]. With an optical near field nonresonant to the electronic state, there are three possible origins of photodissociation [34]: (1) the multiple-photon absorption process; (2) a multiple-step transition process via the intermediate energy level induced by the fiber probe; and (3) multiple-step transition via an excited state of the molecular vibration mode. Case (1) is negligible because the optical power density was less than 10 kW/cm². Case (2) is also negligible because the DEZn was dissociated by ultraviolet–near-infrared light, although DEZn does not have relevant energy levels over such a broad wavelength region. As a result, the experimental results strongly support Case (3). That is, the photodissociation is caused by a transition to an excited state via a molecular vibration mode, which involves three multiple-step excitation processes, as shown in Figure 35.13. Since the system is strongly coupled with the vibration state, it must be considered a nonadiabatic system.

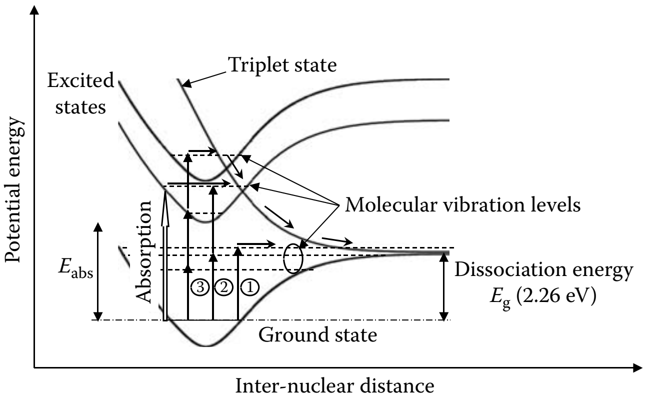

FIGURE 35.13 Potential curves of an electron in DEZn molecular orbitals. The relevant energy levels of the molecular vibration modes are indicated by the horizontal broken lines.

For this consideration, an exciton–phonon polariton (EPP) model was presented in Ref. [32]. The EPP model holds that the optical near fields excite the molecular vibration mode due to the steep spatial gradient. Since the optical near-field energy distribution is spatially inhomogeneous in a molecule due to its gradient, the electrons respond inhomogeneously. As a result, the molecular vibration modes are excited because the molecular orbital changes and the molecule is polarized as a result of the inhomogeneous response of the electrons. The EPP model describes this excitation process quantitatively. The EPP is a quasiparticle, which is an exciton–polariton carrying the phonon (lattice vibration) generated by the steep spatial gradient of the optical field energy distribution. In contrast, since the propagating light energy distribution is homogeneous in a molecule, only the electrons in the molecule respond to the electric field of the propagating light. Therefore, the propagating light cannot excite the molecular vibration.

Zinc-bis(acetylacetonate) ($Zn(acac)_2$) has never been used for conventional optical CVD due to its low optical activity. With NFO-CVD, however, the optical near field can activate the molecule nonadiabatically and the dissociated Zn atom is adsorbed under the fiber probe. Figure 35.14a shows a shear-force topographical image of Zn deposited on a sapphire substrate. The laser power and irradiation time were 1 mW and 15 s, respectively. The Zn dot was 70 nm in diameter and 24 nm high [35,36]. The chemical stability of $Zn(acac)_2$ keeps the substrate surface clean and helps to fabricate an isolated nanostructure. Figure 35.14b shows the shear-force topographical image of a deposited Zn dot that is among the smallest ever fabricated using

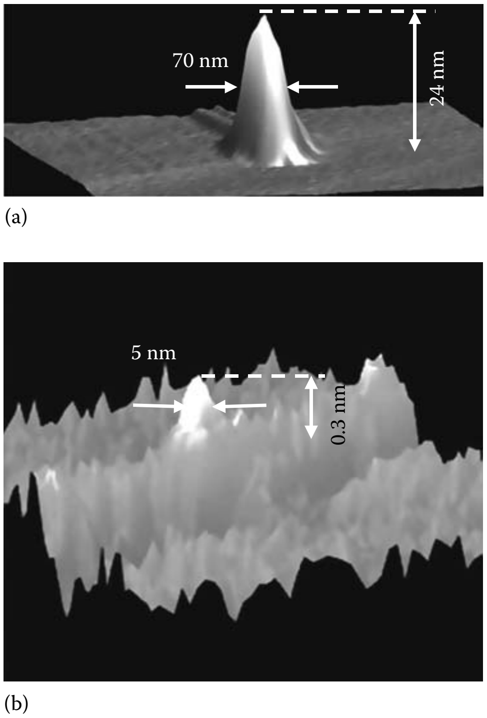

FIGURE 35.14 Shear-force topographical images after NFO-CVD using $Zn(acac)_2$ with a 457 nm-wavelength light source. (a) A deposited Zn dot with a diameter of 70 nm and height of 24 nm. (b) A deposited Zn with a diameter of 5 nm and height of 0.3 nm.

NFO-CVD (5 nm in diameter and 0.3 nm high). The deposition conditions consisted of $Zn(acac)_2$ at a pressure of 70 mTorr in the CVD chamber and a laser wavelength, power, and irradiation time of 457 nm, 65 μW, and 30 s, respectively.

35.4.2 Nonadiabatic Near-Field Photolithography

Section 35.4.1 reviewed a unique nonadiabatic photochemical reaction, which was explained using the EPP model. According to this model, the nonadiabatic photochemical reaction can be considered a universal phenomenon and is applicable to several photochemical processes. This section reviews the application of the nonadiabatic photochemical reaction to photolithography, which can be called *nonadiabatic photolithography* [36,37]. For the mass production of photonic and electronic devices, nonadiabatic photolithography can be used because conventional photolithographic components can be applied to this system.

The wave properties of propagating light cause problems for high-resolution photolithography due to diffraction and the dependence on the coherency and polarization of the light source. To fabricate high-density corrugations, the optical coherent length is too long compared to the separation between adjacent corrugation elements, even when a Hg lamp is used. In addition, the absorption by the photoresist is insufficient to suppress interference of scattered light. Furthermore, since the intensity of the propagating light transmitted through a photomask strongly depends on its polarization, the photomask must be designed while considering these dependences. In contrast, the outstanding advantage of nonadiabatic photolithography is that it is free from these problems.

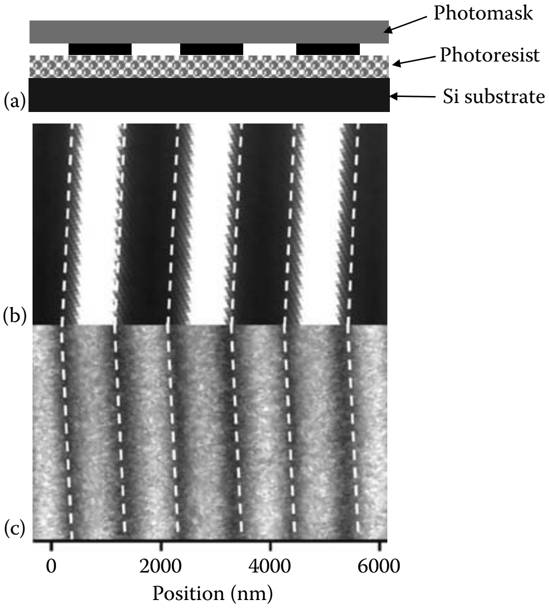

FIGURE 35.15 Experimental results of nonadiabatic photolithography. (a) A schematic of the photomask and Si substrate on which the photoresist (OFPR-800) was spin-coated. (b) Atomic force microscopy images of photoresist OFPR-800 exposed to the g-line of a Hg lamp. (c) AFM images of photoresist OFPR-800 developed after a 4 h exposure with a 672 nm laser.

Figure 35.15a shows a schematic configuration of the photomask and the Si substrate on which the photoresist (OFPR-800: Tokyo-Ohka Kogyo) was spin-coated. They were used in contact mode. Figure 35.15b and c shows atomic force microscopy (AFM) images of the photoresist surface after development. Figure 35.15b shows the result obtained using conventional photolithography. The g-line (436 nm) from a Hg lamp was used as the light source. The fabricated pattern of corrugation was an exact replica of the photomask. Conversely, with nonadiabatic photolithography using a 672 nm-wavelength light source, the grooves on the photoresist appeared along the edges of the Cr mask pattern, as shown in Figure 35.15c. The corrugated pattern was 30 nm deep. The line width was 150 nm, which was narrower than the wavelength of the light source. On the photomask, a steep spatial gradient of optical energy distribution is expected due to optical near fields, while direct irradiation with 672 nm light cannot expose the photoresist. This demonstrated that the photoresist was patterned using a nonadiabatic process.

Figure 35.16 shows AFM images of another photoresist surface (TDMR-AR87 for the 365 nm-wavelength i-line from a Hg lamp: Tokyo-Ohka Kogyo) after development. Figure 35.16a shows the corrugated pattern fabricated using linearly polarized g-line light. Two-dimensional arrays of circles and T-shapes have also been fabricated successfully on this photoresist

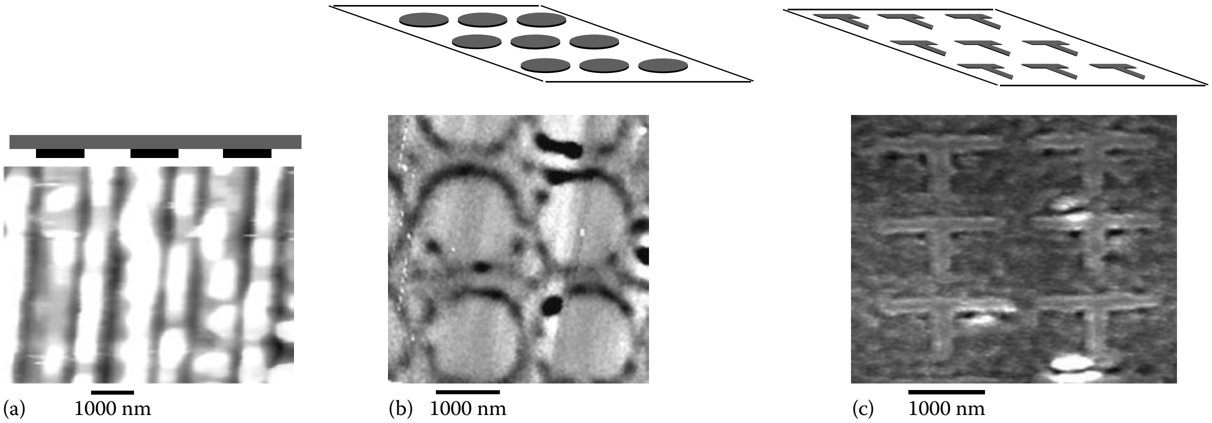

FIGURE 35.16 Experimental results of nonadiabatic photolithography. (a) AFM images of photoresist TDMR-AR87 exposed to the linearly polarized g-line of a Hg lamp for 3 s. (b) AFM images of photoresist TDMR-AR87 exposed to the linearly polarized g-line of a Hg lamp for 10 s using a circle-shaped array photomask. (c) AFM images of photoresist OTDMR-AR87 developed after a 40 s exposure to the g-line of a Hg lamp using a T-shaped array photomask.

(see Figure 35.16b and c). This would be impossible using adiabatic photolithography due to its polarization-dependent nature and interference effects.

An optically inactive electron beam (EB) resist film (ZEP-520: ZEON) can also be patterned nonadiabatically. Figure 35.17 shows an AFM image of the developed EB resist surfaces. The light source was the third harmonic of a Q-switched Nd:YAG laser and the exposure time was 5 min. A two-dimensional array of 1 µm-diameter disks was fabricated successfully, even on the EB resist, which would be impossible using propagating light. The developed pattern had a depth of 70 nm, which is sufficient for the subsequent etching of the substrate. Since the EB resist film has an extremely smooth surface, the homogeneity in the contact with the photomask was improved. This suggests that a smooth organic or inorganic thin film can be used as a photoresist irrespective of its optical inactivity.

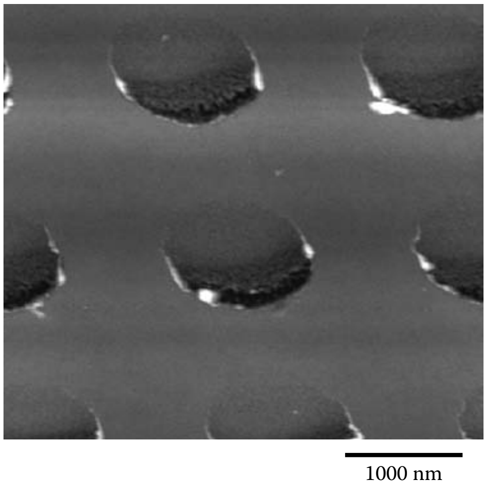

FIGURE 35.17 AFM image of the surface of an electron beam resist exposed for 5 min using a Q-switched laser (355 nm) and a circle-shaped (1 µm diameter) array photomask.

35.5 Summary

After nanophotonics was proposed by M. Ohtsu in 1993 [38], it now exists as a novel field of optical technology in nanometric space. However, the name “nanophotonics” is occasionally used for photonic crystals [39], plasmonics [40], metamaterials [41,42], silicon photonics [43], and QD lasers [44] using conventional propagating lights. For example, plasmonics utilizes the resonant enhancement of the light in a metal by exciting free electrons. The letters “on” in the word “plasmon” represent the quanta, or the quantum mechanical picture of the plasma oscillation of free electrons in a metal. However, plasmonics utilizes the classical wave optical picture using conventional terminology, such as the refractive index, wave number, and guided mode. Even when a metal is irradiated with light that obeys the laws of quantum mechanics, the quantum mechanical property is lost because the light is converted into the plasma oscillation of electrons, which has a short phase relaxation time. Furthermore, the energy transferred via this interaction must be dissipated in the nanometric particles or adjacent macroscopic materials to fix the position and magnitude of the transferred energy. Since plasmonics does not deal with this local dissipation of energy, it is irrelevant for quantitative innovation by breaking the diffraction limit, or for qualitative innovation. Local energy transfer and its subsequent dissipation have become possible only in nanophotonics by using optical near fields [45,46].

Here, we should consider the stern warning by C. Shannon on the casual use of the term “information theory,” which was a trend in the study of information theory during the 1950s [47]. The term “nanophotonics” has been used in a similar way, although some work in “nanophotonics” is not based on optical near-field interactions. For the true development of nanophotonics, one needs deep physical insights into the virtual exciton–polariton and the nanometric subsystem composed of electrons and photons.

References

1. Kagan, C. R., Murray, C. B., Nirmal, M., and Bawendi, M. G., 1996. Electronic energy transfer in CdSe quantum dot solids. *Phys. Rev. Lett.* **76**: 1517–1520.
2. Crooker, S. A., Hollingsworth, J. A., Tretiak, S., and Klimov, V. I., 2002. Spectrally resolved dynamics of energy transfer in quantum-dot assemblies: Towards engineered energy flows in artificial materials. *Phys. Rev. Lett.* **89**: 186802-1–186802-4.
3. Kawazoe, T., Kobayashi, K., Lim, J., Narita, Y., and Ohtsu, M., 2002. Direct observation of optically-forbidden energy transfer between CuCl quantum cubes via optical near-field. *Phys. Rev. Lett.* **88**: 067404-1–067404-4.
4. Sangu, S., Kobayashi, K., Shojiguchi, A., Kawazoe, T., and Ohtsu, M., 2006. Theory and principles of operation of nanophotonic functional devices. In M. Ohtsu (ed.) *Progress in Nano-Electro-Optics V*, Springer-Verlag, New York, pp. 1–62.
5. Tang, Z. K., Yanase, A., Yasui, T., Segawa, Y., and Cho, K., 1993. Optical selection rule and oscillator strength of confined exciton system in CuCl thin films. *Phys. Rev. Lett.* **71**: 1431–1434.
6. Sakakura, N. and Masumoto, Y., 1997. Persistent spectral-hole-burning spectroscopy of CuCl quantum cubes. *Phys. Rev. B* **56**: 4051–4055.
7. Ekimov, A. I., Eflos, AI. L., and Onushchenko, A. A., 1985. Quantum size effect in semiconductor microcrystals. *Solid State Commun.* **56**: 921–924.
8. Itoh, T., Yano, S., Katagiri, N., Iwabuchi, Y., Gourdon, C., and Ekimov, A. I., 1994. Interface effect on the properties of confined excitons in CuCl microcrystals. *J. Lumin.* **60&61**: 396–399.
9. Kobayashi, K., Sangu, S., Ito, H., and Ohtsu, M., 2001. Near-field optical potential for a neutral atom. *Phys. Rev. A* **63**: 013806-1–013806-9.
10. Ohtsu, M., 1998. *Near-Field Nano/Atom Optics and Technology*. Springer, Tokyo, Japan.
11. Suzuki, T., Mitsuyu, T., Nishi, K., Ohyama, H., Tomimasu, T., Noda, S., Asano, T., and Sasaki A., 1996. Observation of ultrafast all-optical modulation based on intersubband transition in n-doped quantum wells by using free electron laser. *Appl. Phys. Lett.* **69**: 4136–4138.
12. Saiki, T., Mononobe, S., Ohtsu, M., Saito, N., and Kusano, J., 1996. Tailoring a high-transmission fiber probe for photon scanning tunneling microscope. *Appl. Phys. Lett.* **68**: 2612–2614.
13. Saiki, T. and Matsuda, K., 1999. Near-field optical fiber probe optimized for illumination-collection hybrid mode operation. *Appl. Phys. Lett.* **74**: 2773–1775.
14. Yariv, A., 1971. *Introduction to Optical Electronics*. Holt, Rinehart & Winston, New York.
15. Ohtsu, M., Kobayashi, K., Kawazoe, T., Sangu, S., and Yatsui, T., 2002. Nano-photonics: Design, fabrication, and operation of nanometric devices using optical near fields. *IEEE J. Select. Top. Quantum Electron.* **8**: 839–862.
16. Kawazoe, T., Kobayashi, K., Sangu, S., and Ohtsu, M., 2003. Demonstration of a nanophotonic switching operation by optical near-field energy transfer. *Appl. Phys. Lett.* **82**: 2957–2959.
17. Kawazoe, T., Kobayashi, K., Sangu, S., and Ohtsu, M., 2003. Demonstrating nanophotonic switching using near-field pump-probe photoluminescence spectroscopy of CuCl quantum cubes. *J. Microsc.* **209**: 261–266.
18. Madelung, O., Schulg, M., and Weiss H., (eds), 1982. *Landolt-Bornstein, Physics of II-VI and I-VII Compounds, Semimagnetic Semiconductors*. vol. 17b. Springer-Verlag, Berlin, Germany.
19. Masumoto, Y., Kawazoe, T., and Matsuura, N., 1998. Exciton-confined-phonon interaction in quantum dots. *J. Lumin.* **76&77**: 189–192.
20. Neogi, A., Lee, C. W., Everitt, H. O., Kuroda, T., Tackeuchi, A., and Yablonovitch, E., 2002. Enhancement of spontaneous recombination rate in a quantum well by resonant surface plasmon coupling. *Phys. Rev. B* **66**: 153305-1–153305-4.
21. Raether, H., 1988. *Surface Plasmons, Vol. III of Springer Tracts in Modern Physics*. Springer-Verag, Berlin, Germany.
22. Kawazoe, T., Kobayashi, K., Akahane, K., Naruse, M., Yamamoto, N., and Ohtsu, M., 2006. Demonstration of nanophotonic NOT gate using near-field optically coupled quantum dots. *Appl. Phys. B* **84**: 243–246.
23. Kawazoe, T., Kobayashi, K., and Ohtsu, M., 2005. Optical nanofountain: A biomimetic device that concentrates optical energy in a nanometric region. *Appl. Phys. Lett.* **86**: 103102-1–103102-3.
24. Nomura, W., Ohtsu, M., and Yatsui, T., 2005. Nanodot coupler with a surface plasmon polariton condenser for optical far/near-field conversion. *Appl. Phys. Lett.* **86**: 181108-1–181108-3.
25. Nomura, W., Yatsui, T., and Ohtsu, M., 2006. Efficient optical near-field energy transfer along an Au nanodot coupler with size-dependent resonance. *Appl. Phys. B* **84**: 257–259.
26. Born, M. and Wolf, E., 1983. *Principles of Optics*, 6th edn. Pergamon Press, Oxford, U.K.
27. Nishibayashi, K., Kawazoe, T., Akahane, K., Yamamoto, N., and Ohtsu, M., 2008. Observation of interdot energy transfer between InAs quantum dots. *Appl. Phys. Lett.* **93**: 042101-1–042101-3.
28. Yatsui, T., Sangu, S., Kawazoe, T., Ohtsu, M., An, S.-J., Yoo, J., and Yi, G.-C., 2007. Nanophotonic switch using ZnO nanorod double-quantum-well structures. *Appl. Phys. Lett.* **90**: 223110-1–223110-3.
29. Yoshimoto, M., Maeda, T., Ohnishi, T., Koinuma, H., Ishiyama, O., Shinohara, M., Kubo, M., Miura, R., and Miyamoto, A., 1995. Atomic-scale formation of ultra-smooth surfaces on sapphire substrates for high-quality thin film fabrication. *Appl. Phys. Lett.* **67**: 2615–2617.
30. Calvert, J. G. and Patts, J. N. Jr., 1996. *Photochemistry*. Wiley, New York.
31. Shimizu, M., Kamei, H., Tanizawa, M., Shiosaki, T., and Kawabata, A., 1988. Low temperature growth of ZnO film by photo-MOCVD. *J. Cryst. Growth* **89**: 365–370.

32. Kobayashi, K., Kawazoe, T., and Ohtsu, M., 2008. Localized photon model including phonons' degrees of freedom. In M. Ohtsu (ed.) *Progress in Nano-Electro-Optics VI*, Springer-Verlag, Berlin, Germany, pp. 41–66.
33. Kawazoe, T., Kobayashi, K., Takubo, S., and Ohtsu, M., 2005. Nonadiabatic photodissociation process using an optical near field. *J. Chem. Phys.* **122**: 024715-1–024715-5.
34. Kawazoe, T., Yamamoto, Y., and Ohtsu, M., 2001. Fabrication of a nanometric Zn dot by nonresonant near-field optical chemical vapor deposition. *Appl. Phys. Lett.* **79**: 1184–1186.
35. Kawazoe, T., Kobayashi, K., and Ohtsu, M., 2006. Near-field optical chemical vapor deposition using Zn(acac)$_2$ with a non-adiabatic photochemical process. *Appl. Phys. B* **84**: 247–251.
36. Kawazoe, T. and Ohtsu, M., 2004. Adiabatic and nonadiabatic nanofabrication by localized optical near fields. *Proc. SPIE* **5339**: 619–630.
37. Yonemitsu, H., Kawazoe, T., Kobayashi, K., and Ohtsu, M., 2007. Nonadiabatic photochemical reaction and application to photolithography. *J. Lumin.* **122–123**: 230–233.
38. Ohtsu, M., (ed). 2006. *Progress in Nano-Electro-Optics V*, Spinger-Verlag, Preface to Volume V: Based on "nanophotonics" proposed by Ohtsu in 1993, OITDA (Optical Industry Technology Development Association, Japan) organized the nanophotonics technical group in 1994, and discussions on the future direction of nanophotonics were started in collaboration with academia and industry.
39. Ho, K. M., Chan, C. T., and Soukoulis, C. M., 1990. Existence of a photonic gap in periodic dielectric structures. *Phys. Rev. Lett.* **65**: 3152–3155.
40. Podolskiy, V. A., Sarychev, A. K., and Shalaev, V. M., 2003. Plasmon modes and negative refraction in metal nanowire composites. *Opt. Express* **11**: 735–745.
41. Shelby, R. A., Smith, D. R., and Shultz, S., 2001. Experimental verification of a negative index of refraction. *Science* **292**: 77–79.
42. Pendry, J. B., 2000. Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**: 3966–3969.
43. Rong, H., Liu, A., Nicolaescu, R., and Paniccia, M., 2004. Raman gain and nonlinear optical absorption measurements in a low-loss silicon waveguide. *Appl. Phys. Lett.* **85**: 2196–2198.
44. Arakawa, Y. and Sakaki, H., 1982. Multidimensional quantum well laser and temperature-dependence of its threshold current. *Appl. Phys. Lett.* **40**: 939–941.
45. Ohtsu, M. and Kobayashi, K., 2003. *Optical Near Fields*, vol. 109–150. Springer-Verlag, Berlin, Germany.
46. Ohtsu, M., 2007. Nanooptics. In F. Traeger (ed.) *Handbook of Lasers and Optics*, Springer-Verlag, Berlin, Germany, pp. 1079–1090.
47. Shannon, C., 1956. The bandwagon. *IEEE Trans. Inform. Theory* **2**: 3.

36

Nanophotonic Device Materials

Takashi Yatsui
University of Tokyo

Wataru Nomura
University of Tokyo

36.1 Introduction **36**-1
36.2 Nanophotonic Devices Based on Quantum Dots **36**-1
Discrete Energy Levels in Spherical Quantum Dots • The Observation of Dissipated Optical Energy Transfer between CdSe QDs • Controlling the Energy Transfer between Near-Field Optically Coupled ZnO QDs
36.3 Nanophotonic AND-Gate Device Using ZnO Nanorod Double-Quantum-Well Structures **36**-7
References **36**-11

36.1 Introduction

Systems of optically coupled quantum structures should be applicable to quantum information processing (Bayer et al. 2001, Stinaff et al. 2006). Additional functional devices, i.e., nanophotonic devices (Ohtsu et al. 2002, Kawazoe et al. 2003, 2005, 2006), can be realized by controlling the exciton excitation in quantum dots (QDs) and quantum-well structures (QWs). This chapter reviews the recent achievements with nanophotonic devices based on colloidal QDs (Section 36.2) and nanorod QWs (Section 36.3).

36.2 Nanophotonic Devices Based on Quantum Dots

36.2.1 Discrete Energy Levels in Spherical Quantum Dots

The translational motion of the exciton center of mass is quantized in nanoscale semiconductors when the size is decreased so that it is as small as an exciton Bohr radius. If the QDs are assumed to be spheres with radius R, with the following potential

$$V(x) = \begin{cases} 0 & \text{for } |x| \le R \\ \infty & \text{for } |x| > R \end{cases} \tag{36.1}$$

the quantized energy levels are given by a spherical Bessel function as

$$R_{nl}(r) = A_{nl}\, j_l\left(\rho_{n,l}\frac{r}{R}\right) \tag{36.2}$$

Figure 36.1 shows the lth order of the spherical Bessel function. Note that an odd quantum number of l has an odd function and it is a dipole-forbidden energy state. To satisfy the boundary conditions as

$$R_{nl}(R) = A_{nl}\, j_l(\rho_{n,l}) = 0 \tag{36.3}$$

the quantized energy levels are calculated using

$$E(n,l) = E_{\mathrm{B}} + \frac{\hbar^2\pi^2}{2mR^2}\xi_{n,l}^2 \tag{36.4}$$

where $\pi\xi_{n,l} = \rho_{n,l}$ is the nth root of the spherical Bessel function of the lth order. The principal quantum number n and the angular momentum quantum number l take values $n = 1, 2, 3, \ldots$, and $l = 0, 1, 2, \ldots$, respectively. $\xi_{n,l}$ takes values $\xi_{1,0} = 1$, $\xi_{1,1} = 1.43$, $\xi_{1,2} = 1.83$, $\xi_{2,0} = 2$, and so on (see Table 36.1) (Sakakura and Masumoto 1997).

Figure 36.2 show schematic drawings of different-sized spherical QDs (X and Y) and confined exciton energy levels. Here, R and $1.43R$ are the radii of spherical QDs X and Y, respectively. According to Equation 36.4, the quantized exciton energy levels of $E(1,0)$ in QD X and $E(1,1)$ in QD Y resonate with each other. Although the energy state $E(1,1)$ is a dipole-forbidden state, the optical near-field interaction is finite for such coupling to the forbidden energy state (Kobayashi et al. 2000). In addition, the intersublevel transition, τ_{sub}, from higher exciton energy levels to the lowest one, is generally less than a few picoseconds and is much shorter than the transition time due to optical near-field coupling (τ_{ET}) (Suzuki et al. 1996). Therefore, most of the energy of the exciton in a QD X with radius R transfers to the lowest exciton energy level in the neighboring QD Y with a radius of $1.43R$ and recombines radiatively at the lowest level. In this manner, unidirectional energy flow is achieved.

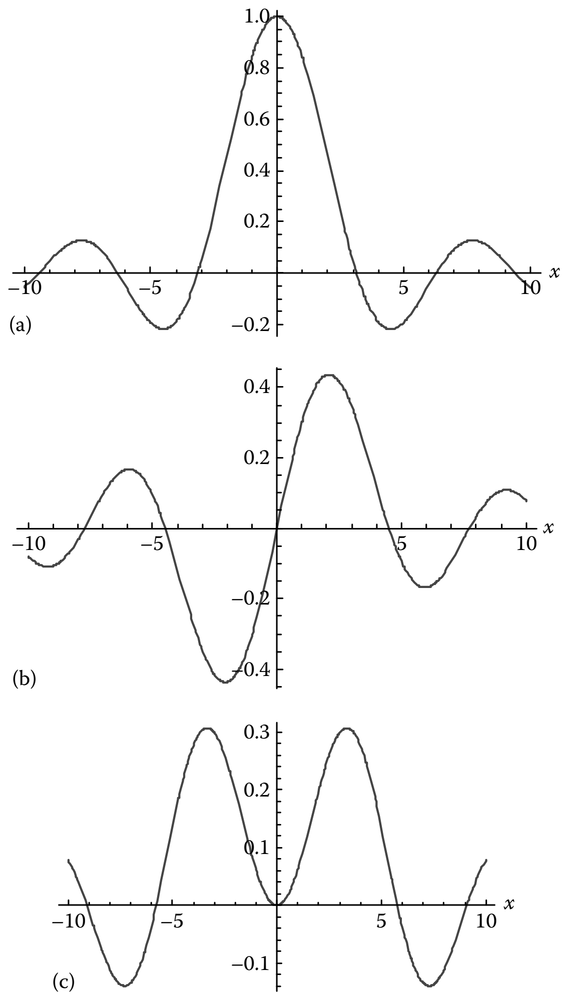

FIGURE 36.1 lth order of spherical Bessel function. (a) $l = 0$ ($j_0(x) = (\sin x)/x$), (b) $l = 1$ ($j_1(x) = (1/x)(((\sin x)/x) - \cos x)$), and (c) $l = 2$ ($j_2(x) = (1/x)\,(((3 - x^2)/x^2) \sin x) - ((3/x) \cos x)$).

TABLE 36.1 Calculated $\xi_{n,l}$ to Satisfy the Condition of $j_l(\pi\xi_{n,l}) = 0$

	$l = 0$	$l = 1$	$l = 2$	$l = 3$	$l = 4$	•
$n = 1$	1	1.43	1.83	2.22	2.61	•
$n = 2$	2	2.46	2.90	3.31	•	•
$n = 3$	3	•	•	•	•	•
•	•	•	•	•	•	•

36.2.2 The Observation of Dissipated Optical Energy Transfer between CdSe QDs

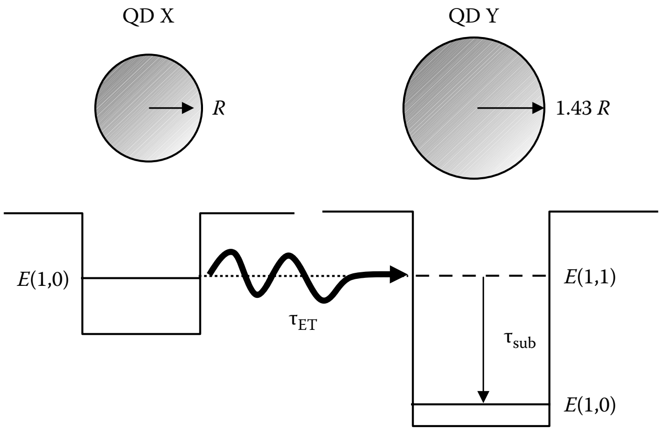

FIGURE 36.2 Schematic drawings of different-sized spherical QDs (X and Y) and the confined exciton energy levels.

To evaluate the energy transfer and the energy dissipation, we used CdSe/ZnS core–shell QDs from *Evident Technologies*. As described in Section 36.2.1, assuming that the respective diameters, D, of the QD_{a1} and QD_{a2} were 2.8 and 4.1 nm, the ground energy level in the QD_{a1} and the excited energy level in the QD_{a2} resonate (Trallero-Giner et al. 1998). A solution of QDs_{a1} ($D = 2.8$ nm) and QDs_{a2} ($D = 4.1$ nm) in 1-feniloctane at a density of 1.0 mg/mL was dropped on mica substrate (see Figure 36.3a), such that areas A and C consisted of QDs_{a1} and QDs_{a2}, respectively, while there were both QDs_{a1} and QDs_{a2} in area B. Using transmission electron microscopy (TEM), we confirmed the mean center-to-center distance of each QD was maintained at about less than 10 nm in all areas, due to the 2 nm-thick ZnS shell and surrounding ligands (2 nm-length long chain amine) of the QDs (Figure 36.3b).

In the following experiments, the light source used was the third harmonic of a mode-locked Ti:sapphire laser (wavelength 306 nm, frequency 80 MHz, and pulse duration 2 ps). The incident power of the laser was 0.6 mW and the spot size was 1×10^{-3} cm^2. The density of QDs was less than 3.5×10^{12} cm^{-2} and the quantum yield of CdSe/ZnS QD was 0.5. Under these conditions, the probability of exciton generation by one laser pulse in each QD was calculated to be 1.6×10^{-2}. Therefore, we assumed single-exciton dynamics in the following experiments.

The energy transfer was confirmed using micro-photoluminescence (PL) spectroscopy. Temperature-dependent micro-PL spectra were obtained. In the spectral profile of the PL emitted from area A, we found a single peak which originated from the ground state of QD_{a1} at a wavelength of $\lambda = 540$ nm, from room temperature to 30 K (broken line in Figure 36.4). From area C, the single peak, which originated from the ground state of QD_{a2}, was found at $\lambda = 600$ nm (dotted line in Figure 36.4). By contrast, area B had two peaks at room temperature, as shown by solid line in Figure 36.4. This figure also shows that the PL intensity peak at $\lambda = 540$ nm decreased relative to that of at $\lambda = 600$ nm on decreasing the temperature. This relative decrease in the PL intensity originated from the energy transfer from the ground state in the QD_{a1} to the excited state in the QD_{a2} and the subsequent rapid dissipation to the ground state in the QD_{a2}. This is because the coupling between the resonant energy levels becomes stronger due to the increase in the exciton decay time on decreasing the temperature (Itoh et al. 1990). Furthermore, although nanophotonic device operation using CuCl quantum cubes (Kawazoe et al. 2003, 2005, 2006) and ZnO quantum wells (Yatsui et al. 2007) has been reported at 15 K, we observed the

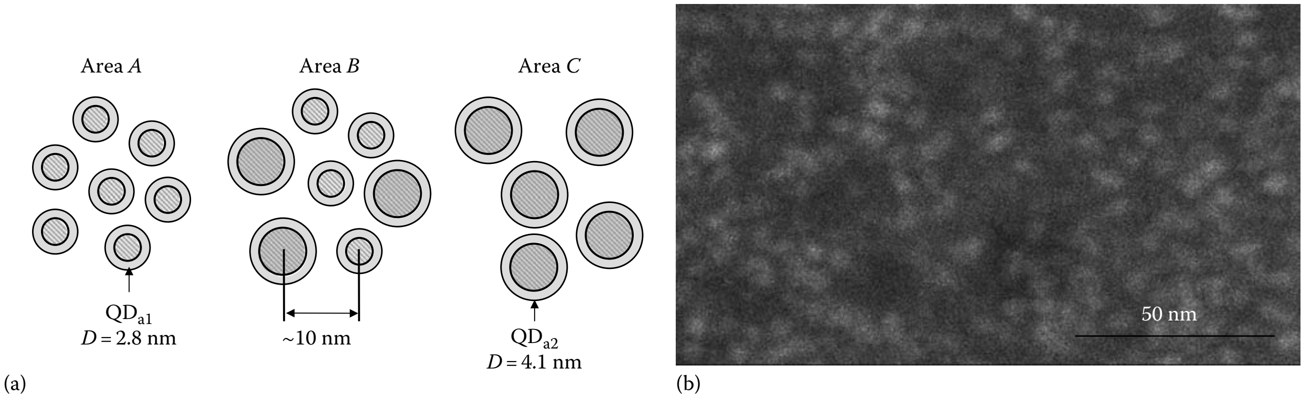

FIGURE 36.3 (a) Schematic images of CdSe QDs dispersed substrate. Areas *A*, *B*, and *C* are covered by QDs_{a1}, both QDs_{a1} and QDs_{a2}, and QDs_{a2}, respectively. (b) TEM image of dispersed CdSe/ZnS core–shell QDs in area *B*.

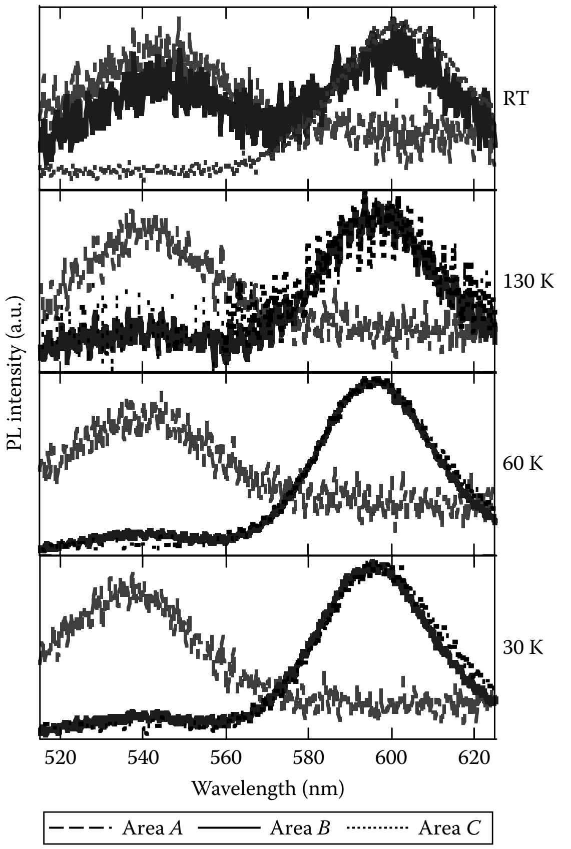

FIGURE 36.4 Temperature dependence of the micro-PL spectra. Broken, solid, and dotted lines show the spectra from areas *A*, *B*, and *C*, respectively.

decrease in the PL intensity at λ = 540 nm at temperatures as high as 130 K, which is advantageous for the high-temperature operation of nanophotonic devices.

To confirm this energy transfer from QD_{a1} to QD_{a2} at the temperature under 130 K, we evaluated the dynamic property of the energy transfer using time-resolved spectroscopy with the time correlation single photon counting method. Circles A_{a1}, squares B_{a1}, and triangles B_{a2} in Figure 36.5a represent the respective time-resolved micro-PL intensities (60 K) from ground energy level in QD_{a1} ($D = 2.8$ nm) in area *A*, QD_{a1} ($D = 2.8$ nm) in area *B*, and QD_{a2} ($D = 4.1$ nm) in area *B*. The peak intensities at $t = 0$ were normalized to unity. Note that B_{a2} decreases faster than A_{a1}, although these signals were generated from QDs of the same size. In addition, although the exciton lifetime decreases on increasing the QD size, owing to the increased oscillator strength, B_{a2} decreased more slowly than A_{a1} over the range $t < 0.2$ ns (see the inset of Figure 36.5a). Furthermore, as we did not see any peak in the power spectra of A_{a1}, B_{a1}, and B_{a2}, we believe that the temporal signal changes originated from the optical near-field energy transfer and subsequent dissipation. Since the QDs_{a1} in area *B* were near QDs_{a2} whose excited energy level resonates with the ground energy level of the QD_{a1} (see Figure 36.5b), near-field coupling between the resonant levels resulted in the energy transfer from the QD_{a1} to the QD_{a2} and the consequent faster decrease in the excitons of the QD_{a1} in area *B* compared with area *A*. Furthermore, as a result of inflow of the carriers from the QD_{a1} to the QD_{a2}, the PL intensity from the QD_{a2} near the QD_{a1} decayed more slowly than that of the QD_{a1}.

For comparison, we also obtained time-resolved PL profiles of different pairs of CdSe/ZnS QDs. Their diameters were $D = 2.8$ nm (QD_{a1}) and 3.2 nm (QD_{b1}), which means that their energy levels did not resonant with each other. Figure 36.6a shows a schematic image of a sample named area *D*, where QDs_{a1} and QDs_{b1} are mixed with a mean center-to-center distance of less than 10 nm. Circles A_{a1} and diamonds D_{a1} in Figure 36.6b show the time-resolved PL intensity (30 K) from the ground energy level in QDs_{a1} from areas *A* and *D*, respectively. There is no difference in the decay profile. This indicates that the excited carriers in QDs_{a1} did not couple with QDs_{b1} due to their off-resonance and, consequently, no energy was transferred (Figure 36.6c). This supports the postulate that Figures 36.4 and 36.5 demonstrate energy transfer and subsequent dissipation due to near-field coupling between the resonant energy levels of the QDs_{a1} and QDs_{a2}.

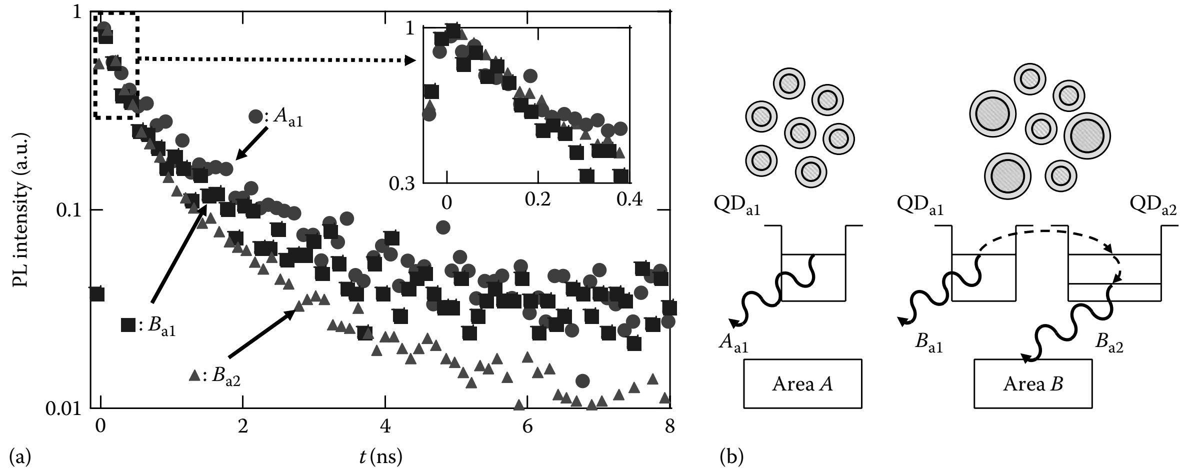

FIGURE 36.5 (a) Time-resolved PL intensity profiles from QDs$_{a1}$ in area A (circles A_{a1}), QDs$_{a1}$ in area B (squares B_{a1}), and QDs$_{a2}$ in area B (triangles B_{a2}). The peak intensities were normalized at $t = 0$. (b) Schematic of the respective system configurations in areas A and B.

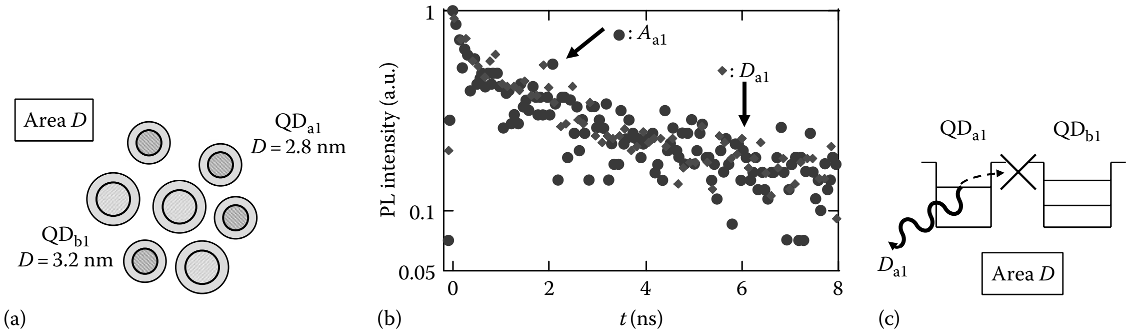

FIGURE 36.6 (a) Schematic image of CdSe QDs dispersed substrate area D. (b) Time-resolved PL intensity profiles from QDs$_{a1}$ in area A (circles A_{a1}) and QDs$_{a1}$ in area D (diamonds D_{a1}). The peak intensities were normalized at $t = 0$. (c) Schematic of the system configuration in area D.

To discuss the exciton energy transfer from QD$_{a1}$ to QD$_{a2}$ quantitatively, we investigated the exciton dynamics by fitting multiple exponential decay curve functions to curves A_{a1}, B_{a1}, and B_{a2} (Bawendi et al. 1992, Crooker et al. 2003):

$$A_{a1} = R_{S1}\exp\left(\frac{-t}{\tau_{s1}}\right) + R_{S2}\exp\left(\frac{-t}{\tau_{s2}}\right), \tag{36.5}$$

$$B_{a1} = R_S \cdot A_{a1} + R_{tS}\exp\left(\frac{-t}{\tau_t}\right), \tag{36.6}$$

and

$$B_{a2} = R_{L1}\exp\left(\frac{-t}{\tau_{L1}}\right) + R_{L2}\exp\left(\frac{-t}{\tau_{L2}}\right). \tag{36.7}$$

We used double exponential decay for A_{a1} and B_{a2} (Equations 36.5 and 36.7), which corresponds to the non-radiative lifetime (fast decay: τ_{S1} and τ_{L1}) and radiative lifetime of free-carrier recombinations (slow decay: τ_{S2} and τ_{L2}). Given the imperfect homogeneous distribution of the QDs$_{a1}$ in area B, some QDs$_{a1}$ lack energy transfer routes to QD$_{a2}$. However, we introduced the mean energy transfer time τ_t from QDs$_{a1}$ and QDs$_{a2}$ in Equation 36.6. In these equations, we neglected the energy dissipation time of about 1 ps (Guyot-Sionnest et al. 1999) because that is much smaller than exciton lifetimes and energy transfer time. Figure 36.7 shows the best-fitted numerical results and experimental data. Here, we used exciton lifetimes of $\tau_{S2} = 2.10$ ns and $\tau_{L2} = 1.79$ ns. The mean energy transfer time was $\tau_t = 135$ ps, which is comparable with the observed energy transfer time (130 ps) in CuCl quantum cubes (Kawazoe et al. 2003) and ZnO QWs (Yatsui et al. 2007). Furthermore, the relation $\tau_t < \tau_{S2}$ agrees with the assumption that most of the excited excitons in QDs$_{a1}$ transfer to excited exciton energy level in a QD$_{a2}$ before being emitted from QD$_{a1}$.

36.2.3 Controlling the Energy Transfer between Near-Field Optically Coupled ZnO QDs

ZnO is a promising material for room-temperature operation of a nanophotonic device because of its large exciton binding energy (Ohtomo et al. 2000, Huang et al. 2001, Sun et al. 2002). Here, we

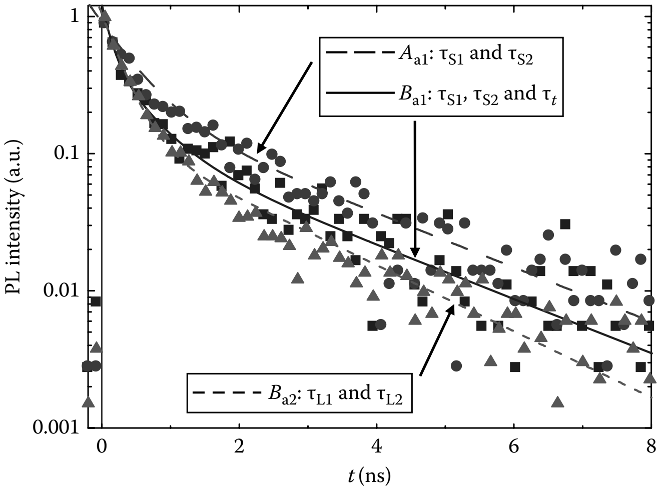

FIGURE 36.7 Experimental results (circles, squares and triangles) and fitting curves (broken, solid, and short broken lines) using Equations 36.5 through 36.7 for the PL intensity profiles. The fitting parameters are $R_{S1} = 0.560$, $\tau_{S1} = 2.95 \times 10^{-10}$, $R_{S2} = 0.329$, $\tau_{S2} = 2.10 \times 10^{-9}$, $R_S = 0.740$, $R_{tS} = 0.330$, $\tau t = 1.35 \times 10^{-10}$, $R_{L1} = 0.785$, $\tau_{L1} = 2.94 \times 10^{-10}$, $R_{L2} = 0.201$, $\tau_{L2} = 1.79 \times 10^{-9}$.

used chemically synthesized ZnO QD to realize a highly integrated nanophotonic device. We observed the energy transfer from smaller ZnO QDs to larger QDs with mutually resonant energy levels. The energy transfer time and energy transfer ratio between the two QDs were also calculated from the experimental results (Yatsui et al. 2008).

ZnO QDs were prepared using the sol-gel method (Spanhel and Anderson 1991, Meulenkamp 1998).

1. A sample of 1.10 g (5 mmol) of $Zn(Ac)_2 \cdot 2H_2O$ was dissolved in 50 mL of boiling ethanol at atmospheric pressure, and the solution was then immediately cooled to 0°C. A sample of 0.29 g (7 mmol) of $LiOH \cdot H_2O$ was dissolved in 50 mL of ethanol at room temperature in an ultrasonic bath and cooled to 0°C. The hydroxide-containing solution was then added dropwise to the $Zn(Ac)_2$ suspension with vigorous stirring at 0°C. The reaction mixture became transparent after approximately 0.1 g of LiOH had been added. The ZnO sol was stored at 0°C to prevent particle growth.
2. A mixed solution of hexane and heptane, with a volume ratio of 3:2, was used to remove the reaction products (LiAc and H_2O) from the ZnO sol.
3. To initiate the particle growth, the ZnO solution was warmed to room temperature. The mean diameter of ZnO QD was determined from the growth time, T_g.

Figure 36.8a shows a TEM image of synthesized ZnO dots after the second step. Dark areas correspond to the ZnO QD. This image suggested that monodispersed single crystalline particles were obtained.

To check the optical properties and diameters of our ZnO QD, we measured the photoluminescence (PL) spectra using He–Cd laser (λ = 325 nm) excitation at 5 K. We compared the PL spectra of ZnO QD with T_g = 0 and 42 h (solid and dashed curves in Figure 36.8b, respectively). A redshifted PL spectrum was obtained, indicating an increase in the QD diameter. Figure 36.8c shows the growth time dependence of the QD diameter. This was determined from the effective mass model, with peak energy in the PL spectra, $E_g^{bulk} = 3.35$ eV, $m_e = 0.28$, $m_h = 1.8$, and $\varepsilon = 3.7$ (Brus 1984). This result indicated that the diameter growth rate at room temperature was 1.1 nm/day.

Assuming that the diameters, D, of the QDS and the QDL were 3.0 and 4.5 nm, respectively, E_{S1} in the QDS and E_{L2} in the QDL resonated (Figure 36.9a). An ethanol solution of QDS and QDL was dropped onto a sapphire substrate. The mean surface-to-surface separation of the QD was approximately 3 nm.

The spectra S and L in Figure 36.9b correspond to the QDS and the QDL, with spectral peaks of 3.60 and 3.44 eV, respectively. The curve A in Figure 36.9b shows the spectrum from the QDS and QDL mixture with $R = 1$, where R is the ratio (number of QDS)/(number of QDL). The spectral peak of 3.60 eV, which corresponded to the PL from the QDS, was absent from this curve. This peak was thought to have disappeared due to energy transfer from the QDS to the QDL, because the first excited state of the QDL resonated with the ground state of the QDS. Our hypothesis was supported by the observation that when R was increase by eightfold, the spectral peak from the QDS reappeared (see spectrum C in Figure 36.9b).

To confirm this energy transfer from the QDS to the QDL at 5 K, we evaluated dynamic effects using time-resolved spectroscopy with the time-correlated single photon counting method. The light source used was the third harmonic of a mode-locked Ti:sapphire laser (photon energy 4.05 eV, frequency 80 MHz, and pulse duration 2 ps). We compared the signals from mixed samples with ratios R = 2, 1, and 0.5. The curves T_A (R = 2), T_B (R = 1), and T_C (R = 0.5) in Figure 36.10 show the respective time-resolved PL intensities from the ground state of the QDS (E_{S1}) at 3.60 eV. We investigated the exciton dynamics quantitatively by fitting multiple exponential decay curve functions (Bawendi et al. 1992, Crooker et al. 2003):

$$\mathrm{TRPL} = A_1 \exp\left(\frac{-t}{\tau_1}\right) + A_2 \exp\left(\frac{-t}{\tau_2}\right) \qquad (36.8)$$

We obtained average τ_1 and τ_2 values of 144 ps and 443 ps, respectively (see Table 36.2). Given the disappearance of the spectral peak at 3.60 eV in the PL spectra, it is likely that these values corresponded to the energy transfer time from the QDS to the QDL and the radiative decay time from the QDS, respectively. This hypothesis was supported by the observation that the average value of τ_1 (144 ps) was comparable with the observed energy transfer time in CuCl quantum cubes (130 ps) (Kawazoe et al. 2003).

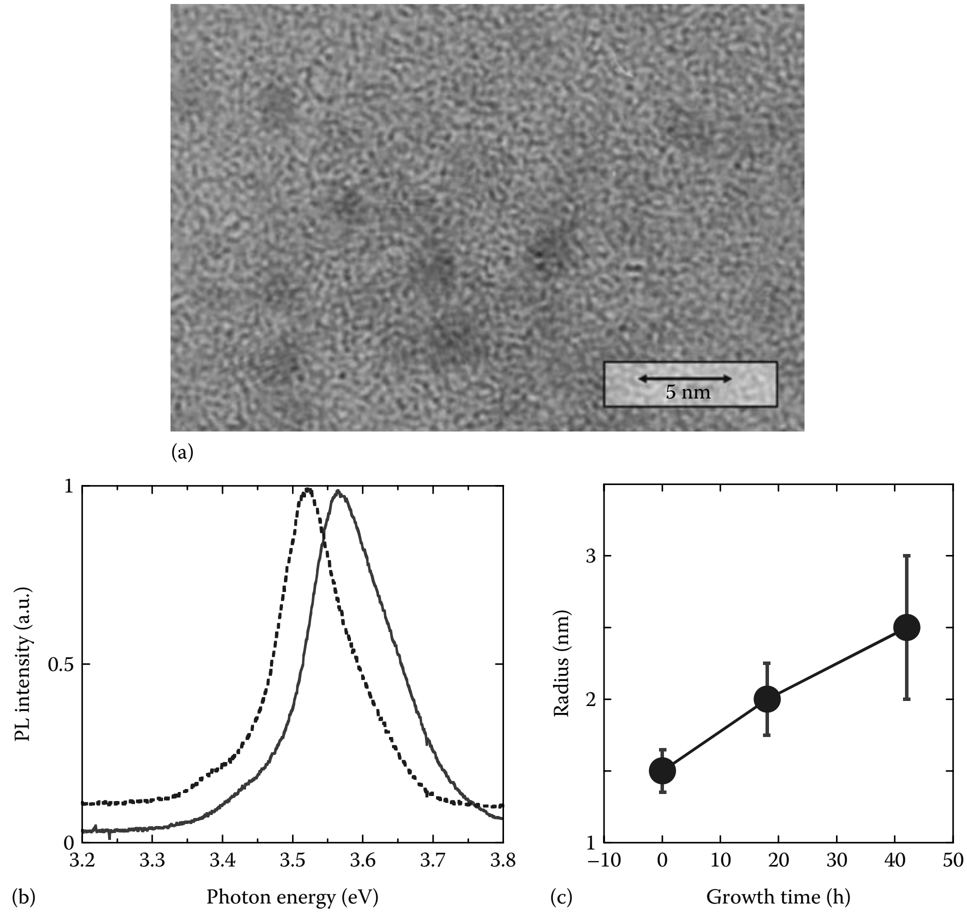

FIGURE 36.8 (a) TEM image of the ZnO QD. The dark areas inside the white dashed circles correspond to the ZnO QD. (b) The PL spectra observed at 5 K. The solid and dashed curves indicate growth time $T_g = 0$ and 42 h, respectively. (c) The growth time dependence of the mean ZnO QD diameter.

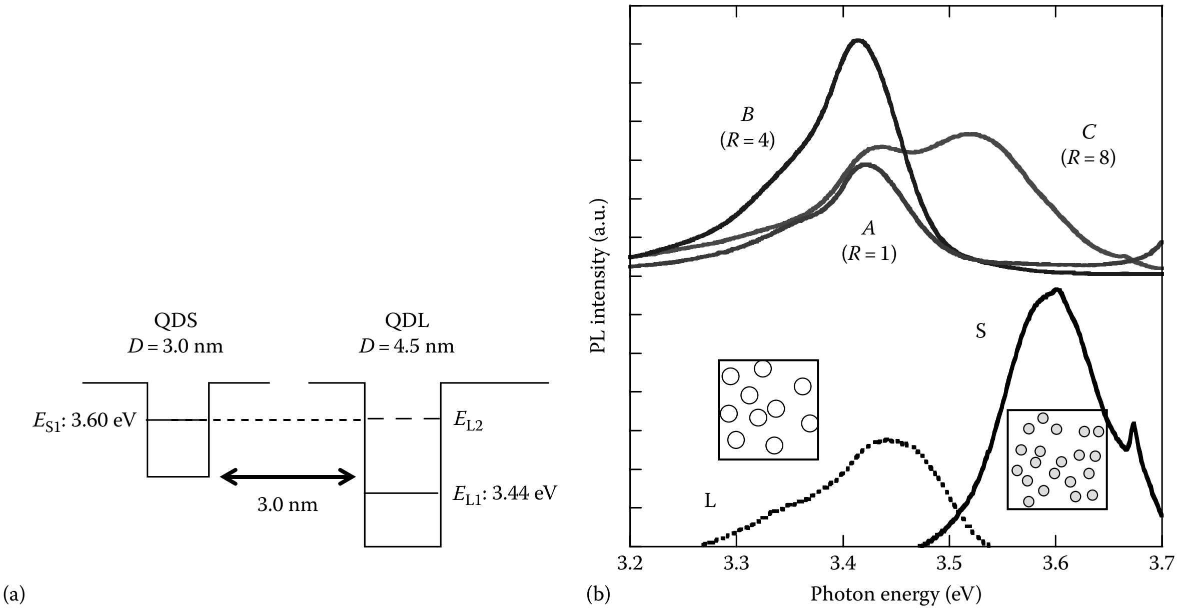

FIGURE 36.9 (a) Schematic of the energy diagram between a QDS and QDL. (b) The PL spectra observed at 5 K. The labels S and L indicate QDS and QDL, respectively. The labels A, B, and C indicate mixes with R ratios of 1, 4, and 8, respectively.

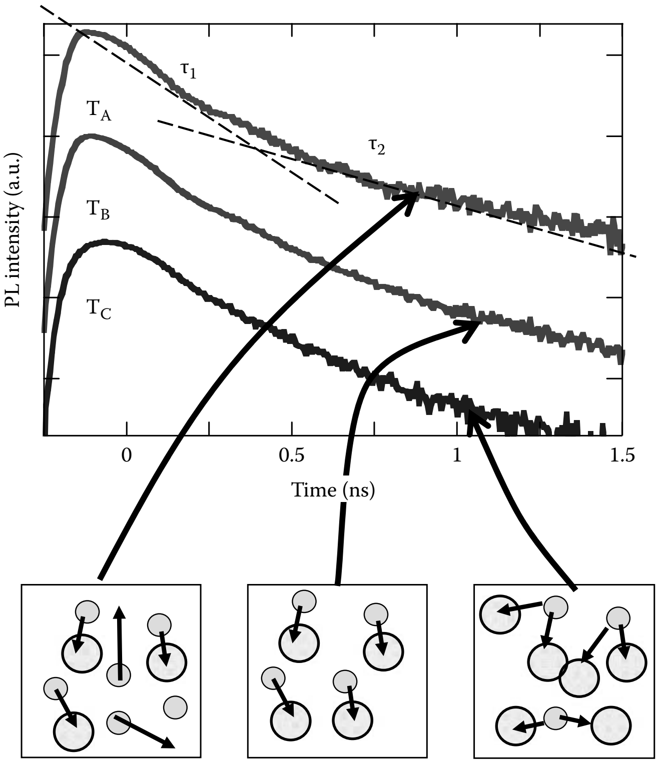

FIGURE 36.10 Time-resolved PL spectra observed at 5 K. The values of R were 2, 1, and 0.5 for curves T_A, T_B, and T_C, respectively.

TABLE 36.2 Dependence of the Time Constants (τ_1 and τ_2) on R as Derived from the Two Exponential Fits of the Time-Resolved PL Signals and the Coefficient Ratio A_1/A_2

R = QDS/QDL	τ_1 (ps)	τ_2 (ps)	A_1/A_2
2	133	490	12.4
1	140	430	13.7
0.5	160	410	14.4
Average	144	443	

We also investigated the value of coefficient ratio A_1/A_2 (see Table 36.2); this ratio was inversely proportional to R, hence proportional to the number of QDL. This result indicated that an excess QDL caused energy transfer from QDS to QDL, instead of direct emission from the QDS.

We observed the dynamic properties of exciton energy transfer and dissipation between ZnO QD *via* an optical near-field interaction, using time-resolved PL spectroscopy. Furthermore, we successfully increased the energy transfer ratio between the resonant energy state, instead of the radiative decay from the QD. Chemically synthesized nanocrystals, both semiconductor QD and metallic nanocrystals (Brust and Kiely 2002), are promising nanophotonic device candidates, because they have uniform sizes, controlled shapes, defined chemical compositions, and tunable surface chemical functionalities.

36.3 Nanophotonic AND-Gate Device Using ZnO Nanorod Double-Quantum-Well Structures

ZnO/ZnMgO nanorod heterostructures have been fabricated and the quantum confinement effect has been observed from single QW structures (Park et al. 2003, 2004). In this section, we review the time-resolved near-field spectroscopy to demonstrate the switching dynamics that result from controlling the optical near-field energy transfer in ZnO nanorod double-quantum-well structures (DQWs). We observed nutation of the population between the resonantly coupled exciton states of DQWs, where the coupling strength of the near-field interaction decreased exponentially as the separation increased (Yatsui et al. 2007).

To evaluate the energy transfer, three samples were prepared (Figure 36.11a): (1) Single-quantum-well structures (SQWs) with a well-layer thickness of L_w = 2.0 nm (SQWs), (2) DQWs with L_w = 3.5 nm with 6 nm separation (1-DQWs), and (3) three pairs of DQWs with L_w = 2.0 nm with different separations (3, 6, and 10 nm), where each DQW was separated by 30 nm (3-DQWs). ZnO/ZnMgO quantum-well structures (QWs) were fabricated on the ends of ZnO nanorods with a mean diameter of 80 nm using catalyst-free metalorganic vapor phase epitaxy (Park et al. 2002). The average concentration of Mg in the ZnMgO layers used in this study was determined to be 20 atm %.

The far-field PL spectra were obtained using a He–Cd laser (λ = 325 nm) before detection using near-field spectroscopy. The emission signal was collected with an achromatic lens (f = 50 mm). The near-field photoluminescence (NFPL) spectra were obtained using a He–Cd laser (λ = 325 nm), collected with a fiber probe with an aperture diameter of 30 nm, and detected using a cooled charge-coupled device through a monochromator. Blueshifted PL peaks were observed at 3.499 (I_S), 3.429 (I_{1D}), and 3.467 (I_{3D}) eV in the far- and near-field PL spectra (Figure 36.12a). These peaks originated from the respective ZnO QWs because their energies are comparable to the predicted ZnO well-layer thicknesses of 1.7 (I_S), 3.4 (I_{1D}), and 2.2 (I_{3D}) nm, respectively, calculated using the finite square-well potential of the quantum confinement effect in ZnO SQWs (see Figure 36.12b) (Park et al. 2004).

To confirm the near-field energy transfer between QWs, we compared the time-resolved near-field PL (TR_{NFPL}) signals at the I_S, I_{1D}, and I_{3D} peaks. For the time-resolved near-field spectroscopy, the signal was collected using a micro-channel plate through a band-pass filter with 1 nm spectral width. Figure 36.13 shows the typical TR_{NFPL} of SQWs (TR_S), 1-DQWs (TR_{1D}), and 3-DQWs (TR_{3D}).

We calculate the exciton dynamics using quantum mechanical density-matrix formalism (Coffey and Friedberg 1978, Kobayashi et al. 2005), where the Lindblad-type dissipation is assumed for the relaxation due to exciton–photon and exciton–phonon couplings:

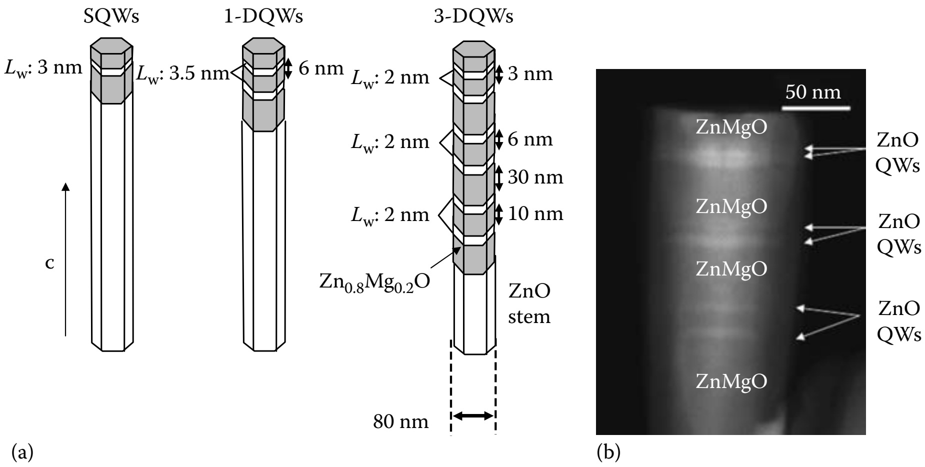

FIGURE 36.11 ZnO/ZnMgO nanorod quantum-well-structures. c: *c*-axis of the ZnO stem. (a) Schematic of ZnO/ZnMgO SQWs, DQWs (1-DQWs), and triple pairs of DQWs (3-DQWs). (b) Z-contrast TEM image of 3-DQWs clearly shows the compositional variation, with the bright layers representing the ZnO well layers. Scale bar: 50 nm.

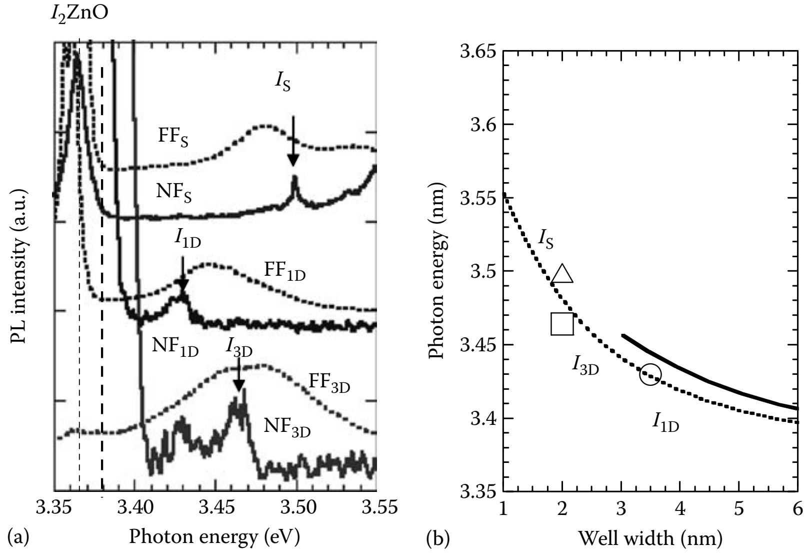

FIGURE 36.12 Near-field time-resolved spectroscopy of ZnO nanorod DQWs at 15 K. (a) NF_S, NF_{1D}, NF_{3D}: near-field PL spectra. FF_S, FF_{1D}, FF_{3D}: far-field PL spectrum of ZnO SQWs (L_w = 2.0 nm), 1-DQWs (L_w = 3.5 nm, 6 nm separation), and 3-DQWs (L_w = 2.0 nm; 3, 6, and 10 nm separation). (b) Well width dependent on the exciton ground state and the first excited state. SQWs: open triangle, 1-DQWs: open circle, 3-DQWs: open square.

$$\dot{\rho} = -\frac{i}{\hbar}[H,\rho] + \sum_{n} \frac{\gamma_n}{2}\left(2A_n\rho A_n^{\dagger} - A_n^{\dagger}A_n\rho - \rho A_n^{\dagger}A_n\right) \quad (36.9)$$

where

ρ is the density operator
H is the Hamiltonian in the considered system
$A_n^{\dagger}$ and A_n are creation and annihilation operators for an exciton energy level labeled n
γ_n is the photon or phonon relaxation constant for the energy level

The exciton population is calculated using matrix elements for all exciton states in the system considered. First, we apply the calculation to a three-level system of SQWs (Figure 36.14a), where the continuum state $\hbar\Omega_C$ is initially excited using a 10 ps laser pulse. Then, the initial exciton population in ZnO QWs is created in $\hbar\Omega_{1S}$, where the energy transfer from $\hbar\Omega_C$ to $\hbar\Omega_{1S}$ is expressed phenomenologically as a Gaussian input signal with a temporal width of $2\sigma_{1S}$ (an incoherent excitation term is added in Equation 36.9), because non-radiative relaxation paths via exciton–phonon coupling make a dephased input signal, statistically. Finally, an exciton carrier relaxes due to the electron–hole recombination with relaxation constant γ_{1S}. Figure 36.14b shows a numerical result and experimental data. Here, we used $2\sigma_{1S}$ = 100 ps, and γ_{1S} was evaluated as 460 ps.

A similar calculation was applied for DQWs. We used two three-level systems, coupled via an optical near-field with a

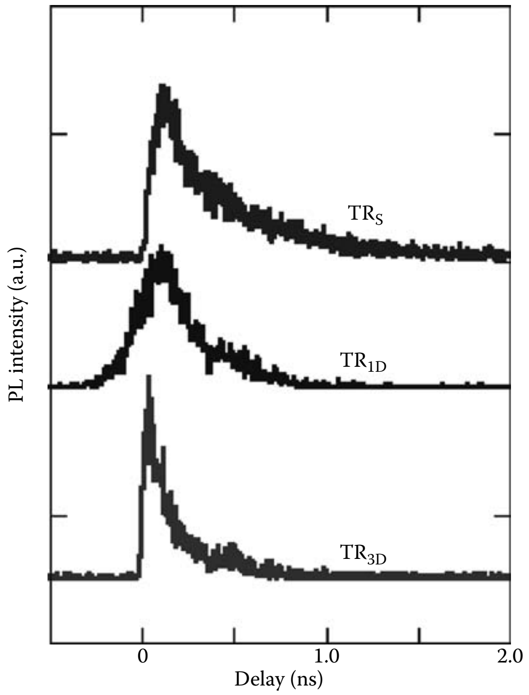

FIGURE 36.13 TR$_S$, TR$_{1D}$, and TR$_{3D}$ show TR$_{NFPL}$ signal obtained at I$_S$, I$_{1D}$, and I$_{3D}$, respectively, using the 4.025 eV (λ = 308 nm) light with a pulse of 10 ps duration to excite the barrier layers of ZnO QWs.

coupling strength of U_{12} (Figure 36.14c). Figure 36.14d shows the numerical results for the exciton population in QWs$_1$ and the experimental data. Here $2\sigma_{1D}$ and $2\sigma_{2D}$ were set at 200 ps, which is twice the value for SQWs, because the relaxation paths extend the barrier energy state in the two quantum wells (QWs). γ_{1D} and γ_{2D} are evaluated as 200 ps. We believe that the faster relaxation for DQWs compared with SQWs reflects the lifetime of the coupled states mediated by the optical near-field. Furthermore, the characteristic behavior that results from near-field coupling appears as the oscillatory decay in Figure 36.14d. This indicates that the timescale of the near-field coupling is shorter than the decoherence time, and that coherent coupled states, such as symmetric and antisymmetric states (Sangu et al. 2004), determine the system dynamics. Furthermore, nutation never appears unless unbalanced initial exciton populations are prepared for $\hbar\Omega_{1D}$ and $\hbar\Omega_{2D}$. In the far-field excitation, only the symmetric state is excited because the antisymmetric state is dipole-inactive. Then, the exciton populations of the two quantum wells are equal and they have the same decay rate. By contrast, in the near-field excitation, both the symmetric and antisymmetric states are excited due to the presence of a near-field probe. Since the symmetric and antisymmetric states have different eigenenergies, the interference of these states generates a detectable beat signal. The unbalanced excitation rate is given by A_1/A_2 = 10 here. From

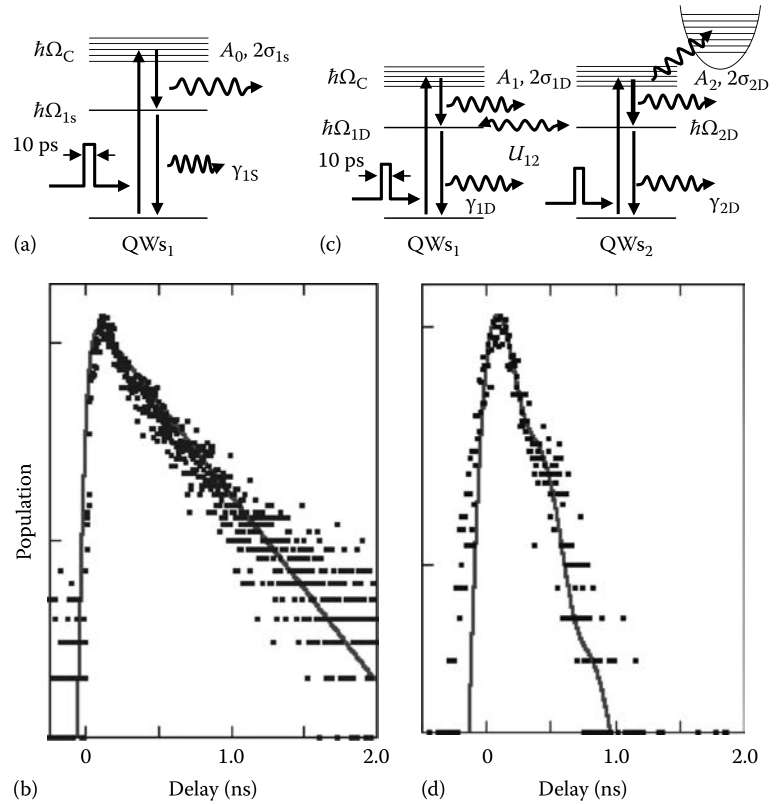

FIGURE 36.14 The schematic depict the (a) SQWs and (c) 1-DQWs system configurations. $\hbar\Omega_C$: barrier energy state with a central energy. Theoretical results on the transient exciton population dynamics (solid curves) and experimental PL data (filled squares) of (b) SQWs (same as curve TR$_S$ in Figure 36.13) and (d) 1-DQWs (same as curve TR$_{1D}$ in Figure 36.13).

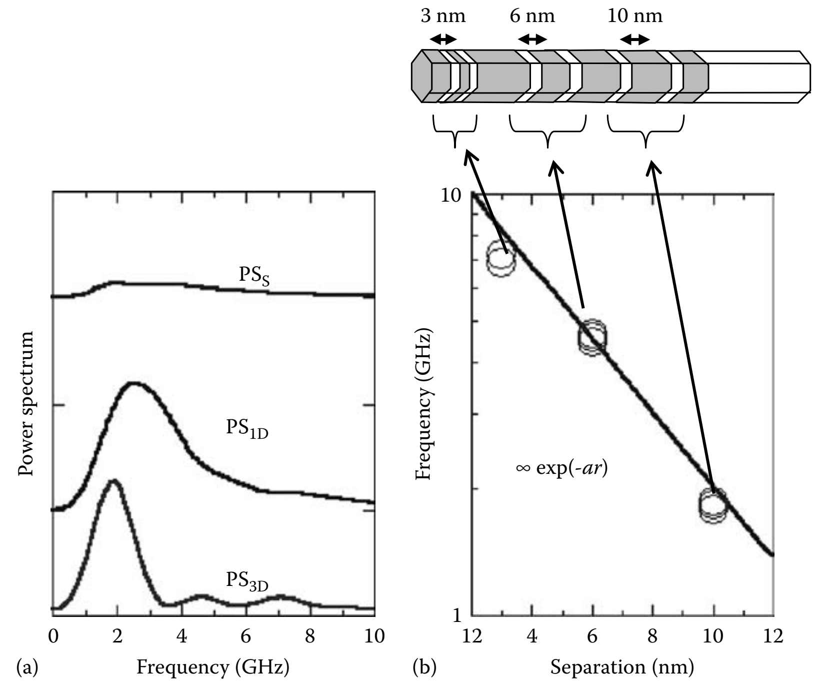

FIGURE 36.15 Evaluation of the nutation frequencies between the QWs. (a) PS$_S$, PS$_{1D}$, and PS$_{3D}$ show the power spectrum of TR$_S$, TR$_{1D}$, and TR$_{3D}$, respectively. (b) Separation D dependence of frequency of the nutation.

the period of nutation, the strength of the near-field coupling is estimated to be $U_{12} = 7.7\,\text{ns}^{-1}$ (= 4.9 μeV).

We evaluated nutation frequencies using Fourier analysis. In Figure 36.15a, the power spectral density of SQWs (PS$_S$) does not exhibit any peaks, indicating a monotonic decrease. By contrast, the power spectral density of 1-DQWs (PS$_{1D}$) had a strong peak at a frequency of $2.6\,\text{ns}^{-1}$. Furthermore, that of 3-DQWs (PS$_{3D}$) had three peaks at 1.9, 4.7, and $7.1\,\text{ns}^{-1}$. Since, the degree of the coupling strength, which is proportional to the frequency of the nutation, increases as the separation decreases, the three peaks correspond to the signals from DQWs with separations of 10, 6, and 3 nm, respectively. Since the coupling strength $\hbar U$ [eV] is given by $\hbar\pi f$ (f: nutation frequency), $\hbar U$ is estimated as 4.0, 9.9, and 14.2 μeV for DQWs with respective separations of 10, 6, and 3 nm. These values are comparable to that estimated above (U_{12} = 4.9 μeV). Furthermore, the peak intensity for the DQWs with 3 nm separation is much lower than for those with 10 nm separation, which might be caused by decoherence of the exciton state due to penetration of the electronic carrier. Considering the carrier penetration depth, the strong peak of DQWs with 10 nm separation originates from the near-field coupling alone. The solid line in Figure 36.15b shows the separation dependence of the peak frequency. The exponentially decaying dependence represented by this line supports the origin of the peaks in the power spectra from the localized near-field interaction between the QWs.

Next, we performed the switching operation. Figure 36.16a and b explains the "OFF" and "ON" states of the proposed nanophotonic switch, consisting of two coupled QWs. QW$_1$ and QW$_2$ are used as the input/output and control ports of the switch, respectively. Assuming L_w = 3.2 and 3.8 nm, the ground exciton state in QW$_1$ and the first excited state in QW$_2$ resonate. In the "OFF" operation (Figure 36.16a), all the exciton energy in QW$_1$ is transferred to the excited state in the neighboring QW$_2$ and relaxes rapidly to the ground state. Consequently, no output

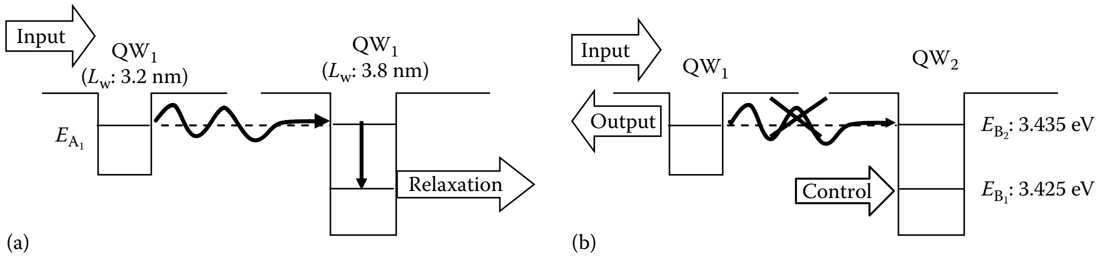

FIGURE 36.16 The switching operation by controlling the exciton excitation. Schematic of the nanophotonic switch of (a) "OFF" state and (b) "ON" state.

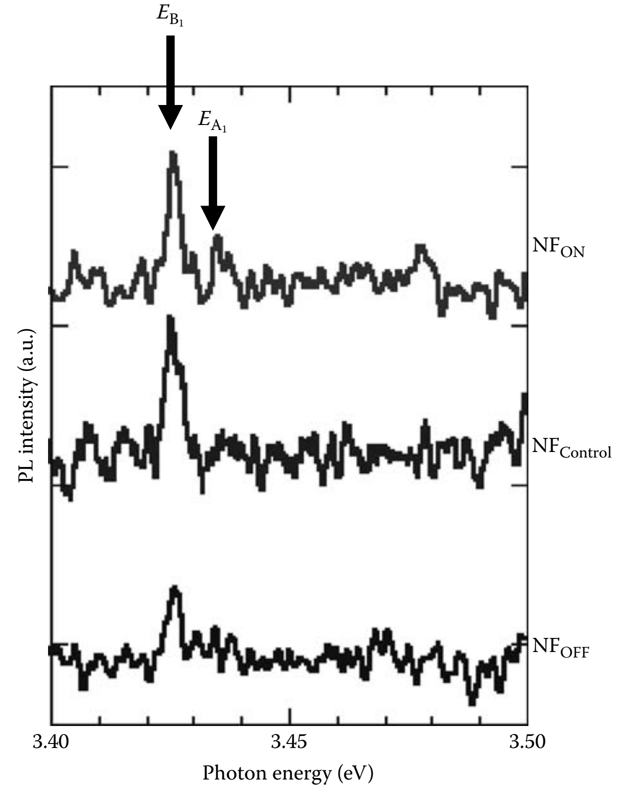

FIGURE 36.17 (a) NF_{ON}, $NF_{Control}$, and NF_{OFF} show NFPL signal obtained with the illumination of input laser alone, control laser alone, and input and control laser, respectively.

signals are generated from QW_1. In the "ON" operation (Figure 36.16b), the escape route to QW_2 is blocked by the excitation of QW_2 owing to state filling in QW_2 on applying the control signal; therefore, an output signal is generated from QW_1.

Figure 36.17 shows the NFPL for the three pairs of DQWs with L_w = 3.2 and 3.8 nm with different separations (3, 6, and 10 nm). Curve NF_{OFF} was obtained with continuous input light illumination from a He–Cd laser (3.814 eV). No emission was observed from the exciton ground state of QW_1 (E_{A1}) or the excited state of QW_2 (E_{B2}) at a photon energy of 3.435 eV, indicating that the excited energy in QW_1 was transferred to the excited state of QW_2. Furthermore, the excited state of QW_2 is a dipole-forbidden level. Curve $NF_{Control}$ shows the NFPL signal obtained with control light excitation of 3.425 eV with a 10 ps pulse. Emission from the ground state of QW_2 at a photon energy of 3.425 eV was observed. Both input and control light excitation resulted in an output signal with an emission peak at 3.435 eV, in addition to the emission peak at 3.425 eV (curve NF_{ON}), which corresponds to the ground state of QW_2. Since the excited state of QW_2 is a dipole-forbidden level, the observed 3.435 eV emission indicates that the energy transfer from the ground state of QW_1 to the excited state of QW_2 was blocked by the excitation of the ground state of QW_2.

Finally, the dynamic properties of the nanophotonic switching were evaluated. We observed TR_{NFPL} signals at 3.435 eV with both input and control laser excitation (see Figure 36.18). The decay time constant was found to be 483 ps. The output signal increased synchronously, within 100 ps, with the control pulse. Since the rise time is considered equal to one-quarter of the nutation period τ (Sangu et al. 2003), the value agrees with those obtained for DQWs with the same well width in the range from $\tau/4$ = 36 ps (3 nm separation) to $\tau/4$ = 125 ps (10 nm separation).

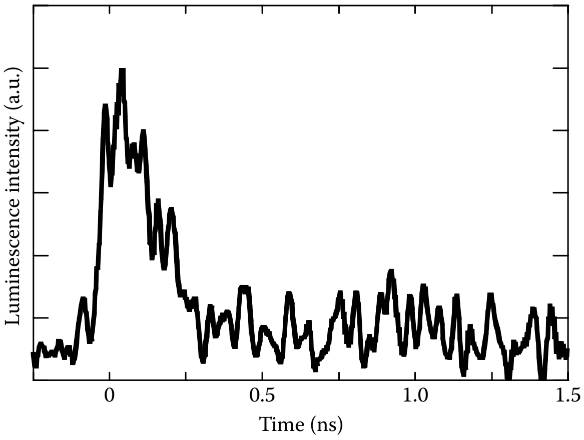

FIGURE 36.18 Near-field time-resolved PL signal with "ON" state.

We observed the nutation between DQWs and demonstrated the switching dynamics by controlling the exciton excitation in the QWs. Examination of the electronic coupling between QWs is now in progress to analyze the detailed switching dynamics. For room-temperature operation, since the spectral width reaches thermal energy (26 meV), a higher Mg concentration in the barrier layers and narrower L_w are required so that the spectral peaks of the first excited state (E_2) and ground state (E_1) do not overlap. This can be achieved by using two QWs with L_w = 1.5 nm (QW_1) and 2 nm (QW_2) with a Mg concentration of 50%, where the energy difference between E_2 and E_1 in QW_2 is 50 meV (Park 2001).

References

Bawendi, M. G., Carroll, P. J., William, L. W., and Brus, L. E. 1992. Luminescence properties of CdSe quantum crystallites: Resonance between interior and surface localized states. *J. Chem. Phys.* 96: 946–954.

Bayer, M., Hawrylak, P., Hinzer, K. et al. 2001. Coupling and entangling of quantum states in quantum dot molecules. *Science* 291: 451–453.

Brus, L. E. 1984. Electron–electron and electron–hole interactions in small semiconductor crystallites: The size dependence of the lowest excited electronic state. *J. Chem. Phys.* 80: 4403–4409.

Brust, M. and Kiely, C. J. 2002. Some recent advances in nanostructure preparation from gold and silver particles: A short topical review. *Colloids Surf. A* 202: 175–186.

Coffey, B. and Friedberg, R. 1978. Effect of short-range Coulomb interaction on cooperative spontaneous emission. *Phys. Rev. A* 17: 1033–1048.

Crooker, S. A., Barrick, T., Hollingsworth, J. A., and Klimov, V. I. 2003. Multiple temperature regimes of radiative decay in CdSe nanocrystal quantum dots: Intrinsic limits to the dark-exciton lifetime. *Appl. Phys. Lett*. 82: 2793–2795.

Guyot-Sionnest, P., Shim, M., Matranga, C., and Hines, M. 1999. Intraband relaxation in CdSe quantum dots. *Phys. Rev. B* 60: R2181–R2184.

Huang, M. H., Mao, S., and Feick, H. 2001. Room-temperature ultraviolet nanowire nanolasers. *Science* 292: 1897–1899.

Itoh, T., Furumiya, M., Ikehara, T., and Gourdon, C. 1990. Size-dependent radiative decay time of confined excitons in CuCl microcrystals. *Solid State Commun*. 73: 271–274.

Kawazoe, T., Kobayashi, K., Akahane, K., Naruse, M., Yamamoto, N., and Ohtsu, M. 2006. Demonstration of nanophotonic NOT gate using near-field optically coupled quantum dots. *Appl. Phys. B* 84: 243–246.

Kawazoe, T., Kobayashi, K., and Ohtsu, M. 2005. Optical nanofountain: A biomimetic device that concentrates optical energy in a nanometric region. *Appl. Phys. Lett*. 86: 103102.

Kawazoe, T., Kobayashi, K., Sangu, S., and Ohtsu, M. 2003. Demonstration of a nanophotonic switching operation by optical near-field energy transfer. *Appl. Phys. Lett*. 82: 2957–2959.

Kobayashi, K., Sangu, S., Itoh, H., and Ohtsu, M. 2000. Near-field optical potential for a neutral atom. *Phys. Rev. A* 63: 013806.

Kobayashi, K., Sangu, S., Kawazoe, T., and Ohtsu, M. 2005. Exciton dynamics and logic operations in a near-field optically coupled quantum-dot system. *J. Lumin*. 112: 117–121.

Meulenkamp, E. A. 1998. Synthesis and growth of ZnO nanoparticles. *J. Phys. Chem. B* 102: 5566–5572.

Ohtomo, A., Tamura, K., Kawasaki, M. et al. 2000. Room-temperature stimulated emission of excitons in ZnO/(Mg, Zn)O superlattices. *Appl. Phys. Lett*. 77: 2204–2206.

Ohtsu, M., Kobayashi, K., Kawazoe, T., Sangu, S., and Yatsui, T. 2002. Nanophotonics: Design, fabrication, and operation of nanometric devices using optical near fields. *IEEE J. Sel. Top. Quantum Electron*. 8: 839–862.

Park, W. I., Yi, G.-C., and Jang, M. 2001. Metalorganic vapor-phase epitaxial growth and photoluminescent properties of $Zn_{1-x}Mg_xO$ ($0 < x < 0.49$) thin films. *Appl. Phys. Lett*. 79: 2022–2024.

Park, W. I., An, S. J., Long, Y.-J. et al. 2004. Photoluminescent properties of $ZnO/Zn_{0.8}Mg_{0.2}O$ nanorod single-quantum-well structures. *J. Phys. Chem. B* 108: 15457–15460.

Park, W. I., Kim, D. H., Jung, S.-W., and Yi, G.-C. 2002. Metalorganic vapor-phase epitaxial growth of vertically well-aligned ZnO nanorods. *Appl. Phys. Lett*. 80: 4232–4234.

Park, W. I., Yi, G.-C., Kim, M. Y., and Pennycook, S. J. 2003. Quantum confinement observed in ZnO/ZnMgO nanorod heterostructure. *Adv. Mater*. 15: 526–529.

Sakakura, N. and Masumoto, Y. 1997. Persistent spectral-hole-burning spectroscopy of CuCl quantum cubes. *Phys. Rev. B* 56: 4051–4055.

Sangu, S., Kobayashi, K., Kawazoe, T., Shojiguchi, A., and Ohtsu, M. 2003. Quantum-coherence effect in a quantum dot system coupled by optical near fields. *Trans. Mater. Res. Soc. Jpn*. 28: 1035–1038.

Sangu, S., Kobayashi, K., Shojiguchi, A., and Ohtsu, M. 2004. Logic and functional operations using a near-field optically coupled quantum-dot system. *Phys. Rev. B* 69: 115334.

Spanhel, L. and Anderson, M. A. 1991. Semiconductor clusters in the sol-gel process; quantized aggregation, gelation, and crystal growth in concentrated ZnO colloids. *J. Am. Chem. Soc*. 113: 2826–2833.

Stinaff, E. A., Scheibner, M., Bracker, A. S. et al. 2006. Optical signatures of coupled quantum dots. *Science* 311: 636–639.

Sun, H. D., Makino, T., Segawa, Y. et al. 2002. Enhancement of exciton binding energies in ZnO/ZnMgO multiquantum wells. *J. Appl. Phys*. 91: 1993–1997.

Suzuki, T., Mitsuyu, T., Nishi, K. et al. 1996. Observation of ultrafast all-optical modulation based on intersubband transition in n-doped quantum wells by using free electron laser. *Appl. Phys. Lett*. 69: 4136–4138.

Trallero-Giner, C., Debernardi, A., Cardona, M., Menendez-Proupin, M., and Ekimov, A. I. 1998. Optical vibrons in CdSe dots and dispersion relation of the bulk material. *Phys. Rev. B* 57: 4664–4669.

Yatsui, T., Jeong, H., and Ohtsu, M. 2008. Controlling the energy transfer between near-field optically coupled ZnO quantum dots. *Appl. Phys. B* 93: 199–202.

Yatsui, T., Sangu, S., Kawazoe, T., Ohtsu, M., An, S. J., Yoo, J., and Yi, G.-C. 2007. Nanophotonic switch using ZnO nanorod double-quantum-well structures. *Appl. Phys. Lett*. 90: 223110.

37

Waveguides for Nanophotonics

Jan Valenta
Charles University

Tomáš Ostatnický
Charles University

Ivan Pelant
Academy of Sciences of the Czech Republic

37.1 Introduction **37**-1
Basic Types of Waveguides for Nanophotonics • Silicon Nanophotonics
37.2 Fabrication of Planar and Rib Waveguides **37**-2
37.3 Experimental Techniques **37**-4
Techniques to Study Internal Photoluminescence Propagation in Waveguides • Techniques to Study Propagation of Light from External Sources • Techniques to Study Losses and Optical Amplification in Waveguides
37.4 Main Experimental Observations in Active Si-nc Waveguides **37**-7
37.5 Theoretical Description of Active Lossy Waveguides **37**-9
37.6 Application of Active Nanocrystalline Waveguides **37**-12
Amplification of Light • Other Applications
37.7 Conclusions **37**-14
Acknowledgments **37**-14
References **37**-14

37.1 Introduction

The term *nanophotonics* may be understood as abbreviation from "photonics of nanostructures." This rapidly evolving research area deals with light interaction in nanostructured materials and their applications in photonic devices like light sources, modulators, detectors, etc.

37.1.1 Basic Types of Waveguides for Nanophotonics

Optical waveguides are structures that are able to confine and guide optical electromagnetic field. *Classical waveguides* (optical fibers, Figure 37.1a) use refractive index difference between the core and the cladding layer to guide light by total internal reflection. When the size of waveguide approaches the wavelength of light, the confinement decreases and losses increase. Therefore, the size of practical classical waveguides is limited to several hundreds of nanometers. Somewhat better confinement can be achieved in more complex waveguides with guiding properties determined by the formation of photonic bands due to their regular spatial structure (Figure 37.1b and c): (1) *photonic crystal waveguides* using "defects" in periodic structures of photonic crystals to confine light and (2) *plasmonic waveguides* based on metallic nanostructures with plasmon resonance (see, e.g., Pavesi and Guillot 2006). Due to limited space, we are going to describe only a special type of the "classical" waveguides formed by luminescing silicon nanocrystals (Si-nc).

37.1.2 Silicon Nanophotonics

The recent years can be characterized by the association of microelectronics with optoelectronics or photonics. While the microelectronics is based almost exclusively on silicon (indirect band-gap semiconductor), photonic light sources are currently made out of direct band-gap III–V compounds (family of GaAs materials). In order to reduce the material diversity, an effort to develop an efficient, electrically pumped silicon-compatible light-emitting device is becoming very strong (Pavesi and Lockwood 2004). This tendency to "siliconize photonics" is driven also by the need to reduce the overheating of silicon integrated circuits due to the ohmic resistance of excessively long multilevel metal interconnects (Pavesi and Guillot 2006). Supplementary advantages like the reduction in charging (RC) time constant (speeding up the circuit performance) and prevention from crosstalks can be obtained as a bonus.

Since bulk silicon, as an indirect band-gap semiconductor, is a very bad light emitter, nanometer-sized silicon nanocrystals, brightly luminescent at room temperature, represent one of the possible solutions (Pavesi and Lockwood 2004). It is then expected that the light signal originating in Si-nc and carrying the required information propagates in a low-loss medium, which assures, at the same time, a good directionality of the radiation. It has been discovered recently that slabs of fused quartz SiO_2 "doped" with Si-nc are able to accomplish both the role of a light emitter and that of a waveguide (Khriachtchev et al. 2001, Valenta et al. 2003a, Khriachtchev et al. 2004). We shall call this type of nanophotonic waveguides (which generate the luminous signal themselves and therefore there is no need to

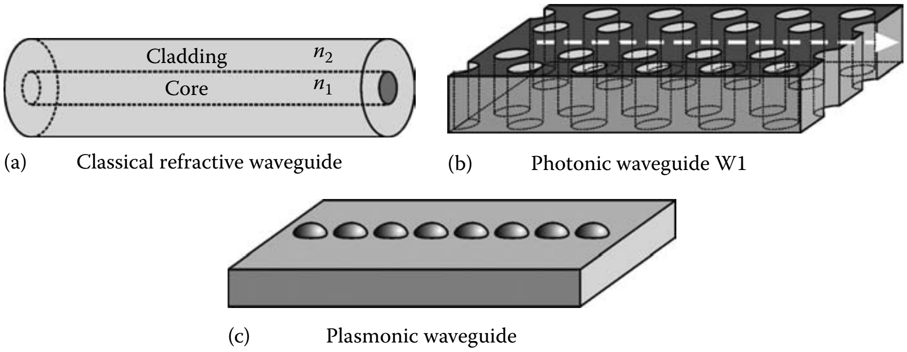

FIGURE 37.1 Nanophotonic waveguides: (a) "Classical" waveguide based on total internal reflection using the core with higher refractive index than the cladding. (b) Photonic crystal waveguide formed by a "defect" in the regular structure—here is the W1 waveguide formed by a row of missing holes. (c) Plasmonic waveguide formed by a row of metallic nanocrystals.

couple the light to them from an external source) "active waveguides." In this chapter, we shall describe the preparation methods of active nanocrystalline waveguides, their experimentally observed properties, and relevant theoretical description. We shall also briefly mention their various application possibilities.

37.2 Fabrication of Planar and Rib Waveguides

There have been many various methods of how to grow thin sheets of luminescent Si-nc embedded in an optically transparent matrix: (1) *Implantation of fused quartz* (SiO_2) slabs with Si^+ ions (Figure 37.2) (Cheylan et al. 2000), (2) *reactive Si deposition onto fused quartz* (Khriachtchev et al. 2002), (3) *plasma-enhanced chemical vapor deposition* (PECVD) of substoichiometric silicon oxide SiO_x thin films on a Si substrate (Iacona et al. 2000), (4) *co-sputtering of a Si wafer and a piece of glass* using fused quartz plates or Si wafers as a substrate (Imakita et al. 2005), (5) *crystallization of Si/SiO$_2$ superlattices* with nm-thick amorphous Si or SiO layers (Riboli et al. 2004), and (6) *embedding of porous silicon grains into sol-gel derived SiO$_2$ layers* (Luterová et al. 2004), to list at least the most frequently used. Most of these techniques comprise high-temperature (1100°C–1200°C) annealing of deposited SiO_x films in order to achieve the phase separation between Si-nc and the SiO_2 matrix. The thickness of the resulting sheets containing Si-nc can vary from hundreds of nanometers to tens of micrometers. Due to the difference in refractive index between the matrix ($n_{\text{silica}} \approx 1.45$) and silicon ($n_{\text{Si}} \approx 3$), such sheets act as planar or rib waveguides. Interestingly, the attractive waveguiding features that we are going to discuss, namely, the wavelength selective guiding of light (which can be also called "spectral filtering") and microcavity-like behavior, are critically dependent upon the preparation method: Till now, they have been discovered only in samples fabricated using the first two methods. We describe them in detail.

Implantation of accelerated Si^+ ions can be applied either to fused quartz (silica) slabs or to a thin SiO_2 layer thermally

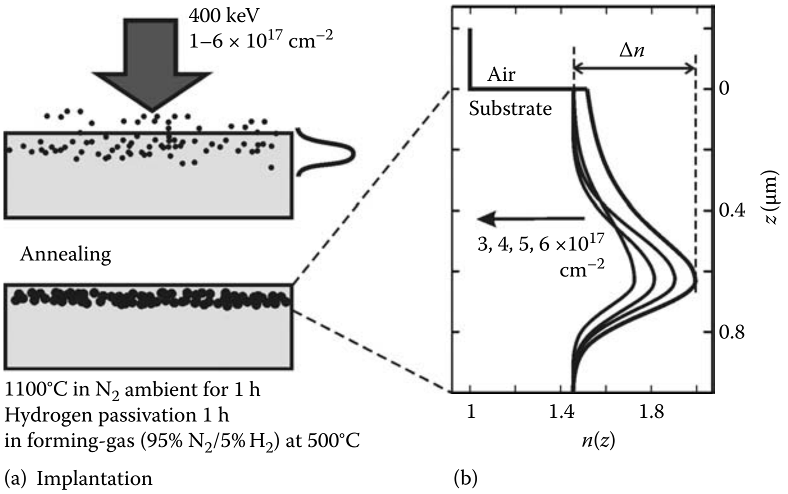

FIGURE 37.2 (a) The schematics of a Si-nc planar waveguide preparation. (b) Refractive index profiles (n as a function of the depth z beneath the surface) of silicon nanocrystalline waveguides (implanted with 400 keV Si^+ ions and different implant fluences), extracted from optical transmission measurements. (After Valenta, J. et al., *J. Appl. Phys.*, 96, 5222, 2004.)

grown on a silicon wafer. Implantation energy varies usually between 30 and 600 keV, and implant fluences (ion doses) are of the order of 10^{16}–10^{17} cm^{-2}. The attractive waveguide properties of fabricated planar waveguides, i.e., pronounced emission line narrowing and high output beam directionality in the visible region, have been found in waveguides prepared from 1 mm thick silica (Infrasil) slabs with optically polished surface and edges, using the implantation energy of 400 keV and the implant fluences 1×10^{17} cm^{-2} to 6×10^{17} cm^{-2}. Because the Si$^+$ ions are almost monoenergetic, their stopping distance beneath the silica slab surface has only a small dispersion, which results in narrow and slightly skewed implant distribution, as reflected in the resulting refractive index profiles shown in Figure 37.2 (Valenta et al. 2004). Peak excess Si concentration was up to 26 atomic%. Spherical silicon nanocrystals with diameter distribution roughly from 4 to 6 nm were formed by annealing the implanted samples at 1100°C in a nitrogen ambient for 1 h. To enhance several times the luminescence intensity of Si-nc, an additional anneal at 500°C in forming gas ($N_2 + H_2$) for 1 h was applied (Cheylan and Elliman 2001). The last operation promotes enhanced hydrogen passivation of non-radiative defects.

The reactive silicon deposition has yielded substoichiometric SiO_x films on circular 1 mm-thick substrate silica plates (Khriachtchev et al. 2001, 2004). Electron beam evaporation and radio frequency cells were used as silicon and oxygen sources. The SiO_x layer thickness was ~2 μm, and the value of the x parameter varied around $x = 1.70$. The as-grown material contained amorphous Si inclusions. Annealing of these "suboxide" films at 1100°C in nitrogen atmosphere resulted in the formation of well-defined Si-nc with diameters of 3–4 nm, as evidenced by Raman spectroscopy. Extensive investigations of this type of silica waveguides containing Si-nc discovered a small effective optical birefringence inside the layers, however, without radically influencing the properties of the waveguides (Khriachtchev et al. 2007). Refractive index profile across the thickness of these waveguides is flat, unlike the preceding case of the implanted layers.

The reason why Si-nc waveguides fabricated using other techniques do not show the spectral filtering effect are not fully clear at present. Of course, the condition *sine non qua* is asymmetry of the waveguide, i.e., the waveguiding layer must not be sandwiched between two materials (cladding layers) having the same effective refractive index, as we shall see in the next section. One can speculate about several further critical parameters: suitable composition and optical quality of the matrix; suitable density of Si-nc (in the order of 10^{18} cm^{-3}); favorable layer thickness (around 1 μm); excellent flatness and parallelism of both interfaces; and, last but not least, certain optimum value of propagation losses (several dB/cm). The dominant losses in the waveguides under discussion are probably due to self-absorption and Mie light scattering on Si-nc clusters, not due to surface roughness (Pelant et al. 2006).

The above discussion has dealt implicitly with planar waveguides. Rib-loaded waveguides containing nanocrystals seem, in principle, even more desirable for nanophotonic applications. They can be also prepared using the ion implantation technique in either thin SiO_2 films thermally grown on Si or in polished fused quartz slabs, as described above. The rib structure can be formed (using standard photolithography and etching) by two possible means: either before or after the implantation procedure. Figure 37.3a represents schematics of such a rib waveguide structure.

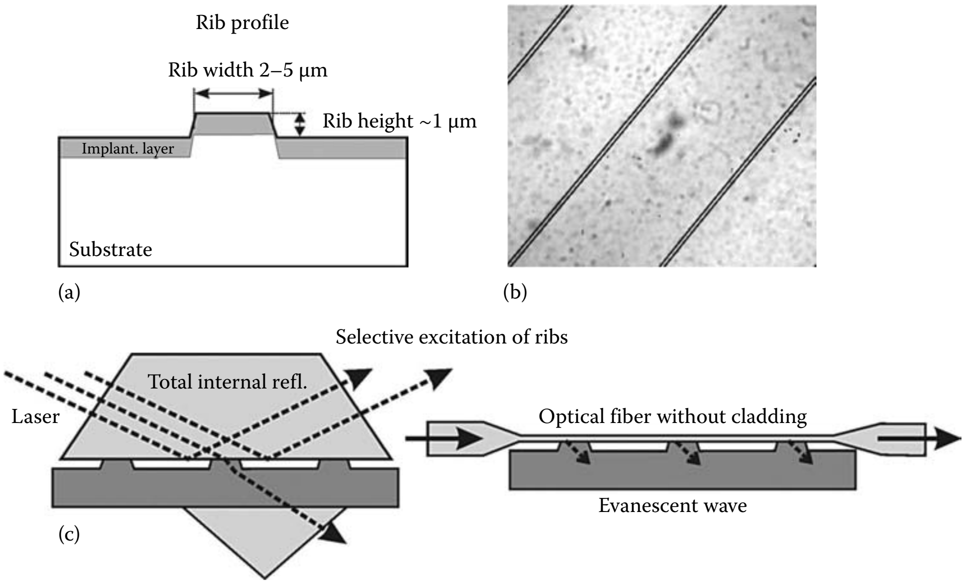

FIGURE 37.3 (a) Sketch of a rib waveguide structure with silicon nanocrystals, prepared on a polished silica slab. The ribs were fabricated by optical lithography before Si$^+$ implantation. (b) Microscopic (reflection) image of a part of the structure. The spacing between the ribs is 100 μm. (c) Methods to selectively excite the rib waveguides through the TIR prism or the evanescent field of a fiber. (After Skopalová, E., Mode structure in the light emission from planar waveguides with silicon nanocrystals, Diploma thesis, Charles University, Prague, Czech Republic, 2007.)

Lateral confinement of the light in a narrow rib is expected, in comparison with the planar waveguides, to have some advantages: saving considerable space in photonics circuits and offering much wider variability in their design. On the other hand, the transition from the two-dimensional to the one-dimensional case can degrade the optical quality of the waveguide and introduce additional losses, such as those due to sidewall roughness or even due to considerably deformed rib cross section. Indeed, measurements of propagation losses in similar waveguides have given values above 10 dB/cm, and this loss coefficient even increases for rib widths below 4 μm (Pellegrino et al. 2005). Therefore, in what follows, we discuss predominantly the two-dimensional nanocrystalline waveguides, and limited space only will be given to the rib structures.

37.3 Experimental Techniques

37.3.1 Techniques to Study Internal Photoluminescence Propagation in Waveguides

One of the most important advantages of active waveguides is that light sources are embedded in the waveguide. In our material, Si nanocrystals forming the waveguide core can efficiently emit luminescence in the orange-red-infrared spectral range when excited with the UV-blue light (usually focused laser beam). Such internally produced light is automatically coupled to all possible modes of the waveguide: radiation, substrate, and guided modes (see Section 37.5). The emission modes may be distinguished by measuring their propagation direction, spectra, and polarization, which imposes requirements to experimental setup.

Two types of setups can be conveniently applied to study luminescence from active waveguides: (1) The micro-imaging-spectroscopy setup is based on a microscope connected to an imaging spectrograph with a CCD camera (Figure 37.4a). For good angular resolution, low numerical aperture (NA) objective lenses should be used (in our experiments, we used mostly the lens with 2.5 × magnification and NA = 0.075, i.e., an angular resolution of about 8.6°). The sample is fixed to an *x*–*y*–*z* table with a rotating holder and excited by a laser beam (325 nm from the cw He–Cd laser), approximately perpendicular to the objective axis. (2) In order to achieve better angular resolution, an experimental arrangement based on a goniometer is employed (Figure 37.4b). Here, the sample is fixed to the center of the goniometer, and the photoluminescence emission is collected by a silica optical fiber (core diameter 1 mm) rotated around the sample at a distance of 50 mm, giving an angular resolution slightly less than 1° (NA ~ 0.01). The output of the fiber can be coupled to the same detection system (a spectrograph with a CCD camera), as described above.

Typical images observed with the microscopic setup (Figure 37.4a) are illustrated in Figure 37.4c and d. Here, the diameter of the excitation spot, located about 1 mm from the sample edge, is roughly 1 mm. One can easily recognize PL emission from the excited spot as a bright ellipsoid. However, there is also a second

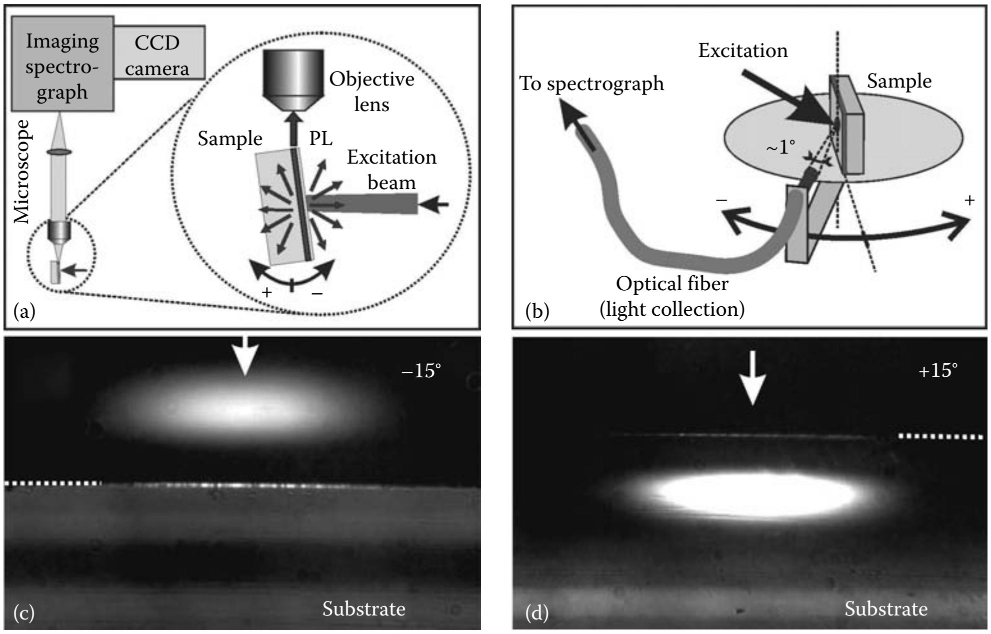

FIGURE 37.4 (a) Micro-spectroscopy setup for studying spatially resolved emission from waveguides, (b) collection of signals from a waveguide using optical fiber mounted on a goniometer. (c,d) Luminescence microimages of the active planar waveguide (obtained with the setup illustrated in panel a), where the elliptical spot is the spot excited by the laser beam and the light line is emission emanating from the sample facet. The inclination of the sample with regard to optical axis is −15° and +15°, respectively. (Adapted from Valenta, J. et al., *J. Appl. Phys.*, 96, 5222, 2004.)

contribution emanating from the facet of the sample. This light is obviously guided in the implanted layer or close to it. The images in Figure 37.4c and d were collected for sample inclination angles of −15° and +15°, respectively, i.e., in a geometry for which the excited spot was observed either directly (Figure 37.4c) or through the substrate (Figure 37.4d). The experimental arrangement shown in Figure 37.4a enables the detection of the PL either from the excited spot or from the edge of sample by positioning the entrance slit of the spectrograph to different locations of the PL image. All experiments described in this chapter were performed at room temperature.

37.3.2 Techniques to Study Propagation of Light from External Sources

The coupling of external light to narrow submicrometer waveguides is a difficult task. The most used approaches are as follows: (1) *Prism coupling of light* from the surface of the sample (Figure 37.5a). Light from Xe lamp, halogen lamp, or LED is collimated into a prism in contact with waveguide. For better optical contact, an immerse liquid should be dropped between the prism and the sample (the best optical contact—i.e., minimizing reflections on interface—is achieved when refractive index of the immersion liquid is between the values for the materials to be connected). A second prism may be placed on the opposite surface of waveguide to couple out the passing beam to avoid total internal reflection. (2) *Direct coupling into the facet* (Figure 37.5b, sometimes called *end-fire coupling*). The edge of the sample can be polished at some angle (here 70°) in order to separate light refracted to the higher index waveguide from light entering lower index substrate. For both the external light coupling setups, the signal can be collected by an optical fiber and guided to the entrance slit of the imaging spectrometer with a CCD detector. The other coupling methods described in literature are *grating coupling* and *evanescent wave coupling* (sometimes called waveguide-to-waveguide coupling) (see Figure 37.5c and d).

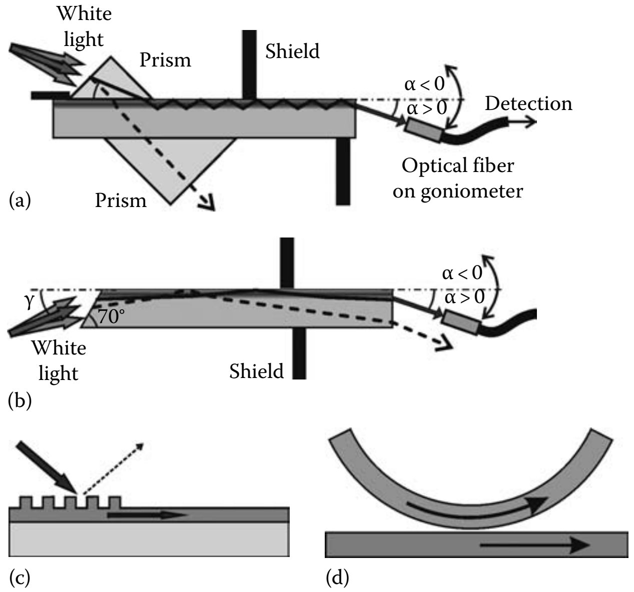

FIGURE 37.5 External light coupling into a narrow waveguide: (a) prism coupling, (b) end-fire coupling, (c) grating coupling, and (d) evanescent-wave coupling. (Adapted from Janda, P. et al., *J. Lumin.*, 121, 267, 2006; Pavesi, L. and Guillot, G. (Eds.), *Optical Interconnects. The Silicon Approach*, Springer-Verlag, Berlin, Germany, 2006.)

37.3.3 Techniques to Study Losses and Optical Amplification in Waveguides

The losses in waveguide are divided into *insertion* (*coupling*) *losses* and *propagation losses*, which contain *scattering*, *radiation*, and *absorption* (Figure 37.6a). If the waveguide material can be pumped to reach population inversion, the guided light may be amplified by stimulated emission. In this case, the absorption coefficient α becomes negative and it is called the gain coefficient $g (g = -\alpha)$.

The most used methods to measure losses in waveguides are (Figure 37.6b through e) (1) *"cutback" method* detecting light output from waveguide with different length; (2) *scattering detection* from different points of waveguide surface; (3) *Fabry–Perot resonance* method, which can be applied to waveguides with good facets; and (4) *shifting excitation spot* (SES) method; which can be applied to active waveguides.

Optical gain is measured by the *variable stripe-length* (VSL) method or by the *pump-and-probe* (PP) method. The principle of the VSL method consists in excitation of a narrow stripe-like region at the edge of a sample (Figure 37.7a) (Shaklee and Leheny 1971, Valenta et al. 2003b, Dal Negro et al. 2004). If the pumping is able to induce population inversion in studied material, photons passing through the excited region can be amplified by stimulated emission. It corresponds to a single-pass optical amplifier. If the emission from the sample edge $I^{\mathrm{VSL}}(l,\lambda)$ is detected as a function of the stripe length l, the net optical gain $G(\lambda)$ is calculated from a simple equation

$$I^{\mathrm{VSL}}(l,\lambda) = \frac{I_{\mathrm{sp}}(\lambda)}{G(\lambda)}\left[\exp(G(\lambda)\cdot l) - 1\right], \tag{37.1}$$

where

I_{sp} is the intensity of spontaneous emission
$G(\lambda) = [g(\lambda) - K]$ is net optical gain (K stands for losses and $g(\lambda)$ is material gain coefficient, i.e., negative absorption coefficient) (Valenta et al. 2002)

Equation 37.1 is valid only when $I_{\mathrm{sp}}(\lambda)$ and $g(\lambda)$ are constant over the whole excited stripe. It means that the exciting power density and properties of sample must be uniform. Also, the coupling of the output emission to a detector must be constant, independent of x, for any part of the stripe. In reality, the geometry and emission properties of the sample are never perfectly constant, and the coupling of light is influenced by directionality of emitted light, limited collection angle (NA), the confocal effect of imaging optics, etc. (Valenta et al. 2003b). Therefore, the SES method has been proposed to correct the VSL method (Valenta et al.

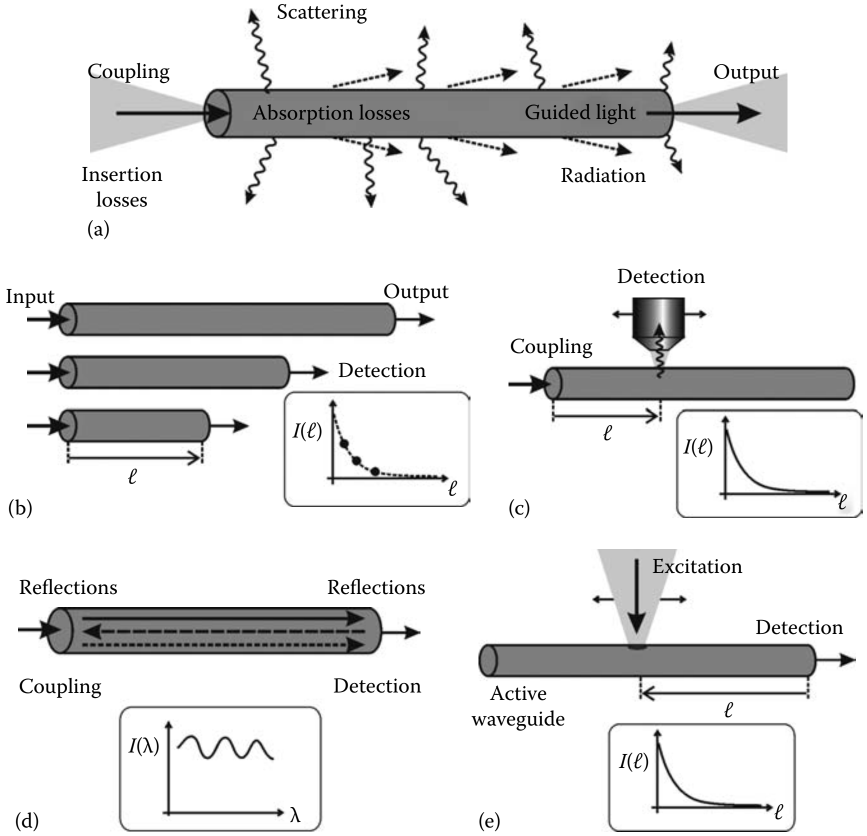
Scattering
Coupling
Output
Absorption losses
Guided light
Insertion losses
Radiation
(a)
Input
Output
Detection
$I(\ell)$
ℓ
(b)
Detection
Coupling
ℓ
$I(\ell)$
(c)
Reflections
Reflections
Coupling
Detection
$I(\lambda)$
λ
(d)
Excitation
Detection
Active waveguide
ℓ
$I(\ell)$
(e)

FIGURE 37.6 (a) Schematic illustration of losses in a waveguide. Methods to measure losses: (b) cutback, (c) scattering light detection, (d) Fabry–Perot resonance, and (e) SES methods.

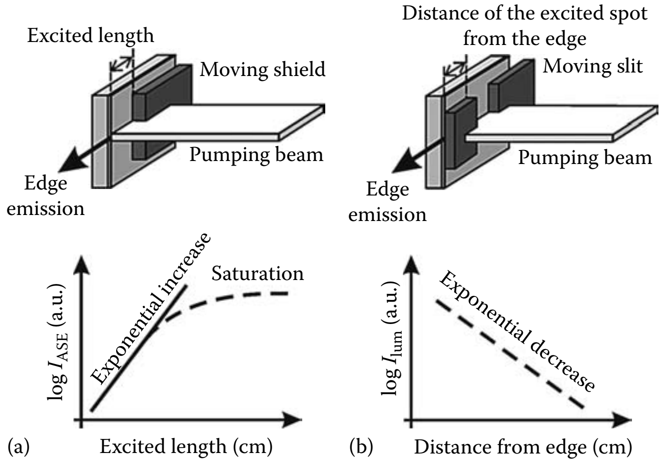
Excited length
Moving shield
Pumping beam
Edge emission
Distance of the excited spot from the edge
Moving slit
Pumping beam
Edge emission
$\log I_{\mathrm{ASE}}$ (a.u.)
Exponential increase
Saturation
(a)
Excited length (cm)
$\log I_{\mathrm{lum}}$ (a.u.)
Exponential decrease
(b)
Distance from edge (cm)

FIGURE 37.7 The principle of the (a) VSL and (b) SES methods and typical dependence of the detected signal as function of the excited length and the distance from edge, respectively (lower panel). (Adapted from Park, J.H. and Steckl, A.J., *Appl. Phys. Lett.*, 85, 4588, 2004.)

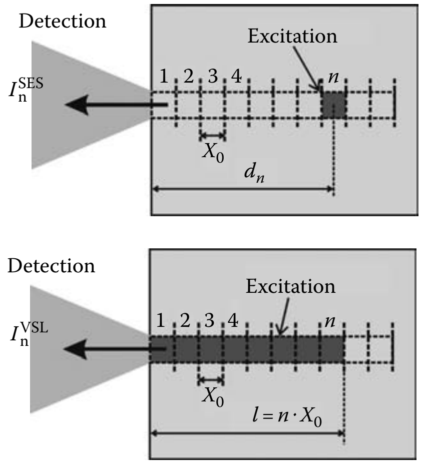
Detection
Excitation
I_n^{SES}
1 2 3 4
n
X_0
d_n
Detection
Excitation
I_n^{VSL}
1 2 3 4
n
X_0
$l = n \cdot X_0$

FIGURE 37.8 Schematic illustration of the principle of combining the SES and VSL experiments.

2002). The principle of the SES method takes advantage of the fact that instead of the whole stripe, only small segments are excited and measured separately (Figure 37.7b).

Figure 37.8 illustrates the procedure to compare the VSL and SES results. The most transparent approach consists of SES measurement with the spot width equal to the shift step. Let us assume that the VSL measurement is done using the same elemental steps x_0 as the SES experiment (obviously, also the stripe width and excitation density must be the same for both SES and VSL). Then we can compare the VSL signal I_n^{VSL} for a stripe consisting of n elemental steps ($l = n \cdot x_0$) with integrated SES signal I_n^{iSES}, i.e., sum of SES signals from n steps

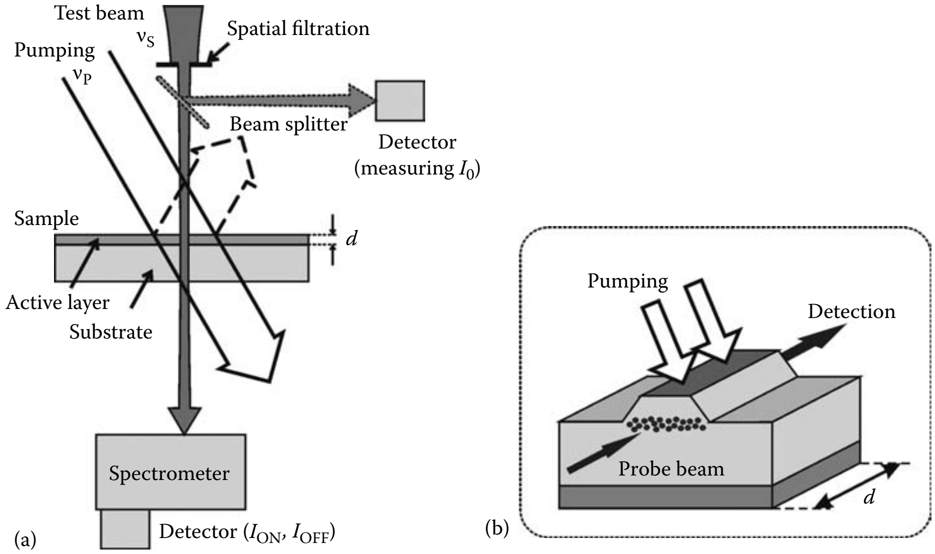

FIGURE 37.9 The basic arrangement of the PP experiment for a (a) planar and (b) rib waveguide.

$$I_n^{\mathrm{iSES}} = \sum_{k=1}^{n} I_k^{\mathrm{SES}}.$$

Integrated SES signal should contain the same amount of spontaneous emission as the VSL measurement (including losses and effects of imperfect experimental conditions), the only difference being that in VSL experiment some photons go through the excited area and might be amplified by stimulated emission or affected by induced absorption. Which effect takes place is clear from plotting both I_n^{VSL} and I_n^{iSES} in one figure.

Similarly, we can calculate the differential VSL intensity I_n^{dVSL}, i.e., the difference of signals obtained for stripe lengths $l = n \cdot x_0$ and $l = (n-1) \cdot x_0$

$$I_n^{\mathrm{dVSL}} = I_n^{\mathrm{VSL}} - I_{n-1}^{\mathrm{VSL}}$$

and compare it with I_n^{SES}. Again, the detected signal comes from spontaneous emission in the nth excited spot that is passing through the excited or unexcited area for the dVSL and SES, respectively. In the ideal case, we can use equations $I_n^{\mathrm{dVSL}} = \mathrm{const}\exp[(g-K)d_n]$ and $I_n^{\mathrm{SES}} = \mathrm{const}\exp(-Kd_n)$ to derive the net optical gain $g(\lambda)$

$$g(\lambda) = \frac{\ln\left(I_n^{\mathrm{dVSL}}(\lambda)/I_n^{\mathrm{SES}}(\lambda)\right)}{d_n},$$

where d_n is the distance of the center of the nth spot from the edge.

The PP technique is based on the application of two beams: the first one is pumping the active sample and the second one is testing the state created by the pumping beam. In case of studying stimulated emission, the test beam may be amplified by stimulated emission. In general, the test beam spot must be situated inside the pumping spot in order to test area pumped as homogeneously as possible. In case of pulsed beams, the temporal coincidence of both pumping and testing pulses must be controlled (the test beam may be delayed after pumping pulses to study decay of the effect). If the thickness of the active layer is small, the induced effect on the test beam may be difficult to detect. Therefore, for waveguide samples, the test beam is often coupled to the waveguide, which is excited from the surface (Figure 37.9b). Then, changes of the outcoupled light with and without pumping are detected.

37.4 Main Experimental Observations in Active Si-nc Waveguides

The most important observations for active planar waveguides are (1) spectral filtering; (2) TE (transverse electric) and TM (transverse magnetic) mode splitting; and (3) high directionality of TE, TM mode emission. These effects are illustrated in Figure 37.10; the panel (a) shows that the PL leaving the edge of waveguide layer is substantially different from the PL perpendicular to the layer. The perpendicular PL is formed by a single band that corresponds to the well-known emission of Si nanocrystals, but the edge PL contains two narrow PL bands. Both narrow lines are linearly polarized, one in the direction of the waveguide layer (TE mode) and the second one perpendicular to the layer (TM mode). The spectral position of the TE and TM modes depends on the exact profile of the refractive index, it means on the implantation fluence (see also Figure 37.16, which shows spectra of another set of samples). The panels (b) and (c) illustrate that the TE and TM modes are emitted only in an extremely narrow angle (a few degrees), basically in the direction of the waveguide plane (only slightly inclined to the substrate, i.e., under positive angles in our notation).

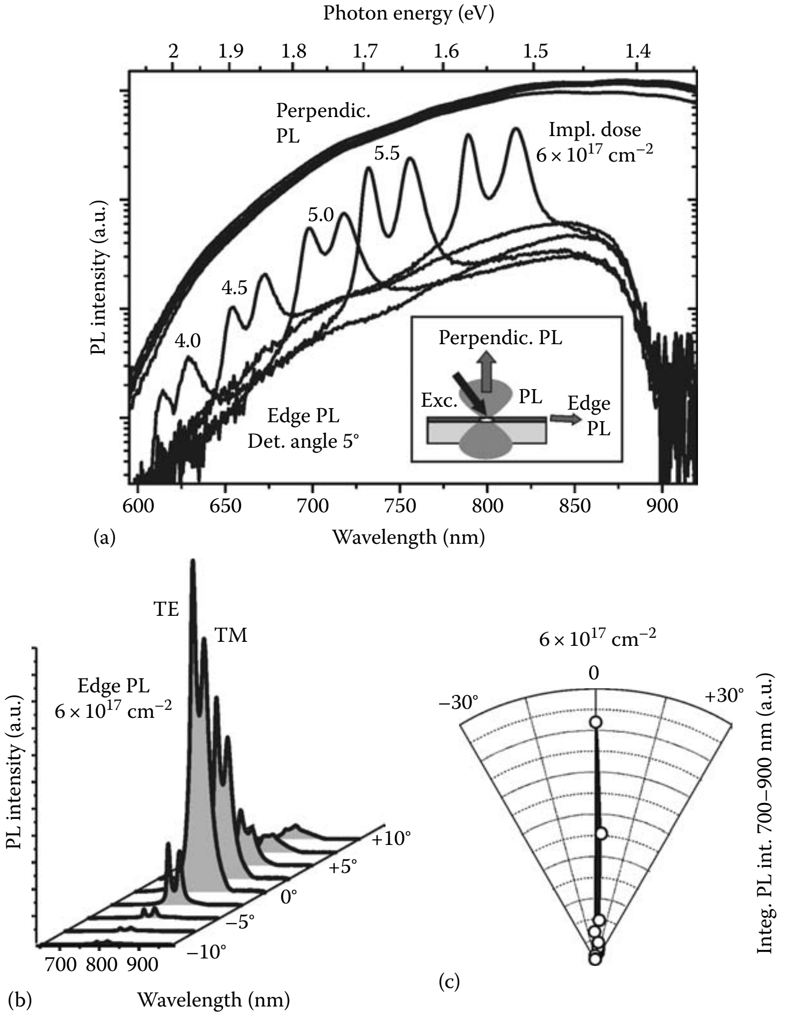

FIGURE 37.10 PL spectra of five fused silica slabs implanted to the fluences of 4×10^{17} cm^{-2} to 6×10^{17} cm^{-2}. (a) Upper curves (a single wide band) correspond to PL emitted in a direction perpendicular to the waveguide, while lower spectra with doublet peaks are facet-PL detected in a direction $\alpha = 5°$ (a sketch of the experimental arrangement is shown in the inset). (b) Angle-resolved facet PL spectra of the sample 6×10^{17} cm^{-2}. (c) Polar representation of integrated PL intensity of the angle-resolved facet spectra from the panel b. Most of the PL intensity is emitted in a direction close to 0°. (Adapted from Janda, P. et al., *J. Lumin.*, 121, 267, 2006.)

The coupling and propagation of *external* light in active Si-nc waveguides were studied by the direct (end-fire) and the prism coupling (Figure 37.5a and b). The results for one sample are illustrated in Figure 37.11. For prism coupling, two broad transmission bands (in the blue and red spectral region) are observed in the measured spectral range. The positions of both bands coincide with those of the PL modes (Figure 37.10a). Our calculations show that the red and blue bands corresponds to the second and third order of substrate modes (the first order being in the infrared region), see Section 37.5 and Figure 37.14. Broadening of the mode structure may be a consequence of the very low number of reflections undertaken by coupled light before escaping to the substrate. Coupling of external light through a truncated facet (Figure 37.5b) gives the best result for a coupling angle $\gamma \sim 20°$, as expected (Figure 37.11, upper curves). In this configuration, the narrow and polarization-split peaks at an output angle $\alpha \sim 2°$ are detected. The peaks are, however, not transmission but absorption peaks. This can be understood if it is assumed that the detected light comes not from substrate modes (which represent a small portion of transmitted light) but from filtered transmitted light propagating almost parallel to the Si-nc waveguide from which a part of power escaped to the substrate modes that are, however, gradually absorbed in the waveguide core. The blue third order modes are much stronger compared to second order because of higher absorption in the blue spectral region.

Experimental observation of rib waveguides indicates (Figure 37.12) that normal guiding of light is improved—see increased intensity of the long-wave edge of PL spectra in Figure 37.12a and b. This is due to the introduction of confinement in the second lateral direction of waveguide (in this case, the confining refractive index profile is symmetrical and the contrast higher (compared to the implanted layer profile) as the surrounding medium is air). On the other hand, roughness of

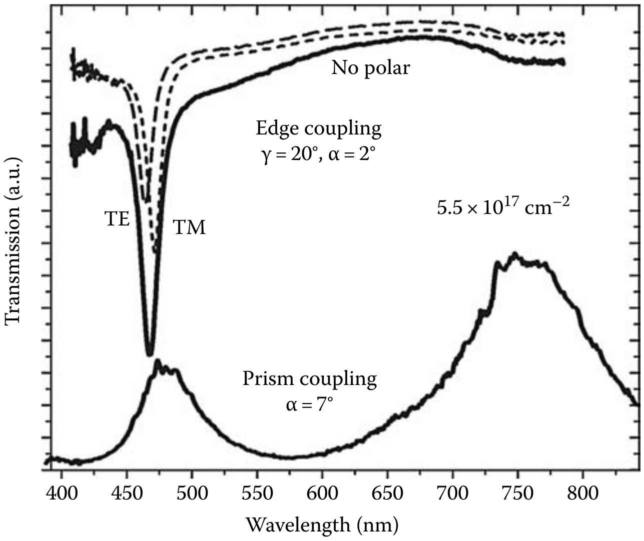

FIGURE 37.11 Comparison of transmission spectra of a sample 5.5×10^{17} cm^{-2} obtained by the direct facet-coupling (upper curves, solid line—no polarization, dashed and dotted lines correspond to TE and TM polarization, respectively) and by the prism coupling (lower spectrum). (Adapted from Janda, P. et al., *J. Lumin.*, 121, 267, 2006.)

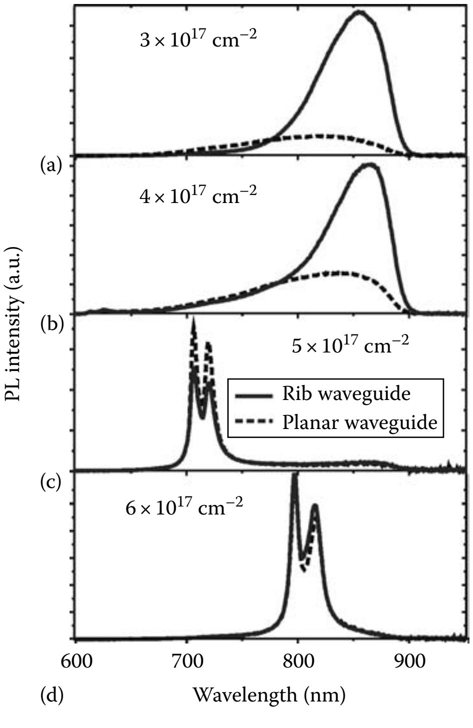

FIGURE 37.12 Comparison of the edge PL spectra from planar and rib waveguides (dashed and continuous line, respectively) in samples implanted to a fluence of (a) 3×10^{17} cm^{-2}, (b) 4×10^{17} cm^{-2}, (c) 5×10^{17} cm^{-2}, and (d) 6×10^{17} cm^{-2}.

the rib edges can introduce significant losses due to scattering. The influence of rib structure on the substrate (leaky) modes is less significant (see Figure 37.12c and d). This is because these modes propagate mostly in substrate, where the effect of side walls is negligible.

37.5 Theoretical Description of Active Lossy Waveguides

Light coupling to a waveguide and its further propagation may be understood in terms of coupling between the waveguide modes and nanocrystals emitting photons. Light field inside the waveguide may be expressed as a coherent superposition of excited waveguide modes, which arise as a solution of a homogeneous wave equation with proper boundary conditions [the waveguide modes are an orthonormal basis of the waveguide optical field (Snyder and Love 1983)]. It is therefore necessary to distinguish between various types of the waveguide modes in order to explain different types of energy transport in the waveguides.

Let us consider a common model structure with three transparent layers (see Figure 37.13): the top *cladding* layer with the refractive index n_1, the bottom *substrate* with the refractive index n_3, and the *core* with the refractive index $n_2 > n_3 > n_1$ and thickness d. We may imagine the modes as rays emerging from inside the waveguide core and propagating toward one of the core boundaries. At the boundary, the ray may be either totally reflected back to the core or it may be partially reflected and partially refracted. The ray then travels toward the other boundary where it totally or partially reflects again. Only those modes that are totally reflected twice during one round-trip may propagate in the core without losses: these are the common *guided modes*. There are, in addition, lossy modes that are called the *substrate radiation modes* (they are refracted only into the substrate) and the *radiation modes* refracted both to the substrate and the cladding.

In many applications, considering traveling of light at large distances compared to its wavelength (and thus working in the far-field limit), we may consider only the guided modes as efficient carriers of energy between a source and a detector. In silicon-based nanophotonics devices, however, the situation may be even more complicated. The typical distances between components on a chip may be comparable to the light wavelength and the far-field approach fails. It is therefore important to analyze an influence of the substrate radiation modes on the response. We may rule out the radiation modes at this stage: compared to the substrate modes, they play only a negligible role in the overall system behavior.

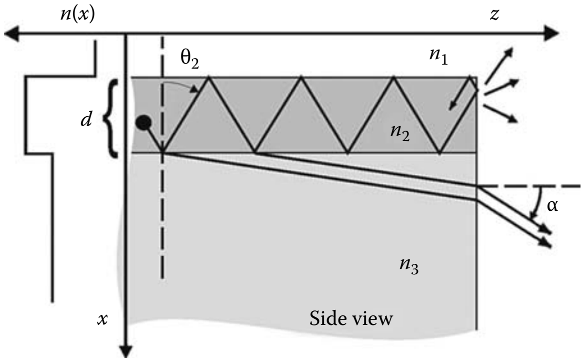

FIGURE 37.13 Schematic representation of propagation and decoupling of substrate modes.

Once light is emitted by a nanocrystal inside the waveguide mode, it undergoes internal reflections at the core/cladding and the core/substrate boundaries. The overall intensity of electric field of an excited mode inside the waveguide core may be written as superposition of the partial waves:

$$E_{\text{mode}}(\lambda,\theta_2) = E_0(\lambda)\left(1 + r_{21}r_{23}\exp\left[\frac{4\pi i n_2 d\cos(\theta_2)}{\lambda}\right] + \cdots\right)$$
$$= \frac{E_0(\lambda)}{1 - r_{21}r_{23}\exp[4\pi i n_2 d\cos(\theta_2)/\lambda]}. \tag{37.2}$$

Here, r_{21} and r_{23} are the reflection coefficients at the respective boundaries, which depend upon the propagation angle θ_2. The wavelength is λ and $E_0(\lambda)$ is an effective emission amplitude of a nanocrystal to the waveguide mode. Considering the guided modes and thus $|r_{21}| = |r_{23}| = 1$, the electric field intensity reveals sharp resonances implying there are well-defined discrete guided modes. The energy carried by each mode is finite, indeed. Unlike the guided modes, the electric field intensity is finite for the substrate modes and there are no sharp resonances in the spectra. We may, however, resolve some weak resonances, depending on the phase factor of the term $r_{21}r_{23}\exp[4\pi i n_2 d\cos(\theta_2)/\lambda]$: the electric field is maximum if this term is positive and real.

It is clear from the ray optics that there are substrate modes that refract to the substrate at the angles near $\pi/2$, which means that they propagate nearly parallel to the core/substrate boundary; these modes may be understood as a crossover between the guided modes and the substrate radiation modes which decouple rapidly from the core, possessing mixed characteristics of the both types of modes. As the modes may be assigned to a distinct guided mode of the nth order (energy of the substrate mode is slightly below the cutoff energy of the nth guided mode), we denote each series of the substrate modes to be of the nth order (see Figure 37.14). The Fresnel formulae give us $|r_{21}| = 1$ and $|r_{23}| \approx 1$ in such a case (Figure 37.15), and it is then clear from Equation 37.2 that these substrate modes reveal sharp resonances in the spectra (for a fixed angle $\theta_2 < \arcsin(n_2/n_3)$). Energy carried by the substrate modes may be therefore comparable to the energy carried by guided modes. The most important point here is that this type of substrate modes cannot be experimentally distinguished from the guided modes at short distances from the source as they propagate near the core/substrate boundary. The substrate modes therefore significantly contribute to the system response and they may play an important role in many experiments as well as in real on-a-chip devices.

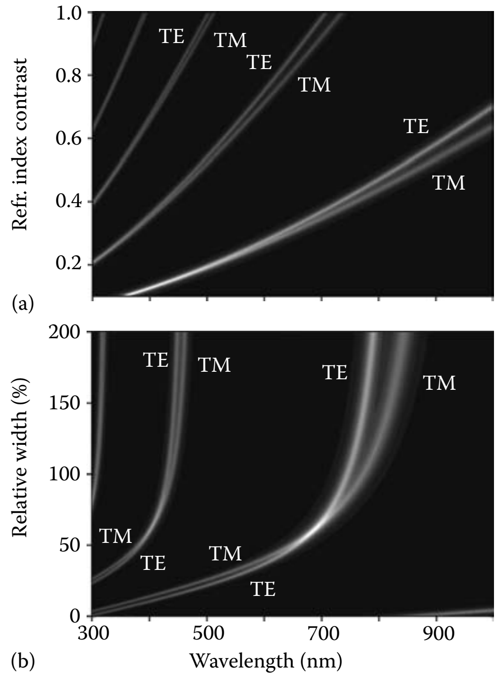

FIGURE 37.14 **(See color insert following page 20-14.)** Calculated spectral positions of the substrate modes as a function of (a) the refractive index contrast $\Delta n = n_2 - n_3$ and (b) the relative thickness of the waveguide core compared to the sample 5×10^{17} Si cm^{-2}. Gray scale indicates intensity from black up to white for the highest intensity. Several orders of modes are seen starting from the first one in infrared region.

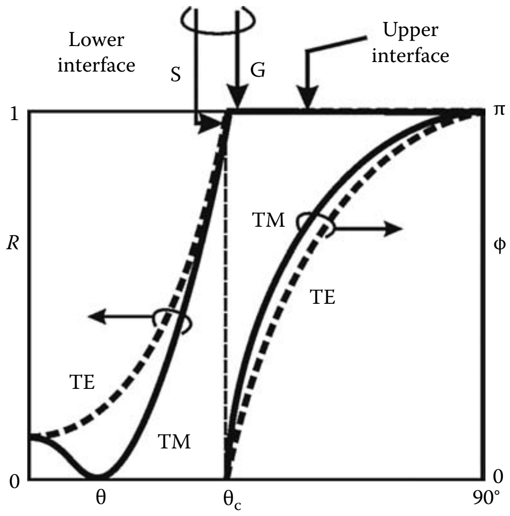

FIGURE 37.15 Reflectance $R = |r_{ij}|^2$ and phase shifts ϕ on the planar boundary between two dielectric media plotted for TE and TM modes versus incident angle θ. S and G stand for substrate and guided modes, respectively. (Adapted from Pelant, I. et al., *Appl. Phys. B*, 83, 87, 2006.)

Although the guided and the substrate modes propagate in a similar direction, they may behave in very different ways because they propagate in two different environments. In our particular case, the guided modes are mostly localized inside the waveguide core doped by Si-nc, providing a possibility of reabsorption or optical amplification. The substrate modes, on the contrary, propagate in a transparent substrate and they are therefore neither amplified nor attenuated. Rigorous derivation of PL spectra in real waveguides accounting for losses or amplification, decoupling of the light from the waveguide and real detection geometry is presented elsewhere (Ostatnický et al. 2008); here, we summarize only the main results.

Considering lossy waveguide core, the substrate modes become dominant in the optical response of a waveguide when detected in the direction parallel to the core/substrate boundary using a

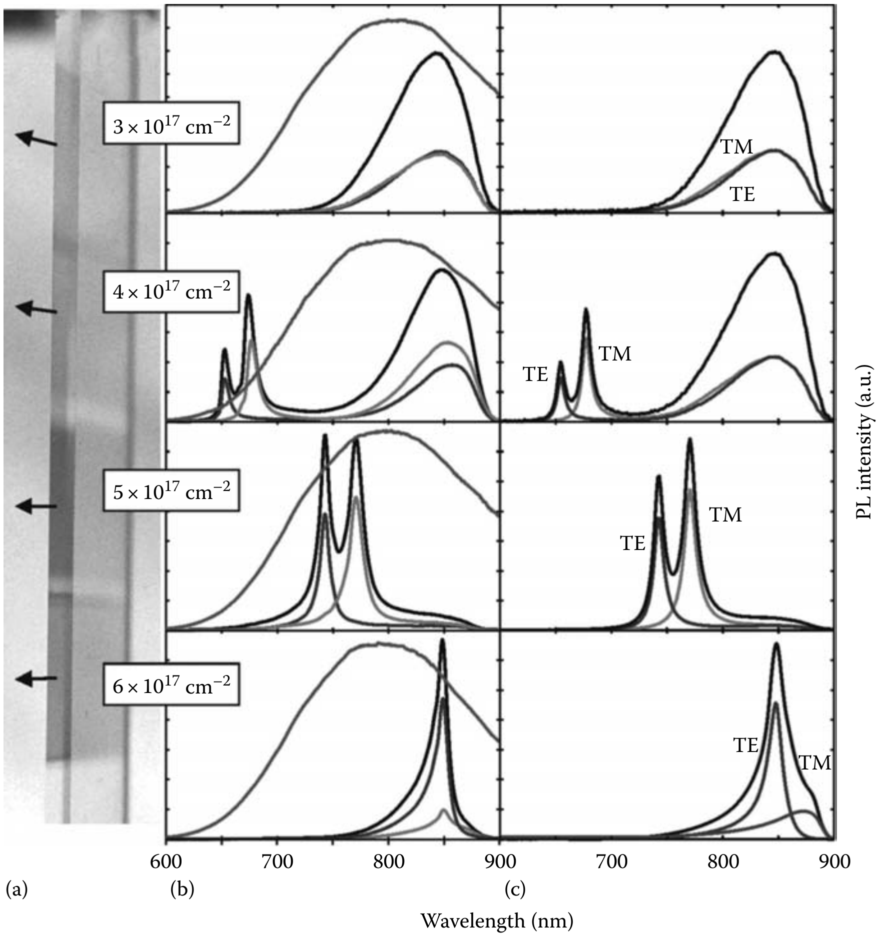

FIGURE 37.16 (See color insert following page 20-14.) (a) Photograph of the edge of a set of Si^+ ion implanted layers with direction of PL indicated by arrows, the edge is on the left. (b) Measured PL from samples implanted to different Si ion fluences in standard (the broadest curves) and waveguiding geometry (black lines, the slightly lighter gray lines stand for TE and TM resolved polarizations). (c) Theoretically calculated PL spectra. We note that these results were obtained on different set of samples than in Figure 37.10. The mode positions are not exactly the same for samples with identical implantation dose because the annealing conditions were slightly different. Therefore, the refractive index profiles are not identical. (Adapted from Pelant, I. et al., *Appl. Phys. B*, 83, 87, 2006.)

detector with a small numerical aperture. This situation is clearly illustrated through the comparison of our experimental data with the theoretical estimates in Figure 37.16 (we considered a continuous refractive index profile in our calculations). The broad part of the emission spectra originates from the (attenuated) guided modes and the narrow peaks are due to the substrate modes. The double-peak structure is further resolved as two single peaks with the respective TE and TM polarizations. In order to explain this point, we should remark that the coefficient r_{23} is real for the substrate modes while the coefficient r_{21} is in general complex. Considering a fixed angle of incidence θ_2, r_{21} gives different values of the phase shift for the respective TE and TM polarizations (see Figures 37.14 and 37.15). Therefore, the resonance condition is met at different wavelengths for different polarizations. From these considerations, the necessity of waveguide asymmetry for the appearance of the TE–TM polarization–resolved modes follows.

The TE–TM splitting was also reported in Khriachtchev et al. (2004), but the respective peaks have an opposite order in the spectra compared to our experiments. Our numerical calculations have shown that the TE–TM splitting depends strongly on the refractive index profile of the waveguide core. For the three-layer structures, the TE mode is positioned always at the high-energy side of the PL spectra. It is, nevertheless, possible to fabricate a structure with multiple layers or with a graded refractive index profile, which provides an arbitrary TE–TM splitting (including both negative and positive values and also their degeneracy); this feature is well illustrated in Figure 37.14b, where the respective TE and TM modes interchange depending on the thickness of the guiding layer.

The appearance of the substrate modes is well controllable by the parameters of the particular layers of the waveguide. Clearly, the optical thickness of the core d determines the resonance condition through the exponent in the denominator in Equation 37.2 and the ratio n_2/n_1 has influence on the phase of the r_{21} term. The latter may be used in detectors integrated on a chip—the cladding can be made of a material that changes its refractive index due to the changes of the surrounding environment conditions, and the position of the emission peak may then

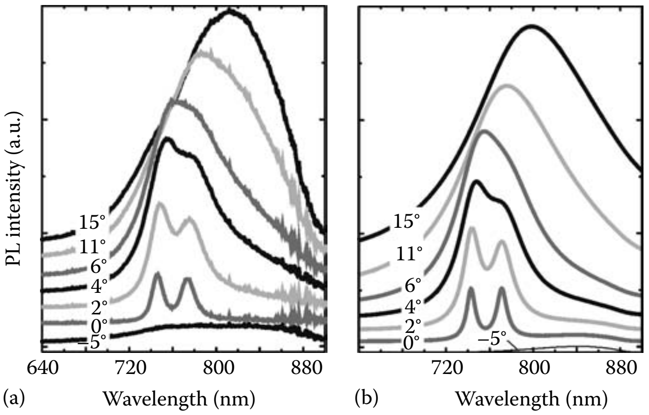

FIGURE 37.17 (a) Measured PL spectra at different detection angles α relative to the waveguide axis. (b) Numerical simulation of the measurements from the panel a. The sample with an implantation fluence of 5×10^{17} cm^{-2}. (Adapted from Ostatnický, T. et al., Guiding and amplification of light due to silicon nanocrystals embedded in waveguides, in Khriachtchev, L. (ed.), *Silicon Nanophotonics*, World Scientific, Singapore, 2008, pp. 267–296.)

be correlated with some external variable (temperature, pH, etc., see Figure 37.20) (Luterová et al. 2006).

As the sharp resonance in the spectra of the substrate modes is restricted only to the region where $r_{23} \approx 1$, we may expect disappearance of the substrate modes at large detection angles, i.e., when the substrate modes do not propagate parallel to the core/substrate boundary. This feature is illustrated in Figure 37.17 in comparison to the experimental and the theoretical data. We see a very good agreement justifying the correctness of our model.

An important aspect of nanodevices is the magnitude of the decoupling length for the substrate modes, i.e., the distance at which the substrate modes completely leak out from the waveguide core. Our calculations (Ostatnický et al. 2008) show that this distance may be few micrometers but also it may be as large as several hundreds of micrometers. As a consequence, the appearance of the substrate modes and their decoupling from the waveguide core may be of a big importance in nanodevices on the contrary to the fiber optics where the optical response to any excitation is provided solely by the guided modes at the typical distances of the order from meters to kilometers.

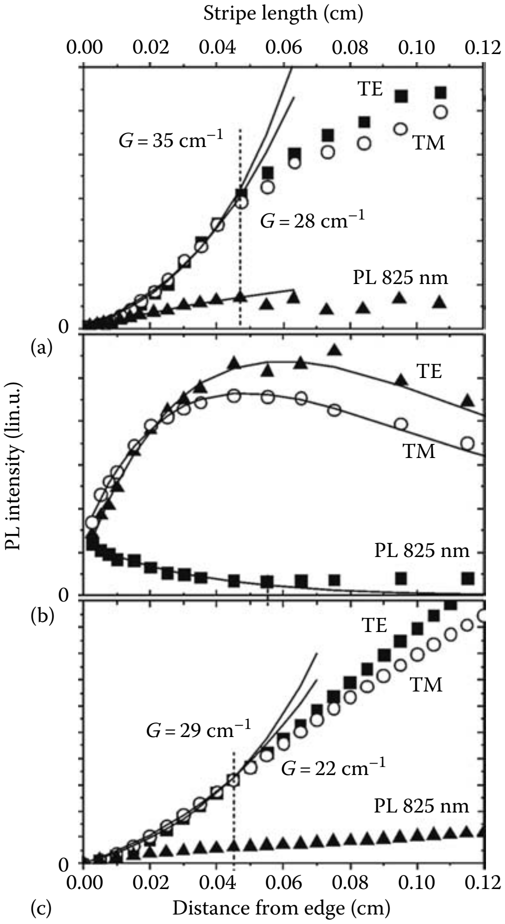

FIGURE 37.18 VSL and SES measurements on a sample implanted to the dose of 4×10^{17} cm^{-2} under continuous wave excitation 325 nm, 0.26 W/cm^2: (a) VSL measurement at the peak of TE and TM modes and for non-guided PL around 825 nm. The fits (lines) give values of $G = 35$ and 28 cm^{-1} for TE and TM modes, respectively, and losses of 11 cm^{-1} for non-guided PL. (b) Results of the SES measurement performed under identical conditions as the VSL. (c) Integration of data from panel (b). The gain fits (lines) give values of $G = 29$ and 22 cm^{-1} for TE and TM modes, respectively. (Adapted from Valenta, J. et al., *Appl. Phys. Lett.*, 81, 1396, 2002.)

37.6 Application of Active Nanocrystalline Waveguides

37.6.1 Amplification of Light

The significant narrowing of substrate leaky mode emission suggests immediately that optical amplification could be responsible for this observation. The VSL method (described in Section 37.3.3) has been applied to the planar asymmetric Si-nc waveguides in several laboratories (Khriachtchev et al. 2001, Ivanda et al. 2003, Luterová et al. 2005). It is tempting to interpret the frequently observed initial weak exponential growth of the VSL curve (Figure 37.18a) as a manifestation of the occurrence of optical gain. High output directionality of the substrate mode emission apparently supports such interpretation. However, it has turned out that, as shown above, it is very difficult if not impossible to evaluate correctly the optical gain magnitude of the substrate modes because the nonlinear growth can originate in the mode leaking itself. Figure 37.18 demonstrates that the evaluation of the VSL experiments can yield a false optical gain if proper comparison between the VSL and SES results is not taken into account. On the other hand, VSL measurements employing a high pulsed laser excitation, performed on a similar sample, revealed a characteristic switch from the light attenuation ($G < 0$) to amplification ($G > 0$) with increasing excitation energy density (Figure 37.19). It may be also of interest here to call the reader's attention to recent articles reporting firmly positive optical gain on leaky

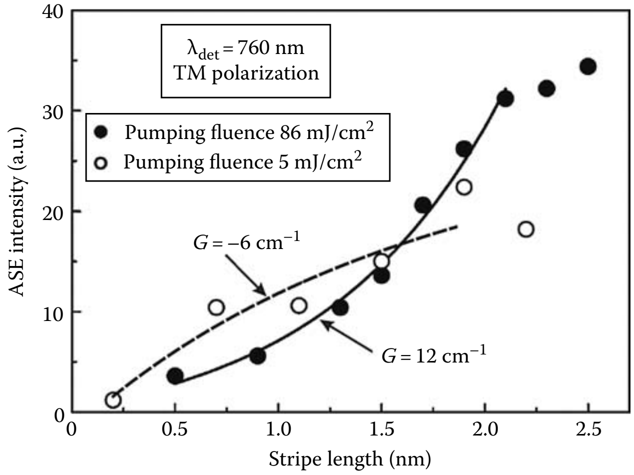

FIGURE 37.19 Time-resolved VSL measurement of optical amplification at position of the TM mode (760 nm) under pumping by 6 ns pulses, 355 nm. The fit with Equation 37.1 gives the net gain coefficient of (−6 ± 6) and (12 ± 2) cm^{-1} for pumping fluence of 5 and 86 mJ/cm^2, respectively. The threshold for positive gain is about 50 mJ/cm^2. Sample was implanted to a dose of 4×10^{17} cm^{-2}. (After Luterova, K. et al., *Phys. Status Solidi (c)*, 2, 3429, 2005.)

substrate modes in asymmetric thin-film organic waveguides (Nakanotani et al. 2007, Yokoyama et al. 2008).

At present, checking the reliability of the VSL method in the case of active asymmetric thin waveguides represents an issue requiring further investigation.

37.6.2 Other Applications

Although the benefit of the waveguide spectral filtering for nanocrystalline thin-film laser design is still controversial, the substrate modes may be attractive for other nanophotonic applications. Firstly, they provide a way how to *generate easily the spectrally narrow, polarization resolved and directional emission* in the wavelength range 650–950 nm without the necessity of building optical (micro)cavities. Moreover, this emission can be spectrally tuned simply by engineering the silicon excess content, as shown in Figures 37.10 and 37.16. Potential feasibility of these properties for *demultiplexing optical signal in photonic circuits* is evident, but improvement in quality of the rib waveguides is anticipated.

Secondly, spectral sensitivity of the radiative substrate modes to surrounding (organic) compounds can be utilized in *photonic sensing* (Figure 37.20): Magnitude of the refractive index

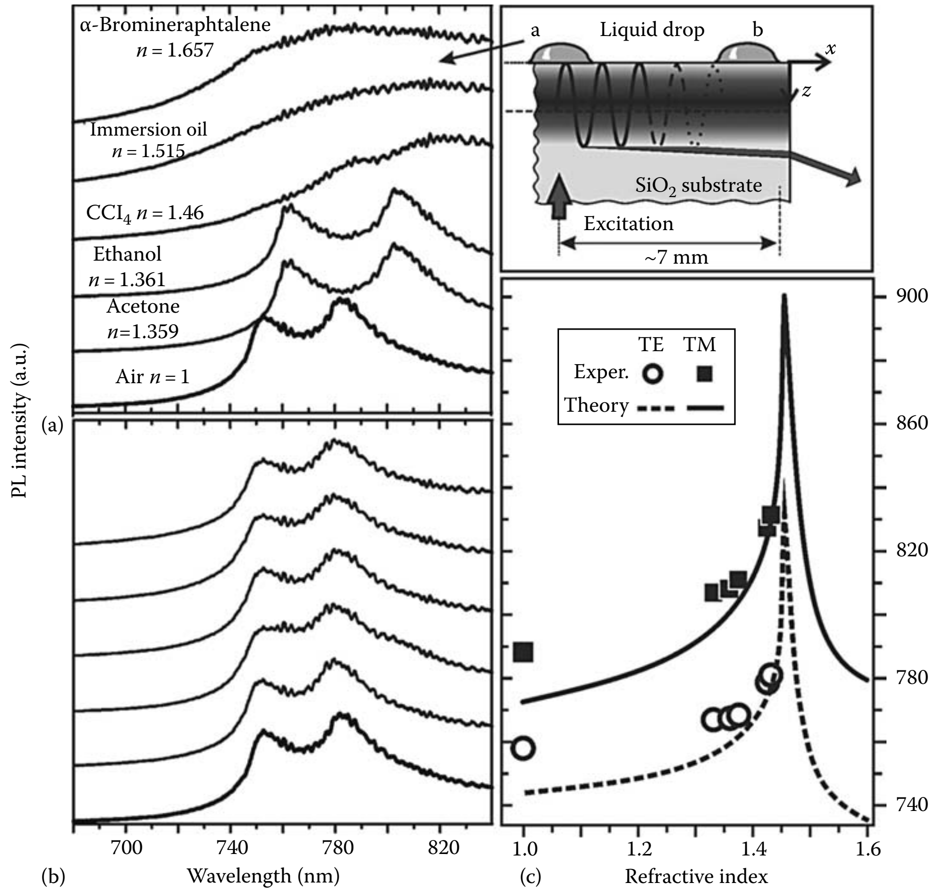

FIGURE 37.20 The influence of a drop of liquid on the PL spectra: (a) the drop is on the excited spot and (b) the drop is out of excited spot (between the spot and the sample edge). (c) Comparison of experimental (points) and calculated (line) positions of PL maxima for different refractive indices of liquid drops. Sample is implanted to a dose of 5×10^{17} cm^{-2}. (Adapted from Pelant, I. et al., *Appl. Phys. B*, 83, 87, 2006; Luterová, K. et al., *J. Appl. Phys.*, 100, 074307, 2006.)

n_1 of the compound (playing role of the cladding layer) affects markedly the spectral position of the TE–TM doublet. When n_1 exceeds the magnitude of the core refractive index ($n_2 \approx 1.45$), the condition of total reflection $|r_{21}| = 1$ is canceled and the TE–TM doublet completely disappears. A more systematic research in this direction is, however, missing.

Finally, the *active waveguides* give the chance to circumvent the *coupling problem* connected with discrete photonic elements, i.e., the problem of how to inject efficiently light from an external source in a waveguide (Orobtchouk 2006). The *active waveguides eliminate the need of any optical couplers* and enable it to design integrated photonic circuits. Theoretical analysis of active Si-nc devices integrating optical emission and waveguiding, compatible with silicon VLSI processing technology, have been submitted quite recently (Milgram et al. 2007, Redding et al. 2008).

37.7 Conclusions

Active waveguides formed by densely packed silicon nanocrystals represent a promising type of nanophotonic waveguides. In contrast to other types of nanowaveguides (based on photonic waveguides, plasmonic structures, etc.), it can be prepared by various technological approaches available and well mastered in many research and industrial laboratories. Its potential applications range from optical amplifiers and filters to optical sensors. The descriptions of characterization techniques and theory of waveguide modes in this chapter are applicable to many other waveguide structures.

Acknowledgments

This work was supported by the Czech Ministry of Education, Youth and Sports through the research center LC510 and research plans MSM0021620835 and 60840770022, the project 202/07/0818 of the Grant Agency of the Czech Republic, and the projects IAA101120804 and KAN401220801 of the Grant Agency of the Academy of Sciences. Research carried out in the Institute of Physics was supported by the Institutional Research Plan AV0Z10100521. The authors thank Prof. R.G. Elliman (Australian National University, Canberra) and Ing. V. Jurka for the fabrication of waveguides.

References

Cheylan, S. and Elliman, R. G. 2001. Effect of hydrogen on the photoluminescence of Si nanocrystals embedded in a SiO_2 matrix. *Applied Physics Letters* 78: 1225–1227.

Cheylan, S., Langford, N., and Elliman, R. G. 2000. The effect of ion-irradiation and annealing on the luminescence of Si nanocrystals in SiO_2. *Nuclear Instruments and Methods in Physical Research: Section B* 166–167: 851–856.

Dal Negro, L., Bettotti, P., Cazzanelli, M., Pacifici, D., and Pavesi, L. 2004. Applicability conditions and experimental analysis of the variable stripe length method for gain measurements. *Optics Communications* 229: 337–348.

Iacona, F., Franzò, G., and Spinella, C. 2000. Correlation between luminescence and structural properties of Si nanocrystals. *Journal of Applied Physics* 87: 1295–1303.

Imakita, K., Fujii, M., Yamaguchi, Y., and Hayashi, S. 2005. Interaction between Er ions and shallow impurities in Si nanocrystals. *Physical Review B* 71: 115440-1–115440-7.

Ivanda, M., Desnica, U. V., White, C. W., and Kiefer, W. 2003. Experimental observation of optical amplification in silicon nanocrystals. In L. Pavesi, S. Gaponenko, and L. Dal Negro (Eds.), *Towards the First Silicon Laser*, NATO Science Series, II: Mathematics, Physics and Chemistry, Vol. 93, pp. 191–96. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Janda, P., Valenta, J., Ostatnický, T., Skopalová, E., Pelant, I., Elliman, R. G., and Tomasiunas, R. 2006. Nanocrystalline silicon waveguides for nanophotonics. *Journal of Luminescence* 121: 267–273.

Khriachtchev, L., Räsänen, M., Novikov, S., and Sinkkonen, J. 2001. Optical gain in Si/SiO_2 lattice: Experimental evidence with nanosecond pulses. *Applied Physics Letters* 79: 1249–1251.

Khriachtchev, L., Novikov, S., and Lahtinen, J. 2002. Thermal annealing of Si/SiO_2 materials: Modification of structural and photoluminescence emission properties. *Journal of Applied Physics* 92: 5856–5862.

Khriachtchev, L., Räsänen, M., Novikov, S., and Lahtinen, J. 2004. Tunable wavelength-selective waveguiding of photoluminescence in Si-rich silica wedges. *Journal of Applied Physics* 95: 7592–7601.

Khriachtchev, L., Navarro-Urios, D., Pavesi, L., Oton, C. J., Capuj, N. E., and Novikov, S. 2007. Spectroscopy of silica layers containing Si nanocrystals: Experimental evidence of optical birefringence. *Journal of Applied Physics* 101: 044310-1–044310-6.

Luterová, K., Dohnalová, K., Švrček, V., Pelant, I., Likforman, J.-P., Crégut, O., Gilliot, P., and Hönerlage, B. 2004. Optical gain in porous silicon grains embedded in sol-gel derived SiO_2 matrix under femtosecond excitation. *Applied Physics Letter* 84: 3280–3282.

Luterová, K., Navarro, D., Cazzanelli, M., Ostatnicky, T., Valenta, J., Cheylan, S., Pelant, I., and Pavesi, L. 2005. Stimulated emission in the active planar optical waveguide made of silicon nanocrystals. *Physica Status Solidi (c)* 2: 3429–3434.

Luterová, K., Skopalová, E., Pelant, I., Rejman, M., Ostatnický, T., and Valenta, J. 2006. Active planar optical waveguides with silicon nanocrystals: Leaky modes under different ambient conditions. *Journal of Applied Physics* 100: 074307.

Milgram, J. N., Wojcik, P., Mascher, P., and Knights, A. P. 2007. Optically pumped Si nanocrystal emitter integrated with low loss silicon nitride waveguides. *Optics Express* 15: 14679–14788.

Nakanotani, H., Adachi, C., Watanabe, S., and Katoh, R. 2007. Spectrally narrow emission from organic films under continuous-wave excitation. *Applied Physics Letters* 90: 231109-1–231109-3.

Orobtchouk, R. 2006. On chip optical waveguide interconnect: The problem of the in/out coupling. In L. Pavesi and G. Guillot (Eds.), *Optical Interconnects. The Silicon Approach*, pp. 263–290. Berlin, Germany: Springer-Verlag.

Ostatnický, T., Rejman, M., Valenta, J., Herynková, K., and Pelant, I. 2008. Guiding and amplification of light due to silicon nanocrystals embedded in waveguides. In L. Khriachtchev (Ed.), *Silicon Nanophotonics*, pp. 267–296. Singapore: World Scientific.

Park, J. H. and Steckl, A. J. 2004. Laser action in Eu-doped GaN thin-film cavity at room temperature. *Applied Physics Letters* 85: 4588–4590.

Pavesi, L. and Guillot, G. (Eds.). 2006. *Optical Interconnects. The Silicon Approach*. Berlin, Germany: Springer-Verlag.

Pavesi, L. and Lockwood, D. J. (Eds.). 2004. *Silicon Photonics*. Berlin, Germany: Springer-Verlag.

Pelant, I., Ostatnický, T., Valenta, J., Luterova, K., Skopalova, E., Mates, T., and Elliman, R. G. 2006. Waveguides cores containing silicon nanocrystals as active spectral filters for silicon-based photonics. *Applied Physics B* 83: 87–91.

Pellegrino, P., Garrido, B., Garcia, C., Arbiol, J., Morante, J. R., Melchiorri, M., Daldosso, N., Pavesi, L., Schedi, E., and Sarrabayrouse, G. 2005. Low-loss rib waveguides containing Si nanocrystals embedded in SiO_2. *Journal of Applied Physics* 97: 074312-1–074312-8.

Redding, B., Shi, S., Creazzo, T., and Prather, D. W. 2008. Electromagnetic modeling of active silicon nanocrystal waveguides. *Optics Express* 16: 8792–8799.

Riboli, F., Navarro-Urios, D., Chiasera A., Daldosso, N., Pavesi, L., Oton, C. J., Heitmann, J., Yi, L. X., Scholz, R., and Zacharias, M. 2004. Birefringence in optical waveguides made by silicon nanocrystal superlattices. *Applied Physics Letter* 85: 1268–1270.

Shaklee, K. L. and Leheny, R. F. 1971. Direct determination of optical gain in semiconductor crystals. *Applied Physics Letters* 18: 475–477.

Skopalová, E. 2007. Mode structure in the light emission from planar waveguides with silicon nanocrystals. Diploma thesis. Prague, Czech Republic: Charles University.

Snyder, A. W. and Love, J. D. 1983. *Optical Waveguide Theory*. London, U.K.: Chapman & Hall.

Valenta, J., Pelant, I., and Linnros, J. 2002. Waveguiding effects in the measurement of optical gain in a layer of Si nanocrystals. *Applied Physics Letters* 81: 1396–1398.

Valenta, J., Pelant, I., Luterová, K., Tomasiunas, R., Cheylan, S., Elliman, R., Linnros, J., and Honerlage, B. 2003a. Active planar optical waveguide made from luminescent silicon nanocrystals. *Applied Physics Letters* 82: 955–957.

Valenta, J., Luterová, K., Tomašiunas, R., Dohnalová, K., Hőnerlage, B., and Pelant, I. 2003b. Optical gain measurements with variable stripe length technique. In L. Pavesi, S. Gaponenko, and L. Dal Negro (Eds.), *Towards the First Silicon Laser*, NATO Science Series, II: Mathematics, Physics and Chemistry, Vol. 93, pp. 223–242. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Valenta, J., Ostatnický, T., Pelant, I., Elliman, R. G., Linnros, J., and Hőnerlage, B. 2004. Microcavity-like leaky mode emission from a planar optical waveguide made of luminescent silicon nanocrystals. *Journal of Applied Physics* 96: 5222–5225.

Yokoyama, D., Moriwake, M., and Adachi, C. 2008. Spectrally narrow emissions from edges of optically and electrically pumped anisotropic organic films. *Journal of Applied Physics* 103: 123104-1–123104-13.

38

Biomolecular Neuronet Devices

38.1 Introduction 38-1
Bacteriorhodopsin Is an Advanced Material for Molecular Nanophotonics
38.2 Some Techniques of Neuro-Molecular and Molecular Information Processing Using Bacteriorhodopsin (Basic Processes, Constructions, Technology) 38-2
Basic Process of Classic Optical Neural Network in Bacteriorhodopsin Medium • Neuronets Based on Multilayered Optical Structures Including Polymeric Bacteriorhodopsin-Containing Layers
38.3 Nanostructuring of Bacteriorhodopsin-Containing Molecular Media....................38-10
Procedure of Complex Estimation of Functional Characteristics of Bacteriorhodopsin-Containing Materials • Functional Characteristics of Nanostructured Bacteriorhodopsin Films • Metal Nanoparticles and Bacteriorhodopsin-Based Hybrid Nanostructures
38.4 Summary....................38-17
References....................38-17

Grigory E. Adamov
Central Scientific Research Institute of Technology "Technomash"

Evgeny P. Grebennikov
Central Scientific Research Institute of Technology "Technomash"

38.1 Introduction

Due to the constant development of nanotechnologies and functional nanomaterials used in information systems, researchers need to look for new ideas and radically new constructive solutions to ensure efficient device operation, record-breaking high speed, and integration of elements. On the other hand, functional elements being formed in nanometer range require landmark approaches. It refers especially to the formation of information-logical devices at the molecular level. That is, the formation of nanodevices by direct impact on separate molecules or atoms is not efficient in the long run. The total time taken for operations with separate molecules leads to tremendous amount of time taken to form the device and huge costs.

Another techno-systematic problem is the necessity to use "macro–nano" and "nano–macro" interfaces to put in the original information and deduce final results. The advantages of the small-size molecular elements cannot help to ensure access to them. Probably the only effective way to arrange the bonding with molecular system is to use optical interaction.

Under the circumstances, it seems quite reasonable to refer to the most competent specialist, whose experience in creating molecular systems cannot raise any doubts, that is, to the nature itself and try to apply bionic principles to engineering and design. The solution seems to be connected with hierarchic structural and functional self-organization of molecular systems that use assemblies of molecules as a basis for molecular information-logical appliances.

Among all the information systems, bionic approach seems to be the most effective, harmonic, and holistic one; it is important to point out molecular neural network appliances. Both adaptive self-organization principles of data processing and self-organization principles of functional nanostructures can be applied to those systems.

Below you will find a detailed description of techno-systematic approaches aimed at implementation of biomolecular neural network appliances. Mentioned below, biological material—protein "bacteriorhodopsin"—has unique technological potential and its optical properties allow the use of optical input–output devices and create molecular appliances based on self-organization principles.

38.1.1 Bacteriorhodopsin Is an Advanced Material for Molecular Nanophotonics

Bacteriorhodopsin (BR)—light-sensitive protein—is similar to the optic rhodopsin of the human eye. BR is obtained from halobacteria containing BR in cellular membranes (so-called purple membranes). During the separation from the bacterial cells, purple membranes save the entire structure (Vsevolodov, 1988; Oesterhelt et al., 1991). The typical size of purple membranes is 500–1000 nm. It is the unique biocrystalline structure capable of saving its permanent properties for a number of years, which consists of dry and polymer films with the thickness from 5 nm (monolayer) up to a few tens of micrometers.

The fundamental BR property is the photochemical cycle availability: after the light quantum absorption, BR molecule passes through the sequence of states and spontaneously returns to the primary form (Figure 38.1). At that point, in compliance with the cycling of BR molecule state, the light-induced cycling

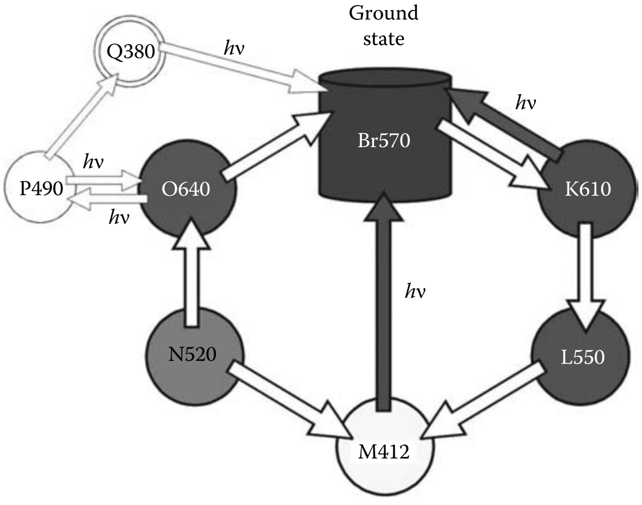

FIGURE 38.1 BR photocycle.

of optical characteristics (refraction and absorption indices) occurs (Haronian and Lewis, 1991; Zeisel and Hampp, 1992). Each of the interstitial states is identified as the intermediate according to its absorption spectrum. The existence of branched photocycles is typical for some BR types (Birge et al., 1999).

The main BR function in purple membranes is light-dependent proton transfer (H^+) over the purple membrane that results in electrochemical hydrogen potential formation on halobacteria membrane. The potential energy is utilized by cell. Ejection of H^+ occurs outside the cell membrane, and the H^+ is captured inside the cell (from cytoplasm). Supposedly, it happens during the formation and the disappearance of intermediate M412 (Siebert et al., 1982; Haronian and Lewis, 1991).

BR spectral sensitivity lies in the optical band (Figure 38.2). Absorption maximum in primary state of BR570 corresponds to wave length 570 nm. M412 in the main intermediate state reaches its absorption maximum at wave length 412 nm (Birge et al., 1999). The absorption of optical emission by BR-containing medium is characterized by certain peculiarities. That happens due to the change of adsorption sites concentration (molecules in form BR570), which is the result of light quantum absorption at wave length 570 nm by these sites and transition to form M412 with a low absorption at wave length 570 nm. As a result, absorption in the yellow range reduces; medium becomes more transparent—bleached. The intensity of BR-containing medium bleach effect depends particularly on the time of intermediate M412 molecules relaxation to form BR570. Relaxation time is characterized by half-value period of intermediate M412 molecules. The light effect at wave length 412 nm results in fast coercive transition of molecules to the primary state. The values of photocycle time parameters lie in the range from microseconds up to tens of seconds.

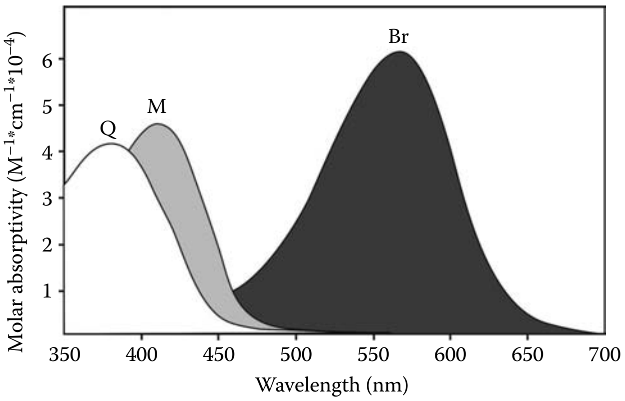

FIGURE 38.2 Absorption spectra of BR and of the main photocycle intermediates.

Thus, BR behaves as a photochromic medium with a quick time of information storage. Optical and dynamic BR characteristics change in wide range by production conditions and matrix composition (medium).

38.2 Some Techniques of Neuro-Molecular and Molecular Information Processing Using Bacteriorhodopsin (Basic Processes, Constructions, Technology)

38.2.1 Basic Process of Classic Optical Neural Network in Bacteriorhodopsin Medium

BR properties listed above allow us to illustrate one of the available basic processes of data transformation in BR-containing media. It is based on reversible light-sensitive changes of absorption index and appears in optical effects considered below.

38.2.1.1 Nonlinear Absorption of Optical Radiation: Medium Bleaching

The absorption of optical radiation in substance is described by some known classic equations. Generalized Bouguer–Beer law associates intensities of the incident light and the light transmitted through the substance layer with the thickness of the layer and molecular concentration of absorption agent:

$$I = I_0 \cdot e^{-D} = I_0 \cdot e^{-\alpha d} = I_0 \cdot e^{-\varepsilon c d}, \qquad (38.1)$$

where

I is the transmitted light intensity
I_0 is the incident light intensity
D is the optical density of the substance
α is the absorption index of the substance
d is the thickness of the substance layer
c is the molar concentration of absorbing substance molecules
ε is the extinction coefficient, characteristic feature of absorbing substance molecule

The values of ε, α, and D depend on the wave length of incident light. When the values of coefficients in Equation 38.1 are invariable, the transmitted light intensity is in direct proportion to the incident light intensity.

Nonlinear absorption of optical radiation by BR-containing media is connected to the change of absorption centers concentration (molecules in form BR570) as a result of light quanta absorption by these centers at wave length 570 nm and transformation to M412 with low absorption at wave length 570 nm. Finally, as was mentioned above, absorption in the yellow range ($\lambda = 570$ nm) decreases, medium becomes more transparent—bleached.

38.2.1.2 Indirect Interaction of Optical Radiation Fluxes in Bacteriorhodopsin-Containing Media

Mediated interaction of optical radiation fluxes appears during the sequential or combined transmission through the same part of BR-containing medium and is at its clearest for monochromatic radiation with wave length 412 and 570 nm.

As a result of interaction between BR and the radiation with wave length 570 nm during the transmission through the medium, the energy of the light flux is absorbed, and in BR-containing medium, the photo-induced allocation of the absorption index forms. The light-induced allocation of the absorption index variation corresponds to energy distribution along the surface of the transmitted light wave front (Figure 38.3A).

Non-modulated along the front surface, the light pulse with the wave length of 570 or 412 nm (as actuating or inhibiting signal) absorbs spatially and nonuniformly in compliance with the changed value of the absorption index (Figure 38.3B and C).

Thus, the energy distribution along the preceding pulse front surface, indirectly, over the BR-containing medium, modulates the energy distribution along the following pulse front surface.

As a result, BR-containing medium is capable of accumulating (summing up) effects signed "+" and "–," correspondingly, at wave lengths 570 and 412 nm. Predicating upon the indirect interaction of optical radiation in BR-containing media, the method of formal neuron creation in such media is available to offer. At that point, all the main functions can be realized by optical technique.

38.2.1.3 One of the Possible Methods of Formal Neuron Main Functions Realization in Bacteriorhodopsin Medium

Realization of the main operations of neuronet algorithms—weighing of input signal vector according to the matrix of weighing coefficients of synaptic bonds; composition of weighed values of input signals; realization of activation (threshold) function by optical method without optoelectronic buffering—permits to simplify the construction and the technological realization of multilayered optical neuronet, to increase the integration of neuro-like elements in device, and to solve the problem of areal density limitation inherent in microelectronic elements and electric connections.

38.2.1.4 Formation of the Neuro-Like Element

The neuro-like element is formed under the exposure of optical emission with the spectrum, corresponding to the absorption spectrum of BR molecules' initial state and BR-containing material medium. This element is a part of the BR-containing medium with photo-induced absorption index. The threshold properties of such neurons are defined by the concentration ratio of molecules in primary and photo-induced forms, and the interaction of neuro-like elements is provided by optical emission.

The example of the similar neuronet realization, based on information conversion of basic process in BR-containing media (Figure 38.4), is considered below.

Construction for optical neuronet formation includes the following:

1. The source of the plane light front (transparent for normal incident light flux) providing the signal formation and transfer to neurons.
2. Photo-detecting layer based on BR-containing material for imaging of the input optical information by photo-induced variation (according to light energy distribution along the surface of input light front) of absorption/transmission in BR-containing medium.

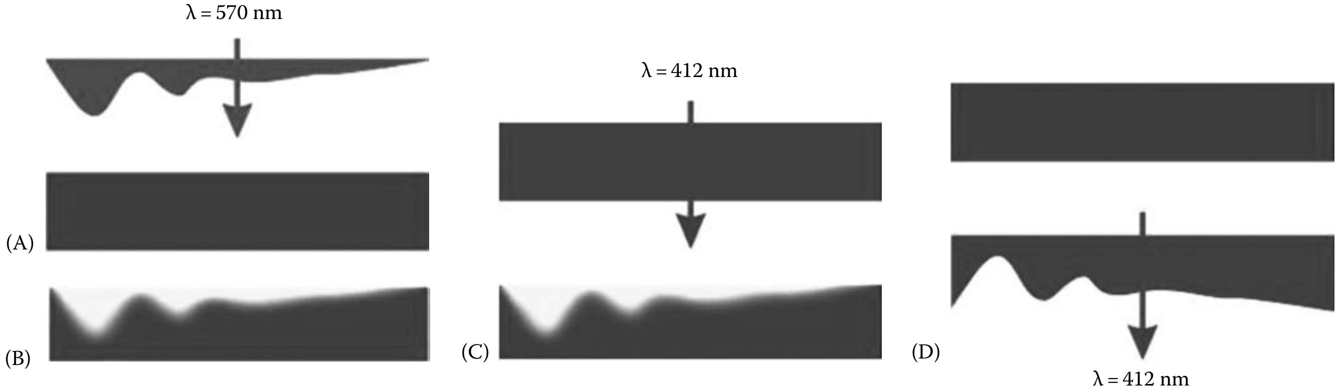

FIGURE 38.3 The indirect interaction of optical radiation fluxes in BR-containing media: (A) green light pulse (modulated along the front surface) acts on a layer of BR-containing medium (nontransparent for green light); (B) modulated allocation of molecules in form M412 (transparent for green light) and BR570 (nontransparent for green light), (C) the action of unmodulated blue light pulse, (D) modulated blue pulse as a result of transmission via the modulated medium (BR-containing medium layer recovered the primary state).

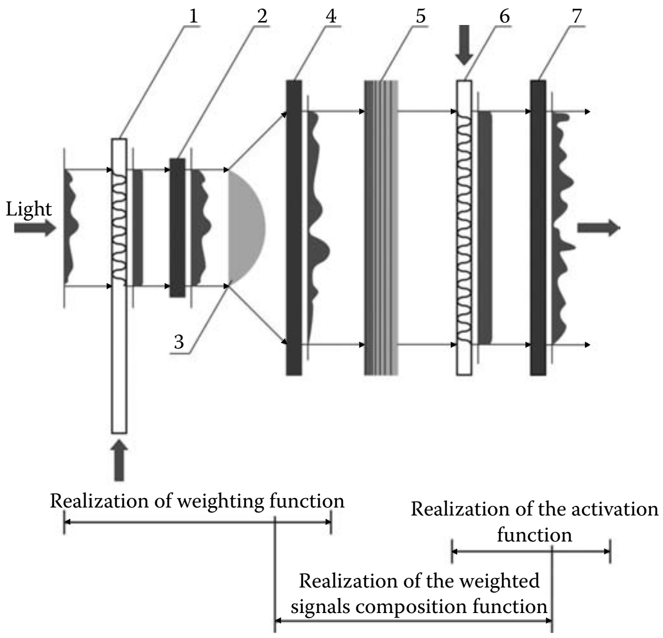
1
2
4
5
6
7
Light
3
Realization of weighting function
Realization of the activation function
Realization of the weighted signals composition function

FIGURE 38.4 Neuronet organization based on the basic process: 1, 6—plane waveguides, including gitters for emission input/output; 2, 4, 7—BR-containing layers (photo-detecting layer, layer of weighting coefficients and neuron layer); 3, 5—cylindrical lenses.

3. Light flux expanding lens.
4. Layer of synaptic bond synthesis (matrix of weighing coefficients) made of BR-containing material.
5. The lens, focusing the light flux on layer 7 and forming the combination of input optical signals based on neuro-like elements in that manner.
6. The source of flat light front (transparent for normal incident light flux) for parallel comparison with thresholds, formation, and transfer to other layers of output neuron signals in layer 7 as light front, modulated by intensity along front surface according to the values of absorption/transmission area of BR-containing medium (corresponding to neuro-like elements).
7. The layer of neuro-like optical elements (being obtained on the basis of BR-containing material) composes input optical signals, traversing matrix 4 and realizing the activation function during the light front traverse from the source 6, and forms thereby output signals.

The input optical information in the form of light flux (input vector) effects the photo-detecting layer 2 that results in the absorption of light flux energy in BR-containing medium of photo-detecting layer and the distribution of the photo-induced absorption index forms.

Distribution of the photo-induced BR-containing medium absorption index along the surface and in depth corresponds to the power distribution along the surface of the effective light front.

Plain light front from the source 1, modulated by intensity according to the contour of absorption index of photo-detecting layer 2 by lens 3, allocates on the surface of weighing coefficients layer 4.

38.2.1.5 Realization of Weighing Function

The input vector weighing function comes around during the transfer process of input light signal—the components of input vector—over the matrix of weighing coefficients 4. By the component of input vector, we mean the quantity of the light energy (the intensity multiplied by exposure—exposition) affecting the matrix section.

Weighing coefficient (by which the input vector component is multiplied) is a transmission coefficient of the corresponding section of BR-containing matrix:

$$I_{w\,ij} \cdot t = \omega_{ij} \cdot I_{in\,i} \cdot t, \tag{38.2}$$

where

$I_{w\,ij}$ is the light intensity transmitted over the ij matrix section or weighed component of the input vector
$I_{in\,i}$ is the light intensity of the light front section or a component of the input vector
t is the exposition time of the corresponding component of the input vector
ω_{ij} is the weighing coefficient or transmission of the corresponding ij-section of the BR-containing matrix

Transmission of the ij matrix section is defined according to Bouguer–Beer law

$$\omega_{ij} = e^{-\varepsilon d c_{i/f\,ij}}, \tag{38.3}$$

where

ε is the BR absorption factor
d is the thickness of BR-containing layer
$c_{i/f\,ij}$ is the concentration of the BR molecules in the initial state in the ij matrix section

The properties of the input vector weighed components are formed by lens 5 in a light flux that gets to the inputs of the corresponding neuro-like elements of the layer 7.

38.2.1.6 Realization of the Weighed Signals Composition Function

Composition function of the input signals is realized in the BR-containing medium in layer 7 by the converging cylindrical lens 5 as a result of the combined effect at the same area of BR-containing medium of the light energy exposition by the corresponding components of the input vector.

Every component of the input vector contributes to the formation of the molecules ensemble in photo-induced state in proportion to the intensity and exposure time:

$$\Delta c_{j\,p/i} = k \cdot \sum I_{w\,ij} \cdot t, \tag{38.4}$$

where

$\Delta c_{j\,p/i}$ is the concentration of BR molecules in photo-induced spectral state

k is the coefficient of proportionality depending on the concentration of BR molecules in the initial state, cross-section interaction, photo response of BR-molecules transition from the initial to the photo-induced form, and the effecting light wave length

$I_{w\,ij}$ is the intensity of the input vector weighed *ij*-component

t is the exposure time of the input vector *i*-component

Thus, $\Delta c_{jp/i}$ contains information about the value of the weighed input interactions sum on the *j*-neuron.

38.2.1.7 Realization of the Activation Function

Magnitude $\sum_i I_{w\,ij} \cdot t$ (the total dose of light energy effecting on the *j*-area of BR-containing medium) assigns the point at the graph (dependence of the transmission value on the sum of weighed input effects) and defines the transmission of the light signal over the *j*-neuro-like element.

The changed transmission $\omega_{n/e}$ value of the BR-containing medium *j*-section according to the *j*-neuro-like element assigns the value of the activation function and the output signal of the neuro-like element in layer 7, according to Figure 38.5. This magnitude depends on the number of molecules possessing the changed spectral properties of the medium section corresponding to the neuro-like element of the layer 7 according to

$$\omega_{n/e\,j} = e^{-\varepsilon d(c_{i/f} - \Delta c_{j\,p/i})}. \tag{38.5}$$

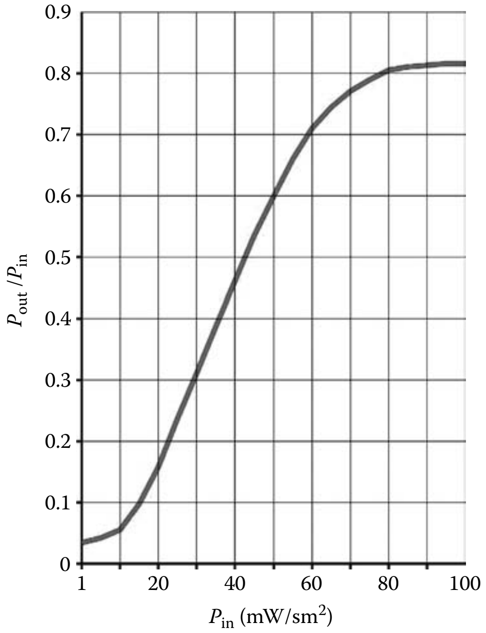

FIGURE 38.5 Transmission (P_{out}/P_{in}) of BR-containing layer depending on the effective emission energy (P_{in}).

The graphic chart on the dependence of the BR-containing layer relative transmission on the emission energy consists of the area with the initial transmission value (unequal to zero), the area of almost linear transmission change, and the saturation area (Figure 38.5). In general, the curve corresponds to the activation function proposed for the neuronet realization by Grossberg (Wasserman, 1989). The similar compressive function automatically provides the output signal range from 0 to 1 and corresponds to the necessary requirements for realization of reconversion algorithm during the neuronet learning, for example, according to scheme (Wasserman, 1989).

The output signal formation of neuro-like element (threshold comparison and realization of the activation function) is carried out by the light front of the specified intensity and duration induced by the source-former 6.

The output signal of the neuro-like *j*-element is formed as an energy portion of the active light signal being transferred over to the corresponding section of BR-containing medium in the layer of neuro-like elements 7 (according to the transmission of the section considered) in conformity with the formula

$$I_{\text{out}\,n/e} \cdot t_{\text{active}} = W_{n/e} \cdot I_{\text{active}} \cdot t_{\text{active}}. \tag{38.6}$$

The minimum value of the output signal is fixed by the transmission of non-firing neuron (the input signals sum value is close to zero) and corresponds to the initial transmission of photochromic medium, and the maximum value is close to the value of active front energy and corresponds to the saturation area of the curve (Figure 38.5) and to the maximum excited state (transmission) of firing neuron. Output signals of neuro-like elements form the continued light front being modulated according to the activation function at every point of BR-containing layer 7. System learning (formation of matrix weighing coefficients) corresponds to the formation of adequate values of transmission coefficients of matrix sections based on BR-containing medium that can be achieved by the inverse transformation method.

The optical version of the inverse transformation method can be simply and effectively realized (failing optoelectronic transformations) by combined presentation of learning pair: the input image in its usual direction and the ordered output as the light front in the counter direction. Due to the reversibility of the light passing, both of light fronts will affect the matrix made of BR-containing material and will change the matrix transmission corresponding to the intensity distribution.

38.2.2 Neuronets Based on Multilayered Optical Structures Including Polymeric Bacteriorhodopsin-Containing Layers

The considered approach of neuro-computer element base formation takes into account the cyclicity of processes in living systems appearing, for example, in spontaneous activity of pacemaking (assigning the rhythm of functioning) neurons. As was conclusively shown in Prigogine (1980), the cyclicity proceeds as a result of the processes' self-organization in open nonlinear nonequilibrium systems and the origination of stable dissipative structures due to which the coordination of trophic processes (providing cell nutrition) is possible in living systems.

The similarity of cyclic processes in the living cell membranes (including neurons) and the processes under optical emission exposure in isolated purple membranes is obvious, particularly, taking into consideration that in that case one of the halobacteria trophic cycles is reproduced.

If the neuron is represented as a structure population (purple membranes) allocated in media and the interactions between the neurons are carried out by light fluxes, the "neurolike medium" in question can be considered as the alternative to the net of neuro-like elements (Grebennikov, 1997).

In the interpretation being stated, the problem of neuronet formation is in simulation of nonequilibrium nonlinear systems in BR-containing media with allocated parameters by optical emission of two different wave lengths corresponding to the absorption maxima of two "long-living" intermediates. It can be solved in the following consequence:

1. Using the spectral sensitivity of BR in optical range to initiate the cyclic light-dependent processes by the light flux and to control the processes by the application of emission with various wave lengths.
2. Stationary and dynamic structuring of BR-containing films by modulated light fluxes to create the conditions of formation and development of neuro-like relations in neuro-like medium and to correlate the multitude of cyclic processes.

Data input and processing in that case is "deformation" of the correlated cyclic processes in BR-containing medium, and the commands of system behavior control are the transient processes originating in the organized neuro-like medium.

On the other hand, analyzing the development tendencies of neuronet technology, the best perspective can be emphasized, and their combination in the same device can provide essential expansion of neuronet data processing:

1. *Optical mechanism of data transmission and processing*, which will allow to construct three-dimensional nets that function simultaneously at high speed. The problem of wide application is in large dimensions, peculiar to optical systems, and the inevitable efficiency losses due to multiple intermediate optoelectronic conversions and the finite size of the optoelectronic elements.

The application of continuous photochromic media with high resolution close to molecular level of data representation and processing can become a solution to the problem. BR is the most available and well-investigated (at present) photochromic material with sufficiently high cyclicity ($>10^6$) and is suitable for the considered purposes. Using BR enables to carry out data processing in optical mode without intermediate optoelectronic conversions and to increase the areal density of neuro-like elements value comparable to the native neuron systems.

2. *Neuro-like structure construction on the principles of self-organization of the processes in the open nonlinear allocated dissipative systems like biological objects.* Although in model version, systems require substance transfer proceeding for nonequilibrium conditions maintenance.

BR application can solve the problem of substance transfer that complicates the construction and limits the continuous working life of neuro-like media processing. Light-dependent properties of that material allow the open nonlinear allocated dissipative systems to simulate effectively. Optical exposure in the range of 520–650 nm is like the input flux and dissipative properties, and in that case can be provided by the component of trophic cycle that remains invariable in composition of artificial medium (for example, in polymeric matrix) and also by emission exposure in blue light range ($\lambda = 400$–420 nm).

3. *Element base adaptability* to losses of some elements during the preparation and exploitation compensated by self-organizing and self-modification of neuro-like structures. Application of continuous media based on BR in conjunction with optical methods permits to form neuro-like elements, and the bonds between it suit requirements according to the light energy allocation.

Classic methods of the light fluxes formation by lens systems result in the loss of advantages expected from significant density of neuro-like elements and dimension restrictions of elements and systems, peculiar to optical computers. Moreover, when the volume of BR-containing medium is greater, the probable process integration is higher. However, the emission access to all the molecule groups is more unfavorable as the medium absorption increases.

In the field of optoelectronic technology, efforts are made to deflate the construction dimension of optical neuro-computers by the application of multilayered structures. Multilayered structures forming neuronets and containing the layer of spatial light modulators as the liquid crystal matrix, the layer of photoconducting material (Engel, 1990) or material with the photovoltaic effect (Akiyama et al., 1995), and the layer of electric transducer-amplifiers (forming and transducing the commands to liquid crystal panels) are proposed.

Essential disadvantage of the proposed constructions is the facility of intermediate conversion of light exposure to electric current or voltage used for the following changes of optical transmission modulator. Furthermore, neuronet function realization requires the application for these purposes of optoelectronic and microelectronic elements. Consequently, the application of metallic conductors in inter-cell links for realization of neuro-systems with a great number of neurons results in delays in communication lines and neuronet processing deceleration due to the capacity influence of inter-cell communication lines. The existing areal density restrictions of optoelectronic elements and electrical bonds inevitably limit the spatial resolution of constructions.

38.2.2.1 Multilayered Constructions, Including Bacteriorhodopsin-Containing Films

We proposed multilayered structures including layers based on the BR for the realization of neuronet medium in nonequilibrium nonlinear dynamically allocated dissipative systems.

Supposedly, the multilayered structures would allow the data processing to continue at the level of BR molecules groups by forming neuro-like elements using optical methods in BR-containing media.

The basic processes in BR medium are defined by light-dependent changes in absorption index allocation profile along the surface of BR-containing polymeric films. In multilayered structures, many light fluxes circulate without interaction. This property is usually proved as an advantage of optical methods enabling data processing and transmission in the three-dimensional space. At the same time, the information arrays in the form modulate indirectly by intensity light fronts over the reciprocal fluctuation of absorption index local value of BR-containing media sections, and the local intensity value of the light front sections realize concurrent information interactions in the three-dimensional space of the multilayered structure.

We would like to consider a possibility of neuro-like element net organization in BR medium by optical method using multilayered structures (Figure 38.6) including BR-based layers, wave guide layers, and reflecting layers. To allocate light fluxes, the system of waveguides, transparent in the optical range, is formed in BR-containing medium. It is possible to input the controlled emission in the form of light front in BR-containing medium, activating at that point the groups of neuro-like elements, by producing the sections with the disturbed conditions of total internal reflection in waveguides.

It is expected that the multilayered structure will provide not only functioning and interaction of neurons and their ensembles but also the generation of new neurons and links (emission output from one layer and penetrating to the other layers) between single neurons and neuronets, and will permit, according to the information (image) at system output, to connect and to correlate the cyclical processes originating in BR medium. At that point, the process of self-organization of data processing system will continue.

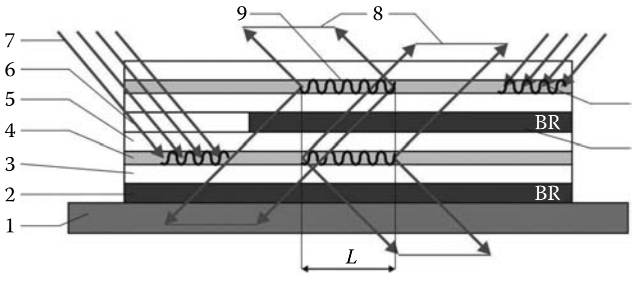

FIGURE 38.6 Fragment of multilayer structure: 1—substrate (glass K-8); 2, 10—layers, containing BR; 3, 5—boundary layers of flat waveguide; 4—guide layer of flat waveguide; 6—emission input area; 7—input emission; 8—output emission; 9—output emission gitter, 11—input emission gitter; *L* is the length of the output emission gitter.

The adaptability principles realization of data processing elements and system self-organization will permit to essentially reduce the requirements of the elements and facility as a whole to provide reliable functioning in case of the single-element failure.

Reduction of technological requirements is achieved by that neuro-like elements and links that are formed in continuous (uniform, i.e., not divided into constructive matrix elements) transparent (without optical dispersion) layer of photochromic material according to the light energy allocation.

38.2.2.2 Expected Parameters of the Element Base

According to the traditional criteria, it is acceptable to evaluate the number of the neuro-like elements in medium containing BR at area 10 mm^2 in quantity not less than 10^6. At the area in question, not less than 10^{11} bonds per second are realized (circuit time 0.1 ms, coefficient of bond formation 10).

38.2.2.3 The Basic Elements of the Multilayered Structures

The multilayered structures (Figure 38.6) for the realization of the basic neuronet data processing include: the system of flat waveguides, the elements of optical emission input as gitter, and the devices of the surface light front formation (Figure 38.7).

BR-containing polymeric films are meant for neuro-like elements formation by the change of the absorption/transmission surface geometry by modulated light flux effect.

The elements of optical emission input in flat waveguides and the elements of output are made as diffraction lattices (gitters) (Zlenko et al., 1975; Unger, 1980). The angles of radiation input and output depend on different layer indexes and glitter spacing; therefore, those angles can be different for various multilayered waveguides.

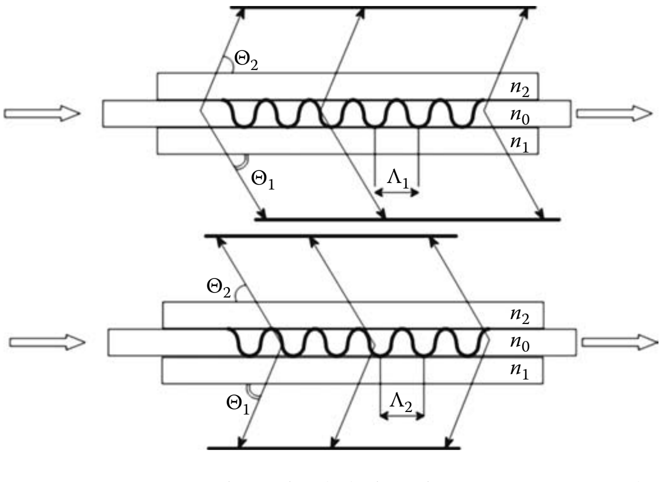

FIGURE 38.7 Device for surface light front formation: Θ_1, Θ_2—angles of emission input; Λ_1, Λ_2—diffraction lattice spacing; n_0, n_1, n_2—refraction indices.

The device for surface light front formation (Figure 38.7) generates directed light fluxes for effective allocation of the light energy in BR-containing polymeric films for neuro-like elements concurrent formation, their concurrent interaction, and output of the data processing results in neuro-like element medium as optical signals.

38.2.2.4 Requirements of BR-Containing Polymeric Films

The following are the requirements of optical and geometrical properties of BR-containing polymeric films.

For photochromic effects (induced changes of refraction and absorption indexes) to significantly appear during functioning process, high optic density and consequently substantial BR concentration in polymeric films are needed. Optical density of such films should be 0.8–1.3. In those conditions, basic characteristics of BR-containing media are utilized under optimum light flux density so that the media could be used for technical purposes. Exposure to light of the fluxes with radiation density equal to 10–100 mW/cm^2 induces the films to experience changes in their absorption and transmission characteristics as much as 10%–50% from the original indexes in as long as 0.1–10 s.

Functioning of films under induced changes in absorption level requires quite homogenous distribution of BR concentration, as optical heterogeneity infringes the conversion of optical information.

Besides, we must ensure repeatability of the main structural parameters: film thickness, surface finish, homogenous BR distribution throughout film surface (1–10 cm^2), etc.

Requirements to physical and chemical properties of polymeric matrix. Selection of matrix material.

To minimize the influence of diffraction divergence on information conversion processes, the thickness of BR-containing films must be 6–14 μm. To reach the specified optical density with that thickness, the volume of BR content in polymeric films should be 40%–50%. Far from all, the polymers transparent in optical range can meet the requirements above. Moreover, only water-soluble polymers can be used to form BR-containing polymeric films.

The comparative studies held to form polyvinyl alcohol- and gelatine-based BR-containing polymeric films have proved that gelatin-based polymeric matrixes have obvious advantages.

Gelatin-based films could apparently have the highest possible BR concentration (up to 50 vol. %) without aggregation of PM fragments due to thermodynamic peculiarities of gelatin polymerization process. Gelatin properties allow us to avoid destruction of BR protein structure while making polymeric mixture and further polymerization.

PM fragments embedded into a gelatin matrix are long-lasting and resistant to many technological factors. Polymerized gelatin creates optimum conditions for BR to function while retaining enough water needed for photochromic cycle. For the same reason, gelatin matrices make it possible to place environment-modifying water-soluble components and to change the photo cycle time frame.

38.2.2.5 Obtaining of Bacteriorhodopsin-Containing Polymeric Films for Multilayered Structures

It is important for processing and conversion of optical information to take into consideration the dispersion of optical emission in BR-containing media conditioned by purple membranes size (500–1000 nm)—comparable to the wave length of the optical range emission. Therefore, the primary suspensions of PM and BR-containing films based on them are optically nonhomogeneous, which results in the functional properties loss.

The dispersion demagnification is reached consequently: at the step of preparation of purple membranes suspension—by PM fragments separation; at the steps of polymeric mixture preparation and polymerization—by elimination of the aggregation process of the purple membranes fragments.

38.2.2.5.1 Preparation of Bacteriorhodopsin Suspension

For the preparation of PM suspension, the triple centrifugal purification (3000 rpm, 5 min) was carried out; pH value and BR concentration in suspension were measured. The pH value is significant for the following polymeric solutions and film obtaining, since the investigations showed that at pH less than 4.1 PM aggregated, the optical transparency of the suspension was not achieved. It was determined that the ultrasound exposure results in the decrease of pH value in suspension at 0.2–0.4. The control of pH value was managed by the addition of 0.01 M borax buffer solution $Na_2B_4O_7 \cdot 10H_2O$, pH = 9.18. During the ultrasound treatment, the suspension temperature must not exceed 36°C.

As the result of technological investigations of ultrasound treatment, optically transparent homogeneous BR suspensions were obtained without detergent addition. The side effect can be the partial melting of protein. Optically transparent PM suspensions were obtained with BR concentration up to 15 mg/mL.

Size evaluation (8.7 ± 0.5 μm) of PM fragments in treated suspensions was carried out by the intensity of Rayleigh scattering and showed that the applied technological mode of suspension treatment permits to separate PM into naturally minimum fragments—trimers without BR protein destruction and the principle possibility of BR-containing medium optical resolution at the level of a few thousand lines per millimeter can be considered.

38.2.2.5.2 Preparation of Bacteriorhodopsin-Containing Mixture

During the preparation of polymeric mixture based on BR, the last exhibited the property of aggregation on the polymer molecules that lead to optical heterogeneity of films. As the result of technological experiments combining thermal parameters and operating pH of components, the conditions selected under that aggregation were not observed and transparent optically homogeneous BR-containing polymeric mixtures were obtained. At the step of polymeric mixture preparation, pH control of gelatine solution was carried out because pH value in gelatine solution depends either on the obtaining method or on gelatine concentration. The component stirring in the mixture was also carried out by the ultrasound exposure. All the modifying components were put into the polymeric mixture at the last step under condition: the final pH value of the polymeric mixture had to be >4.1.

38.2.2.6 The Properties of BR-Containing Polymeric Films for Multilayered Structures

According to the elaborated technology, there are transparent and optically homogeneous BR-containing polymeric films (thickness 6–14 μm with optical density 0.8–1.3 D at $\lambda = 570$ nm) on substrates of glass K-8 and fused quartz (area up to 60 × 48 mm^2) and also on Si plates (76 mm in diameter).

38.2.2.6.1 Evaluation of PM Size Embedded to Film

Since during the film polymerization from the polymeric mixture fragment aggregation is possible, the evaluation of PM fragments size is being embedded into film. The placement of films with the embedded PM fragments in detection system between crossed polarizers resulted in no changes in the initial zero signal of photodetector. It means either no rotation of polarization in the light-pass direction, or no significant dispersion that confirms the PM fragments size to be much less than 0.63 μm.

Optical heterogeneity specified by surface geometry and allocation of BR concentration to the film surface is given below.

The distribution of optical absorption heterogeneity coefficient of BR-containing films is specified by the product of two values: the allocation heterogeneity of BR bulk concentration to the film surface and quality of the film surface as the local dilatation from the average thickness value. In the aggregate it results in the local dilatation of so-called surface concentration and, correspondingly, the optical density.

The optical homogeneity of the film being specified only by the surface quality is provided comparatively easily both for the films being realized by glazing method and for the films being obtained by centrifugation method. It was determined that the typical thickness deviation of BR-containing films being obtained equals less than 50 nm at length 10 mm, that, for example, at film thickness >5 μm equals <1%.

The support of the optical homogeneity being concerned with the allocation of BR bulk concentration along the film surface is hampered by the molecules migration in polymeric solutions during the polymerization process to the area of increased surface tension. Nevertheless, the attained deviation from the average value for allocation of BR bulk concentration does not exceed 3% since the changes of surface concentration and optical density lie in the same range.

38.2.2.6.2 The Structure of Bacteriorhodopsin-Containing Films

It is ascertained that the structure of BR-containing polymeric films surface depends on the obtaining conditions. Depending on pH, the surface is either smooth (Figure 38.8A) or it has punctual (100–300 nm) and rectilinear (400–1000 nm lengthwise, 100–200 nm wide) protuberances (Figure 38.8B and C). The cutting of the surface geometry elements is observed (Figure 38.8B). These investigations confirmed the necessity of pH value maintenance higher than 4.1.

Surface roughness (step height R_z of surface geometry) of the film with particle size of BR phase <10 nm amounts to $R_z \approx 0.1$ μm, with particle size of BR phase ~100–1000 nm $R_z \approx 0.6$–3.8 μm.

The analysis of the films with thickness 20 μm (containing BR-phase cutting) by Roentgen diffractometry led to the missing crystalline phase. Only amorphous gallo is observed on roentgenograms of BR-containing polymeric films, being formed under various conditions in the range of Bragg angles $2\theta = 5°$–8° ($k\alpha$—copper emission).

Clearly defined diffraction maximum appears on roentgenograms of some samples for angles $2\theta = 7.5°$–7.8° (Figure 38.9) corresponding to interplanar spacing $d/n = 1.179$–1.133 nm. For a number of samples, diffraction maximum appears on roentgenograms for angles $2\theta = 18°$–19° ($d/n = 0.467$–0.493 nm). That indicates the ordered placement of protein complexes in polymeric (gelatinous) matrix. Apparently, the axial texturing of lamellar protein complexes exists by the axis normal to the surface of BR plates.

38.2.2.6.3 Refraction Index of BR and BR-Containing Polymeric Films

Information about refraction index in literature (Hampp, 2000) is inconsistent and not exact enough to solve the problem of the multilayered structure construction. Therefore, we made our own measurements.

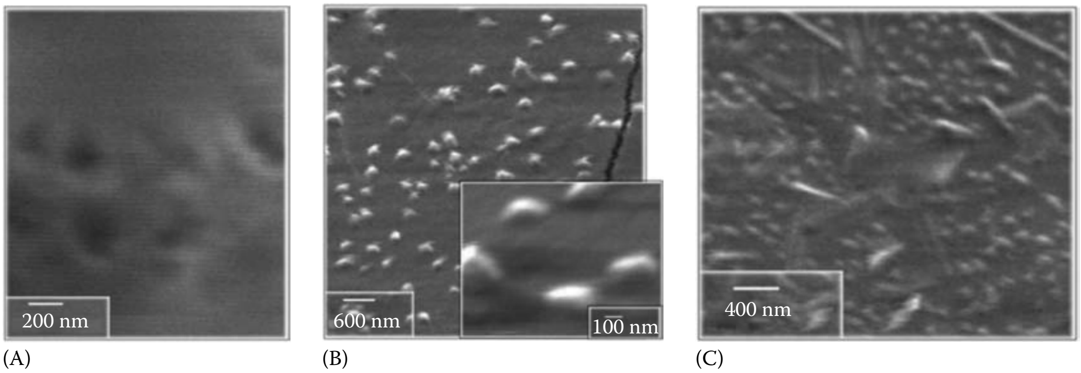

FIGURE 38.8 The structure of BR-containing polymeric film surface being formed from the polymeric mixture: (A) pH = 4.2–4.5; (B) pH = 3.8–4.0 (the enlarged surface fragment is presented at the insertion); (C) pH = 3.5–3.7.

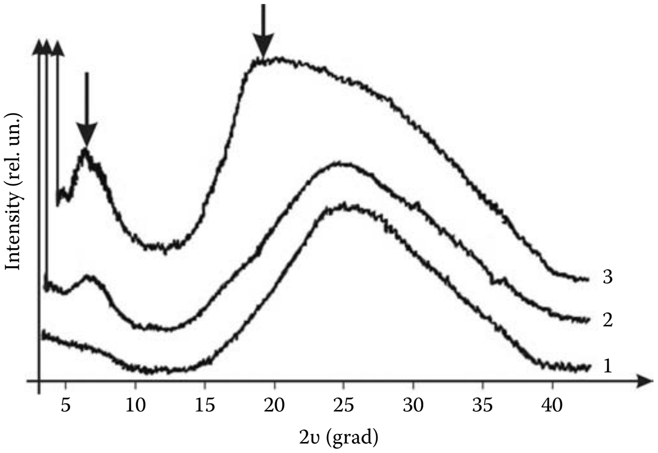

FIGURE 38.9 Roentgenograms, obtained on the samples: 1—glass K-8; 2—glass K-8 with the deposited gelatinous film; 3—glass K-8 with the deposited BR-containing polymeric film.

The refraction index measurement was carried out using refractometer IRF-454B capable of measuring the refraction index in the range from 1.2 to 1.7 in reflected and transmitted light. BR-containing film was formed on the surface of measuring prism by the method of suspension glazing (concentration 15–20 mg/mL) and the following drying at 22°C temperature, relative 60% humidity. The film thickness reached 20 ± 5 μm.

The value 1.534 ± 0.002 of refraction index was obtained in the white light. This result is well reproducible on every one of the samples being prepared from the three different halobacteria strains and conforms to the results of the following refraction index measurements in polymeric BR-containing films. The value of the refraction index in that case is defined by BR concentration in the film.

The refraction index measurement of gelatin was held in order to predict the parameters of the polymeric BR-containing films under construction depending on BR concentration in the film. The refraction index value of polymeric film based on gelatin without BR (formation by the glazing method on the surface of the refractometer measuring prism) got equal to 1.543 ± 0.001. Refraction index was measured at 22°C temperature and relative 60% humidity.

The refraction index value of BR-containing gelatinous film with thickness 50 ± 10 μm lies in the range of 1.539–1.542 depending on BR concentration in gelatin.

38.3 Nanostructuring of Bacteriorhodopsin-Containing Molecular Media

Functional parameters of BR-based multilayered structures on the base of BR (the increase of photochromic sensitivity, management of photocycle duration) can reach their maximum capacity in use by additional nanostructuring of molecular media. Enhancement of functional properties of multilayered structures is caused by introduction of new functional elements—hybrid nanostructures—making it possible to further realize a new class of optical information systems based on self-organizing oscillatory and auto wave hierarchical processes in BR-containing nanocomposites.

Both nanotechnological and nanotechno-systematic approaches are described below.

While developing BR-based media used to process information with neural network methods, two ways of material structuring are applied. The first way means introduction of modifying organic and nonorganic additives into the system to improve functional characteristics of BR. Modification is done on molecular level that is why such material is a molecular one. The second way means building of hybrid functional structures with every component performing its own function. In this way, structuring goes at nanometric range instead of molecular level. Hybrid nanostructures are systems made of three components—colloidal nanoparticles (metal or semiconductor structures based on transition element [TM] chalcogenide), bridge molecules (spacers) with various linear dimensions, and BR molecules.

38.3.1 Procedure of Complex Estimation of Functional Characteristics of Bacteriorhodopsin-Containing Materials

To characterize and compare functional characteristics of BR-based films made by various methods and difference in thickness, optical absorption, structure, and content (including concentration of BR molecules), it is convenient to apply the designated factor for quantity estimation of photochromic sensitivity. For optimization of experimental investigations, we developed the simulator of photo-dependent processes and designated factor $k_{570}(t)$ of the photo-induced transitions of BR molecules from the primary state BR570:

$$k_{570}(t) = \frac{N_2(t)}{N_0} = \frac{N_0 - N_1(t)}{N_0}, \tag{38.7}$$

where

N_1 is concentration of BR570
N_2 is the concentration of BR412
$N_1 + N_2 = N_0$ is the total concentration of BR molecules

That factor $k_{570}(t)$ named by us as the coefficient of photo-induced transition of BR molecules from basic state BR570 was determined from experimental data of changing optical absorption at wavelength 570 nm under the influence of active inducing radiation. This complex parameter takes into consideration the temporary characteristics of excitation and relaxation processes, quantum efficiency, and cross-section of BR molecules contained in the nanocomposite nanostructured films (media).

While studying the influence of technological formation processes on properties of BR-containing materials, it is very labor-consuming to measure each of the above specified characteristics. Therefore, we use a model to describe photo-dependent processes that pass in BR-based materials under radiation.

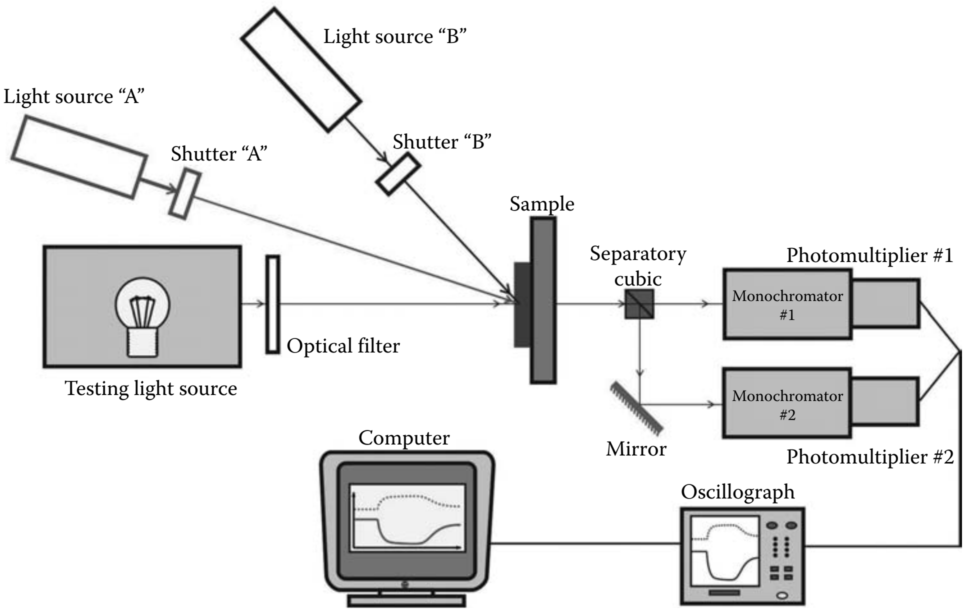

FIGURE 38.10 Schematic diagram of a soft- and hardware complex designed to study functional characteristics of BR-based materials.

We have also designed a method to assess the functional characteristics of those media after one measurement. To carry out all necessary measurements, we can use a soft- and hardware complex (Figure 38.10), making it possible to combine light exposure with two wave lengths in the band of main intermediate absorption (at 570 and 412 nm, respectively).

The light radiated by the source of continuous testing radiation with up to 2 mW/cm^2 intensity passes through the sample and falls onto the monochromator where a narrow spectrum line (at 570 nm) is allocated. The signal goes to the photo-electronic multiplier being adjusted to convert the signal and transfer it to the oscillograph. The sample is exposed to light pulsed by the source of exciting radiation with intensity at least 10 times more than the intensity of testing radiation. The signal registered with the oscillograph goes to the computer for mathematic curve processing as per the below physical and mathematic ratios.

Based on the Buger–Lambert–Beer law, taking into consideration photo-induced changing concentrations N_1 and N_2, the expression has been obtained that allows to calculate the value of $k_{570}(t)$ from the experimental data recording testing light $\lambda = 570$ nm transmission changing under illumination of exciting radiation:

$$k_{570}(t) = \frac{\lg\left(I(t)/I_1\right)}{\lg\left(I_1/I_0\right)}, \tag{38.8}$$

where

I_0 is the intensity of the incident test light
I_1 is the intensity of the transmitted test light in the absence of exciting radiation
$I(t)$ is the intensity of the transmitted test light under exciting radiation at the time moment t

Also the proposed method of the complex estimation is based on the kinetic equation evaluating concentration distribution of BR molecules between forms of BR570 (N_1) and M412 (N_2):

$$\frac{dN_1}{dt} = -\sigma_1 \cdot A_1 \cdot \frac{P}{h \cdot \nu} \cdot N_1 + \frac{1}{\tau} \cdot N_2, \tag{38.9}$$

where

σ_1 is the absorption cross-section of BR570 form (on wavelength of acting light)
A_1 is the quantum yield of photoreaction
P is the power density of acting light
τ is the lifetime of M412 form
h is the Plank constant
ν is the acting light frequency

From the same experimental data (Figure 38.11) based on the solution of equation evaluating concentration distribution of BR molecules between forms BR570 (N_1) and M412 (N_2), the estimation of quantum yield of photoreaction A_1 from the value of the derivative at the point $t = 0$ (the beginning of exciting radiation action) and the estimation of the lifetime of M412 form τ from value of derivative at the point $t = t^*$ (the end of exciting radiation action) may be obtained. As time goes by, film transmission value being influenced by the relaxation of excited molecules approaches to the original transmission level.

38.3.2 Functional Characteristics of Nanostructured Bacteriorhodopsin Films

To form the effective functional structures out of biomaterials, we need to apply various technological methods. In particular, we can use BR-containing materials to prove the necessity to use

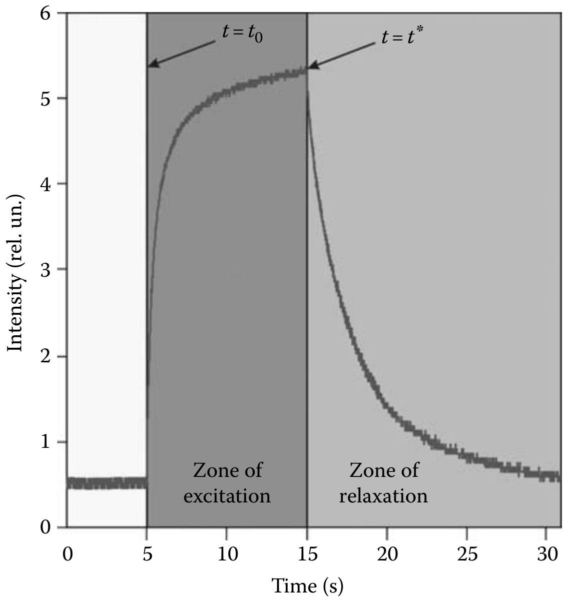

FIGURE 38.11 Typical curve of photo-induced changes of optical transmission of BR-containing media.

the additional nanostructuring of functional biomolecular films subject to storage and neural network processing with optical methods.

As a result of non-modified BR films' investigations, we established that immediately after obtaining the values, $k_{570}(t)$ are in interval 0.5–0.7 and then decrease to value 0.2 for 3–5 h and demonstrate this value during a few years. In the primary formed films, BR molecules constitute the structure that in time is destroyed in consequence of the heat oscillations, resulting in disordered molecule orientation relative to each other and destruction of hydrogen bonds. To keep the high photochromic sensitivity for the whole period, we can create additional linkages (covalent or hydrogenous) between protein molecules by using chemical reagents.

For example, our use of diamine (1.4-diaminobenzen [DAB]) in the ratio of BR:DAB = 1:3, 1:6 и 1:9 (molecular ratio) has resulted in increased values of factor $k_{570}(t)$ (0.35–0.5) in comparison with non-modified BR films (0.2–0.25) during 12 days. The stabilization of the primary formed BR structure occurs due to the fact that 1.4-diaminobenzen is an aromatic amine and capable to interact with the carboxyl groups of glutamic and aspartic acids. In our opinion, the high values $k_{570}(t)$ for BR:DAB films are the result of raising the lifetime of M412 form owing to the fact that 1,4-diaminobenzen screens extract proton groups.

At the same time the use of glutaric dialdehyde (GA) with sodium tetraborate for the alkaline catalysis of the reaction of bonding between amine groups of BR lysine residues and aldehydic groups of GA (the Schiff base) for the samples BR:GA = 1:5 had more high values of factor $k_{570}(t) = 0.55$ (Figure 38.12), whereas BR:GA = 1:10 and 1:30 had $k_{570}(t) = 0.4–0.5$, which is explained by the high content of GA oxidation products by atmospheric O_2 (glutaric acid) that partially destroy BR molecules. The dynamics change of decay factor $k_{570}(t)$ value was studied during 1 year for BR:GA films (Figure 38.12). The samples BR:GA = 1:5 kept the high value of factor $k_{570}(t)$ (≥0.45) for 120 days. After 330 days for all the samples of BR:GA $k_{570}(t) = 0.3$.

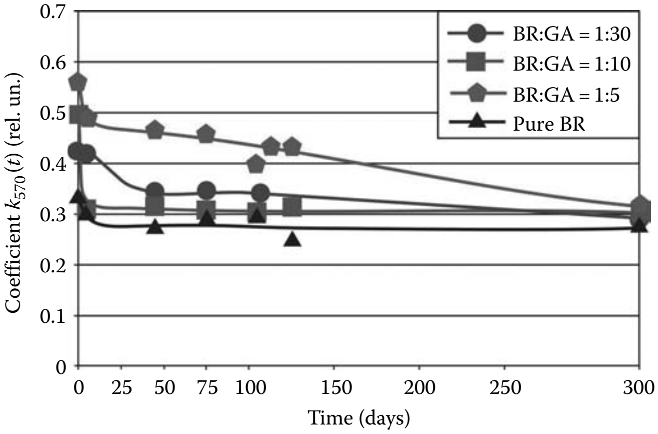

FIGURE 38.12 Changing values $k_{570}(t)$ in the operation process for the films. BR:GA = 1:30, 1:10, and 1:5.

Introduction of amino acids in the content of BR-containing films leads to even better results. The films of BR: glycine, BR: isoleucine, and BR: lysine in molecular ratio from 1:1 to 1:25 had been made. The BR: glycine and BR: isoleucine films were first inhomogeneous optically and characterized by the high light scattering, and consequently had low values of factor $k_{570}(t)$ within the limits of 0.05–0.1. The BR: lysine films were optically transparent and homogeneous. The dynamics of changing value $k_{570}(t)$ for the films with ratio BR: lysine = 1:10 and 1:25 was shown in Figure 38.13. The samples with the ratio BR: lysine = 1:25 kept stable higher values during not less than

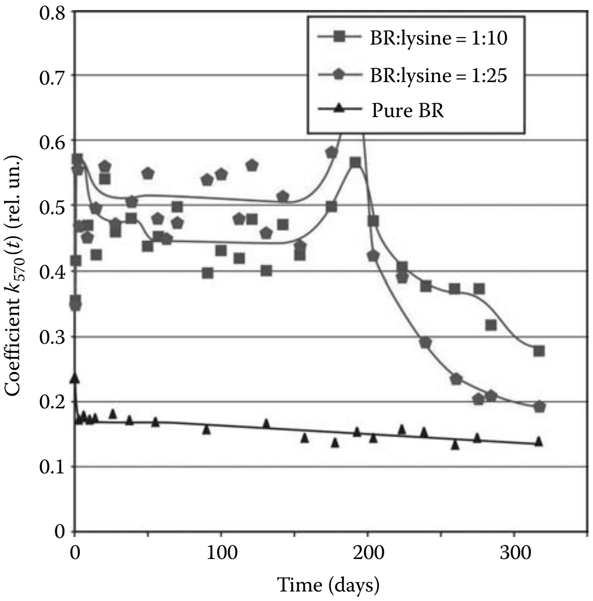

FIGURE 38.13 Changing values $k_{570}(t)$ in the operation process for the films. BR: lysine = 1:10 and 1:25.

80 days—$k_{570}(t)$ (0.4–0.55)—in comparison with the check sample (0.15–0.2). At that optical density and transparency, films did not change throughout that time. The samples with ratio BR: lysine = 1:10 demonstrated higher values $k_{570}(t)$ (0.35–0.5) in comparison with the check sample (0.15–0.2) for 330 days.

The above results prove that there are considerable opportunities to improve functional characteristics of BR-containing materials by nanostructuring (by forming additional bonds between BR molecules).

38.3.3 Metal Nanoparticles and Bacteriorhodopsin-Based Hybrid Nanostructures

The other way to expand the functional opportunities of molecular nanophotonics materials is to form hybrid nanostructures. We suggest the system use of some physical phenomena common for nanoscale structures as the basic processes for photonic information systems. In particular, we can use the effect that nano-objects have on the lifetime of excited atoms and molecules located closer than radiation wavelength. This effect is rapidly enhanced near nanoparticles when strong electromagnetic fields locate at plasma resonance frequency. Strong local fields influence the speed of electron transitions (i.e., the processes of light absorption and spontaneous radiation). Such fields can also change spectral characteristics of those processes and lead to substantial enhancement of various nonlinear optical effects. In particular due to strong local fields we can manage tumescence enhancement and effective suppression (depending on the distance), have radiationless energy transition, and manage the quantum efficiency of BR photoreaction and the lifetime of its spectral intermediates. The specified mechanisms could be useful to design photonic devices, which perform basic functions of information systems on the basis of nonlinear dissipative media with distributed feedbacks. It is to be taken into account that the effect nanoparticles have on the excited state of the functional molecule becomes different with the change of the distance between them. This circumstance creates wide opportunities to crease hybrid nanostructure-based composition materials with bridge molecules of various lengths.

In general, hybrid nanostructures are the systems made of three components—colloidal nanoparticles, bridge molecules (spacers) with various linear dimensions, and functional (photochromic, photoluminescence, electroluminescence, etc.) molecules. Below is the description and diagram (Figure 38.14) to show different aspects of hybrid nanostructures' formation with metal nanoparticles.

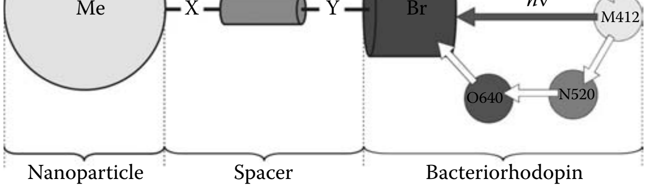

FIGURE 38.14 Schematic composition of hybrid nanostructures.

Formation of the hybrid nanostructures with adjusted distance between a nanoparticle and a functional molecule includes several key stages:

1. Formation and stabilization of nanoparticles in the polar or nonpolar solvent.
2. Modification of nanoparticle surface in order to functionalize and to ensure selective and self-organizing hybrid nanostructures.
3. Binding of functional molecules with nanoparticle surface.

Formation of nanoparticles compatible with BR photochromic protein has its own peculiarities due to the biological nature of the material as water is used as the main solvent.

The number of spacers used to form BR-based hybrid structures is quite limited due to the chemical properties of both components. First, the spacers must be hydrophilic substances, well soluble in water. Second, functional groups are also very limited. As BR is a protein, its structure is based on polypeptide amino acid consequence, including among others asparagic and glutamic acid residues. Those residues have free carboxylate groups taking no part in the formation of peptide bonds. pH area, where BR demonstrates its functional characteristics, is limited to pH > 5. Under those conditions the molecule is negatively charged. Therefore, to bind with BR functional spacer, the groups must be positively charged. The most efficient spacers look like molecules, including free aliphatic positively charged amino acids. Those substances include, for example, symmetrical diamines (1.6-diaminohexane, 1.7-diaminoheptane) or lysine (2.5-diaminopentane acid). Besides, in low-molecular compounds, one can use polymers, including amino acids in their structure. Polylysine is one of such polymers.

Gold nanoparticle-based nanohybrid structures are relatively easy to form as gold nanoparticles are chemically inert in aqueous suspensions. At the first technological stage we apply the borohydride method, which makes it possible to get smaller particles with more homogenous size distribution than nanoparticles obtained by using the citrate method. Besides to reach that we do not need to use any additional technological tricks. To apply the borohydride method, add 1 mL 1% $HAuCl_4$, 0,5 mL 0.1 M K_2CO_3 to 100 mL distilled water. While stirring the mixture vigorously with a magnetic stirrer, add 140 μL 0.1 M $NaBH_4$ solution (in some 10 μL portions during 10 min).

The technological methods of forming hybrid structures that we are developing are based on mechanisms of special bonding and self-organization activated by interaction of spacers with the appropriate functional groups. To build hybrid nanostructures based on formed gold nanoparticles, we have to modify nanoparticle surface with amino-containing spacers, which are able to actively interact with both BR molecules and the surface of nanoparticle. We have selected symmetrical aliphatic diamines, in particular, 1.6-diaminohexane. The surface of nanoparticles is modified by introducing 1.6-diaminohexane water solution with up to 5000 spacer molecules to 1 nanoparticle. The intensity of spacer molecules bonding with the surface of nanoparticles can be indicated by the fact that after introducing 1.6-diaminohexane, excess gold nanoparticles are quick to aggregate (sometimes with residue). The results of that process can be seen while extracting films out of suspensions by methods of atom microscopic investigation (Figure 38.15).

Comprehensive studies have proved that we need to not only modify the surface of gold nanoparticles but also remove the suspension reacting system components that have failed to react during nanoparticle synthesis as those components destruct BR molecules and obstruct synthesis of hybrid structures. To solve this problem, we have suggested the use of dialyze. Place nanoparticle-containing suspension in the so-called dialysis bag, which is a hollow plastic container with some nanometer-size holes. Sink the dialysis bag into the solution of the substance to be introduced into the system (concentration of the solution should be enough to modify the surface of nanoparticles effectively). The volume of that solution is usually at least 10 times more than the volume of the original suspension. During the process, small organic and nonorganic molecules migrate through the hollow walls of the dialysis bag, depending on the concentration gradient of the hole. Nanoparticles that are bigger than bag holes stay inside the dialysis bag. Some time (around 2 days) later thermodynamic balance is reached. That means that concentration of organic and nonorganic substances get homogenous throughout the volume. Therefore, the solution inside the dialysis bag experiences dilution due to contaminating impurities to an extent and solution volume relates to the volume of the original suspension. Concentration of the substance to modify nanoparticle surface is the same as that of the original solution. Using this method, we have obtained gold nanoparticles with 1.6-diaminohexane-modified surface. Those suspensions are stable. The studies of spectral characteristics of such suspensions have proved that dialysis has no effect on the spectrum of gold nanoparticles form (Figure 38.16).

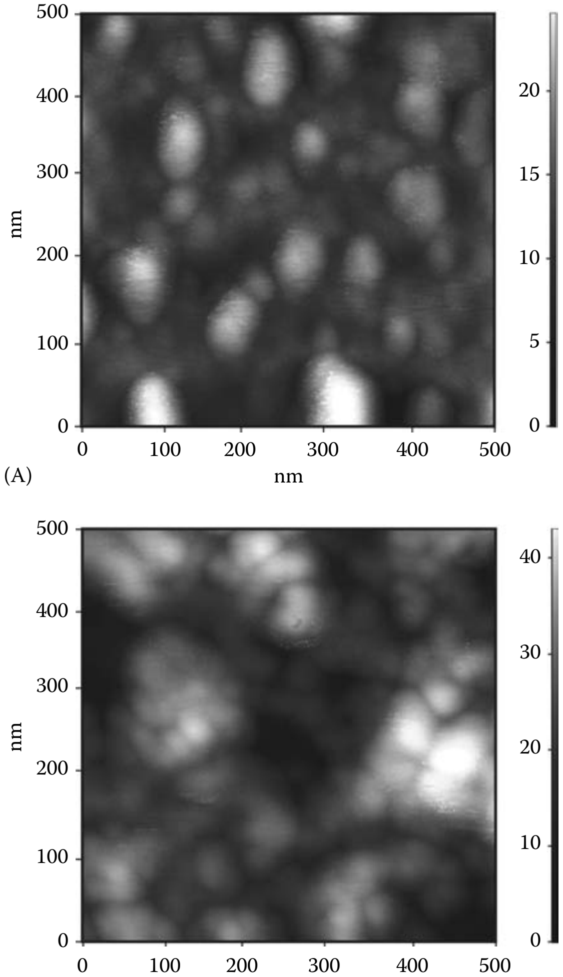

FIGURE 38.15 AFM characterization of the surface of gold nanoparticle-based films extracted from (A) a suspension with optimum 1.6-diaminohexane level; (B) from a suspension with excessive 1.6-diaminohexane level.

We have synthesized hybrid structures with nanoparticles to BR molecules ratio 1:3600. BR suspension (51 mg/mL) is used as an original one. Hybrid structures are formed by vigorously stirring the suspension with a magnetic stirrer for 2 days at the temperature of the reaction mixture equal to 20°C–25°C. Figure 38.17 demonstrates spectral characteristics of the hybrid nanostructures suspension based on gold nanoparticles, 1.6-diaminohexane, and BR. We have to stress that characteristic spectral maximums of certain components do not decompose and therefore make it more complicated to study functional characteristics of those structures. By using extraction, we have obtained films with the surface characterized by atom microscopic investigation (Figure 38.18).

The use of silver nanoparticles providing optimum combination of spectral characteristics in the nanoparticle–BR system is of more interest. At the same time, silver nanoparticles demonstrate substantial chemical and photochemical activity in water suspension and require special stabilization measures.

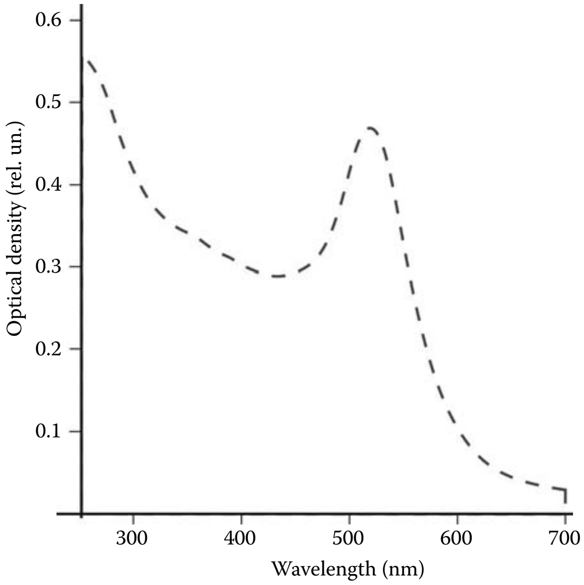

FIGURE 38.16 Spectrum of gold nanoparticles after dialysis with 1.6-diaminohexane solution.

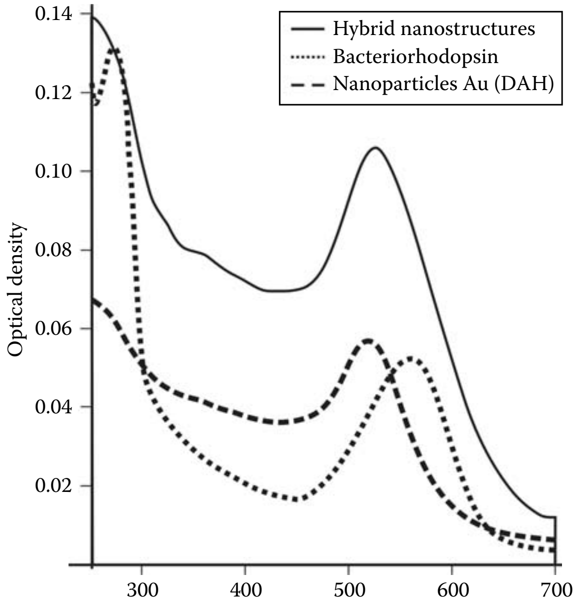

FIGURE 38.17 Absorption spectrum for suspension of hybrid nanostructures based on gold nanoparticles, 1.6-diaminohexane, and BR.

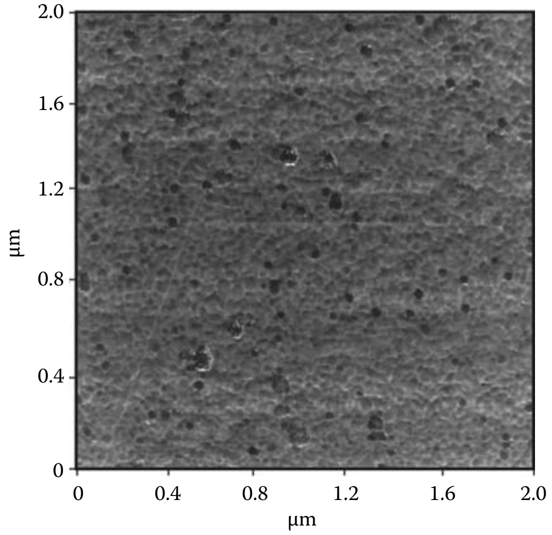

FIGURE 38.18 AFM characterization of the surface of hybrid nanostructures films based on gold nanoparticles, 1.6-diaminohexane, and BR.

To solve this problem, we introduce bridge molecules with various functional groups into the reaction system used to synthesize nanoparticles. This single technological operation includes two stages: formation and stabilization of nanoparticles and modification of their surface.

By using this method, we have obtained water suspensions of hybrid nanostructures based on BR and silver nanoparticles while using cysteine, arginine, lysine, and polylysine as spacers.

To form water colloidal suspension of silver nanoparticles we have used the following original water solutions:

1. Cysteine, arginine, and lysine with concentrations $c = 0.09\%$, 0.118%, and 0.122% (weight) correspondingly
2. Polylysine with concentration $c = 0.125\%$ (weight)
3. Silver nitrate with concentration $c = 0.024\%$ (weight)
4. Water solution of sodium boron hydride with concentration $c = 0.4\%$ (weight) (to stabilize particles we have added 2.5% (vol.) 1 M solution of sodium hydroxide)

To form nanoparticles, first we add the specified volume of spacer solution (up to 0.2% total volume) into silver nitrate solution. Then we introduce 120 μL boron hydride solution (divided into several 30 μL portions to be introduced every 5 min) while stirring the mixture vigorously with a magnetic stirrer. Characteristic absorption maximums in the spectrums of silver nanoparticles suspensions stabilized with lysine, arginine, cysteine, and polylysine are at wavelength 390–400 nm (Figure 38.19). Comparing the bands of absorption of silver nanoparticles stabilized with monomer amino acids and stabilized with polylysine, we can conclude that nanoparticles stabilized with amino acids are smaller in size than nanoparticles stabilized with polylysine. That conclusion is proved by the fact that the spectral maximum of the polylysine-stabilized nanoparticles is located in a longer wavelength area, at $\lambda = 398$ nm, whereas absorption maximums for amino-acid-stabilized nanoparticles are located at $\lambda = 390$–392 nm. Nanoparticles of smaller sizes are likely to occur due to the higher mobility of monomer molecules as synthesis is a quick

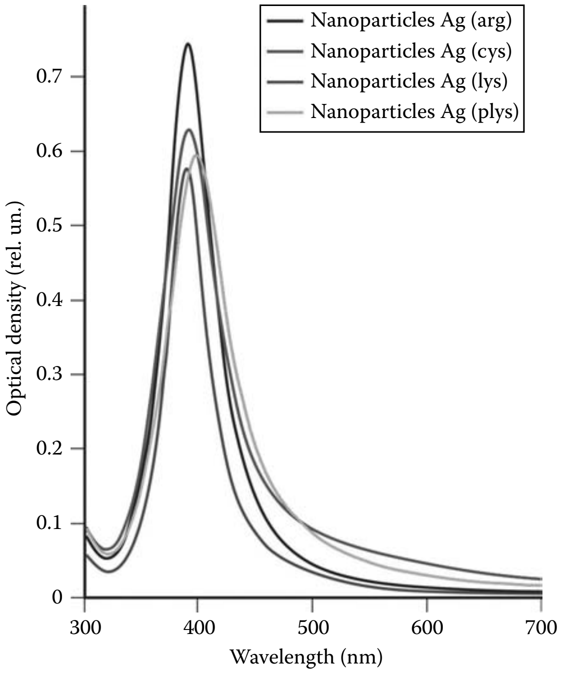

FIGURE 38.19 Absorption spectra for suspensions of silver nanoparticles stabilized with various amino acids (arg—arginine; cys—cysteine; lys—lysine; plys—polylysine).

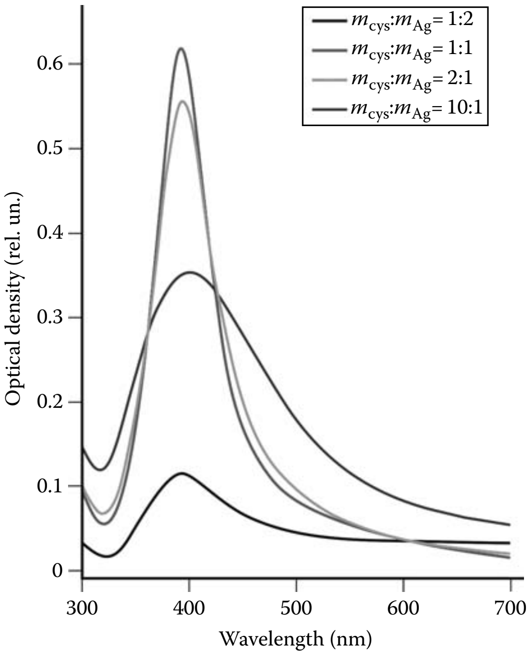

FIGURE 38.20 Absorption spectra for suspensions of cysteine-stabilized silver nanoparticles depending on various spacer concentrations.

reaction, whereas polymer molecules of polylysine are too big to take optimal confrontation that fast. It could also be explained by the fact that silver nanoparticles quite easily form nonspherical structures. This can lead to absorption maximums moving to the long wavelength area (Krutyakov et al., 2008).

The spectral characteristics (location of absorption maximum and absorption intensity) of the silver nanoparticle suspensions also depend on spacer concentration. Nanoparticles of the most homogenous sizes have been formed in suspensions with silver into spacer ratio close to 1:1. The fact is proved by the smaller semi-width of characteristic peaks in the nanoparticle spectrum (Figure 38.20).

Figure 38.19 demonstrates that characteristic maximums in the spectrum of absorption of silver nanoparticles and BR are divided into 160–170 nm to make it easier for the analysis and application of hybrid nanostructure optical properties to form functional nanocomposites to happen.

As mentioned above, the technological methods of forming hybrid nanostructures are based on the mechanisms of special bonding and self-organization activated by selecting appropriate bridge molecules for various functional groups. For example, during formation and stabilization of silver nanoparticle molecules of cysteine, arginine, lysine, and polylysine bond with the surface with the help of thio- (cysteine), guanydo- (arginine), or amino groups (lysine and polylysine). At the same time those molecules are spacers as they include additional amino groups which are able to actively interact with carboxyl groups of glutamic and asparagic acid residues, included in the amino acid consequence of BR. So, by using silver nanoparticles and BR molecules we can obtain hybrid nanostructures.

Such hybrid structures are synthesized with nanoparticles into BR molecules ratio equal to 1:3600. We use 5 mg/mL BR suspension as the original one. The hybrid nanostructures are formed by vigorously stirring the suspension with a magnetic stirrer for 1 day at the temperature of the reaction mixture equal to 22°C–25°C.

Figures 38.21 and 38.22 demonstrate spectral characteristics of hybrid nanostructures using lysine and polylysine as spacers.

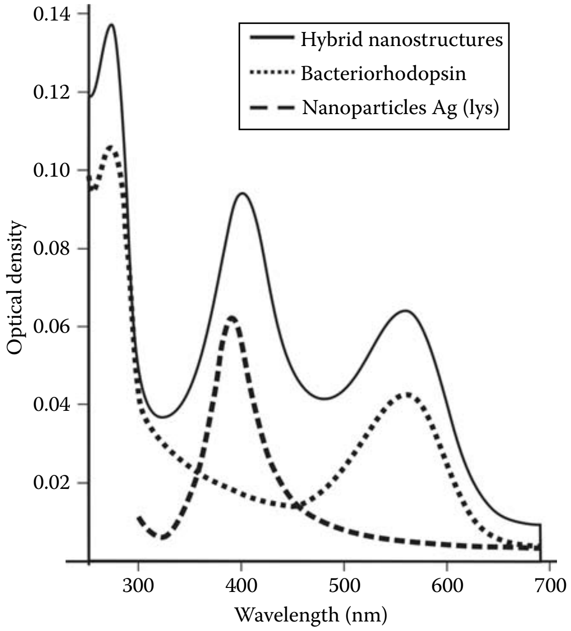

FIGURE 38.21 Absorption spectrum for suspension of hybrid nanostructures based on silver nanoparticles, lysine, and BR.

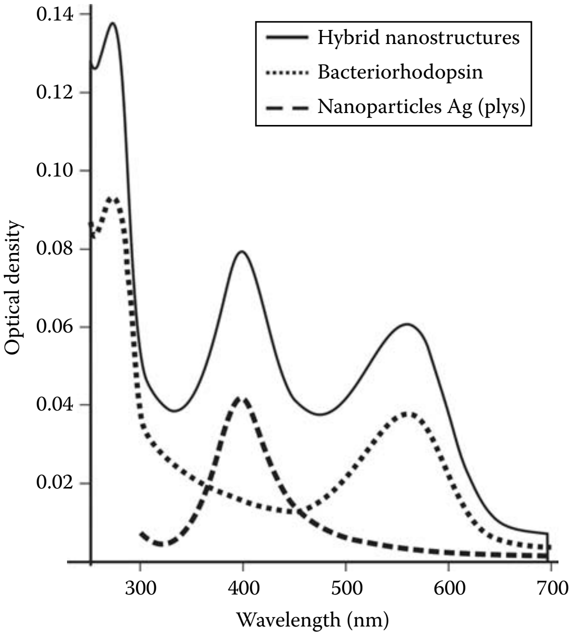

FIGURE 38.22 Absorption spectrum for suspension of hybrid nanostructures based on silver nanoparticles, polylysine, and BR.

While analyzing the technological results of obtaining hybrid nanostructures, we have to pay attention to the following:

The spectral characteristics of hybrid nanostructure suspensions demonstrate two characteristic spectral maximums, corresponding to certain components. It proves that BR and nanoparticle bonding causes no destruction of materials and BR molecules keep their functioning.

1. The characteristic spectral maximums of certain components of hybrid nanostructures may be offset. For example, the suspension of hybrid nanostructures based on silver nanoparticles, lysine, and BR demonstrates absorption maximum offset for silver nanoparticles from 391 to 408 nm (Figure 38.21), whereas the location of absorption maximum for BR stays the same—568 nm. The suspension of hybrid nanostructures based on the silver nanoparticles, polylysine, and BR demonstrates absorption maximum offset for BR from 568 to 560 nm, whereas the location of absorption maximum for the silver nanoparticles stays almost the same 400 nm (Figure 38.22). It could possibly happen because of the different distances between nanostructure components due to various spacer molecules (≈0.75 nm of lysine and ≈1.5 nm of polylysine), the type of bonding with nanoparticle surface, and BR.
2. After measuring the spectral characteristics of hybrid nanostructures, we have found out considerably increased absorption by every component of hybrid nanostructure against the adaptive spectral characteristic calculated by the simple adding of the spectrum of separate components in the same concentration they have in the suspension of hybrid nanostructures. The absorption increase reaches 13% at $\lambda = 394$ nm (cysteine-stabilized silver nanoparticles) and 26% at $\lambda = 568$ nm (BR) (Figure 38.23).

Therefore, synthesized hybrid nanostructures demonstrate functionality, mutual influence of certain components on spectral characteristics of each other, as well as an influence on BR photocycle.

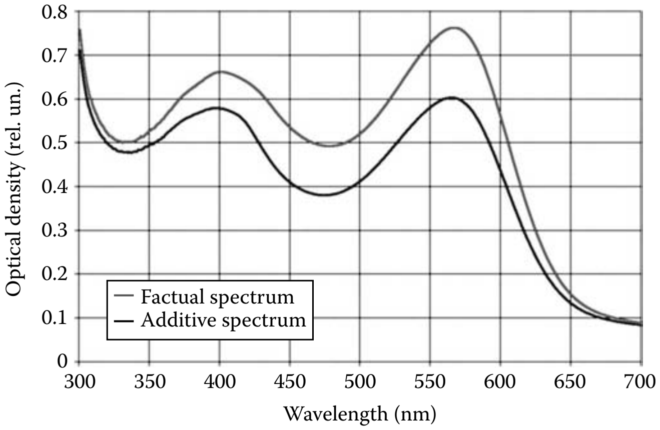

FIGURE 38.23 Factual spectrum of hybrid nanostructure suspension based on silver nanoparticles, cysteine, bacteriorhodopsin, and additive spectrum from spectra of separated components.

38.4 Summary

The use of molecules with conformation-adjusted relative position (distance and angle) of molecular groups including bridge functional groups as spacers can contribute to the development of nanotechnological and nanosystem approaches based on self-organization. Using the above method, we can regulate dipole–dipole interaction of a nanoparticle and a functional molecule. The effect is more obvious for nonspherical particles. Taking into the account that BR molecule performs photo-dependent transit of charge (H^+), the above method of regulating photochromic photoreaction and the lifetime of intermediates must be quite effective. The suggested approach can also be applied for hybrid nanostructures with different composition.

Therefore, we have demonstrated the system and technological aspects of self-organization of functional nanostructures of molecular photonics realizing bionic principles of neural network appliances operation.

The systems designed are expected to have some functional characteristics, making them to sense and analyze like live intellectual neural systems. For example, apart from having distributed and content-addressed memory and the ability to learn and identify steady images, those systems will be able to identify dynamic images, i.e., images of evolving objects. Such abilities are based on new functions built upon self-organization of structures and processes, including: adaptive sensing with feedbacks, regulating activity of various parts of artificial retina; adaptive change of sampling rate and of retina coverage; modeling of tunable retinal receptive fields with spatial filtering; and formation of structures sensitive to the direction and speed of object movements.

References

Akiyama, K., Takimoto, A., and Ogawa, H. 1995. Spatial light modulator and neural network. U.S. Patent 5428711.

Birge, R.R., Gillespie, N.B., Izaguirre, E.W., Kusnetzow, A., Lawrence, A.F., Singh, D., Song, Q.W., Schmidt, E., Stuart, J.A., Seetharaman, S., and Wise, K.J. 1999. Biomolecular electronic: Protein-based associative processors and volumetric memories. *Journal of Physical Chemistry B* 103: 10746–10766.

Engel, S.J. 1990. Neural network processing system. European Patent EP0382230A2.

Grebennikov, E.P. 1997. Light-radiation-induced structure formation of bacteriorhodopsin films for the development of self-organizing information processing systems. *Proceedings of SPIE* 3402: 460–465.

Hampp, N. 2000. Bacteriorhodopsin as a photochromic retinal protein for optical memories. *Chemical Review* 100: 1755–1776.

Haronian, D. and Lewis, A. 1991. Elements of a unique bacteriorhodopsin neural network architecture. *Applied of Optics* 30(5): 597–608.

Krutyakov, Yu.A., Kudrinskiy, A.A., Olenin, A.Yu., and Lisichkin, G.V. 2008. Synthesis and properties of silver nanoparticles: Advances and prospects. *Russian Chemical Review* 77(3): 233–257.

Oesterhelt, D., Brauchle, C., and Hampp, N. 1991. Bacteriorhodopsin: A biological material for information processing. *Quarterly Reviews of Biophysics* 24(4): 425–478.

Prigogine, I. 1980. *From Being to Becoming: Time and Complexity in the Physical Sciences*. San Francisco, CA, W. H. Freeman & Co.

Siebert, F., Mantele, W., and Kreutz, W. 1982. Evidence for the protonation of two internal carboxylic groups during the photocycle of bacteriorhodopsin. *FEBS Letters* 41: 82–87.

Unger, H.G. 1980. *Planar Optical Waveguides and Fibers*. Oxford, U.K., Oxford University Press.

Vsevolodov, N.N. 1988. *Biopigment-Photographic Recorders. Photomaterial Based on Bacteriorhodopsin*. Moscow, Science Publ. (in Russian).

Wasserman, F. 1989. *Neural Computing: Theory and Practice*. New York, Van Nostrand Reinhold Co.

Zeisel, D. and Hampp, N. 1992. Spectral relationship of light-induced refractive-index and absorption changes in bacteriorhodopsin films containing BR-WT and the variant BR-D96N. *Journal of Physical Chemistry* 96(19): 7787–7792.

Zlenko, A.A., Kiselev, V.A., Prokhorov, A.M., Spikhal'skii, A.A., and Sychugov, V.A. 1975. Emission and reflection of light by a corrugated section of a waveguide. *Soviet Journal of Quantum Electronics* 5: 1325–1328.

VI

Nanoscale Lasers

39 **Nanolasers** *Marek S. Wartak* **39**-1
Introduction: Properties of Laser Radiation • Overview of Lasers • Semiconductor Lasers • Rate Equation Approach • Definition of Nanolaser • Nanowire Lasers • Plasmonic Lasers • Photonic Crystal Lasers • Scattering Lasers • Organic Nanolasers • Applications • Conclusions • Acknowledgment • References

40 **Quantum Dot Laser** *Frank Jahnke* **40**-1
Introduction • Electronic States of Quantum Dots • Scattering Processes of Carriers in Quantum Dots • Optical Gain of the Active Material • Laser Emission Properties • Acknowledgments • References

41 **Mode-Locked Quantum-Dot Lasers** *Maria A. Cataluna and Edik U. Rafailov* **41**-1
Introduction • Ultrafast Laser Diodes • Quantum Dots: Distinctive Advantages for Ultrafast Diode Lasers • Mode-Locked Quantum-Dot Lasers: State of the Art • Summary and Outlook • Acknowledgments • References

39

Nanolasers

39.1 Introduction: Properties of Laser Radiation 39-1
39.2 Overview of Lasers 39-2
39.3 Semiconductor Lasers 39-2
39.4 Rate Equation Approach 39-4
39.5 Definition of Nanolaser 39-5
39.6 Nanowire Lasers 39-5
Ring Resonator Nanolasers • Theoretical Analysis of Nanowires • Recent Work on Nanowire Lasers
39.7 Plasmonic Lasers 39-8
Bowtie Structures • Dipole Nanolaser
39.8 Photonic Crystal Lasers 39-10
Edge-Emitting Photonic Crystal Nanolaser • Silicon Nanocrystals
39.9 Scattering Lasers 39-10
Mie Nanolasers • Random Lasers
39.10 Organic Nanolasers 39-11
39.11 Applications 39-12
Nanolaser Device Detects Cancer in Single Cells • Biological and Chemical Detection • Ultrahigh Data Storage
39.12 Conclusions 39-12
Acknowledgment 39-12
References 39-12

Marek S. Wartak
Wilfrid Laurier University

39.1 Introduction: Properties of Laser Radiation

A laser is a device that emits light (or more generally electromagnetic (EM) radiation) through a process called stimulated emission. The term "laser" is an acronym for light amplification by stimulated emission of radiation [1]. The first working laser was demonstrated on May 16, 1960, by Theodore Maiman at Hughes Research Laboratories. It was a pulsed ruby laser. The main characteristics of laser beam are monochromaticity, coherence, divergence, and brightness.

Monochromaticity is associated with the fact that in a laser, only an EM radiation at a particular frequency (or wavelength) is amplified. In practice, it has a certain range, called linewidth, which is determined by homogeneous and inhomogeneous broadening factors. The linewidth is very small compared with typical light sources. Additionally, due to the laser cavity, which forms a resonant system, the oscillation can occur only at the resonance frequencies of this cavity. This leads to the further narrowing of laser linewidth, often by several orders of magnitude.

Coherence is associated with the phase difference. For any EM wave, there are two kinds of coherence, namely, spatial and temporal coherence. Typical laser source emission is very coherent. Most other light sources emit incoherent light, which has a phase that varies randomly with time and position.

Let us illustrate first the spatial coherence. For that, consider two points of EM wave, which, at time $t = 0$, have phase difference ϕ_0. If for any time $t > 0$ the phase difference of EM wave at those points remains ϕ_0, we say the EM wave has perfect coherence between the two points. If this is true for any two points of the wave front, we say the wave has perfect spatial coherence.

Now, consider a fixed point on the EM wave front. If at any time the phase difference between time t and time $t + \Delta t$ remains the same, where Δt is some time delay, we say that the EM wave has temporal coherence over a time Δt. If Δt can be any value, we say the EM wave has a perfect temporal coherence. If this happens only in a range $0 < \Delta t < t_0$, we say it has partial temporal coherence, with a coherence time equal to t_0. Laser light is highly coherent, and this property has been widely used in measurement, holography, etc.

Laser beam is also highly *directional*, which implies that laser light has very small divergence. This is a direct consequence of the fact that laser beam comes from the resonant cavity, and only waves propagating along the optical axis can be sustained in the cavity. The directionality is described by the light beam divergence angle.

The brightness of a light source is defined as the power emitted per unit surface area per unit solid angle. Brightness is inversely proportional to the square of laser divergence; therefore laser light is much brighter than normal light source.

39.2 Overview of Lasers

In gas lasers, the active medium can be regarded effectively as an ensemble of absorption or amplification centers (like, e.g., atoms or molecules) with only electronic energy levels, which couple to the resonant optical field. Other electronic states are used to excite or pump the system.

The generic laser structure is shown in Figure 39.1. It consists of a resonator (cavity) formed by two mirrors and gain medium where the amplification of electromagnetic radiation (light) takes place. A laser is an oscillator analogously like an oscillator in electronics and requires a resonator, which provides feedback. Feedback is provided by two mirrors. Mirrors confine light and provide optical feedback. One of the mirrors is partially transmitting, which allows the light to escape from the device. There must be an external energy provided into the gain medium (process known as pumping). The most popular (practical) pumping mechanisms are by optical or electrical means.

The gain medium can be created in several ways. Conceptually, the simplest one is the collection of gas molecules. The pumping process excites these molecules into a higher energy level. The popular visualization of such collection of molecules is as two-level systems (TLS) (see Figure 39.2). Only two energies (out of many in the case of a molecule) are selected and the transitions are considered within these two energies. As illustrated, three basic processes are possible: absorption, spontaneous emission, and stimulated emission.

Such TLS are found very often in nature. Generally, for an atomic system, for the case under consideration, we can always separate just two energy levels: upper level and ground state, thereby forming TLS.

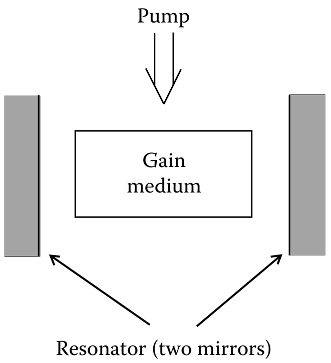

FIGURE 39.1 Generic laser structure: two mirrors with a gain medium in between. The two mirrors form a cavity, which confines the light and provides the optical feedback. One of the mirrors is partially transmitted and thus allows light to escape. The resulting laser light is directional, with a small spectral bandwidth.

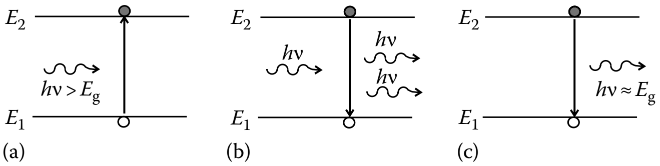

FIGURE 39.2 Possible electronic transitions in two-level system. (a) Absorption, (b) Stimulated emission, and (c) Spontaneous emission.

Electron can be excited into upper level due to external interactions (for lasers through a process known as pumping). Electrons can lose their energies radiatively (emitting photons) or nonradiatively, say by collisions with phonons.

For laser action to occur, the pumping process must produce population inversion meaning that there are more molecules in the excited state (here upper level with energy E_2) than in the ground state. If population inversion is present in the cavity, the incoming light can be amplified by the system (see Figure 39.2b) where one incoming photon generates two photons as the output.

The way how TLS is practically utilized results in various types of lasers, like gaseous, solid state, or semiconductor. Also, different types of resonators are possible as will be discussed in subsequent sections.

39.3 Semiconductor Lasers

A significant percentage of today's lasers are fabricated using the semiconductor technology. Those devices are known as semiconductor lasers. Over the last 15 years or so, several excellent books describing different aspects and different types of semiconductor lasers have been published [2,3].

The operation of semiconductor lasers as sources of electromagnetic radiation is based on the interaction between EM radiation and the semiconductor. Typical semiconductor laser structures are shown in Figure 39.3. These lasers can be classified as in-plane laser where light propagates in a parallel direction [4] and vertical cavity surface emitting laser (VCSEL) [5].

The largest dimension of in-plane structures is typically in the range of 250 μm (longitudinal direction) and as such cannot be considered as a nanolaser. The structure of interest to us is the one where light propagates perpendicularly to wafer's surface and it is known as VCSEL (see Figure 39.3b). The typical diameter of VCSEL cylinder is about 10 μm.

The basic semiconductor laser is just a p–n junction (see Figure 39.4) in which the cross-section along the lateral–transversal directions is shown. Current flows (holes on p-side and electrons on n-side) along the vertical direction, whereas the light travels horizontally and leaves the device on both sides.

Light propagation with amplification is illustrated in Figure 39.5. Mathematically it is described by assuming that there is no phase change on reflection at either end (left and right). The left end is defined as $z = 0$ and right end as $z = L$. At the right facet, the forward optical wave has a fraction r_R reflected (amplitude

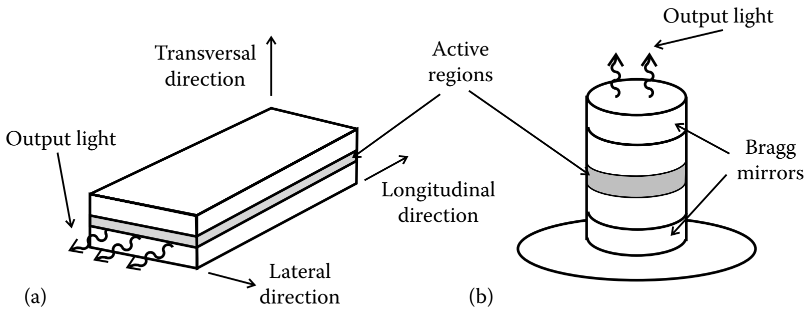

FIGURE 39.3 In-plane laser (a) and VCSEL (b).

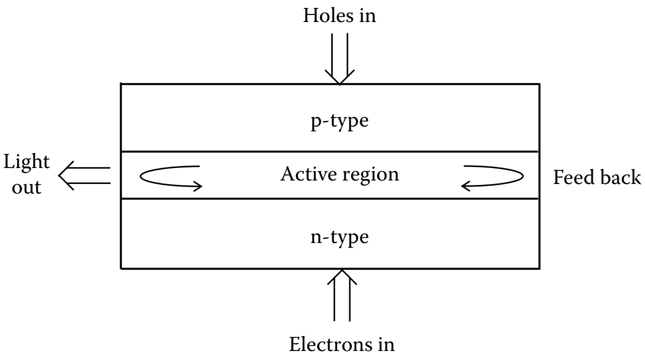

FIGURE 39.4 The basic p–n junction laser.

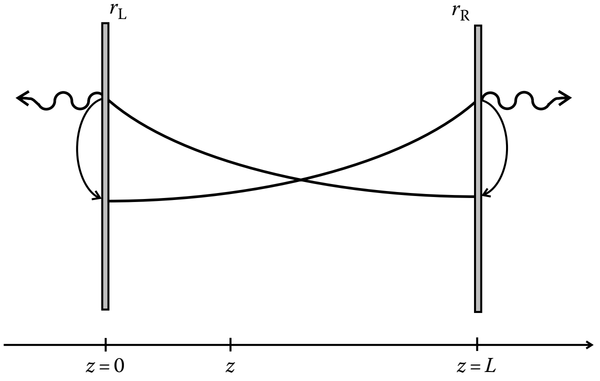

FIGURE 39.5 Schematic illustration of the amplification in a Fabry–Perot (FP) semiconductor laser with homogeneously distributed gain.

reflection) and after reflection the fraction travels back (from right to left).

In order to form a stable resonance, the amplitude and phase of the single round trip must match the amplitude and phase of the starting wave. At arbitrary point z inside the cavity (see Figure 39.5) the forward wave is

$$E_0 e^{gz}\ e^{-j\beta z}$$

where we have dropped e^{jwt} common term and $g = g_m - \alpha_m$. Here r_{R} and r_{L} are, respectively, right and left reflectivities, g is gain (and loss), L length of the cavity, and β the propagation constant.

The wave traveling one full round will be

$$\{E_0\, e^{gz}\ e^{-j\beta z}\}\ \{e^{g(L-z)} e^{-j\beta(L-z)}\}\{r_{\mathrm{R}} e^{gL} e^{-j\beta L}\}\times\{r_{\mathrm{L}} e^{gz} e^{-j\beta z}\} \tag{39.1}$$

The above terms are interpreted as follows. In the first bracket, there is an original forward propagating wave at z, in the second bracket, there is wave traveling from z to L, third bracket describes wave propagating from $z = L$ to $z = 0$, and the last one contains wave traveling from $z = 0$ to the starting point z. At that point, the wave must match original wave and thus one obtains condition for stable oscillations:

$$r_{\mathrm{R}} r_{\mathrm{L}} e^{2gL} e^{-2j\beta L} = 1 \tag{39.2}$$

That condition can be split into amplitude condition

$$r_{\mathrm{R}} r_{\mathrm{L}} e^{2(g_m-\alpha_m)L} = 1 \tag{39.3}$$

and phase condition

$$e^{-2j\beta L} = 1. \tag{39.4}$$

From the amplitude condition, the following relation is obtained:

$$g_m = \alpha_m + \frac{1}{2L}\ln\frac{1}{r_{\mathrm{R}} r_{\mathrm{L}}} \tag{39.5}$$

From the phase condition, it follows that

$$2\beta L = 2\pi n \tag{39.6}$$

where n is an integer. The last equation determines wavelengths of oscillations since

$$\beta = \frac{2\pi}{\lambda_m} \tag{39.7}$$

with λ_m being the wavelength.

In VCSEL, the cavity is formed by the so-called Bragg mirrors and an active region typically consists of several quantum well layers separated by barrier layers (see Figure 39.3). Bragg mirrors consist of several layers of different semiconductors, which have different values of refractive index. Due to the Bragg reflection,

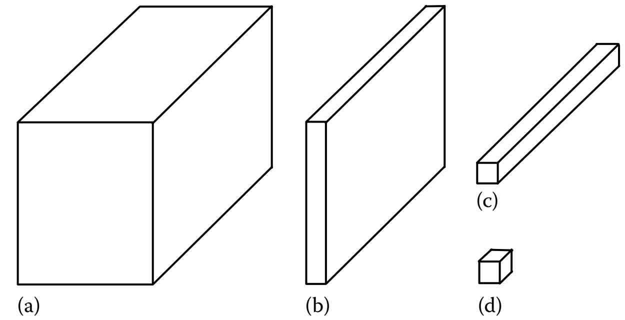

FIGURE 39.6 Illustration of various active regions with different dimensionality. (a) three-dimensional (3D) or bulk structure, (b) two-dimensional (2D) structure known as quantum well, (c) one-dimensional (1D) structure known as quantum wire, and (d) zero-dimensional (0D) structure known as quantum dot. (Adapted from Arakawa, Y. and Sakaki, H., *Appl. Phys. Lett.*, 40, 939, 1982.)

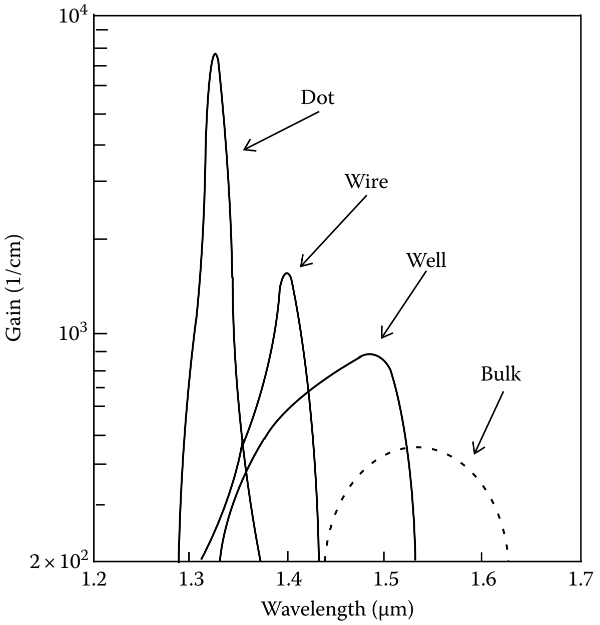

FIGURE 39.7 Calculated gain spectra for quantum dot, quantum wire and quantum well. For comparison, results for bulk crystal are also shown as dotted line. (Adapted from Asada, M. et al., *IEEE J. Quantum Electron.*, 22, 1915, 1986.)

such a structure shows a very large reflectivity (around 99.9%). Such large values are needed because a very short distance of propagation of light does not allow to build enough amplification when propagating between mirrors.

From the electromagnetic analysis of the optical cavity, the quality factor Q of the cavity [6] can be determined as

$$Q = \frac{\Delta\omega}{\omega_0} \tag{39.8}$$

where

$\Delta\omega$ is the cavity linewidth

ω_0 the resonant frequency of the cavity

Gain spectra (also other properties) strongly depend on size of the active region. Different gain media are possible, consisting of 0D, 1D, 2D, 3D (bulk) structures (see Figure 39.6). The typical thickness of a quantum well is about 10 nm. The dimensionality of these structures have profound consequences on laser properties. An illustration of gain spectra obtained with gain media for structures of different dimensionalities is shown in Figure 39.7. As can be seen, the shape of spectrum becomes sharper with increasing quantization dimension. This is due to the variation of the density of states.

39.4 Rate Equation Approach

In a typical laser, there exists two types of subsystems: photons and carriers. A quantitative description of a laser system is given in terms of rate equations. One introduces the number of photons S inside the cavity and number of carriers N (can be also a number of excited molecules for some systems). The rate equations describe the time evolution of S and N, as follows:

$$\frac{dN}{dt} = \eta_i \frac{I}{qV} - \frac{N}{\tau} - v_g g(N) S \tag{39.9}$$

$$\frac{dS}{dt} = \Gamma v_g g(N) S - \frac{S}{\tau_p} + \Gamma \beta R_{\text{sp}} \tag{39.10}$$

We have explicitly indicated that gain g depends on the carrier's concentration.

In Equation 39.9 the first term is responsible for pumping (in this case electrical), with I being the current, the second term accounts phenomenologically for losses, and the last one describes coupling to the photon system.

The last term in Equation 39.10 accounts for spontaneous emission (with the coefficient β). The β coefficient describes the amount of spontaneous emission that contributes to the lasing mode. The β factor is inversely proportional to the number of available modes into which the gain medium can spontaneously emit photons. In typical in-plane lasers, it is a small number (like 10^{-4}) [3]. The value of β is between 0 and 1.

The typical power–current characteristic of semiconductor laser is shown in Figure 39.8 for two extreme cases of β. When $\beta = 1$, all the spontaneously emitted photons end up in the lasing mode. For small values of β, the laser has a well-defined threshold. At threshold, the optical gain compensates for the losses. Above the threshold, the laser operates in a stimulated emission mode and below the threshold, spontaneous emission dominates.

A large β is the key factor in single-photon laser sources [7,8]. The (still hypothetical) case of $\beta = 1$ is often referred to as a thresholdless laser [9]. In the thresholdless laser, all photons participate in the stimulated emission. Such a device would require a small amount of energy to operate. Its dimensions should be very small, say ~λ.

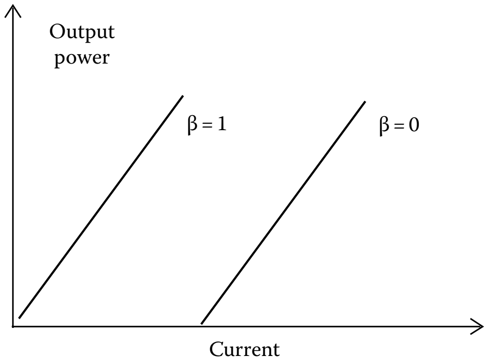

FIGURE 39.8 Current–power characteristics for two values of spontaneous emission factor.

Several issues must be addressed before one could fabricate thresholdless laser [10,11]. Those include: (a) optical modes that induce undesired spontaneous emission should be suppressed where possible, (b) creation of a single-cavity mode with a sufficiently high Q factor and a small modal volume is essential, and (c) excited carriers should be concentrated to emit light coupled to the single-mode cavity.

39.5 Definition of Nanolaser

The need to reduce the size of the semiconductor laser is one of the most active and challenging areas of modern optoelectronics. On the theoretical side, size reduction is important for understanding the basic laser concepts and fundamental light–matter interactions. On the practical side, smaller lasers will find various applications as light sources in integrated optical systems.

If photonics should be compatible with VLSI as for as lasers are concerned, photonic devices must shrink in size to 100 nm (or less) length scales. As indicated in the previous section, the smallest lasers available commercially today are VCSELs. However, in the last few years, the new type of even smaller devices is emerging, namely, nanowire lasers. Their cylindrical dimensions range from few tens up to hundreds of nanometers, whereas their lengths are typically within a few hundreds of microns.

The typical volume of conventional miniature lasers such as VCSELs is determined by the volume of the cavity mode. The effective wavelength in the dielectric should be of the order of the characteristic length of the device. This leads to the existence of the effective modal volume $V > (\lambda/2n)^3$, a condition known as the diffraction limit. Reduction of the volume of the active region is the potential factor for lowering the threshold current.

Recent advances in technology allow the fabrication of optical nanoscale devices where the wave nature of photons becomes one of the most critical variables. It provides the challenges in the realization of a tiny coherent photon source. The localization of the wave is difficult when wavelengths of photons become much larger than the spatial variation of the confinement structure.

As the dimensions of optical nanodevices scale down, devices can be fabricated with effectively only one optical emission mode. These structures could be termed *nanolasers* (see [8]). Alternatively, nanolasers can be defined as structures having dimensions smaller than the wavelength of light in all three dimensions [12]. Recent summaries of research on nanolasers can be found in Refs. [12,13].

In the following sections, we will discuss examples of recently the developed nanolasers (and related) structures:

- Nanowire lasers
- Plasmonic lasers
- Photonic crystal lasers
- Scattering lasers
- Organic lasers

We conclude this chapter with a brief discussion of applications of nanolasers in sensing and medicine.

39.6 Nanowire Lasers

Various types of nanocavities have been fabricated and a coherent laser emission from such structures has been observed. Among others, lasing has been demonstrated in droplets [14,15], silica [16], and polystyrene spheres [17], semiconductor microdisks [18–20], micropillars [21], photonic crystal cavities [22], nanoribbons [23], ZnO arrays [24], GaN nanowires [25], and single-crystal ZnO nanowires [26].

Several methods of the synthesis of semiconductor nanowire heterostructures have been developed, including chemical vapor deposition and the vapor–liquid–solid growth of crystalline semiconductor nanowires. Recent progress in the development of semiconductor nanowires was reviewed by Lauhon et al. [27].

The schematics of a typical cylindrical nanowire laser is shown in Figure 39.9 and the possible schemes of current injection are schematically shown in Figure 39.10. These structures provide photon confinement in volumes of a few cubic wavelengths. A typical lateral dimension of a nanowire is between 20 and 400 nm, with a length in the range of 2–40 μm. As was shown in the journal articles cited earlier, it is possible to grow nanowire arrays with tight control over size (diameter <20 nm) and uniformity (<±10%).

A significant reduction in size of nanometer laser was possible due to large spatial overlap between the active medium (the wire itself) and the guided mode, which propagates along the axial direction. A strong lateral optical confinement is also created. It originates due to the size of the structure itself and also due to large dielectric contrast (difference) between the nanowire material itself and the surrounding material. The end facets of the nanowire form axial Fabry–Perot-type cavity necessary for laser action.

Early experimental evidence for guided modes in nanowires surrounded by air has been reported by Johnson et al. [25] and Duan et al. [28]. Lasing was observed in structures that are rather thick (radius in the 80–200 nm range). In these structures, quantization effects are absent. Such structures can function as the active medium and waveguide at the same time. Structures

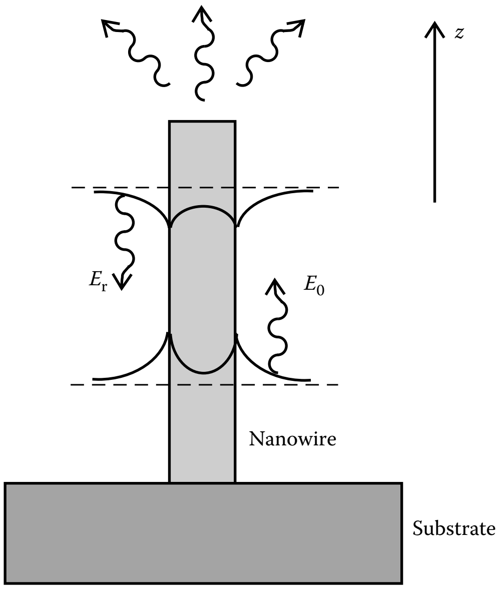

FIGURE 39.9 Side view of a nanowire laser. (Adapted from Maslov, A.V. and Ning, C.Z., *Appl. Phys. Lett.*, 83, 1237, 2003.)

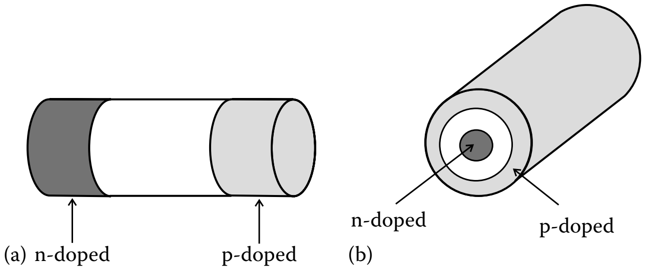

FIGURE 39.10 Possible schemes of current injection in nanowires: (a) through the ends of the nanowire and (b) from the core and shell regions. (Adapted from Maslov, A.V., *J. Appl. Phys.*, 99, 024314, 2006.)

with much smaller dimensions (radius about 1 nm) have also been fabricated [29]. In such samples, quantization effects play a dominant role. As an example, resonant nanowire cavity fabricated by Huang et al. [24] achieved gain and lasing through an excitonic lasing action in ZnO with a threshold of 40 kW/cm^2 under optical excitation. (In this case, because of optical pumping, the density of optical power is provided instead of value of electrical current.) A summary of the properties of basic materials used to fabricate nanowires is shown in Table 39.1.

TABLE 39.1 Basic Properties of Materials Used to Fabricate Nanowires

Materials	Emission Wavelength (nm)	Threshold (nJ/cm^{2})	Gain (cm^{-1})	Reference
GaN	370–390	500	400–1000	[85]
ZnO	370–400	70	1000–3000	[85]
CdS	510	—	—	[28]

39.6.1 Ring Resonator Nanolasers

Microrings are common building blocks in photonics [30]. Formation of nanoring resonator is also possible by converting the nanowire into a ring [31].

This change of shape results in a significant modification of the optical spectra of the cavity. For example, the photoluminescence displays Fabry–Perot (FP) resonances that match those calculated for ring resonance. Also, the dielectric imperfection at the overlapping junction causes each FP mode to split into a doublet, thus breaking the degeneracy between clockwise and anticlockwise photon propagation (see Figure 39.11b).

Microring nanolasers were fabricated and analyzed by Berkeley group led by Yang [31]. The ring structure was fabricated from a linear GaN nanowire (see Figure 39.11). We show here a GaN nanowire before manipulation into a ring configuration (a) and after creation of ring geometry (b). It is not a subwavelength device as can be seen from that figure. In the ring after coupling, a dielectric discontinuity is introduced (since each ends of the original wire are now positioned side by side). This side-by-side overlap is responsible for evanescent coupling between cavity arms (see Figure 39.12).

The circulating optical modes within a resonator cavity containing a defect are theoretically equivalent to a photonic molecule [32] (compare Figure 39.11c). Dielectric discontinuity after coupling leads to perturbation within the cavity and breaks the resonance degeneracy into clockwise and counterclockwise mode propagations. It is known that such perturbed cavity is theoretically equivalent to two perfect cavities that are coupled.

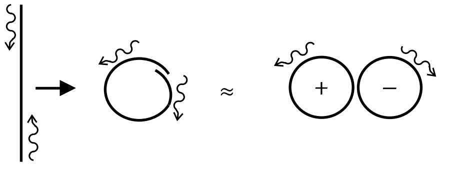

(a) Wire (b) Ring (c) Photonic molecule

FIGURE 39.11 Transition from wire linear cavity to a ring resonator and the theoretically equivalent photonic molecule. (Adapted from Pauzauskie, P.J. et al., *Phys. Rev. Lett.*, 96,143903, 2006.)

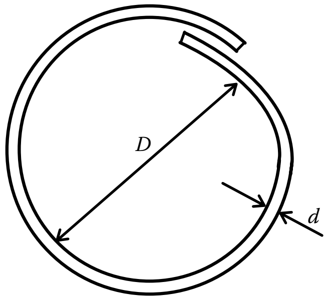

FIGURE 39.12 Cross-section view of the ring structure showing the side-by-side overlap that enables evanescent coupling between cavity modes.

The lasing behavior between nanowires and their resonator counterpart is also markedly different. The emission maximum of ring laser is red shifted up to 10 nm relative to the wire (see [31]).

39.6.2 Theoretical Analysis of Nanowires

For a mathematically oriented reader, we provide here the main elements of theory of nanolasers. This section can be skipped in the first reading.

Theoretical and numerical analysis of nanowire lasers was reported in a series of papers by Maslov and Ning [33–37] and also by Chen and Towe [38]. From experiments, we know that lasing occurs through the end facets of the nanowire (see Figure 39.9 (from Ref. [33].)

The nanowire can support transverse electric (TE_{0m}) and transverse magnetic (TM_{0m}) modes, which have only three field components and no dependence on ϕ and hybrid modes (HE_{nm} and EH_{nm}), which have all six field components [33]. In the axial approximation, the field components are of the type [34]

$$\tilde{E}_z^n(r) = \begin{cases} A_n J_n(\kappa_2 r), & r < r' \\ B_n J_n(\kappa_2 r) + C_n Y_n(\kappa_2 r), & r' < r < R \\ D_n H_n(\kappa_1 r), & r > R \end{cases} \quad (39.11)$$

$$\tilde{H}_z^n(r) = \begin{cases} F_n J_n(\kappa_2 r), & r < r' \\ G_n J_n(\kappa_2 r) + N_n Y_n(\kappa_2 r), & r' < r < R \\ M_n H_n(\kappa_1 r), & r > R \end{cases} \quad (39.12)$$

where J_n and Y_n are the Bessel functions of the first and second kind, respectively. $H_n = H_n^{(1)} = J_n + iY_n$ is the Bessel function of the third kind (Hankel's function). Also, the following definitions of transverse wave numbers were introduced: $\kappa_{1,2}^2 = \varepsilon_{1,2}\omega^2/c^2 - h^2$. The eight unknown coefficients A, B, C, D, F, G, N, and M should be determined from the boundary conditions at the interfaces $r = r'$ and $r = R$.

The numerical approach is based on solving Maxwell's equations using finite difference time-domain (FDTD) method in cylindrical coordinates [35]. Typical computational window is shown in Figure 39.13.

The modes supported in such nanowire structures are similar to those of optical fibers (for a discussion of modes in fiber, see [39]) but are more localized due to high refractive index contrast between the nanowires and the surrounding air.

The Maslov–Ning analysis suggests that the natural facets of the nanowires provide very low quality factors (of the order of hundreds) for nanowires of about 10 μm in length. These factors are very sensitive to the mode type and nanowire radius.

Far-field patterns of the emitted radiation were also discussed [34,35]. It was determined how the radiation pattern depends on the lasing mode. The radiation is emitted in a very broad range of angles with respect to the nanowire axis. Also, the directionality weakens with an increase of the nanowire radius.

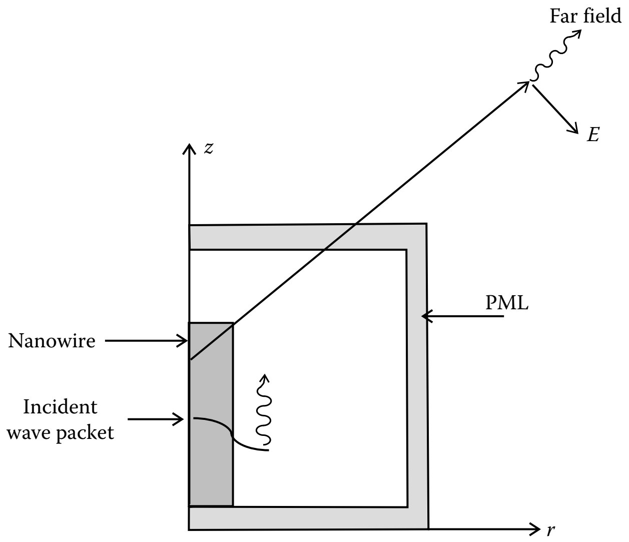

FIGURE 39.13 Schematic of the FDTD computational domain. (Adapted from Maslov, A.V. and Ning C.Z., *Opt. Lett.*, 29, 572, 2004.)

Using their methods, Maslov and Ning recently reported on the numerical analysis of semiconductor nanowire covered with a metal as a possible laser waveguide [40]. Their analysis opens the possibilities of fabricating even smaller nanowire-based lasers. Maslov and Ning analyzed the possible advantages of using a semiconductor nanowire encased in a metal as a laser waveguide. They showed that despite large Joule loss, such structure can be a good candidate for subwavelength laser operating in TM_{01} mode.

Coupled drift–diffusion simulations of nanowire lasers have been recently reported by Chen and Towe [38]. They extended the FDTD approach by including carrier effects. Their method is based on numerical solution of the steady-state 2D drift–diffusion carrier transport equations, which are coupled with the photon generation rate equations. The basic system of equations is as follows:

$$\nabla \cdot \left[-\varepsilon \nabla \psi(x,y) \right] = q\left[p(x,y) - n(x,y) + N_D^+ - N_A^- \right] \quad (39.13)$$

$$\nabla \cdot \mathbf{J}_n = -q\left[G(n,p) - R_{\mathrm{sp}}(n,p) - R_{\mathrm{st}}(n,p) - R_{\mathrm{Auger}}(n,p) - R_{\mathrm{SRH}}(n,p) \right] \quad (39.14)$$

$$\nabla \cdot \mathbf{J}_p = q\left[G(n,p) - R_{\mathrm{sp}}(n,p) - R_{\mathrm{st}}(n,p) - R_{\mathrm{Auger}}(n,p) - R_{\mathrm{SRH}}(n,p) \right] \quad (39.15)$$

$$\mathbf{J}_n = qD_n \nabla n(x,y) - q\mu_n n(x,y) \nabla \psi(x,y) \quad (39.16)$$

$$\mathbf{J}_p = -qD_p \nabla p(x,y) - q\mu_p p(x,y) \nabla \psi(x,y) \quad (39.17)$$

$$G_m S_m - \frac{S_m}{\tau_{p,m}} + R_{sp,m} = 0 \quad (39.18)$$

where $G(n, p)$ is the carrier generation rate, $R_{\mathrm{sp}}(n, p)$ is the local spontaneous recombination rate, $R_{\mathrm{st}}(n, p)$ is the local stimulated recombination rate, $R_{\mathrm{SRH}}(n,p)$ represents Shockley–Read–Hall (SRH) dark recombination rate, G_m, S_m, $\tau_{p,m}$, $R_{sp,m}$ are, respectively, the modal gain, the photon number in the cavity, photon lifetime and for the *m*th lasing mode, and the spontaneous emission rate that couples to the *m*th lasing mode.

39.6.3 Recent Work on Nanowire Lasers

We conclude the section on nanowires with a description of recent works that go beyond single cylindrical nanowire. First, we describe nanowires with metal coating and then the combination of nanowires and quantum wells.

The dimensions of semiconductor nanolasers can be further shrunk by using metal-coated nanowires. Such a structure was first fabricated in 2007 by Hill et al. [41] and this structure was shown to be the smallest electrically pumped nanolaser. They coated a thin semiconductor heterostructure post made of In/InGaAs/InP with a layer of gold as shown in Figure 39.14. The diameter was 210 nm. Electrons are injected through the top of the pillar and holes are injected through a large-area lateral contact.

The device lased in the near-infrared at a wavelength of 1408 nm. The mode was an HE_{11}-like mode oscillating near the InGaAs region cutoff frequency.

All previously described studies have concentrated on fabrication of nanolasers from homogeneous semiconductors such as gallium nitride (GaN). This means that the laser wavelength is determined by the band gap of the used material. In such a design, there is no possibility to tune the properties of the laser.

However, with proper technological tools, structures in which gain and cavity functionalities are decoupled can be designed and fabricated. The purpose of decoupling the gain medium and the cavity is not to separate but to combine the best properties of both subsystems. Thus quantum wells provide the optical gain medium whereas the nanowire acts as the optical cavity.

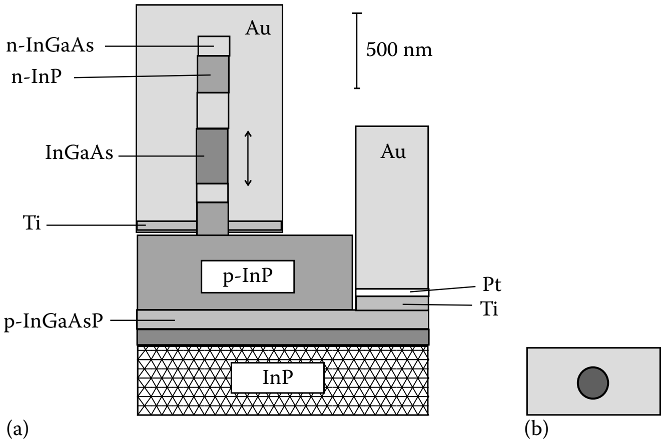

FIGURE 39.14 (a) Structure of the cavity formed by a semiconductor pillar encapsulated in gold. (b) Cross-section of the pillar. (Adapted from Hill, M.T. et al., *Nat. Photonics*, 1, 589, 2007.)

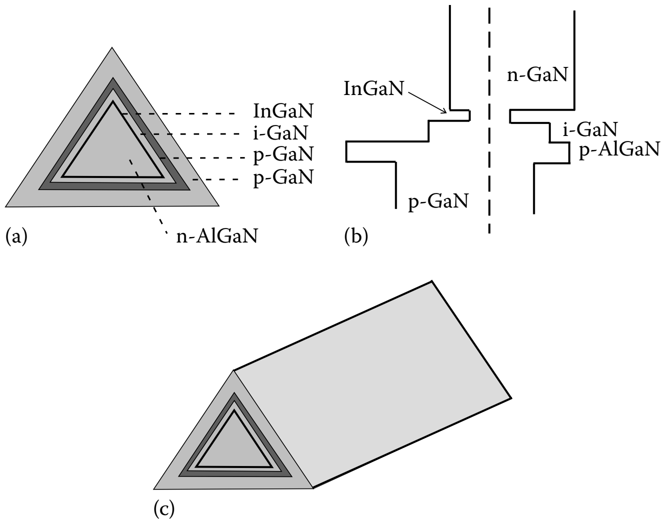

FIGURE 39.15 Cross-sectional view of a (a) GaN-based core/multi-shell nanowire structure and (b) the corresponding energy band diagram. (c) InGaN multiple quantum well (MQW) nanowire structure. (Adapted from Qian, F. et al., *Nat. Mater.*, 7, 701, 2008; Qian, F. et al., *Nano Lett.*, 5, 2287, 2005.)

This new type of nanolasers was described by Qian et al. [42]. The structure is shown in Figure 39.15. It consists of a GaN nanowire core that acts as the optical cavity surrounded by InGaN/GaN multiple quantum well (MQW) shells that serve as a composition-tunable gain medium. By varying the indium content, the emission wavelength can be tuned between 365 and 494 nm, with all devices operating at room temperature. Their nanowire heterostructure contains 3 to 26 quantum wells. Typical nanowires were 200–400 nm in "diameter" and 20–60 μm in length. Although the nanolasers were optically pumped, the authors believe that electrical injection is possible.

39.7 Plasmonic Lasers

Surface plasmon mode confinement has been used to achieve subwavelength modal dimensions at the expense of optical loss [43–45]. Several structures that involve plasmons were proposed. A comprehensive introduction to the main physical aspects involved in plasmonic devices were recently summarized by Dragoman and Dragoman [46].

A summary of the analysis of the plasmonic nanoresonators has recently been provided by Maier [47]. Plasmon cavities can confine electromagnetic energy into both physical and effective mode volumes far below the diffraction limit.

Manolatou et al. [48,49] proposed and analyzed a family of nanoscale cavities for electrically pumped surface-emitting semiconductor lasers that use surface plasmons to provide optical confinement. The analyzed cavities have radii between 100 and 300 nm and the heights of the dielectric part of the cavity between 100 and 250 nm. The metal layers were assumed to be thick enough to prohibit light transmission. The typical circular nanopatch laser analyzed by Manolatou and Rana [49] is shown in Figure 39.16.

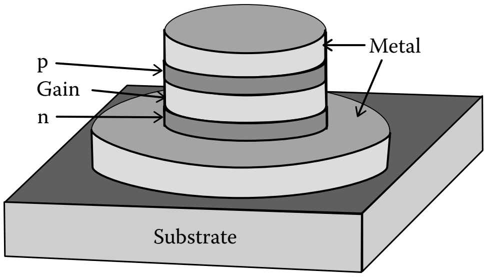

FIGURE 39.16 Circular nanopatch laser.

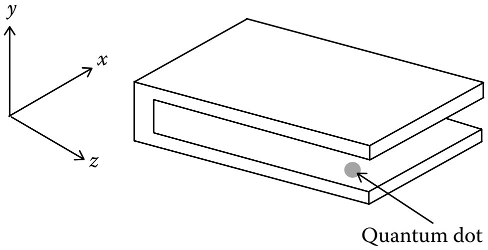

FIGURE 39.17 Horseshoe optical nanoantenna. (Adapted from Sarychev, A.K. and Tartakovsky, G., *Phys. Rev. B*, 75, 085436, 2007.)

Sarychev and Tartakovsky [50,51] proposed plasmonic nanolaser where the metal nanoantenna operates similarly to a resonator. The structure is shown in Figure 39.17. In this type of laser, metallic horseshoe-shaped nanoantenna interacts with a two-level amplifying system (TLS). TLS can represent quantum dot and can be pumped optically or electrically. The size of the proposed plasmonic nanolaser is much smaller than the wavelength.

39.7.1 Bowtie Structures

Bowtie structure consists of two opposing tip-to-tip metallic nanotriangles separated by a gap with an active element in the form of quantum dot in-between or quantum wells below it. The interaction of a single quantum dot with a bowtie antenna was demonstrated by Farahani et al. [52] for visible light. The enhancement of the electromagnetic field in such a structure was described by Sundaramurthy et al. [53]. They also conducted FDTD analysis. The theory of the electrically pumped plasmonic nanolaser based on bowtie structure was recently reported by Chang et al. [54]. They considered both quantum dot and quantum well.

The geometry of a bowtie nanolaser is shown in Figure 39.18. It consists of a metallic bowtie separated (typically) by about 20 nm. Multiple quantum wells are located below the metallic bowtie, which at optical frequencies have negative dielectric constant. The bowtie tips reduce the effective volume of the cavity mode and lead to the field enhancement around the bowtie tips. This results in an increase of the stimulated and spontaneous emission rates and significant decrease of the threshold current.

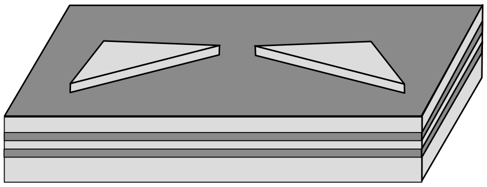

FIGURE 39.18 The geometry of a bowtie plasmonic nanolaser. Typical separation between metallic nanotriangles is 20 nm. (Adapted from Chang, S.-W. et al., *Opt. Express*, 16, 10580, 2008.)

39.7.2 Dipole Nanolaser

Diple nanolaser (DNL) has been proposed by Protsenko et al. [55,56]. It consists of a metallic nanoparticle of size r_0 and a two-level system (TLS) of size r_2 formed by a quantum dot, separated by a small distance r (see Figure 39.19). The device does not need an optical cavity and may have a volume much smaller than λ, the lasing wavelength. An incoherent pump provides population inversion in the TLS. The system can be pumped optically or electrically, in which case pumping is provided by injection of carriers from the bands of a semiconductor material surrounding an embedded quantum dot.

The transition frequency of TLS ω_2 is close to the plasmon resonance frequency ω_p of the metallic nanoparticle. There is a strong interaction (a dipole interaction) between those two systems through a near field. It is known that such coupled interaction significantly modifies optical emission.

To understand the operation of DNL, let us remind ourselves that the usual lasing mechanism involves stimulated emission into the mode of a cavity (for the single mode operation) from a medium in which there is a population inversion. Above the threshold condition, when stimulated emission (gain) exceeds absorption plus internal losses, the energy from incoherent pump is transferred into coherent laser radiation with a narrow lasing spectrum.

In DNL, instead of electromagnetic field, one deals with linear oscillations of polarization of a medium. Such a medium can be excited in such a way that the total energy flux into polarization exceeds losses. This leads to polarization oscillations with a narrow spectrum. The analysis of dipole nanolaser conducted by Protensko et al. [55] is based on equations of motion identical to Maxwell–Bloch equations for a TLS in the electromagnetic field of a cavity. Lasing conditions are created when the population

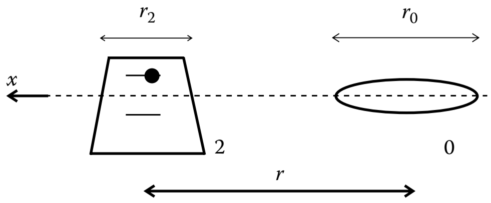

FIGURE 39.19 Dipole nanolaser. Nanoparticle (here on the right) is labeled with "0," TLS (here on the left) has label "2." (Adapted from Protsenko, I.E. et al., *Phys. Rev. A*, 71, 063812, 2005.)

inversion D of energy states of TLS is maintained by optical pumping or injection current to be [56]

$$D > D_{th} \tag{39.19}$$

where $D_{th} = (r/r_{cr})^6$, $r_{cr} = (4|\alpha_0||\alpha_2|)^{1/6}$ and α_0 and α_2 are polarizabilities of the nanoparticle and the TLS. When D exceeds the threshold value D_{th}, the macroscopic dipole moment of two particles appears spontaneously and they emit coherent dipole radiation at the frequency $\omega \approx \omega_{plr}$, the nanoparticle plasmon resonance frequency. Condition (39.19) is fulfilled if the distance between particles is small, which means a strong interaction between them through the near field.

39.8 Photonic Crystal Lasers

Photonics crystal (PC) structure consists of a drilled repeating pattern of holes through the laser material. This pattern is called a photonic crystal. One can deliberately introduce an irregularity, or defect, into the crystal pattern, for example, by slightly shifting the positions of two holes. The photonic crystal structure and the defect prevent light of most frequencies from existing in the structure, with the exception of a small band of frequencies that can exist in the region near the defect.

Photonic crystal defect microcavities can provide extremely small mode volume [22,57]. The schematic of the laser proposed by Painter et al. [22] is shown in Figure 39.20. The interhole separation is 515 nm. It was fabricated from InGaAsP grown by MOCVD on an InP substrate. The active region consists of four (here only two are shown) 9 nm 0.85% compressively strained InGaAsP quantum wells. Two-dimensional photonic crystal hexagonal lattice was formed by etching, resulting in air holes that penetrate through the active region and into an underlying sacrificial InP layer.

Over the last few years, various devices have been fabricated and characterized [58–61]. Such structures are associated with single-photon sources, which are required for quantum computing and quantum communications. Single-photon sources are recent applications of the Purcell effect in quantum-dot microcavities.

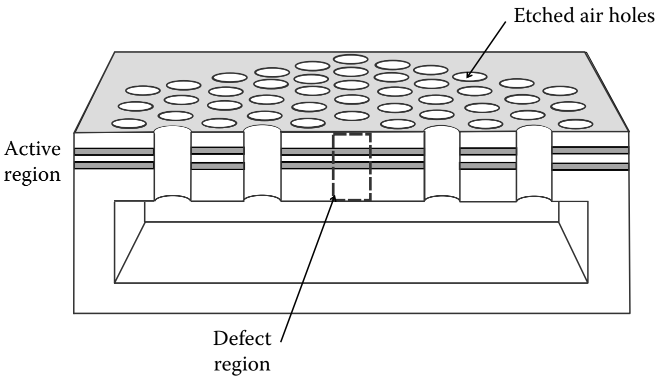

FIGURE 39.20 Cross-section through the middle of the photonic crystal microcavity. A defect is formed (shown in the middle) by removing a single hole. (Adapted from Painter, O. et al., *Science*, 284, 1819, 1999.)

Recently, scientists from the Yokohama National University in Japan [61] have reported interesting results concerning PC nanolasers. They demonstrated high-performing room-temperature nanolaser in the form of PC slab. The laser is made of a GaInAsP. This ultrasmall laser has a modal volume close to the diffraction limit. When operating in a high-Q mode (about 20,000), it will be useful for optical devices in optical integrated circuits. In a moderate-Q (1500) configuration, the nanolaser needs only an extremely small amount of external power to bring the device to the threshold of producing laser light. In this near-thresholdless operation, it might permit the emission of very low light levels, even single photons.

39.8.1 Edge-Emitting Photonic Crystal Nanolaser

Tiny resonators and waveguides based on photonic crystals provide a promising approach to fabricating high-density photonic integrated circuits. Recently, Yang et al. [62] reported the first demonstration of an edge-emitting photonic crystal nanocavity laser integrated with a photonic crystal waveguide. The structure is based on a double-heterostructure photonic crystal nanocavity with a InAs quantum dot active region. The device consists of four photonic crystal sections, each with a slightly different lattice constant. Five layers of InAs quantum dots, each with a quantum dot density of about $2 \times 10^{10}/\text{cm}^2$, were embedded in the 220 nm-thick GaAs membrane. An output waveguide is butt coupled to the mirror with the fewer number of periods. The device was optically pumped with a semiconductor laser diode at 850 nm wavelength. The threshold peak pump power absorbed by the cavity was estimated to be 12 μW.

39.8.2 Silicon Nanocrystals

Compatibility with CMOS materials have stimulated research on nanoscale silicon laser by Jaiswal and Norris [63]. They have theoretically analyzed and numerically simulated various designs based on photonic crystal concept and utilizing Si nanocrystals embedded in SiO_2. Their studies were motivated by observations of optical gain in silicon nanocrystals [64].

Pavesi et al. [64] observed light amplification in silicon itself. Silicon nanocrystals in the form of quantum dots were dispersed in a silicon dioxide matrix. Net optical gain was seen in both waveguide and transmission configurations, with the material gain being of the same order as that of direct band gap quantum dots. Their findings open the possibility for the fabrication of a silicon laser.

39.9 Scattering Lasers

In scattering lasers, the feedback is provided through scattering of light instead of a cavity [65]. These lasers are of two basic types: Mie lasers consisting of only one sphere and random lasers that are formed by many scattering particles. In Mie lasers, the surface of the sphere serves as multiple scatterer, while in random lasers, scattering is provided by randomly distributed particles.

In both types of scattering lasers, the average particle size used in typical experiments was about 5–10 µm [66,67]. They are, therefore, not of the size of true nanolasers.

39.9.1 Mie Nanolasers

In a Mie nanolaser, the mirrors of a conventional laser are replaced by the boundary of a microsphere. Light is multiply scattered at the boundary, and along the boundary, whispering gallery modes (WGMs) at a certain wavelength exist for specific sizes of the sphere.

WGMs occur at particular resonant wavelengths of light for a given droplet size (see Figure 39.21). At these wavelengths, the light undergoes total internal reflection at the particle surface and, after one roundtrip, interferes constructively. It becomes trapped within the particle for timescales of the order of nanoseconds.

The nomenclature of these modes derive from the observation of Lord Rayleigh in the dome in St. Paul's Cathedral in London. He observed sound ("whispers") propagating along the walls and circling around the dome several times.

Theoretical and experimental study of spherical resonators, which form the basis of Mie lasers, is still an open field of research. This interest stems from the analysis of fundamental processes such as scattering, energy propagation through disordered media, and cavity quantum electrodynamics, and from the large number of applications in photonics, chemistry, meteorology, astronomy, and sensing.

39.9.2 Random Lasers

In random lasers, the conventional optical cavity is replaced by light scattering from many particles. They do not have mirrors or optical elements.

In random lasers, the feedback is provided by multiple scattering of light at many scattering points. Random laser does not have a clear feedback mechanism like a conventional laser. It is nonresonant, even disordered and works in qualitatively different way than feedback by a resonant cavity (see [68] for a recent review). A comparison of conventional and random laser is shown in Figure 39.22. The wavelengths of random lasers span from the ultraviolet to the mid-infrared region. The materials used in random lasers include inorganic dielectrics, semiconductors, polymers, and liquids. The size of random lasers can vary from a cubic micrometer to hundreds of cubic millimeters.

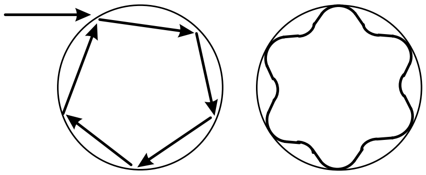

FIGURE 39.21 Whispering gallery mode inside a microsphere.

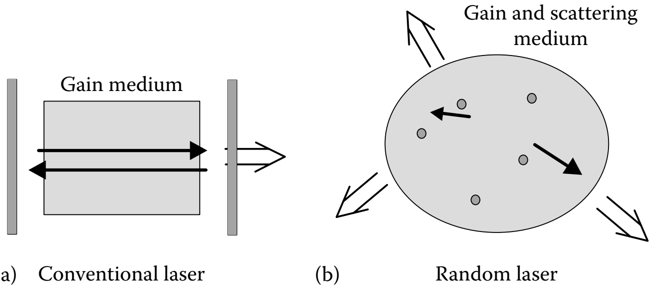

FIGURE 39.22 Comparison of conventional and random laser. (Adapted from Wiersma, D., *Nature*, 406, 132, 2000.)

A short history of the development of random lasers is as follows:

- In 1968, Letokhov [69] described an idea of scattering with "negative absorption."
- In 1994, Lawandy et al. [70] conducted experiments on scatterers in laser dye and observed threshold and line narrowing.
- In 1995, Wiersma et al. [67] conducted multiple scattering experiments in Ti:Al_2O_3 random lasers.

Recently a definite reference on solid state random lasers was published by Noginov [71] who tried to answer the question of what is a random laser by reviewing many types of random lasers. On the theoretical side, we mention two approaches to describe scattering lasers [72,73].

Fratalocchi et al. [72] derived a 3D vector set of Maxwell–Bloch equations, which were solved numerically by employing FDTD method. They performed a series of numerical experiments by investigating the process of laser emission from a single nanosphere covered by a layer of active material for different pumping rates.

Novel theoretical approach to diffusive random lasers has recently been developed by Tuereci et al. [73]. They developed a model by considering all the possible ways in which light can reflect back and forth in the medium. In their approach, they considered coexistence of gain, nonlinear interactions, and overlapping resonances.

39.10 Organic Nanolasers

Ease of fabrication make organic materials attractive for various optoelectronic devices, including lasers. O'Carroll et al. [74] used the flexibility of structuring organic materials to fabricate subwavelength optical nanowire lasers.

Using a melt-assisted template wetting method, they synthesized arrays of semicrystalline nanowires with diameters in a range of 150–400 nm, and with typical length of ~6 µm. Their structures are cylindrical wires with optically flat end facets, which form optical cavities. Lasers were optically pumped by an external pump laser. The observed that lasing wavelength (about $\lambda = 460$ nm) was determined by a standard standing wave condition.

Organic lasers can potentially have various applications [75] as inexpensive and lightweight sources of coherent radiation.

39.11 Applications

To conclude this chapter, we briefly discuss a few recent applications of nanolasers focused on nanowire and photonic crystal-based devices.

39.11.1 Nanolaser Device Detects Cancer in Single Cells

Investigators at Sandia National Laboratories in New Mexico reported a method to rapidly detect cancer in a single cell using nanolaser techniques [76,77]. Their technique, which allows the investigator to distinguish between malignant and normal cells, has the potential of detecting cancer at a very early stage, a development that could change profoundly the way cancer is diagnosed and treated.

The method of determination of malignant cells is based on a technique that rapidly assesses the properties of cells flown through a nanolaser. They observed biophotonic differences in normal and cancer mouse liver cells by using intracellular mitochondria as biomarkers for disease. This difference arises from the fact that mitochondria, the internal organelles that produce energy in a cell, are scattered in a chaotic, unorganized manner in malignant cells, while they form organized networks in healthy cells. This difference produces a marked change in the way that malignant cells scatter laser light.

39.11.2 Biological and Chemical Detection

Optofluidic integration of a new type of photonic crystal nanolaser incorporated into a microfluidic chip was described by Kim et al. [78]. The proposed nanolasers are an ideal platform for high-fidelity biological and chemical detection tools in micro-total-analytical or lab-on-a-chip systems. Its operation is based on the wavelength tunability induced by the refractive index variation of the fluid.

A record high sensitivity utilizing photonic crystal nanolaser was recently reported by Kita et al. [79]. The index resolution limit of their sensor could be $<10^{-6}$.

39.11.3 Ultrahigh Data Storage

Ikkawi et al. [80] reported the development of a near-field optical system capable of focusing light into a spot with diameter of less than 30 nm. To fabricate this type of nanolaser, they deposited a 100 nm thick aluminum film on the emitting edge of a laser diode and then used a focused ion beam to etch various apertures into the film. Their nanolaser technique is capable of storing data beyond 10 Tbit/in^2.

39.12 Conclusions

In this chapter, we provided an introduction to the concept of nanolasers, which are considered to be applicable as main elements, among others, in nanoscale optoelectronics. We discussed numerous examples of recently developed nanolasers.

The field of nanolasers shows rapid progress both on the research side as well as in applications. Examples include nanowire lasers with diameters as small as 100 nm. They find applications in medicine and in biological and chemical sensing. There is no doubt that field of nanolasers is set for interesting research in future and that nanolaser devices will find new and yet unforeseen applications.

Acknowledgment

I would like to acknowledge the financial support provided by the Natural Science and Engineering Research Council of Canada (NSERC).

References

1. A. Yariv. *Quantum Electronics*, 3rd edn. Wiley, New York, 1989.
2. B. Mroziewicz, M. Bugajski, and W. Nakwaski. *Physics of Semiconductor Lasers*. Polish Scientific Publishers/North-Holland, Warszawa, Amsterdam, the Netherlands, 1991.
3. L.A. Coldren and S.W. Corzine. *Diode Lasers and Photonic Integrated Circuits*. Wiley, New York, 1995.
4. G.P. Agrawal and N.K. Dutta. *Semiconductor Lasers*, 2nd edn. Kluwer Academic Publishers, Boston, MA/Dordrecht, the Netherlands/London, U.K., 2000.
5. C.W. Wilmsen, H. Temkin, and L.A. Coldren, (eds.). *Vertical-Cavity Surface-Emitting Lasers*. Cambridge University Press, Cambridge, U.K., 1999.
6. R. Coccioli, M. Boroditsky, K.W. Kim, Y. Rahmat-Samii, and E. Yablonovitch. Smallest possible electromagnetic mode volume in a dielectric cavity. *IEE Proc. Optoelectron.*, 145:391–397, 1998.
7. C. Santori, D. Fattal, J. Vuckovic, G.S. Solomon, and Y. Yamamoto. Indistinguishable photons from a single-photon device. *Nature*, 419:594–597, 2002.
8. S. Strauf, K. Hennessy, M.T. Rakher, Y.-S. Choi, A. Badolato, L.C. Andreani, E.L. Hu, P.M. Petroff, and D. Bouwmeester. Self-tuned quantum dot gain in photonic crystal lasers. *Phys. Rev. Lett.*, 96:127404, 2006.
9. H. Yokoyama. Physics and device applications of optical microcavities. *Science*, 256:66–70, 1992.
10. E. Yablonovitch. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.*, 58:2059–2062, 1987.
11. S. Noda. Seeking the ultimate nanolaser. *Science*, 314:260–261, 2006.
12. P.L. Gourley. Nanolasers. *Sci. Am.*, 278:56–61, 1998.
13. S.S. Mao. Nanolasers: Lasing from nanoscale quantum wires. *Int. J. Nanotechnol.*, 1:42–85, 2004.
14. S.-X. Qian, J. B. Snow, H.-M. Tzeng, and R.K. Chang. Lasing droplets: Highlighting the liquid–air interface by laser emission. *Science*, 231:486–488, 1986.

15. H.B. Lin, J.D. Eversole, and A.J. Campillo. Spectral properties of lasing microdroplets. *J. Opt. Soc. Am. B*, 9:43–50, 1992.
16. M. Cai, O. Painter, K.J. Vahala, and P.C. Sercel. Fiber-coupled microsphere laser. *Opt. Lett.*, 25:1430–1432, 2000.
17. M. Kuwata-Gonokami, K. Takeda, H. Yasuda, and K. Ema. Laser emission from dye-doped polystyrene microsphere. *Jpn. J. Appl. Phys. Lett.*, 2, 32:L99–L101, 1992.
18. B. Gayral, J.M. Gerard, A. Lemaitre, C. Dupuis, L. Manin, and J.L. Pelouardb. High-q wet-etched GaAs microdisks containing InAs quantum boxes. *Appl. Phys. Lett.*, 75:1908–1910, 1999.
19. S.L. McCall, A.F.J. Levi, R.E. Slusher, S.J. Pearton, and R.A. Logan. Whispering-gallery mode microdisk lasers. *Appl. Phys. Lett.*, 60:289–291, 1992.
20. N.B. Rex, R.K. Chang, and L.J. Guido. Threshold lowering in GaN micropillar lasers by means of spatially selective optical pumping. *IEEE Photon. Technol. Lett.*, 13:1–3, 2001.
21. J.L. Jewell, S.L. McCall, Y.H. Lee, A. Scherer, A.C. Gosasard, and J.H. English. Lasing characteristics of GaAs microresonators. *Appl. Phys. Lett.*, 54:1400–1402, 1989.
22. O. Painter, R.K. Lee, A. Scherer, A. Yariv, J.D. O'Brien, P.D. Dapkus, and I. Kim. Two-dimensional photonic band-gap defect mode laser. *Science*, 284:1819–1821, 1999.
23. H. Yan, J. Johnson, M. Law, R. He, K. Knutsen, J.R. McKinney, J. Pham, R. Saykally, and P. Yang. ZnO nanoribbon microcavity lasers. *Adv. Mater.*, 15:1907–1911, 2003.
24. M.H. Huang, S. Mao, H. Feick, H. Yan, Y. Wu, H. Kind, E. Weber, R. Russo, and P. Yang. Room-temperature ultraviolet nanowire nanolasers. *Science*, 292:1897–1899, 2001.
25. J.C. Johnson, H.-J. Choi, K.P. Knutsen, R.D. Schaller, P. Yang, and R.J. Saykally. Single gallium nitride nanowire lasers. *Nat. Mater.*, 1:106–110, 2002.
26. M.A. Zimmler, J. Bao, F. Capasso, S. Mller, and C. Ronning. Laser action in nanowires: Observation of the transition from amplified spontaneous emission to laser oscillation. *Appl. Phys. Lett.*, 93:051101, 2008.
27. L.J. Lauhon, M.S. Gudiksen, and C.M. Lieber. Semiconductor nanowire heterostructures. *Phil. Trans. R. Soc. Lond. A*, 362:1247–1260, 2004.
28. X. Duan, Y. Huang, R. Agarval, and C.M. Lieber. Single-nanowire electrically driven lasers. *Nature*, 421:241, 2003.
29. Y. Gu, I.L. Kuskovsky, M. Yin, S. O'Brien, and G.F. Neumark. Quantum confinement in ZnO nanorods. *Appl. Phys. Lett.*, 85:3833, 2004.
30. C.R. Pollock and M. Lipson. *Integrated Photonics*. Kluwer Academic Publishers, Boston, MA, 2003.
31. P.J. Pauzauskie, D.J. Sirbuly, and P. Yang. Semiconductor nanowire ring resonator laser. *Phys. Rev. Lett.*, 96:143903, 2006.
32. M. Bayer, T. Gutbrod, J.P. Reithmaier, and A. Forchel. Optical modes in photonic molecules. *Phys. Rev. Lett.*, 81:2582–2585, 1998.
33. A.V. Maslov and C.Z. Ning. Reflection of guided modes in a semiconductor nanowire laser. *Appl. Phys. Lett.*, 83:1237–1239, 2003.
34. A.V. Maslov, M.I. Bakunov, and C.Z. Ning. Distribution of optical emission between guided modes and free space in a semiconductor nanowire. *J. Appl. Phys.*, 99:024314, 2006.
35. A.V. Maslov and C.Z. Ning. Far-field emission of a semiconductor nanowire laser. *Opt. Lett.*, 29:572–574, 2004.
36. A.V. Maslov and C.Z. Ning. Modal gain in a semiconductor nanowire laser with anisotropic bandstructure. *IEEE J. Quantum Electron.*, 40:1389–1397, 2004.
37. A.V. Maslov and C.Z. Ning. Band structure and optical absorption of GaN nanowires grown along c axis. *Phys. Rev. B*, 72:125319, 2005.
38. L. Chen and E. Towe. Coupled optoelectronic modeling and simulation of nanowire lasers. *J. Appl. Phys.*, 100:044305, 2006.
39. J.W. Buck. *Fundamentals of Optical Fibers*, 2nd edn. Wiley-Interscience, Hoboken, NJ, 2004.
40. A.V. Maslov and C.Z. Ning. Metal-encased semiconductor nanowires as waveguides for ultra-small lasers. In *Conference on Lasers and Electro-Optics, CLEO*, 6–11 May 2007, Baltimore, MD, pp. 1–2, 2007.
41. M.T. Hill, Y.-S. Oei, B. Smalbrugge, Y. Zhu, T. De Vries, P.J. Van Veldhoven, F.W.M. Van Otten, T.J. Eijkemans, J.P. Turkiewicz, H. De Waardt, E.J. Geluk, S.-H. Kwon, Y.-H. Lee, R. Noetzel, and M. K. Smit. Lasing in metallic-coated nanocavities. *Nat. Photon.*, 1:589–594, 2007.
42. F. Qian, Y. Li, S. Gradecak, H.-G. Park, Y. Ding, Z.L. Wang, and C.M. Lieber. Multi-quantum-well nanowire heterostructures for wavelength-controlled lasers. *Nat. Mater.*, 7:701–706, 2008.
43. W.L. Barnes, A. Dereux, and T.W. Ebbesen. Surface plasmon subwavelength optics. *Nature*, 424:824–830, 2003.
44. M.L. Brongersma, J.W. Hartman, and H.A. Atwater. Electromagnetic energy transfer and switching in nanoparticle chain arrays below the diffraction limit. *Phys. Rev. B*, 62:R16358, 2000.
45. J.A. Dionne, L.A. Sweatlock, H.A. Atwater, and A. Polman. Plasmon slot waveguides: Towards chip-scale propagation with subwavelength-scale localization. *Phys. Rev. B*, 73:035407, 2006.
46. M. Dragoman and D. Dragoman. Plasmonics: Applications to nanoscale terahertz and optical devices. *Prog. Quantum Electron.*, 32:1–41, 2008.
47. S.A. Maier. Effective mode volume of nanoscale plasmon cavities. *Opt. Quantum Electron.*, 38:257–267, 2006.
48. F. Rana, C. Manolatou, and S.G. Johnson. Nanoscale semiconductor plasmon lasers. In *Conference on Lasers and Electro-Optics, CLEO*, 6–11 May 2007, Baltimore, MA, pp. 1–2, 2007.
49. C. Manolatou and F. Rana. Subwavelength nanopatch cavities for semiconductor plasmon lasers. *IEEE J. Quantum Electron.*, 44:435–447, 2008.
50. A.K. Sarychev and G. Tartakovsky. Magnetic plasmonic metamaterials in actively pumped host medium and plasmonic nanolaser. *Phys. Rev. B*, 75:085436, 2007.

51. A.K. Sarychev, A.A. Pukhov, and G. Tartakovsky. Metamaterial comprising plasmonic nanolaser. *PIERS Online*, 3:1264–1267, 2007.
52. J.N. Farahani, D.W. Pohl, H.-J. Eisler, and B. Hecht. Single quantum dot coupled to a scanning optical antenna: A tunable superemitter. *Phys. Rev. Lett.*, 95:017402, 2005.
53. A. Sundaramurthy, K.B. Crozier, G.S. Kino, D.P. Fromm, P.J. Schuck, and W.E. Moerner. Field enhancement and gap-dependent resonance in a system of two opposing tip-to-tip Au nanotriangles. *Phys. Rev. B*, 72:165409, 2005.
54. S.-W. Chang, C.-Y.A. Ni, and S.L. Chuang. Theory of bowtie plasmonic nanolaser. *Opt. Express*, 16:10580–10595, 2008.
55. I.E. Protsenko, A.V. Uskov, O.A. Zaimidoroga, V.N. Samoilov, and E.P. O'Reilly. Dipole nanolaser. *Phys. Rev. A*, 71:063812, 2005.
56. I.E. Protsenko, N.F. Starodubtsev, V.M. Rudoy, O.V. Dementieva, N.N. Naumov, O.A. Zaimidoroga, and V.N. Samoilov. Plasmon resonance, laser generation and photoeffect in thin heterogeneous layers and nanostructures. In V.V. Samartsev (ed.), *Photon Echo and Coherent Spectroscopy*, vol. 6181 of Proc. SPIE, Bellingham, WA, 2005, p. 618116.
57. K.J. Vahala. Optical microcavities. *Nature*, 424:839–846, 2003.
58. H.-G. Park, C.J. Barrelet, Y. Wu, B. Tian, F. Qian, and C.M. Lieber. A wavelength-selective photonic-crystal waveguide coupled to a nanowire light source. *Nat. Photon.*, 2008.
59. K. Nozaki, H. Watanabe, and T. Baba. Photonic crystal nanolaser monolithically integrated with passive waveguide for effective light extraction. *Appl. Phys. Lett.*, 92:021108, 2008.
60. T. Yoshie, M. Loncar, K. Okamoto, Y. Qiu, O.B. Shchekin, H. Chen, D.D. Deppe and A. Scherer. Photonic crystal nanocavities with quantum well or quantum dot active material. In A. Abidi, A. Scherer, and S.-Y. Lin (eds.), *Photonic Crystal Materials and Devices II*, vol. 5360 of Proc. SPIE, Bellingham, WA, 2004, pp. 16–23.
61. K. Nozaki, S. Kita, and T. Baba. Room temperature continuous wave operation and controlled spontaneous emission in ultrasmall photonic crystal nanolaser. *Opt. Express* 15:7506–7514, 2007.
62. T. Yang, A. Mock, J.D. OBrien, S. Lipson, and D.G. Deppe. Edge-emitting photonic crystal double-heterostructure nanocavity lasers with InAs quantum dot active material. *Opt. Lett.*, 32:1153–1155, 2007.
63. S.L. Jaiswal and P.M. Norris. Variations in a design for a nanoscale silicon laser. In P.V. Farrell, F.-P. Chiang, C.R. Mercer, and G. Shen (eds.), *Optical Diagnostics for Fluids Solids, and Combustion II*, vol. 5191 of Proc. SPIE, Bellingham, WA, 2003, pp. 226–233.
64. L. Pavesi, L. Dal Negro, C. Mazzolenim, G. Franzo, and F. Priolo. Optical gain in silicon nanocrystals. *Nature*, 408:440–444, 2000.
65. K.L. van der Molen. Experiments on scattering lasers from Mie to random. PhD thesis, University of Twente, 2007.
66. K.L. van der Molen, P. Zijlstra, A. Lagendijk, and A.P. Mosk. Laser threshold of Mie resonances. *Opt. Lett.*, 31:1432–1434, 2006.
67. D.S. Wiersma, M.P. van Albada, and A. Lagendijk. Coherent backscattering of light from amplifying random media. *Phys. Rev. Lett.*, 75:1739–1743, 1995.
68. D. Wiersma. The physics and applications of random lasers. *Nat. Phys.*, 4:359–367, 2008.
69. V.S. Letokhov. Generation of light by a scattering medium with negative resonance absorption. *Sov. Phys. JETP*, 26:835 840, 1968.
70. N.M. Lawandy, R.M. Balachandran, A.S.L. Gomes, and E. Sauvain. Laser action in strongly scattering media. *Nature*, 368:436–438, 1994.
71. M. Noginov. *Solid-State Random Lasers*. Springer-Verlag, New York, 2005.
72. A. Fratalocchi, C. Conti, and G. Ruocco. Mode competitions and dynamical frequency pulling in Mie nanolasers: 3D ab-initio Maxwell-Bloch computations. *Opt. Express* 16:8342–8349, 2008.
73. H.E. Tuereci, L. Ge, S. Rotter, and A.D. Stone. Strong interactions in multimode random lasers. *Science*, 320:643–646, 2008.
74. D. O'Carroll, I. Lieberwirth, and G. Redmond. Microcavity effects and optically pumped lasing in single conjugated polymer nanowires. *Nat. Nanotechnol.*, 2:180–184, 2007.
75. R.J. Holmes. Nanowire lasers go organic. *Nat. Nanotechnol.*, 2:141–142, 2007.
76. P.L. Gourley, J.K. Hendricks, A.E. McDonald, R.G. Copeland, K.E. Barrett, C.R. Gourley, and R.K. Naviaux. Ultrafast nanolaser flow device for detecting cancer in single cells. *Biomed Microdevices*, 7:331–9, 2005.
77. P.L. Gourley and R.K. Naviaux. Optical phenotyping of human mitochondria in a biocavity laser. *IEEE J. Sel. Top. Quantum Electron.*, 11:818–826, 2005.
78. S.-H. Kim, J.-H. Choi, S.-K. Lee, S.H. Kim, S.-M. Yang, Y.-H. Lee, C. Seassal, P. Regrency and P. Viktorovitch. Optofluidic integration of a photonic crystal nanolaser. *Opt. Express*, 16:6515–6527, 2008.
79. S. Kita, K. Nozaki, and T. Baba. Refractive index sensing utilizing cw photonic crystal nanolaser and its arrayed configuration. In *OSA/CLEO/QELS*, San Jose, CA, page paper CM06, 2008.
80. R. Ikkawi, N. Amos, A. Krichevsky, R. Chomko, D. Litvinov, and S. Khizroev. Nanolasers to enable data storage beyond 10 tbit/in^2. *Appl. Phys. Lett.*, 91:153115, 2007.
81. Y. Arakawa and H. Sakaki. Multidimensional quantum well laser and temperature dependence of its threshold current. *Appl. Phys. Lett.*, 40:939–941, 1982.
82. M. Asada, Y. Miyamoto, and Y. Suematsu. Gain and the threshold of three-dimensional quantum-box lasers. *IEEE J. Quantum Electron.*, 22:1915–1921, 1986.
83. F. Qian, S. Gradecak, Y. Li, C.-Y. Wen, and C.M. Lieber. Core/multishell nanowire heterostructures as multicolor, high-efficiency light-emitting diodes. *Nano Lett.*, 5:2287–2291, 2005.
84. D. Wiersma. The smallest random laser. *Nature*, 406:132–133, 2000.
85. P.J. Pauzauskie and P.D. Yang. Nanowire photonics. *Mater. Today*, 9:36–45, 2006.

40

Quantum Dot Laser

40.1 Introduction **40**-1
Material Systems • Edge Emitter Laser Diodes
40.2 Electronic States of Quantum Dots **40**-4
Localized and Delocalized States in Self-Assembled Structures • Envelope-Function Approximation and Effective-Mass Approach • Continuum Approaches and Atomistic Models
40.3 Scattering Processes of Carriers in Quantum Dots **40**-5
Carrier–Carrier Interaction • Carrier–Phonon Interaction
40.4 Optical Gain of the Active Material **40**-10
Optical Susceptibility • Interband Transitions and Macroscopic Polarization • Optical Gain Calculations • Gain Saturation, Refractive Index, and α-Factor
40.5 Laser Emission Properties **40**-13
Rate Equations for Quantum-Dot Systems • Microscopic Generalization: Semiconductor Effects • Quantum-Dot Microcavity Lasers: Modifications of the Photon Statistics
Acknowledgments **40**-16
References **40**-16

Frank Jahnke
University of Bremen

40.1 Introduction

The success of optoelectronics is largely based on the availability of efficient and reliable semiconductor lasers. Most prominent among the many applications are optical data storage solutions on CDs, DVDs, and BDs or the fiber-based optical communication. The initial success started with conventional laser diodes (Casey and Panish, 1978a,b; Agrawal and Dutta, 1993; Chuang, 1995). More recent microcavity lasers (Li and Iga, 2003) provide a new level of miniaturization and novel emission properties.

In general, the laser emission is based on population inversion of the electronic states, which contribute to the radiative recombination. For semiconductor lasers, it has been realized early (Dingle and Henry, 1976) that the carrier confinement in quantum wells leads to improved laser properties, like a laser threshold reduction. The successful demonstration of this effect was based on advances in the material growth (Tsang, 1982). The use of quantum dots (QDs) as an active material promises even greater benefits. First of all, one gains more freedom to engineer the emission wavelength (Bimberg et al., 1999). Another advantage is the enhancement of the single-particle density of states. For states experiencing a carrier confinement in all three dimensions, discrete electronic energies similar to the spectrum of atomic systems are obtained. This has led to predictions of further increased efficiency for laser applications, which are based on less temperature-dependent operational parameters (Arakawa and Sakaki, 1982), further reduced threshold currents and higher differential gain (Asada et al., 1986), higher modulation bandwidth (Kim et al., 2004) and reduced anti-guiding effects (Newell et al., 1999), as well as a reduced sensitivity to material defects. Also, the miniaturization of the active material in connection with new microcavity laser resonators opens the door for novel applications in quantum information technologies.

The goal of this chapter is to provide an introduction to the physics underlying QD lasers. We will discuss the fundamental properties of the QDs, which are relevant for laser applications, as well as the important interaction processes that determine the laser properties.

40.1.1 Material Systems

Various techniques can be employed for the fabrication of QDs. Each approach has advantages for particular applications. Self-assembled QDs, which are typically obtained in the Stranski–Krastanov growth mode, are the most widely used for optoelectronic devices. Colloidal semiconductor nanostructures, which can be functionalized as biological markers, are less practical for the device integration. Lateral QDs, which are fabricated by etching of quantum wells following electron beam lithography, have the advantage of better controllable size and position. They are favorable for quantum-transport studies. The slow preparation process, the low dot density, and defect formation prevent the broad use of this method in optoelectronics.

For emission in the near infrared and red, including the frequencies for optical fiber telecommunication networks at 1300 and 1500 nm, the GaAs material system (GaAs/AlGaAs, InGaAs/GaAs) is routinely used. The self-organized growth of QDs is well established for this material system. The fabrication starts with the strained growth of a two-dimensional wetting layer (WL) on a substrate with a higher band gap and lower lattice constant. After a critical thickness is reached, partial strain relaxation changes the morphology to a thinner two-dimensional layer and small three-dimensional islands. These islands constitute the self-assembled QDs after their overgrowth with a higher band-gap material.

II–VI compounds with a wider band gap like CdSe, CdTe, or ZnSe provide emission in the green and yellow spectral range (Klude et al., 2004; Sebald et al., 2009). However, issues of the device lifetime due to degradation effects have so far prevented the broad application in light emitting devices. Also the growth mechanism often differs from that in III–V compounds. For example, CdTe/ZnTe has a similar lattice mismatch to InAs/GaAs, but the growth of CdTe QDs starts with two-dimensional platelets on which small dots are formed. Isolated dots are growing, when the material thickness exceeds a critical value.

The group-III nitride material system, like GaN/AlN or InGaN/GaN, received much attention for applications in light emitters due to their extended range of emission frequencies from visible light to UV emission (Nakamura and Chichibu, 2000). While the nonradiative loss of carriers due to trapping at threading dislocations lowers the efficiency of group-III quantum-well light emitters, this effect is reduced by the three-dimensional carrier confinement in QDs. Nitride-based heterostructures with a wurtzite crystal structure are known to have built-in electrostatic fields due to the spontaneous polarization and piezoelectric effects, which lower the spatial overlap of the electron and hole wave functions thereby reducing the interband recombination efficiency.

40.1.2 Edge Emitter Laser Diodes

Two different types of semiconductor lasers need to be distinguished in general. They vary in design, properties, and applications. In conventional edge-emitting laser diodes, the feedback for the laser mode is provided by an optical cavity, which is formed by the cleaved end faces of the device (Agrawal and Dutta, 1993; Chuang, 1995). Correspondingly, the emission occurs from the edge of the semiconductor chip. The low mirror reflectivity due to the jump of the refractive index from the semiconductor material to air results in a short cavity lifetime (large cavity damping). To compensate for the cavity losses, a large propagation length of the laser mode through the active material is necessary. The large cavity length in the range of a few hundred micrometers in turn leads to a small frequency spacing between the cavity modes, so that many modes appear within the gain region of the active material. As a result, edge emitters typically operate multimode. Also the spontaneous emission of the active material is distributed over a large number of nonlasing modes, which leads to a pronounced laser threshold as discussed in Section 40.5.

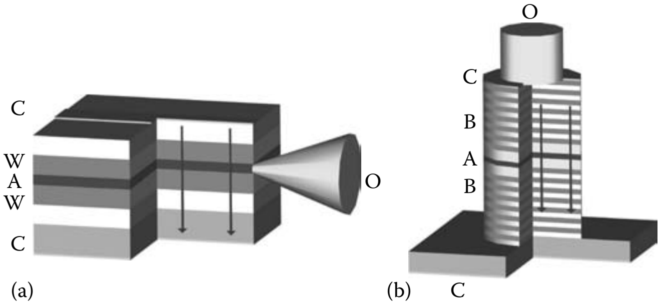

FIGURE 40.1 Edge-emitting laser (left) and vertical-cavity surface-emitting laser (right) with active region (A), waveguide layers (W), upper and lower Bragg mirrors (B), contacts (C), and light output (O). The arrows indicate the direction of the current flow.

The optical gain is achieved by current-injection pumping of the active material. For an edge-emitting laser diode, it is outlined in Figure 40.1 that the electrical injection current flows perpendicular to the optical cavity modes and the emission direction, which is characteristic for this type of devices. To provide an efficient confinement of the pumped charge carriers to the active region, a heterostructure design is used, where the active layer is positioned between materials with wider band gaps. A common heterostructure has a GaAs active layer between AlGaAs barriers. The carrier confinement is increased by designing the laser structure to be very thin in the direction of current flow, which is also the epitaxial growth direction. Additional cladding layers can be used to confine the optical mode in the transverse direction by means of a waveguide geometry and to increase the spatial overlap of the optical mode with the active region. Such a design is called double heterostructure and was independently proposed by Kroemer as well as Alferov and Kazarinov (Casey and Panish, 1978a). The conventional laser configuration works very well, as is evident from its wide ranging applications in many electro-optical systems.

This design allows for a natural integration of self-organized QDs as active material. Since the dots are grown in layers, these can serve as the planes in which the optical mode propagates. To increase the modal gain, often several layers are staggered. Accompanied with the large number of QDs and unavoidable fluctuations of the individual QD properties (mostly geometry and composition) is a large inhomogeneous broadening of the emission spectrum of the active material.

On the other hand, the combination of edge emission and active layer shape give rise to several disadvantages (independent on whether quantum wells or QDs are used as active material). For edge emission, the laser transverse and lateral modes depend on the cross section of the gain region, which is very thin in the transverse dimension for carrier confinement and wide in the lateral dimension for output power. The resulting near field of the laser emission is highly elongated, which does not match well to the circular cross section of optical fibers for possible out-coupling. Since the transverse beam dimension is very small, typically of the order of 1 μm or less, the transverse beam divergence is rather high (typically ≈50° full angle) because of diffraction. Therefore, the far field of the laser emission is also highly elongated. The beam astigmatism together with the high divergence makes design and fabrication of coupling optics rather challenging. Approaches to single-mode operation involve embedded gratings or buried heterostructures, which increase the unit costs and limit the output power.

40.1.2.1 Microcavity Lasers

Microcavity lasers represent a new type of devices. In analogy to the carrier confinement, one utilizes a strong confinement of the electromagnetic field. Cavity dimensions of the order of the light wavelength are characteristic of microcavity lasers. Due to the strongly reduced gain length, this requires very high mirror reflectivities larger 99%. To characterize microresonators, one uses the cavity quality factor $Q = \omega_c/\Delta\omega_c$ with the cavity frequency ω_c and the spectral linewidth $\Delta\omega_c$. The combination of small mode volume and high Q-factor leads to modifications in the density of states of the electromagnetic field. According to Fermi's golden rule, this allows to enhance or inhibit the spontaneous recombination probability, which is known as Purcell effect (Vahala, 2003; Lodahl et al., 2004). Particularly important for lasers is that the small cavity dimensions strongly increase the frequency spacing between the cavity modes. Microcavity lasers can be designed in a way to find only a single high-Q cavity mode within the gain spectrum of the active material, thus allowing genuine single-mode laser operation. The Purcell effect can also be used to suppress the spontaneous emission into other low-Q modes (e.g., modes propagating in other directions). This leads to a strongly reduced laser threshold, and the possibility of thresholdless lasers has been discussed (Yokoyama and Brorson, 1989; Rice and Carmichael, 1994) as a consequence of the novel physics.

The foundation of microcavity lasers lies in the improved engineering of optical resonators in combination with the new physics in the field of cavity-quantum electrodynamics. This includes mastering Maxwell's equations in even more and more complicated materials and geometries and still being able to predict the electromagnetic properties and the response of devices. The photon confinement is based either on Bragg or total internal reflection. A central role plays the concept of photonic bandgap materials. Here, one uses the analogy to the electronic band structure, which emerges for the electron motion in a lattice periodic potential. Photonic structures with periodic modulations of the refractive index can be designed in a way that a photonic band gap for the light propagation, the so-called stop band appears.

For practical device applications, vertical-cavity surface-emitting lasers (VCSELs) are frequently used, since many of the above mentioned disadvantages of the edge-emitting design are avoided. The VCSEL constitutes of a Fabry-Perot resonator, in which the very high reflectivity is achieved with distributed Bragg reflectors (DBRs), consisting of epitaxially grown periodic layers of different refractive indices (Jewell et al., 1991). The typical DBR has alternating layers that are each $\lambda/4$ thick so as to provide constructive interference of the reflected waves from each interface. The spacer layer between the DBRs acts as a defect giving rise to the high-Q cavity mode, which provides the photon confinement.

The basic VCSEL structure is that of an effectively one-dimensional resonator with a length of a few μm. The active region between the mirrors consists either of one or a few quantum wells or layers with QDs, which are embedded between cladding layers. The short cavity length of this structure allows to design a single high-Q (longitudinal) cavity mode within the spectral gain region of the active material. For laser structures, current confinement to a small area of the active region has been introduced by means of oxide apertures.

Since the optical cavity axis is in the vertical (epitaxial growth) direction (see Figure 40.1), the laser emission for a VCSEL occurs out of the surface of the wafer. This surface emitting configuration has several desirable features. Being independent of the gain layer cross section, there is more flexibility in the shape and size of the transverse optical mode. In particular, a circular beam can be realized. With an approximately 6 μm diameter circular output aperture, lowest order (Gaussian-like) transverse mode operation with less than 10° beam divergence is possible (Chang-Hasnain et al., 1991). Other advantages of a VCSEL are even more practical in nature. Unlike an edge-emitter, the VCSEL mirrors are fabricated during the epitaxial growth of the entire wafer, so that the mirrors of hundreds of lasers can be manufactured simultaneously. Edge emitters can only be combined to one-dimensional laser arrays, whereas with surface emitters truly two-dimensional arrays are routinely possible. To test an edge emitter, one needs to expose the edges, which means cleaving the wafer and applying the necessary facet preparations. These are time consuming and therefore costly steps that have to be repeated for each laser diode. Surface emitters can be fabricated, tested, and operated at the wafer level.

The VCSEL design has its own challenges. This gain-length reduction has to be compensated by the high cavity Q, putting great demands on the quality of the DBR mirrors. Having only one high-Q cavity resonance in the spectral gain region, the laser threshold current is very sensitive to the position of this resonance in relation to the peak of the material gain spectrum. From this, a related complication arises: the cavity-gain alignment can be maintained typically only within a small temperature range, because gain spectrum and cavity resonances have different temperature dependences. The gain spectrum shifts in frequency because of the temperature dependence of the bandgap energy and the carrier distributions. The cavity resonances shift with temperature due to thermally-induced changes in the refractive indices of the mirror material and the material within the optical cavity. The result is a temperature dependence of the cavity-gain alignment, which plays an important role in the sensitivity of a VCSEL output to temperature variations.

An additional transverse mode confinement in VCSEL structures can be introduced with the fabrication of micropillars, e.g., due to a combination of electron-beam lithography and various etching methods. The resulting three-dimensional confinement of the optical mode to a very small volume of the order of the cubic light wavelength strongly increases the efficiency of the light–matter interaction. With the three-dimensional photon confinement, the Purcell effect can be used for a strong suppression of the spontaneous emission into non-lasing modes. Even higher cavity-Q factors for three-dimensional mode confinement have been achieved in microdisks or microspheres. They lack, however, the directionality of the light emission. The ultimate control of the photon confinement and record values for the cavity-Q has been possible in photonic crystal structures.

Both in conventional edge-emitters and in microcavity lasers, the QD physics that determines the properties of the gain material is the same and the effects discussed in the subsequent chapters equally apply. Nonetheless, the miniaturization of microcavity lasers can be pushed to the extreme case where only a single QD resonantly interacts with a single high-Q optical mode. This system is the analogous to the single-atom micromaser or the one-atom laser of quantum optics and opens the door for novel applications in quantum-information technologies like quantum cryptography (Fattal et al., 2004) and quantum computers (Scholz et al., 2006). Essential physical effects or mechanisms are the generation of single-photons on demand (Michler et al., 2000) or correlated photons (Moreau et al., 2001), as well as the demonstration of vacuum field Rabi oscillations (strong coupling) (Reithmaier et al., 2004; Yoshie et al., 2004).

40.2 Electronic States of Quantum Dots

The single-particle states will be taken as the starting point for the description of the electronic and optical properties of semiconductor QDs as active material in lasers. In terms of quantum mechanics, single-particle states constitute the basis for the analysis of interaction processes. Interaction matrix elements can be formulated with the corresponding wave functions, and transition processes can be described in terms of the single-particle states.

40.2.1 Localized and Delocalized States in Self-Assembled Structures

Due to the three-dimensional carrier confinement, QDs possess localized states in contrast to the band structure of the bulk material. Nevertheless, the single-particle states are closely related to the properties of the bulk material. Especially the optical selection rules and symmetry properties of the interaction matrix elements reflect properties of the underlying crystal lattice. Also band mixing effects and built-in crystal fields are results of the lattice properties.

Depending on material composition and growth conditions like temperature and material flux, a variety of different QD geometries can be fabricated (Jacobi, 2003). In the following, we consider self-assembled QD systems in a characteristic geometry, which consist of lens-shaped caps positioned on a residual two-dimensional WL, as shown in the left part of Figure 40.2. For the frequently used InGaAs/GaAs material system, one finds typically a diameter of 10–30 nm and a height-to-diameter aspect ratio of 0.3–0.4. The random distribution of QDs on the WL plane has a density of 10^{10}–10^{11} cm^{-2}.

From the geometry of the QD-islands on top of the two-dimensional WL, the single-particle energy spectrum can be understood qualitatively. The WL provides a confinement of carriers in the growth direction and—in the absence of QDs—a free carrier motion in the perpendicular WL plane, as in a quantum well. At the position of the islands, the height of the lower band-gap material increases, thus reducing the ground-state energy and therefore giving rise to an additional in-plane confinement potential. The resulting effective three-dimensional carrier confinement in the QDs leads to a localized state with discrete energies for electrons and holes. These discrete states are located energetically below a quasi-continuum of delocalized states, corresponding to the two-dimensional motion of carriers in the WL. At even higher energies, one finds the barrier states, which are delocalized in all three dimensions. The energy levels of the combined QD-WL system are sketched in the right part of Figure 40.2. The actual number of discrete QD states is limited by the height of the confinement potential as

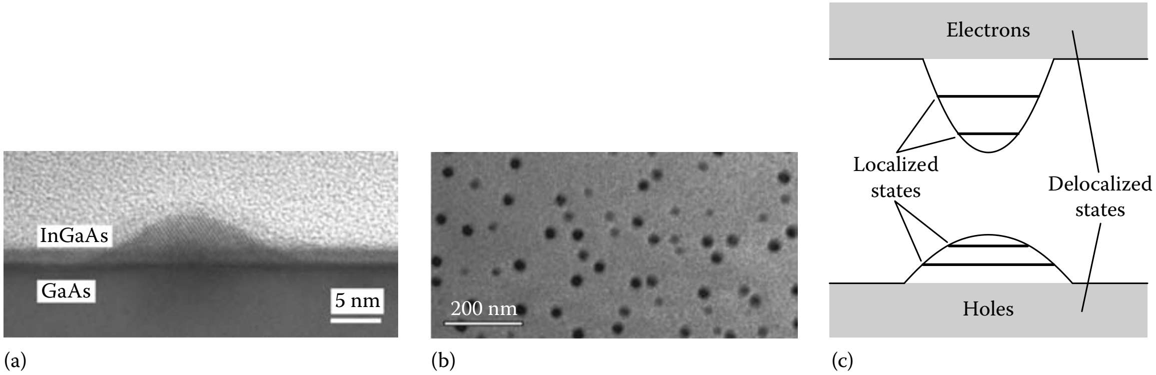

FIGURE 40.2 Cross section (a) and top view (b) TEM image of self-assembled InGaAs/GaAs QDs. The corresponding energy spectrum (c) contains localized states due to the three-dimensional carrier confinement, which are positioned energetically below a quasi-continuum of delocalized states corresponding to the two-dimensional motion of carriers in the WL. (From Anders, S. et al., *Phys. Rev. B*, 66, 125309, 2002.)

well as the geometry of the dots. As we will discuss in Section 40.3, the delocalized states play an important role for the scattering processes in light-emitters based on QDs and, hence, need to be included in the analysis of the single-particle states.

40.2.2 Envelope-Function Approximation and Effective-Mass Approach

The simplest approach to the carrier confinement in semiconductors is based on the envelope-function approximation. The starting point is the Bloch theorem, which represents the wave functions in a three-dimensional crystal in the form

$$\Phi_{\boldsymbol{k}}(\boldsymbol{r}) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}u_\lambda(\boldsymbol{r}), \tag{40.1}$$

describing plane waves modulated with the lattice-periodic Bloch function $u_\lambda(\boldsymbol{r})$, where λ is the band index and $\boldsymbol{k}$ is the carrier wave vector. In a confinement situation, one assumes that the three-dimensional plane wave can be replaced with envelope wave functions that are localized in one, two, or three dimensions.

In effective-mass approximation, one assumes that the actual QD and the two-dimensional WL consist of the same semiconductor material, in which the carrier motion experiences a constant potential energy and a kinetic energy of the carriers that can be expressed in terms of an effective mass m^*. The structure is embedded in a buffer material with a larger band gap. This is described by a step-wise increase of the confinement potential $V(\boldsymbol{r})$ to the outside region. Typically, the effective mass is also different in the buffer material, which is expressed with a step-wise constant $m^*(\boldsymbol{r})$. The Schrödinger equation

$$\left[\frac{\hbar^2}{2m^*(\boldsymbol{r})}\Delta + V(\boldsymbol{r})\right]\phi_n(\boldsymbol{r}) = \varepsilon_n\phi_n(\boldsymbol{r}) \tag{40.2}$$

defines the confinement wave functions $\phi_n(\boldsymbol{r})$, which replace the plane waves in Equation 40.1, and provides the corresponding energies ε_n for the single-particle states.

For the geometry displayed in the left part of Figure 40.2, several simplifications of the solution of Equation 40.2 are possible. The rotational symmetry allows for a classification of the single-particle states according to their angular momentum and to effectively reduce the three-dimensional problem to a two-dimensional one. Details for the solution of this problem are discussed in Wojs et al. (1996). Additionally, when the QD height-to-diameter aspect ratio is small, one can distinguish a weak (strong) confinement situation in the WL plane (growth direction), and in good approximation the confinement wave function can be separated into the corresponding components

$$\Phi_n(\boldsymbol{r}) = \varphi_{l,\lambda}(\varrho)\xi_{\sigma,\lambda}(z)u_\lambda(\boldsymbol{r}), \tag{40.3}$$

with $\boldsymbol{r} = (\varrho, z)$. The in-plane envelope function $\varphi_{l,\lambda}(\varrho)$ represents both localized and delocalized states for the QD and WL contributions, respectively. In a simple but often useful approximation, the in-plane confinement of flat QDs with cylindrical symmetry is described by a two-dimensional harmonic confinement potential with a characteristic equidistant spacing of the corresponding energies (Wojs et al., 1996). In the WL, the in-plane momentum $\boldsymbol{k}$ is a good quantum number if the effect of the localized states on the continuum is neglected. Thus, the quasi-continuum of the WL states (in the absence of QDs) is modeled by using plane waves for the corresponding in-plane envelope functions $\varphi_{l,\lambda}(\varrho)$. However, the localized QD states and the delocalized WL states are solutions of a single-particle problem for one common confinement potential and must, therefore, form an orthogonal basis. By performing a separate ansatz for the QD and WL contributions, this orthogonality is not ensured. It can be enforced, for example, by an orthogonalization procedure of the continuum states. The obtained so-called orthogonal plane waves (OPWs) represent a more correct description of the WL states. For details, we refer the reader to Nielsen et al. (2004, 2005).

40.2.3 Continuum Approaches and Atomistic Models

The effective-mass approach has several shortcomings. Band-mixing effects usually modify the electronic states, and a more detailed description of the valence-band structure is often desired. This can be included in semiempirical continuum approaches like $k\cdot p$-models or atomistic tight-binding approaches. The latter account also for symmetry properties of the atoms in the crystal unit cell, which can influence optical transition rules, and allows to study systematically the combined influence of structure and chemical morphology including alloy fluctuations and defects.

These effects are particularly important in nitride-based heterostructures, and for these systems extensive studies of the electronic states have been performed. The valence band of a wurtzite nitride semiconductor is strongly nonparabolic due to the small spin–orbit and crystal-field splitting (Chuang and Chang, 1996; Vurgaftman and Meyer, 2003). The hexagonal symmetry of the crystal leads to the subbands having the same effective mass along the *c*-axis. Therefore, the confinement of the holes in this direction due to the WL is not able to separate these subbands, and strong subband mixing, and nonparabolicity effects remain present. Also modifications of the wave functions due to built-in fields and their screening due to excited carriers are important.

First-principle calculations are available for the computation of the electronic structure of small QD systems (200–400 atoms). Due to the large number of atoms together with the absence of translational symmetry, which greatly simplifies bulk calculations, density functional theory-based approaches are computationally too demanding to date.

40.3 Scattering Processes of Carriers in Quantum Dots

To utilize the specific optical properties of QDs, we are interested in recombination processes between the conduction and valence band states associated with the three-dimensional carrier confinement. Efficient light emitters like LEDs rely on a large

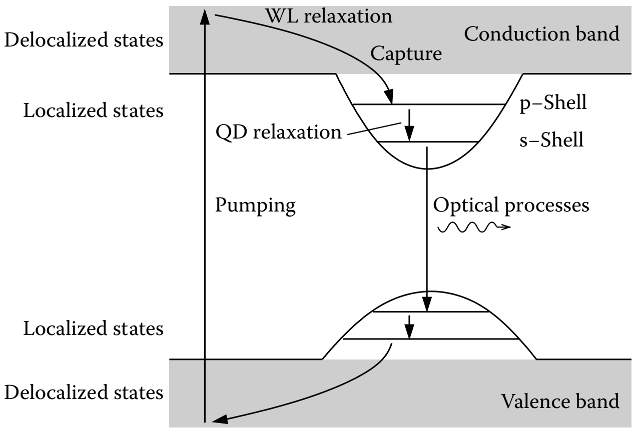

FIGURE 40.3 Fundamental carrier scattering processes in QD-based light emitters.

population of the QD states with electrons and holes. For laser applications, the realization of optical gain even requires population inversion. Electrical current injection leads to the generation of carriers in the delocalized states. These carriers need to be scattered into the localized QD states (carrier capture). Also efficient carrier scattering between the QD states (carrier relaxation) is necessary to obtain a large ground-state population. The corresponding transition processes are schematically shown in Figure 40.3. Likewise, optically pumped QD lasers depend on fast carrier-scattering processes. The two important intrinsic scattering mechanisms contributing to the capture and relaxation are the carrier–carrier Coulomb interaction, which can be assisted by carriers in bound (QD) or extended (WL) states, and the carrier–phonon interaction. Both mechanisms are discussed separately in Sections 40.3.1 and 40.3.2. These scattering processes are also an important source for optical dephasing, which strongly influences the optical properties. Furthermore, this underscores the integral part the WL states play for the physics and application of self-assembled QDs.

Experimental insight into carrier scattering processes can be obtained by ultrafast optical pump–probe spectroscopy, ranging from absorption measurements to advanced two-color pump–probe and four-wave mixing experiments (Shah, 1999). Both at room temperature and at lower temperatures, efficient carrier scattering processes have been demonstrated.

40.3.1 Carrier–Carrier Interaction

The most simple and at the same time most efficient form of carrier–carrier collisions involves the energy transfer between two particles. To analyze the corresponding re-distribution of the carrier population within a quantum-statistical treatment, one can investigate the dynamics of the carrier population $f_\nu(t)$ for the state ν as a function of time t. The corresponding equation of motion for $f_\nu(t)$ can be formulated on different levels of refinement. Frequently used are kinetic equations with Boltzmann scattering rates, which describe the temporal change of the carrier population $f_\nu(t)$ due to various scattering processes into and out of the states ν according to

$$\frac{\partial}{\partial t} f_\nu = (1 - f_\nu) S_\nu^{\text{in}} - f_\nu S_\nu^{\text{out}}. \tag{40.4}$$

To obtain this form, a Markov approximation has been used, where one assumes that the population changes at a given time t depend explicitly only on the population functions $f_\nu(t)$ at the same time. Generally, the scattering rates are proportional to the population f of the initial states and to the nonoccupation $1 - f$ of the final states. Correspondingly, on the RHS of Equation 40.4, the in-scattering (out-scattering) rate, which increases (decreases) the population, is proportional to the non-occupation probability $1 - f_\nu$ (occupation probability f_ν).

For the Coulomb interaction, where two carriers are scattered from the states ν_1 and ν_3 into ν and ν_2 and vice versa, the scattering integral S_ν^{in} is given by

$$S_\nu^{\text{in}} = \frac{2\pi}{\hbar} \sum_{\nu_1,\nu_2,\nu_3} W_{\nu\nu_2\nu_3\nu_1} [W^{\star}_{\nu_2\nu_3\nu_1} \quad - W^{\star}_{\nu\nu_2\nu_1\nu_3}]$$
$$\times f_{\nu_1}(1 - f_{\nu_2}) f_{\nu_3} \delta(\tilde{\varepsilon}_\nu - \tilde{\varepsilon}_{\nu_1} + \tilde{\varepsilon}_{\nu_2} - \tilde{\varepsilon}_{\nu_3}). \tag{40.5}$$

A similar expression for the out-scattering contribution S_ν^{out} is obtained by replacing $f \rightarrow 1 - f$. The population factors account for the availability of scattering partners, as discussed above. The sum involves all available initial states for in-scattering processes as well as all possible scattering partners that can be provided in QD or WL states.

In Markov approximation, the delta function accounts for strict energy conservation of the scattering processes. This approximation is valid in the long time limit, where scattering events are considered for times that are much larger than the inverse scattering rates. Even in this case, it is necessary to account for renormalizations of the single-particle energies $\tilde{\varepsilon}_\nu$ by the Coulomb interaction instead of using free-carrier energies ε_ν as done in the perturbation theory.

The scattering cross section is also determined by the matrix elements of the screened Coulomb interaction $W_{\nu\nu_2\nu_3\nu_1}$. The first and second terms in the square brackets correspond to direct and exchange Coulomb scattering, respectively.

A treatment beyond the Markov approximation is possible within a quantum-kinetic description (Haug and Jauho, 1996). Then the delta-function is replaced by the spectral functions of the involved carriers, and time-integrals over the past of the system account for memory effects. The most simple form of a quantum-kinetic equation requires two further approximations: (1) Within the generalized Kadanoff–Baym ansatz (GKBA), closed equations for single-time population functions are obtained (see Haug and Jauho, 1996 for detailed discussions). (2) The quasiparticle properties of the involved carriers are described by renormalized energies and damping (finite lifetime of the quasiparticle state), which can be summarized in complex energies $\hat{\varepsilon}_\nu$. The latter approximation is related to the Fermi liquid theory

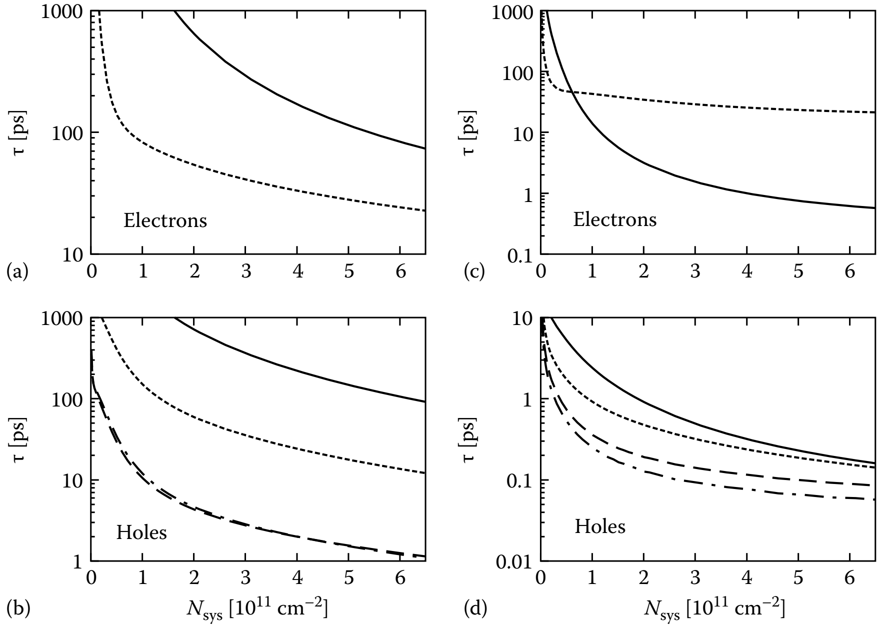

FIGURE 40.4 Capture times for electrons (a) and holes (b) and relaxation times for electrons (c) and holes (d) as a function of the total carrier density in the system N_{sys} for the following quantum-dot states of an InGaN/GaN material system and room temperature: s-shell (solid lines), p-shell (dotted lines), d_0-shell (dashed-dotted lines), and $d_\pm$-shell (dashed lines). (From Nielsen, T.R. et al., *Phys. Rev. B*, 72, 235311, 2005. With permission.)

and valid for large carrier densities, where excitonic effects can be neglected. The quantum-kinetic equation with the discussed approximations has the form

$$\frac{\partial}{\partial t} f_\nu(t) = \int_{-\infty}^{t} dt' \left\{[1-f_\nu(t')]S_\nu^{\text{in}}(t,t') - f_\nu(t')S_\nu^{\text{out}}(t,t')\right\}, \qquad (40.6)$$

where the in-scattering rate, which now depends on two time arguments, is given by

$$\begin{aligned} S_\nu^{\text{in}}(t,t') = {} & \frac{1}{\hbar^2}\text{Re}\sum_{\nu_1,\nu_2,\nu_3} W_{\nu\nu_2\nu_3\nu_1}(t)\left[W^{\star}_{\nu\nu_2\nu_3\nu_1}(t') - W^{\star}_{\nu\nu_2\nu_1\nu_3}(t')\right] \\ & \times f_{\nu_1}(t')\left[1-f_{\nu_2}(t')\right]f_{\nu_3}(t') \\ & \times \exp\left(-\frac{i}{\hbar}(\hat{\varepsilon}_\nu - \hat{\varepsilon}_{\nu_1} + \hat{\varepsilon}_{\nu_2} - \hat{\varepsilon}_{\nu_3}(t-t')\right). \end{aligned} \qquad (40.7)$$

The out-scattering rate is again obtained from $f \rightarrow 1-f$. It turns out, that especially for the calculation of dephasing and its influence on optical absorption and gain spectra, the quantum-kinetic description leads to more realistic results. As discussed in Lorke et al. (2006a), for the description of dephasing in optical spectra, equations similar to (40.6) and (40.7) are used for the polarization dynamics, which will be discussed in Section 40.4.3.

The results for the capture of carriers from the WL into the QD states and for the relaxation between the QD levels are shown in Figure 40.4. Capture and relaxations times have been defined as those times τ_ν, in which the system returns to thermal equilibrium due to the respective processes after a small perturbation. In relaxation-time approximation, one finds $1/\tau_\nu = S_\nu^{\text{in}} + S_\nu^{\text{out}}$, where the scattering rates are evaluated for quasi-equilibrium population functions. With increasing carrier density in the system, the scattering times are decreasing due to the larger number of available scattering partners. For the energetically lower s-shell, direct carrier capture from the WL is less efficient than capture into the p-shell and subsequent relaxation into the s-shell. For further details, we refer the reader to Nielsen et al. (2005).

In the following, we provide additional information on the analysis of carrier–carrier scattering processes. The determination of the Coulomb interaction matrix elements, which enter in Equations 40.5 and 40.7, involves two steps. The bare (unscreened) interaction matrix elements can be calculated from the single-particle wavefunctions $\Phi_\nu(\mathbf{r})$, see Section 40.2, of the involved electronic states according to

$$V_{\nu\nu_2\nu_3\nu_1} = \int d^3r d^3r' \Phi^{\star}_\nu(\mathbf{r})\Phi^{\star}_{\nu_2}(\mathbf{r}')v(\mathbf{r}-\mathbf{r}')\Phi_{\nu_3}(\mathbf{r}')\Phi_{\nu_1}(\mathbf{r}), \qquad (40.8)$$

where $v(\mathbf{r}-\mathbf{r}') = e^2/(4\pi\varepsilon_0\epsilon|\mathbf{r}-\mathbf{r}'|)$ is the Coulomb potential with the background dielectric function ϵ. The electron charge and the vacuum dielectric constant are given by e and ε_0, respectively.

A second step involves the inclusion of screening effects due to excited carriers. From the nature of screening as well as from the excitation conditions at elevated temperatures (considered for optoelectronic applications), it appears natural to account only for screening due to the carriers in the delocalized WL states. A consistent theory of screening for the QD-WL system is developed in Nielsen et al. (2005). On these grounds, a simplified treatment can be justified, which consists in the replacement of $v(\mathbf{r} - \mathbf{r}')$ in Equation 40.8 by a two-dimensionally screened Coulomb potential. The time-dependence of the screened Coulomb matrix elements $W_{\nu\nu_2\nu_3\nu_1}(t)$ is introduced via the carrier population functions $f_\nu(t)$, which determine the screening for the present excitation conditions via a longitudinal dielectric function.

The renormalized single-particle energies, that enter in the scattering rates (40.5) and (40.7), can be obtained from the free-carrier energies by adding Hartree, exchange, and correlation self-energy contributions. The Hartree self-energy involves the electrostatic interaction with all other excited electrons and holes. Only in a system with local charge neutrality, the Hartree terms cancel. In QDs, local charging appears as a result of different electron and hole populations, different envelope wave functions for electrons and holes, as well as due to the charge separation of electrons and holes in a built-in electrostatic fields (Nielsen et al., 2005).

40.3.2 Carrier–Phonon Interaction

For low carrier densities, when carrier–carrier Coulomb scattering is less important, the interaction of carriers with phonons provides the only remaining efficient scattering mechanism. At higher carrier densities, the carrier–phonon interaction remains important, since the Coulomb scattering is unable to dissipate kinetic energy from the carrier system. Note that the carrier–carrier scattering leads to an evolution of the carrier system towards a quasi-equilibrium state for electrons and holes. However, it conserves the average kinetic energy of the carriers. Pumping of carriers with above-average kinetic energy and the preferred recombination of carriers in the highly populated states with below-average kinetic energy effectively heats up the carrier system. This excess kinetic energy can only be dissipated to the crystal lattice via the emission of phonons. A thermalization of the carriers with respect to the crystal lattice is essential for efficient laser operation, since the increasing carrier temperature diminishes the optical gain via a reduction of the population inversion.

40.3.2.1 Perturbation Theory versus Polaron Picture

The early work on carrier–phonon interaction in semiconductor QDs has been based on the application of time-dependent perturbation theory. This method has previously been applied to bulk semiconductors or quantum wells. In these cases, its success is based on the application to carriers in continuum states. For the discussion of carrier–phonon interaction involving localized QD states with discrete energies, one has to consider the fact that the transition matrix elements provide only efficient coupling to phonons with small momenta (Bockelmann and Bastard, 1990; Inoshita and Sakaki, 1997). The strongest contribution to the interaction of carriers with lattice vibrations is typically provided by LO-phonons due to the Coulomb-like coupling of carriers to lattice displacements in polar crystals. The nearly constant LO-phonon energy, at least for the considered small phonon momenta, and the energy-conservation requirement from Fermi's golden rule led to the prediction of inefficient carrier transitions between the QD states, except for the unlikely coincidence of transition energies between the QD states with the LO-phonon energy. The other, usually less important interaction process is provided via deformation potential coupling to LA-phonons. For small phonon momenta, their energy is too small to facilitate efficient carrier scattering processes between the QD states. On these ground, particularly due to the application of perturbation theory, inefficient carrier–phonon scattering processes in QD systems have been predicted (Benisty et al., 1991). To prevent this so-called *phonon bottleneck*, only much less efficient higher-order processes, like those involving combinations of LA- and LO-phonons (Inoshita and Sakaki, 1992; Jiang and Singh, 1998) or the contributions of carrier–carrier scattering processes at elevated carrier densities have been proposed.

For a more advanced description of the interaction with LO-phonons, one has to take into account that freely moving charge carriers in ionic crystals are surrounded by a cloud of lattice distortions. The renormalization of electrons and holes introduced by the corresponding interaction leads to the emergence of a quasiparticle called polaron. In the past, polarons have mostly been studied for semiconductor structures with a quasi-continuum of electronic states, like bulk materials or quantum wells. Then, electronic transitions are already possible within a free-carrier picture. For the usually considered materials with weak polar coupling, this explains why the application of perturbation theory already provides the leading contributions of polaron effects. They consist in a lowering of the carrier energy (polaron shift) in a renormalization of the carrier mass and the dielectric constant for the Coulomb interaction of carriers, and in a small broadening of the electronic states (Schäfer and Wegener, 2002).

On the contrary, it has been pointed out in Inoshita and Sakaki (1997) and Kral and Khas (1998), and Verzelen et al. (2002) that even in materials with weak polar coupling, the interaction of carriers in localized QD states with LO-phonons leads to strong modifications of the electronic properties. In this case, the phonon satellites of one state hybridize with other discrete states, even if the level spacing does not match the LO-phonon energy. The resulting new quasiparticle of the strongly interacting system are called QD-polarons. They cannot be treated within perturbation theory. Therefore, a simple discussion of carrier–phonon interaction based on Fermi's golden rule is not applicable to QDs. Related to this is an ongoing controversy in the QD literature regarding experimental evidence for (Minnaert et al., 2001; Urayama et al., 2001; Xu et al., 2002) and against (Tsitsishvili et al., 2002; Peronne et al., 2003; Quochi et al., 2003) the phonon bottleneck.

Early investigations of QD-polarons have been based on the "random phase approximation" (RPA) (Inoshita and Sakaki, 1997; Kral and Khas, 1998) or on the direct diagonalization of a restricted state space (Verzelen et al., 2002), where only localized states are considered. For a single discrete electronic state, an analytic treatment of the interaction with LO-phonons is possible within the "independent boson model" (Mahan, 1990). The numerical extension to several discrete electronic states is discussed in Stauber et al., (2000). However, in this work, the influence of delocalized WL states with a quasi-continuum of energies, which plays a key role for carrier generation in optoelectronic devices, has been neglected. When investigations are restricted to localized states with discrete energies, RPA results deviate from a more accurate direct diagonalization of a restricted system. On the other hand, the inclusion of WL states provides a natural source of broadening of the electronic states, which justifies the application of the RPA.

The next step in the analysis of carrier–phonon scattering in QD systems was the development of a theory that accounts for the scattering of renormalized quasiparticles (QD-polarons). This became possible within a quantum-kinetic theory that has been applied to the problem of carrier capture and relaxation (Seebeck et al., 2005). Such a quantum-kinetic description naturally incorporates both non-Markovian effects on ultrafast timescales and a nonperturbative treatment of the carrier–phonon interaction in the polaron picture. Polaronic renormalization effects have been found to be the main origin for fast carrier capture and relaxation processes, as well as efficient carrier thermalization on a picosecond timescale. The results strongly deviate from a perturbative treatment using Fermi's golden rule.

40.3.2.2 Polaron States and Kinetics

In the following, the retarded Green's function $G^{\mathrm{r}}_{\alpha}(\tau)$ with the time τ and the state index α is used to describe the quasiparticle renormalizations of QD electrons and holes in the single-particle states α due to the nonperturbative interaction with LO-phonons. In the absence of interaction, this function oscillates with the free-carrier energy ε_α according to

$$G^{\mathrm{r}}_{\alpha}(\tau) = \frac{i}{\hbar}\Theta(\tau)e^{-\frac{i}{\hbar}\varepsilon_\alpha \tau} \tag{40.9}$$

and the Fourier transform

$$G^{\mathrm{r}}_{\alpha}(\omega) = \frac{1}{\hbar\omega - \varepsilon_\alpha + i\eta}, \tag{40.10}$$

with the frequency ω has a pole at the free-carrier energy with $\eta > 0$, $\eta \to 0$. Quasiparticle renormalizations are obtained from the solution of the Dyson equation:

$$\left[i\hbar\frac{\partial}{\partial\tau} - \varepsilon_\alpha\right]G^{\mathrm{r}}_{\alpha}(\tau) = \delta(\tau) + \int\limits_0^\tau d\tau' G^{\mathrm{r}}_{\alpha}(\tau')\sum_\beta G^{\mathrm{r}}_{\beta}(\tau-\tau')D^{>}_{\alpha\beta}(\tau-\tau'), \tag{40.11}$$

where the sum involves all available states β. The Dyson equation contains a self-energy for the many-body description of interaction processes. In Equation 40.11, the so-called random-phase approximation has been used to formulate the interaction term with the phonon propagator

$$D^{\gtrless}_{\alpha\beta}(\tau) = \sum_{\vec{q}} |M_{\alpha\beta}(\vec{q})|^2 \times [n_{\mathrm{LO}}e^{\pm i\omega_{\mathrm{LO}}\tau} + (1+n_{\mathrm{LO}})e^{\mp i\omega_{\mathrm{LO}}\tau}], \tag{40.12}$$

where

n_{LO} is a Bose–Einstein function for the population of the phonon modes (assumed to act as a bath in thermal equilibrium)

ω_{LO} is the LO-phonon frequency

The sum over the three-dimensional phonon wave vector $\vec{q}$ involves all phonon modes. For an introduction into the many-body theory in general and Green's functions in particular, we refer the reader to Schäfer and Wegener (2002) and Mahan (1990).

The Dyson equation (Equation 40.11) takes into account all possible virtual transitions due to carrier–phonon interaction for a carrier in the state α. The single-particle energies for both localized and delocalized states are considered, and the corresponding wave functions enter via the interaction matrix elements $M_{\alpha\beta}$.

The Fourier transform of the retarded Green's function directly provides the spectral function $\hat{G}_\alpha(\omega) = -2\,\mathrm{Im}\,G^{\mathrm{r}}_{\alpha}(\omega)$ that reflects the density of states (DOS) for a carrier in the state α. Only for noninteracting carriers, the spectral function contains a delta-function at the free-particle energies, expressing that each state is associated with a single energy. This picture changes due to the quasiparticle renormalizations.

The dynamics of the carrier population due to interaction with LO-phonons can be described with a quantum-kinetic equation (Schafer and Wegener, 2002; Seebeck et al., 2005)

$$\begin{aligned}\frac{\partial}{\partial t}f_\alpha(t) = 2\,\mathrm{Re}\sum_\beta\int\limits_{-\infty}^{t} dt' G^{\mathrm{r}}_{\beta}(t-t')[G^{\mathrm{r}}_{\alpha}(t-t')]^\star \\ \times\left\{f_\alpha(t')\left[1-f_\beta(t')\right]D^{<}_{\alpha\beta}(t'-t) - [1-f_\alpha(t')f_\beta(t')D^{>}_{\alpha\beta}(t'-t)]\right\}.\end{aligned} \tag{40.13}$$

In comparison to Boltzmann scattering rates, the delta-functions are replaced by convolutions of the phonon propagator with retarded Green's functions. The latter account for the polaronic renormalizations of the initial and final carrier states. Furthermore, memory effects are included by the time dependence of the population factors on the past evolution of the system.

Boltzmann scattering integrals follow as a limiting case, when the Markov approximation is applied (population functions $f_\alpha(t')$ are taken at the external time t) and when quasiparticle renormalizations are neglected with the use of free-carrier retarded Green's functions (40.9).

Examples for ultrafast carrier scattering processes in the polaron picture are displayed in Figure 40.5. Even if the level

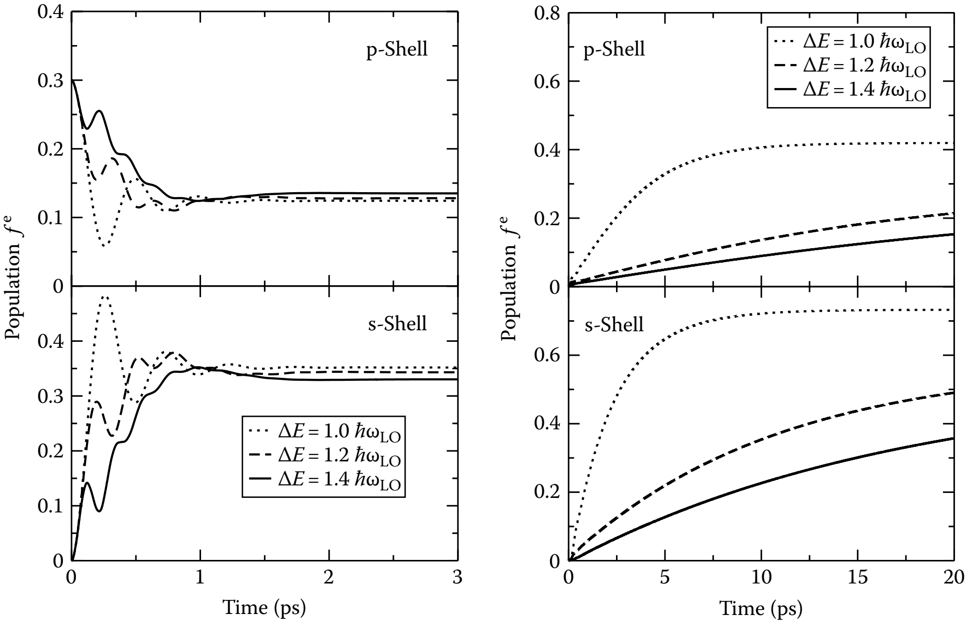

FIGURE 40.5 Temporal evolution of the QD electron population due to scattering of carriers with LO-phonons. Calculations have been performed in the polaron picture for the InGaAs/GaAs material system at room temperature. Left: carrier relaxation from the initially populated p-shell (top) into the initially empty s-shell (bottom). Right: carrier capture from the WL into the initially empty p-shell (top) and s-shell (bottom). Different energy spacings ΔE between the s- and p-shell in units of the LO-phonon energy are compared. (From Seebeck, J. et al., *Phys. Rev. B*, 71, 125327, 2005. With permission.)

spacing between s and p-shell and between p-shell and WL band-edge exceeds the LO-phonon energy by 40%, efficient carrier relaxation and carrier capture are observed, while Boltzmann scattering integrals would predict vanishing carrier transitions for the interaction with LO-phonons.

40.4 Optical Gain of the Active Material

Of central importance for the design process and for various emission properties of lasers are the optical gain of the active material and the corresponding refractive index changes. Both quantities are linked via the optical susceptibility, which is introduced in this section.

In the following, we give some examples for the role of the optical gain. The threshold current of a laser is determined by the transparency carrier density at which the active material switches from absorption to optical gain. The carrier-density dependence of the gain influences the modulation bandwidth, and the temperature dependence of the optical gain is the main factor for the temperature stability of the laser operation. It is also important to study the physical mechanisms of gain saturation, which limits the achievable optical gain.

For the edge-emitting laser structures, the optical mode propagates in the QD plane and refractive index changes, which are induced by excited carriers, directly influence the mode properties like frequency chirp or filamentation, as well as the laser linewidth. In the following, we establish a microscopic picture to determine these gain and refractive index properties in QD systems.

40.4.1 Optical Susceptibility

From Maxwell's equations, a wave equation for the optical field $\boldsymbol{E}(\boldsymbol{r}, t)$ can be derived, which describes how the emitted field from the sample can be traced back to the macroscopic polarization $\boldsymbol{P}(\boldsymbol{r}, t)$ inside the sample:

$$\left(-\nabla\times\nabla\times-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\boldsymbol{E}(\boldsymbol{r}, t)=\mu_0\frac{\partial^2}{\partial t^2}\boldsymbol{P}(\boldsymbol{r}, t). \tag{40.14}$$

The polarization represents the macroscopic dipole density in the medium, which is induced by the field itself, as expressed by the optical susceptibility

$$\boldsymbol{P}(\boldsymbol{r}, t)=\int d^3r'\int dt'\chi(\boldsymbol{r}, \boldsymbol{r}', t, t')\boldsymbol{E}(\boldsymbol{r}', t'). \tag{40.15}$$

In order to fulfill Equations 40.14 and 40.15 simultaneously, a self-consistent solution is necessary: the field determined from the wave equation should be the same function entering the calculation of the polarization. For simplicity, we neglect the tensorial character of χ and consider given polarization directions for $\boldsymbol{P}$ and $\boldsymbol{E}$.

From Equation 40.15 can be inferred that the susceptibility is a response function describing the answer of the medium to the optical field. Mathematically this can be expressed with the functional derivative

$$\chi(\boldsymbol{r}, \boldsymbol{r}', t, t')=\frac{\delta P(\boldsymbol{r}, t)}{\delta E(\boldsymbol{r}', t')}. \tag{40.16}$$

The optical absorption spectrum is commonly defined via the *linear response* of the medium. In this case, χ depends only on the difference of the time arguments (as can be shown from the explicit quantum mechanical calculation of the macroscopic polarization). Then one obtains in Equation 40.15 a convolution in time, which translates into a product in Fourier space and to give

$$\chi_{\mathrm{QD}}(\omega) = \frac{P_{\mathrm{QD}}(\omega)}{E_{\mathrm{QD}}(\omega)}. \tag{40.17}$$

For notational simplicity, we have no longer written the space dependence of the functions. In the considered case of an active material of QDs, the optical susceptibility is given as the macroscopic QD polarization P_{QD} divided by the linear optical test field at the QD position E_{QD}.

The absorption spectrum $\alpha(\omega)$ as well as the refractive index changes of the medium due to resonant excitation $\delta n(\omega)$ of the QD system is given by

$$-K\delta n(\omega) + i\alpha(\omega) = \frac{\omega}{\varepsilon_0 n_B c L}\chi_{\mathrm{QD}}(\omega), \tag{40.18}$$

where

c and ε_0 are the speed of light and permittivity in vacuum
n_B is the background refractive index
L is the thickness of the QD sheet
K is the wave number of the optical field

40.4.2 Interband Transitions and Macroscopic Polarization

The macroscopic optical QD polarization can be determined from the quantum mechanical expectation value of the dipole operator $\langle \boldsymbol{d}\rangle = \mathrm{Tr}\{\boldsymbol{d}\rho\} = \sum_{\nu_1,\nu_2} \boldsymbol{d}_{\nu_1,\nu_2}\rho_{\nu_2,\nu_1}$, which is calculated with the statistical operator ρ. Since $\boldsymbol{d}$ is a single-particle operator, a partial trace over the statistical operator can be performed in a way that only the single-particle statistical operator with the matrix elements ρ_{ν_1,ν_2} contributes. The sum involves all single-particle states of the chosen basis, as discussed for QD systems in Section 40.2.

In the following, we further specify the single-particle density-matrix elements. The diagonal elements represent the population of the state ν and have already been abbreviated by f_ν in Section 40.3. The off-diagonal elements $\Psi_{\nu_1\nu_2}$ are transition amplitudes between the corresponding states. Then the single-particle density matrix has the form

$$\rho_{\nu_1,\nu_2} = \begin{pmatrix} f_{\nu_1} & \Psi_{\nu_1,\nu_2} \\ \Psi^{\star}_{\nu_1,\nu_2} & f_{\nu_2} \end{pmatrix}. \tag{40.19}$$

For the calculation of the optical polarization, we use that the diagonal dipole-matrix elements vanish exactly. Also, we consider only interband transition contributions to the dipole coupling. For the final calculation of the macroscopic QD polarization, we assume that the optical field averages over sufficiently many QDs to obtain

$$\boldsymbol{P}_{\mathrm{QD}}(t) = n_{\mathrm{QD}}\langle \boldsymbol{d}\rangle = n_{\mathrm{QD}}\sum_{\nu_1,\nu_2}\boldsymbol{d}_{\nu_1,\nu_2}\Psi_{\nu_1,\nu_2}(t), \tag{40.20}$$

with the (sheet) density of QDs n_{QD}.

40.4.3 Optical Gain Calculations

A consistent microscopic theory for the calculation of coherent optical properties (related to the interaction of the active material with a classical optical field) can be formulated on the basis of equations of motion for the single-particle density matrix elements. These equations can be derived from the Hamiltonian of the QD system, which contains the contributions of free-carriers, their dipole interaction with the optical fields, as well as further relevant interaction processes like the Coulomb interaction of carriers and the carrier–phonon interaction.

When only the free-carrier Hamiltonian and the dipole interaction with the optical field are considered, the density matrix elements obey the optical Bloch equations (Meystre and Sargent III, 1991). We obtain coupled equations for the interband transition amplitude $\Psi_{\nu_1,\nu_2}(t)$ and for the occupation probabilities $f_{\nu_{1,2}}(t)$.

$$i\hbar\frac{\partial}{\partial t}\Psi_{\nu_1,\nu_2} = \left(\epsilon_{\nu_1} - \epsilon_{\nu_2} - \frac{i\hbar}{T_2}\right)\Psi_{\nu_1,\nu_2} + (f_{\nu_1} - f_{\nu_2})d_{\nu_1,\nu_2}E_{\mathrm{QD}}, \tag{40.21}$$

$$\frac{\partial}{\partial t}f_{\nu_1} = \frac{i}{\hbar}\sum_{\nu_2}[d_{\nu_1,\nu_2}E_{\mathrm{QD}}\Psi^{\star}_{\nu_1,\nu_2} - c.c.] - \frac{f_{\nu_1} - F_{\nu_1}}{T_1}, \tag{40.22}$$

$$\frac{\partial}{\partial t}f_{\nu_2} = -\frac{i}{\hbar}\sum_{\nu_1}[d_{\nu_1,\nu_2}E_{\mathrm{QD}}\Psi^{\star}_{\nu_1,\nu_2} - c.c.] - \frac{f_{\nu_2} - F_{\nu_2}}{T_1}. \tag{40.23}$$

A phenomenological way to include the influence of carrier scattering and dephasing is via time constants $T_{1,2}$, respectively. For the last term in Equations 40.22 and 40.23, a relaxation-time approximation has been used, which describes the evolution of the occupation probabilities $f_{\nu_{1,2}}$ toward quasi-equilibrium functions $F_{\nu_{1,2}}$ on a time-scale T_1. Damping of the interband transition amplitude Ψ_{ν_1,ν_2}, which determines the linewidth of the interband transitions, is accounted for with the dephasing time T_2 in Equation 40.21.

The stationary solution of Equation 40.21 together with Equations 40.17 and 40.20 allows to calculate the optical susceptibility in the form

$$\chi_{\mathrm{QD}}(\omega) = n_{\mathrm{QD}}\sum_{\nu_1,\nu_2}|d_{\nu_1,\nu_2}|^2(f_{\nu_1} - f_{\nu_2}) \times\left[\frac{1}{\hbar\omega - \varepsilon_{\nu_{12}} + \frac{i}{T_2}} - \frac{1}{\hbar\omega - \varepsilon_{\nu_{12}} + \frac{i}{T_2}}\right]. \tag{40.24}$$

The imaginary part of $\chi_{\text{QD}}(\omega)$ describes the absorption spectrum of the system and reveals discrete lines for interband transitions between the QD states $\nu_{1,2}$ with a linewidth $\propto 1/T_2$ at the interband energies $\varepsilon_{\nu_{1,2}} = \varepsilon_{\nu_1} - \varepsilon_{\nu_2}$. When the population in the upper state exceeds that in the lower state, the population factor $f_{\nu_1} - f_{\nu_2}$ changes its sign and the system switches from absorption to optical gain.

It should be noted that the optical Bloch equations (Equations 40.21 through 40.23) and the resulting susceptibility (40.24) represent only a free-carrier picture, which is usually a poor approximation for a quantitative analysis. Important is the additional inclusion of Coulomb interaction effects in the optical Bloch equations for semiconductor systems (Haug and Koch, 2004), which is responsible for different types of effects. First of all, the interband Coulomb exchange interaction between electrons and holes accounts for excitonic effects at low excitation densities and interband Coulomb enhancement of the optical transitions. Secondly, energy renormalization shows up as excitation-dependent shifts of the transition energies, of which the band-gap renormalization is the most prominent signature. A closer inspection reveals, however, nonrigid shifts of all states. As a third modification, the microscopic replacement for scattering terms, expressed by the relaxation approximation in Equations 40.22 and 40.23, has been provided in Section 40.3. For consistency, similar terms need to be used in Equation 40.21 to describe the excitation-induced dephasing, which replaces the constant dephasing time T_2. Furthermore, screening of the Coulomb interaction contributes to a nonlinear dependence of the optical properties on the excitation conditions.

40.4.4 Gain Saturation, Refractive Index, and α-Factor

A microscopic theory for gain calculations in QD systems, which includes excitonic effects as well as excitation-induced energy renormalization and dephasing due to the screened Coulomb interaction and carrier–phonon interaction, has been developed in Lorke et al. (2006a). An example for the transition of a QD system from absorption to gain with increasing excitation density is shown in Figure 40.6. The calculations are performed at room temperature and assume a quasi-equilibrium distribution of the carriers over the QD and WL states due to efficient carrier-scattering processes. For a low excitation density (dotted line), QD s-shell and p-shell transitions as well as the WL exciton resonance at $E - E_G \approx -138$, -80, and -17 meV, respectively, are broadened due to carrier–phonon and carrier–carrier scattering processes. With increasing carrier density, the transition lines are bleached (due to phase-space filling) and further broadened (due to increased scattering efficiency). Note that bleaching reduces the oscillator strength, which needs to be distinguished from the increasing the linewidth (broadening). With population inversion of the QD states, optical gain is realized. Just above the transparency carrier density, only the s-shell transition exhibits optical gain. At higher carrier densities, the optical gain of the p-shell and of the WL are taking over. The WL gain is due to the increased population of the corresponding quantum-well-like states. The combination of band-gap shrinkage and band filling positions the WL gain near the low-density WL exciton resonance. The related small excitation-dependent shifts of the quantum well gain are discussed in Chow and Koch (1999).

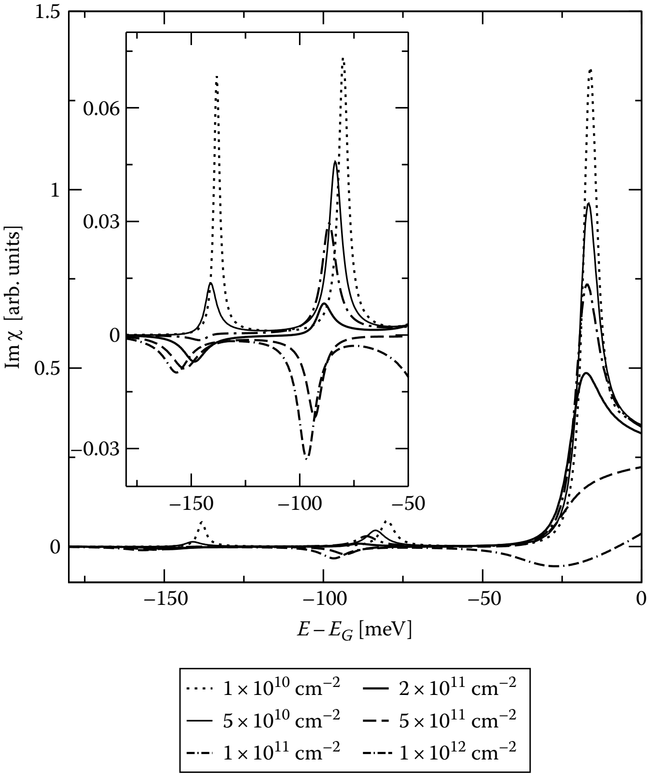

FIGURE 40.6 Imaginary part of the optical susceptibility, representing the absorption spectrum, versus energy relative to the band edge of the WL for a InGaAs/GaAs QD system. The spectrum contains two interband transitions due to confined states, which are magnified in the inset. (From Lorke, M. et al., *Phys. Rev. B,* 73, 085324, 2006. With permission.)

When the QD states are completely populated with electrons and holes, a further increase of the total carrier density in the system does not lead to a larger optical gain at the corresponding transitions. This is clearly visible at the s-shell transition for the largest carrier densities in Figure 40.6. A higher carrier density in the assumed quasi-equilibrium situation only increases the population of the higher QD states and the WL states. The complete filling of the QD states is the origin for the observed gain saturation. Adding even more carriers to the system finally starts to reduce the QD ground-state gain due to further increasing dephasing (Lorke et al., 2006b).

At low carrier densities, the QD interband transition lines can be associated with QD excitons. The terminology might be viewed with reservations, since the electron–hole pairs are restricted to the QDs by the confinement potential and not—like quantum-well or bulk-semiconductor excitons—bound by the Coulomb interaction. The interplay of direct and exchange Coulomb interaction even allows for "anti-bound" states; a biexciton transition with an energy larger than that of two exciton transitions is the most prominent example. With the term "QD exciton," one emphasizes the role of the Coulomb interaction for characterizing the states. For quantum well and bulk semiconductor excitons, it is known that the delicate balance of energy renormalization (mainly band-gap shrinkage) and reduction of the exciton binding energy due to phase-space filling and screening leads to a nearly constant energetic position of the exciton resonance, which is bleached and broadened for increasing carrier density. The same behavior is observed for the WL exciton in Figure 40.6. However, the QD excitons show a pronounced red shift, which persists as a shift of the gain peaks at elevated carrier densities. When the energy renormalization is determined within a many-body theory formulated in a single-particle basis, it is the result of a partial compensation of state-diagonal and state-nondiagonal self-energies. It can be shown that this partial compensation is strongly reduced in QD systems (Lorke et al., 2006a).

The physics of QD gain spectra is expected to play an important role for QD-microcavity lasers with a small number of dots. In edge-emitting devices, inhomogeneous broadening can mask the discussed effects until the sample quality improves or the laser behavior (saturation effects, gain dynamics) is examined in greater detail.

The excitation-induced refractive index changes are directly linked to the corresponding absorption spectra via the Kramers–Kronig relation (Haug and Koch, 2004), since both can be derived from the complex optical susceptibility according to Equation 40.18. Refractive index changes directly influence the mode properties of edge-emitting lasers, including filamentation, as well as frequency chirp and emission spectra. In the past, the α-factor (linewidth-enhancement factor, antiguiding parameter) has been used to characterize the importance of the excitation-density-induced refractive index changes. In the free-carrier picture outlined above, the discrete nature of the QD states leads to symmetric absorption lines and vanishing refractive index changes at the absorption peak. The resulting value of zero for the α-factor has led to the prediction of many beneficial properties of QD lasers. Calculations based on microscopic semiconductor models show deviations from this simple picture, but also smaller α-factors than for quantum-well lasers (Lorke et al., 2007).

40.5 Laser Emission Properties

The optical gain characterizes the active material itself, while the steady-state and dynamical emission properties of lasers depend more generally on the interplay of the photon and carrier systems. The optical gain determines the rate of stimulated emission of photons into the laser mode. Together with the spontaneous emission rate, which is related to the photoluminescence spectrum, and the cavity losses, these processes govern the laser field inside the laser resonator and the light output. The carrier dynamics is determined by the interplay of the pump process, the transfer of carriers into the laser levels, as well as the optical processes.

The most intuitive approach to the steady-state and dynamical properties of lasers is provided by rate equations, which will be introduced in Section 40.5.1. Such a theory is based on restricting the information about the laser output to the mean photon number in the laser mode, as well as on a convenient parametrization of the rates for various processes. A generalization, which allows the systematic inclusion of semiconductor effects, is outlined in Section 40.5.2. The statistical properties of the light emission as well as the control of light–matter interaction in microcavity lasers are addressed in Section 40.5.3.

40.5.1 Rate Equations for Quantum-Dot Systems

In a rate-equation description (Yokoyama and Brorson, 1989; Rice and Carmichael, 1994), one uses two coupled dynamical equations for the number of photons in the laser mode n and the number of excited emitters N, which are QDs for the present purpose:

$$\frac{d}{dt}N = P - \frac{nN}{\tau_l} - \frac{N}{\tau_{sp}}, \tag{40.25}$$

$$\frac{d}{dt}n = -2\kappa n + \frac{(n+1)N}{\tau_l}. \tag{40.26}$$

The pump rate P increases the emitter number, while the cavity loss rate 2κ reduces the photon number. The loss rate is directly connected to the Q-factor of the laser mode, $Q = \hbar\omega_l/2\kappa$, with the laser frequency ω_l. The total rate of spontaneous emission N/τ_{sp}, which includes emission into all available lasing and nonlasing modes, reduces the number of excited emitters, while only the spontaneous emission directed into the laser mode N/τ_l contributes to the increase of the respective photon number. The simulated emission is additionally proportional to the mean photon number n, and the corresponding rate nN/τ_l appears in both rate equations with opposite signs.

An important parameter to characterize laser resonators is the β-factor, which is defined as the ratio of the spontaneous emission rate into the laser mode to the total spontaneous emission according to

$$\beta = \frac{1/\tau_l}{1/\tau_{sp}}. \tag{40.27}$$

In edge-emitting laser diodes, the large number of nonlasing modes leads to small values of $\beta = 10^{-6} \ldots 10^{-5}$, while in microcavity lasers the spontaneous emission into nonlasing modes can be strongly suppressed and $\beta \approx 1$ is approached.

A general evaluation of the rate equations can be readily performed by means of a direct numerical solution in time for a given pump rate P. If the initial condition is the unexcited system and the pump rates are switched on to a constant value, the solution either shows a smooth evolution to the steady-state or damped relaxation oscillations (Milonni and Eberly, 1991), depending on the chosen parameters.

The results for the steady-state solution and various β-factors are shown in Figure 40.7. A set of parameters referring to

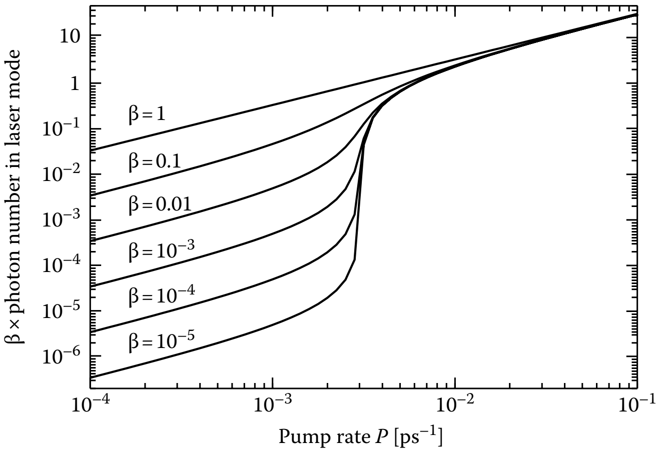

FIGURE 40.7 Calculated input–output curves from the laser rate equations for $\beta = 1$ to 10^{-5} from top to bottom. The photon number and the pump rate are scaled with β in order to have the thresholds appear at equal pump intensities for better comparison. (From Gies, C. et al., *Phys. Rev. A*, 75, 013803, 2007. With permission.)

QD-microcavity lasers as been used: $\tau_{sp} = 50\,\text{ps}$ and $\kappa = 20\,\mu\text{eV}$. The corresponding cavity lifetime is about 17 ps, yielding a Q-factor of ≈30,000. The curves show a typical intensity jump $\propto \beta^{-1}$ from below to above threshold. In the limit $\beta = 1$, the kink in the input–output curve disappears.

The above discussed form of the rate equations is based on various assumptions. Only one optical mode is subject to stimulated emission; otherwise separate rate equations for the photon numbers in different laser modes need to be used and their modal gain determining τ_l needs to be distinguished. The derivation of the rate equations from a microscopic theory is based on the adiabatic elimination of transition amplitudes as well as on the factorization of carrier and photon correlations (Rice and Carmichael, 1994). Furthermore, in the rate equations, one assumes that the spontaneous and stimulated recombination rates are proportional to number of excited emitters N. Our discussion of the optical gain in Section 40.4.3 has already revealed that this is only a rough estimate neglecting optical nonlinearities and saturation effects. Furthermore, each QD is assumed to possess only two possible configurations, the excited and de-excited state. In reality, QDs often contain more than two localized states that can be occupied by several charge carriers. This can lead to various modifications of the simple "atomic approach."

The form of the spontaneous recombination term in the laser rate equations predicts an exponential decay of the photoluminescence after pulsed excitation and in the absence of stimulated emission. Recent experiments with self-assembled InGaAs/GaAs QDs embedded in GaAs-based micropillars (Schwab et al., 2006) showed, however, a nonexponential decay of the time-resolved photoluminescence. Furthermore, this decay was shown to be accompanied by a strong dependence on the excitation intensity. These two effects make it impossible to assign a single spontaneous emission lifetime to the QDs and question the simple picture used in the rate equations. This further underscores the importance of semiconductor models for a more detailed microscopic description of photoluminescence and gain in QD systems.

40.5.2 Microscopic Generalization: Semiconductor Effects

A semiconductor theory for QD lasers can be formulated on the basis of equations of motions for the carrier occupation probabilities of the QD levels and the photon number in the laser mode. For a microscopic derivation, one starts from the Hamiltonian describing the quantized electromagnetic field, the carrier system in second quantization, and the carrier–photon interaction in dipole approximation. Further interaction processes, like the Coulomb interaction of carriers and the carrier–phonon interaction, can be included systematically. Heisenberg equations of motion for the carrier and photon operators can be used to derive dynamical equations for the mean photon number ($n = b^\dagger b$) and the carrier occupation functions $f_\nu^e = \langle c_\nu^\dagger c_\nu \rangle, f_\nu^h = 1 - \langle v_\nu^\dagger v_\nu \rangle$. In these expectation values, $b^\dagger$ and b are photon creation and annihilation operators, $c_\nu^\dagger$ and c_ν are creation and annihilation operators for conduction-band carriers in the states ν and $v_\nu^\dagger v_\nu$ are the corresponding valence-band operators, respectively. The contribution of the interaction of carriers with the laser mode then leads to

$$\left(\frac{d}{dt} + 2\kappa\right) n = \frac{2}{\hbar} |g|^2 \sum_{\nu'} \text{Re}\langle b^\dagger v_{\nu'}^\dagger c_{\nu'} \rangle, \tag{40.28}$$

$$\left.\frac{d}{dt} f_\nu^{e,h}\right|_{\text{opt}} = -\frac{2}{\hbar} |g|^2 \,\text{Re}\,\langle b^\dagger v_v^\dagger c_v \rangle, \tag{40.29}$$

with the light–matter coupling strength g. It can be seen that the corresponding dynamics of the photon number and carrier populations are determined by the photon-assisted polarization $\langle b^\dagger v_\nu^\dagger c_\nu \rangle$, that describes the expectation value for a correlated event, where a photon is created in connection with an interband transition of an electron from the conduction to the valence band. The sum over ν involves all possible interband transitions from various QDs.

The time evolution of the photon-assisted polarization follows from its equation of motion,

$$\begin{aligned}
&\left(\frac{d}{dt} + \kappa + \Gamma + i(\tilde{\varepsilon}_\nu^e + \tilde{\varepsilon}_\nu^h - \hbar\omega_l)\right) \langle b^\dagger v_\nu^\dagger c_\nu \rangle \\
&\quad = f_\nu^e f_\nu^h - (1 - f_\nu^e - f_\nu^h) n \\
&\qquad + i(1 - f_\nu^e - f_\nu^h) \sum_\alpha V_{\nu\alpha\nu\alpha} \langle b^\dagger v_\alpha^\dagger c_\alpha \rangle \\
&\qquad + \sum_\alpha C_{\alpha\nu\nu\alpha}^x + \delta\langle b^\dagger b c_\nu^\dagger c_\nu \rangle - \delta\langle b^\dagger b v_\nu^\dagger v_\nu \rangle.
\end{aligned} \tag{40.30}$$

Here, the free evolution of $\langle b^\dagger v_\nu^\dagger c_\nu \rangle$ is determined by the detuning of the QD transitions at the renormalized energies $\tilde{\varepsilon}_\nu^{e,h}$ from the optical mode ω_l. Cavity losses κ and dephasing processes represented by Γ lead to a damping of the time evolution and to a broadening of the spectral components of the optical processes, which are spontaneous and stimulated emission. When deriving the equation of motion for $\langle b^\dagger v_\nu^\dagger c_\nu \rangle$, one finds that the source term of spontaneous emission is described by an expectation value of four carrier operators $\langle c_\alpha^\dagger v_\alpha v_\nu^\dagger c_\nu \rangle$. For uncorrelated carriers, the Hartree–Fock factorization of this source term leads to $f_\nu^e f_\nu^h$, which appears as the first term on the right hand side of Equation 40.30. It describes a spontaneous recombination probability proportional to the occupation probabilities of electrons and holes, as expected in a free-carrier picture, but opposed to the situation in an atomic system, where the recombination depends only on the electron population. However, electrons and holes will not just contribute as independent carriers. Correlation contributions to the spontaneous emission are included in $C^x_{\alpha\nu\nu\alpha} = \delta\langle c_\alpha^\dagger v_\nu^\dagger c_\nu v_\alpha \rangle$. These correlations describe the joint probability of a two-particle process, where two interband carrier transitions between the states α and ν take place. As mentioned above, the indices include not only the electronic states but also the QD position. Consequently, one must distinguish between correlated transitions within one QD, which can be due to an excitonic population, and correlations of transitions within two separate QDs. The latter case is connected to the phenomenon of superfluorescence or superradiant coupling of different emitters. The Coulomb interaction contributes to the excitonic correlations of electrons and holes, which, in turn, are weakened by screening, phase-space filling, and dephasing effects of the excited carriers. The detailed microscopic description of these correlations is an intricate problem of many-body theory, and for further details we refer the interested reader to Baer et al. (2006).

In a similar fashion, stimulated emission or re-absorption of photons is represented by $(1 - f_\nu^e - f_\nu^h)n$ in the absence of carrier–photon correlations. The latter are represented by $\delta\langle b^\dagger b c_\nu^\dagger c_\nu \rangle$ and $\delta\langle b^\dagger b v_\nu^\dagger v_\nu \rangle$, which require their own equations of motion. Furthermore, the interband Coulomb–exchange interaction with the matrix elements $V_{\nu\alpha\nu\alpha}$ is responsible for excitonic effects similar to its appearance in the semiconductor Bloch equations (Haug and Koch, 2004).

While the semiconductor theory offers the potential to account for various interaction effects in a much more detailed way, this also complicates a numerical analysis significantly. Results from the semiconductor theory for the input–output curves are shown in the lower part of Figure 40.8. Deviations from the rate-equation results involve a different height in the intensity jump at the laser threshold and saturation effects. The semiconductor theory also opens the possibility to calculate parameters entering rate equation models based on single-particle properties and interaction effects.

40.5.3 Quantum-Dot Microcavity Lasers: Modifications of the Photon Statistics

Latest advances in the growth and design of semiconductor-QD microcavity lasers have now attained the regime of β-values close to unity (Strauf et al., 2006; Ulrich et al., 2007). The strongly increased cavity-Q (corresponding to a long lifetime of photons in the cavity) allows to fabricate QD-based lasers with a small number of dots in the active region. The strongly enhanced light-matter coupling and the operation with a small number of photons in the laser mode leads to novel emission properties that are directly related to the quantum-mechanical nature of light. In particular, the photon statistics of the emitted radiation does no longer exhibit the transition from thermal to coherent radiation for increasing pumping, and "nonclassical" properties on the level of few photons can be realized.

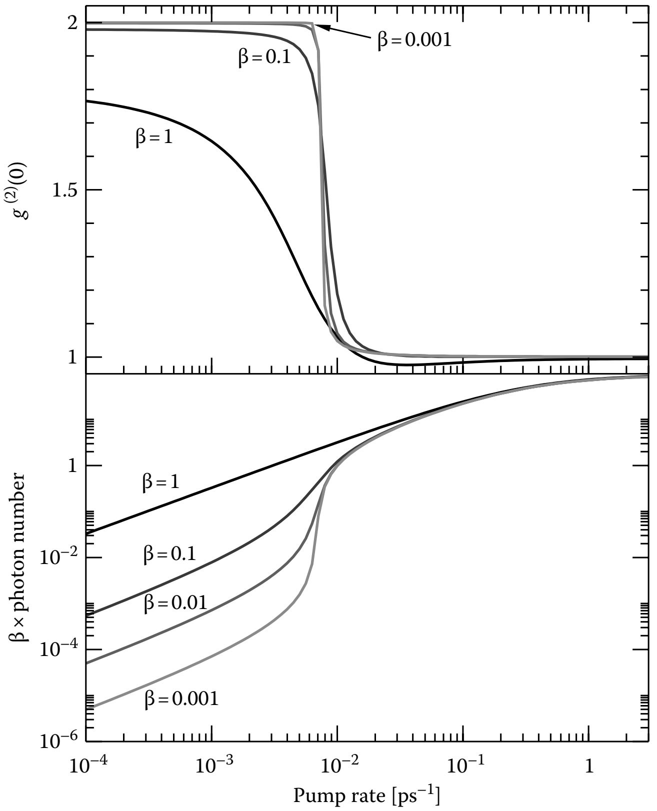

FIGURE 40.8 Input–output curves and auto-correlation functions $g^{(2)}$ ($\tau = 0$) for various β-factors calculated with the semiconductor model. (From Gies, C. et al., *Phys. Rev. A*, 75, 013800, 2007. With permission.)

To characterize these systems, one needs to study the coherence properties of the emitted light and the statistical properties of the photons. Following Glauber, the quantum states of light can be characterized in terms of photon correlation functions. Coherence properties of the electromagnetic field itself are reflected by the (normalized) correlation function of first order,

$$g^{(1)}(\tau) = \frac{\langle b^\dagger(t) b(t+\tau) \rangle}{\langle b^\dagger(t) b(t) \rangle}. \qquad (40.31)$$

Its decay in τ is determined by the coherence time of the emitted light $\tau_c = \int |g^{(1)}(\tau)|^2 \, d\tau$. Here, $b^\dagger$ and b are again the creation and annihilation operators for photons in the laser mode. Information about the statistical properties of the emitted light can be deduced from the correlation function of second order at zero delay time

$$g^{(2)}(\tau = 0) = \frac{\langle n^2 \rangle - \langle n \rangle}{\langle n \rangle^2} = \frac{\langle b^\dagger b b^\dagger b \rangle - \langle b^\dagger b \rangle}{\langle b^\dagger b \rangle^2}. \qquad (40.32)$$

The function $g^{(2)}(\tau = 0)$ reflects the possibility of the correlated emission of two photons at the same time. Often discussed are the limiting cases of light emission from a thermal, a coherent, and a single-emitter light source. Thermal light is characterized by an enhanced probability that two photons are emitted at the same time (bunching), reflected in a value of $g^{(2)}(0) = 2$. For coherent light emission with Poisson statistics, one finds $g^{(2)}(0) = 1$. An ideal single-photon emitter exhibits antibunching with $g^{(2)}(0) = 0$. This, the second-order correlation function, can be used to characterize the emission and to analyze the transition from thermal (or even sub-Poissonian) to coherent light emission.

For atomic systems, various methods have been established to analyze photon correlation functions. A master equation can be used to describe an ensemble of emitters interacting with a single high-Q laser cavity mode and a bath of non-lasing modes (Rice and Carmichael, 1994). For a single-atom laser, a direct analysis of the von Neumann equation for the statistical operator including the coupling to dissipative systems is possible (Mu and Savage, 1992). Semiconductor effects can be included in generalized equations of motion for higher-order carrier and photon correlation functions (Gies et al., 2007). Results of a semiconductor theory for the second-order photon correlation function are displayed in the upper part of Figure 40.8. For smaller β-factors referring to conventional lasers, the output-intensity jump at the laser threshold is accompanied with a sudden change of the correlation function between values representing thermal and coherent light. For increasing values of β, the gradually disappearing jump in the photon number is connected with a smoother transition of $g^{(2)}(0)$ and the presence of correlations for small pump rates.

Acknowledgments

The author thanks P. Gartner and J. Wiersig for the intense and productive collaboration on the covered subjects. The contributions from our PhD students T. Nielsen, N. Baer, J. Seebeck, M. Lorke, and C. Gies are also gratefully acknowledged. This work was financially supported by the Deutsche Forschungsgemeinschaft. We also thank P. Gartner and C. Gies for the critical reading of the manuscript and A. Beuthner for the preparation of Figure 40.1.

References

Agrawal, G. P. and N. K. Dutta. 1993. *Semiconductor Laser*. Norwell, MA: Kluwer.

Anders, S., C. S. Kim, B. Klein, M. W. Keller, R. P. Mirin, and A. G. Norman. 2002. Bimodal size distribution of self-assembled In_xGa_{1-x} As quantum dots. *Phys. Rev. B* **66**:125309.

Arakawa, Y. and H. Sakaki. 1982. Multidimensional quantum well laser and temperature dependence of its threshold current. *Appl. Phys. Lett.* **40**:939.

Asada, M., Y. Miyamoto, and Y. Suematsu. 1986. Gain and the threshold of three-dimensional quantum-box lasers. *IEEE J. Quantum Electron.* **QE-22**:1915.

Baer, N., C. Gies, J. Wiersig, and F. Jahnke. 2006. Luminescence of a semiconductor quantum dot system. *Eur. Phys. J. B* **50**:411.

Benisty, H., C. M. Sotomayor-Torres, and C. Weisbuch. 1991. Intrinsic mechanism for the poor luminescence properties of quantum-box systems. *Phys. Rev. B* **44**:10945.

Bimberg, D., M. Grundmann, and N. N. Ledentsov. 1999. *Quantum Dot Heterostructures*. Chichester, U.K.: John Wiley & Sons.

Bockelmann, U. and G. Bastard. 1990. Phonon scattering and energy relaxation in two-, one-, and zero-dimensional electron gases. *Phys. Rev. B* **42**:8947.

Casey, H. C. and M. B. Panish. 1978a. *Heterostructure Lasers—Part A: Fundamental Principles*. New York: Academic Press.

Casey, H. C. and M. B. Panish. 1978b. *Heterostructure Lasers—Part B: Materials and Operating Characteristics*. New York: Academic Press.

Chang-Hasnain, C. J., J. P. Harbison, G. Hasnain, A. C. von Lehmen, L. T. Florenz, and N. G. Stoffel. 1991. Dynamic, polarization, and transverse mode characteristics of vertical cavity surface emitting lasers. *IEEE J. Quantum Electron.* **27**:1402.

Chow, W. W. and S. W. Koch. 1999. *Semiconductor-Laser Fundamentals*. Berlin, Germany: Springer.

Chuang, S. L. 1995. *Physics of Optoelectronic Devices*. New York: Wiley.

Chuang, S. L. and C. S. Chang. 1996. k·p method for strained wurtzite semiconductors. *Phys. Rev. B* **54**:2491.

Dingle, R. and C. H. Henry. 1976. Quantum effects in heterostructure lasers. U.S. Patent 3982207.

Fattal, D., E. Diamanti, K. Inoue, and Y. Yamamoto. 2004. Quantum teleportation with a quantum dot single photon source. *Phys. Rev. Lett.* **92**:037904.

Gies, C., J. Wiersig, M. Lorke, and F. Jahnke. 2007. Semiconductor model for quantum-dot-based microcavity lasers. *Phys. Rev. A* **75**:013803.

Haug, H. and A.-P. Jauho. 1996. *Quantum Kinetics in Transport & Optics of Semiconductors*, 1st edn. Berlin, Germany: Springer.

Haug, H. and S. W. Koch. 2004. *Quantum Theory of the Optical and Electronic Properties of Semiconductors*, 4th edn. Singapore: World Scientific Publishing Company.

Inoshita, T. and H. Sakaki. 1992. Electron relaxation in a quantum dot: Significance of multiphonon processes. *Phys. Rev. B* **46**:7260.

Inoshita, T. and H. Sakaki. 1997. Density of states and phonon-induced relaxation of electrons in semiconductor quantum dots. *Phys. Rev. B* **56**:4355.

Jacobi, K. 2003. Atomic structure of InAs quantum dots on GaAs. *Prog. Surf. Sci.* **71**:185.

Jewell, J. L., J. P. Harbison, A. Scherer, Y. H. Lee, and L. T. Florez. 1991. Vertical-cavity surface-emitting lasers: Design, growth, fabrication, characterization. *IEEE J. Quantum Electron.* **27**:1332.

Jiang, H. and J. Singh. 1998. Self-assembled semiconductor structures: Electronic and optoelectronic properties. *IEEE J. Quantum Electron.* **34**:1188.

Kim, S. M., Y. Wang, M. Keever, and J. S. Harris. 2004. High-frequency modulation characteristics of 1.3 μm InGaAs quantum dot lasers. *IEEE Photon. Technol. Lett.* **16**:377.

Klude, M., T. Passow, R. Kröger, and D. Hommel. 2004. Electrically pumped lasing from CdSe quantum dots. *Electron. Lett.* **37**:1119.

Kral, K. and Z. Khas. 1998. Electron self-energy in quantum dots. *Phys. Rev. B* **57**:2061.

Li, H. E. and K. Iga, eds. 2003. *Vertical-Cavity Surface-Emitting Laser Devices*, 1st edn. Berlin, Germany: Springer.

Lodahl, P., A. F. van Driel, I. S. Nikolaev, A. Irman, K. Overgaag, D. Vanmaekel-bergh, and W. L. Vos. 2004. Controlling the dynamics of spontaneous emission from quantum dots by photonic crystals. *Nature* **430**:654.

Lorke, M., T. R. Nielsen, J. Seebeck, P. Gartner, and F. Jahnke. 2006a. Influence of carrier-carrier and electron-phonon correlations on optical absorption and gain in quantum-dot systems. *Phys. Rev. B* **73**:085324.

Lorke, M., W. W. Chow, T. R. Nielsen, J. Seebeck, P. Gartner, and F. Jahnke. 2006b. Anomaly in the excitation dependence of optical gain in semiconductor quantum dots. *Phys. Rev. B* **74**:035334.

Lorke, M., F. Jahnke, and W. Chow. 2007. Excitation dependence of gain and carrier induced refractive index changes in quantum dots. *Appl. Phys. Lett.* **90**:051112.

Mahan, G. D. 1990. *Many-Particle Physics*, 2nd edn. New York: Plenum.

Meystre, P. and M. Sargent III. 1991. *Elements of Quantum Optics*, 2nd edn. New York: Springer.

Michler, P., A. Kiraz, C. Becher, W. V. Schoenfeld, P. M. Petroff, L. Zhang, E. Hu, and A. Imamoglu. 2000. A quantum dot single-photon turnstile device. *Science* **290**:2282.

Milonni, P. and J. H. Eberly. 1991. *Lasers*, 1st edn. New York: John Wiley & Sons.

Minnaert, A. W. E., A. Yu. Silov, W. van der Vleuten, J. E. M. Haverkort, and J. H. Wolter. 2001. Frhlich interaction in InAs/GaAs self-assembled quantum dots. *Phys. Rev. B* **63**:075303.

Moreau, E., I. Robert, L. Manin, V. Thierry-Mieg, J. M. Gerad, and I. Abram. 2001. Quantum cascade of photons in semiconductor quantum dots. *Phys. Rev. Lett.* **87**:183601.

Mu, Y. and C. M. Savage. 1992. One-atom lasers. *Phys. Rev. A* **46**:5944.

Nakamura, S. and S. F. Chichibu, eds. 2000. *Introduction to Nitride Semiconductor Blue Lasers and Light Emitting Diodes*, 1st edn. Boca Raton, FL: CRC Press.

Newell, T. C., D. J. Bossert, A. Stintz, B. Fuchs, K. J. Malloy, and L. F. Lester. 1999. Gain and linewidth enhancement factor in InAs quantum-dot laserdiodes. *IEEE Photon. Technol. Lett.* **11**:1527.

Nielsen, T. R., P. Gartner, and F. Jahnke. 2004. Many-body theory of carrier capture and relaxation in semiconductor quantum-dot lasers. *Phys. Rev. B* **69**:235314.

Nielsen, T. R., P. Gartner, M. Lorke, J. Seebeck, and F. Jahnke. 2005. Coulomb scattering in nitride-based self-assembled quantum dot systems. *Phys. Rev. B* **72**:235311.

Peronne, E., F. Fossard, F. H. Julien, J. Brault, M. Gendry, B. Salem, G. Bremond, and A. Alexandrou. 2003. Dynamic saturation of an intersublevel transition in self-organized InAs/InAlAs quantum dots. *Phys. Rev. B* **67**:205329.

Quochi, F., M. Dinu, L. N. Pfeiffer, K. W. West, C. Kerbage, R. S. Windeler, and B. J. Eggleton. 2003. Coulomb and carrier-activation dynamics of resonantly excited InAs/GaAs quantum dots in two-color pump-probe experiments. *Phys. Rev. B* **67**:235323.

Reithmaier, J. P., G. Sek, A. Löffler, C. Hofmann, S. Kuhn, S. Reitzenstein, L. V. Keldysh, V. D. Kulakovskii, T. L. Reinecke, and A. Forchel. 2004. Strong coupling in a single quantum dot-semiconductor microcavity system. *Nature* **432**:197.

Rice, P. R. and H. J. Carmichael. 1994. Photon statistics of cavity-QED lasers. *Phys. Rev. A* **50**:4318.

Schäfer, W. and M. Wegener. 2002. *Semiconductor Optics and Transport Phenomena*, 1st edn. Berlin, Germany: Springer-Verlag.

Scholz, M., T. Aichele, S. Ramelow, and O. Benson. 2006. Deutsch-Jozsa algorithm using triggered single photons from a single quantum dot. *Phys. Rev. Lett.* **96**:180501.

Schwab, M., H. Kurtze, T. Auer, T. Berstermann, M. Bayer, J. Wiersig, N. Baer et al. 2006. Radiative emission dynamics of quantum dots in a single cavity micropillar. *Phys. Rev. B* **74**:045323.

Sebald, K., C. Kruse, and J. Wiersig. 2009. Properties and prospects of blue-green emitting II-VI-based monolithic microcavities. *Phys. Stat. Sol.* (*b*) **246**:255.

Seebeck, J., T. R. Nielsen, P. Gartner, and F. Jahnke. 2005. Polarons in semiconductor quantum-dots and their role in the quantum kinetics of carrier relaxation. *Phys. Rev. B* **71**:125327.

Shah, J. 1999. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures*. Berlin, Germany: Springer-Verlag.

Stauber, T., R. Zimmermann, and H. Castella. 2000. Electron-phonon interaction in quantum dots: A solvable model. *Phys. Rev. B* **62**:7336.

Strauf, S., K. Hennessy, M. T. Rakher, Y.-S. Choi, A. Badolato, L. C. Andreani, E. L. Hu, P. M. Petroff, and D. Brouwmeester. 2006. Self-tuned quantum dot gain in photonic crystal lasers. *Phys. Rev. Lett.* **96**:127404.

Tsang, W. T. 1982. Extremely low threshold (AlGa)As graded-index waveguide separate-confinement heterostructure lasers grown by molecular beam epitaxy. *Appl. Phys. Lett*. **40**:217.

Tsitsishvili, E., R. V. Baltz, and H. Kalt. 2002. Temperature dependence of polarization relaxation in semiconductor quantum dots. *Phys. Rev. B* **66**:161405.

Ulrich, S. M., C. Gies, J. Wiersig, S. Reitzenstein, C. Hofmann, A. Löffler, A. Forchel, F. Jahnke, and P. Michler. 2007. Photon statistics of semiconductor microcavity lasers. *Phys. Rev. Lett*. 98:043906.

Urayama, J., T. B. Norris, J. Singh, and P. Bhattacharya. 2001. Observation of phonon bottleneck in quantum dot electronic relaxation. *Phys. Rev. Lett*. **86**:4930.

Vahala, K. 2003. Optical microcavities. *Nature* **424**:839.

Verzelen, O., R. Ferreira, G. Bastard, T. Inoshita, and H. Sakaki. 2002. Polaron effects in quantum dots. *Phys. Stat. Sol.* (*a*) **190**:213.

Vurgaftman, I. and J. R. Meyer. 2003. Band parameters for nitrogen-containing semiconductors. *J. Appl. Phys*. **94**:3675.

Wojs, A., P. Hawrylak, S. Fafrad, and L. Jacak. 1996. Electronic structure and magneto-optics of self-assembled quantum dots. *Phys. Rev. B* **54**:5604.

Xu, S., A. A. Mikhailovsky, J. A. Hollingsworth, and V. I. Klimov. 2002. Hole intra-band relaxation in strongly confined quantum dots: Revisiting the phonon bottleneck problem. *Phys. Rev. B* **65**:045319.

Yokoyama, H. and S. D. Brorson. 1989. Rate equation analysis of microcavity lasers. *J. Appl. Phys*. **66**:4801.

Yoshie, T., A. Scherer, J. Hendrickson, G. Khitrova, H. M. Gibbs, G. Rupper, C. Ell, O. B. Shchekin, and D. G. Deppe. 2004. Vacuum Rabi splitting with a single quantum dot in a photonic crystal nanocavity. *Nature* **432**:200.

41

Mode-Locked Quantum-Dot Lasers

41.1 Introduction **41**-1
41.2 Ultrafast Laser Diodes **41**-2
Basics of Mode Locking • Mode-Locking Techniques in Semiconductor Lasers • Passive Mode Locking: Physics and Devices • Requirements for Successful Passive Mode Locking • Self-Phase Modulation and Dispersion
41.3 Quantum Dots: Distinctive Advantages for Ultrafast Diode Lasers **41**-5
The Role of Dimensionality in Semiconductor Lasers • Quantum Dots: Materials and Growth • Broad Gain Bandwidth • Ultrafast Carrier Dynamics • Low Absorption Saturation Fluence • Low Threshold Current and Low Temperature Sensitivity • Low Linewidth Enhancement Factor
41.4 Mode-Locked Quantum-Dot Lasers: State of the Art **41**-8
Pulse Duration • Toward Higher Pulse Repetition Rates • Temperature Resilience of Mode- Locked QD Lasers • Mode Locking Involving Excited-State Transitions
41.5 Summary and Outlook **41**-10
Critical Discussion • Future Perspectives
Acknowledgments **41**-10
References **41**-10

Maria A. Cataluna
University of Dundee

Edik U. Rafailov
University of Dundee

41.1 Introduction

Over the last three decades, laser physics has advanced dramatically. Starting from lasers operated in a continuous wave (cw) regime, scientists have developed techniques for generating periodic sequences of optical pulses with ultrashort durations—between a few picoseconds (1×10^{-12} s) and a few femtoseconds (1×10^{-15} s). To put this into perspective, 1 fs compared to 1 s is the same as 1 s compared to 32 million years! Such ultrafast lasers have important applications in medicine, micromachining, optical communications, spectroscopy, and anything else that requires studying physics at extremely high powers or extremely short timescales. For instance, these lasers have been successfully adapted in eye surgery, because ultrashort pulses can make extremely precise cuts with minimum thermal damage.

However, despite the wide range of important areas that can benefit from ultrafast lasers, the use of these lasers is constrained due to several limitations. The ultrafast lasers currently available are often bulky, expensive, and difficult to operate. The ideal ultrafast laser would be a low-cost, handheld, and turnkey laser—features which could be offered by semiconductor lasers. Semiconductor lasers cannot yet directly generate the sub-100 fs pulses routinely available from crystal-based lasers, but they represent the most compact and efficient sources of picosecond and sub-picosecond pulses. Furthermore, the bias can be easily adjusted to determine the pulse duration and the optical power, thus offering, to some extent, electrical control of the characteristics of the output pulses. These lasers also offer the best option for the generation of high-repetition rate trains of pulses, owing to their small cavity size. Ultrafast diode lasers have thus been favored over other laser sources for high-frequency applications such as optical data/telecommunications. Being much cheaper to fabricate and operate, ultrafast semiconductor lasers also offer the potential for dramatic cost savings in a number of applications that traditionally use solid-state lasers. The deployment of high-performance ultrafast diode lasers would therefore have a significant economic impact, by enabling ultrafast applications to become more profitable, and even facilitate the emergence of new applications.

Novel nanomaterials such as quantum dots (QDs) have enhanced the characteristics of semiconductor lasers, greatly improving their performance. QDs are tiny clusters of semiconductor material with dimensions of only a few nanometers. At these small sizes, materials behave very differently, giving QDs distinctive physical properties of quantum nature—for instance, the emission wavelength or "color" depends on the size of the dot! These nanomaterials afford major advantages in ultrafast science and technology, and they can form the basis for very compact and efficient lasers delivering short pulses of the order of hundreds of femtoseconds (Rafailov et al., 2007).

In this chapter, we show how QDs have enabled the generation of ultrashort pulses from compact optical sources based on semiconductor laser diodes. In Section 41.2, the necessary background information is presented on ultra-short-pulse generation from diode lasers. The concept of mode locking is introduced, and an overview of the mode-locking techniques available for semiconductor lasers is provided. The unique properties of QD materials and their suitability for ultra-short-pulse diode lasers are explained in Section 41.3. Finally, a summary of the state of the art in the field of QD mode-locked laser diodes is provided in Section 41.4. The chapter is finalized by a summary and an outlook on the future perspectives of this fascinating field.

41.2 Ultrafast Laser Diodes

41.2.1 Basics of Mode Locking

Mode locking is a technique that involves the locking of the phases of the longitudinal modes in a laser. This results in the generation of a sequence of pulses with a repetition rate corresponding to the cavity round-trip time. This well-established technique enables the production of the shortest pulse durations and the highest repetition rates available from ultrafast lasers, whether they are semiconductor or crystal-based laser systems. In a standing-wave resonator, the pulse repetition rate f_R is given by

$$f_R = \frac{c}{2nL}$$

where

c is the speed of light in vacuum
n is the refractive index
L is the length of the laser cavity

In terms of Fourier analysis, there is an inverse proportionality between the duration of a mode-locked pulse and the corresponding bandwidth of its optical spectrum. The product of both the pulse duration $\Delta\tau$ and the optical frequency bandwidth $\Delta\nu$ is called the time-bandwidth product (TBWP). For a given frequency bandwidth, there is a minimum corresponding pulse duration—if this is the case and the optical spectrum is symmetrical, then the pulse is said to be transform-limited, and the TBWP equals a constant K, whose value depends on the shape of the pulse, whether it is Gaussian, hyperbolic, secant, squared, or Lorentzian. By measuring the full-width at half maximum from an optical spectrum $\Delta\lambda$, it is easy to calculate the TBWP of a given pulse:

$$\Delta\nu \cdot \Delta\tau = K \quad \Rightarrow \quad \frac{c}{\lambda^2}\Delta\lambda \cdot \Delta\tau = K$$

Another important property of mode-locked lasers is that the energy that was dispersed in several modes while in cw operation, is now concentrated in short pulses of light. This implies that although the output average power P_{av} may be low, the pulse peak power P_{peak} can be significantly higher:

$$P_{\text{peak}} = \frac{E_{\text{p}}}{\Delta\tau} \quad \Rightarrow \quad P_{\text{peak}} = \frac{1}{\Delta\tau} \cdot \frac{P_{\text{av}}}{f_R}$$

where E_{p} is the pulse energy.

41.2.2 Mode-Locking Techniques in Semiconductor Lasers

In recent years, mode-locked laser diodes have been at the center of a quest for ultrafast, transform-limited, and high-repetition-rate lasers. To achieve these goals, a variety of mode-locking techniques and semiconductor device structures have been demonstrated and optimized (Vasil'ev, 1995). The three main forms of mode locking can be described as active, passive, and hybrid techniques, as outlined below.

Active mode locking relies on the direct modulation of the gain with a frequency equal to the repetition frequency of the cavity, or to a sub-harmonic of this frequency. The main advantages of this approach are the resultant low jitter and the ability to synchronize the laser output with the modulating electrical signal. These features are especially relevant for optical transmission and signal-processing applications. However, high repetition frequencies are not readily obtained through directly driven modulation of lasers because fast RF (radiofrequency) modulation of the drive current becomes progressively more difficult with increase in frequency.

The frequency limitation imposed by electronic drive circuits can be overcome by employing passive mode-locking techniques. This scheme typically utilizes a saturable absorbing region in the laser diode. In a saturable absorber, the loss decreases as the optical intensity increases. This feature acts as a discriminator between cw and pulsed operation and can facilitate a self-starting mechanism for mode locking. Most importantly, saturable absorption plays a crucial role in shortening the duration of the circulating pulses, as will be explained, thus providing the shortest pulses achievable by all three techniques and the absence of a RF source simplifies the fabrication and operation considerably. Passive mode locking also allows for higher pulse repetition rates that are determined solely by the cavity length.

Inspired by active and passive mode locking, the technique of hybrid mode locking meets the best of both worlds because the pulse generation is initiated by an RF current imposed in the gain or absorber section, while further shaping and shortening is assisted by saturable absorption. The next section explores in more detail the physical mechanisms behind passive mode locking.*

---

* From this point, mode locking will implicitly mean passive mode locking, unless otherwise stated.

41.2.3 Passive Mode Locking: Physics and Devices

So far, a simple frequency-domain picture for mode locking has been provided, where the relative phases are of primary relevance. A physical model for passive mode locking can alternatively and equivalently be described in terms of the temporal broadening and narrowing mechanisms.

Upon startup of laser emission, the laser modes initially oscillate with relative phases that are random such that the radiation pattern consists of noise bursts. If one of these bursts is energetic enough to provide a fluence that matches the saturation fluence of the absorber, it will bleach the absorption. This means that around the peak of the burst where the intensity is higher, the loss will be smaller, while the low-intensity wings become more attenuated. The pulse generation process is thus initiated by this family of intensity spikes that experience lower losses within the absorber carrier lifetime.

The dynamics of absorption and gain play a crucial role in pulse shaping. In steady state, the unsaturated losses are higher than the gain. When the leading edge of the pulse reaches the absorber, the loss saturates more quickly than the gain, which results in a net gain window, as depicted in Figure 41.1. The absorber then recovers from this state of saturation to the initial state of high loss, thus attenuating the trailing edge of the pulse. It is thus easy to understand why the saturation fluence and the recovery time of the absorber are of primary importance in the formation of mode-locked pulses.

This temporal scenario can be connected to the previously described frequency domain description of mode locking. The burst of noise is the result of an instantaneous phase locking occurring among a number of modes. The self-saturation at the saturable absorber then helps to sustain and strengthen this favorable combination, by discriminating against the lower power cw noise.

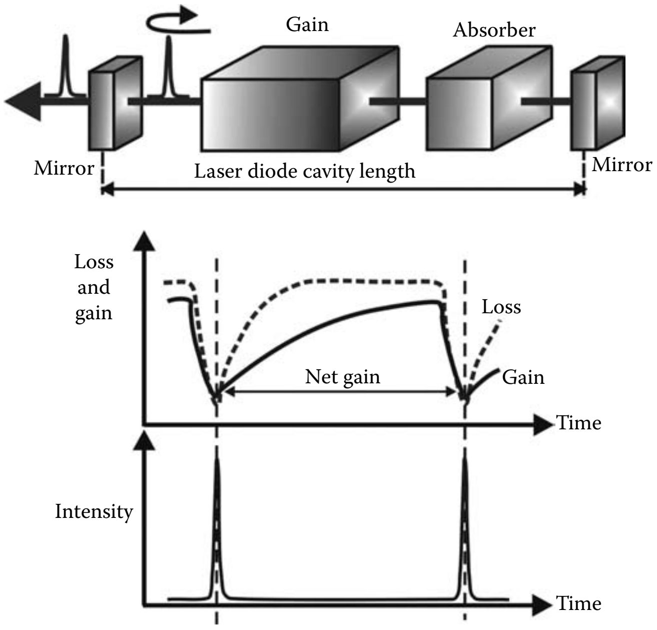

FIGURE 41.1 A schematic diagram of the main components that forms a two-section laser diode (top). Loss and gain dynamics that lead to pulse generation (bottom).

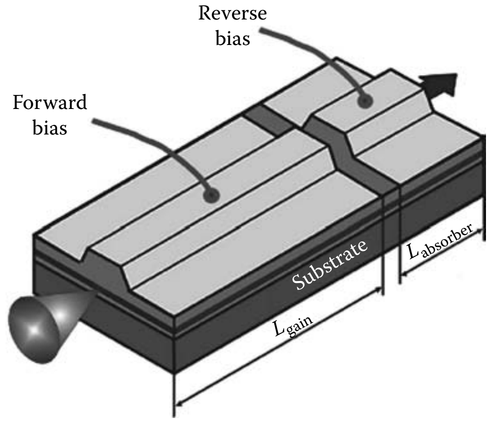

FIGURE 41.2 A schematic of a two-section semiconductor laser diode.

In practical terms, a saturable absorber can be integrated monolithically into a semiconductor laser, by electrically isolating one section of the device (Figure 41.2). By applying a reverse bias to this section, the carriers that are photogenerated by the pulses can be more efficiently swept out of the absorber, thus enabling the saturable absorber to recover more quickly to its initial state of high loss. An increase in the reverse bias serves to decrease the absorber recovery time, and this will have the effect of further shortening the pulses.

41.2.4 Requirements for Successful Passive Mode Locking

Ultrafast carrier dynamics are fundamental for successful mode locking in semiconductor lasers, particularly in the saturable absorber, because the absorption should saturate faster and recover faster. Indeed, the absorption recovery time is one of the determining factors for obtaining ultrashort pulses. In particular, for high-repetition-rate lasers, the absorber recovery time should be much shorter than the cavity period so that the absorber can return to a state of total attenuation prior to the incidence of each incoming pulse. The fast absorption recovery also prevents the appearance of satellite pulses within the window of the net gain. On the other hand, the gain recovery time should be shorter than the cavity round-trip time. Thus, for lasers operating at pulse repetition rates of 20 GHz or more, this means that the recovery times of both gain and absorption should be much shorter than 40 ps.

The saturation dynamics represents another crucial aspect for successful mode locking, as shown schematically in Figure 41.3. Such dynamics can be translated in terms of saturation fluence F_{sat} or saturation energy $E_{sat} = A \cdot F_{sat}$, where A is the optical mode cross-sectional area. The saturation energy is an indication of how much energy is necessary to saturate the absorption or the gain. Indeed, to achieve robust mode locking, the saturation

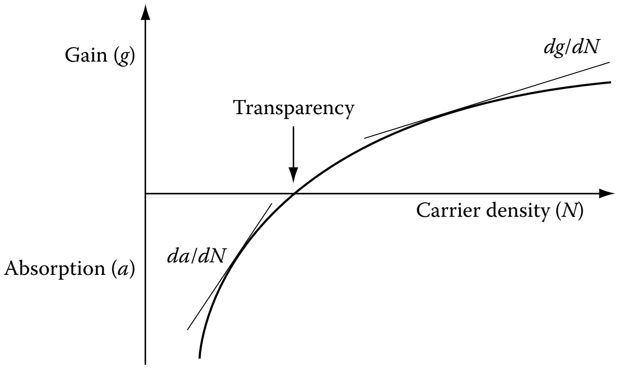

FIGURE 41.3 Dependence of absorption/gain with carrier concentration in a semiconductor laser.

energy of the absorber $E_{\text{sat}}^{\text{a}}$ should be as small as possible and smaller than the saturation energy of the gain $E_{\text{sat}}^{\text{g}}$:

$$E_{\text{sat}}^{\text{a}} = \frac{h\nu A}{\frac{\partial a}{\partial N}} < E_{\text{sat}}^{\text{g}} = \frac{h\nu A}{\frac{\partial g}{\partial N}}$$

where

h is Planck's constant
ν is the optical frequency
$\partial a/\partial N$ and $\partial g/\partial N$ are the differential loss and gain, respectively

The special dependence of the loss/gain with carrier density in a semiconductor laser allows $\partial a/\partial N > \partial g/\partial N$, as shown in Figure 41.3.

This condition implies that the absorber will saturate faster than the gain for a given pulse fluence, thus enabling the creation of the net gain window as already mentioned. The ratio between saturation energies should also be as large as possible to ensure that the losses saturate more strongly than the gain.

41.2.5 Self-Phase Modulation and Dispersion

In a semiconductor material, both the refractive index and gain (or loss) depend on the carrier density and are thus strongly coupled. As the pulse propagates in the gain section,* the carrier density and thus the gain is depleted across the pulse, as the carriers recombine through stimulated emission. This leads to a dynamic increase of the refractive index, which then introduces a phase modulation on the pulse, changing the instantaneous frequency across the pulse. This phenomenon is called self-phase modulation (SPM) and is one of the main nonlinear effects associated with pulse propagation in semiconductor media.

To understand the mechanism of SPM, consider the simple and illustrative example of a plane wave $E(t, x)$:

$$E(t,x) = E_0 \exp i\Phi(t) = E_0 \exp i(\omega_0 t - kx), \quad k = \frac{\omega_0}{c} n(t)$$

where

$\Phi(t)$ is the time-varying phase
k is the wave vector
ω_0 is the optical carrier frequency
c is the speed of light
$n(t)$ is the time-varying refractive index

The instantaneous frequency is the time derivative of the phase and thus can be written as

$$\omega(t) = \frac{\partial}{\partial t}\Phi(t) = \omega_0 - \frac{\omega_0}{c}\frac{\partial n(t)}{\partial t}x$$

From this expression, it is clear that if the refractive index varies with time, then the instantaneous frequency of the plane wave will vary relative to ω_0 and in a manner proportional to the temporal derivative of the index. The time dependence of this instantaneous frequency is called the frequency chirp. An up-chirp† (down-chirp) means that the frequency increases (decreases) with time. An example of a frequency up-chirped pulse is illustrated in Figure 41.4.

SPM is not dispersive in itself, but the pulse will not remain transform-limited when it propagates in a dispersive material such as the laser medium. The effect of dispersion manifests itself in the variation of refractive index for different wavelengths which means that different spectral components will travel at different speeds. For an up-chirped pulse, the frequency is higher in the trailing edge than in the leading edge. When the pulse propagates through a material exhibiting positive (normal) dispersion, the trailing edge of the pulse propagates more slowly than the leading edge of the pulse and so this results in a temporal broadening of the pulse.

In a monolithic two-section mode-locked semiconductor laser, where a saturable absorber and a gain section coexist, the resulting chirp is a balance between the effects caused by the absorber and the gain. In the gain section, a frequency up-chirp results, while the saturable absorber helps to further

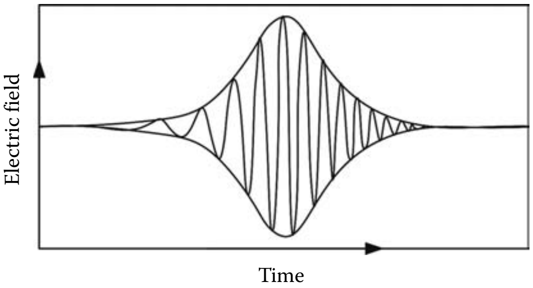

FIGURE 41.4 Illustration of an electric field of a strongly up-chirped pulse, where the instantaneous frequency increases with time.

---

[*] In the following discussion, reference is made mostly to gain to simplify the description. However, all this reasoning can be applied equally well to the absorber.

[†] Up-chirp is also known as blue-chirp or positive chirp.

shape the pulse by contributing with a negative chirp. With a suitable balance between both sections, the chirp can be close to zero thereby leading to transform-limited pulses. Unfortunately, this is the exception rather than the rule, because this usually only occurs for a limited set of bias conditions and/or for given ratios of absorber/gain lengths. Therefore, up-chirp prevails for passively mode-locked lasers, leading to significant pulse broadening as the pulse propagates. The combined effect of SPM and dispersion impose the strongest limitation in the achievable shortest duration of pulses from mode-locked semiconductor diode lasers.

The mechanism of SPM implies that in addition to the original frequency ω_0, there are now more frequencies inside the pulse envelope. This richer spectral content is not necessarily unhelpful because it can provide bandwidth support for shorter pulses, if the chirp of a pulse can be removed by provision of a suitable dispersion-induced chirp of the opposite sign. For up-chirped pulses, a dispersion compensation setup can be configured such that a negative (anomalous) group velocity dispersion is able to slow down the leading edge of the pulse and speed the blueshifted trailing edge to such an extent that at a certain point both edges propagate simultaneously and the pulse is shorter.

To routinely generate pulses that are nearly transform-limited, an alternative could be found in the choice of a material that exhibits lower coupling between refractive index and gain, as described by linewidth enhancement factor (LEF), or α-factor:

$$\alpha = -\frac{4\pi}{\lambda}\frac{dn/dN}{dg/dN}$$

A higher α-factor implies a more significant coupling between gain and refractive index changes with carrier concentration and thus the possibility for higher levels of SPM and frequency chirp.

41.3 Quantum Dots: Distinctive Advantages for Ultrafast Diode Lasers

41.3.1 The Role of Dimensionality in Semiconductor Lasers

The history of semiconductor laser materials has been punctuated by dramatic revolutions. Everything started with the proposal of p-n junction semiconductor lasers in 1961, followed by experimental realization on different semiconductor materials (Basov et al., 1961; Basov, 1964). However, the lasers fabricated at that time exhibited an extremely low efficiency due to high optical and electrical losses. In fact, until the mid-1960s, only bulk materials were used in semiconductor devices, which were functionalized by introducing a doping profile. At the time, pioneers like Alferov and Herbert Kroemer independently considered the hypothesis of building heterostructures, consisting of layers of different semiconductor materials (Alferov, 2001). The classic heterostructure example consists of a lower bandgap layer surrounded by a higher bandgap semiconductor material. Such design results in electronic and optical confinement, because a higher bandgap semiconductor also exhibits a higher refractive index. The enhanced confinement improved notably the operational characteristics of laser diodes, in particular the threshold current density, which decreased by two orders of magnitude.

But another revolution was about to come when it was realized that the confinement of electrons in lower dimensional semiconductor structures translated into completely new optoelectronic properties, when compared to bulk semiconductors. And how small should this confinement be? In order to answer this question, let us recall the concept of the de Broglie wavelength of thermalized electrons, λ_B:

$$\lambda_B = \frac{h}{p} = \frac{h}{\sqrt{2m^* E}}$$

where

h is the Planck's constant
p is the electron momentum
m^* is the electron effective mass
E is the energy

In the case of III–V compound semiconductors, λ_B is typically of the order of tens of nanometers (Saleh and Teich, 1991). If one of the dimensions of a semiconductor is comparable or less than λ_B, the electrons will be strongly confined in one dimension, while moving freely in the remaining two dimensions—this is the case of a quantum well (QW). A quantum wire is a one-dimensional confined structure, while a QD is confined in all the three dimensions. QDs are thus tiny clusters of semiconductor material with dimensions of only a few nanometers, surrounded by a semiconductor matrix that has a higher bandgap.

The spatial confinement of the carriers in lower dimensional semiconductors leads to dramatically different energy–momentum relations in the directions of confinement, which results in completely new density of states, when compared to the bulk case, as depicted in Figure 41.5. As dimensionality decreases, the density of states is no longer continuous or quasi-continuous but becomes quantized. In the case of QDs, the charge carriers occupy only a restricted set of energy levels rather like the electrons in an atom, and for this reason, QDs are sometimes referred to as "artificial atoms."

For a given energy range, the number of carriers necessary to fill out these states reduces substantially as the dimensionality decreases, which implies that it becomes easier to achieve transparency and inversion of population—with the resulting reduction of threshold current density. In fact, this reduction has been quite spectacular over the years, with sudden jumps whenever the dimensionality is decreased (Alferov, 2001).

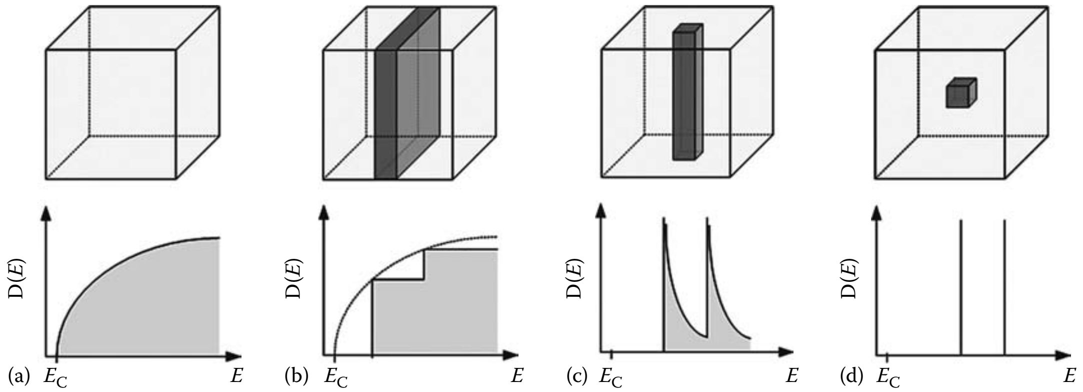

FIGURE 41.5 Schematic structures of bulk and low-dimensional semiconductors and corresponding density of states. The density of states in different confinement configurations: (a) bulk; (b) quantum well; (c) quantum wire; and (d) quantum dot.

41.3.2 Quantum Dots: Materials and Growth

The group of QD materials that has shown particular promise is based on III–V QDs epitaxially grown on a semiconductor substrate. For instance, InGaAs/InAs QDs on a GaAs substrate emit in the 1–1.3 μm wavelength range, which could be extended to 1.55 μm. Alternatively, InGaAs/InAs QDs can be grown on an InP substrate that covers emission in the 1.4–1.9 μm wavelength range (Ustinov et al., 2003).

The remarkable achievements in QD epitaxial growth have enabled the fabrication of QD lasers, amplifiers, and saturable absorbers offering excellent performance characteristics. To date, the most promising results have been achieved using the spontaneous formation of three-dimensional islands during strained layer epitaxial growth in a process known as the Stranski–Krastanow mechanism (Goldstein et al., 1985; Ustinov et al., 2003). In this process, when a film A is epitaxially grown over a substrate B, the initial growth occurs layer by layer, but beyond a certain critical thickness, three-dimensional islands start to form—the quantum dots. A continuous film lies underneath the dots, and is called the wetting layer. The most important condition in this technique is that the lattice constant of the deposited material is larger than the one of the substrate. This is the case of an InAs film (lattice constant of 6.06 Å) on a GaAs substrate (lattice constant of 5.64 Å), for example.

In spite of being an extremely complex process, the Stranski–Krastanow mode is now widely used in the self-assembly of QDs. An advantage of this technique is that films can be grown using the well-known techniques of molecular beam epitaxy (MBE) and metal organic chemical vapor deposition (MOCVD), and therefore the science of QDs growth has benefited immensely from all the previous knowledge gained with this technology. These are also good news for commercialization, because manufacturers do not have to invest in new epitaxy equipment to fabricate these structures.

Due to the statistical fluctuations occurring during growth, there is a distribution in dot size, height, and composition but, at the moment, epitaxy techniques have evolved to such an extent that the amount of fluctuations can be reasonably controlled, and can be as small as a few percent.

If the dots are grown on a plane surface, their lateral positions will be random. An example of such structure is shown in Figure 41.6. In the self-assembly process, there is no standard way of arranging the dots in a planar ordered way, unless they are encouraged to grow at particular positions in a pre-patterned substrate.

At present, the densities of QDs lie typically between 10^9 cm^{-2} and 10^{11} cm^{-2}. The sparse distribution of QDs results in a low value of optical gain. Thus, the levels of gain and optical confinement provided by a single layer of QDs may not be enough for the optimal performance of a laser. In order to circumvent this problem, QDs can also be grown in stacks, which allows an increase in the modal gain without increasing the internal optical mode loss (Smowton et al., 2001), where the various layers are usually separated by GaAs barriers. The GaAs separators are responsible for transmitting the tensile strain from layer to layer, inducing the formation of ordered arrays of QDs aligned on top of each other. Further optical confinement is enabled through the cladding of such arrays within layers of higher refractive index and bandgap energy, therefore forming a heterostructure.

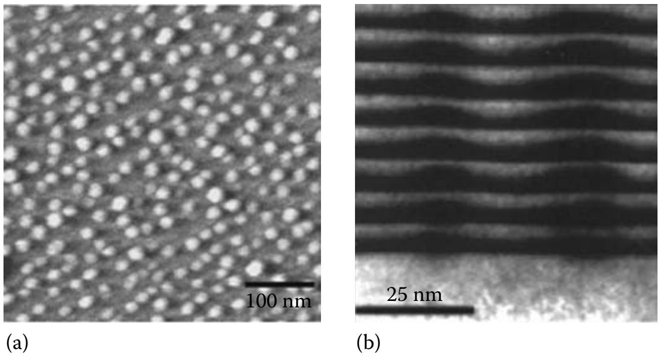

FIGURE 41.6 Photographs of an InGaAs quantum dots grown on GaAs substrate: (a) A TEM image of a single sheet of quantum dots. (b) A TEM image of a cross section of an 8-layer thick stack of quantum dots in GaAs layers.

41.3.3 Broad Gain Bandwidth

A QD laser was proposed in 1976 (Dingle and Henry, 1976) and the first theoretical treatment was published in 1982 (Arakawa and Sakaki, 1982). The main motivation was to conceive a design for a low-threshold, single-frequency, and temperature-insensitive laser, owing to the discrete nature of the density of states. In fact, practical devices exhibit the predicted outstandingly low thresholds (Kovsh et al., 2004; Liu et al., 2005), but the spectral bandwidths of such lasers are significantly broader than those of conventional QW lasers (Rafailov et al., 2007). This results from the self-organized growth of QDs, leading to a Gaussian distribution of dot sizes, with a corresponding Gaussian distribution of emission frequencies. Additionally, lattice strain may vary across the wafer, thus further affecting the energy levels in the quantum dots. These effects lead to the inhomogeneous broadening of the gain—a useful phenomenon in the context of ultrafast applications, because a very wide bandwidth is available for the generation, propagation, and amplification of ultrashort pulses. The effects of inhomogeneous broadening on the density of states are schematically illustrated in Figure 41.7. However, it is important to stress that a highly inhomogeneously broadened gain also encompasses a number of disadvantages, because it partially defeats the purpose of a reduced dimensionality, by broadening the density of states. Indeed, the fluctuation in the size of the QDs has the effect of increasing the transparency current and reducing the modal and differential gain (Qasaimeh, 2003; Dery and Eisenstein, 2005). Therefore, much effort has been put into improving the dots uniformization by engineering the growth and post-growth processes (Ustinov et al., 2003).

The extremely broad bandwidth available in QD mode-locked lasers offers potential for generating sub-100 fs pulses provided all of the bandwidth can be engaged coherently and dispersion effects suitably minimized.

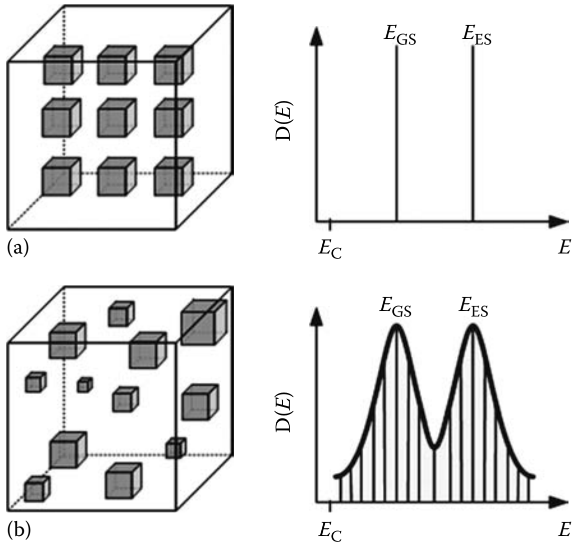

FIGURE 41.7 Schematic morphology and density of states for charge carriers in (a) an ideal quantum-dot system and (b) a real quantum-dot system, where inhomogeneous broadening is illustrated.

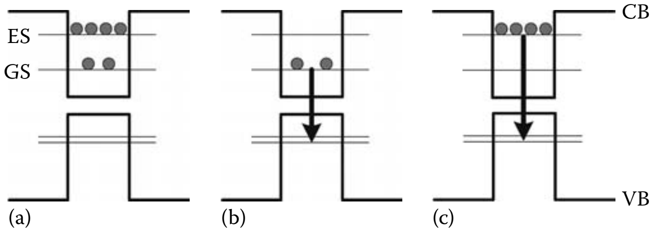

FIGURE 41.8 Schematic of the energy levels in a QD material (a), and radiative transitions via GS—ground state (b) and ES—excited state (c). CB—Conduction band; VB—valence band.

Indeed, it has been shown that there is usually some gain in narrowing/filtering effects in mode-locked QW lasers (Delfyett et al., 1998). With the inhomogeneously broadened gain bandwidth exhibited by QDs, there is support for more bandwidth and this can oppose the effect of pulse broadening that may arise from spectral narrowing. Additionally, due to the particular nature of QD lasers, many possibilities open up in respect of the exploitation of ground-state (GS) and excited-state (ES) bands,* as schematically represented in Figure 41.8. Such versatility has been successfully exploited in a multiple-wavelength-band switchable mode locking (Cataluna et al., 2006c). On the other hand, the interplay between GS and ES can be deployed in novel mode-locking regimes (Cataluna et al., 2006a). Using an external cavity, it is possible to set up tunable mode-locked sources that can operate in the wavelength range that extends from the GS to the ES transition bands (Kim et al., 2006a).

41.3.4 Ultrafast Carrier Dynamics

In the initial studies of QD materials, it was thought that their carrier dynamics would be significantly slower than those in QW materials due to a phonon bottleneck effect (Mukai et al., 1996). Interestingly, experiments have demonstrated quite the opposite. As a consequence of access to a number of recombination paths for the carriers, QD structures exhibit ultrafast recovery both under absorption and gain conditions (Borri et al., 2006). In two evaluations, the absorber dynamics of surface and waveguided QD structures were investigated by using a pump-probe technique (Borri et al., 2000; Rafailov et al., 2004b). This showed the existence of at least two distinct time constants for the recovery of the absorption. A fast recovery of around 1 ps is followed by a slower recovery process that extends over 100 ps (Rafailov et al., 2004b).

More recently, sub-picosecond carrier recovery was measured directly in a QD absorption modulator when a reverse bias was applied (Malins et al., 2006). Absorption recovery times ranged from 62 ps down to 700 fs and showed a decrease by nearly two orders of magnitude when the reverse bias applied to the structure was changed from 0 V to −10 V. This important observation provides significant promise for ultrafast modulators that can operate above 1 THz and for the optimization of saturable absorbers used for the passive mode locking of semiconductor lasers at high repetition rates, where the absorption recovery should occur within the round-trip time of the cavity. Crucially, the shaping mechanism of the fast absorption recovery also enhances the shortening of the mode-locked pulses, and thus QD lasers have the potential for generating shorter pulses than their QW counterparts.

---

* Ground and excited states are also available in quantum wells. However, the δ-like density of states associated with quantum dots enables an easier access to the ES, owing to the faster saturation of the GS in quantum dots.

41.3.5 Low Absorption Saturation Fluence

QD-based saturable absorbers exhibit lower saturation fluence than QW-based materials due to their delta-like density of states. For example, in a QD one electron is enough to achieve transparency and two to achieve inversion. This characteristic facilitates the self-starting of mode locking at modest pulse energies. This feature is particularly important in high-repetition-rate lasers where the optical energy available in each pulse is small. Indeed, it has also been observed that the saturation power is at least 2–5 times smaller for a QD saturable absorber than for a QW-based counterpart when integrated in a monolithic mode-locked laser (Thompson et al., 2004a). In this paper, the authors pointed out that saturation would further depend on the density of dots, reverse bias, and inhomogeneous broadening.

41.3.6 Low Threshold Current and Low Temperature Sensitivity

As devices, QD diode lasers have the advantage of requiring a very low threshold current to initiate lasing (Ustinov et al., 2003). This attribute applies also to operation in the mode-locking regime, because most QD lasers exhibit mode-locked operation right from the threshold of laser emission. (Bistability between the non-lasing state and the onset of lasing/mode locking might be present, as has been shown experimentally (Huang et al., 2001; Thompson et al., 2006b) and numerically (Viktorov et al., 2006).) A low threshold current is clearly advantageous because this can represent a device that is compatible as an efficient and compact source of ultrashort pulses where the demand for electrical power can be very low. Furthermore, having a low threshold avoids the need for higher carrier densities for pumping the laser and this implies less amplified spontaneous emission and reduced optical noise in the generated pulse sequences.

Due to the discrete nature of their density of states, QD lasers also exhibit low-temperature sensitivity (Mikhrin et al., 2005), making them excellent candidates for applications where resilience to temperature effects is important.

41.3.7 Low Linewidth Enhancement Factor

One of the main motivations for the enthusiastic investigation of QD materials in the last few years has been the theoretically predicted potential for very low values of LEF, owing to the symmetry of the gain associated with QD structures. The possibility of a low LEF is very attractive for a number of performance aspects, such as lower frequency chirp in directly modulated lasers, lower sensitivity to optical feedback effects, and suppressed beam filamentation. The potential of a lower effect of SPM in QD lasers also held a promise for the generation of transform-limited pulses. However, disparate reports have been published in the last 3 years, with some reports of LEF values of nearly zero (Newell et al., 1999), and others with values of LEF similar (Ukhanov et al., 2004) or significantly higher than in QW structures (Dagens et al., 2005). Ultimately, the LEF is a characteristic that is highly dependent on the operation conditions of the laser, and as such, its meaning always needs to be contextualized for a set of particular conditions. This is a topic that is currently under intense investigation.

41.4 Mode-Locked Quantum-Dot Lasers: State of the Art

41.4.1 Pulse Duration

The first demonstration of a QD mode-locked laser was reported in 2001, with pulse durations of ~17 ps at 1.3 μm and repetition rate of 7.4 GHz, using passive mode locking (Huang et al., 2001). Hybrid mode locking at the same wavelength was demonstrated in 2003 by the Cambridge University group (Thompson et al., 2003); they reported an upper limit estimation of 14.2 ps for the shortest pulses measured at a repetition rate of 10 GHz. Later in 2004, the same group demonstrated Fourier-transform-limited 10 ps pulses at 18 GHz repetition rate, using passive mode locking (Thompson et al., 2004b).

In 2004, we demonstrated the generation of sub-picosecond pulses directly from a QD laser where the shortest pulse durations were measured to be 390 fs, without any form of supplementary pulse compression (Rafailov et al., 2004a, 2005). These pulses were generated by a two-section passively mode-locked QD laser and this was the first time that sub-picosecond pulses were generated directly from such a monolithic laser.

The generation of sub-picosecond pulses was reported later by several groups (Laemmlin et al., 2006; Thompson et al., 2006b). In one of these reports (in 2006), Thompson and coworkers demonstrated the generation of pulses as short as 790 fs, by using a flared waveguide configuration in a two-section QD laser (Thompson et al., 2006b). Because the beam mode size in the saturable absorber section was much smaller than that in the gain section, the ratio of saturation fluences in the absorber and gain sections was increased. This enhanced the pulse formation mechanisms and allowed for better pulse shaping and shortening.

There has also been much effort in designing QD-based mode-locked sources that could be deployed in the 1.55 μm band (Lelarge et al., 2007). Ultra-short-pulse generation has been achieved from single-section lasers based either on InAs QDs (Renaudier et al., 2005) or quantum dashes (Gosset et al., 2006) grown on an InP substrate. These authors have suggested that there are no fundamental differences between the QDs and dashes in the context of mode-locked laser sources.

41.4.2 Toward Higher Pulse Repetition Rates

To achieve higher repetition rates in mode-locked lasers, it is necessary to decrease the cavity length. This poses a significant challenge to QD lasers because of their lower gain and the operation in short cavities may shift the emission to the ES band (Markus et al., 2003). To avoid this problem, a higher number of QD layers should be deployed in the active region. Using this simple approach, the highest repetition rate directly generated from a passively mode-locked QD two-section laser was 80 GHz (Laemmlin et al., 2006), when a 15-layer structure was used. Another method to boost the repetition rate of mode-locked lasers is to use colliding pulse mode locking. This technique is similar to passive mode locking, but the saturable absorber region is placed at the precise center of the gain section. Two counter-propagating pulses from each outer gain section therefore meet in the saturable absorber region, bleaching it much more efficiently than if just one pulse was present. This process can also result in shorter and more stable pulses. Owing to the device geometry, mode locking is achieved at the second harmonic of the fundamental (round-trip) frequency, and the pulse repetition rate is doubled. A variation of colliding pulse mode locking is harmonic mode locking, where more than two pulses circulate in the cavity, the number being equal to the harmonic. Colliding-pulse mode-locking was first demonstrated for QD lasers in 2005 (Thompson et al., 2005), resulting in a modest repetition rate of 20 GHz. Harmonic mode locking has also been demonstrated with repetition rates of approximately 40, 80, 120, and 240 GHz (Rae et al., 2006).

41.4.3 Temperature Resilience of Mode-Locked QD Lasers

Due to their delta-function-like density of states, QDs offer great potential for designing temperature-resilient devices. If their high-speed performance is also proven to be resilient to temperature, QD lasers can become the next generation of sources for ultrafast optical telecoms and datacoms, because the constraint of using thermoelectric coolers can be avoided, thus decreasing cost and complexity. In this context, we have demonstrated stable passive mode-locked operation of a two-section QD laser over an extended temperature range (from 20°C to 80°C) at relatively high output average powers (Cataluna et al., 2006b).

Additionally, to meet the requirements for high-speed communications, it is important to investigate the temperature dependence of the pulse duration. For instance, in communication systems with transmission rates of 40 Gb/s or more, the temporal interval between pulses is less than 25 ps and so the duration of the optical pulses should be well below this value at any operating temperature. We have shown that the pulse duration and the spectral width decrease significantly as the temperature is increased up to 70°C (Cataluna et al., 2007). The combination of all these effects resulted in a sevenfold decrease of the time-bandwidth product (the pulses were still highly chirped due to the strong SPM and dispersion effects in the semiconductor material).

To account for the decrease in pulse duration with temperature, a model for mode locking in QD lasers was used. It was found that the pulse durations are determined principally by the escape rate of the carriers in the absorber section, which lead to a decrease of absorber recovery time with increasing temperature, thus inducing a decrease in the pulse durations. This has been verified recently using ultrafast spectroscopy to probe the absorber recovery time as a function of temperature (Malins et al., 2007).

41.4.4 Mode Locking Involving Excited-State Transitions

It has been observed that laser emission in QD lasers can access the transitions in GS, ES or both (Markus et al., 2003), as represented in Figure 41.8. Furthermore, sub-picosecond gain recovery has been demonstrated for both GS and ES transitions in electrically pumped QD amplifiers (Schneider et al., 2005). In this reported work, the LEF was shown to decrease significantly for wavelengths below the GS transition, even becoming negative at ES thereby implying a potential for chirp-free operation for the range of wavelengths involved. Laser emission in the ES is also characterized by a higher differential gain than GS, with associated benefits for ultrafast QD lasers. We have demonstrated an optical gain-switched QD laser, where pulses were generated from both GS and ES, and where the ES pulses were shorter than those generated by GS alone (Rafailov et al., 2006). The potential for shorter and chirp-free pulses from ES transitions motivated us to investigate the mode-locked operation of QD lasers in this band. We demonstrated, for the first time, passive mode locking via GS (1260 nm) or ES (1190 nm) in a QD laser, at repetition frequencies of 21 and 20.5 GHz, respectively (Cataluna et al., 2006c). The switch between these two states in the mode-locking regime was easily achieved by changing the electrical biasing conditions, thus providing full control of the operating spectral band. It is important to stress that the average power in both operating modes was relatively high and exceeded 25 mW. In the range of bias conditions explored in this study, the shortest pulse duration measured for ES transitions was ~7 ps, where the spectral bandwidth was 5.5 nm, at an output power of 23 mW. These pulse durations are similar to those generated by GS mode locking at the same power level.

Although pulse durations from ES spectral band have been below expectations so far, it is our opinion that exploitation of the ES transitions—a unique feature of QD lasers—can lead to a new generation of high-speed sources, where mode locking involves electrically switchable GS or ES transitions that are spectrally distinct. This could enable a range of applications extending from time-domain spectroscopy, through to optical interconnects, wavelength-division multiplexing, and ultrafast optical processing.

41.5 Summary and Outlook

41.5.1 Critical Discussion

In this chapter, we have presented the physics and reported on the progress of ultrafast laser diodes based on QD materials.

The results presented in the literature show that monolithic passively mode-locked QD lasers can currently surpass the performance of similar QW lasers in terms of pulse duration (Rafailov et al., 2005, Thompson et al., 2006b). There are other particular features where QD lasers have already been shown to have a superior performance, notably in the case of pulse timing jitter where record low values have been reported (Choi et al., 2006, Thompson et al., 2006a).

We strongly believe that the appeal of QD lasers also resides in the novel functionalities that are distinctive of QDs. These are the exploitation of an ES level as a means to achieve novel mode-locking regimes; the temperature resilience offered by the quantized density of states; lower threshold and higher output power levels; and access to the enlarged spectral bandwidths associated with the inhomogeneously broadened gain features. These characteristics are not only useful from an operational point of view, but also provide some insights into a more comprehensive understanding of the underlying physical mechanisms of mode locking in QD lasers.

41.5.2 Future Perspectives

Although there have been many advances in the control of the growth of QD laser having ultra-low threshold current and temperature resilience, it is not yet understood what is the most advantageous QD structure layout to be used in the regime of mode locking. In particular, it is not clear what is the optimum level of inhomogeneous broadening that results in shorter and higher peak power pulses. Therefore, it is relevant to investigate if and how the inhomogeneously broadened spectral modes are engaged coherently in the generation of ultrashort pulses and how that effect could be used to improve the performance of the lasers toward sub-picosecond pulse durations. Exploiting novel QD materials based on p-doped and tunnel injection structures could also bring advantages in minimizing the effect of any deleterious SPM effects in mode-locked lasers.

Comparison between theory and experiment of undoped and p-doped lasers has shown how this technique can improve the LEF (Kim and Chuang, 2006b). By tuning the level of doping, lasers can exhibit zero and even negative LEF at low current densities (Alexander et al., 2007). Tunnel injection QD structures can also be of great interest for use in mode-locked lasers, as the injection of cold carriers may bring many benefits to the operation of mode-locked lasers (Delfyett, 2006). The LEF has been recently calculated and has been demonstrated to be much less than that reported for other lasers (Mi and Bhattacharya, 2007). Selective excitation of population in these lasers has been demonstrated, which could lead to a mitigation of the inhomogeneous broadening effects and contribute to a narrower spectrum and the production of transform-limited pulses (Bret and Gires, 1964).

The ES spectral band can also be exploited in tunable lasers using QD materials where the inhomogeneous broadening is controlled so as to maximize the overlap between GSs and ESs. While in cw operation, QD lasers have been demonstrated with tunability ranges up to 200 nm (Varangis, 2000), by exploiting the gain available from the GSs and ESs. Combining this tunability with the possibility of generating ultrashort pulses, it will be possible to achieve a new generation of versatile lasers emitting pulses across a wide range of wavelengths—as if we had compressed many different lasers into a single laser! Such disruptive characteristics will offer endless possibilities and enable applications never seen before in science and technology.

Acknowledgments

We wish to thank Dr. D. Livshits, Dr. I. Krestnikov, and Dr. A. Kovsh from Innolume GmbH (Dortmund) for helping to prepare samples and for stimulating discussions.

This work was supported in part by the European Community's Seventh Framework Programme FAST-DOT under grant agreement 224338.

References

Alexander, R. R., Childs, D., Agarwal, H. et al. 2007. Zero and controllable linewidth enhancement factor in p-doped 1.3 μm quantum dot lasers. *Japanese Journal of Applied Physics* 46: 2421–2423.

Alferov, Z. I. 2001. Nobel Lecture: The double heterostructure concept and its applications in physics, electronics, and technology. *Reviews of Modern Physics* 73: 767.

Arakawa, Y. and Sakaki, H. 1982. Multidimensional quantum well laser and temperature dependence of its threshold current. *Applied Physics Letters* 40: 939–941.

Basov, N. G. 1964. Nobel lecture: Semiconductor lasers. http://www.nobel.se/physics/laureates/1964/basov-lecture.html.

Basov, N. G., Krokhin, O. N., and Popov, Y. M. 1961. The possibility of use of indirect transitions to obtain negative temperature in semiconductors. *Soviet Physics JETP* 12: 1033.

Borri, P., Langbein, W., Hvam, J. M., Heinrichsdorff, F., Mao, M. H., and Bimberg, D. 2000. Spectral hole-burning and carrier-heating dynamics in InGaAs quantum-dot amplifiers. *IEEE Journal of Selected Topics in Quantum Electronics* 6: 544–551.

Borri, P., Schneider, S., Langbein, W., and Bimberg, D. 2006. Ultrafast carrier dynamics in InGaAs quantum dot materials and devices. *Journal of Optics A-Pure and Applied Optics* 8: S33–S46.

Cataluna, M. A., Mcrobbie, A. D., Sibbett, W., Livshits, D. A., Kovsh, A. R., and Rafailov, E. U. 2006a. New mode locking regime in a quantum-dot laser: Enhancement by simultaneous CW excited-state emission. *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference paper CThH3*: CThH3, Long Beach, CA.

Cataluna, M. A., Rafailov, E. U., Mcrobbie, A. D., Sibbett, W., Livshits, D. A., and Kovsh, A. R. 2006b. Stable mode-locked operation up to 80°C from an InGaAs quantum-dot laser. *IEEE Photonics Technology Letters* 18: 1500–1502.

Cataluna, M. A., Sibbett, W., Livshits, D. A., Weimert, J., Kovsh, A. R., and Rafailov, E. U. 2006c. Stable mode locking via ground- or excited-state transitions in a two-section quantum-dot laser. *Applied Physics Letters* 89: 81124–3.

Cataluna, M. A., Viktorov, E. A., Mandel, P. et al. 2007. Temperature dependence of pulse duration in a mode-locked quantum-dot laser. *Applied Physics Letters* 90: 101102–3.

Choi, M.-T., Kim, J.-M., Lee, W., and Delfyett, P. J. 2006. Ultralow noise optical pulse generation in an actively mode-locked quantum-dot semiconductor laser. *Applied Physics Letters* 88: 131106–3.

Dagens, B., Markus, A., Chen, J. X. et al. 2005. Giant linewidth enhancement factor and purely frequency modulated emission from quantum dot laser. *Electronics Letters* 41: 323–324.

Delfyett, P. J. 2006. Personal communication.

Delfyett, P. J., Shi, H., Gee, S., Barty, C. P. J., Alphonse, G., and Connolly, J. 1998. Intracavity spectral shaping in external cavity mode-locked semiconductor diode lasers. *IEEE Journal of Selected Topics in Quantum Electronics* 4: 216–223.

Dery, H. and Eisenstein, G. 2005. The impact of energy band diagram and inhomogeneous broadening on the optical differential gain in nanostructure lasers. *IEEE Journal of Quantum Electronics* 41: 26–35.

Dingle, R. and Henry, C. H. 1976. Quantum effects in heterostructure lasers. U.S. Patent.

Goldstein, L., Glas, F., Marzin, J. Y., Charasse, M. N., and Roux, G. L. 1985. Growth by molecular beam epitaxy and characterization of InAs/GaAs strained-layer superlattices. *Applied Physics Letters* 47: 1099–1101.

Gosset, C., Merghem, K., Martinez, A. et al. 2006. Subpicosecond pulse generation at 134 GHz using a quantum-dash-based Fabry-Perot laser emitting at 1.56 μm. *Applied Physics Letters* 88: 241105–3.

Huang, X. D., Stintz, A., Li, H., Lester, L. F., Cheng, J., and Malloy, K. J. 2001. Passive mode-locking in 1.3 μm two-section InAs quantum dot lasers. *Applied Physics Letters* 78: 2825–2827.

Kim, J., Choi, M. T., Lee, W., and Delfyett, P. J. 2006a. Wavelength tunable mode-locked quantum-dot laser. *Enabling Photonics Technologies for Defense, Security, and Aerospace Applications II*: 6243: M2430.

Kim, J. and Chuang, S. L. 2006b. Theoretical and experimental study of optical gain, refractive index change, and linewidth enhancement factor of p-doped quantum-dot lasers. *IEEE Journal of Quantum Electronics* 42: 942–952.

Kovsh, A. R., Ledentsov, N. N., Mikhrin, S. S. et al. 2004. Long-wavelength (1.3–1.5 μm) quantum dot lasers based on GaAs. *Physics and Simulation of Optoelectronic Devices XII*: 5349: 31–45.

Laemmlin, M., Fiol, G., Meuer, C. et al. 2006. Distortion-free optical amplification of 20–80 GHz modelocked laser pulses at 1.3 μm using quantum dots. *Electronics Letters* 42: 697–699.

Lelarge, F., Dagens, B., Renaudier, J. et al. 2007. Recent advances on InAs/InP quantum dash based, semiconductor lasers and optical amplifiers operating at 1.55 mu m. *IEEE Journal of Selected Topics in Quantum Electronics* 13: 111–124.

Liu, H. Y., Childs, D. T., Badcock, T. J. et al. 2005. High-performance three-layer 1.3-μm InAs-GaAs quantum-dot lasers with very low continuous-wave room-temperature threshold currents. *IEEE Photonics Technology Letters* 17: 1139–1141.

Malins, D. B., Gomez-Iglesias, A., White, S. J., Sibbett, W., Miller, A., and Rafailov, E. U. 2006. Ultrafast electroabsorption dynamics in an InAs quantum dot saturable absorber at 1.3 μm. *Applied Physics Letters* 89: 171111–3.

Malins, D. B., Gomez-Iglesias, A., Cataluna, M. A., Rafailov, E. U., Sibbett, W., and Miller, A. 2007. Temperature dependence of electroabsorption dynamics in an InAs quantum dot saturable absorber at 1.3 μm. *CLEO-Europe'07*: CF-6, Munich, Germany.

Markus, A., Chen, J. X., Paranthoen, C., Fiore, A., Platz, C., and Gauthier-Lafaye, O. 2003. Simultaneous two-state lasing in quantum-dot lasers. *Applied Physics Letters* 82: 1818–1820.

Mi, Z. and Bhattacharya, P. 2007. Analysis of the linewidth-enhancement factor of long-wavelength tunnel-injection quantum-dot lasers. *IEEE Journal of Quantum Electronics* 43: 363–369.

Mikhrin, S. S., Kovsh, A. R., Krestnikov, I. L. et al. 2005. High power temperature-insensitive 1.3 μm InAs/InGaAs/GaAs quantum dot lasers. *Semiconductor Science and Technology* 20: 340–342.

Mukai, K., Ohtsuka, N., Shoji, H., and Sugawara, M. 1996. Phonon bottleneck in self-formed $In_xGa_{1-x}As/GaAs$ quantum dots by electroluminescence and time-resolved photoluminescence. *Physical Review B* 54: R5243–R5246.

Newell, T. C., Bossert, D. J., Stintz, A., Fuchs, B., Malloy, K. J., and Lester, L. F. 1999. Gain and linewidth enhancement factor in InAs quantum-dot laser diodes. *IEEE Photonics Technology Letters* 11: 1527–1529.

Qasaimeh, O. 2003. Effect of inhomogeneous line broadening on gain and differential gain of quantum dot lasers. *IEEE Transactions on Electron Devices* 50: 1575–1581.

Rae, A. R., Thompson, M. G., Penty, R. V. et al. 2006. Harmonic mode-locking of a quantum-dot laser diode. *LEOS* 2006: ThR5, Montreal, QC, Canada.

Rafailov, E. U., Cataluna, M. A., Sibbett, W. et al. 2004a. High-power ultrashort pulses output from a mode-locked two-section quantum-dot laser. *Conference on Lasers and Electro-Optics/International Quantum Electronics Conference*: CPDB5, post-deadline, San Francisco, CA.

Rafailov, E. U., White, S. J., Lagatsky, A. A. et al. 2004b. Fast quantum-dot saturable absorber for passive mode-locking of solid-state lasers. *IEEE Photonics Technology Letters* 16: 2439–2441.

Rafailov, E. U., Cataluna, M. A., Sibbett, W. et al. 2005. High-power picosecond and femtosecond pulse generation from a two-section mode-locked quantum-dot laser. *Applied Physics Letters* 87: 81107–3.

Rafailov, E. U., Mcrobbie, A. D., Cataluna, M. A., O'faolain, L., Sibbett, W., and Livshits, D. A. 2006. Investigation of transition dynamics in a quantum-dot laser optically pumped by femtosecond pulses. *Applied Physics Letters* 88: 41101–3.

Rafailov, E. U., Cataluna, M. A., and Sibbett, W. 2007. Mode-locked quantum-dot lasers. *Nature Photonics* 1: 395–401.

Renaudier, J., Brenot, R., Dagens, B. et al. 2005. 45 GHz self-pulsation with narrow linewidth in quantum dot Fabry-Perot semiconductor lasers at 1.5 μm. *Electronics Letters* 41: 1007–1008.

Saleh, B. E. A. and Teich, M. C. 1991. *Fundamentals of Photonics*. New York: Wiley.

Schneider, S., Borri, P., Langbein, W. et al. 2005. Excited-state gain dynamics in InGaAs quantum-dot amplifiers. *IEEE Photonics Technology Letters* 17: 2014–2016.

Smowton, P. M., Herrmann, E., Ning, Y., Summers, H. D., Blood, P., and Hopkinson, M. 2001. Optical mode loss and gain of multiple-layer quantum-dot lasers. *Applied Physics Letters* 78: 2629–2631.

Thompson, M. G., Marinelli, C., Tan, K. T. et al. 2003. 10 GHz hybrid modelocking of monolithic InGaAs quantum dot lasers. *Electronics Letters* 39: 1121–1122.

Thompson, M. G., Marinelli, C., Chu, Y. et al. 2004a. Properties of InGaAs quantum dot saturable absorbers in monolithic mode-locked lasers. *IEEE* 19th *International Semiconductor Laser Conference, Conference Digest*, Matsue-shi, Japan.

Thompson, M. G., Tan, K. T., Marinelli, C. et al. 2004b. Transform-limited optical pulses from 18 GHz monolithic modelocked quantum dot lasers operating at 1.3 μm. *Electronics Letters* 40: 346–347.

Thompson, M. G., Marinelli, C., Zhao, X. et al. 2005. Colliding-pulse modelocked quantum dot lasers. *Electronics Letters* 41: 248–250.

Thompson, M. G., Larson, D., Rae, A. et al. 2006a. Monolithic hybrid and passive mode-locked 40 GHz quantum dot laser diodes. 32nd *European Conference on Optical Communication*: We4.6.3, Cannes, France.

Thompson, M. G., Rae, A., Sellin, R. L. et al. 2006b. Subpicosecond high-power mode locking using flared waveguide monolithic quantum-dot lasers. *Applied Physics Letters* 88: 133119–3.

Ukhanov, A. A., Stintz, A., Eliseev, P. G., and Malloy, K. J. 2004. Comparison of the carrier induced refractive index, gain, and linewidth enhancement factor in quantum dot and quantum well lasers. *Applied Physics Letters* 84: 1058–1060.

Ustinov, V. M., Zhukov, A. E., Egorov, A. Y., and Maleev, N. A. 2003. *Quantum Dot Lasers*. New York: Oxford University Press.

Varangis, P. M., Li, H., Liu, G. T., Newell, T. C., Stintz, A., Fuchs, B., Malloy, K. J., and Lester, L. F. 2000. Low-threshold quantum dot lasers with 201 nm tuning range. *Electronics Letters* 36: 1544–1545.

Vasil'ev, P. 1995. *Ultrafast Diode Lasers: Fundamentals and Applications*. Boston, MA: Artech House.

Viktorov, E. A., Mandel, P., Vladimirov, A. G., and Bandelow, U. 2006. Model for mode locking in quantum dot lasers. *Applied Physics Letters* 88: 201102–3.

Zhang, L., Cheng, L., Gray, A. L. et al. 2005. Low timing jitter, 5 GHz optical pulses from monolithic two-section passively mode-locked 1250/1310 nm quantum dot lasers for high-speed optical interconnects. *Optical Fiber Communication Conference, OFC/NFOEC* OWM4, Anaheim, CA.

Index

A

- AFM, *see* Atomic force microscopy
- Aharonov–Bohm effect, 11-1
- Aluminum and gap engineering, 16-9 thru 16-10
- Ambegaokar–Barato relation, 16-2
- AND and XOR logic gates, 33-10 thru 33-11
- Andreev bound states, 7-3 thru 7-4
- Andreev reflection, 7-2
- Atomic force microscopy (AFM), 6-11 thru 6-12, 35-10
- Atomic-scale technology computer-aided design (TCAD), 12-11 thru 12-12
- Atomistic models, 40-5

B

- Bacteriorhodopsin (BR)
  - absorption spectra, 38-1 thru 38-2
  - basic process
    - activation function realization, 38-5
    - formal neuron main functions realization, 38-3
    - medium bleaching, 38-2 thru 38-3
    - neuro-like element formation, 38-3 thru 38-4
    - optical radiation fluxes, indirect interaction, 38-3
    - weighed signals composition function realization, 38-4 thru 38-5
    - weighing function realization, 38-4
  - nanostructure
    - complex estimation procedure, 38-10 thru 38-11
    - functional characteristics, 38-12 thru 38-14
    - hybrid nanostructures, 38-13 thru 38-17
  - optical neuronets
    - basic elements, multilayered structures, 38-7 thru 38-8
    - cyclicity processes, 38-5 thru 38-6
    - data processing, 38-6
    - element base parameters, 38-7
    - light fluxes formation, 38-6
    - multilayered constructions, 38-7
    - polymeric film property, 38-9 thru 38-10
    - polymeric film requirements, 38-8
    - polymeric mixture preparation, 38-8 thru 38-9
    - suspension preparation, 38-8
  - photocycle, 38-1 thru 38-2
- Berry's phase interference
  - definition, 9-5
  - electron in constant magnetic field, 9-6
  - path integrals, 9-6 thru 9-7
  - in single-molecule magnet transistors, 9-11 thru 9-12
- Biomaterial shaping, 18-6 thru 18-7
- Biomimetics, photonic nanostructures
  - antireflector engineering, 31-1, 31-3
  - cell culture, 31-3, 31-5
  - diatoms and coccolithophores
    - frustule, 31-5
    - *Nitzschia frustulum*, 31-6 thru 31-7
    - photochemical-etching technique, 31-5
    - photoluminescence (PL) emission, 31-5 thru 31-6
    - photonic device, 31-5 thru 31-6
    - porous glass, 31-7
  - eukaryote cell, 31-8
  - iridescent device engineering
    - *Aphrodita* sea mouse, 31-3
    - 2D christmas tree structure, 31-3 thru 31-4
    - metallic colored effect, 31-1 thru 31-2
    - *Morpho* butterfly scale, 31-2 thru 31-4
    - optical characteristics, 31-2
    - photonic crystal, 31-2
  - iridoviruses, 31-7
  - optical reflector types, 31-1 thru 31-2
  - silica deposition vesicle (SDV), 31-7
  - trans-Golgi-derived vesicle, 31-7
- Biomolecular neuronet devices, *see* Bacteriorhodopsin
- Bose–Einstein distribution function, 10-8
- Bowtie plasmonic nanolaser, 39-9
- Bragg mirrors, 39-3
- Bragg MOKE technique, 4-8 thru 4-9
- Break junction experiment, 12-9
- Bucky shuttle memory element, 3-16 thru 3-17
- Buttiker model, 10-5 thru 10-6

C

- Cantor's middle-excluded set, 17-13 thru 17-14
- Capacitance spectroscopy
  - capacitance transient spectroscopy (*see* Deep level transient spectroscopy)
  - depletion region
    - depletion capacitance, 2-13 thru 2-14
    - p–n junction, 2-12
    - Schottky contact, 2-11 thru 2-12
    - width of, 2-13
- Carbon nanotube (CNT) memory elements
  - CNFET-based memory
    - band structure, 3-3
    - charge-storage stability, 3-5 thru 3-6
    - controlling storage nodes, 3-6 thru 3-7
    - Fuhrer's device, 3-4
    - nanotube storage nodes, 3-9
    - optoelectronic memory, 3-7
    - Radosavljevic's device, 3-5
    - redox active molecules, 3-7
    - schematic cross section, 3-2
    - two-terminal memory devices, 3-7 thru 3-9
  - electromigration CNT, 3-16 thru 3-17
  - NEMS-based memory
    - capacitive force response, 3-9
    - feedback-controlled nanocantilevers, 3-12 thru 3-13
    - linear bearing nanoswitch, 3-14 thru 3-16
    - nanorelays, 3-11 thru 3-12
    - nonvolatile random access memory, 3-10 thru 3-11
    - properties, 3-9
    - vertically aligned carbon nanotubes, 3-13 thru 3-14
- Carbon nanotube transistors, 12-6 thru 12-7
- Carbon nanotube weak link
  - experimental realizations, 7-6 thru 7-7
  - Fabry–Perot cavity, 7-6
  - high-frequency irradiation, 7-9 thru 7-10
  - nanotube quantum dot
    - CNT superconducting transistor, 7-7 thru 7-8
    - gate dependence, 7-8 thru 7-9

- Carrier–carrier interaction
  - capture and relaxation time, 40-7
  - Coulomb interaction matrix element, 40-7
  - Markov approximation, 40-6
  - quantum-kinetic equation, 40-7
  - screening effects, 40-7 thru 40-8
- Carrier manipulation (up-converter), 33-7 thru 33-8
- Carrier–phonon interaction
  - perturbation theory *vs*. polaron picture, 40-8 thru 40-9
  - polaron states and kinetics, 40-9 thru 40-10
- Cathodoluminescence
  - band structure, types, 21-2
  - donor–acceptor pair transition, 21-2
  - exciton, 21-3
  - recombination process, 21-2
  - TEM-CL (*see* Transmission electron microscope-cathodoluminescence)
- C_{60} field effect transistors
  - Bell Laboratory misconduct, 15-2 thru 15-3
  - characterization on SiO_2/AlN
    - I–V characteristics, 15-19
    - sub-threshold region, 15-19 thru 15-20
  - fabrication of
    - C_{60} epitaxy, 15-18 thru 15-19
    - device process, 15-18
    - performance measurements, 15-19
    - SiO_2/AlN substrates, 15-18
  - field effect doping, 15-3
  - operation principles, 15-3 thru 15-4
  - solid C_{60} and insulator films interface, 15-3
  - superconductivity, 15-4
- Charge-selective deep level transient spectroscopy, 2-16
- Charge trap memories, 2-6 thru 2-7
- Chemically amplified resist (CAR), 20-12
- Classical Faraday effect, 1-13
- CNT field-effect transistor (CNTFET), 12-7
  - band structure, 3-3
  - charge-storage stability, 3-5 thru 3-6
  - Fuhrer's device, 3-4
  - optoelectronic memory, 3-7
  - Radosavljevic's device, 3-5
  - schematic cross section, 3-2
  - storage nodes
    - control, 3-6 thru 3-7
    - nanotube storage nodes, 3-9
    - redox active molecules, 3-7
  - two-terminal memory devices, 3-7 thru 3-9
- Coherent electron transport
  - electrochemical potential expression, 10-3
  - self-energy parts expression, 10-3
  - through a carbon chain, 10-4
  - via HOMO, 10-2
  - via LUMO, 10-2
- Coherent Néel vector tunneling
  - doped antiferromagnetic molecular wheel, 9-10
  - satellite resonance, 9-10
  - two degenerate classical ground state, 9-9
- Computational micromagnetics
  - boundary element (BE) method, 8-7
  - finite difference (FD) method
    - Ampère field, 8-10
    - anisotropy field computation, 8-10
    - exchange field computation, 8-9
    - exchange length, 8-8
    - Langevin equation, 8-11
    - magnetostatic field computation, 8-10
    - mesh for, 8-8
    - spatial distribution field, 8-11
    - thermal field computation, 8-12
    - *vs*. FE method, 8-8
  - finite element (FE) method, 8-7 thru 8-8
  - pillar *vs*. point-contact, 8-12 thru 8-15
  - spin-transfer torque and giant magnetoresistance, 8-12 thru 8-13
- Confined metallic systems
  - chemical structure influence, 24-10 thru 24-11
  - cluster shape influence, 24-8 thru 24-9
  - dielectric function, 24-6
  - dipolar approximation, 24-10 thru 24-12
  - quantum size effects, 24-6 thru 24-8
  - quasi-static approximation/dipolar, 24-5 thru 24-6
  - scattering, absorption, and extinction, 24-4 thru 24-5
- Conformal deposition and isotropic etching, 17-15 thru 17-17
- Contact arc method, 15-5
- Control ordering and assembly processes
  - crystal growth process, 18-8
  - crystallization process, 18-8
  - full confinement factor, 18-11 thru 18-14
  - partial confinement factor, 18-9 thru 18-11
  - semicrystalline polymers morphology, 18-8
- Cooper-pair transistor (CPT)
  - aluminum and gap engineering, 16-9 thru 16-10
  - band structure
    - critical current modulation calculation, 16-5 thru 16-6
    - effective inductance modulation calculation, 16-6
    - uncertainty principle at work, 16-5
  - Coulomb blockade, 16-4 thru 16-5
  - current switching electrometry
    - quasiparticle poisoning, 16-11 thru 16-12
    - ramped switching current measurement, 16-11
    - single shot measurement, 16-12 thru 16-13
    - switching current measurement, 16-10
  - electron-beam lithography and two-angle deposition, 16-9
  - quasiparticle poisoning, 16-6 thru 16-9
  - single Josephson junction
    - Ambegaokar–Barato relation, 16-2
    - concerns specific to small junctions, 16-4
    - field emission micrograph, 16-2
    - *I–V* curve, 16-2 thru 16-3
    - RCSJ model, 16-3 thru 16-4
  - zero-biased rf electrometry
    - operation beyond 1 GHz, 16-14 thru 16-15
    - quasiparticle poisoning, 16-15
    - relation to the cooper-pair box, 16-16
    - rf measurement setup, 16-13 thru 16-14
- Coulomb blockade, 16-4 thru 16-5
- Coulomb carrier–carrier interaction, 22-6
- Crossbar process, 17-2 thru 17-3
- C_{60} superconducting transistor, 7-18
- Current switching electrometry
  - quasiparticle poisoning, 16-11 thru 16-12
  - ramped switching current measurement, 16-11
  - single shot measurement, 16-12 thru 16-13
  - switching current measurement, 16-10

D

- Davydov transformation, 32-12 thru 32-13, 32-16
- Deep level transient spectroscopy (DLTS)
  - charge-selective deep level transient spectroscopy, 2-16
  - measurement principle, 2-14 thru 2-15
  - rate window and double-boxcar method, 2-15 thru 2-16
- Deep ultraviolet (DUV) lithography
  - CaF_2 lens elements, 19-11
  - depth of focus (DOF), 19-9 thru 19-10
  - double patterning, 19-13 thru 19-14
  - fused silica glass, 19-10
  - 193 nm immersion lithography, 19-11 thru 19-12
  - numerical aperture, 19-11
  - optical-projection lithography system, 19-9
  - photolithography exposure process, 19-9
  - plot of critical feature size, 19-10
  - ultimate resolution limits, 19-14
- Depletion capacitance, 2-13 thru 2-14
- Dipolar approximation
  - expansion coefficients, 24-11
  - extrinsic size effects, 24-12
  - incident, scattering, and internal fields, 24-11
  - quadrupolar mode, 24-12
  - vector Helmholtz wave equation, 24-10
- Dipole nanolaser, 39-9 thru 39-10
- Discharge-produced plasma EUV source, 20-9 thru 20-10
- Dissipated optical energy transfer
  - CdSe QDs, schematic images, 36-2 thru 36-3

- dispersed substrate area, 36-3 thru 36-4
- dynamic property of, 36-3
- experimental results, 36-4 thru 36-5
- exponential decay curve functions, 36-4
- temperature dependent micro- (PL) spectroscopy, 36-2 thru 36-3

Dissipation-controlled nanophotonic devices
- carrier manipulation (up-converter), 33-7 thru 33-8
- energy transfer between two quantum dots
  - density matrix formalism, 33-2 thru 33-3
  - energy diagram, 33-2
  - excitation creation, 33-5 thru 33-6
  - exciton dynamics, 33-3 thru 33-4
  - resonant coupling, 33-4
- nanophotonic switch
  - energy diagram, 33-6
  - exciton population, 33-6 thru 33-7
  - ON- to OFF-state, 33-7
  - switching operation, 33-6
- spatial symmetry, 33-8 thru 33-13

Dissipative electron transport
- contact time, 10-9
- electron conductance *vs.* voltage, 10-10 thru 10-11
- junction power loss, 10-11
- phonon bath, 10-9 thru 10-10
- through large DNA molecules, 10-10

Domain theory, 8-2
Donnan effect, 11-3
Double-boxcar method, 2-15 thru 2-16
Double exposure method, 19-24 thru 19-25
Double patterning, 19-13 thru 19-14
Drude expression, 26-4
Drude model, 24-3 thru 24-4
Dynamic random access memory (DRAM)
- cell array structure, 2-2
- endurance, 2-3
- planar cell layout, 2-2 thru 2-3
- *vs.* Flash memory, 2-4

Dyson equation, 40-9

E

Edge emitter laser diodes
- microcavity lasers, 40-3 thru 40-4
- optical gain, 40-2
- types, 40-2

Effective-mass approach, 40-5
Effective near-field optical interaction
- effective interaction Hamiltonian operator, 32-8
- effective sample–probe-tip interaction, 32-10
- electric displacement operator, 32-9
- electromagnetic field correlations and intermolecular interactions, 32-6 thru 32-7
- exciton polariton, 32-9
- irrelevant macroscopic subsystem, 32-7 thru 32-8
- nanomaterial system, 32-7
- photon–electron–phonon interacting system, 32-7
- P space and Q space, 32-8
- Yukawa functions, 32-11

Electromigration CNT, 3-16 thru 3-17
Electron-beam lithography
- cost of ownership, 19-17
- overlay, 19-16 thru 19-17
- resolution, 19-14 thru 19-15
- throughput, 19-15 thru 19-16

Electron–phonon interactions, 10-2, 10-7, 10-10
Electron-transfer reactions, 10-13 thru 10-15
Envelope-function approximation, 40-5
EPR entanglement, 1-16
EUV imaging system development, 20-2 thru 20-3
EUVL, *see* Extreme ultraviolet lithography
Excitation energy transfer, 35-1 thru 35-3
Exciton–phonon–polariton (EPP) model., 35-7
Extended boundary condition method (EBCM), 30-2
Extreme ultraviolet lithography (EUVL)
- alpha demo tool (ADT), 20-3
- aspheric mirrors, 20-4
- components, 20-1
- defect-free blanks, 20-14
- DPP EUV source-collector module, 20-9 thru 20-10
- EUV device integration, 20-8 thru 20-9
- 2008 EUV focus areas, 20-9
- EUV masks
  - blanks, 20-10 thru 20-11
  - defectivity, 20-11 thru 20-12
  - round-trip shipment, 20-13
  - SEMI P27-1102 specifications, 20-10
  - S-pod, 20-13
- EUV resists, 20-12
- exposure tools
  - ASML, 20-5 thru 20-6
  - CAD model, 20-6 thru 20-7
  - interference lithography, 20-7 thru 20-8
  - scanning electron microscope (SEM) image, 20-7
  - SEMATECH Berkeley MET, 20-6 thru 20-7
  - six-mirror imaging system, 20-6
- high-power pulsed laser, 20-1
- immersion lithography, 20-15
- intermediate focus, definition, 20-9
- LPP EUV source-collector module, 20-10
- multilayer (ML) reflective coating, 20-1, 20-3 thru 20-4
- normal-incidence reflective optics, 20-2
- optics and mask contamination, 20-13 thru 20-14
- phase-shifting point-diffraction interferometer (PS/PDI), 20-5
- projection optics, 20-2 thru 20-3
- Rayleigh equation, 20-14 thru 20-15

F

Fabry–Perot cavity, 7-6
Feedback-controlled nanocantilevers, 3-12 thru 3-13
Fermi Golden rule, 21-6
Ferroelectricity fundamentals
- ferroelectric correlation length, 6-4
- ferroelectric hysteresis loops, 6-4
- perovskite oxide BaTiO3 unit cell, 6-3
- phenomenological definition, 6-2
- second-order ferroelectric phase transition, 6-3
- thermodynamic theory, 6-3

Ferroelectric RAM (FeRAM), 2-4 thru 2-5
Ferromagnetic islands
- experimental techniques, 4-5
- ferromagnetic domains, 4-2 thru 4-3
- ferromagnetic dots
  - longitudinal Bragg MOKE, 4-8 thru 4-9
  - micromagnetic simulation results, 4-7 thru 4-8
  - nucleation and annihilation field, 4-7
  - permalloy dots array, 4-7
- ferromagnetic rings, 4-9
- ferromagnetic spirals, 4-10 thru 4-11
- ferromagnetism, 4-1 thru 4-2
- hysteresis loops, 4-3
- magnetostatic interactions
  - honeycomb structure, 4-13
  - onion state, 4-12
  - open frame system, 4-11
  - Py dipole arrays, 4-12 thru 4-13
  - spin ice state, 4-11
  - symmetrical stable remanent states, 4-12
- micromagnetic simulations, 4-3 thru 4-5
- noncircular rings, 4-9 thru 4-10
- rectangular and elliptical islands, 4-6 thru 4-7
- sub-micrometer-sized island preparation, 4-5

Ferromagnetic molecular magnets, 9-1
Ferromagnetic rings, 4-9
Ferromagnetic spirals, 4-10 thru 4-11
Fick's first law, 11-3 thru 11-4
Figure of merit (FOM), 35-5
Finite difference (FD) method, LLGS
- Ampère field, 8-10
- anisotropy field computation, 8-10
- exchange field computation, 8-9
- exchange length, 8-8
- Langevin equation, 8-11
- magnetostatic field computation, 8-10
- mesh for, 8-8
- spatial distribution field, 8-11
- thermal field computation, 8-12
- *vs.* FE method, 8-8

Flash memories
- electron per cell *vs.* feature size, 2-4
- schematic band structure, 2-3
- *vs.* DRAM, 2-4

Floating nano-dots, 5-2
Fluctuation-dissipation theorem, 8-11
Fokker–Planck equation, 8-11
FOM, *see* Figure of merit
Förster type interaction, 33-3
Fourier transform, 32-2
Fowler–Nordheim tunneling, 6-11
Fractal nanotechnology
  fractals in nature, 17-12 thru 17-13
  producing nanoscale fractals via S^nPT_x, 17-13 thru 17-14
Franck–Condon factor, 10-13, 10-15
Fuhrer's device, 3-4
Fullerene (C_{60})
  discovery of, 15-4 thru 15-5
  FET
    Bell Laboratory misconduct, 15-2 thru 15-3
    characterization on SiO_2/AlN, 15-19 thru 15-20
    fabrication of, 15-18 thru 15-19
    field effect doping, 15-3
    operation principles, 15-3 thru 15-4
    solid C_{60} and insulator films interface, 15-3
    superconductivity, 15-4
  MBE, 15-10 thru 15-11
  micro-clusters, 15-1
  properties of, 15-5 thru 15-7
  solid crystals, 15-1 thru 15-2
    absorption and luminescence spectra, 15-8
    crystal structure, 15-7
    electronic structures, 15-7 thru 15-8
    optical measurements results, 15-9
    photo-conducting spectra, 15-10
    photoluminescence spectrum, 15-8
    structural phase transition, 15-7
    third nonlinear susceptibility, 15-9
    van der Waals interaction, 15-7
  synthesis of, 15-5
  van der Waals epitaxy (*see* van der Waals epitaxy)
Fullerene-based superconducting weak links, 7-17 thru 7-18
Full-width at half-maximum (FWHM), 35-8
Fully depleted silicon-on-insulator (FDSOI) devices, 12-3

G

Gas evaporation method, 15-5
Giant magnetoresistance (GMR), 8-12 thru 8-13, 13-5
Graphene-based superconducting weak links
  proximity effect, 7-15 thru 7-17
  SQUIDs devices, 7-17
Graphene ribbon nanotransistors, 12-7 thru 12-8
Grover algorithm, 1-9 thru 1-12

H

Half pitch scaling, 19-4
Hamiltonian model, 32-11 thru 32-12
Heisenberg's uncertainty principle, 32-1
Hellmann–Feynman theorem, 27-5
Helmholtz equation, 32-4
Hierarchical memory retrieval, 34-7 thru 34-8
Horseshoe optical nanoantenna., 39-9
Hot-carrier diode, 13-1
Hybrid nanostructures
  absorption spectrum, 38-14 thru 38-15
  AFM characterization, gold nanoparticle, 38-14 thru 38-15
  borohydride method, 38-13
  bridge molecules, 38-15
  dialyze, 38-14
  factual spectrum, 38-17
  gold nanoparticles spectrum, 38-14
  schematic composition, 38-13
  spectral characteristics, 38-17
  water colloidal suspension, silver nanoparticles, 38-15 thru 38-16

I

III–V semiconductor quantum dots
  self-organized quantum dots
    based nanomemories, 2-10
    carrier emission process, 2-9
    electron and hole states, 2-9
    fabrication, 2-8
  semiconductor material hetereostructure, 2-7 thru 2-8
193 nm Immersion lithography
  DUV lithography, 19-11 thru 19-12
  photoresist technology, 19-8
Imprint lithography (IL), 17-3 thru 17-4
Incoherent Zener tunneling, 9-7 thru 9-8
Inelastic electron transport
  Buttiker model, 10-5 thru 10-6
  coherent transport
    curves, 10-4 thru 10-5
    electrochemical potential expression, 10-3
    self-energy parts expression, 10-3
    through a carbon chain, 10-4
    via HOMO, 10-2
    via LUMO, 10-2
  dissipative transport
    contact time, 10-9
    electron conductance *vs*. voltage, 10-10 thru 10-11
    junction power loss, 10-11
    phonon bath, 10-9 thru 10-10
    through large DNA molecules, 10-10
  molecular junction conductance and long-range reactions, 10-13 thru 10-15
  polaron effects, 10-11 thru 10-13
  vibration-induced inelastic effects
    electron transmission *vs*. energy, 10-7
    electron–vibron interaction, 10-6 thru 10-07
    inelastic electron tunneling spectrum, 10-8 thru 10-9
    polaronic shift, 10-7
Inelastic electron tunneling spectrum (IETS), 10-8 thru 10-9
InGaAs/GaAs quantum dots
  carrier storage, 2-17
  with additional AlGaAs barrier, 2-18 thru 2-20
Intensity-dependent refractive index (IDRI), 27-2
Interband transitions, 40-11
Intrinsic size effects
  dielectric function, 24-6
  quantum size effects
    absorption cross section *vs*. energy, 24-7
    jellium model, 24-7
    Mie theory, 24-6
    plasmon angular frequency, 24-8
Iridoviruses, 31-7

J

Jaynes–Cummings model, 1-6 thru 1-7
Josephson weak link, 7-2 thru 7-3

K

Kohn–Sham equation, 24-7
Kohn–Sham matrix, 27-4
Kuhn segment, 18-11

L

Landau–Ginzburg–Devonshire theory, 6-1 thru 6-2
Landau–Lifshitz–Gilbert–Slonczewski (LLGS) equation
  general form, 8-6
  solution based on FD method
    Ampère field, 8-10
    anisotropy field computation, 8-10
    exchange field computation, 8-9
    exchange length, 8-8
    Langevin equation, 8-11
    magnetostatic field computation, 8-10
    mesh for, 8-8
    spatial distribution field, 8-11
    thermal field computation, 8-12
    vs. FE method, 8-8
Landau–Zener model, 9-7
Langevin equation, 8-11
Laser emission properties
  microscopic generalization, semiconductor effects, 40-14 thru 40-15
  photon statistics modifications, 40-15 thru 40-16
  rate equations description, 40-13 thru 40-14
Laser-produced plasma EUV source, 20-10

- Laser-vaporization cluster beam mass spectroscopy, 15-4
- Lawrence Berkeley National Laboratory (LBNL), 20-3
- Lift-off technique, 6-7
- Linear bearing nanoswitch, 3-14 thru 3-16
- Liouville equation, 33-3
- Localized surface plasmons (LSPs), 28-5
- Loss–DiVincenzo proposal
  - RKKY interaction, 1-8 thru 1-9
  - using quantum dots and electric gates, 1-7 thru 1-8
- Low-dimensional sp^2 carbon structures, 7-4 thru 7-5
- Low-energy cluster beam deposition (LECBD), 24-13
- Low linewidth enhancement factor, 41-8

M

- Macroscopic polarization, 40-11
- Magnetically ordered materials, 8-1 thru 8-2
- Magnetic-metallothionein (mMT), 11-2
- Magnetic protein folding
  - Donnan effect, 11-3
  - first-order-like state transition model, 11-3
  - of Mn,Cd-MT, 11-2
  - protein folding phase diagram, 11-4
  - quasi-static thermal equilibrium dialysis, 11-5 thru 11-6
  - self-assemble, 11-5
  - three well model, 11-4
- Magnetocrystalline anisotropy, 4-4
- Magnetoresistive RAM, 2-5
- Mask Blank Development Center (MBDC), 20-11
- Maxwell–Garnett theory, 24-14
- MBE, *see* Molecular beam epitaxy
- Metal and semiconductor nanowire
  - dipole orientation, 26-8
  - frequency dispersion, 26-10
  - Gaussian properties, 26-11
  - linear optical phenomena, 26-9
  - nonlinear polarization, 26-8
  - plasmon resonance, 26-10
  - radiation pattern, 26-9
  - second harmonic generation (SHG), 26-7 thru 26-8
  - SHG angular distribution, 26-10
- Metal clusters and nanoparticles
  - alkali clusters, 24-15 thru 24-16
  - bimetallic cluster
    - annealed alloy measurement, 24-19
    - dielectric function, 24-19
    - gold–silver alloy nanoparticle, 24-19 thru 24-20
    - low energy ion scattering (LEIS) measurement, 24-20
    - Ni–Ag cluster optical absorption spectra, 24-20 thru 24-21
  - bulk metals
    - conduction band, 24-2 thru 24-3
    - dielectric function, 24-3
    - Drude dielectric function, 24-4
    - intraband and interband transition, 24-3
    - Kramers–Kronig relationship, 24-4
    - plasma frequency, 24-3
  - cluster source, 24-2
  - confined metallic systems, optical properties
    - chemical structure influence, 24-10 thru 24-11
    - cluster shape influence, 24-8 thru 24-9
    - dielectric function, 24-6
    - dipolar approximation, 24-10 thru 24-12
    - quantum size effects, 24-6 thru 24-8
    - quasi-static approximation/dipolar, 24-5 thru 24-6
    - scattering, absorption, and extinction, 24-4 thru 24-5
  - dielectric confinement, 24-2
  - electronic shells, 24-1
  - minute metal spheres, 24-2
  - nanoparticle synthesis
    - chemical methods, 24-12 thru 24-13
    - physical methods, 24-13 thru 24-14
  - noble metal clusters
    - discrete dipole approximation, 24-19
    - gold clusters, 24-17
    - jellium model, 24-18
    - silver clusters, 24-17
    - SPR, 24-16 thru 24-17
    - TDLDA, 24-18
  - optical spectroscopy, 24-14 thru 24-15
  - single nanoparticle, 24-2
  - spectroscopic techniques, 24-2
- Metallic dot
  - capacitance and coupling coefficient, 5-3 thru 5-4
  - influence of polarization
    - coupling coefficient analytical expression, 5-6 thru 5-7
    - dielectric medium potential, 5-7
  - method of images, 5-4 thru 5-5
  - numerical implementation
    - invertible matrix, 5-7
    - *vs.* classical approximation, 5-7 thru 5-9
  - semi-analytical expression for capacitance
    - B_1 coefficients expression, 5-4 thru 5-5
    - capacitance matricial expression, 5-6
    - partial capacitance expression, 5-6
  - semi-analytical expression for potential, 5-6
- Metal nanolayer-base transistor
  - band diagram
    - band banding conditions, 13-1
    - Co/Si-n Schottky diode characteristics, 13-2
    - hot-carrier diode, 13-1
    - MS contacts, 13-1 thru 13-2
    - TE theory, 13-2
  - fabrication
    - atomic arrangement, 13-4 thru 13-5
    - dependency of the current gain, 13-7
    - epitaxial growth, 13-4
    - glow discharge-deposition process, 13-4
    - magnetic multilayers, 13-5 thru 13-6
    - organic semiconductor as top layer, 13-6
    - test of dependency, 13-7
  - *J–V* characteristic, 13-3 thru 13-4
- Metal nanostructures
  - allowed and forbidden second harmonic emission modes
    - laterally limited light beam, 28-10 thru 28-11
    - plane-wave illumination, 28-9 thru 28-10
  - emission patterns and light polarization, 28-8 thru 28-9
  - field enhancements, 28-1
  - lightning rod effects, 28-1
  - second harmonic generation
    - nanoparticles, 28-5 thru 28-8
    - nanosystems, 28-3 thru 28-5
    - nonlinear optics, 28-2 thru 28-3
    - single gold nanoparticles, 28-11 thru 28-15
  - two-photon photo-luminescence (TPPL) yield, 28-15
- Metal-semiconductor (MS) junction, 13-1
- Microcavity lasers, 40-3
- Micromagnetic theory, 8-2
- Mie nanolasers, 39-11
- Mie theory, 24-6, 24-10, 24-12
- Mode-locked quantum-dot lasers
  - advantages, 41-5 thru 41-8
  - basics, 41-2
  - excited-state transitions, 41-9
  - passive mode locking, physics and devices, 41-3
  - pulse duration, 41-8
  - pulse repetition rates, 41-9
  - requirements, 41-3 thru 41-4
  - self-phase modulation and dispersion, 41-4 thru 41-5
  - semiconductor lasers, 41-2
  - temperature resilience, 41-9
- Molecular beam epitaxy (MBE), 6-5 thru 6-6
  - C_{60}, 15-10 thru 15-11
- Molecular junction conductance, 10-13 thru 10-15
- Molecular magnets preparation
  - magnetic properties, molecular nanostructures, 11-10 thru 11-11
  - magnetic protein folding
    - Donnan effect, 11-3
    - first-order-like state transition model, 11-3
    - of Mn,Cd-MT, 11-2
    - protein folding phase diagram, 11-4
    - quasi-static thermal equilibrium dialysis, 11-5 thru 11-6
    - self-assemble, 11-5
    - three well model, 11-4

  nanostructured semiconductor templates
    direct inkjet and mold imprinting, 11-7
    nanopatterned SAMs for 3D fabrication, 11-7 thru 11-8
    patterned self-assembly, 11-6 thru 11-7
  self-assembling growth
    patterned magnetic molecules, AFM, 11-8
    patterned MT molecules, AFM, 11-9
    Si template, SEM image, 11-9
Molecular nanomagnets
  coherent Néel vector tunneling
    doped antiferromagnetic molecular wheel, 9-10
    satellite resonance, 9-10
    two degenerate classical ground state, 9-9
  incoherent Zener tunneling in Fe$_8$, 9-7 thru 9-8
  phonon-assisted spin tunneling in Mn$_{12}$ acetate, 9-5
  single-molecule magnet transistors, 9-11 thru 9-12
  spin tunneling
    anticrossing of adiabatic eigenvalues, 9-3
    Berry's phase and path integrals (*see* Berry's phase interference)
    generalized master equation, 9-4
    Pauli or master equation, 9-3 thru 9-4
    Rabi oscillation, 9-3
    spin Hamiltonian, 9-2
    uniaxial anisotropy, 9-2
Molecular transistors, 7-18
Molecular tunnel junctions, 12-9 thru 12-11
Monte Carlo simulation, 24-20
Moore's law, 12-1 thru 12-2
Multiferroic tunnel junction, 6-18
Multilayer polymeric structures
  bacteriorhodopsin suspension preparation, 38-8
  basic elements, 38-7 thru 38-8
  containing mixture, 38-8 thru 38-9
  cyclicity processes, 38-5 thru 38-6
  data transmission and processing, 38-6
  element base adaptability, 38-6
  expected parameters, 38-7
  fragments, 38-7
  property, 38-9 thru 38-10
  requirements, 38-8
Multilayer (ML) reflective coating, 20-1, 20-3 thru 20-4
Multispacer patterning technique
  abstract technology
    bodies and surfaces, 17-15
    conformal deposition and isotropic etching, 17-15 thru 17-17
    directional processes, 17-17 thru 17-18, 17-18
  additive route—S^{n}PT$_+$, 17-5 thru 17-7
  addressing cross-point, 17-10 thru 17-11
  concrete technology, 17-18 thru 17-19
  crossbar-architecture fabrication steps, 17-2
  crossbar comparison, 17-11 thru 17-12
  electronics application, 17-12
  energetics application, 17-12
  fractal nanotechnology
    fractals in nature, 17-12 thru 17-13
    producing nanoscale fractals via S^{n}PT$_x$, 17-13 thru 17-14
  multiplicative route—S^{n}PT$_x$, 17-8 thru 17-9
  nonlithographic nanowire preparation
    imprint lithography (IL), 17-3 thru 17-4
    spacer patterning technique (SPT), 17-4 thru 17-5
  three-terminal molecules, 17-9 thru 17-10

N

Nano-dot, *see* Spherical dot in plate capacitor
Nanoelectromechanical systems (NEMS) based memory
  capacitive force response, 3-9
  feedback-controlled nanocantilevers, 3-12 thru 3-13
  linear bearing nanoswitch, 3-14 thru 3-16
  nanorelays, 3-11 thru 3-12
  nonvolatile random access memory, 3-10 thru 3-11
  properties, 3-9
  vertically aligned carbon nanotubes, 3-13 thru 3-14
Nanoelectronics lithography
  advanced lithographic processes
    CD-SAXS, 19-27
    CD-scanning electron microscopy (CD-SEM), 19-26
    double exposure method, 19-24 thru 19-25
    ellipsometry measurements, 19-26
    line edge roughness measurement, 19-27
    overlay measurements, 19-27
    scatterometry measurements, 19-26
    spacer double patterning method, 19-25
  DUV lithography
    CaF$_2$ lens elements, 19-11
    depth of focus (DOF), 19-9 thru 19-10
    double patterning, 19-13 thru 19-14
    fused silica glass, 19-10
    193 nm immersion lithography, 19-11 thru 19-12
    numerical aperture, 19-11
    optical-projection lithography system, 19-9
    photolithography exposure process, 19-9
    plot of critical feature size, 19-10
    ultimate resolution limits, 19-14
  electron-beam lithography
    cost of ownership, 19-17
    overlay, 19-16 thru 19-17
    resolution, 19-14 thru 19-15
    throughput, 19-15 thru 19-16
  NIL
    defect inspection and dimensional metrology, 19-22
    functional materials, 19-19
    nanoimprint materials development, 19-21
    overlay accuracy and control, 19-23
    residual layer control, 19-22 thru 19-23
    technology examples, 19-23 thru 19-24
    template fabrication and availability, 19-21 thru 19-22
    thermal NIL, 19-18 thru 19-19
    tool types, 19-19 thru 19-20
    transfer printing, 19-19
    UV NIL, 19-19
  photoresist technology
    base quencher additives, 19-6 thru 19-7
    EUV lithography, 19-8
    line and space pattern, 19-4 thru 19-5
    material structure advances, 19-7 thru 19-8
    193 nm immersion lithography, 19-8
    photolysis process of photoacid generator, 19-5 thru 19-6
    positive and negative tone resist, 19-4
    residual swelling fraction, 19-7
    resist materials, 19-4 thru 19-5
Nanofibers, light scattering
  angular distribution, 30-9
  cylindrical nanofiber diffraction
    Bessel function, 30-4
    electromagnetic field components, 30-3
    Fourier-transformed amplitudes, 30-5 thru 30-6
    incident beam polarization, 30-4 thru 30-5
    incident wave electric field, 30-3
    inverse Fourier transform, 30-4
  definition, 30-1
  dispersion curves, 30-12
  exciton resonance excitation, 30-10 thru 30-11
  Fresnel reflection coefficient, 30-2
  Hankel functions, 30-1
  He-Cd laser, 30-11
  ideal surface reflection, 30-8 thru 30-9
  inorganic semiconductor materials, 30-1
  multiple-scattering linear equation, 30-2
  normal modes, 30-6
  optical process, 30-1
  photoluminescence measurement, 30-12
  photomultiplier, 30-11
  polarization properties, 30-12
  pulsed Gaussian light beam, 30-12
  scattered field calculation, 30-6 thru 30-8
  scattered light intensity, 30-12
  scattering efficiency, 30-2

- T-matrix method, 30-2
- total scattered intensity *vs.* frequency, 30-10
- total scattered intensity *vs.* incidence angle, 30-9 thru 30-10

Nanoimprint lithography (NIL)
- cavity fill process, 18-4 thru 18-5
- control ordering and assembly processes
  - crystal growth process, 18-8
  - crystallization process, 18-8
  - full confinement factor, 18-11 thru 18-14
  - partial confinement factor, 18-9 thru 18-11
  - semicrystalline polymers morphology, 18-8
- defect inspection and dimensional metrology, 19-22
- functional materials, 19-19
- functional materials nanoshaping
  - antireflective structure creation, 18-6
  - biomaterial shaping, 18-6 thru 18-7
  - interface shaping, 18-5
  - nanopore generation, 18-5
  - optical grating creation, 18-5 thru 18-6
- imprinting resists, 18-4
- molds in, 18-4
- nanoimprint materials development, 19-21
- overlay accuracy and control, 19-23
- residual layer control, 19-22 thru 19-23
- resolution limit, 18-4
- technology examples, 19-23 thru 19-24
- template fabrication and availability, 19-21 thru 19-22
- thermal NIL, 19-18 thru 19-19
- for thermoplastic polymers, 18-3
- tool types, 19-19 thru 19-20
- transfer printing, 19-19
- UV NIL, 19-19

Nanolasers
- applications, 39-12
- definition, 39-5
- laser radiation property, 39-1 thru 39-2
- nanowire, 39-5 thru 39-8
- organic nanolasers, 39-11
- overview, 39-2
- photonic crystal lasers, 39-10
- plasmonic lasers, 39-8 thru 39-10
- rate equation approach, 39-4 thru 39-5
- scattering lasers, 39-10 thru 39-11
- semiconductor lasers, 39-2 thru 39-4

Nanomaterials
- cathodoluminescence
  - band structure, types, 21-2
  - donor–acceptor pair transition, 21-2
  - exciton, 21-3
  - recombination process, 21-2
  - TEM-CL (*see* Transmission electron microscope-cathodoluminescence)
- kinetic energy, 21-8 thru 21-9
- optical spectroscopy
  - global photoluminescence, 22-1
  - luminescence hole-burning (LHB) spectroscopy, 22-1
  - luminescence spectra, 22-1 thru 22-2
  - multiexciton generation, 22-6 thru 22-7
  - nanoparticle quantum dots, 22-4 thru 22-6
  - optical properties, 22-1
  - semiconductor nanoparticles, 22-1
  - single molecule spectroscopy, 22-2
  - SWNT (*see* Single-walled carbon nanotube)
- photoluminescence (PL), 21-1 thru 21-2
- quantum well, 21-9 thru 21-10
- quantum wire
  - density of states, 21-10
  - electron wavefunction, 21-10
  - hole wave-functions, 21-11
  - homogeneous electric field, 21-13
  - intensity and polarization ratio, 21-14 thru 21-15
  - Luttinger–Kohn Hamiltonian, 21-11
  - optical absorption, 21-14
  - photoluminescence excitation (PLE) spectrum, 21-13
  - polarization anisotropy, 21-11
  - quantum confinement effect, 21-12 thru 21-13
  - transition matrix element, 21-12 thru 21-13
- semiconductor quantum structures, optical properties
  - Bloch function, 21-5
  - conversion matrix, 21-6
  - light emission theory, 21-6 thru 21-7
  - Luttinger–Kohn Hamiltonian, 21-5
  - polarization, 21-7 thru 21-8
- semiconductor quantum wires, 21-1
- transmission electron microscope (TEM), 21-1 thru 21-2

Nanomemories
- based on III–V semiconductor QD, 2-7 thru 2-10 (*see also* Quantum dot memories)
- semiconductor
  - charge trap memories, 2-6 thru 2-7
  - single electron memories, 2-7

Nanometer-sized ferroelectric capacitors
- characterization of films and capacitors
  - *P-E* hysteresis loop measurements, 6-10 thru 6-11
  - RBS, 6-8
  - scanning probe techniques, 6-11 thru 6-13
  - XRD, 6-8 thru 6-10
- deposition techniques
  - combinatorial PLD and MBE systems, 6-8
  - MBE, 6-5 thru 6-6
  - PLD, 6-6
  - sputter deposition, 6-6 thru 6-7
- ferroelectricity fundamentals
  - ferroelectric correlation length, 6-4
  - ferroelectric hysteresis loops, 6-4
  - perovskite oxide BaTiO3 unit cell, 6-3
  - phenomenological definition, 6-2
  - second-order ferroelectric phase transition, 6-3
  - thermodynamic theory, 6-3
- low-temperature superconducting Josephson junctions, 6-18 thru 6-19
- multiferroic tunnel junction, 6-18
- novel functional oxide tunnel junctions, 6-18
- patterning
  - electron beam lithography, 6-8
  - lift-off technique, 6-7
  - photo-mask steps, 6-8
- physical phenomena
  - depolarizing-field effect, 6-15 thru 6-16
  - intrinsic size effect, ultrathin films, 6-17
  - strain effect, 6-14 thru 6-15
- "research triangle," 6-1 thru 6-2
- tunneling magnetoresistance (TMR), 6-18

Nanoparticle quantum dots
- Au nanoparticles, 22-6
- CdSe/ZnS nanoparticles, 22-4 thru 22-5
- close-packed monolayer films, 22-5
- decay times, 22-5 thru 22-6
- enhancement and quenching phenomena, 22-5
- light-emitting neutral nanoparticle, 22-4
- PL blinking phenomena, 22-4
- unique exciton energy transfer, 22-5

Nanophotonic buffer memory, 33-11 thru 33-12

Nanophotonic devices
- AND-gate, ZnO nanorod
  - electronic carrier penetration, 36-10
  - exciton–phonon couplings, 36-7 thru 36-8
  - exciton populations, 36-9
  - input laser llumination, 36-11
  - micro-channel plate, 36-7
  - near-field spectroscopy, 36-7 thru 36-8
  - nutation frequency evaluation, 36-10
  - schematic depict, SQW, 36-9
  - single-quantum-well structures(SQW), 36-7 thru 36-8
  - switching operation, 36-10 thru 36-11
- dissipation-controlled nanophotonic devices
  - carrier manipulation (up-converter), 33-7 thru 33-8
  - energy transfer between two quantum dots, 33-2 thru 33-5
  - nanophotonic switch, 33-5 thru 33-7
- operation
  - excitation energy transfer, 35-1 thru 35-3
  - nanophotonic AND gate, 35-4 thru 35-5
  - nanophotonic NOT gate, 35-5 thru 35-6
  - photonic devices interconnection, 35-6 thru 35-7

- quantum dot materials
    - controlling the energy transfer, 36-4 thru 36-6
    - discrete energy levels, 36-1 thru 36-2
    - dissipated optical energy transfer, 36-2 thru 36-4
    - near-filed optically coupled ZnO QDs, 36-4 thru 36-7
- schematic representation, 33-2
- using spatial symmetries
    - AND- and XOR-logic gates, 33-10 thru 33-11
    - nanophotonic buffer memory, 33-11 thru 33-12
    - quantum entangled states manipulation, 33-12 thru 33-13
    - selective energy transfer, 33-9 thru 33-10
    - symmetric and antisymmetric states, 33-8 thru 33-9

Nanophotonics
- devices (*see* Nanophotonic devices)
- fabrication
    - NFO-CVD (*see* Nonadiabatic near-field optical CVD)
    - nonadiabatic near-field photolithography, 35-10 thru 35-11
- hierarchical architectures
    - memory retrieval, 34-7 thru 34-9
    - physical and functional, 34-6 thru 34-7
    - unscalable design, 34-8 thru 34-10
- optical excitation transfer (*see* Optical excitation transfer)
- parallel architecture
    - broadcast interconnects, 34-4 thru 34-5
    - global summation, 34-4
    - memory-based architecture, 34-3 thru 34-4
    - secure signal transfer, 34-5 thru 34-6
- switch
    - energy diagram, 33-6
    - exciton population, 33-6 thru 33-7
    - ON- to OFF-state, 33-7
    - switching operation, 33-6
- versatile functionalities, 34-10 thru 34-12
- waveguides (*see* Nanophotonic waveguides)

Nanophotonic waveguides
- active lossy waveguides
    - cladding layer, 37-9
    - double-peak structure, 37-11
    - electric field intensity, 37-10
    - Fresnel formulae, 37-10
    - substrate modes, 37-12
    - TE–TM splitting, 37-11
- active Si-nc waveguides
    - direct facet-coupling, 37-8 thru 37-9
    - PL spectra, 37-7 thru 37-8
    - prism coupling, 37-8
    - refractive index, 37-7
- external source light propagation, 37-5
- internal photoluminescence propagation, 37-4 thru 37-5
- losses and optical amplification
    - insertion (coupling) loss, 37-5
    - loss measurement methods, 37-5 thru 37-6
    - propagation loss, 37-5
    - pump-and-probe (PP) method, 37-5, 37-7
    - shifting excitation spot (SES) method, 37-5 thru 37-6
    - variable stripe-length (VSL) method, 37-5 thru 37-7
- nanocrystalline waveguides
    - light amplification, 37-12 thru 37-13
    - optical couplers, 37-14
    - photonic circuits, 37-13
    - PL spectra, 37-13
- planar and rib waveguide fabrication
    - advantages, 37-4
    - cladding layers, 37-3
    - fused quartz implantation, 37-2 thru 37-3
    - optical transparent matrix, 37-2
    - reactive silicon deposition, 37-3
    - SiO$_2$ film, 37-3
    - spectral filtering, 37-2
- silicon nanophotonics, 37-1 thru 37-2

Nanorelays, 3-11 thru 3-12

Nanoscale excitons
- definition, 23-1
- density of states, 23-2 thru 23-3
- electron-hole attraction, 23-1
- electron quantum confinement, 23-4
- exciton-binding energy, 23-4
- extended bulk system, 23-5
- Frenkel excitons, 23-1
- optical properties, semiconductor nanostructure, 23-2
- semiconductor quantum dots
    - nonlinear optical properties, 23-5 thru 23-6
    - two-photon absorption (2PA), 23-6 thru 23-8
- single-wall carbon nanotube (SWCNT), 23-4 thru 23-5
- spatial electron confinement, 23-3 thru 23-4
- Wannier–Mott excitons, 23-1

Nanoscale spin valves
- anisotropic effects, 8-4
- computational micromagnetics (*see* Computational micromagnetics)
- energy balance and free energy functional formulation, 8-5
- equation of motion, 8-5 thru 8-6
- exchange interactions, 8-4
- external bias field, 8-5
- magnetization vector, 8-3 thru 8-4
- magnetostatic interactions, 8-4 thru 8-5
- spin-transfer torque effects, 8-6 thru 8-7

Nanoshaping functional materials
- antireflective structure creation, 18-6
- biomaterial shaping, 18-6 thru 18-7
- interface shaping, 18-5
- nanopore generation, 18-5
- optical grating creation, 18-5 thru 18-6

Nanotube-based superconducting quantum interferometers
- magnetometry applications, 7-13 thru 7-14
- quantum information applications, 7-14 thru 7-15
- quantum interference with weak links, 7-10 thru 7-11
- tunability of the phase shift, 7-11 thru 7-12

Nanowire lasers
- basic properties, 39-6
- cavity structure, 39-8
- GaN nanowire cross section, 39-8
- ring resonator, 39-6 thru 39-7
- schematic representation, 39-5 thru 39-6

N$_2$ atmospheric train-sublimation method, 15-5 thru 15-6

Near-field optically coupled ZnO QDs
- energy diagram, 36-6
- growth time dependence, 36-5
- TEM image, 36-5 thru 36-6
- time constants dependence, 36-7
- time-resolved PL intensity, 36-5, 36-7

Near-field photoluminescence (NFPL), 36-7

Negative differential resistance (NDR), 10-13

Nonadiabatic near-field optical CVD (NFO-CVD), 35-7
- electron potential curves, 35-9
- photodissociation, 35-7
- photon flux relationship, 35-9
- sapphire substrate, 35-8
- shear-force topographical images, cross section, 35-7 thru 35-8

Nonadiabatic photolithography, 35-10

Noncircular rings, 4-9 thru 4-10

Nonconventional semiconductor memories
- FeRAM, 2-4 thru 2-5
- magnetoresistive RAM, 2-5
- PCRAM, 2-5 thru 2-6

Nonlinear optics
- average Hirshfeld charge, 27-11
- cadmium selenide clusters ((CdSe)$_n$), 27-5
- cluster materials, 27-5
- density functional theory (DFT), 27-9
- Douglas–Kroll approximation, 27-10
- electric dipole moment, 27-1
- excitation frequency, 27-11 thru 27-12
- frequency-dependent optical polarizability, 27-6
- frequency variation, 27-6 thru 27-8
- generalized-gradient approximation (GGA) technique, 27-6
- hyperpolarizability tensor, 27-1
- inorganic materials, 27-5
- intensity-dependent refractive index (IDRI), 27-2
- lithium tetramer, 27-10
- mean dipole polarizability, 27-5 thru 27-6
- mean second-order hyperpolarizability, 27-9 thru 27-10

  - metallic cluster, 27-5
  - mixed metal cluster, 27-11
  - nanocluster, 27-9
  - Nd:YAG laser frequency, 27-9
  - organic materials, 27-5
  - quantum formulation
    - average polarizability tensor, 27-4
    - density matrix, 27-3
    - dipole interaction model, 27-2
    - frequency-dependent response properties, 27-5
    - Kohn–Sham equation, 27-3
    - Lagrangian multiplier matrix, 27-4
    - Taylor expansion, 27-4
    - time-dependent density-functional theory (TDDFT), 27-3
    - two-state model, 27-2
  - second-order hyperpolarizability tensor, 27-12
  - second-order nonlinear polarization, 27-2
  - semiconducting cluster, 27-5
  - of semiconductor nanostructures (*see* Semiconductor nanostructures)
  - symmetrized fragment orbitals (SFO), 27-11
  - time-dependent Hartree–Fock (TDHF) method, 27-9
- Nonlithographic nanowire preparation
  - imprint lithography (IL), 17-3 thru 17-4
  - spacer patterning technique (SPT), 17-4 thru 17-5
- Nonplanar multi-gate transistors, 12-3 thru 12-4
- Nonvolatile memories (NVM), *see* Flash memories

O

- Optical excitation transfer
  - parallel architecture
    - broadcast interconnects, 34-4 thru 34-5
    - global summation, 34-4
    - memory-based architecture, 34-3 thru 34-4
    - secure signal transfer, 34-5 thru 34-6
  - via optical near-field interactions
    - global summation, 34-2 thru 34-3
    - light–matter interactions, 34-2
- Optical nanofountain, 35-6 thru 35-7
- Optical near-field (ONF), 34-10 thru 34-12
  - coupling strength, 33-4 thru 33-7
  - ONF system, 32-6 thru 32-7
- Optical neuronets
  - basic elements, multilayered structures, 38-7 thru 38-8
  - cyclicity processes, 38-5 thru 38-6
  - data processing, 38-6
  - element base parameters, 38-7
  - light fluxes formation, 38-6
  - multilayered constructions, 38-7
  - polymeric film property, 38-9 thru 38-10
  - polymeric film requirements, 38-8
  - polymeric mixture preparation, 38-8 thru 38-9
  - suspension preparation, 38-8
- Optical-projection lithography system, 19-9
- Optical spectroscopy
  - luminescence hole-burning (LHB) spectroscopy, 22-1
  - luminescence spectra, 22-1 thru 22-2
  - multiexciton generation, 22-6 thru 22-7
  - nanoparticle quantum dots, 22-4 thru 22-6
  - optical properties, 22-1
  - semiconductor nanoparticles, 22-1
  - single molecule spectroscopy, 22-2
  - SWNT (*see* Single-walled carbon nanotube)
- Optical transition matrix element, 21-7
- Organic FET
  - am-bipolar operations, 15-17 thru 15-18
  - device structures, 15-16
  - field effect doping, 15-17
  - metal/organic layer interface, 15-16 thru 15-17
- Organic nanolasers, 39-11
- Ovonic unified memory (OUM), *see* Phase-change RAM

P

- Passivation, 14-7 thru 14-9
- Passive mode locking
  - physics and devices, 41-3
  - requirements, 41-3 thru 41-4
- Pauli equation, 9-3 thru 9-4
- Phase-change RAM (PCRAM), 2-5 thru 2-6
- Phase-matching condition, 28-3
- Phase-shifting point-diffraction interferometer (PS/PDI), 20-5
- Phonon-assisted spin tunneling, Mn$_{12}$ acetate, 9-5
- Photolithography, 18-1 thru 18-3; *see also* Nonadiabatic photolithography
- Photoluminescence (PL)
  - energy-band diagram, 25-1 thru 25-2
  - photon momentum, 25-1
  - porous silicon (PS), 25-2
  - quantum confinement (QC) model, 25-2
  - quantum size effects, 25-3
  - silicon nanostructures
    - Arrhenius plot, 25-11
    - confinement energy, 25-8
    - co-sputtered SiNC, 25-7 thru 25-8
    - Coulomb exchange interaction, 25-11
    - dynamical characteristics, 25-10
    - excitons, 25-8
    - *f*-sum rule/Thomas–Reiche–Kuhn sum rule, 25-13
    - nanocrystal size distribution, 25-12 thru 25-13
    - nonradiative process, 25-14 thru 25-15
    - optical characterizations and luminescence bands, 25-7
    - oscillator strength *vs.* confinement energy, 25-13 thru 25-14
    - PL decay time, 25-10
    - PL yield *vs.* silicon volume content, 25-9
    - radiative process, 25-14
    - relaxation process, 25-11
    - spin-orbit interaction, 25-12
    - surface passivation, 25-9
    - synthesis, 25-3 thru 25-7
  - SiO$_2$, 25-1
  - surface phenomena, 25-3
  - surface-to-volume (STV) ratio, 25-2 thru 25-3
  - vibron model
    - adiabatic approximation, 25-17
    - conduction-ISL (CISL), 25-20
    - core nanocrystal, 25-22
    - electron–phonon interaction, 25-18
    - inter-sub-level (ISL) optical transition, 25-19
    - laser pyrolysis technique, 25-21
    - nanocrystal passivation, 25-16
    - noncoherent surface vibration, 25-19
    - phonon dispersion relation, 25-16
    - photoinduced absorption (PIA) spectra, 25-19 thru 25-20
    - polaron formation mechanism, 25-18
    - polar optical phonon, 25-19
    - resonant coupling condition, 25-17
    - silicon–oxygen bond, 25-16 thru 25-17
    - silicon–silicon dioxide (SiO$_2$) interface, 25-15 thru 25-16
    - Si–O vibration, 25-20 thru 25-21
    - valence-ISL (VISL), 25-20
- Photolysis process, 19-5 thru 19-6
- Photo-mask steps, 6-8
- Photonic crystal lasers, 39-10
- Photonic crystal waveguide, 37-1 thru 37-2
- Photon localization
  - canonical field quantization
    - eigenstates, 32-5
    - Hamiltonian operator, 32-4 thru 32-5
    - harmonic oscillator, 32-4
    - Helmholtz equation, 32-4
    - light–matter interaction, 32-3
    - quantization procedure, 32-5
  - dressing mechanism and spatial localization
    - Davydov transformation, 32-12 thru 32-13, 32-16
    - effective interaction Hamiltonian operator, 32-8
    - effective sample-probe-tip interaction, 32-10
    - electric displacement operator, 32-9
    - electromagnetic field correlations and intermolecular interactions, 32-6 thru 32-7
    - exciton polariton, 32-9
    - Hamiltonian model, 32-11 thru 32-12
    - irrelevant macroscopic subsystem, 32-7 thru 32-8
    - mean field approximation, 32-14
    - nanomaterial system, 32-7

    - phonon-assisted photodissociation, 32-17
    - photon–electron–phonon interacting system, 32-7
    - photon hopping operator, 32-15 thru 32-16
    - photon number operator, 32-14 thru 32-15
    - photon–phonon coupling constant, 32-15
    - pseudo one-dimensional near-field optical probe system, 32-11
    - P space and Q space, 32-8
    - quasiparticle and coherent state, 32-13 thru 32-14
    - virtual photon cloud, 32-5 thru 32-6
    - Yukawa functions, 32-11
  - effective spatial wave function, 32-3
  - Heisenberg's uncertainty principle, 32-1
  - position-representation wave function, 32-1
  - quantum electrodynamics (QED), 32-1
  - virtual cloud effects, 32-2
  - wave function definition, 32-2 thru 32-3
- Photoresist technology
  - base quencher additives, 19-6 thru 19-7
  - EUV lithography, 19-8
  - line and space pattern, 19-4 thru 19-5
  - material structure advances, 19-7 thru 19-8
  - 193 nm immersion lithography, 19-8
  - photolysis process of photoacid generator, 19-5 thru 19-6
  - positive and negative tone resist, 19-4
  - residual swelling fraction, 19-7
  - resist materials, 19-4 thru 19-5
- Piezoresponse force microscopy (PFM), 6-12 thru 6-13
- Plasmon amplification, 26-5
- Plasmon frequency, 26-4
- Plasmonic lasers
  - bowtie structures, 39-9
  - dipole nanolaser, 39-9 thru 39-10
  - horseshoe optical nanoantenna., 39-9
  - surface plasmon mode confinement, 39-8
- Plasmonic waveguides, 37-1 thru 37-2
- PLD, *see* Pulsed laser deposition
- Point contacts and conductance quantum, 12-11
- Poisson–Schrödinger resolution, 5-13, 5-18
- Polarization-sensitive nanowire and nanorod optics
  - absorption coefficient, 26-4
  - anisotropic nanostructure, 26-1
  - core–shell nanowires
    - luminescence amplification, 26-13
    - potential distribution, 26-11
    - spectra and polarization dependence, 26-11 thru 26-13
  - depolarization factors, 26-2
  - electrodynamic effects, 26-2
  - first-kind Hankel function, 26-3
  - Helmholtz wave equation, 26-3
  - light-induced phenomena, 26-1
  - metal and semiconductor nanowires, nonlinear phenomena
    - dipole orientation, 26-8
    - frequency dispersion, 26-10
    - Gaussian properties, 26-11
    - linear optical phenomena, 26-9
    - nonlinear polarization, 26-8
    - plasmon resonance, 26-10
    - radiation pattern, 26-9
    - second harmonic generation (SHG), 26-7 thru 26-8
    - SHG angular distribution, 26-10
  - metal nanostructures, 26-4 thru 26-5
  - optical properties, 26-3
  - photoconductivity, 26-2 thru 26-3
  - planar interface, 26-1
  - plasmon spectra
    - longitudinal plasmon shift, 26-16 thru 26-17
    - self-assembling, 26-15 thru 26-16
  - polarization memory, random nanorod arrays
    - luminescence polarization ratio, 26-14
    - polarization antimemory, 26-15
    - polarization-dependent effects, 26-13
  - polarization ratio, 26-4
  - semiconductor nanostructures, 26-2
    - intensity ratio, 26-6
    - isotropic interband matrix element, 26-5
    - Poynting vector, 26-6
    - radiation characteristics, 26-7
    - refractive index, 26-5
    - ZnO NW luminescence spectrum, 26-7
- Power spectral density (PSD), 20-4
- Proximity effect, 7-15 thru 7-17
- Pulsed laser deposition (PLD), 6-6

Q

- Quantum computing
  - atom–light interaction
    - Jaynes–Cummings model, 1-6 thru 1-7
    - in RWA, 1-4 thru 1-6
  - Loss–DiVincenzo proposal
    - RKKY interaction, 1-8 thru 1-9
    - using quantum dots and electric gates, 1-7 thru 1-8
  - physical implementation conditions, 1-3
  - qubits and quantum logic gates, 1-2 thru 1-3
  - semiconductor quantum dots
    - classical Faraday effect, 1-13
    - electronic band structure, 1-12
    - quantum Faraday effect, 1-13 thru 1-14
  - SPFE
    - conditional Faraday rotation, 1-16
    - GHZ quantum teleportation, 1-16 thru 1-18
    - Jaynes–Cummings Hamiltonian description, 1-14 thru 1-15
    - quantifying EPR entanglement, 1-16
    - quantum computing, 1-18 thru 1-20
  - with molecular magnet
    - double-well potential, 1-11
    - Feynman diagrams, 1-10
    - Grover's algorithm implementation, 1-12
    - single-spin Hamiltonian description, 1-10
  - Zeeman effects
    - strong external field, 1-4
    - weak external field, 1-4
- Quantum dot laser
  - edge emitter laser diodes (*see* Edge emitter laser diodes)
  - electronic state
    - continuum approaches and atomistic models, 40-5
    - envelope-function approximation, 40-5
    - localized and delocalized, self-assembled structures, 40-4 thru 40-5
  - laser emission properties
    - microscopic generalization, 40-14 thru 40-15
    - photon statistics modifications, 40-15 thru 40-16
    - rate equations, 40-13 thru 40-14
  - material systems, 40-1 thru 40-2
  - optical gain
    - gain saturation, refractive index, and α-factor, 40-12 thru 40-13
    - interband transitions and macroscopic polarization, 40-11
    - optical gain calculations, 40-11 thru 40-12
    - optical susceptibility, 40-10 thru 40-11
    - transparency carrier density, 40-10
  - scattering process
    - carrier–carrier interaction, 40-6 thru 40-8
    - carrier–phonon interaction, 40-8 thru 40-10
- Quantum dot (QD) memories; *see also* Self-organized quantum dots
  - hysteresis measurements, 2-22
  - write time measurements, 2-22 thru 2-23
- Quantum entangled states, 33-12 thru 33-13
- Quantum Faraday effect, 1-13 thru 1-14
- Quantum spin tunneling
  - anticrossing of adiabatic eigenvalues, 9-3
  - Berry's phase and path integrals (*see* Berry's phase interference)
  - generalized master equation, 9-4
  - Pauli or master equation, 9-3 thru 9-4
  - Rabi oscillation, 9-3
  - spin Hamiltonian, 9-2
  - uniaxial anisotropy, 9-2
- Quantum teleportation, 1-16 thru 1-18
- Quasiparticle poisoning
  - energetics of the nonequilibrium process, 16-7 thru 16-9

  - in ramped current measurement, 16-11 thru 16-12
  - in rf measurement, 16-15
- Quasi-static thermal equilibrium dialysis, 11-5 thru 11-6

R

- Rabi oscillation, 9-3, 29-16 thru 29-19
- Radosavljevic's device, 3-5
- Random lasers, 39-11
- Rate window method, 2-15 thru 2-16
- RBS, *see* Rutherford backscattering spectrometry
- Rectangular and elliptical islands, 4-6 thru 4-7
- Relevant nanometric subsystem, 32-7 thru 32-8
- "Research triangle," 6-1 thru 6-2
- Resistively and capacitively shunted junction (RCSJ) model, 16-3 thru 16-4
- Richardson's constant, 13-2
- Rotating wave approximation (RWA), 1-4 thru 1-6
- Rutherford backscattering spectrometry (RBS), 6-8

S

- Sawyer-Tower circuit, 6-10
- Scattering lasers, 39-10 thru 39-11
- Schottky barrier, 13-1
- Schottky contact, 2-11 thru 2-12
- Second harmonic generation
  - nanoparticles
    - crystallographic defects, 28-7
    - Feynman diagram, 28-6
    - fundamental wavelength (FW) field interaction, 28-5
    - long-wavelength limit, 28-7 thru 28-8
    - perturbation theory, 28-6
    - retardation effects, 28-6
    - scanning near-field optical microscopy (SNOM), 28-5
    - SHG selection rules, 28-7
  - nanosystems
    - low-symmetry particles, 28-4
    - nonuniform illumination, 28-4
    - random metal nanostructures and rough metal surfaces, 28-5
    - resonant metal particles, 28-4 thru 28-5
    - spherical isolated particles and plane-wave illumination, 28-3
  - nonlinear optics, 28-2 thru 28-3
  - SH electric field polarization, 28-9
  - single gold nanoparticles
    - Fischer pattern, 28-12
    - lithographed nanorods, 28-12 thru 28-13
    - near-field FW properties, 28-13
    - near-field nonlinear optical microscopy, 28-11
    - nonlinear optical behavior, 28-13 thru 28-14
    - second harmonic polarization analysis, 28-14 thru 28-15
    - SHG efficiency, 28-14
    - SNOM, 28-11 thru 28-12
- Second harmonic maps, 28-12 thru 28-14
- Self-organized quantum dots
  - based nanomemories, 2-10
  - carrier emission process, 2-9
  - charge carrier storage
    - hole storage in GaSb/GaAs QD, 2-17 thru 2-18
    - in InGaAs/GaAs QD, 2-17
    - InGaAs/GaAs QD with AlGaAs barrier, 2-18 thru 2-20
    - storage time, 2-20 thru 2-21
  - electron and hole states, 2-9
  - fabrication, 2-8
  - localization energy and carrier storage time determination (*see* Capacitance spectroscopy)
- Self-phase modulation (SPM), 41-4 thru 41-5
- Semi-analytical approach numerical validation
  - convergence, 5-19
  - FEM approach, 5-19 thru 5-21
- Semiconductor lasers
  - Bragg mirrors, 39-3
  - in-plane laser, 39-2 thru 39-3
  - vertical cavity surface emitting laser (VCSEL), 39-2
- Semiconductor nanomemories
  - charge trap memories, 2-6 thru 2-7
  - single electron memories, 2-7
- Semiconductor nanostructures
  - four-wave mixing response, perturbation theory
    - electron–hole pair, 29-11
    - excitation field possess, 29-13
    - excitons, definition, 29-11
    - Green function, 29-16
    - macroscopic exciton polarization, 29-12
    - multiwave mixing response, 29-14
    - Pauli blocking spectrum, 29-15
    - phase factor, 29-14
    - quantum well, 29-11
    - semiconductor nonlinear response, 29-11
    - S-matrix, 29-12
  - Green's functions method, 29-1
  - quantum equations
    - Coulomb energy, 29-8
    - Coulomb gauge, 29-6
    - density matrix, 29-6
    - divergent band-gap renormalization, 29-8
    - electromagnetic field–matter system, 29-6
    - electron–electron interaction, 29-7
    - Heisenberg equation, 29-7
    - Heisenberg representation, 29-6
    - Helmholtz's theorem, 29-7
    - radiative decay, 29-9
    - rotating wave approximation, 29-9 thru 29-10
  - quantum field theory, 29-1
  - rabi oscillations, nonperturbative methods
    - exciton polarization, 29-19
    - Fourier series, 29-18
    - Hartree–Fock approximation, 29-17
    - integro-differential operator, 29-17
    - interband polarization, 29-18
    - spacial decoupling scheme, 29-16
  - semiconductor-field hamiltonian, $\mathbf{k}\cdot\mathbf{p}$-approximation
    - Bloch mode, 29-2
    - Coulomb gauge, 29-4
    - effective mass, 29-5
    - electron field operator, 29-3
    - Fourier series, 29-3
    - heavy–light hole hybridization, 29-5
    - Hermitian equation, 29-2
    - interband coupling, 29-3
    - intraband coupling, 29-4
    - Luttinger–Kohn function, 29-2
    - quantum dynamics, 29-2
    - Schrieffer–Wolff transformation, 29-3
    - semiconductor quantum dot, 29-5
    - spatial field operators, 29-4
    - spatial Fourier spectrum, 29-6
    - spin-orbit interaction, 29-5
- Semiconductor quantum dots
  - nonlinear optical properties, 23-5 thru 23-6
  - two-photon absorption (2PA)
    - amplified stimulated emission and lasing, 23-7
    - CdSe, 23-6
    - continuous-wave (cw) Gaussian beam, 23-7
    - incident laser intensity, 23-6 thru 23-7
    - optical power limiting and biological imaging, 23-7
    - two-photon sensitizer, 23-7 thru 23-8
- Silica deposition vesicle (SDV), 31-7
- Siliconize photonics, 37-1
- Silicon nanostructures
  - Arrhenius plot, 25-11
  - confinement energy, 25-8
  - co-sputtered SiNC, 25-7 thru 25-8
  - Coulomb exchange interaction, 25-11
  - dynamical characteristics, 25-10
  - excitons, 25-8
  - *f*-sum rule/Thomas-Reiche-Kuhn sum rule, 25-13
  - nanocrystal size distribution, 25-12 thru 25-13
  - nonradiative process, 25-14 thru 25-15
  - optical characterizations and luminescence bands, 25-7
  - oscillator strength *vs*. confinement energy, 25-13 thru 25-14
  - PL decay time, 25-10

- PL yield *vs.* silicon volume content, 25-9
- radiative process, 25-14
- relaxation process, 25-11
- spin-orbit interaction, 25-12
- surface passivation, 25-9
- synthesis
  - co-sputtering method, 25-4
  - electrochemical etching, 25-3
  - high angle annular dark field (HAADF) detector, 25-6 thru 25-7
  - laser pyrolysis, 25-5
  - molecular beam epitaxy (MBE) system, 25-6
  - silicon nanocrystals (SiNCs), 25-3 thru 25-4
  - silicon/SiO_2 serial sputtering, 25-6
  - transmission electron microscope (TEM), 25-5

Silicon quantum dot
- boundary conditions, 5-14
- charge density in the nanocrystal, 5-17
- DOF *vs.* maximum mesh size, 5-16
- mesh algorithm, 5-15
- meshing optimization, 5-16
- quarter-structure in FEM, 5-14
- structure for FEM simulation, 5-14
- total energy of system, 5-17
- *vs.* metallic approximation, 5-17 thru 5-18

Single electron memories, 2-7

Single-electron transistor (SET), 12-8 thru 12-9

Single-molecule magnet transistors, 9-11 thru 9-12

Single-photon Faraday rotation (SPFE)
- conditional Faraday rotation, 1-16
- GHZ quantum teleportation, 1-16 thru 1-18
- Jaynes–Cummings Hamiltonian description, 1-14 thru 1-15
- quantifying EPR entanglement, 1-16
- quantum computing
  - Alice's quantum dot, 1-19
  - Bob's quantum dot, 1-19
  - Leuenberger's scheme, 1-18
  - optical Stark effect, 1-20

Single-walled carbon nanotube (SWNT)
- 1D direct-gap band structure, 22-2
- exciton states, 22-3
- Lorentzian function, 22-3
- PL spectrum, 22-3 thru 22-4
- Si/SiO_2 substrate, 22-2
- spectral diffusion, 22-4
- Stark effect, 22-4

Small angle x-ray scattering (CD-SAXS), 19-27

Spacer double patterning method, 19-25

Spacer patterning technique (SPT), 17-4 thru 17-5

Spatial symmetry
- nanophotonic buffer memory, 33-11 thru 33-12
- quantum entangled states manipulation, 33-12 thru 33-13
- selective energy transfer, 33-9 thru 33-10
- symmetric and antisymmetric states, 33-8 thru 33-9
- AND and XOR logic gates, 33-10 thru 33-11

Spherical dot in plate capacitor
- geometry and notations, 5-3
- metallic dot
  - B_l coefficients expression, 5-4 thru 5-5
  - boundary conditions for potential, 5-4
  - capacitance and coupling coefficient, 5-3 thru 5-4
  - capacitance matricial expression, 5-6
  - method of images, 5-4 thru 5-5
  - numerical implementation and results, 5-7 thru 5-9
  - partial capacitance expression, 5-6
  - polarization influence, 5-6 thru 5-7
  - semi-analytical expression for potential, 5-6
- semiconductor dot
  - multiple electrons, 5-12 thru 5-13
  - quantization effects, 5-9 thru 5-10
  - quantum dot with single electron, 5-10 thru 5-11
  - second electron addition, 5-11 thru 5-12
- silicon quantum dot
  - boundary conditions, 5-14
  - charge density in the nanocrystal, 5-17
  - DOF *vs.* maximum mesh size, 5-16
  - mesh algorithm, 5-15
  - meshing optimization, 5-16
  - quarter-structure in FEM, 5-14
  - structure for FEM simulation, 5-14
  - total energy of system, 5-17
  - *vs.* metallic approximation, 5-17 thru 5-18

Spin-transfer torque effect, 8-3
- for both directions of current, 8-6
- and giant magnetoresistance, 8-12 thru 8-13
- LLGS equation, 8-6
- spin-diffusion length, 8-7

Spin transistor, 12-9

Spintronics, 1-1, 8-3, 12-9

Spin-valve system
- active region, 8-3
- nanoscale (*see* Nanoscale spin valves)

Sputter deposition, 6-6 thru 6-7

Static random access memory (SRAM) unit cell, 20-8 thru 20-9

Stratonovich interpretation, 8-11

Sub-*kT/q* switch, 12-8

Substrate radiation mode, 37-9

Superconducting Josephson junction
- Ambegaokar–Barato relation, 16-2
- concerns specific to small junctions, 16-4
- field emission micrograph, 16-2
- *I–V* curve, 16-2 thru 16-3
- RCSJ model, 16-3 thru 16-4

Superconducting quantum interferometers (SQUID), *see* Nanotube-based superconducting quantum interferometers

Superconducting weak links
- carbon nanotube weak link
  - experimental realizations, 7-6 thru 7-7
  - Fabry–Perot cavity, 7-6
  - high-frequency irradiation, 7-9 thru 7-10
  - nanotube quantum dot, 7-7 thru 7-9
- coherent transport, 7-2
- electronic state density, 7-2
- fullerene-based links, 7-17 thru 7-18
- gate control, 7-4 thru 7-5
- graphene-based links
  - proximity effect, 7-15 thru 7-17
  - SQUIDs devices, 7-17
- Josephson effect, 7-2 thru 7-3
- low-dimensional sp^2 carbon structures, 7-4 thru 7-5
- nanotube-based SQUID
  - magnetometry applications, 7-13 thru 7-14
  - quantum information applications, 7-14 thru 7-15
  - quantum interference with weak links, 7-10 thru 7-11
  - tunability of the phase shift, 7-11 thru 7-12
- normal electron transport, 7-3
- transport through Andreev bound states, 7-3 thru 7-4

Surface-architecture-controlled transistors
- depletion width and effective diameter, 14-14
- electrode CS, 14-12 thru 14-13
- I_{DS}–V_G curves, 14-12
- surface band bending, 14-13

Surface plasmon resonance (SPR)/Mie resonance, 24-2, 24-6

Surface-roughness effects, 14-7 thru 14-9

Sweep rate effect, 14-9 thru 14-11

T

Thermal NIL, 19-18 thru 19-19

Thermionic emission (TE) theory, 13-2

Thomas–Reiche–Kuhn sum rule, 25-13

Three-dimensional carrier confinement, 40-4, 40-5

Three-terminal molecules, 17-9 thru 17-10

Time-bandwidth product (TBWP), 41-2

Time-dependent Kohn–Sham equation, 27-3

Time-dependent local-density approximation (TD LDA), 24-7 thru 24-8, 25-15

TMOS, schematic behavior, 5-1 thru 5-2

Transfer printing, 19-19

Transistor structures
- atomic-scale TCAD, 12-11 thru 12-12
- carbon nanotube transistors, 12-6 thru 12-7
- graphene ribbon nanotransistors, 12-7 thru 12-8

- molecular tunnel junctions, 12-9 thru 12-11
- Moore's law, 12-1 thru 12-2
- nonclassical transistor structures, 12-7
- nonplanar multi-gate transistors, 12-3 thru 12-4
- point contacts and conductance quantum, 12-11
- quantum effects, 12-5 thru 12-6
- SET, 12-8 thru 12-9
- silicon processing, technology boosters, 12-2 thru 12-3
- spin transistor, 12-9
- sub-*kT/q* switch, 12-8

Transmission electron microscope-cathodoluminescence (TEM-CL)
- carrier generation and light emission process, 21-3 thru 21-4
- characteristics, 21-3
- diffusion length, 21-4
- gallium arsenide (GaAs) nanowire
  - CL spectra, temperature dependence, 21-20
  - GaAs/AlGaAs nanowire, SEM image, 21-18 thru 21-19, 21-21
  - zinc-blend structure, 21-18
- InP nanowires
  - broad peak blue-shifts, 21-16
  - CL spectra, 21-17 thru 21-18
  - degree of polarization, 21-18
  - electron confinement energy, 21-16
  - emission energy, 21-15
  - metal-organic vapor phase epitaxy (MOCVPE) technique, 21-15
  - phonon scattering process, 21-17
  - SEM image, 21-15 thru 21-16
- Monte Carlo simulation, 21-4
- photomultiplier tube (PMT), 21-5
- resolution factors, 21-3
- scanning electron microscopy (SEM), 21-5
- zinc oxide (ZnO) nanowires
  - CL intensity, 21-22
  - CL spectra, 21-21
  - emission peak, 21-24
  - fitting function, 21-22
  - optical properties, 21-21
  - polarization ratio, 21-23
  - smoke particles, 21-22 thru 21-23
  - wurtzite-type semiconductor, 21-21

Tunneling magnetoresistance (TMR), 6-18
Tunnel magnetoresistance (TMR) effect, 2-5
Two-level amplifying system (TLS)., 39-9

U

Ultrafast laser diodes
- broad gain bandwidth, 41-7
- dimensionality role, 41-5 thru 41-6
- low absorption saturation fluence, 41-8
- low threshold current and low temperature sensitivity, 41-8
- materials and growth, 41-6
- ultrafast carrier dynamics, 41-7 thru 41-8

Ultrahigh data storage, 39-12
UV NIL, 19-19

V

van den Berg's ferromagnetic islands, 4-3
van der Waals epitaxy
- bottom contact structure, 15-11
- initial growth surface
  - hexagonal AlN (0001) surface, 15-14 thru 15-16
  - hexagonal GaN (0001) surfaces, 15-12 thru 15-14
  - H-terminated Si (111) surfaces, 15-11 thru 15-12

van Roosbroeck thru Shockly relation, 21-6 thru 21-7
Vapor–liquid–solid (VLS) mechanism, 14-3
Vapor transport method
- FESEM images, vertically well-aligned nanowires, 14-5 thru 14-6
- furnace system schematic, 14-4 thru 14-5

Vertical-cavity surface-emitting lasers (VCSELs), 40-3

W

Wess–Zumino phase, *see* Berry's phase

X

X-ray diffraction (XRD)
- reciprocal space maps, 6-9
- specular reflectivity method, 6-9 thru 6-10
- synchrotron x-ray scattering measurements, 6-9

Z

Zeeman effects, 1-4
Zero-biased rf electrometry
- operation beyond 1 GHz, 16-14 thru 16-15
- quasiparticle poisoning, 16-15
- relation to the cooper-pair box, 16-16
- rf measurement setup, 16-13 thru 16-14

ZnO nanorod double quantum-well structures, 36-7 thru 36-11
ZnO nanowire field-effect transistors
- crystal structure of ZnO, 14-2
- fabrication and characterization
  - back gate and top gate configurations, 14-7
  - passivation and surface-roughness effects, 14-7 thru 14-9
  - sweep rate effect, 14-9 thru 14-11
- FET, 14-3 thru 14-4
- general growth methods, 14-2 thru 14-3
- N-channel depletion-mode and enhancement-mode, 14-11
- physical parameters of ZnO, 14-2
- surface-architecture-controlled transistors
  - depletion width and effective diameter, 14-14
  - electrode CS, 14-12 thru 14-13
  - I_{DS}–V_G curves, 14-12
  - surface band bending, 14-13
- vapor transport method
  - FESEM images, vertically well-aligned nanowires, 14-5 thru 14-6
  - furnace system schematic, 14-4 thru 14-5

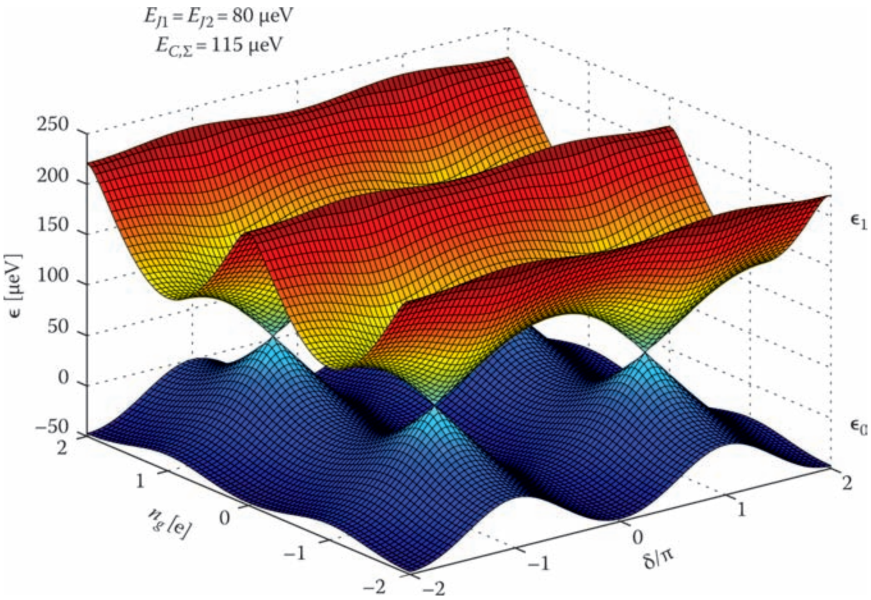

FIGURE 16.4 The ground (ϵ_0) and first excited (ϵ_1) states calculated by numerically diagonalizing Equation 16.11 with the energies indicated in the figure.

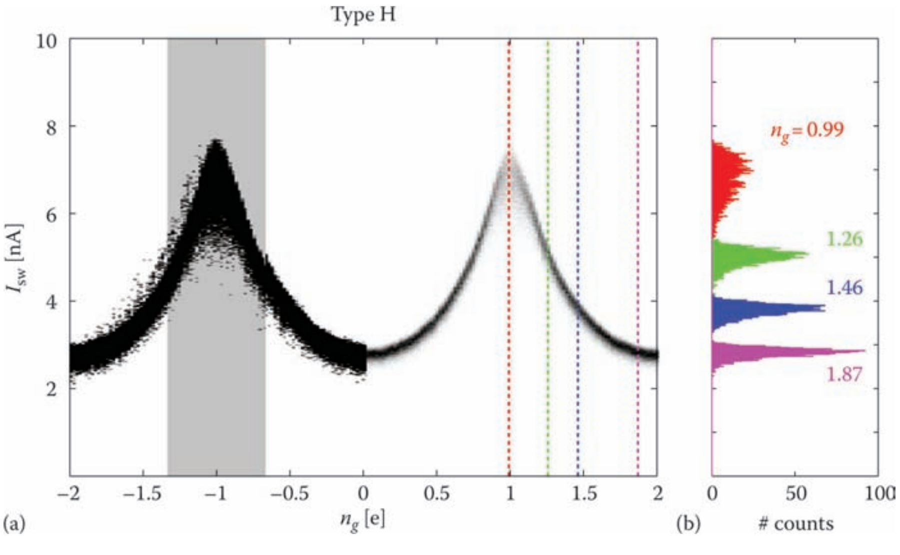

FIGURE 16.11 (a) Type H I_{sw} histograms versus n_g. Histogram height is displayed in grayscale on the right-hand side, whereas all counts are displayed equally on the left-hand side. As in Figure 16.7, the gray box in (a) denotes regions where the island potential is trap like for quasiparticles. (b) Selected histograms corresponding to several gate voltages. Device parameters: $\Delta_i = 246\,\mu\text{eV}$, $\Delta_\ell = 205\,\mu\text{eV}$, $E_C \simeq 115\,\mu\text{eV}$, and $E_{J1} = E_{J2} \simeq 82\,\mu\text{eV}$.

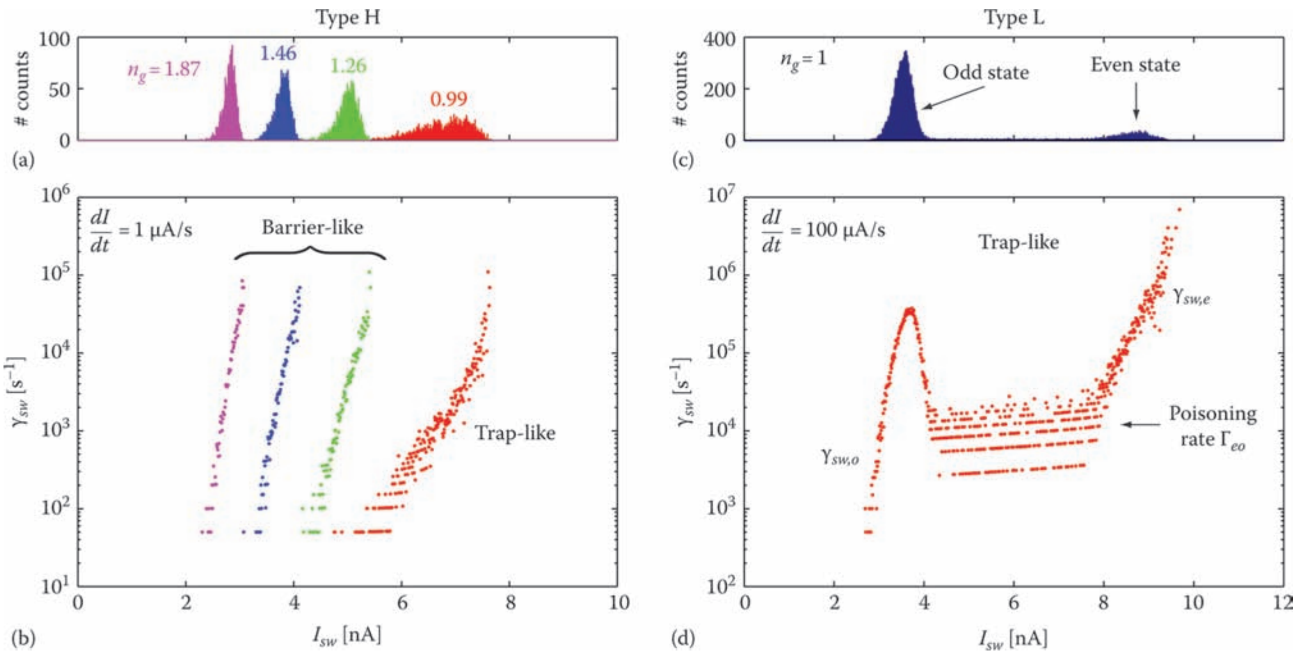

FIGURE 16.13 (a,c) Switching current (I_{sw}) histograms and (b,d) derived switching/escape rates for a type H (barrier-like) and type L (trap-like) CPTs. For the type L device, the quasiparticle trapping behavior is evident in the bimodal I_{sw} distribution. In this case, the poisoning rate Γ_{eo} can be read directly from the derived escape rate in (d) as shown. Although the type H device is barrier-like for most n_g, it still looks like a trap near $n_g = 1$ (see Figure 16.7). This is apparent in the "curvy" structure of the escape curve for $n_g = 0.99$ as compared with the escape rates at other n_g in (b).

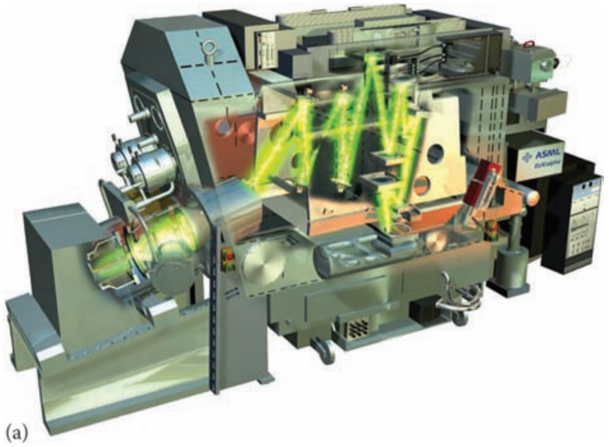

FIGURE 20.6 (a) Artist illustration of ASML's EUV ADT showing a DPP EUV source on the left, illumination optics in the middle, and mask, projection optics, and wafer on the right. (Courtesy of ASML, Veldhoven, the Netherlands.)

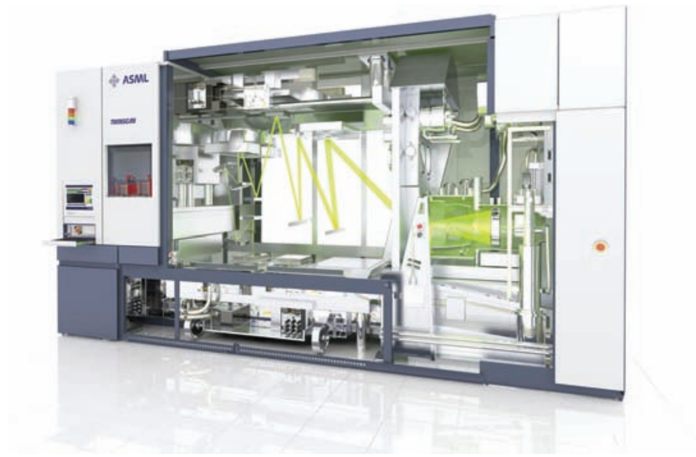

FIGURE 20.18 Artist illustration of ASML NXE: 3100 preproduction EUV exposure tool showing the LPP EUV source on the right, the illuminator optics in the center, and the mask and the projection optics on the right. (Courtesy of ASML, Veldhoven, the Netherlands.)

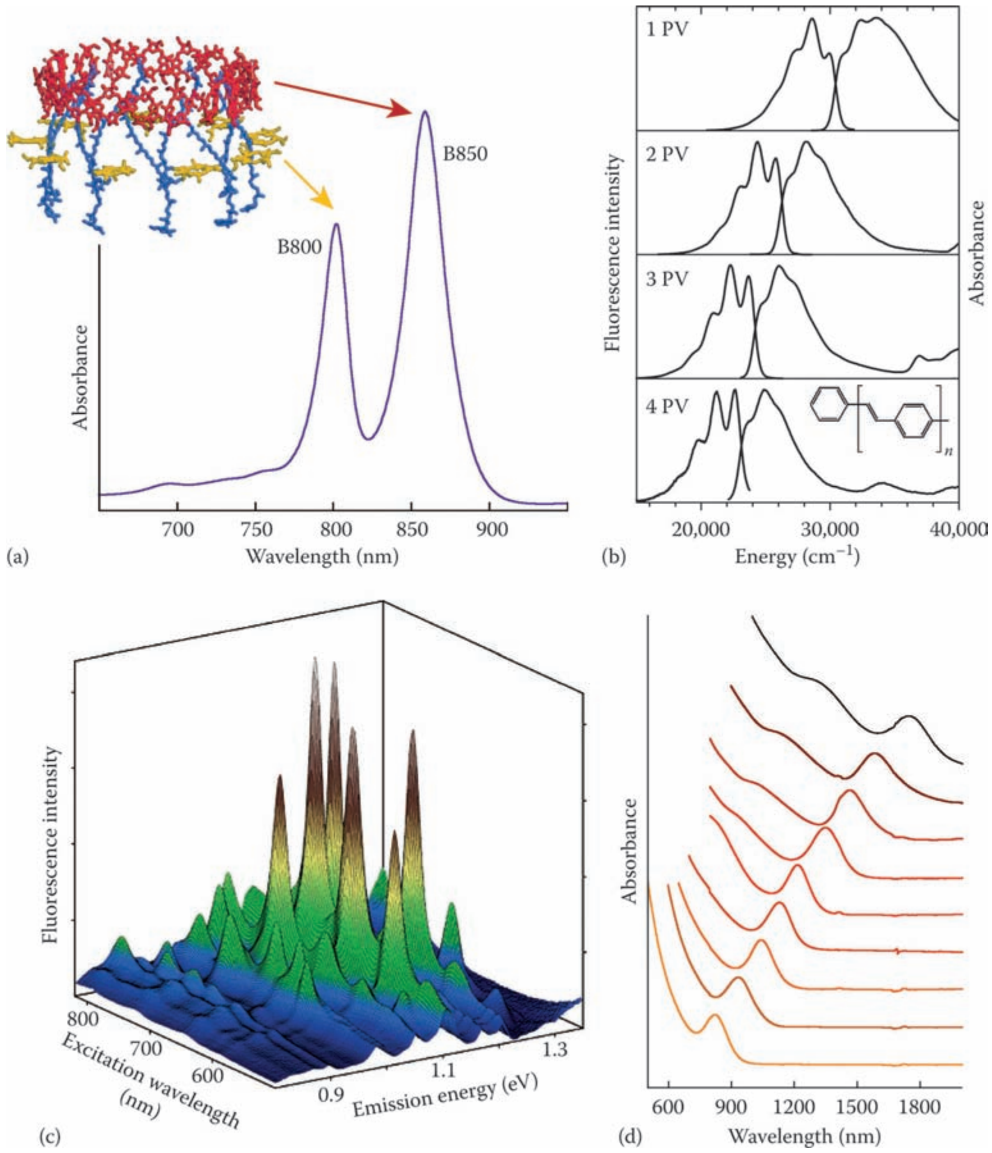

FIGURE 23.1 Excitons and structural size variations on the nanometer length scale. (a) The photosynthetic antenna of purple bacteria, LH2, is an example of a molecular exciton. The absorption spectrum clearly shows the dramatic distinction between the B800 absorption band, arising from essentially "monomeric" bacteriochlorophyll-a (Bchl) molecules, and the redshifted B850 band that is attributed to the optically bright lower exciton states of the 18 electronically coupled Bchl molecules. (b) The size-scaling of polyene properties, for example, oligophenylenevinylene oligomers, derives from the size-limited delocalization of the molecular orbitals. However, as the length of the chains increases, disorder in the chain conformation impacts the picture for exciton dynamics. Absorption and fluorescence spectra are shown as a function of the number of repeat units. (c) SWCNT size and "wrapping" determine the exciton energies. Samples contain many different kinds of tubes, therefore optical spectra are markedly inhomogeneously broadened. By scanning excitation wavelengths and recording a map of fluorescence spectra, the emission bands from various different CNTs can be discerned, as shown. (Courtesy of Dr. M. Jones). (d) Rather than thinking in terms of delocalizing the wavefunction of a semiconductor through interactions between the unit cells, the small size of the nanocrystal confines the exciton relative to the bulk. Size-dependent absorption spectra of PbS quantum dots are shown. (Adapted from Scholes, G. D. and Rumbles, G., *Nat. Mater.*, 5, 683, 2006.)

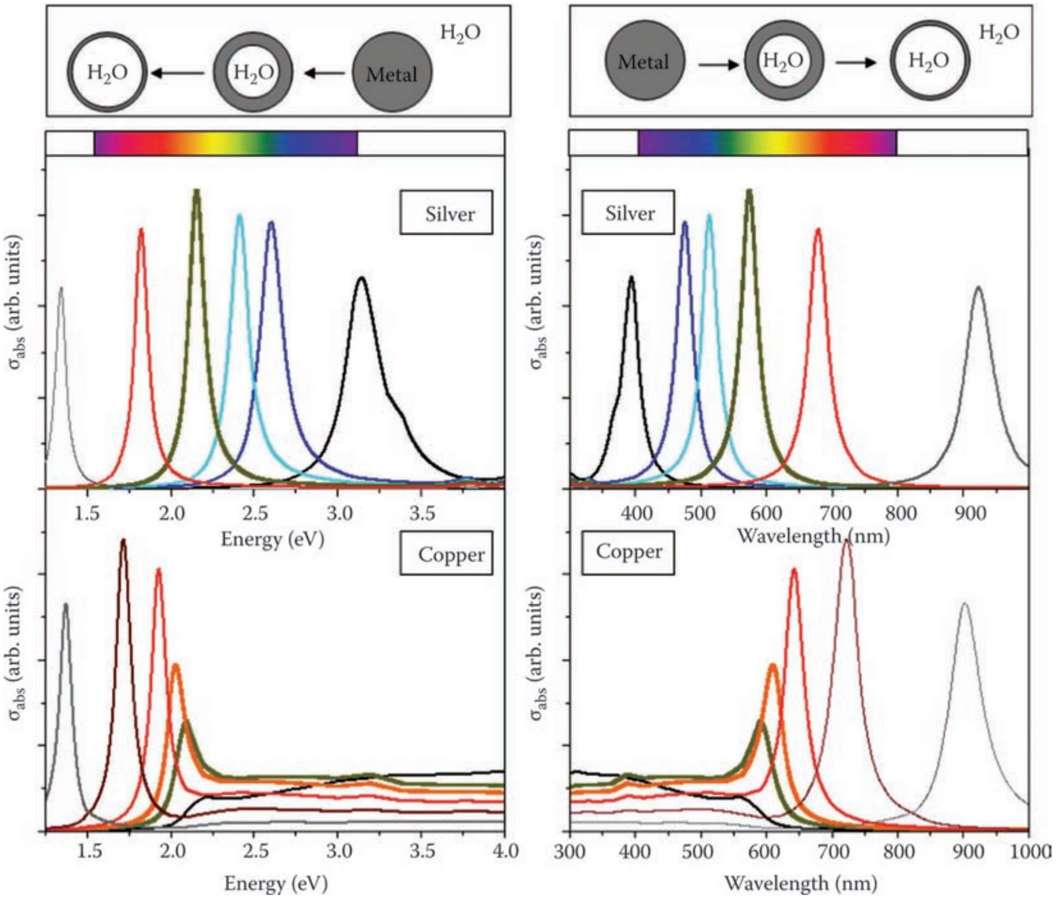

FIGURE 24.9 Evolution of the absorption cross section of a nanoshell of silver (top) or copper (bottom) in the dipolar approximation versus energy (left) or versus wavelength (right) for various thicknesses of the shell. The core is filled with water and the external medium is also water ($n = 1.33$). The dielectric functions of copper and silver have been extracted from Palik (1985–1991). The correspondence between the energy in eV and the wavelength in nm is E (eV) = 1239.85/λ (nm). The total radius of the cluster is always 15 nm and the thickness takes the following values: $e = R - R_c = 5; 4; 3; 2; 1$ nm corresponding to ratios between the shell thickness and the total cluster radius: $(e/R) = 0.33; 0.27; 0.2; 0.13; 0.07$. The spectra in black correspond to the fully homogeneous cluster.

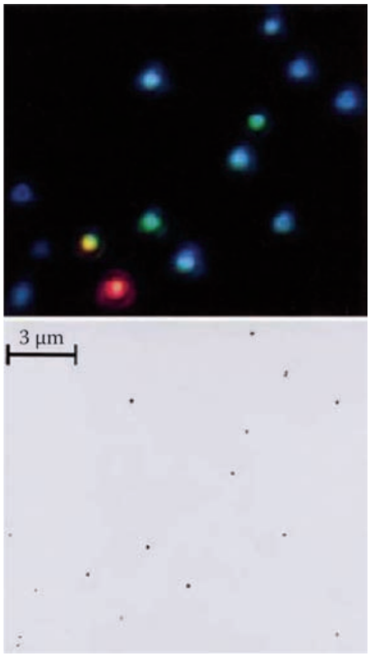

FIGURE 24.15 (Top) Color image of a typical sample of silver nanoparticles as viewed under the dark-field microscope. The brightness of the particles increases from blue to red due to both the intrinsic optical scattering cross section and the spectral output of the light source (the red particle is overexposed). This image is correlated to its electron microscopy image (Bottom). (Reprinted from Mock, J.J. et al., *J. Chem. Phys.*, 116, 6755, 2002. With permission.)

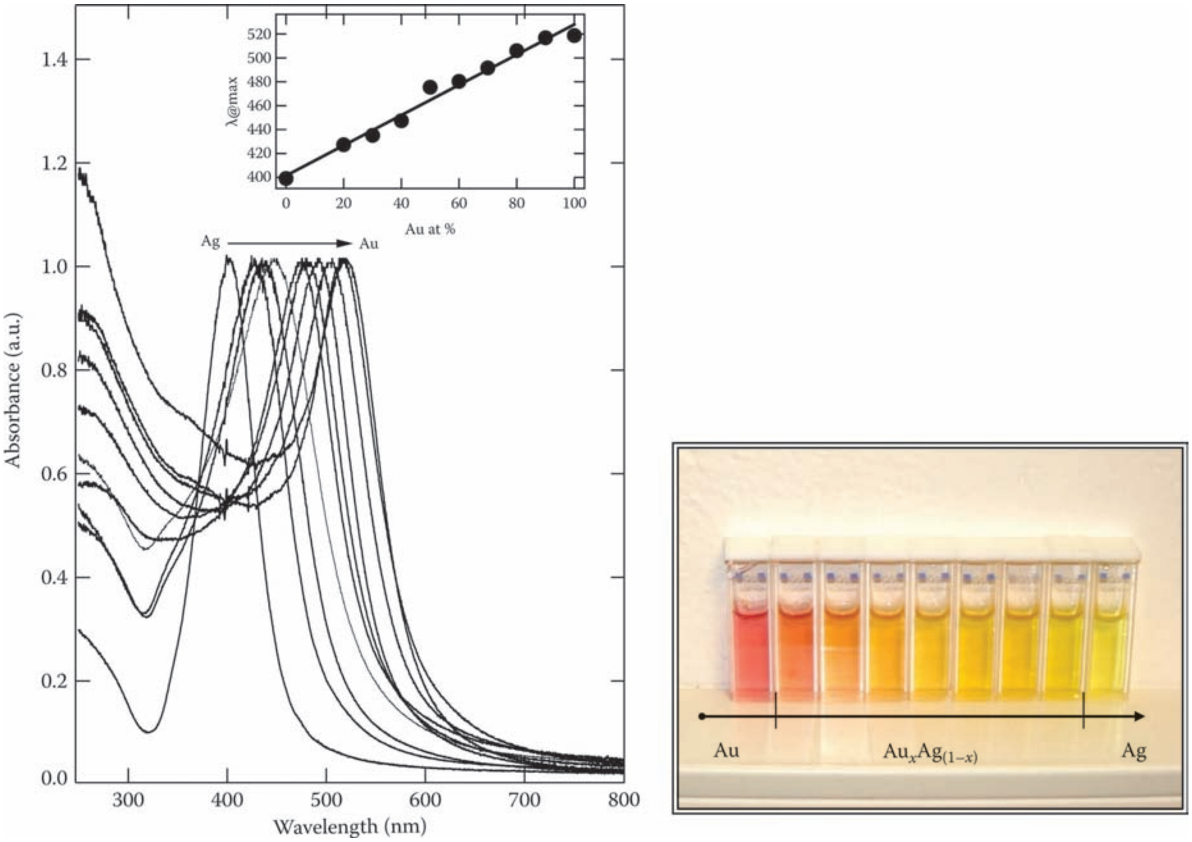

FIGURE 24.16 (Left) Normalized UV–visible spectra of Au–Ag alloy nanoparticles with varying composition. (Insert) Location of the SPR maximum as a function of the gold content. (Right) Corresponding solutions whose colors vary from the red (pure gold nanoparticles) to the yellow (pure silver nanoparticle). (From Russier-Antoine, I. et al., *Phys. Rev. B*, 78, 35436, 2008. With permission.)

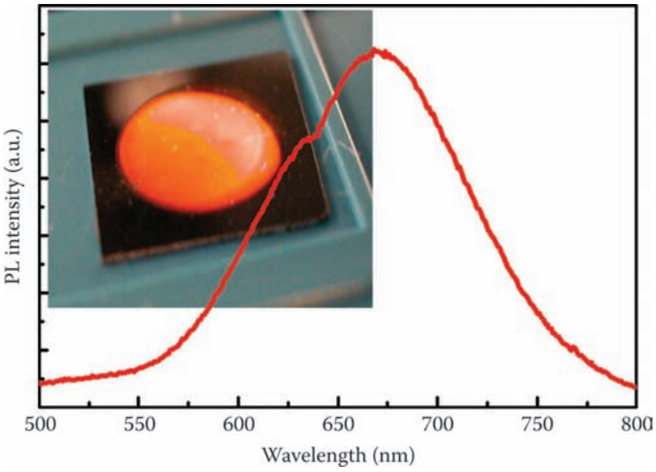

FIGURE 25.2 The PL spectrum from porous silicon with a maximum PL at a wavelength of about 675 nm. The photograph at the inset demonstrates the red color of the emitted PL from a circular layer of porous silicon (the sample has been illuminated with a UV lamp).

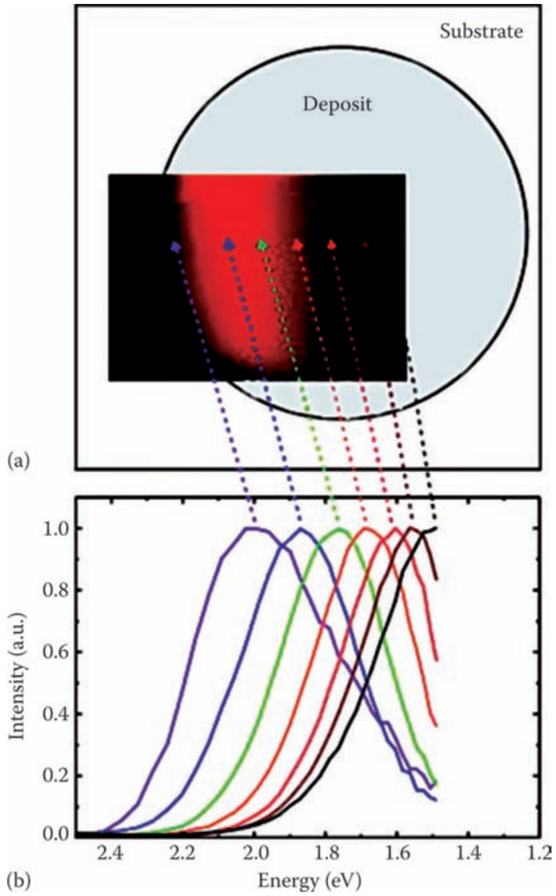

FIGURE 25.6 (a) A photograph showing the PL variation along the substrate from silicon nanocrystals deposited by the laser pyrolysis technique. (b) The normalized PL spectra from different positions along the substrate. (Reprinted from Ledoux, G. et al., *Appl. Phys. Lett.*, 80, 4834, 2002. With permission.)

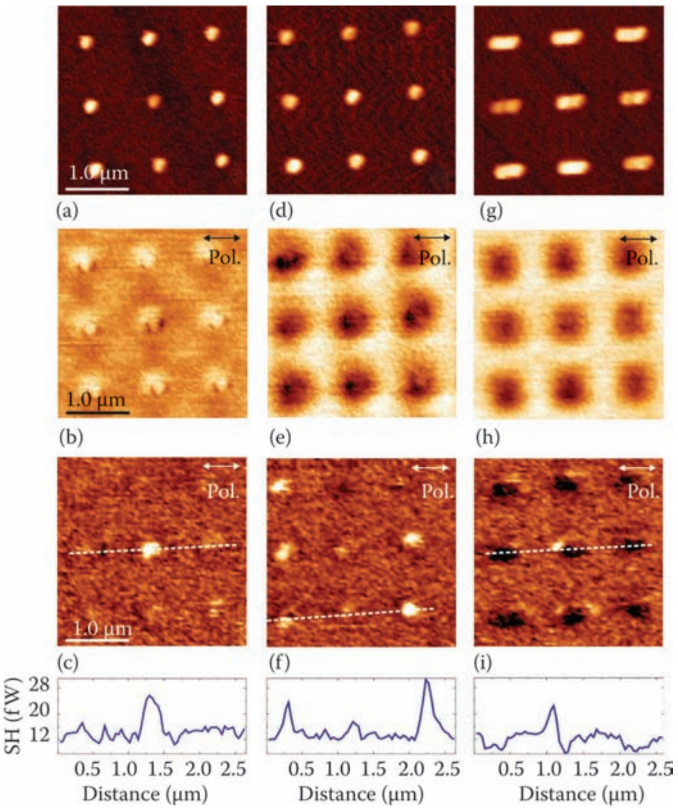

FIGURE 28.10 Nanoparticles: (a), (d), and (g) topography; (b), (e), and (h) FW transmission; and (c), (f), and (i) SH emission SNOM images with corresponding cross sections along the dashed lines, from the raw data. Incident light is polarized parallel to the major axis. The particle major axis lengths are 100 nm (a)–(c), 150 nm (d)–(f), and 400 nm (g)–(i). Image size: 3 × 3 µm^2. (Reprinted from Zavelani-Rossi, M. et al., *Appl. Phys. Lett.*, 92, 093119, 2008. With permission.)

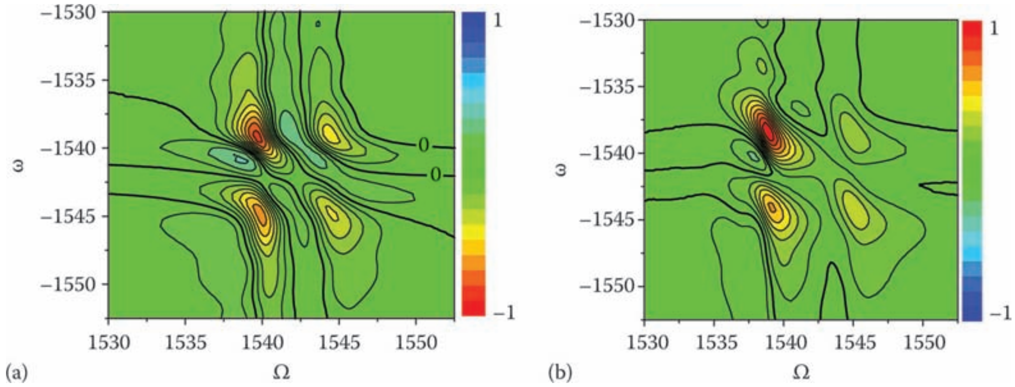

FIGURE 29.3 The imaginary part of the two-dimensional Fourier spectrum, $P(\Omega, \omega)$. (a) The spectrum calculated using Equation 29.92. (b) The experimental results of Zhang et al. (2005).

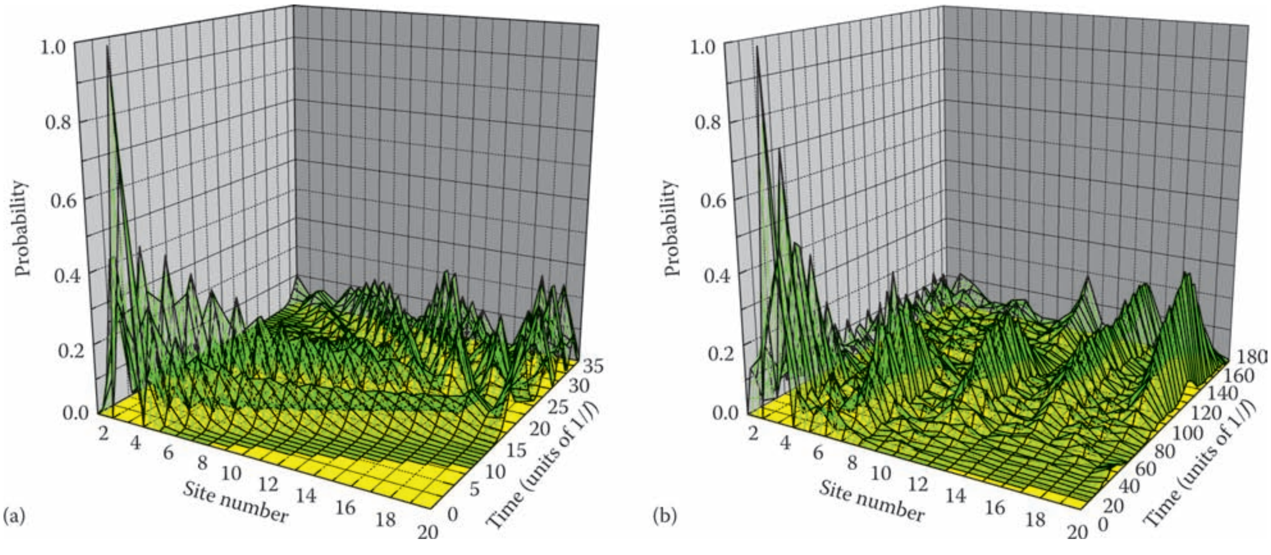

FIGURE 32.5 Probability that a photon is found at each site as a function of time (a) without the photon–phonon coupling, and (b) with the photon–phonon coupling comparable to the photon hopping constant.

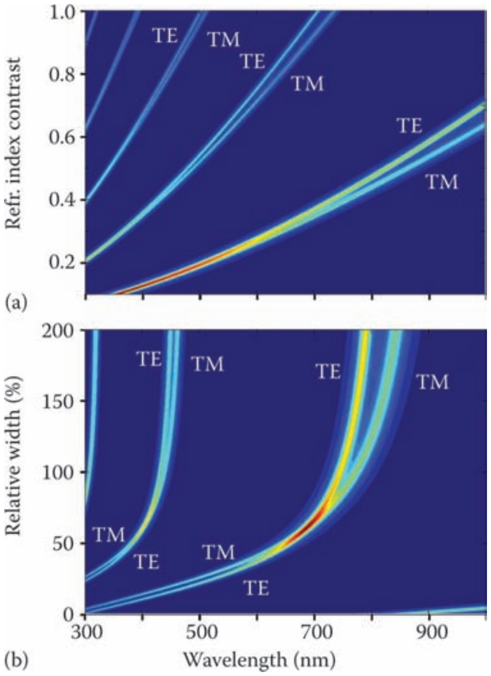

FIGURE 37.14 Calculated spectral positions of the substrate modes as a function of (a) the refractive index contrast $\Delta n = n_2 - n_3$ and (b) the relative thickness of the waveguide core compared to the sample 5×10^{17} Si cm^{-2}. Gray scale indicates intensity increasing from black up to white for the highest intensity. Several orders of modes are seen starting from the first one in infrared region.

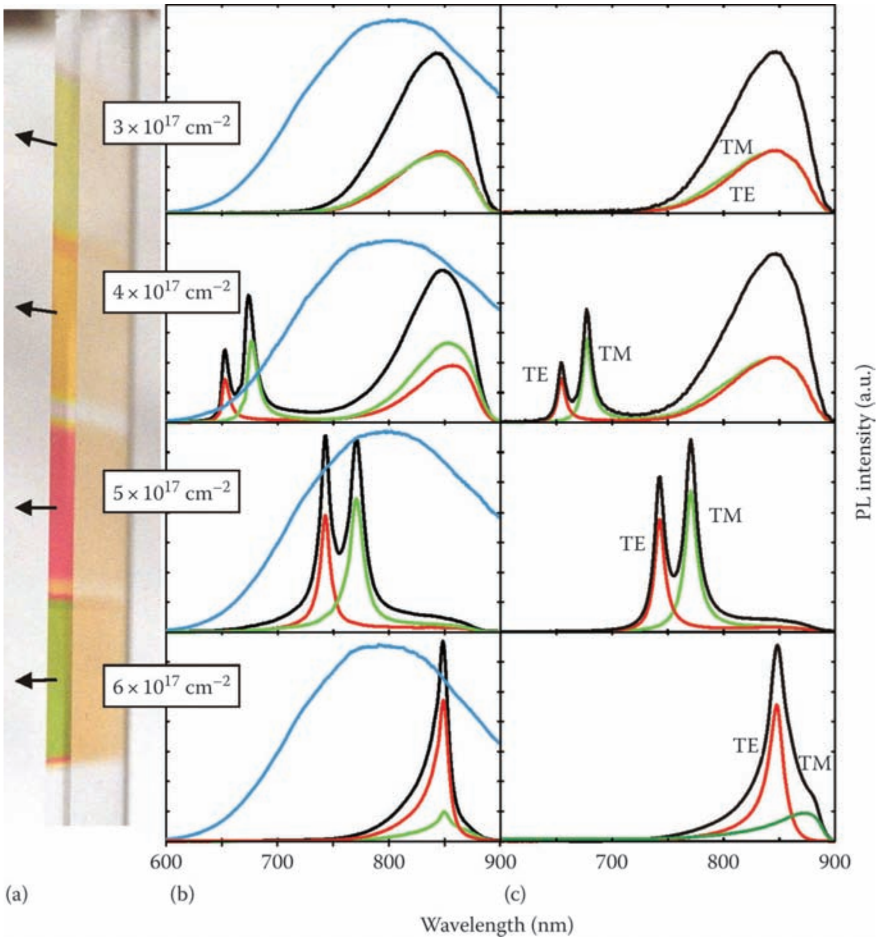

FIGURE 37.16 (a) Photograph of the edge of a set of Si$^+$ ion implanted layers with direction of PL indicated by arrows, the edge is on the left. (b) Measured PL from samples implanted to different Si ion fluences in standard (the broadest curves) and waveguiding geometry (black lines, the slightly lighter gray lines stand for TE and TM resolved polarizations). (c) Theoretically calculated PL spectra. We note that these results were obtained on different set of samples than in Figure 37.10. The mode positions are not exactly the same for samples with identical implantation dose because the annealing conditions were slightly different. Therefore, the refractive index profiles are not identical. (Adapted from Pelant, I. et al., *Appl. Phys. B*, 83, 87, 2006.)

Physics

HANDBOOK OF NANOPHYSICS

Nanoelectronics and Nanophotonics

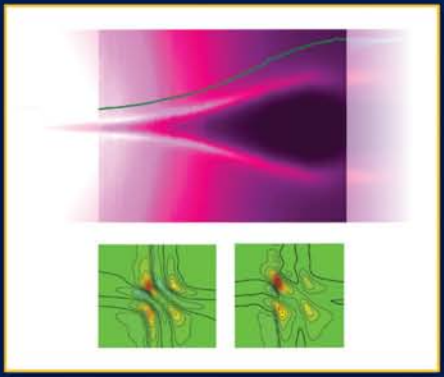

Many bottom-up and top-down techniques for nanomaterial and nanostructure generation have enabled the development of applications in nanoelectronics and nanophotonics. **Handbook of Nanophysics: Nanoelectronics and Nanophotonics** explores important recent applications of nanophysics in the areas of electronics and photonics. Each peer-reviewed chapter contains a broad-based introduction and enhances understanding of the state-of-the-art scientific content through fundamental equations and illustrations, some in color.

This volume discusses how different nanomaterials, such as quantum dots and nanotubes, are used in quantum computing, capacitors, and transistors. Leading international experts review the potential of the novel patterning techniques in molecular electronics as well as nanolithography approaches for producing semiconductor circuits. They also describe optical properties of nanostructures, nanowires, nanorods, and clusters, including cathodoluminescence, photoluminescence, and polarization sensitivity. In addition, the book covers nanophotonic devices and nanolasers.

Nanophysics brings together multiple disciplines to determine the structural, electronic, optical, and thermal behavior of nanomaterials; electrical and thermal conductivity; the forces between nanoscale objects; and the transition between classical and quantum behavior. Facilitating communication across many disciplines, this landmark publication encourages scientists with disparate interests to collaborate on interdisciplinary projects and incorporate the theory and methodology of other areas into their work.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

75500
ISBN: 978-1-4200-7550-2
90000
9 781420 075502
www.crcpress.com

75500_COVER_vol_6_final.indd 1

www.PositiveOffset.com

Electronic Kit & Parts for Educators, Researchers & Hobbyists

30 kHz / 1 kHz Zapper Kit
with Color Illustrated Instructions
$50.00 Add to Cart

Our Zapper Kit contains all the parts necessary to build either a 30,000 Hz or a 1,000 Hz Positive Offset Square Wave Frequency Generator Unit. This breadboard kit is based on Dr. Hulda Clark's zapper circuit design that was recently updated and published in 2003. Copper Pipes and 9 Volt Battery are not included in kit. For copper pipes and individual electronic parts, click the links in the right column.

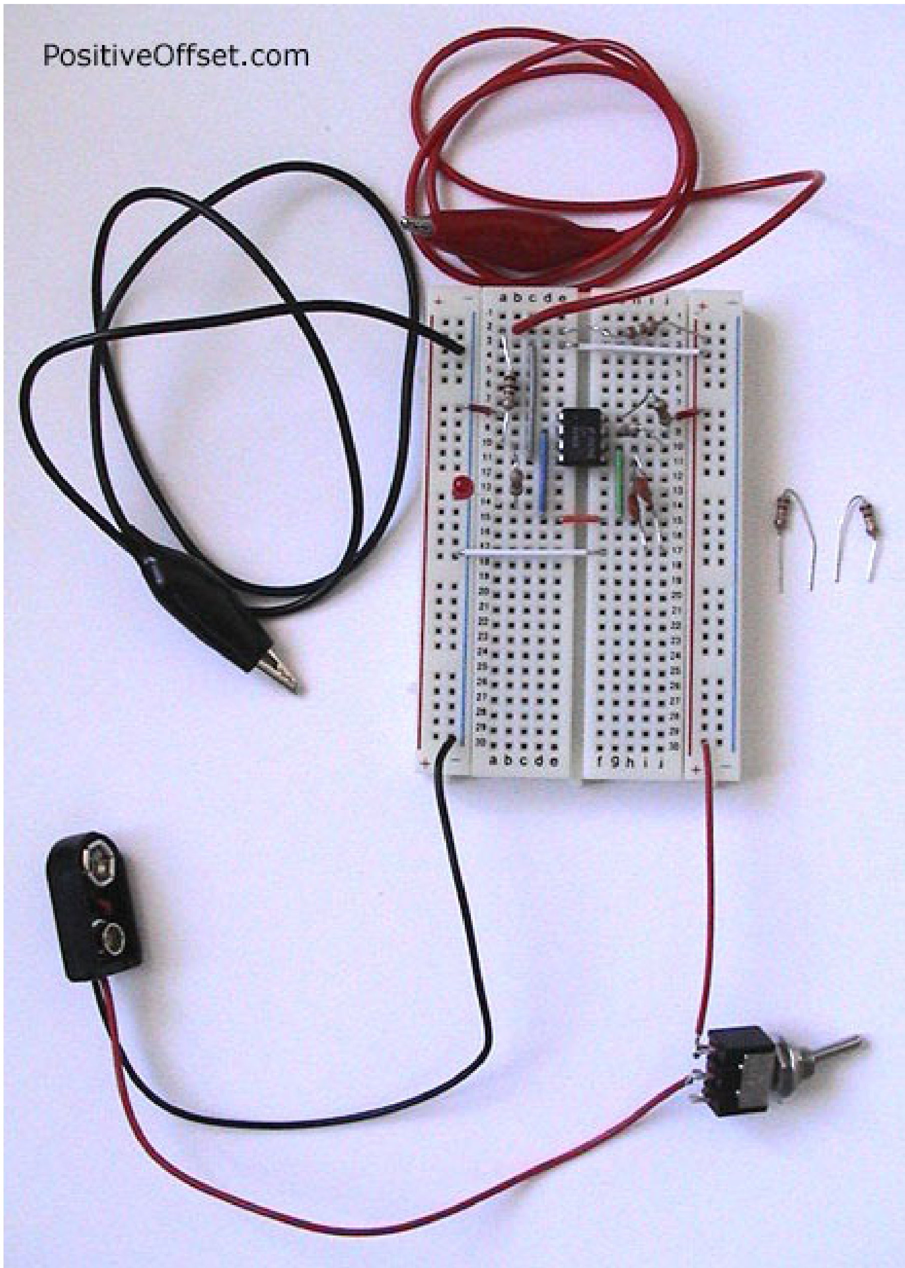

Hulda Clark Zapper Kit with Color Illustrated Instructions
$50.00 Add to Cart

Zapper Kit includes: Breadboard, 555 CMOS Timer, (2) 1K Ohm resistors, 2.2K Ohm resistor, 4.7K Ohm resistor, 39K Ohm resistor, (2) 270K Ohm

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- .0047 uF Capacitors
- 1 uH Inductors
- 1K Ohm Resistors
- 2.2K Ohm Resistors
- 3.9K Ohm Resistors
- 4.7K Ohm Resistors
- 39K Ohm Resistors
- 270K Ohm Resistors
- 555 CMOS Timer
- 8 Pin Wire Wrap Socket
- 9 Volt Battery Snap
- Alligator Clip Converter
- Alligator Leads
- Aluminum Square
- Banana to Alligator Leads
- Breadboard
- Color Coded Jump Wires
- Copper Pipes
- Nuts and Bolts
- On/Off Toggle Switch
- Red Mini LED
- Switching diode 1N914
- Video Instructions

Dual Frequency
1KHz 30KHz
www.FrequencyGeneration.com

www.PositiveOffset.com

VIEW CART

resistors, (2) .0047 uF Capacitors, Mini Red LED, (9) Color Coded Jump Wires, 9-Volt Battery Snap, (2) Alligator leads, and an On/Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit.

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

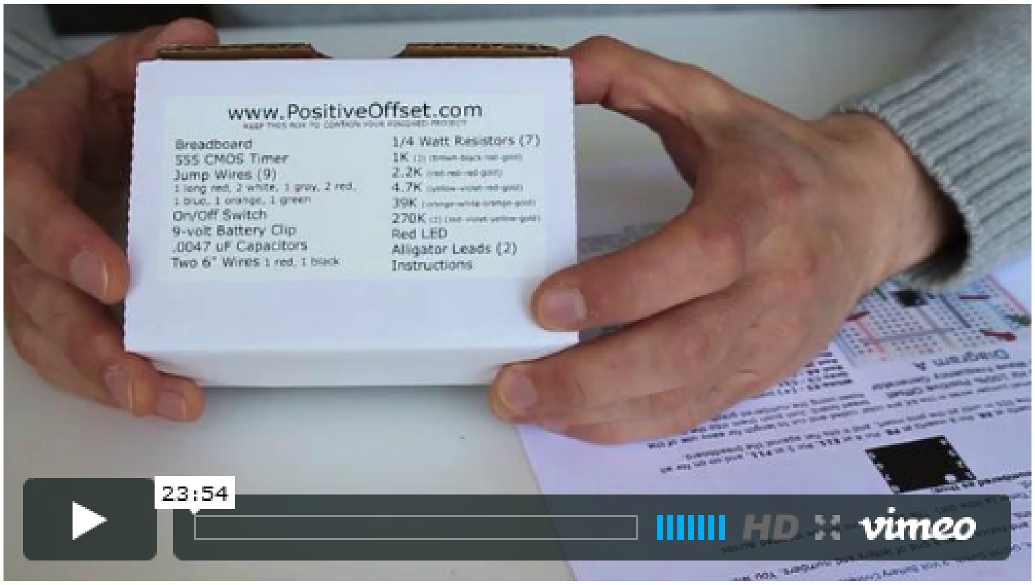
www.PositiveOffset.com
Breadboard
555 CMOS Timer
Jump Wires (9)
1 long red, 2 white, 1 gray, 2 red,
1 blue, 1 orange, 1 green
On/Off Switch
9-volt Battery Clip
.0047 uF Capacitors
Two 6" Wires 1 red, 1 black
1/4 Watt Resistors (7)
1K
2.2K
4.7K
39K
270K
Red LED
Alligator Leads (2)
Instructions
23:54
HD
vimeo

- [YouTube video instructions: 9 easy steps to build this circuit]
- [Build a zapper Frequently Asked Questions and Answers]
- Already have electronic parts? [Download PDF Instructions]

Remember the Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work. If you experience difficulty with your build, or have little experience in electronics, please order replacement parts along with your kit.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement.

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working.

If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

Natural Health Supply .com
Celebrating 17 Years 2001-2018

www.PositiveOffset.com

Electronic Kit & Parts for Educators, Researchers & Hobbyists

Hulda Clark 30 kHz / 1 kHz Zapper Kit
with Color Illustrated Instructions
Order Zapper Kit Here

Hulda Clark Zapper Kit includes: Breadboard, 555 CMOS Timer, 7 resistors: (2) 1K Ohm, 2.2K Ohm, 4.7K Ohm, 39K, (2) 270K Ohm, (2) .0047 uF Capacitors, Red LED, (8) Color Jump Wires, (2) Alligator Leads, and Soldered Battery Snap/On-Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit. Kit does not include copper pipes. Click here to order copper pipes.

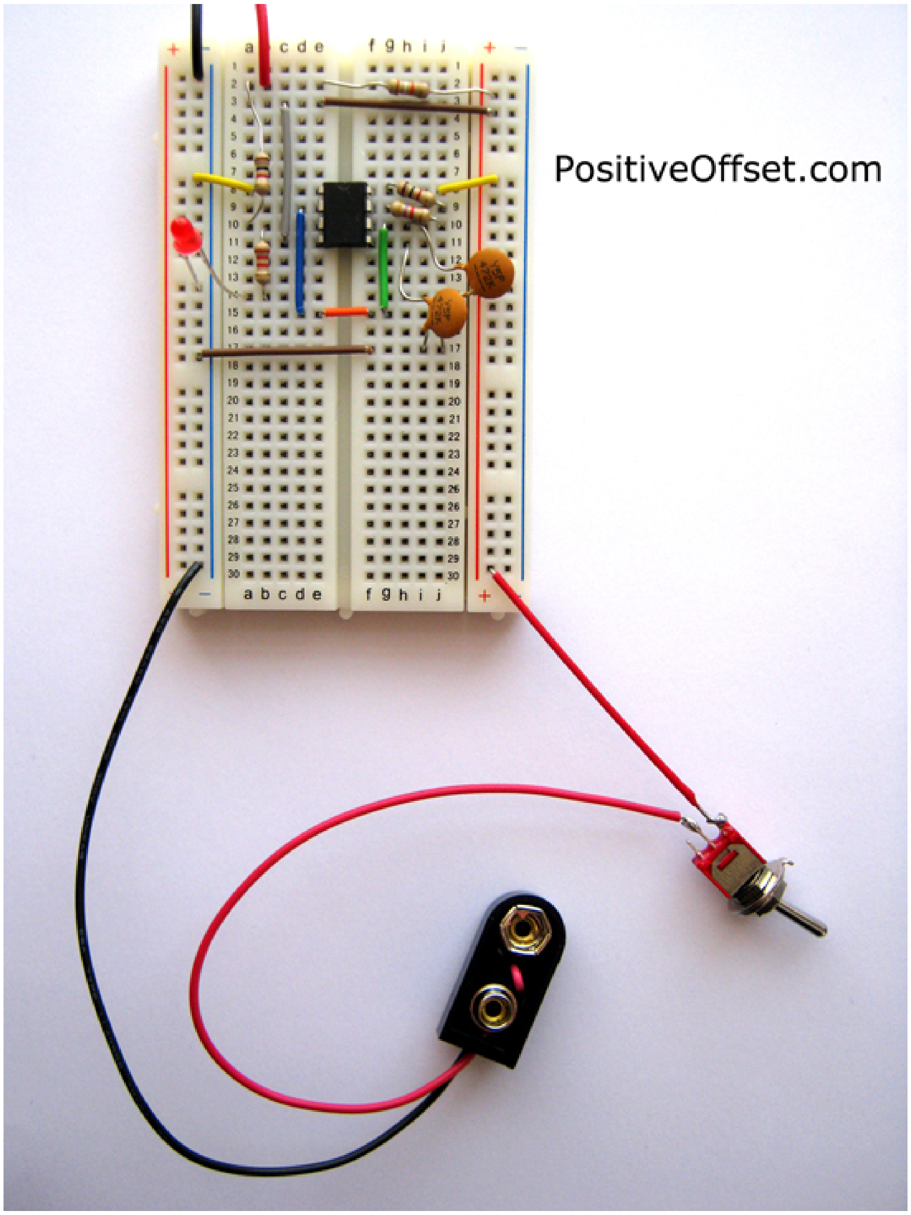

Hulda Clark Zapper Kit with Color Illustrated Instructions
Order Zapper Kit Here

If you are new to building electronics, you may also want to watch the instruction video. The 'Build A Zapper Video' shows how to build this Clark Zapper breadboard kit. Watch the video below. (Note: There are minor updates to the "Build Your Own Zapper" Kit. We have adjusted the jump wires for a

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- 555 CMOS Timer
- Switching diode 1N914
- Mini Alligator Leads
- Copper Pipes
- Dial-A-Zap
- Chair Zappicator
- Pet Zappicator
- Video Instructions

www.ClarkZapper.com

VIEW CART

better fit. We have also pre-soldered the 9-volt battery snap to the on/off switch. These updates are not shown in this video, but are clearly described in the kit instructions.)

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

More Video Instructions

- Build a Zapper, Plate Zapper, Food Zappicator, and Toothbrush Zappicator
- Zapper Kit Frequently Asked Questions and Answers
- Already have electronic parts? Download PDF Instructions

The Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement. If you ever touch the 9 volt battery backwards to the battery snap, it may burn out the 555. You may want to protect your 555 by adding a 1N914 Switching Diode to your kit (diode sold separately).

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working. If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

www.ZapperLab.com

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

Welcome to the Zapper Lab. Here we have the step-by-step instructions for building your very own Hulda Clark Zapper, reprinted with permission from the books by Dr. Hulda Clark. Dr. Clark has improved her original zapper circuit and published this update in 2003. Knowing her readers have varying experience with building electronics, this recent update released several different sets of instructions on how to build a zapper, depending on one's electronic expertise.

- The original published Zapper instructions require soldering or twisting the parts together. It is best when soldered, which does require some special tools and skill. These instructions are found reprinted below from Dr. Clark's 2003 update.
- A newer, easier to build design employs the use of an electronic *breadboard*. A breadboard does not require soldering, as one can push the electronic components into place onto the breadboard to easily build a circuit. Dr. Clark published designs for the 30 kHz Zapper, and the 1000 Hz Zappicator to be built on a breadboard. The instructions for building both breadboard zapper circuits are at ClarkZapper.net along with a nifty on-line video to help guide you through the process.
- And last, for advanced electricians, Dr. Hulda Clark published a schematic for building the updated Zapper (see photo below).

If you do not know anyone who could put one together for you, you can order a pre-built Zapper at the web site, www.ClarkZapper.com. Remember, the Hulda Clark Zapper is not a licensed medical device. It has not been tested by the FDA or AMA for use on humans or animals. The Clark Zapper can be sold and used only as an experimental device, and no claim can be made as to a diagnosis, cure or treatment for any medical condition or disease. Dr. Clark's book states to not experiment with a zapper if you are either pregnant or wearing an electronic pacemaker. Please be sure to read one of Dr. Clark's books before using a Hulda Clark Zapper.

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

How To Build A Hulda Clark Zapper

click here to print the 2003 schematic for building a zapper

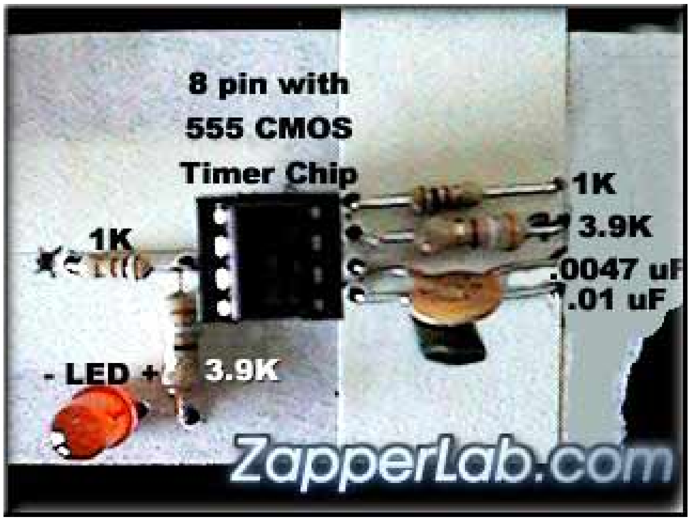

Order A Hulda Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. You can get all the parts you need at online electrical supply stores such as: PositiveOffset.com. If you would rather order a pre-built zapper, just click: Clark Zapper.

Dr. Clark has released new devices called the *Food Zappicator* and the *Toothbrush Zappicator.* For the instructions on how to build this new type of zapper, visit: FoodZappicator.com and ToothZappicator.com

NEW! Build A Hulda Clark Zapper Video Instructions
Nine short videos that take you step-by-step through the process of building a Hulda Clark Zapper.

How To Build A Hulda Clark Zapper

***The Cure For All Cancers* © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.**
Zapper instructions have been updated for 2003 by Dr. Hulda Clark

Hints for absolute novices: Don't let unusual vocabulary deter you. A "lead" is just a piece of wire used to make connections. When you remove a component from its package, label it with a piece of tape. A serrated kitchen knife works best, as does a large safety pin. Practice using the micro clips. If the metal ends are L-shaped bend them into a U with the long-nose pliers so they grab better. Chips and chip holders (wire wrap sockets) are very fragile. It is wise to purchase an extra one of each in case you break the connections. The "555" timer is a widely used component; if you can't locate this one, try another electronics shop.

The Shoebox Way

This circuit has been improved since the one given in earlier books. A resistor has been added that gives every pulse an added Positive Offset of 1/4 volt. You no longer need to operate your zapper so daringly close to a Negative voltage.

Get from Radio Shack, (serial number included to make it easy). Order these parts from their web site at: www.RadioShack.com. If any electrical parts are unavailable at Radio Shack, visit another on-line electrical supplier of these items, such as: www.PositiveOffset.com to buy your electrical parts.

Parts List for the 30,000 Hz Zapper Circuit	
Item	**Radio Shack Catalog Number**
Use Shoebox or Buy a Plastic Project Box	270-1809
9 volt battery	
9 volt battery snap connector	270-324 (set of 5 you need 1)
on-off toggle switch	275-624A micro mini toggle switch
If not available, choose any toggle switch with holes in the contact points or Radio Shack 275-612	
1 K Ohm resistor, brown-black-red-gold	271-312 (500 piece assortment) use 2
3.9 K Ohm resistor, orange-white-red-gold	Use 2 from 500 piece assortment
39 K Ohm resistor, orange-white-orange-gold	From 500 piece assortment
low-current red LED	276-044
.0047 uF capacitor	272-130 (set of 2, you need 1)
.01 uF capacitor	272-131 (set of 2, you need 1)
555 CMOS timer chip (TLC 555)	276-1718 (you may wish to buy a spare)
8 pin wire-wrapping socket for the chip	900-7242
If only 16 pin sockets are available, cut one in half OR leave half empty. Editor's Note: These parts have been discontinued at most Radio Shacks. 8 pin wire wrap sockets are available at many other electrical supply stores such as: www.PositiveOffset.com	
short (12") alligator clip leads	any electronics shop, get 10
If not available, use 14" length from Radio Shack, 278-1156	
Micro clip jumper wires	278-017 (you need 2 packages of 2)
If not available, use mini-clip jumper wires 278-016	
2 bolts, about 1/8" diameter, 2" long, with 4 nuts and 4 washers	hardware store
2 copper pipes, 3/4" diameter, 4" long	hardware store
Sharp knife, pin, long-nose pliers, tape, 4 twist ties or rubber bands.	

Editor's Note:

The 8 pin wire-wrapping socket has been discontinued at many Radio Shacks and readers have asked what to do. We know a great electrical supply store where 8 pin wire wrap sockets are available, as well as easy-to-build science kits for making a 30,000 Hz Frequency Generator with all the parts listed above in one simple package.

A Hulda Clark zapper is basically just a common Frequency Generator set at 30,000 Hz, with a 100% positive offset square wave. The cost for the electrical parts to build one is about $30-$40 for them all. Visit the electrical suppliers at PositiveOffset.com for all your electrical needs. They have everything you'd need to build a basic 30,000 Hz Positive Offset Square Wave Frequency Generator, and their kit is much easier to build than the instructions given below, because it uses a breadboard so there is no need to twist or solder wires, you simply push the electrical parts into place on the breadboard. **-editor.**

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1	8
2	7
3	6
4	5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the

pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Do not attach the battery yet.

10. An inch away from the switch pierce two holes 1/4 inch apart. Mount the 39 K Ohm resistor on the outside and label it inside and outside as "39 K, Positive offset." Flatten the wires on the inside.

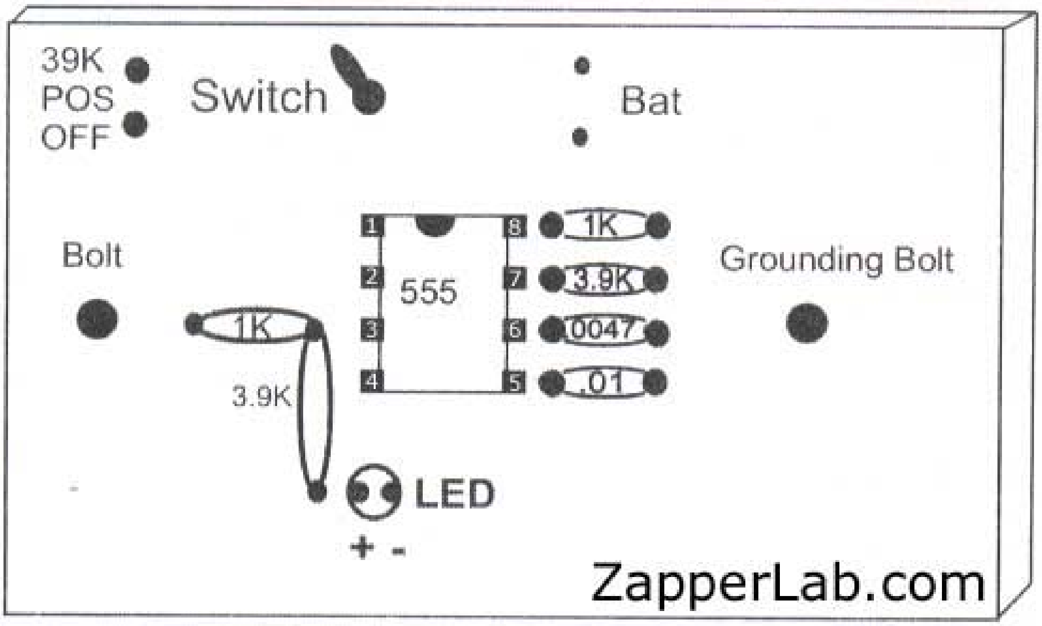

Now To Connect Everything

1. Twist free ends of the two capacitors .01 and .0047 together. Connect this to the Grounding Bolt using an alligator clip.

2. Bend the top ends of pin 2 and pin 6 (which already has a connection) inward towards each other in an L shape. Catch them both with an alligator clip and attach the other end of the alligator clip to the free end of the 3.9 K Ohm resistor by pin 7.

3. Using an alligator clip connect pin 7 to the free end of the 1 K Ohm resistor attached to pin 8.

4. Using three micro clips connect pin 8 to one end of the switch, pin 4 to the same end of the switch, and one end of the offset resistor to the same end of the switch. (Put one hook inside the hole and the other hooks around the whole connection. Check to make sure they are securely connected.) Connect the free end of the offset resistor to the bolt using an alligator clip.

5. Use an alligator clip to connect the free end of the 1 K Ohm resistor (by pin 3) to the bolt. It is the **output resistor**.

6. Twist the free end of the 3.9 K Ohm resistor by pin 3 around the plus end of the LED. Connect the minus end of the LED to the grounding bolt using an alligator clip.

7. Connect pin number 1 on the chip to the grounding bolt with an alligator clip.

8. Attach an alligator clip to the outside of one of the bolts. Attach the other end to a handhold (copper pipe). Do the same for the other bolt and handhold.

9. Connect the minus end of the battery snap connector (black wire) to the grounding bolt with an alligator clip.

10. Connect the plus end of the battery snap connector (red wire) to the free end of the switch using a micro clip lead. Attach the battery very carefully. Before attaching the battery to its snap connector, cover one terminal with tape. After snapping in one terminal, remove the tape to attach the other terminal. This is to prevent accidental touching of terminals in a backwards direction. If the LED lights up you know the switch is ON. If it does not, flip the switch and see if the LED lights. Label the switch clearly. If you cannot get the LED to light in either switch position, double-check all of your connections, and make sure you have a fresh battery. Even if it does light up, check every connection again.

11. Finally tie up the bunches of wire with twist-ties or rubber bands

and replace the lid on the box. Slip a couple of rubber bands around the box to keep it securely shut. For safer storage, place it inside a larger box.

Note: Having gained this much experience, you may prefer to build your next zapper on a piece of cardboard folded and able to fit inside a shoebox or plastic electronic project box for more protection.

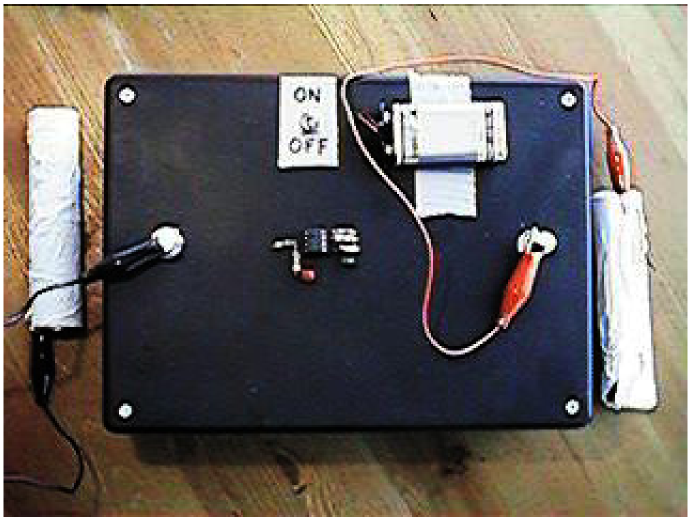

- Optional: measure the frequency of your zapper by connecting an oscilloscope or frequency counter to the handholds. Any electronics shop can do this. It should read between 20 and 40 kHz. The shop can also read the voltage (peak to peak) and the amount of Positive offset (on the .5 volt-per-division scale). The voltage output should be about 8 volts.
- Note: a voltage meter will only read 4 to 5 volts because it displays an average voltage.

Using The Zapper

1. Wrap handholds in one layer of wet paper towel before using. More will reduce the current. Grasp securely and turn the switch on to zap. Keep a bottle of water handy to keep them wet.

2. Zap for seven minutes, let go of the handholds, turn off the zapper, and rest for 20 minutes. Then seven minutes on, 20 minutes rest, and a final seven minutes on. This is the routine for **regular** zapping.

Tell us about your zapper experience at the zapper message board.

To Visit The Zapper Museum - Click Here!

Order A Hulda Clark Zapper - Click Here!

See The '*Build A Zapper*' Video - Click Here!

Learn about Dr Clark's latest zapping technique called: Plate Zapping

Learn about Dr Clark's latest inventions:
The Food Zappicator and The Toothbrush Zappicator

Order Electronic Parts at RadioShack.com and Frequency Generator Kits at PositiveOffset.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease. Visit www.HuldaClark.com for more information on Dr. Clark's Syncrometer™ Science, and the Zapper. Legal Disclaimer Privacy Policy

www.HuldaClark.com

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

KEELYNET

Zapper Schematics
No claims are made by KeelyNet for any of these plans
They are provided for informational purposes only!
KeelyNet 01/10/02

Simple Zapper Schematic

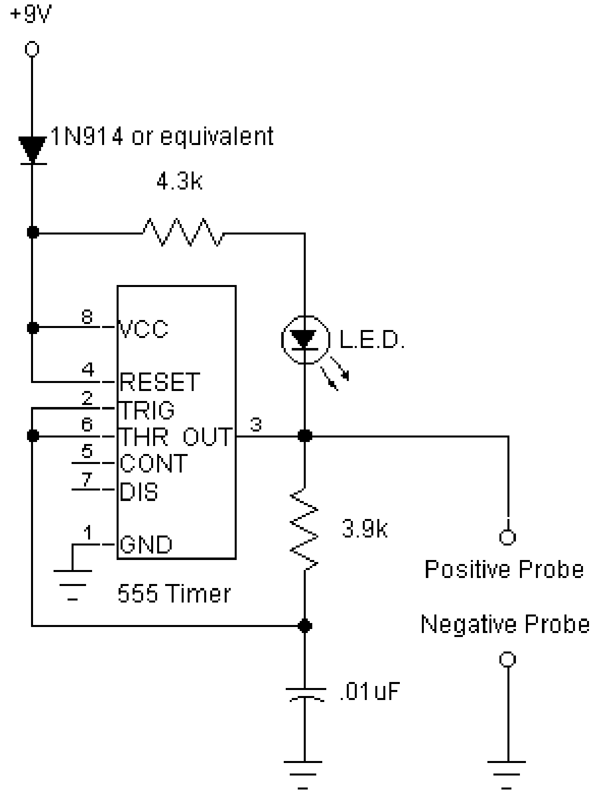
+9V
1N914 or equivalent
4.3k
8 VCC
4 RESET
2 TRIG
6 THR OUT 3
5 CONT
7 DIS
1 GND
555 Timer
L.E.D.
3.9k
Positive Probe
Negative Probe
.01uF

The Research of Dr. Hulda Regehr Clark, Ph. D., N.D.

I have recently (9/15/2000) come across the research of Dr. Hulda Regehr Clark, Ph. D., N.D. In her books, she states that many diseases, among which are cancer and HIV/AIDS, can be cured by the use of, among other things, certain herbs and a frequency generator called a "Zapper" that was invented by her and her son.

Speaking of mothers :), my mother, 62, had fatigue and arthritic symptoms in her hands before I found out about and told her about Dr. Hulda Clark. She is a seamstress, so the arthritic symptoms hindered her work. After doing the Parasite Cleanses and using the Zapper, she doesn't have to take naps like before and her hands are flexible again. She has done the complete Parasite Cleanse only twice but she Zaps practically every day.

Simple Zapper Information

Schematic updated on 12/21/2001

Simple Zapper Information text updated on 12/21/2001.

I forgot to mention that the Simple Zapper that I'm using has a 1 Megohm potentiometer instead of the 2.2k ohm resistor in order to vary the output frequency. If you use a potentiometer with the circuit put it in series with the 2.2k ohm resistor because the chip gets hot when there is too little resistance.

The duty-cycle is pretty consistent with a range of 51% to 57%. With the oscilloscope's frequency analyzer in storage mode, I hold the Zapper probes and put the oscilloscope probes somewhere on my body and then I vary the frequency of the Zapper using the 1 Megohm potentiometer.

When testing my body on a spectrum analyzer while using the Simple Zapper with an output of around 5 Volts (tested on a DC voltmeter), I noticed that the frequencies around 4,000 hz to around 18,000 hz had larger peaks telling me that my body at the time was more conductive (less reactive?) and/or less resistive to frequencies in this range at the given voltage.

With my body being more conductive (less reactive?) at these frequencies, I believe that there is a better dispersal of the Simple Zapper's output within my body. I did this recent test with one Zapper probe in each hand (as opposed to both in one hand) and the oscilloscope probes (RCA to 1/4" phone plug adaptor) in my mouth near my wisdom teeth with my lips around the RCA ground (like a lollipop).

The oscilloscope's frequency analyzer was in storage mode. With the Simple Zapper circuit on this page, change the 2.2k ohm resistor to 4.3k ohm if you want a ~15,000 hz frequency. I have recently built a couple of Zappers with optimized output frequencies of around 11,000 hz. I haven't tested anyone using the Simple Zapper other than myself, so there is the possibility that your body's response may differ.

For more information, check out my Simple Zapper webpage at:

www.members.aol.com/mas1911/index.html

or for really interesting free circuit schematics:

www.members.tripod.com/~mas1911/zappers20011221tripod.html

You can write the Author of Simple Zapper if you have comments or questions.

The following document, written by Luke Parrish, provides instructions on building an astable multivibrator in DC pulse mode, identical in output to the Zapper, described in The Cure For All Diseases by Dr. Hulda Clark.

Plans For Zapper

This page tells how to make a Hulda Clark-style "Parasite Zapper". It's a simpler and more efficient version, but works just as well. We make no theraputic claims for this device. It is for experimental purposes only. We provide a ready made version for those who prefer not to spend so much time building it.

Materials

- A CD4069 hex inverter (Radio Shack link)
- 1 1M resistor, 1 10k resistor, and 1 1k resistor (Radio Shack link) (Resistor color code)
- 1 1000 pF capacitor (Radio Shack link)

NOTE: The size of the resistors and capacitors can vary, because they determine the frequency of the zapper, which doesn't really matter much according to Hulda Clark.

The following parts may be substituted according to preference, though something similar will be needed:
- Solderless breadboard (Radio Shack link)
- 9-volt battery
- 4 alligator clip-leads sets Radio Shack link
- 2 copper pipes (hand-holds)

Instructions

Plug the components into the solderless breadboard according to the following schematic. Make sure none of the wires touch each other. To hook up the power, attach an alligator clip-lead to each side of the capacitor that stretches from hot to ground. Clip the one that goes to hot to the positive terminal on the battery and the one that goes to ground to the negative terminal. The two hand-holds are attached with the other two clip-leads. (One goes to ground, the other to the end of a resistor that goes to pin 12.)

Schematic

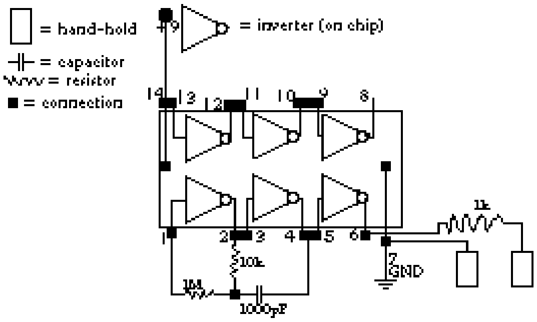

Schematic most recently updated 5-17-2001

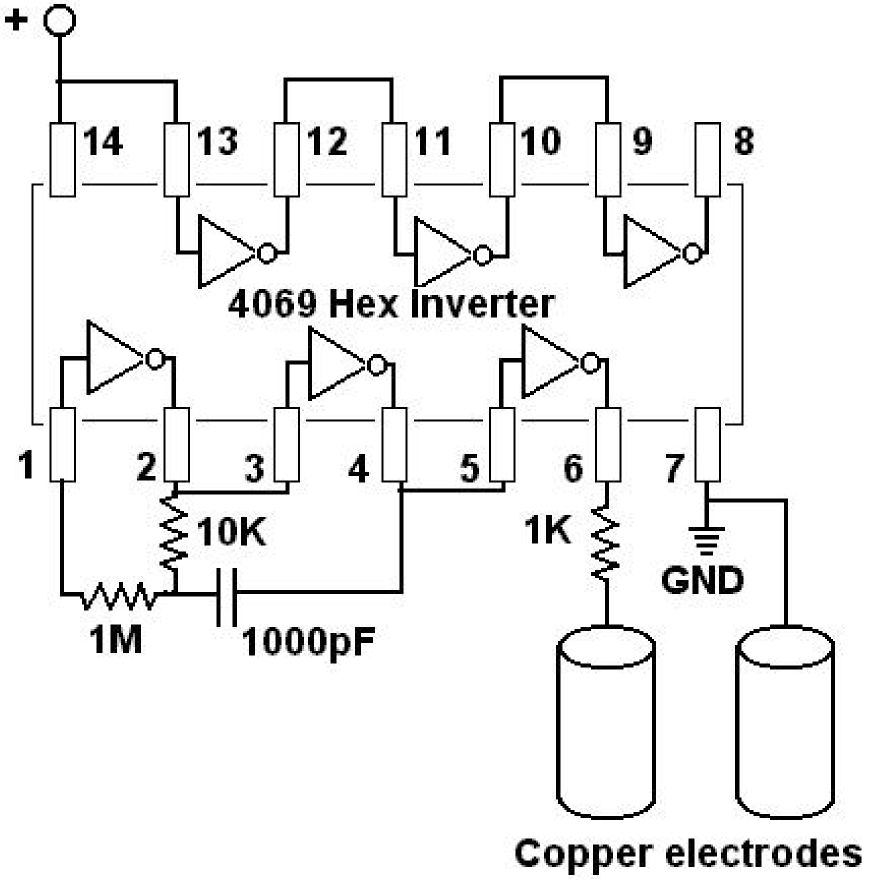

The above instructions are not the only way to do it. You can, for example, solder all the components directly to a 14-pin socket, like we do in our ready made version. This is a lot more trouble and more time-consuming for most people, which is why we recommend a breadboard. The resistors and capacitor don't all have to be the exact sizes mentioned here to make a working Zapper. However, any variation in capacitance or resistivity will change the frequency. (There are certain limits on how high of frequencies the chip can handle, so don't stray too far from the specified values.) If you have an oscilliscope, that's handy to check if it's working.

Ready made Zappers

Free plans from Zapperlab.com

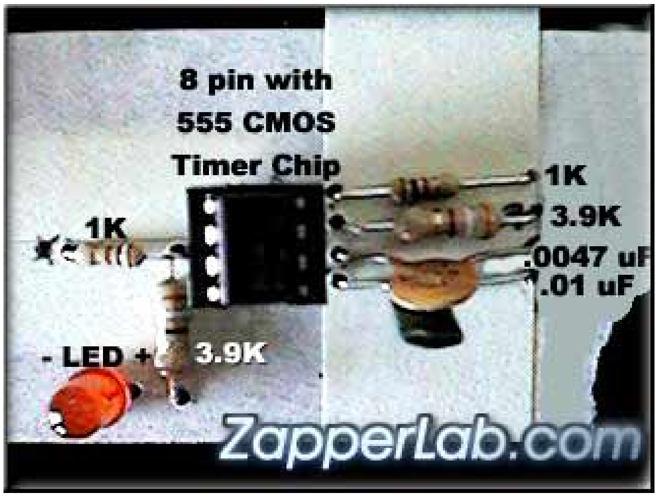

Here we have the easy step-by-step instructions for building your very own Hulda Clark Zapper reprinted with permission from her books (provided below). For advanced electricians there is a schematic for building a Hulda Clark Zapper as well. If you do not know anyone that could put one together for you, you can now order the zapper Hulda Clark uses as described in "The Cure For All Cancers" and "The Cure For All Diseases" at: www.SyncroZap.com. The Dr. Clark Zapper can be sold only as an experimental device, and no claim can be made by us legally as to a diagnosis, cure or treatment for any medical condition or disease.

Build A Dr. Clark Zapper

Order A Dr. Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. If you would rather order a pre-built zapper, just click here.

How To Build A Zapper from *The Cure For All Cancers*:

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

Get from Radio Shack, (serial number included to make it easy).
Order these parts from their web site at:
www.RadioShack.com

___ Black plastic Project Box Radio Shack # 270-1809

___ 9 volt battery

___ 9 volt battery clips: Radio Shack # 270-325

___ On off switch: Radio Shack # 275-624A

___ 1K Ohm resistor: RS# 271-1321

___ 3.9K Ohm resistor: RS# 271-1123

___ low current red LED: RS# 276-044

___ .0047 uF capacitor: RS# 272-130

___ .01 uF capacitor: RS# 272-1065

___ 555 CMOS timer chip: RS# 276-1723

___ 8 pin wire wrapping socket for CMOS chip: RS# 276-1988

___ two packs of Microclip test jumpers: RS# 278-017

___ one pack of 14" alligator clip leads RS# 278-1156C

Get from any Hardware Store:

2 Bolts 1/8" diameter, 2" long with 4 nuts and washers.

2 pieces of ¾" copper pipe, cut to 4 inches long each.

Editor's Note:

Remove 8 Pins From A 16 Pin Wire Wrap Socket

The **8 pin wire-wrapping socket** has been discontinued at many Radio Shacks and readers have asked what to do. Some say it is still available on their website at www.RadioShack.com, if not:

There are a couple options. One can still get the **16 pin wire-**

wrapping socket. Pull out the bottom grouping of 8 pins, and you now have an 8 pin socket. Just make sure when you plug the 555 Timer into it, you line it up with the remaining 8 pins. Very simple.

There is also a **Low Profile 8 pin socket**. The pins are so small one must solder to them.

The wire-wrap socket has long pins for easy twisting onto other wires. The low profile is for those that solder electronics.

If you are soldering your zapper together get the low profile 8 pin socket. Since the socket is just a saddle for the 555 Timer, it is not a vital part of the circuit. It's just a way to put long or short pins on the 555 Timer for easy twisting. If you cannot find one at your local Radio Shack, simply get the 16 pin wire-wrap socket, and pull out 8 pins. -editor.

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1 8

2 7

3 6

4 5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Attach the battery and tape it to the outside.

Now To Connect Everything

Twist free ends of the two capacitors .01 and .0047 together. Connect to Grounding Bolt using an alligator clip.

Bend pin 2 and pin 6 together inward, using an alligator clip, catch them and connect to free end of 3.9K Ohm (by pin 7).

Alligator clip Pin 7 to free end 1K Ohm (near pin 8)

Using two microclips Pin 8 and pin 4 to one end of switch (use hole to attach both microclips).

Free end of 1K Ohm (by pin 3) to the Bolt using an alligator clip.

Alligator clip the free end of 3.9K Ohm (by pin 3) to plus end (long) of LED.

Minus end (short) of LED to Grounding Bolt with an alligator clip.

Alligator clip pin 1 to the Grounding Bolt.

Alligator clip Black battery wire to Grounding Bolt.

Micro clip the Red battery wire to free end of switch.

Hook up the battery, and the light should turn on, click the switch if it does not, check connections if LED does not light up.

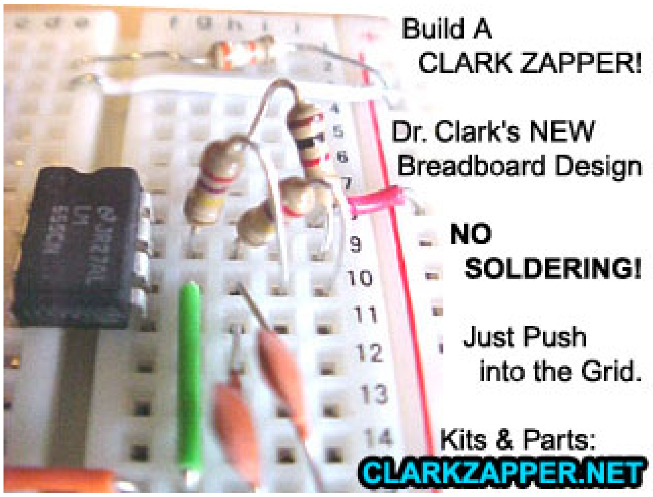

Attach a long lead wire to the grounding bolt & bolt, and then to the copper handles. Wrap handles in paper towel, clip with lead wire, wet handle when you are ready to zap.)

How To Zap

Wrap handles in wet paper towel, or wet cotton sleeves, and hold in each hand. If you have wrist cuffs, run them under water to wet them, attach to each wrist.

Attach handles or wrist cuffs to zapper.

Turn zapper on, zapping for 7 minutes.

At end of 7 minute zapping rest 20 minutes.

Do this a minimum of three 7 minute zapping sessions, with a 20 minute break in between each session. Some people zap five times in extreme cases.

Order A Dr. Hulda Clark Zapper - Click Here!

Learn about Dr Clark's latest zapping technique at: www.PlateZapping.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease.

This Circuit designed by Geoff Clark and approved by Dr. Hulda Clark

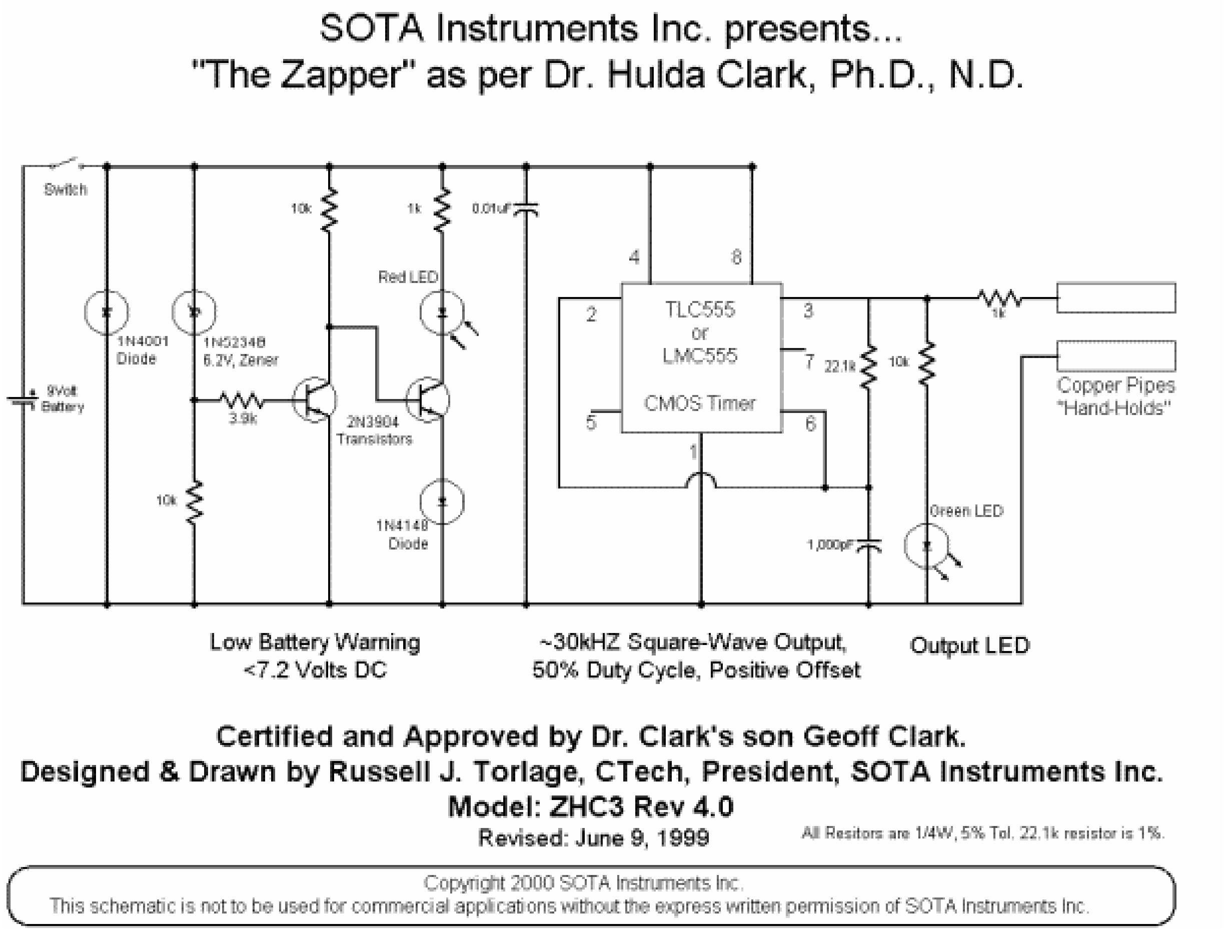

SOTA Instruments Inc. presents...
"The Zapper" as per Dr. Hulda Clark, Ph.D., N.D.

Switch
10k
1k
0.01uF
4
8
Red LED
2
TLC555
or
LMC555
CMOS Timer
3
1k
7
22.1k
10k
1N4001
Diode
1N5234B
6.2V, Zener
9Volt
Battery
3.9k
2N3904
Transistors
5
6
1
Copper Pipes
"Hand-Holds"
10k
1N4148
Diode
Green LED
1,000pF
Low Battery Warning
<7.2 Volts DC
~30kHZ Square-Wave Output,
50% Duty Cycle, Positive Offset
Output LED

Certified and Approved by Dr. Clark's son Geoff Clark.
Designed & Drawn by Russell J. Torlage, CTech, President, SOTA Instruments Inc.
Model: ZHC3 Rev 4.0
Revised: June 9, 1999

All Resitors are 1/4W, 5% Tol. 22.1k resistor is 1%.

Copyright 2000 SOTA Instruments Inc.
This schematic is not to be used for commercial applications without the express written permission of SOTA Instruments Inc.

If you found this file useful or interesting, please consider a donation or a purchase to help keep KeelyNet online and providing free information. Even a dollar will help. Others sell it, we prefer to share it, thanks!

PayPal DONATE

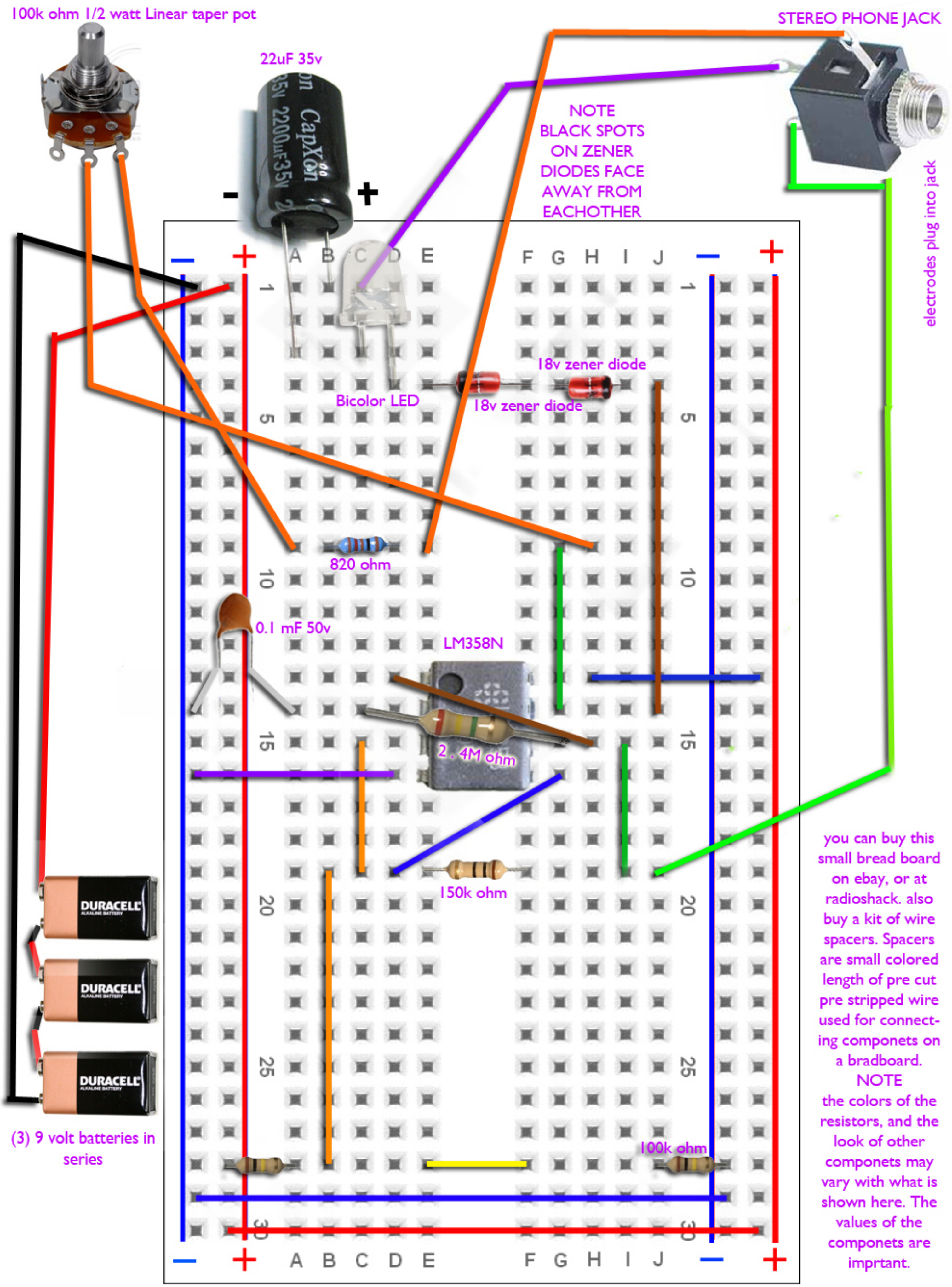
100k ohm 1/2 watt Linear taper pot
22uF 35v
CapXon
35v 2200uF 35v
-
+
STEREO PHONE JACK
electrodes plug into jack
A B C D E F G H I J
1
5
10
15
20
25
30
Bicolor LED
18v zener diode
18v zener diode
820 ohm
0.1 mF 50v
LM358N
2 . 4M ohm
150k ohm
100k ohm
DURACELL
(3) 9 volt batteries in series

NOTE
BLACK SPOTS ON ZENER DIODES FACE AWAY FROM EACHOTHER

you can buy this small bread board on ebay, or at radioshack. also buy a kit of wire spacers. Spacers are small colored length of pre cut pre stripped wire used for connecting componets on a bradboard.

NOTE
the colors of the resistors, and the look of other componets may vary with what is shown here. The values of the componets are imprtant.

COMPONENT DESCRIPTIONS, SOURCES, AND CURRENT PRICES

Revision November 24, 1996. Copyright (c) 1991/1999 Robert C. Beck, D.Sc.

Note: These data are for theoretical, informational and instructional purposes only and are not to be construed as medical advice. Consult with your licensed medical practitioner. Hundreds have been built successfully if duplicated exactly as shown.

LM358 CMOS Operational Amplifier (OP-AMP) chip (generic)		U1	$ 1.00
2.4 megohm 1/4 watt 5% resistor		R1	0.07
150 kohm 1/4 watt 5% resistor		R2	0.07
100 k ohm 1/4w 5% resistor	2 required @ 0.15	R3, R4	0.14
100 kohm linear taper pot, 1/2 watt Cattronics P-68		R5	2.56
150 ohm 1/4 watt 5% resistor		R6	0.07
820 ohm 1/4 watt 5% resistor		R7	0.07
0.1 mF 50 V (or higher) ceramic capacitor		C1	0.45
22 mF 35 V (or higher) electrolytic capacitor		C2	0.25
18 Volt Zener diodes (NTE5027A)	% Watt, 2 @ 0.79	D1 & D2	1.58
Bulb, 6.3V 0.075 A	Type 7377	B1	1.34
Bicolor LED Red/Green	RS# 276-012	LED 1	$ 1.19
Jack for electrode leads (3.5mm or 1/8" mono-phone jack)	RS# 274-251 3/$1.59	J1	0.53
Jack for silver wire leads (2.5mm or 3/32" mono-phone jack)	RS# 274-292 3/$1.59	J2	0.53
3 Alkaline 9 V batteries, type 1604 etc.	3 for		3.75
3 x 9 V battery snaps (clip-on connectors)	RS# 270-325 pkg.	5/$1.29	1.29
Misc. wire, solder, etc.			0.50
Box, if used			2.29
Lead wire with 3.5 mm plug, 6 ft., Mouser or Calrad Electronics			0.35
Electrodes, stretch elastic, Velcro, cotton flannel, alligator clips, etc.			
Total Cost **for all** components for improved **do-it-yourself** project			**$22.47**

All components listed above check your local electronics store.

Ensure the IC chip U 1 (LM358) is wired correctly. The location of Pin #1 is shown below:

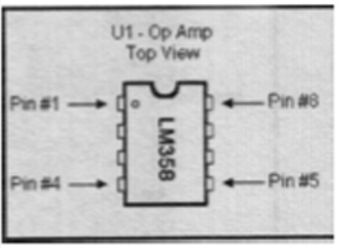

NOTE RESISTOR COLORS NOT TO SPEC. CHECK RESISTOR COLOR VALUES

electrodes: 3/32 x 1 stainless steel cotter pins -hardware store- Snap pins in half across the curve. and wrap the headphone wires comming from the mono plug around the curve of each half of the pin. then hou glue into place.

The Lakhovsky
Multiple Wave Oscillator
Handbook

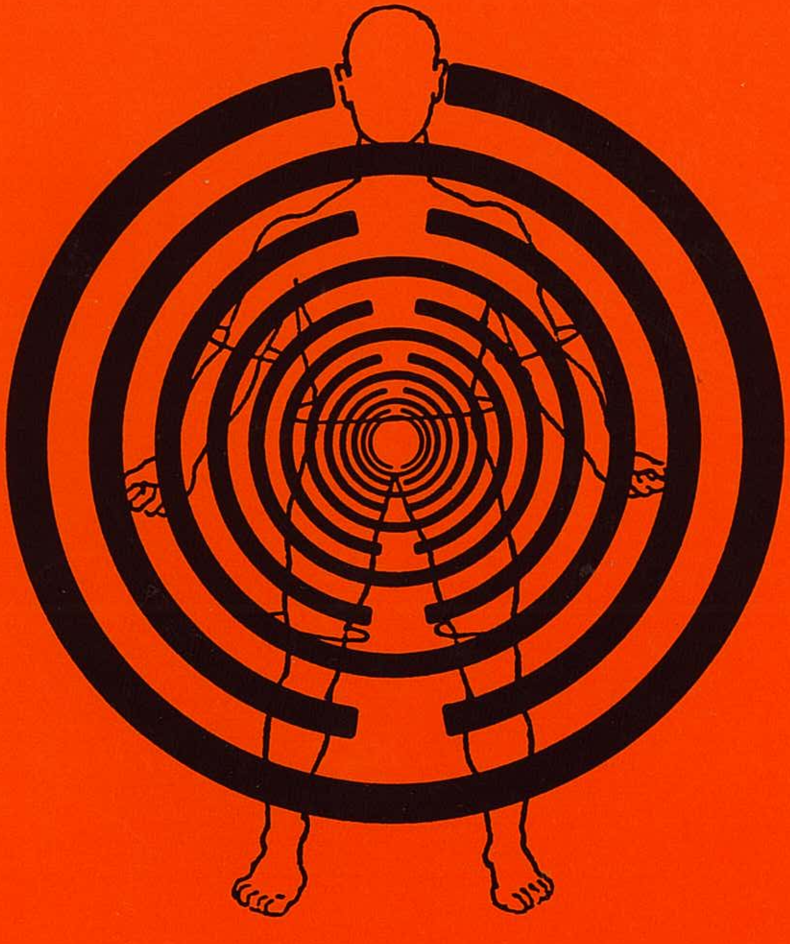

Compiled and Edited by

Thomas J. Brown

The Lakhovsky Multiple Wave Oscillator Handbook

Comprising the
Borderland Sciences Research Foundation
Lakhovsky Multiple Wave Oscillator
&
Radio-Cellular Oscillator
Research Files

Compiled and Edited by
Thomas J. Brown
Executive Director of
Borderland Sciences Research Foundation

Published May 31, 1992
in the Light of further research,
in the interests of BSRF Associates by:

P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

©1992 Borderland Sciences Research Foundation

All rights reserved. No part of this pubication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher except for brief passages quoted in reviews.

ISBN 0-945685-03-3

Library of Congress Catalog Number 92-71664

Third Revised Edition, Published May 31, 1992 by:

Borderland Sciences Research Foundation
Garberville, California 95440-0429

10 9 8 7 6 5 4 3 2 1

DISCLAIMER

This book is a research publication of Borderland Sciences Research Foundation. It consists of our MWO and RCO files which have been put into some semblance of order to preserve the records of many people with new and different ideas on how to study the energetic states of living systems and how these systems may be influenced by externally applied energies, and also to allow for further experimentation in this field by competent researchers. Read this book thouroughly before undertaking any experiments becasue you will find many different opinions, some conflicting, as to how to proceed. High frequency electrical apparatus can cause serious shock injury and should only be handled by knowledgable, competent people. All experiments are conducted solely at your own risk. We disclaim any responsibility for the use or misuse of any information in this book. While this book touches on the subject of human health, all information should be considered as anecdotal and should not in any way be construed as medical advice. We are not qualified to give medical advice and the use of the experimental equipment and research techniques described in this book may provide a false sense of security wherein one may neglect professional medical care. Use your brains. Thanks.

Tom Brown
Editor/Publisher

Table of Contents

PART ONE -- Lakhovsky's Writings and Work

Radiations and Waves – Source of Our Life by Georges Lakhovsky 5-6
Influence of Cosmic Waves on the Oscillation of Living Cells by Georges Lakhovsky 7-8
The Radio-Cellular Oscillator and its Applications by Georges Lakhovsky 9-10
Attempt at Treatment of Experimental Plant Cancer by Georges Lakhovsky.................... 11-13
The Lakhovsky Radio-Cellular Oscillator by Jorge Resines 14-19
Apparatus With Circuits Oscillating Under Multiple Wave Lengths by Georges Lakhovsky, U.S. Patent 1,962,565 20-21
Tube for Producing Multiple Wave Lengths by G. Lakhovsky, U.S. Patent 2,351,055 22-25
Lakhovsky's Vibratory Circle 26-29

PART TWO -- The BSRF MWO Research Reports

The Russian Lakhovsky Rejuvenation Machine by Bob Beck, 1963 30-35
Transistorized MWO by G.W. Steele 36-37
The Lakhovsky Multi-Wave Oscillator Power Supply by Klark Kent................................ 38-45
Excerpts from **The Waves That Heal** by Mark Clement with commentary by Riley Crabb ... 46-52
Curing Cancerous Plant With Ultra Radio Frequencies by Georges Lakhovsky, 1926 52-53
Excerpts from **The Journal of Borderland Research,** 1964-1971 54-83
BSRF MWO 1988 Update 84-94
The Spiral Coil Multi-Wave Oscillator by Louis A. Schad 95-103
Electrostatic MWO by Peter Lindemann 104-105
The Sun-Ether Disc of Oscar Korschelt..106-107
View From the Borderlands by Tom Brown, 1988-90 108-111
Excerpts from **The Journal of Borderland Research,** 1990-1992 111-122

PART THREE -- Reference Materials

High Frequency Oscillators for Electro-Therapeutic Purposes by Nikola Tesla.............123-133
The Magnetic-Arc Instrument of S.S. Knight by Mark Gallert 134
Frequency-Independent Antennas by John D. Dyson 135-137
Short Waves in Medicine: **Medical Aspects of Ultra Short Waves,** C.H. West; **Ultra Short Waves in Medicine,** Dr. F. Noack; **Human Beings as Antennas,** Dr. I. Schliephake.........137-142
Distributed Series Capacity by C.P. Steinmetz 143-144

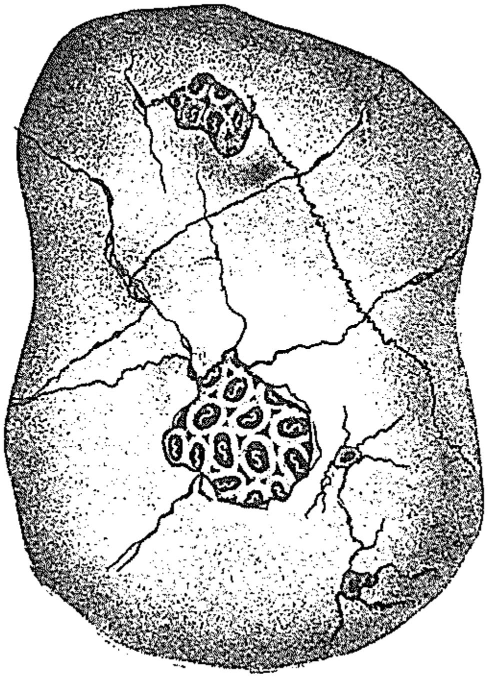

A dead cell. This diagram shows a cell where the chromosomes have stopped oscillating, and for this reason have ceased to live. But the chondromes, visible through the opening, continue to live an independent life. They are surrounded by a tough membrane and are multiplying so as to burst open the outside of the cell, as a baby chick, when it comes to maturity, breaks through the egg shell to come out.

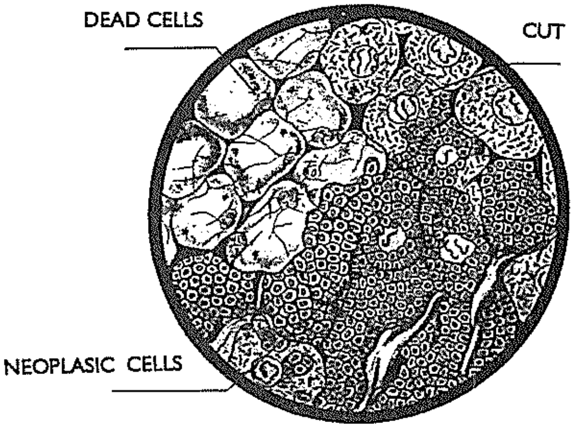

A group of dead cells, but not yet destroyed, in which the neoplasic cells are forming. One can see the transformation in the neoplasic tissue.

Photograph taken during the treatment of Mrs S..., who sits between the MWO antennas at Calvary. Behind the patient is Dr. Marquer, head physician of the hospital, to the left is Dr Z..., assistant, three voluntary Nurses, & to the right is Georges Lakhovsky.

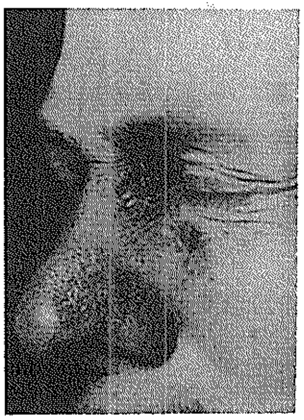

At left, Madame C... *Before Treatment.* Rodent Ulcer in inner angle of left eye.

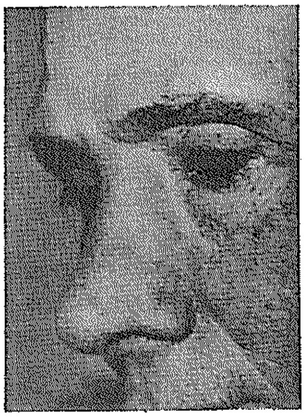

At right, Madame C... *After Treatment.*

Excerpts from:

RADIATIONS AND WAVES

Source of Our Life

by Georges Lakhovsky, 1941

THEORY OF CELLULAR OSCILLATION

Before presenting the therapeutic results obtained with my multiple short wave oscillator and presented at the International Congress of Short Waves in Vienna (July, 1937) it may be useful to give a rapid survey of my theory of cellular oscillation which I have developed fully in a number of books.*

All living cells are composed of two essential elements; the nucleus and the protoplasm in which it is bathed. This nucleus is itself composed of many tubular filaments: the chromosomes. In addition, hundreds of much smaller filaments or chondromes are present in the cytoplasm.

Chromosomes and chondromes are sheathed in an insulating substance (cholesterine, resin, fat, plastrin, etc.) and contain a liquid-like serum with the same mineral content as sea water, and consequently a conductor of electricity. Thus, these filaments constitute ultramicroscopic oscillating circuits capable of oscillating electrically over a wide scale of very short wavelengths.

I have demonstrated in my works that these cellular oscillating circuits, chromosomes and chondromes, vibrate electrically under the stimulus of electro-magnetic waves: cosmic, atmospheric and telluric.

Now, many internal and external influences may upset the oscillating equilibrium of these cells. For instance, a variation or change in the field of cosmic, telluric or atmospheric waves, a demineralization of the organic matter constituting the cellular substance, traumas causing the destruction by shock of the protoplasm or the nucleus.

I have shown in my books, **The Secret of Life** and especially in **The Earth and Ourselves**, that every living cell draws its oscillatory energy from the field of secondary radiations resulting from the ionization of the geological substances of the earth by cosmic radiations.

But certain natural radiations are particularly toxic, especially those originating in geologically-induced geopathic zones. Many cancer cases have been attributed to these toxic radiations which has been proven experimentally, notably in Germany by Dr. Rambeau of Marburg. Therefore, earth radiations sometimes cause disturbance of the cellular oscillatory equilibrium of the organism.

Under these varied circumstances cellular oscillation may cease; the cell is then dead. But within the dead cell, the chondromes sometimes continue to oscillate electrically on their own natural frequencies. Fortunately, this phenomenon occurs rarely, or all humanity would already have perished from cancer.

The chondromes then envelope themselves in a membrane and continue to oscillate and multiply independently of the cell. They may then become neoplasic cells.

THE RADIO CELLULAR OSCILLATOR

To re-establish this equilibrium, I thought of creating, in 1923, a constant compensating field of very short radiations (2 to 10 metre wavelengths) to neutralize the action of the disturbing rays, and give the living cell the necessary stimulation for a return to its normal oscillation.

To this end, in 1923, I constructed my short-wave oscillator, using two triode tubes for very short waves made especially for this apparatus. I tried several cross leakages for this machine using one or more tubes and then multiple triodes with a tube containing oscillatory circuits within the bulb. Finally I adopted the oscillator with symmetrical cross leakage comprising two triodes. The oscillating circuits formed a single spiral, branched between the two grilles and the two anodes. It was fed directly by alternating current from the local supply circuit.

With this short-wave apparatus I was able to cure plants inoculated with cancer. For six years at the Saltpetriere I observed and checked the effects of these short waves.

Using very low power, from 10 to 12 watts, and a limited duration of treatment, I succeeded in curing cancer in human beings, but also had to record some failures.

The news of the success of my experiments became widespread. In many countries, as early as 1928, they began to build short-wave oscillators of considerable power producing thermal effects.

But here was great danger that the chromosomes and chondromes, which are barely a ten-thousandth or twenty-thousandth of a millimeter in thickness, might not survive under a high frequency current. They offer much resistance, even to a low current which is sufficient to dissolve and destroy them.

---

* **The Secret of Life; Contribution to the Etiology of Cancer; The Earth and Ourselves; Cellular Oscillation; Nature and Her Wonders.**

It is simple to prove this by bringing a small bulb of from 2 to 5 volts with a filament of several hundredths of a millimeter, inserted between metal rods forming antennas, within the radius of a short wave transmitter. The bulb will light up and sometimes burn out, if it is brought too near the apparatus.

Moreover, the chondromes and chromosomes of all living cells, which are infinitely finer than the filaments of the bulbs, are sensitive centers of thermal phenomena, which may provoke their fusion. Undoubtedly this method is effective in killing microbes in the organism and in neoplasic cells. But it can also destroy millions of cells of healthy tissue in every irradiation.

THE MULTIPLE WAVE OSCILLATOR

On the other hand, I thought it possible to obtain better results by administering an oscillatory shock to all the cells in the organism simultaneously. Such a shock, very brief, produced by dampened or weakened electrostatic waves, does not provoke thermal and prolonged effects and involves no risk of burning living cells.

I therefore sought to produce an artificial oscillatory shock causing a periodic oscillation of the weak or dead cells.

At first glance this problem seems physically insoluble as there are approximately 200 quintillion cells in our bodies, each oscillating on its own natural wavelength. Theoretically, therefore, we would have to have a different wave length for each cell, so that every cell in the organism would oscillate in resonance on its own wavelength.

After much research I was able to construct an apparatus creating an electrostatic field covering all frequencies from 3 metres to the infra-red, so that every cell can find its natural frequency and vibrate in resonance.

We know that in physics, a circuit fed by damped or weakened high frequency currents creates many harmonics. Consequently, I conceived an oscillator of multiple wavelengths with a broad scale in which every organ, every gland, every tissue, every nerve, could find its natural frequency.

To obtain this result I set up a transmitter composed of a series of circular concentric oscillating circuits linked by a silk cord but not contiguous. These circuits are stimulated by damped high frequency currents from a spark gap. Thus each circuit of the transmitter vibrates not only on its natural frequency, but also on numerous harmonics.

Thus, I built an oscillator with all the basic wavelengths from 10 centimetres to 400 metres, that is, all frequencies from 750,000 to 3,000,000,000 cycles per second. But each circuit also emits many harmonics, which, with their basic waves, their interferences and their effluvia can reach the scale of infra-red and even that of visible light (1 to 3000 trillion vibrations per second.)

Since all the cells as well as the chondromes oscillate precisely at frequencies in that range, they can therefore find, in the output of such an oscillator, the frequencies which cause them to vibrate in resonance.

You know the results I obtained with continuous very short waves using triode tubes at a distance, with no contact electrode....

As early as 1931, I began using this multiple wave oscillator in various Paris hospitals: the Saint Louis, Val-de-Grace, Calvary, Necker, etc... Among the many cures with this treatment, I will mention especially those of various cases of cancer which X-ray and radium treatments failed to improve. These patients, cured six years ago, have had no recurrence and are in perfect health at this date. In all pathogenic cases this treatment gives very good results. As it does not attack the microbes directly, it does not destroy live tissue, but reinforces the vitality of the organism by accelerating cellular oscillation. It is therefore the reinforced organism that successfully resists the microbes and all pathogenic causes.

So, while X-rays and radium destroy microbes, neoplasic cells and healthy tissue at the same time -- which accounts for the serious accidents which occur during and after such treatments -- high frequency radiations (short waves) applied at a distance and without thermal effect cure diseases of all kinds, even those of the prostate to a considerable degree.

Whatever the pathogenic cause, the multiple wave oscillator reproduced the frequencies necessary to re-establish the cellular oscillatory equilibrium.

In general, it is sufficient to seat the patient, or to have him stand in the radius of the apparatus, before the transmitter. The duration of the treatment is usually from five to ten minutes, every other day. These figures are purely arbitrary, since these radiations reinforcing cellular oscillation do not produce organic disturbances, whatever the duration of the treatment may be.

Despite the many cases successfully treated, almost without exception, it must be understood that my oscillator cannot cure all types of cancer in all its stages of development. In many cases, when the cancerous tumor has already destroyed important blood vessels, my apparatus is powerless to rebuild that tissue before the fatal haemorrhage.

INFLUENCE OF COSMIC WAVES ON THE OSCILLATION OF LIVING CELLS

Part 3, Chapter 1, From La Science et le Bonheur, by Georges Lakhovsky, Paris 1930

Translated by Brigitte Donvez, Edited by Tom Brown

© 1991 B.S.R.F.

This part of my study is devoted to the trials, experimentation and applications of the theory of the oscillation of living beings.

I have assembled here short technical reports about the processes and methods employed, as well as a collection of all the papers read at the French Sciences Academy and at various French and foreign Scientific Societies about oscillatory therapeutics and all its domains of application.

We intend first to examine successively the influence of cosmic waves on the oscillation of living cells, taking as a concrete example the influence of solar activity upon the quality of wine and vintage. Then we will examine the construction of an electromagnetic short-wave generator, such as the radio-cellular-oscillator for the treatment of animal and plant organisms.

We will later consider the applications of the radio-cellular-oscillator, particularly in the experimental treatment of plant cancer, diathermy and the artificial production of fever.

Concurrently, we will introduce the application of open oscillating circuits to the treatment of diseases in various living beings, in the absence of any other oscillations but those produced by the natural cosmic and atmospheric waves. We will also review in this chapter the observations of scientists and practitioners who clinically applied my oscillating circuits.

Finally we will take up the very latest applications of oscillating circuits to water and liquid sterilization, and to the activation of the germinative power of plants.

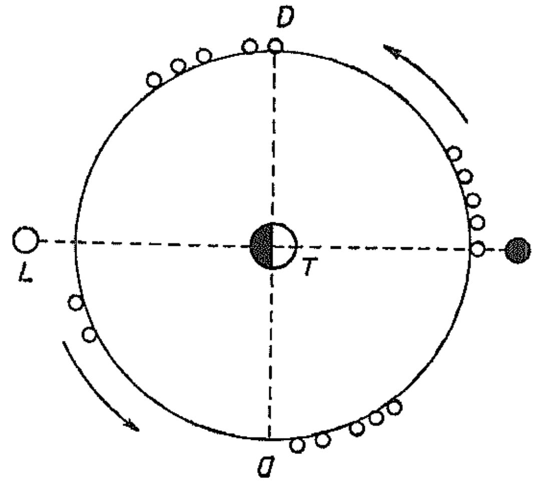

Fig. I -- Distribution of the highest intensity of electromagnetic waves observed by Mr. Vincent, as a function of the moon phases.

Obviously, the number of these applications, presently limited to certain domains, tends to rapidly expand and touch upon the various problems presented by therapeutics, biology, agriculture and zoology.

I have discussed elsewhere the role played by cosmic radiations in the development of living beings, due to the very oscillatory nature of the cell. Now, as a consequence of the earth's rotation and the stars' reciprocal motion, the cosmic and atmospheric radiation field is not constant, but rather variable. Living beings are subsequently under the influence of these variable cosmic radiations, whose nature either weakens or strengthens the cell's oscillatory balance.

I have mentioned elsewhere[1] that the lunar radiation interferes with the waves of radioelectric stations. Every week, the recurrence of moon phases enables us to obtain highest and lowest intensities in the reception of electromagnetic waves (see Fig. 1).

On the other hand, the sun does not only send light, heat and actinic (ultra-violet) radiations, but also electromagnetic waves, especially during the eruptive periods of its prominences (sunspots). The influence of the sun's electromagnetic radiation becomes obvious if we consider the variations in the earth's magnetic field and the spinning of a compass' magnetic needle during magnetic storms, under the influence of the telluric currents induced in the ground. The ionization produced by the sun's radiation also generates atmospherics and statics which disturb radioelectric receptions.

Astrophysicists, notably those of the Meudon Observatoire under the direction of Mr. Deslandres, have been able to establish a correlation between the periodicity of cataclysms on earth (tidal waves, tides, earthquakes) and that of the recurrence of sun-spots in the earth plane (about 27 days).

We can observe that the graphs which render geophysical, electric (ionization, gas conductivity, aurora borealis—northern lights), magnetic (variations of the earth field, compass spinning) and electromagnetic (disturbance in wave propagation, atmospherics and telluric parasites) phenomena, as a function of time, are very appreciably parallel to each other and to the curve of solar activity. An 11½ year periodicity appears on these graphs, confirmed by numerous, previous observations in the domain of meteorology. In particular certain tropical rainfall patterns show a 33 to 35 year periodicity, three times longer than the previous one.

I researched the extent to which cosmic radiations which condition physical and meteorological phenomena, also influence biological phenomena and the development of living beings.

As early as 1801, while cosmic radiations were not even suspected, Hershel pointed out: "the scarcity of vegetation when the sun has no spots". Starvations recur every eleven years in India with alternances of heat and rain. From 1901 to 1909, the canon Th. Moreux noticed that the wheat crop in France and in the world more or less followed the solar activity, so did arthritic conditions, and the nervousness of students and teachers.

I compared the graphs of the solar activity, as recorded at the Meudon Observatory, with the statistics of wine production in the Bourgogne and Beaujolais regions, and I noticed that the outstanding vintages coincided precisely with the year of recrudescent solar activity.

On this topic, Professor d'Arsonval agreed to present in my name at the Academie des Sciences (Academy of Sciences) on March 28, 1927, an original report entitled Influence of Cosmic Waves On The Oscillations Of Living Cells which is entirely reproduced here.

---

[1] L'univerion, p.127, Gauthiers Villars publishers

INFLUENCE OF COSMIC WAVES ON THE OSCILLATION OF LIVING CELLS

(Paper by Mr. Georges Lakhovsky, presented on March 28, 1927 to the Academie des Sciences (Academy of Science) by Professor d'Arsonval.)

In my book The Origin of Life, Radiation and Living Beings, that Professor D'Arsonval did me the honor to present to the Academie des Sciences (Academy of Science), at the sitting of the 15th of February 1926, I exposed my theory about the influence of (cosmic) penetrating radiations on living beings. As a matter of fact, I demonstrated that the nucleus of each living cell, having the form of a tubular filament made of a dielectric matter filled with a conductive substance, could be compared to an oscillating circuit, equipped with self-inductances, electric resistance and capacity. The living cells can thus oscillate at very high frequencies under the influence of cosmic radiation emitted by stars.

I made certain to verify the exactness of this theory by studying the influence of cosmic radiations (sun-spots, comets, interferences of stars' radiations, etc...) on biological matter.

My observations were based on the curves recorded by the astrophysicists of the Meudon Observatoire, curves which revealed since 1845 the activities of the sun-spots, the frequency of magnetic disturbances, and the frequency of aurora borealis -- northern lights.

These three curves are appreciably parallel. I endeavored to study the existing correlation between these cosmic radiations, on the one hand, and the development of vital activity in the plant and animal domains, on the other. As for any given individual, periods of fatigue and illness alternate with periods of health, in the same way there is, for fruit and crops in general, for each species, years of good quality and years of rather mediocre quality.

As to what concerns wine, according to the documentation established by the Chambers of Commerce of the Bordeaux and Bourgogne regions, I observed that, from 1845 until 1915, the outstanding vintages corresponded exactly to the highest activity of the sun-spots, as shown by the curves on Figure 2.

The results for red Bordeaux wines are:

Maximum of 1848: outstanding vintages of 1847 and 1848;
Maximum of 1858: outstanding vintages of 1857 and 1858;
Maximum of 1869: outstanding vintages of 1869 and 1870;
From 1880 to 1889: period of phylloxera.
Maximum of 1893: rather good vintages of 1890 and 1893;
Maximum of 1906: good vintages of 1906 and 1907.

In addition, let's point out that the wine of 1811 was called the wine of the comet, and we could ascribe its excellent quality to this comet's radiation. It is the same with white Bordeaux and Bourgogne wines.

I am sure that there are many other biological phenomena which recur periodically after a few years or a few centuries as a function of the relative revolution of the earth around the stars. I am convinced that all the scourges which ravaged humanity during the Middle-Ages: plague, leprosy, cholera, etc... and disappeared thereafter without traces in other regions will reappear some day, when the earth shall once more occupy the same position as then in relation to the stars, producing thus by interference with the cosmic waves the same oscillatory imbalances of the cells, which at that time, had produced these same diseases.

Fortunately till then, thanks to the oscillating circuits and the new methods that scientific discoveries will put to our disposal without delay, we'll be in a position to fight victoriously against these illnesses provoked by the interference of the stars' radiations.

Figure 2
RED BORDEAUX WINES

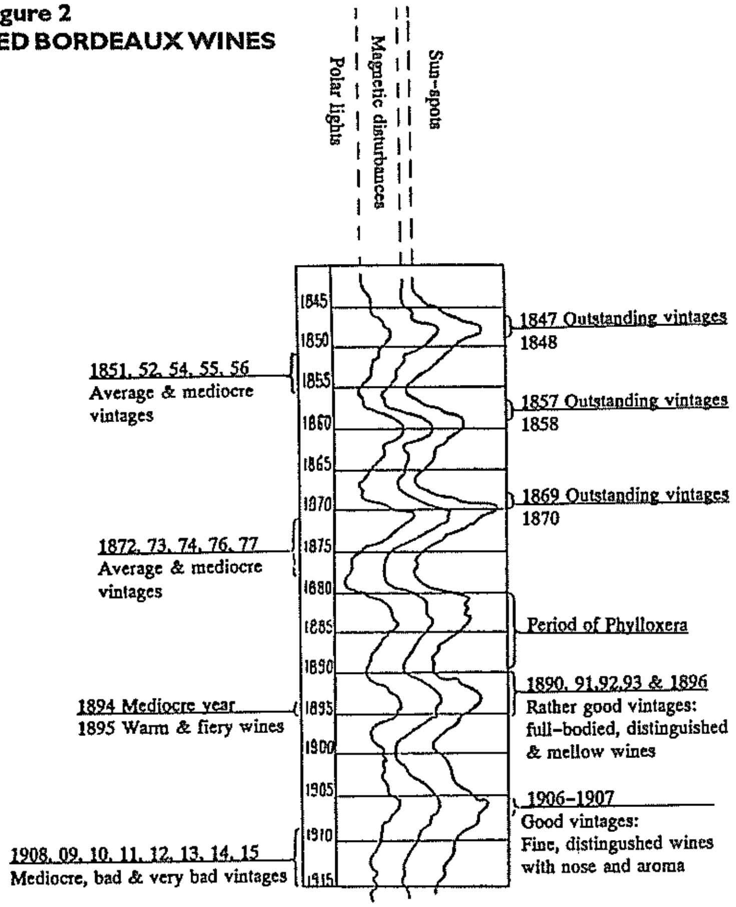

Figure 2 -- Graph showing the coincidence of outstanding vintages with the intensity of the waves emitted by the sun and corresponding to the variation in intensity: 1) of sun-spots, 2) of earth magnetic disturbances, 3) of polar lights. The graph has been established for red Bordeaux wines; the greatest intensities correspond to the outstanding vintages, the lesser intensities correspond to the years of mediocre wines. We can establish a similar graph for the other vintages, notably Bourgogne and Beaujolais wines.

THE RADIO-CELLULAR-OSCILLATOR AND ITS APPLICATIONS

by Georges Lakhovsky

Part 3, Chapter 2, from La Science et le Bonheur, Paris 1930

I demonstrated in the last chapter, how, according to my cellular oscillation theory, alteration in organic tissues came from a variation in the intensity of the cosmic waves' radiations and from these waves' frequency, thus causing the oscillation of the cells.

To reestablish the balance of these natural astral fields, the radiation of which we can no more directly influence, than a meteorologist can influence the weather, I thought of using a local oscillator, generating waves and creating an artificial and auxiliary magnetic field.

Incidentally, I should add that this method is analogous to the methods used in radiotelegraphy, known as heterodyne and superheterodyne methods. These methods use the interferences produced between the waves received and the waves generated by the local generator at the required frequency and intensity.

About 40 years ago (1890) Professor D'Arsonval advocated the treatment of certain illnesses by high frequency electrical currents, that were at the time produced in waves strongly damped by means of a Ruhmkorff coil and a spark gap (arrester).

My radio-cellular-oscillator is a generator of continuous waves of very high frequencies, the first application of triode lamps to biological research and pathological treatments.

We know the difficulty of obtaining very short waves, because of the parasitic capacitance present in electronic lamps due to the electrodes themselves and, within the bulbs, to tube sockets and connective wires.

In 1924, when I studied the possible construction of a triode lamp oscillator for very short waves, shorter than 2 meters, the radioelectrical technique, in France and abroad, was still at the stage of long-waves and the length of the so-called short-waves was in fact from 200 to 500 meters.

Therefore, I had to do my own research in view of studying a type of lamp appropriate to the generation of very short waves. Thus, I have been led to design lamps with multiple electrodes enclosed in the same bulb. These multiple lamps have been patented in France[1] and in Germany[2].

Later, I intentionally abandoned the idea of working this patent and preferred to give it up to the public domain, in a scientific aim, so that no patent claim could oppose disinterested researches undertaken for the radioelectric prophylaxis of diseases, especially cancer.

However, if the French industry did not engage in the construction of these multiple lamps, the problem was reconsidered in Germany three years later and partly resolved as the double or triple "multi-valve" by Dr. Siegmund Loewe[3].

To reduce connections to a minimum, as well as the parasitic capacity between electrodes, to achieve the generation of waves of less than 1.5 meters while avoiding low efficiency, deformation, or distortion of the wave which would jeopardize the purity of the transmission and particularly alter the modulation, I proposed in my patent to enclose the oscillating circuit and its connections within the bulb. This invention, which cannot conveniently be applied to long waves is, on the contrary, easy to carry out for very short wave circuits, of which the bobbins are reduced to several turns of coil, even to a fraction of a turn, and of which the condensers have a very small capacity (Fig. 3).

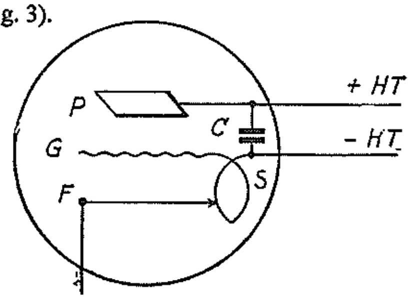

Fig. 3 -- Triode generator of very short waves, after the original drawing of Lakhovsky's patent.

Practically, the oscillating circuit only includes a turn of several centimeters in diameter. The only connections getting out of the tube are the feeders, bringing in the plate voltage (anode voltage) and the heating current (thermo-current).

These tubes can generate oscillations and transmit them by induction to outer circuits (aerials, loops, antennas).

Beside the practical simplification, the improvement of the efficiency comes from the fact that all electronic elements work with the same degree of vacuum. This is what I observed when I grouped within one same glass tubing the bulbs of a 4 triode lamp device (Fig. 4). The equal distribution of the vacuum among the various electrodes increases the purity and stabilizes the functioning.

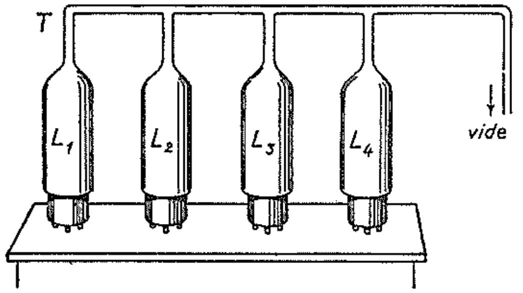

Fig. 4-- Glass tube T helps maintain the same vacuum pressure inside the four triodes of a classic resonance receiver. Author's realization.

According to this principle, I indicated the possibility of constructing not only a certain number of simple or multiple electrical cell oscillators but also all kinds of triodes or electronic lamps contained in a same bulb. Figure 5, quoted from my patent, shows the diagram of a symmetrical oscillator, with double electrodes and filaments connected in parallel. This oscillator is completely enclosed in a same casing.

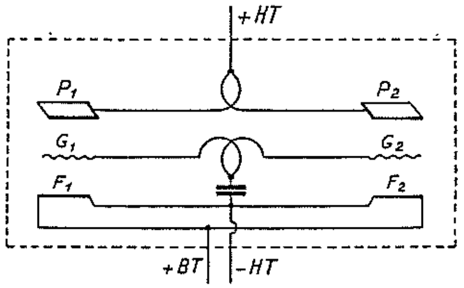

Fig. 5 -- According to the original drawing of the author's patent. Multiple triode including inside the bulb the connections of a symmetrical short-wave transmitter, with filaments set up in parallel.

---

1. G. Lakhvosky. French patent #601,155 of October 18, 1924
2. G. Lakvhosky. German patent #427,596 of May 26, 1925
3. Patent application in France on May 7, 1927

Figure 6, also taken from my patent, presents a double, symmetrical oscillator of which the filaments are connected in series. A possible tuning condenser is anticipated, which would be placed in the same vacuumed bulb as the electrodes and turns of the circuit.

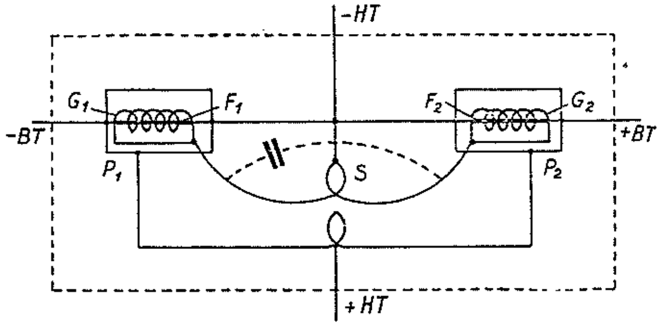

Fig. 6 – Multiple triode symmetrically connected with filaments in series, according to the original drawing of the author's patent.

Figure 7 shows the case of a double oscillator with a unique filament.

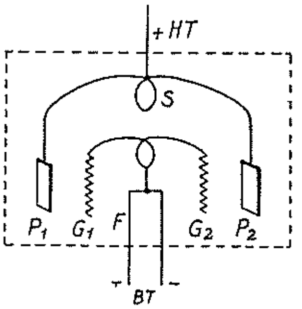

Fig. 7 – Double triode multiple lamp with symmetrical connections and a single filament, according to the original drawing of the author's patent.

For long-waves, moreover, I indicated the possibility of putting together the circuits outside the bulb.

These various constructions are extremely general. We know that for all practical purposes Dr. Loewe has only introduced resistances and capacitances inside his multiple lamps, reducing therefore its use to low and high frequencies amplifications of long waves. These lamps' efficiency is bound to be rather mediocre for the following reasons: great connection resistances, strong plate voltage (about 160 to 200 volts), and impossibility to stabilize the circuits and cause a resonance.

On the contrary, I always anticipated inner couplings through bobbins and capacitances ensuring the greatest efficiency. The resonance is obtained through a variable condenser inside the lamp and commanded from outside without any mechanical element. For example, if you use the gravity field, it suffices to place a counterweight on the sliding condenser and to direct the lamp in space. Or you can use a magnetic field produced by a magnet or an electromagnet. (Fig. 8).

I indicated the means of building an oscillator of good efficiency for ultra-short waves, with the possibility to tune it into a whole range of frequencies.

To resolve this problem on a practical level, before people started to produce multiple lamps, I used a symmetrical oscillator including two triode lamps. It is the radio-cellular-oscillator which I used in 1924, at the wards of the Saltpetriere Hospital in Paris, to treat and heal geraniums inoculated with experimental plant cancer.

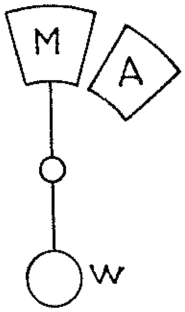

Fig. 8 – Variable sliding condenser M, balanced with a counterweight W within a triode lamp, according to the original drawing in author's patent.

The simplified diagram of the radio-cellular-oscillator is indicated in Figure 9 (Mesny set-up).

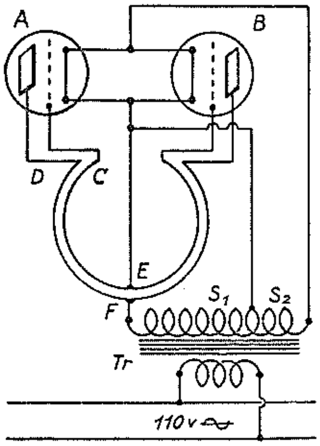

Fig. 9 – Simplified diagram of the Lakhovsky radio-cellular-oscillator.

Plates, grids and filaments parallel connected. The grids and plates circuits all comprise one turn. These two turns are variably coupled between them. Plates and filaments are directly supplied with alternating current of the mains by means of the appropriate transformer. In these conditions, the transmission of waves starts as soon as the plug is connected onto the mains. By using a tuning condenser, one can produce the entire range of waves of a length comprised between 2 and 10 meters. The supply of alternating current creates a great number of harmonics with which the oscillations of the various cells can enter into resonance. The fundamental wave length can moreover be regulated according to the nature of the cells to be treated.

The biological role of the radio-cellular-oscillator is to reestablish the oscillatory equilibrium of the living cell, threatened by any alteration, and in particular by neighboring microbes. When it resumes its normal vigor, thanks to the auxiliary radiation of the radio-cellular-oscillator, the cell can destroy the microbe itself by damping down its pernicious radiations.

The fundamental experiment I presented at the Saltpetriere Hospital demonstrates the functioning of the radio-cellular-oscillator. It dealt with cancerous plants inoculated according to Erwin Smith's method.

This fundamental experiment was the subject, on the 26th of July 1924, of a report to the Societe de Biologie (Society of Biology) and it is reproduced here in extenso, from the society's newsletter.

ATTEMPT AT TREATMENT OF EXPERIMENTAL PLANT CANCER

Report to the Society of Biology

by A. Gosset, A. Gutmann, G. Lakhovsky, J. Magrou

We know that we can produce tumors comparable to animal cancer (Erwin F. Smith)[1] on various plants, by inoculating them with the Bacterium tumefaciens. Using this method[2], one of us obtained experimentally a great number of tumors. These tumors have an indefinite growth; they might partially necrotize but they only die completely when the entire plant or at least the branch infected by the tumor has succumbed to the cachexia. Even when surgically removed, these tumors continued to proliferate.

We intend to study, in this present report, the influence of magnetic waves of great frequencies, obtained by means of a device built by one of us, for therapeutic applications and following his theoretical views[3]: the Georges Lakhovsky Radio-Cellular-Oscillator. This device generates oscillations of a λ wave length = about 2 meters, and corresponds to 150 millions (150,000,000) vibrations per second.

A first plant (Pelargonium zonatum) has been experimented with a month after its inoculation with Bacterium tumefaciens; it showed at that point several, small, white tumors, of the approximate volume of a cherry kernel. The plant has been exposed twice to the radiation, with a 24-hour break and during three hours each time.

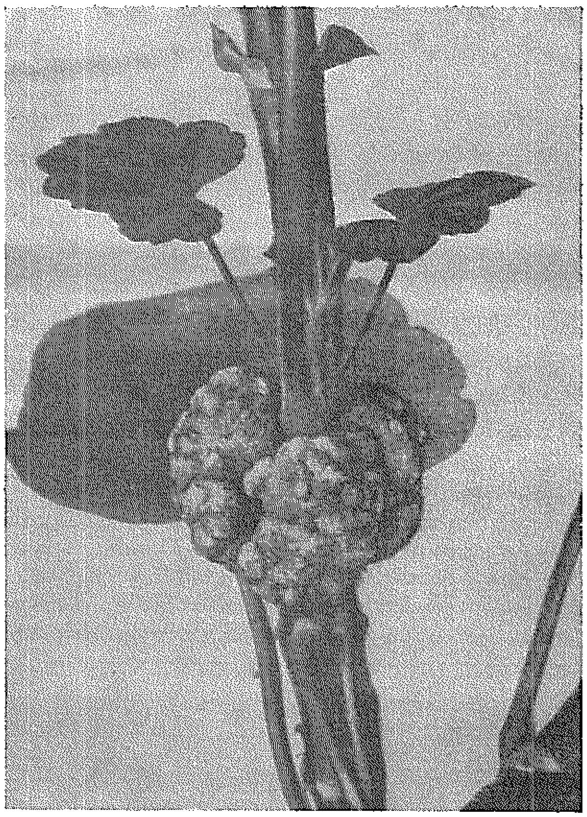

Fig. 10 -- Sight of the tumor of an untreated witness check sample. -- Subject: Pelargonium zonatum inoculated April 10, 1924. We see the growth of the cancerous tumor on the stalk. *(Clinique Chirurgicale de la Saltpetriere)*.

In the days following the treatment, the tumor continued to grow rapidly, like the check sample tumors, forming a great plurilobate mass. About sixteen days later, after the first treatment day, the tumor suddenly started to necrotize. A few days later (about 15 days) the necrosis was complete; the lobes of the tumor, retracted and completely dried out, separated from the stalk bearing them, through furrows of elmination, and the tumor could be taken off by the slightest traction. The necrosis effect of the radiation was extremely elective, limited only to the cancerous tissues which it followed into the very depth and root of the tumor; the healthy parts, stalk and leaves remained unhurt and retained all their vigor.

A second Pelargonium was treated in the like manner, but the time of exposure to the radiations was lengthened (eleven 3-hour sessions); 16 days after the first session, the tumor it carried started to necrotize and several days later it dried out completely. As in the first case, the healthy part remained unhurt.

With a third Pelargonium which underwent nine hours of radiation (three 3-hour sessions) the necrosis of the tumor's lobes followed the same path.

Sixteen control Pelargoniums remained untreated. All bore tumors in full activity which often became enormous.

As a summary, we may say that the Pelargonium which became cancerous after inoculation of the bacterium tumefaciens and for which a surgical intervention could not prevent recurrence seemed to heal under the influence of the magnetic waves indicated above. *(Clinique Chirurgicale de la Saltpetriere.)*

A very clear conclusion can be drawn from this experiment and its counterpart.

On the one side, we observe that, on the control subjects, inoculated with the *Bacterium tumefaciens* and left untreated, enormous tumors grew which by absorbing the plants' vital energy lead them to a quick and certain death.

On the other hand, the plants treated by the radio-cellular-oscillator, and moreover taken at random from among the inoculated subjects, have been healed quickly and acquired a considerable vitality, as is proven by their growth, their continual development and abundant flowering.

We can see it in the photograph (Figure 12), taken one year after the healing of the geranium.

My theory of cellular oscillation, that I had moreover proposed before I realized this experiment, enables us to understand cancer pathology and to direct its therapy onto a new path.

Cancer, mostly reaching older people, comes from the ageing of tissues as a result of the cells and blood decay. This impairing of the cells results from a modification of the

---

[1] Erwin F. Smith, *An Introduction to Bacterial Diseases of Plants*, 1st vol., Philadelphia and London 1920.

[2] J. Magrou, *Revue de pathologie comparee*, March 1924. Two other reports will be soon published in the *Revue de pathologie vegetale et d'entolmologie agricole* and in the *Annales de l'Institut Pasteur*.

[3] Georges Lakhovsky, *Radio-revue*, November 1923 and Conference to *L'Ecole Superieure des P.T.T.*, June 2nd, 1924.

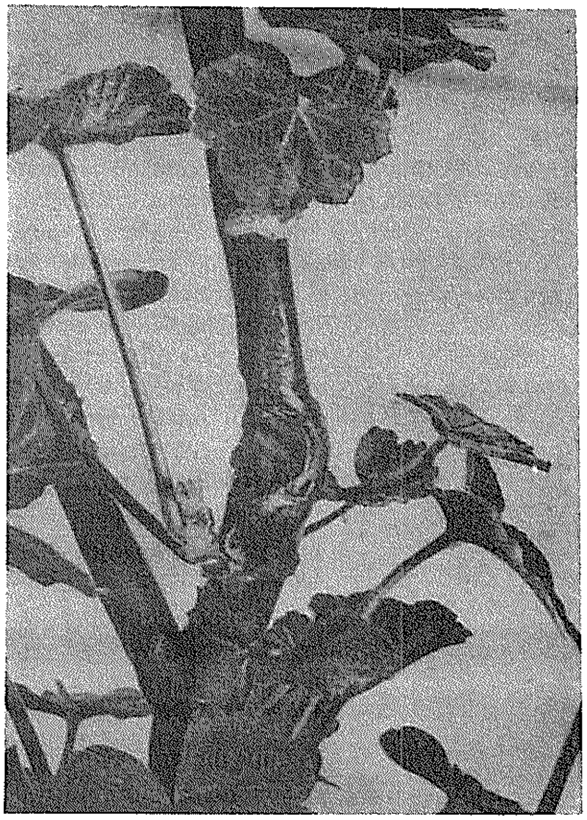

Fig. 11 Sight of the scar of the treated subject. Subject: Pelargonium zonatum, inoculated April 10, 1924 with the Bacterium tumefaciens, treated from May 24, to June 14, 1924, during eleven three-hour sessions with the Lakhovsky oscillator with antennas, photographed after it healed on July 24, 1924. *(Clinique Chirurgicale de la Saltpetriere).*

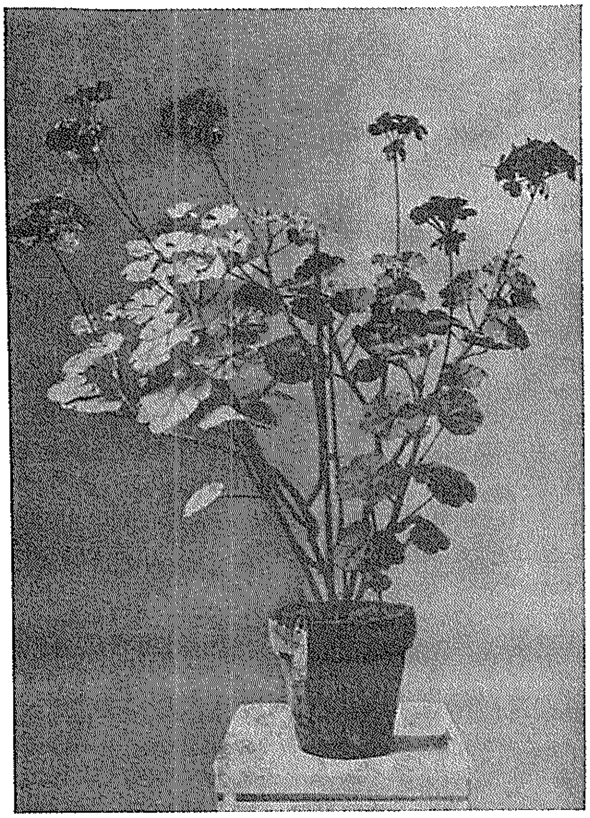

Figure 12 Sight of the treated subject after healing. -- This pelargonium is no other than the one of Figure 10, treated by the Lakhovsky oscillator, on May 24, 1924, healed on the 4th of June, 1924, and photographed in July 1925. As you can see, this plant is in perfect health and abundantly flowered. On the other hand, the sixteen inoculated but untreated check samples died long ago.

chemical composition due to the contribution of foreign bodies which appear as residues of the organism. Such are, for instance, blood haematoblasts, rich in iron and phosphorus, produced by residues of fibrin, leucocytes, and hematites. Each cell destruction brings with it a small mass of cholesterol, an insulating substance forming the armature of the nuclear filament, which later clogs the nucleus. Therefore older people first experience a decrease in the number of active cells and of the white and red corpuscles, which is reflected in the regression of the blood count, and on the other hand, an increase of the quantity of mineral (iron, phosphorus, haematoblasts) and organic (cholesterol) residues. These chemical modifications of the cell result in the variation of the cellular oscillation both in intensity, as a consequence of the changes of electric conductivity and resistance, and in the radiation frequency, because of the changes of capacitance and inductance.

The decayed cell appears like a pendulum with a pendulum-bob one could have lengthened, shortened or damaged. This oscillatory disorder provokes, because of the excess of cosmic waves, a too rapid cellular division, which results in the formation of neoplasms, or cells vibrating at a different frequency from the regular frequency of the tissue.

For long years, it has been wondered whether cancer is contagious and hereditary, or whether it is caused by a microbe. Histological analyses seem to invalidate the existence of a microbe. On the other hand, inoculations of cancer only succeed, and moreover very rarely, with impaired tissues composed of abnormal cells, such as certain moles and this would seen to ruin the hypothesis of contagion and heredity.

In fact, I believe that the true cancer microbe is the neoplastic cell which forces the neighboring cells into oscillating at its own frequency, thus transforming by electromagnetic induction, the healthy cells into neoplasms. To prevent cancer, is to prevent this neoplastic induction and to give back the healthy cell enough vigor to reestablish its oscillatory equilibrium.

I was pleased to observe that, apart from the skepticism of a few, my experiments have been taken up again with great success in France and abroad by various scientists and practitioners who understood the future of oscillation therapeutics.

It seems appropriate to quote the works pursued in Germany by Professor Esau at the Technisch-Psychaliches Institut of Iena.

Although directed toward paths sometimes very different from mine, these works present a very clear analogy with those I have conducted from the use of a very short wave oscillator similar to my radio-cellular-oscillator.

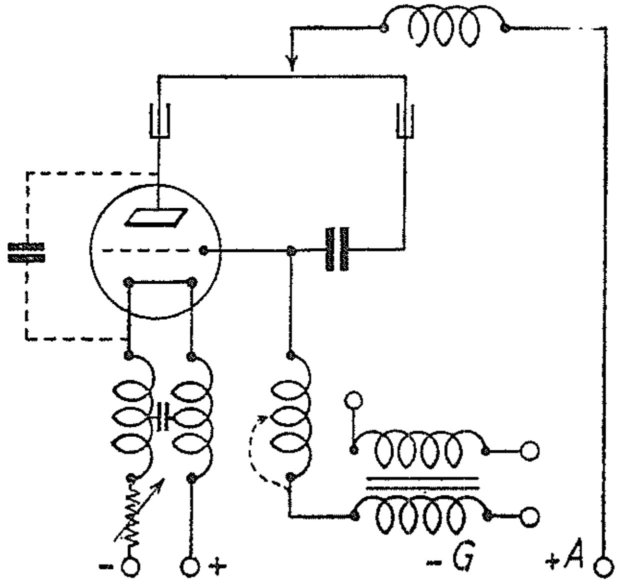

Fig. 13 -- Simplified diagram of Dr. Esau's very short wave oscillator.

The transmitter-receivers used by Professor Esau do not essentially differ from the devices I advocated. I indicated moreover that connections were irrelevant as long as they produced the desired very short waves. Professor Esau's connections, which were described in the technical magazines of Dr. Ernest Busse, assistant at the Ieana Institute[4], are characterized by the use of a single triode lamp, functioning in direct current. The transmitter's filament (Fig. 13) is supplied through a filter preventing high frequency back currents. The coilings are simple fractions of turns, as in the radio-cellular-oscillator, which only includes about half a turn for each oscillating circuit.

A test receiver (Fig. 14) is provided with square loop circuits on which variable plugs can be installed. Another device is provided with a circular turn, of which a radial cursor allows connection to the necessary position.

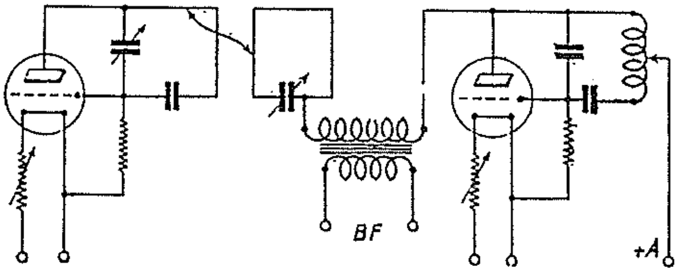

Fig. 14 -- Simplified diagram of Dr. Esau's very short wave receiver.

The works of Professor Esau in the biological domain are rather recent, for during the Radio-electric Symposium in Bremen (August 1928)[5], he stated that he had been doing research, for only a year, on the physiological effects of *ultra-short* waves, or, according to his definition, very high frequency waves comprised between a length of 1.50 and 3 meters, of which I indicated the use in 1924 in my report addressed to the Societe de Biologie, reproduced here. He indicated then that since July 1928, he successfully constructed a very short wave receiver.

About this subject Professor Esau pointed to the fact that the main difficulties came from the electronic lamps, and that he had to use foreign lamps, particularly French lamps, which are more appropriate for very short waves techniques. I indicated earlier that, 4 years before, I had studied and had these special short-wave lamps built, which were designed for my radio-cellular-oscillator.

I shall also point out that a very short transmitter has recently been used by Dr. Frantz Seidel from Vienna, for the destruction of bacteria, particularly in the sterilization of milk.[6] This subject will be investigated at a later time.

---

[4] Dr. Ernest Busse, *Die Erzeugung ultrakurzer Wellen.* (*Radio-Welt*, No. 49, 1928, p. 77. Vienna.)

[5] Oscar Gadamer. *The ultra-short waves and their field of application.* (*Der Deutsche Rundfunk*, August 31, 1928, t. VI, p. 240, Berlin.)

[6] *Les ondes ultra-courte contre les bacteries.* (*Radio-Welt*, January 27, 1929, No. 4, p. 108, Vienna.)

THE LAKHOVSKY RADIO-CELLULAR OSCILLATOR

a synopsis by Jorge Resines

What you are about to read is my resume of the book La Science et le Bonheur, Longevite et Immortalite par les vibrations -- *Science and Good Health, Longevity and Immortality Through Vibrations* -- by Georges Lakhovsky. More specifically, this synopsis encompasses Part III, chapters 1 and 2, being pages from 137-171 of said work, and whatever I translate from it will be forwarded by the words "Lakhovsky said:".

Mr. Lakhovsky begins by indicating the influence of heavenly bodies upon all facets of everyday life, from their influence upon the cells up to how magnetic compasses are affected by magnetic storms originating in the Sun (he refers the readership to his book L'Universion, published by Gauthier-Villars of Paris, France). He says that the French Astronomer Mr. Deslandres with his collaborators at the Observatory of Meudon have established a cycle of about 27 days for sunspots affecting both tides and earthquakes. This lapse is part of another larger cycle of 11½ years (or about) that most usually forms in turn part of another, larger, cycle of 33 to 35 years.

Mr. Lakhovsky also discovered, as indicated by the small circles at the lunar orbit of Figure 1, that lunar radiations provoke interferences with the waves broadcasted from radioelectric stations at said marked positions.

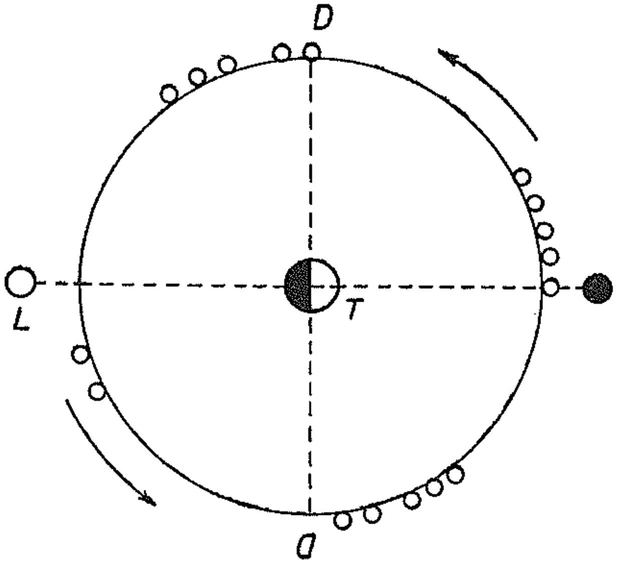

Figure 1

T = Earth
L1 = Full Moon
L2 = Waning Moon
L3 = New (Dark) Moon
L4 = Waxing Moon

Carrying his researches a step further, helped also by the people at the Observatory of Meudon, he correlated the highs-and-lows of three different factors (to wit: a. Sunspot activity, b. Perturbations of Terrestrial Magnetism, and c. Polar Aurora) with the type of wines collected at different French regions. For a single zone and type of wine, his findings are shown in Figure 2 (next page). You will easily notice how the highs of the three factors are in coincidence with the best collections of wines, with a single exception: the 1880-1889 period when the grape-wines were massively attacked at France by the "Phylloxera" plague. Though Mr. Lakhovsky does not explain the reason for this adverse period (when the opposite was expected), he does voice his opinion that cellular and biological phenomena is periodically affected by the same type of radiations (and the generated-sicknesses they produce), proposing at the same time to destroy the scourges by overriding the "astrally-produced radiations" (his words) with this "Radio-Cellular Oscillator."

About the basic concept behind his devices, Lakhovsky said: "To re-establish the balance of these heavenly fields from nature, upon which variations we are able to act directly, unlike Meteorology which cannot make weather change, I will resort to a local oscillating-generator of waves creating an auxiliary artificial electromagnetic field.

"By the way, I must add this method is analogous to that employed in radiotelegraphy and known under the names of heterodyning and superheterodyning. These procedures employ the interferences generated between the received and generated waves by a local generator, at the necessary frequency and intensity."

Mr. Lakhovsky states that his inspiration came from Professor Arsene D'Arsonval, who in 1890 had put forth the opinion that different sicknesses could be eliminated by means of high-frequency currents. He continues on, explaining that by about 1924 he studied the possibility of generating waves under 2 meters in length, but that the electronics of said period considered "short" only those waves from 200 to 500 meters.

To overcome such a hindrance, Lakhovsky developed a vacuum tube with multiple-electrodes and received two patents for it -- French Patent No. 601,155 granted in October 18, 1924 and German Patent No. 427,596 granted in May 26th, 1925 -- afterwards another scientist of that time, Dr. Siegmund Loewe, improved on his ideas and demanded a French Patent on May 07, 1927 (this Patent is not indicated as either granted or not; further investigation is needed here).

Figure 3

+HT = Positive High Voltage
-HT = Negative High Voltage
G = Grid
P = Plate
C = Condenser
S = Loop of Coil
F = Filament

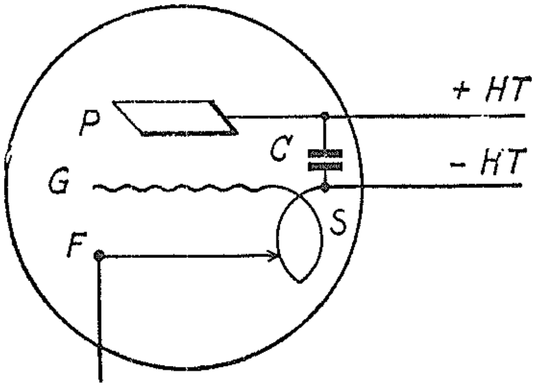

Figure 2 -- RED BORDEAUX WINES

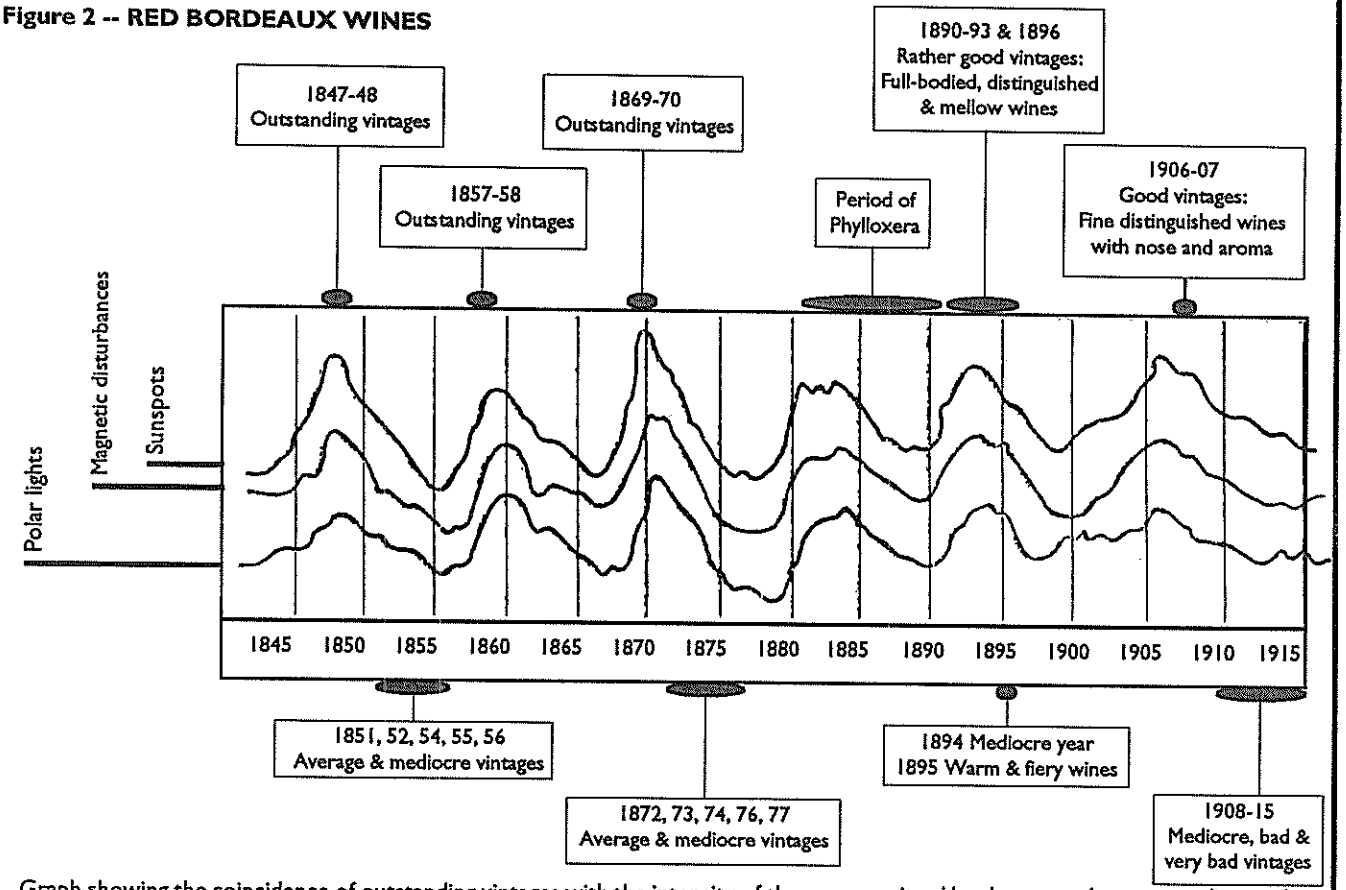

1847-48
Outstanding vintages

1857-58
Outstanding vintages

1869-70
Outstanding vintages

Period of
Phylloxera

1890-93 & 1896
Rather good vintages:
Full-bodied, distinguished
& mellow wines

1906-07
Good vintages:
Fine distinguished wines
with nose and aroma

1851, 52, 54, 55, 56
Average & mediocre vintages

1872, 73, 74, 76, 77
Average & mediocre vintages

1894 Mediocre year
1895 Warm & fiery wines

1908-15
Mediocre, bad &
very bad vintages

Graph showing the coincidence of outstanding vintages with the intensity of the waves emitted by the sun and corresponding to the variation in the intensity: 1) of sun-spots, 2) of earth magnetic disturbances, 3) of polar lights. The graph has been established for red Bordeaux wines; the greatest intensities correspond to the outstanding vintages, the lesser intensities correspond to the years of mediocre wines. We can establish a similar graph for the other vintages, notably Bourgogne and Beaujolais wines.

To produce waves "under 1½ meters" (his words), Mr. Lakhovsky enclosed within a vacuum tube's glass vessel different elements, as is seen in Figure 3. About this new type of tube, Lakhovsky said: "The oscillating circuit does not involve but a loop with a diameter of several centimeters. The only connections going out of the tube are the supply-wires, leading the plate-voltage and the heating current. These tubes can generate oscillations and transmit them to external circuits (square-loop antennas) by induction."

I will stop now the synopsis to introduce a personal comment: It is not clear on the illustration of Figure 3, WHY the filament is connected -- in a controllable fashion as the schematic indicates -- to the loop "S"; neither do I see clearly indicated whether the tube is one of direct or indirect heating and which type of vacuum tube -- existing at the time, with its specifications of that era -- served as the foundation to develop the tube of Figure 3, and how this was made in both constituents, procedure of manufacture, etc.

Neither is it indicated if there was a constitutional modification involving a change of material(s) for the new type of tube Mr. Lakhovsky first employed. I do not want to spoil anybody's fun in experimentation, I am just suggesting caution... Let's go on with Mr. Lakhovsky.

Another feature employed to ensure a high-quality when in use by Mr. Lakhovsky was to put a group of electronic tubes, as illustrated in Figure 4, directly over these lines, in the same vacuum pump. As all the tubes were connected in series, when the apparatus was made to work it produced the same degree of vacuum within them all.

Figure 4

L_1, L_2, L_3, L_4 = Identical Vacuum Tubes
T = joining tube

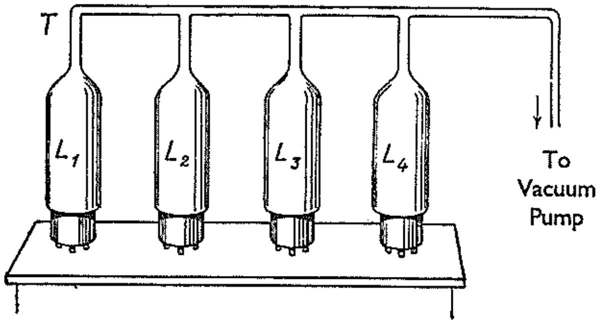

A further development of his idea, "extracted from his patent" he claims without indicating which one (French or German) this is, is the multiple triode illustrated in Figure 5. From its analysis we notice that Mr. Lakhovsky adhered to electronic theory, which says that electricity is a flow of electrons, and in a semiconducting device as the usual vacuum tube there is only a single direction of flow, from the "Cathode" (or negative pole) to the "Anode" (or positive pole). In the multiple triode of Figure 5, the filaments are connected in parallel.

Figure 5

+BT = Positive Filament Voltage
-HT = Negative High Voltage
G_1 = Grid Tube No. 1
P_2, G_2, F_2 = Same for Tube #2
+HT = Positive High Voltage
P_1 = Plate Tube No. 1
F_1 = Filament Tube No. 1
C = Condenser

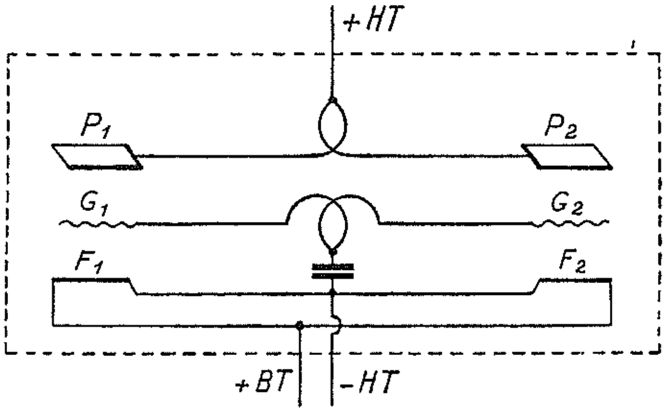

A further improvement of the device is seen in Figure 6, of which Lakhovsky says: "This illustration, also taken from my patent, shows a double and symmetrical oscillator where the filaments are mounted in series. It is also provided for, eventually, to place within a condenser so located as the electrodes and the loops of the oscillating circuit, within the same evacuated glass-vessel".

Figure 6

G_1 = Grid Tube No. 1
P_2, G_2, F_2, = Same for Tube #2
-HT = Negative High Voltage
-BT = Negative Filament Voltage
P_1 = Plate Tube No. 1
F_1 = Filament Tube No. 1
C = Condenser
+HT = Positive High Voltage
+BT = Positive Filament Voltage

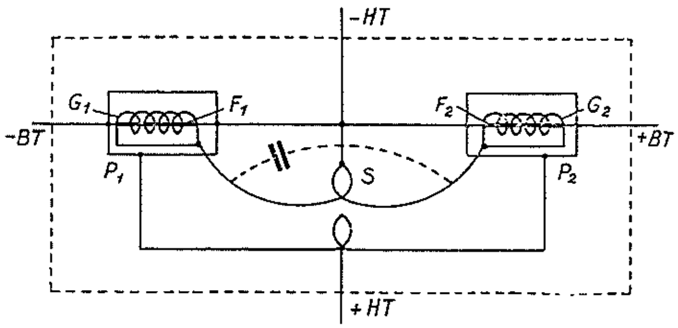

Figure 7 shows us a double-oscillator with a single filament, that is the two lamps of Figure 6 within a single vessel and with just one heat-source.

Figure 7

+HT = Positive High Voltage
+/- BT = Positive and Negative Filament Voltage
F = Filament
G_1 = Grid No. 1
G_2 = Grid No. 2
S = Loop of Coil
P_1 = Plate No. 1
P_2 = Plate No. 2

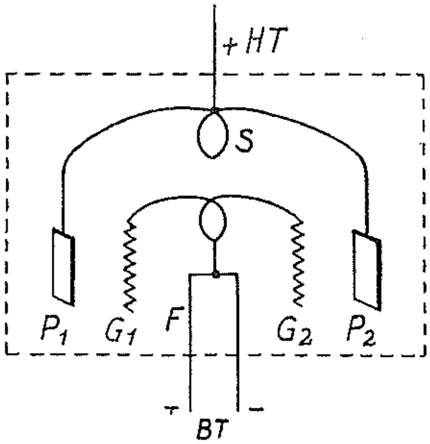

I remind the readership that all the models of vacuum tubes shown so far are for short waves under 2 meters. Lakhovsky indicated that to obtain long waves it sufficed to place the same elements outside a vacuum tube.

After saying that everything so far shown were just "extremely general things" Lakhovsky said: "One can, almost, say that Dr. Loewe just introduced resistances and capacities within his multiple-tubes, where performance is just reduced to amplifying low and high frequencies upon long waves. The performance of these tubes is very mediocre because of the following reasons: a) Heavy linking resistances, b) Strong plate-voltages (about 160 to 200 volts), c) impossibility of coupling circuits and thus producing a resonance."

Figure 8

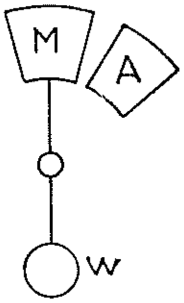

M = Mobile blade of condenser
A = Fixed blade of condenser
W = Weight to move blade "M"
H = Turning point for the combination of "M" with "W"

Here, in the former comment, is an apparent contradiction: Lakhovsky criticized Dr. Loewe for including resistances and capacities within his multiple-tubes, while at the

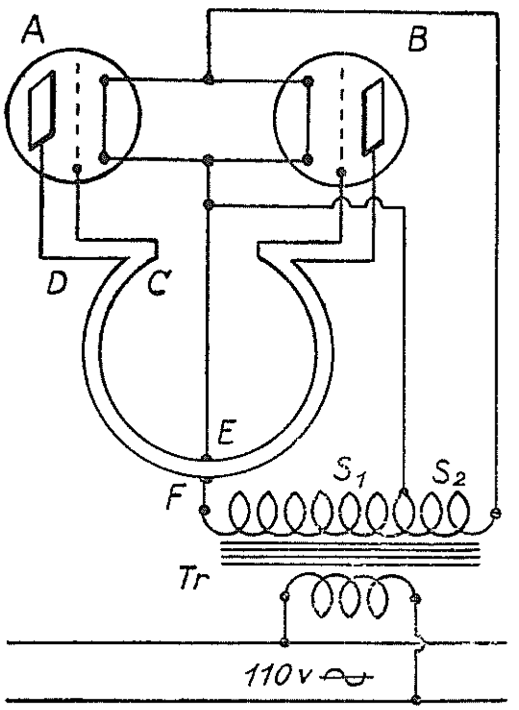

Figure 9

A = a Lakhovsky Triode
B = another Lakhovsky Triode
C = Loop associated to Grids
D = Loop associated to Plates
E = Connection to Filament
F = Connection to High Voltage of Transformer secondary tap S_1
S_1 = High Voltage tap of AC Transformer secondary
S_2 = Filament Voltage tap of AC Transformer secondary
Tr= AC step-up Tranformer with a tapped secondary winding

same time he did the same! But Lakhovsky's capacitors are different to the common types.

What he did was to "separate" a variable condenser's blades and to "counterweight" the moving blades, as indicated in Figure 8, so that the blades would move within the fixed armature of the condenser as the whole tube was rotated and weight "W" responded to gravity! He does not indicate, thus making it an object of personal experimentation on the part of the readership, if there was just one or more of these special variable condensers.

Now, in order to produce a wide range of frequencies with only two of his special triodes, Lakhovsky built the "Radio-Cellular Oscillator" that is shown in schematic form in Figure 9, which he used at the Clinic of Salpetriere in 1924 to treat and heal Geranium plants which he experimentally inoculated with the disease of cancer; it was mounted, he says, according to the "Montage Mesny," thus indicating that one man called Mesny originated the setup at that time (this is another source for research, try finding in old books from either the '20s or '30s anything made by a Mr. Mesny).

This is what, on the device, Lakhovsky says: "Plates, grids and filaments are connected in parallel. The circuits of grid and plate have each a single loop. These two loops are coupled variably between themselves. Plates and filaments are fed directly from the Alternating Current of the place by means of the appropriated transformer. Within these conditions, it is produced the emission of waves because it is branched the intake of current upon the electrified section.

"Using, accordingly, a condenser it is possible to produce the whole spectrum of waves between 2 and 10 meters of wavelength. The supply from AC produces a large number of harmonics upon which can enter into resonance the oscillations of different cells. Therefore, the fundamental wavelength can be regulated according to the nature of those cells to treat.

"The biological role of the Radio-Cellular Oscillator is to re-establish the oscillatory equilibrium in the living cell, under threat by any alteration, particularly by a close microbe. When recovering its normal strength thanks to the auxiliary irradiation from the Radio-Cellular Oscillator, the cell can all-by-itself destroy the microbe by counter-acting its deleterious radiation."

Mr. Lakhovsky goes on indicating that he established the aforementioned concept as the foundation of his device by first inoculating several plants of *Pelargonium zonatum* with the *Bacterium tumefaciens* in order to generate tumors similar to those of animals in plants (his source was: An Introduction to Bacterial Diseases of Plants by Erwin F. Smith, London and Philadelphia, 1920). He was helped with inoculating the plants by Mr. A. Gosset, Mr. A. Gutmann and Mr. J. Magrou, the latter being the one who had more articles published on the subject of plant-inoculation (see the Revue de Pathol. Vegetale et d'entomologie agricole and within that the Annales de L'Institut Pasteur, by this author).

He then applied his Radio-Cellular Oscillator, using waves of about 2 meters, which he equated with a vibration of 150,000,000 cycles per common-second. He broadcasted the oscillator's output upon the infected plants and got them all back to health after a given time. He had already explained how his oscillator worked to his fellow Frenchmen in an article (Radio-Revue, November 1923) and a conference (given at the Ecole Superieur des Postes, Telegraphes and Telephones

Figure 10

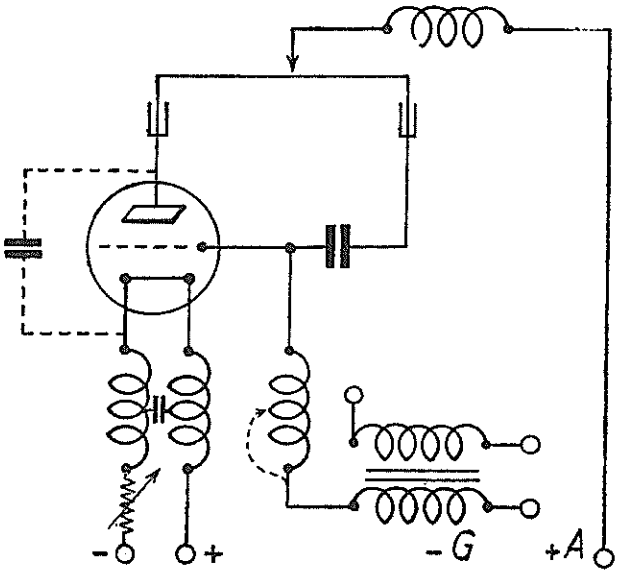

on June 2nd, 1924).

After detailing how the experiment in plant-healing progressed and how the plants were healed, he goes on to explain that -- in his conception -- cancer is provoked by an "aging" of the cells that makes them both reproduce at a slower rate and to evacuate a higher amount of wastes. This, in combination with radiations coming from space at different times, produces an oscillatory disequilibrium in the cells resulting in cancer of different types. Mr. Lakhovsky rejects, on the ground of histological examinations he performed, both the "microbial" theory of cancer and its hereditary transmission.

If you have followed all of the writing up to this point, carefully noting the concepts involved in Lakhovsky's procedures, you will remember that:

(1) His Radio-Cellular Oscillator created an auxiliary electromagnetic field capable of restoring balance to the diseased cell(s) and body.

(2) This was done by means of very short waves (during the period he experimented) of about 2 meters and 150,000,000 cycles per second.

(3) To achieve the necessary shortness of wave he introduced different elements within an evacuated glass-vessel, thus creating a different type of vacuum tube.

(4) He purposely indicated that external elements created long waves that did NOT comply with what he needed for therapy.

(5) He included, within his special tubes, a variable condenser that could be tuned to the desired capacity by moving the tube in space so that the internal variable capacity could generate the coupling needed in therapy.

If all of this has been, so far, an essential group of conditions necessary to produce the short waves of about 2 meters, then why -- as seen in Figure 9 -- does he place two electrified loops OUTSIDE the vacuum tubes of his special design????

In compliance with this "loops outside" line, he goes on quoting the works by one of many foreign scientists who, he claims, have followed his line of research:

Lakhovsky says: "It is convenient to quote the works made by Professor Esau at Jena's Technisch-Physicalishes Institut (Germany). If they are directed by ways somewhat different to mine, his works show the similitude of using an oscillator of very short waves similar to my Radio-Cellular Oscillator. Professor Esau's emitters and receivers do not differ essentially from the apparatus I champion. I have already indicated that the arrangements of constituents are indifferent, because they produce the very short waves needed.

"The setups by Professor Esau -- described in the Institute's technical magazines by Dr. Ernest Busse (see Die Erzeugung Ultakurzer Wellen -- *Research on Ultrashort Waves* -- in Radio Welt, No. 49, page 77, Vienna, 1928), secretary to the Institute of Jena, are characteristic because they employ a single triode working with DC.

"The emitter's filament -- as seen in Figure 10 -- is fed by a filter preventing the return of high-frequency currents. Windings are made with simple fractions of turns, as in the Radio-Cellular Oscillator, which does not involve but a half-turn per oscillating circuit.

"An experimental receiver (see Figure 11) is supplied with a circuit possessing a single square-loop, upon which can be made variable taps. Another apparatus is possessed of a single circular-loop upon which a radial cursor allows to take just the necessary fraction of circuit.

"Works by Professor Esau in the realm of Biology are very recent, because at the August 1928 Radioelectric Con-

Figure 11

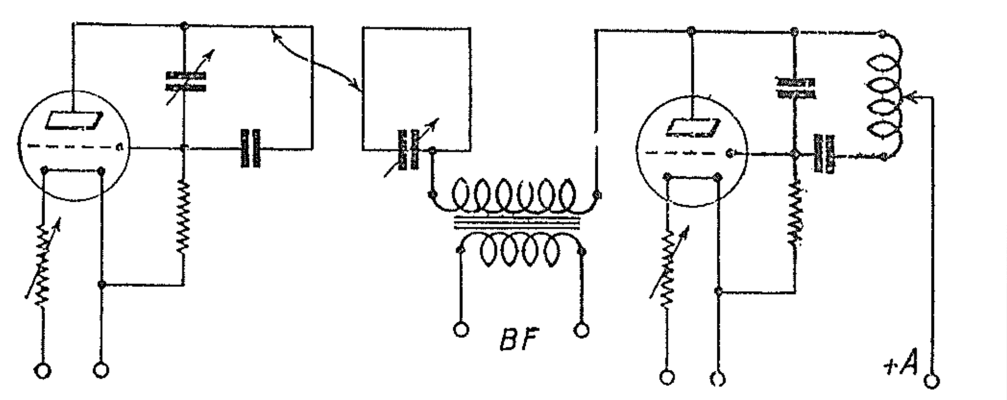

gress of Bremen (see Les ondes ultra-courtes et leur champ d'application -- *Ultrashort-waves and Their Field of Application* -- by Oscar Gadamer in Der Deutsche Rundfunk -- *German Radio* -- of August 31st, 1928, volume VI, page 240, Berlin) this author states that after about a year he investigated the physiological effects of ultrashort-waves that is, according to his definition, waves of very high frequencies between 1½ and 3 meters in length, of which I have indicated how I employ them in a 1924 communication to the Society of Biology (Note: It is on the inoculation and healing of the *Pelargonium zonatum*, as explained on page four of this writing). Prof. Esau then indicated that, since July 1928, he built a receiver of very short waves.

''About this subject, Professor Esau remarked that the main difficulties come from electronic tubes, for he was forced to use foreign tubes, particularly French ones, that are better fitted to the technology of very short waves. I must remind (the readership) that four years before him I studied and had built these special lamps for short waves, to be used in my Radio-Cellular Oscillator.

''I will also point out that an emitter of very short waves has recently been employed by Dr. Franz Seidel of Vienna for destroying bacteria, particularly to sterilize milk (see Les ondes ultra-courtes contre les bacteries -- *Ultra-short Waves Against Bacteria* -- in Radio Welt -- *Radio World*, January 27th, 1929, No. 4, page 108, Vienna).''

We have two definitely contradictory positions in the same chapter dealing with the same type of device:

(1st) In the paragraph translated under Figure 3 of this paper, Lakhovksy emphasized that the elements to produce ultrashort waves (under 2 meters) must be inside the tube. Figures 3 to 8 are dedicated to showing specific constituents for this special type of tube. Though he only speaks on Triodes, there is no reason to suppose he did not experiment with Tetrode and Pentode tubes (plus others of his own design not usually covered in publicly available technical literature).

(2nd) The formerly-indicated conditions change radically when he speaks of his experimental device employed to eliminate tumors affecting inoculated plants, this being the device of Figure 9 of this paper. Notice how the emitting elements in this apparatus are outside the vacuum tube; furthermore Mr. Lakhovsky claims that other, similar, devices developed by Professor Esau at Jena's Technisch-Physicalysches Institut are based in his ''internal'' design when the opposite is quite true!

(3rd) In order to make a better synopsis, I went back to read my copy of The Secret of Life, a translation into Spanish by M. Aquilar (Madrid, 1929). Therein are explained the basics of the Multi-Wave Oscillator but the Radio-Cellular Oscillator is NOT shown. It is reproduced in the 1924 communication to the Society of Biology. Neither source mentioned here reproduced the special vacuum tubes, with internal constituents, as explained in La Science et le Bonheur, the subject of this synopsis.

(4th) In Mark Clement's The Waves that Heal, the Multi-Wave Oscillator is explained as: ''The apparatus consists of a transmitter and a receiving resonator, both arranged so as to set up an electromagnetic field in their immediate vicinity.'' The Multi-Wave Oscillator seems, in my opinion, derived from the ''experimental receiver'' of Figure 11 plus those of Heinrich Hertz (as I have indicated in my Some Free Energy Devices, a BSRF publication, quoting directly from The Secret of Life and introducing complementary information from other sources), rather than from the multiple-tubes shown in Figures 3 through 8 of this paper.

(5th) Considering everything explained from (1st) to (4th), it is my opinion Mr. Lakhovsky produced several types of apparati, along different lines of research, while pretending in his books that only a single-line of investigation was being pursued.

What are the reasons for doing this? There are many:

(*) Trying to disorient his fellow scientists by covering up his actual intentions and actions, so that the actual secret of the apparati rested with him.

(*) Trying to foil any attempt by the ''organised medical profession'' to steal his ''secret'' -- whatever this could actually be.

(*) Having discovered something that he wanted to keep for himself in the course of his researches. This is, in my opinion, the most likely possibility, for Mr. Lakhovsky's researches were well-received in Europe, South America and Asia before WWII began. Even more, he was outspokenly backed by men of science and Academies of great reputation and his enemies were very few (outside of the USA, of course....).

I do not mean to slander him and his researches, I just want to indicate the necessity for greater and deeper research into what he did from all possible sources of information.

VERY IMPORTANT NOTE

In order to facilitate the understanding of the illustrations, I have taken a few liberties to aid the non-French speaking associates:

(1) All the ''lateral letterings'' from Figure 3 to Figure 9 are of my doing, they did not exist in the original text.

(2) I changed into English, keeping the true meaning in translation, the writing in Figure 2.

(3) I made some additional lettering in some illustrations for greater understanding: L1 to L4 in Figure 1; I added the ''C'' in Figure 5, the same ''C'' was added in Figure 6; Finally, I added the ''H'' in Figure 8 for greater clarity.

(4) In French the letters ''HT'' are the abbreviation of ''Haute Tension'' that means in English ''High Voltage''; in Spanish we also use the word ''Tension'' in electricity meaning ''Voltage,'' as the French do. The letters ''BT'' stand for ''Basse Tension'' which I translate not as ''Low Tension'' but as ''Filament Voltage'' for this is the actual meaning.

UNITED STATES PATENT OFFICE

1,962,565

APPARATUS WITH CIRCUITS OSCILLATING UNDER MULTIPLE WAVE LENGTHS

Application November. 13, 1931, Serial No. 574,907

In France, May 2, 1931

7 Claims. (Cl. 250-23)

Fig. 1.

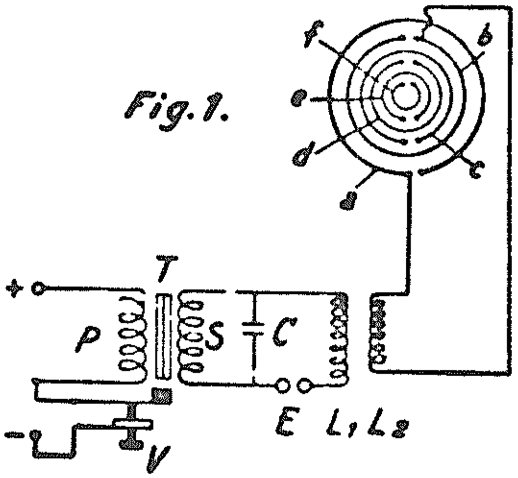

Fig. 2.

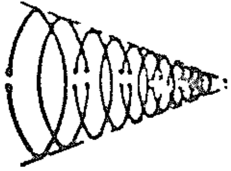

Fig. 3.

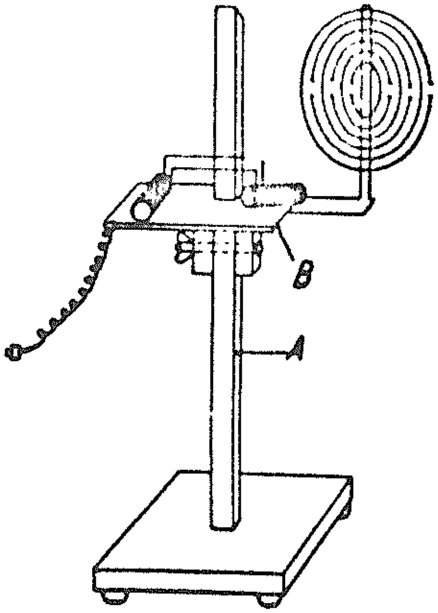

The present invention has for its object an apparatus capable of sending out simultaneously different wave lengths so that among these waves or their harmonics there is found almost always one or several capable of producing the best effect sought for any application whatever.

There is used for that purpose the property of open circuits having well determined wave length if they are excited by electric impulses emanating from discharges of any source.

The radiating part of the system, producing waves of various lengths, is therefore constituted by a certain number of concentric open rings of different diameters, which are maintained insulated by any appropriate means. These rings may or may not be terminated by small spheres forming capacities.

These rings may be arranged either in the same plane or in different planes so as to form any surfaces, such as portions of cones, spheres, paraboloids, ellipsoids, etc..., these surfaces being of a nature to concentrate the waves in a determined direction in the manner of a reflector, for example.

These rings may be fed by any producing device whatever of high frequency, for example, a combination comprising a trembler coil (or any other transformer) and a high frequency circuit provided with a self-induction coil and a capacity.

One of several points of this high frequency circuit is or are connected to the end of one or several of the rings forming the radiating apparatus. The other rings not connected are excited by induction.

The complete apparatus (coil, high frequency circuits, rings, etc....), may be mounted in a box or upon a small board capable of sliding along a guide or rocking so as to be able to be placed in the most convenient position.

In the accompanying drawing, by way of example, not limitative, there is shown:

Fig. 1 a diagram of the system.

Fig. 2 the position of the rings in the form of a cone.

Fig. 3 the apparatus mounted upon a foot or base.

In Fig. 1 there is seen at T the transformer or coil (P the primary, S the secondary of this coil), V the screw of the trembler, C the condenser the discharge of which across the spark gap E produces oscillations in the self-induction coil L1, which in its turn acts by induction upon the self-induction coil L1; this self-induction coil L2 is connected on one side to the first exterior ring a, on the other side to the following ring b. It thus forms an electrostatic field of high frequency which induces oscillations in the other rings c, d, e, f.

As a modification, L2 may be suppressed and the rings a and b may be connected each to one of the ends of L1.

One of the ends, or even any point of the self-induction coils L1 or L2 may likewise be connected to any one of the rings.

Fig. 2 shows the arrangement of the rings in the form of a cone. The open circuits a, b, c....f can be constituted by solid conductors. They can also be tubular circuits, this being advantageous, since, for the same bulk, the radiating surface is larger. Finally, the area of the cross section of each of these various circuits can vary from one circuit to the other for instance, this area can decrease from the periphery (ring a) to the center (ring f) of the radiating system.

Fig. 3 shows the apparatus mounted upon a foot A along which it can slide, in order to be placed at any desired height and upon a hinge B which allows it to be oriented upon the horizon.

I claim:--

1. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters, the split ends being spaced from each other and terminating with small spheres.

2. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and terminated by small spheres forming capacities.

3. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits constituted by concentric insulated split rings of different diameters and arranged in one and the same plane the split ends being spaced from each other and terminating with small spheres, in staggered position as regards adjacent rings.

4. An apparatus of the type described comprising a base, a plate slidably mounted on said base, a support pivoted to the said plate, a series of open and insulated, high frequency radiating circuits mounted on the pivoted support, and a high frequency producing device for the excitation of the said circuits.

5. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits, said circuits being open, insulated, and having dimensions different from each other.

6. An apparatus adapted to produce electric fields of high frequency having multiple wave lengths, comprising a high frequency producing device, a series of high frequency radiating circuits arranged in the same plane, said circuits being open, insulated, and having dimensions different from each other.

7. An apparatus of the type described comprising a transformer, a high frequency energizing circuit fed by the said transformer and provided with a self-induction coil and a capacity, a series of open and insulated high-frequency radiating circuits having dimensions different from each other, and inductive connection between two of these radiating circuits and respectively each of the poles of the outgoing element of the high frequency energizing circuit.

G. Lakhovsky
INVENTOR

By: Marks & Clerk
Attys.

June 13, 1944. G. LAKHOVSKY 2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

Filed Nov. 21, 1941

Fig. 1.

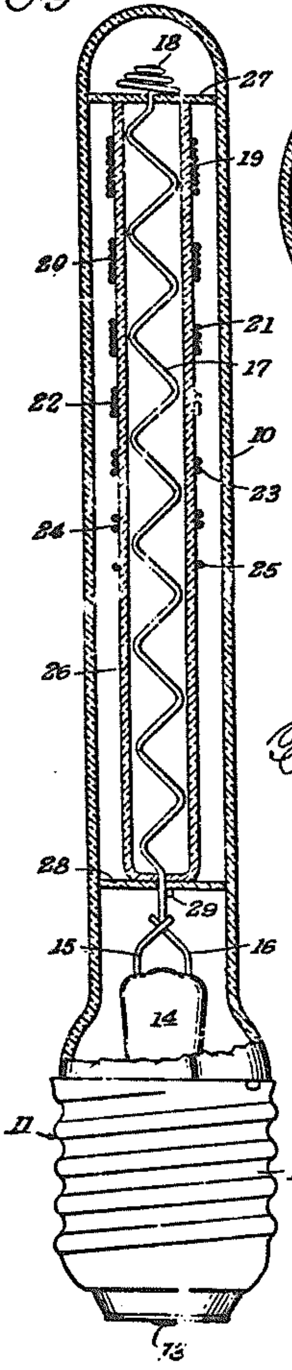
18
27
19
20
21
17
22
10
23
24
25
26
28
29
15
16
14
11
12
13

Fig. 2.

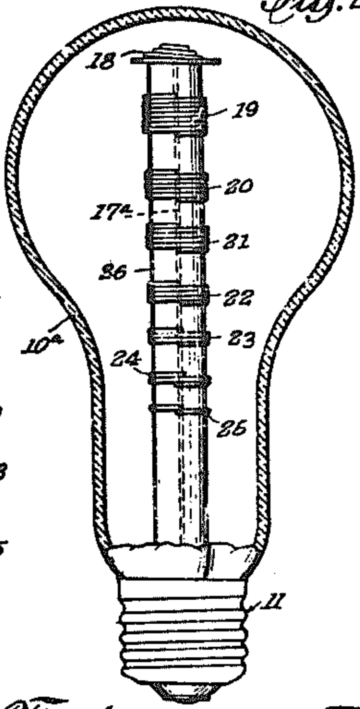
18
19
20
17^{a}
21
26
22
10^{a}
23
24
25
11

Fig. 3.

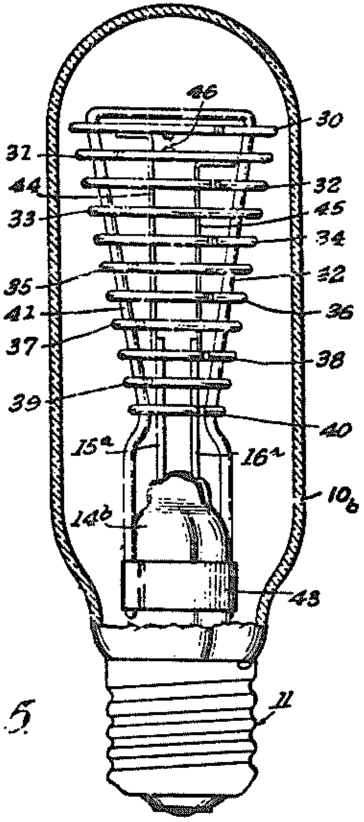
46
30
31
32
44
45
33
34
32
35
36
41
37
38
39
40
15^{a}
16^{a}
10_{b}
14^{b}
43
11

Fig. 4.

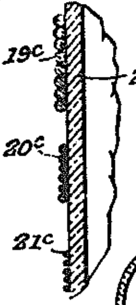
19^{c}
26
20^{c}
21^{c}

Fig. 5.

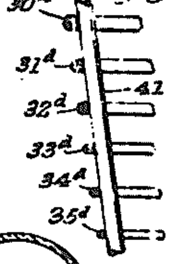
30^{d}
31^{d}
41
32^{d}
33^{d}
34^{d}
35^{d}

Fig. 6.

Fig. 7.

Fig. 8.

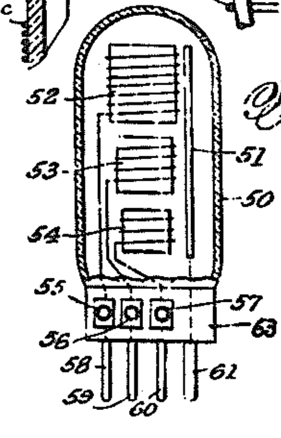
52
51
53
50
54
55
57
63
56
58
61
59
60

Fig. 9.

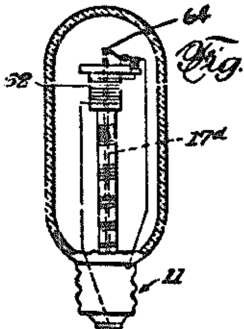

INVENTOR
GEORGE LAKHOVSKY
BY Leon M. Strauss
ATTY.

Patented June 13, 1944

2,351,055

UNITED STATES PATENT OFFICE

2,351,055

TUBE FOR PRODUCING MULTIPLE WAVE LENGTHS

George Lakhovsky, New York, N. Y.; Anne-Marie
Louise Lakhovsky, administratrix of said
George Lakhovsky, deceased, assignor to Henry
S. Blum, Chicago, Ill.

Application November 21, 1941, Serial No. 420,006

11 Claims. (Cl. 250—33)

This invention relates generally to electric devices excited by electric impulses and more particularly to multiple wave length conducting and/or producing means. This invention has for its primary object the provision of such means disposed within a vacuum tube or a tube containing rare gas or gases.

My work over a period of more than 20 years has led me to the belief that there is great need for means conducting and/or producing multiple wave lengths which may be employed in many industrial and medical fields. Suitable apparatus for producing waves of various lengths is disclosed in my U. S. Letters Patent No. 1,962,565, dated June 12, 1934. Embodiments of this apparatus have been in use the world over. Because of the need for simplification in such apparatus so that it may be handy and easily used by persons lacking skill in electrical matters, I have concluded that by providing one unit which incorporates oscillation generating means and another in which the wave lengths are produced, the apparatus may be more universally used in the treatment of cellular life in therapy, in the aging or treating of liquids and in other industrial fields.

The present invention deals with those units, preferably combined with one another and constructed as a portable device and as stated above, has for its major object the provision of a unitary multiple wave length producing and/or conducting device in the form of a vacuum tube.

The present invention further contemplates the incorporation in such a tube, of means for varying the effect of wave lengths, or selectively employing the means therefor enclosed in or forming part of said tube.

This invention further seeks to provide a tube of the indicated type incorporating means for generating oscillations.

Another object of the invention is to provide a single unit which contains an oscillation generator, which produces waves of definite frequency, and further inductance means each permitting emanation of wave lengths of various values.

The structural features of my invention also form a material part of this disclosure, the objects and advantages being attained in structures such as shown in the accompanying drawing, which exemplifies the invention. The following specification, based on said drawing, more clearly points out the purposes and advantages of my invention.

In the drawing:

Fig. 1 is a vertical, partial sectional, partial elevational view of a tube incorporating features of my invention.

Fig. 2 is a similar view of an alternate form thereof.

Fig. 3 illustrates in a similar manner a modified form of my invention.

Fig. 4 is a fragmentary sectional view of a modification made in accordance with the invention.

Fig. 5 is a similar view of another modification.

Figs. 6 and 7 are sectional views illustrating types of wires used in either of the forms of my invention as shown in Figs. 1, 2 and 3.

Fig. 8 is a partial sectional, partial elevational view illustrating a tube incorporating a plurality of grids and selective capacity means for varying the effects of said grids.

Fig. 9 is a similar view of a tube, such as shown in Figs. 1 or 2 and incorporating oscillating or vibrating means.

In that form of my invention shown in Fig. 1, there is provided a sealed preferably transparent envelope 10 made of glass or like material affixed in the usual manner to a screw or Edison plug 11. The latter is conventionally provided with a shell terminal 12 and a central terminal 13. The support or seal 14 of the tube serves to hold wires such as 15 and 16 which connect to the terminals 12 and 13.

Connected to the wires 15 and 16 is a primary inductance member in the form of a loose coil 17 which may be disposed axially of the envelope 10, as shown. The upper end of said member 17 forms a tight wound coil 18 which serves to radiate the strongest induced wave lengths.

Surrounding the primary inductance member 17, there are arranged a plurality of separate coils of wire each being of different length and capacity. Thus, I provide a coil 19 having greatest capacity and successive coils 20, 21, 22, 23, 24, and 25, each having less capacity than the preceding coil. These coils 19 to 25 form secondary inductance means each permitting emanation of a wave length of different magnitude and value than the other.

I may support the secondary coils in several ways. As shown, I prefer to provide a dielectric member such as a glass sleeve 26 surrounding the primary inductance wire member 17 and to mount the secondary coils 19 to 25 on the outer surface of said sleeve, leaving the coil 18 exposed at the top thereof. In order to effectively support and centralize the glass sleeve 26, I provide spacing washers 27 and 28 of mica or the like, top and bottom, and provide a supporting crosspiece 29 on the lower portion of the wire 17.

The tube above described may be employed as indicated for the treatment of cellular structures by connecting it to any one of the rings in the electrostatic field of high frequency, such as shown in said U. S. Letters Patent No. 1,962,565, each of the coils 18 to 25 producing by induction a different wave length, as can be understood.

In the form of my invention shown in Fig. 2, the envelope 10*a* is somewhat differently shaped, and the wire 17*a* straight instead of arranged as an open coil. In other respects, the structure follows that described with reference to Fig. 1.

As shown in Fig. 3, there may be arranged a plurality of open rings 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, and 40 in frustro-conical form, for instance, with the largest of said rings at the top and the smallest one adjacent the bottom. I have shown said rings as supported, for instance, by glass posts 41 and 42, said posts being supported on the seal 14*b*, as by a band 43. I provide extensions for the terminal wires 15*a* and 16*a* and connect said extensions 44 and 45 to the upper ring 30 and the next lower ring 31, respectively.

The rings 30 and 31 may become the primary inductance members, whereas the remaining rings constitute the secondary inductance members as before set forth. Electric current conducted to the suitably spaced rings 30, 31 will cause a spark 48 to be drawn between the rings 30 and 31 providing oscillations emanating from electrical discharges therebetween and causing the radiation of waves of various lengths within the envelope 10*b* and affecting the remaining rings.

In Figs. 1 and 2, I have shown the secondary coils as formed of wires having uniform cross-section. Fig. 4 illustrates how such wires may be made progressively smaller in the succeeding coils 19*c*, 20*c*, and 21*c*, etc. Similarly, I have shown the rings of Fig. 3 made progressively smaller in Fig. 5, said rings being designated by numerals 30*d*, 31*d*, 32*d*, 33*d*, 34*d* and 35*d*, etc. In such various manners, the wave length producing effect of the secondary coils or rings may be arrived at.

The secondary coils or rings may be round or somewhat flattened wire as shown in Fig. 6 or tubular as in Fig. 7, offering a yet greater flexibility of design.

The devices of Figs. 1 and 2 above described may be incorporated in the system shown in said U. S. Letters Patent by connecting one of the rings *a*, *b*, *c*, *d*, *e*, or *f* with plug 11 of Fig. 1, since either terminal 15, 16 is in contact with primary inductance member 17. To this end a conductor (not shown) connects such a ring with said plug or socket.

In Fig. 8, I have shown a modification of the invention which comprises a vacuum tube 50 containing a plate 51 and a plurality of different wave lengths producing grids 52, 53, and 54. I connect each grid with a variable capacity device, such as 55, 56, and 57 carried by said tube 50 and preferably by its base support 63, whereby the grids may be selectively connected into an operating circuit (not shown) by means of the respective prongs 58, 59, and 60, and the plate prong 61. Heating or other exciting means may be incorporated in the tube in a well-known manner. This combination tube 50 is designed to replace a plurality of known tubes each operating with different and variable wave length effect, mere exterior manipulation at said base support brings about change of the capacities 55, 56, and 57 and cutting in one or more of the grids 52, 53, and 54.

The effect of the spark as produced in the form of the invention depicted in Fig. 3 may be further utilized in Fig. 9 in which I also incorporate a vibrator 62 which produces electrical oscillations by a spark intermittently generated between the end of the primary inductance member 17*d* and vibrator armature 64. In other respects the structure follows that of Fig. 1.

It may be noted that any type of spark producing or oscillation generating means may be substituted for the vibrator shown. The tubes shown in Figs. 3 and 9 may be used in the manner described for Fig. 1 by connecting one terminal of the base as set forth to produce an induction effect. These tubes may be also connected across both terminals with a suitable potential to obtain the spark effect.

From the foregoing it may be seen that I have provided tubes in various forms for the purpose of generating and conducting multiple wave lengths. Other forms of the invention may be produced within the spirit and scope of the invention as claimed.

Having thus described my invention, what I claim as new and desire to be secured by Letters Patent, is:

1. A device of the character described comprising a vacuum tube having an envelope and base, primary inductance means, and a plurality of separate secondary inductance means of different length for radiating waves of various lengths extending above said primary inductance means, all said inductance means being arranged within said envelope.

2. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members spaced apart and independent from each other each member surrounding a respective portion of said primary inductance member and within the field of influence thereof.

3. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member axially arranged in said envelope, and a plurality of secondary inductance members each surrounding the primary inductance member and within the field of influence thereof, each of said secondary inductance members having a different wave length effect than the others.

4. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings.

5. A device of the character described comprising a vacuum tube having an envelope and base, a primary inductance member comprising a pair of split rings in inductive relation, and a plurality of secondary inductance members each comprising split rings, said latter rings being arranged in different planes and being of different length to produce a different wave length effect.

6. In a device of the character described, means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of separate wire coils of different lengths on said dielectric means and in the field of inductance of said wire member.

7. In a device of the character described,

means for producing multiple wave lengths comprising a wire member, enclosing dielectric means for said member, and a plurality of wire coils on said dielectric means and in the field of inductance of said wire member, each coil extending over said wire member and having a different number of convolutions than the others.

8. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings.

9. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others.

10. In a device of the character described, means for producing multiple wave lengths comprising two split rings in inductive relation, and a plurality of split rings in inductive relation to the two split rings, the plurality of rings being of smaller diameter than the mentioned two rings and each progressively smaller than the others, all of the rings being arranged in frusto-conical disposition.

11. In a vacuum tube having a primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member and secondary inductance members, a vibrator arranged in the field of influence of said primary inductance member for affecting the induction of said secondary members.

GEORGE LAKHOVSKY.

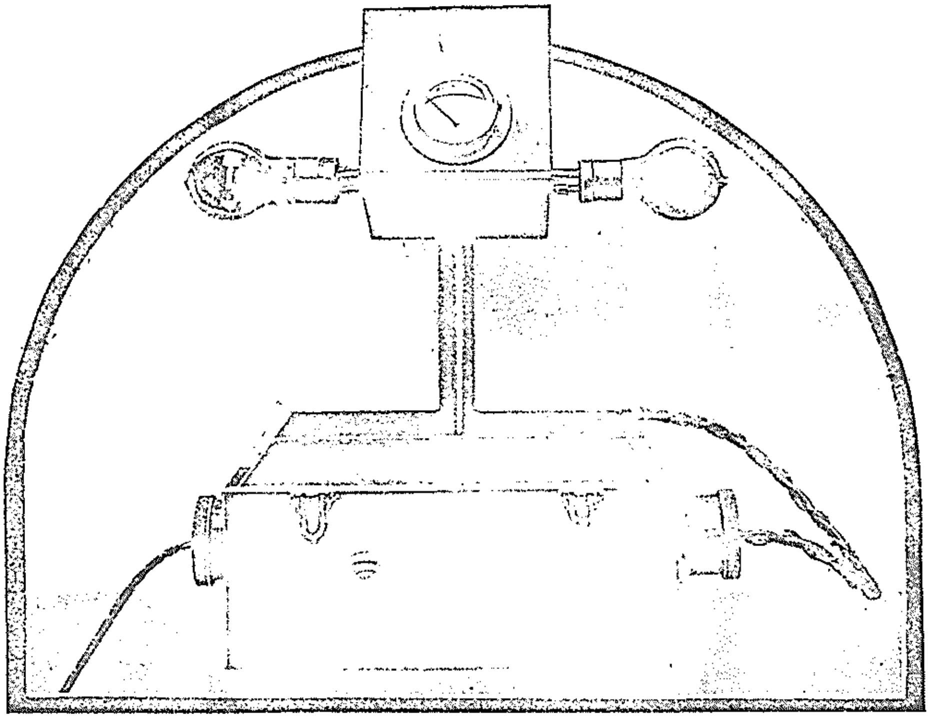

The two-meter oscillator used by the author in his experiments with the supply box containing the necessary controls and transformers.

Picture from *Radio News* for February 1925

Curing Cancer with Ultra Radio Frequencies

by George Lakhovsky

LAKHOVSKY'S VIBRATORY CIRCLE

Author Unknown, Source: BSRF files
Translated from the German by Bodo Capeller, 1988

"We owe to magic all progress in physics and chemistry."
(Jules Lermina, 1915)

The Paris professor with the Russian name Georges Lakhovsky claims to capture, with so called open vibratory circuits (circuits oscillants), cosmic waves and pass them on to humans (neck antenna) and plants. Healing of people has been verified by French and Italian medical practitioners, healing of Cancer injected plants has been proven through photographic records.

The vibratory circles are manufactured from heavy copper wire with a diameter of 5-10 millimeters. Those wires 5-8 millimeters in diameter had a more pronounced effect than the much thinner variety. The spiral (of the open ended circle), which can be single or double, is guided around the plant and is supposed to be insulated against its support by "Ebonit" (1).

Instead of copper wire medical practitioner Dr. Ernst Busse (Garmisch, Germany) recommends the use of thin copper tubing: "because of the short wave length of cosmic vibration only the surface effect (skin effect) is effective (that means electricity is only present on the surface of a conductor, not in the middle) which requires the use of maximum surface area. Besides this the tubing is easier to work with and has less weight. (Private Correspondence 1/10/1953)

Helmut Friedensdorff (Berlin, Germany) did report already in 1932, from himself as well as others, on the effectiveness of an open vibratory circuit against lumbago among other ailments. He inserted a simple isolated copper wire into surgical tubing and positioned it at the problem spot in a manner that both wire ends did not touch (2).

Canon Smythe of Great Britain did proceed in a similar manner: "The only tools required are a one meter length of flexible electrical wire and a sidereal pendulum. The insulation is removed at both ends for a length of 2 centimeters, the wire is stripped and the wire strands are removed except a single strand. Then the wire is laid out on a table and the pendulum is positioned over the wire ends. It will rotate over the positive end in a clockwise motion, over the negative end in reverse. Then one proceeds to make a knot at the negative end for identification purposes. It should be positioned on the left side of the body or in the vicinity of the leg, the bare wire pointing to the left. The wire may be installed over the clothing and can be braided or fastened with safety pins, the ends should point into the proper direction.

"Arthritis in one toe should be treated by using the wire like a stocking holder wound around the leg where the knotted end will point to the left. The author makes no claims for the healing arthritis but reports a lessening of pain.

"With severe headaches the wire is wound just above the eye level around the head so that the negative end rests over the left temple.

"Severe coughing is treated with one wire around the throat, another one around the abdomen."

A just started cold is removed by our associate in a few hours with two wires, one with the positive connected to the negative end of the next wire joined as a 2 meter long wire. "I wound the wire over the shirt under a sweater around the chest until I reached under the arm pits and made sure that the negative end is on left side of the chest pointing to the left, the positive end pointing the opposite way."

Smythe published this in a paper many years ago called "The Other Half of Medicine". He is supported by the research of a recently deceased friend, Dr. Martin of Airaines, France who had been promoted to the medical faculty in Paris in 1932 with a dissertation "Diagnostic Radiesthesique en medecine veterinaire" under the tutelage of professor Nathering--without an examination! In 1935 a Dr. Girard presented a similar thesis before the University at Lyon.

Medical practitioner Arnold Mannlicher (1880-1958) from Salzburg carried in his rear pocket an open vibratory circuit which also acted as a cell detector.

Vibratory circuits proliferated after the announcements by Lakhovsky's discovery as "throat antennas, broadcast jewelry and healing chains" in the market place. I cannot say at the moment if they were of the open or closed kind. Apparently between the two kinds exists a fundamental difference in their effectiveness, open circuits reduce the tension within the body, closed circuits raise it. A closed broadcast chain can easily be transformed into an open one when it is opened and is closed with an insulator inbetween (rubber band, silk thread).

The open circuit is nothing new at all. Dr. Beck, former head of the municipal children's clinic in Bayreuth says: "the open bronze rings of the Hallstadt area from grave sites can be understood as protection against radiation effects very similar to the rings worn by the Negro tribes around the arms, legs and throat." The ancient Germans wore "broadcast jewelry"--in order to coin a modern word--an open solar spiral. Dr. Heermann has revived this. He writes: "invigorating liver and intestines the solar wheel acts like a plated iron spiral with the windings running from the center clockwise."

Smaller--sown into green material--is my usual aid in stomach ulcers and swollen glands. Underneath his illustration (not shown) the researcher explains: "a spiral ending growth rays to the abdomen" (3).

The never tired Dr. Busse since 1954 has also tried out open rings beside vibratory chains and belts. Earlier he replaced acupuncture needles with metal foil band aids and states now in a letter dated 4/19/1954: "instead of metal foil one may arrive at the same effect by wearing open copper rings on certain fingers and joints. When you put a ring that is cut open or manufactured from heavy copper wire on the corresponding finger an immediate change of pulse at the radial pulsation can be observed".

One of his associates who experimented with these open finger rings was surprised by the achieved results.

I am very much indebted to this versatile medical practitioner.

Because copper rings are a tuner (strengthening, invigorating) it was required to find a second metal which reduces organic functions. It turned out to be aluminum. On closed rings it was already reported in the chapter "Amulets". As far as the application of aluminum is concerned the welder Herbert Bauer (born 1921) from Freiburg, Germany has manufactured many years ago the so called "Zelemente".

According to the description it was recently manufactured from a three metal alloy which reminds of the "seven metal (electrum magicum)" of Hermetic fame. We are dealing here with closed chains that are supposed to protect from dangerous radiations of technical and cosmic nature and are therefore named radiation protective antennas. They look flat and are meant to be moistened in the morning and evening.

Whatever the effect is of the vibratory circuit on the world of plants has been shown in 1905 when sick elm trees were healed on a Hamburg street.

The Vienna engineer Lambert Binder (born 1905) caused, during a growth experiment in 1944 on orange trees, accelerated growth by surrounding the tree with an isolated wire spool.

R.E. Esspiau tried out in 1954 the following experiment with the Lakhovsky antenna arrangement. He connected the first on a three stalk head of cabbage to free standing antenna, the second to ground the third one to water which was replaced every 24 hours. In less than a week the first buds began to form, within four weeks other stalks were formed and developed leaves while a control plant withered away (4).

PHOTO TEXT:

Illustration A: This illustration shows a Lakhovsky experiment on a Cancer infected geranium plant. The cancerous growth is clearly visible. Pelargonium zonatum (Geranium) inoculated with Bacterium tumefaciens on April 10th, 1924 and photographed June 6th, 1924. It will be observed that the stem of the plant bears a massive tumour.

Illustration B: An astounding discovery by Prof. Lakhovsky: a free standing antenna ring manufactured from copper that surrounds the geranium caused the plant to drop the cancerous growth (resting on the rim of the container) and made the plant whole again. Photographed June 1925.

Illustration C: The test plant three years later. The antenna ring has completely healed the plant and protected it from other damage. To the side are two unprotected similar control plants of same age that became victims of the same disease. Photographed on March 13, 1928.

Picture credits: These three illustrations are designated tables II, V, and VII in Lakhovsky's Secret of Life.

BIBLIOGRAPHY

(1) Lakhovsky, Georges: Das Geheimnis des Lebens (Secret of Life), Kosmische Wellen und vitale Schwingungen, Munich 1931. Schroedter, Willy: "Der Drahtring um die Geranie, Frankfurter Ilustrierte, #15, 1932, page 356.
(2) Frensdorff, Helmut: Einfache Mittel gegen Hexenshuss und Gliederschmerzen, "Die Zukunft", #31, Berlin 1932.
(3) Heermann, Alex, M.D.: Neues von Strahlen, Stroemen und Wellen, Bad Ausee 1935, page 13.
(4) "La Radiesthesie pour Tous", December 1954, Paris. Schroedter, Willy: Kosmische Strahlen und vitale Schwingungen, "Mensch & Schicksal", #5, 5/15/1953, Villach, page 7

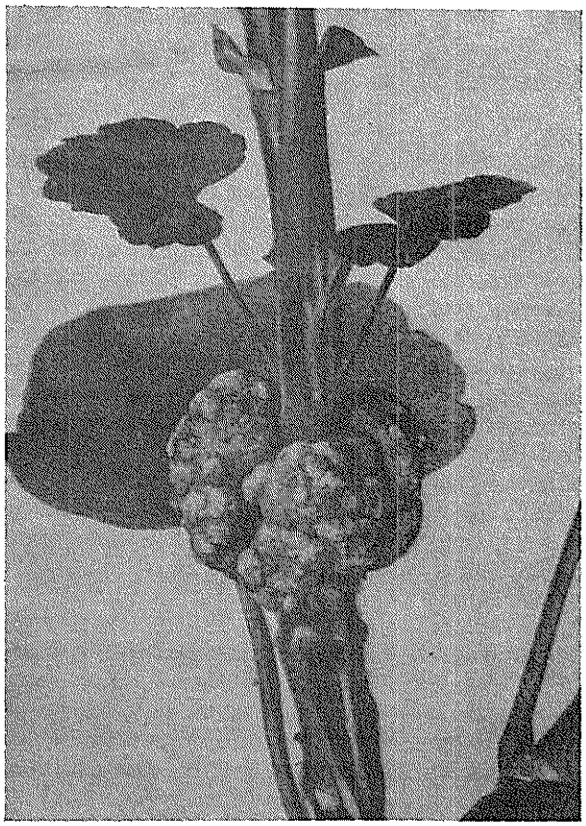

Illustration A

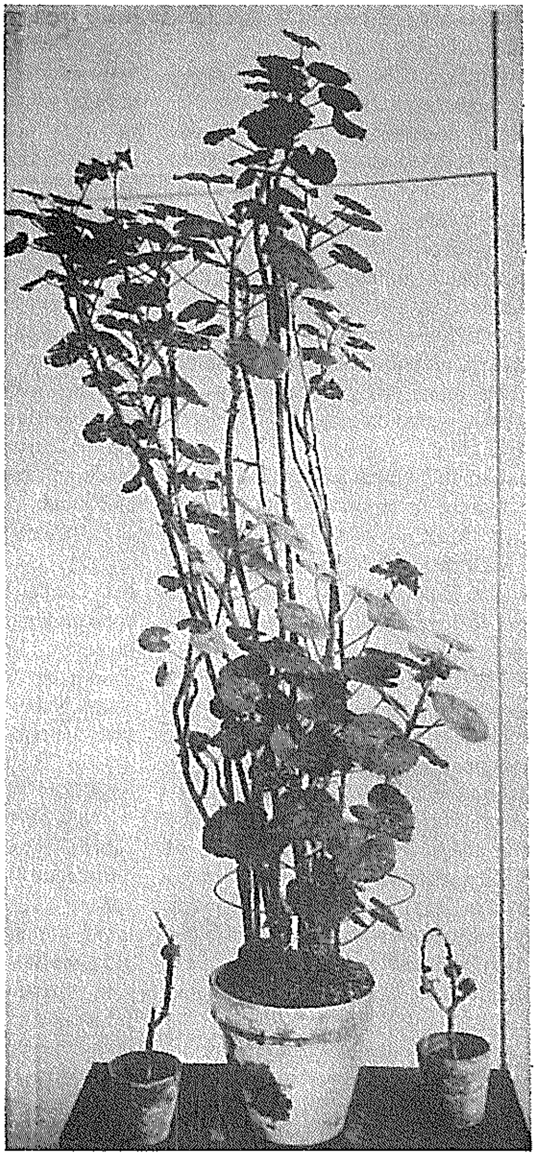

Illustration C

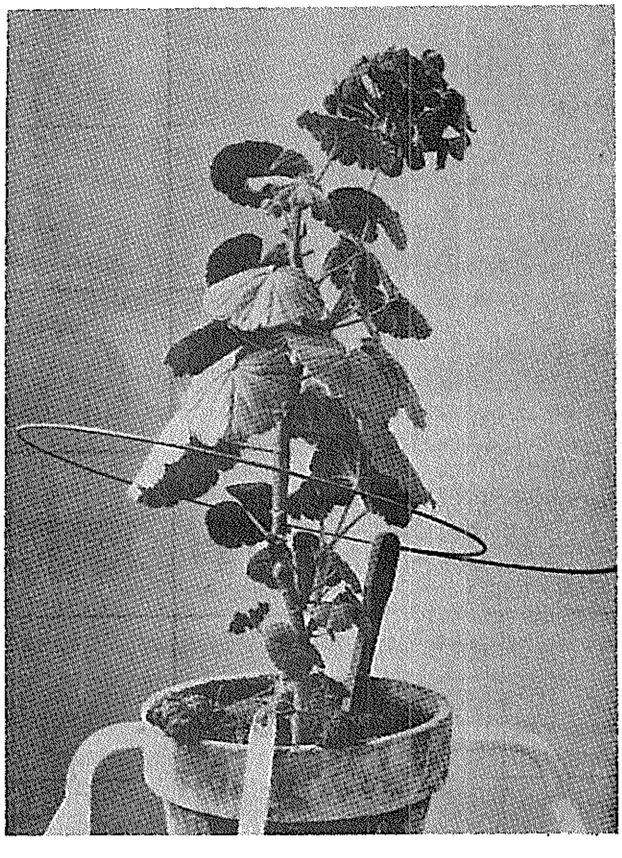

Illustration B

THE RUSSIAN LAKHOVSKY REJUVENATION MACHINE

by Bob Beck

from the November 1963 Journal of Borderland Research

In September 1963, Dr. L.O. Anderson, Anderson Research Center Director and BSR Associate, phoned me about a "marvelous electronic machine" purported to "cure" almost any human ailment. I've looked into dozens of these "modalities" every year; so with extreme skepticism I gathered instruments and took off to check another possible hoax.

I found a heavy spark-gap Tesla Coil inside a four-foot cabinet. This console was rigged to circular "antennas" and was partially shielded inside a home-built Faraday cage in the living room of Dr. Anderson's friend. It hadn't functioned in over a decade, and had been given up by some twenty-odd medical instrument repairmen and engineers as unfathomable.

The machine, which had no nameplate or other identification, turned out to be a Lakhovsky "Multiple Wave Oscillator", or MWO. Fortunately the present owner had found patents and two books describing its operation. We were able to reconstruct the circuits and after several hours had it operating perfectly. The machine was not a hoax.

Lakhovsky did his research prior to world War II and died in 1942. No infringement can be involved in duplicating Lakhovsky's work, since his U.S. Patent #1,962,565 expired in 1951 and is in public domain. I have found from the book **The Waves that Heal** by Mark Clements that MWOs are extensively used in hospitals in France, Germany and Italy, and although tested in a New York Hospital with fantastic success, MWOs remain practically unknown in the United States.

MDs testing Multiple Wave Oscillators reported up to 98% cures of terminal cancer, arthritis and other "hopeless" conditions. Case histories were complete with before-and-after photos. Often rejuvenation was demonstrated, not only in human beings but in plants and animals as well!

So, on Sept. 18, 1963, I designed and built a modern MWO based on Nikola Tesla and Lakhovsky's original work and patents. The transistorized instrument fits into a brief case and weighs under 14 lbs. Experiments with the MWO on some eighteen volunteers have demonstrated almost unbelievable results. At this writing, Oct 20th, every single subject, including myself, reports fantastic gains.

My mother, age 72, who has suffered from hypertension and glaucoma for many years, measured absolutely normal blood pressure the morning after her first treatment. And the eye tension, tested by her MD the week following, had dropped to a lower point than in the previous two years. Another woman with a long-standing back condition says she was permanently relieved in one 15 minute exposure. An asthmatic with chronic bronchial complications claims he is now symptom-free after two exposures and her husband with a bad dermatological condition of seven years history appears to be very definitely improved. Perhaps the most dramatic is Mr. N.V.C. who had broken his leg skiing last December. Because of an old polio complication the broken leg had never healed and flesh surrounding the fracture was in a very poor, waxy condition. After one exposure he phoned to say he had regained feeling in the area. The next day we made an "after" photo slide of nearly normal tissue.

A friend suffering allergies of long standing says she can now eat anything and her MD admits he as startled by her youthful appearance and new vitality. The Lakhovsky literature illustrated many such rejuvenations as an unexpected side effect of cancer cases. Subject's wrinkles disappeared along with the tumors!

THEORY OF OPERATION

Lakhovsky pointed out that all cells capable of reproduction contain in their nuclei "filaments" of highly conductive material surrounded by insulating media. This filament, which may be the RNA-DNA complex, is always in the form of a spiral or helix -- in other words, a coil. Therefore, each will react as a tuned circuit if its resonant frequency can be approximated by an external oscillating coil.

Lakhovsky didn't carry this to it's conclusion; however, I postulate that by exciting the nuclei with electromagnetic energy a "charge" can be induced by the long established principle of electromagnetic induction. This demonstratively raises the energy level and perhaps the vitality of every cell in the field simultaneously. Since each cell is an individual, and of slightly different physical dimensions, the exciting wavelengths must be multiple, and must span a broad frequency spectrum. Diathermy machines, limited to crystal-controlled single frequencies in the 27.255 MC region, can do nothing but heat the tissue; and yet this approach, abandoned by Lakhovsky in the 1930s, can still be found in "modern" doctor's offices!

The Electromotive Force (EMF) produced by the MWO and induced in the cell nucleus, can raise the cell's metabolic rate by Electrolysis, and perhaps jog the RNA-DNA "memory" and reproductive capabilities to their level at an earlier, younger age, thus the rejuvenation. Even more subtle changes might be postulated, such as magnetic "progression" effects as evidenced by heavy water in magnetic fields.

By contrast, any chemical taken into one's system by mouth or by injection -- such as vitamins, food or medicine -- must first be absorbed in proper solution and ratio with unknown tract elements, then be transported by the blood to affected tissue only to be possibly rejected by the target areas.

Perhaps in cancer the EMF induced by the MWO raises the vitality and memory of marginal cells to normal reproduction levels. In the case of other disease, perhaps a cell's given higher energy levels can more readily throw off affliction.

A WAVE FOR EVERY BODY CELL

The Multiple Wave Oscillator described here radiates a bandwidth of radio frequency (RF) energy from the audio frequencies up beyond microwave frequencies. By actual measurement with standard field strength meters, this vast bandwidth of frequencies and harmonics can be shown. In fact, a bluish glow of "brush discharge" surrounds the antenna when operating. A fluorescent lamp held anywhere within several feet of the subject glows brilliantly. Within this multiple-wave range of frequencies, every cell in the body can find its ONE resonant frequency and absorb energy at its own natural wavelength.

Obviously the Electrostatic energy cannot penetrate the body. This is known as the "skin effect". However, the Electromagnetic component of the energy can and does permeate and will induce an EMF in each cell. It is precisely this energy to which Lakhovsky attributes his almost miraculous "cures".

A word of warning, although no deleterious side effects have ever been reported for human reactions, the MWO has input power exceeding 35 watts. It radiates RF interference over a bandpass of 15 Kilocycles to well beyond 250 Megacycles. The Federal Communications Commission has stiff penalties for unlicensed transmitters and radio and TV interference -- not to mention making all your neighbors very unhappy. We operate our machine in a shielded room, and the battery input prevents RF interference from feeding back into the power line.

The deluxe MWO diagrammed here can be built by any intelligent 16-year old with readily available electronic parts. I do not plan to build, manufacture, sell or otherwise exploit this device, as my interest is purely in research. Several Associates given this circuit have successfully constructed MWOs. We would certainly like to hear from any BSR Associates who undertake this project, but please do not communicate until after building the unit. I cannot correspond over details which any intelligent person would be able to thresh out with his or her own technical friends.

At one stage we had planned constructing a much higher-powered unit but the results with the "portable" are so gratifying that Model Two was shelved. Extensive tests are scheduled with plants and animals, and every effort is being made to remove any element of "suggestion" from results. A physician friend is cooperating with the project by sending a few "problem" patients for test. It is much too soon, however, to make a finished statement concerning the MWO's potential. No cures can be claimed even if obviously demonstrated, since our laws are controlled by medical unions forbidding the use of the word except by MDs. But I have satisfied myself with nearly twenty miracles in these few weeks of tests that the MWO might well be the most promising device to come to light in many years.

I have one request, that no one building this device ever be tempted to charge a fee, donation, or reimbursement of any kind whatsoever for its use. I feel most strongly that any device potentially helpful to suffering humanity should be available completely free of charge to anyone needing it.

CONSTRUCTION TIPS

My MWO is built on an 11 x 11½ inch plastic panel. Components extend 3½ in. below this, so the unit fits nicely into a small attache case. The 6 volt battery was a Sonotone lifetime Nickel Cadmium unit; however, five war-surplus NiCads (1.2 v each) wired in series would be less expensive.

My first unit used a transistorized driver for the Tesla Coil but the Ford coil works equally well, is far cheaper, but has the disadvantage of making a buzzing noise. The Ford Model T coils may still be available at Pep Boys stores in California, or can possibly still be purchased from J.C. Whitney & Co., 1917 Archer Ave., Chicago, IL 60616. The spark gap was made from salvaged tungsten relay contact points about 1/4 in. in diameter. (Chevrolet Six timer points, 1950, do very well). The relay spring arm can be tuned by a Nylon 6-32 screw adjusting its natural spring tension. The gap must be carefully tuned for maximum output for every individual, since the subject's absorption seems dependent on his body fluids and weight. *Adjust for maximum sparking at random times and intervals between several of the inner antenna rings.*

The automatic timer, zero to 15 minutes, the meter, the internal battery charger, the RFC circuits, etc. are all simply conveniences and may be left out of the circuit without in any way subtracting from results. By not building in these deluxe features, the entire MWO will cost much less to build. Use plastic instead of metal screws around the high voltage coils. High voltage capacitors can be tracked down at surplus stores or well stocked electronics parts sources. The Tesla Coil must be hand wound as it is not available commercially.

We have determined experimentally that a workable pair of antennas, radiator and resonator, can consist of 16 concentric rings, ranging from an outside diameter of 10-3/4 inches down to 1/2 inch. These were constructed by cementing a one-foot square of

BECK ADAPTATION OF LAKHOVSKY MULTIPLE-WAVE OSCILLATOR, U.S. Patent #1,962,565

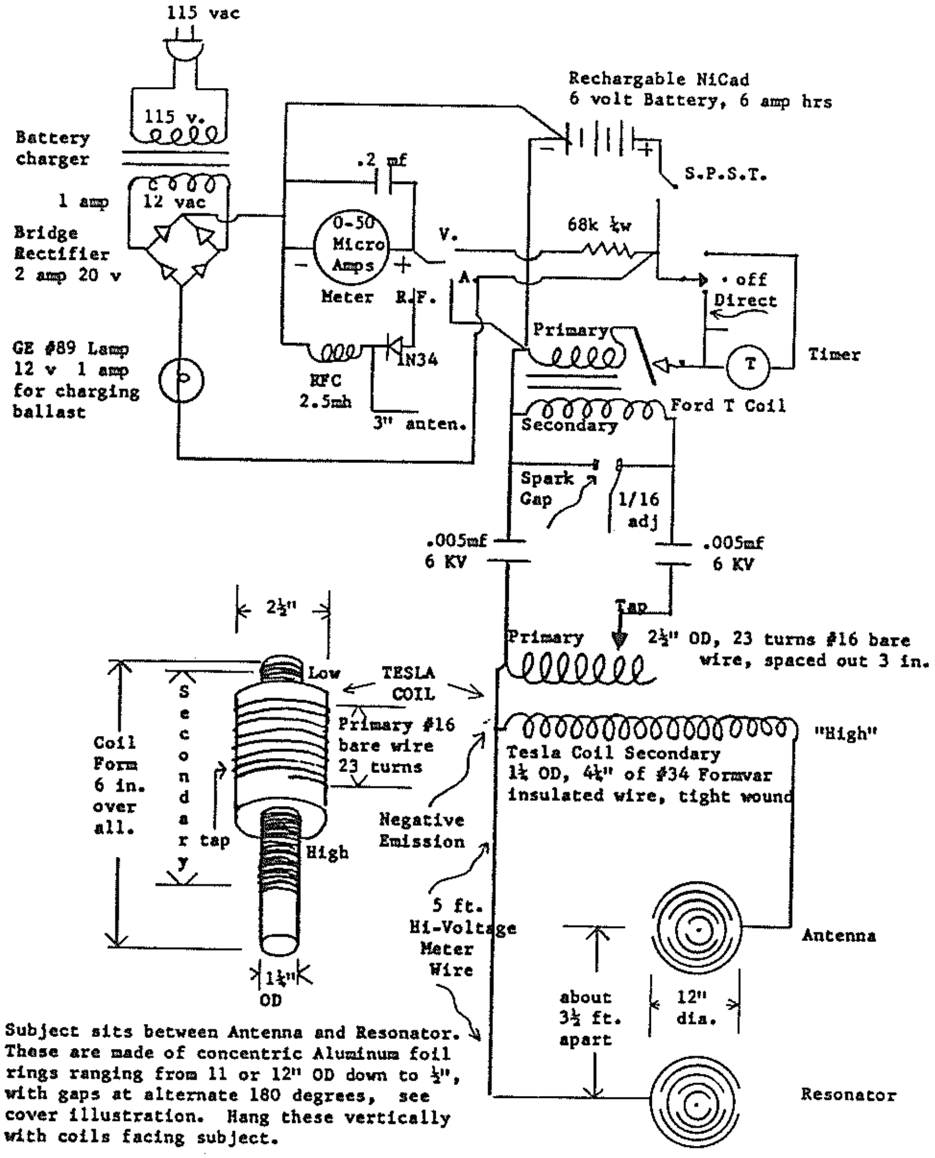
115 vac
115 v.
Battery charger
1 amp
12 vac
Bridge Rectifier 2 amp 20 v
GE #89 Lamp 12 v 1 amp for charging ballast
.2 mf
0-50 Micro Amps
Meter
R.F.
V.
A
RFC 2.5mh
IN34
3" anten.
Rechargable NiCad 6 volt Battery, 6 amp hrs
S.P.S.T.
68k ¼w
• off
Direct
Primary
T
Timer
Ford T Coil
Secondary
Spark Gap
1/16 adj
.005mf 6 KV
.005mf 6 KV
Tap
Primary
2½" OD, 23 turns #16 bare wire, spaced out 3 in.
"High"
Tesla Coil Secondary 1¼ OD, 4¼" of #34 Formvar insulated wire, tight wound
2½"
Low
TESLA COIL
Secondary
tap
Primary #16 bare wire 23 turns
Coil Form 6 in. over all.
High
Negative Emission
5 ft. Hi-Voltage Meter Wire
1¼" OD
about 3½ ft. apart
12" dia.
Antenna
Resonator

Subject sits between Antenna and Resonator. These are made of concentric Aluminum foil rings ranging from 11 or 12" OD down to ½", with gaps at alternate 180 degrees, see cover illustration. Hang these vertically with coils facing subject.

Original cover providing template for aluminum MWO antenna.

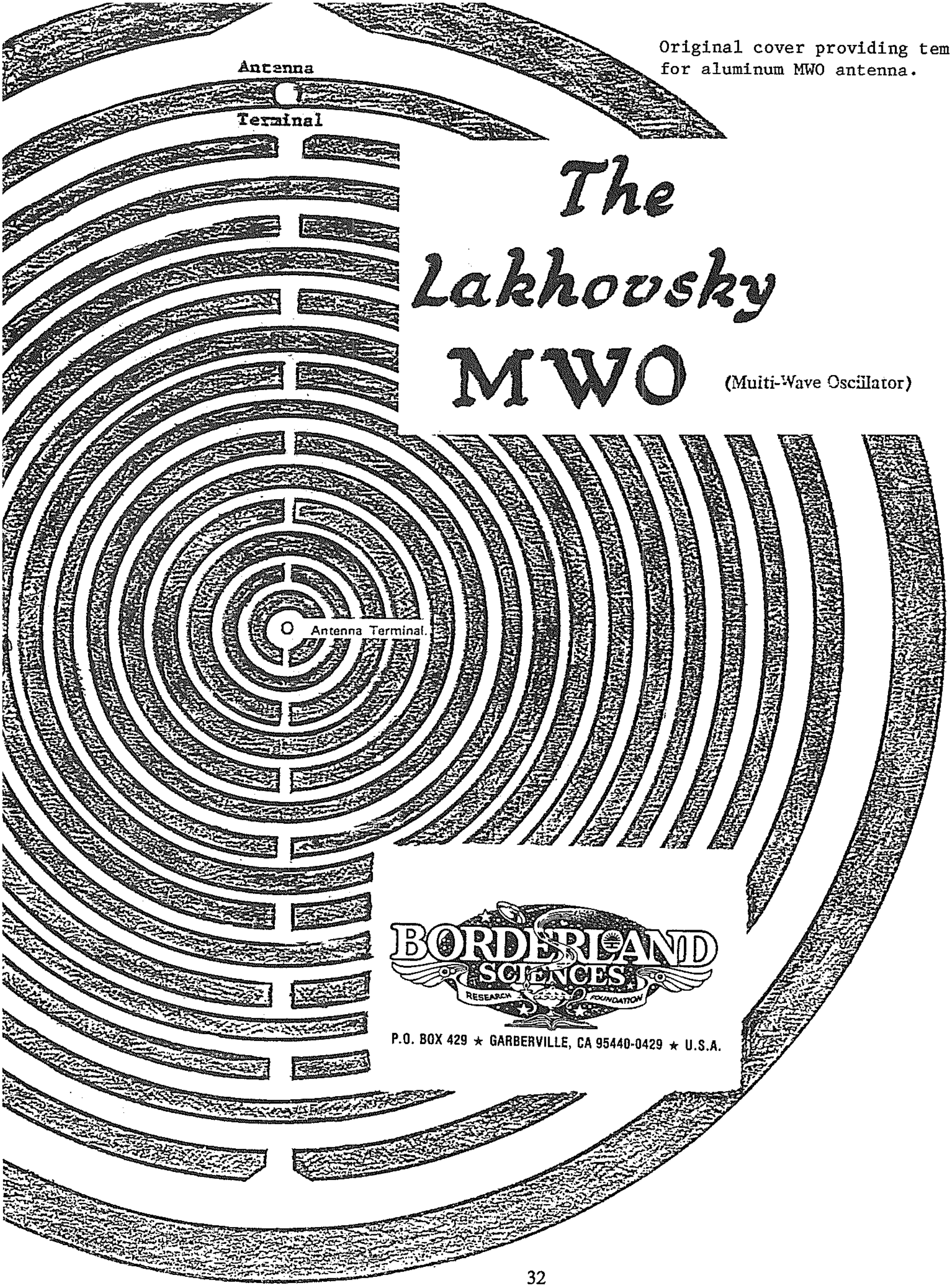
Antenna Terminal
Antenna Terminal.

The Lakhovsky MWO

(Multi-Wave Oscillator)

BORDERLAND SCIENCES
RESEARCH FOUNDATION

P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

Reynolds kitchen aluminum foil to an 11-inch square of thin fiberglass. The outline of each set of circles was carefully scribed with a sharpened compass, razor or Scripto blade. By scoring or cutting the aluminum foil was cut through to the plastic backing and the waste area strips pulled away before the cement hardened. Excess cement was removed with a solvent. The gaps were then cut with the razor. Note that gaps are alternated in position each 180°, as per drawing on original cover page. This was drawn from an actual antenna and you can set your compass distances for each ring from it. Keep gaps between inner rings narrow to encourage random arcing between rings. Trim the corners from the gaps of the three or four outside rings to discourage fixed arcing at these places. Also be sure to clean away any bits of foil between rings. These also will cause fixed arcing.

In use, the ring second from the outside is connected to the high side of the Tesla secondary. (Refer to "antenna terminal" on original cover page illustration) This is the radiator coil. The arcing from this ring to the outside ring causes the "shock excitation" of the fourteen other antenna elements. Since each radiator is of different diameter, the array spans a broad, multiple-wave segment of the Radio Frequency (RF) spectrum. The other antenna element, called the Resonator, is connected with a five-foot, high-voltage meter wire to the ground side of the Tesla coil secondary.

PUTTING THE MWO TO USE

The antenna and resonator are suspended from non-conductive supports about three feet apart, such as wooden chairs with no nails or screws. The subject of the experiment, preferably with all metal removed from their person, including belt buckles, zippers, hooks, fasteners, coins, hairpins, earrings, rings, etc, is seated on a wooden or plastic rest between the antenna and resonator (hung vertically, the foil rings toward the subject. Unless the Resonator is powered by a separate Tesla Coil it will not arc. Metal on the subject will produce eddy currents in the field of radio frequencies, reducing their effectiveness.) Experiments should last from nine to over 15 minutes. A four-day interval between experiments is recommended. Four such experiments over 16 days should suffice to demonstrate anything.

Warning, two-inch sparks can be drawn from antenna; so keep subject at least three inches away. If touched a mild and harmless sting can be felt. HF effects can be demonstrated in a darkened room by holding a fluorescent lamp anywhere near the subject. The RF energy will cause the lamp to light although no wires are touching it. Field strength measurements read several watts of power from the audio frequency spectrum to beyond 250 megacycles.

The illustration to the right is from Clement's **Waves That Heal**, Lakhovsky's illustration of a living cell. "In the centre is a twisted filament which, possessing self-inductance and capacity, constitutes an oscillating circuit. The similarity to a short-wave circuit is manifest; the filament shown here oscillates like a coil having a very small number of spirals." He quotes Lakhovsky: "Every living being emits radiations", and adds: "the living cell can thus play the part of a transmitter or receiver of exceedingly short radio-electric waves which give rise to high frequency currents in the circuits of the cellular nucleus."

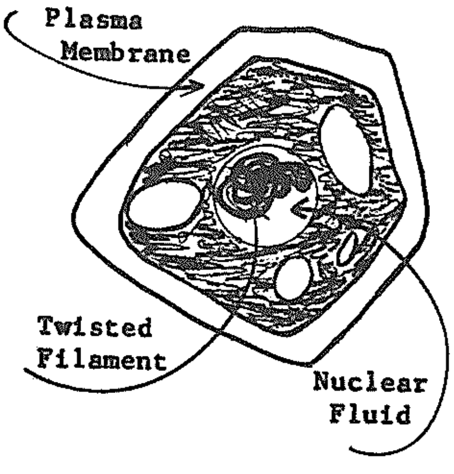

THE FOLLOWING PHOTOS were kindly furnished by Borderland Sciences Research Associates from their own experimental work with the MWO during the years of 1964 and 1965.

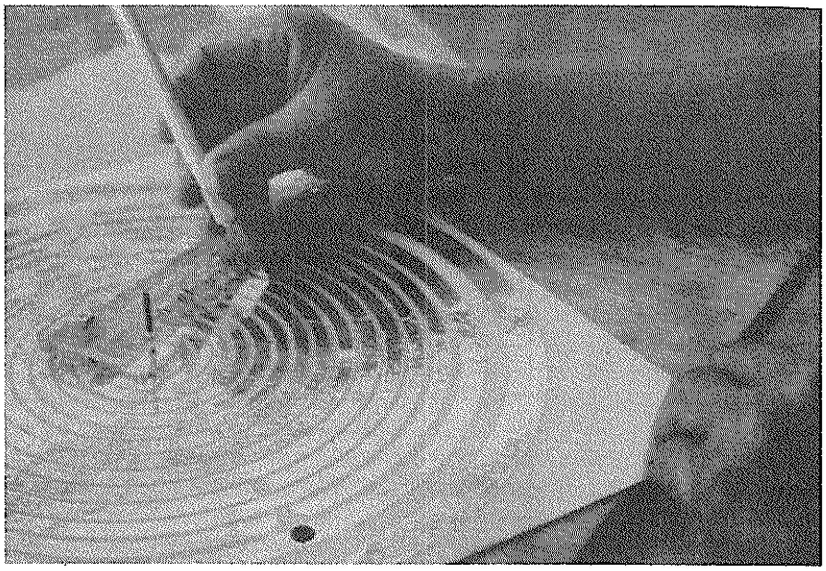

THE PRINTED-CIRCUIT ANTENNA foil is glued to an 11-inch square of 1/16th inch thick fiberglass. Spray one surface of the fiberglass with Hobby Shop spray glue and stretch smooth-surfaced kitchen aluminum foil across it. Press the foil on firmly from the center out, smoothing out repeatedly with a soft cloth. Any remaining air bubbles can be released through a small pin hole. The circles are more easily cut if you make a jig with a 16-inch square of plywood with a headless nail for a center post. Using the antenna pictured on the front cover as a pattern, mark and notch an old ruler or stick where each circular cut is to be made. This marked guide ruler is mounted over the antenna on the plywood jig as shown at right. The foil-covered surface is then turned clockwise with the left hand against the Scripto blade held firmly in the successive notches by the right hand. Be careful to cut through the foil. If you don't, in lifting the cutouts afterward, you may also lift or tear the rings. If a metal grommet is placed in the terminal hole in the second ring this will prevent the high-frequency arc from burning the foil and fiberglass during operation. Copper foil can be used for the printed-circuit antenna rings but deposits build up on it from arcing, wherever it occurs. There is no deposit buildup on aluminum foil.

THE RADIANT ENERGY of the MWO is obvious in this three-minute time exposure during operation in a darkened room. Note the brush discharge of etheric energy all along the edge of the outside ring, going three or four inches. Look at the random arcing between the antenna rings, down to the eighth ring! This antenna received full power by being mounted on the MWO case, directly at the positive end of the Tesla Coil.

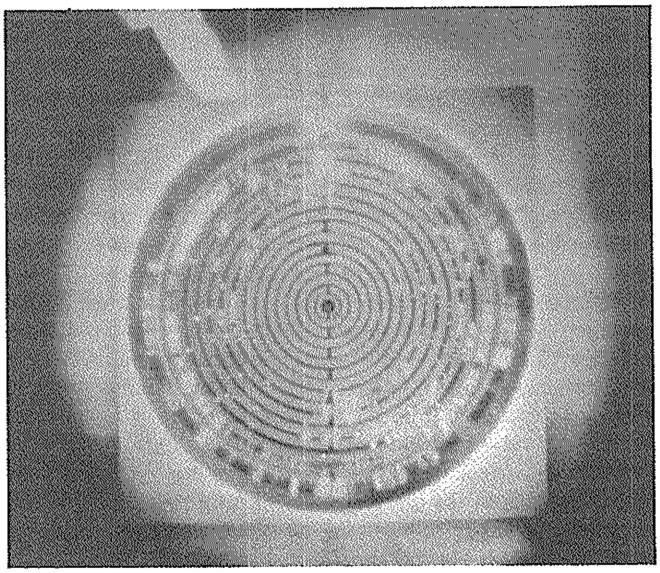

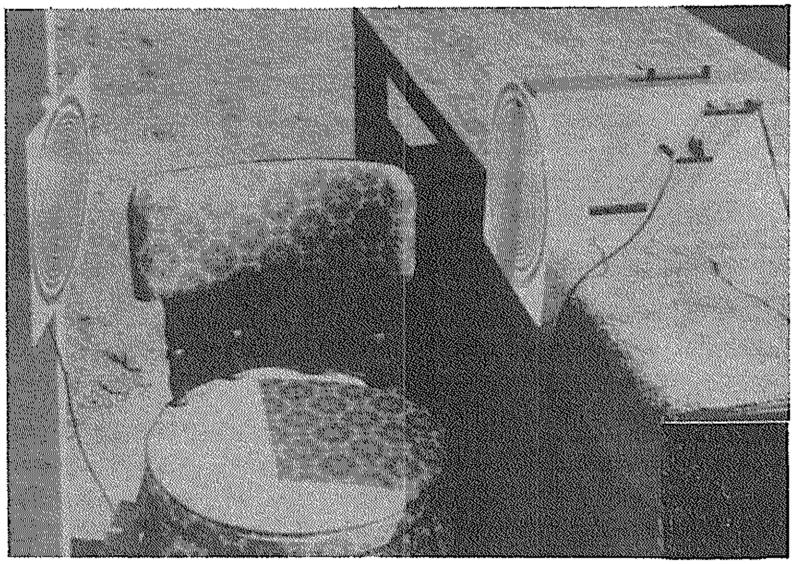

AT LEFT IS A TYPICAL EXPERIMENTAL SETUP, with a chair for "guinea pig" between antenna and resonator. Beck's original equipment called for an 18-inch lead wire from the positive end of the Tesla coil to the antenna. It was found that there was too much loss of power along this wire. The later development shown above has the MWO set upright, with the antenna mounted directly against the positive end of the Tesla Coil in the case.

THE INTERIOR HARDWARE LAYOUT is shown at right in this photo of another MWO, neatly packaged for an attache case. A. The five Ni-Cad batteries, 1-1/4 volts each at full power. B. Timer. C. Tesla Coil. D. 115VAC transformer (for house current operation if desired). E. Spark-gap using modified Chevy Six points. F. Model T Ford Coil. G. Battery Charger.

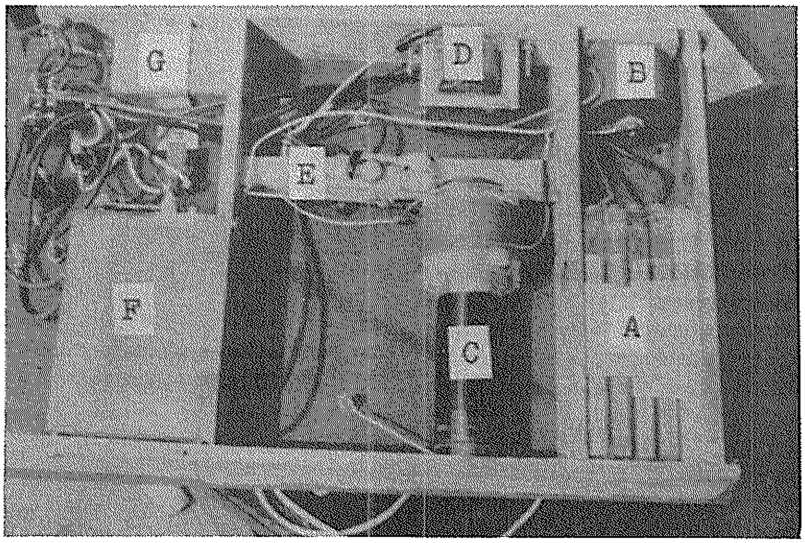

RECTIFYING THE FORD COIL OUTPUT using a 1B3GT or 1G3GT high-voltage tube.

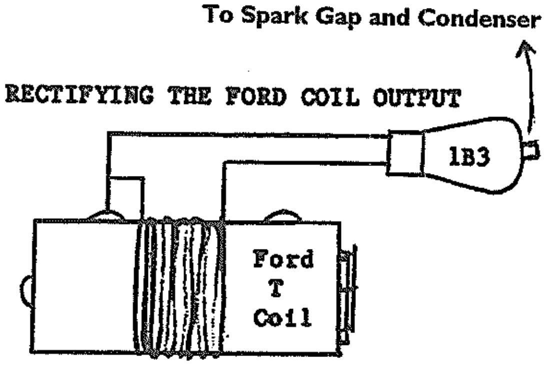

Wind 15 or more turns of fairly heavy insulated copper wire around the middle of the Model T Ford Coil. One lead from the winding goes to the Hi-voltage output of the Coil, and to pin 2 or 7 of the 1B3 tube. The other lead goes to the other pin, as only two pins of the tube, 2 and 7, are used in this hookup. The souped-up or rectified voltage is drawn off the top of the tube to the spark gap and condenser.

MODIFIED VERSION OF THE BECK TESLA COIL

115 volt house current into the black leads of the 6 volt transformer, one green lead to the T Coil negative terminal, the other to an SCR 45 amp Diode to make direct current to the Plus terminal of the T Coil. A 12-volt current to the T Coil heats it up and will not increase the spark. Radio Shack's 6.3 volt, 3 amp transformer does very well for a power supply.

Wrap 16 turns of insulated hookup wire around the middle of the T Coil and feed to pins 2 and 7 of the 1B3 rectifier tube. Feed the plus output of the T Coil into pin 2 or 7 of the 1B3 also.

Hi-voltage from the top of 1B3 goes to Spark plug. Widen gap of plug to 1/8 inch or more for maximum output of Secondary coil to Antenna.

One 10,000 WVDC (Working Volts Direct Current) commercial-type Capacitor, on the plus side, will do very well in this circuit. Capacity can be from .0015 to .005 microfarads.

Beck's tunable Tesla Primary called for 20 or more turns of bare wire for his movable contact but experience shows that 16 turns of insulated #19 wire for a Primary and adjust one turn more or less for maximum spark at Plus end Secondary coil. This is 4-1/4 inches of #34 insulated wire, or 4-3/4 inches of #32 insulated wire, for 420 turns more or less on 1-1/4 inch coil form 6 in. long. The Primary is wound on a 2 or 2½ in. coil form. There should be heavy insulation between Primary and Secondary to prevent power-loss arcing between Coils.

REMEMBER, the goal is a thin blue, or blue-white spark at least 3/4 inch long at the plus end of the Secondary. When fed into the center hole of the Antenna it should light up like a Christmas tree, with arcing between every ring. Feeding the coil output into the center hole gives more power to the shorter wave lengths. We get plenty of the longer wave-lengths from other electrical apparatus and from power lines.

This coil would be a good high-frequency power supply for Kirlian Photography.

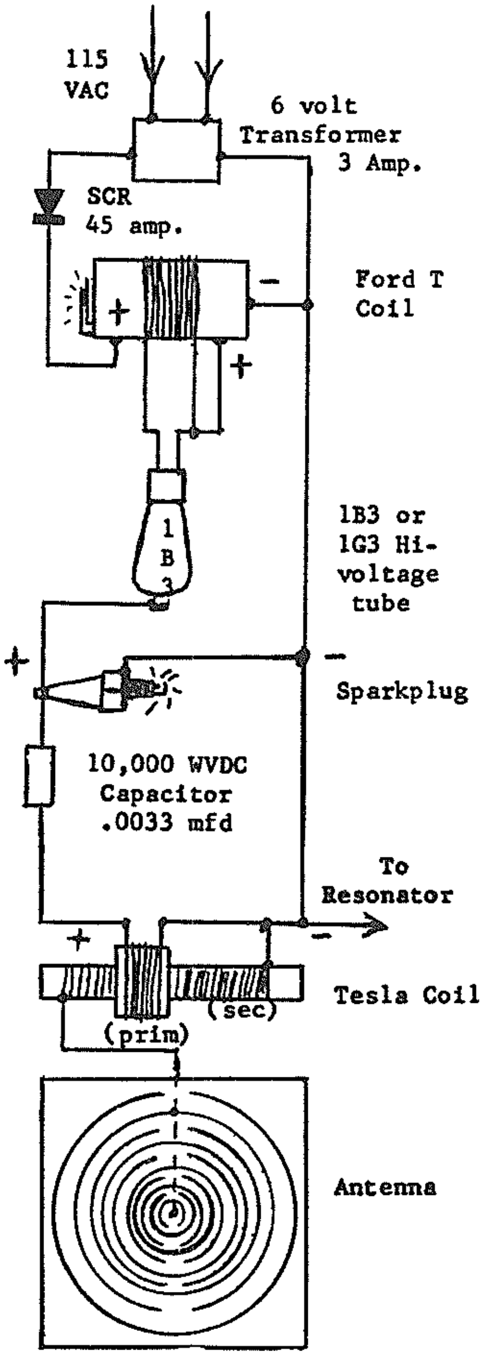

THE TRANSISTORIZED MWO

By Gerald W. Steele

Here is tentative information on the solid-state MWO. I say tentative since this is the 3rd version I've built. The first used an automobile spark coil for T2 but I found this undesirable since it lacked higher frequency harmonics. Even worse, the lower frequency ones could be harmful to living systems.

Version 2 was with a small black and white TV flyback transformer running at a slow (100 pulses per sec.) pulse rate. I eventually sped it up but could never get the required 1" spark from the Tesla coil.

This version uses a color TV flyback and a higher power discharge circuit. Nominal lowest frequency is 1.5 Khz though it could go as low as 270 Hz.

All diodes, SCR, and transistor can be substituted with universal replacement types available at electric part outlets.

As you can see, electronically it is simple. The power metering circuitry in the original MWO can also be used here if desired. I've had enough trouble maximizing the output of this device to worry about monitoring the power.

This device is excellent for use in Kirlian photography several ways: A modified circuit could be used for A, for a wide range of pulse rate settings.

There are two possible voltage sources for Kirlian use, each with unique characteristics. First is the output directly from T2 with or without the Tesla coil circuitry connected. Second is the Tesla coil output itself.

Other modifications include rectifying T2's output. The old MWO manual recommended the use of a tube rectifier out there are solid state types available.

Be sure it is a color TV high voltage rectifier. Connect as shown:

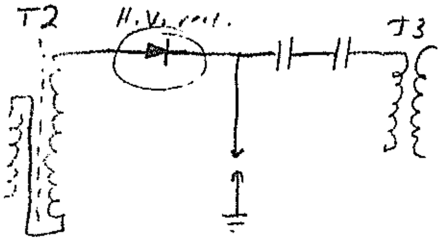

One question may arise in the builders mind: the need to heatsink SCR. No need. Not much heat generated.

Be sure R3 (or combination of resistors forming it) get ventilation. They get hot.

Always short C1 just after turning off power to work on the device. Remove short before reapplying power. The device is perfectly safe if you exercise care. R6 is a safety feature, it discharges C1 when turned off - but still short C1 when servicing, a standard safety procedure.

I am making constant improvements. The alternate values given for R3, C5 may give greater power at a higher pulse rate. I have yet to try them since 225 watt resistors are hard to find and apply. Besides it seems to work quite well as it stands. The only complaint one might have is the cost of construction principally due to the flyback transformer T2. But scavenging could help here.

Enclosing the spark gap in a plastic box could reduce its audible noise level. Ventilation of the spark gap gives more consistent operation. If the spark gap can be physically constructed above R3 the heat will produce air currents giving natural ventilation.

As I said, I am making constant improvements. Its too early to finalize it, but perhaps other associates may be interested in constructing and improving it. The lack of interrupter contacts gives a high level of reliability and consistent output.

Its too early in use for me to comment on its effects except that I think it is quite positive.

CIRCUIT DESCRIPTION:

The circuit consists of: low voltage oscillator, high voltage capacitive discharge generator, and Tesla coil.

The low voltage oscillator consists of Q1 a unijunction transistor operating as a relaxation oscillator. Frequency is variable from 270 Hz to 26 Khz with most power generated at 1500 Hz. VR1 controls this frequency. The DC Power for this circuit comes from a voltage doubler on the power transformer low voltage windings. The low voltage DC value was chosen to be as low as possible with consistent firing of the SCR.

The high DC voltage is rectified and filtered in C1. R1 is surge protection. The high DC voltage charges C5 through R3 until Q1 fires SCR1 whereby C5 is discharged through the primary of flyback transformer T2. D5 protects the SCR from high reverse voltage transients.

The secondary of T2 feeds the adjustable shunt spark gap and then through the Tesla coil resonating capacitors. Two capacitors are used here due to voltage requirements though one 15Kv capacitor could be used. From this point on it is all standard MWO fare.

Transformer T2 is expensive although any used color TV flyback could be tried. In all experimentation or adjustment the goal is maximum spark output at the Tesla coil secondary.

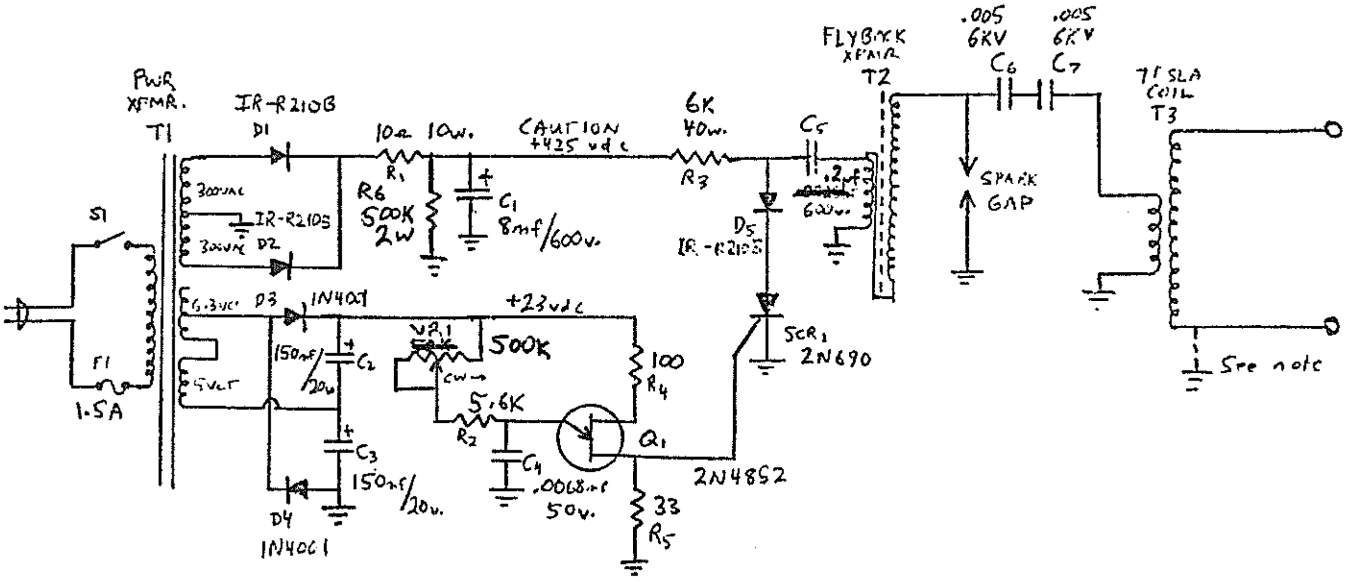
PWR XFMR. T1
IR-R210B D1
10Ω 10w. R1
CAUTION +425 vdc
6K 40w. R3
C5 .2uf 600v.
FLYBACK XFMR T2
.005 6KV C6
.005 6KV C7
TESLA COIL T3
SPARK GAP
S1
F1 1.5A
300vac
IR-R210S
300vac D2
R6 500K 2W
C1 8mf/600v.
D5 IR-R210S
SCR1 2N690
See note
6.3vac D3 1N4001
+23vdc
5vct
150mf/20v C2
VR1 500K
cw
5.6K R2
100 R4
Q1 2N4852
C3 150mf/20v.
C4 .0068mf 50v.
33 R5
D4 1N4001

Note: Optional ground for Tesla coil secondary, if it reduces output do not use.
R3 made from four 25K 10w resistors in parallel. VR1 linear taper. C1 may or may not be polarized. C6, C7 could be single .0025/15KV capacitor. SCR1, Q1 and all diodes can be standard replacement equivelants.
MFR Type: T1 - TRIAD. T2 - Stancor HO 6246, Thordarson fly.
Alternate values: R3 - 1K ohm, 225 watts; C5 - 1 microfarad, 600v.

THE LAKHOVSKY MULTI-WAVE POWER SUPPLY

by KLARK KENT

Many of us have read the Borderland Sciences Research Foundation (BSRF) publication "The Lakhovsky MWO" in which electrical wizard Bob Beck astounds us with tales of wonder. He tells how he was informed of the location of an old Lakhovsky Multi-Wave Oscillator that had been declared "unfathomable". He was able to reconstruct the circuits and after several hours the machine worked perfectly!

Mr. Beck and fellow experimenter Ed Skilling designed and built a modern version of the MWO based upon Nikola Tesla's and Lakhovsky's original works and patents. Georges Lakhovsky, a herald of the New Age, was no doubt a skilled experimenter and investigator of the Life Energy. This is evident by his practical application of Nikola Tesla's famous coil.

Super Science has updated both Lakhovksy's original and Bob Beck's adaptation of the Multi-Wave Oscillator. Looking at the schematic diagram in Bob Beck's article, we will eliminate everything but the 12 volt battery (or 12 volt power supply), switch, Model T Ford Coil, high voltage capacitors, and the Tesla Coil. A large high voltage diode is added to rectify the output of the Model T Ford coil which leaves us with a stripped down but thoroughly functional MWO power supply.

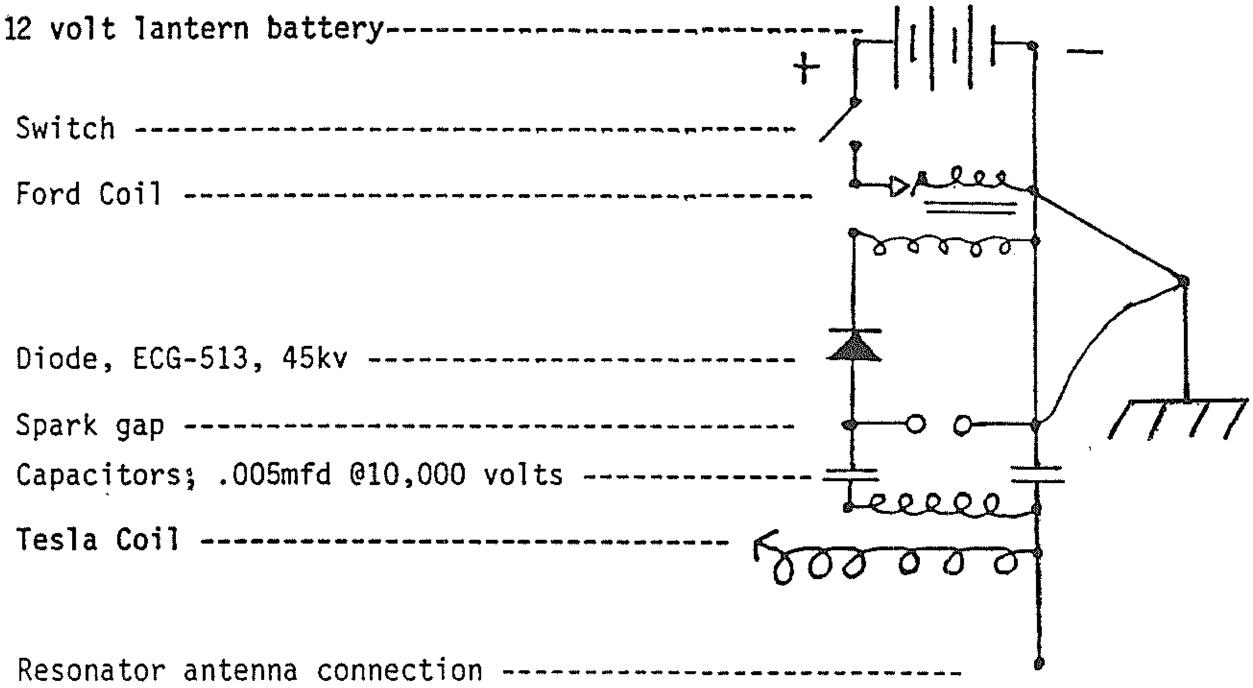

Before proceeding further, let's remember that the MWO is an experimental device and not for any healing or medical purposes! All experimenters should be in at least reasonably good health when building or experimenting with this high voltage, high frequency Tesla Coil Multi-Wave Oscillator power supply.

LET'S BUILD!

To begin our MWO Tesla Coil power supply project, we can round up the needed parts and components. Basic parts list:

-12 volt power supply; this can be a 12 volt lantern battery available from Radio Shack or any plug in power supply rated from approx 2 to 7 amperes.

-Model T Ford Coil; these old time induction coils are still available from J.C. Whitney Co, 1917 Archer Ave, PO Box 8410, Chicago, IL 60680. The MWO power supply will work adequately with the Model T Ford coil but a transistorized driver for a modern "hot rod" auto ignition coil will produce a Tesla Coil power supply with far more power!

-Switch; the on/off switch can be of any type. There is a switch already present on the plug in 12 volt power supply. I prefer to add a second switch of the key type, which prevents unauthorized use of the unit and a piece of vinyl tubing can be fitted over the metal key to create an insulated switch (this is important when turning the unit off).

-Diode, ECG-513 at 45 KV. The diode rectifies the output from the Model T Ford coil or the auto ignition coil by taking out the stray ac signals. This diode is a high voltage device rated at 45 Kilovolts (45 KV) peak inverse voltage.

-Spark Gap; an ordinary spark plug works adequately in this application. Try wrapping the spark plug in a large piece of vinyl tubing to insulate the spark gap. For a better spark gap, use a set of tungsten points. Chevy 6 cylinder 1965 model year works well to fabricate an adjustable gap.

-Capacitors; .005 mfd and rated at least at 10 KV (20 KV is better). We buy our capacitors or make them ourselves. I prefer to buy them. Check with your local electronic supply house to special order these high voltage capacitors; they will not be cheap... An alternative is to scrounge the local surplus stores to fabricate high voltage capacitors from several capacitors of a lower rating. It is easy to round up some 2500 wvdc capacitors with a 5000 volt test rating and .005 mfd capacitance. Wire the capacitors together in a series/parallel arrangement to arrive at the desired value. Series connections add the voltage rating and lessen the capacitance by division of the total number of capacitors in the series. Thus four capacitors rated at 5000 volts wired together in series yield one 20,000 volt capacitor. If the capacitance was originally .005 mfd each, now it will be .00125 mfd for the series. Thus we need to connect four of the series capacitors together in a parallel arrangement to create one 20 KV capacitor at .005 mfd. Parallel connections have no effect on the voltage rating but parallel connections add the capacitance of each individual capacitor.

CAUTION! Capacitors store electrical energy and can shock even when unit is turned off. Be sure to discharge the capacitors before touching them...

TESLA COIL

We make the Tesla Coil ourselves using the basic parameters worked out by Bob Beck in his original MWO adaptation from Georges Lakhovsky. The Super Science Tesla Coil is a close coupled, liquid filled coil that produces a 3" spark discharge when driven by a Darlington transistor and "hot rod" auto ignition coil setup.
For making the Tesla Coil, I like to use pvc pipe for the secondary coil form and acrylic tube for the primary coil form. The acrylic is transparent which permits visual inspection of the interior of the coil to observe any corona leakage or malfunction.

For the secondary winding: cut a 1 1/4" outside diameter piece of pvc pipe seven inches long. For the primary winding: cut a 2 1/2" outside diameter piece of clear acrylic tubing (1/4" wall thickness) seven inches long. Wind the secondary coil form with #32 gauge bare enameled copper wire so that the turns are tightly packed to create a winding approx six inches long with a 1/2" gap at either end. Wind the primary coil form with #14 gauge stranded and insulated copper wire. Wind exactly 18 turns and tightly pack them together. The vinyl insulation on the #14 gauge stranded wire will provide the correct spacing between turns.

Place the secondary winding inside the primary winding and use Super glue to hold the windings in place. For the terminals of the windings, use ordinary banana jacks to create a modular coil with a standard banana plug connector system. There will be tremendous electrical stress between the primary and secondary windings. We could reduce this dialetric stress by using a pancake style primary winding, but the pancake style primary takes up a lot of room. To solve the dialetric breakdown between the two windings, we insulate the space with liquid mineral oil. The liquid mineral oil is self healing should a dialetric breakdown occur. Solid forms of insulation such as paraffin wax are useless once a short circuit path has been created between the two windings.

To use the mineral oil insulation system, finish the coil first. Then drill a small hole in one end and squirt the oil in using a squeeze bottle. Then seal the hole using a piece of acrylic and Super glue. A word about Super Glue! This stuff is quite remarkable and is sold in hobby shops that specialize in flying, gas powered, model airplanes. Use accelerator to fast dry the glue. The brand name of the Super glue is "Gap Filler" and the accelerator is "Insta Set" both made by Bob Smith Industries.

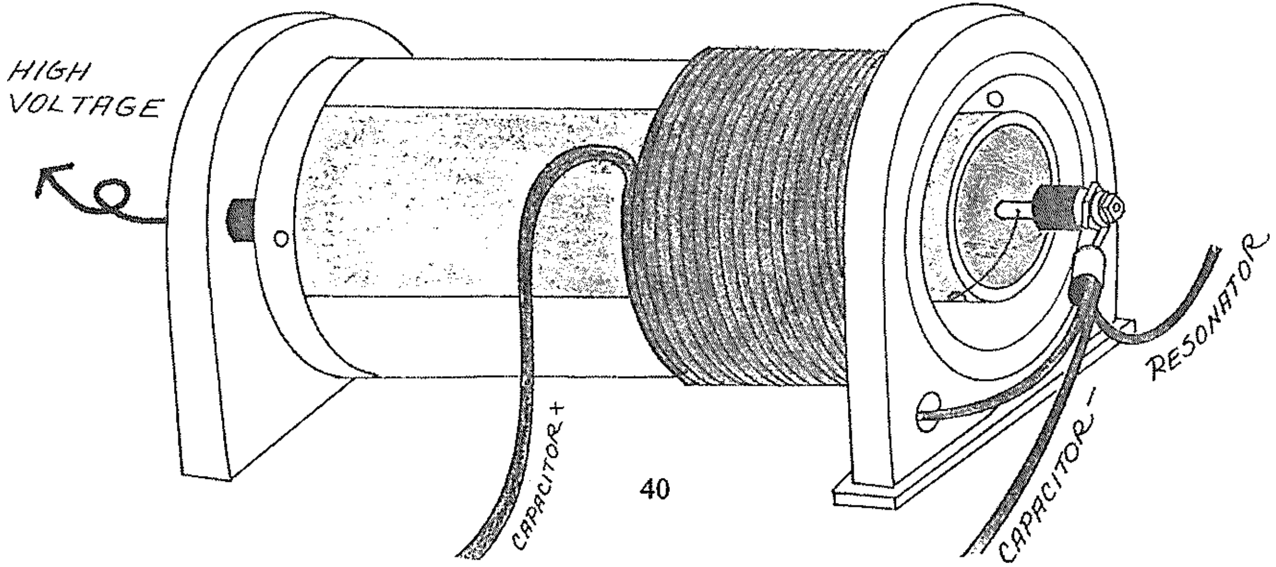

SPARK GAP

Even though an ordinary spark plug will work for a spark gap, we can do much better with an adjustable spark gap made from a set of tungsten auto ignition points. Chevy six cylinder, vintage 1960's are fine. To fabricate the adjustable spark gap, cut the base plate of the point assembly into two parts; make the cut near the spot where the two tungsten contacts normally meet. You should now have two base sections, with a contact point on each section. The reason we cut the base plate is to prevent the high voltage from jumping from the connections to the base and bypassing the breaker points. File down any sharp points where the base was cut. Mount the two tungsten points in such a way that the contact points barely touch each other and the base sections are as far apart as possible. Mount the spark gap assembly on a small piece of acrylic sheet. Wood is not suitable to mount the spark gap because the high voltage current will pass right through the wood base, possibly setting it on fire! Use a large nylon bolt to mechanically adjust the points. Mount the nylon bolt (Toilet seat bolts work well) in the side of the briefcase so that it butts against the spring loaded contact arm; this acts as a spark gap adjuster.

The advantage with an adjustable spark gap is that we can tune the gap while the MWO is in operation. The closer the two tungsten points are to each other, the faster the frequency of the spark jumping the gap. The further apart, the slower the frequency of the spark jumping the gap. Typically the Violet Ray Bulb Assembly calls for a fast frequency spark gap setting; while the High Power Radiator Antenna or Golden Ratio Antenna call for a wider gap spark setting to produce more power.

One side of the spark gap points will be at "ground" potential; while the other side will be at "high voltage" potential. Leave a metal discharge point exposed on the high voltage side of the spark gap assembly. We do this so that the entire MWO power supply may be discharged with a test wire running from the Ground Bus to the high voltage point. This provides a quick way to discharge the capacitors before troubleshooting the unit. Also, cover the spark gap assembly with a visual block made from translucent acrylic. This visual screen will block the ultraviolet light emitted by the spark jumping the gap. Leave the sides of the spark gap assembly open to allow circulation of air around the points; this will "quench" the spark gap and prevent the jumping spark from turning into a constant arc. Also, mount the cooling fan air intake near or directly above the spark gap assembly; this will serve to further quench the gap for better, more bouncy performance...

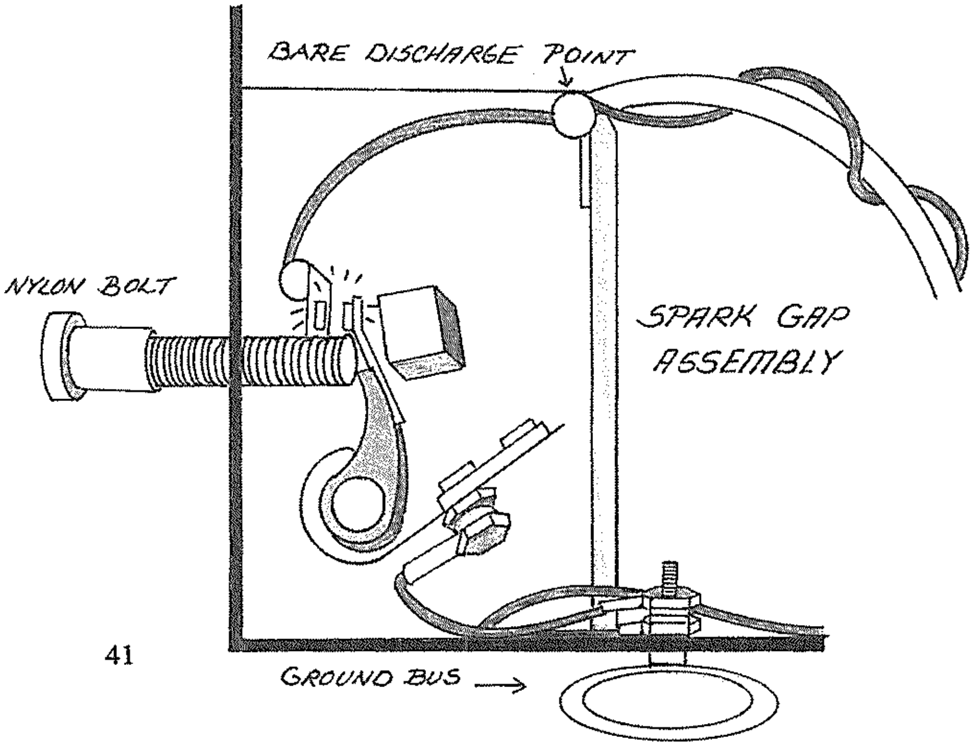

555 TIMER AND TRANSISTORIZED DRIVER

Even though the Model T Ford coil works adequately to power the MWO Tesla Coil power supply, we can really do better with a modern "hot rod" auto ignition coil. Mallory makes an aftermarket coil for drag racing that puts out approx 45 KV. The "MSD Blaster" is another aftermarket auto ignition coil that really gets the job done. Even an ordinary UC-15 replacement coil with its 30 KV rating will outperform the Model T Ford coil.

However there is one consideration to using an auto ignition coil as the primary high voltage source. Since we are working with 12 volts direct current, we need to interrupt or chop the 12 volts going into the auto coil to enable it to function as a transformer. The Model T Ford coil had a built in interrupter for this purpose. We will make a much better interrupter from a 555 timer chip and a high power Darlington transistor. The 555 timer chip will chop or interrupt the 12 volts going into the auto ignition coil from 130 times per second to 1300 times per second while the Darlington transistor handles the power. With the Model T Ford coil we were stuck with the fixed frequency of the interrupter (approx 200 Hz). With the 555 timer system, we can tune the frequency of the primary high voltage from 130 Hz to 1300 Hz. This really makes a difference when we try to get maximum performance from our Tesla Coil when it is driving the MWO Golden Ratio Radiator Antenna or driving the Violet Ray Bulb Assembly for direct stimulation.

Below is the schematic for the 555 Timer/Transistor Assembly. It is shown hooked up to the auto ignition coil. You can hardwire this circuit using a Radio Shack #276-162 project board. I use a large heatsink to mount the Darlington transistor (RCA 8766) and to mount the 50K ohm tuning potentiometer. When the MWO power supply is housed in a briefcase, the shaft of the 50K ohm potentiometer can extend through a hole in the side of the case. Go ahead & slip a short piece of vinyl tubing over the potentiometer shaft both to extend the shaft and to insulate it from high voltage.

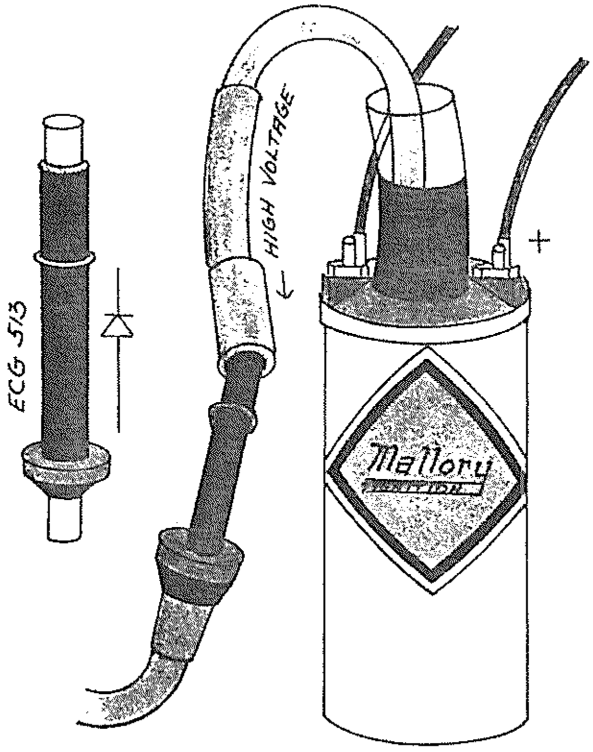

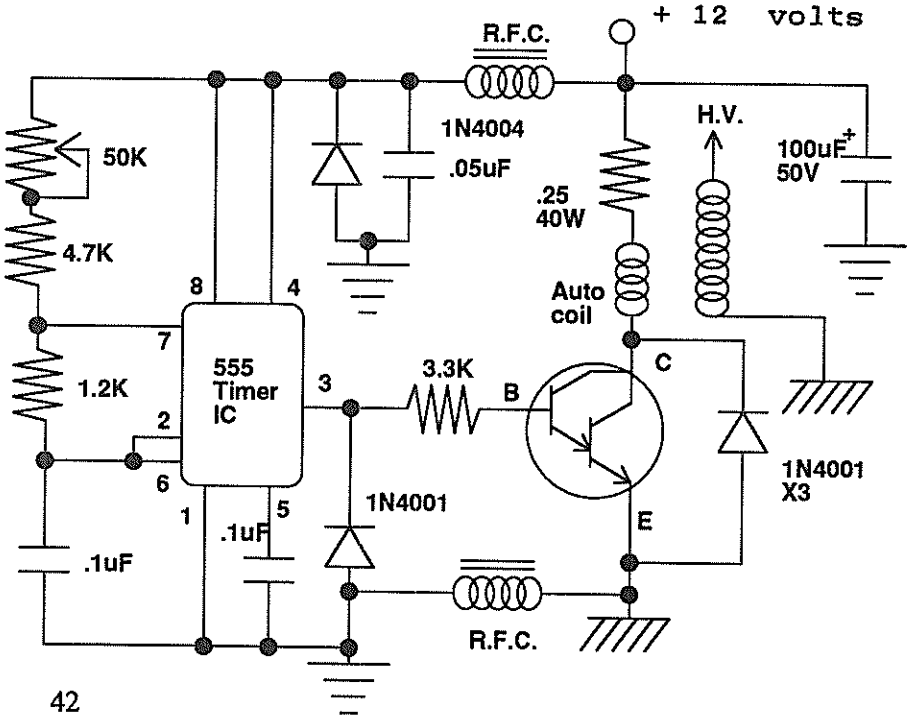

GROUND BUS

An external Ground Bus is important because it lets us reference the high voltage to the zero potential of the Earth itself. Without a true Earth ground, the negative side of the circuit can pick up a substantial voltage by induction from the high voltage potentials in the circuit. This condition is known as a "floating ground" and could result in the experimenter getting shocked. The shock itself is harmless, but the surprise that an accidental shock causes may lead the experimenter to jump back and bump his head! The Ground Bus may be connected by a long wire to a cold water pipe or to a steel fence post or to a piece of rebar driven into the ground. The Ground Bus is a handy reference point when discharging the circuit before troubleshooting.

SAFETY

Try to follow the "one hand rule" when actively adjusting the live high voltage circuit. This means keep one hand in your pocket or one hand behind your back. If an accidental shock should occur, then the electrical current does not travel across the heart when the experimenter follows the "one hand rule". Another safety procedure is to use a 12 volt lantern battery to power the circuit while it is still under construction. Once all the problems have been solved, then a more powerful 12 volt supply can be used.

TESTING THE CIRCUIT

Let the circuit run in darkness and try to observe any corona leaking from exposed high voltage points. There are two good and easy ways to stop high voltage leaks. One method is ordinary vinyl tubing: wrap wires carrying high voltage in vinyl tubing to absolutely stop them from leaking power robbing corona. Auto parts store sell expensive spark plug wire which will withstand the high voltages we are dealing with. If you use spark plug wire, be sure that it is metal wire, not "radio suppressor" wire (it's carbon fiber and impossible to solder or connect to).

Another good method of insulating high voltage is 100% silicone clear caulk! With your caulking gun, apply some 100% silicone clear caulking to exposed wires, exposed high voltage points and other places where you can observe leaking high voltage corona.

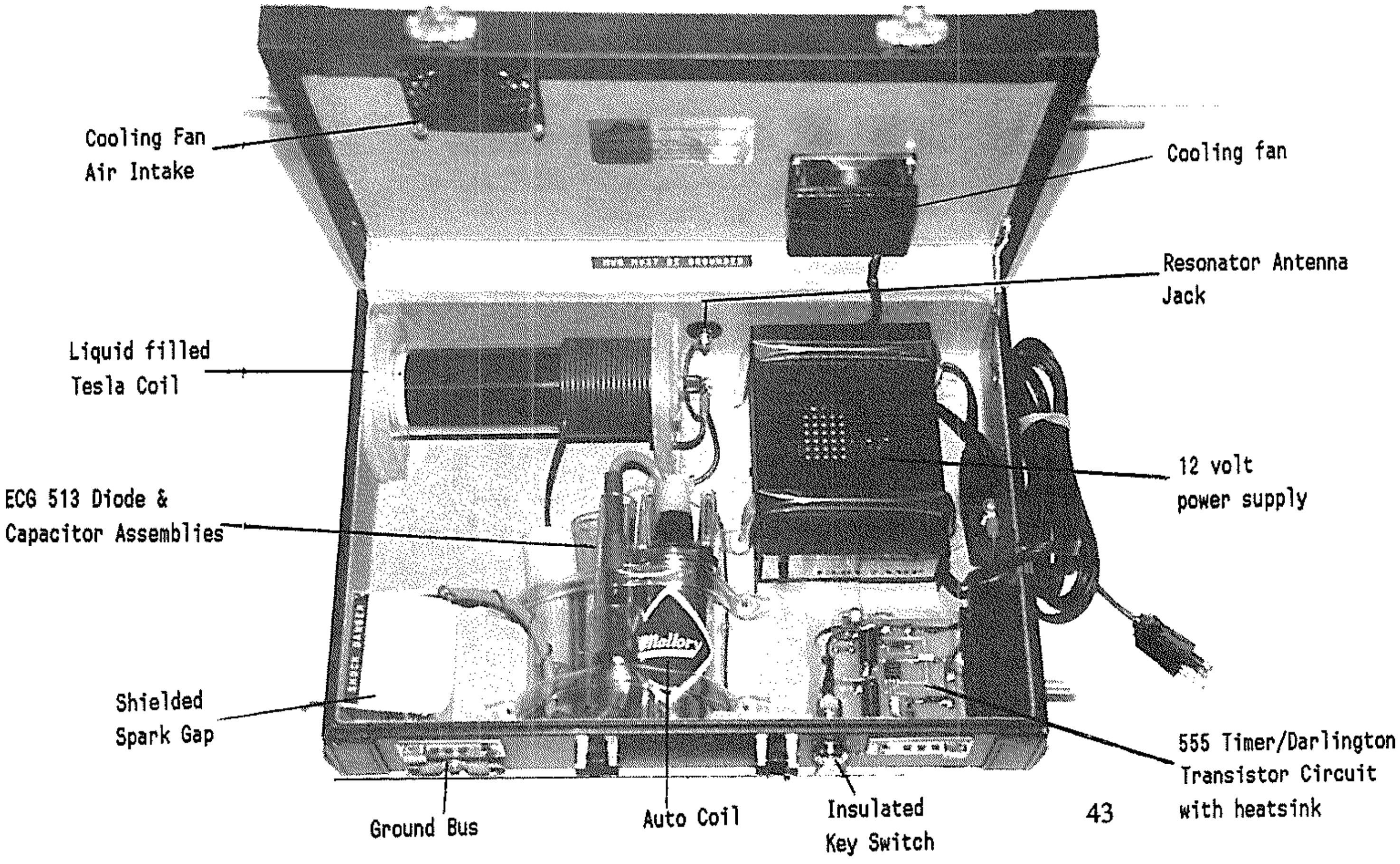
Cooling Fan
Air Intake
Cooling fan
Resonator Antenna
Jack
Liquid filled
Tesla Coil
ECG 513 Diode &
Capacitor Assemblies
12 volt
power supply
Shielded
Spark Gap
555 Timer/Darlington
Transistor Circuit
with heatsink
Ground Bus
Auto Coil
Insulated
Key Switch

LAKHOVSKY MULTI-WAVE OSCILLATOR.

HOW IT WORKS:
The MWO sets up a high voltage oscillating electromagnetic field between the two antennas. Experimental subject sits or stands in this field and every cell in the body is exposed to the resonant effect of the multiple waves. Just as one guitar string will sound when exposed to its resonant tone; so will each cell vibrate to its own natural frequency!

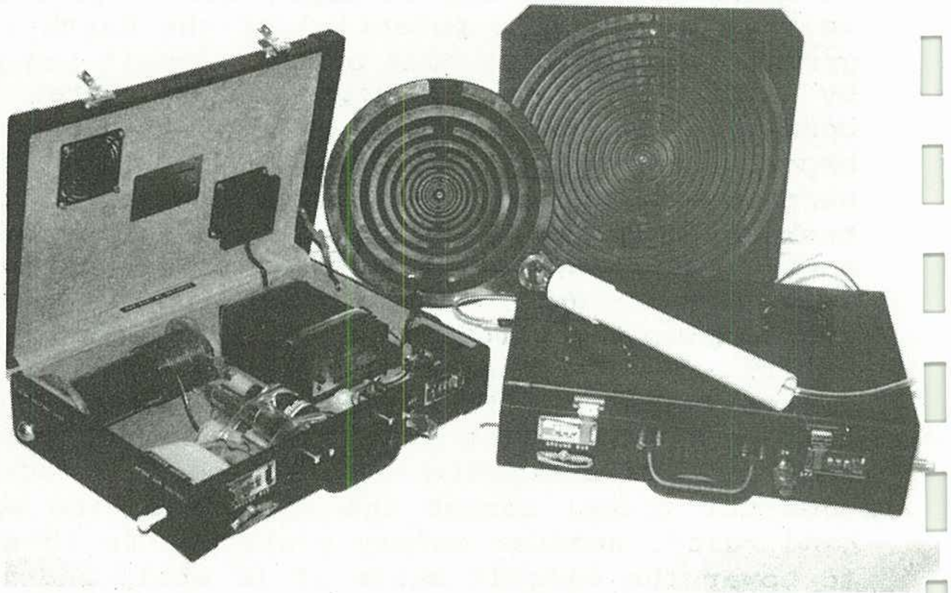

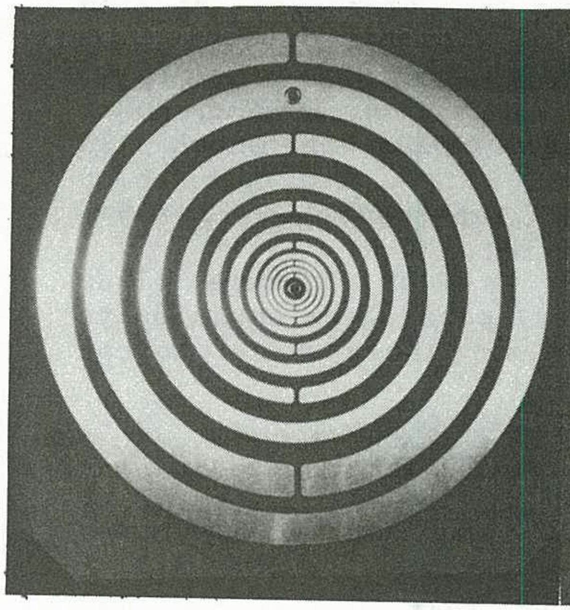

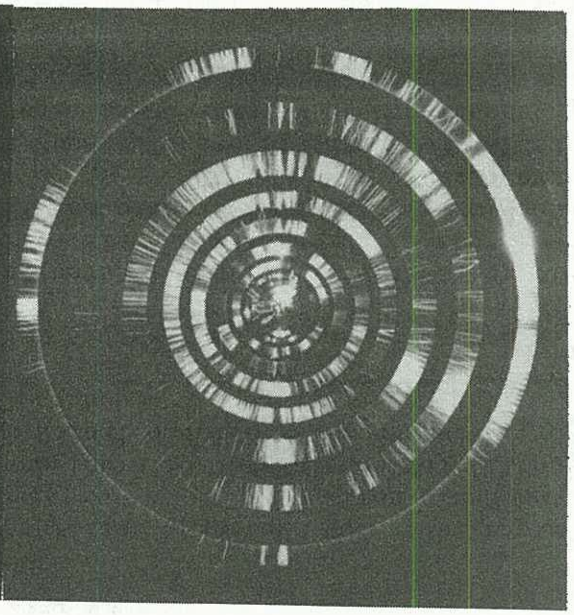

In this time lapse photo, we can see the dramatic effect of the high voltage excitation of each ring of the MWO antenna! Each ring radiates at its characteristic wavelength and frequency to cover a range from approx 200 Hz to 14 Giga Hz!

Modular construction works well for the Tesla Coil in the Super Science MWO power supply. This high voltage Tesla Coil can produce a 3" spark discharge from the smooth HV Discharge Terminal! The unit is a versatile experimenter's kit that can be used for Kirlian photography, or to create a plasma of Argon gas in the Violet Ray Bulb Assembly.

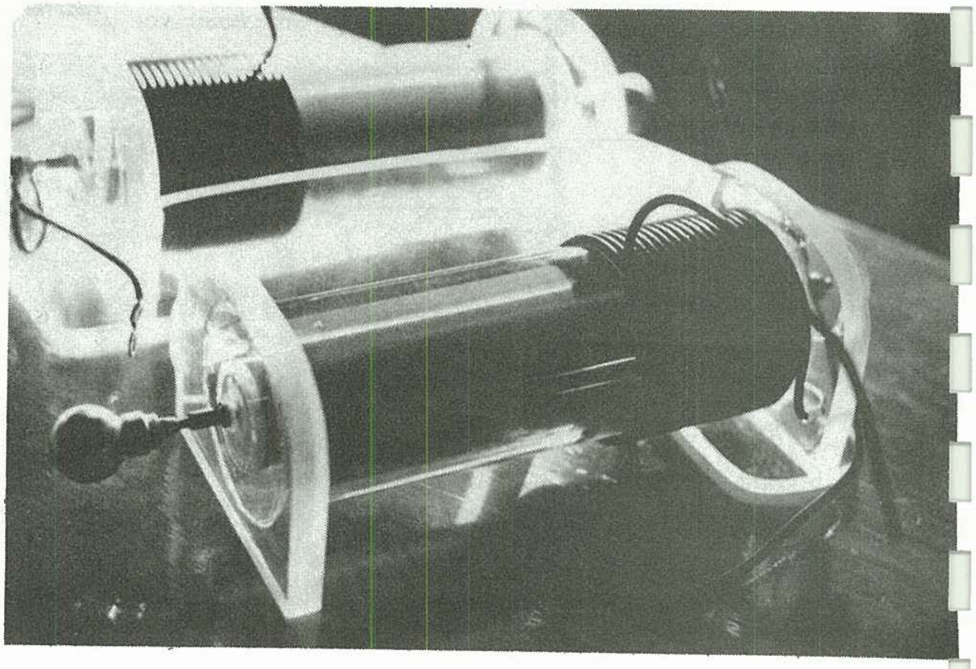

SUPER SCIENCE
P.O. Box 392
Dayton, Ohio 45409
(513) 298-7116

ANTENNA HOOKUP

Experimental subject is placed within the oscillating electrostatic field between the two antennas which are approx 4 to 6 feet apart. All metal such as heavy belt buckles is removed from the body. Clothes are optional. Sessions should last approx 20 minutes; skip at least one day between sessions to allow the body a chance to detoxify.

With the Golden Ratio Radiator Antenna, the use of the second Resonator is not required due to the unique "Scalar" properties of this design.

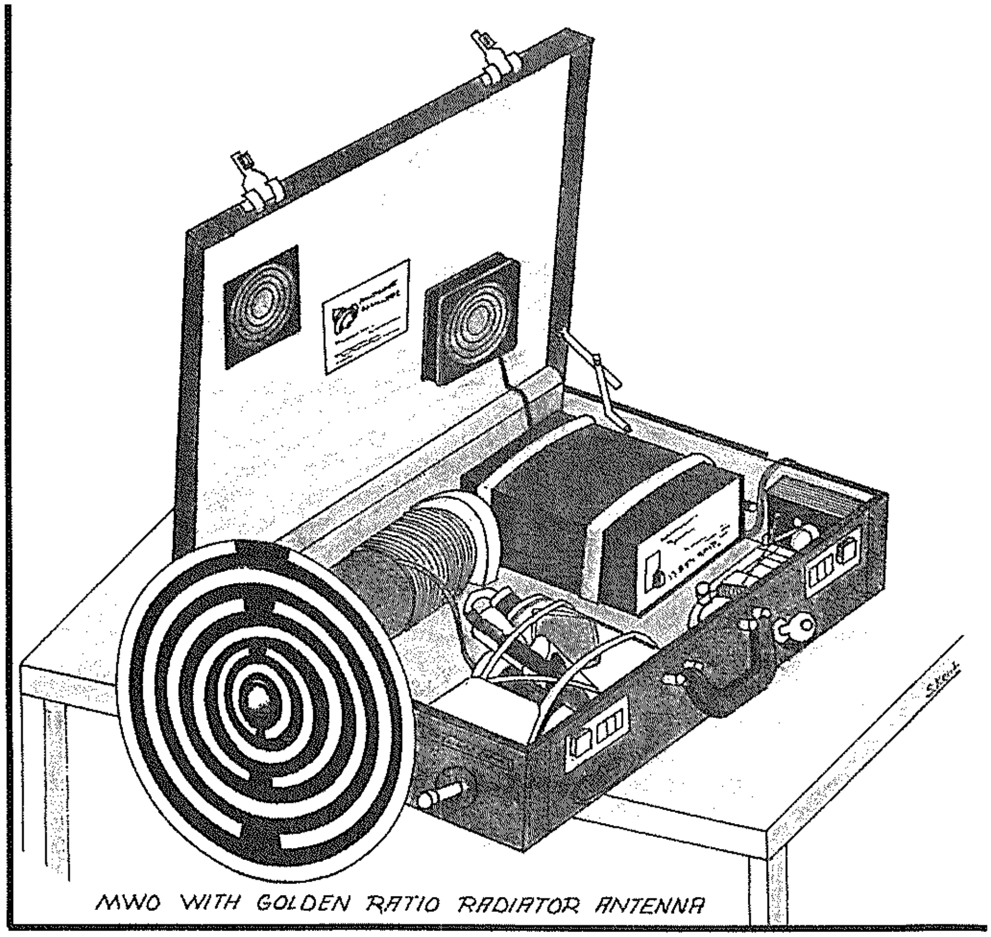

MWO WITH GOLDEN RATIO RADIATOR ANTENNA

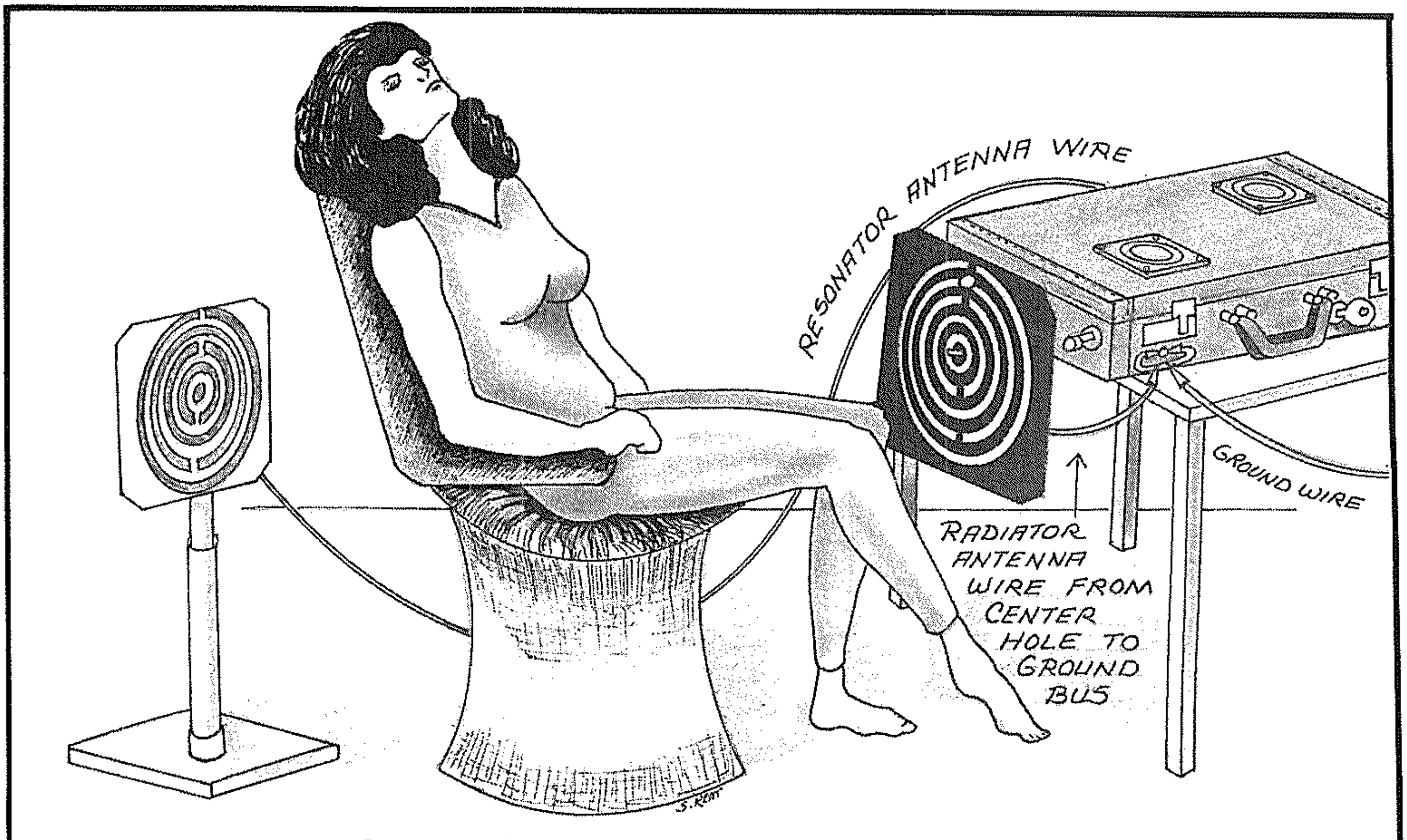
RESONATOR ANTENNA WIRE
GROUND WIRE
RADIATOR ANTENNA WIRE FROM CENTER HOLE TO GROUND BUS

MWO WITH HIGH POWER RADIATOR ANTENNA AND RESONATOR ANTENNA

From the December 1963 Journal of Borderland Research
Excerpts from:

THE WAVES THAT HEAL

The New Science of Radiobiology

by Mark Clement

edited with commentary by Riley Hansard Crabb

Bob Beck has loaned us a copy of the late Mark Clement's brief, illustrated review of Georges Lakhovsky's Multiple Wave Oscillator showing pictures of antennas, before-and-after skin cancers apparently cured by the Oscillator, and of cancerous plants returned to normal health by two means: 1. Radio-Cellular Oscillator, and 2. Single Wire Oscillating Circuits.

"On the day Hitler's mechanized hordes drove into Prague a book was published in London entitled **The Secret of Life** by Georges Lakhovsky, a Russian-born engineer resident of Paris. The book had previously appeared in French, German, Italian and Spanish translations and Lakhovsky's theories were well known on the Continent long before the English Translation was published.

"The evil designs of the Nazis soon gave rise to depressing repercussions in all spheres of human activity and Lakhovsky's remarkable work was left virtually unnoticed in this country (England) by the medical profession and the public alike. A few discerning critics, however, expressed their views with appreciative understanding. The Hospital said 'Lakhovsky's theories are so interesting from the point of view of future developments that it will repay any keen hospital officer to give them his close attention.' Science Forum said 'The admirable translation of Lakhovsky's **The Secret of Life** will be heartily welcomed in progressive circles. The general theory is one which physicists, biologists and medical men cannot afford to ignore.' But these were voices in the wilderness and the field for experimental investigation was left open to a few amateurs devoid of scientific training and lacking the necessary resources.

"The impartial investigation of claims of cure of various diseases by un-orthodox methods demands high qualities of mind and character possessed by very few men indeed. Scientific truth is no respecter of persons or of vested interests, and Lakhovsky's spectacular results in the treatment of plants and animals and human beings were inevitably bound to antagonize orthodox practitioners and exponents of popular medical cults.

"On the Continent Lakhovsky's work attracted a great deal of attention in scientific circles, particularly in Germany and Italy. Italian investigators were among the first to study Lakhovsky's theories and put them to the test in laboratories and clinics. It was, of course, to be expected that the new science of Radiobiology should make a special appeal to the intellectual heirs of Galvani, Volta and Marconi. It is somewhat depressing to observe that in England, the country of Faraday and Clerk Maxwell, Lakhovsky's theories have not yet received the attention they deserve.

REGENERATING OSCILLATING CIRCUITS

"Lakhovsky has been criticized by physicists ignorant of biology and by biologists ignorant of physics! These self-appointed censors of knowledge soon found they were confronted with a redoubtable opponent who fully realized the value of experimental evidence. Its publication resulted in causing annoyance to the custodians of infallible doctrines who made up with carping verbiage what they lacked in clarity of vision. But Lakhovsky's achievements with his oscillating circuits supported by astonishing photographs of regenerated tissues in plants and human beings could not be exploded by mere verbal fireworks, and as evidence accumulated the skepticism of hostile critics relapsed into morose silence.

"The indisputable fact remains that Lakhovsky was the first experimenter to make use of high frequency electro-magnetic waves in the domain of biology. Thus out of the application of radio-electricity to biology his work developed and gradually established the foundations of the new science of Radiobiology. The first international Congress of Radiobiology was held in Venice in 1934. A leading authority on electro-therapy, the late Dr. E.P. Cumberbatch, wrote: 'Although it had been frequently observed that the short Hertzian waves could produce heat at a distance from the transmitter the first scientific investigation from a biological point of view was made by Lakhovsky and his colleagues who published a paper in 1924 on the effects of very short waves on cancer in plants...'

"When France was occupied by German troops, Lakhovsky, being a prominent anti-Nazi, decided to leave Paris and made his way to New York where he died in 1942 at the age of 73.

"In the United States Lakhovsky's theories were put to the practical test with typical American speed. Noticing that patients treated with Lakhovsky's Multiple Wave Oscillator showed signs of rejuvenating appearance, the possibility of exploiting this side-effect proved irresistible to an enterprising 'beautician'. A film was made featuring cases after treatment with Lakhovsky's MWO, which proved to be both interesting and convincing.

"In New York Lakhovsky was approached by several hospitals with a view to testing his apparatus experimentally. The remarkable results obtained . . . in a large New York City hospital, and also by a prominent Brooklyn urologist are fully set out in the text.

"In this country (England), owing mainly to the inauspicious time at which Lakhovsky's major work was published, very little interest has been shown on the part of those best qualified to judge its merits. The medical profession, whose conservatism is the most formidable barrier to progress, has been notoriously slack in investigating the new radio-electrical methods of treating disease as originated by Lakhovsky...."

Mark Clement's observation on the formidable conservatism of the medical profession in 1949 is still true today. Effective electronic diagnosis and treatment of disease is still a closed door here in America as far as the AMA is concerned. Those brave MDs who have tried to keep medical practice abreast of developments in electronic science have been discouraged, if not actually driven out of practice. Dr Abrams is a good example, with his Oscilloclast, developed and proven effective in San Francisco about the time Lakhovsky was developing the MWO in Paris in the 1920s. Radical innovations in any line of endeavor are usually thought up and proven by rank outsiders. Georges Lakhovsky was no doctor, but an electrical engineer!

Doctors, with their own pharmacies and with massive investments in the drug industry, are not likely to show much interest in radical new ways for treating human ailments, especially if the new way is a threat to profitable practice; and, as long as Americans think they can buy health in a pill or a shot in the fanny, doctors will continue to treat them with poisonous drugs at top prices. This is good business.

BORDERLAND RESEARCH

But for those of us who would like to keep our bodies as uncontaminated -- and as healthy -- as possible, Lakhovsky's theories

It is highly significant that about the only rays and ray machines approved for general use by the American Medical Association are the destructive ones of X-rays and atomic radiation for the mis-treatment of cancer.

"Since Lakhovsky's oscillating circuits were first used for the treatment of cancerous geraniums, an experiment fully described in communications to the French Academy of Science, the same method has been successfully applied in Continental hospitals and nursing homes."

Photos of the successful treatment of cancerous geranium plants are included in Clement's book, along with the pictures of the disappearing skin cancers in human beings. So Clement describes the major features of Lakhovsky's work in this first chapter, the simple, one-coil Oscillating Circuit powered by cosmic rays, and the electric-powered Multiple Wave Oscillator. There are photos of the Multiple Wave Oscillator, showing clearly the unique circular antennas. These were made of metal tubing, separated and suspended with silk thread. The simpler, aluminum-foil printed circuit in Bob Beck's adaptation is an easy way to experiment with the idea.

THE OSCILLATING CIRCUIT

"The diagram on this page shows where the oscillating circuits should be worn, i.e. round the neck, waist, and above the elbows and knees. The effect is intensified if more than one circuit is worn. Generally speaking, the neck and waist are the favoured sites and likely to bring about the best results.

"In the light of Lakhovsky's theories the fight between the living organism and microbes is fundamentally a 'war of radiations'. If the radiations of the microbe win the cell ceases to oscillate and death is the ultimate result. If, on the other hand, radiations of the cell gain the ascendant the microbe is killed and health is preserved.

THE SECRET OF LIFE (AT THE FORM LEVEL)

"Broadly speaking, health is equivalent to oscillatory equilibrium of living cells whereas disease is characterized by oscillatory dis-equilibrium. This general principle has given rise to a vast number of experiments covering the whole field of biology. Plants, animals and human beings have been successfully treated by the application of Lakhovsky's oscillating circuits and the reader is referred to his major work The Secret of Life for a full discussion of this important subject.

Those of you who have been studying the Kabala with Mrs. Crabb and me must have been struck immediately by the metaphysical soundness of Lakhovsky's basic theory expressed above. One wonders if this Russian escapee from Communism, and from Naziism, wasn't indeed a student of the Western Mystery Tradition, perhaps even an Initiate! The Tree of Life is the groundplan of our Tradition. On the Tree we show the descent and the ascent of power, oscillating back and forth from left to right and right to left across the Middle Pillar, or balance point.

Now for the final paragraph of Chapter One. "During recent years observations on the part of several investigators have established the fact that most animals, including insects and birds, emit radiations while they are also sensitive to the influence of external electromagnetic waves. These radiations emitted by all living beings cover an indefinite range (infinite range?) and are characterized by a multitude of different wave lengths. In the immense range of existing radiations we can only perceive the luminous octave (visible light) but we know that whole gamuts of radiations exist beyond the narrow limits of the visible spectrum. The limitations of our senses prevent us from perceiving radiations of living beings and this sensory incapacity also excludes from the field of direct awareness the vast gamut of electromagnetic waves traversing our atmosphere. Yet all these radiations and waves exist and affect all forms of organic life in various ways."

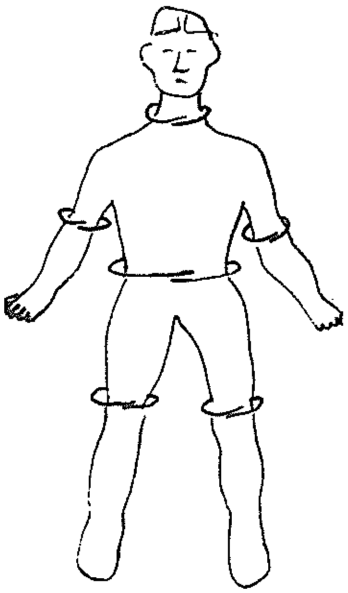

Single wire oscillating circuits around subject's neck, arms, waist and knees.

DR. RAUL ARAUJO'S REPORT

In Chapter Two Mark Clement gives "Medical Reports on Lakhovsky's Oscillating Circuits." We'll give only one, by Dr. Araujo, Physician at the Prophylactic Institute, Montevideo Uruguay. "In 1929, Dr. Araujo was suffering from a long-standing infection of the larynx which his eminent colleagues had failed to cure. While in Paris he met Lakhovsky who advised him to wear one of his oscillating circuits in the form of a collar. Although feeling rather skeptical about this unorthodox method of treatment Dr. Araujo wore the oscillating circuit for several months with the result that he fully recovered and has had no relapse since. This personal experience led him to treat his own patients with Lakhovsky's oscillating circuits. Having treated over 300 patients suffering from a wide variety of diseases Dr. Araujo succeeded in curing or greatly improving cases of rheumatism, asthma, anemia, neuralgia, gastric ulcer, colitis, glandular disturbances, sterility, genito-urinary affections, general weakness and even cancer. Dr. Araujo observed that in cases in which Lakhovsky's oscillating circuits appeared to give no definite results there was invariably a latent infection which required special treatment such as syphilis, cholecystitus, sinusitis, pyelitis, dental sepsis, etc...A full account of Dr. Araujo's Report appeared in a monograph entitled 'Cosmic Rays and Oscillating Circuits' by Dr. P. Rigaux, Paris, 1932."

CHAPTER THREE, THE MULTIPLE WAVE OSCILLATOR

Most of this has been well explained by Bob Beck in his article in the November Journal, though we might quote again in Clement's own words: "After many experiments Lakhovsky succeeded in constructing an apparatus generating an electrostatic field in which all frequencies from 3 metres to the infra-red region could be produced. Hence in this field every cell could find its own frequency in resonance... Thus a type of oscillator was obtained giving all fundamental wave lengths from 10 cms. to 400 metres, corresponding to frequencies of 750,000 to 3 milliards per second...

"Since 1931 the Multiple Wave Oscillator has been used in most European countries and in America for the treatment of many diseases including cancer. No contra-indications for the use of the apparatus nor any harmful effects on patients have ever been reported. This is in striking contrast with short-wave therapy in general, X-rays and radium whose application in the hands of experts has not infrequently been followed by the most serious consequences..."

Then follow several pages of photos of before-and-after cancer cures and medical reports from doctors. This one is typical, from Dr. Nicholas Gentile, Chief Radiologist of the Clinic for Incurables of the Apostolic Ladies of the Sacred Heart of Jesus, offer a way out of the clutches of the medical trust. There is little likelihood that the AMA will support research in this area of electronic science for years to come. There is every probability that there will be murderous opposition to any attempt to exploit the MWO on a national scale. This means that you will have to build and test your own equipment, or make friends with some electronic technician -- engineer, TV or radio repairman -- and prevail upon him or her to build it for you. This is the active kind of borderland research which requires that you get up out of your easy chair and get to work! If you live in an isolated spot, take a mail order course in basic radio construction. If you live near a high school that has an adult education program, sign up for the evening class in radio theory and bench work. I did at Barstow in 1958, so can you. If you want to be a citizen of the New Age you should make some effort to keep up to date, even if the sick industry doesn't!

What are Lakhovsky's general theories? Mark Clement sums them up neatly in Chapter One of **The Waves That Heal**:

"The fundamental principle of Lakhovsky's scientific system may be summed up in the axiom 'Every living being emits radiation'. Guided by this principle Lakhovsky was able to explain such diverse phenomena as instinct in animals, migration in birds, health, disease, and, in general, all the manifestations of organic life.

"According to Lakhovsky the nucleus of a living cell may be compared to an electrical oscillating circuit. This nucleus consists of tubular filaments and chromosomes made up of insulating material and filled with a conducting fluid containing all the mineral salts found in sea water. These filaments constitute minute oscillating circuits endowed with capacity and inductance and capable of oscillating according to a specific frequency. They are comparable to the circuits, coils and windings of radio receivers."

I believe Clement's comparison of a body cell with a radio receiver is only half-true. The cell can also be compared with a radio transmitter when it is releasing energy. We are both radio-receptive and radio-active! Lakhovsky's theory should be acceptable to any occult scientist. The push-pull idea, the positive and the negative, the Law of Opposites, is basic to all forms and to all life at this level of existence. By inductance Clement means that radio-electricity can be induced or put into every cell of your body. By capacity he means that each body cell can hold a charge of radio-electricity, just as the battery of your car can carry a charge of electricity. In the case of your body, the charge is put in through invisible but very real radio waves. These can travel unimpeded through any solid object!

"The living cell can thus play the part of a transmitter or receiver of exceedingly short radio-electric waves which give rise to high frequency currents in the circuits of the cellular nucleus. Now the vibration of an oscillating circuit is maintained by radiated energy. But where does it come from.

COSMIC RAYS

"Scientists have established the existence of highly penetrating waves to which, by reason of their universality, they have given the name of cosmic rays. These rays, generated by the interference of astral radiations, are capable of traversing dense metals and soils to a great depth. Lakhovsky has demonstrated the cellular oscillation of living organisms is maintained by cosmic radiation. According to him disease may be said to result from oscillatory dis-equilibrium due to either a deficiency or an excess of cellular radiation. An important factor in this connection is the nature of the soil. A reinforcement of cosmic rays takes place on soils which are good conductors of electricity such as clay, ferruginous and carboniferous soils.

RESONATE IN HARMONY WITH THE UNIVERSE

"To avoid the harmful effects of loss of oscillatory equilibrium it is necessary to ensure the regularization of the cosmic field round the living organism by filtration. Lakhovsky's own explanation of the problem of filtration is as follows: 'The terrestrial atmosphere is known to be impregnated by a vast range of electro-magnetic oscillations of all wave lengths and intensities as a result of unceasing electrical discharges (thunder, etc.). Besides that we know that electric motors and other electrical machines create in the atmosphere a field of auxiliary waves. Moreover, during the past 25 years the earth has been flooded with radio-telegraphic, radio-telephonic and other waves. Under such conditions it is easy to understand why any oscillating circuit of any form or dimensions is likely to find in this vast field of waves its own resonance with which it can vibrate harmoniously. It is therefore obvious that to make any oscillating circuit oscillate there is no need to have recourse to a local generator such as the Radio Cellular Oscillator used by Lakhovsky in 1924 when he first treated and cured geraniums inoculated with cancer. The oscillating circuit utilized by Lakhovsky in his second series of experiments -- a circuit having no artificial source of excitation -- proved equally effective in curing geraniums rendered cancerous by inoculation."

Yes, in 1924 Lakhovsky 'graduated' from electronics to radionics! He discovered and proved to his own satisfaction that in some ways a simple, one-turn coil of wire or fibre was as effective as the electric powered Multiple Wave Oscillator. The simple Oscillating Circuit, as Lakhovsky called it, has two advantages over the MWO. It doesn't have to be plugged in to a battery or 110 volt home power supply, and it broadcasts no interference with radio and TV sets. It acts as a step down transformer for life-positive cosmic radiation.

Clement says that "it must be particularly stressed that only Circuits made in strict accordance with Lakhovsky's directions should be used, otherwise the results will prove to be entirely negative." Unfortunately we do not have Lakhovsky's directions, and Clement certainly doesn't give them. You are still free to investigate this type of research and many have. If you want to experiment in this area then get yourself some copper wire (solid or stranded, research should be done on both) in various lengths, (8 inches is one wave length which has worked well on tomato plants) and make them into a circle with the ends overlapping, but with the metal ends insulated or separated from each other. Lengths can be determined through radiesthetic means (that is, by dowsing) or perhaps in geometric relationship with the structure of the plant. Considering the radionic and radiesthetic viewpoint geometric shapes would also have an influence and can be experimented with. Always use control samples in your experiments to see what works and what doesn't.

To get back to Clement's text: "Through the action of this constant field of radio-electric waves which emanates from the surrounding atmosphere the oscillating circuit is made to resonate like an instrument in tune with certain sounds. Thus the oscillating circuit creates a field which canalizes and filters, so to speak, the cosmic waves necessary for cellular oscillation.

THE DANGER OF BURNING X-RAYS

"Furthermore, this filtration is a general phenomenon. It has been discovered, through the investigations of Lakhovsky, that radiations such as those associated with light, ultra-violet rays, X-rays, etc., have the property of making cosmic rays act, as it were, to the advantage or disadvantage of the cells. Now the field created by an oscillating circuit is constant and emits a steady and sustained force. It invariably acts in a favorable manner on the oscillation of the body cells, facilitating by the filtration of cosmic rays, regular cellular division, a fact which explains the immunity imparted to the cells and their consequent capacity of resistance to diseases and microbes.

Rome, published in the medical review Medicina Nuova XXVI, N.5., 1935, Rome: "**Effects On The Nervous System.** Generally speaking I have observed a marked analgesic action following the application of the Multiple Wave Oscillator in all affections associated with pain, particularly in the classical cases typical of such affections. For example, a patient suffering from cephalagia, resulting from violent cranial trauma, for a duration of 10 years, was subjected to a great variety of treatments, but all in vain. The patient was definitely cured after 2 months' treatment with the MWO. Improvement was manifest after only 3 applications.''

Dr. Gentile also reported the effects of the MWO on the Central Nervous System, on Chronic Inflammations, on Metabolism, on Disorders of the Cardio-Vascular System, and on miscellaneous cases. He says he adopted the ''technique of localized irradiation. Generally speaking, treatment is given every 4 days for a duration of 5-15 minutes per session. This is the technique which has given me the best results after having experimented with many forms of application. For children, old people and debilitated women, a relatively weaker application is indicated. If results are not apparent after 7 or 8 applications, I continue until 15-30 applications are given at the rate of a single treatment once a week. I have never observed any ill-effects from the MWO treatment, whatever the number of applications was in any given case. Moreover, the general condition of patients is always improved, even if only temporarily, so much so that the patients themselves sometimes insist on having further treatments. . .''

This chapter on the MWO concludes with four pages of reports of notable improvement or cure by American doctors. Chapter Five is concerned with the remarkable results obtained by using the Multiple Wave Oscillator on animals. Chapter Six contains Lakhovsky's prophecy that some time in the future diagnosis of disease would be obtained by television! This is something which Dr. Ruth Drown's Radio-Vision and Homo-Vibra Ray instruments have already accomplished.

THE NOVEMBER CONVENTION TESTING PROGRAM

There were two Multiple Wave Oscillators available. These were generously furnished by Bob Beck and Ed Skilling, and were set up in the shielded seance room. Only short exposures of from 2 to 5 minutes were given, during the lunch and dinner hours on Saturday and Sunday. In that time perhaps 100 people volunteered for tests. Until a follow up inquiry is made, no statistical conclusion can be made, or even begun!

A few individual cases reported to your Director are worthy of note, including my own. By mid-morning on Saturday a spinal injury from my old, hockey-playing days was acting up. Doctors have called it a spasm in the sacro-iliac area. Ed Skilling invited me to sit in the rays from the MWO for a couple of minutes while the seance room was being readied for testing. Within a minute the spasm relaxed completely and the pain was gone. This was not a permanent change, of course, but the analgesic effect was certainly noticeable.

Mrs. Fred Rannenberg, financial secretary of Harmony Grove and a BSR Associate, told me she had had a sharp pain in the area of her left hip for over two weeks at the time of the Convention. One 3-minute exposure to the rays of the MWO removed this completely. When I attended the Grove Board meeting a week later she said the pain had still not returned.

One Associate did report a negative reaction to the MWO. She attributed it to arsenic poisoning. She had a severe case of this earlier in the year. When the machine was turned on she felt a severe pain through her right side, as though someone had jabbed a knife into her. It could be that some of the lead poison was still in her system and the stirring of the cells in the mid-section produced this reaction. It was endurable, however, and she sat out the exposure without a murmur. The pain persisted that night and she had to put cold compresses on her side to get to sleep. The next morning she felt fine and was eager to take another exposure to the multiple waves. This time there was no negative reaction.

The most startling subject of all was the woman who had been awake all Saturday night with a migraine headache, an old and recurring ailment. By Sunday morning, Nov. 10th, she was sick and nauseated. It was suggested that before taking her to a doctor or hospital she be given a short exposure to MWO rays. Within minutes all pain was gone and she returned to her cottage to lay down and sleep like a baby.

THE PSYCHEDELIC EFFECT

Ed Skilling is the only I know who has stayed in the MWO field or rays for longer than 15 minutes. While testing and adjusting his equipment at his microphone factory several weeks ago he deliberately stayed between the antennas for an hour. At the end of that time he was beginning to feel lightheaded and dreamy he thought it a good idea to get out before losing consciousness altogether.

An even more spectacular possibility is the Time-Machine effect in connection with rejuvenation. If, as Bob Beck suggested in his article, the MWO tends to take the body cell back down the time-track to a more youthful, more vital period, what would happen to a person who stayed in the field of the machine for four hours? Would his vibratory rate be speeded up to where he would disappear completely? Like the hero of H.G. Wells' ''Time Machine''? Anybody want to try it? It seems, from hints George Van Tassel has been dropping over the years, that his Integratron at Giant Rock, California might do something like this -- if and when he gets it working! As I recall, both George and the Canadian Saucerian, the late W.B. Smith, spoke of the four dimensions of electricity: 1. is the electricity moving in one direction along a copper wire, 2. is the magnetic field around the wire, 3. is the static charge created by the first two effects, and 4. is Time -- or the ''tempic field'' as Smith called it. The Time Field equates with Space, where consciousness is everywhere present. Seems to me Lakhovsky's Multiple Wave Oscillator is a gadget which gives us the opportunity -- for the first time -- of approaching the Time Field with scientifically acceptable equipment, operating in known frequencies. Remember, that with its circular antennas it operates on the vortex principle. Each individual antenna in the group is a segment of a 60 degree cone. The 1/2 inch gaps, at alternate 180 degrees from each other, conform to the alternating principle and give a damped or pulsed wave.

THE SPACE AGE EFFECT

Copies of our November 1963 **Journal** have gone all over the world. Each day's mail brings new orders. We are confident that Lakhovsky's invention, far ahead of its time in the 1920s and 30s, has been given a new lease on life and will be pushed to its ultimate development on every continent. There will be vigorous opposition from the right-wing conservatives and reactionaries. Their resistance will furnish the platform from which electronic ''medicine'' will eventually sweep the world.

I see the day when a device, similar to the MWO, will be a standard part of every home bathroom or cleansing center. You will consider it as necessary and as routine to take an electronic showerbath, as to take a water bath. Home from work, tired, dirty and discouraged, you strip, flip on the MWO switch, and while shaving, doing your hair or whatever, re-balance your whole system with a two or three minutes exposure to the pulsing waves from the oscillator as they flood the whole room. For the sale of the house of the future it will be as important for the ad to say ''shielded bath with MWO'' as it will to say ''tile bath with radiant heat.''

Mr. Skilling told me he was tired and

"dragged out" after processing dozens of volunteer "guinea pigs" late Saturday afternoon. Before shutting down the equipment he sat in the nail-less wooden chair between the antennas and gave himself a three minute shot of pulsed waves. Refreshed and revived he looked forward to a good dinner and the full evening program.

I don't know of a better Christmas gift we could have given the Associates than the idea and the plans for the Lakhovsky MWO, researched and so generously offered us by Bob Beck in the true borderland science tradition. Now it is up to you to carry it forward in your own individual way. We welcome your comments, observations and research data as you develop equipment and a testing program.

AND NOTHING HAPPENED!

"No doubt you are deluged with correspondence about your articles on the Lakhovsky Rejuvenation Machine. Since you said it is possible for any intelligent 16-year-old to build one that is exactly that which I have had to rely on. Fortunately, my son is fairly well informed on electronics but obviously no expert at his age, 16. He followed your diagram but instead of battery power he used an electric train transformer, which he assures me, is equally efficient. He also found he had to make one change in the hookup of the Tesla Coil. I have taken four exposures and the results! Nil!"

Maybe this lady expected to be rejuvenated back to her early twenties. As far as we know, no such claims have ever been made for the Lakhovsky gadgets -- though it would be interesting if it happened! As for power supply to the Model T Ford coil, any 6, 8, 10 or 12 volt battery or transformer should do, depending on what you have available and where you intend to do your experimenting. We understand the old Model T magneto delivered as high as 18 volts to the four coils. One handy source of 6 or 12 volt electricity is the cigarette lighter on the dash of the car. Plug in there for power when you want to expose yourself while driving or out in the country.

"I had been fighting a virus for several days and the day after the first exposure I felt a great deal better, but the second day I was right back where I had been. But that was the pattern the virus had been taking. One day I'd feel fairly normal, the next miserable again. So it is difficult to know if the Lakhovsky machine could be credited with the temporary improvement."

That's right, ma'am, some people show little or no reaction to the mytogenic waves put out by the Lakhovsky instrument. The results are unpredictable. Perhaps that's why the research was pushed no further in the 1930s by authorities here in the United States. You, like most people, think that health can be bought in a package. Lakhovsky didn't agree with this erroneous conclusion. His theory was that health is a condition of balance of all the forces of the body. Ill health is a condition of imbalance caused by your misbehavior. If you continue to misbehave, any force applied from the outside, whether it be mytogenic waves, manipulation, drugs, medicine or surgery, will only bring temporary relief. Pain is an insistent teacher, urging you to change your mind; but if you can't change your mind, how can you change your body?

"One possible reaction to the exposure I can report is that I took the first one just before going to bed at night. It took me three hours to get to sleep. Several things puzzle me about the machine we have constructed. First of all, although we have experimented with a Tesla coil of the size your diagram calls for, as well as one almost twice the size and strength, we cannot get a two-inch spark from the antenna or resonator, 1/2 to 3/4 of an inch is the absolute maximum."

The mention of the two-inch spark was Mr. Beck's error. 1/2 to 3/4 inch arcing is all this circuitry will produce and that is adequate to affect notable changes in many forms of cell life. More power might produce faster changes, and much harm! Be reasonable, a living cell is so small as to be invisible without a microscope. Only minute amounts of energy are necessary to affect a cell, to stimulate it or to kill it.

"Second, under no conditions do we get the 'bluish glow of brush discharge' from the antenna that you mention; although we have no trouble lighting up a two-foot fluorescent lamp."

Operate your Oscillator in total darkness and you should see some brush discharge at some points on the antenna. True this redesigned equipment puts out nothing like the five-inch brush discharge visible in the photo of some of the original Lakhovsky equipment in Mark Clement's book, "Waves That Heal", and Lakhovsky's antenna appears to be at least two feet across.

But power is not the answer to balanced cell life! Dr. Abrams discovered this with his early electronic experiments in San Francisco in the 1920s. It is the quality and the variety of the pulsed waves that counts! In any event weaker equipment can be compensated by longer exposure, and this may be more beneficial to living matter.

"And thirdly, since we have not constructed any outer shield for the machine, such as a Faraday cage of copper screen, I have been monitoring the operation on our radios and television. On the radio, one notes an annoying and loud buzz, but on T.V. except for a couple of light horizontal lines and a very slight hum, there is no serious effect. So I doubt that we're bothering our neighbors very much. In any case, we try to use the oscillator very early or very late in the day, when few people are using their radios or TV. I have sent for Clement's book, 'Waves That Heal,' and Lakhovsky's 'The Secret of Life', not only to better understand what we are trying to work with but also in the anxious hope that I will find out why our machine does not have the affect it apparently should have."

Let me assure you, good lady, there is no single cure-all for man's ills. The clue to better health is in the profound statement of the Savior of our race: "As a man thinketh in his heart, so is he." Pain is still the great teacher in the world and will be for a long, long time to come. I don't know what your physical condition is and it's none of my business; that's between you and your doctor. If you have a long history of frustration and thwarted ambition, going back to childhood and even into past lives, this mass of energy of crystallized thought-forms, is not going to be broken up and dissipated overnight. If you have consciously and willfully flouted Cosmic laws, or have turned them to your own selfish ends, the consequences of this misbehavior will have to be worked out at this level. The occult scientist calls this Fixed Karma. In this connection, one of my favorite teachers, Dion Fortune, reminds us: "What cannot be cured must be endured."

If the Lakhovsky oscillator does have an analgesic affect -- as it appears to have had in some cases -- it has this advantage over pain-relieving drugs; it does not leave a residue of poison in our already over-loaded liver and kidneys!

WORD OF A SIMILAR INSTRUMENT UNDER DEVELOPMENT

We have a second letter from this same lady Associate, also worthy of note in connection with the Lakhovsky Multi-Wave Oscillator.

"First off, when my brother received the machine we had built for him, using the battery indicated in the schematic, he felt the power source might be too weak. It so happens he has a business contact with one of the largest electronic instrument companies in the area where he lives. He decided to go straight to the president and seek his advice. With the whole idea placed before him this gentleman said, 'This is very interesting. I want you to meet one of our physicists who is working on a very similar instrument.' A meeting with the physicist revealed two interesting pieces of information: first, the physicist had never heard of Lakhovsky or his machine; second as soon as this new instrument was completed, arrangements had been made to test it in a large nearby hospital! Is this latter not a cheering bit of news?"

Both cheering and discouraging. The basic principle discovered by Lakhovsky is certainly available to the mental probing of other sincere researchers, so let us keep it alive by all means. But barring a revolution in the ultra-conservative leadership of the American Medical Association, no electronic developments will be allowed to threaten the high-profit position of the drug industry.

"In their discussion of the power source, the physicist suggested my brother use a wet-cell car battery, #50, 39 plates, 70 A.H. We are using one too and it makes a much better spark from the antenna, although it still is no greater than 1/2 to 3/4 inch. It does seem stronger. Of course the machine is not so readily portable with such a heavy battery. My brother had only one exposure at this writing but he reports a noticeable improvement in the arthritic pain bothering him in his right arm for months. On the other hand, the report on the use of my own machine is somewhat less encouraging. The only positive results concern my husband. After his first ten-minute exposure, within a few hours, a painful bursur in his shoulder cleared up. He has been subject to off-again, on-again attacks of bursitis there for some 15 years; for which he had taken medication, short wave, and even X-ray treatments, when the attacks were severe. He has since taken another exposure, but neither this nor the first one seemed to help an itching-skin allergy which has bothered him for some years.

NOT THE FOUNTAIN OF YOUTH SHE EXPECTED!

"As for myself, I have taken at least a half-dozen exposures and cannot in all honesty detect the slightest difference from them. I'm not sure if it is a matter of 'suggestion' or a desire to prove the capabilities of the machine. On the contrary, I notice such strange contradictions as the fact that although I took an exposure yesterday morning, in the evening the fourth and fifth fingers on my right hand were painful, from some kind of arthritis or rheumatism. I had never had anything like this before. Does it not seem odd that it should develop on the day of the exposure?"

Not at all! Any positive reaction to exposure to the mytogenic rays of the Lakhovsky oscillator indicates that it is working! Mr. Lakhovsky's theory is that the waves or oscillations excite the cell, causing it to release mor energy, also to shake off accumulated poisons and/or tensions. If your system is loaded down with tensions from daily living in America, if your system is loaded down with poisons from processed foods and contaminated air and water, you can expect painful reactions when the junk breaks free. I believe that is what sickness is the body struggling to get rid of poison. For you, things may get worse before they get better!

HOW OFTEN AND HOW LONG?

"With the mention of two successive exposures in two days I get to what is the most important question in the use of the MWO. No one has explained why exposures are suggested at four-day intervals. Is it harmful or useless to take them more frequently? When one is uncomfortable or in pain, one is anxious to rid oneself of the discomfort as quickly as possible. And to have to wait four more days for additional improvement seems unnecessary if there is no harm in the exposures. I have exposed myself at more frequent intervals and still have observed no reactions?"

The suggested exposure rate of fifteen minutes every four days came from Dr. Gentile's extensive researches with Lakhovsky equipment in Rome in the 1930s. This is a typical professional approach to treatment of disease by doctors. From long experience with drugs and other proven forms of therapy, they know that it takes time for the human organism to adjust itself to radical changes -- without harming the organism.

Whether or not this holds true for the operation and effect of the Lakhovsky machine remains to be proven and you'll have to be your own guinea pig. It's your machine; it's your body; it's your life. You can take an exposure every hour on the hour if you like; stay in all day. Who's going to stop you? This is borderland research and you are a borderland scientist, looking for new knowledge which may be of benefit to yourself and the human race, judging from the last lines of your letter.

"I am so excited and hopeful of the efficacy of this MWO machine that I do not want to leave any avenue unexplored that may possibly help in understanding its use and making it freely available to anybody I can possibly help. May I take this opportunity to tell you how tremendously enthusiastic and interested I am in BSRA, not only what you are trying to do but what you are accomplishing, too. From what you imply of the financial situation, it seems to be largely a labor of love and service on your part, for which you deserve great praise and the heartfelt thanks of every BSRA member and others whose lives they touch. Please accept this very, very small donation of the enclosed check, as a symbol of my appreciation."

The donation is most welcome and will be put toward the expenses of carrying on this work, also the words of appreciation. The lady was also most curious about the Lakhovsky circuits, those single bands of wire which are placed around the waist, or knees, or ankles, wrist or neck, and seem to have a subtle but very real affect on the wearers. We have some information on the Lakhovsky Circuits or Belts from Associate A.E. Powell, a retired lieutenant colonel of the British Indian Army:

THE OSCILLATING CIRCUIT

"The Belts can be made by almost anyone, for a few cents. At most hardware stores, for a few cents per foot, one can buy twin electric cable insulated with white rubber, used for lamps, etc. Get the smaller size 1/4 inch wide, having 42 strands of copper wire. Slit an end with a knife and the two wires easily tear apart.

"For the button, get white sheet plastic about 1/8th inch thick, saw it into pieces about 1½ by 5/8th inches, and drill near each end a hole large enough to take two wires. Round the corners of the plastic and bevel edges. Thread the wires, from opposite directions, up through one hold and down through the other, leaving about half an inch projecting beyond each end of the button. Where a circuit can be slipped off the body without unbuttoning, you may overlap the ends of the wire belt an inch or two, bind with fine string and smear with Duco cement for strength.

"Parallelism and overlap of the ends are important. This builds up capacitance, produces inductance and generates an alternating current. Circuits may be worn, as many as you please, round chest, waist, neck, arms, wrists, legs, ankles, neck or head continuously. Water will do them no harm.

"Gladly will I supply Circuits, in any quantity, with buttons or bound ends. To avoid trouble with the Food and Drug people, I make no charge. Donations, however, are acceptable, and go intact into my Fund for the Sick. For long this has been doing fine work and I'm sure you will approve. For reasons stated I make no claims whatsoever for the Circuits, and no guarantee or promise of any kind. If you want to give them a trial, that is wholly your affair." Col. A.E. Powell (Col. Powell died in March 1969).

MWO USED TWENTY YEARS AGO

"The recent article in the BSRA Journal pertaining to the Lakhovsky Oscillator is of special interest to me. First, I am an engineer, Physician and Surgeon, and a Scientific Consultant for the U.S. Government. (Technical Intelligence). I am not in any manner snooping for the government. This is completely personal. We built several oscillators in 1944. These were used in my sanatorium in Dallas where we had some startling results. As we did not have the proper facilities to research the instruments adequately, we became more and more conservative. Many patients would show reactions, resembling the old fashioned Flu, after a few days following exposure to the oscillator. I realized we were using too much power, so cut the level several times with good results; but we did not completely eliminate the side effects. In 1946 I was sent to Germany for the Government to explore Electro-biology and Biophysics in its application to electro-medical equipment. I felt it wise to destroy the machines so injudicious use would not be indulged. The possibilities have intrigued me ever since. It looks as though you have solved most if not all of the problems involved. Hence my renewed interest. The excess power upset the enzyme complex of the body, or as some felt, distorted the double electric layers of charge carried on the cells, hence allowing for the penetration of virii directly into the cells (virii operate within the cells -- bacteria operate between the cells). Cells as well as organs function in a force field so mild oscillating currents not only work through resonance but seem to regenerate these force fields. Here is where application to cancer can be optimistically explored."

Speak for your own state, Doc! Here in California the medical association has decreed that only four forms of treatment for cancer are legal; drugs, surgery, burning rays and death.

"Proof of energy flow from the fingers and thumb is here and must be recognized. We have demonstrated it by its effect on Yeast cells, viewed under the microscope (Girwich research). Whether it comes from the conscious rather than the subconscious mind is harder to prove, but I am impelled to believe it does. This may interest you. Years ago we experimented with crystallization. Precipitated water vapor to snowflakes in a chamber with electrodes planted in two sides. When we threw a current into the chamber at a given frequency the flakes were all the same. Changing the wave length changed the crystallization pattern. Each frequency had its own pattern. Here is the explanation of energy absorption by the chemical used, and the control of the crystalline patterns by the nature of the energy flowing."

Naturopaths are in 100% disagreement with the doctor's above explanation of the side affects of his use of the MWO on his patients. If I understand the theory correctly, according the natural view of colds, flu, virus infections, etc., these are attempts by the human organism to throw off mucous, to eliminate intolerable accumulations of poison from the system. Exposure to the field created by the MWO tends to raise the general tone of the whole system by stimulation. The raised vibrations release an unusual amount of mucous and other junk into the elimination channels, and until this is eliminated one would have a cold or flu and the usual aches and pains which generally go with such a condition.

Another doctor stopped by here not long ago to tell of his sad experience in trying to use spark-gap electronic equipment in his medical practice "for those suffering patients not helped by drugs." This powerful, high-frequency machine produced a two-inch spark! Every time he turned it on, everyone in town seemed to know it and the phone would ring. Even a Faraday cage didn't shield the output of this monster. Finally he was called before the county medical board to explain what he was doing. The board presented a staggering combination of fatuous conceit and stubborn ignorance. "They didn't want to know anything about the affects of radio waves on the body and didn't care!"

CURING CANCEROUS PLANTS WITH ULTRA RADIO FREQUENCIES

by Georges Lakhovsky

From Radio News February 1925

"The new application of short wavelength oscillations described in the following article" wrote the Editor of Radio News, "is one of the most important ever found, and we are pleased to present to our readers this article by Mr. Georges Lakhovsky, the well known French scientist. The experiments described were carried out in collaboration with several doctors and scientists of high standing and it seems from results obtained, that the very high frequency treatments will play an important role in the future. We shall publish in a later issue another article on this new system together with the effects of ultra radio frequencies in the treatment of diseases in human beings."

Since November 1923 I have published in various technical and radio publications, several papers in which I explained my theory that the instinct or special feeling, which permits birds to direct themselves in space, is only the results of the emission and reception of rays by living beings. While developing this theory, I explained how thoroughly I was convinced that science will discover, some day, not only the nature of microbes by the radiation which they produce, but also a method of killing disease bacilli within the human body by means of the proper radiations.

The researches I have made by means of special apparatus have shown such results, that I believe my theory is correct. This theory is that life is born from radiations, kept going by radiation and suppressed by any accident producing the destruction of the oscillatory equilibrium, especially by the radiations of certain microbes, which suppress those of the weaker cells.

Before going any further in our reasoning, it is necessary, in order to present the facts to the uninitiated reader, to imagine what oscillations really are. The motion of a pendulum will be used for this explanation. When a pendulum is displaced from the position of equilibrium, it moves back and forth producing what are known as ochrone oscillations, until the energy stored is entirely exhausted. By means of a motor, a spring, or an electro-magnet, it is possible to keep the motion of the pendulum of constant amplitude, producing undamped oscillations. If, on the contrary, the source of power is removed, the oscillations die down and it is necessary not only to re-apply the power sustaining the oscillations, but also to furnish additional energy to start the pendulum in motion. This oscillation of a pendulum reproduces exactly what happens in the cells of a living being.

Our organs are composed of cells formed of protoplasm containing various mineral matters and acids, such as iron, chloride, phosphorous, etc. It is by the combination of these elements that the cells detect outside waves and vibrate continuously at a very high frequency, probably higher than the period of X-rays or over all other vibrations known and measured today (1925). The amplitude of cell oscillations must reach a certain value, in order that the organism be strong enough to repulse the destructive vibrations from certain microbes.

COSMIC RAYS AND CELL LIFE

The astrophysicians are actually carrying out experiments of great interest on the existence of vibrations, which have been called penetration rays and of which the frequency is higher than that of X-rays and of the alpha, beta and gamma rays of radium. Such rays, according to the theory, are produced by the earth itself and some others come from outer space. Some accurate measurements have proved the correctness of this theory. Therefore, it is quite permissible to believe that these penetration rays, or at least some of them, produce the vibratory motion of living cells and consequently their life.

For instance, let us suppose a cell vibrates at a certain frequency and a microbe vibrates at a different frequency; the microbe begins to fight the cell, and sickness is started. If the cell cannot repel the stronger vibrations and if the amplitude of its own vibration decreases, the microbe gains and its vibrations in decreasing and stopping those of the cells, bring dangerous sickness or death. If, on the contrary, the living cell is started vibrating with the proper amplitude by inside or outside causes, the oscillatory attack is repulsed. Such is my theory.

The problem is somewhat similar to the situation in which a rescuer finds himself when, coming to help a friend in a dangerous situation, finds himself fighting hand to hand against strong aggressors. The rescuer does not dare to fire his gun; fearing to harm his friend mixed up with the aggressors in the melee. Similarly, microbes and healthy cells are exposed to electric or radio-active action, which could be used to destroy the unwanted rays and it is difficult to suppress them without harming or killing at the same time, the cells which are to be treated. In fact, since Pasteur, scientists have been constantly searching for means of destroying microbes. The great difficulty with all methods found was that in destroying the bacillae, the cell was attacked too. The experience gained in cancer and tuberculosis treated with radium, X-ray, or ultra-violet rays, shows how difficult is the work of the investigators.

A RADICALLY NEW METHOD

The remedy in my opinion is not to kill the microbes in contact with the healthy cells, but to reinforce the oscillations of the cell either directly by reinforcing the radio activity of the blood or in producing on the cells a direct action by means of the proper rays. During January 1924 I began to build, according to this theory, and with the purpose of therapeutic applications, an apparatus which I have called Radio-Cellular-Oscillator, with the firm belief that the cells vibrating at extremely short wave-lengths would find their own in the Hertzian waves, which have the properties of producing extremely short harmonics. The cell with very weak vibrations, when placed in the field of multiple radiations, finds its own frequency and starts again to oscillate normally through the phenomenon of resonance. This type of vibration which I propose to use, is harmless, unlike those of X-rays and radium. Their application, therefore, does not present any danger to the operator.

I exposed in front of my apparatus, during long periods, a certain number of microbes in culture, which developed themselves normally. I, myself, have never felt the effect of these ultra radio frequencies, although I remained for a great many days near the apparatus, during the treatment applied to the living cells. It is only when two living beings such as a cell and a microbe, are in contact, that the rays produced by the Radio-Cellular-Oscillator have any direct effect upon cellular structure.

WORK IN A FRENCH HOSPITAL

The experiments which I carried out at the Salpetriere Hospital in Paris, in the service and with the collaboration of Prof. Gosset, were made with plants inoculated with cancer, and the results were described in a paper presented on July 26, last year, before the Biological Society. The text of this paper follows.

"One knows that it is possible to produce by inoculation of Bacterium tumefaciens in plants tumors similar to those of cancer in animals. One of us obtained experimentally by this method, a great number of tumors. These had various degrees of development. Some of them dry up partially, but do not die entirely until the entire plant or at least the limb bearing the tumor dies. Even removed by surgical methods, these grew again on a sick limb.

"We propose to describe in this paper, the action of electromagnetic waves of very high frequency obtained by means of the Radio-Cellular-Oscillator of Georges Lakhovsky. This apparatus produces wave-lengths of the order of two meters and less, corresponding to 150 million cycles per second. A first plant was submitted to the effect of radiation one month after being inoculated with cancer; at this time small tumors of the size of a cherry stone were visible upon it. This plant was submitted to the rays twice, for three hours each time. During the following days, the tumors continued to grow rapidly in the same way as those on plants, which had not been submitted to the effect of radiations. However, 16 days after the first treatment, the tumors began to shrink and dry up. A few days later the tumors were entirely dried up and could be very easily detached from the limb of the plant by merely touching them. The drying action of the radio frequency radiations is selective and affects only the sick part of the plant. Even the inside sick tissues were destroyed, although they were next to healthy cells in the center of the limb, showing that the radiations had not affected the healthy parts.

"Another plant was treated in the same way, except that it was exposed 11 times for three hours each time, to the radiations of the oscillator. Sixteen days after the first exposure the tumors, which were rather large as shown in one of the photographs, began to shrink and dry up and were easily detached from the limb exactly as in the first case. Again in this case, the healthy parts of the plant were not affected in the least. A third plant exposed to the radiations for nine hours, that is, three treatments of three hours each, was cured in the same manner as the two others. Sixteen plants also inoculated with cancer, were left without treatment. They have tumors in full activity, several of which are very large. These experiments show conclusively that plants inoculated with cancer can be treated and cured by means of the ultra radio frequency vibrations, whereas surgical treatment fails.

"Such are the results of my researches with plants. At the present time, similar experiments are being carried out with animals and it seems that the effect on cancerous animals is the same as on cancerous plants.

"I am highly pleased to present my theory and the results of my work in a scientific review of the United States, this great country, which has always been in the lead in the fight against this horrible sickness, cancer, and whose sympathy and help I would greatly appreciate.

"In conclusion I wish to call the attention of the reader to the fact that I have obtained very conclusive results not only with a wave length of two meters, but with longer and shorter wave-lengths. The main thing is to produce the greatest number of harmonics possible."

* * *

This February 1925 article from Radio News magazine shows how long some of Lakhovsky's theories have been known in America. This article was furnished to BSRA by an Associate who had met the Russian-French scientist in New York. He claims also to have visited several times the New York hospital which set up a research program on the Multi-Wave Oscillator and saw the Lakhovsky-type equipment. No information on the machine or its effects could be elicited from the hospital authorities and in the early 1950s the shielded room was dismantled and equipment destroyed.

We feel that Lakhovsky's theories and early researches should be followed up. It should be publicly proven that high-frequency waves, in the proper concentrations, do affect living organisms in a beneficial way, or that Lakhovsky be declared an idle dreamer. Since research of this kind falls within the purposes of Borderland Sciences Research Foundation, Inc., members are encouraged to build their own equipment and, in a suitably shielded or isolated environment, carry on their own research program. If you cannot obtain the necessary materials, BSRF may be able to put you in touch with another borderlander who can be of help. Until the medical authorities come forth with official approval of this radical approach to the state of imbalance or un-health attributed to disease, no one but a fool would go into commercial production of this kind of equipment in America today. It is now said to take a minimum of $300,000 in research money and three years of testing to develop medically acceptable data.

HOW DR. JAMISON WAS TRAPPED

In a letter dated Aug. 29, 1967, Dr. Jamison enclosed this news clip from the **Modesto Bee**: "Turlock – Veterinarian Stanton L. Jamison yesterday was fined $249 in Turlock Judicial District Court after being convicted of three charges stemming from his use of an electronic machine to treat human patients for cancer. A jury of nine women and three men returned a verdict of guilty on one count of practicing medicine without a license and two counts of advertising a device represented to have effect on cancer, stomach ulcers or bone diseases. The decision followed a two day trial in which Jamison defended himself. . .

"They (Jamison and his assistant, Martha Ann Davis) were arrested last December following a lengthy investigation by Turlock police officers and agents of the State Bureau of Food and Drug Inspection. The investigation began when Police Chief John Viarengo requested assistance from the bureau after learning 'the cancer-curing machine' was being operated here.

"During the investigation, Food and Drug agents Willis Worley and Harvey Lee arranged several meetings with Jamison on the pretense they were interested in purchasing machines similar to his. Jamison reportedly organized a church as a front for his operation. The facility, located at 200 West Main St., was known as the Life Aquarium Center. Worley and Lee became bishops of the church as did deputy District Attorney Frank Damrell, Jr., who prosecuted the case. At the time of the investigation, Damrell was employed in the State Attorney General's office.

"Worley said Jamison's intention was to open other treatment centers throughout the state, using his bishops to set up facilities under protection of the church. 'This was the most fantastic investigation I've ever been involved in,' said the agent this morning.

(Yes, it is fantastic to think anyone would want to give or donate healing services free when there are billions to be made on the four kinds of cancer treatment authorized by the California Medical Association: Drugs, Surgery, Radiation and Death. RHC.)

"Worley points out the manufacturer of Jamison's machine was convicted of grand theft two weeks ago in Los Angeles after selling a similar machine to a Food and Drug agent under false pretenses. (This was probably Ralph Bergstresser, Burbank, Calif., who was manufacturing and selling the Bob Beck MWO for $250.)

"After Jamison's conviction yesterday, Damrell called for imposition of a prison term for the veterinarian. If he had received the maximum penalty, he could have been sentenced to 18 months in jail and fined $1,500.

"We felt very strongly on this," said Worley, "because Jamison could easily become a menace to many innocent victims throughout the state."

MWO IN USE AT INDIAN CLINIC

"I am continuing my experiments with the MWO and my experience is that in practically all cases there is an aggravation of the existing symptoms, and previous symptoms are brought to the surface. Hemorrhage was seen in another case and treatment was stopped forthwith. This is one problem which has to be tackled before MWO can be used with confidence. I think MWO treatment should not be given to tubercular cases. Also in cases where there is a hemorrhagic tendency this would do more harm than good.

"Neuritis seems to respond nicely to this treatment. Sciatica is definitely cured by this. Some cases of Rheumatism and Arthritis are benefited, responding at first but the improvement does not last much longer. I cannot understand the reason for this eccentric behavior. I think there must be something with the frequencies."

Don't blame it on the frequencies, Doc, blame it on the patients who stubbornly clings to their fixed ideas and rigid unyielding mental patterns. We have no evidence that the radio frequencies of the MWO reach into the mental world. Each person's aura is loaded with "signatures", national, family and religious concepts hammered into him over the years. His creative energies, flowing freely down from the higher worlds, channel through these deeply cut "grooves" in mind and emotions. The health or illth of the physical body indicate whether the person's ideas are in tune or out of tune with Creation. As a man thinketh in his heart, so is he.

"Can you please let me know what are the ranges of frequencies these machines have? Have they been measured and are they the same in all instruments? Is there any difference in the models prepared in the USA and France? Which model was tried by Lakhovsky in the New York hospital? What are the actual frequencies used by him?"

The only frequency readings we have on an MWO were those submitted by Bob Beck originally, from the output of his own machine. "It radiates RF interference over a bandpass of 15 Kilocycles to well beyond 250 Megacycles. This blocks AM, FM and Television reception for hundreds of yards" But so does my electric razor, and I've been using one for years. I was told by a sailor that electric razors are very effective as "scramblers" for knocking out the electronic brains of guided missiles, and have been so used as defensive equipment during missile exercises on aircraft carriers at sea. No one has furnished us with RF output figures from original Lakhovsky equipment, either at the New York hospital or from Europe. If anyone has these figures I hope they will be kind enough to forward them to you. Lakhovsky's available technical record is very slim.

"An Associate writes 'A very good psychic in Florida gave the following, in a sitting devoted to the MWO. It is not anywhere near developed to its potential. It has unlimited amplifications to be made. The use of the MWO can change the auric field, thus spiritually it can change the individual. This is done with the work of the seven rays of life. The MWO should be used in conjunction with the radionics instrument.' What about this radionics instrument? I would appreciate it if you can throw some light on this."

Sorry Doc, the punitive hand of the American Medical Association has driven radionics builders and users underground here in the United States. No one but a fool like me would stand up on the firing line. England isn't quite the police state America is and you can get information on radionics and radionics equipment from the Delawarre Laboratories, in England, and from Bruce Copen, in Sussex, England. Radionics experts have developed rates for every organ in the body. Equilibrating those rates with specific radio frequencies offers an interesting field for borderland research.

"I have tried both the Vitic and the Eeman Screens. I am still continuing my experiments with the Eeman Screens. Only a couple of days back I gave a set of screens to a patient who is having an encyclopedia of complaints. The MWO did not give her much improvement hence the change. I have one set of Vitic with me and whenever I feel a bit out of sorts I sit for ten minutes and get back my vigor. I am using a very strong pair of magnets and a carbon rod 1½ inch in diameter. After I received the MWO my interest in the other two flagged. I think I will have to cultivate them again."

Dr. A.K. Bhattacharya

From the March 1964 Journal of Borderland Research

ARMY MEDICAL RESEARCH SUPPORTS LAKHOVSKY THEORY!

That's our headline, and it's derived from news clips loaned to us by an Associate. One especially is from the New York "Times" for Friday, Aug. 19, 1960 by John Osmundsen.

The heading says that Radio Microwaves Produce Molecular Changes in Human Beings. This is merely a repetition of what Georges Lakhovsky proved with his pioneering work in Paris hospitals in the 1920s!

"New and possibly important biological effects of radio and radar waves were reported at a three-day scientific meeting that concluded here yesterday. One scientist told of producing 'a profound molecular change' in human gamma globulin with specific wave lengths of radio waves in the high frequency to very high frequency, or VHF range.

"Alterations so created in the molecules that constitute an important part of the body's defense apparatus were accompanied by at least a fourfold increase in their biological activity, the scientist said. 'The changes can be produced under widely varying conditions of voltage, power, pulse width, and pulse repetition rate, provided the frequency is suitable,' according to Lieut. Col. Sven A. Bach of the Army Medical Research Laboratory in Fort Knox, Ky."

If Col. Bach had been working from Lakhovsky's original theories and equipment, he wasn't saying. This is obvious in reporter Osmundsen's next paragraph.

"This is believed to be the first report of a proven frequency-specific effect of radio waves on human chemistry. Dr. Bach's report appeared to cause more comment at the conference than almost any other, for the effect of radio frequencies on human gamma globulin in the test tube were clearly not a result of heating.

It was suggested by Dr. Bach that the increased biological activity of the molecules as measured against a rabbit serum sensitized to them might have resulted from an unfolding of their structure, thus exposing more active spots."

YUP, THEY MIGHT EVEN CURE CANCER

Later in 1960 an electronics magazine revealed even more about the proceedings of this Fourth Annual Tri-Service Microwave Conference held in New York. Our news clip is only half a page which reveals neither the magazine nor the writer, but about Dr. Bach's claim of a four-fold increase in the production of gamma globulin, he says:

"What does this mean to you? Gamma globulin plays an important part in the body's defenses against disease. Speed up its activity fourfold and it may be four times as effective a disease fighter. Some day your doctor may give you 'a shot of microwave' to keep you well.

Dr. Joe Howland of the University of Rochester found that dogs which had been exposed to microwaves were able to withstand heavy doses of X-rays far better than dogs not so treated. No specific application for this technique is now known, but such discoveries frequently lead to undreamed-of advances.

"And finally, some evidence has been gathered to indicate that certain kinds of cells, under the right conditions, will simply disintegrate -- fly to pieces -- in the presence of the right radiation. But surrounding cells are unaffected. Could such a selective beam ever be used to destroy cancer cells while leaving healthy tissue unaffected? Scientists don't yet know."

The writer didn't say which scientists he is speaking for, certainly not for Georges Lakhovsky, who died in that very same New York City 17 years before the Microwave Conference. Before-and-after photos of skin cancer cures effected by the Lakhovsky Multi-Wave Oscillator in Paris hospitals in the 1930s show that the Russian inventor did know just how selective radio microwaves can be. These pictures are reproduced in Lakhovsky's book, "The Secret of Life".

To continue with the magazine article: "said Col. George M. Knauf, one of the foremost experts on the biological effects of microwaves, speaking of these recent advances, 'Our work has progressed sufficiently for us to be fairly certain that the beneficial uses to which this energy may be put will greatly outweigh any harmful results it may produce in men.' All this of course does not mean we should forget that uncontrolled microwave energy can be dangerous. But like fire or W-rays, microwaves can, when used intelligently and with understanding and respect for the hazards involved, be among man's most valuable tools."

MULTI-WAVE OSCILLATOR USE IN A NEW YORK HOSPITAL

Georges Lakhovsky crossed the Atlantic to New York City in 1941, there to observe the effects of his MWO under controlled, technical conditions in a New York Hospital. A brief report of the positive and beneficial results of MWO treatments is appended to the English translation of his book, "The Secret of Life", by Mark Clement, on page 201. These treatments were given from July 1st to Aug. 21st and the name of the hospital and of each patient are withheld for ethical reasons.

Patient	Illness	Treatments	Results
X.	Arthritis, both knees	14	Good improvement
O.P.	Endocervicitis. Abdominal pain	6	Condition good
F.T.	Periarthritis of shoulder	12	Marked improvement after first treatment, condition good
M.M.	Osteoarthritis	7	Marked improvement after two treatments, stiffness diminished
M.K.	General Chronic Arthritis	11	Slight improvement
M.O.	Arthritis of ankle	3	Results good
A.E.	Post-operative fracture	6	Results good
C.V.	Arthritis, both knees	14	Marked improvement
M.D.	General arthritis	8	Marked improvement
J.H.	Epicondylitis of elbow	7	Results good
D.G.	Chronic arthritis	10	No cough, no soreness, marked improvement
S.L.	General arthritis	10	Marked improvement
B.M.	Arthritis and circulatory disturbances	6	Improving
M.L.	Arthritis of ankle and knee	7	Improvement, no recurrence
M.B.	Arthritis of shoulder	7	Marked improvement
M.O.	Tenosynovitis, hand	6	Improvement
S.S.	Arthritis, both knees	7	Improvement

S.C.N.	Congenital hip dislocation	3	Marked improvement
R.L.	Fracture, both ulnae.		
	Synovitis, both knees	5	No pain

Additional Cases

Case 1. G.D., female, aged 25: Patient complained of pain in the chest and muscles of the right arm and back, and constant cough. She did not respond to medical treatment and massage. Treatment with Lakhovsky Multiple Wave Oscillator was instituted. After the fourth treatment coughing stopped, pain in the chest became infrequent, and the tired feeling in the back disappeared. At the end of 12 treatments the patient no longer complained of previous symptoms.

Case 2. M.M., female, aged 51: Patient complained of pain and stiffness in the right hip. X-ray examination disclosed a marked narrowing of this articulation associated with large spurs at the margin of the acetablum which appeared to ankylose the joint. There were proliferative changes in the upper margin of the femoral head. Diathermy and massage failed to relieve the pain and stiffness. After 10 treatments with Lakhovsky's M.W.O. the stiffness markedly reduced.

Case 3. F.T., female, aged 38: For two months the patient complained of pain in both shoulders. Abduction of the right humerus was limited to 45 degrees. A diagnosis of periarthritis was made and treatment with Lakhovsky's M.W.O. was started. After three treatments the patient said she felt much better. The improvement continued. After a series of 11 treatments the patient did not return. Apparently permanent relief had been obtained.

Case 4. A.B., male, aged 60: Patient gave history of pain in both knees beginning in 1940. Diathermy and massage for one year gave only slight relief. Treatment with Lakhovsky's M.W.O. was instituted. After one treatment patient stated that he felt better. After four treatments patient felt so much better that he took only one treatment in two weeks.

Case 5. M.M., female, aged 42: Patient complained of pain in lower left back and hip. X-ray examination of the lumbar spine and pelvis showed calcification at the anterior margins of the inter-vertebral discs between the ninth and tenth and the tenth and eleventh vertebrae. There was also a slight scoliosis of the lumbar spine. Treatment with Lakhovsky's M.W.O. was instituted. After the first treatment patient felt better. After the fifth treatment improvement was very marked. Patient did not return after nine treatments.

Case 6. S.N., female, aged 59: Patient complained of pain in the right shoulder and inability to extend the arm more than 5-10 degrees, accompanied by severe pain. X-ray examination disclosed irregular calcification in the region of the greater tuberosity of the right humerus. Treatment with Lakhovsky's M.W.O. was instituted. After one treatment patient said she felt much better. The improvement continued steadily until 10 treatments were completed.

Case 7. C.P., female, aged 38: Gynecological examination showed uterus slightly enlarged and hard. Right para-metrium tender. Diagnosis --Parametritis. Treatment with Lakhovsky's M.W.O. was instituted on July 17, 1941. After four treatments patient felt better, and after 12 treatments patient was found to be much improved. Examination on Oct. 7th showed the adnexa and parametrium to be painless. Patient was considered cured.

REPORT OF CASES TREATED WITH LAKHOVSKY'S MULTIPLE WAVE OSCILLATOR BY A PROMINENT BROOKLYN UROLOGIST

The following results were obtained by a prominent Brooklyn urologist whose name must be omitted for ethical reasons. He is a Fellow of the American College of Surgeons and enjoys a great reputation as a skilled specialist. He treated hundreds of patients with Lakhovsky's MWO but only a few typical cases can be given here.

H.G.	Acute urinary retention due to enlarged prostate	2	Retention ceased
R.B.	Bleeding from uterine fibroids of two years standing	3	Bleeding ceased
S.M.	Mild multiple sclerosis with halting gate	6	Marked improvement, was able to run on several occasions
R.R.	Cellulitis of nose with pain	2	Cleared completely after two treatments of ten minutes each in one day.
B.T.	Acute urinary retention due to enlarged prostate	2	Retention ceased.
S.B.	Bleeding from cancer of the bladder	3	Bleeding checked. The tumor cleared up nicely and was reduced in size when seen again on cystoscopy three weeks later.
J.G.	Chronic aczema of Axilla	2	Cleared entirely.
L.G.	Severe pains form duodenal ulcer	2	Greatly improved.

In addition to the above cases this Brooklyn specialist treated six cases of enlarged prostate with nocturnal frequency (three to five times per night). After two treatments with Lakhovsky's M.W.O. frequency was reduced "to only once a night or did not occur at all..."

The above American reports were not included in the first edition of Georges Lakhovsky's "The Secret of Life", published in 1935, but were added by the English translator, Mark Clement, to later editions. We are quoting from the 1963 edition published by Health Science Press, Sussex, England. Health Research, (P.O.Box 70 Mokelumne Hill, California 95245) offers this same book.

This early (1941) research with the Lakhovsky equipment seems to nave been ignored by the American medical authorities, who are still concentrating on cure by surgery or medicine. If you want to discover the effects of weak radio waves on living organisms you'll have to dig for yourself. Our BSRA brochure, "The Lakhovsky Multi-Wave Oscillator", will give you some hints about setting up your own research program, and is available from headquarters.

'DOCTORS SAY RADIO TREATMENT IS EFFECTIVE AGAINST ARTHRITIS'

Apparently some doctors didn't ignore the research work done by Georges Lakhovsky. Take this news item from the New York "Post" for Feb. 8, 1962, with an Associated Press dateline from Washington, D.C. "A device theoretically capable of bouncing radio signals off the moon has yielded highly effective results in the treatment of stubborn cases of arthritis and certain other ills, a group of doctors reported today. The machine generates high-frequency, short-wave-length radio waves and emits them as pulses lasting only 1/17,000th of a second each. The pulsed currents are directed into a patient's body, with a maximum achievable penetration of eight inches.

"Use of the technique in treating arthritis, bone infection after hip operations, inflammation of the female pelvis due to gonorrhea and other conditions, and in speeding the healing of surgical wounds was described in reports by Dr. Euclid M. Smith of Hot Springs, Ark.; Dr. Dana Street of the University of Arkansas; Dr. Bruce Camern of Baylor University; Dr. Marshall Lobell of Harlem Hospital, N.Y.; and Dr. Solon N. Blackberg of Chicago. The reports were made last night at a private conference to which several Congressmen and Public Health Service doctors were invited. Blackberg later told reporters the researchers would like to enlist government support for expanded research by themselves and their investigators.

"Smith, in his report on arthritis, said the technique had been employed with "highly "effective results in 99 out of a series of 100 cases treated by him. Declaring that some of the patients treated had arthritis for up to 25 years before receiving radioactive treatment. Smith told reporters that "marked improvement" was achieved in all cases in from three to six months.

PUBLIC RECOGNITION OF RADIOWAVE TREATMENT

This sensational advance in the treatment of human ailments should have been welcomed from coast to coast in every newspaper. Dr. Blackberg and his distinguished fellow scientists from Baylor, Arkansas and Harlem should have received public acclaim from President Kennedy's Dept. of Health and Welfare and from the American Medical Association. But only a thunderous silence followed that early 1962 news item ion the private meeting in Washington -- until the Aug. 24, 1963 edition of the Saturday evening Post! In this national magazine, radiowave treatment of arthritis was branded medical quackery in the article "Hucksters of Pain". There Drs. Smith, Street, Camern, Lobell and Blackberg had their answer to "government support for expanded research". Fortunately, not every American is going to take this kind of libel without striking back at the mouthpieces of the vested interests who lend themselves so readily to control of the mass mind. Read this news clip from the Los Angeles "Times" for Nov. 1, 1963.

"SATURDAY EVENING POST SUED FOR $7.5 MILLION"

"New York -- The Saturday Evening Post, still smarting from one of the largest libel awards in history, was hit Thursday with another libel suit demanding $7.5 million.

"The suit was filed in Federal District Court here by the Diapulse Corp. of America, manufacturer of a machine that employs high-frequency waves in the treatment of infections and inflammations, including arthritis. The company said it was defamed in an article, "The Hucksters of Pain", which appeared in the Aug. 24 edition of the magazine. The article dealt with quack remedies for arthritis. The suit seeks $2.5 million in compensatory damages and $5 million in exemplary damages."

From the June 1964 Journal of Borderland Research

LAKHOVSKY OSCILLATING CIRCUITS

"About eight years ago I sent to the Health Science Press, Sussex, England for a little book, "Quick and Free Healing", and also "The Waves That Heal" and found that there are healing waves in the air we breathe. I am a sensitive with the pendulum. The book told to take an insulated copper wire and find the positive and negative ends with the pendulum, so I would know which end to put to the left side of my body --positive to the right and the ends overlapping. If you can't keep them in place, wrap an elastic band around them. For me, the pendulum swings clockwise for negative and back and forth for positive. First use of the wire came when I skinned my shin with a five-inch gash. The bleeding made me sick and I had to lie down. I thought of my wire circuit, got up and put it on just above the knee. It always had taken a long time for sores to heal on my body but this healed in three days time and showed only a narrow line about the width of small yarn. This convinced me that the oscillating circuit wasn't foolish.

"I loaned my little books to a high school student that was taking science. I thought they would interest him but he made all manner of fun of me, for believing such a thing. None of my family are interested in the good things that I have learned. I have been a member of BSRA since 1951 and think your journal is the best magazine I have ever read. Anyone wanting to learn more about the Lakhovsky circuits can write to Col. A.E. Powell. (Now deceased)

"Forty years ago I was taking treatments from an osteopathic doctor who sometimes diagnosed with the pendulum. When I became pregnant he told me he was going to discover the sex of my unborn child. He took his gold band ring off, tied a string to it and held it over me. It didn't seem to have any definite swing, maybe around once and back and forth. 'That's strange,' he said, 'I've never missed yet!' When the baby came there were two, a boy and a girl! I didn't ask him then where he learned about the pendulum or what it meant. Years afterward I sent to New York for one and found that I could ask questions and get yes or no answers that satisfied me. A medium told me that the pendulum is spirit controlled and mine was controlled by a young girl who was rather mischievous. If you want to put my experiences in the BSRF Journal it is all right with me. I got my education in the little red schoolhouse. I am 73 years old and still learning. They say you never get too old to learn and I find it is very true." Mrs. Earl Holfinger, Piqua, Ohio

A "HOPPED UP MWO" IS DANGEROUS!

"It is far too early to draw final conclusions regarding my experiences with the higher powered Multi-Wave Oscillator, a word of warning is in order. I powered mine with a 250 VA, 10 KV oil burner transformer. This is DYNAMITE and could, in my opinion, be fatal if too much power is used and I didn't miss it by much. My equipment sparks the antenna down to the 4th ring and the resonator to the second, with occasionally a week spark to the third. 10 hours after the first high-power exposure I felt like running instead of walking (age 72) and walked all around town and back, about six miles and no tiredness, whereas two miles would tire me before. That first exposure was about ten minutes for me and four for the wife (age 75). Immediate reaction for me was a moderately strong sensation of tingling from head to foot for about three hours before I could get back to sleep. None of this for the wife. Then, as aforementioned, I felt on top of the world. 24 hours after that, however, the wife and I developed the grandaddy of all colds and flu symptoms that we'd ever had. I hadn't had any real cold or flu in 15 years. This one dredged such an amount of mucus and toxic waste as I'd never seen before in my life. It was rough. It lasted two days before it wore off.

"This all had me guessing until your brochure arrived, with the experience of your Engineer-Physician Associate. He told that many of his patients developed flu symptoms after MWO exposure. His increasing conservatism was well founded. I decided to find out if the doctor's hypothesis of viral penetration would hold up. If so, a second high-powered treatment of equal strength and duration should produce much the same flu symptoms. But, having wound another coil, I used more power than the first time, sparking antenna and resonator down to the fourth and fifth rings. Results: no more flu, only a very slight tingling sensation, no feeling of exhilaration as before. Five days later I took another exposure, of seven minutes, with a little less power. This one did it. No tingle, no exhilaration, only a feeling of rapidly increasing depression, no sleep that night, a rather severe pain in the chest and heart region, extremely weak pulse, hardly pep enough to life an arm, and a nervous shaking all over. For a while I thought, this is it. It took five days for this to wear off and it isn't all gone yet. One night I lay awake until three a.m. figuring out how to pep up the MWO still more. Out of the blue, a voice clear and distinct and very emphatic: 'This is dangerous!' Again very emphatically, "Yes!" Normally I am not the least psychic, clairvoyant or clairaudient; so I choose to heed this as a warning from a discarnate guide or guardian."

Let this Associate's experience be a warning to you, if you feel the urge to build a hopped-up Oscillator, to transform yourself overnight! Bob Beck's original article cautions against this and his original equipment falls well within the safety tolerances set by the Air Force and the telephone companies for personnel working around radar and micro-wave broadcasting and relay equipment. Most of you taking part in this research program have built your equipment only in the last two or three months. It would be wise to stick closely to the original design for a year before altering it for further research.

LAKHOVSKY'S ORIGINAL THEORIES ARE SOUND!

With one or two exceptions all reports in to date show beyond question that Georges Lakhovsky's theories are sound: Radio waves do effect living organisms. In certain concentrations these waves appear to benefit cell growth, even to stimulate it. This BSR Associates have proven to their own satisfaction in only six months from the publication of our first article. Now it is up to others to exploit this knowledge while we turn our attention to other projects.

*** *** *** ***

From the July-August 1964 Journal of Borderland Research

LAKHOVKSY OSCILLATING CIRCUITS

"The danger of strong reaction to a hopped up MWO, as told in the June Journal, reminds me of my experience with the single circuits. Months ago I wrote to Col. A.E. Powell and he sent me single circuits to wear around waist, neck etc. This I did for two months or so, daily. I'd been under the weather enough to hope for benefit. Instead, I had three successive colds -- something I never, never had in my life before, even during the worst of my catarrh years. It never once dawned on me it could have been the Lakhovsky oscillating circuits; now I wonder?"

A California Associate

Yes, it had been the experience of others that the apparently powerless Oscillating circuit, of only one wire looped around the body, can produce a strong reaction on living tissue. A San Fransisco Associate told us during our May trip north that she received a wire loop from Co. Powell and wore it to bed that night, and forgot about it. Early in the morning she awoke with such a violent pain through her middle she thought she'd die! She lay there and stuck it out until the pain left and went back to sleep. In the morning she awoke refreshed and ready to go, with the oscillating circuit still around her middle. It began to dawn on her that this was the cause of the pain, but it also had affected the healing of a long-suffered condition. At the time we were there, weeks later, the condition had not returned. Certainly no permanent cure of any condition can be expected unless the original cause in mind, emotions or spirit is changed, and if a person insists on indulging his vices, use of oscillating circuits may only cause needless pain!

*** *** *** ***

From the OCTOBER 1964 Journal of Borderland Research

HE PREFERS VITIC TO MWO

"Personally, while feeling the passage of the high tension flow from the antenna of the Multi-Wave Oscillator, I found no beneficial effect. On the other hand, the Associate with whom I obtained the MWO realized definite improvement. He has kept up the use and told me lately that he is feeling steadily better and lays the improvement to the use of the outfit. Some years ago I had frequent conferences (thru a medium) with one who was said to be Madame Blavatsky. During the course of these talks she dismissed the subject of electricity as of little value in comparison with a study of magnetism. That in the study of magnetism lay the secrets of many valuable matters. I am led to refer to this because I recently obtained relief, and apparent cure, from a troublesome irritability of the bladder by the use of a magnet. Other treatments had been of no value. I made no report of this until the fact was definite. The use of the magnet was experimental and came from the suggestions of a Chicago University specialist in the subject, relayed to me via various Biomagnetic reports and some correspondence."

Associate A.N. Onymous, Dixie, U.S.A.

This letter from an elderly Associate was dated Sept. 3, 1964. It certainly provides an interesting and stimulating commentary on the material in the preceding pages of the Journal. Interestingly enough, we've had many reports of successful and continued relief of irritating bladder and prostate conditions from exposure to the MWO. In the Vitic device we now have a borderland gadget that is far less troublesome to build and to operate. How effective it is remains to be proven.

THOSE LAKHOVSKY CIRCUITS AT WORK

"About the Lakhovsky wire circuits, there is no other way to account for the consistently maintained energy I receive. I wear one all the time as do several others I know, all of whom report amazing results. One man past 80 finds it peps him up too much, so he wears it intermittently."

Howard D. Clark, Yucca Valley, Calif.

*** *** *** ***

From Jan-Feb 1965 Journal of Borderland Research

HE WORKED WITH GEORGES LAKHOVSKY

"A friend of mine sent me your pamphlet describing the Lakhovsky Multiple Wave Oscillator. I was the first man that bought this instrument in New York in 1940, on the advice of a physician who saw it at a medical congress in Vienna. My young son had developed a face lesion. It was diagnosed as tubercular. I was having a joint office at that time with a progressive MD. I was practicing dentistry. Her success with the MWO was phenomenal, but when she began to treat cancer cases the American Medical Association got after her and me and revoked our licenses. I had translated some of Lakhovsky's works into English from the French. Professor Lakhovsky came often to my New York office. He told me the Pope was using one of his instruments. The success obtained in the hospitals was excellent but the reports were suppressed. The good professor died of a broken heart. Our case was a regular frameup by the AMA, the lawyers and the courts. I was foolish enough to fight the case and spend lots of money, not knowing that the AMA controls even the courts and judges. I still have some copies of the booklet and a photo of my Lakhovsky instrument. If I can locate them I'll send them to you. I sold my instrument to a doctor in Texas and left New York state in 1945. I am a licensed Naturopath in Florida and Chiropractor in California. Am enclosing $ for membership. I belonged to BSRA several years ago."
Dr. N.S. Hanoka, Miami Beach, Florida

*** *** *** ***

From March 1965 Journal Borderland Research

"THE SECRET OF LIFE"

"I loaned my copy of Lakhovsky's 'The Secret of Life' to my chiropractor, also your Journal. He had one build and I've been taking treatments from him on the Multi-Wave Oscillator. After two treatments I got an awful cold and I have never blown so much bloody mucous out of my nose. Also, after each treatment, it would leave me weak. I would have to lie down several times a day. There always was a reaction but my bad sinus condition has cleared up considerable. I also had trouble with my liver and there were so many things that I hadn't been able to eat, for 15 years or more. Now I can eat them and enjoy them. The MWO is certainly worth having and I am improving."
Mrs. A.N. Onymous, Somewhere, Indiana.

*** *** *** ***

From the October 1966 Journal of Borderland Research

HITTING THE HEALING TRAIL WITH THE MWO

"It would be appreciated if you would put an announcement in the Journal that Dr. S.L. Jamison, D.V.M., N.D. (Gr.Br.) announces the opening of a practice of naturopathy in Santa Cruz, California.

"The climate has changed greatly from the old burning-of-books days by the FDA. We cannot all stand back and wait for "George" to do something. I have the advantage of a lot of training in a legitimate field of allopathic medicine. Only a veterinarian would be able to make such a move and escape retribution for it. A regular M.D. would be branded a traitor to his fellows.

"I have moved to Santa Cruz to set up practice as a naturopath, using the M.W.O. unit, spiritual healing and diet -- along with naturopathy measures for the healing of old injuries and chronic degenerative diseases. There will be no charge if there are no results. Since my fees will have no legal status, this will be made clear to the people using the service so they will be under no compulsion to pay, even if results are satisfactory!

"There are no incurable diseases but there are incurable patients!

"With the MWO we have not gotten uniform results classified by the disease the patient had. About two out of ten people do not respond to the MWO unit at all. You had just as well breathe on them. I strongly suspect that these people should consult an allopathic physician, M.D., as they would be helped better by conventional medicine. You might say our cases were selected, as people came to us in the last stages of sheer desperation -- conventional methods (including chiropractic) having failed.

"Some of these (not many) also do not respond to the MWO. They usually respond to spiritual healing methods when this happens. This I have not figured out good as yet. Once in awhile an arthritic will claim to have been worsened by the unit. This occurs usually in an arthritic with arthritis in a number of joints. One treatment seems to break the poison loose in their body, and they fall like the wrath of hell about two days after. Get them in immediately and run them again, and everything is alright.

"To sum it all up, I have got so I do not care what disease a person has got as long as it is not an acute infectious disease, or a recent stroke case. I have been afraid to try it on either of these as all we would need is for somebody to fall over dead while on the unit and we would be wearing striped suits or breathing HCN in a gasper chamber. . . The unit regenerates the whole body." S.L. Jamison, D.V.M.

*** *** *** ***

From the July-Aug 1967 Journal of Borderland Research

A TRANSISTORIZED MWO

"I would like to compare operational notes with some one who has built one of these MWOs because I question some of the statements made in the original article. I'll just give a brief description of mine and some of the things I ended up doing. I started with aluminum elements but had trouble with the gluing process so ended up with using printed circuit boards so my elements (antennas) are copper. I have good arcing between first four rings. Can make 'em all arc if I bring point of wooden pencil near the smallest, but pencil must be held certain way or I'll get knocked on my haunches.

"I ended up with a 9 in. Tesla coil. Tried 7, 6 and 5 but ended up with nine. The trouble here was with leakage between primary and secondary. It was terrific until I took means to reduce to a minimum. The transistorized drive I use is partially a hangover from the days of when I used transistorized ignition in my one vehicle. The ignition coil was modified to increase the step-up ratio and since the coil was mounted originally in oil, it was a 'slippery as an eel' project. The driver uses 5 transistors plus diode and operates on a source power supply which can supply up to 20 volts d.c. The drive seems to function best at about 15 - 16 volts.

"Now for a question. If Mr. Beck, as he indicated in his original article, 'reconstructed the circuits' of an original MWO in a certain man's house, why didn't you people publish a diagram and dimensional data on that original piece of equipment? That would be a project a man could sink his teeth in. Instead this pee-wee MWO which you now say is not much good unless everything arcs. You mean arcing all the way down to the smallest ring?????? When you get your up-dated MWO data published, let me know as I would like a copy. R.A. Rieck, Rochester, Minn.

Arcing between the first four rings, from the outside in on the MWO antenna, is about as good as any we saw when we were researching the instrument in 1964; so it sounds as though your transistorized MWO -- the first we've heard of --is doing very well and should be putting out enough energy to affect living cells. You any wish you had cut aluminum antenna rings after all. For some reason, the arcing between copper rings builds up deposits which must be cleaned off periodically. Bob Beck didn't offer us a schematic of the original Lakhovsky equipment he saw. He said it was housed in a 300 lb. cabinet. The antennas were of circular-formed pipe rings. The outside ring was two feet in diameter. There was a large, separate Tesla Coil driving each antenna. This is a professional model designed to impress the patient. It also probably puts out enough static to antagonize the neighbors for blocks around. We believe Bob Beck deserves all credit for grasping the principle of multiple wave propagation developed by Lakhovsky during radio's infancy in the 1920s and redesigning it into a simpler, smaller device for present-day borderland research.

LATER COMMENT ON THE MWO BY MR. BECK

"The key factor is that the rings MUST arc profusely all around the circumference of the outside and THIRD ring simultaneously. There is no effect, either on a calibrated Hewlett-Packard field strength meter, or physiologically, at the target frequencies if this arcing does not shock-excite the rings. And yet I have seen at least three machines that did not have sufficient voltage from the Tesla Coil to even drive the outside ring, let alone the third element! It would be a shame to invalidate the device because of this. Some of the machines I've seen drive the Model T Coil from an A.C. transformer! The obvious fallacy here is that the coil's vibrator is a resonant device supplying the make-and-break voltages (and magnetic field) to the T coil primary at the designed frequency. Since the transformer is on and off 120 times a second, the chances of the T coil points 'breaking' at the precise instant that the 60-cycle house current is at a POSITIVE peak is about one in fifty. And even if it fires 'sometimes', the T coil is modulated with the 120 cycle line (twice the 60 cycle frequency, since each cycle has two peaks, both a positive and a negative).

"Try it at 300 mc. and 10 kmc. on a field-strength meter some time. You'll see what I mean. The Tesla Coil, incidentally, is just a convenient supply of free electrons. The cascade of electrons or 'high voltage' breaks down the air's resistance (by arcing) and then drives the antenna rings. But if the drive is interrupted, the rings may not absorb and resonate properly. At any rate, it is difficult to kid the Hewlett-Packard meter, and I'd like to see the original design followed, at least for the first year."

AN IMPORTANT MODIFICATION

"The greatest modification to date was the discovery of a box of war surplus very high voltage rectifiers in a bin at a Burbank electronics store. The 30,000 to 60,000 volt selenium stack rectifiers.....eliminates the necessity for a tuneable gap, and roughly quadruples the output of the Tesla Coil, also lowers battery drain to half."

BEWARE THE PUNITIVE HAND OF THE AMA

Turlock, Calif. "Daily Journal", Dec. 2, 1966: "Dr. Stanton L. Jamison, formerly in charge of the state Poultry Pathological Lab here, was arrested yesterday for practicing medicine without a license. The veterinarian is charged with using an electronic device -- with high voltage sparks crackling and jumping about on either side of a patient --with the claim it was beneficial in cancer and bone diseases.

"After posting $1100 bail, Dr. Jamison defended his machine. He said parts for the machine cost less than $50. If someone followed up on the apparatus, he claimed, it would be a 'blessing to humanity'. Police Chief John Viarengo, Lt. Bill Ladd and investigators for the California Bureau of Food and Drug Inspection took Dr. Jamison into custody in the offices of his Life Aquarian Center in the Mercantile Building. Lt. Ladd said a sign reading 'Universal Life Church, Inc.' was prominently displayed in the front office. He said four or five agents had been 'ordained' as ministers in the church by telling Dr. Jamison they wanted to become preachers. Soon their official ordination certificates arrived in the mail, said Lt. Ladd.

"Police described the apparatus as similar to an electric chair, with four-inch metal rings looking like handcuffs hanging over doors to the room. The theory was to determine polarity of a patient with a pendulum and then proceed to change negative polarity to positive with the 70,000 volt device. Chief Viarengo said patients were being treated for a variety of diseases, including stomach ulcers and cancer. Investigation into Dr. Jamison's operations started in May, when police began receiving complaints from relatives of patients. No money was charged for treatment, according to Lt. Ladd, but patients were expected to make donations to the church. Dr. Jamison headed the state poultry lab here for a decade. He will be arraigned on the charges in Turlock Judicial Court next Friday morning.

"In a statement released following his arrest, Dr. Jamison said the 'multiple wave oscillator' picked up by police resembles many electronic gadgets used, usually ineffectively, in the past for treatment of many complaints. However, he insisted, his oscillator appears to be the exception that proves the rule.

"'The unit suffers from a sort of 'guilt by association' with other electronic gadgets to the point where serious doubt would be the reaction of any trained scientist on first observing it,' he admitted. 'I know that this was my reaction.' However, he said he suffered continuous pain following a 1962 automobile accident and, willing to try anything, tried the unit. It helped him, he insisted, and has helped many others. He said the unit had been tested and checked by qualified medical men in Europe in the late 1930s. 'Some progressive healer will follow up my lead in continuing to develop the electronic unit,' he said."

*** *** *** ***

From the November 1968 Journal of Borderland Research

AN MWO PHENOMENON

"Please find enclosed $ for another year's subscription to your wonderful Journal. It has been very enlightening, and I always look forward to receiving it. I have taken treatments on the MWO for a long time and have been puzzled. Soon after I started to take treatments, the hair stopped growing under my arms. I have wondered about this and if anyone else has had the same reactions?"

Mrs. E.H. Piqua, Ohio

This is new to us! If any of you researchers have experienced or observed a similar phenomenon, we'd be glad to hear about it and share it with the rest through the pages of the Journal. The Tesla Coil problem has been solved with the discovery of a good commercial coil made by Master Appliances, Marion Indiana. This coil will drive one antenna very well.

*** *** *** ***

From the Jan-Feb 1971 Journal of Borderland Research

THE PASSING OF CYRIL SCOTT

"Cyril Scott, born Sept. 27. 1879 in England, died first week of 1971, so two friends here inform me. I'm always happy to see one who writes so intelligently regarding diet and health, vindicates his ideas by living that long. Not many do! BUT, I read in our local paper yesterday of the death of a former neighbor who died on 11th January at 93, and she knew nothing regarding diet, positive thinking, etc. So, where are we? I can't ever make 2 & 2 come to four, woe is me.

"Two small books come from Muriel Archdale, England, by Cyril Scott, 'Safe and Saner Remedies', and 'Health, Diet and Common Sense'. On page 162 of the latter he mentions the George Lakhovsky book, 'The Secret of Life', memory of which caused Mr. Scott to suggest using a copper wire around his middle to a man who was too breathless to walk far or uphill. The man wore the Lakhovsky loop for twelve hours, had a copious and offensive bowel movement, then on following day and another uphill walk, the man was able to keep up with Scott! I found my copper flex soon after but it is awkward to keep in place to make sure the negative end is to the left. Cyril Scott gives the homeopathic remedies needed, particularly emphasizes potassium for cancer."

A.F., Santa Cruz, California

COPPER BRACELET CURES "TENNIS ELBOW"

News Item from Crosby Golf Tournament at Pebble Beach, California a year ago. 1/24/70: "Bert Yancey one had to give up golf because of a nervous breakdown. He also came close to quitting the professional tour last year because of a 'tennis elbow'. But he didn't mainly because he found a remedy endorsed by Australian tennis pros John Newcombe and Tony Roche. The so-called remedy is a $7.50 copper bracelet, otherwise known as 'The Absorber' or a 'Voodoo Bracelet".

"Yancey wears one on his right wrist to correct an arthritic condition in his elbow and he has worn it for the past five tournaments, including the $150,000 Bing Crosby National Pro-Am, which he leads by two strokes at 137 going into today's third round.

"'I don't know why it works, but it does,' said Yancey, a 31-year old former West Point cadet who finished third in the 1967 and 1968 Masters and won the Atlanta Classic last year. He said a tennis pro in Napa Valley suggested last summer that he wear the bracelet and got confirmation from Newcombe and Roche, who have suffered from similar ailments, that it would work.

"Yancey strained his right elbow two summers ago lifting a concrete birdbath in the backyard of his home in Tallahassee, Fla. He took a pain-killing drug that enabled him to compete in the U.S. Open at Rochester (where he finished third), but he was warned that the drug w 'dangerous...that it gives you ulcers'. 'I never got ulcers, but I felt a pull in my stomach. That's when I quit using the drug.'

"He said he hasn't needed the pills anyway, since wearing the bracelet. The therapeutic effect of the copper, he said, has to do with chemical displacement within the body. 'I majored in chemistry while taking pre-med courses in college and I think I know enough about copper deficiency in the body to believe that, somehow, the bracelet can help me.'

"Somebody then pressed Yancey for an answer about what makes the bracelet such an effective remedy. 'I don't know for sure,' he drawled, 'maybe it's all psychological.'"

YANCEY WINS THE CROSBY OPEN

Sports writers are always looking for a new angle for stories and Bert Yancey withstood a spectacular charge by Jack Nicklaus on the final round, with a 65, and came in with a hot round of his own a 69, to win the tournament and a cash prize of $25,000. Yancey and his Voodoo Bracelet were headlined all over the country. The LA "Times" called it a "Copper Caper" in its Jan. 26, 1970 story.

As was to be expected, the medical authorities in desperation had to prepare an immediate counter blast of propaganda to keep the public hypnotized. The sales of medical snakeoil went down as arthritis sufferers flocked to buy non-medical copper bracelets! The "Times" and other metropolitan newspapers willingly cooperated. After all, patent medicine advertising is a big source of revenue, and who knows how much profitable pharmaceutical stock is owned by the publishers of the papers?

The "Times medical writer, Harry Nelson, dutifully wrote: "The Arthritis Foundation, a legitimate group interested in getting better treatment for arthritis sufferers (There is a contradiction here, Harry, for medical authorities claim there is no cure for arthritis!) refers to copper bracelets as 'an age-old fraud which has no scientific basis whatsoever'. Yancey is the latest sports figure to get taken by the 'copper caper', which the attorney general of New York last summer called 'one of the oldest and most vicious swindles'."

These opinions by the Arthritic Foundation and the attorney general are of course; utterly worthless. They are not based on the results of intelligent scientific research, but are based on the profit motive of the medical trust. The electrical nature of living cells was established beyond any shadow oaf doubt two generations ago, by the sound theories and research of Dr. Abrams in San Fransisco and Georges Lakhovsky in Paris. Thousands of supporting tests since then have proven conclusively that body cells have polarity, and that in an unbalanced or unpolarized condition, disease or injury, the subtle oscillations of a wire coil or loop will restore polarity. Sad to say, there is nothing of this in the training of medical students today and it is left to Borderland science to carry the torch. The honest, straightforward testimony of public figures like Bert Yancey, Tony Roche and John Newcombe is ridiculed by the witch doctors of authority for medicine.

JUST KEEP TAKING YOUR EXPENSIVE PILLS!

Under orders from above Harry Nelson interviewed Dr. John Calabro, the arthritis authority at University of California in Los Angeles, "He points out that the bracelets may seem to work because arthritis can improve spontaneously and it is easy for the person to associate his improvement with the bracelet.

"'I tell patients it's OK to wear the bracelet providing they do all the other things I ask them to do,' said Dr. Calabro." Including, of course, the purchase of poisonous pain killers at fancy prices. How much scientific integrity does Dr. Calabro have when he and his university accepts fat research contracts from the medical trust? Is it any wonder that today's Aquarian Age students are publicly questioning the morals and ethics of their professors? They are told that the goal of a university is the search for truth -- at least that is what the Chancellors say in their public speeches. But when the student gets to the laboratory he finds that the search is for profitable products for industry -- or help maintain a well-established profit position.

LAKHOVSKY BELTS, LOOPS or OPEN CIRCUITS

"Thank you for the clarification of the Lakhovsky belts described by the late Col. Powell in his publication. I tried what you suggested and after some thought I finally understood what he meant and I believe the belts should look like this:

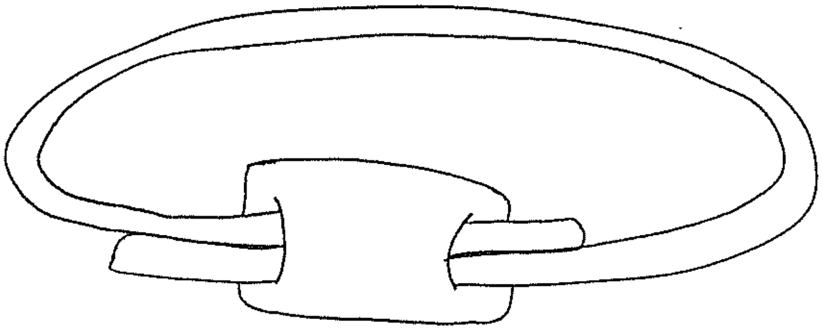

You might want to try experimenting with flat ribbon cable for these belts or coils. It can be bought at most electronic stores and it comes in different widths or colors. I am wondering also, is it really necessary to separate the lamp cord and use only half of it? Could it not be used as it is, with two strands of copper wire, insulated, side by side?" G.B.F., San Fernando, California

Yes, in our one-loop coils of white insulated lamp cord, we now use it just as it comes from the store, without separating the two strands. But you plastic 'buckle' or 'button' shown above may allow too much slippage; so the ones we make for ourselves and for research-minded Associates have separate entrance and exit holes for each end of the insulated wire. The white insulated wire seems preferable from the hygiene angle, seeing that it is usually worn next to the skin.

*** *** *** ***

From the March-April 1971 Journal of Borderland Research

OUR LOOP-THE LOOP PROJECT

"It looks to me like the Lakhovsky loops work homeopathically. It's not the copper element, per se, but the loop the loops it makes, throwing back into the body forces which the body is throwing off. It's an automatic homeopathic prescription, each body being its own pharmacist!

"The Yogi does it when he goes into suspended animation, re-breathing his own air. Urine therapy, drinking one's own urine, as described by Armstrong, is homeopathic. An ostrich putting his head in the sand breathes his own air. Breathing into a paper bag and re-breathing the contents is a specific cure for old age pains and hiccoughs. Dianetics was mental homeopathy, i.e., the original book but not no more. Dianetics is re-thinking one's own thoughts. You imbibe your own aberrations, attaching like with like. Homeopathy is like a feedback system. Gods ate their own offspring. It's Isis picking up the pieces. The smaller the dose the more potent the cure. Etc. Etc.

"Send me two kits of three experimental loops. Maybe I can fatten up on them."
P.F., Florence, Arizona

HE FEELS LIKE A NEW PERSON

"I read that part in the Journal about the copper wire loop; so today I did loop one around me and sure enough the oscillating motion calmed me down. It sure was a sensational feeling because lately at times it got so bad I though I would spin off, even though I do the five Rites nightly. On occasion I also experience a terrific pain in my left side which forces me to sit down, just a great pressure especially after standing over the stove preparing meals for maybe three to four hours at a time. The customers just couldn't understand how I stood up under pressure of serving so many people. Seems like mass production.

"So with this new item I feel like a new person! And I'm going to have you send me one of your experimental kits, to see how it compares to what I hooked up on myself. Of course I just taped the wire on and this is only about six hours since I put it on." J.M.S., Charleston, South Carolina

A DEDICATED BORDERLANDER

"Enclosed is a check for membership renewal and for five sets of experimental loops. I want to compare the effects of the Lakhovsky coil with the Japanese-made magnetic health bands which I sell to my tennis customers. I am very active in tennis playing and teaching. The results with this band are very good."

A.M., Miami, Florida

BEYOND THE PHYSICAL

An Associate writes that she had success in "curing" and arthritic elbow by wearing a copper bracelet on her wrist. Then the pesky condition showed up in her knee! From there is moved to another part of her body. Then it began to dawn on her that their Higher Self was trying to tell her something, that the swollen and painful joints were an outward manifestation of an inner mento-emotional condition of resentment and bitterness over an unhappy marriage. With the cooperation of an understanding doctor and more orthodox heat treatments, etc., and a more positive and cheerful attitude on her part, the condition was finally eliminated. It took months to cleanse the system of the accumulated poisons, physical, emotional and mental.

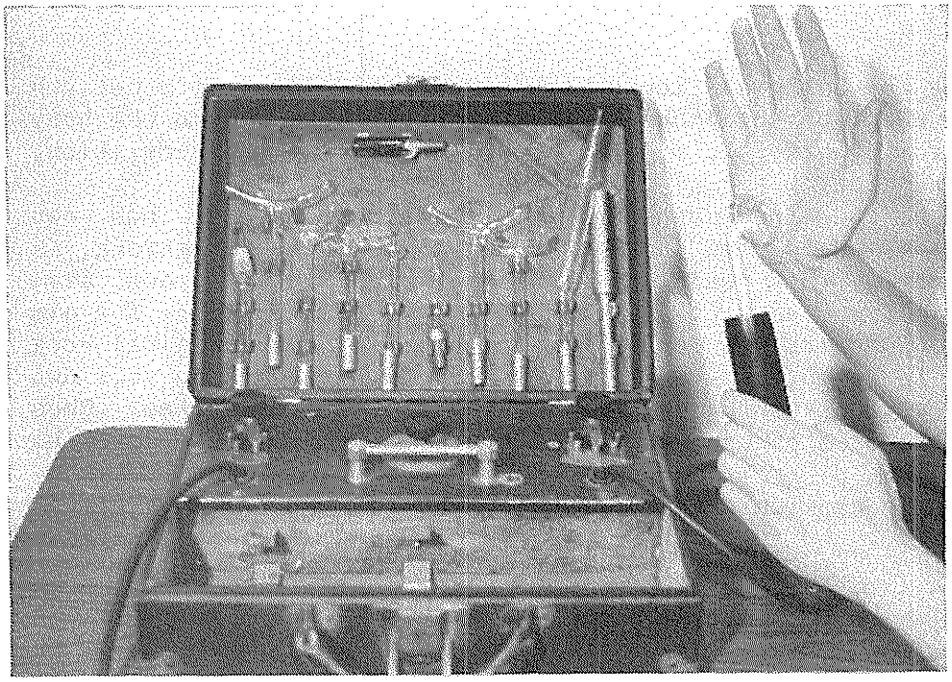

1920's era Violet Ray Massager. From the BSRF Collection.

From the Jan-Feb 1966 Journal of Borderland Research

THE LAKHOVSKY MWO IN INDIA

"The MWO is giving wonderful service and days are passing and I am gaining some new experiences which have not been recorded in the book. I will appreciate if you kindly clear some of my doubts. In practically all the cases I am finding aggravation after three or four sittings. Do you think that it is due to a sudden oscillatory shock that is given to the patient as it so happens when Homeopathic medicine is given? I feel that it should have at least been recorded by some one. One more question. What is the reason for metals being removed from the person of the patient? In case they are not removed will there be any adverse effect? This is very vital because here in Bengal there is a custom that the married ladies must wear one Iron bangle. They will not remove it under any circumstances. What is the remedy for this?

"I find that the appetite improves in practically all cases. Rheumatism and Arthritis cases are definitely benefitted. In one case a Tumour which is there since about 20 years has definitely gone down with only 4 sittings. I am watching the case carefully and if the Tumour disappears it will be a great achievement for the MWO. Debility is markedly reduced in nearly all cases. In a recent case of Spondylitis the MWO has done wonderful work. The flexibility was 8-1/2" from the floor which has come to 6" after 7 treatments and all the pains have practically disappeared. Due to the New Moon and Eclipse there is an aggravation but I hope it will gradually go. Asthma cases have not responded so well but I am not yet disappointed with the results. It is as yet too early for me to assess the capabilities of the MWO on various diseases. As you say I also believe that this instrument will work on all persons irrespective of their complaints but still there must be some specific organs on which there is more influence of this.

"Something which I forgot to mention previously, can you tell me whether you have seen Hemorrhage as a result of MWO treatment? In at least 3 cases I have seen bleeding start where there was no previous history of bleeding. What is it due to? In one case there was profuse bleeding after 2 sittings and that has made me a bit nervous. Just now a patient reported that he was getting blood in the sputum even though there is nothing in the lungs. In the third case blood appeared in the urine. Please give a thought to this as this is a very serious complication if it is caused by the MWO." Dr. A.K. Bhattacharya, West Bengal, India.

The above is a portion of a letter to Associate Al Goeke, Redmond, Washington, which Al passed on to me for comment. I thought the rest of you would be pleased to know that this East Indian doctor has gotten very encouraging results from the very beginning in the use of the Multi-Wave Oscillator. If by aggravation the doctor means a sudden release of mucous, heavy cold or flu, this is a sure indication that the MWO is working and affecting a release of accumulated poisons in the body. This is touched on in our brochure, in the comment from a midwestern doctor which built and used such equipment in the 1940's. When he began producing mild flue -- as he called it -- in his patients, he got scared off! The hemorrhaging by the doctor's East Indian patients is more serious but we believe it is still a result of the same thing, a sudden releasing of accumulated poisons through the shedding of bad blood. We know one oriental who had a heavy nose bleed after the first exposure to the MWO field. Apparently orientals, some of them at least, are more sensitive to the soft radio wavers than Caucasians living in America; so the doctor's prescribed treatments should probably be somewhat shorter in duration and spaced further apart.

Yet I well remember the radio technicians in Honolulu, Hawaii. I was an announcer and producer there for several years. Some of them were of Chinese and Japanese ancestry. When repairs were needed on the radio towers, these men wouldn't hesitate to jump on, climb up and work for hours while the station was putting out a steady five to ten thousand watts of power. I don't recall any of them suffering any ill effects such as internal bleeding. Yet their tools would get so hot because of the strong field that they would have to wear gloves in handling them! It was these eddy currents caused by an inert piece of metal in the pulsing field that Bob Beck was concerned about. There isn't enough power in the MWO to cause heating in bits of metal about the body, if one can't strip down for an exposure. The piece of metal, say a zipper or a ring, might reduce the efficiency of the field in that area but we didn't worry about this when testing the equipment we had for a short while in 1964. I don't think Dr. Bhattacharya should either. The machine seems to produce results anyhow.

*** *** *** ***

From the March 1966 Journal of Borderland Research

BASIC NEW AGE PHYSICS

by Trevor Constable

(Publication of Dr. Bhattacharya's letter in the Journal, page 20, Jan-Feb 1966 issue, elicited the following significant comments by Mr. Constable.)

A word on Dr A.K. Bhattacharya's reports concerning the Lakhovsky Multi-Wave Oscillator he is using successfully at his West Bengal, India clinic. The results are indistinguishable from those systematically recorded and published over a long period by the late Dr. Wilhelm Reich, M.D., concerning his "orgone therapy". Reich was the discoverer of Primary Energy -- pre-material and mass-free -- which he named Orgone.

The improvement of the appetite, the reversal of rheumatism and arthritis even to structural changes, the rapid reduction of tumors and the hemorrhage effects are all to be found in Reich's massive clinical literature on orgone therapy. Dr Bhattacharya's reports, plus others any my own experiences with the MWO serve to confirm my previous evaluation of this unit as published in the Journal over two years ago. In essence, the MWO is a device for orgone therapy. It achieves its startling results because it is a generator of primary energy, or at least a focus for primary energy.

Any person with a modest extension of physical vision who will observe an operating MWO under conditions of indirect, rather dim fluorescent lighting, will see for himself that primary energy from the earth's atmosphere (the Orgone of Reich) condenses prolifically between the coils of the unit. The concentration is many times that which normally surrounds a living organism.

The basic New Age physics involved, for which we are indebted to Reich, are as follows: Wherever secondary energy, i.e. energy obtained from matter in some way, is released, primary energy tends to concentrate to extinguish this energy. Secondary energy is life-negative, i.e. inimical to the living, a fact which is illustrated, in its ultimate extension, in the lethality of atomic energy. Where so-called "soft" radio waves are involved, the antagonism of primary energy to their emission is evidently considerably diminished over the violent effects that accompany radioactive substances in the presence of primary energy accumulators.

In the case of the MWO, the tremendous band of frequencies covered results in an extreme concentration of primary energy, which also has its frequencies and correspondences (the Rates of radionics) with the secondary, electromagnetic spectrum. That this effect is achieved with minimal secondary emission is the beauty of the Lakhovsky unit.

TWO-FOLD FUNCTION OF THE MWO

In my view, the dramatic results obtained are due to a two-fold function of the unit. First, there is the undeniable concentration of primary energy around the body of the patient, particularly potent where a limb or other readily accessible member may be exposed. Secondly, there is the reaction of the specific biological energy of the individual organism, from within, to the stimulus of the secondary waves, which it rushes to combat.

Under the laws of primary energy first discovered by Reich, the patient, as a living organism, is a stronger vital system than the MWO. Accordingly the patient attracts and withdraws the charge of primary energy produced by the device. This is in direct contradiction to formal electrical laws, and as such is typical of the criteria that rule primary energy. The continued emissions of the Lakhovsky unit provide a continuous concentration of primary energy on which the patient draws. Just how long this process should be allowed to continue, per treatment, is where we run afoul of allopathic (MD) thinking.

The acquisition of a strong primary energy charge by the ailing organism results in a reversal, or at least in the reduction of life-negative, i.e. diseased, processes. Reich found with his orgone accumulators that the reduction of tumors was not as great a problem as the auto-infection resulting from the overloading of the excretory apparatus. The debris and bacilli from the diminishing tumors must be eliminated, and unfortunately, because the excretory apparatus is frequently involved in the general putrefaction of which the tumor is the end result, this was the major problem in managing such cases, in Reich's experience.

The medical lectures of Rudolf Steiner, Ph.D., contain one of the master keys to successful therapy, particularly appropriate to the use of the MWO. Steiner points out that what the ailing organism needs is the significant, vital impulse in the right direction, then careful nursing and management as it recovers out of its own forces. This was also the view of another genius, the late Dr Ruth Drown, who knew from experience how important the first impulse back to health is for the sick individual.

Hence, it seems that the old allopathic urge to convulse the patient with therapy, to repeat the massive dose again and again, is something that should be resisted in using the MWO, at least where tumors are being dealt with. The idea of applying the MWO in daily doses may stand in need of modification. We are all to some degree creatures of gross, non-vital thinking that stands behind our allopathic medical civilization, and we will need to carefully train and think ourselves out of the absurdity that the massive dose is the sine qua non of therapeutic success.

THE TRUE-BLUE ROOT CAUSE OF DIS-EASE

As far as unexpected hemorrhaging from the lungs and similar manifestations are concerned, I think the problem is essentially that diagnosis is so inept and inaccurate that the physician does not have a true blueprint of his patient. After my years of association with Dr Drown I am sharply aware of the fact that nobody just has one ailment or problem, but that we are all bearers of a veritable plexus of malfunctions which usually proceed from one primary and obscure root cause. What happens with the MWO is that the injection of primary energy into an ailing organism results in multiple manifestations of the healing impulse, and not just a direct effect upon the object of our attention --which may be a tumor, ulcer or other extreme symptom.

All the effects produced by exposure to MWO units are identical to those experienced in orgone accumulator. The healing agency in both cases is primary energy, and I hold that this is so no matter what plausible arguments might be advanced to reconcile the MWO with formal theories and thus win the neurotically-desired approval of physicists and doctors. It perhaps seems incredible that one can obtain the same effects produced by the MWO by sitting in an orgone accumulator, with absolutely no electronic apparatus connected to is, but it happens to be so. Try it and see!

MWO experimenters owe it to themselves to study Reich's work in full. As long as they wander in these fields, they may as well build into themselves the realization that electromagnetic energy has an inevitable and unavoidable and inescapable concomitant -- primary energy, life-positive, mass-free energy whose criteria and laws are the exact opposite of the energy accepted by formal physics. The best testimony to Reich's tremendous contribution to human advancement is that he died in a U.S. Federal Prison. As all Borderlanders know, innocent bumblers die in bed in their own homes, untroubled by the status quo.

*** *** *** ***

From the April 1966 Journal of Borderland Research

MWO SURGERY AT SHASTRI VILLAGE AND RELATED MATTERS

by Al Goeke

I believe you and the Associates would like to hear the latest comments from Dr. A.K. Bhattacharya, Naihati, West Bengal, India on the Beck adaptation of the Lakhovsky Multi-Wave Oscillator he is using in his clinic there.

"The other day I had an unique experience with the MWO. I think this should be published. I was treating a lady for rheumatism. After four sittings, once a week, she had a pain under the tongue and there was some inflammation with pus coming out. I asked her to take a further sitting. After about five minutes all of a sudden the thing burst inside the mouth and five pieces of stone came out. When I took her medical history I found that she had an operation two years before and a bit of Salivary Calculi was extracted, but the surgeon had told her that some still remained behind. The MWO treatment eliminated the foreign matter which would not have come out without an operation. This then paves the way for research with Gall Stones and Kidney Stones. I would like to know if anyone else had an experience like mine.

"This time I can give you some reports about the special attachment you sent. A patient was having much pain in the knee and there was swelling. I held the Bulb against the knee for five minutes, and after that it was found that the swelling had reduced and the pain also was much less. In an asthma case I applied the Bulb on the chest for ten minutes. The report is the patient had no recurring attack for two days. Yesterday I was having pain in my thigh so I applied the Bulb. In five minutes I got wonderful relief. Gradually I am getting the hang of it and I think I will be using it more and more."

I, too, am using the Argon more and more, and the more I experiment with it, the greater I think it is, also the MWO. I have had two test cases of young men who, after five minutes in the MWO, said they could feel something happening inside their heads. Both cases had a history of head injuries. I take it that the MWO reactivates a dormant, abnormal condition that perhaps one day would cause trouble. Another test case was a mental patient in Texas. The use of the MWO was part of the test program which brought about a marked improvement in only nine days.

Some time ago a BSRA associate spent an evening with me and told of an interesting borderland gadget. People came from miles around to sit in its healing radiations. I built one for experimentation and found the results good. A Texas Associate says that it works! The original was set in the ground, but I wanted it in the house where it would be more convenient, and figured on running a ground wire outside; but this is not necessary. I have found that it corrects polarity of anyone sitting in the same room. The one I built radiates out 25 ft. One length of 6" stovepipe is used, filled with alternate 2" layers of crushed granite, and of charcoal. For a quick charge of energy, touch your foot on wire at lower end and hand on wire or rod sticking out of the upper end. I have checked the walls and ceilings or rooms (with pendulum) and found them negative. After a few minutes with this borderland gadget in a room, the walls and ceilings check positive! Here is something that should be shared with Associates.

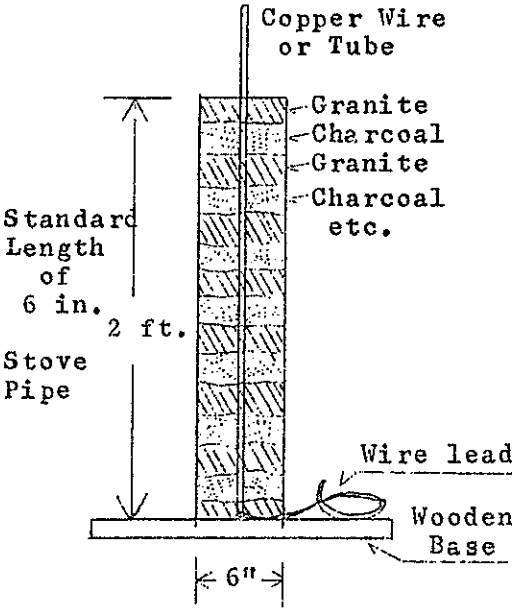

ETHERIC CONDENSER

(Here is our drawing from the rough sketch in Goeke's letter, of the borderland gadget explained above. This is just another of many simple devices which concentrate or condense etheric energies of the higher physical sublevels into a field or vortex of force available or useful in beneficial amounts to a living organism within that field. The drawing is not to scale. The dimensions are not critical. It is the relation of the parts that is important. You can consider the upper end of the copper wire as positive and the lower end coming out of the base as negative, but with the alternating layers of crushed granite and charcoal there must be an alternating or pulsating effect. Remember, charcoal is carbon, so this Etheric Condenser is a Vitic device, but not nearly so concentrated as the field between the Alnico magnets of Vitic. Iron filings sifted in the crushed granite would probably increase the power of this thing. This gadget would augment the power of your Eeman Screens, if you are using them, by adding extra etheric vitality to the circuit. Add long wire leads to the upper and lower ends of the Condenser, so you can hold one in each hand while holding the handles of the Eeman Screen leads. Which Condenser lead should go to each hand? I dunno! This is borderland research and you'll have to experiment on yourself, and be you own authority! This is one of the first principles of the Aquarian Age into which we are moving at breakneck speed. Now to continue Al Goeke's article. RHC)

GERMICIDAL EFFECT OF ULTRA-VIOLET LIGHT

The use of the Argon bulb (General Electric AR-1) as a pain-killer and infection reducer is not a new idea. Nikola Tesla used a gas-filled tube or bulb, excited by high-frequency current to rejuvenate himself every day in his laboratory. 40 years ago the RenuLife Electric Co. of Detroit was making and selling such a treating device to doctors all over the country. The offered a choice selection of vacuum tubes for getting into the body to release rejuvenating, purifying ultra-violet rays. I don't know that their tubes were filled with Argon gas, as is the current General Electric bulb. Maybe this is why the GE bulb is so effective. There is a greater concentration of ultra-violet light in the most beneficial wave-lengths, from pure, excited Argon gas.

We visited one chiropractor on a trip last year who used an MWO machine on his patients, and also uses the Tesla Coil of the MWO to drive the Argon bulb. General Electric designed this little gas-filled bulb for use on the regular 115 volt house current. As a very dimly burning, ultra-violet night light it is a good spook-chaser and can be left on all the time because it uses only pennies worth of electricity in a month! The little bulb has a curved metal plate inside it, as shown here, on which you can see a faint, ultra-violet luminescence when it is turned on, in a standard light socket.

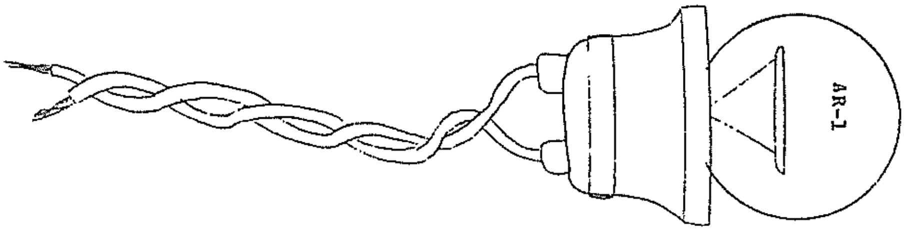

BUT, when you put a hundred thousand volts through it from the MWO Tesla Coil, that Argon gas gets real excited and throws off powerful ultra-violet energy which can penetrate into the body and destroy infection and break up tension, inflammation. Showers of sparks jump from the surface of the clear glass to the skin unless the bulb is held so it touches. Sparks will also jump into your hand unless the socket mounting is well insulated. The mounting or socket shown here is the standard rubber base with heavy, six-inch leads for outdoor lighting. Our local hardware store sells the socket for 39 cents. The Argon bulb needs only one connection -- to the antenna lead of your MWO. The black and white leads from the socket can be twisted together and soldered to a female Banana Plug. This will take the male connection from the Antenna lead. The twisted socket leads can be wrapped with heavy, insulating tape if any of the ultra-high voltage gets through to tickle your hand.

The chiropractor who makes use of this in his work asks his patients if they have any infected teeth, or a toothache. One five-minute treatment over an aching tooth, rubbing the Argon bulb on the cheek over the infected area, stopped a toothache for us. It also relieved a throat, sore from too much lecturing.

General Electric also makes an identical Neon bulb. This gas throws off an orange radiation, which would probably give off more of a surface, heating effect because of the longer wave-length. You won't find these special bulbs in your local stores, probably, but will have to order them through some wholesale dealer, say an electric contractor and builder who has the GE catalog of special bulbs. A compact, professionally built Tesla Coil can be bought from Edmond Scientific Co, Barrington, New Jersey. It doesn't put out as much energy as the MWO coil designed by Bob Beck, but it will cause the GE Argon bulb to emit considerably more ultra-violet radiation than the 115 volt house current.

From the MAY-JUNE 1971 Journal of Borderland Research

A HIGH-POWERED MULTI-WAVE OSCILLATOR

A HIGH-POWER MULTI-WAVE OSCILLATOR

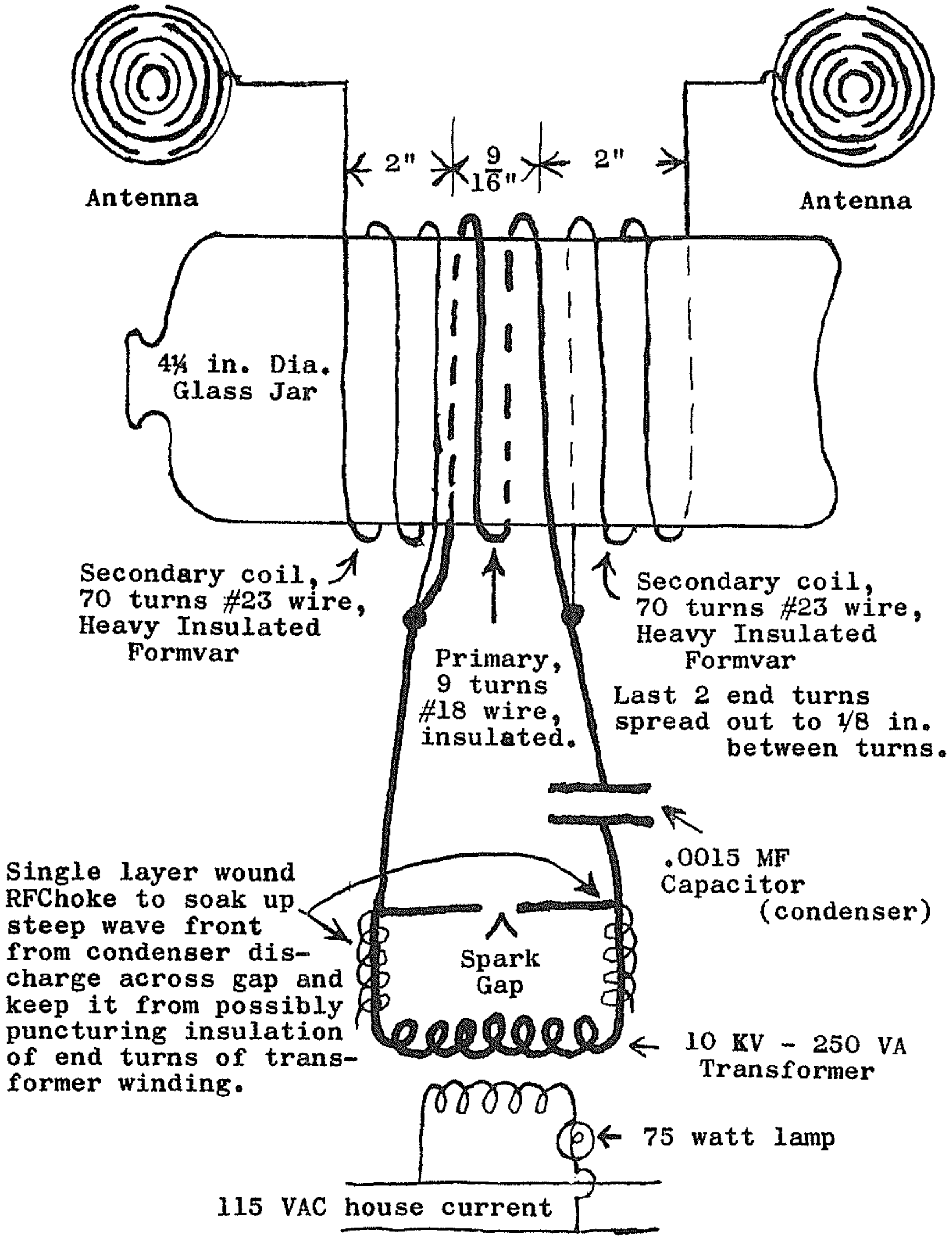
Antenna
Antenna
2"
$\frac{9}{16}$"
2"
4¼ in. Dia. Glass Jar
Secondary coil, 70 turns #23 wire, Heavy Insulated Formvar
Primary, 9 turns #18 wire, insulated.
Secondary coil, 70 turns #23 wire, Heavy Insulated Formvar
Last 2 end turns spread out to 1/8 in. between turns.
.0015 MF Capacitor (condenser)
Single layer wound RFChoke to soak up steep wave front from condenser discharge across gap and keep it from possibly puncturing insulation of end turns of transformer winding.
Spark Gap
10 KV - 250 VA Transformer
75 watt lamp
115 VAC house current

"Your brochure article stating that Georges Lakhovsky's original MWO had a separate Tesla coil driving each antenna is most interesting. This is in line with my experience here in Manila. I find that when livening up both antennas to full voltage, to spark each antenna down to the 3rd ring with an occasional slop-over to the 4th ring, it is much more effective than when keeping one antenna (or resonator) down to near ground potential; and it is least effective when actually grounding this antenna -- as Mr. Beck does --through a 1/10 MF capacitor!

"One five minute treatment a week is enough for me and the equivalent of two 15 minute treatments with only one antenna up to full voltage. However, I drive both antennas up to full voltage with only one Tesla coil. It has rather tightly coupled primary and secondary windings, wound as an auto-transformer, on a 4-1/4 in. diameter glass jar, as per enclosed sketch.

"The present spark gap is 1/8 inch, a fixed, constant gap, needs no turning with variable distances between antennas, subject or no subject. The sparking is much the same and the gap is not at all critical, 1/32nd, plus or minus, works just as well. Gap electrodes are 1/4 inch tungsten. The present primary circuit capacitor is .0015 Micro Farads, but this also is not at all critical, can be varied up to .0025 and still give much the same sparking and effect on subject.

"I tried using #28 Formvar wire, heavy insulated, 70 turns each on the secondaries, to shorten coil and get tighter coupling, primary to secondary, but coil was too short and flashed over from end to end.

"Just plug it in and use it, no tuning manipulation of any kind required. The input of 115 VAC to the 10,000 volt oil burner ignition transformer is only 17 VA through the 75 watt lamp as a limiting resistor. Never use more than a 100 watt lamp. This is all I can take. More power becomes rapidly depressing. And the full 250 VA out put of the transformer I once used, with a 5/8 inch spark gap, was nearly fatal. It took me four months to get over the effects of it. Pass this info on to your Associates."

J. Gilbert E. Wright

Associates can benefit from this dedicated borderlander's experience, which confirms Lakhovsky's warning that healthy cells can be over-stimulated and killed by powerful high-frequency waves -- as well by X-ray and radium waves. It may be easier to locate Neon Sign transformers of 5,000 to 7,500 volts output as power supplies for this MWO, at a cost of $15 to $20. In this case the wattage of the "limiting resistor" electric light could be increased to 150 or 200 watts, or eliminated altogether, if it is necessary to get enough power to cause random arcing down to the third and fourth rings on both antennae. The good doctor says nothing about shielding but care should be taken to prevent radio interference from this powerful spark.

"No data is given in the drawing (May-June 71 Journal) on the size of the antennas. I would suggest that the outer ring be about 18 inches in diameter with a 3 inch gap. Alternate down through the rings separating them about 2 inches apart and held together with plastic electrical tape. Wire should be #10 bare. The secondary leads connects to the second ring from the largest diameter outside ring."

George Van Tassel, Giant Rock, California

An even more spectacular possibility is the Time-Machine effect in connection with rejuvenation. If, as Bob Beck suggested in his article, the MWO tends to take the body cell back down the time-track to a more youthful, more vital period, what would happen to a person who stayed in the field of the machine for four hours? Would his vibratory rate be speeded up to where he would disappear completely? Like the hero of H.G. Wells' 'Time Machine'? Anybody want to try it? It seems, from hints George Van Tassel has been dropping over the years, that his Integratron at Giant Rock, California might do something like this -- if and when he gets it working! As I recall, both George and the Canadian Saucerian, the late W.B. Smith, spoke of the four dimensions of electricity: 1. is the electricity moving in one direction along a copper wire, 2. is the magnetic field around the wire, 3. is the static charge created by the first two effects, and 4. is Time -- or the 'tempic field' as Smith called it. The Time Field equates with Space, where consciousness is everywhere present. Seems to me Lakhovksy's Multiple Wave Oscillator is a gadget which gives us the opportunity -- for the first time -- of approaching the Time Field with scientifically acceptable equipment, operating in known frequencies.

BIOPHYSICIST CONFIRMS LAKHOVSKY'S THEORIES!

LA "Times", March 28, 1971: "The 'magnetic resonance' of the nuclei of atoms (in living cells) has been used to tell the difference between cancer and normal tissues. This resonance is the reaction of the nuclei to electromagnetic energy --first its absorption and then its emission. The process can be timed.

"Raymond Damadian, a biophysicist at the State University of New York's Manhattan Medical Center, said his achievement may open the way to major advances in the diagnosis of cancer in humans... Like other attempts to find a way of making early diagnosis of cancer, Damadian's research began with a hunt for significant differences between cancer and normal cells.

"He relied on a law of physics on which nuclear magnetic resonance is based -- that the atoms in each element in the periodic table absorb and then emit electromagnetic energy on an individual wavelength that is different from that absorbed and emitted by the atoms of any other element. In his system, radio waves are aimed at atoms in strong magnetic fields. If the wavelength is the right one for that atom, the atom will absorb the energy and then emit it, returning to a 'resting' state afterwards.

"The reactions differ according to the kind of molecule the various atoms are incorporated in. They also differ when very subtle changes have taken place in the molecules. The measure Damadian uses is what is called the 'relaxation time' of hydrogen atoms in water molecules within the (body) cells... He found that it takes longer for protons (nuclei) in cancerous tissue to 'relax' than it does for protons in normal tissue... The relaxation time of protons in normal liver tissue, as one example, was .25 of a second, compared with .86 of a second for cancerous tissue.

"Damadian's findings confirm the theory proposed by several scientists that water molecules in a normal cell are held in a fairly well organized structure by electromagnetic attraction exerted by the molecules in the cell. But, it was theorized, when a cell begins the uncontrolled growth of cancer, the structure begins to collapse because of the cell's increasing disorganization.

"Damadian thinks the device for detecting cancer in humans (his experiments were conducted on mice and mouse tissue) that may come out of his work would be a radio frequency coil to emit the electromagnetic waves to be aimed at the atoms in the water molecules under study. The device also would include a magnet to create the field required for measurement.

"The coil would be wrapped around the patient while the magnet would be moved back and forth over his body. The reading of how long it takes the atomic nuclei to relax would be made and then computers would do the matching up and make the diagnosis."

Interesting, isn't it, how the 1970s researches of biophysicist Damadian parallel the pioneering work of electrical engineer Georges Lakhovsky in the 1920s? The goals are quite different, of course, in that Damadian only wants to prove the existence of cancer so the victim can go on to be drugged, operated or radiated with destructive X-rays or Cobalt rays, at fancy prices, a procedure inspired by the soulless monsters who guide the American Medical Association from the Lower Astral plane.

The irony of Damadian's proposed diagnostic technique is that it combines two electro-magnetic healing procedures. If it is used on sick people, sooner or later someone is going to be healed of cancer while being diagnosed! Can you imagine the disagreeable shock this would cause in orthodox medical circles? Extensive surgery would have to be cancelled. The sale of drugs would be lost. Under these circumstances it isn't likely that his diagnostic technique will ever receive the approval of the AMA.

So, borderland science will have to continue to carry the torch of research on electro-magnetics in this area, as George Van Tassel is doing with his Integratron up at Giant Rock, California. In his newsletter "Proceedings" he gives more details of his research program there. We learn that in addition to creating a field of soft, high frequency radio waves, and a magnetic field, the Integratron will also bathe the subject in an atmosphere charged with negative ions.

HOPPING UP THE HIGHER FREQUENCIES

"There is one thing you may wish to try on the MWO. Close the center hole in the driven antenna with a washer and screw. Connect this to the driven ring (second ring from the outside) with a 500 mmf condenser, 10,000 volt rating. This increases the power of the higher frequencies. I think you will be pleased with the results. Mount the condenser at the back of the antenna. Mankind has been saturated with the lower frequencies for years; 60 cycles and their first few harmonics are almost everywhere in more than adequate power. Radio, TV, radar, etc. also saturate our world but the amount of power available to the average person is very small. In 1942 I used to cure my sinus trouble in minutes by getting close to the transmitter section of a radar transmitter. Few get that close, though, and it is a good thing as this power can kill. I have seen birds and larger animals killed instantly. It can happen to humans, too." H.B., Melbourne, Florida

SEND A FEW BUTTONS

"Received the Lakhovsky loops and am pleased to feel the response to the body. I wear my belt only at night. I gave one to a friend who was having a problem with her leg and anticipating hospital treatment for same and the response was also to the good... So all I want now is for you to send me a few buttons so I can make some more loops myself and help to lessen some misery around here, for someone. I tried to have some buttons made here, but no success." J.M.S., Charleston, S.C.

The non-slip loop buttons we make are of 1/16 inch thick, white fiber glass plastic.

MORE POWER NEEDED

"Enclosed is $. Please send me the magnet assembly with carbon rod (March-April 71 Journal). Have not had the hoped for results with the loops. Possibly the assembly, which can be used more directly over the research area, will be more effective." M.S., Ellensburg, Washington

If we've proven one thing in our borderland research, there is no single cure for any or all of mankind's ailments. Injury and disease are outward manifestations of an invisible cause, which may involve three or four levels of consciousness and extend back two or three lives. You Director has found in over thirty years of counselling people on their borderland problems, half of them don't really want to get well, they just want to be free of pain so they can go on indulging their vices. And by the time the hidden vice surfaces as a disease in some target organ of the body, an unhealthy trend has been established in the cell life there. It usually requires a strong jolt or shock to reverse this trend, such as a powerful magnetic field, a drug, chiropractic, even surgery -- anything to break the negative pattern which created the condition. One of the simplest and most effective methods, and cheapest, is fasting for a week or two. This gives the system a chance to purge itself of all filth.

OPEN ENDED COPPER BRACELETS

"Your article on copper bracelets sent me to the typewriter! It is quite well known in England and has figured on TV as anklets for racehorses! Open-ended copper bracelets sold freely in shops for some years here, but not recently. However, I have a copper bracelet from Rhodesia which I wear always, never take it off, and of course no arthritis or rheumatism. But I'd like your experimental loops for throat, waist and wrist.

J.G., London, England

THE PRO AND THE CON

"Thanks for the wire rings. Sorry to report no favorable results. My husband refused to try. I wore them several nights but seemed to feel even worse afterward. It was a good experiment. I guess my trouble is psychosomatic."

Mrs. A.H., Enid, Oklahoma

"I want to thank you for sending me the kit of Lakhovsky coils. I especially appreciate your taking the time to prepare a larger-than-usual coil for my chest. I guess I'm, too young to have any really serious physical ailments -- except for insomnia! And the coils eliminated that the first night I wore them to bed!"

J.T.D, Venice, California

*** *** *** ***

From the September-October 1971 Journal of Borderland Research

UPGRADING CELLULAR ACTIVITY WITH ELECTRO-MAGNETISM

By Aaron H. Steinberg, Ph.D.

For too many years now the scientific knowledge of electro-magnetics and its positive effects on cell life has been collecting dust and cobwebs on the shelves of ignorance, brought on by those special interests who fear the Truth. We are now seeing a rebirth, if you will, with scientists both professional and amateur, who are reviewing the work of their predecessors and making improvements in the mechanical application of magnetic fields.

In 1969 I visited the USSR to see for myself what and how they were utilizing Electro-magnetism. In several country homes for retirees I was most pleasantly surprised to see electro-magnetism used daily on most of the old folks. I should say young folks since may who claimed to be over 100 years young, looked and acted like most people at 50! The main object of the E-M was to reverse the ageing process by altering the cellular structure. I was most anxious to purchase a piece of the equipment, but alas no sale.

Seek and you shall find has always been one of my dominant characteristics. So when I learned that Japan was truly advanced in this field, having been using such equipment since 1936, I decided to go there in 1970. I sought out the one company that had been most productive in this field. They were most kind and cooperative. I was introduced to several scientists who had done much useful research. Their claims were backed up by scientific facts and I was convinced that they had a well constructed instrument, simple to use by anyone so desirous. They named their equipment The Magnetizer.

I returned to the U.S. with a Magnetizer and immediately began a research program. I can report that the results obtained to date have been most favorable, after nine months of experimentation. The magnetic flux generated by the coils can be measured accurately over any part of the body, thus determining those areas where the flux is penetrating as well as those where it is weak.

Chemicals and pollution are surely contributing to an unhealthy alteration in our cellular activity and human magnetic filed. It will be many years, if at all, before this mess can be rectified and controlled. But in the meantime, those who are aware and see the danger signals, will not sit by and wait for the clean up, which may never happen. Instead they will seek to upgrade their cellular activity so that the body's natural resistance may be at peak ohms. There is some indication that the aura force centers known as chakras or vortexes are stimulated by electro-magnetism.

From the Japanese scientists I learned the following about the effects of magnetic flux. It is different from ordinary electric current, which only flows along the surface of matter. It is different from X-rays, which do not penetrate the bones. The magnetic flux of ultra-long wave generated by the Magnetizer penetrates deeply in muscles, fat and bones and has an intensive effect on the nerves.

Magnetic flux never causes unpleasant sensations in the body, such as pain or shock, but instead produces comfortable, warm sensations. These sensations are also known as Joule's heat, which strengthens the function of the cell, corrects spasms and inflammations. When magnetic flux passes through tissues, a secondary electric current called the eddy current is created around the magnetic lines of force in the tissue cells, which ionized the protoplasm and rejuvenates the tissues as a result of activating metabolism. Furthermore, magnetic flux, in the process of penetrating the tissues, works to increase hormone secretions. These maintain youth by providing energy as a result of normalizing function of the internal organs.

Flux strongly stimulates magnetic substances in the blood, like iron. Accordingly, the hemoglobin in the blood vessels moves actively, accompanying the lymph circulation, when the Magnetizer is turned on. The therapeutic effect is not singular but collective, thus eliminating constitutional weakness.

*** *** *** ***

From the May-June 1971 Journal of Borderland Research Clips, Quotes & Comments

OSCILLATING LOOPS AND COILS

"I have a necklace which is definitely an oscillating circuit. I'm sure you have seen them. They are a wire around the neck instead of a chain. The only thing is that they come with a hook arrangement to clasp them. I cut the ends off, and rounded them to make a real oscillating circuit. Then, too, I got a pair of earrings which are simply coiled wires, like this:

One of these days maybe I'll have a headache (this is rare), and I will try either or both of these. Mrs. C.M., Watertown, Wisconsin

REMEDIES, HOMEOPATHIC AND ELECTRO-MAGNETIC

"You might spread the word around that the Copen Vibro-Potentizer actually can make any Homeopathic remedy in any potency from 1X to 10M.

"Since the medical-industrial complex is trying to destroy Homeopathy and its pharmacists, those who know of it and are doing research on its remedies would do well to get such a machine. It can also make a remedy from a radionic rate. I made Skull-Staphylococcus 10M and reduced an intractable sinus infection. The possibilities are fantastic and open up a new field in healing. One can make color remedies and aura remedies with greater power than radionic treatment of the same."

From his own research this Associate is suggesting that homeopathic remedies, being closer to physical matter than radionic treatment rates, may bring physical body changes more effectively. A hundred years ago in Philadelphia, Dr. Pancoast was curing practically every known human disease with colored light, either blue or red. But he (she?) also prepared color-treated dosages of oil or water for internal use, to back up or augment the shining of colored light on the outside of the body.

"May I remind you that the Multi-Wave Oscillator requires an antenna without any edges, to give efficient radiation. Lakhovsky's original had round bars or tubing with spherical ends. Edges lose most of the radiation. In addition the largest ring should be based on that fundamental resonance of the human organism, 3.66 meters as discovered by Bell Laboratories. This could be a quarter-wave antenna length of .91 meters or a circle of pi diameter.

"However, this should be doubled so that the human organism resonance is the first octave. Individual resonances vary slightly above and below 3.66 meters. By giving lower octave and with the variations supplied by the spark gap all wavelengths would be reached. Therefore, the outer ring should be 1.82/pi meters.

"The relative size of the outer rings could be determined by examining one of Lakhovsky's original antennas. The outer ring should be fed current at both ends, quarter wave style, I believe (to a H.V. Coil) not at the middle of the second ring as the Beck antenna does. The frequency of human resonance, 3.66 meters or approximately 84.6 megacycles is the key to radionics. [By my calculations, using the formula wavelength in meters (3.66 meters) equals (=) the speed of light in meters per second (300,000,000 m/s) divided by the frequency in cycles per second, I figure the frequency to be 81.97 megacycles. T.B.] The human sample or witness -- blood spot, etc. radiates around that frequency, varying according to the disease and the person, I believe. The homeopathic remedy also has a resonant frequency and alters the human frequency for better (or worse). In addition there are harmonics and possibly sub-harmonics.

"The resistance rates discovered by Dr. Abrams, followed by Guyon Richards, are still valid even though this work was done 50 years ago. The rate for Aqua Marina (sea water) put on a Copen V.P. produces a very effective remedy good for almost everyone. This rate comes from Richards' book 'Chain of Life', 1934."

S.M.S., Chicago, Illinois

We welcome this Associate's observations on the MWO antenna, radionics rates and homeopathic remedies.

Riley Hansard Crabb, editor JBR 1959-1985

BSRF MWO Update - Issued May 31, 1988

Compiled by Tom Brown from the BSRF Files
In consultation with Eric Dollard & other BSRF Associates
(Contains the complete 1986-87 MWO Update materials)

As there has been a continuing strong interest in the Lakhovsky Multi-Wave Oscillator we felt the need to compile this information into the present book. This documentation is the result of a wide variety of analysis and experimentation with the MWO and covers many years of active research.

There have been many comments made concerning the inherent healing effects of the MWO, but as it cannot be stated enough, the device is merely a tool of research, of which one may draw their own conclusions as to whether or not the device actually has any healing properties. It may be interesting to note that in a personal conversation with Dr. Ita Wegman, Rudolf Steiner stated that electricity, magnetism, and the force upon which nuclear phenomena is based, are all corrupted ethers.

We have found through studying Goethe's indications on the phenomenon of Light and Color in conjunction with Eric Dollard's high frequency researches that electricity is a reflection of the Light Ether. This matter is being discussed further in The Journal of Borderland Research, so it is sufficient for present reference to say that Light has its two poles, red-yellow and blue-violet. The red-yellow pole in nature is longitudinal (direct rays from sun) and the blue-violet pole is a progressive type of transverse wave (rounded sky, Reich's KRW wave/blue orgone). Electricity has its two poles, electro-magnetism: red-yellow-hot, (retarded transverse); and dielectricity: blue-violet-cold (longitudinal). Hence, there are experimental indications that electricity really is a reflection of the Light Ether.

In light of this trail of thought we quote with permission from a 1986 private communication with Trevor James Constable: "The MWO I regard as a false path. The cellular stimulation theories sound and seem good because they are mechanistic, and therefore sympathetic with the mechanistic weltanschaung. With the MWO you can avoid the etheric--or so it seems--and therefore those who feel adrift in the etheric ocean cling to things like this. Healing is one hundred percent an etheric problem. Any dysfunction or hyperfunction not caused by direct trauma has to be tackled via the ether body, and even the direct traumas respond best to etheric correction, such as getting the displaced double back into occlusion through the use of arnica compresses, etc., etc. In between the MWO antennas, there is a general marshaling or focus of light and chemical ethers. Its quite evident even with a minor extension of vision. This is the healing agency, not the sparks. The ether rushes to suppress the EM activity, and because it is low power, the ether accumulates around the spirals and in between them. Any stricken area placed in such a high concentration of formative energy, is going to respond with a return towards normal. BUT, why go to all that trouble when a simple orgone blanket--a layered structure that can be directly applied to a traumatized or dysfunctional area--does a much more efficient, controllable job without that infernal juice. In my time with Ruth Drown, I heard it over and over again, and I pass it on to you: keep that 115V juice away from the organism. The post material EM energy is life negative, but no one seems to feel that they can get results unless they INJECT the alien EM activity into the stricken carcass. It's pure ignorance really. My own experience with orgone accumulators and blankets leaves me with no other choice but to regard the MWO as superfluous. A good rule for any experimenter: If it gets simpler, you are on the right track. Remember always Steiner's dictum: "the spirit of man is always healthy!.....it is the obstruction or dysfunction in the lower sheaths that distorts the manifestation of the spirit and personality via the body."

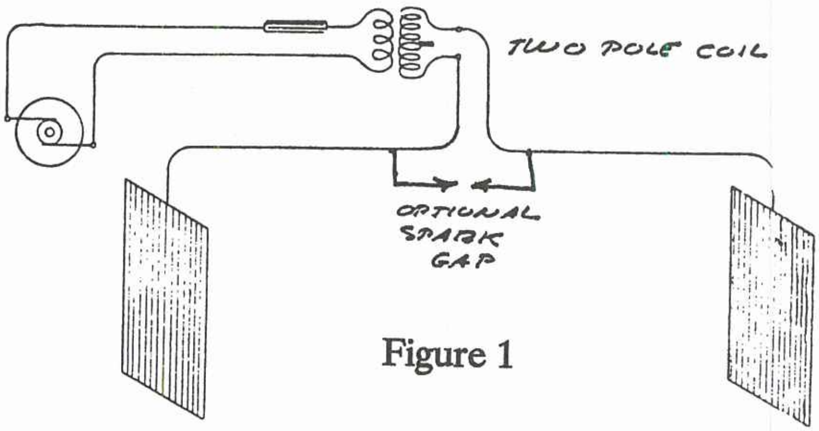
TWO POLE COIL
OPTIONAL SPARK GAP

Figure 1

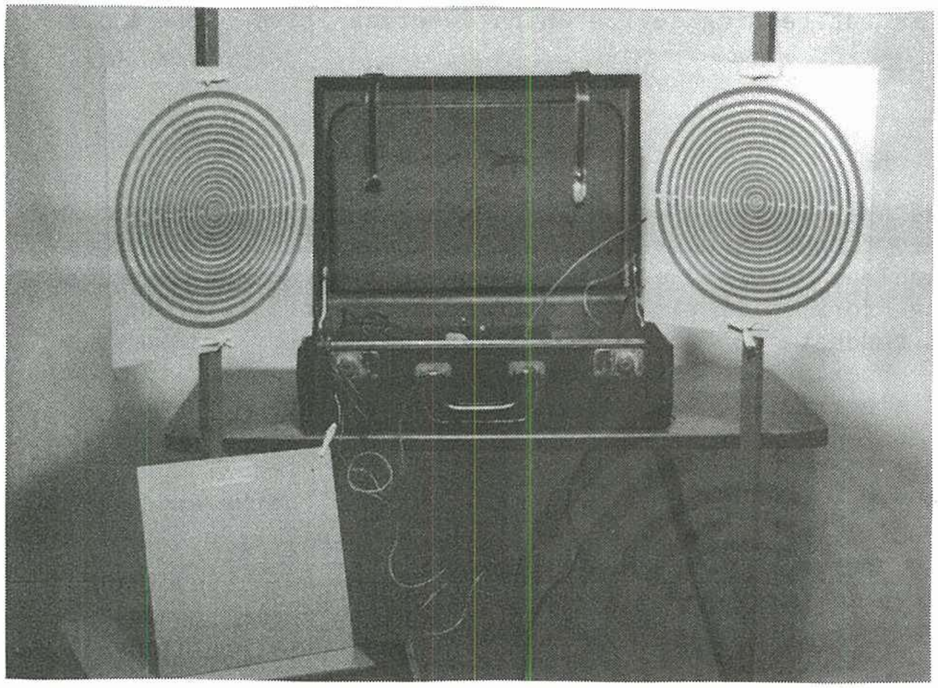

Photo 1
"Beck" style MWO from the BSRF collection (note: foot plate at lower left goes to ground)

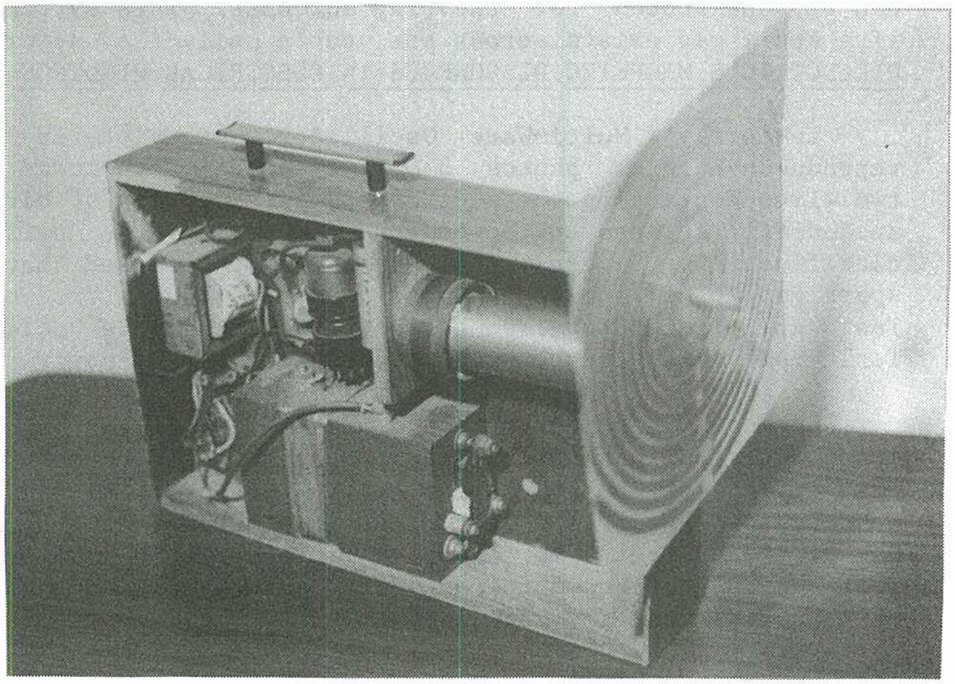

Photo 2
Single antenna version from the BSRF collection (Resonator antenna can be attached at ground)

Not much more need be said in that vein except that of a reinforcement of Trevor's comment. Our present society is not on the path of Nature so therefore some people feel the need for a device such as this. Wilhelm Reich discovered how to concentrate the living energy of the universe, which he termed the Orgone. The Yogic sciences teach us how to accumulate this energy without any outside apparatus, so therefore the simpler and more passive a device is, the closer it resembles Nature.

In the following pages you will find some of the results of MWO experimentation and consideration by BSRF and Associates. We hope the information proves useful, but again it is for research purposes only. We do not recommend the MWO for medical use because we are not qualified to give medical advice, nor do we want to create any false hope where one may neglect professional health care. We do fully support a persons right to investigate and research with devices such as the MWO. No one has all the answers and the MWO is an interesting device which deserves true scientific verification or rejection.

THE BASICS

The general theory behind the MWO is to excite a concentric ring antenna which produces a wide spectrum of radio frequency waves. In this theory it is understood that the RNA-DNA coil in the nucleus of every cell has a resonant frequency within this range. Just as a tuning fork will ring when an identical fork is struck in close proximity, the cells will resonate to their individual frequency pulled from the frequency ocean of the MWO. Most people don't consider that there are different types of waves and two poles of electricity, and these distinction have not generally been noted.

The antennas on the currently popular "Beck" MWO system, serve as simple capacitor plates which transmit dielectric waves out of the Tesla Coil (see figure 1 and photos 1 and 2). The second MWO plate in such a system must be connected to a second Tesla coil or to a double pole (half wave) Tesla coil (May-June 1971 JBR), not a single pole (quarter wave) Tesla coil as per the Bob Beck/Klark Kent style MWO. Action in the MWO antenna is neutralized by the fact that capacitor currents flow radially inward and little electron current flows around the loops. Copper plates can be substituted for the ring antennas. This system operates 1000 times stronger if a spark gap exists across the double coil. A description of the action is in DIELECTRIC & MAGNETIC DISCHARGES IN ELECTRICAL WINDINGS by Eric Dollard from BSRF.

Lakhovsky's Multi-Wave Oscillator, as pictured in THE SECRET OF LIFE and reproduced here as photos 3 & 4, created a structured electro-magnetic field fed by two wires contacting the ends of the outside ring of his antennas. This is obviously an end-fire antenna so the component reaching the person is a novel energy needing clarification. Several researchers have expressed that Lakhovsky used a quenched spark gap which creates damped waves allowing free oscillation in the antenna. A Tesla coil is not required since the Multiple Wave Oscillator IS THE COIL. However a Tesla coil still can be used as the high voltage power supply.

Lakhovsky's patents will cover both types of units so it seems as though he experimented with both, but the pictures in THE SECRET OF LIFE are an indicator as to what he actually used. After having gone over the patents I'm beginning to think that Lakhovsky buried his secret to mislead copycats. That is the purpose for a patent, to protect the inventor. The Tube MWO Patent 2,351,055 gives some interesting insights into the overall picture of Lakhovsky's work and study of it will provide many research leads.

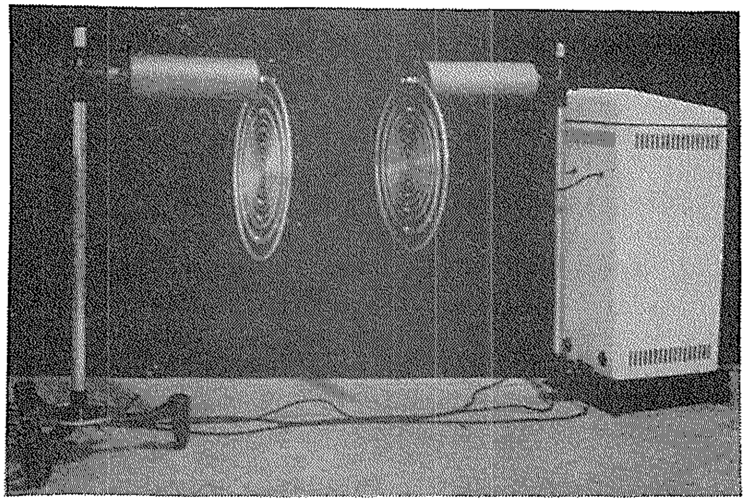

Photo 3
Lakhovsky's Multiple Wave Oscillator

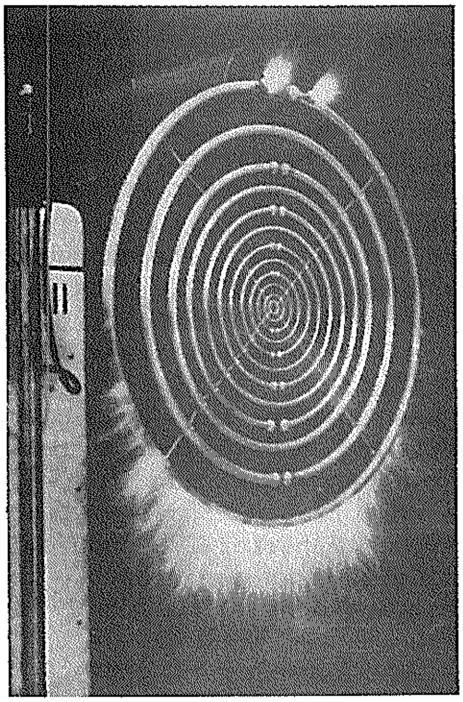

Photo 4
Lakhovsky's MWO. Close-up of transmitter in action, showing *effluve* (electric brush). Note this is end-fire antenna. Regeneration may have come from novel component at 90° to antenna plane.

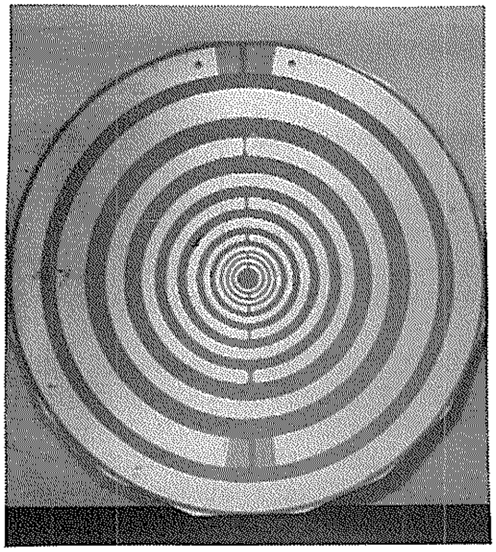

Photo 5
Yagi MWO antenna

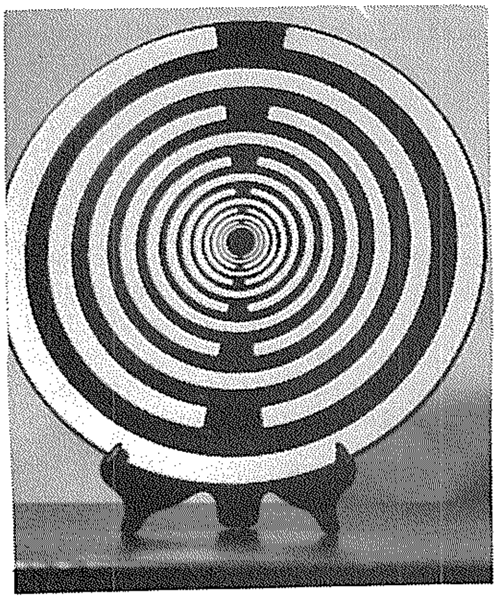

Photo 6
Golden Ratio – Log Periodic
MWO antenna

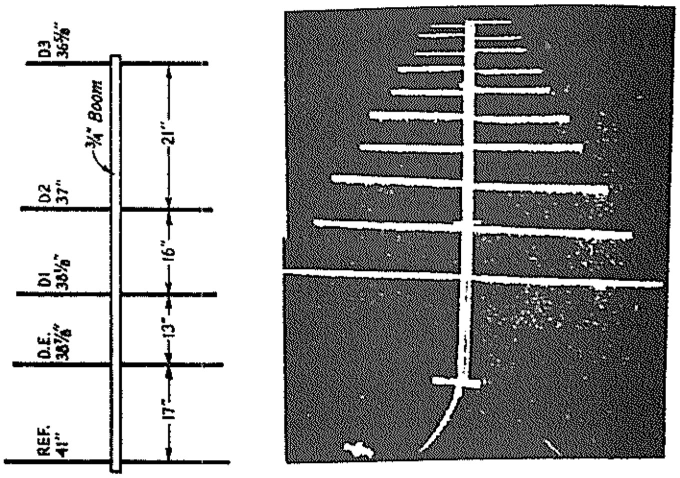
D3 36⅝"
¾" Boom
21"
D2 37"
16"
D1 38⅛"
13"
D.E. 38⅞"
17"
REF. 41"

Yagi antenna

Log periodic antenna

Figure 2

LOG PERIODIC ANTENNAS

Our research indicates that Lakhovsky worked with log periodic antennas and may have been the first to use them. The antenna will create a harmonious structure in the ether. The Beck style MWO produces a dielectric field 90 degrees out of phase with this, and utilizes the dielectric breakdown (arcing) of the field. This is basically a radio frequency noise generator and the most popular derivation of Lakhovsky's first MWO patent.

Scaling the rings non-uniformly using correction factors is not needed according to log periodic antenna theory (see figure 2 and photos 5 & 6). Lakhovsky states that any high frequency radiating circuit will drive the antennas. We are striving for waveforms mirroring the harmonic laws of nature. As an example the old Yagi antenna has non-uniform distribution of the elements. Log periodic design uses the uniform distribution of the elements such as is found everywhere in nature. This presents superior MWO performance. Lakhovsky's Speaker and Microphone Patent indicates that the log periodic design can be applied to sound as well as radio frequency energy.

LAKHOVSKY'S MWO

Figures 3 & 4 are a preliminary attempt by Eric Dollard to accurately reproduce the MWO system as pictured in SECRET OF LIFE. Figure 3 is a simplified version for a single antenna, figure 4 is a double antenna system. Photo 7 is an experimental setup utilizing this design. We feel that Lakhovsky was working with quenched spark gaps which produce damped waves in the form of the antenna. This type of circuit is needed to power log periodic antennas to create their structure in the ether harmonious with the Golden Ratio structure of the human body and life in general.

This information is not definitive, but is the result of experimental and theoretical considerations. We present this so that Borderland researchers can carry on with new input. We are not saying that the "Beck/Kent" models do not work. There are many claims that they do, but based on Tesla's theories, rather than Lakhovsky's. Figure 1 unit works off of longitudinal dielectric waves (displacement current). Figs 3 & 4 work off of Transverse electromagnetic waves (Hertzian) and residual scalar waves.

Figure 5 is a low power driver based on the Steinmetz lightning wave, a reproduction of the wave naturally occurring in the atmosphere. Values given need to be experimented on for optimum pulse formation. A mathematical analysis of this wave by Steinmetz is included as an addendum to this book. See DISTRIBUTED SERIES CAPACITY.

GOLDEN RATIO ANTENNA

Eric Dollard has designed a log periodic antenna based on the Golden Proportion, as found in living systems such as plants and animals. These antennas are two sided, the first side looking somewhat similar to currently used MWO antennas, the second side being capacitive loading (as Lakhovsky mentions, he used spheres on the ends of his antenna rings) based on the Golden Ratio. This design can be seen in photo 8. These antennas are gold plated to provide direct contact with the ether. Their actual size is 12" in diameter.

Figure 3

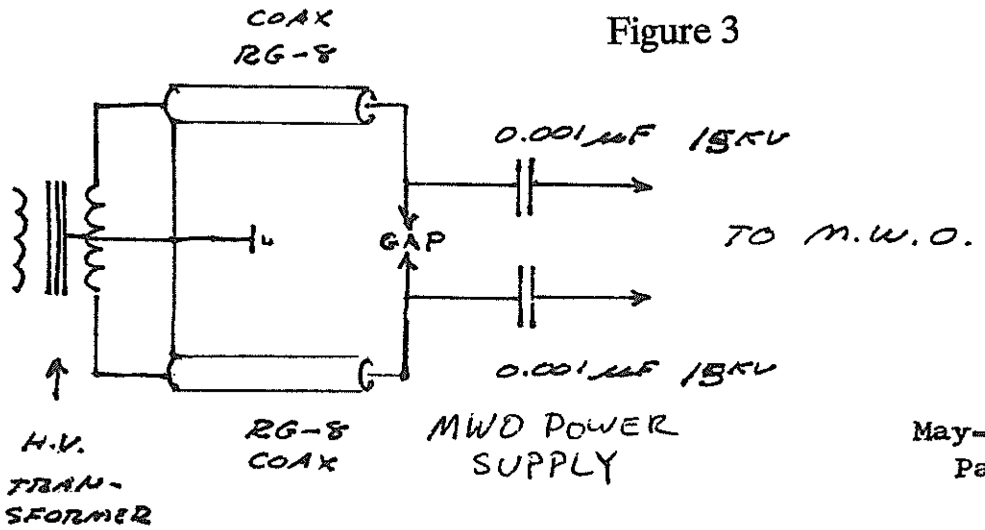
COAX
RG-8
0.001 µF 15KV
GAP
TO M.W.O.
0.001 µF 15KV
H.V.
TRAN-
SFORMER
RG-8
COAX
MWO POWER
SUPPLY

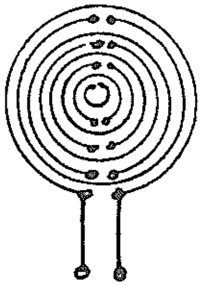

May-June 1986 JBR,
Page 33

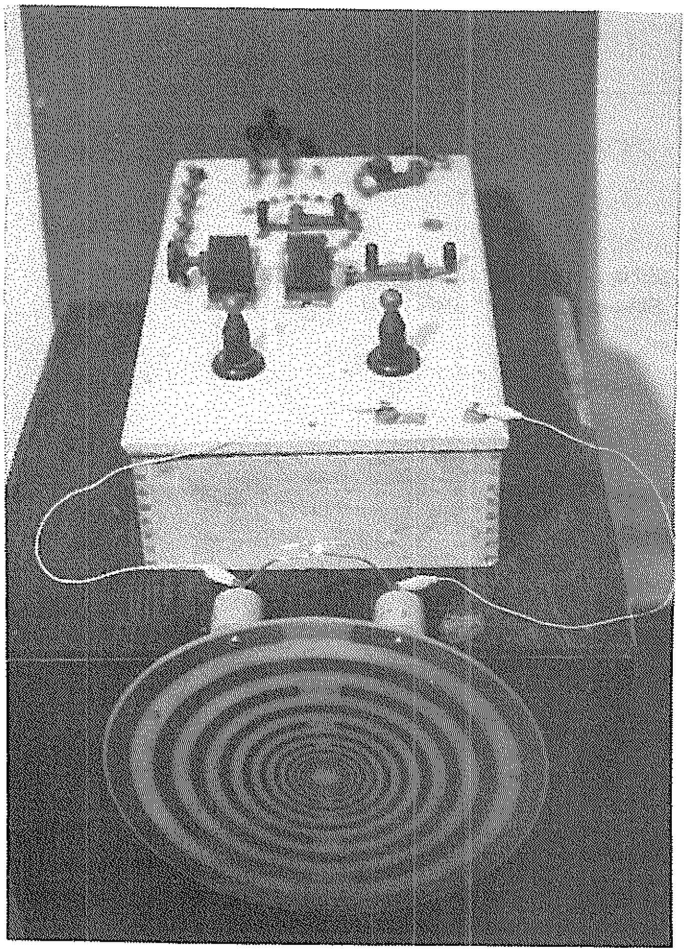

Photo 7
Golden Ratio antenna driven by 1919 era double-phase marble-top medical Tesla Coil. Mid-frequency D'Arsonval currents are utilized in this arrangement. Setup is as per Figure 3 (above), and is our understanding of the arrangement as utilized in THE SECRET OF LIFE. Coil donated to BSRF by John Crane from laboratory of Royal R. Rife.

Action of Lakhovsky setup

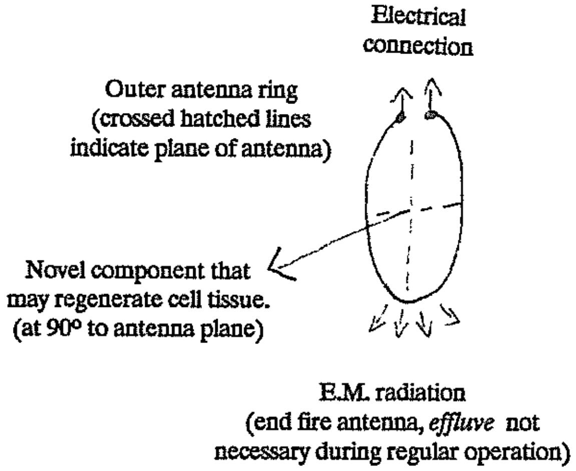
Electrical
connection
Outer antenna ring
(crossed hatched lines
indicate plane of antenna)
Novel component that
may regenerate cell tissue.
(at 90° to antenna plane)
E.M. radiation
(end fire antenna, *effluve* not
necessary during regular operation)

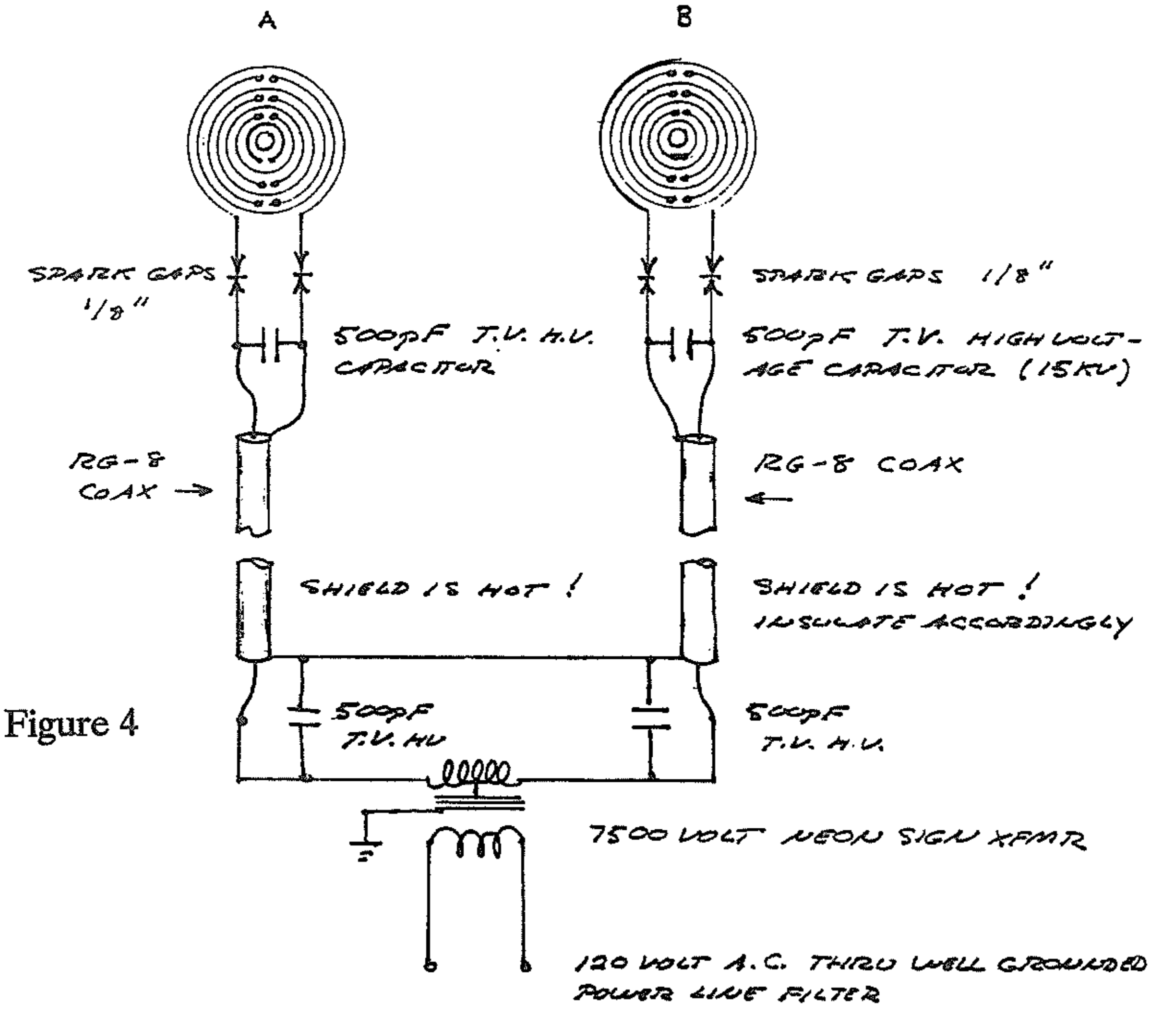
A
B
SPARK GAPS 1/8"
SPARK GAPS 1/8"
500pF T.V. H.V. CAPACITOR
500pF T.V. HIGH VOLTAGE CAPACITOR (15KV)
RG-8 COAX →
RG-8 COAX ←
SHIELD IS HOT !
SHIELD IS HOT ! INSULATE ACCORDINGLY
500pF T.V. HV
500pF T.V. H.V.
7500 VOLT NEON SIGN XFMR
120 VOLT A.C. THRU WELL GROUNDED POWER LINE FILTER

Figure 4

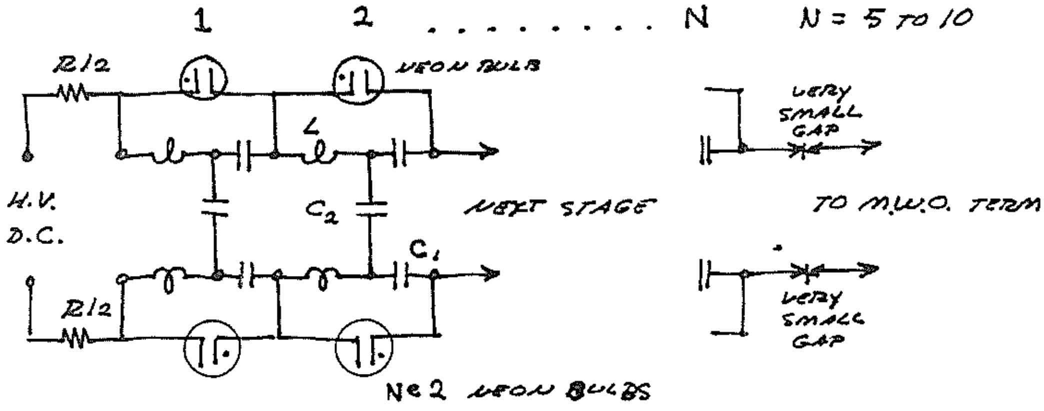
1
2
N
N = 5 TO 10
R/2
NEON BULB
L
H.V. D.C.
C_2
C_1
NEXT STAGE
R/2
VERY SMALL GAP
TO M.W.O. TERM
VERY SMALL GAP
Ne2 NEON BULBS

Figure 5

D.C. INPUT = 240V / STAGE ALL C_1 RATED 600V

R = 100KΩ / STAGE 1 WATT / STAGE

L = 1 μH / STAGE C_1 = 100pF × № OF STAGES

C_2 = 0.02 μF / STAGE 600V × № OF STAGES

The term Golden Mean, Ratio or Section refers to the mathematics of life, which is a logarithmic function. This ratio can be found everywhere in nature. It is the Sacred Geometry used in the architecture of ancient Greek and Roman temples, the Great Pyramid, etc. It has also been concluded by certain Orgonomists that "the Golden Ratio is a basic mathematical property of the orgone energy" discovered by Dr. Wilhelm Reich. (Journal of Orgonomy, V.8, N.2, Rosenblum, THE GOLDEN SECTION). It is evident from Lakhovsky's MWO Tube patent that he was aware of the Log Periodic design.

The MWO has usually been understood as having to be driven by a high-frequency source such as Tesla Coils and neon transformers generate. However, those who have looked into Lakhovsky's work know that he used simple wire circles to cure plants of cancers, and to make plants grow better in general. The Golden Ratio Antenna is a series of Lakhovsky rings because the rear conjugate strips on the antenna close the loops and in effect make Lakhovsky oscillating rings which are a complete circuit for focusing the ethers. The log periodic (following a logarithmic pattern) antenna design carries this effect further and preliminary indications are that the log periodic design creates a structure in the ether even without power, in may be an orgone accumulator of sorts. Futher experimentation in this vein is being undertaken.

Log periodic antennae have the interesting property of creating a virtual (non existent but functioning) antenna extending about 1-1/2 times beyond the actual. It is also frequency independent and is an excellent antenna for Amateur Radio communications.

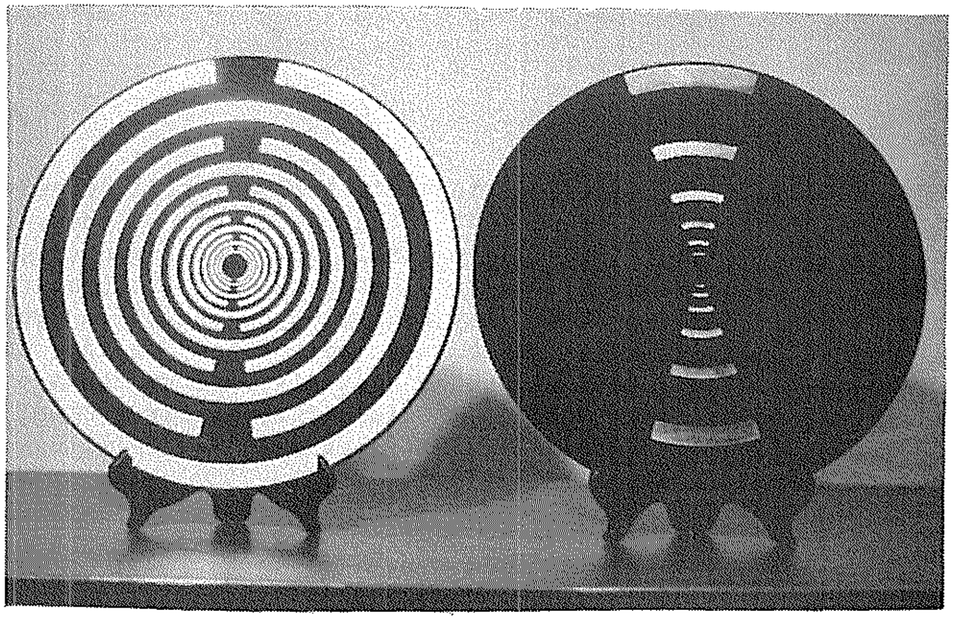

Photo 8
Golden Ratio antennas designed for MWO use by Eric Dollard. Forward side similar to Lakhovsky antenna. Reverse side provides capacitance as did metal spheres at ends of original Lakhovsky antennas. (see Photo 4)

The ethers harmonize to Golden Section Mathematics, (as demonstrated by Trevor Constable in his PRIMARY ENERGY WEATHER ENGINEERING ON THE HIGH SEAS video), so it can be understood that a living organism is a discharge of energy into the Golden Section space. We have observed the discharge of sprout energy into living sprouts when one of these antennas is directed at them. We found that it makes the sprouts grow better. Then we found that simply the pattern makes the sprouts grow better. So here we present you with the Golden Ratio Pattern that you can photocopy and use for experiments. This pattern may be copied as many times as you like for experimentation. We have reports that this pattern keeps the scale out of water heaters and fish tanks.

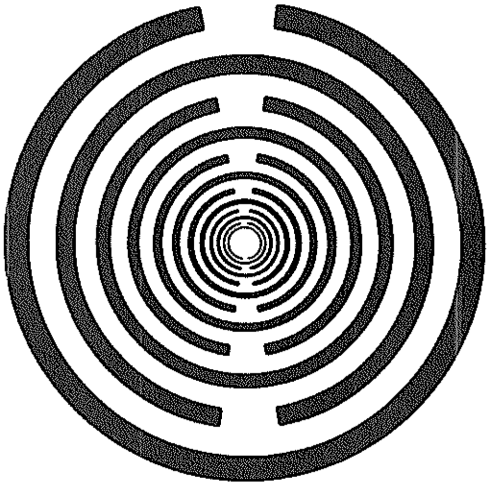

© 1986 BSRF

Log Periodic -- Golden Ratio design for research into the subtle vibrations from patterns.

In 1986 Borderland produced a limited amount of Golden Ratio Antennas in copper and gold. These were precision engineered Log Periodic antennas, and were quickly absorbed into the research network. They can be reproduced if sufficient interest is generated through this book. There is an appendix on log periodic antennas in this book for those who wish to pursue this interesting subject directly related to MWO research.

BUT - DO THEY WORK?

For all the material in this book we still haven't personally seen any solid documentation on the results of either type of MWO that could be re-verified as true to any investigator. I've seen listings and testimonials, but even the photos in THE SECRET OF LIFE look doctored to enhance the cancers (tar can be seen on the wound and the hair!). What was Lakhovsky really up to and which type of device works better? If anyone reading this knows any more than we do then please drop a line. In Lakhovsky's Tube patent he states that his instruments are in use the world over. It is important to check out one of his originals first hand, if any readers have access to such please drop Borderland Sciences a line. We'll make the information available for research. Thanks.

CONCLUSIONS

There are no conclusions at this point in time. Borderland has been publishing MWO information for 25 years now. This book marks the beginning of a new phase of research and development. Perhaps in 1995 after people do some serious research into the subject of the Multi-Wave Oscillator can some conclusions be drawn.

It is important to keep in mind during these researches that just because a frog leg will jump when electricity is applied it doesn't necessarily mean that electricity and life are the same thing. If we labor under this (quite common) delusion we may as well hope for our television sets to "heal" us of our ills. Rather, as astute phenomenological observers, we see that there is a relationship between electricity and life and in the understanding of that relationship we can come to a clearer comprehension of our place in this manifest universe.

Golden Ratio MWO Antenna

Measurements for reconstruction of Golden Ratio MWO Antennas (in inches starting from the outside ring). Template artwork is shown actual size.

Ring	Outside diameter	Inside diameter	Gap
1)	12.00	10.76	+
2)	9.43	8.46	-
3)	7.42	6.66	+
4)	5.83	5.23	-
5)	4.58	4.10	+
6)	3.60	3.22	-
7)	2.83	2.54	+
8)	2.23	2.00	-
9)	1.75	1.57	+
10)	1.38	1.24	-
11)	1.08	0.97	+
12)	0.85	0.76	-

Front & back strips are the same width. Gap on front is an 18° triangle, sections on back are a 36° triangle. Back sections cover gap on rings with 9° overlap on each side of ring. This provides for capacitive loading of the antenna.

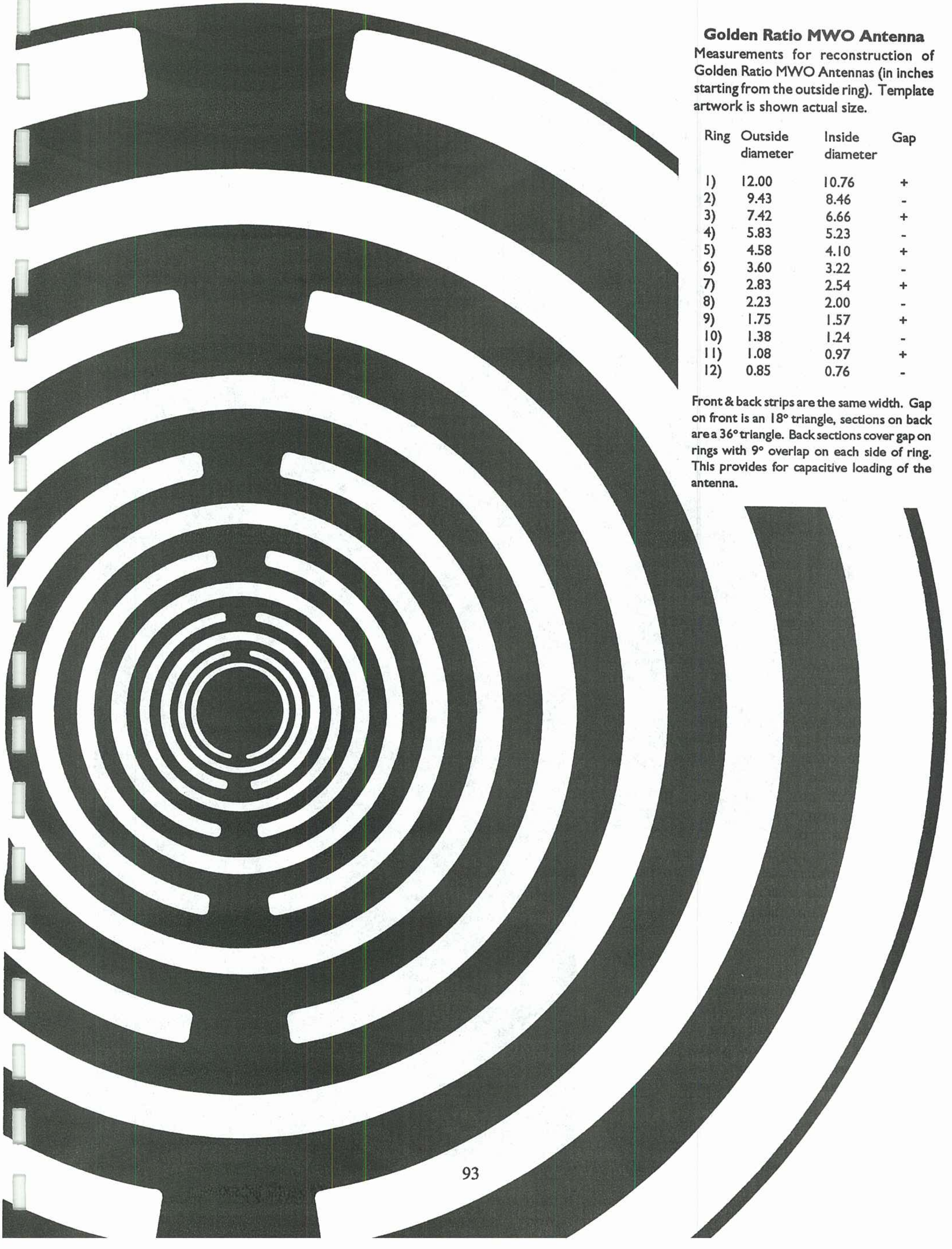

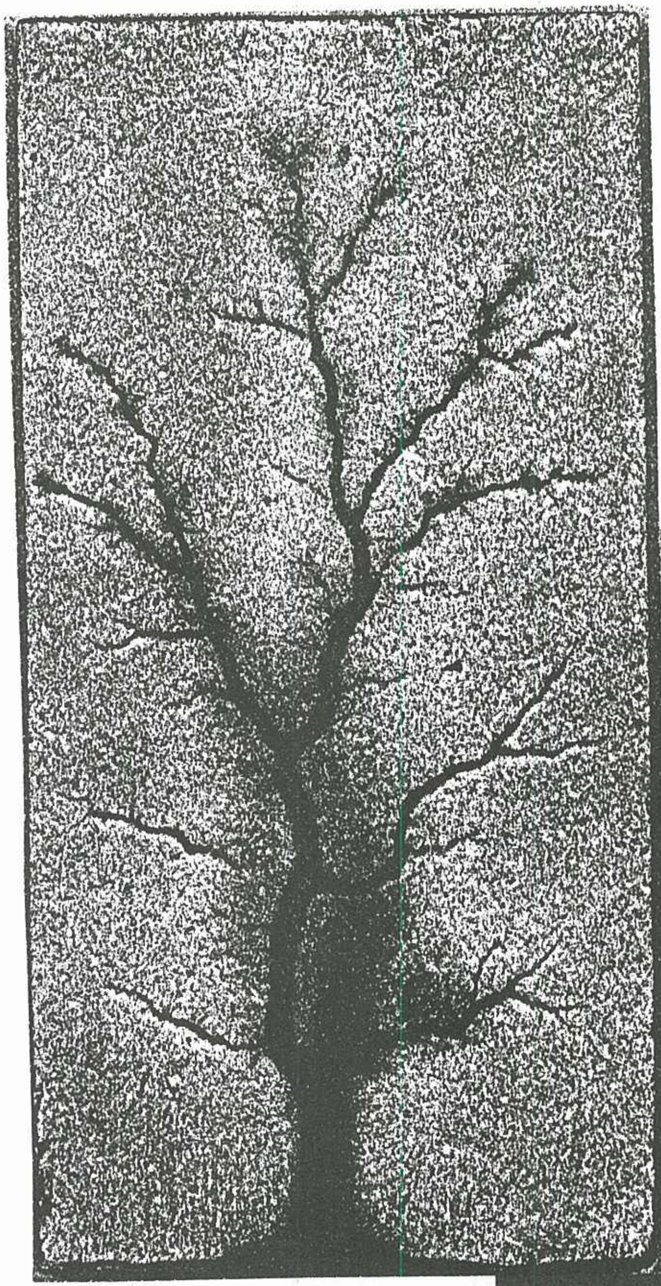

At left is a Golden Ratio pattern burned into particle board using Tesla Coil pictured below, seen here discharging into free space. Scientist Eric Dollard, builder of this coil, has deduced that the Golden Ratio is a property of space itself. The proportions are decided by the amount and type of discharge.

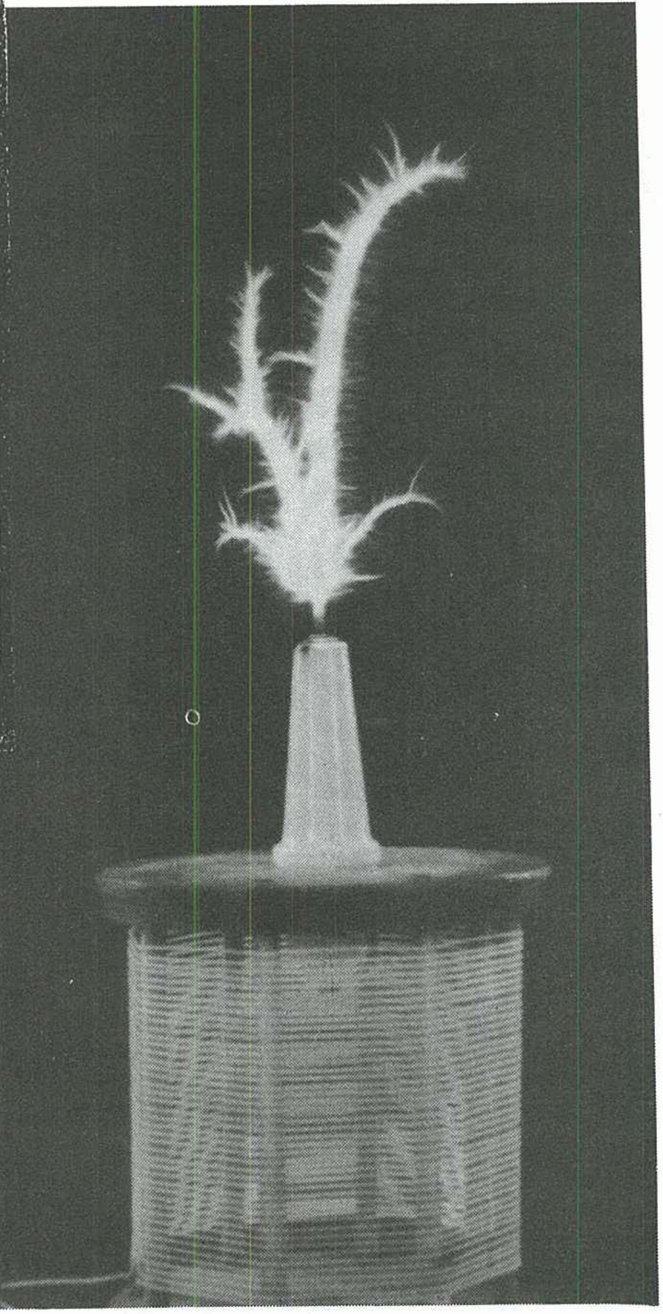

THE SPIRAL COIL MULTI-WAVE OSCILLATOR

Louis A. Schad

DISCLAIMER

Let me make it perfectly clear, I present this information for information and research purposes only and I make no suggestions or recommendations as to its use for healing or medical purposes whatsoever. If you are sick see your doctor. I present to you only what I have learned in my research and observations. I am not a medical doctor and the information I give should be used for education and research only.

To preface this, I have been interested in and studied and collected articles on the unusual sciences for over 30 years and have built many unusual devices. My attention was directed to the Lakhovsky M.W.O. by an article given to me by a friend. I read it several times and studied it and then put it away, (punched holes in it and put it in a notebook). I don't know why -- I can't tell you -- but I couldn't get it out of my mind, so I got it out again and read it some more -- this time going through it sentence by sentence digesting it -- now I was hooked!

I tried to gather all the information I could on the M.W.O., and as usual it wasn't easy -- I had to dig. Bottom line, I started to build Bob Beck's high power multi-wave oscillator -- I still haven't finished it, but I will. The more I studied the M.W.O. (being an average person -- not an electrical engineer, but a mechanical engineer) it seemed to me, complicated, difficult, and had to be adjusted for each person and circumstance. Looking at it now I see that isn't all true -- it was me. Back to square one -- I search for some way simpler, and out of it came the spiral coil M.W.O.

I am a recent member of Borderland Sciences and I have read many fine articles published in this Journal. Many of the articles are very high-tech and delve into the how, where, when and why, and that is how it should be if one is to research. Questions have to be asked and answers have to be found. I am not a professional writer and it isn't one of my greatest talents, so I ask that you have patience with me and I will do my best to tell you about my spiral coil M.W.O.

When I was trying to develop this spiral coil M.W.O. all I wanted was results and I didn't really care how or why it worked, my goal was results, and in doing so I tried many things that I wouldn't have normally tried had I planned each step and tried to fight my way through with pure brain power. In retrospect, I can see that I approached this project with a more light hearted attitude and I have enjoyed developing and building this M.W.O. But at times I did have problems to work out.

I experimented by winding coils, and coils, and coils. I got to be the electric supply house's best copper wire customer, or at least it seemed that way. After boxes and boxes of copper coils, I was ready to throw in the towel and give up -- but having the handicap of being stubborn like I am, I decided to wind one more coil, so I went out to the garage and wouldn't you know it, no more wire. I looked around and all I could find was some old solid strand 16 gauge house wire. It was too light and flimsy for what I wanted, so in desperation I decided to take 2 pieces and twist them together to make 1 heavy piece of wire. Well -- bingo -- I made a coil and got a tingle, that did it. After winding more coils and more coils and experimenting I got to where I am now, so I present to you the Spiral Coil Multi-Wave-Oscillator.

Warning -- this unit is deceiving (it just sits there and looks innocent like it couldn't do much), in fact some of my friends upon first seeing it have laughed at it, and commented you have to be kidding. Another friend of mine asked me to help him build one -- which I did -- and he was lying on his bed in his shorts with a sheet over him and had the unit about 2 foot above the bed and he fell asleep for about 3 hours (he had an infection from a wound in his leg). When he woke up he had a white blister on him about the size of a nickel -- he has done this twice. I warn you, use it with care, it is more powerful than it seems. I think he percolated the infection up out of his leg, as he is fine now.

Now, I know someone is going to say -- I made mine from an old coat hanger and hung it off the coiling with a piece of string, and I'll say -- gee that's nice, do what you wish -- but if you want it to work and work properly build it as it is, use it, experiment with it, and then change it as you wish.

Warning -- persons who have or are diabetic should not use this unit, as it may change the blood sugar level in some people. Diabetes is one of the few things that a M.W.O. will not correct or help. Lakhovsky states this in his book **THE SECRET OF LIFE**. I think I know why but I won't get into it now.

NOTE -- The spiral coils of this unit work all the time -- that is -- they never stop -- they work with or without the vibrator. Who said there is no free energy? If you put it over a bed with an electric blanket, it will charge the blanket and you will probably have difficulty sleeping. If you do not swing it far enough away from the bed at night it will cause unusual dreaming.

As near as I can tell this coil broadcasts at about 120° angle, although its peak energy is centered straight down from the small coil.

Question: is it patented or am I going to patent it? Answer: NO -- as far as I'm concerned it's free to everybody and that's as it should be.

The advantage of using a spiral is that one should get all the frequencies up to the maximum diameter of the outer diameter (O.D.) of the spiral. If you draw a spiral on a paper and place one point of a compass at the center, you can choose any diameter you wish up to the coil O.D. diameter and the scribing diameter will fall some place on the spiral.

This unit can be operated in 6 modes:

1. The coil only.
2. The coil plus the vibrator.
3. The coil plus the vibrator plus the coil shocker.
4. The coil plus the vibrator plus constant coil charge.
5. The coil (no vibrator) plus the coil shocker.
6. The coil (no vibrator) plus the constant coil charge.

So as one can see there is plenty to experiment with here.

NOTE -- I also tried connecting a ion generator to mine -- that really blows me away -- try it if you have one.

If you are experimenting with magnets -- be careful if you use them with this unit as some people get bad reactions. I personally have used magnetic water and this M.W.O. and have had no bad reactions.

Warning -- although I have -- I suggest you do not use this on the head or above the neck.

Warning -- if you have a heart problem do not use this unit as in some people it tends to reduce blood pressure and lower the pulse or heart rate.

Well -- now that I've scared the hell out of you, if you still want to build it I'll tell you the good side. In most all people who have used it or tried it they get a tingle or a feeling of heat (light). Most people get a boost of energy when they use it and it seems to help circulation. I am not sure why. I don't know if it tends to dilate the blood vessels of not. One man that used it can now wiggle his toes and he hadn't been able to do that for about 3 to 4 years.

I know someone will ask what type or kind and amount of energy is emitted by these coils and the answer is I do not know, but I do know that it works. I will leave that to someone who has the proper instruments and the know how to check it, so there is still a lot of work to be done here if someone wishes to do it. Now, before someone starts climbing my ladder, let me say that I give the credit to Lakhovsky -- what I did was to modify, update and expand upon the base of some of his work. You do not have to be gifted in the knowledge of the sciences to build this M.W.O. and do not need special equipment or tools. However, there are little tricks that I have learned that make it easier to build that I will give in another article. The body or base of this unit is made of P.V.C. pipe, also the vertical and swing arm, the vibrator base and also the drop tube that hangs the coils. We want a unit that is low in cost, tough and durable, light in weight, corrosion resistant, portable, easy to fabricate (and repair if necessary) and looks nice, so we will build our unit out of P.V.C. pipe.

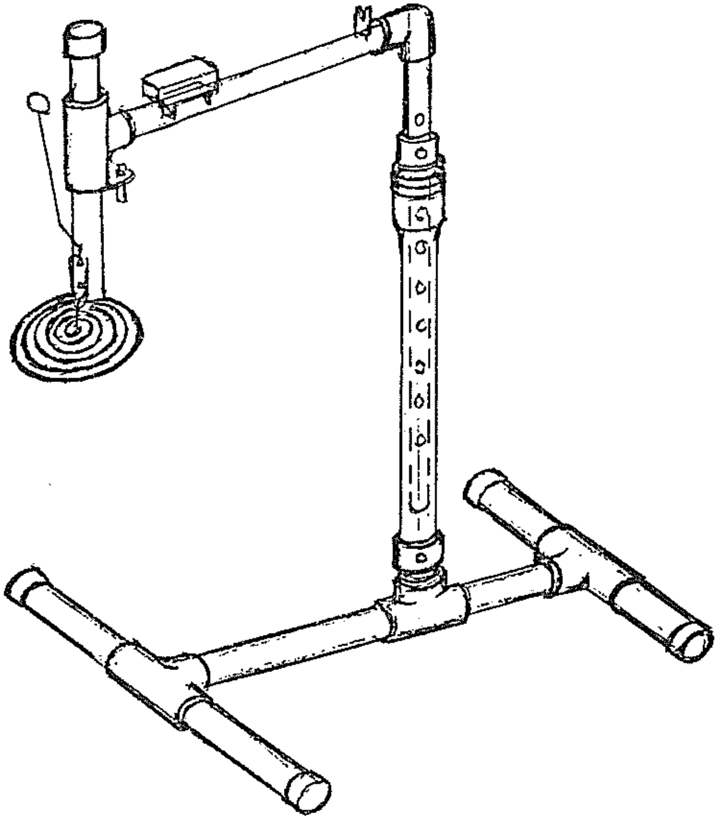

Spiral Coil Multiple Wave Oscillator

ALL COILS ON THIS M.W.O. MUST BE RIGHT HAND COILS, IN RELATION TO THE UNIT AND EACH OTHER. If they are not you will get negative energy. (I'm not speaking of polarity here.)

I use a fish tank air pump (modified) -- (Challenger 1) about $8.50. This provides the oscillatory motion to the coils. I'm not in love with it and would like to find something better -- something of higher frequency, but it works. If anyone has any suggestions on this I would like to hear from them.

I chose to design the unit so it can access any part of the body freely without a lot of excess adjustments. The body when lying down is in one plane or height (approximately), so I built the unit so it would extend over a bed or table, added a swing arm, and a vertical adjustment (holes in the vertical tube arm and a wood dowel or pin stop) -- that seemed simple enough, remember we're trying not to get complicated here.

If you are going to build this M.W.O. I strongly suggest that you read Lakhovsky's book, THE SECRET OF LIFE. When I finally decided to try to build an alternative type of M.W.O., I read everything I could get my hands on about the subject. I read carefully and took notes of key points, functions etc., then I put my books away and studied my notes -- the basics.

M.W.O. -- MULTI-WAVE OSCILLATOR

MULTI -- meaning more than one or many.

WAVE -- energy (given off)

OSCILLATOR -- I looked this up in Webster's -- and simply stated it says -- motion deflected from center to one side -- back to center -- deflected to the opposite side -- back to center, OR -- Vibration.

Now, Lakhovsky in his book stated he made an oscillator of 1 loop or coil of wire errata: This page was inadvertantly left out of the book.

electrical supply house. If you cannot get bare wire, you will have to get plastic coated and strip the coating with a sharp jackknife, taking care not to nick the wire, or yourself.

2. You need a piece of aluminum bar 3/16 x 1-1/2 inces, 20 to 24 inches long.
3. You need a medium size pair of pliers.
4. Soldering iron and acid core solder.
5. Small coil of 19 gauge coated (shellacked) copper wire.

Now – we have 17-1/2 ft. of 10 gauge copper wire, lay it out on a flat clean surface and straighten it out taking the curve out of it and the kinks and bends. Next we need a secure and strong place to fasten the ends of the wire to – like a trailer hitch or bumper of a car or truck etc. After you have found a place to secure your wire, fasten both ends of the wire to it as shown in Fig. 1.

Slowly pull the center of the wire so that you have 2 pieces of wire, put your aluminum bar in the loop and pull tight. With a pair of pliers close the wire down on the bar tight so that it passes between the screws. Pull hard on the wire with the bar – it won't break. Start twisting as shown in Fig. 1 – always keep tension on the wire – do not let go of the bar. (COUNTER CLOCKWISE). Twist the wire, stopping at time, give the wire a couple of good yanks. NO, it won't break. Twist the wire until you have a good tight twist on it. When you think you have twisted it enough – twist it a little more. When you get it where you like it – SLOWLY back the pressure off of the bar.

Remove 1 screw on the bar and slide the loop out or off of the bar. Unfasten the wire where you secured it and cut the spiral wire as close as possible at secured end. Lay the twisted wire down flat. Measure 3" down from the top of the loop as shown in Fig. 2 and mark it with a pencil or pen. Hard or tight bend the wires as shown in Fig 2, 90 degrees. Grip the twisted wire at the bend point with a pair of pliers, and start forming the spiral. Continue forming 1 or 2 loops. Remove the pliers, grip the wire and hand form the spiral as shown in the drawing #2. When you have formed about 4 loops of the spiral, stop and apply light heat to the coil with a hair dryer or similar device to relieve some of the stress in the wire, so that the loops that you have formed will stay where you have bent them. Repeat the operations and continue forming until you have your spiral.

DO NOT BEND END OR TIP IN YET.

Put your spiral in the freezer of your refrigerator for 30 to 45 minutes, and freeze it. This will tend to stabilize the copper some and relieve some more of the stress.

In Fig. 4 we can see that if we bend a piece of metal the inside of the curve is compressed and the outside is stretched, this stress is what we are trying to get rid of. About now some of you will be thinking – GAD – he even tells us how to hold the hammer. I am sure that some of your who are pro's at metal forming and fabrication can build this thing – WHAM-BANG, but please have patience, not all are gifted in some of these skills.

When you have formed your spiral and it is shaped as you want it, hold the loop of the spiral with your thumb and point fingers at eye level height and sight across the spiral, it should hang so that all coils are level and even with each other, so that it is flat – if not form it so it is. Put the coil in the freezer again, repeat freezing operation several times, this tends to age harden the coil some.

With a very hot soldering iron solder the end of the coil, wash it in warm water, dry it, and bend end in as on drawing #2. After you bend the tip in, bend it down a little bit so the bent in tip clears the crossed coil by about 1/4 inch.

Drill 2 small holes in the spiral at points marked on Fig. 6, taking care not to break the drill, use light pressure. Cut 1 piece of coated coil wire (19 ga.) about 4 1/2 feet long. Push one end of wire (about 1-1/2 inches) up through bottom of hole in spiral, and wrap wind (twisted) loops of wire to second hole. Straight wrap 1 loop at hole area, this locks the standing wire. Pull standing wires tight with pliers and cut standing wires to about 1-1/4 inches.

Refreeze spiral and warm to room temperature. Sand or remove shellac from standing wires down about 3/4 of an inch, this makes electrical contact for the coil shocker. You have finished the main spiral coil. You have just done the hardest part, the rest of it is easy. Save your aluminum bar; you may want to build other types and shapes of coils and spirals to experiment with. I have wound and tried 2 wire coils, 3 wire coils, cone shaped spirals, step spirals, spirals with wires attached and hanging down with soldered tips etc. There is no end – you can experiment to your heart's delight, it's only limited by your imagination.

In closing, I would like to tell you a story – a true story about a close relative who gave me permission to tell of his experience with my M.W.O., but prefers to remain anonymous. I will not elaborate, but just state the facts as told to me by him and his nurse.

This relative took a trip by car to the southwest part of the U.S. with some friends. Before he left he called me and said that he would come to California and would stop to visit with me. This was in the late summer of 1990. After arriving at his destination in the southwest, the weather turned very hot for the whole western U.S. Due to this extreme hot spell, he phoned me and said he was returning home and would not come to California.

When I talked to him on the phone before he left, I told him I had built a Multi Wave Oscillator for him and I would give it to him when he came to see me, and he could take it back with him. He did not make it to Calif. but returned to his home in the midwest. After returning home he became ill and went to see his doctor. After being checked by his doctor, he had tests, and the doctors discovered that he had cancer of the large intestine. When I phoned him and found that he had cancer, I asked him if he wanted to try my M.W.O. I told him that I could not guarantee that it would help him – but I said – what do you have to lose. He said yes, he wanted to try it, so I sent it to him.

When he received the M.W.O., it was assembled from instructions I sent him. He used it twice a day for approximately 6 weeks. He had additional medical problems, he had diabetes and congestive heart failure plus the cancer plus a back problem. I had made the M.W.O. for him in hopes that it might help his back. When he went back in 6 weeks to see a cancer specialist doctor they gave him extensive tests and could not find any traces of

of a given diameter that overlapped. Also, I have read that it isn't the intensity or power of the unit that is desired, but the quality of the signal and also the number of frequencies, signals or waves, as the power can be offset by a longer exposure time. It seemed obvious that a spiral should produce all or most of the waves or signals desired and that it would be better than one loop with a limited signal effect. If Lakhovsky could cure cancer in plants and people with one loop it is logical that a spiral should be even better. OSCILLATION -- how about inducing it artificially -- vibrate or agitate the coil or coils.

This unit produces it's peak energy at 12 o'clock at night, but it can be used during the daylight hours too; it's energy diminishes some toward mid-day. Lakhovsky stated this and he is right. I've got to give it to him, he did his homework. Exposure rates or time of treatment with an electronic double screen M.W.O. is said to be 10 to 15 minutes every 3 to 4 days. Exposure time for a spiral coil M.W.O. (rule of thumb) is about 20 to 25 minutes or 45 minutes maximum per day, or less, depending on what is treated.

Lakhovsky also stated that he treated plants, animals and people with belts, collars, bracelets etc., consisting of a single loop of copper wire. From what I can gather, if a subject was wearing a Lakhovsky coil they never took it off.

My spiral coil M.W.O. is half and half -- part 1 loop + of wire, and part electronic. If a subject got 1 exposure every 3 to 4 days with a conventional M.W.O., it means that the subject's system was hit hard or shocked, and it needs time to adjust. A spiral coil M.W.O. requires a longer or more often exposure and this can be an advantage as it gives us more control of the research subject's requirements. The negative side is that the subject has to spend more time more often.

There is one unusual thing that I and others have experienced with this unit. When you first use it the sensation of heat or tingling is light and you require longer exposure time, like about 25 minutes before you feel that you have had enough. After you use it for about 1 to 2 weeks you will find that after 10 to 15 minutes you feel that you have had enough. Simply stated Lakhovsky said that a M.W.O. activates or excites the cells. It's kind of like pushing a car, it takes a lot of energy to get it started, but once it's rolling it's easier to push. Maybe the body just becomes more sensitized to the energy after the M.W.O. is used for a while. In Lakhovsky's THE SECRET OF LIFE on page 75, "The circuit is subjected to any kind of electric or magnetic shock it is then said that it vibrates according to its natural period." So, now what do we have?

1. A spiral coil (Providing Multiple Frequencies)
2. Vibrator (artificial oscillation).
3. Electric current (D.C.) (shock to the coil) provided by a small transformer limited by bulbs turned on and off by a Christmas light blinker or winker.

Now I said previously that most of the unit is made of P.V.C. schedule 40 pipe (white). P.V.C. pipe being what it is has the tendency to sag if placed in a horizontal position, fastened at one end with weight suspended at the other end. We can correct this by inserting a piece of 1 inch diameter .050 aluminum tubing inside the P.V.C. and gluing it with an adhesive or simply by drilling a 1/8 inch hole through both pieces and inserting a #8 sheet metal screw. Remember also I said there were little tricks you can use to build it as given above, so before you get too anxious to build, wait till I can write the HOW TO BUILD article, it may make your job easier and you will not have to change or rebuild the second time. I am trying to save some of you all the mistakes I made.

The heart of this M.W.O. is the large main spiral coil. This coil is in simple terms an antenna. It is a dual antenna, that is, 2 in 1 as it is a receiving antenna and also a broadcasting antenna. The antenna can be shaped in many forms and still retain the right hand spiral. Do we wind it tight toward the center and increase the spacing as we go out? This configuration I have found best, but other configurations can be formed to achieve different effects. I have experimented with different shapes and some have given unusual results. There is still a lot of work that could be done here also. I wound a 3 wire coil that was very effective but I had problems twisting the copper wire and lost several of them. I would suggest that the first coil you make be a 2 wire twist coil. If you want to build the M.W.O. complete you will have a total of 4 coils.

1 large main spiral coil.
1 small spiral coil (that is mounted under the large coil).
2 straight tube coils.

We use these coils because they all have independent functions. We are going to squeeze it and milk it for all we can, to get everything we can out of it. The main spiral coil produces most of the waves or energy, however, due to its design and nature the inside of the spiral cannot be wound tight to create small loops in the coil, so we will add another smaller spiral coil under the large main-spiral, this will give us the higher waves or frequencies we desire. We drill 2 small holes in the main spiral and tight wind a piece of 19 gauge coated coil wire around the twisted copper wire, and pull the wire up through the holes, leaving 2 standing 19 gauge wires. We sand the ends of these and fasten spring clips that hold the tube coils. The other end of the tube coil wire is connected to a banana plug -- into a receptacle, we connect our coil shock to these.

I know there are some of you out there who are like I am, (impatient) and are curious and would like to try this to see if this guy is telling the truth or if he is just full of it. I am aware that the Journal is published every 2 months and maybe some of you don't want to wait 2 months, so for those of you who are like I am -- let me tell you how to build the main spiral coil. This one coil in and of itself won't knock you over -- but you can get started, and you can play with it until I can write the HOW TO BUILD article. This coil will emit energy by itself if suspended horizontally. Fasten it to something, a piece of plastic tube or wood dowel etc. Try it on your feet as the feet seem to be sensitive to these energies. Well -- here we go.

HOW TO BUILD THE MAIN SPIRAL COIL

You will need to purchase a few things.

1. You need 17-1/2 feet of 10 gauge (solid) bare copper wire (cost about 14 cents per foot). You can get this at any

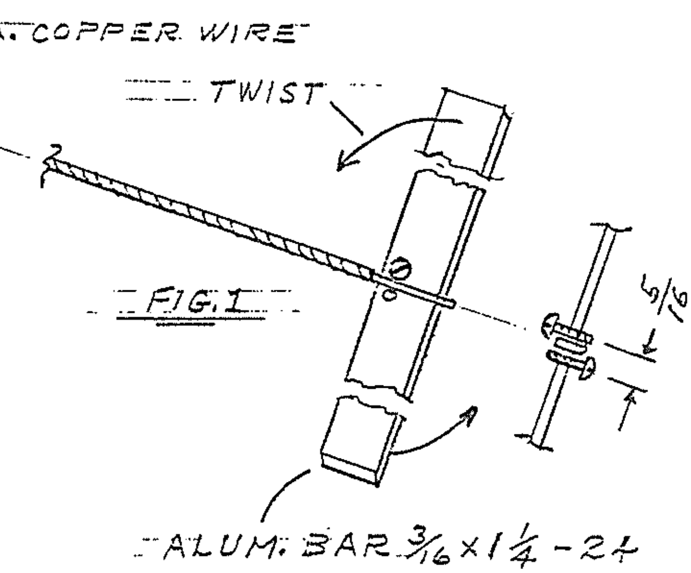
10 GA. COPPER WIRE
SECURE
TWIST
5/16
ALUM. BAR 3/16 x 1 1/4 - 24

FIG. 1

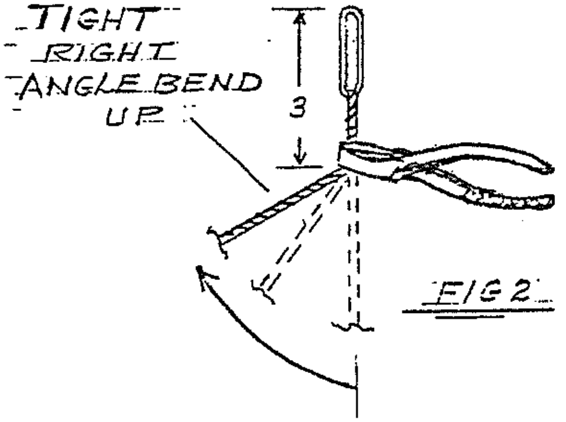
TIGHT RIGHT ANGLE BEND UP
3

FIG 2

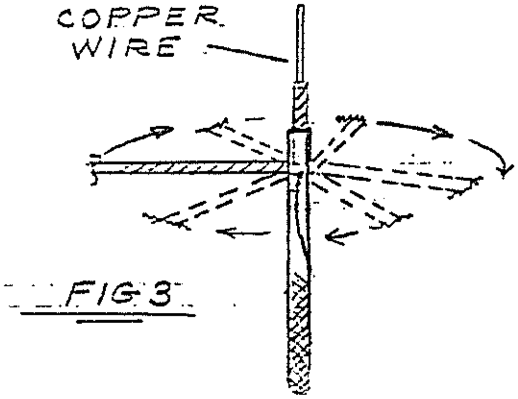
COPPER WIRE

FIG 3

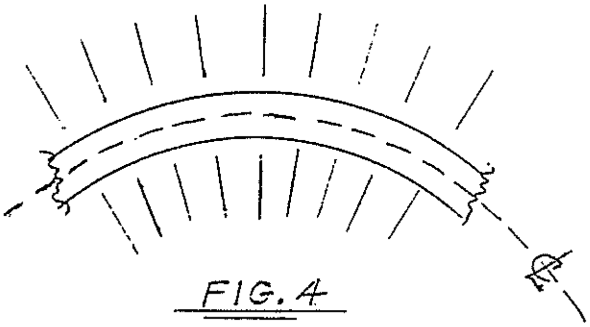

FIG. 4

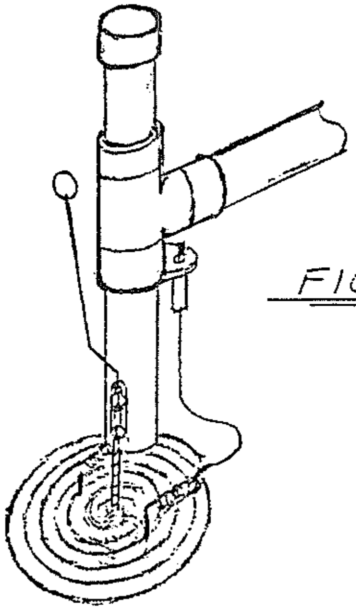

FIG. 5

Spiral Coil Multiple Wave
Oscillator Antenna
Construction Drawings
Figures 1-5

cancer. After his tests he phoned me and said the doctors could find NO traces of cancer whatsoever. I asked him -- did you take any medications whatsoever for your cancer? -- he said NO. I asked him -- did you take any radiation treatments for you cancer? -- he said NO. I asked him -- did you take anything at all for your cancer? -- he said NO. To this date he has not had any return of his cancer.

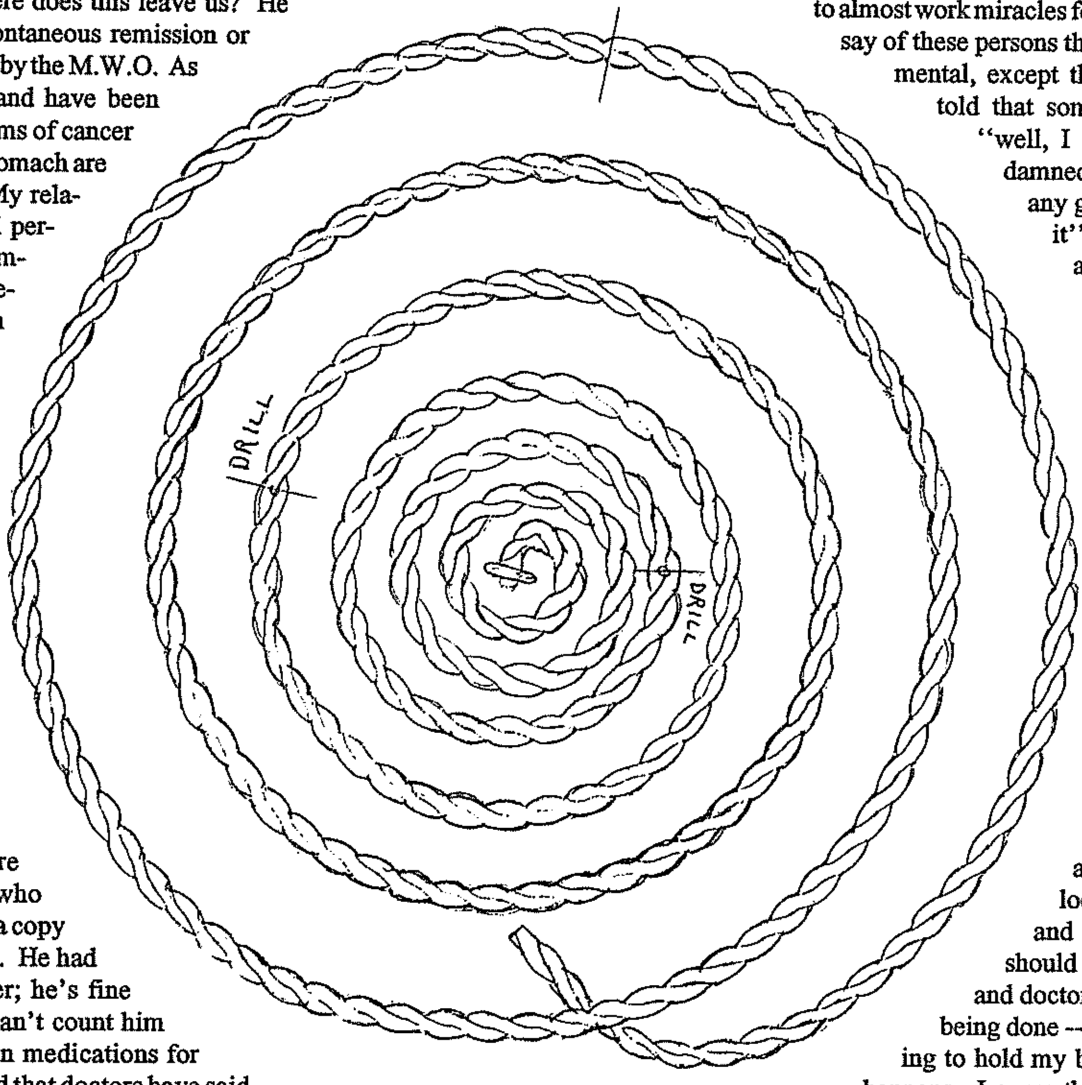

FIGURE 6

Now, where does this leave us? He either had spontaneous remission or he was helped by the M.W.O. As I understand and have been told, some forms of cancer in the upper stomach are inoperable. My relative is the 3rd person who seemingly has had remission from cancer from using my M.W.O. I do not like to use the word cured when speaking of cancer, and I say seemingly as this is all in the experimental stages of research.

Then there is Al Mills who built and used a copy of my M.W.O. He had prostate cancer; he's fine now, but we can't count him as he has taken medications for it. I have heard that doctors have said that prostate cancer is very difficult to treat. Then there also is Simon Lucas who built my M.W.O. and loaned it to a lady who has LUPUS (a cutaneous disease due to the tubercle bacillus). He said she is feeling and doing much better now. I also built a M.W.O. for my mother who lives in So. Sioux City, Nebraska, across the river from Sioux City, Iowa. I sent it to her -- she could not assemble it so I flew back to do it for her. Several years back she was leaving a market and caught her foot in a disposed plastic bag on the sidewalk, fell and broke her hip and leg. The doctors repaired it by wrapping screen around the bone and added a few screws and pins here and there. She was OK for several years and then she got to the point where the pain was so bad she could walk only a short distance. She has used the M.W.O. for about 2 months. She called me the other day and said she took the bus all over town, did shopping, came back, took her pull cart and did her food shopping -- 6 blocks up and 6 blocks back -- and then walked up to the restaurant and got her dinner, all in 1 day and she said she feels fine. The doctors gave her the choice of pain shots or cutting off her leg -- (some choice). NOTE -- She does not have a coil shocker on her unit.

In all honesty I have to say that my spiral coil M.W.O. does not affect all people the same. I have had no negative or bad reports to date from anyone using it with good common sense and discretion. There are a few people that it does not seem to help to any great degree. There are also some people that it seems to almost work miracles for them. I would say of these persons that maybe it was mental, except that I have been told that some have said -- "well, I don't think the damned thing will do any good, but I'll try it". I could go on and on but we'll leave it at that.

If you decide to build my M.W.O. I would welcome any serious and honest experimental input as this unit still has many areas that need to be explored. Sometimes I back way off and just stop and look at all this -- and realize that it should be the scientists and doctors doing what is being done -- but I'm not going to hold my breath until that happens. I guess the only way it's going to get done is for people like you and me getting off our backsides and getting it done. If enough will work at it and add a little bit, it'll get done. I sometimes wonder what the end-run potential of my M.W.O. is. I love the doctors (God Bless them!), we need them, but I wish they were more open minded and that they were not so profit orientated and stigmatized by their organizations. I think these ideas I present here will help.

THE SPIRAL COIL M.W.O.

PART II - HOW TO BUILD

by Lou Schad

In Part One I explained how the spiral coil M.W.O. evolved and how I came to build it and notes on things I have learned about it and its use. I also explained with drawings how to build the main spiral coil. For those of you who wish to build this unit, I have added additional drawings and notes so you can build the rest of the coils, base, etc.

Before I begin I would like to address an issue that has bothered me for some time, it pertains to my M.W.O. and maybe other M.W.O's in general. It is a situation or function that has been ignored by many or all but a few. In my article in Part I, I said all coils on this unit must be right hand coils. What I am referring to is what I call RIGHT HAND SPIN. In my study of Lakhovsky's and others work, I have not been able to find any information pertaining to direction of spin, although he gave very precise directions -- how collars, belts, bracelets etc. were to be made. Did he know and was it his secret? Was he not telling? Lakhovsky stated that his M.W.O. stimulated the twisted filament or nucleus of the cell, and the cell oscillated according to its natural frequency. A weak oscillating cell would increase its oscillations back to normal. The question is, do certain oscillations cause the cell to spin right hand or left hand, or clockwise or counter-clockwise, if the frequencies spin right hand or left hand? Some might say -- what difference does it make. The people who use or practice the Five Tibetan Rites always spin right -- never left. Sir Lawrence Bragg (Nobel Prize winner) discovered that healthy people have "RIGHT HAND" spinning blood. And physicist Dr. E. Mersmann discovered that people exposed to electronic smog have "LEFT HAND" spinning blood, the same as people who have cancer -- they also have "LEFT HAND" spin blood. Maybe if the earth spun in the opposite direction, left hand spin blood would be normal. So, is left hand spin or right hand spin important? Yes, I think so. I have purposely made and used left hand coils and they made me feel sick and ill. There is an instrument that has been built that emits weak pulsed electromagnetic waves that causes the blood to spin right hand in people who have left hand spin blood, however if the unit is removed the blood will go back to abnormal left hand spin.

If you put your hand in a bucket of water you can spin it to the right or to the left -- is this polarity? Does magnetic polarity have an effect on matter causing right or left spin? Yes, it seems to. But how about induced spin direction on a non-magnetic or non-electrical energy or wave? If I hold a magnetic compass near the spiral coil of my M.W.O., I get nothing -- no needle movement. If I hold a pocket radio under the coil with vibrator and tune it off station, I get static on A.M. from about 80 on down, none above 90 and up, but I don't think that static is coming from the vibrator. Warning: I damaged my radio by using if for testing. At first I thought it was the battery, but a new battery did not correct it. Sometimes when I turn it on now it doesn't want to come on, by flipping the switch on and off several times it then works. I had no problem with this radio before I used it for testing.

Whatever the answers are, I would suggest you use right hand coils. Look very carefully on pages 113 and 114, these are right hand coils in relation to the subject, (ref. The Secret of Life). If you build the spiral coil M.W.O. you will have the basic unit. You can experiment and try all types of coils and spirals, as I said in Part I, it is only limited by your imagination. I find it fun and interesting to try unusual and weird coils, it's not expensive and you only have to remove a couple of screws, no big deal.

I have been thinking of building a cheap copy of Lakhovsky's coil of tube and wire, also a foil coil of equal size, to test if the coils when vibrated emit equal energies etc., and what the differences would be. It may answer the question why some of the foil coils don't seem to produce the results as good as some would like. In Lakhovsky's book there are pictures of, and he speaks of, the oscillator coils of Hertz. There is a striking resemblance of Lakhovsky's coils on his oscillator antenna to Hertz's coil. Could he have been influenced by or borrowed the Hertz oscillator coil for his unit -- did he expand upon the base of Hertz's work for his antenna coil? (interesting thought)

Enough digression -- back to our spiral coil building project. If you look at drawing #3 you will see that it is self descriptive, giving diameters, lengths etc. We make most all of our unit out of schedule 40 PVC plastic pipe. Take enough care to make your cuts square and clean, debur the ends of the tubes so they will slide into the glued slip fittings correctly. Now, here comes the stinky part. If you want your unit to look nice, you will have to clean the tube, (after cutting), taking off all the dirt, the ink lettering etc. I use a laundry tub, a small amount of water, laundry detergent, Comet, and some wet or dry #200 sandpaper and a lot of elbow grease.

Warning: PVC glue sets fast, in a matter of seconds. I put glue on the female fittings only -- none on the male side, slide it together firmly and whack it with a short piece of 2 x 4. Glue part #7 to part #8 and drill a 5/16 hole through fitting -- this is so you can tighten or remove the threaded fitting by inserting a phillips screw driver or rod in the hole. This makes your unit portable.

Glue pieces 1-1-5-4-4, also 3-3-5-4-4.

Now glue part 2 to part 6, also part 3 to part 6.

Thread assembly 7-8 (large vertical tube) into assemble 2-6-3.

We will slide assembly 11-5-44 onto tube #2, but we will not glue it. Place this on a flat table, and adjust tube 8 so it is vertical.

When you get it where you like it, glue assembly 33-5-44 to loose tube 3, let this dry for 20 minutes.

Now remove assembly 11-5-44, apply glue to fitting #5 and glue it to tube #2 on your flat table. You now have the base finished.

Glue the rest of the tubes and fittings together on your table. NOTE: Do not glue parts #19 and #20, we leave these loose and just set them in place, as we may add something additional here later.

NOTE: Before you glue tubes #11 and #15, cut some one inch diameter .050 aluminum tube full length of tubes and slide them inside of plastic tubes, this eliminates sag. Drill a 1/8 inch hole through plastic tube and aluminum tube, screw a #8 sheet metal screw through both tubes. Tube #11 has holes in it so we can adjust the height, drill these 2 inches apart, 5/16 inch holes. We insert a 1/4 inch wood dowel through fitting #10 and through tube #11.

We drill 2 holes in tube #21 and thread them 10-24, we fasten our spiral coils and antenna to these. The first hole is drilled 5/16 of an inch up from the bottom edge of the tube and the second hole is drilled 3/4 of an inch from the first hole.

When you buy your 1-1/2 PVC tube, get a piece with a flared end, as we set a reducer fitting in it to make a sleeve for the #11 one inch tube. To keep tube #11 guided at the bottom of tube #8, drill 3 holes at the bottom of tube #11 equally spaced around the diameter and insert 3 #8 sheet metal screws.

Parts #9 and #10 will have to be filed out or machined out so the one inch tube will slide through them.

Part #16, the vibrator base, is made of flat stock PVC. If you cannot buy it from a plastic supply house, cut a couple pieces of 1-1/2 PVC tube about 3-1/2 inches long, cut them again length wise, so you have half tube pieces. Place the half tube pieces between two - 1/8 inch aluminum plates and heat in a stove oven with a brick or weight on top. Let them cool and the plastic will come out flat. Layout and mark parts on flat sheet and cut out.

Glue all the pieces of vibrator base together and fit to tube #15 and glue it tight. A wood match works wood to lay glue in joints.

Make and fit your metal brace from aluminum or non-corrosive metal and fasted with two - #8 sheet metal screws. Cut out part #23, fit it to T - #22, drill two holes for banana receptacles and glue securely, also cord holder #14.

Make your small spiral coil from .0625 low carbon stainless steel heli-arc welding rod -- 36 inches long, do not twist wire, use as is. You can get this from any welding supply; get 4 or 5 pieces. Also make antenna from same wire. To make small spiral use the full length or rod. Start by forming an eye on one end for the 10-24 screw. Move down 3 1/4 inches and form spiral same as large spiral coil, also bending end tip in. Put an offset kink in the stem so it will pass through the main coil.

The tube coils are made from #19 shellac insulated coil wire, 12 to 14 turns. I tight bend a 90 degree bend, so I have a 1 inch long tail. I place this parallel on a pencil or wood dowel and tight wind the coil. Grip the wire tail and dowel with a pair of pliers. Remove the dowel and freeze for 30 minutes, warm to room temperature. Grip the ends of the coil with your thumb and middle finger nails and stretch the coil to the desired length, refreeze for 30 minutes. Bend long tail of coil 90 degrees and fit to banana plug. Trim short end to 3/4 inch, clean and solder pinch clip to coil. The pinch clip is made from light gauge hard stainless or brass. Blank size is 7/32 x 1-1/4.

After you have everything glued and finished put a coat of auto polish on PVC tubes, as this keeps them clean and they will not finger mark or get dirty. Also put a coat of vaseline on the threaded fittings so they can be removed easily as they tend to stick.

If you are, or want to be, the serious experimenter type, buy a note book and keep records and notes. Keeping a log with dates, exposure times, etc. can be a very useful tool to compare results of different coils etc. I would welcome hearing from you on the use of the spiral coil M.W.O., i.e. your success and problems. NOTE: Don't expect results in 2 or 3 exposures, as this unit is more gentle and may require a longer time of weeks or several months to get permanent results. To this date I have not had a single report from anyone for over-exposure if used with common discretion. I have used belts, collars, wire loops etc. with the M.W.O., this seems to multiply the energy. I also have used a spin magnetic unit with a reticular energy filter before and after M.W.O. exposure. This seems to work well for joint problems.

Please Note: Lou Schad does not manufacture or sell this MWO. He is giving his knowledge and experience as a gift to the world. He may be written c/- BSRF.

LEGEND

USE P.V.C. (WHITE) SCHEDULE 40

KEY

L=LENGTH D=DIAMETER C=CUT
T=TUBE T.F.=THREADED FITTING

PART #	**DESCRIPTION** (in inches)
1.	T-D=1-1/2, CL=10
2.	T-D=1-1/2, CL=15-1/2
3.	T-D=1-1/2, CL=8-1/4
4.	CAP-D=1-1/2 (4)
5.	TEE-D=1-1/2 (2)
6.	TEE-D=1-1/2 THD'D
7.	NIPPLE-D=1-1/2 THD'D
8.	T-D=1-1/2, CL=28-1/2
9.	REDUCER - D=1-1/2 TO 1
10.	REDUCER - D=1-1/2 TO 1 INV.
11.	T-D=1, CL=28-1/2
12.	T.F.-D=1 NIPPLE
13.	ELBOW - D=1 - T.F.
14.	CORD HOLDER (ELEC.)
15.	T-D=1, CL=20-1/2
16.	VIBRATOR BASE
17.	METAL BRACE
18.	TEE-D=1
19.	T-D=1, CL=3-1/4
20.	CAP - D=1
21.	T-D=1, CL=5-1/2
22.	ANTENNA
23.	F.T.G. BANANA RECP.

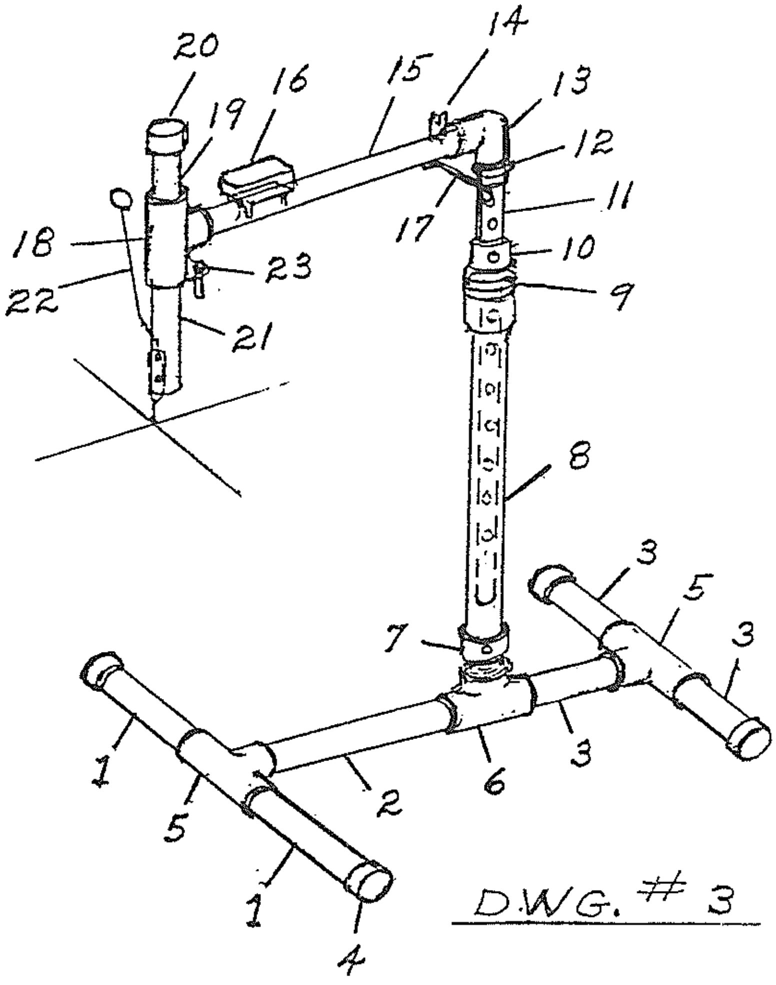

D.W.G. # 3

COIL SHOCKER FOR SCHAD SPIRAL COIL M.W.O.

This unit provides a D.C. shock to the spiral coil. It should be built in a wood box, or grounded metal box for safety sake as we use 115V A.C. power source. All the parts can be purchased from an electrical supply, except auto bulb and socket. Get these from an auto parts supply house.

The Leviton dimmer is not necessary, but it gives more control. I use a ceramic base for 40 watt bulb. A smaller transformer can be used, but I find the 3A-12V best. It is nice to add a handle to the box. I dropped mine once. Put ventilation holes in the box, and rubber bumper feet on the bottom.

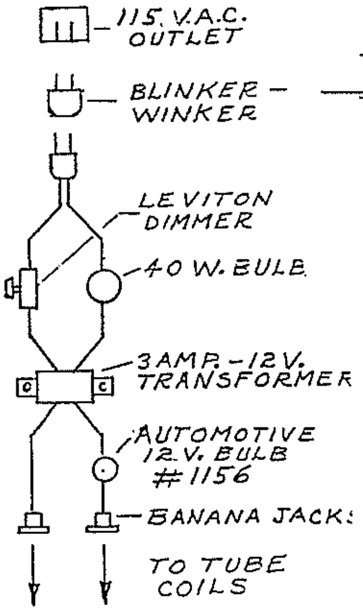
115 V.A.C. OUTLET
BLINKER - WINKER
LEVITON DIMMER
40 W. BULB
3 AMP. - 12 V. TRANSFORMER
AUTOMOTIVE 12 V. BULB #1156
BANANA JACKS
TO TUBE COILS

POOR MANS COIL SHOCKER

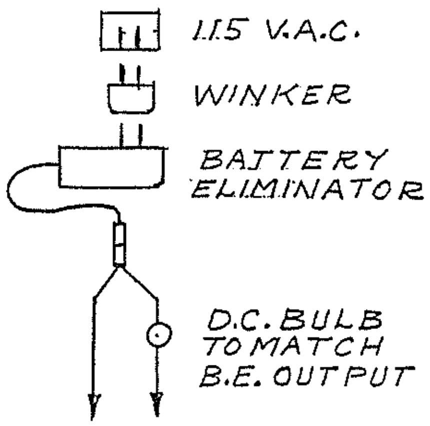
115 V.A.C.
WINKER
BATTERY ELIMINATOR
D.C. BULB TO MATCH B.E. OUTPUT

DWG. # 4

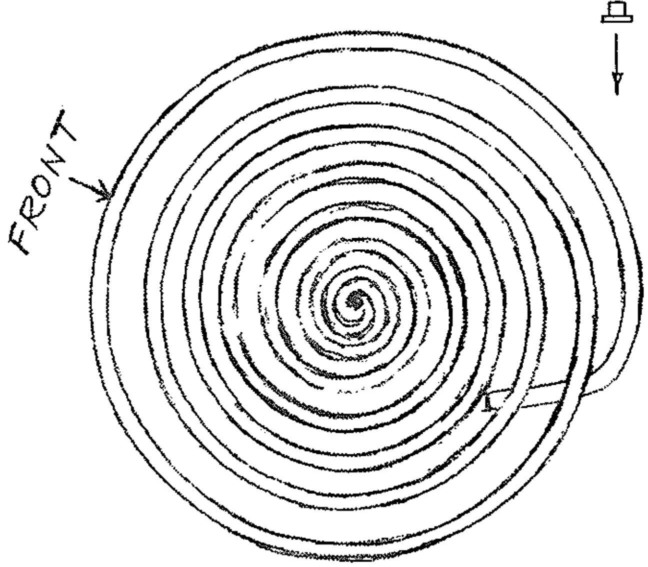

TRACING OF SMALL SPIRAL

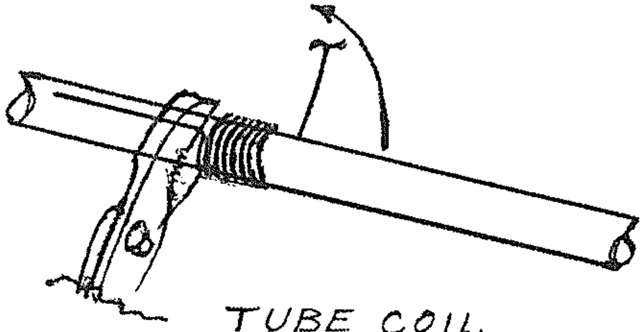

TUBE COIL

#19 COIL WIRE 2 REQ.

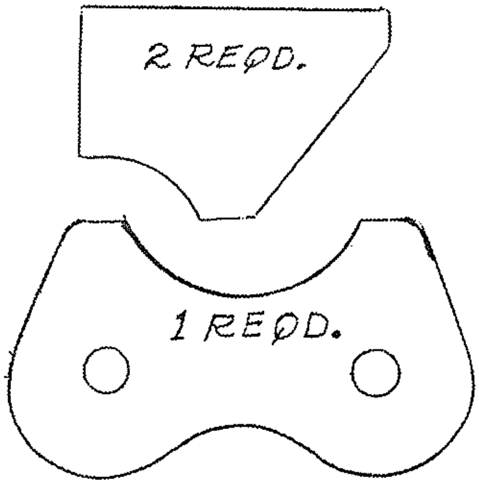
2 REQD.
1 REQD.

DWG. # 5

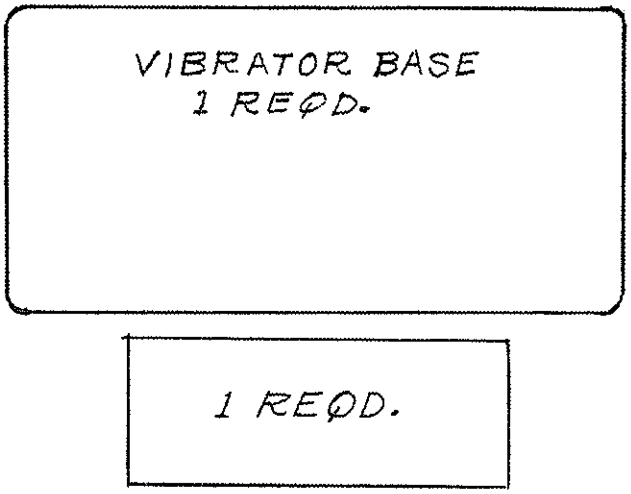
VIBRATOR BASE 1 REQD.
1 REQD.

SPIRAL COIL M.W.O. UPDATE

Lou Schad

This is an update on the work I have done on the spiral coil MWO. I have built and tested several new coils and made changes in design of shapes and twist of the wire and different size of wires etc. I have experimented on plants and myself for the past several months.

In my article in the Journal of Borderland Research of March-April 1991, page 15, figure 1, I stated, to twist the wire counter clockwise, as I was told by someone of authority that this was best. I now find after experimenting and testing that this is not so. I now twist the wire clockwise, causing a left hand spin of the wire when viewed -- or counter clockwise. If the wire is twisted in this manner it more than doubles the energy and the plants and I both seem to like it better. If the wire is twisted as stated it will produce energy not only in a spiral shape but also in single loop rings (overlapped) and also in just a straight length of twisted wire.

I made a coil like a target or a series of one loop rings of twisted wire of different diameters and wire sizes, from about 1 inch to 8 inch and overlapped them about ¾ inch and wired them to a X shaped piece of thin plywood, and it produces more energy than I need. I can stay under it only about 7 to 10 minutes. If I stay under it longer than this, I get over exposure or too much energy in my body, and it makes my feet and hands tingle.

I made a spiral coil from 1/8 inch copper tubing, twisting it like a solid wire, and it produces more energy than a like spiral made from solid wire. I do not know if the energy travels on the inside of the tube and the outside also, but the energy produced is much more than is required for normal use.

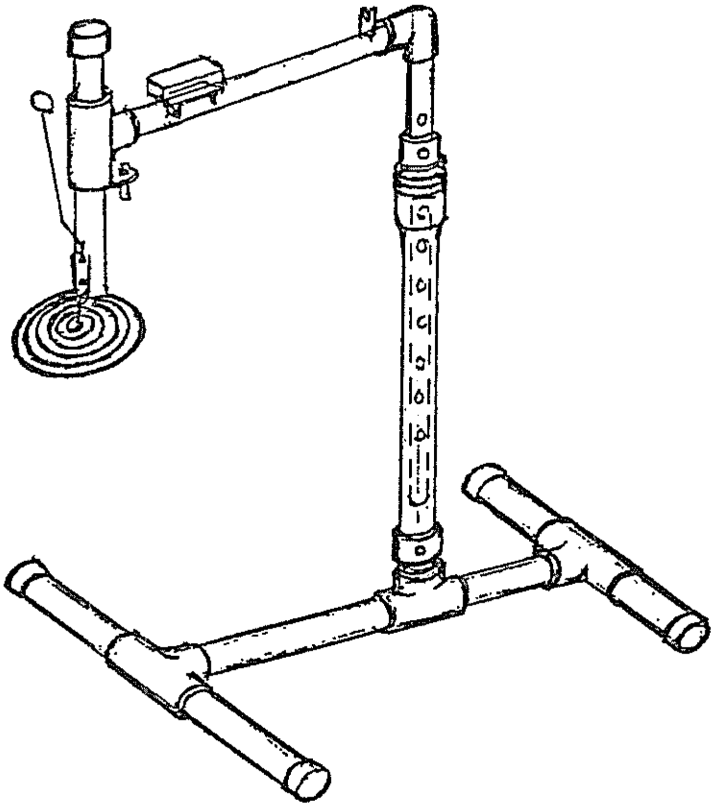

Spiral Coil Multiple Wave Oscillator

I also have been experimenting with new and different ways to pump or project the energy. In my original design I used a vibrator to increase the output. I now find that with the new coils (twisted as stated) that they develop as much or more energy without the vibrator as the old coils did with the vibrator. It seems that the field or energy pattern given off by a straight (twisted) wire is different than those of a single loop (twisted wire) shape or spiral shape.

I am building a new coil that uses both and looks like a spider with 8 legs of different length. It has 2 single loops (overlapped). One loop is 2-3/8 in diameter and the other is 4-1/8 in diameter. The small loop is centered in the large loop and in the same plane. I made a wood jig with nails for pins to hold them in place and soldered twisted straight lengths of wire pointing from the center outward. I will bend the tips of the straight twisted wires down 90° producing a shape like a spider. I'll let you know if it is different from the spiral coil shape in energy output.

It now becomes obvious that a spiral is not the only configuration that will generate or collect and broadcast energy of unknown origin. If other shapes or designs will also produce energy, is their output different? If so, what are the differences? Are the wave lengths different, are there different patterns to the waves or energy? How does it affect the living organism?

When a spiral or coil is placed in front of a reticular energy field it produces unusual results. Coils can be combined with many other forms of energy, magnetics, light, audio, vibration etc.

In my article I said you can design and build coils of all shapes and types, as it is only limited by your imagination. It is my personal belief that someday someone will design and develop a coil of a nature that will be highly beneficial to humanity, we need your input, ideas and help.

If a coil of twisted wire was built in a target configuration of a series of overlapping loops so as to be similar in design to a Lakhovsky coil, and was powered by a Beck MWO output, what would be the results? If a twisted wire coil will produce and project energy, connecting it to the field of a Beck type unit should be interesting. Are there any takers?

ELECTROSTATIC M.W.O.

INTRODUCTION TO A NEW PHASE OF RESEARCH

by Peter A. Lindemann

Lakhovsky's multiple wave oscillator has always intrigued me. I tried to build my first one in 1976 based on Bob Beck's schematics. These early experiments failed miserably, but I remained interested. In 1987, I had the opportunity to see one of Bob Beck's MWOs at a meeting of the Los Angeles Psychotronics Association. I am a healer and a sensitive by nature, and I was looking forward to finally seeing the mythical machine in action. Bob had said that is was a very high power model and he was right about that.

The unit was of the design that has been popular since the 60s; a high voltage spark-gap transmitter driving an antenna of concentric rings. Bob turned it on and ran it for about three minutes. It made a loud zapping noise as hundred of sparks jumped from ring to ring on the antenna.

I was standing in the back of the room, about 30 feet from Bob's MWO. As soon as he turned it on I felt a blast of irritation hit my aura. I could handle the etheric frying of my energy field for only about thirty seconds before I had to leave the room. When I came back in, everyone's energy had been fried. The meeting quickly digressed into a break, with everyone talking to their neighbor. It was quite a sensation!

In thirty seconds I had become convinced of two things. First, this MWO was extremely damaging to the subtle energy field and therefore must be incapable of promoting systemic healing. And second, this MWO must be unrelated to what Georges Lakhovsky was reporting in his book Secret of Life.

Since I drew those conclusions back in 1987, I have been confronted with numerous testimonials of people gaining benefit from this style of MWO. While I do not doubt the accuracy of these stories, I must admit, I do not understand by what process the benefit is derived.

Here is where the MWO languished in my mind until Lou Schad's articles appeared in the March/April and May/June 1991 issues of this journal. Intuitively I felt Lou was on to something very important. On a visit to Borderland Headquarters, I had a chance to evaluate a unit Lou had given to Tom. The energy coming from the antenna was gentle but very penetrating. I was excited. I went home with the intention of building one of Lou's units to experiment with.

After considering Lou's design carefully, I decided to make a few changes and by late May 1991 I had my first unit. As it turned out, my unit did not look at all like Lou's, but it did embody the fundamental principles that he innovated. These included a spiral main antenna that was charged with static electricity and physically vibrated. Lou's antenna was complicated to build. I decided to use an antenna design that I had been developing uses for for three years, a spiral based on the phi ratio that I could have printed right onto a circuit board. That solved the problem of antenna fabrication. Another thing I changed was the method to induce the physical vibrations. Lou had used an aquarium pump. It worked well but was limited to only one rate of vibration. I decided that the rate of vibration might be important to look at so my unit was made with a variable rate of vibration from 2½ vibrations per second up to about 14 vibrations per second. The last change I made was the vertical antenna. Not knowing the height I needed for the optimum effect, I simply used a retractable replacement antenna from Radio Shack, the kind you usually see on portable radios. This gave me a way to vary the electrostatic charge on the spiral, by raising or lowering the antenna.

By early June 1991 I had made an arrangement with a local acupuncturist in town to run some clinical trials with the unit. This medical professional has asked to remain anonymous and so for the purposes of this article, I will refer to this person as "John". John is a gifted sensitive and an advanced bioenergetic diagnostician using equipment from Germany called the VegaTest. John also has had assistance in this evaluation by another gifted sensitive who can see auras in vivid color and detail. I will refer to this second person as "Pam". As of this writing in early November, their evaluation has been on-going for 4½ months or since mid-June. The results are preliminary. No cures are claimed. However, all indications point to systemic detoxification, energizing and rejuvenation, just what you would expect to see from a broadly stimulated natural healing process.

The VegaTest procedure is very sophisticated. It uses a bio-feedback mechanism to establish baselines for toxicity level, systemic stress and specific organ stress as well as the presence and location of specific pathogens and types of toxic substances. Toxicity level is measured on a scale of 0 to 21, where 21 is the most toxic. A "Biological Score" is developed from the toxicity values that come up and can have from 1 to 6 numbers ranging from 1 to 21, 6 numbers being the most toxic. An "Ideal Score" is determined on a scale from 1 to 15 where normal for adults on first testing range from 7 to 11. Stress on the immune system and endocrine system is measured on a scale from 0 to 5.

CASE A is a 21-year old female who has never been treated with other remedies. The patient was VegaTested, then received one 20 minute session with the MWO and then was retested. (See box insert) These scores suggest that even one session can temporarily reduce stress levels significantly and promote detoxification.

SCORES BEFORE SESSION

Ideal Score: 8
Biological Score: 19-16-14-12-9
5 scores with high numbers is considered very toxic.
Immune System Stress: 3
Endocrine System Stress: 5

SCORES AFTER SESSION

Ideal Score: 6
Biological Score: 12-10-8
Immune System Stress: 1
Endocrine System Stress: 0

CASE B is a 72-year old male with extreme intraocular pressure. The patient received only three sessions with the MWO. The pressure within the eyes lessened with each session. Pam reported that before the MWO sessions, the region of the eyes and head was completely dark and devoid of life force. After the first session, the darkness cleared about 40%. With the second session, the darkness cleared about 80% and with the third session, the energy around the eyes was completely vibrant.

Without going into a lot of specific case histories, or naming disease conditions treated, John and Pam are consistently seeing the following:

1. The MWO expands the energy field around the body 3 to 6 feet.
2. It functions to reduce accumulated stress on many levels.
3. It clears the energetic imprints of illnesses.
4. It fills in breaks and cracks in the aura.
5. It builds vitality at a very deep level and combats fatigue.
6. It promotes the drainage of stale and stagnant life energy down the arms and legs and out the hands and feet. Several times

Pam has seen dark yellow or dark green energy drain out of acupuncture points or organs, to be replaced by cleaner and clearer colors.

7. Everyone experienced improvement, regardless of their condition.

METHOD OF USE FOR THE MWO, (developed by John and Pam during this testing period) is as follows:

1. A person using the MWO should sit in a comfortable chair with their feet flat on the floor and their hands apart, laying on their thighs, palms up is best. The chair should be about 6 feet from the MWO. During the session they should periodically rotate their wrists and ankles to help promote the draining.
2. The MWO should be started at slow vibration rates (2-3) and slowly moved to higher vibration rates (5-9) and then at the end, moved to slower ones again. Sessions can be from 10 to 25 minutes.
3. After the MWO is turned off, the person should remain seated for 10-15 minutes to continue to drain while the energy field slowly shrinks back from its highly expanded state.
4. Rooms where the MWO is used may need to be cleared after its use due to the stale energies that drain out. Sun light, fresh air, toning, sage or other means may need to be employed to keep the MWO session room clear.

THE BSRF MWO PROJECT continues to develop, as this and other articles in this issue clearly show. A special thanks to Lou Schad for his brilliant and original departure from the errors of the past and showing us the true Borderland spirit in action.

One final note. We at Borderland are constantly railing against people who invoke the name of some famous researcher, like Rife, Tesla or Lakhovsky, and then sell the public a piece of equipment unrelated to the original work. This is no better than bait-and-switch hucksterism and we deplore it. Therefore, let us state for the record from the outset, that this new Electrostatic MWO is not what Georges Lakhovsky was doing. Neither is it a direct attempt to duplicate his work. Lakhovsky claimed to have had his units in use around the world, but after years of research we have found no physical evidence of this. He never published detailed schematics of any complete devices. This new MWO is a modern attempt to address the same problems that Lakhovsky was grappling with. We call it a multiple wave oscillator because it is, clearly, a device that produces a wide spectrum of oscillatory emissions simultaneously. Lakhovsky pioneered this line of research. This unit is not a "Lakhovsky MWO" but, strictly speaking, no other unit sold in the last 30 years has been either.

As this MWO is part of an ongoing research project, a limited number of these units will be produced and made available to the public.

The new BSRF Electrostatic MWO is for investigational purposes only. It is not a substitute for professional medical care and is not indicated as a specific cure for any disease condition. Any such use goes directly against the intent of the manufacturer. The performance characteristics of this device have not been established.

MULTIPLE WAVE OSCILLATOR

Borderland Sciences Research Foundation is proud to announce that their new Electrostatic Multiple Wave Oscillator is now available for investigational purposes.

TOTALLY NEW APPROACH

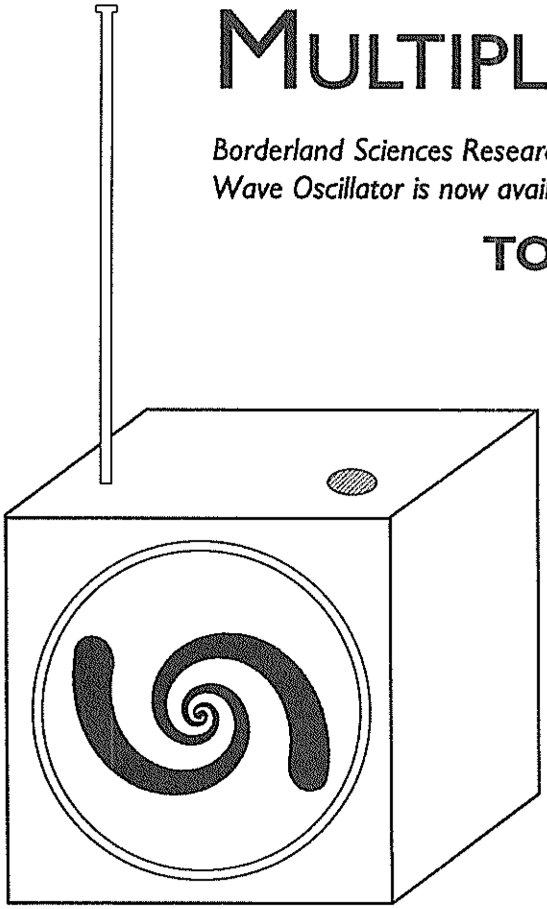

This unique Multiple Wave Oscillator uses a specially designed spiral coil that is both physically vibrated and simultaneously charged with a safe quantity of static potential, unlike designs using high voltage discharges between concentric rings. The result is a wide-spectrum wavefront of electromagnetic and etheric vibrations. It also provides a means to introduce a wide variety of other subtle influences into the broadcast much like a substance biocircuit. The Electrostatic MWO from Borderland Sciences is a versatile investigational tool.

The MWO has a wide range of vibratory rates and charge intensities. It is a one square foot cube with a Golden Mean Spiral MWO antenna on the front and a retractable antenna on the top for tuning. Powered with standard AC. Constructed using no toxic materials.

ELECTROSTATIC MWO

The Electrostatic MWO is a high quality device produced in limited quantity for our research network. Interested researchers may inquire with BSRF as to current availability of units.

SUN-ETHER DISC

By Albert Zock

The German inventor Oscar Korschelt based his cosmic ether accumulator, German Patent #69340, dated July 14, 1891, on the teaching of Wilhelm Weber not realizing that it incorporated the implosion principle as well. Weber professed that all particles of the ether, having an electrical charge, are circling around solid objects without touching them, the molecular particles being negative and the ether particles being positive, according to the law of attraction and repulsion. Since a solid body attracts ether particles, Korschelt searched for a way to collect, condense and to rectify them.

On a wooden disc with a hole in its center, Korschelt fastened spirals of copper wire, on both sides, one wound clockwise and the other, as a mirror image, counter clockwise, and connected the ends together through the hole in the center. His assumption was that the collected energy would be condensed in the center and then be beamed off at a right angle to the disc. He termed the side facing the light positive and the other negative. To increase the condensation and lengthen the antenna, Korschelt wound small coils like key rings and assembled them like links in a chain. He was convinced this device would give additional life energy if the beam was directed at the neck of a person in the height where the 18 cranial nerves are leaving the skull, (medulla oblongata). Such a spiral resembles a Nornen coil, which is stretched out and is used to eliminate undesirable emanations.

An application of a Korschelt coil seems to have its merits. In one case a relative of an elderly man suffering an acute pneumonia, being already unconscious and according to his physician, beyond hope, hung a Sun-Ether Disc over his bed. From this moment, his condition improved rapidly. Fourteen days later, he was almost normal again. In another case, a Sun-Ether Disc hangs over the door in the waiting room of a Naturopath. Every one walking in is amazed over its warmth, and the flowers are flourishing like in a green house, even though it is not heated and faces north. When asked for the reason, the owner points smilingly to the disc and gladly explains how it works and how to build one.

On a plywood disc with a diameter of 50 cm (20-3/4 in.) are two coils, one on each side, wound in a mirror image, the front one counter-clockwise and the other clockwise, almost on top of each other. Both are connected through the hole in the center which has a diameter of 10 mm (6/8 in.). The space between each winding is 15 mm (5/8 in.). Aluminum wire is best. To make the key rings, take a 4 mm (3/16 in.) thick wire, fasten it into an electric drill and wind the armature wire around it tightly, by allowing the rotating wire in the drill to do the spinning. Make sure to wind the links for the front counter-clockwise and for the back clockwise. Then separate every fourth winding and cut it to get the links shaped like key rings. Now attach the rings onto one another like a chain. Each link will lengthen the antenna and work like an oscillator. Fasten the chain onto the board, using nails or screws, but to be accurate, give the chain a light tension so it will not slip off. Twist the ends in the hole together, or better, solder them. The counter-clockwise winding for the front applies to people living in the northern hemisphere; in the southern hemisphere it might need to be reversed.

In Korschelt's time, some made "beam-handbags" carried on a shoulder strap. The material had to be framed to it could not move, then the spiral was embroidered and fasted to the inside of the bag, the whole bag measuring about 20 x 20 cm (8 x 8 in.) and the windings being 1 cm (3/8 in.) apart. Such bags were more powerful than the disc.

From Mar-Apr 1988 Journal of Borderland Research

ETHER RADIATING APPARATUS

An Apparatus For Therapeutic Use Without Direct Or Indirect Suggestion
Oscar Korschelt of Leipzig. German Patent #69340, July 14, 1891
Translated from the German by Bodo Capeller, 1986

This apparatus utilizes the living forces of the ethers for healing purposes. The electro dynamic theory of Wilhelm Weber concerning the structure of smallest parts results in the insight that all ether is identical with Webers electric particles, that all solids rotate around it at a certain distance because of attraction of positive to negative ether particles in relation to solids. Through a corresponding arrangement of solids it ought to be possible to condense the diffuse ether and redirect it. A similar arrangement has been utilized in the following described apparatus which has been named by the inventor "Ether Radiating Apparatus".

This apparatus can be used for the purpose of increasing the intensity of the power of living ether. The difference between several versions of this apparatus is that of shape of materials used. In order to get increased power electric current may be used. Since the same appears to consist of positive and negative particles that try to find equilibrium, the amount of liberated ether particles within the wiring of the apparatus and the diameter of the wire formed ether vortex will increase with the spun off ether particles.

In drawings 1 through 5 an ether radiating device is shown. Within the apparatus the arrangement of the metal wiring is formed in a spiral that will cause ether particles to spin in a predetermined direction so that the living force can be utilized in many ways.

The version in drawing #3 shows the wire is formed in spiral configuration on both sides of plate E and is connected in the center. Through this arrangement the ether particles will accumulate in the center of the disk and be thrown forward. In order to achieve a longer path for the ether vortex one may utilize a chain instead of a simple wire which consists of horizontal displaced links (drawing #5). The disk is intended to cause physiological changes in humans (see drawing #1).

Patent request: An apparatus for therapeutic purposes with direct or indirect suggestion consisting of a non conductive disk (E), free standing or installed in a tube, where both sides of the disk contain spiral wound wire with the side toward the light exposed as the collector and the opposite the radiating portion.

Fig. 1

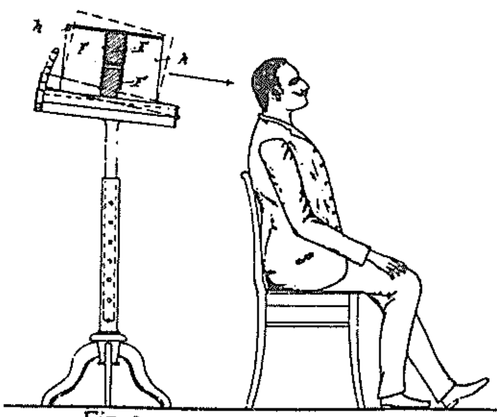

Fig. 2.

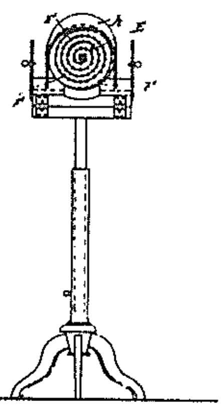

Fig. 3.

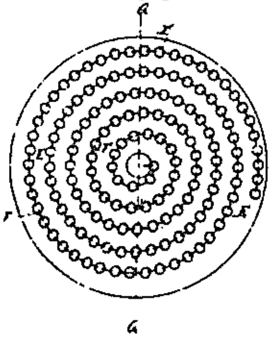

Fig. 4.
(Schnitt a-b Fig 3)

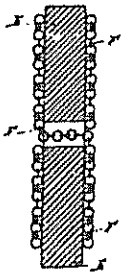

Fig. 5.

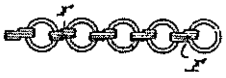

Fig. 6.

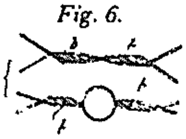

March-April 1988 JBR, VIEW FROM THE BORDERLANDS
by Tom Brown

NATURAL ELECTRICITY

Ernst Lehrs, writing in **MAN OR MATTER** about early views by Volta and Galvani on electric fish, provides insight into the nature of animal electricity. "The electric organ of such a fish consists of many thousands of little piles, each made up of a very great number of plates of two different kinds, arranged in alternating layers. The two kinds differ in substance: in one case the plate is made from a material similar to that present in the nervous system of animals: in the other the resemblance is to a substance present in the muscular system, though only when the muscles are in a state of decay. In this way the two opposing systems of the animal body seem to be brought here into direct contact, repeated many thousands of times.

"In the electric fishes, accordingly, sensation and will are brought into a peculiar interrelation. For as a result of the spatial fixation and the state of deterioration of its bodily foundation the animal's will-pole is related to its bodily foundation in a manner which otherwise obtains only between the nervous system and the psychological processes co-ordinated with it. These fishes have the capacity to send out force-currents which produce in other animals and in man 'concussion of the limbs', or in extreme cases paralysis and even death. When contemplated with pure observation this phenomenon tells us that in cases of minor strength, this force causes processes of a volitional character in bodies affected by it. Through stronger effects the organic system which serves the will of the creature involved passes into disuse, either for a time or for good. By describing the process in this way we realize that electricity appears here as metamorphosed animal will, which takes this peculiar form because part of the animals' volitional system is assimilated to its sensory system in an exceptional manner.

"It is known today that what nature reveals so strikingly in the case of the electric fish is nothing but the manifestation of a principle at work in the bodies of all beings endowed with sensation and volition -- in corporeal terms, with the duality of a nervous and a muscular system -- and therefore at work also in the human body. Observation has shown that the activities of these two systems in man and animal are accompanied by the occurrence of different electric potential in different parts of the body. Manifold in detail though the relevant observations are, they all confirm a definite correlation of negative electricity with the nerve-and-sense pole (sensation) and of positive electricity with the blood-and-muscle pole (volition)." From **MAN OR MATTER**, © 1958 Ernst Lehrs, Rudolf Steiner Press, London.

Thus we see that natural electricity is formed by processes which can be mirrored in non-living matter. But if we can have access to the primary form of the energy why pursue the secondary, and seek out that "magic frequency" that will cure AIDS, cancer, and assorted body clogs that go under fancy medical names? The answer is that everyone has their individual path to travel. Spiritual healing sometimes takes a long time, while an electronic device may get the person back on the track of their day-to-day life. The spiritual lesson of the sickness is buried for the future. Electricity is the subsensible force, but it is a real force with the ability to alter the functioning of the human body.

November-December 1989 JBR, VIEW FROM THE BORDERLANDS
by Tom Brown

MWO NOTES

The work of Georges Lakhovsky continues to be of strong interest. I have been investigating the Lakhovsky MWO for about four years now and my views have consistently changed with new information. BSRF's MWO Handbook is getting quite popular! However, my views have changed somewhat since doing that book, and eventually a new edition will have to be put out to keep up with the new data.

What was Lakhovsky really up to? Well, his basic experiments were with copper wires which were circled around plants. Lakhovsky felt that these rings were activated by cosmic radiation. Lakhovsky was extremely interested in subtle energies, such as those that guided birds on their migrations. The copper rings, when placed around plants, would heal cancers in the plants. THE SECRET OF LIFE, Lakhovsky's best known book in English shows pictures of these experiments. Lakhovsky then moved to bring these healing energies to humans.

Lakhovsky also intuited the DNA spiral coil in the center of every cell long before anyone "discovered" it. He felt that a properly generated multiple wave electrical field would make the coil in each cell resonate at its proper frequency, just as a tuning fork resonates when its frequency is struck on a nearby similar fork.

Today's conceptions of a Multiple Wave Oscillator (MWO) are quite different from what Lakhovsky was really doing. Lakhovsky tried to reproduce the healing energies of the cosmos by electrically energizing coils to produce energy fields which apparently regenerated aberrant cells along the lines of his theories. Lakhovsky first did this by producing an antenna, comprised of concentric copper tubes, the outer one being energized with two wires from his power source. Our impression of his antennas is that they were log periodic, that is logarithmic in nature (Lakhovsky doesn't come right out and say this, but his patents are definitely worded to cover such an arrangement). New data coming in (but not yet verified) tells us that these copper tubes were filled with one or more of the noble gasses (also called rare, or inert gasses).

As anyone could guess, these antennas would be somewhat hard to manufacture. Lakhovsky then miniaturized the antenna and inserted it in a gas bulb. His US patent #2,351,055 says vacuum bulb, but also covers low pressure gasses. There are two main keys here. One is the log periodic antenna. All higher forms of life are based on logarithmic patterns, from ferns to humans. Lakhovsky was creating a structure in the ether which would bring cells into proper functioning. The other is the vacuum / low pressure gas effect. For those who wish to understand the potent effect of the vacuum on life then the reading of The Nature Of Substance by Rudolf Hauschka is required. According to Walter Russell the noble gasses are the octave markers in the harmonic spiral of the elements, which also exists in a logarithmic pattern.

I recently visited with Dr. Norman McVea of Stinson Beach, California. Norman has several high frequency gas tubes built trying to understand what Lakhovsky and Royal R. Rife were up to. Norman has a MWO built of 12 glass concentric rings with various noble gasses inserted in them. When we checked it out it was running at 60 cycles, pulsed at 7.8 cycles. My recommendation to Norman was to run the unit with a radio-frequency oscillator which can be tuned from 1 to 30 megacycles, and modulated with a variety of sources for research. Norman has produced the most unique MWO I've seen in our modern times, and he is certainly on the right track. He is also working to reproduce Rife's high-frequency gas bulbs.

The key to Royal R. Rife's frequency instrument is also the use of the noble gasses energized at high frequency (possibly 4.5 megacycles), which were then modulated with specific musical tones. Lakhovsky's units were energized similarly, but used the frequency-independent log periodic antennas, giving a broad band of treatment, rather than the specifics of Rife's units. In many ways these two pioneers were working with the same keys.

WHO DUNNIT?

Rife and Lakhovsky have another similarity, that is, that people today are building equipment obliquely related to their work and offering them as original, or "improvements." This needs serious examination. The last two years has brought on an underground cult of so called "Rife Frequency Generators." There are hordes of these so-called "Rife" units being sold which are merely fancy variations of TENS (Transcutaneous Electrical Nerve Stimulation) devices. The conception for these units, usually producing square audio waves, comes from John Crane who knew and worked with Rife in his later years. John has affidavits from people who swear that they were healed from cancers and such by his square wave units in the early 1960's. It is extremely important to note that Rife's work in the 1930's through the 1950's was with high frequency gas tubes and all his documentation covers those units. History is rapidly becoming blurred in the race to cash in on people's poor health.

Modern purveyors of the so-called "Rife" equipment are making several errors: The first is that they claim that Rife developed the use of the square wave generators with electrodes. That is incorrect, it was John Crane (which makes Crane a pioneer of TENS research!). The second is that they claim that Crane's equipment was junk and would never work. Maybe the equipment wasn't of the highest quality (we have an original unit in our collection!) but Crane does have a lot of signed affidavits, several of them in our files, claiming that the units worked on specific diseases. In general, people are being misled in order to sell them equipment they probably don't need. That is a bad reflection on the field of electrotherapeutic research.

Perhaps these modern "Rife" (TENS) units have some merit in killing certain types of pathogens. I've spoken with many people who were completely dissatisfied with their so-called "Rife" units, and I've spoken with two people who have claimed that they have killed infections such as foot rot and herpes. Maybe they do, but for all of the thousands of devices being put out where is the documentation? Where is the proof? I have seen Crane's affidavits of the early 1960's, but have seen no medical studies from anyone, not even basic notes!!! There is obviously a very important bit of work to do in researching the effects of pulsed DC on the human body. TENS units kill pain, that has been clinically documented. The claims of the so-called "Rife" units are just that -- claims, *and unsubstantiated to boot.* Until any substantial research work is done in this area I would recommend that people do not follow the mob hysteria which shrewd hucksters have instigated over Royal R. Rife's work.

Lakhovsky's story is similar. All the units being built today are based on Bob Beck's early 1960's interpretation of Lakhovsky's first patent (#1,962,565). After four years of looking at these devices and reading Lakhovsky's patents and books I've come to the conclusion that Bob was misled by the patent. From an electrical engineer's viewpoint Bob's units make sense. From the standpoint of understanding subtle energies within and surrounding the body the units are merely RF noise generators which have questionable effects on the body, and most likely interfere with natural energetic processes (chi flow, etc.). The Beck MWO is a variation of one of Tesla's electrotherapeutic devices, the variation being that Tesla went to great pains to make the antennas spark-proof, and Beck's antennas spark away. Lakhovsky's units didn't spark either!

What makes this more interesting is an article which appeared some months ago in the **Health Freedom News**, titled Who Dunnit? by Mike Brown. Mike gives a weak case for his view that Lakhovsky was inept, that his earlier units didn't work, and who needed Tesla to fix them up and make them work. Perhaps Mike didn't notice that Lakhovsky's SECRET OF LIFE, which he sells to promote his units (in the ad following the article), is based on the units he claims in the article didn't work! Mike refers readers to Tesla's article High Frequency Oscillators For Electro-Therapeutic And Other Purposes (**Lectures, Patents, Articles** by Nikola Tesla, Beograd, 1956), but fails to notice that Tesla's article contradicts what he is doing with his units. I wonder how that article got past the editors?

CORRECT DIRECTIONS

Lakhovsky was a brilliant scientist, highly regarded on the European continent. His name comes up in many books, most recently I found him referenced in Lilly Kolisko's **AGRICULTURE OF TOMORROW**. If he and Tesla met, I'm sure it was for more than to fix poorly engineered units. There are three types of units that I feel should be built for research into this area: One is Tesla's oscillator circuit. The higher quality Beck MWO units, such as those built by Klark Kent, can be converted fairly easily. Use the unit's power supply as is, but for antennas use insulated copper plates (or spheres, as Tesla preferred for maximum effect) coated with heavy dielectric material such as beeswax. Eliminate the sparking! A necessary design change for future units will be to change the ¼ wave coil to ½ wave.

The second is a reproduction of Lakhovsky's Tube MWO. It will take miniaturizing the Golden Ratio antennas that Eric Dollard designed for the MWO and inserting them into low-pressure rare gas bulbs, which will then be driven at high frequencies. It is my estimation that this is the best way to pursue Lakhovsky's high frequency researches. The mathematics for the antenna, along with further reference materials on log periodic antennas, are in The Lakhovsky MWO Handbook.

The third is a reproduction of Rife's original type of equipment being a rare gas bulb driven at high frequencies, modulated with musical tones. This will take RF broadcast equipment and a rare gas tube. No electrodes!

It is time that researchers in this area begin to go in the correct direction. Bob Beck made a noble effort in the early 1960's with his conception of the MWO, and gave people something to work with for 25 years. That phase is definitely over and a new one beginning.

The "Rife generator" craze should be stopped now before it takes people another 25 years to find they've been misled. Serious research into the effects of pulsed DC should be undertaken by those with the wherewithal.

And keep it uppermost in your mind that electricity is a sub-sensible force. In the borderland sciences one should look to the supersensible, such as Dr. Wilhelm Reich's orgone energy (which, by the way, reacts to electricity in such a way that makes it detrimental to the human body), and the human vital energy such as is utilized in Biocircuit research (as begun by L.E. Eeman).

January-February 1990 JBR,

VIEW FROM THE BORDERLANDS

by Tom Brown

THE WONDROUS ELECTRICAL FORCES

Electrical research is a primary area of borderland research. While the term "occult" puts many people off, it should be realized that electricity is an occult force whose effects we only see when equipment is properly engineered. Four to five hundred years ago one would probably be burned at the stake if it would have been possible for them to demonstrate electrical devices.

Electricity is considered by some to be a degraded form of the Light Ether, that the etheric force gets bound up in matter in a way that it can be used. But being bound up in matter it provides mankind with the option of using it properly or improperly. Many articles on the detrimental effects of electric fields have been published, and more studies are no doubt underway on this important topic. There are also some records of electro-therapeutic devices that used this force for healing. Tesla envisioned a worldwide electrical power system and did the basic testing to show himself that it was viable, and needed but to be built. Tesla was a man of great intellect and was a reader of Goethe and other great minds. One would feel that he was in some way familiar with Goethe's *Farbenlehre* (color theory). The type of electricity being sent around our globe on wires is of a different, lower form that that of Tesla's transmission of electricity without wires. It is quite possible that Tesla's electricity was of a beneficial nature. Tesla did research into electrotherapeutics and was no doubt aware of both beneficial and detrimental effects. With the power system he envisioned on line we may not have ended up with the electrical problem we are just beginning to see the "tip of the iceberg" of.

This point needs clarification and elaboration beyond what most people now imagine. We periodically mention certain aspects of Dr. Lilly Kolisko's experiments with the crystallization of metallic salts rising *contra* gravity. These experiments were done to establish whether or not the planets had any direct connection to the metals as alchemical lore supposed.

The connections she investigated are Gold-Sun, Silver-Moon, Mercury-Mercury, Copper-Venus, Iron-Mars, Tin-Jupiter, Lead-Saturn. For example Dr. Kolisko would take a gold chloride solution (1%) and allow a portion of it in a laboratory dish to *climb* a segment of blotter paper. As it dried the fine details of crystalline form could be seen. After many experiments the characteristic form of each metal's salt could be comprehended. It was scientifically established that the actions of cosmic bodies would have direct effects upon the crystallizing actions in the lab. Gold chloride would not crystallize properly during several solar eclipses that Dr. Kolisko reported on (1927, 1947). A darkening effect was noticed on these slides.

Just recently I came across this in the recollections of Dr. Rudolf Hauschka, founder of Wala Pharmaceuticals: "In the course of researches into the delicate changes in matter which are revealed in the capillary-dynamic method, an alarming symptom came to the fore. At one of the Easter conferences of the Anthroposophical Doctors' Working Group in Comburg in 1964, Frau Kolisko, the past-master of the capillary-dynamic method, spoke on 'Gold and the Signs of the Times.' In earlier years she had demonstrated by means of the *Steigbild* (literally rising pictures, also known as chromatograms) the connections between the metals and the planets, and in particular the effect on gold of an eclipse of the sun. During an eclipse of the sun the gold *Steigbild* which normally glows with beautiful colors, shows a darkening and the emergence of grey formations. Frau Kolisko now showed a series of gold *Steigbild* produced in the last years almost hourly and at all times of the day and of the year. Without exception all of them showed the same sinister dark formations which had appeared earlier, in the first half of the century, only at the time of a sun eclipse. She closed her lecture with a question to the doctors, namely whether, considering the change now in gold, they could still be answerable for its use as a medicine. There was no doubt about it: since the middle of the century gold had become different – it was sick. Or, to express it differently, the connection between the spheres of the sun which stream down and the metal which they permeate is disturbed."

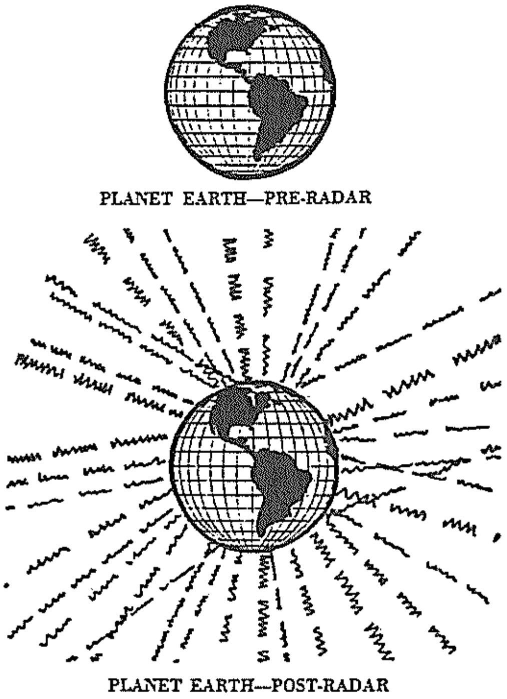

Entire planetary environment permeated with pulsed electromagnetic radiation—an unprecedented condition.

ELECTRICITY: THE MAIN CULPRIT

The main culprit in this sickness of earthly matter is none other than the electrical environment that has manifested on this planet. Certainly there were radio communications going during Dr. Kolisko's early experiments, but as this chart from Trevor Constable's *COSMIC PULSE OF LIFE* shows, the post WWII environment was heavy with these signals, culminating in today's total saturation point. Imagine what the *Steigbild* would have shown in previous centuries before the electrical force began its manifestation.

Nuclear proliferation must also be considered as adding to this phenomenon, but electricity is the degraded form of the light ether, whose source the sun is related to gold. Walter Russell, in ATOMIC SUICIDE, indicates that the spread of nuclear energy will destroy the atmosphere surrounding this planet and now scientists are scratching their heads as the hole in the planetary ozone layer gets larger. (Of course chlorofluorocarbons and New Zealand sheep farts have all been blamed for the ozone depletion, but nuclear energy is a primary cause.) This atmospheric destruction could well aid in cutting off gold from its cosmic wellspring of life.

Rudolf Steiner had many insights into the true nature of electricity, its relationship to humanity and its role in our future. Pertinent to the matter at hand let us consider the following remarks, a portion of his answer to a question about electrical treatment of animal fodder:

"You must consider the whole part played by electricity in Nature ... to the effect that human beings cannot go on developing in the same way in an atmosphere permeated on all sides by electric currents and radiations. It has an influence on the whole development of man. This is quite true; man's inner life will become different if these things are carried as far as is now intended. It makes a difference whether you simply supply a certain district with steam-engines or electrify the railway lines. Steam works more consciously, whereas electricity has an appallingly unconscious influence; people simply do not know where certain things are coming from. Without a doubt there is a trend of evolution in the following direction. Consider how electricity is now being used above the earth as radiant and as conducted electricity, to carry the news as quickly as possible from one place to another. This life of men in the midst of electricity, notably radiant electricity, will presently affect them in such a way that they will no longer be able to understand the news which they receive so rapidly. The effect is to damp down their intelligence. Such effects are already to be seen today. Even today you can notice how people understand the things that come to them with far greater difficulty than they did a few decades ago.

"It is a remarkable fact that whenever something new appears, as a rule in the early stages it is heralded as a remedy--a means of healing. Then the prophets get hold of it. It is strange, where a new thing appears, clairvoyant perception is often reduced to a very human level! Here is a man who makes all sorts of prophecies about the healing powers of electricity, where no such thing would previously have occurred to him. Things become fashionable! No one was able to imagine healing people by electricity so long as electricity was not there. Now--not because it is there, but because it has become the fashion--now it is suddenly proclaimed as a means of healing. Electricity--applied as radiant electricity--is often no more a means of healing than it would be to take tiny little needles and prick the patient all over with them. It is not the electricity--it is the shock that has the healing effect.

"Now you must not forget that electricity always works on the higher organization, the head-organization both of man and animal; and correspondingly, on the root-organization in the plant. It works very strongly there. If... you pour electricity through the foodstuffs--you create foodstuffs which will gradually cause the animal that feeds on them to grow sclerotic. It is a slow process; it will not be observed at once. Electricity will not at first be recognized as the cause; it will be ascribed to all manner of other things.

"Electricity, once and for all, is not intended to work into the realm of the living--it is not meant to help living things especially; it cannot do so. You must know that electricity is at a lower level than that of living things." (from **Agriculture** by Rudolf Steiner)

How pertinent and timely these words are today on many levels of borderland research. Electrical/electronic medicine has the glamorous position on stage at most of the conferences, and in many people's minds, in today's 'borderland' field. People are racing to buy instruments which are designed by electronic engineers with little or no training in the comprehensive constitution of humans. *It is the fashion!* There is no consideration that true healing comes from the spiritual realms, that the spirit is always healthy, and that dis-ease comes from a condition where the various subtle levels of the body are not aligned properly with the physical.

These electrical instruments have the potential to produce apparent results in treating disease symptoms. But what of the long term effects of the subsensible electric current? Considering the alarming state of affairs viewed through the information just presented I pose the question to the reader, "Do you prefer the natural or the unnatural forms of treatment?" Many people flock to electronic medicine because they feel that the petro-chemical drugs that are prescribed by the medical establishment are detrimental to their long term health. Very possibly so, but will electricity be any better? What about your own vital energy and its relation to cosmic life forces?

Electrical stimulation of the body treats a symptom and fails to allow for the body's natural healing process. As electricity is an externalized part of humanity's inner volitional force we can see that there will be definite effects from the subsensible current. But the effects will be relative and the inner healing process put on secondary mode. Electro-therapy, while emerging as a viable medicine of the future, still ignores the natural healing process.

This does not mean that we should not research electro-therapy. In what may seem a contradiction to many readers, I am 100% for further research into this area. It will no doubt be of benefit to many. Our relationship with electricity is one of the greatest questions of our time. People unconsciously use this force, and the results are such as those noted by Lilly Kolisko. Electricity is intimately connected to the Life Force, but great care must be taken not to confuse the two. Electro-therapeutic research helps define the boundaries, and thus is of interest to Borderland Researchers, but it certainly does not mean I recommend it!

Mar-Apr 1990 JBR, BSRF Bulletin Board

MWO QUESTIONS

Your comments re the MWO in the Nov-Dec 89 JBR are of much interest and raise the following questions in my mind:

1. Will you be publishing any information by Dr. Norman McVea?

2. How can the Kent units now in circulation be converted to produce a 1/2 wave? Perhaps Klark would be willing to publish an article on this or be willing to do the converting for a price.

3. Has any member ever tried to obtain any information on the circumstances of Lakhovsky's death? About three years ago a reference librarian made a serious effort for me to obtain any information re Lakhovsky, with absolutely no success. Perhaps there was an obit in the NY Times at the time he died, assuming that is where he was at the time.

4. Where in the world would we obtain copper spheres like Tesla used? Perhaps a member can suggest a source.

P.V.P.
California

In answer to your questions:

1. Norman McVea has sent us in a bit of material on his latest oxygen researches. He has opted for the oxygenation approach instead of the electrical.
2. I don't know about conversion of Kent's units to 1/2 wave, you should check with him on that. It will probably be easier to use it as it is -- they are good power supplies for experimentation -- and in the future engineer units with more refined outputs. After reading my column Klark added a rare gas tube assembly to his "Rife" units. He has always listened to other people and has constantly worked to improve his units.
3. If any readers have info on Lakhovsky's death, or his life for that matter, these pages are open to pursue the matter.
4. Anybody know where to get copper spheres? Tesla also used copper plates coated with a heavy dielectric.

September-October 1990 JBR BSRF Bulletin Board

MWO EXPERIMENTS

I have been very fascinated with your Journal. Thank you for the new world you have opened up to me. For the last few months I have begun to work on an MWO.

I have completed the antennas but did not do as well with the Tesla-coil. I decided to buy the ready made coil from

"Edmonds". I connected the secondary to the center of my antennas and did get some action in the form of a purple glow around the third and fourth smaller rings. I can only assume that I need more power to this unit although I was "Quite" pleased at the limited reaction on my first attempt. I would like to reach out to a fellow member who might be able to help me with some advice at this point.

Any advice you would send my way would be greatly appreciated.
P.D.
New York

Naturally, the quality of output of any electrical device is dependent on not only the quality of materials, but also the geometry of structure and arrangement. The better the coil, the better the antenna, the more refined the fields will be that are put out around the device. As I have pointed out in the article mentioned above in my remarks about the "Rife" devices, the types of "Lakhovsky" MWO's being built today vary quite a bit from what Lakhovsky was actually building and using himself.

Lakhovsky used very refined electrical designs and probably log-periodic antennas with both poles of a Tesla coil feeding the outside ring in such a manner that there was a logarithmic, electro-magnetic structure oscillating in space which apparently had quite beneficial effects on living organisms.

The modern variations, based on Bob Beck's schematic of 1963, feed one pole into an antenna (and the other into a second antenna), creating a dielectric field, which when it collapses, due to the rings on the antennas, creates unnecessary sparking and random radio-frequency noise with questionable effects. It is an organic system alterant, but is quite different electrically from Lakhovsksy's original equipment.

Nov-Dec 1990 JBR, BSRF Bulletin Board

THAT MWO AGAIN

I recently purchased your MWO HANDBOOK. In the book are many references to the "Golden Ratio". I was unable to determine what the mathematical basis for the Golden Ratio is. I have other material which describe the Golden Section, and it is a ratio of .618 to 1. The closest I came to a value for the Golden Ratio is from the MWO antenna data on page 188 of your book the ratio appears to be 1.27:1. Could you please clarify the mathematics of the Golden Ratio antenna?
B.C.S., California

*** I recently purchased THE LAKHOVSKY MWO HANDBOOK. I am in the process of building an oscillator. I have all the parts except 6Kv .005mf capacitors. No electronic store in Alaska has them, or knows how to get them. I would appreciate it if you would send the address of either the manufacturer or an electronic store that carries 6Kv .005μf capacitors. I have one more question. Do you know of a meter or electronic apparatus that measures the amount of energy level the human body produces? It seems that like fingerprints, all human bodies are different in the amount of energy they use or produce. If one can coordinate the energy level of the oscillator to that of the body far less damage would be done and far greater healing potential be made.**
S.N.
Alaska

For the first question, the Golden Ratio antenna uses the golden ratio as a base, but it was engineered to conform to other considerations of antenna propagation. It is what is called a log periodic antenna. These antennas are based on the square root of phi, that is the inner ring is multiplied by the square root of phi to get the next ring, or inversely the outer ring is multiplied by one over the square root of phi to get the next inner, etc. This provides the Golden Ratio mathematics (every other ring is phi related).

MWO NEWS FROM FRANCE

I'd like to mention with regard to "Electricity the Main Culprit" that just after the war, Gurdjieff wrote a chapter in his book ALL AND EVERYTHING entitled "Beelzebub says that man's capture and subsequent destruction of electricity from nature is one of the main causes of the reduction in the human lifespan." And you are lucky in the US. In Europe we have a 230V system that is far more harmful than yours.

A final comment on the MWO which seems as popular as ever. When I lived in Paris I made the acquaintance of Lakhovsky's son and saw two of the MWO's built by his father in operation. The demonstration was very impressive but it didn't last too long because the old units interfered seriously with radio and television reception in the area. They were very big units, about as large as an early TV set. I don't know where they are now since he later loaned them to friends living in the country.
R.C.
France

Very curious quote from Gurdjieff as it concurs with what Steiner indicated, that electricity makes living things sclerotic. If perhaps you wouldn't mind searching your memory a bit more I have a question about Lakhovsky's device: You say it interfered with radio and television reception. Was this from internal spark gaps, or did any unit you see arc between antenna coils. The pictures in THE SECRET OF LIFE indicate that the then necessary spark gaps were interior to the unit, and the pictures of the effluvia coming off the antenna are obviously pictures of corona discharge and not arcing as in the units built in the US today. Bob Beck, in his 1963 article on MWO's claims he built a unit from an original, but his antenna hook-up is quite different from Lakhovsky's pictures. Many years of searching have failed to produce an original unit. The modern units are derived from the most obvious derivation of Lakhovsky's first MWO patent, which is written to both protect his design and cover infinite variations thereof. A detailed account of an original MWO in action would be of great value in this research.

SOUTH OF THE BORDERLAND

I have read Mark Clement's THE WAVES THAT HEAL and discovered a little piece of information not included in the current edition to THE LAKHOVSKY M.W.O. HANDBOOK. In WAVES I noted a short sentence indicating that one set of rings transmits energy through space which falls upon the patient's body and (it must be ascertained through experiment whether this is as I tell here or otherwise) later absorbed by the receiving resonator which resonates it for rebroadcasting into the patient's body. Thus the true Lakhovsky MWO, and UNLIKE the Beck-reconstructed model, would be a more-complex device involving features so-far unthought of. It must be ascertained what KIND of "receiving resonator" Lakhovsky employed, if it had or not special constituents not described at any of his patents included in the MWO HANDBOOK and what other features could be complementary to both.

Another feature to be considered, and which I discovered in the photo on page 30 of Clement's book is that there are two protruding spheres connected to the outermost ring of the Lakhovsky setup. In the copy of the BSRF publication I have, these two photographs are reduced and with not the same kind of resolution of the work by Clement. Each protruding sphere is connected to a cable and following the path of least resistance

the electrostatic current flows both along the metallic path of the outermost ring and the excesses accumulated from attracting ether from the environment are projected towards the patient (in the case of the transmitter) and collected for resonating (in the instance of the receiving resonator). Also notice that there are two “bracing wires” to both keep the rings in place and at the same time to fulfill the 90° division determined for Earth by the metric theorems which are the foundation of the Energy Grid! This adds another feature of complexity to the duplication of the MWO, which carries the Lakhovsky invention, again, farther from the Beck-built model.

Please mention my paper SOME FREE-ENERGY DEVICES for the MWO analysis I made is at best very incomplete, not including the feature of non-uniformity the true MWO does possess.

Jorge Resines
Buenos Aires, Argentina

Thanks for your further insights into the MWO. I feel that a continuing dialogue will eventually produce more evidence as to what type of equipment Lakhovsky used in his successful researches into cell regeneration and how to best reproduce it. The resonator antenna setup on Lakhovsky’s original is more complex than the Beck MWO, as, like the primary antenna, it also has two feeder wires. The two spheres you mention, from which brush discharges can be seen, have always struck me as the nuts that held the heavy antenna onto its stand. Some mechanical contrivance was necessary and the mini-spheres fit the bill with the possibility that they may have been further configured to produce a specific capacitance, as our evidence indicates that Lakhovsky’s antennas were very precisely built, probably log periodic. We have heard that Lakhovsky was a radiethesiest and would dowse the sites where he set his equipment up. He was certainly into the subtle energies, and the French school was very precise in its radiesthetic work, being the source material for much of Christopher Hills Supersensonics work. We have an unconfirmed report that the original tube antennas were filled with a rare gas such as argon. This would seem to be the case as in Lakhovsky’s second MWO patent he reduced the antennas and puts them into vacuum or low pressure gas bulbs (as the first, this patent is written to be very specifically vague). This was meant to be an improvement over the original units, and probably was. What it tells us is that the “Beck” units appear to be far from what Lakhovsky was doing and where he was going, though because of more than 25 years of people building and using Beck models it is hard find information on the original. The “bracing wires” are supposedly silk. Their angle was probably determined solely by mechanical function.

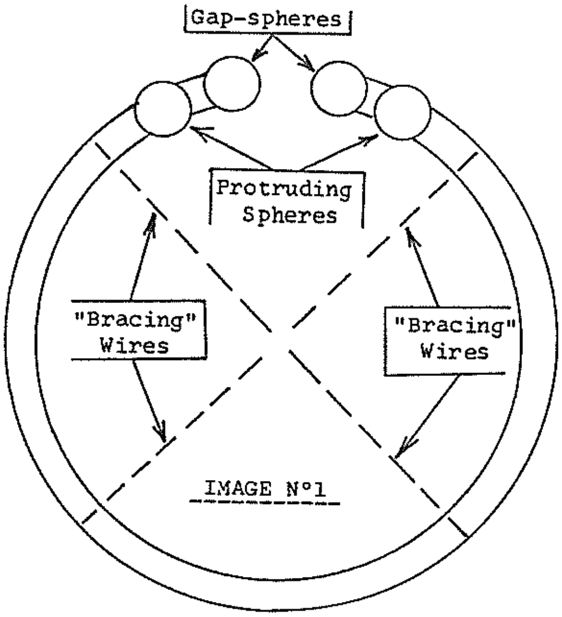

Detail of outer ring on Lakhovsky’s original MWO antenna. Front view above, side view below

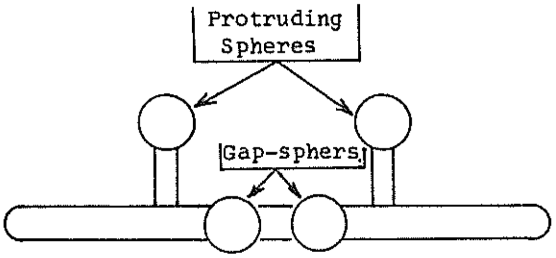

We have actually come under attack for our attempts to clarify the MWO story! In a brochure put out by MWO salesman Mike Brown, Box 25, Pleasant Hope, MO 65725, we come in for the following treatment: “Then there are the characters who, with no formal training whatsoever, denigrate tings [sic] they know less than nothing about. One outfit calling itself Borderland Research sent out a bulletin claiming that no-one today is manufacturing a Lakhovsky Multiple Wave Oscillator. These characters arrived at their conclusion by comparing photos of a 1935 unit to ones being made today. The units don’t have the same external appearance. Quite often people with no formal training in electronics or electricity don’t understand that an electronic device and how it operates are determined by the circuitry contained within the unit, not by the box it comes in.” My comments are that there are perhaps *tings* we don’t know but we do know that what is inside counts as much as the outside on the MWO. I have put five years of serious investigation into the MWO and would have to say that what we published is the result of serious consideration by a number of intelligent people--some highly trained electrical and mechanical engineers--interacting and exchanging ideas. We have published the evidence to support our contention, and have in fact stated that if we can be shown to be wrong then we would like to know, that we have no final conclusions, just some of the facts. A crude attempt to belittle BSRF’s research by ignoring the main facts does not change the facts and the facts are that the units pictured in Lakhovsky’s book **THE SECRET OF LIFE** are quite a bit different than the units being built today, indicated specifically by the antenna hookups, from which basic logic would indicate the need for a different interior setup. A thorough analysis of the basic differences between the two types of units can be found in The Lakhovsky Multiple Wave Oscillator Handbook, edited by myself, Tom Brown (no relation to Mike). Mike’s brochure then goes on to claim that “the equipment necessary to regrow arms and legs can be purchased at Radio Shack for about $10”, that is a “metal plate, copper wire, battery, and 100 K-ohm to 1.5 Megohm resistor.” Heck, if you can regenerate an amputated limb with a only battery and a resistor then why do we need all the electronics in a MWO just to drive out mere cancerous tissue?!

As a final note on this, I would like to point out that electricity is a system alterant. Whether it alters the system for better or worse depends on many other factors, including mental status and diet. As Trevor Constable pointed out back in the 1960s in this Journal, the healing powers attributed to the MWO could be caused by the etheric force rushing to suppress the life-negative electro-magnetic activity. It is also possible that any benefits coming from modern day MWOs are psycho-tronic in nature -- that is the mind is doing the cure and the instrument provides a physical anchor for the psychic energies. The actual operation of the equipment, the sparking and gross electrical output, is superfluous to the matter, other than in convincing the mind that *someting* is happening.

Jan-Feb 1991 JBR, BSRF Bulletin Board

LAKHOVSKY'S ORIGINAL

You ask whether the original Lakhovsky MWO had any arc between antenna coils. I don't remember any but to make sure I checked up with a friend who has worked on the MWO and knows it inside out. He was also present when I saw the device working. He tells me that there is definitely no arc between coils. The arc occurs inside the machine after the "Ruhmkorff coil". He also says that there are two tungsten plates which are drawn apart and brought together manually.

The other thing I remember is a sort of metal wand with a knob at one end which is used between the two units to treat people. At the time I saw it, one of the persons present had a bad back. The two MWO antennas were turned on and were facing each other about 6 ft apart. The "wand" was applied to the sore back. I seem to remember that it kind of gave off sparks, but I'm not sure of this. I touched the wand and I could clearly feel something coming through. My knowledge of electricity is nil, however, if you have further questions, I could always try my friend who should be able to answer them for you.
R.C., France

Many thanks, as this confirms the evidence of Lakhovsky's own pictures used in his books. Would you be so kind as to ask your friend if it is possible to get a schematic (even a rough drawing from memory) of the basic layout of the MWO, which we could reproduce on these pages, also any antenna construction details. Many questions remain such as what components are in the metallic tubes holding the antennas up, and such (Eric Dollard finger-dowsed the pictures of the tubes and felt it was an arrangement of capacitors and spark gaps -- his basic MWO diagram, an intelligent guess of what Lakhovsky was doing, is in this Handbook). A basic schematic would probably be the most important piece of information to date. Please read on, pot has been stirred.....

LAKHOVSKY ANTENNAS

I got my Nov-Dec 89 Journal where it shows an update on Lakhovsky's antennas. On page 181 of SECRET OF LIFE it also shows the same thing. Take a magnifying glass and you can see it. I disagree with all the experts, it seems that they are chasing a phantom ghost. Lakhovsky gave us all his secrets in his book, all we have to do is read carefully.

He could cure cancer and other ailments with a simple piece of copper wire. Lakhovsky wasn't dumb, can you see him telling doctors to cure with a piece of wire they would tell him to get lost, so he built a very complex machine to impress them, also he would have something to sell. I think Beck's machine is close except the antenna. Who came up with this aluminum foil glued to plastic? That's where they fell on their face. If you read page 75 of the same book he repeats VIBRATES 3 times. That's why he hung his rings with silk thread, so it wouldn't damp the VIBRATIONS.

Can you imagine a pipe organ with copper strips glued to a board? I have designed an antenna using copper tube of varying sizes held by fish leader that I think would work fine. I'd bet a penny to 10 bucks that George used thin wall bronze tube as it will resonate better -- but it's expensive to fabricate.
Louis A. Schad
California

Yup, I agree that Georges was no dummy. His book shows plant cancers that have been cured via use of a single copper wire single-coiled around the plant. He felt that the pre-matter cosmic rays (or what I would call the material pole of the etheric instreams) were "ringing" the copper coil, in effect creating a tone that would heal the plant. I got the impression, from studying Lakhovsky's progress towards the vacuum (or low pressure rare gas) patent, that in his equipment Georges tried to invert the process -- exciting the ring with a post-material high frequency current that would create a healing tone. By using log periodic antennas (as would be allowed under the MWO patents) a vibrant chorus would ring in the ether, the tones, via the structure of the antenna, would be such that they could harmonize the functioning of the human body.

The aluminum foil antennas were developed by Riley Crabb, as a low cost way for experimenters to make their own antennas without having to have printed circuit boards made, and many folks were quite happy with the results. The physical evidence now clearly indicates that the Beck unit is quite different from the original Lakhovsky unit for which the cell regeneration results were recorded. What I have termed the "BECK MWO" must stand on its own merits and it can lay no claim to Lakhovsky's recorded results. As stated several times in the past Beck's device is more like one of Tesla's electro-therapeutic oscillators, except Tesla used copper sphere antennas, coated with a heavy dielectric to suppress sparking. Beck's unit would be covered under Lakhovsky's original (and very generalized) patent, but research of Lakhovsky's own material shows it is not what he used for electro-therapeutic researches. To excite the antennas in the Beck manner will not create the same vibrating field that Lakhovsky's antennas excitation circuit will. But it does create an electrical field that some claim is beneficial to a live body placed in its proximity.

It makes much sense that the antenna should also vibrate physically as well as electrically. Ah... if we simply had an original unit to weigh and measure...

Lou has sent us in a MWO of his own design with a single spiral antenna made of two twisted bare copper wires. We will be running a several part article on construction and use starting next issue. This unit is reminiscent of the Sun-Ether disc of Oscar Korschelt, circa 1891. Having sat in front of a number of different operating MWO's over the last six years I have to say that Lou's unit is the most exciting thing I've ever seen like this. I could strongly feel the effect 10-12 inches away, and there was no electrical excitation of the antenna, rather it is physically vibrated. There is a low voltage shocker that can drive the antenna through two smaller copper coils. It's easy to build and use, simpler than the Beck units, and should provide borderlanders with years of workable ideas. Lou will not be selling these devices but will provide complete building plans in the JBR.

We received a call from Patrick Flanagan. He reports that he has a MWO unit built by Lakhovsky's son in Paris. It is a little box with a one tube white noise generator and it has a multi-stranded cable which wraps around the body. It shows that the evolution of Lakhovsky's device has taken a different turn among those who actually knew what he was doing. This unit is based on the Radio-Cellular Oscillator.

Also, Klark Kent (of Super Science, Box 392, Dayton OH 45409) just called to say that he has completed a run of Golden Ratio antennas from the original B.S.R.F. artwork. These are double sided (with inverse gap strips, to simulate the ball capacitance on the original tube antennas, on back) gold plated and are available from Klark for $100 each. (These antennas have produced beneficial effects on plants without any electrical excitation. Many experiments can be done and the results quantified with various protocols such as those used by Rudolf Hauschka and Lilly Kolisko.)

There is no doubt that there is a new phase of MWO research being undertaken. For 25 years people built Beck's model, with many people claiming beneficial results, and others experiencing nothing but no one keeping track of results for evaluation. Thanks to Bob Beck's schematic (published by B.S.R.F. in 1963) the idea was kept alive and now people will have many ideas to choose from in their search for measurable effects from energies ranging from the gross electrical to the fine etheric. Thanks to everyone who has interacted on this ongoing aspect of our researches!

EVEN MORE LAKHOVSKY STUFF

Some time ago I bought your excellent handbook of the Lakhovsky Multi Wave Oscillator and due to pressure of other things have only just found time to read it. Interesting because it parallels some research done by myself and friends, or should I say we duplicated some of the original. This was only to be expected as the entire principle is so outrageously simple, too simple for establishment minds to comprehend.

I should like to offer a few ideas which might be of interest and in return I would be grateful for some additional information.

I helped develop the Betatron now sold by All-Source Systems, PO Box 596, Morton, WA 98356. The Betatron uses a TV Flyback transformer and is very similar to the Master Violet Ray devices. The nicest feature is that the applicator is an ordinary clear light bulb. Even a burnt out one will do. There are differences; some makes of bulb will not work at all, perhaps due to different gas fillings. The Betatron can also be used with the Lakhovsky concentric antenna with spectacular results. This of course has been done in the book's articles although as far as I know it was not actually copied, but was an obvious choice of experiment.

I have tried out an improvement which I did not see in the Handbook or anywhere else, perhaps you will be interested. The Standard circuit whilst apparently covering a very wide range of frequencies, must, because of its inherent fixed inductance and capacitance have its own natural fundamental frequency. Our Betatron is around 21 Kilocycles (I refuse to use the meaningless term Kilo Hertz), and other units seem to be basically tuned to one fundamental even though they do stray and produce an almost infinite range of harmonics.

My first improvement is to substitute a "Jacob's Ladder" for the spark gap. This consists of two 1/4" copper tubes spaced 5/32" apart at the bottom and tapering outwards to 1" at the top then abruptly turning out at 1" radius. The spark travels up the ladder from bottom to top and is extinguished by the sudden increase of gap, then begins again at the bottom. This sequence takes place at about 2 second intervals. The whole point of this is, as the spark progresses up the ladder, the length of the tuned circuit is progressively increased, then suddenly decreased again ditto repeato. As the ladder is about 9" long the length of the tuned primary of the Tesla coil is progressively made 18" longer, then back to its original length. Thereby resonating at infinite numbers of frequencies within its range and delivering virtually the same power at all those frequencies.

A further improvement on this feature is to make the ladder a Spiral Staircase by using a solid rigid central electrode, say a ½" copper tube and the spiral be wound in an ever increasing diameter on the outside, the start being about 5/32" gap and vertical to begin with.

Alternatively the spiral can be the inside element wound on a porcelain former such as is used in radiant "bowl" heaters with a plain copper tube surrounding it or perhaps a perforated tube as used in decorative lighting fixtures. The spark winds its way around the spiral staircase and alters the frequency accordingly. The Spiral Staircase provides a much longer path and more variation in frequency range. I have not yet perfected this and ruined a few porcelain formers in the process. It requires some exact progressive increases in the travelling spark gap but there are no serious obstacles.

Both the Jacobs Ladder and the Spiral Staircase have the advantage of easily cooling the spark gap. We had experienced problems due to overheating of the electrodes in a fixed gap, even spark plugs transmitted the heat to other parts and carbonized insulation. The addition of a cooling fan (the low power kind used in computers) provides a current of air which accelerated the progress of the spark up the Jacobs Ladder and more importantly up the Spiral Staircase. The spark is sometimes reluctant to leave its easy path at the base of the ladder and even more reluctant to wend its winding way up the coil, but the air flow does the trick and keeps everything cool.

The fan also helps remove Ozone from the enclosure. Ozone has the delightful habit of attacking insulation. It has also been blamed for respiratory illnesses. I am skeptical of the latter as I have breathed plenty of Ozone and found it exhilarating.

Master Violet Ray even included an Ozone Inhaler with its de-luxe model!!!, recommended for asthma, bronchitis etc. What's more M.D.s used it on their patients. Certainly Ozone is about the only thing which will quickly digest Carbon Monoxide in a hurry and it deodorizes a room when the cat has misbehaved.

The Betatron and Lakhovsky machine both can be used to destroy troublesome warts and moles, by simply wrapping a wire around the electrode and applying the point of the wire to the offending growth. I had some troublesome warts removed in this way by a Dermatologist. It was not until I got involved with this project that I realized the Dermatologist was using a Lakhovsky device.

He anesthetized the area first but I have since got rid of warts and moles without this. It caused some "Discomfort" which as any mother knows is worse than mere pain, but it was tolerable.

I hope the foregoing will be useful and will not be in the least surprised it has not already been done, but I would like to know what you think of it.

Regarding Spark Coils. I recently found at the local Flea Market, a small spark coil about 3½" overall length 2" dia. with a sturdy vibrator assembly and a turret mounted sideways. It looked like a small steam boiler in fact. It is labelled "Coil, Booster, Battery Operated. Type C1, (24 Volts DC) Spec. #94-32182, Serial #96082, order #A,C 27073, Eclipse Aviation Bendix, N.J. Division of Bendix Corp. I don't know what it is for but on 12 volts it delivers a lively 3/8" spark. The absence of a Zip Code suggests it is fairly old, but this is the ideal driver for a compact tesla coil. I do not know if it contains condensers but I added some anyway. The adjusting screw has a click stop adjustment. Have you come across

one of these? It works well on 6 volts which is just as well as it draws a lot of juice and 8 D cells won't fire it.

Now a request for information about Lakhovsky Treatment and parts. Please Please. On my knees!! Where can I obtain High Voltage Capacitors as used in the Lakhovsky Machine?????? I have tried all the surplus stores I know of in this area but the highest voltage they have is 450. I recently paid over $50 for a 660 volt one for a 12 v trailer power supply and that was ordered specially from New York. HELP!!! I even tried making one from instructions in the Boy Electrician, a Sunday School prize from the 1930s. I used Acrylic 1/8" dielectrics instead of glass. They worked well for a time but soon tracked and failed. I don't fancy using glass plates any way but I desperately need some H V Capacitors at less than Pentagon prices.

Re Treatments. One of the contributors to the Handbook reports that treatment with the Lakhovsky Machine resulted in the dis-appearance of underarm hair. (The report of decreasing underarm hair is from the Nov. 1968 Journal of Borderland Research by Mrs. E.H., Piqua, Ohio.) This is very interesting as I am a heavy construction worker, and after a hard day at the office I like to drop in for a cool beer and fraternize with nice hot ladies. However, underarm odor which results from my daily exertion and defies Right Guard AFTER for 2 hours, is a Repellant. I did find that shaving the offending fuzz helped a lot. Tearing it out by the roots was even better; it seemed to de-activate the sweat glands.

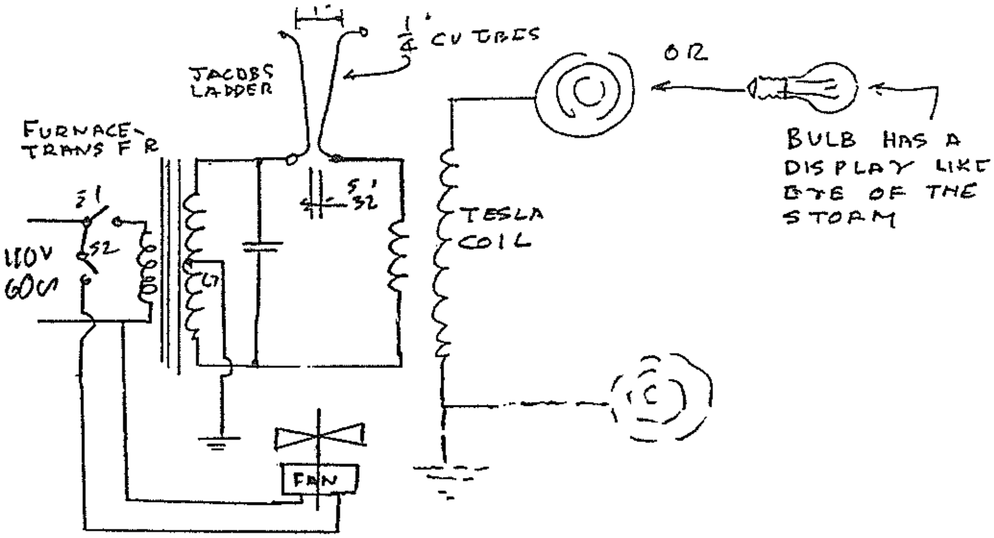

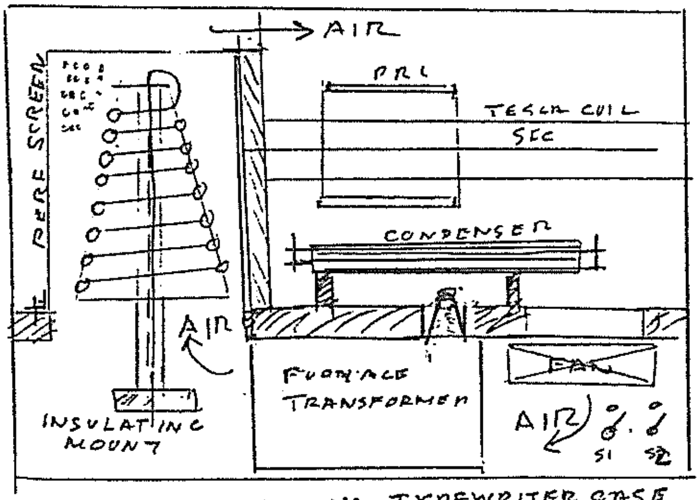

UNIT MOUTED IN TYPEWRITER CASE WITH ACRYLIC FRONT COVER TO DIRECT AIR FLOW

After I sprained my shoulder I used the Betatron all around the joint and have noticed that underarm hair now grows much more slowly, and I have much less trouble with B.O. I used the Betatron frequently and the Ozone removes any residual stink...

NOW the Big question. Has anyone succeeded in using the Lakhovsky machine to stimulate hair growth??? I have hair like Yul Brynner. Maybe some folks think its sexy but I find it repulsive and worst of all it makes me look and feel 20 years older at least without a hat. It seems to me that the machine could well be used to rejuvenate hair cells since it can reverse most other degenerative processes. Its a strange thing that men who are short on top have an overabundance of superfluous body hair in places where it is utterly wasted.

Has anyone done any research on this subject with the Lakhovsky Machine??? If so please let me know. I will be most grateful for any information and will of course inform you of the results of my own experiments.

It occurs to me that such a phenomenon could well dispense with the daily chore of shaving and make Gillette obsolete. I am forever disgusted that such a lot of life force is wasted growing hair on my face instead of where it belongs, On Top! Would be most grateful for your ideas.

I remember the Master Violet ray being demonstrated in a department store in the 1930s, I must have been about 7 years old. My mother who was with me watched the demonstration and I, of course, was fascinated. I said, "Oh Mom do buy one, it's just what you need for your rheumatism". Mother replied, "Oh no, You just want it to play with". How right she was. Well, in spite of all the quack devices which come and go, Lakhovsky in all his forms is still with us. That's recommendation enough. The AMA persecution of course confirms its effectiveness.

A Naturopathic Doctor saw me wearing a copper bracelet. He was treating me for arthritis, effectively with electro therapy etc. and asked if I really believed in the copper. I said "No", then he said "Why wear it?" "Because it works." He laughed and said,

"Look at that green stain on your wrist. That's copper salts and they are being ingested; maybe that's how it works, you might have a copper deficiency, well it's cheaper than pills."
Very Sincerely,
K.V.A.R., Washington

First off, without further ado, let me congratulate you on taking a stand on refusing to use the word Hertz to designate cycles per second. Why use mumbo-jumbo designators when simple English will do. Tesla said that the acceptance of the Hertzian theory would come to be seen by future generations as the greatest aberration of the scientific mind. Of course Tesla is probably rolling in his grave (at 3600 rpm?) at the modern etherless explanations of electricity. He, and others of his day, regarded the (transverse) Hertzian waves as the "waste" put off by coils. The Tesla (longitudinal) wireless system would have been worldwide (including radio and power), with the transverse waves engineered to a minimum, whereas the radio we have today, with the transverse waves engineered to a peak, fades out as one goes around a hill or a decent distance from the emitting source. This is dealt with much more specifically in THE THEORY OF WIRELESS POWER by Eric Dollard. Regarding use of the Jacobs Ladder as a spark gap, that is an interesting idea and thanks for sending it in. As far as your hair growth problems go, perhaps our other readers have further comments. You have provided a lot of food for thought, thanks!

HIGH VOLTAGE SOURCES:

Re: Nov/Dec Journal and inquiry of S.N. Capacitor 6KV (6000V) .0047 uF (which is the standard value, not .005!) available ($4.90) from: Antique Electronic Supply, 6221 S. Maple Ave., Tempe, Arizona 85283. (602) 820-5411.
A.A.
Massachusetts

Source for high voltage equipment: Plastic Capacitors, Inc., 2623 N. Pulaski Rd., Chicago IL 60639-2190. Write for catalog (Dealer for above is Newark Electronics). Also recommended is Homemade Lightning: Creative Experiments in Electricity by R.A. Ford, ISBN 0-8306-3576-9, TAB Books, Blue Ridge Summit, PA 17294-0850. Can be ordered through your bookstore.
Simplified Tech Service
PO Box 2140A, Champaign IL 61825

In the Nov-Dec 90 issue S.N. of Alaska asks about a 6Kv .005 uf capacitor. Radio Shack sells a 2Kv .01 capacitor #272-160. Three connected in series will be 6Kv .003 uf, another 3 connected in parallel across the first three will be 6Kv .006 uf. Six of these costs about $3.

Regarding the MWO, when Hieronymus lived near me in Ft. Lauderdale I asked him about the MWO. He didn't recommend it because the broad spectrum of frequencies had both beneficial and harmful frequencies which could not be separated.
R.B.
Florida

Thanks for all the input on where to get HV capacitors. Also, thanks Ron, for Hieronymus' viewpoint. With the Beck style MWOs I would agree. A mass of radio frequency noise is generated. With Lakhovsky's original units driving structured antennas one could produce multiple wavelengths of a more refined nature. We don't recommend the MWO to anyone, that is not our purpose. This Journal acts as a forum to ideas such as these to be explored, and the MWO refuses to go away, rather interest is growing and it is time we closely examine the field. Read on...

THE BECK MWO

My MWO is a Beck model. It has successfully killed a skin cancer on the back of my hand. Yet I am wondering about its resonator. There seems to be no action, it's of no use and is as if dead. I did not realize that we are having trouble establishing a real Georges Lakhovsky MWO. This [enclosed] information from Europe may help clear up our need.
D.M.
Bryte,

A good puzzle truly has many pieces, so thanks for your experiences. I never said that the Beck MWOs didn't work, but that they are not the same as Lakhovsky built. The enclosed ad was for radionic and related equipment including MWOs from Radionic Instrumentation & Research, Dorset, England. It may be that this is the radionic type of MWO originally sold by Copen Labs. I honestly don't know a thing about these MWOs, so any info would be greatly appreciated especially a schematic. Years ago I had heard them spoken of disparagingly by people who claimed that they wouldn't work because there wasn't enough power, an argument I never bought. It would seem that the more subtle and refined the energy the more beneficial it is to the human organism. The field is still open for new ideas and fresh impulses...

KLARK KENT CHECKS IN!

Just a note to say thanks for the mention in the Journal about the availability of the Golden Ratio Antennas from Super Science. And also to say that the articles on the continuing developments of the MWO are most welcome. Especially interesting is the comparison of the Beck MWO to the Lakhovsky original!

I have never seen an original Lakhovsky MWO. Only from the patent information can we really get schematic information, circuit diagrams, antenna structures, etc. As you know from all of the free energy patents circulating around, a patent wrapper and drawing is not really enough to produce the claimed device. If it were, then we should have many free energy devices in production today!

Several months ago a contact in New York made a generous offer to procure an original Lakhovsky MWO. I passed this information along to you and to others. So far no one has come forth to my knowledge with information leading to the whereabouts of an original Lakhovsky MWO. What this means is that we must grope our own way in replicating the technology first developed by Lakhovsky.

So I guess I owe Bob Beck a word of thanks because he was the one who made it possible for researchers to actually build a MWO. The information left to the world by Lakhovsky is (in my opinion) wholly inadequate to build and construct the device. The patent information is, like most patents, just not adequate to enable someone to replicate the device. And this brings up the question: Just what constitutes the original Lakhovsky MWO? We don't have an original device (and a bounty of $100,000 has so far failed to produce one). The information given in the patents doesn't really tell us enough.

Borderland Sciences can help out by continuing to publish articles about the MWO which emphasize circuit diagrams and construction details. In this way, Borderland researchers can redevelop this important technology.

I enclose a letter we received from a customer. Since the letter is of the testimonial variety, it is best not to reference it in the Journal. But the point is that even a Beck adaptation MWO can seem to produce positive experimental results!
Klark Kent, Super Science,
P.O. Box 392, Dayton, OH 45409

Glad to hear you are open to new research. Lakhovsky left much information -- it just never got translated into English. BSRF will continually publish the results of ongoing MWO research, and will always be on the lookout for the best information available. Here is the letter submitted to Klark, from a person known to us here also:

Dear Mr. Kent,
At long last I get around writing to you and share some of my experiences with the MWO I bought from you.

The unit arrived on May 17, 1990 in excellent condition. Your packaging is superb. I inspected the unit superficially but, because of some professional problems, was not able to make use of it until October.

Let me comment first on general appearance and workmanship. While obviously not industrially produced, this is a well-built high quality machine, using fine components. Having been an electronics engineer part of my life (before I became a philosopher), I feel well qualified to make that judgement.

The instructions are generally adequate but I was glad that I did order your audio cassette, done by you with a fine dramatic flair, worthy of Christopher Reeve, who does resemble you. My suggestion: include the cassette with the MWO as a matter of course.

I used the machine about thirty times, sometimes daily, sometimes skipping a day or two for 35 min. per session, mostly in the High Power Radiator mode.

I suffered over a period of about fifteen years from a mild chronic bronchitis, some coughing in the morning with some phlegm, rest of the day just fine. The medicine men told me that I "just have to live with it". Though I am 70, I judge myself being in great health.

Now, your machine has definitely improved my condition. I know, of course, that my testimonial is worthless to you, since you cannot make, and do not make, any health claims. But still I thought you might want to know.

I do have several radionics machines (De la Warr and Copen) but we all realize that this is sympathetic magic; great stuff, though!
With best wishes, faithfully yours,
NAME WITHHELD

ELECTRONIC ENTITY REMOVAL

Judging by the letters you receive and print in the BULLETIN BOARD people are still experimenting and building MWOs. It should be accepted that no two MWOs will be identical due to different components, circuits, etc. Also, Lakhovsky did not have some mystical secret that is exclusive to his particular instruments; he built one that generated his required multi-frequencies and there are no doubt many ways to effect this.

The large MWO I built in 1966 worked as well as I could expect and, although I never use it now, its circuit can be regarded as a basic Tesla coil, starting out with a 12KV neon transformer. Other circuits I have experimented with unfortunately appeared not to give the same results. Your MWO Handbook should be the starting point for would-be experimenters keeping in mind that most of the 'transistorized' models do not have enough energy to effect the desired results.

A major point that never seems to be mentioned is that the subject should be seated ON AN INSULATED PLATFORM with the antenna about six inches from the back. In this way all the cells of the body are raised to their full potential for, say, ten minutes during which each cell can absorb energy at its one resonant frequency. The subject should ground himself for an instant before stepping off the platform. Then switch off the instrument. I found this procedure mandatory to obtaining positive results and have found it odd that this point has never been mentioned in the literature.

I might also bring up the very real possibility (discussed with Riley Crabb in the past) that at least part of the positive results obtained with the MWO could be attributed to the removal of spirit entities by the high voltage involved. Regrettably, Borderland appears to skirt this issue.

As I have stated before, the Journal is an important platform for the meeting of the minds and I wish you well in your efforts.
R.M.
Nevada

Appreciate your input and valuable ideas. I'm certainly open to the possibility that the results obtained from the MWO could be attributed to the removal of spirit entities. This should also bring up the question of whether or not what is perceived as spirit entities (negative in this sense) could be also related to their potential physical anchors, i.e. virus and bacteria. Maybe we see the same thing but interpret differently.

As far as using different configurations to achieve multiple wavelengths, different arrangements will produce qualitatively different results. A French horn and a clarinet will show the same frequency for the same note played, but the difference in quality of sound should be obvious to all. The major difference between the hookup of Lakhovsky's and Beck's units makes them fall into quite separate categories. A silver bell rings finer than tin.

MWO REPORT FROM FRANCE

I finally managed to get hold of my friend (see BSRF Bulletin Board, Jan-Feb 1991) but he is very, very busy and only gave me some of the information I wanted.

Firstly, he declined to provide any basic layout of the MWO because he say no two units were the same. Apparently Lakhovsky experimented all the time and each machine was a "one-off" model, different from the others.

The origins of the MWO are in Tesla's work. A French professor, Mr. D'Arsonval took some of Tesla's equipment and modified it. Lakhovsky used D'Arsonval's work to build his machine. Tesla heard about D'Arsonval's experiments and, convinced that D'Arsonval was stealing his patents, came to France with the idea of starting legal proceedings against the professor. However, he found D'Arsonval so courteous and charming that he returned to America having dropped the idea of court action.

My friend tells me that the whole secret of the MWO is in the "diffuseur". He further indicated two sources of information, both in French. A series of books have been written and published by the Study Group "Arkall" called "L'Emergence de l'Enel" dealing with all aspects of the MWO. The second source is George Lakhovsky's book "L'oscillation cellulaire" which has been re-edited in Belgium. Unfortunately, my friend did not know the editor, so I suggest you will have to contact the American Embassy in Belgium and they should be able to give you addresses of bookshops where it can be obtained. Apparently, this book gives all the references concerning the people I have mentioned.

I'm sorry I can't be more helpful but the whole question seems to be much more complicated that I at first thought. In short, I'm afraid you will have to wade through "L'oscillation cellulaire" and the "L'Emergence de Enel" to see how Lakhovsky himself worked out the idea of the first MWO's.
R.C.
France

Thank you, your input has been most valuable and stimulating. The "diffuseur" (literally: loudspeaker or nozzle) would probably be the antenna, so the "secret" could very well be (as we have intimated over the years) in his mathematical arrangement -- a much more important consideration under Lakhovsky's antenna feed arrangement than with Beck's.

Jul-Aug 1991 JBR,
BSRF Bulletin Board

MWO CORNER

Received the Journal yesterday and was very excited about the Spiral Coil MWO. I congratulate Mr. Schad on doing a very good job and would submit the following input for future models.

1) No aluminium tubing should be used. Copper or steel tubing should be used instead. Aluminium is wholly negative. It attracts negative radiation like honey attracts files. In this connection, I heard this morning on the British radio that a "mobile home" site had just been devastated in the USA. If I'm right, "mobile home" site means trailer park, in which there is a huge quantity of Al. These are prime targets for tornadoes and all kinds of negative influence. My chief reference for this is Dr. Starbard Webb, Box 200, Orangevale, CA 95662. She has been studying the subject since 1956 and has accumulated a mass of data. Be patient for she travels a lot in the States doing seminars etc.

2) The rest is from a Tibetan lama to whom I showed the drawing. If you're wondering what a lama would know about the MWO I would say that they seem to know about everything that matters. For instance, in my own field (medicinal plants), their knowledge is not great, it is simply colossal, far beyond anything known in the West. The Lama tells me that there should be 12 coils, no more no less (Fig. 6). The "up" wire (see illustration) should be five times longer than the diameter of the coil. No doubt further information will be forthcoming when the whole of the MWO is described.

Yours sincerely,
R.C.
France.

Glad to hear of your suggestions for further research on Lou's spiral coil MWO.. His articles have stimulated a lot of work that should provide a lot of new ideas over time. Peter Lindemann has taken Lou's concept of physical vibration and is experimenting with phi ratioed antennas that are stimulated with sub-audible sound. Also, Lou called to make sure that it was clear that the antenna side facing the subject is a counterclockwise spiral. It is clockwise from the top.

Interesting thoughts on the mobile homes, which are reverse orgone accumulators. A few items congruent with your comments have showed up over the years and perhaps we can get more on that in an upcoming issue.

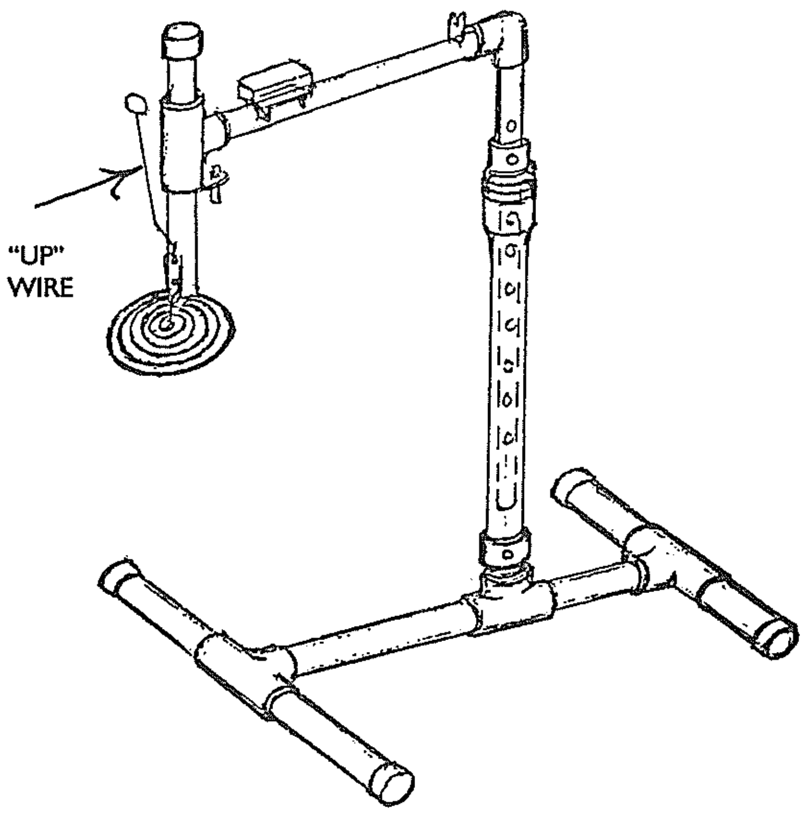

Spiral Coil Multi-Wave Oscillator

LAKHOVSKY COIL INFORMATION

The following was forwarded to us from Health Research, Box 70, Mokelumne Hill CA 95245:

On Health Research's flyer for THE HANDBOOK OF UNUSUAL ENERGIES, the artist shows the copper wire positioned incorrectly through the wood support.

I have made a number of discoveries which I even find hard to believe by using this COILED COPPER WIRE. I expect to write a booklet on my findings and you may want to represent it in your sales catalog. I have proven over and over again what I have done but, to date, have not shared anything with anyone else. I had to be ABSOLUTELY CERTAIN before I 'went public'.

When I first received the books THE SECRET OF LIFE and THE WAVES THAT HEAL from you about 4 years ago, I noted carefully just how the copper wire was COILED around the plant. Please note that the copper wire STARTS from INSIDE THE EBONITE and MOVES COUNTER CLOCKWISE and then ends ABOVE the ebonite... This is shown on page 14 of THE WAVES THAT HEAL and page 110 of THE SECRET OF LIFE.

In the past four years I have made dozens of experiments and I've gone far beyond what even Lakhovsky did or thought of. NOW, to prove my point about the proper way the copper MUST be coiled, I deliberately experimented by coiling the wire the way you have shown it and NOTHING HAPPENED in all my experiments. That's why I KNOW THAT

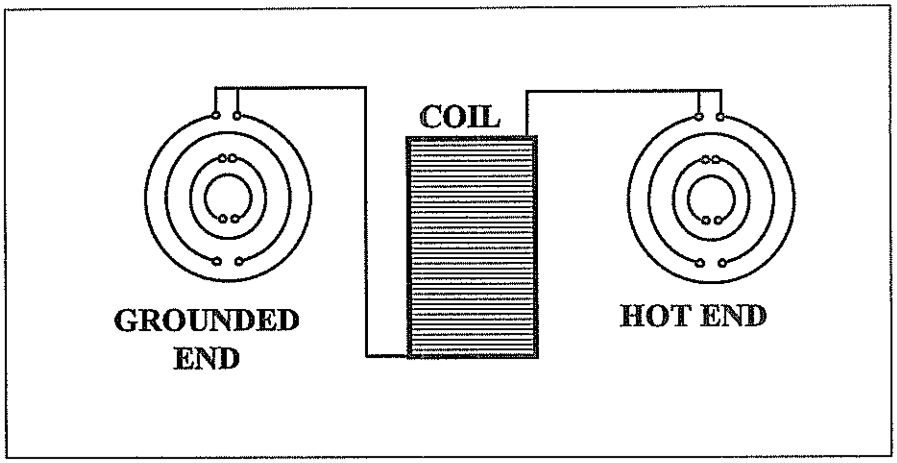

I AM RIGHT and it is imperative that you change your picture to show the correct way. I am not a scientist of a professional person, but I KNOW what I have done and you will have to take my word for it unless you want to YOURSELF, conduct the same experiments I have and PROVE it to yourself... I have not the slightest idea WHY the copper wire MUST be coiled this way but I DO KNOW THAT IT MUST.

You know as well as I that EVERYTHING in the COSMOS is perfect, exact, precise and absolute that the LAWS, FORCES, ENERGIES, POWERS, FREQUENCIES, RAYS... and all the rest function flawlessly and cannot allow any variations even for an instant. Life as we know it will cease forever if even an INSTANT change is made by whatever cause.

I would dearly love to see one of Lakhovsky's BELTS, COLLARS, BRACELETS he calls them OSCILLATING CIRCUITS, of course, and although the picture on page 63 of WAVES THAT HEAL is interesting, I am wondering if there is more to it that what this artist concept shows. I have yet to experiment with a HUMAN BODY as I will need to be very careful to use the RIGHT PERSON who will not think that I have gone off the deep end and headed for the 'funny farm'. But Lakhovsky certainly has plenty of evidence to show. If you know of anyone who has one of the ORIGINAL Lakhovsky oscillating circuits, let me know. It is imperative that the 'human' circuits MUST be done, and applied, correctly.
A.P.
Nebraska

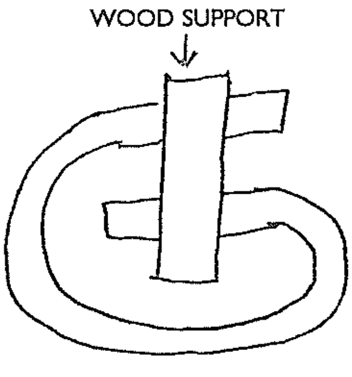

WRONG

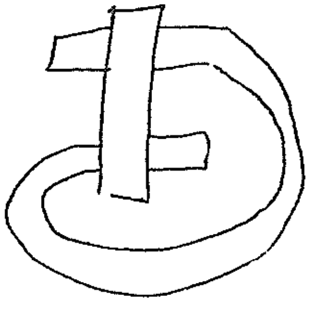

RIGHT

Nov-Dec 1991 JBR,
BSRF Bulletin Board

MWO RESEARCH

Five years ago I duplicated the MWO as best I could with my limited knowledge. I scaled up the antennas from the book WAVES THAT HEAL using copper tubing of all different sizes and soldering brass hex caps of various sizes on the ends after I rounded them off with a grinder. The outside ring was about 2 ft. across.

I was able to duplicate the exact brush discharge pattern in the picture with Tesla coil 4" diameter with about 1,000 turns of regular hook up wire. 500 turns gave a discharge of 2"-3" evenly all around the outer ring. 1,000 turns gave a discharge of 5"-6" at the bottom with less and less towards the top except for some off the two balls at the top, like in the picture.

I found it necessary to run two connections from the feed wire otherwise with one connection the spark would travel around the ring and out the other end burning anything used to suspend the ring. Two connections eliminated this problem forcing all energy towards the bottom of the ring like in the picture.

The resonator was grounded just like the Beck model to the ground end although I used two wires one going to each end of the outer ring. Even so, at this power you could draw a small spark off the outer ring and if you turned the unit off and on you could hear the resonator ring just like it was struck by a small hammer.

I used a 30mA 15,000 volt sign transformer or stronger. These tend to burn out unless you put the primaries of two in series to share the load and restrict current. Actually, I started using half burned out transformers using mid point ground and the good half for 7,500 volts which is much safer for the capacitors. I used two .005 uf 20,000 volt capacitors in parallel for .010. The stationary spark gap was air cooled with a small blower fan.

There was no spark between rings. The side effects from this unit were too strong to continue treatments so I am now working on smaller units. After 15 minutes your head would be full of mucus. The closer the antennas were to each other the stronger the effect to the point of making one very nauseous.

Also, did anyone else purchase and test Dr. Charles T. Turley's square wave sound MWO two printed antennas powered by a sound generator?
J.A.
Ohio

Thanks for your most helpful input. Would be most interested in Dr. Turley's MWO as it may very well relate to Lou Schad's spiral coil MWO, presented several issues back, which worked on physically vibrating the coil. Peter Lindemann has picked up on this and is experimenting with vibrating the antennas with sound waves and through other mechanical means. A whole new field of MWO research is opening up and it sounds like Dr. Turley is working in the same direction. Please send us any information you may have on this. Thanks.

From the March-April 1992 JBR
BSRF Bulletin Board

MWO CORNER!

Here is a nice present for you and your readers. Some tidbits about George Lakhovsky. I phoned my knowledgeable friend on Sat, 14th Dec. To my surprise not only was he at home but he invited me to spend the afternoon with him the following day.

Firstly, it is comforting to know that there are still about four original MWO's within a radius of 200 miles from Paris, and the people who have them are likely to take good care of them.

Secondly, from what I can gather, L. took very great pains with each MWO. I understand now what my friend meant when he said that each one was different. L. would calculate all sorts of parameters relating to the place where the MWO was to be used. He would then do the "fine tuning" with the help, sometimes, of an unusual cleric Abbe Breuil.

L. arrived in France around 1896. He was a friend of Tesla and when the latter came to France the first person he wanted to see was L. I don't know anything about this early period but two episodes occurred later which gave L. quite a bit of money and enabled him to finance his future research himself. 1) During the 1914 war, there was a dearth of methylated spirits in France and L. negotiated supplies, on behalf of the French government, with the Russians. 2) He invented a gadget which was used everywhere on the French railways and which brought in tidy sums in royalties. He was thus able to invite titled people to his table, the Count of this and the Duke of that and, of course, Abbe Breuil. This curious cleric was very well known in France, not for saintly endeavors on behalf of the Church but for his exploration of caves in search of traces of early man. He became the High Priest of Pre-history. However, what interests us here is that he was an extraordinary medium. For instance, L. at table, would say something like this: "Abbe, I've just thought up a new design for the MWO (passing him the drawings) what do you think of it?" And the Abbe might reply: "The antenna's too thick here, you should take 1 millimetre off and here, you are using the wrong kind of metal." In this way, L. gradually refined each model, using the Abbe's intuition to verify his own findings, until he got it as near perfect as possible.

What follows is a translation of the notes scribbled down by my friend yesterday. I offer no comments because I wouldn't know a Hertz dipole if I fell over one in the dark.

These are some of the parameters taken into consideration by L.

1) Diameter: D.

2) Thickness of the wire of the tube (not to be confused with the section).

3) Distance "e" between the two ends.

4) Orientation of the circuit with respect to the horizontal plane (29^{0} 70 N).

5) Kind of circuit. Should it be made of metal or hollow glass containing chemical salts.

6) Orientation of the distance "e" with respect to the magnetic North.

7) Study of the terrain: the mineralogical, geological, geodesic and tectonic factors.

8) Study of the nature of the surface of the circuit and whether and how it should be covered.

I see now that what I had taken for solid rods making up the antenna were in reality hollow tubes.

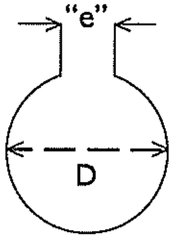

Hertz dipole

MWO antenna

I hope that you and your readers will find this information interesting. It sheds a new light on how L. operated. He used the pendulum a lot and, as we have seen, the mediumistic talents of his friend Abbe Breuil. My best wishes to yourself and to BSRF in 1992.

R.C.
France

Your input on the MWO project has been most helpful to everyone trying to find the truth of the story, and I offer my thanks and I'm sure the thanks of BSRF members the world around. Georges Lakhovsky was a complex individual whose ideas on cell regeneration have ignited many a seeking mind. It seems as though he had some rather interesting friends also, Nikola Tesla, Professor D'Arsonval, and the curious Abbe Breuil! Your friend's comments indicate some important factors in Lakhovsky's considerations: He was very concerned about the orientation towards the natural energetic lines of the earth, even considering latitude (what you have termed the horizontal plane) as a determining factor in the energetic output of his devices. He must have been aware of the interaction of the electric output with the planetary magnetic, electrostatic and etheric fields. He was also concerned about the underlying earth itself. In his **Secret of Life** he proposes his theory that cancers are more predominant in areas where the soils (such as clays) reflect the cosmic radiation back to the surface, causing interference patterns that disrupt the body's energetic flows. It is interesting to note that his antennas were constructed with all these factors involved. It is also very interesting that he relied on the pendulum and the mediumship of Abbe Breuil for his refinements. Lakhovsky was a true borderland scientist! Also of note is your point #5 mentioning the hollow glass circuit containing chemical salts. Something new to figure in our researches. We also have the open question of whether or not some of his tube antennas contained rare gases, which his MWO Tube Patent implies by its refinements in miniaturization. This has been an exciting period of discovering what Lakhovsky was really up to in his vast researches. Please read on....

A REAL LAKHOVSKY SCHEMATIC!

This schematic is from a Lakhovsky Radio Cellular Oscillator manufactured by Lakhovsky's son in Paris long after the untimely death his father in the early 1940s. It shows that the Lakhovsky's own family followed the path of the radio-cellular oscillator and not that suggested by Lakhovsky's turn to damped electrostatic power supplies in his MWO. Our many thanks go to Patrick Flanagan for providing this information to BSRF. Patrick obtained the unit, removed the seal at risk of voiding the warranty, and provided us with the circuit description you see here. Also, a special thanks to Roger Modjeski of RAM Tube Works of Santa Barbara, California for providing the pinouts and tube conversion info. As far as our extensive research notes show, this is the first real, complete schematic of an working Lakhovsky device that has been published in the US!

This unit, which fits in a small box that can be held in the hand, is a very powerful white noise generator with a treatment coil comprising of coax cable which is looped around the subject when in use. The treatment coil loop is 1 meter in length, the center wire and shield are connected. This unit will interfere with electronic communication devices up to the satellite TV range, and may be illegal to operate under FCC rules. The tube is an EL83, a French tube (Telefunken manual gives the German number as PL83) whose US equivalent is given as a 6CK6 video output pentode, but GE never made them. It is very similar in function to a 6CL6 which has a different

Lakhovsky MWO/Radio Cellular Oscillator submitted by Patrick Flanagan

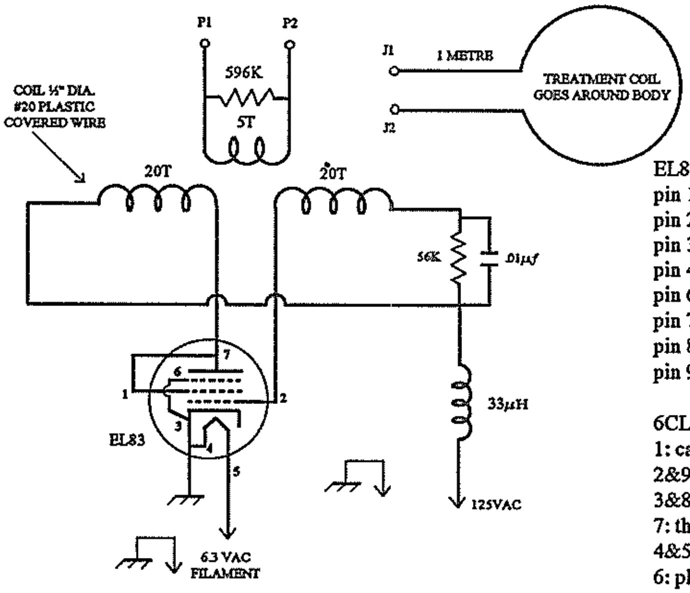

EL83/6CK6 PINOUT
pin 1: grid 2
pin 2: grid 1
pin 3: cathode
pin 4&5: filament 6.3V
pin 6: grid 3
pin 7: plate/anode
pin 8: internal shield
pin 9: not used

6CL6 PINOUT
1: cathode
2&9: first grid
3&8: second grid
7: third grid
4&5: filament
6: plate/anode

pinout. We provide the pinouts for both.

The 5 and 20 turn coils are ½ inch diameter tightly wound with #20 plastic coated wire. They are air core and all three are wound in the same right hand direction. They are spaced tightly and arranged exactly as shown. Varying the physical distance between the primary and secondary coils would vary the output strength. The resistors are ½ watt. The .01 capacitor is ceramic at 500 volts. The choke is 33 µH. The power supply is not shown, but is implied by the circuit.

We have no exact details of the interaction of this generation of Lakhovsky device with the human system, only that suggested by Lakhovsky's original RCO research and its extension through time. This data is provided for research purposes and to record yet another development in radio-biological research.

We would still be most interested in finding the schematic of one of the original MWOs pictured in Lakhovsky's various publications such as that pictured below from his book **La Terre Et Nous (The Earth and Ourselves)**, Fasquelle Editeurs, Paris, 1933, as it would provide the exact antenna feed for the concentric ring antennas and other pertinent details. Note that the antennas on this unit stand above the arm, unlike the well publicized pictures from **The Waves that Heal** and **Secret of Life.** This indicates that the protruding spheres seen in the detail on those pictures have a mechanical function, perhaps more important than their electrical function, though with Lakhovsky's methodology, he would certainly incorporate the optimum characteristics for the main antenna feeds.

We have discovered the basic construction of the Radio-Cellular Oscillator and know that Lakhovsky moved from them to the MWOs with damped electrostatic power supplies and concentric ring antennas because of the potential for thermal shortwave damage from continuous-wave radio frequencies. But then we see that Lakhovsky's son was still manufacturing RCOs after the death of his father! So we can see that the final conclusion has yet to be drawn as to exactly how the original MWOs were built, and of course we would also like to see some actual MWO Tube units that Lakhovsky patented. We also see the RCO as a potential low-cost method of researching Lakhovsky's basic theory of radio-frequency cell regeneration. This MWO Handbook contains the essential keys to all our questions and more. Much research has to be done and Borderland Sciences is willing to help any serious researcher who wishes to pursue these streams of activity.

The Journal of Borderland Research is the publication of record in the English language for these and many other research subjects. Please join in and exchange information with us. Without the help of the many borderland researchers sending us their ideas for over a quarter of a century this book would not exist. Thanks to all and best wishes on your researches -- may they be serious, straightforward, and well documented for others to learn from.

Tom Brown

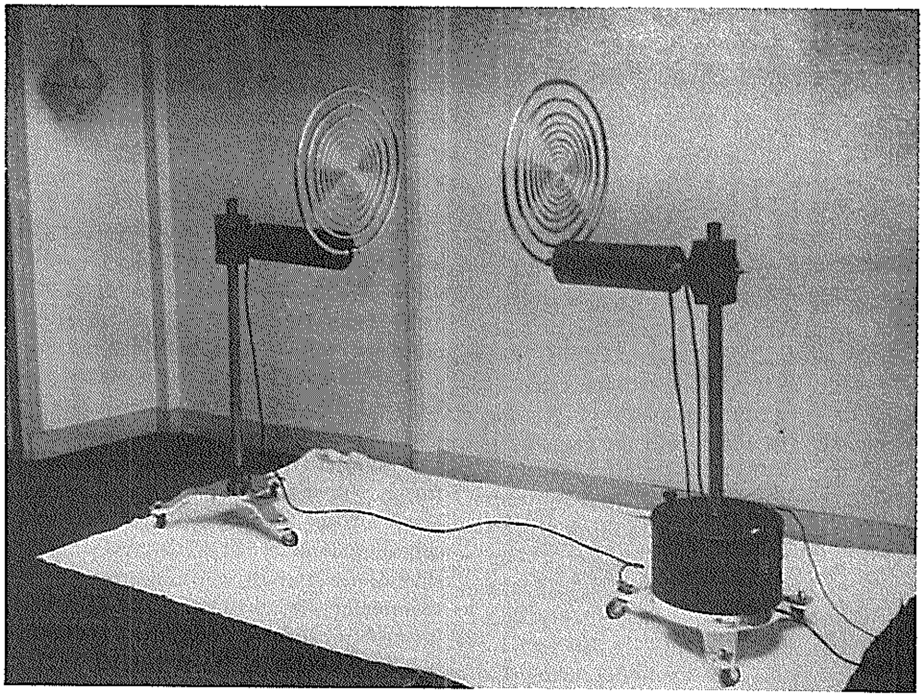

The complete Lakhovsky Multiple Wave Oscillator apparatus which creates an ambient electromagnetic field between its two *resonateurs.*

5

HIGH FREQUENCY OSCILLATORS FOR ELECTRO-THERAPEUTIC AND OTHER PURPOSES*

by Nikola Tesla

Some theoretical possibilities offered by currents of very high frequency and observations which I casually made while pursuing experiments with alternating currents, as well as the stimulating influence of the work of Hertz and of views boldly put forth by Oliver Lodge, determined me some time during 1889 to enter a systematic investigation of high frequency phenomena, and the results soon reached were such as to justify further efforts towards providing the laboratory with efficient means for carrying on the research in this particular field, which has proved itself so fruitful since. As a consequence alternators of special design were constructed and various arrangements for converting ordinary into high frequency currents perfected, both of which were duly described and are now — I assume — familiar.

One of the early observed and remarkable features of the high frequency currents, and one which was chiefly of interest to the physician, was their apparent harmlessness which made it possible to pass relatively great amounts of electrical energy through the body of a person without causing pain or serious discomfort. This peculiarity which, together with other mostly unlooked-for properties of these currents I had the honor to bring to the attention of scientific men first in an article in a technical journal in February, 1891, and in subsequent contributions to scientific societies, made it at once evident, that these currents would lend themselves particularly to electro-therapeutic uses.

With regard to the electrical actions in general, and by analogy it was reasonable to infer that the physiological effects, however complex, might be resolved in three classes. First the statical, that is, such as are chiefly dependent on the magnitude of electrical potential; second, the dynamical, that is, those principally dependent on the quality of electrical movement or current's strength through the body, and third, effects of a distinct nature due to electrical waves or oscillations, that is, impulses in which the electrical energy is alternately passing in more or less rapid succession through the static and dynamic forms.

Most generally in practice these different actions are coexistent, but by a suitable selection of apparatus and observance of conditions the experimenter may make one or other of these effects predominate. Thus he may pass through the body, or any part of the same, currents of comparatively large volume under a small electrical pressure, or he may subject the body to a high electrical pressure while the current is negligibly small, or he may put the patient under the influence of electrical waves transmitted, if desired, at considerable distance through space.

While it remained for the physician to investigate the specific actions on the organism and indicate proper methods of treatment, the various ways of applying these currents to the body of a patient suggested themselves readily to the electrician.

---

* Read at the eighth annual meeting of The American Electro-Therapeutic Association, Buffalo, N. Y., Sept. 13 to 15, 1898

As one cannot be too clear in describing a subject, a diagrammatic illustration of the several modes of connecting the circuits which I will enumerate, though obvious for the majority, is deemed of advantage.

The first and simplest method of applying the currents was to connect the body of the patient to two points of the generator, be it a dynamo or induction coil. Fig. 1 is intended to illustrate this case. The alternator G may be one giving from five to ten thousand complete vibrations per second, this number being still within the limit of practicability. The electromotive force — as measured by a hot wire instrument — may be from fifty to one hundred volts. To enable strong currents to be passed through the tissues, the terminals T T, which serve to establish contact with the patient's body should, of course, be of large area, and covered with cloth saturated with a solution of electrolyte harmless to the skin, or else the contacts are made by immersion. The regulation of the currents is best effected by means of an insulating trough A provided with two metal terminals T' T' of considerable surface, one of which, at least, should be movable. The trough is filled with water, and an electrolytic solution is added to the same, until a degree of conductivity is obtained suitable for the experiments.

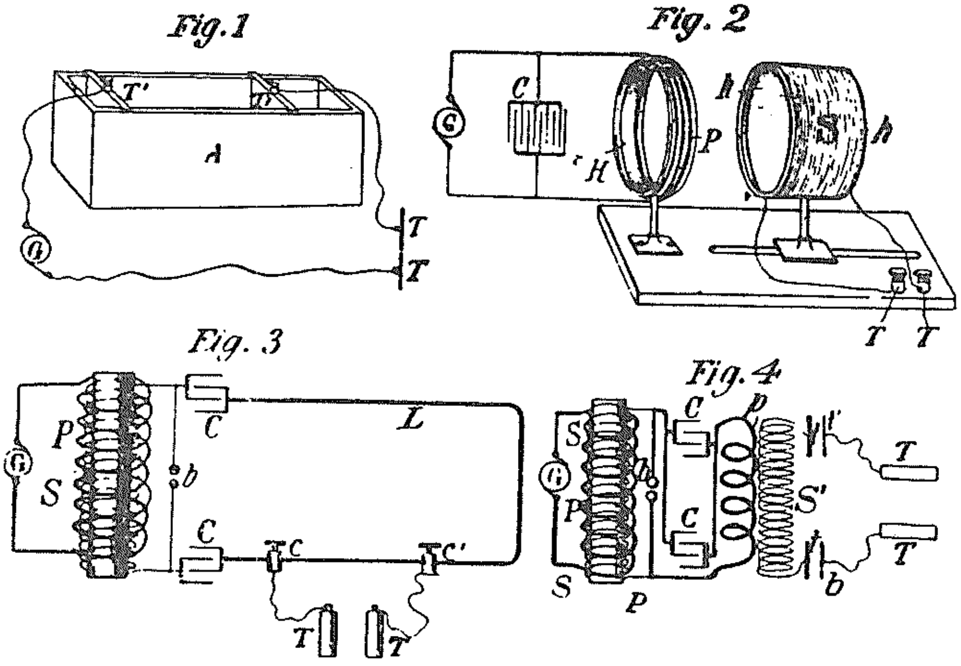

Figs. 1, 2, 3, 4.

When it is desired to use small currents of high tension, a secondary coil is resorted to, as illustrated in Fig. 2. I have found it from the outset convenient to make a departure from the ordinary ways of winding the coils with a considerable number of small turns. For many reasons the physician will find it better to provide a large hoop H of not less than, say, three feet in diameter and preferably more, and to wind upon it a few turns of stout cable P. The secondary coil S is easily prepared by taking two wooden hoops h h and joining them with stiff cardboard. One single layer of ordinary magnet wire, and not too thin at that, will be generally sufficient, the number of turns necessary for the particular use for which the coil is intended being easily ascertained by a few trials. Two plates of large surface, forming an adjustable condenser, may be used for the purpose of synchronizing the secondary with the primary circuit, but this is generally not necessary. In this manner a cheap coil is obtained, and one which cannot be easily injured. Additional advantages, however, will be found in the perfect regulation which is effected merely by altering the distance between the primary and secondary, for which adjustment provision should be made, and, furthermore, in the occurrence of harmonics which are more pronounced in such large coils of thick wire, situated at some distance from the primary.

The preceding arrangements may also be used with alternating or interrupted currents of low frequency, but certain peculiar properties of high frequency currents make it possible to apply the latter in ways entirely impracticable with the former.

One of the prominent characteristics of high frequency or, to be more general, of rapidly varying currents, is that they pass with difficulty through stout conductors of high self-induction. So great is the obstruction which self-induction offers to their passage that it was found practicable, as shown in the early experiments to which reference has been made, to maintain differences of potential of many thousands of volts between two points — not more than a few inches apart — of a thick copper bar of inappreciable resistance. This observation naturally suggested the disposition illustrated in Fig. 3. The source of high frequency impulses is in this instance a familiar type of transformer which may be supplied from a generator G of ordinary direct or alternating currents. The transformer comprises a primary P, a secondary S, two condensers C C which are joined in series, a loop or coil of very thick wire L and a circuit interrupting device or break b. The currents are derived from the loop L by two contacts c c', one or both of which are capable of displacement along the wire L. By varying the distance between these contacts, any difference of potential, from a few volts to many thousands, is readily obtained on the terminals or handles T T. This mode of using the currents is entirely safe and particularly convenient, but it requires a very uniform working of the break b employed for charging and discharging the condenser.

Another equally remarkable feature of high frequency impulses was found in the facility with which they are transmitted through condensers, moderate electromotive forces and very small capacities being required to enable currents of considerable volume to pass. This observation made it practicable to resort to a plan such as indicated in Fig. 4. Here the connections are similar to those shown in the preceding case, except that the condensers C C are joined in parallel. This lowers the frequency of the currents, but has the advantage of allowing the working with a much smaller difference of potential on the terminals of the secondary S. Since the latter is the chief item of expense of such apparatus and since its price rapidly increases with the number of turns required, the experimenter will find it generally cheaper to make a sacrifice in the frequency, which, however, will be high enough for most purposes. However, he only needs to reduce proportionately the number of turns or the length of primary p to obtain the same frequency as before, but the economy of transformation will be somewhat reduced in so doing and the break b will require more attention. The secondary S' of the high frequency coil has two metal plates t t of considerable surface connected to its terminals, and the current for use is derived from two similar plates t' t' in proximity to the former. Both the tension and volume of the currents taken from terminals T T may be easily regulated and in a continuous manner by simply varying the distance between the two pairs of plates t t and t' t' respectively.

A facility is also afforded in this disposition for raising or lowering the potential of one of the terminals T, irrespective of the changes produced on the other terminal, this making it possible to cause a stronger action on one or other part of the patient's body.

The physician may find it for some or other reason convenient to modify the arrangements in Figs. 2, 3 and 4 by connecting one terminal of the high frequency source to the ground. The effects will be in most respects the same, but certain peculiarities will be noted in each case. When a ground connection is made it may be of some consequence which of the terminals of the secondary is connected to the ground, as in high frequency discharges the impulses of one direction are generally preponderating.

Among the various noteworthy features of these currents there is one which lends itself especially to many valuable uses. It is the facility which they afford for conveying large amounts of electrical energy to a body entirely insulated in space. The practicability of this method od energy transmission, which is already receiving useful applications and promises to become of great importance in the near future, has helped to dispel the old notion assuming the necessity of a return circuit for the conveyance of electrical energy in any considerable amount. With novel appliances we are enabled to pass through a wire, entirely insulated on one end, currents strong enough to fuse it, or to convey through the wire any amount of energy to an insulated body. This mode of applying high frequency currents in medical treatment appears to me to offer the greatest possibilities at the hands of the physician. The effects produced in this manner possess features entirely distinct from those observed when the currents are applied in any of the before mentioned or similar ways.

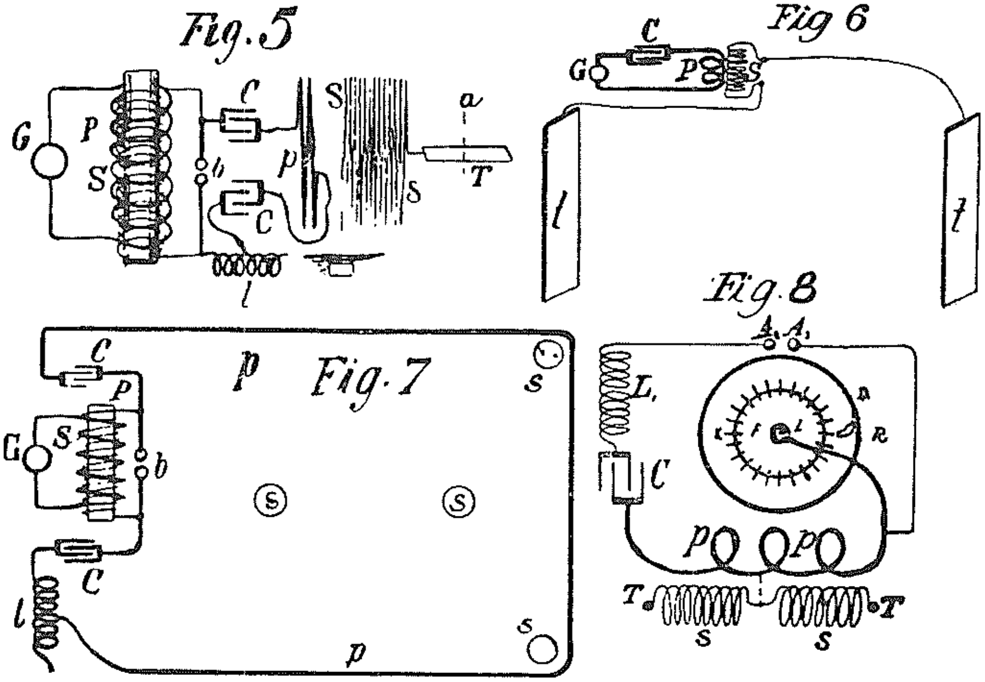

Figs. 5, 6, 7, 8.

The circuit connections as usually made are illustrated schematically in Fig. 5, which, with reference to the diagrams before shown, is self-explanatory. The condensers C C, connected in series, are preferably charged by a step-up transformer, but a high frequency alternator, static machine, or a direct current generator, if it be of sufficiently high tension to enable the use of small condensers, may be used with more or less success. The primary p, through which the high frequency discharges of the condensers are passed, consists of very few turns of cable of as low resistance as possible, and the secondary s, preferably at some distance from the primary to facilitate free oscillation, has one of its ends — that is the one which is nearer to the primary — connected to the ground, while the other end leads to on insulated terminal T, with which the body of the patient is connected. It is of importance in this case to establish synchronism between the oscillations in the primary and secondary circuits p and s respectively. This will be as a rule best effected by varying the self-induction of the circuit including the primary loop or coil p, for which purpose an adjustable self-induction e is provided; but in cases when the electromotive force of the generator is exceptionally high, as when a static machine is used and a condenser consisting of merely two plates offers sufficient capacity, it will be simpler to attain the same object by varying the distance of the plates.

The primary and secondary oscillations being in close synchronism, the points of highest potential will be on a part of terminal T, and the consumption of energy will occur chiefly there. The attachment of the patient's body to the terminal will in most cases very materially affect the period of oscillation in the secondary, making it longer, and a readjustment of the primary circuit will have to be made in each case to suit the capacity of the body connected with terminal T. Synchronism should always be preserved, and the intensity of the action varied by moving the secondary coil to or from the primary, as may be desired. I know of no method which would make it possible to subject the human body to such excessive electrical pressures as are practicable with this, or of one which would enable the conveying to and giving off from the body without serious injury amounts of electrical energy approximating even in a remote degree those which are entirely practicable when this manner of applying the energy is resorted to. This is evidently due to the fact that action is chiefly superficial, the largest possible section being offered to the transfer of the current, or, to say more correctly, of the energy. With a very rapidly and smoothly working break I would not think it impossible to convey to the body of a person and to give off into the space energy at the rate of several horse power with impunity, while a small part of this amount applied in other ways could not fail to produce injury.

When a person is subjected to the action of such a coil, the proper adjustments being carefully observed, luminous streams are seen in the dark issuing from all parts of the body. These streams are short and of delicate texture when the number of breaks is very great and the action of the device b (Fig. 5) free of any irregularities, but when the number of breaks is small or the action of the device imperfect, long and noisy streams appear which cause some discomfort. The physiological effects produced with apparatus of this kind may be graduated from a hardly perceptible action when the secondary is at a great distance from the primary, to a most violent one when both coils are placed at a small distance. In the latter case only a few seconds are sufficient to cause a feeling of warmth all over the body, and soon after the person perspires freely. I have repeatedly, in demonstrations to friends, exposed myself longer to the action of the oscillations, and each time, after the lapse of an hour or so, an immense fatigue, of which it is difficult to give an idea, would take hold of me. It was greater than I experienced on some occasions after the most straining and prolonged bodily exertion. I could scarcely make a step and could keep the eyes open only with the greatest difficulty. I slept soundly afterward, and the after-effect was certainly beneficial, but the medicine was manifestly too strong to be used frequently.

One should be cautious in performing such experiments for more than one reason. At or near the surface of the skin, where the most intense action takes place, various chemical products are formed, the chief being ozone and nitrogen compounds. The former is itself very destructive, this feature being illustrated by the fact that the rubber insulation of a wire is destroyed so quickly as to make the use of such insulation entirely impracticable. The compounds of nitrogen, when moisture is present, consist largely of nitric acid which might, by excessive application, prove hurtul to the skin. So far, I have not noted injuries which could be traced directly to this cause, though on several occasions burns were produced in all respects similar to those which were later observed and attributed to the Röntgen rays. This view is seemingly being abandoned, having not been substantiated by experimental facts, and so also is the notion that these rays are transverse vibrations. But while investigation is being turned in what appears to be the right direction, scientific men are still at sea. This state of things impedes the progress of the physicist in these new regions and makes the already hard task of the physician still more difficult and uncertain.

One or two observations made while pursuing experiments with the apparatus described might be found as deserving mention here. As before stated, when the oscillations in the primary and secondary circuits are in synchronism, the points of highest potential are on some portion of the terminal T. The synchronism being perfect and the length of the secondary coil just equal to one-quarter of the wave length, these points will be exactly on the free end of terminal T, that is, the one situated farthest from the end of the wire attached to the terminal. If this be so and if now the period of the oscillations in the primary be shortened, the points of highest potential will recede towards the secondary coil, since the wave-length is reduced and since the attachment of one end of the secondary coil to the ground determines the position of the nodal points, that is, the points of least potential. Thus, by varying the period of vibration of the primary circuit in any manner, the points of highest potential may be shifted accordingly along the terminal T, which has been shown, designedly, long to illustrate this feature. The same phenomenon is, of course, produced if the body of a patient constitutes the terminal, and an assistant may by the motion of a handle cause the points of highest potential to shift along the body with any speed he may desire. When the action of the coil is vigorous, the region of highest potential is easily and unpleasantly located by the discomfort or pain experienced, and it is most curious to feel how the pain wanders up and down, or eventually across the body, from hand to hand, if the connection to the coil is accordingly made — in obedience to the movement of the handle controlling the oscillations. Though I have not observed any specific action in experiments of this kind, I have always felt that this effect might be capable of valuable use in electro-therapy.

Another observation which promises to lead to much more useful results is the following: As before remarked, by adopting the method described, the body of a person may be subjected without danger to electrical pressures vastly in excess of any producible by ordinary apparatus, for they may amount to several million volts, as has been shown in actual practice. Now, when a conducting body is electrified to so high a degree, small particles, which may be adhering firmly to its surface, are torn off with violence and thrown to distances which can be only conjectured. I find that not only firmly adhering matter, as paint, for instance, is thrown off, but even the particles of the toughest metals are torn off. Such actions have been thought to be restricted to a vacuous inclosure, but with a powerful coil they occur also in the ordinary atmosphere. The facts mentioned would make it reasonable to expect that this extraordinary effect which, in other ways, I have already usefully applied, will likewise prove to be of value in electro-therapy. The continuous improvement of the instruments and the study of the phenomenon may shortly lead to the establishment of a novel mode of hygienic treatment which would permit an instantaneous cleaning of the skin of a person, simply by connecting the same to, or possibly, by merely placing the person in the vicinity of a source of intense electrical oscillations, this having the effect of throwing off, in a twinkle of the eye, dust or particles of any extraneous matter adhering to the body. Such a result brought about in a practicable manner would, without doubt, be of incalculable value in hygiene and would be an efficient and time-saving substitute for a water bath, and particularly appreciated by those whose contentment consists in undertaking more than they can accomplish.

High frequency impulses produce powerful inductive actions and in virtue of this feature they lend themselves in other ways to the uses of the electro-therapeutist. These inductive effects are either electrostatic or electrodynamic. The former diminish much more rapidly with the distance — with the square of the same — the latter are reduced simply in proportion to the distance. On the other hand, the former grow with the square of intensity of the source, while the latter increase in a simple proportion with the intensity. Both of these effects may be utilized for establishing a field of strong action extending through considerable space, as through a large hall, and such an arrangement might be suitable for use in hospitals or institutions of this kind, where it is desirable to treat a number of patients at the same time.

Fig. 6 illustrates the manner, as I have shown it originally, in which such a field of electrostatic action is established. In this diagram G is a generator of currents of very high frequency, C a condenser for counteracting the self-induction of the circuit which includes the primary P of an induction coil, the secondary S of which has two plates t t of large surface connected to its terminals. Well known adjustments being observed, a very strong action occurs chiefly in the space between the plates, and the body of a person is subjected to rapid variations of potential and surgings of current, which produce, even at a great distance, marked physiological effects. In my first experiments I used two metal plates as shown, but later I found it preferable to replace them by two large hollow spheres of brass covered with wax of a thickness of about two inches. The cables leading to the terminals of the secondary coil were similarly covered, so that any of them could be approached without danger of the insulation breaking down. In this manner the unpleasant shocks, to which the experimenter was exposed when using the plates, were prevented.

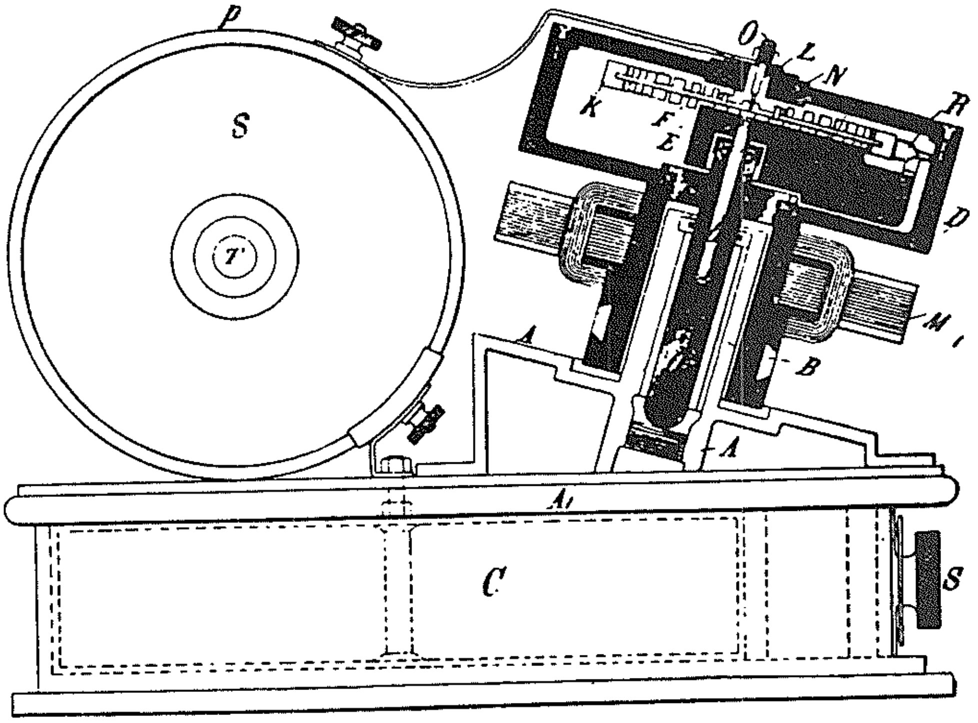

Fig. 9.

In Fig. 7 a plan for similarly utilizing the dynamic inductive effects of high frequency currents is illustrated. As the frequencies obtainable from an alternator are not as high as is desired, conversion by means of condensers is resorted to. The diagram will be understood at a glance from the foregoing description. It only need be stated that the primary p, through which the condensers are made to discharge, is formed by a thick stranded cable of low self-induction and resistance, and passes all around the hall. Any number of secondary coils s s s, each consisting generally of a single layer of rather thick wire, may be provided. I have found it practicable to use as many as one hundred, each being adjusted for a definite period and responding to a particular vibration passed through the primary. Such a plant I have had in use in my laboratory since 1892, and many times it has contributed to the pleasure of my visitors and also proved itself of practical utility. On a latter occasion I had the pleasure of entertaining some of the members with experiments of this kind, and this opportunity I cannot let pass without expressing my thanks for the interest which was awakened in me by their visit, as well as for the generous acknowledgment of the courtesy by the Association. Since that time my apparatus has been very materially improved, and now I am able to create a field of such intense induction in the laboratory that a coil three feet in diameter, by careful adjustment, will deliver energy at the rate of one-quarter of a horse power, no matter where it is placed within the area inclosed by the primary loops. Long sparks, streamers and all other phenomena obtainable with induction coils are easily producible anywhere within the space, and such coils, though not connected to anything, may be utilized exactly as ordinary coils, and what is still more remarkable, they are more effective. For the past few years I have often been urged to show experiments in public, but, though I was desirous to comply with such requests, pressing work has so far made it impossible. These advances have been the result of slow but steady improvement in the details of the apparatus which I hope to be able to describe connectedly in the near future.

However remarkable the electrodynamic inductive effects, which I have mentioned, may appear, they may be still considerably intensified by concentrating the action upon a very small space. It is evident that since, as before stated, electromotive forces of many thousand volts are maintained between two points of a conducting bar or loop only a few inches long, electromotive forces of approximately the same magnitude will be set up in conductors situated near by. Indeed, I found that it was practicable in this manner to pass a discharge through a highly exhausted bulb, although the electromotive force required amounted to as much as ten or twenty thousand volts, and for a long time I followed up experiments in this direction with the object of producing light in a novel and more economical way. But the tests left no doubt that there was great energy consumption attendant to this mode of illumination, at least with the apparatus I had then at command, and, finding another method which promised a higher economy of transformation, my efforts turned in this new direction. Shortly afterward (some time in June, 1891), Prof. J. J. Thomson described experiments which were evidently the outcome of long investigation, and in which he supplied much novel and interesting information, and this made me return with renewed zeal to my own experiments. Soon my efforts were centered upon producing in a small space the most intense inductive action, and by gradual improvement in the apparatus I obtained results of a surprising character. For instance, when the end of a heavy bar of iron was thrust within a loop powerfully energized, a few moments were sufficient to raise the bar to a high temperature. Even heavy lumps of other metals were heated as rapidly as though they were placed in a furnace. When a continuous band formed of a sheet of tin was thrust into the loop, the metal was fused instantly, the action being comparable to an explosion, and no wonder, for the frictional losses accumulated in it at the rate of possibly ten horse power. Masses of poorly conducting material behaved similarly, and when a highly exhausted bulb was pushed into the loop, the glass was heated in a few seconds nearly to the point of melting.

When I first observed these astonishing actions, I was interested to study their effects upon living tissues. As may be assumed, I proceeded with all the necessary caution, and well I might, for I had the evidence that in a turn of only a few inches in diameter an electromotive force of more than ten thousand volts was produced, and such high pressure would be more than sufficient to generate destructive currents in the tissue. This appeared all the more certain as bodies of comparatively poor conductivity were rapidly heated and even partially destroyed. One may imagine my astonishment when I found that I could thrust my hand or any other part of the body within the loop and hold it there with impunity. More than on one occasion, impelled by a desire to make some novel and useful observation, I have willingly or unconsciously performed an experiment connected with some risk, this being scarcely avoidable in laboratory experience, but have always believed, and do so now, that I have never undertaken anything in which, according to my own estimation, the chances of being injured were so great as when I placed my head within the space in which such terribly destructive forces were at work. Yet I have done so, and repeatedly, and have felt nothing. But I am firmly convinced that there is great danger attending such experiment, and some one going just a step farther than I have gone may be instantly destroyed. For, conditions may exist similar to those observable with a vacuum bulb. It may be placed in the field of the loop, however intensely energized, and so long as no path for the current is formed, it will remain cool and consume practically no energy. But the moment the first feeble current passes, most of the energy of the oscillations rushes to the place of consumption. If by any action whatever, a conducting path were formed within the living tissue or bones of the head, it would result in the instant destruction of these and death of the foolhardy experimenter. Such a method of killing, if it were rendered practicable, would be absolutely painless. Now, why is it that in a space in which such violent turmoil is going on living tissue remains uninjured? One might say the currents cannot pass because of the great self-induction offered by the large conducting mass. But this it cannot be, because a mass of metal offers a still higher self-induction and is heated just the same. One might argue the tissues offer too great a resistance. But this again cannot be the reason, for all evidence shows that the tissues conduct well enough, and besides, bodies of approximately the same resistance are raised to a high temperature. One might attribute the apparent harmlessness of the oscillations to the high specific heat of the tissue, but even a rough quantitative estimate from experiments with other bodies shows that this view is untenable. The only plausible explanation I have so far found is that the tissues are condensers. This only can account for the absence of injurious action. But it is remarkable that, as soon as a heterogeneous circuit is constituted, as by taking in the hands a bar of metal and forming a closed loop in this manner, the passage of the currents through the arms is felt, and other physiological effects are distinctly noted. The strongest action is, of course, secured when the exciting loop makes only one turn, unless the connections take up a considerable portion of the total length of the circuit, in which case the experimenter should settle upon the least number of turns by carefully estimating what he loses by increasing the number of turns, and what he gains by utilizing thus a greater proportion of the total length of the circuit. It should be borne in mind that, when the exciting coil has a considerable number of turns and is of some length, the effects of electrostatic induction may preponderate, as there may exist a very great difference of potential — a hundred thousand volts or more — between the first and last turn. However, these latter effects are always present even when a single turn is employed.

When a person is placed within such a loop, any pieces of metal, though of small bulk, are perceptibly warmed. Without doubt they would be also heated — particularly if they were of iron — when embedded in living tissue, and this suggests the possibility of surgical treatment by this method. It might be possible to sterilize wounds, or to locate, or even to extract metallic objects, or to perform other operations of this kind within the sphere of the surgeon's duties in this novel manner.

Most of the results enumerated, and many others still more remarkable, are made possible only by utilizing the discharges of a condenser. It is probable that but a very few — even among those who are working in these identical fields — fully appreciate what a wonderful instrument such a condenser is in reality. Let me convey an idea to this effect. One may take a condenser, small enough to go in one's vest pocket, and by skilfully using it he may create an electrical pressure vastly in excess — a hundred times greater if necessary — than any producible by the largest static machine ever constructed. Or, he may take the same condenser and, using it in a different way, he may obtain from it currents against which those of the most powerful welding machine are utterly insignificant. Those who are imbued with popular notions as to the pressure of static machines and currents obtainable with a commercial transformer, will be astonished at this statement — yet the truth of it is easy to see. Such results are obtainable, and easily, because the condenser can discharge the stored energy in an inconceivably short time. Nothing like this property is known in physical science. A compressed spring, or a storage battery, or any other form of device capable of storing energy, cannot do this; if they could, things udreamt of at present might be accomplished by their means. The nearest approach to a charged condenser is a high explosive, as dynamite. But even the most violent explosion of such a compound bears no comparison with the discharge or explosion of a condenser. For, while the pressures which are produced in the detonation of a chemical compound are measured in tens of tons per square inch, those which may be caused by condenser discharges may amount to thousands of tons per square inch, and if a chemical could be made which would explode as quickly as a condenser can be discharged under conditions which are realizable — an ounce of it would quite certainly be sufficient to render useless the largest battleship.

That important realizations would follow from the use of an instrument possessing such ideal properties I have been convinced since long ago, but I also recognized early that great difficulties would have to be overcome before it could replace less perfect implements now used in the arts for the manifold transformations of electrical energy. These difficulties were many. The condensers themselves, as usually manufactured, were inefficient, the conductors wasteful, the best insulation inadequate, and the conditions for the most efficient conversion were hard to adjust and to maintain. One difficulty, however, which was more serious than the others, and to which I called attention when I first described this system of energy transformation, was found in the devices necessarily used for controlling the charges and discharges of the condenser. They were wanting in efficiency and reliability and threatened to prove a decided drawback, greatly restricting the use of the system and depriving it of many valuable features. For a number of years I have tried to master this difficulty. During this time a great number of such devices were experimented upon. Many of them promised well at first, only to prove inadequate in the end. Reluctantly, I came back upon an idea on which I had worked long before. It was to replace the ordinary brushes and commutator segments by fluid contacts. I had encountered difficulties then, but the intervening years in the laboratory were not spent in vain, and I made headway. First it was necessary to provide for a circulation of the fluid, but forcing it through by a pump proved itself impractical. Then the happy idea presented itself to make the pumping device an integral part of the circuit interrupter, inclosing both in a receptacle to prevent oxydation. Next some simple ways of maintaining the circulation, as by rotating a body of mercury, presented themselves. Then I learned how to reduce the wear and losses which still existed. I fear that these statements, indicating how much effort was spent in these seemingly insignificant details will not convey a high idea of my ability, but I confess that my patience was taxed to the utmost. Finally, though, I had the satisfaction of producing devices which are simple and reliable in their operation, which require practically no attention and which are capable of effecting a transformation of considerable amounts of energy with fair economy. It is not the best that can be done, by any means, but it is satisfactory, and I feel that the hardest task is done.

The physician will now be able to obtain an instrument suitable to fulfil many requirements. He will be able to use it in electro-therapeutic treatment in most of the ways enumerated. He will have the facility of providing himself with coils such as he may desire to have for any particular purpose, which will give him any current or any pressure he may wish to obtain. Such coils will consist of but a few turns of wire, and the expense of preparing them will be quite insignificant. The instrument will also enable him to generate Röntgen rays of much greater power than obtainable with ordinary apparatus. A tube must still be furnished by the manufacturers which will not deteriorate and which will allow to concentrate larger amounts of energy upon the electrodes. When this is done, nothing will stand in the way of an extensive and efficient application of this beautiful discovery which must ultimately prove itself of the highest value, not only at the hands of the surgeon, but also of the electro-therapist and, what is most important, of the bacteriologist.

To give a general idea of an instrument in which many of the latter improvements are embodied, I would refer to Fig. 9, which illustrates the chief parts of the same in side elevation and partially in vertical cross-section. The arrangement of the parts is the same as in the form of instrument exhibited on former occasions, only the exciting coil with the vibrating interrupter is replaced by one of the improved circuit breakers to which reference has been made.

This device comprises a casting A with a protruding sleeve B, which in a bushing supports a freely rotatable shaft a. The latter carries an armature within a stationary field magnet M and on the top, a hollow iron pulley D, which contains the break proper. Within the shaft a, and concentrically with the same, is placed a smaller shaft b, likewise freely movable on ball-bearings and supporting a weight E. This weight being on one side and the shafts a and b inclined to the vertical, the weight remains stationary as the pulley is rotated. Fastened to the weight E is a device R in the form of a scoop with very thin walls, narrow on the end nearer to the pulley and wider on the other end. A small quantity of mercury being placed in the pulley and the latter rotated against the narrow end of the scoop, a portion of the fluid is taken up and thrown in a thin and wide stream towards the centre of the pulley. The top of the latter is hermetically closed by an iron washer, as shown, this washer supporting on a steel rod L a disk F of the same metal provided with a number of thin contact blades K. The rod L is insulated by washers N from the pulley, and for the convenience of filling in the mercury a small screw o is provided. The bolt L forming one terminal of the circuit breaker is connected by a copper strip to the primary p. The other end of the primary coil leads to one of the terminals of the condenser C, contained in a compartment of a box A, another compartment of the same being reserved for switch S and terminals of the instrument. The other terminal of the condenser is connected to the casting A and through it to pulley D. When the pulley is rotated, the contact blades K are brought rapidly in and out of contact with the stream of mercury, thus closing and opening the circuit in quick succession. With such a device it is easy to obtain ten thousand makes and breaks per second and even more. The secondary a is made of two separate coils and so arranged that it can be slipped out, and a metal strip in its middle connects it to the primary coil. This is done to prevent the secondary from breaking down when one of the terminals is overloaded, as it often happens in working Röntgen bulbs. This form of coil will withstand a very much greater difference of potential than coils as ordinarily constructed.

The motor has both field and armature built of plates, so that it can be used on alternating as well as direct current supply circuits, and the shafts are as nearly as possible vertical, so as to require the least care in oiling. Thus, the only thing which really requires some attention is the commutator of the motor, but where alternating currents are always available, this source of possible trouble is easily done away with.

The circuit connections of the instrument have been already shown and the mode of operation explained in periodicals. The usual manner of connecting is illustrated in Fig. 8, in which A_2 A_2 are the terminals of the supply circuit, L, a self-induction coil for raising the pressure, which is connected in series with condenser C and primary P P. The remaining letters designate the parts correspondingly marked in Fig. 9 and will be understood with reference to the latter.

THE MAGNETIC-ARC INSTRUMENT OF S.S. KNIGHT

by Mark L. Gallert

Reprinted from

NEW LIGHT ON THERAPEUTIC ENERGIES, 1976

Mr. Knight, a chemical and electrical engineer of Walnut Creek, California, was impressed with the severity and persistence of pain suffered in certain types of cancer. Realizing that pain is a sensation representative of an electrical phenomenon of excessive nerve-current flow, he decided to experiment to see if there was some way of interfering with that nerve flow through artificial impulses, in the effort to stop or alleviate the pain.

Towards this end, he developed an instrument combining several features of electro-therapy:

1. A wave pulsed at 90 times per minute, taking advantage of data obtained from experience with the older types Faradic machines.
2. A rest interval exceeding 75% of the duration of each pulsing cycle - enhances the beneficial effect.
3. Breaking the treatment by mechanical opening of contacts in such a way as to create an arc at each of the pulsing intervals. An arc produces a very wide band of frequencies, from those of very low frequency, to the highest that can be measured. The multiplicity of frequencies seems to play a part in masking or interrupting the sensation of pain.
4. A unique feature of the Knight instrument is that the arc is broken in a strong magnetic field produced by a large, powerful, permanent magnet. This introduces what is called the "Zeeman effect", that is, a magnification of the arc impulses, and a spreading of frequencies.
5. Use of very low power output - below 1/100 of a watt, so that there will be no perceptible heating of tissue when the complex of impulses from the instrument are fed to the patient through electrodes applied to the skin.

Experiments on cases which had been declared terminal cancer by medical doctors, showed that this device not only brought outstanding relief of pain in the majority of instances, but also produced some surprising "remissions" characterized by shrinking of tumour masses, and restoration of general health to the patient.

Dr. C.P. Bryant of Seattle, Washington, who used these instruments extensively in his practice over a period of years, declared that with their use he was able to restore to health 75% of all cancer cases which had not undergone surgery, X-ray or radium, and 25% of those cancer cases which had been subjected to surgery, X-ray or radium.

The Knight family has never represented these instruments as a "cancer cure". The original aim was to find some way to relieve the pain of cancer. Reports from doctors of apparent cures through use of the instruments were received with interest, as well as reports of benefits from the application of the instrument to conditions other than cancer.

Mr. S.S. Knight, a man with pronounced humanitarian ideals, made the circuit of the instrument available to all inquirers and encouraged the building of instruments by anyone who wished to copy his circuit and specifications. He was keenly disappointed when reports reached him of failure, or lack of therapeutic effect from instruments constructed by others. It developed that adjustments were quite critical, particularly concerning the making and breaking of the arc. Mr. Knight could tell by looking at the arc and listening to it, when it was "right". It was difficult to develop objective criteria for adjustment of the equipment.

A puzzling circumstance, was that the maximum effect from Mr. Knight's instruments was obtained when they were new; after a few weeks of use the effect began to diminish; after some months it was perceptibly weaker, and after a year or more, of very little use. Several years ago, the reason for the lessening of effect was determined. The isolation transformer used to block the direct flow of the arc frequencies from the electric line supply (an isolation required to eliminate radio interference) was being saturated by the magnetism from the large magnet surrounding the arc. The more saturated the transformer became, the less electricity it passed. The equipment was re-designed to place the transformer out of range of the field of the magnet.

The instrument can be classed with experimental devices, as there have not been sufficient tests made under controlled conditions to establish with certainty the percentage of cases of various types of cancer that can be alleviated, arrested, and perhaps cured, but the tests to date indicate at least that this is a promising approach, which should be explored further.

An imitation instrument has been marketed by others, using an electro-magnet instead of a permanent magnet. Mr. Knight was opposed to this substitution, having a number of reasons for believing that the permanent magnet was preferable. Other changes have also been made in constructing these imitation instruments, including the type of electrodes. So far as your compiler has been able to determine, the instruments using electro-magnets have some value in relieving pain in a variety of conditions, but the effect on cancer seems to be slower and less pronounced than that obtained from the genuine Knight instruments when the latter are functioning properly and are in adjustment.

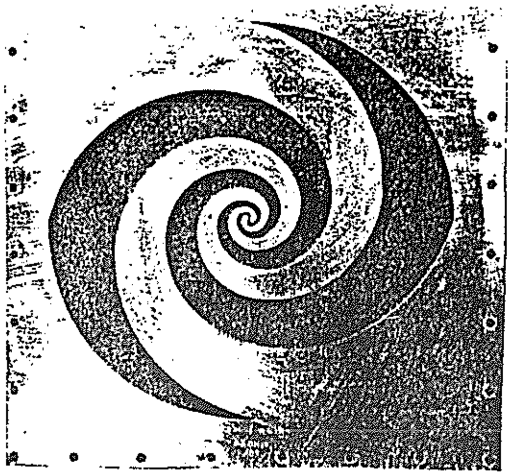

FIG. 1—Balanced planar equiangular or logarithmic spiral-slot antenna (left) radiates a broad circularly polarized beam on each side of, and perpendicular to, the plane of the antenna

electronics
April 20, 1962

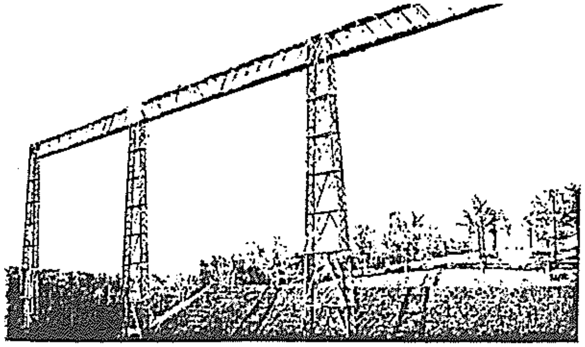

FIG. 3—The University of Illinois radio telescope. The focal line array, 153 feet above the 400-by-600-foot parabolic reflector, is visible below the wooden truss

Frequency-Independent Antennas

SURVEY OF DEVELOPMENT

Antennas now provide continuous coverage of large portions of the frequency spectrum. These very-wide-band antennas offer a challenge to the systems engineer to design equipment that will match bandwidths now available

By JOHN D. DYSON
University of Illinois,
Urbana, Illinois

THE NECESSITY of providing continuous coverage of large portions of the electromagnetic frequency spectrum for defense purposes has exerted great pressure on the antenna engineer to produce antennas that could be used over wider and wider bands of frequencies. Such requirements have led to an important breakthrough in the antenna field. The conception of the frequency-independent antennas at the University of Illinois and their development at this and at other laboratories has produced advances in the state of the art that now puts pressure on the electronics and system engineer to design equipment that will match the available antenna bandwidths. The purpose of this survey is to outline the development of these antennas.

As of 1945, antenna research and development had led to a few structures that operated over a 2 to 1 or slightly greater range of frequencies, with acceptable radiation and impedance characteristics. In a majority of cases, it was the radical changes in the radiation pattern with changes in operating frequency which limited the usefulness of antennas to somewhat less than this 2 to 1 range, since antennas had been developed with an input impedance that remained relatively constant with a change in frequency.

The discone antenna[1] introduced in 1946 was one of the first antennas that could be used over approximately a 4-to-1 range of frequencies. Advances during the next 7 or 8 years were limited, mainly, to increasing the bandwidth of existing antennas by a few percent. The one exception during this period was a modified helical antenna[2, 3, 4]. These unbalanced conical helical structures, fed against a ground plane, had acceptable pattern bandwidths of 4 or 5 to 1.

The spiral antenna introduced in 1953 was a major step forward in wide band antenna design. This construction consists of two flat, narrow, constant-width metallic arms, wound in an Archimedean spiral.[5] Fed at the center in a balanced manner, it radiates a broad circularly polarized lobe on each side of the plane of the antenna. Subsequent development of this configuration produced forms that had stable pattern and impedance characteristics over bandwidths up to 10 to 1.[6-10]

Arrays of the Archimedes spiral antenna provide variable polarization of the radiated field[11] and beam scanning by simple rotation of the antennas.[12-13]

In 1954 a major breakthrough occurred when V. H. Rumsey suggested that an infinite length structure based upon the equiangular or logarithmic spiral could be specified completely in terms of angles, and hence it should have characteristics independent of frequency. A study of balanced planar antennas based on these curves disclosed the equiangular spiral antenna to be the first antenna of practical size that had characteristics associated with an infinite structure.[14, 15] This structure became the first of a class now called frequency-independent antennas.[16]

This descriptive term is applied to antennas set apart from the usual concepts of operating bandwidth. The upper and lower frequencies at which these antennas may be operated are determined by independent paramters, and hence the bandwidth over which the radiation pattern and input characteristics are essentially constant is theoretically unlimited. Actually, the bandwidth is limited only by practical considerations of construction, and such antennas have been built to operate over a bandwidth of more than 42 to 1. There is no fundamental reason why this bandwidth cannot be extended.

The general philosophy behind the frequency-independent antennas is to design a radiating structure such that successive applications of a scaling factor result in structures identical with the original, or which at most, differ by some rotation about an axis through the origin of the original structure. Although the structure must be infinite in size to fulfill the scaling condition of exact equivalence, forms based upon the equiangular, or logarithmic, spiral curve may be truncated to a finite size and still retain, over an extremely wide range of frequencies, characteristics associated with the infinite structure.

A second breakthrough occurred in 1955 when R. H. Du Hamel proposed to force radiation from otherwise angular structures by periodic discontinuities. This led to a linearly polarized antenna with pattern and impedance characteristics that repeat periodically with the logarithm of frequency.[17-20] Although these logarithmic-periodic antennas scale exactly at only certain discrete frequencies, they can be designed so that the radiation characteristics do not change appreciably over the period between these frequencies. Hence they also exhibit nearly frequency-independent characteristics.

Many versions of the log-spiral and log-periodic antennas have been developed. The planar log-spiral antenna shown in Fig. 1 is bidirectional, radiating a broad circularly polarized beam on each side of, and perpendicular to, the antenna plane.[21] This spiral antenna may be made unidirectional, and hence more useful, by an orthogonal projection of the balanced spiral arms upon the surface of a cone (see Fig. 2 left).[22-23] For included cone angles less than about 45 degrees, radiation is confined to one broad circularly polarized lobe with maximum radiation off the apex of the cone. The antenna is balanced, with the feed voltage applied between the two arms at the apex of the cone. This feed voltage may be supplied by a conventional balanced line carried along the axis of the cone, or by a coaxial line and balun.

The characteristics of equiangu-

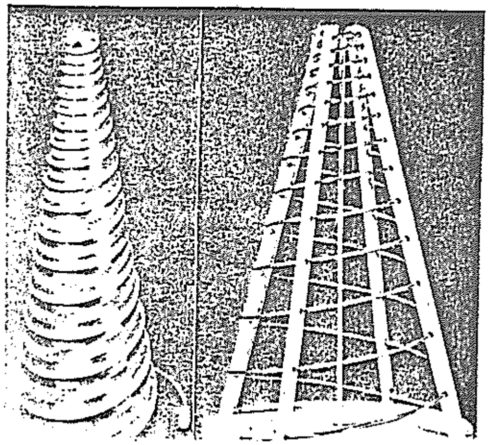

FIG. 2—Balanced conical logarithmic spiral antenna (left) with feed cable carried along one arm; wire version of balanced conical logarithmic spiral antenna (right), fed with coaxial line and balun along axis of cone

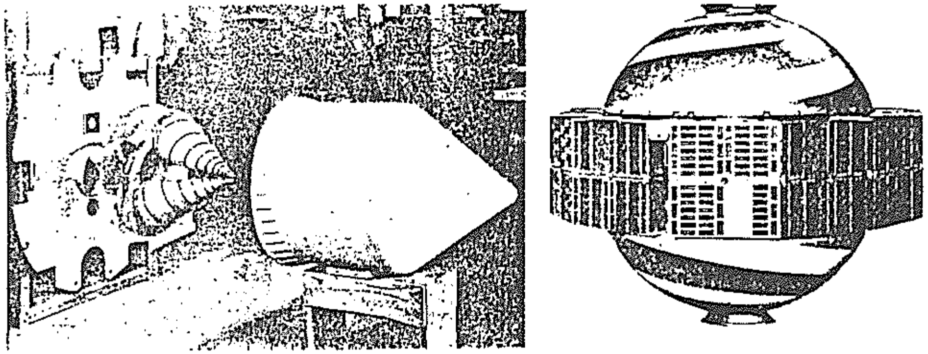

FIG. 4—First production of Dynatronics 500-to-4,000-Mc dual conical spiral Autotrack feed for 40-foot diameter parabolic reflector (left); Transit 3B satellite (right) by Applied Physics Lab. of the Johns Hopkins University. The logarithmic spiral antenna projected onto the spherical surface provides for transmission at 54, 108, 162 and 216 Mc

lar spiral antennas permit a third method. The feed line, a coaxial cable, may be carried along and soldered in contact with one of the arms. The amplitude of the antenna currents on the arms, and also on the outside of the cable, fall off rapidly with distance from the apex, so the ends of the arms where the cable enters carry negligible antenna current. This optional feed, conveniently and automatically, provides a frequency-independent balun, permitting feeding the balanced antenna by means of an unbalanced line.

Conical log-spiral antennas have been constructed to operate over more than 40-to-1 bandwidths. The upper usable frequency is determined by the truncated region at the apex which must remain less than 1/4 wavelength at the highest frequency of operation. The lowest usable frequency is determined by the base diameter of the cone which, for tightly wrapped antennas, must be approximately 3/8 wavelength at lowest operating frequency.

A highly practical conical log-spiral antenna can be constructed by allowing the exponentially expanding arms to degenerate into narrow constant width structures following the logarithmic spiral curve (see Fig. 2 right). For tightly wrapped spirals this version has characteristics similar to the original antenna. This wire version, simple to construct, may be excited by bringing the feed cable up the axis of the cone. If the infinite balun feed is desired, the feed cable itself can become one of the arms, with a dummy cable becoming the second arm of a balanced structure.

The balanced conical log-spiral antenna has characteristics other than bandwith that are attractive. The electromagnetic fields radiated by these antennas are essentially circularly polarized over very wide angles off the axis of the antenna. In addition, the phase of the far field of the two-arm antenna has essentially a one-to-one relationship with the azimuthal angle around the antenna. These two properties were utilized in the line feed for the University of Illinois radio telescope shown in Fig. 3. The focal-line array for this 400-by-600-foot parabolic reflector consists of a linear array of 286 conical logspiral antennas. The beam is steerable in declination ±30 degrees and the phase adjustment to provide this scan is secured by a physical rotation of each element of the array to a prescribed orientation. Thus it is of importance that the beam of each element be circularly polarized over at least a sector ±30 degrees from the axis and that the phase of the radiated field be linearly related to the angle of orientation around the antenna.

Multiarm versions of the balanced conical log-spiral antenna operate in several distinct modes. Thus a four-arm symmetrical structure may be excited to radiate an axial beam with an $e^{j\phi}$ variation, that is, the phase of the far field varies linearly with the azimuthal angle ϕ, around the antenna. It can be excited to produce a conical beam with an $e^{j2\phi}$ and a conical beam with an $e^{j3\phi}$ variation. Since orientation of the maximum of the conical beam may be controlled by varying the rate of spiral of the antenna arms, it can be positioned perpendicular to the axis of the antenna to provide a simple very-wide-band circularly polarized omnidirectional source. The well-behaved phase characteristics of these antennas suggest the basis for many variations of phase-comparison direction-finding systems.

The conical log-spiral antennas have been developed as feeds for dishes used in tracking applications. A modernization program is now under way to equip many of the large parabolic reflector antennas on the Atlantic Missile Range with dual conical spiral feeds designed to cover a 10-to-1 range of frequencies. The first production model of one of these feeds is shown in Fig. 4 (left).

In an interesting modification of the log-spiral geometry, the planar antenna, was projected upon a spherical surface and used on the Transit satellites shown in Fig. 4 (right). Although this geometry is no longer frequency independent, the antenna operates well over four frequencies, spaced from 54 to 216 Mc. Similarly, the planar antenna can be projected upon a myriad of surfaces. However, it can be shown that the conical surface, which includes the plane, is the only surface satisfying the required scaling principles for unlimited bandwidth.

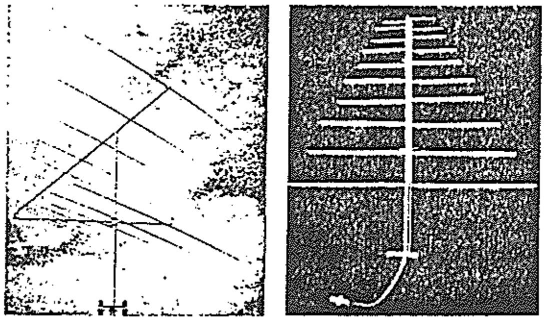

FIG. 5—Trapezoidal tooth nonplanar log-periodic antenna (left) designed by Collins Radio Company to cover 6.5 to 60 Mc; and logarithmic periodic dipole array (right)

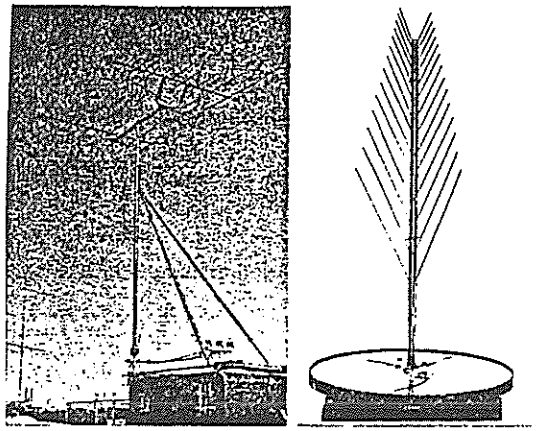

FIG. 6—Militarized version of the log-periodic dipole array (left) by Collins Radio Company; and log-periodic resonant-V array (right), an adaption of the log-periodic dipole array, designed to operate in several different modes

The log-periodic antennas have been extensively investigated. The planar antenna can be modified to provide unidirectional patterns by forming the two elements of the antenna into a V, that is, a nonplanar arrangement. A wire version of the log-periodic elements (see Fig. 5 left), and omnidirectional horizontally polarized, and unidirectional circularly polarized log-periodic antennas have been described. Multielement arrays have been described, and design information for log-periodic antennas as feeds for dishes, and as point-to-point h-f communications antennas has been given.

The dipole array of Fig. 5 (right) is a simple log-periodic structure. The length of each element in this array bears a fixed ratio to the length of the preceding element, and the adjacent element spacings bear the same ratio one to another. The pattern and impedance characteristics of the array at frequency f_n, such that the nth element is resonant, will be repeated at a higher frequency f_{n+1}, which makes the $(n + 1)$th element resonant. The characteristics repeat periodically at all frequencies given by $\tau^n f$, where n is an integer and τ the ratio of the lengths of adjacent elements. The geometry shown in this figure provides an excitation with a 180-deg phase shift between elements. Thus it radiates a single-lobe, linearly polarized beam, directed toward the apex of the array. Analysis and design information for this array has been presented. A militarized version of the antenna is shown in Fig. 6 (left).

In an extension of this work, a log-periodic monopole array has been described in which simple reactive networks produce the proper monopole current amplitude and phase distribution for unidirectional frequency-independent operation. A slot complement of the log-periodic dipole array has also been investigated.

The log-periodic resonant-V array (Fig. 6 right) is an adaptation of the log-periodic dipole array, designed to operate in several modes. In the lowest order mode, the performance approximates the log-periodic array. The active portion of the antenna centers around elements whose lengths are near a half-wavelength at the frequency of operation. As the operating frequency is increased, this active region moves toward the apex. The antenna may be designed so that as the $\lambda/2$ region moves off the apex of the antenna the large elements on the other end approach a $3\lambda/2$ resonance. Thus, as the operating frequency is increased, the active region moves through the array in the $3\lambda/2$ wave mode, then in the $5\lambda/2$ mode, etc. Forward tilt of the individual elements insures a unidirectional beam and provides increased directivity. This antenna has transition regions between modes where the characteristics deteriorate, but it can provide a gain of approximately 18 db over an isotropic radiator in the higher modes.

Triangular or zig-zag log-periodic structures have received considerable attention. A flush-mounted sinuous structure, fed by a waveguide, has been studied, and a balanced zig-zag antenna has been proposed. A vertically polarized log-periodic zig-zag antenna is shown in Fig. 7. The teeth bent parallel to the ground plane serve as a delay line to phase the vertical radiating portions.

Tapering structures, not constructed in a logarithmic periodic manner, have led to antennas of appreciable bandwidth. One tapered ladder antenna consists of symmetrical elements excited by a balanced line. In addition long slot, tapered feed, structures that can be used over a 10-to-1 bandwidth have been proposed. A radiating horn with side walls of periodically spaced wires, when fed from a ridged waveguide, can be operated over a 4-to-1 range of frequencies.

The frequency-independent antennas have opened up a new era in wide-band antennas. However, in view of their complex geometry, the experimental and empirical development of these antennas have outstripped their theoretical analysis.

The theoretical problem of the log-spiral antenna has received considerable attention. This problem has been attacked by calculating the fields due to a logarithmic spiral filament, and the fields produced by an infinite number of planar logarithmic spiral filaments. An approximate analysis for the radiation patterns of a filamentary log-spiral antenna with an assumed current distribution has been published.

A recent comprehensive analysis of the log-periodic dipole array has been reported.

A new and promising approach considers the log-periodic antenna as a locally periodic structure whose period varies slowly, increasing linearly with distance to the origin or apex. This approach indicates that the basic ingredients for a frequency-independent antenna are a slow-wave transmission medium and a series of radiating elements, satisfying the similarity condition, coupled to the transmission medium at points spaced in geometric progression. Such a structure produces a wave substantially radiated toward the feed point with characteristics that vary little as the operating frequency is varied over extremely wide bandwidths.

A general consideration of the log periodic zig-zag antennas and the conical log-spiral antennas as locally periodic, slow wave structures has led to a better understanding of their operation and to some useful design criteria.

This survey has covered the work of many people. It has been impossible to acknowledge the work of many others, however an extensive list of references has been included.

The original work on logarithmic spiral and logarithmic periodic antennas was conducted at the University of Illinois, sponsored by the Aeronautical Systems Division, Wright Air Development Division under contracts AF33(616)-3220 and AF33(616)-6079.

This article is based upon a paper by the author published in the Journal of Research of the National Bureau of Standards, Section D, January 1962.

REFERENCES

(1) A. G. Kandoian, Three New Antenna Types and Their Applications, *Proc. IRE*, 34, p 70w, Feb. 1946.

(2) P. S. Springer, End-Loaded and Expanding Helices as Broadband Circularly Polarized Radiators, *Proceedings National Electronic Conference*, 5, 1949.

(3) J. S. Chatterjee, Radiation Field of a Conical Helix, *Journal of Applied Physics*, 24, p 550, May 1953.

(4) J. S. Chatterjee, Radiation Characteristics of a Conical Helix of Low Pitch Angles, *J. Appl. Phys.*, 26, p 331, Mar. [illegible].

(5) E. M. Turner, Spiral Slot Antenna, Wright Air Development Center, Dayton, Ohio, Technical Note WCLR-55-8, June [illegible].

(6) B. H. Burdine, and R. M. McElvery, Spiral Antennas, MIT Research Laboratory of Electronics, Report No. 1, Mar. [illegible].

(7) B. H. Burdine, and H. J. Zimmerman, The Spiral Antenna, MIT Research Laboratory of Electronics, Report No. 2, Apr. [illegible].

(8) E. M. Turner, Broadband Radiators, *Proc. IRE Nat. Aero. Electronics Conference*, 1958.

(9) W. L. Curtis, Spiral Antennas *IRE Transactions*, AP-8, p 298, May 1960.

(10) J. A. Kaiser, The Archimedean Two-Wire Spiral Antenna, *IRE Transactions*, AP-8, p 312, May 1960.

(11) R. Bawer, and J. J. Wolfe, The Spiral Antenna, *IRE National Convention Record*, p 84, 1960.

(12) J. P. Jones, D. E. Taylor and C. W. [illegible], Design Techniques for a Light Weight High Power Spiral Antenna, *IRE WESCON Convention Record*, 4, Part 1, [illegible], 1960.

(13) J. R. Donnellan, Second-Mode Operation of the Spiral Antenna, *IRE Trans.*, AP-8, p 637, Nov. 1960.

(14) S. R. Jones, and E. M. Turner, Polarization Control with Oppositely Sensed Circularly Polarized Antennas, *Proc. IRE National Aero. Electronics Conf.*, p 558, May 1959.

(15) R. M. Brown, Jr., and R. C. Dodson, Parasitic Spiral Arrays, *IRE International Convention Record*, 8, Part 1, p 51, 1960.

(16) J. R. Donnellan, and R. T. Close, A Spiral-Grating Array, *IRE Trans.*, AP-9, p 291, May 1961.

(17) J. R. Donnellan, A Spiral-Doublet Scanning Array, *IRE Trans.*, AP-9, p 276, May 1961.

(18) J. D. Dyson, The Equiangular Spiral Antenna, *Proc. Fifth Symp. on the USAF Ant. Research and Development Program*, Allerton Park, Monticello, Oct. 22, 1955. (Classified).

(19) J. D. Dyson, The Equiangular Spiral Antenna, *IRE Trans.*, AP-7, p 181, Apr. 1959.

(20) V. H. Rumsey, Frequency Independent Antennas, *IRE Nat. Conv. Record*, Part 1, p 114, 1957. Also Antenna Laboratory Technical Report No. 20, Contract 33 (616)-3220, Dept. of Electrical Engineering, University of Illinois, 25 Oct. 1957.

(21) R. H. DuHamel, U. S. Patent No. 2,985,879, 1958.

(22) R. H. DuHamel, and D. E. Isbell, Broadband Logarithmically Periodic Antenna Structures, *IRE Nat. Conv. Record*, Part 1, p 119, 1957.

(23) J. D. Dyson, The Equiangular Spiral Antenna, AD-145019, Technical Report No. 21, Contract AF33 (616-3220, Dept. of Electrical Engineering, Univ. of Ill., Sept. 15, 1957.

(24) N. Barbano, Phase Center Distributions of Spiral Antennas, *IRE WESCON Conv. Record*, 4, Part 1, p 123, 1960.

(25) J. D. Dyson, Recent Developments in Spiral Antennas, *Proc. IRE Nat. Aero. Elect. Conf.*, p 617, May 1959.

(26) M. S. Wheeler, Design Procedure for Equiangular Spiral Antennas, Seventh East Conf. on Aero. and Nav. Electronics, p 1.2.4, Oct. 1960.

(27) J. D. Dyson, U. S. Patent No. 2,958,-081, 1959.

(28) J. D. Dyson, The Non-Planar Equiangular Spiral Antenna, *Proc. Eighth Symposium, USAF Antenna Research and Development Program*, Robert Allerton Park, Monticello, Ill., Oct. 1958.

(29) J. D. Dyson, The Unidirectional Equiangular Spiral Antenna, *IRE Trans.*, Vol. AP-7, pp. 330-334, October, 1959. Also Antenna Laboratory Technical Report No. 33, Contract 33(616)-6079, Dept. of Electrical Engineering, University of Illinois, July 10, 1958.

(30) M. Nussbaum, Designing Broadband Conical Helix Antennas, *Electronic Design*, Sept. 3, 1958.

(31) H. S. Barsky, Broadband Conical Helix Antennas, *IRE Nat. Conv. Record*, 7, Part 1, p 130, 1959.

(32) J. D. Dyson, and P. E. Mayes, The Log-Spiral Omnidirectional Circularly Polarized Antenna, URSI-IRE Spring Meeting, Wash., D. C., May 3, 1960.

(33) J. D. Dyson, and P. E. Mayes, New Circularly Polarized Frequency Independent Antennas with Conical Beam or Omnidirectional Patterns, *IRE Trans.*, AP-9, No. 4, p 334, July 1961. Also Antenna Laboratory Technical Report No. 46, Contract AF33(616)-3220, Dept. of Electrical Engineering, Univ. of Ill., July 1960.

(34) J. D. Dyson, Multi-Mode Logarithmic Spiral Antennas—[illegible]ssible Applications,, *Proc. National Electronic Conf.*, 17, p 206, Oct. 1961.

(35) J. D. Dyson, The Log-Spiral Antenna in Simple Arrays, *Proc. of the Eleventh Symposium, USAF Antenna Research and Development Program*, Allerton Park, Monticello, Ill., Oct. 1961.

(36) E. M. Turner, The Scimitar Antenna, *Proc. IRE Nat. Aero. Electronics Conf.* 1956.

(37) H. B. Riblet, A Broadband Spherical Satellite Antenna, *Proc. IRE*, 48, p 631, Apr. 1960.

(38) D. E. Isbell, Non-Planar Logarithmically Periodic Antenna Structures, *Proc. Seventh Symp. USAF Ant. Res. and Dev.*, Allerton Park, Monticello, Illinois, October 1957. Also Antenna Laboratory Technical Report No. 30, Contract AF33(616)-3220, Elect. Eng. Dept., Univ. of Ill., Feb. 1958.

(39) R. H. DuHamel, and F. R. Ore, Logarithmically Periodic Antenna Designs, *IRE Nat. Conv. Rec.*, Part 2, p 139, 1958.

(40) R. H. DuHamel, and D. G. Berry, Logarithmically Periodic Antenna Arrays, *IRE WESCON Conv. Rec.*, 2, Part 1, p 161, 1958.

(41) R. H. DuHamel, and F. R. Ore, Log Periodic Feeds for Lens and Reflectors, *IRE Nat. Conv. Rec.*, 7, Part 1, p 128, 1959.

(42) R. H. DuHamel, and D. G. Berry,

Medical Aspects of Ultra Short Waves

By C. H. West

U. S. Public Health Service

Mr. West here describes some of the effects and also some of the dangers of subjecting the human body to powerful ultra short wave fields

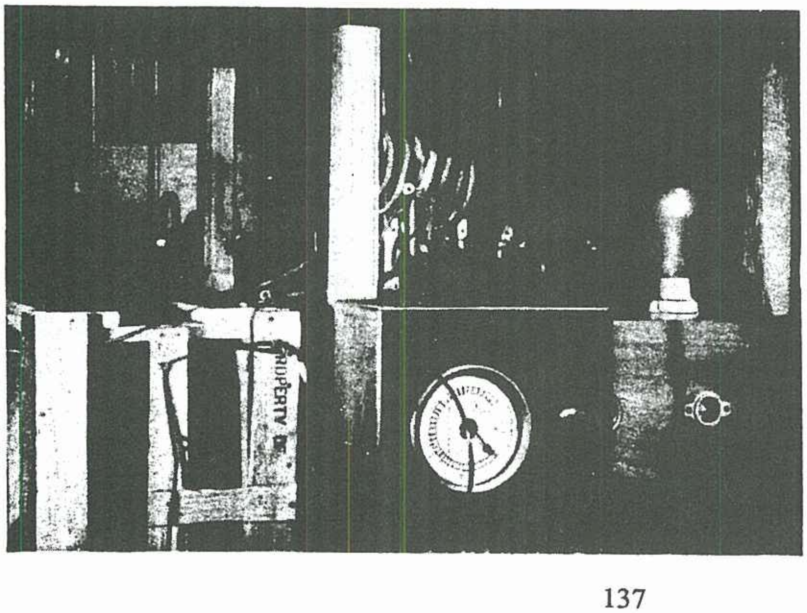

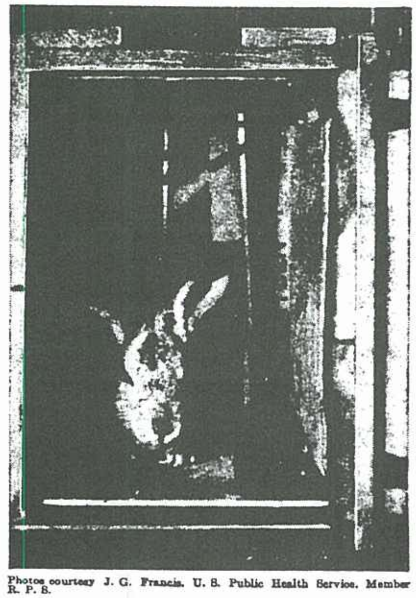

Photos courtesy J. G. Francis. U. S. Public Health Service. Member R. P. S.

Above—A rabbit about to be exposed to a powerful ultra short wave field; the sides as well as the top and bottom linings of the cabinet are of metal.

At left—The appearance of the ultra short-wave apparatus built and tested by Mr. West in his experiments.

● IT SEEMS that the general public is once again taking to radio and the art of building as a pastime. The amateurs are busy developing the five meter band, while the engineers are devoting a goodly portion of their time to dabbling in the mysteries that the ultra short waves have created.

On the other hand, medical-scientific personages are making use of their medical knowledge, plus a large portion of electronics, in efforts to make the world a better place to live in with less aches and pains to bother us.

Many engineers employed in large electrical laboratories have made valuable use of their knowledge pertaining to the ultra short waves, and have constructed oscillators of high power for the purpose of producing fever in the human body at will. It is a known fact that fever is nature's doctor to a certain degree. If it can be produced in cases where there is no cause for a high fever, the patient's chance for recovery and cure is far better than combatting disease by injection of malaria germs, etc., which may produce reactions worse than the disease itself

Unknown Qualities

However, the production of fever by an oscillator is a simple procedure, but the application of these high wave speeds to the human body is an entirely different matter. Secondly, the output or radiation of an ultra high frequency oscillator has never been completely identified.

It seems that for every known element discovered there are many others within—undiscovered. An illustration of this fact was brought to light very recently, in which the writer witnessed a demonstration of photo-electric cell work.

A double-cell apparatus manufactured by the Weston Electrical Instrument Corporation produced readings in foot-candles from the output of a 10 k. w. X-ray tube in which the cells were completely covered by ⅛ inch lead sheet.

The roentgen rays do not penetrate lead, and as the tube was quite a distance from the cells, the reaction in the sensitive meter was not due to heat waves, but simply a current output which registered through the lead upon the faces of the cells. The X-ray is noted for producing very bad burns if not handled properly, and the deflection from the *target* is called and identified as the *roentgen ray* after the name of the discoverer. How does science know whether the burns are caused from this additional element as registered in *foot candles*, for want of a better name?

U. S. W. "Fever" Apparatus

Recently, the writer constructed and put into practice an *ultra short wave oscillator* for the production of fever or temperature. This apparatus consisted of two UX-852 tubes arranged in push-pull, but with the "back-to-back" arrangement of *Mesny-Valuri*. It is a persistent oscillator and is capable of running long hours on wavelengths from 2 to 10 meters.

An auxiliary cabinet was constructed with two leads to couple across the plate inductance. This cabinet has one adjustable side; the top and bottom being one lead, and the two sides the other and forming a large condenser.

Rabbits were used in experiments and their temperature reached 41 degrees (Centigrade) from the usual 38 degrees (normal). However, after a few weeks severe burns were noticed. Since that time research has been carried on by other factions in an effort to ascertain the direct cause of the burns, which do not seem to be due to collections of moisture, or from coming in contact with the condenser plates.

This brings to mind that experiments by the writer with the very short waves disclose the fact that some sort of a photo-electric ray could be produced and would highly expose small dental films. It is believed this elementary action was first discovered by Mr. John Reinartz, the noted radio amateur, who perceived a bluish glow within an auxiliary indicator at a certain frequency.

Before any actual application to humans may be applied, it is our duty to ascertain accurately whether other elements are prevailing. To all indications there are many, and one of them is severe burns, which may be caused by an unidentified element.

Constructional Data and Summary

As will be noted in the photographs, the apparatus is entirely enclosed and surrounded by glass sides to better view the "works" in case things tend to go the "hay-wire" route. The lower portion of the cabinet contains the necessary plate and filament transformers, which operate direct from the 110 volt, 60 cycle line. The plate power is raw A. C. at a potential of 2600 volts, the filament transformer being the customary 10 volt affair. Inductances are removable; the ones shown are at present used for 30,000 kc. work. The circuit is shown in Fig. 1, which is the conventional push-pull method.

However, the auxiliary cabinet is for purpose of holding an animal within the field and without strapping him and causing discomfort. The condenser plates are of sheet metal and insulated, the schematic diagram appearing in Fig. 2.

When a rabbit is placed in this pen and the power applied, he fails to react immediately; but as he "warms" up to the situation the veins in the ears show dilation, and in 20 minutes his temperature starts to rise and keeps going. It is only at a much later date that burns are noticed.

From the foregoing it would hardly seem advisable to subject a human to any lengthy treatment, at least, for the present; but the value of high frequency in various forms have proved essential in many cases. Secondly, the natural heavy perspiration that is produced is essential in one degree in opening the pores (some of which have probably never been opened before) and allowing the natural poisons of the body to escape more quickly than could have been accomplished with potent medicines.

The majority of high frequency apparatus produces this result. In the case of pulsating currents, which produce a series of reactions or "jumps" within the muscular system, this often is very beneficial in activating those muscles which could not have been manipulated manually without a great degree of pain.

Probably the first reaction noticed with reference to the ultra high frequencies were from early experimenters while conducting transmission tests with high power. It was noticed that the body temperature tended to rise at various times, due entirely to their proximity to a powerful oscillator. Science has been looking for a long time for some method to raise the temperature in the human body *quickly*. Ultra high frequency oscillators have solved that problem but have brought many other matters to light. Secondly, to produce the desired results the patient must be enclosed between two or more condenser plates of large proportions. The oscillator must be of high power, capable of delivering 500 watts output and the patient must be wrapped in a suitable covering to prevent his contact with any portion of the charged plates.

Various research workers who have volunteered their services as "trial patients" have received a goodly portion of burns as compensation, and are quite satisfied that ultra high frequency currents are "hot stuff!" Where heat is concerned, one could get equivalent results by setting on a hot steam radiator. The writer knows this to be true, and has experienced many a painful burn which he would have been willing to trade for a shock from a quarter kw. closed-core transformer.*

It could hardly seem probable that burns would prevail where the body is not in contact with any metallic object; but an examination under darkness of a large condenser tapped from the oscillator, will disclose a bluish aurora between the two elements, which is similar to that noted in Tesla coil experiments. Within this bluish discharge there can be many components not yet identified, which have a tendency to attack tender portions of the body and produce an internal burn, which does not come to the surface for identification until some days have passed.

It is open to discussion whether one-half of the electrical apparatus in present use actually cures the patient of pains, or whether it is the psychology of the matter in which the patient is surrounded by many cabinets of mysteries, of which he knows absolutely nothing. What cures one man does not seem to cure another with the identical complaint.

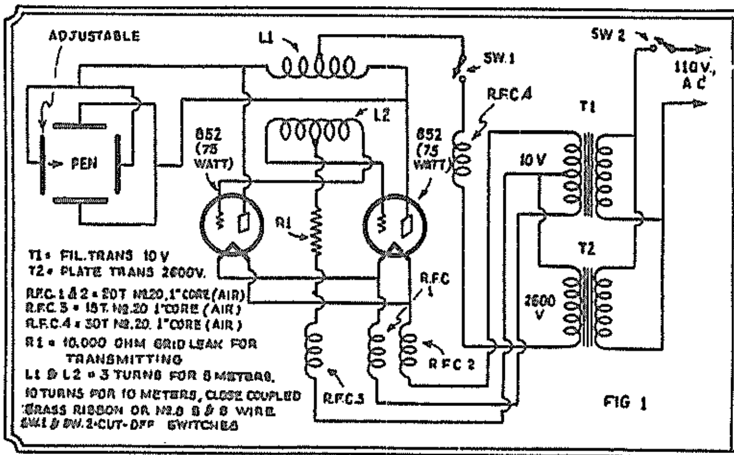
ADJUSTABLE
L1
SW 2
SW.1
110V.
A C
R.F.C 4
L2
T1
852 (75 WATT)
852 (75 WATT)
PEN
10 V
R1
T2
R.F.C 1
2600 V
R F C 2
R.F.C.3
T1 = FIL.TRANS 10 V
T2 = PLATE TRANS 2600V.
R.F.C. 1 & 2 = 20T No.20, 1" CORE (AIR)
R.F.C. 3 = 15T. No.20 1" CORE (AIR)
R.F.C.4 = 30T No.20. 1" CORE (AIR)
R1 = 10,000 OHM GRID LEAK FOR TRANSMITTING
L1 & L2 = 3 TURNS FOR 5 METERS.
10 TURNS FOR 10 METERS, CLOSE COUPLED BRASS RIBBON OR No.8 B & S WIRE
SW.1 & SW.2-CUT-OFF SWITCHES

FIG 1

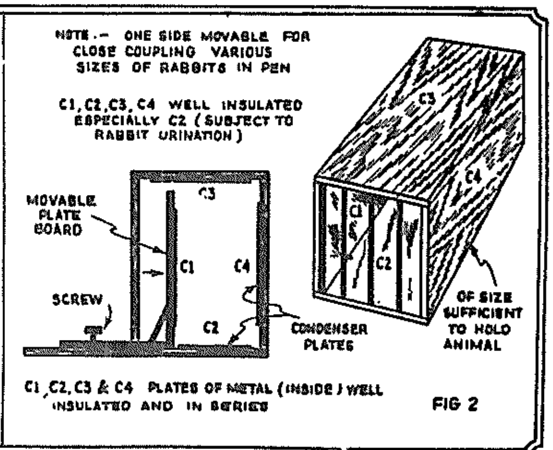
NOTE.— ONE SIDE MOVABLE FOR CLOSE COUPLING VARIOUS SIZES OF RABBITS IN PEN
C1, C2, C3, C4 WELL INSULATED ESPECIALLY C2 (SUBJECT TO RABBIT URINATION)
MOVABLE PLATE BOARD
SCREW
C3
C1
C4
C2
CONDENSER PLATES
OF SIZE SUFFICIENT TO HOLD ANIMAL
C1, C2, C3 & C4 PLATES OF METAL (INSIDE) WELL INSULATED AND IN SERIES

FIG 2

The drawings above provide details of the circuit used by Mr. West in building his ultra short-wave physiological apparatus, together with details of the apparatus itself.

---

*Yes—but it is generally understood we believe, that the high frequency currents or field causes heat to be developed inside the body; even in organs such as the liver, etc., without heating up the whole body, which is not the case with direct application of heat as when seated on a hot radiator, etc.—Editor.

Short Waves in Medicine

ULTRA SHORT WAVES IN *MEDICINE*

By Dr. FRITZ NOACK
(Berlin)

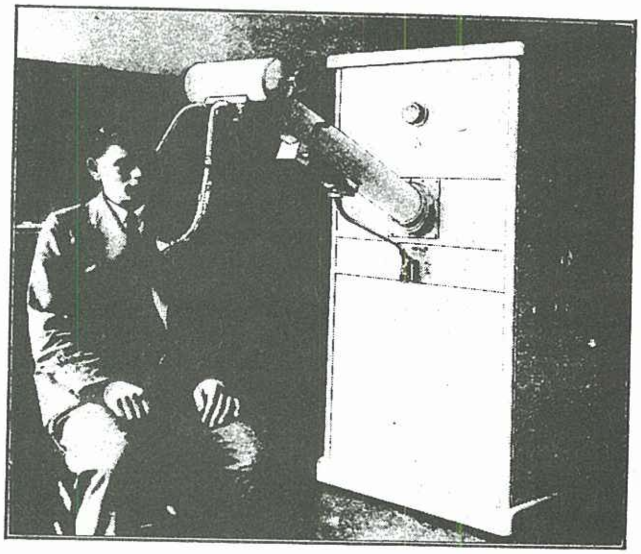

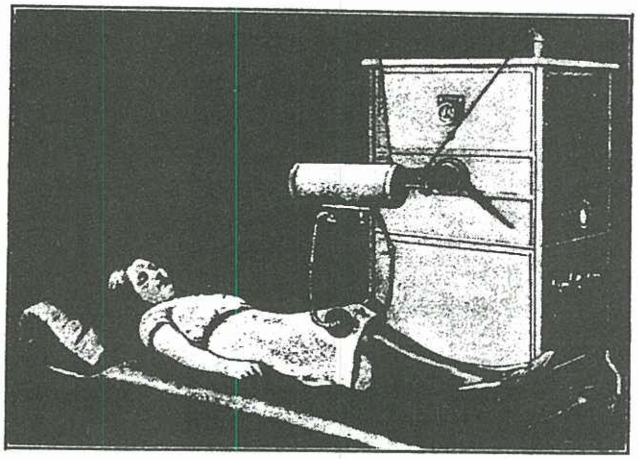

One of the "fair sex" receiving a treatment through the knee by ultra-short waves. Several treatments are given for the average ailment.

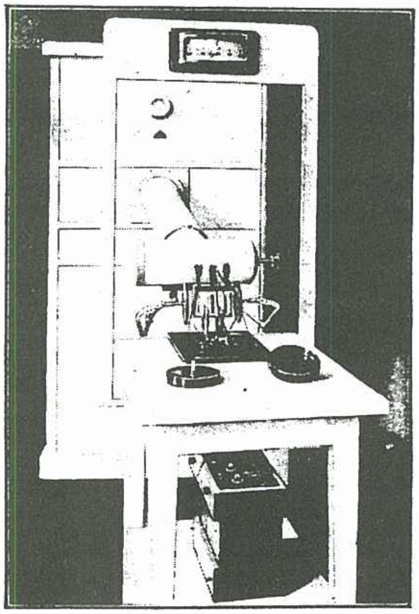

The ultra-short wave medical apparatus with its electrode arm adjusted to permit the waves passing through a solution for experiment.

PROFESSOR ESAU, of the University of Jena, the well-known pioneer in the field of ultra short waves, was the one who nearly two years ago pointed out that very evidently the ultra short wave exercised beneficial medicinal effects.

For further investigation of this effect, small mice were put between the condenser plates of the sender; and Lo! after some time the mice fell victims to a sort of tetanus, which eventually led to death. The result of this fact was that Prof. Esau resolved to commission a physician to make a further study of the physiological effect of these ultra short waves. Naturally the serious investigation of medicinal effects requires experience in physical measurements, likewise an apparatus operating in a constant and easily controlled manner.

Siemens and Halske Interested

It is gratifying that a manufacturer which already possesses very great experience in constructing electro-medical apparatus and which also has the requisite experience in physical measurements, the prominent firms of Siemens and Halske took up the matter. Now this company has manufactured a first ultra-short wave *radiation apparatus* and recently exhibited it to a picked group of scientists and to the press.

The pictures show the new set which, of course, has to be considered as the first of an evolutionary series. There will still be necessary all sorts of experiments to produce the set in such form that it will correspond to all the needs of the medical profession. The present set will, for the time being, be used for the further study of the physiological effect of ultra short waves. It will be the problem of the doctors to investigate the fields of use and the conditions under which the set may be used.

Arrangement of U. S. W. Generator

In the large shielded box are the real ultra short wave generator—(actually resembling a radio transmitter) and the parts needed for its operation. The generator can work on two waves, four and eight meters. The wavelength is adjusted by inserting the proper tuning coils. The arm which projects out of the apparatus has inside it two lead wires which convey the oscillations outward; they are led to the two electrodes, which are placed at the ends of the two visible supports.

In the cross-beam at the end of the arm is a "tuning" device, which tunes the electrode oscillatory circuit exactly to the wavelength of the generator. The exact tuning can be read on a meter, which is above the arm on the box; this shows the direct plate current of the transmitter, which, with correct tuning of the electrode circuit, adjusts to a minimum; the value then indicated by the meter gives, after calibration of the a value for the electrode energy.

To be able to adjust the electrodes, as is necessary from one case to another, the electrode arms are set in ball-joints. To ascertain whether oscillations are actually present in the electrode oscillatory circuit, there is on one electrode arm a little *glow lamp* (Neon tube), which lights up in the presence of oscillations.

Electrodes Swing on Giant Arm

The arm projecting out of the apparatus can be swung in all directions, for the better use of the whole unit. To accomplish this, it is fixed to the box in a universal joint. The end cross-piece of the arm can furthermore be turned, so that the electrodes can assume any desired direction. The right side wall of the box contains the main current switch, as well as a resistor with which the filaments of the tubes are regulated. The apparatus is in fact constructed just like a vacuum tube radio transmitter. The electrodes may take various forms; in accordance with the most recent experiences, they are so shaped that they do not have to be put firmly against the body, but exercise their effect even in the presence of a layer of air between body and electrode. At the same time, the remarkable fact may be determined, that the thickness of the layer of air and the "depth effect" are dependent on each other.

Advantages of Ultra Short Waves

According to all experiences thus far, the medicinal ultra-short wave radiator is a sort of *diathermal* apparatus; therefore it serves to conduct to the part of the body, or the organ to be treated, electric wave energy, which is converted into heat and thereby exercises a healing effect. As compared with the usual diathermal sets (which operate on a wavelength of about 600 meters, using leaden electrodes, which have to be put right on the body), the *ultra-short wave set* offers a considerably better control of the direction in which the heat is to be applied, so that one can better reach any desired organ than was possible before. Furthermore, by changing the wavelength, one can also adjust for a definite deep-lying stratum. Unpleasant accompanying phenomena, which result with the usual diathermal sets, where the electrodes do not rest firmly on the body, do not occur in the case of the ultra short wave radiator. It is especially important to note that one can now, for the first time, perform a "therapy" or treatment directly on the head. The operation of the set is very simple.

Whether, besides the warming of the different strata in the body, still other biological effects occur or can be accomplished, with the new ultra short wave radiator, is, at the present time, still a matter of question.

HUMAN BEINGS *as* ANTENNAS

By DR. ERWIN SCHLIEPHAKE, M.D.

Of the Jena University Medical Clinic

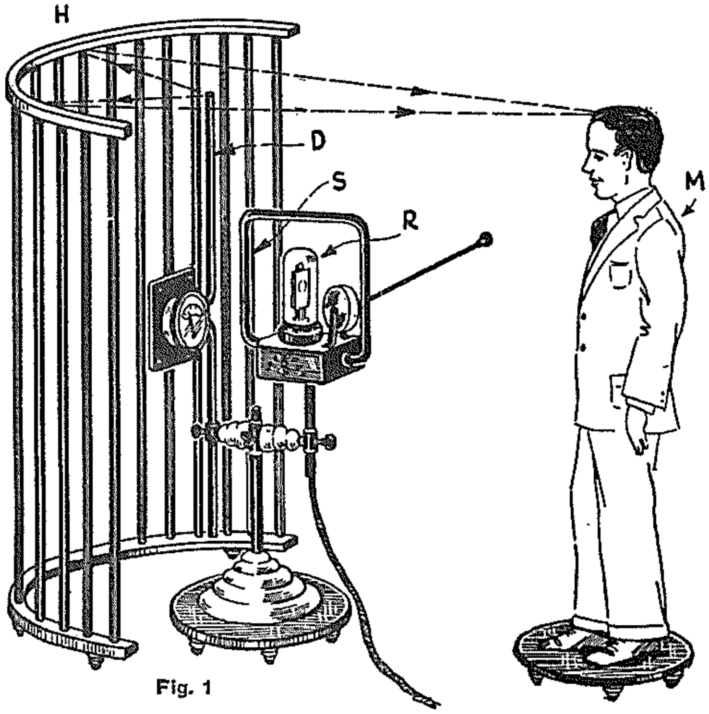

Fig. 1

Arrangement of the apparatus used by Dr. Schliephake in measuring and observing the effects of short waves on the human body, absorption, et cetera. H, parabolic reflector formed of wires; D, antenna, charged inductively by the coil S; vacuum tube at R, and beside it wave changing condenser; subject at M.

IF human beings remain in the vicinity of powerful radio transmitters, they are exposed to a very strong electromagnetic field. One must therefore ask what significance these radiations have for the internal organs, and how the body reacts to them. This question is justified by the fact that all other kinds of electromagnetic rays influence the human body more or less strongly.

The character, however, of the reaction is different, according to the wavelengths. Radium rays and X-rays, for example, because of the extraordinary shortness of the waves and the high effectiveness of energy under these conditions, exercise a very strong effect on the structure of atoms; so that serious injuries to the organism can arise (for instance, under certain conditions, cancer). Less serious are the effects of light rays which, however, in the ultraviolet part of the spectrum, can still occasion serious injuries; while the visible light which we perceive is of too low a frequency to injure healthy persons. There are, however, subjects, made sensitive by illness, who suffer serious affections of the skin from strong illumination. Above the optical spectrum, there follow the infra-red and heat waves, whose effect on atoms is relatively slight; in the case of still longer waves only molecular effects are to be expected. Here we are entering the range of the Hertzian waves, electric waves in the stricter sense. Of these, the shortest are about half a millimeter (.02-inch) long, and therefore just above the heat radiations.

From direct analogy, it is to be assumed that the shortest of these waves must be absorbed by human bodies, as in the case of heat rays, and this absorption can be actually demonstrated.

DO YOU KNOW —

¶ What effect short waves have on the hair and scalp in general?

¶ Whether the current in the exciting antenna increases or decreases when a person stands in its field? Why?

¶ If the person stoops will this affect the reading of the antenna current meter?

¶ Whether 'cold" or "heat" is produced inside the body by an overdose of the high frequency waves?

¶ The effect of short waves on the nervous system?

Absorption of Radio by the Body

With specially-built transmitters, especially such as those which have been described by Kohl, electric waves only a few centimeters in length can be produced; these waves can be reflected by mirrors, exactly like light rays. If one puts the antenna in the focus of a concave mirror, then there is formed a parallel beam of radiation which can be concentrated again in another concave mirror. By means of a lens placed in the course of the radiation, the ray is made very sharp. If a person steps into this path, reception ceases. Since no reflection by the human body or only a very slight one is demonstrable, absorption of the radiation must have occurred; this effect is, moreover, demonstrable in the case of other organic substances or water. Unfortunately, not much can yet be said about the physiological effects of these radiations; the power of the transmitters of extremely short waves is still too low to produce perceptible changes.

It is otherwise with the wavelengths from three meters up, to the production of which powers of several kilowatts can be applied. Here the above described optical phenomena cannot be so well demonstrated, because the diffraction is much greater; but absorption by the human body can be very well shown.

When an antenna is inductively coupled to a 3-meter transmitter, the oscillations can be indicated by a detector even at a considerable distance away. If however, a person puts himself in the place of the antenna, the detector responds much more weakly, although the power consumed by the transmitter remains the same. Accordingly, a part of the power must have been used up in the body. The same phenomenon can be demonstrated as follows: a closed or open oscillation circuit is inductively coupled with the transmitter (see Fig. 1). The ammeter in the circuit shows a definite current strength. If a person places himself on an insulating stool beside it, the current in the oscillation circuit is reduced. This withdrawal of energy, however, depends on the length of the body; for if the subject stoops, or changes to a sitting position, the current in the other circuit increases (Fig. 2). Therefore, it appears as though, by the tuning of the subject, *to about half the wavelength, the power transmitted to him becomes much greater.*

Concentrating Power by Means of a Reflector

By means of a large concave mirror one can also collect the transmitted power to a focus. Such a mirror need consist only of parallel wires stretched between two wooden frames. Its height must be equal to the wavelength, the opening one and one-half times the wavelength. It is best to use elliptical reflectors, with the transmitting antenna at one focus; then the reflected radiation is at the other focus of the ellipse. With a "dipole" ("Hertzian" antenna) containing an ammeter, especially strong concentration of energy at this point can be demonstrated. Here a lessening of the current in the dipole is instantly shown if a person steps into the vicinity. Since the human body is to be regarded essentially as an electrolytic system, with regard to the electric wave, and I have tried to demonstrate the effect in the following manner in a model experiment.

A glass tube, of half the wave's length, was filled with an 0.5% sodium chloride (salt) solution, to which gelatine was added to prevent convection. In this jelly the temperature was measured, at different places, by thermo-elements. It was shown that the heating was greatest in the middle and least at the ends; being half as strong at the quarter points of entire length as it was in the middle. Since the maximum strength of the current is at the middle, the greatest heating is therefore connected with this.

Physiological Effect of Short Waves

Especially noteworthy, also, was a feeling of vibration, which was particularly evident if the hand was raised in front. We could establish this sensation only at our transmitter, which is operated with 50-cycle alternating current. There must be, therefore, a direct influence on the nervous system. Here too we have, therefore, another proof that the ultra-short radio waves exercised an effect on the nerves.

These vibrations are also felt if the hand is placed in a condenser's field. Many persons who remain close to the transmitter also experience remarkable sensations on the head, near the roots of the hair; these are like a peculiar prickling, the hair likewise standing up a little. In many subjects we could also observe slight increases in bodily temperature, which however did not exceed 0.5 of a degree, Centigrade (0.9 degree Fahrenheit). Since the body contains extremely fine regulators, by which the temperature is always kept constant, and since also the amounts of energy which can be conveyed into the body, even by powerful radio transmitters, give (when translated into heat units) only a relatively low number of calories, this increase in heat is probably not attributable to the received energy alone. As is to be shown later, nervous effects may play an important part in this.

The effect on the nervous system is

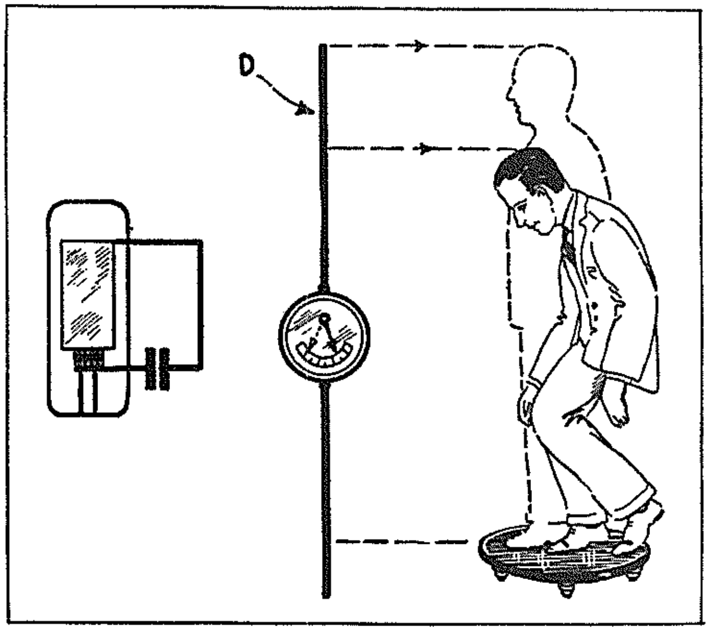

Fig. 2. One of the most interesting effects of placing a human being in the field of a short wave (high frequency) oscillator is that of absorption. The antenna "D" is excited by the power tube and loop circuit at the left. Do you think the antenna current is increased or decreased by bending the body?

Human Beings As Antennas

By Dr. Erwin Schliephake, M.D.

plainly felt by all persons who work a fairly long time with short waves. The sensations are different with the individuals; there are persons who are relatively insensitive, and others who very soon complain about the disturbances of their health. Usually there is first an increasing sleepiness; they are very tired by day, but at night they sleep badly. Several times a night they start out of their sleep, and they are tired and sleepy in the morning.

These phenomena increase more and more. Often there are also intensive headaches, particularly covering the back of the head. Many persons complain of digestive disturbances and pressure on the stomach. Most of them are furthermore easily excited and irritated, being inclined to complaining and to violence. This increased excitability of the nervous system can also be shown by electrical tests of the nerves.

Effects of the Electrostatic Field

It is much easier to study these changes by using the condenser field. Here it is not the electromagnetic wave which is used, but the electrostatic field, which always accompanies it. Here, however, the effect may be much more strongly concentrated. I have generally used plates four inches in diameter, between which the parts of the body in question were treated. The field between these plates suffices to heat 100 cc. of a 0.5% sodium chloride (salt) solution 5 degrees C. (9° F.) in one minute. If parts of the human body are introduced into this field, and the blood is then taken from some part of the body (for instance, from the earlobe) on investigation obvious changes are found.

The number of red corpuscles per cubic millimeter is very greatly increased; i. e., if they formerly amounted to 5 million, the number has risen to 6 million. The same is true of the haemoglobin (a constituent of the blood) and the white corpuscles, the number of which likewise increases. But this is not always true to the same extent, depending upon the strength of the radiation and the place treated. One frequently sees instead a lessening of these amounts, which would therefore correspond to a thinning of the blood. Likewise, the albumens of the blood undergo certain changes which I cannot discuss here.

In the case of these strong effects it is to be assumed that the tissues of the body also undergo changes; which, however, cannot be directly proved in a subject.

Results of Experiments on Animals

In experiments with animals, on the contrary, such changes are plainly recognizable. They occur particularly strong in projecting parts of the animals, such as in the ears and tips of the tails; since there the electric field is distorted. Very often, one sees, some time after the treatment, that the parts in question are dying and falling off. In the case of a rabbit whose leg had been too strongly exposed to the electrostatic field, I noticed a complete disintegration of the flesh in a ring-shaped region, so that only the bone remained; but then, after a while, that also fell off.

That the nervous sytem of the animals is also influenced is shown by the fact that many of them shudder on the switching on or off of the field, at a time when there can be no question of a strong heating effect.

Also very interesting are the disturbances of the internal heat regulation, which one can produce in animals. If the region of the neck and the back of the head of a rabbit are exposed to a limited capacitative field and the temperature is afterward measured, one can soon note a permanently increased temperature, which often lasts for some weeks. If a very powerful dose of "irradiation" is administered the opposite can occur: the bodily heat falls more and more, often below 35 degrees C. (95° F.), until these low temperatures are no longer compatible with life. At the same time, it is interesting to observe that almost all animals which have had such disturbances of their heat regulation after a few weeks developed inflammation of the lungs and pleurisy, afterward dying. It seems as though, by the disturbance of the heat regulation, the resistance of the animals to disease had suffered greatly; so that in this path throughout the central nervous system there was created a special susceptibility to colds. On investigating the spinal marrow of such animals microscopically, serious injuries to the nerve cells have been observed.

Dangers and Beneficial Possibilities

After these experiences, I have not dared to expose entire human beings to a condenser field and in this way produce artificial rises in temperature. The responsibility seemed too great. At the same time, effective heating of the body can be accomplished equally well in other ways; such as with the well-known Apostoli "condenser bed," which can be connected to any diathermal apparatus. With the method previously described, only serious dangers for the patients treated would have been conjured up, without the possibility of producing a fundamentally new effect.

On the other hand, the disturbances of the physical health, which we could observe in the field of free radiation of powerful transmitters, have never been serious. After a period of recovery of a few weeks, with no irradiation, all effects have been observed to vanish. For four years now, I have almost daily worked for several hours at a transmitter with 1½ kw. plate dissipation; and the effects, often very unpleasant, have always gone back to normal on stopping the work.

In these things the wavelength is also certainly of importance, and in fact we have the impression that the disturbances to health became stronger as the wave was shortened. Anyway, the unpleasant sensations appear much quicker with a three-meter wave than with longer ones.

From all these experiences, it is at any rate clear that treatment with electric waves can in no way be regarded as always harmless for the human body. It is plain that their incorrect use can cause serious injuries to health. Certainly such injuries are to be expected only when the frequencies are very high; that is, with ultra-short waves; even then there is nothing to fear except with fairly high transmitter power.

On the other hand, with proper use, the short electric waves seem to be a valuable means of treatment. According to our experiments to date, with bacteria cultures and infected animals, the germs of disease can be killed. There is the added point that certain defensive processes are stimulated in the body. I have also already repeatedly treated human beings; and, in about a hundred cases, I have been able to attain an extraordinarily quick cure of suppuration (pus formation).

Ills Treated by Short Waves

● IN a dispatch to the N. Y. Times, a report from London states that cures for various ailments by the use of short radio waves were claimed by Dr. Erwin Schliephake, a German physician and scientist.

Writing in the *British Medical Journal* he described how he succeeded in treating deep-seated abscess in the human body by passing ultra-short wireless waves through the patient, who was not in immediate contact with any instrument. He found, he said, that various tissues exhibit different degrees of conductivity in the presence of these waves.

Dr. Schliephake declared he has used waves to treat pulmonary abscesses after pneumonia, in pleural empyema, pneumonic tuberculosis, in certain forms of peritonitis, in migraine and acute tonsilitis.

Dr. Willis R. Whitney, research director of the General Electric Company, revealed in April, 1930, that he had developed a radio type of apparatus for killing bacteria in the body.

From THEORY AND CALCULATION OF TRANSIENT PHENOMENON &
ELECTRIC OSCILLATIONS by Charles Proteus Steinmetz, 1920

CHAPTER V.

DISTRIBUTED SERIES CAPACITY.

43. The capacity of a transmission line, cable, or high-potential transformer coil is shunted capacity, that is, capacity from conductor to ground, or from conductor to return conductor, or shunting across a section of the conductor, as from turn to turn or layer to layer of a transformer coil.

In some circuits, in addition to this shunted capacity, distributed series capacity also exists, that is, the circuit is broken at frequent and regular intervals by gaps filled with a dielectric or insulator, as air, and the two faces of the conductor ends thus constitute a condenser in series with the circuit. Where the elements of the circuit are short enough so as to be represented, approximately, as conductor differentials, the circuit constitutes a circuit with distributed series capacity.

An illustration of such a circuit is afforded by the so-called "multi-gap lightning arrester," as shown diagrammatically in Fig. 90, which consists of a large number of metal cylinders $p, q \dots$, with small spark gaps between the cylinders, connected between line L and ground G. This arrangement, Fig. 90, can be represented diagrammatically by Fig. 91. Each cylinder has a capacity C_0 against ground, a capacity C against the adjacent cylinder, a resistance r,— usually very small,— and an inductance L.

The series of insulator discs of a high voltage suspension — or strain — insulator also forms such a circuit.

If such a series of n equal capacities or spark gaps is connected across a constant supply voltage e_0, each gap has a voltage $e = \frac{e_0}{n}$. If, however, the supply voltage is alternating, the voltage does not divide uniformly between the gaps, but the potential difference is the greater, that is, the potential gradient steeper the nearer the gap is to the line L, and this distribution of potential becomes the more non-uniform the higher the frequency; that is, the greater the charging current of the capacity of the cylinder against the ground. The charging currents against ground, of all the cylinders from q to the ground G, Figs. 90 and 91, must pass the gap between the adjacent cylinders p and q; that is, the charging current of the condenser represented by two adjacent cylinders p and q is the sum of all the charging currents from q to G; and as the potential difference between the two cylinders p and q is proportional to the charging current of the condenser formed by these two cylinders, C, this potential difference increases towards L, being, at each point proportional to the vector sum of all the charging currents, against ground, of all the cylinders between this point and ground.

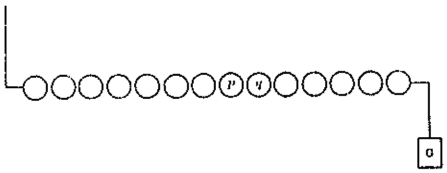

Fig. 90. Multi-gap lightning arrester.

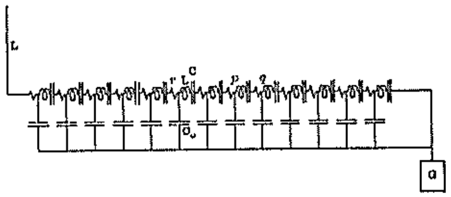

Fig. 91. Equivalent circuit of a multi-gap lightning arrester.

The higher the frequency, the more non-uniform is the potential gradient along the circuit and the lower is the total supply voltage required to bring the maximum potential gradient, near the line L, above the disruptive voltage, that is, to initiate the discharge. Thus such a multigap structure is discriminating regarding frequency; that is, the discharge voltage with increasing frequency, does not remain constant, but decreases with increase of frequency, when the frequency becomes sufficiently high to give appreciable charging currents. Hence high frequency oscillations discharge over such a structure at lower voltage than machine frequencies.

For a further discussion of the feature which makes such a multigap structure useful for lightning protection, see A. I. E. E. Transactions, 1906, pp. 431, 448, 1907, p. 425, etc.

44. Such circuits with distributed series capacity are of great interest in that it is probable that lightning flashes in the clouds are discharges in such circuits. From the distance traversed by lightning flashes in the clouds, their character, and the disruptive strength of air, it appears certain that no potential difference can exist in the clouds of such magnitude as to cause a disruptive discharge across a mile or more of space. It is probable that as the result of condensation of moisture, and the lack of uniformity of such condensation, due to the gusty nature of air currents, a non-uniform distribution of potential is produced between the rain drops in the cloud; and when the potential gradient somewhere in space exceeds the disruptive value, an oscillatory discharge starts between the rain drops, and gradually, in a number of successive discharges, traverses the cloud and equalizes the potential gradient. A study of circuits containing distributed series capacity thus leads to an understanding of the phenomena occurring in the thunder cloud during the lightning discharge.*

Only a general outline can be given in the following.

45. In a circuit containing distributed resistance, conductance, inductance, shunt, and series capacity, as the multigap lightning arrester, Fig. 90, represented electrically as a circuit in Fig. 91, let r = the effective resistance per unit length of circuit, or per circuit element, that is, per arrester cylinder; g = the shunt conductance per unit length, representing leakage, brush discharge, electrical radiation, etc.; L = the inductance per unit length of circuit; C = the series capacity per unit length of circuit, or circuit element, that is, capacity between adjacent arrester cylinders, and C_0 = the shunt capacity per unit length of circuit, or circuit element, that is, capacity between arrester cylinder and ground. If then f = the frequency of impressed e.m.f., the series impedance per unit length of circuit is

$$Z = r + j(x - x_c); \tag{1}$$

the shunt admittance per unit length of circuit is

$$Y = g + jb, \tag{2}$$

where

$$\left.\begin{aligned} x &= 2\pi f L, \\ x_c &= \frac{1}{2\pi f C}, \\ b &= 2\pi f C_0; \end{aligned}\right\} \tag{3}$$

or the absolute values are

$$\left.\begin{aligned} z &= \sqrt{r^2 + (x - x_c)^2} \\ \text{and} \quad y &= \sqrt{g^2 + b^2}. \end{aligned}\right\} \tag{4}$$

If the distance along the circuit from line L towards ground G is denoted by l, the potential difference between point l and ground by $\dot{E}$, and the current at point l by $\dot{I}$, the differential equations of the circuit are *

$$\frac{d\dot{E}}{dl} = Z\dot{I} \tag{5}$$

and

$$\frac{d\dot{I}}{dl} = Y\dot{E}. \tag{6}$$

Differentiating (5) and substituting (6) therein gives

$$\frac{d^2\dot{E}}{dl^2} = YZ\dot{E}. \tag{7}$$

Equation (7) is integrated by

$$\dot{E} = A_1 \epsilon^{-al} + A_2 \epsilon^{+al}, \tag{8}$$

where

$$a = \sqrt{YZ} = \alpha + j\beta, \tag{9}$$

$$\left.\begin{aligned} \alpha &= \sqrt{\tfrac{1}{2}\{yz + gr - b(x - x_c)\}} \\ \text{and} \quad \beta &= \sqrt{\tfrac{1}{2}\{yz - gr + b(x - x_c)\}}. \end{aligned}\right\} \tag{10}$$

---

* See paper, "Lightning and Lightning Protection," N.E.L.A., 1907. Reprinted and enlarged in "General Lectures on Electrical Engineering," by Author.

* Section III, Chapter II, paragraph 7.

Substituting (10) in (8) and eliminating the imaginary exponents by the substitution of trigonometric functions,

$$\dot{E} = A_1\varepsilon^{-al}(\cos\beta l - j\sin\beta l) + A_2\varepsilon^{+al}(\cos\beta l + j\sin\beta l). \quad (11)$$

46. However, if n = the total length of circuit from line L to ground G, or total number of arrester cylinders between line and ground, for $l = n$,

$$\dot{E} = 0, \qquad (12)$$

and for $l = 0$,

$$\dot{E} = e_0 = \text{the impressed e.m.f.} \qquad (13)$$

Substituting (12) and (13) into (11) gives

$$0 = A_1\varepsilon^{-an}(\cos\beta n - j\sin\beta n) + A_2\varepsilon^{+an}(\cos\beta n + j\sin\beta n)$$

and

$$e_0 = A_1 + A_2;$$

hence,

$$\left.\begin{aligned} A_1 &= \frac{e_0}{1 - \varepsilon^{-2an}(\cos 2\beta n - j\sin 2\beta n)}, \\ A_2 &= -A_1\varepsilon^{-2an}(\cos 2\beta n - j\sin 2\beta n), \end{aligned}\right\} \qquad (14)$$

and the potential difference against ground is

$$\dot{E} = e_0 \frac{\varepsilon^{-al}(\cos\beta l + j\sin\beta l) - \varepsilon^{-a(2n-l)}[\cos\beta(2n-l) - j\sin\beta(2n-l)]}{1 - \varepsilon^{-2an}(\cos 2\beta n - j\sin 2\beta n)} \qquad (15)$$

From equation (5), substituting (15) and (9), we have

$$\dot{I} = -\sqrt{\frac{Y}{Z}}\, e_0 \frac{\varepsilon^{-al}(\cos\beta l - j\sin\beta l) + \varepsilon^{-a(2n-l)}[\cos\beta(2n-l) - j\sin\beta(2n-l)]}{1 - \varepsilon^{-2an}(\cos 2\beta n - j\sin 2\beta n)}. \qquad (16)$$

Reduced to absolute terms this gives the potential difference against ground as

$$e = e_0\sqrt{\frac{\varepsilon^{-2al} + \varepsilon^{-2a(2n-l)} - 2\varepsilon^{-2an}\cos 2\beta(n-l)}{1 + \varepsilon^{-4an} - 2\varepsilon^{-2an}\cos 2\beta n}}, \qquad (17)$$

the current as

$$i = e_0\sqrt{\frac{y}{z}}\sqrt{\frac{\varepsilon^{-2al} + \varepsilon^{-2a(2n-l)} + 2\varepsilon^{-2an}\cos 2\beta(n-l)}{1 + \varepsilon^{-4an} - 2\varepsilon^{-2an}\cos 2\beta n}}, \qquad (18)$$

and the potential gradient, or potential difference between adjacent cylinders, is

$$e' = x_c i = e_0 x_c\sqrt{\frac{y}{z}}\sqrt{\frac{\varepsilon^{-2al} + \varepsilon^{-2a(2n-l)} + 2\varepsilon^{-2an}\cos 2\beta(n-l)}{1 + \varepsilon^{-4an} - 2\varepsilon^{-2an}\cos 2\beta n}}. \qquad (19)$$

For an infinite length of line, $n = \infty$, that is, for a very large number of lightning arrester cylinders, where ε^{-2an} is negligible, as in the case where the discharge passes from the line into the arrester without reaching the ground, equations (17), (18), (19) simplify to

$$e = e_0\varepsilon^{-al}, \qquad (20)$$

$$i = e_0\sqrt{\frac{y}{z}}\,\varepsilon^{-al}, \qquad (21)$$

and

$$e' = e_0 x_c\sqrt{\frac{y}{z}}\,\varepsilon^{-al}; \qquad (22)$$

that is, are simple exponential curves.

Substituting (4) and (3) in (21) and (22) gives

$$e' = e_0\varepsilon^{-al}\sqrt[4]{\frac{C_0^2 + \left(\frac{g}{2\pi f}\right)^2}{C^2\{[1 - (2\pi f)^2 CL]^2 + (2\pi f Cr)^2\}}} \qquad (23)$$

and

$$i = 2\pi f C e'; \qquad (24)$$

or, approximately, if r and g are negligible, we have

$$e' = e_0\varepsilon^{-al}\sqrt{\frac{C_0}{C\{1 - (2\pi f)^2 CL\}}} \qquad (25)$$

and

$$i = 2\pi f e_0\varepsilon^{-al}\sqrt{\frac{CC_0}{1 - (2\pi f)^2 CL}}. \qquad (26)$$

47. Assume, as example, a lightning arrester having the following constants: $L = 2 \times 10^{-8}$ henry; $C_0 = 10^{-13}$ farads;

$C = 4 \times 10^{-11}$ farads; $r = 1$ ohm; $g = 4 \times 10^{-6}$ mho; $f = 10^8$ = 100 million cycles per second; $n = 300$ cylinders, and $e_0 = 30{,}000$ volts; then from equation (3), $x = 12.6$ ohms, $x_c = 39.7$ ohms, and $b = 62.8 \times 10^{-6}$ mhos;

from equation (1),

$$Z = 1 - 27.1\,j \text{ ohms};$$

from equation (2),

$$Y = (4 + 62.8\,j)\,10^{-6} \text{ mho};$$

from equation (4),

$$z = 27.1 \text{ ohms and } y = 62.9 \times 10^{-6} \text{ mho};$$

from equation (10),

$$\alpha = 0.0021 \text{ and } \beta = 0.0412;$$

from equation (17),

$$e = 35{,}500\sqrt{\varepsilon^{-0.0042\,l} + 0.08\,\varepsilon^{+0.0042\,l} - 0.568\cos(24.72 - 0.0824\,l)};$$

from equation (18),

$$i = 54\sqrt{\varepsilon^{-0.0042\,l} + 0.08\,\varepsilon^{+0.0042\,l} + 0.568\cos(24.72 - 0.0824\,l)},$$

and from equation (19),

$$e' = 2140\sqrt{\varepsilon^{-0.0042\,l} + 0.08\,\varepsilon^{+0.0042\,l} + 0.568\cos(24.72 - 0.0824\,l)}.$$

Hence, at $l = 0$, $e = 30{,}000$ volts, $i = 64.6$ amperes, and $e' = 2560$ volts; and at $l = 300$, $e = 0$, $i = 57.5$ amperes, and $e' = 2280$ volts.

With voltages per gap varying from 2280 to 2560, 300 gaps would, by addition, give a total voltage of about 730,000, while the actual voltage is only about one-twenty-fourth thereof; that is, the sum of the voltages of many spark-gaps in series may be many times the resultant voltage, and a lightning flash may pass possibly for miles through clouds with a total potential of only a few hundred million volts. In the above example the 300 cylinders include 7.86 complete wave-lengths of the discharge.

Los Angeles Times

Jan 3, 1988

Ham Radio Operators' High Cancer Rate Poses a Puzzle

TACOMA, Wash. (AP)—Amateur radio operators in Washington and California appear to die at abnormally high rates from several forms of cancer, suggesting a possible link between cancer and electromagnetic fields, according to data collected by a state epidemiologist.

Others cautioned that evidence of such a link has been inconsistent and that other factors may be involved.

Dr. Samuel Milham Jr. of the Washington Department of Social and Health Services studied the deaths of 2,485 Washington and California ham operators between 1979 and 1984.

He reported in the American Journal of Epidemiology that 29 leukemia deaths would be expected in a group of people that size, but he found 36 deaths. Statistically, he expected to find 72 lymphatic and blood-forming organ cancers, but found 89. And he expected to find 67.6 deaths from prostate cancer, but found 78.

'Excess Mortality'

The study "indicates that amateur radio operator licensees in Washington state and California have significant excess mortality due to acute myloid leukemia, multiple myeloma and perhaps certain types of malignant lymphoma," Milham reported.

Leonard Sagan, program manager for radiation studies at the Electric Power Research Institute in Palo Alto, Calif., warned that studies like Milham's can be misinterpreted. Milham's findings could be simple associations that have nothing to do with cancer causes among people who work with electricity, he said.

Sagan said his own research hasn't shown whether electricity causes cancer. But he predicted that if it does, it's rare.

"I think it would be unfortunate for people with children to become overly concerned about this," he said. "Evidence is weak and inconsistent and in a few years we will have better evidence."

Here's something to consider while using a Multi Wave Oscillator. Just "any old" radio frequency broadcast energy may not be desireable. Use this handbook in your detective work to discover Lakhovsky's secret.

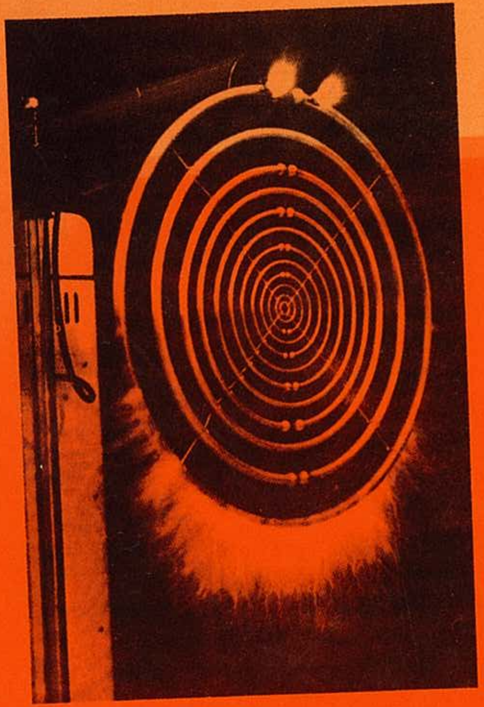

Is This Device
AN ELECTRONIC
FOUNTAIN OF YOUTH?!?

Is it possible to restore aging or diseased tissue? Earlier this century scientist Georges Lakhovsky conducted researches indicating that living cells can be regenerated by applying special fields of electrical energy.

It has been theorized that every cell in your body has a specific vibratory rate. When the cell ages or becomes diseased its rate changes. The Multiple Wave Oscillator produces a full spectrum of vibratory rates that resonate with cells to restore them to oscillatory equilibrium.

Earlier this century this Multiple Wave Oscillator was shown to neutralize certain types of cancer--but no exact plans are available. This book contains the records of many people who have attempted to reconstruct the original equipment and continue with this fascinating research.

This book describes researches such as the enhancement of plant grown through the amplification of cosmic waves with simple wire coils, and the neutralization of cancerous tissue with radio waves. This was done with cosmic radiations ringing wires coiled around plants, and with refined electrical fields ringing specially designed antennas. Lakhovsky's original work into the dramatic impact of the subtle forces of nature included the effects of sunspot cycles on plant and animal health, the nature of cancer and how it appears in clusters over areas which have soils that reflect cosmic radiations, thereby causing interference patters, and how diseased cells would turn to a state of health when specific natural or electrical fields were applied in a proper manner.

Nikola Tesla provided some of the early electro-therapeutic research and Lakhovsky was introduced to this through Professor D'arsonval. Lakhovsky, a brilliant scientist, combined this knowledge with his original work and produced a device known as the Multiple Wave Oscillator.

The Multiple Wave Oscillator is an ongoing research project of Borderland Sciences Research Foundation, begun in 1963. This book is the published record of our search for Lakhovsky's secrets and the potential for reversing the aging process.

ISBN 0-945685-03-3
51795>
9 780945 685036

BORDERLAND SCIENCES
RESEARCH FOUNDATION
P.O. BOX 429 ★ GARBERVILLE, CA 95440-0429 ★ U.S.A.

Electro-Biology
$17.95

THE LAKHOVSKY MUPTIPLE WAVE OSCILLATOR HANDBOOK

NEW RESEARCH FINDINGS show that all diseases have simple explanations and cures once their true cause is known. This book describes the causes of both common and extraordinary diseases and gives specific instructions for their cure.

The Cure for all Diseases

With Many Case Histories

of diabetes, high blood pressure, seizures, chronic fatigue syndrome, migraines, Alzheimer's, Parkinson's, multiple sclerosis, and others showing that all of these can be simply investigated and cured.

Hulda Regehr Clark, Ph.D.,N.D.

Electricity can now be used to kill bacteria, viruses and parasites in minutes, not days or weeks as antibiotics require.

If you have been suffering from a chronic infection or have cancer, or AIDS, learn to build the electronic device that will stop it immediately. It is safe and without side effects and does not interfere with any treatment you are now on.

The Cure For All Diseases

© Copyright 1995 by Hulda Regehr Clark, Ph.D., N.D.
All rights reserved.

Permission is hereby granted to make copies of any part of this document for non-commercial purposes provided this page with the original copyright notice is included.

Published in the United States by New Century Press
1055 Bay Blvd., Suite C, Chula Vista, CA 91911
(619) 476-7400, (800) 519-2465
www.newcenturypress.com
ISBN 1-890035-01-7

Other books by Dr. Clark available from New Century Press:
The Cure For All Cancers
Heilverfahren Aller Krebsarten (Cure For Cancers, German)
The Cure for All Cancers (Japanese Translation)
The Cure For All Cancers (Korean Translation)
Heilung ist möglich (Cure For All Diseases, German)
The Cure For HIV And AIDS
The Cure For All Advanced Cancers

30 29 28 27 26 25 24 23 22 21

Notice to the Reader:

The opinions and conclusions expressed in this book are mine, and unless expressed otherwise, mine alone. The opinions expressed herein are based on my scientific research and on specific case studies involving my patients. Be advised that every person is unique and may respond differently to the treatments described in this book. On occasion we have provided dosage recommendations where appropriate. Again, remember that we are all different and any new treatment should be applied in a cautious, common sense fashion.

The treatments outlined herein are not intended to be a replacement or substitute for other forms of conventional medical treatment. Please feel free to consult with your physician or other health care provider.

I have indicated throughout this book the existence of pollutants in food and other products. These pollutants were identified using a testing device of my invention known as the Syncrometer.™ Complete instructions for building and using this device are contained in this book. Therefore anyone can repeat the tests described and verify the data.

The Syncrometer is more accurate and versatile than the best existing testing methods. A method for determining the degree of precision is also presented. However at this point it only yields positive or negative results, it does not quantify. The chance of a false positive or a false negative is about 5%, which can be lessened by test repetition.

It is in the public interest to know when a single bottle of a single product tests positive to a serious pollutant. If one does, the safest course is to avoid all bottles of that product entirely, which is what I repeatedly advise. These recommendations should be interpreted as an intent to warn and protect the public, not to provide a statistically significant analysis. It is my fervent hope that manufacturers use the new electronic techniques in this book to make purer products than they ever have before.

Dedication

I would like to dedicate this book to all the persons who visited me professionally, from the very first person in 1963, **Mrs. R. Biehl**, to the present. I learned so much from each of you and I appreciate your confidence, your intelligence and your reluctance to be defeated.

Acknowledgments

I would like to express my gratitude to my son, **Geoffrey**, who always listened to my “crazy ideas” on Sundays, right at supper time. He was patient, kind, helpful and willing to share his expertise in electronics and computers. Without him, this book could not have been written.

A big thank you goes to **Frank Jerome, DDS**. Without the loan of his parasite slide collection many of my discoveries could not have been make, and without his development of metal-free dentistry, many of these patient histories could not have ended happily. Thank you to **Linda Jerome** for nurturing us both with personal interest and patience. And thank you to **Edna Bernstein** for her resourceful assistance.

Contents

FIGURES .... **XIII**

PREFACE .... **XVII**

THE PROMISE **1**
Only Two Health Problems 2
Be A Health Detective 3

THE DISCOVERY **5**
Everything Has A Unique Frequency 8
Selective Electrocution 9
Bioradiation 12
Zapping Bugs 13
The Bioradiation Spectrum 16

BUILDING A ZAPPER **19**
Killing The Surviving Pathogens 20
Zapper Parts List 21
Assembling The Zapper 23
Using The Zapper 28
Simple Pulser 28
Using The Simple Pulser 30

PARASITES & POLLUTION **31**
The Worst Parasite 33
Pollution 36
Solvent Pollution 36
Metal Pollution 37
Mycotoxins 38
Physical Toxins 38
Chemical Toxins 40

HOW WE REALLY GET SICK 43
Self Health 46
How To Heal 47
The Road To Wellness 49

PAIN FROM TOE TO HEAD 51
Searching For Bacteria 53
What You Will Find 54
Toe Pain 55
Balance Your pH 57
Foot Pain 59
Heel Pain 61
Leg Pain In Children 69
Leg Pain In Adults 70
Leg Pain Protocol 73
Fibromyositis and Fibromyalgia 75
Joint Pain or Arthritis 78
Osteo or Common Arthritis 79
Rheumatoid Arthritis 82
Allergic Arthritis 84
Thigh Pain 86
Hip Pain 87
Bone Strengthening 87
Groin Pain 89
Low Back Pain 92
Lower Abdominal Pain 97
Stomach Pain 98
Appendicitis 100
Urinary Tract Pain 101
Irritable Bowel Syndrome(IBS), Colitis, and Spastic Colon 102
Crohn's Disease 103
Central Abdominal Pain 108
Interstitial Cystitis 108
Uterine Pain 109
Endometriosis 109

- Contraception ... 112
- The Silent Cervix ... 112
- Menopausal Symptoms, Hot Flushes, PMS ... 113

Infertility ... 117

- Nausea ... 121

Prostate Problems ... 124

Side Pain ... 128

Midabdomen Pain, Stomach Pain ... 131

Stomach Ulcer ... 132

Hiatal Hernia ... 133

Respiratory Illness ... 134

- Asthma ... 134
- Bronchitis, Croup, Chronic Cough ... 140

Breast Pain ... 142

Breast Sensitivity ... 143

- Breast Lumps ... 144

Heart Pain ... 146

- Slow Pulse/Syncope (Passing Out) ... 149

Chest Pain ... 149

Upper Back Pain ... 150

Shoulder Pain ... 151

Upper Arm Pain ... 153

Elbow Pain ... 154

Wrist Pain ... 154

Thumb Pain and Hand Pain ... 155

Finger Pain ... 155

Back Of Neck Pain ... 156

Front Neck Pain ... 156

Temporal Mandibular Joint (TMJ) Problems ... 157

Tooth Ache ... 158

Throat Pain ... 159

Eye Pain ... 160

- Macular Degeneration ... 162

Headache and Migraines ... 162

Earache ... 168

Ear Noises And Ringing ... 169

Scalp Pain 171

NON-PAINFUL DISEASES 173
Diabetes 173
Prediabetic 178
Diabetes Of Childhood 178
Herpes 179
Fatigue 182
Skin Problems 188
Sebaceous Cysts 188
Skin Rash 189
Hives 191
Acne 192
Psoriasis, Eczema 192
Warts 195
Tapeworm Stages 197
The Flu 201
Fever 203
Multiple Sclerosis & Amyotropic Lateral Sclerosis 204
High Blood Pressure 210
Glaucoma 213
Tooth Decay 214
Muscle Diseases 218
Muscular Dystrophy 219
Myasthenia Gravis 221
Universal Allergies 222
Environmental Illness 223
Alcoholism 227
Seizures 230
Weight Problems 237
Sleep Problems 241
Sleep Apnea 241
Insomnia 243
Yeast Infections 246
Diaper Rash 246
Treat Yeast or Fungus the Same 247
Fluke Disease 249

Flukes and Solvents....251
Flukes Not Alone....252
Burning And Numbness....253
Depression....256
Manic Depression....257
Schizophrenia....261
Autism....263
Digestion Problems....266
Alzheimer's....269
Dementias, Memory Loss....272

PUSHING BACK AGE 275
Diet....275
Increase Oxygen....278
Niacin....279
TIA’s, Strokes and Purpura....281
Blood Pressure....281
Air Pollution....282
Anemia....285
Acid Levels....287
Increase Minerals....287
Chelation....289
Pulse....289
Heart Health....290
Incontinence....293
Bad Digestion....295
Chewing....295
Stomach Acid....296
Salmonella and Shigella....297
Other Clues....299
Liver Bile....299
Constipation....301
Poop Your Troubles Away....301
Sugar Regulation....302
Elderly Person’s 7 Day Diabetic Diet....302
Breakfast....302
Lunch....303
Dinner....304

Supper 307
Diabetic Supplements 309
Diabetic Eating Out 309
Extra Diet Tips For The Elderly 309
Salt 310
Tremor 312
Tremor Remedies 313
Weakness 317
Heart Disease 318
Heart/Kidney Relationship 320
Diuretics 321
Feeling Comfortable 322
Hearing Loss 324
Insomnia 326
Healthful Habits 326
Not Old Yet 328
Super Longevity 329
Dying 330

CURING CANCER 331
Another Fluke Disease 331
Teamed With A Solvent 332
Propyl Alcohol Polluted Products 335
Getting Well After Cancer 336
Herbal Parasite Killing Program 338
Parasite Program Handy Chart 341
Maintenance Parasite Program 341
Pet Parasite Program 343
Parasites Gone, Toxins Next 345
Watch For Bacteria 346
Help Your Family, Too 347
Aflatoxin 348

CURING HIV/AIDS 351
Method Of Treatment 353
Benzene Polluted Products 354

Plan For The Future 356

CURING THE COMMON COLD 357

Tapeworm Stage or Mites.. 357

Molds and Colds.. 360

The Cure 363

Mold Free Diet 365

Prevention.. 368

Curing Symptoms 369

True Origins Of Viruses.. 370

BODY WISDOM AND WHY BAD FOOD TASTES GOOD373

Salt.. 373

Fats 374

Starch.. 374

Sugar.. 375

Dislikes.. 375

Outwit The Cravings 377

Back To Normal.. 379

Toxic Food.. 379

Moldy Food 381

Aflatoxin...382

Eradicating Aflatoxin...383

Zearalenone 384

Sterigmatocystin...385

Ergot...385

Cytochalasin B 387

Kojic Acid 388

T-2 Toxin 388

Sorghum Molds 388

Patulin...388

Mold Avoidance 389

Getting Away From Grains..392

Pets Teach Health Lessons.. 394

EASY LIFESTYLE IMPROVEMENTS 397

Living Hand To Mouth.. 397

Better Laundry Habits 398
Better Kitchen Habits 399
Better Housekeeping 402
Windows Open Or Closed? 405
What Kind Of Heat 407
Getting Rid of Mites 407

FOUR CLEAN-UPS 409
Dental Clean-up 409
Guidelines For A Healthy Mouth 411
Horrors Of Metal Dentistry 416
Dental Rewards 419
Diet Clean-up 420
Breakfast 420
Lunch 422
Supper 423
Beverages 424
Horrors In Commercial Beverages 428
Food Preparation 430
Food Guidelines 431
Dining Out 433
Body Clean-up 435
Stop Using Supplements 439
Safe Supplements 441
Home Clean-up 444
Your Basement 444
Your Garage 445
Special Clean-up for Freon (CFCs) 446
Special Clean-up for Fiberglass 447
Special Clean-up for Asbestos 447
Your House 448
Unnatural Chemicals 451

BIOELECTRONICS 457
Making A Syncrometer 457
Making a Hobby Kit Syncrometer 458
Making Test Plates 460
Test Plates Parts List 460

Test Plates Assembly....................461
Using The Syncrometer.................... 462
Resonance....................465
Lesson One....................466
White Blood Cells....................468
Making a White Blood Cell Specimen....................469
Lesson Two....................470
Lesson Three....................470
Lesson Four....................471
Preparing Test Substances....................472
Impure Test Substances....................472
Making Organ Specimens....................477
Making a Complete Set of Tissue Samples....................478
Purchasing a Complete Set of Tissue Samples....................478
Body Fluid Specimens....................479
Lesson Five....................481
Lesson Six....................482
Lesson Seven....................484
Testing the Air....................485
Testing Someone Else....................486
Lesson Eight....................486
Saliva Testing....................487
Lesson Nine....................487
Surrogate Testing....................488
Lesson Ten....................489
Lesson Eleven....................491
Lesson Twelve....................492
Lesson Thirteen....................493
Lesson Fourteen....................493
Lesson Fifteen....................494
Lesson Sixteen....................494
Lesson Seventeen....................495
Sensitivity of Pollutant-In-Product Testing....................497
Microscopy Lesson....................498
Taking Pictures Of What You See....................501
Using A Frequency Generator.................... 502
The Theory....................502
Lesson Eighteen....................504
Lesson Nineteen....................505
Lesson Twenty....................506

Lesson Twenty One 507
Lesson Twenty Two 507
Lesson Twenty Three 508
Lesson Twenty Four 508
Lesson Twenty Five 509
Lesson Twenty Six 509
Lesson Twenty Seven 509
Lesson Twenty Eight 510
Lesson Twenty Nine 510

RECIPES 513
Beverage Recipes 513
Lemonade 513
Fresh Tomato Juice 513
Fresh Pineapple Juice 514
Maple Milk Shake 514
Yankee Drink 514
Hot Vanilla Milk 514
Red Milk 515
C-Milk 515
My Own Soda Pop 515
My Own Super C-Pop 516
Half And Half 516
Buttermilk-C 516
Food Recipes 516
Daily Foods 517
Two Granolas 518
Peanut Butter 519
Sweetening and Flavoring 519
Preserves 520
C Dressing 521
Cheese Sauce 521
Sour Cream-C 521
Yogurt 521
Soups 521
Fish and Seafood recipe 522
Baked Apples 522
5 Minute Ice Cream 522
Seven Day Sample Menu 523
Recipes for Natural Body Products 525

Borax Liquid Soap....526
For Laundry....526
For Dishes....527
In The Dishwasher....527
In The Sink....527
Shampoo....528
Hair Spray....529
Homemade Soap....529
Liquid Soap....530
Skin Sanitizer....530
Deodorant....531
Brushing Teeth....532
For Dentures....533
Mouthwash....533
Contact Lens Solution....533
Lip Soother....534
Foot Powder....534
Skin Healer Moisturizer Lotion....534
Other Skin Healers....535
Massage Oil....536
Sunscreen Lotion....536
Nose Salve....536
Quick Corn Starch Skin Softener....537
Cornstarch Skin Softener....537
After Shaves....537
Personal Lubricants....537
Baby Wipes....538
People Wipes....538
Recipes For Natural Cosmetics....539
Eye liner and Eyebrow Pencil....539
Lipstick....539
Face Powder....540
Blush (face powder in a cake form)....540
Recipes For Household Products....540
Floor Cleaner....540
Furniture Duster and Window Cleaner....540
Furniture Polish....541
Insect Killer....541
Ant Repellent....541
Flower and Foliage Spray....542

Moth Balls 542
Carpet Cleaner 542
Health Improvement Recipes 543
Black Walnut Hull Tincture 543
Quassia recipe 546
Emmenagogue (Menstrual Period Inducer) 546
Bowel Program 546
Constipation Tea 548
Weight Reduction 548
Kidney Cleanse 549
Liver Herbs 552
Liver Cleanse 552
Lugol’s Iodine Solution 559
Vitamin D Drops 560

PATHOGEN FREQUENCIES 561
Bandwidth Of Organism Families 562
Mold, Mold Toxin Frequencies 562
Bacteria and Viruses 563
Roundworms, Flatworms, One-celled Animals 566
Wart Frequencies 569
Tapeworms 569
Mite Frequencies 570
Miscellaneous Frequencies 571
Toxic Elements 571
Solvents 574
Pathogen Frequency Chart 576

SOURCES 581

FINALE 589

INDEX 591

Figures

Fig. 1 Selected animal bandwidths. 17
Fig. 2 Selected pathogen bandwidths. 18
Fig. 3 Zapper schematic. 22
Fig. 4 Finished zapper, outside and inside. 27
Fig. 5 Simple pulser 29
Fig. 6 Ascaris. 32
Fig. 7 (L to R) Human intestinal fluke, pancreatic fluke, sheep liver fluke, and human liver fluke. 33
Fig. 8 Fasciolopsis' normal life cycle 35
Fig. 9 Feed animals out of your yard and away from where children play. Their wastes contain pathogens. 50
Fig. 10 Your kidneys with the adrenal glands sitting like hats on top. The ureters lead to the bladder 60
Fig. 11 All disposable diapers I tested had mercury and thallium! These toxins can be absorbed through babies' skin. 69
Fig. 12 Do not use any commercial personal products, the risk of pollution is too great. 71
Fig. 13 Safe substitutes for personal products 72
Fig. 14 Tooth numbering system. 74
Fig. 15 Trichinella larvae settled in muscles 76
Fig. 16 Hookworms, strongyles, and whipworms 79
Fig. 17 Stone ground corn tortillas, high in calcium. 81
Fig. 18 Your lymph nodes are your best friends 90
Fig. 19 Colon. 97
Fig. 20 Keep personal water bottles sterile. 98
Fig. 21 Estrogen and progesterone levels (pg./ml) 114
Fig. 22 The liver has a large lobe on your right side with the gallbladder tucked inside. The left lobe is smaller. 128
Fig. 23 Gallstones 130
Fig. 24 Stomach, esophagus, diaphragm, gallbladder. 132
Fig. 25 Hiatal hernia. 133
Fig. 26 Lymph node neck glands 157
Fig. 27 Eye. 160
Fig. 28 Bottling equipment should be rinsed with ethyl (grain) alcohol, not propyl alcohol or wood alcohol. 183

Fig. 29 Tight diapers are a modern atrocity, forcing mercury and thallium into the baby's sponge-like skin. 190
Fig. 30 Some cysticercus varieties (types) have multiple heads....... 197
Fig. 31 Emerged cysticercus. 198
Fig. 32 Sheep liver flukes. Black threads in toilet are indicative of fluke remains. 222
Fig. 33 I found all calorie boosters to be polluted with wood alcohol. Make your own. 240
Fig. 34 Alcoholic beverages contain ergot and aflatoxin. Add vitamin C (1/8 tsp.) upon opening and wait 10 minutes for it to act..... 280
Fig. 35 Chlorinated water can cause mental problems. 283
Fig. 36 Red blood cells, top and side view. 285
Fig. 37 Water pitcher with filter. 288
Fig. 38 Sea salt flats are often roosting places for sea gulls. 311
Fig. 39 Too hot and too cold. 323
Fig. 40 Products with propyl alcohol. 334
Fig. 41 U.S. regulations on sterilizing solutions. 348
Fig. 42 Keep both salt and vitamin C powders in closable plastic shakers. 349
Fig. 43 Grilled food has benzopyrenes as do hot dogs and "smoke flavored" foods. 352
Fig. 44 Mexican made candy with no benzene. 354
Fig. 45 Pollutants are in unlikely places. 355
Fig. 46 Every brand of popcorn and corn chips I tested had zearalenone contamination. 355
Fig. 47 This Mexican bread has no molds. It is made from white, unbleached flour and is not wrapped in plastic. 362
Fig. 48 Goods baked in a panaderia (Mexican style bakery) never tested positive to molds. 367
Fig. 49 Starter set of homeopathic remedies. 369
Fig. 50 All U.S. brands of pasta I tested had mold, including health food brands like the one shown here (left). No Mexican brands of pasta, like the one pictured (right), had any molds.......... 385
Fig. 51 All cold cereals I tested were full of mold toxins (besides solvents), health food varieties were worst. 387
Fig. 52 Packaged herb tea is moldy and polluted with solvents. Get yours in bulk from an herb company. 390
Fig. 53 Three safe flavorings. 391
Fig. 54 Don't eat the green on the potato. 392
Fig. 55 Potato harvest of the future. 393

Fig. 56 Smooth floors allow every bit of dust to be removed. 404
Fig. 57 Moving into the summer kitchen got you away from the accumulation of filth from winter! 406
Fig. 58 Mite 408
Fig. 59 Dental replacements 411
Fig. 60 Tops and bottoms of some metal crowns. 412
Fig. 61 Thallium excerpt. 418
Fig. 62 More dental metal 419
Fig. 63 Beautiful plastic mouth. 419
Fig. 64 Ugly metal in mouth. 420
Fig. 65 Unpolluted breakfast cereals. 421
Fig. 66 Lawful uses of solvents in food 428
Fig. 67 Some unsafe beverages. 430
Fig. 68 Some good foods 432
Fig. 69 All breads I tested had mold if they were in plastic 433
Fig. 70 Good restaurant foods 434
Fig. 71 Examples of commercial “food lube.” 437
Fig. 72 Detergents with PCBs. 438
Fig. 73 Safe soaps 438
Fig. 74 Some polluted supplements 439
Fig. 75 Pure salt 440
Fig. 76 Laboratory salt label 441
Fig. 77 Garage fumes. 445
Fig. 78 Our future, unless we act 456
Fig. 79 Syncrometer schematic 458
Fig. 80 Syncrometer parts list. 459
Fig. 81 Bottle with white blood cells taped on top. 470
Fig. 82 Some purchased parasite and tissue slides. 479
Fig. 83 Table of the future. 483
Fig. 84 Microscope, slides and coverslips 498
Fig. 85 Setup for finding resonant frequencies 503
Fig. 86 Function generators 512
Fig. 87 Light colored, roasted peanuts in the shell had no aflatoxin. 519
Fig. 88 Make a bottle of borax liquid to fill your soap dispensers and shampoo bottle. Use citric acid to rinse and condition. 528
Fig. 89 These are gallstones. 554
Fig. 90 Chart of bandwidths for organism families. 562

Preface

The sick have been held hostage for their money or intangible assets since time immemorial. Doctors, even primitive and natural healers, surround themselves with mystery as they use herbs or chemicals and incantations or “prognoses” to help the sick recover. Today, the medical industry (doctors and their suppliers and insurers) take a significant amount of the worker's earnings. Wouldn't it be nice if they could all go back to gardening or some other primitive and useful endeavor? Wouldn't it be wonderful if the sick could join them?

The most promising discovery in this book is the effectiveness of electricity to kill viruses, bacteria and parasites. Does this mean you can cancel your appointment with your clinical doctor? No it does not. Killing your invaders does not make you well instantly. But happily, at your next doctor visits she or he will be removing drugs, not adding them.

You might think that such an invention should be quickly patented. That was my universal advice. But I chose not to. It helps me, my children, and my grandchildren, if you are well. The whole world needs to come out of the dark ages of medicine and illness. And to learn the true causes of infection and disease. We must and can usher in the **new age of disease-free living.**

No diabetes, no high blood pressure, no cancer, no HIV/AIDS, no migraines, no lupus and so on!

Not a single disease is left unconquerable with this new understanding!

The Promise

Step into a new world.

A world without chronic diseases.

Step out of your old world.

It has kept you a prisoner.

Try something new.

The prison has no walls. It has only lines. Lines that mark the ground around you. Inside the lines are your old ideas. Outside are new ideas that invite you to step over and escape your prison. Dare to try these new ideas and your illness promises to recede. In a few weeks it can be gone.

If you are very ill or chronically ill you must have asked yourself many times: why have these problems chosen me? Will there never be a way to conquer them?

You may be quite familiar with your doctor's explanation of your illness or your child's illness. A *Coxsackie* virus has entered your child's brain causing inflammation (encephalitis) there. You pray that your child's immunity will overcome it. You may be familiar, but so very helpless against this microscopic invader.

If you had the proverbial 3 wishes they would be: 1) please spare my child's life; 2) please make it so my child doesn't have permanent damage; 3) please bless and guide the wonderful doctors and nurses who are keeping the oxygen tent going, and are watching my child's temperature and vital signs.

What if you could turn a dial and in 3 minutes kill every Coxsackie virus in your child's body?

What if this had no side effects?

What if the virus never came back?

In this book you will learn how to do that. You will also learn why your child got encephalitis or other disease and how to prevent it forever.

If this is too mind boggling, just take it a step at a time: First, learn about the radio-type broadcasting that all living animals do. Second, find the “station frequencies” that your particular invader(s) broadcast at. Third, learn how to “jam” their frequency until they **expire**: it takes only minutes!

Finally, learn how to make your own diagnostic and treatment devices. The instructions are simple enough for anyone. Only by putting this power in your hands will it be safe from government regulation, however well intended.

Only Two Health Problems

No matter how long and confusing is the list of symptoms a person has, from chronic fatigue to infertility to mental problems, I am sure to find only two things wrong: they have in them **pollutants** and/or **parasites**. I never find lack of exercise, vitamin deficiencies, hormone levels or anything else to be a primary causative factor. So the solution to good health is obvious:

Problem	**Simplest Cure**
Parasites	Electronic and herbal treatment
Pollution	Avoidance

It's a valiant quest: The quest for health. With optimism in one hand and determination in the other, you too can work the miracles for yourself that my clients accomplished in the case histories.

More good news is that it is not expensive. The cost will range from a few hundred dollars to only a few thousand in order to eliminate both problems and cure your chronic diseases.

Be A Health Detective

After curing your own diseases, teach your friends and family how it's done. Families are related and their problems are related. This should make the task easier. Keep a small notebook to become part of the treasured family legacy as much as photographs do. If your aunt, father and brother had diabetes as well as yourself and all were cured after introducing them to this concept and technology, isn't this worthy of notes in your family's history?

Notice what a strong line of inheritance there can be, **not due to sharing genes** but due to sharing a roof, a table, a supermarket, and a dentist!

Many problems can be disinherited. Cure yourself of retinitis pigmentosa, Muscular dystrophy (the “inherited” kind), and break down your family's faith in the gene-concept for these diseases. Bring hope to your family by proving diseases' true etiology. Bring respect back for your loyal genes that bring you hair color, and texture, not hair loss. That bring you eye color, not eye disease. Your genes brought you the good things about your ancestors, not the bad things. Parasites and pollution brought you the bad things.

Killing all your invaders is just the first step, though. It is indeed the life-saving step. But getting well is more than saving your life. Next comes the more tedious task of finding their sources. Where did they come from? Why did they invade you so massively. Why you?

The story of your personal pollution unfolds as in a book. Look closely and you see the whole panorama of your numerous tiny invaders being held at bay by your valiant immune system, your white blood cells. You can see what they are fighting besides the invaders. Your ill-chosen diet and lifestyle products!

Your heart may go out to those tiny white blood cells. Never again, you may say, will you give them arsenic and mercury and lead. Never again, cobalt and asbestos and freon.

That great body of wisdom, your body, the same as listened to your three wishes, will reward you over and over as you co-operate with it, until you have had not 3 but 30 wishes granted, each one seemingly as impossible as climbing Mt. Everest.

- Your chronic yeast infection can go away.
- Your hair can stop falling out–might even grow back.
- Your body can become pregnant–when you had already given up.
- Your fatigue can vanish.
- Your insomnia can be gone.
- Your warts can fall off.
- Your sight and hearing can sharpen.
- Your constant hunger can disappear.

Health isn't just being free of sickness. Health is feeling great, feeling like laughing at funny things. Health is feeling grateful to be alive. It is feeling happy to see the sky and to see growing things and to feel confident in human society's progress. Health is remembering the good parts of childhood and believing you still have a lot of them.

The Discovery

What makes me think I can find things in the human body that a blood test can not? What new technology makes this possible? Why is electronic testing superior in many ways to chemical methods? What are my claims of electrically killing parasites based on?

In 1988 I discovered a new way to scan a body organ. It was electronic. We already can "see" an organ with a sonogram, X-rays, computerized tomography (CAT) scan, or with magnetic resonance imagery (MRI). These techniques can identify abnormal shapes in an organ without having to explore or guess. But my new electronic technique can check for viruses, bacteria, fungi, parasites, solvents and toxins, and in addition is simple, cheap, fast and infallible. Electricity can do many magical things; now we can add detecting substances in our body to that list.

The method rests on radio electronic principles.

If you match, very precisely, the capacitance and inductance properties of an external circuit so that its resonant frequency is the same as the emitted frequency coming from somewhere else, the circuit will oscillate. This means there will be positive feedback in an amplifier circuit. You can hear it. Like when a public address system squeals.

The external circuit I use is called an *audio oscillator*, quite easy to build or buy. Your body provides the emitted frequencies. When you combine the audio oscillator circuit with your body, and you hear *resonance*, then you have detected a match! Something in your body matches something in the circuit on the test plate. By putting a laboratory sample of, say, a virus on the test plate, you can determine if your body has that virus by listening for resonance. Hearing resonance is easy if you're a radio technician or musician. Others must patiently practice. The details are given in the Bioelectronics chapter (page 457).

You do not have to be an expert in anything to learn the electronic detection method. But a keen sense of hearing helps.

In 1988 I learned a way to put anything on my skin, blindfolded, and identify it electronically in a few minutes. I could taste something without flavor and identify it electronically. The system worked fine for detecting things in the skin and tongue. Would it be reliable for internal organs, too?

A whole world of discovery lay ahead of me. I wanted to know what was in my inner ear causing tinnitus, in my eyes causing pain, in my stomach causing indigestion and a thousand other things.

But behind the daily excitement of new discoveries, a gnawing question lingered in my mind. How is this possible without some pretty high frequency energy source, radio frequency in fact, running through my circuit? My audio oscillator was only 1000 Hz (*hertz*, or cycles per second); radio frequency is hundreds of thousands of Hz. And the phenomenon could be produced with an old-fashioned *dermatron*[1], too, that only puts out DC (direct current)–no frequencies at all!

A high frequency energy had to be coming from somewhere. Was it me? *Ridiculous*!

But there was a way to test. If my own body was putting forth the high frequency energy, it could be bled off and diverted into the ground with a correct size capacitor. This should stop the feedback oscillations. This turned out to be true; it was stopped. But *ridiculous* kept ringing in my ears and I tried another test. If there was indeed radio frequency (RF) running through my circuit I should be able to block it with the right snap-on choke. It did block. I thought of a third test. If this was truly a resonance phenomenon I should be able to add a capacitance to this circuit and see the resonance destroyed. Then add an inductance and see the resonance return. It did just that. I made graphs of the relationship between capacitance and inductance. They were entirely reproducible.

---

[1] The dermatron was invented decades ago and made famous by Dr. Voll. Establishment science disdained it!

Then why couldn't I see the RF on my RF oscilloscope? Probably because it was high frequency energy, not high energy frequency, and I didn't know how to amplify it above the background noise level. It was nevertheless not convincing. Yet much too tantalizing to ignore.

I thought of yet a fourth test. If I was really producing RF radiation that could be channeled through a circuit, I should be able to interfere with it by adding another RF radiation from an outside source. I added a frequency from my frequency generator, first at 1,000 Hz. Now there was no resonance. It interfered. Did this mean that my body was not producing radiation at 1,000 Hz? Or was my 1,000 Hz radiation being matched and canceled? I raised the frequency gradually, from 1,000 to 10,000 to 100,000 to 1,000,000 Hz. There was no resonance anywhere, and I couldn't draw any conclusions. It was 5 o'clock on Sunday afternoon. Quitting time. But one last look at my generator reminded me that it could reach 2,000,000 Hz and I was just at 1,000,000. One more quick experiment wouldn't take much time. I cranked it to 1,800,000 Hz. And now a resonance screamed out! Was I "hearing things?" No more interference. I did it over and over. Why was it resonating now and not before? Had I arrived at my body's own *bandwidth* (transmission range), and this was the reason it no longer interfered?

I found the lowest frequency that resonated to be 1,562,000 Hz. All frequencies that I checked (about 2,000) from there up to 2,000,000 (my frequency generator would go no higher) also resonated.

A year later I purchased a better frequency generator to search for the upper end of my bandwidth. Any frequency between 1,562,000 and 9,457,000 Hz could be added to the circuit and produce resonance.

It seemed obvious, then, that the human body broadcasts electrically, just like a radio station, but over a wide band of frequencies and very low voltages, which is why it has not been detected and measured until now.

Everything Has A Unique Frequency

It was a busy year, now 1989. I was determined to find a bandwidth for other living things: I found them for flies, beetles, spiders, fleas, ants. They were between 1,000,000 Hz and 1,500,000 Hz; cockroaches were highest amongst insects I tested.

Then came a dismaying finding. A dead insect had a bandwidth too! Much narrower, and near the top end of the same range it had when living, but distinctly present. So it wasn't altogether a living phenomenon.

But if dead things had a resonant bandwidth, then maybe a prepared microscope slide of a dead creature could be used, and my trips to the garden and telephone calls to abattoirs (for meat parasites) could cease. That was a lucky thought. My first slide was of the *human intestinal fluke*, a huge parasite, scourge of humanity. I had just found it to be present in the liver (not intestine) of every cancer sufferer I saw. The (dead) adult parasite had a resonant frequency around 434,000 Hz. Slides of that parasite's redia resonated nearby (432,000 Hz), as did its other stages.

Dead things still resonated! The entire catalog of biological supply companies, hundreds of specimens of viruses, bacteria, parasites, molds, and even toxins, were now available to research with this new technique!

Suddenly an idea bolted out of the blue. If a person were to hold on to the frequency generator while it was generating 434,000 Hz, what would happen to the adult fluke, if you were infected with it?

I tested this plan that same week on myself, not with the fluke but with *Salmonella* bacteria and *Giardia* and *Herpes* that I carried chronically. After a 3 minute treatment, I retested myself. I could no longer find them in my organs! There were no emissions at their characteristic frequencies. I repeated and repeated. Were they really dead? Maybe they were just numbed or were suddenly hiding. But symptoms were gone quickly too. My *Herpes* lesion stopped tingling. It was all too simple and unbelievable.

But was it safe? Within three weeks I had reliable data regarding the necessary level of electrical treatment. It only took 5 volts for three minutes at the specific frequency. It is not as if you had to use house current which would kill you, along with the parasite.

Selective Electrocution

In twenty minutes (three minutes at six different frequencies) a whole family could get rid of this parasite. Cancer cases showed that in a few hours the universal cancer marker, *ortho-phospho-tyrosine* could be banished from their bodies by killing this same parasite.

"Incurable" HIV cases lost their virus in a few hours, too. Laboratory retesting for HIV came back negative! Most cases of pain got immediate relief if I could identify the correct "bug" and have its frequency found by the next office visit. This seemed to be absolute proof that living things had an essential high frequency output of some kind of energy.

What was actually happening to the bacteria or parasites? If I could kill something as large as an *Ascaris* worm or intestinal fluke, then perhaps I could kill something even larger, like an earthworm or flea, something I could see with my own eyes instead of having to imagine its demise inside my body.

Ten minutes at a frequency chosen near the top of their broadcast range seemed to anesthetize them. But they didn't die. Later I checked the body bandwidth (the range of frequencies they emit) of each. The earthworms had lost a lot of their bandwidth, both at the top and bottom. The fleas seemed hardier; they had only lost a little. However they did not recover, even weeks later, from this loss.

Could it harm humans to douse them with RF frequencies **in their own bandwidth?** Quite probably, if the voltage were high enough. There was no need to experiment, though, because the parasites we want to kill have characteristic frequencies that **do not overlap** the characteristic frequencies of a human. In fact, they are far away (see the chart on page 17).

So my electronic method attacking illness was born. Find the resonant frequency of a bacterium, virus or parasite using a slide or dead bit. Treat the living invaders inside the human body with this frequency and in a matter of minutes they are no longer transmitting their own bandwidths—they are dead or sick and will be removed by our white blood cells.

It was a worrisome truth. Perhaps the department of defense would use this knowledge and develop super high voltage devices to kill people ("enemies") somewhere in the world. But I couldn't let sick people suffer. Besides, it would probably require a voltage much like lightning to kill people from a distance. Possibly a way could be found to shield yourself from frequencies harmful to humans by wearing a choke (inductor) coil which suppresses these frequencies. Remember, there was no recovery, just a slow death for my experimental animals. It must not happen to humans!

Meanwhile, people must be alerted that they can safely kill their invaders and heal their chronic illnesses. Invaders that have been increasing exponentially due to lowered immunity in recent decades. Possibly this is true for all species on our planet. The pollution of the entire biosphere has been increasing and with it the prospect of acquired immune deficiency syndrome (AIDS) for all of us.

Remember, though, that the true challenge is not to kill our invaders but to regain our health and immunity.

More than just parasites are making us sick! Pollution is too. Selective electrocution rarely makes people completely well. Sick people always have an environmental factor that must be corrected also.

How do we do that? The ship of “progress”, of increasingly complex, processed foods and products, must be turned around and simplicity become our goal. Survival is in simplicity of food intake, simplicity of life habits. Did Ralph Waldo Emerson foresee this when he said “To be simple is to be great”? Or will daily parasite and pathogen electrocution become another crutch that makes us just enough better that we can continue a detrimental lifestyle? Yet another “Band-Aid” treatment for our poisoned planet?

Bioradiation

Strange as it appears, it now seems obvious that every living creature broadcasts its presence like a radio station, the sun, or the stars. I have named it *bioradiation*.

Perhaps it is the same energy as the Asian *chi*; perhaps it is merely related to it. Perhaps it is the energy that runs along the meridians discovered eons ago by Asian practitioners.

Perhaps it is the energy that faith healers and religious teachers know how to harness, perhaps not.

Perhaps it is the energy that psychics perceive and that drives occult phenomena, perhaps not.

What is truly amazing is that ordinary persons have discovered such energy well ahead of scientists. Persons using the “art” of kinesiology, pendulums, radionics, dousing rods and many other forms of “strange energy” have no doubt harnessed a part of this bioradiation. It is a tribute to the generally high intelligence of common people and to their open-mindedness that they discovered this energy, **in spite of** opposition from scientists of today.

Over a century ago the scientists of Europe proposed the existence of a “life force” called “élan vitale.” They were scorned out of existence (and out of jobs). Young scientists, (including myself) were systematically taught to scorn this idea. Of course we were also taught that a good scientist was unemotional, does not scorn ideas, has a completely open mind, and does not rule something out until it is disproved to their satisfaction. The youthfulness of college years is so susceptible to prejudices of all kinds, and the desire for acceptance is so great, that special effort needs to be made to teach neutrality. Or at least to distinguish between emotion and fact. Where have these basic pedagogic principles gone? I was indeed inspired with the phrase “search for truth” but then promptly led down the path of “search for acceptance.”

I do not know what bioradiation, this electrical broadcast from our cells, is made of. Only its frequency was noticed and caught (modulated) in such a way as to be measurable. And this frequency, 1,520,000 to 9,460,000 Hz (for a human infant) is in the radio frequency (RF) range[2].

Anyone who is experienced with RF knows its strange behavior. Not strange in the "unknown" sense but in the amazing sense. Circuits don't need to be complete or closed for it to travel. Bodies and objects can "pick it up" without being in the circuit. These amazing properties are due to the capacitive and inductive properties of objects all around us, including ourselves.

Zapping Bugs

By *zapping* I mean selectively electrocuting pathogens. For years I used a commercial frequency generator to "zap" one pathogen after another.

First I made a chart of the frequencies for most of the bacteria and viruses in my collection (over 80, see page 561). Then I would test the sick client for each one of these, and hope they did not have one for which I didn't have a sample. Even persons with a simple cold typically had a dozen they tested positive to (not just *Adenovirus*).

Next it was time to tune in the frequency generator to a dozen frequencies for three minutes each. The total process, testing and treatment, would take about two hours. They frequently got immediate relief. But often the relief would be temporary. What I didn't know at that time was that viruses could infect a larger parasite such as a roundworm. Until you killed your roundworm and your virus, you would keep getting the virus back promptly.

---

[2] AM radio broadcasts are from 540,000 Hz to 1,600,000 Hz (slight overlap with lower end of human band), FM is 88,000,000 to 108,000,000 Hz (out of the human range).

In 1993 my son, Geoffrey, joined me and we tried a new approach. He programmed a computer controlled frequency generator to automatically cover all the frequencies populated by all the parasites, viruses, and bacteria, from 290,000 Hz to 470,000 Hz. It spent about three minutes for every 1000 Hz it covered. This was more efficient, but it meant spending ten hours being zapped.

Again, the results were disappointing. Arthritis pain, eye pain, colds were improved, but not completely cured overnight. Months later I would find that organisms were transmitting as low as 170,000, and as high as 690,000 Hz. My specimen collection was obviously incomplete. To cover this larger range, spending three minutes for every 1000 Hz, would take 26 hours. Still worth doing if it would indeed help all our illnesses. But even this method of zapping was not 100% effective for reasons yet to become clear.

In 1994 my son built a hand held, battery operated, accurate frequency generator. The purpose was to enable everyone to kill the intestinal fluke at 434,000 Hz with a low cost device. Enough benefit would be derived from zapping at various frequencies that I thought everyone should know how to make one. When I tested it on one of my own bacteria, however, **three others at much different frequencies died also!** This had never happened before. When I tested it on others, even though they had dozens of pathogens, all were killed!

Subsequent testing showed it was not due to some unique design, or special wave form produced by the device. It was due to battery operation!

Any positively offset frequency kills all bacteria, viruses and parasites simultaneously

given sufficient voltage (5 to 10 volts), duration (seven minutes), and frequency (anything from 10 Hz to 500,000 Hz).

Before this I had always set my commercial frequency generator to alternate between positive and negative voltage. Now I tried setting it to alternate between positive and zero voltage (*positive offset*). It was just as effective as the battery operated frequency generator my son designed.

Generating positive offset frequencies is the best way to kill all pathogens quickly.

But it takes more than one treatment.

It takes three treatments to kill everything. Why? The first zapping kills viruses, bacteria and parasites. But a few minutes later, bacteria and viruses (different ones) often recur. I conclude they had been infecting the parasites, and killing the parasites released them. The second zapping kills the released viruses and bacteria, but soon a few viruses appear again. They must have been infecting some of the last bacteria. After a third zapping I never find any viruses, bacteria or parasites, even hours later.

Why didn't the virus inside the parasite die with the first zapping? It may be because electricity travels on the exterior of things. The body of the parasite shielded the interior. This is why my earlier, promising work spending hours on a frequency generator gave only partial or temporary improvement—it was only done once, not three times. And it explains why a single treatment with a frequency generator or zapper frequently gives you a cold!

Zapping does not kill shielded organisms

such as those that may be in the middle of your stomach or intestines. The electricity travels along the stomach or intestine wall, not through their contents.

So zapping is still not perfect, but can bring such manifest relief that everyone should buy or make one. Parts cost less than $25.00; the plans are in the next chapter.

The Bioradiation Spectrum

Everything emits a characteristic range of frequencies (bandwidth). In general, the more primitive the organism, the lower its bandwidth. Advanced animals have higher frequencies and the range is wider.

BANDWIDTH OF BIORADIATION OF ANIMALS

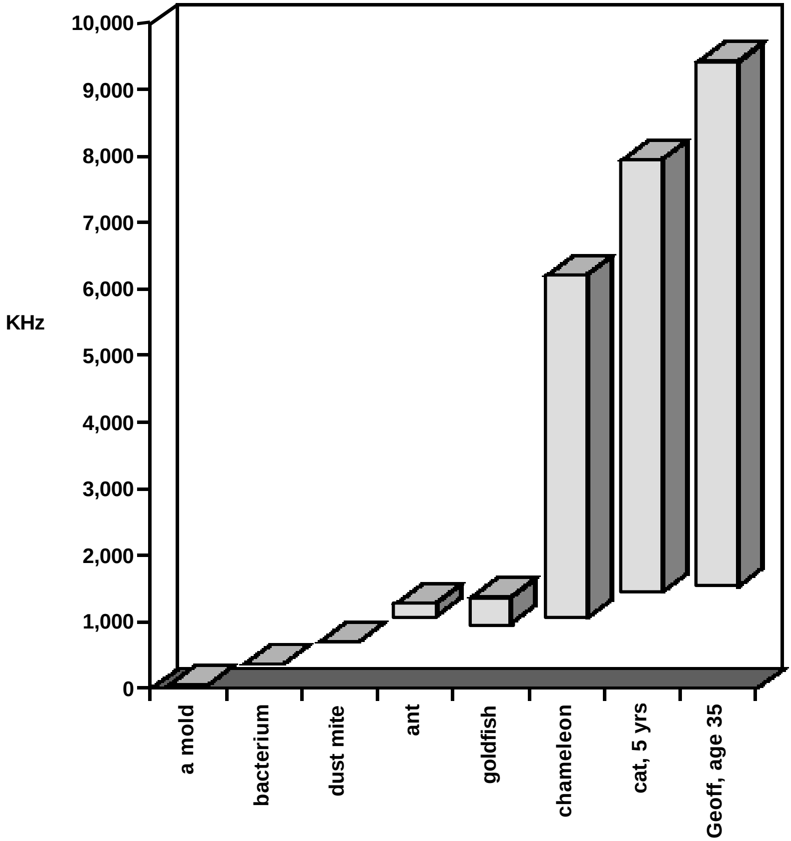

Fig. 1 Selected animal bandwidths.

The human range is from 1520 KHz to 9460 KHz. Pathogens (molds, viruses, bacteria, worms, mites) range from 77 KHz to 900 KHz. Fortunately for us we can work on zapping pathogens in the lower ranges without affecting humans in the upper range.

BIORADIATION OF TYPICAL PATHOGENS

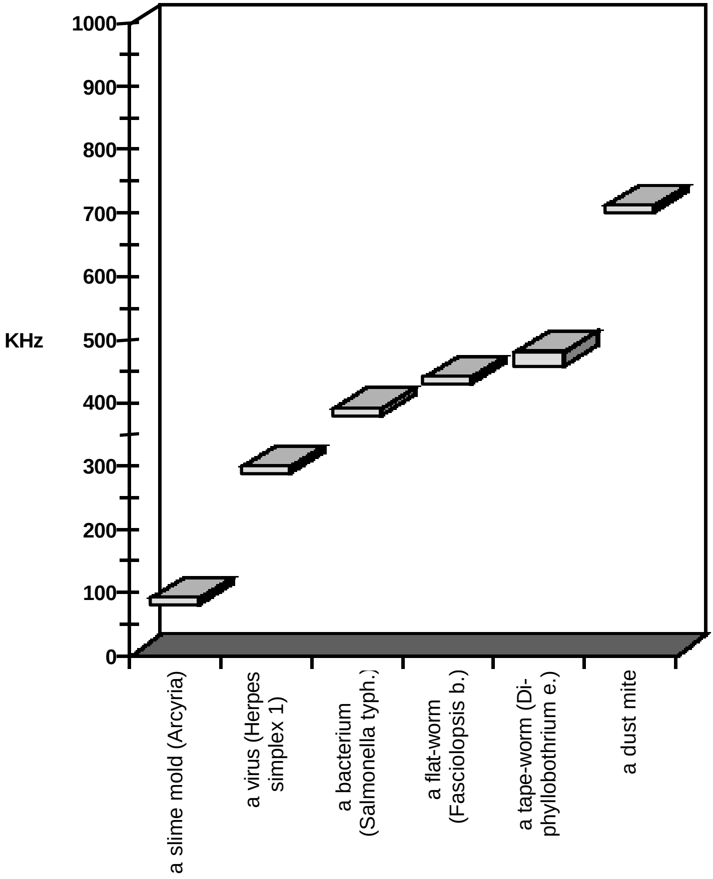

Fig. 2 Selected pathogen bandwidths.

Applying an alternating electrical voltage within an organism's bandwidth injures it. Small organisms with narrow bandwidths are extinguished quite readily (three minutes at five volts).

Positively offset frequencies can kill the entire range of small organisms (viruses, bacteria, parasites) in just seven minutes.

Building A Zapper

Being able to kill your bacteria and other invaders with electricity becomes much more of a panacea when you can do it all in three 7 minute sessions. No need to single out specific frequencies or to sweep through a range of frequencies one KHz at a time. No matter what frequency it is set at (within reason), it kills large and small invaders: flukes, roundworms, mites, bacteria, viruses and fungi. It kills them all at once, in 7 minutes, even at 5 volts.

How does it work? I suppose that a positive voltage applied anywhere on the body attracts negatively charged things such as bacteria. Perhaps the battery voltage tugs at them, pulling them out of their locations in the cell doorways (called *conductance channels*). But doorways can be negatively charged too. Does the voltage tug at them so they disgorge any bacteria stuck in them? How would the positive voltage act to kill a large parasite like a fluke? None of these questions can be answered yet.

Other fascinating possibilities are that the intermittent positive voltage interferes with electron flow in some key metabolic route, or straightens out the ATP molecule disallowing its breakdown. Such biological questions could be answered by studying the effects of positive frequencies on bacteria in a lab.

The most important question, of course, is whether there is a harmful effect on you. I have seen no effects on blood pressure, mental alertness, or body temperatures. It has never produced pain, although it has often stopped pain instantly. This does not prove its safety. Even knowing that the voltage comes from a small 9 volt battery does not prove safety, although it is reassuring. The clotting of red blood cells, platelet aggregation and functions that depend on surface charges on cells need to be investigated. But not before you can use it. Your safety lies in the short period of exposure that is necessary. Viruses and bacteria disappear in 3 minutes; tapeworm stages, flukes, roundworms in 5; and mites in 7. One need not go beyond this time, although no bad effects have been seen at any length of treatment.

The first seven minute zapping is followed by an intermission, lasting 20 to 30 minutes. During this time, bacteria and viruses are released from the dying parasites and start to invade you instead.

The second seven minute session is intended to kill these newly released viruses and bacteria. If you omit it, you could catch a cold, sore throat or something else immediately. Again, viruses are released from the dying bacteria. The third session kills the last viruses released.

Do Not Zap If You Are Pregnant
Or Wearing A Pacemaker.

These situations have not been explored yet. Don't do these experiments yourself. Children as young as 8 months have been zapped with no noticeable ill effects. For them, you should weigh the possible benefits against the unknown risks.

That is all there is to it. Almost all. The zapping current does not reach deep into the eyeball or testicle or bowel contents. It does not reach into your gallstones, or into your living cells where *Herpes* virus lies latent or *Candida* fungus extends its fingers. But by zapping 3 times a day for a week or more you can deplete these populations, too, often to zero.

Killing The Surviving Pathogens

The interior of gallstones may house parasites inaccessible to the zapping. Eliminate this source of reinfection by flushing them out with liver cleanses (page 552).

Although the center of the bowel contents is often unaffected by electric current, which lets bowel bacteria like *Shigella, Escherichia coli (E. coli)* and parasite stages survive, sometimes it is nearly all sterilized by zapping. This results in considerable shrinkage of the bowel movement. Eliminate remaining parasites and bacteria with a single dose (2 tsp.) of Black Walnut Hull Tincture, Extra Strength (see page 543).

There is no way of distinguishing between "good" and "bad" bacteria with either of these methods. However even good bacteria are bad if they come through the intestinal wall, so zapping targets mostly "bad" bacteria. The good news is that perfect bowel habits often result in a few days. Evidently, the good bacteria are benefited by killing the invasive ones. Home-made yogurt and buttermilk (see *Recipes*) are especially good at recolonizing the bowel. But it does not seem wise to culture yourself with special commercial preparations and risk getting parasite stages again when you can become normal so soon anyway. If you do decide to take some *acidophilus* bacteria to replenish your intestinal flora make sure you test for parasites like *Eurytrema* first.

When a large number of parasites, bacteria and viruses are killed, it can leave you fatigued. Try to give yourself a low-stress day after your initial zapping. But there are no significant side effects. I believe this is due to the second and third zapping which mops up bacteria and viruses that would otherwise be able to go on a feeding frenzy with so much dead prey available.

To build your zapper you may take this list of components to any electronics store (Radio Shack part numbers are given for convenience).

Zapper Parts List

Item	**Radio Shack Catalog Number**
large shoe box	
9 volt battery	

9 volt battery clips	270-325 (set of 5, you need 1)
On-Off toggle switch	275-624A micro mini toggle switch
1 KΩ resistor	271-1321 (set of 5, you need 2)
3.9 KΩ resistor	271-1123 (set of 2, you need 2)
low-current red LED	276-044 or 276-041 or 276-045
.0047 uF capacitor	272-130 (set of 2, you need 1)
.01 uF capacitor	272-1065 (set of 2, you need 1)
555 CMOS timer chip	276-1723 (set of 2, you need 1)
8 pin wire-wrapping socket for the chip	276-1988 (set of 2, you need 1) Note: Radio Shack is discontinuing all wire wrap sockets. Find another parts store or use 276-1995 (but the legs are much shorter and harder to attach clips to).
short (12”) alligator clip leads	any electronics shop, get 6
Microclip test jumpers	278-017 (you need 2 packages of 2)
2 bolts, about 1/8” diameter, 2” long, with 4 nuts and 4 washers	hardware store
2 copper pipes, ¾diameter, 4” long	hardware store
sharp knife, pin, long-nose pliers	

Hints for absolute novices: Don't let unusual vocabulary deter you. A “lead” is just a piece of wire used to make connections. When you remove a component from its package, label it with a piece of tape. A serrated kitchen knife works best as does a large safety pin. Practice using the microclips. If the metal ends are L-shaped bend them into a U with the long-nose pliers so they grab better. Chips and chip holders are very fragile. It is wise to purchase an extra of each in case you break the connections.

R1	1K
R2	3.9K
R3	1K
R4	3.9K
C1	.01ì f
C2	.0047ì f
U3	MC1455
LED1	2 ma LED Red
Pin 1	ground
pin 8	power

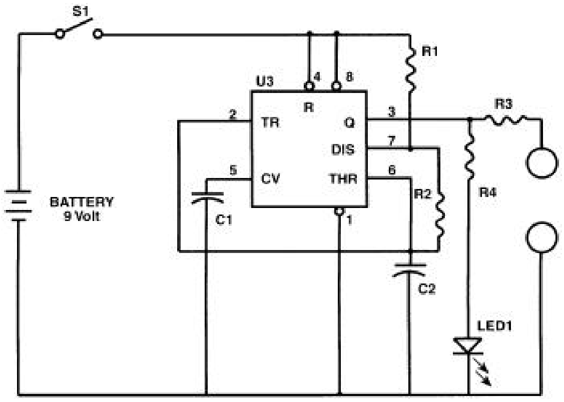

Give this to an electronics person or make it yourself in a shoebox by using the following instructions.

Fig. 3 Zapper schematic.

Assembling The Zapper

1. You will be using the lid of the shoe box to mount the components. Save the box to enclose the finished project.
2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

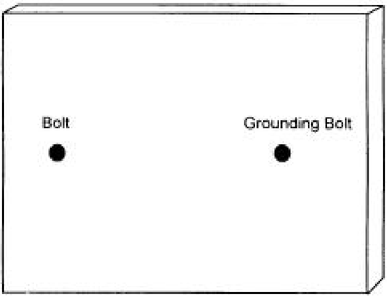

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.
4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1.

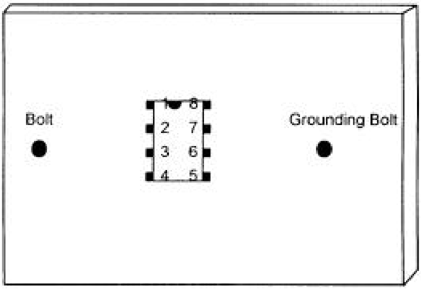

5. Pierce two holes ½ inch apart very near to pins 5, 6, 7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF ca-

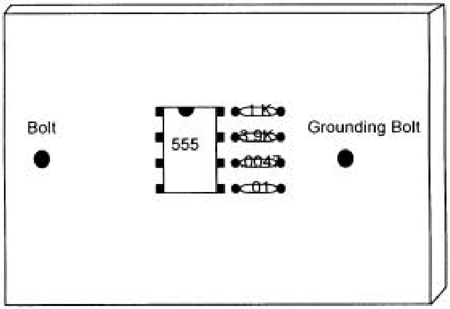

pacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 KΩ resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 KΩ resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 KΩ resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9KΩ resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.
7. Next to the 3.9KΩ resistor pierce two holes ¼ inch apart for the LED. No-

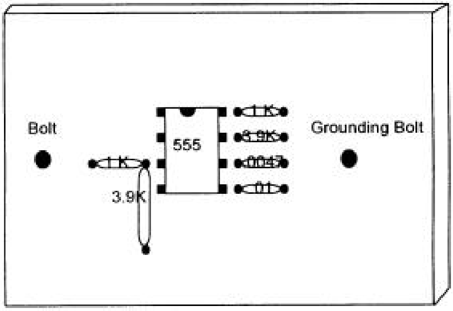

tice that the LED has a positive and negative connection. The longer wire is the anode (positive). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

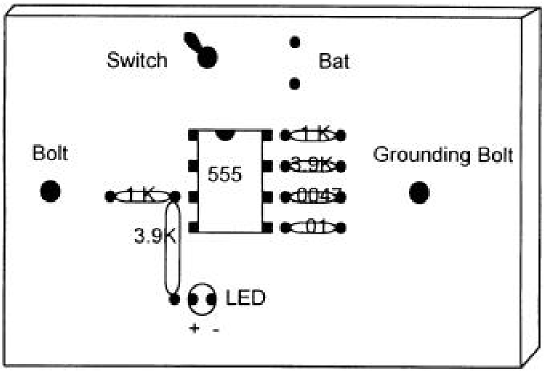

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.
9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Attach the battery and tape it to the outside.

NOW TO CONNECT EVERYTHING

First, make holes at the corners of the lid with a pencil. Slit each corner to the hole. They will accommodate extra loops of wire that you get from using the clip leads to make connections. After each connection gently tuck away the excess wire.

1. Twist the free ends of the two capacitors (.01 and .0047) together. Connect this to the grounding bolt using an alligator clip.
2. Bend the top ends of pin 2 and pin 6 (which already has a connection) inward towards each other in an L shape. Catch them both with an alligator clip and attach the other end of the alligator clip to the free end of the 3.9KΩ resistor by pin 7.

3. Using an alligator clip connect pin 7 to the free end of the 1KΩ resistor attached to pin 8.
4. Using two microclips connect pin 8 to one end of the switch, and pin 4 to the same end of the switch. (Put one hook inside the hole and the other hook around the whole connection. Check to make sure they are securely connected.)
5. Use an alligator clip to connect the free end of the other 1KΩ resistor (by pin 3) to the bolt.
6. Twist the free end of the 3.9KΩ resistor around the plus end of the LED. Connect the minus end of the LED to the grounding bolt using an alligator clip.
7. Connect pin number 1 on the chip to the grounding bolt with an alligator clip.

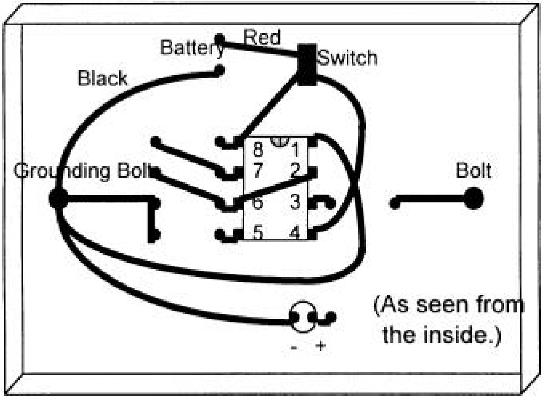

(As seen from the inside.)

8. Attach an alligator clip to the outside of one of the bolts. Attach the other end to a handhold (copper pipe). Do the same for the other bolt and handhold.
9. Connect the minus end of the battery (black wire) to the grounding bolt with an alligator clip.
10. Connect the plus end of the battery (red wire) to the free end of the switch using a microclip lead. If the LED lights up you know the switch is ON. If it does not, flip the switch and see if the LED lights. Label the switch clearly. If you cannot get the LED to light in either switch position, you must double-check all of your connections, and make sure you have a fresh battery.

11. Finally replace the lid on the box, loosely, and slip a couple of rubber bands around the box to keep it securely shut.

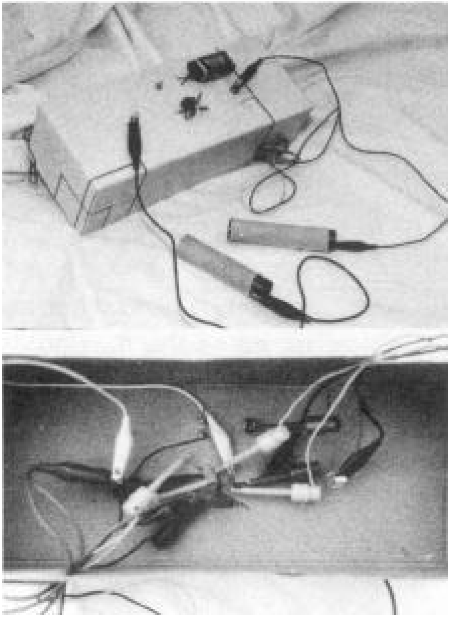

Fig. 4 Finished zapper, outside and inside.

- Optional: measure the frequency of your zapper by connecting an oscilloscope or frequency counter to the handholds. Any electronics shop can do this. It should read between 20 and 40 kHz.
- Optional: measure the voltage output by connecting it to an oscilloscope. It should be about 8 to 9 volts. **Note: a voltage meter will only read 4 to 5 volts.**
- Optional: measure the current that flows through you when you are getting zapped. You will need a 1 KΩ resistor and oscilloscope. Connect the grounding bolt on the zapper to one end of the resistor. Connect the other end of the resistor to a handhold. (Adding this resistor to the circuit decreases the current slightly, but not significantly.)

The other handhold is attached to the other bolt. Connect the scope ground wire to one end of the resistor. Connect the scope probe to the other end of the resistor. Turn the zapper ON and grasp the handholds. Read the voltage on the scope. It will read about 3.5 volts. Calculate current by dividing voltage by resistance. 3.5 volts divided by 1 KΩ is 3.5 ma (milliamperes).

Using The Zapper

1. Wrap handholds in one layer of wet paper towel before using. Grasp securely and turn the switch on to zap.
2. Zap for 7 minutes, let go of the handholds, turn off the zapper, and rest for 20 minutes. Then 7 minutes on, 20 minutes rest, and a final 7 minutes on.

Trying the zapper on an illness to see "if it works" is not useful. Your symptoms may be due to a non-parasite. Or you may reinfect within hours of zapping. The best way to test your device is to find a few invaders that you currently have (see Lesson Twelve, page 492, or Lesson Twenty Seven, page 509). This gives you a starting point. Then zap yourself. After the triple zapping, none of these invaders should be present.

Simple Pulser

If you are ill or want a reliable zapping, make the first model. However, there is another way to make a zapper if you can not afford to build the first model.

An ordinary battery is a source of positive voltage. It is the positive voltage that eliminates so many parasites at once, not a specific frequency. So although the zapper's frequency is about 30 kHz (thirty thousand "zaps" per second), even 5 Hz (five “zaps” per second), **about as fast as you can tap the battery with your hand, is moderately effective!**

You must be connected to both terminals. One will be marked + (positive) and the other – (negative). If you simply touch these terminals with your wet fingers, nothing much happens. That is because your resistance to the current starts going up right away, so less and less current passes through you.

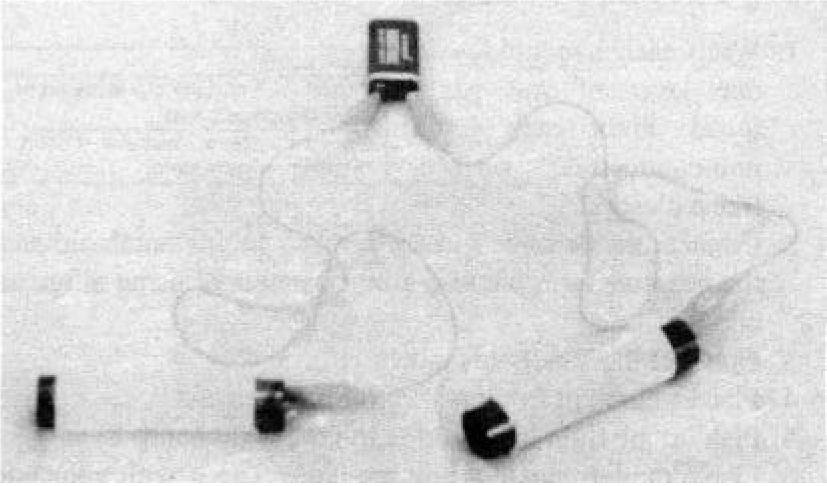

Fig. 5 Simple pulser.

However if you tap the positive terminal with your wet hand, and tap it at a fairly high rate, your body's capacitors come into play. Capacitors only take part in the flow of electricity when they are charged and discharged. Tapping the terminal starts and stops the voltage so capacitors charge and discharge. This kind of resistance to current flow is much smaller.

The faster you tap, the greater the frequency of current pulses and the lower this kind of resistance becomes. Now you can have a considerable sustained current flow through your body.

If you can tap even twice per second (2 Hz) for ten minutes without interruption you can give yourself a zapping that is moderately effective. Remember to take an intermission of twenty minutes and then repeat to avoid catching new viruses. After a second twenty minute intermission repeat zapping a third time.

Using The Simple Pulser

9 volt battery
2 short (12”) alligator clip leads (from any electronics shop)
2 copper pipes, ¾” diameter, 4” long (from a hardware store)

1. Wrap each handhold with one layer of wet paper towel. Place each on a non-conductive surface, like a plastic bag.
2. Connect the positive battery terminal to one handhold and the negative terminal to the other handhold using alligator clip leads.
3. Don't let the handholds touch.
4. Place a clock in front of you to time yourself.
5. Pick up the right handhold with your right hand.
6. Leave the left handhold on the table. Tap it with your left hand, preferably the fleshy part of the palm. You may brace yourself with your fingers on the plastic. Keep up a steady pace as fast as you are able.
7. When you get tired pick up the left handhold with your left hand and tap with your right hand. Keep changing off with the least interruption.
8. Repeat a second time 20 minutes later, and a third time 20 minutes after that.

A single 9 volt battery will wear out rather quickly used this way. Put two together, in parallel, for longer lasting power. This requires two more short alligator clips. Connect positive terminals of the batteries to each other, and the negatives also.

Parasites & Pollution

The word "parasites" is used in two senses. Everything living on you or in you, not just to perch, but to take its food from you is a *parasite*. No matter what its size, it can be called a parasite.

But in some way the big worms need to be distinguished from the medium-sized amoebae, the even smaller bacteria and the smallest of all—viruses. So often the term parasite is reserved for the bigger things, from amoebae on up. In this book, the word parasite will be used in both ways as usual. You can easily guess what is meant.

Parasitic worms are divided into *roundworms* and *flatworms*. Roundworms are round like earthworms even though they may be as thin as hairs (threadworms, *filaria*) or microscopically small (like *Trichinella*). Flatworms are more like leeches. They have a way to attach themselves sometimes with the head (*scolex*) like *tapeworms*, sometimes with a special sucker like *flukes*.

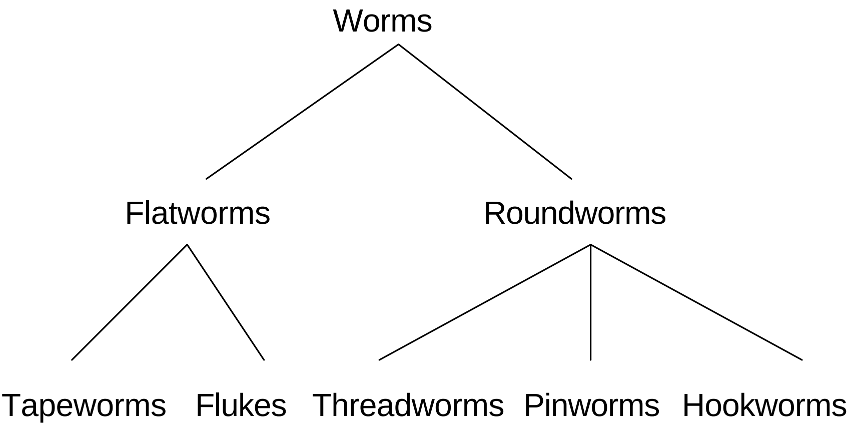

Worm parasites go through stages of development that can look very, very different from the adult.

Roundworms like *Ascaris* (common cat and dog roundworm), are simplest. The **eggs** are swallowed by licking or eating a bit of filth. They hatch into a tiny *larva*. The larva treks to the lungs. You cough it up and swallow it. Meanwhile it has molted a few times. It then crawls to the intestine where it becomes an adult, shedding eggs in your stool.

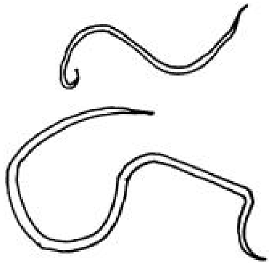

Fig. 6 Ascaris.

Worms usually have preferred locations. The favorite organ for *Dirofilaria* (dog heartworm) is the heart (even human heart). Sometimes the rules can be broken. My tests show *Dirofilaria* can live in other organs, too, if they are sufficiently polluted with solvents, metals and other toxins.

Flatworms like **tapeworms** are much more complicated in their life history. You could eat the eggs accidentally with dirt. After hatching, the tiny larva burrows into its favorite organ. Your body encases it with a cyst. The white blood cells have been taught never to attack your body...and the cyst case is your body! So the tapeworm stage has safe residence for some time. If you are a meat eater, you could eat such a cyst if it happens to be lodged in the meat you are eating! Your teeth break it apart as you crunch. The little larva is swallowed and tries to attach itself to your intestine with its head. Then it grows longer by making segment after segment. The segments with their eggs leave with the bowel contents. I often see dog tapeworm of the small variety in their human family.

Flatworms like **flukes** are also very complicated. The eggs, passed out with bowel contents were not meant to be eaten as such. They were meant to hatch in a pond where snails and minnows eat them. The larva grows up in these new "secondary" hosts. Later, the snail sheds them and they attach themselves to foliage near the pond. They over-winter in a tough *metacercarial* cyst. An unsuspecting browsing animal now eats them. They come out of their metacercarial cyst as a small adult and quickly attach themselves to the intestine with a sucker. They now have "safe haven" and can go about maturing and laying eggs.

Four common flukes are: **human intestinal fluke, human liver fluke, sheep liver fluke, pancreatic fluke of cattle.** Don't let the terms sheep and cattle mislead you. They are all found in humans.

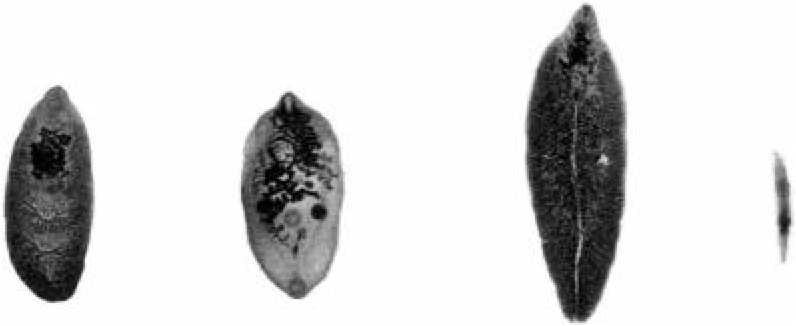

Fig. 7 (L to R) Human intestinal fluke, pancreatic fluke, sheep liver fluke, and human liver fluke.

The Worst Parasite

Fasciolopsis buskii is the fluke (flatworm) that I find in every case of cancer, HIV infection, Alzheimer's, Crohn's disease, Kaposi's, endometriosis, and in many people without these diseases. Its life cycle involves six different stages:

Stage	**Normal Life Cycle**	
1 Egg	Expelled with bowel movement onto soil. Washed by rain into ponds.	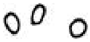
2 Miracidia	Hatches from egg in water. Has cilia, can swim vigorously and must find intermediate snail host in one to two hours or may be too exhausted to invade.	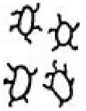
3 Redia	Develop inside miracidia as little balls until expelled. Those are "mother" redia, and each one bears "daughter" redia for up to 8 months, all still inside the snail, and living on the fluids in the lymphatic spaces. Similarly, daughter redia are continually developing cercaria.	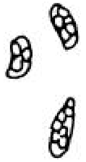
4 Cercaria	Have a tail, use it to exit from snail and swim to a plant. If the snail is feeding on a plant, cercaria can latch onto plant with sucker mouth and start to encyst (form a "cocoon") within minutes. Tail breaks off and swims away to dissolve.	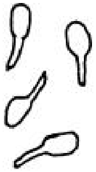
5 Metacercaria	Two-walled cyst. The outer wall is very sticky. But as you eat the plant it is stuck to, the least pressure will break it, leaving the cyst in the mouth. The "almost unbreakable" inner cyst wall protects it from chewing, and the keratin-like coat prevents digestion by stomach juices. However when it reaches the duodenum, contact with intestinal juices dissolves away the cyst-wall and frees it. It then fastens itself to the intestinal lining and begins to develop into an adult.	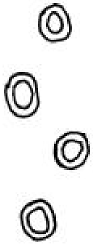

6 Adult	Lives in your intestine and can produce 1000 eggs per bowel movement and live many years.	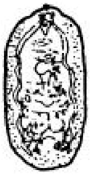

Fig. 8 Fasciolopsis' normal life cycle.

Note that the adult is the only stage that "normally" lives in the human (and then only in the intestine). *Fasciolopsis* depends on a snail, called a *secondary host*, for part of its life cycle. **But when your body has solvents in it, the other five stages can develop in you!**

If **propyl alcohol** is the solvent, the intestinal fluke is invited to use another organ as a secondary host—this organ will become cancerous. If **benzene** is the solvent, the intestinal fluke uses the thymus for its secondary host, setting the stage for AIDS. **Wood alcohol** invites pancreatic flukes to use the pancreas as a secondary host. This leads to pancreatic dysfunction which we call *diabetes*. If **xylene** (or **toluene**) are the solvents, I typically see any of four flukes using the brain as a secondary host. If **methyl ethyl ketone** (MEK) or **methyl butyl ketone** (MBK) are the solvents, the uterus becomes a secondary host and endometriosis a likely result.

This is a new kind of parasitism, based on pollution. I call the diseases caused by fluke stages in inappropriate locations **Fluke Disease**; it is discussed in more detail later (page 249).

Are tapeworms and roundworms affected by solvents this way, too? This is a fascinating and very important question. Search for the answer and help others search for the answer. I do not know yet.

Pollution

Pollutants are all the dead things around us that should not get into your body because they interfere with its work. As long as they don't penetrate your tissues, they won't interfere, like plastic eyeglasses and clothing. But if they are invasive, your body must fight to remove them.

Pollutants can invade your body via the air you breath, the foods and beverages you eat, and the products you put on your skin.

The biggest tragedy is not recognizing when a pollutant is harming you.

Two people can use the same face cream. One develops a rash, the other does not. The one who did not assumes the cream is not harmful to them…that they are like a bank vault, impregnable to that product. A better assumption is that the face cream is somewhat toxic, as evidenced by the rash that can develop, and they escaped the rash only because they had a stronger immune system. The immune system is like money, paid out of the bank vault, for every toxic invasion. When the money is gone, the bank (your health) fails.

Solvent Pollution

Solvents are compounds that dissolve things. Water is a useful, life giving solvent. Most other solvents dissolve fats and are life threatening, because fats form the membrane wall around each of our cells, especially our nerve cells.

The solvent that does the most harm is **benzene**. It goes to the thymus, ruins our immune system, and causes AIDS. The next worst solvent is **propyl alcohol**. It goes to the liver and causes cancer in some distant organ. Other major culprits of disease are **xylene, toluene, wood alcohol, methylene chloride,** and **trichloroethane** (TCE). I'll discuss each one later, with the ailment it's associated with.

Metal Pollution

Biochemists know that a mineral in raw element form always inhibits the enzyme using that mineral. Copper from the meat and vegetables you eat is essential. Inorganic copper, like you would get from a copper bottomed kettle or copper plumbing, is carcinogenic[3]. Unfortunately, the inorganic form of metals is what pervades our environment. We put metal jewelry on our skin, eat bread baked in metal pans, and drink water from metal plumbing.

Another obvious metallic threat is tooth fillings. **Mercury amalgam fillings**, despite the assurances of the American Dental Association, are not safe. And sometimes the mercury is polluted with **thallium**, even more toxic than mercury! **Gold** and **silver** seem to have fewer harmful effects, but no one should have any pure metal in or on their body.

Other prevalent toxic metals include **lead** and **cadmium** from soldered and galvanized plumbing, **nickel** and **chromium** from dentalware and cosmetics, and **aluminum** from food and drink cans, and cooking pots.

---

[3] Haleem J. Issaq, The Role of Metals in Tumor Development and inhibition. From *Carcinogenicity and Metal Ions*, volume 10, page 61, of a series called *Metal Ions in Biological Systems*, edited by Helmut Sigel, 1980.

Mycotoxins

Molds produce some of the most toxic substances known, called *mycotoxins*. One small moldy fruit or vegetable can pollute a huge batch of juice, jam or other product. Although molds are alive, and can be killed by zapping, mycotoxins are not, and must be detoxified by your liver.

But because mycotoxins are so extremely poisonous, a tiny amount can incapacitate a part of the liver for days!

Aflatoxin is the most common mycotoxin I detect. It is produced by molds that grow on quite a variety of plants. For that reason I am always cautioning people to eat only perfect citrus fruit, and never drink commercial fruit juice. Of the thousands of oranges that go into the batch of orange juice you drink, one is sure to be moldy, and that is all it takes to give your liver a setback.

A heavy dose of vitamin C helps the liver recover quickly. It also helps get rid of aflatoxin before it is consumed, right in the food container. So keep a plastic shaker of vitamin C powder handy and use it like salt on all your food.

There are thirteen other mycotoxins I have searched for in our foods. They are described in the section on moldy food (page 381).

Physical Toxins

Breathing in dust is quite bad for you so your body rejects it by sneezing, coughing, spitting up and out. Imagine breathing in broken glass particles. They cut into the lungs in a thousand places and couldn't be coughed up. They would travel. Imagine swallowing a needle or open pin. If the tip was blunt it could move through the intestine. But because it is sharp it gets caught in your tissue, then works its way deeper and deeper.

Would we ever knowingly breathe in broken glass? We are justifiably afraid of it in our food or under our bare feet. We are unaware that it fills our homes when **fiberglass** insulation is left imperfectly sealed off. Any hole made through the ceiling or wall, even if covered with cloth, lets swarms of broken glass bits into the house air. Air currents flow inward, into your living space. So all holes leading to the attic or insulated spaces must be sealed airtight. Of course, fiberglass should never be used in home construction, draperies, or around water heaters. The best advice is to have it all removed while you are away and then vacuum and dust.

Occasional exposures by house builders working outdoors does much less harm. Chronic exposure from a single small hole in the ceiling does a lot of harm, leading to cyst formation. And that cyst is a perfect place for parasites and bacteria to settle and multiply. When the intestinal fluke settles there it becomes malignant!

Cancer patients with *solid tumors* have either fiberglass or asbestos in them.

Asbestos is another tiny bit, sharp as glass, that moves through your body like a swordfish, impaling your cells until it, too, gets routed into a cyst.

We have been led to believe that we no longer have asbestos in our homes because we have outlawed the fireproofing materials it was used in. While that may be true, the source I find most often is all too prevalent: the clothes dryer belt. As it gets hot the belt releases a blast of asbestos particles that are forced through the seams of your dryer, and also openings in your exhaust hose, by the high pressure formed inside. It is now in your air.

Chemical Toxins

Chlorofluorocarbons (**CFCs**) or **freon** is the refrigerant in your air conditioner and refrigerator coils. CFCs are suspected of causing the ozone hole above the South Pole. All cancer sufferers test positive for CFCs in their cancerous organ! I have preliminary evidence that it is CFCs that attract other pollutants–fiberglass, metals, PCBs–to form a growing tumor instead of allowing their excretion. This would make it a “super carcinogen.”. How could you detect CFCs leaking in your home? By the time your air conditioner or refrigerator needs recharging, you have been exposed for a long time. We desperately need an inexpensive, in-home test for this unsuspected killer.

Arsenic is used in pesticide. Why would we poison ourselves along with the cockroaches? Is it because we can't see it happening? Just as we couldn't see the fiberglass floating in the air? Our diligent scientists have studied the mechanism of arsenic poisoning in great detail. Then why are we allowed to put it on our lawns to be carried into our carpets via shoes?

Polychlorinated biphenyls (PCBs), oily compounds with wonderfully useful electrical properties, were originally used in transformers until their inability to break down into less toxic substances in our environment was spotlighted. Banned from use, I find them in most commercial soap and detergents! Is transformer oil being disposed of by selling it to soap makers?

Formaldehyde is used to cure foam. As a result, foam furniture, pillows and mattresses give off formaldehyde for about two years after manufacturing. If you sleep with your nose buried in a new foam pillow all night, you are risking major lung problems.

Every cleanser in your house probably has a toxic warning on its label. Every fluid your automobile uses is toxic. Every pesticide, herbicide and fertilizer you put on your lawn is probably toxic. Every paint, varnish, wax, lubricant, bleach and detergent will send you to the hospital if even a small amount is ingested. Why do we keep them around? See *Recipes* (page 513) for safe, old-fashioned, alternatives.

If you are ill even after zapping, it is toxins still at work. Getting rid of them is a major step toward being well.

I DON'T THINK WE SHOULD LET HUMANS INTO THE HOUSE. WE'LL GET ALL THEIR PARASITES.
YEAH, I JUST CAN'T GET RID OF THIS MANGE.

How We Really Get Sick

What if you invented a device that could search people for the presence of mycotoxins (extremely toxic substances made by food molds)? And what if you found that although many people had them, those who were sick with a cold always had at least one of them. Would you ask whether a sudden buildup of mycotoxins is what really lets colds develop? Why do some people in the same family get the cold while others do not?

- What if you found everyone with **cancer** had the human intestinal fluke in their liver, and no one else did?
- What if you found everyone with **diabetes** had the pancreatic fluke of cattle in their pancreas, and few others did?
- What if you found everyone with **environmental illness** tested positive for Fasciola (sheep liver fluke) in their liver?
- What if you found everyone with **asthma** tested positive for Ascaris in their lungs?

What if you always found every mysteriously ill person had some unsuspected parasite or pollutant?

The device is the Syncrometer™, and these "what ifs" are all true. They forced me to alter my entire outlook on what really causes some of our "incurable", mysterious diseases.

We used to believe that diabetes was caused by over consumption of sugars, a cold by a virus you caught from somebody, cancer from carcinogen exposure, depression from poor parenting. This multicausal concept is what made the study of medicine so difficult that only a few could undertake it. And every year new syndromes are added to the list of human illnesses.

But these diagnoses are based on a description of what is happening at a particular place in your body. This is like calling a mosquito bite behind the ears by one name and a mosquito bite behind the knee by another name. If you never see the true cause, a mosquito at work, this system could be excused as somewhat sensible.

And, until now, the profession of medicine has made some sense. The new truths, however, make the old descriptive system obsolete. You can now find the true causes of all your illnesses. And you can find them yourself by building the electronic diagnostic circuit (page 457)!

Once you have seen a mosquito at work on your body you no longer need to go to the doctor for a red, itchy bump. You don't need to search for the correct diagnosis and an appropriate drug. **You put up screen doors and windows!**

Once you have seen how common house dust is implicated in the common cold you get rid of the house dust. Once you have seen the mold in your food facilitate the cold virus you throw out that moldy food. But only seeing is believing. Nothing is left to faith. The electronic resonance method described in this book will let you see all these things for yourself.

You are not a hapless pawn attacked by bacteria and viruses that dart at you from nowhere to make you ill. You are not at the mercy of diseases all around you, hoping, by chance, to escape, like a soldier hoping to come home from the war. Nature and your body make good sense.

There is no disease that can outwit you if you know enough about it. Not even Lou Gehrig's disease! Nor asthma or diabetes. Read how the people in the case histories made themselves well. Read why some people failed. You have an advantage they did not have. Their instructions were hard to carry out because they had to have faith in them. You don't. You can replace faith with your own hard headed observations by building the diagnostic circuit (Syncrometer). The great convincer is seeing it yourself. When you personally find the mold in your peanut butter, or *Shigella* in your cheese, you have the knowledge, not faith, that convinces and guides you.

All illness comes from two causes, PARASITES and POLLUTANTS.

Only two causes! This is what simplifies the picture to make it possible for you to cure yourself.

We have been taught that illness is largely our own fault. That it is due to “catching something”, not eating what we should, like roughage or vitamins, or not doing what we should, like dressing properly, exercising or going to bed on time. Somehow, it's our own fault. Either by doing something we shouldn't or not doing something we should. When absolutely no fault can be found we are told it's in our genes! True inherited diseases are extremely rare. Our genes have evolved over millions of years to produce healthy humans. Nor are genes that mutate during your lifetime at fault. Pollutants, which are known to be mutagens, are the real culprits.

Neither the parasites nor the pollutants in you are “your fault”. Notice that other people all around us are doing the same things or not doing the same things, and even share our genes and don't have our illness. The current concepts on disease causation blaming our actions and our genes are simply not logical.

But until now they appeared logical. Suppose 1000 people were bitten by a mosquito or flea, it would always be in a different place, and if you were using the location and effect system to diagnose the problem you could have a thousand diagnoses listed for them, including a defective gene. Could you be persuaded to accept a gene replacement for your red itchy bumps? The new gene might be for antihistamine production so the welts don't become large, red and hot, or lead to impetigo. If you were the research doctor, you might be tempted to alleviate a thousand persons' distress with a new gene. Fortunately, you are not and only need to solve your own problems. You can be more logical.

After you have found the parasite interlopers hiding in your body you can kill them electronically. And after you have identified the pollutants stuck in your organs you can stop eating them, breathing them or putting them on yourself. In response, your body will begin to heal, just as surely as a mosquito bite heals.

Heal from multiple sclerosis, emphysema, myasthenia? Yes!! Some healing will be swift. Some healing will be slow. Healing is not understood. It is much faster in young persons. But fast or slow you know it has begun. It will be an exciting adventure to watch yourself lose your symptoms and get stronger.

Self Health

The entire purpose of this book is to enable you to diagnose and treat yourself for any disease. You have three new approaches that make this wish a reality: the understanding that only pollution and parasites make you sick, the quick and inexpensive diagnostic circuit that lets you find which pollutants and parasites they are, and the zapper or herbal recipe that kills the parasites.

Wouldn't it be nice not to have to go to the doctor for your aches and pains? And not to be dependent on the doctor to diagnose and treat you?

Self health means keeping yourself healthy. Doing it **yourself**.

Suppose your doctor has already diagnosed you as having "Atypical Lateral Sclerosis" or "Shoulder-Hip Girdle Metastable Aplasia." Could you re-diagnose this for yourself so as to treat and cure it? And be successful? Why not? You have already succeeded in many enterprises. You learned to talk, walk, read, get along with people. These skills took a few years to master. Learning to keep well is a new skill. It may take a few years too. After you have learned these skills you may pass them on to your children. And so a new gift is given to humanity, like the gift of music or the art of cooking.

How To Heal

Your body has been trying to rid itself of its parasites and pollutants all your life! It had its own ways. It made stones, it made mucus secretions, it made itself toxic dumpsites. These were good tactics but now of course, they are no longer necessary. Can you help your body get rid of these accumulations and sweep itself clean again?

Sweeping your liver clean is the most powerful way of helping your body to heal itself after the parasites are gone. There are thousands of bits of "trash" accumulated in the liver bile ducts. They will turn into stones (gallstones) if left in place.

The kidney, too, has made numerous small stones in its effort to keep your body clear of lead, cadmium, mercury and other impassable pollutants. You can assist the kidney to expel all these.

In days, not weeks or months, you can feel the healing effects of clearing gallstones and kidney stones from your body. But there are miles of bile ducts (50,000 ducts) in the liver; the herbal recipes that do this are used over and over, patiently, until all, the “trash” is removed. This can take several years.

So, although you can stop your disease very quickly from progressing, the healing process may not be complete for years.

Nevertheless, you are healthy again. This means your pains are either gone or greatly reduced. Your organs are functioning better. You have a new sense of well being. Your energy is up. Your desire to live and accomplish something is back.

Organs that have been damaged beyond the ability of our simple methods to reverse can be treated with the magic of modern surgery. Cataracts, bunions, old injuries are examples. Possibly, these too, are “stone” formations. But no recipe has been found to clear them up simply and in a useful time frame.

Killing parasites, removing pollutants and clearing gallstones and kidney stones from your body is a powerful combination of treatments. It is so powerful you can change yourself into a new person in half a year. And then go on improving for years more.

Should you stop taking your prescription medicine while you are treating yourself? NO. Wait until you have cured yourself of the condition that required the medicine. Reduce your medicine and eventually go off it. Will your doctor approve? Find one that will. Remember that the medicine is buying you the time to cure yourself, something to be grateful for.

The Road To Wellness

To review our new understanding of health vs. disease:

- We have only two problems: parasites and pollutants. Parasites are things that live on us, using up our food and giving us their wastes. Pollutants are toxic things in us making it difficult for our organs to do their work. These two things are responsible for all our other problems.
- Our bodies have been trying to rid us of these by making stones, making secretions, giving us swellings, inflammations and benign tumors. We develop deficiencies and disabilities.
- Finally, some permanent damage is done. Our hair turns gray, we develop cataracts, the spine bends, nerves and muscles die. We weaken.

Our strategy to undo all this will be a logical one.

First, we will kill all parasites, bacteria, viruses and fungi.

Second, we will remove the toxic molds, metals and chemicals in our foods and body products.

Third, we will clear away and wash away the stones, secretions and debris already formed, that hinder healing.

Fourth, we will use herbs and special food factors to hasten healing, being very careful to use pure products. (These act more quickly when given intravenously but the emphasis in this book will be on oral consumption.)

Finally, for repairs that are beyond our abilities, we will seek help from health and medical professionals.

What could be more exciting than finding the tremor is out of your arm or the pain is out of your shoulder? Won't it be admirable to correct your pulse and your high blood pressure, by yourself? What an adventure it could be to get rid of all your warts.

But getting rid of pain seems like a first priority to me, since pain often undermines our morale, our initiative to do things, even our interest in getting well!

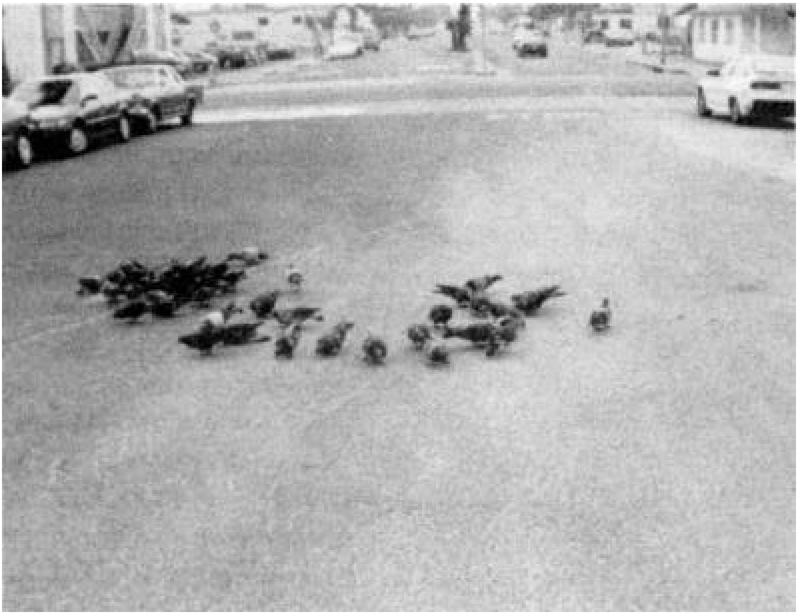

Many of our illnesses are caused by unsuspected sources. I have found pigeon tapeworm in humans, for instance.

Fig. 9 Feed animals out of your yard and away from where children play. Their wastes contain pathogens.

Pain From Toe To Head

I would rather die than endure excruciating, unrelievable pain. That puts **pain** at the very top of my priority list.

Fortunately for us, pain killers are at hand to get us through it and buy us the time it takes to solve the real problem behind it.

The pain killing industry also brought us the addiction industry. As we turn to electrical pain killing the need for addicting drugs should decline. There are other very useful pain killers: acupuncture, massage, listening to music, feedback devices, contemplation, hypnotism, and prayer.

But we will focus on getting rid of the cause of pain and healing the organs that are in pain so none of these methods are needed.

I am not talking about the pain of a broken bone, twisted ankle, bee sting or sunburn. I am not talking about the pain of a misaligned vertebra or stretch trauma in your leg muscles or arm muscles. I am referring to pain that is in one of your organs and refuses to go away. For example, arthritis. Other pains are headache, foot pain, elbow pain, hip pain, chest pain. All of these may have special names like rheumatoid arthritis, cluster headache, fibromyalgia, bursitis, tennis elbow and so on, but they are all the same phenomenon.

Knowing that parasites and pollutants are the real culprits, let us get right down to the job of finding out which they are, where they come from, and how to get rid of them.

The parasites that cause pain are not the large ones, like worms or amoebas. Nor are they the very tiny viruses. They are *bacteria*. Bacteria are the right size to get into the doorways of our cells. Our cells try to keep their doorways tight-shut but, of course, they have to open to let food in, or hormones, or other life-signals. If bacteria are swarming around the outside of cells, some will manage to slip into a doorway while it is open. A battle begins. The cells refuse to let the bacteria in. The bacteria refuse to let go of the door latch. There is probably a specific electrical attraction between them and an exact physical fit. The body's extra forces, the immune system, are called in to help quell the invasion. The bacteria multiply as fast as they can. There is swelling. There is heat produced. There is pressure against other organs. All together it is called inflammation and infection.

The answer is not to deinflame with drugs (like the cortisone variety). The answer is not to reduce swelling or body temperature. **The answer is certainly not pain killer.** The answer is to kill the bacteria. (Even this is not the ultimate answer. We must stop the source of the bacteria and your body's invitation to be invaded. We will get to this later.)

So **Step One** is to search for the bacteria nibbling at our painful regions and identify them. This gives us the clue to finding their source. **Step Two** is to exterminate them electrically. Within minutes they fall out of the doorways to our cells. Your white blood cells are waiting for them, and will gobble them up in a grand feast.

But, remember, there is another cause of pain, pollutants. **Step Three** is to find the pollutants and identify them because this gives us a clue as to their source. **Step Four** is to eliminate our pollution sources. **And the job is done. Pain is gone.**

An intriguing question will pop into your head as you search your organs for parasites and pollutants. Which came first? Pollutants can jam their way into your cell's doorways too. Does this allow the bacteria to swim in? Or do the bacteria come first, jamming open the doorways so the pollutants can enter? Both seem possible. Maybe both events go on simultaneously. Perhaps that is why bacteria and pollutants are always seen together. Viruses can land on your cell's doorways, too, and cause viral diseases, but they are not as often pain producing.

You will also notice something, as you keep testing and watching over your health. **Your body is very good at killing bacteria and viruses**. Your body kills them faster than a wave of a magic wand! The only ones that get away are those that are stuck in doorways and ‘channels with pollutants in them! This seems like evidence that pollutants do the gate-opening. But it isn't proof and we must keep all possibilities in mind.

Fortunately we do not have to know exactly how parasites and pollution make us sick in order to get well.

Searching For Bacteria

In order to find which organs have the bacteria and which bacteria are present you will need to learn the new technology that makes all of this possible. This technology is a simple electronic circuit that is capable of trapping frequencies in such a way that you can hear them. Your body's frequencies, the frequencies of bacteria, viruses and parasites are all different and can be heard as distinctly as a **mooo**, **baah**, **tweet**, or **oink** coming from a farm yard.

But do you have to do all that?

No! You don't. You could simply electrocute all these tiny invaders. But how would you know what to avoid in the future? If your pain returned how would you know if it was the same old bacteria or a new one?

Learning to test takes the guesswork out of diagnosis.

My hope is that you will find it all so intriguing, so absolutely fascinating, that in spite of some chronic pain, you will feel compelled to make the searches yourself. Find someone willing to help. Trade your information. It is less difficult than learning to use a computer.

What You Will Find

First we will study and cure pains of all kinds, starting with the toes and working our way up the body. We don't need to be very specific about the location of the pain since bacteria and toxins flow over to nearby locations anyway.

For each kind of pain, we will look into the causes so you can eliminate them. Pain could come back in an hour if you didn't know the bacteria were coming from cheese and you had a cheese sandwich after *zapping* (killing parasites electrically). In fact, the pain may not have had time to go away before the next onslaught begins and you might conclude, wrongly, that this method doesn't work.

Parasites might also come back to your specially painful place from a few far away places that are hard to reach by your zapper's electrical current. The inside of your eyeball, the testicle, the interior of gallstones, the middle of a tooth abscess or the bowel contents are such places. Your zapper current, because it is high frequency, prefers to “go around” these items, rather than through them. But with repeated zapping, and herbal parasite treatment, you can decimate them, too, and stop reinfecting the rest of your body.

In fact, it is such knowledge about reinfection and sources of our pollutants that is the most important contribution of the case histories.

After dealing with pains we will turn our attention to the diseases that aren't pain producing, like diabetes, myasthenia gravis and so on.

Toe Pain

The ends of your feet get the poorest “service” from your blood supply. They are the furthest away. The blood here has the most accumulated acid and the least oxygen supply. The body produces quite a bit of *uric acid* and this should, of course, be excreted into the bladder by the kidneys. But if the kidneys are doing a poor job of this, levels in the body and blood stream rise. The blood can only hold so much. It holds even less in acid conditions such as the ends of your feet experience. Uric acid begins to settle out or precipitate at our feet.

Hippuric acid, too, is found where pain is found. Hippuric acid is made in large amounts (about 1 gram/day) by the liver because it is a detoxification product. It makes no sense to consume *benzoic acid*, the common preservative, since this is what the body detoxifies into hippuric acid. Read all labels on food you buy. Don't buy any beverages or baked goods preserved with benzoic acid. *Citric acid* is fine. The joints of the big toes are favorite places for pain to develop. This is made even worse when circulation is poor. Take a *cayenne* capsule with each meal to improve circulation. If you cannot find your pulse just below your inner ankle your circulation is poor.

The accumulation of uric and hippuric acid invites bacteria to feed on them. As bacteria multiply the tissue fights back with

inflammation. Now the stage has been set for **pain**. Some people do not have pain although these acids and other deposits are present making their joints knobby and unbending. Bacteria have not found them yet or haven't multiplied sufficiently in them to invade your tissues. Perhaps there are other reasons as well.

If you have **toe pain** your course of action is this:

1. Kill all bacteria possible with your zapper. Repeat daily until no further benefit is seen.
2. Dissolve the deposits away. An herbal recipe, quite elaborate, is capable of doing this. It is called the Kidney Cleanse (page 549). Toe deposits are made of the same crystals as kidney stones, which is why the Kidney Cleanse works for toe pain. But because these deposits are far away from the kidney, it takes longer than merely cleaning up kidneys. It may take six months to make a significant dent on these deposits. This will at the same time remove kidney crystals so that these are no longer a source of bacteria.
3. Get teeth cavitations cleaned (cavitations are bone infections in the jaw where a tooth was pulled; it never healed; see Dental Cleanup page 409). This can "magically" stop the toe pain the same day as they are cleaned. The effect lasts for days afterward showing it is not the dental anesthetic that is responsible. It also teaches you that the bacteria in the toes can come from the teeth. But pain may return as other bacteria find the deposits.
4. Clean the liver of stones using the Liver Cleanse (page 552). This, too, can give immediate pain relief in the toes showing you they are a source for bacteria. Liver cleaning may take you two years to complete! Meanwhile your toe pain is receding. And, of course, this pays extra dividends in health for your body.

5. Reduce the acidity in your toes. First, check your acidity with pH paper meant for testing urinary pH, called Nitrazine™ paper. Ordinary pH paper, as for fish tanks, is almost as accurate and will serve as well. Tear a ½inch piece and hold it in the urine stream. Early morning is the time your urine is most acid. If this reads below 5.5 your feet must have been even lower in the night (lower is more acid).

Once deposits start, it is hard to stop them. If they start forming at 2 a.m., they are likely to continue for several hours even if the pH goes back up to normal. Your tactic will be to go to bed with an alkalizing action. Taking a calcium and magnesium supplement at bedtime, drinking milk at bedtime, using baking soda at bedtime are all remedies to be tried. They should raise your urinary pH to 6 in the morning.

Balance Your pH

Most persons with painful deposits anywhere in their feet have a morning urine pH of 4.5! At 4.5 it is safe to guess that a lot has precipitated again in the night. During the day, your body's pH swings back and forth. The urine gets quite alkaline right after a meal; this is called the *alkaline tide*. Three meals a day would bring you three alkaline tides. During these periods, lasting about an hour, you have an opportunity to dissolve some of your foot deposits. But if you allow your pH to drop too low in the night you put the deposits back again. The net effect decides whether your deposits grow or shrink.

To alkalinize yourself at bedtime, choose one of these options:

1. Two oyster shell tablets, equaling 750 mg. of calcium plus a magnesium oxide tablet, 300 mg (see *Sources*). The magnesium helps the calcium dissolve and stay in solu-

tion. Taking more calcium at one time is not advised because it cannot be dissolved and absorbed anyway and might constipate you. For the elderly only one calcium tablet is advised. Take calcium tablets with vitamin C or lemon water to help dissolve (¼ tsp. vitamin C powder; adding honey is fine).

2. One cup of sterilized milk or buttermilk, drunk hot or cold, plus 1 magnesium oxide tablet, 300 mg. (adding cinnamon is fine).

   If these two remedies work for you, your morning urinary pH will come up to 6.0 but if for some reason they don't, you need to take more drastic measures. Take the supplements and milk earlier in the day and reserve bedtime for:

3. ½tsp. baking soda in water. This is *sodium bicarbonate*. But don't use baking soda from a store because most brands I have tested are polluted with benzene! (See *Sources* for safe baking soda.) Using a combination of sodium and potassium bicarbonate in a ratio of 2:1 is actually a much more healthful potion. You can make your own or ask a pharmacist to make it for you. Mix two parts baking soda and one part potassium bicarbonate (see *Sources*) in a jar. Keep tightly closed. Label it **sodium potassium bicarbonate alkalizer** (this potion is also very useful in allergic reactions of all kinds). Take 1 level tsp. in water at bedtime. If your pH reaches 6 in the morning continue each night at this dose. If it does not, take 1½tsp. Keep watching your pH, since it will gradually normalize and you will require less and less. If you are using plain baking soda, instead of the mixture, watch your pH each morning, also, so you can cut back when the pH goes higher than 6.

Persons with a limit on their daily sodium intake must carefully count the grams of baking soda consumed in this way.

Each tsp. weighs about 2 grams, of which half (1 gram or 1000 milligrams is sodium. The sodium/potassium mixture would only give you half as much sodium (½gram per tsp.). By comparison, the usual daily intake of sodium is about 5 gm., although salt eaters consume twice that amount.

You have done five things to pull the rug out from under the bacteria living in and around the deposits in your toes. Now when you kill bacteria with your zapper, you can expect the pain to go away and stay away.

Locations at the base of toes may be painful due to a *neuroma*. Deposits and bacteria here are even more painful because this is the location of nerve centers. If the build-up is large, you may prefer some surgical help or a cortisone shot rather than wait several years for solid relief.

Foot Pain

This kind of pain does not involve as much deposits as toe pain and is therefore easier to clear up. But trauma to the foot is more important. First, even though your shoes are comfortable, change them. Get wider shoes, longer shoes, lower heels. Alternate two pairs of shoes in a single day. In your home take shoes off.

Keep feet very warm. Wear natural fiber socks, not synthetics. If your circulation is poor, take a *cayenne* capsule with each meal. When circulation is moderately poor, your feet get cold easily. When circulation is very poor, the heart pulse cannot be felt in your feet (take your pulse just below your inner ankle). Again:

- Get teeth cavitations cleaned (Dental Cleanup, page 409).
- Do the herbal Kidney Cleanse (page 549).
- Kill parasites and bacteria with a zapper.

- Then clean the liver (page 552).
- Check body pH in the morning upon rising using Nitrazine™ or other pH paper. Correct it as discussed in Toe Pain. If the urinary pH is 7 or higher, it means you have a bladder infection. Treat it immediately (page 101).

If there is any swelling around the foot or ankle, you are "holding water." It may be called *edema*. This is due to poor adrenal and kidney function. The adrenals are located on top of the kidneys and together they regulate how much salt and water stays in your body. Because they are situated so close together, they share their parasites and pollution. When the kidneys form *kidney crystals* the flow through the kidney tubes is hindered, and less water and salt can leave the body. It stays in your tissues as edema.

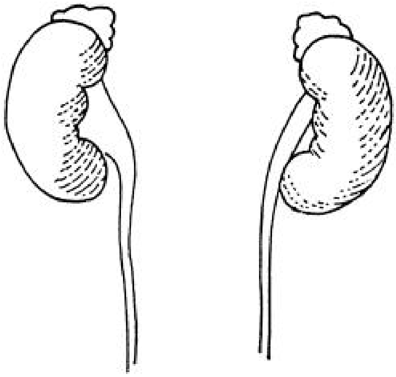

Fig. 10 Your kidneys with the adrenal glands sitting like hats on top. The ureters lead to the bladder.

The kidney herb recipe will dissolve the crystals. But you must remove toxins such as metal from tooth fillings and kill parasites. Continue doing the Kidney Cleanse after the original six weeks is up until all foot pain and edema are gone. Continue killing your tiny invaders with a zapper twice a week. You may need to cleanse the liver several times, too, before all the pain and edema are gone.

The supplement, *pantothenic acid* (see *Sources*), is particularly good for feet. Take 500 mg three times a day for several weeks to see if it helps. Massage and reflexology are also good for them.

Heel Pain

Sometimes heel pain is due to heel spurs, sometimes it is not. Heel spurs are due to deposits. The usual heel deposits are *uric acid* and various *phosphates*. Uric acid deposits become a breeding ground for bacteria. The phosphates give the deposits a rigid structure that is hard to dissolve. You may have to choose a pain killer, get specially built “orthopedic” shoes, or stop your daily walks to get relief from the piercing pains. These will not cure the problem but may “buy you some time” while you make basic changes in your lifestyle.

Stop drinking tea and cocoa because they contain *oxalic acid*. Stop drinking coffee, decafs, fruit juice and soda pop because they are contaminated with solvents. All of these must be detoxified by your body and eliminated by your kidneys. We should spare the kidneys these extra tasks when we wish them to clean up heel spur deposits. You must first dissolve the deposits, then help the kidneys eliminate them.

To dissolve them, you alkalinize your body's pH, watching over it carefully with pH paper or Nitrazine paper. To help kidneys eliminate them, use the Kidney Cleanse and drink lots of water. Only water, not a beverage, helps the kidneys to eliminate.

Drink a pint of water upon rising in the morning, and a pint of water between meals.

Do not buy “special” water. Use your cold tap water only. If you don't like the taste of your own tap water, get it from somebody else's tap. Use only glass or reused polyethylene jugs. Polyethylene is opaque, not clear. It has already lost its plasticizer into the water that stood in it earlier. (Somebody drank it!) Purchased water has traces of solvents from machinery and sterilizing equipment used in its bottling. Your own tap water is not pure (indeed it may have 500 toxic elements), but it never contains solvents in amounts I can detect.

Traces of solvents are worse than traces of other toxins.

Water filters are not the solution. They trap the pollutants and then allow a tiny amount to enter the water on a daily basis. Chronic toxin consumption is much worse for your health than periodic surges of toxins. You can use a small filter, that gets changed every month without much risk. This would dechlorinate the water, at least, and improve the flavor. The pitcher variety (it should be made of hard, inflexible plastic) and the faucet variety are listed in *Sources*.

Bottled water is popular,

and tasty, and has appealing advertising, but it is just not safe.

Why is it easier for everyone to spend dollars per day, for the rest of their life, buying water instead of insisting that their water pipes are metal-free?

Another reason not to drink water from bottles, however convenient, is that it is stagnant and is soon contaminated with our own bacteria from contact with mouth or hands. *Staphylococcus (Staph)* and *E. coli* are commonly seen. The solution is not to add still more chemical disinfectants, the solution is to drink from a flowing source, such as our faucets. If you must carry water, use glass containers; plastic is porous and much more difficult to wash clean (sterilizing in a dishwasher is effective, though).

By drinking a total of four pints of water in a day, the kidneys will notice the assistance. (Any single herb tea without added salt or sweetener counts as water.) The urine will stop having any odor and will become very light colored. This is especially important while you are dissolving the heel deposits since your body is now carrying these in the circulation.

Killing bacteria with a zapper may give you instant pain relief and is, of course, beneficial to your body. But you can't expect the pain to stay away until the deposits are gone and the source of bacteria is also removed. Go after the usual sources: teeth and stones. Another source is dairy food contaminated with *Salmonella* and *Shigella* bacteria. Sterilize all of it by boiling. Even the amount put on cereal in the morning or used in scrambled eggs is enough to reinfect you! Butter and heavy whipping cream also need boiling. Stop eating yogurt and cheese which can't be boiled.

Phosphate deposits are a mixture of three phosphates: *monocalcium phosphate* ($CaH_2 PO_4$), *dicalcium phosphate* ($Ca_2H PO_4$) and *tricalcium phosphate* ($Ca_3 PO_4$). They are formed by eating too much (way too much) phosphate relative to calcium.

Our high phosphate foods are meats, carbonated beverages and grain products like rice, cereals, breads, pastas and nuts.

The body's normal elimination tactic for phosphate is to combine it with calcium and magnesium in order to neutralize it first. Unneutralized phosphate is very acidic and would burn the delicate kidneys. But where shall the calcium and magnesium be taken from? Magnesium is often in very short supply since it comes from green vegetables in the diet and is not stored up in any special organ. So it falls on calcium to be used for this purpose since it is stored up (in your bones and teeth).

The acid condition created by phosphates dissolves your bones and allows large amounts of bone structure to be washed out with the urine: this can be directly seen with an analysis of urine (you can ask your doctor to order this at a lab). If you catch all the urine in a 24 hour period you can measure all the calcium you have wasted. You should not lose more than 150 mg calcium in a day because this is all you can absorb in a day![4] If you do lose more than 150 mg in a day, you are dissolving your bones at a fast clip. This also means there is too much calcium in your blood and lymph, from dissolving so much bone so quickly.

Once you have dissolved your bones it is not so easy to put the calcium back into them. Your body will try to put it back as soon as possible—as soon as your acid condition is gone. But your bones can't do this without *vitamin D*. Vitamin D may "come from" sunshine and from vitamin D in milk but it isn't that simple. Vitamin D must be activated by your kidneys before it can go to work! Remember, though, it was the kidneys that had a problem in the first place, allowing deposits to form! With old kidneys, clogged with crystals, hampered by heavy metal and mold toxins, and beleaguered by bacteria and parasites, is it any wonder that sunshine and vitamin D fortified milk don't supply large amounts of activated vitamin D? It takes large amounts to put back into your bones the large amount of calcium that dissolved out during the acid state you put yourself in by over consuming phosphate food.

If you can't put the calcium back into your bones promptly where is it to go? It may attach itself to tissues that were never meant to be used this way. Your arteries fill with "scale," your kidneys form calcium phosphate crystals, heels form spurs, joints become knobby with deposits.

---

[4] You absorb 5 to 10% of what you eat. If you eat 1 gram (1000 mg) you absorb 50 to 100 mg. But you absorb 25-40% if it comes from milk!

Young persons and children, with healthy unclogged kidneys, make–that is, activate–ample vitamin D, so even if they consume too much phosphate and develop an acid condition that dissolves their teeth and bones, **they can put the dissolved calcium back in its proper place.** They don't develop hardened arteries, heel spurs, and knobby joints, at least not at first. Their bodies can take a considerable amount of abuse without showing it.

In general, people eat way too much phosphate. Meat eaters eat too much meat. Vegetarians eat too much grain. Most everyone drinks phosphated beverages. In this way we set the stage for hardened arteries, joint disease, calcified tissues that no longer have flexibility. We all get kidney crystals that become stones. This is *aging*. All these deposits invite bacteria to live in them and on them, creating pain.

Old age and pain go together as if they were true partners. Yet it is just the result of bad food choices.

Try to undo as much of this false aging as possible.

1. Reduce your meat consumption. Switch to fish which supplies calcium in the tiny bones. It is true, these bones are made of calcium phosphate and one might expect, logically, to be getting a less effective calcium source. Logic isn't necessarily biologically correct. The bones of fish work nicely as a calcium source and their phosphate content is not too great. Such a diet has worked for many primitive societies. Further, I have never seen a case of mercury toxicity from eating fish; amalgam tooth fillings are our truly significant source.

2. Reduce your grain consumption. Instead of cereal for breakfast, add fruit and reduce the cereal to half. Cut bread consumption in half. Cut pasta consumption in half. Cut rice and corn consumption in half. Eat more bananas and other fruits. Eat more vegetables; always choose potato (not potato chips) instead of rice or macaroni. Always choose a leafy salad instead of pasta salad. At any restaurant or salad bar ask yourself: is this wheat, rice or corn? If so, choose potatoes or other vegetables. Choose coleslaw. Choose mixed fruit. You don't need to go off the grains, only reduce them to improve your condition.
3. Take vitamin D as a supplement. Nothing less than 40,000 units has any real impact by the time there are problems. This strength is available by prescription only (usually 50,000 units, which is close enough). To avoid getting a polluted product, ask your pharmacist to follow the recipe on page 560. (In the past some cases of poisoning by overdosing resulted in this regulation. If you overdose you will get joint and muscle pain and nausea but it is reversible.) Take one a day (not more), for the first three weeks, then two a week forever after.
4. Finally, toss the carbonated beverages right out of your diet or make your own (see *Recipes*). It is not the carbonation that is harmful, it is the added phosphate. Drink water, herb teas, homemade fruit and vegetable juices, milk or buttermilk.

Milk, like fish, is full of calcium in the form of calcium phosphate. Again, logic might speak against the effectiveness of this form of calcium. Again, logic is wrong. Evidently, the calcium and phosphate story must be much more complex than I am depicting here.

Milk works best as a calcium source, in spite of its phosphate content. Possibly the *lactose* and other complexities of its composition contribute to this. My recommendation when deposits have formed anywhere in the body, such as heel, toe, arteries, joints, is to switch to milk as a beverage. Compare the calcium level of your urine before and after the switch (allowing several weeks first). Also compare calcium levels in your blood serum. It should move towards normal. This means **up** if it is too low (below 9.2). It means **down** if it is too high (10). If you are monitoring the effectiveness of the kidney herb recipe in dissolving away your phosphate crystals, notice that drinking milk keeps them from reforming. Taking calcium tablets does not! Taking nothing lets them reform the quickest.

The milk must be 2% or higher in butter fat to be effective.

With your body fluids at their proper acid level, with your kidneys able to flush out acids, with heavy metal toxins no longer settling in, with your bone-dissolving stopped, your heel deposits can shrink. Bacteria have no place to feed and breed. You can kill them several times more with your zapper to catch stragglers. And your heel pain becomes history. Be careful not to bruise the sensitive tissue with too much walking or running immediately after the pain is gone. Wear cushioning socks and well cushioned shoes.

Names in the case histories have been changed to ones of the same sex, picked at random from a telephone directory. Other facts may have been altered in non-essential ways.

Walter Jones, a man of 67, was diabetic for 14 years. His feet and upper legs hurt so much for the past 13 years he could barely shuffle along now. They were also cold and clammy. The herb, juniper berry was added to the Kidney Cleanse recipe to make it even more effective for him and he was advised to stop smoking, using alcohol, and caffeine. He also started the parasite killing program (at that time an herbal parasite program was being used instead of a zapper) and a liver cleanse that yielded several hundred small gallstones. Three months from his first visit his legs were better and feet OK, although he still had some trouble walking due to pain in upper thighs.

Dinah Sagun, a minister's wife, had a lot of walking and standing to do besides traveling a lot. She had knee surgery 15 years ago to remove deposits but now they were getting bad again. Also her heels were sore. My tests showed she had phosphate and oxalate deposits in her kidneys. She was started on the Kidney Cleanse followed by the parasite program. She stopped using benzene polluted items, especially her Chap Stick^{TM}. Her diet was changed to include milk and leave out store-bought beverages. In 4½ months her heel pain was gone along with the bad headaches we had not even started to work on.

Julie Fernandez came with a list of 10 serious problems including foot pain and ankle swelling. Her urinalysis showed crystals. My test showed they were uric acid crystals. Her parasites included *Trichinella*. There was a house cat. Her toxic metal test showed tellurium and platinum accumulation from metal dentalware and arsenic from pesticide. She was to start on a Kidney Cleanse and add the parasite program two weeks later. She was to remove all roach hives and other pest killers from her house and arrange for dental metal replacement. At her third visit, four weeks later, she could walk without pain although she still had *Trichinella* and arsenic problems. Two months later she was aching all over again and her kidneys hurt. She still had *Trichinella* and some dental work to do. She was to do a 5-day high dose parasite treatment and treat the cat also and repeat the kidney cleanse. This relieved her pain.

Dorothy Shelley had numerous pains including foot pain in the arch, lower back pain and cramps with her period. She had oxalate and uric acid crystals in the kidneys so was started on kidney herbs. Three weeks later her foot pain and low back pain were gone.

Paul Longtin had toe cramps, calf cramps, and heel pain. After cutting down on smoking, quitting caffeine and soda pop and taking niacin (250 mg time release; 1 capsule each meal and upon rising in morning) and doing a kidney cleanse, all in three weeks, he was much improved and didn't even need his arch supports.

Juan Onley, age 58, came for his gout in both feet and one hand. He could hardly walk, in spite of soft shoes and pain killer. It started 7 years ago. He also had prostate pain and couldn't sit comfortably. My tests showed his kidneys were full of uric acid, oxalate and cystine stones. His prostate was full of *Gardnerella* and *Campylobacter* bacteria. His wisdom teeth were harboring *plantar wart #4 virus* and *Coxsackie B4* virus. *Gardnerella* often comes with fluke parasites and indeed he had intestinal flukes in his kidney! He began with the kidney herb program, then the parasite killing program. It took five months to clear them all. Then he could walk and sit without pain.

Leg Pain In Children

It is commonly accepted as normal(!) in children to have pain in the shins or calves of the leg. They may even be called "growing pains." Children may cry with the pain and never tell anyone the reason for crying. It happens mostly after napping. This may be caused by cramping of the leg or spasms of the blood vessels. *Lead toxicity* is a common cause of both. Test for the presence of both **lead** and **cadmium** in the tap water. Only your own electronic tests are helpful. Water department tests are much too crude.

ULTRAS
ULTRA
Pampers
Pamper
Pampers
ular Thickness
Pamper
Pampers
HUGGIES
HUGGIES

Fig. 11 All disposable diapers I tested had mercury and thallium! These toxins can be absorbed through babies' skin.

If either poison is

found, test the water supply from each faucet in the house, in the morning, before it has been run. Find the offending sources, change the water pipes to polyvinyl chloride (PVC).

Also search for thallium or mercury in the child's saliva. If it is there, remove all dental metal. Stop using all commercial disposable diapers, dental floss, cotton swabs and bandages; they are polluted with mercury and thallium probably from manufacturing them in foreign countries where it is legal to sterilize with mercuric chloride. Test again, several times, after plumbing or dental work has been completed. To relieve pain:

- immerse legs in warm water
- massage legs gently
- give 25-50 mg niacin, not time-release, to dilate blood vessels.

Leg Pain In Adults

Leg pain in adults is usually associated with *cadmium* or *thallium*. Cadmium is present in tap water that runs through corroded galvanized pipes. The cadmium is probably a contaminant of the zinc used for galvanizing. Test the water, electronically, for cadmium. If you have all copper pipes but there is cadmium in the water, there must be a short piece (a **Y** or a **T** joint) made of old galvanized pipe lurking somewhere. Track it down by testing water from all your faucets.

Cadmium causes the blood vessels to spasm and it is made worse by smoking, that's why the condition is sometimes called **Smoker's Leg**. But extremely painful legs are due to chronic thallium poisoning more than any other cause!

It is very important to know exactly how toxic thallium is. Read the clipping on page 417 right now!

Where would you ever get thallium? From your very own mouth! The mercury in fillings is often itself polluted with thallium! Replace your amalgam fillings with composite.

Bandages

Sanitary napkins and tampons

Cotton swabs

limpiecitos
HAS Hg TL

Cotton balls

ACE
First Aid
First Aid
First Aid

Floss

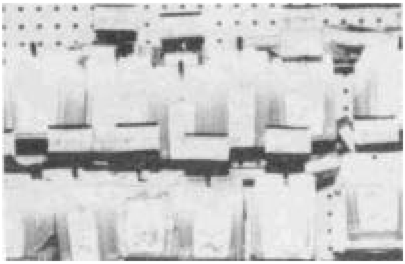

Toothpicks (the one on the right is ten years old and had no mercury or thallium)

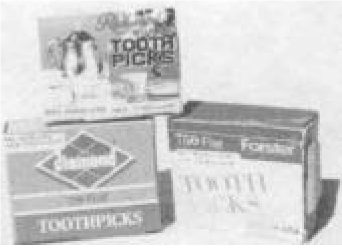
TOOTH PICKS
Diamond
TOOTHPICKS
Forster
TOOTH PICKS

Fig. 12 Do not use any commercial personal products, the risk of pollution is too great.

Thallium has another source: it is riding along as a pollutant in cotton swabs, cotton balls, commercial bandages, toothpicks, floss, gauze, sanitary napkins, tampons, disposable diapers, and paper towels. Evidently these are being sterilized with mercuric chloride which, in turn, has thallium pollution.

- Line **disposable diapers** with a tissue
- Line **sanitary napkins** and pads with a tissue
- Use the **polyester puff** in the top of vitamin bottles in place of cotton balls. Twirl some around a plastic stirrer for a swab.
- Use pieces of **tissue and masking tape** for bandages.

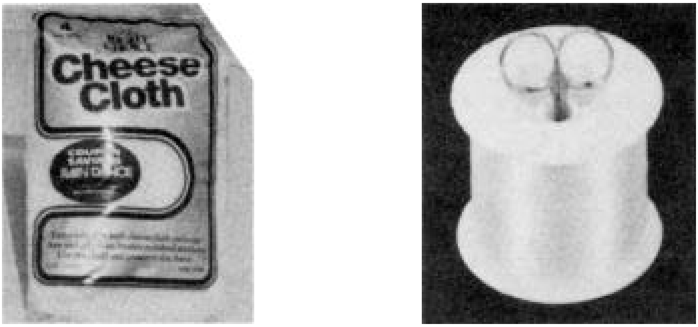

Use cheesecloth in place of gauze.

Use monofilament fish line (2 or 4 lb. test) for floss

Fig. 13 Safe substitutes for personal products.

If you do have thallium in your white blood cells and you haven't used toothpicks, etc. earlier in the day, then it is in your tooth fillings and **you have no higher priority than getting the**

amalgams out. Find a dentist immediately who will remove them, drilling deeply and widely not to miss a speck of it, thereby getting the thallium out, too. You cannot cure your leg pains without removing thallium.

Leg Pain Protocol

1. Stop smoking.
2. Repair plumbing.
3. Do dental cleanup and chelate out the mercury and thallium that has gotten into your tissues with EDTA (see *Sources*). You will need to find a chelating doctor; ask a friendly chiropractor to help you locate one. Or at least take thioctic acid 100 mg, (2 three times a day) and vitamin C (5 gm or one teaspoon) daily for a month.
4. Take magnesium oxide 300 mg (take 1 twice a day).
5. Take niacin, as much as you tolerate—time release varieties are less effective. Try 50 mg with each meal.
6. Change your diet to reduce phosphate and include milk (sterilized).
7. Do the herbal Kidney Cleanse (page 549) followed by a Liver Cleanse (page 552).
8. Zap yourself on alternate days at bedtime. If this zapping makes no difference whatever, your problem is purely spasms. But if you get relief, even if it's very short lived, you must have killed something. Bacteria must come from somewhere. Concentrate your efforts on dental health and better diet.

Jean Booth, age 30, had sore, tired legs and severe three-day headaches. She would get stabbing pains in back of her thighs. After we found thallium and mercury in her kidneys she did a Kidney Cleanse and got all her metal tooth fillings replaced. She then felt fine for one year. Suddenly she got fatigue and heavy legs again with stabbing pain at the outer thigh. She had seen a neurologist. Her chiropractor suggested it was leftover mercury so she came back to us. Indeed, she was toxic with lead, mercury, thallium, but her dentist could not find the leftover metal in her mouth. Three cavitations were cleaned; she was put on thioctic acid; eight varieties of bacteria and viruses were killed with a frequency generator and her legs became well again. Our test showed thallium at 4 teeth, but it was not a big enough deposit to show up on dental X-ray. She may eventually choose to have these redrilled.

Charlie Snelling was a picture of pain: pain in arms, elbows, shoulders, wrist, hands, chest, low back, legs, knees, and feet. He was started on kidney herbs and a few pains were reduced. He was toxic with cadmium so he changed his plumbing to plastic. However, he continued to be toxic with cadmium and thallium throwing suspicion on his numerous old tooth fillings. He got them all replaced and cavitations cleaned. He used our frequency generator to kill *beta Streptococcus*, *Pseudomonas*, *Troglodytella* and *Staphylococcus aureus* all of which were under one tooth (#15). A year later he still had bouts of leg pain. He still had numerous bacteria under his teeth because the jaw bone was not healing. He had not been taking vitamin D, nor magnesium nor drinking milk for the necessary calcium. But he had improved enough to go back to work full time.

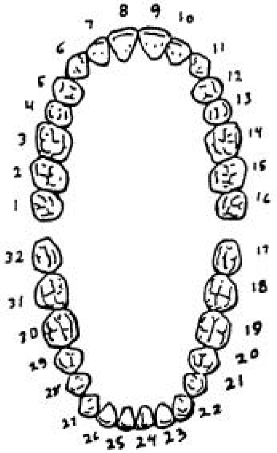

Fig. 14 Tooth numbering system.

Victor Abhay, age 16, could no longer play in high school sports because of knee pain. It began "with a virus" and high temperature two years ago. His knee started to bother him after that. He had *cysteine* kidney crystals and four parasites: *Cryptocotyl*, human liver fluke, *Echinococcus granulosus* cyst and *Echinostomum revolutum* in his white blood cells. He was to start the kidney herb recipe and follow this with the parasite program. Five months later, when we next saw him, his knee was fine.

Kim Murphy, 45, had painful legs, feet and knees. They also were swollen and itched. She was parasitized by *Trichuris*, (dog whipworm). She also had tapeworm stages (*Taenia pisiformis*) and intestinal fluke in the intestine. She started on the kidney herbs, followed by the parasite program. She stopped using zirconium-containing products (deodorant) and barium (lipstick). She stayed out of bus exhaust. (She knew she was allergic to diesel exhaust.) In one month her leg pain and itching were gone; slight swelling remained.

Nancy Tong, 80's, had edematous legs. In fact she could get no stockings on. They appeared like pillars with no taper at all. She was on diuretic medicine from her doctor. It kept her blood pressure down but she was losing ground with water excretion. This raised her general toxicity (blood BUN[5]) which made her feel bad most of the time. Yet she drank enough water, curtailed her salt, used no caffeine and had no really bad habits. She had to wear several pads for incontinence. We found she was toxic with cadmium and lead, which were probably responsible for her huge accumulation of kidney stones. The metals were in her tap water and she was unable to resolve this problem since she lived in a senior citizen center. We advised her to move, or to have her tap water carried in, but she could do none of these. Although the situation was hopeless, she did the kidney cleanse, parasite killing program and changed her metal rimmed glasses and wrist watch to plastic. She gained enough ground from these improvements to be able to wear elastic hose and thereby give some physical assistance to her body. This encouraged her to do a liver cleanse. She had a headache with the cleanse but immediately afterwards she fit into a smaller size “Keds” (elasticized stockings). A half a year from the beginning, her legs had taper to them; she didn't even mind wearing a dress.

Fibromyositis and Fibromyalgia

When pain is widespread, not just in joints or legs but in many muscles and soft tissues of your body your doctor may call it *fibromyositis* or *fibromyalgia*.

---

[5] BUN stands for blood urea nitrogen. It is a body waste and is normally kept low by the kidneys.

For bacteria to get all over your body, they must be riding along with parasites that get all over your body. Microscopically tiny roundworms can do this.

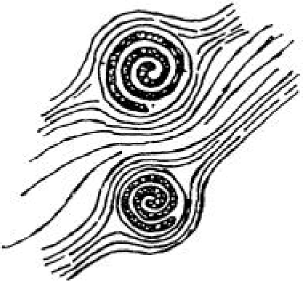

Fig. 15 Trichinella larvae settled in muscles.

Trichinella is the most common cause of these diseases, but sometimes *Ascaris* larvae or *hookworms* or *strongyle* larvae are the main culprits. These wormlets bring hosts of **bacteria** with them, mainly *"Streps"* (*Streptococcus* varieties) and *"Staphs"* (*Staphylococcus* varieties), but also *"Clostridiums"* (*Clostridium* varieties) and "Campyls" (*Campylobacter* varieties). The bacteria are probably the pain causers. By killing all bacteria—*Staphs*, *Streps*, *Clostridiums* and *Campyls*—using a zapper, you may get relief for one hour!

By killing *Trichinella* and *Ancylostomas* (worms) first, followed by the bacteria, you may get relief for several hours. By killing the parasites and bacteria in every household member and the pets at the same time and by never putting your fingers to your mouth, you can expect permanent pain relief. It is interesting to speculate why the other family members, who are also infected with these tiny roundworms don't develop *fibromyalgia*. Perhaps the larvae stay in the intestine or go to the diaphragm (causing coughing) or the eyes (causing "lazy" eye muscles). Perhaps they merely cause **anemia**. *Trichinella*, hookworms and *strongyles* are extremely difficult to get rid of in a family.

These roundworm larvae undoubtedly cross the placenta into the unborn child during pregnancy, too. So they can be "inherited." Try to clear up the whole family before the next pregnancy.

Clearing up pets of these parasites is even harder. The best advice is to give your pets away. They will continue to harbor them even though they are on a pet parasite program. It is impossible to stay free of the parasites your pets have: they will move to your soft tissues immediately, giving you the bacteria and inflammation again.

The next most important advice is to keep fingers out of your mouth (read Hands, page 397). None of these parasites enter through your skin (this is in spite of teachings that hook-worms enter this way), you must put them into your mouth somehow! Consider your mouth off limits to anything but food and kissing.

Finally, if there is a baby in diapers in the family, be patient. When diapering days are over you will have less bowel contact, giving you an opportunity to finish your own treatment. Meanwhile, wash fingernails in Skin Sanitizer (see *Recipes*) after cleaning up children's bowel movements, diapers and yourself. Use borax liquid for soap to leave an antibacterial residue on the skin.

Try to identify your parasites before killing them so you can be on the lookout for them in the future. Get slides or dead cultures of various pathogens and search in your white blood cells. If you can't do this, at least save a saliva sample of your own; keep it frozen or preserved. Also make a saliva sample for your pet. This gives you specimens to test yourself for later. You won't know which parasite is in this saliva specimen, but if you ever test positive for it again you will know you got it back. Then zap yourself.

Brenda Byrd was diagnosed with myofibrositis two years earlier at the age of 36. Her blood test showed triglycerides slightly high (152 mg./DL), indicating the beginnings of urinary tract problems. Her urinalysis stated “hazy” (hazy with bacteria or crystals) instead of clear urine. It also listed white blood cells, red blood cells, and a few bacteria present in her urine. Our tests revealed mercury and numerous other heavy metals distributed in her thyroid, stomach, kidney, lungs, bones and bone marrow. She was also full of beryllium (usually from "coal oil") contained in the hurricane lamps she kept in every room. She had numerous parasites, including *Strongyloides* and hookworms spread through her body tissues. She was thrilled to learn how to get her health back and started with the dental problem.

Marcia Cochran, 36, had muscles twitching all over her body. It was diagnosed as fibromyalgia. Her joints were tender and her chest felt "tight." She had bronchitis twice a year. She was on Amytryptaline™ for muscle twitching and Bentyl™ to calm the intestine (spastic colon). She had depression with it and was on medicine for that. It all started with fever and chills that she thought was the flu but after they went away, she was left with a tremor. Sometimes she felt that little electric shocks were going through her. She had *Ascaris* and hookworm larvae widespread in her body. She had sheep liver flukes in her liver. She was started on the parasite program. She was also toxic with PVC and tellurium (dental metal). The PVC was traced to plastic storm windows applied to the inside of the window and to new shower curtains. Two months passed and she had not solved any of her problems. Then she did her first liver cleanse and got over 100 stones out. This instantly reduced her fibromyalgia to occasional attacks. She was so encouraged she decided to go ahead with dental cleanup.

Joint Pain or Arthritis

Two main kinds of arthritis are recognized clinically, *osteoarthritis* and *rheumatoid arthritis*. In osteoarthritis the joints have bacteria living on the deposits left there. In rheumatoid arthritis the bacteria come from larger parasites—wormlets actually living in these joints. The worms are the common little roundworms whose eggs hatch into microscopic wormlets that travel.

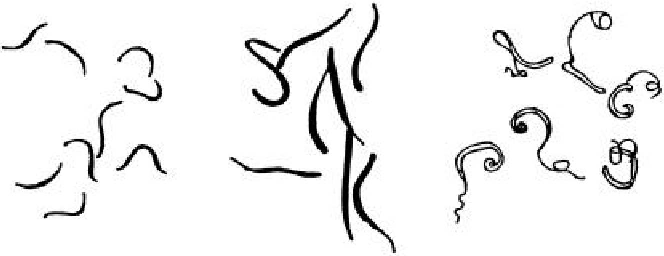

Fig. 16 Hookworms, strongyles, and whipworms.

We have four common roundworms: *Ascaris, hookworm, Strongyloides, Trichinella*. Their life cycle normally directs them to travel to the lungs but in some people they travel through the entire body, including brain, muscles and joints. More research is necessary to explain this. My suspicion is that there are toxins, like **mercury**, **thallium**, **cadmium**, **lead**, as well as **solvents**, distributed through the body, lowering immunity and allowing the tiny larvae to reside there. Once the pathway (routing) to these organs has been established, it continues to be used by other parasites as well. Soon a variety of parasites, their bacteria and viruses, and pollutants are all headed toward these organs.

Osteo or Common Arthritis

When joints are painful it is a simple matter to kill the bacteria with an electronic zapper. Treat yourself with a zapper daily until the pain is gone. Maybe it will stay away, but chances are the bacteria have a steady source.

The most common source for *Staphs* and *Streps* are small abscesses in the jaw bone, under and beside old extractions, root canals and mercury fillings. You may get immediate pain relief just from a dental cleanup, and again disappointment may follow. *Staphs* and *Streps* are such ubiquitous bacteria, they may come not only from jaw bone infections but from gallstones, kidney stones and other parasites. The correct treatment for arthritis is a complete overhaul of body health: a diet cleanup, a body cleanup, and environmental cleanup.

Start with the herbal parasite program and zapping. Follow this with a kidney cleanse, then liver cleanse. If any toxin is overlooked, especially asbestos and fiberglass, it is sure to find your joints and permit bacteria to return and cause pain. Make sure to correct your body acid levels after doing pH measurements of the urine (page 57).

Arthritic deposits contain a large amount of phosphates combined with calcium. This calcium came from some other bone, such as the base of your spine or the wrist. Here the bones are getting weaker due to this calcium loss. Calcium was taken out of your bones for the simple purpose of neutralizing the excess phosphate in your diet. Reduce phosphate consumption (meats, soda pop, grains) by half, eating fish, milk, vegetables and fruit instead. Drink three cups of milk a day. If you are allergic to milk, do several liver cleanses, switch brands of milk, use milk digestant, and use it in cooking and baking. Cheese and cottage cheese are not substitutes for milk (the calcium stayed in the whey). Dairy products must be boiled before consuming and should be no less than 2% butter fat. It takes bile to make calcium absorbable, yet milk with less than 2% butterfat does not trigger the gallbladder to empty it's bile at mealtime.

If you are not used to dairy products, start slowly and work up gradually to the 3 cups a day needed.

Carol Lachance was diagnosed with arthritis of her back and knees. (She had spurs in both places.) She was positive for *Trichinella*, *Ascaris* and *Dirofilaria* (she also had pain over the heart). They had an outdoor dog. Her blood test showed a high phosphate and alkaline phosphatase level showing she was dissolving her bones. After changing her diet to include milk, extra oyster shell calcium (one a day), magnesium oxide and vitamin B_6, and reducing her meat and grain consumption her phosphate level went down to normal (below 4). She did the kidney cleanse and liver cleanse as well as parasite program but still had pain. A toxic element test revealed lead in her bones. It was traced to a drugstore variety multivitamin tablet she had taken daily for years. When she stopped these and added prescription vitamin D (50,000 units) for three weeks to help her bones heal she got relief.

Gail Hildebrand, age 62, had painful arthritic hands and gum disease. She had 12 parasites free-loading on her and was toxic with asbestos. This was traced to her clothes dryer. Four months later, after killing parasites, her hand pain and gums were much better. She had the dryer vent taped up tighter and this got rid of her asbestos problem. She started on kidney herbs and in one month saw that her enlarged knuckles were beginning to go down.

Norma Littrell, age 53, came in for her severe arthritis of six years. Her knees, shoulders and hands were painful. I explained to her that painful shoulders did not belong to the arthritis picture but had a gallstone etiology which she could easily fix in a single night at a later time (liver cleanse). She also had mid-back, upper back and lower back pain; again the upper back pain belonged to the liver problem. She had tricalcium phosphate kidney crystals as well as uric acid crystals. She was told to go off coffee (no decaf either), decrease her meat and grain consumption (phosphate) and increase milk and use stone ground cornmeal products (genuine stone ground tortillas as well as pickled pigs' feet are high in utilizable calcium). She started on the kidney herbs. In 33 days her low back pain had improved a lot, she could wash her own hair again and she could sit down and get up from her living room floor without pain. She was elated but we recommended less stress than such exercises for her joints.

Santa Fe
Santa Fe

Fig. 17 Stone ground corn tortillas, high in calcium.

Patricia Robinson, age 76, had pain in her knees, feet, lower back, hands and wrists. Also in her shoulders and upper back which is not part of the arthritis picture. She had swollen puffy eyelids which is a telltale sign of Ascaris. She had heartburn, cold feet and insomnia as well as high blood pressure. She was on several medicines. We started her on the kidney recipe and ornithine (four to eight as needed at bedtime) for sleep. We hoped she could stop her Ativan™ drug soon. She was to go off coffee, tea and decafs. Her hands were knotted and misshapen at the joints, also tender. Two months later there was little change; her doctor had put her on Prozac™ and she hoped that would do a lot for her (so she stopped the kidney herbs) but it didn't. The parasite test still showed *Ascaris* and she was started on the parasite program. We also found fluoride (from toothpaste), iridium, samarium and palladium, all from her tooth implants. She was to remove as much metal as the dentist could replace, clean cavitations, and take thioctic acid, 2 a day, to help clear metal from her body. In four weeks the sharp pain in her back was gone and in three more weeks the pain in her hands was gone.

Lynne Snyder, 72, had pain in every joint and had to be on pain medicine to keep moving. Her potassium level was very low (3.6—an adrenal/kidney problem), and she was started on kidney herbs, carrot juice, vegetable juice and bananas. In ten days she could feel some new energy but her pains were terrible, especially her knees. She was taken off tomato juice, cranberry juice, citrus, pepper (she was using a lot), and given buttermilk as a beverage which she enjoyed. After three weeks of kidney herbs she was started on parasite killing herbs. In another month her arthritis was much better. She was not on any pain medicine and could get to sleep without it. She thought it was mostly going off pepper that helped.

Rheumatoid Arthritis

When inflammation and swelling affect your joints, besides pain, it is called *rheumatoid arthritis*. In addition, a blood test may reveal “rheumatoid factor” to be present. In this case, the common tiny round worms have invaded your joints. These are *Ascaris*, *Ancylostomas*, *Strongyloides*, and *Trichinella*. Their eggs are everywhere around us, in dust and dirt and the filth under fingernails and our own bowel movements.

Superior sanitation is your first defense. Rinse fingernails in alcohol after cleaning up bowel movements or changing diapers. Never, never, tolerate long fingernails in any family member. If this discipline can't be enforced, do not allow food preparation by “long nails” unless gloves are worn. He or she may not be getting ill (yet) from the family parasites, but you are. Of course, you can kill them with a zapper (internally, not the ones under the nails) but that is after you have been infected. They are easily picked up again. It would be wise to zap for roundworms every week, just in case.

Make sure your pet is treated with parasite herbs or by zapper as well. A pet that goes outdoors will quickly (the very next day) bring these roundworms into the house again. Give away your pets if possible.

Check for dental problems. Do the Dental Clean-Up (page 409). Then do a Kidney Cleanse and Liver Cleanse.

You may relieve your pain and begin to heal immediately after zapping but it is wise to do all the health programs, anyway. Change your diet. Reduce phosphate, start using sterilized milk for calcium. Switch to fish from meats. Drink much more water. Use only harmless beverages (see *Recipes*) and foods. Switch to toxin-free body products. Live in a non-toxic house. Stay on a maintenance parasite program of herbs, and zap regularly. Stay on the kidney cleanse for three to six weeks and repeat a one-week session every few months to keep removing deposits which may also choose these sick joints to settle in. Knees are a favorite location for rheumatoid arthritis. Knees are very dependent on kidneys.

To summarize, do everything as for osteoarthritis, emphasizing the roundworm parasites for elimination.

Allergic Arthritis

A prominent food toxin that is said to affect knees is *piperine*, found in the pepper family which includes black and white pepper (not cayenne). The *Solanaceae* family of plants (potatoes, tomatoes, eggplant, tobacco!) also has a common "allergen" that produces joint pain. Try going off these for two weeks to see if it helps. Two more chemicals that can trigger arthritis-like pain are *hippuric acid* and *phenylisothiocyanate* (PIT).

Your body makes large amounts of hippuric acid, up to a gram a day. It is the product of benzoic acid detoxification by the liver. Quite a few fruits contain natural benzoic acid. But we can easily triple and quadruple our benzoic acid intake by consuming commercial beverages and pastries where benzoic acid is used as a preservative. It is indeed a "natural" preservative. All of it must be detoxified, though, and this gives us way too much hippuric acid. The kidneys are unable to excrete such overloads of hippuric acid, so it distributes itself in our organs. I suspect a simple mechanism could explain its pain-triggering action: hippuric acid molecules could attach themselves to our cell's conductance channels keeping the gates jammed open. This might invite bacteria and viruses to enter there. But there are also hippuric acid-loving bacteria that feed on it. Perhaps hippuric acid can cause pain without the help of bacteria. This requires further study. It is only sensible for persons with chronic pain not to consume benzoic acid (or benzoate) preserved foods.

PIT is a food chemical found especially in chicken, eggs, the cabbage family, peas. PIT is also part of the body's own chemistry, taking place in the liver, and involving detoxification of *cyanide*-containing foods. Many vegetables, notably the cabbage family, contain such cyanides, giving them protection from insects, disease, and grazing animals. It can take the liver a week to detoxify a meal full of these cyanides: in the meantime, PIT levels are higher in the body. PIT is very reactive. In fact, it is the chemical used as a general reactant with amino acids in the well known **Edman degradation** reaction. But now, your body is the reaction flask, supplying the amino acids. Since all organs supply amino acids, it is no wonder they can all react to PIT, giving you multiple allergic reactions and pains.

If you have any kind of arthritis, stop eating the high-PIT foods and clean the liver until you are free of all allergies you are aware of. This suggests that the liver is capable, again, of detoxifying the cyanides for you in a reasonable time and you may eat them again. These foods do have many benefits, of course.

Joint pain, or arthritis, was known in antiquity long before dogs and cats were household pets and giving us their parasites. Pigs and horses harbor these roundworms too and may have been the source at that time.

Herbs and treatments that help arthritis are, therefore, plentiful. Maybe they act by killing roundworms, bacteria, and viruses, or help metabolize hippuric acid and PIT. Homeopathic treatments, as well as massage, heat and electronic devices also help. With this wide range of effective treatments dating to the distant past, why is none of them a permanent cure? The answer is simple. The common roundworms are everywhere about us, sanitation is poor, and our civilized lifestyle leads to deposit formation that invites bacteria. But knowing this, you can stop your pain and remove the causes to become one of the first humans to achieve a permanent cure.

Verna Plumb, age 46, was diagnosed with rheumatoid arthritis four years earlier. Since then she had been continuously on methotrexate™ and prednisone.™ This had caused her to gain a lot of weight from water retention. The drugs were no longer effective and she would need to do something else very soon. She had the typical causes: her body was toxic with mercury and nickel from tooth fillings. Her kidneys were full of five kinds of stones. She had numerous roundworm parasites including two kinds of *Ascaris*, two kinds of hookworm, *Strongyloides* and *Trichinella*. She would have to clean everything up to get relief. She started on the kidney herbs, killed parasites with a frequency generator and in two months noticed her swelling was receding.

Camille Franklin had hands that were swollen and hot and painful. She also had "arthritis" in shoulders and knees. We explained that shoulders were not part of the arthritis picture. She could deal with that in a single evening, soon. She also had bone spurs at sinuses which needed surgical removal. Her kidneys revealed tricalcium phosphate crystals. She was given a diet change; onto milk, fruits and vegetables, off other beverages, less meat and grains. She was started on kidney herbs. In five weeks all the swelling and redness and heat was out of her finger joints.

Thigh Pain

Inner thigh pain often stems from the *sciatic nerve* which is suffering pressure at the lower back. If this is so, chiropractic adjustments should help. The correct treatment, after killing bacteria electronically, is to clean up the entire kidney area using the kidney herb recipe.

If this is not the correct explanation, (and you're not getting pain relief) you may have a trauma—perhaps you overstretched your leg in some exercise. The minor trauma invited bacteria to settle there and give you pain.

If the pain recurs after clearing it several times, there must be a chronic source of bacteria. Since the kidneys are already cleaned, consider the teeth, as well as recurring parasites, and the liver. Clean the liver every two weeks until 2000 or more stones have appeared and no more appear. This could take 1-2 years. Be patient. Do a dental clean up. Keep killing the bacteria so they can't spread. And, of course, let the painful leg rest.

Hip Pain

Hip pain is always due to bacteria. In fact, these bacteria regularly come from two sources: the kidneys and the teeth. This simplifies the treatment since neither of these places takes a long time to clear.

Start yourself immediately on the kidney cleanse (page 549) to clean your kidneys. Continue on the recipe until the dental cleanup has been completed. Both must be clean together to stop exporting bacteria to the hip.

The dental cleanup could take several weeks if extensive work is required. During this time, kill all parasites electronically. Keep killing bacteria, especially *Staphs*, *Streps*, *Clostridia*, *Campyls*.

The dental problem is not always on the same side as the hip pain. The jaw may be very fragile and porous, full of invading bacteria. Cleaning these cavitations may give immediate pain relief in the hip (proving the bacterial source). But getting the jaw bone to heal by taking up calcium again is not guaranteed by the cleaning process.

Give your jaw bone every chance to heal:

- Start taking vitamin D (40,000 to 50,000 u.), every day for three weeks from the day of the dental work or before. After this, take it twice a week forever. **Do not take more.**
- Take vitamin C (1+ gm a day) and B_6 (250-500 mg a day) for healthy gums.
- Get one gram of usable calcium in your daily diet.

Bone Strengthening

Vegetable calcium can't be dissolved by our stomachs (*ruminants*, like cows, can dissolve it—they have an extra stomach loaded with special bacteria who do the actual retrieving of calcium). Tablet-form calcium can't be dissolved well either, especially as we age. We need predissolved calcium if a little is to go a long way.

Primitive people who lived on fish or stone-ground meal ate 4-6 grams of calcium a day.[6] Even if only 20% of this got dissolved, they would still have about 1 gram of utilizable calcium for themselves. Chances are good they dissolved even more, since they were young (life expectancy was less than 50 years). Their skull remains show beautiful, cavity-free teeth. They lived outdoors, mainly, so getting enough sunshine-derived vitamin D was not a problem. Their natural diet supplied enough vitamin B_2 to protect them from UV (ultraviolet) damage from sunshine exposure.

But these are civilized times. Our lives are stretched into old age, when our stomachs no longer produce enough acid to kill bacteria, nor to dissolve the minerals in our food. So they need to be dissolved for us. Milk is a beverage where the calcium has already been dissolved by the other ingredients. The lactic acid in milk formed during digestion gives the calcium the correct chelated structure for absorption by the intestine. Even the bile participates in calcium absorption. Milk also contains phosphate, but not too much to be useful. I recommend milk as a calcium-source to heal the jaw bone after and before dental work. You need 1+ grams a day. One quart of milk has 1 gm (1000 mg) of utilizable calcium. You absorb only 250-400 mg. The rest is excreted and eliminated. Only the absorbed calcium can heal your jaw bone. Use milk in cooking as well as a beverage. The calcium in it is indestructible.

Bones are not made of calcium alone. **Magnesium** is essential. Since magnesium is more soluble and easy to assimilate than calcium, the tablet form (magnesium oxide, 300 mg, see

---

[6] Read Nutrition and Physical Degeneration by Weston A. Price, DDS first published in 1939. At least, gaze at the pictures. Ask your library to buy a copy, available from Price-Pottenger Nutrition Foundation, (800) 366-3748.

Sources) will do. If you are not absorbing the magnesium it will stay in your intestine and act as a laxative. If this happens acidify your stomach during meals: always add fresh lemon, vinegar, or vitamin C to your food or drink to help digest milk and dissolve minerals for you. **Boron** (3 mg. once a day) and **manganese** (15 mg. once a day) are additional bone hardeners.

When your diet is improved, your dental problem is cleared, and your kidneys are clean, your hip pain will stay away and you can stop thinking about hip replacement surgery.

Mary Hammond, 48, had two painful hips. She was diabetic and drinking five cups of coffee a day. She had *Staphylococcus aureus* infections under three teeth that she no longer had. After going off caffeine (caffeine may spread bacteria by making tissues more permeable) and getting some dental work done, her right hip stopped hurting. When cavitations were cleaned, her left hip stopped hurting too, but after a week the pain returned. She still had parasites and their bacteria to kill. She had solvents to eliminate and a kidney cleanse to do but she was quite enthusiastic and enjoyed showing off how well she walked.

Groin Pain

Lymph nodes are situated here, as well as in armpits and around the neck. Lymph nodes are your best friends. They are hives of activity. Your white blood cells “nest” here. Lymph nodes sample your body fluids (lymph) at these locations, much like the water department and health department sample our milk and water, making sure they are pure and sterile. Of course, they never really are. But your white blood cells keep working at it. They are busy removing impurities like zirconium and titanium and pathogens like bacteria and viruses.

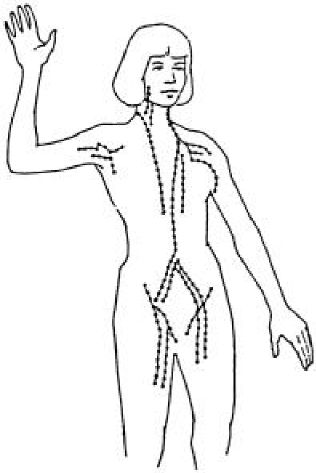

Fig. 18 Your lymph nodes are your best friends.

When an especially challenging dose of parasite or pollution arrives at the node, it enlarges, in order to handle the bigger sized task. This enlargement can be felt. It may press against other organs and you feel the pain of pressure in the groin. This should alert you, of course, to the danger. Kill the parasites and pathogens immediately with your zapper. Start the kidney cleanse. Stop using toxic products and eating toxic food. If you are being effective, the lymph nodes will begin to shrink in a few days, relieving the groin pain. If it does not come back, you probably eliminated the main cause. But if it recurs, try to diagnose it accurately. Test yourself for HIV and AIDS and then to pet saliva, dairy products and other disease specimens. Stop reinfecting from humans or pets or dairy food. Assist your body by cleaning up your dentalware, and environment. Check all supplements for toxins before using them on a daily basis. (If they show up in your white blood cells a few minutes after eating them, there is a toxin.) If your groin pain doesn't go away, you probably missed something!

There are other causes of groin pain, though, such as **hernia**. A physician will help you identify it. Have it surgically repaired. The Shouldice Hospital[7] has the best surgical record and will disclose their statistics! You won't even need full anesthesia there.

Pain from the **ovaries** is often felt near the groin. Sometimes a large cyst in the ovary puts pressure on the sensitive nearby organs. The formation of cysts, *cystic ovaries,* is a common condition of cats, too. Obviously, something is not right in the ovaries.

---

[7] Shouldice Hospital, 7750 Bayview Ave. Box 370, Thornhill, Ontario, Canada L3T-4A3, telephone (416) 889-1125.

Ovaries make hormones: **progesterone** and **estrogen**. If bacteria (like *Gardnerella, syphilis, gonorrhea*) are living there, the ovaries are handicapped and may under-produce or over-produce some hormone. The first treatment should be to kill parasites, especially the flukes. Then kill all other parasites and bacteria, especially *Gardnerella* and enteric bacteria which can migrate easily from intestine to ovary. In spite of all this, the cyst may not shrink. In order to shrink the cyst you must determine what is filling the cyst. Then stop filling it. The cyst will eliminate itself and shrink.

Many kinds of pollution can fill an ovarian cyst. Test for CFCs and PCBs. Gold is another favorite. Particulate pollutants like asbestos and fiberglass are often put into cysts by your body. Your body is wise. If these can't be eliminated through the kidneys or bile, it will at least keep them all together (in the form of a cyst) the same way a toxic dump spares the rest of the landscape. As long as you are adding to it, it cannot get smaller. Fortunately, your loyal white blood cells are trucking toxins away from your cysts every minute of every day and night. All you need to do is stop adding. Remove dental mental, clean up your body, diet, and environment.

You will notice shrinking of cysts in three weeks. What you actually notice is absence of pain, implying shrinkage. Meanwhile pain relief by killing bacteria, herbal ovarian assistance (try wild yam), chiropractic, pain killers are all useful. Decongest the area by means of a kidney cleanse. If your cyst does not shrink you have missed the pollutants. After it does shrink, stay on a regular program of parasite killing, herbal or electronic, and improved lifestyle.

Ovaries may be giving you pain even without a cyst present. By cleaning up parasites and pollutants you can eliminate it quickly; much more quickly than when a cyst has formed.

Synthetic hormones (Provera,™ Premarin,™ Ogen,™ etc.) are often used to clear up ovary problems of various kinds. Don't continue to use them when the need no longer exists. Although they have low toxicity, there are disadvantages such as the need for liver detoxification, and risk of heavy metal pollutants. Cancer acceleration has been seen when taking estrogens.

Low Back Pain

We have been told that lower back pain originates in an inherited weakness of the spine at its base because we humans walk upright instead of on four legs. And we have been told that the bony hooks that keep the spine aligned are flatter in some families, making it harder for them to hold the spine together. We are also told that “proper exercises” could have kept this part of our bodies strong so lower back pain could be avoided.

These theories become obsolete when, without surgery or exercise or change in posture, lower back pain can be made to disappear quickly and permanently.

Acupressure massage and chiropractic can bring “miraculous” relief. The most severely crippled lower back pain sufferer can shuffle lamely into a chiropractor's office and walk out normally, without pain or painkiller after treatments. So although there has been slippage of disc or spine, apparently it goes back into place rather easily.

If muscle relaxation is the clue, we must ask why these muscles spasm so easily. Any muscle spasms if you irritate or injure it suddenly. In fact, your whole body spasms and flinches if a sliver or bit of broken glass is in your shoe. If you remove these objects, the leg can walk normally.

Oxalic acid crystals are as sharp as broken glass. Use the kidney cleanse (page 549) to dissolve them and other stones. All lower back pain can be cured by removing the sharp crystals in the kidneys. It takes about three weeks to dissolve them. In some very severe cases, it may take six weeks.

Whether you have suffered a year or 20 years, the permanent cure is only weeks away.

Our bodies make eight or more different kinds of kidney "stones." The oxalic acid variety is associated with sharp stabbing pains. In its effort to eliminate this extremely vicious acid your body neutralizes it with calcium first to make calcium oxalate. Your kidneys can keep a bit of calcium oxalate in solution but not a lot. The excess hardens into crystals. A glass of regular or iced tea (not herb tea or green tea) has about 20 mg[8] of oxalic acid—way too much for kidneys to excrete. Tea is a toxic drink, not to be considered a beverage. Chocolate is very high in oxalate, too, and should not be used as a beverage (as cocoa).

Children should never drink tea or cocoa. Their delicate kidneys should not be faced with the daily burden of excreting large amounts of oxalic acid. And calcium used to neutralize oxalic acid is wasted. Calcium is a precious nutrient. It should be conserved for children's bone development.

---

[8] Taken from Food Values 14th ed. by Pennington and Church, 1985.

It isn't necessary to find which variety of kidney crystals are causing your muscle spasms. Different herbs dissolve different kinds. And by combining them into a grand herbal mixture you can be dissolving all varieties at the same time.

Wherever oxalate crystals have formed, a particular bacterium, *Proteus vulgaris*, can be found. Does that bacterium somehow thrive on oxalate crystals? Or even help them form? Does *Proteus* itself contribute to lower back pain? Is lower back pain in reality two pains in one—the sharp jabbing of glass-like particles plus the inflammatory effect of bacteria? Fortunately, you can kill *Proteus vulgaris* electronically.

By using your new diagnostic skills, you can test your kidneys for crystals. The kidney stone varieties I have tested for are: **calcium oxalate, uric acid, cysteine, cystine, monocalcium phosphate, dicalcium phosphate, tricalcium phosphate.** All these varieties can be dissolved by the herbal mixture. But all can be formed again in a week!

To prevent **oxalate** formation stop drinking oxalic acid (eating oxalate rich vegetables is not significant—spinach, chard, rhubarb and sorrel all have their place in the diet). Also take magnesium and B_6 supplements (as directed in the kidney cleanse).

To prevent **phosphate** crystals from forming, reduce phosphate consumption and drink milk as a calcium source. Keep your kidneys squeaky clean with herbs and copious water drinking. After drinking one quart of sterilized milk, two pints of water, one-half glass of homemade fruit juice and one-half glass of vegetable juice, there is little desire for additional beverages.

I have no understanding of what may cause **cysteine** or **cystine** stones (the genetic theory does not explain them either, considering that people without *cystinuria* make these stones).

Since they are sulfur-containing, and I have seen them appear after taking sulfa-drugs, it may be wise to avoid use of sulfa-drugs if you have a choice, or go on the kidney cleanse afterward.

If you have severe lower back pain you probably have several kinds of stones. Some persons have all seven kinds!

Rosie Zakar, age 30, came to see us because her mother was cured of lower back pain so severe she could do no housework for 30 years. Rosie had the usual crystals in her kidneys: oxalate, urate and phosphates. She was started on the kidney herb recipe. In three weeks she was so much better she would have missed her appointment if she had not wanted to cure her digestive problem and fatigue too.

Vera Vigneault, age 32, came mainly for help in getting pregnant but she already had lower back pain and mid-back pain. If she had gotten pregnant before clearing this up, she might have developed eclampsia and high blood pressure which are kidney-related disorders. She was started on kidney herbs for these. She chewed gum a lot and had bleeding gums. She stated her bad teeth were hereditary (meaning other family members had bad teeth also). For this she was instructed to stop chewing gum, start drinking three glasses of 2% milk a day and take a vitamin A&D perle. She was to floss her teeth once a day. (All floss varieties are polluted with mercury and thallium. Use monofilament fish line, 2 to 4 lb. weight.) Immediately after flossing she was to brush them with a new very soft toothbrush with five drops of 17½% food grade hydrogen peroxide. She was to avoid brushing teeth with metal fillings to reduce erosion. She was to brush them a second time without flossing first, this time with five drops of white iodine (potassium iodide) made up by the pharmacist, again avoiding the metal. She had only oxalate kidney stones and was to stop drinking regular tea, replacing it with single-herb teas. In five weeks her gums were better although she was still chewing a little gum and the "peroxy" had been too painful for her to use. Her low back and mid-back pains were gone too.

Gerhard Rogers, age 39, came for his lower back pain and leg cramps. He had mono, di and tricalcium phosphate crystals in his kidneys. His diet was changed to reduce phosphate (meat, cereal, breads, pasta, carbonated beverage) and increase calcium and minerals. He was to drink three glasses of 2% milk a day and to start the kidney herb recipe. In 25 days he had only minor improvement. He still had phosphate crystals. He was afraid to drink milk because he had heard so many bad things about it. He was told to boil it first to eliminate these “bad things”. A toxic element test showed a buildup of copper, arsenic, cobalt, cadmium, lead, thallium, vanadium and radon. This could easily explain his leg cramps, headaches and sleep problem, too. The arsenic came from pesticide, cobalt from detergent, thallium and copper from tooth fillings. The vanadium was fixed by having the gas pipes tightened, and radon could be reduced by improving ventilation under the house. He was thankful for the information and set about cleaning up his body and environment.

Alberta Mellos age 52, came in for lower back pain and upper back pain. It was explained to her that lower back pain was simply due to tiny stones cutting into her tissues but upper back pain was due to gallstones. She could clear her low back pain first. The kidney test showed she had oxalate and cysteine crystals. She was started on the kidney herb recipe. Nineteen days later she arrived with a cold but stated that her low back pain was gone.

Glenn Dirk, age 62, called on the telephone to say his urination had stopped, probably due to kidney stones. This had happened once before and now he was in a panic. He started our kidney herb recipe the same day and passed 117 stones the same night without bleeding or enough pain to need painkiller. After this, he could focus on his prostate enlargement and pain with sitting. He had intestinal flukes and other stages in his prostate gland as well as in his intestine. He also had *Clonorchis* (human liver flukes and their eggs) in his prostate. He had carbon tetrachloride methyl butyl ketone and TC Ethylene from food pollution accumulated there too. After stopping grocery store beverages and killing parasites with a frequency generator, he could urinate normally, freely and without pain.

Lower Abdominal Pain

The lower abdomen on the left side has the sigmoid colon as it comes down and bends. This is a favorite location for larger parasites to settle permanently. Flukes, roundworms, parasites of all kinds and their attendant bacteria and viruses can be felt if they produce gas and pain. Sometimes they live perfectly quietly, seemingly in harmony with us.

Moving the bowel more frequently expels them repeatedly and prevents their numbers from getting very high. Nature may help you with this by setting up diarrhea. Diarrhea is your clue that intestinal freeloaders are present.

The small intestine leads into the colon at your lower abdomen on the right side. At the junction is the ileocaecal valve that prevents backwash, and the appendix. The ascending colon goes up your right side then becomes the transverse colon that crosses your abdomen at the belly button level. The colon descends on your left side, leading into the sigmoid.

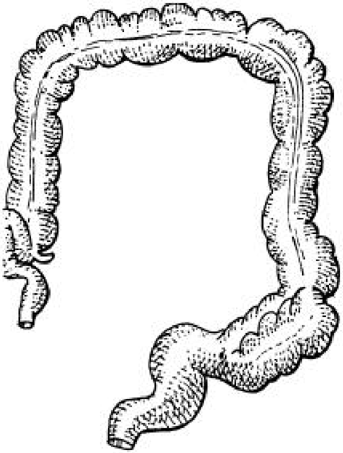

Fig. 19 Colon.

E. coli and *Salmonella* and *Shigella* are “enteric” (they live in your bowel) bacteria that can give you severe abdominal distress and pain. In fact, you can become a chronic sufferer. They can live on hands and under your fingernails, so reinfection from yourself is the most important source. Never, never touch your fingers to your lips. Most importantly, don't try to stop your frequent bowel movements. They will stop on a dime when your parasites and bacteria are dead and gone.

Other sources of *E. coli* are personal water bottles, other people's hands, hands that have changed baby diapers or cleaned bathrooms.

Hands do everything. To eliminate their threat of reinfection, cut out the section on hands (page 397) and paste it on your refrigerator.

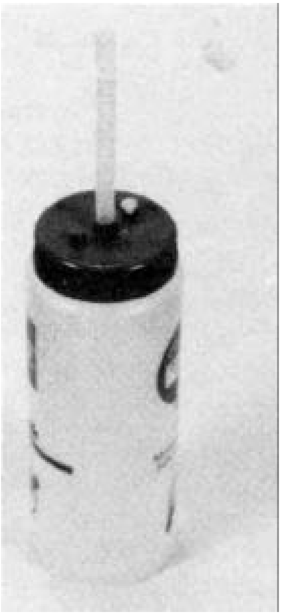

Fig. 20 Keep personal water bottles sterile.

Stomach Pain

Our dairy foods are polluted with *Salmonella* and *Shigella* bacteria. It is impossible to operate a dairy without getting some cow manure into the milk. Although udder wash contains antiseptic it does not kill all manure bacteria. Later, when milk is pasteurized, many heat sensitive bacteria are killed like the “friendly” *streps* and *staphs*, but not all the harmful *Salmonellas* and *Shigellas*. Some survive to colonize the milk, then later infect the consumer. Only milk that is sterilized is safe. A commercial source of sterilized (safe) milk can sometimes be found on the shelf (unrefrigerated). If it had any bacteria, it would not survive shelf life for more than one day!

You may not notice any discomfort from drinking milk, buttermilk, or eating yogurt without sterilizing it. Your stomach acids may be strong enough to kill them, or your liver able to strain them out of your body fluids and dump them, dead, into your bile ducts. Or they may live quietly in some part of the bowel where you feel no effects. Sterilize all your dairy foods by heating at the boiling point for 10 seconds, even if you have no symptoms.

If you are intolerant of milk it doesn't mean that you are lactase deficient, but that you are unable to kill any *Salmonellas* and *Shigellas*. These, in turn, can make you lactase deficient via frequent diarrheas. You can correct this situation by not eating any more bacteria. Give your body clean food, as intended by nature. Don't eat any deli foods, don't eat salad bar food or restaurant food unless it has been baked or cooked. Never eat chicken or red meat at restaurants. They are not thoroughly cooked.

As soon as a new abdominal pain or discomfort, or a gassy condition appears, zap bacteria and try to eliminate your bowel contents. Use the herb, *Cascara sagrada* (follow directions on label) as a laxative, or Epsom salts if necessary. Also start the Bowel Program (page 546).

If you have chronic abdominal problems, make sure you eliminate the bowel contents two or three times a day. There are herbs that can kill enteric bacteria, known to our ancestors of various cultures. **Echinacea** was a treasured herb of American Indians. **Goldenseal** is another favorite. **Turmeric** can kill *E. coli* and some *Shigellas*. **Fennel** can also kill some. But a single dose of Black Walnut Tincture Extra Strength seems to do the most. Make it yourself if possible (see *Recipes*, page 543).

If your body has lost its ability to kill *Salmonellas* and *Shigellas*, all the antibiotics and herbs and good bowel habits cannot protect you from these ubiquitous bacteria. You could ask how you lost your natural protection from them. There is evidence that common **antibiotics** that kill *Streptococcus* and *Staphylococcus* varieties are responsible.

A fraction of your bowel bacteria should be the friendly *Strep. lactis* and *Staph. epidermidis*. After repeated doses of penicillin-like antibiotics (that you may be taking for your strep throat) they are eradicated along with the “bad” *Strep. pneumoniae* bug. No amount of acidophilus culture (which contains active *Lactobacillus*) can replace these *Streps* or *Staphs*.

In mice, it takes a million Salmonella bacteria to start an infection. After giving them streptomycin, it only took 10 bacteria to infect them![9] Your intestines are similarly handicapped after antibiotics, and allow even very small amounts of *Salmonella* and *Shigella* to escape and multiply!

Try to get some natural killing power back. This means improving the stomach's ability to produce acid and the liver's ability to make bile. In turn, this means getting the toxins out of the stomach and cleaning the liver. Certain toxins accumulate in the stomach when the liver and kidneys can't keep up with elimination. Arsenic is a prominent stomach toxin. Get all sources of arsenic removed. Freon is another stomach toxin. Switch to a non-Freon containing refrigerator.

The metals from dentalware: mercury, silver, copper, thallium, first are swallowed and then land in the stomach. Clean up your dentalware. Toxins you inhale such as asbestos, formaldehyde, fiberglass, also are coughed up and swallowed to accumulate in the stomach. Test for them and clean up your environment. Any stomach can recover a significant part of its function by cleaning it up. Even though you regain your tolerance toward minute bits of filth in dairy products, do not go back to unsterilized milk products.

Appendicitis

The lower abdomen on the right side has the valve that separates the small intestine (*ileum*) from the large intestine

---

[9] Sherwood L. Gorbach, M.D., Perturbation of Intestinal Microflora, Vet Human Toxicol 35 (Supplement 1) 1993.

(*caecum*) called the *ileocaecal* valve. It is a common trouble spot because large parasites can attach themselves behind it and keep themselves safe from elimination. (A parasite's biggest worry is your elimination.) It is a favorite location for pinworms in children. It is near this point where the appendix attaches and this, too, is a favorite location of pinworms. With an appendix full of pinworms and their bacteria, is it any wonder when it gets inflamed and causes pain?

Appendicitis refers to inflammation of the appendix. Often, the pain isn't felt over the appendix but over the navel (this is an example of pain being caused at one location but felt at another; it is called *referred pain*). If there are any suspicions of appendicitis, zap pinworms and all enteric parasites and bacteria immediately. Because the current does not penetrate the bowel contents very well, zap every day for two weeks and take 2 tsp. (½tsp. for children) Black Walnut Hull Tincture Extra Strength one time. Make sure bowel movements are regular after this (see the Bowel Program, page 546, for hints) and hands are washed after bathroom use and before eating. Keep fingernails short for the whole family.

If appendicitis does not clear up it can lead to a burst appendix, spewing the dreadful contents into the abdomen. Kill pinworms and roundworms and enteric parasites regularly (once a week) in children. Keep pets on a regular diet of parasite-killing herbs.

Urinary Tract Pain

Urinary tract infections, including bladder, kidney, and urethral infections, are easier to clear up than to test for. Start by drinking a half gallon of water a day. Put yourself on the kidney herb recipe (page 549). This will dissolve the tiny crystals where bacteria hide and multiply.

Common urinary tract pathogens are *Gardnerella*, *Proteus*, *Trichomonas*, *Campylobacters*, *E. coli*, and *Salmonellas*. Sex will give you an instant dose of your partner's urinary tract bacteria. Make sure both of you clean up the urinary tract by zapping and doing the Kidney Cleanse.

Irritable Bowel Syndrome(IBS), Colitis, and Spastic Colon

are all conditions that combine parasite and bacterial problems plus an allergic trigger. If dairy foods trigger yours, you can guess it is not allergic at all but simply *Salmonella* or *Shigella* infection. Boil all dairy foods, stop eating ice cream, cheese and yogurt which you cannot boil. If eating lettuce triggers your intestinal attack, but other roughage does not, it may be a true allergen and cleaning the liver will eventually cure it. Apples, cinnamon and other “allergic” foods can be salvaged the same way. Wheat “allergy” is due to the pancreas being full of pancreatic flukes, wood alcohol, Kojic acid (a mycotoxin), and gold.

All these bowel diseases are quite easily cured by killing all parasites, bacteria, and viruses. Since reinfection is such a big problem, give your pet away until you are completely cured. Have your pet on the herbal parasite program before bringing it back. Clean up your diet, dentalware and environment. Your abdomen will be happy once more and grateful to you for your kind attention. Remember that zapping does not penetrate into the bowel contents. It kills only the outside layer of pathogens. For this reason you should zap daily for several weeks. For this reason, too, I recommend the Bowel Program (page 546) and Black Walnut Hull Tincture Extra Strength even though you may have gotten immediate relief from zapping alone.

Crohn's Disease

is somewhat more serious because the sores are higher up in the digestive tract. This is because sheep liver fluke and pancreatic fluke are commonly the main parasites and these live in the pancreas and liver. They often spill over into the upper intestine. *Salmonella* and *Shigella* are always part of the picture, too, as are various amoebae and fungi. The treatment is the same, kill all parasites and remove all pollutants, especially wood alcohol in commercial beverages. Healing of the digestive tract is very quick, often in a week. Reinfection is very quick too, if the rule about cooking dairy foods is not observed. Keep up meticulous hand sanitation.

Michelle Whorton had stomach pain at the middle of her abdomen, not related to eating. She had occasional very bad diarrhea and also daily headaches. She was started on the kidney herbs for other reasons. Her diarrhea disappeared! We found she had *Ascaris* (probably in her stomach where they cause indigestion and inflammation). She zapped them with a frequency generator set to 408 KHz for three minutes at 10 volts. She was to be very careful with sanitation since they owned a number of farm dogs. Next seen after six weeks, she stated that all her previous problems were gone but she had a different pain in the mid-lower abdomen that got worse during her period and sent pain shooting down both legs.

This turned out to be uterine in origin, but not endometriosis. Her uterus was full of asbestos, arsenic, gold, silver, titanium, propyl alcohol, benzene, styrene, toluene and carbon tetrachloride. This would invite any bacteria toward it!

Mark Lippman, age 51, came in for his irritable bowel syndrome, hoping we would find *Giardia* and put an end to it quickly. Actually, he had intestinal flukes, beginning to invade the liver. He also had propyl alcohol built up in his body giving him a precancerous condition that needed immediate attention. The flukes were killed in twenty minutes, along with *Ascaris* (he had swollen eyelids). He was allergic to milk as could be expected with so much interference with digestion. He was immediately better and did not need to come back.

Billy Henry, a 9-year-old boy, had diarrhea daily and stomach aches. He also wet the bed. There was an indoor pet dog and a bird. Electronic testing showed he had two kinds of *Ascaris* and pinworms. His young body also had a buildup of benzene, moth balls and carbon tetrachloride that he was eating, drinking, and breathing. His bed wetting stopped after killing parasites with the herbal program. His other problems recurred until he was older and could stop licking his fingers when eating.

Tom Ochs, age 36, had chronic stomach problems, alternating constipation and diarrhea, was labeled “lactose intolerant” after an elaborate test, and finally had been diagnosed with irritable bowel syndrome. Actually, he had *Ascaris*, besides other smaller parasites. He was also toxic with cesium from drinking beverages out of clear plastic bottles. This frequently causes depression and he was happy to understand his mood changes. After changing to purer food and products and killing his parasites, he did not need to come back. Five months later he was able to drink all the milk he wanted, no longer had sinus problems and lost his IBS.

Rex Callahan, age 5, had dark circles under his eyes, numerous ear infections until tonsils were removed and tubes put in, and many strep throat infections. Clinically, he was found allergic to dust mites, pollens, and animal dander. His skin got “rashy” if he drank too much fruit juice. He had frequent diarrhea. We found he had *Ascaris* parasites. They were not difficult to clear and he was soon a new person.

David Falls, age 52, had stomach pain and numerous health problems stemming, no doubt, from his diagnosed Crohn's disease. He became allergic to the sulfa drugs commonly used in this disease. We found he had sheep liver flukes and all their developmental stages in his blood and intestine. He was put on the herbal parasite program which he found difficult to follow. Nevertheless, in three months his bowel was nearly normal and the pain in his intestine much less.

Edward Marsili age 7, had bouts of stomach pain. He had intestinal flukes and a build-up of benzene in his body. This would seriously lower his immunity and ability to fight off tiny parasites. He was using a product containing an herbal oil that was polluted with benzene. His parasites were quickly killed with a frequency generator and he was put on the herbal parasite program. One month later his stomach felt much better, but he still had an occasional stomach ache. Testing showed hookworm and rabbit fluke. His benzene was now gone so the tendency to “pick up everything” was gone too. Staying on a child's maintenance parasite program would protect him.

Kim Johnson, almost two years old, had lots of ear infections. It started at eight months so the mother took her off cow's milk and wheat. This stopped her ear infections until mid-winter. She had to be back on antibiotics and a few months ago the doctor began discussing tube implants with her since she was still on antibiotics (six months). Another ear, nose and throat doctor agreed with this opinion, but was willing to wait until Autumn. The baby had been passing a lot of undigested food and was unhappy. They were vegetarians. The baby nursed. Our test showed pancreatic fluke infestation; this would easily lead to bad digestion, especially of milk and gluten in wheat. Fortunately, she was nursed throughout, in spite of going to daycare. Simply killing the parasites (in both mother and baby) solved both problems and she did not need to come back. The ear infections were probably caused by bacteria and viruses brought in by the parasites.

Cynthia Prout, age 36, brought her three children because of their poor health. They all, including herself, had stomach problems, a lot of allergies, asthma, ear infections, and milk intolerance. One boy, age 8, was intolerant of both milk and wheat and hadn't had them for years. He was infested with two kinds of *Ascaris* and pancreatic flukes. His sister, Nola, had itching legs and headaches besides; she was toxic with bismuth and antimony (from shampoo fragrance and laundry fragrance). She also showed a build up of vanadium, implying a gas leak in the home. The youngest, age 5, had frequent stomach aches and vomiting. It was a simple matter to kill *Ascaris* electronically at 408 KHz and the pancreatic fluke with all its stages (421 through 434) after which the children did fine.

Sofia Sobel had extreme ulcerative colitis although she came for her headache. She had been on Prednisone for a month with no relief. We found she had the three large flukes plus *Chilomastix*, dog whip worm, and amoebas in her intestine (but not in body organs).

There were several house dogs. Her stomach and intestines were much too sensitive to accept parasite herbs, or in fact, anything—anything except slippery elm powder. This herb (1 tbs. made into a paste first with water and then drunk as a beverage three times a day) paved the way for acceptance of two oyster shell calcium, one magnesium and one zinc tablet. Her blood test showed high phosphate levels since she was dissolving her bones to get calcium. She added sodium alginate (¼ tsp. to a cup of vegetable broth soup) twice a day to help her tolerate the parasite killing herbs. Her children were given VMF (vermifuge or parasite killing) syrup. By the 12th day of the parasite program she no longer needed colitis medicine; her bowel movements were down to twice a day, soft and formed, but still with a little blood streaking. She loved the alginate mixed with slippery elm. She was able to eat fruits and vegetables but agreed to stay off wheat and corn until her liver was cleansed. In another week she was free from all abdominal complaints except a heaviness over the uterus, possibly due to two missed periods. The thought of pregnancy put her in a panic. She was instructed to induce her period (Emmenagogue, see *Recipes*). She was sure she wanted her period, not a pregnancy and this seemed to be her God-given right. Three weeks later she had a flare up of colitis due to *Salmonella* in food; it also gave her a urinary tract infection. This time she took Quassia herb to kill invaders in addition to the maintenance parasite program which she had begun to neglect. She had been very busy, had lots of energy and wasn't on medications. She treated her urinary tract infection with betaine-hydrochloride (to acidify the stomach), began using plastic utensils to reduce her nickel intake (see Prostate Pain, page 124)) and drank a lot of water. This experience taught her valuable lessons that she was eager to learn, benefiting her family and herself immensely.

Rebecca Goetz, age 53, had ulcerative colitis and her husband had Crohn's disease. She had been on Azulfidine™ and Flagyl™ frequently. Her parasites were only intestinal flukes and their stages, and *Endolimax*, an amoeba. It was a simple task for her to clear her problems by killing them and by sterilizing her dairy foods. Wes, her husband, had three surgeries to remove sections of bowel due to Crohn's disease. He now had a colostomy but was on Advil™ for pain in the rectal area. He could hardly sit. He had been tried on anti parasitic medicine (Cypril™ and Flagyl™) but they did no good. He had intestinal flukes and all their reproductive stages in his body, also pancreatic flukes, *Capillaria* roundworm, and *Diphyllobothrium erinacea* scolex. We interrupted his testing at that point. His kidneys were full of phosphate crystals— he ate no dairy products. He was started on half-doses of kidney herbs and only part of the parasite program in view of his colostomy and possible diarrhea. Two weeks later we continued testing, finding pinworms, *Haemonchus*, *Leishmania tropica*, *Paragonimus*, *Sarcocystis*, *Stephanuris* and *Trichuris* (whip worm.) Quassia was added and doses increased. His blood test showed a high thyroid hormone level (T4), contributory to over activity of his bowel He was started on goat milk, vitamin C (3 gm. daily) and B_{12} shots. He was given magnets to sit on for pain. He was toxic with cadmium, from his old tooth fillings. But in five weeks he could sit comfortably without pain pills. There was less blood in the stool. Dental work would bring him his next big improvement.

Benito Villamar, a middle age man, had severe side pain for several weeks. He was also gassy. He had sheep liver fluke and stages in his thymus and intestine. The thymus is under the top of the breastbone and is a very important organ of immune function. It is easily damaged by benzene. He did, indeed, have benzene accumulated there. He was given a list of benzene-polluted products to avoid and was started on the parasite killing herbs after killing the flukes instantly with the frequency generator. Two weeks later his side was very much better, his benzene was gone and he was eager to rid himself of lower back pain, which he also had.

Al Vickers, age 9, had stomach aches, headaches, a constantly runny nose and asthma. He was on Slo-Bid™ medicine and allergy treatment. He had a sleep problem. He also had two dogs, one rat and two hamsters. The dogs and he had high levels of *Ascaris*. He was zapped for *Ascaris* and the four common flukes (without testing). He was put on vermifuge syrup and Rascal capsules. This ended his problems and began a new chapter of better care for his health by his parents.

Tim Melton, age 16, had several colitis attacks yearly, requiring hospitalization, from third grade to the present. He also drooled constantly, needing to spit a lot. (This is due to mercury toxicity from amalgam fillings. It is better to spit out the mercury than to swallow it.) He had intestinal flukes at a high level for which he was started on the parasite program. One month later he was very much better. He had only one diarrhea session since the last visit. But he still had sharp pains under both buttocks (probably due to kidney stones). He had been an iced tea drinker and had numerous oxalate and cysteine crystals deposited. He was appalled that a common beverage could be so harmful.

Central Abdominal Pain

can be coming from the uterus, bladder, or bowel. It is difficult to tell which is the source. The first step is to simply kill enteric (bowel) free-loaders and get into good bowel habits. Gas and bloating should be gone. If this isn't the solution to the pain there may be special bladder parasites with their bacteria. *Schistosomes* prefer to invade the bladder wall. In fact, very many parasites temporarily invade the bladder because the body is trying to excrete many of them. The whole family should be cleared of these same parasites. Kill them by zapping. Pets should not be kept indoors since they have many of these parasites, too, and they are easily transmitted to us.

Interstitial Cystitis

is one of the most painful conditions described by clients. *Schistosomes* are the real perpetrators but after the bladder wall is weakened, other parasites and their bacteria and viruses accumulate here too. To regain your bladder's health all toxins must be cleared as well. Dental metal, environmental toxins, including radon, asbestos, formaldehyde, must be cleaned up. The diet, body products and home should be carefully searched for toxins. *Schistosomes* are easily zapped but easily picked up off toilet seats and doorknobs. Always wash hands after toileting: a single droplet reinfects you!

Uterine Pain

Endometriosis

Many a woman's dreams have been shattered by her inability to have a child. Endometriosis is often the cause. It starts with painful cramps at period times. They get worse and worse until pain killers are necessary just to get out of bed and move about the house. There are flukes in the uterus! Large intestinal flukes in a rather small organ! Did they migrate to the uterus from the intestine or did they develop there from eggs?

Sometimes sheep liver fluke is seen there. Once an avenue to the uterus is established, numerous other parasites move in the same direction: *Clonorchis*, the human liver fluke and even *Eurytrema*, the pancreatic fluke, can invade the uterus wall. Why have they taken up so abnormal a living place? Because the uterus has solvents in it! This is the green light for flukes. This disarms your organs so they are left helpless against fluke stages left there by the blood and lymph. Stop eating solvent-polluted foods. There are solvents in all cold cereals. Make cooked cereal. There are solvents in purchased drinking water. Drink from your cold kitchen faucet. There are solvents in grocery store bread, grocery baked goods and cholesterol-reduced foods. Eat none of these. Buy baked goods and bread at your local bakery. Stay away from “low cholesterol” foods. There are solvents in decaffeinated and other powdered mixes for beverages. Drink nothing except milk from the grocery store (sterilize it). Milk does not have solvent pollution. The hormones, antibiotics and udder wash can be tolerated—solvents cannot.

Beverages and powdered mixtures sold at health food stores are no exception. Use no powdered mixture intended for weight loss or weight gain, nor vitality supports, nor dietary supplements. They are all polluted. Some solvents (I often see methyl ethyl ketone and methyl butyl ketone) choose the uterus to accumulate in. This sets the stage for endometriosis and fertility problems. Where there are large parasites, smaller ones soon crowd in. All bring their own bacteria and viruses. *Gardnerella*, especially, is found in cases of endometriosis, ovarian cysts and menstrual problems. The flukes evidently travel from the uterus to other parts of your body cavity, distributing bits of the uterine lining as they go. Once this distribution has occurred, can the bleeding (regular menstrual bleeding) at these extra sites ever be stopped?

It stops immediately when the flukes are dead!

Your body knows how to clean up after dead flukes and does the job perfectly. You can be free of pain in time for your next period. Zap to kill the four common flukes, *Gardnerella*, all other common parasites, and urinary tract bacteria (common ones include *Proteus*, *Salmonella*, *Campylobacter*, *Chlamydia*, *Trichomonas*). Avoid reinfection by avoiding solvents! It is impossible not to pick up parasites. If your uterus has solvent in it, they will find their way to it in a day. Without solvent, they will not.

The solvents will leave by themselves. Help your body get rid of them with vitamin C and B_2 (3 grams and 300 mg. respectively, daily, see *Sources*).

To heal the uterus so it no longer attracts parasites, clear up its internal pollution besides solvents. This means mainly the dental metal that has piled up and environmental toxins such as asbestos, arsenic, fiberglass, and formaldehyde. Gold and silver are especially attracted to the uterus. Don't wear gold rings or any metal jewelry touching your skin anymore, and, of course, get all metal out of your teeth. Never try to get pregnant before you have cleared up endometriosis.

The advice given by obstetricians to get pregnant to solve your pain problem is most unwise. Indeed, pregnancy changes your body's metabolism and without periods you get pain relief. But it seems much too risky to grow a baby in an infested, polluted uterus. Fear of birth defects is an intelligent fear. Be careful not to get pregnant while you are killing parasites and getting mercury removed from your teeth.

Healing starts as soon as all the parasites and pollutants are gone.

Joanne Biro, age 22, had severe cramping pain with her periods, diagnosed as endometriosis. She had adult intestinal flukes and the *cercaria* stage in her uterus. She had a xylene (solvent) buildup in both her brain (cerebrum and cerebellum) and uterus. She was started on the herbal parasite program following the kidney cleanse. Her next period was pain free. A check up showed she was free of flukes but had thallium in her immune system. Dental cleanup was next on her agenda.

Denise Leyva, 22, was on birth control pills to control the growth of endometrial tissue. She had laser surgery previously. She had hexanedione and methyl butyl ketone buildup in her uterus supporting the intestinal fluke and its eggs in the uterus. There were also some sheep liver flukes and human liver fluke stages there! She was advised to stop eating cold cereals and commercial beverages and kill the parasites immediately. She had no recurrence.

Anita Pierce, age 32, had numerous surgeries for her endometriosis. She also had chronic fatigue syndrome, and several allergies. She had two beautiful poodles in her home. She could not part with them so she gave them the pet parasite herbs faithfully. She had intestinal flukes, tapeworm stages, *Ascaris* and various other flukes in the uterus. She had all her dental metal removed and home detoxified. Her body was teeming with *Nocardia* and Epstein Barre Virus (EBV). In spite of repeatedly killing the flukes and bacteria with a frequency generator and making herculean efforts she was no better off eight months later. She was unable to solve the problem of reinfection from her dogs.

Christine Solton, age 27, had extremely heavy and painful periods and didn't stop bleeding between periods. A large cyst had been seen by X-ray in the uterine wall. She also had constant bladder pain. Both problems kept her in the bathroom most of the time (90 visits/day). She had the intestinal fluke in her uterus (probable cause of cyst) and *Schistosoma haematobium* (bladder parasite) throughout her body. She was started on the parasite program and in one week her bladder pain was under control but bleeding (from the cyst in uterine wall) continued. *Schistosomes* are very contagious, probably even from toilet seats and the house dust of an infected person. She had them again three weeks later. This time she zapped them and got instant relief. Her bladder and uterus were both full of propyl alcohol, tooth metal, fluoride, cobalt, zirconium, aluminum, antimony, cadmium, and formaldehyde. She was delighted, though, to understand her problem and made the dental appointment.

Contraception

There is an excellent pamphlet available at health food stores, called Wild Yam for Birth Control Without Fear[10] that informs that 3 capsules taken two times a day provides reliable (perfect) contraception provided you give it a two month head start. Also, an emmenagogue recipe is on page 546.

The Silent Cervix

The cervix is a big “trouble spot” for women just as the prostate is for men. It seldom lets you know with pain, however, that bacteria or parasites or toxins are present. Sometimes a brief needle-like pain does alert you to something going on there, but it is easy to miss. The cervix is constantly secreting a little bit of mucous and this helps it stay clean but why give it mercury and copper and gold to secrete? Many a fertility problem has been solved by stopping the toxic pollution of uterus, ovaries, and cervix. Clean up your dentalware and clean up your diet and environment.

---

[10] Willa Shaffer, published by Woodland Health Books, PO Box 1422, Provo, Utah 84603.

Kill parasites and bacteria regularly, every week, with the herbal recipe or by zapping. You should have no pain with menstruation, no bloating, fatigue or headaches. A PAP smear test should always be “good”. If it isn't, hurry to the rescue of your cervix.

Menopausal Symptoms, Hot Flushes, PMS

Insomnia, irritability, PMS (pre-menstrual syndrome), depression, anxiety, nervousness, are all not to be expected at and after menopause. They may certainly be caused by hormone imbalances. It is these imbalances that are not normal.

NO menopausal symptoms are normal

After the ovaries are done with their cycles of estrogen and progesterone production, the **adrenal glands'** hormone production was meant to “kick in” and make up any deficit.

During your fertile years, you were meant to have a peak of 100 picograms/milliliter (pg./ml) of estrogen on day 9 and day 22 of your cycle. Progesterone, on the other hand, only peaks once, on day 22, and it should reach a level 20 to 100 times as high as estrogen! After this ends, your adrenals can still keep your hormone levels regulated. Typical values are 20 pg./ml estrogen and, again, 20 to 100 times as much progesterone. Keeping these two hormones in balance is just as important as the actual amounts. **20 pg./ml estrogen is enough to prevent menopausal symptoms** including hot flushes, and to give you heart protection and bone density protection. Taking synthetic hormones usually gives you exactly such levels. But if your own adrenal glands can supply them, surely it is a better approach. (A blood test can tell you your levels; do it on day 21, 22 or 23 before menopause, after menopause the day is not critical.)

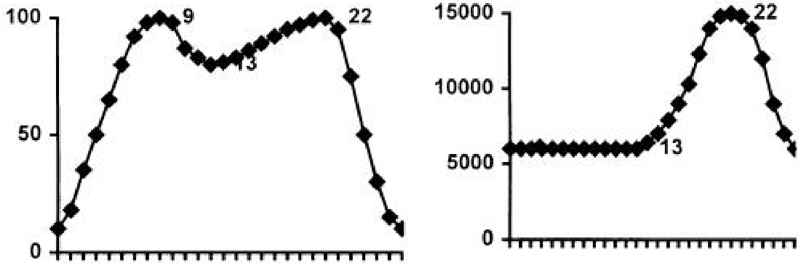

daily estrogen levels

daily progesteron levels

Fig. 21 Estrogen and progesterone levels (pg./ml).

Why aren't your adrenals producing them? Because they are hampered by parasites and pollutants! Kill all the parasites, bacteria and viruses, especially *Gardnerella*, *Proteus*, *Chlamydia*, *Campylobacter*, *Neisseria*, *Treponema*, *Salmonella*. Use the zapper.

Notice that these are also the favorite urinary tract bacteria! This makes good sense, because the adrenal glands sit right on the kidneys and would be geographically close to the kidney bacteria. To avoid getting them back, do a kidney cleanse (page 549) to remove all crystals where they might hide.

After this, hot flushes can be suddenly gone. If not, continue the cleanse. Meanwhile, do some permanent lifestyle improvement. Start drinking two pints of water between meals plus water and milk (sterilized) at mealtime.

Don't drink cranberry juice with its copious *hippuric acid* and its solvent pollutants. You won't need such insufficient help after cleaning up your body, dentalware, and environment. Besides giving you a better hormone supply, your newly revitalized adrenals will get you through stress in better shape and keep your blood pressure normal.

If bacteria are not kept down they will begin to invade other organs. *Gardnerella* goes to ovaries to feast on the corpus luteum after ovulation. This keeps the corpus luteum from making enough progesterone and PMS results.

Menstrual problems and PMS, in general, although they may not be pain-causers, are a sign that all is not well with the uterus and ovaries. Go on a cleanup program. Kill your invaders with herbs or the zapper. Clear up your toxic accumulations from dental metal and environmental sources.

Don't be surprised by a pregnancy! This is not the time to get pregnant, though. If you wish to get pregnant, clean up your body first, being very careful to prevent pregnancy during this time. If this is not under your control do your dental work first. Since every cleanup job increases your fertility, it is best to get the mercury, thallium, copper and nickel out of your body before your risk of conception is raised further by making other improvements. Many an amalgam replacement job had to be halted in the middle due to pregnancy! Couples just couldn't believe they could have a child as a result of cleaning up their bodies so they were careless in spite of my warnings! **A dentist will not take out mercury during pregnancy.**

Nicole Truett, 40ish, had PMS before her periods. She was a returning student and couldn't afford to feel emotionally upset. She also had recurrent yeast infections, *Herpes* and panic attacks. She complained that her thinking was weird, although she was an excellent student (she got her birth date wrong on our office forms!) She had high levels of mercury and we suggested dental work immediately. Two months later she was feeling much better and had all metal removed but was experiencing a slight return of symptoms which panicked her. She still had *E. coli* and *Bacillus anthracus* at four tooth locations, giving her chronic sinus symptoms. She needed her cavitations cleaned. Four weeks later she described how she had gotten immediate emotional relief after two cavitations were done. Our solvent test showed methyl butyl ketone, benzene and carbon tetrachloride (which we found in her Mountain Valley Arkansas Spring Water). They were lowering her immunity giving recurrent Herpes and yeast problems. These disappeared in a week. She was so happy to find the cause of her problems and yet so angry at the nature of the cause that she planned to write to the water company. We need more such environmental activists.

Monica Koziol was on Ibuprofen™ for menstrual cramps. She also got a severe sugar craving and minor depression and headache with her period. She was toxic with silver, copper, platinum, mercury and lead. She also had hookworms, pinworms, human liver fluke and cat liver flukes infesting her. A hormone test showed very low estrogen (57.6 pg/ml) probably due to all these interfering factors. She had all the metal from her fillings replaced and killed parasites with the herbal recipe. This gave her regular normal periods for several months. She decided to get pregnant but couldn't. A follow-up showed she was full of parasites again. She had stopped the maintenance parasite program. She resumed it and began kidney and liver cleanses to get longer lasting benefits.

Barbara Ashby, age 43, had suffered for 1½ years from menstrual pain. She had oxalate crystals in her kidneys and was started on the kidney cleanse. She followed this with the parasite program and dental metal replacement. Then she cleaned her liver and after three cleanses (she got over 1,000 stones the first time!) she said she felt great again.

Terri Entzminger, age 16, had a long list of health problems including painful ovaries and painful periods for which she was put on birth control pills by her doctor. She had several bacteria in her genital tract: *Neisseria gonorrhea*, *Plasmodium cyano*, *Staphylococcus* *aureus* (also at tooth #28 which had a plastic filling), *Streptococcus pyogenes* (also at tooth #28). A parasite test showed intestinal fluke adults in the uterus, not in the intestine or liver. There were also fluke *cercaria* here. She was started on the parasite herbs. Three weeks later there was no improvement. She had a rabbit, a dog and a hamster. She was to get them all onto the pet parasite program. Her diet was changed to exclude solvents. Six weeks later her periods were "great", she did not need the pill and she was keeping the whole household on a maintenance program killing parasites. Three weeks later she had cramps again. This time it was sheep liver fluke in the uterus. She probably got them from the snails in her fish tanks. She was off the maintenance program and drinking caffeine free cola again. This would fill her with solvents that make fluke disease possible. She killed them all with a frequency generator and decided to be more vigilant over parasites as long as she was such an ardent animal lover.

Azar Moya, 57, was on Premarin™ and Provera™ for hot flushes and emotional extremes, Synthroid™ for the thyroid, Xanax™ for nerves and sleep problems, something for diarrhea and something for depression. In five months she needed none of it. She had done a liver cleanse by then and got a commode-full of stones (about 1,000), she had changed her plumbing, got rid of the water softener, killed parasites and cleansed her kidneys. She still had sinus problems and some arthritis and was planning dental metal replacement and cavitation cleaning to clear them up too.

Infertility

An ominous sign in any species, infertility is not just another disease or "problem."

When birds' eggs don't hatch their species is doomed. So we learned from the DDT experiment humans did with birds in the 1960's. The DDT changed the thickness of their egg shells so they cracked when the mother bird sat on them. All changes are experiments whether intentional or unintentional. Nature by itself produces sweeping changes, too, such as droughts, wind storms, fires, ice ages, but usually living things had time or mobility to adapt to them. When there wasn't time, and they couldn't run away, the species went extinct.

Are we humans an exception to this rule of adapt or perish? Nobody feels more helpless and hopeless than the infertile couple. They can not run away, time is limited, and obviously adaptation is not occurring. More likely their lineage will perish. The couples' only wish is “Give us one child, now.” Surely, it is their birthright, as it is any living creatures', to reproduce.

Can we relax with the assurance that our intelligence, through the arm of science, will always rescue us? Are test tube fertilizations, fertility drugs, Cesarean sections, incubators for premature babies all triumphs for science? No, they are signs of reproductive failure for the human species.

When the concern is overpopulation of this planet, reproductive failure might seem less ominous. Maybe it's no worse than the natural way any species curbs its growth rate. Maybe only those who can survive parasitism, pollution and immune deficiency should survive in order to strengthen the species. But when reproductive intervention becomes a necessity, not an option, surely the danger signal is present as it was for the DDT'd birds who saw cracks develop in their eggs. The solution to our reproductive failure is not to find ever more artificial ways to conceive, to give birth, and to care for damaged babies. The solution is to fix the old fashioned way; to safeguard the natural way.

If you are unable to conceive or to provide viable sperms use an intelligent approach. Remove the obstacles. **The obstacles are parasites and pollutants, the same enemies of health we have seen before.**

Kill all large and small parasites with a zapper and the herbal parasite killing program. Don't try to keep a pet parasite free, give it away. Living close to another species is a luxury you can't afford at present. The pet can live with its parasites, you can't. Remember to kill bacteria and viruses too, especially *Gardnerella, Neisseria, Treponema,* the ancient enemies of human reproduction.

Is it safe to kill parasites if you might be pregnant?

The electronic way of killing parasites is safe if you use a frequency generator. The frequencies of parasites and bacteria are far away from human frequencies. The treatment with each frequency is short. There are no side effects.

The zapper has not been tested and should not be used during pregnancy.

The herbal way of killing parasites has been used by pregnant women without bad effects but this is not enough safeguard. I recommend waiting until the baby is born if at all possible. The treatment is long and intense. The growing baby is exposed continuously to herbs. Perhaps this is preferable to the toxins produced by parasites. You must use your own judgment. Obviously it is wiser to take a chance on herbs than to take a chance on inheriting AIDS or “genetic” diseases.

Part two of regaining your reproductive freedom to have a child is removing pollutants. Gold, silver, copper and mercury can accumulate in the reproductive organs, wrecking the delicate hormone balance between estrogen and progesterone, or wrecking the motility of sperm. Research has not been done to search for dental metal in the uterus, ovaries and testicle of infertile couples. You can do this research yourself. Slides of ovary tissue cost less than $10.00 as do other parts of the reproductive system. Search for dental metal yourself. Remove all dental metal from your mouth, and replace it with metal-free composite. Extract teeth with root canals.

A Word Of Warning!

Be extra careful with contraception during the dental cleanup. You could get pregnant the very next day! **This is no joke.** It is a serious hazard to conceive a child while mercury is loose and rampant in your body from the removal process. It may be a higher risk than leaving it untouched. If you are pregnant no dentist will want to finish the job of mercury removal! **Don't try to get pregnant yet.**

You may have tried fertility pills, in vitro fertilization, and other methods for getting pregnant over a ten year time period, all to no avail. Then you start cleaning up your body and taking your mercury out and suddenly you are pregnant before the job is complete! It may seem unreasonable and illogical to have to be careful after ten years of no worries, but play it safe.

If you fail to observe this warning and do get pregnant too soon, you may pray for miscarriage. Otherwise, take vitamin C and thioctic acid and hope for the best. Men should add daily zinc and arginine (60 mg. and 450 mg, respectively) to their diets. Both men and women should add vitamin E (200 mg.), a prenatal multivitamin and multimineral tablet, eat freshly grown vegetables for folic acid, and add vitamin C (at least a gram daily, see *Sources*). **No other supplements**! Supplements polluted with heavy metals or solvents do more harm than good. If you are not sure of their purity, test one by eating it and searching for it in your immune system five minutes later. If it is there, it is harmful; eat no more.

Nausea

of pregnancy is the scourge of expectant parents. After waiting hard and long for the desired pregnancy, the mother-to-be feels rotten, salivates and gags at the thought of food, and wants no more sex. Maybe sex is ill-advised during pregnancy, no matter how reassuring the male or male-oriented obstetrician is! Maybe salivation is actually mercury excretion being attempted by the body. Maybe nausea is all about keeping toxins out of the body and away from the developing child. These are intriguing possibilities, worthy of your research expertise.

A few decades ago the treatment for nausea was a weekly B_6 and B_{12} shot. Ask your obstetrician for this to see if it helps.

An older, herbal remedy was cinnamon tea: 2 tbs. cinnamon (bark or powder) in 2½cups boiling water, steeped for 10 minutes. Strain and add honey to taste. Dose: ¼ cup three times a day before meals.

Nausea invites starch eating—pasta, potatoes, rice and bread. Starches can absorb. Perhaps they absorb the noxious substances causing nausea. Make sure you add vitamin C to grains. In any case you must still eat additional nutritious food to grow your baby. In spite of craving a pickles/chocolate pudding/carbonated beverage lifestyle, you must eat mainly good food. Craving can take strange turns. Search for the taste you crave in good food and in long forgotten childhood foods.

These are all the fertility cases I saw in a year's time. None were left out in order to hide failure. Assess the success rate yourself:

Domilita Renshaw and her husband had been trying for six years to get pregnant. Both had been tested and treated in assorted ways. Domilita's period was irregular, a sure sign that all is not well in the area of reproduction. I gave them the usual warning about not risking pregnancy during their deparasitizing and depolluting procedures they both would be going through. Her hormone test showed slightly high (125 pg/ml) estrogen levels for day 22 (if it really was day 22!) and higher still if it was not yet day 22! Obviously, something was irritating the ovaries into overproduction of estrogen. She had oxalate and urate crystals and was put on the kidney herb recipe. She was switched to milk (3 glasses 2% a day) as her primary beverage besides water. She was toxic with nickel (dental metal) which would invite hordes of urinary tract bacteria, dangerously close to the ovaries. She made her dental appointment. She had sheep liver flukes and was started on the parasite program. She broke out in hives from a new hair spray polluted with praseodymium which got into her ovaries. She prepared to clean her liver for her frequent hives. Then she called to cancel her next appointment because she was pregnant (four months from first visit). Fortunately she had one visit with dentist completed. Nine months later she had a beautiful perfect baby.

Lindy Maloy and her husband had been trying for eight years to have their second child. They all had *Ascaris*, including, of course, their house dog. They wormed the dog monthly and did not want to part with it since they did not believe it mattered. They used the pet parasite program, but five months later she had higher *Ascaris* loads than ever. She also could not rid her uterus of intestinal fluke stages in spite of killing them with a frequency generator and using the parasite herbs. She remained full of solvents, bacteria and platinum from dentalware. Her endometriosis continued. They gave up.

Rosemary Peterson, age 33, had been trying to get pregnant for fourteen years. She had seven laparoscopies for endometriosis and very hard cramps with her period. She had intestinal flukes and sheep liver flukes in her uterus. There were sheep liver flukes and human liver flukes in her liver. There were intestinal fluke redia and cercaria in her saliva. The solvents in her uterus were methyl butyl ketone, acetone, carbon tetrachloride (from drinking store bought water), styrene (from drinking out of styrofoam cups), xylene (from carbonated beverages) and decane (from cholesterol-reduced foods). She also had a chronic yeast infection, treating it constantly with Nystatin^{TM}. She killed the flukes and yeast electronically before leaving the office and started herself on the parasite program and diet restriction. She got pregnant immediately, and did not return.

Elisabeth Tran, age 37, had tried to get pregnant for five years. She did get pregnant recently, on a special “gift” surgery program, but lost it. Her ovaries and uterus were toxic with mercury and thallium from polluted dental alloy. She also had barium and titanium in them, probably from lipstick. We did not see her again, we hope she solved these problems.

Christopher Gravely, a young man of 26 and Frederica, 22, promised faithfully not to get pregnant until their cleanup was complete. He was found by his doctor to have slow moving (low motility) sperm. He was robust and healthy looking but suffered a lot from low back pain—a clue to swarms of bacteria in the lower abdomen. An electronic search of his testicles and prostate (which had been infected once) revealed iridium, platinum and yttrium. This implicated tooth metal. He was also started on kidney herbs. Eight months later he had completed all his tasks, his low back and pain with urination had stopped, and this encouraged him to continue with his fertility program. We started him on thioctic acid two a day and zinc (60 mg.) two a day, switched him to an electric razor so he wouldn't have to use any chemicals on his face, and recommended that he ventilate his garage which was attached to the house, to reduce fumes in his home.

Meanwhile, Frederica, his wife, was also checked for toxic elements. She had antimony (from mascara) in her ovaries and breasts. She had sensitive painful breasts during her periods which were quite irregular. She also had indium and gallium, dental alloys in the ovaries and breasts. She, too, was started on the kidney herbs and instructed to get metal tooth fillings replaced. After two months they canceled their appointment. Frederica was pregnant! Not for long, though, and a wiser couple returned a few months later. Frederica finished her dental work. Both started the parasite program. Frederica's periods became regular. She was started on thioctic acid (one a day) plus zinc, (one a day), until her first missed period. Twelve months later they sent me their baby's picture: he was two months old.

Ginger Hart had been trying for three years to get pregnant. After an endometrial biopsy, a D&C,[11] and laparoscopy she was diagnosed with "inadequate corpus luteum." We found her ovaries toxic with nickel and europium from tooth fillings and strontium from toothpaste. She was delighted to understand her problem and set about correcting it.

Marjory Davis, age 28, had been on the "pill" (synthetic hormones) a long time but was off now and couldn't get pregnant. She actually got pregnant about one year ago but lost it at one month. A toxic element test showed her ovaries and uterus were full of beryllium (gasoline and coal oil), gadolinium and gallium. The metals are alloys of gold used in dentistry. She wore a lot of jewelry, just loved her chains, necklaces, rings, etc. But she agreed to go off all except two rings which did not have these alloys. To reduce fossil fuel fumes in the house she removed all gas cans and the lawn mower from the attached garage. They parked the cars outside. She was started on kidney herbs and promised to use contraception until she was done. She was to drink three glasses of 2% milk a day and take a magnesium tablet and stop drinking other beverages. Three weeks later her husband canceled her appointment because she was too embarrassed and delighted to call herself.

Prostate Problems

If urination is not complete, so you must soon go again, especially in the night, it is suggestive of pressure on the urethra from an enlarged prostate gland. Keeping a little urine from being voided is conducive to bladder and kidney infection, too, because bacteria soon find this "free food."

The prostate collects toxins as if it were a designated dump site, especially for nickel. Urinary tract bacteria quickly find "their" metal, nickel. Any supply of nickel will attract bacteria as quickly as crumbs and cheese attract mice. Urinary tract bacteria are making use of your urea as their food. To digest it, they first break the urea molecule apart into two ammonia molecules. It is the ammonia smell of the urine that gives away their presence. The urine should have no ammonia smell. Our bodies do not make ammonia. Only bacteria can do this! Their digestive enzyme is called *urease*. In order to do its job, this enzyme requires the element nickel.

---

[11] A surgical procedure, called *dilation and curettage*, meaning dilate the bladder with air and scrape away the inner lining.

Nickel is plentiful in the soil which is undoubtedly where these bacteria belong, breaking up and utilizing the urine and droppings left there by animals. They perform an essential task in our environment, destroying animal excrement and thereby cleaning-up the soil around us. What folly it is to load ourselves up with nickel so that in one short hop from the earth they are residing in us! Bacteria are all around us doing their valuable jobs. We cannot stop associating with bacteria. We were not meant to feed them, however. If we did not supply them with nickel, as if we were just another patch of earth, they could not gain a foothold in our urinary tract and then to the prostate.

How can we rid our bodies of nickel? This is the challenge. Are we eating dirt? A small amount of dirt can't be avoided on our food. But we can stop sucking on nickel as if it were a lollipop. Our spoons, forks and knives are made of nickel! Stainless steel is 8% nickel! Does it really come off as you eat? When you stick a knife into the mayonnaise jar, it is stained in a few minutes. Try it. The mayonnaise has reacted with the metal loosening the nickel. When you smear the knife over your bread, this film comes off and you eat it. If you were to put the knife in your mouth, now, you could taste the metal. You will later wash the knife but not before you have eaten enough nickel to supply all the bacteria in your body with the daily allowance of their essential element, nickel. Nickel is not our essential mineral. Even plants keep their nickel levels very low. But due to pollution of animal feed with it, even a hen's egg now has 3 mcg/100 gm of nickel in it.[12] To produce a nickel-polluted egg, the chicken must be polluted.

Especially infants and children should never be given metal cups or cutlery. They need all their immune power to combat the barrage of new bacteria and viruses that is emerging in this age of AIDS. Nickel is part of dirt and belongs there, not in our pots and cutlery.

Another large source of nickel is metal dentalware. It is used to harden gold! If you suffer from prostate problems, remove metal from your dentalware. Nickel is used to make bridges, gold crowns, retainers, amalgams. Exchange it all for plastic ware and composite buildups (see *Sources* for more dental information). Stop eating and cooking with metal utensils; use old fashioned wooden or sturdy plastic cutlery instead. Always use a plastic knife for butter or mayonnaise. Never buy foil-wrapped butter.

Nickel is fat soluble and is stored in your skin fat temporarily when a surge of it enters the body. Your skin oils may be loaded with nickel causing “allergies” in the skin. Male pattern baldness is such an allergy. The sweat tries to excrete it for you. Always wash off your scalp sweat to help with this excretion. The skin oils dissolve nickel from metal jewelry (sometimes leaving your skin with a greenish black color) and transport it into your body. Don't wear metal jewelry. Earring posts should be plastic. Metal watches and metal rimmed glasses should be replaced with all plastic types. Metal rings should be replaced. After lowering your total body nickel levels and your prostate disease is only a memory, you might notice scalp hair returning to sparse areas. Search around the edge of the hair line for the first returnees.

Bacteria cannot live in the prostate without nickel being present. You can cure your prostate problems with the simple tactic of stopping nickel pollution of your body. Notice that you get a fresh attack after accidentally using metal cutlery in a restaurant or eating mayonnaise-style salad with a metal spoon stuck in it. It might be wise to take a **histidine** capsule (about 500 mg., one a day for three weeks, see *Sources*) when this happens. Histidine is a nickel chelator. Taking **zinc** is helpful too (60 mg. two times a day for a month). Possibly, the nickel was poisoning zinc enzymes. Read about the benefits of **flaxseed**, too, but remember to test every product for pollutants before accepting it as a supplement.

---

[12] Food Values by Pennington and Church, 1985

Prostate problems of all kinds clear up when bacteria are zapped, the Kidney Cleanse is done, dental cleanup is done, and the Bowel Program is followed.

Richard Traylor, age 71, had suffered from prostate and urinary tract disease for three years. Scar tissue had to be removed occasionally from prior treatments of them. He was started on the kidney herbs and in two weeks (13 days) he had a considerable improvement in urine flow. At his follow-up visit we searched for toxins. He had radon, chromate, yttrium and strontium in his genital and urinary tract. He got rid of his water softener (such salts are polluted with chromate), toothpaste (strontium source) and opened the crawl space vents (source of radon). This cured these problems in less than two months. It also cured his stomach ulcers for which he had to take medicine. He was so pleased he decided to install a crawl space fan and pursue a parasite program and dental health just to see what extra health improvements he might get.

Omer Whitney, age 45, had always been a strong, healthy, hard worker. He could now barely walk, due to weakness and pains of several kinds; his prostate problems began several years ago. Our tests showed 4 kinds of kidney stones. He was started on only half a dose of the herbal recipe to give them a chance to dissolve more slowly. One month later he still had some stones although his leg cramps were already gone. At this time we found *Ascaris* (both kinds) which he killed with a frequency generator. We also found carbon tetrachloride and pentane (in decaf beverages) built up in him; also gasoline and TCE. He was considerably improved five weeks later and was quite eager to improve further. A toxic element test showed he was full of copper, antimony (from mineral ice massages), cobalt (aftershave), zirconium (deodorant), thulium (vitamin C fortified orange drink) and mercury (very high, from tooth fillings). He planned on cleaning all of it out of his body and regaining his lost strength.

Harvey Van Til, age 35, came in for his prostate and testicle swelling which began shortly after a vasectomy. He ached over the front right side of his abdomen. He was started on the kidney herbs and in four weeks he had eliminated his oxalate crystals and felt considerably better. We next found the adult intestinal flukes and human liver flukes in his prostate gland! After killing them immediately with a frequency generator and getting instant relief of pain, he got his own device and did not need to return.

Clayton Gamino, 26, had pain during urination which he interpreted as a left-over from a prostate infection he once had. He got all the metal out of his mouth, and did a kidney cleanse. A half year later he had no remaining pains and was able to father his first child.

Side Pain

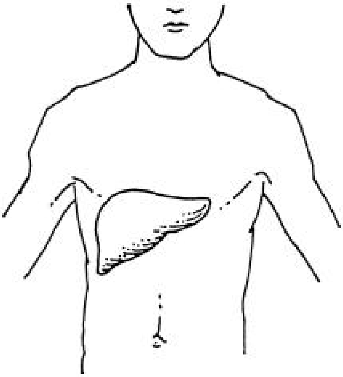

Fig. 22 The liver has a large lobe on your right side with the gallbladder tucked inside. The left lobe is smaller.

Pain on the right side can come from problems at the ileocaecal valve or the appendix or the large intestine itself. It can also come from the liver which is higher up but is sending its pain message to your side. Pursue it as an intestinal problem first, killing parasites and bacteria and normalizing bowel movements with the Bowel Program. If the pain persists, especially if it reaches up the side to the middle of the waist, it is probably due to the liver.

The liver is a large organ, mostly on the right side of the body, but with a smaller lobe on the left side. It is the chemical manufacturing plant for the body. It pulls in the food from the intestine (that you ate hours ago) and makes your body's chemicals from them. Toxic items are changed chemically into non-toxic items that the kidney is able to excrete into the bladder. Fatty things must be made water soluble for them to leave with the urine. The liver also makes bile and sends toxic items along with it to the intestine through the bile ducts. The bile enables calcium and fat to be absorbed. If the liver isn't getting much bile to the intestine, fat is left in the bowel contents and the feces will float in the toilet bowl. That is your clue to bile duct blockage.

Bile is bright green. Mixed with intestinal contents it turns the bowel movement dark brown. If the bile is not arriving in the intestine the bowel movement will stay light colored, even yellow or orange. This is another clue to bile duct blockage. Over a quart of bile should exit the body each day. Since bile is loaded with cholesterol this daily excretion of bile is a major method of keeping cholesterol levels low. If the bile ducts are choked with debris so only half as much (often only a cup instead of a quart!) bile is produced and excreted you can expect cholesterol levels to rise, and digestion to be bad. When food isn't promptly digested and absorbed the ever present, ever-ready bacteria will grab it for themselves.

Taking cholesterol-lowering drugs should be reserved for cases where natural excretion cannot be regained.

E. coli and other intestinal bacteria, which do no harm in moderate numbers, can overgrow in a few hours to give you bloating, gas and pain. Your body produces no gas. Only bacteria can produce gas. If your side pain is accompanied by bloating and gas, you know you have a digestive problem. And that this digestive problem stems from a congested liver if the pain is directly under it or over it, or if the feces are light colored or your cholesterol levels are high. Not everybody has all the symptoms.

To clear the clogged passages of the bile ducts, you simply do the liver cleanse (page 552) over and over until the problem is gone. There is one catch. If there are living parasites in the bile ducts, they will not let the bile ducts clear themselves. They are stuck fast to your ducts as a tick can be to your skin. They must be killed before they will let go. Zap them all, or you may use the herbal parasite program, staying on a twice a week maintenance program. Only after parasites are dead (after day 20 if using the herbal program) will you get a lot of “green stuff” and be able to clear “stones” out of your bile ducts. Only one large duct at a time will clean itself. We have hundreds of larger ducts and thousands of tiny ducts feeding into the larger ones! Stay on a schedule of cleansing the liver every two weeks (unless you are ill) until your side pain is gone, your digestion is normal, and you are bouncing with energy. You may also lose some weight, but only if you are overweight.

Remember that a clogged liver does not necessarily give you pain by itself. It is more likely the bacteria in the gall bladder and bile ducts, causing inflammation there and in your intestine, that cause pain. Don't wait till pain occurs over the liver. Use whatever clues you can to diagnose your clogged condition. Or just assume it is clogged. Do the cleanse, and see if you get any stones out. It can never hurt and can help a lot.

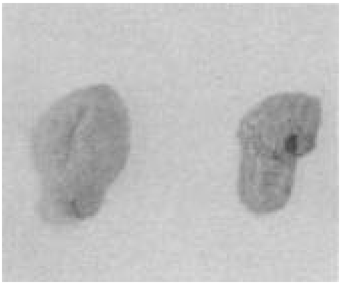

Fig. 23 Gallstones.

Bruce Hearn, 40ish, had severe side pain for several weeks and was rather gassy. He had seen a clinical doctor who found him in good health. But we found sheep liver flukes in the intestine and in the thymus! The thymus is an immunity-giving gland, so anything in the thymus is a very serious matter. He also had benzene in his thymus (inviting AIDS). He quickly switched from drinking soda to drinking milk. He went off everything in the benzene list. He killed the flukes electronically and started on the parasite program. Three weeks later the benzene was gone, his side was very much better and he could begin a kidney cleanse for his low back pain. He hadn't cleansed his liver yet! His improvement was probably due to improving his immunity which then controlled the bacteria.

Midabdomen Pain, Stomach Pain

The colon crosses over from your right side to the left side at the midabdomen. This is also the location for the bile duct to join your small intestine. Most midabdomen pain comes from either the colon or the bile duct connections. Kill parasites and bacteria by zapping or with the herbal recipe. Improve your digestion with diet clean up (off moldy food, boil dairy products). Sometimes the midabdomen pain stems from the stomach itself. The valves at the top and bottom of the stomach are meant to keep the food in. The valve at the top where the stomach joins the esophagus is a favorite location for bacteria.

You seldom feel them here though. This is just under your breastbone. They do their work quietly. Eventually, the tissue there is so weakened, the valve can't shut tightly and food is allowed to go back up the esophagus. This is called *reflux* and you may be told to sleep with your head elevated and to eat small meals, especially at bedtime. You may be given Reglan,™ a drug to help empty the stomach faster. Digestive enzymes will help empty the stomach, too, but may harbor molds. The real solution is to kill the bacteria in this area and keep them from reinvading it. Start by killing the *Salmonellas* and *Campylobacters* (zap them).

Within a day or so, however, the area can be re-invaded. Some bacteria are coming from your teeth! Some come from the liver! Often both sources are supplying bacteria to the stomach. The stomach allows a little bile to enter at the end of each meal, this is normal. But if the bile is full of live parasite stages and bacteria they may try to colonize the stomach, too. If there is insufficient stomach acid to kill them or if there is an accumulation of toxin in the stomach, they will get a foothold. Then they can burrow down deep to escape acid. Eventually, an ulcer can develop.

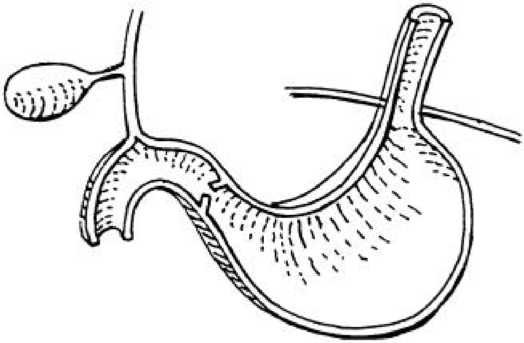

Fig. 24 Stomach, esophagus, diaphragm, gallbladder.

Stomach Ulcer

Often, such an ulcer is painless.

You are, therefore, not alerted to the parasites' presence until they have established themselves in good hiding places. Taking antacids, of course, works in their favor. The solution for both stomach pain and stomach ulcers is to kill parasites and bacteria, followed by dental and liver clean ups. One very common toxin that accumulates in the stomach is the pesticide **arsenic**. You inhale it right along with the flies and roaches you may be trying to kill with arsenic-laced pesticides.

Other inhaled toxins are fiberglass, asbestos, formaldehyde and freon. Your nose and mouth mucous traps a lot of these whereupon you swallow them and they glide into the stomach. Tooth bacteria and tooth metal get into the stomach the same way. You simply swallow them.

Clean up your air, don't use any pesticides (see other methods in *Recipes*). Your dentalware may be cleaned up in a few dental visits but the liver cleanses must go on for a year or two before it is reasonably clean. You may get pain relief in a few weeks but this should not derail your intention to revitalize yourself completely with a cleaned liver and stomach.

Hiatal Hernia

When bacteria have spread to the diaphragm and weakened it, along with the upper-stomach valve, food is allowed to get pushed up right through the diaphragm. Pressure in the abdomen from overeating or sitting in an easy chair pushes it up. Walk after eating. Empty the stomach before going to bed. Don't sit in recliners. Very many of us, about 25%, have a small hiatal hernia. It only hurts if it is inflamed. Work at clearing the inflammation first, to get pain relief. In time, the area will strengthen and the hernia itself may get smaller.

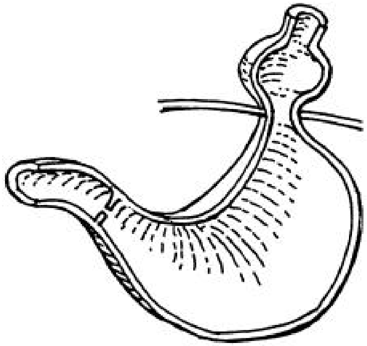

Fig. 25 Hiatal hernia.

Jeffy, 3½ months, had been screaming a lot ever since birth. He was very gassy. He could not be put down without starting to cry. He was congested and had colds. He was on mother's milk exclusively so his chronic *Salmonella* infection must have come from someone who was a carrier, possibly his mother, although she had no symptoms. The mother had platinum and tellurium in her milk (*Salmonella* can be transmitted in milk but this was not checked). The solution was to clean up the mother's milk by getting the dental metal out of her mouth, and washing hands before nursing. Jeffy's father had an *Ascaris* and intestinal fluke infestation. It is quite possible the baby had these also, giving him a nasty tummy ache in addition to the gas pains. The baby's older sister had screamed and cried the first nine months, too, so the parents were told it was a “familial thing.” She also had ear infections, colds and a flaky patch of skin occasionally. She, too, had *Ascaris*. They were very pleased to be able to clear up the whole family's problems by killing parasites and removing toxins.

Ruby Adair, age 14, filled a page with her symptoms. She was also chronically fatigued and had consumed enough antibiotic “to fill a room.” She had shooting pains from under her feet up her legs. A quick check revealed mercury and thallium in her immune system. Instead of being dismayed, this news filled her with hope for recovery. She arranged for dental metal replacement. Then she related her stomach “story”. While at a wedding, a year ago, she began vomiting with diarrhea. She thought it was the stomach “flu” but she didn't recover for six months and had to miss school. The psychologist thought it was emotional. She was still only attending school one hour a day. We found *Fasciolopsis*, the intestinal fluke, in her stomach wall as well as in her intestine. She started the parasite program and in three weeks her appetite was back, insomnia was gone, fatigue was better and a significant improvement was evident.

Respiratory Illness

Asthma

is a very old disease described in the ancient literature. The only progress we have made to date with this disease is to give drugs to soothe the symptoms.

Asthma is associated in **all** (100%) cases with tiny *Ascaris* larvae. As soon as eggs hatch (in the stomach, immediately after swallowing filth) the microscopic larvae travel to the lungs, not the intestines. Imagine the distress of lungs full of tiny worms! One tries to cough them up, of course, but in our misguided effort to be polite we teach children to swallow anything they cough up! Some swallowing is inevitable and the young worms are back in the stomach, this time to set up their housekeeping in the intestine. Some never leave the stomach, causing children stomach aches and, of course, a large entourage of bacteria which, in turn, have their viruses.

Most cases of *Ascaris* infestation also show *Bacteroides fragilis* bacteria which, in turn, carry the *Coxsackie* viruses (brain viruses). Whether or not these bacteria or viruses will thrive in you depends on whether you make a good home for them, namely have low immunity in some organ. The preferred organs for *Bacteroides* are liver and brain (brain tumors always show *Bacteroides*). The preferred organs for *Coxsackie* viruses appear to be tooth abscesses and brain.

Not everybody with *Ascaris* develops asthma, even though they always go through a lung stage. Does it depend on the age of the person when the infection develops? Or how many *Ascaris* are present? Or the time of year when lung infection is present?

That innocent cough of early childhood should not be neglected, as simply "croup." At the first sign of a cough, use a frequency generator set to 408 KHz, the frequency for *Ascaris*, or use a zapper. Pay extra attention to washing hands before meals. Pay extra attention to the animals nearby. Kill their *Ascaris* with a zapper and keep it up daily or put parasite killing herbs in their food.

Asthma sufferers become **allergic** to many air pollutants such as pollen, animal dander, smoke. The production of histamine in the lungs and the vast interconnectedness of histamine to allergies has been well studied scientifically. Although invasion by worms is known to result in both histamine production and high eosinophil counts (over 3), and **asthma clients typically have both**, they are not routinely checked for worms clinically! They are simply given drugs to enable better breathing. More and better (though toxic) drugs have been developed.

But you can put an end to your asthma by terminating your *Ascaris* infection. Then wash your hands and fingernails with grain alcohol, and let no more filth past your lips. Wear plastic gloves for a week to break a nail biting or finger sucking habit. For children wash hands before eating anything, even between meals; keep fingernails short.

Dogs, cats, pigs and horses all get *Ascaris*. Never, never let a child clean up any vomit or mess left by an animal. This could lead to massive infection, the kind that could result not only in asthma but **seizures**. Kill the worms in animal messes before you clean them up too. Never use your dustpan or broom. Use cardboard, newspaper or anything that you can afford to throw away with the mess. Squirt povidone iodine (available at drug stores) on the mess–even if it's outside–and cover with salt before cleaning up. Discard it in outside trash. Wash your hands with grain alcohol, paying special attention to fingernails. If you are a food preparer, you could infect the whole family if you don't wash carefully.

If there is an asthmatic in your family, the whole family should be treated for *Ascaris* with a zapper or with the herbal parasiticides. Even after everybody including the pets have been treated, pets should not be allowed in the bedroom of the asthmatic person. Asthma is more than parasitism. It is also an allergic reaction, to the pet and to other inhaled bits of matter.

Clean the air. Smoke of any kind, fragrance and chemicals of any kind, all household cleaners, polishes, and so forth should be removed. Store essentials in the garage, not the basement, since basement air rises. Clean up the whole house. Persons who must use hair spray or nail polish should do so outside in the summer and in their own rooms with the doors closed in winter. Use a chlorine filter for the water, especially at the shower. Install central air conditioning if possible, with maximum filtering (but never with chemicals added to the filter and never with a fiberglass filter) at the furnace. A room air filter (not fiberglass) is next best. Use it for an hour in the bedroom ahead of bedtime so the air currents can cease. Just moving dust around is worse than leaving it there. Never do dusting when an asthmatic is in the house. Lungs heal quickly when the air is clean and there is no reinfection. The best place to recover is outdoors away from trees and bushes or indoors with total pollution-free air conditioning (free of asbestos, formaldehyde, arsenic, fiberglass, pet dander).

As your asthma lessens, reduce your inhalers, but always keep them on hand. When you suddenly need them, try to identify your source of reinfection or allergens. Use this experience wisely. Try to understand the recurrence of your asthma. Keep notes. It may take half a year with a dozen recurrences to finally learn and conquer! It will feel great to breathe without spraying yourself and taking medicine. If it comes back a year later, figure out what is happening that's seasonal. A pine tree near the house, a flowering bush, the first mown grass? Stay away from these until you are completely healed.

There are traditional herbs for helping lungs. Grow your own **comfrey** and **garlic**. Make **mullein** tea from the dried herb (see *Sources*). Read herb books for more help. Dry some for winter use, being careful to do it right and not let it mold.

Suzanne Carlyle, 45, had asthma from infancy. She was currently on Albutesol™ spray and tablets daily. Now she was beginning to have arthritis too. She was given arginine to replace caffeine. She had two species of *Ascaris* and was allergic to cats and other animals. She was started on the herbal parasite program after killing *Ascaris*, *Bacteroides* and *Coxsackie* viruses with a frequency generator. Her lungs showed kerosene, carbon tetrachloride, mineral oil, benzalkonium (from udder wash, she was also milk sensitive), aluminum from her cookware, and aluminum silicate from her salt. She had two extra lung parasites: *Paragonimus*, a lung fluke, and *Pneumocystis*. She was immediately improved after cleaning up these sources and canceled her future appointment.

Cay Wenkert, 63, had asthma for many years for which she took Proventil™, but this gave her such bad side effects she had to stay home now and not venture out. Her lungs were full of benzalkonium (toothpaste), arsenic (ant poison under kitchen sink), zirconium (deodorant), and nickel from tooth metal. She had dog heartworm in her lungs(!), in addition to her heart where she sometimes felt pain. She had *Ascaris* and *Naegleria*, *mycoplasma*, *Endolimax* and the intestinal fluke in her lungs! She coughed up blood, after her doctor had diagnosed *bronchiestasis* recently, meaning her lungs were not capable of sweeping out the daily refuse we all breathe in. In spite of killing these and cleaning up her environment she got no improvement. She repeatedly got parasitized. She had four or five root canal teeth which she was unwilling to have pulled. Hopefully, the tradeoff between teeth and health will soon begin to look poor.

Amy Newberry, 41, had recently begun to have asthma attacks. She had been on cortisone but now was on inhalers through the day (two puffs Ventillin™ 3x/day, plus three puffs Intel™ 3x/day plus four puffs Asthmacort™ 3x/day). She was often hospitalized for attacks. She also had stomach trouble and sinus problems (had pneumonia recently). She had *Ascaris* larvae, *Endolimax*, *Naegleria* and *Acanthocephala* in her lungs! She also had arsenic and palladium (tooth metal). Going onto homeopathic medicine for stuffiness helped her avoid some hospital visits. It took several months (5 visits) to track her arsenic source to the bedroom carpets (stain resistance!). After steam cleaning it herself and doing a liver cleanse (after first killing parasites) she was amazed at her improvement. She had not been to the hospital in a month and was only using inhalers preventively. She planned to get dental work done. She got all her former health back.

Laura Brewster, 25, lived in a very old house. The slightest exertion would give her an asthma attack. She frequently got pneumonia, too. Her lungs had beryllium (coal oil) and asbestos, and two parasites, *Paragonimus* (lung fluke) and *Ascaris*. She got rid of all the old fashioned lamps and candles in the house, but could not find the source of asbestos. She got rid of the attacks but her cough and pneumonia bouts will continue until she moves from that house.

Brett Wilsey, 70, was congested most of the time, had chronic sinus problems, was getting allergy shots for dust and mold, and was on several inhalers for his asthma plus emphysema. He had asthma for eight years. He had oxalate, urate, and all three phosphate variety crystals in his kidneys. He was started on kidney herbs. His blood test showed high "total carbon dioxide" or "carbonate" showing that his air exchange was not good. His potassium was low, showing that his adrenals couldn't keep his electrolytes (sodium, potassium and chloride) regulated, in turn, giving him muscle weakness. His LDH was very high, showing that his heart muscle) was in distress, too. Fortunately, his RBC was slightly elevated, to help with oxygen delivery. (See Anemia page 285). His eosinophil count was high, as is the rule for asthmatics since they all have *Ascaris* worms. Besides *Ascaris* he had pancreatic and intestinal flukes in his lungs(!) plus *Heterakis* (a chicken roundworm), human hookworm, and *Prosthogonimus*, another fluke. He was started on the parasite herbs. Two weeks later he felt a lot better although he still had some *Ascaris*. He was toxic with barium and hafnium (which were traced to his dentures) nickel, tin, rhenium. He stopped using commercial "denture-soak." His wife stopped using hair spray and they started leaving the car out of the garage and walked around to the front door, instead of using the attached garage. He also started drinking homemade comfrey tea. This made him feel still better.

When he removed his watch he got rid of his nickel source. Nickel has an affinity for lungs (as well as prostate and skin). He was now down to one puff of inhaler instead of two, only four times a day instead of hourly. He could breathe deeper than before. Then the lead in his water was found and traced to a single "sweated" joint in the pipes. He was started on food grade hydrogen peroxide, working up a drop at a time; now his cough became "productive", he was coughing up a lot.

He got the plumbing fixed and did a liver cleanse after which his fatigue lifted. He couldn't stop laughing and joking about the "emphysema" he was so worried over earlier.

Bea Karnes, 49, had asthma from childhood when she also had eczema. She was presently on several medicines plus allergy shots and antibiotics for it. She was toxic with antimony although she used no eye makeup and europium, tantalum, and gadolinium from dental metal. She also had hand swelling in the morning and was started on kidney herbs. She killed her intestinal flukes (in the intestine) and *Ascaris* in her lungs and was not seen for half a year. When we eventually saw her, she said her asthma had been "wonderful." She was faithfully controlling parasites with the maintenance program and occasionally drinking kidney herb tea.

Cynthia Prout's whole family (mentioned previously) had asthma. She was told it was inherited. The three young children and herself were on inhalers, nose sprays, cough syrups and antibiotics. Nola, age 10, also had ear infections and numerous allergies. Lewis, age 8, was a slight, nervous boy; he had been off wheat and milk for many years due to intolerance. Irwin, age 5, seldom went with the family due to his frequent stomach aches and the fact he could vomit without notice. The mother and two children who were with her (Irwin stayed home) had *Ascaris* infection and Lewis also had pancreatic flukes. Their home was toxic with vanadium, namely, a gas leak. The parasites were quickly killed and gas leak repaired. The whole family got well and canceled their next appointment. Some inherited disease!

Asthma is just one of our common respiratory problems. The causes are always a combination of *Ascaris* and other parasites with pollutants (allergies).

Bronchitis, Croup, Chronic Cough

- In **bronchitis** the bronchioles are the site of the problem.
- In **croup** it's further down.
- In cases of **chronic little hacky coughs** it may be heartworm!

Bud Portillo, 62, worked on earth moving machines. He frequently had a "bronchial virus" or "bronchial infection." He was always coughing. He had *Ascaris* infestation and arsenic pollution in his home. He also had palladium toxicity from tooth fillings. As soon as they removed the mouse bait from their home, tore down the hallway wallpaper (arsenic source) and changed wells (the well water had arsenic in it from seepage!) and killed parasites, his cough was gone. He also finished his dental work. All this took six months.

Gene Mizell, age 4, was constantly clearing his throat. His parents wondered if it was a "nervous habit." He had two species of *Ascaris* living inside his small body. He was started on a small dose of parasite herbs immediately, stirred into a daily dose of VMF syrup (see *Sources*). He also had PCBs, aluminum silicate, lutetium from recent painting in the house and xylene and toluene solvents. He was switched off soda pop and onto milk. We saw him five weeks later. He had gone through three treatments with vermifuge syrup. His throat clearing was gone, as well as his hacky cough and the blue circles around his eyes. He still had PCBs, which they later cleared by switching off detergent.

Carmen Castro, 30ish, had a chronic cough without being ill. She also had some heart irregularity. She had *Dirofilaria* (dog heartworm) and *Loa loa*. There were no indoor pets. After killing the parasites with a frequency generator and starting on the parasite herbs she was still coughing a bit but her pulse was down to 80 (from 120). She still had toluene, TC Ethylene solvents and mercury, platinum and lead in her body. Then a gas leak developed in their home. She used moth balls and painted a room. After this she had to be on antibiotics for her extreme cough. This time she had *mycoplasma*, *Haemophilus* and *influenza*. She killed these electronically and her coughing stopped. Until she gets the dental work done she will continue to pick up new infections.

Hope Feldman, 57, had been coughing for half a year. She had seen many doctors including a reflexologist. She had *mycoplasma* and *alpha strep* at tooth #17 (lower left wisdom tooth). As soon as these were killed with a frequency generator her cough stopped. She was advised to wear turtleneck sweaters for extra warmth over her bronchii and get a cavitation cleaned at tooth #17.

Teresa White, 37, had bronchitis several times each winter and was put on antibiotic for the whole season to keep it from breaking out. Her lungs were loaded with tantalum from dental metal, cobalt from detergent and thulium from her vitamin C. She had taken herself off milk, thinking it might be a factor, and used an air cleaner but without results. As soon as she had the tooth metal replaced with plastic (in less than a month) she could go off antibiotics and also was rid of a chronic sinus condition, but still had a bronchitis bout. She had *Ascaris* larvae in her lungs and phosphate and oxalate crystals in her kidneys. Her diet was changed to include milk and fish, magnesium, lysine (500 mg one a day), vitamin B_6 and a vitamin A+D capsule. She started on the kidney cleansing herbs and then the parasite program. In another month the whole problem was gone. Wisely, she got ready to cleanse her liver.

Craig Stewart, 2, had a history of respiratory problems. He had been on Ceclor™ and Ventilin™ for a long time; he had pneumonia the previous year. He had whip worm (*Trichuris*) infesting his intestine which was promptly killed with parasite herbs (as much as his parents could get down him was effective). He was toxic with asbestos. When the clothes dryer belt was changed to a U.S. variety (imported belts contain asbestos which flies into the air when the dryer is used), Craig's health turned around.

Doris Gumb, 26, was on Isoniazid™, Tussionex™ and Rifodin™ for tuberculosis. It started with coughing. She was down to 98 lb. in weight. Her lungs were toxic with beryllium (coal oil fuel), mercury, uranium, and tellurium. She began by clearing all toxic items from her house and basement and then bringing an air sample for testing. She also had *Ascaris* larvae in her lungs and pancreatic fluke in her pancreas. In three weeks she was coughing less and producing less with each cough. The “clean” air sample still had uranium, tellurium, mercury and beryllium. She had no metal tooth fillings, though. In the next three weeks she found a hole in the floor to the crawl space. It was letting up uranium dust and radon. The mercury was traced to the carpet in a child's bedroom. After throwing it out, her sputum cleared up and she was on the way to recovery, although we never found the source of tellurium.

Breast Pain

Although lumps and cancer in the breast produce no pain, they sometimes do give you little warning twinges. These twinges cross into the breast from the armpit or from any direction. It is gone in a second, leading one to believe it couldn't be serious. If the breast has any unusual sensations, painful or not, investigate immediately.

Test yourself for cancer. (You may use a specimen of mother's milk as a cancer test since this has mitotic stimulants in it—see Curing Cancer page 331). If you have purchased a slide of breast tissue (mammary gland) you can search your breast for cancer. If not, but you find “mother's milk” in your white blood cells, assume it is cancer and clear it up immediately.

If you don't have cancer, search for the pollutants giving you these twinges of pain. Search for your deodorant, cosmetics, and soap in your white blood cells. Search for dental metal. The breasts are often full of nickel. Nickel is quite soluble in fat and the breast is mainly fat. Nickel is one of the top carcinogens listed by researchers. It could even explain the high incidence of breast cancer. But titanium and barium from cosmetics, as well as asbestos and fiberglass are also quickly accumulated in the breast. Clean up your dentalware and body products. Check for exposed fiberglass. Change your dryer belt. Buy a new non-CFC refrigerator. Never try to get rid of these pains with pain killers; let the pains show you whether the clean up has been complete.

Breast Sensitivity

Breast sensitivity can be quite uncomfortable to the point of not being able to wear a bra, especially near period time. It may be due to high estrogen levels; this is also conducive to breast lumps and breast cancer. Most of your estrogen is produced by the ovaries before menopause and later by the adrenal glands. Too high estrogen levels plague the modern woman. Certain food molds, particularly **zearalenone**, causes over-estrogenization. It affects men too. I have found it in popcorn and corn chips! And brown rice. Stop eating these. Eat white rice. If you make cooked cereals be sure to add vitamin C to them before cooking (1/8 tsp. per cup), to detoxify food molds. I do not know whether taking vitamin C with your popcorn would detoxify zearalenone. Don't risk it. The excess estrogen compounds must be detoxified by the liver. Yet, the liver may be incapable of this because you ate yet another food mold! See the section on moldy food (page 381).

Over-estrogenized women are over-emotional, seemingly on a roller coaster of enthusiasm and despondence. They can develop a high pitched voice, that almost sounds squeaky. High enough estrogen levels are important for fertility but too high levels can cause infertility. Your body is eager to set the level just right, if only you will clean up the ovaries of parasites and pollution. Don't stop your clean up until the breast feels normal again throughout your cycle and you don't feel over-emotional, even just before your period.

Breast Lumps

Breast lumps may or may not be painful. If you feel one, don't wait to be more certain, don't wait to analyze it with tests, don't wait for a doctor's diagnosis or a mammogram. Obtain a frequency generator or zapper and zap yourself immediately. Also do the herbal parasite program immediately (page 338).

Your body often turns the breast into a collecting station for toxic wastes that have been drawn downward from the top of your body. From your head where shampoo and hair spray and cosmetics leave their daily deposits, from your dentalware with its constant supply of heavy metals, from neck and armpits where cologne, deodorant and soap leave their toxic residues. The lymph nodes under the armpit and the region above the breast, collect it all and let it slip into the breast where it is bundled up in a cyst. Perhaps the kidneys are clogged so toxins are forced to go to a designated dump site instead of out through the bladder. Do a kidney cleanse. Don't rest until all your breast lumps are gone. They will begin to shrink in three weeks if you are removing the correct toxins. Even radon and asbestos go to the breast, so be meticulous with your cleanup.

When the platelet count (in a blood test) is very high (over 400) there is quite a tendency to form cysts or lumps since platelets make your blood clot. The platelet count goes up when parasites are present. Maybe your blood is attempting to clot them! These clots make “nests” for fluke stages which may be why breast lumps often become cancerous. If yours is over 300, (it should be 250,000/cu mm) start patrolling parasites regularly. Stopping the use of caffeine and taking vitamin E (400 u. a day) are helpful in recovery but don't rely on these minor measures. Breast lumps definitely invite breast cancer.

Leslie Yeager, age 37, had breast soreness and "fibrocystic lumps”. She had cerium and nickel accumulated in her breasts. They cleared up in weeks after her dental metal was gone (she simply took out her retainer). Later she replaced it with a partial made of plastic.

Kari Pfeifer, age 36, had numerous cysts in both breasts and uterus. Her estrogen level was too high (187 pg/ml on day 22 of her cycle; the day of testing is important since it varies through the cycle). Her breasts were full of beryllium (coal oil from hurricane lamps) and radon. After these toxins were removed, all her breast lumps got smaller. After she did the kidney and Liver Cleanse, the lumps got softer and breasts were no longer painful. She had several root canals which filled her breasts with numerous bacteria, mainly *Histoplasma cap* (root canals develop infection around themselves). After starting her dental cleanup and killing bacteria with a frequency generator, all her breast lumps disappeared.

Claudia Davis, age 41, had breast soreness ever since a mammogram two years earlier. She had numerous other pains and indigestion. She had intestinal flukes in her intestine and fluke eggs in her blood: a dangerous situation. They might land in the breasts and start developing there. This is how cancer begins. But she did not have cancer yet. She had a buildup of niobium from polluted pain killer drugs and thulium from her vitamin C. She had *Salmonella* and several other bacteria in her white blood cells, which accounted for digestive problems. In eight weeks she had cleaned kidneys, killed parasites and gotten rid of her heavy metals. Her breast pain was better and a lump on her eyelid had also disappeared.

Stephanie Nakamura, 68, had six surgeries to remove breast lumps, going back to youth. Her recent mammogram was O.K. Her breasts were toxic with cadmium, lead, gold, radon, uranium, gallium, silver. Our tests showed she had kidney crystals and she was started on the kidney cleanse. She was given vitamin E, (400 units daily), sodium selenite (150 mcg daily) and vitamin C (1 or more grams daily). Her triglycerides were also very high showing again that she had kidney problems. She was given magnesium (300 mg daily), vitamin B_6 (250 mg daily) and lysine (500 mg daily). She killed parasites and cleaned up everything except gallium, silver, mercury, gold, cadmium. These must have come from her gold crowns. Her dentist advised against removing these and proclaimed they had nothing to do with her developing glaucoma, arthritis and stomach ulcers. It was a tough decision for her and she made the wrong choice. Perhaps if she had been up for the next breast surgery she would have gotten those "gold" crowns replaced with composite too.

Heart Pain

Pain over the heart region is usually quite real, even though an EKG does not find any abnormality. The most common cause is *Dirofilaria*, heartworm of dogs. It often begins as a pain just above the heart but spreads itself over the whole heart region. Kill it with your zapper. Parasite herbs can also be effective. If you did kill them, the pain often intensifies for a day before it leaves. Then the pain should be completely gone.

Heartworm is very easy to pick up again. If you have had heartworm, you should no longer keep a dog for a pet. Give it away.

Another heart parasite, *Loa loa*, is also a filarial worm and may be the causative factor. Both *Dirofilaria* and *Loa loa* can be obtained as slide specimens to use for testing yourself. Heart muscle can also be obtained as a slide specimen, but a chicken heart from the grocery store or snippets of beef heart (make sure to sample all 4 chambers) will do.

Follow up on your heart, even when no pain remains. These tiny heart parasites have stages that you may not be able to purchase in slide form and therefore can't test for. These stages, if not killed, will become adults so a maintenance parasite killing program, herbal or electronic, is essential. Virtually all dogs have *Dirofilaria* in spite of monthly medicine to kill it. They pick it up immediately after their last treatment for it and can give it to you again. The only way to live safely with pets is to give them parasite killing herbs daily in the feed.

Other heart problems such as irregular beat and mitral valve prolapse can clear up along with the pain. Or they may be due to bacteria (see Heart Disease, page 318).

Meredith Zackman, age 53, came for her diagnosed cardiomyopathy (heart disease). She owned a beautiful, old, very big dog, and of course she would never part from him. We knew she would lose her battle against heart disease. She had both *Dirofilaria* and *Loa loa* which we killed instantly with a frequency generator. She was on Lanoxin™, Furosemide™, Captopril™ and Metoprolol.™ We found she also had *Cytomegalovirus*, *Staphylococcus aureus*, *Streptococcus pneumonia* in her heart. The *Staph* bugs were also in tooth #17. She had copper in her heart (from tooth fillings) and cobalt and PCB from her detergent. There was zirconium from her deodorant and fiberglass from somewhere. Her pulse was typically in the 90's. Eight days later, after her dental work to replace metal was done her pulse was normal (low 70's), both worms were gone and she felt much better. But she still carried four bacteria, five viruses and two tapeworm stages: *Taenia solium* scolex in the spleen (she had chronic pain there) and *Taenia pisiformis* in the liver. She was started on Rascal (an herbal combination) for these. Six days later she had all of her problems back including *Loa loa*. She repeated everything, then she had to go off her heart medications because they lowered her blood pressure and pulse too much. She started the dog on the parasite program but continued to be heavily laden with parasites and bacteria that always found their way to her heart. She purchased her own frequency generator and was quite faithful with dog treatments. She may outlive her dog and then regain her health, finally.

Bruce Walby, age 42, had chest pain for three years. We found he had *Dirofilaria* and *Loa loa* in all four chambers of his heart. When he zapped them, the pain left a day or so later.

D'Ann Fonties, age 22, had a lot of chest pain but was told by her doctor it was simply “gas bubbles”. She also had a serious digestion problem. She had *Dirofilaria*, high levels of styrene (from styrofoam drinking cups) and benzene. This information delighted her and she planned to change her habits.

Sheila Osborn, age 27, had chest pain when she lifted objects. Her pulse was slightly elevated (81) and slightly irregular. She had *Dirofilaria*. Five weeks after starting the parasite program she was feeling much better but still had the chest pain. This time she had *Loa loa* (but not *Dirofilaria*).

Wendy Lewellen, age 28, had a chronic cough and chest pain at midsternum (the sternum is the bone attached to the ribs and runs up the middle of the chest). She had *Dirofilaria* in one chamber of her heart (right auricle). She had xylene and toluene solvents which came from her daily beverage, Mountain Dew.™ She was also full of asbestos from her trips to do laundry nearby (this could not be proved, but when she switched laundromats, the asbestos went away). Two months later, after killing parasites, she was free of heartworm and her cough and chest pain were almost gone. She was probably healing very slowly due to the asbestos which was still present.

Lupita Cline, a young mother, had a chronic hacky cough and irregular heart beat. She had *Dirofilaria* in all chambers of her heart and *Loa loa* in her blood but not in her heart. She had carbon tetrachloride, propyl alcohol, hexanedione, toluene, and TC Ethylene buildup from drinking Pepsi™ and Mountain Dew.™ Her pulse was slightly elevated at 80. There were no pets in the house. She was started on the parasite program and two months later was rid of her heartworm but now had *Loa loa* in her heart and was still coughing a bit. She was full of platinum, mercury and palladium from tooth metal as well as vanadium from a gas leak in her home and paradichlorobenzene from using moth balls. She was on antibiotics for a "bronchial infection" and was happy to learn about a better solution.

Slow Pulse/Syncope (Passing Out)

Mason Heckler, 30s, was a mechanic by trade and could not afford to pass out on the job. Yet, it had been happening off and on for 10 years. He had acquired high blood pressure in his teens! This was soon followed by an extremely slow pulse (50 beats/mm). No medicine worked (he had been tried on many) so he was on none. Then he got high blood pressure, it was 160/80 currently. He also had constant chest pain around the left nipple. He had heartworm and was started on the herbal parasite program. He never had a dog. In five weeks his pulse was 72; the parasite was gone. I presume his syncopes were due to sudden blood pressure changes or missing a few heart beats in a row. He began the kidney cleanse next to lower his blood pressure.

Chest Pain

When there is a tightness or just a little pain at the middle of the chest, especially under the breastbone, you may be merely having an allergic reaction. You might be feeling little spasms coming from the esophagus, and reaching up toward the throat from gallstones. You might also have HIV/AIDS disease which has a similar symptom over the sternum. So it is very important to pay attention to even a minor symptom in the chest.

Ruling out HIV disease ranks first in importance. Search for its emission at 365 KHz, or purchase a microscope slide with the dead virus on it as a test specimen (see *Sources*). Or purchase a set of slides representing all the stages of *Fasciolopsis buskii*. Without this parasite you can't get the HIV virus. Purchase a slide of the thymus gland or make your own specimen of throat sweetbreads. Check yourself for benzene buildup in the thymus. (See Using The Syncrometer page 462.)

If you have neither the benzene nor the parasite stages, you have no risk. Your chest distress is due to something else. Improve your air quality so that your lungs are not in distress. This includes radon, chlorine (from the bleach bottle under the sink), colognes, room fresheners as well as the usual pollutants (asbestos, arsenic, formaldehyde, fiberglass, freon).

If you feel waves of pain reaching up to your throat, you probably have a gallstone stuck in a bile duct. Epsom salts can relax that bile duct in 20 minutes. Take a tablespoon in ¾ cup water but only on an empty stomach or you may feel quite ill. Taking a large dose of valerian herb (6 to 8 capsules) may also buy you a little time by relaxing the duct. If you do get relief, you can be sure it was a spasm of some kind. The magnesium in Epsom salts relaxes spasms. I would recommend cleaning the liver (page 552) a number of times to try to dislodge the sticking gallstone. The instructions for liver cleansing advise you to kill parasites and cleanse the kidneys first. But if your throat pain is severe enough, you might just zap and go ahead with it at once.

Upper Back Pain

The main pain may be a dull ache over a shoulder blade, or between the shoulder blades or running right through you from the front to the back of the chest. These are all gallstone pains coming from the liver! Get yourself ready to clean them out. Clean the kidneys first, kill all the parasites that might be residing in the bile ducts and blocking them (flukes, pinworms and roundworms) and get onto a maintenance parasite program. Then mark your calendar for your first liver cleanse. Even if your first cleanse gives you only a dozen bits of green "stuff" you have done well; you have accessed the bile ducts. The pains will probably be "magically" gone the next day, but they might start to return in two days. The bile ducts are having spasms again due to the remaining stones.

After you have cleaned out 1,000 or more, you will get permanent relief. Repeat every two weeks, unless ill, until the upper back pains are gone, permanently. If chest pain or upper back pain is severe, try going off your favorite high fat food (ice cream, butter, cheese). Also try taking 6 valerian capsules, 4 times a day including bedtime to relieve the spasms.

Shoulder Pain

Some shoulder pain is called *bursitis* and some is called *arthritis*. But it always derives from stuck gallstones in the bile ducts of the liver! You can prove this by taking a tablespoon of Epsom salts in ¾cup water at 6 p.m. instead of supper and on an empty stomach (or you may feel quite ill). If some of the pain subsides then you have evidence as to its true cause, because Epsom salts relax the bile duct valves.

Get started cleaning the liver (page 552). By the time you have chronic or acute shoulder pain you have about 3,000 Stones! Count them roughly, as they float in the toilet after the liver cleanse so you know how much progress you have made toward the final goal. Don't start cleaning the liver until you have killed parasites and spent three weeks cleaning the kidneys though. This improves elimination of liquid toxins so a liver cleanse is promptly cleaned up for you.

What emerges from the liver is the most contaminated mess imaginable, full of bacteria and viruses and parasite eggs and stages (all dead we hope) of every kind. It needs prompt clearing from your body. The diarrhea sees to the bowel elimination. But some toxins can only pass through the kidneys. Kidneys should be clean and capable of doing the job. Remember it is unwise to clean the liver before all parasites are dead, especially flukes, because they produce a substance that inhibits any action of the bile ducts!

You are only one day away from freedom of shoulder motion and sleeping on your side again. Permanent improvement, though, depends on progress with your total stone count. Your bursitis can return in a few days or a few weeks. Be patient. You may only cleanse once in two weeks, and not if you are ill. After six cleanses you can be quite sure of being relatively pain free.

Between cleanses use valerian capsules to stop the spasms. It takes 6 capsules 4 times a day to be effective.

Typically only some bile ducts are spasming, and typically those ducts have a single fatty food trigger. Stop eating the high fat food you consume the most (it's probably your favorite). If that hasn't helped in two days choose a different high fat food to omit. The most common culprits are ice cream, potato chips, salad dressings, cheese, butter, cream, and milk.

Perhaps the pain is actually caused by bacteria living in the blocked bile ducts and invading the shoulder. This point has not been clarified. Using your zapper or frequency generator does no good for this pain. Only liver cleaning brings your shoulders back to their youthful mobility.

Upper Arm Pain

Excruciating pain in the upper arm soft tissues can keep your arm hanging straight down for fear of worsening it with motion.

Magnets of high strength (2x5000 gauss) taped to your arm, under your sleeve, can get you through the day. Also try valerian capsules (6 capsules four times a day). Go off all fat in the diet to let the spasm subside. Then start your liver cleanse at 6 p.m. If you used pain killer drugs during the day, the cleanse may not yield anything but it's worth a try anyway. You might be lucky and pop out the chief culprit stone. If not, you should wait several days before trying again; this time avoid pain killers the day of the cleanse. Be sure to zap parasites the day before or earlier.

Peggy Patton, age 60, had shoulder pain and painful feet in addition to aching all over. The aching was due to *Trichinella* which both she and her husband had. It took six months on the parasite program before it stayed away. She had clay colored stools, evidence of bile duct blockage. Then two liver cleanses cured her shoulder pain, nausea and remaining pains. She started gardening again and immediately picked up hookworms and *Trichinella* again. But she learned to sanitize her hands with grain alcohol after washing away dirt and this kept her parasites in check.

Jessica Atkinson, a middle age school teacher, developed a pain in the right cheek quite suddenly. She also had pain over the right mid abdomen and right side at the waist but X-rays and scans showed nothing (she had been X-rayed three times). She struggled for seven years to stay employed. She was having severe pain attacks over the liver and described her stool as almost white after these attacks. She cleaned her liver at least 30 times before she related, one day, that her joy in living had returned. Her gallstones were exceptionally large (½ x ¾ inch). Eventually the abscesses in her upper teeth were found, clearing up her cheek pain and protecting the liver from recurrent infections from these bacteria. Only then did she get permanent and complete pain relief.

Lisa Mattie, 72, had her right arm hanging limply by her side. It was so painful she bent forward to let it hang straight down. But in seven months she had done 6 liver cleanses, getting over 3,000 stones out. All pain was gone although some numbness in that arm persisted. She could also stop using Tums^{TM}, stop coughing, and no longer was bothered by her hiatal hernia.

Elbow Pain

One variety of elbow pain is due to an inflamed tendon there; it is sometimes called "tennis elbow." It is not due to playing tennis or any other arm use. The inflammation is caused by a liver full of stones and parasites, especially flukes which manufacture a chemical that affects tendons. Kill all flukes and cleanse the liver for quick relief. Using your elbows while they are inflamed is traumatic to them, like working with a sore thumb. Don't play tennis or do other arm exercises until they are pain free.

Parasites consume large amounts of your vitamins and minerals. Give yourself vitamin A (25,000 u daily), zinc (60 mg. daily), and B_6 (250 mg twice a day) until the pain is gone.

Wrist Pain

Tendons passing through the wrist can become inflamed from the unnatural chemicals produced by fluke parasites in the liver. Using the wrists to work further traumatizes them (injures them) making it harder for them to heal. A small hole between the tendons lets the nerve and blood vessels through into the hand. Fluke parasites also make chemicals that thicken tendons. When tendons at the wrist thicken, they can squeeze down on the nerves and blood vessels until the hand or fingers feel numb. If you have pain at the wrist or numbness in your hands, killing parasites and cleaning the liver may give you the permanent cure. Wearing a wrist bandage or support can help reduce trauma damage to the wrist while it is healing.

Numbness of hands, without wrist pain, is more often due to a brain problem with parasites and pollutants. Lead, mercury, fluke parasites are the usual culprits.

Thumb Pain and Hand Pain

can be due to liver parasites. Get yourself ready for a liver cleanse. If the pain goes away beforehand, while you are on the kidney cleanse, it shows you had deposits in your joints. You were headed for arthritis in your hands. Read the information on arthritis (page 78) to protect yourself.

Finger Pain

This is pain in a joint, often accompanied by some enlargement or knobbyness of the joint. It is not hard to recognize these as deposits of the same kind as we saw in the toes. You can test yourself to identify the variety. **Uric acid** and **phosphates** are the commonest types. Read the section on toe pain (page 55) for detailed instructions. You can greatly reduce your finger joint deposits and the size of the knobs. In six weeks after starting the kidney cleanse and changing your diet, the knobs may already be shrinking. A large magnet (5000 gauss—used only as directed) may bring pain relief but only dental cleanup and environmental cleanup will give you lasting improvement.

Back Of Neck Pain

The back sides of the neck seem to be highways that run between the teeth and liver. Both contribute to pain at the back of the neck. Pulling an infected tooth or cleaning a cavitation can bring complete relief, only to return the next time a tooth is extracted. Extractions should be followed by cleaning out the cavity created so an infection can't start here. Cleaning the liver can also bring immediate relief, only to find pain and stiffness to return months later. You must cleanse many times for permanent relief.

An allergic reaction to potatoes and tomatoes can express itself in neck pain too. When the liver can no longer detoxify the chemicals (solanine, etc.) in this food family they are free to roam the body with the circulation. Perhaps they prefer to attach themselves at a particular neck site and cause inflammation here. Perhaps an injury was already there, beforehand. Whiplash is often blamed for back-neck pain and indeed chiropractic adjustments can bring total relief. Perhaps the trauma of whiplash first invited all of these contributors. Merely killing bacteria with a zapper is not long lasting. But dental cleanup plus liver cleansing is.

Front Neck Pain

Lymph nodes under the jaw strain your body fluids of the head, removing bacteria and toxins. They are sometimes called “neck glands.” If the stream of bacteria is endless such as when they are coming from a hidden tooth infection, the lymph nodes will enlarge to do a better job. They will also try to remove toxic metals, mouthwash, and toothpaste for you.

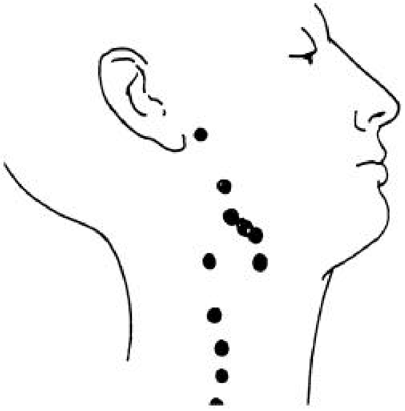

Fig. 26 Lymph node neck glands

To see what is affecting your lymph nodes, purchase a slide of lymph nodes. Since we have lymph nodes in many locations in the body, you can't single out the neck nodes for study. So you will get to see all the toxins affecting lymph nodes everywhere: PCBs in your underwear being removed by groin lymph nodes, lead in your intestinal lymph nodes from the water you are drinking, mercury in your neck lymph nodes from amalgam fillings. Eliminate all these. Don't rest as long as any of your glands are enlarged.

Roland Sanford, 23, had minor pain and a lot of stiffness along the sides of his neck. His arms had some numbness. He only had one metal tooth filling but his whole body was toxic with samarium, beryllium, indium, copper, cesium, and mercury. When it was replaced, his neck pain and numbness got better.

Audrey Doyle had severe neck pains she attributed to sitting all day and sleeping in her wheelchair. She had to take pain killers to get to sleep at night because they were so bad. She knew eating cream and butter made it worse but she had no will power, she said. After a liver cleanse, getting “thousands” of tiny stones, she was free of it.. for one day! But it was enough to convince her and she was determined to be off pain killers.

Temporal Mandibular Joint (TMJ) Problems

Pain at the angle of your jaw is not due to a poorly fitting joint. It fits perfectly but the muscles nearby are pulling it out of joint with their frequent spasms. There is always a hidden tooth infection present. Ask the dentist to search for hidden tooth infections and to clean your cavitations (you will need to find an alternative dentist, and read Dental Cleanup, page 409). Begin immediately to heal these bone lesions with vitamin D (40,000 to 50,000 units once a day for 3 weeks, followed by 2 such doses per week forever), milk-consumption for calcium, and a magnesium oxide tablet. (See bone healing, page 87).

Kill pinworms with a zapper or frequency generator in yourself and family members twice a week and keep everyone's fingernails short for extra hygiene in the family for a month to prevent reinfection. Pinworms can cause tooth grinding and contribute to TMJ dysfunction. For extra muscle relaxation, take two magnesium tablets at bedtime and valerian capsules.

Tooth Ache

Before the pain becomes acute and excruciating, kill bacteria of the mouth, including “tooth decay” and “tooth plaque” frequencies (see frequency list, page 561). Make your dental appointment immediately. If zapping bacteria several times in a few hours relieves pain enough to get you through the night or past the weekend, do not delay a single day. Zapping does not reach into the middle of an abscess—it circles around, so some bacteria are left to repopulate.

Try to understand the problem. If there are teeth with root canals near the location of pain, extract them. The toxic materials of the root canal jeopardize your total body health. Removing them always helps and may let the jaw heal normally where they were extracted. Since the pain is caused by a bulging infection pressing on a jaw nerve, and because each tooth has a related organ(s) it is especially important to clear up all infections to protect these organs.

These relationships are not understood yet. Finding that teeth have the same tissue frequency as some distant organ sheds a little light on the situation. Until the meaning and function of these frequencies is understood we can only guess that they interact somehow. Bacteria have taken advantage of this common resonance and have invaded both organs.

Heart infection by *Staphylococcus aureus* is an example. It is a common heart bacterium, causing much of our heart ailments, particularly mitral valve disease and irregularity problems. They always derive from teeth, whether present or missing! After wisdom teeth are extracted, the hole left in the jaw frequently does not heal, it picks up *Staphylococcus aureus* from the mouth and a chronic infection is started. For this reason it seems ill advised to pull wisdom teeth if it isn't necessary.

Another heart bacterium, *Nocardia*, can originate in the teeth. *Nocardia* can also invade the brain and nerves. Persons with Parkinson's disease always harbor quite a population of *Nocardia*. Zap all these bacteria before and after your dental visits. Remember that killing bacteria in itself is not a panacea. Pain and body damage will return unless you do the proper repair and cleanup work.

Throat Pain

Recurring sore throats are always improved by removing dental metal and root canals, and by cleaning up hidden tooth infections. This is essential for a permanent solution. Bacteria and viruses that cause sore throats are thriving in hiding places under and around dental metal. It is quite difficult to reach the center of such places (abscesses) with electric currents. Even if you could, you would reinfect the very next time you ate non-sterile food! Zapping is not a substitute for cleaning up your dentalware.

Keep your neck warmer than would be perfectly comfortable, gargle with hot water to bring in extra circulation. Use zinc tablets (60 mg. see *Sources*) twice a day; suck on them before swallowing. After cleaning up dentalware you may never have another sore throat!

Children with recurrent sore throats usually do not have a dental problem. They invariably have an air pollution problem. Asbestos, formaldehyde, fiberglass, Freon and arsenic are the chief culprits. Clean these out of the house. Eating moldy food is especially conducive to respiratory illness. The immunity drop can last for weeks! Go off the moldy suspects to raise your family's resistance.

Eye Pain

is caused by parasites.

The eye is a favorite location for many parasites, including *Giardia, amoebas, hookworm, schistosomes, Toxoplasma,* and innumerable others. The eye has two large bodies of watery fluid: the *aqueous humor* and *vitreous humor*, where no blood traverses to bring in extra white blood cells when the need arises. It has its own protective devices, like tears, but some parasites, like *Toxoplasma*, have overcome all the body's

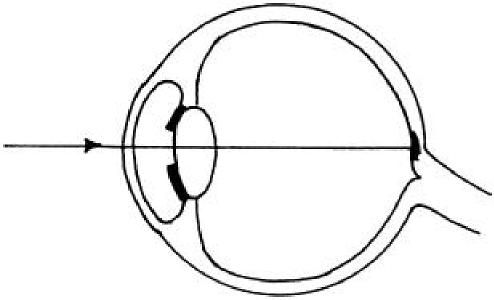

Light travels through the cornea, through the aqueous liquid, the lens, and then the vast vitreous humor, finally striking the super sensitive spot on the retina, called the macula.

Fig. 27 Eye.

defenses and regularly invade the eye if they get into the body somewhere.

We get *Toxoplasma* from cats; the infectious stage is in cat feces.

Toxoplasma infection could be the beginning of a lifetime of eye disease due to weakening of the eyes at an early age. *Toxoplasma* also invades the brain, frequently causing a dull ache or pressure at the back of the head. *Trichinella* is another common eye parasite, invading the eye muscles, so that a muscle is weakened and doesn't allow perfect alignment of the two eyes. Dogs and cats are the source of these.

There should never be a cat box in the house. Never let a child near a cat box.

By killing all the large parasites plus a few bacteria (*Staphylococci*, *Chlamydias*, *Neisserias*) the eye can become pain free in a few days. If pain returns, you missed something or reinfected yourself. Everyone in the family including pets needs to be treated for all the parasites. No indoor pets should be kept by a person of low immunity, since infecting yourself daily and then killing parasites daily is not a solution.

Herpes of the eye is not a single actor. The true host of this virus is a larger parasite, possibly a tapeworm stage. Tapeworm stages should be killed with an herbal preparation, Rascal, or with a zapper. (Not with a frequency generator. Only a zapper can kill all the segments and eggs at once, leaving nothing alive to wander about and find a new tissue to invade.)

Mary Rauch, age 60, filled a page with different pains and problems. Even her eyes had a dull ache around and behind them, sometimes reaching to the back of the head. Her teeth hurt when she ate and she had been getting B_{12} shots once a week for 36 years for pernicious anemia. She was full of *Ascaris*, amoebas and pinworms which kept her legs twitching and jerking in bed at night, even waking her up. She was so much better after the kidney cleanse and parasite program she was eager to cleanse her liver. Before two months had passed her eye pain was gone.

Jessie Healy, middle aged and in good health otherwise, had carried the anxiety of having inherited *retinitis pigmentosa* for forty years. Now her drivers' license was in jeopardy due to fast progression of her disease. Numerous heavy metals had accumulated in her retina, including cerium from dental floss, arsenic from pesticide, tin from health food brand deodorant, PCB from skin salve, cobalt from dish detergent and indium from tooth metal. She had eight parasites in the retina including *Toxoplasma* from association with cats years ago. Removal of dental metal alone arrested the disease process. Two years later she was slightly improved and still driving her car.

Macular Degeneration

Dolores Bollapragada, 50ish, had suffered from clinical ergot poisoning in the past which put her in a coma for several weeks. Ergot is a grain (especially rye) fungus, very toxic to the liver. Its byproducts are used in migraine medicine. She had overdosed, somehow. Perhaps her liver never recovered. It no longer detoxified solvents for her, allowing them to accumulate in her retina. There she had propyl alcohol, benzene, carbon tetrachloride acetone, butyl nitrite, styrene, gasoline, wood alcohol, paradichlorobenzene (moth balls), pentane, methylene chloride and decane. She stopped snacking on cold cereal through the day. She had many habits and products to change. But she was determined to salvage her sight. She did.

Headache and Migraines

Headaches can take the joy out of living. They can lower your motivation level so low that you don't even want to do anything about them! People with daily headache deserve our deepest sympathy since they must carry on with life's daily duties in spite of it. Often, no medicine helps. Although there are common headaches, sinus headaches, migraines, and others, the causes of all overlap a great deal.

Tooth infection, urinary tract infection, bowel problems, and a wormlet, *Strongyloides* are the common causes.

There are also allergic connections which include milk, eggs, citrus and salty foods. But the allergy-approach is difficult to put into practice. It is almost impossible to stay off these foods for long time periods. Eliminating bacteria and *Strongyloides* leads to a permanent solution.

Possibly the parasite wormlet comes first, since even young children can suffer with migraines. *Strongyloides* is a microscopically small threadworm that horses are plagued with, but humans and our pets pick it up easily. Every migraine sufferer (100%) has high levels of this tiny worm. Perhaps it is really the special bacteria it brings with it that cause the blood vessels to seep or to spasm in the brain, causing pain. Zap it immediately.

Bacteria hidden under a tooth filling or root canal or in a space where once a tooth was pulled can be the cause. *Staphylococcus aureus* is a favorite, but various *Clostridia, Streps* and others are often seen, too. You don't feel pain from this small abscess because the pressure isn't building up inside it. It has found a way into your body–namely, your brain!–so no pressure builds up to alert you. Find an alternative dentist with experience cleaning cavitations and finding small hidden abscesses. If you get immediate relief from dental work, only to lose ground again later, the abscess may have formed again (they are notorious for forming again). Go back to the dentist. Irrigate the dental wound site with white iodine (potassium iodide, 12 drops) or Lugol’s (6 drops) to ¼ cup water using a curved-tip syringe.

Do not use iodine if you know you are allergic to it.

The colon is always a thriving mass of bacteria. It should be emptied at least twice a day to keep their levels down. Use the simple herb, *Cascara sagrada*, senna tea, or magnesium oxide tablets (2 to 3 a day) to help you eliminate frequently if your own regularity is lacking.

Gallstones in the liver and a congested liver are sources of bacteria, too. Do several liver cleanses and note the effect.

Is it the toxins made by bacteria or the inflammation from the bacteria or wormlets themselves that produces the headache? Certainly, one can eat the toxins by themselves in foods like yogurt, cheese, wine, sour cream and develop “royal” headaches. Stay away from these **tyramine** containing foods. Boil all your dairy foods to prevent *Salmonellas* and *Shigellas* from swimming into your beleaguered brain. Don't eat dairy foods that can't be boiled for ten seconds.

Begin by killing all *Strongyloides* and other parasites, bacteria and viruses with a zapper. Hopefully, this will only leave a few stragglers behind in abscesses, gallstones and the colon contents. If the colon bacterium, *E. coli*, is your headache cause, start the Bowel Program. Search for the source of your *E. coli* in food or polluted water. Stop reusing water bottles. Most people get their *Strongyloides* back in a few days from pets, other family members, and themselves! Zap every day for three weeks. Try to “clean up” family members and pets. Never let a horse or pet salivate on you. Never put your fingers in your mouth. Always sanitize your hands with grain alcohol after using the bathroom.

To summarize the steps to cure migraine headaches:

1. Kill *Strongyloides* and bacteria by zapping.
2. Do the herbal kidney cleanse.
3. Clean up dentalware.
4. Do the Bowel Program.

5. Boil all dairy foods.
6. Do liver cleanses.

This has never failed to substantially reduce migraine intensity and frequency.

Headaches are also caused by toxins in your environment; especially things you breathe in. Household gas is the most common offender. You don't smell it after you get used to it! Gas pipes are notoriously leaky. Conducting gases through pipes with joints in them, where gases could escape, must be the most ludicrous of all modern “conveniences”. Would you try to conduct water through pipes with holes in them? You would soon see the water on your kitchen or bathroom floor. But gas doesn't land on the floor, it doesn't make a puddle, so you, the consumer, are left helplessly believing you don't have leaks. Every gas pipe that has a seam should have a clear plastic boot around it containing indicator compound to let it be known when gas is escaping. Other methods could be invented to make the gas utility safe. As you will see from the case histories, very many persons are living in a cloud of poisonous gas.

The telltale element is **vanadium**. When your vanadium test is positive, you have a gas leak that your body found, even though the gas company may not. Many gas companies use outmoded equipment to search for it. Four out of five gas companies miss the gas leak. Health Departments and building contractors use modern equipment that detects even the tiniest leak; call them.

If you are a headache sufferer, fixing the pipes is not good enough. Switch from fossil fuels to electric. There will be many dividends. (Remember not to use candles as part of your lifestyle, nor to be a smoker.) Your air needs to be clean: clean of perfumes and colognes, clean of potpourri and air fresheners, clean of air sprays, pesticides, formaldehyde, freon, carpet and car odor, fireplace lighter, and of wood smoke.

Janette Donovan, age 52, had migraines every day but could not tolerate any medication for it. She walked about “like a zombie” most of the time. She frequently had vomiting with them. She was gassy, had pain in her right groin for many years and chronic bladder leakage. Our kidney stone test showed urate crystals. Her urinalysis also showed urate crystals and a slight amount of blood, obviously chronic urinary tract infection. The parasite test showed *Strongyloides*, *Trichuris* and *Fasciolopsis buskii* in the intestines. Her heavy metal test showed beryllium. This was tracked down to hurricane lamps that once held coal-oil in them— it filled her house air unbeknownst to her. Thirty five days later she had done an herbal parasite killing program, done the kidney cleanse, cleared out the hurricane lamps and all fuel containers. Her headaches were “more like pressure” now than pain. She still had *Strongyloides*. Four months later she had some headaches still but not migraines. They were down to once a week. Three months after that she had done a liver cleanse and gotten abut 3,000 stones out! She was still getting some headaches. She tested positive to *Histoplasma* and *Coxsackie virus #4* (a common brain virus) probably stemming from dentalware. She had many root canals and couldn't decide what to do. Keep her teeth and risk return of migraines; or get partial dentures. After eight months of indecisiveness she was back up to ten headaches per month, although not migraines.

Patsy Olsen, age 30, had migraines daily, frequently with vomiting. She had *Strongyloides* as well as *Ascaris*, and other bowel parasites as did her husband and two children. She couldn't tolerate any medication and had to give herself an enema daily for constipation. The whole family was put on parasite killing herbs. A gas leak was found and fixed and the baby's diapers were changed to non fragrant. Ten months later, the whole family still had *Strongyloides*. By one year she was experiencing a couple of good days a month although she still had *Ascaris*, *Coxsackie* viruses, and various tooth-related bacteria. Her two and a half year old had swollen neck glands, was toxic with bismuth from disposable diapers but did not have *Strongyloides*. The eight year old was also toxic with the lotions and fragrance of baby-stuff in the home; she was constantly congested and coughing but became free of *Strongyloides* in six months. After another ten months (the baby had been potty-trained meanwhile) the mother began to have headache-free days.

Donald Schaible, age 14, had migraines, learning disabilities and severe acne. The parasite test showed *Strongyloides*, *hookworm* and *Ascaris*. In spite of being on the parasite program two weeks and zapping, he still had *Strongyloides*.

Kenneth Jones had migraines for thirty five years and had tried all the new medications. They worked for a while, then stopped helping, but he continued taking them anyway. He usually went to the emergency room for the really bad ones, once a week but lived with the constant daily variety. There were two house dogs. They and the whole family had *Strongyloides*. After cleaning up an asbestos problem, killing parasites for five months and clearing kidneys of urate stones, he was down to two to three mere headaches a week. Two months later, he was getting migraines again; they all had *Strongyloides* again. With renewed efforts, one month later his bad headaches were down to one a month, although his low level chronic headache persisted: they had the dogs on a strict schedule of parasite killing herbs as well as themselves. He had not been to the emergency room for a month.

Angelina Gander, age 46, had daily headaches, not migraines. She also had persistent urinary tract infection and sinus infection. She was put on the herbal parasite program and four weeks later was much better. She also lost her chest pain due to heartworm and regained her milk tolerance.

Gracie Arlington had a boy age 6 who wet the bed, a girl age 8 with a behavior problem at school. She was stressed by an unfaithful spouse and thought she should go back to school for a Nursing degree so she could support the family. But she was getting two or three migraines a week in addition to colitis attacks which she feared would make her unable to study. The two cats, a dog, the children and herself all had *Strongyloides*, *Ascaris*, and a variety of other intestinal parasites. The humans were promptly zapped for parasites and the boy was dry that night for the first time in his life. A few nights later he was wet again. This time the animals were zapped and put on the pet parasite program and the children's toileting was carefully supervised. When she dropped her extreme vigilance over all, they all relapsed. After a year of trying, they gave away their beloved dog, put the cat box in the porch

> and the mother did the cleaning herself but nothing prevented relapses. A half year later, her six year old son asked if he could launder his own sheets and blankets. This request broke her heart and she planned to give away one cat. Even with only one cat, the girl's behavior and grades fluctuated extremely, the mother had a migraine a week and the boy wet. During a vacation the mother loaned the last cat, (her cat) to which she was very endeared, to a friend, to see if it made a difference. The boy never wet again, the girl made straight A's, and the mother's headaches became sinus headaches. Some intensive dental work cleared these up. She brought her Nursing diploma to our office two years after beginning. She should have had another diploma for **Intelligent Parenting.**

Earache

Earache is particularly common in children. Bacteria, mainly *Streptococcus pneumoniae,* have built up to a high level. Zap them. They were probably introduced by some larger parasite. But why did they multiply and thrive in your child's ear? There must have been food for them and protection from the immune system. Probably the Eustachian tubes are full of mucous, providing habitat. The mucous is present because some air toxin is irritating the sensitive linings. Clean up the air: not just asbestos, fiberglass, formaldehyde, freon, and arsenic, but perfumes, fragrant school supplies, potpourri.

Being housebound, as in winter, makes the air-toxin problem worse. Use summer time to strengthen your child for the winter—spend summer outdoors. Always wash hands a lot. Stay away from moldy food. Give small doses of niacin (25 mg., niacin thins mucous so it can run away) along with vitamin C (250 mg.) at bedtime.

Putting in “tubes” is a short term solution that is better than perpetually staying on antibiotics, but neither should be necessary.

To hasten healing, as soon as earache is suspected put room temperature olive oil in both ears, tug the earlobe to let bubbles out, and stopper them with cotton wool salvaged from vitamin bottles (other types are polluted with mercury). Put on a light bat that covers the ears.

When adults get earache, the *Streps* are hiding under tooth fillings and in gallstones. Clean up these two sources and zap.

Drinking non sterile milk adds *Salmonellas* and *Shigellas* to the *Strep* ear infection! Small wonder milk is thought to make people mucousy. The ears must now battle them, too. Even a few bacteria consumed in milk can start a whole colony if bowel bacteria have been disturbed by antibiotics. If your child has become “allergic” to milk or gets mucousy, remove cheese and ice cream–not milk–from the diet. Even cheese can be eaten if baked as in pizza or casseroles. Milk must be boiled. Try to reduce the child's *Salmonella* sensitivity by avoiding unnecessary antibiotics. Keep immunity high by avoiding moldy food!

Ear Noises And Ringing

This is also called *tinnitus*. A hissing or buzzing sound is heard in one or both ears. It can be continuous or pulsing off and on. It can be as loud as roaring.

Tinnitus is caused by three things acting in partnership: toxic elements, an allergy to *salicylates* (the aspirin family) and a bacterium *Streptococcus pneumoniae* (the pneumonia bug). This “bug” can be carried in the chronic state after a bout of pneumonia or what seems to be a head-cold. It is always present in earache. It is easily activated by exposure to cold wind or drafts and certain toxic substances. For this reason I recommend keeping the ears warm in winter by keeping them covered or wearing a cotton plug or ear-muffs. The *Strep* bug can also cause *Meniere's syndrome*, congestion, loss of balance and chronic sinus problems. Often an antibiotic of the penicillin family helps immediately, then loses its effectiveness. This shows you that bacteria are involved but cannot be truly vanquished with antibiotics.

Streptococcus pneumoniae often hides in pockets under infected teeth and in holes left where teeth have been pulled—especially wisdom teeth. These can be found by alternative dentists who clean these cavitations. *Strep* also resides in the liver; clean them out with liver flushes.

Certain foods and many medicines, especially cough medicine and lozenges contain salicylate.

To cure the tinnitus we must stop using aspirin or any high level source of salicylates. We must also stop the exposure to certain toxic elements—lead, beryllium, zirconium, benzalkonium. They are present in the air at gas stations and in many of our body lotions, soaps and salves. Switch to safe varieties. Niacin tablets, such as 100 mg, taken three times daily is another treatment, intended to increase the blood supply to the inner ear. Taking a thyroid tablet, by prescription, often helps too.

These methods never fail to improve tinnitus but a complete cure is seldom possible.

William Thall, 47, had a headache daily and was on pain medicine for it daily. He had tinnitus, a loud humming noise in the right ear. He had *Shigella* (producing nerve toxins) and *Borellia* (Lyme disease virus). He had *Gaffkya*, a respiratory bacterium under two wisdom teeth (right side). He also had *Campylobacter* and *Anaplasma* growing on him somewhere. He had *Strongyloides* too. He was toxic with vanadium (gas leak) and asbestos. After ten weeks he had two cavitations on the right side of his jaw cleaned, he had been on the parasite program and his tinnitus was gone, (he still had occasional headaches indicating he still had some *Strongyloides*).

Billie Scott, 26, had a history of ear problems, and lots of headaches. She was started on the kidney cleanse. Then she added the parasite program including their two dogs. She had a cyst in her left buttock, due to PCBs traced to the well water. She stopped using this well and switched off detergents for laundry and dishes. She switched off soda pop, onto milk. In 3½ months her tinnitus was gone and the cyst was shrinking.

Larry Pelegrini, 59, had tinnitus in both ears but was otherwise a strong, tall, intelligent person, who cared for sick people, whether family or not. As soon as he saw how simple it was to kill invaders with a frequency generator, he bought one. Preventing their recurrence was his big challenge since he had neither the means nor insurance to do dental work. He was toxic with aluminum, copper and PCBs. After switching to borax for all washing purposes, he got rid of aluminum and could feel his memory improve. The kidney herbs cleared him of uric acid and oxalate and he felt more supple. He had to go off his favorite beverage to get rid of pentane and methyl ethyl ketone. Twice a week he killed two dozen parasites and bacteria, that just seemed to pop up from nowhere, in order to feel better and reduce his tinnitus. But he lived alone, had to cook, garden, take care of animals and his sick friends which gave him a lot of parasite exposure. Sometimes he would be toxic with arsenic (a new pesticide he tried out) or vanadium (gas leak) but mainly it was tooth filling metal. If only this wonderful man could afford his dental work: what a blessing to society he could be for a long time to come.

Scalp Pain

Infection anywhere in the head can cause sensitive scalp and scalp pain. Even a common cold can cause bouts of scalp pain. Clean up your dentalware and environment.

Lumps on the scalp are often called *sebaceous cysts*, actually this is where your body has sequestered PCBs (polychlorinated biphenyls, extremely hazardous, now banned in the U.S., but often a pollutant in detergents.) Get rid of them by switching off detergents for all purposes. See *Recipes* for dishwasher liquid, dishwasher detergent, and laundry detergent replacements.

Non-Painful Diseases

Explanations for many of these “mysterious” diseases become rather obvious when you see a **common parasite, or pollutant, or both** consistently show up in case after case.

Diabetes

All diabetics have a common fluke parasite, *Eurytrema pancreaticum*, the pancreatic fluke of cattle, in their own pancreas. It seems likely that we get it from cattle, repeatedly, by eating their meat or dairy products in a raw state. It is not hard to kill with a zapper but because of its infective stages in our food supply we can immediately be reinfected.

Eurytrema will not settle and multiply in our pancreas without the presence of **wood alcohol** (methanol). Methanol pollution pervades our food supply—it is found in processed food including bottled water, artificial sweetener, soda pop, baby formula and powdered drinks of all kinds including health food varieties. I presume wood alcohol is used to wash equipment used in manufacturing. If your child has diabetes, **use nothing out of a can, package or bottle** except regular milk, and no processed foods.

By killing this parasite and removing wood alcohol from the diet, the need for insulin can be cut in half in three weeks (or sooner!)

Be vigilant with your blood sugar checks. The pancreas with its tiny islets that produce insulin recovers very quickly. Even if 90% of them were destroyed, requiring daily insulin shots, half of them can recover or regenerate so insulin is no longer necessary. The insulin shot itself may be polluted with wood alcohol (this is an especially cruel irony—the treatment itself is worsening the condition). Test it yourself, using the wood alcohol in automotive fluids (windshield washer) or from a paint store, as a test substance. Try different brands of insulin until you find one that is free of methanol.

Artificial sweeteners are polluted with wood alcohol! Instead of helping you cope with diabetes, they are actually promoting it! Do not use them.

Drugs that stimulate your pancreas to make more insulin may also carry solvent pollution; test them for wood alcohol and switch brands and bottles until you find a pure one. You may not need them much longer, so the extra expense now may soon reward you.

Many persons can detoxify the amount of wood alcohol that pollutes our foods. They do not have a food mold, **Kojic acid**, built up in their bodies as diabetics do. I have found Kojic acid in coffee, and potatoes with gray areas inside. Do not eat discolored potatoes or peels, even if cooked or baked. Being able to detoxify a poisonous substance like wood alcohol should not give us the justification for consuming it. All poisons are bad for us. Do not consume them.

All diabetic persons also carry a virus, *HA virus* in the pancreas. This virus grows in the skin as a wart but is spread quite widely in the body such as in the spleen or liver besides pancreas. It is not necessary to kill this virus since it disappears when the pancreatic fluke is gone. The HA virus undoubtedly belongs to the pancreatic fluke. The question can be asked: Does the fluke or its virus cause diabetes? There might even be a bacterium, so far missed in our observations, that is the real perpetrator.

While recovering from diabetes, it is very important to check your blood sugar every day. Improvement is so rapid, you may suddenly be over-insulinized by your next shot. Cut down your dose to suit your actual need.

There are additional aspects to diabetes that have been studied by alternative physicians. For instance, allergy to wheat and other grains containing gluten is common. Perhaps the pancreas and its islets would heal much faster if grains were out of the diet for a while. Perhaps the 50% improvement that is consistently possible just by killing parasites and stopping wood alcohol consumption could be improved further by a month of grain-free diet. Eating **fenugreek seeds** has been reported to greatly benefit (actually cure) diabetes cases. Are they a specific fluke killer, virus killer, or neither? It seems like a good idea to add this to your diet if you are a diabetic. Wood alcohol also accumulates in the eyes, and there is a connection between diabetes and eye disease. **Bilberry leaves** are an herbal treatment for both diabetes and weak eyes. Do they help by counteracting wood alcohol or detoxifying Kojic acid? Make a tea for yourself, using ¼ cup leaves to three cups water. Drink ½ cup a day. Chromium is another must for diabetics (200 mcg three times a day). It helps insulin enter your cells.

Gold is attracted to the pancreas. Heavy metals should be removed from dentalware including all gold crowns and no metal should be worn next to the skin as jewelry, including all gold items.

Blythe Jenkins was on Micronase™ (5 mg) daily for her diabetes but she still had a morning blood sugar of 183. She had pancreatic flukes and sheep liver flukes in her pancreas, vanadium (a gas leak) in her home and cadmium in her water (old pipes). After killing parasites and cleaning kidneys her morning blood sugar was down to 148. She also got rid of her hot flushes, groin and leg pain. This encouraged her so much she did the rest of her body cleanup and could go off her medicine completely.

Robert Greene, age 65, had been on insulin five years already, getting two shots a day (25 u each), and even this was not controlling his blood sugar which was 288 in the morning. He was raised to 30 units without much improvement. His legs and feet were too painful to walk without a walker to lean on. He had *Loa loa* in his pancreas. (This is another example of a parasite going to the "wrong" location, this time the pancreas instead of the heart. I believe the solvent makes this possible.) Also *mumps virus*, *HA wart virus*, *mycoplasma*, *flu virus*, *chicken pox virus*, *Adenovirus* and *Shigella*, all making their home in his pancreas! This was possible because he had wood alcohol accumulated there, from drinking various beverages and using artificial sweetener. As soon as he stopped this practice and killed everything with a frequency generator his blood sugar fell below 100 in the morning and he had to reduce his insulin to 20 units. He was also on chromium 200 mcg (2 three times a day). Meanwhile, he cleared 2,000 stones out of his liver. He could now walk well again and decided to take a part time job. The whole process took one year.

Ralph Dixon, age 72, had been switched to 30 units of insulin, once a day, after six years on pills for his diabetes. He still had a fasting blood sugar of 242. He had the pancreatic fluke and a host of bacteria and viruses in his pancreas. He had a poodle. After killing the pathogens and cleaning his kidneys, his blood sugar dropped so he cut his insulin to 25 units (blood sugar was at 111) Soon he had to cut it to 20 units. He was also on chromium 200 mcg (2 three times a day), manganese 50 mg. (one a day), fenugreek capsules (2 a day), and bilberry leaves, each supplement by itself. He also lost his angina and fatigue with all this body-cleaning. But if he went off the maintenance parasite program he would promptly get a spike in his blood sugar, showing how easy it was for him to reinfect and how new parasites would immediately find his pancreas.

Melissa Bird, 54, had major illnesses including heart disease (2 angioplasties), numerous other surgeries and diabetes. She took 15-18 units of insulin once a day. Besides pancreatic flukes in her pancreas, she also had intestinal flukes there(!) plus widespread *Ascaris*. She also had warts and candidiasis. Fortunately there were no pets. Her parasites were instantly eliminated with a frequency generator and she was started on kidney herbs for her other problems. Seven weeks later she stated she had to cut down her insulin because her morning blood sugar had dropped to 90. Then she eliminated the decafs and artificial sweetener that were giving her wood alcohol, started the parasite herbs and did a liver cleanse. The day after the liver cleanse her blood sugar went up to 164 but was completely normal after that (under 100) and she did not dare take any more insulin or pills. We advised her to keep monitoring her blood sugar and be very, very vigilant and to please stop smoking.

John Angert, 65, was not on any medicine in spite of his diabetes. He was too embarrassed to tell us his blood sugar. He had a long list of other health problems, too. He was toxic with thulium (vitamin C) and palladium (from tooth fillings). He had a variety of parasites. After doing some dental work and parasite killing his fasting blood sugar dropped to a normal 98. But his leg cramps kept returning in spite of doing a kidney cleanse. Only after changing his diet to include milk did the phosphate crystals stay away and eliminate his cramps.

Cornelius Edens, age 33, came for his diabetes, although he also had fatigue, digestion problems, and headaches. He had numerous other minor symptoms like chest pain over the heart, soreness in testicles, etc. His diabetes was only diagnosed one year ago. he was on insulin (Humulin™ 20 long acting plus 6 regular units) in the morning. He often omitted the evening dose which should be 14 long plus 6 regular. This brought his fasting blood sugar down to 166 in the morning.

He had pancreatic flukes in his testicles besides in his pancreas. His aflatoxin level was very high; he was told to stop eating grocery store bread, eat bakery bread only. He had silver, nickel and very high levels of gold–probably all three coming from his gold crowns–he was to have them all replaced with composite. He had fiberglass in his pancreas. Wood alcohol was present in large amounts in his pancreas as expected. He was to stop drinking all store bought beverages, whether frozen, powdered, or ready to drink. He did not test positive to benzene, propyl alcohol, *Salmonellas*, *Shigellas*, or *E. coli*. But he had *Staph aureus* growing at high levels in his pancreas. The usual source for this is in teeth, so he was told to ask his dentist to search for cavitations and clean up hidden tooth infections while replacing the gold crowns. He was to find the exposed fiberglass and seal it off. He was to zap parasites every other day for 1 month. His supplements were vitamin C (1 tsp. with each meal for constipation; then 1 tsp. daily), vitamin B-50 complex (2 per day), thioctic acid (2 per day). He was to start the Kidney Cleanse recipe for his testicle problem, and after 6 weeks do a Liver Cleanse. He was to follow up with us in 3 weeks, after dental work was completed. He never returned. Four months later we received a phone call he was too embarrassed to make himself. He needed no insulin and was doing fine. Another 3 months later he was still doing well, and off insulin.

Prediabetic

Alyce Dold, 64, came because she was worried about her blood sugar and chest pain. Indeed, a blood test showed her fasting blood sugar to be 136, just beginning to show insufficient insulin production by her pancreas. She had pancreatic flukes and wood alcohol there. Also *mumps virus* and *HA virus*. She had six more solvents accumulated due to eating raisin bran and other cold cereals each day. She was glad to be forced off this routine: she switched to 2 eggs every other day with biscuits or bread (not toast) and cooked cereal in between. Her chest pain was due to dog heartworm and *Staphylococcus aureus* bacteria that originated at teeth #16, 17, 1, 32. The worms and *Staph* were killed with a frequency generator. She was referred to a dentist for cavitations and started on kidney herbs. Two weeks later, there was still a little residual heart pain due to *Staph*; dental work was not yet done. She was given chromium (600 mcg per day) to help her insulin regulate sugar. Her LDH (See tests) was still high from the recent heart stress, no doubt. But she had accomplished a lot and planned to get it all done.

Diabetes Of Childhood

The problem is the same for diabetes of childhood as for diabetes of later onset, but much easier to clear up, provided the whole family cooperates.

Wesley Evanco, age 6, had his onset at age 11 months. Prior to that he had chicken pox and his baby shots. He had pancreatic flukes and their reproductive stages in his pancreas as well as wood alcohol. He had no other solvents accumulated. The problem was clearly due to these two enemies of his small body. Wesley's father did not have pancreatic flukes but his mother had very high levels, along with wood alcohol. She had to clear hers up before Wesley could recover. She chose not to. She couldn't believe the connection. Our most persuasive efforts were not successful. We hope Wesley can forgive us all in due time.

Herpes

The *Herpes* family of viruses includes *Herpes Simplex Virus* (HSV) 1 and 2, *Epstein Barre Virus* (EBV), *shingles or Chicken Pox* (*Varicella zoster*), *Cytomegalovirus* (CMV) and some newly discovered ones.

Herpes simplex virus 1 is the virus that breaks out around or inside the mouth. We call it a cold sore because it often follows a cold in childhood. As children, we may get HSV 1 once or twice each winter. When they come more often the child's immunity is low. Adults who get repeated attacks also have low immunity (this is obvious from a blood test where the white blood cell count is less than 5,000 per cu mm).

Herpes simplex virus 2 breaks out in the genital area. It is often blamed on promiscuous sex but I believe it has quite different origins.

I believe the virus is introduced to our bodies by another large parasite. Perhaps it is pinworms, or *Ascaris*. Perhaps it is a tapeworm stage, picked up from dust and dirt.

I have some evidence that it is released from dog tapeworm stages when these are being killed by your immune system. It is probably the same tapeworm stage as releases the *Adenovirus* (cold virus) which would explain why children frequently get "cold sores" during or after a cold. (See Curing The Common Cold, page 357.)

Nevertheless, you would not get sick from these viral releases unless your immunity was lowered.

Herpes lives in your nerve centers (*ganglia*) and it is from here that you can be attacked after the initial infection. Evidently your immune system can destroy them as quickly as they emerge. But a meal of **aflatoxin** or other **moldy food** suddenly "gags" your white blood cells and lets a viral attack happen.

The viruses can also be "triggered" which lets them out of hibernation (latency) to multiply and travel along the nerve fiber to the skin. Triggers are things that put these nerve centers to work: sudden cold and heat, trauma from chafing and friction. Never drink water with ice cubes in it. Never eat hot soup with a metal spoon. Don't tweeze hairs.

Begin your prevention program by raising the immunity of your skin; this means removing all toxins from the skin. Use only natural lotions, softeners, cleansers on your skin made from recipes in this book. Health food brands are not superior. This will get rid of nickel, chromate, titanium, zirconium, aluminum, and benzalkonium from your skin and probably your whole body! Do laundry with borax and washing soda, only, to eliminate commercial detergent as a source, too.

As soon as you feel that warning tingle or sensation of *Herpes*, zap or use a frequency generator at 293 and 345 KHz (HSV 1) or 360 and 355 KHz (for HSV 2). Immediately take a **cayenne** capsule and 8 **lysine** tablets (500 mg each). The cayenne slows down travel of the virus along the nerve.

Attacks probably occur when the triggers act at the same time as an immune drop occurs. Immune drops happen frequently due to eating moldy food. Of course, mercury in amalgam tooth fillings keeps immunity chronically lowered. Many persons report they stopped breaking out with *Herpes 1* after replacing their amalgams. Stopping wearing tight synthetic underpants helps reduce *Herpes 2* out breaks.

When you get an outbreak, mop up a droplet of the blister fluid and prepare it as a specimen for yourself. If you search for it in your white blood cells when your attack is over, it will not be found because it is in hiding inside your nerve cells. Zapping does not reach them inside your own cells. Nevertheless, you can totally eliminate them by repeated zapping provided you kill them at their earliest warning. Evidently, they haven't multiplied yet, so gradually their numbers go down.

Even after you have been *Herpes* free for a long time, stick to your preventive principles. Avoid the trigger foods, peanuts, and chocolate, or be ready to zap. Avoid cold wind or direct sunlight on your face. Don't eat abrasive or acid foods like popcorn, nuts, toast, crackers, candy, citrus. Although you may stop the virus in its tracks by zapping, healing the lesion takes time. Keep the skin softened with a cornstarch or sodium alginate recipe (see *Recipes*).

A lysine mush helps too: crush a lysine table with a large wooden spoon, add a pinch of vitamin C powder and a pinch of zinc oxide. Save part of this mixture for later use. Wet a small bit of it with a few drops of water to make a paste. Apply to lesions.

Bazezew Hailey, 38, started breaking out in the genital area after a period of antibiotic use. She was started on lysine (500 mg, 8 a day), cayenne caps (one with each meal). She was to fill her prescription for Zovirax™ from her doctor, though. By the time she had it filled, the next day, her lesion had stopped enlarging, and she could reduce her supplements. Her skin was full of mercury, lead, strontium, aluminum. Her white blood cell count was low, showing low immunity. Her ratio of segmental to lymphocyte white blood cells was low, evidence for a chronic viral condition. Her bone marrow contained lead and radon. She stopped using toothpaste (strontium), salt, deodorant, detergents (aluminum). She got the metal out of her mouth and eliminated her radon problem by opening crawl space vents. The lead source was never found, but after she moved to a new house it went away. *Herpes 2* stayed away. A year later she still hadn't opened the Zovirax™ bottle.

Fatigue

Fatigue, whether minor or extreme, is always associated with blood sugar disturbances. The more disturbance, the more fatigue. We have three organs that do most of the sugar regulating: our **adrenals**, the **liver**, and the **islets in the pancreas**. In severe fatigue, that keeps you partly bedridden, all three organs are heavily parasitized. *Epstein Barre Virus* (EBV) is running amok in your body, as a rule, even when clinical culture techniques cannot find it. With your frequency generator and Syncrometer you can find it emitting at 380 KHz. Killing the viruses is not as important as killing the larger parasites and getting your organs functioning for you again. The viruses will go away by themselves.

The liver's role in blood sugar regulation is to get it out of storage when needed. When sheep liver flukes have taken up residence they're spewing their chemicals as well as their own bacteria and viruses into your circulation; it is surprising that the liver has any sugar in reserve and can function at all. Sheep liver flukes are commonly seen in fatigue syndrome cases.

The adrenals (the outer layer called the cortex) help to regulate the blood sugar in a complex way. Some adrenal factors influence the thyroid which is another energy-related organ. Both adrenals and thyroid have toxic buildups in fatigue cases! Their work is hampered.

The heart of sugar regulation is in your pancreas in the tiny islands of cells that secrete insulin, called the *islets of Langerhans*. Here we always find the pancreatic fluke in residence. It actually breeds there when wood alcohol accumulates in it. Wood alcohol is a common pollutant of food, even in artificial sugar which is often recommended to replace real sugar and spare the pancreas. There is wood alcohol in store-bought drinking water, fruit juice, powders meant to be stirred into beverages, even if they are health food varieties. It's probably being used to clean tubes and hoses in the manufacturing plant. The only beverage you can safely buy (not safe unless you sterilize it, though) at a grocery store is milk. Make your own beverages using recipes in this book.

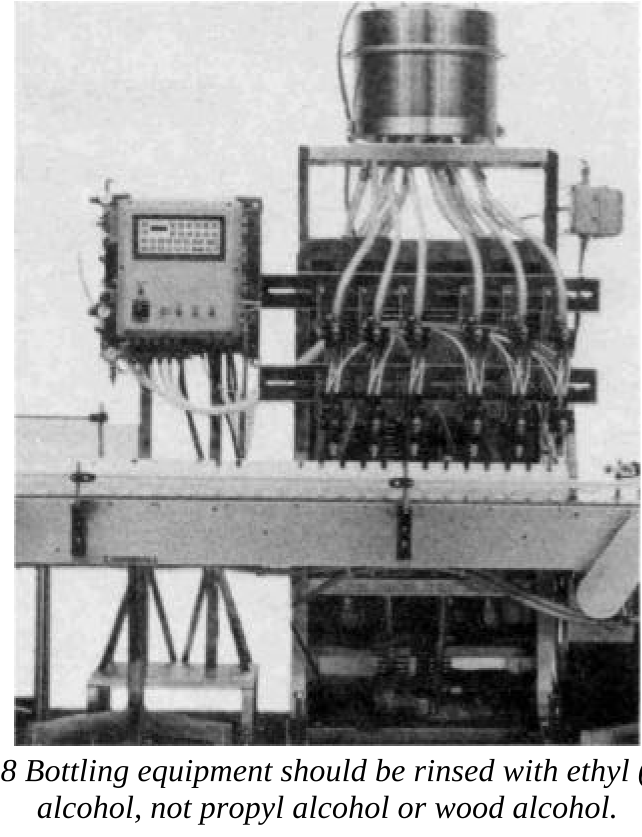

Fig. 28 Bottling equipment should be rinsed with ethyl (grain) alcohol, not propyl alcohol or wood alcohol.

Your first step toward curing your fatigue syndrome is to kill the pancreatic fluke and all other living invaders of the pancreas, liver, adrenals and thyroid. Use a zapper. Drink milk or buttermilk for several days afterward to provide lactic acid for the "good" bacteria to feed and recover on.

Your second step is removing metal from your mouth, particularly gold. Gold from teeth and jewelry readily goes to the pancreas! Do everything in the *Easy Lifestyle Improvements* chapter.

Your energy can bounce back in a few weeks by attending your liver, adrenals and pancreas. Help these organs heal and grow strong again. Avoid food molds (read Moldy Food, page 381). The most powerful assistance to the liver is a cleanse. This will eliminate liver viruses such as EBV and *Cytomegalovirus* (CMV). Be patient. Do it on a schedule until you have over 2,000 stones out. Take **vitamin C** at least 3 gm/day, to help both liver and adrenals. Take **B_2** and **B_6** to help adrenals and kidneys. Take glucose tolerance factor, **chromium** (two 200 mcg tablets 3x/day). See *Sources* for all of these. Take these supplements for three weeks, then cut the dose in half, and take on alternate days only, as a hedge against possible pollution in these.

Although your energy may be normal in three weeks, you are at higher risk for fatigue than the average person. Reinfection with anything will put the new parasites right back where the old ones were. Other bacteria, solvents and toxins will head for the pancreas, liver and adrenals again because these are weakened organs. It could take two years to build your health to its previous level, but is well worth it to have youth, initiative, and a beautiful appearance again. Going back to school is a good use of your time when your initiative has returned but your physical strength is still not up to housework or a job. When your energy comes back to you, it is tempting to overwork: to clean the whole house or to get into some gardening.

It is better to be moderate today than in bed tomorrow.

Anyone who has suffered from Chronic Fatigue Syndrome (CFS) or recurrent EBV has learned that lesson well.

June Timony, age 38, was diagnosed with Chronic Fatigue Syndrome), EBV and *Candida* around 1½ years ago by her family doctor. She also had a thyroid problem and a high estrogen level (165 pg/ml). She had severe depression at times. Our test showed her body was full of bismuth (fragrance) and silver (tooth fillings) especially in the ovaries. She cleansed her kidneys and killed parasites but could not make up her mind to do the expensive dental work.

Janice Brown, age 21, had EBV with chronic fatigue and depression along with a dozen more symptoms. Her skin, kidneys, breasts, brain, ovaries and pancreas were all loaded with mercury, platinum and other metals. She was full of radon and bus exhaust and the plumbing was shedding cadmium. Rather than do all this she and her husband decided to move. Before the moving date arrived she had cleansed kidneys, killed parasites and done dental work and was feeling noticeably better. Then they moved. She immediately was very fatigued again and worried that the move had been in vain. This time she had a liver full of Salmonella and a return of phosphate crystals in her kidneys. But it was easy to clear up and it was a very useful lesson to her to avoid unsterilized dairy products. She was much more careful after this.

Dee Safian, age 36, came especially for her low energy. She had EBV once. It left her extremely nervous. We advised going off caffeine but not substituting for it by drinking decafs. Her tissues were full of arsenic from pesticide; her urinalysis showed kidney crystals and her eosinophil count was high 5.5% (parasites). She had sheep liver flukes and stages in her pancreas due to a buildup of wood alcohol there. In four months after killing parasites and doing a kidney cleanse she was much improved.

Josefina Linzer, age 32, came for her fatigue and depression. Her tissues were full of arsenic, PCBs, chromate (eye makeup), mercury, yttrium, radon and terbium. She needed tooth metal replacement but could not schedule it immediately. She had *Ascaris* and pancreatic flukes in her pancreas and reacted to sugar in her diet quite strongly, so avoided it. She also had *Strongyloides* and *Trichinella*. She killed all these with a frequency generator and started on the kidney herbs. Her wood alcohol buildup was coming from carbonated beverages. She went off. In 6 weeks she had done everything except the mercury removal and was feeling much better.

Brigette Dawn, age 21, had Chronic Fatigue Syndrome with EBV, along with other problems. She cleaned her home and cleansed kidneys, killed parasites, and did two liver cleanses. Still the fatigue would return two days later. Meanwhile, though, her infertility problem got solved (she got pregnant) and this encouraged her to continue the battle against fatigue after the baby was born.

Hector Garcia, age 14, was getting gamma globulin injections every three weeks for his chronic fatigue syndrome. He had pancreatic flukes in his pancreas, sheep and human liver flukes in his liver and intestinal fluke in his intestine. He had a buildup of benzene, propanol, and carbon tetrachloride as well as aflatoxin from his granola breakfasts. He also had *Candida* and measles in his white blood cells. He killed parasites with a frequency generator and went off the solvent polluted items in the propyl alcohol and benzene lists. He immediately (in 20 days) felt much better.

Dana Levi, age 16, had chronic fatigue syndrome and dizziness; he was not in school. He had pancreatic fluke in his pancreas, sheep, human and intestinal flukes in his liver! Both benzene and propyl alcohol were present in his immune system. As soon as the parasites were killed (with a frequency generator) and he changed a lot of his products, he felt better but soon lost his improvement. At the next visit, our tests showed a buildup of vanadium (from burning candles in his bedroom). When he stopped this, he was better again for a while only to relapse again. His propyl alcohol level was up as was aflatoxin. But getting a taste of normal energy gave him the determination to get himself well! He embarked on a liver cleanse program and still more careful food selection.

Dennis Dillard, age 16, was beginning to do poorly at school due to fatigue. He had to stop athletics although his regular doctor pronounced him well. He was also getting chronic sinus infections. His lungs and trachea had accumulated seven heavy metals: vanadium, palladium, cerium, barium, tin, europium, beryllium. His fasting blood sugar was low (73 mg/DL) and LDH very low (90 u/L; it should have been 160). The body makes LDH in response to lactic acid levels. When the muscles aren't making much lactic acid from their normal metabolism, LDH levels will fall too. Was he developing a muscle disease? The gas leak was fixed (vanadium), the garage was sealed off from the house to eliminate barium and beryllium but the other toxic elements came from his dental retainer. As soon as his retainer came out, and they stopped using flea powder on their dog, his energy became normal and sinuses cleared up. A year later he still had not been ill.

Evelina Rojas, age 12, was having extreme fatigue with mood problems and sudden fevers. She killed *Ascaris* and sheep liver flukes with the parasite program but promptly got them back due to a benzene buildup I believe due to using products containing an herbal oil. Her high levels of *Streptococcus pneumoniae* (cause of fevers), *Staphylococcus aureus* and *Nocardia* could not be eliminated until her three baby teeth (with root canals) were pulled. After that, she was well.

Elaine Perkins, 48, came specifically for her low energy and nervousness. She was toxic with arsenic, a substance that replaces energy with nervous excitement and exhaustion. She also had a backlog of antimony (using baby oil), aluminum, rhenium (hair spray), benzalkonium (toothpaste) and radon. In four months, she had the arsenic and three other toxins eliminated and already had more energy.

Neil Youngblood, 53, was so fatigued, he had to brace himself even while sitting. A blood thyroxin level (T_4) of 1.0 instead of the normal 7.5 mcg/DL explained his fatigue. He had a mouthful of assorted dental metals which were accumulating in the thyroid, inviting viruses, particularly CMV, into it. He had to be on four grains of thyroid to feel near normal. After he had all his tooth metal removed, he only needed one grain to feel O.K. This encouraged him to clean up more of his body.

Scott Pennington, 50ish, was on thyroid medicine for his hyperthyroidism. He had been on iodine-radium earlier. He had the *miracidia* of the intestinal fluke, sheep liver fluke, and pancreatic fluke in his thyroid! Adult human liver flukes were also there! His thyroid was toxic with iridium, nickel, tellurium and mercury (metal tooth fillings) and decane, TC Ethylene and pentane solvents. He had been drinking a great deal of regular tea, which let oxalate crystals deposit in his kidney and slow down the excretion of toxins. The parasites were killed with a frequency generator, he changed his diet to get rid of solvents. In two weeks he was feeling better, had more energy and better sleep. This encouraged him to do the dental work.

Skin Problems

Sebaceous Cysts

Your body, in its wisdom, keeps toxins together and out of harm's way by making a cyst out of it. All the *sebaceous cysts* I have seen are filled with *polychlorinated biphenyls* (PCBs).

Get rid of PCB sources in order to clear PCBs from your body. This can take 6 to 12 months. Change all detergents (for dishes, laundry, and body use) to borax and/or washing soda.

Commercial detergents are wonderful cleaners...

but are unquestionably toxic. Whether you have cysts or not, it is always a good idea to use borax and washing soda instead.

If you have a particularly visible cyst, try to poultice it to the surface. Washing away some of it would be a great help to your body. Read herb books on poultices or just take whatever wild leaves you can gather; make a mush out of them with a blender, mix with honey or homemade skin softener to make it spread like peanut butter and cover the cyst with it night and day, covered with a piece of plastic. The wetness has drawing power. The plant juices have other benefits. If you can't poultice and must rely on kidney excretion, be sure to take kidney herbs at the same time.

Skin Rash

Rashes can be caused by many things. To find the cause, follow a logical pattern. It may be due to:

- HIV
- Yeast and Fungus
- Allergy
- Childhood Diseases (rubella, etc.)

HIV is the most alarming possibility. Eliminate it first by testing for it. It's often too early for a clinical test to be positive for HIV, so use the Syncrometer (page 457). I often see the rash disappear within days of eliminating the virus.

Probably the most common cause of rash is **yeast**. *Candida* has a resonant frequency of 384-388 KHz. If you test positive for it, stop all commercial soap and detergent for all possible uses. Zapping *Candida* may drive it away for a few days. The fungus is hosted by another parasite but finds your skin quite satisfactory for a home, at least while your skin immunity is low. It may be low from wearing metal jewelry, having metal tooth fillings, aluminum (from lotions and soaps), cobalt (from shaving supplies), and zirconium (from deodorant.). When all these are removed, the skin will dry up quickly in open air or under a heat lamp.

In babies' skin, immunity is low from diaper chemicals (mercury and thallium), wipes, and soap and detergent chemicals. Tight fitting, mercury treated diapers are a modern atrocity. Always line them with a tissue and keep them loose fitting.

The skin that has rash or fungus should be dried with paper towels, unfragranced and uncolored, in order not to contaminate the cloth towels, and thereby transport the tiny infectious spores to other skin locations. You may use skin healers.

Only the skin healers given in *Recipes* are safe to use. Keep them refrigerated when not in daily use. Commercial lotions, including health varieties, all contain toxic ingredients.

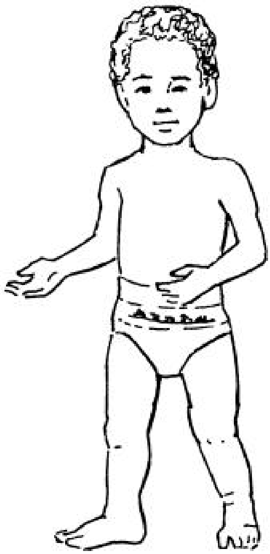

Fig. 29 Tight diapers are a modern atrocity, forcing mercury and thallium into the baby's sponge-like skin.

Many adult rashes are due to **allergies**. Allergy to nickel is common, the reaction is easiest to see under rings or watches. The metal is pulled into the body for elimination. Since nickel is used by so many bacteria, especially urinary tract bacteria, it doesn't get eliminated, it gets taken up by bacteria. It also piles up in kidneys, adrenals, bladder and prostate where bacteria thrive on it. Strangely, it also accumulates in the male scalp (and in women's scalps who have male pattern baldness).

Allergy to strawberries, perfume, deodorant or chlorinated water, however different they are, can all be expressed the same way, in a rash. The liver has refused (been unable) to detoxify the chemicals in these items and allows them to circulate in the body. Not for long, though, since great damage could be done to brain and other tissues. Mercifully, the skin grabs these chemicals into itself. Ultimately, it must still be removed. The allergic reaction and your immune system come to your skin's rescue, although bringing you discomfort.

Why didn't the liver detoxify the chemicals given to it? The answer is based on a simple experiment. Try cleaning your liver (page 552) several times or until 1,000 bits of refuse have been washed out of the bile ducts. This relieves the back pressure on that part of the liver, and allows it to do its work again. The physiological details are not understood. Fortunately, the results are instantaneous. The day before the liver cleanse you would never eat a strawberry or peanut for fear of a reaction. A few days after the cleanse, your body "knows" which food it might tolerate and as you try a bit, you notice no reaction, no rash This simple experiment suggests that the liver couldn't detoxify these foods because of the refuse in its ducts.

Each liver cleanse "cures" a different set of allergies suggesting that the liver is compartmentalized—different parts having different duties. It follows that by getting all the stones out all your allergies will disappear. Experience shows this to be true, although it can take two years to carry out such a program.

Meanwhile, avoiding the offensive food or product is very important. It is quite destructive to bathe the brain in strawberry chemicals or your toes in maple syrup chemicals. Stay off allergy-producing foods and products even if you can tolerate a little or can be "desensitized" to them with shots or homeopathic methods. Use these methods for relief, not license to continue using items that tax your body.

Certain **childhood diseases** produce a rash and this can be diagnosed by testing for the suspected disease with a slide or culture of it. Then use a zapper to kill both the bug and any larger parasites that may have brought it in.

Hives

Sometimes your body will break out in hives instead of a rash. Again, gallstones in the liver and gallbladder are responsible. Do the Liver Cleanse.

Acne

The more severe cases of acne cover parts of the body as well as face and can pit the face so badly there is hardly any clear skin left.

In ten days you can reverse this, so most of the skin is beautifully clear. There are bacteria involved and skin oils of youth feed these bacteria. Acne has been extensively studied by scientists. Perhaps the true culprit was too big to be seen with a microscope or too small (antigen) to be recognized or just too unimaginable. I inevitably find *Trichinella*, one of the four common roundworms that infect humans.

Test yourself to *Trichinella* with the slide called “larvae in tissue.” Search for it in your skin. It is generally believed to reside in muscles, especially the diaphragm, but in acne cases it is in the skin.

If you only have a frequency generator, set it at 404.5 KHz and at neighboring numbers, extending 5 KHz on each side, to be sure to include all eggs and other molt stages. Or zap.

Psoriasis, Eczema

Psoriasis and eczema are both caused by *Ascaris*. Their molting chemicals are quite allergenic; perhaps it is these that are affecting the skin. Since pets pick these worms up daily, there is chronic reinfection in families with pets. Keep zapping.

Bernadette McNutt, 34, had acne on her back, chest and face. She had been treated since teen age with ultraviolet light, Retin A, and antibiotics. She already had a history of shingles and fungus, two more skin conditions. Her skin was toxic with strontium and her kidneys had cadmium, silver and beryllium deposits inhibiting excretion. She had only one parasite, extremely high levels of *Trichinella*. Her children were also infected as was the cat. In spite of using parasite herbs for months she got no improvement until the baby was out of diapers. Then she cleared up.

Royce Hamilton, 17, had acne so dense on his face there was no good spot the size of a dime anywhere. He had it about one year. His urinalysis showed "amorphous" crystals (stones of all kinds) and a trace of protein. He had *Trichinella* worms throughout his organs. He was very fatigued. He was started on kidney herbs so there would be good excretion after killing the *Trichinella*. His thyroid and kidneys were full of zirconium and titanium from all the lotions he used for his skin. He didn't need deodorant. Evidently even his armpit bacteria had been affected. It took four months to clear his *Trichinella* although there were no young children or pets in the house. His face was beginning to heal, but three months later he had a recurrence, although his parent was not a carrier. After this, he cleared it up again and his face looked as beautiful as a child's.

Evan Knight, 36, had psoriasis at elbows and knees from age 9 but now it was spreading to his fingers and scalp. He occasionally had bronchitis and puffy eyelids, indicative of *Ascaris* but at the time of his visit he had *Trichinella* fluke stages and *Echinostomum* in his skin. He was started on the parasite program and in three weeks it was clearing instead of advancing. He switched to milk for his beverage to raise his immunity and removed the arsenic, formaldehyde and thulium (from his vitamin C) by doing the necessary cleanups.

Gerry Chastain, 41, had a red nose, erupted along the sides. He had been on sulfa drugs and Emycin.^{TM} He had *Leishmania tropica*. He killed it in the office with a frequency generator and got immediate improvement but four weeks later it was back. He had four solvents built up in his body: benzene, TCE, TC Ethylene and hexanedione. This situation would make recovery impossible since he was no doubt reinfecting himself. He also had titanium, platinum and silver accumulated in his tissues and needed to replace his dentalware before expecting a permanent cure.

Floyd Oldham, 50s, was getting pimples on his nose. His whole face was red and flushed looking. His bowels were loose and he had some urinary urgency. He harbored two kinds of *Leishmania* (*braziliensis* and *tropica*). He killed the *Leishmanias* with a frequency generator and started himself on the kidney herb program. Five weeks later the pimples were gone but general redness had reappeared. The *Leishmanias* were still gone but this time he was toxic with cobalt.

Cobalt is known to promote skin cancers and also heart disease. This was discovered decades ago when an outbreak of heart disease occurred in England. It was traced to a pub (where they all partook) where cobalt was added to the beer to make the foam rise higher! In fact, the foam will just stay in place like a hairdo, if you add cobalt. It was made an illegal additive. Gradually, it has crept back into consumer's products: first toilet cake (blue), then window washer (blue), then dishwasher detergent, and now even mouthwash. If you see a blue colored product, **stay away from it**. It accumulates in the heart and skin. People with skin cancers often have cobalt build up.

Floyd stopped using all these products and cleared up again.

Grethe Driscoll, middle aged, wore tons of make up, so skillfully applied that scars from a face lift could never be detected. When she had minor breakouts, which usually occurred while away on a trip, it seemed like a catastrophe. She tried everything available but could not get to her parasite herbs until she was back home several weeks later. After one week on them (5 day high dose plus maintenance) her complexion was perfect again.

Crofton Thornton, 15, had an embarrassing case of acne. He had *Ascaris*, hookworm and *Strongyloides* (he also had migraines) all reacting in the skin. He stopped drinking commercial beverages that gave him solvents. He killed parasites electronically and with herbs and got a considerable improvement. But he still had *Strongyloides* one month later. Nevertheless, he had seen the connection and he knew it was just a matter of persistence to a clear complexion. Note, he should have had *Trichinella*—did I miss testing for it?

Warts

Could we get warts from playing with toads in childhood? We don't play with toads anymore. Yet we get warts. We don't know how we get them. But after learning how to get rid of them, you will probably know how you got them. Not all warts are the same. In fact, they might all be different: each one is made up of 5 or 6 different viruses, not just one as we had believed.

Peel a tiny fragment off one of your warts. Prepare it for testing by placing it in a small bottle. Add a few tsp. filtered water and a ¼ tsp. grain alcohol, Label it with the location you got it from: like “left middle finger knuckle. “First, search your body for other locations of this wart (organs that test positive to your sample). You can easily find them in your skin, of course. But also search electronically in your liver, spleen, muscles, stomach, heart, pancreas. **Notice how often they are present in the pancreas**. The pancreas seems to be a wart-virus heaven. Are they in the islets or the rest of the pancreas?

Without a zapper, you will need to find the frequency of each virus to completely destroy it. Attach your frequency generator and search between 400 and 290 KHz. When you find its resonant frequency, kill it by treating yourself for three minutes at 10 volts from a frequency generator. Will your warts fall off?

In a few days one or two of your warts will begin to shred. After a week you may lose one or two completely, and find that several more have become smaller. The remainder are unchanged. Continue to identify and kill them. Notice that they are not necessarily gone from the pancreas or other organs at the same time as they are gone from the skin. Perhaps warts are not the benign entities we have believed them to be. They may, in fact, ride into the body on some common bacteria, like *Salmonella*, or common parasite like pinworms or tapeworm stages.

Zapping doesn't reach all the viruses in a wart either. It takes repeated zappings to start the shredding and gradual killing of warts.

Guy Laird, age 11, had warts on his lips besides fingers. His job was feeding the three outdoor dogs. He was full of *Ascaris*. He had *Taenia pisiformis* and *Taenia solium* bladder cysts in his liver. These were shedding viruses into Guy. He was started on Rascal for six weeks (this was before the zapper was invented). Maybe his benzene buildup was responsible for letting so many parasites (and their viruses) survive and multiply in his body. He stopped using toothpaste, killed *Ascaris* (408 KHz) and some flukes (434 to 421 KHz) and improved his diet. All except one wart came off (without bleeding). He was given different chores, too, to reduce his contact with animals and their parasites.

Georgianna Mills, a middle age music teacher, broke out with warts all over her hands, at least 30 in total. A few months later she was diagnosed with bone cancer; she always wondered if there was a connection. She cleared up her cancer and killed her viruses and bacteria with a frequency generator. Nearly all her warts disappeared. But her indoor pet brought new parasites daily, especially *Moniezia* tapeworm stages. With each *Moniezia* infection (about once a month) she got new warts. She was never able to clear them completely.

I concluded that each wart is actually composed of 3 to 6 viruses and these viruses are distributed throughout our bodies! How satisfying to be able to rid our bodies of them, once and for all even in internal organs. There is a catch. Small remnants of some warts do not disappear in spite of killing most of them. More accurately, they disappear and then reappear in our internal organs. Could this suggest to us their true origin? Could it be a tapeworm stage?

Tapeworm Stages

Our bodies harbor numerous stages of tapeworms. But not the tapeworm itself, which may belong to a dog, cow, or pigeon. Tapeworms lead complicated lives, much like insects with their caterpillars, larvae, larval molts, pupae and eventual adults. Tapeworms shed eggs with the bowel movement of the animal host. The eggs blow in the dust and reside in the earth. A vegetarian animal nibbling vegetation near this filth, or licking dirt and dust off its coat, swallows the eggs. Humans, too, eat plenty of filth by licking their fingers. As children we all eat dirt simply by eating with unwashed hands.

The Jewish society discovered the great importance of washing hands before eating, thousands of years ago. But many of us choose to ignore truths that seem old fashioned. In our own relatively short life times we cannot see the whole picture as well as the prophets and seers of ancient cultures could. We eat plenty of dirt and along with it, the eggs of tape-worms. Dog and cat tapeworms are most prevalent, but sheep, cow, pig, and sea-gull tapeworms are also common.

There is hardly a predator species in existence that doesn't have its own characteristic tapeworm. Whatever animal species you live near, or once lived near, you probably swallowed some of its filth and some tape eggs. The **eggs** hatch in your stomach and the tiny **larvae** burrow into a neighboring organ without any consideration that this is your stomach wall or spleen or muscle. The larva's plan is not to grow into a long worm—that can wait. The larva must simply

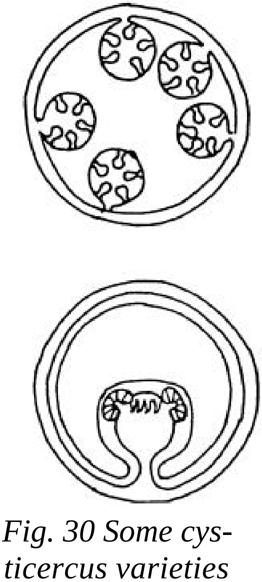

Fig. 30 Some cysticercus varieties (types) have multiple heads.

survive until you can be conveniently eaten! A wolf or a tiger will surely come along! In bygone days it did.

The larva is about ¼ inch long, surrounded by a "sac of waters," like a tiny water balloon. Looking very closely at this sac, called a **cysticercus**, we see a head (scolex), complete with hooks and suckers, turned inside out, inside a bladder.

As the tiger's teeth bite down on the cysticercus, the pressure pops it out. The head is now right side out with hooks and suckers ready for action. Now it grows in the tiger!

It quickly hooks into a loop of intestinal wall so it can't be swept away and begins its growth into a regular long **adult** tapeworm. The tiger is the true or **primary host**. We were merely the secondary or **intermediate host**. Why does the adult tapeworm prefer the tiger instead of us? Only Mother Nature knows. But the best way to get to a carnivore is through its prey.

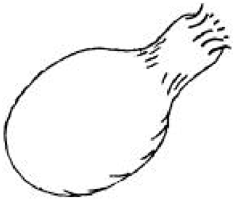

Fig. 31 Emerged cysticercus.

You can find these larval cysts in your organs using slides of the *cysticercus* stage of various common tapeworms. Search in your muscles, liver, stomach, pancreas, spleen, intestine and even brain. **You will not find even little bits of them in your white blood cells**. My explanation for this curious finding is that the tapeworm leaves no debris to be cleaned up by your white blood cells. Evidently your body builds a cyst wall around the larva to tightly encase it and prevent toxins and debris from entering your body. Thus your white blood cells are not alerted in any way. Of course, the larva is much too big to be devoured by tiny white blood cells anyway. Yet, it seems that if a pack of white blood cells had attacked the larva just as soon as it hatched from the egg they would have been able to devour it. Perhaps it enlarges too rapidly. Perhaps our white blood cells are preoccupied. In any case, we begin to load up on tapeworm stages from infancy and by the time we are middle aged we have dozens tucked away in our organs.

Some do die in the course of time. Perhaps their true secondary host is a rabbit or a mouse instead of a human. The short life span of these other hosts might mean that the life span of the *cysticercus* is also quite short, not 40 years! When they die, the white blood cells do clean them up and we can see them in our white blood cells at this time. It can take several weeks for the *cysticercus* to be completely gone by this natural method. During this time, we become ill! Numerous bacteria and viruses spring up, as if from nowhere, in our organs.

Don't be surprised if you are testing yourself during illness to find a tapeworm or two in your white blood cells! It is well worth searching for at such a time. Help your body dispatch the tapeworm stages all together with your zapper. A frequency generator is bound to miss some. Some cysticercus varieties consist of many heads, and each head has even more heads inside it! These might have different resonant frequencies. Only killing them together has the desired effect. Remember bacteria and viruses are released by killing tapeworms, so always follow with a second zapping in 20 minutes, and a third zapping 20 minutes after that. Only then can your tapeworm-related illness disappear.

If you do nothing, your body will be kept busy killing bacteria and viruses as the tape cysticercus wears down and eventually dies. You may not wish to identify all of them (but at least search for *Adenovirus*, the common cold) and just note where you are being attacked: your nose, throat, ears, lungs, bronchi. Internal organs are attacked too. It seldom takes more than three weeks, though, for your body to clean up a tape stage even without any help from a zapper. The attendant illness will be gone by then, too.

Watching these events in your body gives you insight into the very powerful forces at work, called immunity or body defense. The body "knows" a great deal more than we have surmised. There is yet so much to discover.

What initiated the death or dying process of the tapeworm stage in the first place? Has your body been trying all along and finally succeeded? Has the *cysticercus* reached the end of its life span naturally? Have its (the tapeworm's) own viruses and bacteria gotten the upper hand and killed it? Did it accidentally absorb something that killed it?

By taking a herbal combination, Rascal, you can soon find a tapeworm stage in your white blood cells where you could not find it earlier. It is now dead or dying. This proves the effectiveness of Rascal, even though it is slow.

Since we all eat dirt and inhale dust that is laden with dog feces or other animal excrement, we all harbor tapeworm stages, although none may be present in our white blood cells. **Are they harming us?** Perhaps they are living out their lives as quietly as they can in our organs, the way mice or ants try to live in our dwellings. Yet, when tapeworm stages are being killed, either spontaneously by your body or with a zapping device, we see an assortment of bacteria and viruses spread through the body, including the common cold.

Getting rid of the tapeworm stages in your organs seems a very worthwhile goal. Since each of us has been associated with dozens of animal species in our past, we probably have dozens of varieties of tapeworm stages in us. I cannot identify more than a handful due to lack of prepared slides. You can find them without identifying first, though, by listening to their emission frequencies. Their emissions are often extremely weak, possibly due to being encased in a cyst. Search between 510 KHz and 410 KHz. You may wish to "track" them for a while before killing them. You may wish to search for identical frequencies in your pet's saliva. Or you may wish to dispatch them as rapidly as possible. Use the zapper, not a frequency generator. Remember to "mop up" after your tapeworm killing by zapping again to kill bacteria and viruses that have been released from the tapeworm.

You may be disappointed not to feel any different after ridding yourself of numerous tapeworms and their pathogens. Evidently, the tapeworm stage itself doesn't make you sick; it is simply there like a wart is there, without making you sick. Its viruses can make you sick. Depending on which virus it is, it can make you very sick or not sick at all. Different viruses invade different organs. And some of these turn into warts!

The Flu

Influenza is a virus that can cause “the flu.” Does it belong to us as humans or to a larger parasite we are hosting? It is easily transmitted from person to person and in less than a year can spread across the planet. Some flu examples are Influenza A, B, C and Swine flu.

However, much that is called “flu” is actually caused by a bacterium, either *Salmonella* or *Shigella*. If someone in your family is “catching” a flu, test their saliva for the presence of dairy products, implicating the *Salmonellas* and *Shigellas*. Also test for influenza A, B, and C. Children's “flus”, especially when there is a fever, are usually due to *Salmonellas*. Even after zapping it can take an hour for the symptoms and fever to go away.

Go straight to the refrigerator and throw away all dairy products. Throw away all milk, cheesecakes, buttermilk, cream, butter, yogurt and cottage cheese, deli food and leftovers. You may wish to identify the food source of your family's bacteria first, and save the uncontaminated food. Use the sick person as a subject, searching for foods that appear in her white blood cells (or search their saliva sample for the food offender). If the flu is “going around” your neighborhood, you might wish to tell some of your neighbors which foods you found were contaminated. They may have purchased the same food! Obviously, when a contaminated shipment of dairy products arrives in your grocery stores, quite a few people will be consuming it, setting the stage for a "bad flu" that "goes around".

Why can some people eat contaminated food without getting sick? Maybe their *Salmonellas* don't multiply rapidly. Maybe their stomach acid is strong enough to kill most of them. Maybe their bowel movements are frequent enough to expel them quickly. Maybe they haven't been on frequent treatments with anti-strep antibiotics. The answer is not known yet. After a serious bout with *Salmonellas* or *Shigellas* the body does not completely clear itself of them. They stay in hiding somewhere. When a new batch of bacteria arrives, even though very sparse, as in 1 tbs. of milk, the two subtypes can hybridize and produce a much more vigorous offspring. This is called *virulence*. You are made much sicker by more virulent subtypes of bacteria. You may have diarrhea, vomiting, illness. Especially if you believe you have "lactose intolerance," pay attention to *Salmonella* and *Shigella*.

If your flu is due to an influenza virus, kill it with your zapper. Some family members may prefer to take a homeopathic flu "remedy" such as Oscillococcinum^{TM} or Flusolution.^{TM} Others may take herbs. These probably act by prying the viruses out of your cells' gateways and channels so the white blood cells can easily devour them.

But flu due to *Salmonella* is not easily zapped away. Remember, the zapper current does not penetrate the bowel contents, which is exactly where *Salmonella* lives! Besides zapping to clear them from your tissues, you must eliminate them from the bowel by using the Bowel Program (page 546).

Lugol's iodine solution (see *Recipes*) can quite quickly get rid of Salmonella throughout the body. Use 6 drops (small drops from an eyedropper) in ½glass of water four times a day. If no more *Salmonella* is consumed, it will be vanquished in a day or two.

If your flu brings you a fever, use Lugol's.

Fever

Fevers are there to help your body fight the invaders...up to a point. Don't use a fever medication unless the body temperature goes over 102°F, and then only enough to bring the fever down a bit.

Most fevers, especially **"fevers of unknown origin"** are due to *Salmonellas* and *Shigellas*. Your body may be young and strong enough to kill them but not strong enough to kill an everlasting supply of them coming from dairy foods you eat on a daily basis. Stop eating salads at restaurants immediately. Stop eating food made by others' hands unless it is sterilized. Stop eating dairy foods until you have cooked them. Stop eating those that can't be cooked.

Sam Ellis, age 7, had two episodes of severe abdominal pain with fever lasting two weeks. He got hyperactive with milk products, had a frequent cough and stuffy nose. (Here the picture is quite clear. The milk products were bringing him *Salmonellas*, *Shigellas* and other bacteria which grew in his intestine to produce pain. But why only Sam and not his brothers?) Sam had a buildup of benzene from using bathroom soap containing a special herbal oil. Sam also had hookworms, intestinal fluke, and rabbit fluke, probably due to his lowered immunity from the benzene. Then his mother boiled Sam's milk, removed the polluted soap (she planned to use it herself!) and killed his parasites with the parasite herbs. His fever went away and stayed away. He said he enjoyed all this because now he "could play after school" without a stomach ache and he wasn't being sent to the nurse's office because of a fever. Notice the bacteria causing the temperature went away by themselves, probably due to the return of his normally strong immune system.

Jalene McCormick, 46, had been passing lots of kidney stones (hundreds) for years and had a temperature most of the time for which she was on antibiotics. It took her six months on our kidney herb recipe to dissolve and pass so many they no longer showed up on X-ray, and to stop making them. Then her fever left, not to return.

Kristen Jane Johnson, a young mother, had recurrent fevers and fatigue but, besides EBV, her doctors found no cause. We found HIV virus and a lot of bacteria and parasites. The fever-causing bacterium was *Salmonella*. To stop her *Salmonella* attacks she had to raise her immunity besides boiling all dairy products. Moldy foods (pasta) and lunch meats (benzopyrenes) were the source of liver toxicity. Each new *Salmonella* attack immediately invaded the liver so a vicious cycle was set up. When she stayed meticulously on the parasite program, meticulously off unsterilized dairy products, and meticulously off benzene-polluted items, she cured herself of fevers and night sweats and the HIV infection. Perhaps in two years the liver will have recovered enough to kill *Salmonella* that enter it, but she is not taking any chances till then.

Although Kristen was eating food polluted with both *Salmonellas* and *Shigellas* she only “picked up” *Salmonella*, never *Shigella*! Why is that? In contrast, people with multiple sclerosis “pick up” *Shigellas*, not *Salmonellas*.

Multiple Sclerosis & Amyotropic Lateral Sclerosis

Multiple sclerosis (MS) is a disease of the brain and spinal cord. It is called *lateral sclerosis* if the disease is mainly in the spinal cord.

It is caused by fluke parasites reaching the brain or spinal cord and attempting to multiply there. Any of the four common flukes may be responsible. Kill them immediately with your zapper or a frequency generator (434 KHz to 421 KHz). They cannot return unless you reinfect yourself. Stop eating meats, except fish and seafood. All meats are a source of fluke parasite stages unless canned or very well cooked. Pets and family members are undoubtedly carriers of the same flukes, although they do not show the same symptoms. Give away your house pets. Don't kiss your loved ones on the mouth. Make sure your sex partner has also been freed of fluke parasites.

The most important question you must be able to answer is why did these parasites enter your brain and spinal cord? When the brain contains solvents, it allows flukes to multiply there. The solvents, **xylene** and **toluene** are common brain solvents always seen in MS cases. Evidently these solvents accumulate first in the motor and sensory regions of the brain, inviting the parasites to these locations.

Xylene and toluene are industrial solvents used in paint and thinners. It is also a pollutant of certain carbonated beverages (I found it in 7-Up,™ ginger ale and others that I tested). Stop drinking them.

All MS cases I have seen also harbor *Shigella* bacteria in the brain and spinal cord. These come from dairy products. They are manure bacteria. Be absolutely meticulous about sterilizing dairy products. Even one tsp. unsterilized milk added to scrambled eggs could reinfect you. Not even heavy whipping cream or butter is safe without boiling. Kill bacteria every day with a zapper. *Shigellas* produce chemicals that are toxic to the brain and spinal cord. Eliminating *Shigellas* brings immediate improvement.

All large parasites like flukes have their own entourage of bacteria and viruses. Perhaps it is these that initiate the brain's reaction, which is inflammation and scar tissue formation in the outer covering of brain cells and nerve fibers. Perhaps it is the fluke stages themselves. Your brain is trying desperately to heal these lesions, only to be assailed by a fresh batch of solvent and *Shigellas* and another generation of parasites and pathogens.

The other pollutant associated with MS is **mercury** from dental metal. The mercury that is constantly released in the mouth does not all get excreted by the kidneys or eliminated by the bowels. Some of it travels up to the brain and gets into the spinal cord as well. You will be able to eliminate and excrete more mercury by doing a kidney and liver cleanse. The mercury may itself be polluted with **thallium** which is even more toxic. For this reason mercury removal should be done extra thoroughly to be sure no thallium has been left behind.

If you are concerned about MS-like symptoms, purchase slides of the brain regions, cerebrum, and cerebellum. Or purchase pork brains at the grocery store and snip out a portion of the sensory lobe and cerebellum. Prepare these as test substances (sterilize your hands afterward). Test your daily foods and body products for their presence in these brain areas. Also test for parasites, bacteria (especially *Nocardia* and *Shigella*) and other pollutants such as arsenic and pesticides. If the disease (tremor and lack of sensation) has not progressed too far, you can cure it. In all cases you can stop it from progressing further by cleaning up dentalware, the environment and diet.

Brandi Rainey, age 34, of Amish religious culture, was diagnosed with MS four months earlier after an MRI confirmed it although she had symptoms for many years. She was told she had inherited a gene for it and that Amish folk are particularly susceptible to MS for reasons of inbreeding. She had a constant pain running down the side of her neck, and headache. Her legs were getting too heavy to get up stairs. Our tests showed her brain was full of scandium (tooth metal alloy) and fluoride (toothpaste). Her vision was getting worse; her eyes were full of wood alcohol. She lost no time in getting dentures: there were no teeth that could be saved. She had several bacteria growing in her jaw bone: *Strep G* (sore throat bacteria), *Staphylococcus aureus* (this was raising her pulse to over 100), *Clostridium tetani* (causes great stiffness), and *Shigella* (produces nerve toxins). She killed these with a frequency generator. Five weeks later the pain and stiffness in her neck were gone, her pulse was down to 100, her periods were free of pain and her hands seemed to shake less. She was put on the parasite program plus thioctic acid (2 a day) and histidine (500 mg, one a day to keep nickel levels down)and advised to cook and eat with non metal. Four weeks later her pulse was down to 80, her legs were much better. She zapped four remaining bacteria. Two months later the numbness and tremor had left; her legs still felt tired and her hands sometimes shook but she was quite reassured that MS would not claim her life. Nor had a gene betrayed her. A half year later she was walking and working normally, doing liver cleanses and keeping up her vigilance against parasites and pollutants.

Kendra Welch, 56, was diagnosed with MS a year ago, by MRI. She went to a chelating doctor and this cleared up her temporary ischemic attacks (T.I.A's) which were occurring daily. But she had lost her balance, eyesight was getting worse, her feet and hands stung. Her sister also had MS but nobody else in the family did which baffled her doctor. Her brain tissue was full of barium, europium, gadolinium, and platinum. These are dental alloys, although barium could come from bus exhaust (she wore no lipstick). She was advised to have all metal removed from her mouth immediately. Two days afterward she came into the office without any neurological symptoms. She stated she was afraid to stop her new health program, though, and this was good policy.

Lynne Ceretto, age 15, was diagnosed with an “MS-like syndrome” by MRI. She had intestinal flukes and stages, human liver flukes and *Trichinella* in the brain. She had no tooth fillings. But there was benzene in her thymus keeping her immunity low. She also had propane and asbestos in her brain from leaky pipes and a worn washing machine belt. She, too, was told she had a “bad” gene. She was barely able to walk with help. It took several months to track down the source of benzene—the drinking water. Sometimes it was polluted; sometimes it was not! When the pump was oiled, some of it dripped onto the cement platform. Rain washed it to the center and down the pipe in the well. They eagerly removed the platform, found the oil on the water surface, cleaned everything up carefully, until no benzene could be found which put her on the road to recovery. A year later she had recovered further.

Norma Luellen, a young mother, had tingling, numbness and weakness on the entire left side of her body. She was in process of clinical tests for MS. Her body was full of pentane, possibly from her workplace since 12 other persons working there also had MS (such a situation should make our government eager to jump at investigations)! She had intestinal flukes and their stages, not in the intestine or liver or thymus, but in her brain! We found her home air toxic with bismuth too, probably from cosmetics. In spite of staying on the parasite program she got reinfected with sheep liver fluke, probably from eating hamburgers. She was not able to stop her carbonated beverage habit and frequently showed xylene, acetone, methylene chloride in addition to pentane in her white blood cells.

Shannon Synder, age 44, had been getting more numb over her whole body for several years and was presently considered by her doctor to have MS. Her muscles twitched all night, making sleep impossible, and her hands shook. She had intestinal flukes in the brain (cerebrum and cerebellum) but none in the intestine! The brain also had wood alcohol from drinking Diet Coke.™ She also had bismuth (cosmetics), palladium, copper, samarium, and tellurium (tooth alloys) in her brain. She began to improve enough to be off Prednisone by her 10th day of the parasite program.

Erica Blake, age 41, was diagnosed with MS two years earlier although her symptoms went back 13 years. She was on Prednisone™ but her balance was getting so bad she had to be in a wheelchair. Chelation treatments kept her from deteriorating further. Her brain was full of gasoline; she used to work at a gas station and now was getting it from the attached garage. Her hands and feet were completely numb. She had 5 root canals extracted and a few days later was able to stand. She could now walk with a cane. She had human liver flukes, sheep liver fluke and *Trichinellas* and dog tapeworm stages in her cerebellum (motor control center). After killing parasites and starting to take thioctic acid (4 a day) and cleaning up her environment she improved enough to drive a car again, walk without a cane in her home. She regained enough feeling in her hands and legs to do her housework, too.

Kurt Nielsen, age 43, was told he had peripheral neuropathy. His feet were so numb he had to look at the clutch to drive. Also both hands were numb. He was full of kerosene and benzene possibly from fuel oil that he pumped for a living. There were fluke stages in his brain and a dozen bacteria and viruses. He also had mercury and thallium in his immune system which came from tooth fillings. However, he had all his metal fillings replaced two years earlier! He thought there was no mercury left in his mouth! Actually, he had little bits (called tattoos) left somewhere. And they were giving him the classical symptoms: numbness of hands and feet and gradual destruction of his nervous system. His peripheral neuropathy was due to thallium poisoning. The dentist couldn't find tattoos, and he was left in his predicament. (This was before I found mercury and thallium sources in many personal products.)

Duncan Wood, a middle age father of 5 young children, could still slowly shuffle along when he arrived. He could not raise his arms to eat. He had uncontrolled inappropriate laughter every minute. He was diagnosed with MS two years earlier and told he had a “bad” gene. The fact that one child was beginning to show similar symptoms strengthened their belief in the gene theory. Ten days later his inappropriate laughter stopped; he could get his right hand to his face, he walked twice as fast and had very little tremor remaining. Strong chelating treatments obtained at a Mexican clinic had drawn much of the mercury and thallium out of his brain. He killed the flukes and *Shigella* bacteria electronically and stopped consuming unboiled milk. The brain solvents, xylene and toluene were removed quickly, too, as well as asbestos. His fast improvement showed them how important it was to remove the source of these pollutants in his home.

Two days later he regressed considerably which made him feel quite depressed, since his chelating treatments had not stopped. Was there still an unknown factor? It was a return of *Shigella* bacteria! He had inadvertently eaten a non-sterile dairy food: milk added to soup when it was already done cooking! This was a valuable lesson. Nothing else had returned. He was away from the asbestos and xylene from the workshop at home. He zapped the bacteria again and applied greater vigilance to eating only sterilized dairy foods. He recouped his losses in one day.

Then they scheduled their dental work, which had already been done once two years ago! He had leftover mercury and thallium. Now, selecting a dentist with experience in finding tattoos and cleaning cavitations made much more sense to him than it had before. He also planned to do a kidney and liver cleanse after returning home. And to stay out of the workshop until the asbestos-containing belt had been replaced and the furniture painting had been moved to a different building.

Shigella also causes irritability and depression, a frequent problem for MS cases.

High Blood Pressure

High blood pressure is one of the easiest problems to correct without resorting to drugs.

The most important change to make is to stop using caffeine as in coffee, tea, or carbonated beverages. Don't use decaffeinated coffee or tea either because of the solvent pollution in them. Switch to hot milk or hot water if a hot beverage is desired, or any of the beverages given in the recipe section. If being without caffeine leaves you fatigued, take an arginine tablet in the morning (500 mg).

Blood pressure is mainly controlled by the adrenal glands which sit like little caps on top of the kidneys. Whatever is affecting the kidneys is probably affecting the adrenals, too, since they're so close to each other. You must find out what it is.

You could do your search in the kidneys since kidney tissue is available in grocery stores. Adrenal tissue is available on microscope slides. What will you find? Probably **cadmium**.

Search for the cadmium source in your drinking water! Cadmium comes from the metal pipes. In fact, you could scrape a galvanized pipe to get a cadmium test substance. Conducting or storing drinking water in containers of metal is as foolish a practice as eating food off the floor. Water picks up everything it touches simply because it is wet! You may not see what it picked up any more than you can see if it has picked up sugar or salt. The cadmium and other metal is dissolved in the water. The older the pipes the softer, more corroded they are, and the more metal is picked up as the water rushes by. If you find cadmium in your hot or cold water, you will never be able to filter it out. Nor should you switch to bottled water. The amount of cadmium in your clothing from doing laundry with this water is already too much for your adrenals and kidneys.

Change your galvanized pipes to PVC plastic. If you believe you already have plastic pipes or all copper (which leads to leukemia, schizophrenia and fertility problems) you will need to search every inch of plumbing for a very short piece of galvanized pipe left in the system! A piece as short as a 2 inch T or Y can be causing all the trouble.

The toxicity of cadmium, in fact, the high blood pressure connection, has been known a long time. After finding the cadmium start on the kidney cleanse. You might miss the cadmium problem if you don't attend to it first. Also remove all metal from your mouth.

All (100%) cases of high blood pressure I have seen could be easily cured by eliminating cadmium and other pollutants, followed by cleansing the kidneys.

To test whether you still need your blood pressure medicine, wait until your pressure is down to 140/90 or better. Then cut the dose in half. Check it again next day. If it has climbed back up you are not ready; go back to ¾or a full dose of medicine. Try again a few days later. If your blood pressure stays down, cut your medicine in half again (you are now down to ¼ the regular dose) and see if your blood pressure stays improved.

When you are down to 130/80 go off completely. But stay on the kidney herb recipe. At 120/80 try yourself on a few shakes of sea salt. The amount of salt eaten, once the pressure is down, has little influence. In fact increasing salt intake improves energy without raising blood pressure. Take no more than one teaspoon a day (2,000 mg sodium), total, including cooking. Better yet, make a salt that is a mixture of sodium and potassium chlorides (see *Sources*). Mix it for yourself in a 1 to 1 ratio or whatever your taste can accept. The sodium portion could be sterilized sea salt (test and make sure it has no aluminum silicate in it first).

Mold toxins have specific kidney effects! Especially **T-2 toxin**, found mostly in dried peas, beans and lentils. Rinse these thoroughly first, throw away shriveled ones, and add vitamin C to the cooking water. All cases of serious kidney disease show a build up of T-2 toxin. Be extra careful to avoid moldy food (read Moldy Food, page 381).

Bala Cuzmin, age 72, had high blood pressure for ten years but the upper (systolic) pressure remained high in spite of various medicines that were tried. She had three kinds of kidney stones and only one functional kidney. She stopped using caffeine, switching to arginine tablets to get over the let-down. Her diet was changed to reduce phosphate and add calcium, and she took magnesium and Vitamin B_6 to assist the kidneys. She was very anemic and her mean cell volume (MCV) was high due to *Ascaris* infestation. She killed parasites, cleansed kidneys but saw no drop in blood pressure which stayed at 150 to 170 systolic. Her adrenal glands were choked with copper and platinum. She had all the metal in her mouth replaced and promptly saw a blood pressure drop to 145-1 50. Three months later it was at 128 to 133 on half her medicine. She had not been tested for T-2 toxin yet, nor changed her copper water pipes.

Sabrina Patton, 66, had a long list of health problems, including high blood pressure for six years. She was on Corgard™ and diazide drugs which kept it down to 140-160/74-80. She had phosphate crystals in her kidneys and was started on kidney herbs and a diet change to include milk and exclude soda pop. She had high levels of mercury and copper in her immune system. She was feeling so much better after the kidney cleanse that she decided to remove her last fillings and replace her bridge, too, since it was shedding ruthenium. On her way home from the dentist, her ears stopped ringing and soon her blood pressure was down to 126/68. She was still on half a dose of drugs because she was too afraid to go off entirely. But when her pressure stayed down she found the courage to go off completely. This gave her the energy she wanted to play basketball with the grandchildren again.

Rolf Ehrhart, 61, had 80% blockage of heart arteries and high blood pressure for which he was on a Hydropres™ patch, Tenormin™, and Logol™ (diuretic). He had phosphate and uric acid crystals in his kidneys. He was started on kidney herbs followed by the parasite herbs. His *Ascaris* and flukes were zapped. He stopped using store-bought beverages. Then he could cut back on his medicines, measuring his blood pressure daily to guide him. After seven weeks it was down to 140/85, so he decided to do without medicine, a bit early. He was also getting chelation therapy and was now able to walk 2-4 miles a day. His next chore, which he approached gladly, was removal of all metal from his mouth.

Len Gerald, 45, was on Vasotec™ for high blood pressure. He was constantly sleepy; his blood test showed a low thyroid level in spite of being on Euthyroid.™ He was started on kidney herbs followed by parasite herbs. In two weeks, barely into his program, his blood pressure dropped. He had to go off his blood pressure medicine. It stayed at 126/80. He still had some *Ascaris* and other health problems but was highly motivated to clean them up, too.

Glaucoma

In glaucoma the pressure in the eyeball gets too high, putting pressure on fragile retina cells that do your seeing. The first question to ask is: "Is my blood pressure too high?," because there is a link between high blood pressure and elevated eyeball pressure.

Your blood pressure should be 120/80. Your doctor may say 140/85 is "not high." He or she is kindly refraining from giving you drugs until this level of pressure is reached. It is your tip-off, though, that something is not right and you should correct it now, when it is easy, and before other damage is done. Read the section on high blood pressure (page 210) to learn how to reduce it by going off caffeine, checking for cadmium poisoning from your water pipes, and cleansing the kidneys (page 549). Even though your doctor has explained how the tiny tube draining your eyeball is too narrow, you should ask: was it not too narrow before high blood pressure struck? Simply getting your blood pressure to normal is sufficient help for beginning glaucoma.

Antonia Guerrero, age 51, had glaucoma for five years and was deteriorating rapidly. She cleansed her kidneys, killed parasites and changed her diet to the anti-arthritic one since she also suffered from arthritis in her hands for ten years with painful enlarged knuckles. She didn't get relief from taking aspirin. She got rid of her asbestos toxins by bringing her own hair blower with her to the hairdresser. After seven months she had pain relief for her arthritis (without aspirin) and her glaucoma was pronounced stable by her ophthalmologist.

Tooth Decay

The strongest part of our body structure is our bones. The strongest bones are our teeth. How can they decay? We must look at the **enamel**, **dentine** and **root** of the tooth as well as the bone they rest in for some answers.

Scientists have already searched very hard and long for answers. But their work is hampered by commercial interests that try to shape the results. Since commerce determines which research can be done (that is, paid for) sacred territory can be ignored. For example, the effects of sugar-eating, gum-chewing, tooth brushing, fluoridation, tooth filling materials and diet can be ignored if it interferes with product sales. Trivial studies such as comparing shapes of toothbrushes, studying the chemical composition of plaque, and studies of bacterial structure and genes are done instead. Studies “at the molecular level” do not threaten existing industries.

Important research has lapsed since the 40's and 50's. Perspective on tooth health was sound and clear in the mind of Dr. Weston Price in the 1930's. His scientific studies stand as a beacon even today because truths, once found, do not change. He traveled the world over in search of good teeth. Anywhere and anytime he found them, he described the people who had them. This is excellent science. It lets you draw the conclusions. He described what he saw in a book, titled Nutrition and Physical Degeneration.[13] They came to these conclusions from the following observable facts:

1. Skulls of primitive peoples who lived along coastlines, such as Peruvians, Scandinavians and various islanders, and whose staple foods included fish daily, showed perfect teeth; not a single cavity in a lifetime. They had strong bones that didn't break even once in a lifetime of 45 years. Skeletal structure was fully developed, meaning the jaw bone was not undershot or cheek bones squeezed together, forcing the teeth to grow into a smaller than ideal space. Consequently, there was room for the wisdom teeth, and no need to crowd the remainder. They saw no crooked teeth or unerupted wisdom teeth. The authors estimated a daily consumption of **4 to 5 grams of calcium** in their fish containing diet.

Our daily consumption of less than 1 gram calcium daily is small by comparison. Our wisdom teeth erupt poorly, our other teeth are often crooked. But today bad teeth go shamefully unheeded because we don't need to chew our food, we can lap it (ice cream) or suck it, or gum it (applesauce).

2. These primitive peoples got all the calcium, magnesium, phosphate, boron and other bone builders they needed simply from eating (fish) bones. Mexican peoples got 4 to 6 grams of calcium a day from **stone-grinding** of corn for their staple, tortillas, instead of from fish.

---

[13] It is still available from the Price-Pottenger Nutrition Foundation, a non-profit organization that seeks to keep his observations alive. Their address is PO Box 2614, La Mesa, California 91943, (800) 366-3748.

Where do we get our calcium? Milk is our only supply. One quart supplies one gram. There is little excuse for a carnivorous society like ours to regularly throw away the bones of its food animals in view of our dire shortage. It leaves us dependent on milk alone. Milk has so many disadvantages. It is impossible to milk a cow by machine and not get a few manure bacteria, *Salmonellas* and *Shigellas*, into the milk. These bacteria are not completely killed by pasteurization the way more susceptible bacteria are. It takes boiling temperature to kill all of them. Why isn't milk sterilized? Water was sterilized for human consumption in distant decades. Chlorination of water is not ideal but it did sterilize the water. Milk could be sterilized by boiling or flash-heating.

Milk has other disadvantages: dozens of antibiotics, both by feed and by shot, bovine growth hormone, chemicals added in milk processing, the bad effects of homogenization, and allergy to milk. Yet, in a choice between milk drinking and bone loss, one must choose the milk. This would not be necessary if bones were properly salvaged–ground to powder and added back to the meat where it belongs–to offset the acidifying effect of the phosphate in meat. One gram of calcium is not much bone (½tsp.) but it requires a whole quart of milk. Bone powder added back to ground meat, soups, stews could greatly improve our tooth decay problem, bone density problem, and skeletal growth problems.

Softened teeth set the stage for decay; bacteria do the dirty work.

Zapping bacteria does not kill them all. The zapper current does not reach into abscesses under metal filled teeth or around root canals. *Staphylococcus aureus*, which we are constantly stuffing in our mouths as we lick our fingers, finds an immediate hiding place in a crevice where it can't be zapped. Many other bacteria hide here, too: those that cause ear ache, sore throats, bronchitis, stiff knees, joint disease. You can try zapping all the *Clostridia*, *Streps* and tooth decay or plaque bacteria. But the only way to successfully eliminate them is to pry them out of hiding and wash them away. This is a job for the dentist (see Dental Cleanup page 409).

Strep. mutans is considered to be the bacterium that causes tooth cavities. I have found it in milk, evidently another pasteurization escapee. All the more reason to sterilize dairy products.

Frannie LaSalle, 52, was getting compression fractures in her spine, but the weak bone condition was evident in her mouth (many teeth were loose—they could be jiggled!). Her gums were red and inflamed. A low thyroid condition (she needed 2½ grains a day of thyroid—in one day the normal body goes through 5 grains of thyroid products) contributed to this. Her blood phosphate level was high (4.7 mg/DL—should be below 4.0) and her alkaline phosphatase was 205, also high, showing she was dissolving her bones (including tooth sockets) at a rapid pace. Her whole system was too acid, as could be seen in elevated CO_2 levels (28, when 23-30 is normal).

Only the major minerals, sodium, potassium, calcium and magnesium can have an impact on this major disturbance. The dentist said she had to have all her teeth pulled and replaced with dentures. Her kidneys showed all three types of calcium phosphate crystals. She drank no milk. She had only three weeks before her oral surgery appointment. She was started on ½ cup 2% milk, 6 times a day plus 50,000 units of vitamin D (a prescription dose) to make sure she absorbed all the calcium. She also took magnesium oxide (300 mg. once a day) and vitamin B_6 500 mg (one a day).

She was started on the kidney cleanse to help activate the vitamin D and to help the adrenal glands make *estrogen*. Her estrogen level (5.2 pg/ml) was too low to get the calcium deposited back into her bones. She was also given licorice herb for their estrogen-like action to help with this and vitamin C, 1 gram (1,000 mg) 2 to 3 a day.

Her mouth care was to be as follows: potassium iodide (white iodine, made up by dissolving 88 gm potassium iodide in one liter/quart water). Purchase a new very soft toothbrush. Use no toothpaste or store bought floss. Use 2 lb. or 4 lb. (the 4 lb. is coarser) fish line (rinse first). Brush twice a day; floss only once at bedtime before brushing. Use 6 drops of food grade hydrogen peroxide for daytime brushing. Use 6 drops of potassium iodide for nighttime brushing. Use no mouthwash, chewing gum, candy. In three weeks her teeth could not be jiggled. Her dentist was astonished (but was not interested in how she achieved this). In six weeks her mouth looked normal and she could chew some foods. Her vitamin D was tapered as follows: Take 6 a week for the first week (miss one day). Take 5 a week for the second week (miss two days). Take 4 a week for the third week. Then 2 a week indefinitely. She never lost a tooth.

Muscle Diseases

There are a variety of muscle wasting diseases, thought to be genetic in their cause. Yet, what could be more easily inherited than a parasite? Persons living together share food, living habits. refrigerators, and parasites. Their shared genes indeed give them similar susceptibilities but if we take muscle parasites away, muscle diseases "magically" disappear.

Of course, there is no magic involved. It is actually hard work. Hard work to rid the whole family of parasites that are shared and possibly were present even at birth. Parasites that normally don't go to the muscles. For example liver flukes and intestinal flukes. They belong in the liver and intestine! Yet, in muscle disease they show up and reproduce themselves in the muscles. The reason for this becomes clearer when you see that certain solvents have accumulated there. Heavy metals, bacteria, and viruses have accumulated there, too. The host's muscles, instead of kidney and bowel, have taken on the duties of toxic dumping grounds. Could it be that the regular routes of elimination were overwhelmed? Or did the muscles, traumatized by these unusual parasites, invite the toxins? A tantalizing question.

But in seven minutes you can methodically kill everything and anything that is alive in your muscles and shouldn't be there. Zap until you are free of all parasite invaders. Your muscles will feel lighter afterward.

Muscular Dystrophy

In muscular dystrophy the solvents, **xylene** and **toluene** are seen to accumulate in muscles. These also accumulate in brain and nervous tissue! (See Alzheimer's page 269 and multiple sclerosis, page 204). Could it be that these solvents are actually present in the nerves of the muscles?

Fortunately these solvents will leave your body, by themselves, in five days after you stop consuming them! Stop drinking all store bought beverages, including water and powders that you mix, and including health food varieties. Water claims and health food powder claims sound as convincing and strong as a twelve inch plank to walk on. But if the plank leads out over the side of a ship, would you walk it?

Throw all your possible sources of solvents out. Flavored foods are the chief offenders (cold cereals, sweets and candy too). But of course, you should check in your basement or attached garage for cleaning solvents. Places where painting is done or automobiles are worked on should be off limits to you.

If you've been wondering whether you have muscular dystrophy, which I consider to be a fluke disease, search your muscles electronically. Use prepared slides of flukes along with a sample of hamburger meat to represent your muscles. If flukes have already taken up residence in them, you should diagnose yourself as "Positive". If not, but other parasites and toxins are present, you have pre-muscular dystrophy.

Also, the likelihood of finding thallium is quite high, judging by the case histories.

Mel Rickling, age 18, had been seeing a specialist for bouts of muscular weakness for several years, but no diagnosis was given. His condition was not yet severe enough although it was difficult for him to raise an empty glass or get upstairs. It began at puberty, not too long after his first mercury tooth filling. He had asthma in childhood. The flukes attacking his muscles were liver fluke, intestinal fluke, and pancreatic fluke. Other parasites in his muscles were *Leishmanias*, several dog tapeworm cysts, and pinworm.

He also had assorted bacteria in his muscles. The solvents propyl alcohol, benzene, toluene, and xylene were accumulated there too. Ortho-phospho-tyrosine (cancer test) was already positive. His doctors had not searched for cancer in their biopsies. His drinking water contained lead and since he had lived in one house since birth he was probably drinking lead every day of his life. He also had high levels of mercury and some thallium accumulated in his muscles; these came from the tooth fillings in his mouth and could explain why his problems began after his first filling was put in.

His flukes and other large parasites were killed immediately with a frequency generator. He was started on the herbal parasite program to prevent reinfection. His diet and body products were changed to exclude solvent pollution. He could have no commercially prepared beverages except milk which needed to be boiled to kill bacteria. In twelve days his daily stomach pains were gone, so he was able to eat more and gain some much needed weight. The rash on his face was gone, the pain at his right side was gone, his muscle twitches were gone, his joints no longer ached and his mood was much better. The whole family was put on the parasite program and Mel was scheduled for dental cleanup. The plumbing repairs removed lead from the water and he was soon able to walk upstairs, in fact run upstairs.

A young man, seeing himself regain normalcy, wants nothing more than to lead a normal life" which includes reckless behavior. But after several warnings from his muscles he stuck to his restrictions and gained the weight he wanted in order to participate in athletics.

Myasthenia Gravis

is probably a fluke disease. Some chemical, possibly coming from the fluke, may affect the acetylcholine receptors, thereby causing an allergic reaction so they become inefficient. This is most noticeable in the eyelids. They droop from lack of strength to lift them.

The thymus is often involved, too. The thymus is extremely sensitive to benzene and with so much benzene pollution in our products and foods (pollution from gasoline is negligible by comparison), you will probably find benzene accumulated there. Search the thymus and the muscles for parasites, bacteria and tooth metal as well as toxins in the foods eaten daily. Kill invaders twice a week with a zapper or stay on an herbal parasite program until all danger of recurrence is past (one to two years).

Clean up dentalware, diet and environment. Keep no indoor pets since any new parasite, however tiny, will surely find the niche left behind by the flukes and give you a new myasthenia gravis-like disease. The whole family must be parasite-free to protect the member with myasthenia gravis. But it is a task easily accomplished and desirable in its own right, so discuss your plan immediately with family members. Don't delay. The flukes don't waste a single minute. They go right on feeding and breeding.

Carmen Opsal, age 37, was told by her specialist she had her myasthenia from birth since she didn't have the strength to nurse. She had pancreatic fluke stages throughout her body. Her plan was to start on the parasite killing program, clean her kidneys, remove toxic elements, kill bacteria and clean her liver. Long before she accomplished this, in one month, she was feeling better and had return of her strength on some days. She still had the solvent, methylene chloride, from drinking “pure” orange juice and praseodymium from eating foil-packaged foods, also thulium from her brand of vitamin C. She was full of auto exhaust and nickel from dental metal. There was hafnium from nail polish and hair spray and zirconium from deodorant. She planned to get rid of it all, and never need to return.

Universal Allergies

If minor allergies are due to a disabled liver, then extreme allergies must be due to an extremely disabled liver. This is the case for persons suffering from "universal" allergies, namely "everything", like the lacquer on floors, plastic chairs, the neighbor's flowers, and the grocery store.

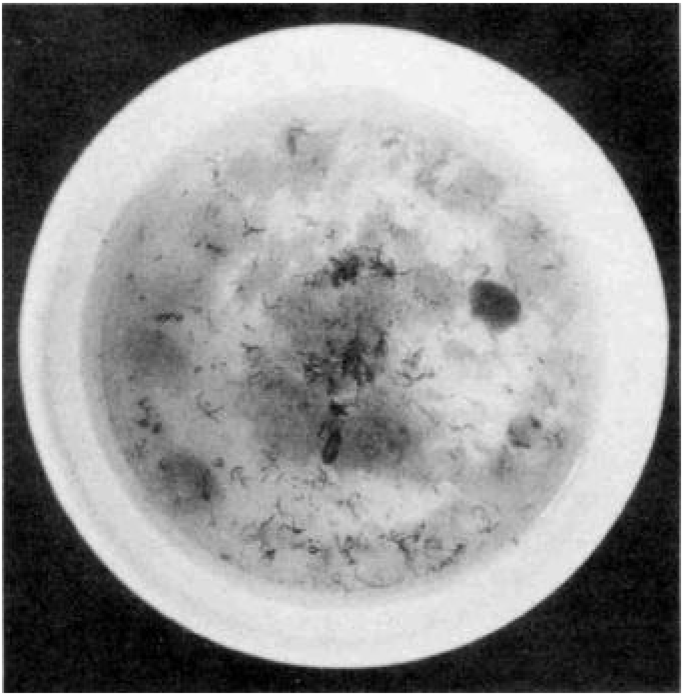

Fig. 32 Sheep liver flukes. Black threads in toilet are indicative of fluke remains.

They have more than merely clogged bile ducts. They have the sheep liver fluke living in the bile ducts! A tip-off to this situation is allergy to wool and wool fat (lanolin). A few flukes might not be noticed but a liver full of flukes that spill over into the intestine can give the worst case of allergy imaginable.

Sometimes the body manages to kill them with its own resources (maybe you ate something even too toxic for them!) They come through the bowel in a torrent. In the water of the toilet bowel they explode, spewing their infectious eggs all over in little black threads. Because these look like hairs, you may believe you passed "things with black hairy legs." These are actually burst flukes with black strings of highly infectious eggs. Why some people are literally taken over by these flukes is unclear. Amongst sheep, only certain sheep will be severely affected, being called "liver-rot." The disease in animals has been extensively studied.

Kill flukes with a frequency generator (434-421 KHz) or zapper. Come to the aid of the liver by avoiding food molds, removing dental metals, stopping chronic *Salmonella* infection and finally cleaning the kidneys and liver.

Environmental Illness

is another name for "universal" allergies. When more than a few flukes are present in the liver, they keep the liver from doing its major job: detoxifying all the food and chemicals that are taken into your body.

Different parts of the liver have different detoxifying jobs. One part detoxifies plastics and solvents, another part detoxifies perfumes and another newsprint ink, and so on. Foods have natural chemicals that need detoxification. By changing our food constantly we avoid overburdening any one of our detoxifying mechanisms. This is probably the basis for wanting different food at each meal and different meals each day. We somehow "know" when we're ready for the same food again.

Less extreme forms of allergy can be due to other flukes in the liver, such as human liver fluke (*Clonorchis*), or just plain clogging with numerous cholesterol crystals.

Anything that is lodged in the bile ducts obstructs the flow of bile. This causes back pressure in that part of the liver so it produces less bile.

The bile duct system is a gigantic tree with lots of interconnecting branches. Remember how "stringy" liver can be when you buy it in the grocery store. These strings are bile ducts. When one is obstructed, others take over its job. But when a whole section of the liver gets obstructed and it can't detoxify a whole set of chemicals you dare not get those chemicals into you again.

What if you do? These chemicals go coursing all over your body! They are taken up by various organs. The brain has special protection, called the *blood brain barrier*. But this can get broken by parasites that burrow. Now chemicals can roam through the brain. Some attach themselves and cause an "allergic reaction". Beryllium, from "coal oil," kerosene and gasoline attaches itself to the brain easily. Then other things attach themselves to the beryllium!

The extreme form of allergies simply requires killing the sheep liver fluke and other flukes inhabiting the liver. They tend to overflow the liver and inhabit the intestine, too. In this case, you might actually see some in the bowel movement after killing them. They won't let go of you as long as they are alive.

They have two attachments to hold onto you, yet, they are not difficult to kill, even with herbs. Use the herbal parasite recipe (page 338), zapper, or a frequency generator.

Sheep liver flukes might actually be breeding, that is, multiplying in the liver of the hyperallergic person. This is not normal. Sheep liver flukes "should" only spend their adulthood in our bodies. When the baby stages are also found in our bodies, (instead of in minnows or snails) there is undoubtedly a specific solvent involved. Environmentally ill persons have quite a few solvents accumulated in their organ tissues. Which one enables the sheep liver fluke to go through all of its development in the human is not known yet.

Obviously, the extremely allergic person, should remove all solvents from their diet and environment. Begin with eliminating propyl alcohol and benzene. The same products have other solvents too. It is processing of foods that puts solvents into them. Go completely natural. Dairy products are free of solvents, except for some cheeses. Remember to boil them to get rid of bacteria. Salt, olive oil, butter, and honey are free of solvents. With electronic technology, you can find solvent free products. Otherwise, if it didn't grow or you didn't make it from scratch, you must assume it has solvents! Cook from scratch, make your own pasta, bread, fruit juices, beverages.

Often, but not always, persons with sheep liver fluke, have a specific allergy to lanolin, a sheep product. Since lanolin is widely used in other products, this becomes a very broad range allergy. Such persons "can get no fat" at mealtimes or wear no wool without a considerable reaction. The allergy to lanolin does not disappear the day the flukes are all dead. But cleaning the liver with several liver cleanses (page 552) after killing parasites will start the recovery process.

Which comes first, the flukes or the solvents? That can't be answered. But what happens next is easily seen. The more flukes, the less able the liver is to detoxify solvents. The more solvents the better able the fluke is to multiply. A vicious cycle is set up that accelerates the illness.

Perhaps neither of these came first. Perhaps something else poisoned the liver so both solvents and flukes are given a home in your liver! Such a powerful liver poison would be a food mold: *aflatoxin*, *cytochalasin B*, *sterigmatocystin*, *zearalenone*, *ochratoxin, sorghum mold, griseofulvin, citrinin, T-2 toxin, Kojic acid, ergot* and others. Avoid food molds—see Mold Free Diet, page 365. The diet must be quite limited at first, to allow the liver time to "regain" its detoxifying capability.

Of course, it is assumed that environmentally ill persons have had their dental metal replaced by metal free composite. This includes gold. Gold accumulates in the pancreas, another organ of digestion. This may mean choosing partial dentures. Read the section on healing the jaw and Bone Strengthening (page 87) to ensure this move brings you success.

The liver is a versatile organ. It can regenerate itself but it won't if food molds block regeneration. Given half a chance it will become like new. After killing parasites do the liver cleanse (page 552). If it has been a month or more since you killed parasites, then go on a high dose parasite herb treatment the week before, or zap. Don't use the herbs the day of the cleanse. With one major allergy gone after each cleanse and by timing liver cleanses two weeks apart, it takes only six months to have a reasonably normal life again. You can endure indoor air again, sit on plastic chairs, read newspapers, wear cotton clothing and leather shoes without reacting. You must still be patient and careful as you take back the world for you to live in.

Delores Flores, 53, was brought by her husband to the driveway in front of the office. There she put on her mask and advanced to the outdoor bench. She did not dare to come in. Without doing any testing her condition was obvious. She must start killing liver parasites. But it seemed too simple to be believable. And she knew she'd be allergic to the parasite killing herbs (this was before the zapper). She decided to do nothing.

Patricia Humphry arrived wearing an industrial painters' mask. It smelled rubbery. Her winter boots smelled moldy. And a faint scent of mothballs came from somewhere. Even her car had an engine problem, spewing exhaust fumes into the driveway. I suggested she begin with some basic reading material on allergies. She did not return either.

Arlene Kelly, 50ish, could eat no fat—not the tiniest snippet. But at Thanksgiving she allowed herself a little gravy. The consequences were swollen eyelids, swollen face, swollen throat: quite a dangerous situation. After killing *Ascaris* and the flukes, and cleansing the liver (all in time for Christmas) she dared a little pie—and got along quite well!

Alcoholism

When the portion of liver that detoxifies ethyl alcohol (the drinking kind) is hampered you are at risk for alcoholism. The other contributors to alcoholism are **beryllium** and **ergot**. Perhaps there are even more contributors.

Beryllium is plentiful in coal products such as "coal oil", and in gasoline to which kerosene or coal oil has been added. Fuel for "hurricane lamps" is a common offender, filling the air with beryllium. Liver blockage can force beryllium to circulate through the body; if it happens to attach itself to the addiction center of the brain, you're in a heap of trouble. Beryllium is very reactive. Any other allergen coursing by can react with it. If this happens to be alcohol–the drinking kind–you will be alcoholic unbeknownst to you.

When the liver is quite disabled, there may still be alcohol coursing through your body the day after you drank even a little bit! It is bound to find the beryllium stuck to the addiction center. Together, they turn the brain into a uncontrolled machine. Neurotransmitters are released that shouldn't be; others not released that should be. Mood is affected in a typically "alcoholic" way. Depression may be lifted—or caused!

Alcoholic persons should remove all fossil fuels from their environment, and never choose a career that exposes them to paint, cleaners, or automotive products. They should do liver cleanses until 2,000 stones or more are out.

Of course, they should never touch a drop of alcohol: not even the Black Walnut Hull Tincture in the herbal parasite program. The aqueous recipe should be made for them. Since alcohol is produced anyway, in the body, the liver should never be poisoned by molds, especially ergot: the very mold that is so abundant in alcoholic beverages! Alcohol and ergot interact to make each more toxic.

To help the brain recover its neurotransmitter status, take glutamine (500 mg.), a B-50 complex, and niacinamide (500 mg—to help detoxify ergot), with each meal.

To prevent alcoholism, protect your liver from food molds, especially ergot. Add vitamin C to nuts, pasta cereals, grains and even alcoholic beverages! Avoid fossil fuel pollution of your home by switching to all electric utilities.

Alcohol Addiction

There are many definitions of addiction. My definition is based on the special brain toxins, *beryllium* and *ergot*.

The brain has a region called the addiction center. If this center is stimulated it produces pleasure-chemicals. It is carefully controlled so that not too much pleasure or happiness can be experienced.

When a toxic substance, beryllium, is inhaled it circulates with the blood to the brain and may land at the addiction center. The more beryllium is inhaled the bigger the chance that it will occupy the addiction center. The brain cells in the addiction center have receptor sites for *glutamate* (the same glutamic acid that comes from the protein in our food). Normally, glutamate activates the addiction center. But when beryllium has “stolen” these seats, the glutamate is powerless to activate the joy and happiness center. The result is a low level chronic depression. The more beryllium there is to clog the receptors the worse the depression. Giving glutamic acid does not help but giving **glu-**

tamine does! Addicted or depressed persons should take glutamine, no less than 3 grams (3000 mg) a day. It comes in 500 mg. tablets. It is completely harmless.

When we drink alcohol or put it on the skin (as in mouthwash, tinctures, medicine) or produce it by fermentation in the intestines (*Candida* produces alcohol) a substance, *salsol,* is formed. Salsol reacts with beryllium. If the beryllium is in the pleasure center it reacts with it there. This reaction has the effect of activating the cells! Now a large amount of pleasure-chemical can be released. The amount is larger than normal because so many clogged cells are activated together. This explains the alcohol "high". In all the alcohol-addicted persons I studied, salsol was present, along with beryllium, on the receptor sites normally activated by glutamate (or NMDA or kainate). As we removed the beryllium we saw that the salsol also disappeared.

The solution to alcoholism is to avoid ergot contaminated food and avoid beryllium inhalation. We also remove the brain beryllium using **thioctic acid**. Stopping the use of alcohol may save a life or career but does not correct the problem. Even after 30 years of abstinence, I still see the beryllium present in the addiction center and the salsol, derived no doubt from endogenous sources, still attached to the beryllium. This is why the addiction is never gone even after years of abstinence.

If any member of the family is, or was, addicted to alcohol the house should be searched for beryllium sources. Hurricane lamps or antique lamps are the most common sources. Remove them permanently. Washing does not clean them. Remove all solvents, cleaners, lighters. Switch to a butane lighter. The air should be tested for beryllium. The garage door to the house should be permanently closed, and the car and lawnmower kept out of it. Addicted persons should not be painters! Nor walk into a dry-cleaning business. Soon you will see a more cheerful disposition in the addicted person and this will be rewarding for the whole family.

Miguel Alcorn's wife cleaned up the environment for Miguel, whose alcoholism was 30 years long. Even lighter fluid was removed. The garage was sealed off from the house. She added vitamin C to his meals so he didn't have to “take anything”. She was very careful about moldy grains. He lost his preoccupation with drinking after killing parasites. She believes he is not sneaking any. Our hats are off to her.

Seizures

are always caused by tiny *Ascaris* larvae in the brain.

I suppose they find their way to the seizure center by accident. It is not normal for them to be in the brain, they typically travel between the stomach and lungs. *Ascaris* eggs are present everywhere in animal filth. Dogs, cats, horses and pigs all get *Ascaris*. Their excrement dries and flies about in the dust, but mostly it resides in the soil. Children playing in the dirt cannot help but pick up *Ascaris* eggs. The eggs hatch in the stomach and the tiny larvae, microscopic in size, travel first to the lungs. Here they go through a *molt*. This causes some coughing.

Whenever a child coughs for part of the day do not assume it is harmless. Use a frequency generator set to 408 KHz, or use a zapper. Children should be treated for *Ascaris* anyway, whether coughing or not, once a week. The tiny larvae are in the cough up. Children should not be taught to politely swallow this. They could be swallowing their own future seizures, asthma, or eczema. Teach children to use tissues for all spit up. The lungs are doing what they can to rid the body of these invaders. Your intelligence must cooperate with your lungs.

Ascaris do not come sweet and clean in themselves. They bring their own bacteria and viruses. One of these bacteria is *Bacteroides fragilis*. *Bacteroides* needs a host like *Ascaris* so it can be oxygen-free. Such a requirement is termed *obligatory anaerobic* meaning "must have absence of air."

Being transported to the brain inside *Ascaris* larvae is probably the means by which *Bacteroides* gets into the brain. Brain abscesses and brain tumors usually have *Bacteroides fragilis* growing there. Brain tumors will not shrink unless all the parasites, bacteria and viruses are dead. *Bacteroides,* in turn, are big enough to house thousands of viruses. Two common viruses seen with *Ascaris* are *Coxsackie* B_1 and *Coxsackie* B_4. Perhaps it is the toxins of the *Ascaris* larvae or *Bacteroides* or *Coxsackies* that induces the seizures. Maybe it is something else about the infestation that induces them. But by killing *Ascaris, Bacteroides* and *Coxsackies* (zapper or frequency generator at 408, 325, 364, 362.5 KHz) you have eliminated the first essential link in the chain of developments that causes seizures.

The brain does not quietly tolerate these invaders. It fights back with its immune system. But the wormlets are already too big to be eaten by white blood cells. The brain fights back by producing **inflammation**.

Inflammations are intended to attract calcium so a wall can be built around the intruders. Inflammations are negatively charged regions so the positively charged calcium can find its way to the inflamed site. But lead and mercury are also positively charged! Perhaps this is how these **toxic metals** are attracted to the brain. All metals are positively charged. Perhaps this is their fateful poisonous attraction to living things. Perhaps they do not poison when no inflammations are present! In seizure cases we see many tooth metals in the brain. These must be removed so the inflamed site can heal.

Other toxic substances have also found their way to the inflammation at the seizure center: **vanadium** from leaking household gas; **PVC** from new carpets; **titanium** from face powder; **zirconium** from deodorant; **asbestos** from the clothes dryer belt or hair blower; **lead** from tainted drinking water. Do a good clean up of your environment.

Solvents accumulate here, too. Especially **toluene** and **xylene**. These are found in paint (persons with seizures should never be around fresh paint) but are also found in trace amounts in carbonated beverages. A person with seizures should drink no commercial beverages: see the Recipe section for homemade carbonated and other beverages. There are several other specific brain irritants that accumulate at the seizure center.

The food fungus, **ergot**, is always seen in seizure cases. Perhaps it can act alone to produce seizures. After all, seizures are an ancient malady, existing long before chemicals and solvents were manufactured.

MSG, which is *monosodium glutamate* should never be used in food. It was already banned in infant food in Europe a decade ago. Why are we still allowing it? Are our children not as precious? Healthy adults may tolerate it better than young children. But what about sick adults? Specifically, adults with a brain illness? Glutamic acid is a natural constituent of protein in our foods. And the sodium glutamate it must form in the body does no harm. But MSG is not biologically produced. It is lab-made. Lab-made amino acids are not the same as biologically made. Perhaps it is these “isomers”, perhaps it is the simple overdose of a natural thing that is brain-toxic. Whatever the mechanism, MSG should not be consumed by anyone, let alone a seizure sufferer. Fortunately, you can ask at restaurants, where (not whether) MSG is used so you can avoid these foods.

BHT and **BHA**, standing for *bishydroxytoluene* and *bishydroxyanisole* are food preservatives and also seizure triggers. They are often put on the boxes of cereals, rather than the cereals themselves, so the cereals can be pronounced preservative-free. Imagine how much the box must be drenched with to prevent oxygen leakage into the interior?

A more insidious seizure trigger is a perfectly natural substance, **malvin**. Malvin is the natural dye found in grapes, strawberries, plums and blueberries. Stop eating strawberry and grape jam or juice. Chickens and the eggs they lay, have lots of malvin too, stop eating chicken and eggs.

Here are foods relatively free of malvin: artichokes, asparagus, almonds, barley, beans of all kinds, green beans, broccoli, Brussels sprouts, cantaloupe, celery, nectarines, citrus, dates, mango, pears, kiwi, pineapple, Granny Smith apples.[14]

To Summarize

To stop your seizures on a dime, and not have another one:

1. Avoid malvin in food. Eat nothing colored red or blue, chicken, eggs, MSG, BHA & BHT. Boil all dairy products or don't eat them.
2. Avoid ergot in food. Eat no whole grain products; take niacinamide 500 mg three times a day to help the liver detoxify tiny bits in other foods.
3. Kill *Ascaris*, *Bacteroides* and *Coxsackie* virus and stay on a maintenance program of killing them. Avoid reinfection. It would be wise to have only outdoor pets.
4. Replace dental metal with metal-free plastic (See Dental Clean-up).
5. Clean up the home environment and body products of toxic substances.
6. Keep your fingers sanitary: spray them with 10% grain alcohol or vodka after bathroom use.

You can often tell by how you feel whether you are near to having a seizure. But some people get no warning. Don't take chances. Keep your drugs with you, even though you no longer need to take them. Drugs like Tegretol™ and Phenobarbital™ are not harmless. Once you can tell that you're better, try reducing your drug dosage slightly. See how you manage. If you have a breakthrough seizure you could fall and break a bone. Don't take chances. Even a year after your last seizure you should carry your medicine with you and have some in your house. This is because it is so easy to pick up *Ascaris* again. The highway to the brain–its seizure center–is still open. They will travel it again. It might only take two days from the time of accidental swallowing of animal filth, to having little larvae in the brain. Use a frequency generator every day at first, or zap daily to guard against this. Notice that killing parasites the first time may actually start a seizure coming. Simply stop zapping if you feel one coming. Do shortened zapping until you can tolerate a regular treatment. Have someone with you while zapping.

---

[14] Taken from *A Guide to the Identification and Treatment of Biocatalyst and Biochemical Intolerances*, 1988 by J. A. Krohn, Los Alamos Medical Center, 3917 West Road Suite 136, Los Alamos, NM 87544.

Say good-bye to seizures forever. But don't forget your susceptibility. I believe you should be extra careful for two years. If you have a recurrence, repeat the whole procedure. This time it should be curable in a single day.

Doug's mother was distraught when she brought him, age 8 months. Her doctor only seemed sympathetic with her purpose to keep the baby's temperature down the next time he catches a cold. But her doctor had referred her to the county social worker. She had been completely honest with her doctor, because she was that kind of trusting person. Since her baby had only experienced one seizure (which was during a fever), she didn't see why her beautiful first born child should be on medicine “the rest of his life”. She wanted her baby to be perfect. But the social worker had called her, talked about “the law” and being an unfit mother. She was all apart. The baby was supposed to be on Phenobarbital™ twice a day. Our tests showed Doug had *Ascaris* plus lead toxicity. He was also getting home made strawberry and grape juice. She promised to put the three cats outdoors, keep the baby off the floor, keep Doug's fingernails short and always wash his hands before eating. The parasites were easily killed. The lead was sporadically present in the water. She planned to move, and until then would filter all the drinking water so her breast milk would be free of it too. His diet was changed to eliminate eggs, chicken and the red and blue fruits. We recommended leaving the state in order to be able to peacefully raise her child. We promised to treat her child free of charge if another seizure should occur. Two years passed and we heard she was doing fine with her child.

Clara Scruggs, 50ish was losing control over her seizures and had to be hospitalized while a new medicine was tried. It changed her personality (again!) which upset her husband, too. She was started on the herbal parasite program but could only increase by one drop of Black Walnut Hull Tincture a week, instead of daily, since each new increase would give her a seizure. In half a year she was fairly free of seizures in spite of being off drugs. After each seizure, a checkup showed she had picked up *Ascaris* again sometimes with additional parasites. She could not bear to put her cat outside; Boots had been a friend in need many times. When she finally got Boots onto a regular parasite program she improved enough to go to church and church events again. When Boots finally “wanted out” she didn't mind. She decided to do a liver cleanse—this, too, gave her two seizures the next day but paid big dividends in other ways. She eventually improved to an incidence of one small seizure (“spacey” time or incoherent speech) in two weeks.

Chun Yee, age 28, has been on Dilantin™ from age 15. Now he was up to five pills a day and it didn't control his seizures. Any activity would trigger it. He was put on a low malvin diet and started on the kidney cleanse. His blood test showed he was anemic with a high MCV (104 cu microns), suggesting the *Ascaris* worms were using up all his vitamin B_{12} too. In six weeks he was down to one or two seizures per week, although he had not yet started the parasite program. He had beryllium built up in him, probably from a kerosene heater. When the pets and family were all treated for parasites he had no more breakthrough seizures and could cut his medicine in half which gave him much more energy. He was no longer in danger of losing his job.

Shiresse Nobel, age 7, was having minimal seizures but the mother did not want to start her on medicine. Shiresse had high levels of mercury in her body, although she had no tooth fillings. The whole family cleared their Ascaris infestation and in three weeks the seizures were gone but aggressive behavior remained. There was mercury in the air of some rooms but it was not in the paint. The bathroom was the worst. When everything in the bathroom was removed, the air cleared and so did Shiresse's behavior problem.

Drew Seaton, age 8, had his first seizure lasting nearly a minute. The parents were very fastidious and extremely conscientious about diet and habits. They were distraught. They all had *Ascaris*. A baby was still being diapered. Drew had arsenic (pesticide under kitchen sink), formaldehyde (some remodeling) and PVC (a new carpet) accumulated in his body. The carpets had to be steam cleaned to get rid of the arsenic. He was started on parasite herbs at once, since he was on medication that would shield him from having another seizure while killing *Ascaris*. Two weeks later everybody, except the mother, was free of *Ascaris*. In another two weeks they were all free and had cleaned up house toxins. They dared to stop his medicine. This let Drew's former happy personality come back to everyone's delight. Two years later there was still no recurrence.

Cosmo Maser, 30ish, was in a hospital across the country. He was having continuous seizures, although he had been there a week. It seemed impossible to transport him but, against doctor's orders, they got him into a station wagon. He had nothing to eat or drink that had any malvin in it (he ate four very well done hamburgers, plain, with lettuce on his trip) and his seizures stopped immediately. They felt a bit sheepish upon arrival 20 hours later since he could sit up, could tell his own story and no longer looked ill. He was without medication, too. They had left in a hurry. They immediately removed all the metal from his mouth; this cleared his mercury problem. He was started on parasite medicine and weathered the small seizures each increase gave him. He could return home in five days with his new diet and thioctic acid daily as a supplement. He occasionally had a seizure (2 a month) until they moved away from the busy street below their apartment. About a year later he could hold a job and go off Social Security support.

Weight Problems

Overweight is not the same problem as *obesity*. In fact, I do not have the answer to either problem, only a part of the answer.

In **obese** women, the ovaries, pancreas and thyroid are all involved. Perhaps the adrenals, the brain satiety center, and the liver are also involved. Maybe it's as simple as gold accumulation in all these places. Perhaps it is bacteria in all these places. When you weigh close to 300 pounds obviously some organ isn't working right. Try several things, but not a starvation diet.

The cause is not eating too much. Try removing all gold: gold teeth, gold jewelry and gold rings. Replace them with non metal varieties. After removing the gold, pull the remaining gold out of your tissues with thioctic acid (2 or 3 a day for several months). Make sure kidneys are able to excrete the gold instead of making crystals by doing a kidney cleanse. Gold accumulates in the pancreas, the brain (possibly in a control center here) and the ovaries (causes some infertility here). Also try clearing the body of all bacteria and parasites by regularly using a zapper. Use the Bowel Program (page 546) to evict the last of the *Shigellas*. Be very careful to avoid nonsterile dairy products. Try cleansing your liver by doing liver cleanses. Get 3,000 stones out.

Make sure you are getting enough nutritious food; make carrot and vegetable juice; use no commercial beverages. Avoid moldy food—don't take risks. If all these measures bring your weight down to the level of mere overweight give yourself good grades.

Overweight is a low energy condition. Your food is being turned into fat instead of energy. The decision not to make energy is being made in the liver mainly, but perhaps other organs as well.

Try cleaning the liver (page 552) until no more stones come out: get at least 2,000 stones. Notice that as the liver gets cleaner you get more and more energy right after each cleanse. Some cleanses have a dramatic effect. Others do not. This suggests that a certain part of the liver is the responsible part. Soon after the cleanse–within a week–the same old lassitude sets in. But during these few days notice how your body feels. It feels light. Your abdomen feels tight, like it's a part of you again. You're mind isn't on food throughout the day. It's very easy to lose weight, in fact you may lose five pounds in these few days without dieting or exercising. Definitely, your long-lost weight regulation is back in force. But then it vanishes. Fortunately, a bit of the weight loss stays with you, and by repeating cleanses (only once in 2 weeks, though) you can shed the pounds you want and gain energy in a permanent way. It is probably the way nature intended.

Try increasing your bowel movements. Notice how cats and dogs seem to derive energy from emptying their bowels. A cat walks to its litter box; after emptying its bowels and carefully covering it up, it jumps from the box and runs away. It now has its playful mood. A body chemical, *acetylcholine*, plays a role in emptying the bowels. Acetylcholine is a necessary operator for many of our muscles. Is there a disturbance in our acetylcholine metabolism in overweight conditions?

Coax your body to release more acetylcholine, at least in the intestines, by using a herbal laxative like Cascara sagrada. Other varieties are useful, too, but don't use a drug laxative: it burdens the liver more. Try to have three bowel movements a day.

Shigella bacteria can cause dreadful constipation. Immediately, the body feels sluggish, abdomen feels disconnected or hanging out. A nameless hunger sets in. Keep a close watch on dairy foods to make sure they are sterile. Zap when in doubt and do the Bowel Program.

Exercise helps as long as you keep at it. Strict dieting works as long as you keep doing it. But there's the problem. When you stop, you lose ground. It doesn't solve the problem of an inefficient body metabolism.

Raising thyroid levels helps but this can be dangerous. Raising thyroid levels naturally, by removing toxins is a very effective method—provided it was low to begin with. Overweight people often have a low body temperature, showing that the thyroid is involved: it is under producing. **But giving the body extra thyroid doesn't solve the thyroid's problem. It only temporarily solves the rest of the body's problem.** The thyroid will have viruses and heavy metals in it. The metal in your mouth drains downward to the stomach passing very close to the thyroid. Perhaps its iodine uptake is inhibited. The chlorine in water and bromine in bread may inhibit iodine uptake by the thyroid, too. After all, they are all halogens. Stop eating bleached bread and filter out the chlorine in your water. If this raises your body temperature you could expect better weight control.

The traditional herb, *Fucus*, was used to treat thyroid problems (and overweight) in days when herbs ruled medicine. Herbalists made a point of discouraging use of plain iodine. Fucus, they said, was much more effective. See *Recipes*.

If all these measures don't work for you, at least you have improved your health trying.

There are some advantages of being overweight. Overweight people seem to weather illness better. They laugh more. But one look in the mirror or at the scales ruins it. Put away the long mirror and scales. Don't ruin your whole life over it. Make a reasonable effort and then let go. Enjoy your stay on this planet.

Underweight can be just as difficult as overweight to correct. Once the stomach has been trained to say “full” or “full enough,” even after a few mouthfuls, it is difficult to heal. *Salmonellas* in the stomach wall are often seen. *Giardia* and other common parasites, that don't belong in the stomach, are seen here in cases of underweight.

The stomach is invaded by bacteria and parasites when its immunity is low. The common culprit is tooth metal draining down continuously into the stomach. Remove every bit of mouth metal. Don't cook or eat with metal. Stimulate your appetite with B-vitamins. And, again, the liver plays a role. If it is toxic with mold, it may say "eat no more" and the body obeys.

A stage of cancer illness is weight loss. This is called *cachexia*. The victim believes he or she is eating enough. They may not even have a poor appetite. Yet weight drops steadily. This metabolic problem has been studied scientifically. A chemical, **hydrazine sulfate** (prescription only), can reverse it to some degree. Use 65 mg three times a day for 30 days. Perhaps a lesson taken from cachexia metabolism could be applied to obesity. For the cachexia sufferer, life would end sooner if high calorie supplements were not used. But don't use canned "calorie boosters". They are polluted—often with the very solvent that makes the condition worse. Instead, make an eggnog: ½cup boiled milk, ¼ cup boiled whipping cream, a raw egg (exterior carefully washed), 1 tsp. olive oil, a banana, honey, cinnamon, cloves, nutmeg to taste. Mix all in a blender. Drink 1 cup a day. Vary it daily to keep it interesting.

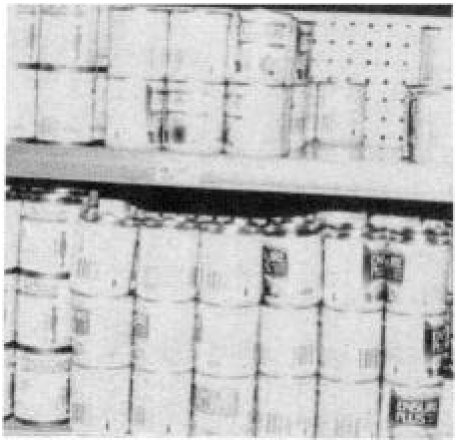

Fig. 33 I found all calorie boosters to be polluted with wood alcohol. Make your own.

If all this fails, give yourself credit for achieving health improvement and shift your focus to a different project. Underweight is not all bad either.

Sleep Problems

All big animals sleep, but some sleep by day instead of night. When humans do this, that is, work the night shift, they don't feel as well. Humans need about seven hours of sleep out of the twenty-four. Younger ones sleep more; newborns sleep much more. When we are deprived of sleep we are grouchy, think less clearly next day and have less energy. In spite of lots of research at “sleep labs” sleep problems are not understood, except for sleep *apnea*. Breathing should be even. When breaths are missed it is called apnea. It is especially disturbing when a baby shows apnea.

Sleep Apnea

Since breathing is regulated by acid levels in the blood and this is influenced by air quality, air toxins should be searched for first. **Cigarette smoke** is an air toxin. **Vanadium** from a gas leak is a very serious air toxin and can go unnoticed. Do your own checking since gas companies give wrong answers four out of five times. Ask a home construction company to check for gas leaks or the Health Department. **PVC** from new carpeting may be polluting the air. **Arsenic** from “treated” carpets and drapes and furniture also pollutes. **Asbestos** from clothes and hair dryers may be the toxin responsible. Perhaps even **fiberglass**, **formaldehyde**, or **freon**.

Adults with sleep apnea show swollen throat tissues: not necessarily pain. This makes the air passage smaller; long gasps of air are taken to try to make up for the missed oxygen and the carbon dioxide build up.

Swelling of the throat is a common allergic reaction. The possibility of allergy should get second consideration after air quality. Drug reactions, even in a nursing baby, where only the mother is using a medicine could be the problem. Allergy to food, chemicals has been suggested, as well as a simple lack of vitamin C (implicating mold and medicine which consume vitamin C in the detoxification process).

A third possibility is infection. Many bacteria and viruses can cause throat swelling. Redness of the throat is a telltale sign. It doesn't necessarily hurt. Kill all invaders with a zapper and try to understand the basis of low immunity in the throat.

Keeping metal in the mouth constantly, is a cause of low throat immunity since it must drain past the throat.

If you snore, you can deduce that your throat is swollen, even if you don't have sleep apnea. Pursue all three possible causes (air toxin, allergy, infection).

Overweight and obesity have been emphasized as causes. This may apply to some cases but certainly not to babies.

Whatever you do, don't do nothing. Keep removing bad things until you find the cause of irregular breathing.

Chester Fannon, 5Oish, was quite overweight and wore a mask at night with an air blower to assist his breathing. He had been referred to a sleep center for sleep apnea. He had extreme dryness of his throat at night and some hearing loss in one ear. He was toxic with arsenic (roach killer), bismuth (cologne), tin (toothpaste), and thallium (polluted dentalware). He was infested with both species of *Ascaris* and had a hacky cough. He had four solvents accumulated in his tissues. He was growing nine pathogens: *Mycoplasma*, *Haemophilus inf.*, *Streptococcus pneu*, *A-strep*, *Nocardia*, *Staphylococcus aureus*, *Bacillus cereus* and Flu virus, over half of them in his throat. These were killed with a frequency generator and a general cleanup was done. After two teeth were pulled he no longer needed his mask, he no longer had apnea.

Peter Day, middle age, was set for throat surgery in a few days. He had a couple of scary nights when he thought he was dying and couldn't risk many more nights with his obstructed airways. He was overweight. His throat was red and swollen, although he felt nothing. He never even had colds. His diet was completely changed, to things he rarely ate (bananas, milk, soup, oatmeal) and off things he ate daily (hamburgers, fries, tea, pancakes with genuine maple syrup). In two days, his throat was quite clear, the tissues having become unswollen. Maybe it was the molds in the maple syrup, maybe it was the oxalic acid in the tea, or something else he could not detoxify in these foods. He was certainly happy not to live the rest of his life with an artificial voice box.

Insomnia

Another sleep disturbance is waking in the night and not being able to go back to sleep for hours. Or not being able to get to sleep.

I believe these problems are caused by a high ammonia level in the brain. This belief is based on two observations. **Ornithine**, an ammonia reducer, induces a wonderful sleep in sleep-deprived persons. It is also observed that after killing parasites, which produce ammonia, sleep is much improved. Our metabolism does not produce ammonia. We produce urea which is excreted by the kidneys along with water and then called urine. When we are parasitized, our metabolism is burdened with ammonia, though, made by the parasites. We have to turn it into urea in the liver and kidneys so we can excrete for them. But this can't be done in the brain! The brain lacks an essential enzyme, *ornithine carbamyl-transferase*, for this bit of biochemistry. The brain was never meant to be parasitized or infected and has no defense. Most of our parasites come from animals we associate with. We weren't meant to live with horses, cows, sheep, pigs, monkeys, guinea pigs, cats, dogs and chickens nor to come in contact with dozens more at a zoo. We do so at our own peril.

It is known that ammonia is a strong brain irritant. In fact, a person can be awakened from a coma by being made to smell ammonia "smelling salts." Ornithine reacts with ammonia, mopping it up like a sponge. **Arginine**, another amino acid, also reacts with ammonia, but does not put you to sleep. So there is more to insomnia than mere inability to reduce ammonia levels. Arginine results in alertness and therefore should be used in the morning, when needed. Ornithine, given at bedtime, may take ½ hour to do its magic. Both are perfectly safe, since they are natural to your body, and a food constituent.

Start by taking two ornithine capsules (each 500 mg.) on the first night. Take four the next night. Take six the night after and choose the dose you like best. Sometimes it takes five days to "catch up" on everything that needs to be done for the brain and get you sleeping. Meanwhile, of course, you are planning to kill your parasites and be done with insomnia in the most effective way of all.

Another sleep aid is herbal. A couple of herbs, **valerian** and **skullcap**, are known for such action. The mechanisms are not understood and this makes for nonuniform action. Some persons sleep well with them, others do not. Simply try them to find out. We are all so different in our metabolism details, we respond differently to herbs. But it is a blessing that the mechanism is not understood. Herbs, a tradition that precedes civilization, need to be forever off limits for intervention by government agencies.

Tryptophane, another amino acid, is about twice as powerful as ornithine, but was taken off the market a few years ago. Some persons taking it daily were seen to become quite ill and some deaths ensued. Since tryptophane had been used in prior years without noticing toxicity, something unusual should have been suspected. My tests showed extreme pollution of tryptophane capsules. They contained PCBs, mercury, ruthenium, strontium, praseodymium, aluminum, and benzalkonium. I can only speculate that a mixing vat broke, dumping its precious load onto the floor—but it was salvaged. Or that the mixing vat wasn't cleaned thoroughly from it's last use.

Persons with illness due to taking tryptophane developed an extremely high eosinophil count in their blood test—an index of parasitism, too. Parasitism, that would have led to insomnia in the first place! Were these unfortunate victims seeing the cause or the result of their tryptophane use? This tragic event should have led to a discovery of the heavy pollution, a revelation of the industrial manufacturing process, and a safeguarding against any repetition. It has not been done (certainly not publicly).

Foreign countries' manufacturing processes do not come under U.S. scrutiny or jurisdiction, although some imported products must pass tests. There are no safeguards against repetition of the tryptophane experience. It behooves us to demand safe supplements and medicines. It is not the list of ingredients that informs. Lot analysis, after bottling, would give us the necessary safeguard. The presence of filth contamination and toxins cannot be completely avoided but the consumer can make informed choices if he or she knows it is there. Disclosure, of course, is the bane of the manufacturing business. Interest rate disclosure was the bane of the money lending business. Such important matters can't be left to "self-regulation" policies. The consumers must simply demand to know what they are consuming.

Ruby Adair, 14, ached all over, had ringing in her ears, sinus problems and chronic fatigue. She couldn't get to sleep, ever; and had been half a year out of school already. She had intestinal flukes in her stomach. In three weeks she had eliminated them with parasite herbs and she could go to sleep naturally.

Yeast Infections

The most common yeast in humans is *Candida albicans*. *Candida* has always been around. It flies in the air, searching for a place to land and reproduce. It can invade a variety of human tissues like the mouth (called thrush), skin (including some kinds of diaper rash), vagina, and the digestive tract. We all have some yeast in our digestive tract, but when it gets out of hand, it's called *candidiasis*.

Yeast is a **fungus**. It needs dampness to survive and sugar to grow.

Our immune system, white blood cells, are capable of eradicating yeast provided it isn't growing too fast. And provided the white blood cells aren't immobilized or preoccupied with something else.

Diaper Rash

A baby's rash is an example of the white blood cells being preoccupied. When chemicals are used in the diaper, the white blood cells go after the chemicals and let the yeast grow. Drying the baby's skin helps since the yeast must have dampness. This should be done with air, sunlight and a heat lamp, not with more chemicals! Certainly not with cortisone containing salves that further reduce the immune competence of white blood cells.

Use a heat lamp for five minutes at a time, several times a day. Switch to cloth diapers; do not bleach them with chlorine bleach, the residual chlorine trapped in the cloth is a chronic irritant, setting the stage for another rash and future chlorine-allergy. Cloth diapers should be sterilized, not bleached. Use the hottest water your laundry system is capable of producing. Add ½ cup borax for the washing process. If you have homemade Lugol’s iodine (made by your pharmacist or by yourself, see *Recipes*), add a tsp. to the wash or rinse. Vinegar is a yeast inhibitor, add it to the rinse. Dry diapers at the hottest setting. Dry to kill. Kill all the yeast spores in the diapers.

To strengthen the baby's skin against future infection, do not put chemicals on the skin. Do not use any soap, fragrance, bath oil, ointment or lotion. Do not use cotton balls or baby wipes. Do not give a daily bath. Wash bottoms gently, with borax followed by a vitamin C rinse. Vitamin C is acid and is our natural healing agent but it will sting on a broken skin surface. Use it as dilute as necessary to be tolerated. Zinc oxide is another natural healer because it competes away the iron that fungus and bacteria need for their reproduction. Never use commercially available zinc compounds though, simply purchase your own zinc oxide powder, mix it with cornstarch and keep in a large old salt shaker, dust it wherever there is moisture or fungus growth.

Treat Yeast or Fungus the Same

Other fungus growths, like *Tinea* (crotch itch), ringworm (not a worm at all), athlete's foot, along with *Candida*, can be similarly eradicated:

1. Deprive the fungus of moisture.
2. Deprive the invaders of iron.
3. Deprive the fungus of sugar.
4. Strengthen the skin's immune power.
5. Strengthen the skin's healing ability.

It may be impossible to deprive the fungus of moisture, for example if your feet sweat and you must wear socks. Take your socks off as soon as you are at home, treat your feet with a heat lamp. Use zinc oxide or cornstarch to powder and dry the skin. Boil your socks when laundering. Dry them to tinder-heat (too hot to touch). Launder with borax only (soaps and detergents contain aluminum which pollutes the skin). Rinse skin with vitamin C water. It takes all these measures used simultaneously to clear up athlete's foot fungus. And great persistence. They may have developed a foothold underneath the toe nail where a steady supply of moisture, iron and sugar is available to them. Nevertheless, your white blood cells will eventually gobble them up if you let them.

In thrush (yeast infection of the mouth) you must again outwit its growth by doing everything possible at one time. Eat no sugar, drink no fruit juice, stay off antibiotic. Avoid trauma like eating abrasive foods (crusts, popcorn, nuts, lozenges) or sucking on things. Floss teeth only once a day (using monofilament fish line), followed immediately by brushing with white iodine (or Lugol’s, but this may temporarily stain). Hydrogen peroxide is not strong enough. Remember to sterilize your toothbrush with grain alcohol or iodine. You may also rinse your mouth with Lugol’s (6 drops to ¼ cup of water). Or apply 6 drops directly to the tongue and rub it in lightly with your lips.

Do not use Lugol’s iodine if you have been told you are allergic to iodine.

Kill *Candida* daily with a frequency generator or zapper. Since reinfection is constant, you must continue to do all the treatments given to permanently cure yourself of fungus disease.

Since *Candida* grows right into your living cells (which you are not attacking!) you cannot kill it all at once. Only surface fungus can be accessed by either Lugol’s or electrical “zapping”. But as the top layer is killed, exposing the next layer, you will make progress. It will take a month of daily treatment to clear it.

Clearing up fungus at one location but not another will not bring you a permanent cure, either. Damp locations like under the breasts, under the belly fold, groin and crotch need to be kept dry with cornstarch daily. Keep it up long after it seems to be cured.

Fluke Disease

Flukes, or flatworms, have a complex life cycle with many stages. Although sheep, cattle, pigs and humans can be "natural" hosts to the **adult** stage, the other stages are meant to develop outdoors and in secondary hosts. When fluke stages other than the adult are able to develop in us, I call it *fluke disease*.

Or, when an adult that "normally belongs" to another species is able to develop in us, I also call that fluke disease. Or even with adult flukes in their "normal" host, when they move from the organ that they "normally" colonize to other organs in the body I call this fluke disease, too.

Four fluke varieties engaged in this extra territorial pursuit are the **intestinal fluke**, **sheep liver fluke**, **pancreatic fluke**, and **human liver fluke**.

As you can see from their names, scientists have studied them well, and know exactly which animals are the "normal" hosts, and which organ in that animal is the adult fluke's "normal" home. Fluke disease is when any of these is "wrong."

Flukes don't have eyes to see with or legs to walk with, so how can they find and travel to the organ they want in the middle of your body? Scientists do not know for sure. However it's concluded from many scientific studies that the liver fluke, *Fasciola*, for example, has no trouble seeking out and colonizing the liver.

Here are some examples of what can happen when flukes go "wrong:"

- Adult flukes (any of the four mentioned) in the uterine wall causes cramping and bleeding when it is not menstrual period time. If an adult crosses the wall to the inside and then manages to get out through the fallopian tubes to the abdominal cavity it takes some endometrium with it—causing **endometriosis**.

- If adults develop in the kidneys, it can cause **lupus** or **Hodgkin's disease**.
- If adults complete their cycle in the brain, **Alzheimer's** disease and **multiple sclerosis** result.
- If the intestinal fluke (*Fasciolopsis buskii*) becomes adult in the liver it causes **cancers** of many (hundreds) kinds.
- If the pancreatic fluke completes its cycle in the pancreas it leads to **diabetes**. This is not an example of flukes straying into the wrong organs, but of having its stages reproducing where they never could before.
- If flukes develop in the thymus, immunity is lowered. If it happens to be the intestinal fluke, **HIV** (Human Immunodeficiency Virus) is released there. In turn, HIV invades other tissues, like penis and vagina.
- These four flukes can also invade the muscles, causing **dystrophies**.

As dissimilar as we always thought these diseases to be, it's obvious to me that they are but one disease—fluke disease

Considering the size of these flukes (adults are easily visible), it is not surprising that they can quickly lay waste a human's organs. Yet a human is big and makes a valiant effort to kill the stages, block access to tissues and otherwise battle them.

But only the human's intelligence can be counted on to defeat them. The intelligent approach is to discover what enables these mighty monsters to do their reproducing in our bodies instead of the pond with its snail/minnow secondary hosts.

Flukes and Solvents

The explanation I have found for two of these flukes is the presence of solvents in our bodies. The presence of *isopropyl alcohol* is associated in 100% of cancer cases (over 500 cases) with reproduction of the intestinal fluke stages in a variety of organs causing cancers in these organs.

The presence of *benzene* is associated in 100% of HIV cases (over 100 cases) with reproduction of intestinal fluke stages in the thymus.

The presence of *wood alcohol* is associated in 100% of diabetes cases (over 50 cases) with reproduction of pancreatic fluke stages in the pancreas.

The presence of *xylene* and *toluene* is associated in 100% of Alzheimer cases (over 10 cases) with the reproduction of intestinal fluke stages in the brain.

Much more work needs to be done to examine the relationship between fluke reproduction, the solvent and the chosen organ. But it seems probable that the solvent allows it all to happen. And our intelligence, to save us, must find a solution.

Stopping use of these solvents seems to me to be the most urgent advice. Finding which foods and products are polluted with them is the first step. It is imperative that you test everything you use or eat for solvent pollution. The Syncrometer makes that an easy task. Ideally, we should all pool our results, adding to the body of knowledge I have begun.

In my observations, when the big sources of solvents are stopped, the body's levels go back to zero. In other words, the minute amounts that we inhale here and there do not accumulate to the point of serious damage. We have to eat, drink or absorb them on a daily basis to injure us! So where can they come from?

The sources of benzene and propyl alcohol that I found are given in special lists (page 354 and 335). The sources of wood alcohol are not as well known, but include commercial beverages, cold cereals, artificial sweetener, vitamins, and drugs. Other solvents are even less studied. But a pattern is emerging: foods and products that require sterilization of bottles and machinery to fill these bottles are polluted with propyl alcohol or wood alcohol. Foods and products containing flavorings or oils are polluted with benzene. Let the buyer (you) be wary! Test your own products if possible. If not, do not purchase them.

There are many other flukes and many other diseases. Are there other fluke/solvent/disease trios? Has fluke disease been going on for a long time or is it a recent phenomenon? Certainly cancer is 100 years old, so is the use of propyl alcohol. Diabetes is quite old as an illness, too, and so is its associated solvent, wood alcohol. But HIV, AIDS and Alzheimer's are recent diseases. Should we conclude that benzene, xylene and toluene were used much less in the past?

Fluke diseases could be eradicated with some simple actions: monitoring of solvents in foods, feeds and products. Hopefully, this will begin. It is in the interest of the consumer to have her or his own independent way of monitoring too. Chemical ways can be devised, besides the electronic way presented in this book. Imagine a small test strip like a flat toothpick which turns color when in contact with propyl alcohol. Keep a pack in your pocket and never be unknowingly dosed again...all in tomorrow's world.

Flukes Not Alone

There are other families of parasites. The roundworms and tapeworms are gaining ground too. Are they associated with solvents? Or with yet undiscovered factors? Are they changing their life cycles to take advantage of our lowered immunity? These are important questions. But you are armed with excellent technology. The answers will be found.

And along with answers there will surely develop a new industry. An industry that not only proclaims purity for its products but provides the proof to your satisfaction.

Burning And Numbness

Burning sensations in the skin let you know that nerves are involved. Mercury is the most common offender. Mercury may have started the trek of a host of other toxins as well into your nervous system: pesticide, automotive chemicals, household chemicals, fragrance and even food chemicals. Some people can get a burning sensation after a car trip, some when exposed to perfume, some when walking down the soap aisle in a grocery store. When the affected nerves don't go to skin but instead go to an organ like toes, you might feel a toe cramp or finger cramp instead. Remove your mercury sources.

Burning skin is an ancient malady—maybe even the basis of concepts like “hell.” St. Anthony's fire was caused by *ergot* (food fungus) ingestion. Molds don't necessarily come singly. Maybe other mold toxins can go to your nerves, too (see Moldy Food, page 381). Maybe the mold toxins interfere with **pantothenic acid** used by your body, because giving pantothenate (500 mg three times a day) can sometimes relieve the condition and, of course, this is good for your body.

Monosodium glutamate (MSG) can cause burning, especially of the face and lips. Sometimes swelling occurs too. MSG is used as a flavor enhancer. It was found decades ago to be a brain toxin and was taken out of baby food. But what about adults? Especially those who already have a brain problem. Throw it all out of your kitchen. Ask at restaurants which foods have it.

Numbness has a similar cause. Numbness of fingers or feet has become quite common since thallium and mercury toxicity has spread so widely.

If you have burning and numbness, can multiple sclerosis (MS) be far away? Remove all the metal in your dentalware immediately, replacing with composite (see Dental Cleanup, page 409). Hopefully, your immune system is still strong enough to clear the bacteria growing around the metal and in pockets in the jaw. Check and clean cavitations. Start your jaw healing with the milk, magnesium, vitamin D diet. Use thioctic acid to help clear your tissues of remaining metal (3 to 6 a day).

The most common nervous system bacterium is *Shigella*. Its beginning can already be seen in cases of burning and numbness. It is deeply entrenched in cases of MS. Three kinds of *Shigella* are readily obtainable on slides: *Shigella dysenteriae*, *Shigella flexneri, Shigella sonnei*. *Shigella flex* causes depression and irritability. All can cause gas and bloating. Zapping doesn't kill all of them because they inhabit the bowel. You must empty your bowels frequently, two or three times a day. And do the Bowel Program until all symptoms are gone. Be very careful not to put contaminated dairy food in your mouth again. Boil all milk products. They must be at boiling point for 10 seconds. This includes cheese, cottage cheese, buttermilk and regular milk. Eat only home made yogurt or home made buttermilk. Butter and whipping cream need this treatment, too. Never use raw dairy products.

Nana Hughes, 48, had numbness of the whole right arm, hand and right side of her head; it was particularly bad in the last four months. She was toxic with PCBs, titanium and dysprosium (paint). She was on thyroid, Xantac™ (for stomach) and chlorazipate medicines. We also found dog heartworm (she had chest pain over the heart). She started on the parasite program, stopped using nail polish, and stopped all detergents for dishes or laundry. In three weeks her numbness was greatly reduced. She still had titanium buildup from dental metal (a partial bridge). Seeing she was on the right track, she stopped use of MSG in her food, switched from concocted beverages to milk and cut down on her smoking. She brought her arm pretty near to normal.

Maria Santana, 45, had numbness in both arms; they would tingle and "go to sleep" a lot. It was spreading to one leg. Her muscles were toxic with thallium. It was in some very old pesticide, still effective and still in use in the house. It was also in the well water, probably, from pesticide seepage. She went off all commercial body products, did a kidney cleanse and killed parasites. She had difficulty getting rid of *Prosthogonimus* but in two months she had everything cleaned up. Her legs, arms, sleep problem, urinary tract problems were all gone and she could focus on her last problem, digestion.

Candy Donaldson, 44, had numbness from her shoulder to the wrist of one arm, it started a year ago. She was toxic with iridium, lithium and vanadium from a gas leak. She was advised to stop caffeine use and switch to milk (her calcium level was low: 9.0 mg/DL) and a magnesium tablet (300 mg daily). Her triglycerides were high, implying a kidney problem. She had urate and phosphate crystals in her kidney. She decreased the phosphate in her diet (meat, nuts, grains, soda pop) and started the kidney cleanse. When the gas leak was fixed, both her lithium and vanadium toxicity disappeared. In six weeks she had also killed parasites and her periods became regular for the first time. An ovarian cyst had disappeared, too, as checked by ultrasound. Soon her PMS was gone, including hot flushes. After four months she had done three liver cleanses and suddenly her numbness improved. After a few more there was no numbness at all.

If cleaning cavitations brings you immediate improvement you know that these bacteria were part of the problem. Have them checked again if problems return; dental bacteria are notorious for returning. If kidney cleansing makes it worse for a day and then better, you know kidney bacteria are partly responsible. If liver cleanses (page 552) make matters worse for a day and then better, you know bacteria are entrenched in the liver. Continue cleansing until none of these sources exist any longer. Kill all bacteria at least once a week electronically. Burning and numbness can be cured, not just arrested. In other words, your nerves can be cleaned up and healed.

Depression

All persons I have seen with clinical depression had small roundworms in the brain. Is it any wonder the brain can't make enough neurotransmitters or gets them out of balance? The usual worms are hookworms (*Ancylostoma*), *Ascaris* of cats and dogs, *Trichinellas* and *Strongyloides*.

Although it is commonly believed that hookworms penetrate the skin when walking barefoot on earth, this appears to be a negligible route. The important routes are eating animal filth and inhaling filthy dust. Our pets pick these worms up daily. We get them and give them in constant exchange with our pets and family members. Diapering babies is an especially hazardous, though necessary, business. Letting little children clean up after their own bowel movements is even more hazardous. Hands should be sanitized with grain alcohol after dealing with bowel contents, whether your own, your child's, or an animal's. If you clean up a messy diaper and then your hands well with soap, then go to make the chopped salad for dinner, you're sure to give each family member a dose of whatever the baby has. It was hiding under the fingernails. Animals clean up the easy way: they simply lick the youngster's bottom. But, we humans are not strong enough to take on a dose of bottom with each meal. **We must sanitize our hands**.

Cleaning bathrooms is also dangerous. Wear gloves. If nobody suffers from depression, you can use bleach (stored in the garage) to disinfect the stool, otherwise use alcohol (50% grain alcohol). Keep doorknobs and faucet handles wiped with alcohol, too.

If you suffer from depression use your zapper to immediately kill these four roundworm species: *Ancylostoma*, *Ascaris*, *Trichinella* and *Strongyloides*.

Other family members should be cleared of these four worms on the same day or as close to it as possible. Reinfection always occurs. In the depressed person, the microscopic parasites travel immediately to the brain. In others, they may simply reside in the intestine or lungs or liver, or other organs. Pathways (routes) to the brain have become established for the depressed person. These must heal before there is any tolerance to reinfection.

Solvents and other toxins, also in the brain, slow down or prevent healing. For this reason the depressed person should do the *Four Clean-ups*.

The bacteria of the *Shigella* family are always seen in depression cases. Kill all *Shigellas* and avoid reinfection by boiling dairy products and not eating those that can't be boiled. Go on the Bowel Program (page 546). Do not eat deli food or hand prepared salads at salad bars. Finally, do the Liver Cleanse, repeating every two weeks. Depression, even of long standing, can lift within days after the brain finally has its territory to itself. Look in the mirror and smile at yourself for your success in vanquishing your invaders. Never again, let these creepy crawlers into the happiness-center of your brain.

Manic Depression

This variety of depression is associated with *Strongyloides*, as the main parasite in the brain. Plus **chlorine** as an allergen! *Strongyloides* is the same worm that causes migraines and other severe types of recurrent headache. It probably depends on where these tiny worms have set up their “housekeeping” which symptoms you get: manic depression or migraines. Maybe chlorine is the deciding factor.

These tiny wormlets can pass the placenta into the unborn fetus. Is it any wonder that these brain disturbances seem to be inherited? Of course, there is ample opportunity to simply eat them off hands, other people's hands, and off floors during childhood. The amazing truth is that some family members do not get infected with it or at least do not get brain symptoms! It is very difficult to eradicate *Strongyloides* in a whole family and thereby let the depressed person get well. It is, in fact, impossible if there is a pet or other animal connection. Getting rid of the chlorine allergy is also very challenging. Step one is to zap all parasites.

I usually see an accumulation of **bromine** as well. Since bromine, fluorine and chlorine are all *halogens,* maybe getting too much of any one could saturate the liver's ability to detoxify chlorine. Stop eating brominated (bleached) bread. Stop taking drugs containing bromides. That is the easy part. You must also stop even washing your face in chlorinated water (use a pure carbon filter system). You inhale it as you wash. Never drink it or use it for any purpose. Of course, there should be no bleach container in the house, even when tightly closed; nor should bleached clothing be worn.

Do the *Four Clean-ups*. Sometimes the cloud lifts and the mind clears on the way home from the dentist. Recovery can be very rapid—less than a week. And recurrence just as rapid when a tiny bit of chlorine is inhaled or drunk. But your careful vigilance pays off. In half a year you can expect no recurrence.

Humans, it seems, must lick fingers with the same compulsion that cows lick their noses and cats lick their rears. The single, most significant advance in human hygiene would most assuredly be stopping the hand to mouth habit. We must always eat with utensils. And never leave a bathroom without washing. These are difficult changes but the new age of parasitism makes it necessary. Together with the new pollutants, solvents, and heavy metals, parasites will overtake us unless we change.

Don't go off lithium and other medicines until your doctor agrees you are ready, about a half year. After this, make sure you still keep it handy. Although you may be free of manic depression in a day, reinfecting yourself weeks later will attack your brain like a hurricane; it has not yet healed, the routes are open. Be patient; healing will happen as it did for these happy persons.

Lena Constantine, age 39, had a history of migraine, TMJ, heavy clotting with periods and numerous pains but it shouldn't have made her try suicide 1½ years ago. She was put on Prozac™ afterward. She was parasitized by intestinal flukes (in the intestine), dog whipworm, *Strongyloides* and human liver flukes. She was started on a parasite program and kidney cleanse. This made her feel so good she took herself off Prozac™ and landed in the hospital for reasons she couldn't remember. After 42 days spent there she got out, wiser than before. She set to work again, leaving no detail undone, because she could remember how good it felt to be free of depression (not drugged out of it). Three months later she still had *Strongyloides* (she had a cat) but she did her first liver cleanse anyway. She got over 500 stones out. Her depression was gone. She substituted 4 ornithine and 2 ginseng capsules daily (more if tension was not relieved) for Prozac and cured her problem.

Mona Zabala, 33, was extremely depressed over her job. She was full of mercury, arsenic, PCBs, chromate from eye liner, wood alcohol from drinking colas and she had *Ascaris*, pancreatic flukes, *Trichinellas* and *Strongyloides*. She thought she was a hopeless case. But in less than three months, when only half her clean-up chores were done, she was already saying positive things about her job.

Acey O'Hara, a young graduate student, was very careful to avoid junk food, caffeine, fragrance, and body products because he had learned the hard way that he didn't tolerate them. When "crying" depression hit him he was not only surprised but angry that his good health habits "hadn't paid off." He was buying his drinking water; it came in clear plastic containers. He had cesium in his tissues and brain. When he switched back to plain tap water (filtered in small quantities) the depression lifted in a week and he was no longer crying over anything.

Leisa Underwood, 46, was clinically depressed and in therapy. She believed it was due to going through menopause. She had a heavy burden of *Strongyloides* but no other worms. Only one of her two dogs had *Strongyloides* (saliva test) and the cat was free of them also. She was full of cesium (from drinking refrigerator water) and vanadium (from a gas leak). In two months she had accomplished the impossible: all pets and herself were free of *Strongyloides*, they had repaired three gas leaks and her depression was just a memory.

Roland Greeley was diagnosed as manic depressive a few years ago and was put on Ativan™ and Prozac.™ This controlled him but did not relieve depression. He had a prominent tremor. He had *Ascaris* and sheep liver fluke stages in the brain. Also *Trichinella*, *Strongyloides* and human liver fluke stages. Styrene (from styrofoam cups), methyl ethyl ketone (beverage) and carbon tetrachloride were in his brain also, probably setting the stage for parasite reproduction. He had high levels of mercury and silver but highest of all–throughout his body–was chlorine (from bleach and tap water). In four weeks he had all the metal removed from his mouth. He stated that this “unchained” him. He could already tell on his way home from the dentist that something special had happened. His depression was simply gone “the way you lift a blanket off a bed”. He resolved to clean up his whole body and recover from his illness using logical methods, like ours. Staying away from regular chlorinated water was a fine challenge to his resolve but with whole house filtering now available he may have done it.

Darren Knox, age 48, had been on Thorazine™ for 36 years but was recently changed to Desyral™ and Valium™ and lithium. He had *Ascaris* and hookworm and two dozen more assorted parasites including fluke stages. His brain was also hosting *Bacteroides* and *Nocardia*. All parasites were killed in half an hour by frequency generator at his first visit whereupon he immediately announced himself free of depression; better than the last eight years. He had many metals in his brain, also chlorine. In four weeks his tremor was way down, after dechlorinating his tap water. His depression never came back after dental work was done.

Schizophrenia

Much more mold toxin was seen in schizophrenic families than in other kinds of illness. They usually had four or more kinds of mold toxins at the same time, meaning that one toxin was not detoxified before the next was already eaten. It may reflect on the liver as much as the food selected. The initial injury, though, to the liver may have occurred in early childhood.

Schizophrenia does not require mercury or other dental metal pollution for its expression. This pattern is logical when it is seen that young children can have schizophrenia. Schizophrenia is an ancient illness, being described in some very old literature, before dentistry existed. But parasites existed. And moldy food existed. And copper existed!

Ergot is always seen in sick persons. Other mycotoxins are also present, including *sterigmatocystin*, *cytochalasin B*, and *aflatoxin*. As the mycotoxin panorama changes, brain symptoms can change from compulsive hand washing to paranoia or from hearing voices to meanness in disposition.

It is my belief that the current increase in violent crime in U.S. society can be attributed to the especially high levels of mold consumption in foods and beverages, with the resultant effect that behavior erupts into violence with almost no provocation, merely some frustration. It would not be difficult or expensive to experiment with a mold-free diet in our prisons.

A bacterium, *Mycobacterium phlei*, may also be a cause because I have seen a high correlation between it and schizophrenia cases. What its specific effects are, I don't yet know. This bacterium may be hiding under teeth. It is also seen in dogs.

Shigella bacteria from nonsterile dairy foods are part of the problem, too. They produce brain and nerve toxins resulting in irritability, depression, anger. In fact, anger can be so intense, it erupts in violence.

Lead and copper are commonly seen in schizophrenia sufferers. The usual source for these is the household water (household plumbing may have lead solder joints). Change plumbing to PVC, then take thioctic acid (100 mg, twice a day).

Parasites always found in schizophrenia are hookworms (4 *Ancylostoma* varieties) in the brain.

Instructions:

1. Stop eating all grain products, nut products and syrups immediately.
2. Sterilize all dairy products.
3. Search for lead and copper in the water.
4. Stop drinking commercial beverages, including bottled water.
5. Zap the parasites in the whole family for three days, followed by repetitions twice a week.
6. Eliminate *Shigellas* with the Bowel Program (page 546).
7. Give away pets. Schizophrenia is too serious to risk reinfection.
8. Do the dental cleanup (page 409).

The need to zap repeatedly arises from reinfection. Nothing needs to be killed twice. But reinfection from self, family members, pets and food happens every hour. One cannot rely on zapping to stay well. Do a thorough diagnostic search of all foods eaten at the last meal, the water drunk, the air breathed. Only when all these are clean, will the brain heal.

When can you eat grains again? When you are well. Find a cereal or pasta that has no mold. Test it for the mold frequencies (77, 88, 100, 126, 131, 177, 188, 232, 242, 277, 281, 288, 295

KHz). Search especially for ergot (295). Remember, ergot makes *lysergic acid diethylamide* (LSD), not a good thing for a schizophrenic brain. Honeys, too, have ergot. Ergot takes longer (up to 20 minutes) to detoxify with added vitamin C. (Read the section on Moldy Food, page 381). If you become ill again after starting grains, go off again. Take 500 mg. niacinamide 3 times a day to speed up ergot detoxification by the liver. Avoidance is much easier.

Healing of the brain is very rapid; in less than one week feelings and behavior are more normal. Perhaps there are herbs that hasten healing; considering how old the illness is, there must surely be several useful herbs. But considering that herbs, too, can be moldy, be very careful to search for molds electronically before using any herbs. Blue vervain, sage, and ginseng are herbs worth trying.

A question may have popped into your mind as well as mine. If the whole family is eating moldy food, why aren't they all schizophrenic? In fact, family members usually do suffer from some symptoms that are similar to the victim. But every person's collection of parasites and pollutants is unique. Certainly, the whole family should obey the moldy food rules, in order to function better.

Autism

Childhood brain disorders suggest an inherited genetic defect. Yet numerous parasites and pollutants are able to pass into the unborn child through the placenta. Even some bacteria and viruses can.

Lead accumulation is always seen in children with autism. Was it acquired before birth?

Mercury can be transmitted from mother to child, too. If mercury fillings were not removed before pregnancy, have them replaced as soon after delivery as possible (never during pregnancy since the removal itself causes a surge of heavy metal to enter the body).

The common tiny worms such as *Ascaris*, hookworm, *Strongyloides* and *Trichinellas* easily enter the brain. They are present in dirt. Dirt enters buildings via shoes. Shoes carry animal filth. Don't let children put their own shoes on until they have learned to avoid touching the soles. Don't let your child crawl on the floor of a public building, even though it looks glossy. Don't set shoes on furniture or table! A preferred habit is to leave shoes at the door.

Once wormlets have found a pathway to the child's brain it is difficult to reroute them. They must all be killed repeatedly since there is daily reinfection from putting hands in mouths. All family members should kill these parasites weekly to protect the child with autism. When lead and parasites are gone consistently for several weeks the pathway to the brain heals and reinfection no longer sends them to the brain and your child can resume a normal life.

I have not treated enough cases to point to a particular parasite or pollutant. For this reason you must do a total cleanup: body, environment, dental, diet (especially solvents and molds).

Leon Dickson, age 10, appeared normal at birth but he crawled and walked late. At age 3½ he began having seizures and was also diagnosed with autism. (Seizures are caused by *Ascaris* larvae, they probably began the routing to the brain). He was started on Phenobarbital™ by his clinical doctor; switched to Dilantin™ then two other drugs were tried. Now he was on Tegretol™ plus Depacote.™ The mother used no anti nausea medicine during pregnancy, no caffeine, no alcohol or nicotine, not even a single aspirin. In spite of his medicines he seemed to have headaches, continued to have seizures which disturbed his sleep and he wasn't trying to talk. He vomited a lot. He would take no pills or drops (no herbs even mixed with honey) and our frequency generator method was not discovered at that time. He couldn't kill his parasites and *Shigellas*. His brain was full of thallium (not mercury!)

besides the roundworm larvae. He must have gotten this from Q-tips™ or cotton balls or some other mercury-sterilized product. Leon's favorite food was grilled chicken which he ate nearly every day, as well as two eggs a day.

His diet was changed to exclude chicken, eggs, bacon, chips, preservatives and colors in foods, grape jelly and strawberry jam. One month later he had not improved, nor had they been able to kill his parasites with the herbal recipe. The diet change was extremely difficult; he was screaming for his favorite junk food and the whole family was upset over his restrictions. But we encouraged the mother to stick to her purpose, get a different baby-sitter who would obey her, and to try to get some parasite herbs and thioctic acid (100 mg. daily, stirred into honey, to take out lead) down him. The first week the new baby-sitter succeeded in getting him to take thioctic acid. He had only one seizure that week. This encouraged the mother to enforce the parasite program and the diet rules. He became attentive in one month and tried to voice sounds.

The problem with eggs is that they are contaminated with *Salmonellas*. I find, however that it is the outside of the eggshell and the carton that is contaminated. The safe way to handle eggs is to remove them and return the carton to the refrigerator, then wash the eggs and your hands before cracking them. Eggs also contain malvin, a known seizure trigger.

Kirk Peeples, age 5, did not have any words yet but he would point to something and voice M-M-M to mean he wanted it (usually food). He was in a special school and “doing well.” He had the brain toxin MSG, and antioxidants BHT and BHA accumulated in him. In other words, his liver was not able to detoxify these common food chemicals. At that time, I had not found the mold toxins yet; I would now surmise that they were responsible for the liver's disability. Besides going off these food additives he was “desensitized” to them with homeopathic drops by an alternative allergist. The result was immediate. In one week he tried several new sounds and managed his first word: “box.” After killing parasites with the herbal program over a four week time period he could speak 19 words. The parents were bewildered with joy. They cleaned the toxins out of their home so thoroughly, there was no place to sit and nothing familiar to eat. But their son could say things and the parents loved each new sound as if it came from a newborn baby. In two months he was saying two-syllable words and putting phrases together. He needed us no more (he was on homeopathic thyroid drops, too).

Geoff Berkely was diagnosed autistic. He went about his business of examining everything in the office without a word. He was infested with both species of *Ascaris* (there was a pet dog) and was started on the herbal parasite program: just a little less than the adult doses. (Children's' syrups are not as effective.) He was toxic with mercury, too. He had 2 or 3 "baby" root canals. These were to be pulled out. He was also taken off food color, MSG, chicken and eggs. The parents accomplished all this very quickly and called to say their child had become normal but they didn't want us to spread the good news to the "autism" club they attended for fear of criticism for doing unorthodox things. Only in the USA could such thinking occur!

Digestion Problems

Burping, bloating, and being gassy are signs that your digestion isn't perfect. What has gone wrong? Burps are gases escaping upward. Bloating is due to bubbles of gas causing pressure. Your body does not make gas! Gas can only be made by bacteria. The immediate conclusion is that bacteria are growing in your digestive tract (stomach and intestines) that should not be allowed to do so. They are likely to be the common enteric (digestive tract) bacteria: *Salmonellas*, *Shigellas*, *E. coli*, *Bacteroides fragilis*.

You may wish to identify them before killing them with the zapper. Or you can sweep through the whole bacterial and viral range killing all with a frequency generator. The good effects can be felt in an hour, although the last gases may take days to get rid of. *Salmonella* and *Bacteroides fragilis* are two bacteria that can eat bile, and can do without oxygen, so they are commonly seen in the liver. *Bacteroides* probably escaped from its roundworm host, *Ascaris*. *Salmonella* probably came with non-sterile dairy food.

If you have an intestinal problem involving digestion or pain, start immediately to boil all dairy foods. Stop eating those that can't be boiled: cheese sandwiches, yogurt, ice cream. It must be boiled for 10 seconds. The bacteria are in the liver because your liver attempted to strain them out of your blood and lymph in order to kill them with bile. Instead, they turned around and “ate” the bile, turning it brown as evidence. Now, every time the liver lets down bile into the intestine (and stomach), a population of these bacteria goes with it. The intestine becomes a seething mass of bacteria, bubbling away as they produce CO_2, SO_2, H_2S, CO (carbon dioxide, sulfur dioxide, hydrogen sulfide, carbon monoxide). Some of these gases are quite toxic.

Help your liver expel its bacterial overload with liver cleanses (page 552) until all the bile is a beautiful bright green. This is evident as dark brown bowel movements. Without the green color of bile added to your intestine, the bowel movement remains light colored, such as tan, yellow or orange! By stopping eating polluted food, killing bacteria and cleansing the liver, digestion becomes normal again.

Of course, there must be enough acid in the stomach and digestive enzymes produced to make good digestion possible. Otherwise, leftover food goes to feed the waiting bacteria.

Persons with a chronic digestion problem may also find they harbor lead, cadmium, or mercury in the intestine! This is very good in a sense. Your body has kept these toxins in the intestine, preventing it from getting into your vital organs. The bad news is that their presence in the intestine could start an intestinal disease. Toxins in the intestine would inhibit your immune system (here the white blood cells are in clumps called *Peyer's patches*), from gobbling up the “bad” enteric bacteria. Maybe the "good" bacteria found your intestine too toxic to live there. Clean up your dentalware. Search for lead and cadmium in your drinking water. After improving your lifestyle, continue to do liver cleanses. Keep your goals high: a flat tummy that feels energetic, no burps or gas.

Stomach ache (page 98) and **hiatal hernia** (page 133) are also digestive problems but are dealt with under pain. (See also lower abdominal pains, page 97, and appendicitis, page 100).

Alan Barth had his two young sons with him. One of them had a very sensitive stomach, a poor appetite, wanting nothing but sweets or chips to eat. The 16 year old had aches and pains that were keeping him out of sports. The father had a tonsil problem; they swelled if he drank milk. One problem was obvious. Their milk was tainted with *Salmonellas* and *Shigellas*, setting up throat problems for the father, stomach problems for one child and a pain syndrome for the other child. Boiling all their milk, not bringing raw chicken into the house (*Salmonella* Source) and stopping eating yogurt and cheese was the solution. There were traces of lead in their tap water and the house air had vanadium in it, announcing a gas leak. They were all heavily parasitized. When these problems were cleaned up, the whole family's health improved.

Kae Nakajima, middle age, had a history of ulcers and stomach problems. She took a lot of Zantac™ and Tagamet™ over the years. She was drinking coffee, tea and colas. She had *Ascaris* and cadmium in her stomach. Five months later she had cleaned up everything except dentalware and was feeling very good. She didn't need any medicines. She still had arthritis and sinus problems but felt so encouraged she had the dental work scheduled.

Sven Lippencott, age 4, had been tube fed for several years due to weak stomach action. He was quite underweight for his age. He had a population of intestinal fluke in his stomach along with arsenic (pesticide). The rest of the family also had intestinal flukes. After killing parasites and washing the upholstered furniture, Sven's appetite went up so he ate more by mouth and grew immediately. But he soon picked up *Ascaris* and cat liver fluke (there was an indoor cat). Then the whole family got intestinal flukes again. They all, including Sven, drank carbonated beverages *ad*

> *libitum* (whenever they wished). With both parents working it would not be possible to discipline the habit. Sven had wood alcohol, methyl butyl ketone, hexanedione, methylene chloride and toluene buildup making his recovery hopeless.

Alzheimer's

What was once a rare disease has become a household word because it now affects so many. Two new pollutants of the brain, inviting an old parasite to a location it would not normally be, is the explanation. The new pollutants are solvents that seek the brain.

Xylene and **toluene** are pollutants of popular beverages, decaffeinated powders and carbonated drinks. At first, the body can detoxify these but with a steady stream of solvent arriving, detoxification slows down and parasites begin to build up in the brain. Common fluke parasites which we eat in undercooked meat and perhaps get from our pets, can now reach the brain and multiply there.

Other toxins are also present, such as aluminum, mercury, freon, thallium, cadmium. Aluminum buildup is seen in all Alzheimer's sufferers (100%). This is undoubtedly part of the true cause. Did it come before or after the parasites?

Whatever the answer, your job is clear. Remove every bit of aluminum from the food and environment. Throw out the pots, the aluminum foil, the cookie sheets, the tea ball. Throw out the kitchen salt, the pickles, the baking powder. Buy things made with baking soda (not baking powder), use a plastic salt shaker, buy salt without added aluminum. Stop using commercial soaps and lotions. Make the soap recipes in this book. Finally, tape over all aluminum handles in the bathroom and elsewhere (e.g. the walker) with masking tape. Then find a chelating doctor to help remove aluminum from the brain. Also use thioctic acid (100 mg; take 2 three times a day).

Kill the four common flukes with the first frequency genrator or zapper you can get your hands on. Prevent reinfection from meats and pets. Stop all commercial beverages, including water. The processing has left xylene and toluene in them. They are not put in intentionally. For the same reason, health food beverages are similarly polluted. Only milk is safely bought from the store. You must still sterilize it, however. Make your own fruit juices. Select beverages from the list of recipes given. Drink water from your cold water tap, filtering it with a small pure carbon filter as in a filter pitcher (see *Sources*).

As much as xylene and toluene are brain-seeking solvents, *Shigella* is a brain-seeking bacterium. The symptoms it causes are not always the same since they depend on the location of infection. Sometimes they cause tremor, sometimes loss of balance, sometimes speech problems. But they are very serious problems. Kill *Shigellas* every day at bedtime with your zapper. Start the Bowel Program (page 546). When improvement is lasting you know you have stopped reinfecting from your own bowel or from polluted dairy products.

Remove dental metal and use thioctic acid as a help to clear tissues of metal. Use vitamin C (3 grams) and B_2 (300 mg) to assist the liver with detoxification. Use B complex (2 a day) to assist the liver generally. Avoid food molds; ergot especially has strong mental effects (see Moldy Food, page 381).

Start a kidney cleanse (page 549) as soon as you can. Follow this with a liver cleanse (page 552). Clean up environment and diet. Your beloved family member or friend with Alzheimer's can regain her or his mental function to a considerable degree. **Most important is stopping the mental deterioration** before it is not reversible.

Lisa Anne Reed, 60ish, was tentatively diagnosed with Alzheimer's 10 years ago. She needed complete care at present but was able to walk (could disappear quickly) and eat. She could occasionally say her name. Her brain had intestinal flukes and their eggs and was toxic with aluminum (cooking pots) and aluminum silicate (salt). One week later she still had the parasites because nobody could skillfully give her the parasite program. She was also toxic with benzene so reinfection was occurring. In addition she was toxic with bromine (from brominated bread?), chlorine (chlorinated water?) lithium, bismuth, vanadium, and tungsten, all known to have strong brain effects. She was also toxic with moth balls and iridium. She harbored *Naegleria*, another brain parasite. In another week there still were no changes due to inability to administer the treatment.

Isabelita Ufford, 77, was in a wheelchair, brought by her two daughters who took turns caring for her. She was on Clanopin™ medicine, did not try to speak and needed total care, including feeding. She had been ill about seven years. She had intestinal flukes and their stages in her brain (the cerebrum) as well as intestine. She also had isopropanol solvent, aluminum, chromate and high levels of arsenic in her body. She was given the parasite herbs plus instruction to get rid of solvents and metals but the plans could not be carried out. The parasites could not be killed without considerably more help than was available. She was a dear, sweet person. The daughters were highly motivated but were overwhelmed with the size of the task.

Beth Hamm, 60ish, arrived led by her ever-vigilant, ever-caring husband. She was started by medical doctors on EDTA chelation to remove aluminum from her brain. My tests showed aluminum, toluene, sheep liver flukes, asbestos and *Shigella* bacteria. The parasites and bacteria were zapped immediately and her husband began the difficult task of excluding non-sterile dairy products from the diet provided. In four days she was able to walk by herself, knowing where she was going. She could finish a short sentence and comply with directions to sit down and get up. Then she had a set back—she had acquired *Salmonellas* in the brain from a bit of dairy food that had slipped by his attention. She was given Lugol’s and she improved further. In ten days she was a new person; an interview of twenty minutes length did not reveal Alzheimer symptoms. Will she be able to hold on to her gains? Only if the aluminum and asbestos are removed from her home environment, his vigilance with dairy food keeps up, and she stays on a maintenance parasite program. But her husband appeared intent on her recovery. She is a person again. And there seems to be a delightful companionship.

Ruben Camberos was brought by his wife and a friend for Alzheimer's. The first day he arrived, the intestinal flukes in his brain were found and killed. He was started on EDTA chelation. They were warned about non sterile dairy foods. His *Shigella* was zapped. Four days later he spoke his first meaningful sentence: it was three words long. He could pay attention to the appointment proceedings. Two days after that he was reading a newspaper. This was utterly shocking to his wife. In another three days, he could hold a conversation consisting of very short sentences. Larger ones became hopelessly garbled. He was started on ornithine (4) and valerian capsules (6) at bedtime: this produced a beautiful nights sleep (especially for his caretakers!) and his days were less agitated. There were still setbacks later but his wife was determined to get him well.

Dementias, Memory Loss

Memory loss is progressive with age but not due to aging. There are plenty of nonagenarians and centenarians with clear minds and good memories to prove that age is not the deciding factor in the dementias. Why do some people deteriorate much sooner? Could you prevent personal deterioration of mental abilities? You probably can. You will know it by noticing memory improvement. Telephone numbers that left you with no recall, unless you wrote them down, number by number, now form groups as you hear them, and you can jot them down the way you always did! This is a good sign of memory improvement. Your writing can improve. The jagged, crooked, misaligned words can be smoothly written again! You can remember things that happened earlier in the day and talk about it later, at mealtime. You can finish your thoughts in conversation.

Mental deterioration of the elderly is not as complicated as is generally believed. Although circulation and blood pressure play a role, **the effect of toxins is much greater**. The action of toxins is greater in age than in youth. The same polluted water and food causes disorientation in the elderly when it only gives a young person a stomach ache.

The liver's detoxification capability may be the real issue. Indeed, the liver may age, in accordance with the calendar date. Perhaps the liver is the only truly aging organ. It may even determine your life span. The answer, then, is to stop giving it toxic substances and shortening your life span.

As the liver is less able to detoxify them, common toxins are allowed to roam the body with the circulation, doing harm to all the organs. The brain feels disoriented or dizzy; there is memory loss. At first, the liver can "catch up" its work and finally clear the toxin for excretion. But, eventually, it can't catch up or keep up. The body, notably the brain, is bathed in toxic chemicals that interfere with its functioning. Now, the elderly person must use a cane for stability, must walk very carefully not to fall, must write everything down to remember it, calls people by their wrong names, can't "find" the right words to speak with, can't finish sentences, must write on a calendar to keep the days straight, starts talking to themselves to help think of things, develops tremors and unsteady gait, acquires a passive personality, loses weight, gets stooped, stops reading the newspaper.

All these signs of aging (dementias) can be reversed by simply removing the common toxins with which we are already familiar.

Of primary significance are food molds. These cause brain hemorrhages. Clean up diet, mouth, body, environment, very meticulously.

Of course, an elderly person cannot bring these changes to herself or himself. If you have a loved one with symptoms of aging, and this person is willing to cooperate with you, you can honestly promise them numerous improvements. Spend a good deal of your effort on persuasion since living longer or being healthier may not seem worth giving up a coffee and doughnut breakfast. On the other hand, they might respond to the goal of needing fewer pills, getting into their own apartment again or becoming freed from a walker.

Walter Heffern, 64, had been to various neurologists but could not find any help. He appeared to have the same kind of mental deterioration as his mother, but at a much earlier age. He couldn't understand an ordinary conversation; he constantly spoke about winning money, walked hesitantly and had to be left undisturbed to accomplish anything—even eating and dressing. He needed a lot of care. We found he had *Ascaris* larvae in the brain—in the cerebrum, where you think. He also had *Acanthocephala*, *Dipetalonema* (a chicken roundworm), amoeba (*Entamoeba histolytica*) and *Fischoedrius* in the thinking part of his brain. He had been in the poultry business all his life: his mother probably shared this exposure, as well as other lifestyle habits that gave them solvents and pollutants besides parasites. He had constant ringing in his ears, this could affect hearing an ordinary conversation. He had a water softener that would have supplied a daily dose of aluminum to the brain, too. There always were dogs in the house. Perhaps the marvel is that he was no worse off, a tribute to human strength in general.

Pushing Back Age

This chapter is a tribute to Jimmy, Nazy, Michele, Suzanne, Marlena and all the others who achieved excellence with daily care-taking of Mary Austin, deceased at 97.

It's true that we have to die sometime. But why die before our life span is up? If many people can live to 100 years, then surely this is the human life span, not three score and ten. Some scientists think the true human life span is closer to 140 years! And that we all lead shortened lives. The shortening is due to failure of some organ in us. Other organs are dependent on the failing organ and begin to fail also. When the brain fails, death occurs, sometimes in five minutes.

If we knew which organ is failing, we could come to its assistance and prevent the collapse of the whole body. Often it is easy to see which organ is failing. But whether this is the true beginning of the body's problems we cannot know. Before death there may have been appetite loss. Before the appetite loss there may have been a broken hip. Before the broken hip, dizziness. Before the dizziness a blood pressure or blood sugar problem. Before these, an episode of “flu” or a dental “repair.” Sometimes we know what started it all. But often we don't. Just make a guess and begin somewhere.

Diet

If your aging friend or relative is in a home for the elderly, you may be able to persuade him or her to choose a diet that is wiser than the average diet people eat there. This can help a lot. Just stopping drinking the coffee, decaf, iced tea and carbonated beverages that are served, and switching to the recipes in this book could get them off some of their medicines.

The beverages to encourage are sterilized milk and hot water—delicious with whipping cream, honey and cinnamon. This gets them away from solvents, oxalic acid and caffeine.

Old age is not a time when you "no longer need milk." Calcium losses increase in old age. Milk has the organic form of calcium, chelated with lactic acid, and it has the cream to promote absorption. For this reason, milk should never be reduced in fat content (not less than 2%). The cream is necessary to improve calcium absorption.

In old age it is downright dangerous to be taking many calcium tablets. The stomach does not have the acid necessary to dissolve them. They pass into the intestine, disturbing its function and acid levels. With tablets, too, one must be careful with dosages, while food is self limiting. No elderly person would be able to drink more than one cup of milk at a time. This contains 250 mg. of calcium.

Milk, however, requires stomach acid to curdle it as the first step in digestion. If there is not sufficient acid, it will pass undigested into the intestine, causing new problems. We must listen to the elderly when they say milk gives them gas or other troubles.

Having the milk warm to hot helps in getting digestion started in the stomach. Milk served hot with cinnamon accomplishes two purposes: it will stimulate acid secretion and the cinnamon is an insulin aid. Milk served hot with honey adds the nutritive value of honey, displacing the need for other unnatural sweets. The meal should always include something sour to curdle the milk. It does not have to be added to the milk; it can simply be included with the meal somewhere.

Lemon juice or vinegar can be put in certain foods but the most reliable way to get it into the diet is to put 1 tablespoon into the water glass along with a teaspoon of honey. This gives the water a "sweet and sour" flavor, enough to make it interesting throughout the meal. The fresh lemon juice or white distilled vinegar and a honey dispenser that is easy to use should always be on the table. Bring these two items to your loved one at the "home" if it cannot be provided regularly and reliably. Pop in at mealtime to check up on it. Powdered vitamin C (¼ tsp.) is another useful acid if the first two are not effective enough.

The lemon and honey habit, alone, can add years (healthier years) to an elderly person. The extra acid taken with lunch and supper (the stomach has its own best supply of acid in the morning, for breakfast) improves overall digestion and helps dissolve the calcium, magnesium, iron, zinc, manganese, and other minerals in the food so they can be absorbed.

The habit of using vinegar and honey in water as a beverage was made famous by Dr. Jarvis in his book *Folk Medicine*, circa 1960. He recommended apple cider vinegar for its extra potassium. In those days, vinegar was made of good apples. Now, all the regular vinegars have mold in them. The toxin, *patulin*, in moldy apples has been carefully studied by scientists. It taints the vinegar as well as apple juice and concentrate made from them. I have not tested patulin to see if it can be detoxified by vitamin C. We must use only white distilled vinegar, even though it lacks potassium, aroma and popularity. Using a variety of honeys can make up for the need to vary the flavor. Get orange blossom, linden blossom, buckwheat, wildflower, and sage honey, besides clover blossom.

But honey is not perfect food. It usually has *ergot* mold, a very serious toxin. To detoxify the ergot, you simply add vitamin C to the honey as soon as it arrives from the supermarket. This gives it plenty of time to react with the ergot before you eat it. Bring your "fixed" honeys to the home.

If your elderly loved one has not tolerated milk in years, start with the vinegar and honey beverage, or lemon and honey, and be patient until that is accepted. Then add only ¼ cup milk to the day's diet, (in the morning, on homemade cereal). Go up very gradually and only when digestion allows it. Of course, the milk must be sterile.

If it is not sterile, the final warming will only increase the bacterial count. You must be sure of its sterility. Boil the milk yourself. Near-boiling is not hot enough. It must be heated until it bubbles up and almost goes over the container for ten seconds. Use a non-metal pot that holds one to two quarts. You may throw away the skin. Then cool and refrigerate. Supply it to the home, too.

Milk that is marketed in paper containers that need no refrigeration has been sterilized; it is safe.

Once the body, even an aged body, finds a nutritious food that does not cause troubles of its own, it asks for more. Your loved one will accept it and drink it without forceful coaxing, if there is no problem with it. As long as your loved one tries to avoid drinking it, your challenge is to find the problem and solve it. It is not a matter of taste or habit. It is a matter of digestibility and lack of toxicity. When your loved one is drinking three cups of milk (or buttermilk or whey) a day and three cups of water, there will be no room (nor request) for the usual coffee and tea and other bad beverages.

We all must die of something. But it needn't be a stroke, or heart failure, or cancer. Choose what seems to be the most pressing problem to work on. Common problems that plague the aged are brain problems, incontinence, bad digestion, diabetes, tremor, weakness, feeling cold, sensitivity to noise, losing the sense of taste and smell, hearing loss, insomnia, kidney and heart failure.

Increase Oxygen

Brain problems include memory loss, communication deficit, dementia (calling things by inappropriate names and saying inappropriate things). The brain is simply not getting enough oxygen and food to work right. It is like having a pocket calculator with rundown batteries: it will give you wrong answers (without telling you they are wrong). Not enough oxygen to the brain is the main cause of memory loss, inability to find the right words, getting words mixed up and not being able to speak in sentences. You can prove this by providing oxygen from a tank; modern equipment is very easy to use and inexpensive. If your loved one responds well to a few hours of oxygen, you have proof of the problem.

How can you increase oxygen in the brain?

1. Open the blood vessels wider.
2. Increase pressure of blood running through the blood vessels.
3. Raise the oxygen level in the air that is breathed. Less carbon dioxide, tobacco smoke and auto exhaust. All these, including a gas leak from the pipes in the house, compete with oxygen.
4. Increase the oxygen delivery system to the brain by raising hemoglobin levels. Cure anemia and low iron levels.
5. Raise oxygen saturation of blood by keeping body acidity down.
6. Correct a slow and irregular heart beat.

Niacin

Open blood vessels wider by giving **niacin**. Give it early in the morning, upon rising, as soon as the feet are set on the floor. Keep it at the bedside, use small capsules or tablets and combine this chore with water drinking. The water should not be cold and should have nothing added to make it a beverage. (Drinking water within minutes of sitting upright may also move the bowels soon.) A 250 mg. time-release niacin tablet (see *Sources*) is a good choice. The elderly have little side effects. Even the niacin-flush, which reddens the face and neck is welcomed since it gives a sensation of warmth.

The flush is intensified by giving hot liquids or acids (even vitamin C) to drink. The flush is reduced by giving cold liquids. The opening of blood vessels by niacin only lasts a few hours. If you see it has a good effect on your loved one, give several a day. Do not use a prescription variety, since they are polluted with heavy metals; use only the brand in *Sources*, or a brand that you have tested pure. You can freely experiment with niacin to find the best dosage and variety; it is not toxic in this amount; but the size of the tablet should not turn it into an unpleasant chore. Reduce it if it seems too large to swallow. Don't cut tablets in half, the rough edges can scratch the throat.

Hawthorn berry is an herb that opens blood vessels, particularly to the heart. Carefully watch the effect on blood pressure when using it.

Food mold, particularly ergot, has the opposite effect of niacin. Brain blood vessels are made narrower, cutting down the oxygen supply. Ergot is a common contaminant of grains: don't provide rye or pumpernickel breads or crackers. Don't provide wine or other alcoholic beverages; they are too contaminated with ergot and aflatoxin. Narrowing the blood vessels in the brain can lead to stroke. If you notice an attack of dementia coming, try a niacin tablet (100 mg, not timed release) immediately.

Fig. 34 Alcoholic beverages contain ergot and aflatoxin. Add vitamin C (1/8 tsp.) upon opening and wait 10 minutes for it to act.

TIA’s, Strokes and Purpura

Sometimes, the elderly person is aware of the onset of a brain attack; they may see stars or peculiar shapes or lose vision temporarily. These are called *temporary ischemic attacks* (TIA's). Immediately give a 100 mg tablet of niacin, 1 gram vitamin C, and a B-complex in this order of importance.

TIA’s are caused by spasming in the brain blood vessels. If this causes them to spring a tiny leak somewhere, a part of the brain will not get its usual oxygen and nourishment. A stroke results.

The spasming was probably caused by sorghum molds. Cooking during the manufacturing of sorghum syrup kills the mold but its toxic byproducts (mycotoxins) are still present. Other syrups may have sorghum added, polluting them. Brown sugar is also polluted with sorghum molds, but fortunately you can detoxify this mold with vitamin C as usual. Mix well ¼ tsp. powdered vitamin C with each new (1 lb.) box of brown sugar.

Purple patches, like bruises, on the hands or arms of an elderly person are called *purpura*, and is also caused by sorghum molds. It weakens blood vessels so they break easily.

When an elderly person has purpura or TIA’s and the mold source isn’t obvious, you must track it down. Test in a saliva sample for all the sweetenings used recently (at least an hour ago). At the very least, keep notes on all the sweeteners being used. Look for a common one at the next attack. Never let the trigger food be eaten again.

Blood Pressure

Increase blood pressure if it is habitually too low. Blood pressures below 110 may prevent strokes but ruin the quality of life. Low blood pressures are often due to toxins from foods. But severe salt deprivation in the diet can also cause this. If the adrenal glands are not conserving salt properly, too much sodium is allowed to leave the body with the urine, resulting in low sodium levels in the blood and a blood pressure drop. Adrenal performance is improved by taking the kidney herb recipe.

The kidney herbs (page 549), at half dose level (½cup a day instead of a whole cup) can be given daily for three weeks and then on alternate days indefinitely. Be very careful to keep the herbal tea sterile by reheating. Elderly people seem to sense the improvement given by kidney herbs. Try other kidney herbs from time to time: **shave grass, cedar berries, juniper berries, butcher's broom, cornsilk**. None interfere with drugs. Remember, they also have a diuretic effect. Be prepared to use extra paper padding in underwear to help catch the extra urine output. The first few nights may be disturbed by extra urination. Interpret this positively.

After the blood pressure comes up to 115 (systolic) mental performance will be greatly improved. Don't try to bring it up to 120 since this raises the stroke potential without giving you much improvement in performance. Use an electronic device to measure blood pressure, one with a finger cuff, not an arm cuff which can itself induce broken blood vessels. Purchase a device that needs no adjustments of any kind and has automatic cuff tension control (see mail order catalogs if your pharmacy does not have one).

Air Pollution

Improve the quality of the air by lowering the pollution level. Check into carbon dioxide, carbon monoxide, fumes from a gas stove pilot light, auto exhaust from an attached garage or nearby highway, arsenic from household pesticide, PVC from new plastic curtains or carpeting, formaldehyde from wearing new clothing before washing it, asbestos from hair dryers, freon from a refrigerator, fiberglass, and chlorine from running tap water. Some of these displace oxygen, some are simply toxic to the body and lungs. Clean up the air according to the general principles of environmental cleanup (see Four Clean-ups, page 409).

For the elderly pay special attention to **chlorine** in the air. Shower water puts a lot of chlorine into the bathroom air which then distributes itself through the rest of the house. Notice whether your elderly person goes into the bathroom in fair shape mentally but comes out confused, unreasonable. Not every day's chlorine exposure will have the same effect. Trust your judgment. Attach a carbon filter to the shower head (see *Sources*). Purchase a variety that has very simple-to-replace cartridges. Figure out how long it should last and write the date for replacement on the outside of it for your own convenience.

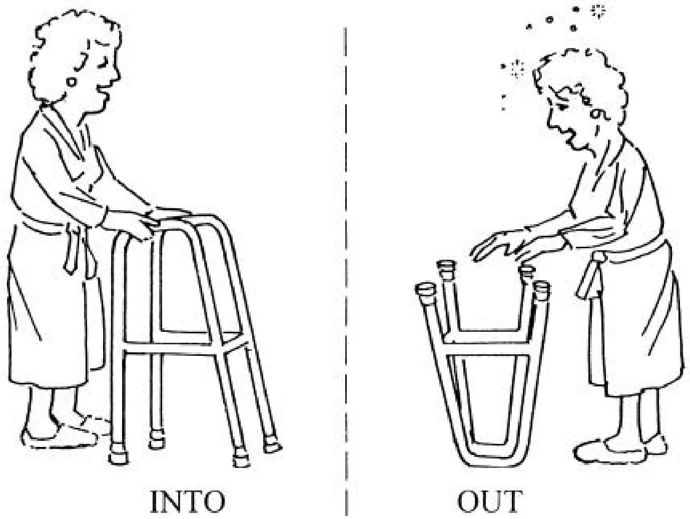

When into and out of the bathroom is a different performance suspect chlorine. Put a filter on your faucets.

Fig. 35 Chlorinated water can cause mental problems.

Washing hands and face in chlorinated water can give off enough chlorine to trigger a manic episode in a manic-depressive person. Certainly, it is enough to cause mental effects in an elderly person. Of course, the chlorine bleach bottle should not be kept under the sink. It should be kept in a closed plastic bag in the garage. It should not be used while the elderly person is in the house and never for his or her laundry. Use chlorine-free bleach.

Don't pollute the air with fresh flowers, potpourris, or room fresheners. These can induce a dizzy spell.

Room air conditioners may have a fiberglass filter! This fills the house with tiny particles of glass to be breathed by everybody. The body makes tumors out of them in order to stop them from cutting through your tissue. Replace the filter with a foam sheet. This sheet is 1/8 inch thick and washable. I have not found these types of foam filters to emit formaldehyde.

Room air filters are not the answer to polluted air. Removing the pollution source is. Air filters may remove some of the toxic elements but by blowing the air (and dust) around vigorously the remaining toxins are made much more vicious in their effect. The noise of a filter motor and fumes it may put out itself adds misery to the simple job of breathing. Using a non-fiberglass filter at the furnace is a better idea.

Make sure all fragrances are removed from the air, even though family members “like” them. They don't belong in air. The lungs treat them like toxins to be coughed up or removed by the kidneys and immune system. This includes colognes, scented tissues, soap and shampoo and shaving supplies. If you can walk into the bathroom blindfolded and know you're in the bathroom, it's not clean enough. Everything in the bathroom pollutes the air of the whole house. People who must use fragrance should apply it outdoors to keep the indoor air less polluted.

Anemia

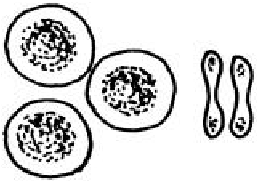

Fig. 36 Red blood cells, top and side view.

Oxygen gets to the brain on a carrier system of tiny rafts called *red blood cells*. They were meant to be an exact shape and size to fit the most oxygen molecules onto them.

Each red blood cell is shaped like a doughnut without the hole. This fits a lot more oxygen, O_2, than round balls would. Yet, if there isn't enough vitamin B_{12}, the dimple isn't put into them to make them doughnut shaped. This reduces the body's oxygen supply and the disorder is called *"pernicious" anemia*. The changed shape of the red blood cells is reflected in a bigger volume called *mean cell volume* (MCV). The correct volume for red blood cells is about 90 cubic microns. Many elderly persons have a MCV over 100!

I have seen pernicious anemia to be associated with *Ascaris* infestation. Kill *Ascaris* on a frequency generator (408 KHz) or zap. The source of *Ascaris* is usually a pet, owned in the past. Once infected, the tiny worms do not leave your body on their own. The infestation may date back to childhood. What a relief for the bone marrow whose job it is to make red blood cells to have enough vitamin B_{12} again! What does *Ascaris* do with your B_{12}? B_{12} is a beautiful rose colored vitamin. Some worms are actually pink from absorbing your B_{12}! Giving B_{12} shots is the current clinical treatment for B_{12} deficiency. Killing *Ascaris* twice a week by zapping and taking B_{12} lozenges (see *Sources*) is a better solution. (If your loved one is getting shots, make sure that no isopropyl alcohol is used on the skin beforehand. Provide vodka yourself in a small pocket flask or 70% grain alcohol for this purpose. Unfortunately, the shot itself may contain traces of this harmful solvent—take a sample home for testing.)

Other kinds of "anemia" can deprive your elderly person of oxygen. A low red blood cell count (under 4.4 million/mm^3) is the "garden" variety. Sometimes iron levels are low and can explain the low red blood cell count. Sometimes they are not.

Most regular anemias, including low iron levels, are associated with hookworm infestations.

Kill all *Ancylostomas* along with *Ascaris* twice a week. It is not wise to take iron pills, even if they do raise hemoglobin levels, except in life-threatening situations. Remove the tiny creatures that cause microscopic bleeding instead. Iron in the form of pills is too easily snatched up by bacteria who also need it, making them more virulent to the body. Use grain alcohol rinse in the bathroom to kill *Ascaris* and hookworm eggs under fingernails. Don't have pets in the house or keep them on daily parasite killing herbs.

It takes nutritious food to build the blood back up to its normal hemoglobin level. Eggs and meats (all very well cooked) are the richest sources of iron and other minerals used in blood building. B_6 and other vitamins are also involved and can be given as a B-complex (see *Sources*). Do not use black strap molasses as an iron source, or any molasses, since it contains toxic molds. (The molds could be detoxified with vitamin C the same way as honey. However, I have not tested enough molasses for solvents and you cannot risk these.) In the past, when the nutrient properties of molasses were discovered, the molasses manufacturing was a small, carefully conducted business. Now it has molds which cause platelet destruction, (purpuric spots) internal bleeding, and immune failure.

Acid Levels

Oxygen must first jump onto its raft, the hemoglobin, in the lungs. Later, in the brain, oxygen must jump off again to enter the brain cells. A difference in acid levels makes this possible. Acid levels operate the latching system that decides whether oxygen will be attached to hemoglobin or let go! Acidity unlatches oxygen. There should be no acidity in the lungs so oxygen can attach here. Sometimes, the entire body is too acid! Diabetics, asthmatics, arthritics, especially, suffer from total body acidity.

Acid was meant to be removed from the blood and loaded into the stomach at mealtime for digestion. When this isn't happening, it was meant to be shipped out of the body with the urine. But the kidneys may be doing a poor job because they are clogged with tiny crystals and because not enough water is drunk, so the body's acid levels rise. You can test total body acidity by measuring the pH of the morning urine. It should not be under (more acid than) 5.5. If the body acid level is too high, help the kidneys excrete it by adding more water to the diet and more minerals to neutralize the acid. The main minerals for this purpose are calcium and magnesium.

Increase Minerals

Adding water to the diet could be the most difficult of tasks if your elderly loved one “doesn't like it.” Calcium should be in the form of milk, magnesium as a tablet. When tablets cannot be swallowed; use magnesium oxide powder (see *Sources*). Use 1/8 tsp. added to cooked cereal, soup, stew, pudding. Magnesium, being a mineral, does not get destroyed as vitamins may. You can add it anywhere in the diet where it won't be tasted. Notice how calming it is to have extra magnesium in this gradual way. And how much better the sleep is at night.

When water "doesn't taste good," there is probably a valid reason. The body may be trying to reject chlorine or other toxins in it. In this case, filter it with a small all-carbon unit that is changed right on schedule. A plastic pitcher (not clear plastic or flexible plastic) with a carbon pack fitted into the top is best. Sterilize it once a week by putting a cup of water and one tbs. of grain alcohol in it and turning it upside down so the filter can soak for 15 minutes. Flush out the alcohol with two pitchers of water. Make sure the temperature suits the person. Temperature can mean everything to the never-thirsty person. Don't allow ice cubes, however, nor beverage making, with the essential water. Adding lemon or vinegar (white distilled) and 1 tsp. honey is probably the best way to stimulate both thirst and appetite.

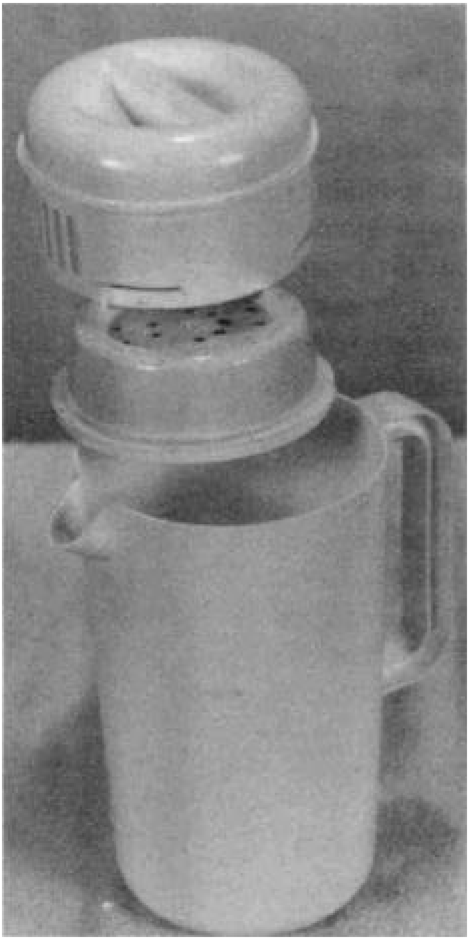

Fig. 37 Water pitcher with filter.

When blood is properly oxygenated it takes on a bright red color, unoxygenated blood is more purple. A chelation doctor can easily see the state of oxygenation.

Chelation

Chelation is a powerful way to quickly improve oxygenation of blood. The most important rule to observe, though, is to take the treatment slowly. Especially if *ethylene diamine tetra acetate* (EDTA) is being used to remove heavy metals, it is important to take the treatment over a two hour time period. Mini doses may be given in a shorter time. Generally, you are in charge of the flow rate. Discuss it with the nurse. Weekly chelations can correct many problems of the elderly that no other treatment could.

Because of hostility from insurance companies who do not wish to add another cost to their ledger and doctors indoctrinated with misinformation, bad publicity is given to this wonderful, life-prolonging mode of treatment. Clinical doctors who have no time to really investigate the statistics of chelation treatments and for whom this is purely competition may feel antagonistic to these treatments. Your loved one should not be the one who must suffer from medical politics. Go to see for yourself what chelation is all about. The receptionist should be glad to show you around. The secret is to talk to the patients themselves. They are usually sitting around a room, eating their lunch and reading as their IV's drip. The pulse, blood pressure and blood chemistry is also carefully monitored. Sit down with them to find out their stories. Get a realistic picture of benefits and costs.

Pulse

The pulse reflects the heartbeat. A **slow pulse** can give weird brain symptoms besides great fatigue. The cause is usually a drug that is being taken to correct a **fast pulse**! Check with the nurse. Read the insert included with packaging for all drugs used. The drugs responsible are usually “beta blockers”, used for the purpose of smoothing out the heart beat, that is, making it regular. Often the drug can be changed.

Less than 60 beats per minute will lead to trouble. For a young person it is a good sign to be as low as 60, provided no drug is involved. But for the elderly it does not reflect a strong athletic heart beat.

The heart is made of four separate “chambers” or compartments each pulsing in turn. They are like four horses pulling a wagon. Unless they pull evenly, the wagon feels jerky, and irregular. The wagon will wear out sooner with jerky pulling. To smooth them out you simply slow them down. Apparently they sense each other better and can pull evenly now.

A heart that is beating 100 times per minute, not unusual for a weak old heart, can be so irregular that it misses every fourth beat. That creates a terrible deficiency of oxygen. Imagine your four cylinder car or lawnmower missing one out of four engine strokes! Beta-blockers have some quite undesirable side effects but heart regularity has a higher priority. So drugs are the immediate choice. Later, when heart health is improved, the heart will beat regularly without drug use. In the meantime, watch over the pulse. When the pulse drops below 60 the new danger is slowness. Take the pulse daily when a new drug has been added, or when you are working on heart health, without getting your loved one anxious about it.

Heart Health

To improve heart health, the first steps of course would be to go off caffeine and to kill parasites and bacteria. This alone could drop the pulse from 120 to 80 in a few days. Obviously, the need for a drug is gone. Cut the drug dosage in half immediately. Don't wait for a doctor's appointment to O.K. it. If you waited another day the pulse could be below 60.

Many common bacteria, especially *Staphylococcus aureus*, choose the heart as their favorite location. Their nesting place, though, will be under a missing tooth in the jaw (cavitation). Heartworm and *Loa loa* are two very common heart parasites. You can have all these killed in a day, without side effects and your heart is once more free to beat regularly. Don't take a chance on over-medicating. As soon as the beat is regular and under 100 per minute, reduce your heart drugs. Stop them when you are regular and under 80. Fatigue will leave and the brain will work better.

Raising potassium levels slows the pulse. Try to do this with diet by eating more potassium rich food and by conserving on potassium losses. The adrenals are in control of losses. Give them a thorough cleaning. The adrenals are situated right on top of the kidneys where all toxic things are being excreted. Being this near to the urinary tract will result in shared toxins. Urinary tract bacteria, small kidney stones, moldy foods and metal from dentalware are the chief offenders. Switch to composite tooth fillings. Use non-metal jewelry. Cook and eat with non-metal ware. Don't handle metal unconsciously throughout the day. Aluminum objects that must be touched should be wrapped in masking tape: this includes walker, shower door, bathroom supports.

Door knobs, taped walker handles, and cane handles should be wiped daily with a grain alcohol solution.

Treat the adrenal glands to 5 supplements

1. **Vitamin C:** shake some into all foods that can absorb a bit of the sour taste, even cooked cereal and vinegar water.
2. **Pantothenic acid:** 500 mg., one daily.
3. **B_6:** 250-500 mg. daily.

4. **Folic acid:** 800 mcg. daily.
5. **B_2:** 150 mg. daily. These are minimum dosages.

If no capsules or tablets can be swallowed put a three day supply in a heavy plastic bag. Pound with a hammer; then roll finer with a big glass jar. Use about 1/3 of the mix each day. If you are trying to do all this in a nursing home, feed it to your loved one while visiting. Put the powder mix in a plastic (not styrofoam) cup, add honey and stir until you get a paste. Feed the paste directly by spoon. Often the elderly prefer it this way in order not to bother with pill taking at meal time. Remind them it will turn the urine yellow.

When the brain problems are corrected for an elderly person, be sure to relate the improvement to him or her. Improvements should not be dismissed lightly. Keep notes. This encourages the elderly, letting them know their existence and quality of life is important to you. Improvements are not merely steps to yet further improvements. Enjoy each bit of progress; it is often too subtle for your loved one to notice even when it is glaringly obvious to you. Before and after a chelation treatment can show a dramatic change in mood, energy, appetite and communication ability, yet get no comment from your loved one. Point it out explicitly so your loved one can look forward to it, too.

Aging is no fun. The elderly long for a brisk gait, laughter, and a picnic at a park again. They dwell on dying, though, because they know it isn't far away and wonder what to do about it. They dare not talk about it because it is too painful a subject for the loved ones. And the immediate problems are too pressing to allow much contemplation of future problems. Talk about aging in its positive aspects. Let your loved one express feelings about it.

Incontinence

plagues most elderly persons. It begins to plague women much earlier—after childbirth, for instance. Surgically shortening the bands that hold the bladder in position (called bladder "lifting") can give temporary relief, but the surgeon may be the first to tell you that it is a temporary fix. Still, it is so shocking not to be able to run a few steps or sneeze or cough without wetting the underwear, that anything seems better than doing nothing. Surgeons will tell you that the bands have been "overstretched."

The real reason why nothing, not even surgery, is permanent is that the support bands are weak. Bacterial invasion causes most of this weakness. Low potassium levels (due to excess potassium losses by the adrenals) causes more weakness. When you kill bacteria (and *Schistosomes* and *Ascaris* and other parasites that bring in bacteria) and blood potassium levels go up, the problem is solved. Overnight you may throw those pads away. Even though you needed three pads to be "safe" you will not need any. Whether you have killed bacteria permanently determines whether you have permanently cured the condition. Make sure all dairy foods are absolutely sterile. Ask that the milk be boiled for ten seconds and other foods that can't be sterilized are not on the menu, like sour cream. Sour cream has too much *tyramine* to be safe. Tyramine is a bacterial by product that is quite toxic; it is rather high in aged cheese, also. With the food bacteria, *Salmonellas* and *Shigella*, out of the way and parasites being killed regularly, you can focus attention on the adrenals which control potassium levels.

Be careful not to rave about the foods that your loved one cannot eat.

Eating more potassium in food is a good nutrition project. Bananas are the top choice. Fresh fruit salad and baked potatoes and soup also provides a lot. Mixing potassium salt with regular salt, half and half, for the shaker is another easy trick, even if you only use it in cooking where the taste cannot be detected. Potassium by prescription is often used by clinicians to conserve body potassium during diuretic use. This need not be stopped (if the pills are not polluted) although taking potassium pills is less useful than salting it in because the adrenals will let any big dose escape anyway. A sign of too much potassium is a slow pulse.

It may be necessary to wear some kind of incontinence underwear. Try to avoid them at night, though, so the skin can breathe freely. Bring a commode near the bed for the night, rather than diapering your loved one (but don't call them diapers; say “underwear”). Absorbent pants of all kinds are heavily chemicalized. This is absorbed by the skin and adds to the toxin level. Less will be absorbed if you powder the skin with cornstarch first. Use them minimally and line them with tissue or paper towel. Chair and bed pads, too, are chemicalized. Don't sit on them with bare skin. To facilitate getting to the commode quickly in the night, dress the elderly in a short night shirt, no pajamas or long gown. Bed socks on the feet help with warmth.

Wash the body parts daily, around the urinary and rectal outlet, using borax water. Follow with 5% grain alcohol. Put washcloth in laundry after a single use. Nothing, not even brain improvement, impresses and encourages an elderly person as much as seeing the incontinence lessen. This bit of progress will put him or her solidly on your side. When they believe in you, it makes your task more rewarding. Remember to enjoy and celebrate your achievements together; don't make a grim business out of it.

Bad Digestion

with its excessive gas and burping is another plague of the elderly. They would rather not go to church nor visit a friend than embarrass themselves in that way.

Chewing

It all begins with the stomach although chewing food well is essential for really good digestion. Dentures should fit perfectly so the mouth does not develop sores. Using denture cream is not a good substitute for correct fit and is toxic. Denture plastic is often toxic, even containing mercury in its composition! Toxins in plastic can seep! Such toxins lower the immunity of the mouth and throat and stomach since it all flows down into the stomach. Low immunity in the mouth permits throat infections to be chronic. If your elderly loved one has a red-looking mouth or throat, instead of pink, an infection is going on in spite of no coughs and no complaints.

It will do no good to keep zapping bacteria when reinfection is so easy. First kill the bacteria in the dentures by soaking in 70% grain alcohol. Then test the dentures for toxins. Soak the dentures in water for several hours. Rinse and soak again in fresh water. Repeat a third time to insure that any toxin found came from the dentures, not the saliva. Save this water for testing. Search for heavy metals in the denture water. If you find any, you know the dentures are toxic! Get new ones, made of uncolored methacrylate (see Dental Cleanup, page 409).

The denture-soak should kill bacteria each night. Plastic has tiny pores where bacteria can hide. Use 70% grain alcohol which you make yourself or plain vodka which is about 50% alcohol. Since alcohol evaporates and is expensive, use a wide mouth jar with close fitting non-metal lid for all this. Fish them out with your toothbrush so it gets sterilized too. It only takes minutes to kill everything. Commercial denture cleaners are much more toxic than grain alcohol; don't use them.

Use food grade hydrogen peroxide or salt water to brush teeth in your mouth, never **toothpaste**. Toothpaste has toxic metals (tin, fluoride, strontium) besides benzene pollution. See the section on brushing teeth (page 532) for details and sources. If you are responsible for this daily chore, use homemade floss (2 pound to 4 pound nylon fish line) first; then brush. If your loved one is seated they may be able to handle the brush by themselves, giving them pride in the achievement.

If an elderly person refuses or can't wear dentures, provide food that is soft and without chunks since this decides whether the stomach can digest it. The stomach is the weak point of the digestive process for the elderly because nearly all don't produce enough acid to get the job done.

Stomach Acid

The body produces hydrochloric acid (HCL) which gets pushed into the stomach from the blood! The enzyme, *carbonic anhydrase*, a zinc enzyme, is involved. Not many ways are known to stimulate this whole process. Drinking water before meals stimulates it in unknown ways but is hard to do for the elderly. Next best is to provide acid.

Because strong HCL would dissolve teeth it is not available as a solution to aid digestion. Ask a pharmacist to make a 1% HCL solution and use 10 drops of it in a beverage at mealtime once a day. HCL as a tablet (“Betaine HCL”) is available but doesn't have enough HCL in it.

Using a lemon or vinegar and honey beverage helps with digestion although this provides citric or acetic acid, not hydrochloric. These acids are completely metabolized so they don't add to the body acid level. But the fact that it is not hydrochloric means that it can't kill bacteria and parasites in the stomach like regular hydrochloric acid could. The stomach becomes a haven for *Salmonellas* and other bacteria and this is **the biggest digestive plague of the elderly**. *Salmonellas* dig deep into the stomach wall, safe from antibiotics and stomach acid and aren't washed away with the food. When they take over the region near the top of the stomach, it weakens the esophageal sphincter and food keeps coming back up a bit—a most uncomfortable development, especially after supper or when lying down.

When the Salmonellas spread out further to invade the diaphragm around the sphincter, the diaphragm weakens, and lets a bit of the stomach up through the hole.

This causes *hiatal hernia* distress. Don't settle your loved one in an easy chair after supper. This presses the stomach upward and the food up, too. Leave them sitting at the table a while, then walk a bit, to get the food down lower. The food will sink lower if some of it can leave the stomach at the lower pyloric end. But if *Salmonellas* are entrenched here, too, the lower end does not have enough action to push the food through the valve. Drugs like Reglan™ are given to speed this up.

What helps most is getting digestion completed. This sets up the natural cues for emptying. Digestive enzyme tablets have been in popular use to help digestion. But they may not be safe since they have not been sterilized. Always try the vinegar and honey method first. Coughing during eating is a sign that the diaphragm is irritated (by a hiatal hernia). If drinking water starts the coughing, omit it at the beginning of meals. Work in sips during the meal.

Salmonella and Shigella

Some *Salmonella* infections can bring dizziness to your elderly person. Dizziness is another plague of the elderly, keeping them from going shopping, getting to church and even from getting around their own homes. Drugs such as Antivert™ are given but only take the edge off the problem. Feeling dizzy can make your loved one home bound and stuck to a walker for every move.

Salmonellas, along with *Shigellas*, produce very toxic substances that cause dizziness. There are three common *Salmonella* varieties: *Salmonella enteriditis*, *Salmonella paratyphi*, and *Salmonella typhimurium* (386, 380, 354 KHz). Kill *Salmonellas* daily for a month by taking Lugol's iodine (6 drops in a half cup water, after meals and bedtime, see *Recipes*). Unfortunately, this will not kill *Shigellas*; follow the Bowel Program (page 546) to get them.

During this time set up a system of sterilizing all dairy products (see Milk, page 425) since this is the source of reinfection. Set up a system of rinsing fingers (and fingernails) in 10% grain alcohol in the bathroom. Deli food and restaurant salads carry *Salmonellas* and *Shigellas*, too. Kill them, routinely, after eating such food due to necessity. A warm stomach full of food at a neutral pH is just the right culture condition for these bacteria. It's like putting yeast into a bowl of warm water, flour and sugar. In half an hour it is overflowing with growth.

Once *Salmonella* is entrenched in an organ it is difficult to eliminate. Only an electric zapper can kill them all (in an organ, not the bowel). If your body has the right conditions (like a low acid stomach) to let them grow you dare not swallow another one! *Shigellas* arrive with dairy foods, too, but prefer the lower intestine as their headquarters. Indigestion that starts right after eating is probably due to *Salmonellas*. If your indigestion comes in the night, this suggests *Shigellas*, since they've had time to reach their favorite place further down.

Campylobacter and *E. coli*, other digestive bacteria, are sometimes the culprits. The Bowel Program is effective against these also. Besides getting digestive improvement you get mental improvement, less depression, less dizziness, less irritability after clearing these up. Remember that eating bacteria and killing them later will not solve the problem. Stopping eating them will.

Other Clues

Digestion problems that remain after eliminating bacteria can be diagnosed in a rational way. Ask these questions:

- Is the stool orangish-yellow, or very pale, instead of greenish brown? If so, bile isn't getting delivered to the small intestine from the liver.
- Is there abdominal pain? (More about this on page 97). It may be due to *Ascaris*, flukes, or other parasites.
- Is there constipation? This will let wastes accumulate, all the longer for bacteria to thrive on them.
- Is there bloating? This is due to gas made by bacteria.
- Does the stool float? If so, it must be lighter than water and contain fat or a great deal of undigested material.

Liver Bile

Bile is necessary for digestion. Absorption of fat and calcium depends on bile mixing with the food. When fat isn't absorbed, it stays in the intestine. Fat is lighter than water; it makes the stool float. Feces should not float. When the stool floats you can assume that calcium isn't being absorbed either, leaving the blood in a deficit which will be taken from the bones.

If the stool floats or is orangish in color prepare your elderly person for a liver cleanse (page 552) to clear a bile duct of obstruction. They get quite fond of these cleanses and will ask to have one. Liver cleanses are completely safe, even for persons in their 80's. One of the stones pictured on page 554 came from a woman age 97. The general rules apply to the extremely elderly: kill all parasites first by zapper if possible, otherwise by herbal parasite killing. Do a kidney cleanse (page 549) first, using half a dose instead of the regular dose, for three to six weeks. Attend your loved one in person for the liver cleanse, have a commode at bedside, protect bedding from accident: use paper underwear if necessary. Share the joy of getting gallstones out painlessly with your loved one; let them see and count them if they wish before you flush them (use a flashlight).

Be extra careful with the skin cleansing. Hot water soothes and heals. Use starch skin soother to dispense onto the wet paper towel, besides borax solution and alcohol. Don't use ordinary soap. The starch skin softener gives the smoothness of soap, and prevents the pain of friction. An elderly person may have no diarrhea at all with the Epsom salts! Evidently the body absorbs all the magnesium so eagerly, none is left in the intestine to absorb water and create diarrhea. It is especially important though to rehydrate your elderly person after a diarrhea. This time they do not balk at water consumption. The liver cleanse, it seems, gives them new thirst as well as new appetite. But it doesn't last long. As the stones from the far corners of the liver move forward, they compact into larger stones and plug the ducts again. Their previous symptoms return. Try to give a cleanse once a month until the dark color of the stool returns and it no longer floats.

The benefits of a liver cleanse will last longer if **valerian herb** is taken the day after the cleanse and from then forward. It may be preventing spasm of the bile ducts. Use 2 oz. of the herb (cut) in 3 cups water. Simmer for 5-10 minutes, let settle or strain. Add honey to sweeten. Give a few tbs. every 4 hours (or 6 capsules) for several days followed by a daily dose at bedtime.

Constipation

If constipation is a problem, use an herbal product rather than a drug until you have removed the cause. **Cascara sagrada** (half dosage) or **prunes** work for many people. Adding **roughage** to the diet is a good solution but often doesn't work. If you try bran, you should add vitamin C and boil it, first, because it is very moldy. But even eating tree branches for supper won't move a bowel that has the wrong bacteria in it.

Bacteria are part of the cause; and part of the result! Constipation increases the bacteria level which causes further constipation! You may solve the constipation problem immediately by zapping. Even though this kills some “good” with some “bad” bacteria, no harm is done. The stool is recolonized in one to two days.

Poop Your Troubles Away

Two bowel movements a day are the minimum necessary for good health. The first one should be in the morning. The morning cup of water, drunk at the bedside has the magical ability to move the bowels. Cold water may fail. But the water effect only works in the early morning. Waiting until after breakfast may not work. Notice the energy lift your loved one gets from this most primitive body cleanse. Take advantage of this to exercise them. Go immediately for the morning walk. This might be the only time of day they can enjoy their walk.

Walking and **liver cleansing** are the most health-promoting activities you can do for your loved one.

Make walking as essential as eating. Walking is not merely walking about the house or shopping. Walking should be done outdoors. Walking is a brisk exercise, done as speedily as possible and lasting at least ½hour. Only if the weather doesn't allow outdoor walking can an indoor walk be substituted. Don't let your elderly person choose whether they will walk that day. To overcome resistance, find a cheerful neighborhood person willing to do this task for pay. The need to respond to a new stranger energizes the elderly more than your persuasion can.

Sugar Regulation

Diabetes is a common development in the elderly. If your loved one is already on a pill for beginning diabetes, take this as your challenge never to let it get worse. It is not a chronic metabolic deterioration. It is a destruction of the pancreas (specifically the islets) by the pancreatic fluke which is attracted to the pancreas by wood alcohol. Zap flukes and eliminate wood alcohol as described in the section on diabetes (page 173).

Use no artificial sweetener and no beverages besides milk, water and the recipes given in this book. This is one regimen your loved one will not resent. They are well motivated to prevent the need for giving themselves daily shots of insulin. Use this motivation to acquire the taste for new foods and beverages.

Elderly Person’s 7 Day Diabetic Diet

Daily supplements: 4 chromium (200 mcg. each), 1 or 2 B-50 complex, vitamin C (1,000 mg. or ¼ tsp.). Take with food.

Breakfast

Choose any one; they need not be in order.

1. Two eggs, (replace carton and wash hands and eggs before cracking), wheat-free, corn-free bread, not toasted (special breads can be found in the freezer section of natural food

stores), with butter and 1 tbs. orange blossom honey. 1 peeled pear, raw, with whipping cream. One cup hot milk with cinnamon.

2. Old fashioned oats, with 1/4 tsp. cinnamon and 1/8 tsp. vitamin C, stirred in just before serving. Mix whipping cream with sterilized milk to make a "half n half" for the cereal. 1 tbs. honey. 1 banana. One half cup milk, one half glass water with honey and vinegar.
3. Fried potatoes with 2 eggs (use only butter, olive oil or lard), 1 cup hot or cold milk. A quartered orange (wash the orange before quartering).
4. Cream of rice, with homemade "half n half" or whipping cream, cinnamon and vitamin C stirred in. 1 cup milk, 1 banana.
5. Cottage cheese, cooked in covered skillet to sterilize. Add chives or peeled fruit (not canned). Wheat-free, corn-free bread, or rice bread with 1 tbs. honey. 1 nectarine or piece of melon. 1 cup hot milk. Water with vinegar and honey.
6. Pancakes or waffles with butter and eggs (no sweetening). Fruit (peeled), hot milk, water.
7. Fruit cup, large bowl of peeled, chopped mixed fruit with whipping cream and 1 tbs. honey and rice bread or other wheat-free, corn-free bread with unsalted butter. 1 cup hot milk with cinnamon.

Remember, all honey must be pretreated with vitamin C. All fruit is peeled and free of blemishes and soft spots. All milk, cream, butter must be sterilized for 10 seconds at full boiling point. Butter must not be "raw". Get wheat-free, corn-free bread at a natural foods store.

Lunch

It is better to have most of the day's calories in the middle of the day than at the end. Arrange for dinner at noon if possible.

Dinner

Choose any one.

1. Green beans with potatoes, meat dish, cabbage apple salad, water with lemon juice and honey, 1 cup hot milk. Water.

   Fresh green beans, especially fava beans contain a substance that is described in old herbal literature to be especially beneficial to diabetics. Don't overcook them—it might harm this substance.

   For the same reason, don't use canned green beans. If “fresh” isn't possible, choose “frozen” but rinse the chemicals off before cooking. Potatoes (not overcooked), peeled to make sure there are no blemishes (contain mold and pesticide) can be cooked with the beans. Cook with onions and oregano for flavoring. Add fresh chopped parsley to the sauce or butter for both green beans and potatoes. Fresh parsley has special herbal goodness (high magnesium, high potassium, diuretic.)

   The meat dish should be overdone. “Fast food” is plopped from the freezer into the boiling grease which browns the outside nicely but can easily leave the inside undercooked. Meat must never be “rare.” There should be no redness near bones! Canned meat is safe from parasites but may have smoke flavoring added (contains benzopyrene) or nitrates. Avoid these chemicals. Avoid MSG too. Whenever a meat dish is not accepted, substitute sardines. Let them choose from a display of six kinds. Purchase the flip-top cans to avoid eating metal grindings from the can opening process.

   Cabbage for salad should be chopped fine enough to be digestible. Add finely chopped apples (peeled) and a few apple seeds and whipping cream for the dressing.

Sweet things are reserved for dessert. Since a diabetic's tissues are not absorbing sugar, they crave it more and more. As the diabetes improves they crave it less. For dessert, serve 1 tbs. of honey to satisfy this craving without endangering their blood sugar regulation. It can be used in the hot milk or in other ways. Undercooking the vegetables also helps slow down the sugar release. Never serve mashed potatoes for this reason.

The drinking water should always have a little vitamin C, lemon juice or vinegar added, and 1 tsp. honey if desired.

2. Asparagus, potato, raw salad, fowl dish, fruit, water with vinegar and honey, 1 cup hot milk.

   The asparagus can be fresh or canned. Bake the potato: not in aluminum foil, not baked until fluffy. Don't let the skin be eaten. Use genuine butter, only, or a homemade sour cream dressing (see *Recipes*). Fresh chopped chives may be added but no regular sour cream since this is very high in *tyramine*, a brain toxin.

   The raw salad should be chopped small enough to be edible by dentures. Use homemade salad dressing with a preference for oil and vinegar styles.

   The fowl dish should be very well done, never “fast food”.

   For dessert, fresh fruit chunks dipped in a homemade honey sauce (honey, water and cinnamon). Less sweets are consumed if you dip the fruit rather than pour the sauce over. Limit the total to 1 tbs. honey. Don't serve grapes or strawberries due to the intense mold problem.

3. Soup, sandwich, fruit, hot milk, water.

   Soup should be homemade from scratch. Add bones and 1 tbs. vinegar (white distilled) or a tomato to the kettle to ensure some calcium leaches out of the bones. A fish chowder serves this purpose very well, too.

The sandwich has lettuce, real butter, and whatever else tastes good (no cheese, bacon bits or condiments). The bread is wheat-free, corn-free, stored in freezer. Homemade salad dressing can be added.

The fruit may be chopped with whipping cream, cinnamon and honey sauce (not more than 1 tbs. honey).

The water may be plain if there was vinegar in the soup.

4. Fish, green beans, potatoes, other greens, fruit, hot milk, water.

   Fried or baked fish is served with lemon or lime. Green beans are served with a cheese sauce so a lot will be eaten. (Cheese sauce: add milk, olive oil to a block of cheese. Melt and cook at least 10 seconds.) Serve au gratin potatoes or scalloped potatoes or any kind of potatoes that will be enjoyed. The extra greens can be beet greens, collards, mustard greens or spinach served with a favorite dressing to make sure it's eaten. (No croutons or bacon bits, though.)

   Never serve dessert if the plate has not been cleared. Your loved one isn't hungry enough. If appetite is very poor, sweets will only worsen the problem. Try to change the menu to stimulate the appetite. Acid foods stimulate; spices and B-vitamins (especially B_1) stimulate; hot foods stimulate. Much appetite is controlled by the liver and brain. Toxins at either location (especially food-derived toxins) tell the body to stop eating. Suspect food molds first, bacteria and chemical additives next.

5. Asparagus, meat dish, white rice (brown rice contains mold), coleslaw, milk, water, ice cream.

   A hot meat dish (no pasta, no wheat flour, no regular gravy) can be fried, cooked or baked, but not grilled. Asparagus is fresh, frozen or canned. Rinse if frozen. Fix it differently than last time. Season rice with parsley and minimal salt and pure herbs like thyme; no MSG or mixed seasoning, make butter sauce. Dessert is homemade ice cream (see *Recipes*).

6. Fish or seafood hot dish. Green peas or peas and onions. Peeled sweet potato with butter (not canned). May switch sweet potato with rice on asparagus day. Sliced tomatoes or cucumber or other raw vegetables with or without dressing. Milk, 1 tbs. honey (can be used on sweet potato).
7. Chili or stew with unlimited rice-bread and butter. If chili produces gas, stay away from it. Serve no canned varieties. Grated carrot salad with stewed raisins added and heavy cream. Milk, water as usual. Blueberry pie, sweet potato pie, custard pie.

If more bread is requested, provide a wheat-free, corn-free variety; but limit bread eating to "after main dish" eating. If not enough milk is drunk: make custard pudding or rice pudding so the daily amount (3 cups) is consumed.

Supper

1. Tuna with salad dressing or tuna salad (no pasta). Non-wheat raisin bread (from natural foods store) and butter. Milk, water.
2. Custard, cooked greens, baked potato. Rice-bread and butter. Herb tea with milk added (single herb only, not mixed herbs).
3. Vegetable soup, homemade, from scratch. Leftover meat sandwich with rice-bread (no deli or cold cuts or luncheon meat). Milk, water.
4. Baked squash with butter, rice pudding with cinnamon, raisins, and honey. Canned salmon or sardines with wheat free, corn free bread. Milk, water.
5. Chili or stew leftovers with wheat free, corn free bread. Custard pudding sweetened with honey, and nutmeg/cinnamon mix. Milk, water.

6. Sardines and rice bread or other wheat free, corn free bread. Homemade tomato juice with celery, strained. Milk, plain water. Pie a la mode (homemade pumpkin or squash pie and homemade ice cream).
7. Potato salad. Leftover meat dish and beans, stewed tomatoes, squash. Baked apple with cinnamon, cream and honey. Milk, water.

Many diabetics lose 50 points (mg/DL) of blood sugar in a few days on this diet. This is why: there is less bread than a diabetic would prefer. There is very little cheese (it must always be boiled in a sauce to sterilize). There is no fruit or vegetable juice except homemade, and not much of that because it crowds out milk and water. If your elderly loved one can't eat all this, make sure there are no snacks consumed between meals that are forbidden.

There is no pasta anywhere. Pasta is mold-ridden. Even after cooking, it may be toxic. There is no wheat or corn bread. The menu is heavy on green beans and asparagus and cinnamon. If by chance, your elderly person hates these and starves themselves to get your sympathy, add a lot more potatoes and rice (never brown) to raise calories.

There is no sweetening other than honey (5 kinds). There are no syrups or sugars. Honey is self limiting—the taste for it is all gone after 1 tbs. Not so for other sugars. The heavy use of cream and butter is offset by no deep fat fried food and little cheese.

Keep in mind that this diet may reduce the need for insulin almost immediately. You may have to cut it in half! But how would you know this? The morning blood sugar test is essential to keep track of changing circumstances. Don't neglect it. Be careful not to use rubbing alcohol when making the finger stick (use vodka or grain alcohol). Your elderly person will feel it is all worthwhile (doing without coffee or pie) if one less insulin shot is needed or if they can go back on tablets instead of insulin. Or even just the knowledge they are staying well controlled and will never have to take insulin shots. Remember to give yourself the credit for a fine accomplishment when your loved one's diabetes improves.

Diabetic Supplements

Several supplements are especially good for diabetics:

- Fenugreek seeds, 3 capsules with each meal.
- Fresh vegetable juice made of raw green beans and carrots (½cup total).
- Bilberry leaves. Maybe they have something in them that helps detoxify wood alcohol, since bilberry leaves are good for eyes, too. Get in capsules or make a tea.

Diabetic Eating Out

Since the rules are always somewhat relaxed when “eating out” a diabetic loved one will badger you to go out with them. If rules are sure to be broken, calculate it into the rest of the day so you can compensate for it.

Extra Diet Tips For The Elderly

Food should taste good.

Eating is a fundamental pleasure of living. In old age there is no other pleasure that can equal enjoyment of food. It is a time when we long for the foods of our own childhood, too. Ethnic foods often had to be given up when children were raised (switched to hot dogs and pizza) but with this diversion gone, a return to family food would be most welcome and most healthful. Ethnic foods were made from scratch. And they certainly were made at home where cleanliness and “persnickitiness” are at their finest! A speck in the batter gets noticed. Not so in a commercial mixing vat. Pots and pans are sanitized with hot water, not chemicals.

Good advice is to return to old fashioned home cooking: with its flour and butter, lard and cream, homemade pasta, olive oil and soup, coarse cereal grains and plain fruit. Gone are the fruit juices, flour mixes, crackers and sweets that fill grocery shelves. What about convenience? Old fashioned cooking took most of the day. It does take 3 or 4 hours to make a soup from scratch. But you then get 3 days off! Each day you reboil it, it is sterile again. Or freeze half of it (take the potatoes out first). It does take a whole morning to make pasta or some ethnic dish. Freeze it in plastic sealed containers so the delicate flavor isn't spoiled. Baking homemade bread is automated now. Do at least this much to get away from the mold-ridden grocery store loaves. Make your own ice cream and nut butters. They last many days and free your schedule.

Time is the great inhibitor but if you have the means or the help, the best advice, nutritionally, is a return to old-fashioned cooking and recipes. Use your new insights to improve them where you can. Don't use your mother's aluminum ware; use her enamel ware or the new glass and ceramic ware. Don't use her copper-bottomed tea kettle or gold-rimmed cups or “silver”. Use her wooden spoons, glass glasses, and plain dishes, her wooden and straw bowls and enamel pots and pans.

Salt

Should you avoid salt? No. But a good salt rule is to either cook with it or have it on the table, but not both. Use aluminum-free sea salt, and make sure the salt is sterilized by heating five minutes at 400°F in a glass pie plate to kill mold. (Sea gulls fly over the salt flats where sea salt is gathered. Their droppings provide a medium for mold.)

Fig. 38 Sea salt flats are often roosting places for sea gulls.

The best salt is a mixture of 1 part of your aluminum-free, sterilized sea salt and 1 part potassium chloride (another kind of salt, see *Sources*). Potassium ousts sodium (salt) from your body, so you can use twice as much of this kind of salt! Also, the extra potassium helps lift fatigue and has other benefits.

Always use a non-metal salt shaker with a closable lid to keep out moisture. Don't put rice in your salt because it invites mold.

Since you, the cook, know where the salt is, (mostly in soups and stews), don't serve as much of these when there is heart and kidney illness or high blood pressure. Don't put salt in cereal, cooking vegetables, or other dishes. Just leave it out! Use herbs instead. **Tip:** encapsulated herbs stay fresher and are more potent. Fenugreek and thyme are the most beneficial of the common cooking herbs. Just open a capsule and season.

Tremor

Tremor is a symptom, not a natural part of aging. The nerves controlling the hands and arms are poisoned. The nerves originate in the brain where the poison has accumulated. What is the poison? Did it happen long ago? No! It could have happened as long as two weeks ago but not longer!

Tremor is the result of ongoing poisoning! It is important to find the poison as soon as you can since the rest of the body will soon be affected, too. Search your memory for the new things that happened in the last two weeks. It is a herculean task but only gets harder each day, so keep notes as you ask: Is there new carpeting? Is there a new furnace? Is there a different water supply? Is there a new hair dresser? Did somebody bring a vase of fresh flowers? Is there a new laundry person? Was the place sprayed for insects? Is there a new medicine (drug) or supplement? Was remodeling done? Is there a new food?

The list is endless and the situation looks hopeless because so many new things can happen in two weeks.

Rather than asking individual questions like these, let's ask only **five general questions** and have the assurance that one of them will catch the culprit.

1. Is it in the **air**? This will catch insecticide, flowers, carpets all together.
2. Is it in the **water**?
3. Is it in the **medicines** or **supplements**?
4. Is it in the **clothing**?
5. Is it in the **food** or on the dishes?

To answer each question, test the item using your Syncrometer searching technique. Make a test substance. Then search a saliva sample for it.

To test the air, take a dust sample off the kitchen counter or table (this gives you fresh dust). Pick up the dust with a few wipes by a small, two inch square of damp paper towel. Place in a resealable baggy.

To test the water, make samples by putting about a tsp. of hot, cold, or filtered water into a resealable baggy with a bit of paper towel in it. Try to get the first morning water before it has run.

To test the medicines and supplements put one of each in resealable baggies.

To test clothing (laundry) use a bit of it, (such as a sock) rolled up tightly.

Testing food is the biggest job. If there are leftovers in the refrigerator or freezer, this helps. You can combine all the leftovers in a single baggy. Frozen things don't need to be thawed for testing. Still, the chance of missing a food culprit is quite high. Be sure to test everything eaten in a two week time period: unusual things like popcorn, candy, crackers, cookies, health foods and special powders. A consolation is that you will find a number of bad foods that are not necessarily the tremor causes but which cause other health problems.

Tremor Remedies

1. Let us imagine that the air (dust) sample proves toxic (resonates with the saliva sample). What is it in the air that is toxic? These are the biggest suspects:

- CFCs (from leaking refrigerant, check refrigerator and air conditioner by removing them from the house and retesting the dust after three days, or simply buy a non-CFC variety)
- vanadium (from leaking house gas from a fuel line—repair)
- arsenic from pesticide (switch to boric acid)
- solvent from a lamp or can of lighter fuel (take them out of the house)

- chlorine from water (use carbon filters; remove bleach bottles)
- asbestos (clothes dryer belt)
- paint thinner, carpet cleaner (remove from house)
- fiberglass from bare insulation somewhere (fix holes in ceiling or wall)
- formaldehyde (new foam bedding, new fabric or clothing)

2. Suppose the water proves toxic (appears in your white blood cells); search for lead, copper, and cadmium. Although municipal water tests occasionally detect small amounts of propyl alcohol, benzene, or wood alcohol, I have never detected them—you need not search for them.

3. If the medicines are toxic just change brands. Your pharmacist can help you find a replacement brand. Find which ones are toxic and stop using them. (Don't use up any of the polluted pills or supplements.) They do more harm than good.

4. If the clothing or dishes are toxic (appear in the saliva), suspect

- cobalt
- PCB
- aluminum

Stop all detergent use. Use borax, or washing soda, or use paper plates and plastic cups (not styrofoam).

5. If the food is toxic, suspect

- food molds
- aluminum
- bacteria

Don't eat that food any more.

6. Any bacterium or toxin that invades certain brain centers can cause tremor. The most common culprits are *Shigella*, mercury, thallium and arsenic. Try to identify these for your sake as well as your loved one.

The *Salmonellas* and *Shigellas* will have come from some food. Be sure all dairy foods, including butter and whipping cream, are sterilized. Parasites come from meats. Be sure no undercooked or fast food meats are eaten. Kill all bacteria, viruses and parasites with a zapper. Sterilize fingernails with grain alcohol to prevent reinfection. Use Lugol's (see *Recipes*) and the Bowel Program (page 546) to clear them up.

Bacteria, coming from teeth and jaw (bone infections, called *cavitations*) may not seem as recent as two weeks. Indeed, they may have been there for ten years. But something recent may have aggravated them, so they now can enter more easily into the blood and brain. You may never find out what caused this. It is wisest to check this possibility with a dentist before doing weeks of other testing. See a dentist who can find and clean cavitations. Do dental repair according to the principles described in the dental section.

Going after a tremor problem in this logical way always finds the cause of tremor

whether its a simple short attack or a situation of long standing tremor with head shaking and drooling. If your situation is extra difficult, you will at least improve it and stop its progression. **This includes Parkinson's cases.**

In cases of Parkinson's disease I often find the bacterium *Clostridium tetani*, well known for causing stiffness. It hides under tooth fillings, too.

Remember, there is a consolation for doing all this work. When you find the culprit, you not only will be stopping the tremor, you will be improving a lot of other conditions along the way. Conditions like hesitant speech, shuffling walk, getting up stiffly and slowly from a chair. These are extra dividends for your loved one. And you have learned which things to avoid for yourself too.

One more thing, don't take no for an answer.

If the problem is a drug, don't accept “No, it can't be changed.” Everything can be changed. Go to higher levels of authority. Doctors are very understanding and sympathetic with your intent. If the problem is leaking house gas, don't accept “We can't find a leak,” implying there is none. Go to a building contractor or the Health Department. Their equipment is better. If you get nowhere, change to electric utilities. If the problem is auto exhaust, don't accept “But we have to use the kitchen door to the garage, because it's cold outside.” Conveniences vs. tremor is no contest. Everyone benefits by excluding auto fumes from the house. Lock the door and cover it with plastic. By the time you have identified the culprits (probably 20 hours of work) surely you have won the right to make changes.

Often others are not impressed. Even when the tremor lessens and the elderly person plainly states they feel better, family members may disregard your recommendations. **Get tough!** It's your loved one. You have something at stake too. Be sympathetic with negative responses. But very firm. Make their choices clear:

- Either the inside door to the garage gets sealed off or the cars and lawn mower get parked outside and anything containing gasoline or solvents gets put in a detached shed.
- Either the girls use their hair dryers in their bedrooms with the door closed or you'll get them all new ones that don't contain asbestos.
- Either the clothes dryer gets a new belt or it is taped securely shut and the laundry is dried on the line or taken to a Laundromat.
- Either the plumbing gets changed (to plastic) or each faucet in the house gets a filter. (If your plumbing is corrod-

ing, not even filters are a good idea—you would have to change them every week!)

Weakness

should not be taken for granted in the elderly. Especially if they themselves complain about it. It isn't normal for them. Sometimes they will describe “spells” of weakness. This is an important clue. Check the pulse immediately. Count for 30 beats at least. Are there missing beats? There should not be! Missing two beats in a row certainly can produce a weak or “sinking” spell. The brain and body need every pulse of blood sent out right on time.

Check into caffeine use first. Take it all away. Caffeine speeds up the heart; then the overworked heart has to “take time out” for itself by missing a beat. Don't switch to decafs because this introduces solvents and new problems. If no other natural beverages appeal, serve hot water with cream and cinnamon. After stopping caffeine use ask: Is the pulse too slow or too fast? The pulse should be between 60 and 80 beats per minute.

If it is lower than 60, a medicine may be at fault. Ask the clinical doctor about it immediately. A slow pulse could certainly bring about weakness. A young athlete may have a slow pulse legitimately, due to having a very strong efficient heart, but your elderly person does not fit this category. The cause must be found.

If the pulse is quite high, over 100 perhaps, this will wear the heart out much sooner than necessary. Ask why it is beating so fast? A probable answer is that it is so weak that it has to beat faster to keep up with its job of circulating the blood. What is making the heart so weak?

Heart Disease

When the heart is enlarged, the valves don't quite close where they should, making the work harder for it and weakening it. It may be called "congestive heart failure." Why is it enlarged? Possibly because it is so weak! Yes, it becomes a vicious cycle, getting worse and worse. But you can break into this cycle and get it all reversed again. The real culprit is parasite invaders and toxic pollutants.

The most common parasite heart invaders are *Dirofilaria*, heartworm "of dogs" and *Loa loa*, another small filaria worm. At one stage these worms are so tiny that they can slide through the smallest blood vessels. They are very contagious. Even persons who don't live with a house dog can pick up heartworm. *Loa loa* is thought to be a tropical parasite but it is alive and thriving in the USA! The source of *Loa loa* seems to be tapeworm stages; this is not a certainty.

Both heartworm and *Loa loa* are very easy to kill with a zapper and both are very easy to pick up again. Treat your elderly person twice a week if there is any heart problem. It makes no difference that the house dog is getting monthly preventive treatments for heartworm. They pick it up daily and have thirty days to develop it and give it to others between treatments. Killing the dog's parasites twice a week with a zapper would be very helpful to you. These heart parasites may not cause any pains, yet disturb the rhythm or the pulse of the heart and cause it to enlarge.

Staphylococcus aureus is a bacterium hiding out in far away places like pockets left under teeth when they were extracted or along root canals. Make sure extractions heal and don't leave permanent cavitations where bacteria can live. Ask a dentist familiar with cavitations to do a mouth search. Once the mouth source is cleaned up, the bacteria do not come back to the heart (after one last zapping). If they do, go back to the dentist!

Killing these three invaders (heartworm, *Loa loa, Staphylococcus aureus*) should cure an irregular heart beat immediately (within a day).

If the elderly person is on a heart-slowing drug, check the pulse twice a day after zapping to make sure it doesn't drop too low. They may need to be off their heart medicine. Nobody will notice the relief of going off this medicine as much as you. The sunshine breaks out! Your loved one can smile again at little things! Even interest in sex returns so watch out! Life is normalized when drugs, especially beta blockers are gone. (Other heart medicines, such as Digitoxin,™ don't have this depressive effect. They are used to make the heart beat stronger, not to affect the rate of beating.)

The pulse should be around 70 beats per minute and perfectly regular.

If it isn't, there is still something wrong. Get rid of toxic body products and house pollutants. Test your air for gas leaks frequently. Gas heat, gas stoves and gas water heaters are notoriously leaky. Weather changes, namely temperature changes make pipes expand or shrink—leaving cracks! The gas is toxic and a small amount can't be smelled. What a predicament! Delivering poisonous house gas to our homes in pipes that are not fail-safe is an archaic practice. Especially when the blood test shows a high “total CO_2" level, near the upper limit, search for an air pollutant like house gas or auto fumes. And read the sections in this book on pulse (page 289) and brain problems (page 278) very closely for more things to check.

With the heart regular again, it will be much stronger, too, since it doesn't have to work against itself. This strength is necessary to push the blood into the farthest “corners” of the body, especially the hands and feet, and warm them up! If your loved one constantly has cold hands or feet, try to improve circulation. Dissolve phosphate crystals with the kidney herb recipe. Give niacin (page 279). Give cayenne capsules (one with each meal). Blood thinning drugs to improve circulation are dangerous—use only if the doctor insists. Monitor blood clotting time if your loved one is on a thinner.

Heart/Kidney Relationship

A strong heart is necessary, too, to push the blood through the kidneys. Often a kidney problem is linked to the heart disorder. Kidneys are made of tubes that get finer and finer. It takes pressure, namely strength, to push the blood through them so wastes and extra water can be let down the kidney tube. Think of the kidneys as a colander full of tiny holes of various sizes that let certain things through them but not bigger things. These holes are constantly being adjusted by the adrenals which sit right on top of the kidneys and "supervise".

If the elderly person is not producing four cups of urine in a day (24 hours), it is not enough. The body cannot get enough cleansing action from less than four cups. More liquid must be consumed. If most of the urine is passed in the night this reflects on unhealthy kidneys. Use the kidney herb recipe—but only half a dose (so it will take six weeks instead of three to see good effects). As the tiny "colander" holes open up there is freer flow and many more trips to the bathroom result. The urine loses its awful odor (no ammonia, acetone and bacteria!) and gets a clear look that shows no sediment. Now that water and wastes (urea and uric acid and other acids) can leave the body quickly through more holes, it takes less pressure from the heart to get blood pushed through the kidneys. This brings relief to the heart because its work is easier. The heart and kidneys work together. Like horse and wagon the heart provides the power and the kidneys follow.

This is why heart medicine and diuretics are commonly given together. *Diuresis* (urine flow) helps the heart and a stronger heart helps the kidneys. Similarly, they fail together. In the old days this was called *dropsy*. Urine that should have left the body is backed up in the tissues. Sometimes it shows up in pockets that hang like giant oranges from the skin.

Even (especially) when the strongest diuretics (Maxzide™, Lasix™) fail to work, even when coupled with strong heart medicine (Digitoxin™), the kidney herb recipe can bail you out of the emergency.

The secret is in the varied actions of different herbs. This makes them work together. Be very careful to keep the herb tea sterile by reheating every fourth day. Freeze unused larger amounts. If too much is drunk at once, especially on the first day, a stomach ache can develop and a pressure felt in the bladder that is most uncomfortable. Go extra slow on the first few days, even though you find it quite tasty, so there is no discomfort (only lots of bathroom visits).

Diuretics

As soon as urine is flowing better, blood pressure may drop. Keep track of this twice a day with a modern electronic finger device (not an arm cuff that itself can break blood vessels). Cut down on drug diuretics gradually, using only ¾dose the first day, then ½dose, then ¼ dose. The amount of urine produced or the weight of the person can be used to assess how effective your method is. **Your goal is to not need any drug diuretics.** Again, mood will improve dramatically when diuretic drugs are removed for your loved one. The sense of humor comes back:

be prepared for some new jokes and new laughter. Give yourself good grades for this accomplishment.

Don't throw out the drug diuretics. After being off them for a while, they become more potent again. So if an emergency (sudden edema) should arise they could again be useful.

With a parasite and pollution-free heart and a low-resistance, freely flowing kidney, some reserve strength will soon be built up. Weak spells are gone and forgotten. Your loved one is walking better, needing less sleep, and a "golden age" finally arrives. It is free of pain, free of medicine, free of shots and doctor visits, free of dementia, free of the dreadful weakness that demands so much help. They are free to enjoy family and friends again. Seeing themselves gain strength and be able to do more for themselves gives the elderly a sense of pride.

Give them the credit for improving their health. When they balk at having to take herbs or vegetable juice, remind them of the days they were on a handful of pills and still had heart failure, pain and kidney disease.

Feeling Comfortable

If all this improvement doesn't warm up their feet or hands (feel them yourself) put extra socks on. Warming up their feet might have the effect of warming up hands, too. But if it doesn't, raise the room temperature.

Being forced to feel cold is an undeserved misery of the aged.

Others must accommodate to the elderly's need for body warmth. The thermometer does not tell all. Comfort is paramount for each of us. Younger persons can undress for comfort. The elderly can't make changes for themselves. Usually, by the time they are complaining of cold, they have suffered a long time, feeling too guilty to request a temperature hike. Ask your loved one. They will appreciate it. A shawl, a lap-blanket, woolen sweater, long underwear and fleecy thermal outerwear help a lot. But if your hand still senses cold feet, the thermostat must still go higher. Cold body temperature is an invitation for fungus and viruses to multiply. Viruses escape into the body daily. Don't give them the advantage.

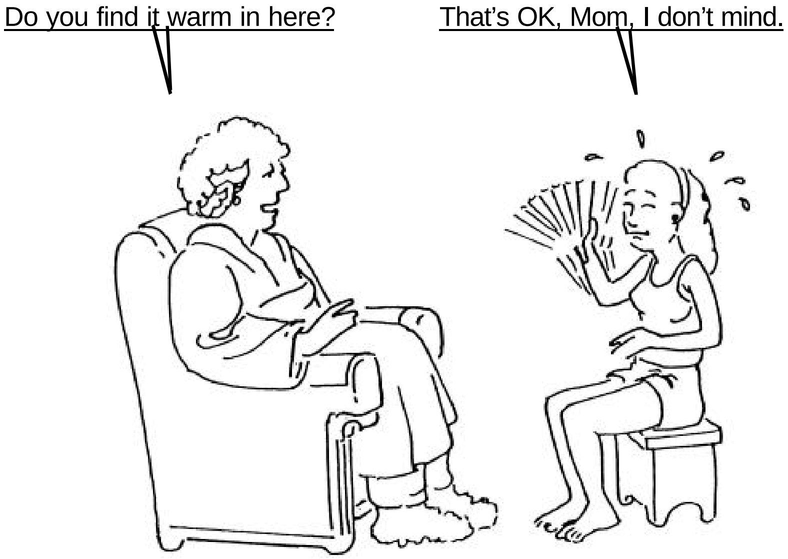
Do you find it warm in here?
That's OK, Mom, I don't mind.

Fig. 39 Too hot and too cold.

But unless all these things are tried, don't raise the temperature. It is much healthier to be warmly dressed and breathe cool air than to be lightly dressed in an 80°F room. The heart, especially, benefits by the cooler air temperature.

For this reason, it is important to have air conditioning during the hot season. Keep your elderly person warmly dressed, away from air conditioner or fan drafts, but keep it cool. As cool as their body temperature and comfort will allow. Don't seat an elderly person under a fan in restaurants. The blowing air not only chills them, but brings filth in dust that blows in the air currents. Bacteria and viruses are soon to follow.

Being comfortable, knowing you are there to care for them, brings out the best in your elderly person. They might get well enough to long for a genuine relationship again. And put you to the test. The test of listening. This can be very rewarding if they are still able to communicate and distill their life experience into wisdom for you. If you can listen and be interested in their distillations or their ramblings their longing for relationship will be fulfilled. There's just one catch, if the new found interest in communication can't be expressed. If they haven't communicated much for a long time, it would be easy for you to miss a simple fact. They can't hear!

Hearing Loss

The hearing deficit in an elderly person is always much greater than they or you realize. Try to understand these communications:

What you (caregiver) say	What the elderly person hears	What they think	What they respond
"Look at the sunshine, isn't it a nice day?"	"Look at something, isn't it a nice bay?"	What a stupid sentence. But she wants an answer. So here goes,	"MmmHmmm" (Meaning, "yes.")
"We'll have such a nice walk."	"Have a…rock."	Rock what? I'm not in a rocking char. I better not say anything.	
"What kind of soup should I make?"	"What kind of hoop should I make?"	I haven't used a hoop for embroidery for a long time. Why is she digging that up. I better say I don't know.	"I don't know."

Don't let this ruin your relationship. It takes the fun out of their life as much as yours. **Get hearing aids**.

Don't ask your loved one if they need a hearing aid. The answer may be “I'm not deaf.” You can go about it more persuasively. Every time you talk, come quite near to the person's ear and speak loudly. They can see that you must come near. If they are fighting against the whole business, they will say things like, “You don't have to shout. I can hear you.” (It's what you're saying that's wrong!)

Arrange for a hearing test. It is free. That will appeal. The results of a hearing test, as it is told by a salesperson, is much more persuasive than you can be. Let the salesperson use his or her special talents to sell your loved one on hearing aids. But you make the choice on quality. Both you and your elderly person deserve the best tone quality that is made. Plus a regular cleaning service. Most companies do offer this but don't tell the customer because of the dreadful amount of time it would take if everybody took advantage of it. You take advantage of it. Clogged hearing aids are the most troublesome feature of any of them—and never mentioned! Make it a rule to buy your batteries at the same hearing aid office where they are cleaned free of charge. This repays them and serves the elderly best. Hearing loss is too subtle to leave to chance; have the hearing aids cleaned each time you buy fresh batteries (about three months). Take your loved one to a nurse for ear cleaning every six months after hearing aids are begun. Wax and debris accumulate behind the aid because the channel is stopped up.

With hearing aids that hear, and kidneys that flush and a heart that beats strongly, your elderly person may choose to attend concerts again, go to church or gatherings—and leave you out of the picture. Give yourself good grades for this achievement. Get them incontinence pants, get regular taxi service. **Do whatever it takes to get your loved one out into the world again!**

If the excitement of a night out keeps him or her from sleeping use **ornithine** and **valerian** capsules. They are good for the health anyway. Hot milk and a piece of cake (homemade, never chocolate) may do as much.

But if insomnia is the rule, not the exception, you need to go after it as a special problem.

Insomnia

is always associated with high levels of parasites and bacteria. This leads me to believe it is their waste products, namely ammonia, that really causes insomnia. Read the section on Insomnia under Sleep Problems for more on this (page 243).

A sound sleep refreshes as nothing else can. Your elderly person will have more energy throughout the day and a better mood if sleep was good.

Of course you must guard against daytime napping if nights are sleepless. Sleep at night is more important than sleep by day. They are not equal. Try to turn your loved one's cycle back to night-time sleeping if it has become turned around by bad kidney function (frequent urination at night) or too much daytime napping. Sleep can be disturbed by taking vitamins at bedtime too. Sleep is enhanced by taking magnesium. Limit bedtime supplements to magnesium, ornithine, valerian (6 capsules) taken with hot milk.

Healthful Habits

If your loved one had his or her way, they would drive the car forever, wear the same cosmetics forever, smoke or chew tobacco forever and eat their favorite dessert forever.

But if you are the caretaker, you know that some things must stop. You also know that gentle persuasion is useless; it merely erodes your relationship. Here are a few tips.

1. Ask your loved one to ask their doctor (clinical doctor or trusted medical advisor) the following question: “Would it be better for my lungs to stop smoking?” Be present so the question does not turn into “Will smoking a few cigarettes once in a while kill me?” Hearing it from the doctor is what's needed.
2. Don't ever purchase something you believe is detrimental to your elderly person. Whether it's coffee, cigarettes, beer or lipstick, say “That is something I can't buy for you; it's against my principles.” Don't be surprised if you cave in a few times to some super ruse they use on you. But the next time, have your answer ready.
3. Let your family and other caretakers know you are no longer supplying these items (the car keys, the wine bottle, the codeine-containing pain pills). Try to get cooperation. Discussing it with your loved one may do more harm than good. If they start the discussion, you end it. This is not a task for the timid! After it's done, you'll wonder what was so difficult.
4. Don't buy a wheelchair if your loved one can still walk with your help. Stay with a cane as long as possible. Then the walker. Stay with a walker as long as possible. Then your personal help. Once a wheelchair has been accepted, the last bit of exercise, walking, is lost. Fight against it. Hide it in a far away closet.

Aging is necessary but chronic illness and pain are not.

If you have managed to free your loved one from having to take pills or from certain disabilities that would soon require pills, you can give yourself great credit. Perhaps you, too, will find the needed natural help when you are aged and have lost your authority and your way mentally. Our lives are all foreshortened, much like the life of a domestic steer's. Does a captive animal learn from seeing its companion disappear? It does nothing to escape its fate. Should we accept our fate with the same docility? None of us can remember how things were in precivilized times. We are eager to believe the present is the best time that has ever been. The steer, too, has its feed provided, its water provided, its shelter for the night provided, seemingly the best time it ever had. Perhaps the price we pay for civilization, like the steer's price, is simply too high. There must be other ways. As a society, we should search for our lost longevity.

Not Old Yet

For persons merely over 40 and not ill a most important rule is: **don't overdo**.

It's easy to injure tendons and muscles by pulling or stretching them too far. Don't try, with almighty determination to open a jar. Leave it for stronger hands. Don't try to stretch a gigantic stretch to reach something on a high shelf. Wait for someone taller to come along and reach it down for you. Don't do exercises, that have left you with strains in the past. Do different exercises. Don't do new exercises, with a determined approach; start moderately; stay moderate. If you start a new job that uses a foot, leg, hand or wrist a lot, more than it was used before, even though it's in ordinary use, it can feel injured. These stresses and strains invite bacteria toward themselves. Bacteria from the liver or your own intestines find these strained tissues immediately and intensify the pain. Kill the bacteria with a zapper, cleanse the liver, and start the Bowel Program if this has already happened to you.

Don't kneel on the floor to do some cleaning job. Don't sit on the floor at a gathering. Don't let your grandchildren take “rides” on you. Don't bend over as far as you can to pick something up. Let your family know you are aiming to reach 100 years in healthy condition.

Super Longevity

We should be able to live to 140 years of age. Middle age might begin at 65, old age at 90. What is the clue to long life? Surely it is keeping all your cells healthy so they can coordinate the constant tasks of nourishing themselves, removing their wastes, plus whatever job that cell was meant to do.

Since your cells divide and therefore start again at age zero, even though you are 90, why do you age at all? Only the nervous system does not divide. Is it the nerve that ages then, and decides death for each of us?

It would never do to live forever. It would clutter the planet. Are evolutionary forces at work preventing this? But we have never cluttered the planet yet. So how could evolutionary forces have “learned” to establish death to prevent overpopulation? Living to age 140 is not living forever. Wisdom comes with age. Wisdom would serve our society well. Perhaps wisdom is needed more than ever for humans to survive. Part of the human tragedy is war. Wisdom gathered from a knowledge of history, might help end wars! But history can't be gathered and understood in less than 100 years. If centenarians can't think well or express themselves, their perspectives are lost. Longevity seems a very useful trait if only it were accompanied by health!

Dying

Although it can't be avoided it can be helped. Dying alone is the most barbaric of all society's practices.

It is still like death in the forest amongst chimpanzees. Unable to feed, to run, to call for help brings pain, fear, loneliness and finally, death. When your loved one is in the mood to talk about dying, listen. Especially if it is about their own imminent death. This doesn't obligate you to carry out any of their wishes. When death approaches the important thing is simply to be there. They may not have requested this, out of a sense of guilt or masochism, or plain dementia. But it is the most primitive of needs, the same as having a loved one nearby during childbirth. It is just a presence that counts. How can you be there if you have a job or are attending a family's needs? It is a time of great frustration for you. All your hard work and successes are culminating in one grand failure!

The good news is that it need not be you who attends your loved one every minute of the last week. Pay for someone to sit—someone who is recognized. But arrange for immediate privacy when you return. No matter how much your loved one admonishes you to go about your business, you will know when you share the final minute that your presence helped. The loneliness of the last coma, the last silence is unspeakable. Give yourself the reward of knowing you shared the pain, fear and silent cry for help.

Curing Cancer

Cancer is no longer the deadly disease it once was. In fact, you can clear it up in less time than it takes to get a doctor's appointment for a check up. If you notice a lump or think you might have cancer, don't rush to see a doctor first. Rush to clear it up, yourself, first. By the time your doctor's appointment arrives, you can be sure of a negative test.

Another Fluke Disease

Cancer is so easily cured because it is a parasite-caused disease. Kill the parasite and you have stopped the cancer. This does not mean you have also stopped being ill. If the cancer damaged your ovaries or prostate, you still must heal these organs. The Ca-125 or PSA test will not drop to normal unless you begin to heal them. These markers reflect the condition of the organ, not its malignant nature. Remember that killing a mosquito does not remove the lump it caused. That will take its course. It will heal beautifully if you let it. But if you scratch it mercilessly, it will take longer. Removing toxins from the affected organs lets them heal.

Fortunately for us, **cancer is not like a fire, unstoppable once it has started**. It takes only 7 minutes to zap all the parasite adults and their stages which cause your cells to multiply.

The responsible parasite is *Fasciolopsis buskii,* the **human intestinal fluke**, a flatworm. It is a human parasite although it can also parasitize other species. It normally lives quietly in the intestine. (The goal for all larger parasites is to live quietly. After all, your demise is their demise.) They were meant to pass their thousands of eggs with your bowel movement, outside, to some pond where snails live. But when the eggs hatch before they leave your body and are allowed to continue their development inside you, the setting is right for cancer to develop.

If the fluke eggs and other stages go through their development in your breast it can become breast cancer. If it is in the prostate it can become prostate cancer. And so on. Each different kind of cancer means the developmental stages of the intestinal fluke are present there. Only one more thing is needed to bring about an avalanche of reproduction, so that hundreds of little larvae turn into hundreds more in a short time: a growth factor. It makes them multiply and your cells are similarly affected. This growth factor, *ortho-phospho-tyrosine* (and possibly, also, *epidermal growth factor* and *insulin-dependent growth factor*) really begins your cancer.

Teamed With A Solvent

The good news is that this growth factor, which is essential for cancer to develop, cannot be made, without the presence of an abnormal solvent, **propyl alcohol** (or more exactly, isopropyl alcohol).

Without taking in propyl alcohol you could never get cancer. It takes two things, together, to give you cancer: propyl alcohol and the human intestinal fluke parasite.

Since it takes a frequency generator (3 minutes at each of 434, 432, 427, 425, 423, 421 KHz) or zapper mere minutes to kill the fluke and its stages you will be stopping the production of growth factors immediately.

Zap yourself every day for three seven minute periods, until after you are completely well.

After killing the flukes, those growth factors already formed will disappear in one hour. Your malignancy is stopped. It cannot return unless you infect yourself with the parasite again!

Getting rid of propyl alcohol is also a simple task. Once you have stopped using it, the last remnants leave your body within three days.

We must marvel at the body's wisdom and capabilities for restoring health. You are not permanently damaged by this large and hungry parasite. Given half a chance your body will throw the rascals out and restore order in your tissues.

Read the list of foods and products that are polluted with propyl alcohol. It is not, of course, usually on the label. If it is used as an ingredient, it is on the label, easy for you to avoid. It is the pollutants not appearing on the label that pose sinister hazards.

INFUSIUM23
INFUSIUM23

All the store bought shampoo I tested had propyl alcohol. Health brands were no exception

CRYSTAL GEYSER

All bottled water is polluted with antiseptics from the bottling procedure. This is a main source of propyl alcohol.

Cosmetics are laced with propyl alcohol. Use the recipes in this book to make your own.

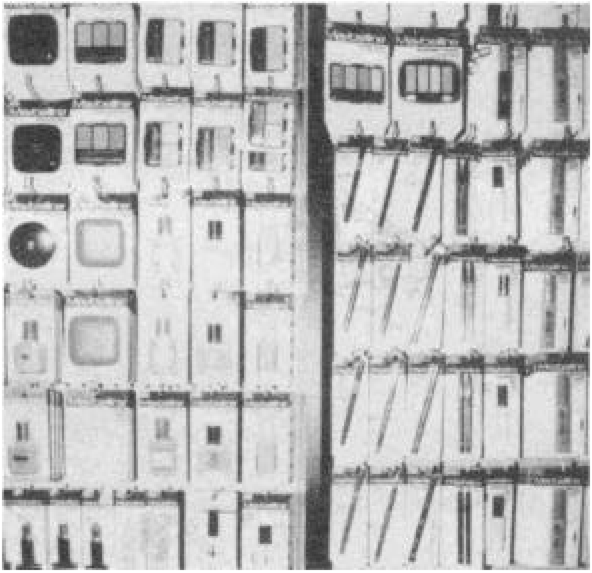

Fig. 40 Products with propyl alcohol.

Propyl Alcohol Polluted Products

THROW THESE OUT

even if propyl alcohol is not listed on the label!

- **shampoo**, even health brands
- **hair spray** and **mousse**
- **cold cereals**, even “natural” granolas
- **cosmetics** (make your own)
- **mouthwash**
- **decaffeinated coffee**, Postum™, herb tea blends (single herb teas are OK)
- **vitamins, minerals and supplements** (unless you test them)
- **bottled water**, distilled water, or spring water
- **rubbing alcohol**
- **white sugar** (brown is OK if detoxified)
- all **shaving supplies** including aftershave
- **carbonated beverages**
- store-bought **fruit juice**, including health food brands

Tear out this page, put it on your refrigerator, and make a copy to stick on your medicine cabinet. Remember propyl alcohol is also called propanol, isopropanol, isopropyl alcohol and rubbing alcohol. You won't drop dead from getting propyl alcohol, but your cancer will flare up with each small addition.

Avoid the entire list, meticulously. Even using one of these, **like your favorite shampoo or bottled water**, will result in failure. Your body will continue to make *human chorionic gonadotropin* (hCG) and the pathology will remain "indeterminate"—not cancerous but not well, either. If you have learned the new bioelectronic technology described in this book, you can test all your foods and products for isopropyl alcohol.

When you find a beverage that is free of propyl alcohol, it may have other pollutants. Xylene and toluene invite parasites to the brain. Wood alcohol invites them to eyes and pancreas. The diseases caused by these are not as frightening, perhaps, as cancer, but entirely avoidable. For this reason, I suggest in this book that you go back (actually "forward") to self made products, unprocessed food and a limited number of tested supplements.

Getting Well After Cancer

The ravages of cancer must be healed once the malignancy has been stopped. This is where *carcinogens* play a role. The lung lesions will not heal unless cigarette smoking, freon, asbestos, and fiberglass exposure is stopped. Carcinogens were thought to be the cause of cancer. Actually, they drew the cancer to the organ. Nickel draws cancer to the prostate. Barium found in lipstick draws cancer to the breast. And so on.

The following toxins can be present in any organ! I consider these to be our most serious threat. Starting with the worst, they are:

1. **Freon** (same as **CFCs** or refrigerant). I have not found a single person to be entirely free of it, including persons without cancer or disease. But in cancer sufferers it is always concentrated in the cancerous organ, and facilitates the accumulation of other toxins. This makes your refrig-

erator, if it leaks even microscopically, the top cancer hazard in the country!

2. **Copper** from water pipes. All cancer sufferers have an accumulation in the cancerous tissue. This makes plumbing the second greatest hazard.
3. **Fiberglass** or **asbestos** is present in about 25% of the cancer victims I see.
4. **Mercury** as in tooth fillings.
5. **Lead** from solder joints in copper plumbing. All colon cancer sufferers test positive.
6. **Formaldehyde** as in foam bedding and new clothing.
7. **Nickel** usually from dental metal. Most prostate cancer sufferers test positive.

At first, tumors are benign—what a relief to find your tumor was benign. Its true nature is still unknown, but it contains freon, other toxins, and later propyl alcohol! Since tumors are often large, many centimeters in diameter, and toxins do not occupy much space, there is much unidentified substance. These tumors can multiply and enlarge, as in fibrocystic breast disease, all without being malignant. But what a convenient place for baby stages of the fluke to hide out and multiply, safe from your immune system. After being colonized by fluke stages, the tumors become malignant. Eventually they also become infected when metastasis sets in.

All malignancies have the same two fundamental causes: intestinal flukes and propyl alcohol. Whether you have a rhabdomyosarcoma or a mesothelioma or melanoma, you can cure it quickly, never to return.

Herbal Parasite Killing Program

Flatworms, roundworms, protozoa, even bacteria and viruses are remarkably easy to kill using a combination of zapping and this herbal program. Thus it is not just for cancer, but a general treatment that can benefit almost every illness

1. **Black Walnut Hull Tincture Extra Strength** (see *Recipes,* page 543):
   Day 1: (this is the day you begin; start the same day you receive it)
   Take one drop. Put it in ½ cup of water. Sip it on an empty stomach such as before a meal.

   Day 2: Take 2 drops in ½ cup water same as above.
   Day 3: Take 3 drops in ½ cup water same as above.
   Day 4: Take 4 drops in ½ cup water same as above.
   Day 5: Take 5 drops in ½ cup water same as above.
   Day 6: Take 2 tsp., all together in ¼ cup water. Sip it, don't gulp it. Get it down within 15 minutes. (If you are over 150 pounds, take 2½ tsp. Do not take more than 3 tsp. because no additional value has been observed.)

   This dose kills any remaining stages throughout the body, including the bowel contents, a location unreachable by electric current. The alcohol in the tincture can make you slightly woozy for several minutes. Simply stay seated until you are comfortable again. You may put the tincture in lukewarm water to help evaporate some of the alcohol, but do not use hot water because that may damage its parasiticide power. Then take niacinamide 500 mg to counteract the toxicity of the alcohol. You could also feel a slight nausea for a few minutes. Walk in the fresh air or *simply rest until it passes. You may add more water or honey or a spice to make it more palatable.*

For a year: take 2 tsp. Black Walnut Hull Tincture Extra Strength every week or until your illness is but a hazy memory. This is to kill any parasite stages you pick up from your family, friends, or pets.

Family members and friends should take 2 tsp. every other week to avoid reinfecting you. They may be harboring a few parasite stages in their intestinal tract without having symptoms. But when these stages are transmitted to a cancer patient, they immediately seek out the cancerous organ again.

You may be wondering why you should wait for five days before taking the 2 tsp. dose. It is for your convenience only. You may have a sensitive stomach or be worried about toxicity or side effects. By the sixth day you will have convinced yourself there is no toxicity or side effects.

Going faster*. In fact, if you are convinced after the first drop of the restorative powers of Black Walnut Hull Tincture Extra Strength, take the 2 tsp. dose on the very first day.*

Going slower*. On the other hand, if you cringe at the thought of taking an herb or you are anxious about it's safety, continue the drops, increasing at your own pace, until you are ready to brave the decisive 2 tsp. dose.*

2. **Wormwood** capsules (should contain 200-300 mg of wormwood, see *Sources*):

Day 1: Take 1 capsule before supper (with water).
Day 2: Take 1 capsule before supper.

Day 3: Take 2 capsules before supper.
Day 4: Take 2 capsules before supper.

Continue increasing in this way to day 14, whereupon you are up to seven capsules. You take the capsules all in a single dose (you may take a few at a time until they are all gone). Then you do 2 more days of 7 capsules each. After this, you take 7 capsules once a week forever, as it states in the Maintenance Parasite Program. Try not to get interrupted before the 6th day, so you know the adult intestinal flukes are dead. After this, you may proceed more slowly if you wish. Many persons with sensitive stomachs prefer to stay longer on each dose instead of increasing according to this schedule. You may choose the pace after the sixth day.

3. **Cloves**:
   Fill size 00 capsules with fresh ground cloves; if this size is not available, use size 0 or 000. In a pinch, buy gelatin capsules and empty them or empty other vitamin capsules. You may be able to purchase fresh ground cloves that are already encapsulated; they should be about 500 mg. Grocery store ground cloves do not work! Either grind them yourself or see Sources.

   Day 1: Take one capsule 3 times a day before meals.
   Day 2: Take two capsules 3 times a day.
   Days 3, 4, 5, 6, 7, 8, 9, 10: Take three capsules 3 times a day.
   After day 10: Take 3 capsules all together once a week forever, as in the Maintenance Parasite Program.

Take **ornithine** at bedtime for insomnia. Even if you do not suffer from insomnia now, you may when you kill parasites.

Parasite Program Handy Chart

Strike out the doses as you take them.

	Black Walnut Hull Tincture Extra Strength Dose	**Wormwood Capsule Dose (200-300 mg)**	**Clove Capsule Dose (Size 0 or 00)**
Day	drops 1 time per day, like before a meal	capsules 1 time per day, on empty stomach (before meal)	capsules 3 times per day, like at mealtime
1	1	1	1, 1, 1
2	2	1	2, 2, 2
3	3	2	3, 3, 3
4	4	2	3, 3, 3
5	5	3	3, 3, 3
6	2 tsp.	3	3, 3, 3
7	Now once a week	4	3, 3, 3
8		4	3, 3, 3
9		5	3, 3, 3
10		5	3, 3, 3.
11		6	3
12		6	Now once a week
13	2 tsp.	7	
14		7	
15		7	
16		7	
17		Now once a week	
18			3

At this point you do not need to keep a strict schedule, but instead may choose any day of the week to take all the parasite program ingredients.

Continue on the Maintenance Parasite Program, indefinitely, to prevent future reinfection.

Maintenance Parasite Program

YOU ARE ALWAYS PICKING UP PARASITES! PARASITES ARE EVERYWHERE AROUND YOU! YOU GET THEM FROM OTHER PEOPLE, YOUR FAMILY, YOURSELF, YOUR HOME, YOUR PETS, UNDERCOOKED MEAT, AND UNDERCOOKED DAIRY PRODUCTS.

I believe the main source of the intestinal fluke is undercooked meat. After we are infected with it this way, we can give it to each other through blood, saliva, semen, and breast milk, which means kissing on the mouth, sex, nursing, and childbearing.

Family members nearly always have the same parasites. If one person develops cancer or HIV, the others probably have the intestinal fluke also. These diseases are caused by the same parasite. They should give themselves the same de-parasitizing program.

Do this once a week. You may take these at different times in the day or together:

1. **Black Walnut Hull Tincture Extra Strength**: 2 tsp. on an empty stomach, like before a meal.
2. **Wormwood capsules**: 7 capsules (with 200-300 mg wormwood each) at once on an empty stomach.
3. **Cloves**: 3 capsules (about 500 mg. each, or fill size 00 capsules yourself) at once on an empty stomach.
4. Take **ornithine** as needed.

	Black Walnut Hull Tincture Extra Strength Dose	**Wormwood Capsule Dose (200-300 mg)**	**Clove Capsule Dose (Size 0 or 00)**
Day	1 time per day, on empty stomach	capsules 1 time per day, on empty stomach	capsules 1 time per day, on empty stomach
1	2 tsp.	7	3
2			
3			
4			
5			
6			
7			
8	2 tsp.	7	3
9			
10			

11			
12			
13			
14			
15	2 tsp.	7	3
and so	**on...**		

The only after-effects you may feel are due to bacteria and viruses escaping from dead parasites! Be sure to zap after taking your maintenance parasite treatment. After-effects also let you know that you did indeed kill something. Try to discover how you might have picked up parasites and avoid them next time.

Pet Parasite Program

Pets have many of the same parasites that we get, including *Ascaris* (common roundworm), hookworm, *Trichinella*, *Strongyloides*, heartworm and a variety of tapeworms. Every pet living in your home should be deparasitized (cleared of parasites) and maintained on a parasite program. Monthly trips to your vet are not sufficient.

> You may not need to get rid of your pet to keep yourself free of parasites. But if you are quite ill it is best to board it with a friend until you are better.

Your pet is part of your family and should be kept as sweet and clean and healthy as yourself. This is not difficult to achieve. Here is the recipe:

1. **Parsley water**: cook a big bunch of fresh parsley in a quart of water for 3 minutes. Throw away the parsley. After cooling, you may freeze most of it in several 1 cup containers. This is a month's supply. Put 1 tsp. parsley

water on the pet's food. You don't have to watch it go down. Whatever amount is eaten is satisfactory.

All dosages are based on a 10 pound (5 kilo) cat or dog. Double them for a 20 pound pet, and so forth.

Pets are so full of parasites, you must be quite careful not to deparasitize too quickly. The purpose of the parsley water is to keep the kidneys flowing well so dead parasite refuse is eliminated promptly. They get quite fond of their parsley water. Perhaps they can sense the benefit it brings them. Do this for a week before starting the Black Walnut Hull Tincture.

2. **Black Walnut Hull Tincture (regular strength):** 1 drop on the food. Don't force them to eat it. Count carefully. Treat cats only twice a week. Treat dogs daily, for instance a 30 pound dog would get 3 drops per day (but work up to it, increasing one drop per day). Do not use Extra Strength.

If your pet vomits or has diarrhea, you may expect to see worms. This is extremely infectious and hazardous. Never let a child clean up a pet mess. Begin by pouring salt and iodine[15] on the mess and letting it stand for 5 minutes before cleaning it up. Clean up outdoor messes the same way. Finally, clean your hands with diluted grain alcohol (dilute 1 part alcohol with 4 parts water) or vodka. Be careful to keep all alcohol out of sight of children; don't rely on discipline for this. Be careful not to buy isopropyl rubbing alcohol for this purpose.

Start the wormwood a week later.

---

[15] “Povidone” iodine. topical antiseptic, is available in most drug stores.

3. **Wormwood capsules:** (200-300 mg wormwood per capsule) open a capsule and put the smallest pinch possible on their dry food. Do this for a week before starting the cloves.

4. **Cloves:** put the smallest pinch possible on their dry food.

Keep all of this up as a routine so that you need not fear your pets. Also, notice how peppy and happy they become.

Go slowly so the pet can learn to eat all of it. To repeat:

- Week 1: parsley water.
- Week 2: parsley water and black walnut.
- Week 3: parsley water, black walnut, and wormwood.
- Week 4: parsley water, black walnut, wormwood, and cloves.

	Parsley Water	**Black Walnut Hull Tincture Dose**	**Wormwood Capsule Dose**	**Clove Capsule Dose (Size 0 or 00)**
Week	teaspoons on food	drops on food, cats twice per week, dogs daily	open capsule, put smallest pinch on food	open capsule, put smallest pinch on food
1	1 or more, based on size			
2	1 or more	1		
3	1 or more	1 or more, based on size	1	
4	1 or more	1 or more	1	1
5 and onward	1 or more	1 or more	1	1

Parasites Gone, Toxins Next

Healing is automatic when you clean up your body tissues. Killing parasites and bacterial and viral invaders is fundamental. Removing toxins which invite them into your organs is even more fundamental.

How do you know which toxins are responsible for your cancer? Unless you use a Syncrometer to test, you can not know. So learn to use a Syncrometer, it will save your life.

The only other alternative is to move to a safer environment. Go on vacation. Stay in hotels less than ten years old (so the plumbing hasn't started to corrode) but not in new construction (give the carpets time to outgas). Do not have a freon containing refrigerator in your room. Do not use the hair drier they supply. Do not let them spray your room with anything. Launder the sheets and towels yourself at a Laundromat with borax and/or washing soda. Don't hesitate to drink the municipal water. Get busy with your dental cleanup.

When you get better on vacation, let that be your inspiration to move from your home. All of the toxins come from a civilized lifestyle. Resolve to leave it behind. Select a warm climate where you can spend your time outdoors in the shade most of the day. Have no refrigerator, air conditioner, clothes dryer, hair dryer, new clothing, detergent. Check that the plumbing is relatively new and that no pesticide is being used. Make the Easy Lifestyle Improvements (page 397). Throw away non-essential health supplements (unless tested) and drugs. The risk is greater than the benefit.

Watch For Bacteria

In the later stages of cancer the tumors are more and more infected with the common bacteria *Salmonella*, *Shigella*, and *Staphylococcus aureus*. Killing parasites prepares a feast for these ubiquitous bacteria. Now, more than ever, must you stay off dairy products (except for boiled milk), do the Bowel Program, take Lugol’s. Remove *Staphs* by doing the dental cleanup (page 409). Don't delay.

Help Your Family, Too

If you had cancer, your whole family should be freed of intestinal fluke parasites to protect you! They may not be getting cancer (yet!) but your closeness puts you at risk. Kissing on the mouth could reinfect you. Request that family members zap themselves and take at least one 2 tsp. dose of Black Walnut Hull Tincture Extra Strength while living with you.

Pets, too, can be a source of cancer fluke stages (in their drool). They get propyl alcohol from their feed. Use unflavored, uncolored feed for them to minimize pollution from processing. Add 1 tbs. vitamin C powder to the top of a 5 lb. bag; it will stick to the pellets as they pour out. Zap your pet along with yourself by holding them and touching a bare spot such as nose or paw.

Never eat rare meats or fast food chicken. Parasite stages survive heat up to the boiling point. If you have taken a risk, zap yourself as soon as you get home to minimize the damage.

Cancer could be completely eliminated in the entire country if laws required testing for solvents in animal feeds and human food and products. Presently it is allowed in the United States Code of Federal Regulations (CFR) (see page 428)!

Another reason for propyl alcohol pollution (and other pollutants) in our food are the chemicals used by manufacturers to sterilize their food handling equipment.

> **21 CFR 178.1010 (4-1-94 Edition) Sanitizing solutions.**
>
> Sanitizing solutions may be safely used on food-processing equipment and utensils, and on other food-contact articles as specified in this section, within the following prescribed conditions:
>
> (a) Such sanitizing solutions are used, followed by adequate draining, before contact with food. **[Note rinsing or drying is not required!]**

> (b) The solutions consist of one of the following, to which may be added components general recognized as safe and components which are permitted by prior sanction or approval.
>
> [Now comes (1) through (43) permissible sterilizing solutions, including several with isopropyl alcohol, like:]
>
> (25) An aqueous solution containing elemental iodine (CAS Reg. No. 7553-56-2), potassium iodide (CAS Reg. No. 7681-11-0), and isopropanol (CAS Reg. No. 67-63-0). In addition to use on food processing equipment and utensils, this solution may be used on beverage containers, including milk containers and equipment and on food-contact surfaces in public eating places.
>
> [Then in paragraph (c)(19) the exact concentration of the iodine is specified. Nowhere is the concentration of the isopropanol specified. It can be as strong as desired.]

Fig. 41 U.S. regulations on sterilizing solutions.

Even if there were regulations governing removal of sanitizing solutions, the overwhelming truth is missed: that **nothing can ever be completely removed after it has been added**. Or perhaps the lawmakers didn't miss this fact. Perhaps they believed that small amounts–too small to measure with an ultraviolet spectrophotometer–could surely do no harm.

Meanwhile, protect yourself by avoiding propyl alcohol. And by observing 2 extra rules:

1. Never eat moldy foods. **2. Always take vitamin C.**

Aflatoxin

A common mold found on bread, nuts and fruit and in beer, apple cider vinegar and syrups, produces *aflatoxin*. This is what prevents you from detoxifying tiny bits of propyl alcohol that get into your body!

Buy bakery bread or bake your own. Roast nuts after washing in vitamin C water to destroy aflatoxin and then store in refrigerator. Keep moldy fruit out of the refrigerator where the spores can spread. Use only white distilled vinegar. Use honeys instead of syrups and even add vitamin C to them.

Vitamin C helps your body detoxify all the mold toxins I have tested, including aflatoxin.

Keep powdered vitamin C in a salt shaker. It belongs on the table with salt and pepper, and at the stove. Put it in everything possible, from cereal to soup to rice (1/8 tsp. is enough). Besides this take 1/8 tsp. powdered vitamin C with each meal (500 mg).

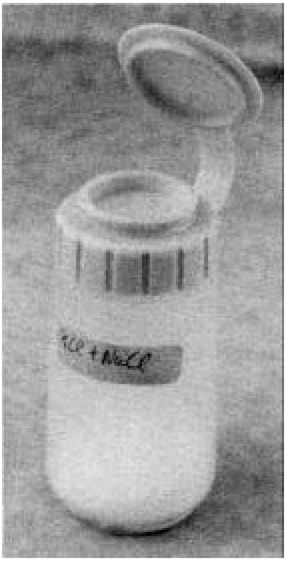

Fig. 42 Keep both salt and vitamin C powders in closable plastic shakers.

Developing cancer is a chain of events. This explains why it is a disease of aging. For ten years or more you poisoned your body with freon, fiberglass, asbestos, mercury, lead, copper, etc. You continually ate moldy food (chips, nuts, etc.) which was toxic to the liver. But the liver regenerated the pieces that died. Eventually a mold toxin prevents the liver from regenerating. Your aflatoxin-ed liver then lets propyl alcohol build up after using it. You use more and more propyl alcohol. Somewhere, over the years, you pick up the intestinal fluke in a hamburger or from a pet or person. The stage is set and cancer is inevitable.

But pull out any one of the links in this chain and cancer is impossible. **Pull them all out for a return to good health**.

So cure yourself, prevent reinfection, heal the damage and go through life without this sword hanging over you.

Over 100 consecutive case histories of cured cancer victims are the subject of another book[16] along with more detailed instructions and suggestions.

---

[16] *The Cure For All Cancers*, by the same author, New Century Press 1993.

Curing HIV/AIDS

HIV and AIDS disease are very similar to cancer. That is why cancer is so often seen with HIV/AIDS. It is caused by the same parasite but the solvent is **benzene** instead of propyl alcohol.

HIV stands for
Human Immunodeficiency Virus

AIDS stands for
Acquired Immune Deficiency Syndrome

Many researchers believe HIV causes AIDS. I observe them both caused by parasites and solvents!

When the body can no longer detoxify benzene it soon may not be able to detoxify propyl alcohol. Food mold, at the base of the propyl alcohol problem, is also at the base of the benzene problem. Zearalenone, a mycotoxin I find in popcorn, corn chips, and brown rice specifically inhibits detoxification of benzene. None of us should be eating these foods untested.

Several common mold toxins inhibit the immune system, too, specifically those white blood cells that are supposed to eat and destroy viruses. Consequently, the HIV virus cannot be checked once it is introduced in the body.

Benzene goes to the bone marrow where T-cells are made, and to the thymus where T-cells are programmed, two big blows to the immune system.

Benzene, a most unthinkable pollutant, is widespread in extremely small amounts. But when **benzopyrenes** are added, the total is entirely too much for the liver and it begins to accumulate in your thymus.

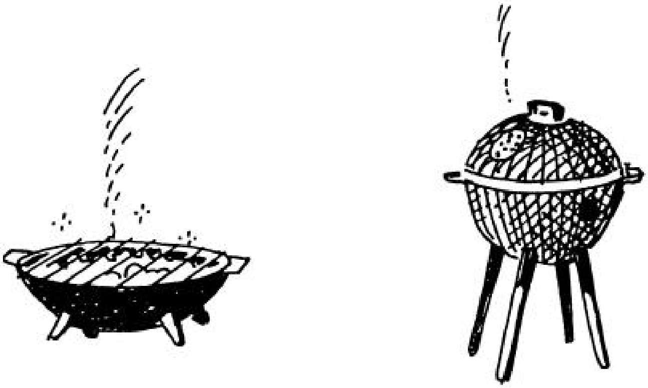

Fig. 43 Grilled food has benzopyrenes as do hot dogs and "smoke flavored" foods.

Benzopyrenes are made right in your food by direct flame-heat. Grilled food, smoked food, hot dogs and lunch meats with "smoke flavor" all have benzopyrenes—even toast has it. Food fried in a skillet does not (even if you burn it black!). As the thymus deteriorates from its benzene burden so does your immunity.

Without benzene in your body
you can't get AIDS.
Without the intestinal fluke in your body
you can't get HIV.

After killing the parasite and its stages the HIV virus will be gone in 2 hours, almost as quickly as the cancer marker, ortho-phospho-tyrosine. It is not necessary to zap the virus itself at 365 KHz.

But the AIDS is not gone. This depends on removing benzene from your body and keeping it out of your body. Keeping it out on alternate days or 5 days of the week will not suffice. It must be out consistently. Only then can your thymus and bone marrow recover.

As your thymus recovers your immunity returns. Immunity to all the diseases that lurk at very low level in our bodies. They are already in us. We have accumulated them in a lifetime. Mumps, measles, chicken pox, CMV, *Staphylococcus aureus*, *E. coli*, are all there. Until now they have been controlled–kept down–by an ever vigilant immune system. If this system fails you are an easy target for any and all of these invaders.

Method Of Treatment

As soon as you find you are HIV positive, don't panic. It was inevitable. It is also inevitable for many others with unexplainable risk factors. If you have a source of infection with the intestinal fluke and at the same time a source of benzene, HIV and AIDS are a logical result. You must stop the source of both.

Purchase or make a zapper or use somebody's frequency generator (434 KHz to 421 KHz giving 3 minutes to each KHz). Kill your flukes immediately. Take vitamin B2 immediately to help detoxify benzene (three 100 mg. tablets 3 times a day, 9 altogether). Zap daily until you feel completely well: no night sweats, no coughing, no symptoms of any kind.

Remember parasites are all around us. We pick them up daily. And your weakened immune system lets everything gain a foothold in your body. Avoiding benzene will restore your immunity. Use the benzene pollution list to guide you.

Benzene Polluted Products

THROW THESE OUT

Your health is worth more than the fortune you spent on them!

- **Flavored food**, yogurt, jello, candies, throat lozenges, store-bought cookies, cakes
- **Cooking oil** and shortening (use only olive oil, butter and lard)
- **Bottled water**, whether distilled, spring, mineral, or name brand. Bottled fruit juice.
- **Cold cereal**, including granola and health brands
- **Toothpaste**, including health brands
- **Chewing gum**
- **Ice cream** and frozen yogurt
- **Pills and capsules**. At least a third of all I test are polluted. This includes herbal extracts and prescription drugs. Test yours and switch brands until you find a safe one.
- **Vaseline products** (Noxzema™, Vick's™, Lip Therapy™), chap stick, hand cleaners
- **Vitamins** and other **health supplements**, unless tested.
- **Rice cakes**, even the plain ones
- **Personal lubricant**, including lubricated condoms
- **Baking soda and cornstarch** (see *Sources*)
- **Soaps, hand creams**, skin creams, moisturizers
- Flavored pet food, both for cats and dogs
- Bird food made into cakes

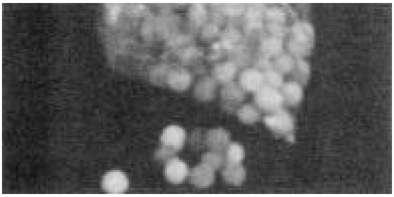

Fig. 44 Mexican made candy with no benzene.

Fig. 45 Pollutants are in unlikely places

It is impossible for me to have tested every batch of every food and product, but so many test positive you simply can not risk any of the foods and products on the list.

Learn to use the Syncrometer to do your own testing. While you are learning, observe the rules perfectly. There is no half way measure with benzene. Take an attitude of over-compliance, not "getting away" with as much as possible. Over-compliance can bring back radiant health, strength you never knew you had, a beautiful normal body again! You will be able to resume your plans for education, professional life, personal relationships, free of the sword hanging over you. You will not be infectious to anybody. Let over-compliance enthuse you.

Fig. 46 Every brand of popcorn and corn chips I tested had zearalenone contamination.

There are 2 extra rules that help eliminate benzene buildup:

1. Never eat moldy or grilled foods including popcorn, corn chips, grocery store breads, pasta, nuts, alcoholic beverages, toast, brown rice (white is OK).
2. Always take Vitamin B_2 (three 100 mg tablets three times a day) and vitamin C (1/8 tsp. with each meal).

Plan For The Future

After you are well again, you may wish to indulge in some philosophy. How did an obscure virus–a snail virus!–become a human virus? This is not so bizarre. After all, rabies virus comes to us from animals, and many encephalitis viruses come from mosquitoes. How did HIV spread so rapidly? How did a pollutant as dangerous as benzene get to be in our very food? Are other parasites getting set to spring on us? What must be done to protect ourselves and loved ones from future disasters? Would getting away from fossil fuel be a big answer? Would getting away from preoccupation with chemistry be a big answer? Would more disclosure of industrial practices be an answer? Should the government agencies responsible for food and product safety be depoliticized? Should public inspection of food manufacturing be a right? Should disclosure of foreign origins of food ingredients be mandatory? Would communicating with other cancer and AIDS survivors be useful? The computer age would make communication possible. Communication leads to answers. Answers might lead to new policies. Policies supported at the grass roots level can bring about change.

Over 50 case histories of cured HIV victims are the subject of another book[17] along with more detailed instructions and suggestions.

---

[17] *The Cure for HIV and AIDS*, by the same author, New Century Press, 1993.

Curing The Common Cold

Sometimes you can zap *Adenovirus*, the common cold virus, at 393 KHz for three minutes with a frequency generator and be rid of your cold, magically, in five minutes. But magic and luck are not really responsible for this. And most of the time the cold will reappear a few hours later.

Homeopathy, too, can immediately banish your cold symptoms. The reason is not mysterious either. But, this time, the cold reappears later in a different location. Choose the right homeopathic remedy for that, and you can chase it away again.

Tapeworm Stage or Mites

The fascinating story of how we really "catch" a cold kept me spellbound for a year. I was hot on the heels of *Adenovirus* wherever it might be in my body. Sometimes I saw it; sometimes I didn't. Sometimes I had cold symptoms; sometimes I didn't. Sometimes I could zap it; sometimes it did no good.

It is now apparent to me that *Adenovirus* isn't our (a human) virus at all! It belongs to other parasites. Parasites as varied as tapeworm stages and mites. Perhaps it belongs to many other parasites, as well. My evidence comes from a tapeworm stage, cysticercus of *Diphyllobothrium erinacea*, the mites *Sarcoptes* and *Dermatophagoides*, and our own colon bacteria, *E. coli*.

The tapeworm stage flies in the dust as eggs, you can trap these by setting out a pint jar with a little water in it. In three days' time you are likely to find its frequency near 487 KHz in your jar. You are also likely to find it on your kitchen sponge, since you wipe up dust each day. To test it, place it in a plastic bag, wet it thoroughly and search from 510 KHz downward, one KHz at a time. The various tapeworm stages emit between 510 and 440 KHz.

If you have a household pet, you will always be able to find a tapeworm stage in your sponge or in a dust sample you collect from the table or kitchen counter in the morning. Gather dust with a damp bit of paper towel, put it in a plastic bag. Then wash your hands or you may accidentally eat some.

This, of course, happens to every household member. Eating the dust off the tables, inhaling the dust, and eating off surfaces wiped by the kitchen sponge happens to everyone. And everyone “catches” colds. If you search for *Adenovirus*, though, in your dust sample, it isn't there!

Similarly, you can search for the mites in your house dust. Search near the frequencies given for them. There is a good chance you will have one that is not given, because the list is so incomplete. Name it after yourself. Compare notes with others; maybe it is common, maybe it's a rare one. Again, you will not find *Adenovirus* beeping its characteristic frequency out of your mite specimen. Why not? Possibly, it is too faint; it must multiply and create a loud chorus before you can hear it. But multiply it will, if given a chance, in you. You must, of course, first eat or inhale the dust.

Then the tape eggs hatch into the *cysticercus* stage, which promptly gets to the liver. Sometimes it gets to other organs, like the muscles, the spleen, the pancreas. Presumably the liver screened it out of the blood originally.

Soon you will zap them, wherever they are. If you are using a slide specimen of *cysticercus* you can locate it in your body. If you are only listening to its beeps, you can't. If you can do both, you may be able to see which organ allows the virus to replicate after it emerges. Maybe only the respiratory organs do; maybe they start to replicate in the organ where they emerge, such as the liver and then get to the respiratory tract. This is a fascinating avenue I have not yet explored.

Mites are inhaled or swallowed or both, just as tapeworm eggs are. They are on your kitchen sponge, and in any food or dishes that stand uncovered anywhere in the home. Never drink water from a glass that has "stood out" all day. After finding one, you will notice it beeping in you for several days. Then the beep disappears; presumably the mite is dead.

The tapeworm stage beeps may disappear in a few days, too, presumably dead. Except in cases of disease. Muscles that are diseased will take in the newcomer and allow it to survive adding to the parasites and pollutants already there! Evidently the immune power of such diseased locations is way down.

On the day that the mite stops beeping, the day it dies, *Adenovirus* appears!

Not many *Adenovirus* at first. You will need to search several times during the day to find it in your white blood cells. (And you have no symptoms, yet, either!) Is it a coincidence that *Adenovirus* appeared directly after a tapeworm stage or mite died?

You can find out by waiting until a time when you have a tapeworm stage or mite and no *Adenovirus*. Then kill your tapeworm stage or mite by zapping for seven minutes. Within minutes after that, the beeping of the tapeworm stage or mite is gone and *Adenovirus* can be heard, loud and clear at 393 KHz! And minutes later you may feel a stuffy nose, a slight congestion developing, a certain head feeling that is different. You are "catching" a cold!

Will you really get this cold? Will it become a full blown cold of the usual kind? After seeing this happen dozens of times after killing a mite or tapeworm stage I concluded that *Adenovirus* really belongs to them; the virus is scuttling its dying host like people jumping off a sinking ship into the ocean. Our bodies are the ocean for them. They too, immediately swim and search for a hospitable island. Our respiratory tract is such an island; perhaps other organs, too.

Yes, this “baby cold” will develop into a full blown cold if, but only if, you have a mold in you!

Molds and Colds

This is part two of the cold story. You may have *Adenoviruses* quietly slipping into your blood stream and tissues from a tapeworm stage or mite you inhaled, or *E. coli* bacteria that strayed into your tissues, and which is being slowly killed by your immune system. Your immune system can keep up with them quite easily provided you don't have a mold in you at the same time. The significance of the mold is that it lowers your immunity, specifically and generally.

Mold eluded, is health improved.

This has already been studied extensively for a number of food molds. There are a variety of ways that mold toxins lower immunity. Some simply kill white blood cells. Others seem to “bind and gag” them so they just can't go about eating viruses.

So with mold toxins present, *Adenovirus*, fleeing the dead tapeworm stage, mite, or *E. coli* is not gobbled up. It has time to get to its favorite organ and enter the cells there. It may get in your lungs if they're full of arsenic or formaldehyde, in your throat if it's full of mercury from your fillings, in your spinal cord if it's full of thallium. Sometimes you feel the viral attack, sometimes you don't.

When *E. coli* is the source of your *Adenovirus*, a question pops up. Why don't you have a perpetual cold, since these bacteria are always in your colon...and should be! As long as *E. coli* stays dutifully in your colon, no *Adenovirus* is seen. But as soon as any cross the colon wall to invade your body, your white blood cells pounce on them. After this, *Adenovirus* appears and again you are catching a cold. They may go to your internal organs where you don't feel them.

One place you do feel an attack is in your respiratory tract: lungs, bronchi, sinuses, nose, Eustachian tubes, inner ear, eyes or head. And the size of the attack depends on whether you recently ate moldy food.

Human food (in general, in the U.S.) is very, very moldy. We do not taste it because manufacturers have been using more and more flavorings in food. This covers up small amounts of mold or “off” flavor. Measures to reduce mold are not effective enough.

Bread is a good example. Calcium propionate is added to bread-stuff to inhibit molding. That's fine. But then the bread is encased in plastic to hold in moisture and keep it “fresh”. The moisture acts to incubate mold spores and overwhelms the inhibitor. Vinegar is used instead of calcium propionate in some breads but, again, the plastic ruins its effectiveness.

Another good mold inhibitor is lime water. This is used in making tortillas. None of the old fashioned tortillas (made with just corn, water, lime) that I tested had any mold, even without propionate added! Other tortillas made of flour and calcium propionate frequently had molds.

Fig. 47 This Mexican bread has no molds. It is made from white, unbleached flour and is not wrapped in plastic.

Bread is such a staple we must correct its mold problem immediately. The two likely sources for the mold spores are: in the flour to begin with, or just flying about the bakery and landing on the newly baked loaves. Bread flour in the grocery store is quite free of mold spores, so maybe it is the bakery that needs to change. Perhaps it is not possible to bake 24 hours a day in the same building, year after year, without bits of flour and moisture accumulating in the millions of tiny cracks and crevices that all buildings have and germinating mold. Yet bread from small neighborhood bakeries does not have mold!

Only buy fresh bread not wrapped in plastic.

As soon as you feel a cold coming, ask yourself: what did you eat recently that might have been moldy? Cold cereal, hot cereal, bread, crackers, cookies, rice, other grains, fresh fruit, store bought fruit juice, nuts, syrups, pasta, honey? This takes up a large part of any person's diet, even in a single meal. The answer is yes, no doubt you ate moldy food recently.

So you can't catch a cold directly from some sneezy companion, or *E. coli*, or tapeworm stages, or mites. You have to eat moldy food first. This lowers your immunity, allowing any *Adenovirus* to invade your weakest tissues. If that happens to be your respiratory system, you get classic symptoms. (Those same “cold” symptoms can be caused by bacteria, for which I have preliminary evidence of arriving in a similar fashion.)

As fascinating as this whole story is, the bottom line is: how can you stop a cold, in record time?

The Cure

Remember, zapping does not kill your cells. So anything hiding in them will not be harmed by the electric current either. Viruses live inside your cells while they reproduce. You can only kill the outsiders: those stuck in your cell gateways. The rest will be killed by your own cells in time. Not much time. Five or six hours at most. Your cells do it with mucous secretion, inflammation and other unknown ways. So zapping is an imperfect solution.

Using a frequency generator to electrocute *Adenovirus* (393 KHz for three minutes) is not effective either because you are not killing the mites and tapeworm stages at the same time, and the *Adenoviruses* are coming from them.

The best advice to stop a cold is to:

1. **Eat sterile food for 24 hours.** Follow the Mold Free Diet on page 365. Do not eat one questionable item.
2. Take **vitamin C** (10 grams or 2 tsp.), a **B50 complex** (2 tablets), and niacinamide (3 500 mg tablets) to help detoxify the mycotoxins already in you. See *Sources*. It will still take five or six hours for your white blood cells to recover their ability to capture viruses, for the "gag" to wear off.
3. **Zap** for seven minutes, killing all viruses, tapeworm stages and mites together. **Wait** twenty minutes to let viruses and bacteria in the dead larger parasites emerge. **Zap** a second time for seven minutes to kill those viruses and bacteria. **Wait** twenty minutes to let any viruses infecting the killed bacteria emerge. **Zap** a third time for seven minutes to kill the last viruses.
4. Now you need only wait for your tissues to decongest and stop making mucous, etc.
5. Immediately start the Bowel Program (page 546) in case yours is an *E. coli* cold. You can't, and wouldn't want to, kill all the bacteria in your bowel. Zapping kills the escapees, though, to give a bit of relief, and the Bowel Program stops the invasive *E. coli*.
6. Do additional zapping as time permits until the Bowel Program has stemmed the invasion.

In five or six hours your cold could be gone.

Of all these measures, stopping mold consumption is the most important. If you eat peanut butter now, your cold will return with a vengeance. If you eat cheese it will add *Salmonella* to your illness and you may develop a fever. Your own immune system is the most powerful cold killer, so stop handicapping it.

Test yourself for the presence of molds to see if you are accomplishing your goal. The various molds I have tested had these emitting frequencies: 77, 88, 100, 126, 133, 177, 181, 188, 232, 242, 277, 288, 295 KHz.

In about five hours, some relief will be felt. The time it takes depends on how much mold you ate. But if you stop immediately and eat only perfectly safe food, your illness will be over in the shortest time. Shorter than merely zapping.

Mold Free Diet

Stick to this for 24 hours, even if you feel better much sooner.

Before starting to cook sterilize your kitchen sponge (microwave it for three minutes), and wash hands.

Breakfast

- 1 or 2 eggs any style. The egg carton and egg exterior have *Salmonella* on them, so remove the eggs, replace the carton, wash the exterior of the eggs and then your hands again before cracking them. You don't want a *Salmonella* infection added to your cold.
- Hash brown potatoes
- Pancakes or waffles with artificial maple syrup made with brown sugar and vitamin C.
- 1 cup herb tea with vitamin C, cinnamon, or other spice.

- Water with a tsp. of vinegar and honey.

Lunch

- Soup without noodles, without rice or any grain product.
- Homemade biscuits or bakery bread with butter.
- Homemade pudding or custard, all ingredients well cooked.

Supper

- Well cooked fish or seafood (canned O.K., like sardines, salmon and tuna).
- Fresh green vegetables, in perfect condition.
- Cooked vegetables with olive oil and salt (pure, see *Sources*).
- Canned beans, any variety.
- Baked or boiled potato (don't eat skin) with olive oil and salt as described on page 310.
- Hot water with whipping cream and cinnamon.
- A perfectly unblemished banana.
- Water with lemon.

You can “mix and match” these safe foods. If you get a hefty dose of mold at the outset of your cold, the toxicity lasts quite a long time. Repeat the diet the next day and the next until you are well. In animal experiments reported by scientists, toxicity from mold usually lasted three weeks. Sometimes the real damage was only seen after three weeks! Keep up your vigilance. When you decide to take some risks, make sure vitamin C has been added to the new food and mixed with it thoroughly.

Fig. 48 Goods baked in a panaderia (Mexican style bakery) never tested positive to molds.

Prevention

What is the best way to prevent a cold? Stop eating moldy food. We were never meant to eat rotting, spoiling food. We were meant to be repelled by the bad odor, the bad appearance, the bad taste. Our parents were supposed to teach us in childhood to distinguish between good and bad food. This system no longer works. By mixing spoiled food with good food, by adding flavors and changing the food consistency, parents can't distinguish substandard foods, nor teach their children to.

We rely on government agency assurances, like beef grades, expiration dates, approved food colors and additives. And when they fail? We land in a debacle such as the present one, where large segments of society are ill with uncontrollable behavior (called crime), suffer from hormone imbalances and sexual disturbances, are sidelined by chronic fatigue and new illness. But the greatest social disaster is lowered immunity—AIDS.

We cannot individually control the destiny of our species. But we can get ourselves out of the path of the falling human tide. We can dissent. We can say **NO** to rotten moldy food.

What about exposure to a cold? Won't you catch one if you're in a roomful of coughing, sneezing people? Or you shake hands with a coughing, sneezing person? Or you drink from the same glass as a person with a cold? Or your child plays with the same toys as sick children?

Yes, you will catch the virus, but you won't get sick. If you are tracking *Adenovirus* using the electronic techniques in this book, you will see that it infects you immediately after eating coughed-on food. It stays in you about twenty minutes. Then it disappears, evidently eaten up by your white blood cells, **provided there is no mold toxin in you**. But if you do have a mold toxin in you, the virus spreads, multiplies and gives you a cold! I have only followed *Adenovirus* behavior. There are numerous other “cold bugs” left for you to track. Each illness in your family could provide you with a specimen to research.

This throws light on self infection, too. When we cough into our hands and then eat with them, we self-infect. But it only makes us sicker if we harbor molds.

Curing Symptoms

Cold symptoms alone can be cured much faster than the cold. Taking drugs proves that. Most of the drugs on the market can cure some symptoms in less than one hour. But not without a price, such as a harmful side effect.

Homeopathy can cure cold symptoms too, without side effects. There are three or four favorite homeopathic remedies for colds and eight or nine less common ones. To use them you read the symptoms listed and take the remedy with the closest match.

Homeopathic Remedy	**For These Symptoms**
Aconitum	early cold with fever, headache, hoarse cough
Allium	clear runny nose with burning of lips or eyes
Arsenicum	sneezing cold, frontal headache, tickling cough
Belladonna	high fever cold with flushed face, throbbing head
Kali bi	thick post nasal drip, colored discharge, sinus headache
Spongia	croupy cough

Fig. 49 Starter set of homeopathic remedies.

There are lots more remedies with fascinating symptoms to try to match with your own. Books suggest that you start with a 6X or a 12X remedy, but success is more certain with 30X. Use three remedies simultaneously. If you get total relief at some point, stop. If the cold has moved or won’t budge, try another set of three after an hour. Don’t take them within 15 minutes of food or anything else. Homeopathy makes very interesting reading.

Homeopaths say they stimulate the immune system specifically. My results show they do much more. They go right to the gateways of your cells and evict the tiny parasite, bacteria or virus stuck to the latch and trying to get in. Your immune system would be able to gobble them all up if they weren't gagged by the food mold you ate.

Different homeopathic remedies go to different tissues, so you can only clear one tissue at a time. If you plan on trying this for yourself, order the set of cold remedies listed above (see *Sources*). The homeopathic method would be a beautiful cold cure if it weren't for the mold intrusion.

Herbs, too, can cure symptoms rather quickly. **Elderberry tea** mixed with **peppermint** is cited in herbal textbooks and it could probably do a lot if it weren't for the mold immunosuppression. If you plan on trying these start with a set of **thyme**, **fenugreek**, **sage** (for throat). Since both herbs and homeopathic remedies work on the principle of ejection, they could eject each other. Maybe the last one to arrive takes over. This is an exciting field for you to explore.

Ultimately, the length of time your own white blood cells are bound and gagged decides how soon you are really cured of your cold. Remember taking vitamin C (10 grams or more) helps detoxify the molds. If you find a recipe that works for everybody in less than five hours, be sure to let everybody know.

True Origins Of Viruses

Your body can eliminate any virus in a short time, such as hours or days. It can even keep up with a steady stream that is coming from tapeworm stages!

Let us apply these new insights to several diseases we are familiar with. Let us speculate what could really be happening.

Epstein Barre Virus (EBV) is a mysterious disease because in spite of building up antibodies, it attacks repeatedly. I only see EBV if the person also has *Eurytrema* (pancreatic fluke) in the pancreas. When the fluke is gone, I no longer find EBV in the white blood cells. This suggests that the virus comes from the fluke. As each fluke dies and is finally removed, the body's white blood cells can catch up with the viruses and you begin to feel better again, which can be as quick as one day. But reinfection with an EBV-carrying *Eurytrema* (plus a wood alcohol containing beverage to allow the fluke to go to the pancreas) can spark the next recurrence.

Shingles is a recurrence of chicken pox. I always find *Ascaris* in persons with shingles! Unfortunately, killing the *Ascaris* does not cure shingles. *Herpes Zoster* (the shingles/chicken pox virus) is known to hide in nerve cells. Perhaps *Ascaris* facilitates it's release, or simply suppresses the immune system in a way that allows it to suddenly multiply.

Polio was once a scourge. At that time, we can theorize that a new large parasite was making its appearance. Was a new animal association taking place in the early 1900's? Life was indeed becoming more urban with horse manure on all streets. Owning a dog for a house pet was becoming an acceptable lifestyle. Could the tapeworms of these animals give us a tapeworm stage that hosts polio virus? Many polio sufferers also had migraines. These are caused by *Strongyloides*, a horse threadworm. Does *Strongyloides* host polio virus?

The **HIV** virus infects us during the time the intestinal fluke, *Fasciolopsis buskii*, is being hosted. Persons test negative to HIV shortly after the fluke and its stages are eliminated.

Coxsackie viruses give us some of our encephalitises. I never see *Coxsackie* viruses without the bacteria, *Bacteroides* *fragilis*, and I never see *Bacteroides fragilis* without *Ascaris* (a roundworm.) I conclude that one lives inside the other!

We may be deriving viruses from all the roundworms, flukes, tapeworms and bacteria that infect us! It would be a fascinating study, simply to examine each of these parasites singly, searching for their viruses with an electron microscope. They could also be searched for using immunological methods.

Fortunately, your health improvement does not need to wait on such studies. Your electronic technique can detect them in your body long before you are made ill by them. You don't need to know their hosts in order to stop hosting them yourself.

It is a time of great change for this planet as pollution spreads from pole to pole. The growth of industrial activity, mining, chemical manufacturing, the food “industry”, and personal habits like smoking have spread new chemicals to every corner of the globe. The element *polonium*, which is radioactive and in tobacco smoke, is harmful to human lungs, but may not be harmful to a small lung parasite, like *Pneumocystis carnii*.

Benzene, which is a solvent and extremely harmful to humans, may not be harmful to fluke parasites living within us. Propyl alcohol facilitates the intestinal fluke but is toxic to us.

Parasites are doing abnormal things. Is this because of pollution?

The tables are gradually being turned against us in favor of our parasites and pathogens. Such large changes are called evolutionary. Is the human species doomed, or will some of us “adapt”. Will some of us survive to pass on our “better” genes to a new population of cancer-resistant, AIDS-resistant, Alzheimer's-resistant, etc.-resistant humans? How many of us will die trying?

The common cold should not be so common!

Body Wisdom and Why Bad Food Tastes Good

Don't let yourself eat junk food just because your body craves something in it. Try to figure out what it is your body craves. Could it be fat? Could it be salt? Could it be starch?

Salt

If it's salt, you might also love popcorn and other salty food. This implicates the adrenal glands. Maybe they're letting too much salt (sodium chloride) leave the body through the kidneys. Maybe they're letting too much potassium chloride through, too. A diuretic pill could certainly have a similar effect.

Help the adrenal glands do their job of regulating sodium and potassium chloride by cleaning them up. Let salt-hunger be your signal to do a kidney cleanse (page 549). This will clean adrenals too. Even a slight drop in sodium and potassium chloride in the blood (body fluids) can make you too fatigued to tie your own shoelaces.

Remember, when your body craves potato chips, it craves something in the potato chips. If you let yourself eat highly salted food while you're giving the adrenals a clean up, at least add potassium chloride to your diet. Make yourself a mixture of equal parts of sodium chloride and potassium chloride. Part of salt hunger is actually potassium hunger. Let your body (your taste) decide on the amount of potassium chloride to add. Maybe one part potassium chloride to two or three parts sodium chloride is a better mixture for you. After mixing, store it in the original containers (re-label them) to prevent caking. If you put it in an ordinary salt shaker, it will cake soon. Use a shaker with a lid that closes.

Fats

Maybe you like French fries because of the fat. If you deprive yourself of the “good” normal greases that come from plant or animal sources which would ordinarily make up 25% of your calories, of course you'll crave grease. But what a bad trade it is. Now you are getting lab-made (hydrogenated) grease with a non-biological structure, and loaded with the carcinogen nickel.

So if you're body tells you that you need grease, go back to olive oil, butter, cheese (baked only), lard, avocados, nuts and nut butters (homemade only) and seeds. Humankind has been eating these natural fats long before cholesterol was vilified. The key to cholesterol control is not fat avoidance, but a liver cleanse!

Starch

If switching to natural greases doesn't satisfy your “fat-tooth”, maybe its the potato in the French fries that your body craves. Plain, pure starch. Do you also love bread and pasta (more pure starch though very inferior to potatoes)? Pure starch is very easy to digest and has a large adsorptive capability for toxins. In fact, if any family member should accidentally eat something poisonous, drinking cornstarch will quickly mop it up and keep it stuck so it can't enter your tissues. (This doesn't work for all poisons.) By craving pure starches, your body could be telling you about a need to improve your digestion (liver disorders) or to eat and breathe less toxic things.

Maybe a stomach-full of baby *Ascaris* is telling you to eat only food that doesn't need a lot of acid: "just potatoes, bread and pasta, please, and skip the sauce.” *Ascaris* inhibits acid production by the stomach. This can result in an aversion to meat.

It doesn't take much acid to digest pure starch and get it on its way out of the stomach. And out of the stomach means relief: relief of the pressure on the diaphragm and liver, heartburn, that too-full feeling, and other digestive disturbances.

Sugar

Your body runs on sugar. If you are short on sugar it will turn fat into sugar. If you are short on both, it will turn your muscles into sugar. However eating more sugar doesn't cure the craving. You have to find out why you are so short, in spite of eating it.

The first thing to try is 1 mg chromium (five 200 mcg tablets, see *Sources*) per day. If you still crave sugar after a week the problem is something else. Perhaps you have pancreatic flukes upsetting your sugar regulation. Kill them and go off commercial beverages that may contain wood alcohol. Sugar regulation is very complex, but these two approaches help most of the time.

Dislikes

Respect your body's opinion when it says, “No, I don't want to eat that.” Our education about nutritive value of food may be sound but there are other facts to consider. We should take a lesson from nursing babies: when they refuse to nurse, there is something unpalatable in the mother's milk. Usually the mother has eaten onions or members of the cabbage family. The baby tries it once, and learns to reject it immediately. The baby's liver, in its wisdom, does not want the baby to eat what it can't properly digest. The mother may feel: “Now, this breast milk is good for you and drink it you must, or you shall go hungry.” Unfortunately, this works for 2-year-olds and up. They are forced to eat carrots, peas, and other vegetables; vegetables that taste terrible, (modern agriculture has ruined the flavor). They alone taste the bitterness of PIT, a cyanide-related chemical, and very difficult for the liver to metabolize. Broccoli and onions may burn the tongue with its sulfur-containing acids. Green beans, onions, garlic, eggplant, all have unique chemicals in them. If you or your child are not ready to eat them, avoid them carefully, so you don't get a surprise dose of the toxic chemical.

The more mold a child eats, inadvertently, in peanut butter, bread, potato chips, syrups, the less capable the liver is of detoxifying foods. This will certainly increase the “pickiness” of a child's appetite. If your child has too many foods on her or his personal “off list”, let this signal you to improve liver function. Stop the barrage of chemicals that comes with cold cereals, canned soup, grocery bread, instant cheese dishes, artificially flavored gelatin, canned whipped cream, fancy yogurts and cookies or chips. Move to a simpler diet, cooked cereal with honey, cinnamon and whipping cream (only 4 ingredients), milk (boiled), bakery bread, canned tuna or salmon, plain cooked or fried potatoes with butter, and slices of raw vegetables and fruit without any sauces, except honey or homemade tomato sauce, to dip into.

It is frustrating to cook “a fine meal” for the family and find everybody likes it except Ms. Picky. The good news is that she can usually think of something she would rather eat. If it's nutritious, be thankful. If it's not say No.

Adults should hide their junk food, including everything off limits to children. Don't “hide” your junk food in the refrigerator and lower level cupboards! Treat yourself as well as your child. If a food tastes bad, don't eat it. If you crave it, try to understand the message.

Outwit The Cravings

Here are some examples. Suppose you crave these items:

- **Pickles**. They supply vinegar and are often loved by persons with little acid in their stomachs or a lot of yeast (vinegar is a yeast inhibitor). Start drinking water with lemon juice or vinegar and honey.
- **Bacon**. The fat soothes the stomach and slows down digestion. Switch to butter and cream, with meals.
- **Sugar coated cereals**. Loved by persons with disturbed sugar regulation. Kill parasites, avoid wood alcohol, use chromium tablets and a lot of cinnamon.
- **Crunchy munchies**. Your jaw and teeth want some work to do. Try salads, an apple, raw sunflower seeds (beware of moldy seeds, nuts and dried fruit).
- **Ice cream**. Ice cold food stimulates the thyroid; loved by low thyroid persons. Clean up the thyroid by doing dental work and liver cleanses.
- **Caffeine**-laced beverages. Stimulate many body tissues, raise blood pressure. Loved by low energy people. Do a general body and environment cleanup. (There are people who say coffee puts them to sleep. Insomnia has better solutions than caffeine, though.)
- **Candy**. The more you eat the more you crave because chromium is being used up as you eat it and yet it is necessary to utilize more sugar. Give yourself chromium (GTF) tablets totaling 1 mg. (1,000 mcg.) a day and watch your sugar craving shrink.
- **Pretzels**. You want salt plus crunch.

- **Potato chips**. You want salt, grease, starch and crunch. No wonder they are so popular!

If your body still has its wisdom, or most of it, why can't it detect the mold in peanuts, crackers and bread for us? There is a very sound reason. Our food manufacturers have gone to great lengths to fool our native senses. Salt and sugar, roasting and flavoring, do most of it.

Your body is accustomed, natively, to interpret sugar, salt, and flavors as “good, good, good.” Of course, the mold is “bad, bad, bad.” But when you mix them, what is your body to read? The “goods” always win; manufacturers don't stop until they do. More flavorings are added. The result is that you can be eating rotten moldy food without knowing it.

Food that is predominantly concocted can't be interpreted by your body wisdom. You must use your second-best ally, your intelligence.

How would you interpret these situations, taken from real life:

- An elderly person can't stand butter, wants and enjoys margarine.
- A child prefers canned spaghetti for the real thing.
- A child wants to eat only sweets, everything else must be coaxed down.
- A young man needs “his” beer to enjoy a cook-out.
- A young man with serious mental illness drinks half a gallon of Mellow Yellow^{TM} a day.
- A child wants ketchup on everything.
- A pregnant woman puts herself on a pickled pigs feet and white bread diet.

Back To Normal

Body wisdom was not meant to substitute for parental teaching. They must both be there. But when there is conflict, trust your wisdom. Will you ever get your primitive body wisdom back and enjoy vegetables, fruit, simple styles of cooking and baking them? Yes, to a considerable extent. You'll hold your nose at all seeds and nuts and most flour (they smell so rancid). You'll back away from cookies and cakes, especially icing (they taste like you imagine shortening tastes). You'll retreat from deep fat fried foods: the dripping grease will just undo the appetite. You'll say “No thanks” to canned food and metal cutlery (you can taste the metal now). You'd have to force yourself to eat grocery bread (it's so doughy and sour smelling). Wheat germ smells terrible (rancidity). Even vitamin tablets may smell awful.

Has life been ruined, now that plain potatoes and butter taste good? Your body wisdom has returned. You are the true gourmet. It is as different as corn on the cob is from the canned variety. Perhaps you are “spoiled” now by eating un-rotten food. Accept the charge with humor and dignity—as long as you're not expected to eat any more spoiled, moldy food.

Toxic Food

Grilled food develops *benzopyrenes* in it that are very toxic. In an age of lowered immunity, it makes little sense to deliberately poison the food with benzopyrenes. Especially for children, who will be faced with new viruses and parasites in their lifetimes. Will they be able to overcome them or succumb at middle age? Only the strength of their immune system decides this.

Benzopyrenes must be detoxified using the liver's valuable benzene-detoxification system. With so many benzene-polluted items, there is hardly enough detoxification capability to get it all taken care of. NAD enzymes (the N stands for niacin) come into play too. These are essential for alcohol detoxification. If you have consumed alcohol, like a can of beer, NAD enzymes must be shared between the alcohol in the beer and the benzene in the beer. It takes longer to detoxify both the benzene and alcohol. The time delay is a time of lowered immunity and facilitates a growth spurt for parasites and pathogens.

Foods that are raised to very high temperatures, made possible with a microwave oven, produce benzopyrenes. Ordinary bread-toasters can do this too! Old fashioned toasters had a layer of metal separating the bread from the wires. Toasting bread in a frying pan or a stove top surface protects it too. But your stove grill, whether electric or flame, will produce benzopyrenes in your food unless there is a separating wall between them.

It does not matter what kind of fuel is used, the benzopyrenes develop due to lack of shielding between the food and heat source. A metal wall between them absorbs some of the heat.

Do not burn your food in a microwave oven. Since the temperature may go higher than your regular oven, you can produce benzopyrenes. Your regular oven is thermostated so that it cannot go higher than it states. Baking and browning is never done above 425°F. This is your safety feature. If anything in your microwave has turned dark brown or black or has melted plastic, throw it out!

Supplements that help your body to detoxify the benzopyrenes are:

1. Niacin or niacinamide. These are NAD-builders. Take 50 mg with each meal.
2. Vitamin B_2 (riboflavin). Take 50 mg with each meal.

For unpolluted vitamins see *Sources*.

It would be wise to teach children the habits that maximize their immune strength. Avoiding food toxins that are specifically immune-lowering is most important. Besides the benzopyrenes, certain mold toxins and solvents do this and are found in foods.

Moldy Food

Everything that's animal or vegetable can get moldy. While living things are alive, the mold attackers can be held at bay. As soon as they are dead, molding begins. First it molds; then bacterial action sets in. This is what makes things biodegradable. It is a precious phenomenon. It does away with filth—in an exquisite manner. Without mold and decay the streets of New York would still be full of horse manure from the days of the horse and buggy and our lakes too full of dead fish to swim in.

Every grain has its molds; every fruit has its molds; tea and coffee plants have their molds; as do all herbs, and vegetables. Nuts have their molds; nuts grown in the ground (peanuts) are especially moldy because the earth is so full of mold spores. But the wind carries these spores high up into trees, and even up to the stratosphere. Molds are not very choosy. They have their preference for certain plants and conditions. But the same molds can grow on many plants. This is why aflatoxin, for instance, is found not just in your cereal, bread and pasta but in nuts, maple syrup, orange juice, vinegar, wine, etc. Where is it not? It is not in dairy products or fresh fruit and vegetables, provided you wash the outside. It is not in meat, eggs, and fish. It is not in water.

Although I find aflatoxin in commercial bread, I do not find it in carefully screened wheat that has had its discolored, shriveled seeds removed before using it for making bread, cereals and noodles. It is not in baked goods bought at bakeries, left open to air. Evidently the system of wrapping baked goods in plastic keeps moisture trapped and starts the molding process. In spite of adding mold inhibitors, American bread-stuff is far inferior to Mexican baked goods in which I do not find aflatoxin!

Here is some good news for cooks: if you bake it yourself, adding a bit of vitamin C to the dough, your breads will be mold free for an extended period (and rise higher).

Aflatoxin

What is so important about molds? Some of them produce very, very toxic chemicals wherever they grow. They produce some of the most toxic chemicals known to exist. Aflatoxin is one of these. My tests show it is always present in cancer patients; in other words it has built up due to the body's inability to detoxify it in a reasonable time. A great deal of research has been done on aflatoxin. Any library would have more information.

Aflatoxin reaches the liver and simply kills portions of it. After a hefty dose the liver is weakened for a long time—possibly years. Hepatitis and cirrhosis cases always reveal aflatoxin. The liver fights hard to detoxify aflatoxin and manage its own survival. It manages for 2 to 3 weeks; then a portion of it succumbs. So the toxic effects of a dose of aflatoxin aren't even noticeable for several weeks! And without a taste or smell to guide you, how would you know to stop eating the moldy peanut butter or spaghetti? The answer is:

1. make and bake things for yourself
2. test the things you dearly love but can't make
3. treat things that are treatable for molds
4. throw the rest out of your diet

Treatments mentioned in the industrial research journals are hydrogen peroxide, strong alkali such as lime-water, metabisulfite (a common reducing agent) and high heat. I have tried heat and vitamin C, which is also a reducing agent.

Just heating a food to the boiling point does not kill the molds. Boiling for many minutes at a higher temperature or baking does kill them (but not ergot, another mold) and also destroys aflatoxin they produced and left in the food. For foods you can't heat that high, for example nuts that are already roasted, or vinegar, vitamin C comes to the rescue. I suppose it acts a lot like the bisulfite; chemically destroying the mold toxin molecules.

Eradicating Aflatoxin

Simply sprinkling vitamin C over **roasted nuts** is not effective because the molds have penetrated the surface. Rinse the nuts in water first (a lot of mold is removed in this simple way). Cover the nuts with water, add about ¼ tsp. vitamin C powder (for a pint of nuts) and mix. Let stand for 5 minutes. The water penetrates the nuts, taking the vitamin C with it and detoxifies them. Pour off the water and dry the nuts in the oven at low heat. (Don't burn them or you will make benzopyrenes.)

Rice and **pastas** can be demolded partly by cooking and partly by adding vitamin C before or after cooking. There is no need to add so much it affects the flavor. Brown rice is especially moldy.

Vinegars can simply have vitamin C added and placed in the refrigerator.

Honey can be warmed and treated the same way (¼ tsp. per pint).

Bread cannot be salvaged. Switch to bakery breads or homemade. Use it up in a few days, left in its paper bag. Or slice and place in a plastic bag in the freezer.

Since all foods have both their own and others' molds, there must be thousands of molds. Very many have been studied besides aflatoxin producers.

Zearalenone

> Zearalenone, an anabolic and uterotrophic metabolite, is frequently found in commercial cereal grains and in processed foods and feeds, and is often reported as causative agent of naturally occurring hyperestrogenism and infertility in swine, poultry and cattle.[18]

What this means is, in animals, “zear” looks likes extra estrogen to the body. Does it affect humans the same way? Are high estrogen levels a problem for us? I find nearly every breast cancer case shows a too-high estrogen level for years before the cancer is found! It starts females maturing too early, too. It could cause PMS, ovarian cysts and infertility. Not everybody gets all of these effects. And what is the effect on men and boys of eating an estrogen-like mycotoxin in their daily diet? This female hormone could have a drastic effect on the maturing process even in small amounts.

Zearalenone ("zear") and aflatoxin both have immune lowering effects. Zearalenone can induce thymic atrophy and macrophage activation.[19] If you have low immunity (low T-cells, low white blood cell count, and so forth), immediately go off moldy food suspects.

---

[18] Bottalico, A., Lerario, P., and Visconti, A., Production of Zearalenone, Trichothecenes and Moniliformin By Fusarium Species From Cereals, In Italy. From *Toxigenic Fungi, Vol 7*, edited by H. Kurata and Y. Ueno, copublished by Kodansha Ltd, Tokyo and Elsevier Science Publishers B.V., Amsterdam, 1984, page 199.

[19] Luster, M.I., Boorman, GA., Korach, K.S., Dieter, M.P., and Hong, L. 1984. Myelotoxicity toxicity resulting from exogenous estrogens evidence for bimodal mechanism of action. Int. J. Immunopharmacol. 6:287-297.

"Zear" is the mycotoxin that prevents you from detoxifying benzene. Every AIDS sufferer I see has a crippled ability to detoxify benzene; they also have zear!

The main zear sources I have found so far are popcorn, corn chips, and brown rice. But it was absent in fresh corn, canned corn, corn tortillas, and white rice, making me wonder how it gets in our processed corn products.

Sterigmatocystin

Sterigmatocystin ("sterig") is plentiful in **pasta**. Emphasize baked pasta dishes, not boiled. This raises the temperature much higher than boiling. Better yet, make your own pasta with a pasta maker. U.S. bread flour is quite free of mold; the mold in our pastas must come from using inferior quality flour. Always add vitamin C to pasta before or after cooking.

Fig. 50 All U.S. brands of pasta I tested had mold, including health food brands like the one shown here (left). No Mexican brands of pasta, like the one pictured (right), had any molds.

Ergot

A food mold that causes strange feelings and behavior is *ergot*. Although laws regulate the amount of ergot allowed in foods,[20] this is not enough protection. Ergotoxins, for example LSD, are active in extremely minute (less than a microgram, about one thousandth of a fly speck) quantities. They are not destroyed by heat and are especially toxic to children. I found traces in cereals, whole grain breads, wines, and honey. It can be detoxified by adding vitamin C but takes longer; about 10 minutes. Detoxify all your **honey** as soon as it arrives in your house. Warm it slightly and add vitamin C (1/8 tsp. per cup). Stir with wood or plastic.

Ergot toxicity could explain "Jekyll and Hyde" behavior in children, commonly attributed to "allergies". In fact, the mechanism, inability by the liver to keep up with detoxification, fits well into the "allergic" concept. If your child has undesirable behavior, try going off the moldy food suspects for three weeks (cold cereals, nuts and nut butters, store bought breads and baked goods, syrups). Substitute cooked cereals, bakery breads, potatoes, and honeys. Add vitamin C to honey, pasta and cooked cereals. Pancakes and waffles made from scratch would be O.K.

Combining alcohol with ergot is more toxic than either is alone. Alcohol seems to drive the toxin deeper into your tissues. I have found ergot and aflatoxin in beer and wine! Perhaps some of the bizarre behavior and speech of intoxication is really due to the mold-alcohol combination. By delaying alcohol detoxification, the mold could even be responsible for deaths "due to" alcoholism. It would be safer to brew your own alcoholic beverages. Start with pristine fruit. Or at least add vitamin C (1/8 tsp. per cup) to the store bought container you are consuming.

Older children and adults are quite susceptible to ergot too. If bizarre behavior shows up, such as saying mean and cruel things, expressing unusual, irrational thoughts, feeling emotionless or unreal, try the same diet changes, but put alcoholic beverages, soy sauces and other sauces, and other grain derived foods on the "off" list. Try this diet on yourself if you have a temper or crying spells or frequent colds! Ergot can make you super religious, hearing voices of command or threat. Ergot also causes seizures!

---

[20] Canada allows one ergot grain per 300 grains of #3 or #4 wheat.

Fig. 51 All cold cereals I tested were full of mold toxins (besides solvents), health food varieties were worst.

Cytochalasin B

Cytochalasin B ("cyto B") is another immune lowering fungus. I find it mostly in pasta. It stops cells from dividing. Dead portions of the liver cannot regenerate as they otherwise would after a toxic encounter!

Kojic Acid

Kojic acid is a mycotoxin that appears to be responsible for wood alcohol build up. In other words, the toxic effect prevents you from detoxifying wood alcohol. This leads to pancreas damage, invites pancreatic fluke infestation, and typically results in diabetes. I find it in potatoes; don't eat potato skins. If you are a potato lover fix your own so you can peel them and remove any gray parts. I have also found it in regular coffee.

T-2 Toxin

T-2 toxin is a mold I have found in all cases of high blood pressure and kidney disease. It is present on dried peas and beans but it can be detoxified in 5 minutes by adding vitamin C to the water they are soaked in. Remember to throw away imperfect ones, first.

Sorghum Molds

Sorghum and millet carry these. Don't buy sorghum syrup. Rinse millet in vitamin C water before cooking, or add vitamin C to the cooking water.

These mold toxins cause hemorrhaging, appetite loss, and inability to swallow. Elderly people are more easily poisoned than others; their hemorrhages show up as strokes and purple blotches on the skin.

Patulin

is the major fruit mold toxin. It is present in most common fruits if they are bruised. It is particularly hazardous since the mold that produces it can actually grow in your intestine in patches. At these locations, bowel bacteria, *E. coli* and *Shigella*, can climb through the colon wall to invade you. These bacteria are then free to spread to regions of injury and tumors.

If you have cancer or bowel disease go off fresh fruit (bananas and lemons are OK) for a few weeks. Then choose your fruit meticulously. Peel everything so you can see and avoid every bruise. Also take a 2 tsp. dose of Black Walnut Hull Tincture Extra Strength. This kills these bowel fungi. But you can reinfect with a single soft grape.

Mold Avoidance

We should be much more critical of our food.

Crackers are notoriously moldy. Never let your child eat crackers. Make crisp things in the oven from left over bakery goods. Just sprinkle with cinnamon.

Dried fruits are very moldy. Soak them in vitamin C water. Rinse and bake to dry again. Then store in the refrigerator or freezer. When fresh fruit gets overripe, don't quickly bake it or preserve it. It's too late.

Peanut butter (store bought) and other nut butters can't be detoxified by adding vitamin C due to the mixing problem, even if you stir it in thoroughly. Make your own. Making your own peanut butter is a great adventure (see *Recipes*). Mix it with home made preserves, honey, marmalade, not very homogeneously so the bright colors and individual flavors stand out in contrast. Having three or four such spreads in the refrigerator will give your children the right perspective on food—homemade is better. Store bought jams are sweeter and brighter in color but strangely low in flavor and often indistinguishable from each other. Let your children eat the polluted foods that friends and restaurants serve (but not rare-cooked meats) so they can experience the difference. Their livers are strong enough to detoxify occasional small amounts.

Tea is quite moldy if purchased in bags. Although I used to recommend single herb teas (tea mixtures have solvents), I can now only recommend single herb teas from fresh sources in bulk (see *Sources*). This also gets you away from the benzalkonium chloride and possibly other antiseptics in the bag itself. When you get them, store them in their original double plastic bag. These herbs are so fresh, you'll only need half as much to make a cup of tea. Use a bamboo strainer (non metal). Bake the strainer occasionally or put through the dishwasher to keep it sterile.

Fig. 52 Packaged herb tea is moldy and polluted with solvents. Get yours in bulk from an herb company.

It comes as a surprise that pure, genuine **maple syrup** has the deadly aflatoxin and other molds. You can often see mold yourself, as a thin scum on the surface or an opaque spot on the inside of the glass after the syrup has stood some time, even in the refrigerator. Some mold spores were in it to begin with. Others flew in. After some time they grew enough to be visible. In my testing, aflatoxin can be cleared with vitamin C but sterig and others need to be treated with a high temperature as well. Fortunately, this is easy to do with a syrup. Heat to near boiling while in the original jar with the lid removed. Keep refrigerated afterwards.

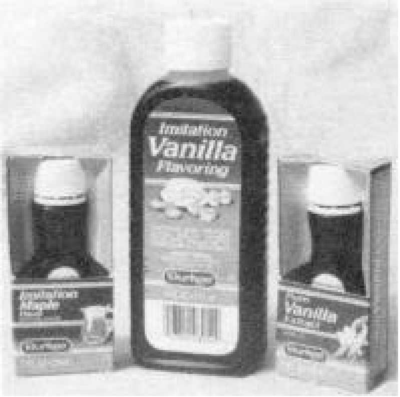

Fig. 53 Three safe flavorings.

Artificial maple flavor did not have benzene, propyl alcohol or wood alcohol, nor molds. **Turbinado sugar** had none of these contaminants either. **Brown sugar** had sorghum mold. **White sugar** had propyl alcohol pollution. You can make your own syrup, safely, with artificial flavor and turbinado sugar. Of course, you'll be missing the taste and nutritional minerals provided by the natural maple product but in a contest between nutritional value and toxicity, always choose the safe product.

The mold in our **hot cereals** can be spotted. Pick out all dark colored, shriveled bits. This represents most of it. Add honey, and salt while it's cooking—this raises the boiling temperature and detoxifies more. At the end, turn off heat and add a sprinkle of vitamin C powder. Rolled oats never showed molds in my testing, although they have their characteristic fungi, too. Don't let grains mold on your shelves simply from aging. Nothing should be more than six months old. Remember you can't see or smell molds when they begin. Molds must have a degree of moisture. As soon as you open a cereal grain, put the whole box in a plastic bag to keep moisture out. This keeps out Weevils too, so you won't have to put the box in the freezer later to kill them.

Anything that is put in the refrigerator or freezer and then taken out develops moisture inside. Store cereals in kitchen cupboards or the freezer.

No government agency can test for all of these mycotoxins in all of our foods. Production and storage methods must be better regulated so as to be fail-safe. Simply sending inspectors out to look into the bins at grain elevators is not sufficient. Crusts of mold, sometimes several feet thick, that form on top of grain bins can be simply shoveled away before the inspector arrives. The humidity and temperature of stored grain should be regulated, requiring automated controls. This would soon be cost effective, too, in terms of reduced spoilage losses and higher quality prices earned. I believe that zear, aflatoxin and ergot require special regulations. Products that are imported should be subjected to the same tests as ours. Test results should be on the label.

Getting Away From Grains

In view of the many molds that are grain-related, and because these cannot be seen or smelled in pastas, breads, cold cereals, it would be wise to steer away from grain consumption. Always choose potatoes, because it is a vegetable instead of a grain, if you have a choice. The potato appears on your plate the way it was harvested. Whereas grain was hulled, stored for quite a long time, perhaps degerminated (the bran and germ picks up

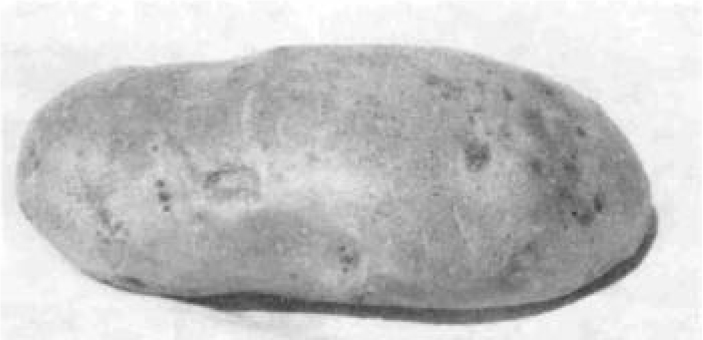

Fig. 54 Don't eat the green on the potato.

mold the fastest). Then it was mixed with assorted chemicals (fumigation, anti oxidants), each polluted in its own way, packaged again and stored again. Grains have a more tortuous history than potatoes that simply get sprayed.

The spray isn't simple, of course. Scrub it off under the tap. If potatoes weren't heavily sprayed they'd be sprouting in the stores. The spray accumulates in the eyes. Cut away all the eyes. By the time you have done this you may as well have peeled them. But no blemish, no cut, no dark spot inside may be left for you to eat. Don't buy potatoes that show a tint of green on them (the green color is due to *scopolamine*; it is toxic). Red potatoes have different chemistry that doesn't produce the green toxin, buy these often. Store potatoes out of the light, to slow down the greening process. They are still a nutritious, vitamin C-rich food—provided you don't fry them in benzene-polluted, hydrogenated grease!

Potatoes have their molds but they are nicely visible. And washing and peeling does away with them. Old literature advises that potatoes should be harvested by moonlight so the green drug isn't produced in the white varieties. With modern mechanized harvesting this should pose no problem. But perhaps this must await the age of robots.

Fig. 55 Potato harvest of the future.

Pets Teach Health Lessons

Dogs don't eat hay and cats don't eat fruit. This is not simply due to their inability to digest them. Nor to training. Their body wisdom guides them. But we can trick them into eating corn and soybeans by adding the flavors they like and thereby defeat their wisdom the same way we defeat ours. A concoction is made for them that is called “complete nutrition” and we feed this meal after meal, day after day, a most unnatural situation. The liver is deluged with the same set of pollutants time after time and never gets a rest. Humans still obey their body wisdom about varying their meals. This gives the liver a chance to catch up with detoxifying one pollutant while the new one builds up. If the liver is absolutely unable to handle something, you are informed quite quickly with an allergic reaction to the food.

Cats and dogs with their monolithic diet get no opportunity to reject food (except by vomiting or starvation). It is not surprising they are getting cancer with increasing frequency, a situation where the liver can no longer detoxify isopropyl alcohol, a common pollutant in their food.

Should we go back to the old days and make their food for them? Yes, they deserve pure food, they deserve variety. Table scraps would be much less toxic for them than their commercial feed. But what if they like and prefer their monolithic “scientific”, “complete”, polluted diet? If our food was doused with sugar for breakfast, lunch and supper, we wouldn't care much about what was under the sugar either. And we'd continue wanting sugar, sugar, sugar the way a pet might want its favorite food and nothing else. Such is the deception of flavorings.

All change should be brought about slowly and with kindness for animals and humans alike. Learn what makes a good pet diet. Cats and dogs are both meat eaters. Cook chicken in a pressure cooker to kill all parasites. Put portions in the freezer. Add table scraps, dressed with a little butter, cheese or lard.

Don't wash the pet dishes with your own—dishwashers don't reach the boiling point. Serve fresh water daily. Standing water picks up bacteria. Don't let food get more than a day old in the dishes. It picks up molds. Don't feed pets at the table, keep them outdoors during mealtime.

After your pets have stopped eating propyl alcohol polluted food and are not getting propyl alcohol in their shampoos, there is no way they can get cancer. Whatever cancer they have will clear up by this change in diet and by giving them the pet parasite program.

Now they are back to a natural state and do not host human flukes. What a relief it is not to worry about reinfection from your pets.

Easy Lifestyle Improvements

None of us likes to change a habit. But once it is changed you are back to an automatic way of doing something. By selecting wise habits your improved lifestyle pays you back for the rest of your life.

Living Hand To Mouth

Hands do everything. They pick up things from the floor. They handle money. They touch other peoples' hands. They clean up bottoms. They touch all kinds of door knobs. And then they pick up food to eat. Some people even lick their fingers when they're sticky or just to turn a page!

What is on the hands that you don't want to eat but can't see? **Bacteria** and **viruses** from coughing and sneezing into your hands! And **cysts**! Cysts are the “eggs” made by parasites. Cysts are so tough not even bleach kills them. They hide under our fingernails when we wash our hands. Then we eat them along with our food. This is called *oral-fecal route*. They hatch in the stomach and go to the intestine to live.

To stop reinfecting yourself the little cysts under the nails need to be killed. Food grade alcohol solution kills them. Buy Everclear™ or Protec (potable) alcohol and make a 5% solution (add ¼ cup of 95% alcohol to a quart of water). Keep it in a small pump bottle at the sink. After using the bathroom and washing your hands, treat your fingernails with alcohol. Pump alcohol into one palm. Put the fingernails of the other hand into it. Scratch a bit. Pour it into the other palm and do the remaining fingernails. Rinse.

- **Don't** eat with your hands! Use a fork.
- **Never** eat food off the floor!
- **Always** wash hands after petting an animal!
- **Never** touch the bottoms of shoes! Keep shoes off couch or bed or chair.
- **Always** cough or sneeze into your clothing or a tissue, not your hands.
- **Keep** your fingers out of your mouth. Don't lick your fingers to turn pages or open plastic bags.

Sick persons need a 50% alcohol solution. Add ½cup 95% alcohol to ½cup cold tap water or buy plain vodka, 80 to 100 proof. Pour the vodka into the pump bottle. Be careful that nobody tries drinking it. If there are teenagers in the house, add a hefty dose of cayenne to it.

Lugol’s iodine will also sterilize your hands. However most commercial Lugol’s is polluted with isopropyl alcohol. Ask your pharmacist to make it from scratch for you (there are only two ingredients and water, see *Recipes*). Then make a solution to wash in (1 tsp. to a quart of water). This can stain some things. Do not use “tincture of iodine.”

Better Laundry Habits

Boil your underwear. In long-ago days, all sheets, towels, table cloths, and underwear were separated and boiled.

With the convenience of our electric washing machine, we tend to overlook the fact that underwear is always contaminated by fecal matter and urogenital secretions and excretions. Mixing these with socks and towels and dishcloths is all right if you are going to kill everything anyway. But if you don't kill them, as in cold water washes, you are mixing the yeast, parasite eggs, bacterial spores, and fungus from underwear with all the other clothing you and your family wear. An enlightened system would be to add an antiseptic to the wash or rinse cycle. Lime water (calcium hydroxide) or iodine based antiseptics seem obviously simple methods to accomplish this. In the absence of this protection, use dryer heat to do your sterilizing. Underwear should be dried until too hot to handle.

Bleach can kill a lot, but doesn't kill *Giardia* spores and a lot of types of fungus. Don't rely on bleaching. Besides, your skin absorbs it from clothing, it is quite toxic to you, and can cause mental effects.

Commercial detergents are polluted with PCBs and have cobalt added. Both of these are easily picked up through your skin. Use **washing soda** and **borax** in the wash cycle. They do not clean quite as well as modern detergents, but there is less static cling, eliminating the need to put more chemicals in your dryer. For spot removal use homemade bar soap.

Better Kitchen Habits

Once a day, sterilize the sponge or cloth you use to wipe up the table, counter tops and sink. This little piece of contaminated cloth is the most infectious thing in the house, besides the toilet. It's more dangerous than the toilet because you do not suspect it. Sometimes it has a slight odor at first, which may warn you, but most pathogens do not have an odor! As we wipe up droplets of milk, we give the milk bacteria, *Salmonellas* and *Shigellas*, a new home to multiply and thrive in. We add crumbs, picking up molds this way. We add dust, picking up parasite eggs and stages. They all feed on the milk and food residue.

As the counter and table and stove get wiped "clean" a film of contamination is left everywhere. A few varieties may die but most of them don't. The general moisture in the kitchen is enough for them to survive. The cloth or sponge recolonizes the kitchen and dining room table several times a day.

No doubt, the last thing you do before leaving the kitchen is squeeze it dry with your hands. Now all the pathogens are on your hands!

Where do your fingers go? To your mouth to remove a hull or bit of something from your teeth. Or to eat a last bite of something. Or to turn a page of the telephone directory. You have just eaten a culture sampling from your own kitchen sponge. In two hours they are already multiplying in the greatest culture system of all: your body! You have given yourself your next sore throat, or cold or headache. The worst possible habit is to wipe a child's face and hands with the kitchen cloth. Or to have a handy towel hanging from the refrigerator handle.

To sterilize the sponge: **drop it into a 50% solution of grain alcohol** at the end of each day. Keep a wide mouth glass quart jar handy just for this. Keep the jar tightly closed and out of the reach of children. Dunk your sponge and plop it onto the sink. If you stand it on end in the sink it will partly dry overnight.

Another way to sterilize the sponge or cloth is to **microwave it**, after wetting it, for 3 minutes. Any shorter time simply warms and cultures the pathogens and multiplies them. Or **boil the cloth** like our grandparents did. Drying out the dish cloth helps kill many–but not all–pathogens. It takes three days of drying to kill all! Another strategy is to **use a fresh cloth** or sponge each day, putting the used one to dry until laundry day.

During the day, set the sponge on end to start drying and slow down culturing.

Don't eat food directly off the counter top or table top. You wouldn't slice a tomato or egg directly on the counter top. It would pick up something: some little particle of dust or dirt. Treat bread the same way. Always on a new clean surface, such as a plate. The counter and table top have on them whatever is in the kitchen dust and on the wipe cloth. Dust is always falling! And the sponge is always culturing. Don't eat the dust!

Keep the cutting board sterile like dishes. Wash it the same way and keep it in the cupboard.

Keep food containers closed. Milk or water glasses are picking up dust as soon as you set them out. Dust is everywhere. Every step on the carpet sends up a puff of dust. Vacuuming sends up a hurricane of dust and distributes bathroom dust to the kitchen and kitchen dust to the bedrooms. So if one person has brought in a new infection, the whole family is exposed to it in hours via the dust.

It is very helpful not to eat the new infectious pathogen. Breathing it is not so damaging. Our noses collect such pathogens and we blow them out again. Touching the infected person is not very damaging either; the pathogens can't get through our skins and since we wash hands before eating we are not at great risk of infection this way. But eating the pathogen is 100% effective in infecting us. The new pathogen is in the dust. The newly contaminated dust drops into your ready and waiting glasses on the table and the open foods. Of course, there is no defense if somebody should cough or sneeze at the table.

Teach children to cough and sneeze into a suitable collecting place like a tissue, not their hands. Pathogens live bountifully on hands. Hands not only provide moisture but often food from the last meal. Hands are second only to the dish cloth in contamination level. If you must cough or sneeze and a tissue is not within reach fast enough, use your clothing! That's what clothing is for—to protect you. Cough and sneeze into your own clothing; this protects the cougher and sneezer, as well as everybody else. A sleeve is handy for children. The inside of your T-shirt for T-shirt wearers. The inside of coat for suited persons. The inside of the neck line for dresses. Of course, paper is best, but in emergency use cloth. Never, never your hands unless you are free to immediately dash into the washroom and clean the contamination off your hands.

Teach children this old rearranged verse:

If you cough or sneeze or sniff
Grab a tissue, quick-quick-quick!
And if you're sitting at the table
Do it in your sleeve if able.

Better Housekeeping

Throw out as much of the wall to wall carpeting as you can bear to part with. It is injurious to everyone's health, even though it's comforting to bare feet and looks pretty.

Carpets clean our shoes. Modern shoes, with their deep treads, bring in huge amounts of outdoor filth which settles deep down into the carpets. In spite of vacuuming every week, the filth accumulates.

Vacuum the carpets when the children, the sick and elderly are out of the house. The dust raised and distributed throughout the house isn't just dirt, it's infectious dirt. It lands on tables and counters. These get wiped with a cloth or sponge and then applied to dishes. The dust in the kitchen falls on open food and into open containers.

Clean carpets with a “steam cleaner”. When you see how much filth is in the water and realize how much dirt you were living with, you might be willing to trade in the “beauty” of carpets for the cleaner living of smooth floors. Don't add chemicals (commercial cleaning solutions) to the steam cleaning machine; these chemicals leave a residue in the carpet which dries and flies up into the air to add to the dust. Popular stain resistance treatments contain arsenic. Cobalt, which adds “lustre” to carpets, causes skin and heart disease after it has built up in your organs. Use borax instead of detergent. Use boric acid to leave a residue that kills roaches and fleas (but not ants). Add vinegar that leaves a residue to repel ants. Nothing controls fleas reliably, except getting rid of the carpets and cloth furniture (keep pets out of bedrooms). Removing all the borax is what brings luster to the carpet. Use citric acid in the rinse water for this purpose. Adding lemon peel to the rinse also adds luster and ant deterrence. Just drop the whole lemon in the tank so it can't block the hoses.

Fleas and other vermin in the carpet simply crawl below the wetness level when you wash the carpet. Spraying a grain alcohol solution with lemon peel in it (it needs to extract for a half hour) on the damp carpet will reach and kill a lot of these, together with the residual bacteria. The damp carpet lets it spread evenly and reach all the crevices.

We are trapped in our dwellings. Primitive peoples were mobile. This got them away from accumulations of filth and rubbish in their living space. Much living was done outdoors, the cleanest space of all.

Now, air conditioning has made indoor living more comfortable. But also has added new hazards. The strong currents of air blow the dust about continuously. Molds and bacteria that grow right on the air conditioning unit get blown about for all to inhale. Never, never use fiberglass as a filter or to insulate your air conditioner around the sides. It is a carcinogen. And the danger of freon escaping from a tiny leak is another major health hazard.

Forced air heating systems are undesirable, too. All dirt brought into the house by shoes gets circulated throughout the house by forced air systems of heating or cooling. Old fashioned radiant heat from radiators or a stove did not distribute the dust so effectively. A return to linoleum floor covering for kitchen and bathroom and hardwood for other rooms would be a good step of progress for a health conscious society. Mopping, instead of the vacuum cleaner, keeps dirt to a minimum. Throw rugs at doors and bedside, easy to clean, would "catch the dirt" as was the original intention. Carpets were intended to help keep filth out of the air. These smaller rugs should be laundered weekly.

Furniture should be wood, cane, or plastic, with cushions to soften the impact. These can be washed weekly if the covers are removable. Modern cloth furniture with its foam interior is a repository of filth and fumes and a constant source of infectious dust.

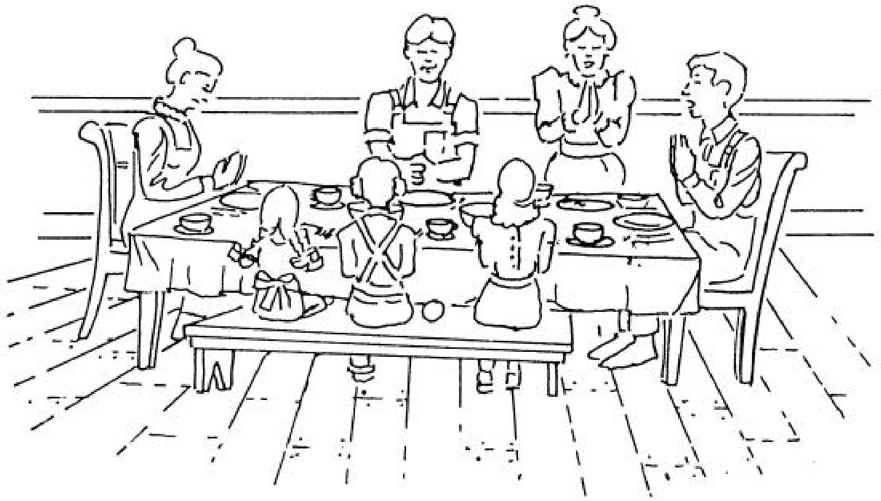

This old fashioned setting is more progressive than our carpeted, modern homes.

Fig. 56 Smooth floors allow every bit of dust to be removed.

Dust your furniture with a damp paper towel. You are picking up and removing highly infectious filth (*Ascaris* and pinworm eggs, pet parasites, "dander" and house mites). Instead of distributing these from room to room, throw the paper towel away after each room is done. Use plain water or vinegar water (50%), not a chemical combination which further pollutes the air.

Clean windows with vinegar water, too. Use a spray bottle.

Keep your dishes in cupboards. This keeps them free of dust. This principle is ancient. It is tempting to leave some of them out. If you must keep the juicer or dishes outside of cupboards, keep them covered or placed upside down so they don't catch dust. Even inside cupboards, store them upside down. When using the “good” dishes or glasses, that haven't been used in a while, wash them first.

Windows Open Or Closed?

In places like Chicago where you can smell the air as you approach the city, it is wiser to keep your windows shut. You can't breath the industrial “soup” all day and night and expect to stay healthy. Of course, it all enters the houses anyway. Central air conditioning and a plain carbon filter at the furnace location (see *Sources*) may be the best solution in spite of blowing dust around the house. Keep the vents to the bedrooms closed to reduce the air turbulence there but leave the cold air return open. Clean the vents in other rooms each week along with floors and carpets by pulling up the grating and reaching down the passage as far as possible.

If you believe the air is free of highway exhaust and industrial smoke open the windows every day. This will let some of the indoor toxins blow away. Asbestos, fiberglass, freon, radon and plain dust can be reduced to a minimum by keeping windows open. If you are ill, sit outdoors (on the porch) as much as you can. Escape to a suitable climate that makes this possible.

Just a few decades ago, many people had summer living quarters that were different from winter living quarters. Gone was all the accumulated infectious dust of half a year of habitation.

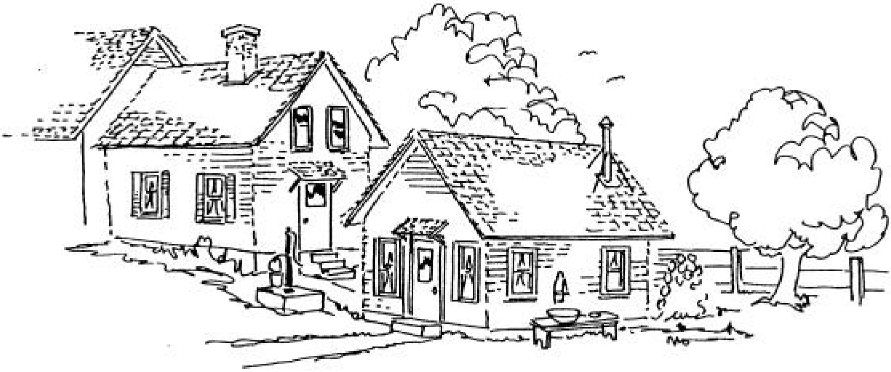

Fig. 57 Moving into the summer kitchen got you away from the accumulation of filth from winter!

Don't have a basement where you stockpile toxic items. Basements invite mold, mice and radon besides toxic things. Fumes travel upward where you live! Keep your toxic things in the attic. If there is no attic, store them in the utility room. Close off the ventilation between utility room and the rest of your house. If you have none of these, perhaps because you live in a senior citizen community or condominium, don't keep any toxic things stored anywhere. Don't save any leftover paints, solvents or cleaners. Buy such small quantities that you can afford to throw it all away when you are done with them.

Live on top of the earth as was intended by nature.

Never have a basement room "finished" for actual living space. Don't buy a house that has a "lower level" built into the earth. This will be the most polluted and dangerous room in your house. If you are ill, move out of such a room. There is no way that it can be “cleaned up”. Move to the other end of the house and furthest away from an attached garage door.

What Kind Of Heat

The worst is coal. The best is none. Breathing coal fumes during the beginning of the industrial age may have brought the new lung diseases: tuberculosis (TB), and pneumonia. It may also have worsened alcohol addiction (beryllium toxicity). Choose electric heat if possible. Even though electricity is based on other fuel consumption, you don't have to breathe those fumes directly.

Wood stoves can be made safe by making sure the chimney works properly. Never use a lighter fluid. Don't fill the house with smoke when stoking.

Minimize your use of fossil fuels in every way you can.

Getting Rid of Mites

We do not tolerate external parasites like bedbugs, lice, ticks, leeches. Bedbugs were once a scourge amongst northern Europeans. I remember our parents spraying for them (kerosene) in the bedroom. This only “controlled” them. What eliminated them was a law against sale of used mattresses. Lice were originally “controlled” by frequent washing, louse combs, and ironing the seams of clothing. What eliminated them was the cutting of long hair as a societal practice. But what about **mites**? They live with us and other animals.

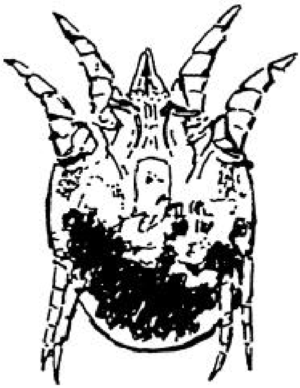

Fig. 58 Mite.

Mites are too tiny to see, tiny enough to ride on a dust particle as if it were a magic carpet. They resemble insects. Chiggers are really mites. Mange in animals is a mite infestation. Dust mites live on our dander (scales of dead skin).

Get rid of their breeding places: beds, cloth covered chairs and soft sofas. Humans leave enough dander behind in these places to support these ultra small insects. Cover mattresses with plastic covers. Use throws on easy chairs and sofas and wash them often. Never allow a pet into the bedroom or the dust will have tapeworm eggs as well as mites. Throw out rugs that have been pet-beds. Spray the air with a mist of 50% grain alcohol before vacuuming. If you have an illness wear a mask to vacuum. Deep, soft, wall to wall carpets compromise an ancient concept: everything should be washable and cleanable, without throwing the dirt into the air for humans to inhale. Vacuuming a carpet blasts mites and tape eggs into the air. Never shake bedding or rugs where the dust will blow back into the house behind you.

Mites don't bite us but we inhale them as they float in the ever present dust in our homes. The mucus in our lungs traps them and in a few days they die, only to release a drove of *Adenoviruses* (common cold virus) in us.

Four Clean-ups

Chronic health problems are not due to exposures of the past. They are ongoing. Your body is constantly fighting to remove pollutants. In order to stay sick, you must be constantly resupplied! These four clean-ups–dental, diet, body, home–are aimed at removing parasites and pollutants at their source. Only then can your body heal.

Dental Clean-up

(This section on dentistry was contributed by Frank Jerome, DDS)

Dr. Jerome: The philosophy of dental treatment taught in America is that teeth are to be saved by whatever means available, using the strongest, most long lasting materials. Long-term toxic effects are of little concern. The attitude of the majority of dentists is: whatever the American Dental Association (ADA) says is OK, they will do.

A more reasonable philosophy is that there is no tooth worth saving if it damages your immune system. Use this as your guideline.

The reason dentists do not see toxic results is that they do not look or ask. If a patient has three mercury amalgam fillings placed in the mouth and a week later has a kidney problem, will she call the dentist—or the doctor? Will they ever tell the dentist about the kidney problem or tell the doctor about the three fillings? A connection will never be made.

It is common for patients who have had their metal fillings removed to have various symptoms go away but, again, they do not tell the dentist. The patient has to be asked! Once the patient begins to feel well they take it for granted, and don't make the connection, either. If everybody's results were instantaneous, there would be no controversy.

Find an alternative dentist. They have been leading the movement to ban mercury from dental supplies. Not only mercury, but all metal needs to be banned. If your dentist will not follow the necessary procedures, then you must find one that will. The questions to ask when you phone a new dental office are:

1. Do you place mercury fillings? (The correct answer is NO. If they do, they probably don't have enough experience in the use of non-metal composites.)
2. Do you do root canals? (The correct answer is NO. If they do, they do not understand good alternative dentistry.)
3. Do you remove amalgam tattoos? (The correct answer is YES. Tattoos are pieces of mercury left in the gum tissue.)
4. Do you treat cavitations? (The correct answer is YES. By cleaning them.) The complete name of *cavitations* is *alveolar cavitational osteopathosis*. They are holes (cavities) left in the jawbone by an incompletely extracted tooth. A properly cleaned socket which is left after an extraction will heal and fill with bone. Dentists routinely do NOT clean the socket of tissue remnants or infected bone. A dry socket (really an infected socket) is a common result. These sockets never fully heal. Thirty years after an extraction, a cavitation will still be there. It is a form of *osteomyelitis*, which means bone infection.

Ninety percent or more of dental offices will not be able to answer ANY of the above questions correctly. If you allow the work to be done by a dentist who does not understand the importance of the above list, you could end up with new problems. Find the right dentist first even if you must travel hundreds of miles. There are 6,000 to 10,000 dentists who should be able to help. Some can do part of the work and refer you to a specialist for the rest. Five hundred to one thousand of these dentists can do it all.

Normal treatment cost is about $1,000 for replacement of 6 to 8 metal fillings including the examination and X-rays. For people with a metal filling in every tooth, or for the extraction of all teeth (plus dentures), it may be up to $3,000 (or more in some places).

Remember, the simpler the treatment, the better. If the dentist says that he or she can change your metal fillings to plastic but it would be better to crown them, say "NO!"

Guidelines For A Healthy Mouth

If you have	**What to do**
Metal fillings	change to plastic fillings
Inlays and onlays	change to plastic fillings
Crowns (all types)	change to plastic crowns
Bridges	change to plastic crowns, partials
Metal partials	change to plastic partials(Flexite™)
Pink dentures	change to clear plastic
Porcelain denture teeth	change to plastic denture teeth
Badly damaged teeth	become extractions
Root canals	become extractions
Braces and implants	avoid
Cavitations	need to be surgically cleaned
Temporary crowns	use plastic
Temporary fillings	use Duralon™

Fig. 59 Dental replacements.

The guidelines can be summarized as:

1. Remove all metal from the mouth.
2. Remove all infected teeth and clean cavitations.

Dr. Clark: *Removing all metal means removing all root canals, metal fillings and crowns. Take out all bridge work or partials made of metal and never put them back in. But you may feel quite attached to the gold, so ask the dentist to give you everything she or he removes. Look at the underside. You will be glad you switched.*

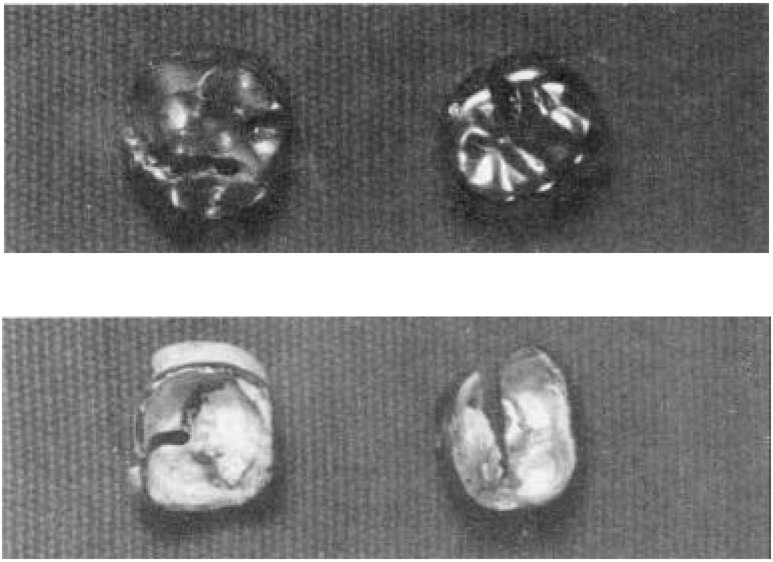

The top surfaces of tooth fillings are kept glossy by brushing (you swallow some of what is removed). Underneath is tarnish and foulness. Ask to see your crowns when they are removed.

Fig. 60 Tops and bottoms of some metal crowns.

The stench of the infection under some teeth may be overwhelming as they are pulled. Bad breath in the morning is due to such hidden tooth infections, not a deficiency of mouthwash!

All metal must come out, no matter how glossy it looks on the surface. Metal does not belong in your body. It is an unnatural chemical. Do this as soon as you have found a dentist able to do it. Find a dentist with experience and knowledge about this subject. It is more than replacing acknowledged culprits like mercury-amalgam fillings. This is metal-free dentistry. Only metal-free plastic should be put back in your mouth.

Dr. Jerome: If your dentist tells you that mercury and other metals will not cause any problems, you will not be able to change his or her mind. Seek treatment elsewhere!

Your dentist should do a complete X-ray examination of your mouth. Ask for the *panoramic* X-ray rather than the usual series of 14 to 16 small X-rays (called full mouth series). The panoramic X-ray shows the whole mouth including the jaws and the sinuses. This lets the dentist see impacted teeth, root fragments, bits of mercury buried in the bone and deep infections. Cavitations are visible in a panoramic X-ray that may not be seen in a full mouth series.

The cost of removing metals should be viewed in the proper light. It took years or decades to get into your present condition. When you do a lot of dental repair in a short time, it can seem to be costly. Unfortunately, many people are in a tight financial position because of the cost of years of ineffective treatment, trying to get well.

Your dentist may recommend crowning teeth to "protect" or strengthen them. Unfortunately, the very concept of crowning teeth is flawed. First, the enamel is removed from a tooth to prepare for the crown. This is permanent and serious damage! Many teeth, up to 20%, may die after being crowned and will need to be extracted. For this reason, you should only get REPLACEMENT crowns and NO NEW crowns. Your metal crowns can be changed to plastic. (Remember, no metal must be left under the crown.)

If you have many crowns, you should have them all removed as quickly as possible. But you should not spend more than two hours in the dentist's chair at any one time. That is too much stress for your body.

Dr. Clark: *Don 't accept intravenous (IV) treatments during amalgam removal. Both IV bags and the supplements used in them are polluted with propyl alcohol, benzene, and wood alcohol.*

Dr. Jerome: It is quite all right to have temporary crowns placed on all teeth that need them in the first visit. You may then go back and complete treatment over the next 6 to 12 months. It is common to find a crowned tooth to be very weak and not worth replacing the crown, particularly if you are already having a partial made and could include this tooth in it.

Dr. Clark: *We are accustomed to thinking that plastic is metal-free. This is wrong. The original dental plastic, methyl methacrylate was metal-free. But modern plastic contains metal. The metal is ground up very finely and added to the plastic in order to make it harder, give it sheen, color, etc.*

Dr. Jerome: Dentists are not commonly given information on these metals used in plastics. The information that comes with dental supplies does not list them either. Most dentists never look at a dental materials book after they graduate. The ADA, however, has a library full of such information.[21]

Dr. Clark: *There are many lanthanide (Rare Earth) metals used in dental plastic. Their effects on the body from dentalware*

---

[21] Call the American Dental Association at (800) 621-8099 (Illinois (800) 572-8309, Alaska or Hawaii (800) 621-3291). Members can ask for the Bureau of Library Services, non-members ask for Public Information.

have NOT been studied. Yet their cancer-promoting ability is known in many cases.[22] *Only metal-free plastic is safe.*

Dr. Jerome: These are the acceptable plastics; they can be procured at any dental lab.

- Plastic for dentures: Methyl Methacrylate. Available in clear and pink. Do not use pink.[23]
- Plastic for partial dentures: Flexite™ Available in clear and pink. Do not use pink.
- Plastic for fillings: Composite Materials. This is the material that has been used in front teeth for 30 years. It has been used in back teeth for 10 years. There are many brands and there are new ones being marketed constantly. The new ones are very much superior to those used 10 years ago and they will continue to improve. They do, however, contain enough barium or zirconium to make them visible on X-rays. There are no alternatives available without these metals.

Dr. Clark: *Composites with barium are not good, but I haven't seen enough barium toxicity from fillings at this time to merit advising extraction instead. Hopefully, a barium-free variety will become available soon to remove this health risk.*

Dr. Jerome: Many people (and dentists too) believe that porcelain is a good substitute for plastic. Porcelain is aluminum oxide with other metals added to get different colors (shades). The metal DOES come out of the porcelain! It has many technical drawbacks as well. Porcelain is not recommended. Sometimes the white composite fillings are called porcelain fillings but they are not. They also require more tooth structure to be removed.

---

[22] Thulium and ytterbium have been studied for their tumor-seeking ability. See page 321 in the book *METAL IONS In BIOLOGICAL SYSTEMS, Vol. 10, Carcinogenicity and Metal Ions*. Editor Helmut Sigel 1980.

[23] The pink color is from mercury or cadmium which is added to the plastic.

If you have a large bridge, it cannot be replaced with a plastic bridge because it isn't strong enough. A large bridge must be replaced with a removable partial (Flexite™).

The methods used to remove metals and infections are technical and complicated. See dental information in *Sources*.

Dr. Clark: *I'd like to thank Dr. Jerome for his contributions to this section, and his pioneering work in metal-free dentistry. I hope more dentists acquire his techniques.*

Horrors Of Metal Dentistry

Why are highly toxic metals put in materials for our mouths? Because not everyone agrees on what is toxic at what level. Just decades ago lead was commonly found in paint, and until recently in gasoline. Lead was not less toxic then, we were just less informed! The government sets standards of toxicity, but those “standards” change as more research is done (and more people speak out). You can do better than the government by dropping your standard for toxic metals to zero! Simply remove them.

The debate still rages over mercury amalgam fillings. No one disputes the extreme toxicity of mercury compounds and mercury vapor. The ADA feels that mercury amalgam fillings are safe because they do not vaporize or form toxic compounds to a significant degree. Opponents cite scientific studies that implicate mercury amalgams as disease causing. Many dentists advocate mercury amalgam fillings simply because they are accepted by the ADA, which they believe protects them from malpractice litigation. Why risk your health and life on their opinions? Remember everything corrodes and everything seeps, so amalgams must too.

Cadmium is used to make the pink color in dentures! Cadmium is five times as toxic as lead, and is strongly linked to high blood pressure.

Occasionally, thallium and germanium are found together in mercury amalgam tooth fillings. Thallium causes leg pain, leg weakness, and paraplegia. If you are in a wheelchair without a very reliable diagnosis, have all the metal removed from your mouth. Ask the dentist to give you the grindings. Try to have them analyzed for thallium using the most sensitive methods available, possibly at a research institute or university.

I was astonished to find thallium in mercury amalgams! It couldn't be put there intentionally, look how toxic it is:

> ***TEJ500*** ***HR: 3***
>
> **THALLIUM COMPOUNDS**
>
> Thallium and its compounds are on the Community Right To Know List.
>
> THR: Extremely toxic. The lethal dose for a man by ingestion is 0.5-1.0 gram. Effects are cumulative and with continuous exposure toxicity occurs at much lower levels. Major effects are on the nervous system, skin and cardiovascular tract. The peripheral nervous system can be severely affected with dying-back of the longest sensory and motor fibers. Reproductive organs and the fetus are highly susceptible. Acute poisoning has followed the ingestion of toxic quantities of a thallium-bearing depilatory and accidental or suicidal ingestion of rat poison. Acute poisoning results in swelling of the feet and legs, arthralgia, vomiting, insomnia, hyperesthesia and paresthesia [numbness] of the hands and feet, mental confusion, polyneuritis with severe pains in the legs and loins, partial paralysis of the legs with reaction of degeneration, angina-like pains, nephritis, wasting and weakness, and lymphocytosis and eosinophilia. About the 18th day, complete loss of the hair on the body and head may occur. Fatal poisoning has been known to occur. Recovery requires months and may be incomplete. Industrial poisoning is reported to have caused discoloration of the hair (which later falls out), joint pain, loss of appetite, fatigue, severe pain in the calves of the legs, albuminuria, eosinophilia, lymphocytosis and optic neuritis followed by

> atrophy. Cases of industrial poisoning are rare, however. Thallium is an experimental teratogen [used to induce birth defects for study]. When heated to decomposition they [sic] emit highly toxic fumes of Tl [thallium]. See also THALLIUM and specific compounds.[24]

Fig. 61 Thallium excerpt.

Thallium pollution frightens me more than lead, cadmium and mercury combined, because it is completely unsuspected. Its last major use, rat poison, was banned in the 1970s. Every wheelchair patient I tested was positive for thallium! One current use for thallium is in Arctic/Antarctic thermostats. When added to mercury the mercury will stay liquid at lower temperatures. Are mercury suppliers then providing the dental industry with tainted amalgam?

The cancer causing or carcinogenic action of metals has been studied for a long time, although it doesn't get attention by our regulatory agencies. A scientific book on this subject was published in 1980.[25] One table from this book is shown on page 431. We can see that chromium and nickel compounds are the most carcinogenic metals. Nickel is used in gold crowns, braces, and children's crowns!

Note that the form of the metal is very important. For instance chromium is an essential element of *glucose tolerance*

---

[24] *Dangerous Properties of Industrial Materials*, 7th ed. by N. Irving Sax and Richard J. Lewis Sr., Van NOSTRAND, Reinhold N.Y. 1989.

[25] The title is *Carcinogenicity and Metal Ions*. It is volume 10 of a series called *Metal Ions in Biological Systems*, edited by Helmut Sigel. A university chemistry library should have this book. It has a fascinating chapter on the leukemias by two scientists from the Academy of Sciences of the USSR, E. L. Andronikashvili and L. Mosulishvili. Their brilliant work and discussion was largely responsible for my pursuit of the whole subject of cancer.

factor, but most of its other compounds are extremely toxic. In general, xenobiotic compounds (foreign) are to be avoided! Metal doesn't belong in our foods or in our bodies.

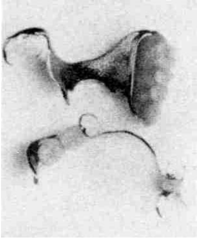

Fig. 62 More dental metal.

Dental Rewards

After your mouth is metal and infection-free, notice whether your sinus condition, ear-ringing, enlarged neck glands, headache, enlarged spleen, bloated condition, knee pain, foot pain, hip pain, dizziness, aching bones and joints improve.

Keep a small notebook to write down these improvements. It will show you which symptoms came originally from your teeth. Symptoms often come back! So go back to your dentist, to search for a hidden infection under one or more of your teeth, or where your teeth once were! That infection can be the cause of tinnitus, TMJ, arthritis, neck pain, loss of balance, and heart attacks!

Dentures can be beautiful. Of course, plastic isn't natural, but it is the best compromise that can be made to restore your mouth. At least it isn't positively charged like metals; it can't set up an electric current nor a magnetic field in your mouth, all of which may be harmful.

Do not be swayed by arguments that plastic is not as strong as metal. You see dentures everywhere and they seem strong enough to eat with. You will be told that “noble” metals like gold and platinum

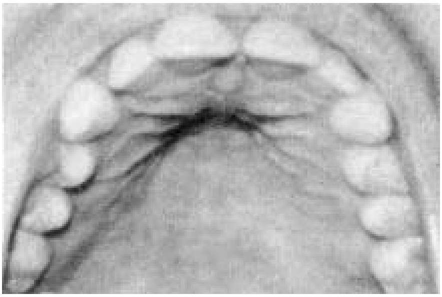

Fig. 63 Beautiful plastic mouth.

and silver are OK, that they are "inert" and do not corrode or seep. Nothing could be more untrue. You may be keeping them glossy by the constant polishing action of your toothpaste. But if you look at the underside, the view is frightful. **Everything tarnishes and everything seeps.** You wouldn't expect even a gold or silver coin that was dropped in a fountain 50 years ago to be intact. As metal corrodes your body absorbs it!

In breast cancer, especially, you find that metals from dentalware have dissolved and accumulated in the breast. They will leave the breast if you clear them out of your mouth (and diet, body, home). **The cysts shrink and are simply gone.** No need to do surgery!

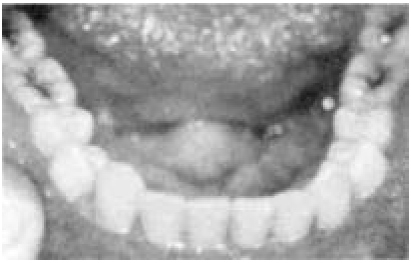

Fig. 64 Ugly metal in mouth.

Diet Clean-up

Breakfast

Cook your cereal from scratch. Don't eat cold cereal; it has numerous solvents and molds. Buy hot cereals that say "no salt added," like cream of wheat, steel cut oats or old fashioned oats,[26] millet, corn meal, cream of rice, or Wheatena. Cook it with milk to add nutritive value. Add your own (non-aluminum) salt and a pinch of vitamin C before cooking. Make granola from a recipe (see *Recipes*). Use honey, or brown sugar. Add raisins that were soaked for 5 minutes in vitamin C water. Use whipping cream or butter (both boiled) if you need to gain weight. Isn't this a delicious way to start your day! Add cinnamon to flavor, or frozen fruit and honey.

---

[26] Rolled oats have 235 mcg nickel per serving of 4 ounces, picked up from the rollers, according to *Food Values* 14th ed. by Pennington and Church, 1985. I have only found nickel in the "one-minute" or "instant" variety of oats, however.

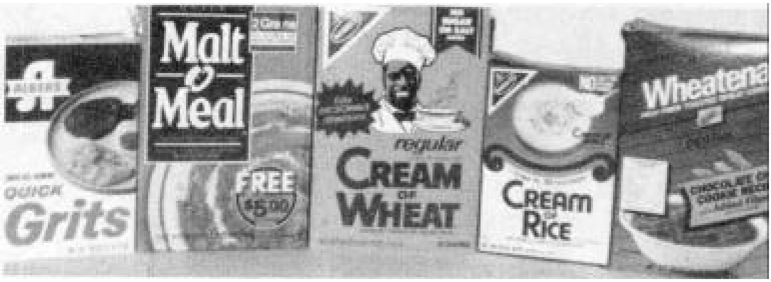

Fig. 65 Unpolluted breakfast cereals.

Or start your day with fried potatoes, an egg, and glass of milk. Don't worry about cholesterol since you will be doing liver cleanses anyway. (We have been told that eggs carry *Salmonella* bacteria. I found *Salmonellas* only on the outside shell and the egg carton—never inside! Could the researchers have accidentally transferred the bacteria from the shell to the inside while they were testing?)

The milk should be 2% or more butterfat because the calcium in milk cannot be absorbed without at least this much fat. Eat homemade yogurt and add honey or homemade preserves yourself. You need 3 cups of a milk product each day. Homemade buttermilk is fine. If you don't tolerate milk, and get diarrhea from it, try a milk digestant tablet to go with it. Start with only ¼ cup at a time. Do not choose chocolate milk. There is no substitute for milk; calcium tablets are not satisfactory. Vegetable matter, although high in calcium, does not give you *available* calcium either, unless you buy a juicer and make vegetable juice out of it. Eating fish can give you a lot of calcium, but it is in the tiny bones hidden in the fish. Don't try to remove them. Canned salmon has a lot of calcium; tuna does not. On a day that you eat fish, you would not need milk. Goat milk is probably better than cows' milk, but more difficult to get used to.

Lunch

Cook your food from scratch. Don't start with cans or packages or frozen items to make some recipe. In fact, don't bother with any fancy recipes. Just cook two or three vegetables for lunch and eat them with butter and salt or homemade sauces. Bread and milk rounds it out, plus fruit (not canned or frozen). Soup is a nice change. Cook it with all the vegetables you can find. Don't start with a can or packet or cube. Use a bit of onion and genuine herbs to give it zest. Thyme and fenugreek, together, make a flavorful combination you can purchase in capsules. Just pull apart and season.

If all this is too much work, make fresh vegetable juice once a week and freeze enough so that you can have a daily nutritious meal just by pouring a glass of it, together with bread and yogurt or milk.

Never diet during illness

This is a rule based on common sense. A weight-loss diet must wait at least two years.

Bake your own bread! I found aflatoxin in commercial bread after just four days in my bread box, but none in homemade bread even after two weeks! Aflatoxin is a most potent carcinogen and immunosuppressant. Aflatoxin is the toxin in your diet that keeps you from clearing propyl alcohol from your body (see 382)! Aflatoxin is a substance made by mold; bread starts to mold on the grocery store shelf. Don't buy plastic-wrapped bread. Don't ever salvage moldy food, whether it is fruit, breads or leftovers in the refrigerator. Throw them out. Buy a bread maker. It can do everything, including baking the bread. Use unbleached (unbrominated) flour and add ½ tsp. vitamin C powder per loaf to help retard mold further (it also makes the bread rise higher).

Supper

Cook your supper from scratch. Emphasize fish for animal food, not beef, pork, turkey or chicken. Don't buy bread crumbs, use your own. Don't buy batter, make your own. Use genuine eggs, not substitutes. Wash your hands after handling raw meat or eggs.

Make your own salad and salad dressing out of olive oil, fresh lemon juice or white distilled vinegar (apple cider vinegar has aflatoxins), honey, salt and herbs to flavor. If your digestion isn't strong enough for raw vegetables or fruit, make juice. Get a sturdy juicer and make your juice about half carrot juice and half from vegetables like celery, squash, lettuce, and broccoli. Make your own tomato sauce with pure herb seasoning, not from ajar or can (home canned foods are fine, of course, as long as they are not made in a big aluminum pot with aluminum-containing salt).

Cook real potatoes, not instant varieties. Peel them to get rid of Kojic acid (mycotoxin) and scopolamine (the green part). Make mashed potatoes from scratch, with milk, not box potatoes, nor chips nor French fries. Chips and fries were made in chemical grease called "hydrogenated." There is a large amount of nickel in hydrogenated fats.[27] Fry your potatoes in butter, lard or olive oil. Find butter that is not wrapped in foil and is not salted. Salt your own butter, using aluminum-free salt.[28] Don't wrap potatoes in foil to bake, coat with butter or olive oil. Don't eat the peels.

Eat no meat that hasn't been cooked as thoroughly as if it were pork.

Other animals are as parasitized as we, full of flukes and worms and *Schistosomes* in every imaginable stage, and if the blood carries these, would we not be eating live parasites if we eat animals in the raw state? We have been taught to cook thoroughly any pork, fish or seafood. Now we must cook thoroughly any beef, chicken or turkey. It must be at cooking temperature (212°F or 100°C) for 20 minutes. Freezing is not adequate. Canned meats are safe from living parasites, but are not recommended due to added chemicals.

Beverages

Drink 6 kinds of beverages:

- milk
- water
- fruit juices
- vegetable juices
- herb teas
- homemade (see *Recipes*)

---

[27] 114 mcg/100 g. Taken from *Food Values* 14th ed. by Pennington and Church, 1985.

[28] Salt has aluminum in it to keep it from caking. Buy salt that has magnesium carbonate as its anti-caking agent. Sea salt must be baked for 5 minutes at 400°F to destroy molds. Or buy chemically pure salt (see *Sources*).

This means getting off caffeine. And if you are already fatigued, this means you might be even more fatigued for a short time. You might have headaches from withdrawal, too. But they will only last 10 days. Mark your calendar and count off the days. Take headache medicine, if necessary, but make sure it does not contain caffeine. For energy, to replace caffeine, take one arginine (500 mg, see *Sources*) upon rising in the morning and before lunch. Soon you won't need it.

Cutting down on coffee, decaf, soda pop and powdered drinks won't do. You must be completely off. They contain very toxic solvents due to careless, unregulated production methods. Much is imported and can't be sufficiently regulated.

Even though grain (drinking) alcohol is the recommended substitute for propyl alcohol, that doesn't mean you may safely drink it. It is inadvisable to drink any form of alcohol at least until you are fully recovered.

1. Milk: 2% or higher, drink three 8 oz. glasses a day. Alternate brands. Buttermilk will do. Homemade yogurt is fine. Goat milk is also fine. Start with ¼ cup and increase gradually, if you are not used to it. If you do not drink milk because it gives you more mucous, try to drink milk anyway. If you have other reactions, like diarrhea, try milk digestant tablets (available at health food stores). Milk is too valuable to avoid: there are many unwanted chemicals in most brands of milk, but it is solvent-free, mold-free and very nutritious. The only exception should be for serious symptoms, like swelling, colitis, flu, or chronic diarrhea.

But all milk, whether goat or cow, is contaminated with *Salmonella* and *Shigella* bacteria as well as fluke parasite stages. Cattle are immunized against *Salmonella* but it does not prevent its persistence in the bowel. All these are very harmful. Pasteurization does not kill all of them. Only heating to a rolling boil makes milk safe. To do this in the easiest way, pour 1 or 2 quarts milk into an enamel double boiler or microwavable glass jar. Stay in the kitchen while the heat is on. When the bubbles have risen to indicate boiling, turn off the heat. You may throw away the “skin”. Pour into glass jar and refrigerate. Another easy way is to use a pressure cooker that holds several pint jars of milk. All dairy products that have only been pasteurized are still contaminated. Ultrapasteurization does not improve matters. Dairy products that cannot be sterilized should not be consumed. It may be possible to find sterilized milk in paper containers on the store shelf—not in the refrigerator; if it wasn't sterile it would go foul in a day! Canned milk has solvent pollution. Powdered milk has both solvent and bacterial pollution.

2. Water: 2 pints. Drink one pint upon rising in the morning, the other pint in the afternoon sometime. The cold water faucet may be bringing you cadmium, copper or lead, but it is safer than purchased water, which inevitably has solvents in it. Let it run before using it. Filters are rather useless because water pollution comes in surges. A single surge of PCB contaminates your filter. All the water you use after this surge is now polluted, so you will be getting it chronically, whereas the unfiltered water cleans up again after the surge passes. Until you can test your own water for solvents, PCBs and metals, no expensive filter is worth the investment. An inexpensive pure carbon filter that is replaced every month may improve your tap water. Inflexible plastic pitchers fitted with a carbon filter pack are available (see *Sources*). Never buy filters with silver or other chemicals, even if they are just added to the carbon. Keep the filter sterile by soaking in diluted grain alcohol weekly.

3. Fruit juice: fresh squeezed only. Some stores make it while you wait. If they freeze some of it, you could purchase the frozen containers. Bottled fruit juices have traces of numerous solvents, as do the frozen concentrates, as do the refrigerated ones, don't buy them. You have to see it being made, but watch carefully: I recently went to a juice bar where they made everything fresh, before your very eyes. And I saw them take the fruit right from the refrigerator and spray it with a special wash "to get rid of any pesticides," then put a special detergent on it to clean off the wash! So instead of getting traces of pesticide, I got traces of propyl alcohol![29] Another grocery store had a machine that squeezed the oranges while you watched. But if you did not watch them filling the jugs, you missed seeing them add a tablespoon of concentrate, from a bottle out of sight, to give it better flavor. It still qualifies as "Fresh squeezed 100% orange juice," but thanks to that concentrate it now has toluene and xylene in it! Best of all, buy a juicer, select completely unbruised fruit, wash with plain water, and make your own juice (enough for a week—freeze it in half pint plastic bottles). For stronger flavor, leave some of the peel in the juice.

4. Vegetable juice: fresh or frozen only. If you or a friend would be willing to make fresh juice, this would be much better than purchased juice. Start with carrot juice. Peel carrots (don't scrape them, it's too easy to miss small dirt spots) and remove all blemishes carefully, then rinse. Drink ½glass a day. After you are accustomed to this, add other vegetables and greens to the juice to make up half of it. Use celery, lettuce, cabbage, cucumber, beet, squash, tomato, everything raw that you normally have in your refrigerator. Then drink one glass a day.

5. Herb tea: fresh or bulk packaged. Tea bag varieties are moldy. Buy a non-metal (bamboo is common) tea strainer. Sweeten with honey or brown sugar with vitamin C added.

6. Homemade beverage. If you will miss your coffee or decaf, try just plain hot water with boiled whipping cream. Sweeten with honey. Please see *Recipes* for many more suggestions.

---

[29] Yes, I took a sample of the wash to test.

Horrors In Commercial Beverages

Commercial beverages are especially toxic due to traces of solvents left over from the manufacturing process. There are solvents in decaffeinated beverages, herb tea blends (not single herb teas), carbonated drinks, beverages with Nutrasweet™, flavored coffee, diet and health mixes, and fruit juices, even when the label states "not from concentrate" or "fresh from the orchard," or "100% pure."

It is allowable to use solvents to clean machinery used in bottling (please look again at page 347)! It is also allowable to use solvents to make spice oleoresins, which are used as flavoring.

> **21 CFR 173.240 (4-1-94 Edition) Isopropyl Alcohol.**
>
> Isopropyl alcohol may be present in the following foods under the conditions specified:
>
> (a) In spice oleoresins as a residue from the extraction of spice, at a level not to exceed 50 parts per million.
> (b) In lemon oil as a residue in production of the oil, at a level not to exceed 6 parts per million.
> (c) [Discusses its use in hops extract.]

Here is a summary of other solvents mentioned:

Solvent	Allowable residue in spice oleoresins	Paragraph in 21 CFR
Acetone	30 PPM	173.210
Ethylene dichloride	30 PPM	173.230
Methyl alcohol	50 PPM	173.250
Methylene chloride	30 PPM	173.255
Hexane	25 PPM	173.270
Trichloroethylene	30 PPM	173.290

Fig. 66 Lawful uses of solvents in food.

I have found all these solvents and others in commercial beverages! Some of the solvents I have found are just too toxic to be believed! Yet you can build the test apparatus yourself (page 457), buy foods at your grocery store, and tabulate your own results. I hope you do, and I hope you find that the food in your area is cleaner than mine! Remember that the Syncrometer can only determine the presence or absence of something, not the concentration. There may only be a few parts per billion, but a sick person trying to get well cannot afford any solvent intake. For that matter, none of us should tolerate any of these:

- **Acetone** in carbonated drinks
- **Benzene** in store-bought drinking water (including distilled), store-bought fruit juice (including health varieties)
- **Carbon tetrachloride** in store-bought drinking water
- **Decane** in health foods and beverages
- **Hexanes** in decafs
- **Hexanedione** in flavored foods
- **Isophorone** in flavored foods
- **Methyl butyl ketone** and **Methyl ethyl ketone** in flavored foods
- **Methylene chloride** in fruit juice
- **Pentane** in decafs
- **Propyl alcohol** in bottled water, commercial fruit juices, commercial beverages.
- **Toluene** and **xylene** in carbonated drinks
- **Trichloroethane** (TCE), **TC Ethylene** in flavored foods
- **Wood alcohol** (methanol) in carbonated drinks, diet drinks, herb tea blends, store-bought water, infant formula

If you allowed a tiny drop of kerosene or carpet cleaning fluid to get into your pet's food every day, wouldn't you expect your pet to get sick? Why would you not expect to be sick with these solvents in your daily food? I imagine these solvents are just tiny amounts, introduced by sterilizing equipment, the manufacturing process, and adding flavor or color. Flavors and colors for food must be extracted somehow from the leaves or bark or beans from which they come. But until safe methods are invented, such food should be considered unsafe for human consumption (or pets or livestock!).

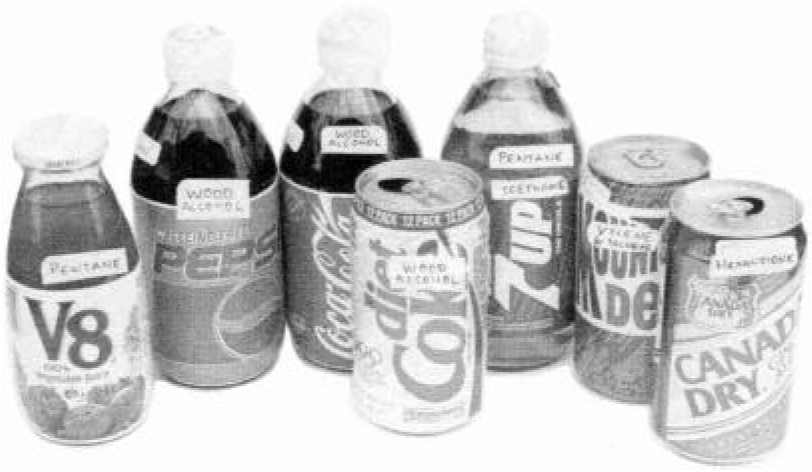

Fig. 67 Some unsafe beverages.

Food Preparation

Cook your food in glass, enamel, ceramic or microwavable pots and pans. Throw away all metal ware, foil wrap, and metal-capped salt shakers since you will never use them again. If you don't plan to fry much (only once a week), you might keep the Teflon™ or Silverstone™ coated fry-pan, otherwise get an enamel coated metal pan. Stir and serve food with wood or plastic, not metal utensils. If you have recurring urinary tract infections, you should reduce your metal contact even further; eat with plastic cutlery. Sturdy decorative plastic ware can be found in hardware and camping stores. Don't drink out of styrofoam cups (styrene is toxic). Don't eat toast (many toasters spit tungsten all over your bread and make benzopyrenes besides). Don't buy things made with baking powder (it has aluminum) or baked in aluminum pans. Choose goods made with baking soda and sold in paper or microwavable pans. Don't run your drinking water through the freezer or fountain or refrigerator. Don't heat your water in a coffee maker or tea kettle. Don't use a plastic thermos jug (the plastic liner has lanthanides) the inside must be glass. Don't drink from a personal water bottle (it begins to breed bacteria) unless you sterilize it daily.

Why are we still using stainless steel cookware when it contains 18% chromium and 8% nickel? Because it is rustproof and shiny and we can't see any deterioration. **But all metal seeps!** Throw those metal pots away. Get your essential minerals from foods, not cookware.

Never, never drink or cook with the water from your hot water faucet. If you have an electric hot water heater the heating element releases metal. Even if you have a gas hot water heater, the heated water leaches metals or glues from your pipes. If your kitchen tap is the single lever type, make sure it is fully on cold for cooking. Teach children this rule.

Food Guidelines

It is impossible to remember everything about every food, but in general do not buy foods that are highly processed. Here are a few foods; see if you can guess whether they should be in your diet or not.

breads	Yes, but only from a bakery, and never wrapped in plastic.
toast	No. It has benzopyrene and tungsten. Yes, if made on a cookie sheet or in a frying pan.
cheese	Yes, if used in baked dishes.
chicken	Only if cooked for 20 minutes at boiling point, as in soup, or canned (never prepare raw chicken yourself).
wine with dinner	No.

peanut butter	Yes if you grind it yourself and add ¼ tsp. vitamin C powder as you grind.
cottage cheese	No, it can't be sterilized easily.
desserts	Yes, but again, only if flavored with safe extracts.
rice	Yes, if vitamin C is added before cooking. Use white only, brown is too moldy.
pasta	Yes, with homemade sauce and vitamin C.
Jell-O™	No, it has artificial flavor and color.
egg dishes	Yes, but not “imitation”, cholesterol-free or cholesterol-reduced varieties.
fish, seafood	Yes!
soy foods (tofu)	No. It's the extensive processing that taints it.
soup	Yes, if seasoned only with herbs (no bouillon cube).
sugar	Yes, turbinado or brown if treated with vitamin C.
herb tea	Yes, if not in a bag and not in a mixture of herbs.
cheesecake	Yes.

Fig. 68 Some good foods.

Choose brands with the shortest list of ingredients. Alternate brands every time you shop.

Fig. 69 All breads I tested had mold if they were in plastic.

Dining Out

Restaurants (excluding fast food) are generally quite safe to eat at. Here are some do's and don'ts;

Do carry your own aluminum-free salt and vitamin C powder with you.

Do ask for plastic cutlery.

Do drink the water if from the tap.

Do ask for boiled, not just steamed, milk.

Don't eat or drink from styrofoam. If getting food "to go," get it in clear plastic containers, or ask them to line the styrofoam container with paper or plastic wrap, and line the styrofoam cup with a plastic baggy.

Don't use their ketchup and condiments (they have been standing out too long).

Here is a list of things that are generally safe to order:

pancakes, French toast, waffles	Don't use their imitation syrup (has benzoate), use honey instead.
eggs	Any style except soft boiled and scrambled. The white should be solid.
hash browns	If lightly fried, not deep fried.
soup	Only if nothing else is available. (It probably came in a can and was cooked in an aluminum pot and is full of aluminized salt.)
vegetarian sandwiches	But no soy products (too processed).
baked or boiled potatoes	Use only cheese sauce, bring your own salt, don't eat the skin.
cooked vegetables	Broccoli, Brussels sprouts, beets, corn, squash, and so forth.
vegetable salads	Don't eat the croutons, bacon bits, and anything that doesn't look fresh.
vegetarian dishes	But no soy ingredients and or sauces. Fresh ketchup OK.
bread and biscuits	White only, not toasted, not “cholesterol-free” varieties.
fish and seafood	Anything but deep fried (the oil may have benzene) is fine: baked, steamed, fish cakes, seafood cocktails, etc.
Mexican food	Any of the numerous baked dishes.
Chinese food	Except dishes with tofu or MSG.
fruit cup	With honey and cinnamon.
fruit pies, cobblers	But not with ice cream (every flavor has benzene).
lemon or lime meringue pie	Indulge yourself.

Fig. 70 Good restaurant foods.

As you see your symptoms disappear, one after another, you will feel the magic of healing. Many sick persons have 50 or more symptoms to start out! They could fill two sheets of paper, one symptom to each line. It can be quite shocking to see a list of all your symptoms.

Sometimes a new symptom appears as fast as an old one disappears. The coincidence makes it tempting to believe that one symptom turns into a different one. But it is not so. If a new symptom appears, it is because another pathogen has become activated due to a new toxin. Try to identify the new item. Stop using any new food, supplement, or body product, even if it is a health variety, and see if it goes away.

Body Clean-up

We are living in a very fortunate time. We are not expected to all look alike! The 60's brought us this wonderful freedom. Freedom to dress in a variety of styles, use make-up or no make-up, jewelry or no jewelry, any kind of hair style, any kind of shoes.

You will need to go off every cosmetic and body product that you are now using. Not a single one can be continued. They are full of titanium, zirconium, benzalkonium, bismuth, antimony[30], barium[31], strontium[32], aluminum, tin, chromium, not to mention pollution solvents such as benzene and PCBs.

Do not use any commercial salves, ointments, lotions, colognes, perfumes, massage oils, deodorant, mouthwash, toothpaste, even when touted as "herbal" and health-food-type. See *Recipes* for homemade substitutes.

---

[30] Breast cancer cases show titanium, zirconium, benzalkonium, bismuth and antimony accumulation in the breast.

[31] Barium is described in the *Merck Index* as "Caution: All water or acid soluble barium compounds are POISONOUS." 10th ed. p. 139 1983.

[32] This element goes to bones.

People are trying desperately to use less toxic products. They seek health for themselves. So they reach for products that just list herbs and other natural ingredients. Unfortunately, the buyers are being duped. The Food and Drug Administration (FDA) requires all body products to have sufficient antiseptic in them. Some of these antiseptics are substances you must avoid! But you won't see them on the label because manufacturers prefer to use quantities below the levels they must disclose. And by using a variety of antiseptics in these small amounts they can still meet sterility requirements. The only ingredient you might see is “grapefruit seed” or similar healthy-sounding natural antiseptic. This is sad for the consumer of health food varieties.

- I have seen rocks sold as “Aluminum-Free Natural Deodorant”. You rub the rock under your arms. It works because the rock is made of magnesium-aluminum-silicate.
- Men's hair color has lead in it.
- Lipstick has barium, aluminum, titanium.
- Eye pencil and shadow have chromium.
- Toothpaste has benzene, tin, and strontium.
- Hair spray has propyl alcohol and PCBs. BEWARE! Stop using it today.
- Shampoo, even health varieties, has propyl alcohol! BEWARE! Stop using it today.
- Cigarettes have lead, mercury, nickel and *Tobacco Mosaic virus*.
- Chewing tobacco has ytterbium

Some of the unnatural chemicals listed are present because of residues in the manufacturing process, but others you will actually see listed on the label!

Propyl alcohol and wood alcohol are present because the tubing used to fill the bottles is sterilized and cleaned with them. Ice cream machines are “oiled” with a gel containing petroleum products. This could explain why I always find benzene in ice cream.

SUPER
TAYLOR-LUBE
TAYLOR
AMOCO OIL CO.
4 FL. OZ.
TAYLOR
LUBE

Fig. 71 Examples of commercial "food lube."

How can propyl alcohol in shampoo get into your body in significant amounts? The skin is more absorbent than we realize, and time and time again I see cancer victims who have gone off every body product except their favorite shampoo. They harbor propyl alcohol until they make that final sacrifice. It is better to switch shampoos than to not need any due to radiation and chemotherapy!

See *Recipes* for easy-to-make, natural cosmetics. But you might consider just stopping them all. Especially if you're going on vacation.

Use nothing that you wouldn't use on a new-born baby. This is a permissive age. You will be the only one feeling "naked." Others won't even notice. Don't forget advertising is aimed at you, even if other people's eyes are not!

Don't even use soap unless it is homemade soap (see *Recipes*) or borax straight from the box. Borax was the traditional pioneer soap. It is antibacterial and can be made into a solution. It is also a water softener and is the main ingredient in non-chlorine bleach. Borax can remove grease, too, and some stains.

But even borax is not natural to your body and it is therefore wise to use as little as necessary. See *Recipes* for anti-bacterial borax soap.

Dishes and clothing are primary source of PCBs.

Fig. 72 Detergents with PCBs.

MULE 20 TEAM
BORAX
100 YEARS

homemade soap pure borax

Fig. 73 Safe soaps.

Don't use toothpaste, not even health-food varieties. To clean teeth, use plain water or chemically pure baking soda (see *Sources*)—but dissolve it in water first, otherwise it is too abrasive. Or brush with hydrogen peroxide food grade, not the regular variety (see *Sources*). Don't use floss; use 2 or 4-pound monofilament fish line. Floss has mercury antiseptics (with thallium pollution!). Throw away your old toothbrush—solvents don't wash away.

Don't use mouthwash. Use saltwater (aluminum-free salt) or food grade hydrogen peroxide (a few drops in water).

Don't use hair spray.

Don't use massage oils of any kind. Use olive oil.

Don't use bath oil. Take showers, not baths, if you are strong enough to stand. Showers are cleaner.

Don't use perfumes or colognes.

Don't use commercial lotions or personal lubricants.

Stop Using Supplements

Stop using your vitamin supplements. They, too, are heavily polluted. This is the saddest, most tragic part of your instructions. I have found solvents, heavy metals and lanthanides in 90% or more of the popular vitamin and mineral capsules and tablets I test. These substances will do more harm in the long run than the supplement can make up for in benefits.

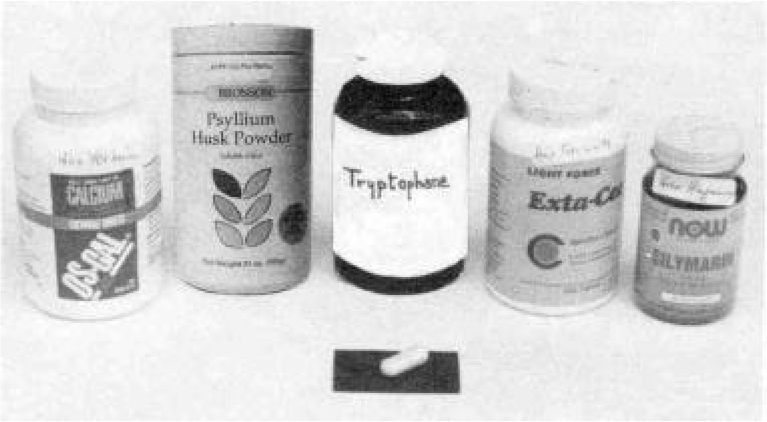

The capsule in the foreground is a notorious tryptophane capsule. It had the following pollutants: PCBs, mercury, ruthenium, thulium, strontium, praseodymium, aluminum, benzalkonium.

Fig. 74 Some polluted supplements.

Most of the varieties of vitamin C that I have tested are polluted with thulium! Until all vitamins and minerals and other food supplements have been analyzed for pollutants, after they are encapsulated or tableted, they are not safe. We need more disclosure on our products. No manufactured product is pure. We can't expect that. But at least we should be able to tell what impurities we are getting, and how much.

It is possible to do detailed analysis of foods or products at a reasonable price. Look at the bottle of common table salt, sodium chloride, that is used by beginning chemistry students to do experiments. It must be thoroughly analyzed for them because minute impurities affect their results. (Those minute impurities, like lead, affect you, too.) Look at the label on the bottle in the picture. Even after all these tests, the cost of laboratory salt is only $2.80 per pound.[33]

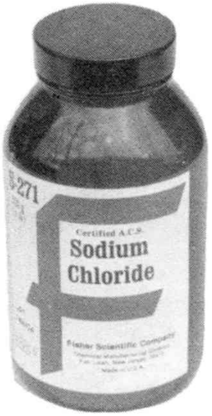

Fig. 75 Pure salt.

It is most important not to be fooled by ingredient claims, like “made from organically grown vegetables”. Sure that's great, but the analysis *I* trust would be done on the final, cleaned, cooked and packaged product on the shelf. The package is a major unlisted ingredient.

Toxic solvents like decane, hexane, carbon tetrachloride and benzene will get more flavor or fat or cholesterol out of things than metabolizable grain alcohol. Of course, the extraction process calls for washing out the solvent later. But it can't all be washed out, and a detailed analysis on the final product would give the public the information they need to make informed choices.

---

[33] You will pay about $8.00 per pound (Spectrum Chemical Co.) for USP (United States Pharmaceutical) grade. But the same analysis is done on the cheaper grades, and my point is that the analysis is cost effective enough that it should be done on our daily foods.

All supplements must be tested for purity by yourself. If this cannot be done,
don't take them.

Polluted supplements do much more harm than good. Get your super-nutrition by juicing vegetables of all kinds and making herbal teas.

Safe Supplements

There are, no doubt, lots of safe supplements to be had. The problem is knowing which they are. The nature of pollution is such that one bottle might be safe, while another of the same brand is not. In view of this, as I found a polluted bottle, I stopped using any more of that brand. That is why I am reduced to recommending only the ones in the *Sources* at this time.

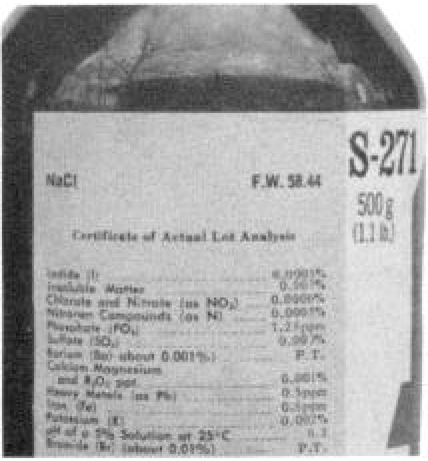

Common salt for student use is thoroughly analyzed for pollution. The label gives you the final "Actual Lot Analysis" of the product. It is not expensive.

Fig. 76 Laboratory salt label.

Vitamin C, in crystal form, is a must in your lifestyle. It helps the liver, and possibly other organs, detoxify things. It also helps retard mold, and perhaps even destroys aflatoxins! Keep some next to your refrigerator so it is handy when you put away groceries. Add 1/8 tsp. to maple syrup, vinegar, cooked cereal, fruit juice, leftovers. Have ¼ tsp. (1 gram) with each meal.

Vitamin B_2, 100 mg. This is the vitamin that helps detoxify benzene! Take one to three a day. While recovering from AIDS you need 3 tablets three times a day.

Vitamin B_6, 250 mg, and B-complex, 50 mg. Undoubtedly these help the liver and kidneys in many ways. One a day.

Magnesium oxide, 300 mg., is another must. Take one or two a day. It is a major mineral; all of our cells need lots of it. Only leafy vegetables provide it.

Hydrogen peroxide, food grade. It is advantageous to kill bacteria and viruses to some extent every day. Hydrogen peroxide lets you do this. It should never come in contact with metal, including its container or metal tooth fillings. If you get a few drops on your skin it may turn white and sting, but does no harm, so simply wash it off. Instructions for its use come with the product.

Herbs. These are excellent supplements, both in bulk and capsules, but not extracts, concentrates, or concoctions. There are many books that describe their uses.

Thioctic acid or lipoic acid. Presumably this chelates (traps and prepares for elimination) heavy metals, and helps the liver in detoxifying obscure and deadly poisons. Everyone would benefit from 100 mg. per day. I find it outstanding, and give it to many ill persons, even when not mentioned in the case histories. It comes as a 100 mg. capsule (see *Sources*). I use it at doses from one capsule, 3 times a day, to 2 capsules, 5 times a day. I have seen no side effects at these dosages, even in very sick persons.

Lugol's Iodine Solution (see *Recipes*) is old-fashioned "iodine." Iodine has a distinctive trait: it hangs up on anything and everything. In fact, it attaches itself so quickly we consider everything it touches as "stained." This is just the property we want to make it safe for use. The amount you use is immediately hung up, or attached, to your mucous and can not be quickly absorbed into the blood or other organs. It stays in the stomach. And for this reason it is so useful for killing vicious bacteria like *Salmonella*.

Do not take Lugol's iodine if you know you are allergic to iodine. It could be fatal.

Six drops of Lugol's solution can end it all for *Salmonella*. If you have gas and bloating, pour yourself ½glass of water. Add 6 drops of Lugol's (not more, not less), stir with wood or plastic, and drink all at once. The action is noticeable in an hour. Take this dose 4 times a day, after meals and at bedtime, for 3 days in a row, then as needed. This eradicates even a stubborn case of *Salmonella*.

Notice how calming 6 drops of Lugol's can be, soothing a manic stage and bringing a peaceful state where anxiety ruled before.

Lugol's is perfectly safe (if not allergic) to take day after day, when needed, because of its peculiar attaching property. It arrives in the stomach, reattaches to everything in proximity. Doomed are all *Salmonellas*; doomed also are eggs of parasites that might be in the stomach (cysts).

Naturally, one would not leave such medicine within the reach of children. Also, one would not use anything medicinal, including Lugol's unless there were a need, like cancer, AIDS, or bowel disease. When the gas and bloating problem has stopped, stop using Lugol's. If one or two doses of Lugol's cures the problem, stop. Store it in a perfectly secure place. In the past, 2/3 of a teaspoon (60 drops) of Lugol's was the standard dose of iodine given to persons with thyroid disease. Six drops is small by comparison.

Turmeric and **fennel** are herbs also used as cooking spices. They can eradicate invasive *E. coli* and *Shigella* bacteria! They are completely harmless, and are part of the Bowel Program.

Other supplements. The concept of supplementing the diet is excellent, but the pollution problem makes it prohibitive. Use only supplements and brands recommended in *Sources*, although the best approach is to test them yourself with your Syncrometer. I can't guarantee the brands in *Sources* will stay pure. If in doubt, leave it out.

Home Clean-up

This is the easiest task because it mostly involves throwing things out. Hopefully your family and friends will jump to your assistance.

- The basement gets cleaned.
- The garage gets cleaned.
- Every room in the house gets cleaned.

Your Basement

To clean your basement, remove all paint, varnish, thinners, brush cleaners, and related supplies. Remove all cleaners such as carpet cleaner, leather cleaner, rust remover. Remove all chemicals that are in cans, bottles or buckets.

You may keep your laundry supplies: borax, washing soda, white distilled vinegar, bleach and homemade soap. You may keep canned goods, tools, items that are not chemicals. You may move your chemicals into your garage. Also move any car tires and automotive supplies like waxes, oil, transmission fluid, and the spare gas can (even if it is empty) into your garage or discard them.

Seal cracks in the basement and around pipes where they come through the wall with black plastic roofing cement. In a few days it will be hard enough to caulk with a prettier color. Spread a sheet of plastic over the sewer or sump pump.

Your Garage

Do you have a garage that is a separate building from your home? This is the best arrangement. You can move all the basement chemicals into this garage. Things that will freeze, such as latex paint, you may as well discard. But if your garage is attached, you have a problem. Never, never use your door between the garage and house. Walk around the outside. Don't allow this door to be used. Tack a sheet of plastic over it to slow down the rate of fume entrance into the house. Your house acts like a chimney for the garage. Your house is taller and warmer than the garage so garage-air is pulled in and up as the warm air in the house rises. See the drawing.

In medieval days, the barn for the animals was attached to the house. We think such an arrangement with its penetrating odors is unsavory. But what of the gasoline and motor fumes we are getting now due to parked vehicles? These are toxic besides! This is even more medieval.

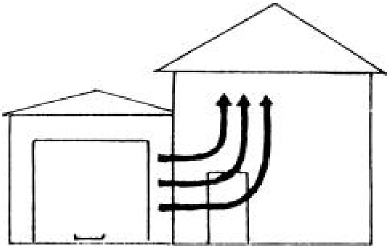

Fig. 77 Garage fumes.

If your garage is under your house, you cannot keep the pollution from entering your home. In this case, leave the cars and lawnmower outside. Remove cans of gasoline, solvents, etc. Put up a separate shed for these items.

Special Clean-up for Freon (CFCs)

Because I consider Freon to be the top health hazard in our home, I recommend turning in your refrigerator for a non-CFC variety. Dispose of all spare units. Remove window air conditioners or test the dust in your home (page 485) for Freon. Have your car's air conditioning system checked. Dispose of old pressurized cans. Even one whiff is too much. It never leaves the body because the body has no detoxification methods for it!

Only one useful reaction with Freon comes to mind. Freon is thought to be responsible for the ozone “hole” at the South Pole. Would Freon react with ozone supplied to your body and thereby become biodegradable? Indeed, it does! But only if you drink it as ozonated water. Other ozone routes, as intravenous or rectal, have not been observed to be as effective.

If you are following your progress with the Syncrometer, you will see that Freon now appears in the liver for the first time. (Before this, it was marooned in the parathyroids, thymus, and other organs.) You may also detect a feeling like indigestion. You must come to the assistance of your liver. Even ozonated Freon is extremely burdensome to the liver.

A combination of herbs (Liver Herb Drink in *Recipes*, page 552) rescues the liver from its plight, and prevents the indigestion. After drinking liver herbs you will see that the Freon now appears in the kidneys. Yet it is marooned there unless you assist them. Take the kidney cleanse to assist the kidneys so they can finally excrete the Freon into the urine.

It's an elaborate detoxifying program and usually takes six to eight weeks to get most of the Freon out. Afterward, continue the programs at one fourth dosage for half a year.

Forane is one of the new refrigerants. Although toxic, at least I observe it in the liver directly, suggesting that your body is capable of handling it. Remember your new refrigerator will still be using a toxic coolant, and it would be best to keep it outside or at least vented to the outside.

Special Clean-up for Fiberglass

Fiberglass insulation has microscopically small bits of glass that are free to blow into the air. When house drafts pull it into the air you will inhale them. They cut their way through your lungs and organs like millions of tiny knives, spreading through your body, since there is no way out for them. You smell nothing and feel nothing. This makes it a very sinister poison. Your body, though, recognizes these sharp, pointed bits and tries to stop their spread by sequestering them in cysts.

Most solid malignant tumors contain fiberglass or asbestos, another glass-like particle. In nearly all cases a hole can be found in the ceiling or walls, leading to fiberglass insulated parts of the house. When these holes are sealed in an air-tight manner the house air no longer is positive for fiberglass. Covering with paneling is not sufficient. Check your dwelling for uncovered fiberglass. Repair immediately. Search for small screw holes intended for pictures, or electric outlet plates that are missing.

Also remove fiberglass jackets from water heater and fiberglass filter from furnace. Replace with foam or carbon. Best of all, hire a crew to remove it all from your home, and replace insulation with blown-in shredded paper or other innocuous substance.

Never build a new house using fiberglass for any purpose.

Special Clean-up for Asbestos

The biggest source of asbestos is not building materials! It is the clothes dryer belt and hair dryer! To be safe, remove the belt from your dryer and check to see if it says “Made in USA” on the belt itself. If so, it is OK. If not, it is imported, and probably contains asbestos. Exchange it for a USA belt (see *Sources*).

Hair dryers, too, may be imported and shed asbestos. It is especially hazardous to be aiming a stream of hot asbestos right at your face! If you can't find a safe model (see *Sources*), or are unsure, don't use any. If you have cancer or are ill, no one in the house should use an unsafe hair dryer.

Turn off radiators and electric heaters and cover them with big plastic garbage bags, or paint them, or remove them. They give off asbestos if their paint is old.

Your House

To clean the house, start with the bedroom. Remove everything that has any smell to it whatever: candles, potpourri, soaps, mending glue, cleaners, repair chemicals, felt markers, colognes, perfumes, and especially plug-in air “fresheners”. Store them in the garage, not the basement. Since all vapor rises, they would come back up if you put them in a downstairs garage or basement.

Do not sleep in a bedroom that is paneled or has wallpaper. They give off arsenic and formaldehyde. Either remove them or move your bed to a different room. Leave the house while this is being done. If other rooms have paneling or wallpaper, close their doors and spend no time in them.

Next clean the kitchen. Take all cans and bottles of chemicals out from under the sink or in a closet. Remove them to the garage. Keep only the borax, washing soda, white distilled vinegar and homemade soap. Use these for all purposes. For exact amounts to use for dishwasher, dishes, windows, dusting, see *Recipes*. Remove all cans, bottles, roach and ant killer, moth balls, and chemicals that kill insects or mice. These should not be stored anywhere. They should be thrown out. Remember to check the crawl space, attic and closets for hidden poisons also.

To keep out mice, walk all around your house, stuffing holes and cracks with steel wool. Use old-fashioned mouse traps. For cockroaches and other insects (except ants) sprinkle handfuls of boric acid[34] (not borax) under your shelf paper, behind sink, stove, refrigerator, under carpets, etc. Use vinegar on your kitchen wipe-up cloth to leave a residue that keeps out ants. Do this regularly. To wax the floor, get the wax from the garage and put it back there. A sick person should not be in the house while house cleaning or floor waxing is being done.

Remove all cans and bottles of "stuff" from the bathroom. The chlorine bleach is stored in the garage. Someone else can bring it in to clean the toilet (only). Leave only the borax soap, homemade soap, and grain alcohol antiseptic. Toilet paper and tissues should be unfragranced, uncolored. All colognes, after shave, anything you can smell must be removed. Family members should buy unfragranced products. They should smoke outdoors, blow-dry their hair outdoors or in the garage, use nail polish and polish remover outdoors or in the garage.

Do not keep new foam furniture in the house. If it is less than one year old, move it into the garage until you are well. It gives off formaldehyde. So does new clothing; it is in the sizing. Wash all new clothes before wearing. If you have a respiratory illness, move all the clothes in the clothes closet out of your bedroom to a different closet.

Do not use the hot water from an electric hot water heater for cooking or drinking. It has tungsten. Do not drink water that sits in glazed crock ware (the glaze seeps toxic elements like cadmium) like some water dispensers have. Do not buy water from your health food store that runs through a long plastic hose from their bulk tank (I always see cesium picked up from flexible clear plastic). Also ask them how and when they clean their tank. Best is to observe that it is done with non-toxic methods.

---

[34] Boric acid is available by the pound from farm supply stores and from Now Foods. Because it looks like sugar keep it in the garage to prevent accidental poisoning.

If your house is more than 10 years old, change all the galvanized pipe to PVC plastic. Although PVC is a toxic substance, amazingly, the water is free of PVC in three weeks! If your house has copper pipes don't wait for cancer or schizophrenia to claim a family member. Change all the copper pipe to PVC plastic immediately. If the pipes are not accessible, ask a plumber to lay an extra line, outside the walls. This is less expensive, too.

If you have a water softener, by-pass it immediately and replace the metal pipe on the user side of the softener tank. Softener salts are polluted with strontium and chromate; they are also full of aluminum. The salts corrode the pipes so the pipes begin to seep cadmium into the water. After changing your pipes to plastic, there will be so little iron and hardness left, you may not need a softener. If the water comes from a well, consider changing the well-pipe to PVC to get rid of iron. While the well is open, have the pump checked for PCBs. Call the Health Department to arrange the testing. If you must have softening after all this, check into the new magnetic varieties of water softener (although they only work well when used with plastic plumbing).

The cleanest heat is electric. Go total electric if possible. If you must stay with gas, have a furnace repair person check your furnace and look for gas leaks before the heating season starts. Don't call the gas company even though it is free. The gas company misses 4 out of 5 leaks! The Health Department does not miss any; call them! House builders and contractors are also reliable in their gas leak detection.

Unnatural Chemicals

And where I found them...

ARSENIC

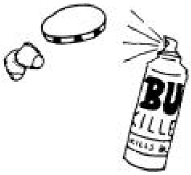

in ant & roach hives, grains of pesticide

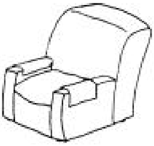

in carpet & furniture "treated" for stain resistance

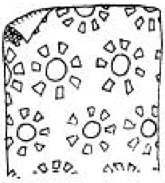

in wallpaper

BARIUM

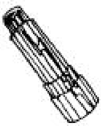

in lipstick

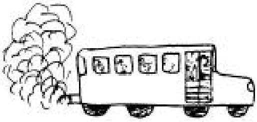

in bus exhaust

MOLYBDENU

M

in "molys"

COBALT

NEW!
WITH
MORE
BLUE
GRAINS

in laundry detergent

in dishwater detergent

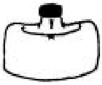

in skin bracer

in mouthwash

ANTIMONY

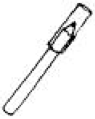

in eye liner

CADMIUM and COPPER

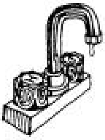

in water running through old metal pipes

TITANIUM

in face powder & other powders, and in metal dental ware

PCB's

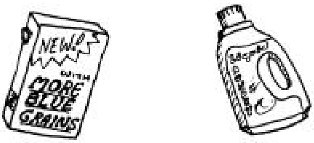

in regular and
health store detergents

LEAD

in men's hair
color restorer

in solder at joints
of copper pipes

CHROMIUM

in eyebrow
pencil

in water
softener salts

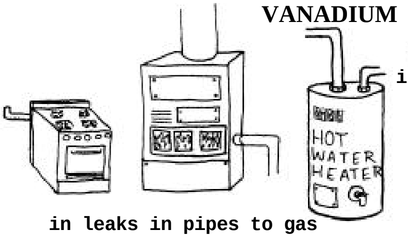

in leaks in pipes to gas
stove, furnace, water heater

VANADIUM

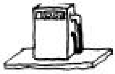

in diesel fuel

in candles
(even when
they're not
burning)

NICKEL

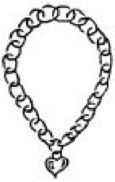

in metal jewelry
worn on the skin

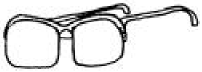

in metal glasses frames

in metal
watch bands

in metal tooth fillings
and retainers

FIBERGLASS

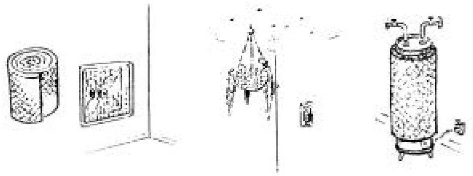

from insulation behind holes in ceiling or uncovered outlets, water heater jackets, stuffed around fans and air conditioners, insulation

CFC's (FREON)

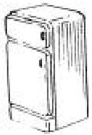

in
refrigerators

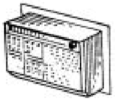

in air
conditioners

in spray
cans

MERCURY and THALLIUM

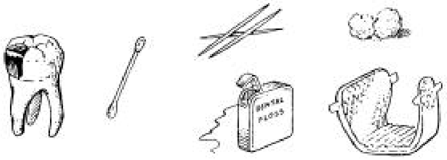
DENTAL FLOSS

in tooth fillings, sanitary napkins, cotton balls, dental floss, toothpicks, cotton swabs

THULIUM

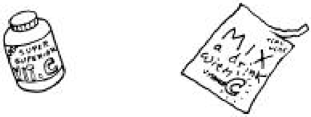
SUPER SUPERIOR VIT. C
MIX a drink with VIT. C

in most brands of vitamin C I tested

DYSPROSIUM and LUTETIUM

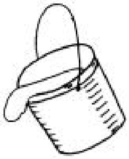

in paint, varnish, shellac

HOLMIUM

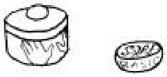

in hand cleaners

HAFNIUM

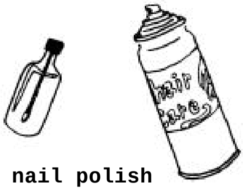

in nail polish & hair spray

RHENIUM

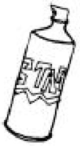

in spray starch

BISMUTH

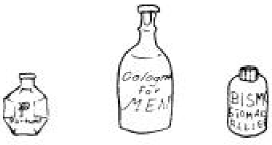
Cologne for MEN
BISM STOMACH RELIEF

in cologne and stomach aids

CESIUM

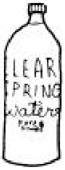
CLEAR SPRING water

in clear-as-glass plastic

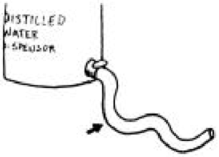
DISTILLED WATER DISPENSOR

after running through long plastic hose

TIN and STRONTIUM

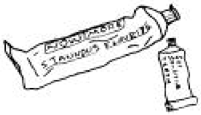

in toothpaste

YTTERBIUM, ERBIUM, TERBIUM

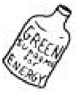
GREEN
ENERGY

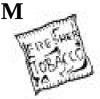
FRESHER
TOBACCO

in foil packaging

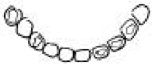

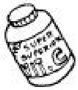

in plastic tooth fillings

in vitamins and supplements

PRASEODYMIUM, NIOBIUM, NEODYMIUM, YTTRIUM

in over the counter drugs

in prescription drugs

RADON, URANIUM, THORIUM

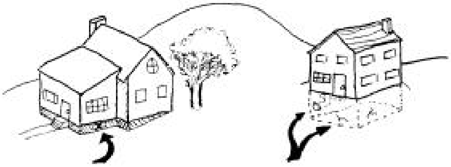

in crawlspace without open vents

in holes and cracks in basement

LANTHANUM

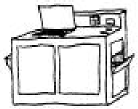

in duplicator and copier ink

BENZALKONIUM and ZIRCONIUM

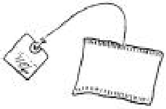

in tea bags

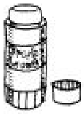

in deodorant

STANNOUS FLUORIDE

in toothpaste

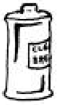

in mouthwash

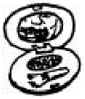

in cosmetics

ALUMINUM

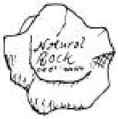
Natural
Rock

in "natural" deodorant advertised as aluminum-free

in salt

in cans

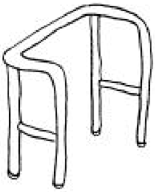

in walkers for the elderly

SKIN
SOFTENER

in lotions

TUNGSTEN

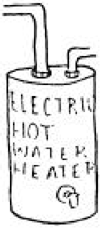
ELECTRIC
HOT
WATER
HEATER

corroded rod in
electric water heaters

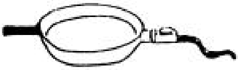

in electric
frying pans

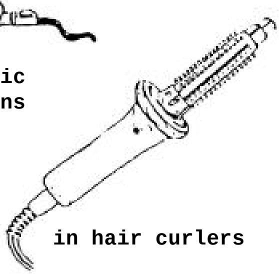

in hair curlers

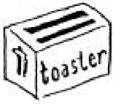

in toasters

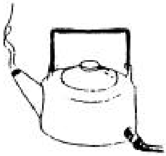

in tea kettles

BERYLLIUM

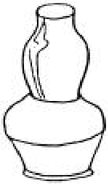

in hurricane
lamps

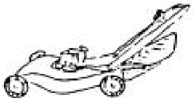

in lawn mowers

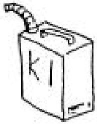

in kerosene

FORMALDEHYDE

in foam
mattresses

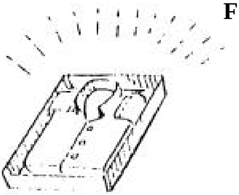

in new
clothing

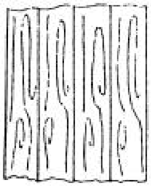

in paneling

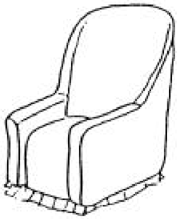

in foam chairs

ASBESTOS

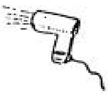

in hair blowers

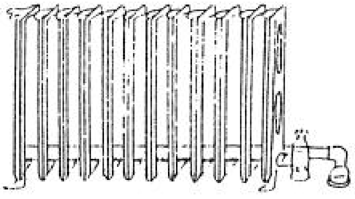

in radiator paint

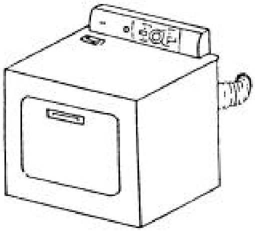

in clothes dryer belts

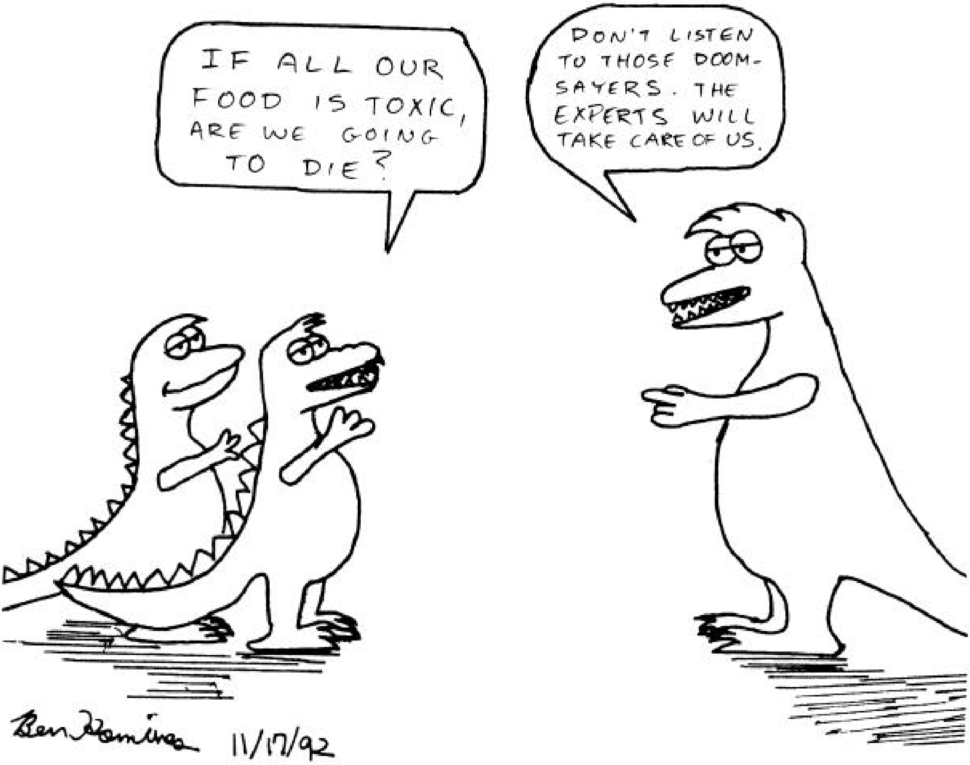
IF ALL OUR FOOD IS TOXIC, ARE WE GOING TO DIE?
DON'T LISTEN TO THOSE DOOM-SAYERS. THE EXPERTS WILL TAKE CARE OF US.
11/17/92

Fig. 78 Our future, unless we act.

Bioelectronics

The most important electronic device to make or buy is a **zapper**—a compact pulse generator operating from a common 9 volt battery whose output is about 30 KHz. It kills all parasites, bacteria, viruses, molds, and fungi even though their individual resonant frequencies are either higher or lower (50 KHz to 900 KHz). Building a zapper is described in an earlier chapter.

The next most useful device to have is a **Syncrometer.™** It lets you diagnose yourself and monitor your progress until you are cured. It consists of an audio oscillator circuit with your body as part of the circuit. Utilizing samples of parasites or pollutants, it lets you test for them in any product or body tissue. I include a design that you can make yourself.

A third very useful device is a **frequency generator**. You can use it to electrocute individual organisms, or together with the Syncrometer™ to find an organism's particular frequency. You need one that operates in the parasite, bacteria and virus ranges, from 50 KHz to 900 KHz. It must also be able to select a particular frequency, like 434 KHz, quickly and accurately. Frequency generators are available for as little as $300.00, but it is worth paying a little more to get a digital display of the frequency.

Making A Syncrometer

This is an *audio oscillator* circuit in which you include yourself by means of a handhold. You listen to the current in your circuit with a loudspeaker. Other oscillator circuits will work, too. A lot of fascinating opportunities present themselves with this concept.

I have previously published three ways to build the Syncrometer circuit.[35] Here is the circuit diagram:

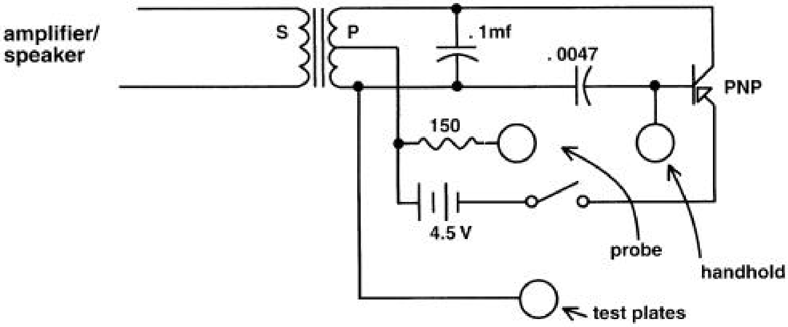
amplifier/
speaker
S
P
. 1mf
. 0047
PNP
150
4.5 V
probe
handhold
test plates

4700pf moly cap
.1 mfd 16v cap
2N2907 PNP
min output transf 8-900Ù
150 .25w res
3 AA batteries
speaker (like Radio Shack 277-1008C)

Fig. 79 Syncrometer schematic.

If you are not an electronics enthusiast, you can still assemble a Syncrometer using a hobby kit. No soldering is required. Here is what you need:

Making a Hobby Kit Syncrometer

Item	**Radio Shack Cat. No.**
200 in One Electronic Project Lab by Science Fair	28-262
3 AA 1½ volt batteries	

---

[35] In *The Cure for all Cancers* and *The Cure for HIV/AIDS*.

Alligator clip test jumpers	You need 2.
Handhold. A four inch length of ¾ inch copper pipe, like for plumbing. These dimensions are critical to assure maximum skin contact.	
Probe. A banana plug.	Precision Mini-Hook Test Lead Set (contains two, you only need one) 278-1160A
Pencil, new.	

Fig. 80 Syncrometer parts list.

Build The Electrosonic Human in the *200 in One Electronic Project Lab*. It takes about 10 minutes.

Later, when you use the probe to press against your knuckle you may find getting the right sound is painful. In this case try substituting the .005 microfarad capacitor for the .01 microfarad capacitor in the circuit.

Attach the Probe. The Archer Precision Mini-Hook Test Lead Set has a banana plug for the probe on one end and a mini-hook on the other end for easy attachment to the circuit. Tape a long, new pencil to the probe to make it easier to hold. Connect the Probe to middle post of the primary side of the transformer (it also connects to the negative battery post). You will not be using the two connections the instructions tell you to hold.

Attach the Handhold. Clip the Handhold to one end of an alligator clip test jumper, and clip the other end to the base (B) of the transistor used in the circuit.

Attach an alligator clip to the post of the transformer that connects to the two capacitors. This will go to the **test plates**.

Final test. Turn the control knob on and keep turning the potentiometer to nearly the maximum. (This reduces the resistance. The schematic has a 150 ohm fixed resistor in place of the potentiometer.) Make sure you have good batteries installed. Test the circuit by briefly touching the Probe to the Handhold. The speaker should produce a sound like popping corn. If it does not, check that your alligator clips are not bending the spring terminals so much that other wires attached there are loose. Finally, turn switches OFF.

Making Test Plates

This is the box you attach to the basic Syncrometer circuit. It has test plates to put your test substances and tissue samples on. The wiring in it is arranged so that you can test for a toxin in a product, as well as search in yourself. This means you can search for *Salmonella* in the milk or cheese you just ate, not just for *Salmonella* in your stomach.

Only if the resonant frequency of an item on one plate is equal to the resonant frequency of an item on the other plate will the entire circuit oscillate or resonate! This implies the two plates have something in common. By putting a known pure sample on one plate you can reliably conclude the other sample contains it if the circuit resonates.

You may build a test plate box into a cardboard box (such as a facial tissue box) or a plastic box. Here are the instructions for the cardboard box model.

Test Plates Parts List

- Stiff paper.
- Aluminum foil.
- A facial tissue box is easiest. A plastic project box, about 7” x 4” x 1½” makes a more durable product, but requires a drill, and you should discard any metal lid it comes with.

- 3 bolts (tapered heads) about 1 inch long, 1/8 inch diameter and 6 washers and nuts to fit.
- toggle switch with OFF-ON positions.
- Alligator clip test leads.

Test Plates Assembly

Cut two 3-1/2 inch squares out of stiff paper such as a milk carton. Cover them with 4½ inch squares of aluminum foil, smoothed evenly and tucked snugly under the edges. You have just made yourself a set of open capacitors. Turn the box upside down and draw squares where you will mount them at the ends of the box. Don't actually mount them, to save wear and tear on them, until the rest of the box is complete.

Mount the ON OFF switch on the front of the box, underneath the right hand plate. Line it up so ON is downward and OFF is up. (An electronics shop can determine this for you at the time of purchase.) Label the box with ON and OFF signs.

Two bolts will be reserved for the plates. The third bolt is used as a terminal where the current from the oscillator circuit will arrive. Make a hole on the side of the box, near the left hand plate and mount the bolt so it sticks half way inside and halfway outside the box. It does not matter whether the head is inside or outside. Tighten it there with a nut on each side of the box. Label it TERMINAL. It merely means connecting place.

Mark the center of each square that you drew and each capacitor you built. Pierce first with a pin; follow with a pencil until a round hole is made at the center. Mount each plate with a bolt, fastening it below with a nut. Washers are optional.

The left side connection (terminal) gets attached to the left plate (bolt) with an alligator clip. Use another clip to attach the same left plate (bolt) to the ON OFF switch (there are two connections, use either one). Finally attach the ON OFF switch connection you didn't use to the right plate (bolt). Make sure the connections at the switch are not touching each other; you might tape them to guard against this.

All these connections should be checked carefully to make sure they are not touching others accidentally. But if you leave the box open so you can see any problems and use clear tape around connections to prevent accidental touching to the wrong connection, it should work OK.

Finally, trace your current. It comes in from the Syncrometer at the main terminal on the left. It is brought to the left plate. When the switch is ON it is simultaneously brought to the right plate. Notice that the plates are not connected to anything else. They are simply capacitors, letting current in and out momentarily and at a rate that is set by the frequency of the oscillator circuit, about 1,000 hertz. This frequency goes up as the resistance (of the circuit or your body) goes down.

The probe and handhold allow you to include yourself in the Syncrometer circuit. You grasp these when testing. This makes you part of the circuit.

The speaker lets you "listen" to the current. As resistance drops, current goes higher and frequency goes up. As frequencies go higher in the circuit, pitch goes higher. You will be comparing the sound of a standard "control" current with a test current.

Using The Syncrometer

Fill a saucer with cold filtered tap water. Fold a paper towel four times and place it in this dish. It should be entirely wet.

Cut paper strips about 1 inch wide from a piece of white, unfragranced, paper towel. Dampen a paper strip on the towel and wind it around the copper pipe handhold to completely cover it. The wetness improves conductivity and the paper towel keeps the metal off your skin.

- Start with the test plate switch at OFF.
- Turn the control knob (potentiometer) on, and to near maximum.
- Touch each plate with the probe, while holding the copper pipe with one hand. Only the left plate should give you a sound from the speaker. Turn the test plate switch ON. Now both plates should give you a sound when the probe touches them.
- Turn the test plate switch OFF again.
- Pick up the handhold, squeeze it free of excess water.
- Pick up the probe in the same hand, holding it like a pen, between thumb and forefinger.

Dampen your other hand by making a fist and dunking your knuckles into the wet paper towel in the saucer. You will be using the area on top of the first knuckle of the middle finger or forefinger to learn the technique. Become proficient with both. Immediately after dunking your knuckles dry them on a paper towel folded in quarters and placed beside the saucer. The degree of dampness of your skin affects the resistance in the circuit and is a very important variable that you must learn to keep constant. Make your probe as soon as your knuckles have been dried (within two seconds) since they begin to air dry further immediately.

With the handhold and probe both in one hand press the probe against the knuckle of the other hand, keeping the knuckles bent. Press lightly at first, then harder, taking one half second. Repeat a half second later, with the second half of the probe at the same location. There is an additive effect and you get two chances to listen to the current. All of this takes less than two seconds. Don't linger because your body will change and your next probe will be affected.

Subsequent probes are made in exactly the same way. As you develop skill, your probes will become identical. Plan to practice for one or two hours each day. It takes most people at least twelve hours of practice in order to be so consistent with their probes that they can hear the slight difference when the circuit is resonant.

For reference you may wish to use a piano. The starting sound when you touch down on the skin should be F, an octave and a half above middle C. The sound rises to a C as you press to the knuckle bone, then slips back to B, then back up to C-sharp as you complete the second half of your first probe. If you have a multitester you can connect it in series with the handhold or probe: the current should rise to about 50 microamps. If you have a frequency counter the frequency should reach 1000 Hz. You should arrive at C-sharp just before the probe becomes painful.

Two things change the sound of the probes even when your technique doesn't change.

1. The patch of skin chosen for probing will change its properties. The more it is used, the redder it gets and the higher the sound goes when you probe. Move to a nearby location, such as the edge of the patch, when the sound is too high to begin with, rather than adjusting the potentiometer.
2. Your body has cycles which make the sound go noticeably higher and lower. If you are getting strangely higher sounds for identical probes, stop and only probe every five minutes until you think the sound has gone down to standard. This could take five to twenty minutes. Learn this higher sound so you can avoid testing during this period.

You may also find times when it is impossible to reach the necessary sound without pressing so hard it causes pain. You may adjust the potentiometer if that helps.

All tests are momentary.

This means less than one second. It is tempting to hold the probe to your skin and just listen to the sound go up and down, but if you prolong the test you must let your body rest ten minutes, each time, before resuming probe practice!

For our purposes, it is not necessary to locate acupuncture points.

Resonance

The information you are seeking is whether or not there is *resonance*, or *feedback oscillation*, in the circuit. If there is the test is **YES** (positive). You hear resonance by comparing the second probe to the first. You can never hear resonance on the first probe, for reasons that are technical and beyond the scope of this book. You are not merely comparing pitch in the two probes. During resonance a higher pitch is reached faster; it seems to want to go infinitely high.

Remember that more electricity flows, and the pitch gets higher, as your skin reddens or your body changes cycle. These effects are not resonance.

Resonance is a small extra hum at the high end of the probe. As soon as you hear it, stop probing. Your body needs a short recovery time (10 to 20 seconds) after every resonant probe. The longer the resonant probe, the longer the recovery time to reach the standard level again.

Using musical notes, here is a NO (negative result): F-C-B-C# (first probe) F-C-B-C# (compare, it is the same sound). Here is a YES (positive) result: F-C-B-C# (first probe) F-D (stop quickly because you heard resonance). In between the first and second probe a test substance will be switched in as described in lessons below.

It is not possible to produce a resonant sound by pressing harder on the skin, although you can make the pitch go higher. To avoid confusion it is important to practice making probes of the same pressure. (Practice getting the F-C-B-C# tune.)

Making Pure Water for Testing Purposes

Since the water you purchase is likely to have solvents in it and since your tap water may be polluted with heavy metals and since your (or a store's) filtration system may be clogged, it is important to make your own pure water.

Purchase a “filter pitcher” made of hard, opaque plastic, not the clear or flexible variety (see Sources). Fill the pitcher with cold tap water, only, not reverse osmosis, distilled, or any other water, since solvents do not filter out as easily as heavy metals. The filter should be made of carbon only. To make test substances, use fresh water in the pitcher and pour.

If your water has lead, copper or cadmium from corroded plumbing, the filter will clog in five days of normal use. So use this pitcher sparingly, just for making test substances and for operating the Syncrometer.

Lesson One

Purpose: To identify the sound of resonance in the circuit.

Materials: Potentized (homeopathic) solutions. Prepare these as follows: find three medium-sized vitamin bottles, glass or plastic, with non-metal lids. Remove any shreds of paper sticking to the rim. Rinse well with cold tap water. Then rinse again with filtered water.

Pour filtered water into the first bottle to a depth of about ½ inch. Add about 50 little grains of table salt using the tip of a plastic knife. This is a “pinch.” Replace the lid. Make sure the outside is clean. If not, rinse and dry. Now shake hard, holding it snugly in your hand. Count your shakes; shake 120 to 150 times. Use elbow motion so each shake covers about an eight inch distance. Shaken samples are different from unshaken ones, that's why this is so important. When done label the bottle on its side and lid: SALT #1. Wash your hands (without soap).

Next, pour about the same amount of filtered water into the second and third bottles. Open SALT #1 and pour a small amount, like 1/4 to 1/2 of a teaspoon (do not use a spoon) into the second and third bottles. Close all bottles. Now shake the second bottle the same as the first. Clean it and label it SALT #2. Do the same for the third bottle. Label it SALT #2 also and set aside for Lesson Four.

These two solutions have unique properties. SALT #1 always resonates. Use #1 to train your ear. SALT #2 shouldn't resonate. Use #2 to hear when you (your body's internal resistance) have returned to the standard level.

1. Turn the Syncrometer ON.
2. Place the SALT #2 bottle on the right test plate.
3. Start with the plate switch OFF.
4. Make your first probe (F-C-B-C#).
5. Flip the plate switch ON, taking only one half second. Brace your hand when switching so it is a fast, smooth, operation.
6. Make the second probe (F-C-B-C#). Total probe time is 2½ seconds. Count it out, "a thousand and one (done with first probe) a thou. (done with switching) a thousand and one (done with second probe)."
7. The result should be a NO (negative). If the second probe sounds even a little higher you are not at the standard level. Wait a few more seconds and go back to step 3.
8. If the first result was NO, remove SALT #2 and put SALT #1 on. Put the test plate switch back to OFF and repeat the test. This time the circuit was resonating. Learn to hear the

difference between the last two probes so that a resonant probe can be terminated early (reducing rest time).

9. The skin must now be rested. When SALT #1 is placed in the circuit there is always resonance whether you hear it or not. Therefore, always take the time to rest the skin.
10. How can you be sure that the skin is rested enough? Any time you want to know whether you have returned to the standard level, you may simply test yourself to SALT #2 (just do steps 3 through 6). While you are learning, let your piano also help you to learn the standard level (starts exactly at F). If you do not rest and you resonate the circuit before returning to the standard level, the results will become aberrant and useless. The briefer you keep the resonant probe, the faster you return to the standard level. Don't exceed one half second when probing SALT #1. Hopefully you will soon hear resonance within that time.

This lesson teaches you to first listen to the empty plate, then to SALT #2, to check for standard state. Then to compare the empty plate to SALT #1 to check for resonance. In later lessons we assume you checked for your standard level or are quite sure of it.

Practice hearing resonance in your circuit every day.

White Blood Cells

Checking for resonance between your white blood cells and a toxin is the single most important test you can make.

Your white blood cells are your immune system's first line of defense. In addition to making antibodies, interferon, interleukins, and other attack chemicals, they also “eat” foreign substances in your body and eliminate them. By simply checking your white blood cells for toxins or intruders you save having to check every other tissue in your body. Because no matter where the foreign substance is, chances are some white blood cells are working to remove it.

It took me two years to find this ideal indicator, but it is not perfect. **Tapeworms are a notable exception.** They can be encysted in a particular tissue which will test positive, while the white blood cells continue to test negative. Also, when bacteria and viruses are in their latent form, they do not show up in the white blood cells. Fortunately, in their active form they show up quite nicely. **Freon** is an example of a toxin that is seldom found in the white blood cells; but typically, the white blood cells are excellent indicators of toxins.

Making a White Blood Cell Specimen

Obtain an empty vitamin bottle with a flat plastic lid and a roll of clear tape. The white blood cells are not going into the bottle, they are going on the bottle. The bottle simply makes them easy to handle. Rinse and dry the bottle. Make a second specimen on a clean glass slide if available. Squeeze an oil gland on your face or body to obtain a ribbon of whitish matter (not mixed with blood). Pick this up with the back of your thumb nail. Spread it in a single, small streak across the lid of the bottle or the center of the glass slide. Stick a strip of clear tape over the streak on the bottle cap so that the ends hang over the edge and you can easily see where the specimen was put (see photo). Wipe the lid beside the tape to make sure all white blood cells are covered. For the slide, apply a drop of balsam and a cover slip (see *Sources*). Both types of preparation will give you identical results. The bottle type of white blood cell specimen is used by standing it on its lid (upside down) so that the specimen is next to the plate. The lid is used because it is flat, whereas the bottom of most bottles is not.

Lesson Two

Purpose: To add a white blood cell specimen to the circuit and compare sound.

Method:

1. Turn the Syncrometer ON.
2. Start with test plate switch OFF.
3. Place the white blood cell specimen on the left plate. Place some junk food in a plastic baggy on the right plate.
4. Eat some of the junk food.
5. After ½minute listen to the current. Flip the plate switch ON and listen again.
6. If the circuit is now resonating, the junk food is already in your white blood cells. It is toxic.

Take vitamin C and a B-50 complex to clear it rapidly; it may have had propyl alcohol or benzene in it. Test every 5 minutes afterward to see how long it takes to clear out.

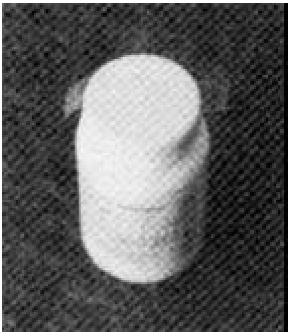

Fig. 81 Bottle with white blood cells taped on top.

Lesson Three

Purpose: To determine the purity of the filtered water you are making.

Method: Pour a few tsp. of filtered water into a bottle or plastic bag. Place your white blood cell specimen on one plate and the water sample on the other. Listen to your circuit. Taste your filtered water. After ½minute, listen to your circuit again, just as in Lesson Two. If it appears in your white blood cells at any time you can conclude the water is not pure. You must have pure water available to you before continuing.

Lesson Four

Purpose: To determine your percent accuracy in listening for resonance.

Materials: The SALT #1 and SALT #2 solutions you made for Lesson One.

Method: Move the SALT #1 and SALT #2 labels to the bottom of the bottles so you can not tell which bottle is which.

1. Turn the Syncrometer ON.
2. Start with the test switch OFF.
3. Mix the bottles up, select one at random, and place it on the right plate.
4. Listen to the current.
5. Flip the plate switch ON and make your second probe.
6. Resonance indicates a SALT #1, no resonance indicates SALT #2. Check the bottom. Remember to rest after the SALT #1, whether or not you heard resonance.
7. Repeat steps 3 through 5 a number of times. Work toward getting three out of three correct. Practice every day.

Trouble shooting:

a) If you repeat this experiment and you keep getting the same bottles "wrong", start over. You may have accidentally contaminated or mislabeled the outside of the bottle, or switched bottle caps.

b) If you get different bottles wrong each time, the plates may be contaminated. Wash the outside of the bottles and rinse with filtered water and dry. Wipe the plates very gently too, with filtered water and dry. Or replace the plates.

c) If all the bottles read the same, your filtered water is polluted. Change the filter.

Preparing Test Substances

It is possible to prepare dry substances for testing such as a piece of lead or grains of pesticide. They can simply be put in a plastic bag and placed on the test plate. However, I prefer to place a small amount (the size of a pea) of the substance into a ½ ounce bottle of filtered water. There will be many chemical reactions between the substance and the water to produce a number of test substances all contained in one bottle. This simulates the situation in the body.

Within the body, where salt and water are abundant, similar reactions may occur between elements and water. For example, a strip of pure (99.9% pure) copper placed in filtered water might yield copper hydroxide, cuprous oxide, cupric oxide, copper dioxide, and so forth. These may be similar to some of the reaction products one might expect in the body, coming from a copper IUD, copper bracelet or the copper from metal tooth fillings. Since the electronic properties of elemental copper are not the same as for copper compounds, we would miss many test results if we used only dry elemental copper as a test substance.

Impure Test Substances

It is not necessary to have pure test substances. For instance, a tire balancer made of lead can be easily obtained at an auto service station. Leaded gasoline and lead fishing weights also make good test substances for lead. There is a disadvantage, though, to using impure test substances. You are including the extra impurities in your test. If your lead object also has tin in it, you are also testing for tin. Usually, you can infer the truth by some careful maneuvering. If you have searched your kidneys for leaded gasoline, fishing weights and tire balancers and all 3 are resonant with your kidneys, you may infer that you have lead in your kidneys, since the common element in all 3 items is lead. (You will learn how to specify a tissue, such as your kidneys, later.)

Using pure chemicals gives you certainty in your results. You can purchase pure chemicals from chemical supply companies (see *Sources*). Your pharmacy, a child's chemistry set, a paint store, or biological supply company can also supply some.

The biggest repository of all toxic substances is the grocery store and your own home.

You can make test substances out of your hand soap, water softener salt, and laundry detergent by putting a small amount (1/16 tsp.) in a ½ounce glass bottle and adding about 2 tsp. filtered water. (Or for quick testing just put them dry or wet in a sealed plastic baggy.) Always use a plastic spoon.

Here are some suggestions for finding sources of toxic products to make your own toxic element test. If the product is a solid, place a small amount in a plastic bag and add a tablespoon of filtered water to get a temporary test product. For permanent use put it in a small amber glass bottle. If the product is a liquid, pour a few drops into a glass bottle and add about 2 tsp. filtered water. Keep all toxic substances in glass bottles for your own safety. Small amber glass dropper bottles can be purchased by the dozen at drug stores (also see *Sources*). Seal your test bottles with tape for safety and to prevent evaporation.

Aflatoxin: scrape the mold off an orange or piece of bread; wash hands afterward.

Acetone: paint supply store or pharmacy.

Arsenic: 1/16 tsp. of arsenate pesticide from a garden shop. A snippet of flypaper.

Aluminum: a piece of aluminum foil (not tin foil) or an aluminum measuring spoon.

Aluminum silicate: a bit of salt that has this free running agent in it.

Asbestos: a small piece of asbestos sheeting, an old furnace gasket, 1/4 inch of a clothes dryer belt that does not say "Made in USA", or a crumb of building material being removed due to its asbestos content (ask a contractor).

Barium: save a few drops from the beverage given clients scheduled for an X-ray. Lipstick that has barium listed in the ingredients.

Benzene: an old can of rubber cement (new supplies do not have it). A tsp. of asphalt crumbs from a driveway.

Beryllium: a piece of coal; a few drops of "coal oil" or lamp oil.

Bismuth: use a few drops of antacid with bismuth in it.

Bromine: bleached "brominated" flour.

Cadmium: scrape a bit off a galvanized nail, paint from a hobby store.

Cesium: scrape the surface of a clear plastic beverage bottle.

CFCs (freon): ask an electronics expert for a squirt from an old aerosol can that used freon as a cleaner. (Squirt into water, outdoors, put the water in a sample bottle.)

Chromate: scrape an old car bumper.

Cobalt: pick out the blue and green crumbs from detergent. A sample of cobalt containing paint should also suffice.

Chlorine: a few drops of pure, old fashioned Clorox.TM

Copper: ask your hardware clerk to cut a small fragment off a copper pipe of the purest variety or a ¼ inch of pure copper wire.

Ergot: a teaspoon of rye grains, or rye bread. Add grain alcohol to preserve.

Ether: automotive supply store (engine starting fluid).

Ethyl alcohol (grain alcohol): the purest "drinking" alcohol available. EverclearTM in the United States, ProtecTM (potable) in Mexico.

Fiberglass: snip a fragment from insulation.

Fluoride: ask a dentist for a small sample.

Formaldehyde: purchase 37% at a pharmacy. Use a few drops only for your sample.

Gasoline: gas station (leaded and unleaded).

Gold: ask a jeweler for a crumb of the purest gold available or use a wedding ring.

Kerosene: gas station.

Lead: wheel balancers from a gas station, weights used on fishing lines, lead solder from electronics shop.

Mercury: a mercury thermometer (there is no need to break it), piece of amalgam tooth filling.

Methanol: paint supply store (wood alcohol).

Nickel: a nickel plated paper clip, a washed coin.

Patulin (apple mold): cut a sliver of washed, bruised apple.

PCB: water from a quarry known to be polluted with it (a builder or electrical worker may know a source).

Platinum: ask a jeweler for a small specimen.

Propyl alcohol: rubbing alcohol from pharmacy (same as propanol or isopropanol). Use a few drops only, discard the rest. Do not save it.

PVC: glue that lists it in the ingredients (polyvinyl chloride).

Radon: leave a glass jar with an inch of filtered water in it standing open in a basement that tested positive to radon using a kit. After 3 days, close the jar. Pour about 2 tsp. of this water into your specimen bottle.

Silicon: a dab of silicon caulk.

Silver: ask a jeweler for a crumb of very pure silver. Silver solder can be found in electronics shops. Snip the edge of a very old silver coin.

Sorghum mold: 1/8 tsp. sorghum syrup.

Styrene: a chip of styrofoam.

Tantalum: purchase a tantalum drill bit from hardware store.

Tin: scrape a tin bucket at a farm supply. Tin solder. Ask a dentist for a piece of pure tin (used to make braces).

Titanium: purchase a titanium drill bit from a hardware store.

Toluene: a tube of glue that lists toluene as an ingredient.

Tungsten: the filament in a burned out light bulb.

Vanadium: hold a piece of dampened paper towel over a gas stove burner as it is turned on. Cut a bit of this paper into your specimen bottle and add 2 tsp. filtered water.

Xylene: paint store or pharmacy.

Zearalenone: combine leftover crumbs of three kinds of corn chips and three kinds of popcorn.

This list gets you off to a good start. Since few of these specimens are pure, there is a degree of logic that you must apply in most cases. If you are testing for barium in your breast, a positive result would mean that a barium-containing lipstick tests positive and a barium-free lipstick is negative.

A chemistry set for hobbyists is a wonderful addition to your collection of test specimens. Remember, however, the assumptions and errors in such a system. A test for silver using silver chloride might be negative. This does not mean there is no silver present in your body; it only means there is no silver chloride present in the tissue you tested.

You are bound to miss some toxins; don't let this discourage you. There is more than enough that you can find.

The most fruitful kind of testing is, probably, the use of household products themselves as test substances. The soaps, colognes, mouthwash, toothpaste, shampoo, cosmetics, breads, dairy products, juices and cereals can all be made into test specimens. Put about 1/8 tsp. of the product in a small glass bottle, add 2 tsp. filtered water and ¼ tsp. grain alcohol to preserve it. For temporary purposes use a plastic baggy and water only. If you test positive to your household products in your white blood cells you shouldn't use them, even if you can not identify the exact toxin.

For a list of toxins and solvents I use, see page 571. To order pure substances see *Sources* for “chemicals for testing.”

Making Organ Specimens

To test for toxic elements or parasites in a particular organ such as the liver or skin, you will need either a fresh or frozen sample of the organ or a prepared microscope slide of this organ. Meat purchased from a grocery store, fresh or frozen, provides you with a variety of organ specimens. Chicken, turkey, beef or pork organs all give the same results. You may purchase chicken gizzards for a sample of stomach, beef liver for liver, pork brains for brain, beef steak for muscle, throat sweet breads for thymus, tripe for stomach lining. Other organs may be ordered from a meat packing plant.

Trim the marrow out of a bone slice to get bone marrow. Scrub the bone slice with hot water to free it of marrow to get a bone specimen. Choose a single piece of meat sample, rinse it and place it in a plastic bag. You may freeze it. To make a durable unfrozen sample, cut a small piece, the size of a pea, and place it in an amber glass bottle (½oz.). Cover with two tsp. filtered water and ¼ tsp. of grain alcohol (pure vodka will do) to preserve it. These need not be refrigerated but if decay starts, make a fresh specimen.

Pork brains from the grocery store may be dissected to give you the different parts of the brain. Chicken livers often have an attached gallbladder or piece of bile duct, giving you that extra organ. Grocery store “lites” provides you with lung tissue. For kidney, snip a piece off pork or beef kidney. Beef liver may supply you with a blood sample, too. Always wash hands and rinse with grain alcohol after handling raw meat.

I use ½ oz amber glass bottles with bakelite caps (see Sources) to hold specimens. However, plastic bags or other containers would suffice. After closing, each bottle is sealed with a Parafilm™ strip to avoid accidental loosening of the cap. You may use masking tape.

To make a specimen of skin, use hangnail bits and skin peeled from a callous, not a wart. A few shreds will do. Remember, they must be very close to the test plate when in use; add 2 tsp. filtered water and ¼ tsp. grain alcohol.

Making a Complete Set of Tissue Samples

My original complete set was made from a frozen fish. As it thawed, different organs were cut away and small pieces placed in bottles for preserving in filtered water and grain alcohol. In this way, organs not available from the grocery store could be obtained. The piece of intestine closest to the anus corresponds to our colon, the part closest to the stomach corresponds to our duodenum. The 2 layers of the stomach and different layers of the eye, the optic nerve and spinal cord were obtained this way.

Another complete set of tissue samples were obtained from a freshly killed steer at a slaughter house. In this way the 4 chambers of the heart were obtained, the lung, trachea, aorta, vein, pancreas, and so forth.

Purchasing a Complete Set of Tissue Samples

Slides of tissues, unstained or stained in a variety of ways for microscope study give identical results to the preparations made by yourself in the ways already described. This fact opens the entire catalog of tissue types for your further study. See *Sources* for places that supply them.

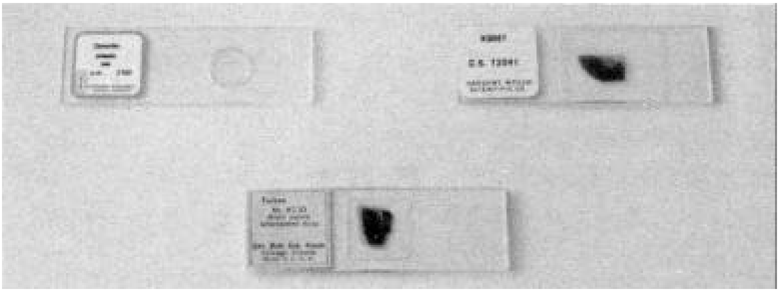

Fig. 82 Some purchased parasite and tissue slides.

You now have a set of organ samples, either fresh, frozen, preserved or on slides. You also have a set of test substances, whether chemical compounds, or elements, or products. Your goal is to search in your own organs and body tissues for the substances that may be robbing you of health.

Keeping yourself healthy will soon be an easy, daily routine.

Body Fluid Specimens

Each of these fluids should be prepared by putting about ¼ tsp. in a ½oz amber glass bottle. Add about 2 tsp. filtered water and ¼ tsp. grain alcohol for preservation. Undiluted specimens do not work for reasons that are technical and beyond the scope of this book. It is important *not* to shake the specimen, but to mix gently.

Urine. It is desired to have a pure, uninfected urine sample as a tissue specimen. Since this cannot be proved with certainty, obtain several urine samples from different persons whom you believe to be healthy and make several test specimens in order to compare results. Label your specimens Urine A (child), Urine B (woman), Urine C (mine), and so forth.

Semen. A sample from a condom is adequate. Aged specimens (sent by mail, unpreserved and unrefrigerated) work well also. Use one to ten drops or scrape a small amount with a plastic knife.

Blood. One to ten drops of blood should be used. Clotted or chemically treated blood is satisfactory. A blood smear on a slide is very convenient.

Milk. Cow's milk is too polluted with parasites to be useful. Electronically, a dead specimen is equivalent to a live specimen, so that pasteurization of the milk does not help. A human milk specimen is preferred.

Saliva. Use your own, if you have deparasitized yourself and test negative to various fluke stages. Otherwise find a well friend or child.

Specifying a tissue is the most powerful investigative technique in your arsenal. Any of your tissue samples can be tested for any of your toxic substances.

You Can Now Test Three Different Ways!

When you test with a substance on one plate and nothing on the other, you are searching your entire body for that substance. Such a test is not very sensitive.

By putting a tissue sample on the other plate you are testing for the substance specifically in that tissue, and this is much more sensitive. The tissue need not be the white blood cells. To find mercury in your kidneys you would use a mercury sample on one plate, and a kidney sample on the other. The technique is the same as when you use white blood cells.

If you put a substance on each plate, a resonating circuit means the two samples have something in common. For example, if you have mercury on one plate and some dental floss on the other, a positive result indicates mercury in the floss.

Lesson Five

Purpose: To watch substances travel through your body.

Materials: Prepare a pint of brown sugar solution (white sugar has propyl alcohol pollution) using filtered water. Use about 1 tsp. brown sugar, 1/8 tsp. vitamin C (to detoxify sorghum mold), and a pint of filtered water. Do not shake it; gently mix. Make a sample bottle by pouring about ½inch into a clean used vitamin bottle. Rinse and dry the outside of the sample bottle. Finally wash your hands with plain water.

Method:

1. Test your skin for the presence of brown sugar, using the newly made sample bottle and your skin specimen. It should not be there (resonate) yet.
2. Prepare a paper applicator by tearing the corner from a white unfragranced paper towel. Fold it to make a wick.
3. Dip the paper wick in the pint of sugar water and apply it to the skin of your inner arm where you can rub freely.

Rub it in vigorously for about 10 seconds (otherwise it takes minutes to absorb). Leave the shredded wick on the skin and tape it down with a piece of clear tape about 4 inches long (this increases the time you have to work). Quickly wash your fingers.

4. Place your skin tissue specimen on one plate and the sugar specimen bottle on the other plate.
5. Probe for resonance every 5 seconds. As soon as you hear resonance, implying that the skin has absorbed the sugar solution (which may take a full minute), replace the skin specimen with one of liver and listen for resonance again. There should be none, yet.
6. Alternate between the skin and liver. Soon the skin will be clear and the liver will resonate. Also check the pancreas and muscles to see how quickly sugar arrives there.
7. Check white blood cells and kidneys. It should not appear here (unless it is polluted with a toxin).
8. After five to ten minutes the sugar will be gone from all of these tissues and your experiment is ended. Wash your arm with plain water.

> Notice that you have only a few minutes to get all your testing done after the skin has absorbed the test substances.

Lesson Six

Purpose: To verify the propyl alcohol and benzene lists.

Method: We will use the Syncrometer to test for a toxin in a product. Assemble the products named in the propyl alcohol list (page 335) and benzene list (page 354)...as many as you can find. Also make sample bottles of benzene and propyl alcohol.

1. Place the propyl alcohol test substance on one plate and your products, in turn, on the other.

2. Listen to the current with only one of the plates in the circuit. Then listen with both plates in (the test plate switch ON). This method can detect one part per quadrillion in concentration. It is not as sensitive as the skin test (Lesson Five).
3. Repeat, with the benzene test substance.

Even tiny amounts of solvents are toxic! They must not be consumed or be left in our environment.

I have found that too many unsuspected products test positive to benzene. This is such a global tragedy that people must protect themselves by using their own tests. Rather than assurances, regulatory agencies should provide the consumer with cheap and simple tests (dip sticks and papers so we need not lug our Syncrometers around). Even if some test should fail, not all tests would fail to find an important pollutant like benzene. It would come to public attention much faster than the present debacle has.

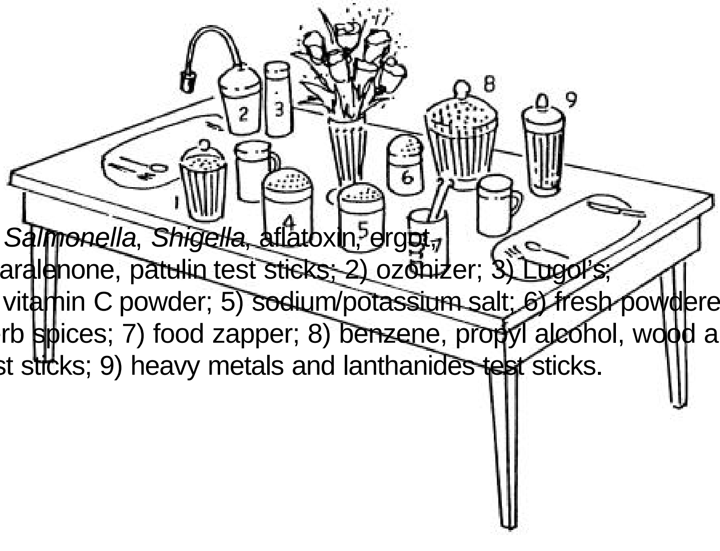

1) *Salmonella*, *Shigella*, aflatoxin, ergot, zearalenone, patulin test sticks; 2) ozonizer; 3) Lugol's; 4) vitamin C powder; 5) sodium/potassium salt; 6) fresh powdered herb spices; 7) food zapper; 8) benzene, propyl alcohol, wood alcohol test sticks; 9) heavy metals and lanthanides test sticks.

Fig. 83 Table of the future.

Lesson Seven

Purpose: To test for the presence of aluminum in your brain and your foods.

Materials: An aluminum measuring spoon, a tsp. of free flowing aluminized salt, a square inch of aluminum foil, a package of pork brain from the grocery store, kept frozen. (Other animal sources will do). Or a stained slide of cerebrum, cerebellum or other brain tissue.

Method:

1. Cut a piece of brain tissue (about 1 tsp.) and place in a plastic bag.
2. Place the aluminum samples in separate plastic bags. Add filtered water to each, about 1 tbs. Keep all surfaces and your hands meticulously clean (do not use soap).
3. Place the aluminum sample on one plate and the brain sample on the other plate.
4. Probe for resonance. If the circuit resonates you have aluminum in your brain.
5. If your aluminum specimen actually has cadmium or copper in it, you are also testing for these in your brain. Repeat the aluminum test with other aluminum objects. If they all resonate, you very, very likely have aluminum in your brain. If you can, test yourself to cadmium and copper, separately. If you don't have these in your brain, the aluminum test result is even more likely to be correct.
6. Of course, it would be desirable to have absolute certainty about this. To achieve this, purchase pure aluminum or an Atomic Absorption Standard. These are available from chemical supply companies.

If you do have aluminum in your brain, where is it coming from?

7. Leave your purest aluminum test substance on one plate, and replace the brain sample with these items, testing them one at a time. Remember to rest after each positive result.
    - a teaspoon of cottage cheese or yogurt taken from the top of a container of a foil-capped variety
    - a piece of cream cheese or butter that was wrapped in foil
    - a chip of bar soap or a bit of hand lotion
    - a piece of cake or rolls baked in an aluminum pan
    - a piece of turkey skin or hot dish that was covered with aluminum foil
    - anything baked with baking powder
    - a carbonated beverage from an aluminum can

Alternative Lesson:

To test for dental metal in your tissues. Use a piece of amalgam from an old tooth filling. This tests for the rest of the alloys in amalgam fillings as well as mercury. If you can't get a piece of mercury amalgam, use a mercury thermometer (don't break it, just put the bulb on the plate). Choose tissues like kidney, nerves, brain, liver, in addition to white blood cells.

I have never dissected human tissues and subjected them to confirmatory laboratory tests. It seems reasonable that because skin and tongue are directly provable, that other tissues work similarly.

Testing the Air

Fine particles and gas molecules that are in the air stick to the dust and eventually fall down onto the table, kitchen counter, and other places. Every night a film of dust accumulates, even though you can't see it. To test for air pollutants, gather some dust. Wipe the kitchen table and counter with a dampened piece of paper towel, two inches by two inches square. Place it in a resealable baggie. Do not get old dust, like from the top of the refrigerator or back shelves, because it does not represent the current air quality.

Testing Someone Else

Seat the person comfortably with their hand resting near you. Choose the first knuckle from the middle or first finger just like you do for yourself. Since you are touching this person, you are putting yourself in the circuit with the subject.

To exclude yourself, you need to add inductance to yourself. A coil of about 10 microhenrys, worn next to the skin, works well and is easily made. Obtain insulated wire and wrap 24 turns around a ball point pen (or something about that size), closely spaced. Cut the ends and tape them down securely. Keep it in a plastic bag, even when in your pocket. A commercial inductance of 4.7 microhenrys, worn touching your skin also works well. It can be worn on a string necklace. (Remember to remove the necklace when testing yourself.) The inductance acts as an RF (radio frequency) choke, limiting the alternating current that can flow through you while testing another person.

Test your inductor in this way. Repeat Lesson One with the coil next to your body. No resonance, even to SALT #1, should occur. If it does, make the coil bigger. Remove the inductor when you are not testing others.

Lesson Eight

Purpose: To detect aluminum in the brain of another person.

Materials: same as previous lesson, you wear the inductor.

Method:

1. Place the aluminum sample on one plate and the brain sample on the other plate.
2. Give the other person the handhold. You use the probe. Hold their finger steady in yours.
3. Probe the other person for resonance. The first probe is with only one plate in the circuit. The second is with both plates in the circuit. Resonance implies there is aluminum in the person's brain.

Saliva Testing

This may become your most useful test. The saliva has in it a bit of almost everything toxic that is in you. But it is not the first tissue to carry the HIV virus or a bit of a tapeworm stage. Nevertheless, *Salmonella* in your liver, mercury in your kidneys, aluminum in the brain all show up in the saliva, too. And saliva can be sent by mail or stored in the refrigerator. It should be frozen for long storage to prevent mold invasion. Or it may have grain alcohol added to preserve it. This test is not as sensitive as having the person present in the circuit, though.

To make a saliva specimen, place a two inch square piece of white, unfragranced paper towel (tear, don't cut) in a lightweight resealable baggy. Hold the open baggy near your mouth. Don't touch the paper towel with your fingers. Drool or spit onto the paper towel until half of it is damp. Zip it shut. Before testing, add enough filtered water to dampen the whole piece of paper.

Lesson Nine

Purpose: To search for shingles or *Herpes*.

Materials: A saliva specimen from the person being tested; they may be thousands of miles away. Also a specimen of the virus. This can be obtained from someone else's lesions—one droplet is enough, picked up on a bit of paper towel. The whole thing, towel and all, can be pushed into a glass bottle for preserving. Water and alcohol should be added. It can also be put on a slide, *Herpes, homemade*. A homeopathic preparation of the virus does not give accurate results for this kind of testing, due to the additional frequency imposed on it by potentizing. (However, homeopathic preparations can be used if the potency matches the tissue frequency where it resides. Hopefully, some way of using homeopathic sources will soon be found.)

Method: Place the saliva specimen in its unopened baggy on one plate. You may wish to open it briefly, though, to add enough filtered water to wet all the paper and add ¼ tsp. grain alcohol to sterilize or preserve it.

Place the virus specimen on the other plate and test as usual (like Lesson Six). A positive result means the person has active *Herpes*.

The main disadvantage of saliva testing is that you do not know which tissue has the pathogen or the toxin. You can only conclude that it is present. Usually this is enough information to carry out a corrective program.

Surrogate Testing

Although saliva testing is so easy, it is also possible to use an adult as a surrogate when testing a baby or pet. The pet or baby is held on the lap of the surrogate. A large pet may sit in front of the person. The handhold is held by the surrogate and pressed firmly against the body of the baby or pet. It can be laid flat against the arm, body or leg of a baby and held in place firmly by the whole hand of the adult. The paper covering should be wet. For a pet, the end is held firmly pressed against the skin, such as between the front legs or on the belly. The other hand of the adult is used for testing in the usual way. The adult must wear an inductor for surrogate testing as well as you, tester.

An ill or bedridden person may be tested without inconvenience or stress. He or she rests their whole hand on the skin of your leg, just above the knee. A wet piece of paper towel, about 4 inches by 4 inches is placed on your leg, to make better contact. You must use an inductor for yourself with this method. You may now proceed to probe on your hand instead of the ill person's.

Lesson Ten

Purpose: To test for cancer.

Materials: Ortho-phospho-tyrosine. Here are four ways to obtain some:

1. Order a pure sample from a chemical company (see *Sources*). Place a few milligrams (it need not be weighed) in a small glass bottle, add 2 tsp. filtered water and ¼ tsp. grain alcohol.
2. All persons with cancer have ortho-phospho-tyrosine in their urine as well as in the cancerous tissue. It is seldom found in other body fluids. Obtain a urine specimen from a friend who has active cancer. Freeze it if you can't prepare it immediately. Keep such specimens well marked in an additional sealed plastic bag. Persons who have recently been treated clinically for cancer are much less likely to have ortho-phospho-tyrosine in the urine.

   Urine cannot be considered a chemical in the same way as a sugar or salt solution. Urine is a tissue and has its own resonant frequency as do our other tissues. If combined with another tissue on the test plates, it will not resonate as if a solution of pure ortho-phospho-tyrosine were used. To use urine as an ortho-phospho-tyrosine specimen, you must:

   a) Pour a few drops of urine into your specimen bottle
   b) Add about 2 tsp. of filtered water
   c) Add a few drops of grain alcohol

   Gently mix, do not shake. Rinse and dry the outside of the bottle. Label it “urine/cancer”.

3. Cancer victims also have other growth factors being produced in their bodies. These are the same as can be found in mother's milk—for example, epidermal growth factor and insulin-like growth factor. Obtain a sample of mother's milk and use it to make another test substance for cancer. A few drops is enough.
4. There is still another way to prepare an ortho-phospho-tyrosine test sample. Common snails from a fish tank or outdoor snails are the natural hosts for *Fasciolopsis buskii* (human intestinal fluke) stages. The stages will produce ortho-phospho-tyrosine when the snails are fed fish food polluted with propyl alcohol. Over half the fish food cans I purchased had propyl alcohol pollution. Buy several brands of fish food. Test them for propyl alcohol and benzene. Obtain some snails, put them in a tank, feed them propyl alcohol polluted fish food. (Feed a separate group of snails benzene polluted fish food to obtain samples of HIV.) After two days put each snail in a zipped plastic bag, and test them individually against someone diagnosed with cancer or their saliva. The snails that the person tests positive to have ortho-phospho-tyrosine. Put these snails in the freezer to kill them humanely, then crush them and place in a specimen bottle with 50% grain alcohol to preserve. The bottles can be kept sealed and at room temperature.

   Similarly, your benzene snails can be tested against someone known to be HIV positive. Any snails that test positive can be used to prepare an HIV test specimen in the same way. The fish food must be tested for both benzene and propyl alcohol pollution, and separated appropriately, or you run the risk of making specimens that have both ortho-phospho-tyrosine and HIV.

Method:

1. Test for cancer by placing the test sample you just made (any of the four) on one plate and a white blood cell sample on the other plate.
2. If you resonate with both samples in the circuit you have cancer. Immediately, search for your cancer in your breast, prostate, skin, lungs, colon, and so forth.
3. To be more certain, test yourself to the other kinds of test samples. You should not resonate.

As you know by now, you can confirm the cancer by testing yourself to propyl alcohol and the human intestinal fluke in the liver. You should eliminate propyl alcohol from use, and zap all parasites. Keep testing yourself for cancer until it is gone. It should take less than one hour. Also continue to test yourself for propyl alcohol and the intestinal fluke in the white blood cells; make sure they are gone. Also test yourself for aflatoxin and freon.

Lesson Eleven

Purpose: To test for HIV.

Materials: Purchase a few milligrams of Protein 24 antigen (a piece of the HIV virus core) or the complete HIV virus on a slide (see *Sources*). You may use the vial unopened if only one test specimen is needed. To make more specimens, use about 1 milligram per ½ ounce bottle. Add 2 tsp. filtered water and ¼ tsp. grain alcohol. Or prepare an HIV specimen from snails as described in the previous Lesson.

Method: Search in the thymus (throat sweet breads), vagina and penis for the virus because that is where it will reside almost exclusively for the first year or two. If you don't have those tissue specimens, you could search in urine, blood, saliva, or white blood cells, but only a positive result can be trusted. Also search for the human intestinal fluke and benzene in the thymus. Of course, a positive test in these tissues is very significant. If you are positive, zap parasites immediately. You should test negative in less than an hour. Remove benzene polluted items from your lifestyle. Also test yourself to several varieties of popcorn, brown rice, and corn chips as an indication of zearalenone, which must be eliminated in order to get well. Follow up on yourself every few days to be sure your new found health is continuing. Test yourself for freon.

Lesson Twelve

Purpose: To test for diseases of all kinds.

Materials: Use slides and cultures of disease organisms. Homemade preparations of strep throat, acute mononucleosis, thrush (*Candida*), chicken pox, *Herpes* 1 and 2, eczema, shingles, warts, measles, yeast, fungus, rashes, colds, sore throats, sinus problems, tobacco virus, and so forth can all be made by swabbing or scraping the affected part. A plastic spoon or bit of paper towel works well. Put a small bit on a slide. Add a drop of balsam and a cover slip. Or put the towel in a bottle, add water and alcohol as described previously. Microscope slides can greatly expand your test set (see *Sources*).

Method: Test yourself for a variety of diseases, using your white blood cell specimen first. Then search in organs like the liver, pancreas, spleen. Notice how many of these common illnesses don't “go away” at all. They are alive and well in some organ. They are merely not making you sick!

Lesson Thirteen

To test for AIDS.

Materials: Benzene sample, slides of tissue samples like thymus, liver, pancreas, penis, and vagina. Also a collection of disease specimens such as the ones used in the previous lesson.

Method: Search in the thymus for benzene. If it is positive throughout the day, you are at risk for developing AIDS, although you may not be ill. Search other tissues for benzene. The more tissues with benzene in them the more serious the situation. Immediately search all your body products and foods for benzene.

Stay off benzene polluted items forever.

Tally up the diseases you tested positive for in Lesson Twelve. Test at least ten. If you had more than half positive you already have AIDS. (50% is my standard, you may set your own; an ideal standard for defining a healthy person should be 0% positive.)

Lesson Fourteen

Purpose: To test for aflatoxin.

Materials: Do not try to purchase a pure sample of aflatoxin; it is one of the most potent carcinogens known. Having it on hand would constitute unnecessary hazard, even though the bottle would never need to be opened. Simply make specimens of beer, moldy bread, apple cider vinegar, and any kind of peanuts using a very small amount and adding filtered water and grain alcohol as usual.

Method: Test yourself for these. If you have all of them in your white blood cells and the liver then you very, very probably have aflatoxin built up. Next, test your daily foods for their presence in your white blood cells. Those that test positive must be further tested for aflatoxin. Notice the effect of vitamin C on aflatoxin in your liver. Find a time when your liver is positive to aflatoxin (eat a few roasted peanuts from a health food store and wait ten minutes). Take 1 gram vitamin C in a glass of water. Check yourself for aflatoxin every five minutes. Does it clear? If not, take 5 or 10 grams vitamin C. How long does it take?

Lesson Fifteen

Purpose: To test for parasites.

Method: If you test positive to your pet's saliva, you have something in common–a parasite, no doubt. You must search your muscles and liver for these, not saliva or white blood cells, because they are seldom seen in these. Zap yourself for parasites until you no longer test positive to your pets' saliva.

Tapeworms and tapeworm stages can not (and should not) be killed with a regular frequency generator. Each segment, and probably each scolex in a cysticercus has its own frequency and might disperse if your generator misses it. Only zapping kills all and is safe for tapeworms.

Be sure to treat your pet on a daily basis with the pet parasite program.

Lesson Sixteen

Purpose: To test for fluke disease.

A small number of intestinal flukes resident in the intestine may not give you any noticeable symptoms. Similarly, sheep liver flukes resident in the liver and pancreatic flukes in the pancreas may not cause noticeable symptoms. Their eggs are shed through the organ ducts to the intestine and out with the bowel movement. They hatch and go through various stages of development outdoors and in other animals. But if you become the total host so that various stages are developing in your organs, you have what I term *fluke disease*. I have found that cancer, HIV, diabetes, endometriosis, Hodgkin's disease, Alzheimer's disease, lupus, MS and "universal allergy syndrome" are examples of fluke disease.

You can test for fluke disease in two ways: electronically and by microscope observation.

Materials: Cultures or slides of flukes and fluke stages from a biological supply company (see *Sources*) including eggs, miracidia, redia, cercaria, metacercaria. Body fluid specimens to help you locate them for observation under a microscope.

Method: Test for fluke stages in your white blood cells first. If you have any fluke stages in your white blood cells you may wish to see them with your own eyes. To do this, you must first locate them. Place your body fluid samples on one plate, your parasite stages on the other plate, and test for as many as you were able to procure, besides adults. After finding a stage electronically, you stand a better chance of finding it physically with a microscope.

Lesson Seventeen

Purpose: To see how sensitive your measurements can be. (How much of a substance must be present for you to get a positive result?)

Materials: filtered water, salt, glass cup measure, 13 new glass bottles that hold at least ¼ cup, 14 new plastic teaspoons, Your skin tissue sample, paper towel.

Method: Some of the best measurement systems available today are immunological (such as an ELISA assay) and can detect as little as 100 fg/ml (femtograms per milliliter). A milliliter is about as big as a pea, and a femtogram is 1/1,000,000,000,000,000th (10^{-15}) of a gram!

1. Rinse the glass cup measure with filtered water and put one half teaspoon of table salt in it. Fill to one cup, stirring with a plastic spoon. What concentration is this? A teaspoon is about 5 grams, a cup is about 230 ml (milliliters), therefore the starting concentration is about 2½(2.5) gm per 230 ml, or .01 gm/ml (we will discuss the amount of error later).
2. Label one clean plastic spoon "water" and use it to put nine spoonfuls of filtered water in a clean glass bottle. Use another plastic spoon to transfer one spoonful of the .01 gm/ml salt solution in the cup measure to the glass bottle, stir, then discard the spoon. The glass bottle now has a 1 in 10 dilution, and its concentration is one tenth the original, or .001 gm/ml.
3. Use the "water" spoon to put nine spoonfuls of filtered water in bottle #2. Use a new spoon to transfer a spoonful of salt solution from bottle #1 to bottle #2 and stir briefly (never shake). Label bottle #2 ".0001 gm/ml".
4. Repeat with remaining bottles. Bottle #13 would therefore be labeled ".000000000000001 gm/ml." This is 10^{-15} gm/ml, or 1 femtogram/ml.
5. Do the skin test with water from bottle #13 as in Lesson Five. If you can detect this, you are one hundred times as sensitive as an ELISA assay (and you should make a bottle #14 and continue if you are curious how good your sensitivity can get). If you can not, try to detect water from bottle #12 (ten times as sensitive as ELISA). Continue until you reach a bottle you can detect.

Calculate the error for your experiment by assuming you could be off by as much as 10% when measuring the salt and water adding up to 20% error in each of the 13 dilutions. This is a total error in bottle #13 of 280%, or at most a factor of 3. So bottle #13 could be anywhere from 0.33 to 3 femtogram/ml. If you can detect water from bottle #13, you are definitely more sensitive then an ELISA, in spite of your crude utensils and inexpensive equipment! Note that the starting error of using 2.5 gm instead of 2.3 gm only adds another 10% error.

If you want to calculate how many salt molecules you can detect, select the concentration at the limit of your detection, and put 2 drops on a square inch of paper towel and rub into your skin. Assume one drop can be absorbed. If you can detect water from bottle #13, you have detected 510,000 molecules (10^{-15} fg/ml divided by 58.5 gm/M multiplied by $6.02x10^{23}$ molecules/M divided by 20 drops/ml). Water in bottle #12 would therefore have 10 times as many molecules in one drop, and so forth. Even if your error is as much as a factor of 2 (100%), you can still get a good idea of what you can measure.

Atomic absorption standards start at exact concentrations; it is easy to make a more exact dilution series with them. When testing for iridium chloride by this skin test method, I was able to detect 3025 molecules!

Troubleshooting:

Always extend your set until you get a negative result (this should happen by at least bottle #18). If you always “detect” salt, then you shook the bottle!

Never try to reuse a bottle if you spill when pouring into it. Get another new bottle.

Sensitivity of Pollutant-In-Product Testing

Get some slides of *Salmonellas* and *Shigellas* and find some milk that tests positive to at least one. Make a dilution series of the milk up to bottle #14, being careful not to shake the bottles. Start with 2 drops of milk in bottle #1. Use an eye dropper to deliver 2 drops to subsequent bottles. Begin testing at bottle #14, using the slide that tested positive. You will learn to search by frequency later. My sensitivity was routinely around bottle #12, for a variety of pathogens. It was the same for toxic elements starting with standard solutions, about 1000 μg/ml, showing this method is less sensitive than skin testing.

Microscopy Lesson

Purpose: To *observe* fluke stages in saliva and urine with a microscope.

Materials:

a. A low power microscope. High power is not needed. A total of 100x magnification is satisfactory for the four common flukes, *Fasciolopsis*, *sheep liver fluke*, *human liver fluke* and *pancreatic fluke*.
b. Glass slides and coverslips.
c. A disposable eye dropper.

PRECLEANED
MICROSCOPE SLIDES
PLAIN GLASS
WARD'S
GLASS SLIDES

Fig. 84 Microscope, slides and coverslips.

d. For sanitation purposes (wiping table tops, slides, microscope and your hands) a 50% to 70% alcohol solution (not rubbing alcohol!) is best. Dilute 95% grain alcohol 7 parts alcohol plus 3 parts water. Vodka or 76% grain alcohol can be used undiluted.

e. Formaldehyde, 20%. Formaldehyde 37% is commonly available at pharmacies. Dilute this with equal parts of filtered water to get 18½%, which is close enough to 20%, for the purpose of "fixing" (killing) the specimens. Store in a glass bottle in the garage, away from sunlight. Label. Specimens that are fixed properly do not lose their life-like appearance.

f Iodine solution. This is only useful for the urine specimens. Lugol's iodine and tincture of iodine are both useful. Ask a pharmacist to prepare *Lugol's Iodine Solution* for you, as follows:

- 44 grams (1½oz) iodine crystals
- 88 grams (3 oz) potassium iodide crystals

Dissolve both in 1 liter (quart) filtered water. This may take a day of frequent shaking.

Method for saliva:

1. Pour the 20% formaldehyde into a small amber bottle or other receptacle to a depth of about 1/8 inch. Keep tightly closed.
2. The person to be tested is asked to salivate into the bottle so the organisms are immediately "fixed" without undergoing cooling first. The total volume should be about double the original amount of formaldehyde used. Make a mark on the container so the subject knows how much to produce. The resultant concentration of formaldehyde will be about 10%.
3. Shake the bottle a few times. Set it aside for 24 hours to settle (less if testing is urgent).

4. With a dropper, draw up some of the bottom settlings. Put one drop on a slide and apply a coverslip.
5. View under low power of microscope. Compare objects you observe with specimens obtained on slides from biological supply companies.

Note: Persons with HIV and moderate AIDS will show about one to ten parasite stages per slide. It requires several hours of searching. Persons with HIV and severe AIDS show 10 or more fluke stages per slide; this makes the task of finding them much easier. Persons with terminal untreated cancer have many more fluke stages than relatively well persons.

Method for urine:

1. Prepare bottles of formaldehyde fixative ahead of time. Put about ¼ to ½ inch of 20% formaldehyde in each. Keep tightly closed.
2. Add freshly voided[36] urine from cancer or HIV sufferers to the formaldehyde in approximately equal amounts, resulting in a 10% formaldehyde solution. Shake immediately. Let settle several hours. The sediment has a higher number of fluke stages. Cancer victims with cervical or prostate cancer will show higher numbers of stages in urine than other cancer types.
3. Staining the slide is optional. It helps to outline fluke stages slightly. Prepare Lugol's solution as described above.

    Slides may be stained in either of these two ways:

    - Put a drop of "fixed" urine on a slide. Add a drop of 50% Lugol's (dilute 1:1 with filtered water). Apply coverslip.

---

[36] Urine that has cooled even slightly below body temperature does not show *miracidia* and *redia* in their original shapes.

- Put a drop of "fixed" urine on a slide. Apply coverslip. Add 1 to 3 drops of 50% Lugol's to edge of coverslip and allow it to seep in.

Note: persons who have been treated for cancer or HIV using any of the known drugs may show only 1 to 2 fluke parasite stages per drop of saliva or urine. For this reason, you may need to search through 20 or more slides to find flukes. Very ill persons may show up to 10 parasites per drop (slide).

Taking Pictures Of What You See

You may be unsure of what you see even if you have the microscope slides of labeled flukes and their stages to study and compare. In real life, they vary so much in shape and size that absolute identification is difficult without experience. Unfortunately in a few hours, just as you are getting proficient, your magnificent specimens will be drying out and unfit for observation. To preserve them longer you can seal the edges by painting around the coverslip with fingernail clear enamel. Or dribble hot sealing wax along the edges and then place them in sealed plastic bags (one per bag). Melt sealing wax in a metal jar lid. Make an applicator from a piece of coat hanger wire bent in the shape of a small square to fit around the coverslip and a handle.

Or take photographs. To take pictures of what you see under the microscope you will need a photomicrographic camera, which costs $200.00 and up (see *Sources*). It is easy to use. Remember to label your pictures so you know which slide they came from.

Even photographs do not scientifically prove identity of parasite stages, but it is very good evidence. Proof would require that the saliva or urine sample could be cultured and seen to produce the known parasite stages.

Using A Frequency Generator

Frequency generators come in all sizes and costs and capabilities. If you can purchase one that reads out the frequency for you in numbers (digital type) and lets you produce a fraction of a kilohertz by turning a dial, it meets your most elementary needs. It should also be possible to set it on positive offset (100% positive) and still give you 5 volts. Then it can be used as a zapper. You should also be able to select a sine wave or square wave.

The advantages of having a frequency generator are that you can do your own diagnosing. You can find, in a few minutes:

1. Which invaders you have by dialing in to their frequencies.
2. At what frequencies you are broadcasting to the world.
3. What frequencies are used by other living things.

The Theory

Every living animal and every cell type produces its own frequencies and responds to these frequencies as well. We may speak of frequencies but we really mean waves, waves of energy. All waves have a frequency associated with them, so it's not really misleading to say frequencies when we really mean waves.

When the animal is alive it produces them, when it is dead it still responds to some of them. It is like the opera singer and the glass goblet. The opera singer produces frequencies in the air. The goblet responds to them because of its structure, not because it's alive. The goblet “picks up” on that particular frequency of sound because its own “resonant” frequency is exactly the same. If the singer sings loud enough the goblet shatters from the vibrations set up in it. An identical goblet, made of plastic, would not have the same resonant frequency.

There is not merely a structural and chemical difference between the living and non living. Living things are transmitting energy of some unique kind. And with your simple device, you will be able to trap this energy. And measure its exact frequency. This is not the same as understanding its makeup and source. We must leave that to others. But we can observe and use our observations to track down bacteria and other parasites. We can track down pollutants. We can measure our health quantitatively and perhaps in the future predict life expectancy.

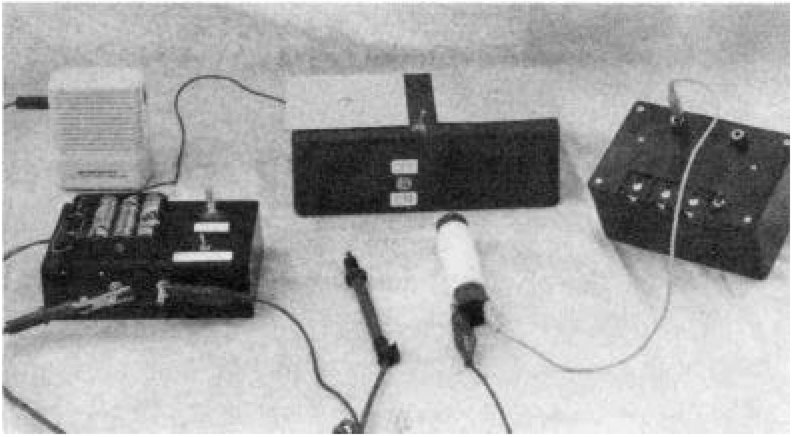

The Syncrometer at left is attached to the handhold. The frequency generator at right is also clipped to the handhold.

Fig. 85 Setup for finding resonant frequencies.

The Syncrometer traps the frequencies that match the ones in the material on the test plates and delivers them to an audio speaker in a range that you are able to hear. **Instead of test tissues or pathogens, we are now going to use pure test frequencies!** Remarkably, few things have overlapping frequencies, making this technique extremely useful for identification, even without a specimen!

Lesson Eighteen

Purpose: To search for the intestinal fluke in your body by listening to its broadcast frequency at 434 KHz.

Method: Turn on the frequency generator, select a frequency a short distance above the one you are interested in, like 438 KHz, turn the voltage (amplitude) down to less than one volt. Select sine waves. The lead coming from the frequency generator will have two connections, usually red and black (ground). We will not use the black (ground), just tape it out of the way. Pick up the handhold and probe in the usual way. Attach the red lead coming from the generator to your handhold. This makes two wires attached to your handhold. Although there is nothing on the test plates, they must be connected as usual with the switch at OFF (one plate is still ON).

Turn the Syncrometer ON. Probe yourself as usual. Your body's waves are being sent to the capacitor (plate) in the test plate box. The frequency from the Syncrometer is sent there, too. And now the 438 KHz waves from the generator are being sent there as well. Three different frequencies are mingling on the plate! If the two from your body and the generator are the same, the circuit will oscillate, and you will hear resonance. Turn the generator to 437 and probe again. Next, 436.

Sometimes, you can hear the resonance start to build. Continue on.

Next try 435. Then 434.

If your body is emitting a frequency of 434 KHz (coming from a live intestinal fluke inside you) it will be reinforced by the generator's 434 KHz. The reinforcement will put oscillations or resonance in the circuit, the same as you are accustomed to hearing with the Syncrometer. If there was none, you don't have the intestinal fluke anywhere in your body. Confirm this by starting at 430 KHz and working your way up.

If you hear resonance, you do have it. You may wish to verify this independently using a prepared slide of the fluke. Kill your flukes immediately as described in the next lesson.

Lesson Nineteen

Purpose: Killing the intestinal fluke with a frequency generator.

Materials: A frequency generator, two handholds with alligator clip leads for them.

Method: Wrap a single layer of paper towel over each of the two handholds. Wet them under the tap; squeeze out excess water. Clip them to the red and black wires of the frequency generator. (We use both wires for this purpose.) Dial up 434 KHz. Set the amplitude (voltage) at 10 volts. Grasp the handholds in each hand and hold on for three minutes. That is all. You have killed whatever tiny invader has a resonant frequency the same as the setting on the generator. Remember to zap all the stages, too; see *Pathogen Frequencies*.

If your frequency generator has a **positive offset** capability, you can use it like a zapper, and a single session will kill all pathogens, provided it is 100% offset and can give at least 5 volts at this setting. When using this technique, the generator can be set to any frequency from 2 KHz to 800 KHz, and you should go for 7 minutes. But even a small percentage of negative voltage will ruin this effect and do more harm than good! To be certain your generator is set correctly it would be best to observe the output on an oscilloscope.

Experiment with other voltage settings. Notice that less than one volt is also effective.

Lesson Twenty

Purpose: To find the bandwidth of a small living animal.

Materials: A fly, beetle or other insect, Syncrometer, frequency generator.

Discussion: Persons using a Syncrometer might have already tried putting a small insect on one of the plates. The circuit always resonates when you join the circuit at the handhold and probe. Even the tiniest ant placed in a glass bottle or plastic baggy will resonate the circuit. Unless it is too far away from the plate. If it has climbed up the side you will lose the resonance. At least one foot must be touching the bottom of the bottle. If the animal is dead this ceases. Obviously the living thing is affecting the circuit differently before and after death. Is it some kind of a wave form energy? To find its frequency you must add another frequency that will reinforce or interfere with the frequency already on the plate. Adding the generator frequency does just that.

Method: Use the same method as described in the last Lesson; however for an ant or fly, start at 1,000 KHz and proceed upward in big steps like 10 KHz. Use the right test plate which is controlled by the ON-OFF switch. Always listen to the current with the switch OFF, first, then ON. Move the frequency up and repeat. Continue until you hear resonance. Stop immediately. Rest your skin and go back down to the nonresonant frequency region. Move up in smaller steps this time. Repeat and repeat until you feel sure you know just where the resonance begins. But where does it end?

Start testing well above the suspected range taking big steps downward until you reach a resonant frequency. Rest and repeat until you find the upper limit of resonant frequencies. Record the bandwidth, for example, 1009-1112 KHz.

Lesson Twenty One

Purpose: To see if similar living things interfere with each other when put on the plate together.

Materials: Two identical living insects or very small living things.

Method: Find the broadcast range of each one separately and then together on the plate.

Note: Identical living things do not interfere with each others' frequencies.

Lesson Twenty Two

Purpose: To see if different living things interfere with each other when put on the plate together.

Method: Find the lower and upper end of the broadcast range of two different living things, such as a fly and a beetle or 2 kinds of flies or beetles. Then put them on the plate together. Notice there is no resonance in the accustomed range for either of them. They are interfering with each other on the plate.

Now add the 2 lower ends, then the two upper ends. Also subtract the 2 lower ends, then the two upper ends. For example imagine two insects, one with a spectrum of 1000 to 1090 KHz, the other with a range of 1050 to 1190 KHz. Adding the lower ends gives us 2050 KHz. Subtracting the lower ends gives us 50 KHz. Adding the upper ends gives 2280. Subtracting the upper ends gives 100. Now search for resonance at 50, 100, 2050, 2280 KHz. (These last two may be outside the range of your frequency generator. Choose more primitive life forms which have lower frequency bandwidths to stay within your limit.)

Notice that you hear resonance at exactly these frequencies and not above or below them. This is evidence for modulation of the frequencies: namely fusing them together.

Lesson Twenty Three

Purpose: To find your own bandwidth of emitted frequencies.

Materials: A frequency generator that goes up to 10 MHz. If yours only goes to 2 MHz you can still investigate the lower end of your band.

Method: You do not need to put yourself on the plate, since you are already there by being in the circuit at the handhold. However, if you are measuring someone else, they can simply touch the plate with a finger. Attach the frequency generator to the circuit at the handhold.

Since human adults begin to emit at about 1560 KHz, start searching at 1550, going upward in 1 KHz steps until you hear resonance.

Younger or healthier humans start emitting at a lower frequency and sometimes end at a higher frequency. In other words, they broadcast on a wider band.

Very young infants begin their band at about 1520 KHz. Could you ever regain this ability?

Most adults terminate at 9375 KHz

By eliminating molds from my diet, killing as many parasites and removing as many toxins as I became aware of, I have been able to expand my bandwidth from an initial 1562-9457 KHz in 1990 to 1520-9580 KHz in 1994! I hope this challenges you to accomplish a health improvement reflected in an even broader bandwidth for yourself.

Lesson Twenty Four

Purpose: To find the effect of a variety of things on the lower end of your spectrum, such as body temperature, eating, time of day, rainy weather, feeling sick. Notice that you may not change for weeks at a time, then suddenly see a shrinking of your bandwidth. You may assume you have eaten a mold.

Search for mold frequencies from 75 KHz to 295 KHz. Or test in your liver with mold samples. If this is positive go on a mold free diet—watching carefully for mold in your white blood cells. Even after removing the mold from your diet, so that no molds appear in your white blood cells, notice that your bandwidth does not recover. It regularly takes 2-3 weeks for mine to recover.

Surely, this sheds light on the poisonous effect of eating bad food.

Lesson Twenty Five

Purpose: To find an emission spectrum using a saliva sample.

Materials: A regular frequency generator.

Method: Search for the bottom of the resonant frequency band as in the previous lesson.

You may store it in the refrigerator for a few weeks without seeing a change. After that the band begins to shrink.

Lesson Twenty Six

Purpose: To observe the effect of dying on the bandwidth.

Method: Freeze the insect you tested in Lesson Twenty to kill it humanely. Repeat the search for its bandwidth. Note the bandwidth has become very narrow.

Note the bandwidth also depends on the accuracy of your particular frequency generator.

Lesson Twenty Seven

Purpose: To find unknown invaders of your body.

Method: Start at 900 KHz and proceed down to 77 KHz in 1 KHz steps, to search for all pathogens. If you find a resonating frequency, go to the Pathogen Frequency Chart (page 561) to identify likely candidates for it. Verify the identity of the invader by using a slide or culture specimen. If your pathogen remains unidentified, add it to the chart. This lets you determine whether the next illness is new or a recurrence of this one. Or just kill it.

Use the frequency generator at one pathogen's frequency. Wait ten minutes and retest all of the ones you found. Only that one will be gone. Now **zap**, wait ten minutes, and test again for all of the ones you found. Notice they are all gone. After one hour, search yet again for the pathogens you had. Any that are back must have come from an internal source not reached by the zapper current, like from the bowel or an abscess.

Lesson Twenty Eight

Purpose: To observe the action of a positive offset frequency on a very small animal. Does the animal die or is it just incapacitated?

Materials: A slug or small earthworm.

Method: Place the small animal in a plastic container like a cottage cheese carton. Add a few tsp. of water to wet the bottom. Attach a metal teaspoon to each of the generator clips. Place them on opposite sides in the carton and fasten with tape. Set the generator to positive offset at a frequency of about 30 KHz and 5 to 10 volts. Experiment with different voltages and compare effectiveness. Measure the time it takes for the animal to seem lifeless. You may try to revive it by keeping it for some time in the presence of food. Retest its emission band.

Lesson Twenty Nine

Purpose: To kill the bacteria in dairy products.

Materials: A glass of regular pasteurized milk, a carton of cottage cheese. A zapper.

Method: Search for *Salmonellas* and *Shigellas* first in the milk and cottage cheese. Search by frequency, using the chart, or with slides of these bacteria. If you don't find any, search different dairy foods until you find some bacteria. Attach metal teaspoons to the red and black leads of the generator. Place them inside the milk glass or cottage cheese carton, across from each other. Secure with masking tape. Attach the zapper. Zap them for 7 minutes. Remove the electrodes and wait 5 minutes. Test again for the same bacteria. They should be gone (but the food is not safe to eat due to the metal released from the teaspoons).

These experiments point to some exciting possibilities. Perhaps water supplies as well as foods and medicines could be sterilized this way. Perhaps sewage could be treated more efficiently, electrically. Best of all, maybe you could protect yourself from unsanitary products. If you do decide to explore this possibility, remember not to put metals in your mouth or food. Nor to use currents greater than 10 milliamps, or for longer than 10 minutes.

There are many commercially available function generators that can meet your needs. Two are shown here.

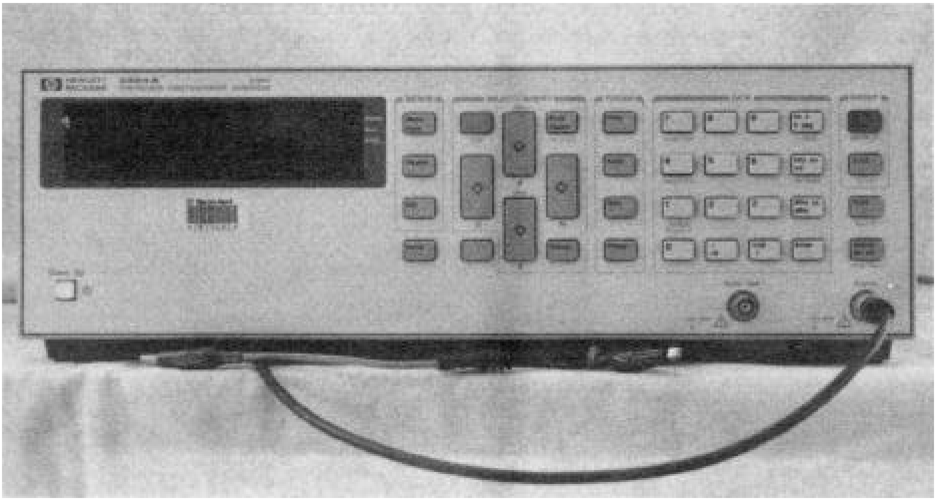

Fig. 86 Function generators.

Recipes

Read old recipe books for the fun and savings of making your own nutritious food. Change the recipes to avoid processed ingredients. Here are some I found:

Beverage Recipes

Anything made in your own juicer is fine. Experiment with new combinations to create different flavorful fruit and vegetable juices. Consider the luxury of preparing gourmet juices which satisfy your own individual palate instead of the mass-produced, polluted varieties sold at grocery stores. Remember to wash all fruit, including citrus before juicing. This removes the ever-present pesticides and common fruit mold.

Lemonade

1 cup fresh lemon juice, 1 cup honey, 1½ quarts water. Bring honey and water to a boil if you plan to keep it several days. Then add lemon juice and store in the refrigerator.

All honey and maple syrup should have vitamin C added to it as soon as it arrives from the supermarket. Warm it first; then stir in ¼ tsp. per pint.

Fresh Tomato Juice

Simmer for ½ hour: 12 medium-sized raw, ripe tomatoes, ½ cup water, 1 slice onion, 2 ribs celery with leaves, ½ bay leaf, 3 sprigs parsley. Strain these ingredients. Season with: 1 tsp. salt (aluminum-free), ¼ tsp. paprika, ½tsp. honey. Serve thoroughly chilled. Makes about 4 servings.

Fresh Pineapple Juice

Peel a pineapple. Remove all soft spots. Cut it into cubes. Extract the juice by putting the pineapple through a food grinder or a blender. There will be very little pulp. Strain the juice and serve it on ice with sprigs of mint. Makes about 1½cups of juice. Mix the pulp with an equal amount of clover honey and use as topping (kept in freezer) for homemade ice cream (below), pancakes, or yogurt.

Maple Milk Shake

For each milk shake, blend or shake together: 1 glass of milk and 2 tablespoons maple syrup.

Remember, all milk gets boiled.

Yankee Drink

Mix together 1 gal. water, 3 cups honey, ½cup fresh lemon juice or distilled white vinegar, and 1 tsp. ginger.

Hot Vanilla Milk

Add one inch of vanilla bean and one tsp. honey to a glass of milk and bring to a near boil. You may add a pinch of cinnamon or other pure spice. You may even use vanilla extract (see *Sources*).

Red Milk

Equal parts fresh carrot juice (use a juicer) and sterilized milk. Save the carrot pulp for salads and soups.

C-Milk

Milk can absorb a surprising amount of vitamin C powder without curdling or changing its flavor. Try ½tsp. in a glass of cold milk.

My Own Soda Pop

Excellent for stomach distress. Put 1 tsp. citric acid, 2 tbs. vegetable glycerin (see *Sources*), 2 tbs. honey, and 1 lemon, juiced by hand, into a quart jar and fill with cold water. Refrigerate until ready to use. Then add 1 tsp. baking soda (chemically pure only, see *Sources*) and shake a few times, keeping the lid tight. Pour over a few ice cubes. Many variations are possible: other fruit concentrates, made in the blender, can be used along with some lemon juice; for example, 2 blended whole apples (peeled), blended pineapple, orange or grapefruit. Always add a bit of lemon to give it zip. You may add a pinch of ginger or other pure spice.

Note: The amount of sodium in ½tsp. baking soda is .476 grams. If you have heart disease, high blood pressure, or edema, use potassium bicarbonate instead. Ask your doctor what an acceptable amount of sodium or potassium bicarbonate is. I would suggest limiting yourself to one glass of soda pop a day, even if you do not have heart disease.

Another Note: the citric acid kills bacteria, while the carbonation brings relief.

My Own Super C-Pop

An excellent way to get lots of vitamin C into a child and relieve stomach distress at the same time. Squeeze 1 slice of lemon and 1 whole orange into an 8 ounce bottle that has a tight lid. Add 1 tsp. vitamin C powder (ascorbic acid), ¼ tsp. citric acid, and 2 tbs. vegetable glycerin (you may also experiment with honey for sweetness). Fill the bottle to the top with cold water. Then add ½tsp. chemically pure baking soda and close tightly. Shake briefly and serve immediately.

Half And Half

Mix equal parts whipping cream and milk or water. Boil and chill.

Buttermilk-C

Stir 1 tsp. vitamin C powder into a glass of milk. Add a pinch of potassium chloride. Additional seasoning may be pepper and herbs. Stir and enjoy.

Food Recipes

Despite the presence of aflatoxins, benzopyrenes, and solvents in many foods, it is possible to have a delicious and safe diet. Many persons need to gain weight, and with the emphasis in today's society on losing weight, consider yourself lucky in this respect. Help yourself to lots of butter, whipping cream, whole milk, avocados, and olive oil. Make your own preserves and baked goods, including breads. Remember, when you are recovering from a major illness it is essential not to diet to lose weight. You must wait two years after you are recovered to try to lose weight.

Daily Foods

Dairy products should contain at least 2% fat to enable you to absorb the calcium in them.

All milk should be sterilized

by boiling it for 10 seconds. If it makes mucous, you already have a chronic respiratory infection. Try to clear this up.

Change brands every time you shop to prevent the same pollutants from building up in your body.

If frying or cooking with fat, use only olive oil, butter or lard (the BHT and BHA preservatives in lard are OK except for seizure sufferers). Mix them for added flavor in your dishes. Never use margarine, Crisco™ or other hydrogenated fats. Do not cook over flames or grill, even when electric.

Eat lots of **fresh fruits and vegetables**. Wash them off only with cold tap water, not commercial food "wash". Scrub hard with a stiff bristled brush. Then cut away blemishes. Always peel potatoes, apples, and carrots. Modern dirt is full of chemicals and is toxic to you.

Be sure to drink plenty of **plain water** from your cold faucet throughout the day, especially if it is difficult for you to drink it with your meals. If you don't like the taste of your own tap water, try to get it from a friend with newer plumbing. Use a polyethylene (opaque) water jug from a grocery store to transport it. Never drink water that has been run through a water softener or copper plumbing or has traveled through a long plastic hose. Don't drink water that has stood in a container for a day. Dump it and sterilize the container. To further improve flavor and to dechlorinate attach a small faucet filter made of carbon only. Or buy a filter pitcher (see *Sources*). Don't drink water that has stood in the filter pitcher very long, either.

Because commercial **cold cereals** are very convenient, but have solvents, here are two replacements.

Two Granolas

7 cups rolled oats (old fashioned, not quick)
1 tsp. salt
1 cup wheat germ (fresh, not defatted)
½ cup honey
½ cup sunflower seeds, immaculate quality
½ cup milk (no need to sterilize, it is being baked)
½ cup melted butter
1 cup raisins, rinsed in vitamin C water

Mix dry ingredients together. Mix liquid ingredients and add gradually, while tossing until thoroughly mixed. Place in large ungreased pans and bake in slow (250°) oven. Stir occasionally, baking until brown and dry, usually 1-2 hours. Store in airtight container in freezer.

6 cups rolled oats
½ cup raw wheat germ
1 cup sesame seeds
1 cup sunflower seeds (raw, unsalted)
1 tsp. cinnamon
½ cup melted butter
½ cup honey

Preheat oven to 250°. Toss all ingredients in mixing bowl. Spread thinly on a baking sheet and bake 20-25 minutes. Stir often in order to brown evenly. When golden, remove and let cool. Makes 12 cups.

If you would like to add nuts to your granola recipes, rinse them in cold tap water first, to which vitamin C powder has been added (¼ tsp. per pint). This removes aflatoxins.

Peanut Butter

Use fresh unsalted roasted peanuts—they will be white on the first day they arrive at the health food store from the distributor. (Ask when they will arrive.) Or shell fresh roasted peanuts yourself, throwing away all shriveled or darkened nuts. Grind, adding salt and vitamin C (¼ tsp. per pint) as you go. For spreadability, especially for children, grind an equal volume of cold butter along with the peanuts. This improves spreadability and digestibility of the hard nut particles. This will probably be the most heavenly peanut butter your mouth has ever experienced.

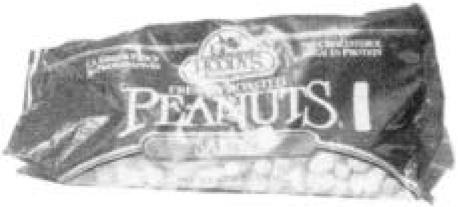

Fig. 87 Light colored, roasted peanuts in the shell had no aflatoxin.

Sweetening and Flavoring

Brown Sugar. Although I am prejudiced against all sugar from a health standpoint, my testing revealed no benzene, propyl alcohol, wood alcohol. However it does contain sorghum mold and must be treated with vitamin C to detoxify it. Add ¼ tsp. to a 1 pound package; knead until well mixed.

Maple syrup. Add vitamin C to newly opened bottle, ¼ tsp. to retard mold. Keep refrigerated and use promptly.

Flavoring. Use maple, vanilla (both natural and artificial), and any pure spice. They are free of molds and solvents.

Honeys. Get at least 4 flavors for variety: linden blossom, orange blossom, plain clover and local or wild flower honey.

Add vitamin C to newly opened jar to detoxify ergot mold (¼ tsp. per pint).

Jams and jellies. They are not safe unless homemade.

Fruit syrup. Use one package frozen fruit, such as cherries, blueberries or raspberries. Let thaw and measure the amount in cups (it might say on the package). Add an equal amount of clover honey to the fruit. Also add ¼ tsp. vitamin C powder. Mix it all in a quart canning jar and store in the refrigerator. Use this on pancakes, cereal, plain yogurt and homemade ice cream too. Use to make your own flavored beverages in a seltzer maker or to make soda pop. If you wish to use fresh fruit, bring it to a boil to sterilize. Use it up in a few days or boil to sterilize it again.

Note for diabetics

Diabetics must not use artificial sweeteners. Nor can they use all the sweeteners listed. Try stevia powder instead.

Preserves

Keep 3 or 4 kinds on hand, such as peach, pineapple, and pear. Peel and chop the fruit. It should not have any bruises. If you use a metal knife, rinse the fruit lightly afterwards. Add just enough water to keep the fruit from sticking as it is cooked (usually a few tablespoons). Then add an equal amount of honey, or to taste and heat again to boiling. Put in sterile jars in refrigerator. Make marmalade the same way, slicing the fruit and peel thinly. Always add vitamin C powder to a partly used jar to inhibit mold. Never use up partly molded fruit by making preserves out of it. Throw it out.

C Dressing

½ cup olive oil
¼ cup fresh lemon juice or white distilled vinegar
1 tsp. thyme, fenugreek or both (capsules are freshest)
1 tsp. vitamin C powder
½ tsp. brown sugar

Combine the ingredients in a clean salad dressing bottle. Shake. Refrigerate. The basic recipe is the oil and vinegar in a 2:1 ratio. After mixing these, add any pure spice desired. Or add fresh tomato chunks for creaminess.

Cheese Sauce

Add milk to cheese in equal amounts. Gradually heat to boiling while stirring. Add more of either to obtain the desired consistency. Boil 10 seconds. Use immediately.

Sour Cream-C

2 cups heavy whipping cream, previously boiled
¼ tsp. citric acid
¼ tsp. vitamin C powder
1 tsp. fresh onion juice or other seasoning (optional)

Stir until smooth, refrigerate 2 hours.

Yogurt

Buy a yogurt maker. Be sure and use boiled milk.

Soups

All home made soups are nutritious and safe, provided you use no processed ingredients (like bouillon), or make them in metal pots. Use herbs and aluminum-free salt to season. Always add a dash of vitamin C or tomato juice or vinegar to draw out calcium from soup bones for you to absorb.

Fish and Seafood recipe

Any kind of fish or seafood is acceptable, provided it is well-cooked. Don't buy food that is already in batter. The simplest way to cook fish is to poach it in milk. It can be taken straight from the freezer, rinsed, and placed in ¼ inch of milk (unboiled is fine) in the frying pan. Heat until it is cooked. Turn over and repeat. Throw away the milk. Serve with fresh lemon and herbs.

Baked Apples

Peel and core carefully. Remove all bruises (this is where the patulin is). Cut in bite-sized pieces, add a minimum of water and cook or bake minimally. Add a squirt of lemon juice when done. Serve with cinnamon, whipping cream and honey.

Ice creams

from the grocery store are loaded with benzene and other solvents. Fortunately there are ice cream makers that do everything (no cranking)! Or try our recipe which uses a blender. Be sure not to add store bought flavors, except vanilla or maple.

5 Minute Ice Cream

(Strawberry) Use 2 half pints of whipping cream, previously boiled, 1 package of frozen strawberries (about 10 oz.), and ½ cup clover honey. Pour frozen strawberries into blender. Pour whipping cream and honey over them. Blend briefly (about 10 seconds), not long enough to make butter! Pour it all into a large plastic bowl. Cover with a close fitting plastic bag and place in freezer. Prepare it a day ahead. Try using other frozen fruits, such as blueberries and cherries. Keep a few berries out of the blender and stir them in quickly with a non-metal spoon before setting the bowl in the freezer. There are many ice cream recipes to be found in old cook books. Avoid those with raw eggs or processed foods as ingredients. You may add nuts if you rinse them in vitamin C water.

Cookies, cakes and pies

Bake them from scratch, using unprocessed ingredients. Use simple recipes from old cook books.

Seven Day Sample Menu

Because processed foods have many toxins, you must cook as much from scratch as possible. So for convenience sake, keep your meals simple in preparation. You may want to prepare ahead and refrigerate your dressings and toppings. Or you could make a hot soup for dinner, refrigerate, and eat the leftovers for lunch. Don't save leftovers more than two days. Make sure they are covered. Try baking several potatoes at one time, refrigerate and put them in a salad the next night. Variety is the spice of life, so combine the allowed foods in the most creative ways you can imagine. And don't forget herbs and spices; learn to use them from old cook books.

	Breakfast	**Lunch**	**Dinner**
Day 1	Granola and honey with milk, half n' half or whipping cream 1 cup fresh squeezed fruit juice Water Milk	Fresh ground peanut butter and preserve sandwich Soup Milk Water	Orange roughy fish Fresh green beans with butter Baked potato with Sour Cream-C topping or fresh chives Pie (homemade) Milk 1 cup fresh squeezed or frozen vegetable juice Water

Day 2	Egg (limit is 2) Fried potatoes 1 glass milk Peppermint herb tea Fresh orange juice	Bagel (from bakery) Sour Cream-C Tomato 1 cup vegetable juice Water Milk	Homemade bean or lentil soup Sardines Dinner roll and butter Salad Homemade dressing Ice cream (homemade) Water
Day 3	Cream of Wheat™ cooked with raisins and milk Banana Peppermint herb tea ½ cup milk Water	Tuna sandwich with olives and butter Soup Milk Water	Baked sweet potato with butter and sweetening. Fresh broccoli with cheese sauce Bread and butter Chopped, peeled pear and whipping cream 1 cup vegetable juice Milk Water
Day 4	French toast with maple syrup Egg Homemade grapefruit juice Milk Water	Avocado and sour cream sandwich ½ cup vegetable juice Bread and butter Water	Lobster or sautéed shrimp Fresh asparagus Potatoes, any style ½ cup vegetable juice Water Milk
Day 5	Cooked cereal 1 glass milk Sliced banana with whipping cream and honey 1 glass water	Cold potato salad with C Dressing Soup ½ cup vegetable juice Custard Water Milk	1 can sardines or salmon in easy-open can (can openers shed metal) Salad of lettuce, tomato, olives, avocado with homemade dressing Bread with butter Ice-cream (optional) ½ cup vegetable juice Water Milk
Day 6	Egg and homemade biscuit with honey and butter Milk Fruit juice (homemade) Water	Homemade peanut butter sandwich ½ cup vegetable juice Milk Water	Gourmet pizza: home baked bread topped with olive oil, sliced tomato or homemade sauce, grated cheese, sardines or anchovies, chopped vegetables, garlic and onion Salad Milk Water

Day 7	Pancakes Banana or chopped fruit with cream Milk Water	Salmon sandwich (from flip top can) ½ cup vegetable juice Milk Water	Stir-fry vegetables: broccoli, carrots, cabbage, in olive oil and butter Bread and butter Pie (optional) Milk Water

Remember, take vitamin C and B-complex with each meal.

Too Sick To Cook, Too Tired To Eat

Pick three meals from the sample menu that need no cooking and eat them every day.

Recipes for Natural Body Products

You can use just borax (like 20 Mule Team Borax™) and washing soda (like Arm & Hammer Super Washing Soda™) for all types of cleaning including your body, laundry, dishes and your house! You don't need all of those products you see in commercials for each special task!

Even if you have dry skin, difficult hair or some other unique requirement, just pure borax will satisfy these needs. A part of every skin problem is due to the toxic elements found in the soaps themselves. For instance aluminum is commonly added as a “skin moisturizer”. It does this by impregnating the skin and attracting water, giving the illusion of moist skin. In fact you simply have moist aluminum stuck in your skin which your immune system must remove. While borax won't directly heal your skin or complexion, it does replace the agents that are causing damage, so that healing can occur.

Borax Liquid Soap

Empty 1 gallon jug
1/8 cup borax powder
Plastic funnel

Funnel the borax into the jug, fill with cold tap water. Shake a few times. Let settle. In a few minutes you can pour off the clear part into dispenser bottles. This is the soap!

Easier way: use any bottle, pour borax powder to a depth of a ½ inch or so. Add water. Shake. When you have used it down to the undissolved granules, add more water and shake again. Add more borax when the undissolved granules get low.

Keep a dispenser by the kitchen sink, bathroom sink, and shower. It does not contain aluminum as regular detergents and soaps do, and which probably contribute to Alzheimer's disease. It does not contain PCBs as many commercial and health food varieties do. It does not contain cobalt (the blue or green granules) which causes heart disease and draws cancer parasites to the skin. Commercial detergents and non-soaps are simply not safe. Switch to homemade bar soap and borax for all your tasks! Borax inhibits the bacterial enzyme *urease* and is therefore antibacterial. It may even clear your skin of blemishes and stop your scalp from itching.

For Laundry

Borax (½ cup per load). It is the main ingredient of non-chlorine bleach and has excellent cleaning power without fading colors. Your regular laundry soap may contain PCBs, aluminum, cobalt and other chemicals. These get rubbed into your skin constantly as you wear your clothing. For bleaching (only do this occasionally) use original chlorine bleach (not “new improved” or “with special brighteners”, and so forth). Don't use chlorine if there is an ill person in the house. For getting out stubborn dirt at collars, scrub with homemade bar soap first; for stains, try grain alcohol, vinegar, baking soda.

For Dishes

Don't believe your eyes when you see the commercials where the smiling person pulls a shining dish out of greasy suds. Any dish soap that you use should be safe enough to eat because nothing rinses off clean. Regular dish detergents, including health brands, are now polluted with PCBs. They also contain harmful chemicals. Use borax for your dishes. Or use paper plates and plastic (not styrofoam) cups.

In The Dishwasher

Use 2 tsp. borax powder pre-dissolved in water. If you use too much it will leave a film on your dishes. Use vinegar in the rinse cycle.

In The Sink

Use a dishpan in the sink. Use ¼ cup borax and add a minimum of water. Also keep a bit of dry borax in a saucer by the sink for scouring. Don't use any soap at all for dishes that aren't greasy and can be washed under the faucet with nothing but running water. Throw away your old sponge or brush or cloth because it may be PCB contaminated. Start each day by sterilizing your sponge (it harbors Salmonella) or with a new one while the used one dries for three full days. Clean greasy pots and pans with a paper towel first. Then use homemade bar soap.

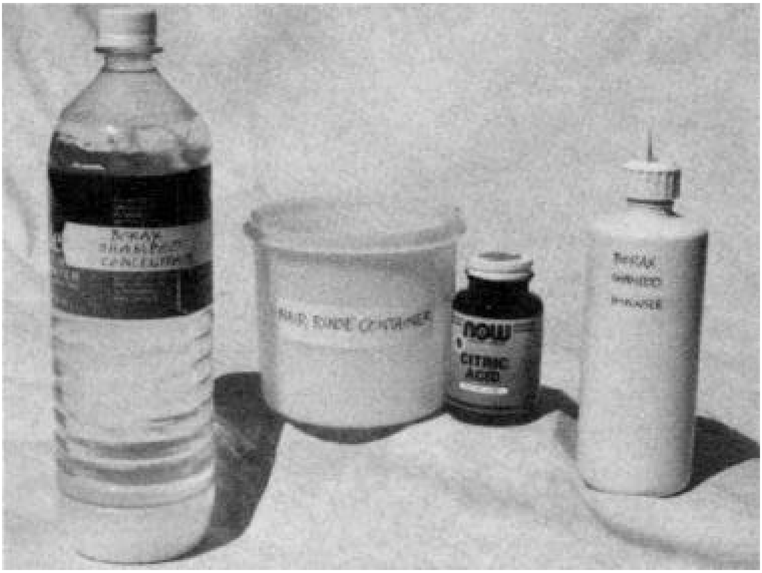

Fig. 88 Make a bottle of borax liquid to fill your soap dispensers and shampoo bottle. Use citric acid to rinse and condition.

Shampoo

Borax liquid is ready to use as shampoo, too. It does not lather but goes right to work removing sweat and soil without stripping your color or natural oils. It inhibits scalp bacteria and stops flaking and itching. Hair gets squeaky clean so quickly (just a few squirts does it) that you might think nothing has happened! You will soon be accustomed to non-lathery soap. Rinse very thoroughly because you should leave your scalp slightly acidic. Take a pint container to the shower with you. Put ¼ tsp. citric (not ascorbic) acid crystals (see *Sources*) in it. For long hair use a quart of rinse. Only **citric acid** is strong enough to get the borax out, lemon juice and vinegar are not. After shampooing, fill the container with water and rinse. Rinse your whole body, too, since citric acid is also anti-bacterial. All hair shampoo penetrates the eye lids and gets into the eyes although you do not feel it. It is important to use this natural rinse to neutralize the shampoo in your eyes. (Some people have stated that citric acid makes their hair curlier or reddens it. If this is undesirable, use only half as much citric acid.) Citric acid also conditions and gives body and sheen to hair.

Hair Spray

I don't have a recipe that holds your hair as well as the bottle of chemicals you can buy at the store. Remarkably a little lemon juice (not from a bottle) has some holding power and no odor! Buy a 1 cup spray bottle. Squeeze part of a lemon, letting only the clear juice run into the bottle. Fill with water. Keep it in the refrigerator. Make it fresh every week. Spraying with just plain water is nearly as good! For shinier hair, drop a bit of lemon peel into the bottle.

Homemade Soap

A small plastic dishpan, about 10” x 12”
A glass or enamel 2-quart sauce pan
1 can of lye (sodium hydroxide), 12 ounces
3 pounds of lard (BHT and BHA are OK here)
Plastic gloves
Water

1. Pour 3 cups of very cold water (refrigerate water overnight first) into the 2-quart saucepan.
2. Slowly and carefully add the lye, a little bit at a time, stirring it with a wooden or plastic utensil. (Use plastic gloves for this; test them for holes first.) Do not breathe the vapor or lean over the container or have children nearby. Above all use no metal. The mixture will get very hot. In olden days, a sassafras branch was used to stir, imparting a fragrance and insect deterrent for mosquitoes, lice, fleas, ticks.

3. Let cool at least one hour in a safe place. Meanwhile, the unwrapped lard should be warming up to room temperature in the plastic dishpan.
4. Slowly and carefully, pour the lye solution into the dishpan with the lard. The lard will melt. Mix thoroughly, at least 15 minutes, until it looks like thick pudding.
5. Let it set until the next morning; then cut it into bars. It will get harder after a few days. Then package.

If you wish to make soap based on olive oil, use about 48 ounces. It may need to harden for a week.

Liquid Soap

Make chips from your homemade soap cake. Add enough hot water to dissolve. Add citric acid to balance the pH (7 to 8). If you do not, this soap may be too harsh for your skin.

Skin Sanitizer

Make up a 5 to 10% solution of food grade alcohol. Food grade alcohols are grain (ethyl) alcohol or vodka. Find a suitable dispenser bottle. Mark it with a pen at about one tenth of the way up from the bottom. Pour 95% grain alcohol (190 proof) to this mark (for 50% grain alcohol or vodka make your mark one fifth of the way up). Add water to the top. Keep shut. You may add a chip of lemon peel for fragrance.

Use this for general sanitizing purposes: bathroom fixtures, knobs, handles, canes, walkers, and for personal cleanliness (but use chlorine bleach for the toilet bowl once a week). Always clean up after a bowel movement with wet toilet paper. This is not clean enough, though. Follow with a stronger damp paper towel. This is still not clean enough; use a final damp paper towel with skin sanitizer added. After washing hands, sanitize them too, pouring a bit on one palm and put finger tips of the other hand in it, scratch to get under nails, repeat on other hand. Rinse with water.

Do not use this recipe, nor keep any bottles of alcohol in the house of a recovering alcoholic.

Deodorant

Your sweat is odorless. It is the entrenched bacteria feeding on it that make smells. You can never completely rid yourself of these bacteria, although they may temporarily be gone after zapping. The strategy is to control their numbers. Here are several deodorants to try. Find one that works best for you:

Vitamin C water. Mix ¼ tsp. to a pint of water and dab it on. Then dab on cornstarch.

Citric Acid water. Mix ¼ tsp. to a pint of water and dab it on. Then dab on cornstarch.

Only a few drops of these acids under each armpit are necessary. If these acids burn the skin, dilute them more. Never apply anything to skin that has just been shaved!

Corn starch. Many people need only this. Dab it on.

Use only unpolluted **cornstarch** (see *Sources*).

Baking soda has been deleted as a deodorant because benzene was found in some boxes.

Lemon juice. This acid is not as strong, use what you need.

Pure alcohol (never rubbing alcohol). The food grade alcohols are grain alcohol and vodka. Dab a bit under each arm and/or on your shirt or blouse, then dab on cornstarch. If the alcohol burns, dilute it with water. Be very careful not to leave the bottle where a child or alcoholic person could find it. Pour it into a different bottle!

Pure zinc oxide. You may ask your pharmacist to order this for you. She or he may wish to make it up for you too, but do not let them add anything else to it. It should be about 1 part zinc oxide powder to 3 parts water. It does not dissolve. Just shake it up to use it. After you get it home, you can add cornstarch to it to give it a creamy texture. Heat 3 tsp. cornstarch in 1 cup of water, to boiling, until dissolved and clear. Cool and add some to the zinc oxide mixture (about equal parts). Store unused starch mixture in the refrigerator. Only make up enough for a month.

Alcohol and zinc oxide. This is the most powerful deodorant. Apply alcohol first, then the zinc oxide, then dab on cornstarch.

Remember that you need to sweat! Sweating excretes toxic substances, especially from the upper body. Don't use deodorant on weekends. Go to the sink and wipe clean the armpits like our grandparents did. Often, just plain cornstarch is enough! These homemade deodorants are not as powerful as the commercial varieties–this is to your advantage.

Brushing Teeth

Buy a new toothbrush. Your old one is soaked with toxins from your old toothpaste. Use only water or chemically pure baking soda if you have any metal fillings. Put a pinch in a glass, add water to dissolve it. Use food-grade hydrogen peroxide (see *Sources*) if you have only plastic fillings. Dilute it from 35% to 17½% by adding water (equal parts). Store hydrogen peroxide only in polyethylene or the original plastic bottle. Use 4 or 5 drops on your toothbrush. It should fizz nicely as oxygen is produced in your mouth. Your teeth will whiten noticeably in 6 months. Before brushing teeth, floss with 4 or 2 pound monofilament fish line. Double it and twist for extra strength. Rinse before use. Floss and brush only once a day. If this leaves you uncomfortable, brush the extra times with plain water and a second “water-only” toothbrush. Make sure that nothing solid, like powder, is on your toothbrush; it will scour the enamel and give you sensitive teeth, especially as you get older and the enamel softens. Salt is corrosive—don't use it for brushing metal teeth. Plain water is just as good.

For Dentures

Use salt water. It kills all germs and is inexpensive. Salt water plus grain alcohol or food-grade hydrogen peroxide makes a good denture-soak.

Mouthwash

A few drops of food grade hydrogen peroxide added to a little water in a glass should be enough to make your mouth foam and cleanse. Don't use hydrogen peroxide, though, if you have metal fillings, because they react. Don't use regular drug store variety hydrogen peroxide because it contains toxic additives. Health food store varieties contain solvents from the bottling process. See *Sources*. Never purchase hydrogen peroxide in a bottle with a metal cap.

For persons with metal tooth fillings, use chemically pure baking soda or just plain hot water. A healthy mouth has no odor! You shouldn't need a mouthwash! If you have breath odor, search for a hidden tooth infection or cavitation.

Contact Lens Solution

A scant cup of cold tap water brought to a boil in glass saucepan. After adding ¼ tsp. aluminum free salt and boiling again, pour into a sterile canning jar. Refrigerate. Freeze some of it.

Lip Soother

For dry, burning lips. Heat 1 level tsp. sodium alginate plus 1 cup water until dissolved. After cooling, pour it into a small bottle to carry in your purse or pocket (refrigerate the remainder). Dab it on whenever needed. If the consistency isn't right for you, add water or boil it down further. You can make a better lip soother by adding some lysine from a crushed tablet, vitamin C powder, and a vitamin E capsule to the alginate mix. If you have a persistent problem with chapped lips, try going off citrus juice.

Foot Powder

Use a mixture of cornstarch and zinc oxide poured into a salt shaker with a lid. Add long rice grains to fight humidity. You may also try arrow root or potato starch. If you don't have zinc oxide use plain cornstarch.

Skin Healer Moisturizer Lotion

1 tsp. sodium alginate
1 cup water

Make the base first by heating these together in a covered, non-metal pan until completely dissolved. Use low heat–it will take over an hour. Use a wooden spoon handle to stir. Set aside. Then make the following mixture:

¼ tsp. Vitamin C (ascorbic acid) (You may crush tablets)
¼ tsp. lysine (crush tablets)
2 tbs. pure vegetable glycerin
2 vitamin E capsules (400 units or more, each)
1 tsp. apricot kernel oil (olive oil will do)

1 tbs. lemon juice from a lemon or ¼ tsp. citric acid (this is optional)
1 cup water

Heat the water to steaming in a non-metal pan. Add vitamin C and lysine first and then everything else. Pour into a pint jar and shake to mix. Then add the sodium alginate base to the desired thickness (about equal amounts) and shake. Pour some into a small bottle to use as lip soother. Pour some into a larger bottle to dispense on skin. Store remainder in refrigerator. (See *Sources* for sodium alginate, vegetable glycerin and apricot kernel oil. Sodium alginate is also available in capsule form at some health food stores.)

Other Skin Healers

Vitamin C powder (ascorbic acid, not the same as citric acid). Put a large pinch into the palm of your hand. With your other hand pick up a few drops of water from the faucet. Rub hands together until all the powder is dissolved and dispensed. It may sting briefly. Do this at bedtime, especially for cracked, chapped hands. Include lips if they need it.

Vitamin E oil. Vitamin E oil from Now Foods was not polluted at the time of this writing, but for the future it would be safer to rely on capsules. Snip open a capsule and rub into skin.

50% Glycerin. Dilute 100% vegetable glycerin with an equal amount of water. This is useful as an after shave lotion.

Vitamin C liquid. Mix ¼ tsp. vitamin C powder in one pint water (crushed tablets will do). This is useful as an after shave lotion and general skin treatment.

Apricot Kernel Oil. This is a very light oil, useful as an after shave lotion and general skin treatment.

Cornstarch (see *Sources*). Use on rashes, fungus, moist or irritated areas and to prevent chafe.

Combining several of these makes them more effective.

Dry skin has several causes: too much water contact, too much soap contact (switch to borax), low body temperature, not enough fat in the diet, or parasites.

Massage Oil

Use olive oil. It comes in very light to heavy textures. Pick the right one for your purpose. Alginate mixtures can be used instead of, or added to, oil. Starch solutions are good, too.

Sunscreen Lotion

Purchase PABA (see *Sources*) in 500 mg tablet form. Dissolve 1 tablet in grain alcohol or vodka. Grind the tablet first by putting it in a plastic bag and rolling over it with a glass jar. It will not completely dissolve even if you use a tablespoon of the alcohol. Pour the whole mixture into a 4 ounce bottle of homemade skin softener. Be careful not to get the lotion into your eyes when applying it. A better solution is to wear a hat or stay out of the sun. Remember to *take* PABA as a supplement, too (500 mg, one a day).

Nose Salve

(When the inside of the nose is dry, cracked and bleeding.)

Pour ½tsp. pure vegetable glycerin into a bottle cap. Add ½ tsp. of water.

Applicator: use a plastic coffee stirrer or straw; cut a slit in the end to catch some cotton wool salvaged from a vitamin bottle and twist (cotton swabs, cotton balls and wooden toothpicks are sterilized with mercury which in turn is polluted with thallium). Dip it into the glycerin mixture and apply inside the nose with a rotating motion. Do each nostril with a new applicator.

Quick Corn Starch Skin Softener

4 tsp. corn starch (see *Sources*)
1 cup water

Boil starch and water until clear, about one minute.

Cornstarch Skin Softener

1 tsp. lysine powder or 8 tablets, 500 mg each
1 tsp. Vitamin C powder (ascorbic acid); or 8 tablets, 500 mg each
3 tsp. cornstarch (see *Sources*)
Vitamin E, 1 capsule 400 mg
¼ tsp. apricot kernel oil (optional)
1 cup water

Boil starch and water until clear, about one minute. Add other ingredients and stir until dissolved. Cool. Pour into dispenser bottle. Keep refrigerated when not in use. Apply after washing dishes and after showering.

After Shaves

Vitamin C. ¼ tsp. vitamin C powder, dissolved in 1 pint water.

Apricot kernel oil.

Vegetable glycerin. Equal parts glycerin and water or to suit your need.

Personal Lubricants

Heat these together: 1 level tsp. sodium alginate and 1 cup water in a covered non-metal pan until completely dissolved. Use very low heat and stir with a wooden spoon handle. It takes a fairly long time to get it perfectly smooth. After cooling, pour into a small dispenser bottle. Keep the remainder refrigerated.

Or, mix and heat 4 tsp. cornstarch and 1 cup water until completely dissolved in a covered saucepan. Use non-metal dishes and a non-metal stirring spoon. Cool. Pour some into dispenser bottle. Refrigerate remainder. This is many person's favorite recipe.

Baby Wipes

Cut paper towels in quarters and stack in a closable plastic box. Run tap water over them, drain the excess. Add 1 tsp. grain alcohol and/or borax liquid on top. Close. Put a dab of the Quick Cornstarch Softener recipe on top of each wipe as you use it.

People Wipes

¼ tsp. powdered lysine (you may crush tablets)
¼ tsp. Vitamin C powder (you may crush tablets)
¼ cup vegetable glycerin
1 cup water

Prepare wipes by cutting paper towels in quarters. Use white, unfragranced towels that are strong enough to hold up for this use. Fold each piece in quarters again and stack in a plastic zippered baggy. Pour the fluid mixture over the stack and zip. Store a bag full in the freezer to take on car trips. If you want to keep them a month or more, add 1 tbs. grain alcohol or vodka to the recipe.

For bathroom use, dampen a roll of paper towels under the cold tap first. Then pour about ¼ cup of the mixture over the towel roll around the middle. Store in plastic shopping bag or stand in plastic waste basket.

Recipes For Natural Cosmetics

Eye liner and Eyebrow Pencil

Get a pure charcoal pencil (black only) at an art supply store. Try several on yourself (bring a small mirror) in the store to see what hardness suits you. You may need to wet it with water or a vitamin E perle first. Don't put any chemicals on your *eyelids*, since this penetrates into your eye. To check this out for yourself, close your eye tightly and then dab lemon juice on your eyelid. It will soon burn! Everything that is put on skin penetrates. Otherwise the nicotine patch and estrogen patch wouldn't work. Not even soap belongs on your eyelids! Charcoal pencils are cheap. Get yourself half a dozen different kinds so you can do different things.

You could also use a capsule of activated charcoal. Empty it into a saucer. Mix glycerin and water, half and half, and add it to the charcoal powder until you get the consistency you like. Use a brush for eyelashes; use a finger for eyebrows.

Lipstick

Beet root powder (see *Sources*)
100% vegetable glycerin

Combine 1 tsp. vegetable glycerin and 1 tsp. beet root powder in a saucer. Stir until perfectly smooth. Then add ½tsp. of vitamin E oil. Snip open vitamin E capsules or buy vitamin E oil (see *Sources*). Very thick olive oil can be substituted. Apply liberally with your finger or a lipstick brush. Do not purse or rub your lips together after application. To make the lipstick stay on longer, apply 1 layer of lipstick, then dab some corn starch over the lips, then apply another layer of lipstick. Store in a small glass or plastic container in the refrigerator, tightly covered in a plastic bag.

Face Powder

Use cornstarch from the original box. You may also try arrow root starch or potato starch. Use your fingers or a tissue to apply because applicators can carry bacteria.

Blush (face powder in a cake form)

Add 50% glycerin to cornstarch in a saucer to make a paste. Slowly add beet root powder to the desired color. Use part of a charcoal capsule to darken it, if desired. A drop of food grade alcohol will also darken it. To make 50% glycerin, add equal parts of glycerin and water. Try to make the consistency the same as your brand name product, and you can even put it back in your brand name container.

Recipes For Household Products

Floor Cleaner

Use washing soda from the grocery store. You may add borax and boric acid (to deter insects except ants). Use white distilled vinegar in your rinse water for a natural shine and ant repellent. Do not add bleach to this. For the bathroom floor use plain bleach water—follow the label. Never use chlorine bleach if anybody in the home is ill or suffers from depression. Use grain alcohol (1 pint to 3 quarts water) for germ killing action instead of chlorine.

Furniture Duster and Window Cleaner

Mix equal parts white distilled vinegar and water. Put it in a spray bottle.

Furniture Polish

A few drops of olive oil on a dampened cloth. Use filtered water to dampen.

Insect Killer

Boric acid powder (not borax). Throw liberal amounts behind stove, refrigerator, under carpets and in carpets. Since boric acid is white, you must be careful not to mistake it for sugar accidentally. Keep it far away from food and out of children's reach. Buy it at a farm supply or garden store (or see *Sources*). It will not kill ants.

Ant Repellent

Spray 50% white distilled vinegar on counter tops, window sills and shelves and wipe, leaving residue. Start early in spring before they arrive, because it takes a few weeks to rid yourself of them once they are established. If you want immediate action, get some lemons, cut the yellow outer peel off and cover with grain alcohol in a tightly closed jar. Let stand at least one hour. Use 1 part of this concentrate with 9 parts water in a spray bottle. Mix only as much as you will use because the diluted form loses potency. Spray walls, floors, carpets wherever you see them. The lemon solution even leaves a shine on your counters. Use both vinegar and lemon approaches to rid yourself of ants.

To treat the **whole house**, pour vinegar all around your foundation, close to the wall, using one gallon for every five feet. Expect to damage any foliage it touches. Reapply every six months.

Flower and Foliage Spray

Food-grade hydrogen peroxide. See instructions on bottle.

Moth Balls

I found this recipe in an old recipe book. Mix the following and scatter in trunks and bags containing furs and woolens: ½lb. each rosemary and mint, ¼ lb. each tansy and thyme, 2 tbs. powdered cloves.

Carpet Cleaner

Whether you rent a machine or have a cleaning service, don't use the carpet shampoo they want to sell, even if they "guarantee" that it is all natural and safe. Instead add these to a bucket (about four gallons) of water and use it as the cleaning solution:

Wash water	Rinse water
1/3 cup borax	¼ cup grain alcohol
	2 tsp. boric acid
	¼ cup white distilled vinegar **or**
	4 tsp. citric acid

Borax does the cleaning; alcohol disinfects, boric acid leaves a pesticide residue, and the vinegar or citric acid give luster. If you are just making one pass on your carpet, use the borax, alcohol, and boric acid. Remember to test everything you use on an unnoticed piece of carpet first.

Health Improvement Recipes

Black Walnut Hull Tincture

This new recipe is four times as strong as the previous one, so it is called **Black Walnut Hull Tincture Extra Strength**.

Your largest enamel or ceramic (not stainless steel, not aluminum) cooking pot, preferably at least 10 quarts
Black walnuts, in the hull, each one still at least 50% green, enough to fill the pot to the top
Grain alcohol, about 50% strength, enough to cover the walnuts
½ tsp. vitamin C
Plastic wrap or cellophane
Glass jars or bottles

The black walnut tree produces large green balls in fall. The walnut is inside, but we will use the whole ball, uncracked, since the active ingredient is in the green outer hull.

Rinse the walnuts carefully, put them in the pot, and cover with the alcohol. Sprinkle on half the vitamin C. Seal with plastic wrap and cover. Let sit for three days. Pour into glass jars or bottles, discarding walnuts, and divide the remaining vitamin C amongst the jars. If the glass jar has a metal lid, first put plastic wrap over the top before screwing on the lid. Potency is strong for several years if unopened, even if it darkens.

You have just made Extra Strength Black Walnut Hull Tincture. It is stronger than the concentrate made with just a few black walnuts in a quart jar (my earlier recipe), because there are more walnuts per unit liquid. In addition, you will not dilute it before use (although when you take it, it will usually be in water).

When preparing the walnuts, rinse only with cold tap water. You may need to use a brush on areas with dirt. If you are not going to use all of them in this batch, you may freeze them in a resealable plastic bag. Simply refrigerating them does not keep them from turning black and useless. The pot of soaking walnuts should not be refrigerated. Nor does the final tincture need any refrigeration.

Exposure to air does cause the tincture to darken and lose potency. To reduce air exposure, fill the pot as much as possible, without touching the plastic wrap, while still keeping a snug fitting lid. Even more importantly, the glass jars or bottles you use to store your tincture should have as little air space as possible, without touching the plastic wrap on top. A large jar should be divided into smaller ones when you are ready to use it. The idea is not to have partial jars, with a lot of air space, sitting for longer than a month or so.

There are several ways to make a 50% grain alcohol solution. Some states have Everclear,™ 95% alcohol. Mix this half and half with water. Other states have Everclear that is 76.5% alcohol. Mix this three parts Everclear to one part water. Yet another method is to buy vodka that is 100 proof. This is already 50% alcohol.

Remember, never use any kind of purchased water to make tincture.

Black Walnut Hull Tincture (Regular Strength)

This is the potency I used originally. It is included here in case you prefer it or wish to treat a pet. The Extra Strength recipe is four times as potent as the original recipe, so it must be diluted in quarters. (Similarly, if you have a lot of the Regular Strength left and want to use it in place of Extra Strength, simply take four times as much.)

Black Walnut Hull Tincture Extra Strength
Grain Alcohol, about 10%

Mix one part extra strength tincture with three parts of the 10% alcohol. Store in glass containers same as described above.

There are several ways to make a 10% grain alcohol solution. Some states have Everclear,™ 95% alcohol. Mix this one part Everclear to nine parts water. Other states have Everclear that is 76.5% alcohol. Mix this one part Everclear to seven parts water. Yet another method is to buy vodka that is 100 proof (50% alcohol) and mix one part vodka with four parts water.

Black Walnut Hull Extract (Water Based)

Because you do not know how commercially available extracts were made, and may not be able to test for solvent pollution, it is wisest to make it yourself!

This recipe is intended for alcoholic persons: cover the green balls in the 10 quart (non-metal) pot with cold tap water. Heat to boiling, covered. Turn off heat. When cool, add vitamin C, cover with plastic wrap, and the lid. Let stand for 1 day. It will be darker than the tincture. Do not dilute. Pour into freezable containers. Refrigerate what you will use in two days and freeze the rest. Add vitamin C after thawing or during refrigeration (¼ tsp. per quart).

For use: in programs calling for Extra Strength Black Walnut Hull Tincture use four times as much of this water based recipe (8 tsp. instead of 2 tsp. Extra Strength).

Important Note: do not use bottled or purchased water to make this tincture or you could pollute it with benzene!

Quassia recipe

Add 1/8 cup quassia chips to 3 cups water. Simmer 20-30 minutes. Pour off 1/8 cup now and drink it fresh. Refrigerate remainder. Drink 1/8 cup 4 times/day, until a total of ½cup of chips is consumed. Flavor with spices.

Emmenagogue (Menstrual Period Inducer)

Here are four herbs that can each bring on your period. They can be started anytime but the most-effective time is before your next calculated period time (count days as if you never missed a period).

1 oz sassafras bark
1 oz rue (cut)
1 oz marjoram herb
1 oz blue cohosh root
4½ cups boiling water

Add the herbs to the boiling water and turn down to simmer, covered, for 20 minutes. Do not boil. Strain and refrigerate in sterile glass jar. Pour one cup for yourself in the morning. Let warm to room temperature, and sip between meals, making it last until supper.

Bowel Program

Bacteria are always at the root of bowel problems, such as pain, bloating and gassiness. They can not be killed by zapping, because the high frequency current does not penetrate the bowel contents.

Although most bowel bacteria are beneficial, the ones that are not, like *Salmonellas* and *Shigellas*, are extremely detrimental because they have the ability to invade the rest of your body and colonize a trauma site or weakened organ. These same bacteria colonize a cancer tumor and delay healing after the malignancy is stopped.

Another reason bowel bacteria are so hard to eradicate is that we are constantly reinfecting ourselves by keeping a reservoir on our hands and under our fingernails.

- So the first thing to do is **improve sanitation**. For a serious problem, use 50% grain alcohol (100 proof vodka) in a spray bottle at the bathroom sink. Sterilize your hands after bathroom use and before meals.
- Secondly use **turmeric** (2 capsules 3 times a day, this is the common spice) which I find helps against *Shigella*, as well as *E. coli*. Expect orange colored stool.
- Third use **fennel** (1 capsule 3 times a day).
- Fourth use **digestive enzyme tablets** with meals as directed on the bottle. (But only as long as necessary, because these frequently harbor molds.)
- Fifth use a single 2 tsp. dose of **Black Walnut Hull Tincture Extra Strength**. Add it to a ½glass of water and sip over a 15 minute period. Stay seated until any side effect from the alcohol wears off.
- Sixth take **Cascara sagrada** capsules if constipated (start with one capsule a day, use up to maximum on label). Remember to drink a cup of hot water upon rising in the morning. This will begin to regulate your elimination.

It can take all six to get rid of a bad *Shigella* problem in a week. Afterward, you must continue to eat only sterile dairy products. Note that the Kidney Cleanse is often effective with bowel problems. Try it also.

You will know you succeeded when your tummy is flat, there is not a single gurgle, and your mood improves!

Constipation Tea

Constipation is often caused by *E. coli* and *Salmonella* from dairy foods, or from killing "good" bowel bacteria with antibiotics (killing a few by zapping actually restores good flora). Eat foods that restore the body's good bowel flora: vegetables, sterilized milk (the milk sugar is essential), lots of water.

There are a lot of remedies for constipation, but many people enjoy this tea:

1 tbs. senna tea leaves
½ tsp. mint leaves

Boil for one minute in a quart of water, add a dash of vitamin C and brown sugar to taste. Sip through the day to avoid "bellyache". It can take years for the body's flora to "right themselves" after an antibiotic session, be patient.

Weight Reduction

Here are two ancient herbal recipes for obesity. I have not personally determined their effectiveness.

Fucus

2 oz *Fucus vesiculosus*, cut (see Sources)
3 cups cold tap water

Boil for 15 minutes, covered. Cool. Dose: ¼ cup four times a day on an empty stomach. After one week increase dose to ½cup. You may add any flavoring desired.

Watch the pot carefully as it comes to a boil. If it boils over, you will have a month of stove-cleaning to do. The odor of *Fucus* boiling is wretched. So is the taste. Maybe garlic (fresh) would improve it.

Fennel

1 oz fennel seed (crushed or powdered is fine)

3 cups cold tap water

Boil water, pour over herb. Steep 30 minutes. Strain. Add 4 oz honey (optional). Drink one cup each day.

You could take them both together, along with the Bowel Program, to be more successful, but the best single weight reducer is the Liver Cleanse.

Kidney Cleanse

½ cup dried Hydrangea root
½ cup Gravel root
½ cup Marshmallow root
4 bunches of fresh parsley
Goldenrod tincture (leave this out of the recipe if you are allergic to it)
Ginger capsules
Uva Ursi capsules
Vegetable glycerin
Black Cherry Concentrate, 8 oz
Vitamin B6, 250 mg
Magnesium oxide tablets, 300 mg

Measure ¼ cup of each root and set them to soak, together in 10 cups of cold tap water, using a non-metal container and a non-metal lid (a dinner plate will do). After four hours (or overnight) add 8 oz. black cherry concentrate, heat to boiling and simmer for 20 minutes. Drink ¼ cup as soon as it is cool enough. Pour the rest through a bamboo strainer into a sterile pint jar (glass) and several freezable containers. Refrigerate the glass jar.

Boil the fresh parsley, after rinsing, in 1 quart of water for 3 minutes. Drink ¼ cup when cool enough. Refrigerate a pint and freeze 1 pint. Throw away the parsley.

Dose: each morning, pour together ¾cup of the root mixture and ½cup parsley water, filling a large mug. Add 20 drops of goldenrod tincture and 1 tbs. of glycerin. Drink this mixture in divided doses throughout the day. Keep cold. Do not drink it all at once or you will get a stomach ache and feel pressure in your bladder. If your stomach is very sensitive, start on half this dose.

Save the roots after the first boiling, storing them in the freezer. After 13 days when your supply runs low, boil the same roots a second time, but add only 6 cups water and simmer only 10 minutes. This will last another 8 days, for a total of three weeks. You may cook the roots a third time if you wish, but the recipe gets less potent. If your problem is severe, only cook them twice.

After three weeks, repeat with fresh herbs. You need to do the Kidney Cleanse for six weeks to get good results, longer for severe problems.

Also take:

- Ginger capsules: one with each meal (3/day).
- Uva Ursi capsules: one with breakfast and two with supper.
- Vitamin B6 (250 mg): one a day.
- Magnesium oxide (300 mg): one a day.

Take these supplements just before your meal to avoid burping.

Some notes on this recipe: this herbal tea, as well as the parsley, can easily spoil. Heat it to boiling every fourth day if it is being stored in the refrigerator; this resterilizes it. If you sterilize it in the morning you may take it to work without refrigerating it (use a glass container).

When you order your herbs, be careful! Herb companies are not the same! These roots should have a strong fragrance. If the ones you buy are barely fragrant, they have lost their active ingredients; switch to a different supplier. Fresh roots can be used. Do not use powder.

- Hydrangea (*Hydrangea arborescens*) is a common flowering bush.
- Gravel root (*Eupatorium purpureum*) is a wild flower.
- Marshmallow root (*Althea officinalis*) is mucilaginous and kills pain.
- Fresh parsley can be bought at a grocery store. Parsley flakes and dried parsley herb do not work.
- Goldenrod herb works as well as the tincture but you may get an allergic reaction from smelling the herb. If you know you are allergic to this, leave this one out of your recipe.
- Ginger from the grocery store works fine; you may put it into capsules for yourself (size 0, 1 or 00).

There are probably dozens of herbs that can dissolve kidney crystals and stones. If you can only find several of those in the recipe, make the recipe anyway; it will just take longer to get results. Remember that vitamin B_6 and magnesium, taken daily, can prevent oxalate stones from forming. But only if you stop drinking tea. Tea has 15.6 mg oxalic acid per cup[37]. A tall glass of iced tea could give you over 20 mg oxalic acid. Switch to herb teas. Cocoa and chocolate, also, have too much oxalic acid to be used as beverages.

Remember, too, that phosphate crystals are made when you eat too much phosphate. Phosphate levels are high in meats, breads, cereals, pastas, and carbonated drinks. Eat less of these, and increase your milk (2%), fruits and vegetables. Drink at least 2 pints of water a day.

---

[37] Taken from *Food Values* 14ed by Pennington and Church 1985.

Cleanse your kidneys at least twice a year.

You can dissolve all your kidney stones in 3 weeks, but make new ones in 3 days if you are drinking tea and cocoa and phosphated beverages. None of the beverage recipes in this chapter are conducive to stone formation.

Liver Herbs

Don't confuse these liver herbs with the next recipe for the Liver Cleanse. This recipe contains herbs traditionally used to help the liver function, while the Liver Cleanse gets gallstones out.

6 parts comfrey root, *Symphytum officinale* (also called nipbone root)
6 parts tanner's oak bark, *Quercus alba* (white oak bark)
3 parts gravel root, *Eupatorium purpureum* (queen of the meadow)
3 parts Jacob's staff, *Verbascum thapsus* (mullein herb)
2 parts licorice root, *Glycyrrhiza glabra*
2 parts wild yam root, *Dioscorea villosa*
2 parts milk thistle herb, *Silybum marianum*
3 parts walnut bark, *Juglans nigra*, (black walnut bark)
3 parts marshmallow root, *Althea officinalis* (white mallow)
1 part lobelia plant, *Lobelia inflata* (bladder pod)
1 part skullcap, *Scutellaria lateriflora* (helmet flower)

Mix all the herbs. Add ½cup of the mixture to 2 quarts of water. Bring to a boil. Put lid on. Let sit for six hours. Strain and drink 1½cups per day. Put the strained herbs in the freezer and use them one more time.

Liver Cleanse

Cleansing the liver of gallstones dramatically improves digestion, which is the basis of your whole health. You can expect your allergies to disappear, too, more with each cleanse you do! Incredibly, it also eliminates shoulder, upper arm, and upper back pain. You have more energy and increased sense of well being.

Cleaning the liver bile ducts is the most powerful procedure that you can do to improve your body's health.

But it should not be done before the parasite program, and for best results should follow the kidney cleanse and any dental work you need.

It is the job of the liver to make bile, 1 to 1½quarts in a day! The liver is full of tubes (*biliary tubing*) that deliver the bile to one large tube (the *common bile duct*). The gallbladder is attached to the common bile duct and acts as a storage reservoir. Eating fat or protein triggers the gallbladder to squeeze itself empty after about twenty minutes, and the stored bile finishes its trip down the common bile duct to the intestine.

For many persons, including children, the biliary tubing is choked with gallstones. Some develop allergies or hives but some have no symptoms. When the gallbladder is scanned or X-rayed nothing is seen. Typically, they are not in the gallbladder. Not only that, most are too small and not calcified, a prerequisite for visibility on X-ray. There are over half a dozen varieties of gallstones, most of which have cholesterol crystals in them. They can be black, red, white, green or tan colored. The green ones get their color from being coated with bile. Notice in the picture how many have imbedded unidentified objects. Are they fluke remains? Notice how many are shaped like corks with longitudinal grooves below the tops. We can visualize the blocked bile ducts from such shapes. Other stones are composites–made of many smaller ones–showing that they regrouped in the bile ducts some time after the last cleanse.

At the very center of each stone is found a clump of bacteria, according to scientists, suggesting a dead bit of parasite might have started the stone forming.

As the stones grow and become more numerous the back pressure on the liver causes it to make less bile. Imagine the situation if your garden hose had marbles in it. Much less water would flow, which in turn would decrease the ability of the hose to squirt out the marbles. With gallstones, much less cholesterol leaves the body, and cholesterol levels may rise.

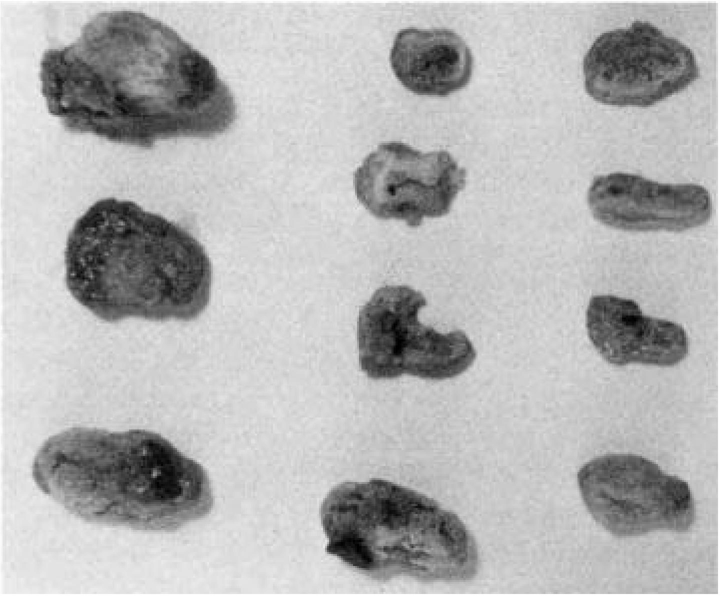

Fig. 89 These are gallstones.

Gallstones, being porous, can pick up all the bacteria, cysts, viruses and parasites that are passing through the liver. In this way “nests” of infection are formed, forever supplying the body with fresh bacteria. No stomach infection such as ulcers or intestinal bloating can be cured permanently without removing these gallstones from the liver.

Cleanse your liver twice a year.

Preparation.

- You can't clean a liver with living parasites in it. You won't get many stones, and you will feel quite sick. Zap daily the week before, or get through the first three weeks of the parasite killing program before attempting a liver cleanse. If you are on the maintenance parasite program, you are always ready to do the cleanse.
- Completing the kidney cleanse before cleansing the liver is also highly recommended. You want your kidneys, bladder and urinary tract in top working condition so they can efficiently remove any undesirable substances incidentally absorbed from the intestine as the bile is being excreted.
- Do any dental work first, if possible. Your mouth should be metal free and bacteria free (cavitations are cleaned). A toxic mouth can put a heavy load on the liver, burdening it immediately after cleansing. Eliminate that problem first for best results.

Ingredients

Epsom salts	4 tablespoons
Olive oil	half cup (light olive oil is easier to get down)
Fresh pink grapefruit	1 large or 2 small, enough to squeeze 2/3 to 3/4 cup juice

Ornithine	4 to 8, to be sure you can sleep. Don't skip this or you may have the worst night of your life!
Large plastic straw	To help drink potion.
Pint jar with lid	

Choose a day like Saturday for the cleanse, since you will be able to rest the next day.

Take no medicines, vitamins or pills that you can do without; they could prevent success. Stop the parasite program and kidney herbs, too, the day before.

Eat a no-fat breakfast and lunch such as cooked cereal with fruit, fruit juice, bread and preserves or honey (no butter or milk), baked potato or other vegetables with salt only. This allows the bile to build up and develop pressure in the liver. Higher pressure pushes out more stones.

2:00 PM. Do not eat or drink after 2 o'clock. If you break this rule you could feel quite ill later.

Get your Epsom salts ready. Mix 4 tbs. in 3 cups water and pour this into a jar. This makes four servings, ¾cup each. Set the jar in the refrigerator to get ice cold (this is for convenience and taste only).

6:00 PM. Drink one serving (¾cup) of the ice cold Epsom salts. If you did not prepare this ahead of time, mix 1 tbs. in ¾cup water now. You may add 1/8 tsp. vitamin C powder to improve the taste. You may also drink a few mouthfuls of water afterwards or rinse your mouth.

Get the olive oil and grapefruit out to warm up.

8:00 PM. Repeat by drinking another ¾cup of Epsom salts.

You haven't eaten since two o'clock, but you won't feel hungry. Get your bedtime chores done. The timing is critical for success; don't be more than 10 minutes early or late.

9:45 PM. Pour ½cup (measured) olive oil into the pint jar. Squeeze the grapefruit by hand into the measuring cup. Remove pulp with fork. You should have at least ½cup, more (up to ¾ cup) is best. You may top it up with lemonade. Add this to the olive oil. Close the jar tightly with the lid and shake hard until watery (only fresh grapefruit juice does this).

Now visit the bathroom one or more times, even if it makes you late for your ten o'clock drink. Don't be more than 15 minutes late.

10:00 PM. Drink the potion you have mixed. Take 4 ornithine capsules with the first sips to make sure you will sleep through the night. Take 8 if you already suffer from insomnia. Drinking through a large plastic straw helps it go down easier. You may use ketchup, cinnamon, or brown sugar to chase it down between sips. Take it to your bedside if you want, but drink it standing up. Get it down within 5 minutes (fifteen minutes for very elderly or weak persons).

Lie down immediately. You might fail to get stones out if you don't. The sooner you lie down the mote stones you will get out. Be ready for bed ahead of time. Don't clean up the kitchen. As soon as the drink is down walk to your bed and lie down flat on your back with your head up high on the pillow. Try to think about what is happening in the liver. Try to keep perfectly still for at least 20 minutes. You may feel a train of stones traveling along the bile ducts like marbles. There is no pain because the bile duct valves are open (thank you Epsom salts!). **Go to sleep**, you may fail to get stones out if you don't.

Next morning. Upon awakening take your third dose of Epsom salts. If you have indigestion or nausea wait until it is gone before drinking the Epsom salts. You may go back to bed. Don't take this potion before 6:00 am.

2 Hours Later. Take your fourth (the last) dose of Epsom salts. Drink ¾cup of the mixture. You may go back to bed.

After 2 More Hours you may eat. Start with fruit juice. Half an hour later eat fruit. One hour later you may eat regular food but keep it light. By supper you should feel recovered.

How well did you do? Expect diarrhea in the morning. Use a flashlight to look for gallstones in the toilet with the bowel movement. Look for the green kind since this is proof that they are genuine gallstones, not food residue. Only bile from the liver is pea green. The bowel movement sinks but gallstones float because of the cholesterol inside. Count them all roughly, whether tan or green. You will need to total 2000 stones before the liver is clean enough to rid you of allergies or bursitis or upper back pains permanently. The first cleanse may rid you of them for a few days, but as the stones from the rear travel forward, they give you the same symptoms again. You may repeat cleanses at two week intervals. Never cleanse when you are ill.

Sometimes the bile ducts are full of cholesterol crystals that did not form into round stones. They appear as a “chaff” floating on top of the toilet bowl water. It may be tan colored, harboring millions of tiny white crystals. Cleansing this chaff is just as important as purging stones.

How safe is the liver cleanse? It is very safe. My opinion is based on over 500 cases, including many persons in their seventies and eighties. None went to the hospital; none even reported pain. However it can make you feel quite ill for one or two days afterwards, although in every one of these cases the maintenance parasite program had been neglected. This is why the instructions direct you to complete the parasite and kidney rinse programs first.

CONGRATULATIONS

You have taken out your gallstones without surgery! I like to think I have perfected this recipe, but I certainly can not take credit for its origin. It was invented hundreds, if not thousands, of years ago, THANK YOU, HERBALISTS!

This procedure contradicts many modern medical viewpoints. Gallstones are thought to be formed in the gallbladder, not the liver. They are thought to be few, not thousands. They are not linked to pains other than gallbladder attacks. It is easy to understand why this is thought: by the time you have acute pain attacks, some stones are in the gallbladder, are big enough and sufficiently calcified to see on X-ray, and have caused inflammation there. When the gallbladder is removed the acute attacks are gone, but the bursitis and other pains and digestive problems remain.

The truth is self-evident. People who have had their gallbladder surgically removed still get plenty of green, bile-coated stones, and anyone who cares to dissect their stones can see that the concentric circles and crystals of cholesterol match textbook pictures of “gallstones” exactly.

Lugol’s Iodine Solution

It is too dangerous to buy a commercially prepared solution. It is certain to be polluted with propyl alcohol or wood alcohol. Make it yourself or ask your pharmacist to make it up for you. The recipe to make 1 liter (quart) is:

44 gm (1½ ounces) iodine, granular
88 gm (3 ounces) potassium iodide, granular

Dissolve the potassium iodide in about a pint of the water. Then add the iodine crystals and fill to the liter mark with water. It takes about 1 day to dissolve completely. Shake it from time to time. Keep out of sight and reach of children. Do not use if allergic to iodine. Be careful to avoid bottled water for preparation.

Vitamin D Drops

1 gram cholecalciferol (see Sources)
10 cups olive oil

Mix in a non-metal container. It may take a day of standing to dissolve fully. Refrigerate. Ten drops contain 40,000 iu. Use within a year.

Pathogen Frequencies

Living creatures emit a range of frequencies, also called *bandwidth*. As they age, the bandwidth shrinks. When they die sometimes all that is left is a single frequency.

Most of the organisms listed below are dead on commercially available and prepared slides (see *Sources* for biological supply companies). However they still exhibit a 5 KHz bandwidth, probably due to testing with a frequency generator that was only accurate to 100 Hz, and also due to using more voltage than necessary (like when a powerful radio station comes in at its own frequencies and ones nearby, too). Some testing was done with a more accurate frequency generator at a lower power level so some bandwidths are reported much more narrowly.

If the same person retests the same specimens with the same equipment within a few days, the results will be absolutely identical (within 1 Hz) 90% of the time. Why a few of the results will not be identical is not known. However different people, and even the same person at different times of the year, can notice that the perceived frequencies shift by as much as 3 KHz (still less than 1% change).

Some specimens have more than one range listed; this may be characteristic of the organism or may be due to having an undocumented organism on the same microscope slide.

Blank locations represent organisms for whom there are prepared slides available, but whose bandwidth has not been determined.

Bandwidth Of Organism Families

In general, the smaller the organism the lower the frequency and narrower the bandwidth. This chart shows the major families studied and where they fall in the spectrum.

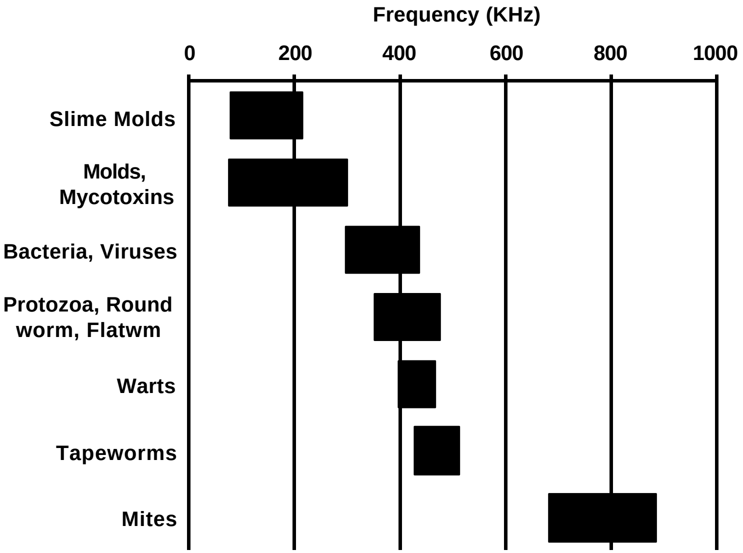

Fig. 90 Chart of bandwidths for organism families.

Mold, Mold Toxin Frequencies

Other molds and mold toxins	**KHz**
Aflatoxin	177,188
Cytochalasin B	77,91
Ergot	295
Griseofulvin	288
Sorghum syrup	277
Sterigmatocystin	88,96,133,126
Zearalenone	100

Slime Molds	**KHz**
Arcyria	81
Lycogala	126
Stemonitis	211

Bacteria and Viruses

Including locations where I commonly found them.

	Low Freq (KHz)	High Freq (KHz)	Use freq gen for 3 min @
Acetobacter aceti			
Adenovirus	393	393	393
Adenovirus (2nd range)	371.45	386.90	
Agrobacterium tumefaciens			
Alcaligenes faecalis			
Alpha streptococcus	369.75	385.4	380,375
Azobacter chroococcum			
Bacillus anthracis causes anthrax in cattle (tooth)	393.5	398.05	395,364,368
Bacillus anthracis (2nd range)	363.2	365.3	
Bacillus anthracis (3rd range)	359.4	370.5	
Bacillus anthracis spores	386.95	391.45	388
Bacillus cereus	373.65	375.85	374.5
Bacillus megaterium			
Bacillus sterothermophilus			
Bacillus subtilis spores			
Bacillus subtilis var. niger	371.85	387.1	385,380,375
Bacteria capsules (capsular strain)	416.05	418.75	417.5
Bacterial capsules	357.6	362.4	360
Bacteroides fragilis found with common roundworm Ascaris	324.3	325.0	325
Bacteroides fragilis (2nd range)	325.7	326.0	
Beta streptococcus (tooth)	380.6	387.4	385
Blepharisma	405.65	407.45	406.5
Bordetella pertussis “whooping cough” (tooth)	329.85	332.25	331
Borellia burgdorferi Lyme disease	378.95	382.0	380
Branhamella (Neisseria) catarrhalis (has hole at 398)	394.9	396.7	396
Brucella abortus			
Cabbage Black Rot			
Campylobacter fetus smear	365.3	370.6	368
Campylobacter pyloridis	352.0	357.2	355
Candida albicans (pure powder) common yeast	384.2	388.4	386
Caulobacter vibrioides			
Central spores (bacillus smear)	372.45	378.65	376
Chlamydia trachomatis	379.7	383.95	381
Clostridium acetobutylicum	382.8	391.15	389,384
Clostridium botulinum (tooth) causes food poisoning	361.0	364.55	362

Clostridium perfringens			
Clostridium perfringens spores	394.2	398.1	396
Clostridium septicum	362.05	365.6	364
Clostridium sporogenes			
Clostridium tetani (tooth) causes tetanus			
Corynebacterium diphtheriae (tooth) causes diphtheria	340	344	342
Corynebacterium pseudodiphthericum			
Corynebacterium xerosis	315.65	316.8	316.0
Coxsackie virus B-1 found with Bacteroides fragilis	360.5	366.1	364
Coxsackie virus B-4 found with Bacteroides fragilis	361.45	363.7	362.5
Coxsackie virus B-4 (2nd range)	363.9	364.9	
Crithidia fasciculata			
Cytomegalovirus (CMV) antigen	408.35	410.75	409
Cytophaga rubra	428.1	432.2	430
Diplococcus diphtheriae	357.95	264.0	361
Diplococcus pneumoniae	351.65	368.45	365,360
Eikanella corrodens	379.5	384.3	382
Enterobacter aerogenes intestinal bacterium	374	374	374
Epstein Barre virus (EBV)	372.5	382.85	380,375
Erwinia amylovora	347.2	352.1	350
Erwinia carotovora	368.1	377.0	373
Escherichia coli (E. coli) intestinal bacterium	356	356	356,393
Escherichia coli (E. coli) (2nd range)	392	393	
Gaffkya tetragena causes respiratory infections	344.85	352.5	350
Gardnerella vaginalis ovarian and genital tract infection	338.0	342.55	340
Haemophilus influenzae bacterial meningitis, infects joints	336.41	336.41	336
Hepatitis B antigen	414.55	420.8	418
Herpes simplex 1	291.25	293.05	292,345.5
Herpes simplex 1 (2nd range)	345.35	345.75	
Herpes simplex 2 (fresh smear)	353.9	362.9	360,355
Herpes Zoster “shingles”	416.6	420.2	418
Histomonas meleagridis (liver)	376.55	378.7	377
Histoplasma capsulatum	298.3	304.85	302
HIV	365	365	365
Influenza A and B (flu shot)	313.35	323.9	320,315
Iron Bacterium Sphaerotilus			
Klebsiella pneumoniae causes pneumonia	398.45	404.65	401,419
Klebsiella pneumoniae (2nd range)	416.9	421.9	
Lactobacillus acidophilus (tooth)	346.05	351.65	349

Leptospira interrogans spirochete	397.05	401.1	399
Lumpy Jaw			
Measles antigen	369.5	373.0	371
Micrococcus luteus			
Micrococcus roseus			
Mumps antigen	377.6	384.65	382
Mycobacterium para TB			
Mycobacterium phlei	409.65	410.65	410.0
Mycobacterium smegmatis			
Mycobacterium tuberculosis (infec nodule) causes tuberculosis	430.55	434.2	432
Mycoplasma	322.85	323.9	323.5,346
Mycoplasma (range 2)	342.75	349.3	
Neisseria gonorrhea causes gonorrhea	333.85	336.5	334
Neisseria sicca			
Nocardia asteroides found in Parkinson's Disease	354.95	355.35	355.1,368
Nocardia asteroides (2nd range)	363.7	370	
Propionobacterium acnes	383.75	389.0	387
Proteus mirabilis	320.55	326.0	324,349
Proteus mirabilis (2nd range)	345.95	352.1	
Proteus vulgaris urinary tract pathogen	408.75	416.45	413,336,328
Proteus vulgaris (2nd range)	333.75	339.15	
Proteus vulgaris (3rd range)	327.2	329.5	
Pseudomonas aeruginosa found in open wounds	331.25	334.6	333
Pseudomonas fluorescens			
Respiratory syncytial virus	378.95	383.15	380
Rhizobium leguminosarum			
Salmonella enteriditis intestinal infection	329	329	329
Salmonella paratyphi	365.05	370.1	368,385
Salmonella typhimurium food poisoning, nervousness, apathy	382.3	386.55	355,386,390
Serratia marcescens	349.45	352.1	351
Shigella dysenteriae intestinal problems	390.089	390.089	390.089
Shigella flexneri depression	394	394	394
Shigella sonnei invades tumors	318	318	318
Sphaerotilus natans	388.4	393.45	391
Spirillum itersonil			
Spirillum serpens	378.35	382.8	380
Spirillum sinuosum			
Spirillum volutans			
Spores in bacteria spore stain			
Staphylococcus aureus (culture)	376.27	380.85	
Staphylococcus aureus (slide) source is tooth infection, causes abscesses, heart disease, invades tumors	381	381	378,381

Staphylococcus epidermidis infects skin and mucous membranes			
Streptococcus lactis occurs in milk	382	387	385
Streptococcus mitis lung infection, tooth infection, abscesses, causes stiff knees	313.8	321.1	318
Streptococcus pneumoniae causes pneumonia and inner ear disease	366.85	370.2	368
Streptococcus pyogenes (tooth)	360.5	375.3	373
Streptococcus sp. group G (tooth)	368.15	368.85	368
Sub terminal spores bac. smear	385.15	385.95	
Terminal spores bacillus smear			
Tobacco mosaic virus (tobacco)	427.15	429.55	428
Treponema pallidum causes syphilis	346.85	347.4	347
Troglodytella abrassari	377.75	385.2	383,419
Troglodytella abrassari (2nd range)	416.9	422.2	
Veillonella dispar	401.75	405.2	403
Vibrio (photobacterium) fischeri			

Roundworms, Flatworms, One-celled Animals

	Low Freq (KHz)	**High Freq (KHz)**	**To kill, use freq. gen for 3 min. at these frequencies**
Acanthamoeba culbertsoni			
Acanthocephala			
Anaplasma marginale	386.4	388.0	387,422
Anaplasma marginale (2nd range)	415.3	424	
Ancylostoma braziliense (adult)	397.6	403.25	401
Ancylostoma caninum	383.1	402.9	400,393,386
Ancylostoma duodenale male			
Anguillula aceti			
Ascaris larvae in lung common roundworm of cats and dogs	404.9	409.15	408
Ascaris lumbricoides (m and f)			same
Ascaris megalocephala (male)	403.85	409.7	408
Babesia bigemina			
Babesia canis smear			
Balantidium coli cysts	458.8	462.9	460
Balantidium sp. trophozoites (from guinea pig) parasitic ciliate			
Besnoitia (lung sect.) protozoan	352.8	361.4	358

Capillaria hepatica (liver sect.)	424.25	430.65	428
Chilomastix cysts (rat)	388.95	390.7	389,426
Chilomastix cysts (rat) (2nd range)	425.2	427.3	
Chilomastix mesnili (trophozoites)			same
Chilomonas, whole mount	393.75	400	398
Clinostomum metacercaria			
Clonorchis metacercariae			
Clonorchis sinensis	425.7	428.75	427
Clonorchis sinensis eggs			
Cryptocotyle lingua (adult)	409.95	416.0	414
Didinium			
Dientamoeba fragilis	401.35	406.05	404
Dipetalonema perstans (microfilaria human blood)			
Dirofilaria immitis dog heartworm	408.15	411.15	409
Echinoporyphium recurvatum	418.55	423.9	421
Echinostoma revolutum	425.5	429.65	428
Eimeria stiedae			
Eimeria tenella			
Endamoeba gingivalis trophozoite	433.8	441.0	438
Endolimax nana trophozoites and cysts	394.25	397.1	396,432
Endolimax nana trophozoites and cysts (2nd range)	430.5	433.35	
Entamoeba coli cysts			
Entamoeba coli trophozoites	397.0	400.35	398
Entamoeba histolytica trophozoite	381.1	387.8	385
Enterobius vermicularis	420.95	426.3	423
Eurytrema pancreaticum	420.35	422.3	421
Eurytrema pancreaticum stages			
Fasciola hepatica	421.35	427.3	425
Fasciola hepatica cercariae	423.8	430.6	427
Fasciola hepatica eggs	422.0	427.6	425
Fasciola hepatica metacercariae			
Fasciola hepatica miracidia	421.75	424.7	423
Fasciola hepatica rediae	420.6	427.5	425
Fasciolopsis buskii adult	427.7	435.1	434
Fasciolopsis buskii eggs	427.35	435.45	434
Fasciolopsis buskii eggs unincubated			
Fasciolopsis cercariae	429.5	436.25	434
Fasciolopsis miracidia	427.35	435.2	434
Fasciolopsis rediae	427.3	433.0	432
Fischoedrius elongatus	441.75	443.2	442
Gastrothylax elongatus	451.9	457.1	455
Giardia lamblia (trophozoites)	421.4	426.3	424
Giardia lamblia cysts			
Gyrodactylus	378.75	381.8	380
Haemonchus contortus	386.8	395.5	393
Haemoproteus			

Hasstile sig. tricolor (adult)	448.05	455.1	453
Heterakis			
Hypodereum conoideum	424.45	429.55	427
Iodamoeba butschlii trophozoites and cysts	437.85	448.5	445,402
Iodamoeba butschlii trophozoites and cysts (2nd range)	398.15	404.75	
Leishmania braziliensis	400.05	405.1	403
Leishmania donovani	398.0	402.65	400
Leishmania mexicana	400.2	403.8	402
Leishmania tropica	402.1	407.4	405
Leucocytozoon	397.45	402.55	400
Loa loa	360.551	360.551	361
Macracanthorhynchus	438.85	442.8	440
Metagonimus Yokogawai	437.35	442.1	440
Monocystis agilis			
Myxosoma	409.6	416.95	414
Naegleria fowleri	356.9	364.35	362
Naegleria fowleri (brain sec.)			
Necator americanus (infect larvae)			
Notocotylus quinqeserialis			
Onchocerca volvulus (tumor)	436.3	442.1	440
Paragonimus Westermanii adult	437.8	454.2	452,447
Passalurus ambiguus	428.8	444.15	441,437
Pelomyxa carolinensis			
Plasmodium cynomolgi	417.3	424.5	422
Plasmodium falciparum smear	372.3	373.8	373.0
Plasmodium vivax smear	438.15	445.1	442
Platynosomum fastosum adult			
Pneumocystis carnii (lung)	405.75	409.15	407
Prosthogonimus macrorchis(eggs)	396.85	404.75	401
Sarcina lutea			
Sarcocystis	450.55	454.95	452
Schistosoma haematobium	473	473	473
Schistosoma japonicum eggs			
Schistosoma mansoni	353	353	353
Stephanurus dentalus (ova)	457.35	463.1	461
Stigeoclonium	404.25	415.25	412,407
Strongyloides (filariform larva)	398.4	402.0	400
Strongyloides parasitic females			
Toxocara (eggs)			
Toxoplasma (human strain)	395.0	395.0	395
Trichinella spiralis (muscle)	403.85	405.57	404.5
Trichomonas muris			
Trichomonas vaginalis	378.0	383.6	381
Trichuris sp. (male)	388.3	408.9	406
Trypanosoma brucei	423.2	431.4	429
Trypanosoma cruzi (brain tissue)	460.2	465.65	463

Trypanosoma equiperdum	434.6	451.25	448,442,438
Trypanosoma gambiense	393.75	398.7	396
Trypanosoma lewisi (blood smear)	424.5	426.0	425
Trypanosoma rhodesiense	423.5	428.55	426
Urocleidus	442.35	450.0	447

Wart Frequencies

(Most of these are from homemade slides.)

	Low Freq	**High Freq**	**Use freq gen for 3 min @**
Wart BS	402	406	404
Wart CC	426	432.35	430
Wart FR	459.3	464.75	462
Wart HA	434.8	444.1	442,437
Wart HRCm	438.9	448.55	446,441
Wart human papilloma plantar	404.7	406.75	405
Wart human papilloma virus	402.85	410.7	407
Wart JB	418.75	422.4	420
Wart L arm	343.65	345.95	344
Wart papilloma cervix smear	404.05	404.6	404.3

Tapeworms

Tapeworms are segmented. The first segment is the head, called the *scolex*. Tapeworms grow by adding a new segment to their body.

Tapeworms can have very large bandwidths (range of frequencies), and it varies by the length of the specimen! It is as if each new segment has a unique, and slightly lower, frequency.

Do not use a frequency generator to kill tapeworms. If you accidentally kill middle segments instead of working your way up from the bottom, you may conceivably promote dispersion! Use only a zapper.

	Low Freq	High Freq
Cysticercus fasciolaris	436.4	440.05
Diphyllobothrium erinacei (Mansoni) (scolex)	467.25	487.55
Diphyllobothrium erinacei eggs		
Diphyllobothrium latum (scolex)	452.9	472.3
Dipylidium caninum (proglottid composite)	439.55	444.3
Dipylidium caninum (scolex)	451.95	472.15
Echinococcus granulosus	451.6	461.5
Echinococcus granulosus (cysts)	441.15	446.5
Echinococcus granulosus (eggs)		
Echinococcus multilocularis	455.85	458.35
Heterophyes heterophyes		
Hymenolepis cysticercoides	478.0	481.75
Hymenolepis diminuta	445	481.15
Hymenolepis diminuta ova		
Hymenolepis nana eggs		
Moniezia (scolex)	430.35	465.2
Moniezia expansa (composite)	430.35	465.2
Moniezia expansa eggs		
Multiceps serialis	453.6	457.8
Pigeon tapeworm		
Taenia pisiformis (cysticercus)	475.2	482.1
Taenia pisiformis eggs (ova)	465.2	469.7
Taenia saginata (cysticercus)	476.5	481.05
Taenia saginata eggs		
Taenia solium (cysticercus)	475	475
Taenia solium (scolex)	444.0	448.9
Taenia solium eggs		

Mite Frequencies

These are the organisms that cold viruses ride in with!

Mite	**KHz**
Demodex folliculorum follicle mite	682
Dermatophagoides dust mite	707
Meal mite	718
Ornithonyssus bird mite	877,878
Sarcoptes scabei itch	735

Miscellaneous Frequencies

	KHz
Blue-green Algae	256
Bryozoa cristatalla	396
Mucor mucedo	288
Rhizobium meliloti	330
Rotifer	1151

It's easy to make homemade slides when you or a family member is ill. Finding out the frequencies of these illnesses helps you identify them (use the Pathogen Frequency Chart) and also lets you know if you are chronically getting them back.

Unidentified pathogens	**Low Freq**	**High Freq**
A cold virus HRC	395.8	395.8
Fungus EW	362.0	364.9
Fungus JWB	397.2	400.75
Tooth decay	384.3	387.2
Tooth decay (N)	367.9	375.05
Tooth decay (N) (2nd range)	326.95	331.5
Tooth decay (N) (3rd range)	293.2	297.4
Tooth plaque I	378.8	383.05
Tooth plaque I (2nd range)	294.7	298.25
Tooth plaque I (3rd range)	233.1	238.2
Tooth plaque II	384.95	387.05
Tooth plaque II (2nd range)	278.75	284
Tooth plaque II (3rd range)	212.15	218
Tooth plaque II (4th range)	340.15	344.8
Tooth plaque II (5th range)	305.5	310.35

Toxic Elements

Although not living, solvents and toxins must exhibit characteristic frequencies, otherwise how could the Syncrometer detect specific ones? This needs further exploration.

Most of the toxic elements I use are metals, heavy metals and lanthanides. But some are not; examples are PCBs and formaldehyde.

Some important elements are missing, like iron, zinc and manganese. This is because I never could find them present in the white blood cells, and I finally gave up searching for them.

Below is a list of the 70 or so toxic elements I use. Most of them were obtained as Atomic Absorption Standard Solutions and are, therefore, very pure. This prevents mistakes in identifying a toxin. They were stored in ½ounce amber glass bottles with bakelite caps and permanently sealed with plastic film since testing did not require them to be opened (they get close enough to the frequency field). The exact concentration and the solubility characteristics are not important in this qualitative test. The main sources of these substances in our environment are given beside each item.

Toxic Substance	Sources
Aflatoxin B	beer, bread, apple cider vinegar, moldy fruit, nuts
Aluminum	cookware, deodorant, lotions, soaps
Aluminum silicate	salt, water softener
Antimony	fragrance in lotions, colognes
Arsenic	pesticide, "treated" carpet, wallpaper
Asbestos	clothes dryer belt, hair blower, paint on radiators
Barium	lipstick, bus exhaust
Benzalkonium chloride	toothpaste
3,4 Benzopyrene	flame cooked foods, toast
4,5 Benzopyrene	flame cooked foods, toast
Beryllium	hurricane lamps, gasoline, dentures, kerosene
Bismuth	colognes, lotions, antacids
Boron	
Bromine	bleached "brominated" flour
Cadmium	galvanized water pipes, old tooth fillings
Cerium	tooth fillings
Cesium	clear plastic bottles used for beverages
Chlorine	from Chlorox™ bleach
Chromium	cosmetics, water softener
Cobalt	detergent, blue and green body products
Copper	tooth fillings, water pipes
Dysprosium	paint and varnish
Erbium	packaging for food, pollutant in pills
Europium	tooth fillings
Europium oxide	tooth fillings, catalytic converter

Fiberglass	dust from remodeling or building insulation
Formaldehyde	foam in mattresses and furniture, paneling
Gadolinium	tooth fillings
Gallium	tooth fillings
Germanium	with thallium in tooth fillings (pollutant)
Gold	tooth fillings
Hafnium	hair spray, nail polish, pollutant in pills
Holmium	usually found in presence of PCBs
Indium	tooth fillings
Iridium	tooth fillings
Lanthanum	computer and printing supplies
Lead	solder joints in water pipes
Lithium	printing supplies
Lutetium	paint and varnish
Mercury	tooth fillings
Molybdenum	auto supplies
Neodymium	pollutant in pills
Nickel	tooth fillings, metal glasses frames
Niobium	pollutant in pills, foil packaging for food
Palladium	tooth fillings
Platinum	tooth fillings
Polychlorinated biphenyl PCB	detergent, hair spray, salves
Polyvinyl chloride acetate (PVC)	glues, building supplies, leaking cooling system
Praseodymium	pollutant in pills
Radon	cracks in basement cement, water pipes
Rhenium	spray starch
Rhodium	tooth fillings
Rubidium	tooth fillings
Ruthenium	tooth fillings
Samarium	tooth fillings
Scandium	tooth fillings
Selenium	
Silver	tooth fillings
Sodium fluoride	toothpaste
Strontium	toothpaste, water softener
Tantalum	tooth fillings
Tellurium	tooth fillings
Terbium	pollutant in pills
Thallium acetate	pollutant in mercury tooth fillings
Thorium nitrate	earth (dust)
Thulium	pollutant in many brands of Vitamin C
Tin	toothpaste
Titanium	tooth fillings, body powder
Tungsten	electric water heater, toaster, hair curler
Uranium acetate	earth (dust)
Vanadium pentoxide	gas leak in home, candles (not necessarily lit)

Ytterbium	pollutant in pills
Yttrium	pollutant in pills
Zirconium	deodorant, toothpaste

Solvents

This is a list of all the solvents I use together with the main source of them in our environment. These are chemicals, very pure, obtained from chemical supply companies, unless otherwise stated. Those marked with an asterisk (*) were the subject of a recent book *The Neurotoxicity of Solvents* by Peter Arlien-Soburg, 1992, CRC Press.

Solvent	**Source**
1,1,1, Trichloro ethane* (TCE)	flavored foods
2,5-Hexane dione*	flavored foods
2 Butanone* (methyl ethyl ketone)	flavored foods
2 Hexanone* (methyl butyl ketone)	flavored foods
2 Methyl propanol	
2 Propanol (propyl alcohol)	see the propyl alcohol list
Acetone	store-bought drinking water, cold cereals, pet food, animal feed
Acetonylacetone (2,5 hexanedione)	flavored foods
Benzene	see the benzene list (page 354)
Butyl nitrite	
Carbon tetrachloride	store-bought drinking water, cold cereals, pet food, animal feeds
Decane	health food cookies and cereals
Denatured alcohol	obtained from pharmacy
Dichloromethane* (methylene chloride)	store-bought orange juice, herb tea blends
Gasoline regular leaded	obtained at gasoline station
Grain alcohol	95% ethyl alcohol obtained at liquor store
Hexanes*	decaffeinated beverages
Isophorone	flavored foods
Kerosene	obtained at gasoline station
Methanol (wood alcohol)	colas, artificial sweeteners, infant formula
Mineral oil	lotions

Mineral spirits	obtained from paint store
Paradichlorobenzene	mothballs
Pentane	decaffeinated beverages
Petroleum ether	in some gasolines
Styrene*	styrofoam dishes
Toluene*	store-bought drinking water, cold cereals
Trichloroethylene* (TCEthylene)	flavored foods
Xylene*	store-bought drinking water, cold cereals

Pathogen Frequency Chart

Use this chart if you know the frequency and wonder what the pathogen might be.

	KHz		5	0			1	0	0			1	5	0			2	0	0			2	5	0			3	0	0			3	5	0
Pathogen	Low	High																																
Cytochalasin B	77.00	77.00			m																													
Arcyria	81.00	81.00				s																												
Sterigmatocystin	88.00	88.00				m																												
Cytochalasin B (2nd)	91.00	91.00					m																											
Sterigmatocystin (2nd)	96.00	96.00					m																											
Zearalenone	100.00	100.00						m																										
Lycogala	126.00	126.00								s																								
Sterigmatocystin (4th)	126.00	126.00								m																								
Sterigmatocystin (3rd)	133.00	133.00									m																							
Aflatoxin	177.19	177.19													m																			
Stemonitis	211.00	211.00																	s															
Sorghum syrup	277.00	277.00																							m									
Griseofulvin	288.00	288.00																								m								
Herpes simplex 1	291.25	293.05																									v							
Ergot	295.00	295.00																									m							
Histoplasma capsulatum	298.30	304.85																									b	b						
Corynebacterium xerosis	315.65	316.80																											b					
Shigella sonnei	318.00	318.00																											b					
Streptococcus mitis	313.80	321.10																											b	b				
Influenza A and B (flusho	313.35	323.90																											v	v				
Mycoplasma	322.85	323.90																												b				
Bacteroides fragilis	324.30	325.00																												b				
Proteus mirabilis	320.55	326.00																												b				
Bacteroides fragilis (2nd)	325.70	326.00																												b				
Salmonella enteriditis	329.00	329.00																												b				
Proteus vulgaris (3rd)	327.20	329.50																												b				
Bordetella pertussis	329.85	332.25																												b	b			
Pseudomonas aeruginos	331.25	334.60																													b			
Haemophilus influenzae	336.41	336.41																													b			
Neisseria gonorrhea	333.85	336.50																													b			
Proteus vulgaris (2nd)	333.75	339.15																													b			
Gardnerella vaginalis	338.00	342.55																													b	b		
Corynebacterium diphthe	340.00	344.00																														b		
Herpes simplex 1 (2nd)	345.35	345.75																														v		
Wart L arm	343.65	345.95																														w		
Treponema pallidum	346.85	347.40																														b		
Mycoplasma (2nd)	342.75	349.30																														b		
Lactobacillus acidophilus	346.05	351.65																														b	b	
Proteus mirabilis (2nd)	345.95	352.10																														b	b	
Erwinia amylovora	347.20	352.10																														b	b	
Serratia marcescens	349.45	352.10																														b	b	
Gaffkya tetragena	344.85	352.50																														b	b	
Schistosoma mansoni	353.00	353.00																															p	
Nocardia asteroides	354.95	355.35																															b	
Escherichia coli (E. coli)	356.00	356.00																															b	
Campylobacter pyloridis	352.00	357.20																															b	
Loa loa	360.55	360.55																																p
Besnoitia (lung sect.) prot	352.80	361.40																															p	p
Bacterial capsules	357.60	362.40																															b	b
Herpes simplex 2 (fresh s	353.90	362.90																															v	v
Coxsackie virus B-4	361.45	363.70																																v

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

KHz

Pathogen	Low	High	3	5	0			4	0	0			4	5	0			5	0	0
Diplococcus diphtheriae	357.95	364.00	b	b																
Naegleria fowleri	356.90	364.35	p	p																
Clostridium botulinum	361.00	364.55		b																
Coxsackie virus B-4 (2nd)	363.90	364.90		v																
Bacillus anthracis (2nd)	363.20	365.30		b																
Clostridium septicum	362.05	365.60		b																
Coxsackie virus B-1	360.50	366.10		v																
Diplococcus pneumoniae	351.65	368.45	b	b																
Streptococcus sp. group	368.15	368.85		b																
Nocardia asteroides (2nd	363.70	370.00		b	b															
Salmonella paratyphi	365.05	370.10		b	b															
Streptococcus pneumoni	366.85	370.20		b	b															
Bacillus anthracis (3rd)	359.40	370.50	b	b	b															
Campylobacter fetus sme	365.30	370.60		b	b															
Measles antigen	369.50	373.00		v	v															
Plasmodium falciparum s	372.30	373.80			p															
Enterobacter aerogenes	374.00	374.00			b															
Streptococcus pyogenes	360.50	375.30		b	b															
Bacillus cereus	373.65	375.85			b															
Erwinia carotovora	368.10	377.00		b	b															
Central spores (bacillus s	372.45	378.65			b															
Histomonas meleagridis (l	376.55	378.70			b															
Staphylococcus aureus (c	376.27	380.85			b	b														
Staphylococcus aureus (s	381.00	381.00				b														
Gyrodactylus	378.75	381.80			p	p														
Borellia burgdorferi	378.95	382.00			v	v														
Spirillum serpens	378.35	382.80			b	b														
Epstein Barre virus (EBV)	372.50	382.85			v	v														
Respiratory syncytial viru	378.95	383.15			v	v														
Trichomonas vaginalis	378.00	383.60			p	p														
Chlamydia trachomatis	379.70	383.95			b	b														
Eikanella corrodens	379.50	384.30			b	b														
Mumps antigen	377.60	384.65			v	v														
Troglodytella abrassari	377.75	385.20			b	b														
Alpha streptococcus	369.75	385.40		b	b	b														
Sub terminal spores bac.	385.15	385.95				b														
Salmonella typhimurium	382.30	386.55				b														
Adenovirus (2nd range)	371.45	386.90			v	v														
Streptococcus lactis	382.00	387.00				b														
Bacillus subtilis var. niger	371.85	387.10			b	b														
Beta streptococcus	380.60	387.40				b														
Entamoeba histolytica tro	381.10	387.80				p														
Anaplasma marginale	386.40	388.00				p														
Candida albicans	384.20	388.40				y														
Propionobacterium acnes	383.75	389.00				b														
Shigella dysenteriae	390.09	390.09					b													
Chilomastix cysts (rat)	388.95	390.70				p	p													
Clostridium acetobutylicu	382.80	391.15				b	b													
Bacillus anthracis spores	386.95	391.45				b	b													
Escherichia coli (E. coli)	392.00	393.00					b													
Adenovirus	393.00	393.00					v													
Sphaerotilus natans	388.40	393.45				b	b													
Shigella flexneri	394.00	394.00					b													
Toxoplasma (human strai	395.00	395.00					p													
Haemonchus contortus	386.80	395.50				p	p													
Branhamella (Neisseria)	394.90	396.70					b													
Endolimax nana trophozo	394.25	397.10					p													
Bacillus anthracis	393.50	398.05					b													
Clostridium perfringens s	394.20	398.10					b													

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																			
			3	5	0			4	0	0			4	5	0			5	0	0
Pathogen	Low	High																		
Trypanosoma gambiense	393.75	398.70						p												
Chilomonas, whole mount	393.75	400.00						p	p											
Entamoeba coli trophozoi	397.00	400.35						p	p											
Leptospira interrogans	397.05	401.10						b	b											
Strongyloides (filariform l	398.40	402.00						p	p											
Leucocytozoon	397.45	402.55						p	p											
Leishmania donovani	398.00	402.65						p	p											
Ancylostoma caninum	383.10	402.90					p	p	p											
Ancylostoma braziliense (	397.60	403.25						p	p											
Leishmania mexicana	400.20	403.80							p											
Wart papilloma cervix sm	404.05	404.60							w											
Klebsiella pneumoniae	398.45	404.65						b	b											
Prosthogonimus macrorc	396.85	404.75						p	p											
Iodamoeba butschlii (2nd)	398.15	404.75						p	p											
Leishmania braziliensis	400.05	405.10							p											
Veillonella dispar	401.75	405.20							b											
Trichinella spiralis (muscl	403.85	405.57							p											
Wart BS	402.00	406.00							w											
Dientamoeba fragilis	401.35	406.05							p											
Wart human papilloma pl	404.70	406.75							w											
Leishmania tropica	402.10	407.40							p											
Blepharisma	405.65	407.45							b											
Trichuris sp. (male)	388.30	408.90					p	p	p											
Ascaris larvae in lung	404.90	409.15							p											
Pneumocystis carnii (lung	405.75	409.15							p											
Ascaris megalocephala (	403.85	409.70							p											
Mycobacterium phlei	409.65	410.65							b	b										
Wart human papilloma vir	402.85	410.70							w	w										
Cytomegalovirus (CMV) a	408.35	410.75							v	v										
Dirofilaria immitis	408.15	411.15							p	p										
Stigeoclonium	404.25	415.25							p	p										
Cryptocotyle lingua (adult	409.95	416.00							p	p										
Proteus vulgaris	408.75	416.45							b	b										
Myxosoma	409.60	416.95							p	p										
Bacteria capsules (capsul	416.05	418.75								b										
Herpes Zoster	416.60	420.20								v	v									
Hepatitis B antigen	414.55	420.80								v	v									
Klebsiella pneumoniae (2	416.90	421.90								b	b									
Troglodytella abrassari (2	416.90	422.20								b	b									
Eurytrema pancreaticum	420.35	422.30									p									
Wart JB	418.75	422.40								w	w									
Echinoporyphium recurva	418.55	423.90								p	p									
Anaplasma marginale (2n	415.30	424.00								p	p									
Plasmodium cynomolgi	417.30	424.50								p	p									
Fasciola hepatica miracidi	421.75	424.70									p									
Trypanosoma lewisi (bloo	424.50	426.00									p									
Enterobius vermicularis	420.95	426.30									p									
Giardia lamblia (trophozoi	421.40	426.30									p									
Fasciola hepatica	421.35	427.30									p									
Chilomastix cysts (rat) (2	425.20	427.30									p									
Fasciola hepatica rediae	420.60	427.50									p									
Fasciola hepatica eggs	422.00	427.60									p									
Trypanosoma rhodesiens	423.50	428.55									p									
Clonorchis sinensis	425.70	428.75									p									
Hypodereum conoideum	424.45	429.55									p									
Tobacco mosaic virus	427.15	429.55									v									
Echinostoma revolutum	425.50	429.65									p									
Fasciola hepatica cercari	423.80	430.60									p	p								
Capillaria hepatica (liver s	424.25	430.65									p	p								

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz																				
			3	5	0			4	0	0			4	5	0			5	0	0
Pathogen	Low	High																		
Trypanosoma brucei	423.20	431.40									p	p								
Cytophaga rubra	428.10	432.20									b	b								
Wart CC	426.00	432.35									w	w								
Fasciolopsis rediae	427.30	433.00									p	p								
Endolimax nana (2nd)	430.50	433.35										p								
Mycobacterium tuberculo	430.55	434.20										b								
Fasciolopsis buskii	427.70	435.10									p	p								
Fasciolopsis miracidia	427.35	435.20									p	p								
Fasciolopsis buskii eggs	427.35	435.45									p	p								
Fasciolopsis cercariae	429.50	436.25									p	p								
Cysticercus fasciolaris	436.40	440.05										t	t							
Endamoeba gingivalis tro	433.80	441.00										p	p							
Onchocerca volvulus (tu	436.30	442.10										p	p							
Metagonimus Yokogawai	437.35	442.10										p	p							
Macracanthorhynchus	438.85	442.80										p	p							
Fischoedrius elongatus	441.75	443.20											p							
Wart HA	434.80	444.10										w	w							
Passalurus ambiguus	428.80	444.15									p	p	p							
Dipylidium caninum	439.55	444.30										t	t							
Plasmodium vivax smear	438.15	445.10										p	p							
Echinococcus granulosus	441.15	446.50											t							
Iodamoeba butschlii troph	437.85	448.50										p	p							
Wart HRCm	438.90	448.55										w	w							
Taenia solium (scolex)	444.00	448.90											t							
Urocleidus	442.35	450.00											p	p						
Trypanosoma equiperdu	434.60	451.25										p	p	p						
Paragonimus Westermani	437.80	454.20										p	p	p						
Sarcocystis	450.55	454.95												p						
Hasstile sig. tricolor (adult	448.05	455.10											p	p						
Gastrothylax elongatus	451.90	457.10												p						
Multiceps serialis	453.60	457.80												t						
Echinococcus multilocula	455.85	458.35												t						
Echinococcus granulosus	451.60	461.50												t	t					
Balantidium coli cysts	458.80	462.90												p	p					
Stephanurus dentalus (ov	457.35	463.10												p	p					
Wart FR	459.30	464.75												w	w					
Moniezia (scolex)	430.35	465.20										t	t	t	t					
Moniezia expansa (comp	430.35	465.20										t	t	t	t					
Trypanosoma cruzi (brain	460.20	465.65													p					
Taenia pisiformis eggs (o	465.20	469.70													p					
Dipylidium caninum (scol	451.95	472.15												t	t	t				
Diphyllobothrium latum (s	452.90	472.30												t	t	t				
Schistosoma haematobiu	473.00	473.00														p				
Taenia solium (cysticercu	475.00	475.00														t				
Taenia saginata (cysticer	476.50	481.05														t	t			
Hymenolepis diminuta	445.00	481.15											t	t	t	t	t			
Hymenolepis cysticercoid	478.00	481.75														t	t			
Taenia pisiformis (cystice	475.20	482.10														t	t			
Diphyllobothrium erinacei	467.25	487.55													t	t	t			

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

	KHz		5	5	0			6	0	0			6	5	0			7	0	0			7	5	0			8	0	0			8	5	0
Pathogen	Low	High																																	
Demodex folliculorum foll	682.00	682.00															x																		
Dermatophagoides dust	707.00	707.00																	x																
Meal mite	718.00	718.00																		x															
Sarcoptes scabei itch	735.00	735.00																				x													
Ornithonyssus bird mite	877.00	878.00																																	

s = slime mold, m = mold, b = bacteria, v = virus, y = yeast, p = parasite (one-celled animals), t = tapeworm, x = mite

Sources

This list was accurate as this book went to press. Only the vitamin sources listed were found to be pollution-free, and only the herb sources listed were found to be potent, although there may be other good sources that have not been tested. The author has no financial interest in, influence on, or other connection with any company listed, except for having family members in the Self Health Resource Center.

Note to readers outside the United States of America:

Sources listed are typically companies within the United States because they are the ones I am most familiar with. You may be tempted to try a more convenient manufacturer in your own country and hope for the best. I must advise against this! In my experience, an uninformed manufacturer most likely has a polluted product! Your health is worth the extra effort to obtain the products that make you well. One bad product can keep you from reaching that goal. This chapter will be updated as I become aware of acceptable sources outside the United States. Best of all is to learn to test products yourself.

Item	**Source**
Amber bottles, ½ ounce	Drug store, Continental Glass & Plastic, Inc. (large quantities)
Apricot Kernel Oil	Now Foods

Arginine	Now Foods, Jomar Labs
Artemesia (wormwood) seed	R. H. Shumway
Baking soda (sodium bicarbonate)	Spectrum Chemical Co.
Beet root	San Francisco Herb & Natural Food Co.
Belts for clothes dryer	Three that tested negative to asbestos are: Maytag™ 3-12959 Poly-V belt, Whirlpool™ FSP 341241 Belt-Drum Dr. (replaces 660996), and Bando™ V-Belt A-65. Bando American makes other belts, some of which might be the right size for your dryer. Call for a dealer near you, make sure it says "Made In America", right on the belt.
Black cherry concentrate	Health food store
Black Walnut Hull Tincture	Self Health Resource Center, New Action Products
Borax, pure	Grocery store
Boric acid, pure	Now Foods, health food store, pharmacy
Cascara sagrada	Natures Way, health food store
Chemicals for testing.	Aldrich Chemical Co., Spectrum Chemical Co., ICN Biomedicals, Inc. (research chemicals only, including genistein), Boehringer Mannheim Biochemicals (research only)
Cholecalciferol	Spectrum Chemical Co.
Citric acid	Now Foods or health food store
Cloves	San Francisco Herb & Natural Food Co. (ASK for fresh)
Cornstarch	Lady Lee is the only unpolluted brand found so far
Dental information	Tooth Truth, by Frank J. Jerome, available from New Century Press.
EDTA	Spectrum Chemical Co.
Electronic parts	A Radio Shack near you.
Empty gelatin capsules size 00	Now Foods, or your health food store

Filters, pure charcoal	Pure Water Products (pitchers), Seagull Distribution Co. (faucet, shower, whole house)
Flaxseed oil	Omega Nutrition USA, Inc.
Fucus	San Francisco Herb & Natural Food Co.
Ginger capsules	Now Foods
Goldenrod tincture	Dragon River Herbals, Blessed Herbs
Grain alcohol	Liquor store, as vodka, Everclear^{TM}
Gravel root (herb)	San Francisco Herb & Natural Food Co.
Histidine	500 mg from Jomar Labs
Homeopathic remedies	Dolisos America, Inc., and others
Hydrangea (herb)	San Francisco Herb & Natural Food Co.
Hydrogen peroxide 35% (food grade)	New Horizons Trust
Iodine, pure	Spectrum Chemical Co.
Lugol's iodine	For slide staining (not internal use) from Spectrum Chemical Co., and farm animal supply store. For internal use must be made from scratch.
Lysine	Bronson Pharmaceuticals
Magnesium oxide	Bronson Pharmaceuticals
Marshmallow root (herb)	San Francisco Herb & Natural Food Co.
Microscope slides and equipment	Carolina Biological Supply Company, Ward's Natural Science, Inc., Southern Biological Supply Company, Fisher-EMD
Microscopes	Carolina Biological Supply Company, Ward's Natural Science, Inc., Edmund Scientific Co.
Milk Thistle	Herb companies.
Mullein tea	San Francisco Herb & Natural Food Co.
Niacin	100 mg or 250 mg time release, Bronson Pharmaceuticals
Ornithine	Now Foods, Jomar Labs
Ortho-phospho-tyrosine	Aldrich Chemical Co.
Ozonator	Superior Health Products, LLC
P24 antigen sample	Bachem Fine Chemicals Inc.
PABA (para amino benzoic acid)	Bronson Pharmaceuticals
Pantothenic acid	Bronson Pharmaceuticals

Peroxy	See Hydrogen peroxide.
Photo-micrographic camera and film	Ward's Natural Science, Inc.
Potassium bicarbonate	Spectrum Chemical Co., or pharmacy
Potassium chloride	Now Foods, Spectrum Chemical Co.
Potassium iodide, pure	Spectrum Chemical Co.
Rascal	Kroeger Herb Products, New Action Products (as Raz-Caps)
Salt (sodium chloride), pure	Spectrum Chemical Co., get USP grade
Sodium alginate	Spectrum Chemical Co. or health food store
Stevia powder	Now Foods
Tea	San Francisco Herbs, buy in bulk, any single herb variety.
Thallium, homeopathic	Dolisos America, Inc.
Thioctic acid (also called lipoic acid)	Maypro Industries
Uva Ursi	Natures Way or health food store
Vanilla extract	Durkee's^{TM} or Mexican brands
Vegetable glycerin	Now Foods or health food store
Vitamin B Complex	Bronson Pharmaceuticals
Vitamin B_2 (riboflavin)	Nutrition Headquarters (250 mg), or Bronson Pharmaceuticals (100 mg)
Vitamin B_6	Bronson Pharmaceuticals, EDOM Labs
Vitamin B_{12} lozenges	Bronson Pharmaceuticals
Vitamin C (ascorbic acid)	Self Health Resource Center
Vitamin D 50,000iu	From dentist, by prescription, or see Recipes.
Vitamin E capsules	Bronson Pharmaceuticals
Vitamin E Oil	Now Foods
Washing soda (sodium carbonate)	Grocery store
Water filter pitchers	Pure Water Products
Wormwood capsules	Self Health Resource Center, Kroeger Herb Products, New Action Products
Zinc	Bronson Pharmaceuticals
Zinc oxide	Spectrum Chemical Co.

Aldrich Chemical Co.
P.O. Box 355
Milwaukee, WI 53201
(414) 273-3850

Bachem Fine Chemicals Inc.
3132 Kashiwa St.
Torrance, CA 90505
(310) 539-4171

Bando American Inc.
1149 West Bryn Mawr
Itasca, IL 60143
(800) 829-6612
(630) 773-6600

Blessed Herbs
109 Barre Plaines Rd.
Oakham, MA 01068
(508) 882-3839

Boehringer Mannheim
Biochemicals
9115 Hague Rd.
P.O. Box 50414
Indianapolis, IN 46250
(800) 262-1640
(317) 849-9350

Bronson Pharmaceuticals
Div. of Jones Medical Industry
1945 Craig Road
P.O. Box 46903
St. Louis, MO 63146-6903
(800) 235-3200 retail
(800) 525-8466 wholesale

Carolina Biological Supply Co.
2700 York Rd.
Burlington, NC 27215
(800) 334-5551
(919) 584-0381

Continental Glass & Plastic, Inc.
841 West Cermak Rd.
Chicago, IL 60608-4582
(312) 666-2050

Dolisos America, Inc.
3014 Rigel Ave
Las Vegas, NV 89102
(800) 365-4767
(702) 871-7153

Dragon River Herbals
PO Box 74
Ojo Caliente, NM 87549
(800) 813-2118
(505) 583-2118

Edmund Scientific Co.
101 E. Gloucester Pike
Barrington, NJ 08007
(856) 573-6250

EDOM Labs, Inc.
860 Grand Boulevard
P.O. Box 780
Deer Park, NY 11729
(516) 586-2266
(800) 723-3366

Fisher Scientific EMD
485 S. Frontage Rd.
Burr Ridge, IL 60521
(800) 955-1177

ICN Pharmaceuticals, Inc.
Biomedical Division
3300 Hyland Ave.
Costa Mesa, CA 92626
(714) 545-0113
(800) 854-0530

Jomar Labs
251-B East Hacienda Avenue
Campbell, CA 95008
(800) 538-4545
(409) 374-5920

Kroeger Herb Products
(wholesale:)
805 Walnut St.
Boulder, CO 80302
(800) 225-8787
(303) 443-0261
(for retail call:)
Hanna's Herb Shop
5684 Valmont Rd.
Boulder, CO 80301
(800) 206-6722

Maypro Industries
2700 Westchester Ave.
Purchase, NY 10577
(914) 251-0701

Nature’s Way
10 Mountain Springs Pkwy
Springville, UT 84663
(800) 962-8873
(801) 489-1500

New Action Products (USA)
P.O. Box 540
Orchard Park, NY 14127
(800) 455-6459
(716) 662-8000

New Action Products (CANADA)
PO Box 141
Grimsby, Ont. L3M4G5 Canada
(800) 541-3799
(716) 873-3738

New Century Press
1055 Bay Blvd., Suite C
Chula Vista, CA 91911
(800) 519-2465
(619) 476-7400

New Horizons Trust
53166 St. Rt. 681
Reedsville, OH 45772
(800) 755-6360
(614) 378-6366

Now Foods
395 Glen Ellen Road
Glen Ellen, IL 60108
(630) 545-9098

Omega Nutrition USA, Inc.
6515 Aldrich Rd.
Bellingham, WA 98226
(800) 661-3529

Superior Health Products, LLC
13549 Ventura Blvd.
Sherman Oaks, CA 91423
(800) 700-1543
(818) 986-9456

Pure Water Products
10332 Parkview Ave.
Westminster, CA 92683
(800) 478-7987

R. H. Shumway
P.O. Box 1
Graniteville, SC 29829
(803) 663-9771

San Francisco Herb & Natural
Food Co.
1010 46th St.
Emeryville, CA 94608
(800) 227-2830 wholesale
(510) 601-0700 retail

Seagull Distribution Co.
3372 Baltimore St.
San Diego, CA 92117
(858) 483-0264

Self Health Resource Center
1055 Bay Blvd., Suite A
Chula Vista, CA 91911
(800) 873-1663
(619) 409-9500 (USA)
(780) 475-2403 (Canada)

Southern Biological Supply
Company
P.O. Box 368
McKenzie, TN 38201
(800) 748-8735
(901) 352-3337

Spectrum Chemical Co.
14422 South San Pedro Street
Gardena, CA 90248
(800) 772-8786
(310) 516-8000

Ward’s Natural Science, Inc.
5100 West Henrietta Road
Rochester, NY 14586
(716) 359-2502
(800) 962-2660

Finale

I hope you have come to the same happy conclusion that I did a few years ago! That we humans don't have hundreds of different maladies and disturbances. **We only have two!** Things that crawl or climb into us. And toxins: unnatural chemicals that we unknowingly inhale or consume.

The living things are both large and small: from worms we can see, to microscopic bacteria, viruses and fungi. The non living things are pollutants in our air, food, dental metal and body products. Taking in a lot of pollutants hampers the body's ability to kill and get rid of the invaders. And so, gradually, as we get older or sicker, the body's invaders get the upper hand and take over. Don't be discouraged if you have lupus, cerebral palsy, cirrhosis, or any complex-sounding disease. Every disease is an example of the same process.

The good news is that our body can reclaim its sovereignty by throwing the rascals out. We must assist by throwing the pollutants out. Fortunately, we don't have to do the whole job ourselves, we only need to assist. Our body has miraculous powers to clean itself up.

By reducing chronic disease to two problems, it becomes manageable. Ill health, even aging, can be reversed. Health comes back and rewards you.

With the new electronic insights and technology, our parasitic invaders can be vanquished with the closing of a switch. Preventing reinfection is the bigger challenge. Similarly, pollutants can be uncovered in as short a time as days, short enough to turn any sick, even terminal, sufferer's verdict around before tragedy occurs.

The tragedies of surgery, organ replacements, radiation, chemotherapies, doses of drugs, even death can be avoided. Reversing illness and turning into a shining example of your former healthy self could be the most exciting adventure of your life.

Killing your invaders is an easy matter: you simply purchase or build the device that can do that and take the proper herbs. Cleaning up dentalware is under your control, too—a financial expense not beyond your reach, hopefully. Trading your body products for unpolluted varieties is a job but not insurmountable. Cleaning your environment may be the stumbling block. If you can't unpollute your air, water, carpets, furniture, move. Move to a healthier dwelling! Get rid of it all. Like a cat that moves her kittens to a safer place, just move.

The healthiest house is no house. If you have been quite ill, move far enough south to avoid heating and cooling. Sit outside in the shade all day. Use your new wisdom and sharp eye to choose a new dwelling as free of pollutants as you can. No refrigerator indoors. No window air conditioner. No fiberglass insulation. No fossil fuels. No attached garage. No carpets or stuffed furniture, no foam bedding, no fresh paint, no pesticide. Go primitive. Health is primitive. You were born primitive—with health. Even if you weren't, you can undo much of the “inherited” damage.

Don't listen to the new doomsayers, who persuade you nothing can be done unless a gene is replaced. Your genes have been reliable for millions of years. Genes are the most reliable of all biological chemicals in your body. They are not faulty. They are hampered in their tasks. They are commandeered by metals and other species’ genes. Those of parasites, bacteria, and viruses. They wouldn't be there if pollutants weren't there. They allow invaders into the most jealously guarded recess of your being: your genes. But now you can throw the rascals out and reclaim your territory. It is not new genes that you need. You simply need your own genes back on the job, directed by your own body, working for you.

Index

A

abdominal pain, 99, 203, 268, 299
abscess, 54, 79, 135, 153, 158, 159, 163, 164, 216, 231, 510
Acanthocephala, 138, 274, 566
acetic acid, 296
acetone, 122, 162, 208, 320
acetylcholine, 221, 238
acidophilus, 21, 100, 564, 576
acne, 167, 192, 193, 194
acupuncture, 51, 465
Adenovirus, 13, 176, 179, 199, 357, 358, 359, 360, 361, 363, 368, 563, 577
adrenal gland, 60, 113, 114, 143, 210, 212, 217, 281, 291, 373
aflatoxin, 38, 177, 180, 186, 225, 261, 280, 348, 349, 381, 382, 383, 384, 386, 390, 392, 422, 491, 493, 494, 519
AIDS, 11, 35, 36, 90, 119, 126, 131, 149, 252, 351, 352, 353, 356, 368, 372, 385, 442, 443, 458, 493, 500
air conditioner, 40, 284, 313, 323, 346, 403, 446, 590
air filter, 137, 284
albuminuria, 417
alcoholism, 227, 228, 229, 230, 386
alkaline phosphatase, 80, 217
alkaline tide, 57
allergy, 85, 102, 105, 107, 111, 126, 135, 139, 140, 163, 169, 175, 190, 191, 216, 222, 223, 224, 225, 226, 242, 246, 258, 386, 495, 553, 558
aluminum, 37, 112, 137, 141, 171, 180, 181, 187, 189, 212, 244, 247, 269, 271, 274, 305, 310, 311, 314, 415, 421, 423, 424, 430, 431, 433, 434, 435, 436, 438, 439, 450, 461, 473, 484, 485, 486, 487, 514, 521, 525, 526, 533, 543
aluminum silicate, 37, 138, 141, 212, 271
alveolar cavitational osteopathosis, 410
Alzheimer's, 33, 219, 250, 252, 269, 270, 272, 372, 495, 526
amalgam, 37, 65, 71, 107, 115, 157, 180, 409, 410, 413, 414, 416, 417, 418, 475, 485
amalgam tattoo, 410
American Dental Association (ADA), 37, 409, 414, 416
amino acid, 85, 232, 244
ammonia, 125, 243, 244, 320, 326
amoeba, 106, 274
Ancylostoma, 256, 257, 262, 566, 578
anemia, 76, 162, 279, 285, 286
anesthesia, 90
angina, 176, 417
animal dander, 104, 135
animal feed, 125, 347, 574
ankle swelling, 68
ant repellent, 540
antibiotic, 134, 141, 170, 181, 248, 548
antimony, 105, 112, 123, 128, 140, 187, 435
antiseptic, 98, 344, 399, 436, 449
aorta, 478
appendicitis, 101, 268
appendix, 97, 101, 128
appetite, 134, 240, 268, 275, 288, 292, 300, 306, 376, 379, 388, 417
apple cider vinegar, 277, 348, 423, 493, 572
apricot kernel oil, 534, 535, 537
arginine, 120, 137, 210, 212, 425
arm, 49, 51, 118, 153, 154, 254, 255, 282, 321, 481, 482, 488, 531, 553, 576
armpit, 89, 143, 144, 193, 531, 532
arsenic, 4, 40, 68, 96, 100, 103, 110, 132, 137, 138, 141, 150, 160, 162, 168, 171, 185, 186, 187, 193, 206, 236, 242, 259, 268, 271, 282, 313, 314, 361, 403, 448
arthralgia, 417
arthritis, 51, 78, 80, 81, 82, 83, 84, 85, 86, 117, 137, 146, 151, 155, 214, 268, 419
artificial sweetener, 173, 176, 177, 252, 302, 520, 574

asbestos, 4, 39, 80, 81, 91, 100, 103, 108, 110, 133, 137, 138, 142, 143, 145, 148, 150, 167, 168, 170, 207, 209, 210, 214, 231, 271, 282, 314, 316, 336, 337, 349, 447, 448, 474, 582
Ascaris, 10, 32, 43, 76, 78, 79, 80, 82, 86, 103, 104, 105, 107, 111, 122, 127, 134, 135, 136, 137, 138, 139, 140, 141, 142, 162, 166, 167, 176, 179, 186, 187, 192, 193, 194, 196, 212, 213, 227, 230, 231, 233, 234, 235, 236, 242, 256, 257, 259, 260, 264, 266, 267, 268, 274, 285, 286, 293, 299, 343, 371, 372, 375, 404, 566, 578
ascorbic acid, 516, 534, 535, 537, 584
aspirin, 169, 170, 214, 264
asthma, 43, 44, 105, 107, 135, 136, 137, 138, 139, 140, 220, 230
athlete's foot, 247, 248
Atomic Absorption Standard, 484, 572
atrophy, 384, 418
Atypical Lateral Sclerosis, 47
autism, 263, 264, 266
auto exhaust, 222, 279, 282, 316

B

back pain, low, 68, 81, 96, 123, 131
back pain, upper, 81, 96, 151, 553, 558
Bacteroides fragilis, 135, 230, 231, 266, 372, 563, 576
baking powder, 269, 430, 485
baking soda, 57, 58, 269, 431, 438, 515, 516, 527, 532, 533
baldness, 126, 190
bandwidth, 7, 8, 10, 16, 18, 506, 508, 509, 561, 562
barium, 75, 123, 139, 143, 187, 207, 415, 435, 436, 474, 476
bath oil, 247, 439
beef, 147, 368, 423, 424, 477
beer, 194, 327, 348, 378, 380, 386, 493, 572
beet root, 539, 540
benzalkonium, 137, 138, 170, 180, 187, 244, 390, 435, 439
benzoic acid, 55, 84, 583
benzopyrene, 304, 431
beryllium, 78, 124, 138, 142, 145, 157, 166, 170, 187, 192, 224, 227, 228, 229, 235, 407
beta blocker, 289, 319
BHA, 232, 233, 265, 517, 529
BHT, 232, 233, 265, 517, 529
bile, 47, 48, 80, 88, 91, 98, 100, 129, 130, 131, 132, 150, 151, 152, 153, 191, 223, 224, 266, 267, 299, 300, 477, 553, 554, 555, 556, 557, 558, 559
bile duct, 47, 48, 98, 129, 130, 131, 150, 151, 152, 153, 191, 223, 224, 299, 300, 477, 553, 557, 558
bioradiation, 12, 13
birth control pill, 111, 116
birth defect, 111, 418
bismuth, 105, 166, 185, 208, 242, 271, 435, 474
black cherry concentrate, 549
Black Walnut Hull Tincture, 21, 101, 102, 228, 235, 338, 339, 341, 342, 344, 345, 347, 389, 543, 544, 545, 547, 582
bladder, 55, 60, 101, 108, 112, 124, 129, 130, 145, 166, 190, 196, 198, 293, 321, 550, 552, 555, 559
bleach, 40, 150, 246, 256, 258, 260, 284, 314, 397, 437, 444, 449, 526, 530, 540, 572
bleeding, 95, 96, 110, 112, 196, 249, 286, 536
bleeding gum, 95
bloating, 108, 113, 129, 254, 266, 299, 443, 546, 555
blood pressure, 19, 49, 75, 82, 95, 115, 148, 149, 210, 211, 212, 213, 273, 275, 280, 281, 282, 289, 311, 321, 377, 388, 417, 515
blood sugar, 173, 175, 176, 177, 178, 182, 187, 275, 305, 308
blood test, 5, 77, 80, 82, 106, 107, 114, 139, 145, 178, 179, 213, 235, 245, 319
blood vessel, 69, 70, 154, 163, 279, 280, 281, 282, 318, 321
body lotion, 170
bone cancer, 196
bone infection, 56, 79, 315, 410
bone marrow, 78, 181, 285, 351, 353, 477
bone spur, 86
Borellia, 170, 563, 577

boric acid, 313, 403, 449, 540, 541, 542
boron, 215
bowel, 20, 21, 32, 34, 35, 54, 77, 82, 83, 97, 99, 101, 102, 103, 104, 106, 108, 128, 129, 152, 163, 166, 169, 197, 202, 219, 223, 224, 238, 254, 256, 267, 270, 298, 301, 331, 338, 364, 388, 389, 425, 443, 494, 510, 530, 546, 547, 548, 558
bowel disease, 102, 389, 443
bowel movement, 21, 34, 35, 77, 82, 83, 98, 101, 106, 128, 129, 197, 202, 224, 238, 256, 267, 301, 331, 494, 530, 558
Bowel Program, 99, 101, 102, 127, 128, 164, 202, 237, 238, 254, 257, 262, 270, 298, 315, 328, 346, 364, 444, 546, 549
braces, 418, 476
breast, 142, 143, 144, 145, 146, 235, 332, 336, 337, 342, 376, 384, 420, 435, 476, 491
breast cancer, 143, 145, 332, 384, 420
breast milk, 235, 342, 376
breastbone, 107, 131, 149
bromine, 239, 258, 271
bronchiestasis, 138
bronchitis, 78, 140, 141, 193, 217
burning, 186, 253, 254, 369, 534
burping, 295, 550
bursitis, 51, 151, 152, 558, 559
butyl nitrite, 162

C

cachexia, 240
cadmium, 37, 47, 69, 70, 74, 75, 79, 96, 107, 112, 146, 175, 185, 192, 210, 211, 213, 267, 268, 269, 314, 415, 418, 426, 449, 450, 466, 484
caffeine, 67, 68, 75, 89, 117, 137, 145, 185, 210, 212, 213, 255, 259, 264, 276, 290, 317, 377, 425
calcium, 57, 63, 64, 65, 66, 67, 74, 80, 81, 83, 87, 88, 93, 94, 96, 106, 129, 158, 212, 215, 216, 217, 231, 255, 276, 277, 287, 299, 305, 361, 399, 421, 517, 521
calcium propionate, 361
calf cramp, 68
Campylobacter, 69, 76, 110, 114, 170, 298, 563, 576, 577
Candida, 20, 185, 186, 189, 229, 246, 247, 248, 492, 563, 577
candidiasis, 176, 246
candy, 181, 218, 219, 313, 354
carbon dioxide, 139, 242, 267, 279, 282
carbon filter, 258, 270, 283, 314, 405, 426
carbon monoxide, 267, 282
carbon tetrachloride, 96, 103, 104, 116, 122, 127, 137, 148, 162, 186, 260, 440
carbonic anhydrase, 296
carcinogen, 40, 44, 374, 403, 423
cardiomyopathy, 147
cardiovascular, 417
carpet cleaning fluid, 429
Cascara sagrada, 99, 164, 238, 301, 547, 582
cat liver fluke, 116, 268
cavitation, 117, 141, 156, 291, 410, 533
cayenne, 55, 59, 84, 180, 181, 320, 398
cercaria, 34, 111, 117, 122, 495
cerebellum, 111, 206, 208, 484
cerebral palsy, 589
cerebrum, 111, 206, 208, 271, 274, 484
cerium, 145, 162, 187
cervix, 112, 113
cesium, 104, 157, 259, 260, 449
cheese sauce, 306, 434, 524
chelation, 213, 271, 272, 288, 289, 292
chemotherapy, 437
chest, 51, 74, 78, 148, 149, 150, 151, 167, 177, 178, 192, 254
chest pain, 51, 148, 149, 151, 167, 177, 178, 254
chewing gum, 95, 218
chicken, 84, 99, 126, 139, 147, 176, 179, 233, 235, 265, 266, 268, 274, 347, 353, 371, 394, 423, 424, 431, 477, 492
chicken pox, 176, 179, 353, 371, 492
chiropractic, 86, 91, 92, 156
Chlamydia, 110, 114, 563, 577
chloride, 37, 70, 72, 139, 162, 208, 221, 269, 311, 373, 390, 428, 429, 440, 475, 476, 497, 516, 573, 574, 584

chlorine, 137, 150, 239, 246, 257, 258, 260, 271, 282, 283, 288, 314, 437, 449, 526, 530, 540
chlorofluorocarbons (CFCs), 143, 313, 446
choke, 7, 11, 486
cholesterol, 109, 122, 129, 130, 224, 374, 421, 432, 434, 440, 553, 554, 558, 559
chromium, 37, 176, 178, 184, 302, 375, 377, 418, 431, 435, 436
chronic fatigue syndrome, 111, 186
cigarette, 336
cinnamon tea, 121
circulation, 55, 59, 63, 156, 160, 182, 273, 320
cirrhosis, 382, 589
citric acid, 403, 515, 516, 521, 528, 530, 535, 542
Clostridium, 76, 206, 315, 563, 564, 577
clothes dryer belt, 39, 142, 232, 314, 447, 474, 572
clotting, 19, 259, 320
cobalt, 4, 96, 112, 128, 141, 147, 162, 189, 194, 314, 399, 474, 526
coffee, 61, 81, 82, 89, 174, 210, 268, 274, 275, 278, 308, 327, 335, 377, 381, 388, 425, 427, 428, 431, 536
cold cereal, 109, 111, 162, 178, 219, 252, 335, 376, 386, 387, 392, 420, 518, 574, 575
cold sore, 179, 180
colitis, 105, 106, 107, 167, 425
cologne, 144, 242
colon, 78, 97, 131, 164, 337, 357, 361, 389, 478, 491
comfrey, 137, 139, 552
composite, 71, 119, 126, 146, 177, 226, 254, 291, 416
condom, 480
congestive heart failure, 318
constipation, 104, 166, 178, 238, 299, 301, 548
cooked cereal, 109, 144, 178, 287, 291, 376, 386, 442, 556
cookie, 23, 269, 431
copper, 22, 26, 30, 37, 70, 96, 100, 113, 115, 116, 119, 128, 147, 157, 171, 208, 211, 212, 261, 262, 310, 314, 337, 349, 426, 450, 459, 462, 463, 466, 472, 474, 484, 517
corn, 66, 81, 106, 144, 215, 282, 302, 303, 306, 307, 308, 351, 355, 356, 361, 379, 385, 394, 420, 434, 460, 476, 492, 532, 537, 539
cornstarch, 181, 247, 249, 294, 354, 374, 531, 532, 534, 537, 540
corpus luteum, 115, 124
cortisone, 52, 59, 138, 246
cosmetics, 37, 143, 144, 208, 326, 335, 437, 476, 572
cottage cheese, 80, 201, 254, 432, 485, 510, 511
Coxsackie virus, 1, 2, 135, 137, 166, 233, 371, 564, 576, 577
cramp, 69, 111, 249, 253
crawl space vents, 127, 181
Crohn's disease, 33, 104, 106
crotch itch, 247
croup, 135, 140
crown, 411, 413, 414
Cryptocotyl, 74
cutlery, 126, 127, 379, 430, 433
cyst, 32, 33, 34, 39, 74, 90, 91, 92, 112, 145, 170, 171, 188, 189, 198, 200, 255
cysteine, 74, 94, 96, 108
cysticercus, 197, 198, 199, 200, 357, 358, 494
cystine, 69, 94
Cytochalasin B, 387, 562, 576
Cytomegalovirus (CMV), 147, 179, 184, 564, 578

D

decaffeinated, 109, 210, 269, 335, 428, 574, 575
decane, 122, 162, 188, 440
dementia, 278, 280, 322, 330
dental cleanup, 73, 78, 79, 87, 120, 127, 145, 155, 156, 220, 262, 346
dental floss, 70, 162, 481
dental metal, 68, 70, 78, 110, 112, 115, 116, 117, 119, 122, 134, 140, 141, 143, 145, 159, 162, 187, 206, 222, 223, 226, 233, 254, 261, 270, 337, 419, 485, 589
denture, 139, 295, 411, 533
denture cream, 295
deodorant, 75, 128, 138, 143, 144, 147, 162, 181, 189, 190, 193, 222, 231, 435, 531, 532, 572, 574

depilatory, 417
depression, 44, 78, 104, 113, 116, 117, 185, 186, 210, 228, 254, 256, 257, 259, 260, 261, 262, 298, 540, 565
dermatron, 6
dessert, 305, 306, 326
diabetes, 3, 35, 43, 44, 55, 173, 174, 175, 176, 177, 178, 250, 251, 278, 302, 305, 309, 388, 495
diabetic, 67, 89, 174, 175, 305, 308, 309
diaper, 189, 246, 256
diaphragm, 76, 132, 133, 192, 297, 375
diarrhea, 97, 103, 104, 107, 108, 117, 134, 152, 202, 300, 344, 421, 425, 558
diesel exhaust, 75
diet clean-up, 80
diet drink, 429
diet, 7 day diabetic, 302
diet, 7 day sample, 523
digestion, 34, 88, 104, 105, 129, 130, 131, 148, 177, 226, 255, 266, 267, 276, 277, 278, 287, 295, 296, 297, 299, 374, 377, 423, 552
digestive problem, 95, 130, 146, 268, 559
digestive tract, 103, 246, 266
Dirofilaria, 32, 80, 141, 146, 147, 148, 318, 567, 578
diuretic, 75, 213, 282, 294, 304, 321, 373
dizziness, 186, 275, 297, 298, 419
dropsy, 321
drug, 44, 82, 131, 136, 234, 238, 289, 290, 301, 312, 316, 319, 321, 322, 344, 393, 473, 533
dry skin, 525
dry socket, 410
duodenum, 34, 478
dysprosium, 254

E

E. coli, 21, 62, 97, 98, 99, 102, 116, 129, 164, 177, 266, 298, 353, 357, 360, 361, 363, 364, 388, 444, 547, 548, 564, 576, 577
ear, 6, 104, 105, 134, 140, 168, 169, 170, 217, 242, 325, 361, 419, 467
ear infection, 104, 105, 134, 140, 169
ear, inner, 6, 170, 361
earache, 169
earthworm, 10, 510
Echinacea, 99
eclampsia, 95
eczema, 140, 192, 230, 492
edema, 60, 322, 515
EDTA, 73, 271, 272, 289, 582
egg dish, 432
elbow, 51, 154, 467
electronic testing, 5
ELISA, 495, 496, 497
emphysema, 46, 139
enamel, 214, 310, 413, 425, 430, 501, 529, 533, 543
encephalitis, 1, 2, 356
Endolimax, 106, 138, 567, 577, 579
endometriosis, 33, 35, 103, 110, 111, 122, 250, 495
endometrium, 250
enema, 166
environmental illness, 43
eosinophil, 136, 139, 185, 245
eosinophilia, 417
Epsom salts, 99, 150, 151, 300, 555, 556, 557
Epstein Barre Virus (EBV), 112, 179, 182, 184, 185, 186, 204, 371, 564, 577
ergot, 162, 226, 227, 228, 229, 232, 233, 253, 263, 270, 277, 280, 383, 385, 386, 392, 520
esophageal sphincter, 297
esophagus, 131, 132, 149
estrogen, 91, 113, 116, 119, 122, 143, 144, 145, 185, 217, 384, 539
ether, 575
europium, 124, 140, 187, 207
Eurytrema, 21, 109, 173, 371, 567, 578
Eustachian tube, 168, 361
exercise, 2, 86, 92, 301, 327
eye, 3, 14, 76, 140, 160, 161, 162, 175, 186, 259, 478, 497, 498, 528, 539, 590
eye disease, 3, 161, 175
eye makeup, 140, 186
eyelid, 146, 539

F

face, 36, 123, 181, 192, 193, 194, 209, 220, 227, 231, 253, 258, 280, 283, 369, 400, 448, 469, 540

face powder, 231, 540
fallopian tube, 250
Fasciola, 43, 249, 567, 578
Fasciolopsis buskii, 33, 150, 166, 250, 331, 371, 490, 567, 579
fatigue, 2, 4, 73, 95, 111, 113, 134, 139, 176, 177, 182, 183, 184, 185, 186, 187, 204, 245, 289, 311, 368, 417
fennel, 444, 547, 548
fenugreek, 175, 176, 370, 422, 521
fetus, 258, 417
fever, 78, 201, 203, 204, 234, 365, 369
fiberglass, 39, 40, 80, 91, 100, 110, 133, 137, 143, 147, 150, 160, 168, 177, 241, 282, 284, 314, 336, 349, 403, 405, 447, 590
fibromyalgia, 51, 75, 76, 78
fibromyositis, 75
filaria, 31, 318
finger, 86, 136, 155, 195, 253, 282, 308, 321, 463, 486, 487, 508, 530, 539
fingernail, 501
fireplace lighter, 165
fish line, 95, 218, 248, 296, 438, 533
flatworm, 33, 331
flavoring, 304, 378, 428, 548, 574, 575
flaxseed, 127
floss, 70, 72, 95, 162, 218, 296, 438, 481, 532
flour, 298, 306, 310, 361, 362, 379, 385, 423, 474, 572
flu, 78, 134, 168, 176, 201, 202, 203, 275, 425, 564, 576
fluke disease, 117, 219, 221, 249, 250, 252, 494, 495
fluke egg, 145, 332
fluoride, 82, 112, 206, 296, 573
foam, 40, 194, 284, 314, 337, 404, 447, 449, 533, 573, 590
foam bedding, 314, 337, 590
foam filter, 284
foam furniture, 40, 449
folic acid, 120
Food and Drug Administration (FDA), 436
foot, 51, 57, 59, 60, 68, 247, 248, 328, 419, 506
formaldehyde, 40, 100, 108, 110, 112, 133, 137, 150, 160, 165, 168, 193, 236, 241, 282, 284, 314, 361, 448, 449, 499, 500, 571
fragrance, 105, 136, 166, 185, 247, 253, 259, 284, 529, 530, 550, 572
Freon, 100, 160, 336, 446, 469
Fucus, 239, 548, 583

G

gadolinium, 124, 140, 207
Gaffkya, 170, 564, 576
gallbladder, 80, 128, 132, 191, 477, 553, 559
gallium, 123, 124, 146
gallstone, 81, 150
galvanized pipe, 70, 210, 211, 450
ganglia, 180
Gardnerella, 69, 91, 102, 110, 114, 115, 119, 564, 576
garlic, 137, 376, 548
gas hot water heater, 431
gas leak, 105, 140, 141, 149, 165, 166, 170, 171, 175, 187, 241, 255, 260, 268, 279, 319, 450, 573
gas pipe, 96, 165
gas stove burner, 476
gas, household, 231
gasoline, 124, 127, 162, 208, 221, 224, 227, 316, 416, 445, 472, 572, 574
gene, 3, 46, 206, 207, 209, 590
genital tract, 116
germanium, 417
Giardia, 9, 103, 160, 239, 399, 567, 578
ginger, 205, 514, 515
ginseng, 259, 263
glaucoma, 146, 213, 214
glucose tolerance factor, 184, 419
glutamic acid, 228
glutamine, 228, 229
gluten, 105, 175
glycerin, 515, 516, 534, 535, 536, 537, 538, 539, 540, 549, 584
goat milk, 107
gold, 102, 103, 111, 113, 124, 126, 146, 175, 177, 184, 226, 237, 310, 412, 418, 419, 420, 475
goldenrod, 549
gonorrhea, 91, 116, 565, 576
gout, 69
granola, 186, 354, 421, 518
gravel root, 552
grilled food, 356
groin, 90, 157, 166, 248

growing pains, 69
gum disease, 81
gums, 81, 87, 95, 217

H

HA wart virus, 176
Haemophilus, 141, 242, 564, 576
hafnium, 139, 222, 573
hair dryer, 241, 282, 316, 346, 447, 448
hair spray, 122, 136, 139, 144, 187, 222, 335, 439, 573
hamburger, 219, 349
hand cream, 354
handhold, 26, 27, 30, 457, 462, 463, 464, 487, 488, 503, 504, 506, 508
hCG, 336
head, 31, 32, 52, 112, 131, 144, 156, 161, 169, 171, 184, 198, 199, 254, 283, 315, 359, 361, 369, 417, 461, 557, 569
headache, 51, 75, 105, 116, 162, 164, 165, 167, 170, 206, 257, 369, 400, 419, 425
hearing, 4, 6, 7, 242, 261, 274, 278, 324, 325, 387, 468, 504
heart, 4, 32, 59, 80, 114, 138, 139, 141, 146, 147, 148, 149, 159, 168, 176, 177, 178, 182, 194, 195, 213, 254, 278, 279, 280, 290, 291, 311, 317, 318, 319, 320, 321, 322, 323, 325, 403, 419, 478, 515, 526
heart disease, 147, 176, 194, 403, 515, 526
heart failure, 278, 318, 322
heartburn, 82, 375
heartworm (see also Dirofilaria), 32, 138, 140, 141, 146, 147, 148, 149, 167, 178, 254, 318, 319, 343, 567
heel, 61, 63, 65, 67, 68
hemoglobin, 279, 286, 287
herb tea, 63, 66, 93, 95, 140, 321, 335, 365, 390, 424, 428, 429, 432, 524, 551, 574
herbicide, 40
hernia, 90, 133, 154, 268, 297
Herpes, 9, 20, 115, 116, 161, 179, 180, 181, 182, 371, 487, 488, 492, 564, 576, 578
hexane, 440
hexanedione, 111, 148, 193, 269, 574
hiatal hernia, 133, 154, 268, 297
high blood pressure, 49, 82, 95, 149, 211, 212, 213, 311, 388, 417, 515
hip, 51, 87, 89, 275, 419
hippuric acid, 55, 84, 85, 115
histamine, 135
histidine, 127, 207
Histoplasma, 145, 166, 564, 576
HIV, 10, 33, 90, 149, 150, 189, 204, 250, 251, 252, 342, 351, 352, 353, 356, 371, 458, 487, 490, 491, 495, 500, 501, 564
hives, 68, 89, 122, 191, 553
hookworm, 78, 79, 86, 105, 139, 160, 167, 194, 260, 264, 286, 343
hormone, 2, 91, 107, 113, 115, 116, 119, 122, 216, 368, 384
hot flush, 114, 115, 117, 175, 255
human liver fluke, 33, 74, 96, 109, 111, 116, 122, 128, 186, 188, 207, 208, 224, 249, 259, 260, 498
hurricane lamp, 78, 145, 166, 227, 572
hydrazine sulfate, 240
hydrochloric acid, 296, 297
hydrogen peroxide, 95, 139, 218, 296, 383, 438, 532, 533, 542
hydrogen sulfide, 267
hyperactive, 203
hyperesthesia, 417
hyperthyroidism, 188
hypnotism, 51

I

ice cream recipe, 522, 523
ileocaecal valve, 97, 101, 128
immune failure, 286
impetigo, 46
incontinence, 75, 278, 294, 325
indigestion, 6, 103, 145, 298, 446, 557
indium, 123, 157, 162
inductance, 5, 7, 486
inductor, 11, 486, 489
infant formula, 429
infertility, 2, 117, 144, 186, 237, 384
inhaler, 139
insect, 8, 506, 509, 529
insecticide, 312
insomnia, 4, 82, 134, 244, 245, 278, 326, 340, 417, 557
insulin, 173, 174, 175, 176, 177, 178, 182, 276, 302, 308, 332, 490
interferon, 468

interleukins, 468
intestinal problem, 128, 267
iodine, 95, 136, 163, 164, 188, 202, 218, 239, 246, 248, 298, 344, 348, 398, 399, 442, 443, 499, 559, 583
iridium, 82, 123, 188, 255, 271, 497
iron, 247, 248, 277, 279, 286, 450, 572
irritability, 113, 210, 254, 262, 298
irritable bowel syndrome (IBS), 102, 103, 104
islets of Langerhans, 182
isopropyl alcohol, 251, 285, 332, 335, 336, 348, 394, 398

J

jam, 2, 38, 52, 233, 265
jaw, 56, 74, 79, 87, 88, 156, 157, 158, 159, 170, 206, 215, 226, 254, 291, 315, 377
jaw bone, 74, 79, 87, 88, 206, 215
joint, 65, 66, 70, 82, 84, 139, 155, 157, 217, 417
joint pain, 84, 417

K

kerosene, 137, 208, 224, 227, 235, 407, 429, 572
kidney stone, 48, 56, 75, 79, 94, 95, 96, 108, 127, 166, 204, 212, 291, 552
knee, 44, 68, 74, 419, 489
knuckle, 195, 459, 463, 464, 486
Kojic acid, 102, 174, 175, 226, 388, 423

L

lactic acid, 88, 184, 187, 276
Lactobacillus, 100, 564, 576
lactose, 66, 104, 202
lamp oil, 474
lanolin, 223, 225
lanthanide, 414
lard, 303, 310, 354, 374, 394, 424, 517, 529, 530
laser, 111
lateral sclerosis, 204
lead, 4, 22, 26, 37, 46, 47, 60, 69, 74, 75, 79, 81, 96, 101, 105, 116, 136, 139, 141, 146, 157, 170, 181, 197, 220, 231, 232, 234, 262, 264, 265, 267, 268, 275, 280, 290, 314, 349, 356, 416, 417, 418, 426, 436, 440, 466, 472, 475, 504
Leishmania, 107, 193, 568, 578
lemonade, 557
lesion, 9, 181
leukemia, 211
lip, 534, 535
lip soother, 534, 535
lipoic acid, also see thioctic acid, 442, 584
lipstick, 75, 123, 207, 327, 336, 476, 539, 572
lithium, 255, 259, 260, 271
liver disorder, 374
liver fluke, 33, 43, 74, 78, 96, 103, 104, 107, 109, 111, 116, 117, 122, 128, 131, 175, 182, 185, 186, 187, 188, 207, 208, 218, 220, 222, 223, 224, 225, 249, 259, 260, 268, 271, 494, 498
Loa loa, 141, 147, 148, 149, 176, 291, 318, 319, 568, 576
loss of balance, 169, 270, 419
lotion, 247, 485, 535, 536
Lou Gehrig's disease, 44
Lugol’s, 163, 202, 203, 246, 248, 271, 298, 315, 346, 398, 442, 443, 499, 500, 501, 559, 583
lung, 40, 135, 138, 336, 372, 407, 477, 478, 576, 578
lung fluke, 138
lupus, 250, 495, 589
lutetium, 141
lye, 529, 530
Lyme disease, 170, 563
lymph, 64, 89, 90, 109, 144, 156, 157, 267
lymphocytosis, 417
lysergic acid, 263
lysine, 142, 146, 180, 181, 534, 535, 537, 538

M

magnet, 155
magnetic resonance imagery (MRI), 5
maintenance parasite program, 83, 105, 106, 116, 151, 176, 271, 340, 341, 555, 558

malignant, 39, 331, 337, 447
malvin, 233, 235, 236, 265
mammary gland, 143
mammogram, 144, 145, 146
manganese, 89, 176, 277, 572
Maple Milk Shake, 514
maple syrup, 191, 243, 365, 381, 390, 441, 513, 514, 524
margarine, 378, 517
marshmallow root, 549, 551, 552, 583
massage, 51, 70, 85, 92, 435, 439
massage oil, 435, 439
mean cell volume (MCV), 212, 285
measles, 186, 353, 492
medication, 166, 203, 236
melanoma, 337
memory loss, 273, 278
Meniere's syndrome, 169
menopause, 113, 114, 143, 260
menstruation, 113
mental problem, 2, 283
mercuric chloride, 70, 72
mercury amalgam, 409, 416, 417, 485
mesothelioma, 337
metabolism, 111, 187, 238, 239, 240, 243, 244
metacercaria, 495
metal tooth filling, 73, 123, 142, 157, 188, 189, 442, 472, 533
metastasis, 337
methanol (wood alcohol), 173, 174, 429
methyl butyl ketone, 35, 96, 110, 111, 116, 122, 269, 574
methyl ethyl ketone, 35, 110, 171, 260, 574
methyl methacrylate, 414
methylene chloride, 37, 162, 208, 221, 269, 574
microscope slide, 8, 149, 210, 477, 501, 561
migraine, 162, 163, 164, 165, 168, 259
milk digestant, 80, 421, 425
millet, 388, 420
miracidia, 34, 188, 495, 500
mite, 358, 359, 360, 408, 570, 580
mitral valve, 147, 159
moisturizer, 525
molasses, 286
mononucleosis, 492
monosodium glutamate (MSG), 232, 233, 253, 255, 265, 266, 304, 434
moth ball, 104, 141, 149, 162, 271, 448
mother's milk, 133, 143, 375, 490
mouthwash, 156, 194, 218, 229, 335, 413, 435, 438, 476, 533
multiple sclerosis, 46, 204, 219, 250, 254
mumps, 176, 178
muscle, 66, 78, 93, 94, 139, 147, 158, 161, 187, 197, 218, 220, 477
muscle pain, 66
muscular dystrophy, 219, 220
myasthenia gravis, 55, 221
mycoplasma, 138, 141, 176
mycotoxin, 38, 102, 261, 351, 384, 385, 388, 423
myofibrositis, 77

N

Naegleria, 138, 271, 568, 577
nail polish, 136, 222, 254, 449, 573
nausea, 66, 121, 153, 264, 338, 557
neck, 89, 144, 156, 157, 160, 166, 206, 280, 402, 419
Neisseria, 114, 116, 119, 563, 565, 576, 577
nephritis, 417
nerve, 36, 59, 86, 154, 158, 170, 180, 181, 205, 206, 262, 329, 371, 478
nervousness, 113, 187
neurologist, 73
niacin, 68, 70, 73, 168, 279, 280, 281, 320, 380
niacinamide, 228, 233, 263, 338, 364, 380
nickel, 37, 85, 106, 115, 122, 124, 125, 126, 138, 139, 143, 145, 177, 180, 188, 190, 207, 222, 374, 418, 420, 424, 431, 436, 475
nicotine, 264, 539
night sweats, 204, 353
niobium, 146
nitrate, 573
Nocardia, 112, 159, 187, 206, 242, 260, 565, 576, 577
nose, 22, 24, 40, 105, 107, 133, 140, 193, 199, 203, 347, 359, 361, 369, 379, 536
numbness, 154, 157, 207, 209, 254, 255, 256, 417
nursing, 134, 242, 292, 342, 375
nut, 23, 25, 262, 310, 374, 386, 389, 461, 519
nut butter, 310, 374, 386, 389

O

obesity, 237, 240, 242, 548
optic nerve, 478
optic neuritis, 417
ornithine, 82, 243, 244, 259, 272, 325, 326, 340, 342, 557
osteoarthritis, 78, 83
osteomyelitis, 410
ovarian cyst, 91, 110, 255, 384
ovary, 90, 91, 92, 119
overweight, 130, 237, 238, 239, 242, 243
oxalate, 68, 69, 93, 94, 95, 96, 108, 116, 122, 128, 139, 142, 171, 188, 551
oxalic acid, 61, 93, 94, 243, 276, 551

P

P24 antigen, 583
PABA, 536, 583
palladium, 82, 138, 141, 149, 177, 187, 208
pancreatic fluke, 33, 35, 43, 102, 103, 105, 107, 109, 140, 142, 173, 174, 175, 176, 177, 178, 179, 182, 183, 186, 188, 220, 221, 249, 250, 251, 259, 302, 371, 375, 388, 494, 498
pantothenic acid, 60, 253
para amino benzoic acid (PABA), 536, 583
paradichlorobenzene, 149, 162
Paragonimus, 107, 138, 568, 579
paralysis, 417
paraplegia, 417
paresthesia, 417
Parkinson's disease, 159, 315
parsley, 304, 306, 343, 344, 345, 513, 549, 550, 551
pasta, 66, 96, 121, 204, 225, 228, 262, 306, 307, 308, 310, 356, 363, 374, 375, 381, 385, 386, 387, 432
patulin, 277, 522
PCB, 147, 162, 188, 314, 426, 475, 527, 573
peanut butter, 45, 189, 364, 376, 382, 389, 432, 519, 523, 524
penis, 250, 491, 493
pentane, 127, 162, 171, 188, 208
perfume, 190, 253
peripheral neuropathy, 209
personal lubricant, 439
pesticide, 40, 68, 96, 132, 162, 171, 185, 236, 253, 255, 268, 282, 304, 313, 346, 427, 472, 473, 542, 572, 590
pet food, 354, 574
pet parasite program, 77, 117, 122, 167, 395, 494
Peyer's patches, 267
phenylisothiocyanate (PIT), 84
phosphate, 63, 64, 65, 66, 67, 68, 73, 80, 81, 83, 86, 88, 94, 95, 96, 106, 107, 139, 142, 177, 185, 212, 213, 215, 216, 217, 255, 320, 551
phosphate crystal, 65, 67, 86, 94, 95, 96, 107, 177, 185, 212, 217, 255, 320, 551
pinworm, 220, 404
piperine, 84
placenta, 76, 258, 263
platelet, 19, 145, 286
platinum, 68, 116, 122, 123, 134, 141, 149, 185, 193, 207, 212, 419
PMS, 113, 115, 255, 384
Pneumocystis, 138, 372, 568, 578
pneumonia, 138, 142, 147, 169, 407, 564, 566
polio, 371
polonium, 372
polyethylene, 62, 517, 532
polyneuritis, 417
popcorn, 144, 181, 248, 313, 351, 355, 356, 373, 385, 476, 492
porcelain, 415
pork, 206, 423, 424, 477, 484
potassium, 58, 59, 82, 95, 139, 163, 211, 217, 218, 277, 291, 293, 294, 304, 311, 348, 373, 499, 515, 516, 559
potassium bicarbonate, 58, 515
potassium chloride, 211, 311, 373, 516
potpourri, 165, 168, 448
praseodymium, 122, 221, 244, 439
pregnancy, 76, 106, 111, 115, 119, 121, 122, 264
probe, 28, 459, 462, 463, 464, 465, 467, 468, 471, 487, 489, 504, 506
progesterone, 91, 113, 114, 115, 119
propane, 207
prostate, 69, 96, 112, 123, 124, 125, 126, 127, 128, 139, 190, 331, 332, 336, 337, 491, 500
prostate cancer, 332, 337, 500

Prosthogonimus, 139, 255, 568, 578
protein, 193, 228, 232, 553
Proteus, 94, 102, 110, 114, 565, 576, 578
Pseudomonas, 74, 565, 576
psoriasis, 193
purpura, 281
PVC, 70, 78, 211, 231, 236, 241, 262, 282, 450, 475, 573
pyloric, 297

Q

quassia, 546

R

rabbit fluke, 105, 203
radiation, 7, 437, 589
radio frequency, 6, 7, 13, 486
radon, 96, 108, 127, 142, 145, 146, 150, 181, 185, 186, 187, 405, 406, 475
rash, 36, 189, 190, 191, 220, 246
rat poison, 417, 418
red blood cell, 19, 77, 285, 286
Red Milk, 515
redia, 9, 34, 122, 495, 500
reflexology, 61
reproductive organs, 119
respiratory, 140, 142, 160, 170, 358, 360, 361, 363, 449, 517
retina, 160, 162, 213
retinitis pigmentosa, 3, 162
rhabdomyosarcoma, 337
rhenium, 139, 187
riboflavin, 380, 584
rice, 63, 66, 121, 144, 303, 306, 307, 308, 311, 349, 351, 356, 363, 366, 383, 385, 420, 432, 492, 534
ringworm, 247
room freshener, 150, 284
root canal, 79, 119, 138, 145, 158, 159, 163, 166, 187, 208, 216, 266, 318, 410, 412
roundworm, 14, 32, 76, 83, 86, 107, 139, 257, 265, 267, 274, 343, 372, 566
rubella, 189
ruthenium, 212, 244, 439

S

salicylate, 170
salmon, 307, 366, 376, 422, 524
salve, 162
samarium, 82, 157, 208
Sarcocystis, 107, 568, 579
sassafras, 529, 546
scalp, 126, 171, 190, 193, 526, 528
scandium, 206
schizophrenia, 211, 261, 262, 450
sciatic nerve, 86
scopolamine, 393, 423
seafood, 205, 307, 366, 424, 432, 434, 522
sebaceous cyst, 171, 188
seizure, 230, 231, 232, 233, 234, 235, 236, 265, 517
semen, 342
sex, 67, 121, 179, 205, 319, 342
shampoo, 105, 144, 284, 333, 335, 336, 437, 476, 528, 542
shaving supplies, 189, 284, 335
sheep liver fluke, 33, 43, 78, 103, 104, 107, 109, 111, 117, 122, 131, 175, 182, 185, 187, 188, 208, 223, 224, 225, 249, 260, 271, 494, 498
shingles, 179, 192, 371, 487, 492, 564
shoulder, 49, 150, 151, 152, 153, 255, 553
silicon, 475
silver, 37, 100, 103, 110, 116, 119, 146, 177, 185, 192, 193, 260, 310, 420, 426, 475, 476
sinus, 104, 116, 117, 138, 139, 142, 163, 167, 168, 170, 187, 245, 268, 369, 419, 492
skin cancer, 194
skullcap, 244, 552
sleep problem, 96, 107, 117, 241, 255
Smoker's Leg, 70
smoking, 67, 68, 70, 73, 177, 255, 327, 336, 372
soap, homemade, 437, 438, 444, 448, 449, 530
soda pop, recipe, 515
sodium, 58, 59, 106, 139, 146, 181, 212, 217, 232, 282, 311, 373, 440, 515, 529, 534, 535, 537, 582, 584
sodium alginate, 106, 181, 534, 535, 537
sodium bicarbonate, 58, 582
sodium carbonate, 584

sodium chloride, 373, 440, 584
sonogram, 5
sore throat, 20, 159, 160, 206, 217, 400, 492
sorghum, 226, 281, 388, 391, 475, 481, 519
soup, 106, 180, 209, 243, 287, 294, 306, 307, 310, 324, 349, 376, 405, 431, 432, 434, 521, 523, 524
sour cream, 164, 293, 305, 524
soy, 387, 432, 434
spasm, 70, 93, 150, 153, 163, 300
spastic colon, 78
speech problem, 270
sperm, 119, 123
spinal cord, 204, 205, 206, 361, 478
spine, 49, 80, 92, 93, 217
spleen, 148, 174, 195, 197, 198, 358, 419, 492
stainless steel, 431, 543
Staphylococcus, 62, 74, 76, 89, 99, 116, 147, 159, 163, 178, 187, 206, 216, 242, 291, 318, 319, 346, 353, 565, 566, 577
starch, 121, 300, 373, 374, 375, 378, 531, 532, 534, 537, 539, 540, 573
Sterigmatocystin, 385, 562, 576
sternum, 148, 149
stomach problem, 104, 105, 268
strep throat, 100, 104, 492
Streptococcus, 74, 76, 99, 117, 147, 168, 169, 170, 187, 242, 566, 576, 577
stroke, 278, 280, 281, 282
Strongyloides, 78, 79, 82, 86, 163, 164, 166, 167, 170, 186, 194, 256, 257, 258, 259, 260, 264, 343, 371, 568, 578
strontium, 124, 127, 181, 192, 244, 296, 435, 436, 439, 450
styrene, 103, 122, 148, 162, 430
styrofoam, 122, 148, 260, 292, 314, 430, 433, 475, 527, 575
sulfur, 95, 267, 376
sweat, 126, 247, 528, 531, 532
sweetening, 303, 308, 524
Syncrometer, 43, 45, 150, 182, 189, 251, 312, 346, 355, 429, 444, 446, 457, 458, 459, 460, 462, 466, 467, 470, 471, 482, 503, 504, 506, 571
syphilis, 91, 566

T

Taenia pisiformis, 75, 148, 196, 570, 579
Taenia solium, 147, 196, 570, 579
tantalum, 140, 141, 475
tapeworm, 20, 32, 50, 75, 111, 147, 161, 179, 195, 196, 197, 198, 199, 200, 201, 208, 220, 318, 357, 358, 359, 360, 363, 364, 370, 371, 408, 487, 494, 570
tapeworm stage, 20, 32, 75, 111, 147, 161, 179, 195, 196, 199, 200, 201, 208, 318, 357, 358, 359, 360, 363, 364, 370, 371, 487, 494
TC Ethylene, 96, 141, 148, 188, 193, 429
tea kettle, 310, 431
tellurium, 68, 78, 134, 142, 188, 208
temporary ischemic attacks (TIA), 207, 281
tendon, 154
teratogen, 418
terbium, 186
testicle, 20, 54, 119, 128, 178
thallium, 37, 69, 70, 71, 72, 73, 74, 79, 95, 96, 100, 111, 115, 123, 134, 189, 190, 206, 209, 220, 242, 254, 255, 264, 269, 314, 361, 417, 418, 438, 536, 573
thioctic acid, 73, 74, 82, 120, 123, 178, 207, 208, 229, 236, 237, 254, 262, 265, 269, 270, 442, 584
threadworm, 163, 371
throat, 20, 100, 104, 105, 141, 149, 150, 160, 199, 206, 227, 241, 242, 243, 268, 280, 295, 354, 361, 370, 400, 477, 491, 492
thrush, 246, 248, 492
thulium, 128, 141, 146, 177, 193, 221, 439, 440
thymus, 35, 36, 107, 131, 150, 207, 208, 221, 250, 251, 351, 352, 353, 446, 477, 491, 493
thyroid, 77, 107, 117, 170, 182, 183, 185, 187, 188, 193, 213, 217, 237, 239, 254, 266, 377, 443
thyroxin, 187
tin, 9, 139, 162, 187, 242, 296, 435, 436, 472, 473, 476
Tinea, 247
tinnitus, 6, 169, 170, 171, 419

titanium, 89, 103, 123, 143, 180, 193, 231, 254, 435, 436, 476
TMJ, 157, 158, 259, 419
toast, 178, 181, 352, 356, 430, 431, 434, 524, 572
toaster, 573
tobacco, 84, 279, 326, 372, 436, 492, 566
Tobacco Mosaic virus, 436
toe, 56, 59, 67, 68, 155, 248, 253
tofu, 432, 434
toluene, 35, 37, 103, 141, 148, 205, 209, 219, 220, 232, 251, 252, 269, 270, 271, 336, 427, 476
tomato juice, 82, 308, 521
tongue, 6, 248, 376, 485
tonsil, 268
tooth metal, 112, 123, 133, 138, 142, 149, 162, 186, 187, 206, 221, 231, 240
toothbrush, 95, 218, 248, 295, 438, 532
toothpaste, 82, 124, 127, 138, 156, 181, 187, 196, 206, 218, 242, 296, 420, 435, 438, 476, 532, 572, 573, 574
Toxoplasma, 160, 161, 162, 568, 577
trachea, 187, 478
tremor, 49, 78, 206, 207, 209, 260, 270, 278, 313, 314, 315, 316
Treponema, 114, 119, 566, 576
Trichinella, 31, 68, 76, 79, 80, 82, 86, 153, 161, 186, 192, 193, 194, 207, 257, 260, 343, 568, 578
trichloroethane, 37
Trichomonas, 102, 110, 568, 577
Trichuris, 74, 107, 142, 166, 568, 578
tryptophane, 244, 245, 439
tuberculosis, 142, 407, 565, 579
tumor, 40, 337, 415, 547
tungsten, 271, 430, 431, 449
turkey, 423, 424, 477, 485
turmeric, 547
tyramine, 164, 293, 305

U

udder wash, 98, 109, 137
ulcer, 132
underweight, 240, 268
universal allergy syndrome, 495
uranium, 142, 146
urate, 95, 122, 139, 166, 167, 255
urea, 75, 125, 243, 320
urethra, 124
uric acid, 55, 61, 68, 69, 81, 94, 171, 213, 320
urinalysis, 68, 77, 166, 185, 193
urinary tract, 77, 102, 106, 110, 114, 122, 125, 127, 163, 166, 167, 190, 255, 291, 430, 555
uterine wall, 112, 249
uterus, 35, 103, 106, 108, 109, 110, 111, 112, 113, 115, 117, 119, 122, 123, 124, 145

V

vagina, 246, 250, 491, 493
valerian, 150, 151, 152, 153, 158, 244, 272, 300, 325, 326
vanadium, 96, 105, 140, 149, 165, 170, 171, 175, 186, 187, 231, 255, 260, 268, 271, 313
vanilla, 514, 519, 522
Vanilla Milk, 514
vasectomy, 128
vegetable juice, 66, 82, 94, 237, 308, 309, 322, 422, 424, 513, 523, 524, 525
vegetarian, 197, 434
vinegar, 89, 276, 277, 288, 291, 296, 297, 303, 305, 306, 348, 349, 366, 377, 381, 383, 403, 405, 423, 441, 444, 448, 493, 514, 521, 527, 528, 540, 541, 542, 572
vitamin A, 95, 142, 154
vitamin B, 80, 88, 142, 146, 178, 217, 235, 285, 353, 551
vitamin B1, 235, 285
vitamin B12, 235, 285
vitamin B2, 88, 353
vitamin B6, 80, 142, 146, 217, 551
vitamin C, 38, 58, 73, 87, 89, 107, 110, 120, 121, 128, 141, 144, 146, 168, 177, 178, 181, 184, 193, 212, 218, 222, 228, 230, 242, 247, 263, 270, 277, 280, 281, 286, 291, 301, 302, 303, 305, 347, 348, 349, 356, 364, 365, 366, 370, 382, 383, 385, 386, 388, 389, 390, 391, 393, 421, 423, 427, 432, 433, 440, 441, 470, 481, 494, 513, 515, 516, 518, 519, 520, 521, 523, 525, 531, 534, 535, 537, 538, 543, 545, 548, 556, 573, 584

vitamin D, 64, 65, 66, 74, 81, 87, 88, 158, 217, 218, 254
Vitamin D Drops, 560
vitamin E, 120, 145, 146, 534, 539
vomiting, 105, 134, 166, 202, 394, 417

W

wallpaper, 141, 448, 572
wart, 69, 174, 176, 195, 196, 201, 478
washing soda, 57, 58, 180, 188, 269, 314, 346, 399, 431, 438, 444, 448, 515, 516, 525, 527, 532, 533, 540
water heater, 39, 319, 431, 447, 449, 573
water softener, 117, 127, 274, 437, 450, 473, 517, 572, 573
weakness, 92, 127, 139, 207, 220, 278, 293, 317, 322, 417
whip worm, 105, 107, 142
white distilled vinegar, 277, 349, 423, 444, 448, 521, 540, 541, 542
white iodine, 95, 163, 218, 248
wine, 164, 280, 327, 381, 386, 431
wood alcohol (see also methanol), 37, 102, 103, 162, 173, 174, 175, 176, 177, 178, 179, 183, 185, 186, 206, 208, 240, 251, 252, 259, 269, 302, 309, 314, 371, 375, 377, 388, 391, 414, 436, 475, 519, 559, 574
worm, 10, 105, 107, 142, 147, 163, 197, 247, 257, 318
wormwood, 339, 342, 344, 345, 582
wrist, 74, 75, 80, 154, 155, 255, 328

X

xenobiotic, 419
xylene, 35, 37, 111, 122, 141, 148, 205, 208, 209, 219, 220, 232, 251, 252, 270, 427, 429

Y

Yankee Drink, 514
yeast, 4, 115, 116, 122, 189, 246, 247, 248, 298, 377, 399, 492, 563
ytterbium, 415, 436
yttrium, 123, 127, 186

Z

zearalenone, 144, 225, 355, 492
zinc, 70, 106, 120, 123, 127, 154, 160, 181, 247, 277, 296, 532, 534, 572
zinc oxide, 181, 247, 532, 534
zirconium, 75, 89, 112, 128, 138, 147, 170, 180, 189, 193, 222, 231, 415, 435

A New Diagnostic Device…

With an audio oscillator circuit and a trained ear, you can detect parasites and pollutants in a way that no one has done before–electronically.

Leads To New Discoveries…

In every case of the “mysterious” disease diabetes, you find the not-so-mysterious parasite *Eurytrema*, and the fairly common pollutant *wood alcohol*. Every case! And never in healthy people! Similarly in cancer, HIV, Alzheimer’s, endometriosis, to name a few, you can find specific parasites and/or pollutants at work.

And New Cures…

You don’t need dangerous, expensive prescription drugs to get rid of the causes of your illness. Once you know what you are fighting you can pick herbal, electronic, or avoidance methods. Except for the cost of dental work, you can get well for under $100!

And New Hope…

Follow the advice in this book preventively, and never worry about your health again! Get your whole family well and keep them well. Start today!

It Took An Independent Research Scientist

to find the answers. Hulda Regehr Clark began her studies in biology at the University of Saskatchewan, Canada, where she was awarded the Bachelor of Arts, Magna Cum Laude, and the Master of Arts, with High Honors. After two years of study at McGill University, she attended the University of Minnesota, studying biophysics and cell physiology. She received her Doctorate degree in physiology in 1958. In 1979 she left government funded research and began private consulting on a full time basis. Six years later she discovered an electronic technique for scanning the human body. Today she puts her methods, results and conclusions before you. Read and recover.

New Century Press
Chula Vista, CA

1 (800) 51902465
www.newcenturypress.com

ISBN 1-890035-01-7
$21.95

About this electronic book

While creating this electronic version of The Cure for all Diseases, I encountered some inconsistencies and (small) errors in it. Below follows an overview.

Niels Willems, december 2002.
niels.w@hccnet.nl
www.rupa.demon.nl

Page	**Line**	**Hard copy**	**Electronic version**
18	Fig.2	(Di-phyllobothruim e.)	(Di-phyllobothrium e.)
25	last-2	both with a alligator clip	both with an alligator clip
255	last-3	liver cleanses(page 549) make	liver cleanses (page 552) make
360	last-7	health improv-ed	health improved
451	5	furniture"treated"	furniture "treated"
452	last	refigerators	refrigerators
467	14	SALT # 1	SALT #1
519	last	honey	honey.
562	last-7	Argyria	Arcyria
563	16	Bac. antracis spores | 391.45 | 386.95	Bac. antracis spores | 386.95 | 391.45
564	6	Corynebacterium diptheriae	Corynebacterium diphtheriae
565	18	Norcardia asteroides (2nd range)	Nocardia asteroides (2nd range)
570	last-4	folicle mite	follicle mite
575	2	Paradiclorobenzene	Paradichlorobenzene
576	(17)	missing: Aflatoxin (2nd) 188 188	(not corrected)
576	last-19	Corynebacterium diptheri	Corynebacterium diphthe
580	4	folliculorum foli	folliculorum foll

Healing
the symptoms known as
Autism

Second
Edition

PDF
E-Book

Kerri Rivera

with Kimberly McDaniel & Daniel Bender

Jim Humble • Dr. Andreas Kalcker
Dr. Marco Ruggiero • Robert L. Sands

Disclaimer

Healing the Symptoms Known as Autism is not intended as medical advice. This book is for informational and educational purposes only. Please consult a medical professional when the need for one is indicated. For obvious reasons neither the author, co-authors, contributing authors, the publisher, nor their associates can take medical or legal responsibility for having the contents herein considered as a prescription for everyone or anyone. You are ultimately responsible for the uses made of this book.

All content, including text, graphics, images, and information, contained in this book or our website, is for general information purposes only. We take no responsibility for the accuracy of information contained herein, and such information is subject to change without notice.

PDF
E-Book

Healing
the symptoms known as
Autism

Second Edition

by Kerri Rivera
with Kimberly McDaniel & Daniel Bender

Healing the Symptoms Known as Autism
Second Edition (PDF E-Book Version)

by Kerri Rivera
with Kimberly McDaniel & Daniel Bender

Copyright © 2014 by Kerri Rivera

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage retrieval system without the written permission of the author. The only exception being by a reviewer who wishes to quote brief passages in connection with a review written for inclusion in a magazine, newspaper, blog, or broadcast.

ISBN: 978-0-9892890-2-3
Library of Congress Control Number: 2013945511 (LOC# of Print Edition)

1st Edition released May 2013
2nd Edition released January 2014
Printed in the United States of America

Mailing Address:
AutismO2
PO Box 10334
Chicago, IL 60611

Feedback Email: kim@cdautism.org

For further information:
www.cdautism.org
www.HealingTheSymptomsKnownAsAutism.com

The Kalcker Parasite Protocol as integrated into Chapter 8 has portions
Copyright ©2013 by Dr. Andreas L. Kalcker & Miriam Carrasco Maceda.
Original English Translation by Mercy Acevedo
Edited & revised by Michael Harrah, Kimberly McDaniel & Daniel Bender

Original Cover Design: Eduardo A. Davad
http://edudes.weebly.com/

Back cover photo of Kerri Rivera by PaulVanVleckPhotography.com

I would like to dedicate this book to the families
of children on the spectrum the world over.

May all of our children
find the healing that they need.

~ Kerri

This book is dedicated to my Dominick.
I love you forever; thank you for being our angel.

~ Kim

Healing
the symptoms known as
Autism
Kerri Rivera
with Kimberly McDaniel

One of our fans currently working towards recovery thanks to the protocol.

Table of Contents

Foreword by Lorna B. Ortiz, PhD.
President of Curando el Autismo, CEA vii

Preface by Kimberly McDaniel ix

Terminology & Units of Measure xiii

Acknowledgements xv

Introduction: Autism is Avoidable, Treatable and Curable xxi

Chapter 1: Kerri's Story 1

Chapter 2: Yes We Can!!! (Testimonials) 19

Chapter 3: Step 1 – The Diet 41

Chapter 4: An Introduction to Chlorine Dioxide 77

Chapter 5: Step 2 – Chlorine Dioxide (CD) 81

Chapter 6: CDS – Another Way of Delivering Chlorine Dioxide 145

Chapter 7: CDH – The Research Continues 155

Chapter 8: Step 3 – The Kalcker Parasite Protocol 165

Chapter 9: Step 4 – Other Supplements 243

Chapter 10: Step 5 – Chelation 255

Chapter 11: Step 6 – Hyperbarics 263

Chapter 12: Step 7 – GcMAF & Autism 293

Chapter 13: Beyond Recovery: The Maintenance Plan 319

Chapter 14: Miscellaneous Information You Should Know 323

Chapter 15: Final Thoughts 347

Chapter 16: Healing Beyond Autism 349

Appendix 1: More Miracles & Testimonials 357

Appendix 2: CD Autism Worldwide 441

Appendix 3: Accidental Cure by Optimists..443

Appendix 4: Autism Treatment Evaluation Checklist (ATEC) 447

Appendix 5: Molecular Mimicry...449

Appendix 6: Measuring the Strength of Your CD, CDS & CDH..................467

Appendix 7: Diluting HCl Concentration...471

Appendix 8: Other Uses of Chlorine Dioxide 473

Appendix 9: Blank Kalcker Parasite Protocol Chart.....................................475

Appendix 10: Lunar Phase Calendar for Kalcker Parasite Protocol...........477

Appendix 11: Health Benefits of Diatomaceous Earth.................................485

Appendix 12: Summary of Protocols...491

Appendix 13: Body Burden – The Pollution in Newborns 499

Appendix 14: Avoiding Autism...505

Appendix 15: Recipes - Cooking with Ana..513

Appendix 16: Websites We Like..517

Appendix 17: Direct Help from Kerri Rivera 521

References 523

Index 529

About the Author 563

Foreword

Autism, not the one described by Kanner in 1943, but the one that we see today diagnosed in 1 of 50 kids is a combination of immune disorders that need to be treated biomedically. There is a lot that we still need to understand about why and how these immune dysfunctions affect our kids' development, causing near complete impairment of their social interaction and communication.

There is a long, hard road ahead to fully comprehend the integrated system that comprises "autism," but our kids cannot wait. It takes more than just medical, professional, or even scientific interests working hard day and night to find a suitable and effective solution to help our kids. It takes drive, passion, and guts to do the right thing; to hear and read the overwhelming stories of parents from all over the world and not turn our backs, but help. It takes an eternal and extreme will to help, even when your own child is on the spectrum.

Kerri's protocols have been indispensible to the full recovery of many of the kids in our Curando el Autismo (CEA) Foundation. These protocols represent a readily available and effective solution to alleviate most of the pathogenic insults to the immune system. Years from now, when "autism" diagnoses do not exist, when we find ourselves fully knowledgeable about the now mysterious immune-brain-behavior connection, I will remember Kerri not only as a friend but as one of the first courageous leaders that dared to change the path for our ill kids. She goes against all odds, sharing knowledge, experience and simply making it happen.

~ Lorna B. Ortiz, PhD.
President of Curando el Autismo, CEA

“Based on parent reports, the prevalence of diagnosed ASD in 2011–2012 was estimated to be 2.00% for children aged 6–17. This prevalence estimate (1 in 50) is significantly higher than the estimate (1.16%, or 1 in 86) for children in that age group in 2007.”

US Department of Health & Human Services
National Health Statistics Reports
Number 65, Pg. 2 - March 20, 2013
www.cdc.gov/nchs/data/nhsr/nhsr065.pdf

Preface

by Kimberly McDaniel

> *Darkness cannot drive out darkness; only light can do that.*
> *Hate cannot drive out hate; only love can do that.*
> ~ ***Martin Luther King, Jr.***

Welcome to the Second Edition of *Healing the Symptoms Known as Autism!* We are absolutely thrilled to share with you the latest protocol updates and everything else that has happened since May, 2013.

You may be thinking: Why a second edition so soon? Our first edition gave us the structure to explain the protocol, and a foundation to build on. This book has already helped many families around the world, and in fact some parents read it and recovered their children without even contacting us until afterwards! Keep in mind that when it was released in May, 2013 it was absolutely up to the minute, but as we mentioned, this protocol will continue to evolve until we have something that is consistently recovering people on the spectrum of all ages. As of January 2014, we again are sharing the latest updates, as well as a whole lot more information that we hope will be as interesting and beneficial to you as it was to us.

Here are some of the exciting new additions:

- Olive Kaiser of www.GlutenSyndrome.net has written a section on gluten and its role in molecular mimicry and autoimmunity. Since many of you are not new to the autism community, a gluten free diet for your children is nothing new. However, you may be interested to find out just how damaging gluten can be for folks who are off the spectrum as well.

- Scott McRae has contributed a chapter on CDH (Chlorine Dioxide Holding [Solution]). A new method of preparing chlorine dioxide that many families are now using with success. This gives us an even wider variety of preparations available to accommodate the needs of our families.

- The Kalcker Parasite Protocol chapter now has some beautiful charts that spell out the timing of all the components for the 18 days a month that a child will be on the parasite protocol. Thanks to Dan Bender, a lot of the confusion surrounding how to make it all fit will be cleared up. You will also find years of lunar calendars to make it easy to see when the protocol is active. You won't have to check Google again to see when the full or new moon is coming.

- The "Worm Whisperer" Robin Goffe shares with us part of her journey towards healing for her 19-year-old son. Her advice is for the Extreme Cases—Self Injurious behavior, aggression, violence, etc. If you are the parent of an older child on the spectrum, or a child who displays these behaviors, or you know someone who is living with a child like this, you owe it to yourself to read Robin's suggestions. They are full of hope and wisdom.

- The one and only Marco Ruggiero, lead researcher on GcMAF, (Gc Macrophage Activating Factor—an immune system supplement) has written an entire chapter on GcMAF and its applications for autism. A must read.

- Last but not least, a whole new crop of testimonials that will make you cry. If after reading this book you still have doubts about giving this protocol a chance, I highly recommend you revisit these testimonials. If I had to pick one part of the book that was absolute favorite, this would be it. It may be because I collect a lot of them, and usually have a little correspondence with the families to get their permission, and genuinely get to feel their excitement, their sense of accomplishment when they see their child start to heal, and not to mention their endless gratitude for being able heal their own children.

  I pray that all of these testimonials find their way to the people who need them and as a result children stop suffering because their parents see them reflected in those words and understand that if other children can heal, their children can as well.

  To me, there is nothing more real than hearing from someone that has walked a mile in your shoes. There is nothing more inspiring that to hear someone say, "I know it's possible, because I did it, I lived it, and I am here to tell you about it."

The families that courageously walked this path and then took the time to share their stories are pioneers and heroes to their own families and to our entire community and beyond. They are blazing a trail for others to follow and countless lives will benefit from their diligence, fortitude and dedication. The stories of healing come from all over the world... from children and adults of all ages. We hope that they move you as much as they moved us. We are eternally grateful for the service these families have provided to humanity and thankful that they were generous enough to take the time out of their lives to pay it forward and share.

- The book now as an extensive index making it much easier to use as a reference.

As of this printing, families are healing their children with autism using this protocol in over 58 countries! Our Facebook groups have over 3,500 members in many of these countries. CD officially knows no borders. It is absolutely thrilling to us that people are coming together with the common goal of healing their children themselves, and to help others do the same! It fills our hearts with joy to be a part of this and to be able to see and feel the love that is shared every single day.

The development of this protocol has been a growing grass roots effort, in sharp contrast to how modern medicine usually works. The reason for this is clear: Modern medicine has not really helped to heal autism and may very well be part of its cause. While we don't have any double blind studies to rest on, we do have a slew of anecdotal evidence, which may not mean much to those in the arena of modern medicine or modern science, that doesn't make it any less real. Time and time again, our *results*—children dropping ATEC points—are being duplicated by families all over the world. For a family who is using this protocol, there is absolutely nothing more real than watching their child come back to them. Ask any autism parent what they would rather have... a long-term double-blind study published in a peer reviewed journal... or a healthy child. My money's on the latter.

As exciting as it is to be a part of this, and as wonderful as it is to hear about gains and read uplifting testimonials, we know there are still families working to heal children that are very sick, and our groups share the highs as well as the lows. Kerri will tell you that some months are better than others and the gains ebb and flow. I urge you to read and reread the testimonials. These are real stories of real healing and if you haven't already I invite you to believe that your child can be one of the success stories in those testimonial pages.

All of our families are proving what many were told was impossible: Children with the symptoms known as autism can heal! This grass roots movement is creating a paradigm shift in the way the world views healing from autism. The second edition of this book will be translated into at least 13 languages including Spanish, Portuguese, French, Flemish, German, Czech, Norwegian, Arabic, Polish, Italian, Hungarian, Bulgarian and Serbian. This is exciting. This is real. And, this is a living protocol. Every family that uses it every single day is helping to shape the future and heal the children who are affected today, and quite possibly prevent children from becoming affected tomorrow. For this we are eternally grateful.

Here's to you and the continued healing of humanity,

Terminology & Units of Measure

Throughout this book we talk about "CD," which is an abbreviation for *chlorine dioxide*, a well-established oxidizer. Chlorine dioxide is also often referred to as MMS, which is the common name given to it by Jim Humble, the man who discovered various applications of chlorine dioxide. There are many books, videos, blogs and articles using the name, "MMS," which is surrounded by a fair amount of controversy. We choose to not get tangled into that debate since our focus is on helping our children recover from autism. For us, our only concern is that (1) it is safe for our children, and (2) that it works. Based on extensive use of CD on thousands of children with autism, we can confidently say that both those statements are true. If this were not so, this book would not exist.

Units of Measure

Throughout this book we talk about various measured treatment components and containers using fluid volumes and weights. Since this book primarily addresses a US audience, we sometimes talk about measures in the US system of pounds and fluid ounces; while also using the internationally recognized metric system, which is frankly... easier to use.

The common abbreviations of measure you will see throughout this book include:

l = liter (volume)
mg = milligram (weight)
ml = milliliter (volume)
lb = pound (weight)
lbs = pounds (weight)
ppm = parts per million (concentration)
fl. oz. = fluid ounce (volume)
net. wt. oz. = net weight ounce (weight)

For easily measuring small amounts, syringes (without needles) are a great tool. Don't bother going to a chain store pharmacy... they won't sell them to you without a prescription. Instead, check your local medical supply store, animal supply store, private pharmacy or protocolsupplies.com. Cost: Surprisingly dirt cheap. The complete set of five below cost about $1.00 in Mexico—less than most candy bars—and not much more in the US. Note: Some brands of syringes have printed on scales that easily rub off, especially if your hands are a little oily. To prevent from losing the markings, cover the scale with some clear tape or clear nail polish.

When it comes to accurately measuring larger volumes, you can buy a set of 7 Polypropylene Graduated Cylinders. Of course this is not a requirement. You can use common kitchen measuring utensils, but these are more accurate and easier to read. If you decide to get them, avoid those with printed on lettering (it rubs off). Raised lettering is best, although sometimes a bit difficult to read. Typical cost on ebay™ or Amazon® for a set is around $25USD. The sizes range from 10ml to 1000ml. You can also buy individual cylinder sizes made of plastic or glass.

Acknowledgements

Gratitude makes sense of our past,
brings peace for today,
and creates a vision for tomorrow.
~ Melody Beattie

Thank you to my Mother who taught me by example the value of volunteerism and to help others who are less fortunate. Thank you for always telling me that I am the best and that I could do anything. I believed you :)

Thank you to my husband of nineteen years for supporting me on this journey. As challenging as it is, you have always stood by me. You are the love of my life, and father of the sweetest boys on the planet. Thank you for giving me the strength to be strong enough to do a mammoth job. It would not be possible without a pillar of your magnitude.

Thank you to Alex, the best big brother in the world who supports and enjoys what we do as a family and who sees the big picture. I couldn't be prouder than I am of you. I am so grateful you chose me as your Mom, and you chose to be on this journey with me, the road less traveled. I love you forever!

Thank you to Patrick for bringing to our family a higher good. You are an angel. You have brought light to all who meet and know you. Thank you for choosing me to be your mother. You and your brother have taught me more about myself and about life than I could have ever dreamed. I am grateful to you and I Love You more than words can say. Through me, you have given people their lives back. You have forever changed the face of autism.

Thank you to Linda for giving me my sister and one of my greatest friends.

Thank you Slimmie for your support, love and dedication, you are my BFF and we speak the same languages. Thank you for the support of my projects and always having me look prepared.

Lorna, thank you for supporting me, believing in me, trusting me, being faithful to your heart, and for being one of the most honest people I have ever known. You are my sister of the heart and I love you very much.

Susan Wiseman, thank you for offering your connection on that day when I most needed it. That one act of humility changed the future forever. I will never forget what you did for me.

Norrah Whitney, thank you for pointing me toward the direction of recovery, and explaining to me what I needed to do. You are without a doubt my first angel on this journey.

Ana Meckes, thank you for being another much needed angel and teaching me how to advocate for my son. I have not stopped ever since.

Anju, thank you for being my friend, for sharing with me, and for believing in me. You are a unique treasure.

Pina, thank you for helping and supporting me, we will be friends forever. I will always remember what you did for me, the clinic, and the kids.

Kenny, thank you for you excellent books full of sage advice and making biomed user friendly. Most of all, thank you for sharing fever therapy with us.

Carolina, Yamileth y El Jefe... gracias por apoyarme, defenderme y creer en mi. Son amigos en todo el sentido de la palabra. Estoy orgullosa de ustedes, moviendo montañas con el amor que solo una familia puede tener por un hijo….y ustedes lo han hecho por un pais. Les quiero mucho.

Bob Sands, thank you for seeing something special in me, thank you for taking me to meet Bernie and Mrs. Rimland. That day changed my life forever. Thank you for giving us the best hyperbaric chamber in the world. With that chamber we have seen so many miracles.

Jim, you have brought light to many and hope to all who have had the fortune to know the miracle that is CD (Chlorine Dioxide). I am grateful to you every day of my life. I have witnessed first-hand the miracle that is autism recovery. Thank you for staying the path that has never been easy. But it is as you say, “The right thing to do...” I love you.

Andreas, my dear friend, thank you for your dedication, your research and for your willingness to help. You are the science, the reason and the truth that is chlorine dioxide, the molecule that has the power to save humanity. Thank you to you and Miriam for your unfailing support over the years. Your contributions to the world of autism are changing the way the world views autism forever. And the lives of so many children are now recovered thanks to your contribution to The Protocol. Without you both, these recoveries would not have been so abundant.

Dan Bender, thank you for seeing the big picture and for giving so generously of yourself to help us help children with autism. Your selflessness has allowed this protocol to reach more families all over the world. Thank you for making the second edition of this book a reality, we never would have gotten it off the ground without you.

Thank you Michael Harrah for standing by me and working tirelessly to share the information that has had such a positive effect on so many families all over the world. Your wisdom and know-how have been invaluable to me in this book, our website and forums. You were a friend when I most needed one. I am grateful that you are in my life.

Dr. Bernard Rimland, even though you have left us all too soon, you moved mountains while you were here. Thank you for letting me train as a DAN! Clinician, and for allowing us to translate the protocol to Spanish. Thank you for giving us Infantile Autism in 1964, changing forever the thinking that autism was caused by the refrigerator mother. And setting into motion biomedical interventions for healing our children I wish we had more like you. I try to think "What would Bernie do?" and I usually get the answer. Always helping and always available. Humility. You set the standard for humanitarian.

Thank you to all the families on our forums for blazing a trail for others to follow and fighting for the health of your children every single day. You are an inspiration.

To all the Mods; Ginette, Caryn, Joy, Alison, Heidi, Michael, Pam, Katya, Carolina, Nilesh, Mirena, Robin, Debbie, Sue, Susan A., Brandi, Don, Clint, Maggie, Claire, Amber, Dawn, Naomi, Maryann, Susan R., Stacey, Jessi, Lina, Boris and Susanne, Olive, Dana, and Pat, for being the best mods in the world. Your help changes lives for the better each and every day. This is for you:

> *It's in our interest to take care of others. Self-centredness is opposed to basic human nature. In our own interest as human beings we need to pay attention to our inner values. Sometimes people think compassion is only of help to others, while we get no benefit. This is a mistake. When you concern yourself with others, you naturally develop a sense of self-confidence. To help others takes courage and inner strength.*
>
> ~ The Dalai Lama

Thank you Joy for sharing with Alison that you had heard about CD for autism through the seminar that Jim had in the Dominican Republic. That gesture opened the gates to the north and from there to everywhere. You are a special healer.

Teri and Ed Arranga, thank you for the platform and for helping so many families find what they need for their children, and never blinking, even when it gets really scary.

Thank you Doll, for a much-needed distraction from autism when I really need it most. BFFs are good at doing that.

Thank you to the Mansours, for believing in my projects and me. Your support made this book and our website possible.

Thank you to the various proof readers who helped with that fun process, including Michael Harrah, Pam Gotcher, Joy Whitcomb, Charlotte Lackney, Don Kalland, James Beyor, Cathy Fuss, Jeremy Horne, Ph.D., Luane Beck, Candace, Andreas Schreiber, Olive Kaiser, Susan, and Clint Melanchuk.

Thank you Mads, for your website magic and our beautiful logo.

Thank you Carolyn Unck for helping get this book into shape and all your support and advice.

Thank you Marco and Stefania for thinking out of the box and preserving the truth because it works.

Thank you so much Pam Gotcher for going above and beyond everyday to make sure our readers get their books!

Thank you Scott McRae, Brenda McRae, and Charlotte Lackney for contributing the CDH chapter. This new method of preparation has already been beneficial to many of our families, and is an exciting new addition to this book.

The purpose of life is to contribute
in some way to making things better.

~ Robert F. Kennedy

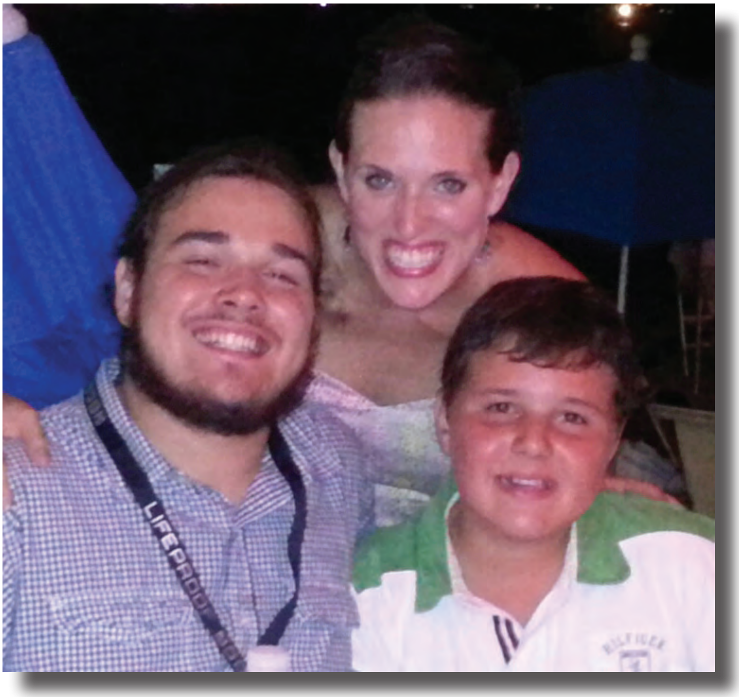

Alex Rivera, Kim McDaniel (Kerri's sister), & Patrick Rivera

Introduction

Autism is Avoidable, Treatable and Curable

Your body's ability to heal is greater
than anyone has permitted you to believe.
*~ **Anonymous***

Congratulations on finding this book, and welcome to the world of autism recovery. This book comes to life as more and more children with an autism diagnosis respond and recover in more than 58 countries around the globe. This book gives families a do-it-yourself guide to an *Autism Spectrum Disorder* (ASD) recovery program with answers all in one place.

I have, on my ASD recovery journey with my son Patrick, personally been frustrated with the lack of information and answers leading to lost time and money. For example, when I first knew that Patrick was no longer developing "normally" (in 2003), I was unable to get a diagnosis. Seven years, dozens of interventions and hundreds of thousands of dollars later I was still searching for pieces to Patrick's autism puzzle.

I learned over the years of many people recovering their children using various protocols and interventions, and I looked into all of them. Some gave us improvements but not recovery (most specifically for Patrick, diet). Some gave us nothing.

My goal with this book is to alleviate that frustration and loss of time and money for other parents.

I became interested in chlorine dioxide (CD) in 2010 but I was unable to find any information about using it with autism on the Internet. Since I knew that almost every child with autism suffers from similar pathogens (viruses, bacteria, candida, and parasites), heavy metal toxicity, inflammation and allergies I researched those conditions in combination with chlorine dioxide—removing "autism" from my vocabulary.

I realized with further research that CD would be excellent for curing the symptoms collectively known as autism. When Patrick was first diagnosed in 2004, his *Autism Treatment Evaluation Checklist* (ATEC) score was 147 and after six years of biomed he was at a 63. (The diets made the biggest difference in those initial dropped ATEC points). 2½ years of CD later he is at a 21. CD has made all the difference in his life, in my life, and in so many lives around the world.

I brought CD to my *Defeat Autism Now!* style Clinic in Puerto Vallarta in 2010. Today more than 115 children globally have lost their diagnosis of autism (meaning an ATEC score of under 10 points). Additionally, thousands of children around the globe have dropped points on their ATEC and are moving towards recovery. 27 children in less than 1 year in Venezuela alone lost their diagnosis of autism with a combination of diet, CD, and ocean water to the surprise of the doctors who diagnosed them in the first place. Many of those same doctors are now looking at CD for other patients.

It is my dream that every family of a child with autism be given this information so that they can decide for themselves whether they want to give it a try.

This book is a protocol for all of us. Some of you may be completely new to autism recovery; some of you may be a veteran like me and/or parents of older children and adults on the spectrum. This protocol works for even the classic "non-responders" and for those who are so close to recovery yet can't seem to get through the door. This book is for you. CD helps the body heal the symptoms we call "autism" across the board—it is an equal opportunity healer.

I know from experience that an autism diagnosis is devastating on many levels. The initial regression of a neurotypically developing baby takes away eye contact, speech, and the emotional connection between parent and child. Then, when strange new behaviors such as flapping, squealing, rocking, spinning or even self-injurious behaviors appear, you know in your deepest mommy (or daddy)-gut that your child was not born like this. It seems to take forever to get the truth about what happened to your happy, healthy baby. It takes even longer and is more confusing when you have to figure out how to heal that sick child. Far too many "health" providers in the autism field are focused on making money so we can't blindly trust anyone. We must do our own homework. The journey itself is a lot of trial and error coupled with misinformation.

Many supposed "autism experts" don't know much about recovery, or the effective order of treatments and end up costing our children time and the parents money. The less time a child spends chronically ill, the easier and faster it is to recover them. Not to mention the child spends less of their lifetime suffering the physical, emotional and mental effects of autism.

I feel that all parents who begin this protocol should expect a full recovery from autism because this protocol treats what causes this diagnosis. Our research indicates that every person with a diagnosis of regressive autism has virus, bacteria, candida, parasites, heavy metals (biofilm), inflammation and allergies. This protocol handles every one of these issues, and that's why it has been so successful. Some recover faster than others. But, every day we are one step closer to the end of autism.

How do you know if the protocol is working and how long does it take to see results?

The ***Autism Treatment Evaluation Checklist*** (ATEC) is our measure. The ATEC is an online survey that evaluates the severity of a child on the spectrum. For more information see Appendix 4, page 447. Many families notice changes from day one, while others take longer. You will get results when you correctly apply the interventions in this book in the proper order and without breaks.

What are the results you can expect from the protocols in this book?

I would love to say that everyone who follows the protocol will get their child down to an ATEC of 10 or less—what we call a recovery—and we have 115 of those already. The majority of those I'm in contact with report substantially significant improvements, even if they have not reached recovery. In the case of my son Patrick, he started at 147 in 2004 and has come down to 26. I'm optimistic as I continue to search for further answers, and will continue to share what I find.

If you are not getting results and have gone through this book including the Frequently Asked Questions (FAQs) and troubleshooting, please contact me through the forum at...

www.cdautism.org

...there are always tweaks we can make to keep moving towards recovery.

I recommend you read this book straight through in the order that it is written as that is the order in which it is to be applied. Jumping around the interventions can lose time for your child and waste money for you. Doing interventions in the correct order, when your child is ready for them, is the best to way to achieve recovery. Diligence and perseverance win the race every time.

My mission is to share with whoever is interested, the blessings that I have received. If the information presented here feels right and resonates with you, then please, give it a try. It might just be what your child needs.

This book in a nutshell:

If you want the best chance at recovery, here's the overview of how to do it:

1. The Diet: Eliminating gluten, dairy, soy, sugar, and toxins, to stop inflammation and reduce the overall toxic burden.
2. The CD Protocol to kill pathogens while using a multimineral like ocean water.
3. The Kalcker Parasite Protocol.
4. Explore and implement other potentially synergistic supplements to aid in speech facilitation, neurotypical behavior and/or seizure reduction.
5. Consider gentle chelators.
6. After 3 parasite protocols and consideration of all above steps, find a hyperbaric chamber (1.75ata).
7. Consider adding in GcMAF.

It is important to consider all of the pieces of information in Chapter 14, *Miscellaneous Information You Should Know* (page 323) and apply them from the beginning when they make sense for your child. You also may find the *Summary of Protocols* in Appendix 12 a good resource when you do not have time to reread a chapter to find something specific.

On the right is what we call the *Stairway to Recovery*. Joy Whitcomb, one of our amazing moms, came up with this so you can see just how each step rests on the previous steps, and without them you would not reach the top step… RECOVERY!

Author's Note: It is never my intention to "change someone's personality" or take away their character by healing autism. I see it quite the opposite. When children start to recover, their personality starts to shine through. The behaviors that we saw before (spinning, flapping, squealing, smearing feces, auto aggression, lining things up, tantrumming, etc.) are not personality traits but symptoms of a sick body. These symptoms start to disappear once the body begins to heal and our children can express who they really are through smiles, eye contact, words, gestures, etc. They can show us what they need and want and play an active role in their own lives. It is my dream that every child has the opportunity to mature and choose the life they want for themselves, and be responsible for making their own decisions. I truly believe this is possible for all of our children and adults on the spectrum, and I want families to have the opportunity to offer healing to their children.

Throughout this book we have used the pronoun "he" when referring in general to "a child on the spectrum." This is not to alienate families with girls or women on the spectrum. It is simply a question of fluidity. To use *he/she* or *his/hers* every time we chose to use a pronoun seemed cumbersome, so therefore we are using "he" or "his" throughout the book. We chose "he" rather than "she" because autism is five times more common among boys than among girls. As of March 2013 the CDC revealed the results of a new study conducted during 2011 and 2012 which surveyed 95,000 families and estimated the prevalence of autism at 1 in 50 children.

The acronym DAN! is no longer applicable to *Defeat Autism Now!*, as it belongs to *Divers Alert Network*. The acronym has been used in several places in this book as several of the personal anecdotes are from a time when its use was still appropriate. Today, a "DAN! Doctor" could be defined as a practitioner who received training through the network formerly known as *Defeat Autism Now!*.

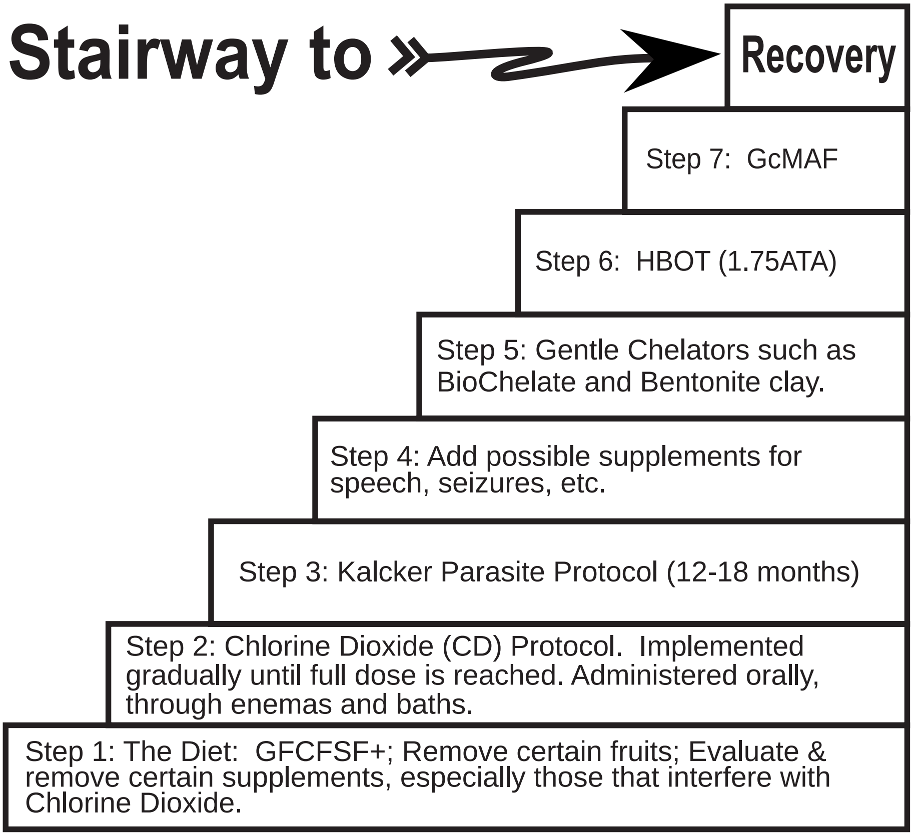

Stairway to » Recovery

Step 7: GcMAF

Step 6: HBOT (1.75ATA)

Step 5: Gentle Chelators such as BioChelate and Bentonite clay.

Step 4: Add possible supplements for speech, seizures, etc.

Step 3: Kalcker Parasite Protocol (12-18 months)

Step 2: Chlorine Dioxide (CD) Protocol. Implemented gradually until full dose is reached. Administered orally, through enemas and baths.

Step 1: The Diet: GFCFSF+; Remove certain fruits; Evaluate & remove certain supplements, especially those that interfere with Chlorine Dioxide.

Concept by Joy Whitcomb.

Everything yields to diligence.

~ Antiphanes

IMPORTANT NOTICE

Please keep in mind that the protocols in this book are still evolving, and will continue to be improved as new discoveries are made. We will release new editions to incorporate those new discoveries. The topics of CDS & CDH are particularly new and rapidly evolving. This book is current as of January 2014. Please be sure to check this book's website for important corrections and updated information beyond this and subsequent editions:

HealingTheSymptomsKnownAsAutism.com

Chapter 1

Kerri's Story

"The impossible is now declared possible as soon as you agree to it. It's just a flip of the mind, letting go really, nothing more."

~ *Stuart Wilde*

"What happened, what did you do to Patrick?" That was the first thing my husband asked me when he first saw our son, after returning from a weeklong trip. This was just five days after our youngest son Patrick received his last vaccine—the DPT (diphtheria-pertussis-tetanus) + Hepatitis B + Influenza B (what is known in Mexico as the *Pentavalente*) on the 13th of August 2002 at two years and one day old.

That question was the first of many to propel us down our path paved with autism. I told Memo (my husband) that we shouldn't worry. The nurse mentioned that we could expect a fever, or he could be listless. These were completely normal reactions. Contrary to what she said, what we observed during those first days and weeks were loss of eye contact, flapping, toe walking, high-pitched marine noises along with excessive drooling—drooling that would soak the front of his clothing.

Patrick had also lost all of the speech he had acquired; Mama, Pa, *agua* (*water* in Spanish), letters of the alphabet, numbers... all of it. The only thing he wanted to do was watch videos while running back and forth in his bedroom squealing the "ambulance sound," flapping and banging his gut, and drooling through all his clothes.

Not being well versed at the time in these problematic symptoms, we chalked it up to a case of the terrible twos. But these particular terrible twos led to loss of sleep for Patrick as well as the rest of the family, antibiotic use for green nose and eye mucous, and raging diarrhea that was so acidic it would burn his skin upon contact. This would be the story for the rest of the third year of his life.

The first of many to associate Patrick's behavior with autism would be my great aunt. She mentioned to me that she believed Patrick had autism after observing him at a family get together in April of 2003 while we were visiting relatives in Chicago. It was the most ridiculous thing I had ever heard.

Nevertheless, I went home that evening and immediately Googled the symptoms of autism. Laid out on the website were symptoms like spinning objects, lining up objects, self-injurious behavior, lack of socialization, and several other factors that I felt in no way described my son. I dismissed the comment, and continued to watch my son display strange behaviors, without understanding why.

I was back in Chicago a few months later in July. While going out for a run, I saw a girlfriend of mine who has a child the same age as Patrick. We stopped to chat and she asked me how he was doing. I said, "fine." She asked me if he was talking yet. I told her he had developed language, but since March he had lost all the vocabulary that he had previously acquired. My girlfriend looked at me and said "Oh…" with a look of concern on her face. I asked her, "What's wrong? What does that mean?" Now I was extremely nervous. "Well," she said reluctantly, "loss of language is a red flag for autism." There was that word again. I told her that we had already gone through all that, because by then I had taken Patrick to a neuro-pediatrician in Guadalajara, a psychologist with a huge center in Guadalajara, and a local psychologist in Puerto Vallarta. All three of these experts had told me that he was fine; they didn't see any problems with his development.

I started on my jog, and about halfway through, it hit me that Patrick did have autism, so I sprinted home, sat down at the computer and went to the *Autism Society of America*'s website and found a checklist of 16 symptoms of autism. The directions said if your child had 12 or more than most likely he has autism, Patrick had exactly 12. Looking back, he most likely had 14 or more, but I wasn't ready to see those just yet.

That same day I called his pediatrician in Puerto Vallarta and I told her I thought my son had autism, and she told me, "No, I have never seen anything like that in your son, but bring him and I will take another look." When we went in she observed him. She said that he didn't line things up, he didn't hit his head, he would still come to me, he was 'playing' with some toys in the waiting room, so therefore my son didn't have autism.

The pediatrician told me to go home that day to wait for him to straighten out. All the experts told me he was the way he was because he was the baby of a bilingual family that—according to them—traditionally produces late talkers; his parents and brother were all late talkers. He was spoiled. He had a nanny. And, he's a boy—boys are late talkers, etc., etc. The pediatrician successfully talked me out of the diagnosis once again.

As nothing was wrong, that fall we put Patrick into a kindergarten. His teacher would tell me Patrick isn't doing this activity or that, and I replied that we had taken him to see all the specialists, and he was just a late talker. She was super sweet about it and every few months she would make a comment to me because she saw him drastically different from his classmates.

Then one day it happened, I got a call from the director of the school saying that her friend, a neuropsychologist was in town from the United States, and she would like her to see my kids. At 6pm on March 12, 2004 I had an appointment with this neuropsychologist. I had already figured that this woman wanted to see Alex, my older son because he hadn't been doing well at school. He hadn't been sleeping well since Patrick hadn't been sleeping, which led to some poor performance in school. I had already gotten the news that Patrick was fine, so when she started focusing on him and his behaviors I was slightly confused. We sat down in his classroom, and she started asking me if he always… runs in circles, flaps, drools excessively, squeals like a dolphin, etc., (we actually used to joke that his mother must have been a dolphin). After that, I told her how we had already gone to all of these specialists and they said he was fine.

I was tired of getting the run around, and then everybody asking me what was wrong with my kid when he was just taking some extra time getting going. That's when she said to me, "I can't believe they didn't tell you your son has autism."

These words changed my life forever. Of course I asked her if she could be wrong, and she said, sure there exists a possibility, but she had done her post grad in autism, had seen hundreds of cases and this diagnosis was her professional opinion. That opened the floodgates for a river of tears that didn't end for years to come.

Being a positive person, I asked her what I should do. She said, "I would like to introduce you to a group of psychologists that are in town." The next day I went with her to a hopeless place, with hopeless people and I asked her if this is something that we can cure, and she said, "No, you can do therapy with these psychologists, and that's it." Children are born with autism and they will die with it, was the basic sentiment. I knew for sure that my son was NOT born with autism. He was the smartest, bright-eyed baby ever, and we had the photos and the videos to prove it. He was not born the ghost of a child that we had now. I knew that I would keep searching until I found something else for Patrick. I became proactive and never returned to that place.

The very next day I saw another girlfriend and she mentioned that she had a book on ADD and autism, so I immediately picked up the book and it was all about diet. Gluten-free, casein-free to be specific, and I decided to start it right away. Truth be told, Patrick's diet was horrible, he had self-restricted to only dairy and wheat. Bread products and cheese were his staple foods, but the good news was he still ate potatoes. Even as an autism novice, I knew I couldn't take him to a fast food chain because those fries were coated in gluten. We started off with just homemade fries with sea salt, because that was the only thing left on his diet that he would eat. After three days of being on the diet he said three words, the first three words he had said in over a year. I knew we were onto something.

The next week I ran into a friend from tennis, I really didn't want to say hi, because I was really depressed, but I heard a voice saying to me "It's not about the road but about smelling the flowers along the way." So, I forced a smile on my face, and went to say hi to her in my great depression. Well, she started to complain about her week, so I listened patiently, and then I told her about my week; Thursday my identity was stolen on the Internet, Friday my son was diagnosed with autism, and Saturday my dog of 14 years had to be put to sleep.

When she heard all this she turned off her car and told me how sorry she was. She said that she would put me in touch with her friend that had opened the *Early Autism Center* in Toronto, Canada. When I woke up the very next morning, I had a long email from Norah Whitney waiting in my inbox. She would turn out to be the first of many autism angels in my life.

I received a lot of precious information from that email, but what was maybe the most important detail for me was that what had happened to my son was an effect of all the vaccines he had received, he wasn't born with autism, as I knew. I also hadn't caused this, and I needed to start letting go of some of my guilt. Norah also told me that we can treat autism and I needed to immediately contact a DAN! Doctor, and Dr. Bobby Newman, Board Certified Behavior Analyst and Licensed Psychologist. Norah said that they were the best, and that this group of DAN! Doctors were curing autism.

I got in touch with everybody and that same month we started our own *Applied Behavior Analysis* (ABA) program. I also took Patrick to the States to see his first DAN! Doctor. When I got back home from that trip I had bought nearly $5,000 (US) of supplements and injectables. That didn't mean I knew how to use them, and by no means was I watching my son improving before my eyes. This was June of 2004, and we went through the rest of that summer with some supplements, other biomedical interventions, and 40 hours a week of ABA therapy.

That fall, some friends mentioned to me that their father was receiving chelation in San Diego, and I had just heard someone else saying that chelation was working for kids with autism, due to their extreme metal toxicity. By the end of the evening I had the phone number for the clinic that was providing his chelation. When I called for information, they told me that they referred all of the children's chelations to Dr. Woeller in Temecula California, so I made Patrick an appointment. In March of 2005, I took the whole family to Temecula to see Dr. Woeller, because basically nothing was happening with the supplements I had been giving to Patrick. I knew I needed to keep exploring other avenues.

Several thousand miles and various tantrums later, we arrived at Dr. Woeller's office. I told the receptionist that we had come to see Dr. Woeller. To my chagrin she said, "No, you have an appointment with some other doctor, Dr. Woeller isn't even in town right now." My husband was sure I had messed up the appointment and was extremely upset with me. Meanwhile, Patrick was screaming, crying and taking his clothes off in the waiting room. They finally let us through to speak with the other doctor and after going through Patrick's autism we explained that we wanted to chelate. We had the understanding at that point that his autism was coming from the mercury in the vaccines. She told us point blank that we couldn't do everything at once, and before we could chelate we would have to clean up his gut. I flew home with my family totally deflated. We started on the impossible journey of cleaning the gut.

My first ever autism conference turned out to be *AutismOne* in late May of 2005. I met a lady who was a rescue angel for *Generation Rescue*, and I told her I wanted to chelate my son with those DMPS drops, which were fashionable at the time. She told me that when it came to my son's recovery I shouldn't take no for an answer. I had to advocate for my son, and not let myself get pushed around. With that wisdom, I called Dr. Woeller's office again and told them I wanted my appointment with Dr. Woeller himself, and I would not be given the run around. The manager listened to me and said that she would put me through to him. After a long c'onversation, Dr. Woeller agreed to take Patrick's case and that he would help me chelate my son. He agreed with me that it was silly to think that we could have the gut completely under control before beginning chelation. Eventually I got the DMPS drops for Patrick that I had so coveted, and I felt we were back on track. However, after about six months of using the drops we still saw no change.

Flashback to that same *AutismOne* conference, where I had attended a lecture about the *Specific Carbohydrate Diet* and the removal of all grains from the diet—a diet that helped Patrick inch forward minimal improvement by minimal improvement.

In November of 2005, I had a phone consult with Dr Woeller, and having been disappointed with the transdermal DMPS I decided to ask if there was anything new in the world of autism. The answer that would end up changing all of our lives was hyperbarics or HBOT (hyperbaric oxygen therapy).

I learned that there was a really nice man by the name of Bob Sands in San Diego who owned a corporation that manufactured hospital grade hyperbaric chambers. I phoned Bob's office because Patrick was going to need 40 sessions right away, and I wanted to check prices and see if there was a package discount for hyperbarics. The answer was yes, in fact there was a discount. I scheduled Patrick's HBOT, and took both of my kids to San Diego, where we stayed for 20 really long days to get Patrick his first 40 sessions of hyperbarics—two a day—everyday, morning and night.

During that time, my husband and mom would call and ask me if Patrick was better, but he was still naked in front of the TV, jumping up and down, flapping and squealing. It wasn't until we got back home that he began to pronounce the first syllables of all of the words for things he wanted, like "ap" for apple. We considered hyperbarics a great success, but like I said that wasn't until a couple of weeks after we finished the 40 sessions. That's when we really began to see the changes in Patrick. Bob always says that HBOT is the gift that keeps on giving; you can see benefits for up to two months after you finish your sessions.

Meanwhile, Bob and I hit it off right away. He had a real jovial and family-oriented atmosphere in the clinic. I told him my story and how I wanted to help people know that autism is avoidable, treatable and curable. I shared with him that there was no information, much less Biomed or autism recovery in Mexico (and most of Latin America). I wanted to help people and to share with people that there is a lot we can do to help our children heal.

The very next day he walked into the office and changed the course of my life forever. He told me that he was in fact friends with Dr. Bernard Rimland, "Bernie," the Grand Godfather of Biomedical treatment for autism, and the author of *Infantile Autism*, *Dislogic Syndrome*, and the founder of the *Autism Research Institute*. I told Bob that meeting Bernie today, would be akin to having met Mick Jagger when I was 15. He then announced that we would be having lunch with none other than Bernie himself and Mrs. Rimland. For the first time in my life, I admitted to Bob that I was so excited that I didn't know what to say. Bob told me that when it was my turn to say something that I should ask Bernie "What can I do for DAN!?"

In a tizzy I ran over to Marshall's, the discount retailer; bought a new suit, hose, and even a pair of high heels. The next afternoon, I put on my new ridiculous clothes, ditched my kids with my friend's maid, and hopped in Bob's Jag to meet Bernie and Gloria (Dr. and Mrs. Rimland)—an absolute dream come true. When we arrived at their favorite restaurant, I was completely overdressed. Gloria ordered a salad, and the guys had tilapia because Bernie didn't really like vegetables. The conversation ranged from Kinotakara patches, to the difference between soft-sided hyperbaric chambers and hard hyperbaric chambers.

I waited until there was a lull in the conversation and I went for it, I asked him what could I do for DAN!, and he said I should translate the DAN! Protocol and take it to all of Latin America. His words left me totally dumbfounded. Had I heard right? Latin America? I was thinking my town of Puerto Vallarta, or maybe Jalisco, (the state we lived in) and maybe in my wildest dreams all of Mexico, but this was much bigger than I had imagined. At this point, though there was no stopping the momentum that we had started. Within months we translated the Protocol to Spanish, and donated it to the Autism Research Institute, for dissemination throughout Latin America. Much later I would find out that one of my dear friends, Yeroline, would heal her son from autism using the translation of the DAN! Protocol, as well as others.

During that same trip my husband and I spoke to Bob about buying one of his chambers for the not-for-profit autism clinic we were planning on opening in Puerto Vallarta. The plan was to run the clinic on a non-profit basis, but anchor the clinic by charging the public for their sessions in the chamber, allowing all the profits to pay for children with autism to go for free in the chamber.

It was a green light all the way. We deposited the money for the chamber in March of 2006, and it arrived on October 31, 2006. AutismO2—Hyperbaric Clinic officially opened its doors on Dec 1, 2006. We threw a party for the inauguration, invited friends, family, Patrick's therapists, and our local priest came to Bless This House. We all wore white, and in honor of Patrick, who was our motivation for opening the clinic, we all had on nametags that said *Kerri—Patrick's Mom*, *Memo—Patrick's Dad*, etc. It was a special evening for all of us, but me especially, as it solidified what would be the path that I would still be walking as I write this book seven years later.

The clinic had a Sand's Hyperbaric Chamber, an allopathic physician with a specialty in Hyperbarics, 2 psychologists, 2 hyperbaric technicians, and me, to initially meet with the parents. As for the chamber, we always knew that it was going to help a lot of kids with autism, but it has helped a lot of others as well. The chamber was what gave language back to Patrick in 2006.

Later that same year we sent our allopathic doctor with our naturopathic hippie doctor to Ixtapa, Mexico for a hyperbaric conference, where they serendipitously met Dr. Giuseppina Feingold, (Dr. Jo) a DAN! doctor using hyperbarics for children with autism. As soon as our guys got back to Puerto Vallarta they started insisting that I had to contact her because she is curing kids with a protocol that included hyperbarics. Never one to waste time, I emailed her immediately, but at the time I didn't know she wasn't an email person and, because I am just not a phone person, we didn't connect.

Fast forward to January 2007, my husband, Memo, was buying a 1951 Desoto on eBay. We had to pay for the Desoto with a check, and therefore started to talk to Bryan, the seller of the car. During their conversation, Memo, my husband, said to Bryan "If you don't want to sell, because it's worth more than your selling price I will understand." Bryan had already decided to let it go, but at the same time he was interested in where the car was going, what Memo did for a living, etc. Memo told him about our life in Puerto Vallarta, and the classified ad magazine business that he owns. That's when Bryan stopped him and said, "I've heard of that magazine!"

It turned out that this Bryan was a nurse who works with a doctor who does hyperbarics and heals kids with autism. Well, that piqued Memo's interest, so he told Bryan about Patrick's autism and then Bryan says you must call Dr. Jo, yes, the very same Dr. Jo that our clinic doctor had met months earlier at the Hyperbaric Conference in Ixtapa. So, I called her immediately and when she answered I said who I was, and I told her Bryan told me I had to call her. I asked her if she believed in God and she said yes. I went on to tell her that in September she had met with my clinic's doctor and how I had emailed, but never heard back. At the time, Dr. Jo had been receiving so many emails that sometimes she couldn't get to all of them.

Finally we had connected, and we hit it off right away. I immediately told her all about the clinic and Patrick, hoping that she could come to Vallarta. She told me that I had to go see her in New York first, and we would get Patrick some treatment. So off we went to freezing cold New York in March of 2007.

We immediately started treating Patrick with IV chelation. The theory at the time was the main factor causing autism were heavy metals coming from vaccines that were damaging the methylation pathways. During my daily treks to Dr. Jo's office I finally met Bryan, the owner of the Desoto that my husband bought on eBay. We started talking about my clinic and what we could do collaboratively if Bryan were to move there. Bryan had already been in nursing for almost 30 years and was an expert in ozone and other alternative therapies. He told me he was ready for some changes, and we were very interested to have someone of Bryan's caliber and relaxed personality to join us at the clinic.

Within a week of my return to Puerto Vallarta Bryan came for his first visit, to see if he could call Vallarta home. Two months later he came back with all of his equipment to set up his office. We were a great fit, and Dr. Jo would fly down from time to time to help out with patients.

From 2007 to 2008 we were treating Patrick with IV chelation on top of his regular GF/CF/SF, diet, supplements, and hyperbarics. In May of 2008, I met a mom who had recovered her autism spectrum child with homeopathy. Almost immediately I started working with a world class homeopath from June 2008 to May 2009, but I didn't see anything to make me feel that the cure to autism was down that path. That same year I met a doctor doing the Yasko protocol, and she had some ideas, so we gave it a try from August 2009 until the end of May 2010. At that point, Patrick honestly looked worse than before we had started giving him 80 supplements a day.

By now I had become disenchanted with *Defeat Autism Now!* based megavitamin protocols. A precious few of the families that we helped recovered their children with diet, supplements, chelation and hyperbarics. However, the overwhelming majority still had an autism diagnosis after tons of work by their parents, and usually a lot of money spent on supplements and treatments. I started feeling like it was a fraud, telling people to follow this protocol, which I knew wasn't going to be enough to recover most of the kids. That's not to say that we weren't seeing vast improvements, but counting all of the children that we worked with, only two children recovered.

By July of 2010, I was totally disillusioned and befuddled and I didn't want to continue doing what I was doing the way I was doing it. So, I asked the Universe/God/Angels—whoever was listening, for help. If my mission truly was to help families recover their children from autism then I was going to need a tool to work with. One that was available on every continent and that was affordable to everyone, because what we had just wasn't doing the job.

No magic voice came, thank God! Because that would have really freaked me out. However, I started to remember these colorful little bottles of chlorine dioxide that I never used. I decided to research their use on Google. Disappointingly, there was absolutely nothing on the Internet about autism and Miracle Mineral Solution (MMS) aka chlorine dioxide (CD). So I started thinking about what autism is made up of. So, I Googled chlorine dioxide with virus, bacteria, candida, heavy metals, blood brain barrier, allergies, and inflammation. The evidence was overwhelmingly positive, which showed me that CD could treat all of the components that make up autism. I had hope once again.

I was especially interested because at the clinic, we specialize in oxidative therapies such as hyperbarics and ozone. As chlorine dioxide is more benign than what we were already using, I decided to investigate further.

No side effects except a possible Herxheimer reaction; which is not a side effect of chlorine dioxide itself, but can happen with any detoxification protocol. I decided to speak to my husband and son Alex, who were also excited. The next day at the clinic, my husband's best friend's cousin and his wife were getting out of the chamber. I said, "Hi!" and she immediately said to me, "I'm taking CD." She didn't say "Hi!" or "Kerri!" just, "I'm taking CD." That was the definitive moment for me—my Aha! moment. I told her I had been researching it for weeks, and I was extremely interested. She was having great results, so my husband said he would try it first. If after three days on the drops he was still alive then we would start Patrick on them.

I contacted Jim Humble, discoverer of CD. I was hoping he would help me to better understand how to dose CD for children with autism. I explained to him that there was nothing on the Internet for kids. He helped me do just that. He gave the following recommendations: 1 drop 8x a day for children under 25lbs, 2 drops 8x a day for children under 50lbs, and 3 drops 8x a day for children under 100 lbs. He told me that the more doses that we can get in one day the better, 8 doses are the minimum.

That first week Patrick vomited (classic Herxheimer reaction), because I went too fast with the dosing. On the Internet, the only protocols I found gave high doses a few times a day, and as I found that week low and slow doses all day was the way to go. However, despite the Herxheimer reaction I caused Patrick (through lack of a low and slow dosing protocol) he was still noticeably better. Seven days later my son had improved eye contact and was asking for things that he had never requested in his life. At 9pm he looked me straight in the eye and said, "I want bed." With my jaw hanging open in disbelief, I followed him upstairs to his room. When we got there he turned to me and looked me straight in the eye again and said, "I want take bath." I knew I was not dreaming, and I had really just heard that. After his bath he looked me straight in the eye and said "I want brush teeth" and the whole time we brushed his teeth he was giggling... so I asked him what he wanted and he said, "I want 'kanket'" so I said, "blanket" and he repeated, "blanket, yes" and ran to the bed, and jumped on it to enjoy the blanket. He had never jumped into the bed in his life before that. This was the first seven days on CD. I was blown away.

By September of 2010, every person that was previously only using herbs or medications to kill viruses, bacteria, candida and other pathogens was about to hear about CD. That's when things really started to happen.

Back in 2007, I learned how Dr. Anju Usman was having great success with her biofilm protocol. She concluded that virus, bacteria, candida, parasites, and heavy metals are all joined together in the biofilm (more on this in Chapter 5, starting on page 117). And when I saw that CD killed pathogens and neutralized heavy metals and so many other things that make up the core of autism, I knew we would kill a lot of birds with one stone. I was also hoping to sidestep a few pharmaceutical pieces of the biofilm protocol, i.e. antifungals, antibiotics, antivirals and be able to use something with no side effects (a Herxheimer reaction is different than a side effect).

I was in the process of discovering how we could use this extremely inexpensive oxidizer that is available all over the world to help the body heal from autism. The other important part of CD is that you don't need to take your child somewhere for treatment, as in the cases of hyperbarics, ozone or IV chelation... there is no doctor needed, or trips to visit doctors in other countries. It's as simple as taking a supplement and you modify your dose depending on what you feel and see. Basically, any family with access to the internet, diet, CD and a few choice supplements can heal their child with autism.

After having limited success with my son on different biomedical treatments, even with the best doctors in the world, it was time for a change. With CD we are attacking the biofilm all day as CD destroys the electron shells of the different molecules making up pathogens, therefore releasing toxins into the bloodstream. This release of toxins is the principal reason one must go slow and build up the dose, to avoid a Herxheimer reaction as so many of these children are very toxic. If we kill too many pathogens at once, too many toxins enter the bloodstream. The body will immediately look to eliminate them, most notably through diarrhea and vomiting. This is unpleasant and totally avoidable.

CD is so benign you can use it on your skin, hair, ears, eyes, orally, rectally, vaginally, inhaled, etc. At the doses we use CD in aqueous solutions; it is not detrimental to healthy cells. It specifically targets pathogens due to their negative charges. Once I understood the basics, and Patrick was still improving, I started to share with others to understand how to use CD. Very quickly we were having success that many other clinicians weren't having. Kids on the spectrum were improving, some started recovering, and we had to sit up and take notice. I was blown away again.

In November, a child recovered and then in December another child recovered. Their families took them to psychiatrists and their doctors to have their diagnosis removed. These were the very important first steps that convinced me that this is something that we must continue to do. We started to spread the word that this was an inexpensive treatment modality available in every continent of the world.

With my background in biomed I learned that you watch for reactions while you load the dose. *Low and slow* is the rule. We arrived at the one-drop at a time dose and as kids recovered, parents shared their stories with other parents and more and more people started using it. It was very grass roots.

This is when the explosion happened. I started to think that this might be the missing piece of the puzzle that we had been looking for. In all seriousness, there is no one cure for all kids with autism, which is why each child's protocol and path to recovery is different. While we have seen great success with CD today I keep working to discover new modalities that help these children heal as non-invasively as possible. Now that we had had success in the Spanish-speaking world, I needed to share these treatments with families of children with autism all over the world.

Around this time something very interesting started happening with the CD enemas... parasites; more specifically roundworms were coming out with the CD enemas of the parents and their children using the Protocol. Today I have hundreds of photos, which were sent to me from parents all over the world (first world and third world nations) who have seen worms passed in stools.

At the moment laboratory testing is woefully inadequate, but a keen veterinarian can easily check for the presence of parasites in a stool sample using a high-powered microscope. The children who have had stool samples reviewed by microscopy have come back positive. Pinworms, roundworms, tapeworms and hookworms are the most common findings. Stool analysis done by laboratories have consistently come back negative even when worms have clearly been seen and photographed and seen under the microscope. In fact, one mother I know sent in a live, moving worm that her child had passed. The result? Negative for parasites! At this point it is simply not enough to trust a coprological analysis when looking for parasites.

Dr. Andreas Kalcker and Miriam Carrasco have been instrumental in this piece of the puzzle and have designed an amazing parasite protocol that has already helped many children including Patrick. In October of 2011, Andreas gave me the first parasite protocol and families in Spain, Mexico, Venezuela and others throughout Latin America began to use it. We'll talk about it in depth in Chapter 8, page 165 and how it has affected my life as well as the lives of so many other families with children on the spectrum.

In January of 2012, I got in touch with Teri Arranga and I was invited to speak at *AutismOne* in May of 2012. This would be the first time I would present at *AutismOne* in English, and obviously the first ever presentation on CD. After eight years of biomed and six years of helping families in Latin America I would be flying above the radar, knowing full well there would be a tradeoff. While I would reach families all over the US, for the first time I would end up taking a lot of flak in the blogosphere. We survived!

When I first came across CD in 2010 and began to watch it work it's miracles with autism, I expected parents, doctors and professionals who dealt with autism to be excited. I assumed they would begin to do research as to how and why the chlorine dioxide molecule was healing/curing autism. Much to my chagrin, many people were disinterested. Some even went so far as to say that what I had seen was impossible or that CD was toxic. Well, to that I would say that healing/curing autism with a toxic substance is impossible. Since then a handful of some of the best doctors in the world have become interested (that number is growing) and quite a few parents are taking note. Hundreds in fact went to my presentation at *AutismOne*. Several told me in hindsight that they had thought about not attending because the title of the presentation sounded too good to be true... *40 Children Recovered in 21 Months*.

The people who did attend were pleased with the information and many began the Protocol. However, what was to happen the days and weeks after my presentation blew my mind and was absolutely beyond my wildest nightmares. Some parents attacked me in print and on the Internet. I received threatening and accusatory emails filled with hate speech and foul language.

In most cases these emails and blogs were from anti-biomed parents. Others in the biomed movement told me not to worry, that these people were notorious for doing this sort of thing to others. They jump on the newest and brightest intervention in order to get their time in the limelight. Stealing attention away from the treatment, altering the truth, and in some cases lying to people about what is happening just to get other parents up in arms. Never could I have imagined anything like what transpired. However, as time passed, so did the threats, negative blog posts, etc.

Update for 2014

One of my favorite quotes states:

> *"All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident."*

Fortunately, the 2013 *AutismOne* conference was a totally different experience—which may indicate we are slipping into that third stage. In May of 2013, we were at 93 recoveries, and during my presentation at *AutismOne,* some courageous parents took the stage with me to share their children's stories of healing and recovery. There were no attacks. As the first edition of this book launched at the conference, many of our wonderful moderators were on hand to answer questions, and assist parents who were interested in getting started. I had a book signing along with the pleasure of meeting many parents I had up until then only known through email or Facebook. By January 2014, the first edition had already sold thousands of copies. If you searched for "autism" in the category of "books" on Amazon, it was showing up in various positions of the first two pages of over 10,000+ results, with the majority being five-star reviews. If you changed the order to "Average Customer Review," it was in the top 10, sometimes in position #1.

This has always been a grassroots parent driven movement, and today help is available in 7 languages online to answer questions and offer support. As with anything, if you attract enough attention, you will also attract some "haters," however CD has already earned itself a place in the treatment modalities that are healing the symptoms known as autism.

CDS (chlorine dioxide solution) was introduced in the first edition, as we were still hoping that it was something better than it turned out to be. It was a better tasting and more tolerated form of CD, and still remains an excellent preparation choice for those who are extremely sensitive, and have trouble tolerating even one drop of classic CD. However, we found that over the long haul, only one child so far has recovered with strict CDS use—the other 114 were with classic CD.

In this edition we are introducing *Chlorine Dioxide Holding [Solution]* (CDH). When this preparation technique was introduced it was touted as something similar to CDS… having a better taste, is better tolerated, along with less Herxheimer reactions. However, there is one big difference; CDH still contains a small amount of the raw materials required for the preparation of CD (sodium chlorite and citric/hydrochloric acid). On the other hand, CDS was chlorine dioxide gas ONLY dissolved in water. That small amount of raw materials in the CDH preparation may be what makes the difference. After 90 days of CDH use with over 70 families, it has not failed. Gains have not plateaued, and parents seem to be having an easier time increasing their children's dose, without any Herxheimer reactions. Another amazing thing about CDH is that the natural sweetener *Stevia* can be added to improve the flavor, while the potency of the dose does not change. This can be a game changer for kids that have taste aversion to classic CD. It should be noted that not all brands of Stevia are created equal and there may be some that can't be used. We are still testing various brands.

CDS and CDH have both earned their place in methods of chlorine dioxide preparation, thus allowing more people to benefit from the healing properties of CD who might not have otherwise been able to tolerate it.

People are always interested to know how my son Patrick is doing, and I am happy to share a little bit about what has been going on in his life lately. This past August 2013, Patrick turned 13. I had expected him to be recovered by now; however, we are still working towards a full recovery. He is better every month and his current ATEC is somewhere between 22 and 24. Patrick is very social, he loves a party. This Halloween, my sister threw a party and it was 11:30pm before he was finally ready to go home. He also loves spending time with his family. Every night he tells me, "I love you Mommy, gimme kissy." That is not only his way of telling me he wants a kiss, but that he wants me to come spend time with him before we go to sleep.

He has been preparing his own food in the kitchen, and while he has always liked to help chop, his being able to place his selection in the toaster oven and heat it himself is new. We didn't show him how to do this. He decided on his own that he would heat and serve himself his dinner one day. Another major advancement is that he is now able to clean himself after toileting, which is something he always asked for help with before. He will even wear headphones when listening to YouTube videos or watching DVDs if someone has to make a phone call.

We don't have any conduct issues, and if no one told you I had a son with autism, and you saw us out, you would never know. Apraxia remains the biggest factor delaying Patrick's recovery. That said, Patrick does communicate more than ever, and attempts more language than ever before.

As far as where I see this movement heading in the future, I believe that if truth does pass through 3 stages then we have finally entered the third stage of "self-evidence." The CD Protocol has now recovered 115 children (as of

December 2013); is used in 58 countries; and has already helped over 5,000 people on the spectrum, with more and more being added every day. The power of social media allows for parents to share with other parents their successes with the protocol, thus forming a stronger bond. Parents in the autism community trust other parents above doctors, and rightfully so.

At this point we are breaking many stereotypes associated with healing autism. For example, we now know that after the age of 9, recovery is still very possible (a 31 year old man is nearing recovery as I write these words). You do not have to be rich to recover your child from autism. You do not have to speak English—there are Facebook groups in 7 languages and this book will be translated into at least 13 languages. We now know that autism is not a psychological disorder. It is biomedical... viruses, bacteria, candida, parasites, and heavy metals cause the behaviors that lead to an autism diagnosis. Once you remove what is causing the symptoms, you can remove the diagnosis.

I witness on a daily basis what was not supposed to be possible: the healing of autism.

The future is bright, and it is up to us to share it!

Best in health,

Alex, Kerri & Patrick Rivera

Fundación Venciendo

Jim, Andreas and myself in Venezuela where 2 of the most amazing women in the world have a foundation to help Venezuela heal from the autism epidemic. I am so grateful to Yamileth Paduani and Carolina Moreno for their foundation and hard work. 28 recovered children from their foundation with this protocol in their first year of service. Thank you ladies. I am so proud to be your sister of the heart.

Healing
the symptoms known as
Autism

Ty reading and healing

Chapter 2

Yes We Can!!!

Without faith, nothing is possible. With it, nothing is impossible.
*~ **Mary McLeod Bethune***

Before we embark on the path of healing autism, it may provide you some comfort and encouragement to hear from those who have gone before you.

Some of the parents of the children who lost their diagnosis through this Protocol were generous enough to share a photo of their children along with a "thank you note." You will see that a couple of the older children wanted to write their own notes and share their success with the world.

① My son gets to play little League!

Thanks to ATEC of 4, autism IS treatable!

THANK YOU KERRI!

②

OMG! I'm still rubbing my eyes. Nathan just played with cars appropriately! He pushed them along while saying, "vroom vroom!" Until now, he would simply mouth, spin or tap toys.

-Week 5 of MMS

③ Today in my son's school they had an "I Can Run" day for good health and to collect food for families in our community. Last year he couldn't make it 1 mile - they run / walk around a track. This year he ran 2 miles... Whhaattt? 7 months on mms and 6 months treating parasites.

④
I love you Miss Kerri,
Thank you for giving me CD.
Matthew

⑤ Wow! Here is something to be grateful for! It was time to do my son's one year ATEC doing MMS (7 months detoxing parasites) and his score was...drum roll please...15! You read it right, a 15! I am tearing up right now writing this. My son is 14 years, 4 months old. He was at a 27 when we started, then went up to a 34 when we started treating parasites (behaviors worsened), then down to 21 three months ago, and now a 15! Thank you so much Kerri! Without you, who knows where he'd be.

⑥ Before MMS we were able to get my sons atec score down to a 24. Although he mainly functioned as a "neurotypical" child, he still would have behavioral problems. He was easily agitated and would struggle with some anxiety and OCD. We were constantly battling yeast, constipation, metals and come to find out, parasites! We began MMS very slowly. We noticed his moods were improving. He was smiling more and over all just really pleasant to be around. We decided to retake his atec after only 1 month. It was a shocking 4!!! It dropped 20 points in 1 month!!! We have our boy back and will forever be grateful! We will be starting the parasite protocol this month and looking forward to an atec of 0!!!

7

I just wanted to share that today my daughter said a full sentence!

She normally only makes two word sentences and often stutters and stammers, with a lot of articulation problems.

My husband came home after a week away for work and she walked up to him as he sat at the computer and she said: "I want... to sit with daddy."

And in the bathtub (I'm doing the MMS steambath as both my kids are coughing right now), she turned to her brother and said "Look brother, cup." and she held the plastic cup in front of him. She has never been this coordinated and 'present', nor articulate. I am so grateful.

Even though we still have a long way to go, I feel like a huge weight has been lifted off my shoulders.
Thank you for your hard work.

8

Thank you for helping me to rescue my son.
God bless you
Kerri Rivera!

9

My son is at a three-day camp with 100 NT 5th graders! We have provided all his food, and go up to give supplements twice a day... other than that, he is on his own sleeping with friends for the first time in his life! I have no words other than Praise The Lord for bringing me Kerri Rivera and also for giving me and my family the strength to never give up! Our little bird had taken flight... thanks be to God!

⑩

Thank you God for your fidelity, thank you my beautiful girl for giving me your gaze and your smile every day, thank you Kerri McDaniel de Rivera for walking with me.

⑪

Hello, I'm Silvia, Alejandro's mom, we live in Spain. I was giving MMS to my son for 10 months without seeing any changes. Kerri told me that I wasn't doing something right… and that was correct. Being the warrior she is, she made me tell her everything I was giving Alejandro and she got it, Alejandro drank pineapple juice all day long, he would take more than a liter a day and that made the MMS not work, I stopped giving him pineapple juice 2 months ago, and he is a new child… He pays attention, his comprehension is almost 100% I must say my baby did not speak, but with this he started saying: "Come on, mom!" "My mommy" All this because I started giving him the MMS as I should. I want to say NO juice near MMS because it anulls its effectiveness. I thank God for putting Kerri in my path. My son improves everyday, slowly but you can see that he is saying goodbye to his little world. Thank you Kerri for being in our lives and for making us see the light. Also thank you JIM HUMBLE for giving hope to our lives.

12

Thank God for putting Kerri Rivera in our path, and giving us our prince back, totally recovered.

13

My son turned 12 today - hooray!! This is truly the very first time since his 1st birthday that my heart didn't shatter into a million pieces because it's another year down the road, and still so far away from recovery!! He decided that he was staying at home today to play with his birthday toys, he told his teacher (several times over), and when she asked if she could take the day off too, he said "nooooo!!. When his grandpa called from Holland to sing Happy Birthday, he sat quietly listening with a smile on his face - the usual is "no singing!" Besides that - here's my brag for the day - his teacher told me that she had to move him and his best class buddy away from each other because they were copying each other's work!!! YAHOO - we were SO excited - yes I know, we're crazy, but hey - if it's typical, I don't care what it looks like, I WILL TAKE IT!!! If you think this is weird, you should've seen me when he bit a child when he was 3!! I want to encourage you all today, look for the small things, check in with the people who are in contact with your kids and keep a journal of the changes. You will be amazed at how much we miss, because we spend every waking moment watching over them! I feel SO blessed!!

⑭

Dear Kerri,

Thank you so much for bringing our son's smile to life. The happiness you see in these eyes comes from listening to you and your amazing protocol.

We were stunned when his ATEC dropped 18 points in 3 weeks, and has now reached a staggering 1 from the original 36 in less than one year.

I thank God every day for having me in that AutismOne lecture in 2012 where you opened my eyes to what else we could do to help him. I want every parent out there to know this is a real Protocol with real results.

Thank you again and God Bless everything you do for our Children.

You Rock! Love Maryann

---

⑮

Greetings from Monterrey.

Just to let you know, we went to visit you in Puerto Vallarta on June 21st, and my child started treatment in July with MMS. As of today my son eats almost all on his own, and wipes his mouth whenever he needs to. For about a month now he has been doing funny things; like covering up his cards when we play as if to hide them, if he sees me squatting on the floor he runs over and jumps on me. He shows expectation on his face when he knows that I am coming to tickle him. He hides behind a wall when we play hide-and-seek.

In both of the institutes where he goes they have told me he is able to pay attention for longer periods of time, his way of being has changed, and they have also said that his eye contact is much better. According to his therapist he is now a candidate for speech therapy because he is able to pay attention.

For us, Kerri, this is a miracle come true, to see our son waking up little by little and to see him eat all on his own. Thank you so much for sharing all of this with us.

⑯

Since starting MMS my daughter's play skills improved, she actually looks into my eyes and gives me eye contact which means a lot. She started answering questions, more engaging and now I started to feel as if the end of the tunnel is not too far. I am so grateful to Kerri for showing me a way to recover my child. GOD BLESS.

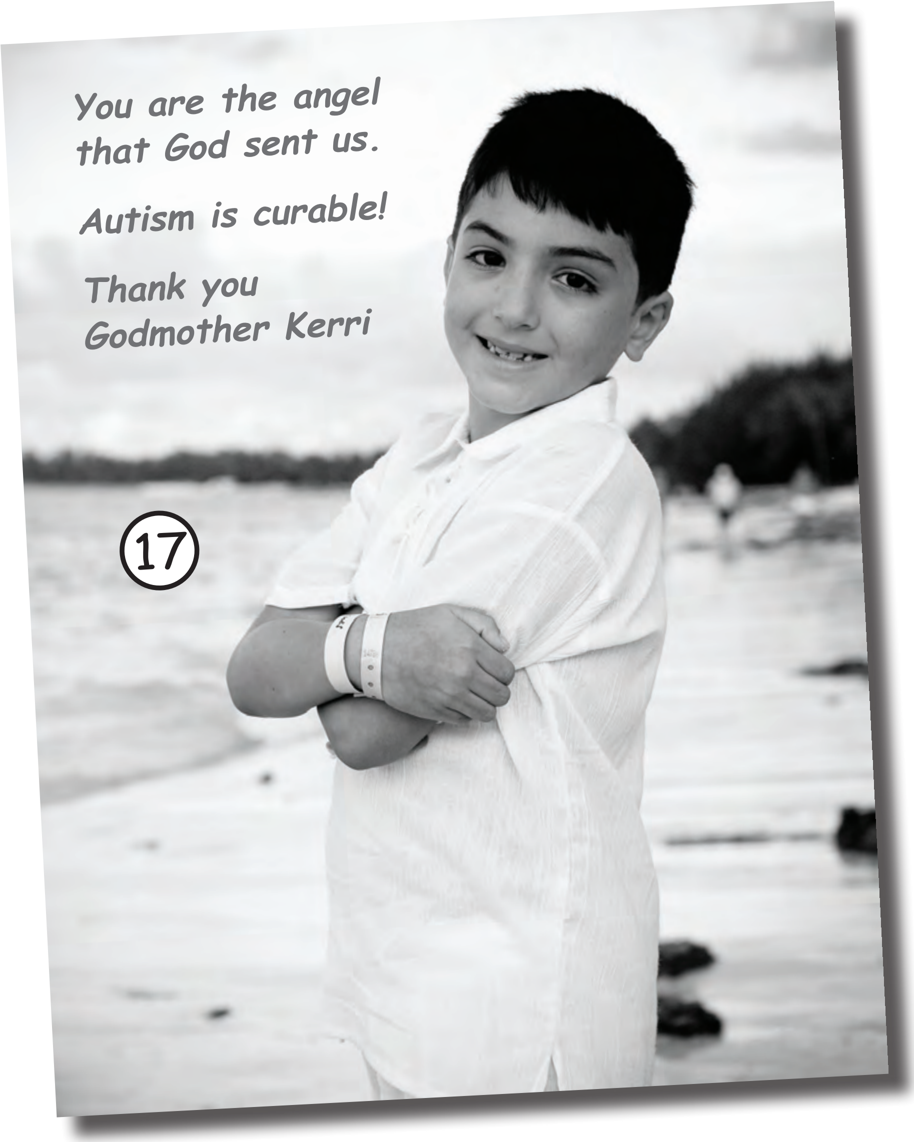

⑰

You are the angel that God sent us.

Autism is curable!

Thank you Godmother Kerri

⑱

Hi Kerri,

I cannot believe it has been 5 weeks since we spoke on the phone.

My son turned 22 yesterday. He actually opened each birthday card, and read the inside. We always knew he could read, but he Never showed an interest in opening cards or gifts. He had some interest in his gifts last night. He has been wearing only light grey shirts for 5 years. This week he wore a blue one and a charcoal grey shirt.

Looking straight at me, and asking for what he wants. No real bad effects yet. We are on 22 drops as of today. 2 enemas a day 300ml - 8 drops - ocean water in the morning.

I have attached 2 pictures from this week, I think these are worms.

Can I hope that he can get better?... After all the treatments we have been through?... well I am. I think, what will I do with my time when he is better (possibly recovered)?... I will yell from the roof tops and dedicate myself to helping other families. Is this what I was meant to do? My purpose? I would love it!

⑲

Thank you for this look.
I share this look with
you, the same look
that you have given
back to my Mom, this
photo would not be, if
it hadn't been for you.
I love you very much,
your Godson Gabriel

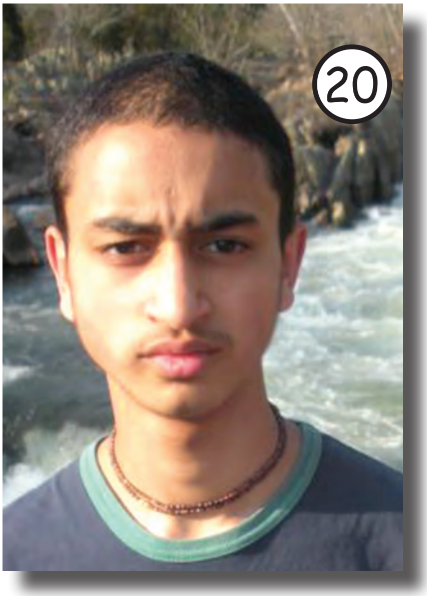
20

My family and I are very grateful to Kerri Rivera for making available the CD protocol to treat ASD which has recovered myself and made a huge difference to thousands of kids and families all over the world.

CD can recover even a teenager like me whom others had given up on. Of all the bio-medical interventions, CD cost the least and made the biggest difference.

Muchas Gracias por salvar mi vida Kerri.

21

This is Benjamin. 4 years 5 months old, current atec of 4 and dropping weekly :) Diagnosis atec at 3 years 2 months was 134. Atec when starting CD protocol was 18 and was stagnant after full on bio-med/diet intervention. Dropped all supps except melatonin when starting protocol and obviously can't be happier with that decision. Thank you to Kerri Rivera of course for giving me a way to save my baby boy... We will continue with protocol until atec = 0 :)

㉒

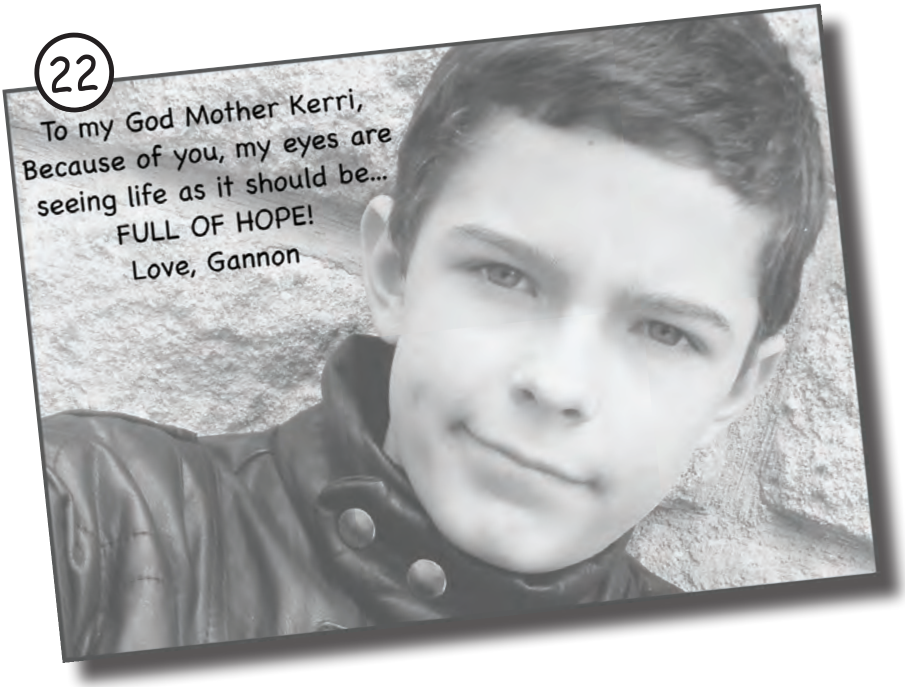

To my God Mother Kerri,
Because of you, my eyes are
seeing life as it should be...
FULL OF HOPE!
Love, Gannon

㉓

My sons started using MMS exactly 2 and a half months ago and we saw improvements almost immediately, my twins attention got better just days after we started the treatment and at 10 days after starting my son Juan Pablo started calling me mom! He has never done that before. (Neither of my twins were verbal, meaning they were not able to speak a single word.) A few days later Jesus Alejandro said "Mom, water" while pointing to the fridge, it was shocking for me.

My two sons can now say 4 or 5 words and their communication skills have gotten a lot better, even if it's not verbal they are very good at expressing what they want. There is a lot of improvement in comprehension, following orders and accomplishing small tasks, they react when you say their names, they are controlling their bowel movements, and in general their whole quality of life has gotten so much better since they started with MMS.

I know we have a long road to go but we are on the right track and I trust that thanks to the science that discovered the great effects of this wonderful formula for our children with autism we will recover them, as it truly does detoxify!

Ing. Artebys Cedeño
Mother of Jesus Alejandro and Juan Pablo

24

Dear Godmother Kerri,

How can we ever thank you enough for crushing autism out of TJ's life? We now have a smart, happy, helathy boy that has great confidence in himself!

We are forever grateful to you!!!

25

We started using MMS a month ago. We are not even at the full dose yet but have been seeing some amazing gains.

I took my son to the mall about a week ago because it is one of his favorite things to do. When we got home he did something he has NEVER done before. He came up to me, hugged me and said "Thanks". Wow.

Two days ago my husband had to go to the local farm supply store because something broke in the barn. So we asked my son if he wanted to go and he literally jumped off the couch and excitedly came with us. While he was there he picked out a couple of books he was interested in and a couple of small vehicle toys. When we got home I was helping him with his coat and boots he said "you're great mom". Never have I ever had spontaneous comments like this and the only thing I can attribute it to is the MMS because we stopped everything else. This gives me inspiration to continue the protocol and I look forward to what the future holds for my son and our family.

Thank you Kerri!

26

If it were not for you Kerri, my mother would still be worried about my future.

You are my Guardian Angel!!!

Autism is curable!

Thank you Kerri
Gracias Kerri!

27

His speech therapist told me today that his sentences are so much better, more organized, also his ability to converse back and forth and also his concentration have improved!! Sooooooooo..... Yay!
Oh, before we went to see his speech therapist he said: Mummy, my speech is getting so much better!
Yup, he said that! He knows!

28

I can't believe it! I just did my daughter's ATEC and in one month it dropped from 72 to 48!!! I am so shocked! Is this a dream? We revised the scores over and over with my husband and accepting that some of our answers could be just very positive feelings, still the improvement is huge!! I am so thankful Kerri Rivera! I have no words!! I am praying this is not a dream!! THANK YOU SO MUCH!!!! I believe there is more to come but right now I am so much in shock!

29

I have to post our Sunday gain, my daughter is and always has been on sensory overload, those of you who deal with this know that teeth brushing, hair combing and showers can be torturous for child AND parent. I literally dread washing her hair. She screams so loud she could break the windows. She smacks her wet thighs so hard that they've been bruised for years. This morning she gets in the tub, sits down and sees me roll my sleeves up (which is usually the moment she loses her mind because she knows I'm about to grab the shower head), she stood up, looked me dead in the eye, smiled at me and said "R. shower". I began to wet her hair down, still convinced that the screaming was about to begin any minute. I started shampooing and she said "well done", I smiled and said "VERY we'll done baby". My husband peeked in to see why it was so quiet and she looked at him and said "Daddy shower", smiling from ear to ear as to say 'I'm taking a shower Daddy'..... what a great way to start our Sunday. HUGE, HUGE deal for her and for us. I feel like I finally exhaled today. MMS rocks and so do all of you!!! Happy Sunday ;-))))

30

Dear Kerri,

Thank you SO MUCH for helping us with our daughter's complicated health issues. The CD/PP protocol allowed my daughter to sleep through the night, for the FIRST time in 7 years! One month on the parasite meds, she began sleeping perfectly and has slept perfectly ever since!

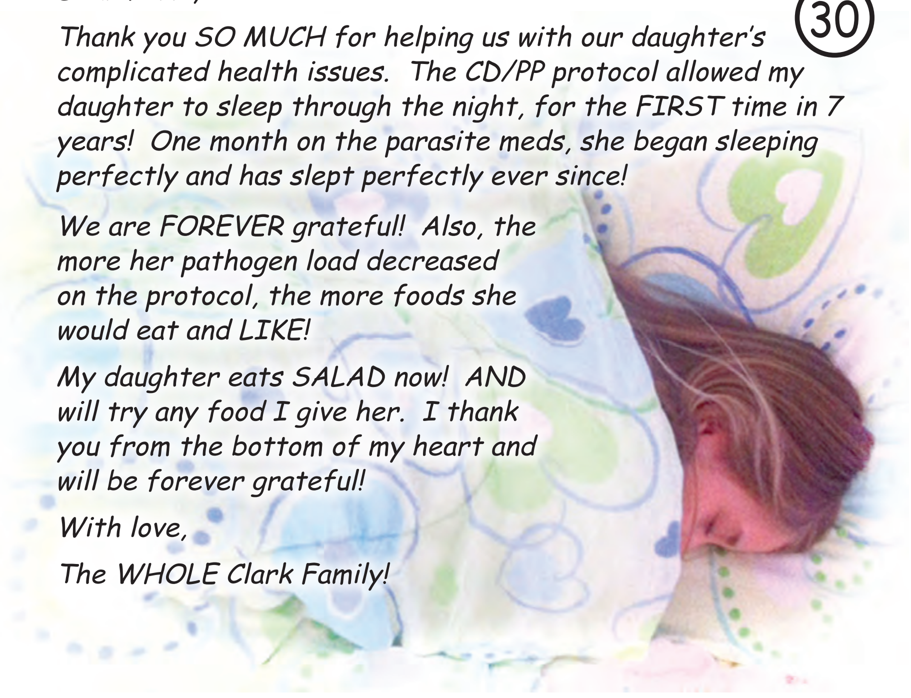

We are FOREVER grateful! Also, the more her pathogen load decreased on the protocol, the more foods she would eat and LIKE!

My daughter eats SALAD now! AND will try any food I give her. I thank you from the bottom of my heart and will be forever grateful!

With love,

The WHOLE Clark Family!

㉛

I was cured,
Yes we can.

I love you
very much
Godmother
Kerri!

㉜

My son had a great week - yesterday we went bowling as a family. It's difficult with him to notice gains sometimes because he is really close - his ATEC was a 7 last time I did it and the changes in him may be subtle and harder to notice. While we were bowling he was calm (despite the loud music, lights and commotion) he sat in between his turn watching everyone, (and the music videos playing) knew when it was his turn and cheered for me (I was beating my husband at the time - lol). He carried his ball down the lane and swung it with one arm - omg!! It was the most NT I have seem him! He now is a brown belt in Tae Kwon Do and this was the first time he was able to swing the ball - I used to have to carry the ball to the ramp, help him lift it and count to 3 for him to push the ball. He was diagnosed hypotonic at 3.5 years, MMS and PP are what helped him get some strength, energy and endurance. I increased his MMS by a drop about 2 weeks ago and we are also ramping up on GcMaf.

㉝ *Great FUA moment today. My son took his first ever lesson.... swim lessons. He did amazing!!! Followed all instructions, waited in line, smiled, talked, and let the instructors help him. Amazing.... I cried almost the entire time. At one point he smiled, waved, and said "hi mommy."*

㉞

Thank you Kerri! Because we followed the protocol with obviously great results.

㉟

My son has continued to improve on his handwriting and his anger has melted away. We had an IEP today and we lost 2.5 hours of services because he has done so well he does not need those services. They were surprised at how well he is doing. And he fed his little brother soup tonight to get him to eat his dinner.

Thank you Kerri for
helping us to recover our son!

㊱

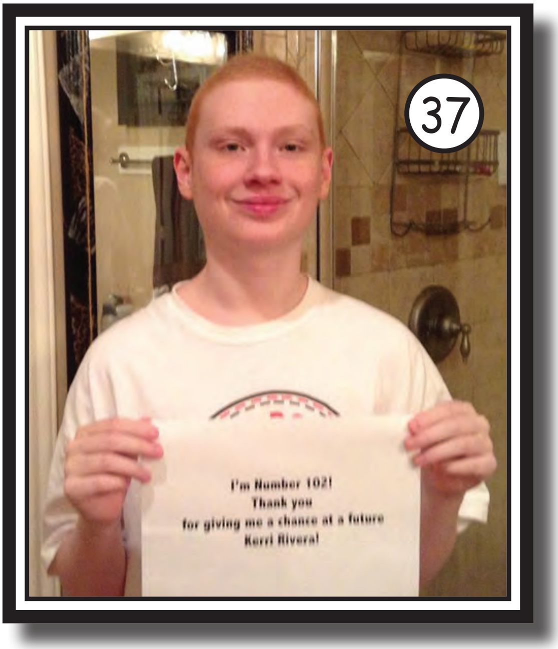
I'm Number 102!
Thank you
for giving me a chance at a future
Kerri Rivera!

37

I'm Number 102!
Thank you for giving
me a chance at a
future
Kerri Rivera!

On any given day, I probably rattle out the words I LOVE YOU, directed at my boys at least a dozen times. My NT boy almost always will reply back. My ASD boy usually says, YES YOU DO as his answer. And I have completely accepted that that is his way of receiving my I love yous.

On very rare occasion he will say I love you to me...

Well, by now you know where this is going.

This morning I was walking past him and he said to me.... I LOVE YOU!

I wasn't at first sure what he said as he didn't shout it, So I asked him, what was that?

He repeated, I love you.

I said thank you son, I love you too!!!

Wait for it...

Bammity bam bam BAM!!!

Celebrating every little big thing. Appreciating my 3 favorite words. Loving every minute of every day. Even loving those big red lips this morning knowing he is detoxing...

38

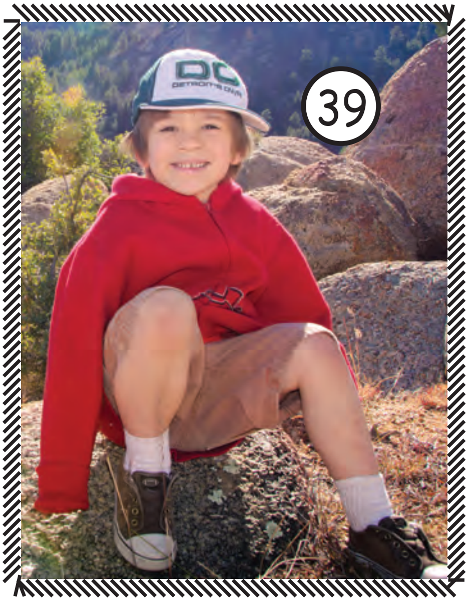

Our son was quickly slipping away from us. We were lost. This protocol gave us direction and ultimately brought our son back to us. He looks at us with smiling eyes and tells us he loves us. He has a full childhood now. Words could never convey the magnitude of our gratitude for Kerri and all who support her efforts to make a future for our kids. See Gunnar's full testimonial on page 378.

Gracias Madrina Kerri!! I now have an ATEC of 10!

㊶

We started the DAN protocol when our 10 year old daughter was 18 months old and remained on it for 9 years. Before starting CD, our daughter appeared neurotypical, but she was heavily dependent on supplements and her progress had stagnated for a few years. She had a distended belly, poor growth, severe constipation, anxiety, a touch of OCD, and focus/concentration/fogginess issues.

Her ATEC before starting CD was 24. Just 26 days later, her ATEC is 7. A 17 point drop in 26 days! We haven't even had time to do a parasite protocol or a 72/2 weekend yet! Her belly is much flatter and her anxiety, minor OCD, and focus issues are greatly improved! We're going to keep her on CD and I plan on starting the protocol myself this week. As a friend said, "DAN got her to third base and CD brought her home!"

㊷

My son just saw me reading this page, and asked what CD Autism was. I told him it was the medicine he's taking. He said, "I think my autism is gone. How do you know someone has it still?" Soon, baby, soon!

Spiro and Peter (9 year old twins)

After years of doing many interventions for our twin boys, and having moderate success at best, we learned about CD. It took about a year before we felt comfortable enough to try it. I can honestly say that, it is the BEST intervention that we've used. We started CDS in mid July 2013. In the 4 short months that we've been using CDS and now CD, we have had more gains in every area, than in the last 6 years of doing so many different treatments. What I love about this protocol, is the simplicity of it, less is more, and that was a huge change in thinking for me.

Just to give you an idea on how well our boys are doing, here are their ATEC scores so far:

Spiro: He was number 100 to recover on the "Healing Train" and the 1st child to recover using CDS.

ATEC prior to starting CDS - 22.

ATEC one month after starting CDS - 16.

ATEC two months after starting CDS - 9.

Peter: He was pre verbal prior to CDS and now uses single words and the odd 2 word combination to communicate. Well on his way to healing.

ATEC prior to starting CDS - 69.

ATEC one month after starting CDS - 55.

ATEC two months after starting CDS - 46.

We finally know, deep in our hearts, that our boys are well on their way towards "true healing". We don't have that stress and anxiety about what the future holds anymore. We finally feel like we can breathe easier, and that day by day, the gains keep coming consistently. The gains stay, and that feels better than anything.

Thank you, to my FB friends who introduced me to MMS and to Kerri Rivera who we have such admiration and respect for. I thank GOD everyday that we were fortunate enough to find out about this treatment. Our boys are coming back to us and their beautiful personalities are emerging. We now look to the future with such HOPE. Kerri, we will forever be grateful to you. xoxoxoxoxoxoxoxoxoxo

㊸

Tim Hortons

㊹

Whoa... My severe kid walked in the door and wrapped his arms around my neck and gave me a normal hug for the first time in his life (6.5 years). Umm... This is crazy. He wiped his butt for the first time EVER... no prompting, no asking! I wasn't even in the room with him! When I got excited and asked, "Did you wipe???" I swear to GOD... I heard "Yeah, Mom!" Kid's never said a word ever. Could be dreaming, but WOW!!! What a day!!!

Please share your experiences!

Testimonials are one of the best ways to share your experiences with this protocol. Perhaps you learned about it by reading or watching a video testimonial?

If you don't tell us your experiences, we can't share them or take action on issues that need improving or correcting...

Send your testimonials to:

testimonials@cdautism.org

Also, let us know if we are free to publish your testimonial, with or without your name.

More miracles & testimonials starting on page 358.

Chapter 3

Step 1 – The Diet

"Let food be thy medicine and medicine be thy food."
~ Hippocrates

If I could choose the one part of *The Protocol* that would be the single most important piece of the recovery puzzle it would be—*The Diet*. By *The Diet,* I mean the dietary plan that I recommend to all of the families that want to start *The Protocol*. It is a combination of the classic gluten-free, casein-free, and soy-free diet along with the elimination of sugar, corn syrup, coloring, preservatives, and other harmful foods. *The Diet* is the basis for the rest of The Protocol; similar to laying the foundation of a house that the rest of the structure will rest on. Adhering to *The Diet* is critical to the effectiveness of the rest of *The Protocol*.

When a family of a child with autism comes to see me, the first thing they want to know is:

What can I do to help my child?

I always, always start with *The Diet*. In fact, I send them away after that first meeting with hope that they will be able to recover their child, but only if they commit to *The Diet* 100%. I explain to the families that they must think about food the same way their great grandparents thought about food. In generations past, food came directly from the Earth, with little processing. Fruits, vegetables, nuts, and meats were dietary staples for our great grandparents and they should be for our children as well. We must think whole foods and not ~~processed foods~~. There is no point in going from regular junk food to gluten-free/casein-free junk food!

After my first meeting with a family, they leave with the list of permitted foods and I have them email me when they have a week straight with zero exceptions or "errors" in *The Diet*. That is where the parents who are truly hungry for recovery are separated from those who are interested in having someone else "fix" their child for them.

One of two things generally happens during that first week. In the first scenario, I get an email from an ecstatic mother or father saying, "I can barely believe it, Johnny slept through the night for the first time in years," or "Johnny had a normal bowel movement," or "Johnny said two new words yesterday!" That's what I hope for.

The second scenario is... I never hear from the family again. There may have been tantrums, or an adjustment period related to *The Diet* that proved too much for them, and they decided to pursue another avenue. That's not to say that everyone sees a miracle, or that there's no middle ground. Some of the results are less obvious, such as more eye contact or less redness in the face, but generally speaking, we see positive changes. Any change is a good sign. *The Diet* is only the first piece of the puzzle. We must continue from here, layering in interventions until we get the desired, end result.

> Kerri... Bless you! When my son was 3 years old he was prescribed Ritalin®, Risperdol® and Clonazepam... can you believe it? Obviously I never gave them to him, it hurt me to see my child "drugged." 2 days after starting the diet he slept through the night, the sparkle is back in his eyes... I can't wait to start with MMS, I am so happy!!!

Throughout this book "miracles," which are actual emails and forum posts from parents about their children's improvements, have been set in text boxes (like the one above) and placed in the chapters that they accent. Please note that as many of these miracles were collected some time ago the contributors have used the term MMS instead of CD. Due to time constraints it was impossible to request permission from these contributors to change the acronym of MMS to CD in their testimonials. MMS and CD are the same substance. Where the acronym MMS is used please know that chlorine dioxide is the substance being discussed and is responsible for the healing.

Author's note: I'm not against all prescription drugs, especially those necessary to bring about healing. However, I don't condone medicating a child with drugs, masking the symptoms known as autism. So, when I see an email like this, I know we've got a great chance to heal that child because we have a parent who has committed herself to healing her child, and who is excited to see what changes the next tool will bring.

Your child's doctor may not have heard of *The Diet* or may be misinformed of its benefits. The references section at the end of the book lists several studies and articles discussing dietary intervention for ASD's. You can use them as a jumping off point for your own study. If you are consulting with a doctor, it is important to choose someone who is familiar with autism and the recovery of autism.

Many doctors do not have time to study up on what's new in autism recovery and if they have zero recoveries then they are in dire need of overhauling their protocols. Another major problem in the mainstream medical community is gigantic egos. If what we are doing is not working to recover our children, then we need to look at what we are doing and why we are doing it. Let's check our egos, update our protocols, and look at what has worked for families who have recovered their children.

Researchers at the New Jersey Medical School's Autism Center found that "children with autism were more likely to have abnormal immune responses to milk, soy, and wheat than typically developing children," which is published in a chapter of Cutting-Edge Therapies for Autism 2011-2012, by Siri and Lyons.[1] In addition, interest is growing in the study of the link between autism and gastrointestinal (GI) ailments. Siri and Lyons also relate a study by the University of California, Davis Health System, where they found that children with autism born in the 1990s were more likely to have gastrointestinal problems, including constipation, diarrhea, and vomiting than children with autism who were born in the early 1980s.

If your doctor is uninformed, or tells you there is no evidence to prove that *The Diet* will help your child, do the research for yourself since only you are in charge of your child's diet. Heck, why not do *The Diet*? It doesn't cost you any money and it just might help your child heal. Whether he eats cheesy puffs or fruits and vegetables will ultimately come down to you. You are the one with the money in your pocket. The only way to know for sure if your child is going to be one of the individuals who recovers with *The Diet*, is to try. It can take gluten six months or more to be removed from the microvilli or "shag carpeting" of the small intestine. As I mentioned previously, some children have obvious changes in two to three days, but even if your child's evolution is taking a little longer than most, DO NOT GIVE UP! At the writing of this book, I have helped about 5,000 or more families of children with autism; all of the children who have lost their diagnosis, as well as those coming close to recovery, have used varied protocols depending on their symptoms. The one thing that they all—unequivocally—have in common is: ***The Diet!***

In my opinion, it makes little difference what other interventions you apply to your child if you can't manage *The Diet* 24/7/365. Hyperbarics, chelation, ABA, etc., simply will not have the desired effects if you are still feeding "drugs" (aka garbage foods) to your child.

Why do I say drugs? Because that is what gluten and casein become in the bodies of our children on the spectrum; more specifically, gluteomorphin (also called gliadorphin) and casomorphin, which are similar to morphine. Gluteomorphin and casomorphin are produced in the gut due to improper digestion of peptides (as we will explain more in detail later). In a person with "leaky gut syndrome" (increased intestinal permeability), they are able to leave the intestine and cross the blood-brain barrier where they act exactly like morphine or heroin. Would you purposely give your child street drugs? NO!

Once we have this information and understand the severity of this issue, we have a responsibility to our children to do better. We must take away the foods that are keeping them ill. Invite your family and your child's school to help heal your child's ailing body. Explain that they can no longer give your child these items, and that if they do, it's like giving them a dose of morphine. If it sounds drastic... that's because it is!

Researchers have found an abnormal amount of these undigested peptides (gluteomorphin/casomorphin) in the urine of children with autism, proving their existence in the body. Among others, Dr. Knivsberg and colleagues in Norway have found that urine samples from people with autism, PDD, celiac disease, and schizophrenia contained high amounts of the casomorphin peptide.[2] Similarly, Gliadorphin (gluteomorphin) has been verified by mass spectrometry techniques to be present in unusual quantities in urine samples of children with autism.[3]

Do *The Diet*!

There are still plenty of food options that are permitted. I promise you—your child won't starve! We have included some recipes in Appendix 15 on page 513 to get you started. *The Diet* is absolutely the most important piece of the puzzle. If we can't remove what is directly linked to brain and gut inflammation, as well as immune-allergic reactions to offending foods, it is nearly impossible to heal a child on the spectrum. I have personally never seen a family recover a child without dietary intervention. That's not to say it hasn't happened, but I have never seen or heard of it.

Like I said before, when we started with my son Patrick, his only legal food was homemade French fries. So that's what he ate. Little by little he began to accept more foods. I assure you, the adjustment period will fade and they do eat. Stop the cycle of inflammation and addiction. Only then will your child begin to heal, and once your child starts to feel better he will accept more foods.

In the case of my own son, his acidic diarrhea and sleepless nights stopped the week we started *The Diet*. From that moment on I was hooked, not only on *The Diet*, but also on biomedical treatments for curing autism. I have never looked back. I strongly encourage you to experience *The Diet* for your child, or even as a family. It is an amazing thing.

> I thought the following might encourage some of you who have recently made big dietary changes for your kids and are struggling with all that involves. I put my son on a GF/CF/SF... well, basically, EVERYTHING-free diet when he was 11 months old, following a major vaccine-induced crash. My choice to put him on a diet was criticized by nearly everyone in my life, beginning with his pediatrician who told me I was "just imagining" that he'd regressed and that he "probably wasn't allergic" to what I had taken out; she called a week later with the test results and an apology, and a long list of other things he was allergic to! He's now just a little over seven and doing GREAT. I just delivered a GF/CF muffin to him, and he said, "You know what I love best about you, Mommy? I love that you are so nice to us, and that you always make sure we have food that doesn't make us sick." So, don't feel bad about not giving your kid that ice cream cone when Mr. Softee comes by or not loading their Easter basket up with food dye this weekend. It may take years, but they will see that you loved them enough to NOT give them these things.

Some other tips include the following:

Review Your Child's Supplement List

Vitam
A
Vitamin B
Folic acid
Vitamin
C
Vitamin
D
Vitamin
E
Vitamin
K
Minerals

Is your child taking 30+ supplements a day? We will talk more about this throughout the book, but this protocol is geared towards eliminating excesses and not about supplementing deficiencies. It's important to carefully review your child's supplements. In order to maximize the benefits from this protocol consider removing any antioxidants, (they kill CD—more on that in Chapter 5, page 88), calcium, magnesium (which feeds the biofilm), as well as iron and B_{12} (which are favorite foods of parasites). Supplements for increasing speech, reducing seizures, enzymes, and probiotics obviously have their place in individual protocols. B_{12}, while known to feed parasites, has also been known to greatly boost speech in some children. If your child saw benefits in speech from B_{12}, it may be worth keeping in. As always, these are decisions that each family has to make for their child on an individual basis.

If a certain supplement proved beneficial to your child, then don't remove it (the exception being antioxidants and the other aforementioned supplements). The goal of this step is to remove unnecessary supplements as they are most likely feeding parasites and creating "excess noise" in the body. Anything that is not serving a purpose needs to be eliminated.

Keep a Journal

Keep a record of the foods you have removed and those you have added in. Then, take note of the types of reactions you observe: rash, more or less hyper, more or less stimmy (self-stimulatory behavior), sleep patterns, frequency and consistency of bowel movements, tantrums, acceptance of new foods, eye contact, language, etc. Write everything down so that you can follow any relationships or patterns that will allow you to identify when something is working, or when it is not. Record these observations in a notebook because it helps guide us, especially those of us who are stressed, frazzled, and sleep deprived.

Don't Stress About Raw Foods or Perfectly Balanced Meals…Yet!

When you are first transitioning, feed your child whatever legal foods he/she will eat. We are taking baby steps. Once your child is on *The Diet* 100%, you can start adding in new foods, bite by bite, if necessary.

Do the ATEC

Keep track of your progress with the ATEC checklist (see page 447), which you can find at...

www.autism.com/index.php/ind_atec

This is an excellent way to see how you are doing. Every time that you begin a new intervention, it's a good idea to do the ATEC, and then repeat it every three months or so to see how your child is evolving. Sometimes our children are recovering right before our very eyes and we don't even know it. The ATEC can be used to measure the effects of all new interventions, not just *The Diet*.

Using the ATEC is a great way to measure improvement. Many of us are tired, burned out, or don't have the best memory. A formal questionnaire can help us to discern when an intervention is working or not, and it only takes about seven minutes to fill out. See Appendix 4, page 447 for a sample survey and more information about ATEC.

Always Read Labels

Read the label on the back of all packages—not just the one on the front that says "gluten-free." Many times they have sugar, yeast, carrageenan, or other items that are not allowed. We must know what we are putting into our children's bodies. If you can't pronounce it... you don't want to put it into your child. Be careful of hidden offenders like malt, natural flavors, artificial flavors, whey, and numbers (red 40, E-441), etc.

Have a Support System

It doesn't matter whether it is a friend, family member, rescue angel, or neighbor. Have a shoulder to lean on. Autism recovery is a marathon not a sprint, and no one should have to go it alone. There are so many amazing parents out there who have walked this path and are willing to help newbies! As of this writing we have several public forums open, including...

www.cdautism.org

and

www.facebook.com/groups/AutismCD

This is an excellent resource for finding the latest info, protocol changes, and share hope and/or frustration with parents all over the world who are walking the same path. Don't forget to sign up for our newsletter as well by visiting the home page at www.cdautism.org!

Take it One Day at a Time

One day
at a
time

Defeatist thinking will only harm your resolve to help your child "I can't do this for the rest of Johnny's life!" or "How will I get through this year?" are defeatist thoughts. Go hour-by-hour or minute-by-minute, and know there are victories every day in the world of autism. We must draw strength from those victories, even when they aren't our own. Both this book and my website have a section of miracles and testimonials (see Chapter 2, page 19 and Appendix 1, page 357), which are real correspondence I have received from parents describing the advances their children have made as a result of doing *The Protocol.* We receive these miracle emails and forum posts every day, but we had to pick and choose which ones to share in the book. You will also see them sprinkled throughout the chapters of this book. Take some time to read them, know that children are recovering every day, and believe your child can be next. If something doesn't go as planned, wake up the next day and know that, *THIS IS the first day of autism recovery.*

Why Does My Child on the Spectrum Need to Go on a Gluten-Free/Casein-Free/Soy-Free Diet?

As I said before I have never seen a child recover without *The Diet.* The following information comes from *www.gfcfdiet.com* and explains why *The Diet* is so important for people with autism.

Scientific studies have shown the presence of high levels of peptides derived from casein and gluten proteins.[4,5] The digestion process is considered "normal" in terms of protein digestion for most people, as they make smaller particles called peptides that are further broken down into amino acids. However, in people diagnosed with autism it is more difficult to properly digest some of these proteins, thus allowing them to enter the blood directly as peptides. This often happens due to a lack of enzymes that help with the optimal assimilation of food and can be a factor in causing exaggerated bowel permeability (aka, leaky gut syndrome). This results in poor digestion, which facilitates the entry of these harmful proteins directly into the bloodstream, where they can cross the blood-brain barrier.

Leaky gut syndrome has been attributed to many causes including viruses; Candida; gluten, which produces zonulin, a protein thought to cause leaky gut; amongst other things.[6]

Gluten-containing foods may destroy the digestive system if they are consumed excessively or are introduced too early into a child's diet. Wheat is hybridized through artificial processing, resulting in inadequately prepared grains. Casein-containing foods may also destroy the digestive system because they are pasteurized and/or homogenized. These processes may result in damage to the enzymes that break down gluten or casein, thus causing incomplete digestion of these proteins.

Recently, Dr. Andreas Kalcker introduced us to the theory that the inability to properly digest these proteins may stem directly from parasites in the intestinal tract. These parasites may contribute to a leaky gut and thereby play a role in the development of allergies towards gluten, casein, soy, etc.

Gluten is found primarily in wheat, rye, barley, oats, spelt, malt, most breads, cakes, muffins, cereals, flour tortillas, pizzas, pastries, and donuts, etc. Gluten is also found in food starches, semolina, couscous, malt, some vinegars, soy sauce, teriyaki sauce, flavorings, artificial colors, and hydrolyzed vegetable proteins. Casein is contained in milk from cows, sheep, goats and any or their derivatives such as yogurt, butter, ice cream, or cheese. No form of cow's milk may be consumed as it causes inflammation and mucous. Even if the product claims to be lactose free, cream free, or casein free—it is not allowed.

Improperly digested gluten and casein fragments can both enter the bloodstream and cross the blood-brain barrier. Because of their opioid properties, these peptides can react with opiate receptors in the brain to cause effects similar to those of an opiate drug such as heroin or morphine.[7] These opiates are called gluteomorphin (or gliadorphin) and casomorphin, and can react with some parts of the brain, for example, the temporal lobes, which are actively involved in the process of the integration of language and hearing. Interestingly, these are two of the areas most affected by autism.

Besides their effects on the brain, opioid-forbidden foods cause inflammation in the gut and brain. When a child has an allergy to a food—in this case gluten, casein, and/or soy, etc.—as it enters the body the immune system sees it as an invader and reacts by trying to protect the body from the substance.

The first time the body is exposed to a food allergen, the immune system makes antibody-specific immunoglobulin E (IgE) against the allergen. IgE antibodies circulate in the blood and adhere to types of immune cells called mast cells and basophils. Mast cells are found in all body tissues, especially in the nose, throat, lungs, skin, and gastrointestinal tract (GI). Basophils are found in the blood and in tissues that have been swollen due to an allergic reaction.

The next time the body is exposed to the same food allergens, the allergen binds to IgE antibodies that bind to mast cells and basophils. These connections between allergens and antibodies direct the cells to release large amounts of chemicals, one of which are histamines. After the liberation of a histamine by an activated mast cell, the permeability of the vessels near the site increases. Therefore, the blood fluids (including leukocytes, which are also involved in immune response) enter the area causing inflammation. Histamine release also causes the release of cytokines and inflammatory mediators by leukocytes. These chemicals, in turn increase the inflammatory response.

Welcome to The Diet!

Now that we've covered the science behind *The Diet*, it's time to get started. First, make sure you complete an ATEC and save the results so you have a baseline score. Next, prepare a shopping list based on the following allowed items.

Author's Note: Organic products are better but not required.

Permitted Foods List

Proteins:

- Beef
- Chicken
- Eggs
- Fish (small not large size)
- Pork
- Turkey
- No processed meats or cold cuts (hot dogs, bologna, etc.)
- No shellfish (preferably—full of toxins)

Fruits:

- Most fresh fruits are permitted (except citrus, mango, pineapple, kiwi and limit berries).
- Frozen fruit without added cream or sugar.
- NO canned fruit (nothing canned ever).
- Be careful of dried fruit as it may contain sugar.

Author's Notes:

Juice is not permitted on this protocol as it has proven time and time again to annul CD and cost families precious time and money. The amount of fruit necessary to make a glass of juice also causes it to be high in sugars, albeit natural, and they can still affect the immune system. Children on this protocol need to drink water and generally speaking most former juice drinkers transition from drinking flavored beverages to water without much issue. If your child must drink a flavored beverage it would be acceptable to blend and strain or juice one apple, for example, and add it to some water and if you choose you can sweeten it with stevia (SweetLeaf® or KAL® brands are ok).

Fruit should not be consumed after a meal as a dessert due to its rapid digestion. If it is eaten after other more slowly digested foods (i.e. meat, grains, etc.), it can ferment in the stomach, causing bloating, gas, or discomfort. Fruit is best eaten before a meal or separate from meals.

Vegetables:

- All vegetables!!!
- Including French fries, however, not frozen fries or fries from fast food chains; these are often coated in flour.

Author's Note: I don't condone the prolonged use of potatoes, as they convert to sugar in the body. However, it was the only "legal" food my son would eat when we began the GF/CF/SF diet.

Nuts:

- Almonds
- Cashews
- Coconut
- Hazelnuts
- Walnuts
- Plus, they all make great milks!

Grains:

- Amaranth
- Buckwheat
- Corn
- Millet
- Quinoa
- Rice
- Sorghum
- Tapioca
- Xanthan gum

Author's Note: I prefer a grain free diet for autism recovery, especially if your child suffers from "grain brain" (addiction to grains, undesireable behaviors and/or inability to concentrate after eating grains). If your child seems to be addicted to carbs, i.e. fruits and starches, you may want to consider the Ketogenic diet or the Rosedale diet.

Even though these grains are allowed, they can be difficult to break down, and can easily ferment in the intestinal tract due to excessive Candida overgrowth, bacteria, and parasites.

Beans:

- All beans—EXCEPT soy
- Garbanzo
- Lentils
- Navy
- Peanuts

Sweeteners:

- Stevia (This is the best of all the sweeteners, but make sure it does not contain erythritol—a sugar alcohol!)
- Agave syrup
- Honey
- Maple syrup (natural/real—not made from corn syrup)
- Xylitol
- NO piloncillo (unrefined sugar)
- NO sugar

Author's Note: Honey is allowed, however, it can create insulin spikes.

Prohibited Foods List

After working with thousands of families, and conducting independent research, it is my personal opinion that the products on the following list should be avoided if your intention is to heal autism:

- Acetic acid (E260)
- Artificial flavoring
- Artificial sweeteners
- Bouillon cubes
- Bread
- Cacao/Cocoa
- Candy
- Cane sugar
- Carrageenan
- Catsup
- Chocolate milk
- Coloring
- Corn flakes
- Corn syrup
- Cow's milk in any form (even lactose-free milk products)
- Flour tortillas
- Gelatin
- Malt
- Margarine

- Mayonnaise
- Microwave popcorn
- MSG (Monosodium Glutamate)
  (MSG goes by many names—too numerous to mention here. A detailed list is available at www.truthinlabeling.org/hiddensources.html)
- Natural flavoring
- Noodle soup
- Oatmeal (except for Bob's Red Mill GF oats)
- Pasta
- Children´s nutritional shakes
- Play-Doh™
  (Contains gluten—Gluten-Free Dough is available at www.discountschoolsupply.com)
- Preservatives
- Processed meats (hotdogs, ham, sausage, cold cuts)
- Sodas
- Soy/fruit beverages
- Soy milk
- Soy sauce
- Sports drinks
- Sugar
- Yeast

NO Cow's Milk:

- Not casein free
- Not lactose free
- Not organic
- Not raw
- Not evaporated
- No No No cow's milk!
- Sorry, no goat's milk either! It too has casein.
- All cow's milk, regardless of what is removed from it, causes mucous production in the body. This, in turn, provides an ideal environment for pathogens.

If you're thinking that once you go GF/CF/SF + that your child will starve, trust me they won't! The list of foods they eat may get shorter, but they will keep eating. As I mentioned earlier, my son Patrick ate homemade French fries for the first three weeks of *The Diet* until he began accepting different fruits, nuts, and chicken again, but we did it... we transitioned. What I have seen in my own son, and in other children, is that when they begin to get healthier in the weeks and months after starting *The Diet*, they become more open to accepting new "legal" foods.

Wow… quite amazing Kerri…

I have always believed that "you are what you eat" and have been very careful with what I give to my kids, so their diet has been quite healthy, except for the occasional ice cream, etc. My son even spits out candies when someone offers him one because he really doesn't like them. He is not into sweets at all, thank God.

But I must say this diet is just what he needed! Yay!

There are a few loopholes to the above list. If you make your own catsup, mayonnaise, etc. with items from the permitted foods list, then your child can obviously eat those foods. The list refers to those items that come from the grocery store that contain white sugar, preservatives, and other problematic ingredients.

There are many items on the forbidden foods list that don't contain gluten, casein, or soy. Here is the explanation behind a few of those items:

White refined sugar: Refined sugar has been found to reduce immune system function as well as contribute directly to obesity and Type II diabetes.

The following list is excerpted from *Suicide by Sugar.*[8]

- Sugar can suppress your immune system.
- Sugar upsets the mineral relationships in the body.
- Sugar can cause juvenile delinquency in children.
- Sugar eaten during pregnancy and lactation can influence muscle force production in offspring, which can affect an individual's ability to exercise.
- Sugar can cause hyperactivity, anxiety, inability to concentrate and crankiness in children.

MSG: An excitotoxin that can literally "excite neurons to death." We need all of our neurons!

Yeast: Feeds Candida and other fungi.

Carrageenan: Contributes to inflammation in the body and studies have linked it to colon cancer in rats.[9]

Artificial sweeteners containing sucralose: In making sucralose, the chlorine in the sweetener bonds to carbon, producing a chemical known as a chlorocarbon. "According to physician and biochemist, Dr. James Bowen, chlorocarbons are never nutritionally compatible with our metabolic processes and are wholly incompatible with normal human metabolic functioning."[10]

Processed Meats (hot dogs, ham, cold cuts, and sausages): The fats found in these meats may contain high amounts of toxins such as heavy metals, pesticides, and herbicides. In addition, processed meats contain sodium nitrite, which can harm the liver and pancreas. They may also contain corn syrup and flavoring.

Natural Flavoring: The exact definition of natural flavorings & flavors from Title 21, Section 101, part 22 of the Code of Federal Regulations is as follows:

> *The term natural flavor or natural flavoring means the essential oil, oleoresin, essence or extractive, protein hydrolysate, distillate, or any product of roasting, heating or enzymolysis, which contains the flavoring constituents derived from a spice, fruit or fruit juice, vegetable or vegetable juice, edible yeast, herb, bark, bud, root, leaf or similar plant material, meat, seafood, poultry, eggs, dairy products, or fermentation products thereof, whose significant function in food is flavoring rather than nutritional.*[11]

Basically if you start with a natural ingredient, you can process or manipulate it any way you choose. No matter how many chemicals or solvents are added it will be labeled as a "natural flavor."

Notice they can come from meat, seafood, dairy, and wheat, etc., can also contain MSG.

Coloring: Food dyes have been linked to allergic reactions, hyperactivity in children, and even cancer. Red No. 2, for example, was banned in 1976 after it was suspected to be carcinogenic. Red No. 40, in a number of tests, was shown to damage DNA in mice.

When we are dealing with children on the spectrum, their immune systems are already compromised, and their detox pathways can be blocked or impaired. The idea is to decrease their burden with life-giving, nutritious foods, rather than add more stress to a body that is already maxed out.

Corn Syrup/High Fructose Corn Syrup: Simply because corn is a permitted food does not mean corn syrup/HFCS are "legal." Both of these shut down the immune system, as refined sugar does. They are chemically similar in composition. There is much controversy surrounding the use of mercury in the refining process, and how much if any, is in the final product.[12]

I wanted to update you on everything with my daughter.

We had a mishap 2 days ago. A substitute teacher gave my daughter a regular granola bar and a flavored beverage containing sucralose. You know what Kerri, when she came home from school that day she was in her own world and was ignoring me. I knew right away she had something she shouldn't have had. I was angry and saddened by what I was witnessing. Is there anything we should do since she had those 2 items that she's not supposed to have. Or just continue to move forward??

After a while on the diet it becomes obvious when an infraction happens, if you are present when it happens you can give an enzyme to help break down the food, if not, we move forward and learn how to prevent it from happening in the future.

Common Errors:

"It's no big deal if we break the diet every once in a while."

Not True! Every time you break the diet, further inflammation is caused in the brain and gut. When gluten and casein proteins are not properly broken down, the resulting peptides reach the brain as gluteomorphin and casomorphin. It takes three days for casein to be eliminated by the body, but months for gluten to be eliminated by the body.

"My child can drink milk as long as it doesn't have casein."

Not True! As long as the milk comes from a cow, your child cannot drink it. It doesn't matter how it's labeled. If it comes from a cow it's off limits. Cow's milk can provoke the body to produce mucous, thus providing an ideal environment for pathogens which can cause chronic inflammation.

"The allergy panel says my child isn't allergic to gluten or casein, therefore he can eat them."

Not True! If your child has autism, or is on the spectrum, he must avoid gluten, casein, and soy. You should also observe your child carefully after adding a new food, or a food he hasn't eaten in a while. In one particular case, even though the child didn't test positive for an orange allergy he continued to produce symptoms of an allergic reaction whenever he ate one. We must remember that the body is changing constantly and that any test is only good for a couple of months, if at all.

I have seen the same story repeated over and over again with mangoes, oranges, bananas, apples, and corn, etc. Observing the conduct of your child after consuming a food you suspect he may have an allergy/intolerance to is the best way to measure whether the food is acceptable or not. As our children move towards healing, foods that at one point produced allergy-type symptoms can be tolerated without an immune system response.

"My doctor says that autism has no cure, and that the diet doesn't work."

Not True! Run from any doctor who says he will take your money even though he feels that autism is incurable. The first question you need to ask a doctor is:

"How many children have your recovered from autism?"

If the answer is zero, keep looking!

The Diet is free—no one earns money if your child is on *The Diet*—so do it! You have nothing to lose, and everything to gain. However, if you choose not to do it, you may further lose the health or your child. Commit 100% to *The Protocol*, and your child may be one of the next recovery stories.

> My son had what can only be described as a miraculous reaction to the GF/CF/SF diet. Since starting 2 weeks ago, all the bumpy red areas on his face disappeared, tantrums went from about a 10 to a 2, and diarrhea is all but gone.

"We tried the diet, but Johnny didn't get any better."

***The Diet* is only one piece of the puzzle.** However, it is the foundation of everything we are going to do. Without maintaining *The Diet*, it is difficult to know which intervention is actually helping.

Candida will be dying and a reaction to *The Diet* may be a Herxheimer. We must go beyond the diet to heal autism, but diet is the first piece of the recovery process.

When I receive an email complaining that a child isn't getting better I ask for a detailed list of exactly what the child is eating. I always find errors in parents' application of the GF/CF/SF + diet. What I typically uncover is that *The Diet* did not fail but that the application was flawed.

"I should remove gluten and casein from my child's diet gradually."

Not True!!! Remove them from the diet immediately and watch your child improve before your very eyes. Foods that cause IgG or IgE allergic reactions cause many other problems in the body including inflammation and psychotic behaviors, in response to gluteomorphin and casomorphin. The sooner you can remove these foods the faster your child will recover.

> As soon as we started working on the diet, my son received a developmental evaluation, although he was 20 months old, the results were that of a child of 15 months. 4 months after implementing the diet, the changes have been dramatic. My boy smiles again, he speaks in his own way, he points, has eye contact, interacts with others, he is very close to being recovered, although he is still missing a little bit. He is occasionally nervous and sometimes distant, but the most important thing is that the therapist reapplied the developmental test and my son is now reacting like a child of 24 months (his current age), even in some areas surpassing it! And this is only with the diet because he hasn't taken any supplements, just a probiotic that was ineffective.

"A gluten-free/casein-free/soy-free diet is good enough."

It's Not! Like I said before, I always find errors when a parent goes step-by-step with me through their child's diet. The list on page 52 is a detailed list of what our children should not eat under any circumstances. Many of these foods contain neurotoxins/excitotoxins. These foods can negatively affect developing and mature nerve tissue. Keep them out of your child's diet.

Beyond GF/CF/SF +

Sometimes, *The Diet* as explained is not enough and we need to go beyond it. If your child suffers from constipation, diarrhea, or seizures it is recommended to do a period on the Specific Carbohydrate Diet™ (SCD™). The following text was sourced from the website...

www.breakingtheviciouscycle.info

...by Elaine Gottschall. A complete list of references is also available at that website.

The Specific Carbohydrate Diet™ has helped many thousands of people with various forms of bowel disease and other ailments vastly improve their quality of life. In many cases people consider themselves cured. It is a diet intended mainly for Crohn's disease, ulcerative colitis, celiac disease, diverticulitis, cystic fibrosis and chronic diarrhea. However it is a very healthy, balanced and safe diet that has health benefits for everyone. The foods that are allowed on the Specific Carbohydrate Diet™ are based on the chemical structure of these foods. Carbohydrates are classified by their molecular structure.

The allowed carbohydrates are monosaccharides and have a single molecule structure that allow them to be easily absorbed by the intestine wall. Complex carbohydrates which are disaccharides (double molecules) and polysaccharides (chain molecules) are not allowed. Complex carbohydrates that are not easily digested feed harmful bacteria in our intestines causing them to overgrow producing by products and inflaming the intestine wall. The diet works by starving out these bacteria and restoring the balance of bacteria in our gut.

Autism & GI Problems

Altered intestinal permeability was found in 43% of autistic patients, but not found in any of the controls (Harvard University). Intestinal permeability, commonly called "leaky gut", means that there are larger than normal spaces present between the cells of the gut wall. When these large spaces exist in the small intestine, it allows undigested food and other toxins to enter the blood stream. When incompletely broken down foods enter the body, the immune system mounts an attack against the "foreigner" resulting in food allergies and sensitivities. The release of antibodies triggers inflammatory reactions when the foods are eaten again. The chronic inflammation lowers IgA levels. Sufficient levels of IgA are needed to protect the intestinal tract from clostridia and yeast. The decreasing IgA levels allow for even further microbe proliferation in the intestinal tract. Vitamin and mineral deficiencies are also found due to the leaky gut problem.

In a healthy intestinal tract the small intestine and stomach are not inhabited by bacteria. When the flora balance in the colon is lost, the microbes can migrate into the small intestine and stomach, which hampers digestion. The microbes compete for nutrients and their waste products overrun the intestinal tract. One of the toxins produced by yeast is actually an enzyme that allows the yeast to bore into the intestinal wall. The yeast also produce other toxins such as organic acids, which can also damage the intestinal wall.

Bacterial growth in the small intestine destroys enzymes on the intestinal cell surface, which prevents carbohydrate digestion and absorption.

The last stage of carbohydrate digestion takes place at the minute projections called microvilli. Complex carbohydrates that have been broken down by the enzymes embedded in the microvilli can be absorbed properly and enter the blood stream. But when the microvilli are damaged, the last stage of digestion cannot take place. At this point only monosaccharides can be absorbed because of their single molecule structure.

In the small intestine, the body should absorb the nutrients needed from what is eaten. But in the case of malabsorption, the undigested carbohydrates left in the small intestine cause the body to draw water into the intestinal tract. This pushes the undigested carbohydrates into the colon where the microbes can feast on it. This allows for even more proliferation of the unwanted microbes and continued increase in malabsorption problems.

Low intestinal carbohydrate digestive enzyme activity was found in 43% of patients with autism. (Horvath) Recent studies point out that ongoing carbohydrate malabsorption keeps the digestive system constantly weakened, leading to systemic disorders. Suspected carbohydrate malabsorption should be treated to ward off further damage to the body's digestive system. (GSDL)

Most intestinal microbes require carbohydrates for energy. The Specific Carbohydrate Diet™ limits the availability of carbohydrates. By depriving these microbes of their food source, they gradually decrease in number. As the number of microbes decreases so do the toxic by-products they create.

The Specific Carbohydrate Diet™ (SCD™) is intended to stop the vicious cycle of malabsorption and microbe overgrowth by removing the source of energy from the microbes. The SCD™ allows simple monosaccharides that do not need to be broken down in order to be absorbed.

By following the SCD™, malabsorption is replaced with proper absorption. Inflammation is decreased and the immune system can return to normal. Once the immune system is returned to adequate levels, it can begin to keep in the intestines microbes in proper balance.

The SCD™ allows simple carbohydrates, but prohibits complex carbohydrates. The diet is started by following an introductory diet, which consists of a limited selection of foods. After the introductory diet, the next stage of the diet allows many more foods, but requires that all fruits and vegetables be peeled, seeded and cooked in order to make them more easily digested. Raw fruits, vegetables, nuts and seeds are added to the diet later. To properly follow this diet, it is imperative to read Breaking the

> *Vicious Cycle by Elaine Gottschall. The book details the progression of allowed foods as well as providing many delicious recipes.*[13, 14]

Salicylates/Phenols

Phenol is a naturally occurring chemical found in many of the foods we eat, such as fruits and vegetables, nuts, and in bioflavonoids and cartenoids (carotene, lutein, lycopene, xanthophylls, and zeaxanthin), etc. Phenols can be found in toothpaste, hair dye, and disinfectants, etc. Many foods have phenols, and they are impossible to avoid completely. Salicylates are a subgroup of phenols, related to aspirin. There are several kinds of salicylate, which plants make as a natural pesticide to protect themselves from insects, fungi, and harmful bacteria. Foods high in natural salicylates are tomatoes, apples, peanuts, oranges, cocoa (chocolate), red grapes, coffee, all berries, and peppers, to name a few.

You may also need to consider a low salicylate and/or phenol diet, as many children on the spectrum may have issues with these items. The following was sourced from *www.scdlifestyle.com*.

> *Dr. Feingold is probably the most widely known individual to study this chemical, as he developed what is now referred to as the Feingold Diet. He started out in the 1960s as a pediatrician and allergist studying children's negative reactions to aspirin. Through his work, he found that many other dietary chemicals were causing physical and even behavioral reactions in his patients. He developed the Feingold Diet to eliminate all food additives, colorings, and salicylates.*
>
> ***Why Do People React to Them?***
>
> *In a normal body that has the correct levels of sulphates and liver enzymes, phenols and salicylates are easily metabolized. The body utilizes what it needs from the chemicals and properly disposes of the rest through the bowels. In those whose levels are not normal or in the case of "leaky gut syndrome", intolerance to this chemical family can occur rather quickly.*
>
> *Many people with gut issues such as yeast/bacteria overgrowth or digestive diseases can develop salicylate intolerance as a result of "leaky gut" syndrome. Leaky gut is a result of various digestive problems and occurs when the small intestine becomes too damaged to properly filter the size and types of food particles or chemicals that enter the bloodstream. [For more on leaky gut syndrome see http://scdlifestyle.com/2010/03/the-scd-diet-and-leaky-gut-syndrome/] When these improper particles are allowed to repeatedly enter the bloodstream the*

body tries to get rid of them by triggering an immune system response. Because phenols/salicylates are so common in most foods, a person with a leaky gut will have much higher than normal levels of these chemicals in their blood and can very quickly develop intolerance´s to these specific particles.

Why Can Phenols Affect Children on the Autism Spectrum More Than Others?

Research by Dr. Rosemary Waring at the University of Birmingham found that children on the autism Spectrum have low levels of the enzyme phenol-sulfotransferase-P (the enzyme that breaks down the phenol and amine families, also known as PST) and the substrate it uses: sulfates. Sulfates are a key tool that the body uses in the process of detoxification and break down phenols such as salicylates. Without normal levels of sulfates in the body, the sulfotransferase enzyme cannot do the task it was created to do: metabolize salicylates. So there are two problems with PST deficiency: low sulfate levels and low enzyme levels. PST deficiency alone can cause problems in children (remember phenols are normal) but factor in any intestinal damage resulting in leaky gut along with it and your child's body can easily be overwhelmed. The end result is a salicylate intolerance and the subsequent physical and behavioral reactions that come with it.

Reactions Caused by Phenols

Salicylates stimulate the central nervous system in people that react to them. This can often bring with it an emotionally extreme high followed by a very low, low. Other reactions to the phenol family can occur anywhere from immediately after consumption up to 48 hours after the consumption of the chemical, depending on the immune response. Physical reactions can include: dark circles under the eyes, red face/ears, diarrhea, headache, difficulty falling asleep at night, night waking, and in some cases excessively tired and lethargic. Behavioral symptoms of a reaction can be: hyperactivity, aggression, head banging or other self-injury, and even inappropriate laughter. Hyperactivity is more common in children's reactions, while adults generally experience symptoms similar to chronic fatigue.[15]

For additional information on leaky gut syndrome you may wish to look at (search for) *Dr. Peter Osborne*:

www.glutenfreesociety.org

If your child suffers from self-injurious behaviors limit these foods that are moderate to high in phenols:

- Almonds
- Apples
- Avocados
- Bananas
- Cacao
- Cantaloupes
- Cherries
- Cider vinegar
- Coconut oil
- Colored fruits
- Dates
- Food coloring
- Honey
- Mangos
- Mint
- Oranges
- Oregano
- Peanuts
- Pineapples
- Powdered chili pepper
- Processed meats
- Raisins
- Strawberries
- Tangerines
- Tomatoes

For more information on this you can visit:

www.scdlifestyle.com

and

www.feingold.org

* * *

She's quite the book worm and has an impressive collection which includes classic favorites such as Brown Bear, Llama Llama Red Pajama, Barnyard Banter and Goodnight Moon... But lately, she has been drawn to her Mama's new favorite read... Perhaps she is hoping that someday her story of recovery will be told too...

Gluten Syndrome

For many parents of children on the spectrum, having their children on *The Diet* is something that is second nature to them. However, as more and more people are using the protocol in this book to heal their own health ailments we have decided to include the following section on how gluten can negatively affect many more than just the autism community. You will get a very interesting look into how the immune system works and why so many sufferers have misunderstood gluten syndrome. Thank you Olive for contributing to our understanding of the subject.

Molecular Mimicry

What It Is & How It Relates to the Gluten Syndrome

by Mrs. Olive Kaiser

Who am I?

I am a married stay at home mom, blessed with a wonderful husband and seven fantastic kids. In 2003, after decades of searching, we learned about gluten reactivity through our daughter's nursing school training and eventually confirmed that we are a gluten syndrome family. Our daughter and my husband had the most obvious symptoms, but we all had manifestations and antibodies. Additionally our oldest son reacted to his MMR vaccination and probably other shots, which added high functioning ASD/ADD to the mix, and he developed type 1 diabetes at age 19. Two other sons had various shades of ADD/ADHD. My own school age vaccinations in the 1950's may have led to repeated bouts of strep throat until I reacted to a strep antibiotic injection about age 10. I developed PANDAS from that reaction (Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcus). What a struggle! Decades later it has responded somewhat to diet changes and now the CD/parasite protocol. I give thanks to God for His guidance along the way.

How did I get into this community and this health project?

We tested for gluten syndrome (we called it celiac disease back then) using standard tests recommended by celiac experts and received confusing results. Then, our daughter had a troubling experience with a gluten challenge that did not match the celiac story we'd been taught. I delved into medical literature and networked extensively with the gluten syndrome community looking for help. In that desperate discovery and prayer process I found practitioners and researchers who stepped outside the "villi damaged celiac only" box. They were able to explain why we received, in the midst of obviously gluten-induced incidents, false negative results from our celiac blood tests and villi biopsy. Those tests significantly led us astray and eventually I put up a website...

www.TheGlutenSyndrome.net

...to warn others of discrepancies we stumbled upon in the diagnostic process.

What is Molecular Mimicry?

Molecular mimicry is a recognized medical theory which explains very nicely why gluten reactions may potentially inflame and damage so many different parts of the body, leading to very different symptoms in different people. It also clarifies why gluten antibodies may cross react with other foods and infections, and why it only takes a small exposure to trigger them.

When we understand molecular mimicry we are better equipped to deal with tempting social situations. Gluten syndrome has its own rules, which do NOT make sense unless this concept is understood.

The following is a brief introduction to Appendix 5, page 449, which goes into much more detail, along with references, about the following questions:

1. How does the gluten syndrome reaction actually damage our bodies?

> Molecular mimicry. The molecular structure of gluten resembles the molecular structure of many of our body tissues. When the immune system attacks gluten it may also attack body tissues that "look like" gluten. Even if you do not read the other detailed answers, learn the details for this question on page 450.

2. Does gluten always damage the villi of the small intestine as the celiac story teaches? Many other tissues such as thyroid, pancreas, liver, joint, brain, nerves, heart, bone, blood vessel walls, etc., are involved in this disorder. Does all that other damage only arise from poor nutrient absorption from injured gut villi?

> No, according to published research, many researchers and practitioners believe the villi are not always damaged in an autoimmune gluten reaction. Where there is no villi damage, injury to other organs CANNOT be due to nutrient deficiencies caused by villi damage. Molecular mimicry provides a mechanism for direct autoimmune gluten damage to many other tissues and organs when the villi are fine, OR other tissues/organs, *including villi*, may be directly damaged through molecular mimicry.

Dr. Vojdani's abstract in his editorial *The Immunology of Gluten Sensitivity Beyond the Intestinal Tract*, supports that the villi are not always injured. To quote his editorial abstract, "Evidence has been accumulated in literature demonstrating that gluten sensitivity or celiac disease can exist even in the absence of enteropathy [gut/villi damage], but affecting many organs."

3. I don't have damaged villi, and my tTG/gliadin tests were negative, but I feel so much better gluten-free. Why?

The tests were likely false negative. That is very common. As gluten digests, it breaks into more pieces than we have tests developed to check them, and the immune system makes a separate antibody for each piece. Standard tests only check 2-3 antibodies. You may have others (Cyrex Labs tests 28 antibodies). Your villi may be fine, but you may be injured somewhere else—for example: thyroid, nerves, heart, etc.

4. Why do many gluten syndrome patients not only react to wheat, barley, and rye but also at times to other foods, particularly oats, milk, corn, soy, egg, yeast, coffee, sesame, rice, chocolate and others?

These foods "look like" gluten closely enough in their structure that the immune system may mistake them for gluten. This situation may also cause your gluten antibodies to run high after you go gluten-free. The immune system may misrecognize other foods, such as yeast, corn or milk, and others for gluten because they resemble gluten molecularly.

5. The diet seems excessively strict? Why does it take so little gluten to start a reaction?

Our perspectives are skewed. We accept that miniscule amounts of venom injected by a bee sting, or a tiny exposure to peanuts in allergic individuals can set off immediate life threatening allergic reactions. Many medications are contained in very TINY pills, but they have powerful effects in our bodies. Immune gluten reactions are also that sensitive. "Crumbs matter."

6. Why do many people react to gluten, proven by antibody tests, but they have few or no warning symptoms for a long time and then they crash with something serious, usually autoimmune?

Gluten is famous for slowly injuring nerves by molecular mimicry, and in many cases, the nerves are silenced by that injury. The patient does not realize there is a problem until the tissue or organ that those nerves supply begins to fail.

7. Why do so many of us react to gluten today, when for centuries most people appeared to be fine with wheat, barley, rye and oats? After all, wheat and barley are mentioned positively in the Bible and other historical documents.

Today's gluten is altered, violently, by nuclear radiation and chemical mutation within the past 60 years*, plus our toxic and poorly nourished bodies do not have optimal digestive capabilities to break it down. Weak, toxic, leaky body barriers/membranes, particularly leaky gut, set the stage for gluten induced molecular mimicry.

*Nina Federoff, *Mendel in the Kitchen*

8. Why do specialists and researchers insist that the gluten-free diet must be life long? Can't we heal this problem and go back to our beloved wheat bagels, croissants, and brownies?

Our scientists still insist that gluten-free is a strict lifelong commitment. I agree. For me it is not worth playing with today's wheat. There is something strange and unpredictable about it. The memory B cells in the immune system never forget what the enemy "looks like", and fresh exposure retriggers antibodies.

9. Traditional peoples soaked and/or sprouted their wheat berries and then made sourdough bread with them. Does that process alter the gluten sufficiently for gluten syndrome patients to safely consume this bread, particularly spelt or einkorn?

No. These processes and ancient wheat grains do make the bread more digestible, but not gluten-free and still unsafe.

10. Should I substitute all the gluten foods I routinely eat with gluten-free substitutes?

No, not routinely. The gluten-free community finds that they are still mainly expensive high carb processed food (i.e., junk food).

11. What are gluten withdrawals?

Occasionally, gluten breaks into specific "pieces" in the gut that resemble opiate drugs. When a person goes gluten-free, they may experience temporary, but unpleasant, withdrawal symptoms for a few days as these pieces disappear from the blood stream.

12. What are the risks of formal gluten challenges?

Many patients avoid these challenges. Occasionally a patient tries the gluten-free diet for an extended period of time and then the patient or doctor decides to run tests to confirm gluten reactivity. The standard advice to restart the production of antibodies is to consume gluten products 4 x per day for 4-6 weeks, and then run the standard blood test, followed by a villi biopsy if the blood work is positive. This is called a gluten challenge and has created some very dramatically unhappy reactions, some of them neurological/psychological.

Please see Appendix 5 (page 449) and www.GlutenSyndrome.net for more info and references. When we understand molecular mimicry our understanding of the gluten syndrome comes into focus. It explains why gluten-free diets and beyond are important tools to reduce inflammation and promote healing. As time goes on, gluten-free diets are easier to manage in public, tests are better and social awareness has grown. The Just Eat Real Food movement and others play into healthy gluten-free dining with wonderful recipes that avoid processed foods and incorporate healthy fats and nutrient density. This is a happy, encouraging era as we watch our children heal and adults find better stability in the midst of a health crisis. Bon Appetit!!!

* * *

The Diet - FAQs

The following FAQ's were printed with permission from gfcfdiet.com:

Slight changes were made to match the formatting of this book; however, the content remains unaltered. Please visit gfcfdiet.com for complete list of the references contained within the following text.

1. My doctor has never heard of any of this and she is extremely skeptical. I'm embarrassed to tell her I'm considering this approach. What do you think?

Skepticism is a good thing in a medical doctor or scientist. However, since there is preliminary evidence to support this safe, non-invasive intervention, it is up to you to educate her, state your wishes, and ask for her support. For a doctor, it is better to wait until all of the data is published in peer-reviewed journals before advocating a treatment. However, for a parent, it is reasonable to want to help one's child without waiting for all of the results of the "double-blind placebo" studies. Because this approach does not include any unusual supplements, invasive drugs, or expensive treatments, your pediatrician should be supportive. Explain that you would like to try this for a few weeks, and agree that you will be objective about recording your child's progress while on the diet.

If you feel that you need to support your case legally with the scientific and medical documentation that is currently available, please see the medical links at www.gfcfdiet.com or at www.autismndi.com.

2. What is casomorphin?

Casomorphin (or caseomorphin) is a peptide derived from casein, a milk protein. Casein is one of the major proteins in the milk of all mammals including cows, goats, and humans. When casein is digested properly, it breaks down into large peptides like casomorphin, and should then be broken down further into smaller amino acids.

However, Dr. Reichelt in Norway, Dr. Cade at the University of Florida, and others found that urine samples from people with autism, PDD, celiac disease, and schizophrenia contained high amounts of the casomorphin peptide in the urine.[4] In its peptide form, casein has opiate properties similar to morphine, and may plug into the same opiate receptor sites in the brain. Researchers have found that these peptides may also be elevated in other disorders such as chronic fatigue, fibromyalgia, and depression based on anecdotal reports of symptom remission after exclusion of wheat and dairy.

3. What is gliadorphin?

Gliadorphin (also called alpha-gliadin or gluteomorphin) is a substance that resembles morphine. Ordinarily, this is a short-lived by-product from the digestion of gluten molecules (found in wheat, barley, rye, oats, and several other grains). Gliadorphin is very similar to casomorphin. Gliadorphin has been verified by mass spectrometry techniques to be present in unusual quantities in urine samples of children with autism, and are believed by many to be a central part of the system of causes and effects that cause autistic development.

The most probable reasons for the presence of these molecules are:

- One or more errors in the breakdown (digestion) process caused by enzyme deficiency and/or
- Abnormal permeability of the gut wall (that would allow these relatively large molecules to enter the bloodstream from the intestine in abnormal quantities).

4. I am confused about allergy vs. intolerance. I understand that our children may be sensitive to corn, soy, and other foods as well as gluten and casein. Does this mean that they will eventually start turning these foods into the morphine-like compounds too? If this were the case, would they show up as an allergy on a RAST test? Or were our children always allergic to these foods (a regular allergy

that may cause behavioral changes in our children), and we just didn't know because the gluten and casein were hiding the allergy?

To a traditional physician or allergist, "allergy" is used to describe a reaction of the IgE part of the immune system, resulting in hives, swelling, or breathing problems. However, the words "allergy" and "intolerance" are often used to describe any inappropriate reaction to foods or substances that should normally be harmless to the body.

There are at least three different ways that a child with autism may have a problem with foods that contain gluten or casein, and it's important to understand the distinction:

1. An IgE ALLERGY commonly results in skin problems, hives, swelling, and breathing problems, etc. This can be tested using a skin test or blood test.

2. An INTOLERANCE (usually mediated by the IgG or IgA part of the immune system, or by an enzyme insufficiency such as lactose intolerance) has more varied or vague symptoms like discomfort, stomach problems, sleep problems, joint pain, ear infections, or hyperactivity and behavior problems. Sensitivity to these substances can be tested with an ELISA blood test.

3. PEPTIDUREA (peptides in the urine) is caused by the inability of the body to properly break down certain proteins. It is hypothesized that certain peptides, notably from milk and wheat proteins, are plugging into the opiate receptor sites of the brain and disrupting brain and nervous system function. Urine testing for this is still experimental, and many parents believe that the best way to find out if this is what is causing a child's autism is a strict trial period on the GF/CF diet.

On the GF/CF diet, gluten and casein are avoided because they are strongly suspected of having a direct pharmacological effect. When these proteins are only partly broken down, some of the resulting fragments can be strikingly similar to morphine, and act in more or less the same manner. (This type of reaction can co-exist along with a classic type of allergy towards the same foods.)

Recent research indicates that protein from both corn and soy may also contain some molecular sequences that could, if the patient has an enzyme deficiency, be broken down into something closely resembling opioid peptides. Even spinach protein has been found to have some opioid activity.

Products made from soy or corn will also often contain metabolic end products made by microscopic organisms like bacteria, molds, or other fungi. Some of these are suspected of being harmful to a small number of people who are genetically predisposed to autism. The amount of danger will depend on individual conditions AND on the quality of the corn or soybeans used in the production process. Soy OIL (lecithin) may be worse than most other soy products since this product will look and taste okay, even when made from moldy raw material, and since it is commonly made from the "bottom grade" of the harvest.

Some people also think that one of the natural pigments in corn (lutein) may cause problems for reasons that are not properly understood (see: "Sara's Diet"). This must be regarded as highly speculative.

5. I don't think my child has allergies, or that allergies could cause autism. Why should I try removing foods from his diet?

Although parents have been reporting a connection between autism and diet for decades, there is now a growing body of research that shows that certain foods seem to be affecting the developing brains of some children and causing autistic behaviors. This is not because of allergies, but because many of these children are unable to properly break down certain proteins.

6. Milk and wheat are the only two foods my child will eat. His diet is completely comprised of milk, cheese, cereal, pasta, and bread. If I take these away, I'm afraid he'll starve.

There may be a good reason your child "self-limits" to these foods. Opiates, like opium, are highly addictive. If this "opiate excess" explanation applies to your child, then he is actually addicted to those foods that contain the offending proteins. Although it seems as if your child will starve, if you take those foods away, many parents report that after an initial "withdrawal" reaction, their children become much more willing to eat other foods. After a few weeks, most children surprise their parents by further broadening their diets.

7. Isn't milk necessary for children's health?

Americans have been raised to believe that this is true, largely due to the efforts of the American Dairy Association, and many parents seem to believe that it is their duty to feed their children as much milk as possible. However, lots of perfectly healthy children do very well without it. It is calcium children need, not milk. Cow's milk has been called "the world's most overrated nutrient" and "fit only for baby cows." There is even evidence that the cow hormone present in dairy actually blocks the absorption of calcium in humans.

Be careful. Removing dairy means all milk, butter, cheese, cream cheese, and sour cream, etc. It also includes product ingredients such as "casein" and "whey," or even words containing the word "casein." Read labels—items like bread and tuna fish often contain milk products. Even soy cheese usually contains caseinate.

For more information on dairy-free living, there's a very good book called *Raising Your Child Without Milk* by Jane Zukin. Another great book called *Don't Drink Your Milk* by Frank Oski (the late head of pediatrics at Johns Hopkins and author of *Essential Pediatrics*). This book cites the results of several research studies that conclude that milk is an inappropriate food for human children. It is available for $4.95 from Park City Press, PO Box 25, Glenwood Landing, NY 11547, ISBN #0671228048.

8. How do I know which foods he's allergic to?

Try an allergy elimination diet. For example, keep common allergens out of his diet for a few days and then re-introduce them, one-by-one. If you see symptoms, either physical or behavioral, try again in a few days. By being systematic you will be more certain of which foods are causing problems and correctly rule them out. Two excellent resources, probably available at your library, are Doris Rapp's book, *Is This Your Child*, and William Crook's, *Solving the Puzzle of Your Hard to Raise Child*.

9. Aren't eggs dairy?

Many years ago, most of us were taught that eggs and dairy were part of the same section of the food pyramid under "dairy." However, they are not. Eggs are free of any dairy. Dairy can be from cows, sheep, or goats. The eggs we eat are normally from chickens. Dairy and eggs happen to be side-by-side in the refrigerator section of the grocery store, which may add to the confusion.

10. What do I do when we go to a party or out to another person's house and are trying to maintain my child on a GF/CF diet? At times we find it impossible to catch him before he ingests something he shouldn't.

It is a good idea to bring some favorite GF/CF goodies with you. Keep them hidden until that moment of "competition." Show the surprise treat to your child before your child eats one of the gluten and/or casein laden foods. Bring more than one goodie to be on the safe side. Also, there are two small pamphlets that are great for handing out to friends and family members that give a simple explanation for your child's special dietary requirements. These are: "Alternative Treatments for Children Within the Autistic Spectrum: Effective, natural solutions for learning disorders, attention deficits, and autistic behaviors," by Deborah Golden Alecson, and "Leaky Gut Syndrome: What to do about a health threat that can cause arthritis, allergies, and a host of other illnesses," by Elizabeth Lipski, M.S., C.C.N. Both of these books cost $3.95 each and can be found at health food stores.

11. Do we need to worry about gluten-containing lotions, shampoos, and toothpaste being absorbed through the skin?

Nicotine patches, birth-control patches, and other transdermal applications of medications are proof that the skin does absorb many things (and pass them to the bloodstream). However, the molecules in gluten are too large to pass through the skin, according to John Zone, MD (a dermatologist quoted in the Spring 2003 issue of *Living Without* magazine).

Most often the problem occurs from hand to mouth (i.e. touching colored modeling dough and then touching one's face), which is a good reason to avoid giving our kids much access to anything that poses a threat. If your child's school has a sensory table, you can request that it be filled with rice or dried beans, instead of macaroni or gluten grains. Likewise, we suggest that you volunteer to be the "Play-Doh™ parent" and keep the classroom supplied with a safe version that you don't have to worry about.

We are not too concerned about shampoo, unless you have a child who is likely to try to drink the stuff. But toothpaste is an entirely different matter, since some of it is likely to be ingested, rather than spit out.

12. I'm already worried about my child's nutrition, and his "allergies" are causing me to further reduce his choices. If apple juice and bananas are the only fruits he will eat, and he's reacting to them, how is he supposed to get by?

Fruit contains water, sugar, fiber, and vitamins. If he will not eat other fruits then he needs to get these things from other sources.

13. I am wondering if anyone else is having problems with school and keeping them on the diet? I have sent a note saying: "Please don't give my son dairy, wheat, corn, or soy." Today I came in and they were giving him Popcorn!

My kid doesn't eat ANYTHING that isn't sent from home. He's almost ten now, and this policy has been good for him because succumbing to temptation outside of the house has simply never been an option. I don't think it would occur to him to take food from anybody without our permission.

We always made sure there are backup snacks and birthday party treats at school for emergencies. I think the school has always taken this seriously because they can see how seriously we take it.

We once got a note from a mom who said that her son had been strictly GF/CF for months without an improvement. I asked what he was eating, and she gave me a long list including expensive GF prepackaged products. Then she ended her letter by saying, "Of course, that's just what he gets at home. I have no idea what he gets in his school lunch, or what they're using for food reinforcers in his ABA program."

Definitely try this—you might see a real difference.

14. What percentage of children will respond to dietary intervention?

DAN! doctors used to try to be conservative about this, and say at least a third of children, and then, after seeing more patients, they said two thirds. Now that they have seen hundreds or thousands of patients, most tell us that they believe that almost every ASD child will benefit from this diet. Many will need further modifications (i.e. removing grains or sugar) before the full benefits are realized.

However, age plays a big part in how quickly results will be seen. We can probably say that the response will be dramatic in more than two thirds of the children under three, and perhaps more subtle, but still helpful, in at least two thirds of older children. We think those are pretty good odds.[3]

Some Additional FAQs...

Kerri, yesterday the therapist was playing with Play-Doh™ with my daughter. She could not resist eating it. We tried very hard to stop her with no luck. In the end we had to stop her from playing with it. People say this is due to a zinc deficiency. Do you know if this is correct? Will giving her Zinc stop her from wanting to eat Play-Doh™? If yes, how much should I give?

She is not eating Play-Doh™ because of a zinc deficiency. Play-Doh™ contains wheat. She is addicted to gluteomorphin, a by-product of her eating gluten, which is contained in Play-Doh™. So, she broke her diet. She cannot touch nor eat Play-Doh™.

What exactly is Candida?

Candida albicans is an opportunistic fungus (or form of yeast) that is the cause of many undesirable symptoms ranging from fatigue and weight gain, to joint pain and gas.

The Candida yeast is a part of the gut flora, a group of microorganisms that live in your mouth and intestine. When the Candida population starts getting out of control it weakens the intestinal wall, penetrating through into the bloodstream and releasing its toxic by-products throughout the body.

What Causes Candida Overgrowth?

Candida is an opportunistic pathogen that can rapidly take over when a person is under a course of antibiotics. Antibiotics destroy beneficial gut flora but have little effect on Candida, giving this normally harmless yeast the chance to take over dominance of the gut environment very quickly.

Babies born via C-section, or to mothers who were treated with IV antibiotics during labor are especially vulnerable to the ravages of Candida overgrowth. This is because they are not exposed to a healthy balance of gut flora on the way through the birth canal prior to the moment of birth.

…disaccharides, or double sugars, are present in many carbohydrates including ALL grains—not just gluten containing ones. An inflamed, imbalanced gut overridden with Candida is unable to digest double sugar molecules completely because the lack of beneficial gut flora has compromised the function of the enterocytes.

According to Dr. Natasha Campbell-McBride MD, author of *Gut and Psychology Syndrome*, and one of the key scientists at the forefront of gut restoration research today, enterocytes are the cells that reside on the villi of the gut wall and produce the enzyme, disaccharidase, which breaks down the disaccharide molecule into easily absorbed monosaccharide molecules. When the enterocytes are not nourished and strengthened properly by adequate beneficial flora, they become weak and diseased and may even turn cancerous. Unnourished, enterocytes cannot perform their duties of digesting and absorbing food properly.[16, 17]

The critical importance of the enterocytes to health cannot be overstated!

Weak and diseased enterocytes also have trouble digesting starch molecules, which are very large molecules that consist of hundreds of monosugars, connected in long branchlike strands. People with weak digestion due to Candida overgrowth and messed up enterocytes have a terrible time digesting these complex molecules leaving large amounts of it undigested—the perfect food for pathogenic yeasts, bacteria, and fungi like Candida to thrive upon. Even the starch that manages to get digested results in molecules of maltose, which is—you guessed it—a disaccharide! This maltose also goes undigested due to a lack of the enzyme disaccharidase and becomes additional food for Candida.

We are a few days into the diet. Can someone comment on what they send to daycare with their kids for snacks? Our daycare is nut free as well. We are ok for lunches but what should we take for snacks. The gluten-free pretzels, etc. we were taking he can't have now because of yeast and soy.

Here a few suggestions: Hard boiled eggs, cut up raw veggies and fruit, air popped popcorn, apples, bananas, dried fruits (with no sugar added). Also you can join "CD/CDS MOMS WHAT IS MY KID EATING TODAY" Facebook group. You can find more ideas there.

Chapter 4

An Introduction to Chlorine Dioxide

by Jim Humble

> *"Condemnation without investigation is the height of ignorance"*
>
> ***-Albert Einstein***

Jim Humble, the man behind chlorine dioxide for healing, was kind enough to write the following introduction to our CD section...

I just can't tell you how pleased and proud I am that Kerri has developed protocols using chlorine dioxide as the main ingredient to bring health and normalcy to many children who had the symptoms called autism. Their resulting healing is one of the greatest stories of this and the last century. Kerri's ability to look at something new and to evaluate the value of the new item without being totally affected by the dictates of present established medical science is truly rare. Kerri decided to do something since modern medicine has yet to come up with a reasonable priced globally available solution to the problem of autism that yielded good results. Medical science says that autism is incurable, that millions of children have to be left to suffer and live without reaching their potential. At this writing 115 children have recovered, hundreds more are near recovery and thousands have had documented improvements as a result of being treated with Kerri's protocol.

Here are some basic points:

Chlorine dioxide (the molecule) was discovered in 1814, with many giving credit to Sir Humphrey Davy for its creation, and today several patents exist for healing various ills with it (see list on page 85).

MMS (Master Mineral Supplement) – I rediscovered chlorine dioxide in 1996 and since then more than 1,020 people have traveled from more than 95 countries to the Dominican Republic, Mexico, and other countries, to be trained in its use. By our estimation, more than 10 million people worldwide have used chlorine dioxide.

The highest cost of a set of bottles of sodium chlorite + activator is less than $25 for 4 ounces of each. This means that almost any place in the world the average dose of chlorine dioxide is less than 4 cents, or less than $.40 a day or $25 for 2 months. In most places it is even less than that. So far there is no one in the world who can't afford it or who wouldn't be able to get it merely by asking if they can't afford it.

The chlorine dioxide molecule is the weakest of all oxidizers used in the human body and thus has very little effect on human cells. This molecule has the unique ability to recognize and oxidize (kill) harmful bacteria. Hans Christian Gram, a scientist of the 19th century, discovered that most harmful bacteria have a negative charge. He was able to dye positive and negative bacteria two different colors. His techniques are still used in laboratories and universities. All oxidizers, including chlorine dioxide, have a positive charge that will attract and kill negatively charged bacteria while repelling positively charged bacteria. Chlorine dioxide is thusly able to kill bad bacteria without destroying the good bacteria. This is simple high school science; like charges repel and unlike charges attract.[1]

Most chemicals and many foods in this society are tested to determine their poison index, which is designated by giving them a toxicity number known as LD_{50}. Sugar, table salt, butter, arsenic, cyanide, vinegar, wine, Clorox, window cleaner, and dozens of other household and industrial chemicals are rated as to their toxicity by force feeding 10 rats with the food or chemical in question until ½ of them die. The amount of chemical or food that it takes to kill 5 of the 10 rats determines the LD_{50} number. Every chemical has a toxic rating and is indicated as an LD_{50} number. All chemicals are toxic in large enough quantities and there are a lot of unlucky rats in our society.

There has never been a recorded death caused by ingestion of chlorine dioxide or sodium chlorite in humans. In the case of chlorine dioxide, it would take 1000 times more chlorine dioxide than is used in a daily treatment dose to reach a toxic dose for a human being. The following information is from the *Toxicological Profile for Chlorine Dioxide and Chlorite* (published by the U.S. Dept., of Health & Human Services) which contains LD_{50} toxic numbers for chlorine dioxide along with documentation that no deaths have been caused by chlorine dioxide (Emphasis in italics is the author's):

> *No information was located regarding death in humans following oral exposure to chlorine dioxide or chlorite. Shi and Xie (1999) indicated that an acute oral LD_{50} value (a dose expected to result in death of 50% of the dosed animals) for stable chlorine dioxide was > 10,000 mg/kg in mice.*[2]

I should say something about bleach as a number of critics insist that MMS (chlorine dioxide) is a bleach and thus is a poison. Chlorine dioxide, the chemical that is MMS, has never been used domestically as a bleach for cleaning toilets. It is an industrial bleach when used 4000 times more concentrated than MMS. It is never used for bleaching purposes in the homes of people of this world.

Legally, ethically, morally, and logically, if any item does not have the characteristics of a particular item, it cannot be said to be that kind of an item. This means that a chemical that does not have the characteristics of bleach cannot be called bleach. So a chlorine dioxide solution being 4000 times weaker than that of bleach cannot be called bleach. This is a fact accepted in courts of law of the world.

I sincerely hope you will take the time to understand the basics of chlorine dioxide as explained here. Any time one has helped someone overcome the symptoms of autism he has helped save a life, as those suffering from the symptoms of regressive autism are not living their full lives. The end result of learning this data is that the time may come when you will need chlorine dioxide to save your own life or that of a loved one.

Jim Humble

Archbishop Jim Humble

A message from Kerri to Jim Humble

"With all my love, respect and gratitude. You have swum against the current for so long, and I am so fortunate that you waited for me. I will continue to recover these generations of children that are the victims of a senseless epidemic. As you have done, I will do. We can't change the past, but the future is ours. You have brought light to where there once was none."

Chapter 5

Step 2 – Chlorine Dioxide (CD)

"All truth passes through three stages. First, it is ridiculed. Second, it is violently opposed. Third, it is accepted as being self-evident."

~ Arthur Schopenhauer

CD (chlorine dioxide) is the most amazing intervention that I have ever tried with my son. It by far packed the biggest punch in the shortest amount of time—we saw the positives in under a week with no side effects! (A Herxheimer reaction is not a side effect.) After seven days on CD, my son was looking at me in the eyes and requesting non-preferred activities using four word sentences. I have never, with any other intervention, seen a change in my son come about so quickly, especially since Patrick was your typical non-responder before CD.

The chlorine dioxide molecule (CD).

Why CD for Autism?

We know that the symptoms known as autism are caused by:

- Viruses
- Bacteria
- Candida (yeast)
- Parasites
- Heavy Metals
- Inflammation
- Food allergies

CD catapults children into recovery because it kills these exact things. As of this writing, over three years since we started using CD in our protocol to heal autism, 115 children have lost their diagnosis. Thousands of children around the world are healing, and that means their families are healing as well, and for this I am eternally grateful.

CD has confirmed my faith and has shown me time and time again that autism IS curable. My life has changed since the day my son was able to look me in the eyes, smile, and request things. I could not keep information of this magnitude a secret. I absolutely had to share it with anyone who would listen.

CD is very affordable as you can purchase a two to three month supply from multiple worldwide suppliers on the Internet for around $25USD. The cost difference between the CD Autism Protocol and a megavitamin/supplement based biomedical protocol for a family living without insurance (or in a third world nation) could be the difference between recovering their child and not recovering their child. The results are also vastly different.

You can think of CD as a shortcut to recovery because it kills all pathogens with one shot. Hypothetically, you could use antibiotics, antifungals, and antivirals to get the same results; however, you could also cause damage due to harmful side effects from these products (liver and kidney stress, etc.). Also, antibiotics, antifungals, and antivirals can only target certain bacteria, viruses, and fungi. CD targets all types of pathogens based on their electrical charge and pH. The particular strain is irrelevant.

CD causes no side effects, leaves no toxic residues, and causes no liver damage. It causes no new stress to already stressed bodies. Everything else we use to kill pathogens leaves something behind. For example, antibiotics also kill your beneficial bacteria, and antiviral use comes with contraindications and side effects. CD is gone from the body one hour after your last dose, leaving no trace behind. We dose hourly throughout the day because CD is only active in the body for up to one hour. We must continue to dose if we want to stop the proliferation of pathogens and eliminate them from the body.

A few children with autism in Venezuela recovered using only CD, The Diet, and a probiotic. That doesn't happen every time, but it can.

False Information and the Fear it Creates

The most common reason that families have told me they are afraid of trying CD is because of false propaganda. The FDA, anti-biomed parents, and others have claimed that MMS (Jim Humbles term for a weak concentration of chlorine dioxide in aqueous solution) is a "poison," comparing it to chlorine or chlorine bleach, which it is not.

If you are interested in a deep dive into the science behind chlorine dioxide therapy (also known as MMS), please spend some time on Dr. Andreas Kalcker's websites...

www.medicasalud.com

...mostly in Spanish, though it has a few pertinent documents in English, and...

www.andreaskalcker.com

Also check Jim Humble's website:

www.jimhumble.org

There, you can also find a forum (G2Cforum.org) of people discussing the use of CD (MMS) for the healing of a multitude of ailments.

In a nutshell, chlorine (Cl) and sodium hypochlorite ($NaOCl$, aka chlorine bleach) are as different from chlorine dioxide (ClO_2) as "required-for-life" oxygen (O_2) and "will-damage-the-lungs" ozone (O_3) are different from each other. Chlorine (Cl) and sodium hypochlorite ($NaOCl$) destroy pathogens through "chlorination," while chlorine dioxide (ClO_2) destroys pathogens by "oxidation." The by-products of chlorination can bond with other molecules and form potentially carcinogenic trihalomethanes. The only byproducts of oxidation by chlorine dioxide are two neutral oxygen atoms and a chlorite ion, which can bond with sodium in the body to form table salt ($NaCl$).

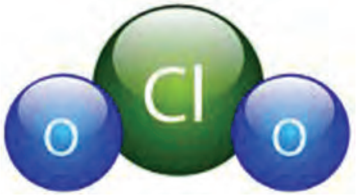

Chlorine Dioxide

≠

IS NOT

CLOROX
SPLASH-LESS

Laundry Bleach

OR

Pool Chlorine

Your doctor is most likely not a chemist, and may or may not be fluent in the language of oxidizers. If you are working with a doctor, please take the science to them when you speak with them about adding CD to your child's protocol. The best thing we can do for our sick children is to make informed decisions. Understanding the difference between chlorine and chlorine dioxide, or chlorine bleach and chlorine dioxide may mean the difference between your child recovering or not.

I have shared this with thousands of families around the globe whose children have used chlorine dioxide in very dilute aqueous solutions, and NONE have had any injuries from CD. Sure, we have had discomfort in the form of a Herxheimer reaction. This is something we have learned to minimize, and can often be prevented by going low and slow. I have learned that what may appear negative at the outset, often turns out to be a healing crisis that leads to incredible gains in the long run.

Many of the first world children who have done the CD protocol have done it in parallel with other autism protocols. They have had regular blood, urine, and stool tests and generally test healthier while using CD than before they started. CD is fortifying immune systems, improving liver function, reducing bacterial and viral loads, *Candida* markers, and inflammation.

> Just did round of blood tests on my son,
> Liver Function Test best ever after 7 months of MMS.

Critics also frequently cite the fact that sodium chlorite and chlorine dioxide can be used for industrial bleaching processes, as a way to frighten parents out of using it. Just because a substance has the power to remove color, or bleach something, doesn't mean it's "bleach." Lemon, sunlight, hydrogen peroxide, and toothpaste all remove color from textiles. We would never call toothpaste bleach, even though it's capable of removing color, nor would we call sunlight bleach, even though that's what it can do to your clothes if you leave them outside long enough. From here we can draw a very similar analogy to using chlorine dioxide for health. The dilutions we are using for internal consumption in aqueous solution (water) are very weak. Day one starts with just one drop in eight fluid ounces of water.

The 1984 study, *Controlled Clinical Evaluations of Chlorine Dioxide, Chlorite and Chlorate in Man*, found no deleterious effects when taking a concentration of up to 50ppm over several hours a day for three weeks.[1] As Jim Humble mentioned (see page 79), the concentrations used for industrial purposes are are thousands of times more concentrated than what families have been using for their children on the spectrum.

Any substance can become toxic if taken in too high a concentration—even water.[2,3] Chlorine dioxide is no exception. We always increase CD dosages, *low and slow*, and have a respect for the power that chlorine dioxide can have over pathogens, always resisting the urge to ramp up our dosing too quickly.

> We started MMS 68 Days ago and my son's ATEC has dropped from 50-23... Now that's a MIRACLE!!! He is So much more "with us" and I'm seeing improvements daily. He is just so much more fun to be with now:) He did have a plateau, but is back on track. We have only done 4 enemas to date, but they made a big difference! Baths do as well:) I can't wait to see what he is like on his 4th birthday (still 39 days away).

Chlorine Dioxide was discovered in 1814 by Sir Humphrey Davy as an effective disinfectant. Since then, numerous companies have patented its use for countless applications, including food and health.

The following is an excerpt from Jim Humble's Nov 21, 2012 newsletter:

> Dr. Andreas Kalcker spoke of patents obtained by different multinational pharmaceutical companies, to cash in on this product or to prevent it from being marketed.
>
> Some of those patents include:
>
> - Nontoxic Antiseptic (Pat 4035483/1977)
> - For combatting human amoebas (Pat.4296102/1981)
> - Against dementia caused by AIDS (Pat.5877222/1999)
> - For curing all types of illness of the skin (Pat 4737307/1988)
> - For disinfecting live blood (Pat. 5019402/1991)
> - For curing injuries more rapidly (Pat. 5855922/1999)
> - For all types of oral care (Procter & Gamble) (Pat. 6251372B1/2001)
> - Against infections caused by bacteria (Pat. 5252343/1993)
> - For treatment of severe burns (Pat.4317814/1982)
> - For the regeneration of bone marrow (Pat. 4851222/1989)
> - Treatment of Alzheimers, dementia etc. (Pat. 8029826B2/2011)
> - To stimulate the immune system in animals (Pat. 6099855/2000)
> - To stimulate the immunological system (Bioxy. Inc.) (Pat. 5830511/1998)

> ...The danger that may occur is if one mistakenly used a highly concentrated form. Highly concentrated forms of virtually every substance can cause death. In his conference, Dr. Andreas Kalcker points out that 70g of even plain table salt can cause death.
>
> There have been thousands of reports worldwide of excessive doses of MMS being ingested, however, with the hundreds of millions of tons of this substance used each year worldwide for water purification, preservation of vegetables, sterilization of slaughtered meat, and hundreds of other uses, there has never been a report of a death caused by ingestion of chlorine dioxide.
>
> This proves chlorine dioxide to be one of the safest products known.

There are many products that contain chlorine dioxide (ClO_2) that are currently approved by the Food and Drug Administration (FDA) in US. They are manufactured by *Frontier Farmaceutical*, *Alcide*, *Bioxy*, and others for skin and oral care. A number of these product names are listed below:[4]

- DioxiRinse™ Mouthwash
- DioxiBrite™ Toothpaste
- DioxiWhite™ Pro Teeth Whitener
- WhiteLasting™ Maintenance Gel
- BioClenz™ Dental Unit Waterline Cleaner
- Penetrator™ Periodontal Gel
- Simply Clear™ Acne Treatment
- DioxiWhite™ Home Teeth Whitener
- Cankers Away™ Canker Sore Cure
- DX7™ Skin Protectant Gel
- Periodontitis Treatment
- DioxiSmooth™ Facial Exfoliant
- Gingivitis Treatment Surface Disinfection
- Fire Fighter™ Burn Pain Reliever
- DioxiGuard™ Spray Disinfectant
- Nail-It™ Nail Protector

Chlorine Dioxide is an Oxidant… But aren't Antioxidants Good for Us?

In 1954, Denham Harman published an article called the "Free Radical Theory of Aging,"[5] claiming antioxidants would slow the aging process. However, he later found that mitochondria determined life span, and that antioxidants do not enter the mitochondria, and subsequently published the

"Mitochondrial Theory of Aging,"[6] in 1972. The "Free Radical Theory of Aging" is still responsible for billions of dollars in antioxidant supplement sales a year. While writing this book, a fascinating article was published in the journal *Open Biology* by Dr. James Watson (of Watson and Crick) called "Oxidants, antioxidants and the current incurability of metastatic cancers."[7]

Watson sets forth a hypothesis that links the presence of antioxidants in the body with late-stage (metastatic) cancer. He points out that some successful cancer treatments use "free radical" molecules to treat cancer—the very same molecules that antioxidants attack and kill. Watson urges his readers of the new paper to consider the following: "Unless we can find ways of reducing antioxidant levels, late-stage cancer 10 years from now will be as incurable as it is today."[8]

The following was excerpted from the same article in *Open Biology*.[8] Please see the original document for references contained within the text.

> Free-radical-destroying antioxidative nutritional supplements may have caused more cancers than they have prevented.
>
> For as long as I have been focused on the understanding and curing of cancer (I taught a course on Cancer at Harvard in the autumn of 1959), well-intentioned individuals have been consuming antioxidative nutritional supplements as cancer preventatives if not actual therapies. The past, most prominent scientific proponent of their value was the great Caltech chemist, Linus Pauling, who near the end of his illustrious career wrote a book with Ewan Cameron in 1979, *Cancer and Vitamin C,* about vitamin C's great potential as an anti-cancer agent [52]. At the time of his death from prostate cancer in 1994, at the age of 93, Linus was taking 12 g of vitamin C every day. In light of the recent data strongly hinting that much of late-stage cancer's untreatability may arise from its possession of too many antioxidants, the time has come to seriously ask whether antioxidant use much more likely causes than prevents cancer.
>
> All in all, the by now vast number of nutritional intervention trials using the antioxidants β-carotene, vitamin A, vitamin C, vitamin E and selenium have shown no obvious effectiveness in preventing gastrointestinal cancer nor in lengthening mortality [53]. In fact, they seem to slightly shorten the lives of those who take them. Future data may, in fact, show that antioxidant use, particularly that of vitamin E, leads to a small number of cancers that would not have come into existence but for antioxidant supplementation. Blueberries best be eaten because they taste good, not because their consumption will lead to less cancer.

Don't Oxidants Cause Oxidative Stress?

Not necessarily. Dr. Andreas Kalcker presented us with another theory explaining why oxidative stress is caused by pathogens and parasites rather than oxidants. If you take an oxidant such as CD, it is likely to upset the parasites living in the intestinal tract, causing them to defecate. Waste from worms can contain many toxic substances including MDA (malondialdehyde), formaldehyde, and ammonia. These toxins are known to cause oxidative stress.[9,10,11,12] When parasites are present in the body they are releasing these toxins, even if we don't kill them. Parasitic presence in the body is responsible for many of the symptoms known as autism.

Oxidants themselves are not directly responsible for oxidative stress in the body. However, they can exacerbate oxidative stress, which can cause parasites in the body to release toxins. The solution is to get rid of the parasites. Please see Chapter 8 for an effective Parasite Protocol.

Chlorine Dioxide & Antioxidants

Although antioxidants have their place in personal health and are common in many foods and supplements, they cannot be taken at the same time with chlorine dioxide—they negate each other. This is why vitamin C, juices, and other antioxidant containing foods are prohibited in *The Diet*.

> Just a testimonial to MMS: For those of you new to the group and still unsure of the MMS protocol I wanted to tell you about us. I was afraid at first, but we decided to try MMS. Within 3 weeks my son went from an ATEC of 36 to 18! That was amazing. Then we were afraid of the enemas. We said no way, we would just do oral MMS. Of course, after seeing others progress here we decided to do them. He would calm down! Then for months we agonized over the PP. We were afraid of it. We were scared of the meds, everything. Finally we decided to try it too. We are in the middle of our first PP and our son, who used to hit, kick, bite and scream at me over the smallest things all the time, has now been a perfect joy. When he gets mad it is over quickly and does not escalate into world war 3. Also these bumps he had on his face have cleared up along with his attitude! So if you are unsure of MMS like I was, here is one more testimonial of just how good the protocol really is. Thanks Kerri, and all the other moms that came before me!

Herxheimer Reaction (Die-Off Reaction)

There are many stories on the Internet about the “side effects” of CD. The definition of a side effect is:

The unintended or secondary effects of a substance.

When people describe what they believe to be side effects of CD, they are actually talking about a Herxheimer reaction. Jim Humble’s original chlorine dioxide (MMS) protocols called for higher doses per hour ramped up much faster than we titrate up for autism. In some cases that resulted in nausea, vomiting, and/or diarrhea. Critics of CD claim that a Herxheimer reaction, which is the result from ramping up too quickly, is the body’s reaction to being ‘poisoned’ by chlorine dioxide.

This is absolutely not the case, as the aforementioned study by Judith R. Lubbers and colleagues found that when healthy individuals were given minimal doses of chlorine dioxide in increasingly higher concentrations, no deleterious effects were noted.[2]

A Herxheimer reaction indicates that the patient is suffering from an overload of pathogens/toxins. If the person taking CD has a fair amount of pathogens, the rapid killing of these invaders produces excess toxins in the bloodstream. It is these substances that the body wants to rid itself of in short order. People with autism often have impaired methylation cycles that reduce their ability to eliminate toxins, making it even more difficult to handle an influx of toxins. So LOW AND SLOW is the way to go.

Children with autism typically don’t (or can’t) tell you they have nausea. However, discomfort may be visible from seeing that they’re eating less, their eating habits have changed, or they’re more listless than usual. That’s usually a sign of detox, and in which case, we won’t give the child any more CD that day. Detox is great, it’s what we want, but without stressing the body further. We need to stay just beneath a Herxheimer reaction.

A Herxheimer reaction can present itself in many forms including:

- Blisters
- Bloating
- Burping
- Chills
- Diarrhea
- Fatigue
- Flatulence
- Flushing
- Headache
- Heart Burn
- Hiccups
- Nausea
- Runny nose
- Sleep changes
- Vomiting

If any of these symptoms appear, suspend the remaining doses of CD for the rest of the day. The following day, return to the number of doses given when the child was stable, which is usually the dose that was given the day prior to the Herxheimer reaction.

To mitigate a Herxheimer reaction from chlorine dioxide, you can give vitamin C or orange juice, as either one of these will cut the effects of CD immediately. To help bind the excess toxins you may want to try activated charcoal, bentonite clay, (be aware that these both can cause constipation), or burbur (a liquid detox extract).

Using Chlorine Dioxide for Autism

As I developed this protocol, it became clear to me that we needed to focus on excesses causing the symptoms of autism, rather than deficiencies caused by them. In the autism biomed world, we are accustomed to supplementing deficiencies (low iron, low B_{12}, low calcium, etc.), without questioning why the deficiency exists in the first place (i.e., presence of pathogens/parasites).

As soon as a family completes one week on *The Diet*, we get right into CD. Everyone starts with the "baby bottle" method, one drop of CD in eight fluid ounces of water, which makes the first day's dose 1/8 of a drop at a time. We have greatly minimized Herxheimer reactions by using this method, but it remains very important to monitor your child while they are using chlorine dioxide. It's very important to go low and slow. Every drop we give is helping to heal our children.

What is Chlorine Dioxide (CD)?

Refer to the diagram on the right for a visual overview.

You can't actually buy CD. You have to make it. Fortunately, it's not a big deal. If you can make coffee... you can make CD, which stands for *chlorine dioxide*.

Chlorine dioxide (ClO_2), is a gas produced through a chemical reaction when two liquids, sodium chlorite ($NaClO_2$) and an acid (such as hydrochloric acid (HCl) or citric acid ($C_6H_8O_7$)), are mixed. The acid brings the combined pH level below five, causing the sodium chlorite to become unstable and release chlorine dioxide (ClO_2).

If any of this sounds complex, it isn't. Go through the following explanation and don't get bogged down by the chemical symbols.

Conceptual Overview of Chlorine Dioxide (CD)

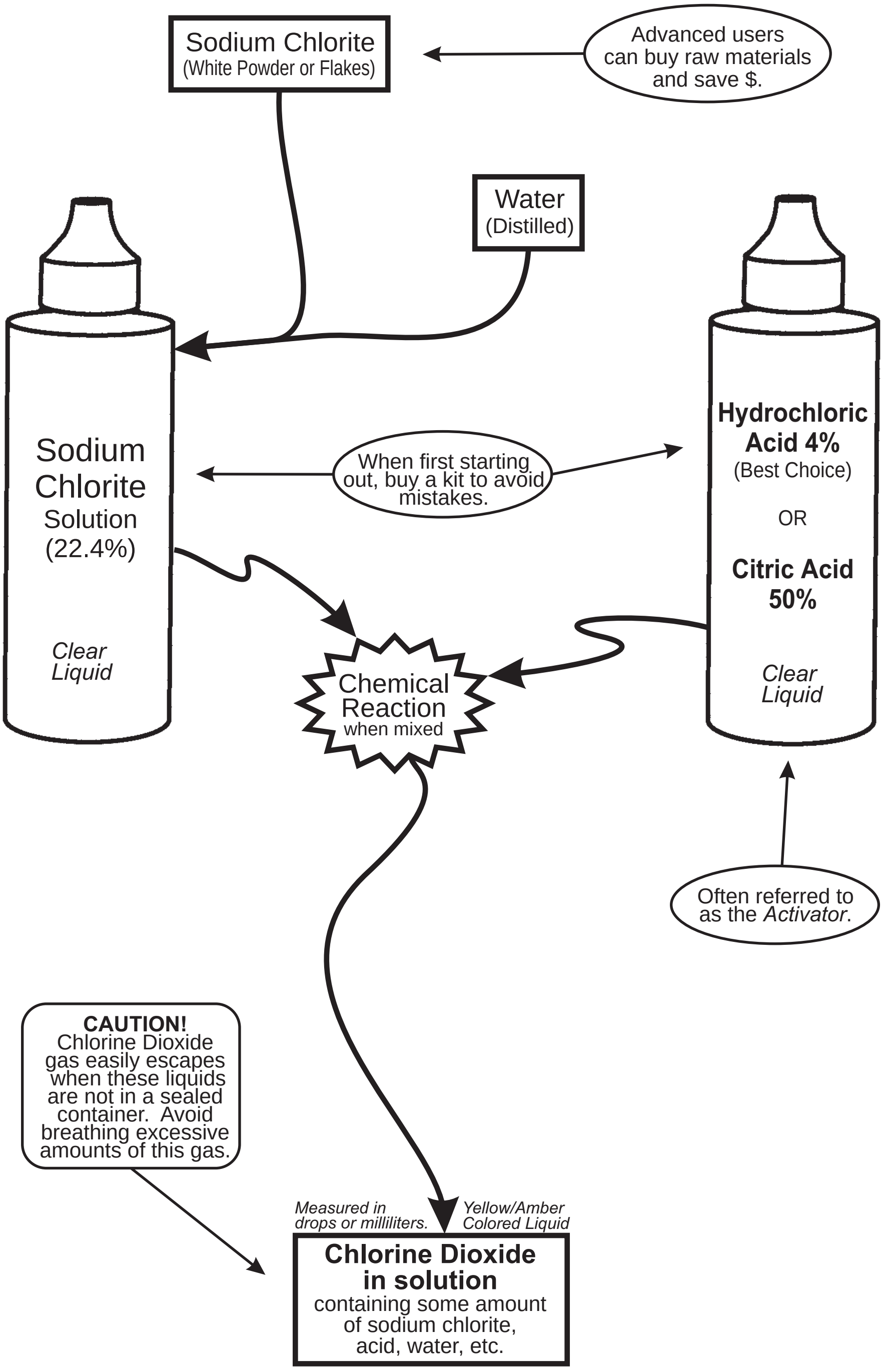
Sodium Chlorite
(White Powder or Flakes)
Water
(Distilled)
Sodium Chlorite Solution (22.4%)
Clear Liquid
Hydrochloric Acid 4%
(Best Choice)
OR
Citric Acid 50%
Clear Liquid
Chemical Reaction
when mixed
Measured in drops or milliliters.
Yellow/Amber Colored Liquid
Chlorine Dioxide in solution
containing some amount of sodium chlorite, acid, water, etc.

Advanced users can buy raw materials and save $.

When first starting out, buy a kit to avoid mistakes.

Often referred to as the *Activator*.

CAUTION! Chlorine Dioxide gas easily escapes when these liquids are not in a sealed container. Avoid breathing excessive amounts of this gas.

Sodium Chlorite ($NaClO_2$)

This is the first chemical you must have to make CD.

You can get it in one of two forms:

- **Sodium chlorite liquid in water:**
    - If you are beginning, get a starter kit of this.
    - Expect to pay about $20-$30 for a 4 fl. oz. bottle set (usually includes acid activator).
    - 22.4% sodium chlorite solution (from 28% technical grade sodium chlorite salts).
    - Recommended: Buy quart or liter sized quantities of liquid once you understand the principles.
- **Sodium chlorite powder or flakes:**
    - FOR ADVANCED USERS ONLY!
    - Cheaper in the long run.
    - Recommend buying 5 lbs. of flakes/powder for a cost of about $100-$200.
    - Look online or check with your supplier for mixing instructions.
    - Avoid buying 90% sodium chlorite flakes. You want 80%. Most suppliers that sell liquid also sell flakes.
    - NEVER mix any acid directly with powder/flakes!

What's the 28%? 22.4%?

This is not a critical thing for you to worry about unless you are mixing your own. But here is the jist: sodium chlorite powder is not 100% sodium chlorite. It's usually 80% sodium chlorite and 20% salt and other inert ingredients. When diluted in water, you end up with 28% solution of powder, but only 22.4% is sodium chlorite. Labels usually show one or both of these numbers.

Sodium Chlorite Flakes vs. Powder

Sodium chlorite flakes and sodium chlorite powder are the same thing. The resulting physical consistency may vary by different manufacturers. We've even seen clumps of powder. It's still the same chemical. The important part is that it is technical grade, and 80% of the powder/flakes should be sodium chlorite ($NaClO_2$). Mixing info is at: *www.youtube.com/watch?v=t0C8XVm9QVE*.

Acid - The Activator

To produce CD, you must add the second chemical, which is a food acid, to the liquefied sodium chlorite (22.4% solution in water). Chlorine dioxide gas is produced when the pH of the sodium chlorite goes below five. We call the acid used to lower the pH of sodium chlorite an "activator" because it activates the sodium chlorite to produce the chlorine dioxide.

There are several activators to choose from. These are food acids in that they are commonly found in food preparations, or are actually food:

- **Hydrochloric Acid (HCl)**
    - The same acid found in your stomach, and therefore, the most biocompatible option.
    - Working concentration is between 4% and 10%.
    - Available on the market in liquid concentrations of 4%, 5%, 10%, 32%, 35%, and 37%.
    - Any concentration over 10% must be reduced to a working concentration of between 4% and 10%. See Appendix 7, page 471. (Note: We use 4% for CD and CDH; & 10% for CDS.)
    - Only available in liquid form.
- **Citric Acid**
    - Commonly available around the world.
    - Often found in soft drinks to give them a tangy flavor.
    - Some people have problems tolerating this acid.
    - Usually used in 50/50 concentration by weight mixed in water.
    - Available in liquid and powder form.
    - Inexpensive; can be purchased online.
- **Lemon or Lime Juice**
    - Consider this an emergency alternative to be used if the first two choices are not available.
    - Can be found in most stores carrying produce.
    - Should be from pure, organic and freshly squeezed lemons (or limes).
    - Relatively weak acidity requiring five drops of lemon juice to every one drop of sodium chlorite solution.
    - Substantially longer activation time.
    - Lower ClO_2 output.

The amount of acid used, and the length of time you wait to dilute the mixture, will depend on the acid activator and its strength. The following chart gives you an idea.

Activator	Strength	Ratio	Time
Hydrochloric Acid	10%	1:1	40 secs
Hydrochloric Acid	4%	1:1	60 secs
Citric Acid	50%	1:1	60 secs
Citric Acid	33% or 35%	1:1	60 secs
Lemon Juice	Pure	1:5	120+ secs

You are unlikely to be using lemon juice, but we've included it here for completeness.

Families should use whichever activator is available to them depending on where they live and what they have access to. All kids are different, so if yours doesn't respond well to 4% HCl, by all means, try citric acid, or vice versa.

If the HCl you purchased is anything other than 4% or 5%, then see Appendix 7 for instructions on reducing concentration.

If you are using one of the two most commonly available activators, then you are likely to be using a 1:1 mixing ratio (1 drop sodium chlorite + 1 drop activator).

In the case of 10% citric acid or pure lemon juice, you would be using five drops of activator to one drop of sodium chlorite, and waiting about three times as long as if you were using one of the newer, more concentrated activators.

Purchasing Product

When you first start out, you should buy a starter kit and get familiar with the chemistry. A starter kit usually comes in two bottles of four fluid ounces each.

1. Sodium chlorite, (4 fl. oz.)
2. Activator (hydrochloric acid or citric acid), (4 fl. oz.)

The bottles look something like those shown in the diagram on page 91. Getting the hydrochloric acid activator is slightly preferable.

You will also need an 8 fl. oz. baby bottle. This is a handy tool for storing a day's worth of CD doses with the added benefit of having marks that show how much it contains. Glass bottles are much better than any kind of plastic bottle. Lifefactory.com (shown on the right) produces a bottle we especially like. It is made of silicon-covered glass and has a reliable top. This particular bottle is marked as having a capacity of nine fl. oz., but we use it up to only eight fl. oz. Plastic should only be used for lids; avoid bottles with metal tops. Over time, CD will oxidize metals so they can't be used.

Another item that most people have at home is a shot glass or wine glass that tapers down in the middle like a "U" shape. This will be used to mix the chemicals and the tapered shape insures that even as few as two drops will find each other. You can also use a cup with a flat bottom, but must tilt it so both solutions pool together.

lifefactory

A standard shot glass with a "U" shaped bottom is the preferred mixing container for CD & activator. A glass baby bottle, such as the one made by Lifefactory®, is the ideal dosing bottle.

Terminology of Available Products

There can be a bit of confusion for some as to what to look for and buy. The following chart shows the actual chemical, and some of the common names found on the label. There may be other names we are not aware of. In most cases, the chemical and its strength are identified on the bottles.

Common Activators & Their Various Names on the Market	
Chemical	**Common Names**
Sodium Chlorite	Chlorine Dioxide – Part 1 (Purifier) Master Mineral Solution Miracle Mineral Solution Miracle Mineral Supplement MMS MMS Drops MMS Solution MMS Water Purification Drops MMS1 NaClO2 Natriumchlorit (German) Pathogen Biocide Sodium Chlorite 28% Water Purification Drops Water Purificator
Hydrochloric Acid	Activator Chlorine Dioxide – Part 2 (Activator) MMS Activator
Citric Acid	Activator Citric Acid – 20% Citric Acid – 50% MMS Activator

Checklist for Preparing a CD Dose

Assuming you have received the starter kit, it's time to make your first dose. Here's the check list of what you now should have on hand before you begin:

- ☐ Bottle of sodium chlorite solution (28% sodium chlorite salts / 22.4% sodium chlorite)
- ☐ Bottle of acidic activator (hydrochloric acid or citric acid)
- ☐ A clean and dry shot-glass with a "U" shaped bottom.
- ☐ 8 fl. oz. glass baby bottle.
- ☐ Distilled or reverse osmosis water (filtered water will do, but no alkaline water!).

A Few Points Worth Noting!

Looking at the short checklist above, it should be apparent that this is not going to be rocket science. Nevertheless, a few pointers are in order:

- In the past, people who used CD for various things would mix a fresh batch each time they took CD. This involved a lot of extra work that could be avoided by producing one batch big enough to last the entire day. For this exercise, we are assuming you are starting out and going to do things low and slow. We are going to produce 8 fl. oz. of water solution containing one drop of CD, with the intention of your child consuming one fl. oz., of water (containing 1/8 drop of CD) eight times throughout the day.

- When you mix sodium chlorite solution and the acidic activator, you are mixing equal drops of each substance (unless you are using one of the less concentrated activators mentioned earlier).
- CD doses are measured in drops of sodium chlorite. So one drop of sodium chlorite + one drop of activator = one drop of CD.
- On your first day, start with one drop of sodium chlorite and one drop of activator (in other words—ONE drop of CD).
- If you put in too many drops of one solution or the other, start over with a clean and dry shot glass.
- A batch of CD is good for 24 hours. However, you must keep the cap on tight!
- Keep the bottle with the daily CD batch out of direct sunlight. A brief exposure won't ruin it, but it will become weaker. Why? The chlorine dioxide gas will come out of the water and sometimes form a cloud in the air pocket of the bottle. No need to worry if you see the cloud—it's a common occurrence.

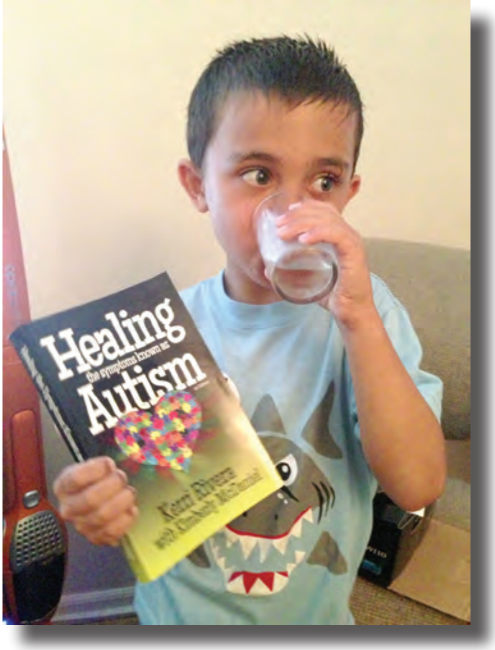

Vincent taking his hourly dose.

Preparing Your First One drop CD Batch for the Day

1.	Start by filling the baby bottle with 8 fl. oz. (237ml) of water (distilled or reverse osmosis). NO alkaline water!	
2.	Place 1 drop of sodium chlorite solution into the CLEAN and DRY shot glass.	
3.	Add 1 drop of acidic activator (hydrochloric acid or citric acid) to the shot glass containing the drop of sodium chlorite. The number of drops may be higher if you are using a weaker activator. See chart on page 94.	
4.	Now wait the appropriate time for the mixture to react (see chart on page 94). You should see the color change from clear to slightly yellow. If there were more drops in the shot glass, the color change is more noticeable. You are also likely to notice the chlorine-like smell coming from the shot glass. Remember, this is NOT chlorine, but rather chlorine dioxide.	
5.	After the activation time has passed, pour a little water from the baby bottle into the shot glass and let it mix. This mostly stops the chemical reaction and insures you get most of the mixture out in the next step.	
6.	Lastly, pour all of the watered down mixture in the shot glass back into the baby bottle, and seal it tightly with the cap—don’t leave it sitting open for any length of time. Think partially used soda pop and how you would want to keep that closed.	

Voilà! You have just produced your first batch of CD!

This is now a ONE drop batch, that makes EIGHT hourly doses—each 1/8 of a drop in strength. If your batch is larger; your hourly dose is stronger:

- 2 drop batch = 8 hourly doses of 1/4 drop
- 4 drop batch = 8 hourly doses of 1/2 drop
- 8 drop batch = 8 hourly doses of 1 drop

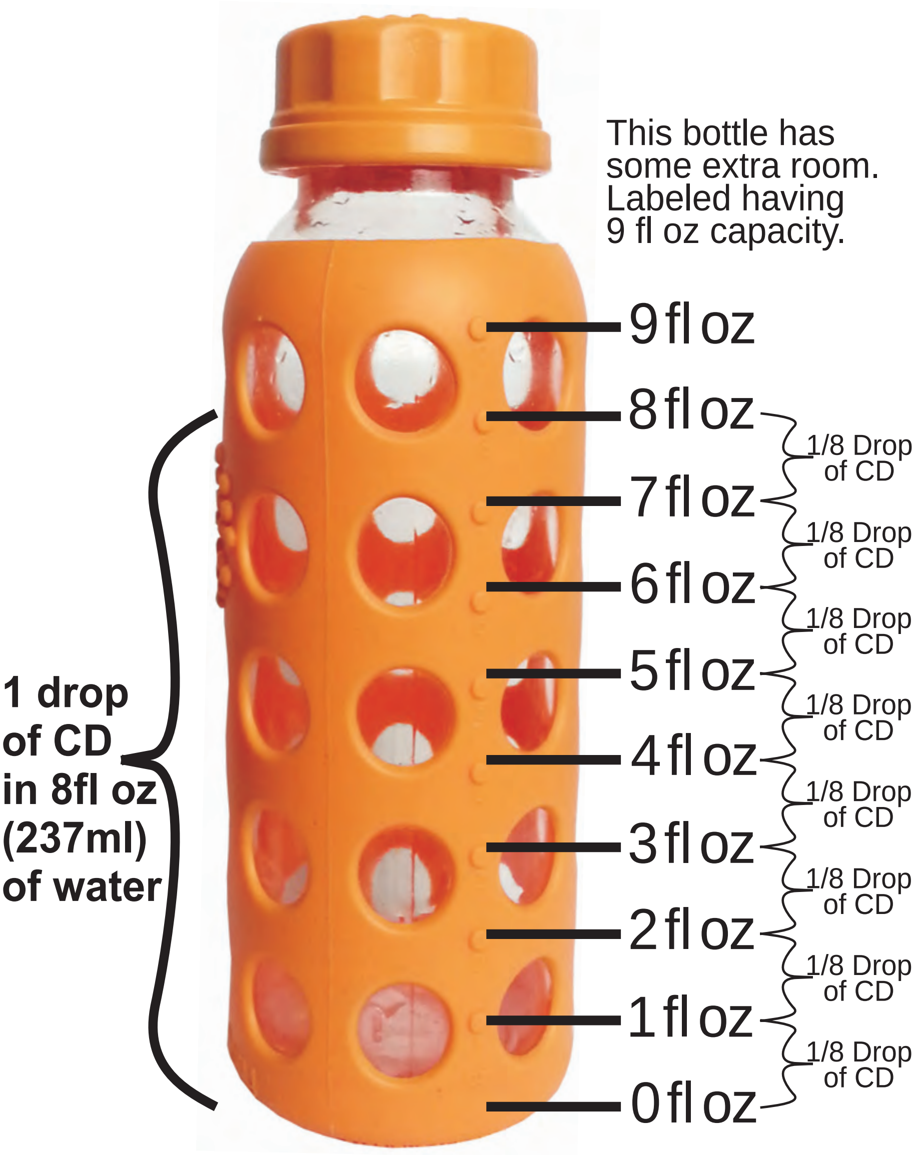
This bottle has some extra room. Labeled having 9 fl oz capacity.
9 fl oz
8 fl oz
7 fl oz
6 fl oz
5 fl oz
4 fl oz
3 fl oz
2 fl oz
1 fl oz
0 fl oz
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1/8 Drop of CD
1 drop of CD in 8fl oz (237ml) of water

Oral Protocol for CD in Autism Spectrum Disorders (ASDs)

In the previous exercise, you learned how to prepare the dose for Day One. Children and adults all start by taking only one drop of CD divided into eight hourly doses. On Day Two, you make a batch of two drops sodium chlorite plus two drops activator and mix that into eight fluid ounces of water. Increase this dosage by one drop per day, unless you see a Herxheimer reaction (see page 89 for more information).

The chart on the right allows you to determine the estimated full oral CD dose based on the weight of the individual. This gives you an idea of the maximum goal number of drops per day for your child. Once the child has been at the full suggested dose for several months, we may need to increase again, by a few drops, to prevent a plateau.

Important Note: Some children can't go up by one drop each day and need to go slower. A child may have a relatively higher toxic load, or he may be extremely sensitive. Whatever the case, resist the urge to try to get to full oral dose as quickly as possible. This is not a race! Some children need to spend a few extra days at the same dose to avoid a Herxheimer reaction. Even if you must remain at one level for a few days or more does not mean you are making no progress. Whatever the level of dosage, you are detoxing the body.

When you get within two to three drops of your child's maximum dose by weight, go ahead and spend three days at each dose, while carefully observing your child, moving up only when your child is stable. For example, if the full dose is 16 drops in the 8 fl. oz. (237ml) baby bottle, stay at 14 drops for a few days to see how that dose level is tolerated, and then increase every third day till reaching the full dose. If the person is better at a drop or two below the calculated full dose, go back to that lower dose that they tolerated best.

> *Just had a fascinating appointment with my son's functional neurologist. My son had this third Qeeg today, which clearly showed that previously under-stimulated parts of his brain are 'waking up'. I told our doctor about CD very early on and he believes that it is THIS that is having such a huge impact on his brain. He got quite emotional and said 'every time you get a worm, you are saving your son. Keep doing what you're doing, cause he clearly isn't autistic. He's poisoned'. Goosebumps!!*

Estimated Full Oral CD Doses by Weight

Use these numbers as a guide only. You may need to go up by as much as 50% or more over the indicated drops. Read chart as: **POUNDS / KILOGRAMS → DROPS OF CD (per 8 fl. oz. water)**

25/11→8	62/28→17	99/45→24	136/62→29	173/78→33	210/95→37
26/12→8	63/29→17	100/45→24	137/62→29	174/79→33	211/96→37
27/12→9	64/29→17	101/46→24	138/63→29	175/79→34	212/96→37
28/13→9	65/29→18	102/46→24	139/63→30	176/80→34	213/97→37
29/13→9	66/30→18	103/47→24	140/64→30	177/80→34	214/97→37
30/14→9	67/30→18	104/47→25	141/64→30	178/81→34	215/98→37
31/14→10	68/31→18	105/48→25	142/64→30	179/81→34	216/98→37
32/15→10	69/31→18	106/48→25	143/65→30	180/82→34	217/98→37
33/15→10	70/32→19	107/49→25	144/65→30	181/82→34	218/99→38
34/15→10	71/32→19	108/49→25	145/66→30	182/83→34	219/99→38
35/16→11	72/33→19	109/49→25	146/66→30	183/83→34	220/100→38
36/16→11	73/33→19	110/50→26	147/67→31	184/83→35	221/100→38
37/17→11	74/34→19	111/50→26	148/67→31	185/84→35	222/101→38
38/17→11	75/34→20	112/51→26	149/68→31	186/84→35	223/101→38
39/18→12	76/34→20	113/51→26	150/68→31	187/85→35	224/102→38
40/18→12	77/35→20	114/52→26	151/68→31	188/85→35	225/102→38
41/19→12	78/35→20	115/52→26	152/69→31	189/86→35	226/103→38
42/19→12	79/36→20	116/53→26	153/69→31	190/86→35	227/103→38
43/20→13	80/36→21	117/53→27	154/70→31	191/87→35	228/103→38
44/20→13	81/37→21	118/54→27	155/70→31	192/87→35	229/104→38
45/20→13	82/37→21	119/54→27	156/71→32	193/88→35	230/104→39
46/21→13	83/38→21	120/54→27	157/71→32	194/88→35	231/105→39
47/21→14	84/38→21	121/55→27	158/72→32	195/88→36	232/105→39
48/22→14	85/39→21	122/55→27	159/72→32	196/89→36	233/106→39
49/22→14	86/39→22	123/56→27	160/73→32	197/89→36	234/106→39
50/23→14	87/39→22	124/56→28	161/73→32	198/90→36	235/107→39
51/23→15	88/40→22	125/57→28	162/73→32	199/90→36	236/107→39
52/24→15	89/40→22	126/57→28	163/74→32	200/91→36	237/108→39
53/24→15	90/41→22	127/58→28	164/74→32	201/91→36	238/108→39
54/24→15	91/41→23	128/58→28	165/75→33	202/92→36	239/108→39
55/25→16	92/42→23	129/59→28	166/75→33	203/92→36	240/109→39
56/25→16	93/42→23	130/59→28	167/76→33	204/93→36	241/109→39
57/26→16	94/43→23	131/59→28	168/76→33	205/93→36	242/110→39
58/26→16	95/43→23	132/60→29	169/77→33	206/93→37	243/110→39
59/27→16	96/44→23	133/60→29	170/77→33	207/94→37	244/111→40
60/27→17	97/44→24	134/61→29	171/78→33	208/94→37	245/111→40
61/28→17	98/44→24	135/61→29	172/78→33	209/95→37	246/112→40

Serving Cup Dilution

Don't have your child drink out of the baby bottle. Pour the hourly dose into a separate cup and add extra water if needed. Make sure to add water to each dose to ensure that there is a **minimum** of one fluid ounce of water per activated drop of CD. However, there is no maximum amount of water for dilution. To be absolutely clear about this, note the following chart which assumes eight fl. oz., daily batches containing eight hourly doses.

Drops of CD in Batch	Hourly Dose	Min. Water Dilution
1	1/8 drop	1 fluid ounce
2	2/8 drop	1 fluid ounce
3	3/8 drop	1 fluid ounce
4	4/8 drop	1 fluid ounce
5	5/8 drop	1 fluid ounce
6	6/8 drop	1 fluid ounce
7	7/8 drop	1 fluid ounce
8	1 drop	1 fluid ounce
9	1 1/8 drops	2 fluid ounces
10	1 2/8 drops	2 fluid ounces
11	1 3/8 drops	2 fluid ounces
12	1 4/8 drops	2 fluid ounces
13	1 5/8 drops	2 fluid ounces
14	1 6/8 drops	2 fluid ounces
15	1 7/8 drops	2 fluid ounces
16	2 drops	2 fluid ounces
17	2 1/8 drops	3 fluid ounces
18	2 2/8 drops	3 fluid ounces
19	2 3/8 drops	3 fluid ounces
20	2 4/8 drops	3 fluid ounces
21	2 5/8 drops	3 fluid ounces
22	2 6/8 drops	3 fluid ounces
23	2 7/8 drops	3 fluid ounces
24	3 drops	3 fluid ounces
25	3 1/8 drops	4 fluid ounces
26	3 2/8 drops	4 fluid ounces
27	3 3/8 drops	4 fluid ounces
28	3 4/8 drops	4 fluid ounces
29	3 5/8 drops	4 fluid ounces
30	3 6/8 drops	4 fluid ounces
31	3 7/8 drops	4 fluid ounces
32	4 drops	4 fluid ounces

CD is active in the body for about one hour and, as result, hourly doses give us the best progress. This is difficult on school days, but we recommend that you dose two to three times before school and the rest after to distribute the pathogen-fighting more evenly. Eight doses a day is the minimum requirement, but the more you can fit in the better. Try to dose every hour as much as possible.

Special Cases: "The Double Dose"

As a child goes through the detox process, they may exhibit conduct which can include hyperactivity, anxiety, OCD, aggressiveness, and night wakings. In these cases, a double dose is 2 fl. oz. (60ml) of the baby bottle given at once. For example, if your child receives one drop per fluid ounce of the baby bottle method, they will be receiving 2 drops. Let's say for example your child wakes up at 1am, give them a double dose. If your child is still awake at 2am, give another dose. The most common scenario is that after 1 or 2 double doses, the child will fall back asleep. This helps oxidize the toxins from the die-off. During the day if your child has a tantrum or other issues with conduct, we can double dose every hour as well.

CD has two functions with pathogens; one is that is oxidizes the pathogen itself, and the second is that it oxidizes the toxins released by the pathogens while they are living, dead, or dying.

Advanced Protocol – The Enema!

In early 2011, we added enemas to the protocol to kill the pathogens causing dysbiosis in the large intestine (we didn't know about parasites yet). We wanted to get the chlorine dioxide into the blood stream so it could kill the biofilm that exists in the blood. In this way, the blood can carry the CD past the blood-brain barrier to kill pathogens in the brain

When we are detoxing, it is absolutely critical to keep the colon moving and avoid the reabsorption of toxins through the intestinal walls. Enemas allow us to do just this. Some toxins can exit the intestine through the intestinal wall (more so if leaky-gut syndrome is present), and cross the blood-brain barrier, therefore affecting cognition and behavior. When we cleanse the colon, we get those out before they can cross into the brain, and we detoxify the lymphatic system, liver, and gallbladder.

The Enema Stigma

When I say the word enema to a family new to the protocol, I often hear things like "*I could never do that!,*" or "*My son would never let me do an enema.*" I assure you it's not traumatic as it has been hyped up to be. Modern society has gotten away from it, but for millennia, cleansing the colon was considered part of daily hygiene. Enemas were mentioned in the Vedas, the Bible, and by the Mayans. The reason for doing CD enemas is NOT for relief of constipation. Of course, when constipation is present, CD enemas can relieve the cause of constipation.

> Great news! First enema today...
>
> My son was so happy that we did and the first thing he said while sitting on the toilet was: "Mommy! IT WORKED!!!" hahaha He is so used to sitting on the toilet and having such a hard time. Poor thing, he was so happy that IT WORKED!
>
> And he even said he wants to do it again tomorrow. He felt so much better and ate so much more for dinner with an empty stomach. I think he got rid of stuff that was there for ages! His tummy is now soft and not hard as usual. Quite amazing and it didn't bother him at all.
>
> Today his molar hasn't bothered him so a perfect day to start the enemas! His appetite has gone up quite a bit. He has seldom asked for food before and now he asks a few times a day for something to eat. The diet is going wonderfully and he is now used to the change of foods. He asks about cheese and milk but I try to talk about something else to distract him so he eventually stops asking. hahaha He hasn't gotten upset about food he wants (that he can't have) not even once. YAY!

I would like to clear up one thing here before we get in the mechanics of the enemas. More than one person has asked in the blogosphere, "*If Kerri Rivera loves enemas so much, why doesn't she do them on herself???*" As a matter of fact I do! I have done my own enemas for years, so has my husband, my sister, and plenty of the parents using this protocol for their children. We understand the benefits of a fit colon, and have no problem applying our own enemas or going for a colonic from time to time.

Actually, I would strongly advise you to practice on yourself. That way you know from first-hand experience that this is really no big deal. You will also discover how much better you feel after cleaning out your colon, and you will understand when your child feels better as well.

When to Start Giving Enemas

Enemas should be started when the child reaches full oral dose.

However, if your child suffers from constipation you can add enemas and baths on day one (see page 113), before you get to the full oral dose. However, alternate them on opposite days or opposite ends of the day.

Frequency of Enemas

I recommend enemas no less than once every other day. You may need to do them more often, particularly if you see parasites coming out. Doing as many as four in a day is not uncalled for if each one results in parasites being released. Some of the older children seem to pass a lot of parasites in one day, typically in or around a full or new moon.

Timing of Enemas

If your child is going to school, avoid giving an enema in the morning since they may need to visit the toilet a few times after. If the child is homeschooled or too young or sick to attend, the timing is up to you. Ideally, do the enema after the child has a bowel movement.

Enemas — What You Need

The first thing you are going to need is some kind of a large plastic, liquid measuring container. Glass is not recommended since it can get knocked over and break in the confines of a bathroom.

You will also need all the same things you used for making CD batches:

- Water. Use reverse osmosis, filtered, or distilled.
  DO NOT use tap water or alkaline water!
- Sodium chlorite solution
- Acidic activator
- Shot glass
- Natural lubricant (coconut oil or any other simple food oil). NO perfumed baby oil or chemical lotions such as Petroleum Jelly or K-Y® Jelly.
- Enema kit (see your options on the following page)

Enema Options (Most to Least Preferred)

Gravity Bag or Bucket with hose

PiC indolor
IRRIGADOR
ENTEROCLISE
KLISTEERSPUIT
ΚΛΙΣΜΑ

Pros:
- Holds plenty of liquid
- Can be refilled without taking it down
- Flow is easy to control
- Can be hung on a door
- Can be used for both enema and douching
- PIC Indolor: Folds for easy travel
- PIC Indolor: Has valve
- PIC Indolor: Two sizes of cathetors

Cons:
- Bucket doesn't fold for easy travel

Enema / Douche Bag / Hot Water Bottle

Hot/Cold
Water Bottle System
HOT/COLD Therapy Anytime... Anywhere!
Safe
Durable
Hygienic

Pros:
- Holds plenty of liquid
- Flow is easy to control
- Can be hung on a door
- Can be used for both enema and douching
- Easy travel
- Has two sizes of cathetors

Cons:
- Can't be refilled without taking down
- Usually made of rubber, which tends to dry out quickly, especially from CD.

Syringe & Catheter

Pros:
- Flow is easy to control
- Can be used for both enema and douching
- Good for small children

Cons:
- Holds only 60ml-140ml
- Need more than one.
- Have to swap them out to go beyond the capacity of one syringe.
- The graduations rub off easily (suggest putting clear shipping tape over them to protect the markings if they are important to you).

Enema Bulb

Pros:
- Portable
- Can be used for both enema and douching
- Good for “implants” of mebendazole

Cons:
- Volume too small for an effective treatment.
- Hard to clean inside
- Can’t tell if dirt is on the inside.
- Holds only about 200ml.

Fleet Enema (Last Resort)

Fleet
Enema
saline
Twin Pack
LAXATIVE
ready-to-use
latex free
Soft, Flexible Comfortip
EASY SQUEEZE

Pros:
- Cheap
- Can find it in most pharmacies
- Portable
- Can be used for both Enema and Douching

Cons:
- Volume too small for an effective treatment.

Enema Kits

As shown in the previous two pages, here are your enema kit options listed from best to less than good. Prices are estimates in US Dollars and are likely to be higher outside the US:

★★★★★ PIC Indolor Gravity Bag (~$15-$30)
★★★★☆ Enema Bucket (~$5-$10)
★★★⯪☆ Multipurpose Enema/Douche Bag/Hot Water Bottle (~$15-$30)
★★★☆☆ 100ml+ Syringes plus Catheter (~$5-$10)
★★☆☆☆ Enema Bulb (~$5-$15)
★☆☆☆☆ Fleet Enema (emptied out) (~$5-$15)

You must have a separate enema kit for each person doing enemas. This insures that you are not cross-contaminating parasites.

In my own home we have recently started using the PIC Indolor enema/douche bag. This has made the process faster, and we can tolerate more water before feeling full and desiring to release the enema. This is probably the most practical kit, especially because of the built-in shut-off valve. The enema bucket is a close second, but the sliding clamp that comes with it doesn't last long and breaks easily. You can always just kink the hose and control the flow that way.

The multipurpose enema/douche/hot water bottle type is a good option, but it is usually made of rubber. This material often goes bad over time and CD just speeds up that process. Rinse it out several times before and after use to insure that no breakdown or CD residue is left. Don't use if you notice the rubber becoming brittle.

Even two-year-olds still use about 400 to 500ml in an enema. The intestine is long. Parents with small children sometimes like the 100ml syringe/catheter method. The enema bulb or fleet enema are not really practical for our purposes due to the small volumes they hold. However, if you want to get started right away, they are usually available in most major pharmacies.

You can find syringes, catheters, gravity bags, multipurpose enema/douche bags on various online sites, pharmacies and medical supply stores around the world.

Preparing and Administering the Enema

Make all your supplies easily accessible in your work area before bringing your child in the bathroom. If this is your first time and you are working out the process, you may wish to ask your spouse or someone else for assistance. It's a good idea to do a dress rehearsal without the child, to insure that it goes smoothly when the child is present. Once you and your child have the process down, it will be easy. Being prepared is the key to success.

Please note: The eight fluid ounce baby bottle is NOT involved in any way with enemas!

1. Ideally, time the enema just after a bowel movement.
2. Start by washing your hands to insure you are not introducing any new bacteria into the colon. Everything you are working with should be clean. Wearing gloves and a face/surgical mask is HIGHLY recommended.
3. Assuming you are using one of the suggested gravity bags, find a good and stable location to hang it up. The best location really depends on your particular bathroom's layout and where you are going to place the child during the procedure. I hang mine from a towel hook in the bathroom, or you can use the shower curtain rod. The idea is to have the bag/bucket about three feet (~one meter) over where the enema will be administered. Most parents put a towel on the floor of the bathroom and lay their child on the towel (smaller children can lay across your lap).
4. Refer to the chart below and measure out the appropriate amount of water for the individual receiving the enema. The ratio is two drops of CD per 100ml of filtered warm water.

Age/Size	Water Volume	Drops of CD
Child	1/2 Liter (500ml)	10
Adolescent	1 Liter (1000ml)	20
Teen/Adult	up to 2 Liters (2000ml)	40

5. Don't add any CD yet. The next step is to heat the water to body temperature (98°F / 37°C). The water should feel warm to the touch—NEVER hot! Similar to the temperature of milk in a baby bottle... test it on your wrist. If you are concerned about getting the temperature right use a quality thermometer. Be aware... many cheap thermometers are notoriously inaccurate!

6. After the heating step, pour the water back into the big measuring cup.

7. Referring again to the chart on the previous page, prepare a CD mixture in the clean dry shot glass that's appropriate for the volume of water needed. For example, if you have an adolescent and are preparing one liter, then you will need to mix 20 drops of sodium chlorite and 20 drops of acidic activator to end up with 20 drops of CD! Note: The most important part of the enema is the ***water***. Resist the urge to add more than two drops per 100ml of water.

8. After the appropriate activation time, pour the concentrated CD from your shot glass into the large measuring cup. You may want to rinse out the shot glass with some water from the large mixing cup so you get all the concentrated CD solution into the water.

9. The bag/bucket should now be in its working location with the hose firmly attached, Place the end of the hose so it is over the top of the bag/bucket. A clothes pin can come in handy to hold it up over the bag/bucket. Now, fill the bag/bucket about 3/4 full with the CD water mixture you just prepared.
10. Make sure the hose is filled with water. You may want to let a bit run out into the sink or some other container and then clamp it. The bucket type enema has a plastic clamp for this purpose. The bag type has a valve where the catheter attaches. What you don't want is for there to be a bunch of air in the hose that then is transferred into the colon.
11. If all of the above is ready, it is time to bring your child into the bathroom.
12. Explain what you are going to do. Some parents say that they are going to put the catheter in his bottom, and that it has "special water" in it to help him feel better. One little boy called it "butt medicine" while another called it "butt wash." We recommend you say that you are the only person who is allowed to give this to him. Just say whatever you need to say to explain what is happening.
13. If the child is on the floor (on a rug or towel), they should lie on their stomach. If the child is small then laying them on your lap is best. You can also have them in a doggy position on the floor.
14. Place one or two drops of coconut oil (or similar) on their anus so the area is well lubricated.

15. Now take the hose and gently insert the tip, (one to two inches is plenty—more is not necessary) as the sphincter muscle will hold the tip of the catheter in place. If you are having any problems getting the catheter in, then you may need to add a bit more oil and/or adjust the angle.

16. Release the clamp or open the valve to allow the liquid to flow. The goal is *fill & release*.—NOT *fill & hold*. We are imitating the colonic process that is made up of gallons of water, yet in the home environment.

17. Your child will most likely hold the contents on their own—they will squeeze hard so as not to "go." This kind of happens naturally.

We do a fill and release method. The first part of the enema is usually stool. After that, we usually see parasites and or biofilm.

It's a learning experience and as time goes on, you will get the hang of it.

> *Quote from a teenager using the Protocol: "I like the enemas. Before them I am mad. But then when I get one, I feel better. My mad leaves in my poop.*

Syringe/Catheter Method

Before I started to use the gravity bag, the syringe/catheter method was my favorite. We have a YouTube video on our channel that shows how to fill the syringes and switch them out while using the catheter. If you don't have access to that video here is an explanation on how to do it:

1. Fill a bowl with body temperature filtered water and fill your syringes. If your child uses 500ml for an enema, fill all five syringes and dump out whatever water is still left in the bowl.
2. Squirt all the water from the syringes back into the bowl. You now have exactly 500ml of purified water.
3. Activate your drops of CD in a clean dry shot glass. If you are using ten drops, activate them and then use the water from your bowl to stop the activation process.
4. Pour the contents of the shot glass back into the bowl.
5. Refill your syringes with the activated CD solution.
6. Lubricate the tip of your catheter with coconut oil, and attach the first syringe.

Note: If you wish to preserve the graduation marks on the syringes, cover them with some clear plastic shipping tape. The printed on marks are often easily rubbed off and the tape will prevent that.

Fleet Enema Method

The third option for enemas is to use a Fleet enema, which can be purchased at any drug store. This method is acceptable if you don't have access to the other items but want to start immediately, as, generally speaking they are easy to find. An enema bulb is a fancy fleet enema without being prefilled with saline solution. Skip step one if that's what you are using:

1. Dump out the contents of the enema (basically saline water). Since a Fleet enema only holds 137ml, two drops of CD are the maximum number you can add. To use more water, empty additional Fleet enemas.
2. Activate your two drops of CD and dilute them with filtered water (not tap water).
3. Funnel the activated CD solution back into the Fleet bottle. Insert the applicator and squeeze the bottle until all or most all of the contents are empty.

Comments & Suggestions from Parents

Most people are surprised at how well this goes. Some parents have asked if they can do the Protocol without the enemas, and I say yes, but why? The results are just that much better with them. Enemas have, in fact, been around for thousands of years, and colon hydrotherapy is becoming a growing trend in autism recovery.

> I was really intimidated at first by the enemas.
> The syringe/catheter system, however, is brilliant. I lube the tip up with coconut oil, attach a filled syringe, and ask my son to put it in. He's 7 and very low verbal. He started inserting the tip by himself maybe on our 3rd or 4th enema with this system and I would depress the syringe. A few months later of every other day (ish) enemas, and he's doing the syringe depressing part himself. All I do is fill and swap syringes out. We are now to 11 100ml syringes, with 3-5 releases per enema (4 months into enemas). He requests enemas frequently on days I don't plan to give him one by getting the box of supplies and handing it to me.

Here are one mom's enema guidelines:

> *Be low-key about it, your attitude sets the tone. Have all your enema gear ready to go and in the bathroom for when the moment strikes. Put a nice, fluffy, washable blanket on the floor; bring in iPad, iPhone, whatever occupies your child for a few minutes. I've been doing my son's enemas after an afternoon poop if I get one because I figure it is easier as he is already in the bathroom, the bowel should be emptier, leaving more room for CD. Also the one time I did an enema close to bedtime, I swear he was up for an extra hour, however, another child might have the opposite reaction. I kneel on the floor, lay a beach towel over my lap, lay my son across me so he's on the ipad, pull his knees up a bit and then in it goes. You can put coconut oil on the anus or on the tip of enema. I ask my son to count to try to hold it in. (On the 4th enema, we only got to 40 seconds.) Then plop them on the toilet, and there you go.*
>
> *If possible, start the enema while lying on the left side, then rolling to the right. In this way, more water is able to reach the colon. (However, this may not be possible for all children.)*

CD Baths

Taking a bath with CD is another good way of detoxing your child. The CD gets into the pores and helps with skin rashes common with autism detoxification. This additional process can be added when you start with the enemas after reaching full oral dose, but avoid giving the child a CD bath right after an enema.

CD baths are done on non-enema days, unless you are doing daily enemas, in which case they are done on opposite ends of the day.

Enemas Every Other Day						
Sun	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**	**Sat**
Enema	Bath	Enema	Bath	Enema	Bath	Enema

Enemas Every Day (Enemas can also be in the morning and baths at night.)						
Sun	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**	**Sat**
Morning Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath	**Morning** Bath
Evening Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema	**Evening** Enema

The CD baths can go from as low as 10 drops to as high as 80 to 100 drops; it just depends on the person, and the size of your tub. With younger/smaller children we start with ten drops. Since we're doing baths every other day, start with ten drops on Monday, 11 drops on Wednesday, and 12 drops on Friday. Just keep going up until you get to 20 drops. The bigger and older the person is, the more drops they can tolerate.

Fill the tub to a level that maximizes skin contact.

If they start getting nauseous/listless or have loose stool, you should back down and find out what their signature dose is on the baths. Some people who have suffered from chronic constipation have done CD baths with up to 100 drops and had a BM during the bath or immediately after, without the use of drugs, laxatives, or suppositories.

CD Steam Bath

Inhalation of CD for children is generally impossible, because you just can't tell a child, especially a child with autism, "just inhale a little bit." That's not going to work. So we do the following:

In a bathtub, with the bathroom door and windows closed, plug the drain, put 20 activated drops on the floor of the tub, and turn on a hot (to tolerance) shower water. The steam will bring the CD into the air, making a really nice steamy bathroom.

When the tub is full, turn it off and place the child to soak in the water for 20 minutes. This way, the child is absorbing it through their skin as well as inhaling it through the air. This technique creates a very dispersed, low concentration dosage. I've seen a lot of children that were going back and forth to the hospital, getting steroid injections and lots of antibiotics due to bronchitis, coughs, asthma, etc., and it was rounds and rounds. This is how this particular method of steam baths came to be a very gentle way of inhaling the CD. It works beautifully; particularly if the child has a really bad cold and is congested, do the oral, enemas, and steam baths. Steam baths are great for asthma, flu, first signs of a cold, runny nose, coughing, etc., putting an end to running to the hospital or pharmacy.

CD Foot Baths

Sometimes it is impossible to give a CD bath, and in some places not all homes have bathtubs. To remedy this you can use a CD foot bath. Fill a small plastic container with hot (to tolerance) water and add your activated drops of CD.

If you have a smaller child, start with ten activated drops, and work up from there. If you have a larger child or an adult, start with 20 drops and go up. Soak feet for 20 minutes every other day, alternating with enemas, or opposite ends of the day if using daily enemas.

72/2 Protocol (Weekend Protocol)

When we give a child eight oral doses of CD throughout the day we are killing more and more pathogens. But at night when we sleep, the pathogens don't. So in order to get ahead of them, we can dose throughout the night. Imagine the progress we can make if we don't give them the chance to gain strength overnight!

Hence, the 72/2 Protocol, which involves giving one dose of CD every two hours for 72 hours straight—including the middle of the night. Here are some additional thoughts and guidelines:

- Start protocol when you pick up your child from school on Friday.
- Give them the last 72/2 dose when you drop them off at school on Monday.
- Avoid giving CD enemas or CD baths during this protocol, UNLESS they are dumping parasites, in which case you might need to reduce the amount of CD on both the oral and enema dose to insure the child doesn't have a Herxheimer reaction.
- Why not give hourly doses during the day at the usual dose? It's simply too much if you are doing it all night as well.
- Watch for improvements on Tuesday or Wednesday each week.
- Ideally, get your spouse or significant other to help with every other nightly dose.

The key is to observe your child and their reactions to the various treatments while keeping the pressure on eliminating the pathogens. You can do this every weekend or not at all. Some families are so happy with the results that they dose throughout the night for months on end.

Ocean Water

This protocol concentrates on excesses and not deficiencies. While we are using CD to reduce pathogen populations and heavy metal toxicity in the body, we need to remineralize the body. Remineralization is extremely important as vital nutrients may be lost along with toxins while detoxing. Children with autism are often mineral deficient to begin with due to these same pathogens and parasites, as well as gut dysbiosis which can prevent adequate nutrient absorption. We need to fortify our cells and tissues so that they can overcome the deficiencies and eventually return to a state of vitality.

Ocean water has helped us achieve mineralization and has led to increased brain connectivity; not to mention the parasites hate it. The importance of ocean water to this protocol cannot be understated. On page 388 of the Testimonials appendix you will see before and after lab tests proving that this protocol, when used correctly, can correct imbalances. More than one family has reported healthy nails, hair and skin as well as appropriate growth and weight gain.

We've all been taught that drinking salt water from the ocean is dangerous. Let me make it clear: We are not advocating drinking glasses full of straight ocean water nor water mixed with table salt. That would most likely make anyone sick, since it would be too much salt for the body. Rather, we are talking about diluted ocean water. Also, there is no comparison between real ocean water and just taking plain salt and mixing it into a glass of water!

The ocean is a complete source of minerals that are bioavailable to the human body. I prefer ocean water to any synthetic minerals, since the body easily recognizes and assimilates it.

Ocean water contains trace amounts of over 74 bioavailable elements including:

- sodium
- iron
- selenium
- sulfur
- calcium
- molybdenum
- cobalt
- boron
- silver
- fluorine
- potassium
- chromium
- zinc
- silicon
- chlorine
- vanadium
- nickel
- lithium
- iodine
- magnesium
- manganese
- copper
- phosphorus
- bromine
- gold
- etc.

We can use water directly from the ocean (diluted to 25%) and drink this to remineralize ourselves. However, the water must be from clean locations where schools of small fish swim and/or near ocean vortices. These locations provide safe and clean ocean water from which we can harvest. Most of us do not have access to oceans. Fortunately, clean ocean water is available commercially from multiple vendors online.

For further information about how and why ocean water is specifically beneficial to all of us, please research Rene Quinton's work on how he healed the sick using ocean water.

Using Ocean Water with CD

You can add ocean water in before starting CD, or after reaching full CD dose. We use a one part ocean water plus three parts drinking water mixed together (1:3). If we take too much ocean water, we can get loose stool. Start low and slow, 5ml ocean water mixed with 15ml drinking water and go up to tolerance.

Use the following chart as a rough guide:

Ocean Water Dosing				
Size	**When**	**Pure Ocean Water**	**Purified Water**	**Total Volume**
Child	Start at	5ml	15ml	20ml
	Go to	30ml	90ml	120ml
Adolescent	Start at	10ml	30ml	40ml
	Go to	50ml	200ml	250ml
Teen / Adult	Start at	15ml	45ml	60ml
	Go to	75ml	225ml	300ml

Ocean water should be given three times a day, with the first dose being given in the morning, after the first dose of CD. You want to separate the dose of ocean water from doses of CD by at least five minutes. Larger children can take anywhere from between 70 and 120ml, depending on the child. Some adults take up to a cup (plus dilution water) per day. It is important to watch for loose stool/diarrhea when you are increasing the dose of ocean water, as that will be an indicator that the person is at their upper limit. The ratio of ocean water to drinking water is 1:3 (one part ocean water to three parts drinking water). It is important to stay under diarrhea. So we need to add slowly, and to tolerance. If someone suffers constipation, with the right amount of ocean water that will quickly change. So low and slow as we increase the ocean water.

Some parents have reported that their children will not drink ocean water, even when diluted, due to its salty taste. You can try adding a little flavored SweetLeaf® Stevia, or if necessary adding it into a beverage or soup. As some ocean water purists say, heating it kills some qualities of the ocean water. Nevertheless, your child will still be receiving all of the minerals.

Biofilm

One of the pioneers in autism recovery, Dr. Anju Usman gave the autism community one of the critical pieces of the puzzle—biofilm. The following information is used with permission from her presentation, *Gut Biology and Treatment.*

What is Biofilm?

A biofilm is a collection of microbial communities enclosed by a matrix of extracellular polymeric substance (EPS) and separated by a network of open-water channels. Their architecture is an optimal environment for cell-cell interactions, including the intercellular exchange of genetic material, communication signals, and metabolites, which enables diffusion of necessary nutrients to the biofilm community.

The matrix is composed of a negatively charged polysaccharide substance, held together with positively charged metal ions (calcium, magnesium, and iron). The matrix, in which microbes in a biofilm are embedded, protects them from UV exposure, metal toxicity, acid exposure, dehydration, salinity, phagocytosis, antibiotics, antimicrobial agents, and the immune system.

Ok, so in layman's terms, these are microbial (bacteria, Candida, virus, etc.) communities that have a protective layer around them, which makes them 100 to 1000x harder to kill than if they did not live in these protected communities.[13]

Why are they so hard to get rid of?

- Microbes impart genetic material to one another to maintain resistance.
- Colonies communicate with one another through the use of quorum sensing molecules.
- Colonies fail to express OMP (outer membrane proteins).[14]

These colonies are fairly amazing in their development. They communicate with each other, share genetic material to prolong their own survival, and don't express outer membrane protections. This last statement is very important to understand. This is the main reason that the immune system doesn't attack these colonies, they don't express themselves as a threat, and therefore the immune system, which in many children with autism is already compromised, is not able to detect or eliminate them.

In addition to this we must be careful with what supplements we choose to give to our children. Some supplements and nutrients may inadvertently feed the biofilm. "When trying to kill bugs, if you take calcium, you may not be making headway," Usman said. "Calcium, iron, and magnesium block our efforts to dismantle the biofilm."

Why CD Works on Biofilm

The microorganism communities attach to surfaces (in many cases the gut), and are held together by polysaccharides (sugars/carbohydrates) and by iron, calcium, and magnesium. EDTA is a known chelator, and can therefore break the bonds (iron, calcium and magnesium) that hold the matrix together. Neither an antibiotic alone nor a chelator alone was effective against a staph (bacterial) biofilm. However, when combined, EDTA broke the bonds exposing the bacteria to the antibacterial agent, allowing the body to expel the biomass.

Chlorine dioxide oxidizes inorganic compounds by effectively removing their charges. Here is how it works: Fe_2+, Ca_2+, and Mg_2+ are positively charged ions that hold the negatively charged matrix together. When these substances are oxidized by CD, the bacteria are then exposed to the CD which can oxidize (kill) the bacteria, allowing the body to finally expel the biofilm mass.

Many parents using chlorine dioxide orally, and with enemas, have found biofilms in their children's stool. Biofilm can look thick, mucousy, cloudy, whitish, greyish, sometimes like pantyhose, etc.

There is some more interesting information available through the AutismPedia about Dr. Usman's work with biofilm.[14]

We had our appointment yesterday and to all our surprise when he tested her for Lyme it was negative. Just 3 months ago and before starting MMS she was testing very positive for it. He was pleased with the outcome. Thank you so much for your guidance and we owe this to the miracle of MMS and your dedication to helping me with my questions. I really, really appreciate your help in bringing us one step closer to the cure.

Also, this article "Lyme-Induced Autism Conference Focuses on Biofilm and Toxicity"[15] by Mary Budinger discusses the effects of pathogens present in the biofilm for Lyme disease and autism. The following is a notable quote from the aforementioned article by Dr. Stephen Fry of Fry Labs, which explores the possible pathogenic causes of chronic disease.

> *I could be barking up the wrong tree, but maybe not. Remember that we used to think stomach ulcers were caused by too much acid production. Then Barry Marshall and Dr. Robin Warren turned medical dogma on its head by proving that a bacterium was the cause. The pair identified*

> *the bacterium H. pylori and proved how it causes inflammation, then ulcers. Maybe in 10 years we will be smart enough to know that the 'auto' in 'autoimmune' actually means pathogen and the whole concept of autoimmunity will change. Chronic inflammation is chronic infection. In autoimmune disease, my model is that there is a chronic infection that cannot be eliminated, thus the immune system is always switched on. The self antibodies are due to apoptosis and death of host cells with host immune response.*

I couldn't agree more, and am so grateful that we have CD to combat pathogens, and restore the body to a state of health.

Mitochondrial Dysfunction

Children on the spectrum are also more likely to have mitochondrial dysfunction than neurotypical children of the same age group. The following is an excerpt from *Science Daily*.[16]

> Children with autism are far more likely to have deficits in their ability to produce cellular energy than are typically developing children, a new study by researchers at UC Davis has found. The study, published in the Journal of the American Medical Association (JAMA), found that cumulative damage and oxidative stress in mitochondria, the cell's energy producer, could influence both the onset and severity of autism, suggesting a strong link between autism and mitochondrial defects.
>
> "Children with mitochondrial diseases may present exercise intolerance, seizures and cognitive decline, among other conditions. Some will manifest disease symptoms and some will appear as sporadic cases," said Cecilia Giulivi, the study's lead author and professor in the Department of Molecular Biosciences in the School of Veterinary Medicine at UC Davis. "Many of these characteristics are shared by children with autism."
>
> Dysfunction in mitochondria already is associated with a number of other neurological conditions, including Parkinson's disease, Alzheimer's disease, schizophrenia and bipolar disorder.
>
> Mitochondria often respond to oxidative stress by making extra copies of their own DNA. The strategy helps ensure that some normal genes are present even if others have been damaged by oxidation. The researchers found higher mtDNA copy numbers in the lymphocytes of half of the children with autism. These children carried equally high numbers of mtDNA sets in their granulocytes, another type of immune

> cell, demonstrating that these effects were not limited to a specific cell type. Two of the five children also had deletions in their mtDNA genes, whereas none of the control children showed deletions.[10]

In our MMSAutism webinar, Dr. Andreas Kalcker mentions that CD changes the mitochondrial membrane electrical potential.[17] This leads to increased power in every cell in the body, including immune system cells. The mitochondria are the electric generator or "powerhouse" of each cell. When we turn off this electrical generator, cancer can develop. If the cell cannot produce electrical energy through oxidation it will use fermentation.

In practice, what I have seen is that children who suffered from lack of energy, listlessness, and other symptoms that can be related to mitochondrial dysfunction have perked up with the use of CD. They are able to do exercise again, walk without having to be carried, and can cognitively sustain activities without tiring. You will see a few testimonials mentioning "mito issues" at the end of the book. It is wonderful to think that this is another means by which CD can help people on the spectrum heal.

ATEC Statistics

Through some of our Facebook groups we have been collecting ATEC data for the past year. Just to give you an idea of what others are experiencing in terms of test results, here are some statistics:

- 246 parents have given us 2 ATEC scores. Of those, the average drop between ATEC1 and ATEC2 is 15 points. We take ATEC data quarterly, so in general that drop is in three months or less. The more impacted a child is the bigger their initial drop – kids starting with an ATEC over 100 averaged a 26 point drop between ATEC1 and ATEC2.
- Of those 246 ATEC reports, we had 32 children (13%) gain in points in the first quarter. The average gain among those 32 was 7 points. All 32 reported ATEC scores three months later, and 29 of them dropped points between their ATEC2 and ATEC3 scores. Further, 13 of those 29 dropped enough points to put them at or below their ATEC1 starting number. Which means at the 6 month mark, only 19 out of 246 (or 7%) had gained points. 93% of children who do this protocol have improved their ATEC score at the six month mark.
- We have now had 163 parents provide four consecutive ATEC scores (encompassing one year on the protocol). Of those 163, the average ATEC drop between ATEC1 and ATEC4 was 24 points.

- Of those 163, a total of 9 or 6% reported an increase in ATEC score between ATEC1 and ATEC4. The average increases was 6 points. 94% of children who do this protocol have improved their ATEC score at the one year mark.
- In a Facebook group collecting ATEC data, we have 25 recoveries as defined by the ATEC dropping to 10 or under in a year. So we have a rate of 15% recovery as defined as ATEC score under 10 in the parents who have taken this protocol through one year.

Thus showing that the longer they are on the protocol the more points they lose.

Although we all would like to see "high level" statistical research done, this is no doubt a major step in the right direction, and demonstrates that the protocol does work. But, research takes money.

DOs & DON'Ts of CD for Autism Recovery

DO review all current supplementation that your child is taking and remove all supplements which contain any antioxidants including vitamin C, vitamin E, vitamin A, vitamin K, ALA, Coenzyme Q_{10}, and colloidal silver as these will neutralize the CD, rendering it ineffective.

DO consider removing supplements that include iron and vitamin B_{12}. These are principal foods for the parasites, which explains why many children are low in them. If you are seeing a chronic iron or vitamin B_{12} deficiency, revisit the parasite chapter. These levels should normalize when parasites are no longer consuming nutrients.

DO use the Baby Bottle Method in an eight-ounce glass bottle with a completely sealing lid and start at one drop, regardless of weight. Increase by one drop/day, until you reach the optimal dosage.

DO dose frequently throughout the day. The less time we leave the pathogens to proliferate the better. Do a minimum of eight doses a day, but try to get in more doses if possible. Go for 16 in cases of PANS/PANDAS and acute situations like colds and flu.

DO activate CD for the proper amount of time for the acid you use. HCl and citric acid both activate sodium chlorite in 60 seconds.

DO supplement minerals using ocean water.

DO completely avoid the following: citrus fruits, corn syrup, fruit juices, green tea, pineapple, vitamins C, E, A, and K. Only give coconut milk and coconut water a minimum of one hour after the last CD dose at night so they will have no chance of affecting the CD. Coconut milk and coconut water are very alkaline.

DON'T give fruit juice of any form—not fresh, organic, homemade or store bought.

DON'T give highly antioxidant foods including chocolate (cacao/cocoa), coffee, green tea, kombucha, citrus fruits, pineapple, mango, or kiwi. Berries (if you have to give them at all) can only be given at night, one hour after the last CD dose.

DON'T mix anything but water with your CD drops.

DON'T give CD with food.

DON'T use a vitamin C shower water filter—yes, there is such a thing to neutralize chlorine in tap water. Unfortunately, it also neutralizes chlorine dioxide.

CD Troubleshooting Checklist

Before you decide that CD isn't working for your child please take a look at following list of common errors. If it looks like you are doing everything right, but you aren't seeing gains, please, before giving up, use one of the support options shown in Appendix 17, page 521. You are NOT alone! We have a network of parents who are ready and willing to help you.

☐ Are you preparing your CD in a clean, dry shot glass, or other type glass container that insures both chemicals are mixing?

☐ Is the CD mixture turning yellowish brown? There should be a chlorine-like smell. If there isn't, something is not right. Perhaps you have bad or incorrect chemistry? Maybe the supplier did something wrong in preparing the solutions.

☐ Are you using the correct ratio of sodium chlorite to acidic activator drops? See chart on page 94 for the correct mixture ratio depending on the type and concentration you are using.

☐ Are you allowing the correct activation time before dilution with water after mixing the sodium chlorite and acidic activator? See chart on page 94 for the minimum activation time you must wait for the chemical reaction to occur.

☐ If you question the potency of your CD, then you may wish to test its strength. See Appendix 6, page 467 for testing information.

☐ Is your CD or activator cloudy or not changing color after they are mixed? If so, there is something wrong with one or both chemicals. Check with your supplier.

☐ Did you leave your baby bottle solution in direct sunlight? A brief exposure to light is no big deal, but if you left it in a hot car for an hour, it may have lost potency.

☐ Are you keeping the same baby bottle solution for too many hours? It should be used on the day it was prepared.

☐ Are you giving antioxidant supplements or a multivitamin containing antioxidants? Vitamin C, vitamin E, etc., should not be given. Cod liver oil and fish oil supplements in general have lots of antioxidants to keep them from spoiling. Make sure those are out as well. Mixing any of these will cancel out the effect of the CD—and the vitamin. You CAN give them a few hours apart.

☐ Have you checked all your labels of supplements and foods to make sure they don't contain antioxidants?

☐ Have you removed all juices and all citrus fruits from the diet (including oranges, pineapple, mango, kiwi)? Berries (if you have to give them at all) can only be given at night, no less than one hour after the last CD dose.

☐ Are you dosing CD 30 minutes apart from food, as a minimum (one hour is optimal, but not always possible)? Grazers do 15 minutes and we dose CD 16 times a day.

☐ Are you mixing anything with your CD, such as baking soda, juice, etc.? The only thing you can add into your dose of CD is more water!

☐ Are you using alkaline water to prepare your CD? Alkaline water kills CD. Some expensive water filtration systems are designed to produce alkaline water, which we must totally avoid!

- ☐ Are you breaking The Diet? Could there be dietary infractions at school or with relatives while you're not there?
- ☐ Are you dosing so high that you are detoxing your child too fast and getting a Herxheimer reaction? Some people need to increase every three days—NOT every day.
- ☐ Are your doses too low to accomplish anything?
- ☐ Are you giving too few doses of CD? We need a minimum of eight doses per day or more.
- ☐ Have you started the Kalcker Parasite Protocol? If you have done three Parasite Protocols (PP) it is time to start looking into supplements, chelation, HBOT, GcMAF, etc.

Frequently Asked Questions (FAQs)

Can I use less than one ounce (30ml) of water to administer a dose of CD?

One ounce of water, or more, is best. You can always add more water, but at least an ounce is the minimum per drop.

We have been on the Protocol and are at five activated drops of CD in the baby bottle. My son's "usual" autistic symptoms got exacerbated to the maximum. He has horrible constipation, stimming, scripting, and he seems to be in a bad shape. Our target dose is 15 drops. What should I do?

The general protocol is to wait with CD enemas and CD baths until one is at their full target oral dose. His negative symptoms are probably from reabsorption of toxins from his unpassed stool. You need to resolve his constipation to stop this vicious cycle. You can administer CD enemas and CD baths on consecutive days. This should help the constipation, and flush the toxins out of his system. Then, once he is stable you can resume slowly titrating up to full dose.

What should I do if my child can handle "double doses" (two ounces of the baby bottle) all day with no herxing?

If your child is taking "double doses" all day, then it is NOT a double dose—it's his actual dose! Depending on his gains and what else you are folding into your protocol you can consider going up from there.

Is 72/2 only if you have done the parasite protocol? We are full oral dose but won't do PP until next full moon.

You can do 72/2 protocol before the Parasite Protocol, but you must have reached full dose of CD.

We are really struggling to get in enough doses of CD with school. We do well to get in 8; we really need more (way, way more). Night dosing is not an option (he won't go back to sleep; our mornings start early enough around here as it is), nor is coming to school during the day to dose him (we both work). Advice?

Some parents are dosing every 45 minutes to get in their doses when they are with their children. However, we need to be killing pathogens around the clock. It is not possible to be effective at killing pathogens when we are giving them 16 hours without a dose of CD. You need to get 1 to 2 doses in before school starts. Then get the other 6 (ideally more) in after school is over and in the evening.

My child hates the flavor of CD. Can I give Stevia with CD to improve the taste?

We recently discovered that some brands of Stevia are compatible with CD, CDS and CDH (i.e., SweetLeaf® & KAL® brand)—it doesn't reduce the potency. However, past tests were not so positive. Some added ingredients may be why this was so in our initial tests. Try to use only pure Stevia or check with the Facebook groups on what others are using successfully.

SweetLeaf
Sweet Drops
Sweetener
STEVIACLEAR
KAL
PURE
Stevia
ORGANIC
extract

Preliminary tests show that some brands of Stevia can be used with CD to improve taste without losing potency. The two brands we have tested in-house are SweetLeaf® and KAL®. Some of our moms report having successfully used flavored versions of SweetLeaf®.

When you say do two back-to-back baby bottles of 11 drops... you mean do a total of 22 drops in 16 hours? Correct? I'm going slowly up on oral CD because of behaviors. If I go from 11 to 22 drops overnight, won't he have strong detox and behaviors?

This is spread out over 16 hours or more, such that every 60 minutes it is the same amount. It is different than putting 22 drops in 8 fl. oz. Behaviors associated with bacteria (PANDAS/PANS) are often reduced with the tolerated dose spread over longer periods of time. Remember that CD is only active in the body for about one hour.

My child seems very uncomfortable. What should I do?

If you see a detox like this, stop dosing for the day. The next day, return to the last dose when the child was stable. Pathogens are dying and toxins are being released into the bloodstream on their way out of the body. This causes discomfort in some people.

How do you know if I should go past full oral dose of CD?

A Herxheimer reaction will be your biggest indicator if you have gone too far. If you have reached full oral dose and it seems like your child has plateaued, please schedule a consult with Kerri to discuss your options. Note: The full oral dose chart (page 101) is only a general guide and not a one size fits all.

Is an hour enough time to separate CD neutralizers from the CD dose?

CD is only active in the body for about one hour, however, supplements like vitamins A, C, E, K and ALA, CoQ10, and GSH (glutathione) need to be avoided. Also, orange juice, pineapple juice, and other high antioxidant fruits are a problem even after one hour. So, the best thing to do is eliminate everything that kills the CD. If not, it is pretty much like not using CD. I have seen it too many times and it's not worth it. Antioxidants destroy CD. There is no period of time that is safe to use antioxidants with CD.

I know you have to start the parasite protocol on a certain day of the month, but how about the first dose of CD?

You can start giving CD any day of the month. That said, it might be easier to start on a weekend, or a day when you will be with your child all day to make sure to get in all of their doses, and keep an eye on them to make sure they tolerate it well, i.e. no Herxheimer reaction.

What is 72/2?

72/2 is when you give a dose every two hours around the clock for 72 hours. We use a baby bottle and a half to get through the night hours. This can be added in once you are comfortable at your full oral dose and have added in enemas and baths. See page 115 for more information.

Do you recommend 72/2 for everyone?

Once a child is at 100% of their target oral CD, enema, and baths then they can do a 72/2 weekend. They are exhausting for the parents, but some kids do wonderfully. A lot of families get a nice boost. But, if you do a couple of weekends and you don't see anything afterwards, then forget it. The benefits are usually seen two to three days after the 72/2 is over.

My son suffers from constipation. Can I start the enemas before I get to his full oral dose?

Anyone who suffers constipation needs enemas right away. These can be combined with oral doses. Still start low and slow, but it is important to get the bowels moving with any detox protocol. When constipation is present, we do not wait to start enemas.

What is the maximum amount of water to use with enemas? How many times do you hold and release?

First of all, we don't encourage holding. Rather we do a fill and release, which is kind of a home colonic. If your child can hold for a few seconds without fuss, then great. The number of cycles depends on the results of each. If all you are getting out is water, you may be finished, unless the person is seriously constipated. The amount of water really depends on the individual's size and the number of cycles administered. My basic philosophy is: *Any enema is a good enema.* So, even if not much comes out, you are still getting some water and CD in.

Is it possible that my son has been getting rid of mucous after his CD enemas? What is up with the mucous? Why is it there in the first place? And, is it a good or a bad thing that it is coming out?

Mucous is always a sign of inflammation and can contain pathogens and parasites. Mucous gives a home to pathogens. The inflammation and mucous can be caused from the allergies and pathogens. It is common to see mucous come out in the beginning of the enemas since our children have so much intestinal dysbiosis.

What is a good enema method for a 17 year old boy? I am having trouble understanding the enema part of the book. The video in the file really helps, but I am having a hard time imagining doing it on my son. How do other people with older children do it? Can it be self-administered eventually? Kerri mentioned a gravity bag, so you don't have to keep changing syringes. Has anyone tried that? Where do you get the bag?

Here is some advice one mom gave to another mom on one of our boards:

My son is 14 and tends toward constipation. I had done suppositories in the past to help him have a bowel movement. So, I told him that the enema was "fanny medicine" to help him have a BM and make him better. I started with just showing it to him after a BM after I wiped him. Then, I just put it next to his anus the next day. Then closer every day. He was still constipated so just inserting the tip would help him poop and then later I was able to push the liquid with him standing next to the sink and it would run all down his leg. After several weeks of this, it was a full moon and he could not poop at all and was lying on his bed in pain. I said let's try it here because it will really help you. I got out a hundred towels and did it right there and he pooped. It was a slow process over two months but now we do them AM and PM always right after the shower. I counted over 20 feet (of parasites) coming out this last PP. And that is just what I could see. I don't dig, just pull out obvious ones that are a foot long. Main advice is keep calm with soothing voice and have tons of patience and keep telling your son how much you appreciate his patience too. I also tell him this will get rid of his autism. He is non-verbal but said with RPM that he is mad he has autism. Also, the soft tip is key. I use syringes because they go fast.

My son is not having daily bowel movements. What should I do?

Do CD enemas every day until the stools normalize to daily. If you absolutely cannot do enemas, use CD baths in high doses (50+ drops depending on the size of the tub). Ocean water in high doses will move the bowels as well. Castor oil also helps.

I am scared of the CD enemas. Will CD harm the intestinal lining?

CD stays active only for about one hour and can gently, but effectively remove biofilm and kill pathogens in the intestine. Our society has shied away from this healing method in the last few decades. However, CD enemas have been proven to be positive turning points in the treatment, time and time again. Lots of kids will actually ask for enemas, as they provide them relief and comfort. If you are nervous, apply an enema to yourself first. You will see it is no big deal, and feel more at ease applying one to your child. Some parents have shown their higher functioning children how to apply their own enemas, so no one has to be there, and to make them feel more comfortable.

What is the difference between colonics and enemas?

Colonics cleanse the entire length of the colon while enemas cleanse the lower part of the colon. Colonics involves multiple infusions of water into the colon, while enemas involve a single infusion of water into the colon (which can be repeated). With colonics, fecal material leaves the body via a tube. Colonics normally involves going to an office and receiving assistance by a trained colonics hydrotherapist, while enemas are free and done in the privacy of your own home.

I have never done an enema before. How do you administer it?

There are several positions that work well, but there will be one that works best for the individual. The enema instructions are on page 109. Some examples would be on all fours or lying on your left side on a towel, or a small child could lay across your lap.

Enema clean up question: What are we boiling? What are we spraying with *Everclear*? What are we throwing away?

After you are finished looking for worms you can throw away any plastic plates/forks you may have used. Pour boiling water on your specimen collector, enema nozzle and catheter. Then spray with *Everclear* grain alcohol. If you are using syringes, you may pour boiling water over them as well.

How far do I need to insert the catheter for an enema?

Not very far at all; one to two inches is plenty. The anal sphincter is about one inch past the anus. As soon as the catheter passes this muscle, it will hold it in place and there is no need to go further.

What is intestinal biofilm supposed to look like?

Biofilm is fluffy, cloudy, and mucousy in appearance. There are no rules on the color. It can look like pantyhose. But, it is usually mucousy and a cloudy whitish/grey color.

How far away should we be doing enema from food? If we miss the morning enema, can I give it right after lunch?

They can go together; there is no need to separate enemas from food. Any enema is a good enema.

I'm having trouble with my syringes (for enemas). One plunges really smoothly, but the other two are almost impossible to move. Do I need to grease them up somehow before each use? How do you clean them?

I use *Everclear* grain alcohol to clean them, but note, it will dry out rubber parts. You can use coconut oil to lube the plunger gasket.

Is it normal to see black specks coming out after using enemas? We saw white specks for a couple of days too.

Black specks can be heavy metals. Oxalates are known to bond to metals and would come out with metals. White dots are generally parasite eggs.

Does ocean water need to be refrigerated after opening?

No it does not, it is fine at room temperature.

If I harvest my own ocean water do I have to sterilize it?

No. Sterilizing is not necessary or desirable in our experienc—we don't want to denature it. Pouring it through a coffee filter to remove any potential particulate matter is fine. Do NOT use charcoal filters (such as a Brita® water filter), because they will alter the characteristics of the natural ocean water. Just make sure to collect the water far from any harbours or large river outlets.

I bought pure sea water in Australia. It sounds like a great mineral drink. Would this be beneficial for killing parasites or just putting goodness back into the body?

Actually, it does both.

How do I avoid feeding the pathogens?

CD kills pathogens. I would use CD to rid the body of pathogens rather than try not to feed them.

Is it possible to use CD, CDS, or CDH during pregnancy? If it is ok, how much can be used?

CD/CDS/CDH are not advisable during pregnancy or nursing.

Can I add CD to breast milk to the bottle to help the baby detox?

CD must be given in water. Mixing breast milk and CD will reduce or potentially cancel out the potency of the CD.

How many drops go in a CD bath?

It depends on the weight of the child (20-100 drops). If he is a little guy I might start at 10 or 15 activated drops and work up over the course of the following couple of weeks. The size of your tub also plays a part. The bigger the tub, the more CD you are going to need. See page 113 for more information.

I've heard that some people mix CD directly in capsules and then swallow them. Can we use that on our children?

NO! Not a good idea! Since many children with autism can't speak, they won't be able to tell you when the capsule gets stuck. CD needs to be diluted in water before taking as instructed in this chapter.

Can I use warm water instead of cold water to store CD and how long will it remain potent mixed in the baby bottle?

CD is a gas dissolved in water and therefore has to be preserved in a sealed container. If not, it will lose its potency, similar to how soda goes flat if the cap to the bottle isn't screwed on tight. Colder is better because the CD will gas out quicker at higher temperatures. If you have doubts about CD potency, get the Lamotte ClO_2 test strips to be sure. See Appendix 6, page 467 for more information.

Are purified, distilled and filtered water all the same thing?

They are different. However, we can use purified, distilled or filtered water. But, alkaline water should never be used!

Must the activated CD water mix be stored in a glass bottle, or is a no-BPA plastic bottle with an airtight cap ok?

I prefer glass to any plastic. Over time, chlorine dioxide can degrade plastic, and that means you or your child will be consuming it. Yes, the bottle needs to remain sealed. Plastic is acceptable for lids only! Metal lids should never be used because they quickly rust, even if covered with a plastic coating.

Is it a problem to transfer one ounce of CD to a steel bottle briefly before drinking it?

That is fine for a transfer, but I wouldn´t store CD in a steel bottle.

Is the GF/CF/SF "clean" diet essential to the CD protocol? My son doesn't have bad reactions and I just do not see how we could implement it.

A healthy, clean diet is very important to the healing process. Unfortunately, our "standard diet" is filled with preservatives, colorings, and other potentially harmful chemicals. When the goal is to heal, the diet is necessary. Dairy especially can cause inflammation and mucous production. Mucous provides protection to parasites.

What is a good CD-friendly multivitamin?

All multivitamins have antioxidants. What we need to supplement during detoxification are the minerals. Ocean water has 90 bio-available minerals, and it is what I prefer for this protocol.

Are there other specific foods, veggies or fruits that need to be avoided in addition to juices?

When you are following the CD Protocol, all antioxidants need to be avoided. Chocolate is a strong antioxidant. As far as vitamins are concerned; A, E, K, and ALA are off limits. High antioxidant supplements such as curcumin have also been a problem for many people. In the case of vitamin D, it is not an antioxidant, but it has caused aggression in previously calm children. Therefore, as a supplement, it must be used with care, if at all.

Vegetables and legumes which are high in antioxidants have not shown to be very detrimental to the effects of CD. However, citrus fruits, pineapple, mangoes must be avoided. Berries are best given at night, one hour after the last dose of CD, if you must give berries. Juice, any sort of fruit juice, is completely prohibited. Not only can the antioxidants involved kill CD, but the high sugar content (albeit natural) shuts down the immune system, which is already impaired in children with autism.

Can our kids drink anything other than water for the duration of the CD protocol or just not drink juices down with the CD?

Some people are not used to drinking water and find it difficult. Some people drink nut milks or rice milk. On occasion, some folks put a piece of fruit in the blender with an apple and blend it, strain it, water it down and have that sometimes. Juicing is taking 5 apples to make a glass of juice and that is a lot of sugar (albeit natural) as well as a lot of antioxidants. Antioxidants kill CD and sugar slows the immune system. You want to totally avoid alkaline water.

Can I use rice milk to give the CD instead of water?

Water is what we need to use to give the CD. If you are giving CD mixed with another beverage you are reducing its potential. By combining CD with liquids other than water we will not get the desired results. Rice milk might be okay, but you would need to test how it affects the potency of the CD. You can get the *LaMotte High Range Chlorine Dioxide Test Strips* (#3002) at www.amazon.com. This way you can see for yourself how mixing CD with any beverage affects its potential. However, I do not recommend putting CD in anything but distilled/purified water.

Can I give my child fresh pressed green juices (celery, kale, cucumber, apple)?

Anything that is very nutritious to the human body is nutritious for the pathogens/parasites. So, a healthy green drink is very healthy for the pathogens. If you are giving "green" juices make sure they are not loaded with sugar from apples, carrots, etc. In my opinion, it is preferable to avoid these for the first few months of the Parasite Protocol (PP), at least.

Can I use coconut and almond milk with CD?

You can drink coconut and almond milk 60 minutes after your last dose of CD, but I would not combine anything other than purified water with the dose of CD.

Can my child drink coconut water if he is taking CD?

If you want to give your child coconut water, give it one hour after your child's last dose of CD, at night, before bed.

Does lemon interfere with CD, even after it is cooked?

Yes, in some cases. A family of a child who suffered from SIB noticed a direct correlation with lemon use in their meals with the return of his SIB. I would avoid it at all costs.

My son barely drinks anything, especially all at once. This is why we have been unsuccessful with detox programs. Is CD different?

You should make sure that your child is well hydrated, but for the CD alone one ounce (30ml) of water is sufficient. I find that it is easier to drink CD cold rather than at room temperature. Also, if the smell is a problem, put the dose in a plastic syringe and squirt it right into your child's mouth. This prevents them from having to smell it before they swallow it. Enemas are a great tool for hydration.

My child's appetite has changed since starting CD. Is this normal?

Yes. On average larger kids tend to slim down, while thinner kids tend to gain some weight.

Can I give green tea or kombucha while taking CD?

That is out of the question in my humble opinion. Green tea is high in antioxidants and caffeine. Kombucha is similar. I don`t know exactly how long the antioxidants in either one are active in the body, so with autism, if we are looking for healing, avoid them both.

Can I use original Nutriiveda™ with CD?

Nutriiveda™ is derived from whey, a dairy product. In humans, all dairy causes inflammation and mucous, which provides a protective refuge for pathogens. That makes this product contraindicated for this protocol, plus it contains antioxidants.

Can I use Chia seeds with CD?

Yes. But do not mix them directly into the CD.

How do I dose my child when he/she is at school from 8:30 am to 3 pm? I work full time so "dropping in to dose" is not an option.

Give a dose upon waking, one at the door of the school at 8:30 and the third dose of the day at 3pm at the door of the school. The last dose of the day is at bedtime, there should be at least four hours between school pickup and bedtime to get in the other four doses. Once you get the hang of it, it is not that hard, it just takes a little getting used to and some prior planning.

What should I do if my son goes away for 5 days with school? Do you think it is possible to prepare a bottle of CDS for 5 days and he takes only one portion each evening?

It's not ideal, but better than nothing.

My son is refusing the CD. What do I do?

Have you tried adding more water with the dose, using cold water, using a syringe or a straw? Here is some advice from a Mom whose daughter refused oral CD:

We just went through a REALLY rough patch with my daughter when she absolutely refused to take oral CD, and I know it gets super stressful for both you and your child if you try to force it… I had to respect her as clearly she was telling me in her own way, "Mom, this does not make me feel good!" for whatever reason and after maybe a week we were back on track. During that time I was able to give her a few doses (dropped from 14 drops of CD to 5) here and there, during really good moments. For example, in the bath or swing that she loves and I think that helped her associate CD with feeling better again. You WILL be back on track in no time!!!!! Good luck! Keep calm and dose on!

I am on 30 activated drops/day with my teenage boy. The taste is simply horrible. He refuses to drink his doses. What should I do?

You can dilute your baby bottle as far as a liter bottle if necessary; it will not impact its effectiveness. Obviously, one dose will be more like a half a cup or a full cup in this case. However, be mindful of your conversion rate so that you get your dosage right. Cold water helps with the taste. You can add approved Stevia brands to cover the taste of CD. See page 126.

If the country I live in doesn't have glass baby bottles, can I use a plastic bottle?

No, do not use plastic for storing CD. If the cap is plastic, we can live with that. In almost any country you can find a glass water bottle (with a plastic cap). You can use a shot glass to measure out eight ounces, and even mark them on the outside of the bottle with permanent marker. Many parents in Venezuela do this, as neither glass baby bottles nor LifeFactory® bottles are available in their country. NEVER use a Stainless steel (or other metal) bottle or container!

Do you think this Protocol could work for a child who already has a low ATEC score (18)?

Yes. I started with a 12-year-old, who last February 2012 had an ATEC of 18. He did not budge from there after years of biomedical interventions. He could not stop with fears, phobias, anxiety (all parasitic in nature) and we started him on CD in February 2012 and the Parasite Protocol in June 2012. His ATEC is now a one!

My son's ATEC is not going down as I had hoped it would by now. It has hovered around 65 the past few ATEC's. I think it is probably due to the fact that if I am not home my husband often forgets to give him his doses or his PP meds. Does anyone have any suggestions on how to stress the importance of this protocol to him?

The following responses are from moms on our public Facebook group:

You will never know what following the protocol will do for your son until you follow the protocol. When I first started the protocol a little over a year ago, I remember I had been using a special juice that so many people were raving about and I was paying $140 monthly for this miracle juice and I really wanted to keep in in my son's protocol. I asked my husband to give it to my son at 4 am when he got up, and that was 4 hours before he would get any CD doses... how could that effect our results? The juice was given 4 hours before starting CD for the day, we were following the protocol exactly otherwise, and even doing the enemas. Well, nothing happened, nothing, until about 3 weeks later when I said, well let's stop the juice and see... BAM! Immediately my son's ATEC results dropped and he blossomed. I mean not only did his constant pacing stop, but his eyes lit up, he not only started talking more, more, more but even his laugh changed to a typical sounding teenager's laugh. It was incredible. His ATEC dropped instantly and kept dropping, 68, 25, 13, 7, 5, 3. It's a 6 today but this a super bad new moon for him. You will never know until you follow the protocol.

We have also seen our son stop pacing, stop suffering minute to minute, in the beginning we made our mistakes, missed a dose, didn't get this or that done, we now stick to it, the cumulative effect of staying on protocol means the difference in drops in ATEC scores and staying the same or going up, it did for us anyway. I could feel the slipping, we are still working on getting this right, it is an evolving protocol that requires the entire family's buy in. I have kept my son home from school to make sure he is dosed properly but that's just us. In November he is supposed to begin going to school full days, if they do not dose him correctly, I will hire a homeschool teacher, that's how important this is. You will not get your child well, Kerri told us 16 doses a day, he got better, she told us to start *enema's early, that made a big difference, we started our first PP before he was at full dose, it's all made a huge difference, email her, keep watching your child to see what's working, this protocol works, it may take time, but what is the alternative????*

My husband is not good about giving meds either. When I'm gone my older girls give it. I text them reminders. Sometimes setting a timer can help. I also set all of the meds that go along in a morning and night dose daily case. Often when I organize it all for them it's much easier for the meds to be given properly and in time.

My child has horrible nasal congestion. Can CD help?

Yes, you can use CD and make nose, eye, and eardrops. Put one activated drop of CD into one ounce of water. Use one drop of the mixture in the nose (eyes or ears) every 15 minutes till symptoms disappear. NOTE: DO NOT USE PURE ACTIVATED CD WITHOUT DILUTING IT!

You can also do a steam bath, which is done by closing off the bathroom (windows and door shut), and putting 20 activated drops on the floor of the tub. Run hot shower water on top of the drops until the tub fills; the air will be full of CD smell. Now, turn off the shower, put the child in the water to soak for 20 minutes, where he will breathe the light CD air.

Is there a seasonal-allergy protocol for using CD? I suffer from extreme itching in my eyes, redness, puffiness, sneezing, sinus congestion etc. Will CD provide relief quickly or does it need to be used over a long period of time?

Being on full CD, ocean water and the parasite protocol the allergies should begin to fade away.

Can I try the CD ear protocol if my child has not started oral doses?

Yes!! Please start the eardrops, if there is an infection, every hour until the symptoms disappear. If you catch an infection at the onset of its symptoms, give one drop of the mix every 15 minutes. We have seen earaches clear up in a couple hours. In any case, work towards a full oral dose, at least one drop eight times a day to combat the infection on all fronts. The eye, ear, and nose protocols are all the same: one activated drop of CD in one ounce of water in a sealed dropper bottle. Apply one drop of this mix every hour until symptoms diminish.

Can CD help arthritis?

Absolutely! CD is amazing for arthritis. Include CD baths in your routine. Use CD in conjunction with DMSO.

Our son has a runny nose and sneezes a lot, but has no fever. He hasn't been sick in well over a year! I'm wondering if it is from doing CD.

This sounds like a very typical die-off reaction. Monitor his progress, make sure he is hydrated, and continue with the Protocol. If his reactions remain manageable, you don't need to back down or change much. This could also be parasites, depending on whether the symptoms are chronic, etc.

Has anyone seen kids who had never been sick start getting sick as they heal?

Yes, it means immune system is waking up and fighting back against the crud they've been carrying around for a long time. It's called a healing crisis and is not uncommon and is actually a good thing.

Does CD, CDS or CDH work on HIV? Also, are there any contraindications of CD and HIV drugs?

We do not focus on HIV or its treatment and therefore we cannot give you a definitive answer. Anecdotally, there are many testimonials that indicate CD having a positive impact on the disease. In addition, we have spoken to doctors in various parts of the world who use CD to treat their HIV/AIDS patients and have positive results from what we understand.

Can I use CD in a Netti pot when doing an ear, eye or nose wash?

No. Do not use a Netti pot with CD. Instead, use the steam bath method. See page 114 for more information.

I´m wondering if CD can be used in a nebulizer? With our daughter's chronic mycobacterium issues, we can never get rid of the cough. I thought maybe it would be helpful to get it into her lungs or is this not safe?

Try a steam bath. Inhaling CD directly has to be done very delicately. This is why the steam bath is recommended, as the particles have time to disperse into a very large area rather than being inhaled directly. Some people are using humidifiers with 35 drops of CD per gallon. See page 114 for more information.

What supplements do I need to avoid while using CD?

Antioxidants (vitamins C, E, A, K, ALA, and CoQ10) need to be avoided because they kill CD. We avoid iron and vitamin B12 because parasites feed on these. So if you do not have something in place to kill pathogens/parasites, like CD, then the majority of your supplementation goes to strengthening the parasites instead of your child.

I really don't understand something basic... Vitamin C in food. I know not to give citrus or mangoes. Also, kiwis are rather high in vitamin C. If my child eats a kiwi or other high vitamin C containing food, how long does this impact CD? When will it wear off?

Kiwis and other fruits high in vitamin C are a tough call. We do our best to avoid the biggest offenders. Unfortunately, I have seen families lose months because they continued to give fruit juice or supplements high in antioxidants. As soon as these were pulled, their children began to improve. A few years ago, a popular omega high in antioxidants became popular for autism. Families that I was helping all over the world noted regression in their children. As soon as we figured out what the common thread was, and pulled that supplement, the improvements began again, but it was a nightmare. To answer the second part of your question, I don't know exactly how long vitamin C from a natural fruit source lasts in the body. A few different websites estimate up to 24 hours.

We have been using gluthatione every day and modified citrus pectin as a binder. Are those compatible with CD?

Gluthatione is a very powerful antioxidant and cannot be used as a supplement with CD. Modified citrus pectin is fine.

I have heard that CD helps with oxalate problems. Is that true?

Rompepiedras (RP) is very effective against oxalate-crystal formation. There are connections between parasites and oxalate problems, which this protocol seems to be helping. Ridding the body of parasites helps with high oxalates.

I know that CD neutralizes heavy metals. What does that actually mean? What is the difference between chelating and neutralizing?

The mechanism of how CD neutralizes and removes heavy metals such as mercury is not clearly known (or understood). However, a number of recovered children have had heavy metal burden tests performed before and after the use of CD, and their heavy metal burden dropped. We also know of an adult in our circles who was extremely toxic with mercury and was able to normalize his levels after all other medical interventions had failed him.

So is CD enough and we don't need to chelate?

For many it is not necessary to chelate while on CD. There are some children though who do IV chelation and CD at the same time. This method seems to have good results; there are big metal dumps, when administering chelation challenge testing. As time goes by, we add interventions until we reach recovery.

Are all chelators compatible with CD?

In many cases there is no need to do both. To answer your question, EDTA is fine as well as DMSA and DMPS with CD.

You may want to consider Bio-Chelat™, bentonite clay or zeolite clay. These are great light chelators.

We have tried every biomed option available, but nothing has helped. What makes CD different?

Biomedical treatments in general provide large amounts of antioxidants. CD is an oxidizer and kills all pathogens in the body. The premise is that once your body does not have to feed and house pathogens, it will work much better and will heal. This protocol focuses on excesses and not deficiencies, which in itself makes it different from many other biomedical interventions. With CD and the Parasite Protocol we are eliminating the pathogens and parasites that cause the symptoms known as autism.

Most biomedical interventions seem to be only applicable to small kids. Does CD work on "older" kids, teenagers or young adults?

Of course! Families with older teens and adults have had success. The body wants to heal at any age be it 7 or 70! This year, two 17-year-olds lost their diagnosis through the use of CD as part of their biomedical protocols. A 32-year-old male is doing excellent on the Protocol as well. Time will tell, and parent/caregiver dedication is the key to success.

How long should I do the protocol to see if it is working for us?

Within 30 days you will know. If you do the Protocol 100%, and in the order laid out here you should see improvements. Some see improvements with their first dose!

Has anyone had an increase in PANDAS/PANS stuff while on CD? Our ATEC didn't budge this time around and it is mainly due to the OCD and other PANDAS stuff that has increased. I get in 16 doses on weekends, but not on weekdays—maybe I get in 12 on a good day during the school week. Has anyone done anything to treat this PANDAS stuff?

Parasites typically coexist with bacteria, so it is very important to treat bacteria and parasites at the same time. It is extremely important to get in 16 doses a day, whether you get in doses earlier in the morning, or send doses to school. Also, many children with PANDAS/PANS need to be treated month long for parasites. If you are giving a probiotic, you can try doing a period without it and see if you child does not improve. We must keep in mind that if you are seeing parasites coming out with the enemas, we may not be seeing gains until we have eliminated a good number of the parasites living in the body. The toxins that they excrete—living and dead—can also cause some of the behaviors we relate with PANDAS/PANS.

Has anyone seen elevated liver enzymes while on CD?

Liver enzymes will not be high from CD. I have watched people do the liver enzyme tests for two years and the kids are in the normal range. However, when they are dumping worms or doing parasite meds is when we may see the enzyme levels go up. Within a couple of weeks they go back to normal, in the handful of times that I have seen them go up. Mebendazole and Combantrin® are non-systemic, and therefore are not absorbed; they basically travel through the intestinal system and are eliminated in the stool and urine.

My son still has a lot of mercury (DMSA provocation). Is it possible that the level can go down from using this protocol? How long should I wait before testing him again?

It is very common for the heavy metal load to go down. If you want to do labs, you can do them in three to four months.

How young is too young for CD?

There is no age that is too young. If a baby is showing signs of cold/flu, etc., you can start with the baby bottle method, one drop in 8 fl. oz. of water, so they are getting 1/8 of 1 drop per dose. The protocols are based on weight.

Ok we have had my son on this protocol for 11 months now. We started at an ATEC of 82 and had a dramatic decrease right away. Our second score was 26! Now we have had 3 consecutive ATECs of 33, most points being in the speech category. Where should I go from here? More diet changes and or supplements? We have been GFCFSF for years, considering going grain free next. We already use omegas and ocean water, and have done nine parasite protocols.

It may be time to look into hyperbarics or GcMAF, as you have *The Diet*, CD and parasite protocol firmly in place. If these 2 options are not financially possible, you need to look at the speech supplement list. Go in order, and see if your child doesn't see gains from some of those. You can also look at chelators. Also, please check that your omega does not have any antioxidants in it, as that will be killing your CD.

Is it normal to see diarrhea in the beginning of the CD protocol?

I consider there to be two types of diarrhea. One is the water faucet, which we want to avoid. If you see water faucet diarrhea, stop for the day, and give a lower dose of CD the next day, when you start up again. The other type of diarrhea is loose, unformed stool, which is normal during detox. As the body is attempting to eliminate the excess toxins quickly, the digestive process will speed up, and not all of the excess water will have a chance to be absorbed through the intestinal tract, causing loose stool.

How do you move CD through inspections at the airport?

Don't carry concentrated sodium chlorite and activator bottles into the cabin. They must go into your checked luggage. Make sure each bottle is tightly sealed. Double bag each bottle SEPARATELY and place at opposite ends of your luggage but not right up against the sides. You may wish to stick each bottle into a pair of socks as extra protection. Surround each bottle with plenty of clothes to insure they will be protected if the bag is roughly handled. According to TSA regulations, you can bring aboard up to 3.4 fl. oz., (100ml) of a liquid in your carry-on luggage, which can hold 3 doses from your baby bottle batch.

Can I use CD as a toothpaste? Should I use regular toothpaste in addition?

I use CD spray (ten drops of activated CD per ounce) on the toothbrush first. Then, brush the teeth, and follow with fluoride-free toothpaste. CD is great for healthy teeth, tongue, and gums.

Is it okay to go swimming when on CD? I worry about the chlorine.

They do have a molecule in common; however one has nothing to do with the other. Some doctors don't let ASD kids swim in chlorinated pools. But, as far as taking CD goes, the chlorine in the pool will not deactivate the CD in your body, nor react with it. A child on CD can swim in a chlorinated pool, just as a child not on CD can swim in a chlorinated pool.

Can I give colloidal silver with CD?

No. Colloidal silver is active for up to 24 hours in the body, and will lower the potency of CD. Therefore the two are not compatible.

Where is the autism clinic in Venezuela?

Puerto Ordaz, Fundacion Venciendo el Autismo. Carolina Moreno is the president and one of my best friends. Her email is venciendoelautismo@hotmail.com. We have over 36 recovered children there. It is wonderful.

My son has been sick and has really bad diarrhea. Can we skip the enemas for a few days or still do them?

We don't use enemas just for constipation, but for overall colon irrigation and health. CD kills pathogens in the colon, and thereby helps heal autism. As the sickness causing the diarrhea is most likely pathogen induced, CD enemas will continue to kill these pathogens and help your son get over this acute situation faster.

My daughter just drank undiluted CDS. What do I do?

First, get her to drink plain water right away. Next, give orange juice or vitamin C to neutralize the CDS. Give burbur or activated charcoal to mop up the toxins that were released. Considering the strong taste, it is unlikely she would have consumed very much.

Can CD be harmful?

In the past 3 years that families have been using CD for autism, we have not seen any children or adults harmed by using this protocol correctly. Liver enzyme tests, nutritional tests, metal porphyrins tests, etc., have consistently shown improvement in the health of the children. Since Jim Humble started using chlorine dioxide for health, no one has died from ingesting CD (or MMS). The story that is often shared around the internet from Vanatu, concerning a woman who unfortunately passed away, was not attributed by the coroner to be the result of ingesting chlorine dioxide. There was another case of a 25-year-old male who attempted suicide with 10g (nearly the contents of a four ounce bottle of sodium chlorite at 22.4% solution) of sodium chlorite (inactivated CD). He developed methemoglobinemia, which he received treatment for, but survived his suicide attempt. That said, as we have mentioned before... anything used incorrectly or recklessly can harm you. Water, table salt, etc. The reason we outlined the protocol in such detail is to prevent mistakes and help families use it responsibly to recover their children with autism.

Does clay or bentonite clay interfere with CD? How far apart should it be taken from CD?

One hour apart from CD is fine. I use diatomaceous earth about ten minutes apart from a CD dose.

Is ibuprofen ok to take while doing CD protocol?

We have never seen a drug that was contraindicated with CD. Always consult your physician for prescription contraindications.

Is it ok to add Epsom salts to a CD bath?

We do not use epsom salts baths in this protocol. We use CD baths. See page 113.

Is it ok to take SAMe (S-adenosylmethionine) with CD? I took out all antioxidants, but not sure about SAMe.

No, it contains magnesium and vitamin C. For a mineral supplement, we take Ocean Water. Vitamin C kills the CD. Magnesium feeds biofilm, and therefore is counterproductive to our goals of healing.

Can I use DMSO in a CD enema to help drive the CD into the parasites?

Warning: Never use DMSO in an enema! Never introduce it into the body rectally. If it is applied rectally, it will carry toxic fecal matter into the bloodstream through the intestinal wall.

How do I use DMSO?

Every chance you can, apply DMSO to a clean body for 20 to 25 minutes and let it soak in. Apply it to affected parts first then remove rings/jewelry to cover all parts of the hands. Allow it to dry while applying to another body area (i.e. right arm, left arm, right leg, left leg). Apply to clean skin with clean hands. Use natural fiber clothing if it is going to come in contact with the DMSO, which can dissolve synthetics. Dr. Stanley Jacob has proven that DMSO brings about healing of rheumatoid arthritis, and it is great for migraine headaches, etc. You can research his work at...

www.dmso.org

...or find a DMSO retailer at...

www.protocolsuppliers.com

The 99% pure stuff is best, but never use that strength on the skin directly. It must be diluted to 70% or less for topical use.

Does DMSO neutralize CD?

No, it does not. However, DMSO has not proven itself to be a tool for autism recovery. In general, we use it for self-injurious behavior (SIB). The "S" in DMSO (dimethyl sulfoxide) is "sulf" as in sulfur. Many children with autism have a spirochete (amongst other pathogens), which feeds on sulfur. We saw many setbacks when adding DMSO. It is used on a case-by-case basis.

Chapter 6

CDS

Another Way of Delivering Chlorine Dioxide

There is more than one way to skin a cat.

~ Seba Smith

In October 2011, I received an email from Jim Humble. After a visit with Dr. Andreas Kalcker, he had some fantastic new information for healing autism. Jim wanted me to fly to the Dominican Republic to see him yet again that year so he could share this new information.

The big news surrounded the creation of *Chlorine Dioxide Solution* (CDS)—a new method of producing and using chlorine dioxide; substantially different from the CD mixing process we were using (described in Chapter 5). Jim was excited about all of the great things that Andreas was developing.

Andreas was contacted by a cattle farmer who was frustrated by the health problems of newborn calves he was receiving. These calves were suffering from infections, diarrhea, ear problems, cysts, coccidiosis, bovine respiratory syndrome, etc. His yearly veterinary drug bill was around €28,000, not including the cost of feed. The farmer had heard about the wonders of MMS in humans and wondered if it might help his animals as well.

Andreas thought, "Sure! Why not?" However, he quickly learned that cattle were very different from humans when it came to their digestive system. Cattle digest through fermentation, which CD disrupts, so he had to find a way to bypass their digestive system and go directly to their blood stream. But, that idea had another big problem. The pH of standard CD was too acidic and therefore not compatible with the calves' blood stream. Injecting even dilute CD caused great pain and could result in damage to the veins. There had to be another way!

After much thought and research, Andreas came up with a distillation process to extract the chlorine dioxide—the key ingredient—from the CD mixture so it would not contain any sodium chlorite or acidic activator. He named the resulting liquid *Chlorine Dioxide Solution* (usually abbreviated CDS). Note: CDS can still contain trace amounts of sodium chlorite and whatever acidic activator used to produce it, but usually not enough to be an issue.

This new CDS solution was injected into the 800 cattle with very positive results. The animals health improved and the vet bill dropped like a rock.

In one case, the farmer called Andreas about a particularly sick cow. He advised him on a dosage, but apparently the farmer misunderstood the instructions and gave the cow 10 times the recommended amount. The result? The cow was a bit "high" with ears and tail standing up, but subsequently became healthy with no ill effects from the accidental overdose.

On the human side, CDS also solves the problem some people have with sensitivity to citric acid as well as the taste of CD. Not having the original chemicals (sodium chlorite/citric or hydrochloric acid) in CDS makes a big difference for some. Another advantage of CDS was a drastic reduction in Herxheimer reactions.

After a long weekend in the Dominican Republic, we went home with new information to try out with Patrick.

I began by swapping out drops of CD for milliliters of CDS. Initially, I didn't notice improvements or regression, but I did notice that Patrick was waking earlier and having a less profound sleep. After about 30 days of trying the new miracle I began to think that, "If it ain't broke, don't fix it." So we went back to the original CD that had given us all of the recoveries and improvements. Patrick went back to sleeping very well and I never looked back. I was a confirmed CD user and CDS, for me, had no obvious advantages.

A little over a year later, Jim was at our home, and on December 26th, 2012, his new assistant showed up from the Middle East to help him. She could be called the "CDS poster child." We differed on this topic right away. But, she was very insistent. She asked me why I didn't like CDS and I told her. So, she told me that I had been using "unstabilized" CDS and now they "stabilize" the CDS.

When the CDS is "stabilized" (with some sodium chlorite added back into the CDS liquid), it gives an extra kick, such that when the chlorine dioxide molecule oxidizes a pathogen, there is sodium chlorite on hand to react with the acid that is created by the death of the pathogen. The result is more chlorine dioxide released on the spot.

I decided to try stabilized CDS with Patrick. We arrived at his full dose based on tolerance. His sleep has been perfect. He sleeps as any pre-teen should—like a rock. And, he has great energy all day long. I can say that there was no noticeable difference over the four months that he was taking CDS, from the days when he took CD. However, he is a little bit "lighter," as if he has no cares in the world. It seems as if something that might have made him uncomfortable no longer bothers him. Little by little, a few families switched to CDS to see if their children did better, worse, or as well as they did with CD.

After a couple of trial months with about 20 families, I think that it is safe to say that CDS has its place in the buffet of different methods of chlorine dioxide administration. We still have the original CD, made with citric acid or HCl. Now, we have CDS and CDH (Chapter 7). Today, we no longer use stabilized CDS, as I feel that CDS is only for hypersensitive individuals and to be used only until they can switch over to CD or CDH. Whatever formula we choose, it will help to heal the body of unwanted pathogens, and helps with parasites as well as heavy metals. Each person is different, and it is up to us as parents to observe and decide which method of chlorine dioxide administration is best.

You are in no way encouraged to switch to CDS if what you're doing with CD is working for you. For some people it may be better, but as of now, only one of our recoveries has come from CDS—all the rest are from CD. Some children may try CDS and come to realize that CD is still better for them.

When to Use CDS

CDS is used when someone just can't get past 1 drop divided in 8 fl. oz., over a day without experiencing a Herxheimer reaction. This allows us to get past the road block and still detoxify the body. Eventually, we want to go back to the standard CD Protocol which I believe to be more effective.

This is not a common situation, but it does occur.

Making vs. Buying CDS

Most people are a bit concerned about making CDS since it involves a distillation process where you produce chlorine dioxide gas and then cause it to go into water. When Andreas first came up with making CDS, it involved two containers, a hose, heat and lots of ventilation. There are still lots of videos on YouTube explaining that original process. However, there is now a much better, easier and safer way to produce CDS at home. So, if you see a CDS making video or read instructions that talk about using a hose—forget it! The new process is much easier and safer and doesn't involve any heat nor hoses.

There are some online sellers of CDS, but the cost of it is high and making your own is simple. It just takes a little effort in learning the process and having the right equipment. In addition, by making your own, you will have a better understanding of the behavior of chlorine dioxide.

Some Important CDS Points

When it comes to CDS, it is all in the concentration of chlorine dioxide in water; measured in *parts per million*—abbreviated **ppm**.

When buying or making your own, the goal is to end up with a bottle of CDS having a concentration strength of 3,000ppm. This bottle should ideally be made out of glass and kept in the refrigerator. You never use this concentrated solution directly. It must be diluted.

When we say, for example, use 10ml of CDS, we mean mix 10ml of 3,000ppm CDS with whatever indicated amount of water, and use that diluted mixture for the indicated purpose.

Make sure you read and understand Appendix 6, page 467 which explains more about concentration and measuring ppm.

There are several names attributed to this new, better and simpler method of making CDS, such as:

- The Shot Glass Method
- The Overnight Method
- The New Method

There are some demonstrations on YouTube showing this new method along with write-ups on various forums. These methods use a plethora of different containers having different shapes and sizes. The specific tools used to produce CDS can have a significant impact on the concentration.

It should also be noted that the concentration strength is reduced every time you open the CDS source bottle as chlorine dioxide gas escapes the liquid and evaporates into the air space above the liquid. When you open the source container, some of the chlorine dioxide gas escapes. So, the strength of a half full CDS bottle can be substantially less than what you started off with when the bottle was full. This is similar in concept to a pop bottle that loses its fizz if not used up quickly.

To substantially reduce the evaporation of the chlorine dioxide gas, the CDS source bottle should be stored in the refrigerator, which stabilizes and significantly reduces the amount of CD gas escaping the CDS solution. But, no matter what the temperature, some gas will always escape and reduce the concentration each time the bottle is opened. A narrow neck bottle helps maintain the integrity of the CDS for longer periods.

It is better to store the concentrated source CDS in smaller 300ml bottles than keep it in a larger one liter bottle. That way you use up what is in one bottle faster with less concentration loss. A good storage bottle is shown and described on page 160.

The New CDS Making Method — The Basic Concept

First, let's go over the basic principles of the CDS making process and then we will go into details.

We start with two containers (both ideally made of glass). One container is larger and has some kind of a lid that can be tightly closed. The other is a smaller container having NO lid. The relative size of each container is such that the smaller container fits comfortably inside the larger container while allowing the lid to properly close on the larger container when the smaller one is inside. See page 151 for an example of a well-made glass kit distributed by wps4sale.com.

The large container is then partially filled with filtered/distilled water. The level of water must be low enough so as to not overflow into the smaller container if it is placed into the partially filled larger container. The photos on page 151 show the concept of the two bottles visually.

Now that you have the basic idea, let's make some CDS...

Using the Overnight CDS Generating Kit

The following directions assume you have the glass *Overnight CDS Generating Kit (1000 ml)* from wps4sale.com (as shown on page 151). You are of course not required to use their kit. However, based on our own experience of looking high and low for a good one, theirs has the key features:

- Made of glass
- Smaller glass fits inside of larger container AND 1000ml of water inside the big container will not flood into the smaller container once placed inside.
- Lid fits tight and no gas escapes during the chemical reaction time.
- Lid is plastic—NEVER USE METAL LIDS!

Note: There are plastic CDS kits that work. You may be able to find a combination of containers made of plastic but not glass where you live. In fact, I made a YouTube video where I show how to make CDS using a plastic container. Just understand that chlorine dioxide will deteriorate plastic containers over time, so it should be avoided. Even if you have to use a plastic jar to make the CDS, store the final product in glass. The less time the CDS is in plastic, the better! As for lids, we have no choice. Only use plastic lids. Metal lids will oxidize VERY rapidly! Even metal lids with a plastic coating on the inside will deteriorate.

Steps to Making CDS:

1. Start by pouring one liter of distilled or reverse osmosis water into the large glass container. The water should be around room temperature.

2. Place the empty dessert cup inside the large container making sure no water spills into the cup.

3. Measure 75ml of sodium chlorite solution (22.4% $NaClO_2$ in water) and pour into the dessert cup, while being careful not to spill any sodium chlorite into the water surrounding the cup.

4. Measure 75ml of 10% HCl or 50% citric acid solution and pour into dessert cup.

5. IMMEDIATELY screw on the lid to the large container while being careful to not jiggle the container which could potentially cause the CD solution

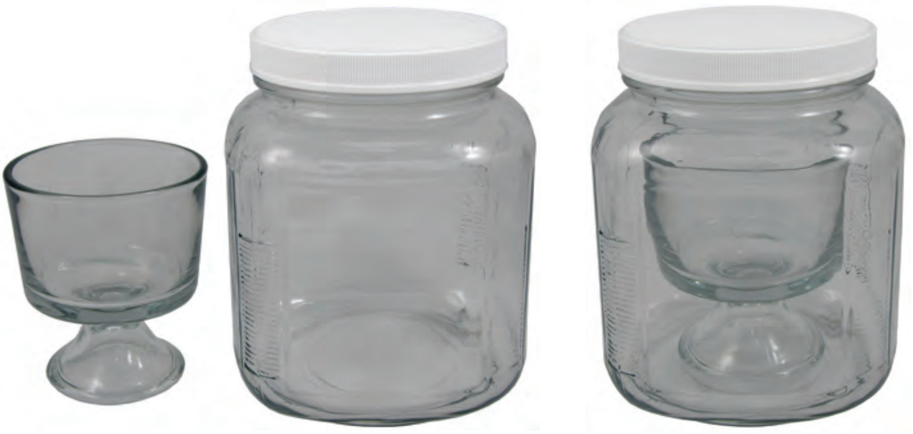

wps4sale.com sells this Overnight CDS Generating Kit that produces 1 liter of CDS at 3,000ppm from 75ml of sodium chlorite plus 75ml of activator. The only part not made of glass is the plastic lid. The glass dessert cup is sized just right to fit inside the larger glass jar with enough room left over for 1 L of water that won't overflow into the dessert cup.

to mix with the surrounding water. Make sure the cover is tightly screwed down so as to not have any leaks. If you smell chlorine dioxide (beyond the little that escapes while you are putting on the lid) then something is wrong with the lid and/or container.

6. If you did everything correctly you should see the CD solution turn a dark amber/brown within about a minute (see below) and notice that the water starts to turn a light yellow several minutes later. It is also normal to see bubbles form in the dessert cup. It is normal for the pressure to change during the chemical reaction. We find that citric acid produces a vacuum and HCl can produce a pressure at first and then a vacuum.

7. Cover the kit with a towel to reduce its exposure to light. It doesn't have to be in a pitch black location, but it definitely should not be in direct sun light. Feel free to check it from time to time and see how the colors change. Of course keep this kit out of the reach of children or anything that could disturb it. I keep mine in the cabinet overnight with a towel on it.

8. When the color in the dessert glass matches that of the surrounding water, the chemical reaction is complete. There is no harm in waiting a bit longer—nothing more will happen. This usually takes from 12-24 hours depending on room temperature and the strength of the acid used.

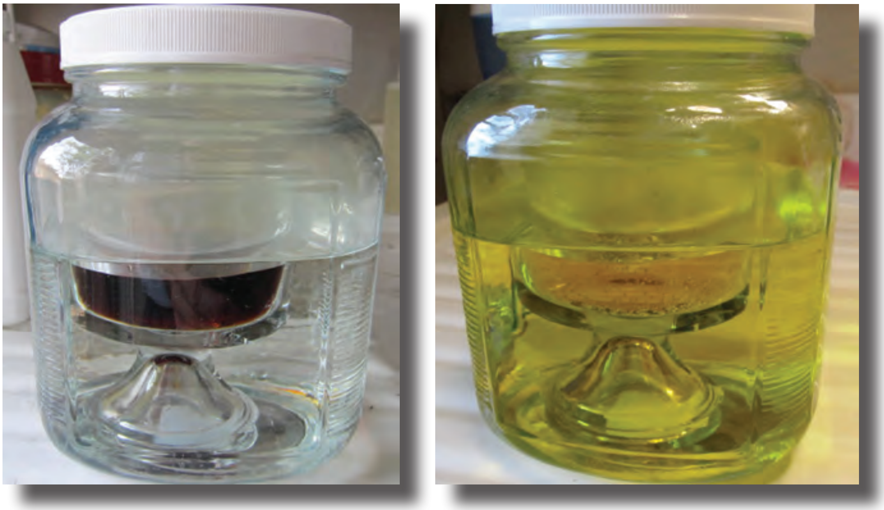

After the sodium chlorite and activator mix, the color will quickly turn a dark amber, almost black color (left). After about 12 hours, the color of the solution in the dessert cup and that of the water surrounding it should be identical (right), which indicates the chemical reaction has completed and the CDS is ready to be poured into a storage container.

9. Have your glass storage bottle(s) clean and ready to receive the CDS solution.

10. Before opening the kit make sure your work area is well ventilated. Perform the steps that follow outside or have the door and/or window open nearby. Having a fan lightly blowing any gas that escapes away from you is a good idea.

11. Slowly open the container being careful to not jostle it which could cause the CD solution to mix with the freshly made CDS. If the chemicals mix, you will have to start over again, so be careful! You will most likely smell the chlorine dioxide gas that was sitting in the air space when you first open it. This is normal, hence the need for good ventilation.

12. Slowly remove the dessert cup while being careful to not spill its contents into the CDS. Immediately dispose of the contents in the toilet and rinse off the dessert cup.

13. Now pour the CDS into the storage bottles while being aware that some chlorine dioxide gas will come out of the water into your working area.

14. Tightly seal the storage bottle(s) and place in the refrigerator.

Note: 10% HCl is the recommended activator when making CDS. However, you can also use 4% of HCl, but the reaction time will be slowed down. It may take as much as a full day before the color of the CD solution and the water match.

Dosing CDS

As mentioned earlier, CDS is particularly well-suited for unusually sensitive individuals who cannot tolerate even low doses of CD.

Start with a baby bottle filled with 8 fl. oz., of water and add 1ml of CDS. This will provide eight doses of 1/8ml. If all is well on the first day, increase to 2ml on the second day. Each day you increase by one milliliter. Increase the amount based on tolerance. If the person is having a problem with a certain level of CDS, then go down to where they were fine and stay there a few days before increasing again.

Tolerance is the key here. Once you can no longer increase you have reached that person's full oral dose of CDS.

Slowly Switching Back to CD

Once you have found the full oral dose where they are stable and cannot go higher, it is time to fold CD back in.

The goal of using CDS is not to replace CD, but rather do some fundamental detoxing so you can bring CD back and reach full oral dose of CD.

This is accomplished by taking out one milliliter of CDS and replacing it with 1 drop of CD. For example, if the individual can tolerate 20 milliliters of CDS, then an ideal transition would be represented in the chart on page 154.

Of course this chart represents an unlikely scenario where everything goes exactly as desired. However, it does demonstrate what we are striving for—switching out CDS for CD and continuing to increase the dose of CD as described in Chapter 5 (page 81).

To accomplish this process effectively requires careful observation of your child and making decisions as to when you can go up; when to hold; and when to back down.

Hypothetical CDS to CD Transition

Transition Day	CDS in milliliters	CD in drops	Comments
1	20	0	Last CDS only day.
2	19	1	Transition starts
3	18	2	Transition continues
4	17	3	
5	16	4	
6	15	5	
7	14	6	
8	13	7	
9	12	8	
10	11	9	
11	10	10	
12	9	11	
13	8	12	
14	7	13	
15	6	14	
16	5	15	
17	4	16	
18	3	17	
19	2	18	
20	1	19	
21	0	20	Transition completes
22	0	21	CD increases each day
23	0	22	
24	0	23	
25	0	24	

CDS Enemas

CDS enemas can be useful for sensitive individuals when starting the protocol. Follow the same enema instructions in Chapter 5, page 103, keeping in mind that 1ml of CDS has a relative strength of approximately 60% of 1 drop of CD. With that in mind you can work up to 40ml of CDS per liter of enema water.

Chapter 7

CDH - Going Beyond CD & CDS

"The progressive development of man is vitally dependent on invention. It is the most important product of his creative brain."
~ *Nikola Tesla*

This protocol is constantly evolving, and until every person with autism recovers, we will continue to search for things to add or tweak to the protocol so we reach that goal. A few months before the release of this second edition, I was made aware of a new method of CD preparation, known as *Chlorine Dioxide Holding (solution)* or simply CDH.

CDH can almost be described as a hybrid between classic CD and CDS. Where CDS has no raw material left in the final product, but only chlorine dioxide gas dissolved in water, CDH contains some raw material (similar to classic CD) along with the chlorine dioxide in the final product. This new process allows the sodium chlorite to react with the acid for a significantly longer period of time, thereby reducing much of the remaining amount of unactivated sodium chlorite and activator. Some people cannot tolerate citric acid so CDH is usually made with 4% HCl. Initial reports indicate that CDH is better tolerated than CD.

Another interesting benefit of CDH is that it appears to mix well with the permitted natural sweetener stevia (SweetLeaf® brand) without reducing the potency of CDH. This can help children who have an aversion to the taste of CD. Many families have also reported that they were able to increase their child's dose without producing a Herxheimer reaction, in contrast to classic CD. Older children and severely affected children have benefitted as well from the CDH preparation; you can read more about that on page 221.

Currently, this new method is being used by a relatively small group of families (around 70 as of November, 2013). Many of them are reporting that CDH continues to produce results for their children on the spectrum, and they are seeing even better things than before.

As with all new things, it is important to test with a diverse group of families over a long enough time period to make sure that gains are sustained. So we ask that you please keep in mind that CDH is literally the bleeding edge of this protocol. While we are excited to share a new option with you, the decision to use CDH must not be taken lightly. If you are doing well with classic CD you may never need to use CDH. Consider the old saying, "If it ain't broke, don't fix it." Of the 115 children who have lost their autism diagnosis, 114 have done so with CD and 1 with CDS. So far, 0 with CDH (4 months in use for ASD). I expect this number to change soon as we are seeing good things from CDH. Time will tell. Stay tuned.

Scott McRae, his wife Brenda, and Charlotte Lackney have been pioneers in developing the CDH method. The following section is written by Scott where he discusses how CDH evolved along with detailed instructions on how to prepare and use it.

* * *

My wife, Brenda, and I first learned about CD early in 2009, through Bhante Vimalaramsi, an American born Buddhist monk. He's now a very close friend of ours, and even as I'm writing this, I'm at a meditation retreat which is based upon his understanding of the open-minded, experiential teaching of the Buddha.

After receiving our first set of CD bottles from an online seller, we started by trying to get to 15 drops of CD activated with lemon juice for 3 minutes, twice a day, which was the protocol at the time. Though we gave it a good effort, we couldn't get past 6 drops before vomiting and having diarrhea, and this caused us to initially stop taking CD. Then, after moving to Jakarta, Indonesia (Brenda's home town), I experienced almost monthly illnesses, which I believe was due to being in a new and tropical part of the world (I'm from San Diego, CA), riding in public transportation twice a day and teaching in a school of over 300 students. After a year and a half of constantly being sick, I remembered our bottles of CD which we had fortunately brought with us from the US, and I started on Protocol 1000. By that time, *Protocol 1000* had been developed to be what it is now—3 drops of CD (activated for 20 seconds with 50% citric acid) per hour, 8 times a day—and so I decided to do the 3 week cleanse. By doing the cleanse, I experienced a huge improvement in my health. Not only did I immediately stop getting sick, but I also felt that my energy level had increased by about 25%! Obviously, I was excited by these great results and so was my wife because she had started taking CD again as well. However,

Scott & Brenda McRae from Jakarta, Indonesia

even though our results were great and our experience was far better than before with the old, "try to get to 15 drops twice a day protocol," we still were having some nausea and bouts with diarrhea when we took CD. This was especially true at the times when we felt the symptoms of a cold or flu and tried to take more than the 3 drops doses in order to overcome it. Still the results we were getting with CD on Protocol 1000 outweighed the nausea and diarrhea that we experienced and so we continued using it, especially when we felt some sickness coming on.

After taking CD for about a year in this way, CDS hit the CD world. Being the experimental type, I decided to give it a try. After searching all over Jakarta for the plastic tubing for a couple of days and finally finding it, I made our first batch of CDS. We both tried CDS for about six months but found it to be less effective than CD, so we stopped with CDS and went back to CD and Protocol 1000.

Nevertheless, there were 2 things I really liked about CDS: (1) it NEVER caused us any nausea or diarrhea, and (2) it was so easy to use since it was pre-made (no mixing of chemicals before each use). The main thing for me was not having any more nausea because I really disliked feeling sick. So, I started thinking about CDS and my experience with making it and I came to the conclusion that possibly, the reason why CDS wasn't nauseating was because there wasn't any unactivated sodium chlorite in the solution—it was just chlorine dioxide dissolved in water. I remembered that when I made CDS myself, even after the activation process had gone on for an hour in a heated condition, if I stirred or swirled the activation chamber bottle, more ClO_2 would still come out of it. This got me to thinking... if the chemical reaction between the sodium chlorite and the 50% citric acid solution used to make the CDS was still able to produce more chlorine dioxide after an hour in a heated condition, then surely the 24 drops that I was using to make my daily CD dosing bottle weren't being fully activated after just 20 seconds.

Therefore, I decided to increase the activation time and tried it out on myself first and then my wife (aren't I considerate?). We both found there was no longer a nausea problem, even when taking more drops per hour than we had ever taken before.

Since I was in an experimental mode, I decided to see if I could make larger quantities of concentrated CD all at once to make it more convenient. I mixed equal amounts of sodium chlorite and 50% citric acid in a bottle, let it activate for about a minute, and then added a specific amount of hot water to further encourage the activation process. In the end, I had a total of 140 ml of this concentrated CD solution. Later, I found that hot water wasn't necessary and that room temperature water worked even better because there was less CLO_2 gas lost in the process.

The finished chemistry was indeed very strong and still didn't cause any nausea, so I was compelled to post my discovery on the Genesis II Forum. I named it *7 Day Fridge MMS (CD)* because it provided a 7 day supply of Protocol 1000. The ingredients added up to 140 ml, so each 20 ml was 1/7th of the total, or a one day supply of "pre-activated" CD similar to Protocol 1000. This made taking CD every day really easy. Just pour out 20ml of the new concentrated solution into a water dosing bottle, and then pour out 1/8th of the bottle every hour into some water in a glass, and drink it.

The *7 Day Fridge MMS (CD)* method worked great for my wife and me. Others on the Forum also tried it and liked it. In addition, I gave it to some people at the school where I work to overcome their colds (usually overnight) as well as other diseases within a short time.

About 18 months later, Charlotte, my Forum friend, began testing the *7 Day Fridge MMS (CD)* process to determine the actual ClO_2 content using her *Sensafe*^{TM} Chlorine Dioxide Photometer. Over several months, we worked together to further refine the 7 Day Fridge method. The result of all of our testing and refinements is this new and exciting CDH product.

The final formula to produce CDH using the 1 Bottle Method came out to be:

22 parts water	91.6%
1 part sodium chlorite	4.2%
1 part HCl (4%) or Citric Acid (35%)	4.2%
Total Solution	**100%**

I believe this new way of making CD will be of great benefit to humanity because it will allow people to gradually double or even triple the amount of CD they can take with little or no stomach upset. By increasing their doses to higher levels, people will be able to overcome CD treatable diseases more quickly than ever before.

As with any new technology, more changes and developments are likely. For example, we now know that CDH tastes much better if activated with hydrochloric acid instead of citric acid.

Producing CDH Using the One Bottle Method

Making CDH is really quite simple. You can use any size of bottle, and produce any quantity you wish, as long as you follow the basic instructions and keep the proportions the same. However, before diverting into different quantities, it is recommended you follow these directions exactly to insure you have the process down correctly.

Note the following volume equivalents:

Fluid Ounce (U.S.) Approximate Equivalents

30ml = 1 fl. oz. US (2 Tablespoons)

660ml = 22 fl. oz. US

720ml = 24 US fl. oz. US

750ml = 25 US fl. oz. US

Equipment Needed

- One 750ml (25 fl. oz. US) glass bottle with an airtight cap. DO NOT use a metal cap (even if it is lined with plastic). The best caps are plastic or even a synthetic cork in good condition. A common 750ml wine bottle is ideal for this as you can get them colored to reduce possible UV exposure (the darker the better). But, a colored bottle is not absolutely necessary and using a clear one will allow you to actually see the chemical reaction as it turns from clear to yellow.

Schweppes
GINGER ALE

- Three 240ml (8 fl. oz. US) bottles ideally made of colored glass to protect the CDH from UV light (the darker the better), but colored bottles are not absolutely necessary. You can also use smaller bottles if you like—these are just used to divide up the resulting 720ml of CDH into smaller bottles to help retain the ClO_2 concentration while opening and closing the bottles during dosing. *Schweppes*™ sells 6-packs of Ginger Ale, Club Soda and Tonic Water in 10 fl. oz. glass bottles with plastic lids that are great for this purpose, and hold 300ml easily. Only the Ginger Ale is in a colored bottle (green), the others are clear.
- One measuring cup or graduated cylinder to accurately measure liquid in either milliliters or fluid ounces.

Ingredients Needed

Ingredients should be at room temperature—not cold. If CD and/or activator are right out of the refrigerator make your water warmer to offset, or allow ingredients to warm up before using.

- 660ml of distilled or purified water at about 70° - 90°F (21° - 32°C).
- 30ml of sodium chlorite (22.4% solution) near room temperature or slightly above.
- 30ml of 4% Hydrochloric Acid (HCl) or 35% Citric Acid ($C_6H_8O_7$) near room temperature or slightly above.

Note: The amounts indicated above add up to 720ml, while the wine bottle easily has room for an additional 30ml and more. Refer to the chart on right if you wish to produce a different batch amount.

The CDH Formulation Table and Different Acid Concentrations

What if you have 10% Hydrochloric acid or 50% Citric Acid (very common)? Or, you wish to use a different size of bottle? Not a problem. These acids can still be used. However, the formula changes accordingly.

The table below is a great tool for determining the formula for a given bottle size. To use the table, start by circling the size of the bottle you wish to fill in the left most column. Next look at the acid you have and its labeled concentration. Match that with one of the 4 options across the top. Below the matching acid/concentration you will find the 3 formulation numbers for water, sodium chlorite (labeled "SC"), and whatever acid you are using. Just go down the appropriate 3 columns to where the bottle size line intersects and you will have the numbers you need. Substitute these numbers in the following preparation instructions if your situation calls for it.

CDH Formulation Table

(Applies to the One Bottle Method of making CDH ONLY!)

SC=Sodium Chlorite / HCl=Hydrochloric Acid / CA=Citric Acid

Water Bottle (ml)	Hydrochloric Acid 4%			Hydrochloric Acid 10%			Citric Acid 35%			Citric Acid 50%		
	Water (ml)	SC (ml)	HCl (ml)	Water (ml)	SC	HCl (ml)	Water (ml)	SC (ml)	CA (ml)	Water (ml)	SC (ml)	CA (ml)
10	9.2	0.4	0.4	9.4	0.4	0.2	9.2	0.4	0.4	9.3	0.4	0.3
20	18.3	0.8	0.8	18.8	0.8	0.3	18.3	0.8	0.8	18.6	0.8	0.6
30	27.5	1.3	1.3	28.3	1.3	0.5	27.5	1.3	1.3	27.9	1.3	0.9
40	36.7	1.7	1.7	37.7	1.7	0.7	36.7	1.7	1.7	37.2	1.7	1.2
50	45.8	2.1	2.1	47.1	2.1	0.8	45.8	2.1	2.1	46.5	2.1	1.5
60	55.0	2.5	2.5	56.5	2.5	1.0	55.0	2.5	2.5	55.8	2.5	1.7
70	64.2	2.9	2.9	65.9	2.9	1.2	64.2	2.9	2.9	65.0	2.9	2.0
80	73.3	3.3	3.3	75.3	3.3	1.3	73.3	3.3	3.3	74.3	3.3	2.3
90	82.5	3.8	3.8	84.8	3.8	1.5	82.5	3.8	3.8	83.6	3.8	2.6
100	91.7	4.2	4.2	94.2	4.2	1.7	91.7	4.2	4.2	92.9	4.2	2.9
150	137.5	6.3	6.3	141.3	6.3	2.5	137.5	6.3	6.3	139.4	6.3	4.4
200	183.3	8.3	8.3	188.3	8.3	3.3	183.3	8.3	8.3	185.8	8.3	5.8
250	229.2	10.4	10.4	235.4	10.4	4.2	229.2	10.4	10.4	232.3	10.4	7.3
300	275.0	12.5	12.5	282.5	12.5	5.0	275.0	12.5	12.5	278.8	12.5	8.7
350	320.8	14.6	14.6	329.6	14.6	5.8	320.8	14.6	14.6	325.2	14.6	10.2
400	366.7	16.7	16.7	376.7	16.7	6.7	366.7	16.7	16.7	371.7	16.7	11.7
450	412.5	18.8	18.8	423.8	18.8	7.5	412.5	18.8	18.8	418.1	18.8	13.1
500	458.3	20.8	20.8	470.8	20.8	8.3	458.3	20.8	20.8	464.6	20.8	14.6
550	504.2	22.9	22.9	517.9	22.9	9.2	504.2	22.9	22.9	511.0	22.9	16.0
600	550.0	25.0	25.0	565.0	25.0	10.0	550.0	25.0	25.0	557.5	25.0	17.5
650	595.8	27.1	27.1	612.1	27.1	10.8	595.8	27.1	27.1	604.0	27.1	19.0
700	641.7	29.2	29.2	659.2	29.2	11.7	641.7	29.2	29.2	650.4	29.2	20.4
720	660.0	30.0	30.0	678.0	30.0	12.0	660.0	30.0	30.0	669.0	30.0	21.0
750	687.5	31.3	31.3	706.3	31.3	12.5	687.5	31.3	31.3	696.9	31.3	21.9
800	733.3	33.3	33.3	753.3	33.3	13.3	733.3	33.3	33.3	743.3	33.3	23.3
850	779.2	35.4	35.4	800.4	35.4	14.2	779.2	35.4	35.4	789.8	35.4	24.8
900	825.0	37.5	37.5	847.5	37.5	15.0	825.0	37.5	37.5	836.3	37.5	26.2
950	870.8	39.6	39.6	894.6	39.6	15.8	870.8	39.6	39.6	882.7	39.6	27.7
1000	916.7	41.7	41.7	941.7	41.7	16.7	916.7	41.7	41.7	929.2	41.7	29.2

Preparation Instructions

Follow the steps of mixing ingredients in this order (assuming you are using the 750ml glass bottle):

1. Pour 660ml of purified water into the 750ml glass bottle.
2. Add 30ml of sodium chlorite to the 750ml glass bottle.
3. Add 30ml of 4% HCl or 30ml of 35% citric acid to the 750ml glass bottle.
4. Immediately cap/cork the bottle tightly so that no ClO_2 gas can escape and give it a good shake to thoroughly mix the ingredients (and a few more times later on if possible).
5. Store the bottle in a dark place with a temperature of 70°F to 90°F (21°C to 32°C) for 12 to 24 hours or more (24 hours or longer if you're in a cold environment—below 70°F (21°C)).
6. After the storage time has passed, place the bottle of CDH into the refrigerator (not the freezer) and allow it to cool down for 3+ hours before opening it for the first time. The temperature of the solution should be no more than 51°F (10.5° C). Note: If you are using a significantly smaller bottle, cool down time can be reduced because the smaller volume of solution will cool faster.
7. Finally, pour the CDH from the 750ml bottle into the smaller bottles, cap tightly and keep refrigerated until ready to use. The smaller bottle are easier to dose from and they also help reduce the number of times gas can escape and reduce the potency of the solution.

Using CDH

The CDH is now ready for use. Each milliliter of CDH solution contains 1 pre-activated drop of CD. It can be used for ANYTHING that CD is used for; viruses, bacteria, yeast, parasites, heavy metals, enemas, tub baths, gums & teeth, skin care, infections, etc.

When used orally, each 1ml of CDH should be added to at least 30ml of water. You can add even more water if you notice slight throat irritation with higher doses. If taste is an issue, a little bit of stevia may be added to each dose to improve taste by sweetening it. Some of the moms are adding SweetLeaf® Natural Stevia Sweetener to their daily bottle and report no negative impact on ppm level.

Although CDH is strong, it has shown to be gentler on folks who have issues with nausea when using traditional CD. You should be able to start at whatever drop dose you were on with CD and switch to an equivalent milliliter dose of CDH and gradually increase to tolerance. Typically, people are able to take 2 to 3 times as much of CDH as traditional CD, without experiencing nausea.

The CDH bottle should be kept refrigerated and only taken out to extract doses. Since 720ml is a fairly large quantity that may take many days to finish, it's a good idea to take the 720 ml and divide it up into 3 smaller 240ml bottles (720ml ÷ 3 = 240ml) so you won't lose much of the ClO_2 each time you open the bottle. By doing this, you will conserve as much of the ClO_2 as possible. It's also easier to extract doses from a smaller bottle using a syringe or pipette.

Also, keep the CDH bottle out of direct or indirect sunlight to prevent loss of ClO_2. If you make a dosing bottle for the day, it's best to keep it cold, but it is not absolutely necessary.

A little note about taste: Most people who complain about the bad taste of CD, CDS or CDH (which does not have any taste at low doses) are actually reacting to the smell of the ClO_2 which can lead to developing a long-term aversion to any of the treatment solutions. So, if you can minimize the gas floating around your nose, you will have an easier time with drinking the CD, CDS or CDH dose. To accomplish this, Charlotte suggested avoiding the use of a cup or wide mouth drinking bottle. Instead, use a bottle with a small opening such as a common water bottle (preferably made of glass). Of course, if the smell doesn't bother you, this is a moot point, but at least you have this little trick if it does.

Well, that's all you need to know to get started on this great new way to make and use CD. May this new CDH formulation bring you and your family much health and happiness.

Scott McRae
Jakarta, Indonesia
November, 2013

* * *

We appreciate Scott, Brenda and Charlotte's contribution to the variety of ways to produce and use chlorine dioxide. Be sure to check the Facebook groups and forums for the latest developments on CDH.

Some of you may feel overwhelmed by what you just read. So, let me give you my "Easy-Peasy" single paragraph method of making CDH: I take a 600ml Lifefactory™ bottle, put in 550ml of water; add 25ml of sodium chlorite; followed by 25ml of 4% HCl (or 50% citric acid). Leave it 12 hours in a cabinet, after which it goes in the fridge for 2 or more hours. Done! ☺

Healing
the symptoms known as
Autism
Kerri River
with Kimberly McDan
Healing the Symptoms Known as Autism
Kerri Rivera
With over 120 testimonials from around the world!

Healing & beating autism.

Chapter 8

Step 3
The Kalcker Parasite Protocol

> *All of our fathers had a treatment for parasites as a part of their cultural practice. We have gotten away from this because of our reliance on modern practice. We would do quite well to relearn the ways of our ancestors in this area and keep ourselves in relatively good health always.*
>
> ***~ Chief Two Trees***

The word "parasite" comes from the Greek word meaning, "one who eats off the table of another." Parasites, to Ancient Greeks, were those who sat at another's table, and paid for their meal with flattery.

As I mentioned earlier in the book, a very interesting thing started happening with the CD enemas. What we had previously believed to be mucous or biofilm coming out with the enemas, turned out in many cases to be worms (helminths) (aka parasites)—in rare cases they were still alive and wiggling in the toilet! We believed the most common were *Ascaris lumbricoides* (roundworm). However, what we are now seeing more and more appear to be rope parasites, a potential new species of helminth discovered by Dr. Gubarev, Dr. Alex Volinsky, and coworkers (submitted January 14, 2013). DNA testing is the only way to definitively say, but at $25,000 USD for each analysis, with a minimum of 100 test cases, it is rather cost prohibitive at the moment.[1]

In addition to Ascaris and rope parasites, parents have also seen hookworms, pinworms, tapeworms, and flukes, among others. This is an extremely important piece of the puzzle for so many of our children. We have been led to believe that in first world nations, parasites are not a problem. This is absolutely not the case.

A well washed parasite. You can almost feel the texture.

Andreas Kalcker,
co–author of the
Kalcker Parasite Protocol

Miriam Carrasco Maceda,
co–author of the
Kalcker Parasite Protocol

I was honored to have Dr. Andreas Kalcker and Miriam Carrasco Maceda share a chapter from their upcoming book, *Parasites: The Silent Enemy*. Andreas explains the importance of lifelong deworming, and shares with us a protocol that has helped many children and adults become healthier; and for some children on the spectrum was the last piece added that led them to recovery. The version included here has been tailored specifically for children and adults with ASDs. The original protocol can be found at:

www.andreaskalcker.com/
index.php/en/health/parasite

Very few details have been changed, but it is important to note that the Parasite Protocol here is what has been proven to help many of our children on the spectrum, including many of the recovered children.

Thank you Andreas and Miriam for your valuable contributions to this movement, selflessly sharing your findings, and for always taking the time to help.

HOW TO DETECT AND TREAT A PARASITIC INFECTION

Parasitic infections are more common than most people think, and may or may not result in serious health problems. We may be infected with multiple types of parasites, which vary in size and location, on or in the body.

Parasites can be classified as either microparasites, such as malaria that are only visible under the microscope, or large macroparasites such as round or flat intestinal worms (roundworms, tapeworms, etc.). These can be seen by the naked eye, and can reach great sizes. Internal parasites are found, not only in the intestines, as is generally thought, but anywhere in the body, including the lung, liver, muscle, stomach, gallbladder, brain, blood, skin, joints, and even in the eyes.

In recent history, the great migratory movements of the human population via rapid transportation and widespread trading have shortened the distances that previously had separated people and diseases. Formerly localized diseases have thus become universal ailments. Parasites previously confined to very specific geographical areas now appear in other locations, far away from their initial homelands. Unfortunately, conditions typical of the lower socioeconomic strata, (under which a large percentage of the global population lives) tend to favor the transmission of diseases and parasites.

A high percentage of the world population suffers from infections by parasites, which the WHO (World Health Organization) estimates are responsible for 15 million child deaths annually. In addition to the great cost represented by deaths, chronic and persistent infections have increased as parasites have developed multiple mechanisms of evasion and resistance to specific immunity. This allows them to circumvent and cancel the host immune response.

Persistent parasitic infection in human hosts leads to chronic immune reactions, which can result in tissue damage and altered immune regulation. Ninety percent of the world population is infected with one or more parasites, and up to five different types may coexist in the same host.

This situation becomes dangerous when the internal balance within the host is upset, the number of parasites skyrockets, and the host begins showing signs of serious illness that may even result in death. However, in some cases, parasitic worm infections do not result in disease, in fact, a number of carriers are found to be healthy.

A really good look into a parasite. You can see the actual intestines of the parasite. Also known as a helminth.

The same parasite (as above) from a different angle.

Considering that most people are not even aware of their infections, parasites have become silent killers, claiming the lives of many unsuspecting victims going about their lives oblivious to the danger. Some doctors in Western Europe and the U.S. seem unwilling to even contemplate that we may be infected.

Taking into account the recent increase in travel, immigration, and trade across continents, it is not hard to see how the problem has now become magnified to an alarming level. Parasites, especially the modern "toxified" versions, may well be causing many of the rare diseases now becoming more prevalent, as well as other recently identified or growing problems such as chronic fatigue, fibromyalgia, and arthritis.

The most common verminosis (infestation with or without obvious symptomatology of disease caused by parasitic worms) is intestinal. People who have them not only suffer from a large quantity of lost nutrients (absorbed by the parasites), but also from perforations made by worms in the digestive tract that can open the door to various infections and possible autoimmune deficiencies. Intestinal worm infections are very common and can affect everyone, not only people with poor hygiene habits. Helminths (worms) are transmitted by ingesting the eggs or larvae of parasites, which then hatch in the intestinal tract.

A parasitic infection or reinfection can be acquired through one or more of the following avenues:

- ▶ From more or less direct contact with an infected person (fecal or sexual).
- ▶ From self-infection, for example, through anal-hand-mouth contact. By scratching the anal area, eggs can become lodged under the fingernails.
- ▶ From congenital transmission (mother to fetus).
- ▶ From commonly contaminated objects.
- ▶ From soil contaminated by human or animal excrement.
- ▶ From eating contaminated raw or undercooked meat.
- ▶ From eating raw fish.

In some countries, raw fish is included in traditional foods. We can avoid the consumption of the larvae or worms by freezing the meat or fish for at least twelve hours, depending on temperature.

A great photo of a dead parasite, believed to be the Ascaris lumbricoides, or possibly a rope worm in the "seaweed" stage.

A very good look at how long some helminths are. The more worms the people pass, the healthier they get, and the improvements come faster and faster. This child passed a lot of parasites in the beginning. Then after a few months was no longer passing them and now has an ATEC score of 5. Meaning, he no longer has the diagnosis of autism.

This is the parasite that measured 32 inches.

Another long, well washed parasite for the collection. The road to recovery is paved with many dead parasites. Adios Autism…

This is a 32-inch parasite that a young boy passed. He went on to have a great day after getting this out.

These are rope parasites as confirmed by Dr. Alex Volinsky.

60cm long worm (Oct 1, 2013) from child; 6 months on CD; 2 drops every 1-2hr; no parasite protocol; 1tblsp DE; 2-3 vials of Quinton.

This 33 inch parasite (aka "Chester") was discovered by a woman using the protocol on herself.

The bubble visible in this photo leads us to believe this is a late stage rope worm.

After passing this parasite, the child went on to have a fabulous day at school. It is so great to feel good and healthy.

The FDA recommends freezing and storing [fish] at -4°F (-20°C) or below for seven days (total time), or freezing at -31°F (-35°C) or below until solid and storing at -31°F (-35°C) or below for 15 hours, or freezing at -31°F (-35°C) or below until solid and storing at -4°F (-20°C) or below for 24 hours is sufficient to kill parasites. FDA's Food Code recommends these freezing conditions to retailers who provide fish intended for raw consumption. Note: These conditions may not be suitable for freezing particularly large fish (e.g. thicker than six inches).[2]

▶ From drinking contaminated water.

▶ From consuming contaminated vegetables or fruits. Often we eat poorly washed (parasite infested) vegetables or fruits. There is a common misconception that vegetables from organic farming are free from any problems, pesticides, or chemicals. The danger is that the eggs or larvae of the worms reach the farm soil through animal waste, decomposed forms of natural compost, and manure (fertilizer) added to the field. There are eggs, such as Ascaris lumbricoides, which can survive in soil under extreme temperatures for five years. It is very important to perform a thorough cleaning of fruits and vegetables, and never eat anything raw and straight from the ground, however healthy it may seem.

▶ From parasite infested animals. Parasitic infections are very easy to spread by contact with pets. Veterinarians are quick to insist upon the quarterly deworming of our animals, but there are steps we must take on our own to avoid contamination.

Parasites come in all shapes and sizes. Of course, they can also come out in pieces.

Suggested: Deworm your pet at least every three months for life, as directed by your veterinarian. During the first month, it should be done every week. Prevent pets from eating raw viscera. If animals eat raw meat or raw bones the best option is to freeze the food in advance for at least 12 hours (See citation above). If the deworming treatment is working, the animal will expel the worms in the feces or vomit, which must then be burned or buried, during the eight-day treatment.

Avoid being licked in the mouth by animals as they are in direct contact with feces, soil, and their own anus. When petting an animal, wash your hands with soap and water before eating or handling food, as the eggs of the parasites remain in the animal's hair.

- Do not walk barefoot or with open toe shoes in soil or sand.
- Avoid Hippotherapy (horseback riding)

Symptoms of Parasitic Infections

The different types of worms and toxic waste produced by parasites in our body may cause the following common problems:

Parasites come in all sizes. Some are very long. These get washed with hot water for photos. This way we get the best look at them.

Blood Disorders & Blood Parasite Diseases

Parasites absorb essential nutrients from the body, such as iron, vitamin B_{12}, and sugars, which may result in certain blood disorders. In addition, some diseases are known to be caused by blood parasites:

Blood disorders:

- Anemia
- Dizziness
- Hypoglycemia
- Weakness

Blood parasite diseases:

- African Sleeping Sickness
- Babesiosis
- Chagas Disease
- Malaria

Fatigue:

The toxic waste produced by the parasites themselves (including ammonia and psychoactive substances), can stress the detox organs and cause disorders of the central nervous system such as:

- Chronic fatigue syndrome (CFS)
- Cold in the extremities
- Dizziness
- Extreme weakness
- Internal cold
- Lethargy
- Low energy
- Night waking
- Restless sleep

Gastrointestinal Symptoms

- Abdominal pain or tenderness
- Blood in stool
- Burning in the stomach
- Chronic constipation
- Chronic diarrhea or diarrhea caused by poor absorption of food
- Colitis
- Cramping
- Cravings for greasy foods and sugary foods, lots of carbs and bread, fruit, fruit juices, alcohol, or vinegar
- Digestive problems
- Distended belly
- Eating more than normal but still feeling hungry
- Excessive bowel movements
- Fever
- Frequent vomiting and nausea
- Gas and bloating (noted after eating)
- Hemorrhoids
- Irritable bowel syndrome (IBS)
- Intestinal irritation
- Intestinal obstruction
- Leaky gut
- Malabsorption syndrome
- Mucous in stool
- Pancreatitis
- Passing a worm in stool

Growth Problems, Weight, & Appetite

Parasites usually live without being detected by the host. They rob the body of many of the essential nutrients in the food consumed. Many overweight people, who are infected with parasites, go hungry for lack of essential nutrients, causing them to eat in excess due to their parasitic infection. Furthermore, depending on the type of infestation, many people are malnourished and cannot gain weight. The following is a list of some possible symptoms:

In children:

- Poor growth
- Poor physical and intellectual development consistent with their biological age

In children and adults:

- Chronic burping
- Craving white flour products; cookies, cakes, pastries, etc.
- Feeling hungry after a meal
- Inability to gain or lose weight
- Long-term obesity
- Loss of appetite
- Obsession and/or compulsion to eat sweets or very specific foods (wheat, sugar, dairy)
- Ravenous appetite
- Uncontrollable hunger to eat more than usual
- Weight gain (specifically around the time of the full moon)
- Weight loss

Mood Problems & Anxiety

Toxins that are released by parasites can irritate the central nervous system. Anxiety and nervousness are often caused by parasites that migrate throughout the body. Some of the problems caused are:

- Anger and irritability
- Anxiety
- Confused thinking (brain fog)
- Depression
- Disorientation
- Forgetfulness
- Lack of coordination
- Mood swings

- Nervousness
- Obsession
- Restlessness
- Slow reflexes

Muscle & Joint Pain

Parasites can travel almost anywhere in the body. When they migrate to the joints and the muscles they can cause cysts and inflammation. These can often be mistaken for arthritis and/or muscle pain.

Toxins from parasites can also accumulate in the joints and muscle tissue causing:

- Chest pains
- Fibromyalgia
- Joint pain
- Muscle cramps
- Muscle spasms
- Numbness of the hands or feet
- Pain in the back, thighs, or shoulders
- Pain in the navel
- Rapid heartbeat
- Restless leg syndrome
- Seizures

Parasites in Children (including children with ASDs)

Parasites can be found in the body in asymptomatic and symptomatic stages. The former are usually found in adults. Symptomatic stages occur mainly in children, in whom we can often observe the following:

- Anorexia
- Anxiety
- Bruxism (teeth grinding)
- Cramping
- Diarrhea that alternates with periods of constipation
- Excessive Flapping
- Growth retardation
- Headaches
- Inability to gain weight
- Itching/Burning/Picking of the anus
- Nasal itching and/or anal urticaria (hives/rash)
- Nervousness and irritability

- Nose Picking
- OCD (Obsessive Compulsive Disorder)
- Rage
- Smearing feces
- Unexplained laughter or weeping
- Verbal stims
- Weight loss

Tapeworms, and some other parasites, have an affinity for B_{12} and iron. Therefore, lab results that show deficiencies in B_{12} and/or iron can be indicators of parasitic infections.[3,4] Due to its size, the tapeworm consumes enormous amounts of food that it obtains by taking the child's food. This can affect the child's normal development.

Treatment is simple, but it requires that the head of the tapeworm be removed, otherwise it will continue to grow. Tapeworm treatment is separate from this protocol and usually requires niclosamide. However, the only way to be sure the head has been removed is to identify it in the stool.

Respiratory Disease

The passage of larvae through the respiratory system or larval invasion in the lungs may cause symptoms such as:

- Acute bronchitis
- Asthma
- Drowsiness
- Dyspnea (shortness of breath; air hunger)
- Chronic/irritative cough
- Pneumonia
- Shortness of breath or respiratory failure

Sexual & Reproductive Disorders

Immune dysfunction as a result of a parasitic infection can lead to:

- Candida - yeast infections
- Cysts and fibroids
- Erectile dysfunction
- Fluid retention
- Male impotence
- Menstrual problems
- Premenstrual syndrome
- Prostate problems
- Urinary Tract Infections

Skin Disorders & Allergies

External parasites (lice, bedbugs, scabies, etc.) that penetrate the skin can cause itching, redness, and/or rashes etc. However, internal parasites can be responsible for skin disorders as well. Parasites create toxic metabolic waste, and because the skin is the largest organ, the body tries to eliminate them through it, resulting in many skin problems.

Some symptoms may include:

- Allergies (to foods, dust, mold, etc.)
- Anal itching
- Brittle hair
- Crawling sensation under the skin
- Dermatitis
- Dry hair
- Dry skin
- Eczema
- Eruptions
- Hair loss
- Itchy nose
- Itchy skin
- Jaundice
- Psoriasis
- Skin ulcers
- Sores
- Swelling
- Urticaria (hives; skin rash)

Sleep Disorders

The body reacts to parasites during rest periods because at night is when parasites are most active. Nocturnal awakenings are common, especially between 2 and 3am, when the liver tries to rid the body of toxins produced by parasites. This in turn may produce:

- Insomnia
- Teeth grinding
- Bedwetting
- Drooling while sleeping
- Sleep disturbances - multiple awakenings during the night
- Restless sleep

Other Problems Associated with Parasites

- Bad breath
- Blurred vision
- Body odor
- Breathing problems
- Chronic infections: viral or bacterial
- Circulatory problems, numbness in the extremities, difficulty in moving
- Cough or coughing up blood
- Difficulty swallowing
- Excessive salivation
- Fever
- Fluid build-up or retention during the time of the full moon.
- Low immune response
- Peritonitis
- Sensation of a foreign body or discomfort in the throat
- Swollen eyes
- Weight gain during the full moon.

Blood Analysis

The following markers may be present when a person is suffering from a parasitic infection or the resulting allergies:

- Anemia/low iron
- Elevated immunoglobulin (IgE)
- Elevated eosinophils (The eosinophil is a specialized cell of the immune system, more specifically it is a proinflammatory white blood cell. According to the Registry for Eosinophilic Gastrointestinal Disorders (REGID), their known functions include movement to inflamed areas, trapping substances, killing cells, antiparasitic, and bactericidal activity, participating in allergic reactions, and modulating inflammatory responses.)
- High ammonia
- High oxalates
- Low vitamin B_{12}

Measures in the Home Environment to Prevent Reinfection

It is important to treat all people and pets that live in the same environment to prevent someone from infecting others. Reinfection occurs via underwear, bedding, towels and household items such as children's toys or animals that have been in contact with eggs. It is important to wash all clothing that has contact with intimate body areas at a temperature not below 60°C (140°F).

All bed linen and underwear must be washed daily (or to the extent possible) while performing anti-parasitic therapy. Affected individuals should not share their swimwear with other members of the family, and should use a separate cloth to wash his anal area. It is best to sleep wearing both underwear and pants to avoid involuntary scratching during the night. This will prevent infection through anal-hand-mouth contact because by scratching the anal area, eggs can become lodged under the fingernails. Keep pets away from the place of rest of their owners, such as beds, sofas, blankets, and cushions.

Thoroughly wash fruits and vegetables in water and soak them in CD or CDS solution for a few minutes. Clean the sink with alcohol, as the eggs of many parasites are immune to the pH of normal cleaning products such as soap or bleach. It is important to note that parasites do not leave any kind of immunity behind in the host, therefore, once cured, the person who has suffered can suffer from them again. The only surefire method of killing the eggs of *Ascaris lumbricoides* is in water above 60°C (140°F) or with 96% grain alcohol (Everclear).

Evolutionary Cycles of Intestinal Parasites

Although there are many more, here is a description of the three most common types of intestinal parasites that can be found in developed countries:

Ascaris Lumbricoides **(Roundworm)**

Ascaris reproduce quickly, as a single female can lay up to 200,000 eggs each day. This parasite is very common, especially in damp conditions, and when hygiene measures are inadequate. It can affect the entire population, but mostly affects children, seriously disrupting their development and growth. It's so infectious that the WHO estimates that there are about 700,000,000 people infected worldwide, of which around 60,000 cases end in death per year, mainly children.[5]

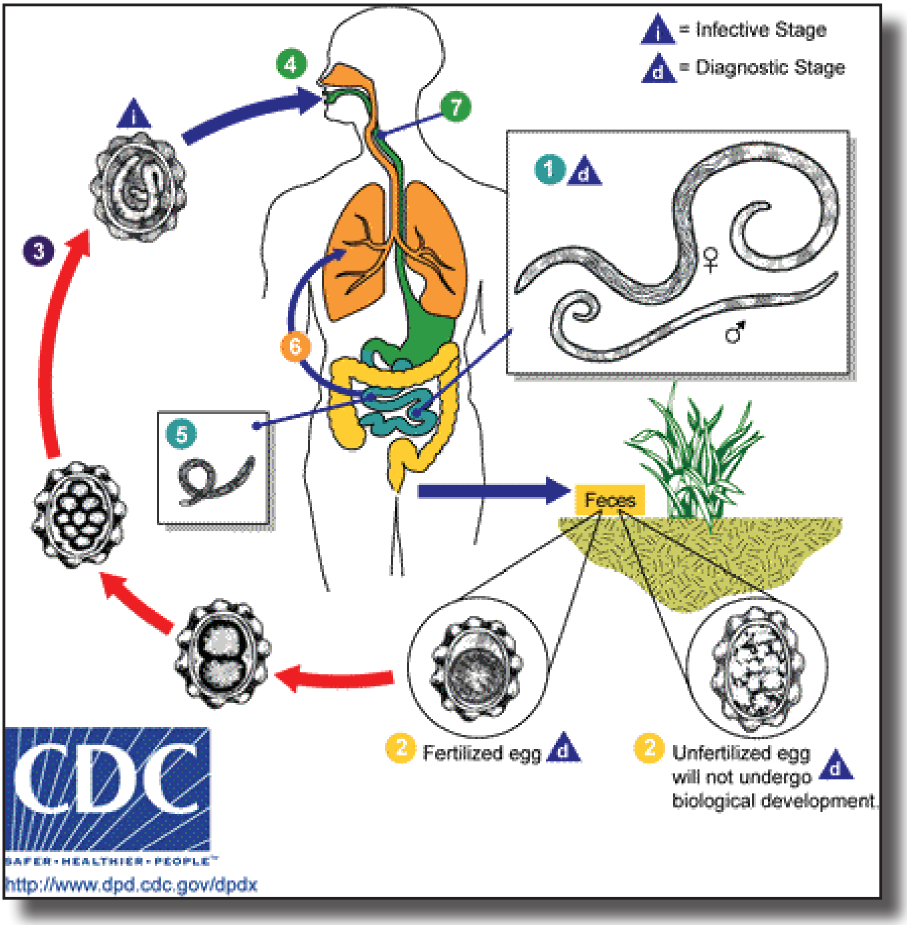
i = Infective Stage
d = Diagnostic Stage
4
7
i
1 d
♀
♂
3
6
5
Feces
2 Fertilized egg d
2 Unfertilized egg will not undergo biological development. d
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx

Parasite eggs reach the duodenum through the mouth of the host. Gastric juices rid the eggs of their shells and release the larvae. These larvae, which are highly mobile, penetrate the duodenal mucosa and migrate to the liver. From there, they continue their migration to the heart, reaching the lungs through pulmonary circulation, and finally become trapped in the pulmonary capillaries. Here, the larvae break the thin walls of the capillaries and penetrate the alveoli, bronchioles, and bronchi at which point they are able to travel up through the pharynx. Once the eggs pass the epiglottis (see diagram above), the larvae are swallowed such that they return to the duodenum, where they complete their maturation process. This process takes approximately two to three months to complete; therefore we calculate that to find ourselves completely parasite free, the initial treatment should be done for at least 12-18 months, possibly longer. From there on out, you may only need to follow a routine deworming two to four times a year.

Keep in mind that eggs are expelled through the feces (human or animal) into the environment where they can survive even in harsh conditions, favoring the persistence of the parasite. They are resistant to low temperatures, desiccation, strong acids, soaps, chlorine, formaldehyde (pH between 2 and 11.5), and can live in planted soils for five or more years, creating a "parasite hotbed" that makes them almost indestructible. Once dry, they are transported through the air, fly in air currents like dust that can be inhaled and/or swallowed. We have recovered eggs from nasal mucus, paper currency, potting soil, dust, and in indoor airborne particles, etc.

Taenia Saginata and *Taenia Solium* (Tapeworm)

Taenias reach humans when humans eat their eggs, by the consumption of tissue from infected cattle or pigs. In infected animals, the larvae are encysted in the muscle tissue. If the infected animal is consumed, development may proceed in the human digestive system. Humans are hosts for *T. saginata* (beef tapeworm) and *T. solium* (pork tapeworm). The tapeworm is considered to be solitary, because usually no more than four species are found in any one individual host. The danger of this parasite is that the larvae can migrate to the brain, or other vital organs (cysticercosis). Tapeworms may be detected by identifying segments in the host's stool that the worms discard as they grow. However, tapeworms may go undetected for many years, living asymptomatically within their host.

According to classification, they can vary in size ranging from 2 to 12 meters in length. They consist of a head called the scolex, which attaches to the intestine by means of suction cups, and a body consisting of repeating units called proglottids. A single Taenia can grow from 1,000 to 2,000 proglottids, depending on the type. A tapeworm produces an average of 720,000 eggs per day.

Pinworms (*Enterobius Vermicularis*)

Humans are considered the only host of what are commonly called pinworms (*Enterobius vermicularis*). This type of worm is the most typical in the family because it propagates easily. It is common for children to become reinfected in schools, through contact with others, or through anal-hand-mouth contact. Pinworms have an elongated shape, are whitish in color, and are about 1cm long. They inhabit the large intestine of humans. Female pinworms leave eggs around the anus. Once deposited, the eggs are infectious for a period of up to 20 days. Once in the intestine, it takes between five and eight weeks to develop into adult worms. The most important symptom is intense itching that occurs in the anal area, especially at night. In women, inflammation of the vulvar area is very common.

A pinworm infection is, generally speaking, not very serious. Unlike other parasites, they infect only humans. Transmission from person-to-person happens by handling clothing, bed sheets, towels, and environmental surfaces (such as curtains, carpeting) contaminated with pinworm eggs, which are so light that they are able to become airborne. A small number of eggs can be integrated into air particles that when inhaled follow the same developmental process as ingested eggs. Enemas are extremely useful in removing this parasite from the large intestine.

Graham's method is a simple method of detection. Just after waking and before a bowel movement, press a piece of tape against the anal folds. The tape will catch the remains of eggs and/or parasites that are situated there. With the naked eye we can see small worms no more than an inch long, but with a microscope, many transparent eggs from females and even other species may be seen.

The Importance of Lifelong Deworming

Once we begin the process of deworming, we should recognize that we must maintain this habit of cleaning for the rest of our lives to enjoy good health. It is common among people who have pets, to follow the recommendations of their veterinarians, and deworm their pets every three months. It is interesting to ask why family physicians do not give the same advice to humans. Perhaps some physicians ignore this information, or simply do not consider it important to eliminate these pests, which are just as harmful to people as they are for animals.

It is true that many parasites are not endemic or common outside certain climates, but human migration and global marketing of food products have facilitated the spread of many parasitic pests silently. It is important that we understand the lifecycle of each parasite, from birth to death including reproductive and death stages. This information is crucial for the complete elimination of the parasite. For example, in the case of intestinal parasites treated here with this protocol, some can live in the host for up to ten years, as in the case of a single Taenia, while others may remain in the host for a lifetime, reproducing again and again, as in the case of the pinworms or the well-known Ascaris.

The Kalcker Parasite Protocol & Lunar Cycle Timing

In the modern civilization in which we live, we have lost touch with much of the ancient wisdom of the past. One of the things we have forgotten is how the natural cycle of the moon influences many of nature's routines. This is especially true for the behavior of parasites. They are known to sync their life-cycle with that of the lunar cycle. Your child may demonstrate extreme behaviors on certain lunar cycle days… especially on the full moon and sometimes even the new moon.

Therefore, to maximize effectiveness, this Parasite Protocol is specifically timed to the lunar cycle. Appendix 10, page 477 provides an easy reference for you to look up the days to perform the Parasite Protocol, which is administered over 19 days—numbered 0 to 18—each month; starting three days prior to the full moon and continue during the waning moon. This period of the moon's cycle is very effective for deworming because many nematodes (parasitic worms) travel back into the intestine to mate at this time.

Length of Treatment

This protocol is not a one-time treatment. You should plan for at least 12 to 18 months to insure you have purged multiple parasite life-cycles and continue beyond 12 months if your child is still expelling parasites.

Building on CD

This protocol builds on what you have already learned using CD. During treatment, it is absolutely necessary to continue CD dosing, CD baths and CD enemas.

Tape Worms

This protocol is specifically designed for the deworming of large intestinal parasites, especially round nematodes such as Ascaris. It is effective for most nematodes, but may not be effective against tapeworms. In the case of infestation by Taenias, the recommended treatment is Niclosamide, the preferred medication due to its low toxicity.

Components of the Kalcker Parasite Protocol

This protocol uses some of what you have already learned and should already be doing, along with a set of ingredients you will need to have on hand before you start (shown in bold below). Here is an overview list of the items you will need to have on hand, and the items are described in detail on the pages that follow:

- Meal time (1, 2, 3)
- CD / CDS (4-19)
- CD Baths (20)
- CD Enemas (21, 22)
- Ocean Water (23, 24, 25) (see page 115)
- **Diatomaceous Earth** (26, 27)
- **Lepidium Latifolium Extract (Rompepiedras)** or **Chanca Piedra (Stone Breaker)** (28, 29)
- **Pyrantel Pamoate (Combantrin®)** (30, 31)
- **Mebendazole** (32-36)
- **Castor Oil** (37)
- **Neem** (39, 40)
- **Probiotic (usually THERALAC®)** (41)

Check the following website for the latest information on where to find these products:

www.ProtocolSuppliers.com

You may have noticed one or more numbers in parenthesis following each of the previous items, such as "(40)" for the probiotic. The timing of what to give when, is covered in great detail in the daily charts starting on page 198. These numbers match those found on the daily sample charts to make it easier for you to connect the dots and also be able to identify related notes under each chart. They have nothing to do with the quantity/dosing of any substance. The use of these numbers allows us to comment on specific items and when and where they come into play as shown on the daily sample charts.

We now detail each one of the items on the list above and discuss what you need to know about them and how to acquire them (including their associated numbers on the charts).

Meals (1, 2, 3)

Obviously meals are part of everyone's day. The purpose in mentioning them here is that many of the following steps are related to meal timing. Some actions or ingredients are taken before breakfast, others during or after.

In our example charts, we make some assumptions which include:

- Breakfast (1) is at 7:30am
- Lunch (2) is at 12 noon
- Dinner (3) is at 5pm

CD, CDS or CDH (4-19)

As you learned in Chapter 5, you continue to dose CD (or CDS/CDH) as before during the Parasite Protocol. Nothing changes in that regard with the Parasite Protocol building on top of those steps already in place.

The sample charts assume your child is going to school, and so you may not be able to administer doses of CD during his school day (7-12) unless you are home-schooling, in which case you are encouraged to give hourly CD doses, even if the total exceeds 8 for the day.

CD Baths (Optional) (20)

Our sample charts assume you administer a CD bath just before bedtime. See pages 113 for more information about CD baths.

CD Enemas (21, 22)

Ideally, give your child a CD enema in the morning (21) and another one in the evening (22). However, if your child is going to school, a morning (21) enema may not be a good idea due to the possibility of an "accident." Therefore, consider the morning enema optional, but the night time enema a must do! See page 103 for more information.

Ocean Water (23, 24, 25)

Supplementing ocean water minerals is important to support the body through the detoxification process. See page 115 for more information on ocean water.

Dosing:

(23) A dose of ocean water should be administered upon waking, but five minutes apart from CD dosing.

(24) One dose of ocean water immediately after school (or at lunch time if at home).

(25) One dose of ocean water, 15 minutes before or after dinner.

Diatomaceous Earth (DE) - Food Grade (26, 27)

Diatoms are unicellular plants that existed by the trillions in our oceans over 300 million years ago. They are encased by a cell wall that is made of silica. When diatoms die, this microscopic coating deposits at the bottom of oceans. Over time, they pile up in banks forming deposits thousands of meters in size. With the receding of the oceans, these deposits have been uncovered. Through compression, and ultimately fossilization, these silica deposits have given rise to a chalk rock called diatomaceous earth.

DE is an inert, nontoxic compound, which contains a number of minerals such as manganese, magnesium, iron, titanium, calcium silicates, and others. Properly ground, the skeletons of microscopic diatoms become sharp silica needles, harmful to parasites, fungi, yeast, worms, and amoebas. However, these needles are harmless to humans and other warm-blooded animals. Although it is safe to consume diatomaceous earth continuously, the best method (as with everything else) is to allow for periods of rest. During the 18-day treatment, take two teaspoons (5ml) twice a day.

Dosing: ½ to 1 teaspoon twice a day for smaller kids, 1 teaspoon three times a day for adults and bigger kids. Mix with a little water and drink. Given on days 1 through 18. DE mixes well with water but never dissolves. Stir the DE/water slurry vigorously and drink immediately before the DE settles to the bottom. Some people take heaping tablespoons in water, but larger amounts are not necessary. DO NOT take dry!

Note: In rare cases, DE can cause constipation, which can usually be managed by reducing the dosages to 1/2, 1/4 or even 1/8th of a teaspoon. If that doesn't resolve the issue, remove DE from the protocol and continue with all other directions.

Source & Cost: Search online for "Food Grade Diatomaceous Earth." Buy at least 1 pound and expect to pay about $20 more or less. Better yet, buy a five pound bag, which will reduce your cost per pound. DE does not expire or degrade, but should be kept in a dry container. Note: Diatomaceous Earth is often used as non-toxic element in pool filtration systems. You DO NOT want to use this kind of DE since it has been processed. Only get "Food Grade" Diatomaceous Earth!

Lepidium Latifolium Extract (aka Rompepiedras or Pepperwort) or Chanca Piedra (aka Stone Breaker) (28, 29)

1.7 fl.oz.
PEPPERWORT
natural extract
Lepidium latifolium L.
SoriaNatural
ROMPEPIEDRAS
PEPPERWORT

50ml℮
ROMPEPIEDRAS
extracto natural
Lepidium latifolium L.
SoriaNatural
ROMPEPIEDRAS
PEPPERWORT
50 ml℮/1.7 fl.oz.
PHARM HERB PHARM
LIQUID HERBAL EXTRACT
STONE BREAKER
(Chanca Piedra)
COMPOUND
HERB PHARM
HERBAL SUPPLEMENT

Lepidium Latifolium (Rompepiedras) and Chanca Piedra (Stone Breaker) both break up hard substances in the body. The reason we use it in this protocol is two-fold: It breaks up the protective outer coating of parasites, and annihilates oxalates, which many of our kids have an abundance of because oxalates are produced by parasites.

Note #1: It is often simply referred to as "RP" in our discussion forums.

Note #2: If you are having trouble finding this particular ingredient, don't let its absence stop you from starting the protocol with all other ingredients.

Dosing: 15 drops for a 100 pound child mixed in with the diatomaceous earth/water. Use seven drops for smaller children.

Source & Cost: You have a choice of two herb extracts; Lepidium Latifolium Extract (aka Rompepiedras or Pepperwort) and Chanca Piedra (aka Stone Breaker). One US source is www.mightyguts.com, which sells a 50ml dropper bottle of Pepperwort for about $30. A primary manufacturer in Europe is Soria Natural from Spain that labels their product Rompepiedras, while they also have an English labeled box that says Pepperwort. They both show Lepidium Latifolium on the box.

Pyrantel Pamoate (30, 31) (Trilombrin/Combantrin®)

Pyrantel pamoate is a broad spectrum anthelmintic, which works by causing a neuromuscular block that produces spastic paralysis of the parasite, and its subsequent expulsion by intestinal peristaltic action, without excitation of the parasites or encouragement of their migration. Pyrantel pamoate acts over a short duration, and tends to be completely eliminated from the body in the feces and urine within three to four days. Pyrantel pamoate is poorly absorbed from the gastrointestinal tract, and approximately 6 to 8% total is found in the urine, with the remainder in the feces. The recommended dose is one daily dose of 10 mg per kilo.

Pyrantel pamoate is incompatible with the use of piperazine, because the two substances neutralize each other. Thus Pyrantel pamoate should not be combined with pumpkin seeds, which contain piperazine, or with antiparasitic drugs that contain piperazine in their formulation.

Dosing: Pyrantel pamoate is given only twice during one cycle of the parasite protocol; once during breakfast on day one (30), and again during breakfast on day five (31). Dose is based on weight and calculated by multiplying your child's weight in kilograms times 10mg of pyrantel pamoate. To make it easy, refer to the following chart:

Pyrantel Pamoate (Trilobrin/Combantrin®) Dosing by Weight		
Pounds	**Kilograms**	**Dose in mg.**
20	9	91
40	18	181
60	27	272
80	36	363
100	45	454
120	54	544
140	64	635
160	73	726
180	82	816
200	91	907
220	100	998
240	109	1089

Pyrantel pamoate (often just referred to as Combantrin®) is available in three forms:

- **Liquid:** where each milliliter contains a certain number of milligrams. For example, one available formulation contains 144mg/ml. So a 100lb child would take 3ml.
- **Tablets:** where each usually contains 250mg.
- **Capsules:** where each usually contains 250mg.

You will have to read the label or the particular product you acquire and determine the milligrams to use.

Source: Combantrin® is available by prescription in the US. Most other countries have it available over the counter.

The preferred source of pyrantel pamoate is a compounding pharmacy, as to avoid coloring and flavoring. If you are unable to find it without coloring/flavoring then I would personally use mebendazole for the entire 18 days rather than risk giving your child an ingredient which may cause regression.

Note #1: Some brands of Combantrin® include mebendazole. You want the stand-alone Combantrin®!

Note: #2 Pumpkin seeds should not be consumed with pyrantel pamoate because they neutralize its effects.

Mebendazole (Vermox®/Lomper®) (32-36)

Mebendazole is a drug used in treating diseases caused by helminths (parasites of the gastrointestinal tract). This drug prevents the parasite from using glucose, which results in a decrease in energy and therefore death of the parasite.

Mebendazole is a non-systemic drug which means it is only absorbed, to a limited extent, in the gastrointestinal tract (approximately 5 to 10%). However, if it is consumed with fatty foods then more absorption occurs.

Approximately 2% of the administered mebendazole is excreted in the urine, while the remainder is excreted in the feces. The appropriate dose of mebendazole may be different for each patient as it depends on the type of parasite causing the infection.[6] The most frequently recommended dose is 100mg for children, 200 mg for adults, two times a day for seven of the first nine days of the Protocol.

Adverse effects from mebendazole are generally rare due to its poor absorption. However, it may cause nausea, vomiting, abdominal pain, and diarrhea. Normally these effects are in fact a result of the release of toxins from the very death of the parasite itself. Anti-parasitic drugs can be administered very effectively by diluting them in water, putting the mixture in a small bulb enema, and inserting the anally. This is especially suitable in the case of oxyuriasis (pinworms). Read more about this "implant method" on page 220.

Dosing:

Note: Indicated weights should only be considered a rough guide.

Small Children (20-40lbs.): Days 2, 3, 4, 6, 7, 8 & 9 — Take as little as 25mg of mebendazole with breakfast and dinner, and do NOT do a blitz on Day 9.

Children (41-70lbs.): Days 2, 3, 4, 6, 7, & 8 — 50mg with breakfast and dinner. On Day 9 they can take 50mg, 50mg and 25mg.

Adolescents (71-100lbs.): Days 2, 3, 4, 6, 7, & 8 — 100mg during breakfast and 100mg during dinner. Day 9 is "Mebendazole Blitz Day" where you administer ONE 200mg dose during breakfast; ONE 200mg dose at lunch; and a final 100mg dose at dinner.

Teens & Adults (101 lbs. and up): Days 2, 3, 4, 6, 7, & 8 — 200mg during breakfast and 200mg during dinner. Day 9 is "Mebendazole Blitz Day" where you administer ONE 500mg dose during breakfast and no dose at lunch or dinner for the remainder of the current cycle.

Source: Mebendazole is available by prescription in the US and over the counter in other countries.

Note #1: Some brands combine mebendazole with Combantrin®. You want the stand-alone mebendazole!

Note #2: Mebendazole is mostly sold in tablet form, but it is also available in liquid. DO NOT buy the liquid form—stick with the tablets! I have seen horrible reactions from the "inert" vehicles used in the liquid products.

Castor Oil (37)

Apollo
CASTOR OIL

Castor oil is extracted from the seed of a plant called Ricinus communis ("Higuera del diablo"). Its seeds contain between 50-80% oil, which has a high content of ricinoleic acid, which has excellent laxative and purgative properties. Once you begin anti-parasitic treatments, spastic paralysis may occur in some parasites and many together may form a "knot" of worms that can cause intestinal obstruction. It is important to help your body purge them by using castor oil. Castor oil should be taken in the morning, two hours after breakfast and other medications. If your child goes to school they take it as soon as they walk in the door from school. A typical dose for a child is 1/2 tsp to 1 tsp, or up to tolerance. The adult dosage is 15 to 30ml (two tablespoons), two hours after breakfast and other medications. If you experience any intestinal distress, mineral purgatives such as Epsom salts, or vegetable purgatives such as senna leaves, can be used.

Castor oil is also available in gelcaps for those who dislike the taste.

Dosing: The amount to administer varies and really depends on the individual's tolerance. A good starting point is ½ teaspoon for smaller children and up to two tablespoons for larger children and adults. Only experimentation will determine the right amount, if castor oil causes diarrhea.

Source & Cost: Readily available in liquid form at most pharmacies in the laxative section. Usually under $10 for 16oz. Also available online.

Neem (Azadirachta indica), Caps or Tea (38, 39)

PREMIUM
Nature's Way
HERBAL
Neem
Leaves
100 Capsules / Certified 475 mg
DIETARY SUPPLEMENT
Recommendation: Take 2 capsules three times daily, preferably with food.
Supplement Facts
Serving Size 2 Capsules
Servings Per Container 50
Total Carbohydrate 1 g
Neem (leaf) 950 mg
Other ingredients: Gelatin (capsule)
Actual Capsule Size

The neem tree is a great natural inheritance of mankind. References in Sanskrit scriptures and ayurvic medicine practices indicate the use of neem since ancient times in Hindu medicine. Even today, Hindus living in rural areas call the neem tree the "village pharmacy" for its ability to alleviate many diseases and is currently endorsed by authorities in India for its use in medicinal preparations. Neem is one of the purifying and detoxifying plants with the greatest potential. Neem has been used to combat all forms of body parasites, external and internal parasites alike. To prepare neem, boil four leaves (normally the contents of an envelope) in one liter of water for five minutes. Drink the tea throughout the day over the course of each parasite protocol.

Dosing: Neem is given each parasite protocol from day 10 through Day 18.

You have a choice of caps or tea. I prefer caps over tea because the taste is strong and unpleasant, so some kids will buck drinking tea.

Caps (assuming 475mg each): An adult takes six in a day, three times two caps at meal times.

Follow the directions on the bottle. Give a full dose for teens and adults 100 lbs. and over. Small children receive ¼ to ½ dose.

If using caps, give one dose during breakfast and one dose at dinner.

Tea: Give four doses throughout the day. Prepare a tea from the leaves, one tea bag in one liter of water (add stevia if needed to cover some of the bitter taste). One tea bag usually contains approximately four leaves. If using loose leaves, then make one liter of tea with four neem leaves. If using crushed leaves, then use approximately one slightly heaping teaspoon.

Neem Tea Dosing

Weight of Person	Total Daily Amount (To be split into 4 doses)
20-34 lbs	100 ml
35-49 lbs	200 ml
50-64	300 ml
65-84	400 ml
85-109	500 ml
110 and over	600 ml

Source & Cost: Neem capsules cost about $8 for 100x 475mg caps. Look in your local health food store. One popular brand in the US is Nature's Way®. BioPure™ sells a product called *Neem Synergy* which contains a few additional herbs that do not affect the neem.

Neem tea can be purchased in bags or as dried leaves.

THERALAC® (Probiotic) (40)

THERALAC
PROBIOTIC MASTER SUPPLEMENT
5+2
BIOTHERAPY
30 CAPSULES

THERALAC® is a probiotic that should be given during the parasite protocol to help reestablish good gut flora. Ideally, it is rotated every other month with THERALAC® TruFlora®.

The key reason for recommending THERALAC® over other probiotics is explained in this paragraph excerpted from their website:

TruFlora
Probiotics + Enzymes
ACID-PROOF

> *THERALAC® probiotics survive transit through the harsh acidity of the stomach and arrive alive in the intestinal tract. THERALAC's ACID PROOF™ technology utilizes sodium alginate from seaweed in a unique formulation that assures survival at pH 1.6 for 90 minutes, the most severe stomach acid conditions normally experienced. Other probiotics that claim acid resistance are tested at pH 2.5 – 3.0, or >10 times less acidic than pH 1.6 – not a fair test. THERALAC's ACID PROOF® technology is protected by US Patents 7,122,370 and 7,229,818. This technology goes beyond resisting stomach acid and involves keeping the probiotic cells together in a viscous alginate-gel moving in group-force, not as separate individual cells like other probiotics, deep into the intestinal tract while retaining key formula ingredients, LactoStim® and Sodium Alginate in close proximity.*

Dosing: One capsule of THERALAC® is given each and every night at bedtime, irrelevent of age/weight. This probiotic is also to be given outside of the 19 Parasite Protocol days.

THERALAC® can be given at the same time as CD if your child swallows capsules. However, if you are using the THERALAC® powder form then give it at least five to ten minutes after the last CD dose of the night. See their website for more information about the powder form.

Special exception: Some people do not tolerate probiotics, in which case we have no choice but to leave them out. In some of these cases, sauerkraut and fermented veggies may help to cultivate beneficial bacteria.

Source & Cost: Amazon carries "THERALAC® 30 caps by Master Supplements Inc." for $37.

Sample Parasite Protocol Daily Calendar

To assist you in understanding the Parasite Protocol and how it changes from day to day, we have prepared the following visual set of daily charts showing how all the pieces fit together. To prepare this set of charts, we made up the following times for example purposes only:

- Your child's awake time is between 6am and 10pm.
- Breakfast is at 7:30am
- Your child goes to school, which starts at about 9am
- School is out at about 2:30pm
- Dinner is at 5pm

As stated earlier, each dose or activity is labeled with a unique number. Those numbers relate to notes below each chart AND they match longer descriptive notes in the previous pages.

You are encouraged to copy/enlarge the blank chart in Appendix 9 (page 475) and workout your real-life schedule based on these example charts.

Refer to the lunar calendar (Appendix 10, page 477) for the exact days the protocol should be followed, as well as to get an idea of behaviors related to parasites around the full and new moon.

After four years of biomed and no major gains, we decided to try mms/cd protocol. We started in April and are currently on our second parasite protocol. Initially, we were really intimidated by the enema part but realized if we were apprehensive our 7 yr old daughter would pick up on that. We told her that what we were about to do would help her feel better we gave her the iPad right before the procedure and a new fav toy (beanie boos) and she was game. After it was over she said "no more booboo in my tummy". We all cried. To make a three-month adventure short and sweet, her team at school cannot figure out how her fragmented speech has gone from 2-3 word utterances to long, drawn out sentences. Her expressive language and receptive have increased. Her auditory processing speed has quadrupled. The most amazing part thus far has been the increase in social skills. Or just the fact that she is interested in engaging. Socially she was nonexistent and is now at a four year old level. We just spent a week with family and she was saying hi and engaging without being prompted!!! She now has a million questions, wants to know when and where we are going and the sequence in which events will take place. All this from a girl who never asked questions before mms! STOP thinking about trying! Just Do It!! For the first time in seven years, I can hear the bells of freedom ringing!!!! Thank you Kerri Rivera !!!!!

Day 0 (3 days before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1											
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3											
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 0 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
40	Probiotic at the end of the day

Day 1 (2 days before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28	30				
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29					
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 1 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
30	Dose of Combantrin with breakfast
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 2 (1 day before Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 2 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 3 (Full-Moon)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 3 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 4

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 4 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 5

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28	31				
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29					
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 5 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
31	Dose of Combantrin with breakfast (NO Mebendazole today!)
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 6

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 6 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 7

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 7 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 8

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		32			
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		33			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 8 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
32, 33	Small Child Dose: 100mg Mebendazole with breakfast and dinner. Teen/Adult Dose: 200mg Mebendazole with breakfast and dinner.
40	Probiotic at the end of the day

Day 9 (Mebendazole Blitz Day)

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28		34			
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM									35			
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29		36			
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 9 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
34	Blitz Day - Small Child Dose 1: 200mg Mebendazole with breakfast. Teen/Adult Dose: 500mg Mebendazole with breakfast ONLY!
35	Blitz Day - Dose 2: 200mg Mebendazole with lunch (or right after school) Teen/Adult Dose: N/A
36	Blitz Day - Dose 3: 100mg Mebendazole with dinner. Teen/Adult Dose: N/A
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
40	Probiotic at the end of the day

Day 10

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 10 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Start dosing Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 11

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 11 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 12

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 12 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 13

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 13 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 14

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 14 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 15

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 15 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 16

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 16 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 17

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM												
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24							
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 17 - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Day 18

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM		4		21	23							
6:30 AM												
7:00 AM		5										
7:30 AM	1					26	28				38	
8:00 AM		6										
8:30 AM										*		
9:00 AM		7										
9:30 AM												
10:00 AM		8										
10:30 AM												
11:00 AM		9										
11:30 AM												
12:00 PM	2											
12:30 PM												
1:00 PM		11										
1:30 PM												
2:00 PM		12										
2:30 PM												
3:00 PM		13			24					37		
3:30 PM												
4:00 PM		14										
4:30 PM												
5:00 PM	3					27	29				39	
5:30 PM												
6:00 PM		16		22								
6:30 PM												
7:00 PM		17			25							
7:30 PM												
8:00 PM		18										
8:30 PM												
9:00 PM		19	20									
9:30 PM												40

Day 18 (Last Day!) - Notes:	
1, 2, 3	Breakfast, Lunch & Dinner
4 - 19	At least 8 CD doses throughout the day, preferably more, at least 30 to 60 minutes away from food.
20	CD Bath at the end of the day.
21, 22	Enema in the morning (optional) and in the evening (mandatory).
23 - 25	3 doses of ocean water at least 5 minutes away from CD.
26 - 29	Diatomaceous Earth and Lepidium Latifolium with breakfast and dinner.
*** 37**	Castor oil 1 hour after breakfast or immediately upon return from school.
38, 39	Neem Caps with breakfast and dinner. If using tea, 4 doses throughout the day.
40	Probiotic at the end of the day

Parasite Protocol *Off Days*

Days "19" through to the next "Day 0" are "Off Days" where you discontinue parasite meds and herbs. Here's a simple chart showing what to continue doing and what to stop during this *off* time:

Continue with these:
CD / CDS / CDH Dosing
CD / CDS / CDH Baths
CD / CDS / CDH Enemas
Ocean Water
Probiotic

Stop these:
Diatomaceous Earth
Lepidium Latifolium
Pyrantel Pamoate (Combantrin®)
Mebendazol
Neem
Castor Oil

Of course this represents a hypothetical case and your situation may call for taking other meds or supplements.

Detection of Parasites in Stool

It is necessary to detect the parasites by observing the stool carefully. For that we use a small plastic basin, and a plastic stick or fork for examination.

Author's Note: One of the moms who is a part of our forum came up with some guidelines for processing your child's stool for parasite identification. She calls it, "Everything You Wanted to Know About Sorting Through Poop." Here are her suggestions:

Supplies:

- rubber gloves
- paper plates
- plastic forks (plastic sticks, chop sticks, or plastic back scratcher)
- a pen
- a coin
- toilet hat

I like to use a plastic "toilet hat" also known as a specimen collector or a specimen collection unit, which goes under the toilet seat and collects the stool before it sinks into the bottom of the toilet (available on Amazon.com).

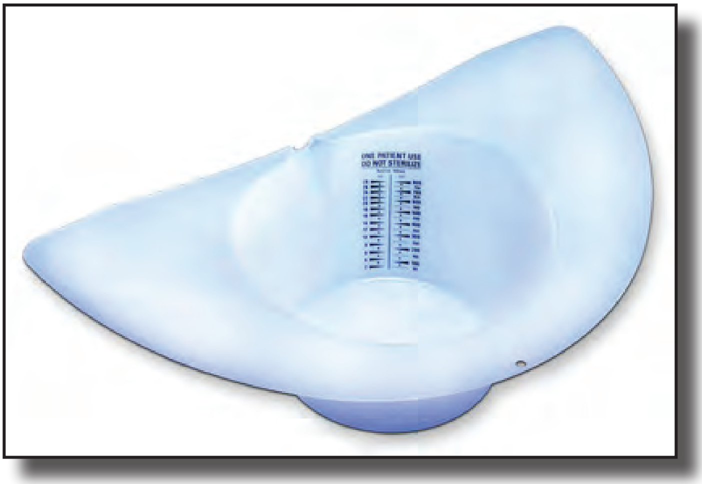

The Collection:

When your child poops, it is collected in the toilet hat (specimen collector). After I get my child cleaned up and taken care of, I remove the toilet hat from the toilet and put the specimen on a paper plate with my plastic fork. I have a look for anything interesting and then transfer that part to a clean plate using my trusty plastic fork. I discard the remainder of the specimen into the toilet, flush and put that dirty paper plate in the bathroom garbage. (We now line with kitchen garbage bags, and I change it after each of these poops.) On the clean plate with the suspected worm, I may add a bit of water and swish it around to get the worm cleaner. I then may transfer the worm to a third plastic plate to get a clear picture. On the clean plate with the washed worm, I write the date, and the initials of the person the worm came from. If you need help identifying the worm, place a penny next to the worm (for size context), snap a picture, and mail the image to kerri@cdautism.org, (Kerri collects the photos for documentation purposes, so send those worm pictures). Then flush the worm and put all paper plates, gloves, and fork in garbage and take it outside. Now, you can go find out what your child has gotten into while you were doing all this.

Clean Up:

Use HOT water (60°C/140°F), and sterilize with 96% (180 proof) grain alcohol (Everclear).

Microscope

It is very useful to have a microscope for diagnosis because it enables you to see both the small parasites that may appear in the blood, as well as eggs or larvae in the feces. This way we are more accurately able to determine if the number of parasites decreases. A simple microscope that costs about $100 is suitable for this kind of identification. The easiest method to determine what you see is to compare your sample to images you find on Google. This way you can enlarge the images, see various samples from different angles, and get a much wider variety of samples than if you were comparing a sample with many textbooks.

Bulb Enemas (aka "Implant")

To prevent anal itching from pinworms, night wakings, etc., you can use a bulb enema or small catheter/syringe with a dilution of 50mg of mebendazole in 10 to 15ml of water for small children, or 100mg of mebendazole in 15 to 20ml of water for larger children/teens/adults. The best way to do this is to introduce the medication together with the water in the rectum immediately before bedtime and hold overnight. If you are using the "implant," a morning enema is mandatory.

So we did the 100mg of Mebendazole with 20ml of water. He kept it in and we did it right before bed... Next day a ton of worms came out!!! Better yet he started singing songs with a tune - 3 different songs over the last 2 days and counting on his fingers very purposefully! His aide at school said he did amazing today - no behaviors, very focused, better articulation and better social interest.

The Worm Whisperer

The following words of wisdom are from a mom who has earned the nickname *The Worm Whisperer* (although her ability to destroy parasites may make the *Worm Ninja* more apropos). Her dedication and diligence have not only made the difference in the life of her son, but countless other children who are what she calls "extreme cases." The information presented here may be the difference between an older/aggressive child living with their family or going to a group home for the care that their family is no longer able to provide them. Thank you Robin, for being a pioneer and for never giving up. Thank you for having the guts, the know-how, and the generosity to share what you have learned with those who need it most.

Extreme Cases by Robin Goffe

This section is for older children who may be violent, self-injurious, destructive, physically aggressive, high risk and/or bedridden.

Some families have children who are older when they first start treatment. My son was 18 years old. We call ourselves the "last chance" group. We say this because our children have lived their whole lives infested, unbeknownst to us. Our children lived as happy, relatively easily manageable, learning-disabled children: perhaps slightly annoying with their routines of movie dialogue scripting or attachment to Disney movies, maps, or little known facts. They were mostly friendless but pretty easy to take care of. However, during the hormonal teen years, things took a grave turn. We may have chalked up their newfound solitude to just giving up socially for lack of friends. They did not fit in. We had no idea that something else was brewing. The hormones within them began to have a bitter war with the parasites living inside them, and a war would soon break out. Their mental state would diminish considerably. What we are faced with are grown, strong children who are mentally ill and violent, self-injurious, and destructive. Some have even become so ill that they had to stay in bed.

This "last chance" is an opportunity to try ONE MORE THING before putting them in a group home; to give them away—for their own safety and the safety of their families.

These things that I am going to share with you are unique to the standard treatment outlined in the book. It is a more aggressive treatment because it is needed. Our children are so highly infested that it is shocking to tell you what we have found. There is a lot to do, but there is a method. It is time

The Goffe family.

consuming and it is a lot of work. But here I will outline what we did to get our son from an aggressive, self-injurious, destructive young man to one who has regained patience, language, understanding, and reasoning; but best of all, the opportunity to remain in our home. With continued treatments, he has the hope for a future: a job perhaps, and maybe even a family of his own.

Beginning treatment for the older child and especially one who is aggressive, self-injurious, violent, and destructive is one done starting low and slow.

Day one = one drop.

I kept a very detailed journal, and I will simply tell you what we did. I will tell you what worked for us and what did not. I will also tell you what I shared with others that worked for them as well. I am not a physician. I am not a chemist. I am a mother in love with my son and his utmost comfort was my main concern. My only goal was to clear his parasites. I had no idea that the things I learned by digging through his stool for 9 months would eventually give me the nickname, "The Worm Whisperer." Although intestinal parasites have been around longer than man, and they are smart enough to live within a human being lifelong, they can go completely undetected. I was ready to plot out a war against them. By the beginning of our 9th month of treatment, I estimate that I had cleaned and examined 35 pounds (16 KG) of solid parasites. Hard to believe? I have 80% of them photographed. The last two months in jars total seven pounds and a running total of the combined length will be at over 200 feet (61 meters), at the publishing of this book. There is not a classroom in the world that can teach what I learned in my bathroom day after day in latex gloves and a hospital mask. There was a fan blowing on me to manage the stench and by the third week my gag reflex was under control.

I examined the parasites, the pieces, how did it die? What was my son feeling at the time? Was he violent? Sweating? Slapping his legs from the pain of the parasites moving around and torturing him from the inside? I needed to eliminate these worms without allowing them to cause my son pain as they were killed. I had no idea the monsters I was to face.

In March of 2013, we started our son on one drop of chlorine dioxide. One of the most feared things that the parents of adult sized children face is the dreaded enema. It is just not something that most of us grew up with as standard care. But even so, my husband and I knew it was something that made sense. There are toxins inside these kids. They must be flushed out. The bowel is the way out. So on the very first day and the very first drop, we also explained to our severely autistic child that we were going to help his belly feel better. Our son, by this time, was so severe in regression that I can only best explain him as non-human. He was no longer speaking, unable to react to us, had stopped answering to his name, and could no longer contain the saliva in his mouth. He walked with a 12-inch drool hanging from his mouth. He ate like an animal, glaring at us. He was frequently violent: jumping on cars and denting them in, kicking down fences and destroying property.

This was the child that we were going to start giving enemas to. If we could do it, anyone can.

Each day we went up a drop; two drops on day two, three drops on day three, and so on. He was edgy and tired. We went for walks, as per usual, to deal with his aggression, but by day seven he began to have a runny nose and had a great deal of fatigue. We were so happy since we knew this was a sign to look for; that the immune system was kicking in. He slept for about 15 hours as we woke him and dosed him every hour and by the next morning there were about 25 white, hairy, thin looking objects in his stool. We knew these were worms and that we had our answer. After three weeks we began to see 4-6 inch parasites in all shapes and sizes, yet the aggression continued. It was here that I learned that what we were experiencing was POWS (Pissed Off Worm Syndrome). By this point we were at 13 drops, and while we were killing the smaller parasites, we were only pissing off the larger ones. The parasites do not like their environment disturbed, and subsequently they cause distress to the host. Parasites excrete ammonia, (possibly, leading to hyperammonemia and possibly seizures), morphine, and a meth-like substance. As these toxins enter the body, they can also cause aggression and anger. We had many possessions broken during these tantrums. We stopped valuing material objects.

It was here that we were told about—double dosing. So even though our son was getting about the equivalent of two drops an hour, if we saw increased aggression, OCD or self-injurious behavior we would give him four (see page 103 for an explanation of double dosing). If he did not calm down, we gave him another four drop dose. This was the perfect return attack for the parasites. Within a few moments the aggression settled. His red face and wild eyes diminished, and we had made it past another hurdle.

We used the double dose method dozens and dozens of times over the following months, and this helped tremendously to curtail the aggression caused by the parasites.

By the third week of treatment, we were concerned about the time during the night that our son would not be getting doses. So for the next six months our family had a dosing schedule, dosing additionally at midnight, 2am, 4am and picking back up to hourly doses from 6am to 10pm. This schedule was shared by not only my husband and I, but also our son's siblings,who would also set their alarms and take their turns, while we all rotated lack of sleep.

By the second month we found that our son's behavior not only changed around the full moon but the new moon as well. The standard protocol uses mebendazole over the full moon. However, with the older, more aggressive children we have had success with including mebendazole over the new moon cycle as well. These mini mebendazole courses feature shorter bursts of 5-6 days each, over the new and full moons beginning four days before each. We also deemed it necessary to start earlier in the moon phase than with the younger kids, since the infestation was greater, and the parasite movement occurred earlier, causing behaviors sooner.

For the child that has aggression, every parent must remember that the behavior is parasite related. During treatment you will see a variety of behaviors stemming from the parasites attempting to control their environment aka—the host. We found that increasing the frequency of the dose dampened the violence and aggression. It can be difficult to trust your child who is lost in a rage. You feel deceived and hurt. Those were our frequent emotions as well. It does pass. It took about 4-5 months for the rages to stop. They happened everywhere and were happening even on the toilet while trying to expel the parasites. Have a plan. Music worked very well for us. So did reading books. Find something to distract, and always give soothing, comforting tones. We frequently told our son how much we loved him and we rubbed his head and back. These herxing behaviors may be avoided through the use of a new method of preparing CD called CDH (see Chapter 7, page 155 for more info on CDH).

CDH: This new preparation method has really been a huge asset for the older kids that need to go higher in drops. At around 20 drops, the volume of the regular CD became unpleasant, making some children nauseous. My son was one of them. With the CDH preparation, I firmly believe that the older, tough-nut children will have more success with getting to the appropriate levels needed to start expelling the larger parasites. It is my personal belief that the amount needed in many of the non-seizure older kids with aggression, violence, self-injury, and destruction to be between 75-100ml of CDH daily. This amount may have a connection to the voltage needed in order for the mitochondria in the white blood cells to be given the power to not only kill the parasites, but to also destroy the bacteria. The job is two-fold, and must be done quickly as once the parasite is dead, the bacteria want to devour them immediately. I say these things because I have seen the condition of the parasites every day. I found that if I can kill the parasite with the least amount of disturbance or preparedness by the parasite, then there is no time for them to fight back and cause herxing. The bacteria also come into play and must be dealt with. Higher amounts of CDH kill both, without so much as a frown or a concern on the face of my son.

Stevia: I found that by adding the sweetener stevia to the CDH, the process can be more pleasant and does not affect the properties or effectiveness of the dose. As proven by Lamotte ClO_2 test strips (see page 467 for information on using the test strips).

Humidifier: In the beginning months, the infestation level is so high that nightly doses may have to be sacrificed in order to start breaking down the will of the parasites. It is not about seeing them every day, but slowly and consistently breaking their will. The constant aggressive irritation of the CDH will kill parasites, stopping them from making gains. There is another layer of nighttime treatments, and that is the humidifier. We fill a cool air humidifier that holds one gallon of water, and put 35 drops of activated CD (NOT CDH) and let it run all night near our son's head. This too was a method we used for maximum parasite elimination.

Spices and Herbs: It is urgent that we work constantly with our older kids to rid the body of parasites and pathogens. We have found that for kids that can easily swallow pills, filling empty gel capsules with the recommended herbs and spices in this chapter under Other Medicinal Plants (page 231) helpful. We have found great results by filling the gelcaps with the variety mentioned there: black walnut, ginger, rue, wormwood, and yarrow and giving two to four different ones with each meal as a toxic feast to the parasites. Don't narrow your selection to those mentioned here—get them all. There is no plan or pattern. Use any combination.

Colonics: Professional colonics have been an important addition to healing and cleansing parasites in our older kids. We started these a few months after our son was accustomed to the routine enemas. This has been a wonderful addition, and we try to do at least one weekly. Some places will allow you to do inserts, and we have taken our CDH there working up to 50-100ml CDH with great success.

Limiting Undesirable Behaviors

Probiotics: If you are using probiotics and your child is aggressive or becomes aggressive after restarting the probiotics, consider removing them as a first step. There could be such a high infestation rate that the child's system does not distinguish between good bacteria and bad bacteria.

Enemas: CD enemas should start on day one of treatment for the aggressive/older child. Since they have such a high rate of infestation, it is necessary to clear the pathogens on a continuing cycle from the very beginning. Please follow instructions for enemas starting on page 103.

Diatomaceous Earth (DE): In some children, the DE may shred the larger parasites to death, causing them to spew additional toxins into the child's body, which results in herx symptoms. Discontinuing DE for a few months may help to ease the distress, and give a cleaner kill.

Salt Baths: There are times that the body is unable to pull out the toxins fast enough. Be careful about killing too fast (dosing without or not enough enemas). If you have done enemas, and there still seems to be distress, you can buy a 40-pound bag of plain pool salt, take 9 pounds of salt and put it into a hot bath. Allow the salt to dissolve and have the child soak for one hour. If the child sweats during the bath, this is a good sign, as it shows toxins are being pulled out. Do not use Epsom salts. They contain magnesium. Magnesium feeds parasites as well as biofilm. Plain pool salt is cheap, and helps to remove toxins.

Dandelion Tea: This is an incredible detoxing source as well. There have been times that our son was in and out of a salt bath all day drinking dandelion tea to get the built-up toxins out. This is a tea ready made by Traditional Medicinals® (traditionalmedicinals.com).

Tips on What to Expect During the Parasite Elimination

Intestinal parasites have been on Earth longer than humans. Their plan is to find a host, and to continue to create life. They have the home advantage by living inside of your child. There are many things that I have learned while at war with them. They have a plan and so should you.

Situations We Have Come Across

Peetox: As you kill the parasites, they are alerted that you have taken away their joyful life. As a result, they do things to the host to show their displeasure. One of the things the parasites do is to spew toxins, including a morphine-like substance that can cause your child to not feel that he needs to urinate or that he is urinating. Naturally, this can lead to accidents. This is not the child's fault, but only a part of the parasite elimination process. One of the most difficult things we did was to put our 18-year-old son in adult diapers. It was heart breaking because it felt that we were moving backwards instead for forwards. However, this was only temporary, and for his own dignity. Especially if they go to school; perhaps being in diapers for a while is the answer. Our son was in and out of them for about two months. This does end.

Sleepytox: As you are going higher in drops, the CD is assisting the immune system to get the parasites out. It will take a lot of energy to do this. Remember, your child is the host and there is a war going on! Sometimes, your child may be tired for days. Our son slept for nearly the entire summer at the beginning of his treatment. Allow the rest they seek. Now our son gets higher doses of CD on the weekends, and he is very tired and sleeps then, too. By Monday morning, we can usually get a big fat rope worm after our efforts. Sleepytox is great because you can get a lot done on your home "to do" list.

Behaviors/OCD/Tics: Some may become alarmed because their child starts having behaviors that they did not have before treatments, or they feel that the behaviors have gotten worse. Parasites have a plan. It is to stay alive and to procreate. They want their eggs to return back into a host to repeat the cycle. So, you may see behaviors of playing with feces, touching their anus to their mouth, trying to put their fingers into your mouth or another person. There are finger and nail chewing, and spitting. These are all parasite controlled behaviors. We have found that eliminating parasites helps to minimize this behavior and it will eventually subside. There may also be a mineral deficiency attached as well. We have found that giving PLENTY of ocean water can help to tame these behaviors. Our 19-year-old son gets over 200ml OW mixed with 400ml spring water per day.

POWS (Pissed-Off Worm Syndrome): POWS presents itself when you have been dosing, and your child seems weepy, is pacing, and/or unhappy. Sometimes it is really that you have pissed off the worms, but are not killing them. Always check with Kerri regarding your concerns. But many times the answer is to go up in drops not down. You need to kill them—not piss them off, or they will fight back.

Clear Zone: This time period can be about five to seven days before the full moon; once infestation is brought down after six months or more of treatment. You may be able to start seeing glimpses of your child being lucid. This is a good time for an ATEC test.

High Volume Eating: Often times, the kids reach a point where they realize that they have been starving. The parasites have been taking a bulk of their nutrients, causing their mental deficiencies. They may even be eating non-stop all day. It is tough to juggle the timing and effectiveness of the CD with the continued eating. Allowing the child to eat high volumes of food while dosing with CD will still kill the parasites, but the CD levels may not be high enough to avoid a herx. Though the parasites may be killed the CD level may not be high enough to also kill off the bacteria eating the dead parasite. One solution we found was to give our son homemade bone broth with ocean water many times a day. This will help to replace the lack of nutrients, meanwhile reducing the overeating enough to continue the CD without risking a herx reaction.

Weight Loss: Our son lost 50 pounds during the first seven months of treatment. He was very over weight at 5’9” and 200 pounds. The weight loss brought concern to people who had not seen him for long periods. During this time I estimate that 35 pounds was solid parasites. His stomach was very bloated, as was his face. Our children are full of parasites and pathogens (i.e., yeast, bacteria, etc.), and all of these things can make them gain weight or be bloated. Our son is now in the normal weight range for his height, and has a flat stomach.

Night Terrors: During the night you may see that your child has trouble settling down, or they may wake in the night screaming your name. They may even do night walking, possibly leaving out the front door in a sleepwalk state. We saw these behaviors years ago, as our son was heading into a higher infestation level. Now that we are on the healing end, many of the behaviors return as sort of a rewind. The behaviors that came into play going down the ladder are sometimes repeated heading back up the ladder to wellness. We feel that our son was experiencing separation anxiety, and the recognition that he had total control again. This new gained independence may be frightening. At a point, our son entered the peetox stage (see above). He had no feeling or sensation to urinate, so he was put into adult diapers. As our son's infestation rate decreased, the normal behaviors like waking in the night and feeling the urge to urinate returned. Then, as he would get up to go to the bathroom, he would suddenly stop, and feel perhaps lost and afraid to continue this task alone. He would scream out or wander around whimpering. Although night wakings were disturbing, we found them to be a necessary passage for bringing our son to reality. In order to give comfort, at night we continued the CD humidifier, as well as CD ear and nose drops. Many times a CD bath before bed was calming as well.

Emotional Blindness: Because the toxins that the parasites excrete are substances that have the effects of morphine, ammonia, and histamine, these toxins can cause emotional confusion in your child. They may be confused by your laughter, smiles, and affection. Your positive and loving encounter may be perceived as threatening or challenging. Your smile may bring a rage. Any spoken words by you no matter how soothing they may be, could be perceived as a confrontation. This is difficult on both parent and child, as the child may see you as a threat and perhaps even as mocking them; while you as the parent feel rejected in your affections. Many times the child's behaviors evolve into a stoned-out loss of reality and presence. These are all related to the toxic build up and defense of the parasites. Their motivation is to remain in the host; they show their displeasure to the discomforts we have given them in our elimination attempts. Many times during these episodes, we have found it best to avoid words or eye contact, but rather to offer double doses as outlined by Kerri, and keep a calm environment. We have found that these toxic levels can be lowered through salt baths and dandelion tea.

Despite the behaviors that may be brought on by a parasite protocol, the results are amazing. In our family we are finally getting our son back. I would recommend staying the course. As we know, these kinds of parasitic infestations in immune-comprised children will not go away on their own. Children do not "grow out of" worms. In fact, the worms will continue to grow inside of them. The following parasites were expelled by my older son, by making the aggressive modifications aimed at older children. By removing these parasites we are seeing the first stages of recovery. The jars on the following page represent seven pounds of worms in less than two months.

Yours in healing,
Robin

Jar #2 Sept 30th
Oct 13 Rope

7 lbs. of 35 lbs. (total) of the parasites extracted in 9 months (160 feet in total above).

39+ inch long worm.

Other Medicinal Plants

Author's Note: The following information is extremely powerful for our continued health and well-being. However, these plants and foods ALONE have NOT been proven to heal autism. Rather, follow the 18-day deworming protocol presented starting on page 187. When this is done for 12-18 months it has proven to be an important part of the protocols that have led to the recovery of many children. With that in mind, some parents of older children, and aggressive/self-injurious children have started to implement a few of the plants and foods listed below, in addition to the aforementioned 18-day protocol. In many cases, these additions have led their children to increased gains through increased parasite elimination.

A number of other plants are also effective for deworming. If, after three months of treatment, the problem persists we can change the type of plant, or repeat any plant that was effective in previous months. We can use them in combinations, mixing several plants at once, or take them individually. Plant formulations that should be considered are alcoholic extracts, in oil or by infusion, and include the following plants:

- Clove (*Syzygium aromaticum*)
- Common Rue (*Ruta graveolens*)
- Dandelion (*Taraxacum officinale*)
- Gentian Root (*Gentiana lutea*)
- Mint (*Mentha sativa*)
- Mugwart/Common Wormwood (*Artemisia vulgaris*)
- Pomegranate Root Bark (*Punica granatum L.*)
- Southernwood (*Artemisia abrotanum*)
- Sweet Wormwood (*Artemisia annua*)
- Sweetflag/Calamus Root (*Acorus calamus*)
- Tansy (*Tanacetum vulgare*)
- Walnut shell (*Juglans*)
- White Fraxinella (*Dictamnus albus*)
- Yarrow (*Achillea millefolium*)

Preventative Food & Diet

There are groups of foods that should be avoided if you have a parasitic infection. For example, dairy products in general, refined sugars (sucrose, fructose, corn syrup), flour (especially refined), and overly sweet foods in general. The list of foods and plants below promote good internal balance of the body, hence becoming our allies. With good production of stomach acid, a normal level of healthy bacteria, and proper bile production, it is impossible for parasites to survive for long. Worms need an acidic environment that comes from the breakdown of sugars and putrefaction from the ingestion of unhealthy or processed foods. It is very important to eat raw vegetables and fruit juices, which provide us with enzymes and other elements necessary to protect us.

Choucroute/sauerkraut (fermented cabbage in salt).

Many people have low levels of stomach acid, which is the cause of many intestinal problems, because the body is unable to defend itself against intruders. Sauerkraut juice or cabbage/sauerkraut is one of the most powerful stimulants for your body to produce stomach acid. The use of unpasteurized fermented foods (water kefir, soy sauce, miso, etc.) is highly recommended for its stimulation of the beneficial bacterial flora that is responsible for generating control over parasites. Take a few spoonfuls of cabbage juice before meals, or better yet sauerkraut juice, because it will do wonders to improve your digestion.

Author's note: I do not recommend soy, in any form, for anyone with an ASD.

Garlic

Garlic, eaten regularly, turns the stomach and intestine into a lethal environment for parasites, providing constant protection. Garlic is the quintessential home remedy to eliminate intestinal parasites naturally. It has been used by many different cultures such as Chinese, Greek, Roman, Indian, and Babylonian.

Garlic is still in use today by practitioners of modern medicine. It is used both fresh and as an oil. The simplest treatment is to eat three cloves of garlic every morning, or take a teaspoon of garlic oil. Alternatively, mix crushed garlic in a little cold water and drink the mixture immediately. Another recipe is to cut and crush four cloves of garlic, place them in milk, and allow the mixture to sit overnight. Take the liquid while fasting the next day.

Pumpkin Seed

Pumpkin seeds contain a substance called piperazine. It acts by paralyzing the parasites, which allows them to be removed easily.

We can find piperazine commercially in pharmacy drug formulations or naturally, as we said, in the seeds of the pumpkin. This traditional method of deworming has been used around the world since man can remember. There are several effective traditional formulas, below we describe one of them:

Use one cup of peeled and mashed pumpkin seeds (about 80 seeds). Mix them with coconut water and two tablespoons of honey. Take the mixture over three hours on an empty stomach. Do not eat during this three-hour period. At the end of the three hours, take castor oil in order to quickly eliminate the parasites.

Papaya and Papaya Seeds

Papain is a digestive enzyme contained in papaya that is capable of breaking down the outer layer of adult parasites. The milky juice of unripe papaya is a powerful agent for destroying roundworms. The adult dose is one tablespoon of fresh green papaya juice, an equal amount of honey, and three or four tablespoons of hot water. Two hours later, administer a dose of castor oil mixed with warm milk. This treatment should be repeated for two days if necessary. For children seven to ten years of age, half of this dose should be administered. For children under three years of age, one teaspoon (5ml) of the mixture is sufficient.

Papaya seeds are also useful for this purpose as they are rich in papain and caricin. For every tablespoon of crushed, fresh seeds, add an equal amount of honey. Take the dose of one teaspoon (5ml) daily in the morning or at night on an empty stomach for ten days, rest five days and repeat the cycle three times. We recommend the use of a purgative.

Ginger

Ginger not only helps to combat intestinal parasites but also reduces nausea and can help calm nerves. For hundreds of years, fresh ginger has proven to be highly successful in destroying intestinal worms. The most common way to consume ginger is raw or by infusion. Ginger extract may also be sprinkled on a variety of foods.

Propolis

Propolis is a resin like substance gathered by bees from the bark and leaf buds of trees, to help disinfect, build and maintain their hives. Propolis has been used for at least 3,000 years. Its use dates back to the Egyptians and the Romans, and remains in use today. To the Greeks we owe the name pro, meaning "before" and polis, meaning "city." This translates as "defenses before the city," or "defender of the city." Thanks to the antibiotic action of propolis, which protects against the activity of viruses and bacteria, the hive is one of the most sterile places known to nature.

Many scientific studies have proven the antiparasitic activity of propolis, therefore it is recommended for treatment of: *Giardia*, amoebas and roundworms, and also for intestinal infections caused by gram-positive bacteria.

Take propolis, diluted in water or fruit juice, for treatment of parasites, for seven days, on an empty stomach. Use Propolis standardized at 30% in either propolis tincture or capsules. Take three drops per kilo of weight, or three capsules one half-hour before each meal. A seven-day treatment cycle should include seven days on, followed by seven days off; repeat three to five times to ensure complete elimination of parasites or bacteria. Repetition of the treatment is essential to halt bacterial reproductive cycles. By repeating the treatment at least three times, the effective elimination of parasites is ensured. The benefits of propolis are that it has no side effects, is well tolerated, and is highly effective.

Pomegranate Bark

Pomegranate bark contains an alkaloid known as punicine, which is highly toxic to earthworms. It is used by decoction of the root bark, stem, or fruit. The root bark is preferable because it contains a greater quantity of the alkaloid than the bark of the trunk. This alkaloid is also highly toxic to tapeworms. A cold decoction of root bark, preferably fresh, should be given in quantities of 90ml to 180ml three times per day (for adults), with one-hour intervals between cups. A purgative should be taken after the last glass. For children, a dose of 20ml to 60ml is appropriate. A decoction is preferably used to eliminate solitaires (tapeworm, *Taenia Solium*).

Carrots

Carrots are another effective home remedy for eliminating intestinal parasites in children. The chemical constituents of carrots attack pests by preventing their development. It is one of the most effective natural treatments for children, when given a small cup of grated carrots each morning until the problem desists.

Condiments

Seasoning plants are also powerful weapons to keep in mind in our everyday cooking. Since time immemorial, mankind has used them to control parasitic diseases. The following are most interesting because of their effects:

- Cayenne
- Cinnamon
- Cloves
- Paprika
- Pepper
- Tarragon
- Thyme
- Turmeric

Thank you Andreas and Miriam for sharing what I know will be very enlightening, not only to the families of children with autism, but to people all over the world suffering from mysterious symptoms consistent with parasitic infections. I want to share one interesting tidbit about parasite infections before we get into the FAQs on parasites and the parasite protocol.

"We have a tremendous parasite problem right here in the U.S. It is just not being addressed."[5] - Dr. Peter Wina, Chief of Patho-Biology in the Walter Reed Army Institute of Research in 1991. (The problem existed in 1991, and with modern globalization parasites are more prevalent than ever, yet they still not being addressed.

Parasite Protocol FAQs

I showed my family practitioner pictures of the worms we have found. He thinks they are not a parasite, but just mucous. How can I be certain?

It can be difficult to identify parasites, and most general practitioners are not trained to do so. Some parents have success with a local vet, who analyzes samples in their office. As a crude test, you can pour boiling hot water onto your specimen. If it falls apart, it is probably biofilm or mucous. If it withstands the wash, it is probably a parasite. Unfortunately, most stool analyses are notorious for false negatives. If you put hot water on parasites, they do not dissolve. Mucous dissolves in any water—hot or cold.

Is it ok to do the Kalcker Parasite Protocol during pregnancy?

No. Don't do any kind of detox or parasite protocol while you are pregnant or nursing. Any detox will release toxins into the blood stream that could potentially negatively impact the developing fetus or nursing baby. If you are planning on getting pregnant, it would be advisable to do whatever detox or deworming procedure before pregnancy.

When is it appropriate to stop the parasite protocol? In other words, does every child on this protocol need both the CD and the other parts of the Parasite Protocol or are there some that only need CD? This is a critical question since these are two protocols joined together and not every child may need both. We have done the Parasite Protocol and have never passed parasites so I am uncertain about my child.

Usually the Parasite Protocol is repeated for 12-18 months, sometimes less. The most important thing is to make sure that there are no fertilized eggs left that could hatch at a later date. As it is very difficult to know if there are remaining eggs, it is important to complete the 12-18 month treatment. The Parasite Protocol is a complete protocol heavily researched to achieve the best results. The Parasite Protocol in conjunction with CD has been shown to be one of the most effective methods to heal regressive autism. Parasites are cyclical and need to be treated as such. You continue with the parasite protocol until no more parasites are seen—be that in behavior or actually in the stool.

What does the process of worms toxifying the brain look like?

Ascaris lumbricoides (roundworm), for example, produces at least five different toxins: Malondialdehyde, ammonia, histamine, formaldehyde, and morphine. Malondialdehyde is responsible for oxidative stress, and is mutagenic.[7] Ammonia, which can lead to hyperammonemia, can be responsible for seizures, tremors, flapping, poor coordination, growth retardation, combativeness, lethargy, and other symptoms. Comparing hyperammonemia with the symptoms known as regressive autism yields overwhelming similarities. Formaldehyde has been shown in some laboratory studies to affect the lymphatic and hematopoietic systems. Morphine inhibits nerve reactions and slows intestinal peristalsis. It also keeps the immune system from finding parasites, and from doing anything about them. This is a reason we often can't identify parasites. We only find parasites when the infection is acute, not chronic. This is because IgE or IgM reactions are altered by morphine. Histamine can lead to chronic inflammation in the body.

As a family, we expect to be doing the CD and parasite protocols for some time. How do we know when to stop? Should we wait until we have a few months of no symptoms?

We do it until full recovery is reached.

We have been doing full oral dose CD and the Parasite Protocol. We are seeing some amazing changes, but we don't see worms. Are we doing something wrong?

CD affects parasites of different sizes. You may be clearing parasites that would be undetectable to the naked eye, or nematodes that are very small and therefore hard to detect in the stool.

Why is my child so deficient in vitamins?

As a general rule, most children with autism are deficient in vitamins. In the first place many pathogens feed off of vitamins intended for the host, and helminthes especially love B12 and iron. Since many of us, especially those of us living in cold climates, do not receive enough sun, we are deficient in vitamin D. Calcium is used by the body as an antagonist for acid inflammation; all acidity in the body is compensated by calcium, and therefore it is usually low with the kids on the spectrum. As our children heal, these deficiencies fade and homeostasis returns.

I just gave mebendazole, Not pyrantel this AM. What should I do?

Don’t stress. Tomorrow is a new day, and you can begin again. Give mebendazole tomorrow as that is tomorrow’s scheduled dose.

Aren’t there some parasites that are good for us and help our immunity and gut healing?

A parasite is defined as a living being that is dependent on a host for survival, to the detriment of that host.

For those of you that continue the parasite protocol thru the new moon, how many days do you stay on it? We are now 3 days past the new moon.

Some people treat month long. The Kalcker protocol itself is 18 days long. Each family has to find a protocol that suits them. Some folks use mebendazole 3 days before the new moon, the day of, and three days after. Other families use herbal remedies on the off days. If you consult with a healthcare provider, this would be a question to ask them. Each child is unique and has different needs when it comes to treating parasites.

Why are live worms not digested? Poisons?

Live helminths are protected by a glucosoid mucous with positively charged ions, making them resistant to stomach acid, or digestive fluids. When they die, the mucous separates from the helminth, leaving them open to digestive enzymes. We often see the mucous, as well as semi-digested helminths, in the stool of people using the protocol.

Does everyone in the family need to do the parasite protocol at the same time? With or without CD, enemas, diatomaceous earth, Rompepiedras/stone breakers, and castor oil? My (neurotypical) daughter will take the anti-parasitic drugs, mebendazole, and pyrantel pamoate, but not these other parasite protocol ingredients, which is why I ask.

Your entire family needs to do the protocol at the same time, or you risk reinfection. Neurotypical family members should do as much of the protocol as possible. If someone will only take the meds, so be it. However, I feel the more the better.

Do I need to separate the neem from CD?

Yes, I give neem with food, and would separate from CD by at least an hour.

If I don’t see any worms, does that mean my child doesn’t have parasites, or are we just missing the eggs?

Just because you don’t see parasites doesn’t necessarily mean you don’t have any. For example, *Toxocara canis* or *Toxocara cati*, which are very common in our pets (and can infect humans as *Toxocariasis*), are not expelled in fecal matter. Some families on the protocol did not see actual worms until month seven.

To what extent do the rest of us need to do the CD protocol? Is it the same as our ASD children? And, is it necessary to do the whole nine yards with the full Parasite Protocol for the whole 12 months? Are there any shortcuts for healthier individuals?

All of the family members, including pets, need to be on anti-parasite treatment for a year and thereafter lifetime maintenance is best. As mentioned before, with neurotypical family members, you do as much as you can, but the anti-parasitic medications are crucial to preventing reinfection.

Is diatomaceous earth a binder and therefore must it be given apart from food, medication, and supplements?

No. It is not a binder. It is fine to give with or without food, and does not affect medication or supplements.

Do you have to do the Protocol forever if you are fully recovered or is there a maintenance plan?

When the child gets to where you want him, you can start to pull things. Then, we do maintenance dosing of CD; one dose on Monday and one on Thursday. We do the Parasite Protocol for a week every three months since we live in a world of parasites. It is good preventative medicine for all. Ocean water is also good for all of us. See Chapter 13 (page 319) for the complete maintenance plan.

Does CD kill parasites?

Yes. CD kills amoebas, *Giardia lamblia*, and other smaller parasites. CD does not kill the larger macroparasites due to their higher oxidative stress resistance.

When is it ok to start the parasite protocol prior to reaching full dose of CD?

Healing autism is a marathon not a sprint. The point is to heal with the least amount of aggravation. Therefore, you should be at a full dose of CD before starting, and have added in enemas and baths. If you are seeing Herxheimer reactions then don't start yet. Obviously, everyone must make his or her own decisions.

Where in the gut do helminths live? Do CD enemas (300 ml) reach them?

It depends on the type of helminth. There are more than 300 different helminths so they could be in many different places. However, most live in the small intestine. Some lay their eggs in the rectum. The size of the person will dictate how high an enema will reach.

Do you tell your kids that they have worms? Do you show them? Are they scared? My son is very interested about everything in the world and asks 4,000 questions a day. What do I tell him?

How to handle this situation is unique to each and every family. There is no right or wrong answer here. The following are some suggestions we have received from families on the protocol:

Nope. My husband was just telling me a story from when he was a kid about another kid who got teased all the time about having worms. He still remembers it. So, no I will not tell my kids.

My kids (ages 10 & 7) know about worms because we have talked about their Lyme disease and other symptoms for a long time. I think it helps them understand why they feel grumpy or sick and why the enemas help. But, I think every child is different. The risk is that they might tell their friends. I tell them that it is private information and that pretty much everyone has worms but some people have more and get sick from them. And, I told them that even most doctors don't know this but slowly people are learning and getting smarter about this and a good diet, etc... My boys like to see the big ones, and try for new records. They like to show their brother the records. I think that's a boy thing.

I have a 10 year-old, and I tell him that we have parasites. I have not showed him anything, I ask him to leave the room before I start to sort the poo. ;)

My son had another bowel movement FULL of worms, pinworms—and I mean hundreds. This is only our second day on CD (1/2 drop in 8 fl. oz.) so a bit thrown to the deep end. Has anyone else had a start like this? How did everything go from there on?

I have seen it a handful of times. I get VERY excited about this. Everyone has autism a bit differently. But, if we see worms early, in many cases the recovery begins early too. That may be the case for your child.

I understand that even organic produce needs to be washed properly. What is the best way to wash my fruits and vegetables?

The Ascaris eggs are resistant to UV, and can withstand a pH from 2 to 11.5. Eggs are killed by heating to approximately 60°C (140°F). Spray your fruits and veggies with CD spray. 10 drops per 1 fl. oz., of water, no need to rinse off.

Anyone notice a huge growth spurt after starting the parasite protocol?

Many parents have reported their children resumed growth as well as weight gain. No real surprise considering they are getting back what the parasites were taking from them.

Do we need to treat our pets if they are taking heartworm medication?

Yes, heartworm medication does not address intestinal parasites, which can infect us and our children. We need to deparasitize our pets, as well as ourselves.

We put "cascara sagrada" in place of castor oil because for my son and me, it is difficult to take the oil. Is it ok?

Cascara sagrada is not the same thing as castor oil and is not part of the protocol. An alternative solution to liquid castor oil is to buy castor oil in softgels. You will need many but it does solve the taste problem.

My son is very high in Trichinella spiralis, which is a roundworm from undercooked meat or pork. It's supposed to be rare in the US, so I'm wondering how my child would have picked this up. Any ideas? No undercooked meat here for the kids, and we don't eat pork.

At this moment we have not seen a relationship between *Trichinella spiralis* and autism. It may be a multi-parasitic infection, or a misdiagnosis. Apart from the standard protocol, your child might need something else. I would consult the practitioner for the appropriate dosage, and only if your child has a definite Trichinella spiralis diagnosis.

How do we know if there are parasites in the brain, and what can we do about it?

Parasites in the brain are very rare. (According to the CDC, *Cysticercosis* is a parasitic tissue infection caused by larval cysts of the pork *tapeworm.* These larval cysts infect brain, muscle, or other tissue, and are a major cause of adult onset seizures in most low-income countries. An individual acquires cysticercosis from ingesting eggs excreted by a person who has an intestinal tapeworm.)[6] Larvae can be seen on a scan. Many people assume that the problem causing behavioral issues, or mental issues, must be located in the brain. However, the problem likely exists elsewhere. If the blood contains toxins, this will affect the brain. For example, if you ingest too much alcohol, your brain, nervous system, etc. will be affected, but the problem is not in the brain itself. We can think the same way about parasites. The chemicals that they produce will have effects on the brain, however, rarely will the parasites be located there.

Do parasites cause autism? If so, why doesn't every child with parasites have autism?

Regressive autism has been called parasitological vaccinosis by Dr. Andreas Kalcker. This is a cross reaction between a child with parasites that receives certain vaccines. Further research is needed for a definitive answer.

Why don't normal lab tests find parasites?

Generally speaking, lab tests need to find living creatures or eggs. The eggs are only present certain days of the month, and even on those days they won't necessarily be present in a particular stool sample. Living worms are extremely rare in stool because they are generally passed only once they die. They are very good at avoiding being excreted in stool. Furthermore, if they die internally, they can be partially or fully digested within our bodies before being expelled.

Are some labs better than others? Is it worth taking a stool sample to a vet if they send it out, or only if they look under a microscope?

We haven't found a lab that consistently finds parasites. Metametrix™ has identified parasites, but in our experience and the experience of the families I have helped, we have not found a lab that is consistent. It is only worth sending a sample to a vet if they are identifying parasites with a microscope.

Combantrin® is available in tablet and liquid form. Which is better?

I hate the liquid and never use it. It may have added colors and flavors. If you absolutely can't find it without additives, I would use only mebendazole instead of using pyrantel with additives.

Why are there no systemic drugs in the protocol?

They are not needed with this Protocol.Treating parasites without systemic drugs is much safer and easier on the body. The suggested treatments using mebendazole and pyrantel pamoato are nearly unabsorbed by the body, meaning that we do not add more toxins to an already overloaded body.

When does a child need systemic drugs?

This is necessary only for certain parasites like cysticercosis, which is caused by tapeworms, hookworms, Trichinella spiralis (from pork), or other hard to kill helminths. These may be identified by a blood test. The practitioner may prescribe systemics depending on the situation.

Should I start on the new moon if my child's behavior declines then?

The full 18-day Kalcker Parasite Protocol always starts before the full moon. However, some parents have found that by treating parasites for 3 days over new moon that they are able to get through the new moon as well as the full moon with limited issues if any.

Is there a time in the moon calendar, where it is normal that there are no worms in the stool? (We don't have any at new moon.)

Further research is necessary to determine this definitively.

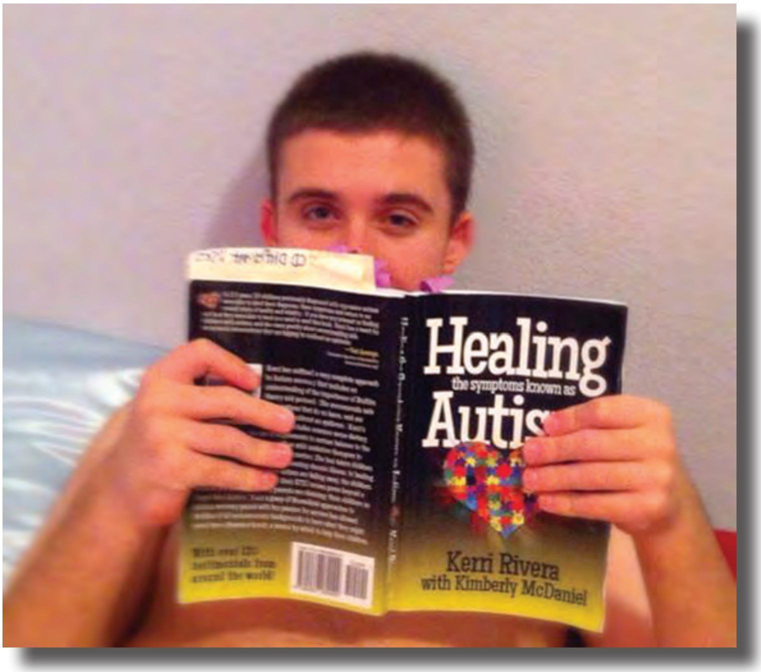
Healing
the symptoms known as
Autis
Kerri Rivera
with Kimberly McDaniel

Hero Guy taking a peek at *The Protocol.*

Chapter 9

Step 4 - Other Supplements

"One of the first duties of the physician is to educate the masses not to take medicine."
~ William Osler

Most of the families that followed this Protocol, and recovered their children, used a combination of these additional supplements along with CD. Each supplement combination is completely unique, depending on the symptoms of the child, and how the child reacts to each intervention. If after *The Diet*, full *CD Protocol*, and three *Kalcker Parasite Protocols* we are still dealing with an autism diagnosis, then families start adding in supplements. This section is about "unfinished business." I like to do three Kalcker Parasite Protocols before adding back in supplements because parasites love supplements, especially vitamin B_{12} and iron.

I feel when we supplement deficiencies in children with underlying parasitic infections, it can result in resistant parasites. We aren't making the child healthier; we're making the parasites healthier and stronger. If significant supplementation does little or nothing to change lab values in subsequent blood, urine, and stool tests, then it's fair to say that we have a flaw in our approach. Pathogens and parasites can cause vitamin/mineral deficiencies. By killing them, instead of feeding them with additional supplementation, we are focusing on the root cause of the symptoms that we call autism.

When I came back from our first *Defeat Autism Now!* doctor visit, I had thousands of dollars in supplements that I did not know how to use. We kept going to different doctors hoping to find the one that had the answer. At one point, Patrick was on a protocol, given to me and supervised by a doctor, which called for 70+ different supplements a day. He looked worse during the entire nine-month period that we followed that protocol. In hindsight, all those years of supplementation made it harder to eliminate Patrick's pathogens.

That is why this protocol focuses on the excesses instead of the deficiencies for children with an autism diagnosis. If by eliminating the excesses of viruses, bacteria, candida, parasites and heavy metals, we can see a reduction in the symptoms known as autism, then eventually we can reach full healing.

Commonly, when a new family contacts me, they send me a list of all the supplements their child is taking. I frequently see lists of 30+ supplements that may be combined in ways that negate their effects. For example, probiotics are given in the morning with breakfast, or GABA with food. Both are incorrect, probiotics go alone at night right before bed, and GABA is always taken without food. If there is a legitimate reason for a certain supplement, then it absolutely must be given correctly or it's just excess noise in the body, and more stress on the detox organs.

The following list of supplements can be considered after *The Diet*, ocean water, CD, and three months of the *Kalcker Parasite Protocol* (except probiotics and enzymes which may start day one). This chapter is to be used as a starting point for your own research on these supplements. This list contains references and the websites used in this research with some of the more pertinent information about products.

Probiotics

The vast majority of the families who have recovered their children with this Protocol have used THERALAC® brand probiotic. It is a prebiotic plus probiotic in a capsule that can withstand stomach acid, which allows it to pass to the small intestine, where the beneficial bacteria are needed. THERALAC® contains three *lactobacillus* and two *bifidobacterium* strains.

The paragraph below was slightly modified from www.theralac.com where you can find more information. Please visit their site for details on references contained within the text.[1] The following two paragraphs were taken from different sources, and help explain benefits of THERALAC®, specifically for children on the spectrum.

> Probiotics are defined as: "Live microorganisms that when administered in adequate amounts confer a health benefit to the host" (World Health Organization 2001). According to the International Probiotics Association (IPA), the benefits of probiotics can include reduction in diarrhea caused by antibiotics and rotavirus, alleviation of symptoms of lactose intolerance, alleviation of symptoms of food and skin allergies in children, reduction of recurrent ear and bladder infections, and other positive indications.

> ### How Does THERALAC® Benefit My Immune System?
>
> THERALAC®'s five probiotic strains work in unison to support the mucosal immune system on the intestinal surface; activation signals are then sent to the body's systemic immune system. This is a gentle and controlled process because THERALAC®'s strains are beneficial. Essentially, the immune system is put on alert so that it is ready to act quickly if there is an appearance of pathogenic microorganisms.[2]
>
> Many children with autism have chronic digestive problems. In fact, gastrointestinal symptoms in autistic children often first appear in conjunction with initial changes in emotion and behavior during the onset of autism, leading researchers to suspect a gut-brain connection.[3]

Normally, proteins are digested in stages by enzymes; first to peptides, and then to smaller amino acid components, which are absorbed into blood capillaries in the gut mucosa. The larger peptides are generally unable to cross this mucous membrane barrier, if they do, however, they can act as opioids affecting neurotransmitters in the brain causing abnormal behaviors and/or activity. These incompletely digested peptides—known as exorphins, casomorphins, and gluteomorphins usually come from milk proteins such as casein, or from wheat (gluten), and are structurally similar to morphine. The formation of excess peptides in the gut is possibly associated with sub-optimal enzyme activity, or an insufficient supply of enzymes required to breakdown these peptides. So if we repair the imbalance of beneficial bacterial organisms in the gut and the gut lining, while killing the pathogens causing the dysbiosis, we have the opportunity to heal children with autism. Probiotics can be used to improve the quality of the gut mucosa.[2]

In addition to aiding in the repair of the gut lining and improving digestion, there is also evidence that probiotics can help with detoxification of heavy metals such as toxic mercury.

The paragraph below was taken from the presentation "Gut biology and Treatment" by Dr. Anju Usman:

> Emerging literature is showing the beneficial effect of oral probiotics on mood and anxiety symptoms. In a double-blind, placebo controlled randomized parallel group study, daily use of probiotics reduced psychological distress.[5] A number of studies have shown the anti-anxiety effects of probiotic use in patients with medical conditions.[6]

> Dr. Vincent Young, University of Michigan (2009): "the gut ecosystem needs to be preserved and that changing the ecosystem through stresses such as antibiotics could irreversibly change the ecosystem, with deleterious effects." Dr. Young has studied the effects of antibiotics on the microbes in our gut. He found that mice when given particularly strong antibiotics completely wiped out all their normal gut microbes. Even more striking, clostridium species and fungal species are then able to overgrow without the bacteria there to fend them off.

There is no doubt in my mind that probiotics helped get my son's gut back in order, helped improve my own health, and that of thousands of families of children on the spectrum. Probiotics should be given without food, directly before bedtime. This way, they have the entire night to proliferate in the small intestine where they are needed.

Special note for those with PANDAS/PANS: We have received some reports where some PANDAS/PANS and/or older children were becoming violent and having more SIBs while on probiotics. Be observant and watch for behavior changes when adding probiotics. In addition, know that we may need to remove them for a period of time!

Omega-3 & Omega-6 Fatty Acids

Omega-3 and omega-6 fatty acids are polyunsaturated fatty acids and are considered essential fatty acids. Essential fatty acids must be consumed because your body cannot produce them. They are important for brain development, immune system and cardiovascular function, and normal metabolism. Omega-3 and omega-6 fatty acids are commonly found in marine and plant oils. While a healthy diet often provides adequate supplies of both, supplementation is sometimes necessary.

Omega-3 fatty acids also help to regulate energy levels, as well as normalizes blood sugar levels. They help improve concentration as well as mental vividness. Using omega-3 fatty acid supplements helps with anxiety, mood swings, and depression. They are especially significant in the treatment of autism. The entire body benefits from ingesting omega-3 fatty acids; from physical development, to relaxation from anxiety and jitteriness, to developing brain cells; all of these aspects can aide in the treatment of autism.[6]

Omega-6 fatty acids promote healthy brain function and assists with skin and hair growth, bone development, and metabolism. A healthy balance of omega-3 and omega-6 fatty acids promotes heart health and minimizes inflammation. Omega-6 may also be used to treat allergies, eczema, osteoporosis and premenstrual syndrome.[7]

Unlike omega-3 fatty acids, excessive consumption of omega-6 fatty acids can have negative effects. Therefore, their consumption must be monitored. Omega-6 fatty acids, in large quantities, can promote inflammation in the body that may lead to flare-ups of eczema, acne, and the aches and pains associated with arthritis. It is recommended that you maintain about a 4:1 ratio of omega-6 to omega-3 fatty acids.

Some research has been done that suggests improvements in overall health, cognition, sleep patterns, social interactions, and eye contact when children on the spectrum were given an EFA (Essential Fatty Acid) supplement.[8]

Furthermore, another study showed the positive effects of EFA's on dyspraxia (a motor disorder frequently associated with cognitive, behavioral, and social challenges). The participants showed improvements in reading, spelling and behavior during the treatment period.[9]

I have personally found that I really like omegas from YES™ Supplements. Dr. Brian Peskin has some very informative videos on YouTube discussing the merits of vegan versus animal based omega sources. I highly recommend you have a look for yourself. My own son, and many other children have built up to taking one tablespoon, two times a day.

L-carnitine

L-carnitine is a vital amino acid that promotes healthy neural levels of acetylcholine, an important neurotransmitter that aids memory and proper brain function. Research suggests that an L-carnitine deficiency may be implicated in a number of conditions, including ME/CFS (chronic fatigue syndrome), diabetes, Alzheimer's, dementia, and autism.

L-carnitine is composed of two essential amino acids, lysine and methionine. It is produced in the liver and kidneys and is contained in most cells of the body. It is required for the proper metabolism of fat and helps with mental concentration, and energy production. It transports long-chain fatty acids across the mitochondrial membrane so that they can be burned to produce energy.

A clinical trial studying the use of L-carnitine as a potential therapy for autism demonstrated the children taking L-carnitine showed improvements in their ability to relate to people, body use, adaptation to change, listening response, verbal communication, sociability, sensory/cognitive awareness, and health/physical behavior. There are a few articles that may be helpful for a better understanding of the importance of this amino acid. Check out an article by Henke Schultz, "L-carnitine helps kids with autism, study finds" and another by Emily Singer, "Defects in carnitine metabolism may underlie autism."

The average dose for L-carnitine is anywhere from 250-1000 mg per day with food, depending on weight and reactions.

GABA

GABA (gamma-aminobutyric acid) is an amino acid and a neurotransmitter (a type of chemical responsible for carrying information from one cell to another). It is produced naturally in the body, but is also widely available in supplement form. Manufacturers claim that GABA supplements can help boost the brain's GABA levels and, in turn, treat anxiety, stress, depression, and sleep problems.

The GABA system acts as something of an information filter to prevent the nerves from becoming over stimulated. It has long been suspected that this filtering process is compromised in many autistic children. Impairment of the GABA system could overwhelm the brain with sensory information, leading to many of the behavior traits associated with autism. GABA is also believed to play a key role in the early development of the brain.[10]

GABA is also involved with the production of endorphins in our brain, which make us feel positive and upbeat. GABA can reduce stress, relieve anxiety, and increase alertness. GABA can be helpful for behavior, language, and possibly even seizures.

GABA is used for control of seizure activity and works on the same receptors as the drug Keppra®. GABA is also useful for speech, helps facilitate language in nonverbal children, and improves language in children that are beginning to speak.

If you are going to give GABA to your child, the max is 5,000 mg per day, divided into two doses, one in morning and one at night, always on an empty stomach. In the evening, GABA can be given 15 minutes after your last dose of Chlorine Dioxide, with your probiotic. I start at 250 mg two times a day, one in the morning and one at night. I then titrate up the dose daily or every other day if everything looks good. If you observe sleepiness or irritability it is a good idea to back down.

5-HTP (5-Hydroxytryptophan)

5-HTP, also known as oxitriptan, is a naturally occurring amino acid. It is a chemical precursor to the neurotransmitters, serotonin and melatonin from tryptophan. 5-HTP works in the brain and central nervous system by increasing the production of the chemical serotonin. Serotonin can affect sleep, appetite, temperature, sexual behavior, and pain sensation.

From *Saving Eli: One Family's Struggle* - Vitamin Research Products website:

> While not clearly understood, researchers know that serotonin pathways are disturbed in autism, contributing to sleep disorders and mood. Tryptophan has been shown to help but has been banned by the FDA since 1989. We found that 5-HTP helped immensely, calming tantrums and increasing communication with our son.[11]

5-HTP is used between 50 mg and 200 mg, divided into two doses, morning and evening with food. I have seen it help children with autism for attention, focus, sleep, and cravings.

L-theanine

L-theanine or gamma-glutamylethylamide or 5-N-ethyl-glutamine, is an amino acid commonly found in tea. Theanine is able to cross the blood–brain barrier, and is reported to have positive effects on mood, stress, and cognition.

The following is excerpted from the L-Theanine website:[12]

> Many people experience stress on a daily basis and are looking for natural and safe ways to manage it. For thousands of years, it has been suggested that drinking green tea will make one relaxed. Recently this relaxation effect was found to be true and works because of an amino acid called L-Theanine which is found in the tea. Clinical research will show that ingesting up to 200mg of L-Theanine will promote the creation of the very important "anti-stress" neurotransmitter in the brain called GABA or gamma-aminobutyric acid. This results in a relaxed, clear and alert mental state.
>
> When used in the correct amounts in supplemental form, L-Theanine may:

- Reduce Stress
- Reduce Occasional, Simple Nervous Tension
- Promote Relaxation without Drowsiness
- Promote Mental Clarity and Focus
- Promote Positive Mood
- Promote Alertness
- Promote Learning and Memory
- Help Prevent Jitters Caused By Caffeine

I have found that when we can reduce these symptoms in individuals with ASDs we can get an increase in speech, learning, focus, concentration, and attention. L-Theanine is given with GABA upon waking, and can also be given at bedtime if needed, always without food. You can work up to 200 or 250mg/day.

Pycnogenol®

Pycnogenol® (pic-noj-en-all) French maritime pine bark extract acts as a potent blend of antioxidants, it is a natural anti-inflammatory, stimulates generation of collagen and hyaluronic acid and help with natural dilation of blood vessels by supporting production of nitric oxide.[13]

Pycnogenol® is not a supplement I tend to use very often, because it is an antioxidant, and therefore cannot be combined with chlorine dioxide. However, in some people it has proved necessary and helpful for speech and reduction of seizures. It should be given in the morning with food at a dose between 25 and 400 mg/day depending on need. Pycnogenol® may not be suitable for children who are sensitive to phenols.

L-Carnosine

L-Carnosine is a dipeptide of the amino acids beta-alanine and histidine. It is highly concentrated in muscle and brain tissues.

L-Carnosine is believed to stimulate the frontal areas of the brain, resulting in overall improved levels of functioning. More and more research shows that the frontal lobes and temporal lobes in the brain control emotion, epileptic activity, cognitive, expressive speech, and abstract thinking. Studies have shown Carnosine to improve language, socialization, and overall level of functioning in individuals within the autism spectrum. It has also been shown in studies to have anti-seizure properties without the side-effects of prescription anti-seizure medications.

Research indicates it prevents the formation of the beta amyloid plaque that is found not only in neurological conditions such as Alzheimer's, Parkinson's and autism, but also in the eyes in degenerative eye conditions and in the pancreas in diabetes. It is a neuroprotectant, with a study indicating it reduced severity of damage in stroke patients.[14]

Chez and coworkers found that after eight weeks on L-carnosine, children showed statistically significant improvements on the Gilliam Autism Rating Scale (GARS). They relate this to the likely ability of L-Carnosine to enhance neurological function, perhaps, in the enterorhinal or temporal cortex.[15] This enhancement in neurological function has led to speech in many of the families that I have worked with.

I have also seen L-Carnosine help reduce seizures in some children and increase language in others. It is given twice a day with breakfast and dinner at a dose of 200 to 400mg, 2 times a day. If you note hyperactivity, remove it.

Taurine

Taurine is an amino acid, a chemical that is a required building block of proteins. Taurine is found in large amounts in the brain, retina, heart, and in blood cells called platelets. It can be consumed through eating meat and fish.

Autism and Low Taurine

"Are You Dangerously Deficient in Taurine"[16] explores possible problems associated with a taurine deficiency. Among the problems noted in the article is autism and low taurine levels. In the article, Leonard Smith MD writes about the benefits of taurine, some of which can be of interest to people dealing with autism spectrum disorders:

- Brain and nervous system function
- Helps eliminate toxins
- Stabilizing the brain (can be effective in treating seizure disorders)

Taurine, like all amino acids, must be given without food. The maximum dose is between 500 and 1,500 mg.

DMG

DMG, or dimethylglycine, is an amino acid that can be found in many common food items like meats (especially liver), various grains, and beans. It has been classified as a type of amino acid that is closely linked to vitamin B.

There have also been interesting developments in the use of a DMG supplement for children with autism. Because many children who have autism are unable to tolerate eye contact, and some have problems forming complete sentences and thoughts studies investigated whether an increase in DMG consumption could help to alleviate many of these problems. In fact, the study found that when children with autism were given supplemental DMG they appeared less frustrated and showed a marked increase in their speaking and cognitive abilities.[17] At present, this research remains novel and therefore further investigation is necessary to determine what the long-term effects of DMG supplementation might have on children on the spectrum.

The following paragraphs were excerpted from *Defeat Autism Now!*[18]

> ### Dimethylglycine (DMG) for Autism
>
> For over 20 years ARI has been hearing from parents who have tried DMG on their autistic children. In many cases remarkably good results have been seen, especially in enhancing speech. In some cases, drug-resistant seizures have been stopped by DMG. (See New England Journal of Medicine, 10-21-82, pgs 1081-82).
>
> There is an extensive research literature on the safety and health benefits of DMG. Many studies have shown that DMG enhances the effectiveness of the immune system, improves the physical and athletic performance of humans and other animals (e.g. race horses) and has, all in all, a very wide range of beneficial effects. It is very safe. I have seen no evidence of any toxic or significant adverse effects.
>
> Many parents have reported that, within a few days of starting DMG, the child's behavior improved noticeably, better eye contact was seen, frustration tolerance increased, the child's speech improved, or more interest and ability in speaking was observed.

A full dose of DMG is 900 mg per day, taken without food upon waking. However, it is best to start at a lower dose and slowly work up to 900 mg over a week or so. If you see an increase of hyperactivity (which is rare), reduce the dose. If no improvement is seen within a month I would switch to TMG.

TMG - Trimethylglycine (Betaine Anhydrous)

TMG (Betaine anhydrous} is a chemical that occurs naturally in the body, and can also be found in foods such as beets, spinach, cereals, seafood, and wine.

How does it work?

A form of betaine called betaine anhydrous helps in the metabolism of homocysteine, a chemical involved in the normal function of many different parts of the body, including blood, bones, eyes, heart, nerves, and the brain. Betaine anhydrous prevents the buildup of homocysteine seen in people who have problems with its metabolism from birth.[19]

The following is excerpted from the *Autism Canada Foundation*.[20] Please visit their site...

www.autism.org

...for full references contained within the text.

The benefits of taking DMG or TMG range from behavioral changes, reduction of seizures, and decreased obsessive-compulsive behaviours to improved language. DMG and TMG have been reported from thousands of families to be quite beneficial to many individuals with autism.

Research on humans and laboratory animals has shown that DMG and TMG enhance the effectiveness of the immune system. Some children and adults with autism have seizures, and there are published reports of decreases in seizure activity as a result of DMG. A double-blind placebo-controlled study by Drs. Shin-siung Jung, Bernard Rimland, and Stephen M. Edelson involving 84 participants documented a significant decrease in behavioral problems.

It should be noted that some kids tolerate DMG but not TMG. TMG is given upon waking without food, in a dose of 500mg. If DMG didn't yield improvements in language, then we switch to TMG.

FAQ's

There are times when I do see undigested food in my son's stools. I am particularly concerned. He never complains of stomach pains. But I do see some at times, especially cashews. Should I bother with an enzyme? Or is it better without one. If so, which brand?

> I love enzymes. Kirkman has one with Isogest®, 851/180 is the number, and it's broad spectrum. Also, Biofilm Defense® is great for dissolving the biofilm. A number of the parents online have used Ness® enzymes Gastric comfort formula #601 very successfully.

How should I administer THERALAC®?

The best time to administer probiotics is at bedtime. Children's THERALAC® is a granular formula so you can sprinkle it on yogurt or mix it into a smoothie and still obtain the same benefits! So how do you take it? We recommend that you take a level 1/4 teaspoon and fold it into yogurt, applesauce, or food of a similar consistency and let it sit for a minute. This is to keep the granules as close together as possible, thus allowing our acid proof gel matrix to form around the product. Visit the THERALAC® website for more info:

www.theralac.com/childrens-theralac.aspx

Supplement Dosing Overview

Supplement	Dose	Time of Day	Empty Stomach (ES) or With Food (WF)
Probiotics	1 cap	Before Bed	ES
Omega-3/ Omega-6	1 Tbsp	With any meal 1-3x / day	WF
L-Carnitine	250-1000mg/day	With any meal 1-3x / day	WF
GABA	Up to 2,500mg 2x/day	Upon waking & at bedtime	ES
5-HTP	50-200mg	Morning & Evening	WF
L-Theanine	Work up to 200-250mg/day	Mornings or Mornings & Nights	ES
Pycnogenol	25-40mg a day/as needed	Morning	WF
L-Carnosine	200-400mg 2x / day	Morning & Evening	WF
Taurine	500-1500mg/day	Morning, Noon & Night	ES
DMG	900mg/day	Morning	ES
TMG	500mg/day	Morning	ES
Enzymes	1 cap w/meals	Morning, Noon & Night	WF

Chapter 10

Step 5 - Chelation

So, let's say we're clipping along; full Diet, CD, Kalcker Parasite Protocol, and we've added in whatever supplements the child specifically needs and the child still has autism. Recovery is still not a reality... yet! At this point, it's time to look at chelation.

Why use Chelation for Autism?

Chelation became very popular in the world of autism about a decade ago, when things were heating up around the Thimerosal/Autism connection. Today, many of our children are still metal toxic. Metal challenges (testing that shows metals in urine) from any of the thousands of children whose families I have helped show the same thing—extremely high levels of mercury, lead, aluminum, as well as sometimes tin, cadmium, and other metals.

These heavy metals can come from various sources such as:

- Coal burning power plants. According to the EPA, coal-fired power plants in the United States emit about 48 tons of mercury into the air every year, where more than half of this mercury falls within five miles of the plant itself. When it reaches the water, microorganisms consume it and convert it into a substance called methyl mercury.
- Drinking water
- Our food supply
- Cookware
- Deodorant
- Beauty products
- Dental amalgams
- Vaccines, etc.

Heavy metals are known to accumulate in different parts of the body including organs, bones, joints, and the brain, etc. Metal toxicity can provoke inflammation, kill neurons, cause behavioral changes, affect the thyroid and other master glands, lower T-cell counts, and cause a myriad of other symptoms. Furthermore, new research shows that fluoride in drinking water makes the aluminum that we ingest more bio-available. As was reported in the journal *Brain Research*, the combination of aluminum and fluoride causes the same pathological changes in brain tissue found in Alzheimer's patients.[1]

Many children with autism also have impaired methylation cycles. The result of this impairment is an inability of the body to rid itself of excess metals, thereby prolonging chronic illness.

What exactly is methylation?

Methylation reactions are those that involve the transfer of a methyl group from one compound to another. The methylation cycle is the name given to a biochemical pathway that contributes to a range of crucial bodily functions, including:

- Detoxification
- Immune function
- Maintaining DNA
- Energy production
- Mood balancing
- Controlling inflammation

Impairments or mutations on the methylation cycle can lead to problems with:

- ASD's
- Alzheimer's
- Diabetes
- Allergies and Asthma

An overload of toxins (including heavy metals), can contribute to the impairment of the methylation cycle, and if the methylation cycle is impaired the body is unable to detoxify as needed, therefore creating a vicious catch 22.

We need to help our children's bodies rid themselves of their excess metals as part of the healing process. Dr. Usman has shown us that heavy metals are present in the biofilm, and Dr. Klinghardt has shown that removing mercury can be directly related to a reduction in chronic infections.[2]

Let's get those metals out!

What exactly is Chelation?

Chelation is a process used to rid the body of heavy metals. It is described as a chemical process in which a substance (chelator) is used to bind molecules, such as metals or minerals, and hold them so that they can be removed from the body.

Some of the most common chelators used in the world of autism are:

EDTA

This is an amino acid that attracts lead, other heavy metals, and some minerals from the bloodstream and expels these toxic elements in the urine. EDTA works to remove excess lead from the body, but it is not specific to mercury or methyl mercury as are DMSA or DMPS. It can be taken orally, by rectal suppository, or IV.

DMSA

This is an FDA approved drug that can be used in children when lead toxicity is suspected, however it can also be effective at removing other heavy metals including mercury and arsenic. It can be taken orally, transdermally, or given as a suppository.

DMPS

DMPS is given to remove mercury from the body. It can be given IV, intramuscularly, subcutaneously, transdermally, or by suppository.

Patrick has been prescribed all of these at different points of his life, and I can't say that I saw miraculous results during the time that I used them. However, some families have seen results, and this is something you may want to discuss with your doctor. After years of using chelators, and seeing other families use chelators, I have opted for a gentle approach. I like to use two products: bentonite clay baths and Bio-Chelat™. These products are strong enough to help the body rid itself of heavy metals, but have not been shown to stress the liver or provoke undesirable detox symptoms.

Knowing now that metals are in the biofilm, I feel it is short sighted to focus solely on metals or heavy metals rather than all the pathogens in the biofilm.

Bentonite Clay Baths

Bentonite clay is sedimentary clay composed of weathered and aged volcanic ash. Bentonites are more widely known as healing clays used for detoxing, cleansing, and drawing out impurities. They are used in many everyday products such as toothpaste, antacids, and cosmetics.

Indigenous people have used bentonite clay for centuries; Dr. Weston A. Price in his book, "Nutrition and Physical Degeneration,"[3] stated that when studying the diets of native tribes he examined their knapsacks. Among the tribes examined in the High Andes, in Central Africa and the Aborigines of Australia, Dr. Price reported that some knapsacks contained balls of volcanic ash clay, a little of which was dissolved in water.

Bentonite is known as "swelling clay." When bentonite clay absorbs water and swells up, it is stretched like a sponge. Toxins are drawn into these spaces through electrical attraction and bound. In fact, according to the Canadian Journal of Microbiology, bentonite clay can reportedly absorb pathogenic viruses, as well as herbicides and pesticides.[5]

One of my personal favorites has been *Even Better Now*®'s product available at:

www.evenbetternow.com

> "EBN® Cleansing Clay is 100% pure sodium bentonite clay which has the highest cation exchange capacity (CEC of 98-107 meq/100g) of any bathing clay that we tested on the market. This clay is high purity air-classified sodium bentonite, selectively mined, consisting of micronized particles, which is a free-flowing powder. EBN® Cleansing Clay is 100% pure, hypoallergenic, and free of viruses, bacteria, yeast, and mold, as well as having a high cation exchange capacity."

There are other quality clays on the market, but EBN® tests every batch for heavy metals when it comes in. No, I do not receive financial gain from the company.

Bio-Chelat™

Another product I really like for gentle (low and slow) chelation is Bio-Chelat™. This is a German product containing a minimal amount of EDTA, which is FDA approved as a food substance. According to a clinical trial carried out in Germany:

> The therapeutic value of the Bio-Chelat™ in the context of other chelators that are currently on the market is seen as follows: Chelators work relatively fast, but they are also very strong with a relative high washout of important trace elements and a high degree of specific side effects. Bio-Chelat™ works much gentler than most common chelators.

> Although during the treatment a significant decrease of the body's heavy metal ion load was seen, this is accomplished without greatly disturbing the mineral and trace element relationships. The reduction of zinc should be looked at with caution and may easily be corrected throughout the treatment.[5]

I have seen very positive changes in many of the children who added in either bentonite clay baths, Bio-Chelate™, or both. As with any other intervention, add them separately while closely observing your child. Keep a log to record any changes, progress, or reactions you observe. Even if you don't see anything right away, I would keep using the products for a minimum of three months. Again, since they work LOW and SLOW changes may not be apparent at the outset, but that doesn't mean they are not helping to heal your child's body.

I use the CD baths every other day and then two times a week I do bentonite clay baths. I don't do the CD bath and the bentonite clay bath on the same day. I do them at opposite ends of the week.

Three days or so later, we'll add in the Bio-Chelat™. Again, follow the instructions on the package, and you can always start low and work your way up. Since Bio-Chelat™ doesn't alter the CD, the drops can be added to a single dose of CD. In addition, since the drops have no flavor, it is suitable to add them to water or any other drink your child might consume throughout the day.

Common Errors in Chelation:

Urine analysis without provocation:

Many families will do urine analysis to see if there are heavy metals present. If you have not done any chelation (oral or intravenous) you really don't get a good picture of what kinds of metals are actually in the body. Simply because the metals don't show up in urine, doesn't mean they are not present in the bones, the intestinal tract, the brain, and organs, etc. If we don't do any sort of provocation, we may end up with a false negative, leading a family to believe that their child is not metal toxic, when in fact they are. Bentonite clay baths and Bio-Chelat™ are gentle enough that they will not cause any harm to a child (metal toxic or no) but can greatly help reduce the heavy metal load, as well as other toxins in the body.

Healing the gut before starting chelation:

A doctor once told me that we had to heal the gut before we could start the chelation process. However, in the children that I have seen recover, probably up to the last month before they recover, their gut might still not yet be perfect. For example, one of the little girls that recovered continued to suffer from constipation at 30 days before she lost her autism diagnosis. However, at this stage something seemed to switch, she started having regular bowel movements and within 30 days she lost her diagnosis.

Basically we're healing the intestines as we're chelating, as we're healing everything else. They are all parts of the puzzle. I would never put chelation before the CD or parasite protocols. However, it's definitely something that's part of our road when, after these other approaches, children have not yet recovered. At some point after we have the CD, ocean water, parasite protocol and some supplements is when we can add in chelators. It's not like I'm saying, "Oh we'll talk about that in a couple of years." We want our children recovered in less time. It's like a marathon, we keep the pace, never lose our heads, or break into a sprint.

Do not chelate with constipation:

If the child suffers from constipation, it's really important that we start moving the intestines when we're using chelators. There are a few options open to us for moving the intestines. I have found that the most effective is through the use of CD enemas. If your child isn't having a daily bowel movement, I would do an enema once a day. If you absolutely can't do enemas, get that child in a CD bath. We're finding that 50 to 100 drops in a bathtub is allowing children who previously had constipation to have bowel movements.

There are a few children/adults on the protocol that will not do enemas, or the parent will not. Whatever their situation may be, we have to get the bowels moving again. In this case, we do 50 to 100 drops in their bath water along with their max oral dose every day. The result has been that they're having regular bowel movements. That's one way we can move it along and then you can keep chelating. It's really important to have a daily bowel movement with any kind of chelation because otherwise the toxins are simply reabsorbed through the intestinal wall, whether they're heavy metals or other pathogens.

IV chelation is not the only means by which to cure autism:

I think at one point between 2006-2008, we were all focused on the problem in the world of autism being the heavy metals, because there was mercury/Thimerosal in the vaccines. We had to get the mercury out of our children. Well, part of the problem is they're not detoxing. As mentioned before, methylation cycles are impaired in children on the spectrum. I met a young girl whose parents were both allopathic doctors and they really believed at that point, as most of us did, that heavy metals were the problem with autism. If we could get the heavy metals out then they would recover. The parents did 122 IV chelations and their child never improved. At this point I have lost touch with them, but these are things that we have to learn along the way. I know now that it is NOT just about chelation, and why metals are NOT the only important piece. All of the pieces play a role and likely interact to cause autism in our children. There is a symbiotic relationship between the viruses, bacteria, candida, and parasites that exist in the biofilm. Since CD can break down biofilm and kill the pathogens living in it, as well as neutralize heavy metals, it has been a very powerful tool for us.

Doing a blood test to look for heavy metals:

Generally speaking, within three days of heavy metals entering the bloodstream, the body will deposit the heavy metals into tissues, organs, and eventually bones. For this reason, a blood test is not the best place to look for heavy metals in children with autism. Unless the child has very recently been exposed to heavy metals, we are unlikely to see elevated levels of heavy metals through a blood analysis.

The articles at the end of this book (Appendices 13 & 14) will provide further insight as to why chelation can be so important for our children on the spectrum.

Healing
the symptoms known as
Autism
Kerri Rivera

Luca

Chapter 11

Step 6 - Hyperbarics

by Bob Sands (Introduction by Dr. Beatrice Golomb)

"What Oxygen is to the lungs, such is hope to the meaning of life."
~ *Emil Brunner*

We were fortunate enough to have the amazing Bob Sands write this chapter on hyperbarics explaining its importance in healing ASD's.

Note: Throughout this section of the book, the author, Robert Lyne-Sands uses the actual names of the people with whom he came in contact with. In some instances it is with their permission, in others it is not. The accuracy of quotations, opinions expressed within this chapter, and the names used are solely the responsibility of Robert Lyne-Sands. Great care has been taken to remain within the bounds of current journalistic integrity and practice.

Robert Lyne-Sands new book, *OXYGEN: The First Medicine, Volume 1 CANCER*, is expected to be released January 2014.

Robert Lyne-Sands author of the Hyperbaric section of the book, with one of his young patients, getting ready for a session.

Robert Sands is one of only about five licensed manufacturers of hyperbaric oxygen therapy (HBOT) machines in the US, with extensive experience in clinical hyperbaric treatment, whose expertise will be an exceptional asset to this project. He has designed and manufactured HBOT machines for the US, Australia, Britain, Thailand, and many other nations, for naval/military use, civilian diving use, and medical use. He has the theoretical understanding of HBOT therapy that derives from intimate involvement in the process of designing and manufacturing these machines. He has the practical experience that derives from running a set of HBOT centers, and personally overseeing greater than 20,000 hours of HBOT sessions over 35 years.

~ Dr. Beatrice Golomb, MD, PhD.
Professor of Medicine. UCSD 2011.

Beatrice Alexandra Golomb, MD, PhD (PI) is singularly qualified to lead this study. She has expertise in GWI including authorship of four RAND reports related to GWI29-31, and work in the scientific literature. She has been a member, since its inception, of the Department of Veterans Affairs Research Advisory Committee on Gulf War Veterans' Illnesses, and served as Scientific Director of this Committee. As a licensed internist, she is a primary care physician for a panel of ~280 veterans, including GWV, at the VA San Diego Healthcare Center.

Recently, I sought the answer to something that was puzzling me. My friend, Dr. Bernard Gunter, is a revered and renowned psychologist in his eighties and, if anyone would know the answer, Bernard would.

"Does God, Great Spirit in the Sky, The Universe or whatever unseen deity you worship give you freedom for personal choices or is everything predestined?"

Bernard's eyes twinkled. He was never in a hurry to answer a question. In my mind, Dr. Bernard Gunter is sort of the American version of the Dali Lama.

He thought for a moment, selecting his words carefully, and then replied, "Funny you should ask that—just the other day I was talking with Eckhart Tolle and asked him that very same question. Tolle replied, 'God allows you to think that you have free will and free choice—and you do, by and large. However the roadmap is laid out for you beforehand.'"

Just as I suspected.

Dr. Bernard Rimland, Kerri Rivera, and the Hand of God.

On January 20th, 2006, a tall, glamorous woman walked into my San Diego clinic; her name was Kerri Rivera. She had brought her small five-and-a-half-year-old son, Patrick all the way from Mexico for hyperbaric oxygen therapy. The little fellow was absolutely gorgeous, but had a vacant stare and made no eye contact with any adult, rarely even with his mother. Day after day Patrick spent time in the chamber.

Kerri explained to me that Patrick developed normally for almost the first two years of his life. Then, to comply with vaccine regulations, Patrick received the DPT, HepB, and Influenza B (Pentavalente) vaccines in one day. Patrick fell into a complete zombie-like state after the injections. By the time Kerri arrived in San Diego, she had already done her homework. She blamed the vaccinations and had already undertaken the task of healing her son.

After her own experience with autism, Kerri wanted to help people and she was already doing the DAN! Protocol, etc., as Dr. Rimland has done so much for the children and she had a lot of respect and admiration for him and his work.

One day, I walked into the clinic and told her, "We're having lunch with Bernie tomorrow. Get a babysitter." At that point, totally shocked, she said, "Oh my God, yes, but I don't have anything to wear!"

"Trust me, what you have on would be just fine." A flustered Kerri rushed off to purchase some dressy clothes—she was, after all, a woman originally from Chicago. The very next day, we were sitting with Bernie and his wife Gloria in a quaint little restaurant, *The Green Tomato*. If the reader wishes to see that day, just go to your computer and look up Bernie on Wikipedia and there you will see Kerri in her black business suit (http://en.wikipedia.org/wiki/Bernard_Rimland).

Over lunch, Bernie and Kerri bonded immediately. None of us at the table realized that Bernie was suffering from cancer and had less than a year to live. Driving the short distance back to my clinic I made the observation to Kerri that I suspected she was about to become a scientist, and no longer just a trophy wife. "Be careful," I warned her, "Bernie Rimland's mantle will be heavy on your shoulders."

Kerri turned her face toward me and said, "I will carry that mantle willingly and with honor."

After a few days, Bernie told me that he thought Kerri had what it takes to carry the torch and help families heal their children with autism. I saw a lot of Bernie in the next nine months. He would come down to the clinic for regular hyperbaric treatments because it eased the pain of his cancer. Just before his death, he told me, "I will not die before I have found the reason for and defeated autism." But die he did on November 21, 2006, at the age of 78.

As I drove to Bernie's grave, I anticipated many, many people lining the way with some hundreds even thousands thronging the memorial park, similar to Giuseppe Verdi's funeral, in 1901, where 200,000 people gathered. At his graveside, the thousands of mourners spontaneously burst into the "Slaves Chorus" from his Opera, Nabucco.

I was stunned upon arriving. Probably less than thirty or so weeping caregivers of children with autism were present at the burial, along with Gloria and Bernie's three children. WHERE WERE ALL THE PEOPLE, THE MOTHERS AND FATHERS, THE RESEARCHERS AND SCIENTISTS? I asked myself. It was my first Jewish funeral and I think that the booming sound of the shovel full of earth that I threw down onto his plain casket made me look heavenward and ask, "God, where are you now?"

I drove home, eyes unfocused with tears, and stone eggs of the birds of unhappiness in my heart, stopping only to purchase two fine bottles of wine to toast Bernie. I knew one would not be enough. Arriving at my house, I found the fine crystal goblet my Mom had left me, filled it, and played Neil Diamond's, Morningside over and over, with each playing, another glass, held high to Bernie...

Morningside
The old man died
And no one cried
They simply turned away

And when he died
He left a table made of nails and pride
And with his hands,
He carved these words inside
'For my children'

Morning light
Morning bright
I spent the night
With dreams that make you weep
Morning time
Wash away the sadness
From these eyes of mine
For I recall the words an old man signed
'For my children'

And the legs were shaped with his hands
And the top made of oaken wood
And the children
That sat around this great table
Touched it with their laughter
Ah, and that was good

Morningside
An old man died
And no one cried
He surely died alone
And truth is sad
For not a child would claim the gift he had
The words he carved became his epitaph
'For my children'

In fact, as I sit here writing this, I play Diamond's melody and listening to the words makes my heart swell and my eyes become damp.

The morning after Bernie's funeral, hung-over and miserable, I pondered: Did my beloved friend Bernie die feeling that he failed? I now doubt it for he had passed his mantle to carry on his search for the cure without fear, without the need for accolades and rewards and to do so with integrity. I did not know it, but before his passing, Bernie and Kerri had been in communication. Bernie had given her permission to translate his *Defeat Autism Now!* Protocol into Spanish, Kerri's other language, as well as other tasks.

In review, I now see God's Hand at work—and it started one afternoon at the Green Tomato. Thus, my question to my friend on free will versus predestination right at the start of my notes on this section.

Brain overload and frozen fingers.

At the time, when Kerri asked me to write this section of her book about hyperbaric oxygen therapy and the treatment of challenged children, my mind went into overload. Why? Kerri was (in my view and that of others) the most efficient and effective health scientist with a high normalization rate of challenged children (autism, etc.). With her own hospital-quality hyperbaric oxygen center, Kerri had first-hand knowledge that to normalize such an afflicted person required a lot more than some trips in a oxygen chamber or some Silver Bullet response from a pharmaceutical company.

My brain whistled like an express train rushing through a tunnel. I thought of the dozens of children that hyperbaric oxygen therapy has helped at my different centers. Some of the little guys were so badly damaged they would repeatedly punch at their own head, or chewed on their own wrists. I'm thinking of Marco as I write this. Ten years later, he still punches. A clearly defined casualty of a hotshot thimerosal vaccination, normal until three, his parents won a multimillion-dollar lawsuit against Big PHARMA. There has been some progress over the years; but it will be a matter of unrelenting attention to Marco.

I remember another little guy; a handsome, articulate smiling boy of about 12 years of age named Chase. Meeting him and his parents, you would never know he had autism or ADHD. Chase was a "runner." When his treatment in the chamber was over, he would promptly disappear out of the clinic doors. Eventually, Chase would be carted back down by one of the upstairs nurses. "Chase" by name and chased by nature.

After one such session in the chamber, he rocketed out and into the restroom of the clinic. I was sitting at the controls when he came back. Chase climbed up onto my knee and said, "I love you Bob," and he rubbed his hands across my face. From the terrible odor I knew in an instant what was all over my face and my lips. Chase had used his hands instead of toilet paper and bypassed the washbasin on the way out.

"Oh goodness," came from his horrified mother. "He got you. I am so sorry. It is one of his favorite jokes."

"Not to worry," I said, "I have eaten and tasted this in a variety of forms most my life." So, whenever I hear somebody say, "Sands is not a Doctor so he doesn't know s—t from Shinola," I think of Chase.

Normalizing kids is so rewarding, but you have to take a chance on s**t.

My fingers were not ready for the keyboard as I tried to wade through dozens of experiences with these children, my own personal experiences, and finally adding input from experts in the field of pediatric abnormalities. (Many of these experts became my close, personal friends as the years went by.)

For any parent or caregiver to delve into all of the opinions and conclusions of the so-called experts, it is a bewildering journey through a lot of contradictory and self-serving educated guesses and erroneous assumptions. To arrive at something that the reader will identify with, and believe in, I have to go way back. After 70 years on the face of this planet, I have reached certain conclusions about the treatment of children that enables me to write this section. Mind you, I'm not stuck in my ways and I am open to change my thoughts and conclusions as the science improves on treating children with autism.

Little Patrick Rivera, Little Bobby Sands Something in Common?

Here is something that I doubt my adult children even know about. In 1947, polio was raging throughout Australia. I was seven years of age and had developed a sore throat and fever. My father, a wealthy man, called the very best doctor to my bedside—yes physicians did make house calls in those days. I can even remember the doctor's name, Dr. Bradfield. His father had been the engineer that designed the Sydney Harbor Bridge. My dad, with all of his money drew a lot of water in those days and could get almost anything he wanted. Bradfield came in and examined me briefly; after all he had other house calls that night.

Bradfield told my Mom and Dad that he did not think I had polio, but perhaps diphtheria and that I should spend time in isolation in the infectious disease hospital overlooking Botany Bay, "just in case." The next morning an ambulance came and, in spite of my fear, pleading, misery, and tears they carted me away. The three-hour journey became a six-hour journey. The ambulance broke down outside the Sydney airport. I remember looking out of the rain spattered window and seeing the aircraft clearly visible over the fence. We sat on the side of the road for almost two hours before the replacement ambulance came to transport me, and my journey continued to a huge colonial building complex, built by convicts and perched on rolling, green lawns that looked down onto the Pacific. The complex was known as, The Coast Infectious Disease Hospital, which later became Prince Henry Hospital. I was terrified and clutched my teddy bear, *Threadbare*, under the blankets. Threadbare became my only friend for the next four months, and a glimpse of airplanes over the fence became an enduring memory of the trip.

My parents were already there and had obviously filled out all of paperwork. I never got to say goodbye to them since it was an infectious disease hospital. As they walked past a big hopper window, they paused momentarily and Mum waved through the window. I saw them maybe three times in the next four months when the nurses wheeled my bed up against the window that looked out into the hall. They would tap on the glass and smile at me but because my father was quickly bored, after about five minutes they would wave and be gone for another couple of weeks. It was a long way for them to travel.

The night of my arrival, a team of doctors and nurses arrived and they rolled me onto my side. Full syringes lay on a tray; in those days they were glass with little thumb and finger holes on each side and each needle was reused and supposedly resharpened before their next use. This meant that the needles were thick and long. One-by-one these needles were inserted into my upper thigh on the left-hand side. The pain was agonizing and I'm sure I screamed and screamed. I remember at least three nurses holding me down. At seven, I was just a little fellow. There had to be at least 20 of these very large syringes. Checking the history books, I believe that they were mostly penicillin and sulfa drugs. I might add, it was already known to Doctor Bradfield that I was highly allergic to sulfa and remained so for my entire life.

The next morning, when I awoke, I found that I was paralyzed from the feet up to just under my arms. Threadbare was my only friend for the next four months. The days all blurred together. One thing I do remember vividly is that I used to deliberately wet the bed. I did so only when a lovely red-haired nurse with a band of freckles across her nose was on duty. This nurse was kind and gentle, probably in her twenties. After she had changed the sheets, she would stroke my cheek and tell me that everything was going to be okay and not to worry about wetting the bed. The other older nurses scolded any of the other pediatric patients if they had an accident. At some point later, my gentle nurse told me that I really did not need to wet the bed to get her attention. I soon stopped and I still got a cuddle from her. Such a display of care and affection was unknown to me from my parents back at home.

After that stay in the hospital, I was never the same. I know that there was constant discussion amongst my family members that I should be placed in some sort of an institution because my behavior was "contrary and stubborn." I was either talking my head off, trying to make my family love me, or distant, and as such, I was said to be "uncooperative." Soon after, I was placed in a Hogwarts-style boarding school.

Could that childhood experience have made me what I am today, forever alone, but rarely lonely? As the years ticked by, to my surprise (and I think everybody else's), the school counselors found that I had a genius IQ. Amongst the 1,200 boys, I was second from the top.

After seventeen years of treating so-called challenged children, I have yet to find one that was not as smart as a whip in something.

I managed to burn every social bridge at the boarding school, thrown out of every special social club at Newington. On Saturday afternoon, after mandatory sports such as Cricket or Rugby, the other boarders would get to do things they wanted like billiards, photography, or gardening (these were clubs that you had to apply for and join). However, I had been expelled from each of them so I wandered alone each Saturday afternoon and finally picked the lock on a large room-sized cage full of Sulfur-Crested Cockatoos and other Australian parrots. I would sit alone with them and be jeered at by the boys in the garden club that were planting and weeding. Within weeks, these birds would happily climb all over me, enjoying my company. The birds loved me, but the boys and masters did not. To me, it was a two-way street. Animals made better friends than people. Eventually some of the other boys would come and beg to be allowed to accompany me into the cage so they too could have that same privilege.

Do not get the wrong impression of my beloved Newington College in Sydney Australia; even today I dream of the sanctuary of that place. It was better than what I had with my family back home.

The really curious thing is that my mind had blotted my hospital stay out, or had hidden it in some black cabinet in my brain where the really bad things are kept to allow me to get on with the reality of NOW. I remembered the illness and isolation only when Kerri asked me to write this. The details have returned with clarity: the old iron beds, the smiling face and freckles of my favorite nurse, being cold at night, the colors on the wall, the distant sight of the ocean through a window on the other side of the big ward, the frame of the window so far away that it looked like a painting hanging there. If I close my eyes as I write I can see it now.

I now ask myself, “Is this the way children with autism feel?”

I don’t know. What I do know is that my old friend and confidant, teddy bear, is sitting on the desk as I write this. I reach out to rub Threadbare’s head and a great calmness washes over me. The little fellow is 64-years-old now and there is a great competition amongst my adult children to see who gets him after I die.

Botched diagnosis
The incorrect pharmaceutical response.

With almost 40 years of working in medicine behind me, and now in my retirement, I do know that neither diphtheria nor polio were the problem. The suffocating “Bull Neck” or the “Strangling Angel of Children,” diphtheria, had not happened to me, nor did the withered limbs of poliomyelitis set in after paralysis.

Simply put, Dr. Bradfield cocked it all up in a big way by guessing, not taking enough time and arriving at the wrong medical response on his visit. The relentless attack on me with sulfa drugs caused paralysis, which was probably what is commonly known today as Guillain–Barré Syndrome, a disorder affecting the peripheral nervous system. Ascending paralysis, weakness beginning in the feet and hands and migrating towards the trunk, is the most typical symptom; the disease is usually triggered by an infection.

It sure sounds similar to what happened to little Patrick, and probably to the reader's challenged child as well.

Migrating to America in 1990 with my tribe of offspring.

I arrived in the United States over twenty years ago. If the reader is interested enough, a quick look at a funny but classy "roast" presentation is on YouTube. You can find it by searching "Robert Lyne-Sands" on YouTube.

After completing the United States Navy transportable chamber development project, I arrived in swank, Pacific Palisades, on the outskirts of Los Angeles; living amongst all of the Hollywood greats. I was going nuts with boredom and became the "Bob Clampett" of the area; jogging down the leafy streets I always said Gudday to any celebrity. I rarely remembered their names but knew their faces. They never replied on the first go past, but then I would jog backwards, keeping pace with them (big stars like Madonna always ran with a group of minders) and I would yell out "I said Gudday. My goodness, your manners are appalling." I might add, it was not just me; they never ever nodded to each other when their paths crossed either.

Around this time, I opened the very first freestanding hyperbaric oxygen therapy center. It was more or less hidden away in Santa Monica on 6th Street. Tucked away in what used to be an old motel, you could almost refer to it as a "speakeasy" center, akin to something from the days of prohibition, where you had to knock on the door three times before being let in. I would only treat people who had been referred by another patient. Most of these people were from the entertainment industry in Hollywood, many with the HIV virus, and lots of folks with chronic fatigue syndrome.

The kiss of death in Hollywood is to be known to be ill. Hyperbaric oxygen therapy works so well that the word soon got around, and I was inundated with patients, some famous and well-known, and others just hardworking folk in the movie industry. Thus, I ended up being known for some years as "Mr. Oxygen," and because of the remarkable effects of HBOT, I befriended many of these celebrities, who subsequently invited me as their guest to many of the big Hollywood functions (the only one that I missed was the Academy Awards).

One of the first of these talented people that became my close friend was Harold Michelson. Harold was an award-winning art director working out of DreamWorks. He was suffering the aftermath of a mild stroke. His wife Lillian had told me that her Harold loved crossword puzzles, but since his stroke he would put the across answer in the down column. This caused him immense frustration since his answer was usually correct, just in the wrong place. Five treatments in the little chamber were all it took for Harold to get his verticals and horizontals in the right place. The big thing that he noticed was that his blood pressure stabilized, his lung problems normalized, and he needed less medication for his diabetes.

Harold still stood well over 6-feet–tall, though stooped with old age. We hit it off immediately because Harold had a quick wit and loved a good laugh. The stories and banter between us were constant. Harold was extremely talented yet self-effacing. He really was a hero in the truest sense of the word. As young man, he made over 40 flights across the English Channel to bomb Germany. As navigator, Harold sat in the nosecone of the B-17 and either froze with the cold on the trip or froze with fear when making the lower-level bombing run. "Bob, I was so frightened of the flak from the antiaircraft fire from the Germans on the ground. As the puffs of smoke would erupt around the front of the aircraft, shrapnel would pierce plexi-glass windows in front of me. I would hold the navigational map in front of me and try to believe that the thin paper would stop the lethal fragments."

The first challenged child to be treated in a hyperbaric chamber—Emily.

I need to jump forward in time for a moment before getting back to Harold Michelson and his wife, Lillian. I went back into production of hyperbaric chambers after I realized the great need for hospital-grade dispensing of oxygen to the public. There were no rules preventing me from doing this even though I'm not a physician. All my patients needed was a prescription from a registered physician; that was the easy part. My first center was in San Diego.

In the first few months of patient treatments we received a call from a father in Canada. His daughter, Emily, had cerebral palsy. The father had read that hyperbaric oxygen therapy would perhaps help his little girl normalize, and so when he asked me what I thought I gave him my usual answer, "I do not know, I am not a physician." And then I added, "I would hate for you to waste your money on a trip from Canada to San Diego if it doesn't work."

The father replied, "I have two choices. I can go to England where there is snow on the ground or I can come to San Diego. It is an easy choice to make."

So father and daughter duly arrived. Emily was the most delightful child. When she walked through the door she had a crash helmet on, knee guards, and wrist-guards to prevent serious injury when she would crash, face first, into the floor. We learned a lot in the first ten days of treating Emily. One of the first things was the fact that fathers can never properly do a good job on their daughter's hair. I felt for the father since Emily had lots and lots of hair and was never still for a moment (I have a few daughters myself). Emily would arrive looking like a rag doll; her little blonde head would have pins, bits of string and ribbon through it. So the very first task of our nurses would be to grab Emily (usually tearing around the clinic), and comb her hair out to make her more presentable.

Now, the really big thing that we noticed was that on day ten she arrived at the clinic without any of her body armor. Yes, she still "toe-walked" because her heels did not touch the floor. To compensate, she would park her little fists against her chest, push her elbows out to keep a center of gravity, and rush around the clinic. On day 11, father and daughter spent five hours at the San Diego Zoo without the usual body protection. Emily never lost balance or tumbled even once. Follow-up visits over the last 17 years find that Emily is living a normal life. Yes, she still has some walking problems, but never needed to undergo the mutilating surgery that was recommended to her by her Canadian doctors. During 20 days of hyperbaric oxygen treatments in San Diego, something had changed in Emily's brain.

Likely the most important introduction in my life... The Great-Spirit-in-the Sky's (God's) hand at work.

Listening to the improvements that little Emily had made, Harold Michelson was both enchanted and excited.

"Did you know that Lillian and I have an autistic son? We need to introduce you to a couple of people in San Diego. Dr. Eric Courchesne, who runs the autism research center for the University of San Diego, known as the Autism Center of Excellence or ACE."

Lillian chimed in and said, "I really must call them about you. Eric is a gem and his wife Rachel is also a research doctor. They work side-by-side and will be so helpful."

With Emily's positive results, and more children making gains at the center, I called the Courchesne couple at the ACE and offered to provide them with the use of a $150,000 hospital-grade hyperbaric chamber for one full year. There was no response to my offer.

While you may imagine that the UCSD-ACE Research Center is located in some lofty center of learning, it is in fact to be found in a delightful millionaire's village known as La Jolla in Southern California, a block or so from the Pacific Ocean. On the top floor of a three-story building, an elegant balcony ran across the front of the building, while red and white umbrellas and tables make it look more like a posh restaurant than a research center.

I waited about a month and called again and repeated my offer of a hyperbaric chamber, operated by trained technicians for a year, or whatever time it took to make a decision on efficacy, at zero cost to the ACE. About an hour after making the second call, my secretary told me that a Dr. Rachel Courchesne was on the line and needed to speak to me on an urgent matter. I picked up the telephone, anticipating a good conversation but this is the way that it went, all one-way at me. Dr. Rachel was in fine fury and, looking back, the only way I could have put a word into that conversation would have been to fold the word flat and slip it into the tirade, sideways. She told me that it was amateurs such as myself that muddied the scientific waters of autism research and that, essentially, I should crawl back into whatever hole I came out of.

She made it clear to me that the mission of the ACE was to perform research to find a cure with pharmaceuticals. In fact, the ACE receives funding from drug companies, and that I should not bother them again. I was taken aback, to say the least.

A decade later, Dr. Eric is apparently doing fine work. For example, in his recent research, quoted widely in the media, he has found that the frontal lobes of an autistic child's brain is abnormally large. This is an observation, perhaps useful in the future, but only time will tell. But does it matter to you the reader whether this is so? Maybe yes, maybe no, so let's hold off on that sort of thing until we get into the scientific part of the book.

Let's get back to Dr. Rachel's harangue of me. In my mind, there is no doubt about the fact that for her, at least, it was all about the money and the potential for a patented silver bullet for Big PHARMA. If successful, a drug company could charge you a huge amount of money to heal your child with autism. At worst, there would be the ability to raise more money for research despite scant evidence of clinical results on children.

An example of this ability to raise money on a potential medical marketplace is the fact that in the year 2012, the National Institute of Health (NIH) allocated $100 million over a five-year period to a bunch of ACE research across the United States. So much for Lillian's glowing endorsement of Dr. Rachel.

However, the Michelson's other introduction proved to be a link in the chain.

Autism research with integrity—Dr. Bernie Rimland.

"A dear friend of ours, Dr. Bernard Rimland, also has a research center in San Diego which he calls *Defeat Autism Now!*, or DAN! for short. Bernie's son, Marc, is a true autistic savant. The movie with Dustin Hoffman and Tom Cruise, Rain Man, was modeled around Bernie and Gloria's son. Bernie was the technical advisor for the film. You must meet Bernie."

Lillian was especially grateful and fond of Bernie and his wife, Gloria. Back in the early 1950s, autistic children were attributed to cold, hard, heartless mothers. Many psychologists and psychiatrists agreed with the "Emotional frigidity" hypothesis, which suggested that the mothers were the cause of what we now know as autism. A lot of children in the United States were taken by force from their parents, and placed in isolated institutions to "protect" them from their parents. Bernie Rimland alone stood against this hypothesis; after all Bernie did have degrees in psychology, including a PhD. In fact, it was Bernie that took away the shame, the blame, and the guilt of many loving parents who were bullied into believing labels put upon them by the so-called "learned."

Driving forces behind autism research centers.

Unlike the La Jolla location for the ACE, Bernie's Autism Research Institute (ARI) was about five minutes drive from my San Diego hyperbaric oxygen dispensary; two minutes of freeway, a quick turn into Adams Avenue, and I had arrived. Again with an erroneous assumption, I had thought that Bernie's ARI would be similar to the ACE La Jolla location. In fact, I drove past Bernie's place three times before I found it. An unassuming, clay-colored, converted storefront, with a little sign on the front door that said, "Autism Research Institute." I knocked on the door and walked in. Instantly, the smell of antiquity and old books overwhelmed me and beguiled me, like something out of a Harry Potter movie. Stacks of papers were high on the floor almost in a willy-nilly fashion. A gentleman dressed in a checkered shirt came towards me. "Bob Sands? The Michelson's said to expect you." He peered over his glasses and gray whiskers at me. In that one glance, we created a bond of friendship that lasted for more than a decade. At least once or twice a month we would lunch together along with his wife, Gloria; always two blocks from his Center, at the *Bleu Boheme*. Bernie and Gloria would carefully examine the menu and then they would both look out and recommend, "Let's all order the fried green tomatoes."

Bernie was interested in clinical results. He would turn over any stone that would perhaps unlock the mystery of autism. When I first met Bernie, 1 in 850 children born in the United States was afflicted with this mysterious and terrifying disorder. As I write, autism spectrum disorders have spiked to 1 child in every 88. So when the phone rang at the ARI, Bernie would often answer it and ask many questions, if the person believed that they had found a cure or a clue. He never discarded anything, in stark contrast to my encounter with Dr. Rachel at the ACE in La Jolla. I once listened in when he spoke with a lady who believed that you could cure autism by filling crystal glasses to different levels and rubbing your fingers around the edges to create pure tones.

Bernie put the telephone down and said to me, "It could be worth looking into. Music does create new neuronal wiring in small children, for example, or a mother singing to a child." I had already mentioned to him that when a child with autism was in the chamber, if you played unfamiliar sounds such as whale songs, or bagpipes, the unresponsive child would often brighten up and look around.

"Maybe new neuronal connections are being established," Bernie pondered as he rubbed his whiskers.

In those days, I worked on weekends. If I wanted to chat with Bernie, Sunday was the best day. No staff to answer the telephone, so he always did. I once asked him whether he took a day off and his response was succinct:

"When autism takes a day off so will I."

Bernie, a $2.6 million donation and his ARI.

Now, here is the most interesting of Bernie Rimland's driving forces—to me at least. It was not about dollars or ego, not in the slightest.

At that time, I was also friends with another Bernie, the renowned anthropologist, Dr. Bernard Aginsky. To avoid confusion between the two Bernie's I will refer to him as Dr. Aginsky. He was in his late 90s at the time, as rich as Croesus. For many years, he had leased the penthouse of the La Jolla Shores Hotel, right on the beach of one of the most desired parts of Southern California. Again, like most of my friends, Dr. Aginsky was a humble man and thought nothing of his wealth other than as a means of accomplishment. Out of the blue, one day in my office, Dr. Aginsky looked at me thoughtfully and then asked, "I have $2.6 million and I want to give it to somebody that would make a difference. Can I give it to you?"

I immediately thought of Bernie and the children with autism and the nonprofit status of the Autistic Research Institute. "Hold that thought Dr. Aginsky. I will give you an answer in a couple of days."

A couple of days later, Bernie and I were traveling to take part in a medical internal review board. I was wearing a shirt with French cuffs and a silk tie. Bernie had on his usual checkered shirt. He confided in me once that he had to make a choice between wearing a tie or trimming his Santa Claus-gray beard. The beard always won. Wondering about his shirts, I respectfully asked him whether he ever received a salary from his ARI.

"Good gracious. I could never pay myself a salary from ARI funds and donations. You must remember that I am on a full Navy pension so it would be wrong to double dip."

After the meeting, as we traveled back along the I-405 to San Diego, I broached the subject of Dr. Aginski's offer of a $2.6 million donation for the Autism Research Institute. Bernie looked out the window for a few moments and fiddled with his beard. I could see he was calculating in his head. He turned to me and said, "We have so much money in the ARI at this time, donations and such, that I think that amount of money could be used elsewhere. Find another recipient."

Dr. Bernard Aginsky, a widower, passed away in January 2000. I have no idea where his donation ended up. I do know whether Bernie Rimland ever had a second thought about not taking that money. He never mentioned our conversation again.

DAN! changing to parallel other autism research.

Research costs money. Bernie used his resources wisely and was not interested in empire-building. He just wanted to find the cause and the answers. Bernie cured kids.

On the anniversary of his death, I went to his Kensington ARI center. I had two bottles of good wine, a bunch of wine glasses, and a whole lot of chocolates. I walked in and was astonished. The whole tenor and tone of the place had changed. The clutter was gone. Bookshelves were neat and organized. The smell of scholarly interest in documents and old books had evaporated. It was as if the ghost of Dr. Bernard Rimland seemed to have moved to another place completely. The staff were delighted to see me, and happily munched on the yummy chocolates that I had brought for them. I commented on the neatness and the vast changes that I observed. One of the staff shrugged her shoulders, grimaced, and said,

"You have no idea just what changes have taken place, and what is about to happen."

An example of the transformation in attitude is exemplified by an offer for inflatable, aka soft-sided, chambers for DAN! while Bernie was still alive. One day, he called me up and explained to me that a group was coming to do a presentation on these soft chambers, and suggested that I plant one of my staff at the ARI to listen to the blurb from the manufacturer, Oxyheal. My secretary, Crystal was waiting amongst Bernie's staff to greet the technicians, doctors, and such that accompanied the owner of the company. They arrived at the ARI, set up the inflatable chamber, and gave their sales pitch, extolling the virtues of hyperbarics at low pressure and at a low cost.

Crystal reported back to me that Bernie looked dour during the presentation and asked, "What would happen to this with a razor blade?" Bernie and I discussed the concept of inflatable chambers the next time he was at my center, after he had finished his treatment in a chamber. He was not convinced about the soft-sided chambers.

"I worried that the scientific data that they quoted was all from hospital quality chambers. I think you are correct Bob, the physics do not hold up. I could not recommend this device to parents for their children."

The "discovery" (observation) of God's gunpowder + Bernie Rimland and Kerri Rivera.

Throughout history, air was taken for granted. You just breathed in and out and there it was; bugs and birds flew in it, mostly useful to keep your floors and ceilings apart. Even in this day and age, many folks think of it in the same way.

Let´s revisit what is now called the Age of Enlightenment, the 17th and 18th Centuries. About 1774, the kindly and soft spoken clergyman/scientist Joseph Priestly and his close friend Benjamin Franklin. They spent time drinking coffee together once a week in a London coffee house along with other scientists. Together they realized that air was actually a mixture of gases, but mostly nitrogen and oxygen. Priestly and Franklin could kill a mouse in a glass jar in less than five minutes by depriving it of one of those gasses. Put the poor little mouse in a glass jar with a lit candle (which consumed the oxygen) and the flame of life in the mouse departed by about the time the candle went out.

Using precise scales, Joseph Priestly and the French chemist Antoine Lavoisier, gave oxygen its name. It could be weighed on a scale just like you would weigh out a cup of rice. Each unseen gas has a weight. Baros is the Greek word for weight. Words that we are familiar with come from this: barometer, barometric pressure, and hyperbaric (as in hyperbaric oxygen therapy, hyper meaning more than and baric meaning weight of the gas).

Weigh oxygen?

The weight of air (mixed oxygen and nitrogen) is 14.7 pounds per square inch at sea level. Most folks like me would prefer to use the word pressure, however, the math is easier if we use the doctors measure (the same as that used in measuring blood pressure – millimeters of mercury or mmHg). An average blood pressure measure is typically 120/80 mmHg. Let's look at it this way; if air at sea level is 760 mmHg, then 20% of that air (oxygen) is 20% of that amount. So multiply 760 by 0.20 and you will realize that vital, life-giving oxygen going into your lungs is about 150 mmHg.

Oxygen - You cannot see it, or smell it but you have to "eat" a kilogram of it each day to stay alive.

Make no mistake. If you take a substance into your body, ALL of your 60 to 90 trillion cells will be affected. Snort it, chew it and swallow it, breathe it, or inject that substance and it will have a positive or negative effect on your whole body. Cells "consume" the substance. Another word for consume is "eat." We will discuss this process, known as oxidative metabolism shortly and, especially how to produce stem-progenitor cells.

Eat oxygen?

Oxygen is a cellular food. The adult human consumes approximately 0.6 kg of oxygen each day (about 1.2 pounds). We can live about three weeks without the food that we chew and swallow, survive three days without water, but we can only live approximately three minutes without oxygen. Body cells start to die—the brain is the most vulnerable. Even a minute or so without oxygen can cause permanent damage.

Don't worry about the earth running out of oxygen. Back in the 1700s Ben Franklin and Joe Priestly also discovered the fact that sunlight on green vegetation produces oxygen—a process called photosynthesis. They showed this with the mouse and candle in the glass jar. As the candle flame seemed to flicker, about to go out and the mouse lay, apparently dead, if there was a sprig of a green plant, both the candle and mouse came back to life. All of the oxygen that you are going to eat in the next 24-hours is replaced by just one square meter of grass producing oxygen through photosynthesis during that same length of time. In populated areas, where there is not much greenery, life-sustaining winds carry oxygen over the oceans and hills to all humans and animals.

Captain Cook, Joseph Priestly, Bernie Rimland, and Kerri Rivera—Explorers

Before we stop talking about Joseph Priestly, he made another observation about gas and this applies to our subject of autism, Bernie, and Kerri.

Living next door to a brewery, Priestly was intrigued about the gas given off by the fermentation of beer in the barrels. He put a dish of water under the vent-pipe coming out of the barrel. When he tasted the water it was full of refreshing bubbles we now call fizz. In 1767, Priestly discovered how to carbonate a liquid. He taught the Navigator, Captain Cook, how to carbonate water and, thinking that drinking that bubbly water would prevent scurvy, Cook set sail with it in 1772. It did not prevent scurvy, but, just like Bernie and Kerri, everything is worth investigating.

We now know that any gas becomes soluble if it is concentrated or made thick. In this instance, the gas was an extremely light gas known as carbon dioxide (yes, the stuff we breath out and, unlike the 21% oxygen, there is only about .033% of carbon dioxide in the air we breathe). However, if you concentrate it to 100% some of it melts into the water and becomes carbonic acid, as is the case with soft drinks. Pop the lid and the acid turns back into carbon dioxides gas, forming little bubbles in the liquid. Joseph Priestly should have patented his observation. Ten years or so later, Joseph Schweppes patented the process in Germany and most folks think of him, not Priestly, when they have a gin and tonic.

Thereby hangs another Newtonian rule about gas that cannot be changed. "Gas under pressure becomes soluble in a liquid according to its density (or thickness)." – Henry's Law.

And this is how oxygen works in a hyperbaric chamber – Newtonian Gas Laws

Stay with the soda can thought and the thicker gas melting into the sugary drink. When a human climbs into a hyperbaric oxygen chamber, we substitute oxygen for carbon dioxide. Believe it or not, you do not feel squashed by the pressure other than you have to "pop" or equalize your ears. Humans are mostly water and other fluids that cannot be compressed like a gas. What is happening in the chamber is that the oxygen (it should be 100% oxygen) is being made thicker or concentrated. And, just like the soda, Henry's Law kicks in. The oxygen melts and the body supersaturates with oxygen.

Another curious thing about a well-oxygenated body is that it becomes slightly alkaline. Your powers of hydrogen (pH) read about the same as seawater, which is alkaline because of the dissolved oxygen in it. Yes, even fish need their daily meal of oxygen!

The word "saturate" comes from the Latin word meaning "to completely fill." Since only our red blood cells (RBC) carry oxygen, they are almost always full of oxygen in a healthy human. The little pulse-ox device that clips on your finger when you are in hospital almost always reads 99% unless you are anemic.

It takes between one and three minutes for each RBC to leave your heart, make its complete circuit, and be back in your lungs for its next load of oxygen. Remember that Henry's Law also applies here, therefore, oxygen is a gas in your lungs but turns into molecular oxygen when it enters your body's fluids. This is so complex that there is no chemical symbol for it, but rest assured, your body's cells feed on it with a great appetite. (More on oxidative cellular metabolism shortly.)

Supersaturate?

When you fill a glass of water to the brim, it is full or saturated. Right? Yes. Holding it under the faucet and running more water into the glass will only mean that the water slops over the edges. The glass is still full. However, this is not the case with any gas, particularly oxygen breathed in by a human. Sure, just like a glass of water, oxygen can be measured by volume, but there is plenty of room to over-fill our bodies with oxygen—supersaturated.

To illustrate a complex process, let's paint a simple picture—An underground subway system.

Think of your lungs as a subway station. Passengers get off the carriages and others get on. In this instance, remember that air is made up of (for easy math calculations) 80% nitrogen and 20% oxygen. Visualize each gas molecule of either type as a gas-person. The body is already full of nitrogen at sea level and since it does nothing at all (it is inert), the nitrogen-people just come in and out your lungs. However the oxygen-people are dynamic and bustle to climb on board the little red-blood-cell carriage (which has a special oxygen-people magnet—an iron molecule), and away the red blood cell rushes with the oxygen-people ready to jump off to the hungriest of the body's cells.

Since Henry's Law allows a gas to melt into clear fluids (and our body has lots of clear fluids), there is a whole lot of room to add more oxygen-people. For example, if we barred entry to our lungs of the nitrogen-people (simple to do—just put on a mask and breath oxygen out of a 100% oxygen tank), then our lung-subway station would have five times more oxygen-people and that would mean that we increased the oxygen-people with each breath from 160 mmHg to 760 mmHg. Remember that they are dynamic, and that each red-blood-cell carriage has an oxygen magnet in it. The thick crowd oxygen-people all bustle to get onto that carriage, through the wet soapy surfactant in the wall of your lungs, even though it was already full.

Essentially, the oxygen-people fill the red-blood-cell carriage, and into the plasma where they are quickly swept along in that clear plasma fluid. Then, when even more oxygen-people continue to crowd on back at the subway station (i.e. the lungs), the plasma, which normally does not carry oxygen-people (molecular oxygen), then needs to put the overload of oxygen somewhere. From here the oxygen-people go out through the rail tunnels (i.e. the circulatory system) and into the clear fluid (i.e. interstitial fluid) that bathes our body cells. As this fluid fills with oxygen, it dumps it off into the body's largest water compartment—the body's cells (the intra-cellular compartment). These cells now have a veritable feast of oxygen available.

About 90 minutes after you start breathing 100% oxygen, you are supersaturated with five times the normal amount of molecular oxygen. No, you do not swell up since we are talking about molecules in existing body fluid, not added fluid. Curiously, if you put the little pulse-ox back on your finger, it would still only read 99% since it is just measuring the oxygen in your red blood cells. However, using high-tech oxygen measuring equipment (transcutaneous oxiometry), you can even measure molecular oxygen in your urine, saliva, and tears—something once thought impossible.

Now comes the hyperbaric chamber part. You climb in, the door is shut, and 100% oxygen floods in. The technician can look at the physician's prescription and increase the thickness of the gas in that chamber. Double the thickness (i.e. weight or pressure) to 1520 mmHg; ninety minutes later you will climb out of that chamber with ten times the number of oxygen-people (molecular oxygen) in your body. Or, put another way, since your body's cells are going to consume (eat) that additional oxygen, this is the only way you can increase the weight of oxygen-food available in a 24-hour period. One hundred percent pure oxygen at sea level and you will "eat" between just under half a kilogram to just over one kilogram of oxygen.

The deeper/thicker the oxygen is in the chamber the more you get. In fact, following Dalton's Universal Gas Laws, by doubling the pressure, you can actually get your daily oxygen meal at almost two kilograms.

Now the nifty thing about oxygen supersaturation.

Unlike many pharmaceuticals that are weight dependent, (too much of a certain drug will sicken or kill a smaller patient, not enough of the same drug and there will be no therapeutic benefit to the patient) with hyperbaric oxygen no individual can suffer from an "overdose" of oxygen, regardless of whether the patient is a mouse, rabbit, horse, or human. All climb out of the chamber, (depending on the pressure/thickness of oxygen they were treated at) with the same amount of molecular oxygen in their body. Not one person has suffered from an "overdose" of oxygen.

Our oxygen dispensary centers have treated, according to their physician's prescription, tiny babies, the elderly, and everyone in between; we have never seen a negative side effect. Instead, we have observed clinical positive effects for all.

What happens when we actually give people an extra "oxygen feast" by putting them in the chamber? Again, an extremely complicated process has to be simplified for understanding. However, before making things simpler we must first look at the science. Chatelier's principle predicts the effect of a change in conditions on a chemical equilibrium. The principle is named after Henry Louis Le Chatelier, who observed that if a chemical system at equilibrium has a change in concentration, temperature, volume, or partial pressure then the equilibrium shifts to counteract the imposed change and a new equilibrium is established. Any change in the status quo prompts an opposing reaction in the responding system.

Time to change just a couple of words for better understanding. Let us start with the word equilibrium and now call it "the way it was." We can also use that term for status quo and counteract can be changed to "pushing back." So what is really being said here is, if we change something in any chemical environment it will push back hard to be the way it was. Now it is time to make things even easier.

We are all chemical systems.

We have already mentioned cola in a can and advised against giving it to anybody, your child or yourself, because of the acidity of the liquid in the soda. But there is an even more compelling reason—the sugar. There are 39 grams of sugar, about ten teaspoons of sugar, in one can of soda. Would you feed your child that much dry sugar?

"Come on," you might think to yourself, "No way. It does not taste that sweet." It is true. The manufacturers of sodas use a lot of phosphoric acid to disguise the sickly sweetness of the drink. If you still doubt this, take an ordinary penny, or an egg and put it in a glass full of cola overnight. Within a few days, you will find all tarnish removed and instead a bright, shiny penny; the eggshell will be soft and pliable, all its protein dissolved.

Visualize what happens when you allow your small child a glass of cola (or any soda drink). Your little kid becomes supercharged, bouncing off the wall, and for some time you cannot get the child to sit still. In other words, your child has become hyperglycemic. Too much sugar, of any sort, whether it be corn fructose, beet or cane sugar, the results are the same according to Chatelier's principle. Remember, we are all chemical systems and by adding all of the sugar to a small child's body, all of the cells in that body will push back to return to the way they were before the additional sugar. The most efficient way to return to this state is to increase physical activity, causing cells to consume all of the additional added sugar.

Run, baby, run!

Another way, of course, is to rely on the body to use its own chemicals to burn off the sugar. This is predominantly the job of the hormone insulin. If an adult drinks a lot of soda each day (or even has one gin and tonic, which has about half as much sugar as a soda) and does not increase physical activity, diabetes eventually sets in as the body become insulin exhausted, unable to produce more. Then the risk of blindness in old age, amputations, and other terrible consequences may become a reality.

The process of cells using the nutrients we consume is called cellular metabolism. Overloading cells with nutrients that we do not need such as excessive raw sugars (or even food that is easily converted by the body into glucose) and fats, results in obesity; your brain will become sluggish. Now, add extra oxygen and your body will certainly consume it. This is called oxidative metabolism. The added oxygen will push all cells into overdrive; repair cells use the added oxygen to work harder in an effort to restore their state prior to entering the hyperbaric chamber. Repair cells help get rid of the toxins, acids and, excess/unneeded sugars.

After people have been in a hyperbaric chamber, healthy adults comment on the fact that their brain is working more efficiently. Memory becomes vivid and problems are easier to solve. This state is commonly known as the "cappuccino" effect or "brain brightening."

As for small children with autism, almost all of them have a neuronal oxygen transport problem. So the additional feast of oxygen actually makes the good cells, the repair cells, work harder for just a little while after the treatment. This increase in oxidative metabolism dissipates after a couple of hours, but that little bit of improvement lasts: brain cells wire up into complex circuits, blood vessels within the brain become robust, and stem cells are produced. (More on that shortly.)

This is why all oxygen treatments need to take place over consecutive days. There is no point in attending one hyperbaric treatment and then not following it up for another week. Sort of like a seed sprouting from the earth. With each treatment, the positive changes begin to accumulate, and positive results are visible with the passing days.

Can oxygen therapy alone cure autism?

The short answer is no; it cannot. I make this statement based on the treatment of scores of challenged children that have come to my centers across America, and in other parts of the world. Something more is needed—an interdisciplinary approach.

How many treatments in the chamber does a challenged child need to be part of that recovery program?

As noted earlier, we treated our first challenged child, Emily, in one of my centers in the United States. Just one treatment per day at 2 ATA, the pediatrician's decision, made a vast difference to little Emily. Then along came a physician, Dr. Paul Harch, a very bright health scientist. He came to the conclusion 15 years ago that two 60-minute treatments each day at 1.75 ATA for 20 days could cure children with autism. He was so sure of himself that he even took out a patent on that protocol.

I was nonplussed and puzzled. I did not think that you could patent a natural phenomenon such as the effects of hyperbaric oxygen therapy or gravity. Of interest is the fact that the Supreme Court came to the same conclusion in 2012 with what is now known as the "Prometheus" ruling. In essence, the courts had taken all of the enforceable teeth out of Dr. Harch's protocol. This is not to say that there isn't some merit in the treatment pressure, nor is it to imply that Dr. Harch was ignorant of the science of hyperbaric rights in the therapy. So my centers took notice of the fact that he is a physician, and for a little while at least, we did use the Harch protocol on children.

While our little patients saw useful and positive results, there was never a normal recovery by any of the children in that 20-day double ride. This left a lot of the parents disappointed, some angry at what they felt were false claims about hyperbaric oxygen therapy, and rightfully so.

It is significant to note that Dr. Harch, and other clinicians, no longer use the two treatments a day protocol for challenged children, and he no longer claims that recovery is possible in just 20 days.

Now, onto our wonderful Jennifer; the ballet dancer and champion ice-skater. Her mom, Vickie, a bright spirit and highly intelligent woman (actually that describes the whole family), independently arrived at the conclusion that diet was a most important part of the normalization program. All together, over a 12-year period, Jennifer had 135 rides in the hyperbaric chamber. She ice skated at 2am (when the rink was clear), and focused on diet. In a nutshell, while the extra oxygen was feeding one side of cellular metabolism in spurts of 20 treatments at a time, Jennifer's cells were getting the correct diet of nutrients on the other. All of Jen's cells were giving off the right amount of heat and energy, particularly her repair cells.

Bernie and Kerri proved scientifically that the addition of proper nutrients, combined with the removal of deleterious/toxic foods is probably more important than hyperbaric oxygen therapy for children and young adults on the spectrum. Yes, the oxygen boost helps propel the patient to "normal," but the other nutritional changes are really the rocket fuel. It is a long journey, not all that expensive, and certainly worthwhile.

When a hyperbaric chamber is not a hyperbaric chamber.

One final comment about treating neuronal disorders at 1.75 ATA pressure is of utmost importance; this is that 100% oxygen is essential to the protocol. Since Dr. Harch arrived at this treatment depth, there have been thousands of inflatable chambers, so-called mild hyperbaric chambers, sold in the United States and around the world. These are not the same as hospital grade, hard-sided chambers. These bag-chambers do not deliver the needed 100% oxygen. In my opinion, at the very best, if you modify them (against FDA rules) you will only get 34% oxygen.

What have we learned thus far? Time for a review and then THE BIG SECRET REVEALED!

- Oxygen cannot be seen, but we can weigh it.
- Our bodies "eat" oxygen every day because it is a food for our cells.
- The only way we can get additional oxygen is in a hyperbaric chamber.
- This is because any gas becomes soluble and melts into a fluid (i.e. Henry's law); the same reason we can carbonate soda drinks.
- In a single treatment in a hyperbaric-oxygen-therapy chamber, we cannot overdose (and I have never witnessed any negative side effects).
- With the extra oxygen feeding the cells, our bodies get a boost of energy, particularly the repair cells.
- For oxygen in the chamber to be effective, it must be 100% oxygen. Inflatable chambers do not deliver 100% oxygen.
- Hyperbaric oxygen therapy alone will not cure autism, but it does make a big difference.
- There is no such thing as a "one-size fits all" treatment protocol in a hyperbaric chamber that works for all patients. For example, twice a day in the chamber for 20 days will make some difference but not necessarily "normalize" a challenged child.

Did you miss the big secret?

It was already mentioned about one page back—the discoveries of Bernie and Kerri. Let us repeat the secret for you and then push the edges of their really stunning discovery that is backed up by the latest science.

Bernie and Kerri proved scientifically that the addition of the proper nutrients, in the absence of offending foods, in addition to eliminating pathogens and parasites, are most likely more important than hyperbaric oxygen therapy for children and young adults. Yes, the oxygen boost helps propel the patient to "normal" but the other dietary interventions are really the rocket fuel. Not to mention freeing the body of offending pathogens/parasites which wreak havoc on the body. It is a long journey, not all that expensive and, certainly worthwhile.

Nutritional "padding." Hyperbaric oxygen therapy and the production of progenitor and stem cells.

When a baby is growing inside its mother, it has progenitor cells and stem cells. The function of these cells is considered to be, as yet, undecided. Although studies continue to elucidate exactly which mechanisms are involved in determining the fate of these cells, it is generally accepted that their surrounding cells will dictate their development. For example, the big arteries that crisscross the chest area make a decision sooner or later to turn into a heart with all of its compartments. Then, as development moves on different parts of the baby appear. These are called progenitor or stem cells.

There are various estimates on how many different types of cells the developed human has in total; however, it is predicted there are somewhere between 210 and 300. Mind you, when the baby takes its first breath of air, all but about four of these different cell groups stop reproducing. For example, the brain, with all of its neuronal pruning and circuitry needs about 20 year's worth of stimulation to properly wire all of the connections. Surprisingly, a lot of the neurons will be shed or "pruned" since they will not be put to use in the adult. However, neuroplasticity allows new skills to be added or repairs to be made to and by the brain in a lifetime.

However, medical scientists have now realized the great need for progenitor or stem cells (cells that can make decisions to replace worn or damaged cells within the body), and the search is on for these decision-making cells. Many scientists are looking for something they can patent and as a result are missing the point that it is already here. Or they are ignoring it, because but they cannot patent it.

It is called hyperbaric oxygen therapy.

> Penn Study Finds Hyperbaric Oxygen Treatments Mobilize Stem Cells
>
> **Science Daily** — According to a study to be published in the American Journal of Physiology-Heart and Circulation Physiology, a typical course of hyperbaric oxygen treatments increases by eight-fold the number of stem cells circulating in a patient's body. Stem cells, also called progenitor cells are crucial to injury repair. The study currently appears online and is scheduled for publication in the April 2006, edition of the American Journal.
>
> Stem cells exist in the bone marrow of human beings and animals and are capable of changing their nature to become part of many different organs and tissues. In response to injury, these cells move from the bone marrow to the injured sites, where they differentiate into cells that assist in the healing process. The movement, or mobilization, of stem cells can be triggered by a variety of stimuli including pharmaceutical agents and hyperbaric oxygen treatments. Where as drugs are associated with a host of side effects, hyperbaric oxygen treatments carry a significantly lower risk of such effects.
>
> "This is the safest way clinically to increase stem cell circulation, far safer than any of the pharmaceutical options," said Stephen Thom, MD, PhD, professor of emergency medicine at the University of Pennsylvania School of Medicine and lead author of the study. " This study provides information on the fundamental mechanisms for hyperbaric oxygen and offers a new theoretical therapeutic option for mobilizing stem cells."
>
> "We reproduced the observations from humans in animals in order to identify the mechanism for the hyperbaric oxygen effect," added Thom. "We found that hyperbaric oxygen mobilizes stem/progenitor cells because it increases synthesis of a molecule called nitric oxide in the bone marrow. This synthesis is thought to trigger enzymes that mediate stem/progenitor cell release."[1]

Nutritional padding: It has long been known that when hyperbaric oxygen therapy is applied, existing cells, the good ones, the repair cells, turn the body into its own pharmaceutical company, providing the right sort of chemicals at the right time. Above all, there is a dramatic increase in the production of the molecule called nitric oxide from the bone marrow. Probably the best way of describing the addition of hyperbaric oxygen therapy, as an adjunct to the correct nutritionals, is simply by putting fuel onto an already existing smoldering fire. Out of the glow of embers and smoke you will see a vigorous combustion as your fireplace heats up and the flames appear.

Now imagine other ingredients that have yet to be discovered as beneficial. For example, stem cells that repair the damaged brain of a person with autism. It could be that adding external stem cells to the recipe of recovery and normalization, with simple nutritional low-cost padding, and putting the little patient in the chamber would be a faster way of achieving what all parents want for their children—to eventually become fully functional and independent adults.

All of this by just a little bit of nutritional padding and helping the body to rid itself of pathogens/parasites.

Now, it is time to turn back to Kerri so she can tell you what she has seen when hyperbaric oxygen therapy is added to the correct nutritional padding, along with chlorine dioxide to keep pathogens under control. That is the big secret to helping your child through the challenges that living in this century have visited upon you.

* * *

As Bob mentioned, I have seen miracles happen when hyperbaric oxygen therapy was applied to a child who is ready. Two girls, in particular, lost their diagnosis completely within weeks of completing treatment in the chamber. As I mentioned before, hyperbaric oxygen therapy gave my own son his speech back. Before his first 40 dives he had several single-syllable sounds, but after, he began stringing words together to request his favorite foods.

Today we are seeing that one dive a day at 1.75 ATA (atmospheres absolute) for 90 minutes at depth is better than two 60-minute sessions were first prescribed for Patrick seven years ago. Every three months after the first treatment of 20 sessions, apply ten sessions of 90 minutes, each at 1.75 ATA, until your child is recovered.

A lot of parents ask when is a good time to do hyperbarics. To this I reply that we must first have diet and chlorine dioxide totally under control, and be at least three months into the Parasite Protocol. At that point it's time to start looking for a chamber.

Make sure to do an ATEC before your child's hyperbaric oxygen therapy sessions, and repeat two months after the sessions are over; the results will be obvious.

Healing
the symptoms known as
Autism

This is my Asperger son Tobias, 10 years old, and me... when we received the book this summer.

Chapter 12

Step 7 - GcMAF & Autism: State of the Art & Future Perspectives for a Natural Cure

by Marco Ruggiero, MD, PhD.
Specialist in Diagnostic Radiology
Full professor of Molecular Biology
Dept. of Experimental & Clinical Biomedical Sciences
University of Firenze, Italy

> *"The voyage of discovery is not in seeking new landscapes but in having new eyes."*
>
> ~ ***Marcel Proust***

My family and I have been in biomed for so long, that there aren't too many interventions we haven't already tried. A few years back, Dr. Usman mentioned to me that GcMAF might be something good for Patrick. She explained that GcMAF is a human protein that occurs naturally in a healthy body. It is a special supplement that replaces the missing part of the immune system, and also acts as the body's own internal medicine.

IMMUNE GCMA
NG/ML 2.2ML
TORE AT 4°C
TCH M1210-25
ETAINED VIAL

Based on her recommendation, we ordered a vial, and applied it for about two and a half months. I saw less obsessive compulsive eating, but when the vial ended and I hadn't seen amazing results, I wasn't extremely motivated to continue. Not only that, funds were scarce, so continuing really wasn't an option.

Two years went by and I spoke to Dr. Usman again. She said we really should give it four to five months before making a final decision. So, we scraped together the money for a second order. We started again in November of 2012, and reached full dose by December, 26th 2012. Seemingly, as time went on, Patrick started to do new things.

By springtime of 2013, Patrick's BFF's mom was saying how well he was doing, and reminded me that during the prior month Patrick had not been doing the things that he was now doing. As I write this in October of 2013, all of my money now goes to GcMAF. ☺

One of the things that I feel makes GcMAF so exciting is the fact that it helps regrow dendrites in a matter of hours after it is injected. The dendrites are the very things that got fried when mercury entered our children's bodies. With the help of GcMAF, we are now able to restore what was lost. Not to mention, GcMAF boosts the immune system, which we have come to find is often sorely lacking in people with ASD's.

Many people have asked me if GcMAF is now an official part of the protocol. When working towards recovery, I feel that we need to try everything at our disposal that has been proven to heal autism—in the correct order, or course. GcMAF fits that description. According to www.gcmaf.eu, "As of February 2013, on the American National Library of Medicine alone, 142 eminent scientists from 8 nations have published 59 major GcMAF research papers, which can be viewed on the US Government's Pubmed system." There is now enough research to prove that GcMAF has secured its place as a powerful biomedical intervention for healing autism. Therefore, after three months on the Kalcker Parasite Protocol, it is time to start thinking about adding in supplements, HBOT and GcMAF.

We are blessed to have the one and only Dr. Marco Ruggiero, the lead GcMAF researcher, write the remainder of this chapter about this substance, it's benefits and applications in autism and beyond.

The information he presents is extremely important because I feel that GcMAF is the future of medicine—today. What other intervention boasts a cancer recovery rate of 80%? None! Dr. Bradstreet is reporting 85% responders and 15% recovery rate from autism with GcMAF.[1]

Imagine adding GcMAF to the protocol already outlined in this book at the right moment (after working down some of the pathogen load), we should start seeing even higher rates of recovery than 15%. I look forward to seeing what transpires during the upcoming year as more families add GcMAF to their children's protocols. I am excited for more and more people to lose their autism diagnoses.

For the past 20 years, Dr. Marco Ruggiero has worked on the vitamin D axis, a metabolic pathway that includes the vitamin D binding protein-derived macrophage activating factor or GcMAF. In the past 3 years he has published studies on the immunotherapeutic effects of GcMAF in cancer, HIV/AIDS, chronic fatigue syndrome and neurological conditions.

Marco Ruggiero holds a PhD in Molecular Biology, is a certified Medical Doctor specializing in Clinical Radiology and is a professor of Molecular Biology at the Department of Experimental and Clinical Biomedical Sciences of the University of Firenze, Italy.

He served in the Army as Medical Lieutenant with specific training in chemical, biological and nuclear warfare. He worked at Burroughs Welcome Co. North Carolina, USA publishing a seminal paper with Nobel Laureate Sir John Vane and, subsequently, at the National Cancer Institute of the NIH in Bethesda, working with Dr. Stuart A. Aaronson and Dr. Peter Duesberg.

Since 1992, he holds the chair of Molecular Biology at the University of Firenze where he leads a research group of about 10 researchers. He has published more than 150 peer-reviewed scientific papers on signal transduction in a variety of experimental and spontaneous pathologic system related with cancer, chronic kidney disease, chronic fatigue syndrome and neurological conditions.

The role of nutrition in the treatment and eradication of the symptoms of autism is well established and other chapters of this book treat this topic in detail. What is far less known is the fact that certain natural components of the diet, certain nutrients, have profound effects on all the systems that are affected in autism and in particular on the brain and the immune system. Their effects are so dramatic that for about 20 years, these natural components have been labelled as "drugs" even though they are essential parts of our bodies, just like hormones or neurotransmitters. I am referring here to the family of molecules that is known as GcMAF, an acronym that stands for Gc-protein-derived Macrophage Activating Factor.

GcMAF has been the object of intense research in the past 20 years for its therapeutic role in cancer, autoimmune diseases, and HIV infection (for review, see: Anticancer Res. 2012 Jan;32(1):45-52). But, only one year ago, in 2012, its dramatic effects on the eradication of the symptoms of autism were published in a prestigious peer-reviewed scientific journal (Autism Insights 2012:4 31–38).

In this chapter I shall provide an overview on GcMAF with particular reference to its therapeutic role in autism and to the future perspectives deriving from the basic and applied research that is revolving around this fascinating molecule.

A Brief History of GcMAF

The first publication on GcMAF in a peer-reviewed journal indexed in the US National Library of Medicine is dated 1994 (J Immunol. 1994 May 15;152(10):5100-7). In this article, a group of researchers from the Department of Biochemistry at the Temple University School of Medicine in Philadelphia, USA, described the effects of a protein, defined as GcMAF, on rat macrophages. Macrophages are key elements of the immune response and the research group led by Dr. Nobuto Yamamoto, the first author of this paper, showed that GcMAF activated these cells. Given the central role of macrophages in the control of all immune responses, GcMAF was considered a powerful regulator of the immune system and the research on GcMAF initially was focused on its immune-stimulant properties. In fact, one year after the initial observation on rats, Dr. Yamamoto, who in the meantime had moved to the Laboratory of Cancer Immunology and Molecular Biology of the Albert Einstein Cancer Center in Philadelphia, authored a paper suggesting that a defect in endogenous GcMAF production contributed to immunodeficiency in AIDS patients (AIDS Res Hum Retroviruses. 1995 Nov;11(11):1373-8.). Consequently, he provided GcMAF to Medical Doctors treating HIV patients and, 14 years later, after 7 years of follow-up, he published a seminal paper describing eradication of HIV infection through GcMAF administration (J Med Virol. 2009 Jan;81(1):16-26).

Immunodeficiency, however, is not unique to HIV/AIDS and, back in 1996, Yamamoto and colleagues, hypothesized that deficient endogenous GcMAF production was responsible for the relative immunodeficiency that is typically observed in cancer patients (Cancer Res. 1996 Jun 15;56(12):2827-31). Quite logically, this observation of 18 years ago, led to the proposal of administering GcMAF to cancer patients with the goal of empowering the immune system that in turn would fight the cancer growth. This approach is known as "immunotherapy" and it is much older than GcMAF since, in modern times, it stems from the observation of Dr. William Coley in 1891. However, it could be argued that the idea of stimulating the immune system to fight diseases is even older since, as it is stated in this recent article, "…cancer regresses when associated with acute infections such as bacterial, viral, fungal, protozoal, etc. Acute infections are known to cure chronic diseases since the time of Hippocrates…" (Indian J Cancer. 2011 Apr-Jun;48(2):246-51).

Whatever the case, administration of GcMAF to patients with metastatic advanced cancer proved efficacious and there are now scores of papers published in the peer-reviewed literature describing the dramatic therapeutic effects of GcMAF in all types of experimental or spontaneous tumors.

Nowadays, the interest for GcMAF in the immunotherapy of cancer is so high that the prestigious scientific journal "OncoImmunology" dedicated the cover of its August 2013 issue to GcMAF. The molecular rendering of GcMAF is on the left of the cover with a caption linking it to an article by Thyer and colleagues published in the same issue and describing successful treatment of advanced cancer patients (Figure 1).

Figure 1

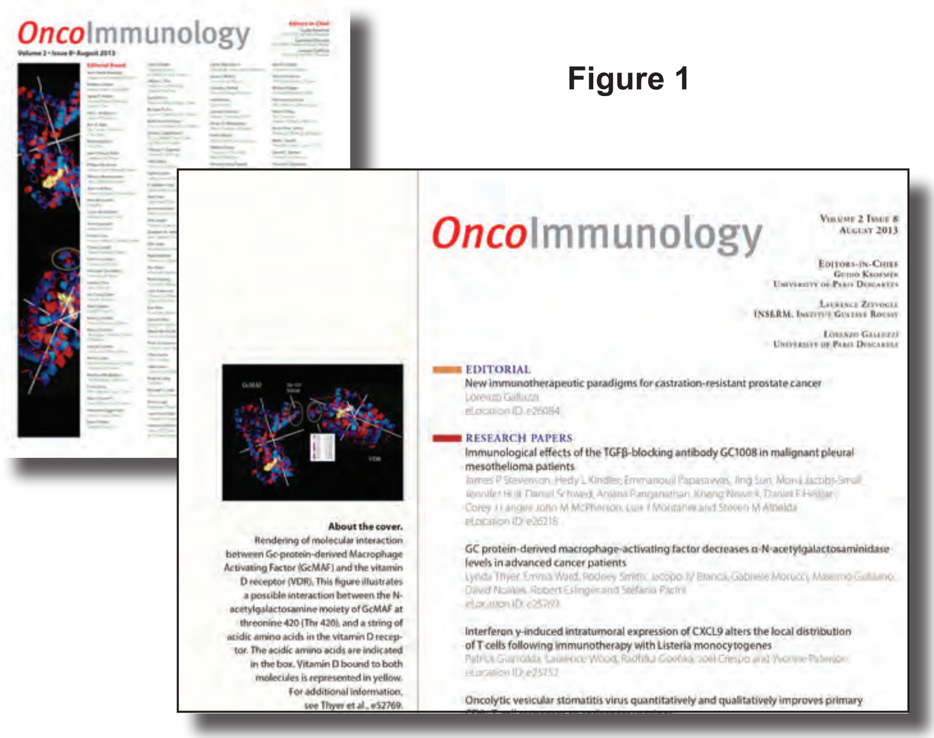

OncoImmunology

Volume 2 Issue 8
August 2013

Editors-in-Chief
Guido Kroemer
University of Paris Descartes

Laurence Zitvogel
INSERM, Institut Gustave Roussy

Lorenzo Galluzzi
University of Paris Descartes

EDITORIAL

New immunotherapeutic paradigms for castration-resistant prostate cancer
Lorenzo Galluzzi
eLocation ID: e26084

RESEARCH PAPERS

Immunological effects of the TGFβ-blocking antibody GC1008 in malignant pleural mesothelioma patients
James P Stevenson, Hedy L Kindler, Emmanouil Papasavvas, Jing Sun, Mona Jacobs-Small, Jennifer Hull, Daniel Schwed, Anjana Ranganathan, Khang Newick, Daniel F Heitjan, Corey J Langer, John M McPherson, Luis J Montaner and Steven M Albelda
eLocation ID: e26218

GC protein-derived macrophage-activating factor decreases α-N-acetylgalactosaminidase levels in advanced cancer patients
Lynda Thyer, Emma Ward, Rodney Smith, Jacopo JV Branca, Gabriele Morucci, Massimo Gulisano, David Noakes, Robert Eslinger and Stefania Pacini
eLocation ID: e25769

Interferon γ-induced intratumoral expression of CXCL9 alters the local distribution of T cells following immunotherapy with Listeria monocytogenes
Patrick Guirnalda, Laurence Wood, Radhika Goenka, Joel Crespo and Yvonne Paterson
eLocation ID: e25752

Oncolytic vesicular stomatitis virus quantitatively and qualitatively improves primary

About the cover.
Rendering of molecular interaction between Gc-protein-derived Macrophage Activating Factor (GcMAF) and the vitamin D receptor (VDR). This figure illustrates a possible interaction between the N-acetylgalactosamine moiety of GcMAF at threonine 420 (Thr 420), and a string of acidic amino acids in the vitamin D receptor. The acidic amino acids are indicated in the box. Vitamin D bound to both molecules is represented in yellow. For additional information, see Thyer et al., e52769.

Source: https://www.landesbioscience.com/journals/oncoimmunology/oncoimmunology_2-8.pdf

Consecration of GcMAF in the Olympus of science, however, occurred a few days later, on September 14, 2013, when another article by Thyer and colleagues was published in "Nutrients", an international, peer-reviewed open access advanced forum for studies related to human nutrition (Nutrients. 2013 Jul 8;5(7):2577-89). In fact, this article describing the effects of GcMAF on human breast cancer cells, was ranked in the top 5% of all articles ever tracked by Altmetric. The Altmetric score is a general measure of the attention that an article, book or dataset has received online and reflects the quantity of attention received as well as the quality of that attention. By the time that the GcMAF was ranked in the top 5% of all scientific articles ever tracked, Altmetric had analyzed 1,510,524 articles across all scientific journals. Therefore, we may safely state that in this chapter we are describing one of the hottest topics in today's world science (Figure 2).

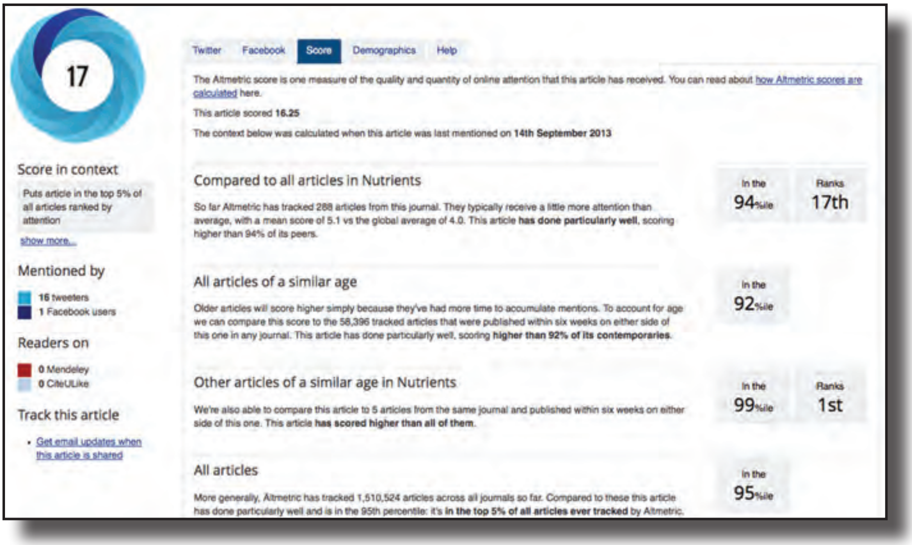

The Altmetric score is one measure of the quality and quantity of online attention that this article has received. You can read about how Altmetric scores are calculated here.

This article scored **16.25**

The context below was calculated when this article was last mentioned on **14th September 2013**

Score in context

Puts article in the top 5% of all articles ranked by attention

Mentioned by

Readers on

Track this article

- Get email updates when this article is shared

Compared to all articles in Nutrients

So far Altmetric has tracked 288 articles from this journal. They typically receive a little more attention than average, with a mean score of 5.1 vs the global average of 4.0. This article **has done particularly well**, scoring higher than 94% of its peers.

All articles of a similar age

Older articles will score higher simply because they've had more time to accumulate mentions. To account for age we can compare this score to the 58,396 tracked articles that were published within six weeks on either side of this one in any journal. This article has done particularly well, scoring **higher than 92% of its contemporaries**.

Other articles of a similar age in Nutrients

We're also able to compare this article to 5 articles from the same journal and published within six weeks on either side of this one. This article **has scored higher than all of them**.

All articles

More generally, Altmetric has tracked 1,510,524 articles across all journals so far. Compared to these this article has done particularly well and is in the 95th percentile: it's **in the top 5% of all articles ever tracked** by Altmetric.

Figure 2

The interest for the therapeutic effects of GcMAF in a variety of conditions derives from the early observation by Yamamoto and colleagues in 1995, who showed elevated level of serum alpha-N-acetylgalactosaminidase (Nagalase) in the serum of AIDS patients. Nagalase is an enzyme that deglycosylates (degrades) the precursor of GcMAF that is the circulating Gc-protein, also known as vitamin D-binding protein (AIDS Res Hum Retroviruses. 1995 Nov;11(11):1373-8). This results in the loss of GcMAF precursor activity and consequent dysfunction of the immune system. Increased Nagalase activity, however, is not observed uniquely in AIDS patients. For example it was demonstrated that the increase of Nagalase activity in the serum of cancer patients is due to the fact that cancer cells release Nagalase and, therefore, Nagalase activity reflects tumor burden, aggressiveness and progression of the disease. Consequently, determination of Nagalase activity is currently proposed as a reliable way of evaluation of cancer severity (Cancer Lett. 2009 Oct 8;283(2):222-9).

In short, in cancer, HIV/AIDS, and in many other chronic conditions, elevated serum Nagalase, degrades Gc-protein that is the precursor of GcMAF. Therefore, GcMAF is not produced inside the body to a sufficient extent and this leads to immunodeficiency. Serum Nagalase, however, does not degrade GcMAF, and this justifies the administration of exogenous GcMAF in all those conditions where its endogenous production is insufficient due to elevated Nagalase activity.

Until 2002, it was thought that GcMAF exerted its effects only on macrophages and that its therapeutic efficacy was due to stimulation of macrophages that in turn would attack and destroy cancer cells as well as virus-infected cells. In 2002, however, a novel effect of GcMAF was described. Kanda and colleagues, working at the Department of Molecular Microbiology and Immunology, Division of Endothelial Cell Biology, Nagasaki University Graduate School of Medicine, Nagasaki, Japan, published a paper, co-authored by Yamamoto, describing the inhibitory effects of GcMAF on angiogenesis (J Natl Cancer Inst. 2002 Sep 4;94(17):1311-9.). Angiogenesis is the process of building new blood vessels that provide the tumor mass with the required supply of blood and nutrients that it needs to grow and metastatize. Therefore, since 1971 it was postulated that inhibition of angiogenesis could be a successful strategy to deprive the growing tumor mass of its supply of blood and nutrients (N Engl J Med. 1971 Nov 18;285(21):1182-6). While there are scores of inhibitors of angiogenesis, some of which in use in the therapy of cancer, GcMAF offers the advantage of performing more than one anticancer effects; in fact, not only it inhibits tumor-induced angiogenesis (Cancer Immunol Immunother. 2011 Apr;60(4):479-85), but it also stimulates macrophages that attack and destroy the cancer cells (Nutrients. 2013 Jul 8;5(7):2577-89).

If these combined anticancer effects were not enough, in 2010 Gregory and colleagues, working at the Department of Ophthalmology and Visual Sciences of University of Kentucky in Lexington, Kentucky, USA, demonstrated that GcMAF directly inhibited proliferation and metastatic potential of human prostate cancer cells (PLoS One. 2010 Oct 18;5(10):e13428). Slightly more than one year later, we were able to demonstrate that GcMAF not only inhibited proliferation and metastatic potential of human breast cancer cells, but it also reverted their neoplastic phenotype; in other words, cancer cells treated with GcMAF became normal (Anticancer Res. 2012 Jan;32(1):45-52).

On the basis of the peer-reviewed scientific literature quoted above, we can now state that GcMAF exerts multiple effects that are responsible for its efficacy in anticancer therapy:

1. It stimulates macrophages that attack and destroy cancer cells, as originally postulated by Yamamoto and colleagues.

2. It inhibits tumor-induced angiogenesis thus depriving the growing tumor mass of blood and nutrients.

3. It directly inhibits cancer cell proliferation and metastatic potential and it retro-transforms cancer cells into normal healthy cells.

Given this plethora of effects on cancer, it is no wonder that the majority of GcMAF research has been focused on its anticancer properties. There are, however, a few research articles suggesting that GcMAF might exert also other biological actions with therapeutic implications. In addition to the studies on HIV and AIDS quoted above, GcMAF has received attention for its potential role in the therapy of bone disorders (Blood. 1996 Oct 15;88(8):2898-905), and autoimmune diseases such as Lupus Erythematosus (Clin Immunol Immunopathol. 1997 Mar;82(3):290-8).

Despite these scanty evidences on the possible role of GcMAF in conditions other than cancer and AIDS, it was not until 2012, however, that GcMAF proved dramatically effective in the therapy of autism.

GcMAF & Autism

GcMAF, being derived from Gc-protein, also known as vitamin D binding protein, belongs to the so called “vitamin D axis”, a nutritional metabolic pathway described in detail in a recent review (European Nephrology, 2011;5(1):15–9). My research group had been working on the vitamin D axis since 1996 (Epidemiol Prev. 1996 Apr-Sep;20(2-3):140-1), that is since the beginning of GcMAF research. Just like our colleagues in the USA and Japan, we were mainly interested in the role of the vitamin D axis in bone metabolism and cancer (Radiol Med. 1996 Nov;92(5):520-4. Oncol Res. 1998;10(1):43-6). Therefore, it came as a surprise when Dr. Nobuto Yamamoto first mentioned to me the effects of GcMAF on neurodegenerative diseases such as Parkinson’s and Alzheimer’s diseases.

In the hot summer of 2010 we visited Dr. Yamamoto in his hometown of Philadelphia, and we had the honor and the pleasure to spend two very intense days with this old-style exquisite gentleman. Dr. Yamamoto gave us a detailed history not only of his decade-long research activity, but also of the entire field of immunotherapy, starting with the pioneering work of Dr. William Coley at the end of the nineteen century. During these intense and fruitful talks, Dr. Yamamoto mentioned the dramatic effects that he had personally observed by administering GcMAF to a patient suffering from Parkinson’s disease. According to Dr. Yamamoto, the effects had been almost immediate and dramatic up to the point that he himself was amazed. Regrettably, he had been unable to clinically document the case and, therefore, this remained an anecdote. From this anecdote, however, an entire new field of GcMAF research was about to rise; that is the study on the role and implications of GcMAF in treating neurologic and neurodegenerative disorders.

The first solid evidences for a role of GcMAF in treating a neurodegenerative disorder were simultaneously presented by Dr. Paul Cheney and Prof. Kenny de Meirleir at the meeting of the International Association for Chronic Fatigue Syndrome and Myalgic Encephalomyelitis (CFS/ME) held in Ottawa in September 2011. A few days later, in collaboration with Dr. Paul Cheney, we presented in Padova, Italy, a study demonstrating that food-based GcMAF dramatically improved the symptoms of CFS/ME, and this work was published in the official journal of the Italian Society of Anatomy and Histology, one of the oldest and most respected European scientific societies, founded at the beginning of the twentieth century (It. J. Anat. Embryol. Vol. 116, No 1, 2011).

CFS/ME is a complex disorder that shares many features of autism. In fact, it is characterized by immune system dysfunction, widespread inflammation, and multi-systemic neuropathology (J Intern Med. 2011 Oct;270(4):327-38. In Vivo. 2013 Mar-Apr;27(2):177-87). Dysfunction of the immune system involves abnormal functions and distributions of T lymphocytes, B lymphocytes, natural killer cells, and monocyte/macrophages (Nihon Rinsho. 1992 Nov;50(11):2625-9. Brain Behav Immun. 2012 Jan;26(1):24-31). The etiology of CFS/ME, just like the etiology of autism, is still unknown and multiple factors are thought to be responsible for its onset and progression, thus lending credit to the hypothesis that both etiology and pathogenesis are multifactorial (J Intern Med. 2011 Oct;270(4):327-38). Heavy metal exposure and viral infections are among the factors that contribute to CFS/ME etiology and pathogenesis, and a role for human endogenous retroviruses has been recently hypothesized (In Vivo. 2013 Mar-Apr;27(2):177-87). Both chronic heavy metal exposure and viral infections are considered responsible for the immune system dysfunction and neuropathology that are typical of CFS/ME (Med Hypotheses. 2012 Sep;79(3):403-7. In Vivo. 2013 Mar-Apr;27(2):177-87).

In addition to these areas of overlapping, it had been noticed by Doctors treating autism and/or CFS/ME, that in several families the two conditions are often associated with parents of autistic children suffering from CFS/ME. Interestingly, this happens also in families with adopted children with autism, thus lending credit to the hypothesis that environmental or infectious factors may contribute to both conditions.

Whatever the case, it had been observed that CFS/ME patients had elevated levels of serum Nagalase together with a number of immunological alterations, and it was hypothesized that they could benefit from GcMAF treatment. The results presented in Ottawa and in Padova confirmed this assumption and ever since immunotherapy with GcMAF is a stronghold of CFS/ME treatment.

At about the same time, Dr. Bradstreet, a leading authority in the field of autism, began to study the levels of Nagalase in autistic children and their responses to GcMAF treatment. The results obtained in a first cohort of 40 subjects were reported in a seminal paper published in 2012 (Autism Insights 2012:4 31–38). This paper represents a turning point in the understanding of autism pathogenesis and treatment.

The first significant finding is the observation that the average level of Nagalase in autistic subjects was more than two folds higher than the normal level, thus lending credit to the hypothesis that immunodeficiency or immune system dysfunction is a hallmark of autism. The second finding, probably the most important in recent years, is the observation that GcMAF treatment significantly improved the symptoms of autism and this improvement was associated with a decrease of Nagalase levels.

The therapeutic response to GcMAF was statistically significant and showed the typical bell-shaped curve that is expected when a biologically effective treatment is at work. About 15% of subject did not respond; this means that their symptoms were not ameliorated by GcMAF treatment. However, more than 15% of children responded so dramatically that all the symptoms of autism were eradicated in a matter of weeks up to the point that the children could no longer be defined as "autistic." According to the very word of the authors, in the children in this group, "This response was demonstrated at school, during therapies, home and outside the home as substantial improvement to the point that many or most of the criteria of autism were no longer present."

The remaining 70% of children were distributed according to the so-called Gaussian distribution in the groups showing "slight," "moderate," or "considerable" improvement.

This paper demonstrated for the first time that GcMAF treatment was able to eradicate the symptoms of autism in a matter of weeks, thus reinforcing the idea brought forward in many chapters of this book that autism is a curable condition indeed.

Quite obviously, the dramatic results obtained by Dr. Bradstreet did not go unnoticed and, in April 2013, he presented further results at the First GcMAF Immunology Conference held in Frankfurt, Germany. According to Dr. Bradstreet, as the time of follow-up and the number of subjects increased, a shift toward the right (that is the "best") part of the curve could be observed. This means that the percentage of those showing "considerable" to "very considerable" improvement increased.

From the point of view of the general practitioner, the issue could be considered closed: GcMAF cures a significant percentage of autistic children now and that figure is rising.

However, since I am a researcher in addition of being a Medical Doctor, the successes of Dr. Bradstreet in treating autism and the research papers published by Immuno Biotech laboratories are in the middle of an exciting series of observations that are leading to a completely new understanding of how GcMAF works and what autism truly is.

In fact, the undisputable efficacy of GcMAF in treating autism raises a number of questions on the mechanisms involved at the cellular and molecular levels. The most important questions are:

1. The efficacy of GcMAF in curing autism is due only to its immune-stimulating effect?
2. Or, is GcMAF acting directly also on neurons as it does on cancer cells?
3. Does GcMAF revert the neuro-anatomical alterations that are typical of the autistic brain?
4. Why a certain number of autistic subjects do not respond to GcMAF?
5. Can we devise strategies to have these non-responders respond?
6. Can we devise strategies to further improve this already dramatic therapeutic effect?

The Effects of GcMAF on Human Neurons

GcMAF is a protein and, as such, it was thought that it would stimulate cells from the outside. In other words, until 2013, it was conjectured that, on their external surface, cells had a receptor that bound GcMAF and conveyed the signal to the inside of the cell and, ultimately, to the nucleus and the DNA, thereby modifying the cell behavior. However, despite 20 years of research, such a receptor for GcMAF had not been identified. And this might appear truly odd considering that the entire human genome had been sequenced, that is studied in every single detail, more than 10 years ago.

In our first publication with Dr. Yamamoto, presented at the XVIII International AIDS Conference in Vienna in 2010, we had demonstrated that the response to GcMAF in human mononuclear cells (that is macrophages) was dependent on polymorphisms (that are the individual variations among humans) of the vitamin D receptor (VDR) gene. This observation was not surprising considering that vitamin D, its receptor, and GcMAF all belong to the vitamin D axis.

In order to further investigate this relationship between GcMAF and the VDR we began studying the molecular assembly of GcMAF that is the "shape" that the molecule assumes in its physiological conformation. Having done this, we looked for complementary areas in the GcMAF and VDR molecules and… voilà! The two molecules complemented each other as two elements of a puzzle.

We published this observation in the "Nutrients" paper that is now in the top 5% of all articles ever tracked, and the picture of the two molecules interacting with each other is in the front cover of the OncoImmunology August 2013 issue. The molecular mechanisms through which GcMAF and VDR interact and the consequent complex web of intracellular signaling are topics for specialists and rather difficult to divulge; the basic point, however, is that we had demonstrated how is it possible that GcMAF exerted so many different effects that can be exploited to treat a number of diseases so different from each other.

From this observation we deducted that all cells having the VDR would have responded to GcMAF and, given our interest for the effects of GcMAF in neurological conditions such as autism and CFS/ME, we began our study on human neurons. In fact, human neurons express (i.e. "have") the VDR and therefore, they are candidate to be stimulated by GcMAF.

We cultured human neurons in Petri dishes and we challenged them with ultra-pure GcMAF; in this type of experiments there were no other variables. That is, there were only neurons and GcMAF; the cells of the immune system or other molecules were not involved. By doing so, we would have been certain that whatever effect we observed, this was due uniquely to GcMAF.

As I have shown at the congress "Curando el Autismo" in San Juan de Puerto Rico in August 2013, GcMAF at extremely low concentration significantly increased neuronal cell viability and metabolic activity. These effects on cell viability were associated with dose-dependent intracellular cAMP production (cAMP is an intracellular second messenger). This means that GcMAF increased the production of energy inside human neurons and made them more active, more viable. Increased viability and metabolic activity following GcMAF stimulation as well as cAMP formation were accompanied by morphological changes, meaning that stimulated neurons changed their shape.

In the absence of GcMAF, human neuronal cells under the microscope appeared as small, relatively undifferentiated cells with large nuclei. After 24h stimulation with GcMAF, neurons showed a significant change in morphology that was consistent with the induction of neuronal differentiation and increased connectivity. The cytoplasm was enlarged and cytoplasmic elongations could be observed. After 72h incubation with GcMAF, these morphological changes were more evident and well differentiated cells could be observed.

After incubation with GcMAF, the cells appeared to establish contacts with each other. Taken together these results indicated that GcMAF increased neuronal cell viability, metabolic activity and differentiation, with the first effects being observed at 24h. It is worth noticing that the assay that we used to determine cell viability, measured mitochondrial activity, and mitochondrial dysfunction is known to be hallmark of ME/CFS (Int J Clin Exp Med. 2012;5(3):208-20), and possibly of autism as well. Therefore, the effects of GcMAF that we observed explain the reason why GcMAF is so effective in CFS/ME and autism; it counteracts the molecular alterations that are the pathogenetic basis of both conditions.

I wish to spend a few words here on the effects of GcMAF on neuronal connectivity. It is well assessed that this is one of the basic alterations in the brain of autistic subjects; neurons do not establish contacts with each other and therefore the neurological signal cannot be transmitted from one area of the brain to another thus causing a "disconnection" (Rev Neurol. 2012

Feb 29;54 Suppl 1:S31-9). GcMAF re-establishes neuronal connectivity by stimulating the formation of what are called "cytoplasmic elongations" that are like small "arms" that protrude from the neurons and establish contact with each other as if neurons were tending arms to "touch" each other and exchange information.

In Figure 3A, human un-stimulated human neurons in culture can be observed; they are small and roundish. After stimulation with GcMAF (Figure 3B), they are much larger because increased energy production leads to increased protein synthesis in their cytoplasm. But, perhaps more important, they have a highly irregular, elongated, triangular, shape with spikes that touch each other. The areas of contact between the two neurons can be clearly observed in the central part where they establish connecting bridges. Please notice that both pictures were taken with the same magnification.

Figure 3

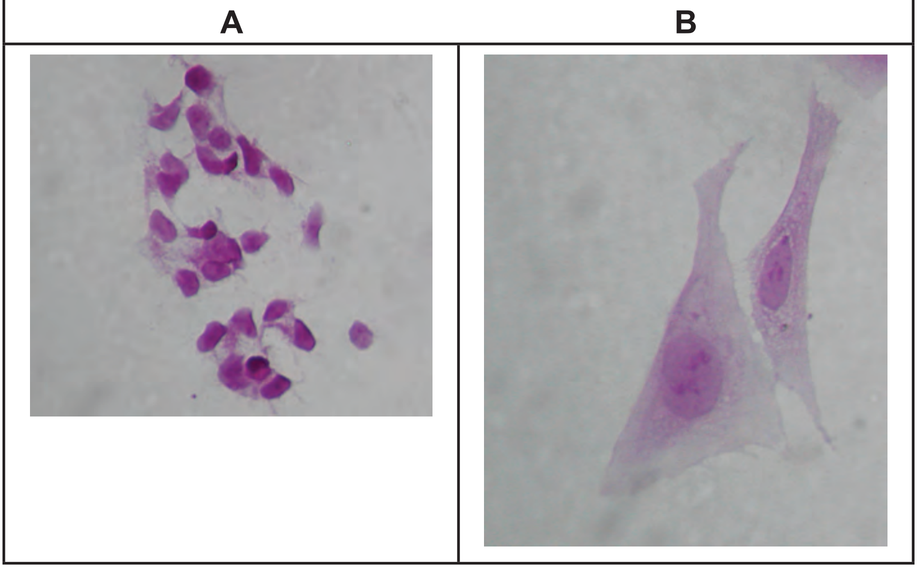

These observations provide an answer to the first two questions raised by the article published by Dr. Bradstreet: the therapeutic effects of GcMAF in autism can be ascribed to at least two different actions of GcMAF. On one side, it rebalances the immune system that is typically dysfunctional in autistic subjects; on the other side, it directly stimulates human neurons increasing their metabolic activity and their connectivity. But, as it always happens in research, these "answers" led to other "questions": does GcMAF cross the blood brain barrier? That is, does it arrive to the neurons of the central nervous system? Before challenging this question further, however, let's proceed with order and let's try to provide an answer to the other questions that were raised by the Bradstreet's paper.

GcMAF and the Neuro-Anatomical Alterations of the Autistic Brain

If the etiology of autism is still undefined, the pathogenesis and the brain alterations associated with autism are better understood. Several studies using sophisticated imaging techniques demonstrated that there is atypical structural brain connectivity within cortical white and grey matter and it is postulated that such an alteration of brain connectivity may represent one of the principal neural substrates underlying autistic symptoms (PLoS One. 2013 Jun 18;8(6):e67329. Proc Natl Acad Sci U S A. 2013 Jul 22.).

The observed atypical brain connectivity is localized to particular areas of the brain and the alterations in the fronto-temporal regions are associated with the severity of social and repetitive symptoms of autism (Proc Natl Acad Sci U S A. 2013 Jul 22.). Atypical brain connectivity in autism is also associated with reduction of grey matter volume in different areas of the brain (Cerebellum. 2013 Apr 10.). In the bilateral superior temporal gyrus significantly decreased functional connectivity was accompanied by the strongest trend of grey matter volume decrease (PLoS One. 2013 Jun 18;8(6):e67329), thus reinforcing the hypothesis that study of grey matter volume may provide indication of functional connectivity.

However, the studies on brain connectivity and grey matter reduction quoted above were conducted with sophisticated Magnetic Resonance Imaging (MRI) equipments using complex softwares. Because of their cost and complexity, these techniques are not suitable for routine examination and this appears to be in contrast with the need to provide the general practitioners with the instruments to study autism. In fact, given the current epidemiological trend, it is highly probable that even primary care providers will see with increased frequency a patient with autism spectrum disorder in their clinic (Curr Probl Pediatr Adolesc Health Care. 2013 Jan;43(1):2-11). This consideration prompted us for the search of accessible methods to study brain alterations in autism and to study the response to different treatment including GcMAF.

To this end, we developed an easily accessible method to study the brain cortex using transcranial ultrasonography. This technique was first applied to the diagnosis and treatment CFS/ME a syndrome that, as we know, share neuropathological similarities with autism (Toxicology. 2008 May 2;247(1):61-72). In particular, we proposed to use ultrasounds to study the grey matter in the temporal lobe, and to evaluate signs of neuronal layer disorganization and brain inflammation (Med Hypotheses. 2012 Sep;79(3):403-7). It is worth noticing that these same neurological alterations, i.e. minicolumn structural changes and inflammation are frequently observed in autism (Front Immunol. 2013 Jun 10;4:140. doi: 10.3389/fimmu.2013.00140. Print 2013). Therefore, we hypothesize that ultrasounds could be used in autism in order to study the brain alterations that have been described using more sophisticated and less accessible techniques.

First of all, we assessed the safety of the technique; to this end we used sub-thermal ultrasounds. These are considered inherently safe and have been used for fetal imaging in utero, and virtually every part of the body, including brains of newborn babies through fontanelles. Their inherent safety and any lack of correlation with the onset of autism have been confirmed by meta-analysis of thousands of cases. A recent randomized controlled trial on over 1.400 cases demonstrated beyond any reasonable doubt that there is no link between the prenatal ultrasound scans and the autism phenotype (J Autism Dev Disord. 2012 Dec;42(12):2693-701).

Then, in a collaborative work with Dr. Bradstreet, we decided to focus our attention on the temporal cortex in order to compare our results with those already published by Hameroff and colleagues and by ourselves using this technique (Med Hypotheses. 2012 Sep;79(3):403-7. Brain Stimul. 2013 May;6(3):409-15). In addition, having to choose a particular area of the brain for this initial study, the temporal lobe was particularly interesting because of its accessibility to ultrasounds and its involvement in the pathogenesis of autism (PLoS One. 2013 Jun 18;8(6):e67329. Print 2013).

In the typical adults subject, meninges appeared as a well-organized array of layers of about 2.6 mm thickness. The meninges were separated from the cortex by a structure that shows echogenic and morphological feature consistent with the sub-arachnoid space. The cortex of the temporal lobe appeared as a well-organized array of alternate, hyper-echogenic/hypo-echogenic, layers. The thickness of the cortex was 5.0 mm. This value of thickness and the anatomical positioning of the probe led us to hypothesize that we were observing the temporal areas designated by von Economo as TG and TE, i.e. those areas involved in the control of eye movements and balance in standing position (area TE), social behavior, mood and decision making (area TG) (von Economo C. Temporal lobe. In: Triarhou LC, editor. Cellular structure of the human cerebral cortex. Basel; 2009. p. 114-132). It is worth noticing that the layers within the column-like structure of the healthy subject appeared well-organized and parallel to each other.

However, the images obtained in a representative autistic subject showed that the thickness of the cortex of the temporal lobe was significantly reduced, as it had been observed with MRI (PLoS One. 2013 Jun 18;8(6):e67329. Print 2013). Other images showed a concomitant increase in the spinal fluid in the sub-arachnoid space that was twice as much as in the healthy subject. These findings are consistent with the very recent observation that infants who developed autism spectrum disorder had significantly greater extra-axial fluid characterized by excessive cerebrospinal fluid in the sub-arachnoid space (Brain. 2013 Jul 9.).

Taken together, the results obtained in autistic subjects demonstrate that our technique was able to visualize the excess extra-axial fluid in a manner comparable to MRI; however, at variance with MRI, the observation with ultrasounds was immediate and intuitive and could be easily performed with no discomfort during a routine examination.

Another finding that we observed was a relative disorganization of the neuronal layer arrangement in the temporal cortex of autistic subjects. It was possible to observe that the neuronal layers had a different orientation and that there were black areas of about 0.4 mm that could represent areas of poor connectivity.

Since this technique has been applied only recently to the study of autism, there is no statistical analysis indicating whether GcMAF or any other effective treatment can reverse the anatomical alterations that we have observed in the brain of autistic subjects. Therefore, there is no sound scientific answer to this question; there are only some anecdotic evidences suggesting that this might indeed be the case. Autistic subjects treated with GcMAF and showing considerable clinical improvement showed a normal brain anatomy. Future observation will tell whether these anecdotic evidences have a statistical meaning.

Ultrasounds, however, in addition to their role in early diagnosis and follow-up, might also have a role in the therapy of autism and in particular in improving the response of those subjects who do not appear to benefit from GcMAF treatment.

GcMAF Non-Responders

As odd as it may appear to non-researchers, the presence of non-responders is often the best assurance that we are dealing with a real biological phenomenon and not with placebo effects or worse. In fact the bell-shaped curve presented in the article by Bradstreet and colleagues it typical of any effective treatment and it confirms the validity of the results. This having said, however, there is the need to provide an answer to those who do not respond. And, in the effort of providing an answer to non-responders, improvements in the overall efficacy of the treatment for everyone will eventually follow.

At the moment, there are two hypotheses that could help explaining why certain subjects do not respond to GcMAF in terms of improvement of symptoms of autism.

The first hypothesis is related with the VDR gene polymorphisms. As we proposed at the AIDS conference in Vienna in 2010, and later published in 2012 (J Nephrol. 2012 Jul-Aug;25(4):577-81), the individual responsiveness to GcMAF depends on the individual variations (polymorphisms) in the VDR gene. In other words, there are subjects harbouring variants that are associated with little or no response. However, this lack of response is relative and not absolute. In fact, it has been observed with vitamin D that these individuals respond to higher doses of the vitamin, thus indicating that the VDR is "less" responsive, but still functioning. This would mean that the so-called non-responders might need a higher dose or a longer time of GcMAF treatment to obtain the desired therapeutic effects. Prolonged observation of a higher number of cases will tell whether this hypothesis is correct.

The second hypothesis is closely associated with the first one as well as with the question briefly outlined in a preceding section: "does GcMAF cross the blood brain barrier?" In a recently submitted paper, we proposed that the interconnection between GcMAF and VDR signaling is responsible for the transport of GcMAF through the blood brain barrier. In fact, transport of macromolecules across the blood brain barrier requires both specific and nonspecific interactions between macromolecules and proteins/receptors expressed on the luminal and/or the abluminal surfaces of the brain capillary endothelial cells (Int J Cell Biol. 2013;2013:703545. doi: 10.1155/2013/703545). Since VDR signaling appears to enhance brain to blood transport at the blood brain barrier through both genomic and non-genomic actions, it can be hypothesized that the interaction between GcMAF and VDR (Nutrients. 2013 Jul 8;5(7):2577-89. doi: 10.3390/nu5072577) may favor GcMAF transport into the brain. This means that subjects with a less responsive VDR because of unfavorable genetic variants, would also have a less efficient transport of GcMAF to the brain.

This latter consideration, however, opens the way to a possible solution that might also help increasing the effectiveness of GcMAF in those who already respond to it.

In fact, now that we know the physiological molecular assembly of GcMAF, we can design specific GcMAF molecules that are more efficient in crossing the blood brain barrier and, therefore, are more effective in treating neurological conditions such as CFS/ME and autism. These molecules exploit the fatty acid binding site in the GcMAF molecule, and use one of the most beneficial fatty acids, oleic acid, the basic component of olive oil, to target brain cells.

Most interestingly, GcMAF/oleic acid molecules do not only stimulate human neurons with a much higher efficiency than GcMAF, but they also regulate glial cell metabolism. Glial cells are the connective tissue of the brain and glial cell dysfunction is involved in the pathogenesis of autism. In particular, glial cell dysfunction seems to be responsible for the neuronal underconnectivity typical of the autistic brain (Neuron Glia Biol. 2011 May;7(2-4):205-13). In short, we can postulate that the development of these novel GcMAF/oleic acid molecules will prove beneficial not only for the GcMAF non-responders, but also for improving the therapeutic efficacy of GcMAF in all neurological conditions where GcMAF treatment is indicated.

In addition to these novel, more physiological, brain-targeted, molecules, there are other strategies that can improve GcMAF transport into the brain with consequent increased therapeutic efficacy. Just a few days ago, in a journal dedicated to advanced radiological research, it was published a report demonstrating that ultrasounds, just like those described in the preceding section, transiently open the blood brain barrier. This phenomenon is exploited to favor transport of chemotherapeutic drugs into the brain of individual with brain tumors and it proved safe and efficacious after extensive testing.

On the basis of this observation, one could devise a simple treatment protocol using the same ultrasound system that is currently used for transcranial ultrasonography. In analogy with what is performed to deliver chemotherapeutic drugs to brain tumors, ultrasounds could be used to facilitate the transport of GcMAF through the blood brain barrier and to focus its delivery into those areas of the brain that are known to be altered in autism.

In this manner, during a regular examination, the health professional could exploit our technique to assess the neuro-anatomical alterations in the autistic brain and to deliver GcMAF to those areas. The use of more physiological, brain-targeted GcMAF molecules that can be administered sublingually, would greatly improve the efficacy of the treatment and it will most likely reduce the percentage of the so called non-responders to GcMAF.

As noted above, however, the use of ultrasounds to treat neurological conditions is not completely new. In fact, we and others have demonstrated that non-thermal ultrasounds have significant effects on mental states (Brain Stimul. 2013 May;6(3):409-15), modulate neurotransmission (Neuropsychobiology. 2012;65(3):153-60), cardiac frequency and blood pressure and improve skeletal muscle strength (The Journal of IiME, vol. 6 (1), p 23-28, 2012). These effects can be observed with ultrasounds alone, that is in the absence of other treatments.

In a protocol for the treatment of autism and CFS/ME, we propose to use our ultrasonographic technique together with timely administration of sublingual GcMAF/oleic acid molecules. In this manner, we are able to exploit the direct effects of ultrasounds on the neuronal microtubules (Brain Stimul. 2013 May;6(3):409-15), together with the ability of ultrasounds to favor the transport of GcMAF/oleic acid into the brain and, more precisely into those areas where the neuroanatomical alterations are observed.

What is GcMAF?
A drug, a hormone, a natural component, a nutrient?

How to define a molecule that is formed every time a wound is licked? This is the most evolutionary conserved protective behavior in animals, humans included. The enzymes in the saliva convert the Gc-protein in the blood that is on the wound into active GcMAF that prevents infection of the wound and promotes repair (PLoS One. 2013 Jul 18;8(7):e69059).

For many years researchers considered GcMAF only a key regulator of the immune system because of its activity on macrophages and its denomination reflects this early conjecture. However, as it often happens in science, the original denomination of a molecule proves incorrect with the passing of time and the accumulation of scientific evidences. Vitamin D is a notable example; it is not an vitamin and it does not contain any amine group (even though the suffix “vita” applies, in that it is indispensable for life). Now it is rightly considered a hormone produced by the body and, to be more precise, a secosteroid hormone that has a mechanism of action similar to that of classical steroidal hormones. GcMAF, being a part of the vitamin D axis might soon become another example of scientific mislabeling. In fact, we now know that it exerts its actions in many other cell types in addition to macrophages and that its role go well beyond activation of macrophages.

And even though there are no doubts that it is a powerful stimulant of the immune system, it would not be surprising to realize that this might not be its most relevant physiological role. In this respect, GcMAF shares the destiny of its precursor, the Gc-protein. Gc-protein is a serum alpha 2 glycoprotein composed of a single polypeptide chain, structurally related to serum albumin, that performs a variety of actions in several cell types. It is synthesized in the liver and has been detected on the surface of several cell types, yolk sac endodermal cells, and some T lymphocytes. In B cells, it participates in the linkage of surface immunoglobulins. In addition to vitamin D, it binds actin and acts as an actin scavenger. These characteristics are of paramount importance in human pathology since actin is the most abundant protein in eukaryotic cells and it is a major cellular protein released during cell necrosis that may cause fatal formation of actin-containing thrombi in the circulation if the actin scavenging capacity of Gc-protein is exceeded. Not surprisingly, therefore, several studies suggest an association between Gc-protein and resistance or susceptibility to chronic obstructive pulmonary disease, thyroid diseases, diabetes, multiple sclerosis, and sarcoidosis (for a review on Gc-protein and the vitamin D axis, see: European Nephrology, 2011;5(1):15–9).

From these considerations it derives that the potential indications for GcMAF treatment might be far too numerous than those currently envisaged. And also this parallels the fate of vitamin D, once thought only to prevent rickets and now involved in almost all aspects of human physiology and pathology up to the point that a (serious) scientific article on this subject was entitled: “Does vitamin D make the world go ‘round’?” (Breastfeed Med 2008; 3: 2392–50.).

Apart from issues of biomedical terminology, it is now emerging the concept that GcMAF is an essential component of that nutritional pathway that is the vitamin D axis. It is not a case or a coincidence that the article on GcMAF that has received the greatest attention was published in a scientific journal named “Nutrients”. With the proceeding of research it is becoming apparent that GcMAF participates in such a variety of physiological events that each proposed definition is limited.

But, if we accept the concept that GcMAF is a natural nutrient, indispensable for brain and immune system development and function (and possibly for many more essential physiological events), two questions spontaneously arise: “Where do I find it in nature?” and “How can I consume it?”

The second question already has an answer: now that the physiological molecular assembly of GcMAF is known, there are on the market several GcMAF products that can be used sublingually or simply eaten. Modern biotechnology has reproduced the physiological conformation of GcMAF so that it can be absorbed in a natural way without recurring to intravenous or intramuscular injections. GcMAF injections are reserved for those cases where a strong and rapid stimulation is required.

The other question is more intriguing and leads to the final section of this chapter.

My Own Personal Experience with GcMAF

As it happens in many people of my age exposed to the radiations from the Chernobyl incident, I have several thyroid nodules. Being a radiologist, it was easy for me to perform ultrasonography on my thyroid and what I saw on the screen was not nice. In addition to small lesions in my right lobe, in my left lobe I observed a massive mixed (solid and cystic) protruding nodule substituting for most of the lobe. The anterior-posterior diameter was about 3 cm, whereas the longitudinal diameter was greater than the length of the probe (4 cm). The structure was grossly inhomogeneous with cystic areas and hypo-reflecting areas alternated with iso- or hyper-reflecting areas. The margins bordering the residual normal thyroid tissue appeared well defined with an area of edema (inflammation) surrounding the nodule. The blood vessels were located around and inside the nodule and few or no blood vessels could be observed in the large hypo-reflecting areas suggesting the presence of metabolically inactive tumor tissue. Consistent with these findings, laboratory tests showed signs of immunodeficiency and anemia. And, it has been reported that anemia is among the predictive elements of malignancy in thyroid nodules (Tunis Med. 2002 Sep;80(9):536-41).

It was immediately clear to me that I needed a powerful stimulant of the immune system as wells as an anticancer agent, and GcMAF was my first thought. However, at that time, GcMAF was synthesized only in academic research laboratories and it was not available on the market as the pure, highly active, quality controlled product that we know today. Therefore, we decided to look for natural sources of GcMAF.

We already knew that GcMAF is naturally formed when a wound is licked because saliva contains the enzymes that convert blood Gc-protein into active GcMAF. Licking my thyroid, however, quite obviously, was not an option.

Because of these hindrances, we hypothesized that GcMAF could be obtained through a natural food fermentation process that exploits the characteristic of certain microbial strains to produce the enzymes that convert food Gc-protein into GcMAF (Hollis and Draper, 1979). According to our hypothesis, fermented food-derived GcMAF would activate the macrophages widely diffused in the walls of the entire gastrointestinal tract and, in particular, in the Waldeyer's-Pirogov tonsillar ring (or pharyngeal lymphoid ring). Our first goal was to assess the feasibility of obtaining natural GcMAF through home-made food processing procedures. To this end, we devised a strategy to prepare a fermented food product using those microbes, known to mankind for their healthy properties for millennia, that express the enzymes involved in GcMAF synthesis (Hollis and Draper, 1979).

After having prepared this fermented food at home, I began consuming it on a daily basis.

Since at that time we too considered GcMAF mainly for its immune-stimulating properties, the first objective was to assess the efficacy of this procedure on certain markers of immunodeficiency, the so called CD4 lymphocytes. The general public became familiar with CD4 lymphocytes because of their involvement in HIV infection and AIDS, and CD4 cell number is considered a reliable marker of immune system function. The role of CD4 cells in the immune system has been widely popularized up to the point that the webpage of the AIDS Authority of the Government of the United States of America states; "CD4 cells or T-cells are the "generals" of the human immune system. These are the cells that send signals to activate your body's immune response when they detect "intruders," like viruses or bacteria. Because of the important role these cells play in how your body fights off infections, it's important to keep their numbers up in the normal ranges..."

In our observation, the reference laboratory considered normal value for CD4 cell count 500 - 1500 cells/μL. Before consumption, my CD4 count was low, close to the threshold of AIDS definition (372 cells/μL). CD4 count rose from the abnormal value of 372 to the normal value of 609 cells/μL after three weeks of consumption of the fermented food that I had prepared. Such an increase continued up to eight weeks with CD4 cell count reaching 853 cells/μL. After many years, they are still in the normal range.

The trend toward normalization of blood parameters was observed also in other abnormal haematological values. My values (low or abnormal) changed as follows:

	Before Consumption	**After 3 Weeks**	**Normal Values**
Red Blood Cells			
(x106/uL)	4.55	4.85	4.6 - 6.2
Haematocrit, %	42.9	45.9	40 - 52
Haemoglobin, g/dL	14.8	15.6	13 - 18
Platelets/uL	165,000	206,000	140,000 - 440,000

The neutrophil cell count decreased from 77% to 55% in about 12 weeks, indicating a normalization of parameters suggestive of a pro-inflammatory state. Overall, the observed changes in haematological parameters were consistent with those positive modification of immune responses in humans that might represent the potential mechanism by which food fermented by probiotic bacteria confer health benefits.

Being a researcher, I am very well aware that this report is a mere anecdote that, in the era of evidence-based medicine, has low value or is not even considered to be medical evidence. However, a very recent study on the evaluation of clinical practice challenges the deprecation of stories like that reported above. Thus, it is well known that some studies present large and impressive statistics obtained from many observations while others report a small number of noteworthy events, as we did in this study. However, according to this authoritative epistemological approach, "all of these stories become evidence of what works in medicine" (J Eval Clin Pract. 2011 Oct;17(5):920-6).

Conclusions

In August 2013, a paper published in the very prestigious American Journal of Immunology reported a series of clinical cases successfully treated with GcMAF (Am. J. Immunol., 9: 78-84. 2013). Among these, there were pathologies already known to be responsive to GcMAF treatment, such as cancer, autism, and CFS/ME. Other neurodegenerative diseases such as multiple sclerosis and amyotrophic lateral sclerosis, however, had not been treated with GcMAF before, and this is the first report on such a successful treatment. This paper, together with all the other recognitions reported above, places GcMAF as one of the most effective tool to fight diseases that were considered incurable until very recently.

In other chapters of this book it is clearly stated that autism IS curable, and the successes of GcMAF in curing autism further strengthen this concept. What is emerging and what I hope to have been able to convey, is that GcMAF is not an exogenous drug that "fixes" some problem, but rather a very natural component of our body that is essential for, at least, brain and immune system development and function. In many conditions, including autism, GcMAF is not produced to a sufficient extent and therefore it has to be administered from the outside. With the advent of a courageous little biotech company that has perfected the synthesis of natural, fully physiological, GcMAF molecules that can be eaten, GcMAF treatment will become all the more and flexible to fit the individual needs. Eventually, GcMAF will become part of an integrated nutritional plan as it was intended by nature during evolution.

Consistent with this concept, I wish to conclude with an example from the realm of nature.

According to the Genetic Science Learning Center of the University of Utah, "Some mother rats spend a lot of time licking, grooming and nursing their pups. Others seem to ignore their pups. Highly nurtured rat pups tend to grow up to be calm adults, while rat pups who receive little nurturing tend to grow up to be anxious. It turns out that the difference between a calm and an anxious rat is not genetic—it's epigenetic. The nurturing behavior of a mother rat during the first week of life shapes her pups' epigenomes. And, the epigenetic pattern that mom establishes tends to stay put, even after the pups become adults."

It seems that licking their pups while breastfeeding in the first week of life is the decisive factor; in the saliva there are the enzymes that convert milk Gc-protein into GcMAF. What is more natural than this?

* * *

Subcutaneous GcMAF has been available through FIRSTiMMUNE for several years. Their website is:

www.gcmaf.eu

They now also produce a sublingual version of GcMAF called GOleic. They also have Bravo probiotics, and more. Please visit their website for more information.

"If you're not stubborn, you'll give up on experiments too soon. And if you're not flexible, you'll pound your head against the wall and you won't see a different solution to a problem you're trying to solve."

~ Jeff Bezos
Founder of Amazon

Chapter 13

Beyond Recovery: The Maintenance Plan

"Success is not final, failure is not fatal:
it is the courage to continue that counts."
~ Winston Churchill

We see these questions on our boards from time to time:

Will my son have to take CD for the rest of his life?
or
Will we be sorting through poop forever?

The answer is No!

Once your child has recovered, you can move to the maintenance phase. Since we live in a toxic world, it is important to continually detox the system, and keep the body free of parasites. While this can look different for every child and every family, the following are some general guidelines designed to keep our kids healthy once they are recovered.

Diet

While some families go back to giving their children foods that aren't permitted here, many do not. Wheat, dairy, and soy aren't "healthy" foods for any of us to be consuming. After reading Olive Kaiser's section on gluten, you may have a tough time giving your child any bread products ever again. That said, once the gut is healed, it should be able to tolerate many of these foods again, to the extent that a "healthy" person would.

CD

Even into the recovered stage of autism, many of the families keep using CD (oral, baths, enemas) to reach an ATEC of zero... or until there are no behaviors left from the autism. Once that is accomplished, then a maintenance plan of one oral dose on Monday and one oral dose on Thursday is enough. Typically, such a dose is 3 drops for children and 6 drops for adults—all in one dose on each of those 2 days. However, each family does this a little differently depending on weight and what the optimum dose was while on the full protocol.

The more I learn about how toxic our planet is (i.e. GMOs, toxins in the air, drinking water, parasites in the environment, etc.), the more I realize how critical it is to be on a at least a limited long-term detoxing program. This is especially true for someone who was, or may still be immunocompromised. I do believe that once a child recovers from autism, their immune system can do the job that it was designed to do. However, healing takes time, and every case is different. Some people ask me if recovered children still take CD once recovered—and to be honest—many do not.

Parasite Protocol

It is impossible to avoid parasites in our world. As Andreas and Miriam explained to us, their larvae are nearly indestructible in nature and can live for years. A strong immune system, as well as sufficient stomach acid, may prevent a parasitic infection from taking hold in the body. To make sure, it is best to do a "mini protocol" every 3 months. You can do the full 18 days every 3rd month or a 1-week mini version—whichever you prefer.

Other Supplements

While ocean water is necessary when detoxing with CD, it is extremely beneficial to all of us as our soils are almost completely devoid of nutrients and are contaminated with heavy metals and toxins. Also, it has been theorized that many GMO crops are lacking in nutrition as well. The 90 bioavailable nutrients in OW can certainly go a long way in keeping our bodies healthy. I especially love that it helps with conductivity in the brain. And, it is conducive to speech and cognitive function as well.

Depending on each individual's needs it may behoove a family to keep giving MCT oil or Vegan 3,6,9 omegas. Certain supplements may still be required if there is a history of seizures.

Chelation

Gentle chelators like Bio-Chelate™ can continue to be used into the recovery phase as heavy metals are present in our air and water.

Hyperbarics

Nothing brings down inflammation quite like a hard hyperbaric chamber. However, 10 to 20 sessions can be very cost prohibitive for some families, making it a very personal decision as to whether or not to continue with HBOT after recovery.

GcMAF

I never felt better than when I did a course of GcMAF injections on myself. Energy went up and so did my ability to concentrate. Again, GcMAF is a fairly expensive intervention and that will factor into each family's decision as to whether to keep using it. or not

In a nutshell, I would at least continue with CD two times a week, and a mini parasite protocol every 3 months, just as a preventative. Staying healthy is as important as getting healthy.

> Um, YEAH so just took an ATEC. We have officially been on the protocol for 12 months now and finishing up our 11th PP. ATEC of THREE!!!!!!!!!!!!!! 57 to 3 in 12 months. Henry's speech is what is keeping us from 0. He has great conversation with people that are patient and don't interrupt him so we are so close to a BIG ZERO!!! 57 to 3 in 12 months!!!! Far cry from the severe autism diagnosis he got at 2.5 years old. Here is to the doctor that told me my son should be medicated, would never speak, and will probably need to be institutionalized. I hope YOU never say those things to another person!!!!! All you need is hope, determination, and LOVE.

Healin
the symptoms kno
Auti
Edition
l Rivera
mberly McD

Chapter 14

Miscellaneous Information You Should Know

"Always bear in mind that your own resolution to succeed
is more important than any one thing."
~ Abraham Lincoln

Seizures

An estimated 40% of children with autism suffer from some form of seizures. Seizures can be absolutely terrifying for both child and family and deleterious to the health of a child who is already suffering from various issues. I have talked to many different professionals about seizures and in turn have heard just as many different opinions on why this happens. Many doctors are unsure, while some simply don't express their opinions. Some say it is a virus in the brain or heavy metals, inflammation, or pathogens/parasites. In a presentation by Dr. Andreas Kalcker, he stated that high levels of ammonia in the body can cause seizures and the high levels of ammonia in the body come from parasites. Since we know that CD kills pathogens/parasites, neutralizes heavy metals, and helps to reduce inflammation, it is not surprising that we have noticed reduced, or even eliminated seizure activity in children with seizure disorders that use CD in combination with the Parasite Protocol.

MCT (medium chain triglyceride) oils, which come from the Ketogenic Diet out of UCLA, are also a very important piece to healing the body and eliminating seizure disorders. There is plenty of information on the Internet that discusses the results of people (on the spectrum and otherwise) using MCT oils for seizure disorders.

The following is a list of things that parents on The Protocol have used successfully to reduce/eliminate seizures in their children.

- CD
- Parasite Protocol
- MCT oil

- GABA
- Hyperbarics at 1.75 ATA
- Ketogenic diet
- SCD diet (grain free)
- L-Carnitine
- P5P with magnesium
- L-Carnosine
- Taurine
- DMG
- Pycnogenol®

The first four in this list have proven to be the most important and have shown consistent results. If you have a child suffering from seizures, you would be wise to review and consider these interventions immediately. The others have also proven very helpful and can be added in one-by-one, (always leaving three days between each new supplement) while you watch for a decrease in seizures.

When we introduce GABA, we increase the dose daily in search of the appropriate dose for the individual. While establishing the correct dose for one intervention it is not appropriate to add in other supplements. Watch out for sleepiness or undesired behavior with GABA. In these cases, I would personally discontinue its use, as with everything there is not one formula that is right for every child. Similarly, if you see hyperactivity or undesired behavior with DMG discontinue its use.

All of the above supplements/interventions have proven very important for eliminating seizures, in different combinations, for different families. Whether the seizures are grand mal (tonic clonic) or petit mal (absence), they start to fade away as a result of the Protocol.

Sometimes, when a family starts corresponding with me, their child may already be taking some of these supplements, and they continue to do that. They may have already taken some of these and seen no improvement. Some other families live in parts of the world where these supplements are unavailable, and in that case we focus on CD and killing parasites.

As I mentioned, a critical piece for some of the children with seizures has been the implementation of the Parasite Protocol. This has also led to a reduction or cessation of seizure activity in many children with seizures.

Interestingly, once some moms were made aware of the parasite/seizure connection it was easy to time seizure activity/anxiety/aggression to full and/or new moons. As Dr. Andreas Kalcker talked about in the parasite section, the full moon period is when parasites are mating, many in the gastrointestinal tract, and therefore "wormy behaviors" (anxiety, SIBs, depression, tantrums, aggression, and teeth grinding to name a few) will generally be higher around that period. New moons can also cause an increase in these behaviors.

Children who have a seizure disorder or are self-injurious need to avoid food coloring, eggs, tomatoes, pork, and chocolate. When adding in the supplements for the seizures that we listed above, follow the same rules as before, never adding more than one supplement every three days, and always carefully observing and keeping track of any and all changes in your notebook so we know what causes what.

After starting anti-parasitics with the help of her daughter's doctor, one veteran mom introduced the CD protocol; a combination that has been extremely beneficial for her daughter, who suffered from frequent grand mal (tonic clonic) seizures.

Her Advice:

Most of my understanding of how parasitic infections can lead to seizures involves my own tracking of seizure activity patterns in my daughter, which coincided with new moon and full moon periods. Then, since MOST doctors want references to what they feel are 'reliable sources,' I started by Googling "NIH parasites seizures," which brought me to:

ncbi.nlm.nih.gov/pubmed/18754958
Helminthic parasites and seizures. [Epilepsy. 2008][1] *- PubMed - NCBI*

I would strongly suggest that you document patterns of seizure activity against the Farmer's Almanac of Full Moons (you can even look up past Full Moons) and marry that information with info from sites like the one above. Also, I would highly recommend you hunt down a pediatric infectious disease specialist in your local area (start by researching pediatric staff at local hospitals) and bring with you all information/insight to your first appointment.

Her Testimonial:

From her very first drop of activated CD in early July 2012 (actually her very first one-eighth drop to be exact), I witnessed an improvement in my daughter. Within about two hours of drinking her first ounce of the activated CD water solution (one activated drop of CD mixed into eight ounces of water), I witnessed her sitting still in our kitchen and observing me in such a way unlike she had ever done before. She was calmly observing me instead of running around the kitchen. She was taking my actions in. She was studying me. Her eyes were focused on me, rather than constantly scanning everything else in the room. Both her facial expression and her body language were calmer than I had ever witnessed before. She was more focused, more centered. She appeared content.

This change that I witnessed in her was so noticeable and so sudden, that I couldn't help but think to myself, "So this must be why the first word in MMS is Miracle." It was that amazing of a change in those moments.

Shortly after those incredible moments, my daughter went on to eating contently and then jumping on the neighbor's netted trampoline. She has enjoyed the neighbor's trampoline for years; yet on that particular day, only a few hours after her first drink of a tiny amount of CD, I noticed something quite different in her behavior. As I observed her from the windows of our house, I could see that she was jumping less than usual and instead looking around a lot more. She was calmly looking around at the neighbors' houses, down at some birds in the grass, up at the trees, the clouds, etc. It was all as though she was perhaps truly noticing them for the first time. She had glanced at these things before that day, sure, but she never before appeared to be studying them in the manner she was in those moments on the trampoline that July afternoon. She did engage a bit in some of her tic-like spitting movements with her mouth while on the trampoline, but to nowhere near the intensity, nor duration that would more typically have consumed her prior to that day.

When she came back into the house after using the trampoline and observing her world, I asked her, "why were you spitting out there?" To my surprise, she then very calmly replied, "because it's sensory." Wow! She truly was more connected somehow to her world than ever before. Something good was happening to my girl—only hours after drinking one-eighth of a drop of activated CD with one ounce of water. From those moments on I felt encouraged in my heart and soul that CD could possibly be a very important catalyst toward improving my daughter's health and well-being.

With the help of a caring local doctor, we had already begun some trials of anti-parasitic medications on our daughter in early May 2012, prior to starting CD, in response to behavioral struggles that had been consistently intensifying around full moons every month. (Later, we noticed new moon behavioral intensities as well). We had seen some improvement with anti-parasitic medication, yet CD seemed to somehow boost our daughter toward further improvements that much quicker. In fact, our daughter's ATEC score before starting CD measured 79 on July 8th, 2012 (the day before starting CD), but as of September 30th had dropped down to 55 (only two and a half months later).

Now, in early December 2012, our daughter still appears to be on an overall upward trend of improving health and stability, attributable to the synergy of CD and anti-parasitic medications. Natalie's interpersonal skills have improved, her growth rate has improved, her focus has improved, and her frequency of seizures has lessened.

In fact, at this writing she is 53 days seizure free, which is quite an improvement considering before she was on a terrifying pace of grand mal (tonic clonic) seizures occurring every 2-3 weeks. These are very dangerous, life-threatening events. Her seizures were rapid onset, heading toward grand mal within seconds. They could happen anywhere, i.e. while in her car seat coming back home from a Dr.'s appointment, while driving away from Costco, while running in the backyard playing with the dog, or swinging high on the swing set in the backyard with a neighbor girl. Yet another occurred while wading in the lake (with life vest on), another hit while in our neighbor's four-foot swimming pool with me and her brother. She was right next to me, walking around and then suddenly, she was in the throes of a rapid-onset grand mal seizure that required me to lift her out of pool into her NT (neurotypical) brother's arms. May these be the last my daughter ever knows of seizures.

It is our hope that, in time, CD will be able to take on the lead role as the primary health-support for our daughter, if not, perhaps even someday the sole healthcare support to her. We're looking forward to the days when we may be able to remove or reduce many, or perhaps even almost all, of her current supportive medications (i.e. antibiotic, antiviral, anti-seizure, etc.). With the caring support of parents of The Thinking Moms' Revolution, as well as some excellent local doctors, we suspect that our daughter's life will continue to improve. In the almost 14 years of caring for our daughter in her various states of medical fragility, we have never ever had this much Hope. We feel fortunate to have been swept up into this Revolution and embraced warmly by so many Thinking Moms who are also in the process of working toward more healing for their children. Working to heal illness is so much nobler than working to mask illness, wouldn't you agree? God Bless Jim Humble, Andreas Kalcker, Kerri McDaniel de Rivera and all the TMR families!

UPDATE: January 30th, 2013

I too am so thankful for Kerri and her whole 'crew'. Thanks to Kerri for bringing CD to all of our kids on the spectrum worldwide. My Natalie is now OFF of her former roller-coaster ride of prescription anti-fungals, OFF of her former roller-coaster ride of prescription antibiotics, OFF of prescription antiviral, OFF of prescription low-dose naltrexone, etc. Her belly distention has disappeared, she's gained weight, she's gained height and she's now almost 103 days seizure free (as of 6:16pm tonight). She still has a ways to go toward more improvements, but she's happily climbing up from bottom of the trenches; she's climbing up from what seemed like the very bottom of the abyss that vaccine injury/injuries had placed her in. Thank you Kerri, Andreas Kalcker, Jim Humble and all of the moderators on the supportive websites!!!

UPDATE: February 1st, 2013

My Natalie's long run of 105 seizure-free days has ended. After being off of Albenza® less than 24 hours, BOOM! a grand mal (tonic clonic) seizure occurred yesterday evening. This same thing happened the last time we stopped Albenza®, sooooo we're not sure if it's withdrawal related or parasitic activity related. We're going to closely watch her behaviors this week for signs of parasitic burden. We put her on Alinia® now, just to give her some sort of parasitic coverage while we wait to likely restart Albenza®, soon. She did fine on Alinia® 200mg BID in the past. I'll keep you all posted, but for now we unfortunately need to restart the clock on my Natalie's 'seizure-free days' tracking. I'm bummed, but thought I should keep you in the loop on this; as I've been soooooo excited over the past three months about her freedom from seizures, which I believe was due to the synergy of CD and Albenza®. I'll touch back if/when we learn more.

UPDATE February 2nd, 2013

O.K. so, I just wanted to kind of round out this long thread a bit with my thoughts, for whatever they're worth: I really truly suspect that perhaps too sharp of a drop off (lack of weaning) from the systemic med, Albenza® lead to the seizure. Natalie is stable right now on a low-dose interim of Alinia® 200mg BID for now until all insurance approvals are in for the delivery of more Albenza®. This leads me to believe that the several courses of Albenza® over the past months likely HAS lowered her overall parasitic load/burden fairly well thus far (it's a pretty broad-spectrum anti-parasitic, from what I understand). However, I don't think for a *minute that Natalie's broad spectrum anti-parasitic medicine needs are finished. For this reason, we'll resume Albenza® as soon as we're able.*

Although further discussion/negotiation between myself and Natalie's pediatric infectious disease doctor is required, my thought is that she may need an entire year on Albenza. However, certainly in the future, whenever we stop Albenza®, I'm going to plan for a slow weaning off period so as not to disrupt Nat's delicate nervous system. Also, I'm kind of thinking that perhaps if our kids 'wig out' when CD is eventually dropped that maybe we've not lowered the overall parasitic burden enough, as yet to allow their immune system to 'kick in gear' and properly function??? I don't know, obviously, just theorizing with the brain matter God has given me.

SUMMARY & UPDATE FOR SECOND EDITION: December 1st, 2013

From birth to age 7 years, my daughter Natalie unfortunately had her share of health and developmental struggles. Her diagnoses began with 'Failure to Thrive' shortly after birth, followed thereafter by a Mitochondrial Encephalopathy Partial Complex III Deficiency diagnosis by age 7 months, then Hypothyroidism as well as global developmental and cognitive delays obvious from there. Despite significant medical and developmental challenges, Natalie entered a Kindergarten designed for the 'mild cognitively impaired' (formerly referred to as educatably mentally impaired 'EMI') with daily opportunities for mainstreaming with the typical Kindergarten classroom next door. After a fall 2005 Flu shot early in that same Kindergarten year, however, all bets were off. She regressed very rapidly in every way: medically, developmentally, and behaviorally. Within months of that shot, a laundry list of additional diagnoses began to form: Sensory Integration Disorder, Autism, Epilepsy, P.A.N.D.A.S., P.A.N.S., and Combined Immune Deficiency, to name a few. After much study by a large team of educational specialists, Natalie was reclassified by school professionals by the following April of 2006 to the 'SXI" category of disabilities (Severely Multiply Impaired), having deteriorated terribly from her former classifications of POHI (Physically or Otherwise Health Impaired) and 'CI' (Cognitively Impaired) within only months of that Flu shot. Her significant healthcare needs and behavioral challenges since have necessitated assignment of a one-on-one Para-educator to her case, even while she's been schooled within high adult-to-student ratio basic education special- needs classrooms over the past several years.

In 2011, between the ages of 12 and 13, I began to notice patterns of apparent P.A.N.S. flares in Natalie, which seemed to coincide with the New Moon and Full Moon weeks. When I began to study this phenomenon a bit more closely, I also recognized a pattern of seizure activity that tracked these timeframes as well. In reading about potential parasite activity that is thought to link to the New Moon and Full Moon timeframes, I requested anti-parasitic medications for Natalie. Thankfully, a local Pediatric Infectious Disease doctor kindly accommodated my request. Each time that the doctor prescribed various anti-parasitic medications for Natalie, she would improve. The improvements were noticeable both at school and at home. It was incredible to witness. However, each and every time that she was off of such medications, she would regress back to her former state of illness (displayed primarily by significant behavioral regressions). So, we found that we needed to try to keep her on various anti-parasitic medications more often than off of them, in order for her to enjoy overall improved health and neuropsychiatric stability. Even so, there was still some sort of physiological strain that she was very *definitely experiencing, but we were unable to decipher what it could be. This was frustrating, of course; to witness her continued rollercoaster of ups and downs in health and her various psychiatric disturbances each month.*

It wasn't until a friend suggested we get Natalie started on CD, that the health and behavioral 'rollercoaster ride' began to show signs of smoothing out. In July 2012, after her first drink of CD (only one-eighth of a drop), Natalie began to show initial signs of relief from whatever it was that was still ailing her. Her connectedness and overall psychological state of being changed within hours of that first one-eighth drop of CD. After that first day, we of course then continued with a very slow and conservative ramp-up of CD (orally, then baths, then enemas) alongside various anti-parasitic medications and all of the components of the Kalcker PP (with frequent swap out of Mebendazole with Albendazole, predominantly). With each passing month over the past sixteen months, we've been able to witness improvements in Natalie; perhaps a bit more slowly than some of the younger and/or potentially less health-compromised kids that Kerri's protocol has healed, but improvements nonetheless. Before CD, Natalie's ATEC score was a 79. As of this writing, her ATEC score is a 39. Even more astounding to us than that improved ATEC score have been the changes witnessed thus far in her overall health status and health regimen. Here is a list of some of the improvements that I can very easily recall as I write this (there are likely many more that I could comment upon, if I had more time):

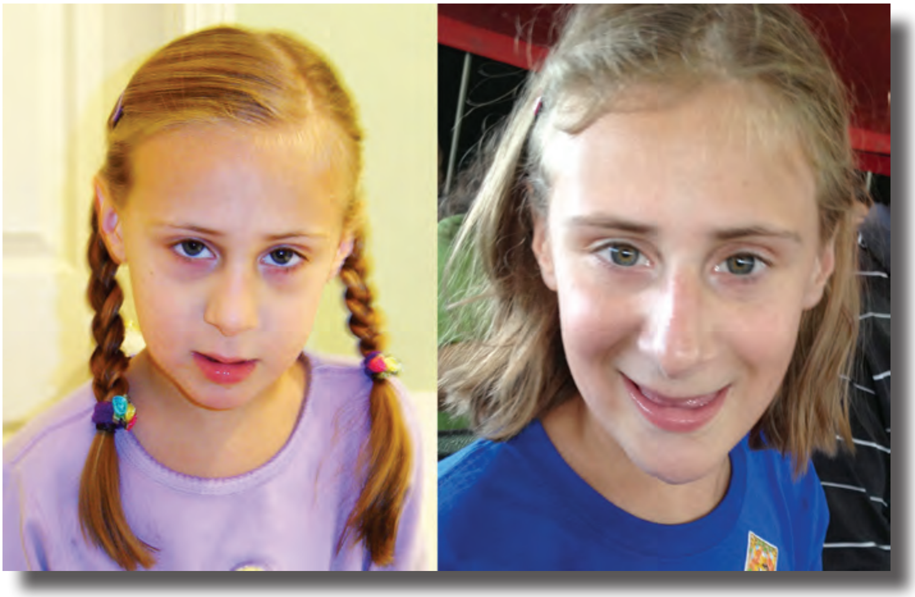
Natalie in 2006 (left) and in 2013 (right)

1. *Daily prescription anti-fungal medications no longer needed.*
2. *Daily prescription antibiotic medications no longer needed.*
3. *Daily prescription antiviral medication no longer needed.*
4. *Daily prescription LDN (low dose naltrexone) immune modulator removed.*

5. *Daily prescription Leucovorin removed without ill effects noticed, as well as several other daily supplements that might have been ultimately not helpful.*
6. *Prescribed Immune Globulin (weekly subcutaneous, formerly monthly intravenous for years) removed.*
7. *Distended belly gone & improved posture (no longer hunched over).*
8. *Formerly very frequent tongue sores/lesions lessening more & more.*
9. *Seizures formerly as high as every 17-21 days in her worst times; now she's enjoyed seizure-free stretches as high as 105 days (3.5 months) followed shortly thereafter by a seizure-free stretch of 168 days (5.5 months). Prescription anti-seizure medication reductions have also been enjoyed as well as fewer headaches resultant of those reductions.*
10. *P.A.N.S. flares controlled with the combination of CD and anti-parasitic meds. Neuropsychiatric symptoms of flares very dramatically declined.*
11. *Overall appearance of improved health, improved language and interpersonal skills, improved attentiveness (per Neurology's opinion at recent appointment).*
12. *Improved stamina and strength (core, legs, etc.) noticed at Special Needs Performance Cheerleading practices, physical therapies, etc. Her ability to peddle adaptive bike has also improved.*
13. *No more nose bleeds, no more ear infections, no more apparent seasonal allergies, fewer headaches (as noted above).*
14. *Incredibly improved sleep and need for naps only once in a while now (as opposed to formerly daily).*
15. *Improved growth rate and normalized appetite.*
16. *Improved language (more descriptive words, more sentences).*
17. *Healthier-looking skin overall; softer, less dry, improved color & ability to tan.*
18. *Improving oral muscular abilities, able to swallow pills/capsules.*
19. *Capable of successful dentist appointments and now even Orthodontist appointments, with calm and ease.*
20. *Increased success rate during therapies (Speech Therapy, Occupational Therapy, Physical Therapy, etc.).*
21. *Less rigidity overall, fewer transition issues, cooperates more often than she ever used to.*

With all of the well-documented improvements noticed in Natalie at home, at school and in doctors' offices, I don't have the same level of fear as other parents might surrounding our open use of CD, as Natalie's doctors are all well informed and quite supportive. We've also formally notified her social worker (both verbally at a home meeting and in writing) of our use of Kerri's CD Protocol (including Kalcker's Parasite Protocol and of course a GFCFSF/low sugar diet) while under Natalie's local medical team's support and medical monitor. We're 'out of the closet', so to speak, when it comes to our use of CD for improving Natalie's health. It's been quite obvious to Natalie's doctors; actually, that Kerri's protocol has been extremely complementary to their medical efforts for improving Natalie's health and overall quality of life.

There's of course no way to predict with certainty how far my sweet Natalie will be able to ultimately improve via this protocol, but based upon what we've witnessed thus far, I do feel that her current best hope for continued improvement lies within the pages of this book; as these healing strategies of Kerri Rivera's have brought Natalie toward more health (physical, neuropsychiatric, emotional and even some initial cognitive improvements) than any other medical effort attempted since her 2005 Flu shot injury. All other medical efforts for Natalie over the years, I feel, perhaps only temporarily patched her severe illness. Kerri's methods actually work to attack disease, kill it and remove it. Yes, that's right. Eight years after Natalie's tragic rapid spiral downward, it's now finally a time of healing for Natalie. Thank you, Kerri Rivera, for all of your help given to our girl and to so many children around the world.

Self-Injurious Behaviors (SIBs)

Self-injurious behaviors (SIBs) can be absolutely devastating for families with children on the autism spectrum. Common self-injurious behavior includes: headbanging, hand biting, pinching, slapping head/ears, eye poking, hair pulling, scratching/rubbing, throwing one's body on the floor, etc. The horror of watching your own sick child further inflict pain onto himself, and the helplessness of not knowing what to do to help him is heartbreaking. I have met with families through the clinic who have had terrifying episodes of SIBs in front of me as well as at home, and many of these families as well as others are now using the Protocol to help their children with very good results.

There are many theories as to why some children have SIBs. The article "Self-Injurious Behavior" by Stephen Edelson on ARI's website discusses several different theories and supporting research.[2] The theory that most fits what I see on the ground is that of pain alleviation. Many of our kids are in chronic pain and cannot tell us. Sadly, we haven't always had answers to help them.

If we follow this thinking, and assume that many of these children suffer from severe gastrointestinal pain, or headaches (supported by head banging, gut clutching, etc.) we can hypothesize that much of this pain can come from pathogenic overload, parasites wreaking havoc on the body and the resulting inflammation.

Families applying CD in combination with the Kalcker Parasite Protocol are reporting a dramatic reduction in SIBs in their children. I believe there is an important connection to be made here, as we eliminate pathogens and parasites SIBs decreases. More research is needed, but for families facing the horrors of self-injurious behavior, this combination could be of great value. (More testimonies on decreased SIB's can be found in Chapter 2, and the Miracles and Testimonials section at the end of the book starting on page 357.)

> Prior to starting CD, we had tried almost every autism biomedical protocol. We saw several DAN and GI specialists and numerous other practitioners. We spent tens of thousands of dollars on consults, supplements, diets, medicine and medical procedures not covered by insurance. My daughter either had no response or negative ones and she continued to get worse. She developed extreme self-injurious behaviors that progressed to the point of her needing to be restrained several times a day in an attempt to prevent her from pinching or biting herself until she drew blood. Her body was covered with bruises and cuts. She would also pull out big clumps of her hair. She would be up for hours screaming in the night with these behaviors. She was severely malnourished despite the fact that two registered dieticians had deemed her daily caloric intake adequate. She was 14 years old and weighed a little over 70 pounds. I honestly did not think that CD would be any different than anything else we had ever tried but figured things could not get much worse.
>
> Since starting CD her self-injurious behaviors are almost entirely gone. She went from having to be restrained almost daily to once in the last 5 months. Her former multiple daily 45 minute self-injurious tantrums are now a few minutes of crying or protest and happen once a week if at all. We have noticed gains in language and cognition (20 point drop in her ATEC score). After years of fluctuating between loose stools and constipation, she is now regular and has gained almost 10 pounds. She sleeps through the night and we are able to take her places that we would have never dared to before. If my daughter can make gains this drastic in 5 months at age 14, I can't wait to see what the future holds.
>
> Julie, Colorado, USA

PANS, PANDAS and PITAND

PANS stands for *Pediatric Acute-onset Neuropsychiatric Syndrome*. It encompasses PANDAS (Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcal Infections) and PITAND (Pediatric Infection Triggered Autoimmune Neuropsychiatric Disorders), where PANDAS is a subset of PITAND. Dr. Susan Swedo, Chief of Pediatrics & Developmental Neuroscience Branch at the National Institute of Mental Health (NIMH), and her research team identified a subtype of OCD in children during the 1990s that is triggered by infection. Streptococcal infection has been most studied; however, any infection can trigger this condition.[3]

I have run across these diagnoses a lot since I have started helping families in first world nations. There seems to be an overlap in some autism and PANS/PANDAS/PITAND cases. I have heard that around 25% of children with ASD's also suffer from PANS/PANDAS, and/or PITAND. Some practitioners would put that figure even higher. However, since further research is needed, and this field is still developing it is impossible to say for certain at the moment.

According to the International OCD (Obsessive Compulsive Disorder) Foundation some of the symptoms of PANDAS/PANS may include:

- Acute sudden onset of OCD.
- Challenges with eating, and at the extreme end, anorexia.
- Sensory issues such as sensitivity to clothes, sound, and light.
- Handwriting noticeably deteriorates.
- Urinary frequency or bedwetting.
- Small motor skills deteriorate - a craft project from yesterday is now impossible to complete.
- Tics
- Inattentive, distractible, unable to focus and has difficulties with memory.
- Overnight onset of anxiety or panic attacks over things that were no big deal a few days ago, such as thunderstorms or bugs.
- Suddenly unable to separate from their caregiver or to sleep alone.
- Screaming for hours on end.
- Fear of germs and other more traditional-looking OCD symptoms.

I know several moms who have been able to reduce/control/eliminate their children's symptoms with CD. In fact, families began to eliminate the antibiotics that they were using as they reached full dose of CD. I am by no means suggesting anyone reduce or remove a prescription medication without a doctor's supervision. I am simply reporting on what I have witnessed.

We know that CD kills bacteria, so we could hypothesize that CD could have a similar effect on the PANS/PANDAS/PITAND symptoms that an antibiotic would. What has worked in these cases is upping the number of doses of CD given a day so as to kill pathogens around the clock. For PANDAS/PANS/PITAND use two baby bottles a day. If your max dose is eight drops in an eight ounce bottle, and a PANDAS flare presents, prepare two baby bottles with eight drops in each and give one dose every hour over the course of 16 hours. Some families are doing 16 doses a day even if there is no flair and the child is doing excellent.

As noted previously, in the absence of research, we cannot recommend doing nothing for these children. Their suffering is intense and the risk of long-term damage cannot be ignored. We also do not know if some of these children remit spontaneously later in adolescence. We do not know if some of them grow into adults with the most treatment resistant OCD due to permanent basal ganglia damage. While we research those questions, offering treatment options at the local level, especially in the first few months after onset, could be critical to the long term success for these kids.[4]

We parents are pioneers in the research surrounding PANS/PANDAS/PITAND, so it is critically important that we document what we are doing with our children and what is working for them. In fact, A MOM was the first to discover the connection between strep and OCD/tics. Her child was the first documented PANDAS child. Dr. Susan Swedo reinforces that physicians should listen to what the mothers have to say![5] Speak up to your doctor, share your findings, join our forum to learn, and share what is working for you. Parents are driving the biomed movement (and always have) for ASDs as well as PANS/PANDAS/PITAND. In fact Dr. Bernard Rimland was a parent and the first to suggest biomedical interventions for people on the spectrum. We as parents can help countless other families by sharing our experiences.

You will find more info on PANS/PANDAS/PITAND at:

www.pandasnetwork.org

and

ocfoundation.org/PANDAS

Fever Therapy

Dr. Kenneth Bock and Nellie Sabin, in their groundbreaking book *The Road to Immunity: How to Survive and Thrive in a Toxic World*, explain how to use fever therapy and how it is an incredible tool for healing.[6] When a mom emails me to tell me her child has a fever, it tells me that the immune system is reacting to an invader, and we have an excellent opportunity to kill pathogens in the body. Having a fever is part of an effective immune response. This is also true in infants as well as small children. Using over-the-counter (OTC) fever reducers interferes with the body's own ability to fight infection.

This section was excerpted from the aforementioned book.

> Fever has long been recognized as an important tool in the process of healing. The process of heating the body is called "hyperthermia." Hyperthermia, or ordinary fever, can be very effective against cancer cells, various pathogens, or invading organisms that cannot "tolerate the heat." Hyperthermia may not kill all pathogens in the body, but may reduce the number of invaders in the body to a manageable level for the immune system. Hyperthermia may also be responsible for stimulating the immune system to produce cytokines i.e. antibody production, and release of toxins from fat cells.

What Temperature Constitutes a Fever?

According to Dr. Robert Sears, the following are normal temperatures in children over 3 months:[7]

Normal temperature: 97° to 99° F (36° to 37.2° C)
Low-grade fever: 99° to 100.9° F (37.3° to 38.3° C)
Common fever: 101° to 103.5° F (38.4° to 39.7° C)
High fever: 103.6° F (39.8° C)

Fever Therapy in Practice

If your child develops a fever, it is not a failure of the body that needs to be corrected, but a mechanism that brings the body closer to healing. Rather than administer an OTC remedy to reduce fever, apply a cool, damp washcloth to the forehead and neck, replace them as they warm, and let the body do its job. Some parents put damp socks on their child's feet.

During a fever you can dose orally, as well as do one CD bath and one CD enema in one day. Keep dosing CD every hour (16 to 24 doses a day), unless the person stops eating. Even if you must reduce the number of drops, even if you are only giving one drop, keep dosing every hour. CD baths, as well as baths in general, are also your friend if your child has a fever. Some parents have given several baths throughout the day while a fever is present with great results. Bath the child if the fever spikes. Always be vigilant that the temperature does not rise too high or too fast so as to cause febrile seizures, we have to use common sense and trust our instincts. Bath and keep a cool damp wash cloth going 24/7 until the fever pases.

According to the National Institute of Neurological Disorders and Stroke febrile seizures are convulsions brought on by a fever in infants and small children. The majority of children with febrile seizures have rectal temperatures greater than 102° F. Approximately 1 in 25 children (4%) will have one febrile seizure. Although febrile seizures can be frightening to parents, the vast majority of febrile seizures are fast and harmless. There is no evidence that short febrile seizures cause brain damage. Large studies have found that children with febrile seizures have normal school achievement and perform as well on intellectual tests as their siblings who don't have seizures.[8]

If your child has a febrile seizure, experts recommend that you try to stay calm, keep the environment around your child safe (place them on the floor if they are not, and clear objects out of the way), and allow the seizure to resolve spontaneously without restraint. Try to look at your watch so you will know how long the seizure lasted. Do not put anything in your child's mouth as it may break, obstructing their airway.

Skin Issues

I have been hearing about skin issues right and left: rashes, hives, redness, acne, eczema, psoriasis, pustules, etc. It is very easy to assume that since skin is on the outside of our bodies that these reactions are caused by exterior factors (i.e. irritation from a detergent, mold, etc.). However, the skin is our largest detox organ, and many pathogens seek exit through the skin causing many of the above conditions.

The scenario that I keep hearing is the following: We just started the Kalcker Parasite Protocol. Little Johnny has a rash with white heads on his legs. What do I do? When the families keep up with the parasite protocol, and continue dosing daily with CD these eruptions disappear, and this is generally accompanied by improvements in behavior, or hyperactivity/anxiety.

We need to think about skin issues as outward manifestations of inner toxicity, and use our protocols accordingly. Skin issues can be a good indicator that something is being flushed from the system, or something needs to be flushed from the system. The Kalcker Parasite Protocol combined with CD tends to eliminate skin issues at the root.

The following paragraphs were excerpted from leakygut.co.uk, written by Dr. Gloria Gilbere ND DA hom PhD[9]:

> Experts agree that most chronic skin disorders, specifically psoriasis, is the external manifestation of the body's attempt to eliminate internal toxins that have accumulated within the lymphatics and blood stream by "seeping" through intestinal walls in a condition known as leaky gut syndrome.
>
> In working with clients around the world, the protocols that have been most effective are those that include repairing the intestinal walls by first beginning a protocol to eliminate candida yeast and parasites.
>
> The most effective cleansing protocol is obtained by clearing out these "uninvited" health-depleting microorganisms in order to release the accumulated pressure on the liver and the intestines—allowing free passage and thus reduction of overall toxic load.

The following paragraphs, authored by Dr. Edward F. Group III, DC, ND, DACBN, DABFM, were excerpted from parasite-cleanse.com/intestinal-parasite-symptoms.html[10]:

> **Symptoms of Intestinal Parasites Related to Skin**
>
> The presence of intestinal parasites, at times, manifest as allergies, tumors, and various other skin conditions.
>
> Intestinal parasites cause the human body to release hormones and immune system defenders, which in turn irritate the skin and cause various skin conditions such as acne, hives, rashes, itching, weeping eczema, ulcers, swelling, sores, lesions, blisters on lips, and dermatitis. It also deprives the skin of its usual glow.
>
> Intestinal parasites cause irritation and inflammation in the stomach making it difficult to digest certain types of foods. The presence of undigested food particles in the stomach causes the body to produce increased levels of eosinophil, an immune system defender. These eosinophils inflame the body's tissue and cause skin rashes.
>
> Intestinal parasites trigger the body's immune system to develop a tumor-like mass known as granuloma to encase parasitic larvae or eggs. Granulomas usually develop in the colon, rectal wall, lungs, liver, peritoneum, and the uterus.

Btw, you were bang on re his allergies and PP (parasite protocol). He had this horrible red lip thing because of food allergies. He would make it worse by rubbing it. It would get really bad. After first PP (Parasite Protocol) - NO MORE RED LIP ISSUES!

Families often tend to panic when dealing with sudden "unexplained" skin issues. If you are seeing chronic or acute skin issues I think there is enough evidence to look towards parasites as the main or at least a contributing culprit. Therefore, if we work towards removing the parasites causing the symptoms, rather than treating the symptoms topically we can expect an alleviation of the symptoms.

If you are seeing acute or chronic skin issues and you have not started a Parasite Protocol, I would suggest you revisit the parasite chapter and go over the recommendations with your healthcare provider. Keep dosing CD, in some cases upping the dose has helped resolve the issue faster, and in the case of severe irritation using a CD spray topically can provide relief (10 activated drops per ounce of water in a spray bottle) and a CD bath can also help. I have seen most, if not all skin issues resolved with CD and/or the Kalcker Parasite Protocol.

Healing Adults with Autism

Common thinking in the biomed community and the autism community at large is that autism is only "curable" or "recoverable" up to a certain age. I have heard that two to nine years old is the "golden window" for autism, and that after nine years of age improvement would be possible, but not recovery. The neural pathways would be set, the damage done, and the child would retain the autism diagnosis for life.

Not anymore. This year a 15-year-old recovered from autism, as well as two 17-year-olds. In addition, I am helping an amazing mother of a 31-year-old with autism that began the Protocol and for the first time in 28 years was seeing real progress with her son. In fact that young man's ATEC score dropped from a 65 to somewhere in the 20s in his first three months on The Protocol. I want to prove that autism recovery is possible at any age. It may take longer, and the road may prove more challenging, but it's already happening. Families that have been using biomedical interventions for more than ten years are finding the Protocol and sharing their gains.

Older children and adults on the spectrum start the same way as the little kids—clean up the diet first, and then on Day one of CD, give one drop in eight ounces of water over eight doses throughout the day. It is so important to start low and slow even with adults because they have often been chronically ill much longer than young children and have deep rooted pathogenic and parasitic infections. This means that when we start giving CD they might experience a detox reaction considering the amount of pathogens that will be dying.

Working with Neurotypical Adults

Can CD benefit parents and other adults not on the spectrum? Absolutely. You can literally find thousands of testimonials on the Internet (Jim Humble's website, www.jimhumble.org or www.curezone.com, etc.) on people who successfully detoxed, or healed a myriad of illnesses using CD.

Many parents who have seen improvements in their children have decided to use CD, or the Kalcker Parasite Protocol for themselves. I have done CD oral, baths, and enemas on myself, especially when I feel "something coming on" and with great results.

It's important to keep in mind that just because we weigh more than our children doesn't necessarily mean we should jump in and go to full dose from one day to the next. The best way to start an adult who may have 40 or 50+ years of accumulated toxins/pathogens/parasites is low and slow. Use the baby bottle method and start day one with one drop spread out over eight doses and go up from there. Chart your progress, see how you feel, go low and slow. When we as adults use the protocol it can be great for us, but also a window into how our children are feeling.

I love to see a mom or dad begin CD with their child. Both begin at one drop on day one in eight ounces of water, 1/8 or a drop per dose, for eight doses throughout the day. Starting with our children gives us a much better understanding as to how our child is feeling, and when we up the dose how that feels as well. This includes enemas. Once parents see how good they can feel after an enema, it helps garner confidence, if we had doubts about using them for our children. Applying our own enemas can help to remind us just what a powerful tool they are for healing.

Many of the parents doing the protocol for themselves write me to tell me about their remission from fibromyalgia, chronic fatigue syndrome, Lyme disease, life-long constipation, etc.

We all do the parasite protocol in my house so as not to re-infect my son, as parasites can be highly contagious. Last year I did a round of parasite medications prescribed to me by a doctor. During the course of treatment I felt a tremendous amount of anxiety and some depression (feelings I had never felt before) right before I passed worms, I can assure you it felt awful. That experience motivated me to encourage parents to get these invaders out of their children's bodies.

As soon as I started passing worms and detoxing more with CD, I started to feel good again. The cough that I had had on-and-off for years miraculously disappeared. Seriously, go low and slow, a toxic body can only process so much at a time. Too much too fast can cause headaches and diarrhea. The way I see it, there is no need for that, listen to your body and go at whatever pace you need. Don't just "push through," low and slow wins the race. Always remember to drink water when doing any sort of detox!

Tax on a Marriage

During the first five years after an autism diagnosis married couples face an estimated 80% divorce rate, and a 90% chance of divorce beyond the tenth year of marriage. My family was almost a statistic. Even before we had a diagnosis, our family was suffering. Due to Patrick´s constant screaming and tantrums, he wasn't sleeping at night, and that meant that no one else was sleeping either. He had also stopped being the affectionate baby that he was for the first two years of his life, which was emotionally challenging for all of us. We were all over stressed and sleep deprived, not to mention having symptoms of post traumatic stress disorder (PTSD).

If I could give any advice to families dealing with a new diagnosis it would be to start The Protocol and CD immediately. The sooner the affected family member starts to heal, so will the rest of the family, similar to the sun coming out after a storm—this too shall pass.

Financial Stress

Not only is an autism diagnosis emotionally stressful on a marriage, but it can be extremely financially stressful as well. The Protocols outlined here can be extremely cost effective, even if including ocean water. Recently, one of the moms using The Protocol did some calculations comparing what she spent previously (see table on page 342) using another biomedical intervention with many supplements, etc., to what she spends now (see table below).

The Protocol - Estimated Cost

Product	Size	Price	Reorder Schedule	Current Cost per Month
CD	1 bottle	$22	1/mo	$22.00
Diatomaceous Earth	2.5 lb	$15	1/3mo	$5.00
Rompepiedras	50 ml	$24	1/2mo	$12.00
Vortex Ocean Water	1 L	$40	1/mo	$40.00
THERALAC®	30 caps	$36	1/mo	$36.00
Natures Way® Neem	100 caps	$4.09	1/5mo	$0.75
Pyrantal Pamoate (Combatrin®)	32 oz	$35	1/6mo	$5.00
NOW® Castor Oil	16 oz	$8.23	1/6mo	$1.38
Mebendazole	90 tabs	$28	1/5mo	$5.60
TOTAL				**$127.73**

***Defeat Autism Now!* based Protocol - Estimated Cost**				
Product	**Size**	**Price**	**Reorder Schedule**	**Current Cost per Month**
Actifolate	60	$15.00	1/2mo	$7.00
Calm PRT	120	$89.00	1/5weeks	$78.00
Caprilex™	90	$16.00	1/6weeks	$10.60
Cytoflora™	4 oz	$79.00	1/mo	$79.00
Designs for Health Lithium Synergy	120 caps	$22.00	1/mo	$6.00
DMG	90	$16.00	1/5mo	$10.60
Douglas B-Complex	60	$21.00	1/6mo	$21.00
Douglas CoQ10	60	$49.81	1/6mo	$12.50
Enhansa™	150 caps	$58.00	1 / 75 days	$29.00
Enzymedica Virastop™	120 caps	$32.11	1 / 4 months	$8.00
Garden of Life Probiotics	81g	$32.17	1/mon	$32.17
GI Revive	225g	$52.00	1/mon	$52.00
Homocycsteine Supreme	60 caps	$20.00	1/4months	$5.00
Kirkman Digestive Enzyme	200 caps	$60.00	1/3months	$20.00
Kirland Vit E	500	$13.65	1/16months	$0.84
Klaire Multivitamin	180	$28.00	1/3months	$9.00
Liver Life	4 oz	$50.00	1/2months	$25.00
Magnesium	240	$13.25	1/6months	$7.00
Metagenics Adreset®	60	$33.00	1/4months	$8.00
Milk Thistle	100	$25.00	1/3months	$8.00
NDF+	1 oz	$80.00	1/6months	$15.00
NeuroProtek®	4 bottles	$140.00	4/4months	$35.00
NeuroScience Avipaxin	60	$58.00	1/2months	$28.00
NeuroScience EndoTrax	1 bottle	$27.00	2/month	$54.00
NeuroScience Kavinace	120	$65.00	4/4months	$32.50
New Beginnings Copper	50 ml	$20.00	1/year	$1.75
NOW MCT	32 oz	$17.00	1/6months	$2.80
Nutribiotic GSE	4 oz	$16.99	1/6months	$3.00
OSR	30	$60.00	1/2months	$30.00
Oxycell Glutathione Cream	1 jar	$33.00	1/2months	$16.50
Pharmax Cod Liver Oil	8 oz	$13.33	1/mon	$13.33
Source Naturals B6 (p5p)	240	$15.00	1/6months	$2.60
Source Naturals Evening Primrose Oil	90	$7.45	1/6weeks	$4.97
Source Naturals Gaba	100g	$18.00	1/4months	$4.50
Source Naturals Lysine	200	$10.00	1/mon	$10.00
Source Naturals Quercitin	200	$25.80	1/3months	$8.25
Speak EFA	60	$42.00	1.5/month	$63.00
TheaNAQ	90	$54.00	1/month	$54.00
Therabiotic	120	$69.00	1/2months	$35.00
TravaCor Jr.	120 caps	$24.00	1/4months	$7.00
Vit B12 shots	1 refill	$45.00	1/3months	$15.00
Wobenzyime®	120	$60.00	1/2months	$30.00
TOTAL				**$894.91**

Tips for a Non-Toxic Home

Many of our children on the spectrum have chemical sensitivities, and even if they don't, it's wise to limit their exposure to toxins during the healing process as much as possible. Toxins are not good for anyone at any point in their life. The benchmark investigation "Bodyburden: The Pollution in Newborns" by the Environmental Working Group[11] (see Appendix 13, page 499) explains why this is so important for children with developing brains and immune systems. The following is a list of products to use instead of toxin filled, name brand cleaning products.

Floor cleaner: Add one cup of white vinegar and baking soda to a bucket of water (add a squirt of liquid soap if you like).

Air Freshener: To remove odors from carpets spread on baking soda and then vacuum the carpet. Simmer cloves and cinnamon as air freshener.

Removing Toilet odors: Borax

Trashcans: Spread baking soda in the bottom, then spray to disinfect with a solution of borax and water.

Dish soap: I use liquid dishwashing soap and white vinegar to cut grease. If you buy dish soap make sure it's biodegradable.

Disinfectants: Borax with water is an excellent disinfectant, or spray CD.

Laundry detergent: Soak the clothes for an hour in borax soap before putting into the washer.

Oven Cleaner: Dissolve two teaspoons of liquid soap (not detergent) and two teaspoons of borax with enough warm water to fill a spray bottle.

Windows and Mirrors: Club soda in a spray bottle

Toilets and Tubs: Baking soda with apple cider vinegar

96% Alcohol: Kills parasites on enema catheters, etc., and is great for cleaning bathrooms.

Videos

I have had many parents ask me if I think it is a good idea that their child watch "Genius Baby" or other videos that are supposed to help with their development. I do not believe these videos to be appropriate for a developing brain. Studies were done in Cuba that found that excessive TV watching before three years of age can cause the symptoms of autism. When the TV was removed from the children that presented symptoms of autism, the symptoms disappeared. They also reported that TV watching could cause neurological discharges in the brain.

I can also tell you that the only time my son stims or flaps is when he is allowed to watch TV. When he was younger, and deep in the throes of autism, the only thing he wanted to do was watch TV, while running back and forth and squealing. Any other activity was met with screaming and tantrums. He is much more interested in other activities now, and easily transitions from one to the next, including school. However, he still requests to watch YouTube in the evenings when all of his activities are complete, and that is truly the only time he flaps his hands anymore. To me it is very obvious that TV is a source of over stimulation for him. However, it is hard to deny him a truly preferred activity, especially when he enjoys videos on cooking and foreign language! The least amount of screen time (TV, computer, iPad, etc.) the better, however, we do what we can.

Therapy Options:

ABA is not the only therapy for children with autism.

Many parents worry because ABA/IBI is either not available, or not affordable to them. I firmly believe that something is better than nothing. At the same time, I have seen many children recover without any ABA or IBI. Here are some of the options that have worked for families in conjunction with the protocol:

- ABA - Applied Behavior Analysis
- Occupational Therapy
- DIR® Developmental - Individual Difference, Relationship-based Model
- Floortime Therapy
- TEACCH - Treatment and Education of Autistic and Related Communication Handicapped Children
- Son-Rise®
- PROMPT - Prompts for Restructuring Oral Muscular Phonetic Targets
- Sensory Integration Therapy
- RPM - Rapid Prompting Method
- Supportive Typing

There are many other options not mentioned here. Human interaction is key for our children on the spectrum to find joy with us in our world.

I have known many families to have varying degrees of success with all of the above therapies. Some are better for younger children and some for older children. Depending on the areas most lacking in your child's development, you can investigate any one of these types of therapies. Families in many countries where insurance doesn't cover any type of therapy have had to be trained themselves in how to apply therapy. Sometimes an aunt, sister, cousin, or neighbor steps up to help out. No ONE therapy has been fundamental to all of the recovered children and for that reason I say something is better than nothing when it comes to therapy.

Even household chores can be therapeutic and task oriented. Simply because a person has autism doesn't mean they can't have responsibilities and learn new skills while helping out in the home.

A word on laboratory testing

I can't stand to watch people waste money that they could be using for healing their children through various interventions on laboratory testing. If you have all the money in the world, and want to chart your child's progress through laboratory testing, so be it. However, the rest of us sometimes have to choose between expensive testing and treatment. So many families have come to me after spending thousands of dollars they didn't have on testing to tell me their child has viruses, bacteria, and candida. Am I surprised? Never. I have yet to see a child with regressive autism that does not have pathogens/parasites. No matter which pathogens are present, we use CD.

Testing for parasites is woefully lacking at this time, and in fact, the only practitioners who have consistently diagnosed parasites have been veterinarians and practitioners of alternative medicine (also Metametrix labs have identified parasites in some cases).

Generally speaking, if a child doesn't have viruses, bacteria, yeast, parasites, heavy metals and allergies they don't have autism.

Stay Up-To-Date

Subscribe to our Newsletter!

Just go to www.cdautism.org and enter your name and email address...

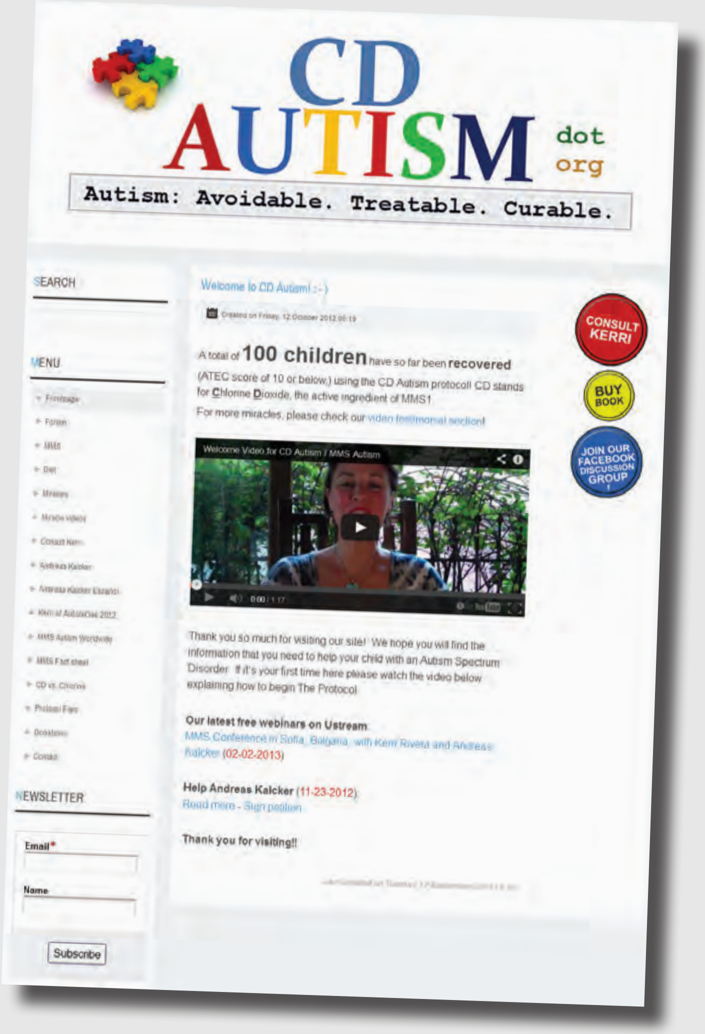
CD AUTISM dot org
SEARCH
MENU
Frontpage
Forum
MMS
Diet
Miracles
Miracle videos
Consult Kerri
Andreas Kalcker
Andreas Kalcker Español
Kerri at AutismOne 2012
MMS Autism Worldwide
MMS Fact sheet
CD vs. Chlorine
Donations
Contact
NEWSLETTER
Email*
Name
Subscribe
CONSULT KERRI
BUY BOOK
JOIN OUR FACEBOOK DISCUSSION GROUP!
Welcome Video for CD Autism / MMS Autism
0:00 / 1:17

Autism: Avoidable. Treatable. Curable.

Welcome to CD Autism! :-)

Created on Friday, 12 October 2012 06:19

A total of **100 children** have so far been **recovered** (ATEC score of 10 or below,) using the CD Autism protocol! CD stands for Chlorine Dioxide, the active ingredient of MMS1

For more miracles, please check our video testimonial section!

Thank you so much for visiting our site! We hope you will find the information that you need to help your child with an Autism Spectrum Disorder. If it's your first time here please watch the video below explaining how to begin The Protocol

Our latest free webinars on Ustream:
MMS Conference in Sofia, Bulgaria, with Kerri Rivera and Andreas Kalcker (02-02-2013)

Help Andreas Kalcker (11-23-2012)
Read more - Sign petition

Thank you for visiting!!

Chapter 15

Final Thoughts

If you knew how powerful your thoughts were
you would never think a negative thought again.
~ Peace Pilgrim

What would be your message to families of children with autism who have just read the book, but not started the protocol?

Do not wait! Get started now! Autism is Avoidable, Treatable and Curable! You don't have to spend fortunes on lab testing, imported foods nor supplements. Even if the medical community can't decide if your child has autism, PDD-NOS, ADHD, Asperger's, PANDAS, etc.... don't wait! Start the diet. Even if you cannot do more than that, it will help. Some children have lost their diagnosis with just diet. There are parents all over the world doing biomedical protocols for autism, get together with them, join the forum at...

www.cdautism.org

...or the CDAutism Facebook group. The Internet helps us all to be close and share information. Don't let anyone tell you the diet doesn't work, just keep going and don't stop. There is never an excuse to not do what is right for your child. "Their future depends on our courage."

The final chapter of this book is yet to be written. Until the day we have recovered every victim of regressive autism, there is no success. Just progress. Here I have expressed what I have seen and what the road of autism recovery has brought to me. I have learned a lot and continue to learn. Our children need us to keep learning and searching until we find what unlocks them from their autism. Not everything works for everyone. However, I have seen that this protocol works for most. Especially when it is done exactly as it should be. If you have found this book you have an advantage that I did not have with my son. When I first learned of Patrick's diagnosis I lost precious time and money that I didn't have to lose, trying interventions that didn't yield results. In the nine years that my son has been in recovery I have yet to come across another protocol that has helped as many children to recover from autism in such a short amount of time—less than three years, to be exact! As of the date of this writing in May of 2013, 93 children have lost the diagnosis of autism and every month more and more are recovering.

Please remember that if your journey is autism you will never walk alone. I am here too. Just one click away...

www.cdautism.org

I wish everyone who is ill to have the fastest recovery possible from whatever it is that ails them.

To be continued…

Dr. Andreas Kalcker, Kerri Rivera, Jim Humble, in Puerto Vallarta, Mexico 2012.

Chapter 16

Healing Beyond Autism...

When TRUTH is found, understood and accepted is when we realize it was not even close to what we thought it would be...
~ Miriam Delicado

As Jim Humble has been proving for many years, chlorine dioxide is a molecule capable of healing many common ailments. When combined with the other elements of this protocol, in the correct order, it has shown to be even more effective. We have decided to include this bonus chapter because it is time to start sharing this information with everyone, as pathogens and parasites are the root cause of many of the diseases that are causing suffering in our world today.

As parents saw improvements in their children's health from the protocol, some of them decided to use it on themselves as well. Similarly to when the flight attendant tells you to put on your own oxygen mask first before helping others, it is important to take care of yourself if you want to be able to help your children heal, grow, learn and mature.

Over time, parents who followed in their children's footsteps started to have some undeniable gains in their own health. This was an added bonus many didn't expect, and we couldn't resist sharing their stories with you.

After reading the parasite chapter, you may have been surprised to see how many common symptoms can be caused by parasitic infections. Imagine what would happen if we could get this information into the hands of people who have seen many doctors, used various drugs, and applied countless protocols for years while only masking the symptoms of the underlying cause. So, there should be no surprise if going after the real cause was life changing in real ways.

The following testimonials are a small sample of how some of these people's lives have been changed. Serious afflictions such as Lyme disease, fibromyalgia, chronic fatigue syndrome (CFS), depression, cancer, and chronic infections, are just some of the conditions that improved or completely cleared up. By cleaning up toxic pathogens and parasites, as well as removing foods that feed these invaders, it should be no surprise that disease triggering inflammation is reduced so true healing can take place.

I am so excited for you to read these, and I know that some of you will recognize yourselves here. Please know that healing is possible! We have a public Facebook group called *CD Health* where adults using the protocol for themselves can share questions, doubts and success stories. I highly recommend that you check it out if you are looking for support or some guidance getting started.

Testimonials Beyond Autism

After we saw our son recover from ASD using CD, I tried CD myself and it helped me recover from Lyme and its co-infections. Some of our extended family heard about our success and they have had very good results using CD for treating Hypertension, Diabetes and Liver problems. My younger NT son used to bang his head in sleep since he was born, CD has now resolved the problem. I think anyone who tries CD sincerely, can only benefit in treating so many hard problems. One very beneficial side effect I have seen with CD is the resolution to seasonal allergies and improvement in the immune system to the point where kids do not fall sick or catch colds. In this regard CD is much better than yearly flu shots which have so many toxic elements in them and still do not protect adequately.

—Nilesh

To say I was anxious about the impending birth of Baby #2 is an understatement. I felt like I had failed our first born on so many levels and was a mess believing that I'd do the same for our new daughter.

But alas, she was born a perfect and healthy baby girl. So was Callum, our son, but unlike Lila who had zero interventions, Callum had antibiotics from my labor, the Vitamin K shot, HiB, and DTaP. And it was the DTaP vaccine at 15 months that took the shine from our sweet boy's eyes.

We discovered Kerri's protocol when Lila was 3 months old and more worries cropped up about how to best protect her from the die off that was going on in our house.

At 5-months, she developed an eczema patch underneath her chin, became chronically constipated, and her normally cheerful disposition became irritable.

As soon as she hit the 20-lb mark, I started giving her 20 drop CD baths, low dose oral CD, and 3 drop CD/150 mL water enemas. To my horror and surprise, Lila started passing 1" baby roundworms almost daily.

In a family with both NT and ASD children, the NT child always has some type of quirky behavior. Now we know why! Parasites do not only affect the ASD

population, they affect everyone. Having CD in my tool chest gives me the power to prevent so many things that may have been in my daughter's future.
Don't be afraid to treat your babies! Her eczema is gone and she's pooping regularly once more.

I write and share my story in that it might help or save someone else's life. That it may just be the key to help someone see what may be what is happening for them as it was and is happening to me. If it helps just one person onto their path to recovery and health then it will have been worth sitting down at this computer and sharing it.

For the past 10 years or so my world has been collapsing in on me, and getting smaller and smaller, and accelerating the past 5 years. I have felt as if I were under a thick coffee mug on the edge of a table about to fall off. On a scale of 0 to 20 with 0 being death and 20 being fully alive and cognizant, I felt as thought I was at 3 staring at 0. For the past 5 years at least I have done so much research, listened to so many radio interviews, done so many different modalities, bought so many products, paid a lot of money to different therapists, ate weeds, made my own concoctions, etc. My mother would always say to me wow you should be so healthy to which I would reply no mum, actually I am not. I began feeling like a phony. I ate real good took all the right stuff, got plenty of rest foraged for my own food and spring water, got exercise, walked barefoot, etc, but always felt exhausted and looked worse and worse. I was fighting with everyone it seemed and chasing people away including family.

Four months ago my wife decided we should go away on a weekend getaway and relax. We fought the whole drive up there and when we got there she threw me my room key and said you're on your own. I spent the whole weekend in bed and on the trip home we both decided to end our 20 year relationship and marriage. I was good with that and looking forward to being on my own. I was finally chasing the last person who mattered to me away.

A week later an angel came into my life. I have been listening to oneradionetwork.com for the last 5 years and on this day Patrick Timpone had a lady by the name of Kerri Rivera on and she spoke about how she was curing kids with autism and that the kids had parasites. I looked up the symptoms of autism and realized I had quite a few of them. I immediately began the mms enema protocol which she recommends. I could not believe what then came out of me. I kept this up for a week and still there was no let up with the amount of parasites, biofilm, and who knows what coming out of me. After a week of this and seeing no let up on what was coming out of me, and having just done an mms enema I laid on my bedroom floor and the grief just poured out of me. My wife came in and lay down next to me and just held me. I knew then that something had changed, that my life and my marriage were now on a healing path. That the pain, the toxins and parasites were now leaving my body.

It has now been just over 3 months and there has been no let up on what has been coming out of me. If I were to have scooped it all up and put it in a jar I am sure it would fill a one gallon jar with hardly any water in it. I now feel I am at about a 9 with 20 in my sights. I thank God that this info came into my life and I took action on it. It has saved my life and my marriage. I love my wife, and cannot imagine not having her with me. And I love my life and am looking forward to not only getting *to the 20 mark but what is beyond the 20. What is possible when one's brain is at full capacity and not being inaundadated with toxins? Much love and appreciation goes out to Patrick Timpone at oneradionetwork.com, Kerri Rivera and her work with autistic kids (adults too Kerri), and to Andreas Kalcker. Thank you all for showing me and pointing to the keys I needed to free myself from what has been a nightmare.*

A big part of my healing has been reviewing my life and realizing that this has been going on for most of it and answers so many questions I have had. As a child growing up I had pain in my stomach and what the doctors eventually did was take out my appendix, which most of us know now is needed to keep the healthy flora in our intestines. And they told my parents that it was all in my head. So whenever I complained about stomach aches and pains was told that it was all in my head. School was a nightmare for me, as socially I felt I didn't fit in or belong, and a lot of easy concepts that most others got, I just couldn't get. I always had the feeling like I was missing something. I felt as though I was stupid or something just wasn't right with me. I now know there are many just like me. The words Autism, add, adhd, ocd, etc were not a reality back then. And I know those are just words or diagnosis of toxins and parasites. Get rid of the toxins and parasites and the diagnosis goes away. Easy peasy. Thank you all for listening to this chapter of my story. I look forward to many more chapters in my life story. Please feel free to pass this along to anyone who it may help. Much love.

—Bruce

When starting the protocol for my son's Autism, I decided to do it along with him.
I was suffering from Hypothyroidism, Adrenal Fatigue, a Hormonal Imbalance, and Severe Anemia. My cortisol was at a 3. One week into the protocol, my cortisol was at a 14, and my iron levels were normal without supplements. My thyroid medication dosage was decreased twice. My vitamin levels were all normal. My uterine fibroid cysts were completely gone after a few months on the protocol. I continue to see improvement in my health. This protocol has been a Godsend for my family.

—Maggie Kaye

CD has been life changing for my 18 year old son who has autism and Lyme, although his ATEC today is only a 3, down from a 68 at the start of cd, but I never thought it would be so instrumental in my dad's life also.

My dad, William, has a malignant melanoma in his eye and has received radiation for treatment for the last few years. I talk about cd all the time but they never took this treatment seriously, at least for anything besides autism until now. My dad gets terrible ear infections that he sometimes has to go to the hospital for. They tell him each time that he has "swimmers ear" and they give him abx, sometimes it seems to work, sometimes not, and he goes back to the drs or hospital for more abx....over and over. Recently he was leaving here for his 7-hour trip home, and he was very worried that he would not make it home with the painful infection that had begun, *and he would have to stop on the way and visit a strange hospital. So I mixed up some cd, and told him to drop a few drops of the cd mixture in his ear periodically on his way home. To his surprise, the infection cleared up by the time he got home and there was no need to even call a doctor.*

Of course, my secret plan is to get him on the cd protocol to finally deal with the tumor in his eye and that's the plan we have started. Whether he uses it and shrinks the tumor and saves his life with cd or not, he's a true believer now that cd cures infection.

I have been taking CD and the parasite protocol for 5 mo., and here are my observations. I have had long-term chronic sinus issues which has caused sever sleep problems. Also, a few months ago, I had a severe toothache in the upper left area. The dentist examined and found no fault with the tooth, even though there was severe swelling in the area. The infection originated from my sinus, and by taking my CD regularly hourly, was quickly solved. (My hourly CD doses were compromised by travel) I have also noticed improvement in my sinus condition, resulting in better sleep, although I have not fully recovered yet. Right now my sleep is somewhat interrupted by peetox. I will continue CD and PP for the 18 mo. Previously; I had used several naturopaths, with either negative results or limited results that disappeared when the medications were discontinued. So far I am happy with CD and would recommend it to people to try.

—Ed Canada.

A friend started eardrops with her 13-year-old pug 3 weeks ago. The pug Lucy was almost deaf from ear infection after ear infection, on constant drops, had seizures. After 3 weeks on CD eardrops, went to the vet, no ear infection, hearing is improving!! Winning!!

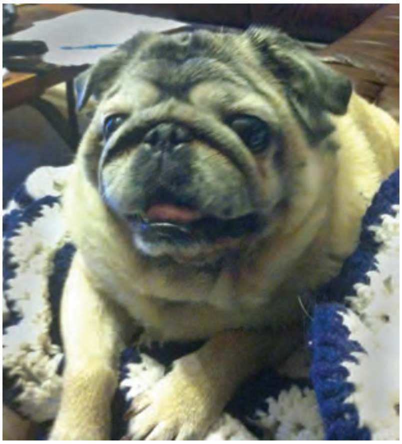

Since I was a child, I had been anemic and reading the book "Healing the symptoms known as autism" on the Parasite chapter, I saw that anemia is a sign of parasite infection. Since my son is taking CD I also took CD to know how my son feels and maybe cure my anemia. I had no problem going up to 24 drops in less than 2 weeks. I stopped taking my iron pill after 1 month of taking CD. I also did parasite protocol as I do not want to reinfection my son. Since, I stopped taking my iron pill I never had any dizzy spells and I feel great.

I used to wake up at night too to use the bathroom, now with CD no more night waking. In 2009, I was informed by my Doctor that I need to go and see a specialist as there was some suspicious stuff in my Pap Test and if not removed it *can lead to cervical cancer. I was booked in the hospital for the procedure. After taking CD reaching full dose, I had been discharging this horrible smell, like rotten egg, whenever I urinate. It took about a month before it stopped. Whatever that was, CD cleansed me and I think that discharge is related to my diagnosis on the pap test result. I am very grateful for CD for giving me my health back. I am still continuing taking CD as I want to be healthy and I will also try to complete the Parasite protocol for 12 months, I just finished my 5th PP and have been seeing lots of parasites mainly pinworm and ascaris. Now, I know the root cause of my anemia especially. I would encourage people to try the Protocol and see where the healing leads you. Thanks to Kerri Rivera, Jim Humble, Andreas Kalcker and people who used the CD before me as you gave me courage to use the CD.*

—G.M, Canada.

Two days before our first set of cd bottles arrived (we bought it to start Kerri's protocol with our autistic daughter), our son (6) got sick with scarlet fever. We didn´t give him antibiotics because we just finished our research on cd and were excited to use it instead. My sons face was swollen and red, he had a moderate fever and rash all over his body. After the first dose of 1 activated drop of cd within one hour his swollen red face subsided completely and in the evening the fever was gone. His rash got dry and after 3-4 days it was not visible anymore. Our physician told us that scarlet fever without antibiotics lasts up to three weeks. We were done within 4 days, without damaging his gut flora, as it would have using antibiotics. CD has taken care of many infections in our family since then and most important - brought our daughters ATEC down from 100 to 46 in nine months. I´m beyond grateful for its discovery.

For the past six years, my life has been plagued with anxiety/fear, depression, pressure in my chest, making it difficult to lay flat and sleep, bloating which caused me to look 7 months pregnant and which made my abdomen very tight and uncomfortable. Headaches, weight gain, sore joints, no energy, fibromyalgia pain, and dizziness causing me to feel drugged and off balance.

The anxiety and fear were debilitating literally making it impossible for me to sleep. I would go night after night not sleeping, just dead on my feet suffering with so much anxiety and very frightened wondering what in the world was wrong with me.

When Kerri's book, Healing The Symptoms Known As Autism came into our lives, I read it wanting to help or autistic son. I soon realized I had so many of the symptoms that I needed to do the protocol.

I started on the cd taking a very small amount, 2 drops every hour, about 12 doses per day, and worked up to between 60 and 90 drops a day, I also took cd baths and did enemas. I increased the cd every day by just a drop or two. Very soon I started to calm down significantly. I started sleeping! I started losing weight! For about 5 weeks I thought I had died and gone to heaven. In two months I lost 20 lbs.

But when my anxiety started coming back I set up a consultation with Kerri. I learned I needed to begin the parasite protocol to kill the parasites. The cd just killed the little ones.

With Kerri's help in several consultations I was able to tweak the protocol in her book and apply it to me for my personal needs. I started on low doses and increased it for what I needed to sleep.

As of today I have only been taking Mebendazole, the medication that kills parasites, for 3 and a half weeks. I sleep at night. Some nights are a little choppy but I get 4-5 hours of sleep. I am not depressed. My emotions are not all over the place. I am coping. I'm functioning without sever headaches, only mild ones. I still have bloating but not as severe. I feel like I will live and that I am not dying anymore. I hardly ever feel depressed.

To date I take 200mg Mebendazole at breakfast, lunch, dinner and bedtime, plus a 200mg implant right before I get into bed. The implant has been the turning point in my healing. (Consult with Kerri for directions regarding the implant.) I also take 2 neem tea capsules with each meal. Also, one dose of Combantrin everyday at lunch. I take castor oil as needed each day 1-2 Tablespoons to keep things moving, plus a cd bath at bedtime, 150 drops soaking for 20 minutes.

Trying to differentiate what is die-off and what are actual symptoms from the parasites is challenging. Pay close attention to how you feel after you do or take something. After an hour note how you feel. Worse, better? You have to listen to your body. Consult with Kerri often and keep her informed by email. You may have to tweak the protocol every few days so write everything down, day, time, how much, how you feel etc.

All of my symptoms are not gone 100% but they have greatly improved. The anxiety, which was the worst and most debilitating symptom, is the greatest improvement I have had. I know I will be able to heal completely and get my life back. I have just begun!! I have had so much progress in just a few short weeks that I still cannot believe it!! And...I have lost 2 more pounds!!

Thank you Kerri for all of your love, support and guidance!! I could not have attempted this without your knowledge and help!! You are an angel!!!! I AM HEALING!!!!!!

Your devoted friend, Marcie

My mother 84 yrs of age- who has dementia was very, very serious last week. Her lungs were filled with mucous and she could not spit it out. We thought this was the end and my sister invited the priest to bless her. The doctors suggested that I place her in a hospice, but I refused! They also offered antibiotics, but I bought some CDS and gave it to her every hour. I placed a humidifier with 40 drops of activated CD by her bedside day and night. And massaged different parts of her body with CD and DMSO. soaked her feet just once in CD as it is difficult. She cant sit straight. To everyone's surprise she made it through. Just a little bit more to clear up! Thank you everyone for sharing wonderful tips on loving and caring for our loved ones! This is a wonderful extended family! love the CD!!!!!!!!!!

* * *

I want to thank all of those who were kind enough to share their personal stories with us. There are many more, but you get the idea. I am so excited for you to read them, and I know that some of you will recognize yourselves here. It just goes to show that healing challenging illnesses is possible. We have a public Facebook group called *CD Health* where adults using the protocol for themselves can share questions, doubts and success stories. I highly recommend that you check it out if you are looking for support or some guidance getting started.

Appendix 1

More Miracles & Testimonials...

The only limits you have are the limits you believe.
~ Wayne Dyer

To date, 115 children have lost their diagnosis of autism by using CD and some or all of this Protocol. Autism IS curable!

This section is comprised of testimonials, miracles, and short blurbs on gains made by children using The Protocol. In many cases, these are candid emails from parents to Kerri sharing their joy at their child healing. These were gathered over the past 3 years, and have come from families all over the world. Some were translated by the parents of the child to English, and some were translated by my sister or I. We have done our best to honor these exactly how they came to us, unless a correction was needed for clarification, these were not edited. Many are anonymous for the protection of the families and their children. Names were used when permission was given. These families were generous enough to share their experiences with all of us to spread hope and joy, as well as to prove that children all over the world of all ages are healing from the symptoms know as autism. Please enjoy and share.

Please note that many of these testimonials were collected some time ago and therefore some contributors use the term MMS instead of CD. Due to time constraints it was impossible to request permission to change the acronym of MMS to CD in their testimonials. As we mentioned previously, MMS and CD are the same substance. Where the acronym MMS is used in the following testimonials please know that chlorine dioxide is the substance responsible for the healing.

We have added many new testimonials since the first edition came out. Those new testimonials are located at the end of chapter 2 and continue on the next page, followed by those appearing in the first edition.

(continued from Chapter 2, page 40)

45. *Things I wish I'd known before starting this protocol:*
    - * *My son would drop juice and pick up a water regimen immediately.*
    - * *My son would sleep through the night every night.*
    - * *My son would not be constipated again.*
    - * *My sons ATEC score would be halved in 7 weeks.*
    - * *My son's speech would be at age level.*
    - * *My son wouldn't need his daily rx med again.*
    - * *My son would be excited about and participate in his birthday celebration.*
    - * *My son would come into the bathroom and say, "I'm ready to wash my butt" unprompted.*
    - * *My son would drink ocean water.*
    - * *My son would get acclimated to school and would initiate conversation with his peers.*
    - * *My son's dark circles under his eyes would disappear. He would start to look and act like a healthy boy, and others would take notice.*
    - * *My son would be affectionate, tell me he loves me, crack jokes, and be connected to me like never before.*
    - * *The people who opposed our diet and protocol choices would be unable to deny that this works and we're on the right track.*
    - * *My marriage would be saved by rolling our sleeves up and tackling this together.*
    - * *I would find a way to heal my own ailments.*

    The future is bright

---

46. *My son with autism is 31 years old. Although he is fully verbal, he had a lot of issues: SIB, Aggression, he was not able to sleep. We have been trying different kinds of protocols since he was 3 years old. Unsuccessfully. I heard about Kerri Rivera and the CD Protocol last September. We started right away. We were continuously in touch with Kerri who kept giving us the faith, that even at that old age it is possible to heal and even recover. In September, when we stared CD he had an ATEC of 87, 4 months later we measured 63, and now, after 8 mounts, he has an ATEC of 28. There are some SIB's and aggression left, but it happens only rarely. If we could get rid of it, my son would be recovered. He became an interested young man, his memory improved. He keeps giving proper answers to my questions, something he never did before. I can see his real personality now: he became a nice, lovable, open-minded young man from a very sick raging, roaring, screaming, suffering boy. There are no words to say thank you to Kerri to show us the end of the tunnel. I think I'm on the best way to recover my 31-year-old son!*

    Mark, 31, Hungary

---

47. *Dear Kerri, I am very grateful to you for your benevolent service of helping to recover kids with Autism like myself. You have really improved the quality of lives of many hopeless kids. I really feel the difference now that my ATEC score is a*

I and I am still making inroads in executive functioning and social skills. Without someone like you, we would have not known about CD and it would have taken several years to cure me solely with chelation. By that time, I would have been out of high school and probably in a community college taking rudimentary general education courses and would have had a bleak future ahead of me. But now, I am doing well in school and am taking college level courses. And I am not the only one in my family that has had benefitted from using CD. My brother used to have mild seizures in the form of head banging and most of that has now stopped. My father has used it because he contracted Lyme's disease twice. And my grandparents have used it to treat Alzheimer's, Diabetes, Blood pressure and Hepatitis C. So again, I cannot thank you enough for everything you have done that has not only changed my life forever but has made a profound difference in the lives of others afflicted with Autism.

---

48. My seventeen-year-old son was a healthy happy baby until he received vaccinations at age three. The day after the vaccinations he began having horrible anxiety problems. Day by day he lost skills and regressed. He developed problems with fine and gross motor skills, and social skills, memory and sensory integration. He also had multiple learning disorders and struggled in school. He spent all of his time in his room by himself and he did not make eye contact.

We tried many interventions to try and help him. Biomedical interventions, antibiotics, therapies homeopathy and Biofeedback all had limited benefits for my son. The only thing that helped him some was a gluten free, casein free, sugar free diet. While reading Healing the Symptoms Known as Autism we learned that most of the interventions we tried did not work because we were feeding the parasites that were causing my son's autism with vitamins and supplements.

While I was watching videos about Chlorine Dioxide I came across a video of Kerri Rivera talking about the protocol she uses with autistic children. As I watched the video everything began to make sense and I knew the protocol would help my son.

We bought Kerri's book Healing the Symptoms Known as Autism, gathered the supplies and my whole family started the protocol. To our amazement we saw improvements in my son overnight. We had taken the ATEC test for my son right before we started the protocol and it was a 43. In just one month of Kerri's protocol my son's ATEC dropped to a 5! We will continue with the protocol for the recommended time and we expect to see my son's ATEC drop to zero before we finish.

His social skills, motor skills, thinking, behavior, attention span and stimming have all improved immensely. His sleep patterns, digestion, physical appearance and weight have also improved tremendously.

Kerri Rivera's book gave us the tools we needed to reach in and pull my son away from the grip that autism had on him. We will forever be grateful to Kerri. She devotes her life to helping children with autism and always has an open line of communication for anyone who needs help. She is the voice for so many children that don't have a voice. Thank you from the bottom of my heart!

---

49. FUA!!!! BIG moment... crazy morning, ughhh we miss the Bus I was so mad because my other daughter had preschool at 9:30 so my morning was a mess, I had to leave in Pj's and make them eat breakfast in the car....but then in our way to school my son saw a School Bus passing by our side and said "school Bus is gone, L is sad" OMG OMG OMG......did he just said this 30 words sentence?????

(actually 7 words but seems like 30 to me!!!) Ok...WAIT then I get to school not mad anymore, I walked with him still in shock, I couldn't believe what I just heard....and when he saw the line of kids by the wall he point at them and said "my friends" OMG OMG OMG....ok WAIT....then we get there and he takes the teacher hands, gives me a kiss and said "bye mommy" OMG OMG OMG .... WAIT then the teacher ask him L, remember back pack by the wall, so he goes to the wall with his back pack in his back and does exactly what he was ask to do!!!! O M G

FUA...THAT'S RIGHT.... 223 days in this Protocol, searching poop, dosing, enemas, writing in his food diary every little thing he eats, therapies ABA, speech, occupational, hours hours of work, hard work...all just to witness this miracle of healing...to see him think, react, analyze a situation, be happy, spontaneous and smart by himself (no prompt or reinforcer necessary)....THANK YOU Kerri Rivera since the first day I found you in YouTube I have never stopped believing....I LOVE YOU WITH ALL MY HEART!!!! YES COMADREEEEE!!!!! I can feel it...we are close...it's happening.... so happy I missed the Bus this morning... this is the best day in my son's way to recovery so far!!!!!

HOPE FOR ALL!!!!

50. My son is 23 and we have tried many, many approaches to heal him over the years. Typically we would see some initial progress, but then the rate would slow down and he would plateau again, slipping further and further behind. Lyme disease and co-infections, as well as various viral, fungal and metabolic issues complicated everything. We had to resort to psych meds, even, for his depression, anxiety and mood disregulation.

We started Kerri's protocol in the summer of 2012, and had to cut the number of drops again and again to something he could tolerate, since he has MTHFR genetic issues with detoxing, until he was at 1/64th of a drop per dose. By October 2012, he was at full dose and after two weeks, began to talk about learning to drive, to offer to do a chore at home, to draw again. So it has been about a year for us, and the improvements are continuing, his ATEC score went from 77 pre-meds to 11 now, and our lives are all much better thanks to this book and Kerri's dedication to the cause of healing the children. Please, if you are reading this, don't be afraid to try one more thing!

*51. My son started Kerri's protocol 10 months ago. At the time we started the protocol he was doing a home based therapy program for 60 hours a week. We did not send him to school because he was not able to handle all of the outside stimulus, unable to follow directions, and unable to communicate effectively. He stimmed most of his day and only spoke in 2-3 word sentences to get needs met only. When he started the protocol his ATEC was 57. It quickly dropped and the improvements were mind blowing. In 5 months time his ATEC was a 9. Today at 10 months after starting the protocol his ATEC is a FIVE!!!! He now attends school with out any kind of extra support and exceptions made for him. He is truly just one of the other children. I have been volunteering in his school and I have had 5 different office people tell me that he stops and greets them on his way to different specials (p.e, library, art, gardening). They told me he initiates the greetings and says: hello, good morning, buenas noches, and all with a big smile on his face. I have even been told by 2 of them that he greets them by name. Did I mention that he leaves the classroom on his own to go to specials??? Yep, my one time *severely autistic* child finds his name and places it in that specials category* *basket, leaves the classroom and goes to specials on his own???? When he is done with specials he goes back to his classroom and does his work. Holy CRaP!!!! A dream come true.*

Thank You Kerri!!

---

52. *My 6 year old son's ATEC went from 45 to 22!!! Wow I knew he was doing better but more then half!!*

    Thanks Kerri Rivera!!

---

53. *I don't know what to chalk it up to--we just finished our 3rd PP and did our 2nd shot of GcMAF on Saturday--but my 9 y/o son has started taking pictures with my iPhone. They are not just random shots; he is finding objects or angles and taking pictures. They are FASCINATING! He has also figured out how to use the reverse feature, and has taken a series of self-portraits, which he flips through and says to himself "Who's that?"*

---

54. *My 16-year-old son is doing great on your CD protocol. ATEC down from 89 to 37. Best thing that ever happened after doing 8 years biomed with him.*

---

55. *We started CD with my precious 5 year old 'O-ster' in June 2013. Some of the completely new developments we have seen since we started....... He is aware of how to burp and toot, and thinks it is funny. He is apologizing for things he does wrong. Even a day later, he will remember and apologize. He goes to the bathroom on his own. He is completely pottytrained, even at night. He does household chores happily. He sent people on my contact list emails that said 'I love my mommy.' He has better problem-solving, more motivation to draw pictures. If he gets messy, he just goes to the bathroom and washes his hands and wipes off his shirt by himself without saying anything. He dresses himself. He does somersaults (just a few months ago, it was impossible for him to even get in somersault position), can open baby gates, can get in crab walk position, can jump off of furniture with both feet. He greets us when we come home and is excited to see us. He MISSES us when we are gone and talks about it during the day. He's making up 4 line rhymes. He cares what he wears and picks out his clothes! He is not running back and forth as much. We have normal trips to restaurants. I can leave our front door open now, and not worry that he will run away. He tries to help put our dog in his crate and fetches with him. He wants to come with me if I leave the house to go somewhere. He will now eat almost anything I put on his plate. He used to vomit at the mere thought of things like GF pasta or rice, and he wouldn't allow any food on his plate to touch another type of food on his plate. He will now eat bowls full of tomato sauce, onions, burger, rice, garlic and peppers all mixed together. Our family can eat dinner together now. I no longer have to make separate meals. He even eats raw peppers! He yells 'watch me!' and does things like diving onto the bed or jumping off of something. He looks at cloud formations to see what pictures he sees in them. Last week he heard a toddler crying in Target. He wondered why the child was crying and wanted to go find her. He misses his little sister when she is not at home. He wants her to play with us when we are playing. He is completely aware of what is going on now. He is much more aware of danger. He notices the weather. There is a sense of calmness about him that he did not have before. His tantrums are much less severe and much less frequent. I was supposed to be his aide this year in preschool, but I have been able to just drop him off at school. There are many more new things that he is saying and thinking and doing.*

Our quick biomed history- He regressed after vaccines at 12 months. Starting at 18 months of age, he was gfcfsf for 1 year, then scd for almost 3 years. It feels like we have tried everything with supplements over the last 4 years. Some things helped a little, but nothing really tackled his pathogen load, even though he was on an awesome diet. His ATEC was 66 in Dec 2010. Low dose chelation brought it down to 33 in 3 months. The ATEC stayed at 33 for 2.5 years. We did 112 rounds/weekends of low dose chelation during that time. ATEC June 2013 was 34 (before starting CD), July 2013 was 30, October 2013 was 20! We are on our way down. This is the first time that I have physical proof that we are getting rid of the pathogens that have controlled and poisoned his body and brain for most of his life. It feels so good to finally help him. I am hoping to write our complete recovery story in Kerri's next book. I thank God that we saw Kerri at AutismOne.

56. *Nick texted me yesterday for the first time, on his own. "Hi from Nick" - that's a 2013 FUA fo sho. – Alison*

    (this text was accompanied by an ATEC drop from 69 at 6 years old to 14 at 8.5 years)

57. *Yesterday my son told me I worried too much, I said you worry too much Too Noah! He responded, my hair is red, yours is gray, so who is the worrier, hahaaa! The windshield wipers keep wiping away the autism, what's staying around it seems, is completely lucid speech and unbelievable comebacks! He never could say things like that before, not at all, he really was in there somewhere.*

58. *Something really beautiful is going on in our household right now...Caleb is totally blossoming socially on every level. There is a calmness in his body. He is alert, responding to his name, sharing his toys with other kids, wanting to give affection without prompting, soothing Lilia when she is crying rather than batting at her. We even have gotten a few spontaneous silly phrases!!!! Oh baby, oh baby, oh baby!!! We are on our way!!!!!*

59. *We started the diet on the 1th of March, CD on the following week. Adam started to ask questions in April. Since then he keeps asking continuously…His ATEC dropped down from 49 to 38.*

    Tunde, Hungary

60. *OMG OMG OMG!!!!! My son is singing!!! He was ZERO verbal... I mean ZERO!! His shadow teacher told me a week ago that he started to mimic, I said to myself "yeah he was babbling and she thought he is mimicking!..." Yesterday she told me he started to sing and I said to myself, "Yeah right, she is absolutely illusionist or exaggerating!!!" She knew I will not believe her because he is just humming at home! So she sent me a video which was a shock for us. He is singing the song "A is for apple, a a apple... B is for ball, b b ball... C is for cat, c c cat... D is for dog, d d dog..." I don't have enough words to thank you and thank Kerri Rivera and this incredible protocol!!! Thank you for saving my child. xoxo*

61. *...My son is doing his first CD PP. We picked him up from preschool he was extremely happy. Hugs & kisses. Then...runs up to a classmate and touches his shoulder and says, "bye Johnny", runs to another classmate and says, "bye Kevin." Then he skips to the car. These improvements in his socialization is new. We are so pleased and I wanted to pass along these social gains with you all. His atec at start is 58.*

62. *My little sweetpea is almost 4...and completely NON VERBAL...upon picking her up from therapy today, her teacher said she was having a great last few days... and both therapists said her eye-contact has been different...like in a really good way!!! I've also noticed her looking at me...straight into my eyes for what seems like forever...almost as if she has never seen me before?! They also said that she has been looking at them and has been opening her mouth trying to make sounds...as if she wants to say something very badly...but can't...but with a lot of prompting...she did say "Hi!"...all I know is that somewhere in there is this amazing person waiting to be let out! I often worry if we are ever gonna make it...finances and healing your family can be very overwhelming...especially on one income. But for the love of my daughter we just cannot give up...feeling very hopeful and blessed today...and very grateful that something is finally working.*

---

63. *Hi Kerri,*

I wanted to write and thank you for discovering and sharing your protocol. This protocol has saved my almost 5-year-old son.

My son suddenly started having absence seizures that would freeze his breath. From the day he had his first noticeable seizure, he was hospitalized every seven weeks for uncontrollable status epilepticas, a life threatening condition. They would pump him with seizure meds, Ativan, and Valium for hours before the seizures would finally stop.

Since starting CD five months ago my son has not been hospitalized once. He went ten weeks seizure free. He had a few mild seizures that he brought himself out of without any extra medications. He currently broke this record and has made it another thirteen and a half weeks seizure free. I true miracle.

Everything we searched for to treat his Lyme disease in the past would either cause seizures or not help to prevent them and I now know why. As you predicted after starting your protocols we discovered that he not only has Lyme, he is also incredibly loaded with parasites. So much so, that it scares me to think of what may have happened to him if we didn't find you.

I'm so amazed to watch him subtly emerge. Since we started, he went from saying about 20 words to saying 69 words, along with two and three word combinations, and over the past month he has begun to answer simple questions. His vestibular system has gotten much better; he'll actually swing. He just started jumping off curbs. His eye contact is perfect. He's gained two pounds in the past three months, something he hasn't' done since being diagnosed with failure to thrive at a year old. He is playing with his three year old sister; it's so adorable to watch them chase each other around the house, hug, chase each other, hug, get in the dog crate together, hug, take turns pushing each other in their wagon, and kick the ball back and forth. Last night the three of us played catch together for over 10 minutes. I'm so happy for her to have her brother to play with.

This protocol changed his life, my life, and our family's life.

There are no words to thank you enough for all your help along the way – he's on his way…

Love, Love, Love

Update: as of December 3rd, 2013 this little guy is 18 weeks seizure free!

64. *My name is Dawn. Below is a shortened version of what we've been through prior to and since Kerri Rivera coming into our lives and changing our family forever.*

I am Noah and Eli's mom, caregiver, advocate etc., that's been my job for 17 years. Prior to having children with special needs I had a career that I loved and my plan was to continue my career, how deluded I was, my life has been forever changed by my children and their symptoms as my entire existence over these years has been to save my son.

My son Noah born August 19, 1996 has struggled and suffered with what is now known as the symptoms of autism. His symptoms began at 8 months of age with a rare blood disorder known as neutropenia, he had a low white blood count. As he continues to improve, I too, along with our entire family seem to be healing as we bore witness to his daily struggles of self injurious behaviors, aggressions and rages, learning disabilities, speech and auditory delays, 2 steps forward and 4 steps back.

Healing the symptoms know as autism with the Chlorine Dioxide and Kalcker Protocol have been a cathartic exercise for me as well, no amount of therapy would have soothed my soul like witnessing my son's return from the daily brink of death.

My hopes and dreams for my son dashed with vaccines before he was old enough to speak. I'm 50 this year, however I'm older then those years, and my views and outlook on life have changed forever. Imagine fighting for my son's life against those trusted factions of our society supposedly put in place to protect us, i.e., government agencies, churches, schools, teachers, principles, school boards, doctors, hospitals. I'm changed forever by the daily battles, but somehow they are small now in comparison to the beast of the symptoms that ravaged my boy with red hair, often referred to by me as my favorite Martian. As he continues on his healing path, so do those of us who have loved him so dearly, despite his outwardly awkwardness, he is the innocent victim left to suffer for what is often referred to as the "greater good", he's responsible for so much more in my life then just the symptoms known as autism.

It was a Saturday morning in May of 2003. Noah, at the time 6 yrs. old, had slept with us because of seizures on and off through the night. Like so many times before, I was very concerned he would go 'status epileptus', a state of continuous seizures. He ate cantaloupe for breakfast that day. I remember it vividly even 11 years later. His seizures would come about every 20-25 days. First he would complain of a stomach ache, and I could hear the gurgle of his stomach. They usually lasted a few minutes and then went away. This day, the Saturday before Mother's Day, he didn't come out of the seizure. We administered his anti-seizure medication but it didn't work. We called 911, the fourth such call in Noah's brief life. When EMS arrived, Eli, then 3 years old, ran to his bedroom upstairs. I remember the technicians working on stabilizing Noah. They ripped the shirt from his body as they worked on him and carried him out to the ambulance. As I followed them to the ambulance I happened to look back to see Eli watching from his upstairs window as they placed his brother in the ambulance. He stood motionless. Darryn scooped him up and followed us as we drove swiftly through traffic to Texas Children's emergency room, about 25 miles away. I remembered thinking, why don't people heed the right of way? I looked back as they worked on Noah. His seizures continued until we arrived at the hospital. Once we arrived he was taken straight to the ER, ICU room, where *four doctors began to work on him. Darryn and Eli arrived as I stood next to the bed while the doctors worked to save his life. I took Eli out of the room and we sat in a chair looking in at what was happening. His eyes were as big as I have ever seen them and he sat perfectly silent in my lap, we colored. I overheard the doctors explain to my husband that they were having problems stopping the seizure. They wanted permission to use a drug that could stop his heart. Darryn asked if there were any other options. The doctor shook his head. He asked what they would do if the heart stopped. The paddles. So he was given the drug. No heart stoppage. His breathing became shallow and the seizures finally stopped after 6-7 straight hours. A couple of hours later Noah woke up in his hospital bed, connected to monitors and the like, in a drugged stupor. By this time my parents had arrived (like so many times before) and had taken Eli home. Darryn stood next to Noah's bed rubbing his chin nervously. Noah asked if he could use the restroom. Darryn said yes, and in the next second Noah leapt from the bed, almost a perfect swan dive, toward the floor. Luckily, Darryn was there to catch him. Noah's life was saved twice in the same day. After the drugs wore off, he was released from the hospital. Noah missed a week of school while he learned to walk again.*

I never knew the impact this had on Eli until Father's Day that same year. About a month after the seizure we planned to attend with some friends one of those workshops at Home Depot where the kids make a gift for Dad. He was excited to do the workshop. As we approached the store we noticed the local fire station had an area of the parking lot roped off for tours of ambulances and fire trucks. Eli became frightened and started crying, and began backing up, finally screaming and contorting. This from a child that hadn't given us a peep of trouble before. I had to leave our friends standing in the parking lot as my second son became inconsolable. At the time it hadn't occurred to me that he remembered the incident with his brother just a month prior. He too suffered, and sadly he was 3 years old at the time, unbeknownst to me, Eli has probably remembered more then I realize even today.

Noah had his first seizure in March 1998 at 18 months of age, shortly after a round of vaccines. (His doctors continually assured us that vaccines had nothing to do with Noah's problems, but I know otherwise.) Seizures were not Noah's only problem, however. He was diagnosed with neutropenia (the lack of neutrophils or a type of white blood cells) in 1997. He has suffered from acid reflux for as long as we can remember. He was hospitalized for 4 days at 4 years old for a roto-virus infection, something most children get through at home in a couple days. He had a cyst removed from his thyroid. His tonsils were removed due to frequent infections. In 2005, he was diagnosed with Lyme disease. In 2008 his pediatrician discovered his blood was missing carnitine, a key component for the body's metabolism. We took him to a specialist who diagnosed him with mitochondrial disease. His seizures stopped after we started giving him medication for the carnitine deficiency. But his behavioral symptoms continued and even worsened which always indicated a deeper underlying illness plaguing Noah. But what?

Early in 2013 Noah began to show some psychiatric symptoms that frightened us to the point we had to consider putting him in a home. He had frequent panic attacks (all day, minute to minute), tics, and sudden bursts of outrage over insignificant things. He often didn't make sense while speaking, and secluded himself in his bedroom. He fantasized uncontrollably about his videos, and spoke over and over about what we were going to do about the characters (not what *we would do if we were in the stories). He had repeated events of unrealistic fears. We took him to see an energy medicine specialist. The energy medicine protocol involved various supplements and dietary changes, as well as electrical pulses to adjust his body voltage (an approach similar to acupuncture). During one of our visits we learned of MMS, which this doctor indicated could be used in small doses to treat colds. Symptoms improved, but very little. But then they worsened again dramatically. We took him to his pediatrician who diagnosed PANDAS, and prescribed antibiotics. This helped but only temporarily. His psychiatrist prescribed anti-anxiety medications, with only marginal results. He was unable to attend school through all of this, and needed someone by his side constantly. His pediatrician prescribed Albenza thinking parasites may be involved (despite a negative stool test). This improved his behaviors some but only marginally. There was also the side effect of hair loss. Noah became completely bald, like a cancer patient.*

To see your son regress into psychotic behavior is something no human can ever imagine. Facing people everyday as they watched Noah regress and comment, it was a slow slip and then suddenly he was in a dive toward insanity. I felt he was slipping away. I scoured the Internet for answers. I had heard of Kerri Rivera but thought antibiotics were the only treatment for my son, like the vaccines, I had been told by doctors that the only treatment for Lyme was antibiotics. I emailed her and she answered immediately to "start the protocol." We began the CD protocol in June 2013. The protocol takes complete dedication, it requires steadfast commitment, and it's the best gift I have ever given Noah. I have learned so much about illness and it's roots through treating my son with the protocol.

Noah's behavioral symptoms started to improve immediately. He had fewer and fewer tics and outbursts. He was noticeably calmer. He now answers the phone, takes messages, and helps me scan food items at the grocery store. He even got behind the wheel of our car for a practice drive in the school parking lot. He asks a lot of questions. So much has improved in such a brief time. Prior to the CD protocol I scored Noah as a 38 on the ATEC test. Since then, he has improved to an 8. Darryn (Noah's Dad) scored Noah a 55 before we started, and two months after beginning it was a 22. It's amazing that a few brief months ago we were considering having him put him into an institution. Although we try not to think too far in advance, now we can hope our son can someday lead the normal life he was meant to.

I can't help one final story about Noah. Yesterday he asked me what autism was. We've never explained to him that he has it, because he would not have been able to understand it even if we tried. While I searched for an answer, he said "you know, Mom, Sponge Bob couldn't get his boating license because of his autism." I belted out a loud laugh and said "Ok. Noah, you know you have a little bit of autism too." And he said "Yes Mom, I know that," and smiled.

While he still has much to make up, years of life to regain, experiences to live, we're so pleased and grateful to Kerri and the others responsible for these protocols for these improvements, his life is being restored. I would encourage anyone to follow the protocols as closely as possible and do not delay. It is so fitting that the only opportunity I've ever been given to tell a little of what my family has been through would be afforded to me by the very person that actually saved my son…she gains nothing from what she has done for us, except the knowledge that she saved a boy's life that she has yet to meet, by saving my son, she has saved my family.

65. *Shay is doing so well it's surreal. In the past week the balance has shifted so that he is now recovered more of the time than not. He is having a 2-4 hour block every day or two where he is back in his old state (weeping, raging, zero-to-flipout, shrieking, obstinate, defiant, OCD, mean, potty-mouthed, etc.). The rest of the time he is lucid, has perfect mental clarity, is as funny as can be, does his math quickly, has perfect handwriting, gets along with his sister, is generous, expresses himself clearly, plays well, understands how others feel, etc. It is a mind blower Kerri.*

    Although Shay doesn't have the awareness to know when he is in the autistic state, he is now clearly aware when he is in the recovered state. He comes up to me and says quietly, "Mom, it's gone again. Do you think it will come back?" I tell him yes, that it will be like this for a while, coming and going, until one day it will be gone and it won't come back. Shay and I have been discussing which "number" he is going to be in the line of children that recover. :) He is also planning a big "Goodbye Autism" party for himself.

    Bella is getting over her meltdowns much more quickly. Sometimes I say "no" to something and then flinch, waiting for the screaming, but it's small, or sometimes she cries and gets over it within minutes. This is a big improvement for her. In her own words, she has noticed that she feels less anxious.

---

66. *We started the protocol 9 months ago. My son had 2, 3 horrible days every month. He wouldn't stop crying all day long, and he could not get any sleep at night. After talking to Kerri, we realized that these terrible days were always the days of the full moon. We saw improvement straight after the first PP. After 6 months, all of the bad behaviors were gone. I realized that the PP helped my child to improve month by month.*

    And to prove that it is not only me who can see the changes, my son's therapist from school told me the other day: "Well, when is full moon again? So that you would start the Mebendazol? I love when he gets it, he is doing so well with it".

    Staring atec in February was 50, now his atec is 34. Thank you so much, Kerri!

    Betti, Hungary

---

67. *We are celebrating 3 months on CD and 2 PP's done! Our little angel has (or I should say had) extreme ADHD...we are not seeing much of that anymore. Our life altering issue is her lungs. This child has undergone over 35 surgeries in 8 years. 17 of those to surgically remove mucous from her lungs. She is hospitalized every 6 weeks with pneumonia. We have O2 in our home. She even had a port placed for IV meds at home. Our life has revolved around illness. In 8 years the docs have done every test in the book to figure out the cause. We've turned up nothing. In July, we flew to NYC to see a medical mystery doctor. He told us it's all nutritional and had us do some parasite testing. The testing turned up zilch, but thanks to Jean Marie, we ended up here. Long story short, we are making huge strides. We are not there yet, but to give you an idea, normally her pneumonias take 4 months to clear, this time it was only one month! She's only needed one antibiotic in 3 months, normally it's 8 antibiotics in 3 months. She's doing better in school too, which is just a bonus. Today she came home and told me she was able to run in gym, something she hasn't done for years! Thank you Kerri Rivera and Jean Marie and all the mods for welcoming us to this group and for helping us to take back this child's life. Bless it! Bless all of you!*

68. *30 days of gratitude, Day 4...I am thankful for an evening alone with my son. It's rare that we have time together alone. He jumped on the bed and danced around while I packed for my trip in the morning, talking and singing the whole time. The dinner table was lonely by ourselves so we decided to eat in front of the fire. I said, "This is fun, isn't it?" and he said, "I love this and I love you, Mom" Moments like this were non-existent before our journey brought us to Kerri Rivera who is an angel on earth. She gave us the direction we needed to heal our son.*

---

69. *I observed my kiddos at school today. Get this........my once severely autistic 6 year old choose and received a Spanish lesson that lasted 27.5 minutes (I timed it). The whole lesson he was sitting on his bottom, following instructions, answer questions correctly (in Spanish), smiling, great eye contact, and really enjoying his time with the Spanish teacher. FUA to the max! This time last year Henry was speaking in 1-2 word sentences to get needs met. Now he is talking in complete sentences in English and Spanish!!!!!!!*

---

70. *In the past, if I gave Nate a hug and told him I loved him, he would sort of hug me back, awkwardly jutting his backside out (to reduce physical contact) and patting me on the back and saying, "I love you, too." He would then escape as soon as possible.*

    After a few days of being on 7 drops (last week) Nate approached me with full eye contact, threw his arms around me, gave me a prolonged bear hug and spontaneously said, "I love you, Mom". I hugged him back and we stood there, just enjoying the moment, while I quietly balled my eyes out.

    Thirteen years. I had waited 13 years for that moment.

    And now he has been doing it almost every day since that first hug.

    This alone is the world.

---

71. *There's a lot of people jumping on the "Thankfulness Train" this time of year. Here's my contribution. There's only one major thing I can think of this year that I am thankful for. Of course, there's so much to be thankful for, but one thing that really stands out. That would be this protocol. God led us to see Kerri Rivera at Autism One this year. I only went because it was basically the one thing I didn't know much about. The day before Kerri presented, I saw my friend Laura at Karaoke Night (so fun by the way!). She told me she was doing CD and I'm pretty sure I looked at her like she was crazy! Saw Kerri the next day and I was hooked. We started the protocol June 3rd. Laura got me in this group. Thank you Laura, for your patience with me and for letting me know about this group! Thank you, Debbie who has talked to me many times about the protocol, about my son's aggression, reassuring me that things will get better and they have! Much, much better. We still have further to go, but there is light at the end of the tunnel. My four kids and myself are doing the protocol. We all have Lyme. We are all feeling so much better. We are no longer spending thousands of dollars a month on empty treatments. We are getting our lives back; our finances under control, and my son with autism went from an ATEC of 47 to 20 in four months. Thank you and love you all. And to Kerri Rivera, there are no words. Thanks and love isn't enough. You can be assured of my prayers for you, Patrick and your family for life. That's the least I can do for all you have done for all of us.*

72. *Small progress this week! My son had his ABA therapy yesterday and at the end the therapist gives him a "high five" before he leaves. The therapist asked my 3-year-old normally non-verbal son for a "high five". My son looked at him, gave him a high five and said "HIGH FIVE!" plain as day. My husband almost fell over sideways in shock! He also put his toys away without being asked to, and that too is a big change. CD is the best!"*

---

73. *We have been on the protocol since Aug 30th of this year. We've done one PP. My DS is 8 with severe apraxia and autistic.*

    ATEC: 4/2013: 62 starting some biomed

    ATEC: 8/15/2013: 51 before CD

    ATEC: 10/6/13: 24

    Some WOW moments for us so far:

    * *Playing soccer with other kids at school three weeks after the start of the protocol. *Really "on" in conversations...adding information when we didn't think he was listening because he was playing on his iPad.*

    * *Talking about other things in his day besides Minecraft or Plants Vs Zombies.*

    * *Going from a battle with homework that would last over three hours to picking him up at daycare with his homework already finished. And correct. That he did on his own.*

    * *His ST saying he is able to follow three step directions better and was able to talk/ask Who and What questions for nearly the entire hour...I actually had to tell the ST what we are doing because it was obvious to her it was more than just diet.*

    He's just happier. We're all just happier.

---

74. *We came back from our Intensive in December of last year on a high. After a week with the Son-Rise® team, Jordan was saying 20 different words; he was more confident, and more mature. He had such an amazing time learning and growing. Every morning he would take his facilitator by the hand and walk him/her into the playroom. He really grew up that week. I loved watching him grow through love and laughter. What an amazing, wonderful thing to witness.*

    We, also, learned at our IAHP appointment that Jordan was reading at about a third grade level. What a trip! The evaluator held up a paragraph for Jordan to read. It was so fast I didn't even have time to read it. Then, she held up a question for him to read, then, a sheet with the answers. She, then, asked him to point to the answer and he did. He was three for three. My son taught himself to read. What?! We were, also, very surprised to find out that between running our Son-Rise® program and incorporating exercises from IAHP, Jordan improved 433% from September to December. WOW!

    Between the two experiences on the east coast, I felt like I didn't know my own son. I always believed he can accomplish anything and yet I had tremendously underestimated his ability. I was so excited to get home and start his program. I knew that if we focused and put everything into his therapy for two years, he would be ready to go to mainstream kindergarten.

    I was so excited; I was beyond hopeful; I believed he was on his road to healing.

    On Christmas Eve, Jordan was crouching down and just froze there. I called his *name, no response. I ran over to him and noticed his lips were turning blue. Immediately, I tipped him over, hit him on the back and did a swipe in his mouth to get out what he was choking on. There was nothing there, but he gasped for air and began breathing again. Did he just have a seizure?*

January 15 has become a day I will never forget. We woke up to Jordan throwing up and he didn't want to get out of bed. He just wanted to sleep; he must have caught the flu and needed to sleep it off? Finally, around 10am he got up and ran around the house like nothing was wrong so we began therapy. While I was in the playroom with him we were jumping on the trampoline and he suddenly stopped and looked at me with a panicked look on his face and reached out for me. Clearly, jumping was too active for his upset stomach so I brought him to the family room with Kaiyan and my helper.

Suddenly, he looked off and froze again. He then threw up again and passed out. Oh shit, he just had a seizure, and another one, and another one, and another one. All the while staying unconscious. I brought him to the ER where he slept and had another mild seizure. While he was passed out they gave him a CT head scan to make sure there wasn't a bleed or something in his brain. Everything was fine.

Shortly after, he woke up. He was playing, running around the ER as if nothing had happened. He went through what little water I had brought and was asking for food and water. A good sign, I thought. The ER wouldn't allow him to eat or drink anything "just in case". Anyone who knows anything about seizures knows that no food and drink could very well cause a seizure…and it did.

This time the ER Doctor went into full-blown panic mode, probably because she saw what happened. They began trying to get blood out of him. On a good day this is almost impossible. Now, they are trying on a kid that didn't have any fluids in his system to help expand the veins.

And so it went, needle poke, dig around in the arm searching for a vein, Jordan would hold his breath from the pain and have a seizure. Seizures are typically the brains way of trying to get oxygen; the pinpricking was inadvertently inducing seizures because of the breath holding. This went on for two hours: needle poke, dig, hold his breath, seizure. They poked him 27 times: in his neck, head, ankles, thigh, anywhere they could think of.

This of course wasn't enough. They insisted on doing a spinal tap to check for meningitis, which we had been declining all day. Now they are telling us that the way he had been seizing for the past two hours is classic for meningitis and not doing it could be fatal. They crunched him in the tightest ball they possibly could and held him so tight he couldn't move…or breath. After the fifth poke, I began to get faint and had to sit down. They didn't know what they were doing. When I sat down I could really see how they were holding him and asked if he was breathing. The nurse looked down and said, "he's blue, get the oxygen" and they put oxygen near his face.

I immediately thought, "they're killing him. They don't know what they're doing." That's when I reached out to every FB group I'm on and asked for prayers. That was the only thing that was going to save him.

When they, finally, let him go he had the worst seizure he's had to date. He was actually convulsing. As they went for blood the 28th time, I told the nurse that's enough. This has gone on far too long. She agreed, ran and got the Dr., they wheeled him into to the CPR room, and called for a crash cart. I think I said, "what the fuck?" She said it's "just in case".

They proceeded to get a drill and drill a hole through his shin and into his shinbone so they could put in an IO (IV into the bone). They pumped him full of all kinds of drugs and, at long last, drew blood.

He looked up at me with a longing in his eyes and reached out for me. I went over and put my hand on his chest and he passed out. He was transferred by ambulance to our local children's hospital; we were allowed to follow the ambulance, but couldn't ride with him. He remained unconscious until he arrived into his private room at the hospital.

The first thing he did when he woke up was reach out for my hand and give me a kiss. My little boy!

The blood they had finally taken revealed low electrolytes and bicarbonate at a life-threatening low level. Screaming and crying can drop bicarbonate, which he did for two hours in the ER; lack of food and drink can drop bicarbonate levels, which he wasn't allowed to have in the ER. Low bicarbonate levels can cause seizures. What a recipe for disaster!

The hospital ran every test they could think of and no reason was found for the low bicarbonate or the seizures. We came home after five days, followed by several follow up appointments over the next two weeks. Everything seemed to check out fine until I brought his labs to his DAN (Defeat Autism Now) Dr., she looked at the blood test, pointed to something and said, "Lyme, every time I see this in one of my kids, it's Lyme."

We had the (one of many) $800 test at the best lab in hopes we wouldn't get a false negative, which is very common with Lyme. Jordan came back positive! Now we had something to go on. I was excited to finally have an answer to what was going on with my little boy. Lyme is known for seizures and abnormal electrolyte levels.

We scheduled an appointment with a Lyme Literate Medical Doctor (LLMD) and were ready to start healing our guy. Two days before the appointment, we woke up to Jordan vomiting. He didn't want to get out of bed. He finally did and ran around like nothing was wrong and then he had another seizure. We were told to take him for blood tests immediately the next time he had a seizure. We took him to his DAN Dr. an hour away, but knew they would be able to get his blood without difficulty. We went to lunch, came home, he had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, passed out. His DAN Dr. just happened to call to check on him in the middle of this and told us to get him to the hospital before he had a heart attack. We went to the ER...

And so the story goes every seven weeks TO THE DAY, seizure, turn blue, vomit, pass out, seizure, turn blue, vomit, pass out, admitted to the hospital where he was pumped full of drugs for hours to try and stop the seizures, and they ran the same tests over and over again. The hospital didn't think he had Lyme even with a positive test and thought that seizures every seven weeks to the day was a coincidence, although, they could still not offer any explanation.

We tried treating him with ABX that the LLMD prescribed. One teaspoon and he had a seizure, turned blue, threw up, passed out, had a seizure, turned blue, threw up, and passed out. This time the Valium I administered did stop the seizure cycle, most likely because it was induced from the ABX and not part of his regular seven week seizure cycle. He had multiple mild seizures every day for two weeks from that one-teaspoon.

Clearly, ABX were not going to work. I emailed with a woman who coauthored the book, "The Lyme Autism Connection" and tried the homeopath that she used down in San Diego. This was definitely working with detoxing, but it wasn't helping with the seizures. He went from seven weeks to the day to three weeks to the day. This treatment was working on opening Jordan's detox pathways so his body could rid itself of the Lyme. I liked the idea of treating this way, but the process could take four years. I didn't want Jordan having what the neurologist told us were life-threatening seizures for four years. I was desperate to find something to help him.

We knew what was wrong, but needed something safe to treat him. Nothing was covered by insurance and we were running out of money. I had lost hope. Every treatment seemed to make him worse and I couldn't bear to see my boy suffer the way he had been for the past six months. I didn't know what to do.

I, literally, got on my knees and prayed for help every night. In the same week, a woman who attended the same Son-Rise® start-up emailed me about a newly discovered treatment she was doing. I had signed up for a Lyme conference that mentioned the same treatment, and I had signed up for an autism telesummit that unbeknownst to me had Kerri Rivera as one of the speakers. (Kerri is the woman who discovered this treatment for autistic kids.) I emailed back and forth with Kerri, I posted on the FB group, trying to get all my questions answered. I was terrified to treat Jordan with anything given his seizure history.

Then, he was hospitalized again for uncontrollable seizures and I realized I couldn't let fear get in the way of helping him. I started the day we were home from the hospital at a really low dose that Kerri recommended. She has held my hand through this process to help prevent him from having any seizures.

Since we started in May, he went 10 weeks seizure free and had several mild seizures that he brought himself out of without any extra medication and NO hospital. He is currently at a record-breaking 15 weeks seizure free.

Through this treatment we have learned that in addition to Lyme he has three different visible parasites. He is loaded—he passes more parasites than most of the teenagers in the FB group. It scares me to think what would have happened to him if we weren't led to this treatment. Thank God for Kerri Rivera sharing what she has learned and being so gracious and generous with her time.

I'm incredibly grateful we attended the Son-Rise® Intensive before all of this began. His team of facilitators created a base for him that continued to blossom even through this tumultuous year. Coupled with the parasite treatment, Jordan has gone from 20 to 90 words—he's picking up new words almost daily. His balance is amazing, his muscle tone is getting better, and he understands everything. Even more exciting, he's playing interactively with his sister. He is on the road to recovery.

We are finally able to start picking up the pieces of where we left off back in December. I felt stuck; I had PTSD; I needed help believing we can move past (what I had deemed) the horror of this year. The first thing I did was turn to Son-Rise®; I started having monthly dialogue sessions with Bears (Barry Neil Kaufman)—the co-founder of Son-Rise®, which has been incredible. This is one of those times that happen in life that can make you a better person or break you. Bears is guiding me to be better. I just arrived back from another Son-Rise® training where I realized that staying in fear and being nervous about getting excited or anxious about feeling hopeful again will prevent me from being the *bridge to healing that I want to be for Jordan. I felt a relaxation that I haven't felt all year. I've been wound real tight, waiting, being on the look out for that next seizure.*

This year has had a huge impact on Kaiyan as well. She'd see her brother sick and then her mommies would disappear. I'm noticing she's wound pretty tight as well. Her favorite moments are when the entire family is together. I'm hoping talking with her about what happened and spending more one on one time with her will help. She asks daily is Jordan throwing up, is he having a seizure? She doesn't really know what it all means, but is really aware it's not good. She, recently, said she wanted to have autism. After I explained to her that autism is why Jordan isn't talking or riding a bike, she decided she didn't want autism. She wants to be a boy and insists on wearing Jordan's clothes almost daily. I think it's her way of healing and feeling included in the process.

We've been doing Son-Rise® again and it feels great. I do feel like we have some time to make up for since this year has been spent dealing with health issues rather than cognitive issues so we are working up to having a full time program, having quarterly consultations, and doing another Intensive.

Son-Rise® and Kerri's protocol's are not only bringing our son back, they're bringing me back.

---

75. *Healing From Autism*

Hospital birth, bottle fed, fully vax until age 4. Digestive disorders beginning at age 1. Low tone, speech and language delay. Mild early education interventions. GF since age 2, GF, Soy free, sugar free since age 4. Majority of autism symptoms presented in 2013; in the fall of 2012, our son did not qualify as being in need of special education, and thus, could not attend school in our district at all.

Long winter days in a tiny home had me sending him out to play, in our tiny backyard, which had a patch of grass and a bush, under which small animals and birds lived. In his boredom, he sprinkled the dirt all over himself; his head, face, clothes, shoes, on several different days.

Later in the early spring, the symptoms of autism appeared more and more obvious: disobedience, disinterest in sitting for a story or coloring, lost interest in puzzles of more than 24 pcs, jumping on and off furniture for hours, despite scoldings and time-outs. Sudden relapse in bedwetting. Then by summer, the symptoms were very obvious: longer tantrums, rages, anger issues, hitting, fighting, a mean spirit; running out of the house without permission, with no respect for danger or regard for safety. Weepiness, hypersensitivity to touch, suddenly repeating the same phrase, over and over all day. Lost balance and coordination, bumping into walls and furniture, stubbing toes. Shoulder shrugging, all day, but not able to stop or say why he was doing it. Lost dexterity—could not put cards in a box. Lost organizational skills—could no longer put toys or clothes away. Extreme nighttime drooling and bedwetting. The need for sameness, and to have all food served cold. He frequently did not seem to be in his right mind.

In desperation, I asked a group of special needs moms for advice, and ordered a handful of books on Autism, including Healing the Symptoms of Autism by Kerri Rivera.

It was the best money I ever spent.

I read the book thoroughly, ordered the supplies and began the CD Protocol on July 8, 2013. Starting ATEC: 46. We ramped up slowly on drops, serving doses 8 times per day. My son has had two low dose/partial PP's, to minimize the emotional meltdowns, the gas and discomfort. I did not realize how very deficient in minerals he was, and did not give him enough ocean water. Now I order the big 10-liter jug of ocean water, so I am not tempted to be stingy. He needs more than 100 ml ocean water per day for his digestive system to handle the parasite dumping during the PP. His previous digestive trouble was unexplained chronic constipation, since age one, and that is really linked to his high need for the OW. Interestingly, his doctors had all said to keep giving Miralax, and that he might possibly need that laxative every day for the rest of his life. We stopped the laxative when we started the CD program, and began the enemas right away. When I sifted the stool and saw the number of parasites, I understood finally why he couldn't poop well on his own for the last 4 years, despite diet interventions. We are looking forward

to PP #3, in which I hope to ramp up to full dose on all components.

Poop sifting revealed what 6 pediatricians and 2 GI specialists could not tell us. Our son is infected with at least 6 different kinds of parasitic flukes and 3 kinds of parasitic worms. His stool has had a very large amount of sediment, with odd, tiny, but defined shapes. A search for "rodent and avian parasites" reveals photos of very tiny creatures.

I just completed his ATEC at Kerri's request: November 12, 2013: he now scores a 10, with a 1 in language, 2 in social, and 7 in health.

What does his recovery look like? We enrolled our son in a small private school this fall, with one teacher for 6 students, no aide. He now memorizes poems, Bible verses, and songs. He sings, knows his alphabet and all the sounds of the letters, and can count to 47 with precision. He shows kindness to others, is mindful of safety, and has glowing reports from his NT school. He sits with me for stories, and brings books to 'read' to his baby sister. He is not literate quite yet, but enjoys word search games, where he can complete levels and advance to more challenging puzzles. Twice now he has played the word search game on Leap Frog for 45 minutes. He has regained his balance and coordination. The rages, anger, and shoulder shrugging disappeared within the first 10 days on CD. He identifies printed numbers and letters, and does well in math, phonics, penmanship and reading class. He is learning to write numbers and letters in cursive, and follow directions on drawing worksheets. He operates the TV and DVR unaided, recording and watching programs he chooses.

He is happy; he has a patience and joyfulness about life that was previously lacking. He tells me jokes, and now has the coordination to operate a hand puppet while speaking for the puppet!! He is thoughtful and obedient. He has regained fine and gross motor skills, and motor planning skills.

There are still improvements we hope for: further speech, language, and vocabulary improvements. Social graces. More improved digestive health; I can see that he still has parasites in his stool. We will continue the CD Protocol, and we look forward to further healing. Thank you all for your support and encouragement.

The CD Autism Program has done more than cure our son of his brief affliction with Autism; it addresses the pervasive, underlying gut disorder and parasite issues that have plagued him all his life. Where complex testing in expensive medical centers has failed, a simple home health program succeeds. Thank you Kerri Rivera!!

76. *Well...we too did an ATEC! Our son was 11.5 when we started the CD Protocol July 2012 First fully loaded (All PP meds on board) was November 2012*

    VERY First ATEC: Jan. 2006= 124

    ATEC in July 2012: 64

    ATEC August 2012: 52

    ATEC Oct. 2012: 48

    ATEC Dec. 2012: 45

    ATEC May 2013: 40

    ATEC August 2013: 39

    ATEC NOV. 2013: 33

    I actually did one myself as I have done throughout all the previous, but I had hubby do one and combined and divided by the 2! Resulting in the 33! Oddly, it's the "SPEECH" that remains consistently around 14 & 15 throughout ALL of these! All the others continue to go down! We've had some ups and downs, but with continued evacuation of some MIGHTY big worms...we continue to see major gains!

    This is the ONLY Protocol that has EVER given us the progress we've seen in one year's time let alone in 6 months! Thank you Kerri Rivera and especially Patrick Rivera for being such pioneers in this journey to recovery! You are SOOO loved and appreciated beyond words! Thank you for never giving up! Thank you for giving so much, and bringing healing to our son! He's come a long way, but we have a ways to go! But with the progress he's making...I honestly feel it will seem like nothing from where we've come!

    Again THANK YOU! XOXOXO!

---

77. *We had a great day today. For the first time, he joined me in throwing something at a target (blocks at carpet circles). He laughed and smiled and pulled me to play with him better than he has in weeks. I heard a lot more sounds and approximations today. When I couldn't figure out what he was saying, he kept trying until I did.*

    My favorite part of the day, and I know this is odd, was when his head collided with my mouth. I exclaimed in pain and he was clearly worried/empathetic. He kept trying to get close to see my lip and giving me kisses. I've seen him concerned before but never that clearly empathetic.

---

78. *Just wanted to share that on 21/8/13 my sons ATEC was 46, just redid because we're at the start of our first pp & his score has come down to 31! The big improvement came after introducing CD enemas 3 weeks ago. Although he's been on oral CD for about 2 months, it was always the case that if I missed a dose within 1/2 hour he'd be sensory, inflexible, super-anxious, shouting, aggressive etc & I'd have to syringe it down him but within 10mins of dosing the lightswitch would flick on & he'd be in control & a calm, loving boy again. Since the 3 enemas I gave him 3 weeks ago (only 3 because then his brother had one, panicked, screamed & totally freaked out my 4yr old who up til then had said he liked them bc they made him feel good afterwards) his personality has been shining through again, there's less panick/anxiety, less sensory issues & there's no 'reaction' if I miss a dose or more of CD. I've seen more improvements this week, on Monday he decided to sit down & colour & proceeded to colour neatly,*

inside the lines - he was so pleased with himself! (He used to colour so neatly but about 6 months ago his colouring became scribbly & he frustrated that he couldn't do it neat anymore). Then tonight, when I asked him to hop out of the bath, HE DID! He stood up, stepped out & as I handed him a towel he took it & said 'thank you mummy' and wrapped himself in it & walked out - all completely relaxed - like it was usual!!! WWWOOAHAHH! Yes! He has never done this, it's always tantrums, always arguing, needs to bathe longer, has to be made to get out, refuses to step out, has to be lifted, needs me or hubby to wrap & dress him, & usually all while panicking, crying, demanding, arguing or screaming because it's cold, different, he doesn't want to, it's not right etc. I just can't believe what happened tonight! I'm still astounded. It's like the Finlay we 'know' isn't Finlay & now we're beginning to see who he really is!

So I'm REALLY excited about our 1st pp. We're already seeing full moon behaviours & although tough it's been great to see them because it's a reminder of what Finlay was like every day until 3 weeks ago!

He's already healing. It's soooo exciting! Thank you Kerri - may God bless you & all the children on this protocol. Healing, here we come!

---

79. *True Hope by Maggie Kaye*

Most people don't know that babies are born with about 100 million brain cells. That is more cells with more connections than there are stars in the galaxy! This is difficult to picture when you consider an infant's brain only weighs about 12 ounces at birth, about 25% of it's full size. But, there is a logical explanation for this: Brain cells; like the brain itself, have yet to grow. As neurons (brain cells) mature, they become larger, stronger, and form tentacle-like branches that form synapsis (connection and communication) with other brain cells. Brain growth is directed by neural activity. The neural activity of the cerebral cortex (the largest, most complex part of the brain) molds the way a child will think, behave, and acheive.

Conventional wisdom long held the belief that the human brain cannot change… that it is hard-wired at birth. Scientists today have proven that this is not the case. They have found that the brain is malleable. It has the ability to change shape, size, number of branches, number of connections, and even the strength of it's connections over time. The potential growth, or neuroplastisity (neuro means neurons or brain cells, plastic meaning changable) is far-reaching. At birth, the brain is simply a blueprint of what it will become. No one can predict the amount of growth that a brain has yet to acheive.

For the parents of a child with Autism, the diagnosis can be devastating. But, the biggest and most harmful blow can come from the prognosis: There is no cure. Kenn and I were given a diagnosis of an Autism Spectrum Disorder for our son, but we were not given this prognosis. In fact, the word "Autism" was avoided at our son's diagnostic evaluation until the therapists heard us say it first, and realized we were comfortable with the word. For many parents, a diagnosis is the end all, be all. For far too many parents, the words "No cure" are scary and harmful. We did not hear these words nor would we have accepted them as truth if we had. We do not want to change our son. We want him to be his best self. We are not looking for a cure. We are in search of treatment and *healing for his symptoms. We know and understand that a Therapist is not God. A Therapist cannot predict the future for our son. He or she can simply assess where Gunnar is developmentally and give us options for a treatment plan for a healthy, successful future. Our son's diagnosis was not an ending. It was simply the begininning of a journey through a world of strength, challenges, support, opposition, and the love of a parent that cannot ever be fully defined.*

Our son, Gunnar was born with an Autoimmune disorder, Hypothyroidism. He was tested at birth, was diagnosed, and has been treated by an Endocrinologist since he was three days old. We were not upset by the Hypo diagnosis, but rather ready for the course of action to regulate all of Gunnar's bodily systems. What did surprise us was the constant shrugging off of our concerns over our son's developmental delays and gastrointestinal symptoms over the last few years. We were continually disappointed by the compartmentalization of our son's physical and mental challenges by every Pediatrician and Specialist we saw. To us, Gunnar's symptoms all had to be as connected as are all of his body parts. But, we were treated as if our concerns were dumbfounded…as if digestive, behavioral, and physical challenges were not intertwined.

Let's be clear...The brain does not function in isolation. It is a team player. It needs vital nutrients and informational input to perform at optimum levels. To fill these needs, the brain relies heavily on complex interactions between the immune, endocrine, and gastrointestinal systems. Gunnar, like so many other children on the Autism spectrum, suffers from gastrointestinal issues uncluding chronic constipation, diarrhea, and abdominal discomfort. Scientific studies and countless parent reports are pointing to gut inflammation caused by casein, gluten, and soy in our foods. Gunnar has shown clear physical and behavioral reactions to these foods, and improvement in these areas with the removal of Gluten, casein, and soy from his diet. This is not an allergy from the perspective of a traditional allergist. This is what is called a T Cell Inflammatory response.

The use of a gluten, casein, and soy free diet as a means of treatment for Autism symptoms has gotten mixed responses. Many parents and doctors who deal with Autism, Immune, and Endocrine disorders rave about it's benefits, including me and my husband. Still, others call it "junk science". Clearly, if someone is looking purely at one body part at a time, they will not be open to the truth that the body and brain are codependent. I, personally, have been challenged by the choice to follow this diet for my own thyroid disease and for my son's Autism. I was even told on an online support forum that the GFCFSF diet is "fake science" and is causing "false hope" for parents of children on the Autism Spectrum.

False Hope?

This may be one of the biggest oxymorons I've ever heard. This ranks right up there with "God Hates". I cannot say what causes people to be so oppositional to our choice to follow this eating plan. Perhaps the resistance to this way of eating is due to a simple misunderstanding of the leaky gut syndrome that plagues so many children on the Autism Spectrum. This leaky gut syndrome does not cause food allergies. It causes food sensitivities.

In food allergies that many parents are famililar with, the symptoms arise immediately in the form of hives, watery eyes, sneezing, or difficulty breathing. They can even produce life-threatening anaphylactic shock. Only about 10% of *autistic kids have this type of allergy. Some 85% of children with Autism have food sensitivities. Rather than immediate physical symptoms, these sensitivities produce and inflammatory response that results in more subtle mental and behavioral symptoms that can take anywhere from 6-72 hours to appear. The brain's primary fuels are oxygen and glucose which are manufactured from nutrients in our food supply. It only makes sense that we get out of our brains what we put into our bodies.*

When it comes to the treatment of any neurological disorder, many parents and doctors focus solely on the obvious cognitive impairments and fail to realize that correcting their child or patient's underlying intestinal imbalances can; and often does, lead to significant overall improvement. Leaky gut syndrome which is so prevalant in children with Autism Spectrum Disorders causes a long list of vitamin and mineral deficiencies because the inflammation process damages the various carrier proteins normally present in healthy GI systems. Our plan is to treat our son from the inside out; Reduce the inflammation, test for vitamin and mineral deficiencies, provide necessary supplements, test for toxin poisoning and remove if necessary, provide therapy as needed, and set Gunnar up for the best possible outcome. When you feel better, you perform better. Who can argue with this rationale?

Many people have differing opinions on what does and doesn't work for families facing Autism. There is a saying in the Autism community, "If you have met one child with Autism, you have met one child with Autism" There is a broad spectrum with severe to mild physical and behavioral symptoms. One person's treatment protocol might not work for another person. I am not judging what others choose to do to treat their child, I am simply sharing what has worked for us thus far. I am not searching for or promising a cure, I am finding a treatment for and healing the symptoms that affect our son. Whether or not you agree with how my family and I choose to approach this journey does not change the outcome. If we have helped one person understand this complex neurological disorder, we have helped spread understanding, and that makes us happy. And, if you are a family facing Autism and don't know which avenue of treatment to pursue first, our advice is: Go with your gut.

---

80. Recovery #107

Having a sick child is a full time job. Even when they are well, your mind is filled with ways to keep them from getting sick again, ways to improve their current state, ways to help others who were where you were just a short time ago. It is my hope to do all three.

Autism entered my life a few years ago, but I didn't recognize it. My son, Gunnar was three when my husband and I realized he wasn't progressing as he should. He wasn't asking why, what, where, or when questions. He wasn't growing. He was lining up his trains and cars, screaming if we touched one of them. Gunnar was oppositional and refused to go into a crowded place. He didn't look anyone in the eye, and it seemed to make him uncomfortable to try. He had several sensory issues such as covering his ears when music played, even when the volume was low. He had an aversion to touch and would often yell, "ouch!" when someone touched his arm softly. He was constipated. He smeared his feces on the wall. He often woke in the middle of the night screaming. My husband or I *would find him in the fetal position and he couldn't or didn't respond when we asked what hurt. Gunnar went ballistic when we tried to wash or cut his hair or brush his teeth. He was indifferent to me leaving for work, and what was most painful for me, he ran and hid when I returned. He didn't much care for his own mom and that was heart breaking for me.*

We also noticed some regression. Gunnar stopped showing interest in potty training. He went from saying, "Gotta go pee pee!" to not even recognizing he'd soiled his pants. He started refusing foods he previously enjoyed. He went from using several words in succession to not speaking more than two words strung together. He started to be extremely picky when it came to the foods he would eat and spent most of his time at the dinner table staring out of the window. We were losing our son.

Our son was born with an autoimmune condition, Hypothyroidism. He started taking medication just three days after he was born. Initially my husband and I interpreted some of our son's Autism symptoms as side effects from his medication dosage fluctuations. After all, we voiced our concerns to his Pediatric Endocrinologist, and she assured us he was fine. She said the medication was likely making him lethargic. She said boys took longer to potty train. As time went on, Gunnar's symptoms multiplied and worsened, and other caregivers took notice and shared their concerns. At our son's next Endocrinologist appointment, I insisted on a referral for a developmental evaluation.

Our son was diagnosed with an Autism Spectrum Disorder in January of 2013. He was 4 years old. I immediately went into research mode reading everything I could get my hands on that was related to healing Autism. Two weeks after his diagnosis, we started a gluten, dairy, soy free diet and saw digestive issues disappearing. The night-waking was going away. He turned toward us when we called his name. He spoke in short sentences, requesting to watch a movie or eat a snack.

A month after Gunnar's diagnosis we had further testing done to see where he was on the spectrum. When we showed up for the appointment, the diagnostic team was impressed with our son's progress. He responded when they asked him to point to specific items, and he had no trouble matching items together. I remember thinking he was much smarter than I'd realized. Clearly the problem with our son was physical.

The test results showed Gunnar's speech and language to be at a 23 month old's level with his gross and fine motor skills not much further along. We were determined to heal our son and scheduled an appointment to have all kinds of testing done including food allergies, comprehensive stool analysis, and genetic mutation tests. These results showed that our son was sick with food allergies, leaky gut caused by an overgrowth of bacteria, and pathogens. We knew what we were up against. We were on the right track with diet, but had to find out how to get rid of the bacteria and pathogens. In my research, I stumbled upon the CD forum. It was the first treatment program that made complete sense to me.

We received the book in June. It was the first book our son ever showed interest in. He kept saying, "This is it!" and, "This book is the best!" He didn't want to put the book down. My husband and I looked at each other with tear-filled eyes. We knew Gunnar was right; this was it.

We started the CD Protocol immediately...just five months after our son's diagnosis. His starting ATEC was an 80. Immediately we saw regular bowel movements, diminished dark circles under his eyes, and healthy skin coloring. Gunnar began to sleep through the night. Four days into the protocol, our son started taking turns and sharing with other kids. That night, he danced at a graduation party. Every day, his vocabulary improved.

Four weeks into the protocol, our son no longer needed his thyroid medication he'd been on his entire life. He started using the bathroom unprompted. He dressed himself unprompted. He brushed his teeth on his own. Gunnar didn't mind getting his hair cut. He began to engage in reciprocal conversation.

Seven weeks into the protocol, we saw so many behavioral improvements; we decided to take the ATEC test again. Gunnar's score was a 40. It was cut in half in seven weeks!!! Eight weeks into the protocol our son celebrated his 5th birthday. It was the first time he showed anticipation for a party. He was excited to open his gifts and thanked everyone individually. Normally, Gunnar would run and play by himself when guests came over. He had always covered his ears and yelled, "Stop singing!" at previous parties. Gunnar was happy when everyone sang Happy Birthday to him. He was healthy and present. It was one of the happiest days of our lives.

Twelve weeks into the protocol, our son started mainstream preschool. All previous attempts to take him to school resulted in anxiety and complete meltdowns. He started speaking in complete sentences and his speech was relevant to the situation. He started showing interest in books. Fifteen weeks in, our son read his own bedtime story with no help. He picked out his own Halloween costume. He brought home his artwork and proudly showed it to us. His teachers continually voiced their pleasure at his progress. His speech therapist said he was at age level for speech and language. (that's 3 years worth of progress!)

Twenty-four weeks into the protocol, our son's ATEC score is a 10. He continues to go to mainstream preschool and is interested in learning. He is reading a variety of books. He is counting in English and Spanish. He knows his US geography, and does puzzles that are advanced for his age. He asks questions. His speech is appropriate and relevant with continued improvement in vocabulary and articulation. He loves to sing, dance, and paint. Our son is smart, funny, and very affectionate. He spontaneously hugs and kisses me and tells me he loves me. And the best part of all? When I come home from work, Gunnar yells, "Moooommmmyyyy!", runs up to me, jumps into my arms, and wraps his little arms around my neck. I have my son back.

I could never sufficiently thank Kerri Rivera for all of her selfless help. She and her team of parents give freely of their time and energy to bring kids and their families out of the depths of Autism. It is my hope to pay it forward. All children deserve a healthy childhood. If you are holding this book, you are likely in the position my family was in just 10 months ago. This is the way to turn your feelings of desperation into determination. This is the way to change the future for your family. This is the way out of Autism.

—Maggie Kaye

81. *Ok, my brother is 17 and started CD at the end of october. We are on 9 drops as of today. Non verbal. My mom called me crying and my brother said "I want snack please." My mom asked if he wanted something and he "no, no, no." He got his yogurt (df) out and said "thank you." This is just so amazing. Thank you God for putting us on this path and around so many amazing warrior parents to help us with our journey. Happy healing to all!!!*

---

82. *At the age of 3 our son was absent. He talked a lot... to himself. Scripting. It was almost impossible to get him to look into our eyes when we spoke to him. He almost never answered back, and when he did, it was scripting again (from a show on TV or a movie he was into.) He played by himself. He refused to potty train. He would wake up in the middle of the night crying and holding his stomach but was unable to tell us what was wrong. He stopped eating whatever we put in front of him and seemed to only want fruit snacks, American cheese, hot dogs, chicken nuggets and fries. And thats it. He didn't seem to be growing anymore, both physically and mentally. He seemed to be slipping away from us. You could see it in his eyes. Those eyes that were pointed off in a different direction. It was gradual...hard to notice right away. But definitely a change, and not one we were happy with.*

Up until this time he was progressing like any other child. He ate well. He slept well. He played well. Was even starting to potty train. A typical child. But not anymore. Then there was the diagnosis. Autism. No parent wants to hear that, but we knew. Every parent knows what I mean. You just know. And at this point we (as parents) have to make a choice.

We can choose to do nothing. Just live with it. We can listen to our doctor, or a specialist in the field. You know, someone who claims to know everything about you and your child in the 5 minutes they spend with you. Or you can talk to other parents. Parents that spend every moment with their children. Not just a few minutes every six months. You can network. You can research. Ask whats working for them, or more importantly what's not.

We chose the latter, and because of this our son is recovering. He looks into our eyes. He asks questions. He sleeps at night. He's potty trained. He sets the dinner table with his older sister and asks to be excused when he's done eating. He's in pre-school now and plays with all of the other kids. He calls them by name. He draws. He paints. He reads books. He does all of the things a typical five year old does.

Our son is recovering, and I need to give two heartfelt thanks for this. The first to my wife for the countless hours of reading and research she did immediately following our sons Autism diagnosis. She is an incredible woman who refuses to give up on anything. Ever. She refused to believe that we had to accept what was handed to us. She knew in her heart and mind that we could make a difference. We just had to do it. Because of this determination, she found Kerri Rivera, whom is the next person I want to thank. She is a true pioneer, someone who refused to keep thinking inside the box, someone who is helping families like ours all over the world.

"We have the world at our fingertips. Every book, every encyclopedia for us to read at our local library, in our own homes, and now on (most) of our cell

phones. It's a beautiful tool for us to use. Use it. Do your own research. Talk with parents. Find out what works and do it.

The only thing I really have to say about Kerri's book and the protocol it explains is that it works. We have our son back.

—Kenn Kaye, Father of Gunnar Kaye

---

83. Last night, my 19 year old son came out into the kitchen, picked up a purple straw from the dish drainer, and pretended to play it like a clarinet with great enthusiasm. Pretend play. For the first time in 19 years.

---

84. We've been playing the "healing game" with my son for some time, (7 years 11 months old) but we've never really healed anything. To be honest we've never known what we were up against...until now. Like everyone here, my child is very sick. He doesn't grow, he has maldigestion and malabsorption issues and he's chronically constipated. His history of constipation is long and extensive. ANYTHING that disrupted his gut flora would result in this extreme constipation; whether it was a cold, a new food, treating the GI. To say he has leaky gut is a tremendous understatement. So, when we started Kerri's protocol 6/15 my biggest health markers for initial healing were growth and for him to poop daily on his own. With his history being what it is I wasn't surprised to see he was instantly constipated. How could he not be? We are killing pathogens and parasites like gang busters. We started enemas on day two and have done them daily. But, an amazing thing is happening! He is healing. He has grown two inches in these last three months and he starting to poop on his own. BTW, these aren't just lame little poops they are total TROPHY poops. They are soft, formed, have no undigested food and there is lots (I'd show you a picture but we barely know each other)

Yep... It's taken 3.5 months to get here, but his little; fragile and damaged body is doing what it is supposed to do.

Trust the process... It's not easy; it's not always fast. There will be ups and downs. But, my gosh, watching the body heal is an extraordinary thing. It's nothing short of a miracle.

---

85. Hi Kerri,

I started my son on the CD last night. I'm taking it with him.

Well, I for one, feel fabulous. My lungs are clearing up, the low-level depression I felt is evaporating. Wow!!!

I've noticed some subtle changes in my son. He is more "with us" in the world and not looking out the window at cars quite as much. Also, a little less flapping his hands.

He and I have been doing a ketogenic version of the GAPS diet (no sweet vegetables, no dairy, no sweeteners) since January of 2013, and I think that's why we've been so successful with the CD! :)

I was thinking of upping it to two drops per 8 oz bottle--one for him, one 8 oz bottle for me.

I am hoping that Sulfur is Ok to still give him. We used to cut it with Kombucha, but that is now a no-no.

Thanks and Huzzah!

86. *Ryan is the first of our 3 children to recover. Ryan was not considered autistic by most standards but did score a 19 on the ATEC test prior to starting the CD protocol. He is 9 years old and has been in mainstream classes without an IEP. We did hold him and his twin brother back one year prior to starting kindergarten so they could have more time to heal from Lyme Disease which we had been treating since they were 4.*

Ryan was a healthy baby but I did notice after saying his first words at the appropriate age, he stopped talking a few months later. (I now know that is a symptom of autism!) And both he and his brother were very late with language despite knowing their alphabet at a very young age. Ryan did eventually talk but was often hyperactive and had a hard time sitting still in class or at the dinner table. It was often hard to get his attention and to get him to listen to requests. He was constipated since he was a baby and was a very picky eater. I dreaded mealtimes because it was so hard to get him to eat. We often had to spoon-feed food into his mouth even when he was 7 years old. Treatment for Lyme Disease helped many of his symptoms but he still had some remaining symptoms such as weakness, stomachaches, seeing orange and blue spots at night (more noticeable in the dark), nightmares, bedwetting, difficulty falling asleep and early awakening, and overly emotional and irrational behavior. These symptoms worsened dramatically when our house well water became contaminated with parasites and bacteria. It took us two years to figure out that the water was the cause of our deteriorating health. During this time, Ryan became unable to play sports or have play dates without melting into irrational tantrums when things didn't go his way. When playing mini golf, he refused to finish the round after getting a score of "10" on one hole and had a crying tantrum. In tennis class, he would sit down on the court and start crying if he didn't get to hit as many balls as the other children on the court with him. We didn't even attempt to put him in team sports for fear that he would melt down and storm off, refusing to play the rest of the game.

Then we discovered CD and Kerri's protocol, and his and our lives changed forever. Within three weeks of taking CD, Ryan's ATEC dropped to an 8. This improvement was the difference in Ryan being able to play sports. He was able to play on a baseball team and while he did get angry if he struck out or made an error, I was there with a dose of CD and that would keep him from a complete meltdown. As the season continued, Ryan's emotions became more and more stable and normal. His energy increased greatly and now after 6 months on the protocol, Ryan is playing three different sports, taking piano lessons and excelling in above grade level reading and math. His ATEC is a 3 but we will continue until it is 0. His first report card since starting the protocol was the best of his life. We were most excited to see that his scores for behavior and attention were above expectations. Last year, these same scores were all below expectations because Ryan was so hyperactive and unable to focus due to the parasites. Ryan still sees a few orange spots at night but they are diminishing and we know they will be gone so long as we stick with this protocol until he is completely healed. We are so grateful that Kerri blazed the way for us all by developing this protocol. I have been searching for the past 7 years for a way to fully heal my children and the CD was the last but most significant piece of the puzzle for us. We will be forever thankful to you Kerri!

87. *I first heard about CD, then called MMS, to treat autism in May 2012 and wanted to try it out with my then 10-year-old son. I bought supplies, and did the then-current protocol for a few months, but then school ended and summer vacations started and I put CD on the back burner or did it very sporadically. My son was verbal and high functioning, but still had lots of autistic behaviors. His ATEC score in 2012 was 27.*

    I started reading about CD again in the summer of 2013. This time there was more information out there. Kerri had written her breakthrough book and lots of facebook friends in the autism community were mentioning it. I bought the book and dove in, first using my old supplies (CD with citric acid) then buying HCl and trying that. My son did very well with the new protocol, starting low and getting to full dose around the time of his 12th birthday at the end of August. I started seeing worms before we even started our first Parasite Protocol! When I started the Parasite Protocol and added in ocean water, things kept getting better! No set-backs or regressions. School was seemingly much easier for my son. He was in a mainstream class with resource room help for reading and writing for the second year, but this year, math homework was no longer taking several hours, but 20 minutes! His reading comprehension was improving as well, and more importantly, I didn't worry about hearing about behavioral issues anymore!

    I wanted to complete three months of CD before I took another ATEC. When I filled in out in November, at the start of our third Parasite Protocol, I was shocked to see a score of 7! My son had dropped 20 points from his 2012 score! Seeing the single digit score was so amazing! My son had been diagnosed with autism at age 2 ½ and there were no options given to me at the time. An early IQ test given by a speech therapist had shown he was mentally retarded and that was that. However, I looked into everything I could, and was not going to give up on my son. So glad I didn't! Nine years of autism, but the beginning of his 12th year marks the end of it! I know his ATEC score will be 0 soon!

---

88. *"thank you for your help in recovering me"*

    This is what my son just typed to me this morning on his iPad with the app we use for supportive typing!

    Thank you Kerri & Patrick for all you have done and continue to do!

    Without you BOTH, we would NEVER be in the place we are right now... HEALING & RECOVERY!

    Just know you are LOVED!

    —Melissa in Indiana

---

89. *Sophia was born on March 9, 2006. She was a typical happy baby who slept well and seldom got sick. She met all of her developmental milestones and laughed and made great eye contact. When she turned one just after her vaccines she became violently ill and vomited with fever for 2 days straight. After that she was never the same, her smile was rare and words were gone. When she was nineteen months after more shots, we took her to an ear doctor out of concern that she was not responding to her name or speaking clear words. Her hearing was tested and we were told she was fine and to give her time. After a few months of no change her pediatrician determined that she was developmentally-*

delayed and had all the symptoms of Autism. We were stunned and unprepared for this news. I went into warrior mode and ended up with a DAN dr, who started her on lots of supplements after initial lab results showed heavy metal toxicity and high amounts of intestinal yeast. We followed the program and advice of this doctor for the last 4 years with the only change being dietary fluctuations. Sophia did improve but hit a plateau where no new gains were made in the last year or so, so in May 2013 her DAN doctor suggested that we do another round of stool testing after a urine acid test showed bad bacteria overgrowth and terrible mito issues. Her comment to me was "this kid's DNA is so messed up, I haven't seen a test this bad since the 1990's." This was after thousands of dollars and too many supplements to count.I was done at that point. There were no answers there obviously. I came across this protocol on a Facebook page for autism parents and purchased Kerri's book in September 2013. I immediately knew that this was the answer for my daughter. I had a gut instinct that she would be healed . We started the diet immediately and CD just after. My daughter had been a juice addict before this and I was worried about her taking the CD and being able to drink the water. This was only an issue for a day or so. I have not missed a day of giving her 8 or more doses. We get up very early and dose her ocean water and 2 CD doses before school. On weekends we are doing 16 doses/day. In the second week we started seeing some herxheimer/die off, very sleepy followed by fevers. This is when the transformation started! I followed the fever protocol and did enemas and she got better. She also had an episode around week 7 of very red throat and fever which we double dosed for and resolved. We have seen a withdrawn, self -focused child with little interest in the world outside of herself disappear and turn into a child interested in friendship with peers, reading books and understanding the story (huge!), and having relevant conversation with great eye contact. She gets all A's at school and has made a complete turnaround with behavior. We have so much hope for her future and she is blossoming into a beautiful, happy girl!

---

90. *Here are a few of the emails I mentioned from my son's therapist in regard to what's happening at therapy! :)*

This is an email from the exact day I restarted the CD, which was August 21 of 2013.

We discovered a new game today. He was being a turkey and running up the hall to try to get around me to get places he wasn't supposed to go. I started singing (to the tune of rolling down the river) about whatever he was doing "walking, walking, walking down the hallway" "jumping, jumping jumping down the hallway" climbing, climbing, climbing up the ladder. He thought it was hilarious, especially when I'd change in the middle of a verse "running running, jumping down the hallway.

—Your son's therapist

Great day today! Vail seems to be understanding first, then sequencing and I think I finally have a bribe for him--he loves our fish.

Also, he made a friend today. One of our other pt's was playing in the gym and he was watching him, imitating him a little and said his name 2x!

—Your son's therapist

We tried to play the game and he wasn't interested. Instead we ended up *pushing the fisher price horse, sheep, goat, and cow off the treehouse. He would "randomly" produce sounds which we took for requests for the animal associated. Then we would give him the animal "requested." Reviewed his sound cards and he produced the /n/ and /h/. Coulda sworn he said go a couple of times, and mom said she heard a whispered "home." Listen and be watchful for whispers! Oftentimes children with apraxia can get one system working but not the other, ie. articulate with no sound, or vocalize with intonation and no articulation, like when he says "ah uh oo" for "I love you." Yep, he does that!*

Cheers!

91. *Hi had to share this moment today, just wish daddy wouldn't have missed it!*

Well, we were seeing daddy off for work this evening, when Vail decided that he wanted to go run around in the grass. I told him we needed to put on socks and shoes and then we could play outside. He walked straight to the door, we walked in, I pointed at his socks and shoes, and FOR THE FIRST TIME EVER (always knew he could, just to stubborn to do it unless he really wants too!) he promptly walked over, picked up his socks and shoes and turned and handed them to me!!! (all the while doing it with an attitude!!!"lol)

This is Just from MOM to therapists. :)

CD and Kerri have been life changing for our entire family. We now can give Vail choices, usually of two items. For example I showed him two pairs of pants and asked which he wanted to wear, and he grabbed one pair and started to put them on even though he was still in his pj's. I showed him two shirts, and he picked a shirt. Same thing with his shoes and socks, he is able to choose now on his own. I look forward everyday to him doing something differently or completely new, and he almost always does now!!! This was not possible for CD!!! THANK YOU KERRI !!! WE LOVE YOU FOR SHARING AND FIGHTING THE FIGHT!!!

LOVE,

—The parents of Vail

92. *Best day of the entire school year! My son initiated verbal interactions 5 times today and responded to his teacher in a group over 5 times, WITHOUT A SINGLE PROMPT! Most days his TA is poking him to pay attention and ask other kids for directions, rather than an adult, all day long! She is tracking these interactions as they are his IEP goals.*

93. *Casey is a young man who is 17 years old today, and the estimated ATEC score from 2006 is from when he was 10 years old. Proving once again that recovery is possible at any age.*

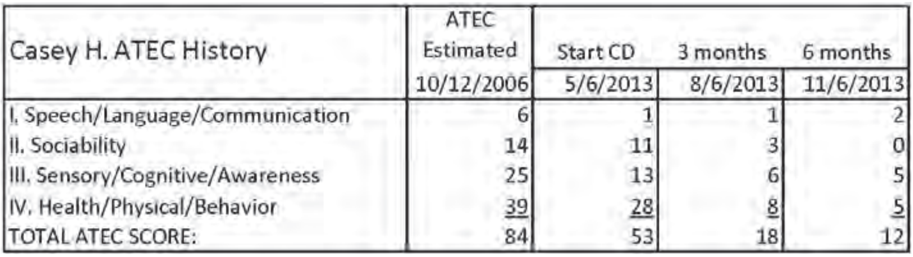

Casey H. ATEC History	ATEC Estimated	Start CD	3 months	6 months
	10/12/2006	5/6/2013	8/6/2013	11/6/2013
I. Speech/Language/Communication	6	1	1	2
II. Sociability	14	11	3	0
III. Sensory/Cognitive/Awareness	25	13	6	5
IV. Health/Physical/Behavior	39	28	8	5
TOTAL ATEC SCORE:	84	53	18	12

94. *Daniel - beginning ATEC June 2013 - 101. Today ATEC Nov 12, 2013 - 44! We are thru our 3rd PP and starting speech supplements. Speech, by far, has been the hardest thing. In all other areas he is under "10" on the ATEC! Kerri Rivera words are coming. This week alone he is saying "Hi" and "Bye" and trying to say others. I can't wait to hear "I love you Mom"! You are the Bomb Diggitty Kerri. We all love you!*

UPDATE: December 3rd, 2013

I officially want to tell the world that my son, Daniel, who has autism is nonverbal no more! He is starting to talk. Praise God! This battle has been the struggle of my life, but i have learned to trust God, be patient, and enjoy the journey! Where some take a lot for granted regarding their children, I have been waiting almost 7 yrs to hear his beautiful voice! And thank God he has one! Autism is treatable and there is hope!

---

95. *We had a super exciting trip, we went to a therapist on the 28th who teaches Rapid Prompting Method. We have tried since by daughter was diagnosed to get her to speak, but haven't had any luck yet, she is now 11. We thought we would give RPM a try to see if the therapist could get her to communicate.*

We not only found out that she is able to make correct choices, but can actually SPELL! Her very first word that she spelled was CLOUD!

Later on in the session the therapist asked her what her favorite color was... she spelled out PINK!

I was sitting on the floor in awe... For the first time in 11 years, I finally know what my girl's favorite color is... it was surreal... it was amazing... she is so smart, but had no way of letting us know... I am soooooooo proud of her!

---

96. *Diagnosed at 20 months old in August 2011*

Started supplements under Dan in September 2011

Started CD / PP on July 1, 2013, after seeing rope in stool

Starting ATEC on July 1, 2013, was 46

Present ATEC 22

Gained 5 pounds from July 1st to present!

Too many improvements to list!!!!!

Also doing GcMAF which has also helped……but, I believe GcMAF may have helped more with the viral load. I suspect this because we actually started IV GcMAF in January 2013, before CD . Huge gains with GcMAF but after 35 shots, her gut was still a mess and GcMAF definitely didn't help with the gut. Found Kerri in July 2013, and began our healing journey.

The following lab results are proof of just how well the child described above is doing since starting the protocol. Here is what her mom had to say:

> *"The original test was taken last August of 2012, which showed a deficiency in almost everything despite the fact that my daughter had been on supplements for over a year at that point. The re-test was just taken October 2013, only 4 months after CD / PP and removal of ALL supplements. Wow!"*
>
> *You can see in the second test that her high viral titers were reduced as well.*

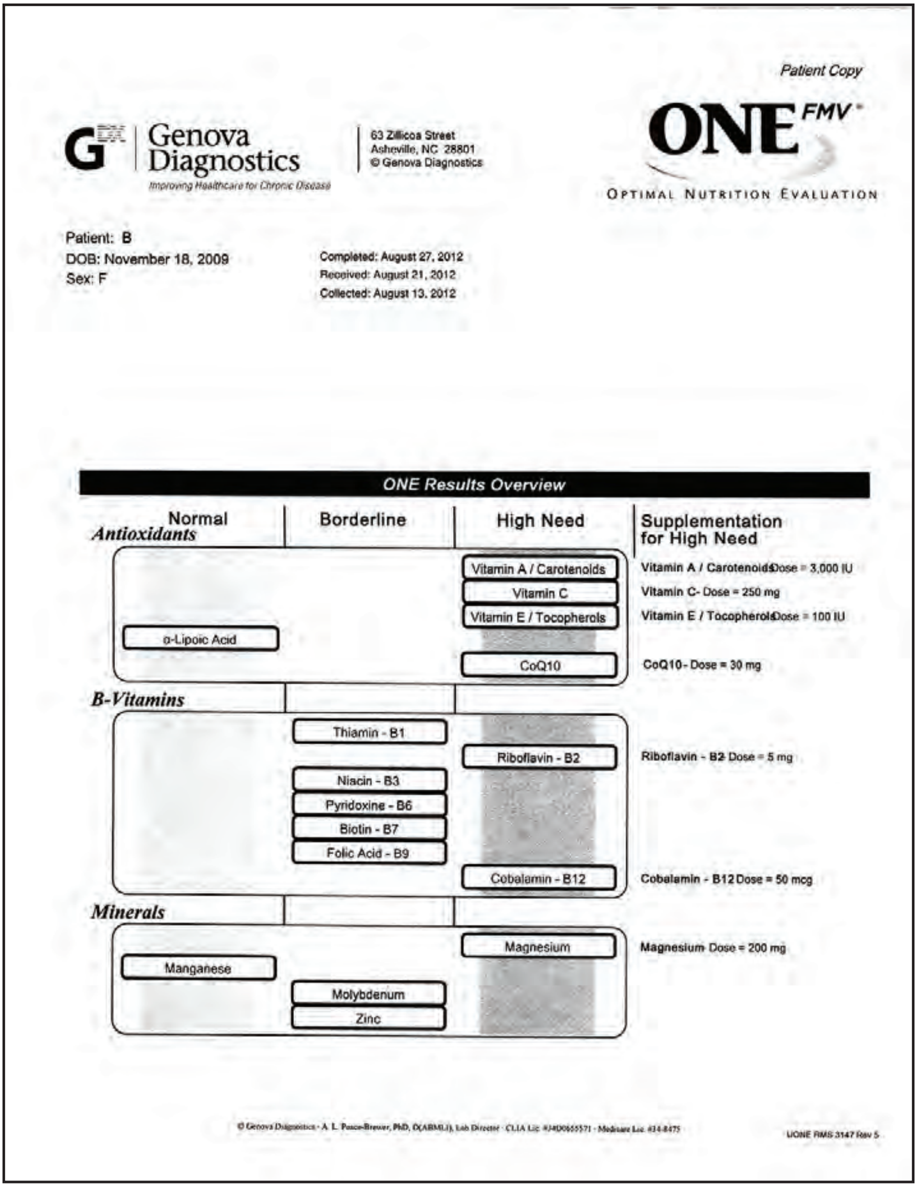

Patient Copy

Genova Diagnostics
Improving Healthcare for Chronic Disease

63 Zillicoa Street
Asheville, NC 28801
© Genova Diagnostics

ONE FMV®
OPTIMAL NUTRITION EVALUATION

Patient: B
DOB: November 18, 2009
Sex: F

Completed: August 27, 2012
Received: August 21, 2012
Collected: August 13, 2012

ONE Results Overview

	Normal	Borderline	High Need	Supplementation for High Need
Antioxidants				
			Vitamin A / Carotenoids	Vitamin A / Carotenoids Dose = 3,000 IU
			Vitamin C	Vitamin C- Dose = 250 mg
			Vitamin E / Tocopherols	Vitamin E / Tocopherols Dose = 100 IU
	α-Lipoic Acid			
			CoQ10	CoQ10- Dose = 30 mg
B-Vitamins				
		Thiamin - B1		
			Riboflavin - B2	Riboflavin - B2 Dose = 5 mg
		Niacin - B3		
		Pyridoxine - B6		
		Biotin - B7		
		Folic Acid - B9		
			Cobalamin - B12	Cobalamin - B12 Dose = 50 mcg
Minerals				
			Magnesium	Magnesium Dose = 200 mg
	Manganese			
		Molybdenum		
		Zinc		

© Genova Diagnostics · A. L. Peace-Brewer, PhD, D(ABMLI), Lab Director · CLIA Lic #34D0655571 · Medicare Lic #34-8475

UONE RMS 3147 Rev 5

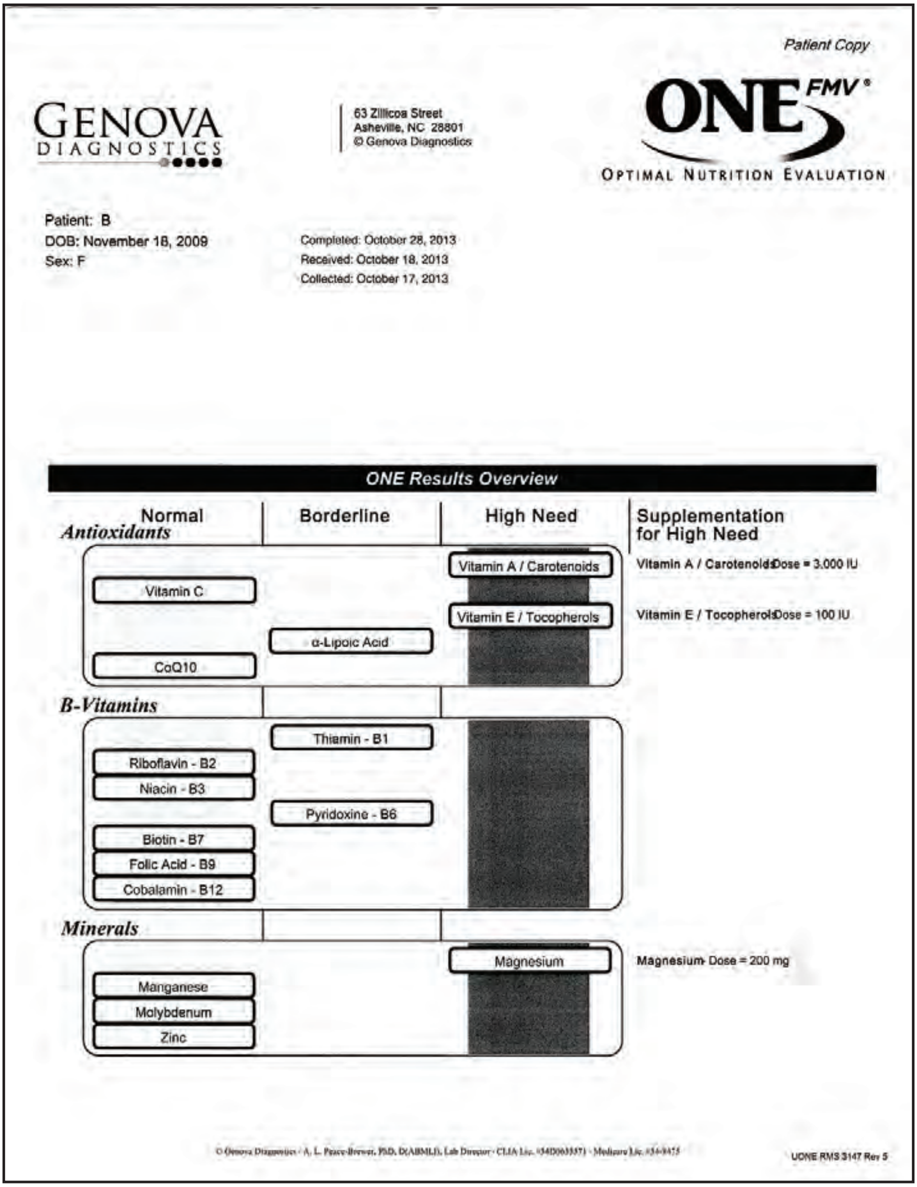

Patient Copy

GENOVA DIAGNOSTICS

63 Zillicoa Street
Asheville, NC 28801
© Genova Diagnostics

ONE FMV®
OPTIMAL NUTRITION EVALUATION

Patient: B
DOB: November 18, 2009
Sex: F

Completed: October 28, 2013
Received: October 18, 2013
Collected: October 17, 2013

ONE Results Overview

	Normal	Borderline	High Need	Supplementation for High Need
Antioxidants	Vitamin C CoQ10	α-Lipoic Acid	Vitamin A / Carotenoids Vitamin E / Tocopherols	Vitamin A / Carotenoids Dose = 3,000 IU Vitamin E / Tocopherols Dose = 100 IU
B-Vitamins	Riboflavin - B2 Niacin - B3 Biotin - B7 Folic Acid - B9 Cobalamin - B12	Thiamin - B1 Pyridoxine - B6		
Minerals	Manganese Molybdenum Zinc		Magnesium	Magnesium Dose = 200 mg

© Genova Diagnostics · A. L. Peace-Brewer, PhD, D(ABMLI), Lab Director · CLIA Lic. [illegible] · Medicare Lic. [illegible]

UONE RMS 3147 Rev 5

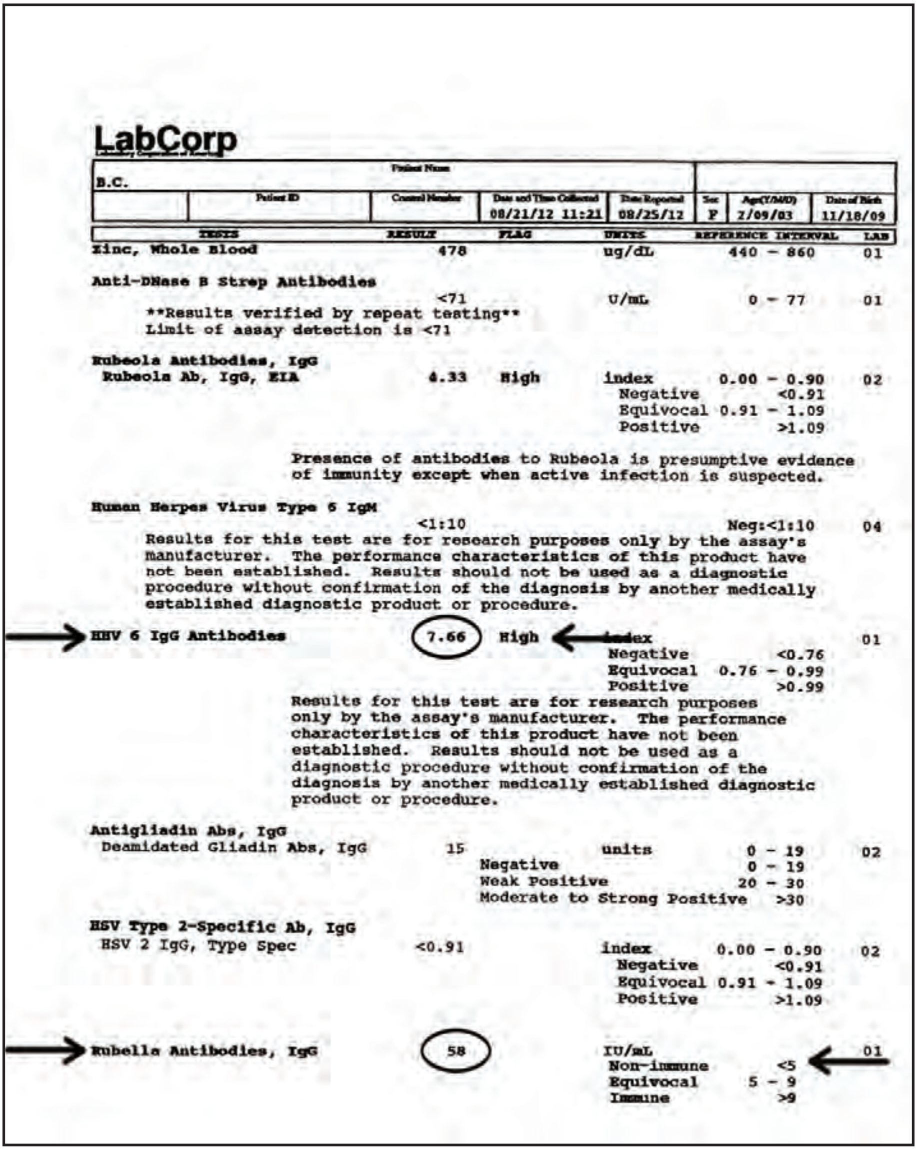

LabCorp

Patient Name							
B.C.							
	Patient ID	Control Number	Date and Time Collected	Date Reported	Sex	Age(Y/M/D)	Date of Birth
			08/21/12 11:21	08/25/12	F	2/09/03	11/18/09

```
              TESTS                  RESULT      FLAG        UNITS       REFERENCE INTERVAL    LAB
Zinc, Whole Blood                     478                    ug/dL          440 - 860          01

Anti-DNase B Strep Antibodies
                                      <71                    U/mL             0 - 77           01
     **Results verified by repeat testing**
     Limit of assay detection is <71

Rubeola Antibodies, IgG
 Rubeola Ab, IgG, EIA                 4.33       High        index       0.00 - 0.90           02
                                                               Negative          <0.91
                                                               Equivocal 0.91 - 1.09
                                                               Positive          >1.09

                   Presence of antibodies to Rubeola is presumptive evidence
                   of immunity except when active infection is suspected.

Human Herpes Virus Type 6 IgM
                                    <1:10                               Neg:<1:10          04
     Results for this test are for research purposes only by the assay's
     manufacturer.  The performance characteristics of this product have
     not been established.  Results should not be used as a diagnostic
     procedure without confirmation of the diagnosis by another medically
     established diagnostic product or procedure.

HHV 6 IgG Antibodies                  7.66       High        index                             01
                                                             Negative            <0.76
                                                             Equivocal  0.76 - 0.99
                                                             Positive            >0.99
                   Results for this test are for research purposes
                   only by the assay's manufacturer.  The performance
                   characteristics of this product have not been
                   established.  Results should not be used as a
                   diagnostic procedure without confirmation of the
                   diagnosis by another medically established diagnostic
                   product or procedure.

Antigliadin Abs, IgG
 Deamidated Gliadin Abs, IgG           15                    units           0 - 19            02
                                             Negative                        0 - 19
                                             Weak Positive                  20 - 30
                                             Moderate to Strong Positive      >30

HSV Type 2-Specific Ab, IgG
 HSV 2 IgG, Type Spec                <0.91                   index       0.00 - 0.90           02
                                                               Negative          <0.91
                                                               Equivocal 0.91 - 1.09
                                                               Positive          >1.09

Rubella Antibodies, IgG                58                    IU/mL                             01
                                                             Non-immune        <5
                                                             Equivocal      5 - 9
                                                             Immune            >9
```

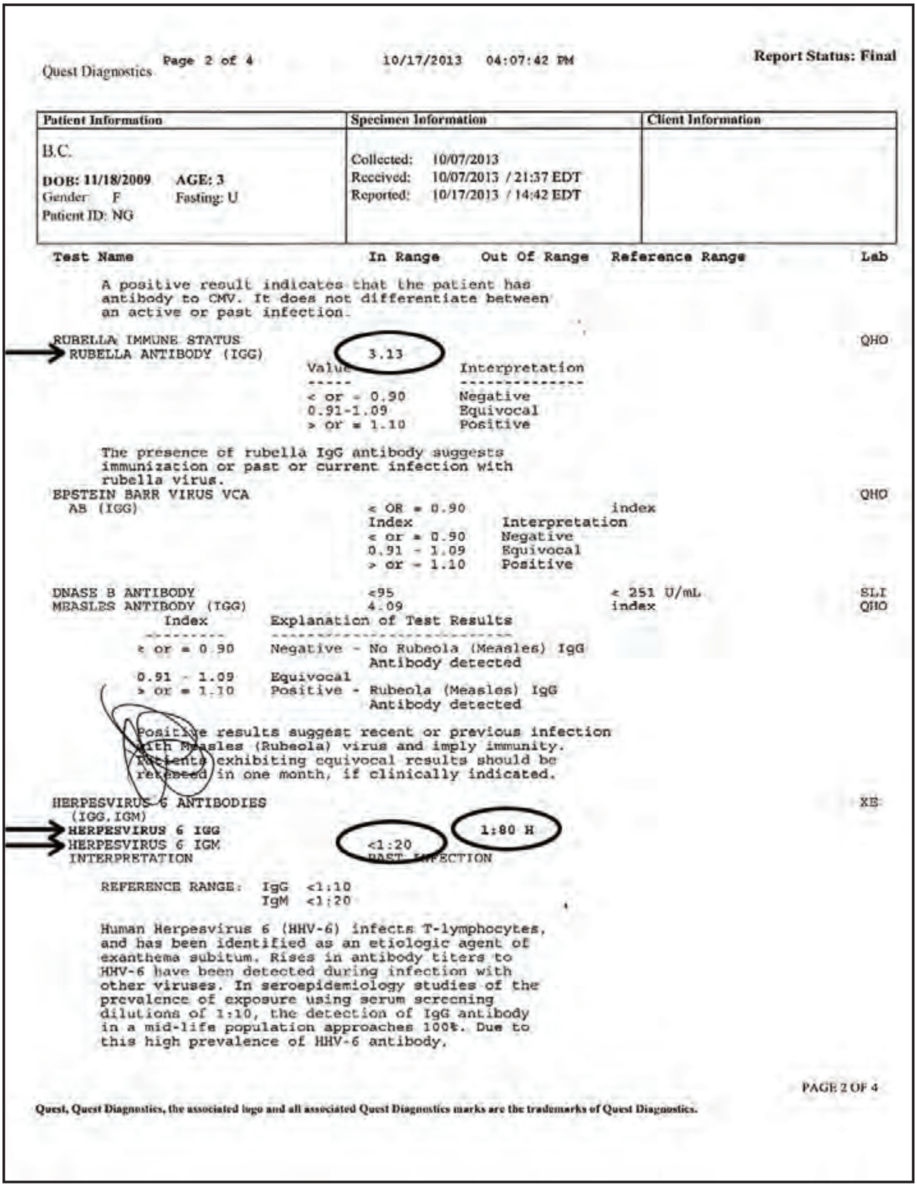

Quest Diagnostics Page 2 of 4 10/17/2013 04:07:42 PM **Report Status: Final**

Patient Information	**Specimen Information**	**Client Information**
B.C. **DOB: 11/18/2009** **AGE: 3** Gender: F Fasting: U Patient ID: NG	Collected: 10/07/2013 Received: 10/07/2013 / 21:37 EDT Reported: 10/17/2013 / 14:42 EDT	

```
Test Name                         In Range     Out Of Range  Reference Range      Lab

       A positive result indicates that the patient has
       antibody to CMV. It does not differentiate between
       an active or past infection.

RUBELLA IMMUNE STATUS                                                              QHO
  RUBELLA ANTIBODY (IGG)                3.13
                             Value              Interpretation
                             -----              --------------
                             < or = 0.90        Negative
                             0.91-1.09          Equivocal
                             > or = 1.10        Positive

       The presence of rubella IgG antibody suggests
       immunization or past or current infection with
       rubella virus.
EPSTEIN BARR VIRUS VCA                                                             QHO
  AB (IGG)                              < OR = 0.90                  index
                                        Index        Interpretation
                                        < or = 0.90  Negative
                                        0.91 - 1.09  Equivocal
                                        > or = 1.10  Positive

DNASE B ANTIBODY                        <95                          < 251 U/mL    SLI
MEASLES ANTIBODY (IGG)                  4.09                         index         QHO
             Index           Explanation of Test Results
           ---------         ----------------------------
          < or = 0.90        Negative - No Rubeola (Measles) IgG
                                        Antibody detected

          0.91 - 1.09        Equivocal
          > or = 1.10        Positive - Rubeola (Measles) IgG
                                        Antibody detected

          Positive results suggest recent or previous infection
          with Measles (Rubeola) virus and imply immunity.
          Patients exhibiting equivocal results should be
          retested in one month, if clinically indicated.

HERPESVIRUS 6 ANTIBODIES                                                           XE:
  (IGG, IGM)
  HERPESVIRUS 6 IGG                                 1:80 H
  HERPESVIRUS 6 IGM                     <1:20
  INTERPRETATION                        PAST INFECTION

       REFERENCE RANGE:  IgG  <1:10
                         IgM  <1:20

       Human Herpesvirus 6 (HHV-6) infects T-lymphocytes,
       and has been identified as an etiologic agent of
       exanthema subitum. Rises in antibody titers to
       HHV-6 have been detected during infection with
       other viruses. In seroepidemiology studies of the
       prevalence of exposure using serum screening
       dilutions of 1:10, the detection of IgG antibody
       in a mid-life population approaches 100%. Due to
       this high prevalence of HHV-6 antibody.
```

PAGE 2 OF 4

Quest, Quest Diagnostics, the associated logo and all associated Quest Diagnostics marks are the trademarks of Quest Diagnostics.

Testimonials from the First Edition Continue Here...

97. *Meeting a Rescue Angel*

The way OUT of AUTISM

My 5 year old little daughter, Anna's story has been long and paved with lots of suffering and doubt, compared to her age, but prayers have always helped us in our journey. We are aware that it may take years to get out of the dark depths of Autism but now we know we are going in the right direction. Finally! Thanks to Kerri Rivera, the Rescue Angel of so many suffering children. She showed us the way out!

We have wasted so much time by trying to face the fact, that unfortunately there is a problem with our perfectly healthy-born daughter. After the consultation with many doctors, they first suspected a hearing problem. As the result of this, she went under many surgeries and antibiotic-treatments, which made her state worse mentally and physically as well. The medical professionals could not find any evidence of physical abnormalities, so they choose the easy way out, the one word diagnosis of Autism! Dear Mom, please accept it, you cannot do anything about it. Try some developmental therapies, but do not put your hopes high!

It is very hard to accept the unexplainable. What does autism mean? If you google it, you will find the verdict immediately: "Autism is a lifelong developmental disability characterized by impaired social interaction and communication, and by restricted and repetitive behavior" How is it possible that a perfectly developing child changes overnight? She stops saying the words she said before; she stops paying attention when called by her name and simply forgets some of her skills she knew so well before. How can one accept that? For us, it was impossible. We took up the gauntlet!

We spent so many sleepless nights on the Internet searching for solutions, to be shortly shocked to learn that in the US, there is a totally different concept and approach to autism. A group of medical professionals and scientists are proving the crucial role of environment and vaccines in the rapid rise of autism, accepting genetics as not the single cause but a factor of susceptibility to the disease. More and more so called DAN (Defeat Autism Now) clinics and practitioners are treating children on the spectrum all over the world with different biomedical interventions with various success rates. Whichever method one chooses in their realm, the base line to all is a strict diet, especially free from gluten, casein and sugar. There is no way out without eliminating these from the diet of an autistic child.

We immediately changed her diet from a bread and butter with chocolate milk style to a much stricter one. The first items to remove were sugar, gluten and casein, but these were shortly followed by potato, rice, corn and all artificial additives. We slowly changed to SCD (Specific Carbohydrate Diet). A few days into the diet and Anna slept through the night, and what's more, the painful eczema spots that had been on her face since she was a baby have shortly disappeared. Parallel with the diet modifications we also started the DAN interventions. We spent a fortune on lab testing, supplements and to see a DAN practitioner in Germany. Unfortunately none of these brought us significant improvements.

But I knew and believed that there has to be a solution.

And yes! I cannot keep it to myself and I would like to spread the word happily, powerfully and self-confidently, that there is a solution! This year in August, I have found the presentation of Kerri Rivera on the internet. Autism is preventable, avoidable, treatable and curable."

At the beginning "Kerri Rivera rose against the wind rather than with it when, after a frustrating slew of hits and misses, her youngest son, Patrick, was diagnosed with autism nine years ago. He was two at the time. Surrounded by literally no local expertise or any type of peer support circle, she did what most mothers would do: She became informed and involved in the biomedical community. It was through the Autism Research Institute (www.autism.com), a non-profit organization dedicated to autism research, diagnosis and treatment for the past 40 years, that she learned about a specialized biomedical protocol involving a special diet and supplements, physical therapy, applied behavioral analysis, chelation and hyperbaric treatment. There was light for Patrick, Kerri's son at the end of the tunnel."

But Kerri did not stop here! As she told me, she felt that there was a missing piece to the autism puzzle. A huge piece! It shortly turned out to be the MMS piece. She developed MMS protocol which can be done at home. Following Kerri's protocol, 68 children lost their autism diagnosis in the past two years! This result is outstanding, considering the fact that this is said to be an incurable disease.

The day I saw her presentation I e-mailed her and I got an answer within a half an hour. Kerri Rivera's MMS protocol is the missing piece of the puzzle to cure autism, ADHD, asthma, allergies, seizure disorders. Even mothers of children with Down Syndrome have reported improvements with this protocol.

My whole family was in Mexico in August. We spent two months with the person, who is also a mother of a child with autism, who sacrificed part of her life to help others. For free, out of faith, humanity and love. Approximately 8-900 children are following Kerri's protocol currently all over the world. The protocol is available for everybody and can be done at home (www.mmsautism.com).

Anna is getting better every day. Although we have heard very promising sounds from her, speech is not developed yet. However she has improved a lot in the last 3 months in many other areas like mood, communication, learning ability, focus and attention span are all better and she started pointing. There were no signs of these at the beginning of the summer.

We are fully aware that there is still a long journey ahead of us, but now we know we are on the right track. Thanks to God and Kerri Rivera to guide us to recovery!

---

98. Hi Kerri,

PLEASE use our testimony. MMS kept my daughter's PANDAS under control after switching from Silver Hydrosol BUT the biggest is the PP's (parasite protocols) in combo with MMS. She began sleeping through the night, THEN, sleeping in her own room after sleeping in our room for 9 months. THEN she began putting herself to sleep for the first time in 7 years. We always laid with her--her WHOLE life--I KNEW it was NOT "behavorial", knew it was due to *something going on interally, AND NOW, we read, kiss, and WALK OUT of her room! It's SURREAL really. Seriously SURREAL!*

Also, she was a chronic chewer, nose picker. Almost GONE! Peaks at new and full moon, BUT goes away. She was chewing about 20 straws a day before PP! Mangling them! UNREAL! I have pics of the straws... cant even recognize them as a "straw" once she was done with them. CLEARLY parasites. CLEARLY on the right track. NO other interventions. Except for how the MMS PP protocol is outlined MINUS Rhompe etc.. and Neem. NO sleep intervention supps etc.. None.

---

99. *Favy is my God sent angel. He is 4 years old and since he was born he came to change lives, to make families stronger and to help others forgive. As many of your stories, everything was "perfect" (besides of milk intoleration, delay speech and sleeping problems) the rest was great, milestones on time.. Then gradually after 13to 18 month something was clearly wrong, we loose him and then the bomb!: diagnosed in Puerto Rico by a developmental pediatrician and a neurologist as PDD-NOS on December 6, 2010.*

We moved fast and start the diet GFCFSFYF no sugars.. Etc Thanks to CEA non- profit organization who help parents with kids with autism. Soon we saw positive changes, then probiotics and a natural anti fungal. Difficult days... But always seeing little good things. On march 2011 I met an angel: Kerri Rivera and we start MMS protocol!!! Then we start to see the light! We moved to FL and they diagnosed him (now I know he was having a die off when reevaluate) he was diagnosed as Autism. For me , not a big differences, still in the spectrum and he still sick!!! We need to cure him!!!

After MMS the big changes came ,and still coming!!

first from 2 to 4 hours straight per night to 6to 9 straight during night plus a nap during day!!!!! Eye contact was much better and he had less tantrums. He learned fast 3 signs to communicate. He was more happy. I didn't see more great results because I was using Mms incorrectly for almost a year (with juice, antioxidants and vit c) When a couple of months ago I started doing it right (just water and no vit c nor antioxidants etc) he started to improve again and better.. After a lot of efforts, dedication, patience and hours of hard work, with Kerri Rivera MMS protocol Favy to improved in many areas.

This is all he have improved:

Educational:

- *Respond to his name when called and say his name when ask.*
- *Recognize people: Mommy, Daddy, Sister, Teachers, Therapists, friends*
- *Name and identify some body parts: eyes, nose, mouth, ears, belly, feet*
- *Recognize some animals (Ipad application or tv) : cow, horse, duck, chicken, zebra turtle.*
- *Recognize some colors: blue, yellow, purple, red.*
- *Recognize and count 1-10 (can repeat 11, 13)*
- *Recognize and/ or use some common objects: plate, spoon, fork, bottle, chair, pants, shirt, shoes, backpack, lunch box, blanket, pillow, tv, books, some toys, ball, cars.*
- *Stack objects like blocks, books*

* *Match pictures, objects or things.*

* *Color independently but refuses, he can trace but hand over hand.*

Social / Emotional:

* *He enjoyed for the first time his birthday Party, he was happy! He opened his presents and show emotions!!!*

* *He is beginning to interact more with others.*

* *He can follow direction of 1 step and some cases of 2. (clean up your plate to the sink and push your chair.)*

* *He is affectionate : kisses, hugs, eye contact is better.*

* *Wave and say bye.*

* *Greet with a good morning or good night.*

* *Play around and with others: chase, tickles, peek-a-boo,*

* *Dancing and imitating some song rituals like: The wheels on the bus, If you are happy, head shoulders knees and toes, and some in Spanish too.*

* *Throw, push, blow, pull, jump, run, walk, go up and down stairs, climb, open, close, zip, unzip,*

* *He can watch an ABC DVD seated or lay down , dancing or repeating.*

* *He can be seated for almost 5 to 8 minutes with his favorite books.*

* *He enjoys to go to playgrounds, go up and down slide or swing.*

* *He enjoyed the trick or treating this year, going house to house picking candy and wore all night his costume.*

Comunication:

* *I think this is the area of more improvement. From 0 to almost 35 words and different ways to label, request, interact and communicate!*

* *He can answer questions with yes or no (yeah- nah)*

* *He signs (modified) more, all done, juice*

* *He can use 3 words at a time (mommy water please, mommy come here, I love you, mommy ipad please, mommy hungry, eat)*

* *Say independently—not all of the sound clear but very similar sounds.*

* *Mommy*

* *Dadda*

* *On*

* *Yes(yeah)*

* *No (nah)*

* *Please (peas)*

* *All done (ah dah)*

* *More (mor)*

* *Water (at-ter)*

* *go*

* *Four*

* *Three*

* *Tickles*

* *Zebra (eebah)*
* *Turtle (totlo)*
* *Uh oh!*
* *Oh No!*
* *I Love You!*
* *Under*

When Prompting or Repeat:

* *Tv (tetete)*
* *Dvd (dededah)*
* *Off*
* *Come (she)*
* *Itch (each)*
* *Scratch (ach)*
* *Bubbles (babah)*
* *Bye (bah)*
* *Swing (win)*
* *Push (pah-sh)*
* *Five (hi)*
* *Nine (nun)*
* *Ten (ttn)*
* *Baby*
* *Ble (buh)*
* *Purple (por-po)*
* *Yellow (ye-yo)*
* *Chase (ch sound)*
* *Play (ay or pah)*
* *He can imitate some animal sounds like tiger, horse and wolf.*
* *He make choices between to objects . (playing, dressing, eating)*
* *Independent Functioning (he needs some prompting)*
* *He eats independently with spoon and clean up after finished.*
* *He can drink of a regular cup.*
* *With pictures or social stories he can transition to different activities.*
* *He is showing understanding of routines and rules. (dvd or toy stay home, 2 minutes and we are all-done)*
* *He can pull up and down his pants (multiple prompt)*
* *He can put his shirt on but need assistance to take it off.*
* *He can pick up and put on hamper his clothes.*
* *He can put his shoes (crocs) independently and cooperates with other type.*
* *He move furniture to reach what he wants.*

Health:

Favy still in the Gluten, Casein, Soy, Yeast, Antibiotics and Hormones free diet. He is trying new textures, flavors and consistency in a variety of food allowed.

He is drinking more water. MMS help us every time he have a cold, a mosquito bite aor nasal allergies. Is the best product we can ever find!

We are excited about all he learned, do and know during this last year because of the use of the protocol of MMS. We are proud of his efforts, but we still have areas to improve. We will continue with MMS, diet and wil start the parasite protocol with faith because MMS Rocks!!!!

We love you Godmother Kerri! God Bless you and your Family.

---

100. Conversation at pick up:

Aide: "What did you spike his breakfast with this morning? I've never seen him like this!"

*Me: *smile* "Oh glad he had a good day"...in my head...'three doses of MMS before drop off'...*

This child is going to figure out he has wings and start to SSSSOOOOAAAARRR!!!!

---

101. I have a 4-year-old son on the autism spectrum and MMS has been the most effective treatment we have tried for him. He had bad eczema on his legs for over a year that I was never able to get rid of. We tried diet, medication, cream and nothing seemed to clear it up. Within 2 weeks of starting MMS it was completely gone and has not returned. My son has made great strides in the past 5 months of being on MMS and the parasite protocol. He is more interactive with all the members of the family, talking more, and becoming more successful with his OT goals.

---

102. Kerri; Thank you all for your great job. Because through the Webinar—Autism and MMS—the information has come to me and gave me the strength to start a treatment.

---

103. Hi Kerri,

My wife and I did an ATEC on 8 year old son and it came in at 7. Wow.

We're learning so much from you about how this whole recovery thing works. The child's health takes three steps forward then two steps back then three steps forward then two steps back...but over time that one extra step forward seems to slowly add up. His pee is smelly again but, as you taught us, it's pee-tox and this is a good thing. His focus is still a bit of a challenge at school and he still won't shut up :-) but we are feeling that he truly is slowing getting better.

We've got a ways to go yet but we feel that we are truly moving forward. Thanks so much for all of the help you've given us and any help you might be able to give in the future.

---

104. My son is 19 years old. We started MMS in July. For 2 months my son did not sleep. The end of August at full oral dose we started the baths. At night you could see he was uncomfortable. I don't know how we did it but we talked him into doing enemas. My husband does it. They induce him to have a bowel movement.

Before enemas my son would have nights where you heard him walking back and forth from his bed to the toilet. In the morning I would say you had a bad night *last night, what was your problem? He would spell constipation. He doesn't seem to be doing this since the enemas.*

On Saturday 11-10 our son went to the first day of the season Special needs Ice Hockey. He is just learning. There were no problems with him about anything. In fact for the first time he put on his hockey gloves without objecting, which we have been trying to do the last 2 years. He followed directions from a volunteer on the ice, using his hockey stick to hit the puck back and forth to him and towards the goal. When hockey was over we went outside and hung out at the car and did nothing for about 20 minutes and had a dose of MMS. He did not care one bit. He entertained himself by watching the sea gulls. He was not impatient.

Next we walked over to the bowling alley, which is next door to the ice rink. He participated in special needs bowling. They have the bumpers up so the kids don't get gutter balls. My son actually threw some balls that didn't hit the bumpers. There were times in the past where every ball he threw hit the bumpers. Again there was not a single problem, incident, or signs of agitation from him while we were there. This is the first time in his life where we were able to go from one activity to another and not have an upset. He enjoyed himself.

What was really interesting was that he usually stims with a container of pumpkin seeds when he is in the car or at the bowling alley. This Saturday he left the container of seeds at home and not once did he even ask for them while we were out.

My son asks for a heating pad, which he has been doing for years. We wonder is it worms that have been causing all this pain for him? We would love to eliminate the pain, the repetitive behaviors, and the prompting through self-help and activities. He does lay around a lot. Everything he does is at a slow pace. He lacks stamina and speed for appropriate activities.

---

105. Hi Kerri, my daughter's first atec was 85. We just did it again and it is now 55! :) :) (approximately 2 months on MMS)

---

106. Hi Kerri,

I just wanted to let you know that we are up to 30 drops per 8 oz bottle and now I'm seeing some new things. He used future tense this morning (he sticks larry boys super suction ear to his window and waits for him to fall) This morning he said "Larry boy will fall." Usually when he falls he says in his high pitched voice "are you ok". Today when Larry Boy fell he said in a deep 11 year old voice "ouch". I just burst out laughing.

---

107. Dear Kerri,

In the spirit of giving thanks, I wanted to say how deeply grateful I am to you and your amazing work with mms and recovering children. I have never felt such hope for my son and such joy sifting through poop!!!

What a truly tremendous gift you have given us all...

Sending love and Thanks

---

108. My son is getting tiny bit more normal every day!!! I could kiss you!!!!!

109. *There was once a day not too long ago when all my son would do is scream and cry. No words at all. No communication. Today he asked me "mommy who holds up the clouds so they don't fall?" :) xoxo love that question! It means so much!! Autism is big. But God is bigger!! Don't give up ladies!! God led us to mms & to Kerri for a reason!!*

110. *Kerri is fabulous. She is very loving and supportive and will do anything to help our daughter recover. Using her MMS and parasite protocols have resulted in our eleven year old girl's ATEC scores dropping by half in less than sixth months. She is fully connected and much more responsive and affectionate. She herself is now motivated to succeed! We look forward to full recovery through Kerri's guidance.*

111. *I think we are over the initial die off hump!!! We are getting good reports from school about better focus and independence. Also, my son shampooed and combed his hair independently this morning. He even used a Q tip in his ears! I only asked him to try and shampoo by himself. He rewarded me with the combing and Q tip! Awesome mms!*

    Thank you again!

112. *I am so grateful to have found you!! MMS here in our house is a MUST and my boy gets better by the day because of it.*

113. *6 months ago, my son was on antibiotics for PANDAS, he was suffering with OCD and anxiety, as well as completely distracted - like a zombie.....his teachers came out one day with tears in their eyes because my son said he feels like monsters are inside his head and they are scaring him! They were so worried about him, as we all were!*

    We started MMS in April. My son got up to full dose relatively quickly and we added in MMS baths and MMS enemas and then I pulled the antibiotics—cold turkey! I was emailing Kerri at the time and was afraid to stop the antibiotics. She guided me and gave me the confidence that was needed! As suggested by Kerri, we also did night dosing of MMS. This really helped my son. Keeping a constant flow of MMS was the key.

    My son is flourishing!! He is HAPPY and confident. His focus has increased greatly! Since the start of the new year, he has had great day reports everyday in school! SO many gains in all areas! His coordination and strength have improved - he can do 11 sit ups (In June he couldn't do 1) He was chosen as Student of the Month in his Tae Kwon Do school (he is now a purple belt). The best news to report is his ATEC is now a 10!!!!

    Our family is so grateful to Kerri!!! This protocol has made the BIGGEST impact on my son's recovery!

    Thank you so much Kerri - My son's Fairy Godmother!!!

    Xoxoxo

114. *I was praying for a new direction to go in to heal my boy. I literally woke up the next morning with a link to Kerri's presentation in my inbox from a friend whose daughter is making beautiful gains with mms. We are on day 3 here and things are going great, I am so excited! Thank you!!*

115. *Yes, my son is making real progress - slow and steady but real progress. The MMS & Parasite Protocol has been very beneficial for him...that's why I am trying to stick to it very closely. Also, I have wanted to tell you that one of the best things that we have ever done for my son is the enemas!!! He is a gut kid with chronic constipation and your enema program has really improved his daily life.*

Thanks again for all of your help and guidance—always truly appreciated!

---

116. *Hi, I'm the mother of Lizbeth Hernandez Saray Miramontes and I want to share my testimony in case it is useful to you.*

My 13-year-old daughter started The Diet in January 2012, in June the full moon protocol to deparasite and withinin two months, she expelled a worm. This helped her regain 8 kilos of weight. (She was 10 kilos underweight when beginning the protocol) Her attitude has changed completely, from being very passive and low energy, she began to smile and be more active.

Thank you and God bless!

---

117. *From the ATEC at www.autism.com*

Total Score: 5

I. Speech/Language/Communication: 1

II. Sociability: 1

III. Sensory/Cognitive Awareness: 2

IV. Health/Physical/Behavior: 1

Started at 56 score in March.

I´m still having problems in environmental awareness, because in his mind sometimes he is Super-Hero (he wants to help to all the people in disasters, his own words). Sometimes he is eating portions like me, not such big portions but he is burning calories in the swimming classes. His meals are so healthy.

So, my homeopathic doctor told me (and recommends me) to wait until January next year to start the PP. He wants to see the effects of MMS, Probiotics and some homeopathy that he will provide at the end of this month.

We continue using as I said MMS, Theralac, GFCFSF diet, NAET and A.B.A therapy. I am so happy to share it with you.

Hugs my dear God Mother

---

118. *My daughter is doing well on the MMS. Her ATEC went from 53 to 26*

in three months. The special education teacher at her school was very impressed with how well she is doing. She told me to keep doing whatever it was that I was doing, because it is working!

Thank you for your help.

---

119. *Kerri, good day*

I want to tell you about the progress of my child, the treatment began in July and the progress that we have seen is:

-He has gone from not wanting to eat ANYTHING, TO EATING BY HIMSELF, SITTING DOWN he grabs his spoon and he cleans his mouth every time he gets dirty

- *He has stopped wearing a diaper, and goes to the bathroom by himself, almost like any other typical child.*

- * *He lets people other than my wife and I hug him.*
- * *He now gives hugs and has started imitating a few other things.*
- * *He can now obey simple instructions that he couldn't before.....no, pick it up, get down, get up, give me your hand, etc.*
- * *At the institute where he receives therapy they have told us that they are seeing great improvement in his eye contact, as he now focuses his vision more clearly on an object.*
- * *He still will not play with other children, but at least he is where they are.*

Thank you for everything and we are together in this fight day after day, winning small battles everyday to win the war.

God bless

---

120. My son was a non responder. We worked with a DAN! Dr. for 14 years. We did the DAN! protocol tirelessly with very, very little gains. Did 3 years of ABA and 2 of SonRise. Did the diet. Still do the diet. We started the MMS protocol in April of this year. My son has had more gains since April of this year than with 14 years of DAN protocol combined. My local children's hospital specialist told me to put my son in a home and forget I ever had him. What a jerk. We pressed on and I could write a book about the gains since April. My son's Dr said he had two near impossible "tough nut kiddos to crack open...one being my son." Well, MMS works. If my son can improve I know yours can too. Whatever you have to do to do enemas, and get him to drink it...do it. My son is drinking 35 drops a day. Trust me, it is yuck. But he understands that if he drinks it, it makes the autism go away and he slams it down and says, "yuck." He totally knows I am helping him and cooperates. Be encouraged.

---

121. Hi Kerri,

We're now at 1/4th of a drop per oz and today my daughter for the first time EVER answered a 'Where is...' question. I was putting out the garbage and when I came back in I remembered I had forgotten to put away a black marker

and when I found it was missing, I could only hope the property damage would not be too bad this time. I checked the living room and the kitchen but could not find any traces and meanwhile my daughter had taken the photo camera

and was taking pictures of herself in the mirror. Amazingly she seemed to have a way better understanding of the thing. I asked her twice where the black marker was, and after the second time she pointed towards the kitchen and said "There."

SHE HAS NEVER DONE THIS BEFORE! I checked the kitchen and found the black market next to the fruit. The pears, melon and mango all had black circles and stripes. I was a bit overwhelmed I should say.

Thank you, thank you, thank you for never having given up your search!

---

122. My son Fernando Latin is 8 years old and was diagnosed at age 3 and half. When I started this struggle with autism all the relevant studies were performed. We visited a neurologist, psychologist and a gastroenterologist, he began taking the supplements according to DAN! protocol, but the list of pills he was taking was increasing and my boy was restless. Then I went to a homeopath, *the treatment worked a little but it was so expensive I couldn't continue, it is needless to say how much money and time I spent on appointments and tests, plus the hyperactivity problem was growing each day. Until God put in my path an extraordinary friend MARIA CAMPERO and she talked to me about a foundation in Port Ordaz "VENCIENDO EL AUTISMO" FOUNDATION (Carolina, Yamileth and all that beautiful team) they brought Kerri Rivera to lecture here in Venezuela, on 04/02/2011. I will never forget these words AUTISM IS PREVENTABLE, TREATABLE, AND CURABLE. A new hope that a mother's heart never looses… In that conference she also talked about MMS and its benefits I was lucky Kerri evaluated my child, I started giving him MMS plus some supplements and have won this since:*

- *Better pronunciation, for example he used to say Dowg and now he says Dog.*
- *He obeys orders.*
- *Spends more time sitting.*
- *Evacuates normally with no bad smell.*
- *He stopped biting his clothes.*
- *He stopped making noises like: wiiiii yiiiii.*
- *He stopped playing with his hands.*
- *We go somewhere and he is relaxed.*
- *If we take a picture, he looks at the camera.*
- *He is more aware of his environment.*
- *He is improving a lot and really fast with his ABA therapy*

These are all the benefits I've seen with my son after starting with the MMS, I know there is a lot to do but know a have real hope. Some advice for parents that are starting this road: please accept what is happening, inform yourselves. I always think that to get your children through this you need the four-legged table:

- *Diet*
- *MMS, Supplement treatment*
- *Therapy*
- *Help at home*

I've always asked God to give me enough wisdom to help my child and to put me in a direction were I could find people that can help me, and believe me he has listened: Venciendo al Autismo Foundation, Kerri Rivera, my sons therapist Milagros Alcanta, Danny her teacher-shadow, Maria Campo, a woman not only committed to her son but to all of those who she can help and guide me, including Marilyn Aparcedo and her team. All of them are my bases to keep going and as Kerri and Carolina say: This is an endurance test and there is no giving up!

I have nothing but thanks these wonderful and professional people, Thank you!

Norkis Pinto, El Tigre—Venezuela.

---

123. Alexander was a normal baby until he was 6 months old, after that he suffered a change that, when we got home from work it was like no one came in, he was very uneasy, we talked to him but he didn't seem to care, first we thought that he didn't listen, the pediatrician and neurologist recommended hearing tests, *we did them and his results were normal. The pediatrician also recommended a psychiatrist but I thought it was very early for that so I took him to a psychologist, but he didn't diagnosed him, instead he referred me to a psychiatrist specialized in autism, and with a simple test he confirmed the suspicions: your child has autism and this has no cure, take him to special classes and therapy. I quickly signed him to a special camp and school, but they they just took my money and my sacrifice to travel 60 kilometers daily, they didn't do what they promised, they charged me triple just to give me a diagnosis and I believe he didn't even receive therapy. In my despair I went to the internet, I made an email account, a facebook page and started to look for parents that were going through the same things I was, then I contacted a Dominican mom living in Puerto Rico and she told me about Kerri Rivera and the diet. She told me that Kerri was the person to go to about this and Kerri answered immediately and asked me if my boy was on the diet and I said he was only 50% and she told me that if I wanted my boy to get better I had to do 100% so I wrote her back when I was doing the diet entirely. She gave me the first three supplements MMS1, THERALAC AND QUINTON WATER. Everything has been so difficult to get but if you persevere you get want you're looking for. I got the MMS1 but I was scared because of the strong smell of chlorine but I gave it to him with faith, I took it and my youngest baby girl started taking it. She had had a cold for about a month and it disappeared, Alex started to get better in every aspect but the biggest changes began 15 days ago, since I give him his MMS1 treatment. I start at 3 o'clock and finish at 12 at night. On weekends I start at 1 and I complete 12 doses of one drop mixed with citric acid and give him that daily. My 3 year and 4 month old baby boy is saying his first words. He is more connected, pays attention, he is aware of what happens to him and he is more loving than ever. I am so happy; I thank God, Kerri Rivera, Zeneida, Cynthia. But I mainly thank MMS1, because with a supplement so cheap, I'm getting my son back, the one that Autism wanted to steal from me.*

---

124. Juan Eduardo is 9 years old and he was diagnosed with Autism at age 2. It has been a long road in which we have tried everything. 7 years after trying and failing, in April of this wonderful 2011 we started in the right way. MMS came to our lives. Juan Eduardo was a self-centered boy and with a lot of fixations to objects. His improvement in school has been slow, but in 5 months he has improved what he improved in 5 years. MMS in our lives has been wonderful, despite the hyperactivity and lack of sleep at the beginning of the protocol, the life of Juan Eduardo and our lives have changed drastically in a very positive way. He will start regular school at the beginning of next year. He is a happy boy, he sleeps all night, he talks a lot more even if there is a long way to go, his quality of life is so much better. He shares with whoever is around him, a lot of times he surprises us with his qualities. He has adapted to the MMS in a great way. We took the panel viral test and the results were amazing, there was no virus. When I saw the great positive effects the MMS has made, I started to do protocol 1000 for maintenance with 100% success. Last year I suffered from several episodes of flu in which I had to take antibiotics, but this year I feel excellent, Thank God and Kerri and MMS I haven't been ill. I thank God, Kerri for putting you in our lives; I'm sure he has given back everything you selflessly have done for everybody. The smiles and looks in our children's eyes are priceless. Thank you Kerri and Jim Humble. A big hug.

---

125 Yes, I have increased the drops. Honestly, since we last spoke, he has had some of his best days...super happy. He actually read a book on his own on the way to school this morning. Seems very calm. Thank you.

126. *Hello Kerri! I have not written you in some time. I must tell you I have already been using MMS for two months. We have seen a lot of improvement. He suddenly speaks, says single words. If I sing a song or the vowels, he tries to imitate me, although it is still not very clear. But he tries to imitate. He smiled back at me when I smiled at him. And so those are several things. Tantrums are also less.*

---

127. *She is doing excellent! Every day we see small but important progress. I am not sure if she has noticed the changes in herself, the best will be when she starts school and classmates and teachers can give me their opinions. Family members have also commented on the changes we have seen in her.*

She is also answering questions about what she wants... With yeah, nah or the sign for all done with and then says dddon!

Now she wanted to see Mickey and told me ttttttiii and I showed her the movies and she said "ickey!" ;) We are communicating!!

---

128. *We are a family that lives in a country where living is not easy at all. The maternal death rates and child deaths are very high, and there is hardly any chance of getting access to special medical treatments and therapies. Our son started to show signs of autism since he was 2 years 8 months old. He had begun speaking but at that moment he stopped talking, he started to show no expression with his eyes, to roll around trees, to scream and to tantrum, and to not allow contact with anyone. Little by little he started to be very hyperactive and to have serious sleeping problems (he didn't want to wake up and he could not go to sleep at night).*

The start of my son's autism went along with colds, severe coughing and infectious diarrhea. Our son's teacher had always had excellent reports about our son's behavior, but since this severe cold with synovitis in his hip, the transformation begun. Since then, the awful diagnosis and the absence of guidance led us to an abyss. Our son presented total absence of self, almost no language, hyperactivity and many sicknesses. 6 months went by in which he met with a real parade of doctors and therapists that gave us bad treatments, bad diagnoses and almost no information, our life became a living hell. We went into a state of denial in which we didn't want to see anything on the Internet regarding Autism. In that time our son had been diagnosed with severe sinusitis, complications with his adenoids, and limitations that he developed because of his hearing capacity.

By the end of the year 2007, I alone broke the pact I had made with my husband, and started to search the Internet and found Yeroline Ruiz's page where she commented about the biomedical protocol through which she had gained her son back in months. Little by little I tried to convince my husband, who at the beginning was completely reluctant to the possibility, but by December 2007 we made our first order of supplements and quelators (magnetic clay from the United States). The supplements took forever to get to us and when they arrived we gave them to our son without any clue or guidance and hardly any diet. The results were horrible, he got worse and worse, and this lasted about 8 months. During this time the therapists gave us a terrible prediction about our son's future. Our family problems were also getting worse and the only guide we had was Yeroline and the Curando el Autismo foundation. Thank God we didn't give up and we passed the hard times.

We eliminated some of the food that is forbidden in the protocol and we started to use Enzymes. Our son started to improve, his health got a lot better and we stopped visiting traditional doctors and I decided to take care of his treatment, the process was not easy because there aren't even the basics for the protocol in my country, from the flour to the milk, everything is really hard to get. Obviously a DAN! doctor or supplements are unthinkable. However even if his health had improved, the Autism symptoms were still present, the speech therapy did not help and sometimes even made it worse. With a lot of effort we brought Rosa Dominguez (Tomatis Therapist) and we begun applying this technique to our son. The economic effort was very big but the results were worth it, he was more in control of his body, the clearest sign of this was the control of his sphincter that happened almost magically. His language also improved even if he still didn't use it socially, but little by little we verified that our son had no sign of retardation compared to other kids his age. Even with the improvement, the tantrums, frustration and limitations of his language remained, we decided then to make a new effort and traveled to Panama to the Stem Cells Institute to make a stem cell transplant.

Our debt was huge, and despite the effort our baby showed two contrary signs: a positive one that was that our boy was completely connected to his environment (he is never outside of the world anymore) and a negative one: he became aggressive. He had never shown aggression, auto-aggression or violent behavior, but from the moment the anesthesia was applied these began.

Our child started hitting himself when things didn't go his way and he bit other people out of frustration. We remained on our path with despair because we could not see the light. We tried hyperbarics for 70 days in a row, eqinotherapy, and others. In 2011 we came across Puerto Rico's CEA conference and we found Kerri's topic and the SCIA topics very interesting. We decided to do the SCIA tests and we confirmed that our boy also had virus and an inflammation even though his immune system didn't seem very affected. Because of our country's limitations we couldn't complete the SCIA tests but then we had a lead on how to start a new journey. All we needed was a doctor that could apply the SCIA protocol.

But then there was a miraculous day in which we saw Kerri's talk and hope grew inside us, we contacted her immediately and she answered immediately too, with good will and all the indications as to how give him MMS and how to apply the protocol. After two weeks we started seeing a change in our son. Eye contact was way better; he speaks fluid sentences and gives structure to his words.

We are now three weeks into the protocol, we can see almost no aggressions and tantrums, we teach him letters and he recognizes them, reads and writes his name (he does it with some difficulty) his receptive language is %100 active, he is very loving, specially with me. (His mom), he hugs me and kisses me and tells me he loves me. We know we still have a lot to go but I'm certain that we will make it. My son has now Borderline Autism; his connection with the world is total. We are still thinking about doing all the SCIA tests but just to find out our child is cured. The light and hope that Kerri and the MMS have given us does not compare to anything. Thank God! Kerri is a blessing in our lives!

129. *This email was after an ATEC drop of 64 to 44:*

This is truly AWESOME!!! He lacks in the "Speech/Language/Communication" area the most!!

I honestly thought when I was filling this ATEC out that he would have regressed because I just thought he might have...it seems everything else we have EVER done it is always too good to be true! It actually is not this time!! Praise the Lord Jesus Christ! \o/ I even had my oldest daughter give her thoughts on the questions too so that I doubled checked myself regarding his behavior etc...you know to make sure I was not out of touch! :)

We definitely have a ways to go, but I have never had my son at this level of 'awareness' before! We are doing RDI (Relationship Development Intervention) and have been since Feb. 2011...he's made great gains with it, but he is FLYING now only because of the MMS!!

~Melissa, IN, USA

---

130. *Hello, my name is Silvia and I have a 5-year-old son named Ricardo who was diagnosed with Autism 2 ½ years ago. From that moment he has been on a gluten and casein free diet and a therapy routine that help him day by day. Three months ago I met Kerri Rivera and the Venciendo el Autismo Foundation, and started giving Ricardo MMS. From that moment on he has improved significantly, his language is better and his behavior is a lot more controlled, plus with the biomedical treatment he goes to the bathroom regularly, this makes me very happy because he was always constipated.*

Silvia Morales — Caracas.

---

131. *Since I was a baby I was a very sick child ... I remember I had many gastrointestinal problems, I made my mom's life miserable with all the diarrhea and vomiting ... I couldn't gain weight and I had no appetite. All this accompanied me for many years; the diarrhea became chronic constipation, fluid retention and dry skin.*

All of this without knowing I had a serious problem, until I gave birth to my first child. The constipation was a nightmare, my baby got ill, when I fed him breast milk he would get a rash, and then he had convulsions for about 20 seconds, when I fed him again the same would happen but now the convulsions lasted for nearly a day. The doctors didn't know what was going on either. In one of my researches I ran into a page that talked about Autism and Candida being one of the main causes, this is when I find out that my son had chronic candidiasis and that he got it from me when I breast fed him. I tested my child and myself and we both had Candida, and I have had it since I was a baby.

One day I heard about the existence of a woman named Kerri Rivera who is devoted to helping children with Autism and I contacted her and she gave me all the help to treat my child. I will be forever grateful. She talked to me about the MMS and how to give it to my son. I also started taking it and in a matter of days I had a white layer in my tongue, it was the Candida getting out. I have gained my energy back, the headache I had in the mornings is gone, I feel vivid and almost all the Candida is gone. I've been with MMS for a month and I've never felt so satisfied with a treatment. This simple and yet powerful remedy has given my health and family back and with this I can assure, by my own experience that MMS works.

132. *Hi Kerri, I want to tell you that Erik has started with the speech therapist, and psychologist, when she finished the session with Erik, I asked if she thought that Erik could still be diagnosed with autism. She looked me in the eye and said, autism? Are you sure that he was correctly diagnosed? I found her the forms and she told me today Erik has eye contact, he socializes, and is cooperative. She said he does seem small for his age, and that we were going to have to work on strengthening the muscles of his mouth and the tongue as well because they are very week. She said that not chewing would affct the development of one's language, and Erik has only been chewing for 8 months, which started after we did the chamber with you in Mexico. That was the answer she gave me, and that's something, the fact that she told me that this is not autism is really something, even though she doesn't want to contradict the diagnosis they gave two years ago. The truth is, I am so glad he has come to improve, and that is since you KERRI have started to help my child. It's been a year since we started with the MMS protocol and the improvements have been incredible, he still lacks language as he mispronounces many words, but the worst is over. You can publish this if you want, thank you very much for all your help.*

Kisses

---

133. *We had first formed stool at 4 drops. It was so inspiring. Since then somedays we have formed stool, some days we have loose stool but I relate this to detox - I can clearly see his gut healing. His expressive speech hasn't changed yet but he obviously understands much, much better and is able to follow complex comands like "come and give me one teaspoon from the drawer" or "go and put this in the dishwasher" or "wait for daddy to get out the baby from the car and then you can lock the door and put on the alarm". Things like that, that were unthinkable before. I hope he will go beyond his few words soon and will begin to talk.*

Although speech is the thing that interests me most of all, perhaps the most impressive thing that MMS brought to us all of a sudden is the sense of smell - our boy never seemed to feel any smell, good or bad, in his entire life. He got this sense a week ago. We are thrilled - something is waking up in his brain, perhaps some pathogens in some part of his brain or in his nose or whatever died and now he is able to sense a smell. Things are getting moved.

Thank you Kerri.

---

134. *My son is 8 1/2 years old. I have spent six years and well into six figures to help my son. While I would see gains, nothing was really moving him forward. Three months with MMS, DE and ocean water; the junk that is coming out of him is both horrifying and thrilling. I think he is finally starting to feel better, and because he feels better he is more with us. He is trying so hard to speak and everyday he is saying at least four new words. He also watched a movie and laughed at a funny part!*

---

135. *We started MMS 68 Days ago and my son's ATEC has dropped from 50-23... Now that's a MIRACLE!!! He is So much more "with us" and I'm seeing improvements daily. He is just so much more fun to be with now:) He did have a plateau, but is back on track. We have only done 4 enemas to date, but they made a big difference! Baths do as well:) I can't wait to see what he is like on his 4th birthday (still 39 days away).*

Kelly, CA, USA

136. *Is it just an amazing day or is the MMS already working? Its just day 1! I know people say that once your child starts talking, they never stop, but holy cow... Today in the grocery store, it took me 3x's longer to shop because he was talking about everything he saw around him and needed a response on all of it from me! So unlike anything we have ever experienced with him before. Could it really be the MMS already? Also, potty training has been going well (slow & and steady), but today, he didn't need much prompting! Many times, he just stopped what he was doing and ran to the bathroom to go all by himself! This is amazing, because he never stops a preferred activity to go potty! When i thought it had been too long since he went last and said "its time to go potty." he just ran in there and peed, washed hands etc. No tantrums or weight dropping, he just ran in and went! God is answering my prayers! :) I'll keep you posted on what happens tomorrow!*

---

137. *You know Kerri, I had to email you directly I waited until today to retake my son's atec...just did his first Parasite Protocol and did the 72/2 hr this weekend, so waited a few days to pass by before we did...so this atec was based SOLELY on this past monday, tuesday & today. I got an 11, hubby got a 13. so average: 12. (he was 32 before MMS, then 27 after 3 weeks of MMS...now 12) I had to email you to thank you b/c well...because. What is there really to say :) cant think of the right words really :) thank you?? not really enough if u know what i mean :) ...But to think i was going to skip your presentation at A1 b/c I thought it's too good to be true. But you caught my attention when you started tearing up as you got up on stage :) you were the only person/speaker to show emotion like that. AND have NOTHING TO SELL ON YOUR TABLE :) You were there only for our children...getting nothing back in return! :) I know I've been emotional, but I'm prego and my sons getting better so cut me some slack :)*

I'm SO grateful for all you do Kerri. If there is something I can do in return, Please let me know. I know you have a lot on your plate. You have a heart of gold; such a giving, compassionate person. a "rare-ity" (this word is in my dictionary!!) You & Patrick are always in our prayers. I seriously want to go visit you in Mexico w/my family so my son can meet you when he's recovered. I told my hubby & he's up for it hehe :)

Anyways, long email but really just wanted to let you know how well he's doing and to 'thank you'. It blew my mind after I got the results. He still has issues in socialization and getting his vocab/speech up to his 4 year old level. I think much of that will change as he learns English better: he'll interact better with his peers.

Thank you Kerri from the bottom of me & husbands heart :) You are loved & covered in prayer. :)

---

138. *We started to see a setback in January with our 1 year 8 month old girl. Until this date she had had a normal development, like her sisters (she is our third child) but then, she stopped talking, she just screamed and twirled. She would twirl like 20 times a day for no reason until she got dizzy and she fell to the ground. She didn't look at anybody in the eyes and could be behind a curtain hours comparing her toothbrushes; she wasn't hungry nor asked for her milk, we had to push the food into her mouth so she would eat it. I was desperate so I went in to a forum and a friend from Venezuela told me to look for Kerri Rivera in my country so I did. She immediately answered and gave me hope, since the*

three doctors I had seen before told me that Autism had no cure, that all we could do is to teach her how to survive and be a little bit more independent.

I found my hope in Kerri, I read the testimonies on her page and sent her a message that she answered right away. She told me to do the diet, remove milk and everything else, I went to her clinic and within a week of the diet Abby could look at me and smile (she had stopped doing that long ago) When I went to her clinic Autismo2 in Puerto Vallarta she looked at Abby with all the patience in the world, she gave me MMS, the supplements and all the hope in the world. Everything was new to me but I had faith that I was in the right place and from that moment on, while always counting on Kerri's advice, my baby girl started to get much better. Her first ATEC was 95 and a year and a half after it's a 15. My girl now smiles, talks, dances, she obeys and stopped screaming, her only tantrums are of a normal girl her age. I did hyperbarics and am still doing MMS, supplements and deworming her. She is now diagnosed with Pervasive Developmental because the Autism symptoms are gone. She doesn't run from one place to another, she plays, looks for other kids to play with and has friends. The only thing left is for her to establish a conversation, and to put together the words to make sentences, but I know she is going to make it.

Once she sat down on ant hill and a bunch of ants bit her, we had to go to the hospital and the doctors gave her cortisone and she was seriously ill, now when a bug stings or a mosquito bites she tells me where it bit her, and she asks me to apply a calendula ointment that I use for the mosquito bites. She sets the table and can eat by herself, she doesn't use a diaper anymore, we can go shopping or to parties and she is well behaved, there is no more tantrums or fixations, this is wonderful! I know there is a long way to go, I've been in this fight for a year and a half, but with KERRI RIVERA'S HELP I will rescue my baby from autism, even though some say that Autism has no cure I know it does. Kerri has been a blessing for our lives and I thank God for putting her in my path.

Thank you Kerri, I have no way to repay what you have done for Abby.

---

139. Our 10 year old son, Ben, showed symptoms of autism since he was an infant. Until he was 6, he never connected normally with family or peers. As his mother, I ached for a hug, to be called "Mommy," or for a lingering look of love. I never got them.

Until he was diagnosed with autism at age five, I was utterly befuddled. What had I done? What was wrong with Ben? Why didn't he want to play with other children? Why did he drool endlessly, laugh constantly for no reason, and bolt from play dates and parties with no warning? Why was he constantly lethargic, to the point where we called him "the playdough boy?"

Ben was always happy, dwelling far from us, in his own perfect little heaven. My husband and I were depressed, even despondent, despite attempts at gluten-free, casein-free diets, ABA therapy, and many supplements, because we saw no substantive and lasting changes in our child.

Finally we found the Son-Rise Program®, which brought Ben back to us. After three-plus years of a mostly full-time Son-Rise Program®, Ben became highly interactive, engaged, and interested in his peers. That alone was a palpable miracle. As we say in Hebrew, Dayeinu! That would have been sufficient. But we constantly experienced regressions back into autistic symptoms. For no apparent reason, and despite endless biomedical interventions, including the DAN!

protocol—to the tune, I might add, of tens of thousands of dollars, multiplied over four to five years—Ben would suddenly revert to autistic behaviors, and we would lose contact. He became once again non-responsive, uninterested in family or peers, and seemingly unable to connect. Only the day before, he had been totally connected.

After being on the SCD and GAPS diets, and finding even them ineffective to alleviate these regular and devastating returns to autism, we were advised by Dr. Campbell-McBride that it sounded to her as if Ben had parasites. A friend of mine who was also a Son-Rise mom had mentioned to me that she and another Son-Rise mom friend of ours were trying something called "MMS." When I heard the words for which the acronym stood, I gave an internal scoff. "Master Mineral Solution? Is this a joke?"

I can tell you, in all seriousness, that now, five and a half months into our MMS autism protocol, MMS is definitely no joke. Since starting MMS, we have seen nothing but progress in Ben. We are still doing GAPS, but avoiding foods we know Ben cannot yet digest. Ben's ATEC score has dropped from 43 when we started MMS to a startling 20. We are seeing sustained, natural, unsolicited and consistent eye contact, like never before. Ben is consistently responsive, certainly on par with most 10 year old boys. He is playing hockey in our driveway with friends. He is solicitous of others, and compassionate when others are hurt in any way. He is funny, expressively loving, and eager to please and be loved. It makes me cry to write this. I wanted this so badly, and now, with MMS leading the other interventions, I have what I prayed for.

We are not perfect on the MMS. Ben is in his first year of school post-Son-Rise, and dosing during the school day is not easy. But we are tremendously hopeful and excited.

I truly believe, as Kerri Rivera has often said, that parasites are the missing piece of the autism puzzle. Like so many other warrior parents, we have done everything. But everything minus one was not enough for Ben. Now, I think we truly have the tools, and it is only a matter of time, faith, and love, to fully and permanently recover our beloved Ben.

---

140. Caitlin has been on the MMS protocol for 3 months now. Caitlin is now off her 100 plus supplements a day that were the glue holding her together so to speak, with no regression. She is happier, more affectionate and focused. She no longer has a distended belly. She used to have terrible smelling gas on a daily basis and very foul smelling bowel movements. This is gone! She has increased muscle tone (has a diagnosis of low muscle tone and bilateral hip anteversion). She has been seizure free as well! She continues to make slow but steady gains in her reading and math skills at school. She isn't conversational in her speech yet but can comment on things and express her wants in longer and more elaborate sentences. She is falling asleep faster and sleeping longer.

---

141. Eczema!!

At last my son (4 yrs) is living now without eczema, he was born with severe eczema all over his body, we used every allergic cream and med and even the magical Dead Sea Clay but it didn't work, but now after using mms and pp (parasite protocol) for 2 months... ITS GONE!! Thank you MMS!!"

Thank you!

---

142. *This is the story of our son who is 4 years and 9 months old today. Our story is a desperate struggle to defeat Autism, the same struggle that a lot of families are going through to free their sons and daughters. Guillem started to have ear inflammation and laryngitis when he was a few months old and we medicated him with antibiotics, he had his vaccines according the vaccine calendar, and by the time he was 9 months old he was showing signs of Autism. He would put his fingers in his throat until he choked, he would hit his head on the wall and he was always mad, his tantrums were frequent, longer and more intense, he started to look at the lights and to twirl with no sense, he showed less interest in people day by day and he would get really mad if someone said anything to him. He was hyperactive, couldn't stop running and going up and down the stairs. The nights were terrible too, he didn't sleep well and woke up screaming as loud as he could plus he could not say a word.*

My father and my brothers blamed me for this; they said that my boy wanted to dominate me and to get my attention. I had a hard time recognizing that it wasn't this, that my boy had something else. Once I accepted it I spent hours and hours reading on the Internet, and I read about the gluten-free diet and that was our first step. When Guillem was 2 and half years old we changed the diet and it worked. Guillem started to respond to orders, there was less screaming and his first words appeared. Three months into the diet we contacted Kerri, this has been a key moment in our history, because she has been our guide to get our son out of this long tunnel. The first thing she did was to check what our son was eating, I thought that I was doing a perfect diet but I wasn't, there were some foods with sugar, soy and additives, the next thing we did was to talk about the supplements and MMS, but I was very scared to use it and it wasn't until three months later, the day Guillem turned 3 that I started giving it to him. Now we have been using MMS for a year and 9 months. We give him MMS daily and every weekend we can but minimum one weekend a month we do the 72/2. His improvement was amazing, his language got better, his dressing skills, he learned how to use the bathroom, his sleep was normal, and the screaming was almost gone. The next step was to start with the MMS enemas, every enema made Guillem more alert, more present, we started with the enemas 4 months after starting with the MMS. We saw that the enemas expelled a great amount of biofilm and the more he expelled it, the more abilities Guillem gained back, I saw the importance of healing the intestine and that has been my obsession since.

The next important step was the sessions in the hyperbaric chamber, his first sessions were at age 3 and 9 months, and the changes we saw were mainly in the obsessions he had, he paid much more attention to everything, his social skills were very affected but after this he made his first friend. Now we do hyperbarics with seawater every three months. Another piece in the puzzle were the parasites but with Kerri's help we've been on the protocol against parasites for 7 months and every deworming cycle our son improves. I know that we have to keep fighting but I also know that we are closer to the exit. Now Guillem sings, he plays with his brothers, we go to the movies, we do puzzles, we paint, and he is very happy, he has learned to spell his name and he walks joyfully down the street telling you about his day. He goes to school, plays football and more importantly, we are a family again.

For all of this we are deeply thankful to Kerri, because she has always been there for us, without Kerri and MMS our story would be completely different, so we owe them EVERYTHING. I know that we will someday say Guillem is cured, and all the other children with Autism can say that too.

UPDATE: As of December, 2012 Guillem no longer has a diagnosis of autism, here is an email his Mother sent to Kerri when she knew he had recovered:

> *KERRI!!!*
>
> *My son has an ATEC of 10!!!!*
>
> *I have it repeated several times because I did not believe it...*
>
> *I have copied it below...I'm going to have it framed!*
>
> *The truth is that this Christmas we have enjoyed our whole family, we've been to the Zoo, to the fair, and during family meals he was very well behaved and enjoyed the Magi and the magic of Christmas. Christmas last year was totally different. For all this we give infinite thanks! There are not enough words to tell you how we feel! Without you our story would be different. We will continue working to get to zero!*
>
> *A big hug!*

---

143. I started with MMS on April 2012, I was afraid but had a lot of faith. I did the ATEC for my son and it was 92. He could not use the bathroom alone nor say a single word. The day I started I thought there was nothing different, but when we were at the beach on vacation I noticed that when I called his name he would turn, and come to where I was. All this with one drop! Before that day I always had to stand up and find him so he would pay attention to me. He wouldn't look at me but he would listen to instruction. Days went by and when we came back from the beach he went straight to the bathroom… I couldn't believe it; I sat down and asked my self WHAT IS THIS? What happened? And I silently thanked God and Kerri. This was decisive for me to continue. After this he improved day by day, after 5 weeks his ATEC had gone down to 76 points.

I made a few mistakes at the beginning because I didn't know the protocol very well, and wanted to rush things. Again, Kerri with her infinite patience and love for others led me to the right path, and I started to see more, way more.

My son started writing, he knows his letters, he reads some words (he has memorized them) he is great in the bathroom, he listens, his eye contact has improved significantly, his last ATEC was last month and it was 60 points. I know we have a long way to go, but I every month there is improvement and even though there are some bumps and setbacks the good things overcome the bad.

My son has also started saying words, names of movies, he can say over 100 words (names of movies or things he likes). Some will think it doesn't mean anything to know the name of a movie but for me, it is to know how my son's voice sounds and that, is priceless. Now my next goal is to hear him say "MOM"

~Myrna Sterling, Mother of Gianmarco, age 5.

Myrna is the president of the GM foundation for Autism

---

144. My wife and I have always been fond of traveling, getting to know different cultures and people, this is why we had our little one we decided t o keep traveling but to adapt to his rhythm.

When he was 9 months old he went on his first international journey. Everything was perfect, he was the happiest baby on the plane, he played with the people in the seats behind us, the stewardess wanted to stay with him to play, all the plane ride went along perfectly.

Something happened between his 15th and 18th month, his smile was gone, we no longer had happiness in our home, only nerves and less language. Maybe the excessive vaccines were destroying his immune system and leaving the door open for parasites to take hold.

When he was 18 months we went on a second trip and it was an absolute nightmare. He wouldn't stop crying and screaming on the plane, he didn't want to play with anybody. We spent a few days getting to know this new place while our baby fell further into this abyss in front of our astonished eyes.

Something was wrong, and after an episode of auto aggression we knew that this was not normal. After watching a couple of videos on the Internet we started to understand, and we went to a psychologist who didn't give us a diagnosis because our son was so young.

Later on we went to a neuro-pediatrician who also used the excuse of our son's young age to dismiss us, and finally a last doctor who didn't explain anything either.

Not even our family listened because they thought we were exaggerating. We were left alone and our biggest ally was and still is the Internet.

After the medical failure and their "anti-diagnosis" we took the best and most important decision of our lives, that was not to listen to any of the diagnoses, or to wait another 6 months to be called by the neuro-pediatrician again. We started down this road without an official diagnosis but with all the symptoms and behaviors that define autism.

We started with the casein, gluten and soy free diet with no experience on how to do it and our baby suffered awful withdrawal symptoms. We couldn't believe what we were seeing, his nervousness and aggression took us to our limits. After a few days on the diet his tantrums went down little by little until he was stable. But we needed something else because everything was out of control; this is when we went back to the Internet.

After three months of false doctors and pseudo medical experts we found Kerri Rivera.

She gave us hope, and filled us with confidence, but most importantly she was clear with us from the beginning. Everything she told us was logical and everything made sense as to why it was done so we trusted her without a moment's doubt.

He got better and better everyday. Kerri told us how to do the ATEC and what it meant. Our first result was 57 points when he was 2 years old, which was very high for us because of his young age. We have been doing everything Kerri suggests, improving the diet, and using everything she recommends, and the results have been spectacular, right now he is 2 years and 10 months old, his ATEC is 8 and he is doing great. Our goal is to have him at 0, and I'm certain that once we free his intestines of all the parasites, he will be a strong, healthy and happy boy.

He doesn't have any behaviors like the ones he had one year ago, now he smiles again, he talks and asks for things, he plays a lot and he is happy.

This was possible thanks to his Godmother Kerri.

Kerri, you have done and given us everything in exchange for nothing, and that is your defining quality as a human being and as a person.

Every day you give hope to all that YES WE CAN defeat Autism and now we have a lot of puzzle pieces to complete our puzzle. We may find a new one soon because there are a lot of people investigating and that makes us feel alive, to fight against the entire circus that has sprung up around this problem.

I would recommend that if anyone is not seeing progress in their children, to come to Kerri, and try everything she says word for word. She has done it all and has the best investigators and doctor on her side. The ones that aren't intimidated by the absurd laws of the upper echelon, those who only want our children to be sick all their lives so they can sell medicine to us.

We will keep up the fight.

---

145. Hi Kerri, I just got the third MMS Autism newsletter, and I tell you it couldn't have arrived at a better moment. For the last few days my nephew has had a strange infection, he was hospitalized and the doctors couldn't figure how to fight it because it was resistant to many antibiotics. I told my brother-in-law about MMS and he started using it. After two days my nephew was healed, the doctor asked me for a little so he could examine it, and his answer was that the way MMS functions is amazing, he couldn't believe that such a simple and powerful product hasn't come out to the public eye.

---

146. My husband and I were on the couch having our morning coffee and my son came bolting out of the room and ran up to my husband's face and said..."Oh Hi Daddy, I'm back!" And then ran back to his room...this is our once in a while "word" little man. After 3 non-progressive PPs (parasite protocols) the 4th one was a charm. A nightmare, but quite an eliminator... Lots of Worms Gone! And he is slowly coming back...we are seeing it in fact he is even reassuring us that he "is back".

~The previous email was accompanied by an ATEC drop from 95 to 35 in 6 months.

---

147. Dear Kerri,

It's a pity I don't speak English and that's why I can't write you on my own. Currently, a lot of good things are happening to my son (31 years old). I have followed all your advice and as a result he has completely calmed down. I have learned to prepare CDS at home with the help of Jim Humble's video. The CDS is a 600 ppm solution; my son drinks 30 ml per day with the feeding bottle method. I suppose this is the daily dose for a person of his weight. We are past the third PP and preparing for the fourth one.

Unfortunately I won't be able to attend the congress in Prague, since I can't find anyone to look after my son. If it were in the summer I would risk going by car and taking him with me.

I am very grateful for your attention to our life and progress and for your help in curing my son. God bless you for helping children with autism by sharing your knowledge and experience with the world. I will try to give a regular account on the progress of my son. I have no doubt he will be cured, too.

148. *Three more tiny little gains on our step by step path to recovery....yesterday Nick got a haircut and did not cry at all AND was relaxed enough that it was the first time it wasn't like a sheep shearing, he actually got to get a little style (so cute!). Also, he was doing some pretty terrible screaming in the car in the parking lot at Whole Foods while I tried to find a parking space (I wanted to scream too) and I let him know he scared me. When we got into the store he spontaneously said "I'm sorry" -total first! Lastly, He has requested a hug from me three times recently - HUGE new thing and I love it. Thank you MMS and Kerri.*

---

149. *Thank you Kerri. For all of your help, answering my emails, and for all that you're doing to help all of us, help our kids!!! I was at a breaking point. Nothing was helping her anymore, her organs continuing to get worse, hadn't grown in 4 years, and had become Jeckyll/Hyde this past 4 years...A violent, raging Hyde being increasingly present more and more each year, each month. I knew something had a hold of her all these years. And it amazes me how 2 endocrinologists, 3 pediatric GI specialists, and many other doctors had no clue.*

Thank you for ALL you are doing to get the truth out there. You are an angel!!!

Merry Christmas - xoxo

---

150. *Puerto Ordaz, November 4th, 2011*

My name is Carolina Moreno and my husband is Rafael Colmenares, we are the lucky parents of a little 6 year-old whose name is Ana Victoria, she came to this world to give us the world's greatest lesson, through her we have learned the real value of life, to be steadfast, and brave. We learned that it's our job to fight for our dreams if we want to accomplish them, and over all we've learned that God is who holds us up and gives us the strength and tools to move forward.

My journey with Autism began the 27th of march of 2007, my daughter Ana Victoria was 23 moths old, her initial development was normal, but little by little she stopped talking, she didn't understand anything, she didn't miss her parents, she didn't play with other children or even her toys, she didn't like to be kissed or hugged, she acted as she was deaf but we didn't know that all this were symptoms of autism. It wasn't until the day that we went to a neurologist in our area, that someone told us that this was Autism, an incurable condition, and that our daughter would have it all her life. It was a very hard moment for us; all of our plans to be a happy and healthy family went down in a matter of hours.

As parents we didn't want to give up, after recovering from the horrible news, we began looking for information on how to help Ana Victoria. In our country of Venezuela there were 4 specialists in Autism and they were all using the gluten and casein free diet, vitamin therapy, as well as expensive tests performed in the United States to check candida and bacteria. Thanks to them I learned how important the diet was to get Ana back from Autism, without it, recovery would be impossible.

We did some testing through Great Plains labs, and the results indicated candida, bacteria, inflammation and much more . We started with Ketazol and Flegyl. My daughter was taking up to 400 mg of Ketazol even though she was barely three years old. Now with everything I've learned I can not believe that a physician could prescribe such a high dose to a child who is only 3 years old and weighs only 16 kg. That alone should be punished and criticized

However, I felt that I wasn't doing the right thing, years passed, Ana Victoria was 5 and her comprehension and attention levels were very low, my husband always thought there had to be something else we could do!

One day my dear friend Carolina Garcia talked to me about Kerri Rivera's blog, she told me that it explained the reason for the supplements and the hours they had to be taken, for example that the antifungals had to be given away from food and the pro-biotics worked better if they are given at night. After I read her blog and visited her website in July 2010 I called the Clinic Autism02 and Kerri Rivera herself answered me. I explained my daughter's case, and she asked me to give her an opportunity to help, and if in 2 months we didn't see anything different with our daughter we could continue doing what we always did. I told my husband and we agreed to try. I felt a great sense of peace when when Kerri told me, "I'm here, and I promise I will walk with you until we recover your daughter"

In August Ana Victoria was on Kerri's full protocol and in November God gave me the opportunity to go to Puerto Vallarta, México. An entire month in this city were my daughter received the hyperbaric chamber treatments, we started with MMS, plus we got together with Kerri at her home every night with pen and paper, she explained to us step by step how we could heal our children. From that moment a beautiful friendship was born between Kerri and I.

This is when I understood that Autism is caused by a weakness of the immune system and this brings as consequences; viruses, bacteria, candida, intestinal and brain inflammation, food allergies and contamination with heavy metals, and that it is necessary to attack every aspect to be able to heal autism.

The 23rd of November Ana Victoria Started to take MMS and in 30 days of treatment we noticed the firsts changes. My daughter's stools were consistent; there were no food residues or foul smells. When I came back from Vallarta I was full of new knowledge and a lot of MMS bottles, Ana Victoria had already a very well defined protocol and the best thing was that it wasn't a long list of supplements, only 7, MMS and Theralac. Everyone in my family noticed a great change in our daughter, better eye contact, controlled hyperactivity, she started saying her first words with sense and purpose; I want park please, I want food please, I want to sleep. She was doing much better in therapy and at school. Anita had begun her way back to the world she left when she was 1 and half years old.

When I saw this improvement in my daughter I decided to share my experience with other members of my community and it was there with my dear friend Yamileth Paduani and her husband Alejandro Teran that Fundación Venciendo el Autismo was born. Mr. Alejandro asked if Kerri could come to Venezuela and talk to parents in a simple way so everyone could understand that Autism is curable. Kerri was delighted to accept the invitation and in only 7 moths of work we had 2 conferences and many evaluations and nowadays there are 700 kids on Kerri's protocol that is: casein, soy, gluten and sugar free diet, some supplements depending on every child, MMS and Theralac. The results are very positive and there hundreds of families that have started with their children's recovery.

Then came the great discovery of the parasite protocol, for me, it was the missing piece in my daughter's puzzle. The first few months of the PP were severe (fever, laughter, hyperactivity) and many worms in her stool. However, after all these worms came out out I saw how her mind cleared, and month by month we waited for the full moon to deworm ...

Ana Victoria is now improving very quickly, she is a girl that enjoys being with her family, the park, a day out, she goes to school and she says to me that at school she now does her work in silence without crying, when I arrive home she greets me with a smile, she runs to me and calls me Mom, she tells me what she wants and what she doesn't, she can use the computer like an expert, she takes her own shower and dresses herself, she even picks out her clothes, she knows when Christmas is coming and that there will be gifts. Her therapists and her shadow teacher are really happy to see how Anita is progressing day by day.

Yesterday we were all at the table eating lunch ... and she points with her finger …"Papa's food, Mama's food and Ana Kikota's (Victoria's) food…yummy… heeheehee. My husband and I were laughing with excitement and happiness. Some time ago we felt that speech would be nearly impossible for Ana Victoria.

My advice: do not be afraid of MMS ... be afraid of autism; that which is unforgiving and has the power to destroy lives and families if we don't attack it on time.

We are grateful to God because He has never abandoned us and has always given us the strength and tools to fight and to believe that it is possible to recover our children. We thank God for putting Kerri in our path and we ask God to bless her everyday for having such a big and noble heart, she who is always helping others in a selfless way. We also thank Jim Humble and Andres Kalcker for helping us find the missing piece in this great puzzle that is Autism. We can do it!!!

To all of you, many sincere thanks for helping the children in my country.

Blessings,

Carolina Moreno, President of Fundación Venciendo el Autismo. (The Defeat Autism Foundation)

Venezuela.

151. *Good Day, first of all I want to thank God for giving me all the necessary tools to be able to help my son, Jesus Soto who is 8 years old today. Jesus was born normally with no complications, loved by his parents and family, everything was going fine, but when he turned 18 months old he became ill; laryngitis, and fever, amongst other things.*

He was diagnosed with Autism at age 3. As a worried mother, I took him to the psychologist because I observed several behaviors that were not normal. After the terrible diagnosis we started to take him to a specialist after specialist, they were all very good doctors, his DAN! doctor in Venezuela, and my sons gastroenterologist were very helpful. I met a very special friend Mayerling Aparcedo who brought the ABA method to Venezuela. My son had many improvements but we still felt that we needed something else. I spent my time on the internet finding out as much as I could until I found Kerri Rivera that without even knowing me helped me via Internet, talked to me about MMS and about the supplements and how to take them. My friend Carolina Moreno had the joy to travel to Puerto Vallarta and wrote to me from there, telling me that MMS was helping a lot of children. The 14th of December I ordered my first MMS and the 27th of January it arrived to my country, from that date on my son is taking MMS.

The 2nd of April Kerri arrived to Venezuela and I went to see her, I took all the tests I had and she checked them, after only two months of taking MMS his viral *load went down to near normal levels, and his liver profile showed no problems. I made some adjustments that day and I continued with the treatment seeing a lot of improvement in his language, socialization, understanding of things, his hyperactivity has gone down and the laughter for no reason stopped. I'm very happy with the results and now all the family is taking MMS.*

I'm very grateful to God for putting my husband in my life, as he has been a very special part of this battle that we are winning, I also want to thank our family, Kerri, Carolina, Yamileth, Mr. Alejandro and everyone from the Fundación Venciendo al Autismo.

Yours truly,

Maria Campero

---

152. *My baby girl Yuliangel Nazareth Quijada Montero was diagnosed with autism when she was one and a-half years old. From that moment she was given a treatment to allow her to sleep and to control her tantrums: "Risperdal" which had absolutely no effect, I gave her up to three pills before bed and even them nothing happened, I was just damaging her brain. She didn't sleep night or day, I was worried because she screamed all the time and she hit herself.*

The 2nd of April 2011 I attended the conference of Fundación Venciendo el Autismo, where I heard for the first time that Autism had a cure. Kerri explained in a very simple way how I could cure my girl, even though I didn't have all the economic resources necessary for her treatment, the Foundation helped me with a part of the treatment, MMS and multivitamins. From there I started noticing changes, with a lot of sacrifice and effort my girl was on a strict casein, soy and gluten free diet. When I started giving her the MMS I would give her 72/2 on weekends and then I noticed that there was a lot of mucus in her stool, from there she started talking, she slept the entire night, and took naps in the afternoon, she pays attention, she knows how to count and all the colors with only three years of age. For me MMS is a miracle, since I started giving it to her she is a very healthy little girl, with no colds or any of the diseases she used to have, she is now considered to be very high functioning (14 pts. on the ATEC scale) with MMS, Vitamins, and the diet my child is getting out of the Autism spectrum. She does not go to therapy because of my finances but my sister Yenitze Montero interacts a lot with my girl, and her grandparents Raiza and Julian help me a lot with the diet, my nephews Jesus and Luis are also helping me with the home therapy and of course her dad and me fill her with love and care. My daughter is healthy, she has meaningful speech, she recognizes herself, she sleeps well with no pills and we are all very happy with the results. God Bless the entire team at Fundacion Venciendo el Autismo, especially Kerri Rivera for bringing MMS to my country, it is absolutely miraculous. Blessings.

Yurliana Montero and Jonatha Quijada

---

153. *Martin is 6 years old, weighs 19 kilograms and is diagnosed with severe non-verbal Autism. He was diagnosed when he was 2.*

He almost made no progress with therapy, when he turned 4 everything seemed to be lost, but when he was 5 we tried a new therapy and he started to say a few words. He also got very sick almost everyday, he coughed a lot, had an ear inflammation and fever, he couldn't go to school.

He started taking MMS the 29th of June 2011, when he was only taking two doses of one drop and two drops every third day has intestines were not moving and he had a psychotic attack (something that I've never seen before) and this lasted 2 hours.

Children with Autism are severely intoxicated so his reaction to the MMS was very clear. After 15 days of taking 8 takes of a drop a day Martin stopped coughing, he is more alert, connects all the time, his expression changes, he started to make friends, he starts to play with the computer, I hear from school that he'll be a preschooler because all the improvement he has had and that he didn't have all year. (I was almost going to get him out of school) he was attending school with a shadow teacher but had no apparent social communication or increase in cognitive capabilities.

This past October, in a meeting with his team of therapists, I asked them all what they thought about Martin's improvements and they were very happy, I asked 2 of his speech therapists who have worked with him for more than three years, "Did Martin have severe autism?" "Yes." They both responded. "And now?" I asked them. "Noooo" they both agreed. The rest of the team, who have only known Martin for a few months looked at each other, and said that they could not imagine Martin like that.

The path out of autism is scary, but MMS is giving us great joy and hope.

Martin has not been sick anymore, the last time he was sick, before MMS, we had to admit him to the hospital, his cough was so bad that he was vomiting non-stop.

First off we give thanks to God, we thank Jim, Andreas, my dear Kerri who guides me with all her love, and to every single mother that is in this struggle. I also thank Mr. Luis from Buenos Aires, Argentina who has made it easy for me to get MMS; he is a wonderful human being who helps those in need.

We are on the way to full recovery for Martin, I'm sure of that.

---

154. So I just redid my daughter's ATEC. She was a 71 a month and a half ago and is now a 60!!! I thought she was making nice improvements but it really set in when I saw her score. Some things that have changed for her are now she is pooping on the toilet every day and telling me. She used to just go every few nights in her diaper. There's no more diarrhea, she's not constipated, she's sleeping great. She is just starting to show signs of imaginative play and a big one for us is that she just learned to blow. She couldn't blow no matter what we did to try to teach her. Now she's got it. Hopefully more good things to come.

---

155. My now 8-year-old daughter was dx with autism at 18 months old. Started GF/CF diet and started with a DAN! She slowly made progress with speech, had a few words, a few 2 and 3 word sentences, and despite having autism, she was always a happy, calm, sweet little girl. Shortly after her 3rd birthday, she "changed.' She became more of a Jekyll / Hyde personality, and all progress in speech and improvements in eye contact and sociability started to fade. Seemed like everything we did was just like yo-yoing back and forth between improving, and then crashing yet once again.

Last year at this time, we had just left the doctor we had been taking her to. For years she was on antivirals, antifungals, antibiotics on and off, SSRI's, Tenex, and when her violent SIB had become so bad the second year into treatment with him, he put her on Abilify. SSRI's and the Abilify did nothing for her except make *her behave worse. Neither did any of the other things he had her on. The first year she was doing pretty good, but possibly because the antivirals were acting as an anti-inflammatory. She always had chronic constipation with loose stools/diarrhea. Nothing helped that either. We finally left that doctor who yelled at us in his office for 20 min straight, in front of my daughter, because he did not like that I had been challenging him on all of this for the past 6 months. I was no longer going to let him put the blame on us and everyone else... that her ABA and speech therapists didn't know what they were doing, and telling us she acted out this way because we didn't know how to give proper time outs or how to correctly give "pep talks" at bedtime. I was convinced that she had PANDAS/PANS and/or parasites that were causing her ever increasing SIB, loss of all progress she had made in speech over the past 4 years, and the fact that she had not grown in 4 years...did not gain 1 pound, feet did not grow at all, and she only grew 1/2 inch in height in that 4 year span. We had taken her to a very well respected endocrinologist the previous summer who ran all the tests. Could find nothing wrong as to why she wasn't growing. But he decided to put her on HGH anyway. Thankfully, we left and decided not to give her the HGH. We also took her to one of the best Ped Psychiatric doctors at UCLA who specialized in autism that summer. He observed her, her SIB happened right there in his office for him to see. He said her SIB is coming from pain, discomfort, something medical going on inside, most likely stemming from her gut. He suggested we get her scoped upper and lower. So we did. The only thing that came up was on the upper endoscopy, and that was that she had esophageal gastritis. So the GI dr gave her Nexium.*

All the parasite stool tests we had done over the years had always come back neg except for one at age 3 came back with Giardia. But that was it. So last January, when we had finally left that her doctor of 3 years, her liver AST and ALT were at 130 and 135, her Creatinine (kidneys) were severely elevated at a dangerous level, she now had hypothyroidism, and after testing her cortisol thru blood and saliva we discovered she was barely making any at all. She had dark circles under her eyes, grinding her teeth for 3 years now, and agitated beyond belief ALL the time. Mind you when we started with him 3 years prior, her thyroid, liver, and kidneys were all in normal range. She did have dark circles, many many food allergies, some SIB, but nothing compared what we were dealing with now, or what was to come.

So now, what to do. We took her to a very highly recommended Pediatric GI in Los Angeles who was very nice. Listened to what we had to say. Ran the most sensitive and comprehensive stool analysis. Came back with nothing at all. Then took her to the head of pediatric Infectious Disease at Cedars Sinai. She obviously had her mind made up before she even walked in the door. She saw the word AUTISM on my daughter's chart, and that was it. I showed her every blood and stool test, explained the not growing at all for 4 years despite all the endocrinology labs said everything related to growth was in normal range, her severe SIB. She said to us..."I know you want to help your child. You want to find something "medical" that has caused her autism so you can fix it. But the truth is, your daughter has autism. You need to accept that, and take her to a pediatric psych dr who can help her with these SIB's. There are lots of meds that can help her be calmer." Had she not listened to one damn word I told her!!! All the meds we had tried in the past did nothing, already took her to a psych dr who specialized in autism telling us something internal was wrong. She would not run any blood tests, or any other kind of Infec Disease tests. So with that, my *husband, my sister, and I completely dumbfounded started to walk out. The dr then had the nerve to say..."She hasn't been vaccinated since 18 months (she was 7 here) Let's get her caught up right now before you go. I can give her 6 vaccines today." I looked at her and said you have got to be crazy! I am not giving my very sick child vaccines full of the poisons that did this to her."*

So a month later we started with doctor who "gets it." Who listens, and treats parents with respect. Values their opinions. And the networking I had been doing for the past 6 months on a few different bio-med autism groups with other moms was paying off. The moms are the ones who KNOW!!! That's how I found my daughter's current doctor and how I found Dr Maile Pouls last June to help her with nutritional and metabolic healing. Dr Pouls ran a $100, 24 hour collection urine analysis, and we found out she had severe malabsorption, Ph was too Alkaline, she was catabolic, extremely electrolyte and mineral deficient, severely Vit C and D deficient, and had severe bowel toxicity. Working with Dr Pouls and her new doctor lead them to suspect parasites/worms. I was encouraged to look into mms, but of course I had heard nothing but bad things like it's bleach. I was very hesitant to try it, but I researched it, and sought out other moms on Facebook who were doing it. The toxins from the parasites had completely taken over her body and brain. She was SIB almost all day, every day by now. It was pure hell. We thought we were going to lose our minds it was so bad. We had stopped taking her anywhere but school and dr's appointments for the past 2 years because she would just suddenly out of nowhere, for no reason, violently freak out and there was no way to help her calm down. Her doctor said she's so toxic, and her body is so sick that she can't detox. That the mms will not only help kill, but will neutralize the toxins and help her calm down. So we started mms, and I could not believe how much calmer, happier, and more present she was in just a week. MMS is literally saving her life, and bringing her back to us. Before starting mms, we did just parasite meds like Alinia and Mebendazole. The parasite meds alone did not help her, even though she was dumping worms, but she was still completely psychotic, and getting more crazy and manic every day. Since starting mms she has dumped hundreds of worms, some 10-12" long, ascaris eggs, TONS and TONS of the shedded skins of the worms, tons of liver flukes, and hundreds of tapeworm segments. We had not done any PP yet. This had all been with MMS only! We actually started Albendazole for treating tapeworm about a month ago and saw immediate improvements with it. It really seems like this is her biggest beast right now to deal with.

These are her labs from before MMS, and after starting MMS:

2/7/2012 - one month after leaving scumbag dr of 3 years

AST = 79 ALT = 108 Creatinine = 1.24 EOS = 11.2 Sed Rate = 9

~ Started MMS 11/23/2012 ~

12/4/2012

AST = 73 ALT = 64 Creatinine = 1.32

1/15/2013

AST = 61 ALT = 60 Creatinine = 1.30

2/27/2012

AST = 51 ALT = 42 Creatinine = 0.92 EOS = 5.1 Sed rate = 12

Ref Range for AST is [15 - 46] ALT is [3 - 35]

Ref Range for Creatinine (kidney) is [0.60 - 1.20] Shows how well kidneys are working. Anything close to or over the 1.20 is considered to be of serious concern

*Ref Range for EOS (eosinophils) is [0.00 - 3.0] * High EOS are always seen with parasites*

Ref Range for Sed Rate (marker for inflammation) is [0 - 10]

Her doctor said her Sed Rate is likely high because when killing off pathogens and detoxing, inflammation will go up temporarily.

So I say never give up. Don't listen to doctors who don't listen to you. Kerri, thank God you saw those bottles of MMS at the clinic that day and asked what they were. And that you bought some out of curiosity and tried it. And that you selflessly have taken what you've learned and experienced, and shared it. Given so much of your time to help others, to help heal and recover their kids. You are truly an amazing person. We finally have the answers to our daughters autism after almost 7 years of numerous doctors and specialists, so many tests, so much money and valuable time wasted... and now we finally have real hope...real results. We know without a doubt she is going to get healthy, which will in turn give her the happy joyful life back that she once had before she got "autism." ~ xoxo

---

156. *Today was just wonderful! We have been on MMS since 12/11 and this is our second round of Ivermectin from our Dr, just started today. My son has been very rigid lately, but not today. For the first time in over 6 months we went to the park and had a great time (no stimming), then we were able to run an errand and pick up books from the library. (he usually gets upset unless we do one specific route in the car). Felt like a regular mom running kids to the park and one errand. So lovely!!!*

---

157. *Just a testimonial to MMS. For those of you new to the group and still unsure of MMS protocol I wanted to tell you about us. I was afraid at first, but we decided to try MMS. Within 3 weeks my son went from an ATEC of 36 to 18! That was amazing. Then we were afraid of the enemas. We said no way, we would just do oral MMS. of course after seeing others progress here we decided to do them. He would calm down! Then for months we agonized over the PP. We were afraid of it. We were scared of the meds, everything. Finally we decided to try it too. We are in the middle of our first PP and our son, who used to hit, kick, bite and scream at me over the smallest things all the time, has now been a perfect joy. When he gets mad it is over quickly and does not escalate into world war 3. Also these bumps he had on his face have cleared up along with his attitude! So if you are unsure of MMS like I was, here is one more testimonial of just how good the protocol really is.*

Thanks Kerri, all the other moms that came before me!

---

158. *Dear Kerri and the readers,*

I would like to say that we started giving MMS to my son Filip, autistic/PDD NOS, when he turned 5 years old, after having tried lots of things (diet, many different supplements, HBOT, some sort of behavioral therapy Teach and ABA), we have seen enormous changes already in the first month.

I had heard about MMS from a mom of an autistic child. I started reading about it, but was not impressed by the comments I read on the internet. I contacted *a few mums that I knew from the Son-Rise programme in the US and they have reassured me that it works miracles.*

I was however very sceptical of MMS, as I have read so many "negative" comments on the internet. This is why I decided to start it myself for a few months to see what would happen to me. Also, I had to figure out how I was gonna dose MMS 8 - times a day, when I am working 10 hours a day and Filip is going to the kindergarden.

In fact, I was just feeling great when I started using MMS.

I wanted to start with my son when on holiday, but my husband displaced MMS bottles by mistake. I was so mad. I have only found them the last day when we were returning home. I have immediately decided to give one drop to Filip before we started to drive. I need to underline that my husband was very sceptic. Filip would at that time speak, but it would mostly be scripting and rarely would he say things in context. Otherwise he also had other autism symptoms- tantrums, no friends, low muscular tone, no drawing, obsessive occupations (in Filip's case watching cartoons and reading books), no sense of danger, wandering out, nose picking, scratching but, toe walk etc. But suddenly Filip started commenting everything in the car, asking questions, observing the things he would see from the car etc. Me and my husband looked at each other and my husband commented " Maybe it is working after all".

Filip's ATEC was 48 at the time we started. At start, my mother-in-law helped out in dosing and we were able to dose 8 times a day. Filip's ATEC fell to 32 the first month. Unfortunately, my mother-in-law read some horrible stories on the internet and the FDA's advice was detrimental for her decision to stop dosing Filip, even though she was seeing wonderful results, she was scared and said that she is only believing the official medicine, admitting that the official medicine has nothing to offer our son.

I was very disappointed, but continued to dose as much as I could, however not reaching the recommended 8 times a day. I was therefore only able to dose 4-5 times a day. I started telework on Friday's in order to at least get 8-10 doses in on Fridays, Saturdays and Sundays and 4-5 doses the rest of the week. My son was great in November (the volunteers working on Son-Rise with Filip were commenting on unprecedented success), but his ATEC only fell to 28, I would say due to low dosing the progress was slower. Filip was having some hard times in January, when he was more hyperactive. At that time we have also introduced the GcMaf injections, which could have added to Filip's hyperactiveness.

But since two weeks now Filip is back to his shape. He is so communicative, uses unseen vocabulary, amazes us every day, he started drawing without being prompted, communicates with everyone, was invited to a birthday from a girl at the kindergarden, started to dress without prompting, started to tell me:" Common mummy, hurry up, let's go." or "Where are you mummy? What are you doing?" or when he occasionally wakes up during the night " Is it day already? or " I should not go on the street, because a car can hit me" or " Daddy, you are not sad, you look angry" or just yesterday when he draw a picture in the kindergarden he said to his teacher " I want to show it to my mum". The list goes on and on and every day I have to thank the all mighty for having led me to Kerri and to the MMS protocols. I know we are on the right path to recovery and I am not scared of the future any more. And, finally, i just did his ATEC to check out his ATEC- it is 16!!!!!! LOVE IT! LOVE IT!!!!!!

Kerri, thank you!!!!!!!!!

159. *My daughter has had really bad SIB for over 4 years. We've been in the ER for massive goose-egg bumps on head, and she always had severe bruising on her tailbone and spine from dropping herself down, butt first and legs in a "W" position, onto the floor, and then would fling and flail herself into walls, tables, furniture, bang her head on all of the above and the floor, a rage like she was the girl from the Exorcist. Nothing could stop it. It just had to run its course...like a tornado. Comes out of nowhere, and it will disappear when it's ready to. Nothing can stop it.*

    But after starting MMS, it was 80% better right away. Then after 3 weeks it started coming back. For a month it was really bad again, and we could not move up past 2 drops without her going crazy. Then after you had us switch to the HCl, she was remarkably better in about 5-6 days. She has had a few "episodes" in the past 2 weeks, but now they can last 30 min max, where before mms they would last anywhere from 15 min - 2 hours literally non-stop. The year before starting mms, there were times I had to lay her on the bed or couch and just sit on top of her, like on her butt so she couldn't hurt herself. I would just lose strength after about 30-45 min of trying to hold her.

---

160. *Hey everyone! Just wanted to give a positive report about the HCl activator. My son is 12, non-verbal, very hyper and lots of food intolerances. I got mine in the mail today, gave a quick 3 drop dose to my son before he went out the door to school, noticed he was a bit more compliant and calm than usual after school. have given 2 more doses since coming home and he has been doing his usual, but he is much more calm today than he has been in months!!! I gave him a snack and he took it and started to eat and I told him: can you say thank you mom? and he actually responded with a sorta grumbled "graaaa mooooom" which was to me "thanks mom."*

---

161. *Thanking God for Kerri Rivera and happy here too. My 15 y/o is in the chorus of his HS play. He has to do 5 costume changes, his own hair and makeup and several song/dance numbers, which he did perfectly on opening night! He is not your typical MMS/autism kid. He was pretty much recovered, starting with biomed at age 2.5, until age 11 when he hit puberty and he started getting rigid and anxious. We had some supplements going and were starting to look at stem cells when we found Kerri. He is not on the strict MMS protocol but takes it when he can and does PP and enemas each month. He is so much better since starting this last August! So thankful for all the info and support from this group.*

---

162. *Hi Kerri,*

    Last week I contacted you regarding our 8 year old son and his focus issues and the disappearing act his mathematical skills have taken, and we've done our best follow through on your advice to up the number of doses to 16 per day. I visited my son's Integrative Medicine doctor this morning, who has been his biomed doctor for the last 3 or so years to do a live cell microscopy, and to see if he had any ideas about the focus issues our son has been having. I told my son's doctor in detail what we've been up to over the last 10 months (MMS Protocol, Parasite Protocol) and he took it all in.

    When he looked under the microscope he was shocked to see that our son's nasty dysbiosis issues were gone! He told me that he's had almost no luck completely getting rid of dysbiosis in a case like my son's. The yeast beast is gone! He still has evidence of leaky gut but the drainage looks great. He was so excited, our 20 minute scheduled visit turned into 40 minutes as he wanted *to know all about what we did so he can find a way to use it in his practice. He also remarked that you must be one bright lady to come up with the protocol, but we already knew that. :-) He had us fill out an ATEC when we first visited him back in 2010 and our son was at 55. He dropped into the 30's when we started him on Carnitine but the MMS protocol has brought him down to the 6. The doctor also noticed the distinct lack of autistic behaviours, so it's nice to get some outside feedback to help confirm what we are showing on the ATEC. I asked him directly if he thought we should change anything at this point and was direct...don't change anything keep going. And so we continue on our quest for the zero ATEC. :-) About two years ago we did a chelation challenge test (DDI) and found our son extremely high in lead and so we proceeded with IV chelation at that time. He ants us to do a chelation challenge test to check the metals status just to rule out lead and the possibility that metals might be still at the heart of our son's lack of focus and more ADHD like symptoms.*

He wants me to send him some information on your MMS protocol. He attempted to use chlorine dioxide about 10 years ago and found that he couldn't get the results that you are getting. I'll send him to MMS Autism website but can I include your contact info (email)?

Thank you for being so patient with us and everyone. As you know full well when it involves your own kid the emotions come out and it's hard to remain logical.

---

163. Drum roll....

I have been dying to share...

Been off MMS for a few weeks...AND HOLY SMOKES we are ALLLLL GOOD!!!!!!!!!!!!!!!!

Summary...

My daughter—PANDAS—MAJOR sleep issues—fluctuate between diarrhea and constipation

Me—Acne—horrible—and nail fungus

ALL GONE. ALL.

NOBODY could EVER, ever, ever tell me why I broke out so much esp. with our squeaky clean diet.

First PP, OMG it was REALLY bad—worst ever—and became cystic—painful—so for SURE worse before better in my case BUT believed in protocol and KEPT GOING.

By month 4, GONE and yet to return

Fungus—gone in month 1, never to return

My daughter:

Well, you know, AMAZING!!!!!

YOU, Kerri, YOU helped me figure out FINALLLLLY after a zillion docs, WHY MY KID could NOT SLEEP! EVER! 7 LONG years... long long—exhasuting years—and it was PARASITES NO DOUBT in my mind! Month 1 of PP, literally like a light switch, SLEEPING —falling asleep ON HER OWN—(not with a circus routine to "get her to sleep") AND staying asleep ALL NIGHT LONG! UNREAL! This is HUGE!

AND bye bye PANDAS (Siver/Aloe nipped it, MMS maintained her gains, *BEAUTIFULLY!) SO I do believe in our Silver Aloe protocol for PANDAS, 100% worked for us! AND MMS. Both.*

AND since OFF MMS, my daughter is 100% pooping NORMAL—OMG first time in her WHOLE LIFE!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! (on MMS it was still fluctuating...)

AND she is eating things she would have NEVER ever touched! EVER! Raw veggies, nuts, trying EVERYTHING I make... unreal... this is HUGE too... the "healthier" you get, your tastes DO change... and one becomes less "picky"

I have learned that SOME probiotics can literally throw a child into a PANDAS rage--and also brought back MY symptoms—literally within 12 hours of consuming. (Like "Regarding Caroline" posted) This clearly happened to us. Mental note: NOT all prob's are appropriate for everyone. We react to fermented food—so I am going to experiment with no prob's and with THeralac __ONLY THeralac BTW

Now,

We are back to silver/aloe, daily, as prevention

We will still take DE (prevention)

Vit D3

Trace Mineral Drops

Enemas as needed

AND I am going to go BACK to the herbs for 3 months around full moon "just in case"—

I CANNOT WAIT to meet you in person, THANK YOU, and share our story with EVERYONE that will "hear me".

Again, my daughter did not have 'autism' per se, but autoimmune issues, meaning that, perhaps, her treatment protocol is 'shorter' then someone who is "more sick' etc. if that makes sense.

We also did a boat load of work prior to this. Meaning diet in place, for instance, for 4 years already—so healing began. This was the final layer for us. The frosting on the cake!!!!!

I have 2 brand new bottles of MMS in my cabinet :) AND will ALWAYS have it on stock!

With MUCH LOVE and respect to you Kerri.

-Jeni, IL, USA

---

164. My daughter was born Sept 9 2010 and she had pretty good health until she had 15 months, she was able to say five to ten words by then, she had all her vaccines on time as recommended by our pediatrician, and she was really lovely and happy but suddenly every start to change slowly, we noticed she start losing the words she had learned, she use to wave people saying hi, and suddenly she was not interested on waving anymore and she start to isolate herself little by little to the point she stop being a happy girl, she barely looked at us and we started to worry about it, anyway, I believe the first thought every parent may have think in this very same situation is that her personality is different than other kids and is something temporal.

When she had 18 months we went to the regular doctor check and they asked us to get the opinion of a development pediatrician since her social and language skills were behind the expected, and this started the worst nightmare ever.

It seems like every single development pediatrician has a waiting list huge enough to have more than a couple months totally scheduled, we called on May and the first available appointment was 4 months later, we called like 4 or 5 of the recommended pediatricians and institutions but it was all the same thing, but we finally got a pediatrician, not form the recommended list, but he made room for us in a couple of weeks and we were able to see him in a couple of weeks.

The diagnostic was mild autism, the recommended treatment was ABA therapy and the follow up was going to be 1 year later since he did not have may expectations of positive changes in the next 6 months, we were devastated by this, my wife was in denial, and I started our journey to recovery the very same day, I started researching and suddenly I found tons of information about autism and people who claimed to have recovered their children from autism.

I sent emails everywhere and I never imagined I would have such a great response from every person I wrote extending me a helping hand on all my doubts with the most sincere and honest recommendations, 2 or 3 days later I found Autismo2.com website and I saw all the videos available.

I bought 3 or 4 books on amazon, I saw many hours of recordings and seminars in Spanish and English, and I was amazed about some kids losing their diagnostic, but also about how they were affected by heavy metals, bacteria, viruses and all sort of things.

At the beginning I thought my daughter could not be sick, she has never had a single cold, how come? but a home heavy metal screen test gave us the first clue, and some lab tests confirmed the very same problems all our kiddos share, mercury levels on the roof, candida, yeast, you name it.

I started the GFCFSF diet the first weekend after her confirmed diagnosis and just one week later we started noticing changes, then I started with baths with bentonite clay, and noticed more improvements, we also started ABA on July, and little by little our daughter started to improve, later on we introduced MMS baths, Quinton water, DE, probiotics, and she improved every day.

I honestly believe that the early diagnosis and prompt reaction from our side played a big role on her fast recovery, It happened that we never canceled the first appointment that was made with the most recommended development pediatrician, by then we were looking for a second opinion and a new evaluation due to her huge improvements, we never imagined this pediatrician will say what.

Although my history with this disease or condition has been very short nobody knows how much pain, stress and suffering had caused to our family, we now enjoy our daughter more than ever when we see her playing with her old sister and taking away every single concern we had, we continue keeping an eye on her development but she is doing really well, I wish nothing more than every single mother I have hear on the discussion groups I follow have the very same results we have had with all the guidance from such smart and courageous parents like Kerri Rivera and many others that shed us the light on our darkest night.

I am more than grateful with all of those parents, I would do anything on my reach to help them, we need to help every single kid and family that goes through this disease to recover our kids.

UPDATE:

I'm very happy to forward the last ATEC from my daughter, today we came from her 2 year check up with the new pediatrician and she found her pretty well, she just told us" It's pretty hard to believe she was diagnosed with mild autism keep *doing whatever you are doing, since I would say she is acting like a regular kid for her age."*

I would say that keeping in mind that most of her ATEC score is on the Language section, for a 2 year old and from what we came from she is on her way to full recovery (Previous ATEC was 55, at this writing was a 23)

---

165. I found Jim on the Internet 2 years ago, and I immediately started Jim's protocol for MMS. Unfortunately we were so loaded with mercury, so when I reaches 8 drops of mms, and my son 2 drops (after more than 2 months on slow upgrading) we were really sick, again. All the heavy metals that where released with mms could not get out of us. That´s when I found zeolites, and that was our way back.

Now, two years later, my arthritis is gone, and now I know it was mms that killed the clostridium bacteria that I had been carrying since I was 13 (I am 52 now). 6 months ago my son fell off his bike and got an open wound. It was infected and some hours later the blood was black in his hand. Now I had a choice, to go to the doctor and give him antibiotics, which I knew would kill the bacteria but make him worse, or do it my own way. I gave him 2 x 3 drops of mms, and the blood was ok in some hours, I thought. Some days later he started to wave his hand. I thought he had eaten something else that he could not digest, but it got worse. After 4 weeks he was not only waving/flapping both his hands, but also his head. Then I started get suspicious. I tested him for bacteria (I use a Rayonex PS 10 for testing, bioresonance) and I found Clostridium botulism.

Now I had a choice again. I had a fully booked week, but his father came and took him home to his island, where they could use MMS every day and be in a calm environment. He got 2 x 3 drops of mms, for 6 days. After 3 days his father called and said "Eva this is scary, the flapping is getting worse". And I smiled and calmed him down, the bacteria are dying and the toxins are being released, I said. The zeolites took care of the toxins, and after 4 week he was ok again.

My lesson in this was, apart from the clostridium toxins, that I also found an imbalance in my son on Condrogenesis with the Rayonex. When I was 13 I stepped on a rusty nail. The blood was poisoned in the foot, but I did not dare tell my parents. After 3 days the whole leg up to my knee was black. I got antibiotics, of course. At 18 I was diagnosed arthritis in my knee. I think I had small amounts of bacteria left, as my son had, and the toxins from clostridium bacteria degenerates the cartilage near the infection. After some years one develop arthritis. I did, but it's gone now!

Eva, Sweden

---

166. Drum roll please.......my son is 17 years old. He's been on mms for 5 months now, I'm so excited I don't even know, could it be 6? I think we started late July. He started with a 63.......he is now a 7. He went from 63 to 25 in just a few weeks, then a 13 a couple months later, now a 7. I feel like I'm cheating every time. I go back and look, argue with myself, nitpick....but whether he's a 5, 7 or 9, it's nothing short of a miracle because just one year ago, he was WAY over 100.

The ATEC is not perfect, not even close. There are things that the ATEC does not show. All that is true, but this is incredible improvement that no doctor I know would have ever been able to accomplish. Been there, did that, spent the 425K+.

And no, it's not just MMS, but that has been the main treatment and everything else we do supports MMS and general health because I believe that in the end, *it's a battle between your immune system...and all that's attacking it. We are winning and I sure do wish they would STOP CALLING IT AUTISM. That alone hurts so many.*

Here's a little background on my son: Matthew was very high functioning from age 6-12. That was after diet since 3, floortime play therapy, some aba, AIT, IVIG, chelation for a year, OIG, years of infusions of glutathione, vitamin C, B-com, etc...immune stuff. Then he did very well GFCF ages 6-12, ran track (very well), 2nd degree black belt, boxing lessons, etc.... Then puberty hit, and what I did not know about, Lyme and co-infections RAGED, and for around 3 years, I lost him cognitively, worse, worse, worse. At 15, he was like an advanced Alzheimer's patient, ADVANCED, he had lost ALL short-term memory, it was horrible, he struggled to think and could not.

He became more and more violent; choking me, kicking me, did I say he was a blackbelt? Even though he lost everything cognitively, like the ability to answer a question, those karate skills popped back into his head and he nearly killed me.... really close a couple times.

That was a year ago when his ATEC was over 100. So, here's what worked: 1. Low amylose diet helped with the constant urinating. (he acted like a diabetic and had 14 of 16 PANDAS symptoms so auto-immune) 2. Biofeedback helped calm him some. 3. Cholestyramine helped him psychologically - I could see him clear up mentally in minutes. (chemical/mold/dust sensitivities big time) By this time, his doc had him on Ketamine for pain. 4. PEMF (pulsed electromagnetic field device) stopped the pain in 6 weeks, got him off of Vicodin, which was not working anyway at double doses, and all the Ketamine. I hated drugging him. Got him off the psych drugs over 6 months time. Then he got a lot better but communication did not really come around until 5. MMS. BAM. Big improvements in communication and PERSONALITY. With MMS, even his laugh changed to a typical sounding teenager's laugh, and for the first time in years, he can sit still and stop pacing, pacing pacing....the parasites were eating him alive....and now more recently, I know 6. Hydro-colonics is taking us to another level of clear communication. Now there's lots of spontaneous language.

I write things on Facebook so that I can remember and maybe write a book one day. I don't care how bad things get, believe me, your kids can get better and it's amazing, they really do store up all this info that you didn't think was there. It is there, and they will share it with you one day.

I hope one day Matthew can share his perspective with others. It's just that it has taken a very long time for us to help him understand that the things he did back then were not in his control so I'm careful not to talk about it. For months he would out of the blue apologize and feel terrible, over and over. Then there's memories of his dad losing it also....protecting me.

Thankfully, Matthew has blacked out or forgotten a lot of it. He did say he remembers being locked in his room. We had to turn the lock around and sometimes lock him in there. Bad memories. We were all traumatized. But not anymore. He is a different person today but he's always been the sweetest boy I know. I knew that, even back then. One day, he'll probably talk more about it.

---

167. Hi Kerri,

First of all a HUGE big thank you for what you are doing with our children! My daughter has been on MMS two weeks now and she is like a different child!!!!!!!! Every day she gets better...By leaps and bounds. She WILL be one of your recovered ones ;)

UPDATE:

My daughter has now been on MMS for three weeks. ATEC down from 78 to 44. This IS a miracle. We have been in absolute desperation for almost three years trying to figure out what in the world is going on with our child. After hundreds of thousands of dollars, experts all over the world, LA, New York, San Fransisco... and nothing would work. Minor improvements but nothing that I could say really worked. Everyone always agreed that she probably does not have "autism" but what is making her behave that way?!?!?! So a long story short... If I could I would wrap my arms around all of you helping my daughter and so many others through MMS until I was sure you truly understand what your dedication means and that you are saving lives...literally <3 This is by far the biggest thing that has ever happened to my family. We are forever thankful.

UPDATE #2

I sent you my daughter's ATEC a week ago but redid it today. We started MMS in December and I decided I will do the test every month. Here are the results, 4 more points gone!!!! In 1 month from 78 to 40... Almost half off!!!!!! Her team is blown away ;) all we need now is that language score to come down... Kerri my goal is to have her recovered by A1!!!

Update #3

We are seeing such amazing improvements every single day. I think the true testament is the daily feedback from people working with her who do not know we are doing MMS. Here is part of a note from her SLP " her communication on the whole is truly remarkable- words, iPad, eye contact...everything is improving!!!!!!! So exciting. Has anyone told you lately that your daughter is a superstar? You should be so proud. :)" And I am...

Our main challenge has always been anxiety and today we had a breakthrough on that front. Instead of clinging to me for her dear life when I separate from her at school she took her friends hand, waived AND said goodbye and walked away... Here is the best part, she then turned her head, waived and gave me the biggest smile . Her face at that moment is forever engraved to my heart.

Also for the first time EVER my son was able to have a friend come over from school in our car without her having a major meltdown. Instead she shared her iPad with the friend to show him her program and use it TOGETHER!

Honestly I think our major challenge at the moment is for the rest of the family to heal with her... We had lost all hope and settled for life with Autism. We can now look at the future very differently and actually look forward to it! Love our new life

UPDATE #4

I have never really posted gains but this is too huge not to share. We started MMS 3,5 months ago and at that point I could hardly take my daughter to the grocery store. And even if I did it was a guaranteed meltdown. Well we just got home from a two week long vacation that included travel by air, time difference, different language, going from hot weather to freezing temps and piles *of snow, totally different foods, nothing but strangers to her, visit after visit to new places...The list goes on and on. The BEST part - the vacation was full of excitement, fun, laughter and happiness every single day I have to pinch myself to believe this is my life after such a short time on MMS. Thank you all for your support and guidance I have my life back and can't wait for the future now! We still have a looooong way to go but we'll get there.*

Today is a wonderful day already

---

168. MMS Works!

It's my number one pathogen fighter. Our family continually suffers from Strep, my son specifically from Pandas. MMS remains and continues to keep our strep symptoms at bay. I have used everything for our bacterial flares (Abx, herbals, supplements even homeopathics) and nothing has worked so effectively like MMS does for us.

Recently I changed over to the HCL activator and within in a few days we heard my severely apraxic son's first WORD "Mama" Its been just over a month on the HCL activator and he is now stringing two words together!! MMS is a huge part of our recovery protocol. I feel relief beyond measure that I found a post on MMS and decided to try it, and mostly I'm immensely grateful to Kerri for putting the MMS protocol together and for her love for our kids.

---

169. We just had a good meeting with all of Amor's teachers and therapists. In the last few months there has been big jumps in Amor's cognitive skills.

I used to worry that Amor might have dyslexia. I never pushed her to read even if was behind her peers. I know frustration will only make things worse. Her 'x' and 'k', 'b' and 'd' were always mixed up. I have been observing this mix up for the past 2 years. Early this year, her speech therapist agreed that dyslexia might be an issue and we were ready to start studying dyslexia management.

However, Amor now shows interest in reading both phonics and sight words. She can answer letter names and sounds precisely. The same speech therapist now says that we shouldn't look at dyslexia anymore. It might have simply been a mix-up as part of early learning. Amor's dyslexia issues in 2 years all of a sudden went away in 2 months??

Amor is also doing well with number recognition and improving in one-to-one correspondence. This is something her former speech therapist (back in Japan) was working on for a year...with little progress.

All these seem so simple but these are great strides from someone who has vision issues. Amor also wears glasses.

If I were to plot Amor's improvements with time, the biggest jumped occurred in the past few months, which of course can easily be explained by MMS.

MMS clears the gut of bad bacteria, yeast and parasites. All these affect sensory processing thus inhibiting the child's ability to understand her environment and people around her. Clean the gut, remove autism...in Amor's case, remove GDD.

We are also running a son-rise lifestyle program. I.e. we don't put fixed hours into her program but use the son-rise attitude and son-rise her in the weekends.

170. *Luca came into our lives July 26, 2010, he was premature, but all my kids were born in week 35, so I was not surprised. He was perfect, I have thousands of videos of him looking into my eyes, smiling, playing with his sisters, following daddy to the door when he was ready to leave to work and saying bye bye. He walked at one, and was a perfect little baby. It is hard to know exactly when he changed but I can tell you when I started to get worried, and saw him behaving differently. I had a neighbor the same age as my son, and when they got together to play I was surprised at how my neighbors son followed simple directions and looked into my eyes every time I called his name. When I had play dates with my friends all the moms were so relaxed, and while they were talking and kids were playing with toys, I was always chasing Luca and trying to keep him in the same room because he was not interested in toys or playing with other kids.*

My first trip to the Zoo with my three kids was a disaster, Luca didn't want to move from the Monkey's area and had his first melt down, all over the floor screaming and crying. I remember seeing all the families with kids the same age as my son, and seeing how they understood, and paid attention to what their mothers or fathers were saying. For me that was enough to feel for the first time something was not right, and it was not that I was a bad mother, I knew deep in my heart it had nothing to do with discipline.

I enrolled him in a school thinking that would help. The first day when I picked him up the teacher said they needed to move him with the babies because he was not ready for the 2-year-old class. That day I came home just knowing - and despite all my family, friends, husband and Pediatrician thinking I was crazy, we got him evaluated and the rest is history. On November 15th, 2012 it was official and we got the papers. Classic Autism, by that time we had read so much, watched hundreds of videos and searched every day on the Internet for interventions, schools, therapy etc.

I remember I wanted to start ABA the same day that I got the diagnosis, but the waiting list was really long and we would have to wait. So we waited and started an early intervention program in the Kent District. Surprisingly, I think we took the diagnosis really well; we just wanted to help him as soon as possible.

After the first month I stopped asking for a miracle or cure. I even asked all family and friends to do the same. I needed to accept it!!! My miracle was already here. My husband wanted to do the Gluten free diet and I was not even 100% in agreement...why??? He is such a blessing; let our miracle eat whatever he wants!!! I agreed to do the diet anyway, but wasn't really interested. I just had the feeling my husband was not "getting IT" or accepting autism = our new precious miracle. Now I can recognize and see clearly I was in total denial too. I was so afraid to let anybody know, even myself, that I didn't want this for Luca. I tried and tried and tried, but a voice inside of me was not in peace. I spent a lot of time crying and looking for recovery videos on YouTube, imagining Luca being one of them, and was fascinated with those stories.

Then one day I found Kerri Rivera's conference in Bulgaria, it was a "Godcidence," (like my father always says) I wasn't looking for MMS, I didn't know anything about that. I remember I watched the conference 3 times, and I felt butterflies for the first time since I got the diagnosis. It was like that voice was finally quiet and listening to every single word! "I want this."

I did my homework and found a lot of negative information about MMS but I did not care…that voice was so strong that I knew, the same way I knew about

Luca being autistic, that this was something I needed to explore. I looked for her in FB and to my surprise she answered me in minutes, when I asked her about the testimonies she connected me with some moms with recovered kids, REAL moms that decided to believe too. After I saw pictures from before and after and reading about their journeys I felt Hope!!! I knew deep in my heart this was the way and the path we should take. So I got my case together and when I was ready I talked to my husband about it.

It was easier than I expected, we wanted so badly to recover our son!!! It was not an easy process but finally after almost 3 months I GOT IT…WE GOT IT...THERE IS NOTHING WRONG TO WANT THIS, that is ALL we want for Luca!!!! For the first time I recognized my kid's autism was not a blessing, he was sick; GI problems, extreme constipation, food issues, sensory problems... and that voice (God voice) sounds really clear and loud; I NEED TO HELP MY KID TO RECOVER!!!!!!

On February 18th we started a Dietary Intervention with all Kerri's recommendations, the one diet that we were doing was full of sugar and carbohydrates, so we started to be really serious about all the food choices. We kept a very detailed food diary and ordered the MMS. By March 12th when our bottles arrived at our doorstep we were ready to start Kerri Rivera's Baby Bottle Protocol. March 13th we did our first bottle and that same day I asked Luca to look for a ball in the other room, and give it to me, and he did.

It was like he was sleeping soundly, and then he woke up. He was extremely alert and he was looking in a different way. At first I thought I was crazy, but then my husband noticed it too and the changes keep coming. He improved dramatically in all his therapy sessions, his waiting time, sitting at the table, responding to commands, pointing, eye contact, everything improved!

Kerri recommended doing an ATEC when we started back on February 18th, so I did, his score was 64 points. By April 1st he dropped 20 points! He now follows simple directions, plays with his sisters, recognizes them by name, hugs me and says mommy and daddy, has close to 200 words, and every day is a surprise for us.

I'm proud to say my son is on his way to RECOVERY AND HE WILL RECOVER, HE WILL BE RECOVERED!!!!!!!! I'm loving Kerri Rivera's Protocol, I'm loving my kid dropPing 20 points in his ATEC in a month, I'm loving searching his diaper, and inspecting his poop finding parasites, I know it seems impossible but I truly get excited! (Probably some mothers would understand) I'm loving this side of the picture FULL of hope. HOPE. I don't think I can leave this road now, I don't think I want my son to live in his "perfect" world and not in mine... I'm a fighter and thanks to Kerri I'm a believer ...THAT... is the blessing right there HOPE AND FAITH IN GOD IS ALL I HAVE AT THE MOMENT… Luca is really special, but it's not the autism that makeS him special, autism sucks!!! Overcoming it and all the obstacles he encounters every day, and doing what he is doing now is what makes him THE MOST SPECIAL KID ON THE PLANET, HE ALWAYS WILL BE SPECIAL TO ME NO MATTER WHAT, AUTISM OR NOT….I see it in his eyes now, he is a fighter like me, he is telling me what to do, recovery is soooooo in the picture!!!!!! I can actually see it!!!! I'm a HAPPY Mom!!!!!!!!!!!!!!!!!!!!!!!

-Alma

171. *Today is a wonderful day already. My daughter got up to her alarm, took a shower without help, got dressed, packed her lunch, woke up her sisters, and is now getting breakfast for everyone ready. It amazes me how independent she is becoming, and how much our life is changing because of her new independence.*

She is 10 and we started MMS in Sept. She had done 5 PP and following the PANDAS protocol the last week after emailing Kerri Rivera last weekend with concerns about eye fluttering. She is also on Fish oil, garlic, Magnesium, and GABA. Before she was very hyper, spacey, a lot of language, but not using it correctly. She was extremely impulsive, and would fixate on object to play with and ignore others. Now she does reciprocal play, she talks back and forth with her friends, she cracks jokes that are now actually funny. She is focused with her school work and not hyper - except in typical situations.

This week my husband forgot to pick her up to take her to therapy and she came to my classroom to tell me how irresponsible he was, and that she was disappointed in his inconsistent behaviors. I had to laugh, but at the same time I was so proud as I explained to her the whole situation because we where having a fabulous discussion about life. She is doing so well. Thank you to all of you who have restored some of my hope that there is a cure.

---

172. *We are now in our 7th month of using MMS (CD) and in our 4th year of treating B.D., our son, who was diagnosed Autism. Our first week of MMS (CD) brought on a skill that we had spent years trying to accomplish, over night potty training. Week 1 of MMS (CD) and B.D. was out of pull-ups and in underwear. We weren't even close to full dose and already seeing positives. We knew we were on to something big at this point. One month later, we all witnessed my son blow out his birthday candles and open birthday gifts. Probably the first time he knew this event was a birthday and it was meant for him. As a gift to him, we took a trip to Disneyland. Not only did B.D. walk the entire park at our side, but he never once cried, whined, or objected to anything on the trip. Prior to this, he had to be in a stroller or wagon at all times.*

B.D.'s skills continued to flourish over the next few months. We enrolled B.D. in swimming lessons hoping he would feel comfortable without a life jacket on and clinging to an adult. This hope turned into a major accomplishment. B.D. has passed all Beginner levels in swimming and is now on to Intermediate. He is now swimming unassisted in the deep end of the pool! I used to hold my breath during his sessions hoping he wouldn't do anything that would scare the other kids in the pool. I still hold my breath, but now I'm at the end of my seat recording him as he's giving every single bit of effort to learn from his instructor and pass another level!

Just the other day, we went to a birthday party that was held in a warehouse full of screaming kids, loud music, inflatable jumpy houses and slides. This place can over stimulate just about anyone! In the past, my son would vocally stim, cover his ears, run behind the jumpers where the generators are (trying to turn them off) and require me, or his Dad to keep him engaged with the rest of the kids. Well, not anymore! He blended right in with the rest of the kids! Not a single stim, didn't cover his ears, stood in lines, listened and followed the rules and instructions, stood in lines and PLAYED with the other kids. His Dad and I were able to hang out with all the other parents, watching him run and whiz by with all the other kids as if that is how it has always been. Only it hasn't. He has come so very far in the past 7 months. We have a 6 year old son who no longer has sound sensitivity, loves to eat whatever is placed in front of him, DRINKS *tons of WATER, has manners, shows empathy, is careful and aware of other kids (especially babies and toddlers), making progress academically, and so very affectionate. B.D. isn't recovered… YET, but we expect it to happen now more than ever. We have to put our shades on because the end of the tunnel just got closer and brighter!*

---

173. *My husband and I were told to come to the 2nd grade assembly this morning. They gave awards to 2 kids in each of the 4 second grade classes. One was for outstanding in academics, and the other was for most improved in Self Control, which is the "character trait of the month" for the whole school. Kayla was given the award for most improved in Self Control in her class! She walked up and got her award and stood up there with the other kids holding theirs. Up until a few months ago, she would have violent, self-injury tantrums at school every day... throughout the day, lasting anywhere from 10 - 90 min. She could never get through a class singing performance without losing it and tantruming, screaming through the whole thing. Her teacher came up to me on Wed. and said she just can't get over how happy my daughter is lately, and that all the teachers and staff in the school have really noticed too. FUA!!!! I'm so proud of my girl!*

---

174. *Our third child (K) was diagnosed with moderate autism at 21 months of age. K was very ill. He suffered from vomiting and diarrhea mixed with severe constipation for years. At first, we began every intervention we could afford. The first step was removing all gluten, caseine and soy. We saw a decrease in stimming behaviour and an increased ability to learn. We implemented many vitamins and B12 injections, cranio sacral therapy, anti-fungals, HBOT and an even stricter diet removing all artificial colours, flavours, preservatives and most sugar. Our son was making slow but steady gains. ABA therapy continued from when he was two until just this past March (4.5 years). K was seeming much better but despite these gains, was never close to "normal".*

In July of 2012 a friend suggested MMS. I did my research and so it began. In three days K started to sleep through the night. After less than one week at a very low dose, K stopped the constant verbal scripting and obsessive speech- his most evident sign of autism. In that same week he began dressing himself independently. In the coming months, K's ATEC began dropping steadily. He began to make jokes, initiate social interactions, show empathy, cope with noise and busy places and best of all- understood safety. The risk of him running away was a memory. We could go places and he would stay near us. Teachers and Educational assistants were raving about his progress, unaware of the protocol we had begun. His most significant changes happened after a three months of the Parasite Protocol, using mms, Vermox, rompepiedras, neem and diatomaceous earth. Through enemas, he expelled worms as long as 24 inches to my husband and family doctor's disbelief. What K expels in his stool now looks normal and ordinary.

His ATEC in July 2012 was 46. Ten months later he has a score of 6.

He is undecipherable within his classroom and amongst his peers. He laughs, plays and acts like any other child his age. His academics are at par with his peers.

Our son does not appear to have autism anymore and only a trained eye would suspect that he ever had it to begin with. We thank Kerri Rivera, her Facebook group for all their support and MMS for being the final, and most significant leg in our sons' journey to recovery.

Maria A., Ontario Canada

175. *My severe autistic boy has been now doing wonderful!!! This protocol has completely changed my life!! He's been dumping tons of parasites since we started MMS about three months ago we're now in the second pp...he's now much calmer more happier!! His ATEC dropped down from 80 to 48 in three months only...that was truly amazing he is now gaining weight, I feel like I'm getting my child back to normal again!! I wish I could tell the whole world how Kerri Rivera has changed thousands of people's lives!!! God bless u Kerri for all your hard work to help others!!!*

---

176. *My son Nick started MMS just after his 7th birthday. Within 6 months he had cut his ATEC score in half from 66 to 37. Nick showed speech gains from the very first drop of MMS after having made no changes in speech for two years. Really, Nick hadn't made any improvements at all for two years. He was in a total stall until we began MMS. We raised Nick's oral dose of MMS very slowly taking two months to get to his target dose and never experienced die off that was unreasonable. Three significant changes are obvious in Nick's health at this point; he has been off of his Mitochondrial cocktail (for several years he has had blood labs pointing towards Mito dysfunction) for over 8 months and we haven't seen any of the lethargy that used to concern us so much, he was able to wean off of his seizure medication for his absence seizures and we haven't seen any return of seizure activity, and his thyroid appears to be beginning to shift from his long standing diagnosis of hypothyroid. The nicest changes we see are in the vibrancy of his health and how happy he is! His eyes sparkle and he has lots of energy for the day. He is no longer sitting on the couch staring out the window waiting for someone to figure out how to help him feel better. Since beginning MMS, every day Nick shows us just a little bit more of the boy that he is underneath the veil of autism and sickness. He tells us what his favorite things are, what he wants to do, and his sense of humor is coming shining through. MMS has given me a ton of hope that we just might get our boy back!"*

---

177. *After 7 years of biomedical treatments, numerous therapies, and more doctors than I care to think about, our son, who was diagnosed with autism, could speak a full sentence, was potty trained, and could finally go to the mall and movie theater without screaming and putting his hands over his ears.*

Victory??...yes...however, we knew we had so much farther to go. You see, as your child gets older, you realize that he is missing the key things in life that are crucial for independence. Our son had become "stuck" if you will, he was not socially maturing, and still lacked empathy, mature connections with peers, and proper communication with teachers and the other adults in his life. He would still yell when he didn't win a game, and he did not always react in an age appropriate manner, if something did not go his way.

My research went on as always...searching, searching, searching...we were not getting anywhere with his current doctor, and I knew that we had to change our current plan for treatment.

A good friend of mine called me to tell me about a new product that some of our warrior mom friends were trying. The letters MMS were spoken to me for the first time in March of 2012. My son was 11 years old and I was desperate for help. The first thing I asked was, "How much does it cost?" $25.00...what?? It can't be! We have spent our life savings, and now I am looking at something that cost less than a night out for dinner?? I decided to check it out!

I emailed Kerri Rivera and told her that I wanted to start the MMS autism protocol, and I'll never forget her quick response..."YIPEE", she said, let me know if you have any questions, I am here for you.

I read, studied, and called my friend about twenty times over the next couple of days, and then we began the next chapter in our journey.

About three weeks into the protocol, our son seemed calmer, as well as aware of others on a more mature level...we were excited! Once he was at full dose, we began, baths and enemas, and he continued to improve. Our next move was to try the 72/2 protocol, which is giving MMS every two hours around the clock for 72 hours.

Well, it was as if someone had kicked our sons maturity clock into high gear. We saw HUGE gains!

"Hey mom, I like your dress...you look really nice today"....WHAT??

"I'll try that new kind of chicken for lunch"...WHAT??

"I feel like I'm beginning a new life"...WHAT??

"It's OK that I didn't win the race, I had fun"...WHAT??

I wanted to call the Good Morning America Show and say, LOOK AT WHAT MY SON IS DOING!

He also began to grow and put on weight for the first time in three years! He put on 10 pounds and 2 shoe sizes in about 2 months!

The gains continue to come, and our boy is coming alive...recovering...finally it's here! My eyes well up with tears as I write these words...the words I have dreamt about for eight long years.

He celebrated his 12th birthday on January 3rd, and has told us that 2013 is going to be the best year of his life...YES...it is!

He is our angel, our fighter, our miracle. Thanks be to God for MMS!

We are blessed and forever grateful~

---

178. Once upon a time there was a little boy who had a strange mind, he hooted and he danced and he did not learn like the other children around him. So, he was sent to a special school. For his whole educational life he was separated, because he could not express correctly the answers put before him.

As he grew older he seemed sad and very much in pain. He appeared to try to stop the pain by clawing at his face and arms. He also stopped talking. We, his mom and dad, were very sad.

Then an angel came into our lives and told us of a magic potion.

"But" said the angel. "You must give him the potion every hour of the day and night for him to be better."

So the little boys' mom and dad have been doing that now for 9 weeks and 4 days. Every hour.

Now in those hours the mom and dad still cry. Some, because the little boy has been hurting for so long, and..... now they know why. But they also cry because they see that their little boy is not hurting like before. They are happy for his healing, and that he does not claw at his body any more.

And they are also happy because their little boy is talking again. And he is laughing. And the things that he knows surprise them. We thought he was not learning…turns out he was.

179. *Feeling really good still about my Natalie's progress from the combination of anti-parasitic medication, the supportive components of the PP and CD (HCL version). Natalie is about 63-64 lbs and she takes 2 drops of CD an hour both at home and at school. We don't wake her for CD in the middle of the night; but if she wakes for the restroom or something, then we give her another 2 drops (we just keep it pre-mixed in the bathroom near her bedroom). CD has been the single most important catalyst to Natalie's improvements/healing over the past 10.5 months; utilizing all three modes of delivery (oral, baths and enemas). Kerri having escorted CD into the ASD community is like a Miracle, in my opinion. Sorry... I don't mean to be corny nor impose my Beliefs on anybody else; but I sometimes find it hard to contain my Joy now and I get chills just thinking about all of the children now feeling better and/or beginning to feel better and better with CD. Before CD, it was hard to feel Hope for my special little Natalie; who was formerly suffering very extreme PANS flares and re-flares monthly as well as rapid-onset grand mal (tonic clonic) seizures every 2.5-3 weeks on top of her ASD and global developmental disabilities. Now Natalie's PANS 'flares' if/when they occur are waaaayyy less intense and better-controlled/'snufffed out' with CD than the former roller-coaster of oral antibiotics that her Drs had tried for years, her seizure activity is decreasing significantly and we've even begun a slow reduction in her anti-seizure med Lamictal ODT. We're optimistic that Natalie will be able to remain stable on less and less of that med, as we feel it might be causing her some nasty side effects. Our Hope for her is that she might even be able to come fully off of Lamictal ODT, as we continue CD and PP for her. Time will tell. While she's still a 'work in progress' [Note: She's 14 years old and been ill since infancy and it's a bit harder to heal longterm infection(s) and/or infestation(s)], we do see lowering ATECS in Natalie's future and brighter and brighter days. Patience is a virtue. So Grateful. Thank you Kerri and all you supportive Moms & Dads. More and more healing to come to ALL of our children!*

---

180. *Hi Kerri,*

Thank you for helping my daughter, age 15, recover from Autism. Thank you for the MMS/CD program and the support to carry it out. Here is a summary of her recovery:

June 19, 2012 Atec=52

Started with MMS slowly, adding one drop a day

GF/CF/SF diet

It took her 1 1/2 months to get to full dose.

Noticed improvement in physical health & behavior

Began parasite program

Sept. 18, 2012 Atec=24

Continued with 6-8 doses/day & baths

More improvements

School discontinued IEP, because she was doing so well academically and socially

January 15, 2013 Atec=19

My daughter, who is not vaccinated, comes down with Chicken Pox

No one at school or in community had it before her. It appears to be an isolated case.

It may have been in her system and released as parasites died?

Fever of 102 degrees for two weeks was the lead up to the Chicken Pox. Seemed like her immune system woke up! After chicken pox, huge gains! That is when we noticed she was recovered!

April 25, 2013 Atec=6

Yeah!

* * *

Please share your experiences!

Testimonials are one of the best ways to share your experiences with this protocol. Perhaps you learned about it by reading or watching a video testimonial?

If you don't tell us your experiences, we can't share them or take action on issues that need improving or correcting...

Send your testimonials to:

testimonials@cdautism.org

Also, let us know if we are free to publish your testimonial, with or without your name.

It's in our interest to take care of others. Self-centredness is opposed to basic human nature. In our own interest as human beings we need to pay attention to our inner values. Sometimes people think compassion is only of help to others, while we get no benefit. This is a mistake. When you concern yourself with others, you naturally develop a sense of self-confidence. To help others takes courage and inner strength.

~ Dalai Lama

Appendix 2

CD Autism Worldwide

Since September of 2011 over 5,000 families in 58 countries have used CD to help with an autism spectrum disorder. Here is a list of all the countries who have at least one family using CD to help heal the symptoms known as autism. This list is constantly updated at www.CDAutism.org.

- Argentina
- Australia
- Austria
- Belgium
- Bolivia
- Brazil
- Bulgaria
- Canada
- Chile
- China
- Colombia
- Costa Rica
- Cuba
- Czech Republic
- Dominican Republic
- Ecuador
- Egypt
- England
- France
- Germany
- Ghana
- Greece
- Hungary
- India
- Iraq
- Ireland
- Italy
- Japan
- Jordan
- Kenya
- Kuwait
- Malaysia
- Mexico
- Montenegro
- Morocco
- New Zealand
- Norway
- Panama
- Peru
- Philippines
- Poland
- Portugal
- Puerto Rico
- Qatar
- Russia
- Scotland
- Singapore
- Slovakia
- Slovenia
- Somalia
- South Africa
- Spain
- Swaziland
- Sweden
- Tanzania
- Turkey
- United Arab Emirates
- United States
- Venezuela

If you are using CD for autism in a country not listed above please email us so we can add you to the list... kim@cdautism.org.

Conformity is the jailer of freedom and the enemy of growth.

~ Robert F. Kennedy

Appendix 3

AutismOne on Healing Autism: Accidental Cure by Optimists

by Simon Yu, MD

Accidents happen. We accept the reality of them as is and we usually move on. On the other hand, accidental discovery is another story. For some people, they ask questions like, “why me?” Or keep asking for a deeper meaning: Is this really an accident or a message to understand? An accident is a fertile ground to find out if you are a pessimist or optimist.

Tom Jacobs, story teller from Kansas City, recently told a story about a pessimist and optimist: Two sales men were assigned to Africa in the early years of the shoe industry to sell their shoes. One week later, one salesman telegrammed to his boss, “Business situation in Africa is hopeless. Nobody wears shoes”. On the other side of the continent, the other salesman excitedly telegrammed to his boss, “Unlimited potential in Africa. Nobody wears shoes!”

I don’t know about you but I would rather be an optimist. I always encourage my patients to look at the bright side and be cautiously optimistic no matter how grim their conditions might be. Hope is a powerful driving force to promote healing. Pessimists do not call themselves pessimists. Just realists.

Today, I saw a 77 year old patient with a history of chronic pulmonary fibrosis, bronchiectasis and rheumatoid arthritis with a recent diagnosis of stage one bladder cancer. She has been depressed and joined a cancer support group because of feeling a sense of hopelessness.

After a long discussion, I told her how lucky she is that her lung and rheumatoid conditions have been stable and she has only a relatively benign early stage of bladder cancer. I told her if you are going to have cancer, this is the kind of cancer to have and she should be thankful. All of a sudden, her doom and gloom mood was lifted with a broad smile. The rest of our session was uplifting.

The AutismOne 2013 conference was held in Chicago during Memorial Day weekend. I was invited to give a talk on parasites, allergies, and autism. Autism is not my field. I usually see adults as an Internist. However, beginning in 2012, I have been seeing autistic children with rather interesting responses. In May 2013, I wrote an article about a medical hypothesis of a relationship between parasite infection and autism in preparation for my lecture for the AutismOne conference.

I attended a full day of lectures and was awestruck by the thousands of autism parents from all over the country attending the lectures. Multiple lectures were conducted by many autism specialists at the same time covering a large variety of topics. This conference was driven by a grass roots movement of parents with a hope to help their children. Here is a short synopsis of the lectures.

Dr. Anju Usman, MD from the Chicago area covered gut-brain connection and biofilms by pathogens in the gut. Symptoms such as depression, anxiety, poor attention and focus, and obsessive compulsive behaviors may be related to the delicate balance of bugs which produce a mucous slime known as biofilm.

Andreas Ludwig Kalcker, Ph.D. Bio-Physicist from Spain, who studied under German bio-physicist Fritz Albert Popp, Ph.D., discussed parasites in depth. He demonstrated a successful treatment for more than 65 children around the world with his parasite protocol over a one year period.

Kerri Rivera, from Mexico, one of the main leaders of the group running the AutismOne conference, showed many cases of successfully treated autistic children including her child. She has been using chlorine dioxide in conjunction with diet, nutritional supplements, detox, and hyperbaric oxygen.

By the time I presented my medical hypothesis on parasite infection and autism, they had already had numerous discussions about parasites and had been saying all along how parasite infection might be one of the major underlying problems for autism that has been overlooked. I felt like I was repeating what had already been presented.

Dr. Andreas Kalcker and Kerri Rivera collaborated using chlorine dioxide for two years with prescribed parasite medications, albendazole and pyrantel pamoate, to turn around autistic children. Professionally, I have no experience using chlorine dioxide on my patients. The difference with my therapy was that I was using acupuncture meridian assessment as a guide to detect and treat parasites.

If what they are reporting is even partially true for these autistic children, it would be a major breakthrough in the autism community. Some of the audience was crying in excitement but some were skeptical and saying it sounded too good to be true. I could feel the excitement of the audience but also a sense of fear that Kerri Rivera and Andreas Ludwig Kalcker will be attacked by special interest groups. Chlorine dioxide is too inexpensive to a fault.

Is this an accidental finding leading into an accidental cure for autism? It is too early to tell but any chronic medical conditions like cancer, heart disease, diabetes, arthritis, Alzheimer's dementia, or autism seems driven by epigenetic influences from environmental toxins, parasites, hidden dental problems, and faulty diet and nutrition.

I spent several hours with Andreas Kalcker after our lectures were over. We found a common ground for treating parasites: his experience as a bio-physicist and my experience as a military medical officer. Few people truly understand and are aware of the magnitude of parasite problems. This man was very passionate to rescue these children from the scourge of autism.

Kerri Rivera just published a book called [Healing the Symptoms Known as Autism]. I just finished the book. This book is for everyone but especially for both autism parents and all medical professionals involved in the care of autistic children. The book contains much important information that is not available in the main stream autism community.

Most pediatricians might be sympathetic but they are rather pessimistic regarding the care of autistic children. Andreas Kalcker and Kerri Rivera are true optimists in the midst of skeptics and pessimists. They see unlimited potential for the cure for autism based on diet, nutrition, and parasite eradications using chlorine dioxide and parasite medications. If you want to know more about autism and what is possible, I highly recommend reading [Healing the Symptoms Known as Autism] by Kerri Rivera. The book explains in detail the Kalcker parasite protocol.

Dr. Simon Yu, M.D. is a Board Certified Internist. He practices Internal Medicine with an emphasis on Alternative Medicine to use the best each has to offer. For more articles and information about alternative medicine as well as patient success stories, and Dr. Yu's revolutionary health book Accidental Cure: Extraordinary Medicine for Extraordinary Patients, visit his web site at...

www.PreventionAndHealing.com

...or call Prevention and Healing, Inc., 314-432-7802. You can also attend a free monthly presentation and discussion by Dr. Yu on Alternative Medicine at his office on the second Tuesday each month at 6:30 pm. Call to verify the date. Seating is limited, arrive early.

Appendix 4

Autism Evaluation Treatment Checklist (ATEC)

The next page contains a copy of the original *Autism Evaluation Treatment Checklist* (ATEC) developed by the late Dr. Bernard Rimland, the founder of the *Autism Research Institute*. We are not suggesting you use this particular copy. It is only included here for those not familiar with the evaluation so you have an idea what the ATEC involves since we mention it throughout this book.

The survey consists of a series of questions to be answered by a parent about what they are observing in their child's CURRENT behavior. It is not uncommon for both parents to take the survey separately and compare/average the scores.

The best way to take the survey is online at the *Autism Research Institute's* website. Here is the direct link to the current survey:

www.autism.com/index.php/ind_atec_survey

After completing the questions, a score is generated that can range from 0 to 180. 0 means no autism while 180 would mean that the answer to every question was in the most negative extreme.

We consider a child recovered from autism if their score is between 0 and 10.

You are encouraged to complete an ATEC before starting the Protocol, and every 3 months thereafter so you can clearly see the progress you are making.

Also, please share your results with us; we collect data on the children whose parents follow the Protocol as explained in this book. Send your name, date of the survey, general location (city, state, country) and score to:

atec@cdautism.org

If you have a history of ATEC results while on the protocol, please include them all.

NOTE: Do not send questions to this email address! You are unlikely to get any response from this address! It is just a collection point.

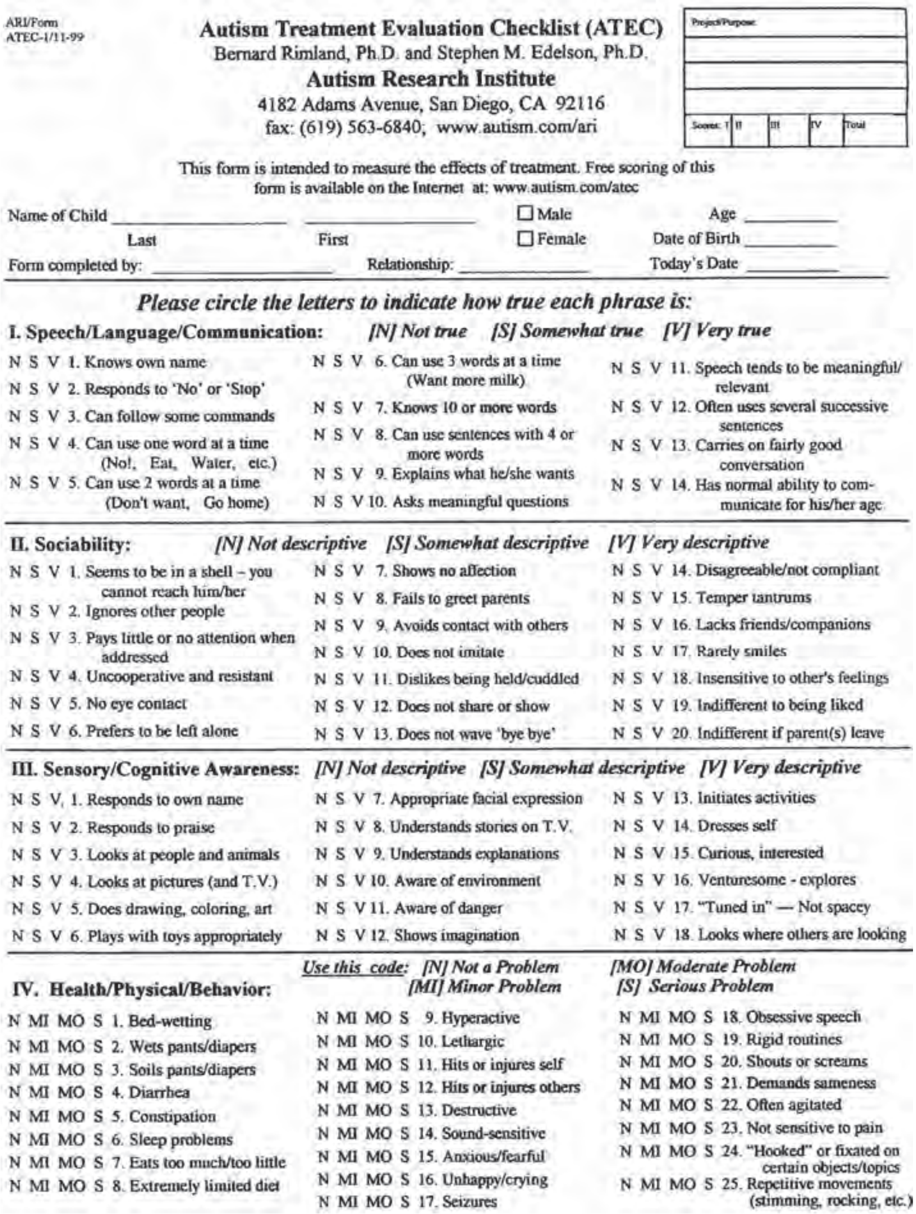

ARI/Form
ATEC-1/11-99

Autism Treatment Evaluation Checklist (ATEC)

Bernard Rimland, Ph.D. and Stephen M. Edelson, Ph.D.
Autism Research Institute
4182 Adams Avenue, San Diego, CA 92116
fax: (619) 563-6840; www.autism.com/ari

Project/Purpose				
Score: I	II	III	IV	Total

This form is intended to measure the effects of treatment. Free scoring of this form is available on the Internet at: www.autism.com/atec

Name of Child ______________ ______________ ☐ Male Age ________
Last First ☐ Female Date of Birth ________
Form completed by: ______________ Relationship: ______________ Today's Date ________

Please circle the letters to indicate how true each phrase is:

I. Speech/Language/Communication: ***[N] Not true [S] Somewhat true [V] Very true***

N S V 1. Knows own name
N S V 2. Responds to 'No' or 'Stop'
N S V 3. Can follow some commands
N S V 4. Can use one word at a time (No!, Eat, Water, etc.)
N S V 5. Can use 2 words at a time (Don't want, Go home)
N S V 6. Can use 3 words at a time (Want more milk)
N S V 7. Knows 10 or more words
N S V 8. Can use sentences with 4 or more words
N S V 9. Explains what he/she wants
N S V 10. Asks meaningful questions
N S V 11. Speech tends to be meaningful/ relevant
N S V 12. Often uses several successive sentences
N S V 13. Carries on fairly good conversation
N S V 14. Has normal ability to communicate for his/her age

II. Sociability: ***[N] Not descriptive [S] Somewhat descriptive [V] Very descriptive***

N S V 1. Seems to be in a shell – you cannot reach him/her
N S V 2. Ignores other people
N S V 3. Pays little or no attention when addressed
N S V 4. Uncooperative and resistant
N S V 5. No eye contact
N S V 6. Prefers to be left alone
N S V 7. Shows no affection
N S V 8. Fails to greet parents
N S V 9. Avoids contact with others
N S V 10. Does not imitate
N S V 11. Dislikes being held/cuddled
N S V 12. Does not share or show
N S V 13. Does not wave 'bye bye'
N S V 14. Disagreeable/not compliant
N S V 15. Temper tantrums
N S V 16. Lacks friends/companions
N S V 17. Rarely smiles
N S V 18. Insensitive to other's feelings
N S V 19. Indifferent to being liked
N S V 20. Indifferent if parent(s) leave

III. Sensory/Cognitive Awareness: ***[N] Not descriptive [S] Somewhat descriptive [V] Very descriptive***

N S V 1. Responds to own name
N S V 2. Responds to praise
N S V 3. Looks at people and animals
N S V 4. Looks at pictures (and T.V.)
N S V 5. Does drawing, coloring, art
N S V 6. Plays with toys appropriately
N S V 7. Appropriate facial expression
N S V 8. Understands stories on T.V.
N S V 9. Understands explanations
N S V 10. Aware of environment
N S V 11. Aware of danger
N S V 12. Shows imagination
N S V 13. Initiates activities
N S V 14. Dresses self
N S V 15. Curious, interested
N S V 16. Venturesome - explores
N S V 17. "Tuned in" — Not spacey
N S V 18. Looks where others are looking

IV. Health/Physical/Behavior: *Use this code:* ***[N] Not a Problem [MI] Minor Problem [MO] Moderate Problem [S] Serious Problem***

N MI MO S 1. Bed-wetting
N MI MO S 2. Wets pants/diapers
N MI MO S 3. Soils pants/diapers
N MI MO S 4. Diarrhea
N MI MO S 5. Constipation
N MI MO S 6. Sleep problems
N MI MO S 7. Eats too much/too little
N MI MO S 8. Extremely limited diet
N MI MO S 9. Hyperactive
N MI MO S 10. Lethargic
N MI MO S 11. Hits or injures self
N MI MO S 12. Hits or injures others
N MI MO S 13. Destructive
N MI MO S 14. Sound-sensitive
N MI MO S 15. Anxious/fearful
N MI MO S 16. Unhappy/crying
N MI MO S 17. Seizures
N MI MO S 18. Obsessive speech
N MI MO S 19. Rigid routines
N MI MO S 20. Shouts or screams
N MI MO S 21. Demands sameness
N MI MO S 22. Often agitated
N MI MO S 23. Not sensitive to pain
N MI MO S 24. "Hooked" or fixated on certain objects/topics
N MI MO S 25. Repetitive movements (stimming, rocking, etc.)

Appendix 5

MOLECULAR MIMICRY
What It Is
&
How it Relates to
The Gluten Syndrome

by Mrs. Olive Kaiser

After reading the short answers in *The Diet* Chapter (page 64), we understand (simply) the basics of how some of these departments of the immune system function. It is easier to grasp how, through molecular mimicry, gluten can damage so many different tissues in different people, and sometimes cause other foods that *look like* gluten to also be reactive. Before we go any further, I would like to thank the following professionals for their contributions to our family's well-being and a wider understanding of the gluten syndrome and other related topics. If you would like to conduct further research into this topic, the body of work that the following professionals have contributed would be a great place to start. There is so much more to learn.

Dr. Alessio Fasano, MD, shook America awake on gluten awareness in the 1990's and in 2003 published a landmark paper[1] on one small subset of the gluten syndrome, villi damaged celiac disease. That study provided the impetus that brought gluten awareness to the table (or sadly, OFF many tables) across our nation.

Dr. Thomas O'Bryan, DC, Functional Medicine, and international gluten syndrome educator - www.thedr.com. Dr. O'Bryan is my functional medicine doctor. He taught me much of what I learned about gluten and other topics and led me to his mentor, Dr. Aristo Vojdani. Later Dr. O'Bryan's amazing research review seminars on gluten contributed hugely to my knowledge base and bolstered my confidence to manage our new lifestyle in a then difficult social era.

Dr. Aristo Vojdani, PhD., MsC, Immunologist, CEO owner of *Immunosciences Lab*, Los Angeles, CA. Dr. Vojdani is also chief scientific advisor of *Cyrex Laboratories*, Phoenix, AZ. Dr. Vojdani has published nearly 150 excellent research papers.[2,3,4] They fit the community like a glove.

Dr. Rodney Ford, MD, pediatric gastroenterologist, NZ, grasped the extent of neurological injury,[5] and "silenced nerves phenomenon."[6] This compellingly suggests why many of us do not recognize symptoms of gluten damage until we are in deep trouble. This builds on the previous work of Dr. Marios Hadjivassiliou, Professor of Neurology, Sheffield, UK.[7]

Dr. Kenneth Fine, MD, gastroenterologist, Dallas, Texas, owner of an investigative research lab, *Enterolab* (www.Enterolab.com). His lab ran the only accurate gluten antibody tests our family received back in 2004.

Without the courage of these astute researchers, we would still be wandering in the dark. Thank you all, and other individuals who thought outside the box. Kerri you are one of those thinkers. Thank you!

Lastly, special thanks to **LuEllen Giera**, my support group leader, for leading me to Dr. O'Bryan and Dr. Vojdani.

1. What is Molecular Mimicry? How does gluten damage us?

The structure of gluten resembles the structure of many of our body tissues. When the immune system attacks gluten or partly digested "pieces" of gluten it may also attack body tissues that "look like" those pieces of gluten. There may also be other processes that we do not yet understand.

What is gluten? - Gluten is a stretchy protein found in some bread grains. The problematic types are found in wheat, barley, and rye, and now early research suspects possibly rice, corn and oats (Dr. Peter Osborne, DC, CCN, www.glutenfreesociety.org).

What are proteins? - Proteins are a class of materials found in living tissues, such as hair, nerves, enzymes, etc. Molecularly, all proteins look like necklaces of beads strung into various color sequences. The different sequences make the proteins different, and the "colored beads" represent 22 separate amino acids. Our digestive system uses enzymes to cut up these necklaces into single beads so they are small enough to cross the gut wall properly and be restrung into new proteins. See image on page 452.

Unfortunately, many folks today are toxic and poorly nourished and do not have strong digestion, so the gluten "necklaces" may never completely break down. As weaker bonds between the beads break, "pieces", called "peptides," of gluten are formed. Gliadin is a well known gluten piece and there are many others.

Our immune system - has "departments" to protect us in various ways, including IgA, IgG, IgE, IgM, IgD and others. IgE can cause immediate allergic reactions to bee stings or peanuts, etc. IgA, IgG and IgM may react more slowly with less drama. They all manufacture "workers" or "soldiers", called antibodies, each custom designed to patrol our bodies, looking for the bead sequence of one particular enemy. When they find that bead sequence, they "tag or stick onto" it. Our killer white blood cells interpret the antibody tag as a condemned sign and know to surround and destroy that protein.

If an antibody test lists "Gliadin – IgA", "Gliadin – IgG", and"Gliadin – IgM" it means the test checked for gliadin antibodies in the IgA, IgG and IgM departments. If the antibodies are high it means the immune system is does not like gliadin and is working to destroy it.

Weak, leaky barriers - Our gut wall and other barrier membranes such as the skin, lung, placental, and blood brain barriers are held together with "tight junction proteins" that act like velcro. Inflammation, parasites, gluten, medications, infections, electrosmog, etc, may damage the "velcro" or open them up too much. Substances may slip through them into places they should not be and cause trouble.

Unfortunately if the gut wall does not hold together well, i.e., is "leaky", pieces of incompletely digested gluten strings (and others) may slip through and run into the immune system—our dutiful "guard dog" on the other side. Due to their too-large size "he" may raise the alarm. The invader strings are "frisked out" i.e., examined. If they are rejected, one or more of the immune departments make matching antibodies to "tag" them so the killer cells know to go after them.

It is at this point more problems may arise. A gluten antibody may "run by" a natural body tissue, for example a nerve in the heart. It may see in that innocent nerve tissue a sequence of "beads" that matches or partly matches the gluten sequence it was designed to "tag" and stick to that section of the nerve protein instead. This, unhappily, attracts killer cells to the misidentified nerve, resulting in autoimmune injury to the nerve. This is molecular mimicry.

Other foods and proteins beside gluten may also initiate molecular mimicry when they prematurely cross a leaky gut, and may prompt allergies (IgE) or intolerances (IgA, IgM, or IgG). However specialists agree that gluten's particular "bead" (amino acid) sequences are uniquely guilty in their ability to upset the immune system and instigate molecular mimicry between "their" antibodies and our tissues.

Molecular mimicry can take place between non-food proteins and body tissues also. Infectious microbe sequences such as strep, for example, are believed to partially match, (in the case of strep), heart and joint proteins and so injure the heart muscle or valves, and (rheumatic) joints. Ditto is suspected for root canal and cavitation infections that may circulate and injure specific tissues, including the heart. Flu microbe sequences can resemble gluten and may trigger or surface gluten syndrome. There are numerous other examples of mimicry between infections, foods and body tissues.

Consider the image below. An antibody may be made to seek the shorter "bead" sequence of a gluten piece that crossed the gut wall, but it might also recognize a similar sequence, a partial match, in the longer nerve protein sequence. The antibody "tags" or "sticks to", the part of the nerve tissue that matches the gluten piece that it seeks.

<table>
<tr><td colspan="16">Gluten Peptide (and antibody) Sequence
Note: the different colors represent different amino acids.</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>pink</td><td>blue</td><td>white</td><td>brown</td><td>brown</td><td>brown</td><td>red</td><td>red</td><td>green</td><td></td><td></td></tr>
<tr><td colspan="16">Nerve Sequence</td></tr>
<tr><td>orange</td><td>orange</td><td>yellow</td><td>green</td><td>red</td><td>red</td><td>blue</td><td>white</td><td>brown</td><td>brown</td><td>brown</td><td>red</td><td>orange</td><td>green</td><td>blue</td><td>yellow</td></tr>
</table>

It gets more complicated, but eventually the victim tissue is attacked by killer cells on an ongoing basis. This process may be slow but stealthy. In time, the nerve may malfunction and the organ it serves may also begin to fail; for example, heart failure. This is termed autoimmunity. The body attacks its own tissue.

Molecular mimicry can occur in hundreds of places in the body since gluten can break up into numerous pieces having amino acid sequences that partly match many of our natural tissues. This may explain why gluten syndrome presents so differently in different people.[2]

Once this process starts, like elephants who never forget, the memory B cells that manufacture the antibodies never forget the sequence of the "invader" gluten pieces. If gluten is removed, antibody manufacture eventually stops, but any time the immune system sees even a tiny amount of that protein sequence, antibody production is retriggered. This is why specialists insist the gluten-free diet is strict and lifelong.

Gluten removal stops production of the initial antibody, which decreases the inflammatory process significantly. Literature shows the gluten-free diet may reverse or improve many serious conditions, but it may not always arrest an advanced autoimmune disease. Prevention is key.

Example: Once a ball is pushed down a hill, further pushing will only send it downhill faster, but it will roll on its own. To correlate, molecular mimicry between the gluten antibody and victim tissue starts the ball rolling. Removal of gluten stops pushing the process, and the sooner the better. The longer the exposure[8], and to other factors as well in a triad of autoimmunity (a. environmental stress, toxins, infections, etc.; b. a faulty blood brain barrier; and c. susceptible genetics), the higher the chance of autoimmune disease.

Dr. Vojdani comments, "If it is detected early enough and steps are taken early enough the condition may be reversed. If any condition is advanced enough you can reach a point when simple removal may not be enough. Autistic children generally have autoimmune reactivity rather than full-blown autoimmune disease, which is why they show great improvement upon the removal of gluten."

2. Does gluten always damage the villi of the small intestine as the celiac story teaches? Many other tissues such as thyroid, pancreas, blood vessels, joints, brain, nerves, liver, bone, etc. may be involved in this disorder. Does all that other damage arise due to poor nutrient absorption from injured gut villi?

No, according to published research many researchers and practitioners now believe the villi are not always damaged in an autoimmune gluten reaction.[2] Where there is no villi damage, injury to other organs cannot be due to nutrient deficiencies caused by villi damage. Molecular mimicry provides a mechanism for direct autoimmune gluten injury to many other tissues and organs when the villi are fine. It is also possible that several tissues and organs, *including villi,* may be injured and then nutrient deficiencies from poor villi absorption may affect other areas.

Dr. Vojdani's abstract in his editorial, *The Immunology of Gluten Sensitivity Beyond the Intestinal Tract,*[2] supports that the villi are not always injured. To quote his editorial abstract, "Evidence has been accumulated in literature demonstrating that gluten sensitivity or celiac disease can exist even in the absence of enteropathy" [gut/villi damage], "but affecting many organs."

3. I do not have damaged villi, and my tTG/gliadin blood tests were negative, but I feel so much better gluten-free. Why?

The tests were likely false negative. That is very, very common. As gluten is processed in the digestive system, it can break into more pieces than we have tests developed to check them, and the immune system makes an antibody for each separate piece. The test probably missed your antibodies because they were different from the one or two the test checked. Many patients have proven this because they received negative results to the standard test, and then found a number of positive gluten related antibodies when they ran a more comprehensive panel. Cyrex Laboratories (www.cyrexlabs.com), runs 28 gluten related antibodies in 3 areas of the immune system.

Your properly performed* biopsy was probably negative because likely your gut villi are fine.** Your gluten injury may have targeted other organs or tissues, not the villi.*** For instance, if gluten has damaged the heart, snipping intestinal villi will not help find the heart injury. Damage in either place is not ok.

More detail on False Negative Tests

The blood and saliva tests - Standard celiac blood and saliva tests only check tissue transglutaminase IgA, (tTG – IgA) as an initial screener. Most doctors give up if that test is negative. However, the literature shows that standard tTG-IgA tests are only elevated when the villi are completely destroyed. Dr. Vojdani finds that form of tTG is often not found in other gluten injured tissues. Therefore the standard tTG-IgA test is a poor screener for most patients, and returns many false negatives. More forms of tTG have now been discovered.

Doctors might also order a standard test for deamidated gliadin IgA and possibly IgG. Unfortunately these tests also have a miserable failure rate for picking up gluten syndrome because many patients have other antibodies beside those two. The gluten "string" can break up into many other pieces, plus other "departments" of the immune system may react.

---

* If there IS villi damage it is possible for the biopsy to miss it if the damage is patchy or further down the duodenum than the villi samples are taken. Most good "gluten aware" gastroenterologists take a number of samples to try to avoid this possibility.

** It is possible that very early villi damage may slowly accumulate long before it shows up on a pathology report. An elevated IEL count, (intraepithelial lymphocyte, an early sign of inflammation) of the villi tissue indicates this process is underway. Ask the gastroenterologist to specify that the pathologist do an IEL count.

*** I am careful to specify the "villi of the gut". There may actually be gut damage, but not to the villi. There are many types of tissue in the gut, such as nerves, for instance. The villi may be fine, but it is still possible that some other gut tissue may be damaged by molecular mimicry or other processes we do not yet understand. For instance, if the nerves that control peristalsis (wave like gut motion) are injured, it could result in chronic constipation.

Villi biopsy vs. positive antibodies for diagnosis? - To add insult to injury, even if antibodies are discovered, many celiac, villi focused specialists may still insist that the villi biopsy is the final word for "celiac" diagnosis and they trust a negative villi biopsy over positive antibodies. Many folks have positive antibodies and suspicious symptoms, including improvement upon gluten removal, but no apparent villi damage. Their doctors assure them they can eat gluten! This is a serious mistake for many patients.

More on the villi biopsy – By now it is easy to understand why even properly performed* villi biopsies are only useful for the relatively small subset of patients in which the villi of the duodenum are injured. Most patients do not happen to have damaged villi. Their damage is somewhere else in their body, other organs, nerves, etc., or some enzyme or functional item has been damaged by gluten antibodies. Scientists such as Dr. Aristo Vojdani[2,3,4,] do NOT recommend villi biopsies for gluten syndrome diagnosis (there may be other reasons to scope the gut, such as tumors, etc). Dr. Vojdani explains that if there are elevated antibodies to gluten, the immune system is screaming "I do NOT want this substance! I am manufacturing antibodies to tag and destroy it." That is sufficient reason to remove gluten from the diet. It is less expensive to check an extensive array of gluten antibodies than undergo an endoscopy anyway, and is much less invasive.

4. Why do many gluten syndrome patients not only react to wheat, barley, and rye but also sometimes to other foods, particularly oats, milk, corn, soy, egg, coffee, sesame, yeast, chocolate and others.

These foods have similar amino acid sequences to gluten. Now that we grasp molecular mimicry, this is logical. The immune system may misrecognize them for gluten, causing cross reactions which may keep the gluten antibodies running high even on a gluten-free diet. Happily, this does not always occur. Cyrex labs, Array # 4, checks a list of cross reactive and gluten substitute foods, IgA and IgG.

5. The diet seems excessively strict? Why does it take so little gluten to start a reaction?

We understand that miniscule amounts of bee venom, or peanut, can trigger emergency allergic reactions, and very tiny medication pills can cause major effects on our bodies. This is also true of gluten reactions, both immediate allergic IgE, and delayed IgA, IgG, and IgM.[3] "Crumbs matter." The difference is that delayed reactions are not as obvious as the allergic reaction. They may go unnoticed for hours to decades, but may be a long, slow, serious process. Their lack of drama robs these reactions of the compliance respect they deserve.

6. The silent syndrome - Why do many people react to gluten, proven by antibody tests, but they have few or no warning signs, or seemingly unrelated symptoms they do not recognize or connect with their diet for a long time. Then they crash with something serious, often or usually autoimmune?

Gluten is famous for slowly injuring nerves[5] by molecular mimicry, and in many cases the nerves are silenced by that injury.[6] The patient does not realize there is a problem until the tissue or organ that those nerves supply begins to fail. Furthermore some areas of the body do not contain many pain nerves, so we may not feel the damage. Slow silent damage is understood in other diseases. Heart damage, cancer, and aortal aneurysms are examples of conditions that develop silently and then suddenly flare. Prevention is best.

For years, celiac literature recognized 2 neurological conditions, peripheral neuropathy and gluten ataxia. In the wider perspective of gluten syndrome, nerve damage may be one of the most important areas of injury.[5]

Dr. Rodney Ford, MD, suggests in his book, *The Gluten Syndrome: Is Wheat Causing You Harm*,[6] that this is primarily a neurological disease[5], injuring and in some cases silencing nerves, compromising the health and function of the tissues they serve. This idea was reinforced by an astute observation made by Dr. Ford of one of his patients, an elementary school child who had not achieved bowel control. After she went gluten-free, the problem resolved. Dr. Ford realized the child now recognized the signal to visit the toilet, and accidents were avoided. The nerves in the lower bowel apparently "woke up" once the antibodies that injured them disappeared. This is an interesting and logical theory to explain "silent gluten injury", and it fits the community. Here is an abstract of Dr. Ford's published paper which discusses a possible widespread neurological focus:[5]

The Gluten Syndrome: A Neurological Disease, by R.P. Ford

Hypothesis: gluten causes symptoms, in both celiac disease and non-celiac gluten-sensitivity, by its adverse actions on the nervous system. Many celiac patients experience neurological symptoms, frequently associated with malfunction of the autonomic nervous system. These neurological symptoms can present in celiac patients who are well nourished. The crucial point, however, is that gluten-sensitivity can also be associated with neurological symptoms in patients who do not have any mucosal gut damage (that is, without celiac disease). Gluten can cause neurological harm through a combination of cross reacting antibodies, immune complex disease and direct toxicity. These nervous system affects include: dysregulation of the autonomic nervous system, cerebella ataxia, hypotonia, developmental delay, learning disorders, depression, migraine, and headache. If gluten is the putative harmful agent, then there is no requirement *to invoke gut damage and nutritional deficiency to explain the myriad of the symptoms experienced by sufferers of celiac disease and gluten-sensitivity. This is called "The Gluten Syndrome."*[5]

7. Why do so many of us react now, when for centuries most people appeared to be fine with wheat? After all, wheat is spoken of positively in the Bible and many other historical documents.

There may or may not be a conclusive answer to this question, but a few factors may play a role.

a. Many folks have higher toxin levels now, their nutritional status is worse, and digestive strength is weaker.

b. Today's wheat is different. Gluten grains have been subjected to a lot of changes, some genetically violent, according to Nina Federoff, a pro GMO scientist. She asserts in her book, *Mendel in the Kitchen*, that gluten grains were altered with nuclear radiation and chemical mutation by the 1950's – 1960's.[9] Recently a stash of old blood samples from that era, stored in a freezer by the military, were checked with standard celiac antibody tests. The incidence of positive antibodies was much less in those samples than is typically found in the general public today.

c. Wheat seed is sometimes treated with mercury. Might this play into unintended results?

8. Why do specialists and researchers insist that the gluten-free diet must be life long? Can't we heal the gut and go back to our beloved wheat bagels, croissants, and brownies?

Our scientists insist that gluten-free is a strict lifelong commitment. The memory B cells in the immune system never forget what the enemy "looks like." Today's gluten is a violently altered substance according to the very scientists who defend genetic alterations.[9] It is not worth playing with today's wheat. There is something strange and unpredictable about it. Even if a leaky gut has healed, future circumstances, toxins and stress might injure it and retrigger the syndrome, perhaps silently. Researchers say gluten creates leaky gut for a few hours in everyone.

Particularly since gluten appears to be the "bad boy" that predisposes to other intolerances, I prefer to walk away for life and concentrate on other nutrient dense foods. However, it is wise to consume whole foods rich in B vitamins, (ex., liver or bee pollen), and silica (the herb horsetail, "equisetum hymale" species, is a source,) as gluten grains contain those nutrients.

9. Traditional peoples soaked and/or sprouted their wheat berries and then made sourdough bread with it. Does that process alter the gluten sufficiently for gluten syndrome patients to safely consume it today, particularly organic spelt and einkorn?

No. These methods make bread more digestible, but the fermented products and ancient gluten grains still contain gluten and can trigger reactions in research trials. Even if there are no visible reactions, silent injury cannot be ruled out without long term research. Additionally, the preparation process before fermentation is completed is definitely gluten based, so significant cross contamination issues come into play in the kitchen. Healing the gut is a challenge. Other areas crop up that need cleanup too, such as the parasite issues. Gluten specialists advise us to go gluten-free, stay that way, and move on.

10. What are gluten withdrawals?

Rarely, a few days to a few weeks after going gluten-free, or after being glutened, a patient may experience a few hours to a couple of weeks of a parade of varied and unusual symptoms including dizziness, black pit depression, crying, physical or emotional exhaustion, even as in difficulty getting up to use the rest room, and other odd symptoms. In severe cases they may experience an inability to socialize, make eye contact, make decisions or hold a normal conversation. Children or teens may act out in extreme ways during this situation. Often patients cannot bring themselves to discuss this experience afterward. However, a patient reluctantly described her experience 2 years later as "encountering an empty white board with nothing on it." The rest of life around her seemed to be "across the Grand Canyon." This appears to be a temporary crisis that resolves anywhere from a few hours to a couple of weeks or so according to folks who contact me about them. It is assumed that this phenomenon is due to particular pieces of gluten strings called gluteomorphins with amino acid sequences which resemble opiate drugs. When these gluten pieces disappear from the blood stream the patient may experience "withdrawal," very much like a drug withdrawal. Another theory for why this may happen involves changes in blood flow to the brain that may create a temporary neurological crisis. Autistic children may suffer these withdrawals and may take longer to stabilize, but they usually make nice gains after the crisis passes. Happily, once withdrawal is over the patients are usually much better, and they are VERY vigilant with their diet.

In the rare event that this type of reaction might occur, family and friends of the person do well to understand that the person may (or may not) be able to prepare their own food, for instance, but not be able to verbally communicate much, make eye contact, hold a conversation, answer questions, and may be uncharacteristically snappy particularly if others attempt to communicate with them. Family members may wish to be quietly and unobtrusively nearby until the person passes through this stage. If the patient has children to care for, help may be needed, and also solitude, rest, and simply prepared "real" nourishing food that is gluten-free and easy to digest such as bone broth. Probiotics or old fashioned fermented raw veggies or sauerkraut from the refrigerated section of the health food store may be helpful. Family and friends should not take the person's temporarily withdrawn personality personally.

For details on these unstudied rare reactions, or if you need support during a crisis, see...

www.theglutensyndrome.net
/Adverse_reactions.htm

...or contact me at *jka8168@sbcglobal.net*. I collect testimonials, so feel free to contribute if you have experienced this type of reaction. We hope these reactions will eventually be studied.

11. Should I replace all the gluten foods I routinely eat with gluten-free substitutes?

No, most gluten-free substitutes are still mainly expensive processed food (aka junk food). For the first few weeks it is normal for newbies to look for substitutes to replace their "old gluten friends" and it helps them make the very real emotional transition, but there are better, more nutrient dense food choices. The substitutes are still high carb, (potato, corn, tapioca, rice), and most contain sugar, processed gums, GMOs and ingredients that do not help us get well. We may actually eat more of them BECAUSE they are gluten-free. Additionally they are expensive and not always easy to prepare at home.

Many experienced gluten-free folks gradually wean themselves off a constant diet of the rather junky substitutes in favor of other real unprocessed foods, such as meat, eggs, veggies, fruits, nuts, fermented vegetables, bone broths, and so on according to their digestive abilities. The gluten-free substitutes become the occasional treat, such as pizza crust. Gluten free families are wise to find the healthiest GF versions of just the items they miss the most, and skip the rest. For example, at our house, I had served gluten spaghetti for years, but when I dropped that habit, no one noticed. However, my husband wanted his breakfast toast, so we found a gluten free brand he liked and continued that tradition. Each family works out these adjustments for themselves.

Lettuce wraps are crunchy and yummy instead of sandwiches. Hamburger patties without the bun but with onion for the onion lovers, are still tasty and they feel better afterward. In our family we now use watermelon or a watermelon basket for birthday "cake", complete with candles and a bow in the summer. Fruit pizza with a nut flour or nut/date crust, avocado chocolate or other dairy-free or fruit based pudding for the sauce, and an artistic fruit/coconut topping works for parties. There are lots of ways to celebrate happy healthy gluten-free, junk free birthdays.

12. Warning! Formal gluten challenges for testing purposes – risky!

A word of caution. Antibody tests must be performed while the patient is still consuming gluten or shortly after going gluten-free. Sometimes doctors advise patients who are already gluten-free to go back on gluten, 4 slices a day for 4-6 weeks to restart the antibodies for testing purposes. Patient experience has shown that once the system is fairly clean of gluten, going back to perform a formal gluten challenge for testing purposes may be risky. The secondary reactions for some individuals can be significant, particularly neurologically. Challenges have been known anecdotally to trigger psychological black pit crashes, fibromyalgia, and other organ injury. Autistic children may have a hard time when they just go gluten-free initially or consume gluten accidentally. A planned challenge for them may be very unwise.

13. Intermittent infractions (aka cheats) are seriously unwise and may increase injury.[10]

This is not a fad, or cheater's diet. The gluten-free diet is a medical diet to treat or control serious autoimmune, inflammatory, and often neurological diseases. There is no room for casual infractions. Research suggests that repeated, intermittent cheats, even every few weeks, over time may actually influence mortality rates.[10] This is not a reason to avoid the diet, but to take it seriously. Once the strictness of the diet is embraced the patient or family adjusts and discovers that it is doable. Gluten-free food bars, nuts, a packet of gluten-free soy sauce, and GlutenEase™ or other brands of DPP IV enzymes are good to keep in the glove compartment or backpack to handle emergencies. NOTE: DPP IV (pronounced DPP 4) digestive enzymes help break down gluten but they do not stop a reaction and are NOT a reason to cheat. However if there is a possibility of exposure it makes sense to take them for whatever help they might afford. In the case of a confirmed gluten exposure, take the enzymes, keep the bowels moving, stay calm and deal with it. Worry and drama makes everything worse.

14. Discrepancy – The celiac focus uses villi biopsy for diagnosis. The wider perspective relies on positive antibodies or improvement on the diet. Question? How did villi damage become the gold standard?

The first place the medical profession found conclusive damage by gluten in the 1950's – 60's was to the villi of the gut. After the endoscopy tool was developed, villi damage and subsequent healing upon gluten removal could be observed. They concluded that gluten damaged the villi. This was true. The particular subset of cases they scoped had villi damage, but their conclusion that the villi were the only target of damage for everyone with gluten syndrome was too narrow according to research today.[2] In most patients the target damage was NOT the villi, but the bones, joint lining, heart, thyroid, pancreas, liver, brain, almost any organ, blood vessel walls, nerves almost anywhere in the body, and so on. The patients might even have injury to multiple tissues, BUT NOT NECESSARILY ALWAYS TO THE VILLI. When the villi biopsy was declared the gold standard for gluten syndrome diagnosis, it cut out most of the patients who were reacting to gluten. Snipping villi does no good if the damage is in the thyroid or brain. For the next 60 years very few patients were prescribed a gluten-free diet because most of them did not have damaged villi, (or the doctor never thought to look at all). Their gluten-induced injury was somewhere else in their body, so they were never diagnosed.

15. Discrepancy - The villi damaged celiac disease story teaches that celiac disease is autoimmune and much worse than "non autoimmune" non-celiac gluten syndrome (NCGS). The wider gluten syndrome perspective teaches that both are autoimmune and serious.[2,3,4]

A significant disagreement exists over the autoimmunity of non-celiac gluten syndrome (NCGS). In the beginning, antibodies to gluten could not be found in NCGS patients, therefore it was assumed that NCGS was not autoimmune. However, Dr. Vojdani insists NCGS IS autoimmune. The NCGS patients have plenty of antibodies, just different ones than the standard tests check. His tests, which check and find more antibodies,**** the illnesses these patients develop, and recoveries or improvements on the diet all prove his point. He also asserts that NCGS can indicate a gut wall in worse shape than the celiac villi damaged subset.[2,3,4] The damage is simply somewhere else, not to the villi. See the link below for diagrams of these reactions, and compare the condition of the gut wall in the celiac diagram with the gut wall in the gluten intolerance diagram.

**www.TheGlutenSyndrome.net/
VojdaniDiagrams.htm**

**** Dr. Vojdani's research found several more forms of both tTG, and gliadin antibodies, (alpha, gamma, omega), gluteomorphins, glutinins, and others. His saliva and blood panels check 12 separate antibodies in 3 different immune departments and 2 mediums, totaling 28 gluten related antibodies. The two panels, run together, nearly always find antibodies if they are present in the patient, translating to far fewer false negatives.

> This disagreement fosters confusion among NCGS patients that villi damaged celiac disease is the "big bad boy" to avoid (true, it is bad), but that NCGS is less severe, not autoimmune. This translates in real life to gluten birthday cake for the Friday night party and then back on the gluten-free bandwagon Saturday morning, "so I won't eventually develop villi damaged celiac disease." Yes, that could happen also, particularly if the person seesaws on the diet, but the notion that NCGS is not as bad as villi damaged celiac disease is a misunderstanding according to Dr. Vojdani.[2,3,4] To repeatedly seesaw off and on gluten indiscriminately is unwise according to medical literature,[10] and community experience. Gluten exposure needs to be an accident, which happens occasionally even to the most vigilant. Casual cheats are a more risky mindset, usually meaning "more frequent."[10] Take it seriously, nerves, blood vessels, and organs are precious.

Gluten Tests with a Good Track Record

Some families cannot afford to test or good tests are not available. In the case of autism this protocol and most others require gluten-free and that is the end of the matter. Kerri takes this position partly due to cost and the miserable record of false negatives that standard tests return. The protocol does not work without the diet. Period.

Some adults are also willing to go gluten-free without a test. Their bodies tell them what they need, they listen, and are happy to find answers. Social pressure does not sway their decision or ability to comply.

In many other situations, a positive test is very meaningful. It gives patients confirmation that gluten-free is right for them, silences critics, and helps them comply with the diet. In the case of children it provides proof to the other parent, grandparents, therapists, doctor, etc. It may also come in handy if the child turns into an invincible teen who doubts he ever needed the diet in the face of pizza and beer, ditto for his future spouse and in-laws. The catch is that the test must be adequate. A false negative misvalidates the skeptics. These are decisions every individual or family must evaluate for themselves. Thankfully there are test panels with good track records now.

Absolutely, regardless of any test result, if the body is able to communicate a reaction to gluten or its removal, then that is the final answer. Be grateful and go gluten-free.

This testing section is for those who wish to test for their own confirmation or social support.

NOTE: If a patient has been ill for a very long time it is possible for the immune system to be so worn out that few antibodies are manufactured. Lower antibody counts might show a false negative but not prevent injury to innocent tissues. Also, due to the wide variety of antibodies a patient may happen to make, it is possible to run any antibody panel and miss those particular antibodies. The more antibodies that are checked the less likely this may occur, but should be considered in the event of a negative test that the patient or practitioner questions. In this case a gene test may be helpful.

Enterolab - Stool and 1 part gene tests

For 10 years Enterolab's mail in home collection research stool test stood in the gap for thousands of patients who received false negative standard tTG and gliadin blood tests. It has saved many lives, and gave our family the social confirmation we needed. This unpublished research test checks stool for tTG-IgA and gliadin-IgA only. The use of stool as the testing medium appears to pick up those antibodies most of the time, much much more often than standard tTG or gliadin IgA blood tests. However, since only two antibodies are checked, it may miss in some cases.

Dr. Fine's lab also offers a one part gene test which he believes is adequate for a reasonable price. Dr. Fine finds if a gene is present, nearly always so are the antibodies, and two genes are worse. Villi damaged celiac specialists recognize only HLA DQ 2 and 8 as gluten related, but Dr. Fine includes 1 and 3 and their subsets 5, 6, 7 and 9. In fact according to Enterolab, HLA DQ 4 is the only DQ gene that does NOT correlate with gluten syndrome. According to him, a patient needs 2 copies of the HLA DQ 4 gene to miss the predisposition. This translates to 81% of the Caucasian population with a predisposition to trigger gluten syndrome at some point in their lifetime, including before they are born.

Cyrex Labs - Better blood and saliva tests

Cyrex Labs (www.CyrexLabs.com) opened their doors in 2010 in Phoenix, Arizona. They run much more complete antibody panels designed by Dr. Aristo Vojdani, PhD., their scientific advisor, an immunologist, and researcher. Dr. Vojdani found a wider variety of gliadin antibodies, (alpha, gamma, omega) and variations of tTG in other tissues, plus gluteomorphins, and several others. He also checks an IgM antibody due to possible malfunctions in that system. Cyrex blood and saliva panels, Array #1 and #3, combined, test for 28 gluten related antibodies between 3 immune departments, 2 mediums, plus IgA insufficiency. They rarely miss a diagnosis because they look for so many antibodies. (Note: Vojdani believes stress, toxins and infections, i.e. environment, can trigger a gluten reaction without the genes.)

Gluten Free Society - Two Part Gene Test

A third approach espoused by Dr. Peter Osborne, *The Gluten Free Society*, Sugarland, Texas, (www.glutenfreesociety.org), is to run only the gene test, since any antibody panel may theoretically miss the particular ones the patient may have. He uses a 2 part test, the most complete method, and looks for both celiac and gluten sensitivity genes. A positive gene test does not prove a current immune response as do antibodies, but predisposition to it. Presence of 2 genes indicates a more severe case. Gene tests have an advantage in that they can be run anytime, gluten consumption is unnecessary, and depending on the results, useful information can be gleaned for immediate and extended family members.

Elimination Diet

The elimination diet is inexpensive. Often/usually it demonstrates improvement upon removal of gluten, or worsening upon reintroduction. Many "heads up" practitioners accept this as reason enough to go gluten-free.

There are occasional complications or interpretation issues for the elimination diet as follows:

a. Reintroduction of gluten (gluten challenge) can trigger stronger, sometimes risky reactions.[10] Stop a challenge upon negative symptoms, including depression and emotional instability, or best, don't challenge. An accidental infraction may come up that provides insight.

b. Occasionally it takes several months to see the difference, or a silent reaction may mask symptoms.[6]

c. If the patient later decides to test and is already "clean" on the gluten-free diet, blood tests will not work unless gluten is reintroduced for many weeks.[10] No! Stool/gene tests are safer.

d. There is no lab confirmation to silence naysayers.

Other Related Tests

Cyrex Labs Array # 2 - Intestinal Permeability Panel focuses on specific causes of leaky gut. This helps strategize treatment, and is an improvement over the old lactulose/mannitol test.

Cyrex Labs Cross Reactive Foods and Gluten Substitutes, Array # 4. - This specific list checks foods that commonly cross react with gluten and also foods commonly used to replace gluten. It helps customize an anti-inflammatory diet.

Cyrex Labs Predictive Antibody panels determine if or which tissues are currently under antibody attack. This predicts autoimmunity years ahead of time and gives the patient advance notice in order to address trouble spots.

Enterolab (www.Enterolab.com) offers several stool based food antibody panels and gut related stool tests.

What are the lab instructions?
Do I need to consume gluten for testing?

Antibody tests, (blood, saliva, stool), prove a reaction and require recent gluten consumption. Ideally, test first, then go gluten-free. If the patient is off gluten, call the lab for advice on the time window before the test will not work.

Cyrex Labs (www.CyrexLabs.com) tests require prescriptions and a saliva specimen and/or blood draw. If a doctor is needed to write the script, check www.thedr.com for a partial list of practitioners who are familiar with Cyrex Labs. Results are sent to the prescribing doctor.

Enterolab stool specimens are ordered online, kits are sent, home collected and mailed. No prescription is required. Results arrive on email. Enterolab's test works for several months after going gluten-free.

Gene tests do not require gluten consumption or a script and can be run at any time. Genes prove a predisposition to gluten reaction.

Enterolab's gene test is a relatively inexpensive "one part" test. It is a mail in cheek swab and reports the patient's actual genes.

Gluten Free Society (www.Glutenfreesociety.org) gene test is a "two part" (more complete) mail in cheek swab and reports yes or no for both celiac and gluten sensitivity genes.

For information on testing (I have no financial interests) see:

www.TheGlutenSyndrome.net

The latest version of this article and a complete list of references is available at:

www.TheGlutenSyndrome.net
/Molecular_Mimicry.pdf

Author's note: While the information Olive has provided can be extremely helpful for adults or children who are not on the spectrum or who are suffering from other ailments, a gluten-free diet is non-negotiable for healing autism. We need to remove foods that cause inflammation and mucous, and which can produce gluteomorphin in the body. I have seen many families waste time and money doing testing only to get confused by false negatives—time and money they could have spent on recovering their child.

Online Gluten Summit – A Chance to Support CDAutism

Recently, in November 2013, Dr. Thomas O'Bryan, an internationally recognized gluten educator, put together an amazing online Gluten Summit. Twenty nine iconic experts on Celiac Disease and its wider perspective, Gluten Sensitivity, also called Gluten Syndrome, covered every possible aspect before 115,000 appreciative viewers. This set of interviews plus extra materials is available for purchase at only $3.30 per speaker and is a priceless addition to all our libraries. It also is a great gift and a good way to inform friends and family. Often our closest loved ones accept new information better from a respected professional rather than from us.

Arrangements are in place to contribute half of your purchase price of The Gluten Summit interviews to the CDAutism project. Go to the link below to order and also credit CDAutism with your contribution.

gg110.infusionsoft.com/go/tgso/Kerri

Appendix 6

Measuring the Strength of Your CD, CDS & CDH

by Charlotte Lackney

CD, CDS, and CDH work because all of them contain chlorine dioxide (ClO_2). ClO_2 is made when a 22.4% solution of sodium chlorite ($NaClO_2$) is activated with an acid, usually 4% hydrochloric acid (HCl) for CD and CDH; and 10% HCl for CDS.

Sometimes it might be useful to measure the amount of ClO_2 in those solutions, although that has not been done in the past when using CD. And, it may only be necessary to know the amount of chlorine dioxide if you are not getting the expected results. There could be a problem with the ingredients or process and knowing the amount of ClO_2 in the solution could be helpful in determining what might be wrong.

The amount of ClO_2 in a solution is measured in *parts per million* (ppm). It is always necessary to specify the volume of solution when talking about ClO_2 ppm because the ppm will vary depending on the dilution of the solution.

If using CDH or CDS with the protocol in this book, it is assumed that you are preparing them at 3000ppm. If using CDS, please keep in mind that the relative strength of one milliliter of CDS is approximately 60% of one drop of CD.

However, if the ingredients are of good quality and the CD or CDH protocol is closely followed, there is usually no need to know or measure the ClO_2 concentration.

Both CD and CDH may continue to activate inside the body from whatever sodium chlorite has not been activated outside the body. That is not the case with CDS, as it has no sodium chlorite to continue activating inside the body. Some believe that this is why many people can tolerate a higher dose of CDS than an equivalent dose of CD.

Measuring the concentration of ClO_2 can easily be done at home. You will need to purchase ClO2 test strips made by *LaMotte* or others. I have only used LaMotte Insta-Test® High-Range Test Strips, 0 to 500ppm (code #3002). The container's label has a chart showing seven colors on it wih each color indicating a ClO_2 concentration in ppm. Note: LaMotte recently changed the appearance of their labels, as shown on the right. Both are the same product. For more information, see their website:

www.lamotte.com/
en/water-wastewater/test-strips/3002.html

Directions for how to use the LaMotte test strips are printed on the side of the container and need to be followed in order to get an accurate ClO_2 concentration reading.

The numbers below each of the seven colors represent 0, 10, 25, 50, 100, 250 and 500ppm. Many people who use the test strips think the 50ppm color is the best one to use for color matching, because they can see the difference in color below and above that color more easily than the others.

If the ClO_2 solution you want to measure is suspected to be 3000ppm, how do you measure that if the test strips highest ClO_2 concentration color is 500ppm? Recall that I had said that it is always necessary to specify the volume of solution when talking about ClO_2 ppm, because the ppm will vary depending on the dilution of the solution.

So if we dilute the suspected 3000ppm ClO_2 stock solution with more water, the same amount of ClO_2 will now be evenly spread out in the larger amount of solution if stirred. There will be less ClO_2 in any one spot than before dilution, because the ClO_2 is now dispersed into a larger volume of water. Visualize what happens when a drop of red food coloring falls into a glass of water. It disperses throughout the water if stirred, and that dark red color is now much lighter because it is diluted in the water.

In order to read 3000ppm with a test strip, and have it match the 50ppm color on the container, we will need to dilute a small sample of the stock solution.

Stock solution (1ml)	÷	Desired dilution in ppm	=	Volume of water for dilution.
3000ppm		50ppm		60ml

Insta-TEST
50 Test Strips
High Range
Chlorine Dioxide
For testing sanitized water
Just dip & read!
LaMotte
InstaTest ANALYTIC
High Range
Chlorine Dioxide
50 Test Strips
Code 3002
LaMotte

The chlorine dioxide test strips (code #3002) made by LaMotte are available on the market with two different labels, although both are the same product. Both labels show the same measuring color chart on the outside of the container.

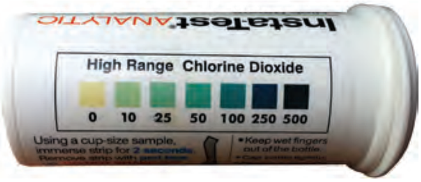
InstaTest ANALYTIC
High Range Chlorine Dioxide
0
10
25
50
100
250
500
Using a cup-size sample, immerse strip for 2 seconds.
Keep wet fingers out of the bottle

So, if we take a 1ml sample of the stock solution and add it to 60ml of water, then the stock solution is diluted 60 times its original volume. Technically, one should add the 1ml of stock solution to 59ml of water, for a total of 60ml, but using 60ml is easier to measure and you would not see the difference in ClO_2 readings if 59ml of water was used.

To measure, pour 60ml of distilled water into a small glass and then add 1ml of stock solution. Stir well to evenly distribute the ClO_2 sample in the water. Then take one of the test strips—taking care not to touch the pad of chemicals on the end of the strip—and dip it in the solution for two seconds. Keep the strip in one place in the solution and do not move it around during the two-second period.

Without flicking any of the solution off the strip, remove the strip from the test solution with the pad facing up and wait ten seconds. Now compare the color of the test strip pad to the color chart on the side of the test strips container (as shown above).

If the stock solution is 3000ppm, then the color on the test strip will match the 50ppm color on the container. Multiply 50 times 60 (the amount of dilution) and you get 3000ppm.

If the color on the test strip does not match the 50ppm color on the container, try to estimate the ppm number between container colors and multiply that number by 60 to get the ClO_2 concentration.

Tip: You can cut the strips lengthwise to get 100 strips instead of 50; just be sure not to touch the strip's pad with your fingers or your readings may not be accurate. Also it is important to keep the container tightly capped to keep moisture out which can affect the readings.

Dana one year after starting Kerri's protocol.

Healing
the symptoms known as
Kerri Rivera

Appendix 7

Diluting HCl Concentration

The protocols in this book use two different concentrations of hydrochloric acid (HCl): 10% & 4%. Some suppliers sell both concentrations, while others only one. You may also find a source with a different concentration and wish to dilute it to 4 or 10%. The math to determine how much to dilute your stock solution is rather simple. You should have no problem if you know how to multiply, divide, add and subtract. Note: The same calculation works for citric acid if you have 50% and need to go down to 35%.

What you will need is:

- Bottle of concentrated HCl (10% or higher)
- Receiving bottle to store the resulting diluted HCl
- Graduated cylinder or other measuring cup

On the next page is a graphical representation of the dilution equation. It may look intimidating at first glance, but it is actually quite simple and is made up of three simple mini-calculations. Each box is labeled with a letter which refers to the graphic. Just fill in the numbers with a pencil and follow the directions.

We'll go over one example here: Let's say you have 10% HCl and want to make 1 liter of 4% HCl. Here's what you would have to calculate:

Desired Volume	×	Desired HCl %	=	Temporary Number
1000ml		4%		4000

Temporary Number	÷	Stock HCl %	=	Amount of 10% HCl
4000		10%		400ml

Desired Volume	–	Amount of 10% HCl	=	Amount of water
1000ml		400ml		600ml

So, in this example, you would need 400ml of 10% HCl and 600ml of water. to make 1 liter of 4% HCl. If you add them together, you should have 1000ml (1 liter).

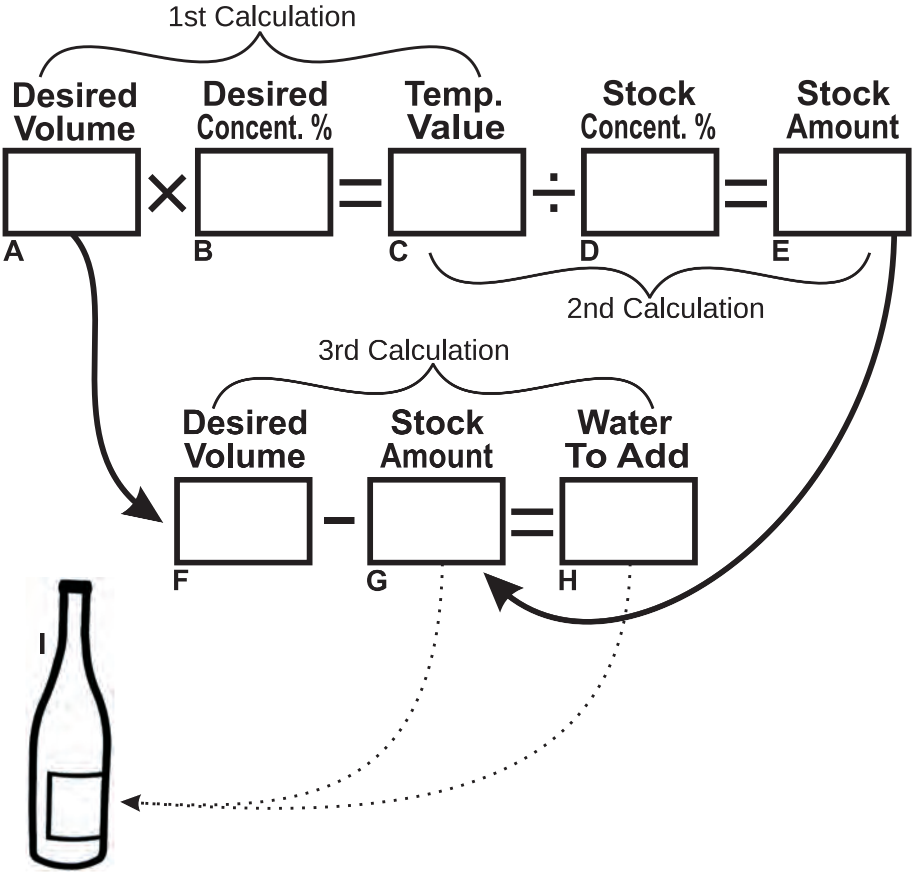
1st Calculation
Desired Volume
Desired Concent. %
Temp. Value
Stock Concent. %
Stock Amount
×
=
÷
=
A
B
C
D
E
2nd Calculation
3rd Calculation
Desired Volume
Stock Amount
Water To Add
–
=
F
G
H
I

A	Determine the volume of the receiving bottle in milliliters. For example, if you have a 1 liter bottle, you would enter the number “1,000” for 1,000ml.
B	This is where you enter the % of the final solution. If you are making 4%, then enter “4” in this box.
C	Multiple **A** and **B** and enter the result into box **C**. This is a temporary value you will need for the next calculation.
D	This is where you enter the percentage of your concentrated stock solution–the HCl you want to dilute. If you have 10%, then enter a “10” in this box.
E	Now divide the temporary value in **C** by **D** and enter the result in **E**. This is the amount of concentrated stock HCl (in milliliters) you will need.
F	Copy what you entered in **A** to **F** for the next calculation.
G	Copy the result from **E** to **G**.
H	Do the subtraction of **F** – **G** and enter he value in **H**. This is the number of milliliters of distilled or filtered water you will need.
I	Measure the resulting volume of water calculated in **H** and pour it into the receiving bottle **I**. Finally, measure out the value in **G** and pour it into the receiving bottle **I**.

Appendix 8

Other Uses of Chlorine Dioxide

Bladder Infection Protocol (for Adults)

The protocol for bladder infection is three oral drops of CD eight times a day with a douche of 10 drops CD three times the first day, and then once every day thereafter, for 30 days (for females).

Ear, Eye, Nose Protocol

Make a mixture that contains one ounce of water for every one drop of CD. Use one drop of that mixture in the affected orifice every hour until symptoms disappear. If the issue is acute you can use one drop every 15 minutes. This method can be used in the ears for infections, in the nose for congestion/infection, and the eye for infection or conjunctivitis.

General Fever Protocol

If the person has a fever, or if you want clarification of any kind, email Kerri, but in the meantime, this will hopefully get you started:

1. Give baths interspersed with enemas, approximately eight hours apart (e.g. bath in the evening and enema in the morning or vice versa).
2. Give oral CD doses (using the person's usual number of drops) 12 to 16 times or more a day (essentially every hour); unless they stop eating, in which case, stop oral dosing. Enemas and baths can continue.
3. Have them drink lots of water.
4. Cool clothes on forehead and neck.

Please see Chapter 14, page 335 for more information on fever therapy.

Teeth

Ten activated drops of CD per one ounce of water. Put this in a spray bottle and spray on the toothbrush.

Throat Spray

Ten activated drops of CD (chlorine dioxide) in one ounce of water. Spray on throat once/hour. Recommended in the case of a sore throat.

Vaginal Douching for Yeast Infections:

Put six activated drops of CD in 60 ml of water. Hold this in for two minutes every day for a few weeks or between periods. Six drops MAX in the vagina.

CD Skin Spray

Use ten drops of activated CD per one ounce of water in a spray bottle.

CD spray is great for a variety of ailments.

- Skin
- Dandruff
- Scrapes
- Cuts
- Rashes
- Toe Fungus

Burns

Any burn should be sprayed with full strength sodium chlorite solution (no acid added) directly from a small spray bottle. Do not add citric acid or HCl at all! If you don't have a spray bottle available drip sodium chlorite solution directly onto the burn, making sure the area is soaked with it.

Wait up to five minutes, but no longer before rinsing off with room temperature water. If you fail to rinse off the sodium chlorite solution, the burn will continue to hurt and get worse.

This technique will cause the burn to heal in 1/4 the time normally required. The pain should stop almost immediately or reduce to almost zero within several minutes.

Sunburns should be treated the same way. Spray the red area, wait 1 to 5 minutes, and rinse off. If the area is still sore, in about an hour spray the area again and wait 5 minutes before rinsing off. Remember, DO NOT allow the sodium chlorite to remain in place. It must be rinsed off! The pain should be gone in a couple of minutes. Generally two doses will overcome most sunburns, but on rare occasions if the discomfort is not all gone you can use a third dose. Be sure to rinse it off.

Note that sodium chlorite is alkaline and burns are acidic. Therefore the sodium chlorite neutralizes the acidity that resides in the burned areas. This is part of the reason why burns heal rapidly after sodium chlorite applications.

Appendix 9

Blank Kalcker Parasite Protocol Chart

The next page contains a blank Kalcker Parasite Protocol Daily Chart identical to those found on pages 199 through 217. You are encouraged to copy this chart for use in your parasite protocol planning. Setting the copy machine to enlarge at 125% should fill a standard letter-sized page.

Time	Meal Time	CD / CDS / CDH	CD / CDS / CDH Bath	CD / CDS / CDH Enema	Ocean Water	Diatomaceous Earth	Lepidium Latifolium Extract (Rompepiedras / RP)	Pyrantel Pamoate (Combantrin®)	Mebendazole	Castor Oil	Neem	Probiotic
6:00 AM												
6:30 AM												
7:00 AM												
7:30 AM												
8:00 AM												
8:30 AM												
9:00 AM												
9:30 AM												
10:00 AM												
10:30 AM												
11:00 AM												
11:30 AM												
12:00 PM												
12:30 PM												
1:00 PM												
1:30 PM												
2:00 PM												
2:30 PM												
3:00 PM												
3:30 PM												
4:00 PM												
4:30 PM												
5:00 PM												
5:30 PM												
6:00 PM												
6:30 PM												
7:00 PM												
7:30 PM												
8:00 PM												
8:30 PM												
9:00 PM												
9:30 PM												

Appendix 10

Lunar Phase Calendar for Kalcker Parasite Protocol

As covered in Chapter 8, the Kalcker Parasite Protocol is administered over a period of 19 days and scheduled according to the lunar cycle. We have included an easy to follow calendar starting on the next page for the years 2014 through 2016.

Each chart page shows 6 months with each month broken up into 3 columns: Column 1 is the day of the month; column 2 indicates the lunar phase; and column 3 shows the corresponding Parasite Protocol day.

The full moon is identified with an "F" in the second column and the new moon is identified by an "N" in each month section, which usually occurs towards the end of each protocol cycle.

We have used New York as our center point since it is roughly the center of the majority of those using this book in the US and Europe. The moon phase may occur 1 day before or after the dates shown if you are in another part of the world, but the difference is not great, so feel free to use the dates shown. However, you are welcome to be precise for your particular location. The Farmer's Almanac is a great source and provides an online calendar that can be adjusted to your location:

www.almanac.com/moon

You may also wish to install a moon phase app into your smart phone. Many free options exist.

2014 - Lunar Phases for New York, NY																	
Jan			Feb			Mar			Apr			May			Jun		
1	N	18	1			1	N	18	1			1	N		1		
2			2			2			2			2			2		
3			3			3			3			3			3		
4			4			4			4			4			4		
5			5			5			5			5			5		
6			6			6			6			6			6		
7			7			7			7			7			7		
8			8			8			8			8			8		
9			9			9			9			9			9		
10			10			10			10			10			10		0
11			11		0	11			11			11			11		1
12		0	12		1	12			12		0	12			12		2
13		1	13		2	13		0	13		1	13		0	13	F	3
14		2	14	F	3	14		1	14		2	14		1	14		4
15	F	3	15		4	15		2	15	F	3	15		2	15		5
16		4	16		5	16	F	3	16		4	16	F	3	16		6
17		5	17		6	17		4	17		5	17		4	17		7
18		6	18		7	18		5	18		6	18		5	18		8
19		7	19		8	19		6	19		7	19		6	19		9
20		8	20		9	20		7	20		8	20		7	20		10
21		9	21		10	21		8	21		9	21		8	21		11
22		10	22		11	22		9	22		10	22		9	22		12
23		11	23		12	23		10	23		11	23		10	23		13
24		12	24		13	24		11	24		12	24		11	24		14
25		13	25		14	25		12	25		13	25		12	25		15
26		14	26		15	26		13	26		14	26		13	26		16
27		15	27		16	27		14	27		15	27		14	27	N	17
28		16	28		17	28		15	28		16	28		15	28		18
29		17				29		16	29	N	17	29		16	29		
30	N	18				30	N	17	30		18	30	N	17	30		
31						31		18				31		18			

2014 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1			1			1			1			1			1		
2			2			2			2			2			2		
3			3			3			3			3		0	3		0
4			4			4			4			4		1	4		1
5			5			5		0	5		0	5		2	5		2
6			6			6		1	6		1	6	F	3	6	F	3
7			7		0	7		2	7		2	7		4	7		4
8			8		1	8	F	3	8	F	3	8		5	8		5
9		0	9		2	9		4	9		4	9		6	9		6
10		1	10	F	3	10		5	10		5	10		7	10		7
11		2	11		4	11		6	11		6	11		8	11		8
12	F	3	12		5	12		7	12		7	12		9	12		9
13		4	13		6	13		8	13		8	13		10	13		10
14		5	14		7	14		9	14		9	14		11	14		11
15		6	15		8	15		10	15		10	15		12	15		12
16		7	16		9	16		11	16		11	16		13	16		13
17		8	17		10	17		12	17		12	17		14	17		14
18		9	18		11	18		13	18		13	18		15	18		15
19		10	19		12	19		14	19		14	19		16	19		16
20		11	20		13	20		15	20		15	20		17	20		17
21		12	21		14	21		16	21		16	21		18	21	N	18
22		13	22		15	22		17	22		17	22	N		22		
23		14	23		16	23		18	23	N	18	23			23		
24		15	24		17	24	N		24			24			24		
25		16	25	N	18	25			25			25			25		
26	N	17	26			26			26			26			26		
27		18	27			27			27			27			27		
28			28			28			28			28			28		
29			29			29			29			29			29		
30			30			30			30			30			30		
31			31						31						31		

2015 - Lunar Phases for New York, NY

Jan			Feb			Mar			Apr			May			Jun		
1		0	1		1	1			1		0	1		1	1		2
2		1	2		2	2		0	2		1	2		2	2	F	3
3		2	3	F	3	3		1	3		2	3	F	3	3		4
4	F	3	4		4	4		2	4	F	3	4		4	4		5
5		4	5		5	5	F	3	5		4	5		5	5		6
6		5	6		6	6		4	6		5	6		6	6		7
7		6	7		7	7		5	7		6	7		7	7		8
8		7	8		8	8		6	8		7	8		8	8		9
9		8	9		9	9		7	9		8	9		9	9		10
10		9	10		10	10		8	10		9	10		10	10		11
11		10	11		11	11		9	11		10	11		11	11		12
12		11	12		12	12		10	12		11	12		12	12		13
13		12	13		13	13		11	13		12	13		13	13		14
14		13	14		14	14		12	14		13	14		14	14		15
15		14	15		15	15		13	15		14	15		15	15		16
16		15	16		16	16		14	16		15	16		16	16	N	17
17		16	17		17	17		15	17		16	17		17	17		18
18		17	18	N	18	18		16	18	N	17	18	N	18	18		
19		18	19			19		17	19		18	19			19		
20	N		20			20	N	18	20			20			20		
21			21			21			21			21			21		
22			22			22			22			22			22		
23			23			23			23			23			23		
24			24			24			24			24			24		
25			25			25			25			25			25		
26			26			26			26			26			26		
27			27			27			27			27			27		
28			28			28			28			28			28		0
29						29			29			29			29		1
30						30			30		0	30		0	30		2
31		0				31						31		1			

2015 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1	F	3	1		4	1		6	1		7	1		8	1		9
2		4	2		5	2		7	2		8	2		9	2		10
3		5	3		6	3		8	3		9	3		10	3		11
4		6	4		7	4		9	4		10	4		11	4		12
5		7	5		8	5		10	5		11	5		12	5		13
6		8	6		9	6		11	6		12	6		13	6		14
7		9	7		10	7		12	7		13	7		14	7		15
8		10	8		11	8		13	8		14	8		15	8		16
9		11	9		12	9		14	9		15	9		16	9		17
10		12	10		13	10		15	10		16	10		17	10		18
11		13	11		14	11		16	11		17	11	**N**	18	11	**N**	
12		14	12		15	12		17	12	**N**	18	12			12		
13		15	13		16	13	**N**	18	13			13			13		
14		16	14	**N**	17	14			14			14			14		
15	**N**	17	15		18	15			15			15			15		
16		18	16			16			16			16			16		
17			17			17			17			17			17		
18			18			18			18			18			18		
19			19			19			19			19			19		
20			20			20			20			20			20		
21			21			21			21			21			21		
22			22			22			22			22		0	22		0
23			23			23			23			23		1	23		1
24			24			24		0	24		0	24		2	24		2
25			25			25		1	25		1	25	**F**	3	25	**F**	3
26			26		0	26		2	26		2	26		4	26		4
27			27		1	27	**F**	3	27	**F**	3	27		5	27		5
28		0	28		2	28		4	28		4	28		6	28		6
29		**1**	29	**F**	3	29		5	29		5	29		7	29		7
30		2	30		4	30		6	30		6	30		8	30		8
31	**F**	3	31		5				31		7				31		9

2016 - Lunar Phases for New York, NY

Jan			Feb			Mar			Apr			May			Jun		
1		10	1		12	1		11	1		12	1		12	1		14
2		11	2		13	2		12	2		13	2		13	2		15
3		12	3		14	3		13	3		14	3		14	3		16
4		13	4		15	4		14	4		15	4		15	4	N	17
5		14	5		16	5		15	5		16	5		16	5		18
6		15	6		17	6		16	6		17	6	N	17	6		
7		16	7		18	7		17	7	N	18	7		18	7		
8		17	8	N		8	N	18	8			8			8		
9	N	18	9			9			9			9			9		
10			10			10			10			10			10		
11			11			11			11			11			11		
12			12			12			12			12			12		
13			13			13			13			13			13		
14			14			14			14			14			14		
15			15			15			15			15			15		
16			16			16			16			16			16		
17			17			17			17			17			17		0
18			18			18			18			18		0	18		1
19			19		0	19			19		0	19		1	19		2
20		0	20		1	20		0	20		1	20		2	20	F	3
21		1	21		2	21		1	21		2	21	F	3	21		4
22		2	22	F	3	22		2	22	F	3	22		4	22		5
23	F	3	23		4	23	F	3	23		4	23		5	23		6
24		4	24		5	24		4	24		5	24		6	24		7
25		5	25		6	25		5	25		6	25		7	25		8
26		6	26		7	26		6	26		7	26		8	26		9
27		7	27		8	27		7	27		8	27		9	27		10
28		8	28		9	28		8	28		9	28		10	28		11
29		9	29		10	29		9	29		10	29		11	29		12
30		10				30		10	30		11	30		12	30		13
31		11				31		11				31		13			

2016 - Lunar Phases for New York, NY

Jul			Aug			Sep			Oct			Nov			Dec		
1		14	1		16	1	N	17	1		18	1		18	1		
2		15	2	N	17	2		18	2			2			2		
3		16	3		18	3			3			3			3		
4	N	17	4			4			4			4			4		
5		18	5			5			5			5			5		
6			6			6			6			6			6		
7			7			7			7			7			7		
8			8			8			8			8			8		
9			9			9			9			9			9		
10			10			10			10			10			10		0
11			11			11			11			11		0	11		1
12			12			12			12			12		1	12		2
13			13			13		0	13		0	13		2	13	F	3
14			14			14		1	14		1	14	F	3	14		4
15			15		0	15		2	15		2	15		4	15		5
16		0	16		1	16	F	3	16	F	3	16		5	16		6
17		1	17		2	17		4	17		4	17		6	17		7
18		2	18	F	3	18		5	18		5	18		7	18		8
19	F	3	19		4	19		6	19		6	19		8	19		9
20		4	20		5	20		7	20		7	20		9	20		10
21		5	21		6	21		8	21		8	21		10	21		11
22		6	22		7	22		9	22		9	22		11	22		12
23		7	23		8	23		10	23		10	23		12	23		13
24		8	24		9	24		11	24		11	24		13	24		14
25		9	25		10	25		12	25		12	25		14	25		15
26		10	26		11	26		13	26		13	26		15	26		16
27		11	27		12	27		14	27		14	27		16	27		17
28		12	28		13	28		15	28		15	28		17	28		18
29		13	29		14	29		16	29		16	29	N	18	29	N	
30		14	30		15	30	N	17	30	N	17	30			30		
31		15	31		16				31						31		

Three things cannot be long hidden:
the sun, the moon, and the truth.

~ ***Buddha***

Appendix 11

Health Benefits of Diatomaceous Earth

by Dee McCaffrey, CDC

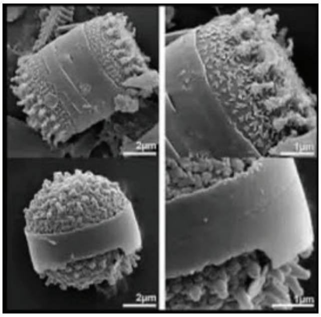

You've probably never heard of one of the best and most economical supplements to improve your health and cleanse your body. In fact, this supplement is perhaps one of the best kept secrets ever. This relatively unknown "food" supplement is called Diatomaceous Earth, also known as DE. It is a completely natural substance that is rich in naturally occurring silica, a mineral whose list of documented health benefits continues to grow as more research is being conducted.

Some of the most recent studies show that it can strengthen bones and joints, prevent osteoporosis and restore bone health if you already have osteoporosis, boost the immune system, ward off Alzheimer's, prevent premature aging and wrinkling of the skin, and strengthens the arterial walls to maintain good heart health. Another of the benefits of silica is that it helps to destroy bad fats in the body. Used as a daily treatment, diatomaceous earth can alleviate the potentially deadly risks of high cholesterol, high blood pressure, and obesity.

Additionally, because of its physical structure, diatomaceous earth is a highly effective anti-inflammatory and internal cleansing agent for the body. It can remove intestinal bacteria, parasites, e-coli, viruses, pesticides, heavy metals, and other toxins. It has also been known to assist with vertigo, headaches, tinnitus, and insomnia.

Sound too good to be true? Read on to learn more about this overlooked multi-purpose supplement and how it can improve many aspects of your health at a fraction of the cost of other supplements.

What is Diatomaceous Earth?

Diatomaceous earth is the fossilized shells of microscopic water-dwelling organisms known as diatoms. Diatoms are one-celled plants (algae or phytoplankton) that are the primary food source for marine life. These plants have been part of the earth's ecology since prehistoric times, yet their species still exists today in both salty and fresh waters all over the world.

Diatoms use soluble silica (sand) from their environment to make their shells, or exoskeletons. As diatoms die, their shells fall to the bottom of the bodies of water in which they live. It is believed that 30 million years ago the diatoms piled up to form thick beds of chalky fossilized sediment. These beds, known as diatomite or diatomaceous earth, were discovered after the waters receded.

Today, hundreds of large deposits of diatomaceous earth occur all over the world. Some are still underwater and some are found in ancient dried lake bottoms. It has been estimated that one cubic inch of diatomaceous earth may contain as many as 400 million shells, that's how small diatoms are!

Diatomite is mined and ground into a powder that looks and feels like talcum powder. This all natural powder is called Diatomaceous Earth (DE) or fossil shell flour. It is made up of approximately 33% silicon, 19% calcium, 5% sodium, 3% magnesium 2% iron and many other trace minerals such as titanium, boron, manganese, copper and zirconium.

What Do Fossilized Shells Have to Do With Improving Your Health?

It's all about the silica.

As I mentioned earlier, the shells of diatoms are made up mostly of silica (known by chemists as silicon dioxide). Silica is one of the most abundant minerals on the planet, but most of it is in a form that is unabsorbable by humans, and there are limited numbers of foods that contain an adequate amount of the absorbable form to supply the quantity our body needs. Silica is the most important trace mineral for human health. It plays an important role in many body functions and has a direct relationship to mineral absorption. For optimal health, the average human body needs to hold approximately seven grams of silica, a quantity far exceeding the figures for other important minerals such as iron and calcium.

Silica can be found in such foods as alfalfa, beets, brown rice and oats. Bell peppers and leafy green vegetables also provide silica, while asparagus, Jerusalem artichokes, parsley, sunflower seeds and grain husks, such as barley, millet and wheat contain smaller amounts.

Years ago, the silica found in our foods was adequate, but with today's depleted soils, only about one-third of the silica needed by our body is supplied in our food.

In our youth, our tissues absorb and maintain high levels of silica— enabling our bodies to remain flexible, resilient, and energetic—but as we age, and as our dietary sources of silica are not meeting our needs, our silica levels steadily decline until they become almost non-existent. In fact, 80% of all of our body's silica is used up by the time we become adults.

Our bodies need silica regardless of our age and even when diet is not the primary factor in cases of deficiencies, we often become deficient in this essential trace element simply thorough the aging process. The effect of this steady decrease in silica levels is a progressive decline in health, increasing fatigue and acceleration of the aging process.

Diatomaceous earth, which is very high in the absorbable form of silica, can replenish silica levels in the body, which can vastly improve your health and reverse many chronic problems. Studies show that high levels of silica can:

Improve Bone Health and Stop Osteoporosis: Calcium and vitamin D alone are not sufficient for bone growth, density, strength, and flexibility. In fact, the body cannot absorb and use calcium without the presence of silica. Recent data suggests that instead of promoting healing, calcium supplements actually speed up the leeching away of bone calcium and accelerate the degenerative process of osteoporosis and similar diseases that affect the connective tissues in the human body!

To re-mineralize and repair damaged bones, it is now advised that a sufficient silica supplement be taken daily because bones are composed of mainly of the minerals phosphorus, magnesium and calcium; however, these minerals need the presence of silica to be deposited into the bones, especially calcium. Silica hastens the healing of fractures and also diminishes scarring at the location of a fracture. A great deal of research evidence indicates that silica has the ability to "morph" itself into calcium through a transmutation process. Yes, silica actually has the ability to be turned into calcium when there is a calcium deficiency and the body needs it!

That fact alone should be reason enough to supplement with DE!!

Ward Off Alzheimer's: Scientists and researchers have long hypothesized that Alzheimer's disease is linked to a build up of aluminum in the brain, and links between aluminum in drinking water supplies and Alzheimer's have now been ascertained. A factor that had been overlooked is that silica reduces the accumulation of aluminum. When researchers added silica to aluminum-laced water supplies, it inhibited the aluminum from being absorbed. It also caused a proliferation in the excretion of aluminum in urine and lowered aluminum concentrations in the brain, liver, bone, spleen and kidneys. Silica, therefore, may be important in supporting neurological health.

Improve Heart and Lung Health: Silica can hinder the effects of coronary disease by fortifying blood vessels. Studies confirm that with age, silica disappears from the aorta, the heart's key blood vessel—thus weakening its critical connective tissue, and resulting in a greater cardiac risk. Studies have shown diatomaceous earth to significantly lower cholesterol by removing plaque and keeping arteries and veins supple. It also helps to regulate blood pressure. Silica also aids in the repair and maintenance of vital lung tissues and defending them from pollution.

Stop Premature Wrinkling and Sagging Skin: Silica is excellent for supporting bones and connective tissue. Your body needs healthy connective tissue for internal organs and the largest organ which is your skin. And you probably know that degrading connective tissue, with age is why wrinkles appear and skin begins to sag. Perhaps it is not age, but rather not enough silica that makes this happen.

Collagen, which is mostly made up of silica, is the glue that holds us together. Collagen accounts for up to 75% of the weight of the dermis and is responsible for the resilience and elasticity of the skin. Our connective tissues consist of collagen, elastin, mucopolysaccharides and mucous carbohydrates which aid in moisture retention. Their capacity to hold on to moisture keeps the connective tissue resilient and has apparent importance in the prevention of premature aging. All these valuable molecules house large quantities of silica. Also, many people with advanced arthritis suffer from bone deformation when tendons and ligaments in the joints lose flexibility due to loss of collagen. Tendons and ligaments need silica for health and flexibility.

Boost the Immune System: Silica may play an important role in the immune system and its biological response to harmful stimuli. Silica is necessary for the body to produce antibodies that fight off viruses, bacteria, allergens and other invaders that the body views as foreign.

Other Ways Diatomaceous Earth Can Improve Your Health

One of the most sought after benefits of diatomaceous earth is to cleanse the digestive tract. Many people consume a variety of processed foods, which contain various harmful chemicals and toxins. Plus, as air quality has decreased, we are constantly putting more toxins in our body that need to be removed.

Diatomaceous earth can purge any parasite, virus, bacteria, or toxin that is clinging to our digestive tract. While diatomaceous earth feels like a soft powder, the truth is that diatomaceous earth is actually a small cylinder with extremely sharp edges. As we consume diatomaceous earth, these sharp edges scrape away parasites, toxins, and viruses clinging to the lining of our digestive tract. Plus, any toxins or bacteria floating in our digestive tract are absorbed and trapped by diatomaceous earth. It has also been shown to cling to bad fats in the body, effectively reducing bad cholesterol. Diatomaceous earth is then expelled through our bowel movements and these harmful materials are removed.

Our skin and organs are strong and thick enough not to be affected by these sharp edges, which is why we feel no pain taking diatomaceous earth. After a few days, the body's digestive system can be thoroughly cleansed and can operate much more efficiently.

Are All Forms of Diatomaceous Earth the Same?

No! There are two types of DE—Food Grade and Pool Grade.

Most deposits of diatomaceous earth are from salt water sources, while only a few are from fresh water sources. Of the hundreds of DE deposits that exist worldwide, only 4 of them can be called food grade. The DE sourced from fresh water beds are of extremely high purity—so pure that it is called Food Grade Diatomaceous Earth, or also known as fossil shell flour. Food grade diatomaceous earth, the only form safe for human consumption, is a white powder, while other forms of diatomaceous earth may have a brown/reddish tint.

Also, there are two types of naturally occurring silica—crystalline silica and non-crystalline (amorphous) silica. Crystalline silica can be very dangerous, especially to our lungs if inhaled, while non-crystalline silica is completely safe, even for human and animal consumption. The fresh water diatoms contain mainly non-crystalline silica, while salt water diatoms can contain both types, with higher amounts of the crystalline form.

PERMA-GUARD™ is the trade name known worldwide for using a grade and quality of Diatomaceous Earth (DE) that is extremely pure. Brands that use Perma-Guard™ food grade Diatomaceous Earth contains less than 0 .5% crystalline silica. It is important to have a consistent shape diatom and no unwanted sediment. The shape of the diatom must be tubular with holes on the walls. It must be from fresh water because the fresh water diatoms form a harder shell and are less fragile that those from salt water. Salt water deposits contain a mix of diatom species. These deposits shapes and sediments are inconsistent making them unusable for safe human consumption.

Other types of diatomaceous earth have other industrial uses. It is heated to a very high temperature (about 1000°C or 1800°F). This type of DE is called "calcined." It is used for pool filters and other types of filtering but it is also used as a filler and can end up in paints, cosmetics, drugs, chemical insecticides, and other things. Food grade DE is never heated.

Bottom Line: ONLY use Food Grade Diatomaceous Earth for health purposes!!! Food grade diatomaceous earth is the purest form of diatomaceous earth and can be consumed and used by humans. Food grade diatomaceous earth is heavily regulated and must contain less than a certain amount of specific minerals.

Pool grade DE is calcinated and is a stronger, more potent form of DE. It is not to be used by humans and pool grade DE must be handled with gloves and a mask to ensure no throat irritations occur.

All DE should have a label that tells you if it has been calcined and how much crystalline silica it contains.

How Much to Take—How Safe Is It?

Food grade diatomaceous earth is a very fine powder and is very light due to its high porosity. It mixes easily into liquids and foods. Most people take a teaspoon or a tablespoon two or three times a day (up to a total of one quarter cup per day) for best results. Mix it into water, juice, smoothies or other foods.

It sort of resembles putting a spoonful of baking soda in water and drinking it!

Silica is water soluble; hence, once you get it into your body it easily absorbed via the intestinal wall and is also rapidly excreted. It does not accumulate in the body, so consistent daily supplementation is important. Studies have not found any negative side effects from too much silica. Its safety and extensive range of uses makes silica one of the most important minerals used in complementary therapy and alternative medicine. Regular supplementation could make a significant difference in your health.

Being approved by the FDA, Diatomaceous Earth has absolutely no dangerous side effects. It can be used by anyone, as long as the person doesn't suffer from a serious illness. As with any other health supplement, talk to a doctor before using DE. You might have some intolerance to it, so it's advisable to not risk your health. Pregnant and breastfeeding mothers can use it as long as their doctor has agreed.

What Other Uses Does DE Have?

One of the most commonly known uses for diatomaceous earth is as a non-toxic bug killer. DE is almost pure silica (with some beneficial trace minerals); under a microscope, it looks like shards of glass (glass is made from silica). On any beetle-type insect that has a carapace, like ants, fleas and cockroaches, the DE works under the shell and punctures the body, which then dehydrates their innards and the insect dies. DE is totally nontoxic. There is no buildup of tolerance like there is to poisons because the method of killing is PHYSICAL, not chemical. But rest assured, food grade DE doesn't hurt people or animals. In fact, it makes a great flea killer for your pets, and can also be added to their food as de-wormer. Many farmers use it to keep farm animals healthy, both inside and out. Food grade DE is also added to grains in storage because it keeps bugs from eating the grain. So if you eat any grains, you're probably already eating a little bit of DE!

Where to Get It!

You can buy food grade diatomaceous earth online. Processed-Free America sells it through its online store.

Written by Dee McCaffrey, CDC, who lost 100 pounds and has kept it off for over 20 years. For more info, go to: www.processedfreeamerica.org

Appendix 12

Summary of Protocols

The purpose of this list is to have all the protocols in one place, so you don't have to be hunting and flipping pages to find something. This information is also available on our website www.cdautism.org if you would like to print anything full size.

Diet

Permitted Foods List:

Note: Organic products are better but not required.

Proteins

☐ Beef
☐ Chicken
☐ Eggs
☐ Fish (small not large size)
☐ Pork
☐ Turkey

Fruit

☐ Most fresh fruits are permitted (except citrus, mango, pineapple, kiwi, and limit berries).
☐ Frozen fruit without added cream or sugar.
! NO canned fruit (nothing canned ever).
! Be careful of dried fruit as it may contain sugar.

Vegetables

☐ All vegetables are fine!!!
☐ Including French fries, however, not frozen fries or fries from fast food chains; these are often coated in flour.

Nuts

☐ Almonds
☐ Cashews
☐ Coconut
☐ Hazelnuts
☐ Walnuts

Grains

☐ Amaranth
☐ Buckwheat
☐ Corn
☐ Millet
☐ Quinoa
☐ Rice
☐ Sorghum
☐ Tapioca
☐ Xanthan gum

Beans

☐ All beans—EXCEPT soy
☐ Split Pea
☐ Garbanzo
☐ Lentils
☐ Navy
☐ Peanuts

Sweeteners

☐ Stevia (Best choice of all sweeteners)
☐ Agave syrup
☐ Honey
☐ Maple syrup (without added sugar)
☐ Xylitol

Prohibited Foods List

- Acetic acid (E260)
- Artificial flavoring
- Artificial sweeteners
- Bouillon cubes
- Bread
- Candy
- Cane sugar
- Carrageenan
- Catsup
- Chocolate milk
- Coloring
- Corn flakes
- Corn syrup
- Cow’s milk in any form (even lactose-free milk products)
- Flour tortillas
- Gelatin
- Malt
- Margarine
- Mayonnaise
- Microwave popcorn
- MSG
- Natural flavoring
- Noodle soup
- Oatmeal (except for Bob’s Red Mill GF oats)
- Pasta
- Piloncillo (unrefined sugar)
- Children´s nutritional shakes
- Play-Doh™
- Preservatives
- Processed meats (hotdogs, ham, sausage, bologna, cold cuts)
- Shellfish (full of toxins)
- Sodas
- Soy/fruit beverages
- Soy milk
- Soy sauce
- Sports drinks
- Sugar
- Yeast

Preparing Your First One drop CD Batch for the Day

1.	Start by filling the baby bottle with 8 fl. oz. (237ml) of water (distilled or reverse osmosis). NO alkaline water!	
2.	Place 1 drop of sodium chlorite solution into the CLEAN and DRY shot glass.	
3.	Add 1 drop of acidic activator (hydrochloric acid or citric acid) to the shot glass containing the drop of sodium chlorite. The number of drops may be higher if you are using a weaker activator. See chart on page 94.	
4.	Now wait the appropriate time for the mixture to react (see chart on page 94). You should see the color change from clear to slightly yellow. If there were more drops in the shot glass, the color change is more noticeable. You are also likely to notice the chlorine-like smell coming from the shot glass. Remember, this is NOT chlorine, but rather chlorine dioxide.	
5.	After the activation time has passed, pour a little water from the baby bottle into the shot glass and let it mix. This mostly stops the chemical reaction and insures you get most of the mixture out in the next step.	
6.	Lastly, pour all of the watered down mixture in the shot glass back into the baby bottle, and seal it tightly with the cap—don't leave it sitting open for any length of time. Think partially used soda pop and how you would want to keep that closed.	

Estimated Full Oral CD Doses by Weight

Use these numbers as a guide only. You may need to go up by as much as 50% or more over the indicated drops. Read chart as: **POUNDS / KILOGRAMS → DROPS OF CD (per 8 fl. oz. water)**

25/11→8	62/28→17	99/45→24	136/62→29	173/78→33	210/95→37
26/12→8	63/29→17	100/45→24	137/62→29	174/79→33	211/96→37
27/12→9	64/29→17	101/46→24	138/63→29	175/79→34	212/96→37
28/13→9	65/29→18	102/46→24	139/63→30	176/80→34	213/97→37
29/13→9	66/30→18	103/47→24	140/64→30	177/80→34	214/97→37
30/14→9	67/30→18	104/47→25	141/64→30	178/81→34	215/98→37
31/14→10	68/31→18	105/48→25	142/64→30	179/81→34	216/98→37
32/15→10	69/31→18	106/48→25	143/65→30	180/82→34	217/98→37
33/15→10	70/32→19	107/49→25	144/65→30	181/82→34	218/99→38
34/15→10	71/32→19	108/49→25	145/66→30	182/83→34	219/99→38
35/16→11	72/33→19	109/49→25	146/66→30	183/83→34	220/100→38
36/16→11	73/33→19	110/50→26	147/67→31	184/83→35	221/100→38
37/17→11	74/34→19	111/50→26	148/67→31	185/84→35	222/101→38
38/17→11	75/34→20	112/51→26	149/68→31	186/84→35	223/101→38
39/18→12	76/34→20	113/51→26	150/68→31	187/85→35	224/102→38
40/18→12	77/35→20	114/52→26	151/68→31	188/85→35	225/102→38
41/19→12	78/35→20	115/52→26	152/69→31	189/86→35	226/103→38
42/19→12	79/36→20	116/53→26	153/69→31	190/86→35	227/103→38
43/20→13	80/36→21	117/53→27	154/70→31	191/87→35	228/103→38
44/20→13	81/37→21	118/54→27	155/70→31	192/87→35	229/104→38
45/20→13	82/37→21	119/54→27	156/71→32	193/88→35	230/104→39
46/21→13	83/38→21	120/54→27	157/71→32	194/88→35	231/105→39
47/21→14	84/38→21	121/55→27	158/72→32	195/88→36	232/105→39
48/22→14	85/39→21	122/55→27	159/72→32	196/89→36	233/106→39
49/22→14	86/39→22	123/56→27	160/73→32	197/89→36	234/106→39
50/23→14	87/39→22	124/56→28	161/73→32	198/90→36	235/107→39
51/23→15	88/40→22	125/57→28	162/73→32	199/90→36	236/107→39
52/24→15	89/40→22	126/57→28	163/74→32	200/91→36	237/108→39
53/24→15	90/41→22	127/58→28	164/74→32	201/91→36	238/108→39
54/24→15	91/41→23	128/58→28	165/75→33	202/92→36	239/108→39
55/25→16	92/42→23	129/59→28	166/75→33	203/92→36	240/109→39
56/25→16	93/42→23	130/59→28	167/76→33	204/93→36	241/109→39
57/26→16	94/43→23	131/59→28	168/76→33	205/93→36	242/110→39
58/26→16	95/43→23	132/60→29	169/77→33	206/93→37	243/110→39
59/27→16	96/44→23	133/60→29	170/77→33	207/94→37	244/111→40
60/27→17	97/44→24	134/61→29	171/78→33	208/94→37	245/111→40
61/28→17	98/44→24	135/61→29	172/78→33	209/95→37	246/112→40

Advanced Protocols (After Full Oral Dose is Reached)

CD Enemas:

I recommend enemas no less frequently than every other day. You can do them more often particularly if you see parasites coming out. The ratio is 1-2 drops chlorine dioxide per 100 ml of filtered warm water. Work your way up to 500 mL – 1.5 L of water per enema, depending on the size of your child. You have the following options for methods of administering the enemas: a catheter and syringes, a multipurpose enema/douche bag, empty out a fleet enema, or a gravity bag. (note, if your child suffers from constipation you can add enemas and baths in before you get to full dose, alternating them on opposite days or opposite ends of the day).

Enemas should be started when the child reaches full oral dose. However, if your child suffers from constipation you can add enemas and baths on day one (see page 113), before you get to the full oral dose. However, alternate them on opposite days or opposite ends of the day. The following chart is an approximate amount of water based on the size of the person, and calculated the maximum of 2 drops of CD per 100mL of water.

Age/Size	Water Volume	Drops of CD
Child	1/2 Liter (500ml)	10
Adolescent	1 Liter (1,000ml)	20
Teen/Adult	up to 2 Liters (2,000ml)	40

CD Baths:

10-100 Activated drops of CD in a hot (to tolerance) bath with enough water to cover the child's body. Soak for 20 minutes. Apply on alternate enema days, unless you are doing daily enemas, in which case apply on opposite ends of the day.

The CD baths can go from as low as 10 drops to as high as 80 to 100 drops; it just depends on the person, and the size of your tub. With younger/smaller children we start with ten drops. Since we're doing baths every other day, start with ten drops on Monday, 11 drops on Wednesday, and 12 drops on Friday. Just keep going up until you get to 20 drops. The bigger and older the person is, the more drops they can tolerate. Fill the tub to a level that maximizes skin contact.

72/2 Protocol

Give one dose of CD every 2 hours for 72 hours straight, including the middle of the night. Apply every possible weekend.

This involves giving one dose of CD every two hours for 72 hours straight—including the middle of the night. Here are some additional thoughts and guidelines:

- Start protocol when you pick up your child from school on Friday.
- Give them the last 72/2 dose when you drop them off at school on Monday.
- Avoid giving CD enemas or CD baths during this protocol, UNLESS they are dumping parasites, in which case you might need to reduce the amount of CD on both the oral and enema dose to insure the child doesn't have a Herxheimer reaction.
- Why not give hourly doses during the day at the usual dose? It's simply too much if you are doing it all night as well.
- Watch for improvements on Tuesday or Wednesday each week.
- Ideally, get your spouse or significant other to help with every other nightly dose.

Supplement Dosing Overview

Supplement	Dose	Time of Day	Empty Stomach (ES) or With Food (WF)
Probiotics	1 cap	Before Bed	ES
Omega-3/ Omega-6	1 Tbsp	With any meal 1-3x / day	WF
L-Carnitine	250-1000mg/day	With any meal 1-3x / day	WF
GABA	Up to 2,500mg 2x/day	Upon waking & at bedtime	ES
5-HTP	50-200mg	Morning & Evening	WF
L-Theanine	Work up to 200-250mg/day	Mornings or Mornings & Nights	ES
Pycnogenol	25-40mg a day/as needed	Morning	WF
L-Carnosine	200-400mg 2x / day	Morning & Evening	WF
Taurine	500-1500mg/day	Morning, Noon & Night	ES
DMG	900mg/day	Morning	ES
TMG	500mg/day	Morning	ES
Enzymes	1 cap w/meals	Morning, Noon & Night	WF

Parasite Protocol

Please see the complete explanation of the parasite protocol starting on page 187. Day-to-day protocol charts can be found on page 198.

Chelation:

I suggest that when you decide to add clay baths into your child's protocol that you do so on the off days of the CD baths.

Follow the instructions on the package, and follow septic tank precautions if you have one.

Three days or so later, you can add in the Bio-Chelat. Again, follow the instructions on the package, and you can always start slowly and work your way up. Since Bio-Chelat™ doesn't alter the CD, the drops can be added to a single dose of CD. In addition, since the drops have no flavor, it is suitable to add them to water or any other drink your child might consume throughout the day.

Hyperbarics:

Option #1:

Two sessions of 60 minutes each for 20 days for a total of 40 sessions at 1.75 ATA.

Option #2:

One 90-minute session a day for twenty days at 1.75 ATA.

Instead of thinking outside of the box, get rid of the box.

~ Deepak Chopra

Appendix 13

Body Burden
The Pollution in Newborns

When I came across the following investigation I knew it was the perfect precursor to the incredible "Avoiding Autism" by Dr. Anju Usman and Beth Hynes. Babies are exposed to toxins in the womb and this study by the Environmental Working Group demonstrates so perfectly why the advice in the "Avoiding Autism" article is so very crucial to health of our children and for our species as a whole.

Body Burden — The Pollution in Newborns

A benchmark investigation of industrial chemicals, pollutants and pesticides in umbilical cord blood.

Environmental Working Group, July 14, 2005

Summary. In the month leading up to a baby's birth, the umbilical cord pulses with the equivalent of at least 300 quarts of blood each day, pumped back and forth from the nutrient- and oxygen-rich placenta to the rapidly growing child cradled in a sac of amniotic fluid. This cord is a lifeline between mother and baby, bearing nutrients that sustain life and propel growth.

Not long ago scientists thought that the placenta shielded cord blood — and the developing baby — from most chemicals and pollutants in the environment. But now we know that at this critical time when organs, vessels, membranes and systems are knit together from single cells to finished form in a span of weeks, the umbilical cord carries not only the building blocks of life, but also a steady stream of industrial chemicals, pollutants and pesticides that cross the placenta as readily as residues from cigarettes and alcohol. This is the human "body burden" — the pollution in people that permeates everyone in the world, including babies in the womb.

In a study spearheaded by the Environmental Working Group (EWG) in collaboration with Commonweal, researchers at two major laboratories found an average of 200 industrial chemicals and pollutants in umbilical cord blood from 10 babies born in August and September of 2004 in U.S. hospitals. Tests revealed a total of 287 chemicals in the group. The umbilical cord blood of these 10 children, collected by Red Cross after the cord was cut, harbored pesticides, consumer product ingredients, and wastes from burning coal, gasoline, and garbage.

This study represents the first reported cord blood tests for 261 of the targeted chemicals and the first reported detections in cord blood for 209 compounds. Among them are eight perfluorochemicals used as stain and oil repellants in fast food packaging, clothes and textiles — including the Teflon chemical PFOA, recently characterized as a likely human carcinogen by the EPA's Science Advisory Board — dozens of

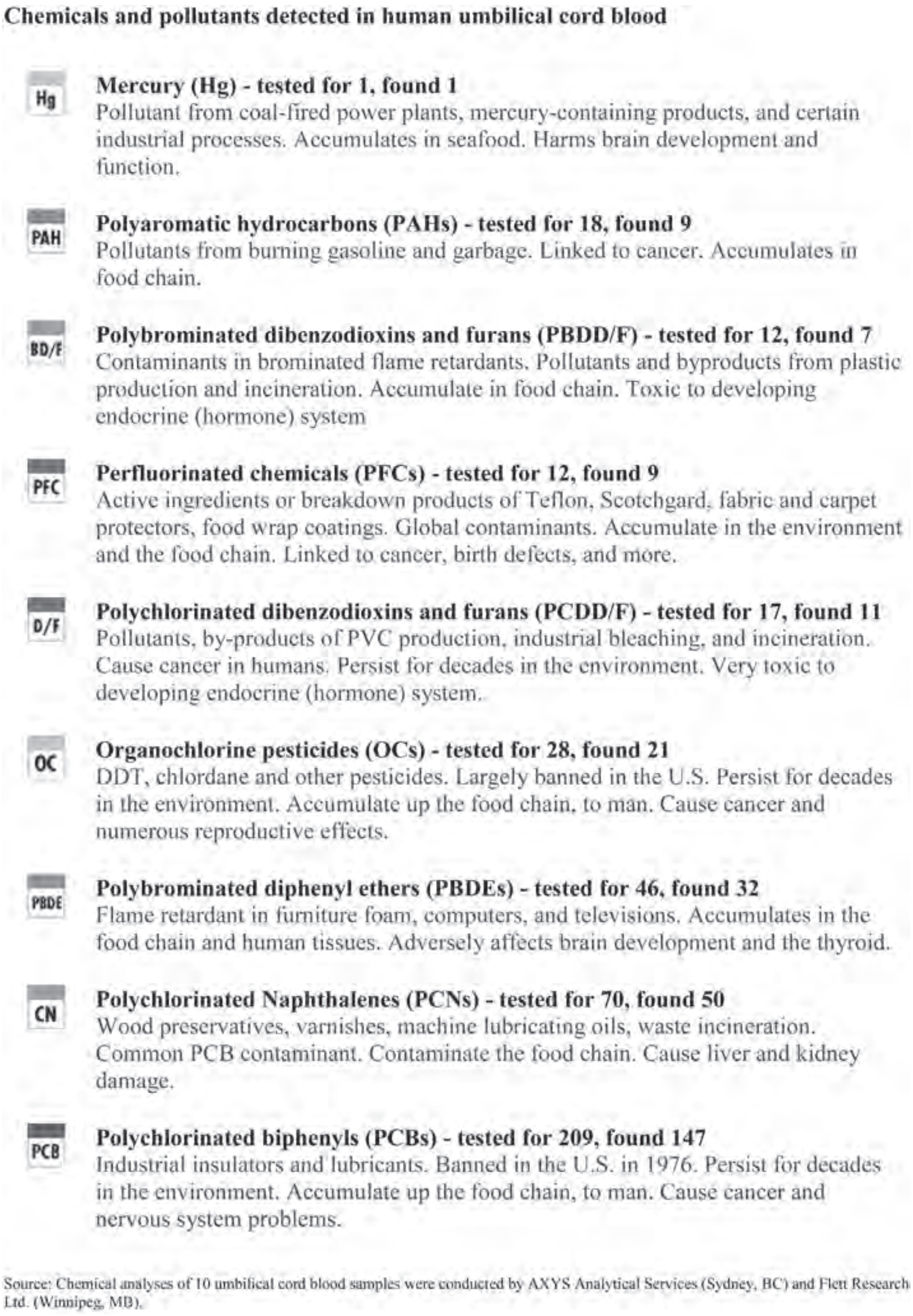

Chemicals and pollutants detected in human umbilical cord blood

Hg **Mercury (Hg) - tested for 1, found 1**
Pollutant from coal-fired power plants, mercury-containing products, and certain industrial processes. Accumulates in seafood. Harms brain development and function.

PAH **Polyaromatic hydrocarbons (PAHs) - tested for 18, found 9**
Pollutants from burning gasoline and garbage. Linked to cancer. Accumulates in food chain.

BD/F **Polybrominated dibenzodioxins and furans (PBDD/F) - tested for 12, found 7**
Contaminants in brominated flame retardants. Pollutants and byproducts from plastic production and incineration. Accumulate in food chain. Toxic to developing endocrine (hormone) system

PFC **Perfluorinated chemicals (PFCs) - tested for 12, found 9**
Active ingredients or breakdown products of Teflon, Scotchgard, fabric and carpet protectors, food wrap coatings. Global contaminants. Accumulate in the environment and the food chain. Linked to cancer, birth defects, and more.

D/F **Polychlorinated dibenzodioxins and furans (PCDD/F) - tested for 17, found 11**
Pollutants, by-products of PVC production, industrial bleaching, and incineration. Cause cancer in humans. Persist for decades in the environment. Very toxic to developing endocrine (hormone) system.

OC **Organochlorine pesticides (OCs) - tested for 28, found 21**
DDT, chlordane and other pesticides. Largely banned in the U.S. Persist for decades in the environment. Accumulate up the food chain, to man. Cause cancer and numerous reproductive effects.

PBDE **Polybrominated diphenyl ethers (PBDEs) - tested for 46, found 32**
Flame retardant in furniture foam, computers, and televisions. Accumulates in the food chain and human tissues. Adversely affects brain development and the thyroid.

CN **Polychlorinated Naphthalenes (PCNs) - tested for 70, found 50**
Wood preservatives, varnishes, machine lubricating oils, waste incineration. Common PCB contaminant. Contaminate the food chain. Cause liver and kidney damage.

PCB **Polychlorinated biphenyls (PCBs) - tested for 209, found 147**
Industrial insulators and lubricants. Banned in the U.S. in 1976. Persist for decades in the environment. Accumulate up the food chain, to man. Cause cancer and nervous system problems.

Source: Chemical analyses of 10 umbilical cord blood samples were conducted by AXYS Analytical Services (Sydney, BC) and Flett Research Ltd. (Winnipeg, MB).

widely used brominated flame retardants and their toxic by-products; and numerous pesticides.

Of the 287 chemicals we detected in umbilical cord blood, we know that 180 cause cancer in humans or animals, 217 are toxic to the brain and nervous system, and 208 cause birth defects or abnormal development in animal tests. The dangers of pre- or post-natal exposure to this complex mixture of carcinogens, developmental toxins and neurotoxins have never been studied.

Chemical exposures in the womb or during infancy can be dramatically more harmful than exposures later in life. Substantial scientific evidence demonstrates that children face amplified risks from their body burden of pollution; the findings are particularly strong for many of the chemicals found in this study, including mercury, PCBs and dioxins. Children's vulnerability derives from both rapid development and incomplete defense systems:

- A developing child's chemical exposures are greater pound-for-pound than those of adults.
- An immature, porous blood-brain barrier allows greater chemical exposures to the developing brain.
- Children have lower levels of some chemical-binding proteins, allowing more of a chemical to reach "target organs."
- A baby's organs and systems are rapidly developing, and thus are often more vulnerable to damage from chemical exposure.
- Systems that detoxify and excrete industrial chemicals are not fully developed.
- The longer future life span of a child compared to an adult allows more time for adverse effects to arise.

The 10 children in this study were chosen randomly, from among 2004's summer season of live births from mothers in Red Cross' volunteer, national cord blood collection program. They were not chosen because their parents work in the chemical industry or because they were known to bear problems from chemical exposures in the womb. Nevertheless, each baby was born polluted with a broad array of contaminants.

U.S. industries manufacture and import approximately 75,000 chemicals, 3,000 of them at over a million pounds per year. Health officials do not know how many of these chemicals pollute fetal blood and what the health consequences of in utero exposures may be.

Had we tested for a broader array of chemicals, we would almost certainly have detected far more than 287. But testing umbilical cord blood for industrial chemicals is technically challenging. Chemical manufacturers are not required to divulge to the public or government health officials methods to detect their chemicals in humans. Few labs are equipped with the machines and expertise to run the tests or the funding to develop the methods. Laboratories have yet to develop methods to test human tissues for the vast majority of chemicals on the market, and the few tests that labs are able to conduct are expensive. Laboratory costs for the cord blood analyses reported here were $10,000 per sample.

A developing baby depends on adults for protection, nutrition, and, ultimately, survival. As a society we have a responsibility to ensure that babies do not enter this world pre-polluted, with 200 industrial chemicals in their blood. Decades-old bans on a handful of chemicals like PCBs, lead gas additives, DDT and other pesticides have led to significant declines in people's blood levels of these pollutants. But good news like this is hard to find for other chemicals.

The Toxic Substances Control Act, the 1976 federal law meant to ensure the safety of commercial chemicals, essentially deemed 63,000 existing chemicals "safe as used" the day the law was passed, through mandated, en masse approval for use with no safety scrutiny. It forces the government to approve new chemicals within 90 days of a company's application at an average pace of seven per day. It has not been improved for nearly 30 years—longer than any other major environmental or public health statute—and does nothing to reduce or ensure the safety of exposure to pollution in the womb.

Because the Toxic Substances Control Act fails to mandate safety studies, the government has initiated a number of voluntary programs to gather more information about chemicals, most notably the high production volume (HPV) chemical screening program. But these efforts have been largely ineffective at reducing human exposures to chemicals. They are no substitute for a clear statutory requirement to protect children from the toxic effects of chemical exposure.

In light of the findings in this study and a substantial body of supporting science on the toxicity of early life exposures to industrial chemicals, we strongly urge that federal laws and policies be reformed to ensure that children are protected from chemicals, and that to the maximum extent possible, exposures to industrial chemicals before birth be eliminated. The sooner society takes action, the sooner we can reduce or end pollution in the womb.

Copyright © Environmental Working Group, www.ewg.org. Reprinted with permission.

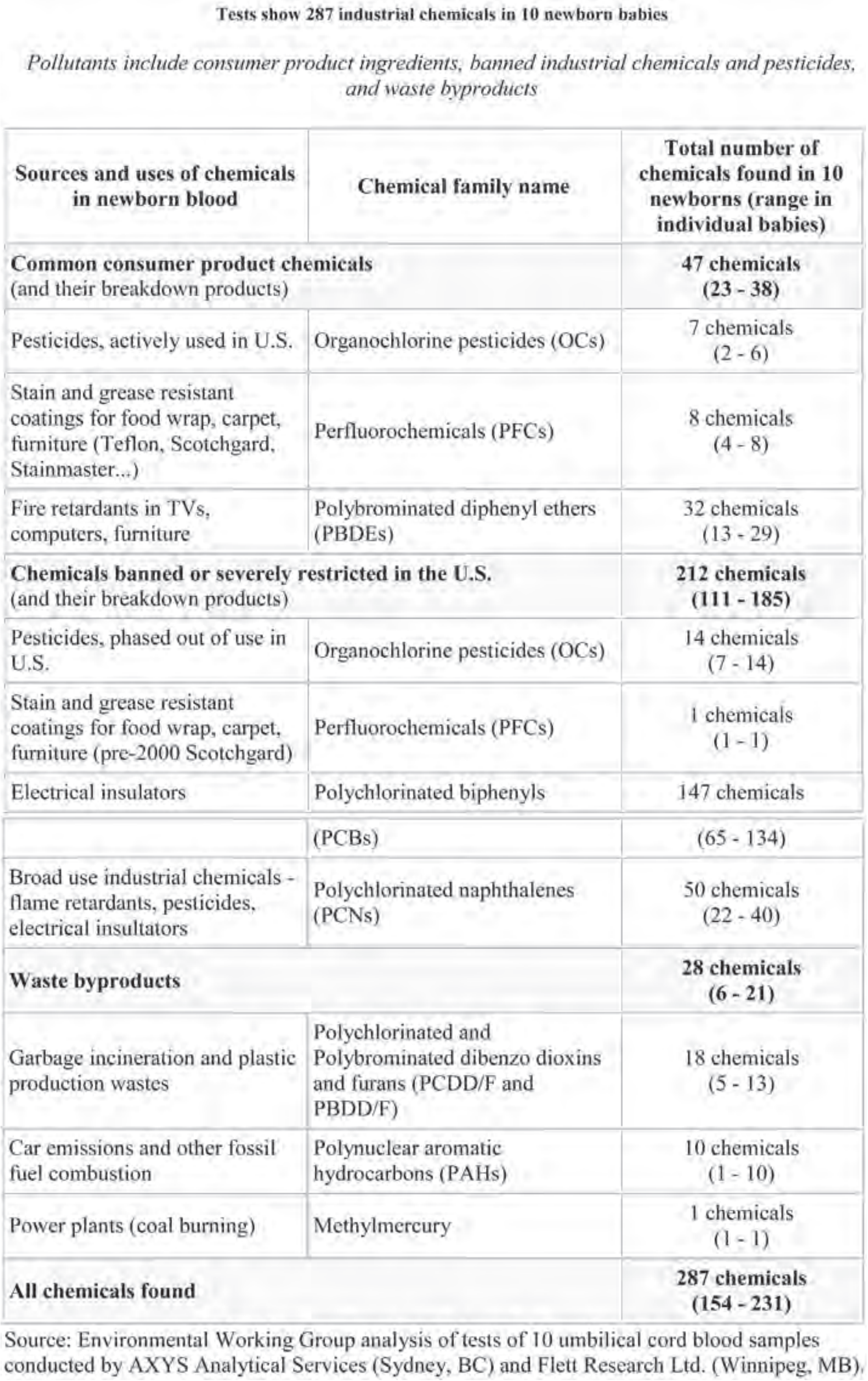

Tests show 287 industrial chemicals in 10 newborn babies

Pollutants include consumer product ingredients, banned industrial chemicals and pesticides, and waste byproducts

Sources and uses of chemicals in newborn blood	Chemical family name	Total number of chemicals found in 10 newborns (range in individual babies)
Common consumer product chemicals (and their breakdown products)		**47 chemicals** **(23 - 38)**
Pesticides, actively used in U.S.	Organochlorine pesticides (OCs)	7 chemicals (2 - 6)
Stain and grease resistant coatings for food wrap, carpet, furniture (Teflon, Scotchgard, Stainmaster...)	Perfluorochemicals (PFCs)	8 chemicals (4 - 8)
Fire retardants in TVs, computers, furniture	Polybrominated diphenyl ethers (PBDEs)	32 chemicals (13 - 29)
Chemicals banned or severely restricted in the U.S. (and their breakdown products)		**212 chemicals** **(111 - 185)**
Pesticides, phased out of use in U.S.	Organochlorine pesticides (OCs)	14 chemicals (7 - 14)
Stain and grease resistant coatings for food wrap, carpet, furniture (pre-2000 Scotchgard)	Perfluorochemicals (PFCs)	1 chemicals (1 - 1)
Electrical insulators	Polychlorinated biphenyls (PCBs)	147 chemicals (65 - 134)
Broad use industrial chemicals - flame retardants, pesticides, electrical insultators	Polychlorinated naphthalenes (PCNs)	50 chemicals (22 - 40)
Waste byproducts		**28 chemicals** **(6 - 21)**
Garbage incineration and plastic production wastes	Polychlorinated and Polybrominated dibenzo dioxins and furans (PCDD/F and PBDD/F)	18 chemicals (5 - 13)
Car emissions and other fossil fuel combustion	Polynuclear aromatic hydrocarbons (PAHs)	10 chemicals (1 - 10)
Power plants (coal burning)	Methylmercury	1 chemicals (1 - 1)
All chemicals found		**287 chemicals** **(154 - 231)**

Source: Environmental Working Group analysis of tests of 10 umbilical cord blood samples conducted by AXYS Analytical Services (Sydney, BC) and Flett Research Ltd. (Winnipeg, MB).

The ultimate ignorance is the rejection of something you know nothing about and refuse to investigate.

~ Dr. Wayne Dyer

Appendix 14

Avoiding Autism

By Anju Usman, MD and Beth C. Hynes JD, MBA

Whether or not you are the parent of a special needs child, please pay special attention to this article, we do believe autism is totally preventable and that starts with pregnancy. Here's to your health!

Every parent's dream is to have a healthy baby. Over the past ten years of working with families devastated by the diagnosis of autism, we, along with our patients' parents, have realized one key fact: if we knew then what we know now, the diagnosis of the disease called autism could very well have been prevented. Considering that the cause(s) of the current epidemic of autism spectrum disorders (ASD) are unknown (and scant research dollars are being dedicated to finding the most likely culprits), there is no surefire method to avoid having your infant regress into this devastating state. That sad fact, however, does not mean that there are no proactive steps parents can take to try to mitigate the chances of this disease overtaking their baby and stealing their hopes and dreams for that child's life.

Having reviewed the medical and laboratory test results of thousands of children afflicted with ASD, clinicians using a biomedical approach to treating autism see repeating patterns in these children—impaired methylation and detoxification, mitochondrial dysfunction, gastrointestinal distress, and immune dysregulation compounded by chronic viral, fungal and bacterial infections and a burden of heavy metals and other toxins. From this perspective, we have come to view autism as a multifactorial medical disease that can be treated and overcome and, therefore, possibly avoided. Coming straight from the heart, this advice is for thinking, active parents and prospective parents who want to know what they can do to reduce the risk that their infant will regress into autism. In our practice, we have helped many parents go on to have healthy children even though they already have a child with autism. Considering that the information provided here is not harmful but, to the contrary, is helpful in promoting improved maternal and fetal health, there is little downside to pursuing these strategies while simultaneously minimizing the risk of having your child slip away into the autism epidemic. According to the latest statistics out of California, that epidemic continues to grow from the already shocking 1 out of 150 children. The aim here is to provide an analytical construct that will guide parents in making choices for both their own health and that of their baby before, during, and after pregnancy.

Think of the analogy of the straw that broke the camel's back. Human beings start life as infants with some "straws" in their saddle – these are predispositions to be harmed by certain chemicals, to be susceptible to certain diseases, or to have a less robust detoxification capacity. The goal of gestating and parenting an infant in today's toxic world is to not accumulate additional straws during your baby's gestation and infancy, thereby avoiding adding the straw that breaks the camel's back. Autism seems to occur when the infant's body is so overwhelmed by toxins, viruses and/or pathogenic elements that typical development is arrested and/or derailed.

Children diagnosed with autism have a variety of medical problems, the underpinnings of which can be ameliorated and/or healed. The symptoms associated with autism occur through malfunctioning of the infant immune, gastrointestinal, and central nervous systems ("Infant Body Systems") – all of which, when functioning properly, contribute to the healthy development of the brain. Viruses, heavy metals, toxins, fungi, and bacteria together form a weapon that disrupts the development of, and harms the functioning of, these critical Infant Body Systems. As researchers have recently found, "a single toxicant may promote different immune-associated diseases that are dependent upon the specific window of early life exposure, the gender of the exposed offspring, and the genetic background of the offspring." Our construct will consist of actions you can take to promote the healthy development of these Infant Body Systems as well as steps you can take to avoid upsetting the healthy development of those systems.

Due to the symbiotic relationship between a mother and her developing fetus, care must be taken by the mom before, during, and after pregnancy (if nursing) to both avoid exposure to harmful elements and to promote optimal maternal detoxification processes so that the host body remains as clean an environment as possible within which vigorous infant development can unfold.

Only recently have researchers begun closely linking environmental factors with developmental delays in our children. A host of studies show how environmental factors including diet, nutrition, air quality, clean water, heavy metals (e.g., mercury, lead, arsenic, and aluminum), xenobiotics (e.g., chemicals, pesticides, plasticizers, and perchlorates), pharmaceuticals, and vaccinations can change your child's inborn risk.

This area of study is called epigenetics – the molecular changes in our DNA caused by environmental factors. In 2003, researchers at Duke University found that rats carrying a gene for obesity, cancer, and diabetes, when fed a diet high in B-12, folic acid, choline, and methyl-rich foods, had babies that were born healthy and with a reduced risk of developing disease; this research revealed that these genetic changes persisted across four generations. Thus, when you eat a healthy diet, you are affecting not only the health of your own kids but also your grandchildren, great-grandchildren, and great-great-grandchildren. Researchers are finding hundreds of unwanted and potentially toxic chemicals in the placenta and cord blood of newborns as well as in mothers' breast milk. That means that today's children are born with a toxic burden their infant bodies must contend with from day one. Studies out of China, Texas, and San Francisco show that if pregnant mothers can decrease their exposure to pollutants during gestation, they can improve health outcomes for their children – especially brain development. Compounds such as mercury, cadmium, nickel, trichloroethylene, and vinyl chloride in the air around the birth residence caused a 50% increase in the risk of autism.

Numerous studies document the neurotoxic effects of heavy metals – especially dangerous to the developing brain is mercury. Yet, many of our children are actually injected with the mercury-containing preservative thimerosal in their infant vaccines. Although thimerosal has been removed from, or perhaps phased out of, many childhood vaccines, it still remains in the flu vaccine, which is recommended by health authorities to pregnant woman as well as to children under age 2. Mercury exposure is such a concern that pregnant women are advised by health authorities not to eat fish because of potential mercury contamination – yet injecting thimerosal is much more damaging than the exposure suffered through eating it in food such as fish. Vaccines also contain aluminum, formaldehyde, and other toxic elements as well as live viruses that, in a subset of infant bodies, can trigger harmful impacts. Heavy metals have also been identified as factors affecting human fertility. Diagnosing and reducing the heavy metal burden of women improved the spontaneous conception chances of infertile women. Women with many dental amalgams had a higher incidence of miscarriages and a higher excretion of mercury when given the chelating agent DMPS (2,3-Dimercapto-1-propanesulfonic acid). It has been found that DMPS was a useful and complementary method to increase fertility compared to hormone therapy in infertile women.

With all this evidence mounting against various environmental toxins affecting human development, an international assembly of scientists, doctors, toxicologists, and researchers sponsored by the

World Health Organization (WHO), the European Environmental Agency, the Centers for Disease Control (CDC), and the National Institutes of Health (NIH) gathered in the Faroe Islands in May 2006. Their consensus statement was clear: “it is time to take action, now.” As expected, the declaration highlighted the results of hundreds of studies that showed early fetal exposures to toxic substances, even at low concentrations, can cause health problems later in life. There are a growing number of chemicals that affect the developing embryo, fetus, and infant, including the pesticides DDT, atrazine, methoxychlor, and vinclozolin. Also suspect are plastics and the epoxy resin bisphenol A, plasticizing agents called phthalates, mercury, lead, arsenic, organotoxins, polychlorinated biphenols (PCBs), carbon monoxide, smog, tobacco smoke, and alcohol. The health problems caused by exposure to these substances comprise a dizzying array of maladies including cancer, diabetes, obesity, asthma, and allergies as well as reproductive, cardiovascular, neurological, cognitive, endocrine, psychological, immune, and respiratory troubles.

Other troubling chemicals include parabens (found in various creams and lotions) and triclosan, an antibacterial agent (found in toothpaste, soaps, and cleaning products). Although parabens and triclosan are not ingested, absorbing chemicals through the skin can be far more dangerous than swallowing because transdermal applications are transported directly to the bloodstream. Most women absorb 51 pounds of chemicals every year from their cosmetics alone! Additional exposures emanate from daily lifestyle, profession, and locale. For those elements for which we have the flexibility to make choices, we must choose wisely. After all, it is not just our health that is being affected by our choices but the health of our children and progeny as well.

Basic principles to guide you in decision making surrounding pregnancy are as follows:

1. You are what you eat (and drink); therefore, make healthy choices in what you consume.
2. Skin is the largest organ in the body, so be very careful of what you rub into yours and your baby’s.
3. Think beyond “green”: What is “green” for the environment is not always what is healthiest for the body, but what is healthiest for the body is always “green.”

Suggestions to promote these principles are below and are cumulative, so keep following each set of suggestions as this journey into and through parenthood progresses.

Pre-Pregnancy

While planning a pregnancy, we recommend that you clean up any toxicity in your body and begin to follow a more organic, healthy lifestyle. Remember, the less toxic you are, the better for you and your future baby. Undertake a sequential detoxification program that targets the liver and colon; this type of program can take six months or more and should not be done while pregnant. In a careful manner, with an experienced dentist, remove amalgams from your teeth, which also can take six months or more. Take the time now to find organic, nontoxic makeup, hair, and body products that you like and start integrating them into your daily life. Ask your doctor to run some tests to determine any additional specific supplementation you may need to optimize levels within your body; a good place to start is your copper-zinc ratio (1:1 is the ideal ratio), thyroid function (TSH, free t3, free t4), Vitamin D 25 OH, vitamin A levels, and total cholesterol (low cholesterol is associated with preterm births).

Next, clean up your living and working environment. Remove all harmful chemical cleaning agents from your cleaning routine at home and at the office, and instead use cleaning products labeled level 1 by the EPA. Do not forget to include products for dishwashing and clothing detergent in your cleanup, and avoid toxic dry cleaning as much as possible. Finally, improve your nutrition with a targeted vitamin supplementation program to include omega-3 essential fatty acids, sublingual methyl B-12, folinic acid, vitamin D3, zinc, and antioxidants.

DO	DON'T
• Eat organic, hormone-free food • Drink organic green tea, filtered water, and antioxidant rich organic juices • Use stevia, raw organic honey, and xylitol as sweeteners • Go for walks and get some sunshine daily • Use aluminum-free natural deodorant in them • Use natural hennas to color your hair • Use cast iron, glass, or stainless steel cookware • Use chemical-free cleaning products in your home	• Consume fish or foods w/MSG or food dyes • Drink soda, carbonated beverages, or alcohol • Consume artificial sweeteners • Be exposed to lawn chemicals or second-hand smoke • Use moisturizers or makeup with chemicals or parabens • Use chemical dyes, perms, or other such hair treatments • Cook with pans that are non-stick, Teflon coated, or made from aluminum

During Pregnancy

We would suggest that you find a holistic health practice to guide you during pregnancy and delivery as there are many decisions to make during this time. In general, we suggest avoiding medications to the extent possible as well as acetaminophen because it hinders normal detoxification. Add zinc, calcium, essential fatty acids, and prenatal vitamins to your daily supplement intake. Discontinue use of nail polish and any makeup (including lipstick) products that contain parabens and other toxins. Use fluoride-free toothpaste as fluoride interferes with iodine metabolism, which is an issue implicated in mental retardation worldwide.

DO	DON'T
• Yoga and engage in stress management techniques, such as massages and listening to soothing music • Eat fermented foods, cook with organic coconut oil, use organic raw apple cider as salad dressing, and consume healthy fats, and cold pressed oil • Use natural remedies for pain, like homeopathic arnica • Drink kombucha and take probiotics • Run an air filter in your bedroom while you sleep	• Start a rigorous exercise program, sit in a sauna, or get dental work (not even cleanings) • Wait until labor arrives to discuss NOT subjecting your baby to vaccines • Talk on a cell phone without a headset or work with a laptop computer on your lap • Undertake a major home renovation (concern is for lead and other toxins in the process)

Infancy

Healthy babies grow and develop perfectly when nature's biochemical processes are allowed to unfold uninterrupted. The sequence of events that leads an infant to begin to focus, develop gross and fine motor skills, begin to speak and to walk is based on a delicate but deliberate series of biochemical processes and chemical interactions. A primary goal of raising a healthy infant is to avoid disruption or interference with this sequence. Babies look fragile and, indeed, they are. An infant's immune system, which guards the rest of the body from harm, is immature and requires time and peace to mature maximally. Thus, help your infant progress uninterrupted by avoiding the introduction of toxins, viruses, allergens, and heavy metals into their bodies. In light of this, parents should consider an alternate, gradual vaccine schedule of carefully thought through vaccines for infants being raised in the domestic United States.

If you have amalgams and plan to nurse, you should send a sample of your breast milk to a specialty lab for heavy metal testing. If you plan to use formula, use those containing DHA (docosahexanenoic acid), an essential fatty acid critical to the healthy development of the central nervous system. In terms of food introduction, organic baby food is recommended. Start feeding with organic rice cereal. Avoid the introduction of soy, gluten, or dairy until after the baby turns two years of age. After one year, supplement your baby with a quarter teaspoon of mercury-free cod liver oil – again this is to increase the supply of omega-3 essential fatty acids.

In terms of coping with baby colds and other minor illnesses, unless symptoms are severe, less is more when it comes to treatment. For fevers over 101˚F, treat with a tepid bath or dye-free ibuprofen. Fevers, while nerve-wracking for new parents, are the response of a healthy immune system reacting to kill off an invading virus through heat. Antibiotics should be used sparingly and only for confirmed bacterial infections (they do not alleviate viral infections). Remember, antibiotic use disrupts the normal gut flora and promotes the overgrowth of yeast and resistant organisms that, in turn, harms the optimal functioning of the immune system. Bear in mind that most ear infections are viral and are thus not treatable with antibiotics. Use homeopathic ear drops to help ease the symptoms associated with ear infections and colds.

DO	DON'T
• Invest in an organic baby matress, bedding and pillows and hypo allergenic encasements • Bathe your baby daily in warm filtered water — enjoy the experience with your baby! • Feed with all organic and hormone-free products • Feed baby using glass bottles • Wearing a hat, walk outside with baby to get 10 to 15 minutes of sunshine daily • Run an air filter in baby's bedroom	• Clothe the baby in pajamas soaked in flame retardant chemicals • Use soaps, moisturizers, or other "baby products" on the skin as such items are unnecessary and contain harmful chemicals • Introduce dairy, gluten, or soy until after age two • Feed baby from plastic bottles or cups or microwave formula or breast milk • Take your infant into polluted and heavily populated locations

Source: http://www.autismfile.com/what-is-autism-facts/autism-symptoms/avoiding-autism

Reprinted with permission.

When a truth is not given complete freedom,
freedom is not complete.

~ Vaclav Havel

Appendix 15

RECIPES

Cooking with Ana

Since 2001 Ana Bobadilla has been a chef to Kerri Rivera and family, including her son Patrick, who has a diagnosis of regressive autism. Over the years she has mastered gluten-free/casein-free to vegan to Atkins and back again. Her dishes always impress and nourish guests and family members alike.

One of the most important pieces to the autism recovery puzzle is diet. As Kerri always says, if your goal is to heal autism, you must do *The Diet.* However, that is easier said than done for many families whose children have self-limited to grilled cheese sandwiches and pizza.

A book is in the works containing dozens of delicious recipes using natural ingredients that can be found in almost any country. Ana has found unique ways to make foods that children on the spectrum can enjoy, without breaking the bank. The ingredients that she uses can be found all over the world, not just the US. They are presented here in a very approachable way so even if you are not a whiz in the kitchen your child will be able to enjoy what you serve.

The following is a sneak peek from *Cooking with Ana*:

Almond Milk

Amount: 1 L of milk

- 1 cup of peeled almonds
- 1 L of filtered water
- Stevia and vanilla to taste

Preparation:

To a saucepot of boiling water, add the almonds. Remove immediately from the fire. Allow almonds to soak for 20 minutes, remove them from water, rinse, dry and blend in a liter of filtered water. Strain the liquid through a piece of cheesecloth until all the water is drained. Add agave syrup and vanilla to taste.

Vegetable Fries

3 Servings

Ingredients:

- 3 Carrots
- 1 cup oil
- Salt to taste

Preparation:

Peel the carrots and then form very thin slices using the same peeler. Pat dry with a piece of cloth to remove any excess water. Heat oil in a frying pan and fry carrots. When frying be very careful because they can quickly burn, be sure to have a slotted spoon ready to remove from the oil as soon as they are crisp. After frying carrots remove excess oil with a paper towel and sprinkle some salt on top of them.

You can also do this recipe with banana, potato, sweet potato or jackfruit. This is a great alternative to French fries or if you want to eat something crunchy.

Beef or Chicken Tostadas

4 Servings

Ingredients:

- 5 ground chicken breasts or beef filets
- 1 cup of rice or coconut flour
- 1 cup of oil
- Salt and pepper to taste

Preparation:

In a mixing bowl, combine the ground chicken, flour, salt, pepper and a ½ cup of oil. Mix thoroughly until a uniform dough results. Take approximately 2 tablespoons of the dough and form small tortillas/flat circles with it. They should be approximately ¼ inch think and 4 inches wide. Fry in a layer of oil until cooked through, and a thin outer crust forms, or bake in a 350 degree oven, after pouring a dollop of oil on top of each tostada.

Coconut Bars

15-18 bars

Ingredients:

- 3 cups of vegan butter (or coconut oil)
- 1 cup of honey
- 1 cup of raisins
- 1 cup of grated of coconut
- 1 cup of walnuts
- 1 tablespoon of salt

Preparation:

In a saucepan over medium heat combine honey and butter, stirring constantly for a few minutes until thoroughly combined. Remove from heat and mix with all other ingredients. Spread in a 9x13 casserole pan without oil, and bake in an 350°F (180°C) over for 25 minutes until it sets. Allow to cool, and cut into small rectangles.

Get people back into the kitchen
and combat the trend toward
processed food and fast food.

~ Andrew Weil

Appendix 16

Websites We Like

General:

- **Our website containing the latest information about our research, videos, forums, etc.**
  www.cdautism.org

- **The Spanish version of our website.**
  www.autismo2.com

- ***Curando el Autismo* is a Puerto Rican organization created by three mothers of children recovered from autism who have an interest in helping the community learn about their options.**
  www.curandoelautismo.com

- **Defeating Autism - Non-Profit Organization in Venezuela that utilizes Kerri's Protocols.**
  www.fundacionvenciendoelautismo.blogspot.mx

- **Autism Research Institute originally founded by Dr. Bernard Rimland. Home of the ATEC survey.**
  www.autism.com

- **Our public Facebook page.**
  www.facebook.com/groups/AutismCD

- ***CD Health*: Facebook group dedicated to using the protocol for non-autism related issues.**
  www.facebook.com/groups/mojoother

Diet:

- **Website forcused on the *Specific Carbohydrate Diet*™ (SCD) in English.**
  www.pecanbread.com

- **Top 10 FAQs about the *Specific Carbohydrate Diet* (SCD) in Spanish.**
  www.pecanbread.com/pandenuez/10preguntas.html

- **Autism Network for Dietary Intervention**
  www.autismndi.com

- **Gluten Free/Casein Free Diet Intervention website.**
  www.gfcfdiet.com

- **Intestinal health through diet with the Specific Carbohydrate Diet™**
  www.breakingtheviciouscycle.info

- **Resource for naturally healing digestive disease, reducing stress and living a long, healthy life.**
  www.scdlifestyle.com

- ***The Feingold® Association* is a non-profit organization helping provide dietary management and generate public awareness of the role food and synthetic additives play in behavior, learning and health problems.**
  www.feingold.org

Chlorine Dioxide (CD):

- **Jim Humble's MMS (CD) website.**
  www.jimhumble.org

- **Another popular site about MMS (CD).**
  www.mmswiki.org

- **Spanish site selling various books about MMS (CD)**
  www.voedia.com

- **Andreas Kalcker's website (in English)**
  www.andreaskalcker.com

- **CD research site by Andreas Kalcker (in Spanish)**
  www.medicasalud.com

- **A huge public health forum with thousands of discussions about every imaginable health topic. Also includes surveys about various topics.**
  www.curezone.com

- **Good sources of CD and other related supplies.**
  www.wps4sale.com
  www.mightyguts.com

- **The latest information on CDH.**
  http://www.mmsinfo.org/infosheets/infosheet_cdh.pdf

Hyperbarics:

- **Good source of information about Hyperbaric Oxygen Therapy.**
  www.hyperbaric-oxygen-info.com

Other Valuable Source of Information:

- **PubMed comprises of more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books.**
  www.ncbi.nlm.nih.gov/pubmed

- **Parents Helping Parents - P.A.N.D.A.S. Network: A Resource Library of Medical Studies and Case Histories**
  www.pandasnetwork.org

- **International OCD Foundation**
  www.ocfoundation.org/PANDAS/

Videos / Audio of Interest:

- **Video Testimonials. This link contains various health testimonials related to MMS, many are about autism.**
  www.youtube.com/mmstestimonials

- **AutismOne 2012: 38 Children Recovered in 20 months with MMS by Kerri Rivera**
  www.autismone.org/content/38-children-recovered-20-months-mms-0

- **Video about Kerri, produced by Daniel Bender in 2012.**
  www.youtube.com/watch?v=HnCiX5bll-Y

- **Kerri speaks in Bulgaria, 2013 (Voice in English, slides in Bulgarian)**
  www.youtube.com/watch?v=4OxNGIEWzds

- **Kerri on the Robert Scott Bell Show**
  http://www.youtube.com/watch?v=-1zJhr5VhxE

- **MMS Autism Webinars**
  http://www.ustream.tv/channel/mms-autism-webinar%20

- **Patrick Timpone interviews Kerri Rivera (8/15/2013)**
  http://oneradionetwork.com/health/kerri-rivera-healing-the-symptoms-known-as-autism-children-are-not-born-with-autism-so-they-should-not-die-with-it-august-15-2013/

- **Patrick Timpone interviews Kerri Rivera (11/18/2013)**
  http://oneradionetwork.com/health/kerri-rivera-encore-interview-healing-the-symptoms-known-as-autism-using-parasite-protocols-and-chlorine-dioxide-november-18-2013/

- **Bulletproof Executive Radio with Dave Asprey (11/27/2013)**
  http://www.youtube.com/watch?v=R7s7lYj2SCs

Appendix 17

Direct Help from Kerri Rivera

Consults

Since the release of the first edition of the book, we have found that when families decide to implement The Protocol some do not need further assistance beyond reading the information in this book, using our online support groups, YouTube videos, and attending conferences. However, there are some of us who need a more personalized approach and support to navigate the information, or would like ideas on dosing or timing.

For those of us who need or want that additional support, consults are perfect. We can work together to map out a plan for the parent/family and child, brainstorm through challenges, or just know that you will never walk alone. Whatever the case may be I am available if you need me. I am fluent in English and Spanish, and we can use a translator for other languages. We can speak over the phone or on Skype.

As time goes on, more and more practitioners will be able to support families while using this protocol. In the meantime, If you would like to set up a consult with me, please click on the "Consult Kerri" button on...

www.CDAutism.org

...and follow the instructions there.

> *Kerri Rivera's undivided attention + discussing your child + $100 = priceless. Tonight, instead of living it up in Mexico, Kerri spent her Friday evening on the phone doing a consult with me. Not only was the consult informative but also highly enjoyable. Kerri's energy is contagious, her vision boundless and her sense of humor edgy (just the way I like it). Kerri thank you for rocking our world! You are an amazing lady and my family is eternally grateful for your love of our children. Peeps, if you are new, stuck, confused or you want a better understanding of the protocol PLEASE, PLEASE, PLEASE take advantage of this opportunity. I've done a million consults over the years but this was by far my favorite.*

Note: A consult with Kerri does not constitute medical advice, nor is it a substitute for a consult with your doctor or healthcare provider. As always, you are responsible for your health, and for the health of your children.

might not be tonight. tomorrow or the next day.. but everything's gonna be ok
e.s
Healing the symptoms known as Autism
1st Edition
Kerri Rivera

References

Chapter 3 - Diet

1. Siri, Ken, and Tony Lyons. Cutting-Edge Therapies for Autism 2011-2012. New York: Skyhorse, 2011. Print.

2. Knivsberg, A.M., K.L. Reichelt, T. Høien, and M. Nødland. "A randomized, controlled study of dietary intervention in autistic syndromes. Nutritional Neuroscience 5.4 (2002) : 251-261. Print.

3. "Frequently Asked Questions." The GFCF Diet Intervention – Autism Diet. TheGFCDiet.com, n.d. Web. 19 Mar. 2013.

4. Reichelt, Karl L., Dag Tveiten, Anne-Mari Mari Knivsberg, and Gunnar Brønstad. "Peptides´ role in autism with emphasis on exorphins." Microbial Ecology in Health & Disease 23 (2012). Web. 11 May 2013.

5. Shattock, Paul, and Paul Whiteley. "Biochemical aspects in autism spectrum disorders: updating the opioid-excess theory and presenting new opportunities for biomedical internvention. Expert Opinions on Therapeutic Targets 6 (2002) : 175–183. Print.

6. Fasano, A. "Dr. Fasano on Leaky Gut Syndrome and Gluten Sensitivity. " Gluten Free Society, n.d. Web. 11 May 2013.

7. "How does gluten and casein relate to the problem of autism?" University of Florida. Department of Pediatrics Division of Genetics and Metabolism, n.d. Web. 11 May 2013.

8. Appleton, Nancy, and G.N. Jacobs. Suicide by Sugar A Startling Look at Our #1 National Addiction, New York: Square One, 2009. Print.

9. Tobacman, J.K. "Review of harmful gastrointestinal effects of carrageenan in animal experiments." Environmental Health Perspectives 109.10 (2001) : 983-994. Print.

10. McMillan, Tricia. "Does Splenda sugar substitute have chlorine in it?" Livestrong.com, Demand Media. 11 June 2011. Web. 11 May 2013.

11. "Food and Drugs." 21 CFR 102.22. 2012. Web. 11 May 2013.

12. Wallinga, David, Janelle Sorensen, Pooja Mottl, and Brian Yablon. Institute for Agriculture and Trade Policy. "Not So Sweet: Missing Mercury and High Fructose Corn Syrup." 2009. Web. 11 May 2013.

13. Gottschall, Elaine. "About the Diet." Breaking the Vicious Cycle, n.d. Web. 11 May 2013.

14. Gottschall, Elaine. "Autism & GI Problems." Breaking the Vicious Cycle. n.d. Web. 11 May 2013.

15. Reasoner, Jordan. "Phenols and Salicylates: What They Are and Why it Matters." SFK. Web. 2010. 11 May 2013.

16. Campbell-McBride, Natasha. Gut and Psychology Syndrome: Natural Treatment for Autism, Dyspraxia, A.D.D., Dyslexia, A.D.H.D., Depression, Schizophrenia. 2nd ed. Cambridge UK: Medinform. 2010. Print.

17. Sarah. "Don't Waste Your Time: Why the Candida Diet Doesn´t Work." The Healthy Home Economist. Austus Foods. 15 Nov. 2012. Web. 11 May 2013.

Chapter 4 - An Introduction to Chlorine Dioxide

1. Port, Tami. "Gram Negative (Gram-) Bacteria." Suite101. 27 July 2007. Web. 11 May 2013.

2. Toxicological Profile for Chlorine Dioxide and Chlorite. U.S. Department of Health and Human Services. Public Health Service. Agency for Toxic Substances and Disease Registry. September 2004. Pg 25. 3.2.2 Oral Exposure. 3.2.2.1 Death. Web. 11 May 2013.

Chapter 5 - Chlorine Dioxide (CD)

1. Lubbers, Judith R., Sudha Chauan, and Joseph R. Bianchine. "Controlled Clinical Evaluation of Chlorine Dioxide, Chlorite and Chlorate in Man." Environmental Health Perspectives 46 (1982) : 57-62. Print.

2. Ballantyne, Coco. "Strange but True: Drinking Too Much Water Can Kill You." Scientific American. 21 June 2007. Web. 11 May 2013.

3. "Woman died from too much water´." BBC News. 12 Dec. 2008. Web. 11 May 2013.

4. Humble, Jim. The Master Mineral Solution of the Third Millennium. 5th ed. Jim V. Humble 2011. Print.

5. Harman, D. "Aging: a theory based on free radical and radiation chemistry." Journal of Gerontology 11.3 (1956) : 298-300. Print.

6. Harman, D. "The biologic clock: the mitochondria?" Journal of the American Geriatrics Society 20 (1973) : 145–147. Print.

7. Watson, J. "Oxidants, antioxidants and the current incurability of metastatic cancers." Open Biology. 12 April 2013. Web. 11 May 2013.

8. Tarr, Peter. "Nobel laureate James Watson publishes novel hypothesis on curing late-stage cancers." Cold Spring Harbor Laboratory. 7 January 2013. Web. 11 May 2013.

9. Kilic, Eser, Süleyman Yazar, Recep Saraymen, and Hatice Ozbilge. "Serum malondialdehyde level in patients infected with Ascaris lumbricoides." World Journal of Gastroenterology 9 (2003) : 2332-2334. Print.

10. Kosenko, Elena, Yuri Kaminski, Oksana Lopata, Nikolay Muravyov, and Vicente Felipo. "Blocking NMDA receptors prevents the oxidative stress induced by acute ammonia intoxication." Free Radical Biology & Medicine 26 (1999) : 1369-1374. Print.

11. Lee, J.K., E.H. Choi, K.G. Lee, and H.S. Chun. "Alleviation of aflatoxin B1-induced oxidative stress in HepG2 cells by volatile extract from Allii Fistulosi Bulbus." Life Science 77 (2005) : 2896-2910. Print.

12. Nielsen, F., B.B. Mikkelsen, J.B. Nielsen, H.R. Andersen, and P. Grandjean. "Plasma malondialdehyde as biomarker for oxidative stress: reference interval and effects of life-style factors." Clinical Chemistry 43 (1997) : 1209-1214. Print.

13. Anwar, H., M.K. Dasgupta, and J.W. Costerton. "Testing the susceptibility of bacteria in biofilms to antibacterial agents." Antimicrobial Agents and Chemotherapy 34.11. (1990) : 2043-2046. Print.

14. Usman, Anju. AutismPedia. 12 May 2010. Web. 11 May 2013.

15. Budinger, Mary. "Lyme-Induced Autism Conference Focuses on Biofilm and Toxicity." Public Health Alert. 4.8. Publichealthalert.org. Aug. 2009. Web. 11 May 2013.

16. "Children With Autism Have Mitochondrial Dysfunction, Study Finds." Science News. 30 November 2010. Web. 11 May 2013.

17. MMS Autism Webinar - May 10, 2012, Andreas Kalcker & Kerri Rivera.

Chapter 8 - The Kalcker Parasite Protocol

1. Volinsky, Alex, A., Nikolai V. Gubarev, Galina M. Orlovskaya, and Elena V. Marchenko. "Developmental stages of the rope´ human intestinal parasite" Submitted to arxiv.org 13 Jan. 2013. Web. 11 May 2013.

2. United States. Department of Health and Human Services, Food and Drug Administration, Centre for Food Safety and Applied Nutrition. Fish and Fishery Products Hazards and Controls Guidance 4th Edition. Florida: Government Printing Office, 2011. Web. 11 May 2013.

3. Zam, S.G., Martin, W.E., and Thomas Jr., L.J. "In Vitro uptake of Co^{60} –vitamin B_{12} by *Ascarias suum."The Journal of Parasitology 49* (1963): 190-196. Print.

4. Vasudevan D.M., Sreekumari, S., and Vaidyanathan, K. "Textbook of BIOCHEMISTRY for Medical Students. 6th ed." *Section D: Nutrition*. New Delhi, 2011. 404. Print.

5. "Water related diseases; Ascariasis." Water Sanitation Health. WHO. 2013. Web. 11 May 2013.

6. "Mebendazole Dosage." Drugs.com. n.d. Web. 11 May 2013.

7. Wina, Peter. "Clay Uses – Parasite Removal – Yes, You Do Have Worms!" aboutclay.com. n.d. Web. May 13 2013.

8. "Parasites – Cysticercosis." Centers for Disease Control and Prevention. n.d. Web. 11 May 2013.

9. Nielsen, Flemming, Mikkelsen, Bo Borg, Nielsen, Jesper Bo, Andersen, Helle Raun, and Grandjean, Philippe. "Plasma malondialdehyde as biomarker for oxidative stress: reference interval and effects of life-style factors." Clinical Chemistry 43 (1997) : 1209-1214. Print.

Chapter 9 - Other Supplements

1. "What are probiotics?" Theralac. Master Supplements. n.d. Web. 11 May 2013.
2. "Frequently Asked Questions About Theralac." Purest Colloids. 2013. Web. 11 May 2013.
3. "Autism." The Probiotic Revolution. Freedom. n.d. Web. 11 May 2013.
4. Messaoudi, Michaël, Robert Lalonde, Nicolas Violle, Hervé Javelot, Didier Desor, Amine Nejdi, Jean-François Bisson, Catherine Rougeot, Matthieu Pichelin, Murielle Cazaubiel, and Jean-Marc Cazaubiel. "Assessment of psychotropic-like properties of a probiotic formulation (Lactobacillus helveticus R0052 and Bifidobacterium longum R0175) in rats and human subjects." British Journal of Nutrition 105 (2011) : 755–764. Print.
5. Silk, D. B. A., A. Davis, J. Vulevic, G. Tzortzis, and G.R. Gibson. "Clinical trial: the effects of a trans-galactooligosaccharide prebiotic on faecal microbiota and symptoms in irritable bowel syndrome." Alimentary Pharmacology & Therapeutics 29 (2009) : 508–518. Print
6. "Omega-3 fatty acids." University of Maryland Medical Center. A.D.A.M. 2011. Web. 11 May 2013.
7. "Omega-6 fatty acids." University of Maryland Medical Center. A.D.A.M. 2011. Web. 11 May 2013.
8. Bell, J.G., E.E. MacKinlay, J.R. Dick, D.J. MacDonald, R.M. Boyle, and A.C.A. Glen. "Essential Fatty Acids and Phospholipase A2 in Autistic Spectrum Disorders." Prostaglandins Leukotrienes and Essential Fatty Acids 71.4 (2004) : 201-204. Print.
9. Richardson, A.J., and P. Montgomery. "The Oxford-Durham Study: A Randomized, Controlled Trial of Dietary Supplementation with Fatty Acids in Children with Developmental Coordination Disorder." Pediatrics 115.5 (2005) : 1360-1366. Print.
10. "GABA." Healing Autism & ADHD. n.p. n.d. Web. 11 May 2013.
11. English, Jim. "Saving Eli: One Family´s Struggle with Autism." Jim English. 2012. Web. 11 May 2013.
12. "L-Theanine Information." L-Theanine. n.d. Web. 11 May 2013.
13. "Pycnogenol." Horphag. n.d. Web. 11 May 2013.
14. "1 in 50 US. Kids Have Autism." Autism Coach. n.d. Web. 11 May 2013.
15. Chez, Michael G., Cathleen P Buchanan, Mary C. Aimonovitch, Marina Becker, Karla Schaefer, Carter Black, and Jamie Komen. "Double-Blind, Placebo-Controlled Study of L-Carnosine Supplementation in Children With Autistic Spectrum Disorders." Journal of Child Neurology 17 (2002) : 833-837. Print.
16. Smith, Leonard. "Are You Dangerously Deficient in Taurine?" drlwilson. n.d. Web. 11 May 2013.
17. "DMG supplement." Healthcare Information Directory. iHealth Directory. n.d. Web. 11 May 2013.
18. "Dimethylglycine (DMG) for Autism." Autism Research Institute. n.d. Web. 11 May 2013.

19. "Find a Vitamin or Supplement Betaine Anyhydrous." WebMD. Therapuetic Research Faculty. 2009. Web. 11 May 2013.

20. "Nutritional DMG/TMG Treatment Overview." Autism Canada Foundation. 2011. Web. 11 May 2013.

Chapter 10 - Chelation

1. "Why Cleanse Metals." Baseline of Health Foundation. n.d. Web. 11 May 2013.

2. Klinghardt, Dietrich. "D: Amalgam/mercury detox as a treatment for chronic viral, bacterial and fungal illnesses." Explore 8 (1997) : 13-16. Print.

3. Price, Weston A. Nutrition and Physical Degeneration. 8th ed. Lemon Grove, CA: Price Pottenger Nutrition, 2008. Print.

4. "Bentonite clay for internal healing." Healing Daily. 2005. Web. 11 May 2013.

5. Clinical Evaluation of Bio-Chelate™." Bio-Chelat™. 09 Dec. 1987. Web. May 11, 2013.

Chapter 11 - Hyperbarics

1. University of Pennsylvania School of Medicine. "Penn Study Finds Hyperbaric Oxygen Treatments Mobilize Stem Cells." ScienceDaily. 28 Dec. 2005. Web. 11 May 2013.

Chapter 12 - GcMAF & Autism

1. Bradstreet, James Jeffrey, Vogelaar Emar, and Thyer Lynda. "Initial Observations of Elevated Alpha-N-Acetylgalactosaminidase Activity Associated with Autism and Observed Reductions from GC Protein—Macrophage Activating Factor Injections." Original Research. "Autism Insights 4 (2012): 31-38. Print.

Chapter 14 - Miscellaneous Information You Should Know

1. Garcia, H.H, and M. Modi. Helminthic parasites and seizures. Epilepsia 49 Suppl 6. (2008) : 25-32. Print.

2. "Self-Injurious Behavior." Autism Research Institute. n.d. Web. 11 May 2013.

3. PANDAS-PITAND Awareness & Research Support. 2011. Web. May 11 2013.

4. Jenike, Michael, and Susan Dailey. International OCD Foundation. 2012. Web. May 11 2013.

5. Jenike, Michael, and Susan Dailey. "Sudden and Severe Onset OCD – Practical Advice for Practitioners and Parents." Beyond OCD. 2013. Web. 11 May 2013.

6. Bock, Kenneth and Nellie Sabin. The road to Immunity: How to Survive in a Toxic World. New York: Pocket Books, 1997. Print.

7. "Fever." AskDrSears. 2013. Web. 11 May 2013.

8. "NINDS Febrile Seizures Information Page." National Institute of Neurological Disorders and Stroke. 2013. Web. 11 May 2013.

9. Gilbère, Gloria. "Skin & Intestinal Disorders." Leaky Gut Syndrome. Dawn Mellowship. 2005. Web. 11 May 2013.

10. Edward, F. "14 Common Symptoms of Parasites." Parasite Cleanse Resource Center. 10 June 2009. Web. 11 May 2013.

11. "Body Burden: The Pollution in Newborns. A Benchmark Investigation of Industrial Chemicals, Pollutants and Pesticides in Umbilical Cord Blood." Environmental Working Group. 14 July 2005. Web. 11 May 2013.

Appendix 5 - Molecular Mimicry

1. Fasano A, Berti I, Gerarduzzi T, Not T, Colletti RB, Drago S, Elitzur Y, Green PH, Guandalini S, Hill ID, Pietzak M, Ventura A, Thorpe M, Kryszak D, Fornaroli F, Wasserman SS, Murray JA, Horvath K. "Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicentre study." Arch Intern Med. 163.3 (2003): 286-292. Print.

2. Vojdani A, O'Bryan T, Kellermann GH. "The immunology of gluten sensitivity beyond the intestinal tract." European Journal of Inflammation." 6.2 (2008) : 0-0.

3. Vojdani A., O'Bryan T, Kellermann GH. "The Immunology of Immediate and Delayed Hypersensitivity Reaction to gluten." European Journal of Inflammation Vol 6.1, (2008) :1-10

4. Vojdani A, O'Bryan T, Green JA, McCandless J, Woeller KN, Vojdani E, Nourian AA, Cooper EL. "Immune Response to Dietary Proteins, Gliadin and Cerebellar Peptides in Children with Autism." Nutr Neurosci. 7.3 (2004) : 151-161. Print.

5. Ford RP. "The gluten syndrome: a neurological disease. Med Hypotheses." 73.3 (2009) : 438-440. Print.

6. Ford, RP. The Gluten Syndrome: Is Wheat Causing You Harm. Dr. Rodney Ford MD MBBS FRACP Smashwords 2011 ebook

7. Hadjivassiliou M, Sanders DS, Grunewald RA, Woodroofe N, Boscolo S, Aeschlimann D., "Gluten sensitivity: from gut to brain." Lancet Neurol. 9.3(2010):318-30.

8. Solaymani-Dodaran M, West J. Logan RF. Long-term mortality in people with celiac disease diagnosed in childhood compared with adulthood: a population-based cohort study. Am J Gastroenterol. 102.4 (2007): 864-870. Print.

9. Fedoroff, Nina V, and Nancy Marie Brown. Mendel in the Kitchen: A Scientist's View of Genetically Modified Food. Washington, DC: Joseph Henry Press, 2006. Print.

10. Corrao G, Corazza GR, Bagnardi V, Brusco G, Ciacci C, Cottone M, Guidetti CS, Usai P, Cesari P, Pelli MA, Loperfido S, Volta U, Calabró A, Certo M. "Mortality in patients with coeliac disease and their relatives: a cohort study." Lancet. 358.9279 (2001) : 356-361. Print

Index

5-Hydroxytryptophan 249
  dosing 249
5-N-ethyl-glutamine. *See* L-theanine
7 Day Fridge MMS 158
72/2 protocol 115, 126, 128

A

ABA. *See* Applied Behavior Analysis (ABA)
absorption
  proper 60
abstract thinking 250
Accidental Cure by Optimists (article) 443
ACE. *See* Autism Center of Excellence
acetic acid 52
acetylcholine 247
acid exposure 118
acidic activator 93, 96, 105, 110, 123, 146
  citric acid 90, 93, 94, 146, 147. *See also* citric acid
  hydrochloric acid 93
  lemon juice 93
  lime juice 93
acidity in the body 236
acne 86, 247, 337, 338
actifolate 342
activated charcoal 143
activator 78, 93. *See* acidic activator
  buying 94
acupuncture 445
ADD 64, 352
additives, food 61
ADHD 64, 268, 347, 352, 367, 393, 425
affection 229
African sleeping sickness 176
agave syrup 52
aggression 62, 103, 133, 325
aging
  premature 485
Aginsky, Dr. Bernard 277, 278
agitation 20
AIDS 85, 138, 296, 298, 300, 315
air freshener 343
air quality 506
ALA 122, 133, 138
albendazole 327, 328, 329
Albenza®. *See* albendazole
alcohol 240, 343, 508
Alecson, Deborah Golden 73
alertness 248, 250
alfalfa 486
algae 485
Alinia® 327
allergens 50, 488
allergic reactions 50, 55, 66, 451
  emergency 455
allergies xxi, xxiii, 10, 45, 49, 56, 57, 69, 70, 71, 73, 128, 137, 181, 247, 256, 338, 345, 444, 508
  food 338
allergy panel 56
allergy-type symptoms 57
almond milk 134
almonds 51, 63
alpha-gliadin. *See* gliadorphin
alpha-N-acetylgalactosaminidase (Nagalase). *See* Nagalase
aluminum 255, 256, 487, 506, 507
Alzheimer's Disease 85, 120, 247, 251, 256, 300, 445, 487
amalgams 509
amaranth 51
American Dairy Association 71

American Journal of Physiology 290
amino acids 48, 69, 245, 248, 249, 250, 251, 252, 450, 458
amino acid sequences 452, 455
ammonia 88, 176, 229, 236, 323
  high 182
amoebas 85, 190, 233, 238
amyotrophic lateral sclerosis 316
anal itching 179, 181, 220
anemia 176, 182, 314
anger 34, 178
angiogenesis 299
anorexia 179, 333
antacids 257
anthelmintic, broad spectrum 192
antibacterial agent 508
antibiotics 11, 75, 82, 118, 119, 244, 246, 327, 329, 333
  use 1
antibodies 50, 451, 452, 455, 457
  production of 335
  tests for 66
antibody panel 463
antibody-specific immunoglobulin E 49
anticancer agent 314
antifungals 11, 82, 327
anti-inflammatory 250, 485
    antimicrobial agents 118
antioxidant foods 123
antioxidants 45, 86, 87, 88, 124, 127, 132, 138, 139, 143, 250
anti-parasitic medications 326, 328
anti-seizure 327
  medications 251
antiseptic, nontoxic 85
antivirals 11, 82, 327, 329
anxiety 20, 54, 103, 136, 178, 179, 246, 248, 325, 333, 337, 340, 444
anxiety, separation 228
aortal aneurysms 456
appetite 249, 330
appetite, ravenous 178
apple juice 73
apples 51, 57, 61, 63, 76, 133
Applied Behavior Analysis (ABA) 4, 43, 74, 344
Apraxia 15
ARI. *See* Autism Research Institute
Arranga, Teri 13
arsenic 78, 257, 506, 508
arterial wall, strength of 485
arthritis 73, 137, 169, 179, 247, 445
  rheumatoid 144, 443
artichokes 486
artificial colors 49
artificial flavoring 47, 52
artificial sweeteners 52, 55
Ascaris lumbricoides (roundworm) 12, 165, 167, 174, 183, 187, 233, 236, 240
  longevity 174
ASD 64
asparagus 486
Asperger's 347
aspirin 61
asthma 114, 180, 256, 508
ATEC xxi, xxiii, 46, 47, 136, 140, 141, 319, 339
  not going down 136
  statistics 121
athletic performance 252
atrazine 508
autism 44
  adults with 339
  avoiding 505
  curing 303
  environmental factors 506
  epidemic 505
  regressive 236, 347
  statistics viii, xxv
Autism Canada Foundation 253
Autism Center of Excellence 274, 275, 276
AutismO2 Clinic 7
AutismOne (conference) 5, 13, 14, 24, 444
AutismPedia 119

Autism Research Institute 6, 7, 276, 278
Autism Society of America 2
Autism Treatment Evaluation Checklist. *See* ATEC
autoimmune 462
autoimmune deficiencies 169
autoimmune disease 295, 453, 460
autoimmune gluten reaction. 453
autoimmune reactivity 453
autoimmunity 452, 453
autonomic nervous system 456
avocados 63, 460
azadirachta indica. *See* neem

B

Babesiosis 176
baby 505, 509
baby bottle 98, 109, 493
    Lifefactory.com 95
    method 90
bacteria xxi, xxiii, 10, 11, 52, 59, 70, 76, 78, 81, 82, 85, 118, 140, 162, 225, 244, 258, 261, 334, 345, 488
    beneficial 82
        imbalance of 245
    growth 59
    harmful 61
    healthy 231
    intestinal 485
    load 84
    negatively charged 78
        CD kills 78
    overgrowth 61
    positively charged 78
bad breath 182
bagels 67, 457
baking soda 124, 343, 490
balance 274
bananas 57, 63, 73, 76
banging 1
bark 55
barley 49, 66, 69, 455, 486
barometer 280
barometric pressure 280
basal ganglia damage 334
basophils 50
bath 131. *See also* salt baths; *See also* CD baths
bathtub 114
B-Complex 342. *See also* vitamin B
beans 52, 252
    navy 52
beauty products 255
bedwetting 181, 333
beef 50
bee pollen 457
bee stings 66, 451. *See also* bee venom
beets 253, 486
bee venom 455
behavior 228, 247, 248, 324
    changes 253, 256
    problems 20, 70
    undesirable
        probiotics, remove 226
behavioral
    reactions 62
bell peppers 486
bentonite clay 140, 143, 257, 258
    baths 257, 259
        schedule 259
Bernie. *See* Rimland, Dr. Bernard
berries 61, 123, 133
beta-alanine 250
beta amyloid plaque 251
beta-carotene 87
betaine anhydrous. *See* trimethylglycine (TMG)
bifidobacterium 244
bile 231
Bio-Chelat™ 140, 257, 258, 259, 320
biofilm xxiii, 11, 45, 103, 111, 117, 118, 119, 129, 130, 143, 165, 226, 235, 257, 261, 444
    appearance 130
    protocol 11
    what is 118
    why CD works on 119
bioflavonoids 61
biomedical intervention 341

biomedical protocols 347
biomedical treatments 11, 140
biomed movement 334
biopsy 454, 455
bipolar disorder 120
birth-control patches 73
bisphenol A 508
biting 332
  hand 331
black pit depression 458
black walnut 226
bladder infections 244
blank chart 198
bleach 79, 84
blisters 89, 338
blitz day. *See* mebendazole
bloating 51, 89, 177
blood 50, 85, 253, 299, 310, 507
blood analysis 182
blood-brain barrier 10, 48, 49, 451, 453
blood cells 251
blood disorders 176
blood flow
  changes in 458
blood panels 461
blood parasites 176
blood pressure, high 485
blood stream 48, 60, 75, 103, 145, 337
blood tests 67, 70, 243, 261, 454
blood vessels 250, 453
blood vessel walls 461
blueberries 87
Bobadilla, Ana 513
Bock, Dr. Kenneth 335
body language 325
body odor 182
body products
  non-toxic 509
bologna 50
bone 65, 253, 255, 259, 261, 453, 461, 485
  disorders 300
  health 487
  marrow 85
  metabolism 300
bone broth 459
borax 343
boron 116, 486
bottle
  wine 162
bouillon cubes 52
bovine respiratory syndrome 145
bowel movements 28, 42, 46, 105, 260
  excessive 177
bowel permeability 48
Bowen, Dr. James 55
Bradfield, Dr. 269, 270, 272
Bradstreet, Dr. 302, 303, 306, 308, 310
brain 49, 65, 248, 249, 250, 251, 253, 255, 259, 295, 305, 307, 310, 320, 323, 344, 453, 461, 487, 507
  connectivity, increased 116
  development 246
  fog 178
  function 247
bread 4, 49, 52, 71
Breaking the Vicious Cycle (book) 60
break the diet 56
breastfeeding 317, 490
breast milk 507
breathing problems 70, 182
bromine 116
bronchiectasis 443
bronchitis 114, 180
brownies 67, 457
brown rice 486
Bruxism. *See* teeth grinding
buckwheat 51
bud 55
Budinger, Mary 119
burbur 143
burns 85, 474
burping 89, 178
butter 49, 72, 78

C

C6H8O7. *See* citric acid
cabbage juice 232
cacao 63, 123
Cade, Dr. 69
cadmium 255, 507
caffeine 134, 250
cake 49
calamus root 231
calcium 45, 90, 116, 118, 119, 236, 486, 487
calcium silicates 190
Calm PRT 342
calves 145
Campbell-McBride MD, Dr. Natasha 75
cancer 54, 55, 87, 121, 295, 296, 298, 299, 300, 316, 335, 444, 445, 456, 507, 508
  bladder 443
  breast 297
    inhibition of 299
  colon 54
  immunotherapy of 297
  metastatic 87
candida xxi, xxiii, 10, 11, 49, 52, 54, 57, 75, 81, 84, 118, 180, 244, 261, 337
  antibiotics 75
  c-section 75
  overgrowth 75, 76
candy 52, 54
canker sores 86
cantaloupes 63
CaprilexTM 342
carbohydrates 59, 60, 75, 119
  complex 59, 60
  digestion 59
  malabsorption 60
carbon 55
carbon dioxide 282
carbon monoxide 508
carcinogenic 55
cardiovascular 246
  problems 508
caricin 233
carnitine 248
carnosine 250
carotene 61
carotenoids 61
carrageenan 52, 54
Carrasco Maceda, Miriam 13, 166, 235
carrots 133, 234
cascara sagrada 240
casein 44, 48, 49, 53, 54, 56, 58, 69, 70, 72, 245
caseinate 72
casein-free 4, 41, 45, 48, 49, 57, 58, 70, 72, 74, 513
cashews 51
casomorphin 44, 49, 56, 58, 69, 245
  peptide 69
castor oil 129, 188, 194, 233, 237, 341
  source & cost 195
catsup 52, 54
cattle 145
cayenne 234
CD 9, 10, 11, 42, 77, 78, 81, 83, 84, 89, 90, 103, 111, 114, 115, 121, 145, 188, 244, 319, 336, 445, 446. *See also* sodium chlorite
  72/2 protocol 115, 126, 128
  activation time 110, 124
  activator 78
  antioxidants 124
  appetite 134
  aqueous solution 84
  ATEC 136
  babies 131
  back-to-back baby bottles 127
  baths 113, 114, 125, 131, 187, 189, 229, 260, 319, 336
  best bottles to store 132
  bottle caps 132
  bottles 135
  bottle, steel 132
  breast milk 131
  buying 82
  capsules 132
  chelator compatibility 140
  chia seeds 134
  child refusing dosing 135
  Controlled Clinical Evaluations of

  Chlorine Dioxide, Chlorite and Chlorate in Man 84
  controversy xiii
  critics 84
  Davy, Sir Hymphrey 77
  deaths 78
  diarrhea 89, 141
  die-off reaction 138
  diet requirements 132
  discomfort 127
  DOs & DON'Ts 122
  doses, missed 136
  dosing 102, 187
  dosing schedule 135
  dosing too low 125
  double doses 103, 125
  double dosing 103, 125
  drug interactions 143
  ears 11
  emergency 142
  enemas 12, 125, 187, 189, 226
  eyes 11
  false information/propaganda 82
  first dose 127
  flying 142
  food 124
  foods to avoid 133
  foot baths 114
  forum 83
  full oral dose 125, 126, 128
    going past 127
  green tea 134
  hair 11
  harmful? 143
  health 84
  heavy metals 139
  humidifier 225
  industrial bleaching processes 84
  industrial purposes 85
  inhaled 11
  internal consumption 84
  intervention 81
  introduction to 77
  killing parasites 238
  kombucha 134
  life expectancy of 124
  maintenance dosing 238
  Master Mineral Supplement 77
  mercury 141
  molecule 77
  multivitamins 132
  nausea 89
  nebulizer 138
  neutralizers 127
  night dosing 115
  Nutriiveda 134
  oral protocol 11, 100
  overdose 142
  patents 77, 85
  pH 145
  poison index 78
  pregnancy 131
  prescription contraindications 143
  protocol xxiv
  rectal 11
  school 126, 135
  school field trip, dosing on 135
  serving cup dilution 102
  side effects 81, 89
  skin 11
  smell, chlorine like 98, 493
  sodium chlorite 78
  steam bath 114
  stevia 126
  strength 467
  sunlight 124
  supplements 138
  taste 135, 225
  testing 124
  test strips 133, 225
  toxicity 78
  toxic rating 78
  toxic residues, no 82
  troubleshooting checklist 123
  users of 78
  vaginal 11
  vomiting 89
  water 125, 132
  water temperature 132
  weight gain/loss 134
  yellow color 123
CD4 315
CDH 14, 155, 156, 158, 225, 226
  baths 162
  Easy-Peasy method of making 163
  Formulation Table 161
  making, detailed instructions 159
  strength 467
  taste 163
  testing 467
  using 162
CDS 14, 145, 146, 147, 157, 188
  buying 147
  CDS to CD transition table 154
  color of solution 151, 152
  concentration strength 148
  deteriorate plastic (avoiding) 150

- detoxing with 153
- distillation process 145, 147
- enemas 154
- hose, making with a 147, 148
- instructions on making 150
- making 147
  - New Method, with 148
  - Overnight Method, with 148
  - Shot Glass Method, with 148
- metal (avoid) 150
- overdose 146
- Overnight CDS Generating Kit 150, 151
- sleep 146
- sodium chlorite, added back in 146
- stabilized 146
- strength 467
- switching back to CD 153
- testing 467
- time it takes to make 151
- unstabilized 146
- waking time 146

CEA. *See* Curando el Autismo Foundation
celery 133
celiac disease 44, 59, 65, 69, 454, 456, 461
- celiac antibody tests 457
- villi damaged 449

cell-cell interactions 118
cellular metabolism 286
Centers for Disease Control (CDC) 508
central nervous system 62, 176
cereal 49, 71, 253
cerebella ataxia 456
cerebral cortex 309
cerebral palsy 273
cerebrospinal fluid 309
CFS/ME 305, 311, 316
- aetiology 301

chagas disease 176
chanca piedra 191
- dosing 191
- source & cost 191

chart
- Estimated Full Oral CD Doses by Weight 101, 494
- hourly dose / mimimum water dilution 102

Chatelier's Principle 284, 285
cheese 4, 49, 71, 72
cheese puffs 43
chelation 5, 43, 139, 141, 255, 257, 259, 260, 320
- IV 261
- what is 257

chelator xxiv, 119, 257, 258
chemical mutation 457
chemicals 50, 55, 174, 506
- absorption through skin 508

Cheney, Dr. Paul 301
cherries 63
chest pains 179
chia seeds 134
chicken 50, 53, 72
children
- growth spurt 239
- nutritional shakes for 53
- older 140, 155

chili pepper 63
- powdered 63

chills 89
chlorination 83
chlorine bleach 55, 83, 84, 116, 142
chlorine dioxide. *See also* CD
- antioxidants 88
- patents 85
- testing 467

chlorine dioxide gas 147, 149
Chlorine Dioxide Holding (Solution). *See* CDH
chlorine dioxide photometer 158
Chlorine Dioxide Solution (CDS). *See* CDS
chlorite molecule 83
chlorocarbon 55
chocolate 61, 66, 123, 133, 325, 455, 460
- milk 52

cholesterol 509
- high 485

choline 507
choucroute 232
chromium 116

chronic fatigue syndrome 62, 69, 169, 177, 247, 272, 340
chronic illness 256
chronic immune reactions 167
cider vinegar 63
cinnamon 234
citric acid 96, 98, 155, 493
citrus 123, 124, 133, 139
clay 257
cleaning products 508
  harmful 509
clear zone 228
ClO2. *See* CD
clonazepam 42
clorox 78
clostridia 59
clostridium 246
clothes
  sensitivity to 333
cloves 231, 234
coal power plants 255
cobalt 116
coccidiosis 145
cocoa 61, 123
coconut 51, 134, 460
coconut milk 123
coconut oil 63
coconut water 123, 134, 232
cod liver oil 124, 342
coffee 61, 66, 123, 455
cognition 247, 249
cognitive
  abilities 252
  decline 120
  impairment 328
  problems 508
cold 157
  internally 177
cold cuts 50, 53, 55
colds 114
Coley, Dr. William 296, 300
colitis 177
collagen 250, 488
colloidal silver 122, 142
colon 59, 60, 103, 104, 130, 338, 509
  cancer 54
  hydrotherapy 112
colonics 104, 111, 130, 226
  hydrotherapist 130
colored fruits 63
coloring 41
Combantrin®. *See* pyrantel pamoate
combativeness 236
Combined Immune Deficiency 328
comprehension 28
concentration 30, 247
condiments 234
confused thinking 178
congestion 114
  sinus 137
constipation 43, 58, 104, 125, 128, 129, 177, 179, 260, 340
consults with Kerri Rivera 521
cookware 255
coordination 21
  lack of 178
  poor 236
copper 116, 342, 486, 509
CoQ10 122, 127, 138, 342
cork 160, 162
corn 51, 57, 66, 69, 70, 74, 455, 459
corn flakes 52
corn syrup. *See* high fructose corn syrup (HFCS)
cosmetics 257, 489
  absorption of 508
cost of protocol (estimated) 341
cough 114, 180, 182, 340
Courchesne, Dr. Eric 274, 275. *See also* Autism Center of Excellence
Courchesne, Dr. Rachel 275, 277. *See also* Autism Center of Excellence
couscous 49
cow's milk. *See* milk, cow's
cramps 177, 179

crankiness 54

cravings 177, 178, 249

cream 49, 50

cream cheese 72

creams 508

crohn's disease 59

croissants 67, 457

Crook, William 72. *See* Solving the Puzzle of Your Hard to Raise Child (book)

Cruise, Tom (actor) 276

crust 460

crying 332, 458

crystalline silica 489

C-section 75

cucumber 133

Curando el Autismo Foundation vii, 305, 404, 517

curcumin 133

curezone.com 339

cyanide 78

Cyrex Laboratories 450, 454, 463, 464

Cysticercosis 240

cystic fibrosis 59

cysts 145, 179, 180, 240

CytofloraTM 342

cytokines 50, 335

D

dairy 4, 55, 69, 72, 74, 134, 231, 319. *See also* milk

dairy-free 460

Dalton's Universal Gas Laws 284

DAN!. *See* Defeat Autism Now! (DAN!)

dandelion 231
  tea 227, 229

dates 63

Davy, Sir Humphrey 77, 85

daycare 76

DDT 508

DE. *See* diatomaceous earth

Defeat Autism Now! (DAN!) xxii, xxv, 4, 6, 7, 8, 9, 74, 243, 252, 265, 267, 276, 278, 279, 332
  estimated costs 342

defeatist thoughts 48

deficiencies 116
  iron 180
  vitamin B12 180

deglycosylates 298

dehydration 118

dementia 85, 247, 445

dental amalgams 255. *See also* mercury

dental problems 445

deodorant 255

Department of Health & Human Services (DHHS) 78

depression 69, 178, 248, 325, 340, 443, 444, 456, 458, 464

dermatitis 181, 338

dessert 51

detergent 337
  clothing 509
  dishwashing 509

detox 89, 103, 127, 257, 340

detoxification 62, 256, 505

detoxification program 509

detoxing 115

detoxing program 320

detox organs 176

detox pathways 55

detox reaction 339

developing fetus 235

developmental delay 456

Developmental - Individual Difference, Relationship-based Model (DIR®) 344

deworming 166, 231
  importance of lifelong 186

diabetes 54, 247, 251, 256, 313, 445, 507, 508
  type 1 64
  Type II 54

diapers 227

diarrhea 11, 43, 57, 58, 59, 62, 89, 141, 142, 145, 157, 177, 179, 244, 340

diatomaceous earth 143, 188, 190, 226, 237, 238, 341, 485
  calcined 489
  deposits of 486
  dosing 190
  food grade 489
  medication, affect on 238
  source & cost 190
  sources 490
diatomite. *See* diatomaceous earth
diatoms 190, 485
die-off reaction 89
diet xxi, 4, 11, 56, 74, 506
  preventative 231
dietary intervention 42
Diet, The xxiv, 41, 42, 43, 44, 45, 46, 47, 50, 53, 57, 90, 244, 319, 462, 513
digestion 49, 51, 59, 69, 245
  diseases of 61
  enzymes 237
  poor 48
  problems, chronic 245
  system 49, 60
digestive system 454
dimethylglycine. *See* DMG
dimethyl sulfoxide 144
disaccharidase 75
disaccharides 59, 75, 76
discomfort 51, 70, 127
dishwashing detergent 509
dishwashing soap 343
disinfectants 61, 343
disorientation 178
distant 58
distended belly 177, 330
distillate 55
distractible 333
Divers Alert Network xxv
diverticulitis 59
divorce rate 341
dizziness 176, 177, 458
DMG 252, 253, 324, 342
  dosing 252
DMPS 6, 140, 257, 507
  drops 5
DMSA 140, 257
DMSO 137, 144
  application 144
DNA 120, 256, 304
DNA testing 165
Don't Drink Your Milk (book) 72
donuts 49
double-blind placebo xi, 68
double sugars 75
douche bag 108
dough 73
DPP IV enzymes 460
DPT (diphtheria-pertussis-tetanus) 1, 265
drinking water 255, 320
  fluoride 256
drooling 1
  while sleeping 181
drowsiness 180
drugs 295, 489
dry cleaning 509
duodenum 184
dysbiosis 103, 245
dyspnea 180
dyspraxia 247

E

ear
  infections 70, 244, 330
  problems 145
Early Autism Center (Toronto, Canada) 4
ears
  red 62
eating, high volume 228
EBN® Cleansing Clay 258
e-coli 485
ecosystem 246
eczema 181, 247, 337, 338
Edelson Dr. Stephen M. 253, 331
EDTA 119, 140, 257, 258
egg 455
eggs 50, 55, 66, 72, 325, 459
  hard boiled 76
elastin 488

electron shells 11
electrosmog 451
elimination diet 464
ELISA blood test 70
embryo
  chemicals that affect 508
emotion 245
  extreme high 62
  extreme low 62
emotional confusion 229
emotional frigidity hypothesis 276
emotional instability 464
endocrine problems 508
endorphins 248
enema 12, 26, 85, 88, 103, 104, 105, 108, 113, 115, 119, 128, 129, 130, 134, 162, 165, 220, 226, 237, 239, 260, 319, 340
  administering 109
  black specks 131
  bucket 106, 108
  bulb 107, 108, 112, 220
  catheter 108, 110, 111, 130
  CDS 154
  clean up 130
  colonic process, similar to 111
  comments & suggestions from parents 112
  constipation 129
  diarrhea 142
  distractions 113
  DMSO 144
  douche bag 106, 108
  enema bucket 108
  enema bulb 108, 112
  Everclear grain alcohol 130
  explaining to child 110
  fill & release 111
  fleet enema 107, 108
  fleet enema method 112
  frequency 105
  gravity bag 106, 108, 111, 129
  heating water 110
  history 104
  hose 110
  hot water bottle 106, 108
  how far to insert 130
  how to fill the syringes 111
  kids ask for 129
  kits 105, 108
  location 109
  lubricant 105, 110
  mucous 128
  oil 110, 111
  one mom's enema guidelines 113
  options 106
  position 110
  sphincter muscle 111
  stigma 104
  syringes 107, 108, 129, 130
  syringes & catheter 108
  teens & adults 129
  temperature 109
  timing 105, 109, 113, 130
  water 128
    amount of 109
    mixture 110
  when to start 128
energy 60, 147, 156, 247, 256
  lack of 121
  low 177
  production 247
EnhansaTM 342
enterobius vermicularis 185
enterocytes 75
Enterolab 450, 463, 465
enteropathy 454
environment
  beyond thinking green 508
  living 509
  working 509
environmental factors 507
environmental stress 453
environmental toxins 507
enzymes 45, 48, 49, 56, 59, 232, 244, 245, 342, 460
  deficiency 69
enzymolysis 55
eosinophil 338
EPA (Environmental Protection Agency) 255
epigenetics 507
epilepsy 328
epileptic activity 250
epoxy resin 508
Epsom salts 143, 226
equisetum hymale 457
erectile dysfunction 180

eruptions 181
essence or extractive 55
essential fatty acid 247
essential nutrients 178
essential oil 55
excitotoxins 58
exercise 54, 121
  intolerance 120
exhaustion 458
exorphins 245
extracellular polymeric substance 118
eye contact xxii, 42, 46, 58, 247, 252, 458
eyes 137, 253
  dark circles under 62
  degenerative conditions 251
  poking 331

F

face
  red 62
Facebook xi, 14, 16, 47, 76, 121, 122, 126, 136, 163, 347, 350, 356, 517
facial expression 325
failure to thrive 328
families
  advice to 341
farm animals 490
farmers 490
Farmer's Almanac 325, 477
Fasano, Dr. Alesio 62, 449
fat cells
  releasing toxins from 335
fatigue 75, 89, 176
fats 55
  bad
    destroying 485
fatty acids 246, 311
fatty acids, essential 509
FDA (Food and Drug Administration) 86, 249, 257, 490
  food code 174
fears 136
Federoff, Nina 67, 457
feet 115
Feingold Diet 61
Feingold, Dr. Ben 61
Feingold, Dr. Giuseppina 8
fermentation 55, 121, 458
  digestion 145
  food, of 51
fever 177, 182, 335, 336
  protocol 473
  reducers 335
  temperature 336
  therapy 335
fever therapy 336
fiber 74
fibroids 180
fibromyalgia 69, 169, 179, 340
financial matters 341
Fine, Dr. Kenneth 450, 463
fish 50, 251
fish oil supplements 124
flapping xxii, 1, 6, 179, 236
flatulence 89
flat worm 167
flavorings 49, 55
floor cleaner 343
Floortime Therapy 344
flora 76
flora balance 59
flour 49, 51, 52, 231
flu 114, 452
fluid retention 180, 182
flukes 165
fluoride 142, 256
fluorine 116
flushing 89
flu vaccine 507
  injury 331
focus 249, 250, 333, 444
folic acid 507
folinic acid 509
food
  acids 93
  addiction 71
  additives 61
  allergy 49, 50, 57, 70, 81, 244
  choices 508

- coloring 52, 55, 61, 63, 325
  - fast 51
  - intolerance 69
  - mercury in 507
  - nutrient dense 457
  - preventative 231
  - processed 41, 231
  - pyramid 72
  - raw 46
  - reaction to 70
  - supply 255
  - to be avoided 133
- food bars 460
- foot baths. *See* CD, foot baths
- Ford, Dr. Rodney 450, 456
- foreword vii
- forgetfulness 178
- formaldehyde 88, 236, 507
- fossilized shells 485
- fossil shell flour 486, 489
- fractures 487
- Franklin, Benjamin 279, 280
- Frank Oski. *See* Don't Drink Your Milk (book)
- frazzled 46
- free radical molecules 87
- Free Radical Theory of Aging (article) 86, 87
- french fries 44, 51
- french maritime pine bark extract 250
- fructose 231
- fruit 41, 43, 50, 51, 53, 55, 60, 61, 73, 76, 127, 139, 459
  - beverages 53
  - canned 50
  - dried 50, 76
  - frozen 50
  - juice 51, 55, 123, 124, 133, 139, 231, 490
  - raw 60
- frustration 252
- Fry, Dr. Stephen 119
- Fry Labs 119
- full moon. *See* parasite protocol, full moon
- fungus 54, 61, 71, 75, 76, 82, 190, 246

G

- Gaba 342
- gallbladder 103, 167
- GAMA. *See* gamma-aminobutyric acid
- gamma-aminobutyric acid 244, 248, 249, 324
  - dose 248
- gamma-glutamylethylamide. *See* L-theanine
- Garbanzo 52
- garlic 232
- garlic oil 232
- gas 51, 75
- gastrointestinal distress 505
- gastrointestinal (GI)
  - ailments 43
  - cancer 87
  - pain 331
  - problems 59
  - symptoms 177
  - tract 50, 325
- GcMAF 32, 125, 141, 293, 295, 321
  - AIDS Conference 304
  - alpha-N-acetylgalactosaminidase (Nagalase) 298
  - Altmetric score 297
  - anticancer properties 300
  - autism 295, 300, 301, 302, 305, 306, 307, 308, 311
  - autism, curing 303
  - autism improvement 302
  - autism is a curable condition 303
  - autism pathogenesis 302
  - autism treatment 302
  - bacterial 296
  - biotech company 317
  - blood brain barrier 306
  - bone disorders 300
  - brain 317
  - brain inflammation 308
  - breastfeeding 317
  - cancer 296
  - CFS/ME 305
  - conclusions 316
  - DNA 304
  - energy, increase of 305
  - eradication of HIV infection 296
  - fermented food product 315

  first publication on 296
  food fermented by probiotic bacteria 316
  fungal 296
  Gc-protein 298, 300
  grey matter 307
  haematological parameters 316
  history of 296
  hottest topics in today's world science 297
  human neurons 304
  immune system 306
  immune system development 317
  immunotherapy 296
  injections 314
  licking 317
  Lupus 300
  metabolic activity, increase of 306
  metastatic advanced cancer 296
  most effective tool to fight diseases considered incurable 316
  MRI studies 307
  Nagalase 302
  natural component of body 317
  natural nutrient 313
  neuronal connectivity 305
  neurons 305
  neutrophil cell count 316
  non-responders 303, 310
  normal brain anatomy after treatment 309
  personal experience 314
  protein, is a 304
  protozoal 296
  radiation exposure 314
  saliva 312, 315, 317
  sublingually taken 314
  therapeutic effects of 298
  transcranial ultrasonography 307
  tumours 299
  ultrasound 308, 311
  viral 296
  vitamin D 300
  what is? 312
Gc-protein 298, 313, 315
Gc-protein-derived Macrophage Activating Factor. *See* GcMAF
gelatin 52
gelcaps 226
Generation Rescue 5
gene test 463, 464
genetic material 118
genetics 453
Genova Diagnostics 388, 389
gentian root 231
germs
  fear of 333
gestation 507
Giardia lamblia 233, 238
Giera, LuEllen 450
Gilbere, Dr. Gloria 337
Gilliam Autism Rating Scale 251
ginger 226, 233
gingivitis 86
Giulivi, Cecilia 120
glands 256
glass bottles 160
gliadin 451, 454
gliadin antibodies 461
gliadin blood test 463
gliadin-IgA 463
gliadorphin 44, 49, 69
glial cell dysfunction 311
glutathione 127, 342
gluten 4, 43, 44, 49, 53, 54, 56, 58, 64, 66, 67, 69, 70, 73, 75, 245, 449, 451, 454, 455, 464
  accidental exposure 462
  antibodies 65, 66
  ataxia 456
  challenge 64
  challenge testing 67
    risks of 460, 464
  cheats 460, 462
  cross reactive foods 465
  diet 455
  elimination diet 464
  gluten-free 454, 456, 458, 459
  gluten-free diet 4, 41, 45, 47, 48, 57, 58, 70, 74, 453, 457, 513
    diet 67
    lifelong commitment 67
    pretzels 76
  grains 73, 457
  infractions (dietary) 460
  junk food 459
  molecules 69
  mortality rates 460
  peptide sequence 452
  pieces 453, 458
  proteins 48

- reactions 455
- related tests 464
- sensitivity 456
- silent gluten injury 456
- silent syndrome 456
- substitutes 465
- syndrome 65, 66, 67, 449, 452, 454, 458
  - diagnosis 455, 461
  - reaction 65
- testimonials 459
- what is 450
- withdrawals 67, 458

GlutenEase™ 460

gluten-free diet 464

Gluten Free Society 464

gluten reactivity 64

Gluten Syndrome: A Neurological Disease (paper) 456

Gluten Syndrome: Is Wheat Causing You Harm?, The (book) 456

gluteomorphins 44, 49, 56, 58, 69, 75, 245, 458, 461

gluthatione 139

glutinins 461

GMOs 320, 457, 459

gold 116

Golomb, Dr. Beatrice 263, 264

Gottschall, Elaine 58, 61

grain free diet 51

grains 51, 69, 75, 252, 324, 490
- grain-free 141
- grain husks 486

Gram, Hans Christian 78

granola bar 56

granulocytes 120

granuloma 338

grapes 61
- red 61

grass roots movement 444

green juices 133

green tea 123, 134

Group III, Dr. Edward F. 338

growth rate 330

GSE 342

Guillain-Barré Syndrome 272

gums 162, 459

Gunter, Dr. Bernard 264

gut 49, 246, 319
- clutching 331
- dysbiosis 115
- flora 75
- Gut and Psychology Syndrome (book) 75
- gut-brain connection 245
- healing the 260
- inflammation 44
- microbes 246
- mucosa 245
  - improve 245
- villi 65

Gut Biology and Treatment (presentation) 117

gut villi 454

gut wall 451

H

Hadjivassiliou, Dr. Marios 450

haematological values 316

hair 116
- brittle 181
- dry 181
- dye 61
- loss 181
- pulling 331, 332

hair products
- non-toxic 509

ham 53, 55

hamburger patties 460

handwriting 34
- poor 333

Harch, Dr. Paul 287

Harman, Denham 86

Hazelnuts 51, 52

HBOT 6, 125

HCl 90, 94, 96, 98, 155, 467, 493

headaches 62, 89, 179, 330, 331, 340, 456, 485
- migraine 144

head banging 62, 331

healing crisis 84. *See also* Herxheimer reaction

hearing 49

heart 65, 251, 253, 451, 452, 454, 461, 485, 487
  damage 456
  disease 445
  failure 452
  muscle 452
Heart and Circulation Physiology 290
heart burn 89
heartworm medication 240
heavy metals xxiii, 10, 11, 55, 81, 139, 141, 147, 162, 244, 245, 255, 257, 259, 260, 261, 320, 323, 345, 485, 505, 506, 507. *See also* metals
  exposure 301
  load 259
  toxicity xxi, 115
helminths 165, 193. *See also* parasites
hemorrhoids 177
Henry's Law 281, 282, 283
Hepatitis B 1, 265
herbicides 55, 258
herbs 11, 55, 226
heroin 44, 49
Herxheimer reaction 10, 11, 14, 57, 81, 84, 89, 90, 100, 125, 127, 146, 147, 155, 225, 226, 228, 339
HFCS. *See* high fructose corn syrup (HFCS)
hiccups 89
high fructose corn syrup (HFCS) 41, 52, 55, 123, 231
high oxalates 182
hippotherapy 175
histamine 50, 229, 236
histidine 250
HIV 138, 272, 298, 300, 315
  infection 295
hives 70, 337, 338
Hoffman, Dustin (Actor) 276
homeopathy 9
homeostasis 236
homocycsteine 342
homocysteine 253
  metabolism of 253
homogenized 49
honey 52, 63, 232, 233
hookworms 12, 165
hopelessness 443
hormones 295, 312
horsetail 457
hot dogs 50, 53, 55
H. pylori 120
Humble, Jim xiii, 10, 22, 77, 84, 85, 89, 143, 145, 327, 339, 348
  message from Kerri 80
humidifier 225
humidity 225
hunger 177, 178
hyaluronic acid 250
hydrochloric acid. *See* HCl
hydrogen peroxide 84
hydrolyzed vegetable proteins 49
hydrotherapy. *See* colon hydrotherapy
Hynes JD, MBA, Beth C. 505
hyperactivity 46, 54, 55, 62, 70, 103, 251, 324, 337
hyperammonemia 236
hyperbaric chamber 267, 321
  inflatable chamber 279
hyperbaric (oxygen) chamber xxiv, 282, 284, 286, 287, 288
hyperbaric oxygen therapy 6, 280, 287, 290, 291
hyperbarics 6, 11, 43, 141, 263, 267, 321, 324
hyperglycemia 285
hypersensitive individuals 147
hyperthermia 335
hypoglycemia 176
hypothyroidism 328
hypotonia 456
hypotonic 32

I

IBS. *See* irritable bowel syndrome (IBS)
ibuprofen 143
ice cream 49, 54
IgA 59, 451, 454, 455

IgA blood tests 463
IgD 451
IgE 70, 236, 451, 455
  allergic reactions 58, 70
  antibodies 50
IgG 58, 451, 454, 455
IgM 451, 455, 463
IgM reaction 236
immune cells 50
immune complex disease 456
immune dysregulation 505
immune function 256
immune globulin 330
immune gluten reactions 66
immune response 50, 296
  effective 335
  low 182
immune-stimulant 296
immune system 49, 51, 54, 60, 64, 66, 67, 70, 84, 85, 118, 120, 133, 227, 245, 246, 252, 253, 295, 296, 313, 320, 335, 338, 451, 453, 454, 455, 457, 463, 488
  dysfunction 301, 302
  function 54
  problems 508
  response 57, 62
immunodeficiency 298, 314, 315
Immunology of Gluten Sensitivity Beyond the Intestinal Tract, The (editorial) 453
Immunosciences Lab 450
immunotherapy 296, 300
implant, the 220
impotence, male 180
inattentive 333
Infant Body Systems 506
infants 335
infections 85, 120, 145, 162, 169, 182, 335, 451, 452, 453
  bladder 244
  chronic 167, 256
inflammation xxi, xxiii, 10, 49, 50, 53, 54, 56, 58, 60, 81, 84, 120, 128, 132, 134, 179, 236, 247, 256, 314, 323, 331, 338, 451, 485
inflammatory
  disease 460
  mediators 50
  process 453
  response 50
Influenza B (Pentavalente) 1, 265
insecticides 489
insects 61
insomnia 485
insulin 52
  spikes 52
International Probiotics Association 244
interventions 324
intestinal
  damage 62
  dysbiosis 128
  obstruction 177
  peristalsis 236
  peristaltic action 192
  permeability 59
  tract 49, 52, 59, 60, 259
  wall 59
intestinal wall 103
iodine 116
iron 45, 90, 116, 118, 119, 122, 138, 176, 180, 190, 243, 486
irritability 178, 248
irritable bowel syndrome (IBS) 177
irritation 338, 339
Is This Your Child (book) 72
itching 338
  anal 179
  nasal 179, 181
IV chelation 8, 9, 11, 139, 261

J

Jacob, Dr. Stanley 144
Jacobs, Tom 443
jaundice 181
jitteriness 246
jitters 250
Jo, Dr.. *See* Feingold, Dr. Giuseppina
joint lining 461
joint proteins 452

joints 65, 167, 179, 255, 453, 485
  pain 70, 75
journal, keeping a 46
Journal of the American Medical Association 120
juice. *See* fruit juice
juicing 133
Jung, Dr. Shin-siung 253
junk food 41, 67, 459
Just Eat Real Food movement 68
juvenile delinquency 54

K

Kaiser, Olive 64, 319, 449
Kalcker, Dr. Andreas 13, 49, 83, 85, 86, 88, 121, 145, 166, 235, 240, 323, 324, 327, 348, 444, 445
Kalcker Parasite Protocol. *See* parasite protocol
kale 133
Kanner, Leo vii
kefir 232
Keppra (drug) 248
ketogenic diet 51, 323
kidney 82, 487
killer cells 452
kiwi 123, 124, 139
Klinghardt, Dr. Dietrich 256
Knivsberg, Dr. 44
kombucha 123, 134

L

LabCorp 390
labels
  LaMotte Chlorine Dioxide Test Strips 469
  reading 47, 72
  Rompepiedras 191
  sodium chlorite 92
  supplements 124
laboratory testing 12, 345. *See also* tests
Lackney, Charlotte 156, 158, 163
lactation 54
lactobacillus 244
lactose 53
  intolerance 70, 244
  lactose-free 49
  lactose-free milk products 52
LaMotte High Range Chlorine Dioxide Test Strips (#3002) 133, 468
language 46, 49, 248, 250, 253, 330, 332
laughter 229
  inappropriate 62
laundry bleach 83. *See also* chlorine bleach
laundry detergent 343
Lavoisier, Antoine 280
laxative 195
L-carnitine 247, 248, 324
  dose 248
L-carnosine 250, 324
LD50 number 78
lead 255, 257, 506, 508
leaky gut syndrome 44, 48, 49, 59, 61, 62, 67, 73, 177, 337, 451
learning 250
learning disorders 456
lecithin 71
legumes 133
lemon 84, 134
lemon juice (CD activator) 94, 156
lepidium latifolium extract 139, 188, 191, 237, 341
  dosing 191
  Source & Cost 191
lesions 338
lethargy 62, 177, 236
lettuce 460
Leucovorin 330
leukocytes 50
ligaments 488
light
  sensitivity to 333
lime juice 93
lips 338
Lipski, M.S., C.C.N., Elizabeth 73
listening 248

listlessness 89, 121
lithium 116, 342
liver 55, 65, 82, 103, 181, 252, 313, 338, 342, 453, 457, 461, 487, 509
  enzymes 61, 141
  enzyme tests 143
  flukes 165
  function 84
Living Without (magazine) 73
localized diseases 167
Lomper® 193. *See also* mebendazole
lotions 73, 508
low intestinal carbohydrate digestive enzyme activity 60
L-theanine 249
Lubbers, Judith R. 89
lunar cycle 187
lunar phase calendar 477
lungs 50, 138, 338, 451, 487
lupus 300
lutein 61, 71
lycopene 61
Lyme disease 119, 239, 340
lymphatic system 103
lymphocytes 120, 315
Lyne-Sands, Robert. *See* Sands, Bob
lysine 247, 342

M

macromolecules 310
macroparasites 238
macrophages 296, 299, 304, 312, 315
magnesium 45, 116, 118, 119, 143, 190, 226, 324, 342, 486, 487
maintenance plan 319
makeup
  nontoxic 509
malabsorption 60
  problems 60
  syndrome 177
malaria 167, 176
malnourishment 178, 332
malondialdehyde 88, 236
malt 47, 49, 52
maltose 76
manganese 116, 190, 486
mangoes 57, 63, 123, 124, 133, 139
maple syrup 52
margarine 52
marriage 341
Marshall, Barry 119
mast cells 50
maternal detoxification 506
mayonnaise 53, 54
McCaffrey, CDC, Dee 490
McRae, Scott & Brenda McRae 156, 157
MCT 323, 342
meals 188
measurements in
  fluid ounces xiii
  pounds xiii
meats 41, 50, 51, 53, 251, 252, 459
  processed 53, 55, 63
mebendazole 141, 188, 193, 220, 237, 329, 341
  blitz day 194, 208
  dosing 194
  source 194
medications 451
medicinal plants 226, 231
medium chain triglyceride. *See* MCT
megavitamin protocols 9
Meirleir, Prof. Kenny de 301
melatonin 27, 249
memory 247, 250, 333
memory B cells 67, 452, 457
Mendel in the Kitchen (book) 457
menstrual problems 180
mental clarity 250
mercury 55, 139, 141, 245, 255, 257, 261, 457, 506, 507, 508, 509
metabolic end products 70
metabolism 246, 248
  of proper fat 247

metabolites 118
Metagenics Adreset® 342
metals 255, 257. *See* heavy metals
  porphyrins tests 143
  toxicity 5, 118, 255
Metametrix™ Clinical Laboratory 345
methemoglobinemia 143
methionine 247
methoxychlor 508
methylation
  cycle 89, 256, 261
  impaired 505
methyl mercury 255, 257. *See also* mercury
metric system xiii
Michelson, Harold 273, 274
microbes 59, 60, 118, 246
  overgrowth 60
  proliferation 59
  proliferation of the unwanted 60
  proper balance 60
microorganisms 244, 245
microparasites 167
microscope 12, 167, 186, 220
microvilli 43, 60
migraine 456. *See also* headaches
milk 43, 49, 51, 52, 53, 56, 66, 70, 71, 455. *See also* dairy
  cow's 52, 53, 69, 71, 72
  evaporated 53
  goat 49, 53, 69, 72
  human 69
  nut 133
  protein 69, 245
  raw 53
  rice 133
  soy 53
milk thistle 342
millet 51, 486
minerals 116, 122, 132, 257
  absorption 486
  deficiencies 59, 115, 243
  supplementation 143
  synthetic 116
mint 63, 231
miracle 48
Miracle Mineral Solution (MMS) 9. *See also* CD
miscarriages 507
miso 232
mission, Kerri's xxiii
mitochondria 86, 120, 121, 225
  voltage 225
mitochondrial
  defects 120
  dysfunction 120, 121, 505
  membrane 247
    electrical potential 121
  Mitochondrial Encephalopathy Partial Complex III Deficiency 328
  Mitochondrial Theory of Aging (article) 87
MMS xiii, 42, 77, 86. *See also* CD
modified citrus pectin 139
mold 70, 258, 337
molecular mimicry 64, 65, 67, 449, 451, 452, 455
  what is 450
molybdenum 116
monosaccharide molecules 75
monosaccharides 59, 60
monosodium glutamate (MSG) 53, 54
monosugars 76
mood 20, 249, 250, 256
mood swings 178
moon 187
morphine 44, 49, 69, 70, 229, 236, 245
mouthwash 86
MRI 309
MS. *See* multiple sclerosis (MS)
MSG. *See* monosodium glutamate (MSG)
mucopolysaccharides 488
mucosal gut damage 456
mucosal immune system 245
mucous 49, 128, 132, 134, 165, 235, 444
  in stool 177
mucous membrane barrier 245
muffins 49
mugwart 231
multiple sclerosis (MS) 313, 316

multivitamins 132, 342
muscle 250
  cramps 179
  spasms 179
muscle & joint pain 179
muscular abilities 330
music 277
mycobacterium 138

N

NaClO2. *See* sodium chlorite
Nagalase 298, 302. *See also* alpha-N-acetylgalactosaminidase
nails 116
  chewing 227
naltrexone 327, 329
National Health Statistics Reports viii
National Institute of Mental Health (NIMH) 333
National Institute of Neurological Disorders and Stroke (NINDS) 336
National Institutes of Health (NIH) 276, 508
natural flavoring 47, 53, 55
natural pesticide 61
nausea 89, 157, 177
navel, pain in the 179
NCGS. *See* Non-Celiac Gluten Syndrome (NCGS)
NDF+ 342
nebulizer 138
needles 190
neem 188, 195, 237, 341
  dosing 196
  source & cost 197
nematodes 187
nerves 65, 66, 253, 451, 453
  protein sequence 452
nervousness 58, 179
nervous tension 250
netti pot 138
neurodegenerative disorders 301
neurological
  disease 456, 460
  function 251
  health 487
  problems 508
neuronal connections 277
neurons 256, 305, 306
neuropathology 301
neuroprotectant 251
NeuroProtek® 342
neuropsychiatric symptoms 330
NeuroScience
  Avipaxin 342
  EndoTrax 342
  Kavinace 342
neurotoxins 58
neurotransmitters 245, 247, 248, 249, 295
neutrophil cell count 316
newborns 507
New England Journal of Medicine (NEJM) 252
New Jersey Medical School (NJMS) 43
New Jersey Medical School's Autism Center 43
Newman, Dr. Bobby 4
new moon. *See* parasite protocol, new moon
newsletter 47
Newtonian Gas Laws 282
nickel 116, 507
niclosamide 180, 187
nicotine patch 73
nightmares 228
night terrors 228
night wakings 103, 220
nitric oxide 250
nitrogen 283
non-celiac gluten-sensitivity 456
Non-Celiac Gluten Syndrome (NCGS) 461
non-crystalline silica 489
non-responder xxii, 81, 310, 311
non-systemic drug 193
noodle soup 53
nose 50
  bleeds 330

nuclear radiation 67, 457

numbness 182

numbness of the hands or feet 179

nut flour 460

nutrients 59, 60, 295, 299
- absorption 115, 453
  - poor 65
- deficiencies 453
- lost 169

Nutrients (publication) 297, 304, 313

Nutriiveda™ 134

nutrition 295, 506
- deficiency 457
- poor 451
- shakes 53

nutritional tests 143, 388, 389, 390, 391

Nutrition and Physical Degeneration (book) 258

nuts 41, 51, 53, 60, 61, 459, 460

O

oatmeal 53

oats 49, 66, 69, 455, 486

obesity 54, 178, 286, 485, 507, 508

O'Bryan, Dr. Thomas 449

obsession 179

Obsessive Compulsive Disorder (OCD) 20, 103, 140, 180, 227, 333, 334, 352, 444
- decreased 253

occupational therapy 330, 344

OCD. *See* Obsessive Compulsive Disorder (OCD)

ocean 116

ocean water xxii, 26, 115, 116, 122, 129, 132, 141, 143, 188, 189, 238, 244, 260, 282, 320, 341
- care of/handling 131
- charcoal filters 131
- dilution 117
- dosing 117
- parasites 131
- Quinton, Rene 116
- timing of chlorine dioxide dosing 117

oils
- marine 246
- plant 246

oleic acid 311

oleoresin 55

olive oil 311

Omega-3 & Omega-6 Fatty Acids 246, 247
- excessive consumption 247

omegas 320

OMP. *See* outer membrane proteins

OncoImmunology (scientific journal) 297

onions 460

opiate receptor sites 69

opiates 71

opioids 245

opium 71

oral care 85

orange juice 127

oranges 56, 57, 61, 63, 124

oregano 63

organic 53, 123, 509

organic acids 59

organisms, invading 335

organotoxins 508

organs 259, 261

Ortiz, PhD., Lorna B. vii

Osborne, Dr. Peter 450, 464

OSR 342

osteoporosis 247, 485, 487

outer membrane proteins 118

oven cleaner 343

oxalates 131, 139

oxidants 86, 88

oxidants, antioxidants and the current incurability of metastatic cancers (article) 87

oxidation 83, 120, 121

oxidative
- metabolism 286
- stress 88, 120, 236

oxidizer xiii, 78, 84

oxitriptan (5-HTP) 249

oxygen 83, 272, 280, 281, 282, 283, 284, 286, 288

oxygen chamber. *See* hyperbaric chamber

Oxyheal 279

ozone 11, 83

P

P5P 324

pain 129, 249, 331
  back, thighs, or shoulder 179

pancreas 55, 65, 251, 453, 461

pancreatitis 177

PANDAS 64, 127, 140, 328, 333, 334, 347
  probiotics 246

panic attacks 333

PANS 127, 140, 328, 330, 333, 334
  probiotics 246

papain 233

papaya 233
  seeds 233

paprika 234

parabens 508

parasite protocol xxiv, 13, 88, 125, 126, 127, 133, 136, 140, 141, 165, 166, 187, 189, 244, 260, 291, 320, 323, 337, 338, 339, 340, 444, 477
  breakfast 189
  CD 189
  CDS 189
  chanca piedra 191
  chlorine dioxide 187
  components of 187
  daily calendar 198
  diatomaceous earth 190
  dinner 189
  family members 237
  family requirements 238
  full moon 126, 129, 178, 182, 187, 224, 228, 241, 325, 326, 477
  length of treatment 187, 235, 236
  lepidium latifolium extract 191
  lunch 189
  maintenance plan 238
  meals 188
  new moon 187, 237, 325, 326, 477
  not seeing worms 236
  off days 218
  pregnancy 235
  supplies 218
  toilet hat 218
  what to expect 227
  when to stop 235

parasites xxi, xxiii, 11, 12, 20, 46, 49, 52, 81, 88, 103, 105, 111, 115, 128, 129, 140, 147, 162, 165, 167, 169, 174, 181, 183, 190, 244, 261, 320, 323, 324, 331, 337, 345, 444, 445, 458, 485
  abdominal pain or tenderness 177
  african sleeping sickness 176
  allergies 181
  altered immune regulation 167
  ammonia 176
  anal itching 181
  anemia 176, 182
  anger & irritability 178
  animal excrement 169
  animal waste 174
  anorexia 179
  anus, itching, burning, picking 179
  anxiety 178
  appetite, loss of 178
  appetite, ravenous 178
  arthritis 169, 179
  ascaris lumbricoides 183
  asthma 180
  autism 240
  babesiosis 176
  bad breath 182
  bedwetting 181
  bleach 183
  bloating 177
  blood 167
  blood analysis 182
  blood disorders 176
  blood parasites 176
  body odor 182
  bowel movements, excessive 177
  brain 167, 236, 240
  brain fog 178
  breathing problems 182
  bronchitis 180
  burping 178
  CD 238
  chagas 176
  chest pains 179
  chlorine 185
  chronic fatigue 169
  chronic fatigue syndrome 177
  clothing 183

cold, internally 177
cold in the extremities 177
colitis 177
collection 219
confused thinking 178
congenital transmission 169
constipation 177, 179
contaminated raw or undercooked meat 169
contaminated vegetables or fruits 174
contaminated water 174
coordination, lack of 178
cough 180, 182
cramps 177, 179
cravings 177, 178
cysts 179, 180
definition of 237
depression 178
dermatitis 181
desiccation 185
detox organs 176
deworming of our animals 174
diarrhea 177, 179
disorientation 178
distended belly 177
dizziness 176, 177
doctors in Western Europe and the U.S. 169
drooling (while sleeping) 181
drowsiness 180
dyspnea 180
eczema 181
eggs 131, 169, 174, 183, 184, 185, 227, 338
elevated immunoglobulin (IgE) 182
energy, low 177
enterobius vermicularis 185
eosinofils, elevated 182
erectile dysfunction 180
eruptions 181
essential nutrients 176
everything you wanted to know about sorting through poop 218
evolutionary cycle 183
eyes 167
    swollen 182
farm soil 174
fatigue 176
fecal contact 169
fertilizer 174
fever 177, 182
fibroids 180
fibromyalgia 169, 179
fish, raw 169
fluid retention 180, 182
forgetfulness 178
formaldehyde 185
freezing 174
freezing meat or fish 169
fruits & vegetables 183
gastrointestinal symptoms 177
glucosoid mucous 237
good for us? 237
growth retardation 179
hair
    brittle 181
    dry 181
    loss 181
headaches 179
hemorrhoids 177
high ammonia 182
high oxalates 182
hippotherapy 175
host immune response 167
hunger 177, 178
hypoglycemia 176
immune response, low 182
impotence, male 180
infections 182
inflammation 179
insomnia 181
intestinal
    skin problems 338
intestinal irritation 177
intestinal obstruction 177
iron 176
irritability 179
irritable bowel syndrome (IBS) 177
itching, nasal 179, 181
jars of 230
jaundice 181
joints 179
lab tests 241
large numbers in BM 239
larvae 169, 174, 184
leaky gut 177
lethargy 177
liver 167
location of 238
lunar cycle timing 187
lung 167
magnitude of problem 445
malabsorption syndrome 177
malaria 176
manure 174
mating 325
menstrual problems 180

mood swings 178
muscle 167
muscle cramps 179
muscle & joint pain 179
muscle spasms 179
nausea 177
navel, pain in the 179
nervousness 179
niclosamide 180
not visible 237
numbness 182
numbness of the hands or feet 179
obesity 178
obsession 179
organic farming 174
outer coating 191
pancreatitis 177
perforations made by 169
peritonitis 182
pets 175, 183, 240
photos 165, 168, 170, 171, 174, 175, 230
pinworms 185
pneumonia 180
poor absorption of food 177
poor growth 178
poor hygiene habits 169
poor physical and intellectual development 178
premenstrual syndrome 180
prostate problems 180
psoriasis 181
psychoactive substances 176
rapid heartbeat 179
reflexes, slow 179
respiratory disease 180
restless leg syndrome 179
restlessness 179
saliva, excess of 182
seizures 179
self-infection 169
sexual contact 169
sexual & reproductive disorders 180
shortness of breath 180
skin 167
    crawling sensation under 181
    dry 181
    itchy 181
    ulcers 181
sleep
    disorders 181
    restless 177, 181
smearing feces 180
soaps 183, 185
soil contamination 169
sores 181
stomach 167
    burning in 177
stool
    blood in 177
    detection in 218
    mucus in 177
sugar 176
swallowing, difficulty 182
sweets, cravings for 178
swelling 181
symptons 175
taenia saginata 185
taenia solium 185
tapeworm 180, 185, 187
teeth grinding 181
telling children about 239
temperatures 185
test for 235
toxic waste produced by 176
toxins 179, 229
unexplained laughter or weeping 180
urinary tract infections 180
urticaria 181
veterinarian 241
vision, blurred 182
vitamin B12 176
vitamin B12, low 182
vomiting 177
waking at night 177
wash produce, best way to 239
weakness 176
weakness, extreme 177
weight gain 178
weight gain during full moon 182
weight, inability to gain or lose 178, 179
weight loss 178, 180
yeast infection 180

parasitic infection 167
    supplementing deficiencies 243
parasitological vaccinosis 240
Parkinson's disease 120, 251, 300
parsley 486
parts per million 148, 467
pasta 53, 71
pasteurization 49, 232
pastries 49
Pathogen Biocide. *See* chlorine dioxide

pathogenesis 301
pathogenic overload 331
pathogens xxi, 11, 82, 85, 89, 103, 115, 120, 128, 147, 245, 257, 260, 261, 323, 334, 335, 337, 444
  dying 127
Patrick. *See* Rivera, Patrick
Pauling, Linus 87
PCBs 508
PDD. *See* Pervasive Developmental Disorder (PDD)
PDD-NOS 347
peanuts 52, 61, 63, 66, 451, 455
Pediatric Acute-onset Neuropsychiatric Syndrome. *See* PANS
Pediatric Autoimmune Neuropsychiatric Disorder Associated with Streptococcal Infections. *See* PANDAS
pediatricians 445
Pediatric Infection Triggered Autoimmune Neuropsychiatric Disorders. *See* PITAND
peer-reviewed journals 68
peetox 227
Pentavalente (Spanish). *See* DPT (diphtheria-pertussis-tetanus)
pepper 234
peppers 61
pepperwort. *See* lepidium latifolium extract
peptides 44, 48, 49, 56, 69, 245, 451
peptidurea 70
perchlorates 506
peripheral neuropathy 456
peritoneum 338
peritonitis 182
permitted foods list 50
Pervasive Developmental Disorder (PDD) 44, 69
Peskin, Dr. Brian 247
pessimists 443
pesticides 55, 174, 258, 485, 506, 508
pets 85, 183, 186, 353
pH 82, 282
phagocytosis 118
pharmaceuticals 506
phenols 61
phenol-sulfotransferase-P 62
phobias 136
phosphoric acid 285
phosphorus 116, 487
photosynthesis 281
phthalates 508
Physically or Otherwise Health Impaired (POHI) 328
physical therapy 330
phytoplankton 485
PIC Indolor Gravity Bag 108
piloncillo 52
pinching 331, 332
pineapple 123, 124, 133
  juice 22, 127
pinworms 12, 165, 185, 194, 220
  eggs 186
piperazine 192, 232
Pissed-Off Worm Syndrome (POWS) 228
PITAND 333, 334
pizza 49, 460
placental 451
plants, medicinal 226, 231
plasticizers 506
plasticizing agents 508
plastics 508
plateau 85
platelets 251
Play-Doh™ 53, 73, 75
PMS. *See* premensrual syndrome
pneumonia 180
pollutants 507
pollution 487
polychlorinated biphenols (PCBs) 508
polysaccharides 59, 118
polyunsaturated fatty acids 246
pomegranate bark 234

pomegranate root bark 231
pool 142
pool chlorine 83
popcorn 74
  air-popped 76
  microwave 53
Popp, Ph.D., Fritz Albert 444
pork 50, 325
post traumatic stress disorder 341
potassium 116
potatoes 4, 51, 459
poultry 55
powers of hydrogen. *See* pH
POWS. *See* Pissed-Off Worm Syndrome (POWS)
PP. *See* Kalcker Parasite Protocol
ppm 467
prebiotic 244
pregnancy 54, 235, 490, 508, 509
premenstrual syndrome 180, 247
prescription medication 42
  removing a 333
preservatives 41, 53, 54
preterm births 509
Price, Dr. Weston A. 258
Priestly, Joseph 279, 280, 281
primrose oil 342
probiotics 45, 58, 140, 188, 197, 226, 244, 245, 248, 342, 459
  PANDAS/PANS 246
  when to give 246
produce
  wash, best way to 239
Prompts for Restructuring Oral Muscular Phonetic Targets (PROMPT) 344
propolis 233, 234
prostate problems 180
protein 48, 49, 50, 70, 245, 251, 451
  digestion 48
  hydrolysate 55
proteins
  beads 450
  what are 450
Protocol 1000 157
psoriasis 181, 337
PST. *See* phenol-sulfotransferase-P
psychoactive substances 176
psychological distress 245
psychological problems 508
psychotic behaviors 58
pudding 460
pulmonary disease 313
pulmonary fibrosis 443
pumpkin seeds 192, 193, 232
punicine 234
purgative 233
pustules 337
Pycnogenol® 250, 324
  dose 250
pyrantel pamoate 141, 188, 192, 193, 237, 341, 445
  dosing 192
  source 193

Q

quercitin 342
Quest Diagnostics 391
quinoa 51
Quinton, Rene 116

R

radiation 67
rage 180
Rain Man (movie) 276
Raising Your Child Without Milk (book) 72
raisins 63
rapid heartbeat 179
Rapid Prompting Method (RPM) 344
Rapp, Dr. Doris 72
rash 46, 337, 338
RAST test 69
rats 78
reactions 46
reading 247
recipes 61

recovery xxii, 12
rectal suppository 257
rectal wall 338
redness 337
Red No. 2 55
Red No. 40 55
reflexes, slow 179
Reichelt, Dr. Kalle 69
relaxation 250
remineralization 115, 487
reproductive problems 508
respiratory disease 180
respiratory problems 508
restless leg syndrome (RLS) 179
restlessness 179
retardation
  growth 236
retina 251
retroviruses 301
rheumatoid arthritis 443
rice 51, 66, 459
Ricinus communis. *See* castor oil
ricinus oil. *See* castor oil
rickets, prevent 313
rigidity 330
Rimland, Dr. Bernard xvii, 6, 253, 264, 265, 266, 267, 276, 278, 279, 281, 289, 334
  funeral 265
  song in memory of 266
Risperdol® 42
Ritalin® 42
Rivera, Patrick xv, xx, xxi, xxiii, 1, 3, 4, 5, 6, 8, 10, 11, 13, 15, 17, 44, 81, 82, 146, 257, 264, 265, 269, 272, 341, 347
RLS. *See* restless leg syndrome (RLS)
Road to Immunity: How to Survive and Thrive in a Toxic World, The (book) 335
rocking xxii
Rompepiedras. *See* lepidium latifolium extract
root 55
root canal 452
rope parasites 165
rope worm 165, 170, 173. *See also* rope parasites
Rosedale diet 51
rotavirus 244
roundworm. *See* Ascaris lumbricoides (roundworm)
rubbing 331
rue 226, 231
Ruggiero, MD, PhD., Marco 293, 467
runny nose 89, 114, 138
rye 49, 66, 69, 455

S

Sabin, Nellie 335
S-adenosylmethionine 143
salicylate intolerance 61, 62
salicylates 61, 62
salinity 118
saliva 317, 461
saliva tests 454
salivation, excessive 182
salt 78, 83, 116
  table 86, 116
salt baths 226, 229
salt water 116
SAMe 143
sand 486
Sands, Bob xvi, 6, 263
  hospital 269
  polio 269
Sand's Hyperbaric Chamber 7
sandwiches 460
Sara's Diet 71
sarcoidosis 313
sauce 460
sauerkraut 197, 232, 459
sausages 53, 55
SCD. *See* Specific Carbohydrate Diet (SCD)
schizophrenia 44, 69, 120
school 74, 105, 126, 135, 302

Schultz, Henke 248
Schweppes, Joseph 281
Schweppes™ Bottles 160
scratching 331
screaming 332, 333, 341
scripting 125
seafood 55, 253
Sears, Dr. Robert 336
seasonal allergies 137, 330
sea water. *See* ocean water
seaweed 197
sedimentary clay 257
seeds 60
seizures 45, 58, 120, 179, 236, 240, 248, 251, 252, 253, 323, 324, 325, 326, 328, 330, 336
  febrile 336
  full moon 324
  grand mal 324, 325, 326
  new moon 324
  parasite protocol 324
  parasitic infections 325
  petit mal 324
selenium 87, 116
Self-Injurious Behavior (article) 331
self-injurious behavior (SIBs) xxii, 62, 144, 325, 331, 332
  probiotics 246
semolina 49
sensory/cognitive awareness 248
Sensory Integration Disorder 328
Sensory Integration Therapy 344
sensory overload 31
sensory table 73
separation anxiety 228
serotonin 249
  pathways 249
sesame 66, 455
Severely Multiply Impaired 328
sexual behavior 249
sexual & reproductive disorders 180
shakes 53
shampoo 73
sheep 49, 72
shellfish 50
Shi and Xie 79
shortness of breath 180
shot glass 96, 98, 105, 110, 493
shower 114
SIBs. *See* self-injurious behavior (SIBs)
side effects 11, 82
silenced nerves phenomenon 450
silent gluten injury 456
silica 457, 485, 486, 490
silicon 116, 486
silicon dioxide 486
silver 116
Singer, Emily 248
Siri and Lyons 43
skepticism 68
skin 50, 73, 85, 116, 330, 451, 488, 508
  allergies 244
  care 162
  crawling sensation under 181
  dry 181
  itchy 181
  problems 70
  test 70
  ulcers 181
  wrinkling of 485
skin issues 337
  chronic 338
slapping head/ears 331
sleep 31, 42, 62, 147, 247, 249, 330
  changes 89
  deprived 46
  patterns 46
  problems 70, 248
  restless 177
sleepiness 248, 324
sleepytox 227
small intestine 43, 59, 60, 61
smearing feces 180
smile 229
Smith, Dr. Leonard 251
smog 508
smoothies 490

sneezing 137
soaps 183, 508
sociability 248
social interactions 247
socialization 250
socializing
    problems with 458
soda 53, 285
sodium 83, 116, 486
sodium bentonite 258
sodium chlorite 78, 84, 90, 92, 105, 123, 146, 155, 157, 467
    buying 94
    powder or flakes 92
    solution 96
    technical grade (salts) 92
sodium hypochlorite 83
sodium nitrite 55
soft chambers 279
solvents 55
Solving the Puzzle of Your Hard to Raise Child (book) 72
Son-Rise® 344
sores 181, 338
sorghum 51
sound
    sensitivity to 333
sour cream 72
sourdough bread 67, 458
southernwood 231
soy 43, 49, 52, 53, 54, 56, 66, 69, 70, 74, 76, 319, 455
    cheese 72
    milk 53
    oil 71
    sauce 49, 53
    soybeans 71
    soy-free 41, 48, 57, 58
soy sauce 232, 460
spaghetti 459
Specific Carbohydrate Diet (SCD) 58, 59, 60, 324
speech xxii, 1, 45, 46, 141, 248, 250, 251, 252
    therapist 30
    therapy 24, 330
spelling 247
spelt 49
spices 55, 226
spinach 70, 253
spinal fluid 309
spinning xxii
spirochete 144
spitting 227, 326
spleen 487
sports drinks 53
squealing xxii, 1, 6
stairway to recovery xxiv
stamina 330
starch 49, 76
starch molecules 76
statistics, autism viii, xxv
steambath 21, 114, 138
steroid injections 114
stevia 51, 126, 155, 162, 225
    KAL® 51, 126
    SweetLeaf® 51, 117, 126
stimming 125
stimmy 46
stomach 51, 59, 338
    burning in 177
    problems 70
    ulcers 119
stone breaker. *See* parasite protocol, chanca piedra
stool
    analysis 12
    blood in 177
    loose 141, 332
    mucous in 177
    sample 12, 241
    tests 243, 463, 465
strawberries 63
strength 330
strep 334
strep throat 64
streptococcal infection 333
stress 248, 249
    reduce 250
stressed 46
stroke 251
sublingual methyl B-12 509
sucralose 55, 56

sucrose 231

sugar 41, 47, 50, 51, 52, 53, 54, 74, 78, 119, 133, 176, 231, 285, 459
  cane 52
  cravings for 178
  natural 51
  refined 55

Suicide by Sugar (book) 54

sulfotransferase enzyme 62

sulfur 116, 144

sulphates 61

sunflower seeds 486

sunlight 84

supersaturate 282

supplement list 45

supplements xxiv, 4, 11, 21, 45, 58, 87, 118, 138, 139, 243, 260, 320, 324, 341
  focus on excesses 90

support beyond this book 521

Supportive Typing 344

support system 47

suppositories 129

swallowing, difficulty 182

Swedo, Dr. Susan 333, 334

sweeteners 52, 55

sweetflag 231

SweetLeaf® stevia. *See* stevia

sweets 54
  cravings for 178

swelling 70, 181, 338

swelling clay. *See* bentonite clay

swimming 142

syringe/catheter method 111

systemic immune system 245

T

taenia saginata 185

taenia solium 185

tangerines 63

tansy 231

tantrums 42, 46, 57, 103, 249, 325, 332, 341

tapeworms 12, 165, 167, 180, 185, 234
  eggs 185

tapioca 51, 459

tarragon 234

taurine 251, 324

t-cells 256

tea 249. *See also* green tea

teenagers 140

teeth 86, 162, 473
  amalgams 509
  grinding 179, 181, 325
  mercury 509
  root canal 452

temperature 249

temperature, body
  common fever 336
  high fever 336
  low-grade fever 336
  normal temperature 336

temporal cortex 308

tendons 488

teriyaki sauce 49

testimonials 19, 42, 357
  curezone.com 339

tests
  antibody 456, 460
  blood 454, 464, 465
  celiac antibody 457
  false negative 65, 454, 462, 463
  gene 464, 465
  gluten
    good 462
  laboratory testing 345
  parasites 345
  saliva 454, 465
  stool 465

TheaNAQ 342

therabiotic 342

THERALAC® 188, 197, 244, 245, 341
  dosing 197
  dosing time 254
  source & cost 197

thermometer 109

thimerosal 255, 261, 507
  vaccination 267

Thinking Moms' Revolution, The 327

Thom, MD, PhD, Stephen 290

throat 50

thyme 234
thyroid 65, 256, 453, 461
  disease 313
  function 509
  nodules 314
tics 227, 333, 334
tin 255
tinnitus 485
tiredness 62. *See also* sleep
tissues 261
titanium 190, 486
TMG. *See* trimethylglycine (TMG)
toast 459
tobacco smoke 508
toe walking 274
toilet odors 343
tomatoes 61, 63, 325
tongue sores/lesions 330
toothbrush 142
toothpaste 61, 73, 84, 86, 142, 257, 508
tortillas
  flour 49, 52
toxic by-products 60
toxicity 451, 456
Toxicological Profile for Chlorine Dioxide and Chlorite (report) 78
toxic substances, exposures to 508
toxin levels 457
toxins 11, 50, 55, 59, 88, 89, 103, 115, 143, 259, 320, 337, 453, 485, 505
  environmental 445, 507
  oxidize 103
trace minerals 258, 486
transdermal applications of medications 73, 508
transglutaminase 454
TravaCor Jr. 342
Treatment and Education of Autistic and Related Communication Handicapped Children (TEACCH) 344
tremors 236
Trichinella spiralis 240
trichloroethylene 507
triclosan 508
trilombrin. *See* pyrantel pamoate
trimethylglycine (TMG) 253
  dosing 253
tryptophan 249
tTG 66, 454, 461, 463
tTG-IgA 454, 463
tumors 338
tuna fish 72
turkey 50
turmeric 234
TV watching 343, 344

U

UC Davis 120
UCLA 323
ulcerative colitis 59
ulcers 120, 338
ultrasound 308
  anatomical alterations, observed 309
  autism, treatment of 309
  autistic subject appearance 309
  drugs, transport of 311
  meninges 308
  mental states, effect on 312
  safety 308
undigested carbohydrates 60
units of measure xiii
University of California Davis Health System 43
University of Pennsylvania School of Medicine 290
updates and corrections to this book xxvii
urinary tract infections (UTI) 180
  treating 473
urination
  frequent or bedwetting 333
urine 257, 487
  analysis 243, 259
  sample 44, 69
urticaria 181
US Department of Health & Human Services viii, xxv, 78
US Food and Drug Administration 174

Usman, Dr. Anju 11, 117, 245, 256, 444, 505
uterus 338
UTI. *See* urinary tract infections (UTI)
UV exposure 118
UV light 160

V

vaccines 1, 4, 45, 255, 261, 265, 506, 507
  induced crash 45
  injury 327
  MMR 64
vaginal douche 474
vanadium 116
vegan 513
vegetables 41, 43, 51, 55, 60, 61, 133, 459
  fermented 197, 459
  juice 55
  leafy green 486
  raw 76
Venezuela Autism Clinic 142
venom 66
verbal
  communication 248
  stims 180
verminosis. *See* parasites
Vermox®. *See* mebendazole
vertigo 485
veterinarians 12, 174, 175, 186, 241
villi 65, 453
  biopsy 67, 455, 461
  damage 65, 460
Vimalaramsi, Bhante 156
vinclozolin 508
vinegar 49, 78, 343
vinyl chloride 507
viral load 84
viruses xxi, xxiii, 10, 11, 49, 81, 82, 118, 162, 244, 258, 261, 323, 345, 485, 488, 507
vision, blurred 182
vitality 116
vitamin A 87, 122, 123, 127, 133, 138, 509
vitamin B 252, 457
vitamin B6 342
vitamin B12 45, 46, 90, 122, 176, 180, 236, 243, 507
  low 182
vitamin B12 shots 342
vitamin C 87, 122, 123, 124, 127, 138, 139, 143
  water filter for shower 123
vitamin D 133, 236, 300, 304, 310, 312, 313, 487, 509
  aggression 133
vitamin deficiencies 59, 236, 243
vitamin E 87, 122, 123, 124, 127, 133, 138, 342
vitamin K 122, 123, 127, 133, 138
vitamins 74
Vojdani, Dr. Aristo 66, 449, 450, 453, 454, 455, 461, 463
volcanic ash 257
Volinsky, Dr. Alex 165, 172
vomiting 11, 43, 89, 177

W

waking at night 177
walnuts 51
walnut shell 231
Walter Reed Army Institute of Research 235
Waring, Dr. Rosemary 62
Warren, Dr. Robin 119
waste products 59
water 51, 105, 124, 133, 160, 340, 487, 506
  filter 123
    Brita® type 131
  purification 86
  reverse osmosis or distilled 96
watermelon 460
water purification drops. *See* CD
Watson, Dr. James 87
weakness 176
  extreme 177
weekend protocol. *See* 72/2 protocol

weight
  gain/lose 75, 116, 134
  inability to gain 179
  inability to gain or lose 178
  loss 228
wheat 4, 43, 49, 55, 66, 67, 69, 71, 74, 319, 455, 457, 486
  bagels 67, 457
  berries 67
  proteins 70
wheat berries 458
whey 47, 72, 134
whimpering 229
Whitcomb, Joy xxiv
white blood cells 225, 451
white fraxinella 231
white heads 337
Whitney, Norah 4
WHO. *See* World Health Organization (WHO)
whole foods 41
Wina, Dr. Peter 235
window cleaner 78
wine 78, 253
withdrawal reaction 71
Wobenzyime® 342
Woeller, Dr. Kurt 5, 6
World Health Organization (WHO) 167, 183, 244, 508
worms 26, 88, 165, 190. *See also* parasites
wormwood 226, 231

X

xanthan gum 51
xanthophylls 61
xenobiotics 506
xylitol 52

Y

Yamamoto, Dr. Nobuto 296, 298, 299, 300, 304
yarrow 226, 231
Yasko Protocol 9
yeast 20, 47, 53, 54, 55, 59, 61, 66, 75, 76, 81, 162, 190, 258, 337, 345, 455
  edible 55
  infection 180
yogurt 49
Young, Dr. Vincent 246
Yu, Dr. Simon 443

Z

zeaxanthin 61
zeolite clay 140
zinc 75, 116, 259, 509
zirconium 486
Zone, Dr. John 73
zonulin 49
Zukin, Jane. *See* Raising Your Child Without Milk (book)

About the Author

Kerri Rivera is a native of Chicago, but has lived the last 19 years of her life in Puerto Vallarta, Mexico with her husband. Both of her sons were born there, and 13-year-old Patrick is currently in autism recovery. Kerri is the founder of AutismO2, a non-profit autism clinic based on what is formerly known as the *Defeat Autism Now!* approach. Located in Puerto Vallarta, the clinic opened in 2006. Kerri is the clinic's biomedical consultant as well as a consultant for Curando El Autismo (Latin America), and Venciendo el Autismo (Venezuela).

At the request of Dr. Bernard Rimland, she was responsible for translating the *Defeat Autism Now!* protocol to Spanish and donating it to the Autism Research Institute, so it could be applied throughout Latin America. Since then she has helped over 3,500 families in over 58 countries to improve the lives of their children on the autism spectrum. She has become the foremost expert on the use of CD (chlorine dioxide) for spectrum disorders. Since the addition of CD to the biomedical protocol she has seen 115 children recover from autism in the past 2.5 years.

Since 2007 Kerri has lectured internationally on the Biomedical Protocol for autism and CD for autism. She has been featured on *One Radio Network* with Patrick Timpone, *The Bulletproof Executive* with Dave Asprey, *The Mother Cub Show*, *Thought for Food*, *Enlightened Health Radio* with Justin Elledge, *Voice America*, and *The Robert Scott Bell Show*. She is the Mexican Liaison for AutismOne. Kerri graduated as a Certified Homeopath in June of 2013.

Our greatest weakness lies in giving up.

The most certain way to succeed is always to try just one more time.

~ *Thomas Edison*

The world of autism is rocking again. Kerri Rivera has done it. In this comprehensive book she has condensed and simplified the core elements of the biomedical approach, those that work 90% of the time on 90% of the children. Her focus on using a unified and affordable approach to reduce the body burden of chronic infections and infestations is addressing the very core issue of most autistic children and adults.
It has made her method the most powerful tool to this day to help many children recover safely, inexpensively and solidly. Her research is validated by the enormously positive response from parents from all over the world.

—Dietrich Klinghardt MD, Ph.D.

EXCITING UPDATES TO THIS EDITION INCLUDE:
A new method of CD preparation • New parasite protocol charts
Lunar calendar • Many new testimonials
A section on gluten's role in molecular mimicry and autoimmunity
Information on helping older or violent / aggressive children

In the seven months since the release of the first edition of *Healing the Symptoms Known as Autism*, 22 more children have lost their autism diagnosis and returned to a state of health, for a total of 115 recovered children and counting. Hundreds more have lost ATEC points, as well as shown many cognitive, behavioral, emotional and physical gains thanks to the protocol outlined in these pages. This edition includes all protocol updates and a whole lot more information that can benefit families of children and adults on the spectrum.

In 3 years, 115 children previously diagnosed with regressive autism were able to shed their diagnosis, their symptoms and return to an overall state of health and vitality. If you have any interest in finding out how they were able to heal you need to read this book. Kerri has a heart for children with autism, and she cares greatly about recommending safe, economical interventions that are helping to confront an epidemic.

—Teri Arranga
Executive Director of AutismOne (www.autismone.org)

With over 180 TESTIMONIALS from around the world!

ISBN 978-0-9892890-2-3
52499>
9 780989 289023

LANDES
BIOSCIENCE

Vademecum

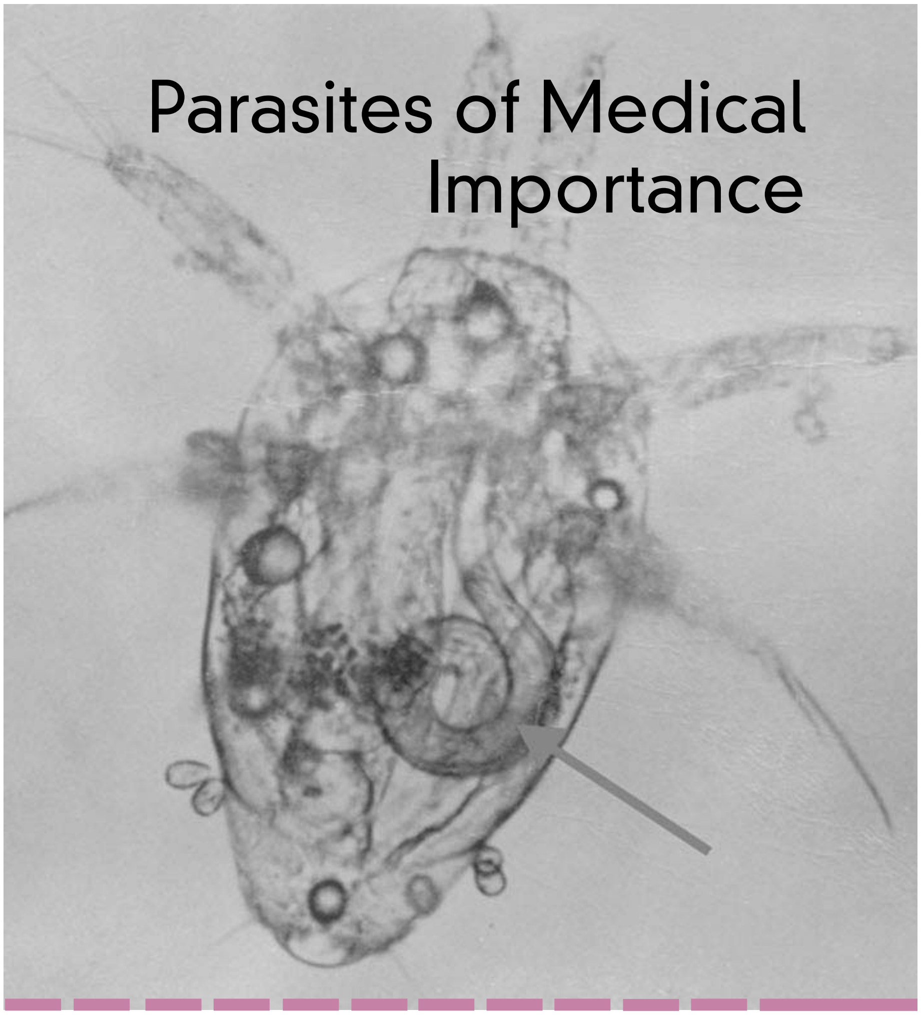

Parasites of Medical Importance

Anthony J. Nappi
Emily Vass

vademecum

Parasites of Medical Importance

Anthony J. Nappi, Ph.D.
Department of Biology
Loyola University
Chicago, Illinois, U.S.A.

Emily Vass, Ed.D.
Department of Biology
Loyola University
Chicago, Illinois, U.S.A.

LANDES
BIOSCIENCE
GEORGETOWN, TEXAS
U.S.A.

VADEMECUM
Parasites of Medical Importance
LANDES BIOSCIENCE
Georgetown, Texas U.S.A.

Copyright ©2002 Landes Bioscience
All rights reserved.
No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.
Printed in the U.S.A.

Please address all inquiries to the Publisher:
Landes Bioscience, 810 S. Church Street, Georgetown, Texas, U.S.A. 78626
Phone: 512/ 863 7762; FAX: 512/ 863 0081

ISBN: 1-57059-679-4

Library of Congress Cataloging-in-Publication Data

CIP information applied for but not received at time of publishing.

While the authors, editors, sponsor and publisher believe that drug selection and dosage and the specifications and usage of equipment and devices, as set forth in this book, are in accord with current recommendations and practice at the time of publication, they make no warranty, expressed or implied, with respect to material described in this book. In view of the ongoing research, equipment development, changes in governmental regulations and the rapid accumulation of information relating to the biomedical sciences, the reader is urged to carefully review and evaluate the information provided herein.

Dedication

With the hope that it will live up to her high standards and expectations, this book is dedicated with affection to Emily, co-author, colleague and friend. She was a young scholar who always strived to learn more than the basics. She enjoyed her work, and it was a joy to work with her. Á toute á l'heure.

Emily Vass
11 February 1967 - 19 April 2002

Contents

1. Interspecific Interactions 1
Specificity in Host-Parasite Relations 2
Modes of Infection 2
Clinical Effects of Animal Parasitoses 3
Prevalence of Parasitic Diseases 4

2. Major Groups of Parasites of Humans 5
Major Groups of Parasitic Protozoa 5
Protozoan Reproduction and Life Cycles 6
Parasitic Flagellates 7
Hemoflagellates: *Trypanosoma* 7
Hemoflagellates: *Leishmania* 14
Flagellates of the Digestive and Reproductive Passages 18

3. Sarcodina 19
Ciliate Parasites 23

4. Apicomplexa: Sporozoa and Piroplasmea 24
Introduction to Sporozoa 24
Piroplasmea 33
Other Apicomplexa 33

5. Digenetic Trematodes: Flukes 36
Life Cycle 36
Intestinal Flukes 40
Hepatic Flukes 42
Pulmonary Flukes 46
Blood Flukes 46

6. Cestodes 54
Developmental Stages and Life Cycles 55

7. General Morphology of Parasitic Nematodes 70
Trichuris trichiura 71
Trichinella spiralis 74
Strongyloides stercoralis 77
Hookworms 78
Cutaneous Larva Migrans 82
Visceral Larva Migrans 87
Filarial Worms 93

8. Arthropods 100
Types of Injury Caused by Arthropods 100
Arthropods as Vectors of Disease 101
Chelicerates (Arachnids) 101
Insects 110

Glossary 127

Index 145

Preface

Virtually every organism serves as the host for a complement of parasites. Parasitism is so common that it is rare to find classes of animals without members that have adopted a parasitic mode of living. Evidence gained from various archeological studies indicates that parasitic diseases existed in prehistoric human populations. Since there is no evidence to suggest that our long and intimate association with parasites will ever end, it seems reasonable to propose that the study of human parasites warrants some consideration. However, the study of parasites is a very challenging endeavor. Host-parasite associations involve complex biochemical, physiological, behavioral and ecological adaptations that very likely have co-evolved independently and on many different occasions. These complex and intimate interactions are continually evolving as counterstrategies in both host and parasite populations, thus limiting our ability to adequately study the factors that influence immune competency, parasite virulence, adaptability, epidemiological diversity, and drug resistance. However, the most important challenge facing parasitologists derives not from technical or experimental difficulties, but from the fact that most of the parasitic diseases that have a major impact on humans are largely associated with the rural poor in tropical, developing countries, which typically attract little interest from strictly commercial enterprises and other agencies that fund research.

Today, the extent of human suffering due to parasites is incalculable and intolerable. The physiological, pathological and economic problems caused by parasites are global concerns, and it is imperative that health professionals have some understanding of the complex interactions between humans and their parasites. Inexplicably, many medical schools fail to offer a curriculum that contains a formal course in parasitic diseases, or, in some cases, even to provide a single lecture on the topic. It is our belief that the collaborative efforts of parasitologists and medical professionals are urgently needed to improve efforts to treat parasitic infections. *Parasites of Medical Importance* is designed primarily for health professions and students interested in pursuing careers that will address the growing threat current and emerging parasitic diseases pose to the global population. In preparing this textbook we assumed that it would be a first exposure to the study of parasites for those who have had little or no formal instruction in parasitic diseases. Thus, emphasis has been placed on parasite life cycles and host pathology, with limited discussions of parasite morphology, taxonomy, and pharmacological treatments.

The authors assume full responsibility for omissions or any mistakes that appear in the book, and will correct such issues in subsequent editions.

Acknowledgements

We wish to thank Dr. Pietro Carmello of the Carlo Denegri Foundation, Torino, Italy, for granting permission to use several of the illustrations that are maintained by the Foundation. We wish to acknowledge the Centers for Disease Control, Division of Parasitic Diseases, Atlanta, Georgia and the Bayer AG Company, Leverkusen, Germany, for providing several illustrations. A special thanks to the following colleagues who provided us with original photographs: Harvey Blankespoor, Robert Kuntz and Dianora Niccolini. A portion of the effort spent on finishing the textbook was made possible because of research support from The National Institute of Health (GM 59774), The National Science Foundation (IBN 0095421) and Research Services at Loyola University Chicago.

CHAPTER 1

Interspecific Interactions

The term symbiosis, which means the “living together” of two species, was first used in 1879 by the botanist Heinrich Anton de Bary to describe the relation between certain species of fungi and algae living together to form lichens. Based primarily on the type of dependency that exists between the interacting species, three types of symbiosis are distinguished; commensalism, mutualism, and parasitism.

Commensalism is a type of symbiotic association which is beneficial to one species and at least without any detectable adverse effect on the other species. The basis for such a relation may be food, substrate, space, or shelter. The commensal is usually the smaller of the two species and may be attached to the exterior of the host (ectocommensal), or live within the body of the host (endocommensal). Examples include certain tropical commensal fishes, which are protected from predation by living among the tentacles of certain sea anemones, and the pilot and remora fishes, which associate with sharks, sea turtles, or other species of fish usually feeding on “leftovers”. If the association is only a passive transport of the commensal by the host, the relationship is referred to as phoresy. Phoresy is essentially an accidental association with no metabolic dependency or interaction between the two individuals.

Mutualism is an association of two species that are metabolically dependent on each other. Examples of mutualism include flagellates living in the gut of wood roaches and termites, lichens, and the cultivation of fungi by various species of insect. Parasitism is an association of heterospecific organisms during which the parasite, usually the smaller of the two species, derives its nutrient requirements directly from (and at the expense of) the host. In some heterospecific interactions it is difficult to determine the nature of the symbiotic association because variations exist in the degree of intimacy, pathogenicity, and permanency of the association. Parasites living within the body of their hosts are termed endoparasites, while those attached to the outer surface of the body are called ectoparasites. The term infection is commonly used when discussing endoparasites, and infestation when reference is made to ectoparasites. Parasitoses is the infection or infestation of a host with animal parasites.

Parasitism may be the only option for an organism, or it may be an alternative way of life. If an organism is completely dependent on its host during all or a part of its life cycle and cannot exist in any other way, the parasite is known as an obligatory parasite. A facultative parasite is an organism that does not depend on the parasitic way of life at any stage during its development, but may become parasitic if provided the opportunity to do so. Multiple parasitism occurs when a host is infected (or infested) by two or more species of parasites, whereas superparasitism is the infection of a host by more individuals of a single species of parasite than the host can support. The host may be so severely injured by the heavy infection that, if it does not succumb, it provides such an inadequate environment for the parasites that they fail to develop completely and eventually die. The term superinfection is used when an infected

---

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

host is reinfected with the same species of parasite. If two or more hosts are involved in the life cycle of a parasite, the host in which the parasite reaches sexual maturity is termed the final or definitive host. Hosts associated with larval or juvenile stages of a parasite are referred to as intermediate hosts. A biological vector is a host that is not only required for the development of the parasite, but also for transferring the parasite to another host. A transfer or paratenic host is one that is not absolutely necessary for the completion of the parasite's life cycle, but serves as a temporary refuge and/or mechanical vector for transfer to an obligatory host. Hosts that serve as a direct source from which other animals can be infected are known as reservoir hosts. The term zoonoses refers to those diseases transmittable to humans from other animals.

Specificity in Host-Parasite Relations

Specificity refers to the mutual adaptations that restrict parasites to their hosts. A high degree of host specificity indicates that the parasite is unable to survive in association with any other species. The human pinworm, *Enterobius vermicularis*, and the beef tapeworm, *Taenia saginata*, are examples of parasites that are very host specific. Some of the factors that prevent a parasite from infecting an organism other than the host species include host immunity, seasonal, behavioral, or geographic barriers, or the absence of specific metabolites, intermediate hosts or biological vectors that are required for parasite development.

Host specificity may be a function of physiological, ecological, and/or behavioral adaptations. The conditions determining the degree of host specificity often are markedly different for the various developmental stages of a parasite that uses different hosts to complete its life cycle. Parasites with two or more intermediate hosts are less specific than those with one intermediate host. Also, parasites that infect the host by penetrating the skin tend to be more host-specific than those that are ingested by the host. Even within a single host the physiologic demands of the different stages of a parasite may be so different that there is site specificity (blood, liver, etc.) at different times during development. Generally, a parasite that has a high degree of host specificity requires a specific site within its host in which to develop, while a parasite that is not host specific lives in various host tissues. The beef tapeworm, which is specific for humans, can live only in the small intestine. On the other hand, the roundworm *Trichinella spiralis*, which infects various warm-blooded animals, can live in different host tissues. Unfortunately, very little is known of the factors that determine the localization of parasites within their hosts. The host tissue-specific sites occupied by parasites represent specific niches, and complex behavioral and physiological adaptations regulate the precise migratory routes followed by the parasites in locating these sites for their development.

Modes of Infection

The life cycles of parasites are characteristically complex, with many specific requirements for development and survival. Parasites with a direct life cycle develop in or on the body of only a single, definitive host. These parasites generally have a free-living stage away from the host, and adaptations for the successful transfer of this stage often include a protective covering (i.e., cuticle, thickened cell wall or cyst) and/or locomotor structures that propel the parasite in the environment. Parasites with indirect life cycles contend not only with environmental problems, but also with different biotic requirements of the definitive and intermediate host(s) that often belong to different phyla. Natural transfer of the infective stage(s) of a parasite may be accomplished by ingestion of contaminated food or water, inhalation, inoculative transmission during feeding of an infected host (e.g., trypanosomes, malaria), or by the active penetration of the host body by the parasite (cercariae of blood flukes, hookworm larvae). There may be transplacental transmission (*Toxoplasma gondii*), as well as via sexual intercourse (*Trichomonas vaginalis, Treponema pallidum*). Parasites may escape from their hosts by actively penetrating their tissues and by passage through the digestive, urinary, respiratory, or reproductive systems.

Clinical Effects of Animal Parasitoses

The adverse effects a parasite has on a host depend on numerous factors including host age, health, immune competence, nutritional state, site of attack, number of parasites, and the interaction of various environmental factors. In some host-parasite interactions there may be no pathological symptoms of infection (asymptomatic), while in others the parasites may produce clinically demonstrable effects. Unfortunately, the pathologies caused by animal parasites are not always diagnostically specific, and these may be mistaken for a variety of bacterial, fungal, or viral infections. Hence, positive identification of the parasite is always essential for effective treatment. Some examples of parasite-induced injuries include:

1. **Tissue Damage.** Injuries to tissues may occur during and/or after penetration of the host. Examples include scabic mites, fly maggots, ticks, penetration of hookworms, and mosquito punctures. The migration through the host body of eggs and larval stages of various helminths produce tissue lesions. Also, lytic necroses may result from enzymes released by tissue-inhibiting parasites.
2. **Stimulation of Host Cellular and Tissue Reactions.** Parasites and/or their metabolites may induce various inflammatory and immune responses by the host. Blood disorders may include eosinophilia, erythropoiesis, anemia, polymorphonuclear leukocytosis, and leukopenia. The salivary and venomous secretions of insects and other arthropods may provoke systemic responses such as allergic and neurological reactions in addition to localized skin inflammation at the site of the wound. Tissue abnormalities may involve fibrosis, granulomatous growths, metastasizing sarcomas, and carcinomas. In various cell types there may be evidence of hyperplasia (accelerated rate of mitosis), hypertrophy (increase in size), and metaplasia (abnormal cellular transformations). The production of antibodies (immunoglobulins) and the mobilization of phagocytic cells may in part characterize the immune response to various parasites.
3. **Mechanical Interference.** The invasion of numerous parasites into the body may cause partial or total obstruction of the digestive system and associated organs, circulatory system, and the lymphatic system. Considerable necroses of these organs are also manifested in heavy infections.
4. **Nutritional Disturbances.** Parasites acquire nutrients by consuming a

portion of the food ingested by the host, and/or by feeding directly on host cells, tissues, or body fluids. Host metabolism may be severely disturbed by the presence of parasites, and symptoms of a chronic nature such as gradual loss of weight and progressive weakness may develop. Parasite-induced pathogenicity may be manifested in response to inadequate host nutrition.

5. **Secondary Infections.** Many parasites produce ulcerations and wounds as they enter the host. These areas subsequently become sites for infection by microbial pathogens. Secondary microbial infections may be more serious than those caused by the parasites.

Prevalence of Parasitic Diseases

The incidence of human infection with parasites is staggering. These global problems are magnified because numerous other parasites ravage livestock, reduce agricultural productivity, and contribute greatly to serious nutritional deficiencies in underdevelped countries. The extent of suffering due to parasites is incalculable. Various sources estimate that approximately one billion persons are infected with the roundworm *Ascaris lumbricoides*, 700 million suffer from filariasis, 270 million have schistosomiasis, and 20 million suffer from trypanosomiasis. Each year, between 300 and 500 million people contract malaria, of whom between 1.5 and 2.7 million die. The World Health Organization estimates that one-fifth of the world population is under threat from this disease. Malaria and other mosquito-borne diseases (e.g., dengue fever, yellow fever, meningitis, filariasis) cause a death every 30 seconds. Either our present medical technology is inadequate to cope with these parasitic infections or our priorities need to be altered. Tropical medicine does not occupy a position in the mainstream of biomedical and clinical research because parasitic diseases generate little journalistic attention, and because the billions of people who suffer from tropical disorders are mostly poor, illiterate, and are seldom heard from. These problems represent global concerns with billions of individuals at risk.

CHAPTER 2

Major Groups of Parasites of Humans

I. Protozoa. Unicellular eukaryotic organisms
II. Helminths. Parasitic worms
Phylum 1. Platyhelminthes: Flatworms
Flukes
Tapeworms
Phylum 2. Nemathelminthes: Nematodes or unsegmented roundworms
III. Arthropods. Animals with chitinous exoskeleton and jointed appendages
Insects, spiders
Mites, Ticks
Scorpions, Lice
Fleas

Major Groups of Parasitic Protozoa

Phylum 1. Sarcomastigophora

These protozoans possess monomorphic nuclei, and flagella, pseudopodia, or both types of locomotor structures. They reproduce asexually by binary and multiple fission, typically without spore formation, and sexually by fusion of isogametes or anisogametes.

Subphylum Mastigophora (Flagellates)

Important species infecting humans include *Giardia lamblia*, *Leishmania tropica*, *L. braziliensis*, *L. donovani*, *Trichomonas vaginalis*, *Trypanosoma rhodesiense*, *T. gambiense* and *T. cruzi.*

Subphylum Sarcodina (Amoebas)

Important parasitic amoebas include *Acanthamoeba*, *Endolimax nana*, *Entamoeba histolytica*, *Entamoeba polecki*, *Entamoeba gingivalis*, *E. coli*, *E. hartmanni*, *Hartmannella*, *Iodamoeba butschlii* and *Naegleria fowleri*.

Phylum 2. Ciliophora

Very few parasitic forms are present in this subphylum. Simple cilia or compound ciliary organelles are present at some stage in their development. In most members the nuclei are of two types. Asexual reproduction is by binary fission; sexuality involves conjugation, autogamy, or cytogamy. The single important species is *Balantidium coli.*

Phylum 3. Apicomplexa

All members of this group are intracellular parasites without locomotor organelles. A complex system of organelles is present in the apical end at some stage. One or

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

more micropores is generally present. The organisms reproduce sexually and/or asexually, and in some members a cyst stage is present.

Class Sporozoa.

Subclass 1. Coccidia

Important species include *Isospora belli*, *Plasmodium malariae*, *P. vivax*, *P. falciparum*, *P. ovale*, *Sarcocystis lindemanni* and *Toxoplasma gondii.*

Subclass 2. Piroplasmia

Two important species are *Babesia bigemina* and *Theileria parva*.

Other Apicomplexa

The taxonomic status of some species remains questionable: One important member is *Pneumocystis carinii*.

Protozoan Reproduction and Life Cycles

Protozoans are typically microscopic and unicellular, and possess one or more nuclei and other organelles comparable to the cells of metazoan organisms. Protozoan parasites cause more suffering, debilitation and death than any other group of pathogenic organisms. The success of this group is attributed in large measure to their high reproductive potential.

Parasitic protozoans reproduce by asexual and/or sexual methods. Asexual methods include schizogony, or multiple asexual fission, and budding. In schizogony the nucleus and certain other organelles undergo repeated divisions before cytokinesis; the nuclei become surrounded by small amounts of cytoplasm, and cell membranes form around them while they are within the mother cell which becomes known as a schizont. The daughter cells, termed merozoites, are liberated when the cell membrane of the schizont ruptures. If multiple asexual fission follows the union of gametes, the process is termed sporogony. Budding involves mitosis with unequal cytokinesis.

Sexual reproduction in parasitic protozoa involves reductional division in meiosis resulting in a change from diploidy to haploidy, with a subsequent restoration of diploidy by the union of gametes (syngamy) derived from two parents (amphimictic), or from a single parent (automictic). When only haploid nuclei unite, the process is called conjugation. Gametes may be similar in appearance (isogametes) or dissimilar (anisogametes). Marked dimorphism is frequently seen in anisogametes. The larger gamete (female) is termed macrogamete, the smaller, generally more active gamete (male) is the microgamete. Fusion of gametes produces a zygote. Frequently, the zygote is a resting stage that overwinters or forms spores that enable survival during transfer to different hosts.

Another mechanism of transfer between hosts is encystment. In some parasitic forms, the normal feeding or vegetative stage (trophozoite) cannot infect new hosts because it cannot survive the transfer. Such protozoans secrete a resistant covering around themselves and enter a resting stage or cyst. In addition to protection against unfavorable conditions, cysts may serve for cellular reorganization and nuclear division. Possible adverse conditions within the host favoring cyst formation include desiccation, deficiency of essential host metabolites, changes in pH, temperature, or tonicity. In the group Sporozoa, the cyst is termed an oocyst. Within the oocyst sporogony and cytokinesis occur to produce infective stages termed sporozoites. The oocyst may serve as a developmental capsule for the sporozoites within the host, or it may be the resistant stage that is transmitted to new hosts.

Parasitic Flagellates

Flagellates constitute the largest group of parasitic protozoa. Typically, the body of a flagellate is elongate and slender with a single flagellum. However, some species are spheroid in shape, possess more than a single flagellum, or lack flagella entirely. The flagellum, which arises from a basal body or kinetoplast, may originate near, and extend freely from, the anterior end, or it may run along the free margin as an undulating membrane attached to the side of the organism. In some cases the flagellum passes through the entire body, extending beyond the anterior end of the organism as a free structure. Based on characteristics of the flagellum, as many as four developmental stages occur in the life cycles of flagellates that are transmitted by insects to humans (Fig. 1). The amastigote is a spheroid form, devoid of an external flagellum. A small internal flagellum extends only slightly beyond the flagellar pocket. This stage is found in the life cycles of the three species of *Leishmania* parasitizing humans. The promastigote is an elongate form with a kinetoplast located in front of the nucleus (antenuclear), near the anterior end of the organism. A short flagellum arises near the kinetoplast and emerges from the anterior end of the organism. The epimastigote is an elongate form. The kinetoplast is close to the nucleus (juxtanuclear) with a flagellum arising near it and emerging from the side of the organism to run along a short undulating membrane. The trypomastigote is an elongate form with a post nuclear kinetoplast and is the definitive stage of the genus *Trypanosoma*. The flagellum emerges from the side of the organism to run along a long undulating membrane, which is directed anteriorly. Two additional morphological stages of flagellates are known, the choanomastigote, and opisthomastigote. The choanomastigote form is slightly ovoid and has an antenuclear kinetoplast. The flagellum emerges from a wide funnel or collar-like reservoir at the anterior end of the body. The opisthomastigote is an elongate form with a post nuclear kinetoplast. The flagellum passes through the organism and emerges from its anterior end. There is no undulating membrane present. Neither the choanomastigote nor the opisthomastigote form occurs in the life cycle of any flagellate parasite of humans.

The flagellate parasites of humans generally reproduce asexually by longitudinal binary fission. Based on their location within their hosts, two medically important groups are recognized; flagellates of the blood and connective tissues (hemoflagellates), and flagellates of digestive or reproductive systems. Hemoflagellates require a blood-sucking arthropod as a biological vector, while flagellates of the digestive and reproductive passages do not.

Hemoflagellates: *Trypanosoma*

The flagellates that parasitize the blood and tissue of vertebrates belong to the family Trypanosomidae. There are two important genera of *Trypanosoma*tids, *Leishmania* and *Trypanosoma* (Table 1).

Most of the Trypanosomatids that parasitize terrestrial vertebrates require a blood-sucking arthropod as a biological vector. Two mechanisms of transmission of

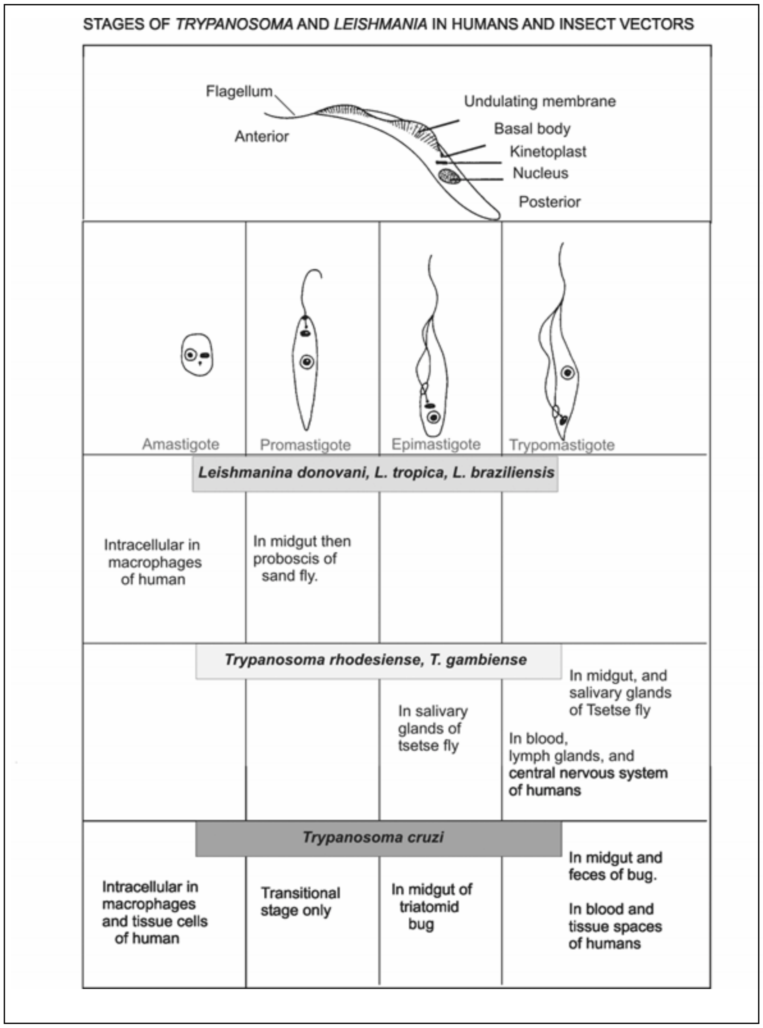
STAGES OF *TRYPANOSOMA* AND *LEISHMANIA* IN HUMANS AND INSECT VECTORS
Flagellum
Anterior
Undulating membrane
Basal body
Kinetoplast
Nucleus
Posterior
Amastigote
Promastigote
Epimastigote
Trypomastigote
Leishmanina donovani, L. tropica, L. braziliensis
Intracellular in macrophages of human
In midgut then proboscis of sand fly.
Trypanosoma rhodesiense, T. gambiense
In salivary glands of tsetse fly
In midgut, and salivary glands of Tsetse fly
In blood, lymph glands, and central nervous system of humans
Trypanosoma cruzi
Intracellular in macrophages and tissue cells of human
Transitional stage only
In midgut of triatomid bug
In midgut and feces of bug.
In blood and tissue spaces of humans

Figure 1. Morphological stages of *Trypanosoma* and *Leishmania* found in humans and insect vectors. Modified from Beaver PC, Jung RC. Animal Agents and Vectors of Human Disease. Philadelphia: Lea and Febiger, 1985.

hemoflagellates occur with blood-sucking arthropod vectors. In one, the parasites pass from the mouth parts of the blood-feeding vector directly into the definitive host. This inoculative transmission is referred to as infection from the anterior station. In the second method, the parasites are voided in the feces of the biological vector, and infection occurs when the parasites are inadvertently rubbed into the wound produced by the vector when it bites the definitive host. This mechanism of infection resulting from wound contamination is referred to as infection from the posterior station. The only trypanosome of vertebrates not transmitted by an animal vector is *T. equiperdum*. This flagellate, which causes dourine in horses and mules, is transmitted during coitus.

Genus Trypanosoma

Most species of the genus *Trypanosoma* are parasites of the blood, lymph and tissue fluids of vertebrates. In these hosts, they appear in the trypomastigote form and divide by longitudinal binary fission. One notable exception is *T. cruzi,* which has become adapted to intracellular life in the amastigote form and does not multiply in the trypomastigote form. There are two major human diseases caused by trypanosomes, sleeping sickness, a disease found in Africa, and Chagas' disease which occurs in Central and South America, and parts of the United States.

African Trypanosomiasis (Sleeping Sickness)

Based on their separate geographic distributions and generally different clinical manifestations, two forms of African sleeping sickness are distinguished; Gambian (chronic) or West African form caused by *T. gambiense*, and a more virulent Rhodesian (acute) or East African form caused by *T. rhodesiense*. These two species are morphologically indistinguishable from each other, and from a third species, *T. brucei*, which infects many domestic and natural game animals, but apparently does not parasitize humans. *Trypanosoma gambiense* and *T. rhodesiense* are transmitted to humans by both sexes of the tsetse fly *Glossina*. *Glossina palpalis* and *G. tachinoides* are the principal biological vectors of *T. gambiense*, while those of *T. rhodesiense* are *G. morsitans*, *G. pallidipes*, and to a lesser extent, *G. swynnertoni.*

The trypanosomes of humans typically are found in the blood, lymph, spleen, liver, and cerebrospinal fluid (Fig. 2). When the tsetse fly bites and takes a blood meal from an infected individual, the flagellates are taken into the midgut of the insect where development begins. The trypanosomes later migrate into the proventriculus, labial cavity, and then into the salivary glands where they develop to the infective or metacyclic stage. The complete life cycle in the insect requires 2 to 3 weeks. Human infection occurs during host feeding when an infected tsetse fly injects the parasites contained in the saliva into the skin. In the area of the inoculation the parasites initiate an interstitial inflammation that gradually subsides within a week. Occasionally ulcerations appear at the site of the puncture with the formation of an indurated, painful chancre, which slowly disappears. Within 1 to 2 weeks after infection, the parasites gain access to the circulatory system and cause a heavy parasitemia, chills, fever, headache, and occasionally nausea and vomiting. Congenital infection is also possible with the parasites passing through the placenta. Breast milk from infected individuals also may be a source of infective trypanosomes.

The Gambian form of sleeping sickness involves primarily lymphoid and nervous tissues. Marked lymphadenitis occurs with the painful enlargement of the posterior cervical lymph nodes (Winterbottom's sign). Slaves from Africa en route to the Caribbean exhibiting such enlarged cervical lymph nodes were routinely thrown overboard by slave traders. Some of the more pronounced clinical manifestations as the disease advances include edema, enlargement of the spleen and liver, anorexia,

Table 1. Major blood and tissue-dwelling flagellates of humans

Parasite	Epidemiology	Location in Host	Mode	Symptoms of Infection
Leishmania tropica	Mediterranean area, Asia, Africa, Central America	Skin	Bite of *Phlebotomus* (sandfly)	Skin lesion
Leishmania braziliensis (espundia)	Central and South America	Skin and mucocutaneous tissue	Bite of *Phlebotomus*	Skin lesions, enlarged liver and spleen, death
Leishmania donovani (kala-azar)	Mediterranean area, Asia, Africa, South America	Skin and somatic organs	Bite of *Phlebotomus*	Skin lesions, enlarged liver and spleen, death
Trypanosoma gambiense (Gambian sleeping sickness)	West Africa	Blood, lymph nodes, central nervous system	Bite of *Glossina* (tsetse fly)	Lymphadenopathy (Winterbottom's sign), meningoencephalitis enlarged liver and spleen, lethargy, death
Trypanosoma rhodesense (Rhodesian sleeping sickness)	Eastern and Central Africa	Blood, lymph nodes, central nervous system	Bite of *Glossina* (tsetse fly)	Enlarged liver and spleen, Glomerulonephritis, meningoencephalitis, death
Trypanosoma cruzi (Chagas' disease)	United States, Central and South America	Cardiac muscle, blood and other tissues	Reduviid bugs by posterior station, congenital, ingestion of infected mothers' milk	Muscle pain, lymphadenitis, meningoencephalitis, myocarditis, tachycardia, death (Romana's sign)

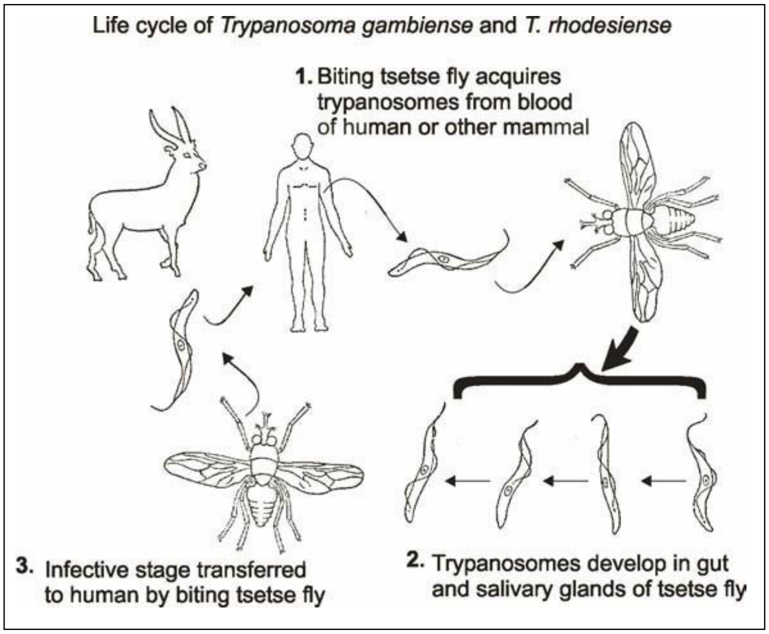
Life cycle of *Trypanosoma gambiense* and *T. rhodesiense*
1. Biting tsetse fly acquires trypanosomes from blood of human or other mammal
2. Trypanosomes develop in gut and salivary glands of tsetse fly
3. Infective stage transferred to human by biting tsetse fly

Figure 2. Life cycle of *Trypanosoma gambiense* and *T. rhodesiense.* Modified from Belding, D. L. 1958. Clinical Parasitology. Appelton-Century-Crofts, Inc., New York.

extreme weakness, rapid loss of weight, disturbed vision, meningoencephalitis, fibrillation of facial muscles, tremor of the tongue and hands, mental deterioration, paralysis, convulsions, and finally coma and death. The complete course of the disease may extend over several years. Rhodesian trypanosomiasis is a more rapid form of the disease than the Gambian form, usually resulting in death within a few months. Generally, there is little or no neurologic involvement associated with the disease, since rarely does the host live long enough for the parasites to attack the central nervous system. Domestic animals serve as reservoir hosts for both Gambian and Rhodesian trypanosomiasis. Native game animals are believed to serve as reservoirs for *T. rhodesiense*, but not for *T. gambiense*.

South American Trypanosomiasis (Chagas' Disease)

Trypanosoma cruzi is a parasite that lives in the blood and reticuloendothelial tissues of humans and many domestic and wild mammalian reservoir hosts, including dogs, cats, bats, raccoons, foxes, opossums, squirrels, monkeys and pigs. The geographic range of the parasite extends from southern parts of the United States through Mexico, Central and South America. Approximately 12 million persons are infected with *T. cruzi*. The principal vectors of *T. cruzi* are various reduviid bugs, including *Panstrongylus megistus*, *Triatoma infestans* and *Rhodnius prolixus*. The insects are notorious, nocturnal household pests, having a penchant for biting around the

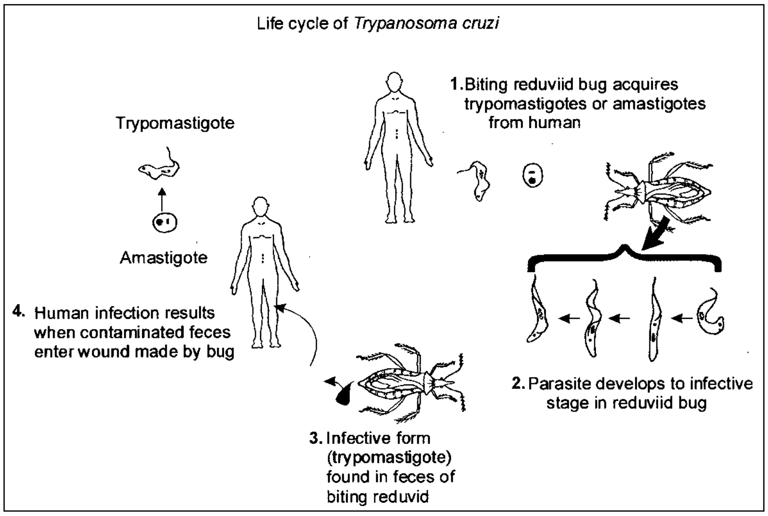
Life cycle of *Trypanosoma cruzi*
1. Biting reduviid bug acquires trypomastigotes or amastigotes from human
Trypomastigote
Amastigote
4. Human infection results when contaminated feces enter wound made by bug
2. Parasite develops to infective stage in reduviid bug
3. Infective form (trypomastigote) found in feces of biting reduvid

Figure 3. Life cycle of *Trypanosoma cruzi*. Modified from Belding, D. L. 1958. Clinical Parasitology. Appelton-Century-Crofts, Inc., New York.

eyes and lips of sleeping individuals. When feeding on the blood of infected vertebrates, the reduviids obtain the parasite either as free flagellates in the trypomastigote stage, or as intracellular amastigotes within host macrophages. With the blood meal the parasites pass first into the midgut of the insect where development transforms the flagellates into epimastigotes. The latter migrate to the hindgut where they are further transformed into infective or metacyclic trypomastigotes. The complete cycle in the insect requires about 2 weeks. Parasitized insects can retain an infection for several months (Fig. 3). While infected bugs feed, they defecate, voiding feces containing infective trypomastigotes. Human infection occurs when contaminated feces enters the skin through punctures made by the biting bugs, through skin abrasions, or through mucous membranes of the eye and mouth that are rubbed with contaminated fingers. Human infection may also occur through ingestion of the insect vector or its contaminated feces. The parasites also may be transmitted through the placenta and in breast milk. In endemic areas transmission may occur from infected donors during blood transfusions.

Entrance of the infective trypomastigotes initiates an acute local inflammation. During the early stages of infection, the trypomastigotes are abundant in the blood, but they do not undergo multiplication there. They eventually invade, and/or are engulfed by reticuloendothelial cells of the liver, spleen, and lymphatics, glial cells, and myocardial and skeletal muscles. Other tissues infected include nervous, gonadal, bone marrow and placenta. Within the various host cells, the trypanosomes rapidly transform into amastigotes that repeatedly multiply by binary fission producing numerous individuals. The parasites transform successively into promastigote, epimastigote, and trypomastigote stages, and are liberated when the destroyed host

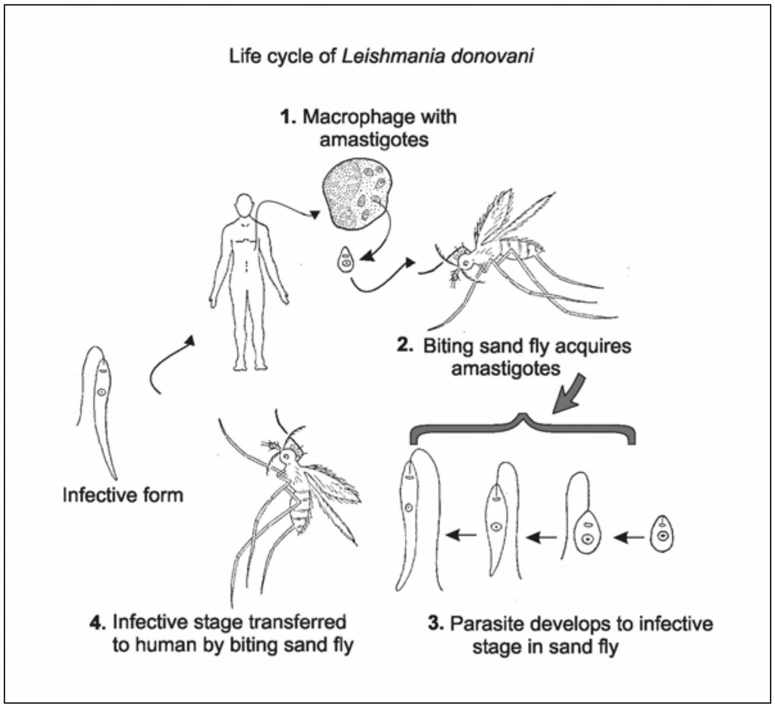
Life cycle of *Leishmania donovani*
1. Macrophage with amastigotes
2. Biting sand fly acquires amastigotes
3. Parasite develops to infective stage in sand fly
4. Infective stage transferred to human by biting sand fly
Infective form

Figure 4. Life cycle of *Leishmania donovani.* Modified from Belding, D. L. 1958. Clinical Parasitology. Appelton-Century-Crofts, Inc., New York.

cells rupture. The released trypomastigotes are infective to other host cells, as well as to the insect vectors. A generalized parasitemia accompanies the release of trypanosomes from host cells into the blood. Although almost every type of tissue is susceptible to invasion by *T. cruzi*, the flagellates demonstrate a preference for muscle and nerve tissues.

Chagas' disease appears in an acute stage primarily in young children and in a chronic form in adults. Frequently, early symptoms of the acute form appear as inflammatory swellings or nodules (chogomas) at the sites of the insect bite, unilateral edema of the eyelid and conjunctiva, and swelling of the pre-auricular lymph nodes (Romana's sign). Later, there is enlargement of the spleen, liver and lymphatic tissues, anemia, fever, and headache. Myocardial and neurological dysfunctions represent more severe manifestations of the chronic form of the disease. The heart becomes markedly enlarged and flabby. In endemic areas, the disease accounts for approximately 70% of the cardiac deaths in adults. Chagas' disease has been reported as the most important cause of myocarditis in the world. Additional manifestations include enlargement of the esophagus and colon, resulting in impaired peristalsis.

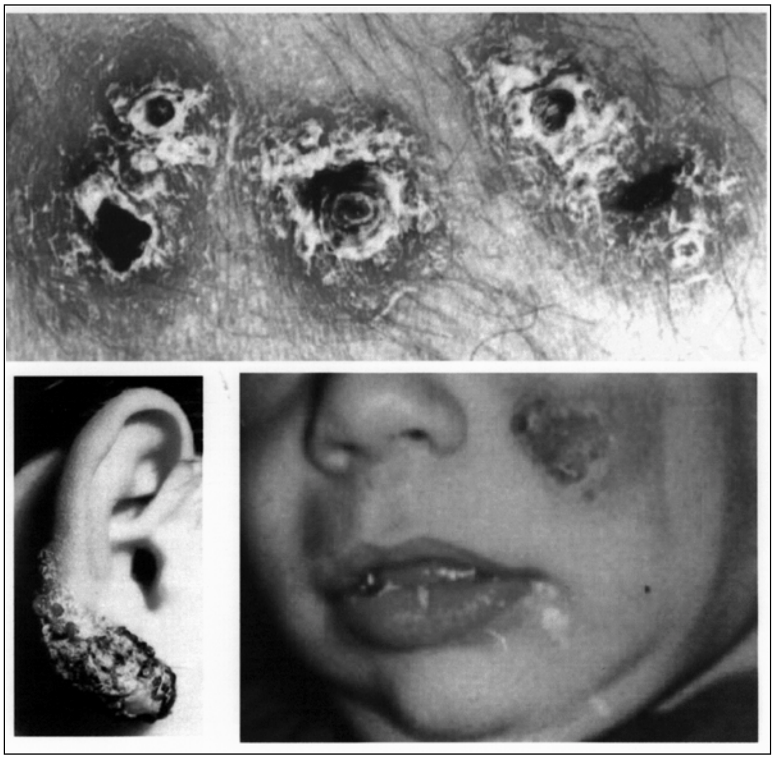

Figure 5. Examples of cutaneous Leishmaniasis caused by various species of *Leishmania*, including *L. tropica, L. mexicana* and *L. major.* Courtesy of Drs. Joseph El-On and Luis Weinrauch, Ben-Gurion University of Negev, Israel, and the Carlo Denegri Foundation, Torino, Italy.

Hemoflagellates: *Leishmania*

Genus Leishmania

Species of the Genus *Leishmania* occur in tropical and subtropical areas, where they are transmitted to humans and reservoir hosts (dogs, rodents) by female sand flies belonging to the genera *Phlebotomus* and *Lutzomyia*. The parasites occur in the amastigote stage within macrophages and reticuloendothelial cells of subcutaneous tissues, mucous membranes, liver, spleen and lymph nodes. Infected host cells rupture, liberating amastigotes that are engulfed by other phagocytes. When a sand fly sucks blood from an infected animal, amastigote forms are taken into the midgut where they transform into spindle-shaped promastigotes. The promastigotes multiply by binary fission, and migrate into the pharynx and buccal cavity from which they are injected into the skin of a vertebrate host when the fly again takes a blood meal. Mechanical transfer through the bite of stable flies (*Stomoxys calcitrans*) has been reported. Contact infection is possible when infected flies are crushed into the

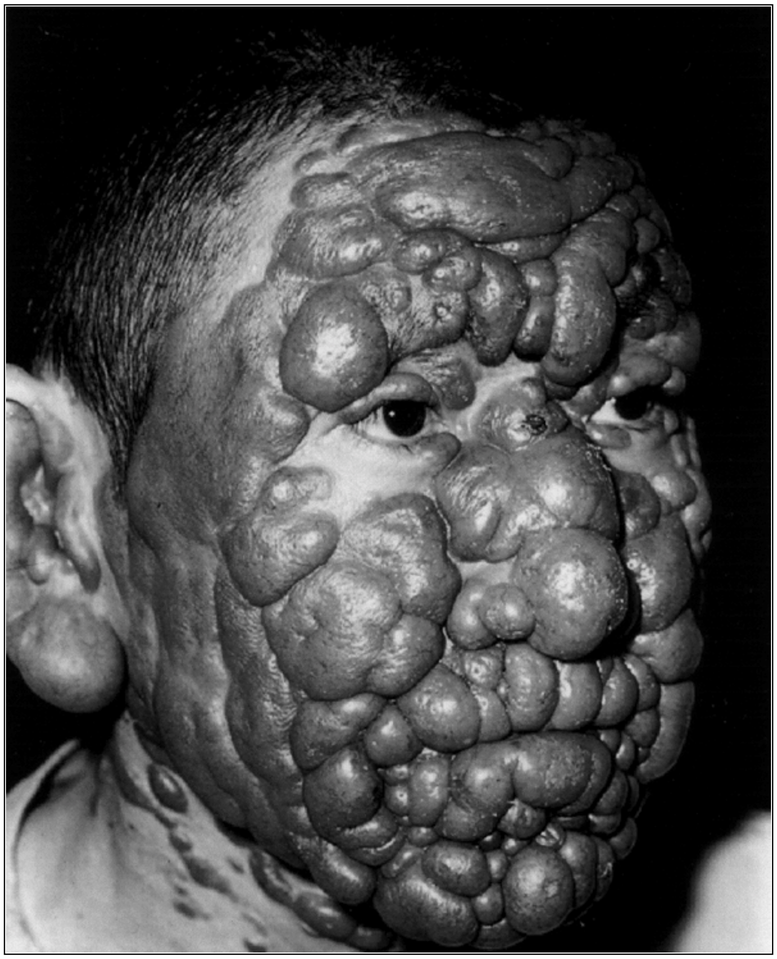

Figure 6. A severe dermal, post-visceral manifestation of "kala-azar" infection caused by *Leishmania donovani*. Photograph courtesy of Dr. Robert Kuntz.

skin or mucous membranes. Infection also may be possible by fecal contamination, since promastigotes have been found in the hindgut of some flies. After being introduced into the skin, the promastigotes are phagocytosed by macrophages, in which cells they undergo transformation to amastigotes, and begin to multiply. Heavily infected cells rupture, liberating amastigotes that are engulfed by other host macrophages, and parasite reproduction continues (Fig. 4).

The medically important species of *Leishmania* include *L. tropica*, *L. major*, *L. donovani*, *L. braziliensis* and *L. mexicana*. The parasites are morphologically indistinguishable and have virtually identical life cycles. They differ clinically and

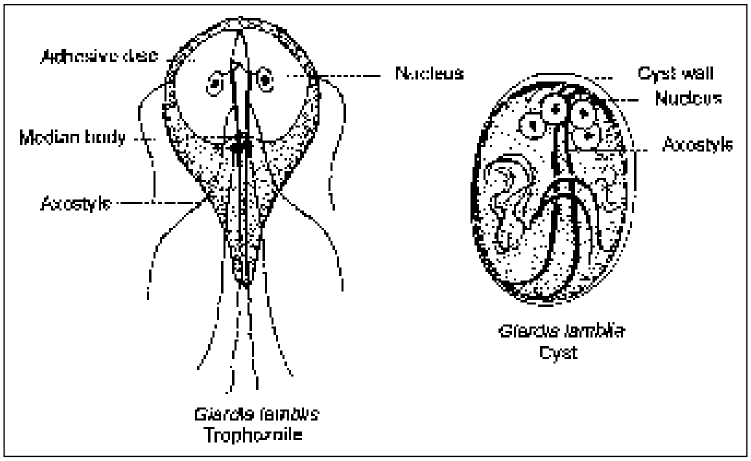

Figure 7. Diagrams of the trophozoite and cyst stages of *Giardia lamblia.*

serologically, but at times these characteristics overlap, thus species distinctions are not always clearly observed. *Leishmania tropica* and *L. major* are the etiological agents of cutaneous Leishmaniasis, also known as oriental sore, Jericho boil, Aleppo boil, or Delhi boil (Fig. 5). The disease occurs in Africa, the Middle East, southern Europe, Asia, India, Central and South America. The sand fly, *Phlebotomus papatasii*, is the important biological vector of cutaneous Leishmaniasis. After an extremely variable incubation period ranging from several weeks to three years, a small red sore or papule appears at the site of inoculation. Multiple sores may appear because of several infective bites or as a result of early contamination of other areas. The organisms may also disseminate within the host producing subcutaneous lesions of the face and appendages. Early papules gradually increase in size and become scaly. Ulceration occurs and spreads circularly. The lesion remains shallow, with a bed of granulation tissue, and surrounded by an area of red induration. The surrounding lymph nodes may become enlarged, especially if there is secondary bacterial infection. Rarely do the parasites infect adjacent mucocutaneous areas. Untreated infections gradually heal within several months to a year, leaving flattened and depigmented scars.

Leishmania donovani causes visceral Leishmaniasis also known as dum-dum fever or kala-azar. The flagellates infect cells of the reticuloendothelial system throughout the body. Infections occur primarily in the spleen, liver, bone marrow, and visceral lymph nodes. *Leishmania donovani* occurs in the Mediterranean region, southern Russia, China, India, Bangladesh, Africa, and Central and South America. The parasites are transmitted by various species of *Phlebotomus,* including *P. argentipes*, *P. longipalpis* and *P. orientalis*. The incubation period varies from a few weeks to eighteen months. The parasites initially colonize the dermis and later enter the blood, lymphatics and then the viscera where they are engulfed by macrophages. Typically, the liver and spleen become greatly enlarged (hepatosplenomegaly). There is an increased production of macrophages, decreased erythropoiesis, and thrombocytope-

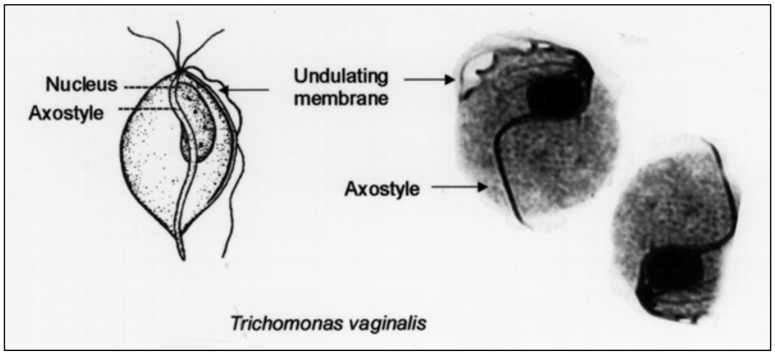

Figure 8. Diagram and stained preparations of *Trichomonas vaginalis*.

nia, which results in multiple hemorrhages. As the disease progresses there is a gradual loss of weight, the abdomen becomes swollen by the enlargement of the viscera. Other symptoms include edema, especially of the face, breathing difficulties, chills, fever, vomiting, and bleeding of the gums, lips, mucous membranes, and intestinal mucosa. In some individuals, there develops a post-kala-azar dermal leishmanoid condition, characterized in part by reddish, depigmented nodules in the skin (Fig. 6). The mortality in untreated cases may reach 95%. A fatal outcome is common in infected infants and young children. Death generally occurs within three years after infection.

Leishmania braziliensis is the etiological agent of a disease variously known as mucocutaneous Leishmaniasis, espundia or uta. The geographical range of the parasite is from Mexico to Argentina. The clinical manifestations of the disease, reservoir hosts, and species of sand flies involved in transmission, vary considerably from one location to another. Among several species of phlebotomine sand flies that serve as vectors are *Lutzomyia flaviscutellata*, *L. intermediua* and *L. tropidoi*. At the site of inoculation, a primary lesion, similar to oriental sore, occurs. This primary lesion heals within 6-15 months. Secondary lesions, characterized by epithelial hyperplasia, inflammation, and edema, may develop on the ear (chiclero ulcer), causing erosion of the earlobe cartilage. Secondary lesions may also occur in the mucous membranes of the mouth and nose (espundia), causing erosion of the lips, nasal septum, palatine tissues, pharynx, larynx, and trachea. The time of appearance of secondary lesions varies from before the primary lesions heal to many decades after infection. Death may result from secondary infections and/or respiratory complications.

Leishmania mexicana produces a disease with cutaneous, nasopharyngeal mucosal, and visceral manifestations. The cutaneous form of the disease is called “chiclero ulcer”, and is common in individuals harvesting gum from chicle trees. The parasite is found in Texas, Mexico, and parts of Central America. Sand flies of the genus Lutzomyia are biological vectors. The disease is a zoonosis with rodents as the principal reservoir host.

Flagellates of the Digestive and Reproductive Passages

Giardia lamblia is the most common flagellate of the human digestive tract. The parasite is cosmopolitan, but the disease, giardiasis, is more commonly found in children than in adults, and in individuals residing in warm climates rather than in cold climates. In some areas in the United States the incidence of infection may be as high as 20% of the population. The pathogen has both trophozoite and cystic stages in its life cycle (Fig. 7). The trophozoites are confined essentially to the duodenum, but occasionally invade the bile ducts. The trophozoite is 'tear-drop shaped', with a convex dorsal surface and a concave ventral surface ('adhesive disc') which makes contact with the intestinal mucosa. The trophozoite possesses two nuclei, and four pairs of flagella. In heavy infections, the intestinal mucosa may be carpeted with the parasites. A single diarrhetic stool from a heavily infected individual may contain several billion parasites. The flagellates penetrate into mucosal cells and also interfere with the absorption of fat and fat soluble vitamins. Heavy infections may be characterized by extensive ulcerations of the intestinal mucosa. Biliary disease sometimes occurs when flagellates pass up the bile duct. The trophozoites multiply by longitudinal binary fission in the small intestine and eventually encyst. Mature cysts, which are tetranucleate, are found in stools. Infection of new hosts occurs when mature cysts are ingested with contaminated food or water. Following excystation in the duodenum, the tetranucleate parasite undergoes cytokinesis forming two binucleate daughter trophozoites, which then adhere to epithelial cells and feed. Symptoms of this highly contagious disease include diarrhea, abdominal pain, the passing of blood and mucus, hypoproteinemia with hypogammaglobulinemia, fat-soluble vitamin deficiencies, and the production of copious light-colored fatty stools.

Trichomonas vaginalis is a cosmopolitan parasite that resides in the male and female urogenital tracts. Transmission of the infective trophozoite stage is chiefly by sexual intercourse, and because of its potentially pathogenic nature, the disease is regarded as a serious venereal disease (Fig. 8). Transmission may occur from female to female through contaminated clothing or toilet facilities. The parasite has been found in newborn infants. In males, infection is frequently asymptomatic, but severe symptoms involve not only the urethra (urethritis) and bladder, but also the genital organs and glands, including the prostate. A discharge from the urethra containing the flagellates may occur. Although the vagina is most commonly infected (vaginitis), the trichomonads may spread to all parts of the urogenital tract of the female. A frothy, creamy discharge is frequently observed in infected females. The disease may be complicated by concurrent fungal, bacterial, or spirochetal infections.

Some Nonpathogenic Flagellates

Trichomonas tenax is a nonpathogenic species confined to the mouth, especially in pyorrheal pockets and tarter along the gumline, and in tonsillar crypts. Transmission of trophozoites may result from kissing or the use of common drinking or eating utensils. Other nonpathogenic intestinal flagellates include *Dientamoeba fragilis*, *Chilomastix mesnili*, *Retortamonas intestinalis*, *Enteromonas hominis* and *Pentatrichomonas hominis*. Frequently, the presence of these nonpathogenic forms is an indication of direct fecal contamination.

CHAPTER 3

Sarcodina

Amoebae belong to the subphylum Sarcodina, a group comprised mostly of free-living organisms living in a wide variety of habitats. Symbiotic amoebae commonly found in the intestinal tract of humans and domestic animals belong to the genera *Entamoeba, Endolimax* and *Iodamoeba* (Fig. 1). Species infecting extraintestinal sites include members of the genera *Naegleria, Hartmannella* and *Acanthamoeba*. The life cycle may involve just a motile, feeding trophozoite stage, or both trophozoite and cyst stages. In most members the trophozoite is amoeboid only, moving by means of pseudopodia. In some life cycles, both flagellate and amoeboid trophozoites are represented. Reproduction involves mitotic divisions of the nucleus followed by binary fission. Cysts may be ingested, or possibly even inhaled. Excystation generally occurs in the small intestine of the host.

Entamoeba histolytica is a pathogenic, tissue-invading amoeba. Perhaps 80-95% of the one-half billion human infections by this parasite are symptomatic. The highest incidence of the disease, termed amoebiasis, occurs in persons living in warm climates, in rural areas where poor sanitary conditions exist, and in crowded institutions such as prisons and asylums. As with most enteric diseases, amoebiasis is commonly associated with sewage contamination of drinking water. The virulence of *E. histolytica* varies considerably. Unfortunately, the factors which determine the virulence of the parasite are not completely understood. Dietary deficiencies appear to influence the incidence and severity of infection.

Infection occurs by the ingestion of mature cysts of the parasite in contaminated food or drinking water, and by hand-to-mouth contact (Fig. 2). Flies and cockroaches may mechanically transport the cysts and contaminate food and eating utensils. Following excystation in the ileum, the emerging trophozoites, occasionally referred to as metacysts, multiply by binary fission and adhere to the intestinal mucosa. They lodge in the crypts of the lower portions of the small intestine and in the large intestine, where some invade the mucosal epithelium by elaborating a proteolytic enzyme which lyse the cells. The parasites may invade deep into the wall of the intestine, feeding, eroding tissues, and forming ulcers. The majority of intestinal lesions occur in the cecal and sigmoid-rectal areas. Invading amoebae may enter mesenteric venules or lymphatics and be transported to the liver, lungs, brain, and other extraintestinal organs where they continue feeding, causing severe lesions and tissue necrosis. Some of the intestinal trophozoites do not invade the gut wall but instead form cysts (a process termed encystation) which are later passed outside the host. Immature cysts contain one or two nuclei; mature cysts contain four nuclei. Immature cysts are able to mature in the external environment and become infective. Trophozoites passed in the feces are unable to encyst outside the host and are not infective.

Based on the anatomic site of infection and on clinical manifestations, two major types of symptomatic amoebiases are recognized: (1) Intestinal or primary amoebiasis, including both dysenteric and non-dysenteric forms; and, (2) extraintestinal or

---

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

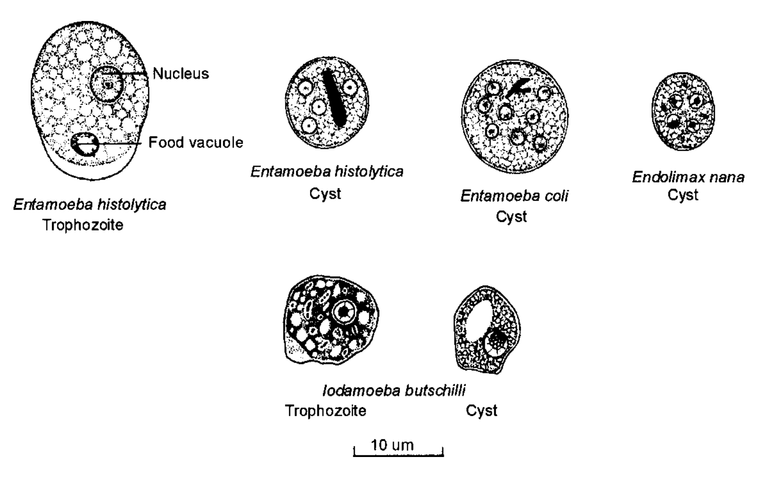

Figure 1. Drawings of representative amoebae.

secondary amoebiasis, including hepatic, pulmonary, cutaneous, cerebral, coronary and urogenital sites. Symptomatic intestinal infections of *E. histolytica* are termed amoebic colitis. Common, relatively mild symptoms include diarrhea, dysentery (blood and mucus in stools), abdominal discomfort, flatulence (gas in stomach or intestine), anorexia (loss of appetite), and loss of weight. The pathology of intestinal amoebiasis is destruction of the intestinal epithelium and invasion of the gut wall by the trophozoites. The initial invasion by the parasites into the intestinal mucosa may initiate an inflammatory response. The perforation of amoebic ulcers and resulting peritonitis characterize moderately severe infections. In the liver, single or multiple abscesses may occur, and the liver frequently becomes enlarged. Extensive invasion of the liver results in the destruction of parenchymal tissue by cytolysis. Hepatic infection may spread through the diaphragm to the lungs (amoebic pneumonitis) and respiratory passages where abscess formation frequently follows. Abscess formation in these areas is frequently accompanied by the infiltration of leukocytes and fibroblasts.

Naegleria fowleri causes a serious disease in humans termed primary amoebic meningoencephalitis. The parasite rapidly invades and multiplies in the brain and meninges causing extensive hemorrhage and tissue destruction. Infections, which are nearly all fatal, have been reported from New Zealand, Australia, East Africa, Europe, and America. The disease resembles bacterial meningitis. The first signs of infection are mild fever, headache, nausea, vomiting, and in some cases, a sore throat. Later, as the headache and fever persist and increase in severity, the patients become disoriented and usually comatose. Death usually occurs within seven days of the onset of the symptoms. The infective trophozoite stage is diphasic in that it exists not only as an amoeba, but under certain conditions, such as when transferred from a culture medium to distilled water, it develops two flagella to become a flagellate. In

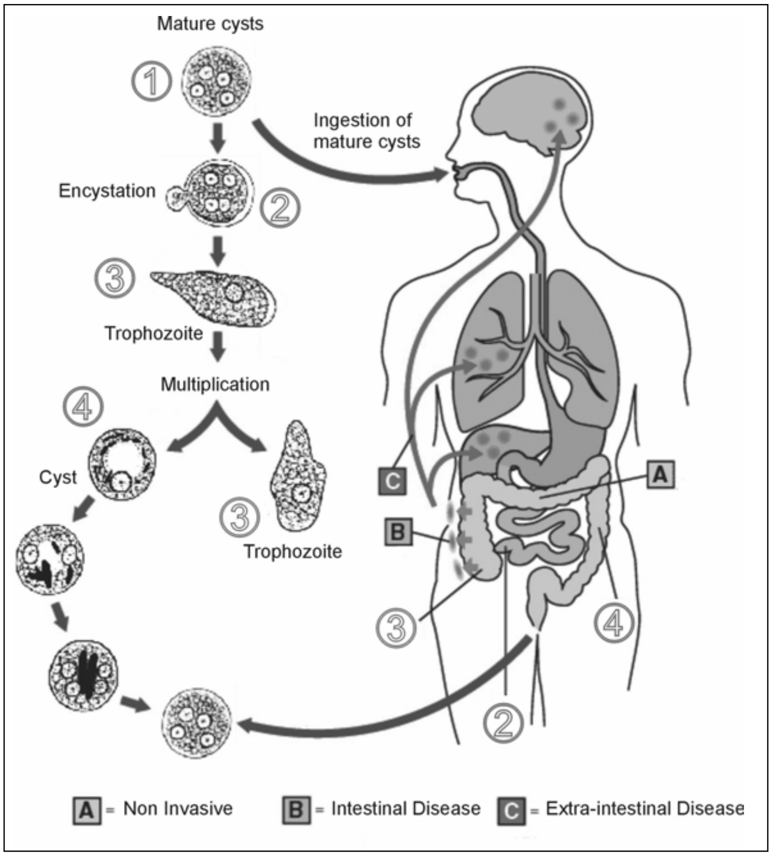
Mature cysts
①
Ingestion of mature cysts
Encystation
②
③
Trophozoite
Multiplication
④
Cyst
③
Trophozoite
C
B
A
③
④
②
A = Non Invasive
B = Intestinal Disease
C = Extra-intestinal Disease

Figure 2. Life Cycle of *Entamoeba histolytica*. Following the ingestion of cysts (1), excystation occurs in the small intestine (2). The trophozoites (3) may remain in the intestinal lumen of individuals (A), who then become asymptomatic carriers that can disseminate the cysts, or the trophozoites invade the intestinal mucosa (B), enter the cisrculation and establish extraintestinal infections (C). Courtesy of the Centers for Disease Control, Division of Parasitic Disease, Atlanta, Georgia.

human tissues, the parasite is present only in the amoeboid phase. A resistant cyst stage is formed only in laboratory cultures. Intracerebral infection is generally acquired while the individual is swimming in fresh or brackish water, with the mode of entry into the brain and meninges through the olfactory mucosa and cribiform plate. Apparently, some infections are acquired by inhaling airborne trophozoites (Fig. 3).

Species of *Hartmannella* and *Acanthamoeba* also invade the central nervous system. Members of these two genera are similar morphologically to the amoeboid stage of *Naegleria*, but do not possess flagella at any stage in their developments. *Hartmannella* has been observed in the respiratory tract of humans. *Acanthamoeba*

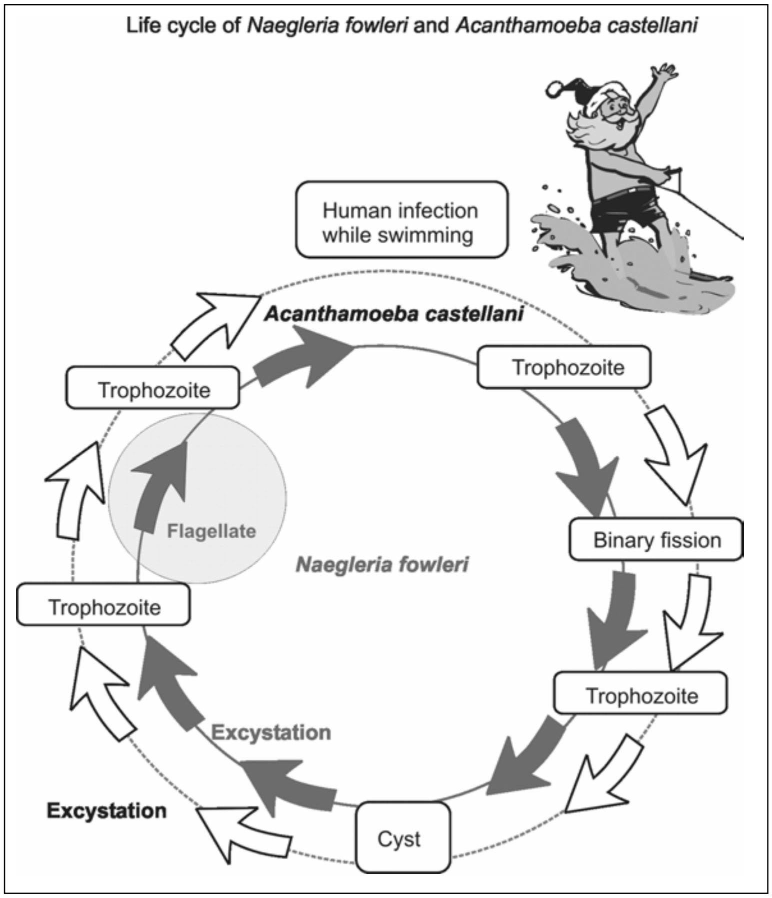
Life cycle of *Naegleria fowleri* and *Acanthamoeba castellani*
Human infection while swimming
Acanthamoeba castellani
Trophozoite
Trophozoite
Binary fission
Flagellate
Naegleria fowleri
Trophozoite
Trophozoite
Excystation
Excystation
Cyst

Figure 3. Comparison of the life cycles of *Naegleria fowleri* and *Acanthamoeba castellani*.

is believed to cause ulcerations of the cornea, resulting in blindness, and lesions and granulomatous formations of the skin, especially in immune compromised individuals (Fig. 4). Cysts of *Acanthamoeba* have been found in cases of meningoencephalitis.

Entamoeba gingivalis is a common commensal found in the mouth, especially in the gingival areas, in the tartar near the base of the teeth, and in the crypts of the tonsils. No cysts are formed by *E. gingivalis*; only the trophozoite stage has been described. Transmission is by droplet spray, intimate oral contact, or by the sharing of eating utensils. Although *E. gingivalis* is frequently associated with peridontitis, there is no evidence that the organism is pathogenic.

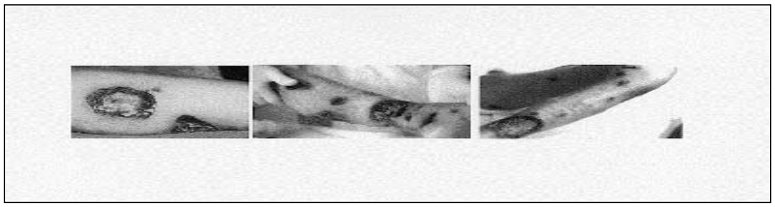

Figure 4. Cutaneous acanthamoeba infection in a 35 year-old individual with AIDS and previous episodes of *Pneumocystis carinii* pneumonia and toxoplasmic encephalitis. The lesions are initially nodular and reddish, but later become ulcerative, necrotic and more numerous. Courtesy of Drs. Antonio Macor, Ezio Nigra, Romina Ruffatto, Alberto Pisacane and M.L. Soranzo, at the Amedeo di Savoia Hospital, Torino, Italy and the Carlo Denegri Foundation, Torino, Italy.

Entamoeba coli, E. hartmanni, E. polecki, Iodamoeba butschlii and *Endolimax nana* are all regarded as nonpathogenic intestinal amoebae. Their life cycles are similar, with infection occurring via excystation in the intestine. Although generally harmless, the presence of the organisms indicates ingestion of contaminated food or water.

Ciliate Parasites

Balantidium coli is the only ciliate parasite of humans. It is cosmopolitan, inhabiting the large intestine, cecum, and terminal ileum, feeding on bacteria. The parasite may also invade the intestinal mucosa causing ulceration. Extraintestinal spread to the liver, lungs, and urogenital tract is rarely observed. Infection occurs by ingestion of cysts in contaminated food or drink. Less than one percent of the human population is infected with *B. coli*. Pigs represent the usual source of infection for humans. Symptoms of the disease, which is termed balantidiasis, may range from mild colitis and diarrhea to clinical manifestations resembling severe amoebic dysentery.

CHAPTER 4

Apicomplexa: Sporozoa and Piroplasmea

Introduction to Sporozoa

All members of the Class Sporozoa are parasites. The Class consists of two subclasses, Gregarina and Coccidia. The Gregarina are parasites of the intestinal tract, body cavity, and reproductive organs of invertebrates only. The coccidians live in the intestinal mucosa, liver, reticuloendothelial cells, blood cells, and other tissues of vertebrates and invertebrates. The latter group includes parasites causing serious coccidioses of domestic animals, and malarias of humans and other vertebrates. The major species parasitizing humans belong to the genera *Isospora, Toxoplasma, Cryptosporidium, Babesia, Pneumocystis* and *Plasmodium*.

Coccidians

In many coccidians, reproduction is both asexual, by either binary or multiple fission, and sexual, by either isogamous or anisogamous union. Spore formation frequently follows sexual reproduction. The infective stage is the sporozoite. Transmission from host to host is accomplished either by vectors, usually biting arthropods, which transmit the infective sporozoites directly, or by ingestion of highly resistant cysts (oocysts) with internal spores (sporocysts) containing the infective sporozoites.

The typical Coccidian life cycle pattern consists of three developmental phases: sporogony, schizogony, and gamogony (Fig. 1). Different morphological types result from each developmental phase; sporozoites from sporogony, merozoites from schizogony, and zygotes from gamogony. The only diploid stage in the life cycle is the zygote. Following meiosis, the zygote becomes a multinucleate sporont and undergoes multiple fission (sporogony). Depending on the species, the haploid cells formed are either free infective sporozoites or sporozoites enclosed within an oocyst or spore, from which they subsequently escape. Within the host, infective sporozoites become intra- or extracellular vegetative trophozoites. The trophozoites develop first into multinucleate schizonts, which then undergo multiple fission (schizogony) producing numerous daughter cells or merozoites. Merozoites either continue the infection of the host with repetition of asexual multiplication and invasion of other host cells, or they develop into sexual stages termed gamonts. Gamonts undergo multiple fission (gamogony) and give rise to gametes. The latter join in pairs in syngamy to produce diploid zygotes.

Variations in the basic life cycle pattern among different species result from the absence of any one phase. When sporogony is omitted, the zygotes develop directly into sporozoites, and when gamogony is omitted, merozoites develop into gametocytes directly. In monoxenous coccidians, all stages in the life cycle occur in a single host, with the oocyst eliminated from infected individuals. In heteroxenous forms,

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

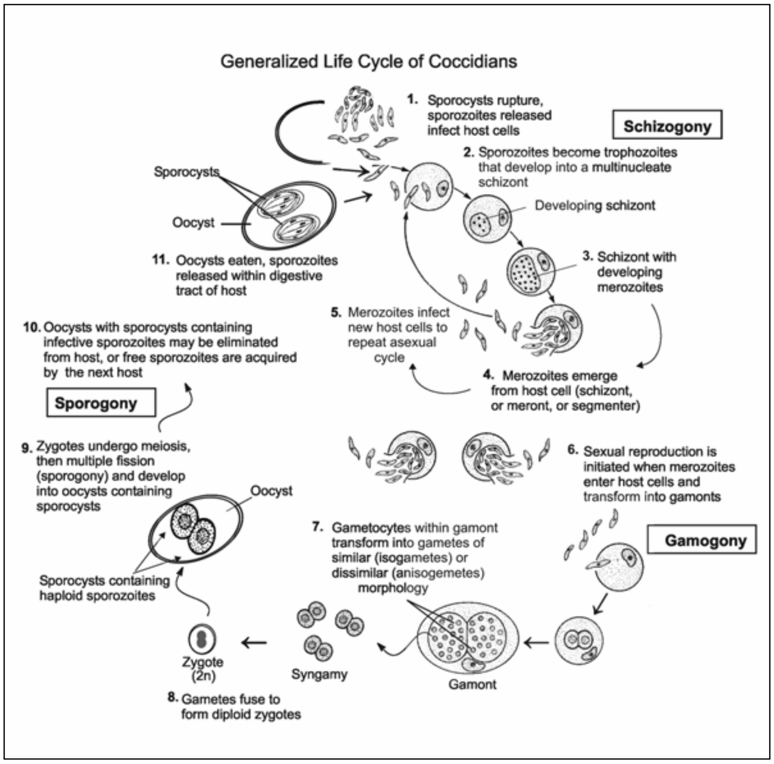
Generalized Life Cycle of Coccidians
1. Sporocysts rupture, sporozoites released infect host cells
Schizogony
2. Sporozoites become trophozoites that develop into a multinucleate schizont
Developing schizont
3. Schizont with developing merozoites
4. Merozoites emerge from host cell (schizont, or meront, or segmenter)
5. Merozoites infect new host cells to repeat asexual cycle
6. Sexual reproduction is initiated when merozoites enter host cells and transform into gamonts
Gamogony
7. Gametocytes within gamont transform into gametes of similar (isogametes) or dissimilar (anisogemetes) morphology
Gamont
Syngamy
Zygote (2n)
8. Gametes fuse to form diploid zygotes
Sporocysts containing haploid sporozoites
Oocyst
9. Zygotes undergo meiosis, then multiple fission (sporogony) and develop into oocysts containing sporocysts
Sporogony
10. Oocysts with sporocysts containing infective sporozoites may be eliminated from host, or free sporozoites are acquired by the next host
11. Oocysts eaten, sporozoites released within digestive tract of host
Sporocysts
Oocyst

Figure 1. Developmental stages in the life cycle of coccidians.

sporogony occurs in an invertebrate host, and following the transfer of sporozoites by biting vectors, schizogony and a part of gamogony occur in the vertebrate host.

Genus Isospora

Isospora belli is a parasite of the intestinal mucosa of humans. Although widely distributed, *I. belli* is more prevalent in warm climates than in cool climates. The life cycle of this species is believed to be similar to those of *I. canis* in dogs and *I. felis* in cats. Infection occurs when mature or sporulated oocysts are ingested. The freed sporozoites invade the intestinal mucosa where all three developmental phases occur. Asexual stages terminate with the production of oocysts, which are eliminated in the feces. The oocysts of *Isospora* contain two sporocysts, and within each sporocyst there are four infective sporozoites.

Apparently, the majority of human infections are without adverse manifestations, but symptoms ranging from mild gastrointestinal discomfort, nausea, and anorexia, to severe dysentery or diarrhea have been reported. The stools are often loose and pale yellow, indicating, as in giardiasis, the inability of the patient to adequately absorb fat. Infections may last up to four months.

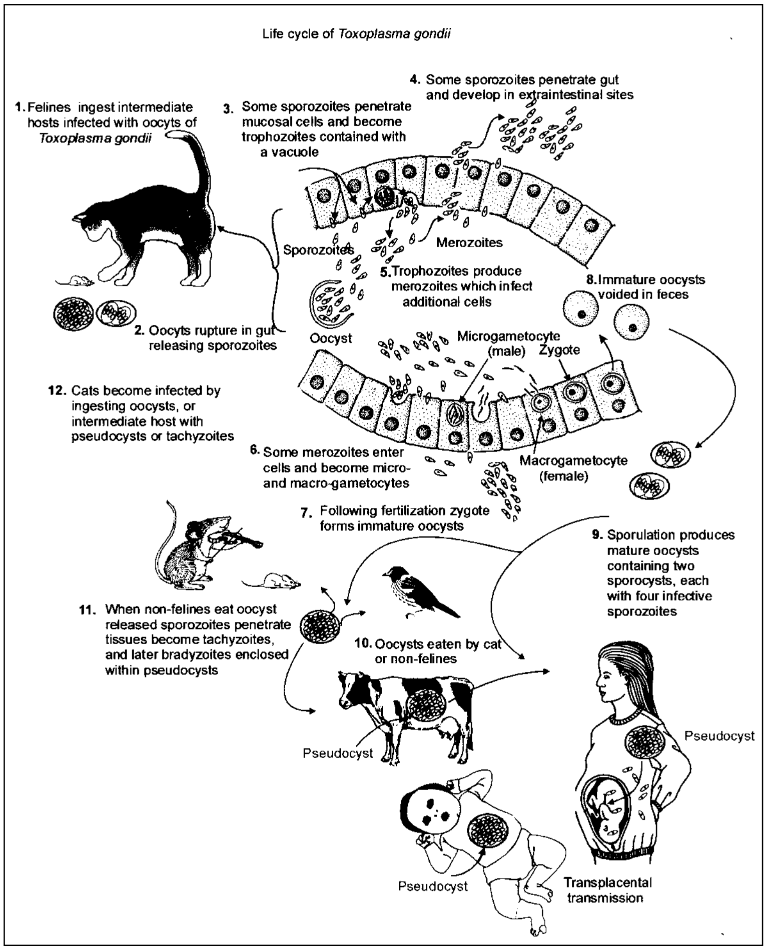
Life cycle of *Toxoplasma gondii*
1. Felines ingest intermediate hosts infected with oocyts of *Toxoplasma gondii*
2. Oocyts rupture in gut releasing sporozoites
3. Some sporozoites penetrate mucosal cells and become trophozoites contained with a vacuole
4. Some sporozoites penetrate gut and develop in extraintestinal sites
Sporozoites
Merozoites
5. Trophozoites produce merozoites which infect additional cells
Oocyst
Microgametocyte (male)
Zygote
8. Immature oocysts voided in feces
12. Cats become infected by ingesting oocysts, or intermediate host with pseudocysts or tachyzoites
6. Some merozoites enter cells and become micro- and macro-gametocytes
Macrogametocyte (female)
7. Following fertilization zygote forms immature oocysts
9. Sporulation produces mature oocysts containing two sporocysts, each with four infective sporozoites
11. When non-felines eat oocyst released sporozoites penetrate tissues become tachyzoites, and later bradyzoites enclosed within pseudocysts
10. Oocysts eaten by cat or non-felines
Pseudocyst
Pseudocyst
Pseudocyst
Transplacental transmission

Figure 2. Life cycle of *Toxoplasma gondii.* Modified from Schmidt, G. D. and Roberts, L. S. 1996. Foundations of Parasitology. Wm. C. Brown Publishers.

Genus Toxoplasma

Toxoplasma gondii is a cosmopolitan parasite of a variety of mammals and birds. The domestic cat and other felines are definitive hosts, and a large number of mammals, including humans, and birds serve as potential intermediate hosts. Animals become infected when they ingest the infective stages of the parasite, which include mature (sporulated) oocysts containing sporozoites, pseudocytes containing bradyzoites, or free tachyzoites. The course of development of the parasite following its entry into the small intestine varies, and is dependent on whether a definitive or an intermediate host is infected. In cats and other felines parasite development involves an enteroepithelial cycle and an extraintestinal cycle. There is no enteroepithelial cycle in non-feline animals, which may serve as intermediate hosts (Fig. 2).

In the enteroepithelial cycle, some of the sporozoites liberated from oocysts in the lumen of the intestine of the cat penetrate the mucosal cells, become enclosed in a parasitophorous vacuole, and transform into trophozoites. The trophozoites reproduce asexually by schizogony producing numerous merozoites, which in turn infect other cells to repeat the asexual cycle. Eventually some merozoites undergo gamogony and give rise to sexual stages, (i.e., microgametocytes and macrogametocytes). Intracellular fertilization occurs when microgametes released from one cell enter other host cells containing macrogametes. Following fertilization, the resulting zygotes develop into immature oocysts that are voided with the feces. Peak oocyst production generally occurs approximately one week post-infection. Outside the host, sporulation occurs within one to five days, producing mature oocysts containing two sporocysts, each enclosing four infective sporozoites.

In the extraintestinal cycle, some of the sporozoites liberated in the small intestine enter the blood and are disseminated to various regions of the body, including the mesenteric lymph nodes, lungs, spleen, voluntary and cardiac muscles, retina, and cells of the nervous system. In these areas, the sporozoites invade cells and transform into rapidly dividing merozoites called tachyzoites. In acute infections there may be several cycles of cellular invasion, parasite multiplication, and host cell destruction. As the disease becomes chronic, infection is manifested by the formation of pseudocysts that envelope numerous parasites that are referred to as bradyzoites. Pseudocysts with infective bradyzoites may last for years in nerve tissue. Presumably pseudocyst formation is induced by host immune responses. As immunopathology decreases, the parasites are released from the cysts and infect other host cells. Thus, an alternation of proliferative and cystic phases occurs in infected intermediate hosts. Cats and other carnivorous animals readily acquire infections when they ingest oocysts or consume prey whose tissues contain pseudocysts or infective free parasites.

Human toxoplasmosis may result from either contact with cat feces contaminated with mature oocysts, or from ingesting infected meat or milk. Meat of various domestic animals may be infected, including pork, mutton, beef and poultry. Only asexual development of the parasite occurs in humans. Merozoites arising from asexual development enter the blood and lymphatics and form intracellular cysts in various tissues of the body. Symptoms vary greatly, with the majority of human infections benign or asymptomatic. The most serious are transplacental and neonatal infections. Clinical manifestations include hepatosplenomegaly, pneumonitis, retinochoroiditis, encephalomyelitis, cerebral calcification, hydrocephalus or microcephaly. Rarely is there complete recovery in infected children. The symptoms of postnatally acquired toxoplasmosis are frequently mild and may mimic those of infectious mononucleosis, with chills, fever, fatigue, headache, lymphadenitis, and myalgia (muscle pain).

It is estimated that approximately 13% of the world population is infected with *T. gondii.* In some European countries where raw meat commonly is eaten, the prevalence of human toxoplasmosis may be as high as 50%. In the United States approximately 4,000 infants are born annually with toxoplasmosis.

Genus Plasmodium

Species belonging to the genus *Plasmodium* are causative agents for malaria of humans and other animals in various tropical and subtropical regions of the world. The life cycle of these parasites involves an asexual phase (schizogony) alternating with a sexual one (gametogony), followed by sporogony. About forty species of female anopheline mosquitoes are definitive hosts, in which sexual reproduction occurs. Male mosquitoes do not feed on blood, and thus play no direct role in malarial transmission. Asexual reproduction occurs in the tissues of vertebrates, which are thus considered intermediate hosts. There are four species of malarial parasites infecting humans: *Plasmodium falciparum, P. vivax, P. malariae* and *P. ovale*.

Plasmodium falciparum is the cause of malignant tertian malaria also known as aestivo-autumnal or subtertian malaria. The parasite causes 50% of the malarial cases world wide. Falciparum malaria often terminates fatally. *Plasmodium vivax*, which causes benign tertian malaria, has the widest geographical distribution of the four species infecting humans. Indigenous cases of malaria occur as far north as England, Siberia, and Manchuria, and south into Argentina and South Africa. *Plasmodium malariae* is a relatively uncommon parasite producing quartan malaria. It is prevalent in tropical regions except South America. The parasite causes 7% of all malaria in the world. *Plasmodium ovale* is a rarely encountered species, but it has been reported from tropical and subtropical regions of many continents.

General Life Cycle of Plasmodium

A general account of the development of malarial parasites may be divided into three phases (Fig. 3):

1. Pre-erythrocytic or exoerythrocytic schizogony. This stage involves the development of merozoites from sporozoites introduced into humans when infected female mosquitoes bite. Within 30 minutes after infection, the sporozoites disappear from the peripheral circulation and invade parenchymal cells of the liver. Within the liver cells the parasites first develop into vegetative trophozoites. The parasites undergo one (*P. falciparum*) or two schizogonic cycles producing numerous exoerythrocytic merozoites. The merozoites escape from the liver cells and pass into the blood stream to invade erythrocytes. In *P. vivax, P. malariae* and perhaps *P. ovale*, exoerythrocytic trophozoites and merozoites may persist for years, causing relapse by producing numerous intermittent invasions of the bloodstream, especially in cases of declining immunity or ineffective drug therapy. Relapses do not occur in human infections with *P. falciparum* since exoerythrocytic multiplication is limited to a single generation of merozoites before invasion of the blood (Figs. 4A and 4B).
2. Erythrocytic schizogony. Merozoites that pass from the liver into the bloodstream invade erythrocytes and undergo schizogony. Within the red blood cells, growth of the parasite proceeds through the trophozoite and erythrocytic schizont stages. Infected erythrocytes enlarge as the intracellular parasites multiply. Eventually, the schizonts consist of numerous, fully formed merozoites. An erythrocytic schizont consisting of mature merozoites is termed a meront or segmenter. The segmenters

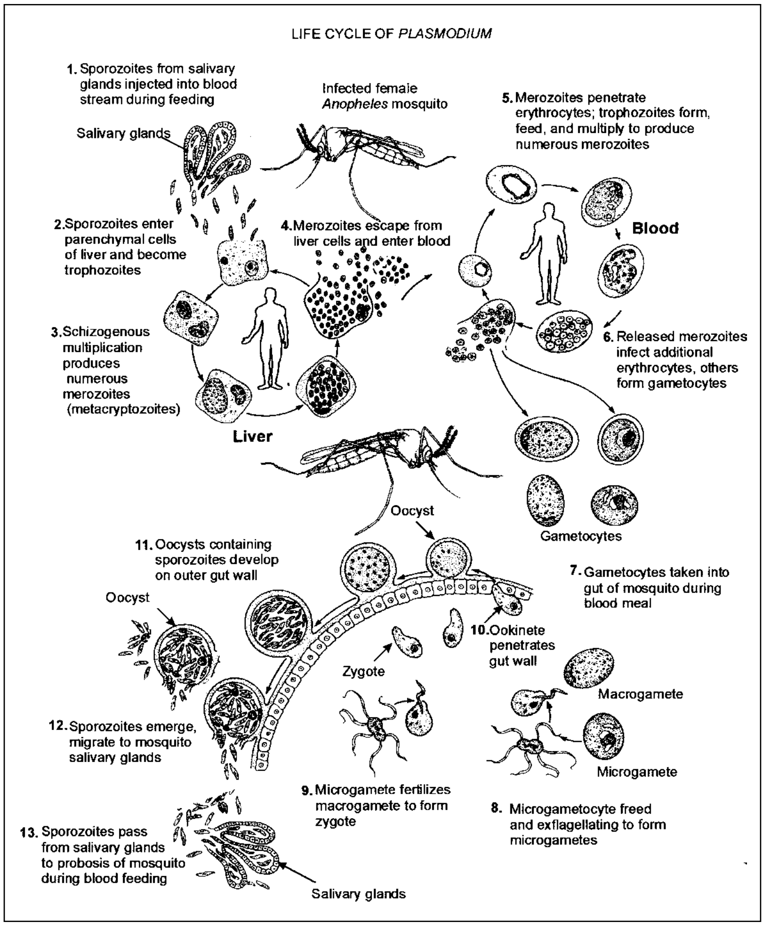
LIFE CYCLE OF PLASMODIUM
1. Sporozoites from salivary glands injected into blood stream during feeding
Salivary glands
Infected female Anopheles mosquito
5. Merozoites penetrate erythrocytes; trophozoites form, feed, and multiply to produce numerous merozoites
2. Sporozoites enter parenchymal cells of liver and become trophozoites
4. Merozoites escape from liver cells and enter blood
Blood
3. Schizogenous multiplication produces numerous merozoites (metacryptozoites)
6. Released merozoites infect additional erythrocytes, others form gametocytes
Liver
Oocyst
Gametocytes
11. Oocysts containing sporozoites develop on outer gut wall
7. Gametocytes taken into gut of mosquito during blood meal
Oocyst
10. Ookinete penetrates gut wall
Zygote
Macrogamete
12. Sporozoites emerge, migrate to mosquito salivary glands
Microgamete
9. Microgamete fertilizes macrogamete to form zygote
8. Microgametocyte freed and exflagellating to form microgametes
13. Sporozoites pass from salivary glands to probosis of mosquito during blood feeding
Salivary glands

Figure 3. Life cycle of *Plasmodium*.

rupture almost synchronously, liberating the erythrocytic merozoites, which then invade other red blood cells repeating the asexual cycle. The cycle of schizogonic development and liberation of merozoites occurs typically every 48 hours in *P. vivax* and *P. ovale* infections, every 72 hours in *P. malariae* infections, and about every 36 to 48 hours in *P. falciparum* infections. The first visible manifestations of the disease occur with the synchronized lysis of red blood cells and the release of merozoites into the bloodstream, causing malarial paroxysms consisting of chills, burning fever, followed by sweating. Eventually, some erythrocytic merozoites develop into micro- and macrogametocytes instead of asexual schizonts.

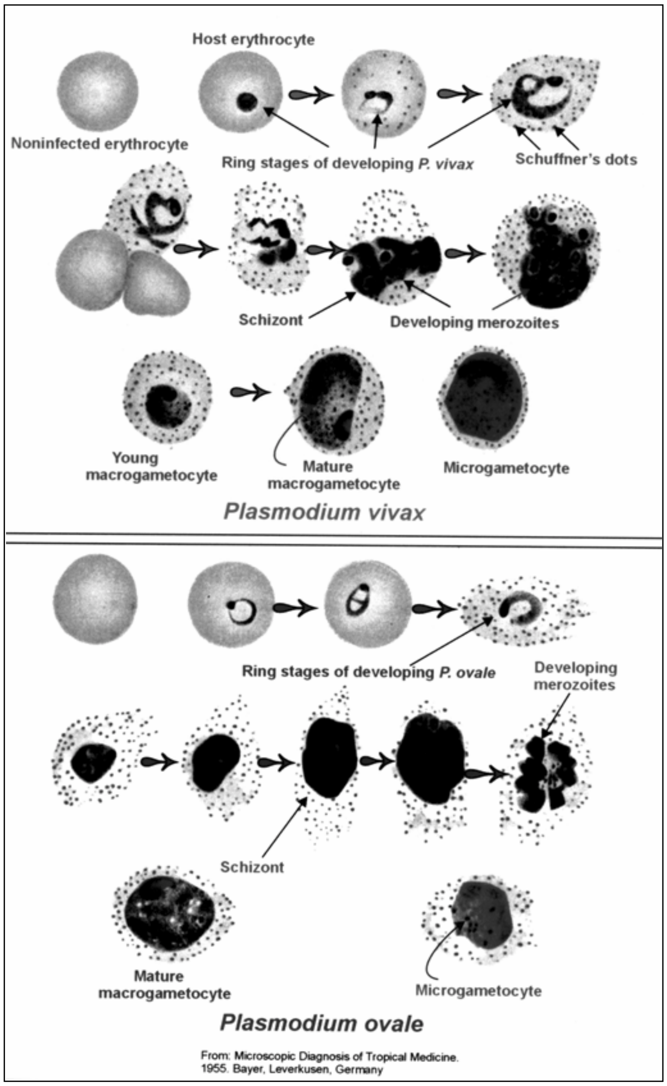
Host erythrocyte
Noninfected erythrocyte
Ring stages of developing P. vivax
Schuffner's dots
Schizont
Developing merozoites
Young macrogametocyte
Mature macrogametocyte
Microgametocyte
Plasmodium vivax
Ring stages of developing P. ovale
Developing merozoites
Schizont
Mature macrogametocyte
Microgametocyte
Plasmodium ovale

From: Microscopic Diagnosis of Tropical Medicine.
1955. Bayer, Leverkusen, Germany

Figure 4A. Intracellular developmental stages of *Plasmodium.* From Microscopic Diagnosis of Tropical Medicine. 1955. Bayer, Leverkusen, Germany.

Unfortunately, very little is known of the factors influencing gametogenesis. These immature gametocytes typically are present in circulating red blood cells after a few to several erythrocytic schizogonic cycles. They remain as gametocytes within the erythrocytes and do not mature further

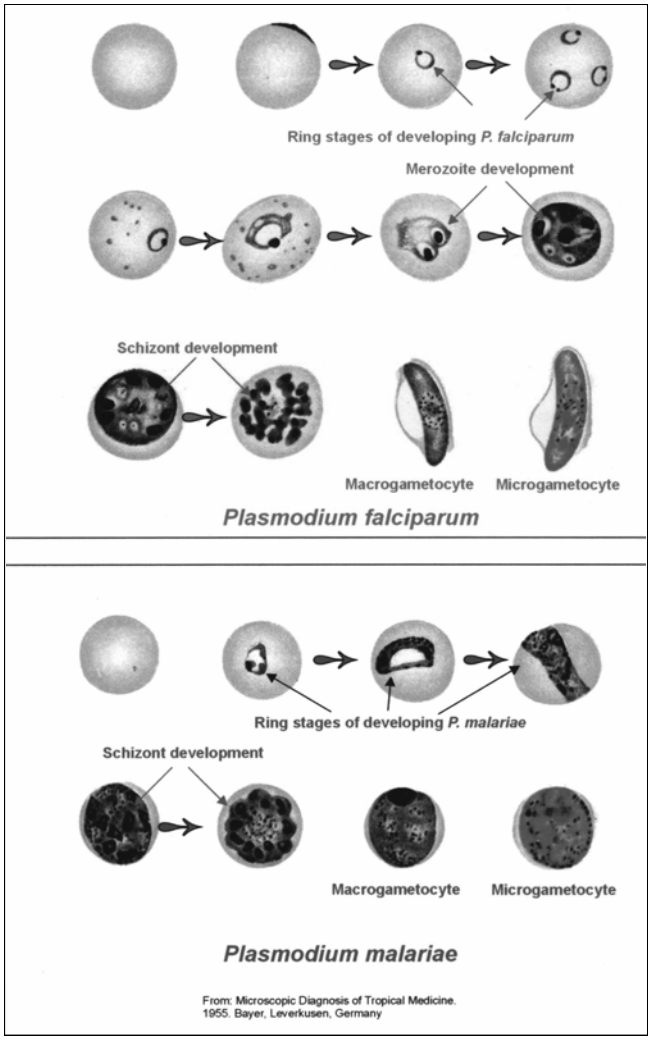
Ring stages of developing P. falciparum
Merozoite development
Schizont development
Macrogametocyte
Microgametocyte
Ring stages of developing P. malariae
Schizont development
Macrogametocyte
Microgametocyte

Plasmodium falciparum

Plasmodium malariae

From: Microscopic Diagnosis of Tropical Medicine.
1955. Bayer, Leverkusen, Germany

Figure 4B. Intracellular developmental stages of *Plasmodium.* From Microscopic Diagnosis of Tropical Medicine. 1955. Bayer, Leverkusen, Germany.

in the vertebrate intermediate host. The remaining sexual phase of the life cycle, involving fertilization and sporogony, occurs in the female mosquito.

3. Sexual phase. Further development of the gametocytes occurs in the gut of the female mosquito where the parasites are liberated from digested erythrocytes taken in with a blood meal. Microgametes fertilize macroga-

metes and the resulting zygotes develop into mobile elongated organisms, 15-20 µm in length, called ookinetes. Within a few days after the blood meal, the ookinetes penetrate the gut wall of the mosquito and develop as oocysts between the epithelium and the basement membrane. The oocysts enlarge rapidly, and within 2-3 weeks produce large numbers of spindle-shaped sporozoites. Mature oocysts rupture, releasing the sporozoites into the body cavity or hemocoel of the insect. Some sporozoites migrate to the salivary glands, which they invade. A new infection is established when the insect next feeds and sporozoites are introduced with the saliva into a cutaneous blood vessel.

Symptomology

A primary clinical attack of malaria has its onset about 2-3 weeks after infection. Seldom are typical paroxysms evident during the initial stages of attack, instead patients exhibit sustained or irregularly remittent fever. However, within a week of the primary attack typical paroxysms are experienced. The characteristic chills and fever of a malarial attack result from the release into the blood of necrotic debris from ruptured erythrocytes, together with merozoites and their metabolic by-products. With each successive schizogony, numerous additional erythrocytes are destroyed. Primary *P. vivax, P. malariae* and *P. ovale* infections generally develop suddenly with a shaking chill or rigor. Initially, the chill lasts for up to 20 minutes, and may gradually increase with successive paroxysms. During this period, the patient's temperature is actually elevated. High fever follows and may be accompanied by headache, muscular and abdominal pain, anorexia, nausea, vomiting, and increased pulse and respiratory rates. After continuing for several hours, the hot stage terminates, and is followed by a profuse sweating stage, which also may last for several hours. At the end of this stage, the patient is usually weak, but feels marked relief until the onset of the next paroxysm. The primary attack, consisting of a series of several paroxysms, may extend over a period of a month or more before the parasites completely disappear from the blood and the symptoms are terminated. As the primary attack wanes, paroxysms frequently become less severe and irregular in periodicity before they disappear. Relapses of vivax malaria may continue for a period of five years before the infection is completely eliminated. Relapses of quartan malaria may develop years after the onset of the initial malarial paroxysm. Serious complications are rarely encountered in vivax and quartan malaria. Falciparum malaria, however, frequently produces severe complications. Manifestations of cerebral involvement include hyperpyrexia, convulsions, coma, and death resulting from shock and anoxia. Gastrointestinal problems involve vomiting, diarrhea, and hemorrhage. Lysis of erythrocytes plus enhanced phagocytic activity results in anemia and enlargement of the spleen and liver. The liver may be congested and contain deposits of pigment (hemozoin) derived from hemoglobin of infected erythrocytes. In severe infections, the number of erythrocytes may be reduced by 20%. Infected erythrocytes may form multiple thrombi in various capillaries. The capillaries of the lungs, brain, and kidneys frequently become thrombotic with accumulations of infected erythrocytes, pigment and macrophages, and may rupture producing multiple extravasations.

Piroplasmea

The class Piroplasmea is comprised of a single order, Piroplasmida, whose members are of considerable veterinary importance. Apparently, the organisms reproduce only asexually, by binary fission or schizogony, in the blood cells of vertebrates. Infection is transmitted by various species of ticks, in which a sexual multiplication cycle occurs.

Genus Babesia

Human babesiosis is a fulminating hemolytic disease similar to malaria. Of the approximately 25 cases reported in humans, two have been attributed to *Babesia divergens*, a parasite of cattle, and the others to *B. microti*, a parasite of wild rodents, which can also infect dogs and cats. Most of the *B. microti* infections in humans have been reported from Nantucket Island, Martha's Vineyard, Shelter Island (near Long Island, New York), and Montauk, Long Island, New York. In the mammalian host, the parasites occur only in the trophozoite stage in the erythrocytes where they multiply by binary fission. Upon destruction of the blood cells, the parasites escape and invade other host erythrocytes continuing the cycle. Ticks become infected when they feed on the blood of an infected vertebrate and ingest the intraerythrocytic parasites, but they do not themselves transmit the disease. Instead, the parasites penetrate the gut of the tick and enter the eggs developing in the ovaries, thus infecting the young ticks, which eventually hatch from these eggs. Within the young ticks, the parasites migrate to the salivary glands, and are injected into the vertebrate host by the feeding tick. The transmission of parasites from an infected female to her offspring through the eggs is referred to as transovarian transmission (Fig. 5).

Clinical manifestations of babesiosis are noted within 10 days after infection and may last for several weeks. The illness, which mimics malaria, is characterized by fever, chills, drenching sweats, myalgias, fatigue, weakness, hemolytic anemia, and enlargement of the spleen and liver. Death results from the accumulation of toxic metabolites and anoxia from capillary occlusions.

Other Apicomplexa

Genus Pneumocystis

Pneumocystis carinii causes is an extracellular parasite found in the interstitial tissue of the lung where it causes interstitial plasma cell pneumonia. Only an intrapulmonary cycle is known for this parasite. The parasite has a cosmopolitan distribution and is found in infants, in patients with certain immunologic disorders such as leukemia, Hodgkin's disease, and hypo- or agammaglobulinemia, and in individuals undergoing immunosuppressive therapy. The organism occurs as a uninucleate pleomorphic trophozoite with a doubly contoured outer membrane, or as a cyst-like form containing as many as eight intracystic sporozoites. The intracystic sporozoites rupture from the cyst and develop into haploid trophozoites. The haploid trophozoites combine to form diploid trophozoites. Asexual development follows giving rise to precyst and mature cyst forms.

In young children, the mortality rate attributed to *P. carinii* infections may range from 30 to 40 percent. The organism is believed to be present in many humans, but may be incapable of producing a disease in healthy, immune competent hosts. The incubation period ranges from two to eight weeks. The disease is considered highly contagious. It is believed that the mode of transmission among humans is by

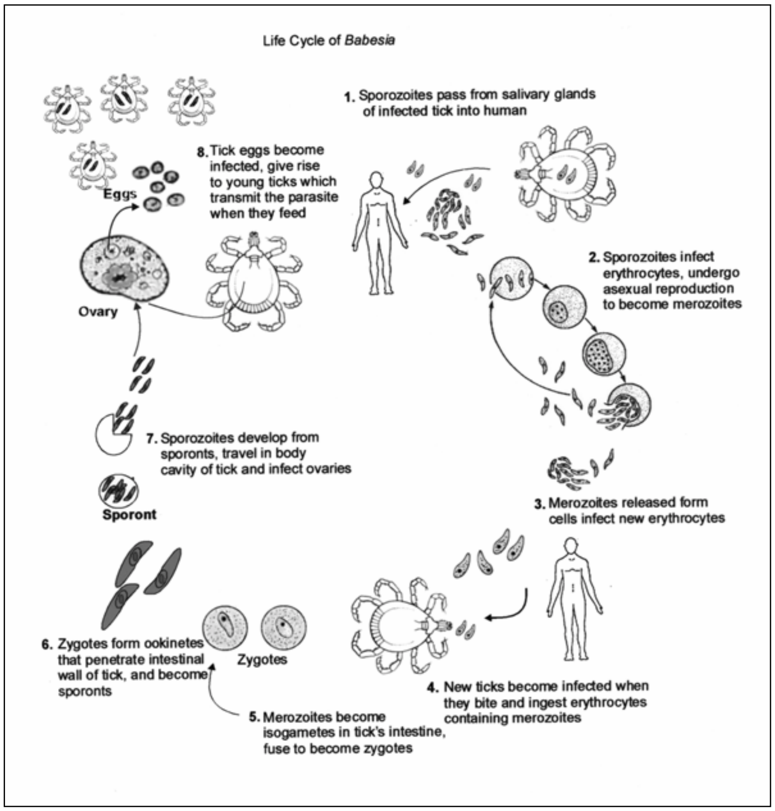
Life Cycle of *Babesia*
1. Sporozoites pass from salivary glands of infected tick into human
2. Sporozoites infect erythrocytes, undergo asexual reproduction to become merozoites
3. Merozoites released form cells infect new erythrocytes
4. New ticks become infected when they bite and ingest erythrocytes containing merozoites
5. Merozoites become isogametes in tick's intestine, fuse to become zygotes
Zygotes
6. Zygotes form ookinetes that penetrate intestinal wall of tick, and become sporonts
Sporont
7. Sporozoites develop from sporonts, travel in body cavity of tick and infect ovaries
Ovary
Eggs
8. Tick eggs become infected, give rise to young ticks which transmit the parasite when they feed

Figure 5. Life cycle of *Babesia.*

inhalation of cysts, and by transplacental transmission. The clinical course of the disease is marked by progressive dyspnea, cyanosis, tachypnea, and infiltration of the lungs, with death resulting from asphyxia. The recent increase in the incidence of pneumonocystis pneumonia is coincident with the occurrence of AIDS. Pneumocystis is the most common cause of death among AIDS patients, affecting about 60% of the patients with this immune deficiency disease.

Cryptosporidium parvum

The coccidian *Cryptosporidium parvum* causes cryptosporidiosis, a common gastrointestinal disorder. The disease, which occurs in a wide variety of hosts including primates, cattle, rodents and birds, is of clinical importance in immune compromised individuals. Symptoms include diarrhea, nausea, abdominal pain, vomiting, biliary disorder, and pneumonitis. In immune compromised individuals, infections may last several months and terminate in death. Human infection results from the ingestion of oocysts present in contaminated food or water. Sporozoites excyst in the intestine and invade the epithelial cells of the gut and respiratory system. Sporozoites develop into feeding trophozoites seen within parasitophorous vacuoles. The merozoites that are produced by asexual reproduction escape and invade nearby cells to repeat the infection cycle. Ultimately, microgametocytes and macrogametocytes are produced. These differentiate into gametes that fuse to form oocytes that are voided with the host feces.

CHAPTER 5

Digenetic Trematodes: Flukes

Digenetic trematodes comprise a large group of endoparasites variable in size, shape, and habitat. The body of most adult flukes is flattened dorsoventrally, and covered with a smooth or spiny, resistant cuticle. Adult flukes inhabit the digestive tract, bile passages, lungs, or blood of their vertebrate hosts.

Adult digenetic flukes have two suckers or attachment organs, an anterior oral sucker surrounding the mouth, and a more posterior ventral sucker or acetabulum. The mouth or buccal cavity leads to a muscular pharynx and esophagus, and then into branched cecae (Fig. 1). The cecae run parallel to each other and end blindly, either in the posterior portion of the worm, or about halfway down the body. The parasites feed on mucus, blood, and other host tissues. Except for schistosomes or blood flukes, the trematodes infecting humans are hermaphroditic, and also capable of self-fertilization. Typically, the male reproductive system consists of two testes, seminal ducts (vas efferens) that connect to form a vas deferens, seminal vesicle, ejaculatory duct, and a muscular cirrus or copulatory organ that terminates at the male genital pore located within a common genital atrium. The genital atrium, which is commonly found on the midventral surface and anterior to the acetabulum, opens to the exterior via a gonopore. The cirrus may be enclosed within a cirrus pouch, and evaginated for sperm transfer to the female system. The female reproductive system consists of a single ovary, and oviduct, seminal receptacle, an ootype or egg chamber, and a uterus that extends to the female genital pore. The female genital pore is usually located near the male genital pore within the genital atrium. Self- and cross-fertilization are possible modes of reproduction. Following copulation, spermatozoa are stored in the seminal receptacle and released as required.

Life Cycle

The typical life cycle of a digenetic trematode is comprised of an asexual phase in molluscs, usually gastropods, and a sexual phase in vertebrates, the definitive hosts. Many species have a second intermediate host in which the sexual phase begins. The asexual phase includes an egg that is either unembryonated when discharged from the definitive host, or it is embryonated and contains a fully formed ciliated larval stage called a miracidium (Fig. 1). Most eggs have an operculum or cap at one end through which the miracidium emerges. In the majority of species, further development is possible when the miracidium hatches and penetrates the soft tissues of a suitable species of snail. In other cases, the eggs do not hatch outside the definitive host, but rather are ingested by the intermediate host. It is possible that miracidia respond to various conditions, including light, temperature, salinity, pH, and gravity, that bring the parasites into the range of a suitable molluscan host. There is some evidence that some miracidia are attracted to their molluscan host by chemotactic stimuli. Penetration of the host is affected by a boring action of the larva, a process

---

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

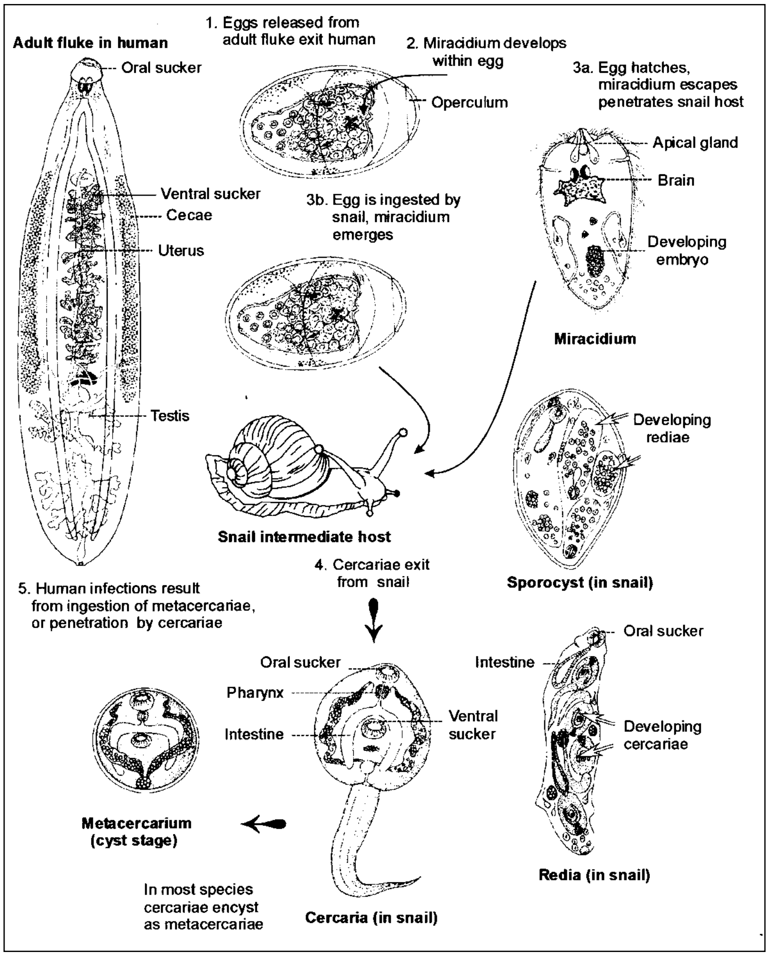
Adult fluke in human
Oral sucker
Ventral sucker
Cecae
Uterus
Testis
1. Eggs released from adult fluke exit human
2. Miracidium develops within egg
Operculum
3a. Egg hatches, miracidium escapes penetrates snail host
Apical gland
Brain
Developing embryo
Miracidium
3b. Egg is ingested by snail, miracidium emerges
Developing rediae
Snail intermediate host
Sporocyst (in snail)
4. Cercariae exit from snail
5. Human infections result from ingestion of metacercariae, or penetration by cercariae
Oral sucker
Pharynx
Intestine
Ventral sucker
Intestine
Oral sucker
Developing cercariae
Metacercarium (cyst stage)
In most species cercariae encyst as metacercariae
Cercaria (in snail)
Redia (in snail)

Figure 1. Developmental stages of digenetic trematodes. Modified from Hegner RW, Engemann. Invertebrate Zoology. New York: The Macmillan Company, 1968.

that is facilitated by the release of proteolytic enzymes from glands located in the apical region of the body of the miracidium.

Within the miracidium, a mass of germinal tissue provides the cell lineage for succeeding stages in the life cycle. Upon entering the molluscan intermediate host, the miracidium sheds its ciliated covering, elongates and transforms into either a saclike first-generation (mother) sporocyst, or in some species a slightly more advanced stage called a redia. Germ cells within the body of the mother sporocyst or redia develop into either second generation (daughter) sporocysts or rediae. Germinal cells in the daughter sporocysts or rediae eventually give rise to cercariae, the final

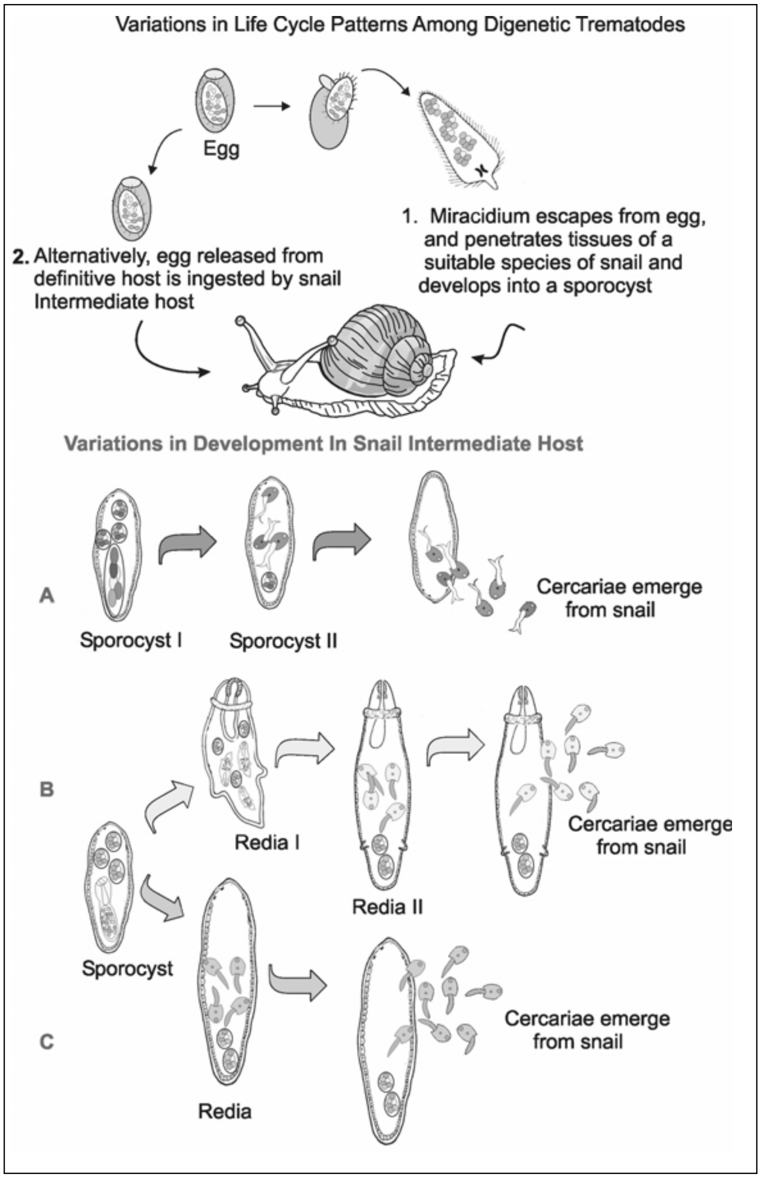
Variations in Life Cycle Patterns Among Digenetic Trematodes
Egg
1. Miracidium escapes from egg, and penetrates tissues of a suitable species of snail and develops into a sporocyst
2. Alternatively, egg released from definitive host is ingested by snail Intermediate host
Variations in Development In Snail Intermediate Host
A
Sporocyst I
Sporocyst II
Cercariae emerge from snail
B
Redia I
Redia II
Cercariae emerge from snail
Sporocyst
C
Redia
Cercariae emerge from snail

Figure 2. Variations in the life cycle of digenetic trematodes.

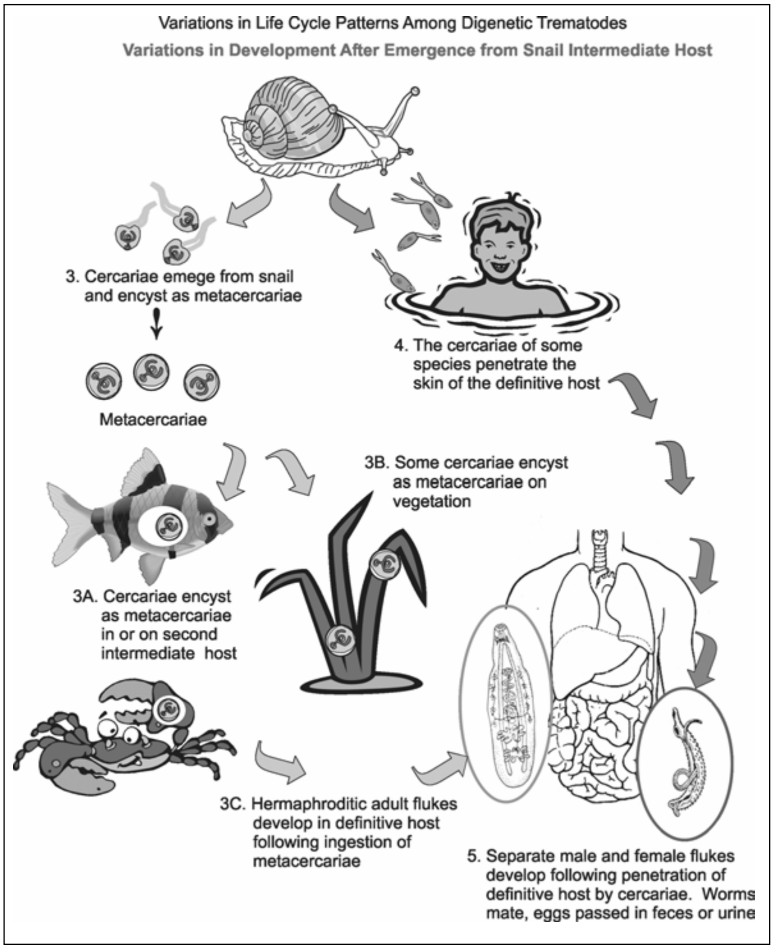
Variations in Life Cycle Patterns Among Digenetic Trematodes
Variations in Development After Emergence from Snail Intermediate Host
3. Cercariae emege from snail and encyst as metacercariae
Metacercariae
4. The cercariae of some species penetrate the skin of the definitive host
3B. Some cercariae encyst as metacercariae on vegetation
3A. Cercariae encyst as metacercariae in or on second intermediate host
3C. Hermaphroditic adult flukes develop in definitive host following ingestion of metacercariae
5. Separate male and female flukes develop following penetration of definitive host by cercariae. Worms mate, eggs passed in feces or urine

Figure 3. Variations in the life cycle of digenetic trematodes.

product of the asexual phase in the molluscan host. The cercariae are essentially young flukes. They possess a mouth, gut, suckers, penetration glands (secretions from which facilitate the invasion of the molluscan intermediate host), and/or cystogenous glands (whose secretions may provide a temporary protective layer). During their development, propagatory cells, derived from the germ cells, produce the stem cells of the adult reproductive system. In most species, the cercariae escape from the first intermediate host and encyst on objects in the water, or in the body of a second intermediate host, and then transform into metacercariae which develop and become infective to the definitive host. The definitive host becomes infected when it ingests the plant or animal harboring the metacercariae. Species specific

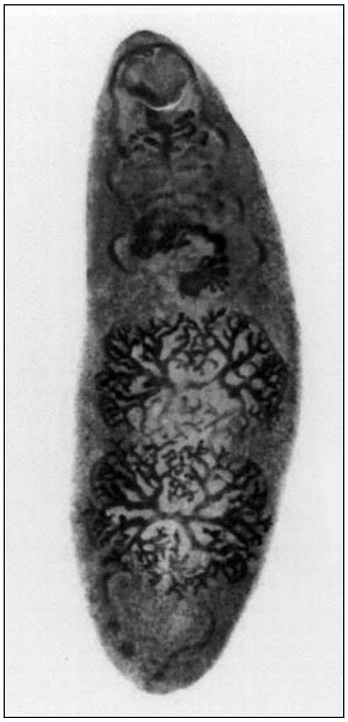

Figure 4. *Fasciolopsis buski* adult, an intestinal fluke.

variations in the life cycle pattern include: (1) more than one generation of sporocysts or rediae; (2) deletion of either sporocyst or redial stages; and, (3) penetration by cercariae of the definitive host (Figs. 2 and 3).

Intestinal Flukes

Fasciolopsis buski is an intestinal parasite of pigs and humans. This fluke is widespread in China, Vietnam, Thailand, and Indonesia. The parasite is a large worm, measuring up to 8 cm in length (Fig. 4). Each worm may produce as many as 25,000 eggs per day.

Infection with *F. buski* is acquired by ingestion of metacercariae encysted on freshwater edible plants such as bamboo shoots, water chestnut, water hyacinth, and water caltrop (Fig. 5). After excystation in the duodenum, the larvae attach to the intestinal mucosa and in about three months develop into adult parasites. Embryonation occurs in fresh-water after the eggs pass with the feces from the host. Hatched miracidia penetrate the soft tissues of certain planorbid snails (*Segmentina hemispaerula, S. trochoides, Hippeuitis contori, Gyraulus* sp.) and develop into sporocysts. Two redial generations are followed by the formation of numerous cercariae, which pass from the snail and after swimming about, move onto aquatic vegetation and encyst. At the site of attachment, these large flukes produce a local inflammation and ulceration, occasionally accompanied by hemorrhage. In severe infections,

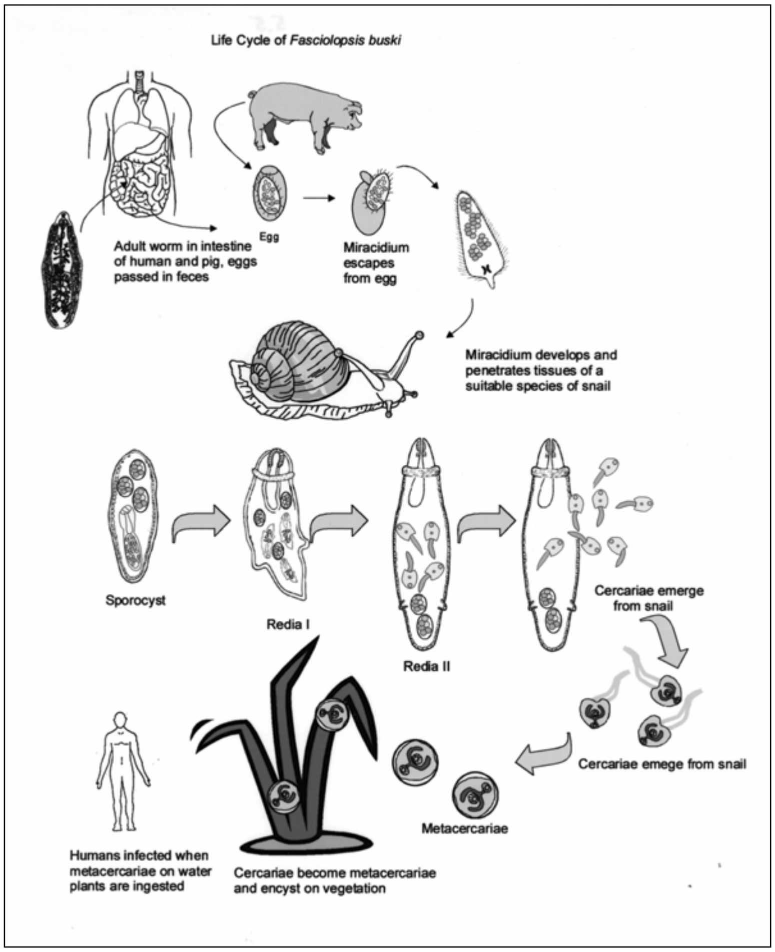
Life Cycle of *Fasciolopsis buski*
Adult worm in intestine of human and pig, eggs passed in feces
Egg
Miracidium escapes from egg
Miracidium develops and penetrates tissues of a suitable species of snail
Sporocyst
Redia I
Redia II
Cercariae emerge from snail
Cercariae emege from snail
Metacercariae
Humans infected when metacercariae on water plants are ingested
Cercariae become metacercariae and encyst on vegetation

Figure 5. Life cycle of *Fasciolopsis buski.*

there may be acute intestinal obstruction, abdominal pain, diarrhea, anorexia, nausea, vomiting, edema of the face, abdomen, and lower extremities, and yellow profuse stools indicative of malabsorption.

Heterophyid flukes are small, ovoid-shaped flatworms that live attached to the intestinal mucosa of fish-eating birds and mammals. The flukes are generally less than 3 mm in size, and produce operculate eggs. Hatching does not occur unless the eggs are ingested by appropriate species of freshwater snails. Within the snail, a single sporocyst generation is followed by two generations of rediae, and then cercariae. After escaping from the snail, the cercariae encyst as metacercariae either on the underside of the scales of certain fresh-water fishes or in the muscles of the fish.

Human infections are acquired by eating uncooked fish containing encysted metacercariae.

Two heterophyids infecting humans are *Metagonimus yokogawai* and *Heterophyes heterophyes*. The parasites occur in China, Japan, Korea, the Philippines, and Israel. *Heterophyes* also occurs in parts of India and Egypt. *Metagonimus* has also been reported from Indonesia, Spain, Russia, northern Siberia and the Balkans. In severe infections, there may be superficial ulceration of the intestinal mucosa and persistent mucous diarrhea. Occasionally, some heterophyids penetrate deep into the intestinal mucosa. Their eggs enter the lymphatic and circulatory passages and are transported to various organs, including the heart, brain and spinal cord, where granulomatous responses are provoked. Heart damage (heterophyid myocarditis) may result from the accumulation of numerous eggs in the organ.

Echinostomate flukes are characterized by a collar of spines on a disk surrounding the oral sucker. The parasites have been reported from the Philippines and China, where dogs are commonly infected. Operculated eggs are voided in the feces of the definitive host. Planorbid and lymnaeid snails are first intermediate hosts, into which the miracidia penetrate and develop into sporocysts. Cercariae, which are produced from a second generation of rediae, escape from the snail in water and encyst as metacercariae either on aquatic vegetation, or in various aquatic organisms. Human infections frequently result from the raw consumption of the second intermediate host, usually edible fresh-water snails (*Pila conica, Viviparus javanicus*), in which the metacercariae are found. *Echinostoma ilocanum* is a common parasite of humans. The worm measures about 7 mm in length, by 1.5 mm in width. The parasite produces inflammatory reactions and ulcerations at the sites of attachment to the wall of the intestine. The usual symptoms are abdominal discomfort, diarrhea, and anemia. Echinostomiasis can be prevented by adequate cooking of food.

Hepatic Flukes

The flukes inhabiting the biliary passages of humans are commonly referred to as liver flukes. The worms produce fibrosis and hyperplasia of the biliary epithelium leading to portal cirrhosis. The liver flukes of significant medical importance belong to the following three genera: *Fasciola, Clonorchis* and *Opisthorchis*.

Fasciola hepatica is commonly known as the sheep liver fluke. The parasite was the first trematode to be described, and the first for which a complete life cycle was elucidated. The parasite has a widespread distribution throughout many sheep- and cattle-raising countries of the world. Numerous human infections have been reported from Latin America, France, Algeria, England, Germany, Poland, China, Russia, Hawaii, parts of Africa, and in some areas of the southern continental United States.

The adult fluke is flattened and leaf-shaped, with an anterior conical projection and a broadly rounded posterior. The adults, which measure up to 3 cm in length and 1.3 cm in width, produce operculate eggs. Human fascioliasis is acquired by ingesting vegetation, usually freshwater cress, on which encysted metacercariae are found. The metacercariae excyst in the gut, and the parasite penetrates the intestinal wall, migrates through the peritoneal cavity and enters the bile ducts and liver parenchyma. Occasionally, some adult flukes wander in the peritoneal cavity and other ectopic areas producing necrotic foci with fibrosis. While grazing, infected sheep and cattle contaminate vegetation and water sources with feces containing

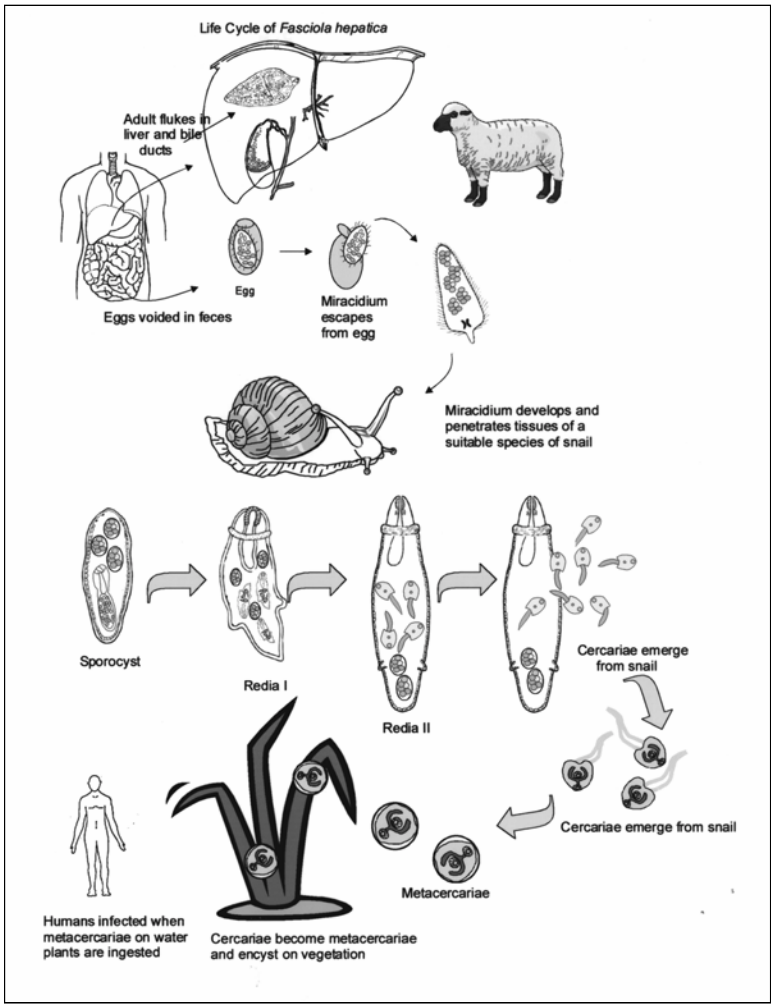
Life Cycle of *Fasciola hepatica*
Adult flukes in liver and bile ducts
Egg
Eggs voided in feces
Miracidium escapes from egg
Miracidium develops and penetrates tissues of a suitable species of snail
Sporocyst
Redia I
Redia II
Cercariae emerge from snail
Cercariae emerge from snail
Metacercariae
Humans infected when metacercariae on water plants are ingested
Cercariae become metacercariae and encyst on vegetation

Figure 6. Life cycle of *Fasciola hepatica.*

fluke eggs. Upon hatching, the miracidia have about 24 hours in which to find a suitable lymnaeid snail host. A miracidium penetrates the snail, sheds its ciliated epithelium, and transforms into a sporocyst. Two generations of rediae are followed by cercariae, which begin emerging from the snail 5-7 weeks post infection. The cercariae attach to any available object and transform into metacercariae (Fig. 6).

Severe infections may be characterized in part by mechanical obstruction, irritation, and inflammation of host tissues. Frequently, there are hyperplastic changes in

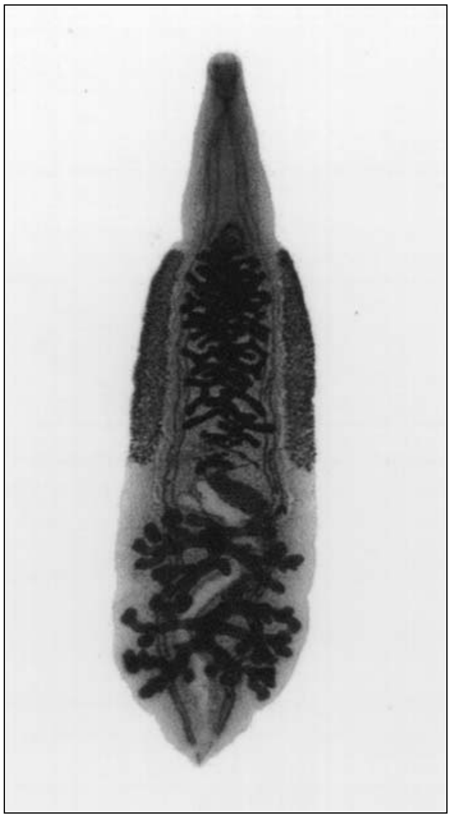

Figure 7. Adult *Clonorchis sinensis,* a hepatic fluke.

the biliary epithelium, fibrosis and/or erosion of the walls of the bile ducts, atrophy of the liver parenchyma, and cirrhosis. Migrating larvae may produce lesions in the eye, brain, lungs, skin and other ectopic foci.

Clonorchis sinensis, the Chinese liver fluke, occurs in China, Taiwan, Japan, Korea, and Vietnam. Dogs and cats are important reservoir hosts. Adult worms are elongate and flat, measuring as much as 2.5 cm in length by 0.5 cm in width (Fig. 7). They occupy the bile ducts, most frequently in the more distal regions, just under the surface of the liver. The eggs are discharged into the biliary passages and pass in the feces of the host. The average daily production of eggs per worm is about 2000. The eggs hatch when ingested by a suitable species of operculate snails (*Bulimus, Alocinma* and *Parafossarulus*), and the miracidia transform into sporocysts. A single generation of rediae is followed by the production of cercariae. After emerging from the snails the cercariae penetrate the tissues of certain fresh-water fishes and become encysted. Human infection is acquired when uncooked fish containing the encysted metacercariae is eaten (Fig. 8). Upon excystation in the duodenum, the larvae enter the smaller biliary passages by migrating through the ampulla of Vater.

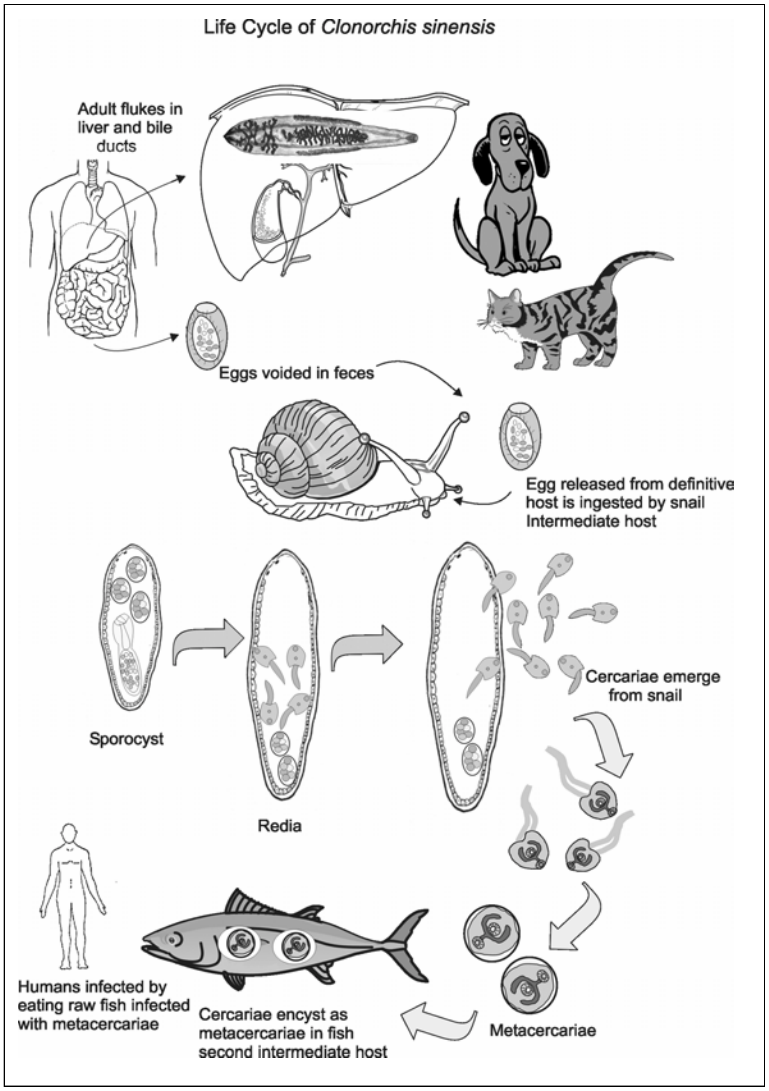
Life Cycle of *Clonorchis sinensis*
Adult flukes in liver and bile ducts
Eggs voided in feces
Egg released from definitive host is ingested by snail Intermediate host
Sporocyst
Redia
Cercariae emerge from snail
Metacercariae
Cercariae encyst as metacercariae in fish second intermediate host
Humans infected by eating raw fish infected with metacercariae

Figure 8. Life cycle of *Clonorchis sinensis.*

Chronic clonorchiasis is generally asymptomatic. In severe infections, the acute period lasts less than a month. The infection may be characterized by chills, fever, epigastric pain, diarrhea, enlargement and tenderness of the liver, congestive splenomegaly, hyperplasia of bile duct mucosa, biliary obstruction, cholangitis, and liver abscesses. The degree to which these symptoms are manifested is directly related to the intensity of infection.

Two other flukes inhabiting the bile ducts of humans are *Opisthorchis felineus* and *O. viverrini*. The symptoms of opisthorchiasis are very similar to the disease picture produced by *Clonorchis sinensis*. Except for different species of intermediate hosts, their life cycles are similar to that of *C. sinensis. Opisthorchis felineus* has been reported from Siberia, Poland, southeastern Europe, Asia, and India. Several species of fresh-water fishes serve as secondary intermediate hosts. Dogs, cats, and other fish-eating mammals are naturally parasitized. The snail *Bulimus tentaculatus* is the first intermediate host. *Opisthorchis viverrini* is endemic to northeastern Thailand, where fish-eating mammals are affected.

Pulmonary Flukes

Flukes inhabiting the lungs of various mammals belong to the genus *Paragonimus*. The best known species infecting man is *P. westermani*, the cause of endemic hemoptysis or paragonimiasis in regions of Thailand, Indonesia, China, Central and South America, Africa, and the South Pacific. The adult worms, which measure up to 1.6 cm in length and 0.8 cm in width, are characteristically found enclosed in cystic structures near the bronchi. The eggs are coughed up and expectorated in the sputum, or swallowed and later evacuated in the feces. In fresh-water, the eggs embryonate in about two weeks, then hatch. The free swimming miracidia penetrate appropriate snails (*Semisulcospira libertina, Melania amurensis, M. obliquegranosa* and *Brotia asperata*) and transform into sporocysts, in which rediae are developed (Fig. 9). Each redia subsequently produces a brood of cercariae. Emerging cercariae next invade the tissues of suitable fresh-water crustaceans, especially crabs or crayfish (*Cambarus, Pseudotelphusa, Potamon, Paratelphusa* and *Eriocheir*), and form metacercariae. Human infection results from eating raw crabs or crayfish containing encysted metacercariae. Following excystation in the duodenum, young worms penetrate the intestinal wall, migrate through the peritoneal cavity, pass through the diaphragm and enter the peribranchial tissues. The circuitous migratory route through the body cavity may take about three weeks before the worms enter the lungs, in which they develop into adults in 5-6 weeks.

Paragonimus westermani has been found in many abnormal or ectopic sites, such as various muscles, heart, liver, brain, mesenteric lymph nodes, testes, and pleural or peritoneal cavity. In these locations, fibrous inflammatory cysts are formed around the worms. Some of these lesions may be ulcerative. Cerebral paragonimiasis frequently results in intracranial calcification, impaired vision, and epilepsy. Lung infections may be characterized by occasional coughing with the discharge of blood in the sputum (periodic hemoptysis), dyspnea, fever, fatigue, and anorexia.

Blood Flukes

Trematodes belonging to the genus *Schistosoma* inhabit the circulatory system of humans causing one of the more serious diseases of helminth origin, schistosomiasis (bilharziasis). The schistosomes differ from most other trematodes in that the body is cylindrical, and the sexes are separate. The worms exhibit sexual dimorphism, with females slightly longer and more slender than males. Males possess a ventral sex canal (gynecophoric canal) in which the female reposes (Fig. 10). Adult schistosomes

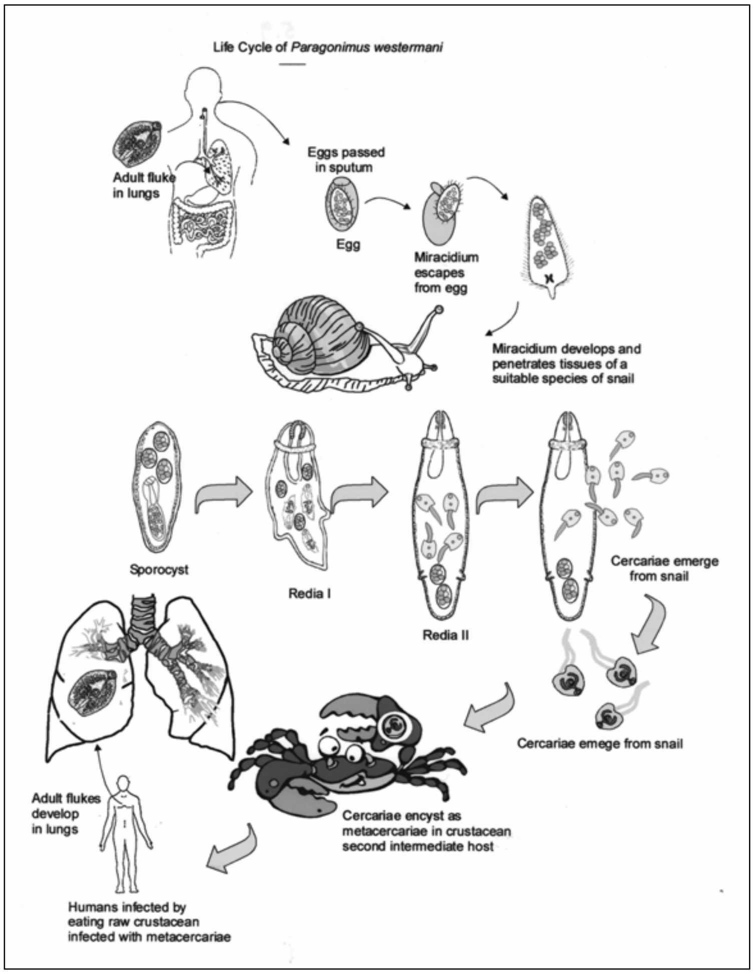
Life Cycle of Paragonimus westermani
Adult fluke in lungs
Eggs passed in sputum
Egg
Miracidium escapes from egg
Miracidium develops and penetrates tissues of a suitable species of snail
Sporocyst
Redia I
Redia II
Cercariae emerge from snail
Cercariae emege from snail
Cercariae encyst as metacercariae in crustacean second intermediate host
Adult flukes develop in lungs
Humans infected by eating raw crustacean infected with metacercariae

Figure 9. Life cycle of *Paragonimus westermani.*

characteristically live in pairs in the portal venous blood vessels or in the vesicle venules of the caval system, where females lay non-operculate, partially embryonated eggs. The worms live on the average of a few years, but in exceptional cases they can persist for as long as 30 years or more. The females produce prodigious numbers of eggs, which are forced through the walls of either the intestine or urinary bladder and are discharged in the excreta. The eggs hatch in fresh-water and the miracidia penetrate suitable species of snails, in which they develop into sporocysts (Fig. 11). The cercariae, which possess a forked tail, are produced from second generation sporocysts. The cercariae escape from the snail and, on contact with

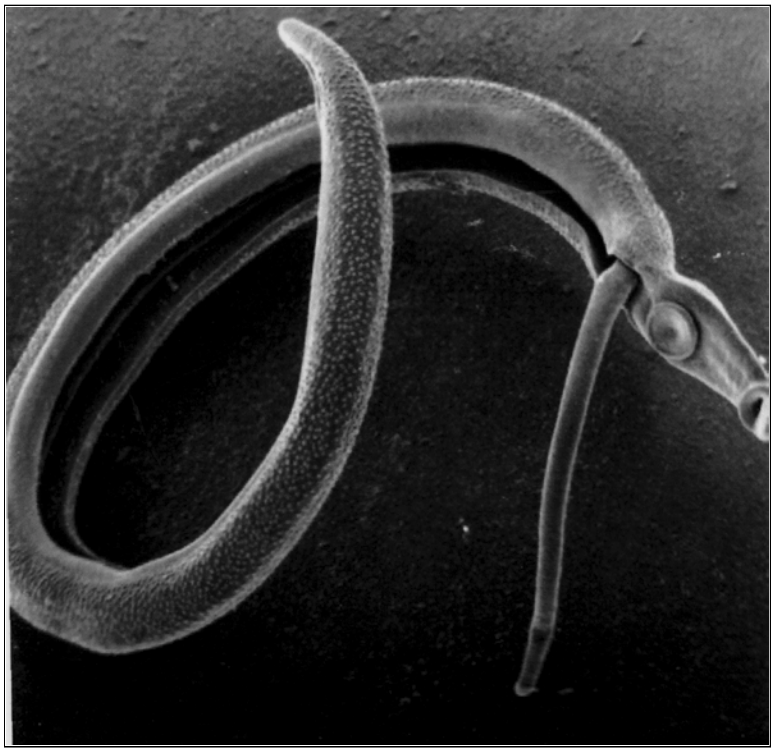

Figure 10. Male and female *Schistosoma mansoni* in copula. Scanning electron micrograph courtesy of Harvey Blankespoor.

human skin, penetrate the body (Fig. 12), enter the cutaneous blood vessels and initiate infection. Young flukes are first transported to the liver sinusoids where they feed and grow for a period of 5 to 6 weeks. When mature, they migrate into the portal system to their final location in mesenteric or vesicular veins.

Three species of schistosomes infect humans:

1. *Schistosoma mansoni* occurs over extensive areas of Africa, Egypt, the Arabian peninsula, and parts of South America and the West Indies. Various primates, insectivores, rodents, and marsupials are also infected with this fluke. Male parasites measure up to 1 cm in length, females about 1.6 cm. The eggs are elongate and oval and possess a conspicuous lateral spine projecting near one pole (Fig. 13). Adult worms are usually found in smaller branches of the inferior mesenteric vein, intrahepatic portal blood, vesicle venules, and pulmonary arterioles. The pulmonate snail *Biomphalaria glabrata* is the intermediate host of *S. mansoni*.
2. *Schistosoma haematobium* occurs in Egypt, Africa, Malagasy, Arabian peninsula, Iran, India, Israel, Portugal, Cyprus, Syria, and Lebanon. Male

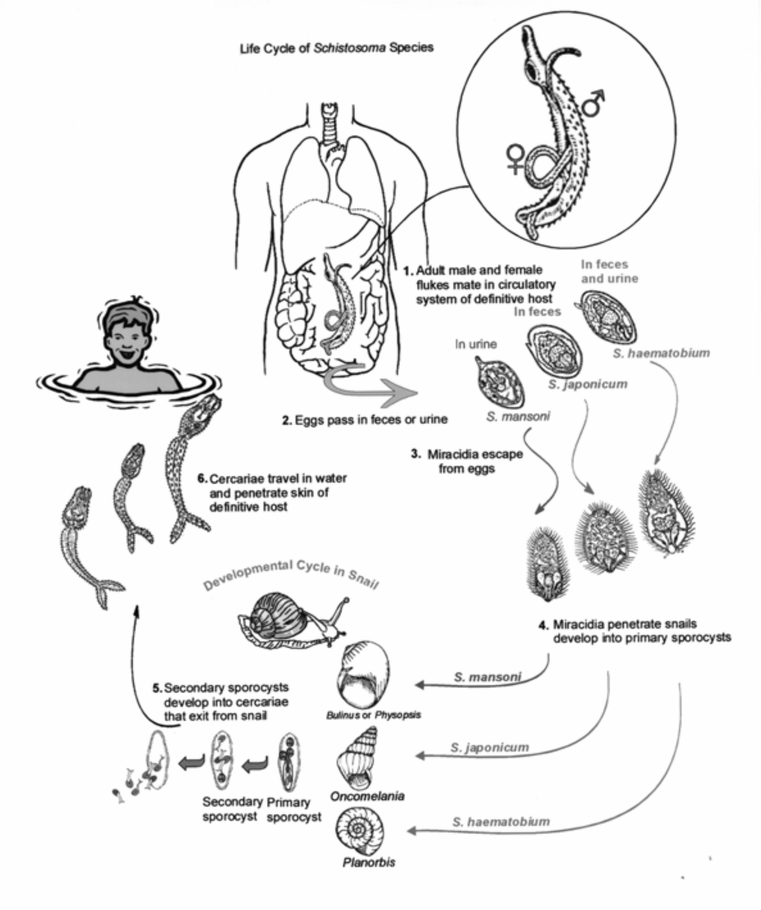
Life Cycle of *Schistosoma* Species
♂
♀
1. Adult male and female flukes mate in circulatory system of definitive host
In feces and urine
In feces
In urine
S. haematobium
S. japonicum
S. mansoni
2. Eggs pass in feces or urine
3. Miracidia escape from eggs
6. Cercariae travel in water and penetrate skin of definitive host
Developmental Cycle in Snail
4. Miracidia penetrate snails develop into primary sporocysts
S. mansoni
5. Secondary sporocysts develop into cercariae that exit from snail
Bulinus or Physopsis
S. japonicum
Oncomelania
Secondary sporocyst
Primary sporocyst
S. haematobium
Planorbis

Figure 11. Life cycle of *Schistosoma.*

worms may attain a length of 1.5 cm, and females a length of 2 cm. The eggs have a conspicuous terminal spine (Fig. 13). The adult worms occupy the blood vessels surrounding the bladder. The parasite is carried by pulmonate snails of the genus *Bulinus*.

3. *Schistosoma japonicum* occurs in Japan, Korea, Formosa, China, Taiwan, Laos, Cambodia, Thailand, the Philippines, and Celebes. Male worms may attain a length of 2.2 cm and females 3.0 cm. The spherical eggs are smaller than those of the other two schistosome species and possess a minute, blunt projection may be seen on some eggs near one pole (Fig. 13). Adult worms live within the branches of the superior mesenteric

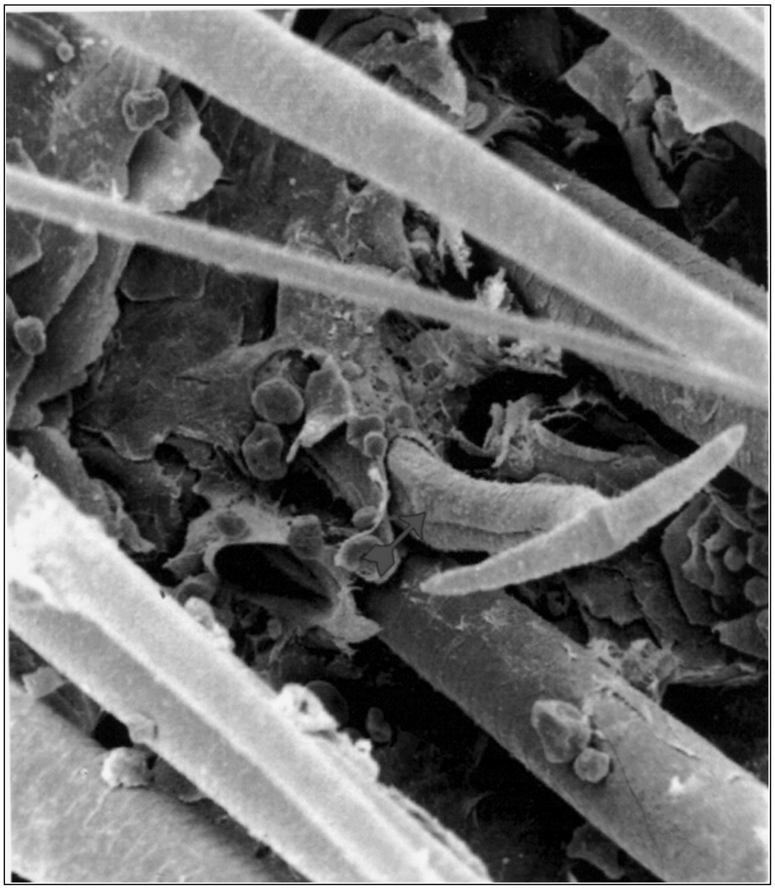

Figure 12. Caudal end of *Schistosoma mansoni* cercaria penetrating epidermis of definitive host shortly after emergence from snail intermediate host. Scanning electron micrograph courtesy of Harvey Blankespoor.

vein, near the small intestine. As the disease progresses, the inferior mesenterics and the caval system may also be invaded. Gill-breathing snails of the genus *Oncomelania* are the principal intermediate hosts for *S. japonicum*.

Early symptoms of schistosomiasis include localized dermal hemorrhages, edema and pruritus at the site of penetration. These symptoms typically disappear in less than one week. During the succeeding month, fever, toxic and allergic manifestations, accompanied by abdominal distress may develop. Migration of the parasites through the lungs may cause spasmodic cough, hemoptysis, and severe pleuritic chest pain. Progressive pathologic conditions include bronchitis, fibrosis, and pleu-

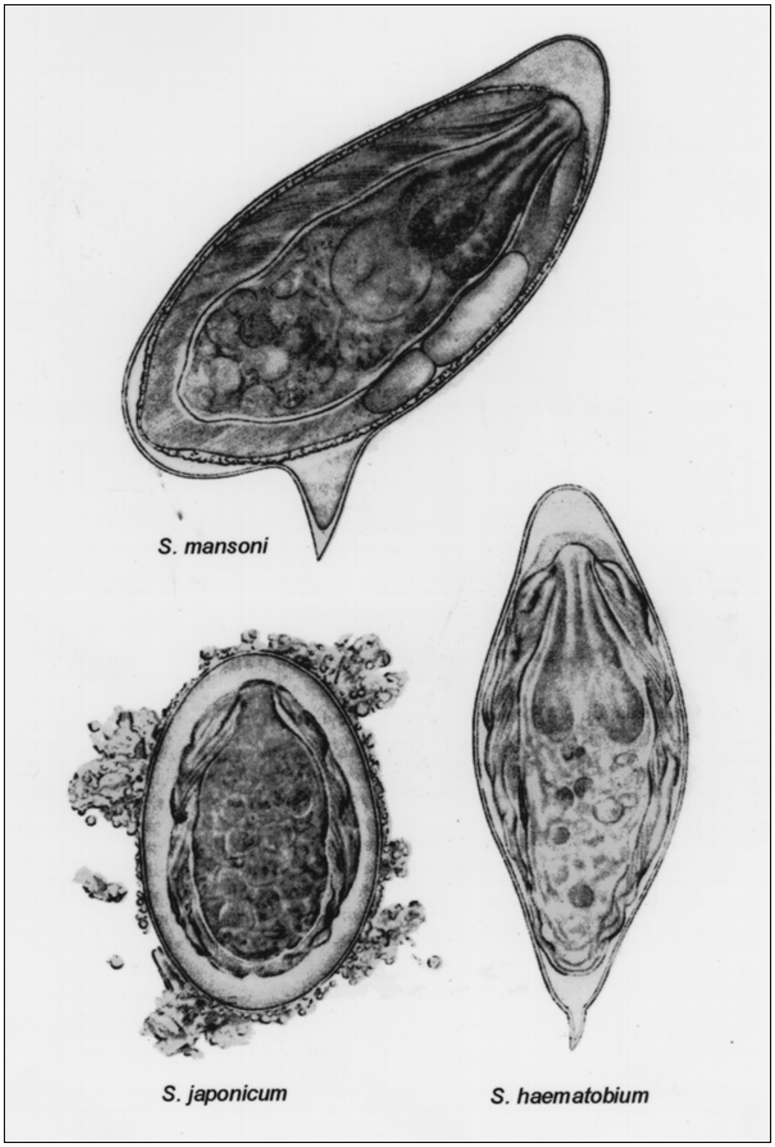
S. mansoni
S. japonicum
S. haematobium

Figure 13. Eggs of *Schistosoma* species. From Microscopic Diagnosis of Tropical Medicine. 1998. Bayer, Leverkusen, Germany.

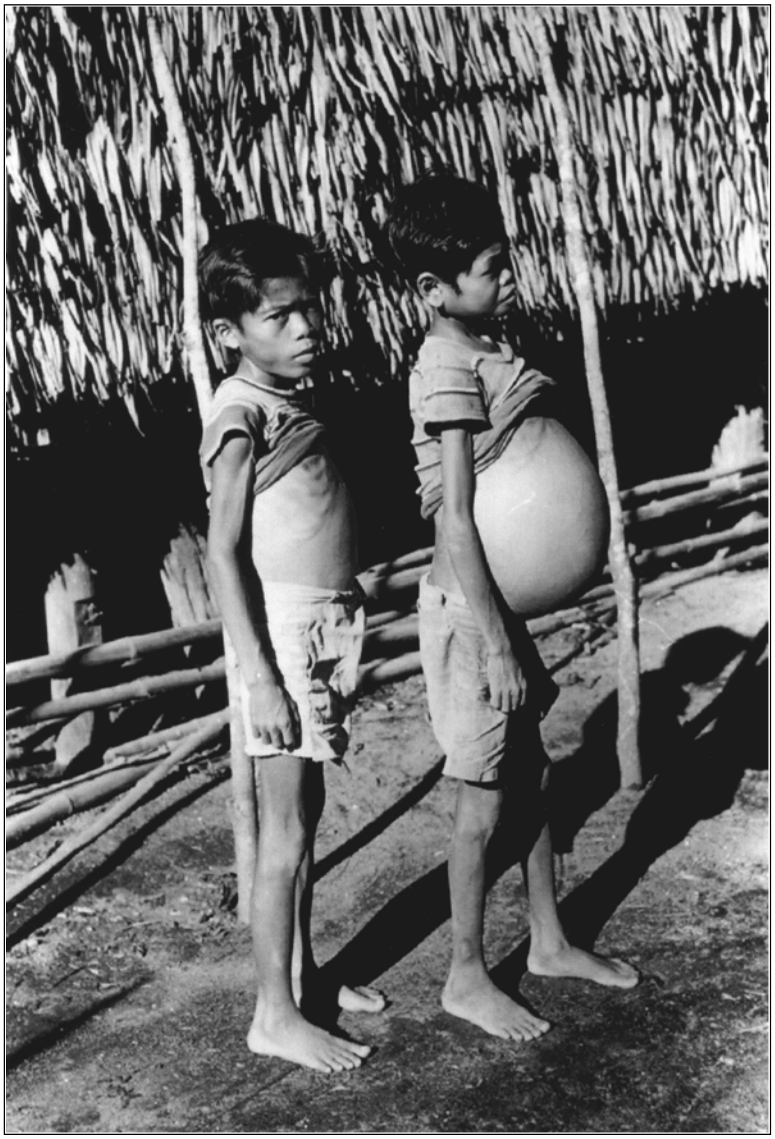

Figure 14. Ascites in advanced schistosomiasis japonica (Leyte, Philippines). This is an example of dwarfing (individual at right) caused by the parasite. The male of the left is 13 years old, the individual on the right is 24 years old. Photograph courtesy of Robert E. Kuntz.

ral effusion. Acute hepatitis is generally manifested soon after the parasites invade the liver. Other symptoms appearing during advanced stages of infection include fever, uticaria, epigastric discomfort, diarrhea, dysentery, the accumulation of fluid in the abdominal cavity (ascites) (Fig. 14), hepatosplenomegaly and liver dysfunction. Hematuria (blood urine), cystitis (inflammation of the bladder), and dysuria are common symptoms in patients infected with *S. haematobium.* There may be extensive tissue damage due to the extrusion of eggs through the wall of the intestine or urinary bladder. Eggs lodged in various tissues provoke inflammatory reaction, with leukocytic and fibroblastic infiltration. Fibrous nodules (granulomas) enclosing small accumulations of eggs are frequently formed on the serosal and peritoneal surfaces. Necrosis, ulceration, and malignant changes may follow. Neurological pathologies, involving coma and paralysis, result from parasite eggs invading the brain.

Schistosome dermatitis or “swimmer’s itch” in humans results from the penetration of the skin with non-human (i.e., bird, cattle, rodent) schistosome cercariae in fresh and brackish water. Fortunately, the parasite does not produce a permanent infection. Infection produces an initial severe prickling, local edema, generalized urticaria, and the development of macules and pustules. Schistosome dermatitis has been reported from many regions of the continental United States, Hawaii, Europe, Latin America, India, and Thailand.

CHAPTER 6

Cestodes

Cestodes or tapeworms are parasitic during all or a major portion of their lives. With few exceptions, adult cestodes possess an elongated tape-like body, and they lack a digestive tract. The latter feature separates these worms from trematodes and nematodes. Typically, the habitat of the adult tapeworm is the intestinal tract of its host. Cestode larvae, however, invade a wide range of host tissues, although most larvae demonstrate a preference for particular, species-specific, sites.

The body of a tapeworm consists of an anterior attachment organ or scolex, followed by an unsegmented neck, succeeded by a chain of proglottids ("segments") termed the strobila (Fig. 1). The number of proglottids may vary from 3 to 4 in the hydatid worm (*Echinococcus granulosis*), to more than four thousand in the broad or fish tapeworm (*Diphyllobothrium latum*). The scolex may be equipped with various holdfast organs, which secure the worm to the mucosa of the host's small intestine. There are essentially three types of adhesive structures: (1) Bothridia are broad, leaf-like structures with flexible margins. They usually occur in groups of four and project from the dorsal or ventral side of the scolex. (2) Bothria are dorsal or ventral grooves of weak muscularity. Usually two are found on the scolex. (3) Acetabula are highly muscular, cup-shaped adhesive structures (suckers). There are frequently four acetabula on a scolex. With the exception of the broad tapeworm, which possesses bothria, all tapeworms of humans have four cup-shaped suckers on the scolex. In addition to suckers, most tapeworms have keratinaceous hooks that anchor the scolex to the intestinal wall. In acetabulate cestodes, the hooks may be arranged circularly on a protrusible cone-like structure, the rostellum. In some cases, the rostellum lacks hooks, and is termed "unarmed".

In many cestodes, the strobila grows continuously throughout the life of the worm by asexual budding (strobilization) of new proglottids in the neck region. Each proglottid moves posteriorly as a new one is formed. As new proglottids are added, the strobila elongates so that in some species enormous lengths are attained. As the proglottids move from the neck region, the reproductive organs mature and the eggs are fertilized. The most recently formed and immature proglottids are found nearest the scolex, while the larger, mature proglottids are found near the middle of the strobila. The terminal portion of the strobila contains ripe or gravid proglottids filled with eggs. Typically, each proglottid contains one or more sets of reproductive organs. Mature segments contain both male and female reproductive organs, and thus are capable of self fertilization. Cross fertilization also is possible between different segments of a single worm, or between segments of two worms living together within the host. Usually the male organs mature first and produce sperm that are stored until eggs are manufactured (protandry or androgyny). In some species, the ovary matures before the testes (protogyny or gynandry).

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

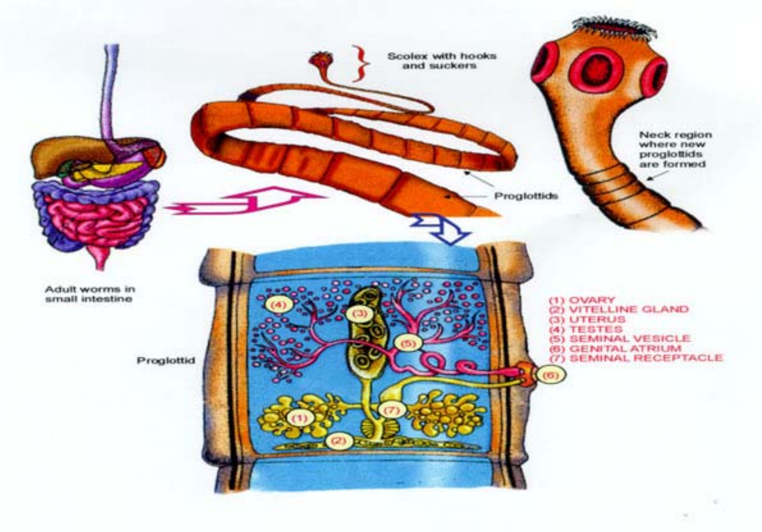
Scolex with hooks and suckers
Neck region where new proglottids are formed
Proglottids
Adult worms in small intestine
Proglottid
(1) OVARY
(2) VITELLINE GLAND
(3) UTERUS
(4) TESTES
(5) SEMINAL VESICLE
(6) GENITAL ATRIUM
(7) SEMINAL RECEPTACLE
(1)
(2)
(3)
(4)
(5)
(6)
(7)

Figure 1. Morphology of a typical cestode.

Developmental Stages and Life Cycles

There are two basic life cycle patterns among the cestodes parasitizing humans; one exhibited by pseudophyllideans, the other by cyclophyllideans. In both, the egg gives rise to an oncosphere or hexacanth larva, so-called because it possesses six elongated hooks at the posterior pole. An egg membrane or embryophore surrounding the embryo may be ciliated, in which case the parasite is termed a coracidium. Virtually every known tapeworm has an indirect life cycle involving at least one intermediate host that ingests the eggs that pass from the intestinal tract of the definitive host. One notable exception is *Vampirolepis (Hymenolepis) nana*, a cyclophyllidean with a direct life cycle in which both larval and adult stages occur in the definitive host (either mice or humans).

Pseudophyllidean Life Cycle

The eggs of pseudophyllids are generally unembryonated when shed from the definitive host, but in water they soon embryonate and hatch. The oncosphere emerging from the egg has a ciliated embryophore and is called a coracidium (Fig. 2). Within the first intermediate host the coracidium sheds its ciliated epithelium and actively penetrates the gut. Within the body cavity it develops into an elongated, larva called a procercoid. The embryonic hooks are retained and, in some species, are borne on a posterior structure termed a cercomer. However, the definitive holdfast has not yet developed. When the first intermediate host is consumed by the second intermediate host, usually a freshwater fish, the larva penetrates the gut and migrates into the connective tissue and skeletal muscles, and transforms into the last larval stage, the plerocercoid. The embryonic hooks are lost, and rudiments of the

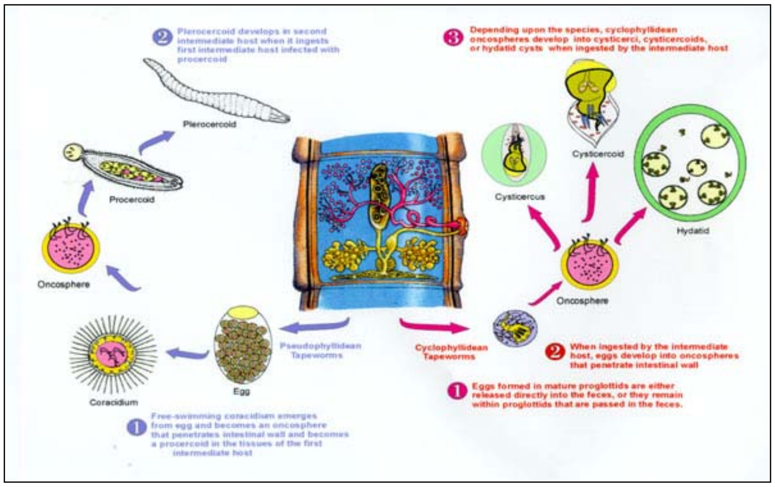
② Plerocercoid develops in second intermediate host when it ingests first intermediate host infected with procercoid
Plerocercoid
Procercoid
Oncosphere
Coracidium
Egg
Pseudophyllidean Tapeworms
① Free-swimming coracidium emerges from egg and becomes an oncosphere that penetrates intestinal wall and becomes a procercoid in the tissues of the first intermediate host
③ Depending upon the species, cyclophyllidean oncospheres develop into cysticerci, cysticercoids, or hydatid cysts when ingested by the intermediate host
Cysticercoid
Cysticercus
Hydatid
Oncosphere
Cyclophyllidean Tapeworms
② When ingested by the intermediate host, eggs develop into oncospheres that penetrate intestinal wall
① Eggs formed in mature proglottids are either released directly into the feces, or they remain within proglottids that are passed in the feces.

Figure 2. Variations in the life cycle patterns of cestodes.

holdfast develop on an adult-like scolex. Human infection occurs when the second intermediate host is ingested.

Cyclophyllidean Life Cycle

The tapeworms in this group produce eggs that are usually fully embryonated and infective when discharged from the definitive host, and only rarely is there an aquatic stage in the life cycle. The eggs do not hatch until eaten by an intermediate host. The oncosphere that hatches from the egg penetrates the host's gut and develops into one of several morphological types of larvae characteristic of different species of tapeworms (Fig. 2). The following types of larvae are distinguished:

1. Cysticercoid: A cysticercoid larva consists of an anterior vesicle containing the scolex, which is not invaginated, and a tail-like posterior region containing the embryonic hooks. This is the larval stage in the life cycles of *Hymenolepis nana*, *H. diminuta* and *Dipylidium caninum*. This stage is usually found in invertebrate intermediate hosts. The term cercocystis is usually applied to a cysticercoid with a tail-like appendage from the bladder.
2. Cysticercus: A cysticercus is a larval stage characterized by a fluid-filled bladder surrounding a single scolex, retracted and invaginated within itself. This is a stage in the life cycles of some species of *Taenia*. The larval stage is usually found in vertebrates.
3. Coenurus: This larval stage is comprised of groups of scolices (protoscolices) which are produced by budding directly from the inner wall of a large bladder (occasionally called a cyst). The protoscolices, however, remain connected by a stalk to the wall of the bladder. This larval stage is characteristic of *Taenia multiceps*, and is found in vertebrates.

4. Hydatid: This larval stage is comprised of many (up to several million) protoscolices which usually develop by budding from the inner wall of vesicles of "brood capsules". Several liters of fluid may be produced within a hydatid. This larval form is known only for the members of the genus *Echinococcus* (Fig. 2).

Pseudophyllideans

Diphyllobothrium latum is the broad or fish tapeworm of humans. The parasite has a worldwide distribution, but a high incidence of infection occurs in Scandinavia, Finland, Alaska, and around the Great Lakes in both the United States and Canada. Other endemic areas include central Africa, parts of Asia, northern Chile and Argentina, and in New South Wales, Australia. Definitive hosts include humans, dogs, cats, pigs, and other fish-eating mammals.

The adult tapeworm lives primarily in the ileum, attached to the intestinal mucosa by an elliptical scolex bearing two bothria. The worm may be ten or more meters in length, and possess up to 4000 proglottids. Usually only one tapeworm is found per host. The life cycle requires two intermediate hosts. Unembryonated eggs, measuring about 40-50 µm wide by 60-70 µm long are continuously discharged from the gravid proglottids through the uterine pores and pass in the feces. Embryonation occurs in fresh-water and, in one to two weeks, the ciliated coracidium hatches through the operculum and swims about. In order for development to proceed, the coracidium must be ingested (within 12 to 24 hours) by the first intermediate host, copepods of the genera *Diaptomus* or *Cyclops*, in which development of the procercoid is completed (Fig. 3). The procercoid averages about 500 µm in length. When an infected copepod is ingested by freshwater fish, including pike, perch, trout, and salmon, the liberated procercoid burrows through the intestinal wall and into the flesh of the fish, where it develops into a plerocercoid. Mature plerocercoids vary in length from a few millimeters to several centimeters. Infection of humans occurs when fish harboring plerocercoids are consumed. Upon their release in the intestine of the definitive host, the parasites attach to the intestinal wall and develop to sexual maturity in 5-6 weeks.

Except for some nonspecific abdominal symptoms, most infections are symptomatic. Clinical B12 deprivation occasionally develops as a result of attachment of the parasite to the proximal portion of the jejunum. Quinicrine hydrochloride and niclosamide have been used to treat diphyllobothriasis.

Two other species, *D. chordatum* and *D. pacificum*, parasitize sea lions in northern and southern hemispheres and are apparently acquired by humans from eating raw marine fish. *Diphyllobothrium ursi* of bears also occurs in humans in Alaska. *Diplogonoporus grandis*, a parasite of whales, is frequently encountered in humans in Japan. Infection is probably acquired from eating raw anchovies, sardines, or other marine fish containing plerocercoids.

Sparganosis

Sparganosis is a disease of humans caused by migrating plerocercoids or spargana of certain diphyllobothrid tapeworms that, as adults, normally parasitize other mammals. Several species of *Spirometra* are intestinal tapeworms of feline and canine hosts. The first intermediate host is Cyclops, and various vertebrates such as frogs,

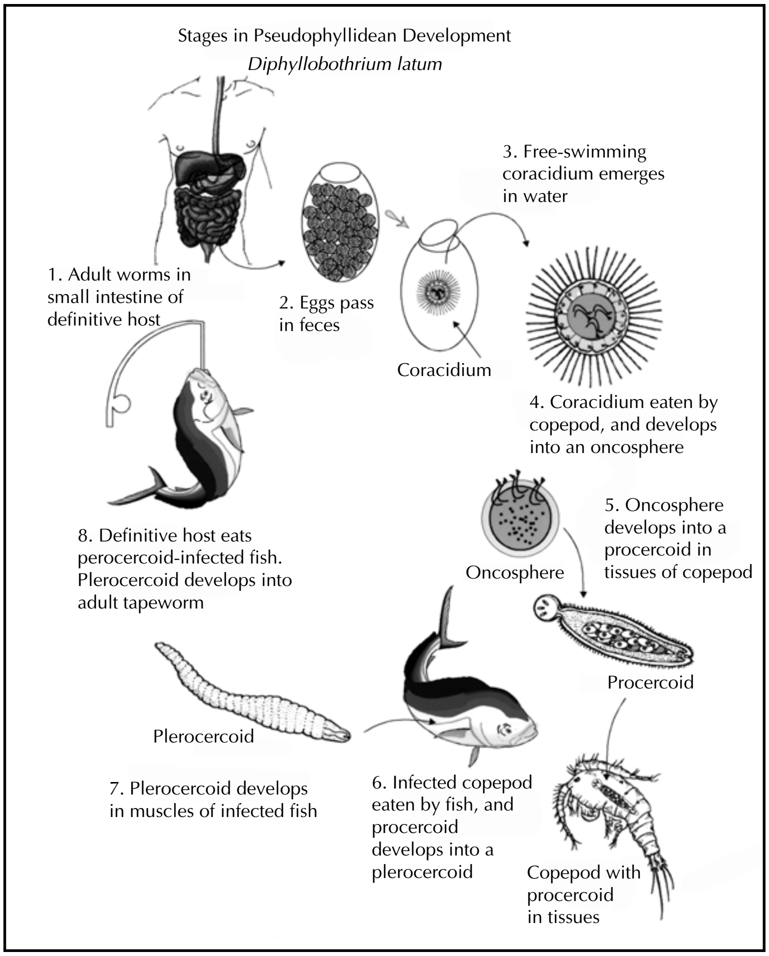
Stages in Pseudophyllidean Development
Diphyllobothrium latum
1. Adult worms in small intestine of definitive host
2. Eggs pass in feces
3. Free-swimming coracidium emerges in water
Coracidium
4. Coracidium eaten by copepod, and develops into an oncosphere
5. Oncosphere develops into a procercoid in tissues of copepod
Oncosphere
Procercoid
Copepod with procercoid in tissues
6. Infected copepod eaten by fish, and procercoid develops into a plerocercoid
Plerocercoid
7. Plerocercoid develops in muscles of infected fish
8. Definitive host eats perocercoid-infected fish. Plerocercoid develops into adult tapeworm

Figure 3. Life cycle of the pseudophyllidean *Diphyllobothrium latum*, the fish tapeworm of humans.

snakes, birds, pigs, and other mammals normally serve as second intermediate hosts by harboring the plerocercoids. Human infection is usually acquired by drinking water containing procercoid-infected copepods. The released procercoids migrate through the gut and develop into spargana in muscles or subcutaneous tissues. Infection may result also from the consumption of inadequately cooked tissues of second intermediate hosts. The viable plerocercoids present in these hosts may become established in humans. In many regions of the Orient, snakes and tadpoles traditionally are consumed raw as a therapeutic measure for various illnesses. Infection may be acquired from the Oriental practice of poulticing open wounds, inflamed eyes or vagina, and numerous other lesions, with frog or snake flesh incidentally contaminated with plerocercoids. The larva penetrates the poulticed lesion and becomes established.

Spargana may persist in human tissues for several years and attain a length up to several centimeters. During development, they may produce severe inflammatory responses in surrounding tissues. In China and Vietnam, ocular sparganosis is a very serious infection, producing intense pain, periorbital edema, excessive lacrimation, nodule formation and corneal ulceration. Treatment of most cases of sparganosis is by surgical removal of the parasite (Fig. 4). A few cases are treated by chemotherapy.

Cyclophyllidean Life Cycles

Taenia saginata, the beef tapeworm, is the most common taeniid of humans. The parasite has a cosmopolitan distribution, occurring where inadequately cooked beef is eaten and also where sanitation is given little attention. Adult worms are up to 25 meters long, but worms slightly less than one-third this size are more common. The scolex bears four muscular suckers. The terminal gravid proglottids, each of which contains about 100,000 eggs, detach from the strobila in the intestine of humans and pass out with the feces. Some proglottids exhibit marked individual activity, migrating out through the anus and onto the skin in the perianal region. At this stage, they may be mistaken for adult trematodes. When the evacuated proglottids rupture, fully embryonated eggs are discharged. The spherical eggs remain infective for several weeks. Various ruminants serve as intermediate hosts, including cattle, goat, bison, sheep, giraffe, and llamas.

Following ingestion of the eggs by a ruminant, the outer shell disintegrates in the duodenum releasing the six-hooked oncosphere. Histologic secretions from the oncosphere facilitate its penetration through the mucosa and into the circulation. The parasite enters muscles in various regions of the body and develops into an infective cysticercus in about 2 months. *Cysticerci*, which measure up to 10 mm long by 6 mm wide, remain viable up to 9 months. Eventually, the cysticerci are destroyed by what appears to be a host response involving the development of a fibrous capsule around the parasites. Prior to being identified as a larval stage, the cysticercus of *T. saginata* was given a separate species status and referred to as *Cysticercus bovis*. The disease in cattle was known as cysticercosis bovis.

Human infection is acquired through the ingestion of raw or inadequately cooked beef containing cysticerci (Fig. 5). A period of about three months is required for the ingested cysticerci to develop into sexually mature worms in the intestine. Although autoinfection is possible if eggs voided in the feces are swallowed, human cysticercosis, involving *T. saginata*, is rarely encountered. Cattle and other ruminants become infected when their feed and feeding areas are contaminated with the egg-laden feces of infected humans. The indiscriminate release of raw sewage into lakes and rivers contributes significantly to the dissemination of the eggs of the parasite.

Taeniasis caused by *T. saginata* may be characterized by nausea, persistent abdominal pain, localized sensitivity to touch, diarrhea, or alternating diarrhea and constipation, loss of appetite and weight, chronic indigestion, and hunger pains. Intestinal obstruction is rarely observed. Psychological stress resulting from gravid proglottids migrating out of the anus of an infected individual may be severe.

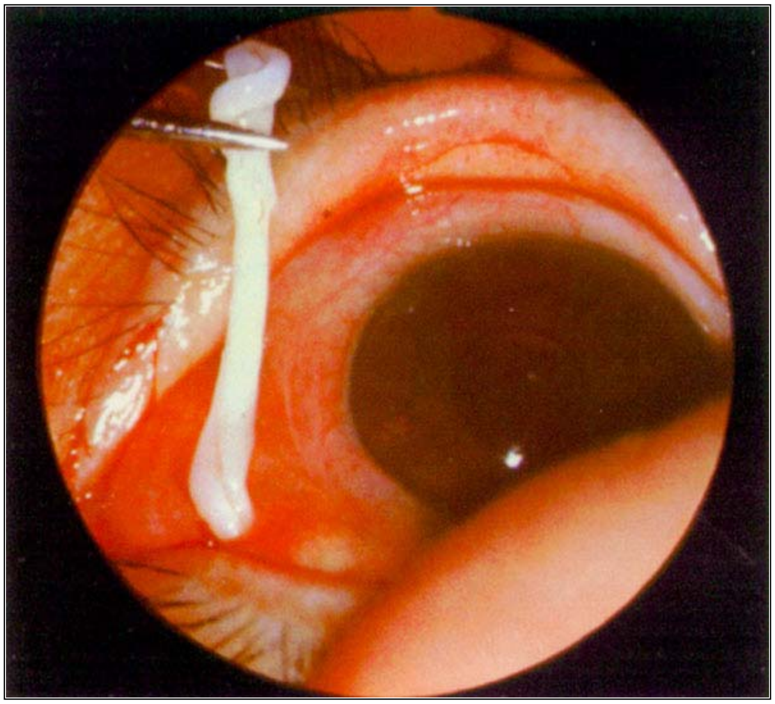

Figure 4. Ocular sparganosis. FromYamaguchi, T. (Ed.) 1981. A Color Atlas of Clinical Parasitology. (Wolfe Medical Publ. Ltd.) Lea & Febiger, Pennsylvania.

The pork tapeworm, *Taenia solium*, has a worldwide distribution. Adult worms, which are parasitic in humans only, are up to 7 meters long. The scolex bears four suckers and a double crown of prominent rostellar hooks. Gravid proglottids detach from the strobila and are evacuated with the feces. Each proglottid contains about 30,000 eggs. The eggs, which are indistinguishable from those of *T. saginata*, are fully embryonated when discharged from the host's intestine (Fig. 6). The life cycle is similar to that of *T. saginata* except that pigs are the intermediate hosts harboring the cysticerci (*Cysticercus cellulosae*) (Fig. 5).

Taenia solium is potentially the most serious tapeworm of humans, primarily because individuals harboring the adult worm may also develop cysticercosis (Fig. 7) through autoinfection. Unlike *T. saginata*, the cysticerci of *T. solium* develop readily in various tissues of humans, which then become potential intermediate hosts (Fig. 8). Both internal and external autoinfection are possible. External autoinfection results when embryonated eggs are ingested. This frequently results from the fecal contamination of food and drink. Internal autoinfection occurs when detached proglottids are carried by reverse peristalsis from the small intestine into the stomach. Here the released eggs hatch and the oncospheres migrate into the body and undergo development to become cysticerci. Light infections may be asymptomatic.

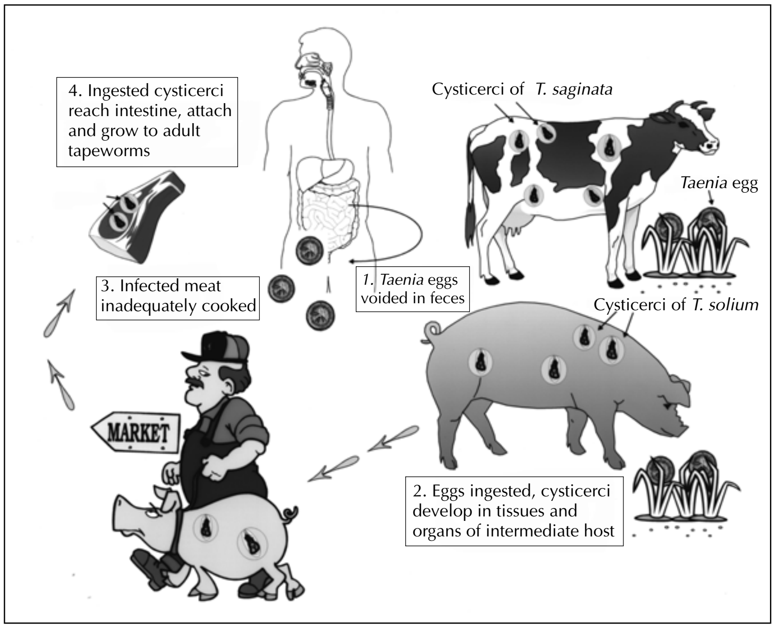
4. Ingested cysticerci reach intestine, attach and grow to adult tapeworms
Cysticerci of *T. saginata*
Taenia egg
3. Infected meat inadequately cooked
1. *Taenia* eggs voided in feces
Cysticerci of *T. solium*
MARKET
2. Eggs ingested, cysticerci develop in tissues and organs of intermediate host

Figure 5. Life cycles of the beef tapeworm, *Taenia saginata*, and the pork tapeworm, *T. solium*.

Apparently, no tissue or organ is resistant to penetration by cysticerci. In heavy infections, the larvae may develop in the brain, heart or eyes, causing numerous serious problems such as paralysis, epileptic attacks, disequilibrium, blindness, and hydrocephalus as a result of blockage and inadequate drainage of cerebrospinal fluid. Ocular cysticercosis may be mistakenly diagnosed as a malignancy, resulting in the unwarranted removal of the eye. In various tissues and organs, host cellular reactions occur, and as a result the cysticerci become calcified and/or enveloped by a fibrous capsule. Cysticerci may remain viable for a few years before they begin to degenerate, whereupon an intense local inflammatory response is initiated by the host.

Adults of *Taenia (Multiceps) multiceps* parasitize carnivores, particularly dogs and other canines in many parts of the world. The larval stage, which is similar in morphology to a hydatid and is termed a coenurus, occurs in herbivorous mammals. In sheep, the most common intermediate host, coenuriasis is characterized by a vertigo resulting in an unstable gait or giddiness, hence the designation "gid" for the infection. Occasionally, coenuri develop in the subcutaneous tissues, muscles, brain, spinal cord, and eyes of humans. These infections result from the accidental ingestion of eggs. The oncosphere released from the egg penetrates the intestinal mucosa and enters the general circulation. Multiple scolices then bud from the inner wall of each parasite to form a coenurus. Unfortunately, no effective therapeutic procedure is available.

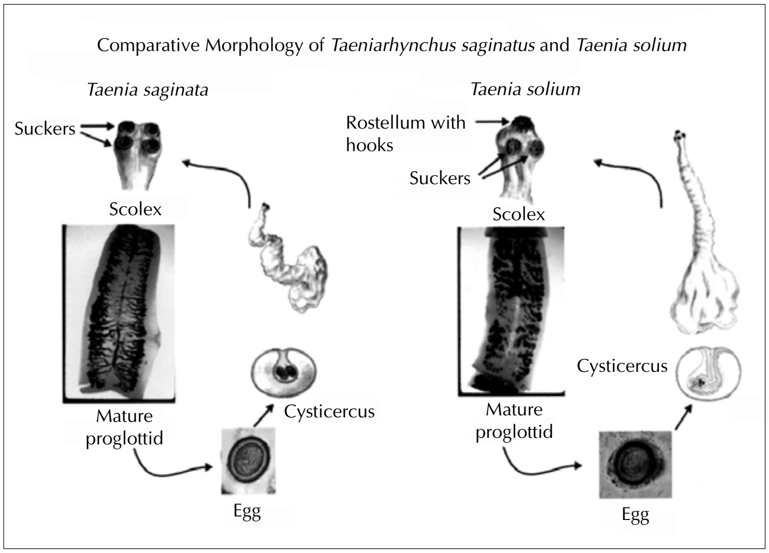
Comparative Morphology of *Taeniarhynchus saginatus* and *Taenia solium*
Taenia saginata
Taenia solium
Suckers
Scolex
Mature proglottid
Cysticercus
Egg
Rostellum with hooks
Suckers
Scolex
Mature proglottid
Cysticercus
Egg

Figure 6. Comparative morphology of *Taenia saginata* and *T. solium*.

Echinococcus *Species and Hydatidosis*

The adult tapeworms of the genus *Echinococcus* are the smallest taeniids. The entire worm is usually less than a centimeter long and is composed of a scolex with an armed rostellum, a neck region, and 3 or 4 proglottids. The adult worms live attached to the intestinal mucosa of various carnivores, especially dogs and other canines, and felines, but not in humans (Fig. 9). The larval stage, or hydatid, is the causative agent of hydatidosis, a disease of considerable medical and veterinary importance. Hydatids occur in virtually any mammal that ingests the eggs of *Echinococcus*. When swallowed by the intermediate host, the eggs hatch and the oncospheres penetrate the gut wall and enter the venules of the hepatic portal veins. Hydatids may develop in any organ or tissue, but commonly are found in the liver, lungs, marrow cavity of long bones, kidneys, spleen, muscles, and brain (Fig. 10). Growing hydatids occupy space, exert pressure, and destroy surrounding host tissues. From the inner germinative surface of the hydatid there develop secondary vesicles or brood capsules, within which the larval stages with immature, inverted scolices (e.g., protoscolices) develop. Occasionally, the parent hydatid cyst forms 'daughter cysts', which then produce brood capsules. Hydatid cysts commonly range in size from 5 to 10 cm and contain several liters of fluid, but cysts up to 50 cm and containing up to 18 liters of fluid have been found. The reproductive potential of the organism is immense, with each liter of fluid containing up to one million protoscolices. Eventually, the brood capsules and daughter cysts break down, and the liberated components, termed hydatid sand, float free in the hydatid fluid. When infected viscera are ingested by canine definitive hosts, the cyst wall is digested,

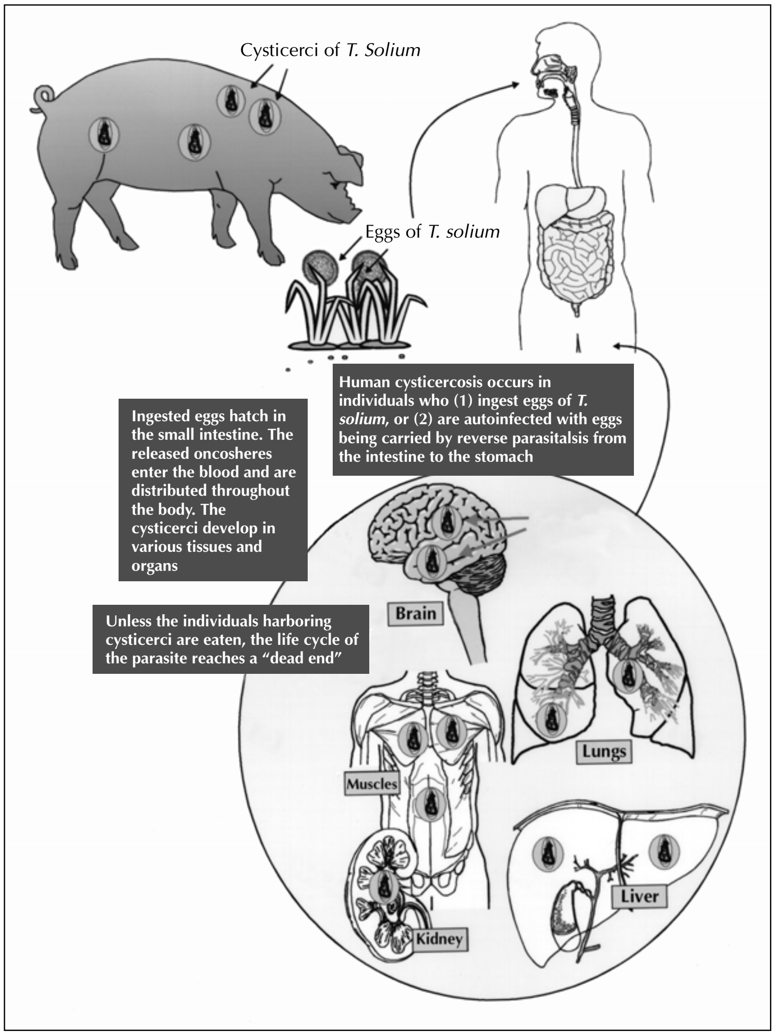

Human cysticercosis occurs in individuals who (1) ingest eggs of *T. solium,* or (2) are autoinfected with eggs being carried by reverse parasitalsis from the intestine to the stomach

Ingested eggs hatch in the small intestine. The released oncosheres enter the blood and are distributed throughout the body. The cysticerci develop in various tissues and organs

Unless the individuals harboring cysticerci are eaten, the life cycle of the parasite reaches a "dead end"

Figure 7. Development of human cysticercosis.

liberating the protoscolices, which then evaginate, attach to the intestinal wall, and develop into adult parasites.

Three species of *Echinococcus* are known to cause hydatidosis in humans. *Echinococcus granulosis* matures in dogs and other canines (but not foxes) and uses mainly herbivores as intermediate hosts. The parasite is cosmopolitan in distribution, but it is especially common in areas where domestic herbivores, such as sheep, pigs, goats, horses, rabbits, camels, and reindeer, are raised in association with dogs. Dogs ac-

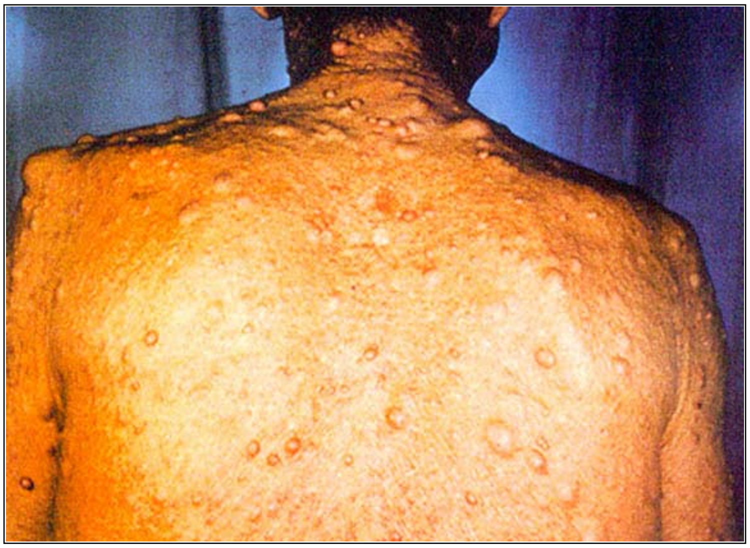

Figure 8. Human cysticercosis. This individual harbors the larval stage of *Taenia solium*, and thus serves as a potential intermediate host for the tapeworm. FromYamaguchi, T. (Ed.) 1981. A Color Atlas of Clinical Parasitology. (Wolfe Medical Publ. Ltd.) Lea & Febiger, Pennsylvania.

quire the infection when they consume the offal of butchered animals, while herbivores become infected when they feed on vegetation contaminated with egg-laden dog feces. Human infection occurs when *Echinococcus* eggs are ingested, usually as a result of fondling infected dogs or handling their scats. A second species, *E. multilocularis*, matures mainly in foxes, but cats, dogs, and other canines also serve as definitive hosts. Small rodents, such as mice, voles, and lemmings, are intermediate hosts. The parasite has been reported from parts of Europe, North and South America, and New Zealand. A third species, *E. oligarthrus*, is a parasite of jaguars, pumas, and other wild felines. Human infections with the hydatid of this parasite have been reported from South and Central America.

The symptoms of hydatidosis depend on the type and location of the hydatid. Hepatic hydatidosis produces jaundice, ascites, and splenomegaly, which frequently results in fatal intrahepatic portal hypertension. Osseous hydatids progressively erode bone causing fractures and eventual decay. The symptoms of cranial hydatids mimic those of intracranial tumors. Chemotherapy is ineffective in hydatid disease. Surgical removal of the cyst is the only successful method of treatment, providing the hydatids are situated in operable locations. Considerable care is required during surgery to prevent the hydatid fluid from escaping into the body cavity, as this may cause secondary infections and fatal anaphylactic shock. Moreover, scolices liberated from ruptured hydatids may become attached to the peritoneum and produce multiple secondary growths. In some cases, it may be possible to first aspirate the fluid with a syringe before excising the hydatids.

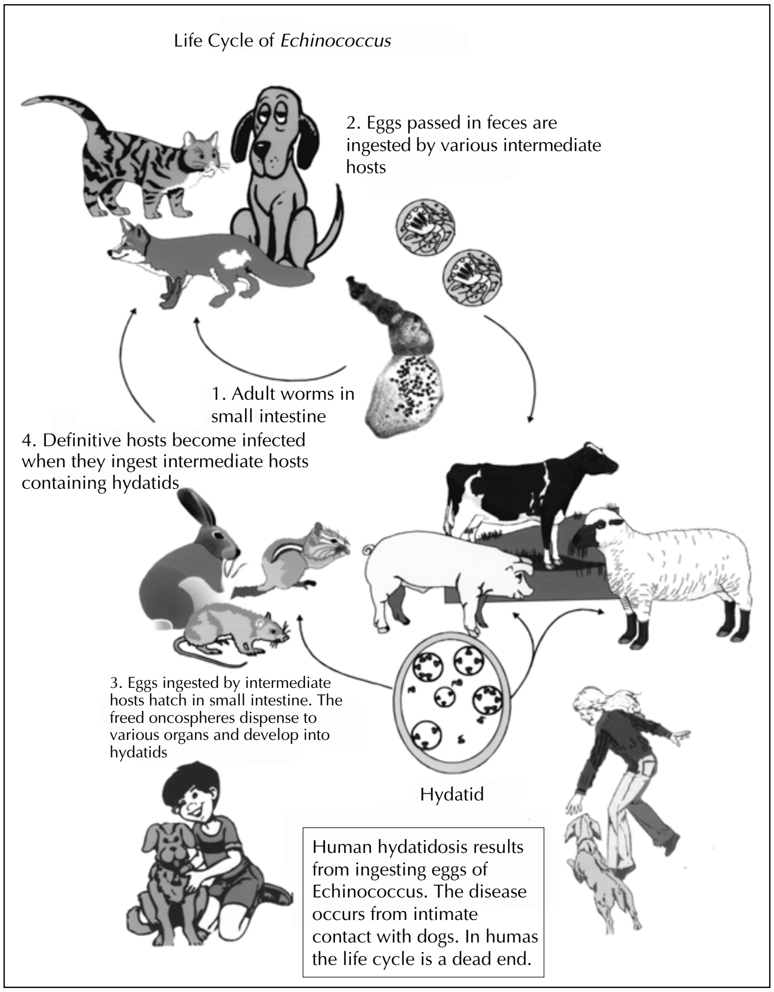
Life Cycle of *Echinococcus*
2. Eggs passed in feces are ingested by various intermediate hosts
1. Adult worms in small intestine
4. Definitive hosts become infected when they ingest intermediate hosts containing hydatids
3. Eggs ingested by intermediate hosts hatch in small intestine. The freed oncospheres dispense to various organs and develop into hydatids
Hydatid
Human hydatidosis results from ingesting eggs of Echinococcus. The disease occurs from intimate contact with dogs. In humas the life cycle is a dead end.

Figure 9. Life cycle of *Echinococcus.*

Dipylidium caninum is a cosmopolitan tapeworm of dogs and cats, measuring up to 50 cm in length. The scolex possesses four conspicuous suckers and a retractable rostellum with several rows of minute hooklets. Detached gravid proglottids are passed with the feces or actively migrate out the anus. On reaching the soil, the proglottids begin to desiccate, releasing egg capsules, each of which contains several embryonated eggs. Fleas, chewing lice, and possibly other arthropods serve as intermediate hosts when they ingest egg capsules (Fig. 11). The parasite invades the host's hemocoel and develops into a cysticercoid. The definitive host becomes in-

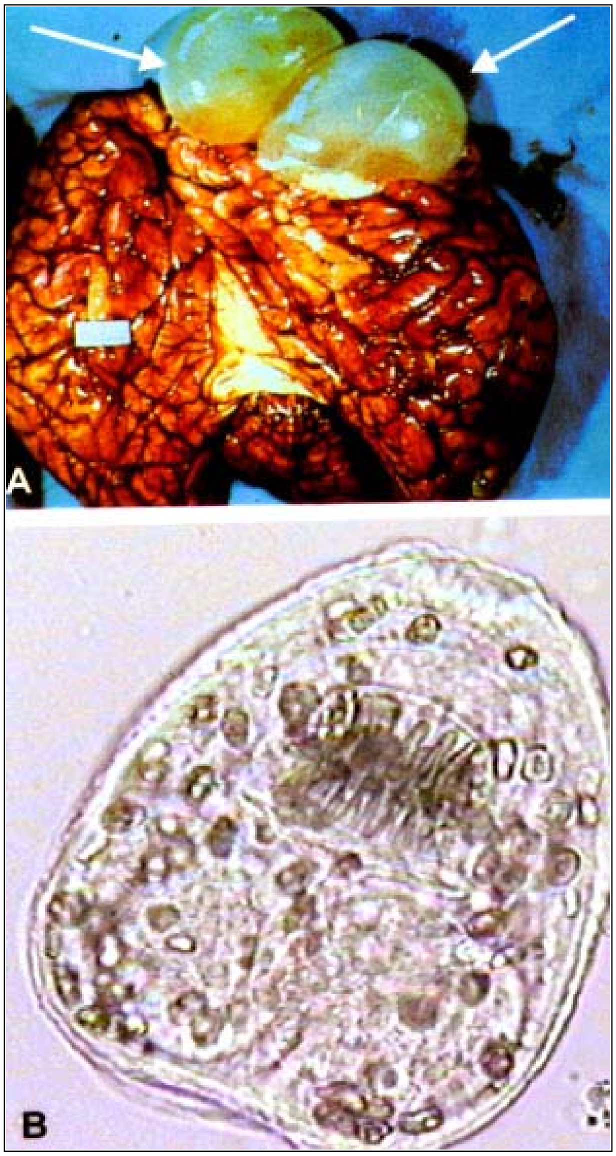

Figure 10. (A) Two fluid filled hydatid cysts (arrows) situated on the surface of the brain. FromYamaguchi, T. (Ed.) 1981. A Color Atlas of Clinical Parasitology. (Wolfe Medical Publ. Ltd.) Lea & Febiger, Pennsylvania. (B) Photograph of a hydatid cyst showing numerous protoscolices that can form daughter cysts. Rupture of the cysts can cause secondary hydatidosis. Courtesy of Dr. Pietro Caramello, and the Carlo Denegri Foundation, Torino, Italy.

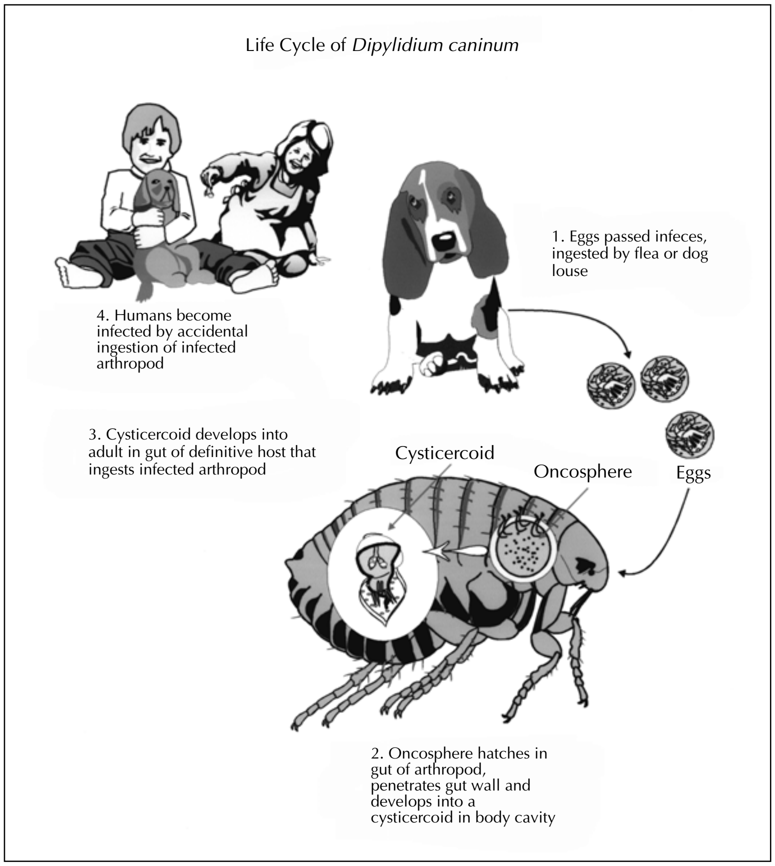
Life Cycle of *Dipylidium caninum*
1. Eggs passed infeces, ingested by flea or dog louse
4. Humans become infected by accidental ingestion of infected arthropod
3. Cysticercoid develops into adult in gut of definitive host that ingests infected arthropod
Cysticercoid
Oncosphere
Eggs
2. Oncosphere hatches in gut of arthropod, penetrates gut wall and develops into a cysticercoid in body cavity

Figure 11. Life cycle of the dog tapeworm, *Dipylidium caninum.*

fected when infected arthropod hosts are consumed. The cysticercoids escape in the intestine and develop directly to adult worms in about one month. Human infections with *D. caninum* frequently involve children who fondle infected animals and accidentally ingest infected fleas.

Hymenolepis (Vampirolepis) nana, the dwarf tapeworm, is a cosmopolitan parasite of rodents and other mammals including humans. The parasite is the most common human tapeworm in the world, occurring most frequently in young children. The parasite has a worldwide incidence of about 1 percent. The adult tapeworm ranges from 0.1 to 10 cm in length. In heavy infections, where crowding affects size, the worms are small, averaging less than 3 cm. The scolex bears four deep acetabula and a retractable rostellum armed with a crown of hooks. The eggs are usually liberated from the gravid proglottids before they detach from the strobila. Eggs are infective to the definitive host, in which both larval and adult stages

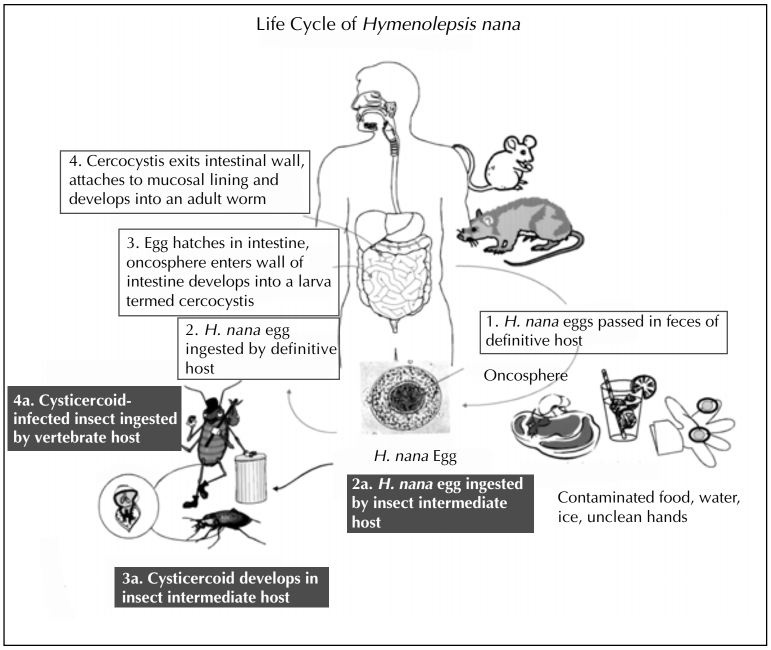
Life Cycle of *Hymenolepsis nana*
4. Cercocystis exits intestinal wall, attaches to mucosal lining and develops into an adult worm
3. Egg hatches in intestine, oncosphere enters wall of intestine develops into a larva termed cercocystis
2. *H. nana* egg ingested by definitive host
1. *H. nana* eggs passed in feces of definitive host
Oncosphere
4a. Cysticercoid-infected insect ingested by vertebrate host
H. nana Egg
2a. *H. nana* egg ingested by insect intermediate host
Contaminated food, water, ice, unclean hands
3a. Cysticercoid develops in insect intermediate host

Figure 12. Life cycle of the dwarf tapeworm, *Vampirolepis (Hymenolepis) nana*.

may develop. When ingested by the definitive host, the eggs hatch in the duodenum and the liberated oncospheres penetrate into the mucosa where they rapidly develop into cysticercoids (Fig. 12). Within one-week post-infection, the cysticercoids emerge into the gut lumen, attach to the mucosa and, in about 2 weeks, develop into sexually mature worms. Human infection is believed to be acquired most commonly by the ingestion of eggs that are voided in the feces. Heavy infections can occur through internal autoinfection due to the hatching of eggs released from adult worms in the small intestine. The life cycle of *H. nana* is of interest in that the parasite may develop to the cysticercoid stage in intermediate hosts, which include a large number of insects (e.g., adult grain or flour beetles, cockroaches, fleas). When these intermediate hosts are consumed by humans (or other suitable definitive hosts), the adult tapeworms develop in the small intestine. Light infections are usually asymptomatic. Abdominal pain, diarrhea, anorexia, vomiting, pruritus of the nose and face and urticaria generally characterize heavy infections.

Hymenolepis diminuta is a parasite of rats and occasionally humans. It is a larger species than *V. nana*, measuring between 20 and 90 cm in length. The scolex possesses the typical tetrad of suckers, but lacks an armed rostellum. Completion of the life cycle requires an arthropod intermediate host, commonly flour moths, stored grain beetles, earwigs, and rodent fleas. The intermediate hosts, which harbor the cysticercoids, become infected from eating flour or cereal foods contaminated by the egg-laden dropping of infected rats. Human infections are associated primarily with the contamination of cereals and other food by infected grain beetles. Infections are generally mild, with occasional indigestion and slight abdominal pain. In some humans, the adult worms are spontaneously evacuated, suggesting an unsuitable host environment.

CHAPTER 7

General Morphology of Parasitic Nematodes

Nematodes are unsegmented worms that typically are elongate and cylindrical in shape with tapered ends. Roundworms vary considerably in size from microscopic to over a meter in length. They possess a fluid-filled cavity or pseudocoel, and a complete digestive system comprised of an anterior mouth, a muscular esophagus (pharynx), an intestine, and a rectum that terminates posteriorly at the anus. The mouth may be surrounded by lips and a buccal cavity with cutting plates or teeth. Parasitic nematodes commonly feed on the semi-liquid contents of the host's alimentary canal, intestinal mucosa, blood or other body fluids, and various lysed tissues. In some forms, the muscular esophagus is cylindrical and virtually of a uniform diameter throughout, and the parasite is termed filariform. This type of esophagus generally characterizes infective stage larvae. In certain other nematodes, the esophagus is expanded posteriorly into a valved bulb, and the parasite is termed rhabditiform (Fig. 1). The latter type frequently characterizes the free-living larval stages. Some nematodes have both filariform and rhabditiform stages in their life cycles. The body is covered by a non-cellular cuticle, which may exhibit longitudinal ridges, striations, wart-like structures, lateral expansions anteriorly and posteriorly, and spines. Before reaching sexual maturity, all nematode larvae undergo a series of four molts or ecdyses. Cuticular structures of some importance are amphids and phasmids. Amphids are a pair of minute sensory organs, considered to be chemoreceptors that open on each side of the head. Phasmids are a pair of caudal (post-anal) organs similar in structure to amphids. Some phasmids are glandular and serve an excretory function, while others are sensory and believed to be involved in chemoreception.

With few exceptions, nematodes are dioecious and exhibit sexual dimorphism, with males generally smaller and possessing a more curved tail than females. Monoecious species may be either parthenogenic or self-fertilizing hermaphrodites. The gonads are tubular cords of cells continuous with the ducts that transport the gametes to the external environment. In males, the intestine and the reproductive tract open into the cloaca. There is a single tubular testis from which extends a vas deferens that terminates in a muscular ejaculatory duct. Most males have a pair of sclerotized copulatory spicules originating from within the cloacal wall that serve as holdfast structures during copulation. The tail may be drawn out into longitudinal ridges called caudal alae, which also assist in holding the female in opposition during copulation. The female reproductive system also is tubular and frequently highly coiled within the body. The system may be composed of a single set of reproductive structures, but commonly there are two sets, and occasionally more, with the following regions differentiated; ovary, oviduct, seminal receptacle, uterus, ovijector, and vagina. The vulva (genital pore) is located ventrally and is independent of the alimentary

---

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

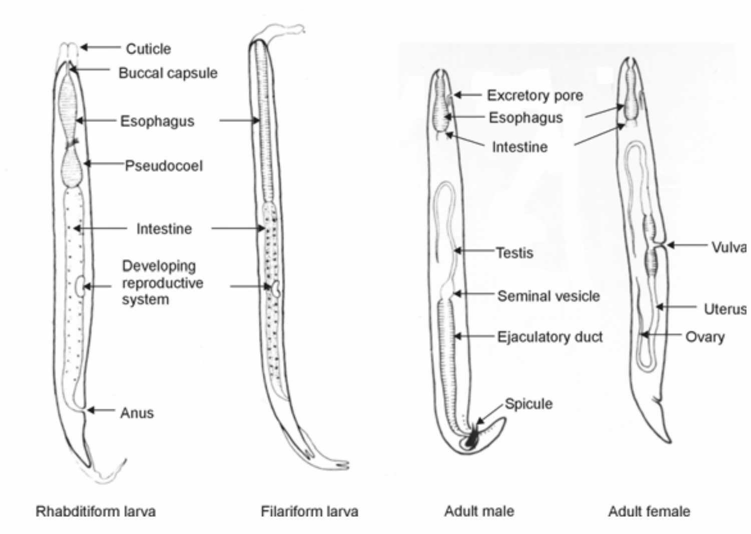
Cuticle
Buccal capsule
Esophagus
Pseudocoel
Intestine
Developing reproductive system
Anus
Excretory pore
Esophagus
Intestine
Testis
Seminal vesicle
Ejaculatory duct
Spicule
Vulva
Uterus
Ovary
Rhabditiform larva
Filariform larva
Adult male
Adult female

Figure 1. Generalized diagram of the larval and adult stages of Nematodes. Modified from Meyer, M. C. and Olsen, O. W. 1980. Essentials of Parasitology. W. C. Brown Co., Dubuque, Iowa.

canal. Experimental alterations in the environment have been shown to induce changes in the sexual characteristics of certain nematodes, including sex reversals.

There are two basic types of life cycles among parasitic nematodes, direct (monoxenous) with only a single host, or indirect (heteroxenous) with two or more hosts involved. Nematode larvae develop through a series of four stages, each separated by a molting of the cuticle. The third stage is generally infective to the final host. In some species one or more larval stages may occur within the egg capsule. In many parasitic species with a direct life cycle, the first three larval stages are free in the soil, where the first two stages feed on soil components. The third stage retains the shed cuticle of the preceding stages as an enclosing sheath and is unable to feed. Oviparous species produce eggs that are released from the definitive host with the feces. In ovoviviparous species, the eggs hatch *in utero* and the larvae are passed in the feces. Some of the variations in the life cycle patterns of parasitic nematodes are given in Table 1.

Trichuris trichiura

Trichuris trichiura, or whipworm, has a cosmopolitan distribution, but is more frequently encountered among the poor in the tropics and subtropics, and where living conditions are crowded and sanitation is poor. In parts of the southeastern United States the incidence of trichuriasis may reach 25%, with young children most commonly infected. The adult worms measure from 3 to 5 cm in length, with females larger than males. The anterior three-fifths of the body is thread-like or

Table 7.1 Some Variations in the Life Cycles of Human Nematodes

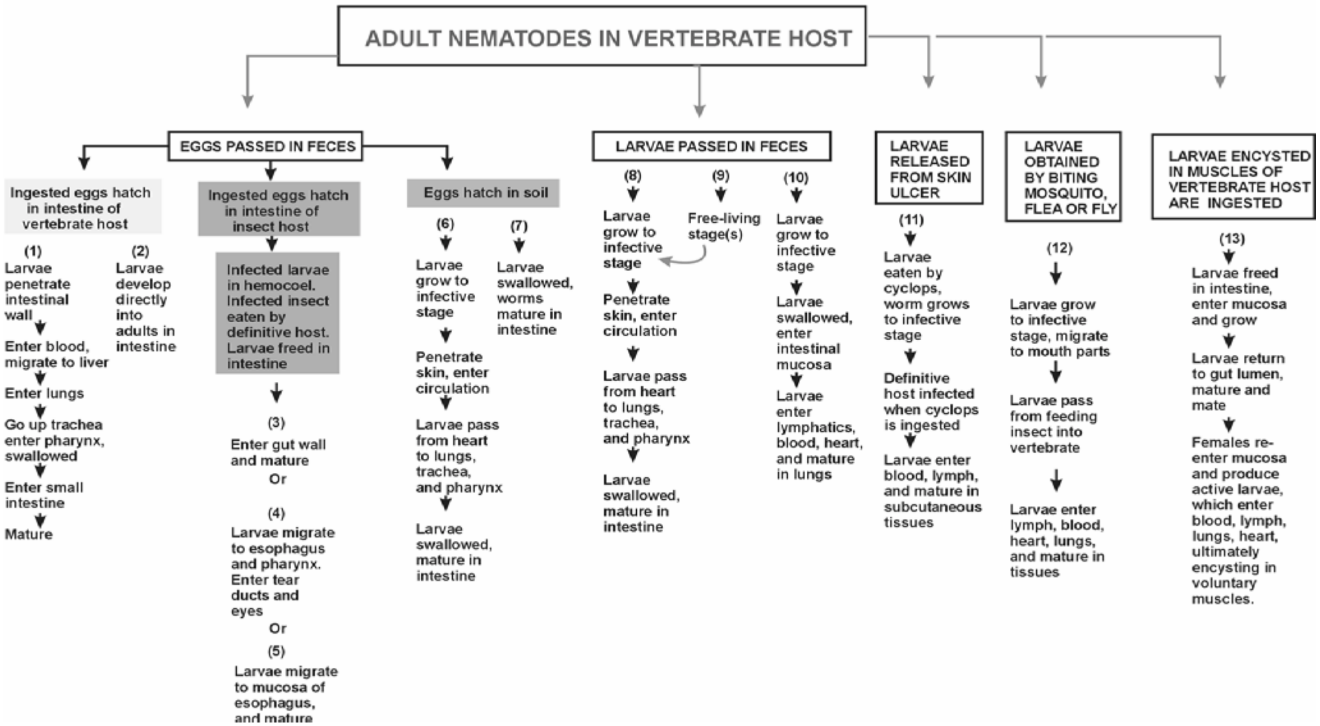
ADULT NEMATODES IN VERTEBRATE HOST
EGGS PASSED IN FECES
Ingested eggs hatch in intestine of vertebrate host
(1)
Larvae penetrate intestinal wall
Enter blood, migrate to liver
Enter lungs
Go up trachea enter pharynx, swallowed
Enter small intestine
Mature
(2)
Larvae develop directly into adults in intestine
Ingested eggs hatch in intestine of insect host
Infected larvae in hemocoel. Infected insect eaten by definitive host. Larvae freed in intestine
(3)
Enter gut wall and mature
Or
(4)
Larvae migrate to esophagus and pharynx. Enter tear ducts and eyes
Or
(5)
Larvae migrate to mucosa of esophagus, and mature
Eggs hatch in soil
(6)
Larvae grow to infective stage
Penetrate skin, enter circulation
Larvae pass from heart to lungs, trachea, and pharynx
Larvae swallowed, mature in intestine
(7)
Larvae swallowed, worms mature in intestine
LARVAE PASSED IN FECES
(8)
Larvae grow to infective stage
Penetrate skin, enter circulation
Larvae pass from heart to lungs, trachea, and pharynx
Larvae swallowed, mature in intestine
(9)
Free-living stage(s)
(10)
Larvae grow to infective stage
Larvae swallowed, enter intestinal mucosa
Larvae enter lymphatics, blood, heart, and mature in lungs
LARVAE RELEASED FROM SKIN ULCER
(11)
Larvae eaten by cyclops, worm grows to infective stage
Definitive host infected when cyclops is ingested
Larvae enter blood, lymph, and mature in subcutaneous tissues
LARVAE OBTAINED BY BITING MOSQUITO, FLEA OR FLY
(12)
Larvae grow to infective stage, migrate to mouth parts
Larvae pass from feeding insect into vertebrate
Larvae enter lymph, blood, heart, lungs, and mature in tissues
LARVAE ENCYSTED IN MUSCLES OF VERTEBRATE HOST ARE INGESTED
(13)
Larvae freed in intestine, enter mucosa and grow
Larvae return to gut lumen, mature and mate
Females re-enter mucosa and produce active larvae, which enter blood, lymph, lungs, heart, ultimately encysting in voluntary muscles.

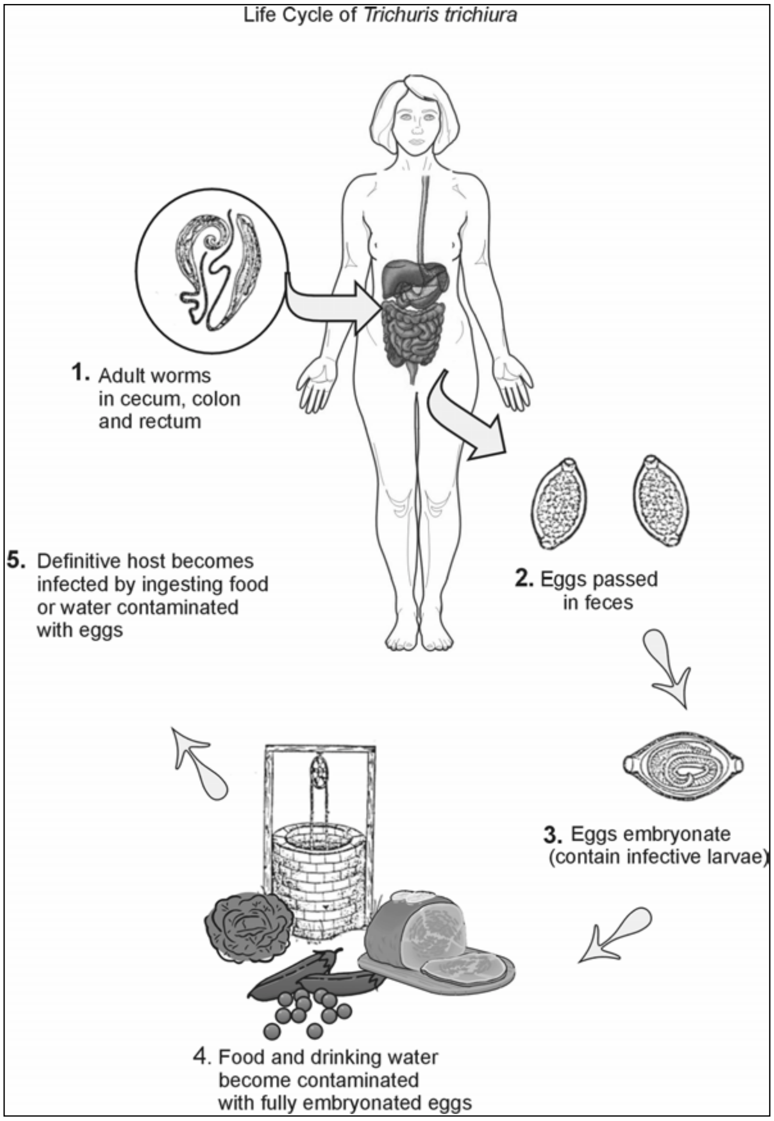
Life Cycle of *Trichuris trichiura*
1. Adult worms in cecum, colon and rectum
2. Eggs passed in feces
3. Eggs embryonate (contain infective larvae)
4. Food and drinking water become contaminated with fully embryonated eggs
5. Definitive host becomes infected by ingesting food or water contaminated with eggs

Figure 2. Life cycle of the whipworm, *Trichuris trichura*.

filariform, with the remaining portion abruptly expanded, resembling a lash or whip with a thick handle.

Adult whipworms may live several years embedded primarily in the mucosa of the cecum (Fig. 2). In heavy infections, however, the worms can be found elsewhere

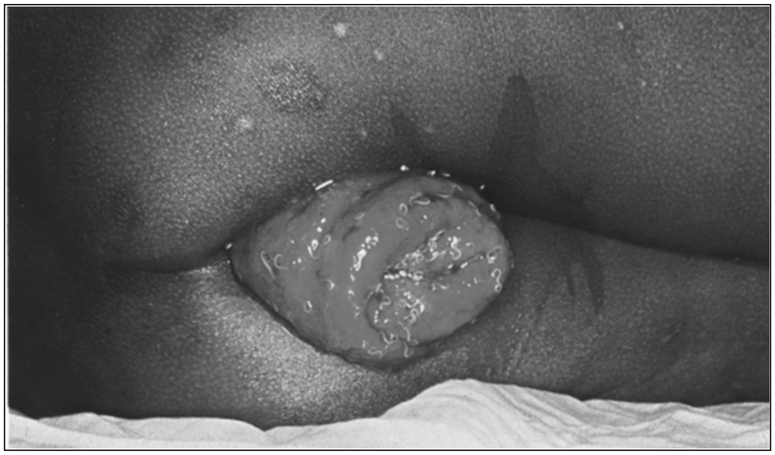

Figure 3. Prolapse of the rectum due to *Trichuris trichura*. From Zaman, V. 1980. Atlas of Medical Parasitology. Singapore University Press.

in the colon, including the rectum. The females may produce up to 7,000 eggs daily. Embryonation occurs outside the body of the host and requires a period of up to 3 weeks. Infection is acquired by ingestion of fully embryonated eggs. The eggs are elongate with a prominent plug at each end. In the small intestine of the definitive host, the eggs hatch and the larvae temporarily penetrate the intestinal wall. After a brief period of development, the larvae re-enter the intestinal lumen and migrate to the cecum where they attach and mature in a few months. The anterior regions of the worms are embedded in the mucosa, where the nematodes feed essentially on blood and lysed tissues. Light infections may be asymptomatic, or characterized by abdominal pain, diarrhea, constipation, vomiting, flatulence and fever. Symptoms of heavy infections may include bloody diarrhea, colitis, and prolapse of the rectum (Fig. 3).

Trichinella spiralis

Trichinella spiralis is one of the more clinically important roundworm parasites. In addition to humans, other natural vertebrate hosts include swine, dogs, cats, mice, raccoons, muskrats, squirrels, opossums, bears, foxes, bobcats, and coyotes. The life cycle is of interest in that one organism serves as both intermediate and definitive host, with all stages of the parasite present, but in different organs. New infections are dependent on the ingestion of infective larvae encased in the muscles of the first definitive host. The parasite has an extensive distribution, but is more common in areas where garbage containing uncooked pork scraps is fed to swine. It is estimated that 35 million persons in the United States are infected with trichinosis.

Adult *T. spiralis* is thread-like in shape and barely visible to the unaided eye. Females measure 3 to 5 mm in length, and males average slightly less than 2 mm in length. The parasites have a life span of a few weeks to three months. Human infection occurs by the ingestion of raw or insufficiently cooked meat, chiefly pork, con-

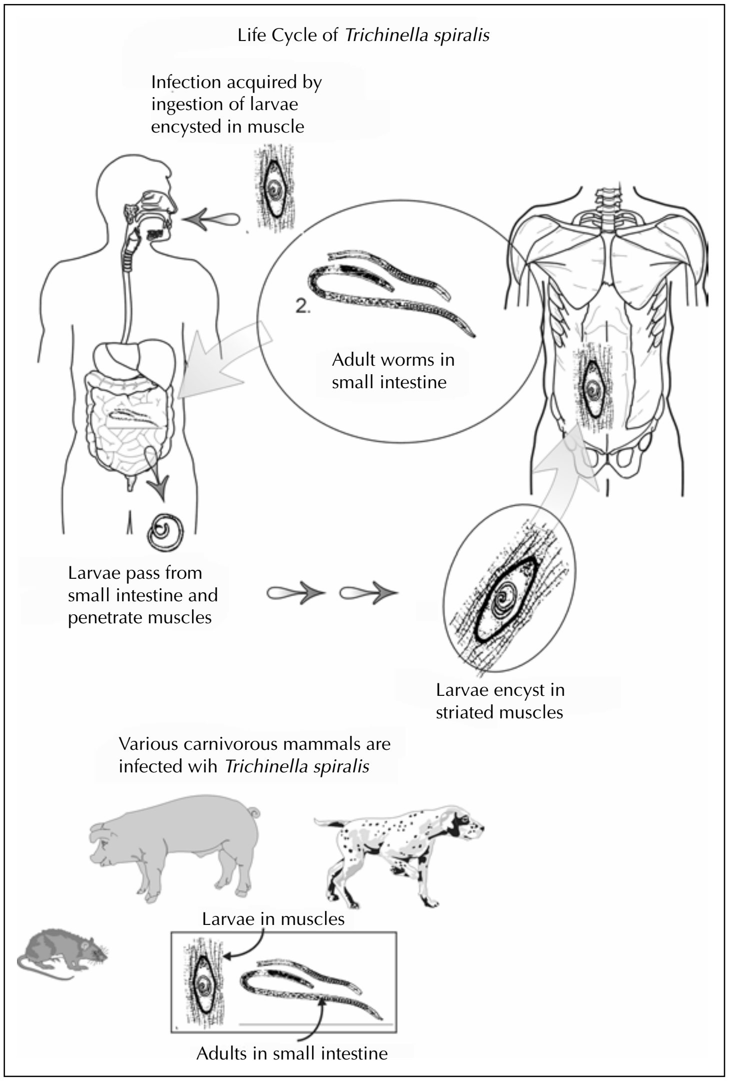
Life Cycle of *Trichinella spiralis*
Infection acquired by ingestion of larvae encysted in muscle
2.
Adult worms in small intestine
Larvae pass from small intestine and penetrate muscles
Larvae encyst in striated muscles
Various carnivorous mammals are infected wih *Trichinella spiralis*
Larvae in muscles
Adults in small intestine

Figure 4. Life cycle of *Trichinella spiralis*.

taining the infective larvae encysted in the muscle. Occasionally, infections result from ingestion of bear or walrus meat, and beef contaminated by a meat grinder previously used to grind infected pork. The larvae excyst in the small intestine and become attached to the mucosa and grow to the adult stage in less than two days, at which time mating takes place (Fig. 4). The intestinal phase of the infection generally

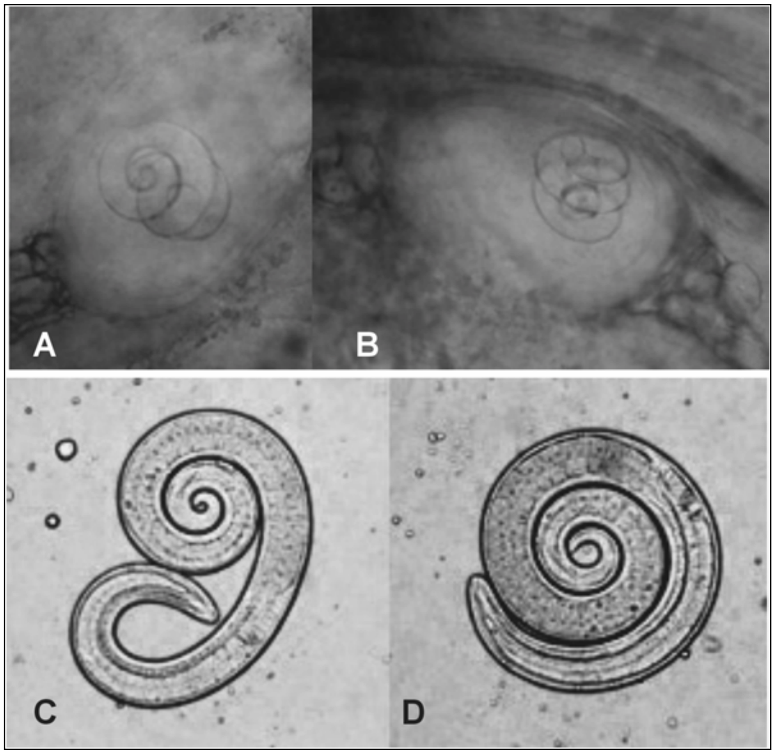

Figure 5. Larvae of *Trichinella* coiled and encysted in muscle (A and B), and freed from their cysts (C and D). Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

lasts less than two months. Within three days after copulation, the ovoviviparous females burrow deep into the intestinal wall and deposit larvae, which first enter the lymphatics and then are transported through the general circulation to various organs. The larvae penetrate the sarcolemma of skeletal muscle fibers in which they coil up and become encysted (Fig. 5). Fully developed larvae measure about 1 mm in length. The life cycle is completed when the encysted larvae are consumed by a suitable definitive host, in which the larvae are freed from their capsules (Fig. 5). Some of the encysted larvae may remain viable and dormant for many years, even after the capsules become calcified, which usually occurs 6-18 months post-infection. The muscles most commonly infected include the diaphragm, intercostals, abdominal, pectoral, gastrocnemius, deltoid, larynx, and base of the tongue.

Manifestations of trichinosis vary from asymptomatic cases to fatal infections. Fever and gastrointestinal complaints, particularly diarrhea, may be common during the early intestinal phase. During muscle invasion and encystment there may be severe myositis producing difficulty in breathing, mastication and speech, hemor-

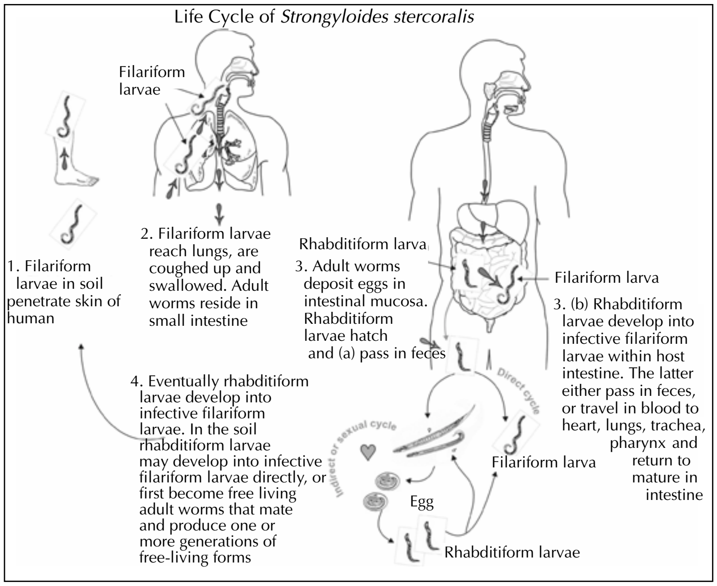
Life Cycle of *Strongyloides stercoralis*
Filariform larvae
1. Filariform larvae in soil penetrate skin of human
2. Filariform larvae reach lungs, are coughed up and swallowed. Adult worms reside in small intestine
Rhabditiform larva
3. Adult worms deposit eggs in intestinal mucosa. Rhabditiform larvae hatch and (a) pass in feces
Filariform larva
3. (b) Rhabditiform larvae develop into infective filariform larvae within host intestine. The latter either pass in feces, or travel in blood to heart, lungs, trachea, pharynx and return to mature in intestine
Direct cycle
4. Eventually rhabditiform larvae develop into infective filariform larvae. In the soil rhabditiform larvae may develop into infective filariform larvae directly, or first become free living adult worms that mate and produce one or more generations of free-living forms
Indirect or sexual cycle
Filariform larva
Egg
Rhabditiform larvae

Figure 6. Life cycle of *Strongyloides stercoralis*.

rhages of the skin, mucous membranes, lungs, circumorbital edema, vasculitis, and neurological manifestations. On rare occasions, the larvae may penetrate the cardiac muscles, and pass through the cerebral capillaries causing motor and psychic disturbances.

Strongyloides stercoralis

Strongyloides stercoralis is an intestinal parasite of humans, dogs, cats, and other mammals. The parasite is commonly found in tropical and subtropical areas and is also sporadically reported from temperate and cold climates. A second species, *S. fuelleborni,* occurs commonly in central and east Africa. Both parasitic and free-living generations are formed. Parasitic females are protandrogonous, their reproductive organs developing after the male reproductive organs have disappeared. These females, which measure up to 2.5 mm in length, burrow into the intestinal mucosa where they lay partially embryonated eggs. Occasionally, the parasites are found in the bronchial passages, and biliary and pancreatic systems. The eggs hatch within the submucosa or during passage through the lumen of the intestine, liberating rhabditiform larvae, which are then voided with the feces. However, if delayed in their movement down the digestive tract, the rhabditiform larvae may molt and transform into filariform larvae, which penetrate the colonic mucosa or perianal skin, migrate via the blood to the lungs and eventually to the small intestine where they mature to the adult stage. This developmental cycle, termed autoinfection, may account for some infections persisting for several years.

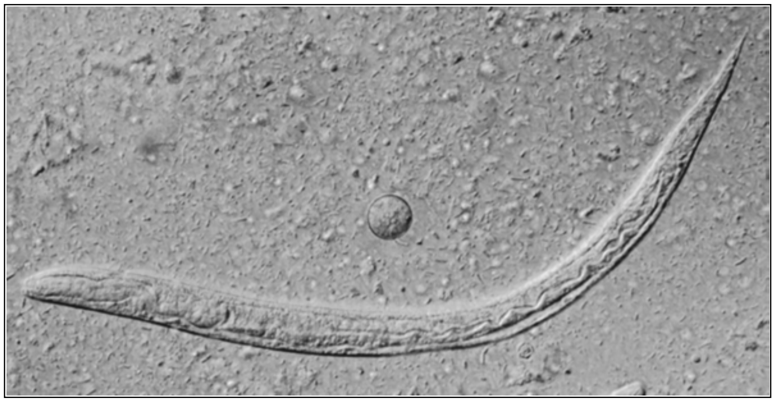

Figure 7. *Strongyloides stercoralis.* Note also cyst of *Entamoeba coli* (arrow). Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

The rhabditiform larvae which pass out of the host in the feces may either transform into infective filariform larvae directly (homogonic cycle), or when environmental conditions are optimal, develop into free-living adults that mate and produce eggs that hatch into rhabditiform larvae (Figs. 6 and 7). This indirect or sexual cycle (heterogonic cycle) may be repeated indefinitely before infective filariform larvae are produced. In the soil, the filariform larvae are unable to develop further. The life cycle continues when the filariform larvae enter a host, either by skin penetration or ingestion. If infection occurs by penetration of the skin of the host, the larvae enter the blood and are transported to the lungs where they pass from the capillaries into the alveoli, up the bronchial passages to the pharynx where they are coughed up, swallowed, and later complete their maturation in the small intestine. A lung migration apparently does not occur when infection is acquired by ingestion of the filariform larvae.

Slight hemorrhage, swelling and itching are sometimes noted at the sight of penetration of the skin by the invasive filariform larvae. Other symptoms of strongyloidiasis include pneumonitis, anemia, abdominal pain, ulceration and sloughing of the intestinal mucosa, and occasionally death resulting from septicemia. Intestinal malabsorption syndrome with steatorrhea is sometimes observed. Thiabendazole is currently the drug used in treatment of strongyloidiasis.

Hookworms

Several species of hookworm infect domestic and wild animals. Two species, *Ancylostoma duodenale* and *Necator americanus*, are parasitic in humans. Both species are found around the world, but *A. duodenale* or Old World hookworm occurs mainly in Europe, Asia and Africa, and is found only in scattered areas of the Caribbean Islands, South America and the United States. *Necator americanus*, the New World or American hookworm is the predominant species in the United States. In *Ancylostoma*, the buccal capsule contains paired tooth-like processes (Fig. 8), while in *Necator* the buccal capsule is provided with semilunar cutting plates. Adult males

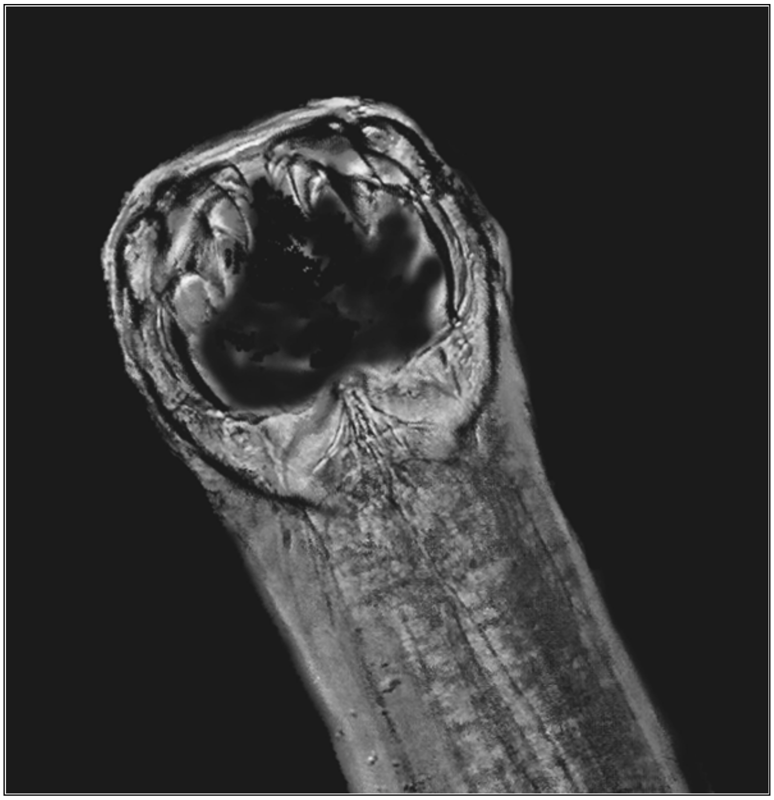

Figure 8. Buccal capsule of the hookworm *Ancylostoma caninum* illustrating three pairs of cutting teeth. From Zaman, V. 1980. Atlas of Medical Parasitology. Singapore University Press.

are 7 to 11 mm long, and females 8 to 13 mm long. The adults live in the small intestine attached to the mucosa, feeding on blood and tissue fluids. The eggs are passed in the feces and under optimal conditions of moisture and temperature, rhabditiform larvae hatch within 24 hours. The rhabditiform larvae feed on fecal matter, bacteria, and organic material in the soil. Eventually, they transform into third-stage, non-feeding, filariform or infective larvae. The filariform larvae penetrate the skin of the host, and enter the circulatory system. The larvae are carried by the blood to the heart and into the lungs (Fig. 9). They emerge from the capillaries into the alveolar spaces, move up the bronchial passages to the pharynx and are swallowed. The worms attach to the intestinal mucosa where they molt twice to become sexually mature. The worms mate and produce eggs that are passed in the feces. Three to six weeks are required from the time of infection to the appearance of eggs in the feces. In some cases, transplacental transmission and transmission in breast milk are known to occur.

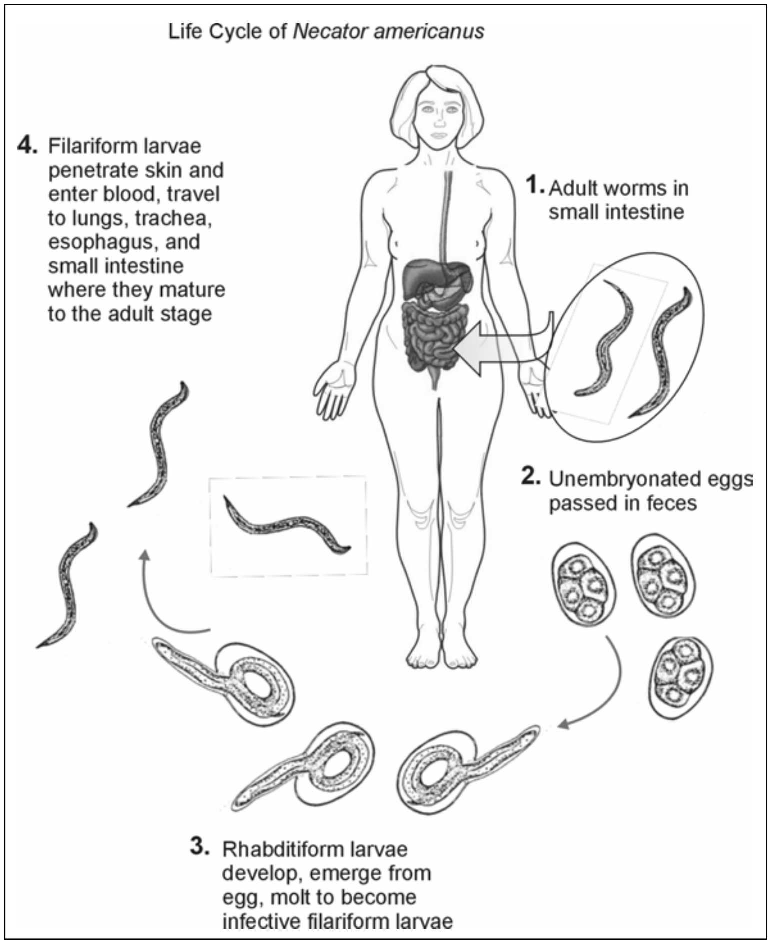
Life Cycle of *Necator americanus*
4. Filariform larvae penetrate skin and enter blood, travel to lungs, trachea, esophagus, and small intestine where they mature to the adult stage
1. Adult worms in small intestine
2. Unembryonated eggs passed in feces
3. Rhabditiform larvae develop, emerge from egg, molt to become infective filariform larvae

Figure 9. Life cycle of the hookworm, *Necator americanus.*

Hookworms are among the more prevalent parasites of humans. Clinical manifestations of the disease are dependent on the number of worms present, the age and nutritional state of the infected individual, and the immune capabilities of the host. Penetration of the skin by the filariform larvae may cause an allergic or urticarial reaction known as ground itch. Secondary infection with pyogenic bacteria may occur. During larval migration through the lungs, patients may experience fever, headache, nausea, dyspnea, excessive coughing and pharyngeal soreness. Pneumonitis may occur in severe infections as a result of the parasites rupturing through the capillaries and invading the alveoli. During the intestinal phase of the infection, the symptoms may include such nonspecific gastrointestinal disorders as intermittent

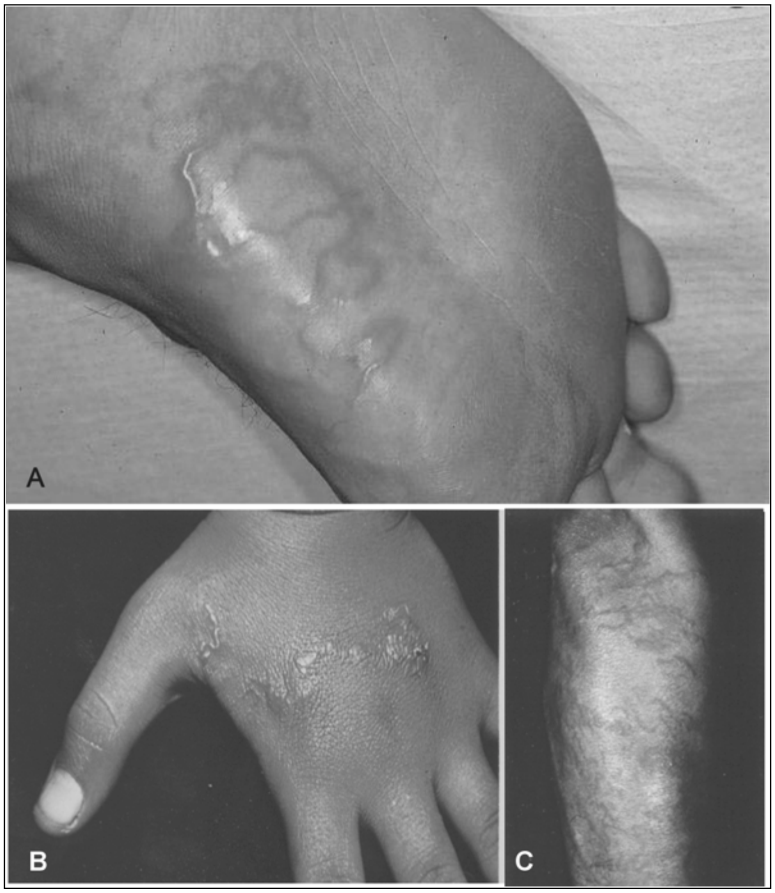

Figure 10. Cutaneous larva migrans infection ("creeping eruption") by cat or dog tapeworms. (A) Courtesy of Dr. Pietro Caramello, and the Carlo Denegri Foundation, Torino, Italy. (B) From Zaman, V. 1980. Atlas of Medical Parasitology. Singapore University Press. (C) From Yamaguchi, T. (Ed.) 1981. A Color Atlas of Clinical Parasitology. (Wolfe Medical Publ. Ltd.) Lea & Febiger, Pennsylvania.

abdominal pains, nausea, loss of appetite, vomiting, flatulence, diarrhea and/or constipation. Intestinal malabsorption is not common with hookworm infection. Major pathological manifestations of hookworm infection are iron-deficiency anemia, protein deficiency, and immunodeficiency. Since as much as 200 ml of blood per day may be lost by patients with heavy hookworm infections, a large parasite burden can gradually deplete a patient's serum proteins and iron reserves. Chronic malnutrition, especially in young patients, compounds the problems of hookworm infection, and may adversely affect physical and mental development.

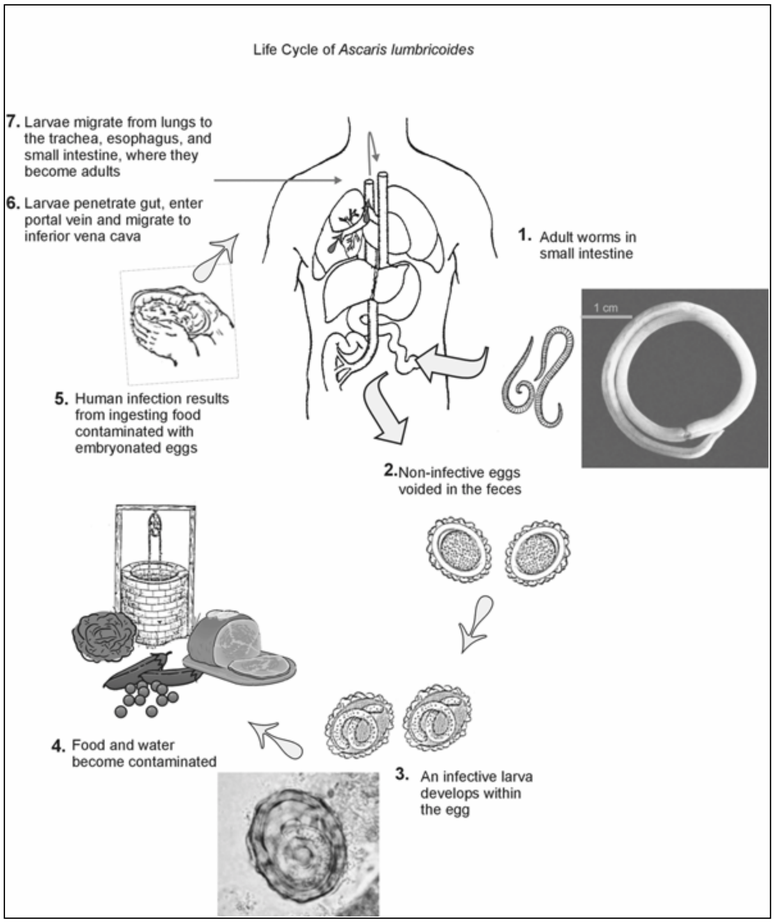
Life Cycle of *Ascaris lumbricoides*
7. Larvae migrate from lungs to the trachea, esophagus, and small intestine, where they become adults
6. Larvae penetrate gut, enter portal vein and migrate to inferior vena cava
1. Adult worms in small intestine
1 cm
5. Human infection results from ingesting food contaminated with embryonated eggs
2. Non-infective eggs voided in the feces
4. Food and water become contaminated
3. An infective larva develops within the egg

Figure 11. Life cycle of *Ascaris lumbricoides.*

Drugs for the treatment of hookworm infection include bephenium hydroxynaphthoate (Alcopara), pyrantel pamoate, and tetrachloroethylene. The only definitive diagnosis of hookworm infection is the early examination of feces and the identification of parasite eggs. Old stools or feces from constipated patients may contain, in addition to eggs, hatched rhabditiform larvae.

Cutaneous Larva Migrans

Two common hookworms of domestic dogs and cats, *Ancylostoma braziliense* and *A. caninum*, have been found in humans on several occasions. Two other hookworm species that have been found in humans are *A. ceylanicum*, which normally parasitizes carnivores in parts of Asia and the East Indies, and *A. malayanum,* which

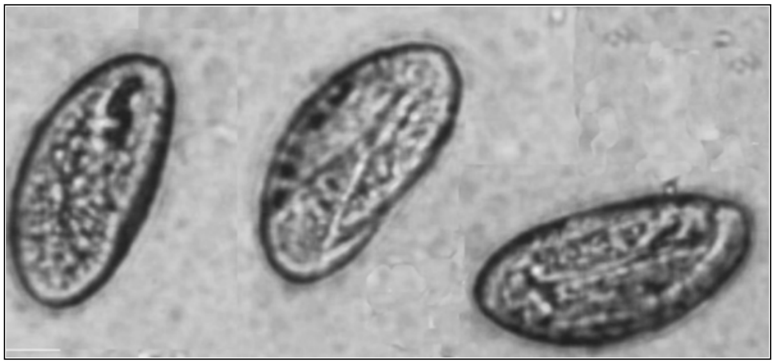

Figure 12. Eggs of *Enterobius vermicularis.* Sizes range from 20-30 mm by 50-60 mm. Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

is parasitic in bears in Malaysia. When nonhuman hookworm filariform larvae penetrate human skin, they may be trapped within the integument, unable to complete their migratory cycle to the lungs and intestinal mucosa. Instead, they migrate randomly through the subcutaneous tissue forming serpiginous burrows and causing intense itching, a condition referred to as cutaneous larva migrans or creeping eruption (Fig. 10). The wounds formed by scratching are frequently subjected to secondary bacterial invasion. The dog and cat hookworms, *A. braziliense* and *A. caninum,* are the principal causative agents of cutaneous larva migrans.

Ascaris lumbricoides

Ascaris lumbricoides is the largest intestinal nematode infecting humans. The parasite, commonly known as the “human roundworm”, occurs endemically in many parts of the world, but is most prevalent in tropical and subtropical areas. The intensity of infection may reach staggering levels, with several hundred worms present in a single host. The adults live free and unattached in the small intestine for about one year, during which time the daily egg production per female may average 20,000. Female worms generally range from 20 to 40 cm in length. The smaller males measure 12-30 cm in length, and are readily distinguished from females by a ventrally curled tail. Unembryonated eggs pass in the feces and become infective in about three or four weeks. At moderate temperatures, the eggs will remain viable in moist soil for up to 5 years. Ingestion of contaminated salads and other foods resulting from indiscriminate defecation or the use of feces as a fertilizer is the common means of infection. Transmission of infection by windborne dust carrying eggs is also possible. When fully embryonated eggs are ingested, they hatch in the small intestine. The larvae burrow into the mucosa wall, enter the circulation and undergo an extensive migration to the liver, heart, and lungs. Upon reaching the lungs, the larvae break out of the pulmonary capillaries, enter the alveoli, ascend the bronchi, are swallowed, and eventually reach the small intestine where they develop into adults

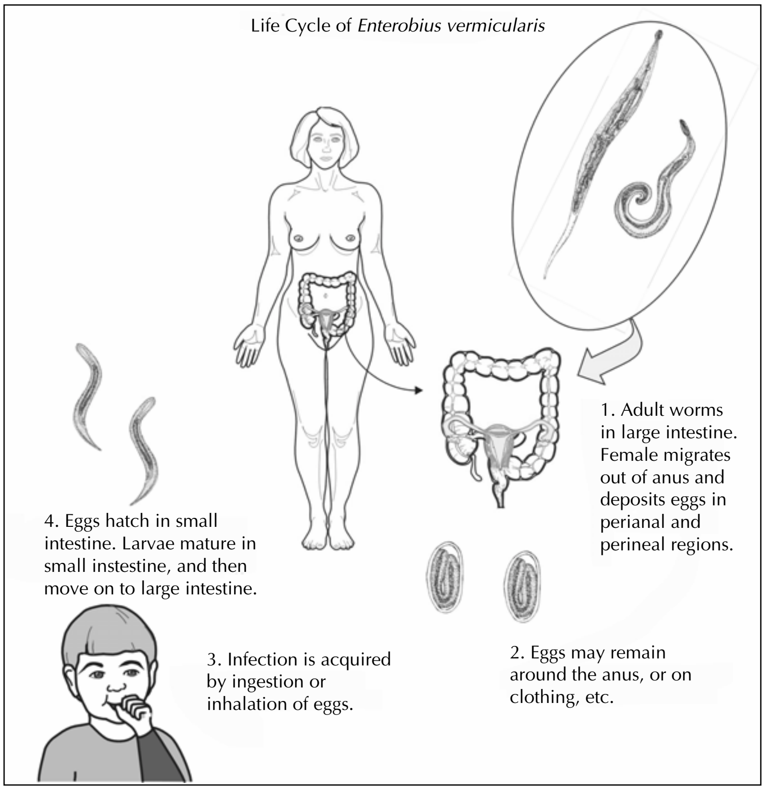
Life Cycle of *Enterobius vermicularis*
1. Adult worms in large intestine. Female migrates out of anus and deposits eggs in perianal and perineal regions.
2. Eggs may remain around the anus, or on clothing, etc.
3. Infection is acquired by ingestion or inhalation of eggs.
4. Eggs hatch in small intestine. Larvae mature in small instestine, and then move on to large intestine.

Figure 13. Life cycle of the pinworm, *Enterobius vermicularis.*

and mate. Adult worms begin laying eggs in the intestine about two or three weeks after infection. Transplacental migration of the parasite into a fetus is also (Fig. 11).

Symptoms of ascariasis are variable, often vague, mild or absent. Congestion, irregular respiration, spasms of coughing, edema, and a bloody sputum commonly characterize heavy pulmonary infections. Numerous parasites in the intestine may cause bowel obstruction, abdominal pain, nutritional and digestive disturbances, vomiting, restlessness, and disturbed sleep. Fatal peritonitis occurs rarely. The larvae may wander into such anomalous sites as the brain, spleen, liver, gallbladder, bile ducts, lymph nodes, peritoneal cavity, eustachian tubes, and the middle ear, causing inflammation, lesions, blockage of circulation or drainage, and death. The larvae may even exit from the body through the nasal passages. Live worms passed in the stool or vomitted are frequently the first manifestations of infection. Intestinal obstruction caused by the presence of numerous parasites may require surgical removal of the ascarids.

Pyrantel pamoate, piperazine hexahydrate or piperazine salts, levamisole, and thiabendazole are effective drugs for the treatment of ascariasis. Some side effects of

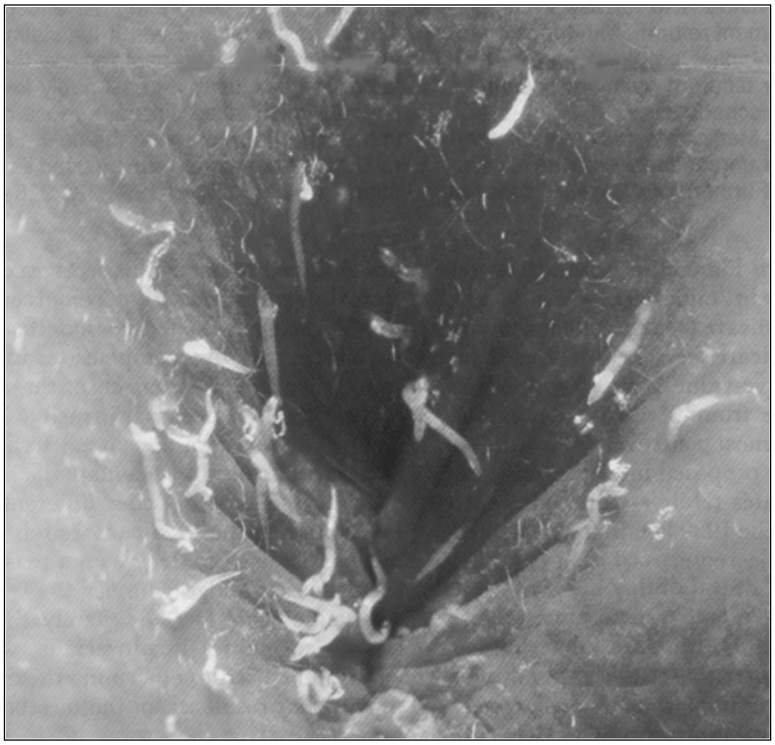

Figure 14. Pinworms in the perianal region. From Weber, M. Children's Hospital, Hannover Medical School, Hannover, Germany.

drug therapy of ascariasis include nausea, vomiting, diarrhea, and vertigo. When other intestinal worms are also present, *Ascaris* should be treated first to avoid the stimulation to migrate by the drugs.

Enterobius vermicularis

Enterobius vermicularis is the cosmopolitan pinworm, or seatworm, of humans. The parasite is the most common worm infection in the United States, with between 5 and 15% of the population harboring the parasite. Prevalence is highest in children of school age, and in individuals confined in institutions. The male worm is 2-5 mm in length, and females 8-13 mm in length. The posterior end of the male in strongly curved, while the female has a long, thin tapering tail. The adults characteristically congregate in the cecum and colon, but they do migrate within the lumen of the gastrointestinal tract from the stomach to the anus, feeding on the mucosa. Occasionally, adult worms in the stomach are regurgitated into the mouth. Gravid females migrate nocturnally out of the anus and deposit eggs on the perianal skin. The ovoid eggs are fully embryonated and infective within 6 hours after depo-

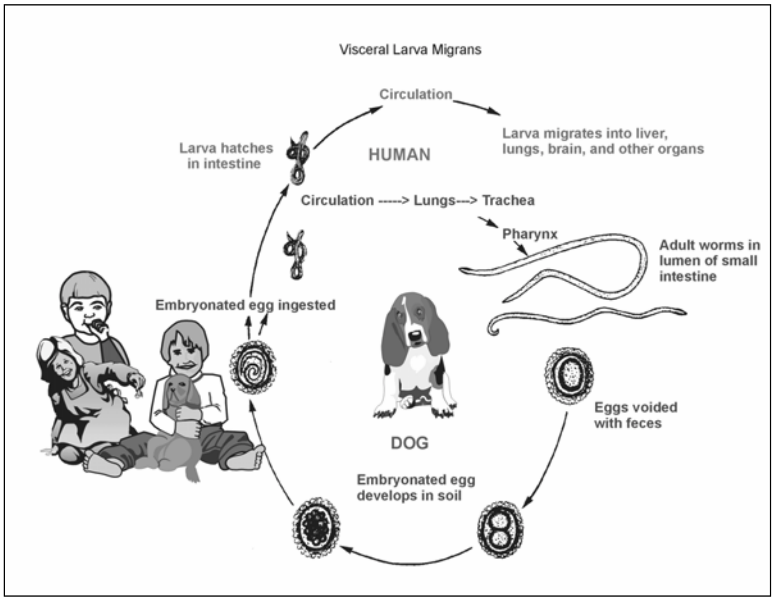
Visceral Larva Migrans
Circulation
Larva hatches in intestine
HUMAN
Larva migrates into liver, lungs, brain, and other organs
Circulation -----> Lungs---> Trachea
Pharynx
Adult worms in lumen of small intestine
Embryonated egg ingested
DOG
Eggs voided with feces
Embryonated egg develops in soil

Figure 15. Life cycle of the intestinal roundworm of dogs (*Toxocara canis*) or cats (*T. cati*), agents that can cause visceral larva migrans in human.

sition (Fig. 12). Each female may deposit about 10,000 eggs, which can survive for a few weeks under conditions of high humidity and moderate temperature (20-25°C).

Infection results from direct transfer of infective eggs by hand from the anus to the mouth of the same or different host, or indirectly through contaminated food or articles. Viable eggs have been collected from clothing, bedding, towels, furniture, and house dust (Fig. 13). Infection by inhalation may occur in heavily contaminated environments. Domestic animals are free of pinworms, and thus cannot transmit the disease. When swallowed, the eggs hatch in the duodenum, and the larvae molt twice and mature in the ileocecal region in 2-6 weeks. Retroinfection occurs when the eggs deposited in the perianal region hatch and the larvae wander back into the intestine through the anus (Fig. 14).

The majority of pinworm infections are asymptomatic. Children with heavy worm infections often experience mild gastrointestinal irritation, nausea, vomiting, irritability, and restless sleep. Intense scratching of the perianal region may produce lesions and secondary bacterial infections. Inflammation of the intestinal mucosa occurs at times. Occasionally, pinworms migrate into the female genital tract and peritoneal cavity causing irritation and infection. Encapsulated worms have been found in the peritoneum, liver, and lungs, but these complications occur only rarely. Eggs inhaled or introduced into the nose by contaminated fingers may hatch, with the larvae surviving for a time in the nasal passages producing some irritation.

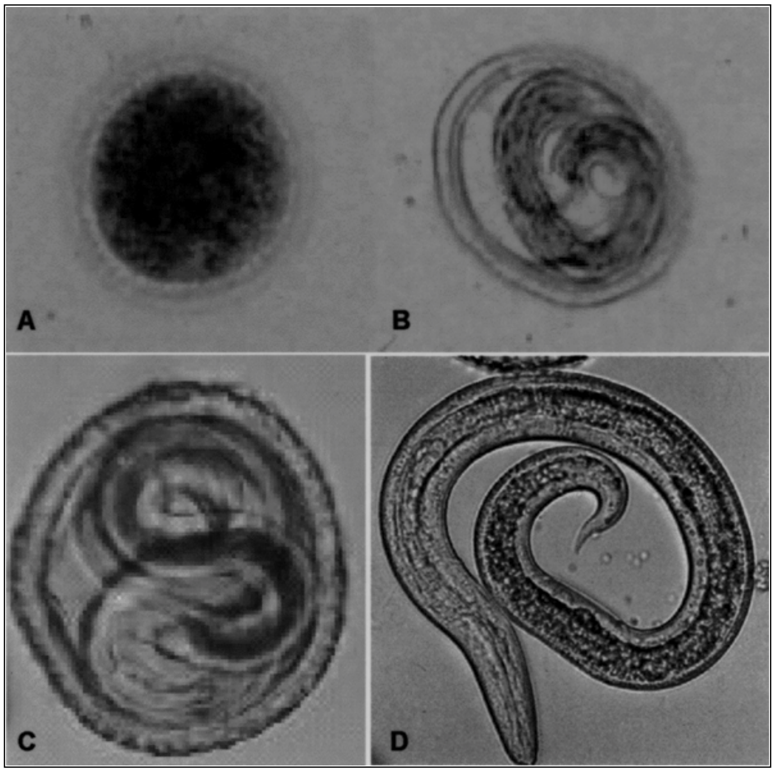

Figure 16. (A) Fertilized and unembryonated egg of *Toxocara canis* passed in dog feces. (B, C) Embryonated eggs of *T. canis*, each with a well developed larva. (D) Larva of *T. canis*. Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

Visceral Larva Migrans

Visceral larva migrans is a condition resulting from the accidental ingestion of the eggs of non-human nematoid larvae, the adults of which normally infect dogs and cats. The freed larvae are unable to complete their development in humans and wander erratically into various tissues and organs causing various pathologies. Frequently, the parasites become inactive and produce granulomas. Human visceral larva migrans frequently result from ingestion of the eggs of *Toxocara canis* or *T. cati*, two cosmopolitan intestinal parasites of domestic dogs and cats, respectively, and related carnivores. The disease, which is termed toxocariasis, is common in young children who always seem to have fingers going into their mouths (Fig. 15). Symptoms of infection typically include fever, eosinophilia, pulmonary infiltrates, and hepatomegaly. Granulomas have been found in the liver and retina. Death may result from extensive migrations of larvae, especially through the brain. In normal hosts, unembryonated eggs are passed in the feces and mature in the soil to the infective stage in about one week (Fig. 16). When ingested, the eggs hatch in the intestine, and the larvae burrow into the intestinal mucosa. The migratory route taken by the nematodes depends on the age and immunity of the host. In young dogs without previous exposure to the parasite, and thus with no established immunity against *T. canis*, the worms migrate through the portal system to the pulmonary circulation, up the trachea and back to the intestine. In mature dogs with some immunity from a previous infection, the larvae generally fail to complete the pulmonary migration and bypass the tracheal route leading to the small intestine. Instead, the larvae enter the systemic circulation and become distributed in various tissues and organs where they remain alive but inactive, encapsulated by a granulomatous host reaction. These dormant, encapsulated larvae appear to be activated by host hormones during pregnancy, and pass from the bitch into the fetal bloodstream to complete a pulmonary migration en route to the intestine. Parasite development occurs so rapidly that sexually mature worms may be found in young pups less than three weeks old. Except for the apparent absences of a lung migration and a prenatal infection route, the life cycle of *T. cati* is similar to that of *T. canis*. Rodents may serve as transport hosts when they ingest eggs of *T. canis* or *T. cati*. In rodents, the nematodes develop only to the second stage and become dormant in the tissues. It is estimated that 20% of the adult dogs and cats in the United States are infected with *Toxocara*. The infection rate may exceed 95% in pups and kittens.

Angiostrongyliasis

Two other nematode species causing human visceral larva migrans include *Angiostrongylus cantonensis* and *A. costaricensis*. The normal definitive hosts for these two species are rats. In humans, *A. cantonensis* causes eosinophilic meningoencephalitis. The disease is endemic in Hawaii, Tahiti, Japan, Taiwan, the Philippines, Australia, and Madagascar. Aquatic and terrestrial snails and slugs serve as intermediate hosts when they eat the feces of infected rats. Human infection is acquired by eating the infected intermediate or transport hosts, or raw vegetables or fruit containing viable parasite larvae expelled from these hosts. Symptoms of angiostrongyliasis may include severe headache and stiff neck (meningoencephalitis), marked eosinophilia in peripheral blood and cerebrospinal fluid, fever and coma. Although infection is usually benign, some fatalities have been reported. Temporary facial paralysis is occasionally observed. Human infections with *A. costaricensis* have been reported from Central America and Mexico. The terrestrial slug *Vaginulus plebeius* and various physid snails are intermediate hosts. Human infection results from eating contaminated vegetables. The worms characteristically cause abdominal pain and fever, thrombosis, inflammation and necrosis in the mesenteric arteries, and lesions in the walls of the small and large intestine. Occasionally, eosinophilia is found in infected individuals.

Anisakiasis

Anisakiasis is a disease of the gastrointestinal tract caused by ascarid larvae of the genus *Anisakis*. The adult nematodes are normally intestinal parasites of various marine fishes, birds, and mammals, such as whales, seals, dolphins, sea lions and porpoises. Many species of fish serve as transport hosts. The worms measure up to

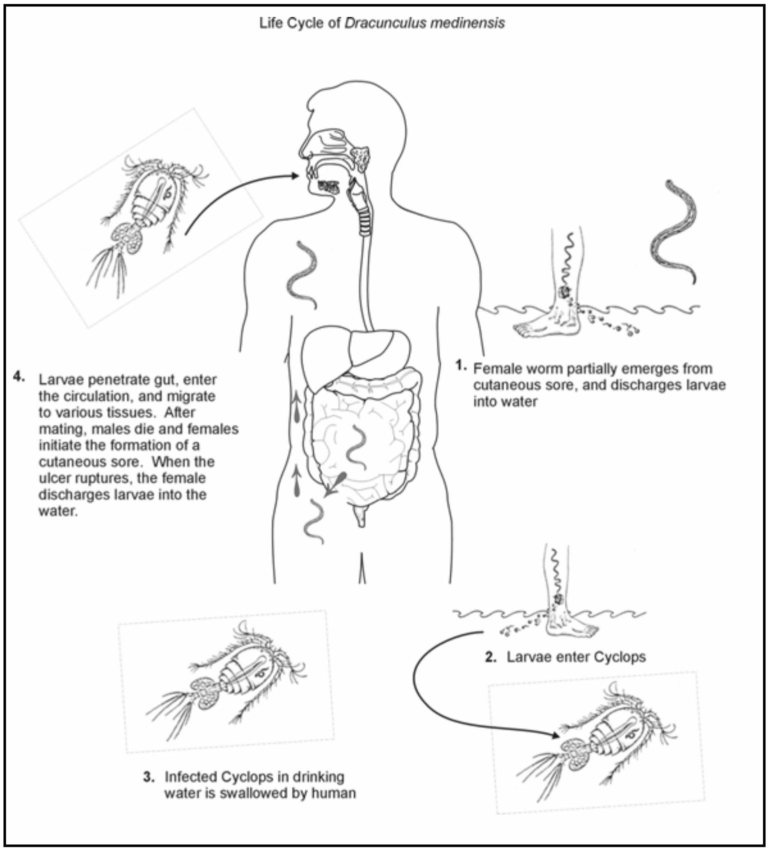
Life Cycle of *Dracunculus medinensis*
1. Female worm partially emerges from cutaneous sore, and discharges larvae into water
2. Larvae enter Cyclops
3. Infected Cyclops in drinking water is swallowed by human
4. Larvae penetrate gut, enter the circulation, and migrate to various tissues. After mating, males die and females initiate the formation of a cutaneous sore. When the ulcer ruptures, the female discharges larvae into the water.

Figure 17. Life cycle of the Guinea worm, *Dracunculus medinensis.*

3.5 cm in length. The larval stages are acquired by humans when raw or inadequately cooked infected hosts are eaten. The disease occurs predominately in Japan and adjacent areas, in various parts of Europe and Scandinavia, where raw, pickled, or smoked fish (herring, cod, salmon) are commonly eaten. The larvae do not mature in humans, but may cause acute nausea, vomiting, intestinal obstruction, abdominal symptoms mimicking those of appendicitis or ulcers, abscesses, gastric tumor-like growths, and peritonitis. Some infections have been reported to terminate fatally. There has been a marked increase in the incidence of anisakiasis in the United States due to the growing popularity of eating sushi.

Gongylonemiasis

Gongylonema pulchrum is a cosmopolitan parasite primarily of ruminants and swine, but the nematode has also been found in monkeys, hedgehogs, bears, and occasionally humans. Human infections have been reported in Europe, Russia, China,

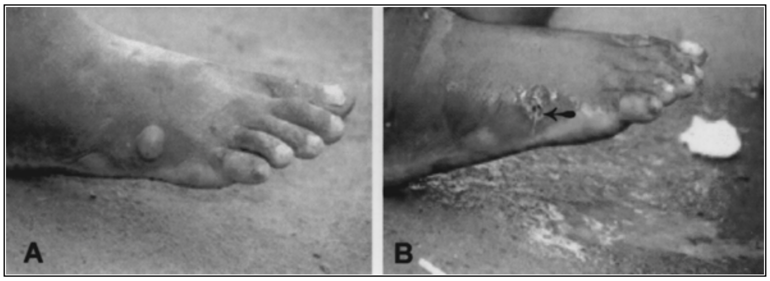

Figure 18. The female Guinea worm induces a painful blister (A) in the region where it will exit from the host. After the blister ruptures, the adult worm emerges (arrow in B). Frequently, the blister becomes secondarily infected. Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

New Zealand and the United States. The adult worms are found within the oral or esophageal epithelium. Males may reach a length of about 6 cm, and females about 15 cm in length. The eggs are fully embryonated when they pass from the females en route to the intestine to be voided with the feces. The life cycle involves an intermediate host, either dung beetles or cockroaches. When ingested by the intermediate host, the eggs hatch and the larvae penetrate the intestinal wall and become encapsulated in the hemocoel of the insect. When the intermediate host is ingested by the definitive host, the larvae migrate from the intestine up to the esophagus or oral cavity and burrow into the mucosa where they mature. In human infections the worms do not mature. The larvae may wander into the tongue, gums, hard and soft palate, lips and other areas of the oral cavity. Symptoms of gongylonemiasis include irritation, bleeding from the mouth, pharyngitis and stomatitis. The only effective treatment consists of surgical removal of the worms.

Gnathostomiasis

Gnathosoma spinigerum is the principal etiologic agent of human gnathostomiasis externa. The disease usually is acquired by humans through ingestion of raw or improperly processed or cooked freshwater fish infected with third-stage larvae of the nematode. The parasite is found in Japan, China, Thailand, Malaysia, Sumatra and the Philippines. Usually cats and dogs serve as definitive hosts. The adult worms are found embedded in tumor-like growths or nodules in the stomach wall. Male worms are 11-30 mm long, females 11-54 mm long. The eggs are unembryonated when passed in the feces. If deposited in water, the eggs mature in about one week and hatch. Two intermediate hosts are involved in the life cycle. The first intermediate host is a freshwater copepod that ingests the free-swimming parasite larva. The second intermediate host may be crustacea, freshwater fishes, amphibians, reptiles, birds, or mammals. When the infected copepod crustacean containing the second-stage larva is ingested by the vertebrate second intermediate host, the parasite penetrates the intestine of the new host, migrates to the connective tissues or muscles and molts to the infective third-stage larva. When the second intermediate host is

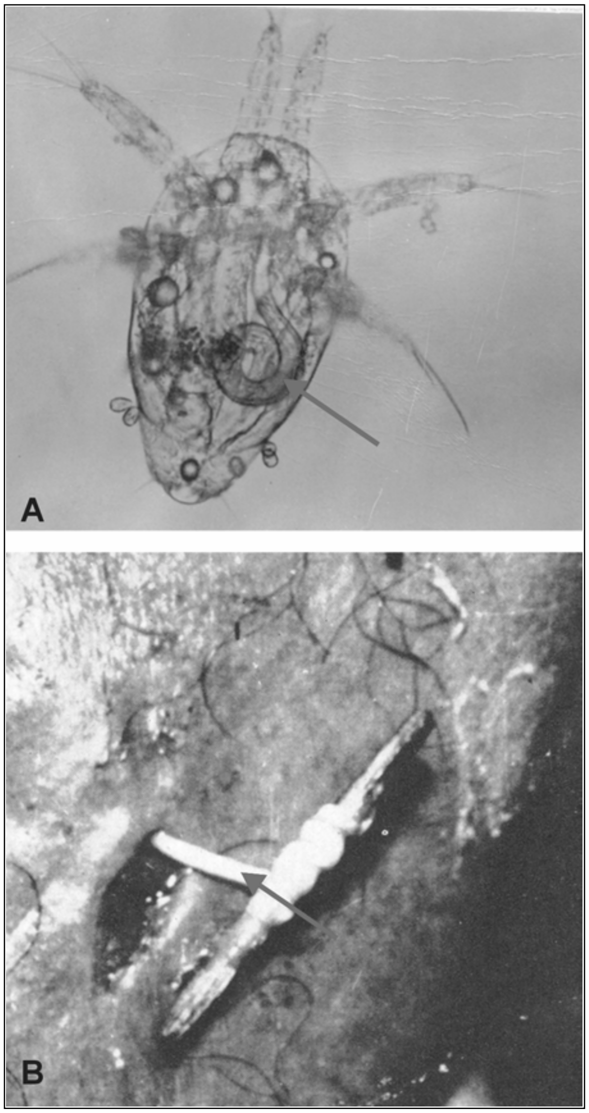

Figure 19. (A) Crustacean with larva of *Dracunculus* (arrow) in the body cavity. From Muller, R. 1971. *Dracunculus* and dracunculiasis. Advances in Parasitology (Ed. B. Danes). Volume 9, Academic Press, New York. (B) Guinea worm adult being removed from a leg by winding around a matchstick. From the Institute of Parasitology and Malariology, University of Teheran School of Medicine.

eaten by the definitive host, the parasite penetrates the wall of the stomach or intestine, invades the liver, muscles, and connective tissues. The worm later re-enters the gastric wall and becomes embedded in a nodule. Through a small opening in the nodule, the eggs pass into the lumen of the stomach and are then evacuated with the feces. When the second intermediate host is consumed by an unnatural host, the nematode penetrates the gut wall and migrates erratically in the host's tissues, usually in the subcutaneous layers, where they may produce endomatous areas and transient inflammatory lesions or abscesses which resemble those of cutaneous larva migrans caused by hookworms. Occasionally the worms will erupt from cutaneous abscesses spontaneously. The larvae may invade the brain producing cerebral lesions and death. Ocular damage and blindness also have been reported. Surgery is presently the only effective treatment for gnathostomiasis.

Dracunculus medinensis, or the guinea worm, is believed to be the "fiery serpent" reported in the Old Testament infecting Israelites during their travels into the Sinai Penninsula. The parasite is prevalent in India, Africa, and the Middle East, and it is estimated that about 48 million humans are presently infected. Guinea worm infections have also been reported from humans in the United States. The parasite in North America is probably a related species, *D. insignis*, which has been reported from dogs, raccoons, and other carnivores.

The adult female worms may reach 80 cm in length, while males are about 4 cm in length. The adults develop in the body cavity or visceral connective tissues (Fig. 17). Within a few months after copulation, the males die, become encapsulated and degenerate. The gravid female, which is ovoviviparous, migrates to the subcutaneous tissues, primarily of the extremities, where a dermal blister forms (Fig. 18). The blister, which develops where the head of the worm penetrates the dermis, is commonly found on the feet, ankles, calf, thigh, and knee joint, but it may also form on other areas of the body including the arms, trunk, buttocks, and scrotum. Eventually the blister ruptures, exposing the worm and a portion of the uterus, which has prolapsed through the body wall of the worm. When the blister is immersed in water, the worm and uterus protrude through the wound and numerous larvae are discharged. Emptied (spent) portions of the uterus disintegrate and new portions containing viable larvae move into the ulcer. After the uterus is emptied, the worm may be expelled, or it may withdraw and penetrate to deeper tissues before dying. Evacuation of the entire worm may take 3 weeks.

When ingested by a copepod the larvae penetrate the gut and enter the hemocoel of the crustacean where they molt twice and develop to the infective stage in about 2-3 weeks (Fig. 19). If the infected copepods are swallowed with contaminated drinking water, the larvae penetrate the intestine of the definitive host, and move via the lymphatics to the abdominal muscles and visceral connective tissues where they mature. Approximately one year after infection, the gravid female migrates to the extremities producing a dermal ulcer.

Although single infections of *Dracunculus* are most common, some individuals have been known to harbor as many as 50 worms. Migration of the gravid female to the extremities may initiate an allergic reaction, which causes erythema, urticarial rash, and intense pruritus. After several weeks, these symptoms usually subside or disappear. Female worms frequently fail to reach the surface and discharge larvae. Such non-emergent worms generally die and become absorbed or calcified, with no

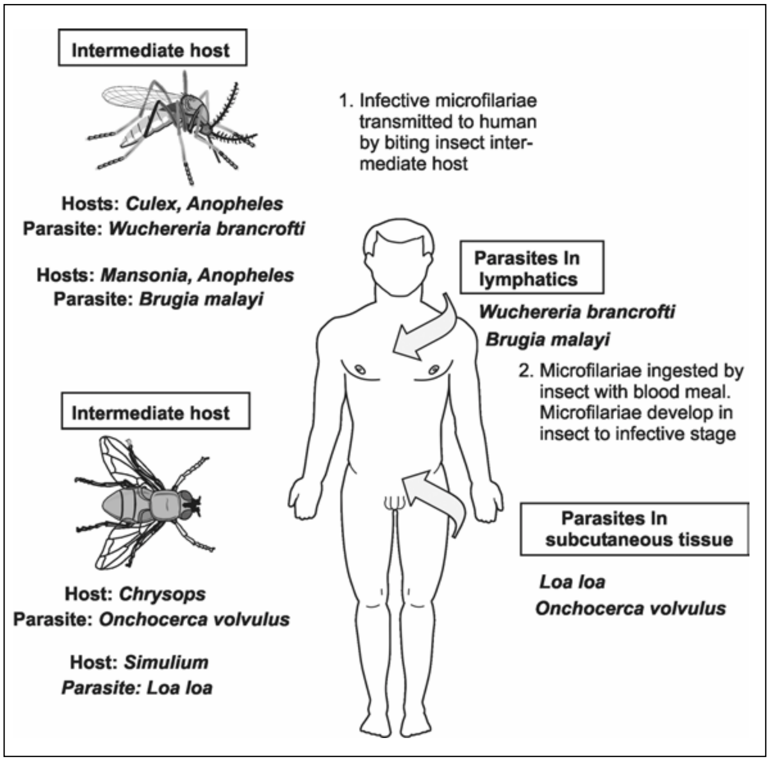
Intermediate host
1. Infective microfilariae transmitted to human by biting insect intermediate host
Hosts: Culex, Anopheles
Parasite: Wuchereria brancrofti
Hosts: Mansonia, Anopheles
Parasite: Brugia malayi
Parasites In lymphatics
Wuchereria brancrofti
Brugia malayi
2. Microfilariae ingested by insect with blood meal. Microfilariae develop in insect to infective stage
Intermediate host
Parasites In subcutaneous tissue
Loa loa
Onchocerca volvulus
Host: Chrysops
Parasite: Onchocerca volvulus
Host: Simulium
Parasite: Loa loa

Figure 20. Generalized life cycles of filariids.

apparent adverse affect on the host. However, non-emergent worms may form pus-filled abscesses, which can lead to arthritis if located in or near a joint. The abscesses can be very large, containing numerous larvae and up to 0.5 liter of fluid. Rare complications result from adult worms in the urogenital system, thoracic cavity (causing pericardiasis), and central nervous system (causing paraplegia). Secondary bacterial infections may also cause serious complications.

The most common technique for removing a guinea worm is to slowly wind the parasite on a stick. Each day only a few centimeters of the worm are wound about the stick, with care exercised as not to rupture the worm and cause secondary infections (Fig. 19). Surgical removal of the worm after local anaesthetic is widely practiced in India and Pakistan.

Filarial Worms

Filariasis refers to infection with one of several species of filarial worm. Human filariids are slender, blood- and tissue-dwelling parasites that cause some of the most disfiguring and debilitating diseases known. Female worms do not lay eggs but instead, give birth to larvae termed microfilariae. The ingestion of microfilariae by

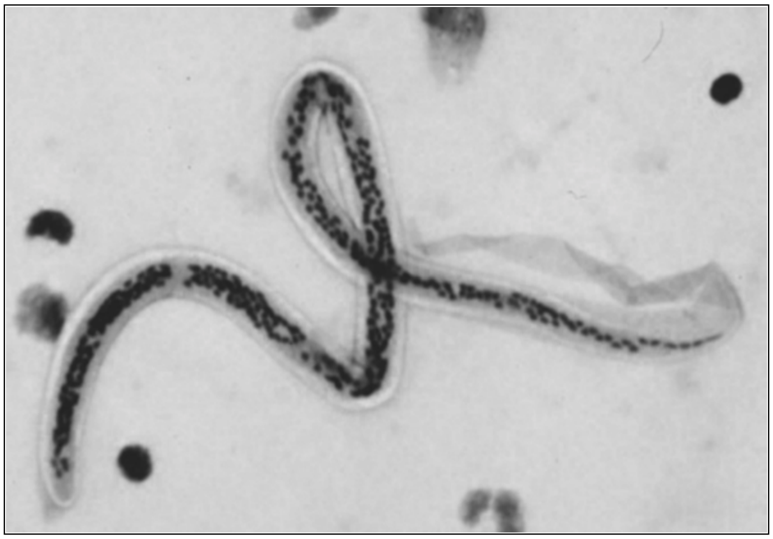

Figure 21. Sheathed microfilaria of *Wuchereria brancrofti* in a blood smear. Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

blood sucking insect vectors, which also serve as intermediate hosts, provides for the transmission of the filariids (Fig. 20).

Some important filarial worms of humans are: *Dracunculus medinensis, Wuchereria brancrofti, Brugia malayi, Loa loa, Onchocerca volvulus* and *Dirofilaria.*

Wuchereria brancrofti

Wuchereria brancrofti is the etiologic agent of Brancroftian filariasis or elephantiasis. The disease is the most widespread of the filariases of humans, occurring in Central and South America, Africa, Asia, West Indies, part of Europe and the Pacific islands. At one time the parasite was reported from humans living in the region of Charleston, South Carolina. Humans are the only known host for *W. brancrofti*. Female worms are 5 to 10 cm long, while males measure 2 to 4 cm in length. The adult worms live in tightly coiled masses in the lymphatic system where thousands of microfilariae are released by the females. The microfilariae eventually leave the lymphatic system and move into the blood via the thoracic duct. The microfilariae may be nocturnally periodic, being present in large numbers in the peripheral blood at night, and in low numbers during the day, or occur in greater concentration in the peripheral circulation in the daytime (diurnal periodicity). Variations in periodicity occur in different geographic strains of the parasite. Changes in host body temperature, activity, and in oxygen tension of the blood may be responsible for the filarial periodicity. Numerous species of mosquito belonging to the genera *Aedes, Culex, Monsonia* and *Anopheles* serve as intermediate hosts. The discovery by Manson in 1878 of the transmission of *W. brancrofti* by *Culex fatigens* in

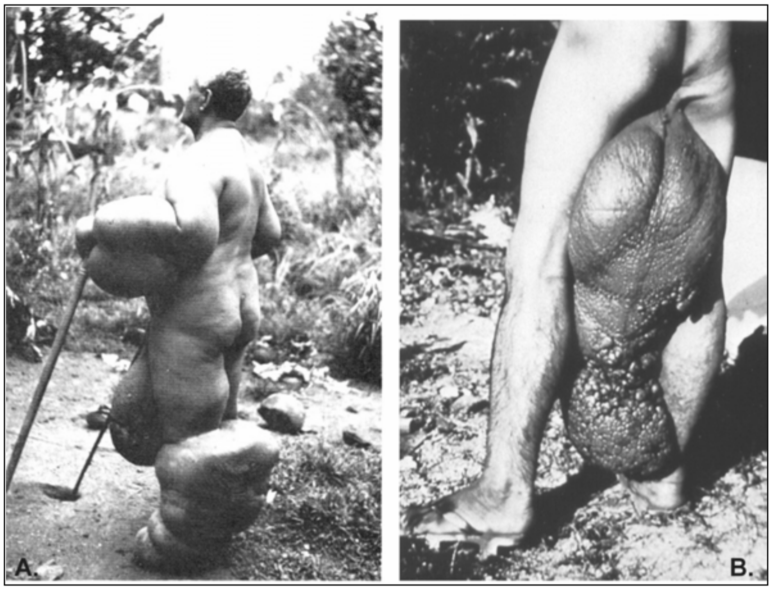

Figure 22. Extreme cases of elephantiasis of limbs and scrotum. (A) Photograph by Dr. John F. Kessel. From Markel, E. K., Vogue, M. and John, D. T. 1986. Medical Parasitology. W. B. Saunders Co., Philadelphia. (B) From Yamaguchi, T. (Ed.) 1981. A Color Atlas of Clinical Parasitology. (Wolfe Medical Publ. Ltd.) Lea & Febiger, Pennsylvania.

China was the first demonstration of an arthropod as a biological vector of a parasite (Fig. 20).

When the microfilariae are ingested by the mosquito along with a blood meal, the parasites penetrate the insect's gut, migrate to the thoracic muscles, and mature to the infective filariform stage. Infective larvae then migrate throughout the hemocoel and enter the proboscis of the mosquito, from which they escape when the insect again feeds. The larvae enter the skin of a new host through the puncture made by the feeding insect. The infective larvae move first to the peripheral lymphatics, then to the lymph nodes and larger lymphatics where they mature. Maturation of the worm may take up to one year from invasion of the skin until microfilariae appear in the blood (Fig. 21).

Clinical manifestations of Brancroft's filariasis result from the presence of adult worms in the lymphatic vessels. Light infections may go undetected, with the only physical signs of infection being slightly tender and swollen lymph nodes. Parturient females and microfilariae may cause intense lymphangitis accompanied by lymphadenitis, fever, chills, and toxemia. Lymphangitis commonly affects the arms and legs, and the epitrochlear and femoral lymph nodes are usually involved. Acute tissue reactions, consisting of hyperplasia of the vessel wall surrounding the parasite, and the accumulation of histocytes, eosinophils, and lymphocytes in the lumen of

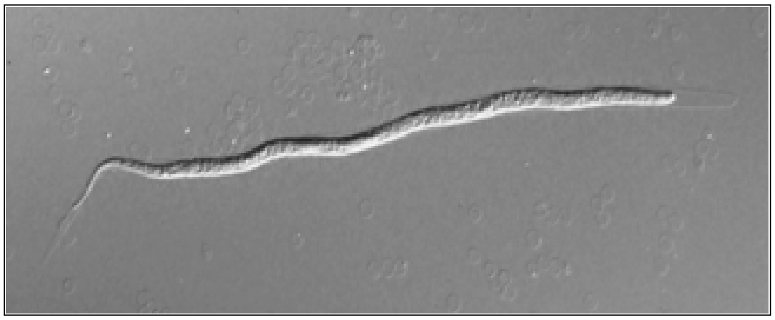

Figure 23. Microfilaria of *Brugia malayi* with a sheath extending beyond the anterior and posterior ends. Courtesy of the Centers for Disease Control and Prevention, Division of Parasitic Diseases, Atlanta, Georgia.

the vessels gradually obstruct the flow of lymph. Affected lymphatic vessels are distended, tender, and painful, and the overlying skin erythematous and hot. As affected areas become progressively infiltrated with fibrous connective tissue, a chronic lymphedema and thickening of the skin develops. Chyluria, or lymph in the urine, is a common symptom. In males, there may be extensive inflammatory involvement of the scrotum, with orchitis, epididymitis, and hydrocele (forcing of lymph into the tunica vaginalis of the testes or spermatic cord). In females, the vulva and breasts are sometimes affected. The development of elephantoid limbs and organs with dermal hypertrophy and verrucous growths represent uncommon and extreme manifestations of the disease, occurring in some individuals with repeated infections (Fig. 22). The swollen organs are composed primarily of fibrous connective tissue and fat. Many infected individuals never develop symptoms more severe than microfilaremia, localized edema, recurrent attacks of lymphangitis and fever, lymphadenopathy with splenomegaly, and transient pulmonary infiltrates and hypereosinophilia. Worms may be present for years in dilated or varicose lymphatic vessels before they die and are absorbed or become calcified. Occasionally, advanced cases of elephantiasis require surgery to correct lymphatic obstructions. Elephantoid tissues may also be treated by using elastic or pressure bandages, which force the lymph from the swollen areas. Surgical treatment of scrotal elephantiasis is commonly effective, whereas surgical treatment of elephantoid limbs is typically unsuccessful.

Brugia malayi

Brugia malayi is the etiologic agent of malayan filariasis. The parasite differs only slightly in morphology from *W. brancrofti,* and the life cycles of both species are nearly identical, with mosquitoes of the genera *Aedes, Mansonia* and *Culex* serving as intermediate hosts and vectors (Fig. 23). Females are 8 to 10 cm long. Males are about one-third the size of females. Malaysia is one of the major endemic areas, but the parasite also occurs in China, Korea, Japan, southeast Asia, Sri Lanka, the East Indies and the Philippines. Unlike *Wuchereria*, for which humans comprise the only known host, *B. malayi* also occurs in monkeys, dogs, and cats. The clinical and

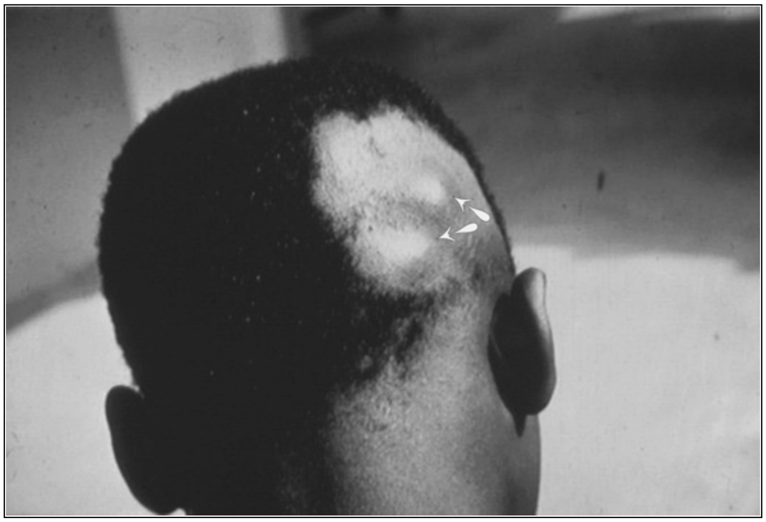

Figure 24. Cranial onchocercariasis. Adult *Onchocerca volvulus* larvae develop to adult worms in subcutaneous nodules (arrows). Courtesy of Professor Wallace Peters and the Carlo Denegri Foundation, Torino, Italy.

pathological features and treatment of malayan filariasis are similar to those described for the brancroftian variety. However, elephantiasis caused by *B. malayi* is typically not as severe as that caused by *Wuchereria*, and there is rarely any involvement of the genitalia.

Loa loa

Loiasis is a disease resulting from infection with *Loa loa*, the African eyeworm. The parasite, which is found only in Africa, is transmitted to humans by one of several species of mango or deer flies belonging to the genus *Chrysops*. Monkeys and other primates appear to be the only other definitive hosts. Female worms are 5 to 7 cm long, and males are approximately one-half the size of females. The sheathed microfilariae of *L. loa* undergo a developmental cycle in mango flies similar to that of *Wuchereria* in mosquitoes, with infective larvae passing from the fly's proboscis and entering the skin of the definitive host when the insect bites and draws blood. Following infection, the larvae develop into adult worms which migrate through the subcutaneous and deeper connective tissues of the definitive host. The microfilariae are periodic, appearing in large numbers in the peripheral circulation during the day, and in the lungs at night. Infections are generally asymptomatic, although occasionally migrating worms provoke localized inflammatory responses. Localized, transient subcutaneous nodules termed "calabar swellings" develop in areas where the worms remain stationary for brief periods of time. The swellings are believed to be allergic manifestations of the infection, and they generally disappear within a few days when the worms move to other areas. Clinical manifestations

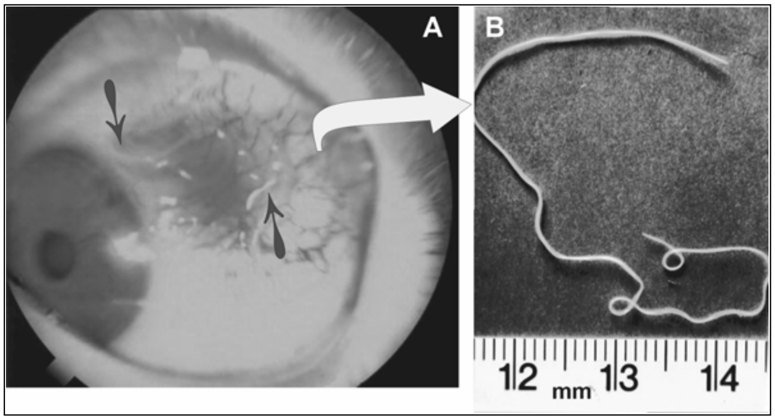

Figure 25. Ocular dirofilariasis caused by *Dirofilaria repens*. (A) A single worm coiled under the bulbar conjunctiva (small arrows) of the right eye. (B) Excised worm. Courtesy of Drs. Alberto Biglino and Angelo Casabianca, Asti General Hospital, Asti, Italy, and the Carlo Denegri Foundation, Torino, Italy.

result when the adults migrate into the facial area, conjunctival tissue, and cornea. When the worms are located in suitable superficial locations, they can be removed surgically, but this procedure is generally unwarranted because of effective chemotherapy.

Onchocerca volvulus

Onchocerca volvulus is the etiologic agent of onchocercariasis or river blindness, a disease found only in humans in central Africa, and northern South America. Female worms are 33-50 cm long by 0.2-0.4 mm wide. Males are 19-42 cm long by 0.1-0.2 mm wide. Adult worms are located in the dermis and subcutaneous tissues, where, as a result of host immune responses, they become enclosed in fibrous capsules that are usually visible as small nodules under the skin. The nodules (onchocercomas) develop on any part of the body, but they are frequently found on the trunk, hips, elbows, and scalp. Unsheathed microfilariae escape from the nodules and remain in the connective tissue of the skin. Rarely do the microfilariae enter the general circulation. Any one of several species of black fly belonging to the genus *Simulium* serves as the intermediate host and vector. The microfilariae are ingested by the biting flies when the insects feed on tissue fluids of the definitive host. The parasites penetrate the intestinal tract of the fly and move directly to the thoracic muscles, where they molt to the infective filariform stage. The parasites then pass to the mouthparts and enter a new host when the simuliid feeds again. Following introduction into the new host, the parasites migrate through the subcutaneous tissues and eventually become encapsulated.

Superficial nodules may be disfiguring, but generally are not painful (Fig. 24). Microfilariae migrating from the nodules throughout the dermis and connective tissues may cause sensitization reactions, severe pruritic dermatitis, ocular lesions, and blindness. In certain endemic regions of Africa, 30% of the population has impaired vision due to onchocercariasis. Thick, wrinkled, and hyperpigmented skin are also common symptoms. In extreme cases, infected skin around the pelvic region loses its elasticity, and a condition known as "hanging groin" or adenolymphocele occurs. The fold of skin, often containing lymph glands, may hang down to the knees. Infection with *O. volvulus* may also cause inguinal and femoral hernias. The parasite is also believed to cause pituitary damage and dwarfism in individuals living in certain parts of Uganda.

Dirofilaria immitis

Dirofilaria immitis, the heartworm of dogs, cats, foxes, and wolves, is widely distributed in the United States, and throughout most of the world. Adult worms are commonly found in the right ventricle and pulmonary arteries, but they also invade the subcutaneous tissues and eyes of the host (Fig. 25). Females are 2.5-3.1 cm long, and males measure 1.2-2.5 cm long. The microfilariae are nocturnally periodic. Mosquitoes belonging to the genera *Anopheles, Aedes* and *Culex,* and fleas, serve as intermediate hosts and vectors. Following ingestion by the insect, the larvae enter the cells of the malpighian tubules and develop to the infective stage. The larvae later migrate to the fat body and eventually reach the mouthparts, from which they escape and infect a new host when the insect again feeds. Infected animals may suffer from chronic endocarditis, pulmonary complications, enlargement of the liver, and inflammation of the kidney.

Dirofilaria immitis has been reported from humans several times, as has *D. tenuis* and *D. repens*, two filariids commonly found in the subcutaneous tissues of raccoons and dogs, respectively. The parasites have been found alive in the heart and lungs, and in solitary nodules or abscesses in various subcutaneous locations including the eyes, lips, arms, and breasts. The worms may reach sexual maturity in humans, but rarely do they produce microfilariae.

CHAPTER 8

Arthropods

Arthropods comprise approximately 78% of all known species of animals. They possess jointed appendages and a segmented body that is covered with a chitinized exoskeleton. There are two major groups of arthropods, chelicerates and mandibulates. The chelicerate arthropods include arachnids (spiders, scorpions, mites, and ticks), xiphosurids (king crabs), and pycnogonids (sea spiders). These organisms differ markedly from one another, but generally they possess chelicerae as the first pair of appendages, and they lack antennae. Mandibulate arthropods, which include insects, crustaceans, millipedes, and centipedes, typically possess antennae and mandibles, but lack chelicerae. Some of the characteristics distinguishing mandibulates and chelicerates are given in Table 1.

Arthropods are of medical and veterinary importance not only because they serve as intermediate hosts and vectors of disease, but also because they themselves are causal agents of disease. Some of the more important arthropod vectors and the diseases they transmit are given in Table 2. The remaining portions of this chapter consider arthropods as pathogenic agents.

Types of Injury Caused by Arthropods

The types of injury caused to humans by the adults and/or young stages of arthropods may be classified as follows:

1. Cutaneous or mucocutaneous injury. The development of blisters or lesions on the skin or mucous membranes may result from the discharge of body fluids (vesication), the contact with hairs, spines, or other cuticular projections (urtication), or the introduction of various secretions from biting or stinging arthropods (venenation).

2. Systemic injury. Various pathologies result from the introduction of secretions into the body by venenating arthropods, including parasitemia, fever, hemorrhage, tachycardia, and respiratory difficulties.

3. Tissue injury. Mechanical damage to visceral organs and tissues frequently results from the penetration of the body by the adult or young stages of arthropods. Blood-sucking arthropods have been classified into two groups according to the method used to obtain blood. Solenophagic feeders insert their mouthparts directly into the lumen of a blood capillary and thus feed almost exclusively on blood. Adult mosquitoes are solenophagic feeders. Telmophagic feeders insert their mouthparts indiscriminately into the skin. The enzymatic secretions initially cause lysis of the tissues, and the animal feeds first on tissue fluids until the walls of the dermal blood vessels are broken down and a blood pool is produced from which they feed. Ticks are good examples of telmophagic feeders.

Parasites of Medical Importance, by Anthony J. Nappi and Emily Vass.
©2002 Landes Bioscience.

Table 1. Some general characteristics of the major groups of arthropods

	Mandibulates				Chelicerates
	Crustaceans	Insects	Centipedes	Millipedes	Arachnids
Body divisions	Usually cephalothorax and abdomen	Head, thorax, abdomen	Head and trunk	Head, thorax, abdomen	Cephalothorax and abdomen
Mouth parts	Mandibles, maxillae (2 pairs)	Mandibles, maxillae (1 pair)	Mandibles, maxillae (2 pairs)	Mandibles, maxillae (1 pair)	Chelicerae, pedipalpi
Legs	1 pair per segment, or less	3 pairs on thorax	1 pair per segment	2 (or 1) pairs per segment	4 pairs on cephalothorax
Antennae	2 pairs	1 pair	1 pair	1 pair	None
Respiration	Gills or body surface	Tracheae	Tracheae	Tracheae	Book lungs, or tracheae
Genital openings	2, hind part of thorax	1, end of abdomen	1, end of abdomen	1, third segment near head	1, second segment of abdomen
Development	Usually with larval stages	Usually with larval stages	Direct	Direct	Direct, except mites and ticks

Arthropods as Vectors of Disease

Arthropods are of medical and veterinary importance not only because they serve as intermediate hosts and vectors of disease, but also because they themselves are causal agents of disease. A list of some of the more important arthropods and the diseases they transmit is given in Table 1. The remaining portions of the chapter consider arthropods as pathogenic agents.

Chelicerates (Arachnids)

Scorpions

Scorpions are the largest of the arachnids, measuring up to 20 cm in length. They are cosmopolitan in distribution through tropical and subtropical regions. Scorpions are largely nocturnal carnivores feeding on insects, centipedes, spiders, and small mammals. The abdomen terminates in a sting that can penetrate the integument of the prey, into which venom is injected (Fig. 1). The venom of most scorpions is harmless to humans. The wound produced by the sting is painful, and is usually accompanied by local swelling and discoloration. A radiating, burning sensation is generally experienced at the time of venenation. A few species (*Centruroides, Androctonus, Parabuthus, Buthacus, Buthotus, Scorpio* and *Tamulus*) are capable of inflicting systemic reactions, which may cause death. Young children are most liable to venenation and experience a high mortality. The venom affects the nervous system and also causes pulmonary disorders. Symptoms may include fever, vomiting, frothy salivation, extreme thirst, muscle spasms, convulsions, impaired respiration, tachycardia, paralysis and death. Early therapeutic measures include placing an ice pack over the wound, the application of a tourniquet proximal to the site of venenation, and the removal, by suction, of as much venom as possible. If available, anti-venom should be administered in severe cases involving systemic infections.

Spiders

Spiders rank among the most feared of animals, an emotion that is unjustified, since the vast majority of species are completely harmless to humans. The body of a spider is comprised of two distinct regions, an anterior cephalothorax and a posterior abdomen (Fig. 2). The cephalothorax bears four pairs of walking legs, a pair of pedipalps, and a pair of chelicerae. One to four pairs of silk-spinning organs or spinnerets occur near the terminal portion of the abdomen. Most spiders have eight eyes arranged in two rows of four. The chelicerae are comprised of two segments, a basal portion and a distal fang, which is used to puncture prey and introduce venom. Spiders feed on insects, other arthropods, and in the case of some tropical spiders, small vertebrates. Following the initial puncturing by the chelicerae, the prey is subjected to extra-oral digestion by enzymes secreted from the mid-gut, and from glands found in the pedipalpal coxae. The liquid tissues are then sucked from the prey by means of the pharynx.

Few spiders are medically important. Many attacks on humans occur only accidentally when an exposed part of the body comes in contact with the spider. Depending upon the species of spider, two types of pathophysiological disorders may be produced, necrotizing lesions or systemic injury. The most important spiders causing necrotic arachnidism belong to the genus *Loxosceles*, and are commonly known as “brown” or “violin” spiders, since some species have a fiddle-shaped marking on the dorsal surface of the abdomen. *Loxosceles laeta* is widely distributed in southern South America. *Loxosceles reclusus*, the brown recluse spider, is the chief agent of loxoscelism in the mid-western United States. The bite of the spider is painful, and swelling and tissue necrosis usually develop. In mild loxoscelism, the cutaneous lesion heals leaving a disfiguring scar several centimeters in diameter. Manifestations of systemic injury may include fever, hemorrhage, erosion of mucous membranes, hematuria, and cardiac failure. Corticosteroids may alleviate the pain and prevent the development of necrosis and systemic damage.

Many of the spiders producing systemic injury belong to the genus *Latrodectus*. The “black widow”, *L. mactans*, is distributed in Canada, the United States, and parts of South America. Other species of *Latrodectus* occur in these areas, as well as in the Mediterranean area, Africa, New Zealand, parts of the Caribbean, Asia and Europe (Fig. 3). The venom acts as a peripheral neurotoxin. Immediate sharp pain occurs at the site of the bite, which later becomes reddened. Symptoms include pain in the abdomen and chest, weakness, motor disturbances, spastic contractions, delirium, and convulsions. Death from latrodectism is rare, but may occur in young children or the elderly, and frequently results from respiratory and/or circulatory failure. Analgesics and sedation generally suffice as treatment for most cases. In se-

Table 2 Some human diseases transmitted by arthropods

Disease	Pathogen	Vector	Vertebrate Reservoir
	Bacterial, Rickettsial, and Related Diseases		
Bubonic plague	*Pasteurella pestis*	Fleas, especially the rat flea	Rodents
Tularemia	*P. tularensis*	Deer flies, ticks, fleas, body lice	Rabbits
Anthrax	*Bacillus anthracis*	Horse flies	Mammals
Epidemic typhus	*Rickettsia prowazekii*	Body louse, rat flea, rar mite	Rodents
Endemic or murine typhus	*R. prowazekii,* *R. mooseri*	Fleas, lice, mites, rodent ticks	Rodents
Scrub typhus or tsutsugamushi	*R. tsutsugamushi*	Harvest mites (chiggers)	Rodents
Spotted fever	*R. rickettsii*	Ticks	Rodents
Q-fever	*Rickettsia* spp.	Ticks	Cattle, sheep, goats
Bartonellosis or oroyo fever	*Bartonella bacilliformis*	Sand flies	?
	Protozoan Diseases		
African Trypanosomiasis	*Trypanosoma gambiense, T. rhodesiense*	Tsetse flies	Herbivores
South American Trypanosomiasis	*Trypanosoma cruzi*	Assassin bugs	Rodents, carnivores
Epizootic (Leishmaniasis)	*Leishmania braziliensis*	Sand flies	Dogs, cats
Oriental sore (Leishmaniasis)	*L. tropica*	Sand flies	Dogs, cats
Kala-azar (Leishmaniasis)	*L. donovani*	Sand flies	Dogs, cats
Malaria	*Plasmodium vivax, P. malariae, P. falciparum, P. ovale*	Mosquitoes	None

continued on next page

vere cases, the intravenous injection of specific antivenom, if available, should be administered.

Mites and Ticks (Acarina)

The order acarina is comprised of mites and ticks. Ticks differ from mites in their comparatively large size and their exclusive habit of feeding on vertebrate blood during each of their developmental stages. The acarine body is comprised of a head region (capitulum or gnathosoma), which bears the mouth parts, and a posterior region (idiosoma), which is subdivided into a leg-bearing part and an abdominal part. The first developmental stage of is a six-legged (hexapod) larva that emerges from the egg. With one or more intervening molts, the larva becomes an eight-legged nymph, which after one or more molts, becomes an adult.

Table 2. Cont.

Disease	Pathogen	Vector	Vertebrate Reservoir
		Viral Diseases	
Yellow fever		Mosquitoes (*Aedes aegypti*)	Monkeys
Dengue fever		Mosquitoes (*Aedes aegypti*)	?
Encephalitis		Mosquitoes (spp. *Culex, Aedes, Mansonnia*)	Horses, birds
Pappataci fever		Sand fly	Humans
Colorado tick fever		Ticks	Rodents
		Spirochaetal Diseases	
Relapsing fever	*Borrelia recurrentis*	Ticks	Rodents
Relapsing fever	*B. duttonii*	Body louse	None
		Fluke Diseases	
Lung fluke	*Paragonimus westermani*	Fresh-water crabs, crayfish	Carnivores
		Tapeworm Diseases	
Dipylidiasis	*Dipylidium caninum*	Dog flea	Dogs, cats
Hymenolepiasis	*Hymenolepis diminuta*	Rat flea, grain beetles, earwigs	Rats, mice
Diphyllobothriasis	*Diphyllobothrium latum*	Water fleas	Dogs, bears
		Roundworm Diseases	
Brancroftian filariasis	*Wuchereria brancrofti*	Mosquitoes (*Aedes, Culex, Anopheles, Mansonia*)	?
Malayan filariasis	*Brugia malayi*	Black flies	Felines, monkeys
Loaiasis	*Loa loa*	Deer flies	Monkeys
Onchocerciasis	*Onchocerca volvulus*	Black flies	None
Guinea worm	*Dracunculus medinensis*	Water flea	Carnivores

Mites

Mites usually measure less than 1 mm in length. The chelicerae and pedipalps are located on the capitulum. A median hypostome is situated below the mouth. In scavenging and predatory mites the chelicerae are chelate and used for prehension. In parasitic forms, the chelicerae are stylet-like. The chelicerae open an incision, and the hypostome penetrates into the tissues. In some mites, external digestion results from secretions from the paired salivary glands. The liquid food is drawn into the

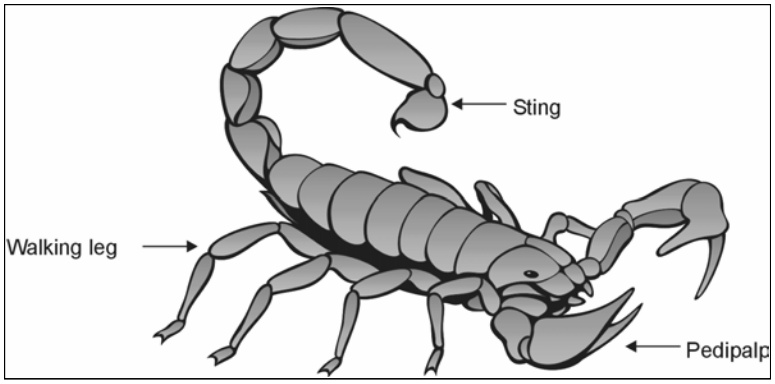

Figure 1. Scorpion.

pre-oral food canal by the suctorial action of the pharynx and then passes to the esophagous and gut. A few mites, such as mange and follicular mites, burrow into the tissues of the host.

Red bugs or "chiggers" are larval mites of the family Trombiculidae. Only the larval stage is parasitic, sucking tissue juices from various vertebrates, including humans. Upon hatching from eggs laid on the ground or vegetation, the hexapod larvae attach to the skin of vertebrates, inject a salivary secretion that lyse the integument, and suck tissue fluids until engorged. After feeding, which may last from a few days to several weeks, the larvae drop to the ground and molt to become free-living nymphs and finally adults, all of which feed on the eggs of insects. Trombiculids cause a type of dermatitis, which may be characterized by pruritus and bleeding. Areas frequently attacked are the ankles, legs, groin, external genitalia, waistline, axillae and breasts. The common species of trombiculids in the United States belong to the genus *Trombicula*. In parts of Asia, larvae of *T. akamushi* and related species serve as biological vectors of *Rickettsiae tsutsugamushi*, the etiologic agent of scrub typhus or tsutsugamushi fever. The disease is transmitted by the mites from rodent reservoir hosts to humans.

Several species of mites are known as mange mites. They infest the skin of mammals causing severe irritation, cutaneous lesions, and mange by burrowing into the skin. Psoroptic mange or "scab" is caused by mites belonging to the genera *Psoroptes* and *Chorioptes*. These mites feed on the surface of the skin, and remove portions of the epidermis as they spread upon the raw dermal layer under the scabs which form. Sarcoptic mange and "itch" is caused by burrowing mites. The "itch mite", *Sarcoptes scabiei*, forms cutaneous tunnels a few millimeters to several centimeters in length, and feeds on host tissues (Fig. 4). The itch mites of cattle, hogs, sheep, dogs and humans are believed to be varieties of this species. The adults, which range in length from 0.2 to 0.45 mm, are found on the surface of the skin at night, and can easily be transferred by personal contact and by clothing to new hosts. Gravid females deposit eggs and feces as they excavate tunnels in the epidermis. Three to five days after oviposition, the eggs hatch and the hexpod larvae emerge and either burrow off the

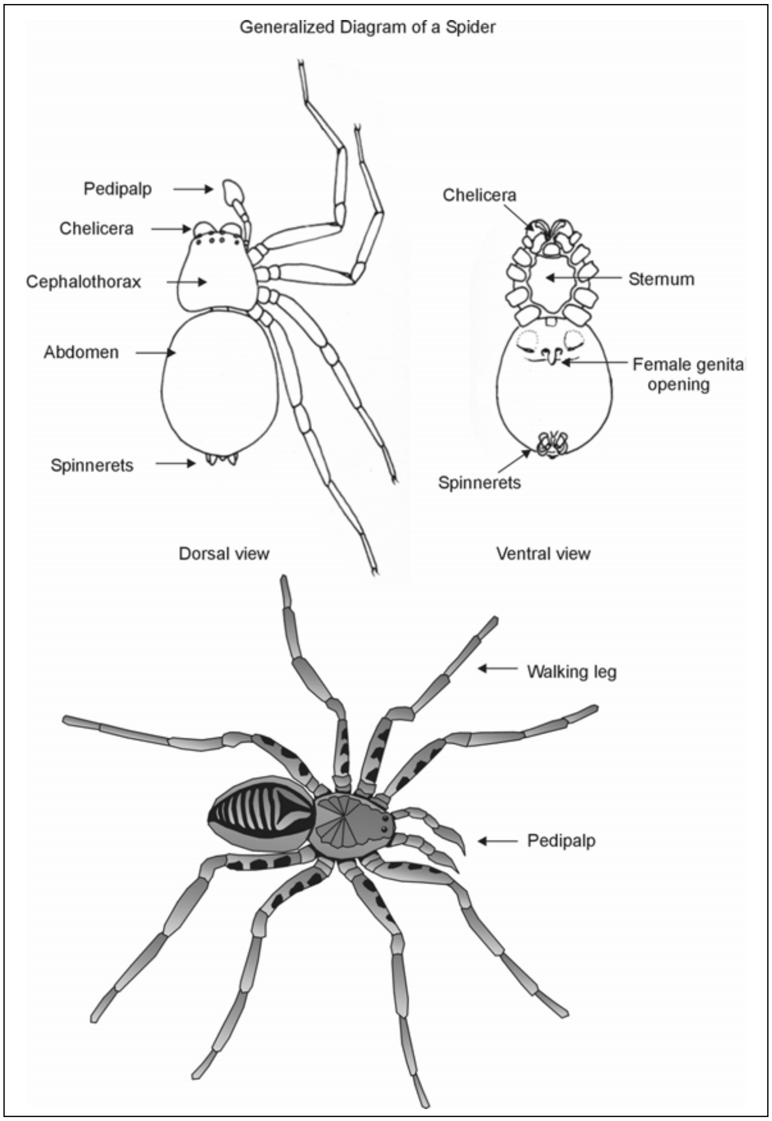
Generalized Diagram of a Spider
Pedipalp
Chelicera
Cephalothorax
Abdomen
Spinnerets
Chelicera
Sternum
Female genita opening
Spinnerets
Dorsal view
Ventral view
Walking leg
Pedipalp

Figure 2. Morphology of a typical spider.

main channel or exit and form new tunnels. The larvae molt to become nymphs, and later adults. The life cycle is completed in one to two weeks.

The initial infection is generally mild and asymptomatic, but after a few weeks the disease spreads and is accompanied by intense itching. Scratching causes lesions

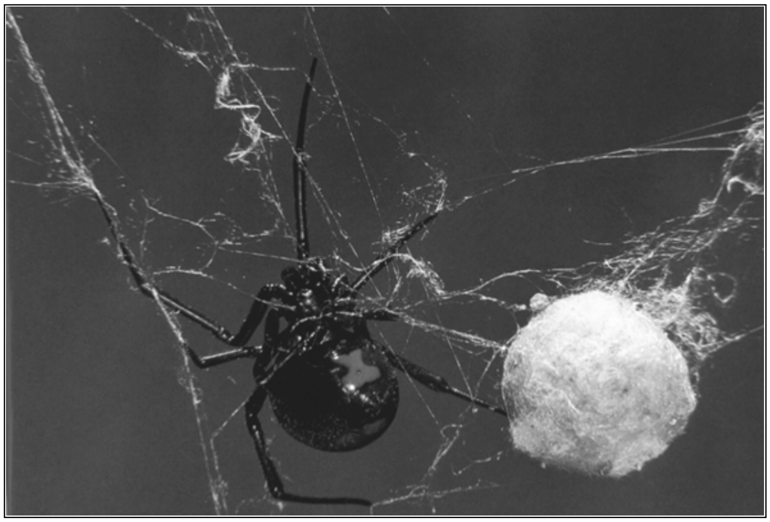

Figure 3. Black widow spider, *Latrodectus mactans*. From Knopf, A. A. 1997. National Audubon Society: Insects and spiders. Chanticleer Press, Inc., New York.

to form in the dermis, and secondary bacterial infections to develop. The lesions are most commonly found between the fingers and toes, on the back of the hands, in the groin, axillary regions, and external genitalia.

Demodex folliculorum, or follicle mite, is a small (0.1 to 0.4 mm), elongated, worm-like parasite that burrows into the hair follicles and sebaceous glands of humans (Fig. 5). The burrows are shallow, and the mites themselves generally produce only a mild pruritus, acne, or local keratitis. Demodectic mange in dogs results from bacterial complications following invasion of the skin by mites.

Various species of mites live on poultry and rodents. Those of medical and veterinary interest belong to the genera *Dermanyssus, Liponyssoides, Allodermanyssus* and *Ornithonyssus.* The common red chicken mite, *D. gallinae*, is found on various types of foul, and occasionally attacks humans. The mite is a nocturnal feeder, sucking blood and tissue fluids of the host. The introduction of toxic salivary secretions into the skin during feeding produces a severe dermatitis and papular eczema. The mite harbors viruses that cause St. Louis encephalitis and western equine encephalitis. The mouse mite, *L. sanguineus*, also produces a severe dermatitis in humans. In addition, the mite is the vector of *Rickettsiae akari*, the etiologic agent of human rickettsial pox. Mites transmit the disease from the reservoir host, the house mouse (*Mus musculus*), to humans. The rat mite, *O. bacoti*, and the tropical fowl mite, *O. bursa*, also produce an annoying dermatitis in humans. In addition, *O. bursa* transmits western equine encephalitis, and *O. bacoti* vectors *Rickettsia typhi* (the cause of endemic typhus), *Rickettsia akari* (the etiologic agent of rickettsial pox), and the virus that causes Q fever.

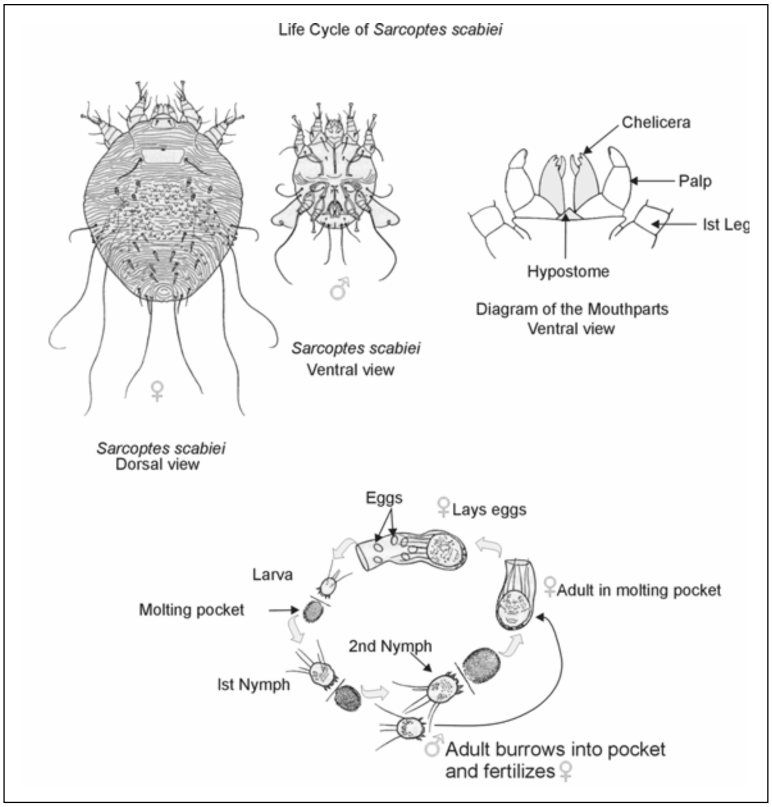
Life Cycle of *Sarcoptes scabiei*
♀
Sarcoptes scabiei
Dorsal view
♂
Sarcoptes scabiei
Ventral view
Chelicera
Palp
Ist Leg
Hypostome
Diagram of the Mouthparts
Ventral view
Eggs
♀Lays eggs
Larva
Molting pocket
♀Adult in molting pocket
2nd Nymph
Ist Nymph
♂Adult burrows into pocket
and fertilizes ♀

Figure 4. Life cycle of the itch mite, *Sarcoptes scabiei.*

Various species of itch mites are agricultural pests that inhabit warehouses and grain mills, or are predaceous on insects which infest grain crops. Dermatitis and allergic reactions variously termed “grocer’s itch”, “copra itch” or “miller’s itch”, result when humans come in contact with mite-infested plant products such as cereals, grains, cheese, dried fruits, and sugar. The mites generally crawl under the epidermis and produce temporary pruritism, but some have been found in the lungs, ears, intestinal and urinary passages.

Ticks

There are two distinct types of ticks, hard or ixodid ticks, and the soft or argasid ticks. Ixodid ticks have a dorsal shield or scutum covering virtually the entire back in the males, and only a small portion anteriorly in the females. Argasid ticks lack plates or shields, and have a leathery cutical. In ixodid ticks, the mouth is anterior in position, and the adults feed only once. In argasid ticks, the mouth is ventral, and the adults feed continually. The eggs are deposited on the ground, and the young hexapod larvae, termed “seed ticks”, climb up on vegetation and attach to their

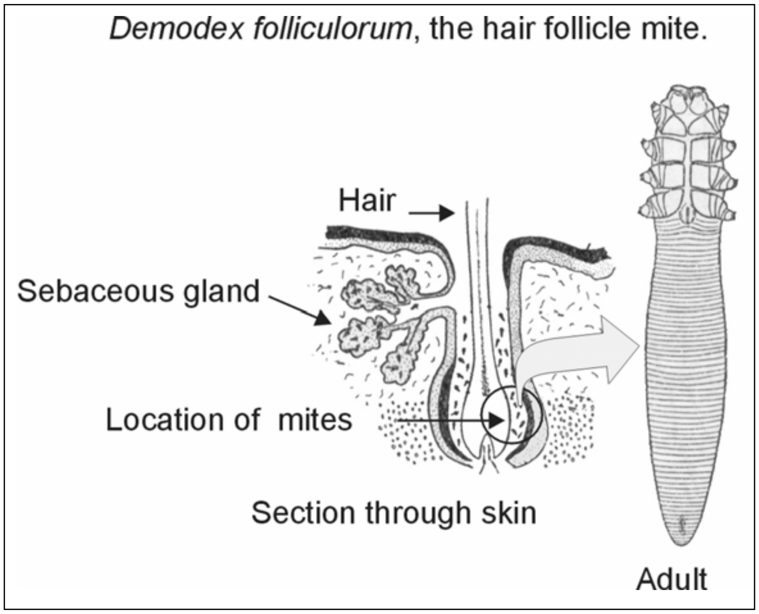

Figure 5. The hair follicle mite, *Demodex folliculorum*.

hosts on contact with them. The young tick makes an incision into the skin with the cheliceral digits, and the chelicerase and hypostome are then inserted into the wound (Fig. 6). The barbed hypostome does not aid in skin penetration, but prevents the animal from being dislodged from the skin. In addition to producing a histolytic secretion, which liquifies the host tissues, some species produce a salivary secretion, which hardens to a latex-like consistency around the embedded mouth parts, preventing the tick from being removed during the feeding period. The liquid food is sucked into the gut by the action of the muscular pharynx. After feeding for a few days, the larvae drop to the ground and molt to become eight-legged nymphs. The nymphs, in turn, will attach to their host and again feed. After feeding, the nymphs will return to the soil and molt to the adult stage. The adult ticks repeat the process, attaching themselves to their hosts and feeding.

Ticks are of considerable medical importance primarily because of their role as vectors of diseases caused by bacteria, viruses, and rickettsia. The argasid ticks of medical importance belong to the genera *Ornithodorous, Otobius* and *Argas.* Unlike the hard ticks, the argasids are nocturnal feeders, which hide in various cracks and crevices during the day. The medically important ixodid ticks belong to the genera *Dermacentor* and *Amblyomma*. In addition to being vectors of diseases, saliva introduced by some species into the wound during feeding may cause a type of sensitization reaction, which is systemic. Much more serious is tick paralysis, which at times is fatal. The paralysis appears to be an occasional host reaction to a toxin presumably elaborated from the eggs of feeding female ticks. Apparently, it occurs

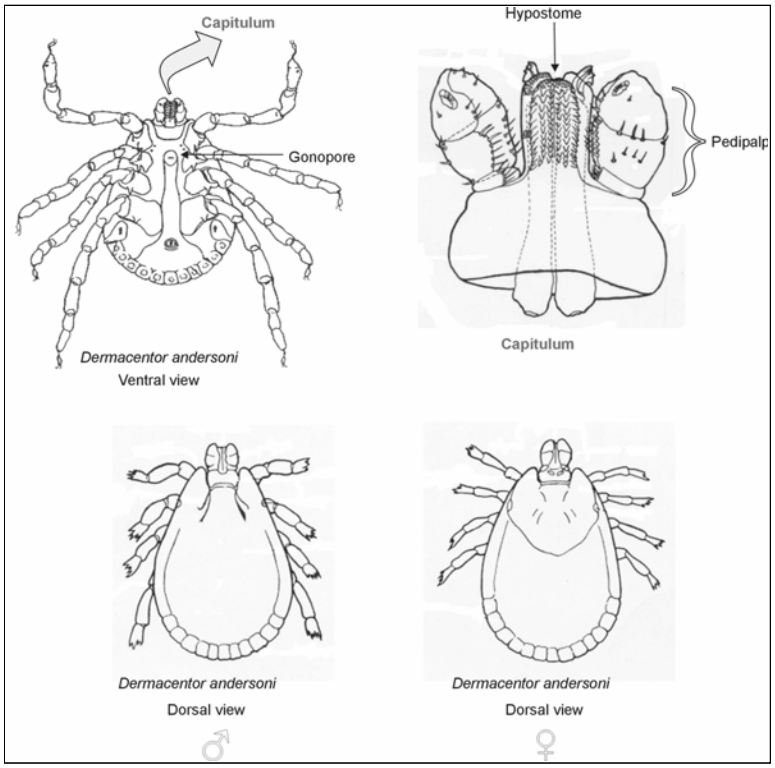
Capitulum
Gonopore
Dermacentor andersoni
Ventral view
Hypostome
Pedipalp
Capitulum
Dermacentor andersoni
Dorsal view
♂
Dermacentor andersoni
Dorsal view
♀

Figure 6. The ixodid tick, *Dermacentor andersoni.*

most frequently when ticks feed for extended periods of time at the base of the skull or back of the neck and along the spinal column. The ixodid ticks, which cause paralysis, are *D. andersoni* in western North America, and *D. variabilis, A. americanum* and *A. maculatum* in eastern and southern United States. The paralysis is characterized by ascending motor paralysis, which originates in the legs and moves up to include the chest and neck, and impairment of respiration, speech and swallowing. There is no specific treatment for tick paralysis. Paralysis terminates and recovery follows if the tick is removed before the heart and respiration are affected. Secondary infections with lesions and systemic poisoning may result if the capitulum is broken off when improper attempts are made to remove the embedded tick. The application of various oils or creams to the body of the tick generally initiates its withdrawal from the skin of the host.

Insects

Insects comprise about 70% of all known species of animals. Both economically and medically, insects represent the most important group of animals. The orders of insects containing species of medical importance include Coleoptera (beetles), Orthroptera (cockroaches), Hymenoptera (ants, bees), Anoplura (sucking lice), Hemiptera (true bugs), Siphonaptera (fleas) and Diptera (flies, mosquitoes, midges).

Coleoptera (Beetles)

Blister beetles produce a vesicating fluid, which is present in their body tissues. When the beetle is accidentally crushed on the skin, the substance causes irritation and blistering at the site of contact. A common vesicating beetle is *Lytta vesicatoria* or "Spanish fly". The cuticular hairs or small body fragments of some dermestid beetles may act as allergens producing asthma-like reactions when inhaled. The accidental ingestion of infected beetles, which serve as intermediate hosts of helminths can produce infections in humans. Examples include the tapeworm parasite of rodents, *Hymenolepis diminuta*, and the intestinal roundworm of pigs and ruminants, *Gongylonema pulchrum*.

Orthroptera (Cockroaches)

Cockroaches are extremely common wherever food or garbage is found. In addition to being effective transmitters of enteric pathogens, such as *Entamoeba histolytica*, they also serve as intermediate hosts for the cestode *Hymenolepis diminuta*, and the nematode *Gongylonema pulchrum*.

Hymenoptera (Ants, Bees)

Wasps and certain other Hymenoptera affect humans adversely chiefly by injecting venom into the skin. In some species the modified ovipositor of the female serves as a stinging apparatus (Fig. 7). The sting may cause only local irritation, but some species can inflict serious injury. The host reaction to the venom may be immediate or it may be delayed for hours. The fire ant, *Solenopsis* sp., which is found in the southern United States, produces a very painful, burning sensation following its sting. Some individuals become sensitized to components of certain bee and wasp venoms, and subsequent stings may cause fatal anaphylactic shock.

Anoplura (Sucking Lice)

Sucking lice are small, dorsoventrally flattened ectoparasites of mammals (Fig. 8). The mouthparts of these insects are adapted for piercing and sucking, and their legs are modified for clasping hairs or fibers. Lice are not only host-species specific, but also site specific. All developmental stages live permanently on the clothing or body surfaces of their hosts to which the eggs are cemented. Louse infestation (pediculosis) is most common among individuals living in crowded environments, such as military camps, prisons, and mental institutions, and where facilities for personal hygiene are poor or lacking. The human pubic louse, *Phthirus pubis*, attaches its eggs to the hairs of the pubic region, chest, eyebrows and eyelashes. The body louse, *Pediculus humanus*, occurs as two habitat-specific strains. One strain, designated *P. humanus capitis* (head louse), lives only on the head, attaching its eggs ("nits") to the hair of the head or neck (see Fig. 8). The second strain, *P. humanus corporis* (body louse or "cooties"), cements its eggs to clothing fibers. The two strains are not completely genetically isolated, and can interbreed easily. Saliva introduced by lice into the skin causes pruritus, and scratching of the lesions results in an eczematous dermatitis. *Pediculus humanus corporis* and *P. humanus capitis* are vectors of the organisms causing epidemic typhus, trench fever, and relapsing fever. Transmission

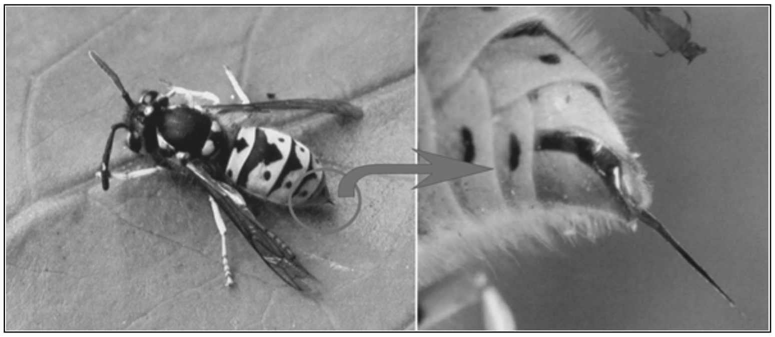

Figure 7. The insect order hymenoptera includes wasps, bees, and ants. In some species, the ovipositor is modified as a sting. From Knopf, A. A. 1997. National Audubon Society: Insects and spiders. Chanticleer Press, Inc., New York.

of the pathogens occurs when infected lice are crushed on the skin, or by venenation when the insects feed.

Hemiptera (True Bugs)

The order Hemiptera is comprised of insects known as true bugs. The most conspicuous external feature of these avid blood suckers is the proboscis, which is used to pierce tissues and suck fluids (Figs. 9 and 10). Two medically important families are the Cimicidae (bedbugs), and the Reduviidae (triatomids or assassin bugs). Bedbugs are dorsoventrally flattened, bloodsucking ectoparasites of birds and mammals. The two most common species attacking humans are *Cimex lectularius* of temporate climates, and *C. hemipterus* of the tropics. The former species occurs also on chickens, rabbits and bats. The adults and immature stages of bedbugs may be found in cracks and crevices, or under carpeting during the day, and emerge at night to feed. Inflamed, cutaneous lesions may be produced at the puncture sites, accompanied by systemic sensitization. The reduviids which attack humans and other vertebrates belong to the genera *Rhodnius, Triatoma* and *Panstrongylus*. Some assassin bugs commonly bite on the lips and near the eyes producing painful wounds and localized bleeding. Many reduviids also vector *Trypanosoma cruzi*, the etiologic agent of Chagas' disease.

Siphonaptera (fleas)

Fleas are ectoparasites with laterally compressed bodies. They are important as pests of humans and domestic animals and as vectors of disease. Fleas generally are not very host-specific, and unlike lice and acarines, fleas have a non-parasitic larval stage. Female fleas deposit eggs on the ground near sleeping areas of their hosts. The larvae, which have chewing mouthparts, feed on organic waste and fragments of hair and epidermis of the host. The adults have mouthparts adapted for piercing and sucking, and long, powerful legs providing them with excellent jumping ability (Fig. 11). The fleas of medical importance include the human flea (*Pulex irritans*), the dog flea (*Ctenocephalides canis*), the cat flea (*C. felis),* and the fleas of rats, mice, and other rodents (including *Xenopsylla cheopis, Nosopsyllus fasciatus* and *Leptopsylla*

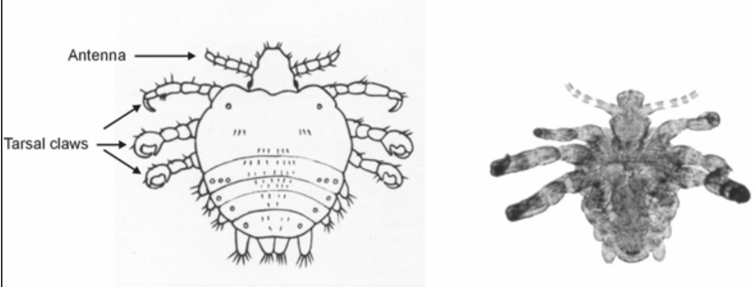

Pubic louse (*Phthirus pubis*)

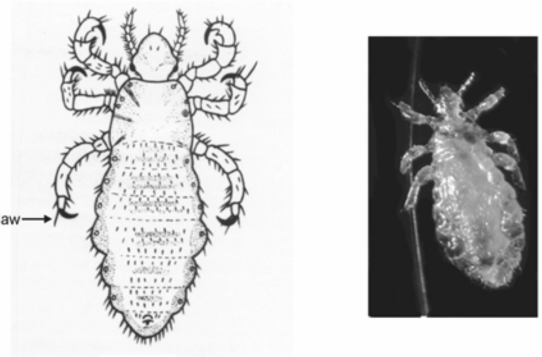

Body louse (*Pediculus humanus*)

Figure 8. Sucking lice *Phthirus pubis* and *Pediculus humanus.*

segnis). In some humans, fleas cause a local sensitization reaction at the site of infestation. The lesions may become secondarily infected as a result of scratching. Fleas from dogs and cats frequently produce a dermatitis when they attack humans. The dog flea also serves as an intermediate host of the dog tapeworm, *Dipylidium caninum*, which also can be transmitted to humans. The cysticercoid of *Hymenolepis diminuta* develops in larvae of several species of fleas. Rodent fleas transmit two serious diseases, bubonic plague and endemic typhus. The sticktight flea, *Echidnophaga gallinacea*, lives partly embedded in the skin of poultry, usually on the head near the eyes and bill. Occasionally the flea invades the ears of dogs and cats, and may also attack children. In addition, the flea can transmit plague among birds and rodents.

The sand flea or chigoe, *Tunga penetrans*, lives in sandy soil in tropical and subtropical America and parts of Africa. Unlike other fleas, which puncture the skin of humans and suck blood, *T. penetrans* invades the skin, usually burrowing under the toenails. The females engorge themselves by feeding on tissue fluids, and their abdomens become enormously distended. The fleas attach to a wide range of hosts, but are most commonly found on hogs and dogs. Humans walking barefoot on contaminated soil easily become infected. The parasite produces a painful sore which frequently becomes inflamed and secondarily infected and may result in lameness to

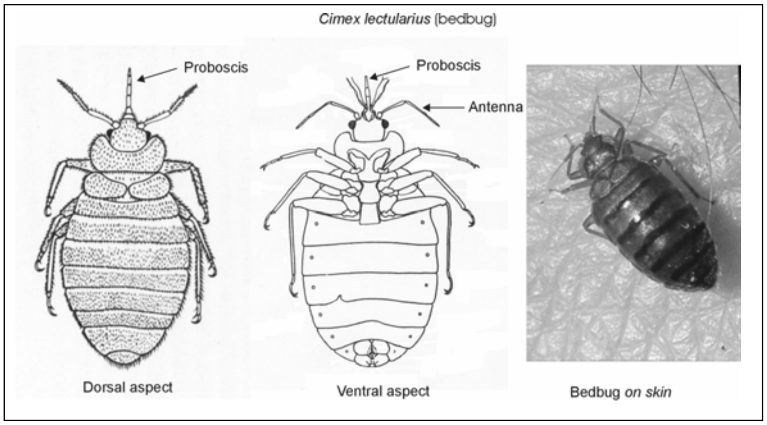
Cimex lectularius (bedbug)
Proboscis
Proboscis
Antenna
Dorsal aspect
Ventral aspect
Bedbug *on skin*

Figure 9. The bedbug, *Cimex lectularius.*

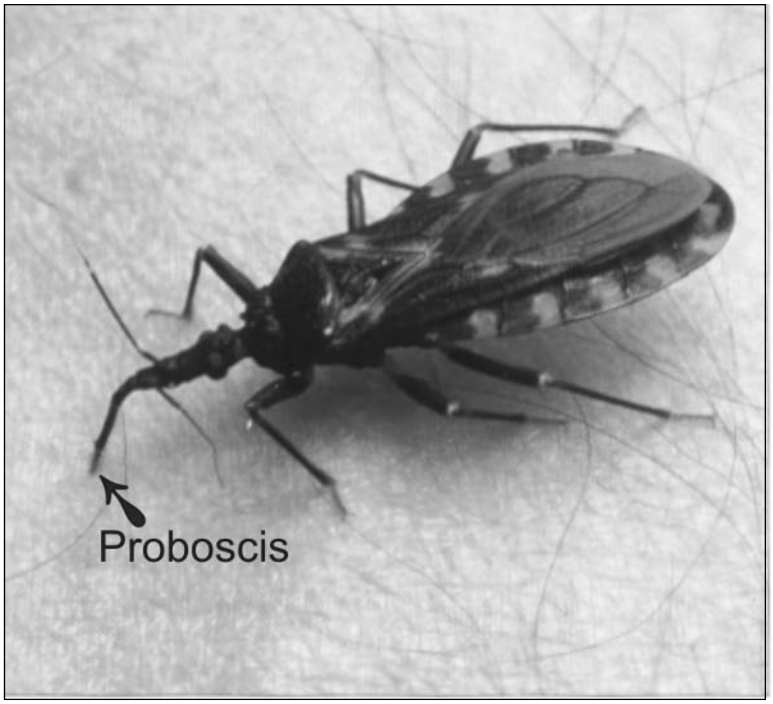

Figure 10. A reduviid bug.

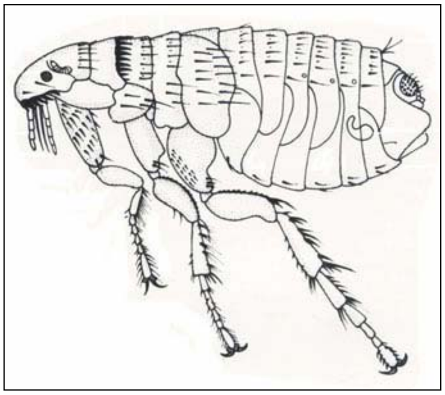

Figure 11. A flea.

the host. Death from gas gangrene and tetanus have been reported as complications of chigoe infestation.

Diptera (Flies, Mosquitoes, Midges)

Insects belonging to the Order Diptera are the most medically important arthropods. They are important ectoparasitic bloodsuckers that serve as biological and mechanical vectors of numerous diseases, and their larvae (maggots) are causal agents of myiasis. The adults of most species are characterized by having only one pair of functional, membranous wings, the forewings. The posterior wings are reduced to knoblike or club-shaped structures termed halteres which, in some species, are known to function as organs of balance.

The Family Psychodidae is comprised of sand flies (*Phlebotomus*), which are predominately nocturnal feeders. Only the females bite and suck blood. The insects produce an irritating pruritus at the site of attack. Allergic reactions develop in some hosts. Sand flies are biological vectors of the organisms that cause Leishmaniasis and bartonellosis (verrura peruana or Oroya fever).

The Family Ceratopogonidae is comprised of various small insects known as biting midges (“punkies”, “gnats” and “no-see-ums”). Only the females puncture the skin and feed on blood. The majority of medically important species belong to the genus *Culicoides.* They usually attack humans at dusk, inflicting painful bites which may be followed by severe pruritus. Some species transmit to humans and other vertebrates the filariids *Dipetalonema perstans* and *D. streptocerca*.

Mosquitoes belong to the Family Culicidae. They are not only serious pests, but also important vectors of disease. *Anopheles, Aedes, Culex, Mansonia* and *Psorophora* are among the more important genera affecting humans. Some of the major diseases transmitted by mosquitoes include malaria, filariasis (caused by the nematodes *Wuchereria brancrofti* and *Brugia malayi*), dog heartworm (caused by *Dirofilaria immitis*), dengue fever, equine encephalitis, west Nile fever, and yellow fever.

The Family Simulidae is comprised of black flies, or buffalo gnats. The adults have an elongated thorax (hump-back) and rasping mouthparts. They are diurnal feeders, which attack various birds and mammals. They feed on the abraded tissues, blood, and tissue fluids of their hosts. Only the female black flies suck blood. The adults are generally found close to streams and lakes, where the larval and pupal stages develop attached to rocks or to plants underwater. The females attach their eggs to objects at, or just below, the water surface. Occasionally, the adults occur in tremendous swarms causing considerable annoyance to humans and domestic animals. The bites may produce intensely pruritic lesions. Hypersensitization may occur in persons that are attacked repeatedly by black flies. Some species of *Simulium* are biological vectors of the medically important filarial worm *Onchocerca volvulus*, which causes river blindness.

Members of the Family Tabanidae are robust flies commonly known as horse flies, deer flies, or mango flies. Only the females pierce the skin of their hosts and suck blood. They are mostly diurnal species, which inflict very painful bites. Two bacterial diseases, tularemia (*Pasteurella tularensis*) and anthrax (*Bacillus anthracis*), are transmitted to humans and other animals by members of the genus *Tabanus*. *Chrysops*, which also carries tularemia, is the vector of the filarial worm *Loa loa*.

The family Chloropidae is comprised of small flies known as eye flies or fruit flies. The flies cluster about the eyes, mucous membranes, and open sores of humans and domestic animals. The adult flies are frequently encountered in sandy or moist areas, and near irrigated and cultivated farm lands. Although incapable of piercing the skin, they do produce small lesions with their spinose labella. These flies are believed to transmit pinkeye (acute contagious conjunctivitis) and yaws, a painful and disfiguring disease caused by the spirochaete *Treponema perenue.* The larval stages of eye flies breed in moist soil and in feces. The eye gnat, *Hippelates pusio*, occurs in the southern United States. A related species, *H. pallipes*, is common in the West Indies.

The Family Muscidae includes house flies, face flies, stable flies, horn flies, and tsetse flies. The house fly, *Musca domestica*, is a cosmopolitan pest, and a mechanical vector of the etiologic agents of typhoid (*Salmonella typhosa*), yaws (*Treponema pertenue*), dysentery (*Entamoeba histolytica*) and cholera (*Vibrio comma*). The fly breeds in decaying organic materials. Adults feed upon a wide variety of foods, which they liquify with saliva, and then lap up with their proboscis. The larvae have been reported as causing accidental cutaneous, gastrointestinal and urinary myiasis in humans.

Other medically important species of *Musca* include the face flies, *M. autumnalis, M. vicina* and *M. sorbens*, which crawl upon the face and near the eyes, mouth and nares. Some species of *Musca* serve as intermediate hosts of *Thelazia* spp., the eye worms of cattle, dogs and humans. The larval nematodes escape from the proboscis of the fly when the insect feeds on or around the eyes of the vertebrate host. Some species are capable of transmitting conjunctivitis. The stable fly, *Stomoxys calcitrans*, is a ravenous bloodsucker that feeds on horses and other domestic animals. The fly frequently attacks humans, inflicting painful bites. Both sexes suck blood. The females oviposit in manure. The larvae develop in the manure and later pupate in the surrounding soil. Instances of cutaneous and intestinal myiasis have been reported. The repeated feeding of adults with contaminated mouthparts on different hosts may result in the mechanical transmission of trypanosomiasis and cutaneous Leishmaniasis.

The tsetse fly, *Glossina*, is the vector of the etiologic agents of African sleeping sickness. Both males and females feed on cattle, wild animals, and humans. Female flies give birth to fully developed larvae. The eggs, which are produced one at a time, hatch in utero. Special intrauterine glands ("milk glands") secrete nourishment for the larvae, which develop through three larval stages. Fourth-stage larvae are deposited in the soil where pupation occurs. The principal vector of Gambian trypanosomiasis is *G. palpalis*, while *G. morsitans* and *G. swynnertoni* are the major vectors of Rhodesian trypanosomiasis.

The Family Calliphoridae includes the blow flies, screw worm flies, and bluebottle flies. The adults are not bloodfeeders, but the larvae of several species cause myiasis in domestic animals and in humans (Fig. 12). Female screw worm flies deposit and glue their eggs in masses at the edges of open wounds. After the eggs hatch, the larvae ingest host tissues, penetrate the wound, invade the sinuses, eyes, ears, mouth, and genital passages, producing painful lesions which characteristically become secondarily infected. The species of major importance include *Callitroga hominivorax* (the primary screw worm), and *Chrysomya* spp. (generally a secondary invader of sores). The former species is widely distributed from the southern United States to northern Chile. The latter species occurs throughout southern Asia and in parts of Africa. The tumbu fly, *Cordylobia anthropophaga*, is distributed throughout much of Africa. Eggs are deposited on the ground, and the larvae actively penetrate the skin of humans or other animals in contact with contaminated soil. The larvae feed on host tissues and produce indurated, cutaneous swellings from which they eventually emerge, drop to the ground, enter the soil and pupate.

Flesh flies, which belong to the family Sarcophagidae, are parasitic only during the larval stages, and thus are capable of producing myiasis. Female sarcophagids are larviparous, depositing their young in masses on carrion, feces, open wounds, or on healthy, undamaged body surfaces. Myiasis-causing species include *Sarcophaga* spp., *Titanogrypha* spp., *Wohlfahrtia magnifica* and *W. vigil.* Only the latter species larviposits on healthy skin, which the larvae actively penetrate. The larvae of all species characteristically produce small raised lesions as they burrow in the subcutaneous tissues of their hosts. The lesions communicate with the surface through a small pore.

The Family Hippoboscidae is comprised of louse flies, which are permanent bloodsucking parasites of birds and mammals. Adult hippoboscids look neither like lice nor flies, but more like six-legged ticks. The sheep ked, *Melphagous ovinus*, is a wingless ectoparasite of sheep and goats. Both sexes are voracious bloodsuckers. The parasite is distributed worldwide except in the tropics. The entire life cycle is spent on the host. Gravid females larviposit approximately one larva per week over a period of four months. Almost immediately after deposition, the larvae attach to the host's wool and pupate. Heavy infestations cause severe irritation, emaciation and anemia. The insects readily attack sheep shearers and other persons associated with infested animals, inflicting painful bites. Another hippobosid that occasionally bites humans is *Pseudolynchia canariensis,* or pigeon fly. The parasite is common on pigeons throughout most of the warm, temperate areas of the world.

The names warble fly and bot fly frequently are used interchangeably for flies belonging to the families Cuterebridae, Gastrophilidae, Oestridae, and Hypodermatidae. Warble flies are so-called because their larvae live for a period of time just under the skin of their vertebrate hosts and produce a swelling or "warble". *Dermatobia hominis*, or the human bot fly, is a member of the family Cuterbridae that occurs in Central and South America. Unlike other myiasis-causing flies, *D.*

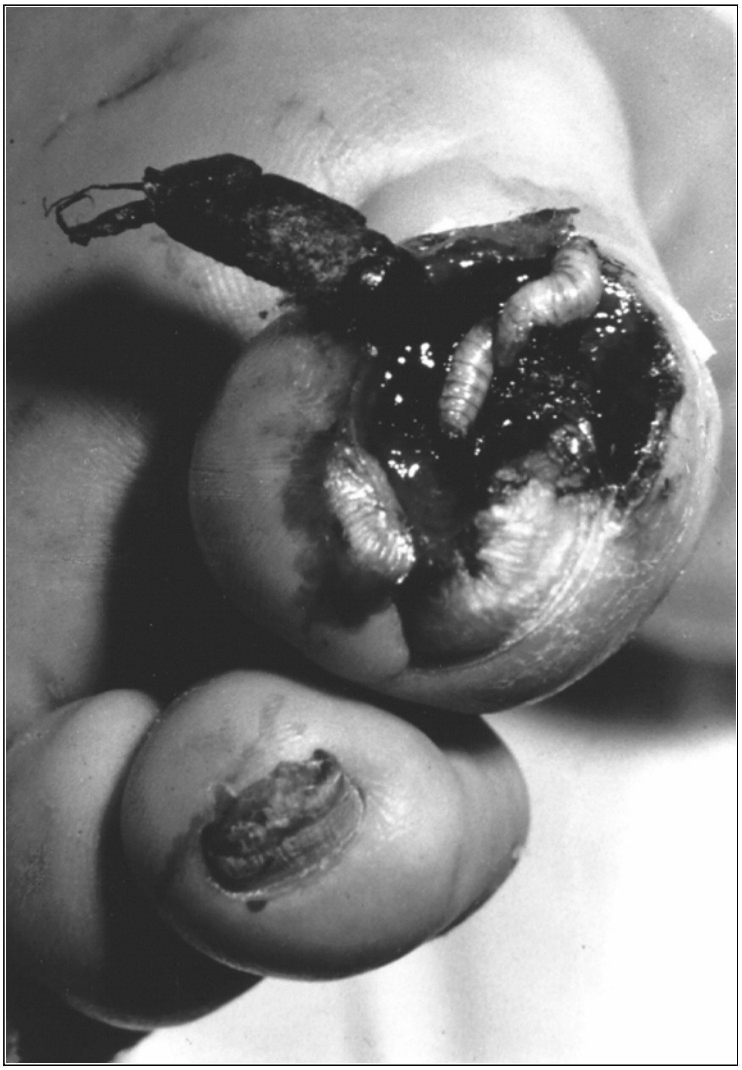

Figure 12. Dipterous larvae feeding on an open wound. Photograph courtesy of Dianora Niccolini.

hominis does not lay its eggs directly on its host. Instead, the non-biting female cements her eggs to the bodies of ticks, mosquitoes, or other dipterans, engaging these arthropods in transporting the eggs to humans or other suitable mammals. After the bot fly eggs hatch, the larvae penetrate the host's skin. The larvae migrate to the subcutaneous tissues and produce inflamed, boil-like pockets, each with a separate opening to the surface. After a period of one to two months, the mature larvae emerge from the skin, drop to the ground and pupate. In humans, cutaneous lesions frequently develop on the wrists, ankles, neck and face. Heavy infestations of cattle and dogs may terminate fatally. Early surgical removal of the larvae is an effective treatment. The family Gastrophilidae comprises the stomach bots of equids, elephants, and rhinoceroses. The adults, which are similar to honeybees in size and appearance, are nonfeeders. The parasitic larvae feed on the mucous of the host's stomach and intestines, causing enteric myiasis. Heavy infections cause severe damage to the host's gut, and if not treated early terminate fatally.

Three species are common in North America and throughout most of the world; *Gastrophilus intestinalis*, the horse bot, *G. nasalis*, or throat bot and *G. haemorrhoidalis*, the nose bot. Females of *G. intestinalis* attach their eggs to the hairs of the horse, usually on the knees and lower legs. When the horse licks these areas, the friction, warmth and moisture provided by the tongue stimulate hatching. The first-stage larvae penetrate the lips and tongue and tunnel down to the stomach and small intestine. When mature, the larvae detach from the mucosa of the alimentary canal, pass from the host in the feces and undergo pupation in the soil. The eggs of *G. nasalis* are usually attached to the hairs on the underside of the jaws, and the larvae crawl into the mouth where they feed for a time in pockets between the teeth before migrating to the gut. The eggs of *G. haemorrhoidalis* are deposited around the lips. The route of larval migration to the gut parallels that of *G. intestinalis* except that third-stage larvae attach to the anus for a short time before passing out with the feces. Occasionally, larvae of all three species penetrate human skin and mucous membranes, causing a creeping myiasis as they migrate under the epidermis. The serpiginous tunnels formed by these migrating fly larvae resemble cutaneous larva migrans produced by hookworm larvae. Human infestations usually occur on the extremities, and the maggots can be easily removed surgically.

Hypoderma bovis, the cattle bot, and *H. lineatum*, the heel fly, are two cosmopolitan species belonging to the family Hypodermatidae. The adult flies, which resemble bumble bees in size and appearance, do not feed. Females cement their eggs on the legs or body hairs of cattle. On hatching, the larvae penetrate the skin and, for a period of several months, migrate through the body, invading the intestines, liver, heart, esophagus, and other organs. Eventually, the larvae migrate to the back of the hosts, where they produce subcutaneous swellings termed warbles. The larvae make small holes in the skin through which they obtain air. The larvae remain in the skin for about two months, during which time they feed and grow to a length of about 25 mm. When mature, the larvae exit the skin and pupate in the soil. Cutaneous myiasis frequently develops in humans who are in contact with heavily infested cattle. The lesions, which are frequently inflamed and very painful, are usually on the extremities and in the neck region. Larvae of *Hypoderma*, unlike those of *Gastrophilus*, can successfully migrate and develop in humans.

Head maggots, which belong to the family Oestridae, are similar in size and shape to honeybees. The adult flies do not feed, but the larvae invade and parasitize the nasal cavities and sinuses of sheep, goats, horses, and other hoofed animals. The sheep bot, *Oestrus ovis*, is a cosmopolitan parasite of domestic sheep and goats. The female is larviparous and deposits maggots in the nares of their hosts. The larvae then enter the sinuses and attach to the mucous membranes and feed. The mature larvae emerge from the nostrils and drop to the ground to pupate in the soil. *Rhinostrus purpureus* is a head maggot of horses in Europe, northern Africa, parts of Asia and Central and South America. Females larviposit onto the eyes and nostrils of the host, and occasionally cause opthalmic myiasis. The larvae, which do not mature in humans, cause inflammation, conjunctivitis, and in severe cases blindness.

SELECTED REFERENCES

The following references include relatively recent articles of a general nature that address varied topics concerned with tropical medicine, hygiene, and host-parasite relationships.

1. Abath FGC. Development of vaccines against human parasitic diseases: Tools, current status and perspectives. Expert Opinion on Investigational Drugs 2000; 9:301-310.
2. Abath FGC, Montenegro SML, Gomes YM. Vaccines against human parasitic diseases: An overview. Acta Tropica 1998; 71:237-254.
3. Ambroise-Thomas P. Emerging parasite zoonoses: the role of host-parasite relationship. Internat J Parasitol 2000; 30:1361-1367.
4. Ambroise-Thomas P. Parasitic diseases and immunodeficiencies. Parasitology, 2001; 122:S65-S71.
5. Arya SC. Global warming and the performance of drugs used to treat parasitic and other diseases. Ann Trop Med Parasitol 1999; 93:207-208.
6. Ashford RW. Current usage of nomenclature for parasitic diseases, with special reference to those involving arthropods. Med Vet Entomol 2001; 15:121-125.
7. Aspock H, Auer H, Picher O. Parasites and parasitic diseases in prehistoric human populations in Central Europe. Helminthologia 1999; 36:139-145.
8. Ayles HM, Corbett EL, Taylor I, Cowie AGA, Bligh J, Walmsley K et al. A combined medical and surgical approach to hydatid disease: 12 years' experience at the Hospital for Tropical Diseases, London. Ann Royal Coll Surg Eng 2002; 84:100-105.
9. Baker DG, Bryant JD, Urban JF, Lunney JK. Swine immunity to selected parasites. Vet Immunol Immunopathol 1994; 43:127-133.
10. Barcinski MA, Costamoreira ME. Cellular response of protozoan parasites to host-derived cytokines. Parasitol Today 1994; 10:352-355.
11. Barcinski MA, DosReis GA. Apoptosis in parasites and parasite-induced apoptosis in the host immune system: A new approach to parasitic diseases. Brazilian J Med Biolog Res 1999; 32:395-401.
12. Barillas-Mury C, Wizel B, Han YS. Mosquito immune responses and malaria transmission: lessons from insect model systems and implications for vertebrate innate immunity and vaccine development. Insect Biochem Mol Biol 2000; 30:429-442.
13. Barry M, Maguire JH, Weller PE. The American Society of Tropical Medicine and Hygiene initiative to stimulate educational programs to enhance medical expertise in tropical diseases. Am J Trop Med Hyg 1999; 61:681-688.
14. Beerntsen BT, James AA, Christensen BM. Genetics of mosquito vector competence. Microbiol Mol Biol Rev 2000; 64:115-+.
15. Bell A. Microtubule inhibitors as potential antimalarial agents. Parasitol Today 1998; 14:234-240.
16. Bergquist NR. Vector-borne parasitic diseases: New trends in data collection and risk assessment. Acta Tropica 2001; 79:13-20.
17. Blancou J, Meslin FX. A brief historical overview of zoonoses. Revue Scientifique Et Technique De L Office International Des Epizooties 2000; 19:15-22.
18. Borst P, Ouellette M. New mechanisms of drug resistance in parasitic protozoa. Ann Rev Microbiol 1995; 49:427-460.

19. Burkhart CN. Ivermectin: An assessment of its pharmacology, microbiology and safety. Vet Hum Toxicol 2000; 42:30-35.
20. Bynum WF. Transmissible diseases at the end of the XIXth century: Vectors and pathogenic agents. Bulletin de la Societe de Pathologie Exotique 1999; 92:408-410.
21. Carlier Y, Truyens C. Influence of maternal infection on offspring resistance towards parasites. Parasitol Today 1995; 11:94-99.
22. Chakraborty AK, Majumder HK. Molecular biology of *Leishmania*: Kinetoplast DNA and DNA topoisomerases as novel therapeutic targets. Current Science 1999; 76:1462-1472.
23. Chang KH, Han MH. MRI of CNS parasitic diseases. J Mag Reson Imag 1998; 8:297-307.
24. Christensen BM, Ferdig MT. The field of vector biology—A perspective on the first 100 years. Parasitol Today 1997; 13:295-297.
25. Chu DTW. The future role of quinolones. Expert Opinion on Therapeutic Patents 1996; 6:711-737.
26. Colley DG. Parasitic diseases: Opportunities and challenges in the 21st century. Memorias do Instituto Oswaldo Cruz 2000; 95:79-87.
27. Combes C. Fitness of parasites—Pathology and selection. Internat J Parasitol 1997; 27:1-10.
28. Coop RL, Kyriazakis I. Nutrition-parasite interaction. Vet Parasitol1999; 84:187-204.
29. Couturier E, Hansmann Y, Descampeaux C, Christmann D. The epidemiology of infective endocarditis. Medecine et Maladies Infectieuses 2000; 30:162-168.
30. Dailey MD. Parasitic diseases. CRC Handbook Of Marine Mammal Medicine. Second Edition. 2001.
31. Dedet JP. Update on leishmaniases. Presse Medicale 2000; 29:1019-1026.
32. Diouf S, Diagne I, Moreira C, Sy HS, Faye O, Ndiaye O et al. Iron, vitamin A deficiencies and parasitic diseases: Impact on Senegalese children developement. Archives de Pediatrie 2002; 9:102-103.
33. Dodd RY. Transmission of parasites by blood transfusion. Vox Sanguinis 1998; 74:161-163.
34. Duvic C, Nedelec G, Debord T, Herody M, Didelot F. Imported parasitic nephritis: Review of medical literature. Nephrologie 1999; 20:65-74.
35. Ebert D, Hamilton WD. Sex against virulence—The coevolution of parasitic diseases. Trends Ecol Evol 1996; 11:A 79-A 82.
36. Ebert D, Herre EA. The evolution of parasitic diseases. Parasitol Today 1996; 12:96-101.
37. Edwards G, Winstanley PA, Ward SA. Clinical pharmacokinetics in the treatment of tropical diseases—Some applications and limitations. Clinical Pharmacokinetics 1994; 27:150-165.
38. Eyckmans L. Teaching tropical medicine—A personal view in perspective. Ann Trop Med Parasitol 1997; 91:743-746.
39. Feldmeier H, Poggensee G, Krantz I, Hellinggiese G. Female genital schistosomiasis—New challenges from a gender perspective. Trop Geograph Med 1995; 47:S 2-S 15.
40. Gallup JL, Sachs JD. The economic burden of malaria. Am J Trop Med Hyg 2001; 64:85-96.
41. Giamarellou H. AIDS and the skin: Parasitic diseases. Clinics in Dermatology 2000; 18:433-439.
42. Gilles HM, Lucas AO. Tropical medicine—100 years of progress. Br Med Bull 1998; 54:269-280.

43. Gomez-Lus R, Clavel A, Castillo J, Seral C, Rubio C. Emerging and reemerging pathogens. Internat J Antimicrob Agents 2000; 16:335-339.
44. Gubler DJ. Prevention and control of tropical diseases in the 21st century: Back to the field. Am J Trop Med Hyg 2001; 65:V-XI.
45. Harms G, Feldmeier H. Review: HIV infection and tropical parasitic diseases—Deleterious interactions in both directions? Trop Med International Health 2002; 7:479-488.
46. Hellinggiese G, Feldmeier H, Racz P, Hickl EJ. Female genital schistosomiasis (FGS)—Literature review and two case presentations. Geburtshilfe und Frauenheilkunde 1997; 57:136-140.
47. Heyworth MF. Parasitic diseases in immunocompromised hosts—Cryptosporidiosis, isosporiasis, and strongyloidiasis. Gastroenterology Clinics of North America 1996; 25:691 ff.
48. Hoberg EP, Alkire NL, de Queiroz A, Jones A. Out of Africa: Origins of the Taenia tapeworms in humans. Proceedings of the Royal Society of London—Series B: Biological Sciences 2001; 268:781-787.
49. Houszka M. Metastrongylosis as an agent in the population decrease of wild boars. Medycyna Weterynaryjna 2001; 57:638-640.
50. Hutubessy RCW, Bendib LM, Evans DB. Critical issues in the economic evaluation of interventions against communicable diseases. Acta Tropica 2001; 78:191-206.
51. Ijumba JN, Lindsay SW. Impact of irrigation on malaria in Africa: Paddies paradox. Med Vet Entomol 2001; 15:1-11.
52. Innes EA. Emerging parasitic diseases, bioterrorism and the New World order. Parasitol Today 1999; 15:427-428.
53. Irwin PJ. Companion animal parasitology: A clinical perspective. Intern J Parasitol 2002; 32:581-593.
54. Jacobson ER. Causes of mortality and diseases in tortoises—A review. J Zoo Wildlife Med 1994; 25:2-17.
55. James WD. Imported skin diseases in dermatology. J Dermatol 2001; 28:663-666.
56. Johnston DA, Blaxter ML, Degrave WM, Foster J, Ivens AC, Melville SE. Genomics and the biology of parasites. Bioessays 1999; 21:131-147.
57. Jokiranta TS, Jokipii L, Meri S. Complement resistance of parasites. Scand J Immunol 1995; 42:9-20.
58. Kaplan JE, Hu DJ, Holmes KK, Jaffe HW, Masur H, Decock KM. Preventing opportunistic infections in human immunodeficiency virus-infected persons—Implications for the developing world. Am J Trop Med Hyg 1996; 55:1-11.
59. Kimber KR, Kollias GV. Infectious and parasitic diseases and contaminant-related problems of North American river otters (Lontra canadensis): A review. J Zoo Wildlife Med 2000; 31:452-472.
60. Kinnamon KE, Engle RR, Poon BT, Ellis WY, McCall JW, Dzimianski MT. Polyamines: Agents with macrofilaricidal activity. Ann Trop Med Parasitol 1999; 93:851-858.
61. Kristensen TK, Brown DS. Control of intermediate host snails for parasitic diseases—A threat to biodiversity in African freshwaters? Malacologia 1999; 41:379-391.
62. Le Souef PN, Goldblatt J, Lynch NR. Evolutionary adaptation of inflammatory immune responses in human beings. Lancet 2000; 356:242-244.
63. Liance M. Progress toward the development of vaccines against parasitic diseases—A review based on studies performed by participants of the laveran seminar. Parasite-Journal de la Societe Francaise de Parasitologie 1994; 1:197-203.
64. Lightowlers MW, Gauci CG. Vaccines against cysticercosis and hydatidosis. Vet Parasitol 2001; 101:337-352.

65. Londershausen M. Approaches to new parasiticides. Pesticide Science 1996; 48:269-292.
66. Luder CGK, Gross U, Lopes MF. Intracellular protozoan parasites and apoptosis: Diverse strategies to modulate parasite-host interactions. Trends Parasitol 2001; 17:480-486.
67. Mackenzie CD. Anthelmintic therapy—Current approaches and challenges. Current Opinion in Infectious Diseases 1993; 6:812-823.
68. Mahe A. Bacterial skin infections in a tropical environment. Curr Opin Infect Dis 2001; 14:123-126.
69. Maizels RM, Holland MJ. Parasite immunity—Pathways for expelling intestinal helminths. Current Biology 1998; 8:R 711-R 714.
70. Marton LJ, Pegg AE. Polyamines as targets for therapeutic intervention. Ann Rev Pharmacol Toxicol 1995; 35:55-91.
71. Michael E, Grenfell BT, Isham VS, Denham DA, Bundy DAP. Modelling variability in lymphatic filariasis—Macrofilarial dynamics in the *Brugia pahangi* cat model. Proceedings of the Royal Society of London—Series B: Biological Sciences 1998; 265:155-165.
72. Mitchell G, Nossal G. Funding options for research: Facing the market as well as government. Internat J Parasitol 1999; 29:819-831.
73. Molyneux DH. Vectorborne parasitic diseases—An overview of recent changes. Internat J Parasitol 1998; 28:927-934.
74. Morgado MG, Barcellos C, Pina MD, Bastos FI. Human immunodeficiency virus/acquired immunodeficiency syndrome and tropical diseases: A Brazilian perspective. Memorias do Instituto Oswaldo Cruz 2000; 95:145-151.
75. Mott KE, Nuttall I, Desjeux P, Cattano P. New geographical approaches to control of some parasitic zoonoses. Bulletin of the World Health Organization 1995; 73:247-257.
76. Murray M. The parasites, predators, places and people I have known: A great adventure. Vet Parasitol 1999; 81:149-158.
77. Nozais JP. Parasitic diseases and fecal peril: Diseases due to helminths. Bulletin de la Societe de Pathologie Exotique 1998; 91:416-422.
78. Osman M, Lausten SB, Elsefi T, Boghdadi I, Rashed MY, Jensen SL. Biliary parasites. Digestive Surgery 1998; 15:287-296.
79. Ouaissi A, Ouaissi M, Sereno D. Glutathione S-transferases and related proteins from pathogenic human parasites behave as immunomodulatory factors. Immunol Lett 2002; 81:159-164.
80. Pantanowitz L, Telford SR, Cannon ME. Tick-borne diseases in transfusion medicine. Transfusion Medicine 2002; 12:85-106.
81. Papadopoulou B, Kundig C, Singh A, Ouellette M. Drug resistance in leishmania—Similarities and differences to other organisms. Drug Resistance Updates 1998; 1:266-278.
82. Papadopuloseleopulos E, Turner VF, Papadimitriou JM, Bialy H. AIDS in Africa—Distinguishing fact and fiction. World J Microbiol Biotechnol 1995; 11:135-143.
83. Patel R, Paya CV. Infections in solid-organ transplant recipients. Clinical Microbiology Reviews 1997; 10:86 ff.
84. Patz JA, Graczyk TK, Geller N, Vittor AY. Effects of environmental change on emerging parasitic diseases. Internat J Parasitol 2000; 30:1395-1405.
85. Peng SL. Rheumatic manifestations of parasitic diseases. Seminars in Arthritis & Rheumatism 2002; 31:228-247.
86. Perlmann P, Bjorkman A. Malaria research: host-parasite interactions and new developments in chemotherapy, immunology and vaccinology. Curr Opin Infect Dis 2000; 13:431-443.

87. Pointier JP. Invading freshwater gastropods: Some conflicting aspects for public health. Malacologia 1999; 41:403-411.
88. Poli G, Vicenzi E, Ghezzi S, Lazzarin A. Cytokines in the acquired immunodeficiency syndrome and other infectious diseases. Internat J Clin Lab Res 1995; 25:128-134.
89. Repetto R, Baliga SS. Pesticides and immunosuppression—The risks to public health. Health Policy & Planning 1997; 12:97-106.
90. Roberts CW, Walker W, Alexander J. Sex-associated hormones and immunity to protozoan parasites. Clin Microbiol Rev 2001; 14:476-+.
91. Roberts T, Murrell KD, Marks S. Economic losses caused by foodborne parasitic diseases. Parasitol Today 1994; 10:419-423.
92. Rodriguez JB. Specific molecular targets to control tropical diseases. Curr Pharm Des 2001; 7:1105-1116.
93. Ruppel A, Doenhoff MJ. Vector biology and the control of parasitic diseases. Parasitol Today 1998; 14:299-300.
94. Ryan TJ. Women in dermatology—Gender and tropical diseases. Internat J Dermatol 1995; 34:226-235.
95. Sangster NC, Gill J. Pharmacology of anthelmintic resistance. Parasitol Today 1999; 15:141-146.
96. Sansom C. New drugs needed for tropical diseases. Lancet Infect Dis 2002; 2:134.
97. Scully C, Monteil R, Sposto MR. Infectious and tropical diseases affecting the human mouth. Periodontology 2000; 18:47-70.
98. Sepulvedaboza S, Cassels BK. Plant metabolites active against *Trypanosoma cruzi*. Planta Medica 1996; 62:98-105.
99. Stanley SL. Protective immunity to amebiasis: New insights and new challenges. J Infect Dis 2001; 184:504-506.
100. Stanley SL, Virgin HW. *Scid* mice as models for parasitic infections. Parasitol Today 1993; 9:406-411.
101. Szenasi Z, Endo T, Yagita K, Nagy B. Isolation, identification and increasing importance of free-living amoebae causing human disease. J Med Microbiol 1998; 47:5-16.
102. Taverne J. Transgenic mice and the study of cytokine function in infection. Parasitol Today 1994; 10:258-262.
103. Thomas PK. Tropical neuropathies. J Neurol 1997; 244:475-482.
104. Thompson RCA. The future of veterinary parasitology: a time for change? Vet Parasitol 2001; 98:41-50.
105. Trouiller P, Olliaro P, Torreele E, Orbinski J, Laing R, Ford N. Drug development for neglected diseases: A deficient market and a public-health policy failure. Lancet 2002; 359:2188-2194.
106. Trouiller P, Olliaro PL. Drug development output: What proportion for tropical diseases? Lancet 1999; 354:164.
107. Uilenberg G. Integrated control of tropical animal parasitoses. Trop An Health Prod 1996; 28:257-265.
108. Ullman B, Carter D. Hypoxanthine-guanine phosphoribosyltransferase as a therapeutic target in protozoal infections. Infectious Agents & Disease 1995; 4:29-40.
109. Valerio L, Sabria M, Fabregat A. Tropical diseases in the Western world. Medicina Clinica 2002; 118:508-514.
110. Vanhamme L, Pays E. Control of gene expression in trypanosomes. Microbiologic Rev1995; 59:223-240.

111. Vermeulen AN. Progress in recombinant vaccine development against coccidiosis—A review and prospects into the next millennium. Internat J Parasitol 1998; 28:1121-1130.

112. Vial HJ, Traore M, Fairlamb AH, Ridley RG. Renewed strategies for drug development against parasitic diseases. Parasitol Today 1999; 15:393-394.

113. Vuitton DA, Godot V, Harraga S, Liance M, Beurton I, Bresson-Hadni S. Echinococcoses: A parasitic model for understanding allergic diseases? Revue Francaise d Allergologie et d Immunologie Clinique 2001; 41:285-293.

114. Wasson K, Peper RL. Mammalian microsporidiosis. Vet Pathol 2000; 37:113-128.

115. Werbovetz KA. Target-based drug discovery for malaria, leishmaniasis, and trypanosomiasis. Current Medicinal Chemistry 2000; 7:835-860.

116. Zarnbrano-Villa S, Rosales-Borjas D, Carrero JC, Ortiz-Ortiz L. How protozoan parasites evade the immune response. Trends Parasitol 2002; 18:272-278.

117. Zingales B, Rondinelli E, Degrave W, Dasilveira JF, Levin M, Lepaslier D et al. The *Trypanosoma cruzi* genome initiative. Parasitol Today 1997; 13:16-22.

GLOSSARY

Acanthella
The larval stage of an acanthocephalan following the acanthor and prior to the cystacanth.

Acanthor
The larval stage of an acanthocephalan; that develops inside an egg capsule and possesses bladelike hooks.

Accidental or incidental host
An organism, other than the normal host species, in which a parasite may or may not continue its development.

Acetabulum
Ventral sucker or holdfast structure of digenetic trematodes.

Acoelomate
The absence of a body cavity. The internal organs (viscera) lie embedded in parenchyma.

Acquired immunity
A host's immune response to a previous infection.

Ala (pl. alae)
Cuticular wing-like projection in certain nematodes.

Allergy
See hypersensitivity.

Alveolus (pl. alveoli)
An air sac in the lung where gaseous exchange occurs.

Amastigote
A small, ovoid, form of hemoflagellates that develops intracellularly.

Ametabolous
A type of insect metamorphosis in which there is no significant morphological change as the organism proceeds through a series of molts to the adult stage.

Amphids
Sensory depressions or pits, believed to be chemoreceptors, located anteriorly on the body surface of certain nematodes.

Anaphylaxis (anaphylactic shock)
An exaggerated or hypersensitive reaction by a host in response to a foreign protein (allergen) involving histamine-release.

Anapolysis
The process in which terminal, gravid proglottids or segments of a tapeworm are not shed.

Anisogamete
Morphologically different male and female gametes.

Anorexia
Loss of appetite.

Anterior station
The development of protozoan parasites in the anterior part of an insect vector, with transmission to the definitive host occurring when the vector bites.

Anthelmintic
A chemical used to remove worms, usually from the intestinal tract.

Anthrax
A bacterial disease of humans, cattle, and sheep, transmitted by tabanid flies.

Antibody
A specific serum protein (immunoglobulin) synthesized by B lymphoid cells (plasma cells) in response to an antigen.

Antigen
A substance, usually a protein, capable of inducing the host to synthesize antibodies.

Antigenic mimicry
Acquisition or production of host-like molecules by a parasite so that it is not recognized by the host as foreign, thus circumventing an immune response.

Apical complex
A combination of secretary structures found in the apical region of sporozoites and merozoites of members of the phylum Apicomplexa.

Apicomplexa
A phylum containing animals whose life cycles include feeding stages (trophozoites), asexual multiplication (schizogony), and sexual multiplication (gametogony and sporogony).

Apolysis
The process in which terminal, gravid or egg-filled proglottids are detached and shed from tapeworm strobila.

Arachnida
A class in the phylum Arthropoda containing ticks, mites, spiders, and scorpions.

Arbovirus
A virus transmitted from one human to another by an arthropod.

Arthropoda
A phylum comprised of animals having a chitinous exoskeleton and paired jointed legs. Includes insects, crustaceans and arachnids.

Atrium (pl. atria)
An opening into the body (e.g., mouth, urethra).

Autogeny
The ability of some bloodsucking arthropods to lay eggs without having had a blood meal (adj. autogenous).

Autoinfection
Reinfection of a host by the progeny of an existing parasitic organism residing within the host.

Axoneme
An intracellular microtubular portion of the flagellum or cilium.

Axostyle
A tube-shaped sheath of microtubules, observed in many flagellates (e.g., trichomonads), that usually extends from a basal body to the posterior end, and functions as a supporting cytoskeleton.

B cell
A specialized lymphocyte that produces antibodies.

Basal body (= Blepharoplast)
A centriole-like organelle from which the flagellum or cilium originates.

Biological vector
A host that is required for the development of the parasite and for transferring the parasite to another host.

Biramous
A structure divided into two branches.

Blackwater fever
Massive lysis of vertebrate erythrocytes that occasionally accompanies falciparum malaria.

Bothridium (pl. bothridia)
A muscular, leaf-like adhesive groove on the scolex of certain tapeworms of the order Pseudophyllidea.

Bradyzoites
Slowly multiplying intracellular trophozoites of *Toxoplasma gondii* that form pseudocysts in immune hosts.

Brood capsule
A structure within the daughter cyst in Echinococcus granulosus, in which numerous scolices develop. In the definitive host each scolex can develop into an adult tapeworm.

Buccal capsule (cavity)
The oral cavity of roundworms; it may contain either teeth or cutting plates.

Bug
An insect of the order Hemiptera.

Bursa (pl. bursae)
A muscular copulatory structure.

Calabar swelling
A transient, subcutaneous swelling caused by the nematode *Loa loa*.

Capitulum
A collective term referring to the mouthparts of ticks and mites extending forward from the head of the arthropod.

Carrier
A host harboring and disseminating a parasite but exhibiting no clinical symptoms.

Cell mediated reaction
The effect produced by specialized T lymphocytes.

Cellular immunity
A specific response to an antigen in which lymphoid cells are the primary effectors.

Cercocystis
A modified cysticercoid larva of *Hymenolepis nana* found in the intestinal villus of the definitive host.

Cercomer
Tail-like appendage on tapeworm procercoid and cysticercoid larvae frequently possessing hooks of the hexacanth embryo.

Cestoda
A class within the phylum Platyhelminthes comprised of tapeworms.

Chagas' disease
A disease, also known as American trypanosomiasis, caused by the flagellate *Trypanosoma cruzi*.

Chigger
A mite of the family Trombiculidae.

Chitin
An insoluble polysaccharide found in the exoskeletons of arthropods.

Chromatoidal body (or bar)
A rod-shaped structure in the cytoplasm of some ameba cysts.

Cilia
Hair-like processes attached to the cell surface and used for motility through fluids.

Ciliophora
A phylum containing animals that move by means of cilia and that have two dissimilar nuclei.

Cirrus
The penis or ejaculatory duct of a flatworm.

Coenurus
A larval tapeworm comprised of numerous scolices that bud from internal germinal epithelia.

Commensalism
The association of two different species of organisms in which one partner is benefited and the other is neither benefited nor injured. A type of symbiosis in which there is no discernible damage to the host.

Conjugation
A temporary union of two ciliated protozoans for the exchange of nuclear material.

Copulatory spicules
Needlelike structures possessed by some male nematodes used during copulation.

Coracidium
Free-swimming, ciliated embryophore of tapeworms of the order Pseudophyllidea.

Costa
A striated, rod-like structure that lies just under the recurrent flagellum of certain protozoa of the order Trichomonadida.

Creeping eruption
The irritation and rash caused by the migration under the skin of non-human hookworm larvae.

Crustacea
A class in the phylum Arthropoda that is comprised of crabs, water fleas, lobsters, shrimp, and barnacles.

Cryptozoite
The exoerythrocytic stage stage in the life cycle of *Plasmodium* spp. developing in liver cells.

Ctenidium (pl. ctenidia)
Comb-like structures found on the head region of fleas. Genal combs are located just above the mouthparts. Pronotal combs are located immediately behind the head and extend posteriority on the dorsal surface.

Cutaneous larva migrans
A disease caused by the migration under the skin of humans of larvae of *Ancylostoma* spp. (frequently dog or cat hookworms) or other roundworms. Cutaneous larval migration, which also is termed creeping eruption, is marked by thin, red, papular lines of eruption on the skin.

Cuticle
The outer protective surface of helminths and arthropods.

Cyst
A general term used to describe the resistant stage in the life cycle of an organism. This stage is frequently infective to a new host.

Cysticercoid
A tapeworm larva in which the non-inverted scolex is surrounded by a fluid-filled bladder. This stage is characteristic of tapeworms belonging to the families Hymenolepididae, Dilepididae, or Anoplocephalidae.

Cysticercosis
Infection with cysticercus larvae.

Cysticercus
A tapeworm larva in which the inverted scolex is surrounded by a fluid-filled bladder. This stage is characteristic of the cyclophyllidean family Taeniidae.

Cystogenous glands
Secretory cells in the cercariae of some digenetic Trematodes that give rise to metacercarial cysts.

Cytostome
Rudimentary mouth.

Definitive host
The animal in which a parasite passes its adult existence and/or sexual reproductive phase.

Delayed hypersensitivity
Increased reactivity to a specific antigen, mediated by cells rather than antibodies, usually requiring up to 24 hours to reach maximum intensity.

Dengue (= blackwater fever)
A disease caused by a mosquito-transmitted virus.

Dermatitis
Inflammation of the skin.

Diagnostic stage
The developmental stage(s) of a pathogenic that aids in its identification.

Differential diagnosis
The clinical comparison of different diseases that exhibit similar symptoms designed to determine from which the patient is suffering.

Direct life cycle
A life cycle in which only a single host is required for the successful development of a parasite or pathogen.

Diurnal
Occurring during the daytime.

Dysentery
A form of diarrhea involving the discharge in the feces of blood and mucus.

Ecdysis
The molting or shedding of an outer layer or covering (cuticle) and the development of a new one.

Ectoparasite
A parasite that lives on the exterior surface or in the integument of a host.

Ectopic site
Abnormal or unexpected site of infection.

Endoparasite
A parasite established within the body of its host.

Edema
Swelling resulting from the abnormal accumulation of fluid in cells, tissues, or tissue spaces.

Elephantiasis
Overgrowth of the skin and subcutaneous tissue due to obstructed circulation of lymph in the lymphatic vessels caused by the roundworm *Wuchereria bancrofti*.

Embryonation
The development of a fertilized helminth embryo into a larva.

Embryophore
The shell of *Taenia* and other tapeworm eggs.

Endemic
A disease or disease agent that occurs in a human community at all times.

Endodyogeny
A special form of merogony in which two daughter cells are while still in the mother cell. This processes occurs in certain members of the phylum Apicomplexa (e.g., Toxoplasma).

Endoparasite
A parasite that lives inside the host.

Endosome
The small mass of chromatin within the nucleus, comparable to a nucleolus of metazoan cells (also termed karyosome).

Endosome
Inflammation of the intestine.

Entomology
The branch of zoology dealing with the study of insects.

Enzootic
A disease or disease agent that persists in an animal population.

Eosinophilia
A disease manifested by high levels of blood eosinophils.

Epidemic
A disease or disease agent that spreads rapidly through a human population.

Epidemiology
The study of the relationship of the various factors that determine the frequency and distribution of an infectious process or disease in a community.

Epimastigote
A flattened, spindle-shaped form of a hemoflagellate, possessing a short undulating membrane, and a kinetoplast that lies anterior to the nucleus. Epimastigotes are seen primarily in the gut (e.g., in the reduviid bug) or salivary glands (e.g., in the tsetse fly) of the vectors in the life cycle of trypanosomes.

Epizootic
A disease or disease agent that spreads rapidly through an animal population.

Espundia
A disease caused by *Leishmania braziliensis*, also known as mucocutaneous Leishmaniasis, uta, pian bois, and chiclero ulcer.

Excystation
Transformation from a cyst to a trophozoite after the cystic form has been swallowed by the host.

Exflagellation
The process whereby a sporozoan microgametocyte releases haploid flagellated microgametes that can fertilize the macrogamete and thus form a diploid zygote, such as *Plasmodium*.

Exoskeleton
A hard, chitinous structure on the outside of the body, providing support for internal organs.

Feral
Pertains to parasites that occurs in the wild as opposed to an urban sites.

Filariform larva
Infective, non-feeding, sheathed, third-staged larva of nematodes.

Flagellum (pl. flagella)
An extension of ectoplasm that serves in locomotion.

Flame cell (=protonephridium)
The terminal cell of the excretory system in platyhelminths containing a group of luminal cilia that moves fluid through the tubule.

Gamogony (=gametogony)
Formation of gametes.

Gamete
A mature sex cell.

Gametocyte
The malaria sexual cell in human blood. Gametocytes produce gametes on the mosquito's stomach.

Gametogony
The phase of the development cycle of the malarial parasite in the human in which male and female gametocytes are formed.

Gastrodermis
The tissue lining the digestive tract, as found in digenetic trematodes.

Genital atrium (=genital pore)
A area in the body into which open male and female genital ducts.

Genus (pl. genera)
A taxonomic category subordinate to Family and superior to Species.

Granuloma
A swelling composed of cells (leukocytes), fluid, and connective tissue, frequently representing a host reaction against a foreign-body.

Gravid
A pregnant individual having developing eggs, embryos, or larvae.

Ground itch
Skin penetration by non-human hookworm larvae causing localized irritation and rash.

Gynecophoral canal
Longitudinal groove on the ventral surface of male schistosomes in which the female worm lies.

Helminth
A term for worms.

Hematophagous
Bloodsucking; usually refers to the feeding habits of various insects and ticks.

Hematuria
Blood in the urine.

Hemimetabolous
A type of insect development in which there is a gradual change in the external structure as development proceeds to the adult stage.

Hemocoel
The body cavity of arthropods, typically containing blood or hemolymph.

Hemozoin
Granules, seen in erythrocytes infected with Plasmodium malariae.

Hermaphroditic
Having both male and female reproductive organs within the same individual.

Heterogonic
The term used to describe a life cycle in which free-living generations may alternate periodically with parasitic generations. Reproduction in which sexual and asexual generations alternate, as in the nematode *Strongyloides*.

Heteroxenous
Requiring more than one host to complete a life cycle.

Hexacanth embryo
A tapeworm larva having six hooklets (see onchosphere).

Holometabolous
A type of insect development in which there are distinct morphological changes as the insect develops through one or more larval stages, a pupal stage, and the imago or adult stage.

Homogonic
The term used to describe a lifestyle that is consistently either parasitic or free-living.

Homoxenous
Refers to a parasite that has a direct life cycle, or one that requires only a single host.

Host specificity
The extent to which a parasite can exist in more than one host species.

Host, definitive
The host in which a parasite attains sexual maturity.

Host, intermediate
The host in which a parasite undergoes developmental changes but does not become sexually mature.

Host, paratenic
A transfer host in which a parasite resides without developing.

Host, reservoir
Usually a non-human host in which a parasite lives and remains a source of infection but usually shows no symptoms.

Host
An animal or plant that harbors a parasite, providing the latter with some metabolic resource.

Humoral immunity
A specific host response to an antigen in which the principal effectors are circulating antibodies that immobilize or destroy the antigen.

Hydatid
A larval form of tapeworm in which numerous scolices of potential tapeworms bud from secondary cysts. This larval form is characteristic of tapeworms in the genus *Echinococcus*.

Hydatid sand
Granular material consisting of free scolices, hooklets, and daughter cysts in the fluid of hydatid cysts of Echinococcus granulosus.

Hyper-
A prefix meaning more than normal, over or above.

Hyperemia
An abnormally large amount of blood in a tissue.

Hyperplasia
An abnormally high number of cells in a tissue.

Hypersensitivity (=allergy)
A condition in which an organism is sensitized to a particular substance and manifests an abnormally strong reaction upon subsequent exposures to the substance.

Hypertrophy
An abnormal increase in the size of a tissue.

Hypodermis
The tissue that secretes the overlying cuticle.

Imago
The sexually mature adult insect.

Immunity
A specific response, cellular and/or humoral, to a foreign protein (antigen).

Immunopathology
An immune response that is damaging in itself.

Immunosuppression
Depressed immune responsiveness.

In vitro
Observable in a test tube, or other non-living system.

In vivo
Within the body.

Incidence (of infection)
The proportion of a population infected or showing disease over a given period of time.

Incidental parasite
Parasite found in a host other than its usual one.

Incubation period
The interval of time from initial infection to the onset of clinical symptoms of a disease.

Indirect life cycle
A life cycle in which more than one host is required for the parasite or pathogen to complete its development.

Infection
Invasion of the body by a parasite or pathogen.

Infective stage
The stage of a parasite that is capable of entering the host and continuing development within the host.

Infestation
The establishment of a parasite on the external surface of a host.

Inflammation
A response to a physical, chemical or biological insult causing pain, reddening, increased temperature, and swelling at the site of injury.

Instar
A stage in the life cycle of an insect, such as a larval or nymphal instar.

Intermediate host
An animal that serves as host for only the larval or sexually immature stages of a parasite. An intermediate host is required for the successful development of the parasite.

Invertebrates
Animals having no spinal column.

Juvenile
A sexually immature stage of an organism.

Kala-azar
A disease, also known as visceral Leishmaniasis or dum-dum fever, caused by *Leishmania donovani*.

Kinetoplast
A mitochondrial-like organelle characteristic of protozoa of the order Kinetoplastida.

Kinetosome
See basal body.

L.D. body (Leishman-Donovan body)
Each of the small ovoid ametabolous forms found in macrophages of the liver and spleen in patients with *Leishmania donovani* infection.

Larva (pl. larvae)
An immature or pre-adult stage in the development of certain insects or nematodes.

Laurer's canal
A canal, originating on the surface of the oviduct near the seminal receptacle in some digenetic trematodes, that may represent a vestigial vagina.

Leishmaniasis
A disease caused by members of the genus *Leishmania*.

Lumen (pl. lumina)
The central cavity of an organ.

Lycophore
The ten-hooked or decacanth larva that develops within the egg capsules of tapeworms of the subclass Cestodaria.

Lymphokine
Any of several chemical mediators, released by T cells, that react with other cells essential to the inflammatory process.

Macrogametocyte
The cell that gives rise to a macrogamete.

Malpighian tubules
Excretory organs of terrestrial insects and most acarines.

Mastigophora
A subphylum of Protozoa containing organisms that possess one or more flagella.

Maurer's dots
Aggregates in cytoplasm of erythrocytes infected with Plasmodium falciparum.

Mechanical vector
A host that is not necessary for the development of a parasite, but one that serves as a temporary refuge for transfer of the parasite to an obligatory host.

Merogony
A type of asexual reproduction in which there is nuclear replication without plasmotomy, and then two to many daughter cells (i.e., merozoites) are produced simultaneously.

Merozoite
Usually an elongate stage in the life cycle of sporozoans produced by merogony that infects host cells to undergo additional merogony or to ultimately form gametes by gamogony.

Metacercaria
The larval stage between cercaria and adult in the life cycle of many digenetic trematodes.

Metacestode
Tapeworm stage following the oncosphere, but one not yet sexually mature.

Metacystic trophozoite
A small trophozoite of Entamoeba spp. that emerges from the cyst in the intestine of the host.

Metamerism
Segmental repetition of homologous parts; in each metamere or segment there are identical structures such as muscles, neural ganglia, and nephridia.

Metamorphosis
A change of shape or structure involving the transition from one developmental stage to another. In incomplete metamorphosis the immature stages resemble adults in structure and are termed nymphs. In complete metamorphosis, the immature stages (larvae and pupa) do not resemble the adults.

Metazoa
A subkingdom of animals consisting of all multicellular animals, in which cells are differentiated to form tissue. The subkingdom includes all animals except Protozoa.

Microfilaria
The juvenile, first-stage larva of filarial nematodes.

Microgametocyte
The cell that gives rise to microgametes.

Miracidium
The ciliated larva that emerges from the egg of digenetic trematodes.

Molecular mimicry
The acquisition by a parasite of a body surface that manifests host-like molecules.

Molt (=ecdysis)
Shedding of an external covering (e.g., integument, cuticle or exoskeleton), and the replacement of a new one that accommodates for the growth and development of the pre-adult stages.

Monoecious
Both male and female sex organs in one individual (i.e., hermaphroditic).

Monoxenous
Having a single host in the life cycle.

Monozoic
Refers to tapeworms that possess only a single set of reproductive organs.

Multiple parasitism
Infection or infestation of a host by more than one species of parasite.

Multivoltine
Having a number of generations in a year.

Mutualism
A symbiotic relationship in which both symbiotic individuals benefit from each other.

Myiasis
A condition caused by infestation of the body with fly larvae.

Nagana
A disease of domestic ruminants, caused by *Trypanosoma brucei brucei, T. congolense* and *T. vivax*.

Naiad
The aquatic, pre-adult stage of an insect that has incomplete metamorphosis.

Natural immunity
The type of immunity conferred by the presence in an organism of certain naturally occurring antibodies to specific antigens.

Nymph
The terrestrial, pre-adult stage in the life cycle of an insect with incomplete metamorphosis.

Obligate parasite
A parasite that requires a host for the completion of its life cycle.

Oncosphere
The motile, six-hooked embryo (hexacanth) of a tapeworm that is contained in the egg membranes.

Oocyst
A stage in the life cycle of certain members of the phylum Apicomplexa in which the zygote secretes a resistant covering around itself. The encysted form of the ookinete which occurs on the stomach wall of *Anopheles* spp. mosquitoes infected with malaria.

Ookinete
The zygote in the life cycle of certain members of the phylum Apicomplexa following the fusion (syngamy) of macro- and microgametes. The term ookinete most often refers to the motile stage of *Plasmodium*, which is seen in the midgut of the mosquito shortly after syngamy.

Ootype
A specialized region of the flatworm oviduct that is surrounded by Mehlis' gland.

Open circulatory system
The system in which blood flows slowly through large sinuses (hemocoel) back to a dorsal, tubular heart.

Operculum
A lid-like structure at one end of the eggshell of many digenetic trematodes and some cestodes.

Oriental sore
A disease, also known as cutaneous Leishmaniasis, caused by *Leishmania tropica*.

Oviparous
Producing eggs that hatch after leaving the body of the mother.

Ovoviviparous
Producing eggs with persistent membranes through which the young escape while still within the body of the mother.

Ovum (pl. ova)
The female germ cell.

Parabasal body
The Golgi complex of the protozoan order Trichomonadida.

Parasitemia
The presence of parasites in the blood (e.g., malaria schizonts in red blood cells).

Parasitism
An association between two species in which the smaller (parasite) is physiologically dependent on the larger (host). The host may be adversely affected by the parasite.

Parasitophorous vacuole
A clear space between an intracellular parasite and the host cell cytoplasm.

Paratenic host (=transport host)
A host in which a parasite resides but does not develop and which is not physiologically essential for the completion of the life cycle.

Parenchyma
In flatworms, the mesodermal tissue filling all available body spaces.

Paroxysm
The fever-chills syndrome in malaria occurs cyclically every 36 to 72 hours depending on the species of *Plasmodium*. The onset of fever corresponds to the release of merozoites and toxic materials from infected red blood cells. The shaking chills are manifested during schizont development.

Parthenogenesis
Development of an organism from an unfertilized egg; common in insects such as aphids and in some nematodes such as *Strongyloides*.

Parthenogenic
Capable of unisexual reproduction; (i.e., without fertilization).

Pathogenic
Production of tissue changes or disease.

Pathogenicity
The ability to produce pathogenic changes.

Periodicity
Recurring at regular time periods.

Peritoneum
A thin membrane of mesodermal origin that lines the body cavity of vertebrates (and some other higher metazoans) and supports the viscera or organs of the body cavity.

Peritrophic membrane
A covering that forms around the blood meal of a hematophagous insect; digestion usually takes place within the membrane; the membrane sometimes serves as a barrier to a parasitic agent moving from the blood to the tissues of the host.

Phasmid
Sensory pit located on the posterior part of nematodes of the class Secernentea.

Phoresis
A form of commensalism in which one organism is mechanically transported by another. The relationship is nonobligatory.

Plasma cell
An effector B cell that secretes into the circulation antibodies of the same specificity as its cell surface receptors.

Plerocercoid
A larva or metacestode developing from a procercoid in the life cycle of pseudophyllidean tapeworms. This larval stage, which is infective to humans, develops in fresh-water fish.

Polar filament
A coiled filament or tubule in the spore of certain protozoa (e.g., Myxozoa and Microspora). The filament is extruded when the spore is ingested by a host.

Polyembryony
A process in which a zygote gives rise to more than one embryo.

Polyzoic
Consisting of more than one zooid or animal or proglottid.

Procercoid
The larval stage that develops from the coracidium of the pseudophyllidean tapeworm *Diphyllobothrium latum*. The procercoid is found in the body of a freshwater crustacean that serves as the first intermediate host.

Proglottid
One of the segments of a tapeworm strobila containing male and female reproductive organs when mature.

Promastigote
A morphological form of flagellates in which the kinetoplast is located at the anterior end of the organism, and there is no undulating membrane.

Prophylaxis
Procedures that are carried out to prevent the transmission of disease.

Protandry
The condition in which the male gonad matures before the female gonad, or a single gonad first produces sperm and then ova.

Protoscolex (protoscolices)
The immature scolex found in coenurus and hydatid larvae of tapeworms.

Pruritus
Intense itching.

Pseudocoelom
A body cavity of a metazoan that is b;t completely lined with mesoderm.

Pseudocyst
A cluster of am2‘tigotes of *Trypanosoma cruzi* in a muscle fiber. A cyst-like structure filled with bradyzoites of *Toxoplasma gondii*.

Pseudopod
A protoplasmic extension of amebae which allows them to move and engulf food.

Pupa (pl. pupae)
The encased resting stage (e.g., cocoon) between the larva and adult (imago) stages of certain insects.

Quarantine
Limitation in the freedom of movement of humans or animals in order to contain the spread of a disease.

Radula
Rasping structure inside the mouth of mollusks of the classes Gastropoda, Cephalopoda, Amphineura, and Scaphopoda; the pattern of teeth on the radula is characteristic of a species.

Rectal prolapse
Weakening of the rectal musculature resulting in a “falling down” of the rectum; occasionally seen in heavy whipworm infections, particularly in children.

Redia (pl. rediae)
A larval form of digenetic flukes that arises asexually from within a sporocyst or a primary redia.

Relapse
The recurrence of symptoms of a disease after an abatement of weeks or months.

Reservoir host
An animal that harbors a species of parasite from which humans may become infected.

Resistance
The ability of an organism to withstand infection or the successful establishment of a parasite.

Retrofection
A process of infection in which a parasite leaves the body of the host and then returns almost immediately, frequently by penetrating the skin (e.g. the nematode Strongyloides).

Rhabditiform larva
Non-infective, feeding, first-stage larva of some nematodes.

Romano’s sign
Early symptoms of Chagas’ disease, consisting of unilateral, periorbital swelling (edema) and conjunctivitis.

Rostellum
The rounded protuberance on the apex of the scolex of some tapeworms that usually bears a circular row hooks.

Sarcodina
A subphylum of Protozoa containing amebae that possess pseudopodia for movement.

Schiffner’s dots
Fine granules distributed throughout erythrocytes infected with *Plasmodium vivax*.

Schizogony

A form of asexual reproduction characterized by multiple nuclear divisions followed by cytoplasmic divisions and the resulting formation of a large number of daughter cells.

Schizont

A multinucleated cell undergoing schizogony prior to cytoplasmic division.

Scolex (pl. scolices)

Anterior end of a tapeworm that attaches to the host intestinal wall. The scolex may be comprised of suckers and hooks.

Scrub typhus

Rickettsial disease transmitted by chigger mites.

Scutum

A chitinous shield or plate covering the dorsal surface of hard ticks.

Serology

The study of antibody-antigen reactions in vitro, using host serum for study.

Serum

The fluid part of vertebrate blood after the fibrin has been removed.

Shell (yolk) glands

Clusters of cells that synthesize substances essential to eggshell formation.

Sleeping sickness

A disease caused by *Trypanosoma rhodesiense* or *T. gambiense* in Africa, or any one of several arboviruses transmitted by mosquitoes and causing encephalitis.

Sparganosis (or sparganum)

Infection with a plerocercoid or larval stage of a tapeworm belonging to the order Pseudophyllidea.

Spicule (=copulatory spicule)

A sclerotized structure of male nematodes used during copulation.

Spore

A resistant stage that is formed internally by the mother cell.

Sporoblasts

Cells that divide into sporozoites.

Sporocyst

The larval stage of digenetic trematodes developing in the first intermediate host.

Sporogony

Sexual reproduction of Apicomplexa. Production of spores and sporozoites.

Sporozoite

The infective or transfer stage of *Plasmodium* and other members of the phylum Apicomplexa

Stenoxenous

Having a narrow host range.

Strobila

In tapeworms, a chain of proglottids or segments formed by budding.

Superparasitism

The infection or infestation of a host by more individuals of a single species of parasite than the host can support.

Swimmer's itch (=cercarial dermatitis)
A skin hypersensitivity reaction in response to the penetration or attempted penetration by non-human schistosome cercariae.

Sylvatic
Refers to an animal (or disease) cycle that exists in the wild.

Syngamy
The union of gametes.

Symbiosis
The intimate association of two different species of organisms exhibiting metabolic dependence by their relationship.

Syzygy
End-to-end joining of two or more gamonts of members of the subclasses Gregarina and Coccidia.

T cell
A specialized lymphocyte, processed through the thymus, that elicits cell-mediated reactions.

Tachyzoite
A form of merozoite in Toxoplasma, found in parasitophorous vacuoles of vertebrate hosts. Rapidly growing meront or zoites characteristic of the early stage of infection with Toxoplasma and related organisms of the phylum Apicomplexa. Rapidly growing intracellular trophozoites of *Toxoplasma gondii*.

Temporary host
A host on which an arthropod (adult or larval form) reside temporarily in order to feed on blood or tissue.

Temporary parasite
A parasite that visits a host at intervals and only for relatively short period; examples are mosquitoes and ticks.

Transport host
See paratenic host.

Trophozoite
The motile, growing, and feeding (vegetative) stage of a protozoan stage of a protozoan that maintains the population within the host.

Trypomastigote
A hemoflagellate form with a kinetoplast located posterior to the nucleus and an elongated undulating membrane extending along the entire body. This form is seen in the blood of humans with trypanosomiasis and as the infective stage in the insect vectors.

Tularemia
A bacterial disease of humans transmitted by tabanid flies and for which rabbits often serve as sylvatic reservoir hosts.

Undulating membrane
That portion of a plasma membrane or cytoplasm of a flagellate that is drawn away from the cell like a fin along the outer edge of the. The membrane moves in a wave-like pattern.

Uniramous
Having one branch or ramous.

Univoltine
Having one generation per year.

Vagina
An organ that joins the oviduct and carries sperm from the genital atrium to the oviduct.

Vector
Any organism that actively transmits a disease-producing organism from an infected to a non-infected individual. A mechanical vector is one in which the parasite neither multiplies nor develops (passive transmission). A biological vector is one in which the parasite either multiplies or develops;

Vermicle
The infective stage, analogous to sporozoite, of Babesia spp. from ticks.

Vertical transmission
Transmission of a parasite from one generation to the next through the egg or in utero.

Virulence
The ability of a parasite to produce pathogenic effects or to invade the host and become established.

Viscera
Any of the large organs in the body cavities of vertebrates.

Visceral larva migrans
Migration of second-stage larvae of nematodes in the internal organs of unnatural hosts.

Vitellaria
Clusters of cells that synthesize substances essential to eggshell formation.

Winterbottom's sign
Symptom of African sleeping sickness characterized by enlarged, sensitive cervical lymph nodes.

Yaws
A fly-transmitted spirochete disease.

Yellow fever
A viral disease transmitted by the mosquito *Aedes aegypti*.

Zoonosis (pl. zoonoses)
A disease of animals that can be transmitted to humans.

Zoonotic agent
An organism that causes a zoonosis.

A

Acetabulum 37
Alae 71, 121
Allergy 121, 129
Alveoli 78, 80, 81, 121
Amastigote 7, 9, 12-16, 130
Amphids 71, 121
Amphids 71
Anaphylactic shock 68, 111, 121
Anisogamete 5, 6
Anorexia 11, 21, 26, 33, 42, 47, 70, 121
Anterior station 9, 121
Anthrax 102, 116, 122
Antibody 136
Antigen 122, 123, 125, 129, 132, 136
Apicomplexa 5, 6, 25, 34, 122, 126, 133, 137
Apolysis 121, 122
Arbovirus 122, 136
Arthropoda 122, 124
Atrium 37, 122, 127, 138
Autoinfection 61, 62, 70, 77, 122

B

B cell 122, 134
Basal body 7, 122, 123, 130, 133
Biological vector 2, 7, 9, 16, 18, 93, 105, 115, 116, 123, 138
Biramous 123
Blackwater fever 123, 125
Bradyzoites 27, 28, 123, 135
Brood capsule 58, 64, 65, 123
Buccal capsule (Buccal cavity) 15, 37, 71, 78, 79, 123
Bursa 107, 123

C

Calabar swelling 97, 123
Capitulum 103, 104, 110, 123
Carrier 23, 123
Cercocystis 58, 123
Cercomer 57, 123
Cestoda 123, 131
Chigger 102, 105, 124, 136
Chitin 5, 100, 122, 124, 127, 136
Cilia 5, 24, 37, 38, 44, 56, 57, 58, 124, 127, 132
Ciliophora 5, 124
Cirrus 37, 124
Coenurus 58, 63, 135, 124
Commensalism 1, 124, 134
Conjugation 5, 6, 124
Copulatory spicules 71, 124
Coracidium 56-58, 124, 134
Costa 76, 87, 88, 124
Creeping eruption 81, 82, 124, 125
Crustacea 47, 89, 91, 92, 100, 122, 124, 134
Cutaneous larva migrans 81, 82, 90, 119, 125
Cuticle 2, 37, 71, 72, 125, 126, 129, 132
Cyst 2, 6, 7, 17, 18, 20, 22-28, 33-35, 58, 64, 65, 67, 68, 78, 123, 125, 127, 131, 135
Cysticercoid 57, 58, 69, 70, 113, 123, 125
Cysticercosis 60-62, 64, 65, 125
Cysticercus 58, 60, 62, 125
Cystogenous glands 40, 125

D

Definitive host 2, 9, 37, 40, 41, 43, 51, 56-58, 69, 72, 74, 76, 89, 91, 96, 97, 121, 123, 125
Dengue 4, 105, 116, 125
Dermatitis 54, 98, 105, 107, 108, 111, 113, 125, 137
Direct life cycle 2, 56, 72, 125, 128
Diurnal 93, 116, 125
Dysentery 21, 24, 26, 54, 116, 125

E

Ecdysis 126, 132
Ectoparasite 117, 126
Edema 11, 13, 17, 42, 51, 54, 60, 76, 82, 95, 126, 136
Elephantiasis 93, 95, 96, 126

Embryonation 41, 58, 73, 126
Embryophore 56, 57, 124, 126
Endemic 12, 14, 47, 58, 87, 96, 98, 102, 107, 113, 126
Eosinophilia 3, 86, 87, 88, 126
Epidemic 102, 112, 126
Epidemiology 10, 126
Epimastigote 7, 13, 126
Epizootic 102, 127
Espundia 10, 17
Excystation 18, 20, 23, 24, 41, 45, 47, 127
Exoskeleton 5, 100, 122, 127, 132

F

Flagella 5, 7, 18, 21, 22, 127
Flagellum 7, 122, 123, 124, 127

G

Gamete 6
Gametogony 29, 122, 127
Gamogony 25, 26, 28, 127, 131
Genital atrium 37, 127, 138
Genital pore 37, 71, 127
Gravid 55, 58, 60, 61, 68, 69, 84, 91, 105, 118, 121, 122, 128
Ground itch 80, 128

H

Helminth 47, 126, 128
Hematophagous 128, 134
Hematuria 102, 128
Hemocoel 33, 69, 89, 91, 93, 128, 133
Hemozoin 33, 128
Hermaphroditic 37, 128, 132
Heterogonic 78, 128
Heteroxenous 25, 72, 128
Hexacanth embryo 123, 128
Homogonic 77, 128
Host 1-4, 6, 7, 9-16, 18, 20, 25-28, 31, 34, 36-38, 40, 41, 43-45, 47, 49, 51, 55-58, 60, 62-65, 69, 70, 72-74, 76-81, 84, 86, 89-91, 93, 96-99, 105, 107, 109-113, 115, 117-138
Host specificity 2, 128
Hydatid 55, 58, 63-68, 129, 135
Hydatid sand 65, 129
Hyperplasia 3, 17, 43, 46, 94, 129
Hypersensitivity 121, 125, 129, 137
Hypertrophy 3, 95, 129

I

Imago 128, 129, 135
Immunity 2, 28, 29, 86, 121, 123, 129, 132
Immunopathology 28, 129
Incidence of infection 58
Incidental host 121
Incubation period 16, 17, 34, 130
Indirect life cycle 56, 130
Infection 1, 3, 4, 9, 10, 12, 14-18, 20-22, 24-26, 28, 29, 33-35, 41, 44-48, 51, 54, 57-61, 63, 65, 69, 74-76, 78-88, 91, 93, 94, 96-98, 106, 121, 125, 126, 129, 130, 132, 135-137
Infective stage 3, 25, 71, 86, 91, 98, 130, 138
Infestation 1, 111, 113, 115, 130, 132, 137
Inflammation 3, 9, 12, 17, 41, 44, 54, 82, 84, 88, 99, 120, 125, 126,130
Intermediate host 2, 3, 27, 31, 37, 38, 40, 43, 47, 49, 51, 56-58, 60, 63, 65, 70, 89, 90, 97, 113, 130, 134, 137

K

Kala-azar 10, 14, 16, 17, 102, 130
Kinetoplast 7, 126, 130, 135, 138

L

Larva 37, 56, 57, 60, 81, 82, 85, 87-90, 92, 103, 104, 118, 119, 123, 125-128, 130-132, 134-136, 138
Leishmaniasis 15-17, 102, 115, 117, 127, 130, 131, 133
Lumen 23, 28, 69, 74, 77, 84, 89, 94, 100, 131

M

Malpighian tubules 98, 131
Mastigophora 5, 131
Mechanical vector 2, 116, 131, 138
Merogony 126, 131
Merozoite 131, 137
Metacestode 131, 134
Metamorphosis 121, 131, 132
Microfilaria 95, 97, 132
Microgametocyte 127, 132
Miracidium 37, 38, 44, 132
Molt 77, 79, 84, 91, 97, 105, 106, 109, 132
Monoecious 71, 132
Monoxenous 25, 72, 132
Multiple parasitism 1, 132
Mutualism 1, 132
Myiasis 115-117, 119, 120, 132

N

Nymph 104, 132

O

Oncosphere 56, 57, 60, 63, 131, 133
Oocyst 6, 7, 25, 28, 133
Ootype 37, 133
Operculum 37, 58, 133
Oriental sore 16,17, 102, 133
Oviparous 72, 133
Ovoviviparous 72, 75, 91, 133

P

Parasitemia 9, 13, 100, 133
Parasitism 1, 132, 133
Parasitophorous vacuole 133
Paratenic host 2, 133, 137
Parenchyma 43, 45, 121, 133
Paroxysm 33, 133
Parthenogenic 71, 134
Pathogenic 6, 18, 20, 23, 100, 101, 125, 134, 138
Pathogenicity 1, 4, 134
Periodicity 93, 134
Peritoneum 68, 85, 134
Phoresy 1
Plasma cell 34, 134
Plerocercoid 57, 58, 134, 136
Procercoid 57, 58, 60, 123, 134
Proglottid 55, 61, 134
Promastigote 7, 13, 135
Protandry 55, 135
Pruritus 51, 70, 91, 105, 107, 111, 115, 116, 135
Pseudocyst 28, 135
Pupa 132, 135

R

Redia 38, 47, 135
Reservoir host 107, 135
Rostellum 55, 63, 68, 69, 70, 136

S

Sarcodina 5, 20, 136
Schizogony 6, 25, 26, 28, 29, 33, 34, 122, 136
Schizont 6, 29, 134, 136
Scolex 55, 57, 58, 60, 61, 63, 68-70, 123, 125, 135, 136
Scrub typhus 102, 105, 136
Scutum 108, 136
Serum 81, 122, 136
Sleeping sickness 9, 11, 117, 136, 138
Sparganosis 60, 61, 136
Spore 5, 25, 134, 136
sporocyst 26, 38, 41, 42, 44, 135, 137
Sporogony 6, 7, 25, 26, 29, 31, 122, 137
Sporozoite 25, 137, 138
Strobila 55, 60, 61, 69, 122, 134, 137
Superparasitism 1, 137
Sylvatic 137, 138
Symbiosis 124, 137
Symbiotic 1, 20

T

Trophozoite 6, 17, 18, 20, 21, 23, 29, 34, 127, 131, 138
Trypomastigote 7, 9, 12, 13, 138
Tularemia 102, 116, 138

U

Undulating membrane 7, 126, 135, 138

V

Vector 2, 7, 9, 12, 16, 93, 97, 102, 105, 107, 112, 116, 117, 121, 123, 131, 138
Virulence 20, 138
Viscera 65, 121, 134, 138

Y

Yaws 116, 138
Yellow fever 4, 105, 116, 139

Z

Zoonosis 18, 139

LANDES
BIOSCIENCE

Vademecum

Table of contents

1. Interspecific Interactions
2. Major Groups of Parasites of Humans
3. Sarcodina
4. Apicomplexa: Sporozoa and Piroplasma
5. Digenetic Trematodes: Flukes
6. Cestodes
7. General Morphology of Parasitic Nematodes
8. Arthropods

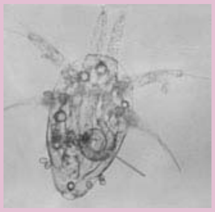

The Vademecum series includes subjects generally not covered in other handbook series, especially many technology-driven topics that reflect the increasing influence of technology in clinical medicine.

The name chosen for this comprehensive medical handbook series is Vademecum, a Latin word that roughly means "to carry along". In the Middle Ages, traveling clerics carried pocket-sized books, excerpts of the carefully transcribed canons, known as Vademecum. In the 19th century a medical publisher in Germany, Samuel Karger, called a series of portable medical books Vademecum.

The Landes Bioscience Vademecum books are intended to be used both in the training of physicians and the care of patients, by medical students, medical house staff and practicing physicians. We hope you will find them a valuable resource.

All titles available at
www.landesbioscience.com

ISBN 1-57059-679-4

DIAGNOSING MEDICAL PARASITES:

A Public Health Officers Guide to Assisting Laboratory and Medical Officers

AIR EDUCATION & TRAINING COMMAND

Compiled and edited by:
Michael J. Cuomo, Maj, USAF, BSC (PH)
Lawrence B. Noel, Maj, USAF, BSC (PH)
Daryl B. White, Capt., USAF, BSC (Lab)

Contents:

Chapter 01. The Ameba

Chapter 02. The Ciliates, Coccidia, and Microsporidia

Chapter 03. The Flagellates

Chapter 04. The Cestodes (Tapeworms)

Chapter 05. The Nematodes (Roundworms)

Chapter 06. The Trematodes (Flukes)

Chapter 07. Unusual Tissue Dwelling Nematodes

Chapter 08. Larval Cestodes and Nematodes which Infect Man

Chapter 09. Malaria

Chapter 10. The Blood Nematodes

Chapter 11. Babesia, Trypanosomes, and Leishmania

Chapter 12. Arthropod Vectors

Chapter 13. Artifacts and Confounders

Chapter 1.

The Ameba

1.1 An Introduction to Parasitology

Parasitology is the study of parasites and as such does not include bacterial, fungal or viral parasites. Human parasites are separated into intestinal and blood-borne parasites. For a parasite to be defined as intestinal it must have an intestinal life-cycle stage, though it may have life-cycle stages in the heart, circulation, lung, tissue, other animals or the environment.

Parasites found in the intestines can be categorized into two groups: Protozoa and Helminths.

Protozoa are single celled organisms. There are four classes of Protozoa commonly found in concentrated fecal samples. These are differentiated by the method of motility. Protozoa include *Entamoeba*, *Giardia*, *Trichomonas*, *Cryptosporidium*, *Isospora*, *Pneumocystis* and *Balantidium*. There are two diagnostic life-cycle stages commonly seen in parasites - the cyst and the adult trophozoite stage. The trophozoite stage is analyzed directly on a slide without concentration. Cysts require concentration. The key diagnostic factor is that Protozoan cysts are typically 5-30µm (µm = microns or micrometers) in diameter, and as such are smaller than most Helminth eggs. Due to the size they are particularly difficult to see under the microscope if the sample clarity is bad.

The medically important Helminths are nematodes (roundworms), cestodes (tapeworms) and trematodes (flukes). Genera include: *Fasciola*, *Schistosoma*, *Ascaris*, Hookworm, *Trichuris*, *Taenia* and *Enterobius*. The normal stage for examination is the egg stage, although larvae may develop in some organisms (Strongyloides); the diameter of the eggs range from 30µm - 150µm.

The other major grouping of parasites is known as blood-borne parasites which are transmitted through an arthropod vector. By far the most important arthropod for transmitting parasitic infections is the mosquito. Mosquitoes are known to carry malaria and filarial nematodes. Different types of biting flies transmit African trypanosomiasis, leishmaniasis and several kinds of filariasis.

Most protozoan and helminthic infections that are transmitted by arthropods can readily be diagnosed, on clinical grounds alone, but are usually identified by fairly simple techniques designed to present the presence of the causative parasite by microscopy. Sophisticated techniques are also being employed including highly sensitive and specific simple monoclonal antibody tests, DNA probes, and PCR primers.

1.2 Infections Acquired Through the Gastrointestinal Tract

Many of the infections of the gastrointestinal tract (GI) are caused by parasites that are cosmopolitan in distribution. Protozoa can be directly infectious for man when they are passed in the feces into the environment, but helminths require a period of maturation while in the soil, where they become infectious. Others such as *Taenia saginata* require the involvement of an intermediate host during their life cycle.

Infections of the GI tract account for a high proportion of deaths in infants where the standards of hygiene and nutrition are low.

Fecal-oral transmission of the pathogens is the most common mode of GI infections, whereby water, food and hands become contaminated with fecal material which then come in contact with the mouth.

A number of GI infections can reach epidemic proportion, protozoal pathogen *Cryptosporidium parvum*, has been known to cause the severe water-borne epidemics, even in first-world countries such as the United States and the UK. Other infections such as amebiasis or enterobiasis can be more localized, infecting households or institutions.

Some of the rarer, protozoal infections such as the microsporidia are only now being understood as they are appearing as concomitant infections in people with depressed immune responsiveness, e.g. AIDS.

The Ameba

Ameba (or Amoeba) are characterized by possessing clear protoplasm which form pseudopodia. These pseudopodia are the means by which these organisms move and engulf bacteria and red blood cells for feeding purposes. The most common amebas seen in the intestinal tract are E*ntamoeba histolytica/dispar, Entamoeba coli, Entamoeba hartmanni, Endolimax nana* and *Iodamoeba bütschlii*. All but *Entamoeba histolytica* are thought to be non-pathogenic. The cysts can be identified in an ethyl acetate concentrate by the addition of iodine to reveal the characteristic inclusions and also by measuring the cyst using an eyepiece graticule. The trophozoites can be seen in a fresh saline preparation of the stool although accurate identification is on a permanently stained fecal smear.

Entamoeba histolytica

Introduction

There are a large number of species of ameba which parasitize the human intestinal tract. Of these, *Entamoeba histolytica / dispar* is the only species found to be associated with intestinal disease. Although many people harbor this organism world wide, only about 10% develop clinically invasive disease, thus the parasite has been shown to present as two very differing clinical presentations.

1. The commensal or non-invasive luminal form where the parasite causes no signs or symptoms of disease.
2. The pathogenic or invasive form where the parasite invades the intestinal mucosa and produces dysentery or amebiasis and may give rise to extra-intestinal lesions via the blood, mainly to the liver.

Sargeaunt and Williams (1978) conclusively proved that invasive and non-invasive strains of *E. histolytica* could be differentiated by isoenzyme electrophoresis and the application of molecular biology has confirmed the presence of two distinct species with the same morphological features. The pathogenic or invasive species has retained the name *E. histolytica* and the non-pathogenic, non-invasive species has been named *E. dispar.*

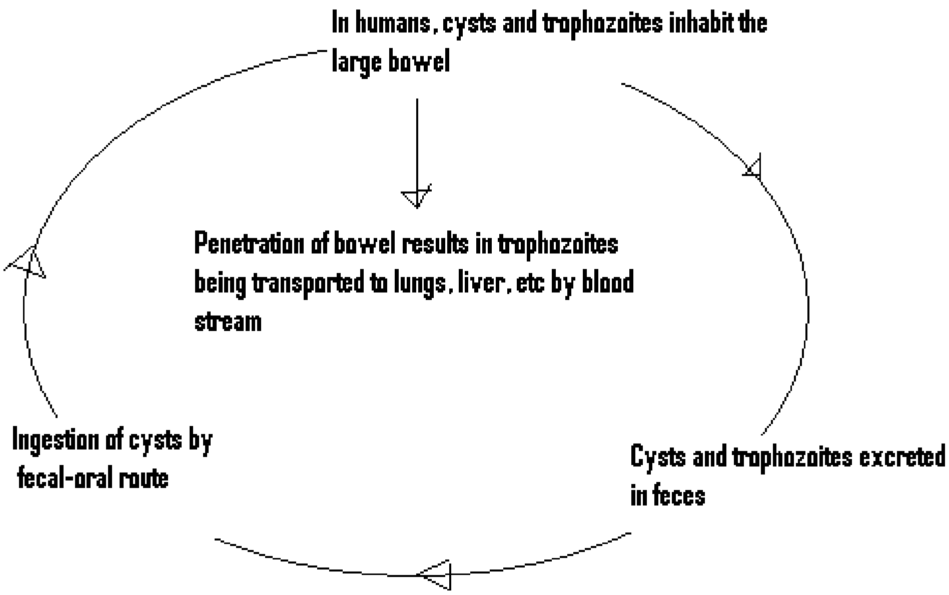
In humans, cysts and trophozoites inhabit the large bowel
Penetration of bowel results in trophozoites being transported to lungs, liver, etc by blood stream
Cysts and trophozoites excreted in feces
Ingestion of cysts by fecal-oral route

Illustration 1-1. Life Cycle of *Entamoeba histolytica:* A protozoan in which its life cycle consists of two stages; cysts and trophozoites (Cuomo)

Morphology of Trophozoites

The trophozoites of *E. histolytica / dispar* recovered from dysenteric stools exhibit ingested red blood cells and clear pseudopodia. Those of *E. dispar* will have no ingested red blood cells. They can be up to 60µm in diameter and motility is rapid and unidirectional. On a permanently stained fecal smear e.g. Trichrome or Iron hematoxylin, the morphological features are more visible. When using Trichrome stain nuclei, chromidial bars, chromatin, red cells and bacteria stain red cytoplasm stains blue-green and background and yeasts stain green. The presence of a small centrally placed karyosome is clearly visible. With Iron hematoxylin, nuclear chromatin and the karyosome will be stained immensely black. The remainder will be varying shades of grey/black.

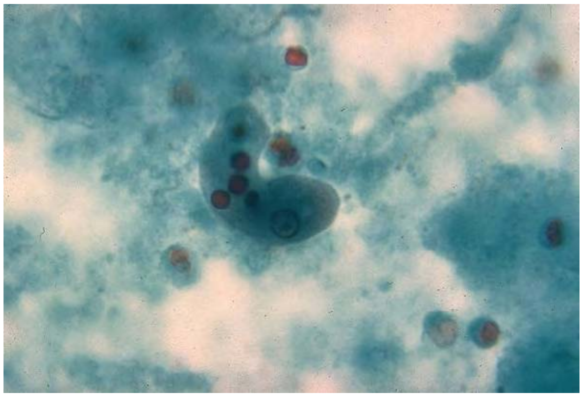

Image 1-1. *Entamoeba histolytica* trophozoites (SOURCE: PHIL 336 -CDC / Dr. N.J. Wheeler, Jr.)

Morphology of Cysts

Cysts of *E. histolytica / dispar* are 10-15µm in diameter and contain one to four nuclei. Chromatoid bodies are usually present in young cysts as elongated bars with bluntly rounded ends. Glycogen is usually diffuse, but in young cysts it is often present as a concentrated mass, staining reddish brown with iodine.

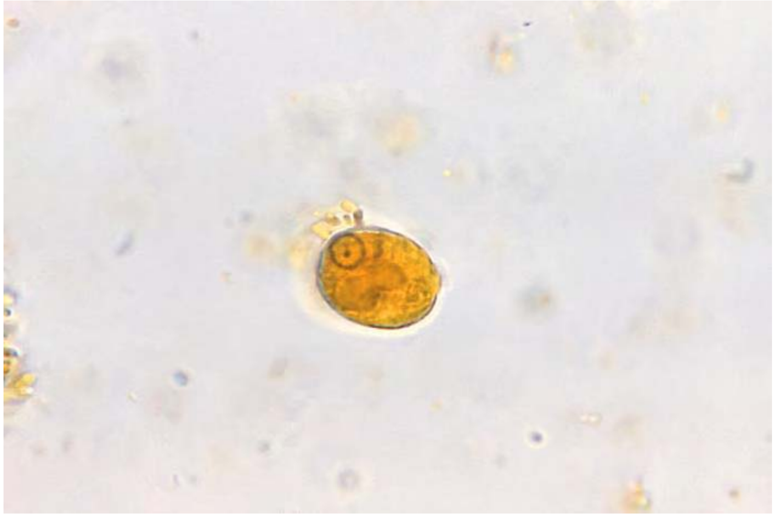

Image 1-2. *Entamoeba histolytica* cyst. (SOURCE: PHIL 531 - CDC/ Dr. L.L.A. Moore, Jr.)

Clinical Disease

Amebiasis is an infection usually caused by the pathogenic *Entamoeba histolytica / dispar*, and is commonly an infection of the colon. It has a world wide distribution where environmental sanitation is poor. The parasite may behave as a commensal (causing no harm to the host) or it may act as a parasite (harming the host). It is a disease of human beings, although some monkeys can become infected and the infection is then transmissible to humans.

Intestinal Disease

Patients with intestinal disease may exhibit a number of symptoms including profuse diarrhea with blood and mucus, fever and dehydration. Amebic ulcers may develop in the large colon and can also be found in the rectal area. The ulcers are usually "flask shaped" with a small opening on the mucosal surface and a larger area below the surface. **Image 1-3.** illustrates *E. histolytica* trophozoites in the intestine, resulting in amebiasis.

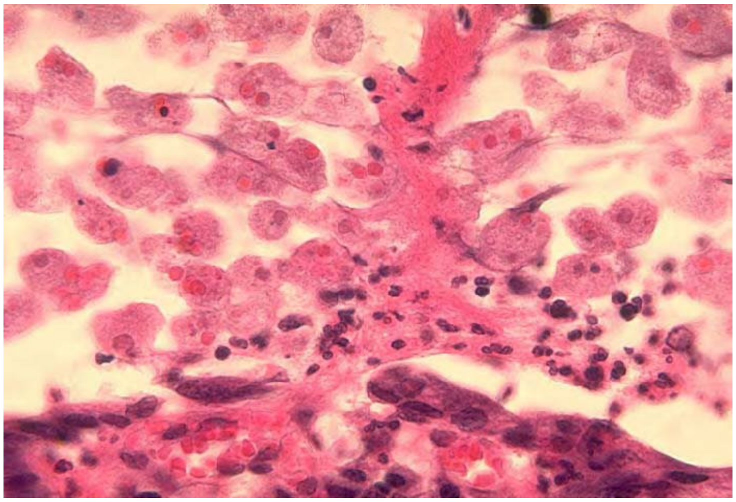

Image 1-3. *Entamoeba histolytica*, intestine. (SOURCE: PHIL 623 - CDC)

Hepatic Disease

Trophozoites are transported from the intestine to the liver and liver disease is characterized with abdominal pain, fever, hepatomegaly and tenderness. If the abscess ruptures, there is spreading to the brain, pericardium and other sites. If left untouched the abscess will grow normally until it reaches a surface where it can discharge, e.g. the skin, the peritoneum, the pleural cavity or the pericardium. The stretching of the liver is presumably the main source of the pain.

Laboratory Diagnosis

Laboratory diagnosis is made by finding the characteristic cysts in an iodine stained, formol-ether concentration method or by detecting the characteristic trophozoites in a wet preparation or a permanent stained preparation.

Microscopy

Where amebic dysentery is suspected, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed. On cooling the ameba stop moving which then become very difficult to identify. Direct microscopy should be done by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of motile trophozoites of *Entamoeba histolytica / dispar* and can also provide information on the content of the stool (i.e., the presence of leucocytes and red blood cells). On search e.g. primarily for cysts, not for ameba, several stool samples are required to be examined, by direct microscopy and a sensitive concentration technique. Three negative stool samples are required before it can be accepted that there is no amebic infection. Microscopic examination of an amebic abscess aspirate e.g. in the liver or lungs, may reveal hematophagous trophozoites. It must be examined immediately by mixing a drop of warm saline with some aspirated pus on a microscope slide.

Serology

If visceral or hepatic amebiasis is suspected, serological tests should be done as microscopic methods do not always reveal the characteristic trophozoites. The tests of choice are indirect fluorescent antibody test (IFAT), counter immunoelectrophoresis (CIEP) enzyme linked immunosorbent assay (ELISA), and Rapid Antigen testing by enzyme immunoassay.

The search for *E. histolytica* / *dispar* is mainly carried out because there is a natural concern to ensure that patients, even in the absence of symptoms, are not harboring parasites that may lead to serious complications later on.

Entamoeba coli

Introduction

Entamoeba coli are a non-pathogenic ameba with world wide distribution. Its life cycle is similar to that of *E. histolytica* but it does not have an invasive stage and does not ingest red blood cells.

Morphology of Trophozoite

The trophozoite is larger than that of *E. histolytica* ranging from 15-50μm in diameter. It exhibits blunt pseudopodia with sluggish movement. A permanently stained preparation shows a nucleus with a moderately large eccentric karyosome with the chromatin clumped on the nuclear membrane. The cytoplasm appears granular containing vacuoles with ingested bacteria and other food particles.

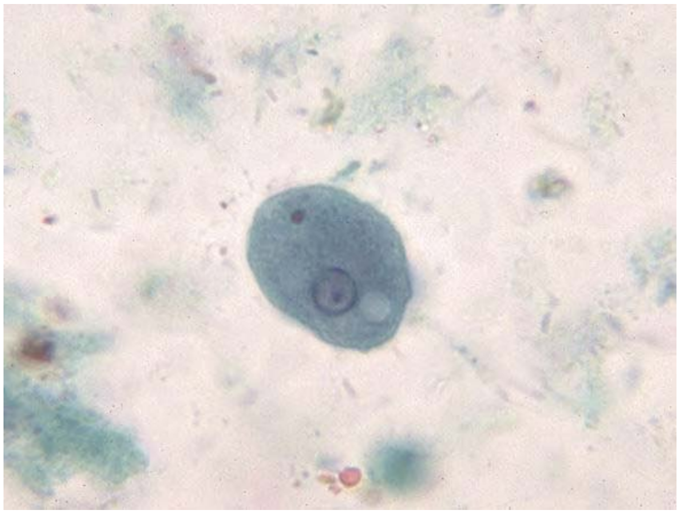

Image 1-4. *Entamoeba coli* trophozoite with ingested bacteria. (SOURCE: PHIL 605 - CDC)

Morphology of Cysts

Cysts of *E. coli* are 15-30μm in diameter and contain one to eight nuclei with irregular peripheral chromatin: karyosomes not central. Chromatoid bodies are not frequently seen but when present they are usually splinter-like with pointed ends. Glycogen is usually diffuse but in young cysts is occasionally found as a well-defined mass, which stains reddish brown with iodine.

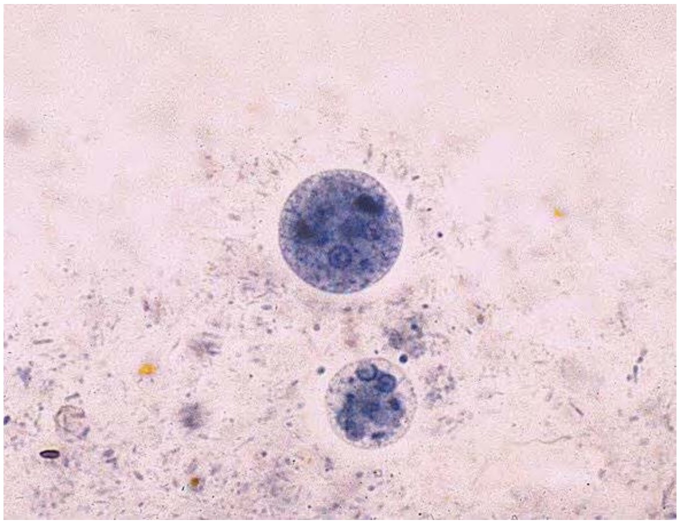

Image 1-5. *Entamoeba coli* (larger) and *Entamoeba histolytica* (smaller) cysts. (SOURCE: PHIL 442 - CDC/Dr. Mae Melvin)

Laboratory Diagnosis

Laboratory diagnosis is made by finding the characteristic cysts in an iodine stained, formol-ether concentration method or by detecting the characteristic trophozoites in a wet preparation or a permanent stained preparation.

Microscopy

Where amebic dysentery is suspected, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed. On cooling the ameba stop moving which then become very difficult to identify. Direct microscopy should be done by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of motile trophozoites of *Entamoeba coli* and can also provide information on the content of the stool (i.e., the presence of leucocytes and red blood cells). On search e.g. primarily for cysts, not for ameba, several stool samples are required to be examined, by direct microscopy and a sensitive concentration technique. Three negative stool samples are required before it can be accepted that there is no amebic infection. Microscopic examination of an amebic abscess aspirate (e.g. in the liver or lungs), may reveal hematophagous trophozoites. It must be examined immediately by mixing a drop of warm saline with some aspirated pus on a microscope slide.

Entamoeba hartmanni

Introduction

Entamoeba hartmanni is a non-pathogenic amoeba with world wide distribution. Its life cycle is similar to that of *E. histolytica* but it does not have an invasive stage and does not ingest red blood cells.

Morphology of Trophozoites

Morphology of the trophozoites is similar to those of *E. histolytica / dispar* but they do not contain ingested red blood cells and the motility is less rapid.

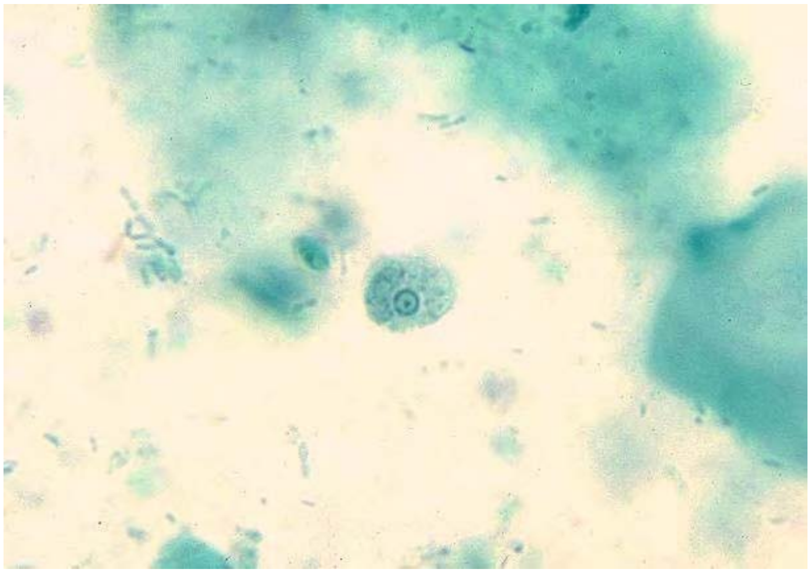

Image 1-6. *Entamoeba hartmanni* trophozoite (SOURCE: PHIL 528 - CDC/Dr. L.L.A. Moore, Jr.)

Morphology of Cysts

Cysts of *E. hartmanni* 7-9μm in diameter and contain one to four nuclei. Chromatoid bodies are usually present in young cysts as elongated bars with bluntly rounded ends. Glycogen is usually diffuse, but in young cysts it is often present as a concentrated mass, staining reddish brown with iodine.

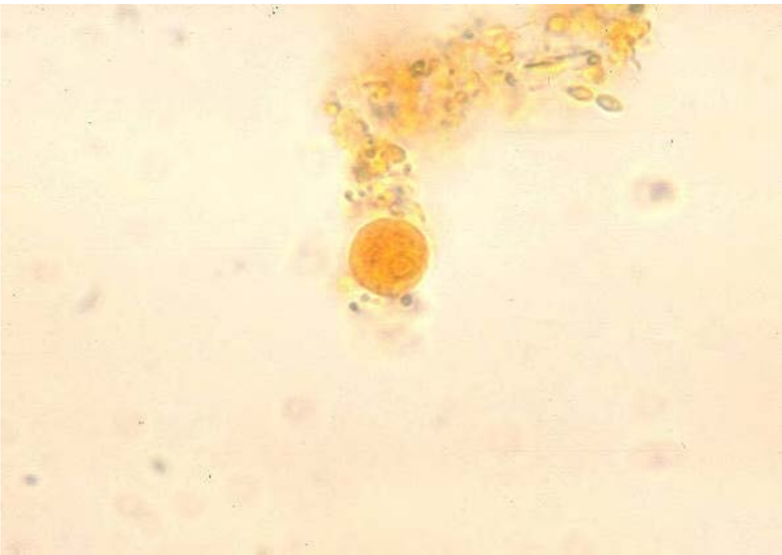

Image 1-7. *Entamoeba hartmanni* cyst (SOURCE: PHIL 533 - CDC/Dr. L.L.A. Moore, Jr.)

Laboratory Diagnosis

Laboratory diagnosis is made by finding the characteristic cysts in an iodine stained, formol-ether concentration method or by detecting the characteristic trophozoites in a wet preparation or a permanent stained preparation.

Microscopy

Where amebic dysentery is suspected, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed. On cooling the ameba stop moving which then become very difficult to identify. Direct microscopy should be done by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of motile trophozoites of *Entamoeba hartmanni* and can also provide information on the content of the stool (i.e., the presence of leucocytes and red blood cells). On search e.g. primarily for cysts, not for ameba, several stool samples are required to be examined, by direct microscopy and a sensitive concentration technique. Three negative stool samples are required before it can be accepted that there is no amebic infection. Microscopic examination of an amebic abscess aspirate (e.g. in the liver or lungs), may reveal hematophagous trophozoites. It must be examined immediately by mixing a drop of warm saline with some aspirated pus on a microscope slide.

Endolimax nana

Introduction

Endolimax nana is a small non-pathogenic amoeba with world wide distribution. Its life cycle is similar to that of *E. histolytica* but is non-invasive.

Morphology of Trophozoite

Trophozoites of *E. nana* measures from 6-12µm. Motility is sluggish with blunt hyalin pseudopodia. In a permanently stained preparation, the nucleus exhibits a large karyosome with no peripheral chromatin on the nuclear membrane.

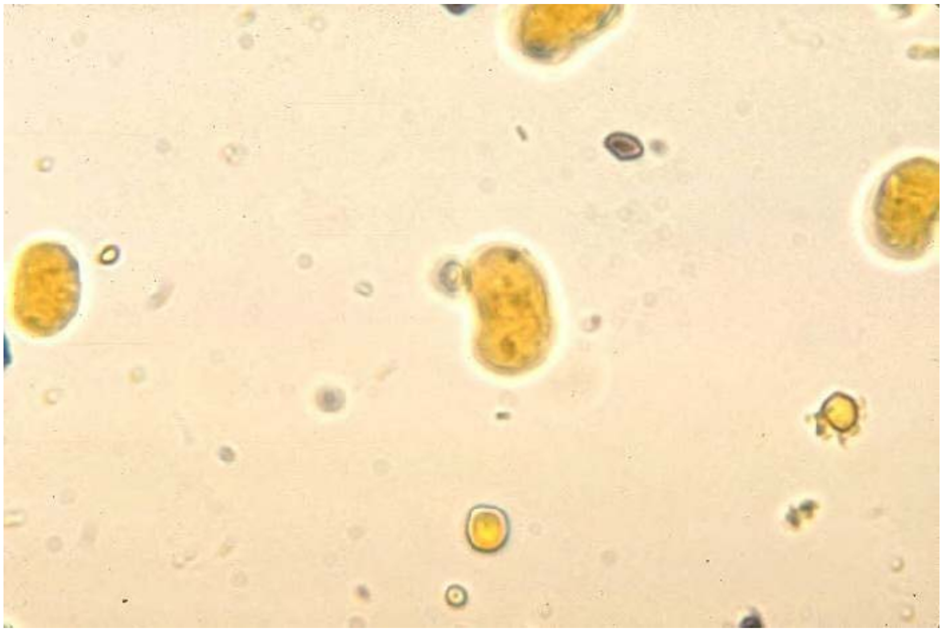

Figure 1-8. *Endolimax nana* cyst. Iodine stain. (SOURCE: PHIL 530 - CDC/Dr. L.L.A. Moore, Jr.)

Morphology of Cysts

Cysts of *E. nana* are 6-9µm in diameter. They can be spherical or ovoid in shape and contain four pinpoint nuclei, which are highlighted by the addition of iodine. Chromatoid bodies are not found and glycogen is diffuse.

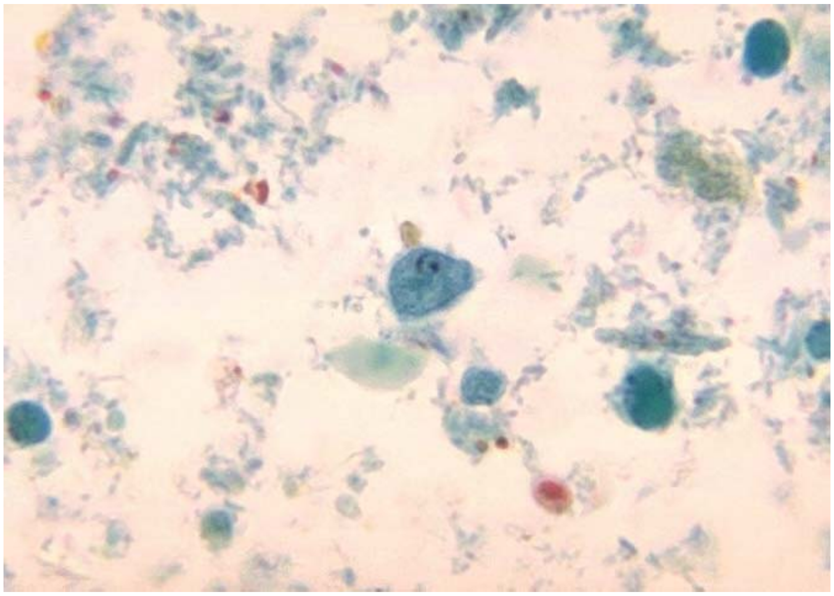

Image 1-9. *Endolimax nana* trophozoite (SOURCE: PHIL 1463 - CDC/Dr. Mae Melvin)

Laboratory Diagnosis

Laboratory diagnosis is made by finding the characteristic cysts in an iodine stained, formol-ether concentration method or by detecting the characteristic trophozoites in a wet preparation or a permanent stained preparation.

Microscopy

Where amebic dysentery is suspected, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed. On cooling the ameba stop moving which then become very difficult to identify. Direct microscopy should be done by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of motile trophozoites of *Endolimax nana* and can also provide information on the content of the stool (i.e., the presence of leucocytes and red blood cells). On search e.g. primarily for cysts, not for ameba, several stool samples are required to be examined, by direct microscopy and a sensitive concentration technique. Three negative stool samples are required before it can be accepted that there is no amebic infection. Microscopic examination of an amebic abscess aspirate (e.g. in the liver or lungs), may reveal hematophagous trophozoites. It must be examined immediately by mixing a drop of warm saline with some aspirated pus on a microscope slide.

Iodamoeba bütschlii

Introduction

Iodamoeba bütschlii is a non-pathogenic amoeba with world wide distribution although not as common as *E. coli or E. nana*. Its life cycle is similar to that of *E. histolytica* but is non-invasive.

Morphology of Trophozoites

Trophozoites of *I. bütschlii* are 8-20µm and are actively motile. On a permanently stained fecal smear, a nucleus with a large karyosome is evident. Chromatin bodies form striations around the karyosome. The cytoplasm appears granular containing vacuoles with ingested bacteria and debris.

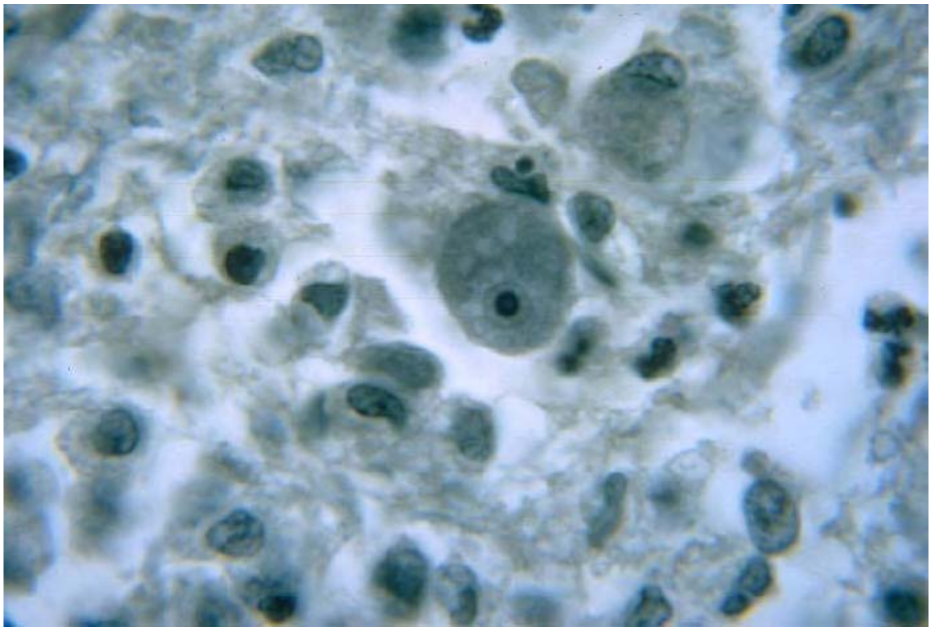

Image 1-10. Histopathology of amebiasis of brain due to *Iodamoeba bütschlii*. (SOURCE: PHIL 323 - CDC/Dr. Martin D. Hicklin)

Morphology of cysts

Cysts of *I. bütschlii* are 9-15µm in diameter and have one nucleus in mature cysts usually eccentrically placed. Chromatoid bodies are not present. Glycogen is present as a compact well defined mass staining dark brown with iodine.

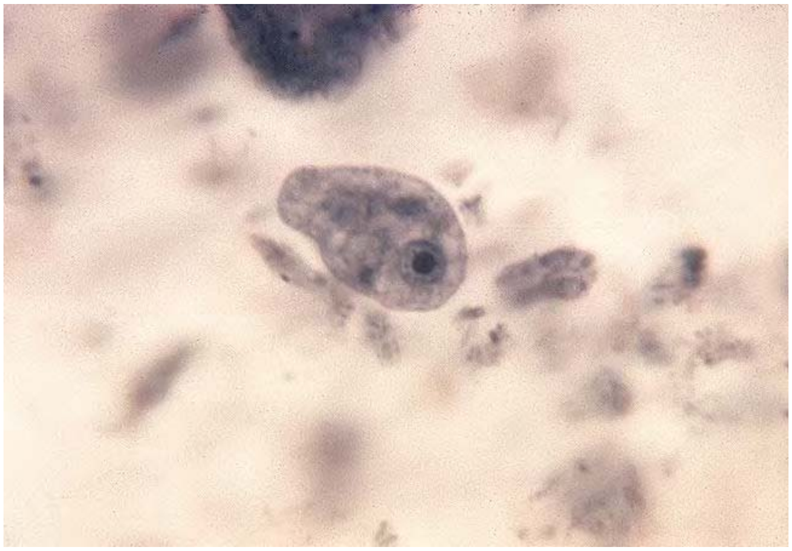

Image 1-11. *Iodamoeba bütschlii* cyst. (SOURCE: PHIL 587 - CDC/Dr. L.L.A. Moore, Jr.)

Laboratory Diagnosis

Laboratory diagnosis is made by finding the characteristic cysts in an iodine stained, formol-ether concentration method. Trophozoites are difficult to detect in a wet preparation.

Microscopy

Where amebic dysentery is suspected, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed. On cooling the ameba stop moving which then become very difficult to identify. Direct microscopy should be done by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of motile trophozoites of *Iodamoeba bütschlii* and can also provide information on the content of the stool (i.e., the presence of leucocytes and red blood cells). On search e.g. primarily for cysts, not for ameba, several stool samples are required to be examined, by direct microscopy and a sensitive concentration technique. Three negative stool samples are required before it can be accepted that there is no amebic infection. Microscopic examination of an amebic abscess aspirate e.g. in the liver or lungs, may reveal hematophagous trophozoites. It must be examined immediately by mixing a drop of warm saline with some aspirated pus on a microscope slide.

Blastocystis hominis

Introduction

B. hominis is an inhabitant of the human intestinal tract. It is capable of both sexual and asexual reproduction by binary fission and of pseudopod extension and retraction.

Morphology

The classic form that is usually seen in stool specimens varies in size from 6-40µm and is characterized by a large membrane bound central body which occupies 90% of the cell. It has no internal nuclear structure and a rim of peripheral granules the function of which is not known.

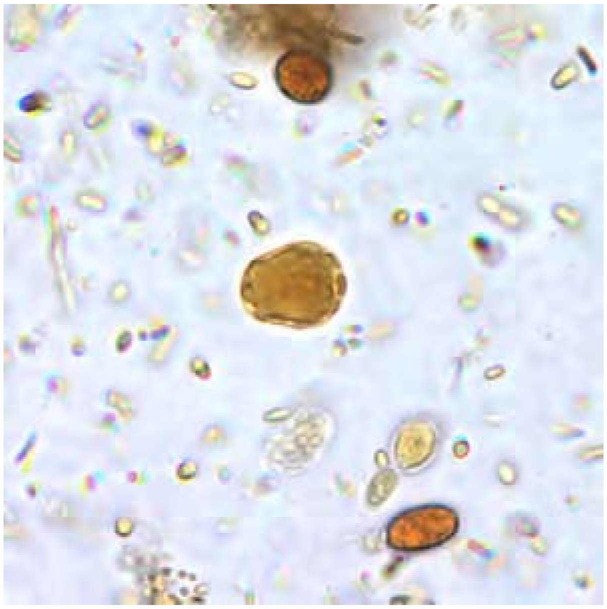

Image 1-12. *Blastocystis hominis* cyst-like forms in a wet mount stained in iodine. (SOURCE: CDC)

Clinical Disease

The pathogenic potential of *B. hominis* is unclear, though this organism has been associated with nausea, fever, vomiting, diarrhea and abdominal pain.

Laboratory Diagnosis

Permanently stained preparations of fecal smears is the procedure of choice for identification, although the organism can be seen in wet preparations. The recommended stains are Fields' and Giemsa.

1.3 Direct Fecal Parasite Concentration Methods

Diagnosis of intestinal parasites is confirmed by the recovery of protozoan trophozoites and cysts, helminth eggs and larvae in the clinical parasitology laboratory. Microscopic examination of feces is essential for the recognition and identification of intestinal parasites. Due to the low density of the parasites in the feces, direct microscopy is useful for the observation of motile protozoan trophozoites and the examination of cellular exudate, is not recommended solely for the routine examination of suspected parasitic infections. It is essential to increase the probability of finding the parasites in fecal samples to allow for an accurate diagnosis. Therefore, a concentration method is employed. (Direct wet mount examination should not be entirely excluded as the trophozoites are usually destroyed during the concentration procedure and therefore, microscopic examination of wet mounts should be performed).

Conventional Methods

Ridley-Allen Method

The concentration procedure used in hospitals requires the use of ether or ethyl acetate as a lipid removing agent and formalin as a fixative. The process involves the use of either expensive brass sieves or the use of tea strainers as the filter element.

Tea strainers have a very open pore structure of at least 600 micron and due to the shape of the strainer it is a non-linear pore size.

The fecal matter is filtered directly through these meshes in a dead stop manner, and hence there is the tendency for occlusion of the filter. There is also a formation of a secondary filter layer, which retains eggs and allows the extrusion of particles (particularly fibers) into the sediment. The net result is a reduction in egg yield and in sample clarity.

a. Using orange sticks, select a quantity of feces (approx. 1g) to include external and internal portions.
b. Place in a centrifuge containing 7ml of 10% formalin.
c. Emulsify the feces in the formalin and filter through the brass/plastic filter into the dish.
d. Wash the filter and discard any lumpy residue.
e. Transfer the filtrate to a boiling tube-add 3ml of ether and mix well on a vortex mixer for 15 seconds or mix by hand for 1 minute.
f. Transfer back to the centrifuge tube and centrifuge at 3,000 rpm for 1 min.
g. Loosen the fatty plug with an orange stick and pour the supernatant away by quickly inverting the tube.
h. Allow the fluid on the side of the tube to drain onto the deposit – mix well and transfer a drop to a slide for examination under a coverslip. (WHO Basic lab methods in Medical Parasitology)

The advantages of this method are that it will recover most ova, cysts and larvae and retain their morphology (thus facilitating identification). It has the disadvantage of destroying trophozoite stages and distorting cellular exudate. Liquid feces do not concentrate well, thus it is necessary in these cases to examine the stool by direct microscopy. Since the sieves are not disposable there is a problem with cleaning for re-use. The system is also open so there is a biohazard and odor issue.

Flotation Method

This technique is predominantly used in veterinary laboratories. By exploiting the density of the parasites, particularly eggs, it allows the parasites to float to the top of a dense solution (final specific gravity of about 1.20) and can then be skimmed from the top of the tube. The most commonly used reagent is zinc sulphate. Operculated eggs as well as schistosome and infertile Ascaris eggs are not easily recovered by this method. Also trophozoites are killed due to the high specific gravity and certain other fragile eggs such as *Hymenolepis nana* become distorted.

1. Crush 10-20g (about 1 teaspoon) of feces with applicator sticks and mix well with 10-12ml of saline. Filter the mixture through two layers of dampened surgical gauze into a 15ml conical centrifuge tube.
2. Centrifuge the suspension at 1500 rpm for five minutes. Decant the supernatant into disinfectant. Resuspend the sediment and recentrifuge in saline if there is excessive debris in the sample.
3. Resuspend and thoroughly mix the sediment in 12ml of zinc sulphate solution (specific gravity, 1.18 to 1.20, as verified with a hydrometer).
4. Centrifuge for one minute at 2500 rpm. Place tube in a rack in a vertical position and slowly add enough zinc sulphate with a dropper pipette to fill the tube so that an inverted meniscus forms.
5. Without shaking the tube, carefully place a 22 x 22µm coverslip on top of the tube so that its underside rests on the meniscus. The meniscus should not be so high that fluid runs down the side of the tube carrying parasites away from the cover glass.
6. Allow the tube to stand vertically in a rack with the coverslip suspended on top for ten minutes.
7. Carefully lift the coverslip with its hanging drop containing parasite eggs and cysts on the underside and mount on a clean slide, liquid side down. A small drop of iodine stain may be placed on the slide prior to adding the coverslip. The slide is gently rotated after adding the coverslip to ensure a uniform mixture. The slide is then thoroughly examined microscopically. (Clinical Laboratory procedures)

Chapter 2.

The Ciliates, Coccidia, and Microsporidia

2.1 Infections Acquired Through the Gastrointestinal Tract

In vertebrates, by far the most favorable sites for intestinal parasites are the duodenum, ileum, cecum and large intestine. To survive to reproduce in the gastrointestinal tract the parasites have to adapt to continuous physiological changes relative to the feeding habits of the host, the battery of protein, fat and carbohydrate-splitting enzymes, pH changes and the almost oxygen-free environment. Despite these features many parasites, both protozoa and helminths are capable of colonizing the gastrointestinal tract successfully.

Protozoa are single-celled animals which resemble a single cell of a higher organism. However, the protozoan cell is capable of carrying out vital functions such as reproduction, feeding, and locomotion. Intestinal protozoa include species which can live in the lumen of the intestine and others which additionally live and reproduce in the cells of the intestinal walls.

The protozoa make up a wide spectrum of organisms which have different life cycles and variable characteristics.

The Ciliates

2.2 The Parasites

The ciliates belong to the family Ciliophora. They possess simple cilia or compound ciliary organelles, two types of nuclei and a large contractile vacuole. The only member of the ciliate family to cause human disease is *Balantidium coli*.

Balantidium coli

Introduction

Balantidium coli is widely distributed in warmer climates, which is where human infections most commonly occur. The organisms inhabit the large intestine, cecum and terminal ileum where they feed on bacteria. The most common hosts being humans, pigs and rodents. Human infection is usually from pigs and is rare.

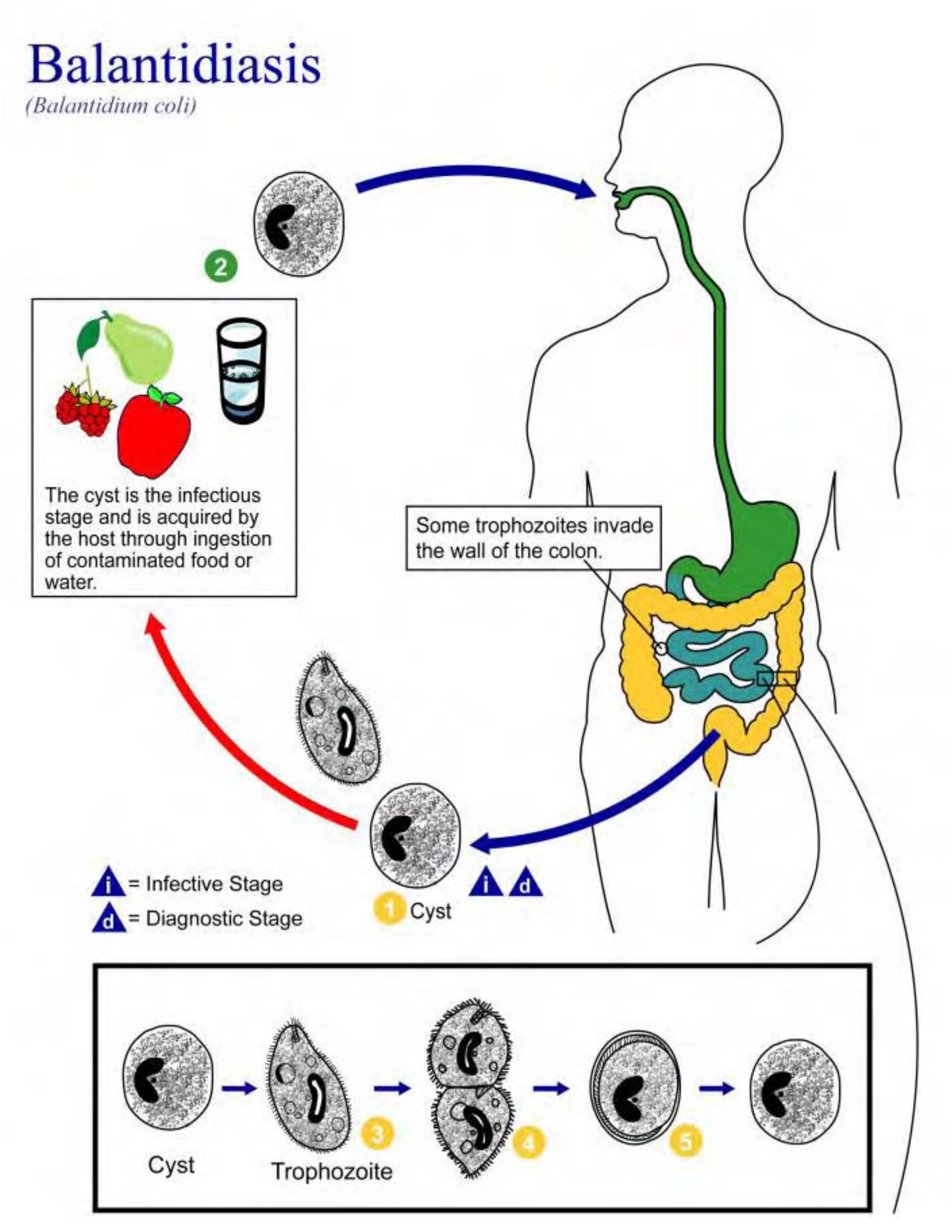
Balantidiasis
(Balantidium coli)
2
The cyst is the infectious stage and is acquired by the host through ingestion of contaminated food or water.
Some trophozoites invade the wall of the colon.
i = Infective Stage
d = Diagnostic Stage
1 Cyst
i d
Cyst
Trophozoite
3
4
5

Illustration 2-1. Life cycle of *Balantidium coli*. (SOURCE: PHIL 3380 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology of the Cyst

The cyst is spherical or ellipsoid and measures from 30-200µm by 20-120µm. It contains 1 macro and 1 micronucleus. The cilia are present in young cysts and may be seen slowly rotating, but after prolonged encystment, the cilia disappear. Cysts form when diarrhea subsides and the rectal contents become formed. The cyst, ingested by a fresh host, excysts to liberate the trophozoite.

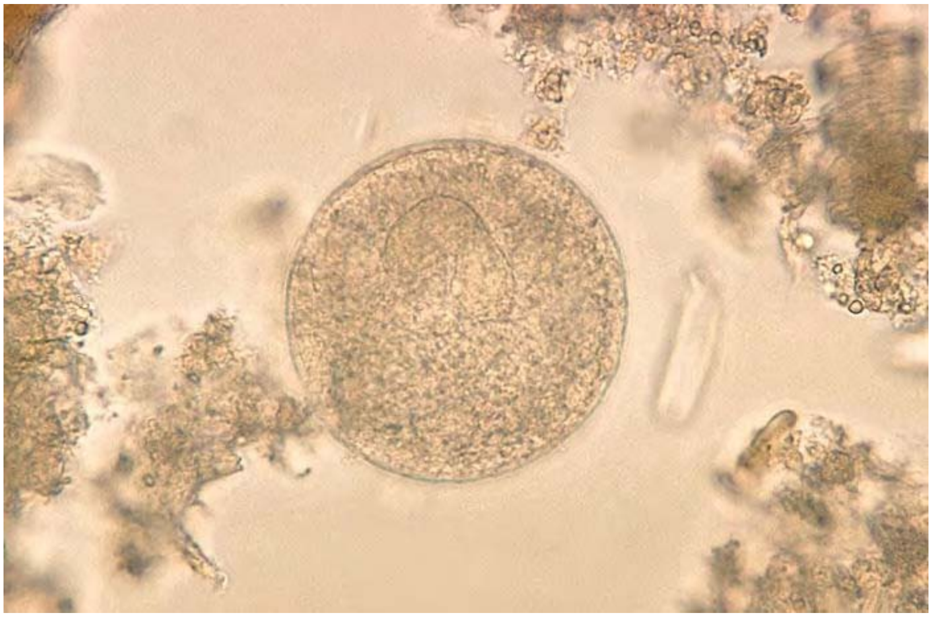

Image 2-1. *Balantidium coli* cyst. (SOURCE: PHIL 584 - CDC/Dr. L.L.A. Moore, Jr.)

Morphology of the Trophozoite

Trophozoites of *B. coli* measure approximately 30-150µm in length x 25-120µm in width but have been known to attain lengths of up to 200µm. They are oval in shape and covered in short cilia. A funnel shaped cytosome can be seen near the anterior end. Multiplication is by binary fission in the trophozoite stage. In an unstained preparation, the organisms are easily recognized because of their size and rapid revolving rotation. In a stained preparation, the characteristic macro and micronuclei may be observed.

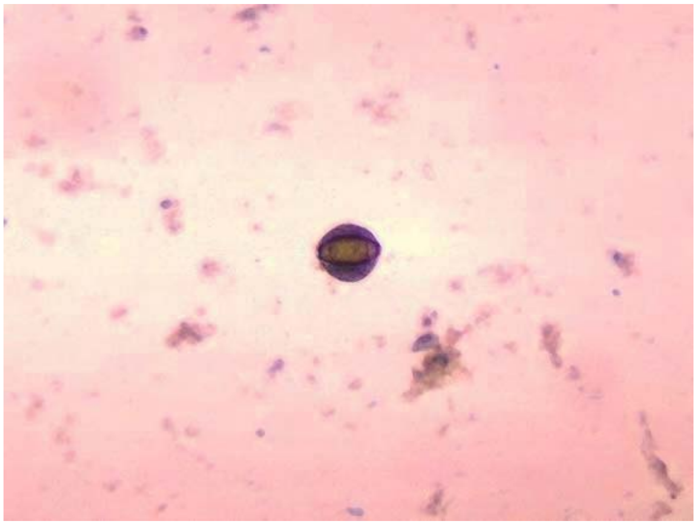

Image 2-2. *Balantidium coli* cyst with *Trichuris* egg inside. Unstained wet mount. (SOURCE: PHIL 1460 - CDC/Dr. Mae Melvin)

Clinical Disease

Severe *B. coli* infections may resemble amebiasis. Symptoms include diarrhea, nausea, vomiting, and anorexia. The diarrhea may persist for long periods of time resulting in acute fluid loss. *Balantidium coli* also has the potential to penetrate the mucosa resulting in ulceration just as those of *Entamoeba histolytica,* but perforation is more common. Metastatic lesions do not occur. Extra-intestinal disease has also been reported, but is rare.

Laboratory Diagnosis

Wet preparations of fresh and concentrated stool samples reveal the characteristic cysts and motile trophozoites. They are easier to identify in direct-smear saline preparations than permanently stained fecal smears.

The Coccidia

The Coccidia are a group of organisms which parasitize the epithelial cells of the intestinal tract. This group includes *Cryptosporidium parvum, Cyclospora cayetanensis* and *Isospora belli*. Most of the coccidian infections in man are zoonoses (having the potential to infect animals or arise from animals). In immunocompetent individuals, they usually produce mild, self-limiting infections.

Cryptosporidium parvum

Introduction

Cryptosporidium species, are coccidian protozoa, which are cosmopolitan in distribution, occurring in both developed and underdeveloped countries and causing infection in both humans and their livestock. *Cryptosporidium parvum* is the species responsible for human infection.

Development of *Cryptosporidia* occurs in a parasitophorous vacuole located on the microvillous surface of the epithelial cells.

Life Cycle and Morphology

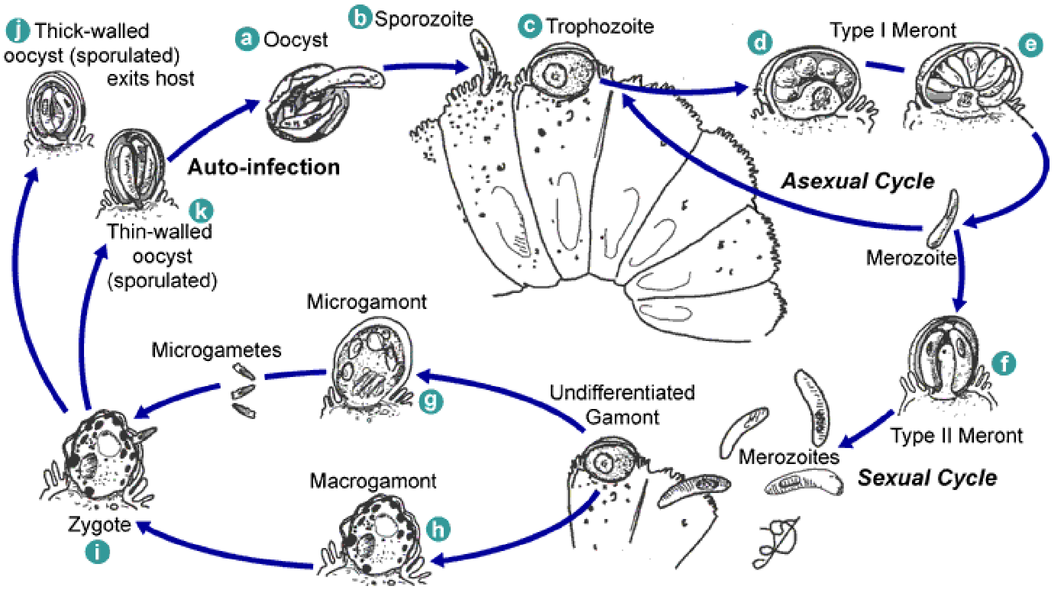

j Thick-walled oocyst (sporulated) exits host
a Oocyst
b Sporozoite
c Trophozoite
d
Type I Meront
e
Auto-infection
Asexual Cycle
k Thin-walled oocyst (sporulated)
Merozoite
Microgamont
Microgametes
g
f
Undifferentiated Gamont
Merozoites
Type II Meront
Sexual Cycle
Macrogamont
h
Zygote
i

Illustration 2-2. Life cycle of *Cryptosporidium* **sp.** Sporulated oocysts, containing 4 sporozoites, are excreted by the infected host through feces and possibly other routes such as respiratory secretions. Following ingestion (and possibly inhalation) by a suitable host, excystation ⓐoccurs. The sporozoites are released and parasitize epithelial cells (ⓑ, ⓒ) of the gastrointestinal tract or other tissues such as the respiratory tract. In these cells, the parasites undergo asexual multiplication (schizogony or merogony) (ⓓ, ⓔ, ⓕ) and then sexual multiplication (gametogony) producing microgamonts (male) ⓖand macrogamonts (female) ⓗ. Upon fertilization of the macrogamonts by the microgametes (ⓘ), oocysts (ⓙ, ⓚ) develop that sporulate in the infected host. Two different types of oocysts are produced, the thick-walled, which is commonly excreted from the host ⓙ, and the thin-walled oocyst ⓚ, which is primarily involved in autoinfection. Oocysts (measuring 4-5µm in diameter and containing 4 sporozoites) are infective upon excretion, thus permitting direct and immediate fecal-oral transmission. (SOURCE: PHIL 3386 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Clinical Disease

C. parvum is now widely recognized as a cause of acute gastro-enteritis, particularly in children. The infection produces a persistent, watery, offensive diarrhea often accompanied by abdominal pain, nausea, vomiting and anorexia. In immunocompetent persons, symptoms are usually short lived (one to two weeks). The small intestine is the site most commonly affected, symptomatic *Cryptosporidium* infections have also been found in other organs including other digestive tract organs, the lungs, and possibly conjunctiva.

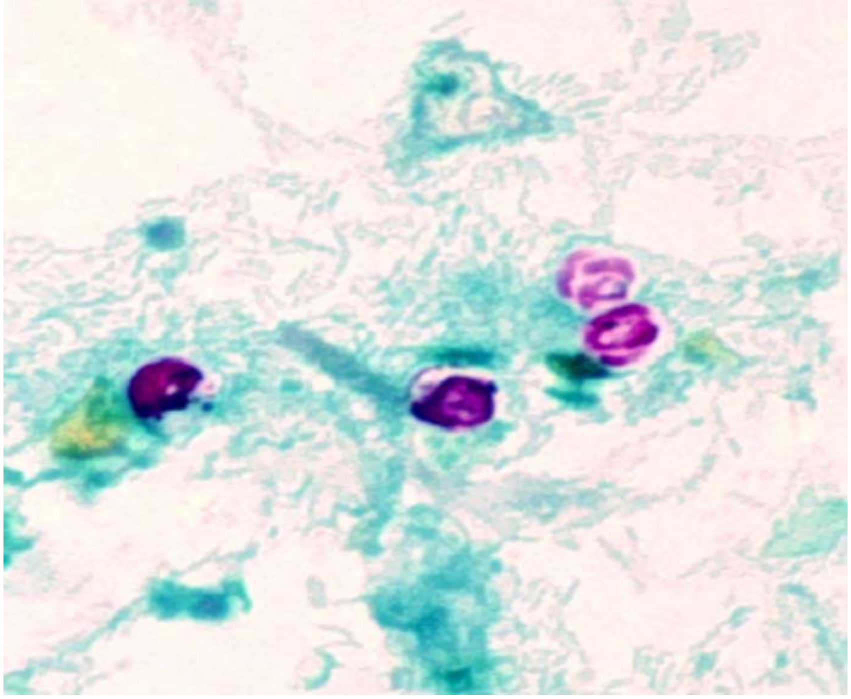

Image 2-3. *Cryptosporidium parvum* oocysts. This photomicrograph revealed the morphologic details of *Cryptosporidium parvum* oocysts, i.e., encapsulated zygotes, which had been stained using the modified acid-fast method. These oocysts exhibit a bright red coloration when using this staining technique, and in this case, you'll note the sporozoites that were made visible inside the two oocysts on the right. Sporozoites are the nucleated, motile stage of development through which many protozoans pass such as *C. parvum*, on their way to becoming adults, and represent a very infectious form of these organisms. The sporozoites will be released from these *C. parvum* oocysts. (SOURCE: PHIL 7829 - CDC/ DPDx - Melanie Moser)

Cryptosporidiosis in immunocompromised individuals, especially in HIV patients, can be life threatening, as many as 10% may pass oocysts of *C. parvum*. Infections are characterized by the production of frequent, large volume watery stools and sometimes there is invasion of the pancreas, biliary or respiratory tract.

Oocyst excretion and symptoms may fluctuate during the course of infection. Asymptomatic infections are commonly found in developing countries with poor hygiene, where there is close contact with livestock.

Laboratory Diagnosis

Definitive diagnosis of cryptosporidiosis is by finding the characteristic spherical oocysts in fecal samples. They do not concentrate well using standard concentration techniques and are identified using various staining techniques.

Using the modified Ziehl-Neelsen staining method (fuschin followed by methylene blue), the oocysts are acid fast. However, staining results within a smear and between specimens are diverse, varying from unstained to partial red staining and complete staining. Fully sporulated forms can be seen in which the red staining sporozoites are within an unstained oocyst wall. When staining the fecal smear with phenol-auramine/carbol-fuchsin, the oocysts appear as bright yellow discs with an "erythrocyte" pattern of staining against a dark red background. Detection of the oocysts can also be achieved by using specific polyclonal or monoclonal antibodies conjugated to fluorescein. These tests are now commercially available and offer a high degree of sensitivity. However, caution must be exercised when they are used to detect oocysts in the fecal smears distributed by NEQAS parasitology. Such specimens are preserved in formalin, which interferes with the fluorescent staining of the parasites, and are thus difficult to detect.

Oocysts in stool specimens (fresh or in storage media) remain infective for extended periods. Thus stool specimens should be preserved in 10% buffered formalin or sodium acetate-acetic acid-formalin (SAF) to render oocysts non-viable. (Contact time with formalin necessary to kill oocysts is not clear; we suggest at least 18 to 24 hours).

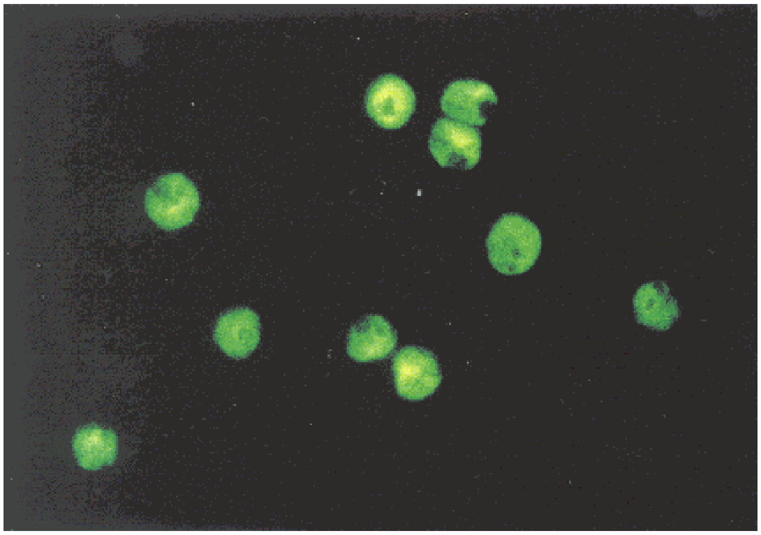

Image 2-4. Staining of *Cryptosporidium parvum* oocysts in a stool smear with monoclonal antibodies conjugated to fluorescein. The *Cryptosporidium* oocysts appear with a peripheral green fluorescence. This technique could be of interest in epidemiological inquiries. (x 670) (SOURCE: Unknown)

Detection of the oocysts can also be achieved by using rapid antigen detection testing kits that specifically detect antigen released by *Cryptosporidium parvu.* These tests are now commercially available and offer a high degree of sensitivity even on specimens that are preserved.

Isospora belli

Introduction

Isospora belli is a coccidian protozoan of cosmopolitan distribution, occurring especially in warm regions of the world infecting both humans and animals.

Life Cycle

The life cycle of *I. belli* involves an asexual (schizogonic stage) and a sexual (sporogonic stage).

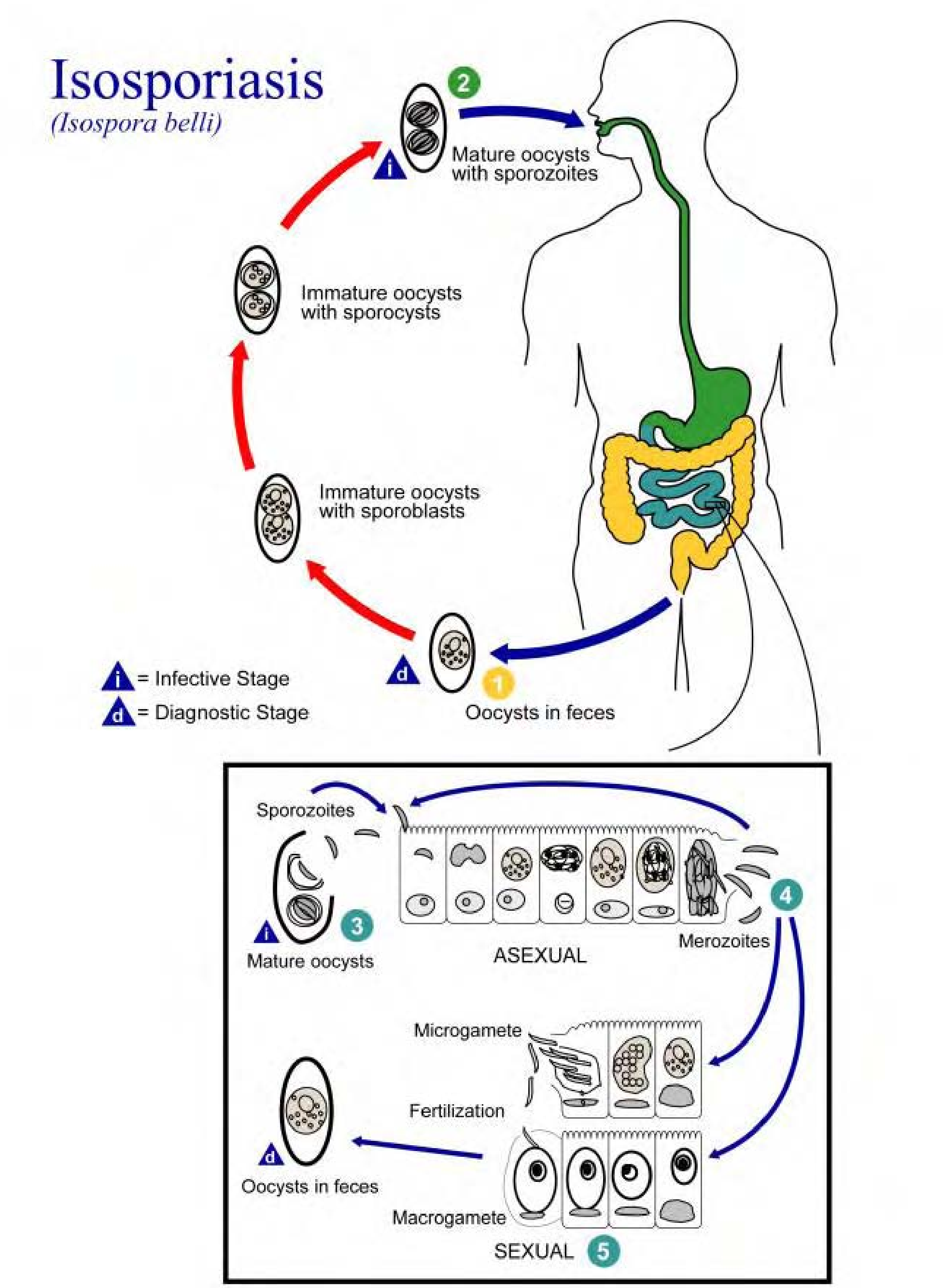
Isosporiasis
(Isospora belli)
2
Mature oocysts with sporozoites
i
Immature oocysts with sporocysts
Immature oocysts with sporoblasts
i = Infective Stage
d = Diagnostic Stage
d
1
Oocysts in feces
Sporozoites
i
3
Mature oocysts
ASEXUAL
Merozoites
4
Microgamete
Fertilization
d
Oocysts in feces
Macrogamete
SEXUAL
5

Illustration 2-3. *Isospora belli* life cycle. At time of excretion, the immature oocyst contains usually one sporoblast (more rarely two) ①. In further maturation after excretion, the sporoblast divides in two (the oocyst now contains two sporoblasts); the sporoblasts secrete a cyst wall, thus becoming sporocysts; and the sporocysts divide twice to produce four sporozoites each ②. Infection occurs by ingestion of sporocysts-containing oocysts: the sporocysts excyst in the small intestine and release their sporozoites, which invade the epithelial cells and initiate schizogony ③. Upon rupture of the schizonts, the merozoites are released, invade new epithelial cells, and continue the cycle of asexual multiplication ④. Trophozoites develop into schizonts which contain multiple merozoites. After a minimum of one week, the sexual stage begins with the development of male and female gametocytes ⑤. Fertilization results in the development of oocysts that are excreted in the stool ①. *Isospora belli* infects both humans and animals. (SOURCE: PHIL 3398 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

Infection with *Isospora belli* occurs in both immunocompetent and immunocompromised patients and begins when the mature oocyst is ingested in water or food.

Morphology of oocysts. The mature oocyst contains 2 sporocysts, each containing 4 sporozoites measure on average 35 x 9µm.

Figures 6. and 7. demonstrate fecal smears of oocysts. The sporulated oocyst is the infective stage of the parasite and they excyst in the small intestine releasing sporozoites which penetrate the epithelial cells, thus initiating the asexual stage of the life cycle. The sporozoite develops in the epithelial cell to form a schizont, which ruptures the epithelial cell containing it, liberating merozoites into the lumen. These merozoites will then infect new epithelial cells and the process of asexual reproduction in the intestine proceeds. Some of the merozoites form macrogametes and microgametes (sexual stages) which fuse to form a zygote maturing finally to form an oocyst.

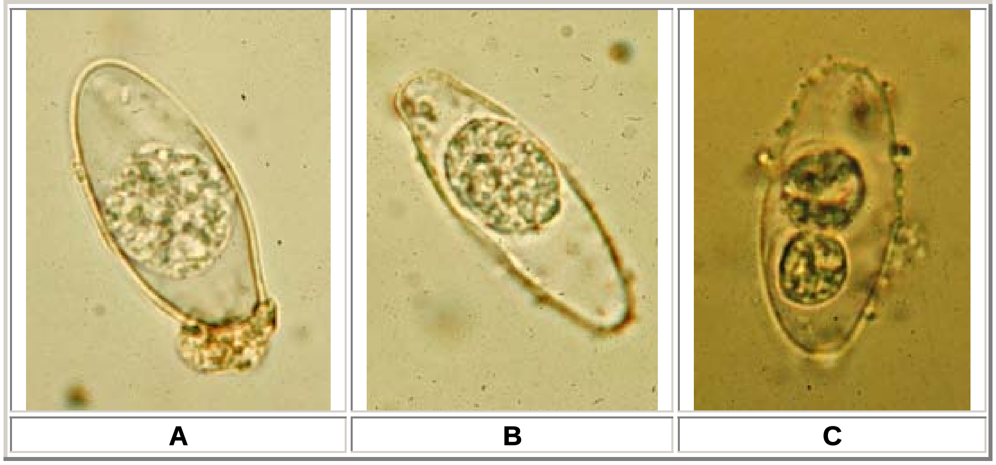

Image 2-5. **A**, **B**, **C:** Oocysts of *Isospora belli*. The oocysts are large (25 to 30 µm) and have a typical ellipsoidal shape. When excreted, they are immature and contain one sporoblast (**A**, **B**). The oocyst matures after excretion: the single sporoblast divides in two sporoblasts (**C**), which develop cyst walls, becoming sporocysts, which eventually contain four sporozoites each. (Images contributed by Georgia Division of Public Health. (CDC))

Clinical Disease

In the immunocompetent, infection is generally asymptomatic or a self-limiting gastro-enteritis. However, in chronic infections, severe non-bloody diarrhea with cramp-like abdominal pain can last for weeks and result in fat malabsorption and weight loss. Eosinophilia may be present (atypical of other protozoal infections).

In immunocompromised individuals, infants and children, infection ranges from self-limiting enteritis to severe diarrheal illness resembling that of cryptosporidiosis.

Laboratory Diagnosis

Oocysts are thin walled, transparent and ovoid in shape. They can be demonstrated in feces after a formal ether concentration where they appear as translucent, oval structures.

Alternatively, oocysts can be seen in a fecal smear stained by a modified Ziehl-Neelsen method, where they stain a granular red color against a green background, or by phenol-auramine.

Cyclospora cayetanensis

Introduction

Cyclospora cayetanensis, a coccidian protozoan, has been described in association with diarrheal illness in various countries, in particular Nepal, Pakistan and India. Infection results in a disease with non-specific symptoms. Quite often the disease is the cause of unexplained summer diarrheal illness and similar illness following travel to tropical areas.

Life Cycle and Morphology

The life cycle of this organism is unknown, however environmental data suggest that *Cyclospora*, like *Cryptosporidium* species, is a water-borne parasite. The oocysts of *C. cayetanensis* are spherical, measuring 8-10µm in diameter and the mature oocyst contains 2 sporocysts. Oocysts of *Cyclospora cayetanensis*, are twice as large in comparison with *C. parvum* and are not sporulated (do not contain sporocysts - upon excretion).

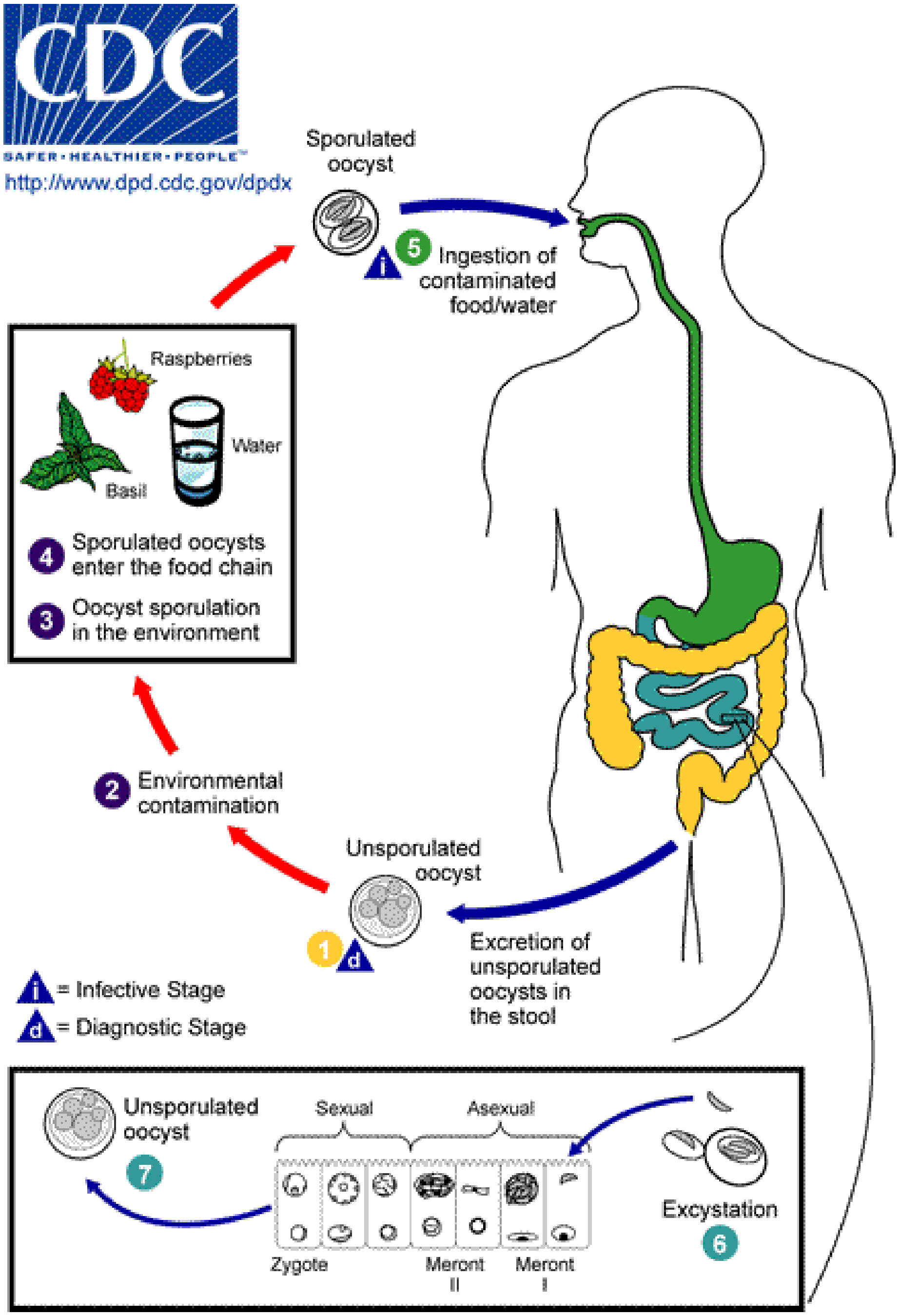
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Sporulated oocyst
5 Ingestion of contaminated food/water
i
Raspberries
Water
Basil
4 Sporulated oocysts enter the food chain
3 Oocyst sporulation in the environment
2 Environmental contamination
Unsporulated oocyst
1
d
Excretion of unsporulated oocysts in the stool
i = Infective Stage
d = Diagnostic Stage
Unsporulated oocyst
7
Sexual
Asexual
Zygote
Meront II
Meront I
Excystation
6

Some of the elements in this figure were created based on an illustration by Ortega et al. *Cyclospora cayetanensis*. In: Advances in Parasitology: opportunistic protozoa in humans. San Diego: Academic Press; 1998. p. 399-418.

Illustration 2-4. When freshly passed in stools, the oocyst is not infective ①(thus, direct fecal-oral transmission cannot occur; this differentiates *Cyclospora* from another important coccidian parasite, *Cryptosporidium*). In the environment ②, sporulation occurs after days or weeks at temperatures between 22°C to 32°C, resulting in division of the sporont into two sporocysts, each containing two elongate sporozoites ③. Fresh produce and water can serve as vehicles for transmission ④and the sporulated oocysts are ingested (in contaminated food or water) ⑤. The oocysts excyst in the gastrointestinal tract, freeing the sporozoites which invade the epithelial cells of the small intestine ⑥. Inside the cells they undergo asexual multiplication and sexual development to mature into oocysts, which will be shed in stools ⑦. The potential mechanisms of contamination of food and water are still under investigation. (SOURCE: CDC)

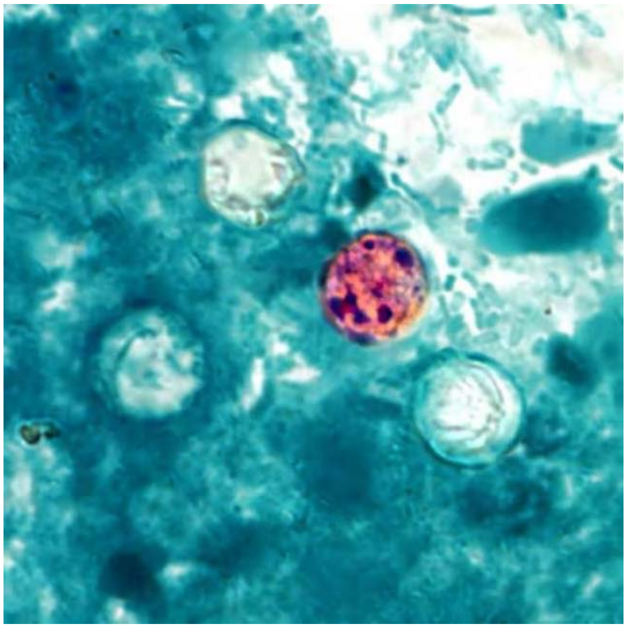

Image 2-6. This photomicrograph of a fresh stool sample, which had been prepared using a 10% formalin solution, and stained with modified acid-fast stain, revealed the presence of four *Cyclospora cayetanensis* oocysts in the field of view. Compared to wet mount preparations, the oocysts are less perfectly round and have a wrinkled appearance due to this method of fixation. Most importantly, the staining is variable among the four oocysts. (SOURCE: PHIL 7827 - CDC/ DPDx - Melanie Moser)

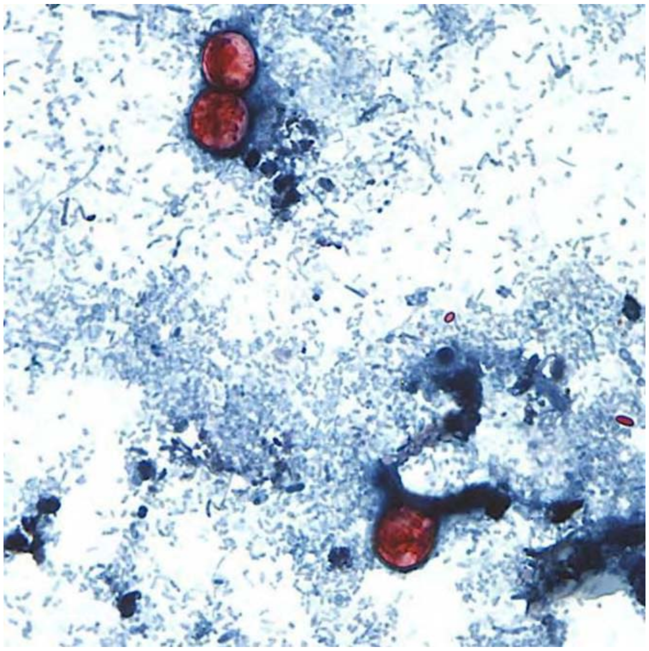

Image 2-7. This photomicrograph of a fresh stool sample, which had been prepared using a 10% formalin solution, and stained with safranin, revealed the presence of three uniformly stained *Cyclospora cayetanensis* oocysts in the field of view. (SOURCE: PHIL 7828 - CDC/ DPDx - Melanie Moser)

Clinical Disease

Patients from whose stools the organism has been isolated have reported nausea, vomiting, weight loss and explosive watery diarrhea. Flatulence and bloatedness, nausea and vomiting, myalgia, low-grade fever, and fatigue are associated symptoms. The site of infection is the small bowel. The disease is usually self-limiting to three to four days but untreated infections can last from several days to a month or longer, and may follow a relapsing course. Some infections are asymptomatic.

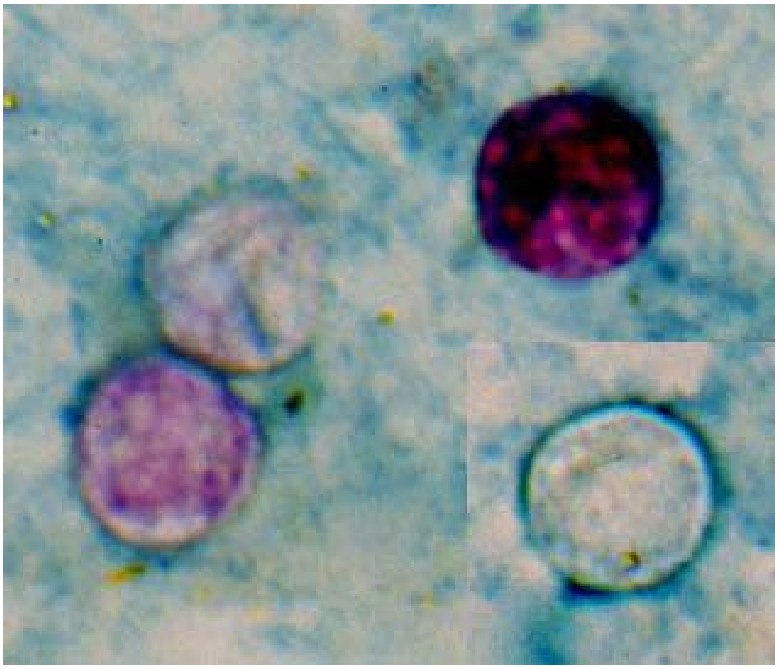

Image 2-8. Four *Cyclospora* oocysts from fresh stool fixed in 10% formalin and stained with modified acid-fast stain. Compared to wet mount preparations, the oocysts are less perfectly round and have a wrinkled appearance. Most importantly, the staining is variable among the four oocysts. (SOURCE: CDC)

Laboratory Diagnosis

The oocysts of *C. cayetanensis* are spherical as can be seen in formol-ether concentrated stool samples by light microscopy. They are refractile spheres which exhibit blue autofluorescence under ultraviolet light. It is important to note that UV microscopes set up for FITC and auramine microscopy only (450-500nm) will fail to detect the autofluorescence of the oocyst. Iodine-quartz microscopes do not produce UV wavelength below 400nm, while both mercury vapor and xenon vapor microscopes must be fitted with a 340-380nm excitation filter to demonstrate autofluorescence.

The oocysts are variably acid-fast when stained by the modified Ziehl-Neelsen method. Some cysts are acid-fast whereas others appear as round holes against a green background. They do not stain well with phenol-auramine.

Microsporidia Species

Introduction

The term microsporidia is also used as a general nomenclature for the obligate intracellular protozoan parasites belonging to the phylum Microsporidia. To date, more than 100 genera and 1,000 species have been described as parasites infecting a wide range of vertebrate and invertebrate hosts. There are at least seven microsporidian species that are well characterized as human pathogens. (**Table 2-1.**)

Microsporidia are characterized by the production of resistant spores and the polar tubule (or polar filament) which is coiled inside the spore as demonstrated by its ultrastructure.

They have recently come to medical attention as opportunistic pathogens in humans with Acquired Immune Deficiency Syndrome (AIDS) and have been implicated in conditions ranging from enteritis to keratoconjunctivitis.

Life Cycle and Morphology

Microsporidia are primitive organisms. They possess no mitochondria and have prokaryotic like ribosomes. Classification is based on the ultrastructural features, which include the number of coils in the polar tubes, the configuration of nuclei and the spore size 1-4µm, depending on the species.

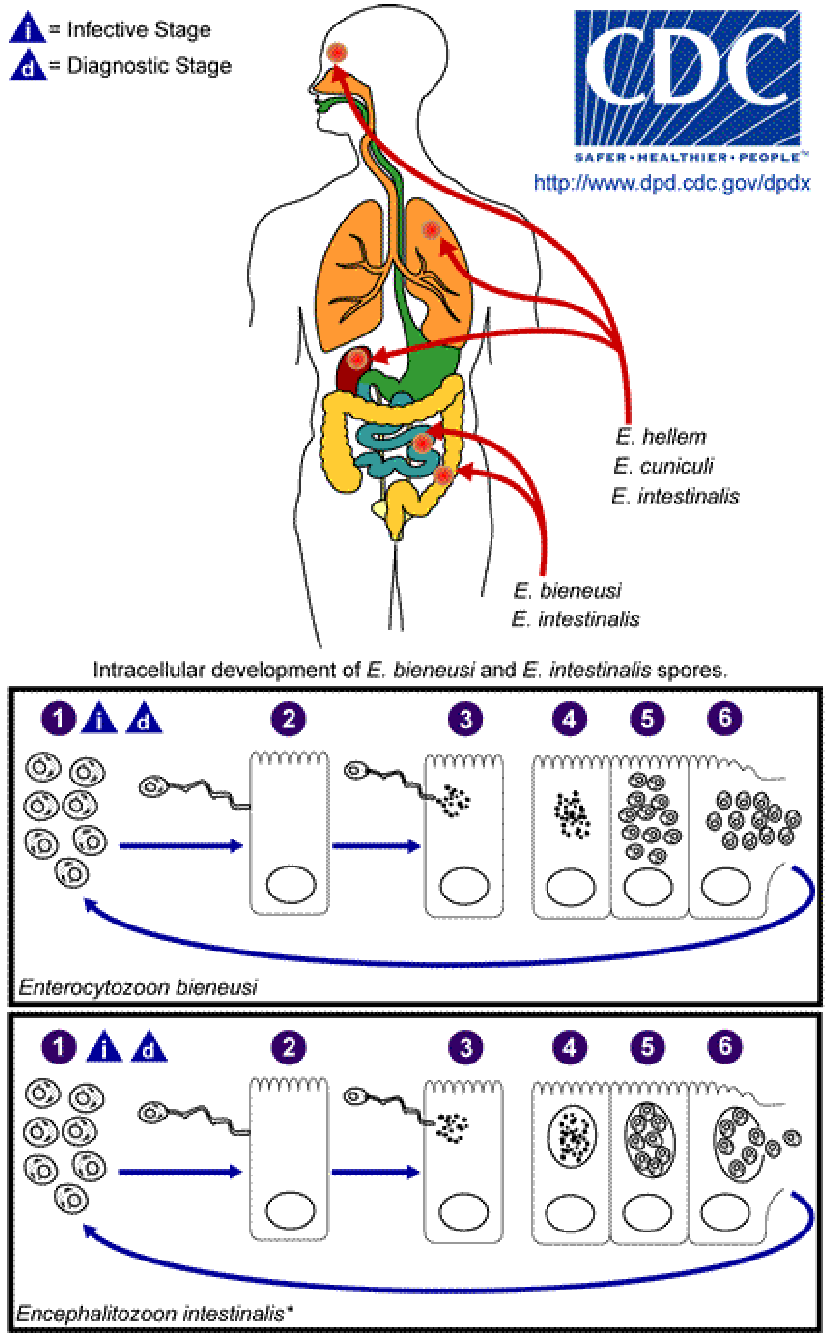
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER · HEALTHIER · PEOPLE™
http://www.dpd.cdc.gov/dpdx
E. hellem
E. cuniculi
E. intestinalis
E. bieneusi
E. intestinalis
Intracellular development of E. bieneusi and E. intestinalis spores.
1
i
d
2
3
4
5
6
Enterocytozoon bieneusi
Encephalitozoon intestinalis*

*Development inside parasitophorous vacuole also occurs in *E. hellem* and *E. cuniculi*.

Illustration 2-5. Life cycle of *Microsporidia* **sp.** The infective form of microsporidia is the resistant spore and it can survive for a long time in the environment ❶. The spore extrudes its polar tubule and infects the host cell ❷. The spore injects the infective sporoplasm into the eukaryotic host cell through the polar tubule ❸. Inside the cell, the sporoplasm undergoes extensive multiplication either by merogony (binary fission) or schizogony (multiple fission) ❹. This development can occur either in direct contact with the host cell cytoplasm (e.g., *E. bieneusi*) or inside a vacuole termed parasitophorous vacuole (e.g., *E. intestinalis*). Either free in the cytoplasm or inside a parasitophorous vacuole, microsporidia develop by sporogony to mature spores ❺. During sporogony, a thick wall is formed around the spore, which provides resistance to adverse environmental conditions. When the spores increase in number and completely fill the host cell cytoplasm, the cell membrane is disrupted and releases the spores to the surroundings ❻. These free mature spores can infect new cells thus continuing the cycle. (SOURCE: CDC)

Microsporidia	Size	Associated Disease
Enterocytozoon bieneusi	1µm x 1.5µm	Gastrointestinal and biliary tract infections
Encephalitozoon intestinalis	1.5µm x 2.5µm	Gastrointestinal tract and systemic infections
Encephalitozoon hellem	1.5µm x 1µm	Keratopathy (corneal edema), respiratory tract infection
Encephalitozoon cuniculi	1.5µm x 1µm	Central nervous system disease
Nosema connori	2µm x 4µm	Systemic infections
Nosema corneum	2µm x 4µm	Keratopathy
***Pleistophora* species**	2.8µm x 3.4µm	Myositis

Table 2-1. Microsporidia found in humans and their associated disease. (Cuomo)

In addition to the species in **Table 2-1.**, above, there are other, not well-characterized microsporidian human pathogens. These are designated as Microsporidum, a collective taxon that includes *Microsporidum africanum* and *Microsporidium ceylonensis.*

Clinical Disease

The most common microsporeans found in patients with AIDS are *Enterocytozoon bieneusi, Encephalitozoon intestinalis* and *Encephalitozoon hellem*. Patients with these infections tend to be severely immunodeficient with a CD4 count less than 100 x 10^6/L. Additionally, cases of microsporidiosis have been reported in immunocompromised persons not infected with HIV and in immunocompetent individuals. The clinical manifestations of microsporidiosis are very diverse, varying according to the causal species, with diarrhea being the most common.

Enterocytozoon bieneusi

Infections with *E. bieneusi* are restricted to the enterocytes of the small intestine, resulting in villous atrophy and malabsorption. Clinical symptoms include chronic watery, non-bloody diarrhea, malaise and weight loss.

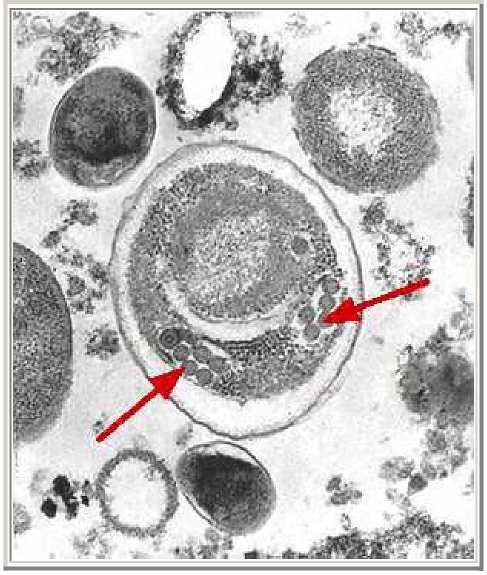

Image 2-9. Electron micrograph of an *Enterocytozoon bieneusi* spore. Arrows indicate the double rows of polar tubule coils in cross section which characterize a mature *E. bieneusi* spore. (SOURCE: CDC)

Encephalitozoon intestinalis

Infection with *Encephalitozoon intestinalis* occurs in the enterocytes of the small intestine but is more widely disseminated than *E. bieneusi* and has been found in the colon, liver and kidney.

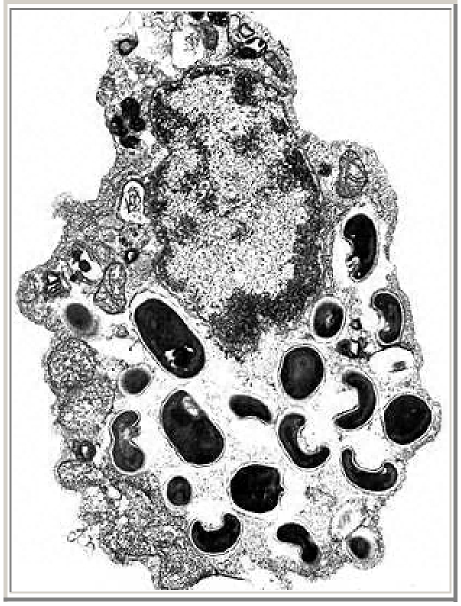

Image 2-10. Electron micrograph of an eukaryotic cell with *Encephalitozoon intestinalis* spores and developing forms inside a septated parasitophorous vacuole. The vacuole is a characteristic feature of this microsporidian species. (SOURCE: CDC)

Encephalitozoon hellem and *Encephalitozoon cuniculi*

These organisms have also been found in disseminated microsporidiosis. Clinical symptoms may include sinusitis, nephritis, hepatitis, keratoconjunctivitis and peritonitis.

Nosema corneum

This organism has been detected in AIDS patients with keratoconjunctivitis.

Laboratory Diagnosis

Initially, the diagnosis of intestinal microsporidiosis depended on tissue biopsies which were stained with Gram's stain and examined by light microscopy. However, in order that ill patients were not subjected to unnecessary invasive procedures, non-invasive diagnostic procedures were developed. The modified Trichrome stain and the Fungiqual fluorescent stain are the stains of choice. Immunofluorescence assays (IFA) using monoclonal and/or polyclonal antibodies are being developed for the identification of microsporidia in clinical samples.

2.3 Examination of fecal specimens for Parasites

Many intestinal disorders are due to intestinal parasites which cannot be diagnosed symptomatically. Laboratory investigation is therefore required and the staff responsible should have adequate expertise in examining fecal specimens for parasitic organisms.

2.4 Relevant information required

The request form should always state the patients clinical symptoms and signs and whether the patient had recent overseas travel. If the patient has had no recent history of overseas travel, examination for *Cryptosporidium, Giardia* and Microsporidia, if immunocompromised should be considered. If overseas travel has been undertaken, it is important to note is the patient ill or whether a routine post-tropical screen is requested. The geographical location is also important as it may indicate these parasites which could be present.

2.5 Collection of samples

If a fecal sample is not properly collected and taken care of before examination, it will be of little or no value for accurate diagnosis. This is especially true if protozoa are present. Amoebic trophozoites begin to degenerate one to two hours after passage, as do flagellate trophozoites. Cysts will deteriorate if the fecal specimen is left standing for many hours or overnight, especially at high temperatures.

Helminth eggs and larvae are less affected by the age of the specimen than are protozoa. Nevertheless, changes may occur that could affect their identification; hookworm larvae may become embryonated and larvae may hatch from the eggs risking confusion with *Strongyloides* larvae. Larvae themselves may disintegrate thus making their identification difficult.

To ensure that good specimens are provided for examination, it is important to note the following points.

1. A clean dry container must be used for the collection of fecal samples. Urine and water will destroy trophozoites, if present, and the presence of dirt also causes identification problems.
2. Ideally the specimen should be brought to the lab as soon as it is passed, to avoid deterioration of protozoa and alterations of the morphology of protozoa and helminths.
3. The specimen container should be clearly labeled with the patient’s name, date, and time of passage of the specimen.
4. An amount of stool adequate for parasite examination should be collected and a repeat sample requested if too little is supplied. The smallest quantity that should be accepted is about the size of a pigeon’s egg.
5. Diarrheal specimens, or those containing blood and mucus, should be examined promptly on arrival in the laboratory. The specimens may contain motile amoebic or flagellate trophozoites and may round up and thus be missed if examination is delayed. Where amoebic dysentery is suggested, the laboratory should be informed that a "hot stool" is being supplied so that it can be examined within twenty minutes of being passed.
6. With the exception of "hot stools" if specimens cannot be examined as soon as they arrive, they should be put in the refrigerator.

2.6 Visual observation of the fecal sample

It is important to observe the macroscopic appearance of the stool as this can give a clue to the type of organisms present. Therefore the consistency; formed, unformed or liquid; the color and the presence or absence of the exudate are reported. The presence of adult worms can also be seen in a freshly passed stool e.g. adult stage of *Ascaris lumbricoides* and *Enterobius vermicularis*. Proglottids of *Taenia* species can also be seen.

2.7 Routine procedure for the microscopic examination of fecal samples for parasites

1. Direct microscopy should be done on all unformed and liquid samples by mixing a small amount of the specimen in 0.9% sodium chloride solution. This permits detection of trophozoites of *Entamoeba histolytica* and *Giardia lamblia*. It can also provide information on the content of the stool i.e. the presence of leucocytes and red blood cells.
2. A formol-ether concentrate should be done on all fecal samples examined for parasites. This reveals the presence of most protozoan cysts, eggs of nematodes, cestodes and trematodes and also the larval stages of some nematodes.
3. A permanently stained direct fecal smear should be used for all bloody, liquid or semi-formed stools. The smear can reveal the presence of intestinal parasites that can be either destroyed or missed by the formol-ether concentration method e.g. *Dientamoeba fragilis*.
4. Specimens from patients with HIV should be left in 10% formalin for one hour before proceeding with parasite examination.

2.8 Principals of Diagnostic Methods for the Identification of Parasites

The principal of the successful identification of fecal parasites is based upon,

1. Measurement - The use of an eyepiece graticule is of the utmost importance, especially for cyst identification.
2. Morphology - In protozoan cysts, the number of nuclei and the presence of inclusions e.g. glycogen mass and chromidial bar, aid the identification of protozoa. In trophozoites, the presence of red cells in amoebae is diagnostic of *Entamoeba histolytica* and flagella also aid identification of some protozoan trophozoites.
3. Appearance - In helminth eggs, the shape of the egg, the thickness of the shell, the color of the ovum and the presence or absence of features such as an operculum, spine or hooklets are diagnostic pointers to the identity of the parasite.
4. Stains also aid in identification of the parasite.

The addition of iodine to formol ether concentrates highlights the internal structure of cysts and helps distinguish between vegetable matter and cysts. Permanently stained fecal smears are useful in demonstrating the nuclear pattern of cysts.

2.9 Problems of identification

Many things in stool specimens look like parasites but are not.

Epithelial cells and macrophages can be confused with amoebic trophozoites, especially macrophages that show slight amoeboid movement and may contain red blood cells. Pus cells can be confused with amoebic cysts. The nuclei appear as 3 or 4 rings and usually stain heavily. The cytoplasm is ragged and the cell membrane is often not seen. Amoebic cysts have a distinct cell wall.

Hair and fibers may be confused with larvae, but they do not have the same internal structure as larvae.

Plant cells can be confused with cysts or eggs. Though plant cells usually have a thick wall and cysts have a thin wall.

2.10 Reporting of Parasites

Ideally, the presence of all parasites should be reported, whether they are pathogens or non-pathogens. This particularly applies to the presence of cysts. However, if it is laboratory practice to report all cysts, the report should state whether they are pathogenic or non-pathogenic.

The stage of the parasite should always be reported. For the protozoa, whether cysts or trophozoites are present; the stage of larvae as in *Strongyloides;* and whether adult stages or eggs of helminths are present.

Chapter 3.
The Flagellates

3.1 Infections Acquired Through the Gastrointestinal Tract

The Flagellates belong to the Magistophora and possess more than one flagellum. Beating these flagella enable them to move. A cytosome may be present which helps in the identification of the species.

Flagellates possess one advantage over their ameboid relatives in that they can swim. Therefore, enabling them to invade and adapt to a wider range of environments unsuitable for other amoebae. They are able to change from a flagellated free-swimming environment to a non-flagellated tissue dwelling stage and *vice versa*.

Flagellates are known to inhabit the reproductive tract, alimentary canal, tissue sites and also the blood stream, lymph vessels and cerebrospinal canal.

There are pathogenic and commensal species of flagellates. The flagellates which are encountered in the intestinal tract are *Giardia lamblia*, *Dientamoeba fragilis, Chilomastix mesnili, Trichomonas hominis, Retortamonas intestinalis* and *Enteromonas hominis* (the latter two being less common). The trophozoites are easily recognized in saline preparations by their motility. However, accurate identification is done on a permanently stained fecal smear. Cysts are more commonly seen than the trophozoite.

Giardia lamblia

Introduction

Giardia lamblia is a flagellate of world-wide distribution. It is more common in warm climates than temporal climates. It is the most common flagellate of the intestinal tract, causing *Giardiasis*. Humans are the only important reservoir of the infection. The infection is most common in parts of the world where sanitation is at its lowest. *Giardiasis* is an infection of the upper small bowel, which may cause diarrhea. Only *Giardia* spreads disease.

Morphology of the Trophozoites

The trophozoites of *G. lamblia* are flattened pear shaped and are an average size of 15μm long, 9μm wide and 3μm thick. When stained, the trophozoite is seen to have two nuclei, two slender median rods (axostyles), and eight flagella arising from the anterior end. They have been described as looking like tennis rackets without the handle (they are often seen has having a comical face-like appearance when looking at the front view).

The movement of the trophozoites is described as tumbling leaf motility, using their four pairs of flagella for locomotion. They attach themselves to the surface of the jejunal or duodenal mucosa by their disc-like suckers which are found on their ventral surface. They multiply in the gut by binary fission. Once the trophozoites drop off the mucosal surface they are normally carried in the intestinal contents down the gut where they usually encyst.

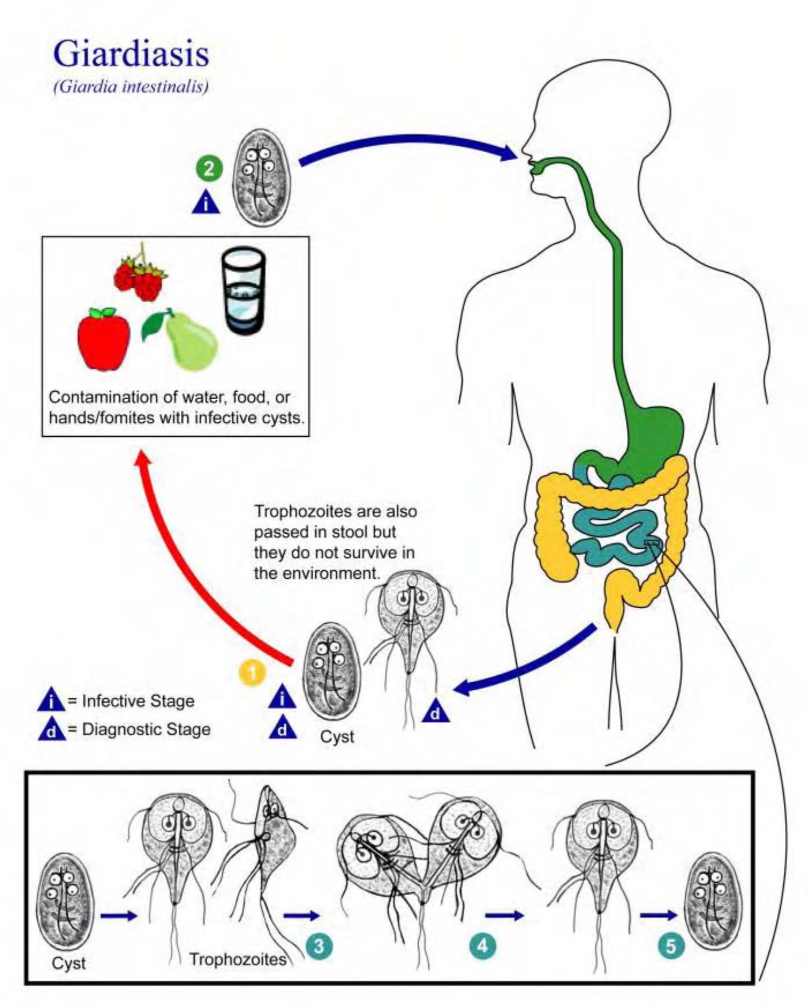
Giardiasis
(Giardia intestinalis)
2
i
Contamination of water, food, or hands/fomites with infective cysts.
Trophozoites are also passed in stool but they do not survive in the environment.
1
i
d
Cyst
d
i = Infective Stage
d = Diagnostic Stage
Cyst
Trophozoites
3
4
5

Illustration 3-1. Life cycle of *Giardia lamblia.* Cysts are resistant forms and are responsible for transmission of Giardiasis. Both cysts and trophozoites can be found in the feces (diagnostic stages) ①. The cysts are hardy and can survive several months in cold water. Infection occurs by the ingestion of cysts in contaminated water, food, or by the fecal-oral route (hands or fomites) ②. In the small intestine, excystation releases trophozoites (each cyst produces two trophozoites) ③. Trophozoites multiply by longitudinal binary fission, remaining in the lumen of the proximal small bowel where they can be free or attached to the mucosa by a ventral sucking disk ④. Encystation occurs as the parasites transit toward the colon. The cyst is the stage found most commonly in nondiarrheal feces ⑤. Because the cysts are infectious when passed in the stool or shortly afterward, person-to-person transmission is possible. While animals are infected with *Giardia*, their importance as a reservoir is unclear. (SOURCE: PHIL 3394 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology of Cysts

The cysts of *G. lamblia* are 8-12µm in length and are ellipsoid in shape. They contain four nuclei which tend not to be obvious. Longitudinal fibrils consisting of the remains of axonemes and parabasal bodies may also be seen. Cysts may appear to shrink from the cell wall. The cysts are infective as soon as they are passed.

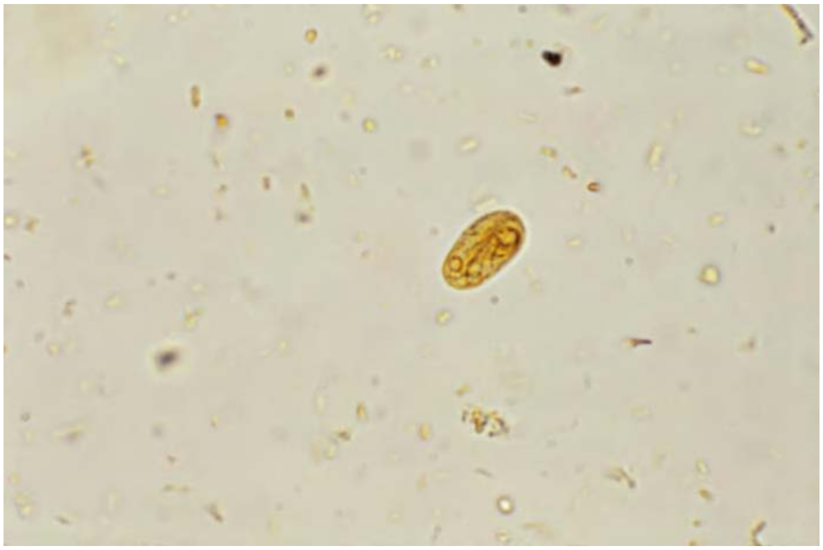

Image 3-1. *Giardia lamblia* cyst iodine stained. (SOURCE: PHIL 3741 – CDC/ Dr. Mae Melvin)

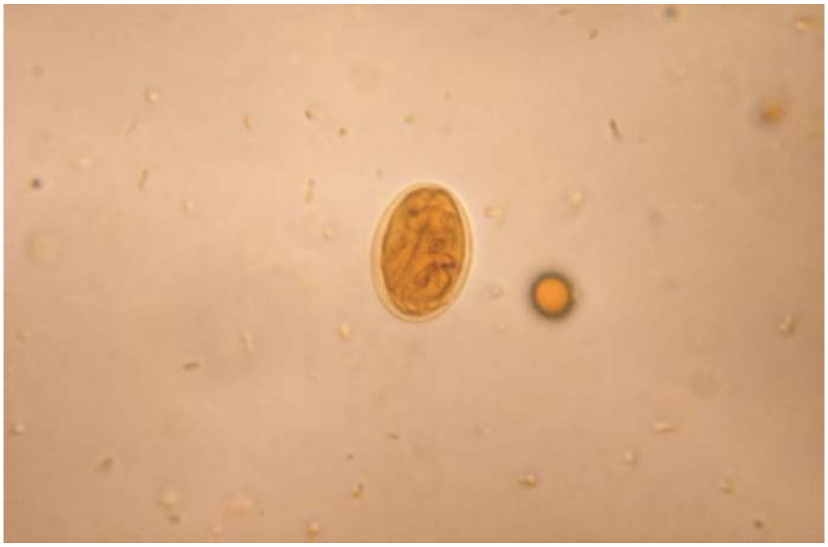

Image 3-2. *Giardia lamblia* cyst iodine stained. (SOURCE: PHIL 3742 – CDC/ Dr. Mae Melvin)

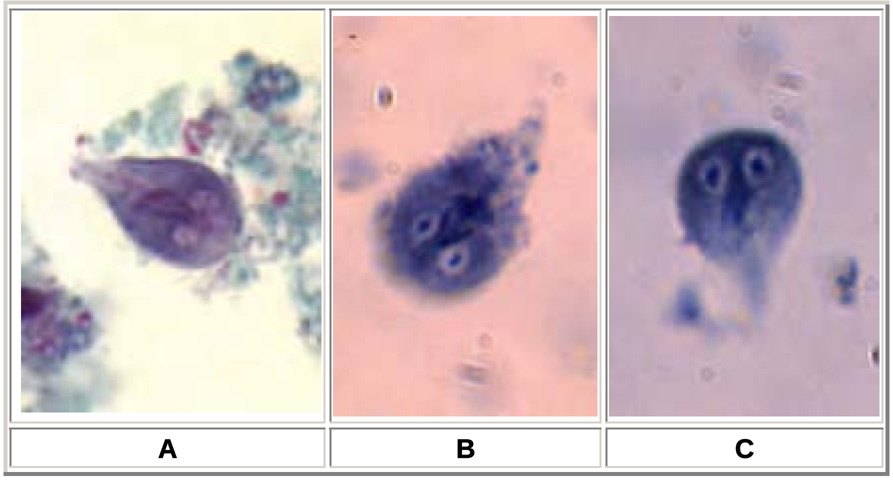

Image 3-3. A, B, C Three trophozoites of *Giardia intestinalis*, stained with trichrome (**A**) and stained with iron hematoxylin (**B** and **C**). Each cell has two nuclei with a large, central karyosome. Cell size: 9 to 21 µm in length. (SOURCE: CDC)

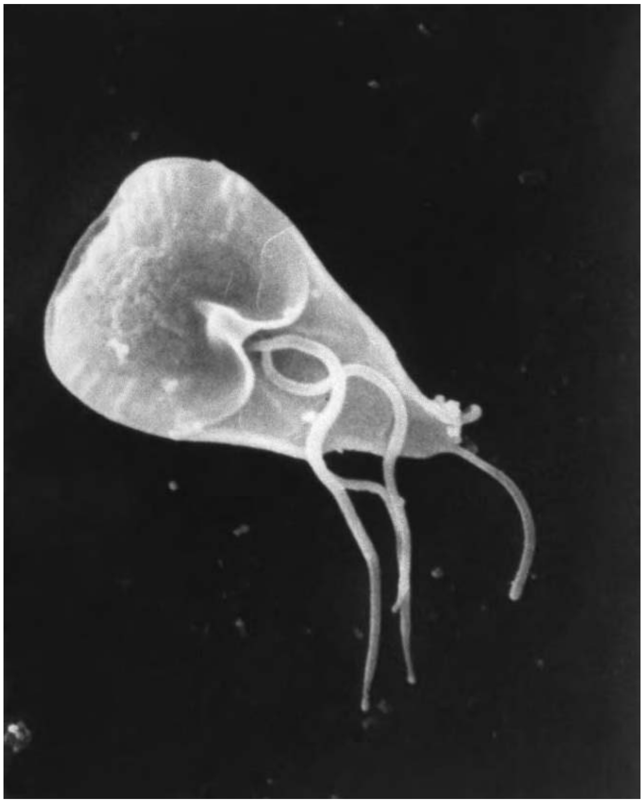

Image 3-4. This scanning electron micrograph (SEM) revealed some of the external ultrastructural details displayed by a flagellated *Giardia lamblia* protozoan parasite. *G. lamblia* is the organism responsible for causing the diarrheal disease "giardiasis". Once an animal or person has been infected with this protozoan, the parasite lives in the intestine, and is passed in the stool. Because the parasite is protected by an outer shell, it can survive outside the body, and in the environment for long periods of time. (SOURCE: PHIL 8698 - CDC / Janice Carr)

Clinical Disease

Giardia lamblia colonizes the small intestine where the trophozoites adhere to the mucosal surface by means of their sucking disc. Cysts are produced as the parasites descend the intestinal tract although trophozoites can be passed in the feces in severe infections*. G. lamblia* is transmitted through ingestion of cysts in contaminated water or food. Cysts can survive outside the body for several weeks under favorable conditions. The main symptoms are abdominal pain, flatulence, and episodic diarrhea with steatorrhea and periodical soreness in severe cases. No blood or mucus is normally seen. However 50% of *G. lamblia* infections are symptomless, although severe infections may develop in immunocompromised hosts. What determines susceptibility is poorly understood. After swallowing cysts for the first time, symptoms commonly develop 2-6 weeks later.

Laboratory Diagnosis

Cysts can be found by examination of the deposit of a formol-ether concentrate of a stool preparation. The oval cysts with thick walls serve as characteristic features for these organisms. (**Keys 3-1. & 3-2.**) The flagella disintegrate and form a central 'streak' which becomes visible when stained with iodine or MIF (merthiolate-iodine-formaldehyde). Cysts may be excreted intermittently; therefore it is important to examine more than one stool. Stools are usually passed 3-8 times / day and are usually pale, offensive, rather bulky and accompanied by much flatus.

Trophozoites are found by examination of saline wet preparations of fresh, diarrheic stool, duodenal or jejunal aspirate or in a permanently stained fecal preparation.

Trophozoites can also be found in the jejunal aspirate. These can be recovered by the String Test or Enterotest capsule and the material examined microscopically for motile trophozoites.

Trophozoites and cysts can be found to be scarce in chronic infections. Serological methods of diagnosis are proving to be useful as means of diagnosis. An ELISA to detect IgM in serum provides evidence of a current infection. A polyclonal antigen-capture ELISA can be used to demonstrate submicroscopic infections in feces and an IgA-based ELISA will detect specific antibodies in saliva. **Table 3-1.** details useful morphological features that are similar between species of flagellate and are used in laboratory diagnosis. Detection of the trophozoites and cysts can also be achieved by using rapid antigen detection testing kits that specifically detect antigen released by *Giardia lamblia.* These tests are now commercially available and offer a high degree of sensitivity even on specimens that are preserved and can be detected days before any trophozoites and cysts are shed.

Dientamoeba fragilis

Introduction

Dientamoeba fragilis is an amoeba-flagellate with a cosmopolitan distribution. The life cycle is not completely known.

Morphology of Trophozoites and Lifecycle

D. fragilis are relatively small, varying from 3-22μm in diameter and there can be considerable variation in size among organisms in the same fecal sample. The organisms have only a trophozoite stage and in a permanently stained preparation, one, two or rarely three nuclei can be seen, two being the most common. The nuclear chromatin is usually fragmented into three to five granules but these have not been visualized by Giemsa Stain, and there is normally no peripheral chromatin on the nuclear membrane. The cytoplasm is usually vacuolated and may contain ingested debris as well as some large uniform granules. The cytoplasm can also appear uniform and clean with a few inclusions. *D. fragilis* live in the lumen of the cecum and upper colon.

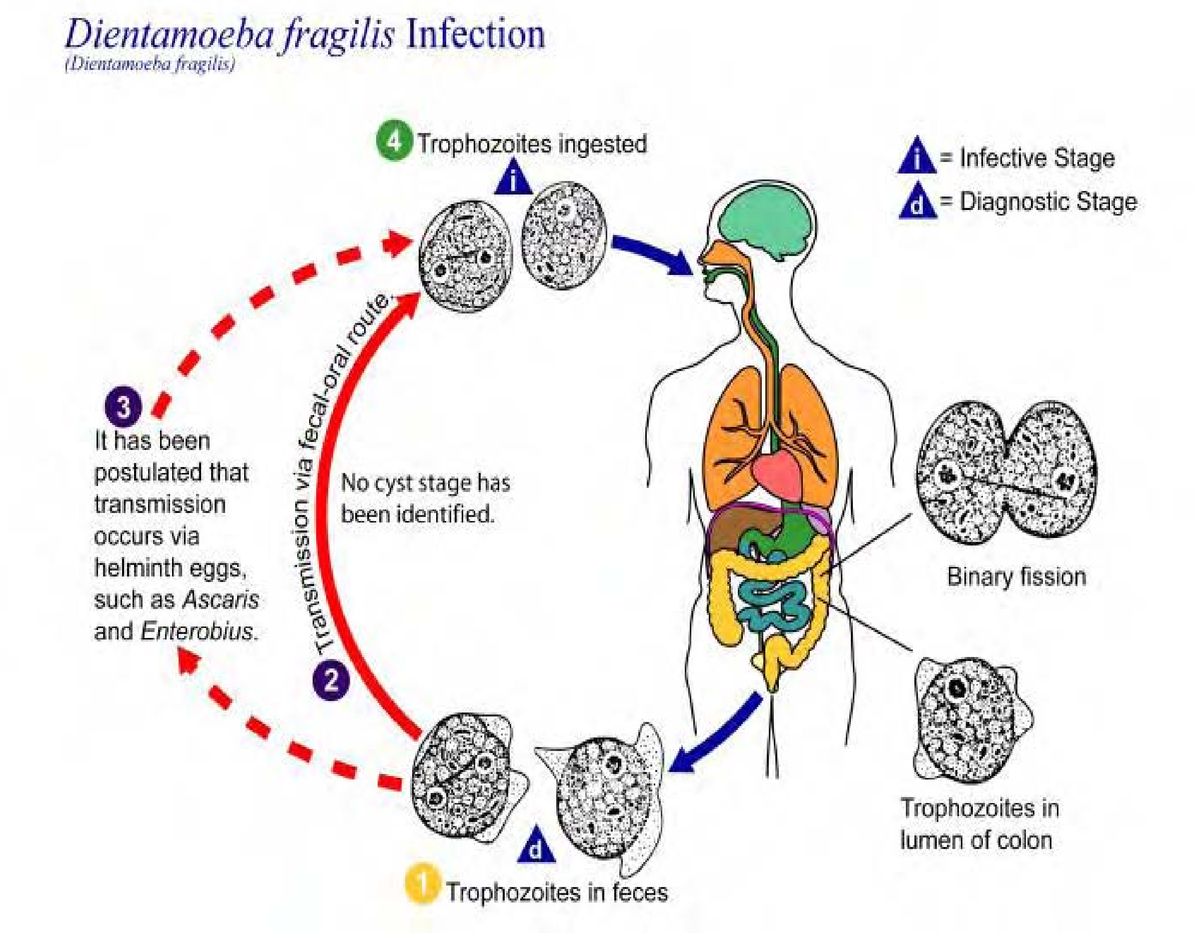
Dientamoeba fragilis Infection
(Dientamoeba fragilis)
4 Trophozoites ingested
i
i = Infective Stage
d = Diagnostic Stage
3
It has been postulated that transmission occurs via helminth eggs, such as *Ascaris* and *Enterobius*.
Transmission via fecal-oral route
No cyst stage has been identified.
2
Binary fission
Trophozoites in lumen of colon
d
1 Trophozoites in feces

Illustration 3-2. This is an illustration of the assumed life cycle of *Dientamoeba fragilis*, the cause of a protozoan parasitic infection. (SOURCE: PHIL 3389 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

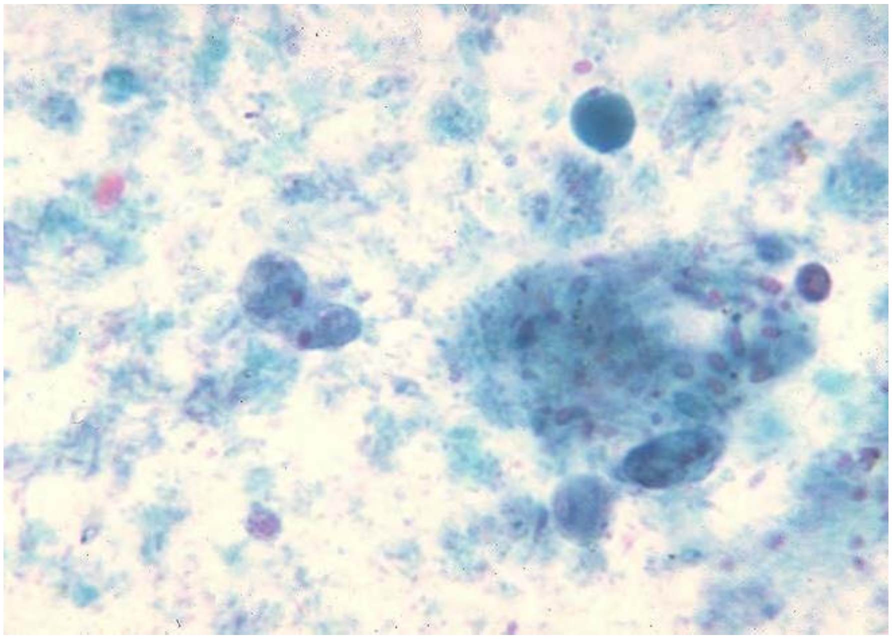

Image 3-5. *Dientamoeba fragilis*. Trichrome stain. (SOURCE: PHIL 548 – CDC)

Pathogenesis

This is a controversial area. The organism has been reported in association with mucous diarrhea, abdominal pain and tenderness. Nausea, vomiting and low-grade fever have also been reported in a number of cases. The precise role of this organism as a cause of disease remains to be determined.

Laboratory Diagnosis

Diagnosis is dependent on examination of the fresh direct wet preparation or permanently stained smears prepared from unformed or formed stools with mucus. It is particularly important that permanently stained smears of stool preparations should be examined, because survival times of the organism in terms of morphology, is very limited and specimens must be examined immediately or preserved in a suitable fixative as soon as possible after defecation. The recommended stains are Fields' and Giemsa stain (trophozoites are destroyed in a formol-ether concentrate). **Table 3-1.** details useful morphological features that are similar between species of flagellate and are used in laboratory diagnosis.

Trichomonas hominis

Introduction

This flagellate is cosmopolitan in its distribution. It is thought to be non-pathogenic although it has been associated with diarrheic stools. It is the most commonly found flagellate next to *G.*

lamblia and *D. fragilis.* Found in a wide host range including non-human primates, cats, dogs and various rodents.

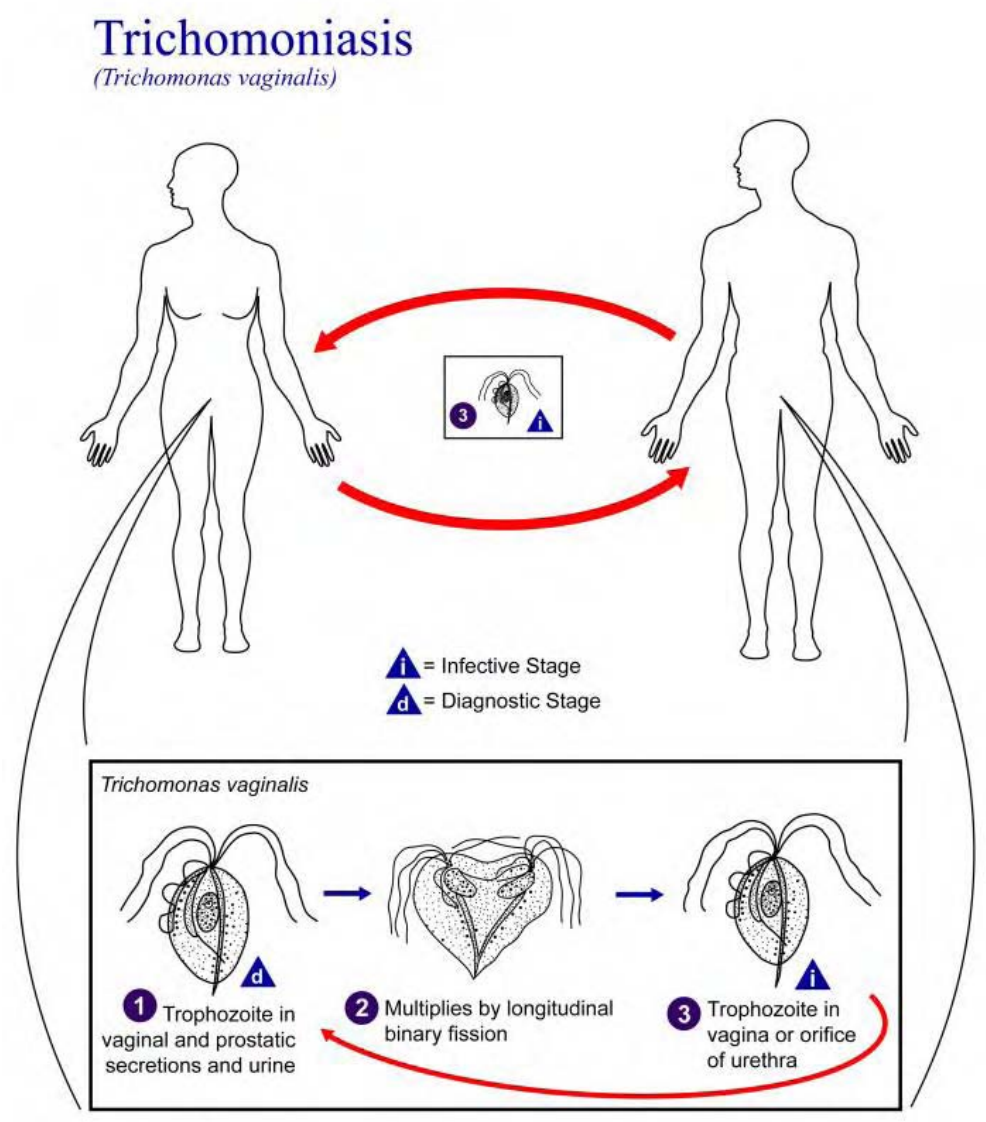
Trichomoniasis
(Trichomonas vaginalis)
3
i
i = Infective Stage
d = Diagnostic Stage
Trichomonas vaginalis
d
1 Trophozoite in vaginal and prostatic secretions and urine
2 Multiplies by longitudinal binary fission
i
3 Trophozoite in vagina or orifice of urethra

Diagram 3-3. This is an illustration of the life cycle of *Trichomonas vaginalis*, the causal agent of Trichomoniasis. (SOURCE: PHIL 3423 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology of Trophozoites

Trichomonas hominis do not have a cystic stage. The trophozoites measure from 5-15μm in length by 7-10μm in width. The shape is pyriform and has an axostyle which runs from the nucleus down the centre of the body and extends from the end of the body. They also possess an undulating membrane which extends the entire length of the body and projects from the body like a free flagellum (this feature distinguishes it from other trichomonads). The characteristic number of flagella is five; there is some deviation from this number. They also have a single nucleus at the anterior end. Trichomonads swim with a characteristic wobbly movement, which makes them unmistakable during diagnosis.

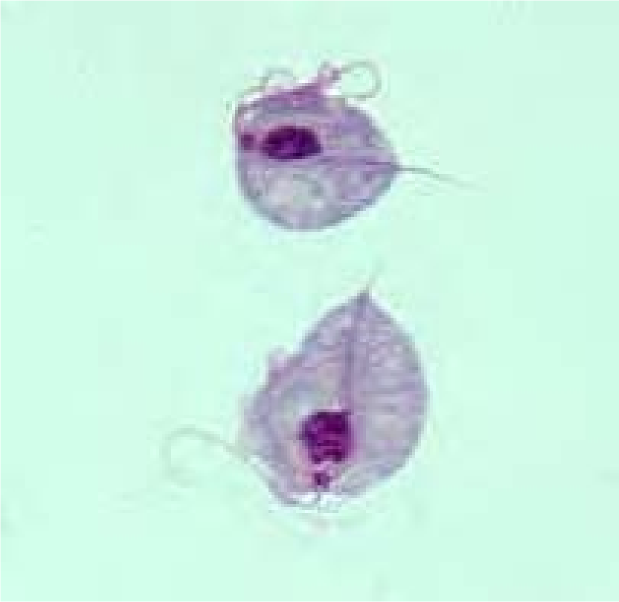

Image 3-6. Two trophozoites of *Trichomonas vaginalis* obtained from in vitro culture. Smear was stained with Giemsa. (SOURCE: CDC)

Laboratory Diagnosis

In a fresh stool, the flagellates move very rapidly in a jerky, non-directional manner. The axostyle and undulating membrane are diagnostic. The flagellates are difficult to stain; however, the axostyle can be seen on a stained preparation and is diagnostic. **Table 3-1** details useful morphological features that are similar between species of flagellate and are used in laboratory diagnosis.

Chilomastix mesnili

Introduction

Chilomastix mesnili is cosmopolitan in distribution although found more frequently in warm climates. It is thought to be non-pathogenic although the trophozoite has been associated with diarrheic stool. This is the largest flagellate found in man with an incidence of 1-10% being in the large intestine.

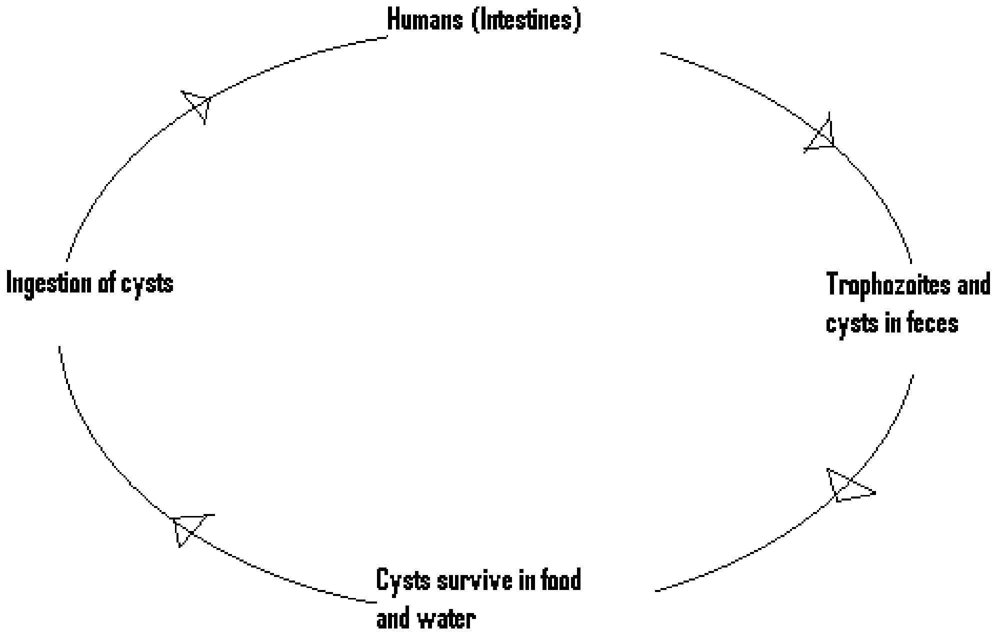
Humans (Intestines)
Trophozoites and cysts in feces
Cysts survive in food and water
Ingestion of cysts

Illustration 3-4. The life cycle of *C. mesnili* (Cuomo)

Morphology of the Trophozoite

The trophozoites of *C. mesnili* are pear shaped and measure 6-20μm in length. They have one large nucleus with a small karyosome and three flagella that extend from the nucleus at the anterior end of the parasite. A distinct oral groove or cytosome can be seen near the nucleus with its sides being supported by two filaments. They are known to move in a directional manner.

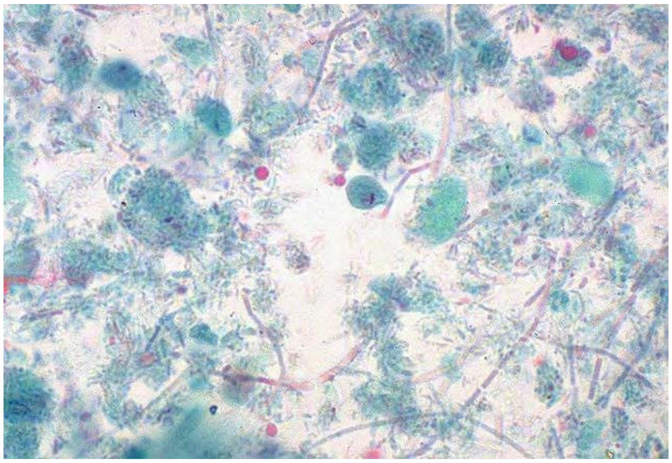

Image 3-7. *Chilomastix mesnili* cyst.(SOURCE: PHIL 426 - CDC /Dr. Mae Melvin)

Morphology of cysts

The cysts are 6-9μm; they have a large single nucleus with a large karyosome. They also have a prominent side knob giving it a characteristic lemon shape. The cytosome is evident with a curved shepherds crook fibril. It also has a characteristically coiled filament which when stained is darker in color.

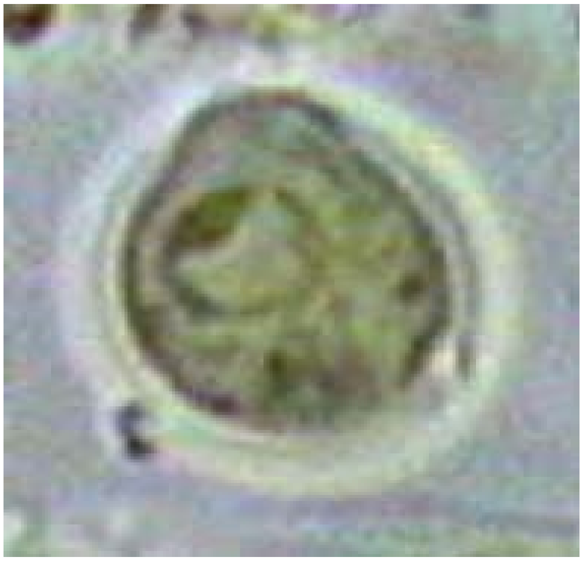

Image 3-8. *Chilomastix mesnili* cysts are excreted with feces and constitute the transmission form of the micro-organism. The uninucleated lemon shaped cysts are seen with a little proturberance at one end and a prominent cytostome. (Iodine stained). (6µm) (CDC)

Laboratory Diagnosis

The characteristic lemon shaped cysts can be seen in a formol-ether concentrate. Motile organisms can be seen in a wet preparation of a fresh stool however the characteristic morphology is evident in a permanently stained preparation. **Table 3-1** details useful morphological features that are similar between species of flagellate and are used in laboratory diagnosis.

Enteromonas hominis

Introduction

Enteromonas hominis is a small flagellate and is rarely encountered in man. It is found in both warm and temperate climates and is considered to be non-pathogenic.

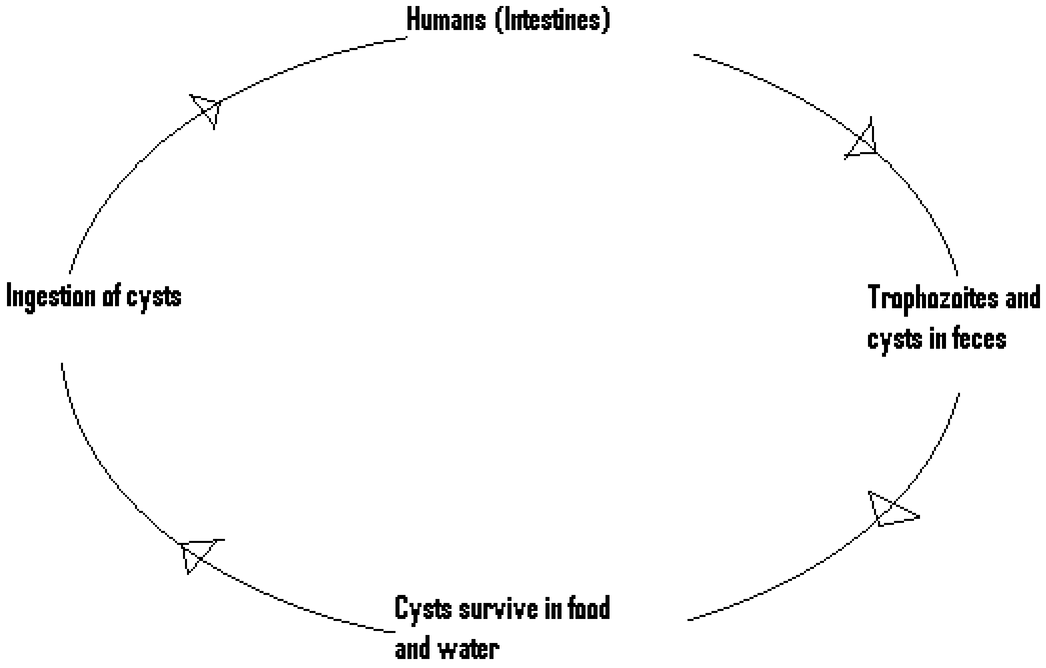
Humans (Intestines)
Trophozoites and cysts in feces
Cysts survive in food and water
Ingestion of cysts

Illustration 3-5. The life cycle of *E. hominis* (Cuomo)

Morphology of the Trophozoite

The trophozoites are oval and 4-10μm in length. They have four flagella, three anterior flagella and one adheres to the body ending in a tail, producing a jerky rotational movement. They have one nucleus with a large karyosome that is evident in a stained preparation.

Morphology of the Cyst

The cysts are oval and range between 6-8μm in length. They have up to four nuclei with a bipolar tendency.

Laboratory Diagnosis

The cysts are seen in a formol-ether concentrate. The cysts have no distinguishing characteristics and thus can be confused with *E. nana* or even yeasts. The characteristic trophozoites can be seen in a permanently stained fecal smear. **Table 3-1.** details useful morphological feature's that are similar between species of flagellate and are used in laboratory diagnosis.

Retortamonas intestinalis

Introduction

Retortamonas intestinalis like *Enteromonas hominis* is a small flagellate and is rarely encountered. It is found in both warm and temperate climates and is considered to be non-pathogenic.

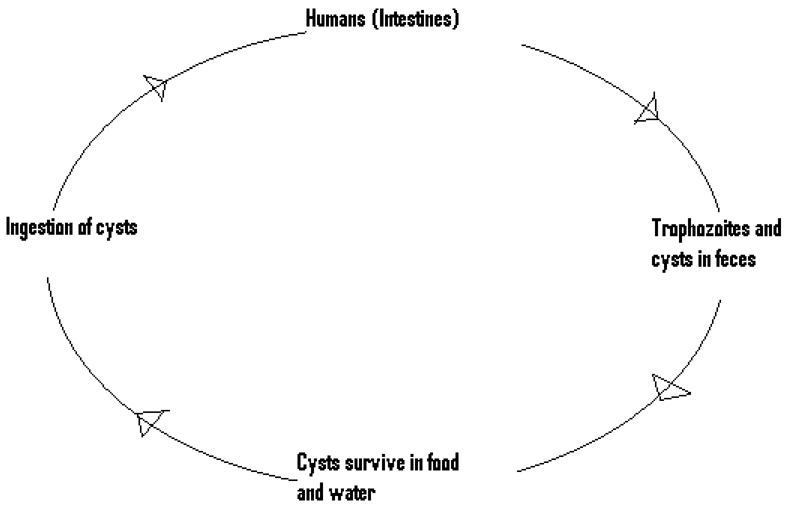
Humans (Intestines)
Trophozoites and cysts in feces
Cysts survive in food and water
Ingestion of cysts

Illustration 3-6. Life cycle of *Retortamonas intestinalis* (Cuomo)

Morphology of the Trophozoite

The trophozoite is small, measuring between 4µm and 9µm. Its movement is jerky and rotational and has two anterior flagella and a prominent cytosome that can be seen in an unstained preparation. It has a relatively large nucleus at the anterior end with a small compact karyosome.

Morphology of the Cyst

The cysts are small and pear shaped. They range in size between 4-7µm with one large nucleus frequently near the centre. The fibril arrangement from the nucleus is suggestive of a birds beak. This is characteristic of *R. intestinalis* cysts.

Laboratory Diagnosis

The small pear shaped cysts are uncharacteristic in an unstained formol-ether preparation. However, the addition of iodine reveals the characteristic bird beak fibrillar arrangement in the pear shaped cyst. In a fresh stool preparation, the two anterior flagella and cytosome can be seen in the trophozoite. In a permanently stained preparation, the large nucleus with small central karyosome is diagnostic. The below Tables detail useful morphological features which are similar between species of flagellate and are used in laboratory diagnosis.

3.2 Identifying Flagellates

It is important to know and understand the morphological features which differentiate each species of flagellate from one other. The below Tables details the important features that are used when identifying flagellates found in human stool samples. Trophozoites and cysts can be seen in saline mounts of fresh feces. On occasions, species identification may require stained preparations.

Species	Size (Length)	Shape	Motility	Number of Nuclei	Number of Flagella*	Other Features
Trichomonas hominis	8-20 µm. Usual range. 11-12 µm.	Pear shaped.	Nervous, jerky.	1 Not visible in unstained mounts.	3-5 anterior. 1 posterior.	Undulating membrane extending length of body.
Chilomastix mesnili	6-24 µm. Usual range, 10-15 µm.	Pear shaped.	Stiff, rotary.	1 Not visible in unstained mounts.	3 anterior. 1 in cytosome.	Prominent cytostome extending 1/3-1/2 length of body. Spiral groove across ventral surface.
Giardia intestinalis	10-20 µm. Usual range, 12-15 µm.	Pear shaped.	"Falling leaf."	2 Not visible in unstained mounts.	4 lateral. 2 ventral. 2 caudal.	Sucking disk occupying 1/2-3/4 of ventral surface. Median bodies lying horizontally or obliquely in lower part of body.
Enteromonas hominis	4-10 µm. Usual range, 8-9 µm.	Oval.	Jerky.	1 Not visible in unstained mounts.	3 anterior. 1 posterior.	One side of body flattened. Posterior flagellum extends free posteriorly or laterally.
Retortamonas intestinalis	4-9 µm. Usual range, 6-7 µm.	Pear shaped or oval.	Jerky.	1 Not visible in unstained mounts.	1 anterior. 1 posterior.	Prominent cytostome extending approximately 1/2 length of body.

* Not a normal feature for identifying species in routine stool samples

Table 3-1. Differential Morphology of Protozoa Found in Stool Specimens of Humans: Flagellates-Trophozoites (SOURCE: CDC)

Species	Size (Length)	Shape	Number of Nuclei	Other Features
Trichomonas hominis	No cyst.			
Chilomastix mesnili	6-10 µm. Usual range, 8-9 µm.	Lemon shaped with anterior hyaline knob.	1. Not visible in unstained preparations.	Cytostome with supporting fibrils. Usually visible in stained preparations.
Giardia intestinalis	8-19 µm. Usual range. 11-12 µm.	Oval or ellipsoidal.	Usually 4. Not distinct in unstained preparations. Usually located at one end.	Fibrils or flagella longitudinally in unstained cysts. Deep staining fibers or fibrils may be seen lying laterally or obliquely across fibrils in lower part of cyst. Cytoplasm often retracts from a portion of cell wall.
Enteromonas hominis	4-10 µm. Usual range, 6-8 µm.	Elongated or oval.	1-4, usually 2 lying at opposite ends of cyst. Not visible in unstained mounts.	Resembles *E. nana* cyst. Fibrils or flagella are usually not seen.
Retortamonas intestinalis	4-9 µm. Usual range, 4-7 µm.	Pear shaped or slightly lemon shaped.	1. Not visible in unstained mounts.	Resembles *Chilomastix* cyst. Shadow outline of cytostome with supporting fibrils extends above nucleus.

Table 3-2. Differential Morphology of Protozoa Found in Stool Specimens of Humans: Flagellates-Cysts (SOURCE: CDC)

Flagellate trophozoites are best identified in fresh saline mounts, allowing you to observe the way that they move. Use **Key 3-1.** to help to identify stained flagellate trophozoites. Iodine solutions are used primarily to stain flagellate cysts, this makes it possible to see the structure of the nuclei. Use **Key 3-2.** to help to identify amebic and flagellate cysts:

Key 3-1. Flagellates (Adapted and redrawn, WHO, 1991)

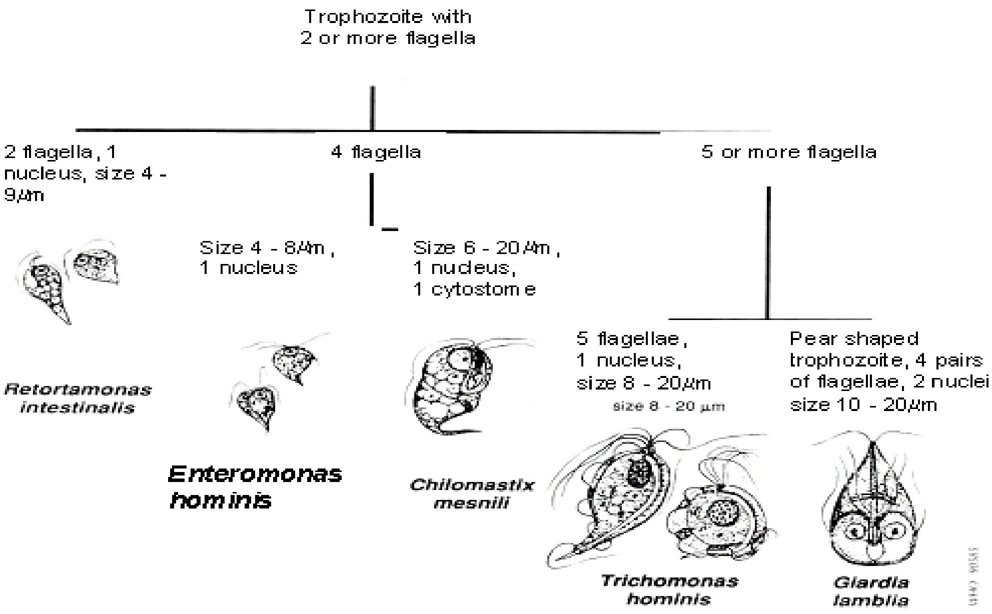
Trophozoite with
2 or more flagella
2 flagella, 1 nucleus, size 4 - 9µm
4 flagella
5 or more flagella
Size 4 - 8µm, 1 nucleus
Size 6 - 20µm, 1 nucleus, 1 cytostome
5 flagellae, 1 nucleus, size 8 - 20µm
size 8 - 20 µm
Pear shaped trophozoite, 4 pairs of flagellae, 2 nuclei, size 10 - 20µm
Retortamonas intestinalis
Enteromonas hominis
Chilomastix mesnili
Trichomonas hominis
Giardia lamblia

Key 2. Key to differentiating flagellate and amoebic cysts. (Adapted and redrawn, WHO, 1991)

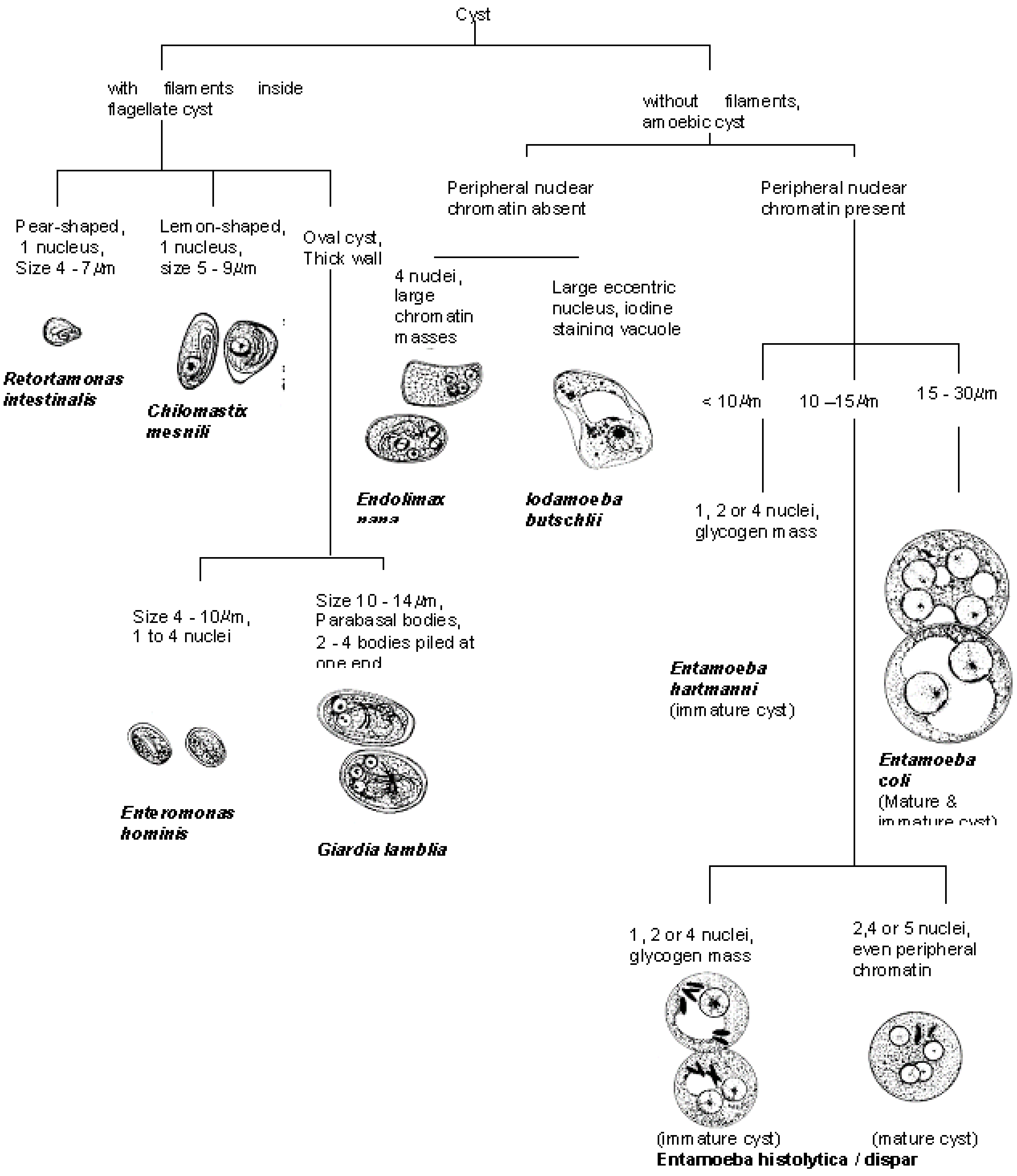
Cyst
with filaments inside flagellate cyst
without filaments, amoebic cyst
Pear-shaped, 1 nucleus, Size 4 - 7 µm
Lemon-shaped, 1 nucleus, size 5 - 9 µm
Oval cyst, Thick wall
Retortamonas intestinalis
Chilomastix mesnili
Size 4 - 10 µm, 1 to 4 nuclei
Size 10 - 14 µm, Parabasal bodies, 2 - 4 bodies piled at one end
Enteromonas hominis
Giardia lamblia
Peripheral nuclear chromatin absent
4 nuclei, large chromatin masses
Large eccentric nucleus, iodine staining vacuole
Endolimax nana
Iodamoeba butschlii
Peripheral nuclear chromatin present
< 10 µm
10 –15 µm
15 - 30 µm
1, 2 or 4 nuclei, glycogen mass
Entamoeba hartmanni (immature cyst)
Entamoeba coli (Mature & immature cyst)
1, 2 or 4 nuclei, glycogen mass
2,4 or 5 nuclei, even peripheral chromatin
(immature cyst)
(mature cyst)
Entamoeba histolytica / dispar

Chapter 4.
The Cestodes

4.1 Identifying Intestinal Helminths

Helminth Parasites - The word "worm" is used loosely to describe organisms with elongated bodies and a more or less creeping habit.

Although the word "Helminth" does mean "worm," in zoological terms it is more restricted to members of the phyla Platyhelminths, Nematoda, and Acanthocephala.

There are three groups of medically important helminths; Cestodes (tapeworms), Nematodes (roundworms) and Trematodes (flukes).

These parasites live in both the body spaces (gut lumen, bile ducts, lungs, oral cavity, etc.) and in tissues (blood, muscles and skin).

4.2 Infections Acquired Through the Gastrointestinal Tract

The cestodes (or tapeworms) form a group of worms, exhibiting two unmistakable morphological features; they all possess flat, ribbon like bodies and lack an alimentary canal.

Adult tapeworms usually inhabit the alimentary canal of their hosts (though they occasionally are found in the bile or pancreatic ducts) and attach themselves to the mucosa by means of a scolex. Despite the lack of a digestive system they do absorb food from the hosts intestine; thereby providing the tapeworms a habitat that is associated with high nutritional levels, feeding the tapeworms high growth rate.

Larvae on the other hand show a wide range of habitat preferences, being found in almost any organ of both vertebrate and invertebrate hosts. Though most larval species show a preference for a particular site.

This lack of an alimentary canal markedly separates tapeworms from nematodes and trematodes. The outer tegument of the body must serve not only as a protective coating but also as a metabolically active layer through which nutritive material can be absorbed, along with secretions and waste material to be transported out of the body.

The body consists of a chain of segments or proglottids, which can be immature, mature or gravid; the latter of which contain a fully developed uterus packed with eggs. Therefore, each tapeworm is made up of a ‘string of individuals’ having a complete set of reproductive organs in progressive degrees of sexual maturity and budding off from a body attached to the host tissue by a head or scolex.

Except for *Hymenolepis nana,* which can develop directly in the same host, the lifecycle of tapeworms involves both an intermediate and definitive host.

Species	**Size**	**Shape**	**Color**	**Stage of Development When Passed**	**Specific Features and Variations**
Cestodes					
Taenia saginata ***Taenia solium***	35 µm. Range, 31-43 µm.	Spherical with thick striated shell.	Walnut brown.	Embryonated. 6-hooked oncosphere present inside a thick shell.	Thick, striated shell. Eggs of *T. solium* and *T. saginata* are indistinguishable and species identification should be made from proglottids or scoleces. "*Taenia*" spp. should be reported if only eggs are found.
Hymenolepis nana	47 µm x 37 µm. Range, 40-60 µm x 30-50 µm.	Oval. Shell consists of 2 distinct membranes. On inner membrane are two small "knobs" or poles from which 4 to 8 filaments arise and spread out between the two membranes.	Colorless, almost transparent.	Embryonated. 6-hooked oncosphere inside shell.	Polar filaments.
Hymenolepis diminuta*	72 µm. Range, 70-86 µm x 60-80 µm.	Round or slightly oval. Striated outer membrane and thin inner membrane with slight poles. Space between membranes may appear smooth or faintly granular.	Yellow.	Embryonated. 6-hooked oncosphere inside shell.	Resembles *H. nana* but lacks polar filaments. Poles are rudimentary and often hard to see.
Dipylidium caninum*	35-40 µm. Range, 31-50 µm x 27-48 µm.	Spherical or oval. 5-15 eggs (or more) are enclosed in a sac or capsule.	Colorless.	Embryonated. 6-hooked oncosphere inside shell.	Eggs are contained in a sac or capsule which ranges in size from 58 µm to 60 µm x 170 µm. Occasionally capsules are ruptured and eggs are free.
Diphyllobothrium latum	66 µm x 44 µm. Range, 58-76 µm x 40-51 µm.	Oval or ellipsoidal with an inconspicuous operculum at one end and a small "knob" at the other end.	Yellow to brown.	Unembryonated. Germinal cell is surrounded by a mass of yolk cells which completely fills inner area of shell. Germinal cell is usually not visible.	Egg resembles hookworm egg but has a thicker shell and an operculum.

Table 4-1. Differential Morphology of the Diagnostic Stages of Helminths Found in Humans: Eggs (Cestodes) (SOURCE: CDC)

Taenia species

Introduction

Taenia species are the most common cestode parasites of humans. More than 60 million people are infected with *T. saginata* (beef tapeworm) world wide and about four million are infected with *T. solium* (pork tapeworm). The life cycle of a *Taenia* species can be seen in **Illustration 4-1.** *T. saginata* has a cosmopolitan distribution, but is more common in developing countries where hygiene is poor and the inhabitants have a tendency of eating raw or insufficiently cooked meat. *T saginata* is the most common adult tapeworm found in man. *T solium* is virtually extinct in Europe and the USA.

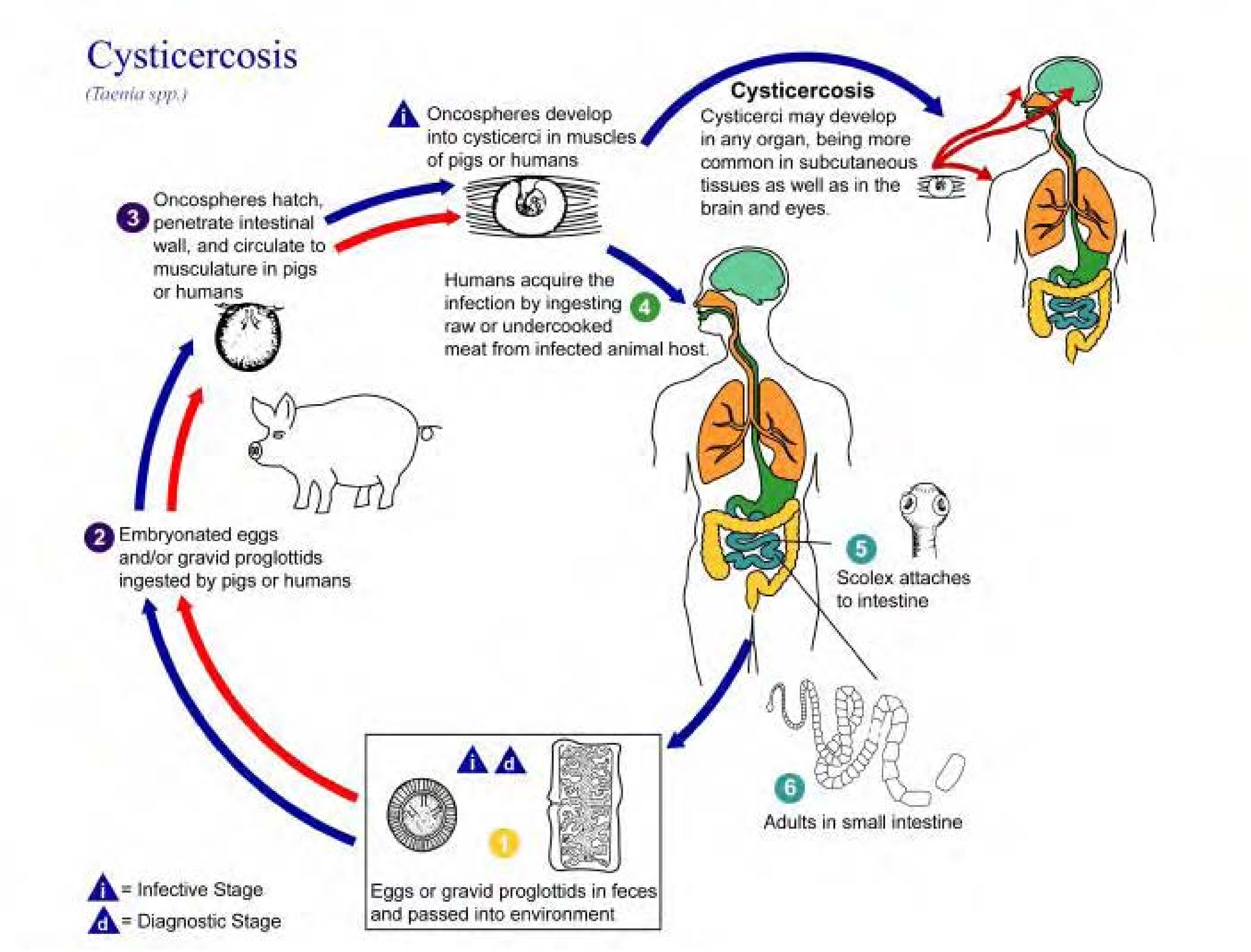
Cysticercosis
(Taenia spp.)
Oncospheres develop into cysticerci in muscles of pigs or humans
Cysticercosis
Cysticerci may develop in any organ, being more common in subcutaneous tissues as well as in the brain and eyes.
3 Oncospheres hatch, penetrate intestinal wall, and circulate to musculature in pigs or humans
Humans acquire the infection by ingesting raw or undercooked meat from infected animal host. 4
2 Embryonated eggs and/or gravid proglottids ingested by pigs or humans
5 Scolex attaches to intestine
6 Adults in small intestine
i d
1
Eggs or gravid proglottids in feces and passed into environment
i = Infective Stage
d = Diagnostic Stage

Illustration 4-1. This is the life cycle of *Taenia spp.*, the causal agents of Cysticercosis. Cysticercosis is an infection of both humans and pigs with the larval stages of the parasitic cestode, *Taenia solium*. This infection is caused by ingestion of eggs shed in the feces of a human tapeworm carrier ①. Pigs and humans become infected by ingesting eggs or gravid proglottids ❷, ❼. Humans are infected either by ingestion of food contaminated with feces, or by autoinfection. In the latter case, a human infected with adult *T. solium* can ingest eggs produced by that tapeworm, either through fecal contamination or, possibly, from proglottids carried into the stomach by reverse peristalsis. Once eggs are ingested, oncospheres hatch in the intestine ❸, ❽invade the intestinal wall, and migrate to striated muscles, as well as the brain, liver, and other tissues, where they develop into cysticerci ❾. In humans, cysts can cause serious sequellae if they localize in the brain, resulting in neurocysticercosis. The parasite life cycle is completed, resulting in human tapeworm infection, when humans ingest undercooked pork containing cysticerci ❹. Cysts evaginate and attach to the small intestine by their scolex ❺. Adult tapeworms develop, (up to 2 to 7 m in length and produce less than 1000 proglottids, each with approximately 50,000 eggs) and reside in the small intestine for years ❻. (SOURCE: PHIL 3387 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Both humans and cattle or pigs are necessary to the complete life cycle of *Taenia* species. (In Europe and the USA cattle are the normal intermediate hosts, but in the tropics several other ruminants, e.g. goat, sheep llama and giraffe, may serve as the intermediate hosts.) Eggs ingested by the intermediate hosts usually contain oncospheres. The oncospheres then hatch out in the duodenum, pass into the intestine where they penetrate the intestinal wall and are then carried by the circulation to be deposited in tissues (usually muscle). There they develop into cysticerci larva which are white and ovoid, measuring approximately 8 x 5µm.

Humans become infected by ingesting inadequately, cooked beef or pork with cysticerci, containing an invaginated protoscolex. The protoscolexes evaginate and pass into the small intestine where they attach themselves to the mucosa and develop into adult worms.

Eggs and proglottids are passed out in the feces, and are then eaten by the intermediate host, thus, perpetuating the life cycle.

Morphology

Ova of *Taenia* species are spherical, yellowish brown and measure 31-34µm in diameter. The shell is thick and radially striated. Within the shell, the oncosphere has 3 pairs of hooklets. However, the microscopic appearance of the ova of *T. saginata* and *T. solium* are identical. (**Table 4-1.** highlights some of the differences between the two species)

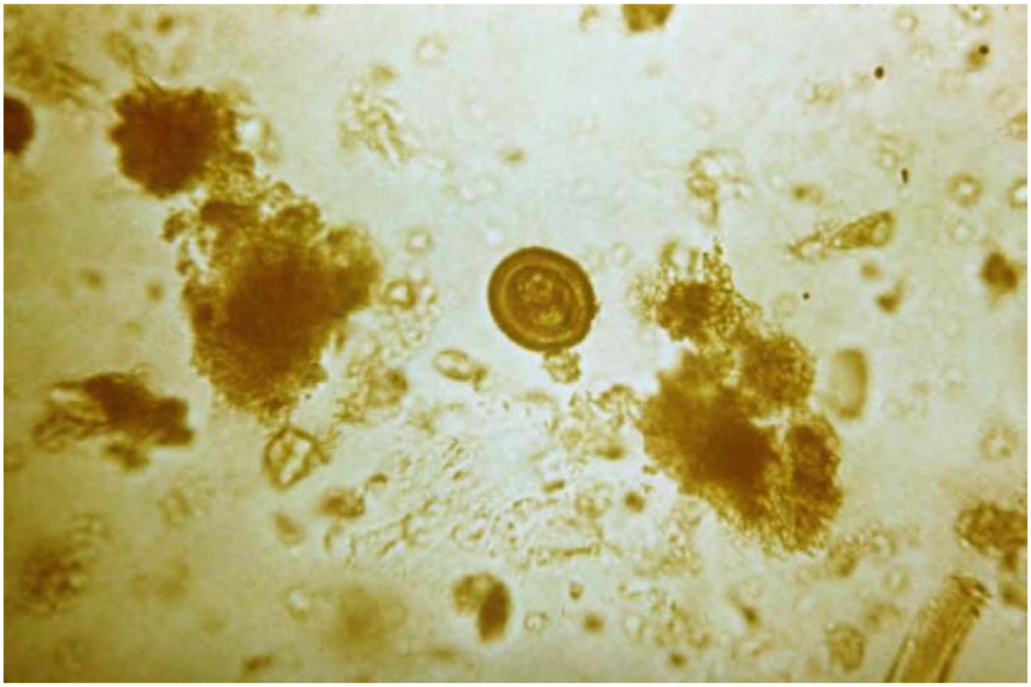

Image 4-1. The eggs *Taenia saginata* and *Taenia solium* are rounded or subspherical, with a thick radially striated brown shell. The diameter is 31 - 43 µm. Inside each shell is an embryonated oncosphere with 6 hooks. (SOURCE: PHIL 4832 – CDC)

The length of the adult *T. saginata* is 4-8 meters long and that of *T. solium* is 3-5 meters long.

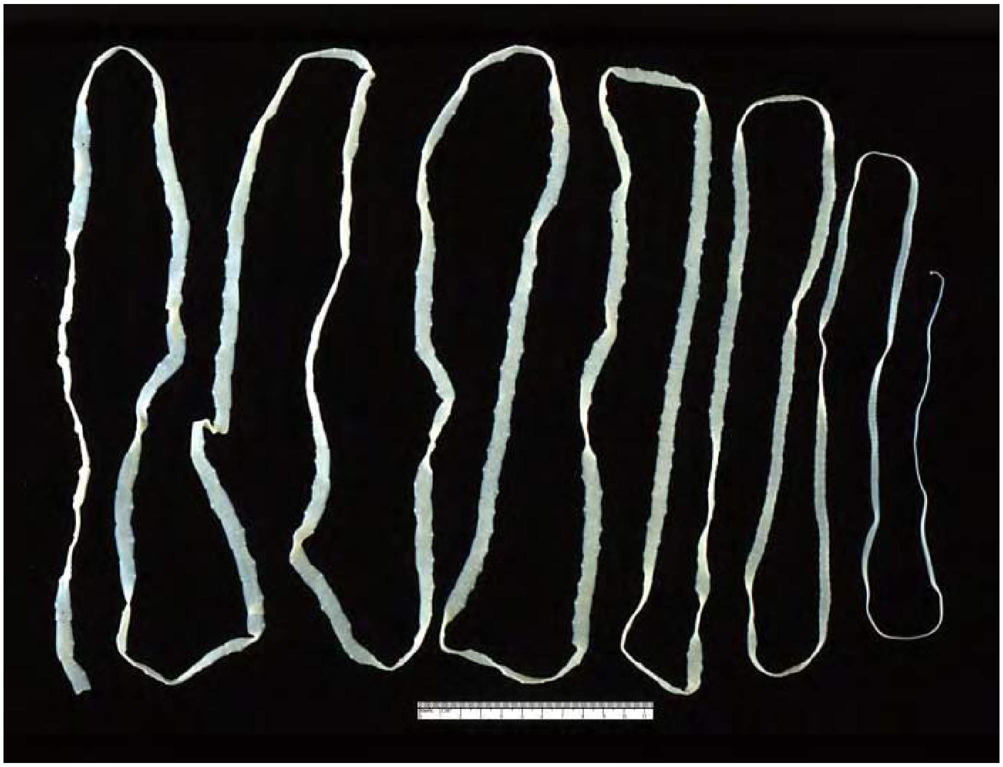

Image 4-2. This is an adult *Taenia saginata* tapeworm. (SOURCE: PHIL 5260 – CDC)

The proglottids of *Taenia* species can be identified by the number of uterine branches; 7-13 for *T. solium* and 15-20 for *T. saginata*.

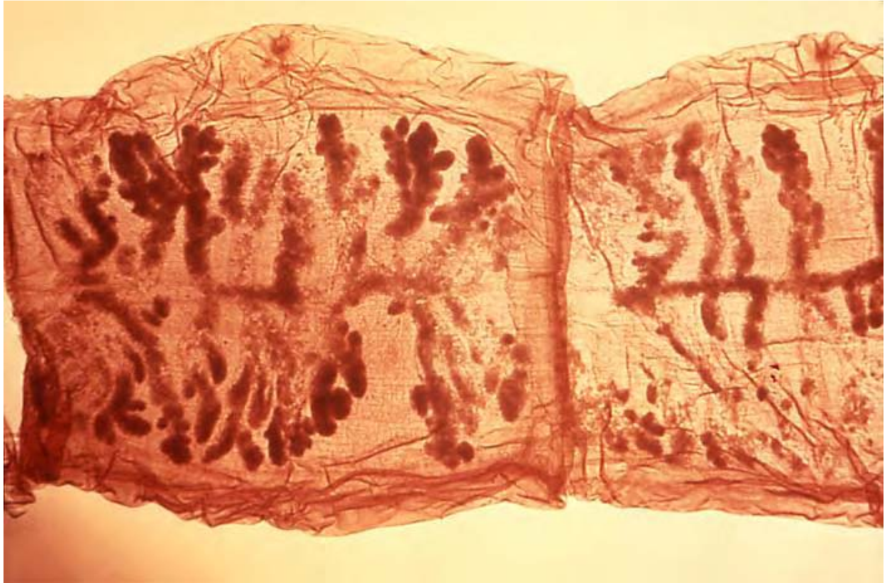

Image 4-3. This micrograph reveals the organ morphology within *Taenia solium* tapeworm proglottids. (SOURCE: PHIL 5261 – CDC)

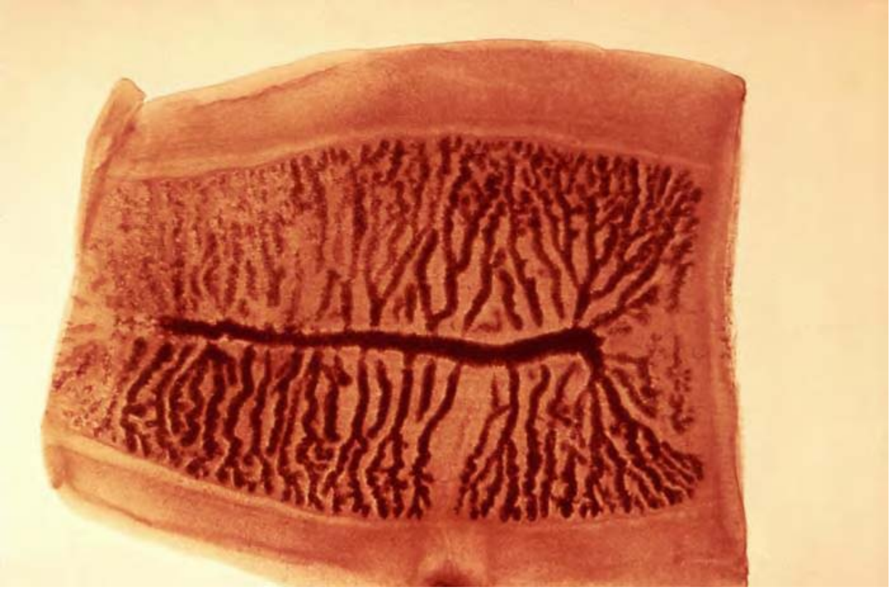

Image 4-4. This micrograph reveals the morphology of a gravid proglottid from the cestode *Taenia saginata*, a tapeworm. (SOURCE: PHIL 5259 – CDC)

If the scolex is recovered, the four suckers and rostellum of hooklets of *T. solium* will distinguish it from *T. saginata,* which has four suckers but no hooklets.

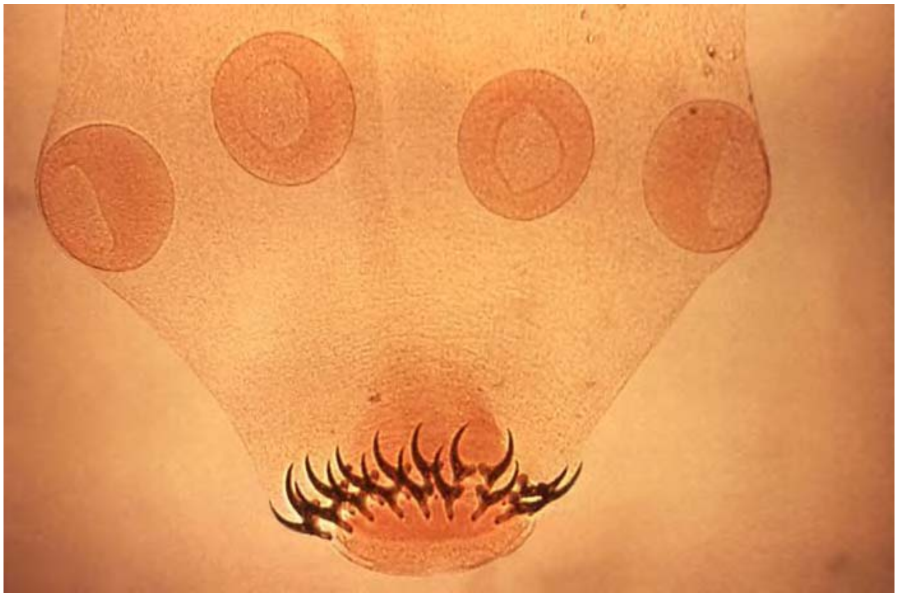

Image 4-5. This is an adult *Taenia solium* taperworm scolex. (SOURCE: PHIL 5262 – CDC)

Clinical Disease

The presence of the adult worm rarely causes symptoms apart from slight abdominal irritation with diarrhea, constipation or indigestion. The accidental ingestion of the embryonated ova of *T. solium* may result in cysticercosis in man. An infection due to an adult *Taenia,* in man or animals, is referred to as *taeniasis.* *T. saginata* (the beef tapeworm) **does not** cause human cysticercosis.

When the embryonated eggs are ingested, the embryos hatch out, migrate through the intestinal wall and are carried around the body in the circulation and deposited in various tissues. Muscle and subcutaneous tissues are usually infected, but cysticerci can infect most organs and tissues. Human cysticercosis is usually asymptomatic unless the infection is particularly heavy or cysticerci are formed in some vital area e.g. the brain, resulting in neurological sequelae.

Table 4-2. Some characteristics differentiating *T. saginata* from *T. solium*.

Characteristic	*Taenia saginata*	*Taenia solium*
Intermediate Host	Cattle, reindeer	Pig, wild boar
Site of Development	Muscle, viscera	Brain, skin, muscle
Scolex: adult worm	No hooks	Hooks
Scolex: cysticercus	No rostellum	Rostellum & hooks
Proglottids: uterine branches	23 (14 – 32) *	8 (7 –11) *
Passing of proglottids	Single, spontaneous	In groups, passively
Ovary	Two lobes	Three lobes
Vagina: sphincter muscle	Present	Absent

***** No universal agreement to the number of uterine branches in these two species. As a rough guide, specimens with more than 16 branches are likely to be those of *T. saginata* and those with less than ten branches are likely to be of *T. solium*. (SOURCE: CDC)

Laboratory Diagnosis

Diagnosis of intestinal taeniasis can be made by recovery of the characteristic ova in the stool. However, the ova of *T. solium* and *T. saginata* are identical and diagnosis is made by the recovery of the segments or scolex.

The diagnosis of cysticercosis depends upon serology. MRI scans may reveal the presence of lesions in the brain. Calcified cysticerci are less often seen in the brain: in about one-third of cases, 10 years or more after infection. Occasionally, the diagnosis is made histologically on surgical specimens. Calcification in muscles usually appears three to five years after initial infection, and are most typically seen as spindle-shaped calcifications, most numerous in the thighs.

Western Blots

Various Immunodiagnostic tests appear to give good results on serum or CSF.

Diagnosis using an immunodiagnostic test can be achieved using an *in vitro* qualitative assay for the detection of IgG antibodies in serum reactive with *T. solium* antigens present on a membrane.

Infected individuals develop a predominately IgG response to the parasite. ELISA has been used as a screening test, but low sensitivity and frequent artifactual cross reactions, or cross reactions with antibodies from other parasitic infections, limit its usefulness as a confirmatory diagnostic test. The Western Blot assay (US Patent No. 5,354,660) developed by Tsang *et al*, at the U.S. Centers for Disease Control has been shown to provide a reliable method for evaluation of sera from patients with clinically diagnosed active cysticercosis. Field studies support a sensitivity of 98% and specificity of 100% for this assay.

This assay is known as the QualiCode^{TM} Cysticercosis Kit, the principle behind the test is that it is a qualitative membrane-based immunoassay manufactured from *T. solium* proteins. The *T. solium* proteins are fractionated according to molecular weight by electrophoresis on a polyacrylamide slab gel (PAGE) in the presence of sodium dodecyl sulfate (SDS). The separated *T. solium* proteins are then transferred via electrophoretic blotting from the gel to a nitrocellulose membrane. This antigen-bearing membrane has been cut into strips for testing of individual samples. Sera are tested at 100X dilution.

During the procedure, the strips containing the *T. solium* proteins are incubated with serum specimens and washed to remove unbound antibodies. Visualization of human immunoglobulins specifically bound to *T. solium* proteins is performed by sequential reaction with goat anti-human immunoglobulin-alkaline phosphatase conjugate and BCIP/NBT substrate. Bands corresponding to the positions of the resoled *T. solium* proteins will be visualized on the strip, indicating the presence in the serum sample of IgG antibodies direct against *Taenia* antigens. Band positions are compared to those on a reference strip developed using the cysticercosis positive control.

Hymenolepis nana

Introduction

Hymenolepis nana, the dwarf tapeworm, is the smallest tapeworm to infect humans. This cestode belongs to a large family known as Hymenolepididae. The diagnostic features of this family are: scolex armed with one circlet of five hooks; one to three large testes and sacciform uterus. In addition to the *H. nana*, three other species, *H. diminuta, H. microstoma* and *H. citelli* have been used extensively for studies on cestodes.

Hymenolepis nana has a cosmopolitan distribution and is thought to be the most common tapeworm throughout the world. The infection is more frequently seen in children although adults are also infected, causing *hymenolepiasis*.

Life Cycle

The lifecycle of *H. nana* does not require an intermediate host, complete development occurring within the villi of a single host, resulting in a ‘direct’ life cycle. It can also utilize an insect as an intermediate host.

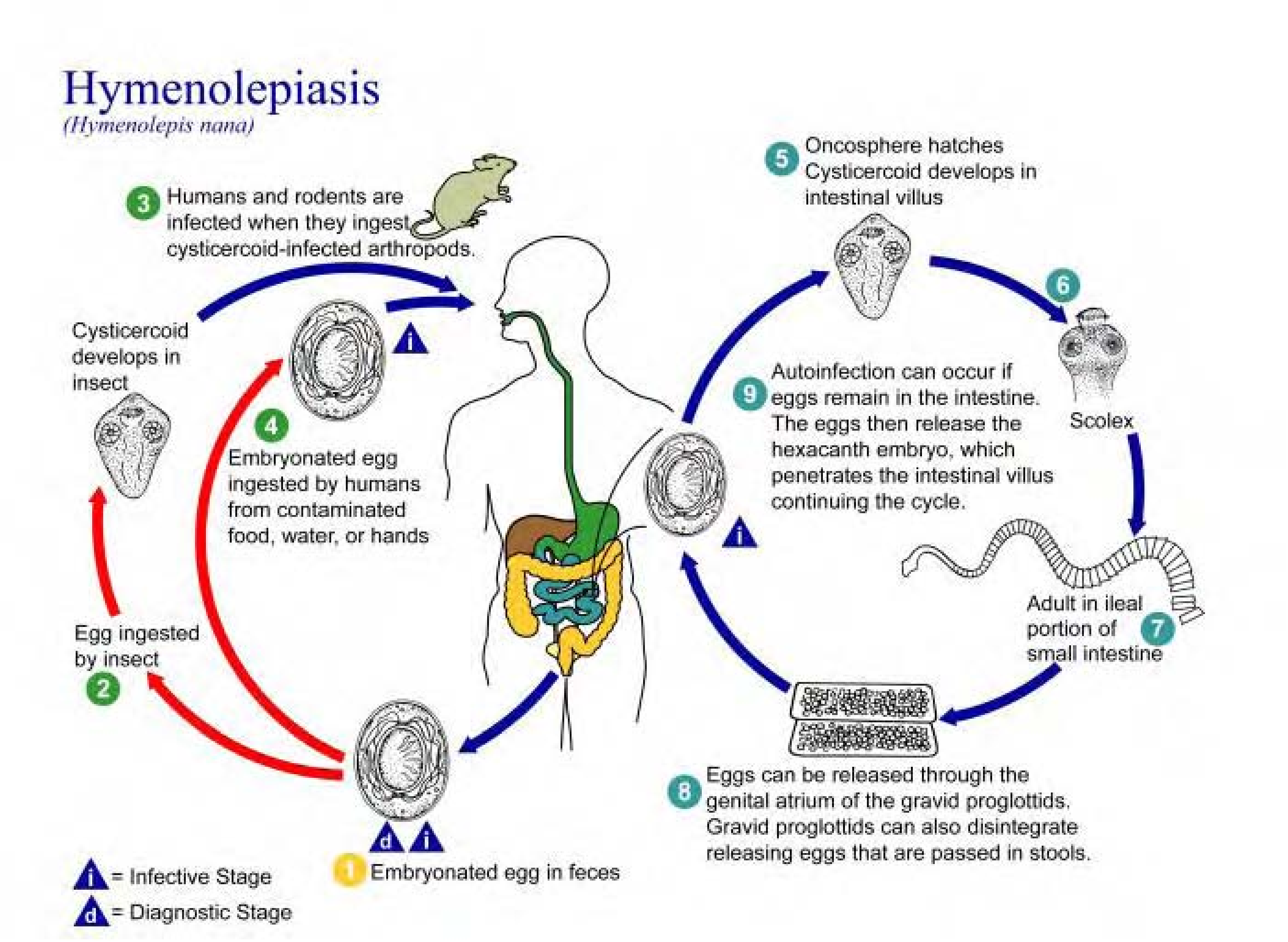
Hymenolepiasis
(Hymenolepis nana)
3 Humans and rodents are infected when they ingest cysticercoid-infected arthropods.
Cysticercoid develops in insect
i
4 Embryonated egg ingested by humans from contaminated food, water, or hands
Egg ingested by insect
2
d i
1 Embryonated egg in feces
i = Infective Stage
d = Diagnostic Stage
5 Oncosphere hatches Cysticercoid develops in intestinal villus
6
Scolex
9 Autoinfection can occur if eggs remain in the intestine. The eggs then release the hexacanth embryo, which penetrates the intestinal villus continuing the cycle.
i
Adult in ileal portion of small intestine 7
8 Eggs can be released through the genital atrium of the gravid proglottids. Gravid proglottids can also disintegrate releasing eggs that are passed in stools.

Illustration 4-2. This is an illustration of the life cycle of *Hymenolepis nana*, the causal agent of Hymenolepiasis. (SOURCE: PHIL 3396 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

The eggs that are released from mature proglottids in the upper ileum are usually passed out in the feces. If swallowed by another human they develop into hexacanth oncospheres and burrow into the villi of the small intestine. This is where they develop into tailless cysticercoids and then migrate towards the ileum and attach to commence the formation of proglottids. The eggs which are ingested by insects, such as fleas, beetles or cockroaches hatch to form tailed cysticercoids which remain unmodified as long as they are inside the insect. If they are accidentally swallowed by a human they pass down to the ileum and establish themselves.

Morphology

The egg containing the oncosphere bears three pairs of hooklets and is surrounded by a membrane. This membrane has two polar thickenings from which arise threadlike filaments extending into the space between the membrane and the colorless hyaline shell, unlike those of *H. diminuta* which do not possess any filaments.

The adult tapeworm is normally 2.5-4cm long. The scolex is knob like in shape, has a rostellum with hooklets and four suckers. The segments are wider than they are long. Ova are spherical or ovoid measuring 30-47µm in diameter. This is what distinguishes it morphologically from *H. diminuta.*

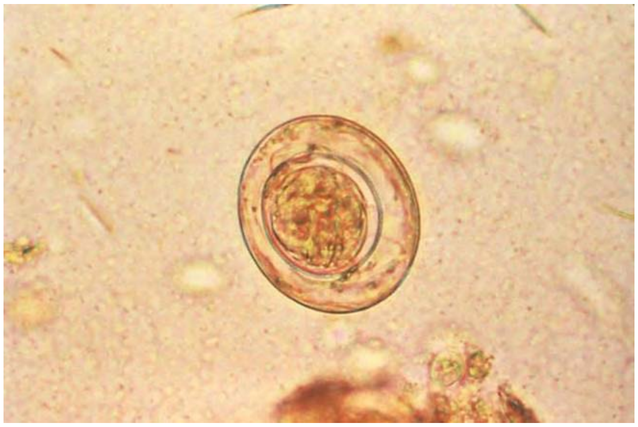

Image 4-6. This micrograph depicts an egg of the dwarf tapeworm *Hymenolepis nana*. On the inner of the two membranes surrounding the *Hymenolepis nana* egg are two poles, from which 4 to 8 polar filaments spread out between the two membranes. The oncosphere, or larval stage, has six hooks. (SOURCE: PHIL 5265 – CDC)

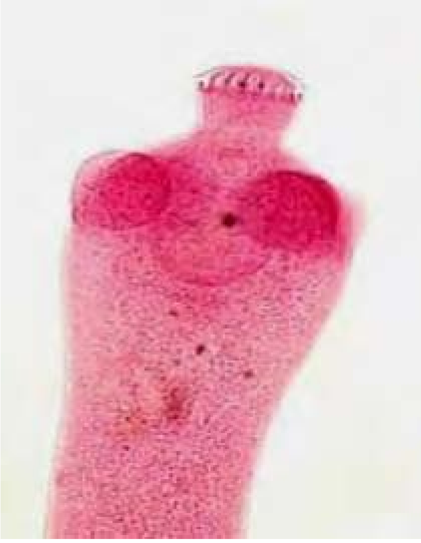

Image 4-7. *Hymenolepis nana* scolex. Stained to show the scolex with a knob like rostellum bearing a ring of hooklets. They possess four suckers, two of which can be seen just below the protruding rostellum. (SOURCE: D. Scott Smith, M.D./CDC)

Clinical Disease

Infections due to *H. nana* may cause no symptoms even with heavy worm burdens. However, symptoms of restlessness, irritability, anorexia, abdominal pain and diarrhea have been reported. Heavy worm burdens may be caused by auto-infection which can be a problem in the immunocompromised.

Laboratory Diagnosis

Diagnosis is based on recovery and identification of the characteristic ova in a formol-ether concentrate of feces. Adult worms and proglottids are rarely seen in stool samples.

Hymenolepis diminuta

Introduction

Hymenolepis diminuta is a small tapeworm commonly found in rats and mice. It has a world wide distribution in these hosts but is infrequently found in humans, with only sporadic cases being reported.

Life cycle and Transmission

The life cycle of *H. diminuta* requires an intermediate arthropod host e.g. earwigs, larval fleas and various beetles. Human infection occurs by the accidental ingestion of an infected arthropod, which contains the cysticercoids.

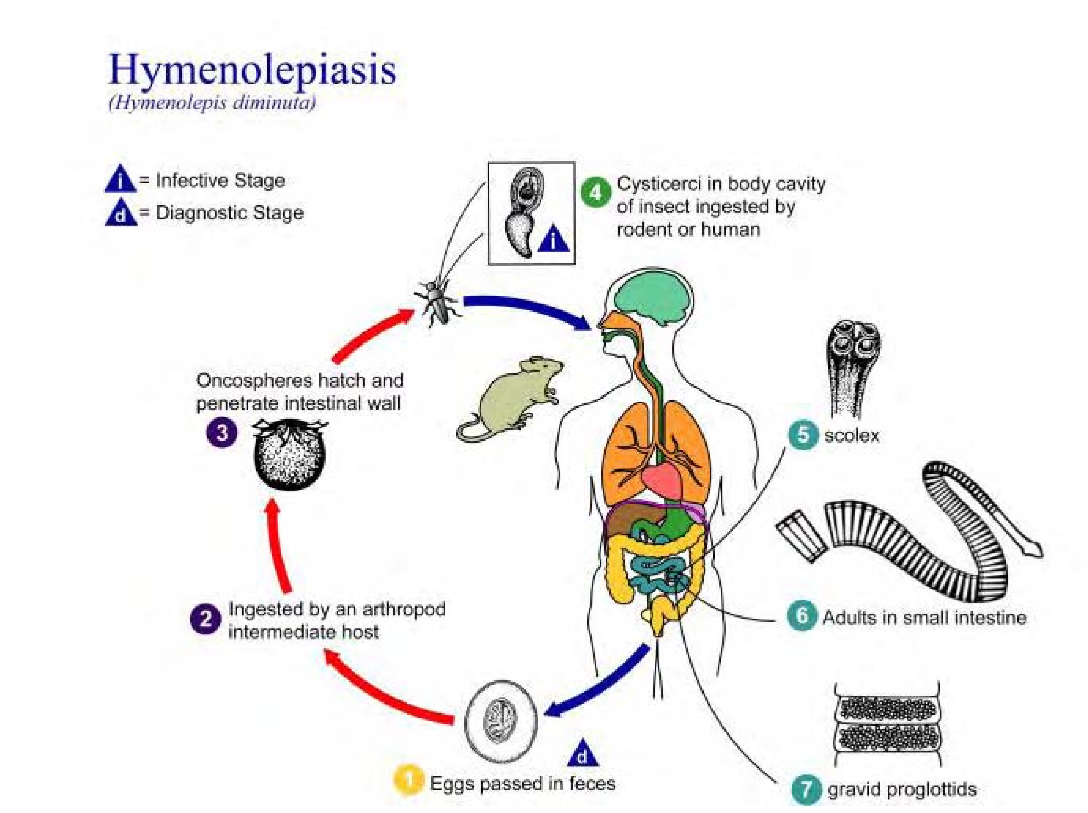
Hymenolepiasis
(Hymenolepis diminuta)
i = Infective Stage
d = Diagnostic Stage
4 Cysticerci in body cavity of insect ingested by rodent or human
i
Oncospheres hatch and penetrate intestinal wall
3
5 scolex
6 Adults in small intestine
2 Ingested by an arthropod intermediate host
d
1 Eggs passed in feces
7 gravid proglottids

Illustration 4-3. This is an illustration of the life cycle of *Hymenolepis diminuta*, the causal agent of Hymenolepiasis. (SOURCE: PHIL 3395 - CDC /Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The ova are large, ovoid and yellowish with a moderately thick shell. They contain an oncosphere with six hooklets and a clear area between the oncosphere and the shell. They measure 70-85µm by 60-80µm.

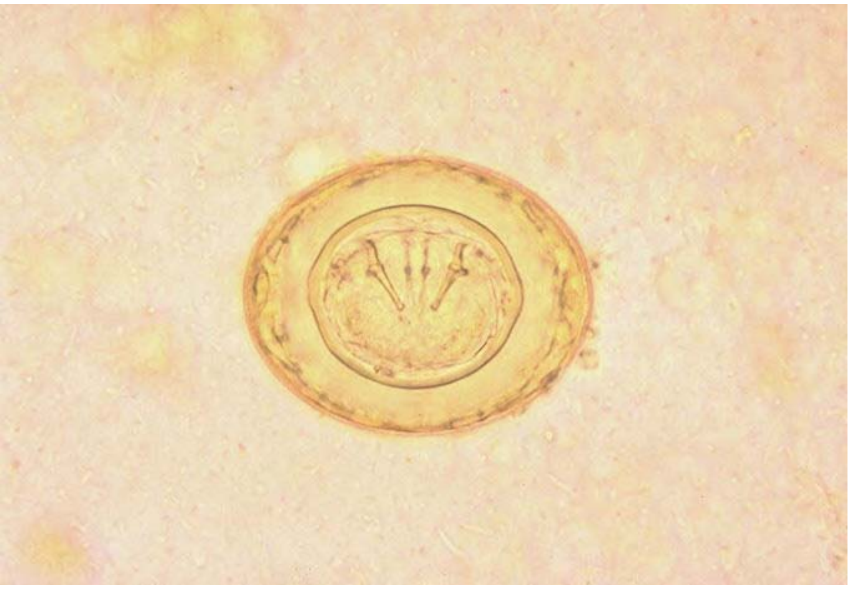

Image 4-8. *Hymenolepsis diminuta* egg. (SOURCE: PHIL 658 - CDC/Dr. Mae Melvin)

The adult worm is 20–60cm long. It has a knob like scolex with a rostellum but no hooklets and four suckers (in contrast to *H. nana*). The rostellum can be withdrawn into a rostellar sac. The tapeworm contains about 1000 proglottids, each of which is wider than long.

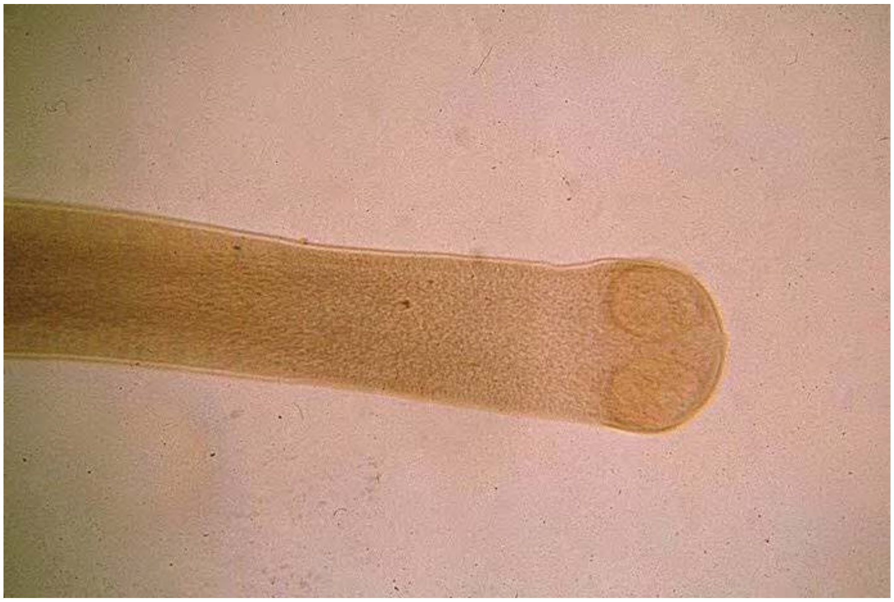

Image 4-9. *Hymenolepsis diminuta* scolex. (SOURCE: PHIL 370 – CDC)

Clinical Disease

The symptoms associated with *H. diminuta* infections are few if any.

Laboratory Diagnosis

Diagnosis is based on recovery and identification of the characteristic ova in a formol-ether concentrate of feces. Adult worms and proglottids are rarely seen in stool samples.

Diphyllobothrium latum

Introduction

Members of this order, commonly known as pseudophyllids, are chiefly parasites of fish-eating mammals, birds and fish. They typically are found with a scolex which is characterized by two shallow elongated bothria situated with one dorsally and one ventrally. The proglottids are flattened dorsoventrally.

Diphyllobothrium latum is an intestinal tapeworm, known as the human ‘broad’ tapeworm. It is the largest tapeworm found in man. The term ‘broad’ relates to the fact that the proglottids are generally wider than they are long. It is an important human parasite. The adult worms of two other species of the genus, *D. dendriticum* and *D. ditremum* are chiefly parasite of fish-eating birds and mammals.

The tapeworm, *D. latum* has a wide distribution, occurring especially in countries bordering the Baltic Sea (Finland, Sweden etc.): and also in Russia, Switzerland and North America. It is in these countries where the populations are known to eat uncooked or partly cooked (i.e. smoked) fish.

Apart from man they are found in many other hosts, especially the dog, cat and pig. This is due to the host countries allowing the domestic animals access to the offal from the infected fish.

Life Cycle and Transmission

The life cycle of this tapeworm requires two intermediate hosts.

The eggs are passed out in human feces, once in water they hatch out into small ciliates coracidium larvaem which swim until ingested by Copepods. It is in these intermediate hosts that growth and development of the 1st larval stage are completed (They are now known as procercoids). When these crustaceans (fresh water) are eaten by fish, the procercoid larvae continue to develop in the flesh of the fish and become known as plerocercoid larvae. It is this stage of the larvae which develops in man when they eat undercooked fish and they grow into adult worms in the small intestine.

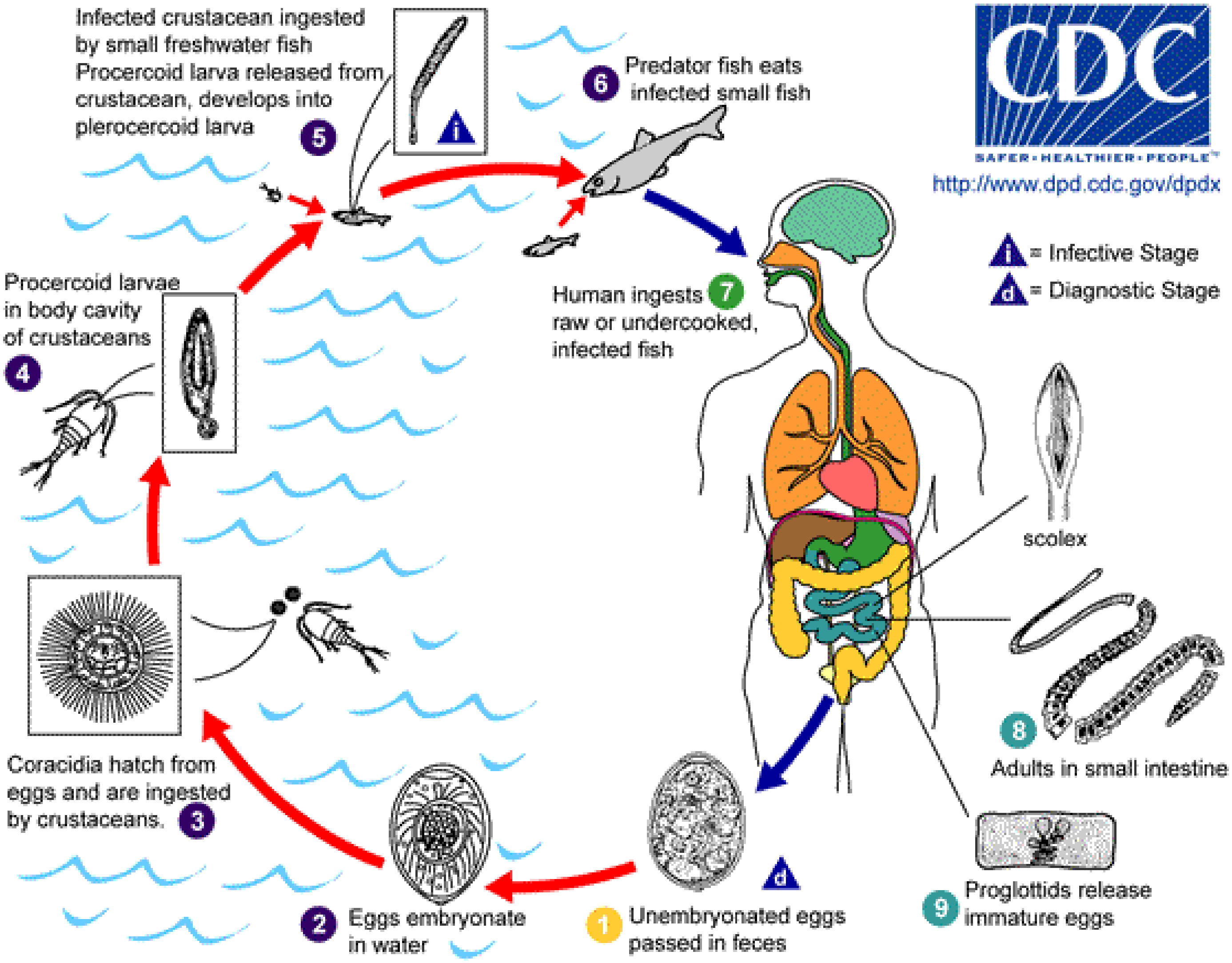
Infected crustacean ingested by small freshwater fish Procercoid larva released from crustacean, develops into plerocercoid larva 5
i
6 Predator fish eats infected small fish
CDC
SAFER • HEALTHIER • PEOPLE™
http://www.dpd.cdc.gov/dpdx
i = Infective Stage
d = Diagnostic Stage
Procercoid larvae in body cavity of crustaceans
4
Human ingests 7 raw or undercooked, infected fish
scolex
8
Adults in small intestine
Coracidia hatch from eggs and are ingested by crustaceans. 3
2 Eggs embryonate in water
1 Unembryonated eggs passed in feces
d
9 Proglottids release immature eggs

Illustration 4-4. Immature eggs are passed in feces ①. Under appropriate conditions, the eggs mature (approx. 18 to 20 days) ②and yield oncospheres which develop into a coracidia ③. After ingestion by a suitable freshwater crustacean (the copepod first intermediate host) the coracidia develop into procercoid larvae ④. Following ingestion of the copepod by a suitable second intermediate host, typically minnows and other small freshwater fish, the procercoid larvae are released from the crustacean and migrate into the fish flesh where they develop into a plerocercoid larvae (sparganum) ⑤. The plerocercoid larvae are the infective stage for humans. Because humans do not generally eat undercooked minnows and similar small freshwater fish, these do not represent an important source of infection. Nevertheless, these small second intermediate hosts can be eaten by larger predator species, e.g., trout, perch, walleyed pike ⑥. In this case, the sparganum can migrate to the musculature of the larger predator fish and humans can acquire the disease by eating these later intermediate infected host fish raw or undercooked ⑦. After ingestion of the infected fish, the plerocercoid develop into immature adults and then into mature adult tapeworms which will reside in the small intestine. The adults of *D. latum* attach to the intestinal mucosa by means of the two bilateral groves (bothria) of their scolex ⑧. The adults can reach more than 10 m in length, with more than 3,000 proglottids. Immature eggs are discharged from the proglottids (up to 1,000,000 eggs per day per worm) ⑨and are passed in the feces ①. Eggs appear in the feces 5 to 6 weeks after infection. In addition to humans, many other mammals can also serve as definitive hosts for *D. latum*. (SOURCE: CDC)

Morphology

The egg is usually ovoid and has a small knob at the opercular end and is yellowish-brown in color with a smooth shell, of moderate thickness. They measure 58-75µm by 40-50µm in size.

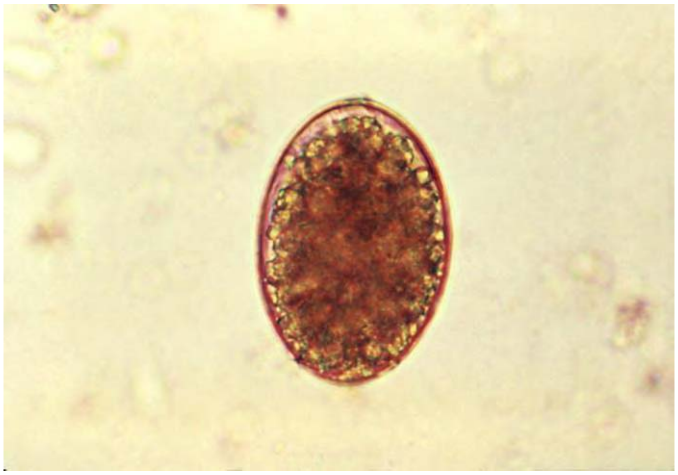

Image 4-10. This micrograph reveals an egg of tapeworm cestode parasite *Diphyllobothrium latum*. *Diphyllobothrium sp.* unembryonated eggs passed in the stool, are oval or ellipsoidal with an operculum at one end that can be inconspicuous. At the opposite (abopercular) end is a small knob that can be barely discernible. (SOURCE: PHIL 5258 – CDC)

Adult worms can reach up to a length of 10 meters or more and may contain up to 3,000 proglottids. The scolex is spatulate with no rostellum or hooklets. It has two shallow grooves or bothria, which are unlike the typical four suckers seen on the *Taenia* species. The proglottids measure 3µm long and 11µm wide and have a rosette shaped central uterus.

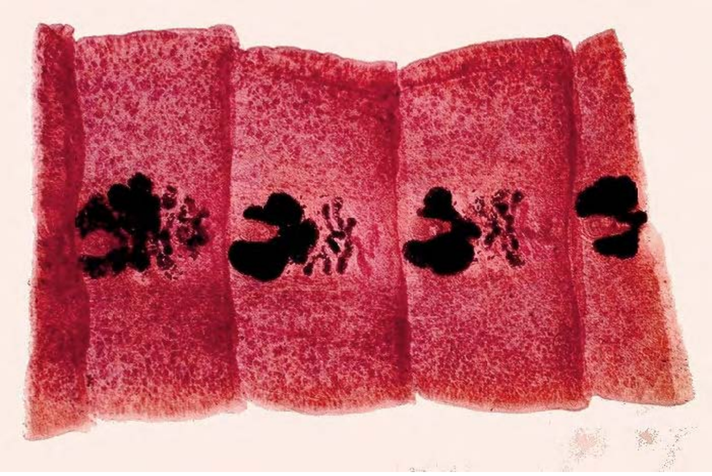

Image 4-11. Gravid proglottids of *Diphyllobothrium latum*. (SOURCE: PHIL 1516 - CDC /Dr. Mae Melvin)

Clinical Disease

The infection caused by *D. latum* is due to the ingestion of raw, poorly cooked or pickled fresh water fish. The symptoms associated with *D. latum* infection may be absent or minimal with eosinophilia. There may be occasional intestinal obstruction, diarrhea, and abdominal pain. The most serious symptom is the onset of pernicious anemia. This is due to a vitamin B_{12} deficiency, caused by excessive absorption of the vitamin by the adult worm and the absorption of cobalamins from the host intestine (occurring only in a small percentage of people).

Laboratory Diagnosis

Laboratory diagnosis depends on the recovery of characteristic eggs from a formol ether concentrate of feces. Proglottids may also be seen in fecal samples usually in a chain of segments from a few centimeters to about half of a meter in length.

4.3 Identifying Intestinal Helminths

The usual diagnostic stages for identifying medically important helminths are the eggs and larvae. Occasionally, adult worms like *Ascaris* and *Enterobius* may be seen and segments or proglottids are used for diagnosing certain tapeworms.

If an egg, is found with the following features as described below, it should be carefully observed in order to make a specific identification.

1. Size: The length and width are measured and are generally within a specific range.

2. Shape: Each species has its own particular shape.

3. Stage of development when passed: In some species, the eggs consist of a single cell; in some, there may be several cells; and some species are usually embryonated (i.e., they contain a larva) when passed in the feces. Occasionally, if the stool specimens are several hours or 1–2 days old, eggs may develop to more advanced stages. *Ascaris* eggs usually have only one cell when passed in the feces; however, the single cell may divide and, in old specimens, eggs with two or four cells may be seen. Hookworm eggs in specimens that are several hours old may contain 16, 32 or more cells. In 12–24 hours, the egg may be embryonated and later still the larvae may hatch. Therefore, when observing the stage of development of helminth eggs, be sure that the stool specimen is freshly passed. If it is several hours or a day old, expect to see changes in the stage of development of some species. Ideally only fresh samples should be accepted for diagnosis.

4. Thickness of the egg shell: Some species, like *Ascaris*, have thick shells; others, like hookworm, have thin shells.

5. Color: Some eggs are colorless (e.g., hookworm, *Enterobius*), others are yellow or brown (*Ascaris, Trichuris*).

6. Presence of characteristic like opercula (lids), spines, plugs, hooklets, or mammillated outer coats.

Chapter 5.
The Nematodes

5.1 Infections Acquired Through the Gastrointestinal Tract

Nematodes (or round worms) are non-segmented helminths known as make up a large assemblage of relatively simple structured organisms. They possess bilateral symmetry and a complete digestive tract with oral and anal openings; they taper to a relative point at both ends. They are also found to have separate sexes, with the male being smaller than the female, ranging in size from a few millimeters to over a meter in length. Their cylindrical non-segmented bodies allow them to be easily distinguishable from other helminths.

Nematode infections have a widespread distribution being found in both Temperate and Tropical climates. They can be found in fresh water, in the sea and the soil, successfully invading both animals and plants. The nematodes found in man invade the body fluids such as the blood or lymph channels and also the intestine. The ones that successfully invade the intestine are generally larger but, the nematodes which invade the tissues can grow to relatively enormous lengths.

Once hatched in the intestine they undergo an incredible migration. The larvae initially burrow into the mucosa, penetrate blood vessels and appear as second stage larvae in the liver within six hours post-infection. Here they remain for several days and develop into third stage larvae, L_3. These larvae then migrate to the heart and are carried to the lungs via the pulmonary arteries, arriving within four to seven days. From there they break out of the capillaries into the alveoli and finally work their way up the trachea to the pharynx and reach the small intestine on the 8th or 10th day post-infection.

Within the intestine, the larvae begin their third molt and become fourth stage larvae by the tenth day. The pre-patent period of *A. suum* in pigs (40–53 days) is less than that of *A. lumbricoides* (54–61 days) in humans. Two to three months after ingestion of the eggs, the females lay eggs in the intestine.

The fertilized female can lay about 200,000 eggs per day. Eggs require oxygen and moisture to embryonate and the worm is often found associated with *Trichuris trichiura* (see the *Trichuris trichiura* section).

Species	**Size**	**Shape**	**Color**	**Stage of Development When Passed**	**Specific Features And Variations**
Nematodes					
Enterobius vermicularis	55 μm x 26 μm Range, 50-60 μm 20-32 μm.	Elongated, asymmetrical with one side flattened, other side convex.	Colorless.	Embryonated. Contains C shaped or tadpole-like embryo.	Smooth, thin eggshell with one flattened side. Occasionally may contain a fully developed larva. (More readily found on anal swabs than in feces).
Ascaris lumbricoides					
fertile egg	60 μm x 45 μm. Range, 45-70 μm x 35-45 μm.	Round or ovoidal. with thick shell.	Brown or yellow brown.	1 cell, separated from the shell at both ends.	Mammillated albuminous coat or covering on outer shell. Coat is sometimes lost and decorticated eggs have a colorless shell with gray or black internal material. Eggs may be in 2, 4, or more cells, or contain a fully developed larva.
infertile egg	90 μm x 40 μm. Range, 85-95 μm x 35-45 μm.	Elongated, occasionally triangular, kidney shaped or other bizarre forms. Shell often very thin.	Brown.	Internal material is a mass of irregular globules and granules that fills shell.	Mammillated covering attenuated or missing in many cases.
Trichuris trichiura	54 μm x 22 μm. Range, 49-65 μm x 20-29 μm.	Elongated, barrel-shaped with a polar "plug" at each end.	Yellow to brown. "Plugs" are colorless.	1 cell or unsegmented.	Polar plugs are distinctive. Eggs occasionally are oriented in a vertical or slanted position and may not be readily recognized. A gentle tap on the coverslip will usually reorient the egg. On rare occasions, atypical eggs lacking polar plugs may be seen.
Hookworm					
Ancylostoma duodenale	60 μm x 40 μm. Range, 57-76 μm x 35-47 μm.	Oval or ellipsoidal with a thin shell.	Colorless with grayish cells.	4- to 8-cell stage.	Occasionally, eggs in advanced cleavage (16 or more cells) or even embryonated may be seen. Rhabditiform larvae may be present if the specimens are old. Species identification can not be made on eggs alone; therefore, eggs should be reported simply as hookworm.
Necator americanus	65 μm x 40 μm. Range, 57-76 μm x 35-47 μm.				
Trichostrongylus species	90 μm x 40 μm. Range, 75-95 μm x 40-50 μm.	Elongated with one or both ends more pointed than hookworm.	Colorless with grayish cells.	May be in advanced cleavage or morula stage.	Egg resembles hookworm egg but is larger and more pointed at the ends.

Table 5-1. Differential Morphology of the Diagnostic Stages of Helminths Found in Humans: Eggs (Nematodes) (SOURCE: CDC)

Ascaris lumbricoides

Morphology

Ascaris lumbricoides is the largest of the intestinal nematodes found in man. The male measures 15cm with a diameter of 3-4µm and has a curled tail with protruding spicules. The female is 20-35cm long with a diameter of 5µm with a straight pointed posterior end. The mouth has one dorsal and two ventral lips. Both are creamy white and the cuticle has fine circular striations.

cm
1
2
3
SPEC.
DATE

Image 5-1. Adult worm of *Ascaris lumbricoides.* They appear creamy white with fine circular striations, for diagnosis they require to be distinguished from earthworms. Females measure approximately 20-35cm long (SOURCE: CDC)

The ova can be unfertilized, fertilized or decorticated and can show considerable variation in shape and size. They measure 85-95µm by 43-47µm. The fertilized ova are easily recognized, oval in shape with a thick wall showing an irregular bumpy surface. They measure 45-75µm by 35-50µm. The outer covering has an albuminoid coat, stained golden brown by bile. The outer wall lies directly on top of a thick smooth shell, which is not easily distinguishable. Some have lost their albuminoid wall. The unfertilized ova are longer and narrower than the fertile ova, measuring 75-85µm by 35-50µm.

The shell layers of the egg provide a very resistant structure which can withstand many chemicals which make them ideal parasites of the intestine.

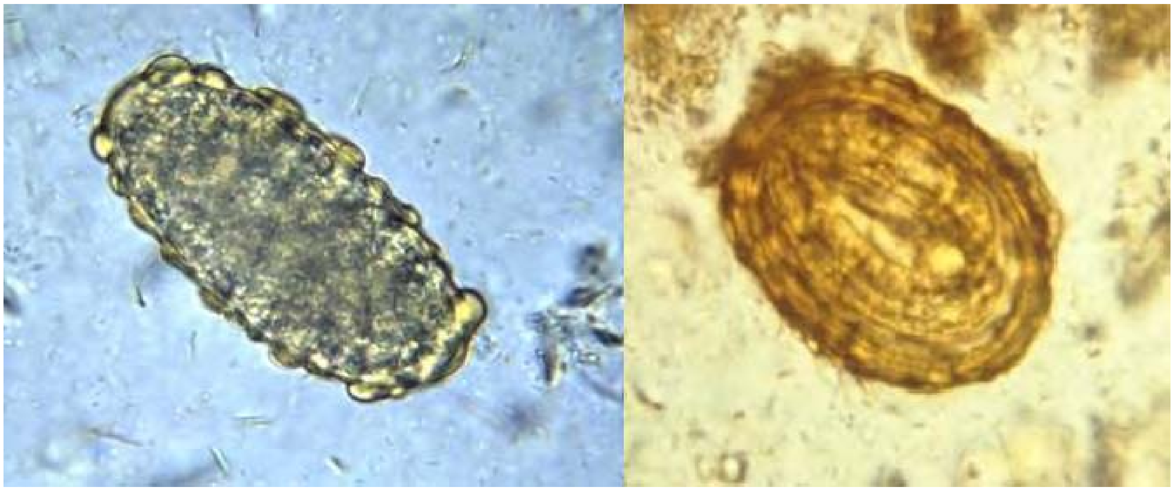

Image 5-2. Unfertilized (left) and fertilized (right) *Ascaris lumbricoides* eggs. The unfertilized eggs are longer and narrower than fertilized eggs, measuring 75-85µm by 35-50µm wide. The fertilized eggs have a thick and bumpy outer wall which is stained golden brown with bile. (Saline wet prep) (SOURCE: PHIL 411/4821 - CDC/Dr. Mae Melvin)

Clinical Disease

Small burdens of worms in the intestine may cause no symptoms. The patient may have symptoms of pneumonitis with cough and low grade fever during the migration of the larvae through the liver and lungs. This can be accompanied by wheezing, coughing and eosinophilia. In heavy worm burdens the adult worms actively migrate in the intestine resulting in intestinal blockage, vomiting and abdominal pain but infections may also be asymptomatic. The worms can penetrate through the wall of the intestine, or into the appendix, travel up the common bile duct, which may become blocked or they may then enter the gal bladder or liver. A heavy worm burden in children may lead to severe nutritional impairment and retardation in growth.

Laboratory Diagnosis

The adults of *A. lumbricoides* may be expelled through the anus, mouth or nose. It is important to distinguish the adult worms from earthworms which are segmented and are often collected as a contaminant from toilets.

The microscopic examination of stool deposits after concentration reveals the characteristic bile stained ova. Eggs may be difficult to identify if an excess of iodine is added to the wet preparation as they retain the stain thus resembling debris. Ova may also become decorticated. In most symptomatic cases identification is easy due to the vast number of eggs, which can be found within a few seconds of starting to scan the slide.

Hookworm Species

Introduction

Hookworms infective to man comprise of two species, *Necator americanus* and *Ancylostoma duodenale*. They are classed as one of the most destructive of human parasitic helminths. There is no intermediate host, with man being the only definitive host.

It is estimated that there are some 900 million cases of infection world wide (Crompton, 1989). The infection is serious where the worms insidiously undermine the health of their hosts.

They occur in areas where sanitary and environmental conditions favor the development of the eggs and larval infections (e.g., warm, damp soil).

The geographic distributions of the two species are remarkably divided into:

Necator americanus which predominately is a New World hookworm, where it was introduced from Africa to the Western Hemisphere. It can also be found in the Far East, Asia, Africa, South America, and Oceania.

Ancylostoma duodenale is an Old World hookworm; it is the only species of Europe and areas bordering the Mediterranean. It can also be found in the Middle East, North China, Africa, Asia, and South America.

Life Cycle

The adult worms live in the small intestine, attached firmly to the mucous membrane of the gut lining, and feed on blood and tissue. The adult females deposit their eggs whilst in the gut (they can produce up to 20,000 eggs per day), the eggs are then passed out in the feces. The rhabditiform larvae hatch in warm, damp soil (light sandy loam), feeding on bacteria. After about one week during which they have gone through two molts become infective and climb into a suitable position waiting for a suitable host to pass by. The larvae enter the host by penetrating unbroken skin (it is now recognized that *A. duodenale* can successfully enter man by oral ingestion, this may be more important for this species than skin penetration). The larvae then enter blood vessels and are carried to the heart, lungs and trachea. They are then swallowed and develop into adult worms in the small intestine. Larvae that are initially swallowed may not show this migration.

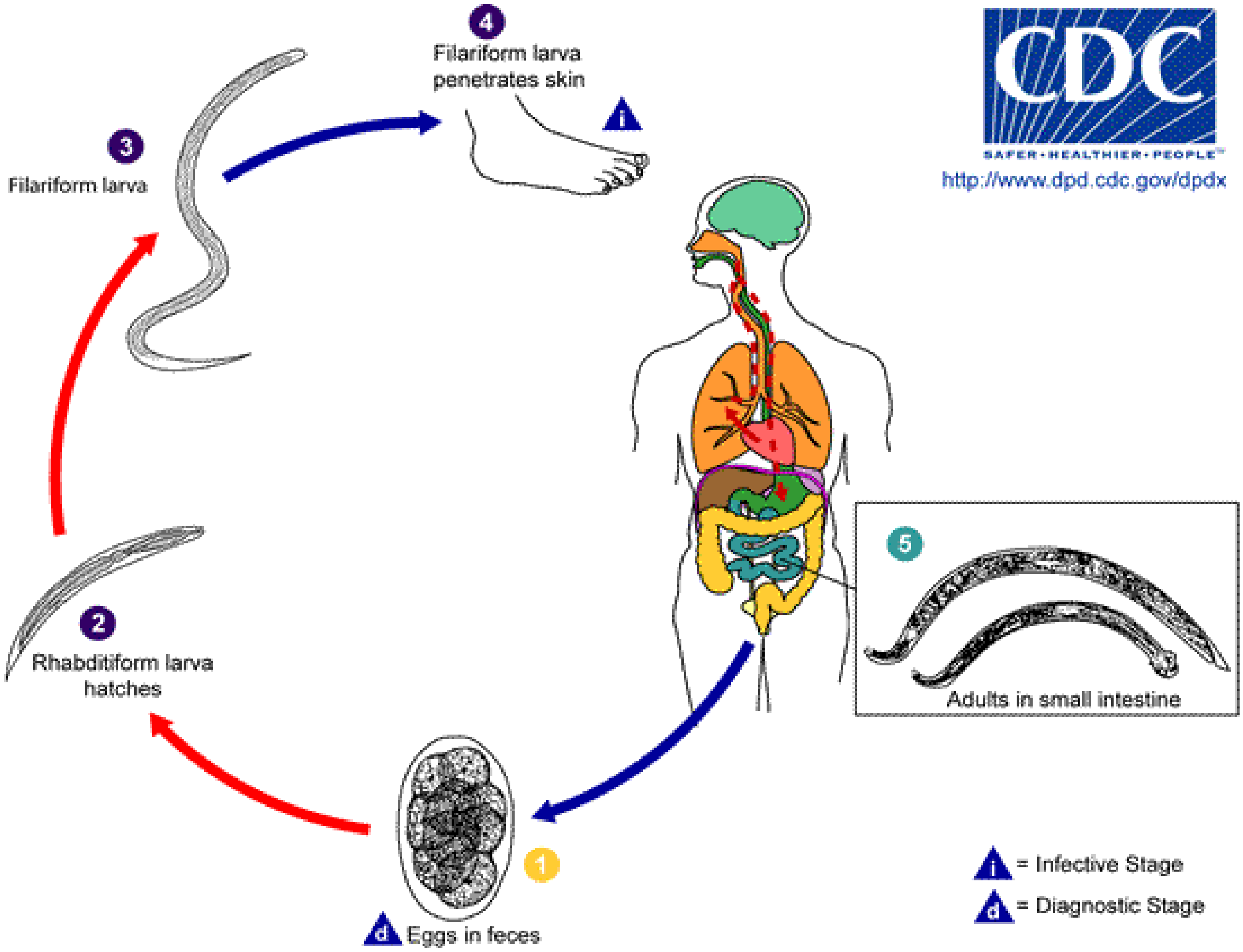

④
Filariform larva penetrates skin
i
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
③
Filariform larva
②
Rhabditiform larva hatches
⑤
Adults in small intestine
①
d Eggs in feces
i = Infective Stage
d = Diagnostic Stage

Illustration 5-1. Eggs are passed in the stool ①, and under favorable conditions (moisture, warmth, shade), larvae hatch in 1 to 2 days. The released rhabditiform larvae grow in the feces and/or the soil ②, and after 5 to 10 days (and two molts) they become become filariform (third-stage) larvae that are infective ③. These infective larvae can survive 3 to 4 weeks in favorable environmental conditions. On contact with the human host, the larvae penetrate the skin and are carried through the veins to the heart and then to the lungs. They penetrate into the pulmonary alveoli, ascend the bronchial tree to the pharynx, and are swallowed ④. The larvae reach the small intestine, where they reside and mature into adults. Adult worms live in the lumen of the small intestine, where they attach to the intestinal wall with resultant blood loss by the host ⑤. Most adult worms are eliminated in 1 to 2 years, but longevity records can reach several years. Some *A. duodenale* larvae, following penetration of the host skin, can become dormant (in the intestine or muscle). In addition, infection by *A. duodenale* may probably also occur by the oral and transmammary route. *N. americanus*, however, requires a transpulmonary migration phase. (SOURCE: CDC)

Larvae live for an average of 3–6 weeks in the tropics (*A. duodenale* can live at lower temperatures than *N. americanus* can, and so is found in more temperate climates).

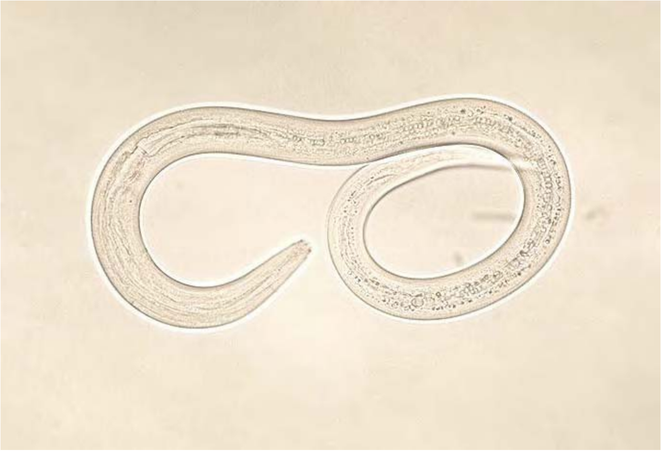

Image 5-3. Hookworm larva. Both *Ancylostoma duodenale* and *Necator americanus* larvae have similar general morphology and measuring approximately 10-13µm for females and 8-11µm for males. (SOURCE: PHIL 1513 – CDC/Dr. Mae Melvin)

Morphology

Both species have similar general morphology and measure approximately, females 10–13µm and males 8–11µm. The general morphology of the two species resemble those of *Nippostrongylus brasiliensis*, the rat hookworm, but they are approximately twice the size of the rat hookworm (*Nippostrongylus brasiliensis* a species not discussed here).

The male species has a posterior copulate bursa which is absent from the female. The females though possess a vulva opening which is found almost one third of the body length from the posterior end, they also have two ovaries. Most of the female body is occupied with eggs.

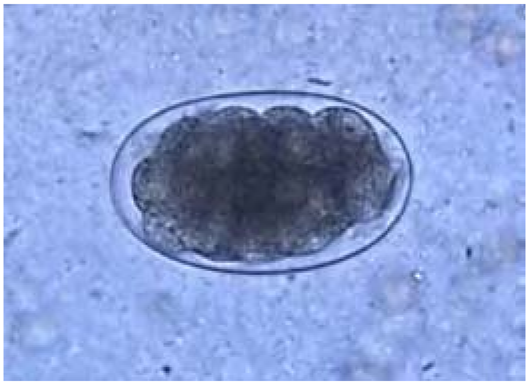

Image 5-4. Hookworm egg, normally passed in feces during the 4 - 8 cell stage. The ova of both species are similar with the shape being oval and transparent with a smooth thin shell and measuring 56-75µm by 36-40µm. (SOURCE: PHIL 5220 – CDC)

The mouth (or buccal cavity) of the two species show a conspicuous pair of chitinous plates on the dorsal surface. *Ancylostoma duodenale* buccal cavity bears two hook like teeth on the top and two triangular cutting plates on the bottom. While the mouth of *N. americanus* has four cutting plates, two on the ventral and two on the dorsal surfaces. The head is curved in both species but *Necator* adults it is finer but more pronounced forming a definite “hook” at the anterior end. The buccal cavity is used to attach the worms securely to the mucosa of the small intestine. With the teeth and cutting plates used to pierce the mucosa.

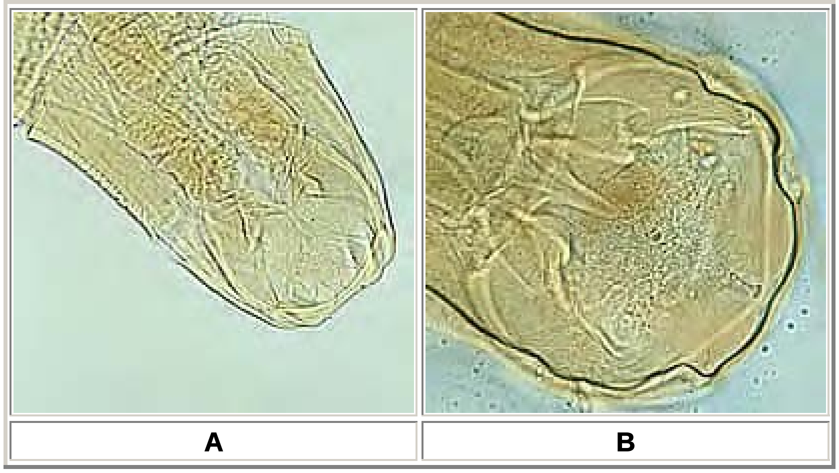

Illustration 5-2. A: Adult worm of *Ancylostoma duodenale*. Anterior end is depicted showing cutting teeth. **B:** Adult worm of *Necator americanus*. Anterior end showing mouth parts with cutting plates. (SOURCE: CDC)

The ova are oval and transparent with a smooth thin shell and measure 56-75µm by 36-40µm. They are usually passed in the 4-8 cell stage in feces and may be embryonated. The ova of both species of Hookworm are similar.

The bursa (the characteristic external feature which forms an umbrella-like extension surrounding the cloaca) of both male species is well developed, *Necator americanus* adults are distinguished from *Ancylotstoma duodenale* by the split dorsal rays and the close arrangement of the lateral rays.

Simplistically the life cycle has three phases:

The parasitic adult females lay eggs while they are in the duodenum where they hatch producing rhabditiform (non-infective) larvae.

1. The larvae can have two fates in life, one where they are passed out in the feces to continue down the free-living path or they develop into infective filariform larvae whilst traveling down the small intestine.

2. The larvae which, develop in the environment can also undergo different development. Some larvae undergo direct development (homogonic) or indirect development (heterogonic).

The non-infective first stage (rhabditiform) larvae develop into free living adults in the soil within 2–5 days and produce infective third stage or filariform larvae which can penetrate exposed skin (heterogonic development). This phase is common in moist, warm tropical countries.

The non-infective rhabditiform larvae which are excreted in the feces develop into infective filariform larvae in the soil (homogonic development). These infective larvae penetrate exposed skin. There is no development of free living adult worms and this phase is common in temperate zones. The larvae never undergo sexual maturity.

Both types of larvae can become established in the host by penetrating the skin or by oral ingestion.

The larvae which infect the host by penetrating the skin undergo a migration through the dermal tissues and into the circulation to the heart and lungs, then up the bronchi and trachea, where they are eventually swallowed and pass down into the intestine. On reaching the mucosa of the duodenum the females develop and produce eggs. Adult males are unable to attach themselves to the mucosa, therefore, for any copulation to take place they must mate in the lumen of the intestine.

3. The non-infective rhabditiform larvae develop into infective filariform larvae while passing down the small intestine. Autoinfection occurs when the larvae reinfect the host by penetrating the intestinal mucosa or the perianal or perineal skin. The larvae migrate to the lungs via the circulatory system and then return to the intestine.

From initial infection to maturity usually takes less than four weeks.

Hyperinfection Syndrome

The autoinfective capability of larvae may be responsible for long term infections which persist for many years. The parasite and host reach an equilibrium state where neither host nor parasite suffers any adverse reactions. If this equilibrium is disturbed e.g. immunosuppression, the infection proliferates with immense numbers of larvae migrating to every tissue in the body, especially the lungs. This condition is referred to as disseminated strongyloidiasis. This results in tissue damage, pneumonitis, brain damage or respiratory failure.

Laboratory Diagnosis

Microscopy

Laboratory diagnosis depends on finding larvae in stool, sputum or duodenal aspirates.

Strongyloides larvae may be present in the stool in very small numbers and culture methods may be needed to encourage the rhabditiform larvae to develop into filariform larvae and migrate from the sample. The Enterotest or string test can be used to recover larvae from duodenal aspirates.

Larvae must be distinguished from hookworm larvae especially if it is an older sample. Rhabditiform larvae are most commonly seen.

A good concentration technique is essential to increase the chances of seeing larvae, though they are easily killed making diagnosis more difficult.

Serology

Serological tests are of value in the diagnosis of strongyloidiasis when larvae cannot be found. An enzyme linked immunosorbent assay (ELISA) using larva antigen, is usually employed.

Little can be used to distinguish between the two species, but using the curved shape of the head is a good indicator.

Clinical Disease

Larval penetration of the skin may lead to pruritis, often termed as ‘ground itch’ at the site of penetration. Respiratory symptoms may arise during the larval migration.

The adult worm in the intestine may cause intestinal necrosis and blood loss as a result of the attachment of the adult to the intestinal mucosa. Patients with acute infections may experience nausea, vomiting, abdominal pain, diarrhea and eosinophilia.

Chronic infections may lead to iron deficiency and anemia resulting from the excessive loss of iron. Heavy worm burden in children may have serious consequences including death.

Laboratory Diagnosis

Adults of Hookworm species may be passed out spontaneously in feces. The microscopic examination of stool deposits after an iodine stained, formol-ether concentration method concentration reveals the characteristic ova.

Cutaneous Larva Migrans

If man comes in contact with hookworm larva of the dog (or cat), A. braziliense or A. caninum, penetration of the skin may take place. The larvae are unable to complete the migration to the small intestine and become trapped. Trapped larvae may survive for weeks or even months, migrating through the subcutaneous tissues.

Trapped larvae have been known to produce severe reaction, forming tunnels through the tissues, causing intense itchy skin eruption, producing a red, track under the skin which demonstrates accurately the wanderings of the larvae.

Often intense pruritis and scratching may lead to secondary bacterial invasion, known as ‘creeping eruption’ or ‘cutaneous larval migrans’.

Trichuris trichiura

Introduction

Trichuris trichiura, more commonly known as the Whip Worm, due to the whip-like form of the body. These nematodes are most commonly seen in tropical climates and in areas where sanitation is poor. They seem to occur in areas particularly where Ascaris and Hookworms are found due to the eggs requiring the same conditions to allow for embryonation. Both species can be found in humans together. There are several species within this genus each infecting specific hosts, but only T. trichiura infects man. Causing human trichuriasis. It is a parasite that infects many more people than is generally appreciated, up to 800 million people throughout the tropics and temperate regions.

Life cycle eggs require a warm, moist environment with plenty of oxygen to ensure embryonation, but once they have embryonated they are extremely resistant to environmental conditions.

Adult worms are found in the cecum and upper part of the colon of man. In heavy infection they can be found in the colon and the terminal ileum. They attach to the mucosa by the anterior end or by embedding the anterior portion of the body in the superficial tissues, obtaining nutrition from the host tissues.

Once fertilized the female worms lay several thousands of eggs, which are unsegmented at the oviposition and are passed out in the feces. Once they have been passed out they require an embryonation period in the soil which may last from two weeks to several months, after which they become infective.

When embryonated eggs are swallowed by human hosts larvae are released into the upper duodenum. They then attach themselves to the villi lower down the small intestine or invade the intestinal walls. After a few days the juveniles migrate slowly down towards the cecum attaching themselves to the mucosa, reaching their final attachment site simultaneously.

The larvae reach maturity within three weeks to a month after infection, during which they undergo four molts. There is no lung migration and the time from ingestion of infective eggs to the development of adult worms is about three months.

Infection is achieved by swallowing soil that contains embryonated eggs. Therefore, children are most commonly seen to possess the infections, as they are more likely to swallow soil.

Morphology

The adult worms of *T. trichuria* are characterized by the enormously elongated capillary-like esophagus (anterior end); with the anus situated in the extreme tip.

The thin anterior portion of the worm is found embedded in the mucosa. There are no lips and the vulva is at the junction of the thread-like and thickened regions of the body. The posterior end is much thicker and lies free in the lumen of the large intestine.

The female measures 35-50µm long and the male 30-45µm long.

The ova are characteristically barrel shaped, bile stained with bipolar plugs. They measure 50-54µm by 20-23µm.

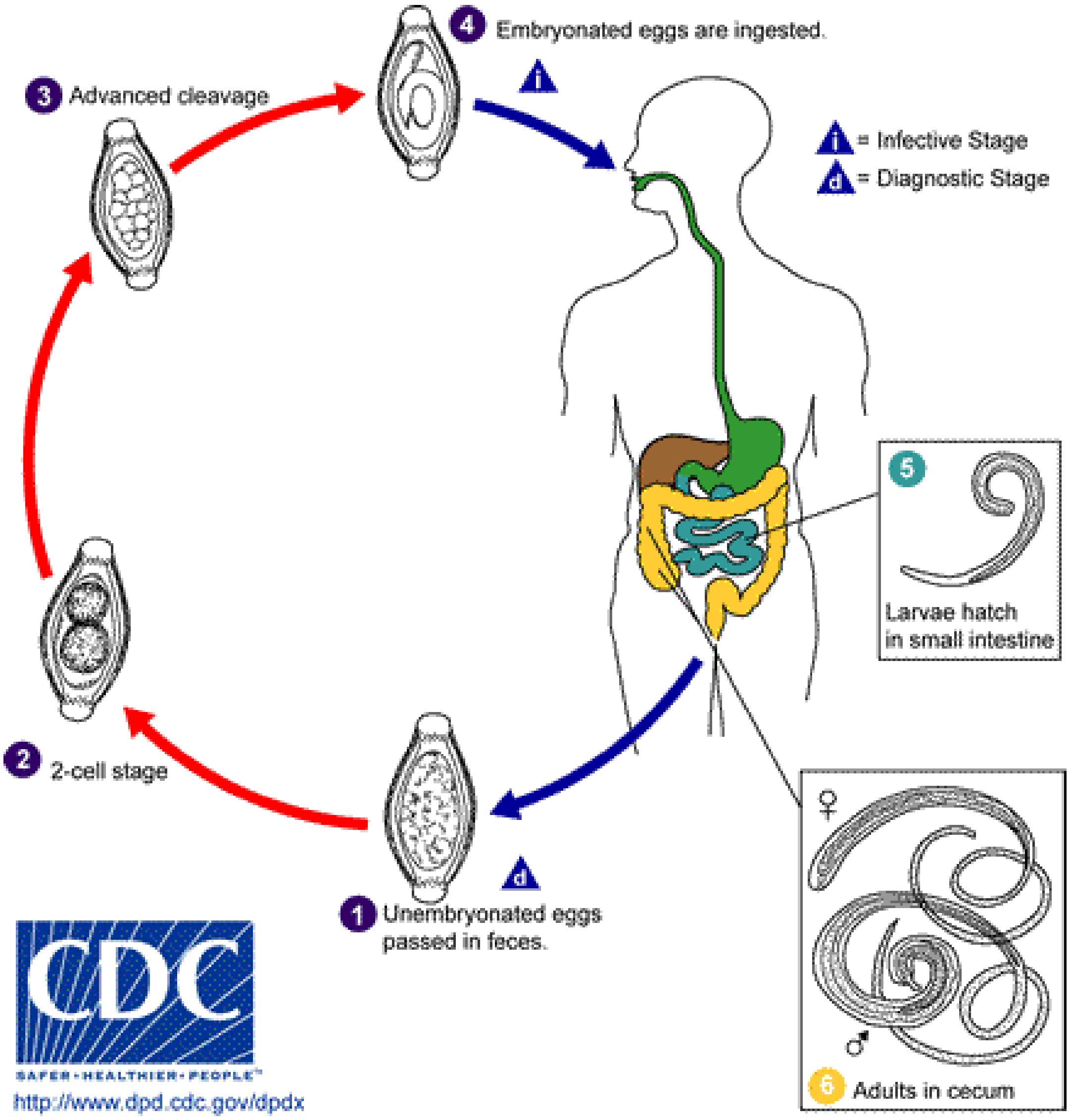
4 Embryonated eggs are ingested.
i
3 Advanced cleavage
i = Infective Stage
d = Diagnostic Stage
5
Larvae hatch
in small intestine
2 2-cell stage
♀
d
1 Unembryonated eggs
passed in feces.
♂
6 Adults in cecum
CDC
SAFER·HEALTHIER·PEOPLE™

http://www.dpd.cdc.gov/dpdx

Illustration 5-3. The unembryonated eggs are passed with the stool ❶. In the soil, the eggs develop into a 2-cell stage ❷, an advanced cleavage stage ❸, and then they embryonate ❹; eggs become infective in 15 to 30 days. After ingestion (soil-contaminated hands or food), the eggs hatch in the small intestine, and release larvae ❺that mature and establish themselves as adults in the colon ❻. The adult worms (approximately 4 cm in length) live in the cecum and ascending colon. The adult worms are fixed in that location, with the anterior portions threaded into the mucosa. The females begin to oviposit 60 to 70 days after infection. Female worms in the cecum shed between 3,000 and 20,000 eggs per day. The life span of the adults is about 1 year. (SOURCE: CDC)

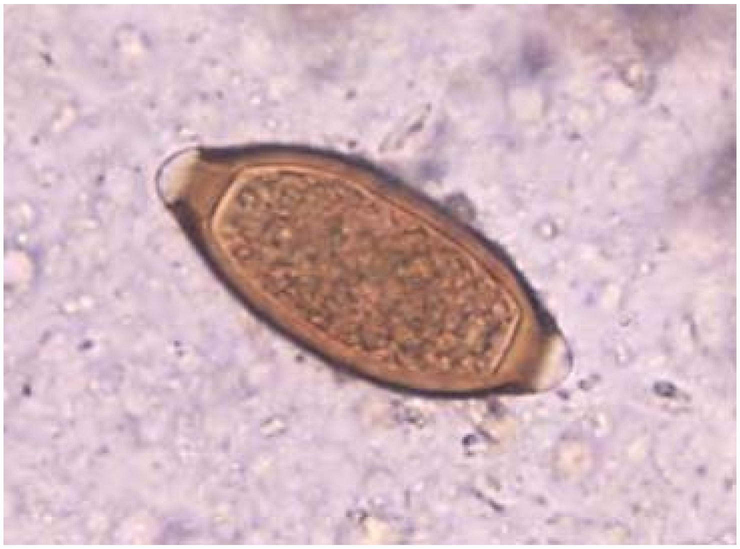

Image 5-5. *Trichuris trichuria* ovum with its opercular plugs, shown as white gaps at either end of the egg. (Saline wet prep) (SOURCE: PHIL 652 – CDC/Dr. Mae Melvin)

Clinical Disease

Most infections due to this nematode are light to moderate with minimal or no symptoms. However, a heavy worm burden may result in mechanical damage to the intestinal mucosa due to the adult worm being threaded into the epithelium of the cecum. Abdominal cramps, tenesmus, dysentery and prolapsed rectum may occur in these cases.

If a prolapsed rectum is observed, many worms may be seen adhering to the mucosa of the rectum.

Symptomatic infections are usually only seen in children. The majority of infections are chronic and mild, with nonspecific symptoms like diarrhea, anemia, growth retardation, eosinophilia.

Laboratory Diagnosis

The adult worms of *T. trichiura* are rarely seen in the feces. The microscopic examination of stool deposits after an iodine stained, formol-ether concentration method concentration reveals the characteristic barrel shaped ova. In symptomatic infections numerous numbers of eggs can be seen due to the prolific nature of the female worms, even in light infections many eggs can be seen in the smear.

Strongyloides stercoralis

Introduction

Strongyloides stercoralis is an intestinal nematode commonly found in warm areas, although it is known to survive in the sub-tropics (hot and humid conditions). The geographic range of *Strongyloides* infections tend to overlap with that of Hookworm due to the eggs requiring the same environmental conditions to induce embryonation.

This parasite is interesting in that it contains a free-living stage (exogenous) and a parasitic stage (endogenous) where the larvae undergo development in both stages.

Life cycle

The life cycle of *S. stercoralis* is a complex one as demonstrated in the diagram below.

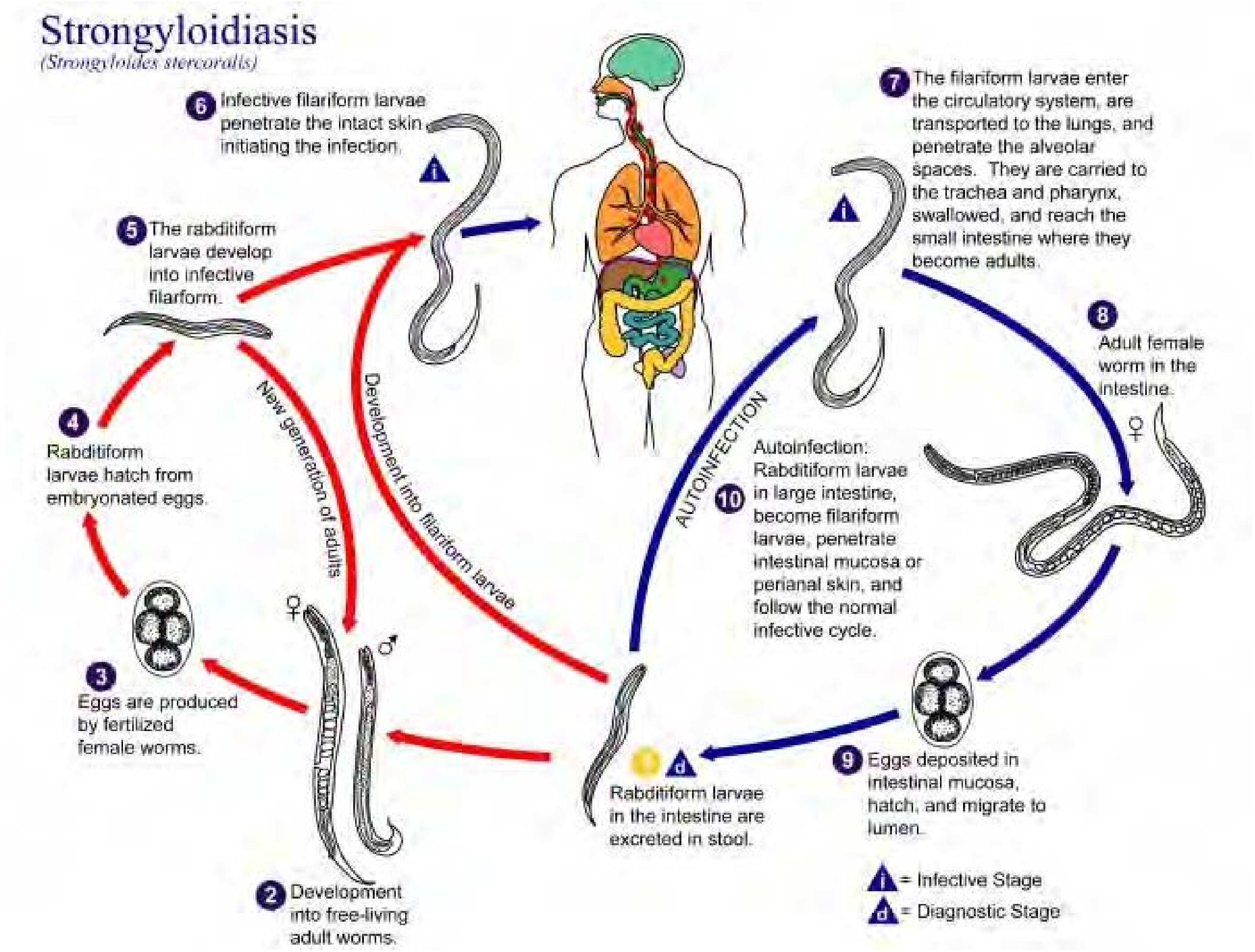
Strongyloidiasis
(Strongyloides stercoralis)
6 Infective filariform larvae penetrate the intact skin initiating the infection.
i
7 The filariform larvae enter the circulatory system, are transported to the lungs, and penetrate the alveolar spaces. They are carried to the trachea and pharynx, swallowed, and reach the small intestine where they become adults.
i
5 The rabditiform larvae develop into infective filarform.
8 Adult female worm in the intestine.
♀
Development into filariform larvae
New generation of adults
4 Rabditiform larvae hatch from embryonated eggs.
AUTOINFECTION
Autoinfection: Rabditiform larvae in large intestine, become filariform larvae, penetrate intestinal mucosa or perianal skin, and follow the normal infective cycle.
10
♀
♂
3 Eggs are produced by fertilized female worms.
d
Rabditiform larvae in the intestine are excreted in stool.
9 Eggs deposited in intestinal mucosa, hatch, and migrate to lumen.
2 Development into free-living adult worms.
i = Infective Stage
d = Diagnostic Stage

Illustration 5-4. Diagram of the life cycle of *Strongyloides stercoralis.* (SOURCE: PHIL 3419 – CDC/ Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The first stage rhabditiform larvae measure approximately 250µm long by 20µm wide. They have a bulbed esophagus and a short buccal cavity. In an old specimen, rhabditiform larvae of *S. stercoralis* must be differentiated from those of hookworm which have a longer buccal cavity. The third stage or filariform larva is approximately 500µm long and has a notched tail (see below) compared with that of hookworm which is sheathed and has a long slender tail.

Adults are slender and possess and extremely long esophagus which in the female extends1/3 to1/2 of the body. The anal opening is ventral and the tail is pointed.

	RHABDITIFORM LARVA (First Stage. Has bulbed esophagus.)			**FILARIFORM LARVA (Third Stage. Lacks prominent bulb in esophagus.)**		
Species	**Size**	**Genital Primordium**	**Buccal Cavity**	**Size**	**Length of Esophagus**	**Tip of Tail**
Strongyloides stercoralis	225 µm × 16 µm. Range, 200- 300 µm × 16-20 µm.	Prominent. Is an elongate, tapered, or pointed structure located along ventral wall about the body length.	Short, about 1/3-1/2 as long as the width of the anterior end of the body.	550 µm × 20 µm. Range, 500-550 µm × 20-24 µm.	Extends approximately 1/2 length of body.	Notched.
Hookworm	250 µm × 17 µm. Range, 200- 300 µm × 14-17 µm.	Inconspicuous. Rarely distinct. When seen, is small, located nearer the tail than that of *Strongyloides*.	Long. Approximately as long as the width of the body.	500 µm. Range, 500-700 µm × 20-24 µm.	Extends about 1/4 1ength of body.	Pointed.

Table 5-2. Differential Morphology of the Diagnostic Stages of Helminths Found in Humans: Larvae (SOURCE: CDC)

Eggs are rarely found in the stool as they hatch in the intestine. They are oval and thin shelled, resembling those of hookworm but are smaller measuring 50-58µm by 30-34µm.

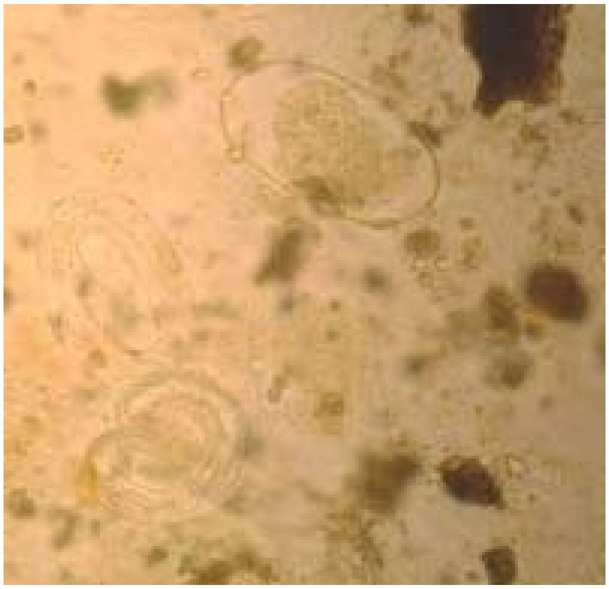

Image 5-6. Thin shelled ova of *Strongyloides stercoralis.* Measuring 50µm by 30µm, they are smaller than Hookworm eggs. (SOURCE: CDC)

Clinical Disease

Disease associated with infections due to *S. stercoralis* is varied, ranging from some patients being totally asymptomatic to the hyperinfection syndrome. There are three areas of involvement in *Strongyloides* infections; skin, lungs and intestine.

1. Initial skin penetration of the filariform larvae usually causes very little reaction, however with repeated infections the patient may mount a hypersensitive reaction thus preventing the larvae from completing its life cycle. The term larva currens is used when there is a rapidly progressing urticarial track.

2. The migration of larvae through the lungs may stimulate an immune response which can result in a cough, wheezing and fever.

3. Symptoms associated with intestinal strongyloidiasis may mimic a peptic ulcer due to ulceration of the intestinal mucosa. In heavy infections the intestinal mucosa may be severely damaged resulting in malabsorption. There may also be lower gastrointestinal bleeding. Eosinophila may be high.

Chapter 6.
The Trematodes

6.1 Infections of the Alimentary Canal and Associated Organs

Introduction

The trematodes (or flukes) are leaf shaped with an outer cover called the tegument which may be smooth or spiny. There are two suckers or attachment organs, an anterior oral sucker and a posterior ventral sucker. The suckers form a characteristic feature of the group, from which the name Trematode is derived from the Greek word for "hole." They can occur in a variety of host environments, with the majority being endoparasites but some are found to be ectoparasitic.

Most trematodes are hermaphroditic and most of the body consists of reproductive organs and their associated structures. The digestive system is well developed; they generally feed on intestinal debris, blood, mucus and other tissues, depending on the host environment.

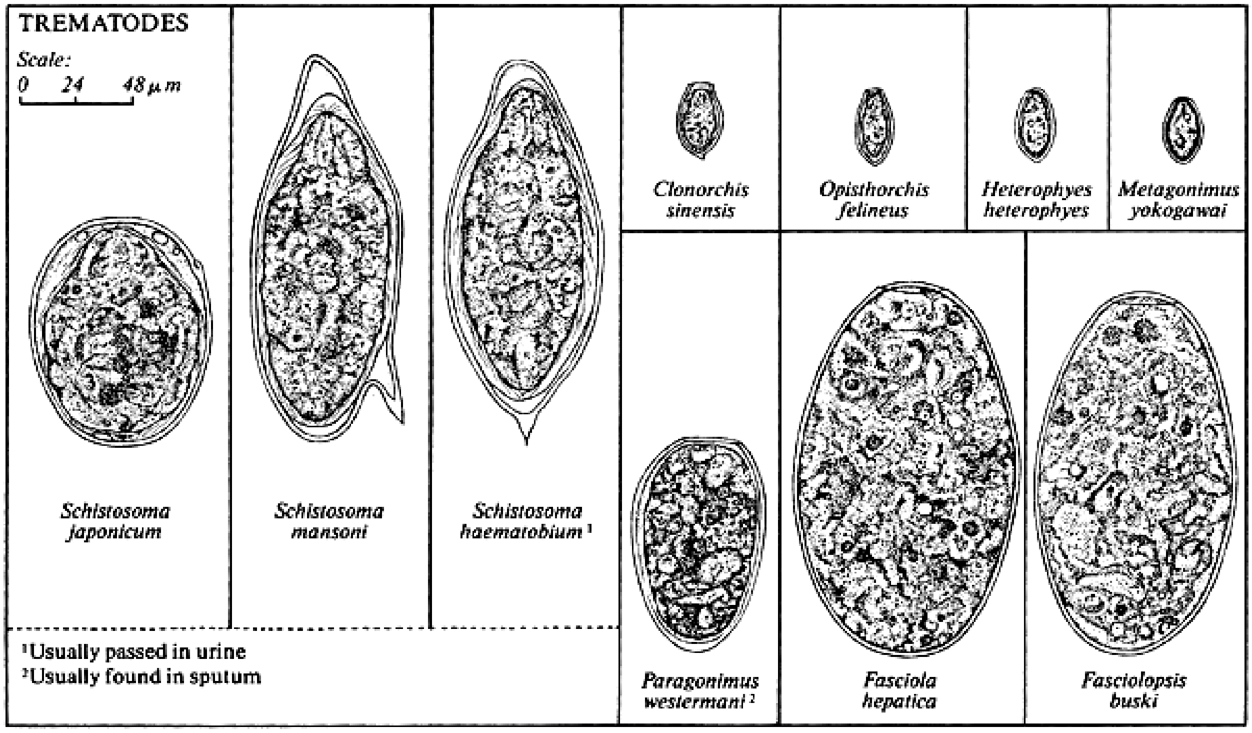
TREMATODES
Scale:
0 24 48 µm
Schistosoma japonicum
Schistosoma mansoni
Schistosoma haematobium ¹
Clonorchis sinensis
Opisthorchis felineus
Heterophyes heterophyes
Metagonimus yokogawai
Paragonimus westermani ²
Fasciola hepatica
Fasciolopsis buski
¹Usually passed in urine
²Usually found in sputum

Illustration 6-1. Trematode Eggs Found in Stool Specimens of Humans. (SOURCE: CDC/Adapted from Melvin, Brooke, and Sadun, 1959)

Species	Size	Shape	Color	Stage of Development When Passed	Specific Features and Variations
Trematodes					
Schistosoma mansoni	140 µm x 66 µm. Range, 114-180 µm x 45-73 µm.	Elongated with prominent lateral spine near posterior end. Anterior end tapered and slightly curved.	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Lateral spine. Found in feces; in rare cases, in urine also. Eggs are discharged at irregular intervals and may not be found in every stool specimen. Are rare in chronic stages of infection.
Schistosoma japonicum	90 µm x 70 µm. Range, 68-100 µm x 45-80 µm.	Oval. Small lateral spine is often seen or may appear as a small hook or "knob" located in a depression in the shell.	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Found in feces. Often coated with debris and may be overlooked.
Schistosoma haematobium	143 µm x 60 µm. Range, 112-170 µm x 40-70 µm.	Elongated with rounded anterior end and terminal spine at posterior end.	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Terminal spine. Found in urine, occasionally in feces. Egg often covered with debris.
Schistosoma intercalatum	175 µm x 60 µm. Range, 140-240 µm x 50-85 µm.	Elongated with tapered anterior end and terminal spine. Sometimes "spindle-shaped."	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Terminal spine long, slender with bent tip. Resembles *S. haematobium* egg except it is longer, is thinner, and has a longer spine. Found in feces. May have debris adhering to shell.
Schistosoma mekongi	69 µm x 56 µm* Range, 51-73 µm x 39-66 µm.	Spherical. Small lateral spine, not always visible or may appear as a small "knob" in a depression in the shell.	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Found in feces. Closely resembles *S. japonicum* egg except it is smaller. May be coated with debris.
Clonorchis sinensis	30 µm x l6 µm. Range, 27-35 µm x 11-20 µm.	Small, ovoidal, or elongated with broad rounded posterior end and a convex operculum resting on "shoulders." A small "knob" may be seen on the posterior end.	Yellow brown.	Embryonated. Contains mature miracidium.	Small size, operculum and "knob" on posterior end. Shell often is covered by adhering debris.
Opisthorchis felineus	30 µm x 12 µm. Range, 26-30 µm x 11-15 µm.	Elongated with operculum on anterior end and pointed terminal "knob" on posterior end.	Yellow brown.	Embryonated. Contains mature miracidium.	Lacks prominent shoulders characteristic of *Clonorchis* and has more tapered end.
Heterophyes heterophyes	28 µm x 15 µm. Range, 28-30 µm x 15-17 µm.	Small, elongated or slightly ovoidal. Operculum. Slight "knob" at posterior end.	Yellow brown.	Embryonated. Contains mature miracidium.	Resembles *Clonorchis* egg but with less distinct shoulders. Operculum is broader than in *Clonorchis*.
Metagonimus yokogawai	28 µm x 17 µm. Range, 26-30 µm x 15-20 µm.	Small, elongated or ovoidal. Operculum. No "shoulders" at anterior end. Small "knob" often seen on posterior end.	Yellow or yellow brown.	Embryonated. Contains mature miracidium.	Resembles *Clonorchis* and *Heterophyes* eggs. Shell is slightly thinner than *Heterophyes*. Operculum is broader than *Clonorchis*.
Paragonimus westermani	85 µm x 53 µm. Range, 68-118 µm x 39-67 µm.	Ovoidal or elongate with thick shell. Operculum is slightly flattened and fits into shoulder area of shell. Posterior end is thickened. Egg often asymmetrical with one side slightly flattened.	Yellow brown to dark brown.	Unembryonated. Filled with yolk material in which a germinal cell is imbedded. Cells are irregular in size.	Found in sputum, occasionally in feces. Resembles egg of *D. latum* but is larger, slightly asymmetrical and the operculum is smaller and flatter. The widest part of the *Paragonimus* egg is usually anterior to the center ; in a *D. latum*, the widest area is around the center.

Fasciola hepatica	145 μm x 80 μm. Range, 120-150 μm x 63-90 μm.	Ellipsoidal, thin shell. Small, indistinct operculum.	Yellow to light brown.	Unembryonated. Filled with yolk cells in which an indistinct germinal cell is imbedded.	Large size. Broadly oval eggs.
Fasciolopsis buski	140 μm x 80 μm. Range, 130-159 μm x 78-98 μm.	Ellipsoidal, thin shell. Small, indistinct operculum.	Yellow brown.	Unembryonated. Filled with yolk cells in which an indistinct germinal cell is imbedded.	Large size. Resembles *F. hepatica* egg and cannot be easily distinguished from *Fasciola*.

Table 6-1. Differential Morphology of the Diagnostic Stages of Helminths Found in Humans: Eggs (Trematodes) (SOURCE: CDC)

Trematodes require an intermediate host in their life cycle with vertebrates being the definitive host. Larval stages may occur in either invertebrate or vertebrate hosts.

There are three groups of trematodes:

- Monogenea, which typically are external parasites of fish with direct life cycles
- Aspidogastrea, these are endoparasites with the entire ventral surface as an adhesive organ
- Digenea, these are endoparasites with simpler adhesive organs and life cycles involving one or more intermediate hosts (indirect life-cycle). This section concentrates on the Digenean trematodes.

Most Digenean trematodes inhabit the alimentary canal of vertebrates and many of the associated organs, such as the liver, bile duct, gall bladder, lungs, bladder and ureter. These organs are rich in cavities containing food such as blood, mucus, bile and intestinal debris.

The Digenean trematodes have a complex life cycle, with rare exceptions, always involve a mollusk host. There may be six larval stages – the miracidium, sporocyst, redia, cercaria, mesocercaria (rare) and the metacercaria (the majority have 4 or 5 stages).

Trematode eggs have a smooth hard shell and the majority of them are operculate.

Fasciola species

Introduction

Fasciola, *Fasciolopsis* and *Echinostoma* species are trematodes which parasitize the liver and intestines of a variety of vertebrates. They are hermaphroditic and their distinguishing characteristics are shown in **Table 6-2**.

Fasciola hepatica trematodes are not thought to infect man but in fact man is not an unusual host, with infections being reported in many countries including Europe and the USA. The eating of unwashed watercress appears to be the source of infection, with them ending up in the liver. The most common host is sheep where they can cause severe disease.

Fasciolopsis buski (giant intestinal fluke) is a duodenal parasite infecting both man and pigs. They are found widespread in Asia and China, but they have been found to be endemic in Taiwan, Thailand, Bangladesh and India. Night soil (human excreta) is used as a fertilizer in these countries on plants such as water chestnut and caltrops. The snails graze on these crops and also the definitive hosts eat them raw and unwashed, peeling the edible water plants with their teeth.

Infection with *Echinostoma* species is thought to be contracted by ingestion of fresh water snails containing metacercaria. Such as *Echinostoma ilocannum* which occurs in the Philippines. The metacercariae infect the large snail *Piola luzionica* and in return are eaten raw.

Despite the large numbers of these flukes they are of little medical importance, the most important being *F. buski.*

Species	Geographic Distribution	Reservoir Hosts	Location of adult in host	Size of Ova
Fasciola hepatica	Cosmopolitan	Sheep	Bile Ducts	130-150µm by 63-90µm
Fasciola gigantica	Africa, the Orient and Hawaiian Islands	Camels, Cattle and Water Buffalo	Bile Ducts	160–190µm by 70-90µm
Fasciolopsis buski	Far-East and Indian Sub-continent	Pigs, Dogs and Rabbits	Intestine	130–140µm by 80-85µm
Echinostoma species	South East Asia and Japan	Variety of Mammals	Intestine	88–116µm by 58-69µm

Table 6-2. Table describing the characteristics which differentiate the various *Fasciola* species which are important to man.

Life Cycle and Transmission

The life cycles of *Fasciola*, *Fasciolopsis* and *Echinostoma* species are complex, requiring more than one intermediate host.

Adult worms inhabit the liver or bile ducts of the definitive host (human), where they lay many eggs which are deposited into the environment in the feces. They are immature when passed. If they are passed into water they become mature in nine to 15 days at the optimum temperature of 22-25°C.

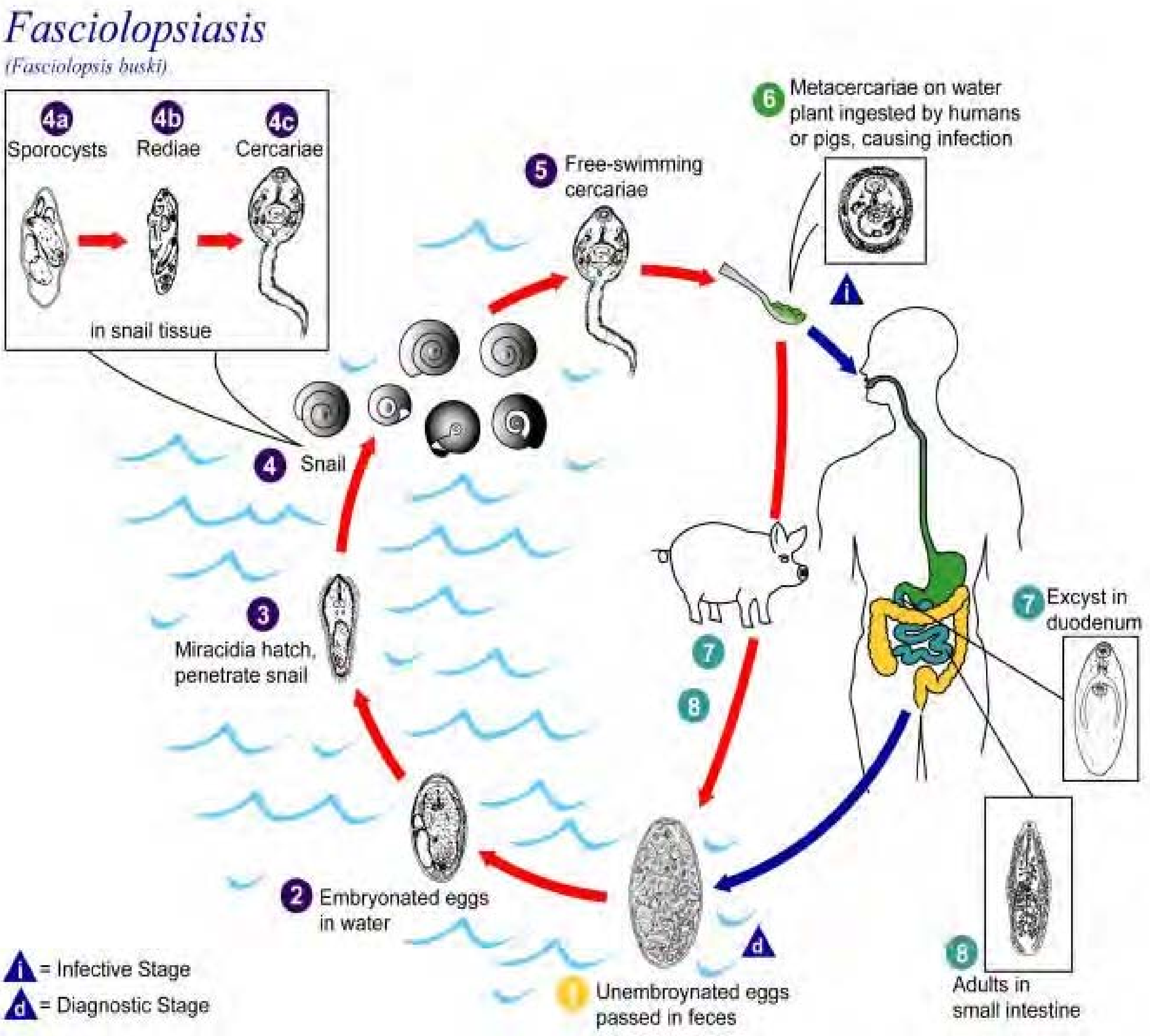
Fasciolopsiasis
(Fasciolopsis buski)
4a Sporocysts
4b Rediae
4c Cercariae
in snail tissue
4 Snail
5 Free-swimming cercariae
6 Metacercariae on water plant ingested by humans or pigs, causing infection
i
3 Miracidia hatch, penetrate snail
7
8
7 Excyst in duodenum
2 Embryonated eggs in water
d
1 Unembroynated eggs passed in feces
8 Adults in small intestine
i = Infective Stage
d = Diagnostic Stage

Illustration 6-2. The general life cycle of *Fasciola*, *Fasciolopsis* and *Echinostoma* species. Immature eggs are discharged in the biliary ducts and in the stool ①. Eggs become embryonated in water ②, eggs release miracidia ③, which invade a suitable snail intermediate host ④, including many species of the genus *Lymnae*. In the snail the parasites undergo several developmental stages (sporocysts (4a), rediae (4b), and cercariae (4c)). The cercariae are released from the snail ⑤and encyst as metacercariae on aquatic vegetation or other surfaces. Mammals acquire the infection by eating vegetation containing metacercariae. Humans can become infected by ingesting metacercariae-containing freshwater plants, especially watercress ⑥. After ingestion, the metacercariae excyst in the duodenum ⑦and migrate through the intestinal wall, the peritoneal cavity, and the liver parenchyma into the biliary ducts, where they develop into adults ⑧. In humans, maturation from metacercariae into adult flukes takes approximately 3 to 4 months. The adult flukes (*Fasciola hepatica*: up to 30 µm by 13 µm; *F. gigantica*: up to 75 µm) reside in the large biliary ducts of the mammalian host. *Fasciola hepatica* infect various animal species, mostly herbivores.

The life cycle of the *Echinostomes* differs by one minor point: the cercariae (4c)encyst wither within the tissues of the intermediate host ④in which sporocysts (4a) and rediae (4b)develop, or penetrate and encyst in other animals such as amphibians or fish. (SOURCE: PHIL 3393 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The morphology of the adult flukes of *Fasciola, Fasciolopsis* and *Echinostoma* species is well documented. They are large leaf-shaped parasites about 2–3cm long. There are two suckers, an anterior oral sucker surrounding the mouth and a ventral sucker (acetabulum) on the ventral surface.

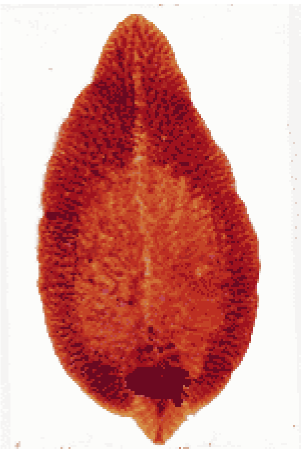

Image 6-1. Adult fluke of a *Fasciola* Trematode. Their morphology shows a large leaf-shape about 2–3cm long with two suckers, an oral and a ventral one. (SOURCE: CDC)

The outer tegument is covered in tiny spines which face backwards enabling them to attach themselves along with their suckers to the tissues.

Ova are all thin shelled, ellipsoid, quinone colored (bile stained) with an operculum that is often inconspicuous. Although ova of *Echinostoma* species can usually be differentiated by size due these flukes being much smaller in size than *F. Buski* and *F. hepatica*, there is much cross-over in the size of *Fasciola* and *Fasciolopsis* species.

Pathogenesis

Light infections due to *Fasciola hepatica* may be asymptomatic. However, they may produce hepatic colic with coughing and vomiting; generalized abdominal rigidity, headache and sweating, irregular fever, diarrhea and anemia.

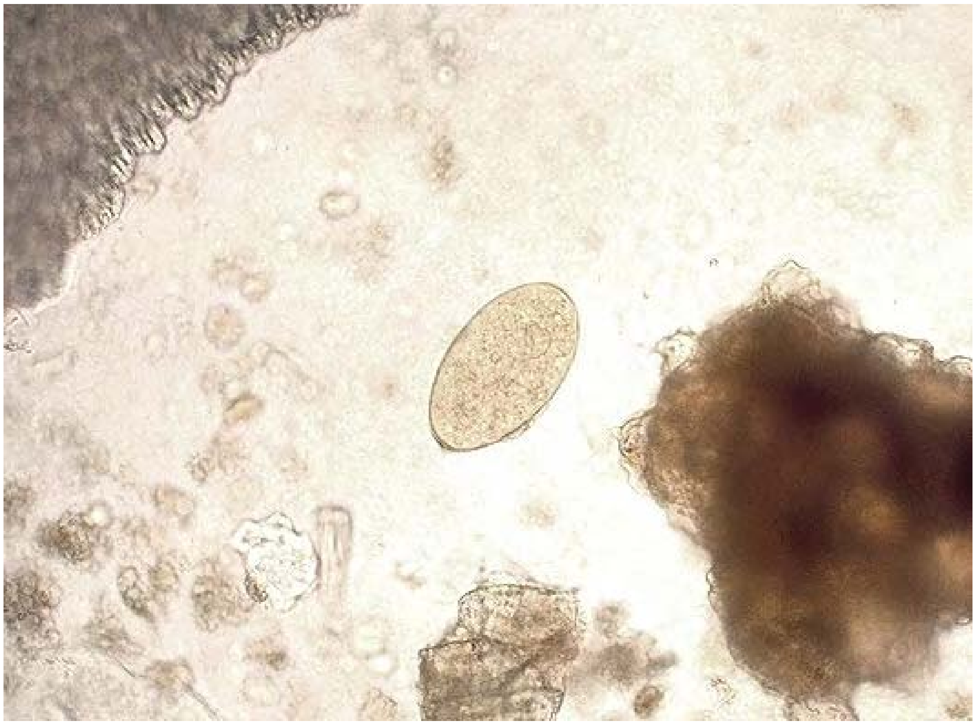

Image 6-2. Ova of *Fasciola* are ovoid in shape, quinone color and often showing an inconspicuous operculum. *Fasciola hepatica* ova measure 130 - 150µm by 63 - 90µm. There is much cross-over in ova size between all of the *Fasciola* species. (SOURCE: PHIL 1540 - CDC/Dr. Mae Melvin)

Infections due to *Fasciola gigantica* occur mainly in cattle raising areas and cause clinical symptoms similar to those of *Fasciola hepatica* although human infections are less common.

The adult flukes of *Fasciolopsis buski* attach to the intestine, resulting in local inflammation and ulceration. Heavier infections may subsequently lead to abdominal pain, malabsorption and persistent diarrhea, edema and even intestinal obstruction. Marked eosinophilia may be seen.

The adult flukes of *Echinostoma* species attach to the intestine resulting in little damage to the intestinal mucosa. Light infections are generally asymptomatic and heavy infections may produce light ulceration, diarrhea and abdominal pain.

Laboratory Diagnosis

Definitive diagnosis is made by observing the ova in feces, since the flukes are very prolific any significant infection will be easily picked up. Where identification cannot be made from the size of the ova, clinical information and the source of infection may help to provide a diagnosis. Serological techniques are available for the diagnosis of *Fasciola hepatica.*

Clonorchis sinensis

Introduction

Clonorchis sinensis, also known as the Chinese (aka Oriental) liver fluke is a narrow elongate liver fluke found in the Far East, mainly Japan, Korea, China, Taiwan and Vietnam.

It belongs to the group of Oriental liver flukes where there are three main species which commonly infect man. The other two species are *Opisthorchis felineus* and *Opisthorchis viverrini*. (**Table 6-2**) The three species are so similar in their morphology, life cycles and pathogenicity that they are very rarely discussed as separate species.

All members of this group are parasites of fish-eating mammals, particularly in Asia and Europe. Man is the definitive hosts and water snails and fish are the intermediate hosts. Infections can be easily avoided by man not eating raw fish since this is the only way that infection can be passed on.

Clonorchis sinensis parasitize the biliary duct in humans who become infected by eating raw or undercooked fish. Dogs and cats are the most important reservoir hosts.

Life Cycle and Transmission

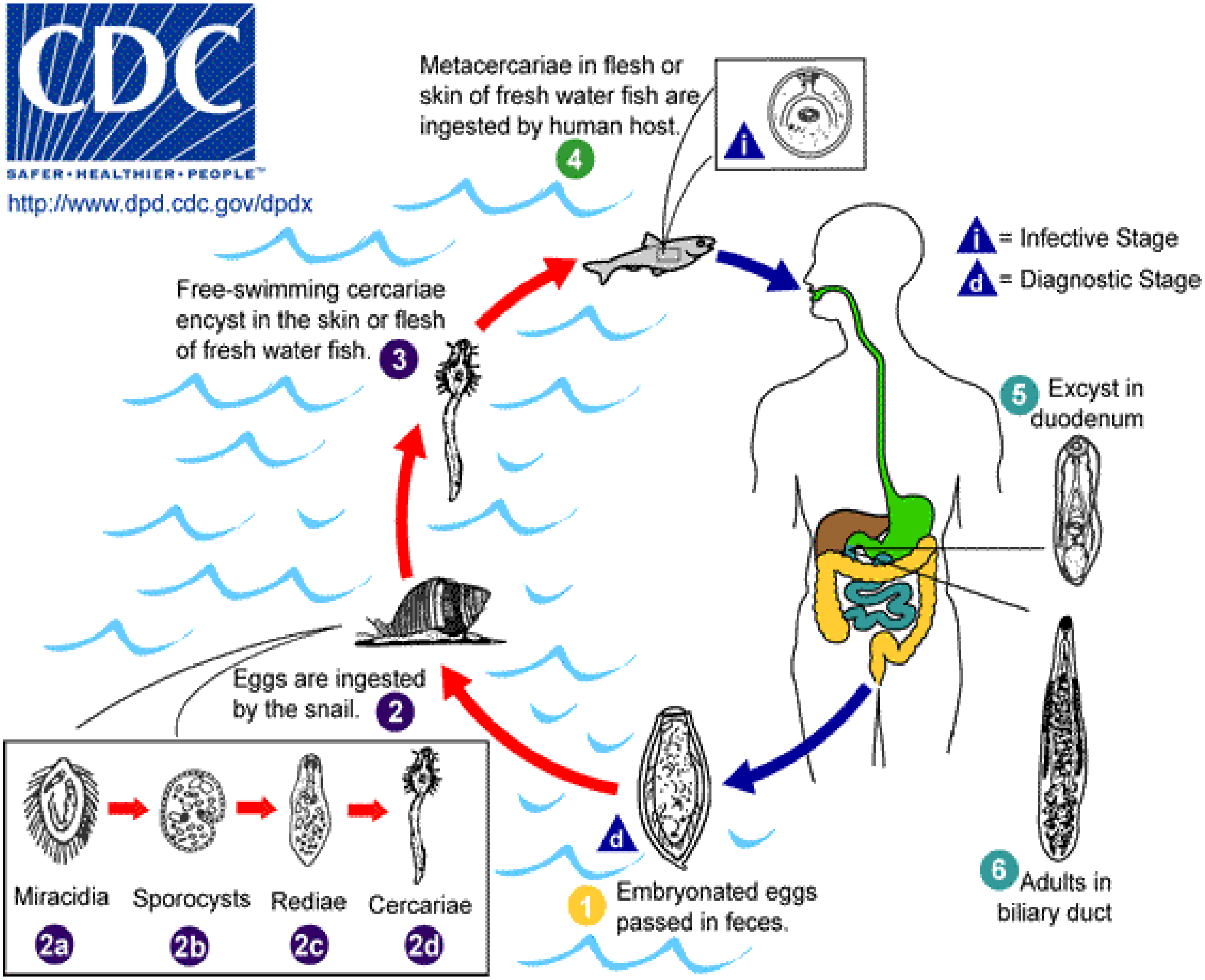
CDC
SAFER • HEALTHIER • PEOPLE™
http://www.dpd.cdc.gov/dpdx
Metacercariae in flesh or skin of fresh water fish are ingested by human host.
4
i
i = Infective Stage
d = Diagnostic Stage
Free-swimming cercariae encyst in the skin or flesh of fresh water fish.
3
5 Excyst in duodenum
Eggs are ingested by the snail.
2
Miracidia
Sporocysts
Rediae
Cercariae
2a
2b
2c
2d
d
1 Embryonated eggs passed in feces.
6 Adults in biliary duct

Illustration 6-3. Diagram illustrating the life cycle of *Clonorchis sinensis* (Chinese Liver Fluke). This parasite requires the involvement of two intermediate hosts (fresh water snails and fish) to complete the life cycle. Embryonated eggs are discharged in the biliary ducts and in the stool ①. Eggs are ingested by a suitable snail intermediate host ②; there are more than 100 species of snails that can serve as intermediate hosts. Each egg releases a miracidia (2a), which go through several developmental stages (sporocysts (2b), rediae (2c), and cercariae (2d)). The cercariae are released from the snail and after a short period of free-swimming time in water, they come in contact and penetrate the flesh of freshwater fish, where they encyst as metacercariae ③. Infection of humans occurs by ingestion of undercooked, salted, pickled, or smoked freshwater fish ④. After ingestion, the metacercariae excyst in the duodenum ⑤and ascend the biliary tract through the ampulla of Vater ⑥. Maturation takes approximately one month. The adult flukes (measuring 10 to 25 µm by 3 to 5 µm) reside in small and medium sized biliary ducts. In addition to humans, carnivorous animals can serve as reservoir hosts. (SOURCE: PHIL 3385 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The adult flukes measure 11–20µm by 3–4.5µm and are lanceolate in shape, translucent and brownish in color. They are all hermaphroditic. Keeping in common with other flukes they possess two suckers.

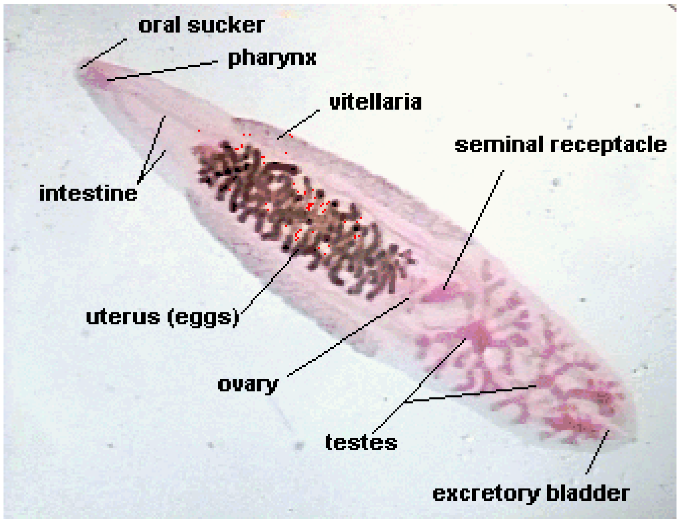
oral sucker
pharynx
vitellaria
seminal receptacle
intestine
uterus (eggs)
ovary
testes
excretory bladder

Image 6-3. Diagram illustrating the internal morphology of the Oriental liver fluke, *Clonorchis sinensis.* (SOURCE: SFSU.EDU)

The ova of *Clonorchis sinensis* contain fully developed miracidia and possess prominent opercular shoulders (flask shaped egg) and are operculate. They are bile stained and measure 29µm by 16µm. In wet mounts they are transparent and you can quite easily see their anatomy. There can be up to 6,000 worms present and a daily egg output of 1,000 eggs per microliter of bile or 600 per gram of feces.

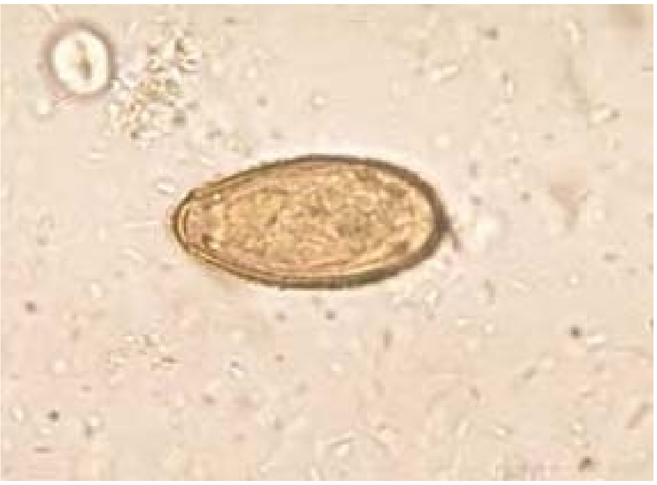

Image 6-4. Ova of *Clonorchis sinensis.* Showing the prominent opercular shoulders which makes identifying this trematode easy. They are described as flask shaped, bile stained. (SOURCE: PHIL 695 – CDC/Dr. Mae Melvin)

The cercariae possess eyespots, the penetration and cystogenous glands are also well developed.

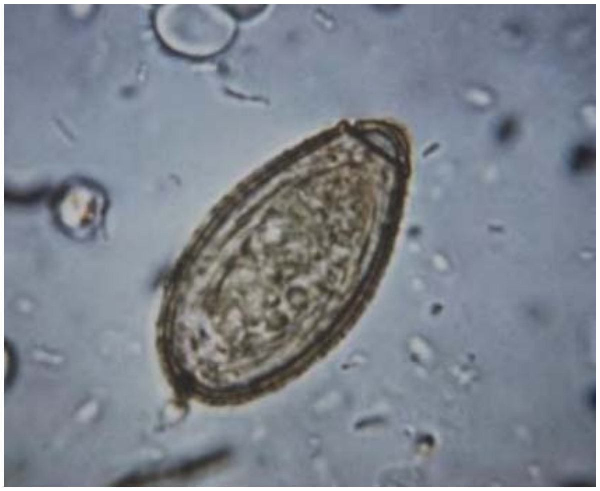

Image 6-5. Saline smear showing the characteristic flask shape of the *Clonorchis sinensis* ova. They are bile stained with a smooth outer coat. (SOURCE: PHIL 4845 – CDC)

Pathogenesis

Many millions of people become infected every year but only a minority suffers from any illness. The pathology is related to the number of parasites present. Light infections of up to 50 eggs or more are usually asymptomatic. A heavy infection of 500 or more eggs may cause serious illness.

Acute infections may be characterized by fever, diarrhea, epigastric pain, enlargement and tenderness of liver and sometimes jaundice. The invasion by these worms in the gall bladder may cause cholecystitis, due to flukes becoming impacted in the common bile duct.

Laboratory Diagnosis

Definitive diagnosis is made by observing the characteristic ova in feces following an iodine stained, formol-ether concentration method of the feces or from duodenal aspirates when there is complete obstructive jaundice or from the Entero-Test.

	Heterophyes heterophyes	***Metagonimus yokogawai***	***Opisthorchis viverrini***	***Dicrocoelium dendriticum***
Geographic distribution	Far East	Far East	Thailand	Far East
Location of adult in host	Small intestine	Small intestine	Liver and bile ducts	Liver and bile ducts
Size of ova	26.5-30μm by 15-17μm	26.5-30μm by 15-17μm	26.7μm by 15μm	38-45μm by 22-30μm
Shape of ova	Prominent opercular shoulders Bile stained	Prominent opercular shoulders Bile stained	Prominent opercular shoulders Bile stained	Dark brown, thick shelled and large operculum
Infection acquired by	Eating raw or pickled fish	Eating raw or pickled fish	Eating raw fresh water fish	Eating infected ants
Symptoms	Occasionally diarrhea and vomiting	Occasionally diarrhea and vomiting	Malaise and right upper quadrant pain	Biliary and digestive problems

Table 6-3. Table summarizing the less common flukes that are known to infect man. (CDC)

Paragonimus westermani

Introduction

Paragonimus westermani is a lung fluke found in both humans and animals. The adults are 12µm long and are found in capsules in the lung. Although they are hermaphroditic, it is necessary for worms to be present in the cyst for fertilization to occur. The disease is seen in the Far East, China, South East Asia, and America.

Life Cycle

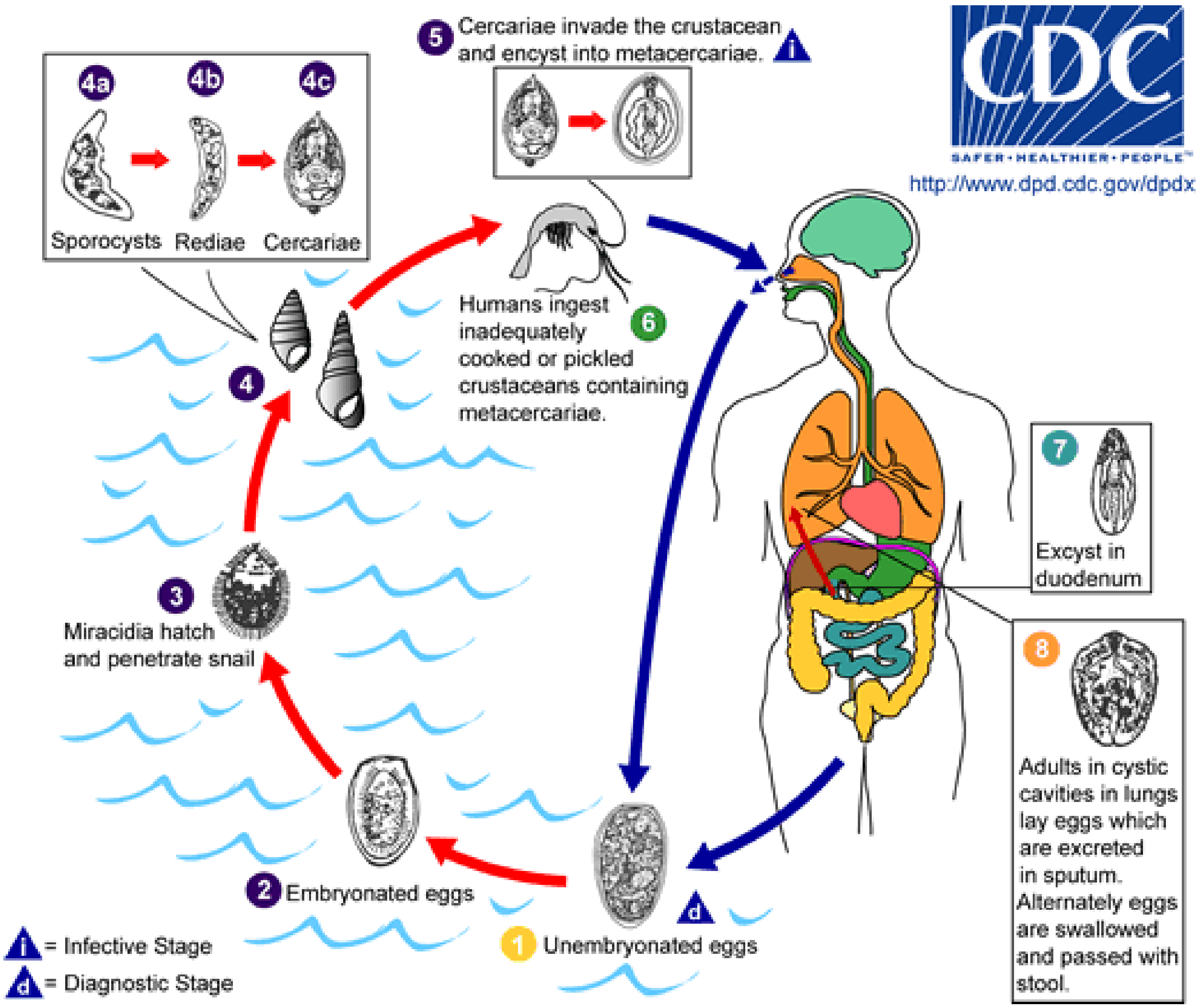
5 Cercariae invade the crustacean and encyst into metacercariae. i
4a 4b 4c
Sporocysts Rediae Cercariae
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Humans ingest inadequately cooked or pickled crustaceans containing metacercariae. 6
4
7 Excyst in duodenum
3 Miracidia hatch and penetrate snail
8 Adults in cystic cavities in lungs lay eggs which are excreted in sputum. Alternately eggs are swallowed and passed with stool.
2 Embryonated eggs
d
1 Unembryonated eggs
i = Infective Stage
d = Diagnostic Stage

Illustration 6-4. Diagram of the general life cycle of the lung fluke, *Paragonimus westermani*. The eggs are excreted unembryonated in the sputum, or alternately they are swallowed and passed with stool ❶. In the external environment, the eggs become embryonated ❷, and miracidia hatch and seek the first intermediate host, a snail, and penetrate its soft tissues ❸. Miracidia go through several developmental stages inside the snail ❹: sporocysts (4a), rediae (4b), with the latter giving rise to many cercariae (4c), which emerge from the snail. The cercariae invade the second intermediate host, a crustacean such as a crab or crayfish, where they encyst and become metacercariae. This is the infective stage for the mammalian host ❺. Human infection with *P. westermani* occurs by eating inadequately cooked or pickled crab or crayfish that harbor metacercariae of the parasite ❻. The metacercariae excyst in the duodenum ❼, penetrate through the intestinal wall into the peritoneal cavity, then through the abdominal wall and diaphragm into the lungs, where they become encapsulated and develop into adults ⑧(7.5 to 12 µm by 4 to 6 µm). The worms can also reach other organs and tissues, such as the brain and striated muscles, respectively. However, when this takes place completion of the life cycles is not achieved, because the eggs laid cannot exit these sites. Time from infection to oviposition is 65 to 90 days. (SOURCE: PHIL 3415 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Infections may persist for 20 years in humans. Animals such as pigs, dogs, and a variety of feline species can also harbor *P. westermani*.

Morphology

The adult worm is an ovoid, reddish brown fluke about 12µm long.

The eggs are ovoid, brownish yellow, thick shelled and operculated. They measure 80-100µm by 45-65µm and may be confused with the ova of *Diphyllobothrium latum*.

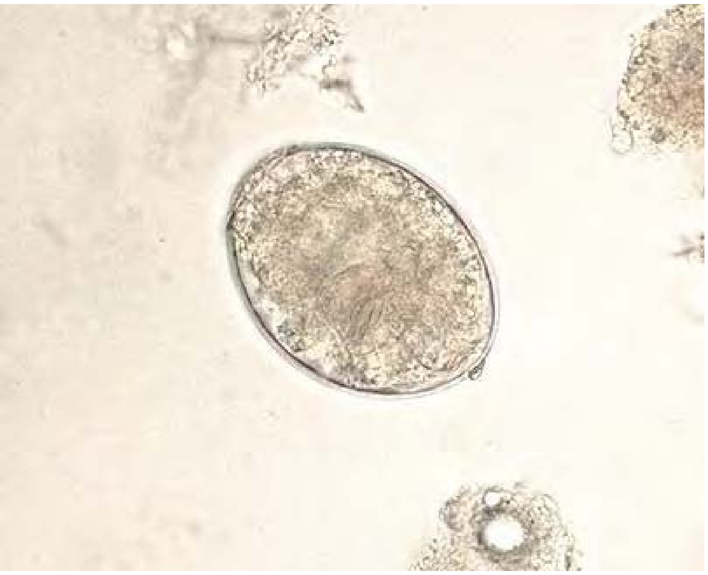

Image 6-6. Saline smear of *Paragonimus westermani* egg. The egg shells are thick and operculated. (SOURCE: PHIL 1541 – CDC/ Dr. Mae Melvin)

Clinical Disease

As the parasites grow in the lung cyst, inflammatory reaction and fever occurs. The cyst ruptures and a cough develops resulting in an increase in sputum. The sputum is frequently blood tinged and may contain numerous dark brown eggs and Charcot-Leyden crystals. Hemoptisis may occur after paroxysms of coughing. Dyspnea and bronchitis develop with time. Bronchiectasis may occur and pleural effusion is sometimes seen. The disease resembles pulmonary tuberculosis. Cerebral calcification may also occur.

Laboratory Diagnosis

Diagnosis is based on finding the characteristic eggs in brown sputum. The eggs can also be found in the feces due to swallowing sputum. A chest x-ray may show cystic shadows and calcification. Serological tests, in particular, the ELISA method, are useful diagnostic tests.

Schistosomes

Introduction

The Schistosomes are blood trematodes belonging to the Phylum *Platyhelmintha*. They differ from other trematodes in that they have separate sexes. The male worms resemble a rolled leaf where they bear the longer and more slender female in a ventral canal (the gynaecophoric canal). They require definitive and intermediate hosts to complete their life cycle. There are five species of *Schistosomes* responsible for human disease; *S. mansoni, S. haematobium* and *S. japonicum* with *S. mekongi* and *S. intercalatum* being less common.

They are the only trematodes that live in the blood stream of warm-blooded hosts. The blood stream is rich in glucose, and amino acids, so along with the plasma and blood cells, it represents an environment which is suitable for egg producing trematodes.

Over 200 million people are infected over at least 75 countries with 500 million or more people exposed to infection. With the disease spreading due to improved water supplies being created therefore, forming potentially new habits for snails. The disease caused is called schistosomiasis or Bilharzia and is the most important of helminth diseases.

Infection by the three most common species is the same in both sexes and in all age groups. Though, *S. mansoni* and *S. haematobium* is seen to occur more often and most heavily in teenagers especially males.

Life Cycle

Adult worms of *S. mansoni* live in the plexus of veins draining the rectum and colon, and in branches of the portal vein in the liver.

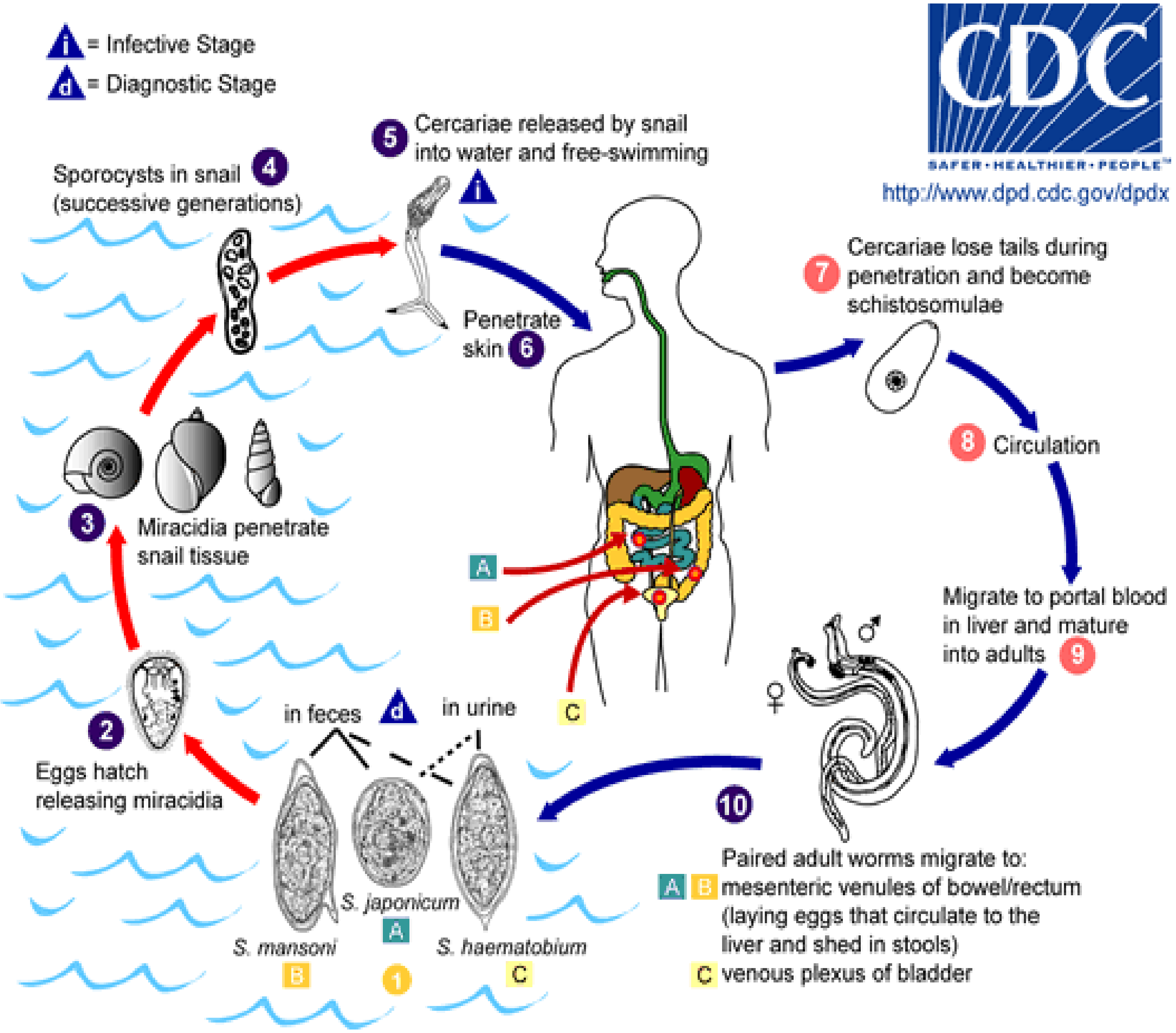
= Infective Stage
= Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Sporocysts in snail ❹
(successive generations)
❺ Cercariae released by snail
into water and free-swimming
i
Penetrate
skin ❻
⑦ Cercariae lose tails during
penetration and become
schistosomulae
⑧ Circulation
❸
Miracidia penetrate
snail tissue
Migrate to portal blood
in liver and mature
into adults ⑨
A
B
C
♂
♀
❷
Eggs hatch
releasing miracidia
in feces
d
in urine
❿
Paired adult worms migrate to:
A B mesenteric venules of bowel/rectum
(laying eggs that circulate to the
liver and shed in stools)
C venous plexus of bladder
S. japonicum
A
S. mansoni
B
S. haematobium
C
①

Illustration 6-5. Diagram illustrating the general life cycle of the *Schistosomes*. Eggs are eliminated with feces or urine ①. Under optimal conditions the eggs hatch and release miracidia ❷, which swim and penetrate specific snail intermediate hosts ❸. The stages in the snail include 2 generations of sporocysts ❹and the production of cercariae ❺. Upon release from the snail, the infective cercariae swim, penetrate the skin of the human host ❻, and shed their forked tail, becoming schistosomulae ⑦. The schistosomulae migrate through several tissues and stages to their residence in the veins (⑧, ⑨). Adult worms in humans reside in the mesenteric venules in various locations, which at times seem to be specific for each species ❿. For instance, *S. japonicum* is more frequently found in the superior mesenteric veins draining the small intestine A, and *S. mansoni* occurs more often in the superior mesenteric veins draining the large intestine B. However, both species can occupy either location, and they are capable of moving between sites, so it is not possible to state unequivocally that one species only occurs in one location. *S. haematobium* most often occurs in the venous plexus of bladder C, but it can also be found in the rectal venules. The females (size 7 to 20 µm; males slightly smaller) deposit eggs in the small venules of the portal and perivesical systems. The eggs are moved progressively toward the lumen of the intestine (*S. mansoni* and *S. japonicum*) and of the bladder and ureters (*S. haematobium*), and are eliminated with feces or urine, respectively①. Pathology of *S. mansoni* and *S. japonicum* schistosomiasis includes: Katayama fever, hepatic perisinusoidal egg granulomas, Symmers' pipe stem periportal fibrosis, portal hypertension, and

> occasional embolic egg granulomas in brain or spinal cord. Pathology of *S. haematobium* schistosomiasis includes: hematuria, scarring, calcification, squamous cell carcinoma, and occasional embolic egg granulomas in brain or spinal cord. Human contact with water is thus necessary for infection by schistosomes. Various animals, such as dogs, cats, rodents, pigs, hourse and goats, serve as reservoirs for *S. japonicum*, and dogs for *S. mekongi*. (SOURCE: CDC)

Adults of *S. japonicum* live in the anterior mesenteric blood vessels and in the portal vein in the liver; while the adults of *S. haematobium* live in the vesical plexus draining the bladder.

Once the eggs are laid by the adult female worms the majority of them first pass through the veins of the blood vessel in which the worm is living, and then into the lumen of the intestine and are passed in the feces (*S. mansoni* and *S. japonicum*), or into the lumen of the bladder, and are then passed in the urine (*S. haematobium).* Those eggs that reach fresh water hatch, releasing a miracidium which, to develop further must infect a snail of the correct species within 24 hours. The eggs of each species are markedly different but each produce virtually identical miracidium.

Asexual multiplication takes place in the snail, and results in the release of cercariae (minute in size with forked tails, 200µm long) into the water about 3–6 weeks later. Cercariae actively swim around and when they have located, or come into contact with, a definitive host they actively penetrate the skin. They can stay active looking for a host for 24–48 hours after which if they don't find a host they will die. The head of the cercariae migrates to the liver and develops into either adult male or female worms (flukes), here they pair up and then migrate to their region of the venous blood system (species specific sites). The females leave the males and moves to smaller venules closer to the lumen of the intestine or bladder to lay her eggs (about six weeks after infection). The majority of adult worms live from 2–4 years, but some can live considerably longer.

Schistosoma mansoni

Introduction

S. mansoni occurs in West and Central Africa, Egypt, Malagasy, the Arabian Peninsula, Brazil, Surinam, Venezuela and the West Indies. The intermediate host is an aquatic snail of the genus *Biomphalaria*. Man is the most common definitive host, occasionally baboons and rats are infected.

The adult worms live in smaller branches of the inferior mesenteric vein in the lower colon.

Morphology

The adult males measure up to 15 millimeters in length and females up to 10µm. The schistosomes remain in copula throughout their life span, the uxorious male surrounding the female with his gynaecophoric canal. The male is actually flat but the sides roll up forming the groove. The cuticle of the male is covered with minute papillae. The female only posses these at the anterior and posterior end as the middle section being covered by the male body. Oral and ventral suckers are present, with the ventral one being lager serving to hold the worms in place, preventing them being carried away by the circulatory current.

The ova of *S. mansoni* are 114-175µm long by 45-68µm wide. They are light yellowish brown, elongate and possess a lateral spine. The shell is acid fast when stained with modified Ziehl-Neelsen Stain.

A non-viable egg is dark colored and shows no internal structural detail or flame cell movement. Eggs can become calcified after treatment and are usually smaller, appear black and often distorted with a less distinct spine.

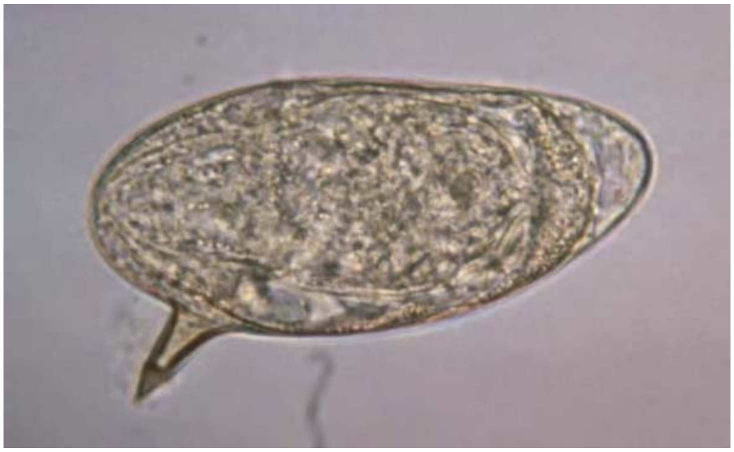

Image 6-7. Micrograph of a *S. mansoni* ova, clearly showing its lateral spine which is a good distinguishing factor when identifying *Schistosome* ova. They range in size between 114-175µm long by 45-68µm wide. (SOURCE: PHIL 4841 – CDC)

The schistosomes differ from other trematodes in that they are dioecious, digenetic, their eggs are not operculate and infection is acquired by penetration of cercaria through the skin.

Clinical Disease

The clinical disease is related to the stage of infection, previous host exposure, worm burden and host response. Cercarial dermatitis (swimmers itch) follows skin penetration and results in a maculopapular rash which may last 36 hours or more.

After mating, the mature flukes migrate to the venules draining the large intestine. Their eggs are laid and they penetrate the intestinal wall. They are then excreted in the feces, often accompanied by blood and mucus.

It is the eggs and not the adult worms, which are responsible for the pathology associated with *S. mansoni* infections. The adult flukes acquire host antigen which protects them from the host's immune response.

The host's reaction to the eggs which are lodged in the intestinal mucosa, leads to the formation of granulomata and ulceration of the intestinal wall. Some of the eggs reach the liver via the portal vein. The granulomatous response to these eggs can result in the enlargement of the liver with fibrosis, ultimately leading to portal hypertension and ascites. The spleen may also become enlarged. Other complications may arise as a result of deposition of the eggs in other organs e.g. lungs.

Katayama fever is associated with heavy primary infection and egg production. Clinical features include high fever, hepatosplenomegaly, lymphadenopathy, eosinohilia and dysentery. This syndrome occurs a few weeks after primary infection.

Laboratory Diagnosis

Microscopy

Laboratory confirmation of *S. mansoni* infection can be made by finding the eggs in the feces after an iodine stained, formol-ether concentration method. When eggs cannot be found in the feces, a rectal biopsy can be examined.

Serology

Serological tests are of value in the diagnosis of schistosomiasis when eggs cannot be found. An enzyme linked immunosorbent assay (ELISA) using soluble egg antigen, is employed at HTD.

Schistosoma japonicum

Introduction

Schistosoma japonicum is found in China, Japan, the Philippines, and Indonesia. It causes disease of the bowel with the eggs being passed out in the feces.

It differs form *S. mansoni* and *S. haematobium* in that it is a zoonosis in which a large number of mammals serve as reservoir hosts; cats, dogs and cattle playing major roles in the transmission of the disease.

The life cycle is not very different from that of *S. mansoni*, the intermediate hosts are from the subspecies *Oncomelania hupensis*. Sexual maturity is reached in about four weeks and eggs may be seen in the feces as quickly as five weeks.

The worms live coupled together in the superior, mesenteric veins and deposit 1500–3500 eggs per day in the vessels of the intestinal wall. The eggs infiltrate through the tissues and are passed in the feces.

Morphology

The adult worms are longer and narrower than the *S. mansoni* worms. The ova are about 55-85μm by 40-60μm, oval with a minute lateral spine or knob.

Clinical Disease

The main lesions are again due to the eggs, occurring in the intestine and liver. The eggs which are sequesters in the intestine mucosa or submucosa initiate granulomatous reactions, resulting in the formation of pseudotubercles.

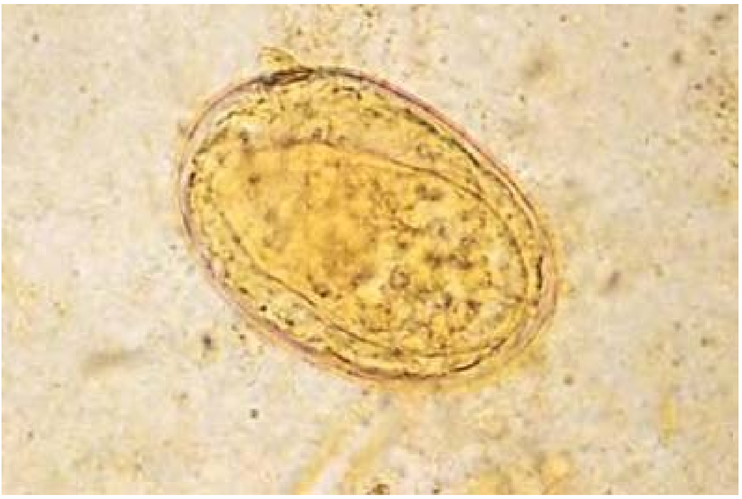

Image 6-8. Unstained micrograph of a *S. japonicum* ova. They are oval in shape with a minute lateral spine or knob. (SOURCE: PHIL 649 – CDC/Dr. Moore)

Due to the number of eggs released by the females the infection is more severe than one with *S. mansoni*. This is also due to the parasite being less well adapted to man, therefore, the circumoval granuloma is very large. The initial illness can be prolonged and sometimes fatal.

Laboratory Diagnosis

Microscopy

Laboratory confirmation of *S. japonicum i*nfection can be made by finding the eggs in the feces after an iodine stained, formol-ether concentration method. When eggs cannot be found in the feces, a rectal biopsy can be examined.

Other Intestinal *Schistosome* species

Other Schistosome species which are responsible for human disease are *S. mekongi and S. intercalatum.* These two species cause similar symptoms to that of *S. mansoni* and can be summarized in **Table 6-3**.

	S. mekongi	***S. intercalatum***
Geographic location	Mekong River basin	Central and west Africa
Diagnostic specimen	Stool, rectal biopsy, serology	Stool, rectal biopsy, serology
Egg size	30-55µm by 60-65µm	140-240µm by 50-85µm
Egg shape	Oval, minute lateral spine or knob	Elongate, terminal spine

Table 6-4. Table describing the other less common intestinal Schistosome species that are known to cause disease in man. (SOURCE: CDC)

Schistosoma haematobium

Introduction

Schistosoma haematobium is different from the other two species previously mentioned in that it causes urinary schistosomiasis. It occurs in Africa, India and the Middle East. The intermediate host is the *Bulinus* snail.

Just like *S. mansoni,* its distribution runs parallel to the irrigation projects and in areas which favor the intermediate hosts. They are exclusively parasites of man.

The mature worms live in copula mainly in the inferior mesenteric veins and the females deposit their eggs in the walls of the bladder and finally making their way into the urine. The life cycle is very similar to that of *S. mansoni*, with sexual maturity being reached within 4–5 weeks, but eggs may not appear in the urine until 10–12 weeks or even later.

Morphology

The adult worms are longer than those of *S. mansoni.* The ova are relatively large, measuring 110-170µm in length and 40-70µm in width. They have an elongated ellipsoid shape with a prominent terminal spine.

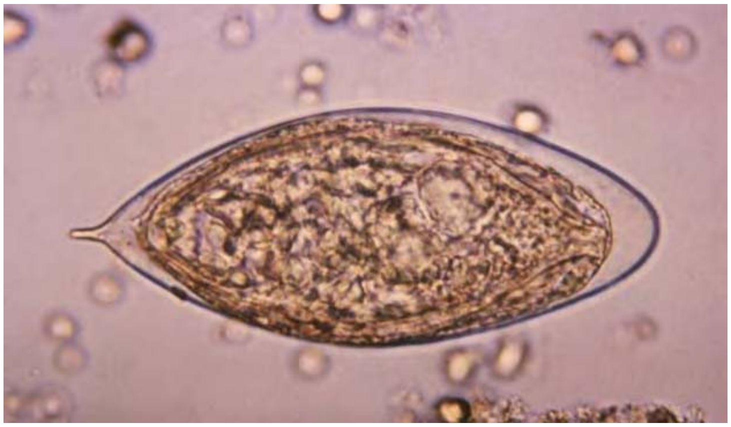

Image 6-9. *Schistosoma haematobium* eggs are elongated with a prominent terminal spine. The larva inside the egg produces an enzyme that passes through the egg-shell. (SOURCE: PHIL 4843 – CDC)

Clinical Disease

The clinical disease is related to the stage of infection, previous host exposure, worm burden and host response. Cercarial dermatitis (Swimmer's Itch) following skin penetration, results in a maculopapular rash and can last 36 hours or more. The mature flukes of *S. haematobium* migrate to the veins surrounding the bladder. After mating, the eggs are laid in the venules of the bladder and many penetrate through the mucosa, enter the lumen of the bladder and are excreted in the urine accompanied by blood. Thus hematuria and proteinuria are characteristic, though not invariable features of urinary schistosomiasis.

As with all *Schistosoma* species, it is the eggs and not the adult worms which are responsible for the pathology associated with *S. haematobium*. In chronic disease, eggs become trapped in the bladder wall resulting in the formation of granulomata. Following prolonged infection, the ureters may become obstructed and the bladder becomes thickened resulting in abnormal bladder function, urinary infection and kidney damage. Chronic urinary schistosomiasis is associated with squamous cell bladder cancer. Heavy infections in males may involve the penis resulting in scrotal lymphatics being blocked by the eggs.

Laboratory Diagnosis

The definitive diagnosis of urinary schistosomiasis is made by finding the characteristic ova of *S. haematobium* in urine. Terminal urine should be collected as the terminal drops contain a large proportion of the eggs. The urine can either be centrifuged and the deposit examined microscopically for ova. Eggs can sometimes be found in seminal fluid in males.

A bladder biopsy is seldom necessary to make the diagnosis. A rectal snip may show the presence of ova as they sometimes pass into the rectal mucosa.

Serological tests can be of value when eggs cannot be found in clinical samples. An enzyme linked immunosorbent assay using soluble egg antigen to detect antischistosome antibody is most sensitive.

There is a marked periodicity associated with the time when most eggs are passed out. Higher numbers of eggs are encountered in urine specimens passed between 1000 and 1400 hours, presumably as a result of changes in the host’s metabolic and physical activities.

Chapter 7.
Tissue Dwelling Nematodes

7.1 Unusual Tissue Dwelling Nematodes

Many species of nematodes are known to indirectly infect man causing an array of symptoms and diseases. Many of the larvae of these parasites cause serious zoonosis in man. These parasites are known to infect many intermediate hosts during their life cycle with a variety of animals being the definitive host, while man is normally an accidental definitive host.

Angiostrongylus (Parastrongylus) cantonensis

Introduction

Angiostrongylus cantonensis was first described in 1944 in Taiwan from a boy with suspected meningitis. Since then it has been recognized throughout the Pacific areas with sporadic cases reported in other parts of the world including Central America and Cuba It is associated with eosinophilic meningitis. The species *A. cantonensis* and *A. costaricensis* have now been reclassified into the genus *Parastrongylus.*

Angiostrongylus cantonensis is also known as the rat nematodes where the main cause of eosinophilic meningitis is infection with the larvae of this nematode. Man and rats become infected by eating infected mollusks of the giant land snail (intermediate hosts) or food which is contaminated by the snails' bodies.

Life Cycle and Transmission

Adult worms of *A. cantonensis* live in the pulmonary arteries of rats. The females lay eggs that hatch, yielding first-stage larvae, in the terminal branches of the pulmonary arteries. The first-stage larvae migrate to the pharynx, are swallowed, and passed in the feces. They penetrate, or are ingested by, an intermediate host (snail or slug). After two molts, third-stage larvae are produced, which are infective to mammalian hosts. When the mollusk is ingested by the definitive host, the third-stage larvae migrate to the brain where they develop into young adults. The young adults return to the venous system and then the pulmonary arteries where they become sexually mature. Of note, various animals act as paratenic (transport) hosts: after ingesting the infected snails, they carry the third-stage larvae which can resume their development when the paratenic host is ingested by a definitive host. Humans can acquire the infection by eating raw or undercooked snails or slugs infected with the parasite; they may also acquire the infection by eating raw produce that contains a small snail or slug, or part of one. There is some question whether or not larvae can exit the infected mollusks in slime (which may be infective to humans if ingested, for example, on produce). The disease can also be acquired by ingestion of contaminated or infected paratenic animals (crabs, freshwater shrimp). In humans, juvenile worms migrate to the brain, or rarely in the lungs, where the worms ultimately die. The life cycle of *Angiostrongylus (Parastrongylus) costaricensis* is similar, except that the adult worms reside in the arterioles of the ileocecal area of the definitive host. In humans, *A. costaricensis* often reaches sexual maturity and release eggs into the intestinal tissues. The eggs and larvae degenerate and cause intense local inflammatory reactions and do not appear to be shed in the stool.

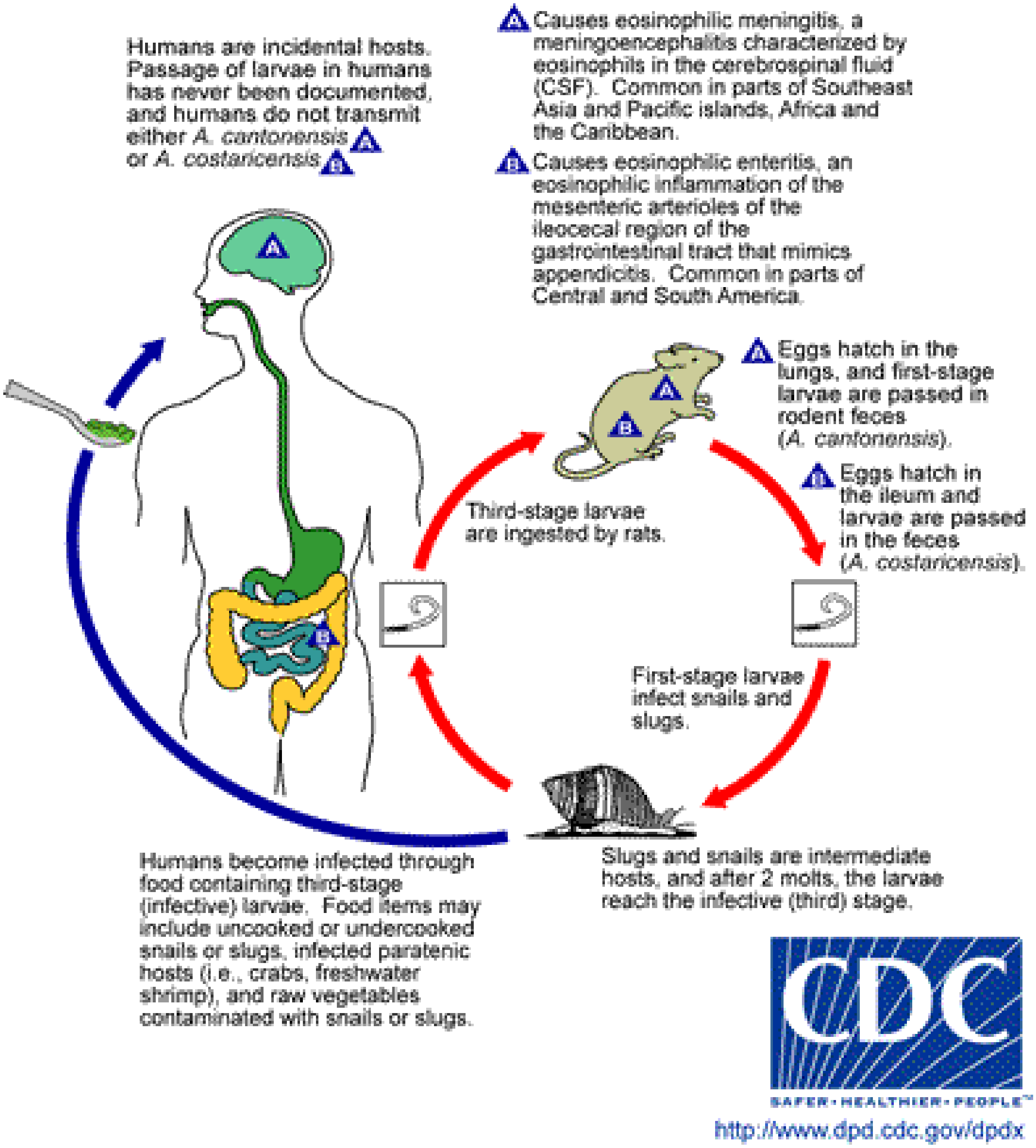
Humans are incidental hosts. Passage of larvae in humans has never been documented, and humans do not transmit either *A. cantonensis* A or *A. costaricensis* B
A Causes eosinophilic meningitis, a meningoencephalitis characterized by eosinophils in the cerebrospinal fluid (CSF). Common in parts of Southeast Asia and Pacific islands, Africa and the Caribbean.
B Causes eosinophilic enteritis, an eosinophilic inflammation of the mesenteric arterioles of the ileocecal region of the gastrointestinal tract that mimics appendicitis. Common in parts of Central and South America.
A
A Eggs hatch in the lungs, and first-stage larvae are passed in rodent feces (*A. cantonensis*).
A
B
B Eggs hatch in the ileum and larvae are passed in the feces (*A. costaricensis*).
Third-stage larvae are ingested by rats.
B
First-stage larvae infect snails and slugs.
Humans become infected through food containing third-stage (infective) larvae. Food items may include uncooked or undercooked snails or slugs, infected paratenic hosts (i.e., crabs, freshwater shrimp), and raw vegetables contaminated with snails or slugs.
Slugs and snails are intermediate hosts, and after 2 molts, the larvae reach the infective (third) stage.
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx

Illustration 7-1. Life cycle of the rat nematode *Angiostrongylus cantonensis.* Adult worms of *A. cantonensis* live in the pulmonary arteries of rats. The females lay eggs that hatch, yielding first-stage larvae, in the terminal branches of the pulmonary arteries. The first-stage larvae migrate to the pharynx, are swallowed, and passed in the feces. They penetrate, or are ingested by, an intermediate host (snail or slug). After two molts, third-stage larvae are produced, which are infective to mammalian hosts. When the mollusk is ingested by the definitive host, the third-stage larvae migrate to the brain where they develop into young adults. The young adults return to the venous system and then the pulmonary arteries where they become sexually mature. Of note, various animals act as paratenic (transport) hosts: after ingesting the infected snails, they carry the third-stage larvae which can resume their development when the paratenic host is ingested by a definitive host. Humans can acquire the infection by eating raw or undercooked snails or slugs infected with the parasite; they may also acquire the infection by eating raw produce that contains a small snail or slug, or part of one. There is some question whether or not larvae can exit the infected mollusks in slime (which may be infective to humans if ingested, for example, on produce). The disease can also be acquired by ingestion of contaminated or infected paratenic animals (crabs, freshwater shrimps). In humans, juvenile worms migrate to the brain, or rarely in the lungs, where the worms ultimately die. The life cycle of *Angiostrongylus (Parastrongylus) costaricensis* is similar, except that the adult worms reside in the arterioles of the ileocecal area of the definitive host. In humans, *A. costaricensis* often reaches sexual maturity and release eggs into the intestinal tissues. The eggs and larvae degenerate and cause intense local inflammatory reactions and do not appear to be shed in the stool. (SOURCE: CDC)

Morphology

Angiostrongylus cantonensis is a long slender worm measuring between 17 and 25 µm long by 0.26 to 0.34 µm wide.

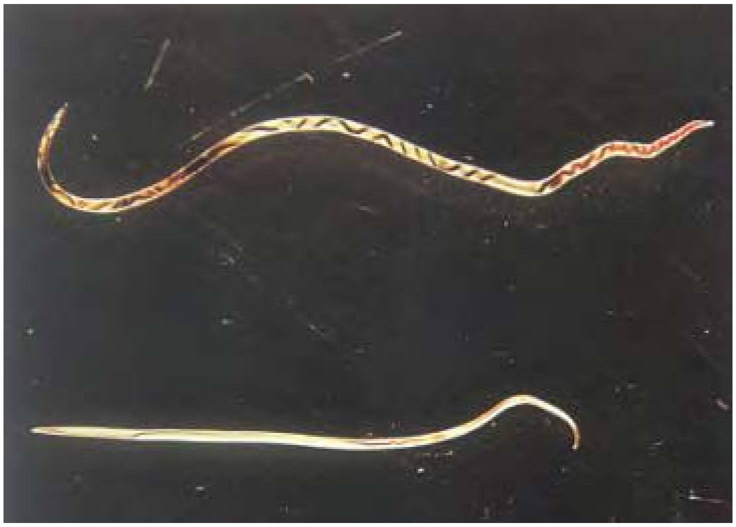

Image 7-1. The adult worms of *Angiostrongylus cantonensis.* They measure between 17–25µm long and reside in the pulmonary arteries and arterioles of the definitive hosts. (x 3.5) (SOURCE: CDC)

Symptoms

The incubation period is usually about 20 days but can be up to 47 days. Infection in man is usually self-limiting but can result in fatalities. The main symptom is severe headache but other symptoms include convulsions, vomiting, facial paralyses, paresthesia, neck stiffness and fever. It is accompanied by moderate eosinophilia in the CSF (i.e. the cell count is 100 – 2000/µm^3). Blood eosinophilia is also common. Eye involvement is accompanied by visual impairment, ocular pain, keratitis and retinal edema. Living worms have been reported and are removed surgically.

Laboratory Diagnosis

Presumptive diagnosis can be made on the basis of the patients' symptoms (i.e. fever, meningitis, ocular involvement and severe headache). Eosinophilia in CSF and peripheral blood is also suggestive. Larvae or young adults can be recovered in the CSF. ELISA methods can also provide confirmation.

Angiostrongylus (Parastrongylus) costaricensis

Introduction

Angiostrongylus costaricensis was first described in 1971. Human infections are most common in Costa Rica but have been reported in Mexico, Central and South America.

Life Cycle

The life cycle is similar to *A. cantonensis,* (above) with the human being an *accidental* host by consumption of snails and salad vegetables which have become contaminated by infective larvae that have been shed by slugs in their mucus. Unlike *A. cantonensis,* the larvae of *A. costaricensis* penetrates the intestinal wall and results in inflammatory lesions of the bowel wall. Here the life cycle ends.

Morphology

The eggs are 90µm, oval, thin shelled and can be embryonated. The adult female measures 42 x 300µm and the males measure 22 x 140µm.

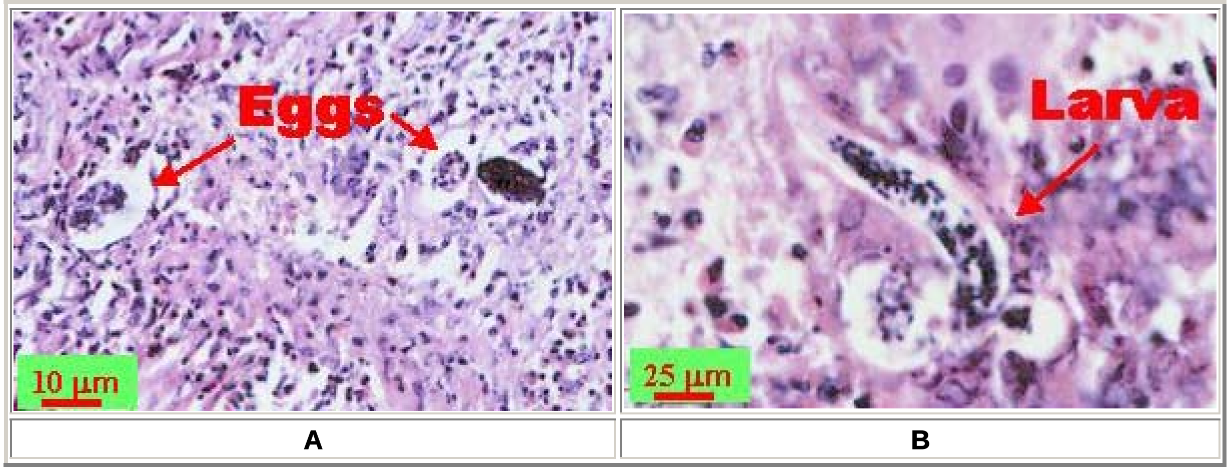

Image 7-2. **A:** Eggs of *Angiostrongylus costaricensis.* **B:** Larva of *A. costaricensis.* In humans, eggs and larvae are not normally excreted, but remain sequestered in tissues. Both eggs and larvae (occasionally adult worms) of *A. costaricensis* can be identified in biopsy or surgical specimens of intestinal tissue. The larvae need to be distinguished from larvae of *Strongyloides stercoralis*; however, the presence of granulomas containing thin shelled eggs and/or larvae serve to distinguish *A. costaricensis* infections. (SOURCE: CDC)

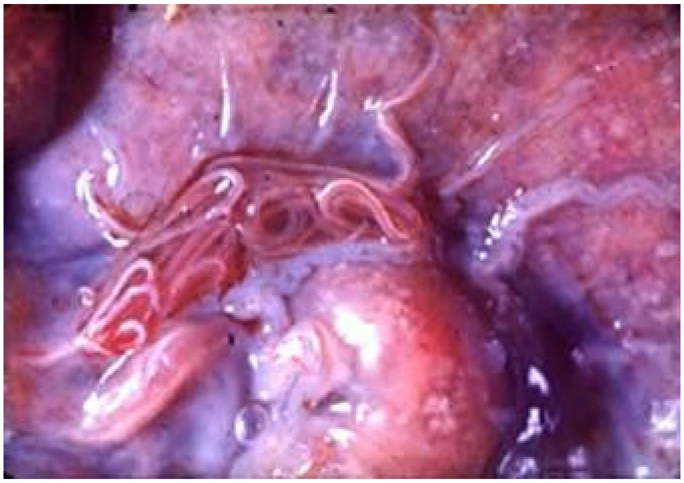

Image 7-3. Picture showing the adult *Angiostrongylus costaricensis* worms in the natural host of a rat. The worms normally localize within the mesenteric arteries, especially those of the ileocecal region of the natural definitive host. (SOURCE: CDC)

Clinical Disease

The most common symptoms are pain and tenderness with fever, vomiting and diarrhea. A tumor-like mass is often palpable and can mimic malignancy. The symptoms of abdominal pain, vomiting and diarrhea, and anorexia, are often mistaken for those of appendicitis. Worms can be found in the regional lymph nodes and mesenteric arteries. They can also be found in the spermatic arteries causing testicular obstruction and necrosis. Occasionally the larvae and the ova reach the liver and symptoms may mimic visceral larvae migrans. Eosinophilia is also present.

Anisakiasis

Introduction

Anisakiasis is a collective name for infections belonging to the genera of nematodes, *Anisakis, Phokanema, Terranova* and *Contracaecum*. The adult worms from the many species belonging to this family, *Anisakidae*, occur in various sea mammals, seals, dolphins, porpoises and whales, the larvae can cause a serious zoonosis, *anisakiasis*, in man. Anisakiasis was first reported in the Netherlands in 1960 and since then cases have been reported from Japan, North America, Canada, Chile, and the United Kingdom with the increase in popularity of ‘sushi’.

Two species of nematode have been especially implicated worldwide in causing human anisakiasis, *Anisakis simplex* and *Pseudoterranova osculatum.* The larvae of *P. osculatum* can be distinguished from *A. simplex* by the possession of a cecum.

Life Cycle

These nematodes are parasites in a large number of marine animals including seals, sea-lions, whales, and dolphins. It is in these mammals that the adult worm is found. Eggs are passed out in the feces of these mammals, they embryonate and hatch to liberate the first stage larvae into the sea water. These larvae then become ingested by crustaceans and molt to become L_2 and L_3 larvae. If fish and squid eat the infected crustaceans, the third stage larvae become liberated and penetrate the intestine or muscles of that host. These larvae then encyst. Over 100 species of fish can act as intermediate hosts. The marine mammals ingest the fish and thus the cycle is completed.

Humans become infected by consuming raw or improperly cooked fish which contain the third stage larvae.

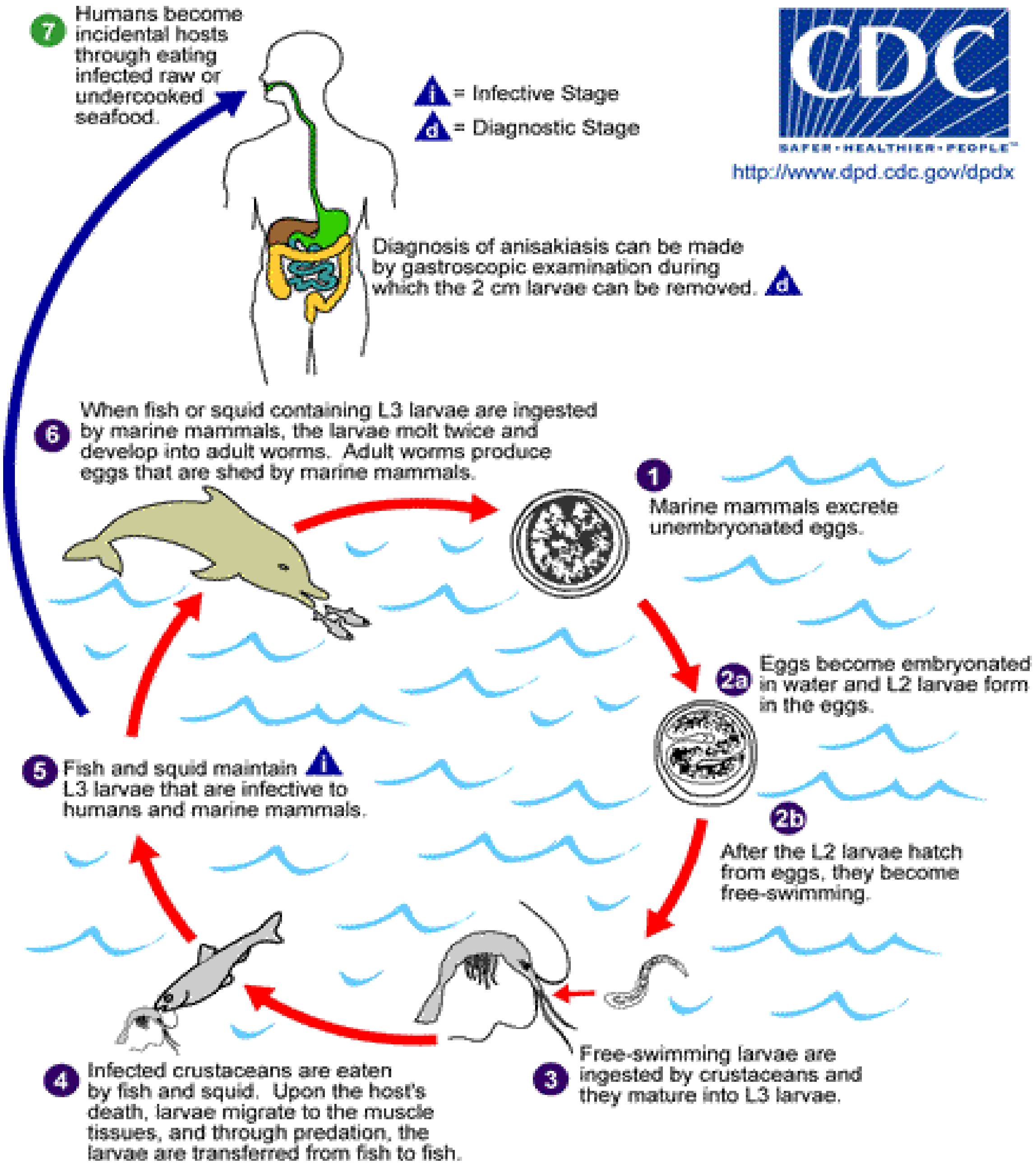
7 Humans become incidental hosts through eating infected raw or undercooked seafood.
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Diagnosis of anisakiasis can be made by gastroscopic examination during which the 2 cm larvae can be removed. d
6 When fish or squid containing L3 larvae are ingested by marine mammals, the larvae molt twice and develop into adult worms. Adult worms produce eggs that are shed by marine mammals.
1 Marine mammals excrete unembryonated eggs.
2a Eggs become embryonated in water and L2 larvae form in the eggs.
5 Fish and squid maintain i L3 larvae that are infective to humans and marine mammals.
2b After the L2 larvae hatch from eggs, they become free-swimming.
4 Infected crustaceans are eaten by fish and squid. Upon the host's death, larvae migrate to the muscle tissues, and through predation, the larvae are transferred from fish to fish.
3 Free-swimming larvae are ingested by crustaceans and they mature into L3 larvae.

Illustration 7-2. Adult stages of *Anisakis simplex* or *Pseudoterranova decipiens* reside in the stomach of marine mammals, where they are embedded in the mucosa, in clusters. Unembryonated eggs produced by adult females are passed in the feces of marine mammals ❶. The eggs become embryonated in water, and first-stage larvae are formed in the eggs. The larvae molt, becoming second-stage larvae (2a), and after the larvae hatch from the eggs, they become free-swimming (2b). Larvae released from the eggs are ingested by crustaceans ❸. The ingested larvae develop into third-stage larvae that are infective to fish and squid ❹. The larvae migrate from the intestine to the tissues in the peritoneal cavity and grow up to 3 cm in length. Upon the host's death, larvae migrate to the muscle tissues, and through predation, the larvae are transferred from fish to fish. Fish and squid maintain third-stage larvae that are infective to humans and marine mammals ❺. When fish or squid containing third-stage larvae are ingested by marine mammals, the larvae molt twice and develop into adult worms. The adult females produce eggs that are shed by marine mammals ❻. Humans become infected by eating raw or undercooked infected marine fish ❼. After ingestion, the anisakid larvae penetrate the gastric and intestinal mucosa, causing the symptoms of anisakiasis. (SOURCE: CDC)

Morphology

The larvae reach a length of 50µm with a diameter of 1-2µm. Classification of the Anisakids is made by the structure of the digestive tract.

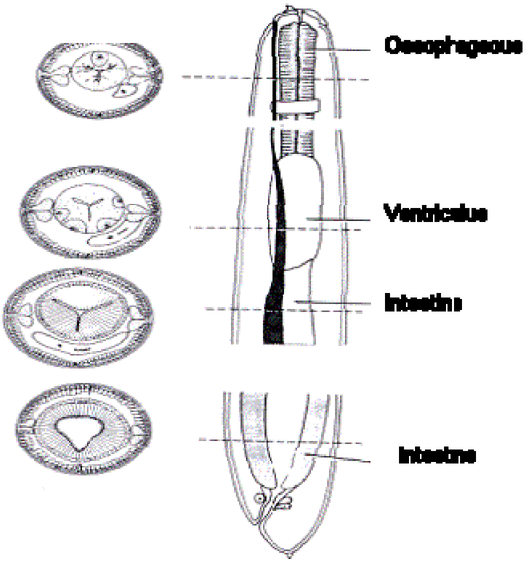

Illustration 7-3. Line diagram representing the morphology of the third stage larvae of *Anisakis simplex* (from fish), it is the causative organism of *anisakiasis* in man. (SOURCE: Smyth, J.D, 1994)

Clinical Disease

After ingestion of the raw fish by humans, the larvae penetrate the intestinal wall resulting in inhabitation of the stomach or duodenum, but can be found in any part if the alimentary canal or outside the gut in various viscera. The end result is abdominal pain, nausea, and sometimes vomiting and diarrhea, often occurring within 6 hours of eating an infected meal. Abdominal irritation may mimic gastric ulcer, carcinoma, appendicitis or other conditions requiring surgery. Eosinophilic granulosus may result.

Transient anisakiasis which has been reported in North America is characterized by some vomiting and distress within a few hours of ingesting the fish and then quickly subsides. The larvae can be coughed up a few days later. Low grade eosinophilia and occult blood in the stool are common.

Laboratory Diagnosis

Diagnosis can only be confirmed by endoscopy and the removal of the worms by biopsy forceps and microscopic identification.

Epidemiology and Prevention

Anisakis larvae are usually found in herring, mackerel, and North American salmon. *Pseudoterranova* are found in cod, halibut, rockfish, sardines, and squid.

Human infections result from eating raw fish, insufficiently cooked or smoked fish, marinated or salted fish. Freezing fish at -20°C (-4°F) for a minimum of five days kills the larvae and thus the fish is suitable for consumption in dishes like sushi. Smoking the fish at temperatures >65°C (149°F) renders it suitable for consumption. Marinating or salting the fish cannot be depended upon to kill the larvae, however adequately cleaning the fish can render it safe for consumption. Human anisakiasis has been virtually eliminated from the Netherlands due to the mandatory freezing of herring.

Gnathostoma spinigerum

Introduction

Several species of the genus Gnathostoma are responsible for zoonotic infections in man. *Gnathostoma spinigerum* is a nematode found in dogs, cats, and several other carnivores. Human infections of the disease have been reported from Japan, China, Thailand, the Far East, and the Philippines, with man acquiring the infection from eating various freshwater fish.

Life Cycle

The life cycle of this parasite involves two intermediate hosts, cyclops and birds, snakes, fish and frogs where they mature before developing into adults in the definitive hosts.

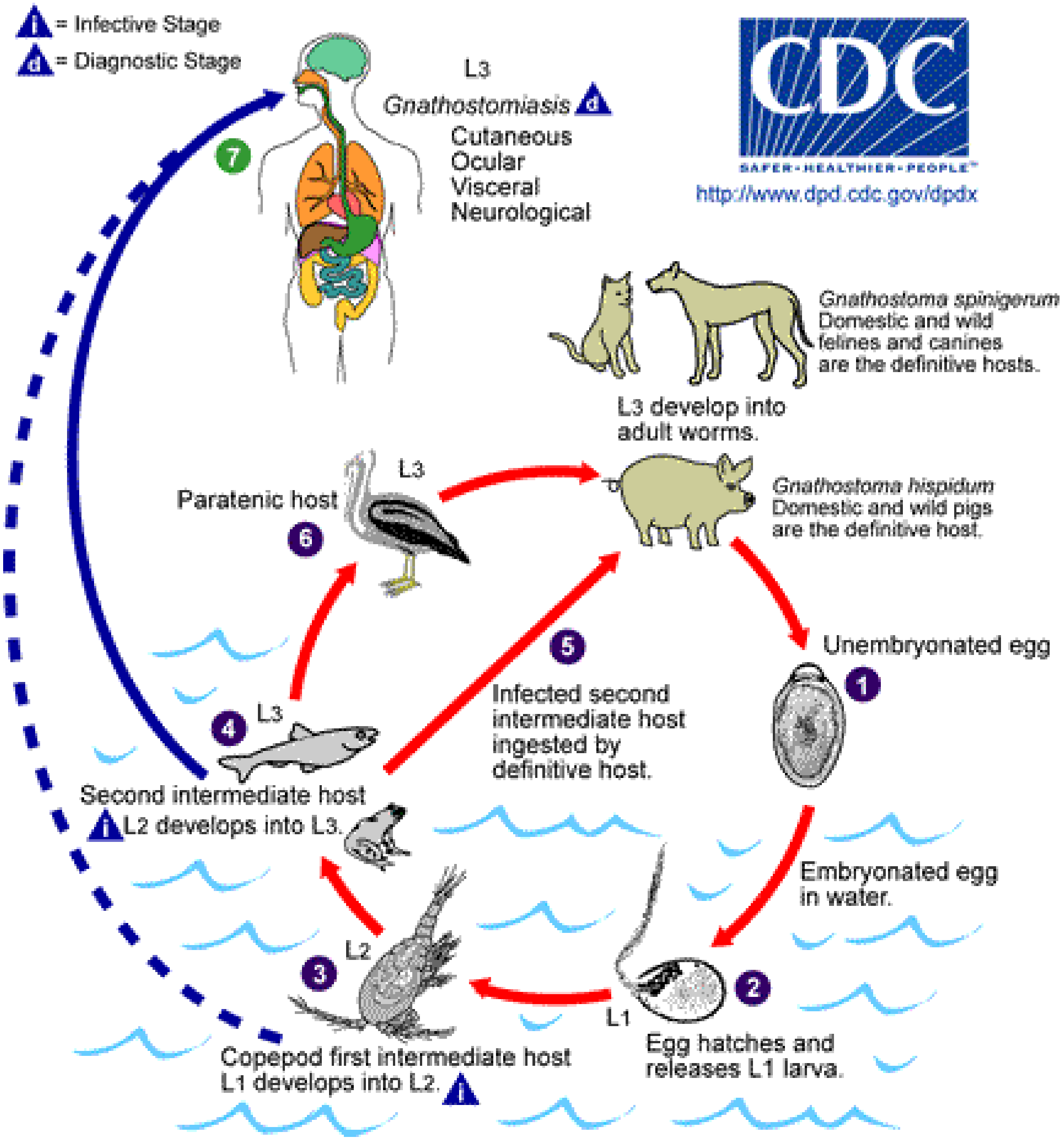
i = Infective Stage
d = Diagnostic Stage
L3
Gnathostomiasis d
Cutaneous
Ocular
Visceral
Neurological
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
7
Gnathostoma spinigerum
Domestic and wild felines and canines are the definitive hosts.
L3 develop into adult worms.
L3
Paratenic host
6
Gnathostoma hispidum
Domestic and wild pigs are the definitive host.
5
Unembryonated egg
1
Infected second intermediate host ingested by definitive host.
4
L3
Second intermediate host
i L2 develops into L3.
Embryonated egg in water.
L2
3
2
L1
Egg hatches and releases L1 larva.
Copepod first intermediate host
L1 develops into L2. i

Illustration 7-4. Diagram illustrating the life cycle of *Gnathostoma spinigerum.* In the natural definitive host (pigs, cats, dogs, wild animals) the adult worms reside in a tumor which they induce in the gastric wall. They deposit eggs that are unembryonated when passed in the feces ❶. Eggs become embryonated in water, and eggs release first-stage larvae ❷. If ingested by a small crustacean (*Cyclops,* first intermediate host), the first-stage larvae develop into second-stage larvae ❸. Following ingestion of the *Cyclops* by a fish, frog, or snake (second intermediate host), the second-stage larvae migrate into the flesh and develop into third-stage larvae ❹. When the second intermediate host is ingested by a definitive host, the third-stage larvae develop into adult parasites in the stomach wall ❺. Alternatively, the second intermediate host may be ingested by the paratenic host (animals such as birds, snakes, and frogs) in which the third-stage larvae do not develop further but remain infective to the next predator ❻. Humans become infected by eating undercooked fish or poultry containing third-stage larvae, or reportedly by drinking water containing infective second-stage larvae in *Cyclops* ❼.

If the infected fish or frogs are eaten by other hosts, apart from the definitive hosts, (paratenic host), such as herons, pigs and man, they do not mature but migrate through the subcutaneous tissues causing visceral and cutaneous larva migrans.

Morphology

The adult female worms measure 25-54µm whereas the male measures 11-25µm. The anterior half of the worm is covered with leaf like spines. The male worms have a red tail while the larger females have a more curved tail. The larval worms are 4µm long.

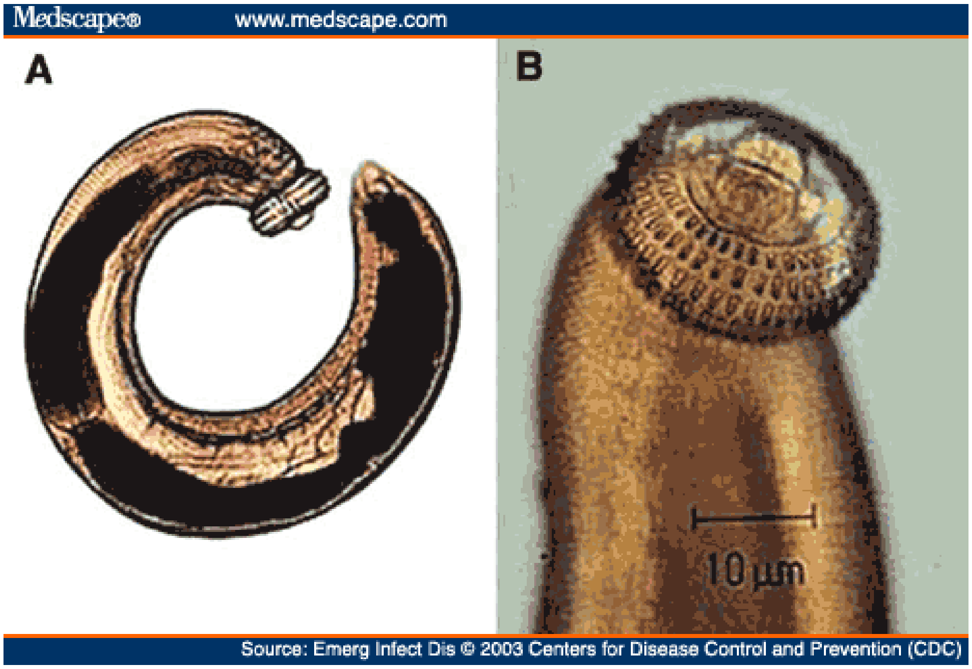
Medscape®
www.medscape.com
A
B
10 µm
Source: Emerg Infect Dis © 2003 Centers for Disease Control and Prevention (CDC)

Image 7-4. Adult worms of *Gnathostoma spinigerum.* The males are smaller than the females and possess a red tail. (x 1.3) (SOURCE: CDC/Medscape)

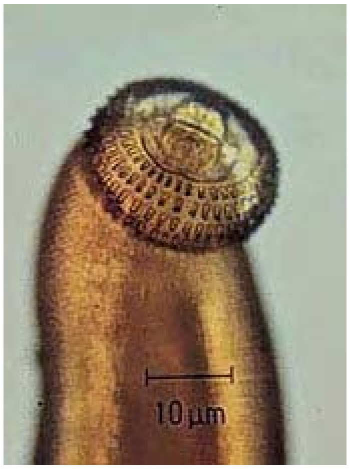

Image 7-5. Third stage larval Head of *Gnathostoma spinigerum.* (SOURCE: CDC/Medscape)

Clinical Disease

Humans are accidental hosts and after ingestion, the larvae do not mature but migrate throughout the body via the intestinal wall. Symptoms include epigastric pain, vomiting and anorexia. These symptoms subside as the larvae continue their migratory path through the cutaneous tissue. Evidence of migration appears as either lesions similar to cutaneous larvae migrans or migratory swellings accompanied by inflammation, redness or pain. The swelling is hard and non-pitted and lasts for several days. These migratory lesions may be accompanied by pruritis and pain. There is marked eosinophilia in patients with cutaneous involvement. Ocular involvement resulting in blindness may occur in serious disease. Eosinophilic myeloencephalitis may result from migration of the worms along the nerve tracks. Symptoms may include pain, paralyses, seizures, coma and death. The CSF may be xanthochromic (yellowish discoloration) or bloody.

Laboratory Diagnosis

Presumptive diagnosis may be made on the basis of clinical symptoms. Definitive diagnosis is the recovery and identification of the worm since the symptoms may be suggestive of Sparganosis, paragonimiasis and cutaneous larvae migrans and myiasis. A bloody spinal fluid or xanthochromia may resemble infection with *Angiostrongylus cantonensis.*

Trichinella spiralis

Introduction

Trichinella spiralis was first seen by James Paget, but was named and described by his Professor, Richard Owen. The family Trichinellidae contains only one single genus; *Trichinella,* and was originally thought only to contain the one species; *Trichinella spiralis,* which causes the serious and often fatal disease in man known as *trichinosis* (Trichinellosis). It is a parasite of carnivorous animals and is especially common in rats and in swine fed on uncooked garbage and slaughter house scraps. Humans become infected by eating raw pork, with sausages being the most common cause of infection. It is a cosmopolitan parasite and is common in China, though prevalent in many European countries as well.

It is now thought that there are four varieties of this species that exists worldwide;

- *Trichinella spiralis spiralis* - Temperate zone – high infectivity for pigs, rats and man.
- *Trichinella spiralis nelsoni* - Tropics - low infectivity for pigs and rats and high infectivity for lions, hyenas.
- *Trichinella spiralis nativa* - Arctic - low infectivity for pigs, found in polar bears, resistant to freezing.
- *Trichinella spiralis pseudospiralis*- New Zealand – low infectivity for pigs, rats and mice.

Trichinella spiralis is a ‘domestic’ parasitic nematode long recognized to cause a zoonosis transmitted to man by the ingestion of infected pork.

Life Cycle

Infection in the definitive hosts is acquired by the hosts eating raw or undercooked flesh (e.g. pork, containing encapsulated larvae). Rats are probably the most highly infected 'natural' hosts and pigs become infected by eating infected pork scraps or occasionally rats which inhabit their stalls. For man, sausages are a dangerous source of the parasite as a small fragment of infected pork, (after mincing), may become widely distributed among a number of sausages.

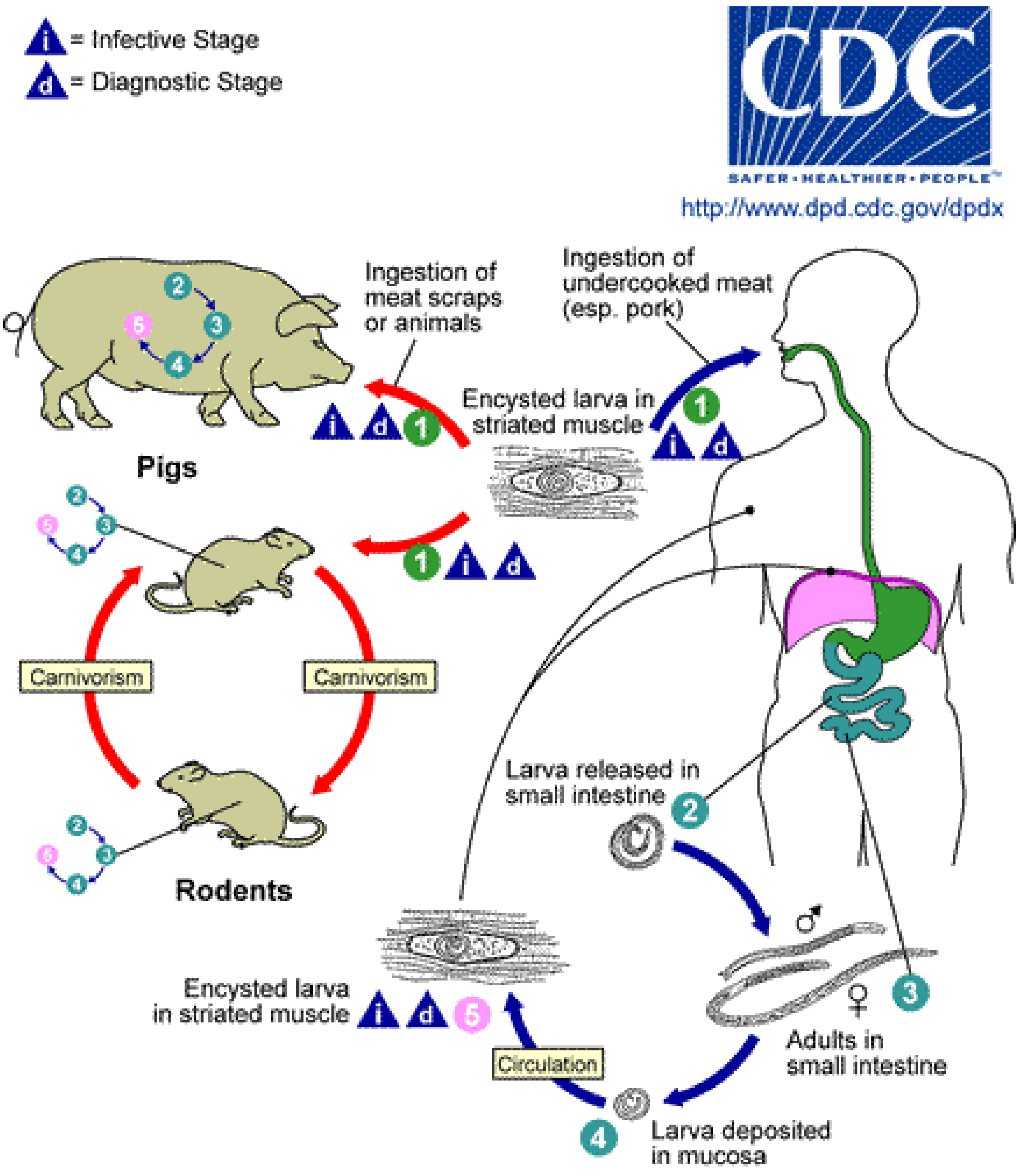
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Ingestion of meat scraps or animals
Ingestion of undercooked meat (esp. pork)
Encysted larva in striated muscle
Pigs
Carnivorism
Carnivorism
Rodents
Larva released in small intestine
Adults in small intestine
Larva deposited in mucosa
Circulation
Encysted larva in striated muscle
1
2
3
4
5
6

Illustration 7-5. The life cycle of *Trichinella spiralis*. Trichinellosis is acquired by ingesting meat containing cysts (encysted larvae) ① of *Trichinella*. After exposure to gastric acid and pepsin, the larvae are released ② from the cysts and invade the small bowel mucosa where they develop into adult worms ③(female 2.2 mm in length, males 1.2 mm; life span in the small bowel: four weeks). After one week, the females release larvae ④that migrate to the striated muscles where they encyst ⑤. *Trichinella pseudospiralis*, however, does not encyst. Encystment is completed in 4 to 5 weeks and the encysted larvae may remain viable for several years. Ingestion of the encysted larvae perpetuates the cycle. Rats and rodents are primarily responsible for maintaining the endemicity of this infection. Carnivorous/omnivorous animals, such as pigs or bears, feed on infected rodents or meat from other animals. Different animal hosts are implicated in the life cycle of the different species of *Trichinella*. Humans are accidentally infected when eating improperly processed meat of these carnivorous animals (or eating food contaminated with such meat). (SOURCE: CDC/DPDx Trichinellosis)

Morphology

The adult female worm is about 2-3µm long and 90µm in diameter. The male is smaller measuring 1.2µm long by 60µm in diameter.

The female adult worms are ovoviparous (born alive from the mother but held as an unshelled egg internally until birth, after having no umbilical connection to the mother) and up to 1500 larvae may be released by a single worm.

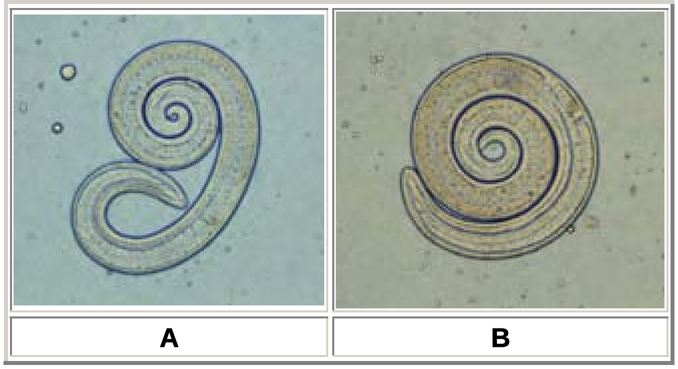

Image 7-6. A, B: Larvae of *Trichinella*, freed from their cysts, typically coiled; length: 0.8 to 1.0 µm. Alaskan bear. (SOURCE: CDC)

Clinical Disease

Symptoms during the intestinal phase may go unnoticed or may be severe. Epidemics can result in outbreaks of gastro-enteritis, two to seven days after the ingestion of raw pork. Diarrhea with or without abdominal pain may last for several weeks. Eosinophilia and fever occur in most cases. Leukocytosis is common and hyperglobulinaemia (an autoimmune, chronic blood disease) is characteristic. Myocytosis and circum orbital edema are classic signs. There can also be central nervous system involvement

Pathogenicity

The primary pathogenic effect of *Trichinella* comes from the destruction of the striated muscle fibers in which it encysts. There can be neurological manifestations of trichinosis and death may be ascribed to myocarditis, encephalitis or pneumonitis.

Laboratory Diagnosis

Diagnosis of trichinosis depends on the clinical signs, such as myalgia, periorbital edema, fever and eosinophilia in a patient with a history of eating pork or sausages.

Serological tests are available but may be negative if carried out within three to four weeks post infection. Circulating antibodies to *T.* spiralis appear from two to four weeks after infection. Redefined diagnostic antigens for their detection are currently being developed. A simple IFAT employing fragments of larvae as antigen is a useful diagnostic tool. Latex tests with extracted larval antigens have also proved valuable in the acute stage, during which high antibody titers develop.

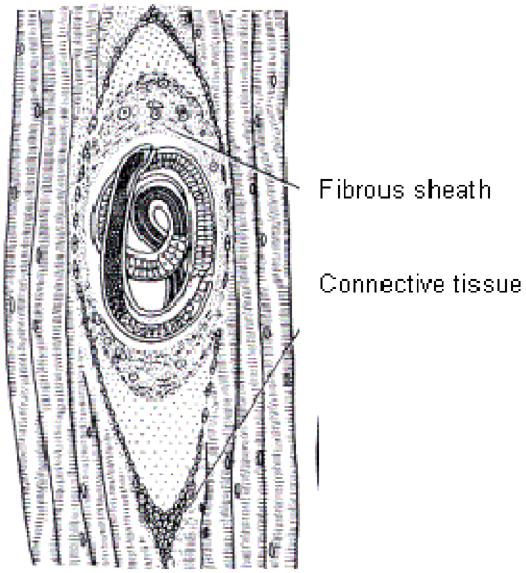

Illustration 7-6. Diagrammatic representation of the *Trichinella* larvae encapsulated in striated muscle. (SOURCE: Smyth, J.D, 1994)

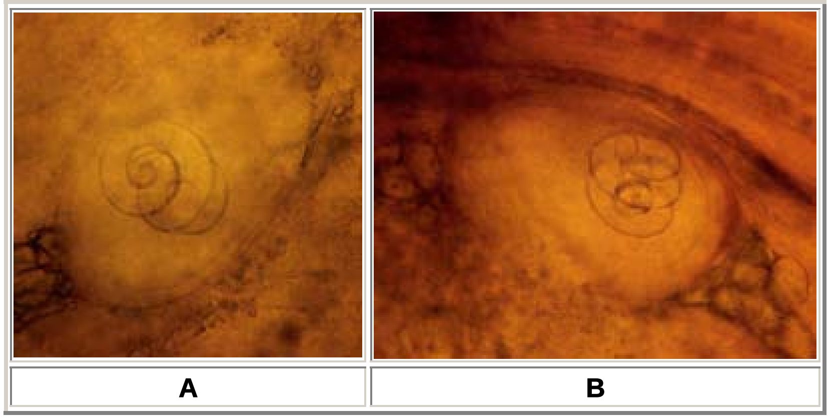

Image 7-7. A, **B:** Encysted larvae of *Trichinella* in pressed muscle tissue. The coiled larvae can be seen inside the cysts. (SOURCE: CDC)

Muscle biopsy is available with the muscle being digested in pepsin, which frees the encapsulated larvae or by a simple device whereby the muscle sample is compressed between two glass plates to make it semi-transparent, allowing you to see any encapsulated larvae using a ‘trichinoscope’ (a simple magnifying system).

Chapter 8.

Larval Cestodes and Nematodes which Infect Man

8.1 Larval Cestodes which Infect Man

Infections in man with *Echinococcus granulosus, Echinococcus multilocularis,* and *Multiceps multiceps* are caused by the accidental ingestion of eggs which are excreted by the definitive animal host. The disease that is produced due to the invasion of these parasites is caused by the larval stages or hydatid cyst, is known as hydatid disease or hydatidosis.

Echinococcus granulosus causes cystic echinococcosis, and is the form most frequently encountered; *E. multilocularis* causes alveolar echinococcosis; *E. vogeli* causes polycystic echinococcosis; and *E. oligarthrus* is an extremely rare cause of human echinococcosis.

Each cestode possesses an elongated tape-like body which lacks an alimentary canal. The adult tapeworms are strings of individuals having a complete set of reproductive organs (proglottids) in progressive degrees of sexual maturity and budding off from a body attached to the host tissue by a head or scolex.

The larval stage, show a wide variation being found in almost any organ of both vertebrate and invertebrate hosts.

Echinococcus granulosus

Introduction

Echinococcosis or Hydatid disease in man is caused by the larval stage of the dog tapeworm, *Echinococcus granulosus*. Hydatid disease is most extensively found in East Africa, North Africa, South Africa, the Middle East, and parts of South America, and Australia. The intermediate hosts are cattle, sheep, pigs, goats, or camels and the definitive host for this disease is the dog or other canids.

Life Cycle

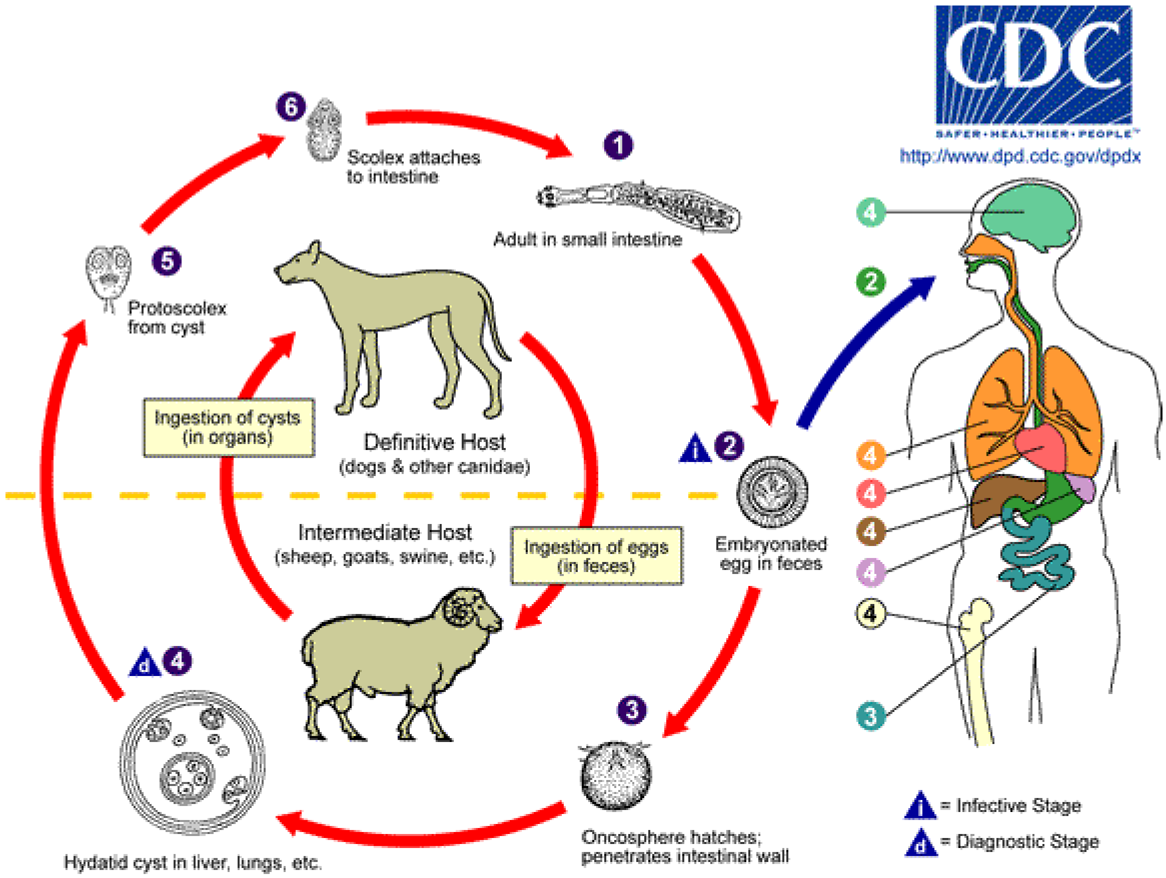
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
⑥
Scolex attaches to intestine
①
Adult in small intestine
⑤
Protoscolex from cyst
Ingestion of cysts (in organs)
Definitive Host (dogs & other canidae)
i ②
Embryonated egg in feces
Intermediate Host (sheep, goats, swine, etc.)
Ingestion of eggs (in feces)
d ④
③
Oncosphere hatches; penetrates intestinal wall
Hydatid cyst in liver, lungs, etc.
④
②
④
④
④
④
④
③
i = Infective Stage
d = Diagnostic Stage

Illustration 8-1. Life Cycle of *Echinococcus granulosus*. The adult *Echinococcus granulosus* (3 to 6 mm long) ①resides in the small bowel of the definitive hosts, dogs or other canids. Gravid proglottids release eggs ②that are passed in the feces. After ingestion by a suitable intermediate host (under natural conditions: sheep, goat, swine, cattle, horses, camel), the egg hatches in the small bowel and releases an oncosphere ③that penetrates the intestinal wall and migrates through the circulatory system into various organs, especially the liver and lungs. In these organs, the oncosphere develops into a cyst ④ that enlarges gradually, producing protoscolices and daughter cysts that fill the cyst interior. The definitive host becomes infected by ingesting the cyst-containing organs of the infected intermediate host. After ingestion, the protoscolices ⑤evaginate, attach to the intestinal mucosa ⑥, and develop into adult stages ①in 32 to 80 days. The same life cycle occurs with *E. multilocularis* (1.2 to 3.7 mm), with the following differences: the definitive hosts are foxes, and to a lesser extent dogs, cats, coyotes and wolves; the intermediate host are small rodents; and larval growth (in the liver) remains indefinitely in the proliferative stage, resulting in invasion of the surrounding tissues. With *E. vogeli* (up to 5.6 mm long), the definitive hosts are bush dogs and dogs; the intermediate hosts are rodents; and the larval stage (in the liver, lungs and other organs) develops both externally and internally, resulting in multiple vesicles. *E. oligarthrus* (up to 2.9 mm long) has a life cycle that involves wild felids as definitive hosts and rodents as intermediate hosts. Humans become infected by ingesting eggs ②, with resulting release of oncospheres ③in the intestine and the development of cysts ④, ④, ④, ④, ④, ④ in various organs. (SOURCE: CDC/DPDx Echinococcus)

Morphology

The adult worm measures approximately 3–8.5µm long. The scolex has four suckers and a rostellum with hooks, the latter becoming tightly inserted into the crypts of Lieberkühn. The mature strobila has only 3–4 proglottids, one is immature, one is mature and the final one is gravid; when gravid the eggs are expelled in the feces.

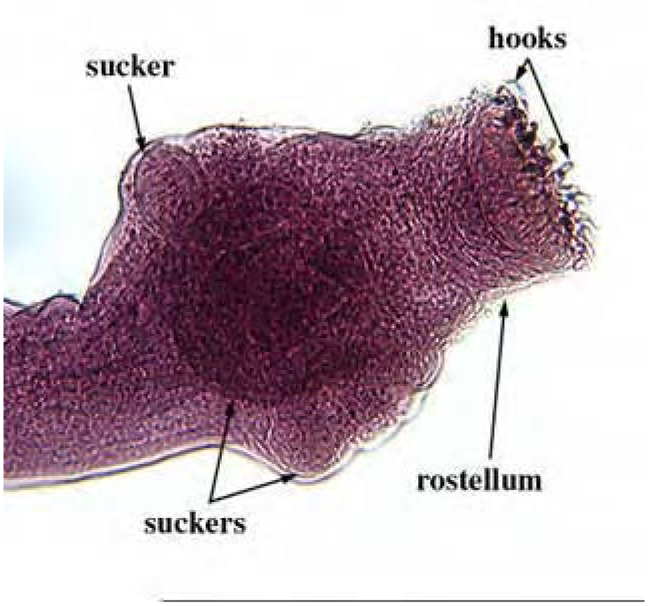
hooks
sucker
rostellum
suckers

Image 8-1. *Echinococcus granulosus* scolex. (SOURCE: CDC)

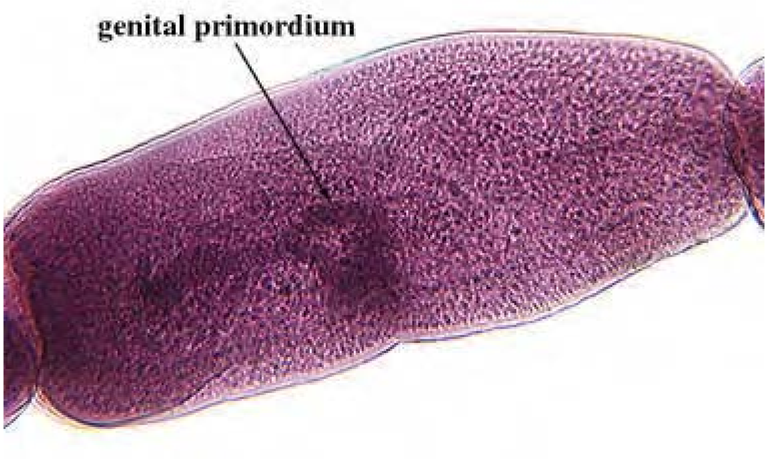
genital primordium

Image 8-2. *Echinococcus granulosus* immature proglottid. (SOURCE: CDC)

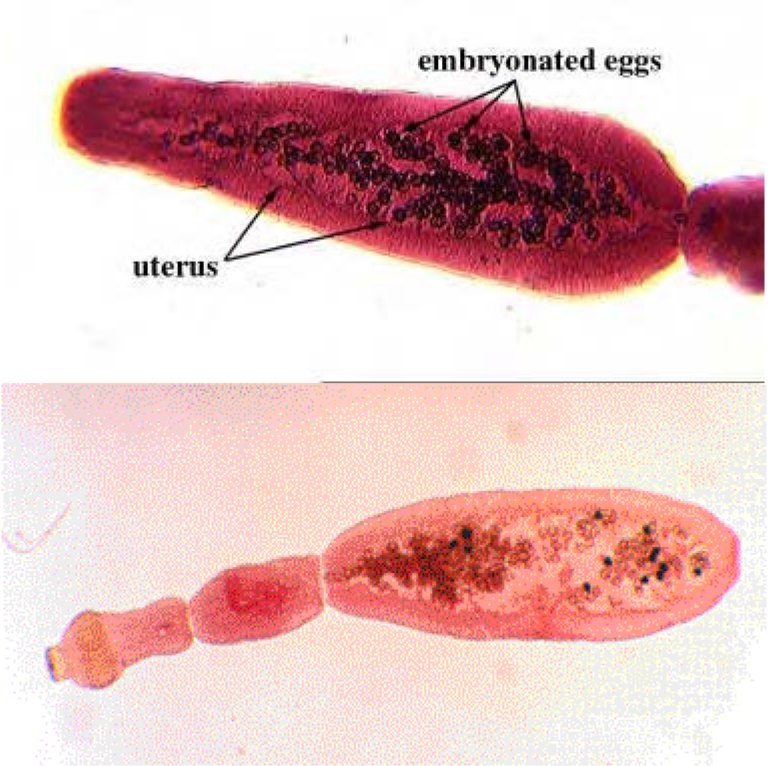

Image 8-3. Adult female worms of *Echinococcus granulosus* with gravid proglottids. (SOURCE: CDC)

Due to the close similarity of the eggs to other *Taenia* species found in dogs they were until recently thought to be morphologically indistinguishable.

The larvae in man develop into a unilocular cyst which gives rise to unilocular hydatid disease. This is characterized as having only one bladder or many completely isolated bladders, each enclosed in its own well-developed envelope. The latter consists of several layers, the most prominent being the laminated layer. Within this again is the germinal membrane from which the brood capsules arise inside which develop thousands of larvae or protoscoleces, the whole being suspended in a hydatid fluid.

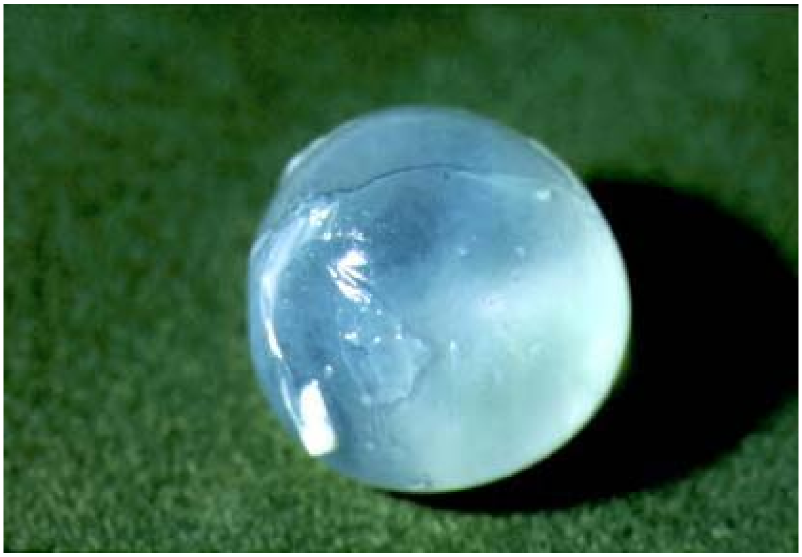

Image 8-4. Unilocular cyst of *Echinococcus granulosus*. This is characterized as having only one bladder or many completely isolated bladders, each enclosed in its own well-developed envelope. Inside which develop thousands of larvae or protoscoleces, the whole being suspended in a hydatid fluid. These cysts in man give rise to unilocular hydatid disease.

Clinical Disease

Hydatid disease in humans is potentially dangerous depending on the location of the cyst. Some cysts may remain undetected for many years until they become large enough to affect other organs. Symptoms are then of a space occupying lesion. Lung cysts are usually asymptomatic until there is a cough, shortness of breath or chest pain. Hepatic cysts result in pressure on the major bile ducts or blood vessels. Expanding hydatid cysts cause necrosis of the surrounding tissue.

Slow leakage of the hydatid fluid results in eosinophilia and rupture of an abdominal hydatid cyst results in severe allergic symptoms.

Symptoms may not manifest themselves for five to 20 years after the infection.

Laboratory Diagnosis

1. Imaging and serodiagnosis are the mainstay of diagnosis. Serological tests include Enzyme linked immunosorbent assay (ELISA), an indirect hemagglutination test a complement fixation test and a Western Blot system.

2. Microscopic examination of the cyst fluid to look for the characteristic protoscoleces which can be either invaginated or evaginated. The cyst fluid will also reveal free hooklets and tissue debris. One percent eosin may be added to the fluid to determine the viability of the protoscoleces. Viable protoscoleces exclude eosin whereas nonviable protoscoleces take up the eosin.

3. Histological examination of the cyst wall after surgical removal.

Western Blots

One serological test which has proved to be of value to diagnosing Hydatid disease is the Western Blot. The test presents a definitive means for detection of human antibodies to the cestode *E. granulosus.*

Diagnosis can be achieved using the Western Blot assay for the detection of IgG antibodies in serum reactive with *E. granulosus* antigens present on a membrane. Field studies support a sensitivity of 80% and specificity of 100% in patients with hepatic cysts.

This assay is known as the Qualicode™ Hydatid Disease Kit; the principle behind the test is that it is a qualitative membrane-based immunoassay manufactured from *E. granulosus* proteins. The *E. granulosus* proteins are fractionated according to molecular weight by electrophoresis on a ployacrylamide slab gel (PAGE) in the presence of sodium dodecyl sulfate (SDS). The separated *E. granulosus* proteins are then transferred via electrophoretic blotting from the gel into strips for testing of individual samples.

During the procedure, the strips containing the *E. granulosus* proteins are incubated with serum specimens and washed to remove unbound antibodies.

Visualization of human immunoglobulins specifically bound to *E. granulosus* proteins is performed by sequential reaction with goat anti-human immunoglobulin-alkaline phosphatase conjugate and BCIP/NBT substrate. Bands corresponding to the positions of the resoled *E. granulosus* proteins will be visualized on the strip, indications the presence in the serum sample of IgG antibodies direct against *E. granulosus* antigens. Band positions are compared to those on a reference strip developed using the Hydatid disease positive control.

Prevention

- Safe disposal of dog feces.
- Education to prevent feeding uncooked offal to dogs.

Echinococcus multilocularis

Introduction

The larvae of *Echinococcus multilocularis* is a particularly dangerous species causing multilocular (alveolar) hydatid disease in man and animals and is common in the highlands of Europe i.e. Switzerland and Germany, in Canada, Alaska and Northern Russia. The most common definitive hosts are foxes and wolves in addition to domestic cats and dogs when they have access to infected rodents.

Life Cycle

Foxes (1) are the primary definitive hosts although in domestic circumstances dogs (1a) and cats (1b) can act as the definitive host. Rodents (4) are the intermediate hosts. Man (4a) is an accidental host by the ingestion of eggs (2) where multilocular cysts are formed (3). In these cysts, the limiting membrane is thin and the germinal epithelium may bud off externally resulting in proliferation in any direction (5). Metastases may occur. Unlike *E. granulosus*, there is little fluid in the cysts of *E. multilocularis.*

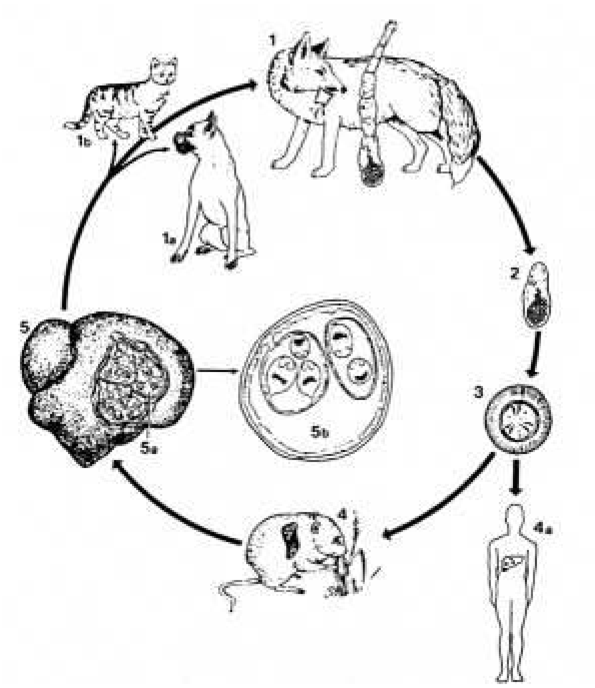

Illustration 8-3: Life cycle of *Echinococcus multilocularis*. (SOURCE: Unknown)

Morphology

Larvae - The larval *E. multilocularis* is very different from that of *E. granulosus*. In this case the 'cyst' grows invasively by external budding, forming a diffuse growth through the infected organ, replacing that organ's tissues. The growth itself, (it cannot truly be called a cyst as there is no real cyst wall), is composed of numerous cavities containing a gelatinous matrix within which protoscolices and numerous brood capsules are produced, and which in its behavior, most closely resembles a malignant neoplasm. In contrast to *E. granulosus* this growth is also very rapid. Infective protosocialises being present after only wo to three months, as compared to the one to two years in the related metacestode.

Adults - The adult parasite is very similar to *E. granulosus*, being slightly smaller, with a maximum length of approximately 4µm, and consisting of four to five proglottids.

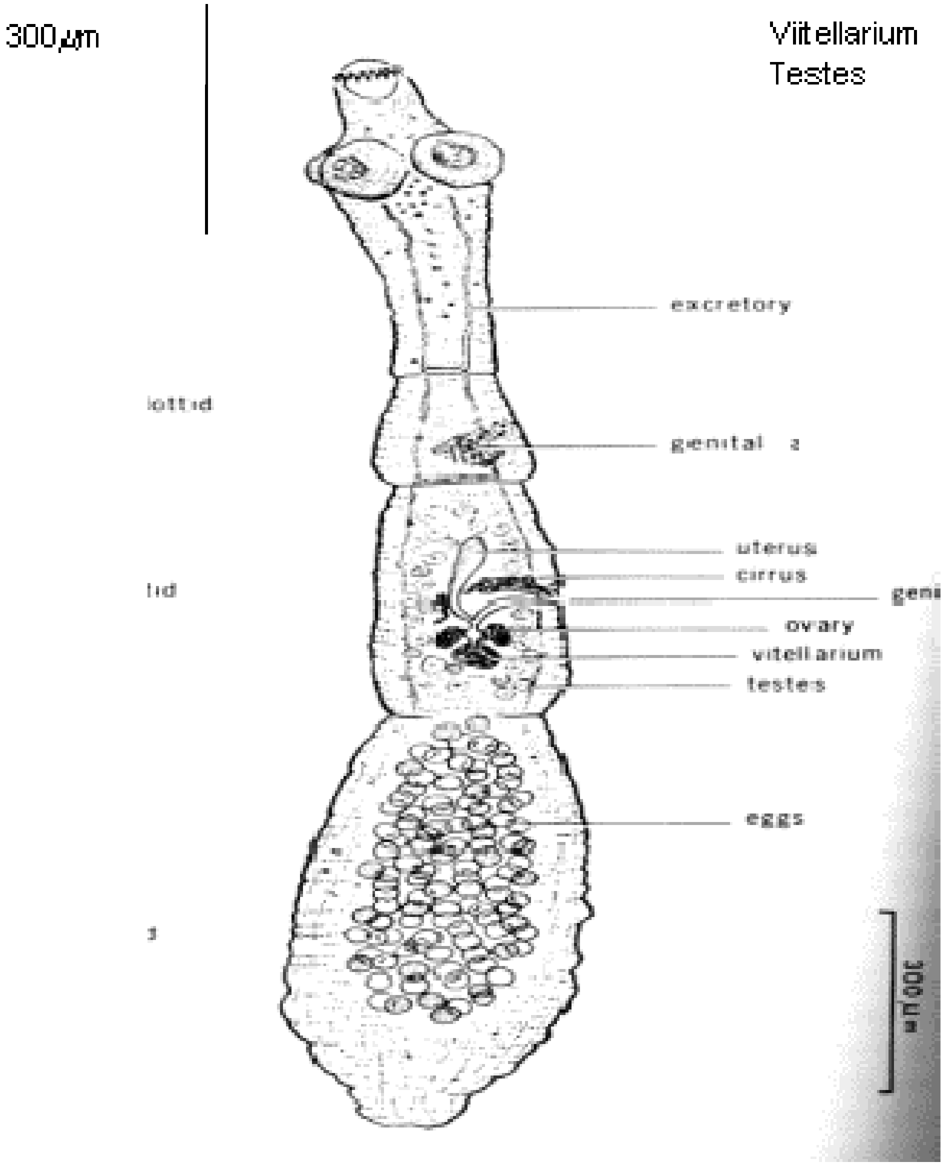
300µm
Viitellarium
Testes
excretory
iottid
genital 2
lid
uterus
cirrus
geni
ovary
vitellarium
testes
eggs
300 µm

Illustration 8-4. Diagrammatic representation of the adult worm if *Echinococcus multilocularis.* (SOURCE: Unknown)

Clinical Disease

Cysts form primarily in the liver and growth in the vena cava or portal vein results in metastases in the lung or brain. Clinical disease is similar to that of *E. granulosus*.

Laboratory Diagnosis

1. Laboratory diagnosis is can be made by ELISA.
2. Clinical diagnosis is made by ultrasound.

Echinococcus granulosus	*Echinococcus multilocularis*
Slow development of cyst	Rapid development of cyst
Cysts have thick-walled chambers	Cyst has thin-walled chambers
Separated by connective tissue	Not separated by connective tissue
Cyst is fluid filled	Cyst is gelatinous filled
Cyst is free of host material	Cyst is contaminated by host material

Table 8-1. Differences between the hydatid cysts of *E. granulosus* and *E. multilocularis*. (SOURCE: White)

Multiceps multiceps

Introduction

Multiceps multiceps, is a cestode of cosmopolitan distribution and causes coenuriasis in man. The dog is the common definitive host in the USA and UK, with the intermediate stages developing in many ungulates, especially sheep.

Life Cycle

The adult worm is found in dogs and other canidae. The intermediate hosts comprise of a number of herbivorous animals, where the cysts develop in the brain and spinal cord causing a disease called ‘staggers’, which affects the balancing powers of the animals. Man becomes infected by the accidental ingestion of eggs. The oncosphere hatches and penetrates the intestinal wall and the embryo is carried by the blood stream to various parts of the body including the central nervous system where it lodges and the cyst or coenurus develops. Multiple scolices burst from the inner surface of the cyst wall. The cyst resembles that of a cysticercus (*Cysticercus cerebralis)*. It is filled with fluid, semi-transparent and glistening white. The cysticercus possesses unusual asexual multiplication, forming a bladder (or coenurus) which gives rise to hundreds of daughter protoscoleces directly from its inner wall.

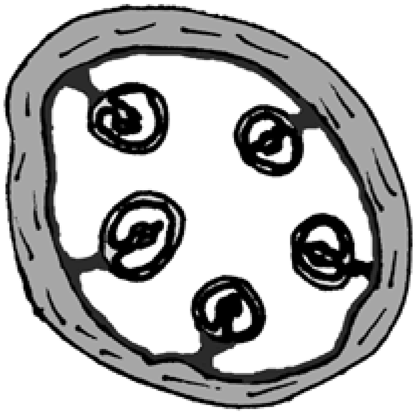

Illustration 8-5. Diagrammatic representation of a *Multiceps (Taenia) multiceps* coenurus cyst. (SOURCE: Unknown)

Clinical Disease

In humans, coenuri are most frequently found in the brain and spinal cord but also in the subcutaneous tissue. Symptoms include headache, vomiting, paraplegia seizures and eye problems. The coenurus may cause serious damage or even death, but only a few have ever been reported.

Laboratory Diagnosis

There are no serological tests available. Diagnosis is achieved by the finding of the coenurus cyst in the brain and spinal cord, or in the subcutaneous tissue.

8.2 Larval Nematodes which Infect Man

Intestinal capillariasis caused by *Capillaria Philippinensis* appeared first in the Philippines and subsequently in Thailand, Japan, Iran, Egypt, and Taiwan, but most infections occur in the Philippines and Thailand. As established experimentally, the life cycle involves freshwater fish as intermediate hosts and fish-eating birds as definitive hosts. Embryonated eggs from feces fed to fish hatch and grow as larvae in the fish intestines. Infective larvae fed to monkeys, Mongolian gerbils, and fish-eating birds develop into adults. Larvae become adults in 10 to 11 days, and the first-generation females produce larvae. These larvae develop into males and egg-producing female worms. Eggs pass with the feces, reach water, embryonate, and infect fish. Autoinfection is part of the life cycle and leads to hyperinfection. Humans acquire the infection by eating small freshwater fish raw. The parasite multiplies, and symptoms of diarrhea, borborygmus, abdominal pain, and edema develop. Chronic infections lead to malabsorption and hence to protein and electrolyte loss, and death results from irreversible effects of the infection. Treatment consists of electrolyte replacement and administration of an antidiarrheal agent and mebendazole or albendazole.

Capillariasis Philippinensis is considered a zoonotic disease of migratory fish-eating birds. The eggs are disseminated along flyways and infect the fish, and when fish are eaten raw, the disease develops.

Capillaria Philippinensis

Introduction

Intestinal capillariasis was first observed in the Philippines in 1962 and since then it has been noted, although less frequently in Thailand, with scattered reports from Taiwan, Japan, Egypt, and Iran. *Capillaria Philippinensis* is a small nematode which burrows into the intestinal mucosa of its human host.

Humans and birds become infected when they eat uncooked or poorly cooked fish. Therefore, infections are seen in endemic proportions where uncooked fish is considered a delicacy.

Life Cycle

The complete life cycle is not known however, human infection is initiated by the consumption of raw fish. The infective larvae are found in the intestine of the fish. When infective fish are eaten by humans, the larvae mature and the adult worms live in the intestinal mucosa, mainly in the jejunum, where they are usually present in large numbers. Larval stages, oviparous and larviparous females are also found in the host's intestine, which suggests that the nematode multiplies in the intestine resulting in auto-infection as in *Strongyloides* infection. The eggs are passed out in the feces and embryonate in the soil, a process which takes about 12 days. The eggs are ingested by fresh-water fish, they hatch and develop into the infective form in the intestine of the fish.

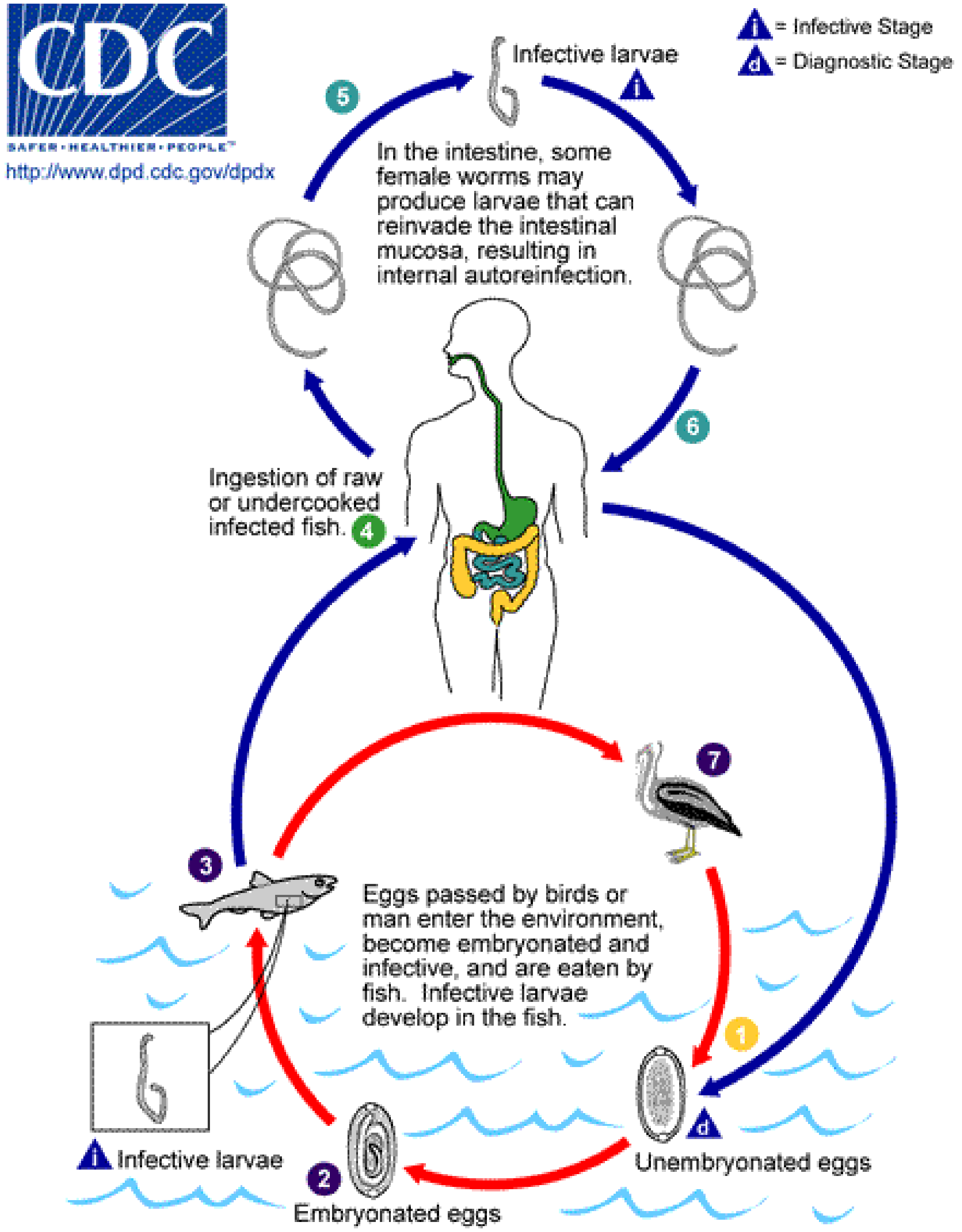
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
i = Infective Stage
d = Diagnostic Stage
Infective larvae
i
5
In the intestine, some female worms may produce larvae that can reinvade the intestinal mucosa, resulting in internal autoreinfection.
6
Ingestion of raw or undercooked infected fish.
4
7
3
Eggs passed by birds or man enter the environment, become embryonated and infective, and are eaten by fish. Infective larvae develop in the fish.
1
d
Unembryonated eggs
i
Infective larvae
2
Embryonated eggs

Illustration 8-6. Diagram of the life cycle of *Capillaria Philippinensis.* Humans become infected from eating infected fish. The disease in man is known as Capillariasis. Typically, unembryonated eggs are passed in the human stool ❶and become embryonated in the external environment ❷; after ingestion by freshwater fish, larvae hatch, penetrate the intestine, and migrate to the tissues ❸. Ingestion of raw or undercooked fish results in infection of the human host ❹. The adults of *Capillaria Philippinensis* (males: 2.3 to 3.2 mm; females: 2.5 to 4.3 mm) reside in the human small intestine, where they burrow in the mucosa ⑤. The females deposit unembryonated eggs. Some of these become embryonated in the intestine, and release larvae that can cause autoinfection. This leads to hyperinfection (a massive number of adult worms) ⑥. *Capillaria Philippinensis* is currently considered a parasite of fish eating birds, which seem to be the natural definitive host ❼. (CDC 3382 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The adult female worm measures 2.5-4.3µm, while the males are marginally smaller measuring 2.3-3.2µm. Adult females can produce immature and mature ova as well as free larvae that can auto-infect.

The eggs measure 45µm x 21µm, and resemble those of *Trichuris* but have less prominent polar plugs. They have a thick striated shell.

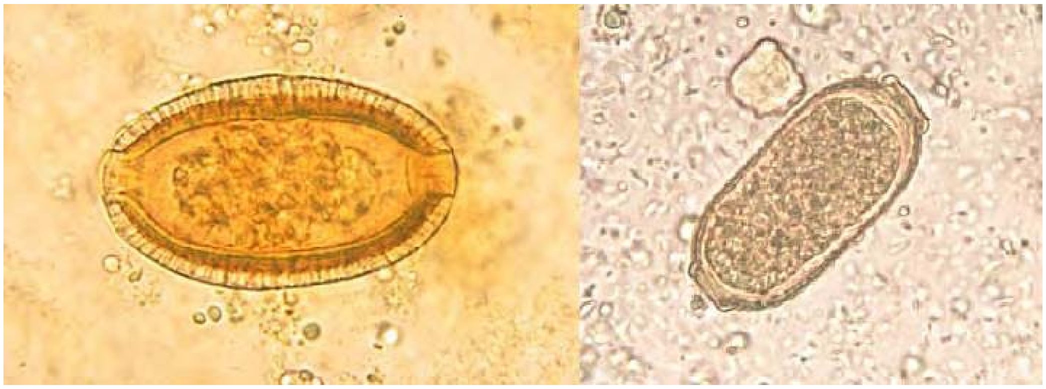

Figure 8-5. Eggs of *Capillaria Philippinesis.* They are oval in shape and closely resemble the eggs of *Trichuris* but have less prominent plugs. (SOURCE: PHIL 1376/653 - CDC/Dr. Mae Melvin)

Clinical Disease

Many cases may be asymptomatic, but any manifested symptoms will be related to the worm burden. The most common symptoms are watery stools with large fluid loss, anorexia, nausea, vomiting, and hypotension. Abdominal distension and edema may also develop. Death can ensue from pneumonia, cerebral edema, hypokalemia and heart failure. The large worm burden can cause electrolyte imbalance, plasma protein imbalance, and fluid loss, proving fatal within 4–6 months if untreated. Villi blunting leads to malabsorption but eosinophilia is not a feature of this disease.

Laboratory Diagnosis

Diagnosis depends on finding the characteristic eggs or larvae in the feces. Pulmonary capillariasis diagnosis relies on finding adult worms or eggs in lung biopsies.

Chapter 9.
Malaria

9.1 Blood Parasites

Red Blood cells offer parasites an excellent environment for invasion and survival. Haemosporina are the only protozoan parasites which can invade the red blood corpuscles of vertebrates. Most, if not all, have multiplicative phases in the reticulo-endothelial system.

The red blood cells are thin-walled and constantly moving, so absorption of food materials and elimination of waste products of metabolism are relatively easy to achieve. They also contain rich supplies of protein and oxygen.

Malarial parasites do not actually penetrate the red blood cell, but enter the cell membrane by endocytosis and enclose in a parasitophorous membrane.

Introduction

Malaria is the most important tropical disease known to man. It remains a significant problem in many tropical areas, especially in sub-Saharan Africa. Malaria is spreading as a result of environmental changes, including global warming, civil disturbances, increasing travel and drug resistance (Greenwood, B.M, 1997). There are approximately 100 million cases of malaria worldwide with about 1 million of these proving fatal.

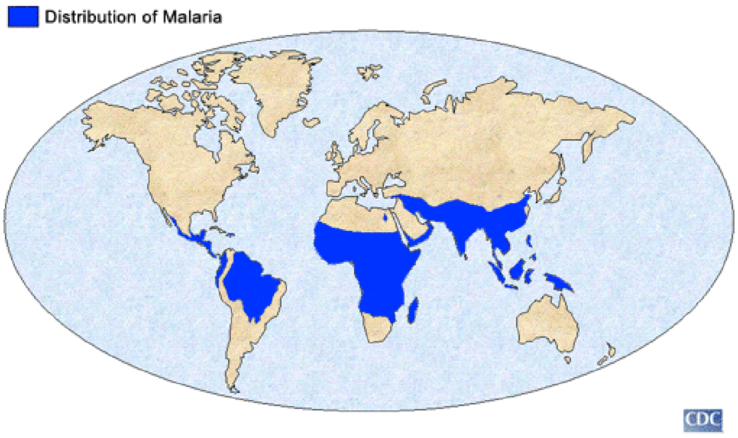

Illustration 9-1. Map illustrating the enormous distribution of malaria throughout the world. (SOURCE: CDC)

Malaria is caused by protozoa of the *Plasmodium* species. There are four species which infect both humans and animals; *Plasmodium malariae* (quartian malaria), *Plasmodium vivax* (benign tertian malaria)*, Plasmodium falciparum* (malignant tertian malaria, subtertian malaria) and *Plasmodium ovale* (ovale tertian malaria).

The transmission of the protozoa, *Plasmodium* requires two hosts, an intermediate invertebrate host (vector), and a definitive vertebrate host (mammals, birds and lizards).

All *Plasmodium* species undergo the general haemosporina developmental cycle which can be summarized as:

- initial or continual schizogony (*reproduction by multiple asexual fission*) in the vertebrate host with initiation of gametogony (*the formation or production of gametes*);
- formation of gametes in the arthropod host and subsequent fertilization and formation of a zygote;
- formation of sporozoites from the zygote by repeated nuclear division followed by cytoplasmic divisions. (Smyth, J.D, 1994)

There is no requirement for resistant stages since the transfer of the parasites between the vertebrate and invertebrate hosts is made by withdrawal or injection during the bloodsucking act, there is little or no exposure to the hazards of the outside world; thus by blood transfusion or inoculation, via the blood stages of the parasite.

Life Cycle

Malaria is transmitted by the female anopheline mosquito. The life cycle of all species of human malaria parasites is essentially the same. It comprises an exogenous sexual phase (sporogony*)* with multiplication in certain Anopheles mosquitoes and an endogenous asexual phase (schizogony) with multiplication in the vertebrate host. The latter phase includes the development cycle in the red cells (erythrocytic schizogony*)* and the phase taking place in the parenchyma cells in the liver (pre-erythrocytic schizogony).

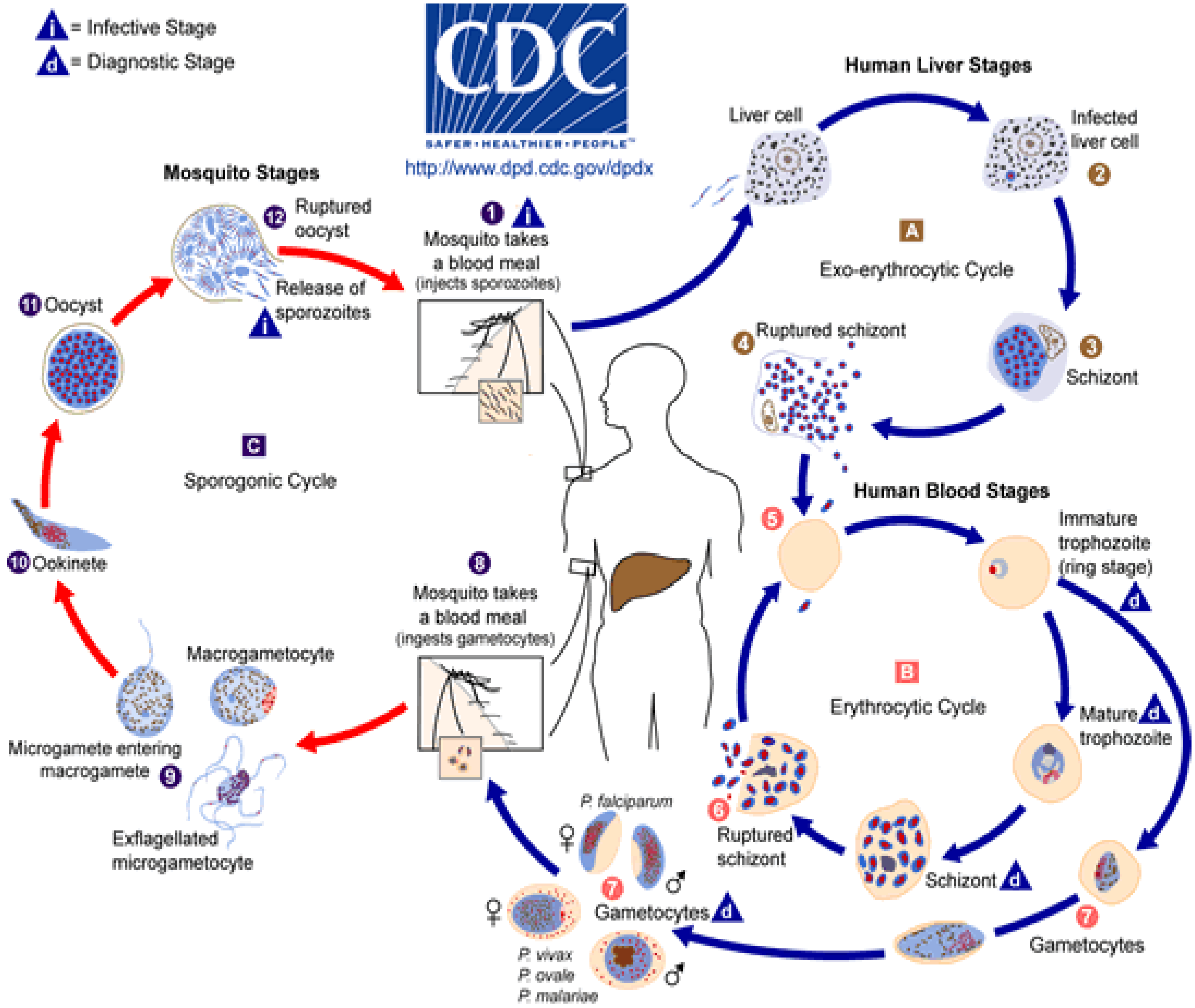
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
Human Liver Stages
Liver cell
Infected liver cell
②
A
Exo-erythrocytic Cycle
④ Ruptured schizont
③ Schizont
Mosquito Stages
⑫ Ruptured oocyst
Release of sporozoites
i
① i Mosquito takes a blood meal (injects sporozoites)
⑪ Oocyst
C
Sporogonic Cycle
⑩ Ookinete
Human Blood Stages
⑤
Immature trophozoite (ring stage)
d
B
Erythrocytic Cycle
Mature trophozoite
d
⑧ Mosquito takes a blood meal (ingests gametocytes)
Macrogametocyte
Microgamete entering macrogamete ⑨
Exflagellated microgametocyte
P. falciparum
♀
♂
⑦ Gametocytes d
⑥ Ruptured schizont
Schizont d
♀
P. vivax
P. ovale
P. malariae
♂
⑦ Gametocytes

Illustration 9-3. The malaria parasite life cycle. The malaria parasite life cycle involves two hosts. During a blood meal, a malaria-infected female *Anopheles* mosquito inoculates sporozoites into the human host ①. Sporozoites infect liver cells ②and mature into schizonts ③, which rupture and release merozoites ④. (Of note, in *P. vivax* and *P. ovale* a dormant stage [hypnozoites] can persist in the liver and cause relapses by invading the bloodstream weeks, or even years later.) After this initial replication in the liver (exo-erythrocytic schizogony A), the parasites undergo asexual multiplication in the erythrocytes (erythrocytic schizogony B). Merozoites infect red blood cells ⑤. The ring stage trophozoites mature into schizonts, which rupture releasing merozoites ⑥. Some parasites differentiate into sexual erythrocytic stages (gametocytes) ⑦. Blood stage parasites are responsible for the clinical manifestations of the disease.

The gametocytes, male (microgametocytes) and female (macrogametocytes), are ingested by an *Anopheles* mosquito during a blood meal ⑧. The parasites' multiplication in the mosquito is known as the sporogonic cycle C. While in the mosquito's stomach, the microgametes penetrate the macrogametes generating zygotes ⑨. The zygotes in turn become motile and elongated (ookinetes) ⑩which invade the midgut wall of the mosquito where they develop into oocysts ⑪. The oocysts grow, rupture, and release sporozoites ⑫, which make their way to the mosquito's salivary glands. Inoculation of the sporozoites into a new human host perpetuates the malaria life cycle ①. (SOURCE: PHIL 3405- CDC/Alexander J. da Silva, PhD/Melanie Moser)

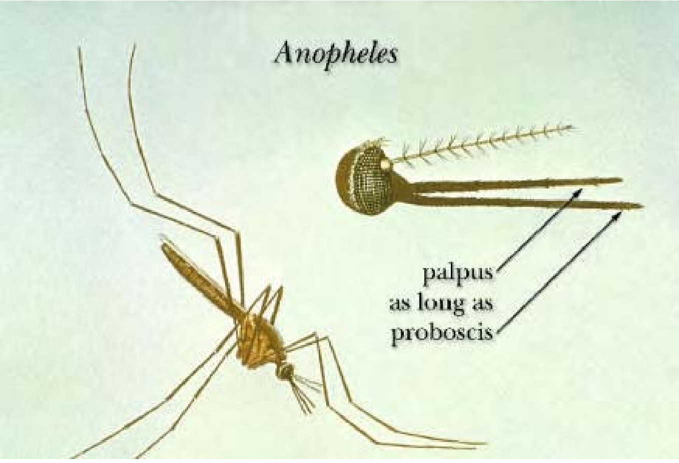

Illustration 9-2. Distinguishing characteristics of the Anopheles mosquito.

Image 9-1. An Anopheline mosquito, the vector of the protozoa group *Plasmodia*, the parasite known to cause malaria in both man and non-human primates. Malaria is transmitted by female *Anopheles* mosquitoes to the definitive host while the mosquito blood-feeds on its victims. (SOURCE: PHIL 2070/6765 - CDC/ James Gathany)

When a female Anopheles mosquito bites an infected person, it ingests blood which may contain the mature sexual cells (male and female gametocytes) which undergo a series of developmental stages in the stomach of the mosquito. Exflagellation (*the extrusion of rapidly waving flagellum-like microgametes from microgametocytes*) occurs resulting in the production in a number of male and female gametes. Fertilization occurs producing a zygote which matures to an ookinete. This penetrates the stomach wall of the mosquito where it grows into an oocyst and it further matures to become a motile sporozoite.

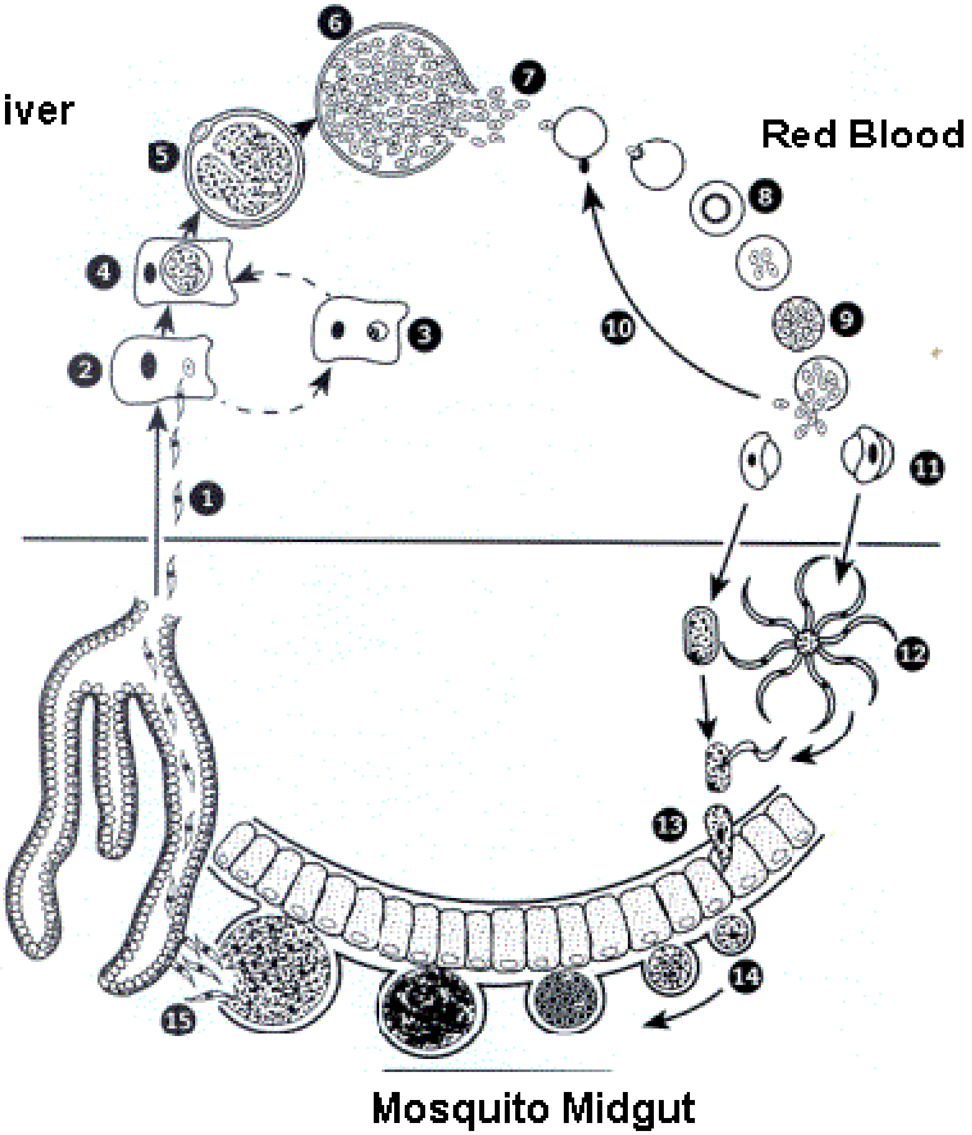

Illustration 9-4. Diagram of the malaria life cycle. **1)** Sporozoites, injected through the skin by female anopheline mosquito; **2)** sporozoites infect hepatocytes; **3)** some sporozoites develop into hypnozoites (*P. vivax* and *P. ovale*): **4)** liver stage parasite develops; **5 – 6)** tissue schizogony; **7)** merozoites are released into the circulation; **8)** ring stage trophozoites in red cells; **9)** erythrocytic schizogony; **10)** merozoites invade other red cells; **11)** some parasites develop into female (macro-) or male (micro-) gametocytes, taken up by the mosquito; **12)** mature macrogametocyte and exflagellating microgametocytes; **13)** ookinete penetrates gut wall; **14)** development of oocyst; **15)** sporozoites penetrate salivary glands. (SOURCE: Unknown)

The length of the developmental stage in the mosquito not only depends on the *Plasmodium* species but also the mosquito host and the ambient temperature. This may range from eight days in *Plasmodium vivax* to as long as 30 days in *Plasmodium malariae*.

The sporozoites migrate from the body cavity of the mosquito to the salivary glands and the mosquito now becomes infective. Sporozoites enter into the blood stream of a host when the mosquito feeds on blood. Following the inoculation, the sporozoites leave the blood within 40 minutes and enter the parenchymal cells of the liver (hepatocytes). In all four species, asexual development occurs in the liver cells, a process known as pre-erythrocytic schizogony**,** to produce thousands of tiny merozoites which are released into the circulation after about 16 days. However in *P. vivax* and *P. ovale* some sporozoites differentiate into hypnozoites which remain dormant in hepatocytes for considerable periods of time. When they are “reactivated” they undergo asexual division and produce a clinical relapse.

In *P. falciparum* and *P. malariae* hypnozoites are not formed and the parasite develops directly into pre-erythrocytic schizonts.

Once in the circulation, the merozoites invade the red cells and develop into trophozoites. In the course of their development they absorb the hemoglobin of the red cells and leave as the product of digestion a pigment called hemozoin*,* a combination of hematin and protein. This iron-containing pigment is seen in the body of the parasite in the form of dark granules, which are more obvious in the later stages of development.

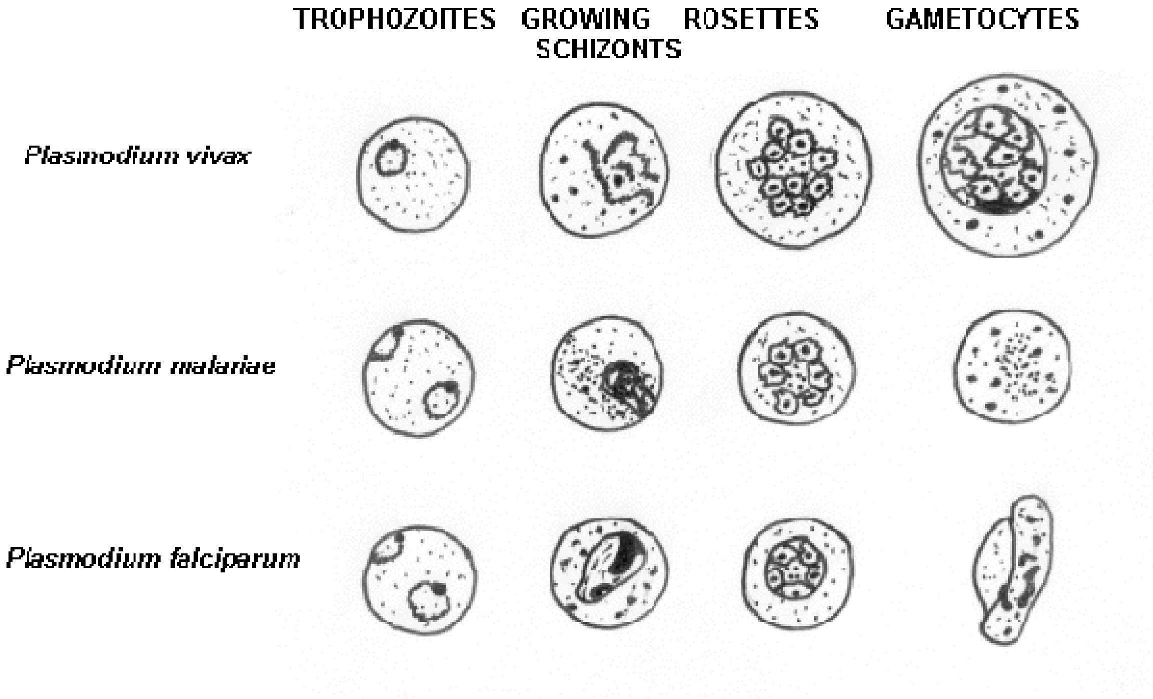

Illustration 9-5. Diagram illustrating the various stages of the three common species of malaria which infect man. (SOURCE: Unknown)

After a period of growth the trophozoite undergoes an asexual division, erythrocytic schizogony. When the mature trophozoite starts to divide in the red blood cell, separate merozoites are formed resulting in a schizont. When fully developed, the schizont ruptures the red blood cell containing it, liberating the merozoites into the circulation. These merozoites will then infect new red cells and the process of asexual reproduction in the blood tends to proceed. Some of the merozoites entering red blood cells do not form trophozoites then schizonts but develop into gametocytes and this process takes place in deep tissue capillaries. This erythrocytic cycle of schizogony is repeated over and over again in the course of infection, leading to a progressive increase of parasitemia.

Infections with all four strains of malaria have many clinical features in common. These are related to the liberation of fever-producing substances, especially during schizogony. The common features are:

Fever: Often irregular. The regular pattern of fever does not occur until the illness has continued for a week or more; where it depends on synchronized schizogony.

Anemia: The anemia is hemolytic in type. It is more severe in infections with *P. falciparum* because in this infection cells of all ages can be invaded. Also, the parasitemia in this infection can be much higher than in other malarias.

Splenomegaly: The spleen enlarges early in the acute attack of malaria. When a patient has been subjected to many attacks, the spleen may be of an enormous size and lead to secondary hypersplenism.

Jaundice: A mild jaundice due to hemolysis may occur in malaria. Severe jaundice only occurs in *P. falciparum* infection, and is due to liver involvement.

9.2. Species Specific Characteristics

Plasmodium falciparum

Introduction

Plasmodium falciparum exists in the tropics and sub-tropics, and is responsible for approximately 50% of all malaria cases. The incubation period of *P. falciparum* malaria is the shortest, between eight and 11 days and has a periodicity of 36–48 hours. It can be differentiated from the other species by the morphology of the different stages found in the peripheral blood. In infections with *Plasmodium falciparum* usually only young trophozoites and gametocytes are seen in peripheral blood smears, the schizonts are usually found in capillaries sinuses of internal organs and in the bone marrow. The disease it produces runs an acute course and often terminating fatally. It is a significant cause of abortions and stillborns and even death of non-immune pregnant women.

Life Cycle

The aspects of the life cycle which are specific to *P. falciparum* are as follows:

- It attacks all ages of erythrocytes so that a high density of parasites can be reached quickly. In extreme cases up to 48% of the red blood cells can be parasitized.
- Multiple infections resulting in several ring forms in a corpuscle are not uncommon.
- The latter stages in the asexual cycle do not occur in the peripheral blood as in other forms of malaria, so that only rings and crescents are found in blood films. After 24 hours the ring forms and older trophozoites show a tendency to clump together and adhere to the visceral capillary walls and become caught up in the vessels of the

heart, intestine, brain or bone marrow in which the later sexual stages are completed.

- Sporulation is not as well synchronized as in other malaria forms so that the fever may last longer.
- Exo-erythrocytic forms do not persist in the tissues and hence relapses do not occur.

Morphology of Trophozoites

Red blood cells in *Plasmodium falciparum* infections are not enlarged and they may have a spiky outline which is common in cells which have dried slowly. The typical arrangement cytoplasm in young trophozoites is the well-known ring formation which thickens and invariably contains several vacuoles as the trophozoite develops. Chromatin is characteristically found as a single bead, but double beads and small curved rod forms frequently occur.

Maurer's dots are slow to appear and are first seen as minute purplish dots, 6 or less in number. The points become spots, still few in number and are now characteristic enough to be recognized. Maurer describes them as fine ringlets, loops or streaks. They are seldom absent from the red blood cells containing large rings but the staining of the spots is very sensitive to pH and are seldom seen if the pH falls below 6.8.

Trophozoites of *P. falciparum* can be found on the edge of the red blood cells. These are known as accole forms and are found as three distinct types:

1. **Common:** The single chromatin bead lies on the edge of the cell with most of the cytoplasm extended along the edge on both sides of the bead.

2. **Rim:** The complete parasite lies in a thickened line along the edge of the cell with no evidence of ring formation.

3. **Displaced:** The parasites are displaced beyond the edge of the host cell. All degrees of displacement may occur, from partial to marked displacement with most of the parasite lying beyond the cell margin.

Pigment is not a characteristic finding in the early stages of *P. falciparum* infections.

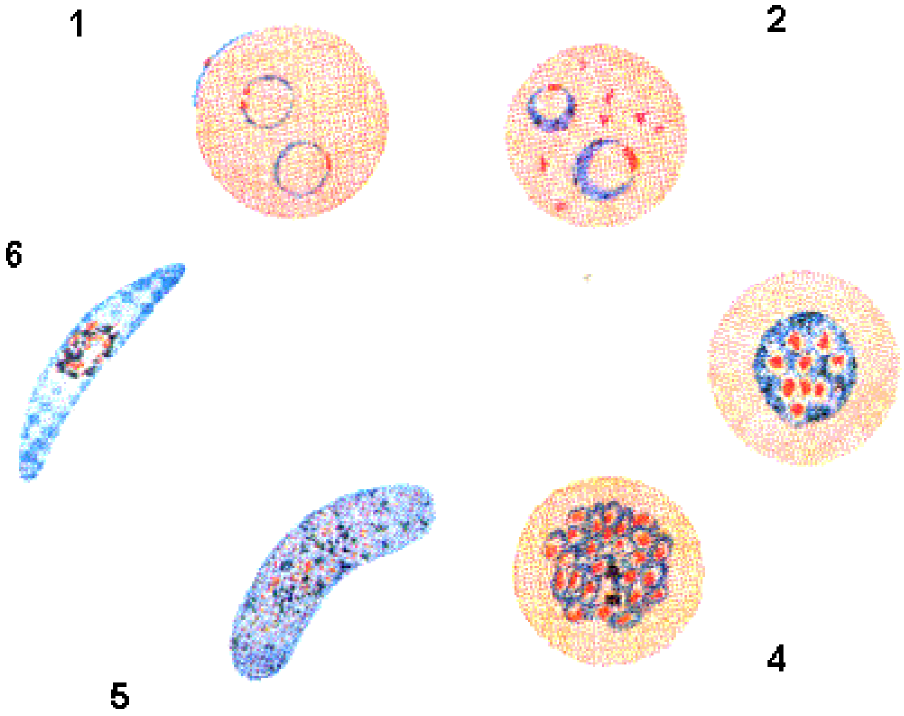

Illustration 9-6. Diagrammatic illustration of the morphology of the different stages of the *Plasmodium falciparum* life cycle in thin blood films. **1)** *P. falciparum* early trophozoites / ring forms. **2)** Developing trophozoites (rarely seen in peripheral blood). **3)** Immature schizonts (rarely seen in peripheral blood). **4)** Mature schizonts, almost fill the red blood cell. **5)** Microgametocytes, large numbers appear after 7–12 days. **6)** Macrogametocytes, large numbers appear after 7-12 days.

Morphology of Gametocytes

Gametocytes are the sexual stage of the malaria parasite. *Plasmodium falciparum* gametocytes appear in the peripheral circulation after 7-12 days of patent parasitemia and by then, they have assumed their typical crescent shapes. They soon reach their peak density, and then decline in numbers, disappearing in about three months as a rule.

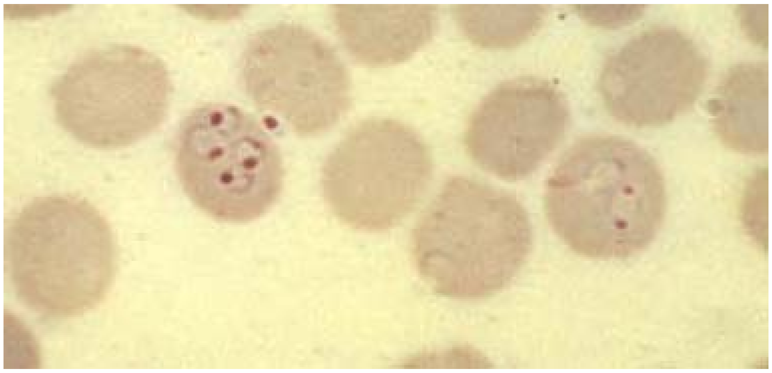

Image 9-2. Young trophozoite / ring stage of *Plasmodium falciparum*. The ring thickens and invariably contains several vacuoles as the trophozoite develops. Maurer's dots are slow to appear and are first seen as minute purplish dots. (Giemsa stain). (SOURCE: PHIL 5946 - CDC/ Steven Glenn, Laboratory & Consultation Division)

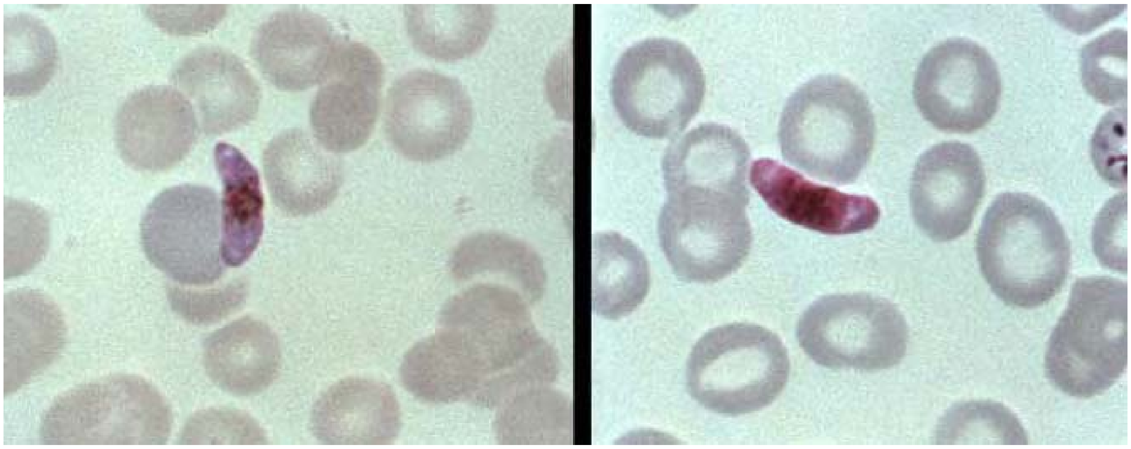

Image 9-3. *Plasmodium falciparum* gametocytes appear in the peripheral circulation after 7-12 days of patent parasitemia and by then, they have assumed their typical crescent shapes. (Giemsa stain) (SOURCE: PHIL 5941 - CDC/ Steven Glenn, Laboratory & Consultation Division)

The female form, or macrogametocyte, is usually more slender and somewhat longer than the male, and the cytoplasm takes up a deeper blue color with Giemsa stain. The nucleus is small and compact, staining dark red, while the pigment granules are closely aggregated around it. The male form, or microgametocyte, is broader than the female and is more inclined to be sausage shaped. The cytoplasm is either pale blue or tinted with pink and the nucleus, which stains dark pink, is large and less compact than in the female, while the pigment granules are scattered in the cytoplasm around it.

In humans, gametocytes do not multiply, nor cause symptoms but they are the forms which are infective to the mosquito. When a female Anopheline mosquito takes a blood meal, the male and female gametocytes continue their sexual development.

Morphology of Schizonts

Schizonts are rarely seen in the peripheral blood and their presence may indicate a potentially serious parasitemia. Schizonts have 8-36 merozoites and a large mass of golden brown pigment (hemozoin) is seen in the pre-schizont and schizont stage.

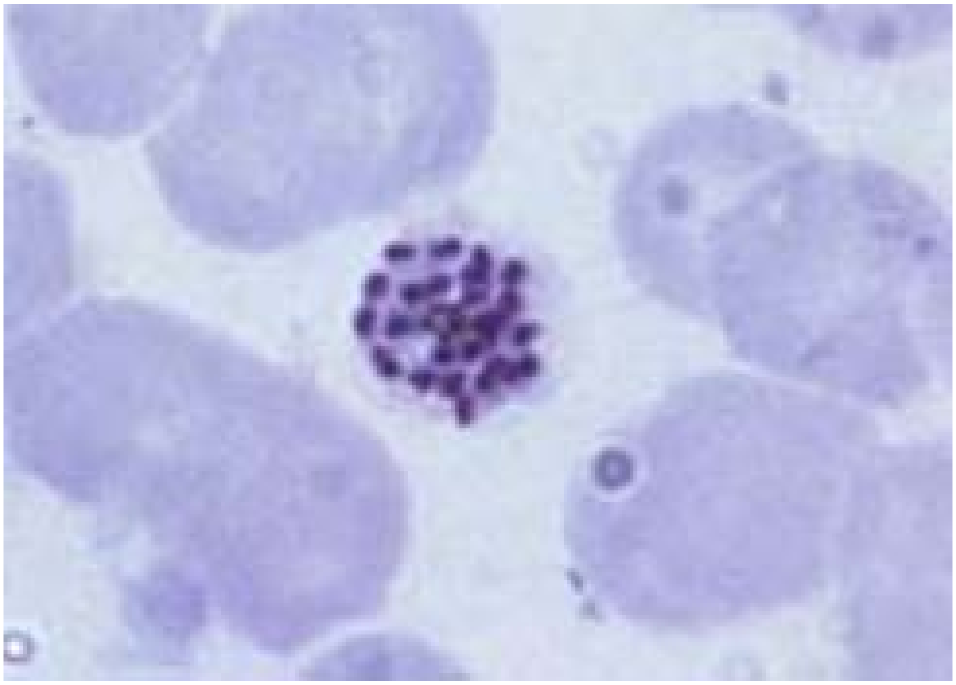

Image 9-4. *Plasmodium falciparum* schizont. Rarely seen in the peripheral blood, a good indicator of a potentially serious parasitemia. They have 8 – 36 merozoites and a large golden brown pigment. (Giemsa stain) (SOURCE: PHIL 5854 - CDC/Steven Glenn, Laboratory & Consultation Division)

Clinical Disease

Symptoms include headache, photophobia, muscle aches and pains, anorexia, nausea and vomiting. Complications include severe anemia cerebral malaria, renal disease, black water fever, dysentery, pulmonary edema and tropical splenomegaly syndrome.

Plasmodium vivax

Introduction

Plasmodium vivax is found almost everywhere malaria is endemic and is the most predominant of the malaria parasites. Causing 43% of all cases of malaria in the world, it also has the widest geographical distribution. Although the disease itself is not usually life threatening, it can cause severe acute illness.

Plasmodium vivax does not infect West Africans due to the fact that West Africans do not possess the Duffy Antigen on the red blood cells which the parasite requires to enter the red blood cell. It has an incubation period of between 10 and 17 days which is sometimes prolonged to months or years due to the formation of hypnozoites. It has a periodicity of 48 hours. *Plasmodium vivax* infections are usually characterized by the presence of more than one developmental stage in the peripheral blood film. The parasites parasitize young enlarged erythrocytes and Schüffner’s dots develop on the erythrocyte membrane.

Life Cycle

The aspects of the life cycle which are specific to *P. vivax* are as follows:

- The degree of infectivity is low, only the young immature corpuscles are infected; about 2% of erythrocytes are parasitized.
- The periodicity of the asexual cycle is closely synchronized.
- Hypnozoites develop in the liver, so that relapses may occur.

Morphology of Trophozoites

Most trophozoites of *P. vivax* are already several hours old when they appear in peripheral blood and by that time the Schüffner's dots are already visible. The trophozoites are actively amoeboid and contain single or sometimes double chromatin dots that are either circular or ovoid. As the trophozoites mature, the Schüffner's dots increase in number and size and the parasite changes from large irregular rings to rounded or ovoid forms in mature trophozoites.

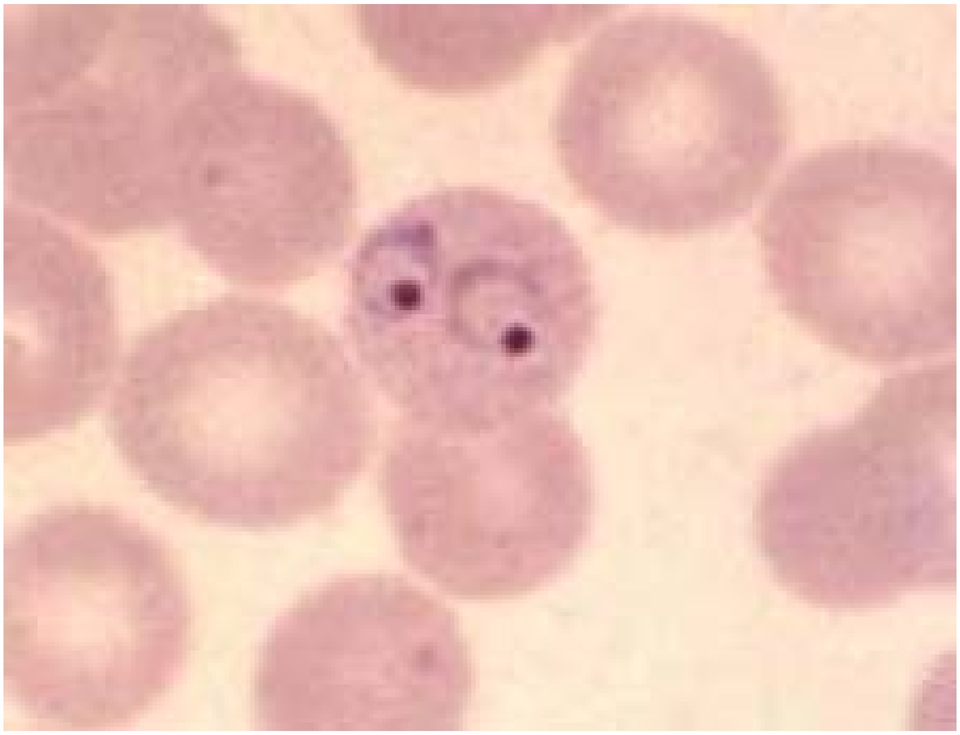

Image 9-5. Trophozoites of *Plasmodium vivax* are already several hours old when they appear in the peripheral blood and therefore, you can already see the Schüffner's dots. They contain single or sometimes double chromatin dots. (Giemsa stain) (SOURCE: PHIL 5928 - CDC/ Steven Glenn, Laboratory & Consultation Division)

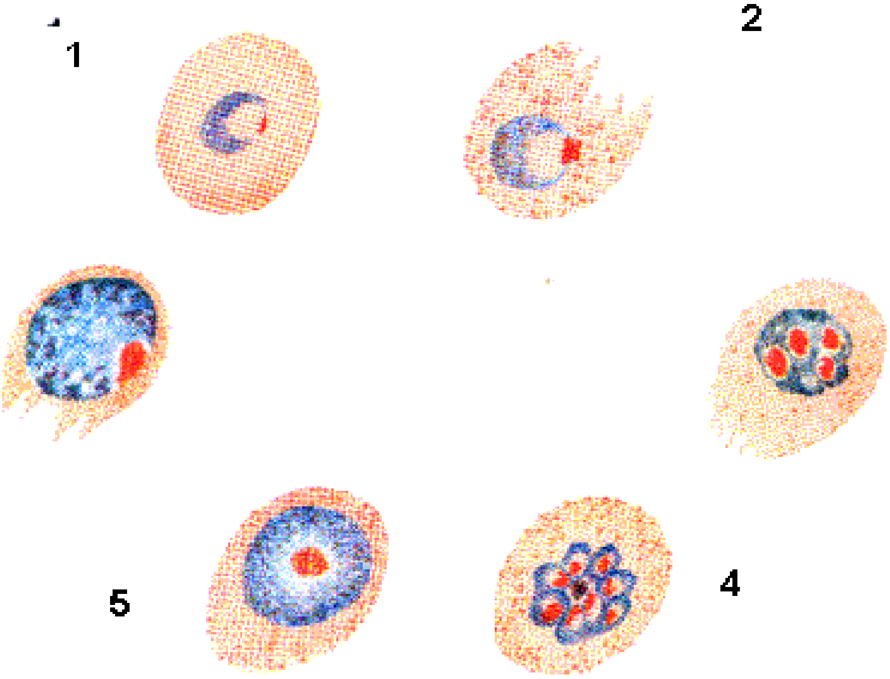

Illustration 9-7. Diagrammatic illustration of the morphology of the different stages of the *Plasmodium vivax* life cycle in thin blood films. **1)** Early trophozoites / ring forms (accole forms, not shown here, are occasionally seen). **2)** Developing trophozoites are large and irregular with a prominent vacuole. **3)** Immature schizonts, are amoeboid and almost fill the red blood cell. **4)** Mature schizonts, almost fill the red blood cell. **5)** Microgametocyte, large numbers appear after 3–5 days. **6)** Macrogametocyte, large numbers appear after 3–5 days. (SOURCE: Unknown)

Morphology of Gametocytes

Mature female gametocytes are large rounded parasites which fill or nearly fill the host cell. The cytoplasm is blue and fairly homogenous. The nuclear chromatin is a single, well-defined purplish mass, varied in form and usually peripheral in distribution. Mature male gametocytes can be distinguished from females by the large, loose and ill-defined mass of chromatin and by their paler color and smaller mass.

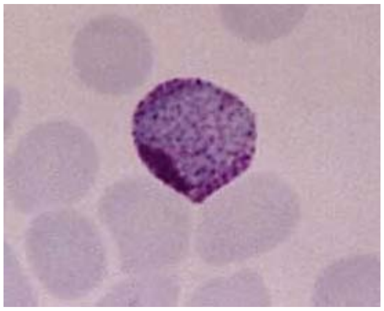

Image 9-6. Mature female *Plasmodium vivax* gametocytes are large rounded parasites which fill or nearly fill the host cell. The cytoplasm is blue and fairly homogenous. The nuclear chromatin is a single, well-defined purplish mass, varied in form and usually peripheral in distribution. (Giemsa stain) (SOURCE: PHIL 5138 - CDC/ Dr. Mae Melvin)

Morphology of Schizonts

The parasitized red cells are much enlarged containing Schüffner's dots. The parasites are large, filling the enlarged red cell. There are between 12-24 merozoites in the schizonts (usually16). The pigment is a golden brown central loose mass.

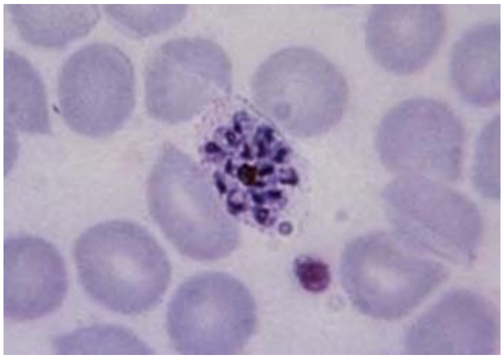

Image 9-7. A schizont of *Plasmodium vivax.* The parasites are large, filling the enlarged red cell. There are between 12 - 24 merozoites in the schizonts (usually16). The pigment is a golden brown central loose mass. (Giemsa stain) (SOURCE: PHIL 5925 - CDC/ Steven Glenn, Laboratory Training & Consultation Division)

Clinical Disease

Symptoms include headache, photophobia, muscle aches and pains, anorexia, nausea and vomiting. Complications due to *P. vivax* are relatively rare and arise from a previous debility or pre-existing disease.

Plasmodium ovale

Introduction

Plasmodium ovale is widely distributed in tropical Africa especially the west coast, but it is not often encountered. It has also been reported in South America and Asia. It has an incubation period of 10–17 days which is sometimes prolonged to months or years due to the formation of hypnozoites. It has a periodicity of 48 hours; the fever it produces is milder than that caused by the benign tertian *P. falciparum.*

Life Cycle

The features of the life cycle which are specific to *P. ovale* are as follows:

- It morphologically resembles *P. malariae* in most of its stages.

- The changes produced in the erythrocytes in general are similar to those produced by *P. vivax,* but Schüffner's dots appear considerably earlier (in the ring stage) and are coarser and more numerous.
- In the oocyst the pigment granules are (usually) characteristically arranged in two rows crossing each other at right angles.
- Hypnozoites develop in the liver so that relapses may occur.

Morphology of Trophozoites

Parasites of *P. ovale* are usually found in enlarged and stippled red blood cells (James's dots), similar to those found in *P. vivax* infections. Host cells show an oval shape, particularly those containing younger stages of the parasites and the host cell may also show "spiking" or fimbriation.

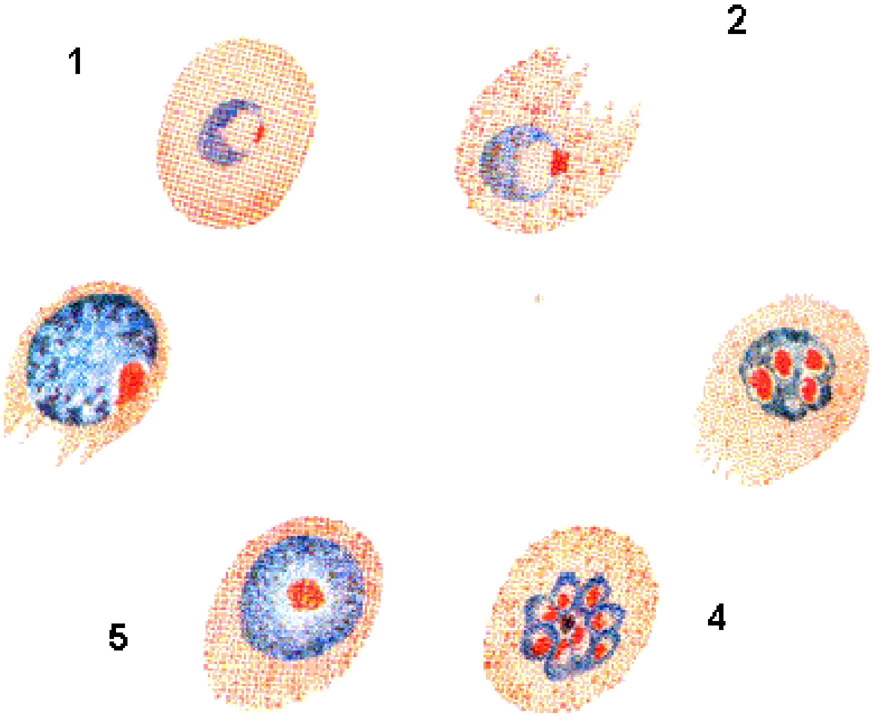

Illustration 9-8. Diagrammatic illustration of the morphology of the different stages of the *Plasmodium ovale* life cycle in thin blood films. **1)** Early trophozoites / ring forms, are dense rings with well- defined masses of chromatin. **2)** Developing trophozoites, small and compact with an inconspicuous vacuole. **3)** Immature schizonts, compact and almost fill the red blood cell. **4)** Mature schizonts, fill ¾ of the red blood cell. **5)** Microgametocytes, low numbers appear after 12-14 days. **6)** Macrogametocytes, low numbers appear after 12–14 days.

Young trophozoites are found as compact rings in cells containing Schüffner's dots. The trophozoite rings remain compact as they develop and show little of the amoeboid features common in *P. vivax*. Small, scattered pigment granules can be seen in developing trophozoites which disperse as the trophozoite matures. Late trophozoites are round and consolidated with an increase in cytoplasm, they are very similar to *P. vivax* at this stage.

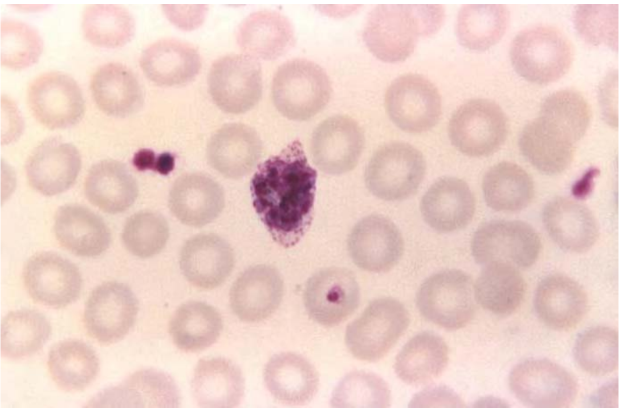

Image 9-8. Trophozoite of *Plasmodium ovale.* Young trophozoites are found as compact rings in cells containing Schüffner's dots. The trophozoite rings remain compact as they develop. Late trophozoites are round and consolidated with an increase in cytoplasm, they are very similar to *P. vivax* at this stage. (SOURCE: PHIL 5935 - CDC/ Steven Glenn, Laboratory & Consultation Division)

Morphology of Gametocytes

The mature gametocytes are round, filling two thirds of the red cell. The red blood cell is slightly enlarged and stippled and contains pigment which has a distinct arrangement of concentric rods, mostly at the periphery.

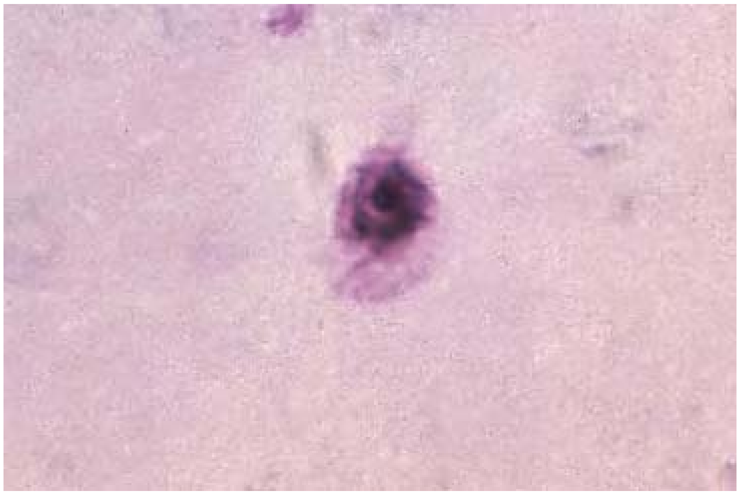

Image 9-9. Gametocyte of *Plasmodium ovale.* The mature gametocytes are round, filling two thirds of the red cell. (Giemsa stain) (SOURCE: PHIL 3474 – CDC/Dr. Mae Melvin)

Morphology of Schizonts

The parasite is smaller than red blood cells and contains 6-12 merozoites, usually 8 in a single ring. The pigment is a brown / greenish central clump. The red cell slightly enlarged, stippled, frequently oval and fimbriated.

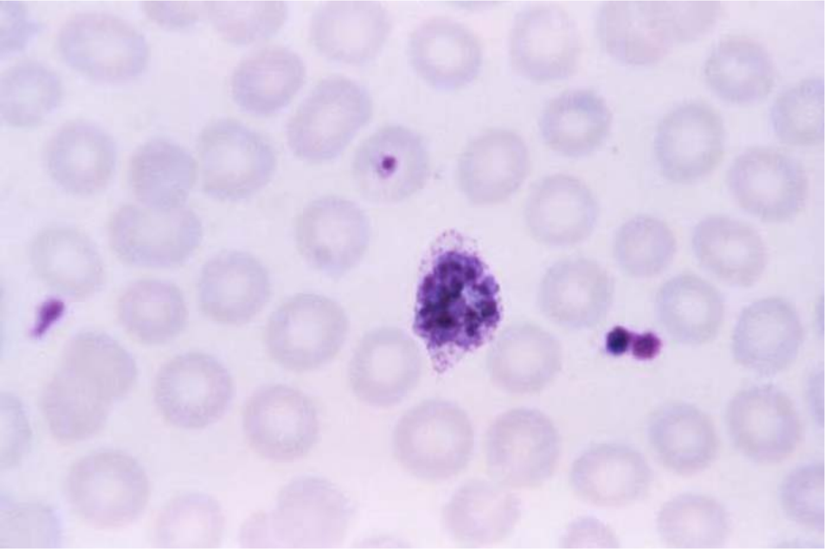

Image 9-10. Schizont of *Plasmodium ovale.* The parasite is smaller than the red blood cell and contains 6 – 12 merozoites. The red cell is slightly enlarged, stippled, frequently oval and fimbriated. (Giemsa stain) (SOURCE: PHIL 5846 - CDC/ Steven Glenn, Laboratory & Consultation Division)

Clinical Disease

Symptoms, like those of *P. vivax,* include headache, photophobia, muscle aches and pains, anorexia, nausea and vomiting. Complications due to *P. ovale* are relatively rare and arise due do a previous debility or pre-existing disease.

Plasmodium malariae

Introduction

Plasmodium malariae occurs mainly in the subtropical and temperate areas where *P. falciparum and P. vivax* occur. It is responsible for approximately 7% of all malaria in the world. It has an incubation period of 18–40 days and a periodicity of 72 hours.

Life Cycle

The features of the life cycle specific to *P. malariae* are as follows:

- Infected erythrocytes are not larger than uninfected ones and sometimes even smaller.
- Mature erythrocytes are attacked, but reticulocytes generally are not
- The density of parasites is very low; about 0.2% of erythrocytes are parasitized.
- It is often difficult to distinguish between a large trophozoite and an immature gametocyte.

Morphology of Trophozoites

Parasites of *P. malariae* are typically compact heavily pigmented parasites which are usually smaller and more deeply stained than normal. They tend to parasitize small, old red blood cells, they do not contain any inclusion dots and the parasitaemia is usually low.

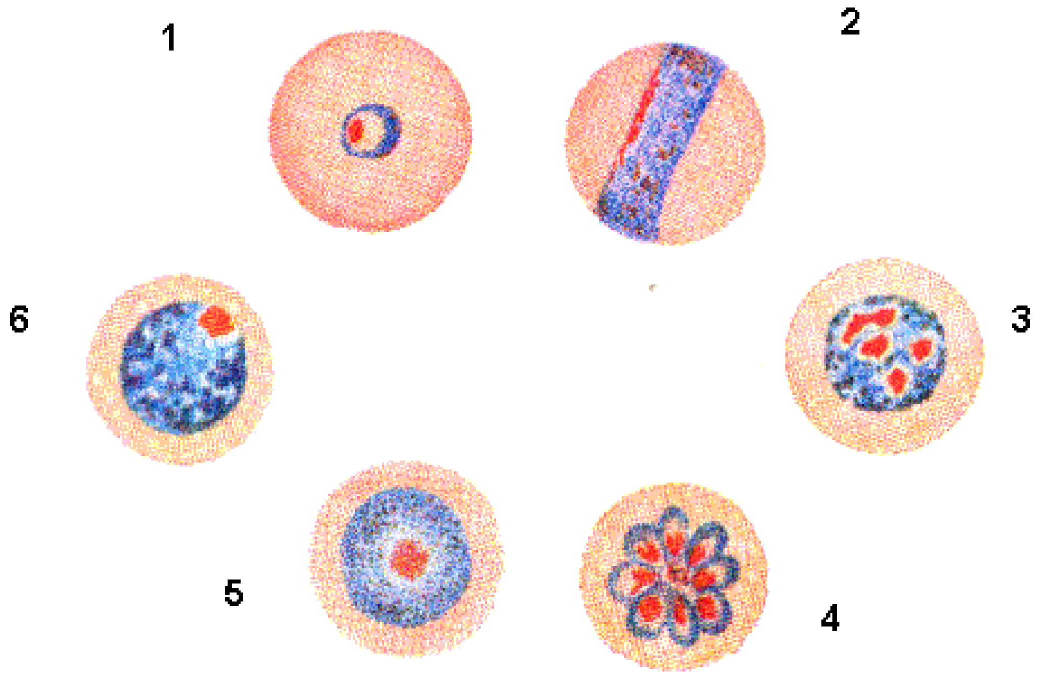

Illustration 9-9. Diagrammatic illustration of the morphology of the different stages of the *Plasmodium malariae* life cycle in thin blood films. **1)** Early trophozoites / ring forms, compact rings containing one mass of chromatin. **2)** Developing trophozoites, small and compact (often band forms) with an inconspicuous vacuole. **3)** Immature schizonts, compact and almost fill the red blood cell which contains scattered pigment. **4)** Mature schizonts, almost fill the red blood cell. **5)** Microgametocytes, low numbers appear after 7–14 days. **6)** Macrogametocytes, low numbers appear after 7-14 days.

Trophozoites are found as fairy large fleshy rings with a single chromatin dot. These can be very distorted and can often take the form of bands across the cell. All trophozoites have a single chromatin dot and contain pigment.

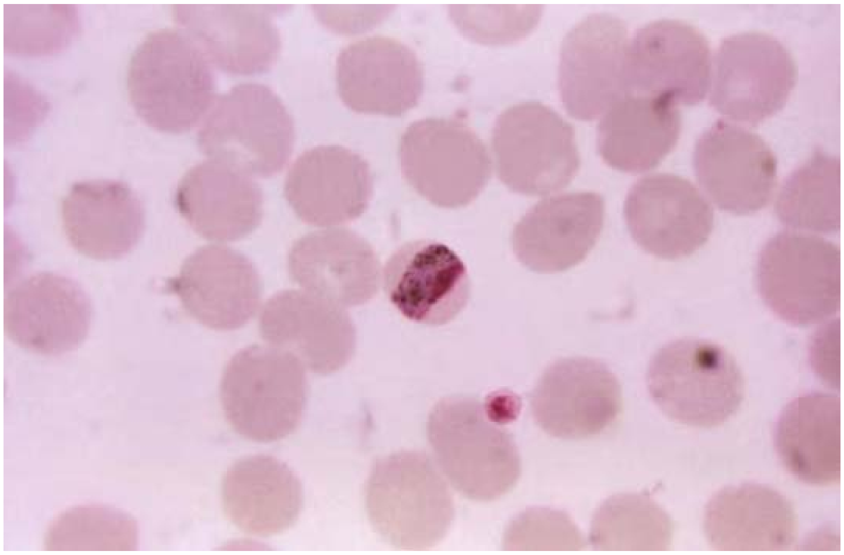

Image 9-11. Trophozoite of *Plasmodium malariae.* These can be very distinct and distorted by taking the form of a band across the cell. (Giemsa stain) (SOURCE: PHIL 639 - CDC/ Dr. Mae Melvin, Steve Glenn)

Morphology of Gametocytes

Gametocytes contain large amounts of black pigment, with chromatin present as a compact mass in females and diffuse in males. They occupy less than two thirds of the red blood cell.

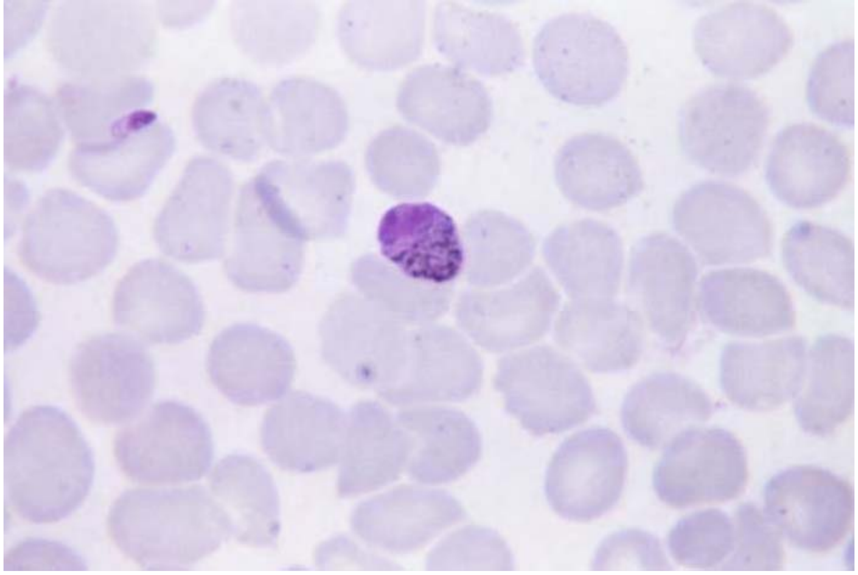

Image 9-12. *Plasmodium malariae* gametocyte. They contain large amounts of black pigment, with chromatin present as a compact mass in females (macrogametocyte) and diffuse in males (microgametocyte). (Giemsa stain) (SOURCE: PHIL 5837 - CDC/ Steven Glenn, Laboratory & Consultation Division)

Morphology of Schizonts

Schizonts are usually few in numbers with 6-12 large merozoites in a single ring. Pigment is usually present as a central black mass. The parasites present are generally only found at one stage of schizogony development.

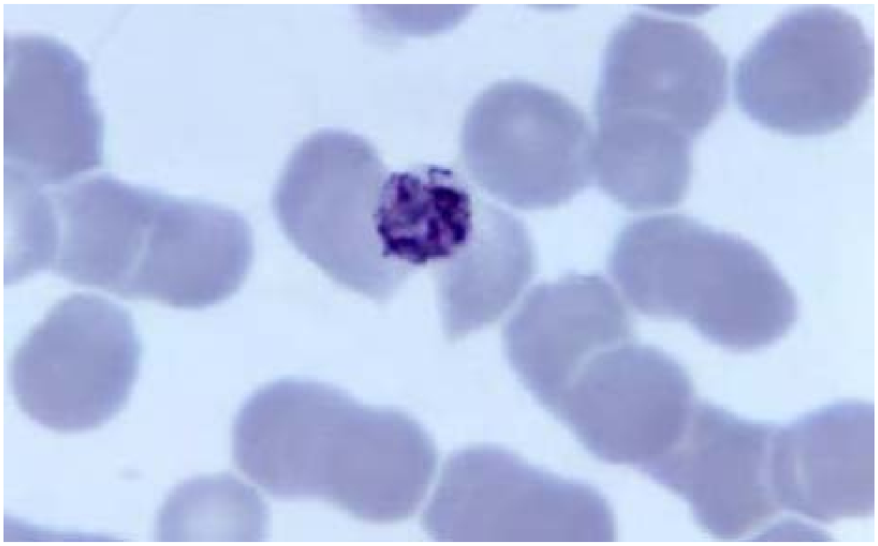

Image 9-13. Schizont of *Plasmodium malariae.* They are usually few in numbers with 6–12 large merozoites in a single ring. Pigment is usually present as a central black mass. (Giemsa stain) (SOURCE: PHIL 4826 - CDC/Dr. Mae Melvin)

Clinical Disease

Symptoms include headache, photophobia, muscle aches and pains, anorexia, nausea and vomiting. *Plasmodium malariae*, like *P. vivax* and *P. ovale* are relatively benign. However, chronic infections in children may lead to nephrotic syndrome due to immune complexes depositing on the glomerular wall.

9.3. Laboratory Diagnosis of Malaria Parasites

Introduction

The definitive diagnosis of malaria infection is still based on finding malaria parasites in blood films. In thin films the red blood cells are fixed so the morphology of the parasitized cells can be seen. Species identification can be made, based upon the size and shape of the various stages of the parasite and the presence of stippling (i.e. bright red dots) and fimbriation (i.e. ragged ends). However, malaria parasites may be missed on a thin blood film when there is a low parasitemia. Therefore, examination of a thick blood film is recommended. With a thick blood film, the red cells are approximately 6-20 layers thick which results in a larger volume of blood being examined.

Thick Blood Films

In examining stained thick blood films, the red blood cells are lysed (destroyed), so diagnosis is based on the appearance of the parasite. In thick films, organisms tend to be more compact and denser than in thin films.

Field's Stain Method for Thick blood Films

The method recommended for staining thick blood is Field's Stain which is made from two components. Field's A is a buffered solution of azure dye and Field's B is a buffered solution of eosin. Both Field's A and B are supplied ready for use by the manufacturer.

Method

1. Place a drop of blood on a microscope slide and spread to make an area of approximately 1 cm^2. It should just be possible to read small print through a thick film.

2. The film is air dried and NOT fixed in methanol.

3. The slide is dipped into Field's stain A for three seconds.

4. The slide is then dipped into tap water for three seconds and gently agitated.

5. The slide is dipped into Field's stain B for three seconds and washed gently in tap water for a few seconds until the excess stain is removed.

6. The slide is drained vertically and left to dry.

Microscopic Features of the Field's Stained Thick Blood Film

- The end of the film at the top of the slide when it was draining should be looked at. The edges of the film will also be better than the centre, where the film may be too thick or cracked.

- In a well-stained film the malaria parasites show deep red chromatin and pale blue cytoplasm.

- White cells, platelets and malaria pigment can also be seen on a thick film. The leukocyte nuclei stain purple and the background is pale blue. The red cells are lysed and only background stroma remains. The occasional red cell may fail to lyse.

- Schizonts and gametocytes, if present, are also easily recognizable.

- A thick film should be examined for at least 10 minutes, which corresponds to approximately 200 oil immersion fields, before declaring the slide negative.

- As a result of hemolysis of the red blood cells due to staining of an unfixed film, the only elements seen are leukocytes and parasites, the appearance of the latter being altered. Consequently:

1. The young trophozoites appear as incomplete rings or spots of blue cytoplasm with detached chromatin dots.

2. The stippling of *P. vivax* and *P. ovale* may be less obvious although occasionally ghost stippling may be seen.

3. The cytoplasm of late trophozoites of *P. vivax* and *P. ovale* may be fragmented.

- Caution should be exercised when examining thick blood films as artifacts and blood platelets may be confused with malaria parasites.

Thin Blood Films

When examining thin blood films for malaria you must look at the infected red blood cells and the parasites inside the cells.

1. Rapid Field's Stain for Thin Films

This is a modification of the original Field's stain to enable rapid staining of fixed thin films. This method is suitable for malaria parasites, *Babesia* sp., *Borrelia* sp. and Leishmania sp.

Method

1. Air dry the film

2. Fix in methanol for one minute.

3. Flood the slide with 1 ml of Field's stain B, diluted 1 in 4 with distilled water.

4. Immediately, add an equal volume of undiluted Field's stain A, mix well and allow to stain for 1 minute.

5. Rinse well in tap water and drain dry.

Uses

This is a useful method for rapid presumptive species identification of malarial parasites. It shows adequate staining of all stages including stippling (mainly Maurer's dots). However, staining with Giemsa is always the method of choice for definitive species differentiation.

2. Giemsa Stain for Thin Films.

Method

1. Air dry thin films
2. Fix in methanol for one minute
3. Wash in tap water and flood the slide with Giemsa diluted 1 in 10 with buffered distilled water pH 7.2. The diluted stain must be freshly prepared each time.
4. Stain for 25-30 minutes.
5. Run tap water on to the slide to float off the stain and to prevent deposition of precipitate on to the film. Dry vertically.
6. Examine the film using the x100 objective.

Microscopic Features of the Thin Blood Film

1. Examine the tail end of the slide where the red cells are separated into a one-cell-layer thick.
2. An alkaline buffer pH 7.2 is vital for clear differentiation of nuclear and cytoplasmic material and to visualize inclusions such as Schüffner's / James's dots in the red cells. Acidic buffer is unsuitable.
3. The red cells are fixed in the thin film so the morphology of the parasitized cells and the parasites can be seen.
4. On a well stained film the chromatin stains red/purple and the cytoplasm blue. Leukocytes have purple nuclei. The red stippling, if present, should be clearly visible.

Infected Red Blood Cells

1. Look at the size of the infected red blood cells.
2. Are there any Schüffner's dots present or not?

Species	Stages found in circulating Blood	Appearance of Red Blood Cells	
		Size	Stippling
Plasmodium falciparum	Trophozoites Gametocytes	Normal	Maurer's dots or clefts infrequently seen
Plasmodium vivax	All: Schizonts Trophozoites Gametocytes	Enlarged, Maximum size may be 1.5-2 times normal	Schüffner's dots may be present
Plasmodium ovale	All: Schizonts Trophozoites Gametocytes	Enlarged, Maximum size may be 1.25-1.5 times normal	Schüffner's dots may be present
Plasmodium malariae	All: Schizonts Trophozoites Gametocytes	Normal	Ziemann's dots rarely seen

Table 9-1. Stages and appearance of *Plasmodium* species in blood. (SOURCE: CDC)

Species	***P. falciparum***	***P. vivax***	***P. malariae***	***P. ovale***
Host CELL				
Size	Not enlarged	Enlarged	Not enlarged	Enlarged
Shape	Round and sometimes crenated	Round or oval (frequently bizarre form)	Round	Round or oval, often fimbriated
Color	Normal but may become darkened	Normal but inclined to be pale	Normal	Normal
Stippling	Maurer’s dots (large red spots)	Schüffner's dots (small red dots)	Ziemann’s dots (few tiny dots not important for diagnosis)	James’s dots (numerous small red dots)
Pigment	Usually black or very dark brown	Fine golden brown granules seen in cytoplasm	Black or brown coarse granules	Resemble more closely P. *malariae*
PARASITE				
General features	Small, compact dark, staining parasite. Multiple infections of single RBC	Large light staining parasite. Many trophozoites, may be amoeboid	Regular shape and moderate size. Strong tendency to form a band across the infected RBC	Regular shape. Size in between *P. vivax* and *P. malariae*
Common. Stages found in smear	Only rings and gametocytes	Trophozoites, Schizonts, Gametocytes	As in *P. vivax*	As in *P. vivax*
Ring stage	Delicate, small, 1.5 µm Double chromatin and multiple rings common. Accole, wing and marginal forms	Large 2.5 µm, usually single. Prominent thicker chromatin	Similar to *P. vivax* but thicker	Similar to *P. vivax*, more compact
Trophozoite	Compact, small, vacuole inconspicuous, seldom seen in smear	Large, irregular actively amoeboid prominent vacuole Chromatin as dots or threads	Characteristic band form, vacuole inconspicuous	Compact coarse pigment, chromatin as large irregular clumps
Schizont	Small, compact seldom seen in blood smear	Large, filling the RBC, segmented, yellow brown pigment	Nearly fills RBC, like segmented, daisy head, pigment is dark brown	Fills three fourth of RBC, segmented Dark yellow brown pigment
Micro-gametocyte	Larger than RBC, kidney shaped with blunt round ends, cytoplasm reddish blue, fine granules scattered, many in number in smear	Fills enlarged RBC, round or oval, compact cytoplasm, pale blue, Abundant brown granules	Smaller than RBC, very few in PBF, round compact, cytoplasm pale blue. Pigment and chromatin as in *P. vivax*	Of the size of RBC round, compact very few in PBF, cytoplasm pale blue, chromatin and pigment as in *P. vivax*

Table 9-2. Characteristics of *Plasmodium* parasites (SOURCE: CDC)

Rings of the four main species of malaria may look alike. If you see rings, look for older stages. Patients with a *P. falciparum* infection only, rings are usually seen; older stages are present only in severe infections.

In poorly stained slides, Schüffner's dots may not be visible, so it is essential that correct staining methods are used. Also Schüffner's dots may not be seen in the earlier rings of *P. vivax* or *P. ovale.*

Estimation of Percentage Parasitemia of *Plasmodium falciparum*

Counting of red blood cells infected with parasites of *P. falciparum* is essential and the percentage parasitemia should always be reported as this has implications for prognosis and the pattern of treatment employed.

The recommended procedure for estimating the percentage parasitemia in a thin blood film is by expressing the number of infected cells as a percentage of the red blood cells e.g. three parasitized red cells / 100 red blood cells or 3% parasitemia.

A red blood cell infected with multiple parasites counts as one parasitized red cell. The percentage parasitemia should be calculated by counting the number of parasitized red blood cells in 1000 cells in a thin blood film.

Method 2

Alternatively, the World Health Organisation recommends a method which compares the number of parasites in a thick blood film with the white blood cell count.

The parasitemia is estimated by first counting the number of parasites per 200 white blood cells in a thick blood film and then calculating the parasite count / μl from the total white blood cell count / μl.

Knowledge of either % parasitemia or total parasite count is essential for the correct clinical management of *P. falciparum* malaria.

Effects of Anticoagulant on the Microscopic Diagnosis of Malarial Parasites

Thin blood films for malaria diagnosis are best prepared from venous or capillary blood taken directly from the patient, without the addition of anticoagulant. However, this is not usually possible in a clinical laboratory, as many samples are received from general practices and other hospitals. All anticoagulants have some effect on the morphology of malaria parasites and the red blood cell they inhabit. This effect depends on the stage of the parasite, the time taken for the blood to reach the laboratory and the type of anticoagulant used. If it is necessary to use an anticoagulant, the films should be prepared as soon as possible after the blood has been taken. If the films cannot be made immediately, potassium EDTA is the anticoagulant of choice. However if the blood is left for several hours in EDTA, the following effects may be seen.

1. Sexual stages may continue to develop and male gametocytes can exflagellate, liberating gametes into the plasma. These can be mistaken for organisms such as *Borrelia*. Gametocytes of *P. falciparum* which have a characteristic crescent shape, may round up and then resemble those of *P. malariae*.

2. Accole forms, which are characteristic of *P. falciparum,* may be seen in *P. vivax* because of attempted re - invasion of the red blood cell by merozoites.

3. Mature trophozoites of *P. vivax* may condense when exposure becomes prolonged and in cases of extreme exposure, red blood cells containing gametocytes and mature schizonts may be totally destroyed along with the contained parasites. The malaria pigment, hemozoin, always remains and can provide a clue to the presence and, to an expert eye, identity of the parasite.

4. The morphology of the red blood cell may be altered by shrinkage or crenation.

Malaria species	***P. falciparum***	***P. vivax***	***P. ovale***	***P. malariae***
Red Cell Changes	Maurer's dots	Schüffner's dots	James' Dots / Fimbriation	Ziemann's dots
Trophozoite – Ring form	Cytoplasm very fine in young rings; thick and irregular in old rings. Accole forms, multiple infections common.	Cytoplasm fine in young rings. Red cell unaltered in size.	Cytoplasm thicker than that found in *P. vivax.* Red cell unaltered in size.	Cytoplasm noticeably thicker. Red cell unaltered in size.
Trophozoite – Growing form	Red cell unaltered in size, sometimes stippled with Maurer's dots. Parasite is compact; pigment is dense brown or black mass.	Red cell enlarged, stippled. Parasite - ameboid, vacuolated; pigment fine and scattered, golden brown.	Red cell unaltered in size, or slightly enlarged. Stippled; may be oval and fimbriated. Parasite – compact, rounded; pigment fine brown grains.	Red cell unaltered in size. Parasite – compact, ugly, rounded or band-shaped; dark brown / black pigment often concentrates in a line along one edge of the band.
Mature schizonts	Red cell unaltered in size. Parasite – merozoites 8 – 36; pigment clumped, black. Rare in peripheral blood	Red cell much enlarged, stippled. Parasite – large, filling enlarged red cell. Merozoites 12 - 24, usually 16; pigment golden brown central loose mass.	Red cell slightly enlarged, stippled, frequently oval and fimbriated. Parasite – smaller than red cell. Merozoites 6 – 12, usually 8 in a single ring; pigment, brown / greenish central clump.	Red cell unaltered in size. Parasite – fills red cell. Merozoites 6 - 12, usually 8, sometimes forming a rosette; pigment, brown / black central clump.
Gametocytes	Red cell distorted. Parasite – crescent. Rare in early cases < 10 days.	Red cell enlarged, stippled. Parasite - large, fills red cell.	Red cell slightly enlarged, stippled. Parasite – round, filling 2/3 of the red cell.	Red cell unaltered in size. Parasite – small, round, fills the red cell.

Table 9-3. Differential diagnostic features of human *Plasmodia* species - Giemsa stained thin film of peripheral blood.

Chapter 10.

The Blood Nematodes

10.1 The Blood Nematodes

These nematodes are known as filariae and consist of a group of nematodes which have successfully invaded the blood stream, connective tissue or serous cavities of vertebrates. They are long thread –like nematodes.

Many of them are of medical and veterinary importance attacking man and various domestic animals being transported by various vectors, including mosquitoes. The nematodes from this order require intermediate hosts for the completion of their life cycle.

The morphology of these nematodes consist of a cylindroid pharynx with an anterior muscular portion and a posterior glandular portion; the males have well-developed alae (a flat, wing-like anatomic process or part) and spirally coiled tails.

Sexually mature female worms release microfilaria, which are pre-larval stages. These are released into the bloodstream. Most species are known to be ovoviviparous and some have ‘sheathed’ microfilaria.

The filarial nematodes which parasitize man consist of *Wuchereria bancrofti*, *Brugia malayi, Brugia timori* and *Loa loa*, *Onchocerca volvulus, Mansonella perstans, Mansonella streptocerca* and *Dipetalonema streptocerca.*

They inhabit a range of locations within the body; lymph glands, deep connective tissue, subcutaneous tissues or mesenteries. Invasions of these tissues usually result in inflammatory reactions which is a typical symptom of a human filarial infection. In some cases these result in fleshy deformities known as elephantiasis.

It has been estimated that approximately one billion people in tropical and subtropical countries are exposed to the risk of filarial infections and at least 200 million are infected with filariasis. The species which are primarily responsible for these human filarial infections are *Wuchereria bancrofti, Brugia malayi* and *Onchocerca volvulus*.

Wuchereria bancrofti

Introduction

Wuchereria bancrofti is a nematode causing lymphatic filariasis throughout the tropics and subtropics and is transmitted by the mosquito. There are two strains of *W. bancrofti*;

1. The nocturnal periodic strain which is widely distributed in endemic regions (i.e. Africa, India and the Far East and also parts of China, Korea and Japan) with the microfilariae being in their highest concentrations between the hours of 10pm and 2am.

2. The sub-periodic strain which is found in the Pacific region, and has a microfilaremia all the time with the highest numbers being detected between noon and 8pm.

Humans are the only known reservoir host of *W. bancrofti*. Infection rates in some communities in East Africa exceed 30% of adults causing revolting swellings of the legs or genital system, known as elephantiasis in man. The adult worm occurs in tightly coiled nodular masses in the major lymphatic ducts.

The main vector is *Culex quinquefasciatus*, a mosquito that is particularly common in towns and cities, breeding in organically polluted water, resting in houses and feeding by night on their human occupants. Typical breeding sites include: storm drains blocked with domestic refuse, accumulations of domestic waste water, inadequately covered septic tanks and pit latrines.

In rural areas throughout Africa, *Anopheles gambiae* and *Anopheles funestus* are involved in transmission. Elsewhere other anopheline mosquito species may transmit bancroftian filariasis in rural areas. In Papua New Guinea, *Mansonia* species may act as a vector.

Life Cycle

Microfilariae enter the host during a blood meal when the vector, a mosquito, punctures the skin. The infective larvae enter through the wound and migrate to the peripheral lymphatics where they grow to mature male and female worms. They can live there for several years. After mating, the gravid females release sheathed microfilariae into the peripheral blood where they can be detected 8-12 months after the initial infected bite.

The mosquito acquires the infection by ingestion of the microfilaria in the blood meal. The microfilariae lose their sheath on arrival in the stomach of the mosquito due to gastric juices. The larvae migrate to the thoracic muscles and develop into infective larvae over a period of 6-14 days. The larvae then migrate to the mouthparts of the mosquito which infects the host during a blood meal.

The blood stages of filariae, microfilariae, vary in the times when they are present in the peripheral blood, corresponding with the peak biting time of the vector. Thus, in nocturnally periodic forms the microfilaria are present in the peripheral blood circulation at night; during the day they reside in the deep tissues, particularly the lungs.

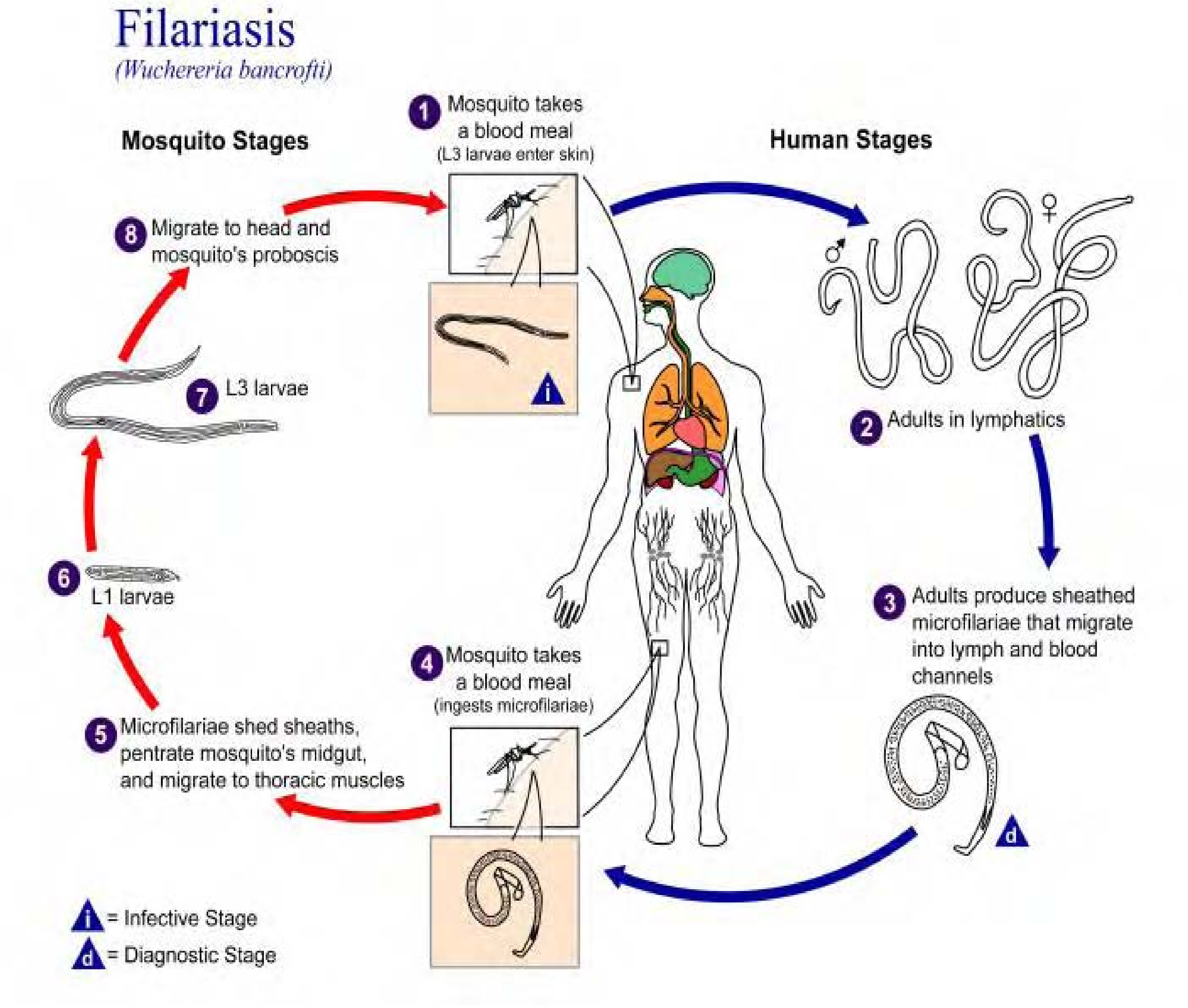
Filariasis
(Wuchereria bancrofti)
Mosquito Stages
Human Stages
1 Mosquito takes a blood meal (L3 larvae enter skin)
2 Adults in lymphatics
3 Adults produce sheathed microfilariae that migrate into lymph and blood channels
4 Mosquito takes a blood meal (ingests microfilariae)
5 Microfilariae shed sheaths, pentrate mosquito's midgut, and migrate to thoracic muscles
6 L1 larvae
7 L3 larvae
8 Migrate to head and mosquito's proboscis
♂
♀
i
d
i = Infective Stage
d = Diagnostic Stage

Illustration 10-1. Diagram showing the life cycle of *Wuchereria bancrofti,* the filarial nematode known to cause the disfiguring disease, elephantiasis, in man. (SOURCE: PHIL 3425 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The adult worms are white and threadlike. The male measures between 2.5–4cm whereas the female is larger, measuring between 8-10cm.

The microfilariae are 230-275μm in length. The tail of the microfilariae of *W. bancrofti* tapers to a delicate point and exhibits no terminal nuclei. The sheath the microfilariae of *W. bancrofti* stains with hematoxylin stain.

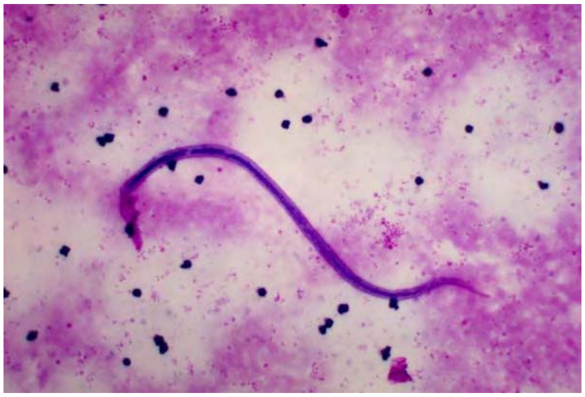

Image 10-1. Microfilaria of *Wuchereria bancrofti*. The microfilaria is sheathed, its body is gently curved, and the tail is tapered to a point. The nuclear column (the cells that constitute the body of the microfilaria) is loosely packed, the nuclei can be visualized individually and do not extend to the tip of the tail. (Micrograph) (SOURCE: PHIL 3008 - CDC/Dr. Mae Melvin)

Clinical Disease

Many patients are asymptomatic. Patients may present with fever, lymphangitis and lymphadenitis. Lymphangitis commonly affects the lower extremities and there may also be genital and breast involvement. An inflammatory reaction occurs in the lymphatic vessels that harbor the adult worms. Edema develops which may resolve after the first few attacks. A late complication resulting in thickening and verrucous changes in the skin known as elephantiasis may occur after recurring lymphangitis. Secondary bacterial and fungal infections may occur in patients with long-standing elephantiasis.

Obstruction of the genital organs may result in hydrocoele formation and scrotal lymphoedema. Obstruction of the retroperitoneal lymphatics may cause the renal lymphatics to rupture into the urinary tract producing chyluria.

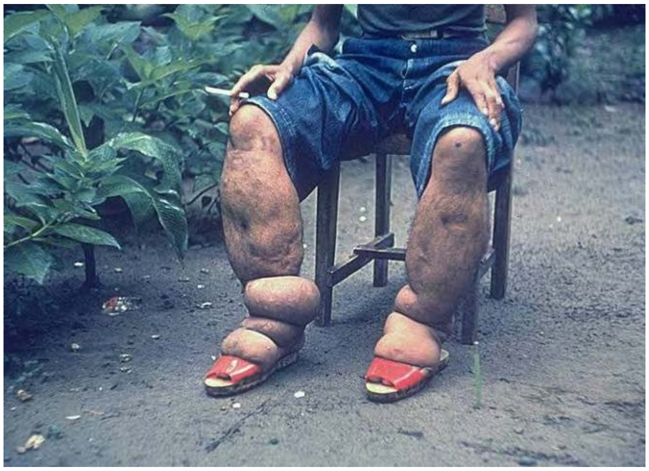

Image 10-2. Lymphatic filariasis: Elephantiasis is the last consequence of the swelling of limbs and scrotum. (SOURCE: PHIL 373 – CDC)

Some patients with filariasis do not exhibit microfilaremia but develop tropical pulmonary eosinophilia which is characterized by peripheral eosinophilia, wheeze and cough. High eosinophilia, high IgE level and high anti-filarial antibody titers are features of this syndrome.

Laboratory Diagnosis

See section 10.2 below

Sheath may or may not stain with Giemsa; does stain with hematoxylin stains. Discrete nuclei. Empty space between the nuclei and the body wall. No nuclei in tip of tail. Innerbody is rarely visible in Giemsa. Does not stain with hematoxylin. Cephalic space as long as it is broad. Tip of tail may be bent underneath the body. Found in blood.

Brugia malayi

Introduction

Brugia malayi is a nematode causing lymphatic filariasis in South East Asia. There are two strains of *B. malayi*;

1. The nocturnal periodic strain which is widely distributed in Asia, the microfilariae being in their highest concentrations between the hours of 10pm and 2am.

2. The sub-periodic strain which is found in Malaysia, Indonesia and the Philippines where humans exhibit a microfilaremia all the time with the highest numbers being detected between noon and 8pm.

Nocturnally periodic Brugian filariasis is primarily a rural disease, being transmitted by various *Anopheles* species of mosquitoes and also by *Mansonia* species, a mosquito that usually bites during the night.

Nocturnally sub-periodic *B. malayi* is transmitted almost exclusively by *Mansonia* species, often different species than those involved in transmitting the periodic form. *Mansonia bonneae* are important vectors in Malaysia, breeding in swamp forests and biting by night, although sometimes by day as well.

This species like *W. bancrofti* also parasitizes the lymph nodes and lymphatics; the adults of the two species are indistinguishable, causing Malayan filariasis.

Life Cycle

The adult worm inhabits the lymphatics and the female produces sheathed microfilariae which circulate in the peripheral blood. The mosquito acquires the infection by ingestion of the microfilaria in the blood meal. The microfilaria loses their sheath on arrival in the stomach of the mosquito. The larvae migrate to the thoracic muscles and develop into infective larvae over a period of 6-14 days. The larvae then migrate to the mouth parts of the mosquito and enter the skin of the definitive host through the puncture wound when a blood meal is taken. The infective larvae enter the peripheral lymphatics where they grow to mature male and female worms.

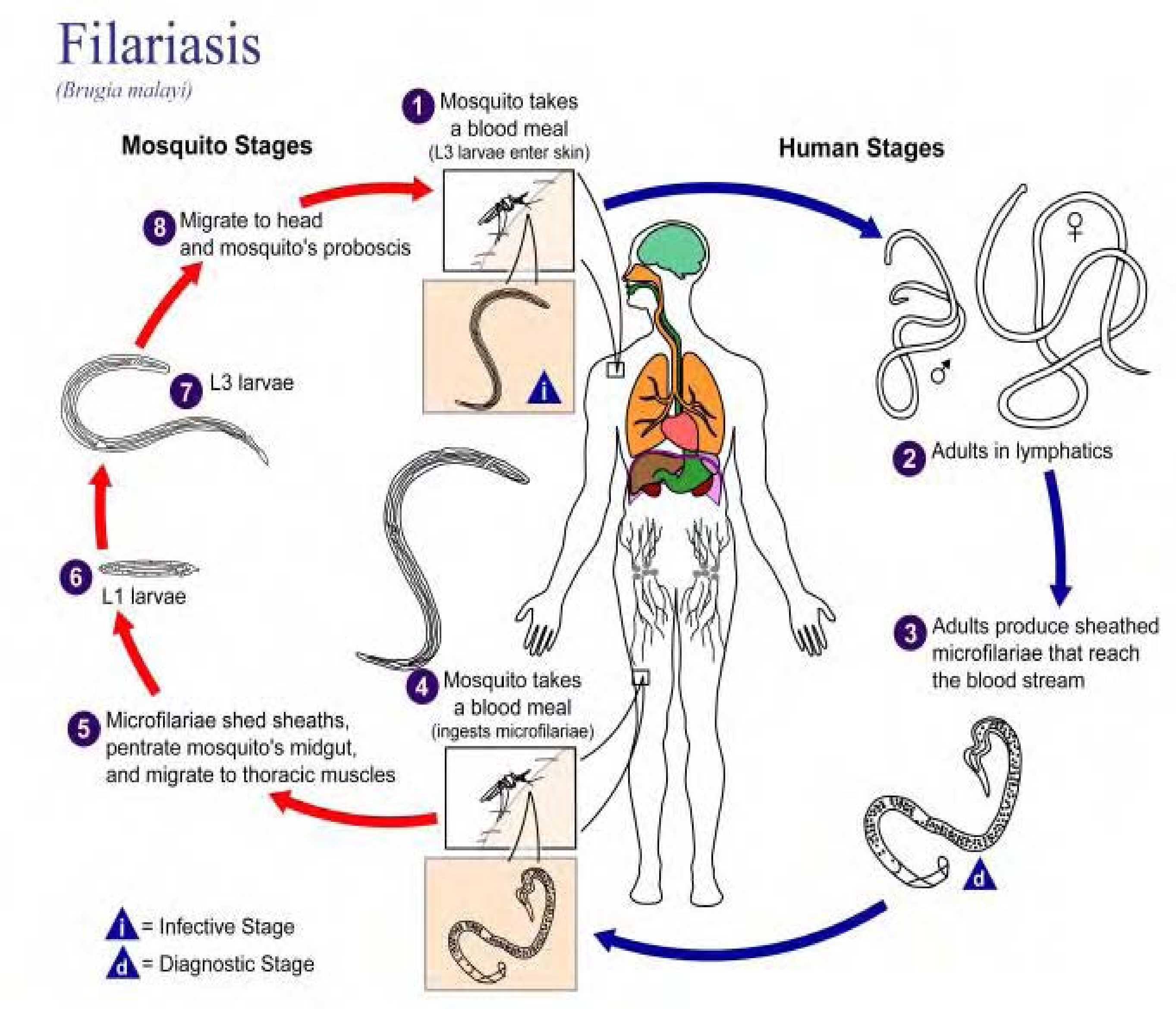
Filariasis
(Brugia malayi)
Mosquito Stages
Human Stages
1 Mosquito takes a blood meal (L3 larvae enter skin)
2 Adults in lymphatics
♀
♂
3 Adults produce sheathed microfilariae that reach the blood stream
4 Mosquito takes a blood meal (ingests microfilariae)
5 Microfilariae shed sheaths, pentrate mosquito's midgut, and migrate to thoracic muscles
6 L1 larvae
7 L3 larvae
8 Migrate to head and mosquito's proboscis
i
d
i = Infective Stage
d = Diagnostic Stage

Illustration 10-2. Diagram showing the life cycle of *Brugia malayi*, a nematode which causes lymphatic filariasis. (SOURCE: PHIL 3379 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

The adult worms of *B. malayi* are smaller than those of *W. bancrofti*. The microfilariae of *Brugia malayi* are 170–230µm in length and have two terminal nuclei that are distinctly separated from the other nuclei in the tail. The last terminal nucleus is quite small and is at the tip of the tail. The sheath stains deep purple with hematoxylin stain.

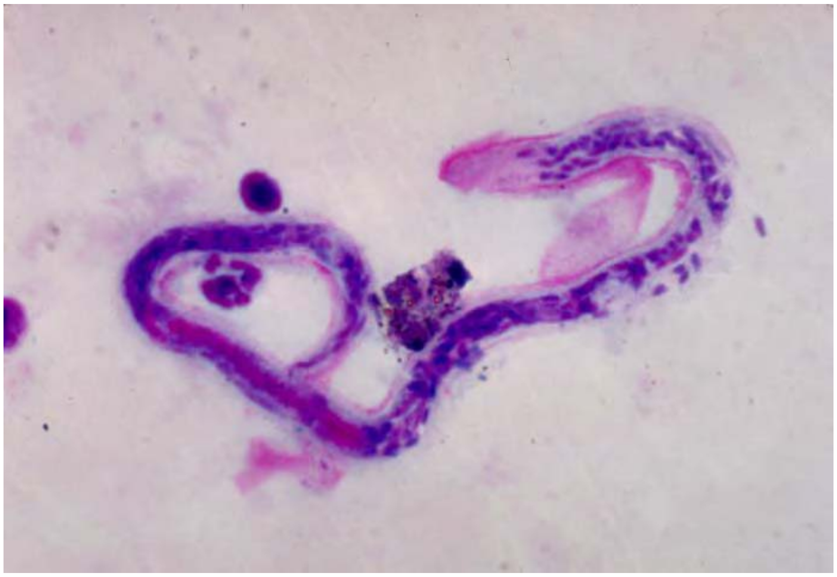

Image 10-3. Microfilaria of *Brugia malayi.* These microfilariae are sheathed measuring approximately 170–230μm in length and have 2 terminal nuclei that are distinctly separated from the other nuclei in the tail. The microfilariae in this species are tightly coiled, and the nuclear column is more tightly packed, preventing the visualization of individual cells. (Blood Smear) (SOURCE: PHIL 3003 - CDC/Dr. Mae Melvin)

Clinical Disease

Clinical features of *B. malayi* are similar to those of *W. bancrofti*, however in *B. malayi*, unlike *Wuchereria bancrofti*, genital involvement, hydrocoele and chyluria are rare.

Many patients are asymptomatic. Patients may present fever. Lymphaginitis and lymphadenitis develop in the lower extremities. An inflammatory reaction occurs in the lymphatic vessels that harbor the adult worms. Edema develops which may resolve after the first few attacks. However, in long standing disease after several episodes of lymphaginitis, thickening and verrucous changes in the skin known as elephantiasis occurs.

Some patients with lymphatic filariasis do not exhibit microfilaremia. However, they do have high eosinophilia, high IgE level and high anti-filarial antibody titers.

Laboratory Diagnosis

See section 10.2 below

Kinked microfilaria. Sheath stains deep pink with Giemsa stain. Does stain with hematoxylin stains. Nuclei crowded and fill the whole body. Empty space between nuclei and body wall. Cephalic space twice as long as it is broad. Innerbody may or may not stain; when it does, it is prominent. Found in blood.

Brugia Timori

Brugia timori is found in the islands of Indonesia and exhibits a strictly nocturnal periodicity. The lifecycle and disease closely resembles that of *Brugia malayi*. However, the microfilariae can be distinguished from those of *B. malayi* in that they are about 310µm in length. The sheath satins pink with Giemsa and the nuclei at the tip of the tail are similar to those of *B. malayi*

Loa loa

Introduction

Loa loa, also known as the African eye worm, is a filarial nematode endemic in the rain forests of West and Central Africa. It is transmitted by *Chrysops* species, also known as mango flies or horse flies and humans are the only known reservoir. It is estimated that 2-13 million humans are infected with the larvae.

Adults migrate in the subcutaneous tissues of man and monkeys, with them eventually migrating across the eyeball under the conjunctiva.

Life Cycle

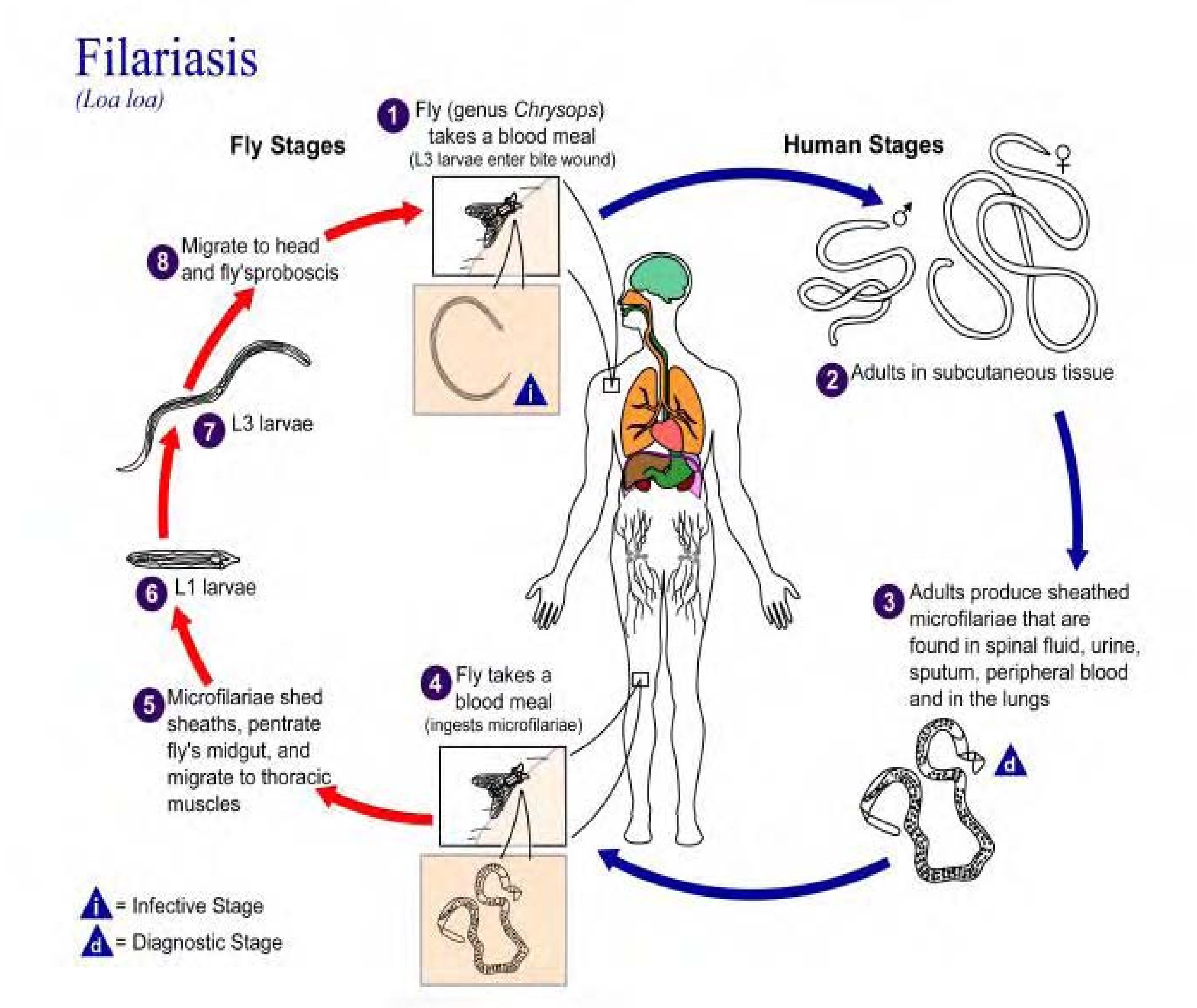
Filariasis
(Loa loa)
Fly Stages
Human Stages
1 Fly (genus Chrysops) takes a blood meal (L3 larvae enter bite wound)
2 Adults in subcutaneous tissue
♀
♂
3 Adults produce sheathed microfilariae that are found in spinal fluid, urine, sputum, peripheral blood and in the lungs
4 Fly takes a blood meal (ingests microfilariae)
5 Microfilariae shed sheaths, pentrate fly's midgut, and migrate to thoracic muscles
6 L1 larvae
7 L3 larvae
8 Migrate to head and fly'sproboscis
i
d
i = Infective Stage
d = Diagnostic Stage

Figure 10-3. Diagram showing the life cycle of *Loa loa,* the African eye worm. (SOURCE: PHIL 3399 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

The adult worms live in the subcutaneous and deep connective tissues and the microfilariae are found in the peripheral blood, where they can be in ingested by the *Chrysops* fly (day biting fly) The adults can live in the tissues for up to 17 years.

Once the microfilariae have been taken up by the Chrysops during a blood meal they develop within the fat body. They develop through to L_3 within 10–12 days. The microfilariae, L_3 re-enter the hosts blood stream when the fly takes another blood meal. They reach adult worms within 4- 6 months living in the subcutaneous and deep connective tissues.

The microfilariae exhibit diurnal periodicity, the highest numbers being detected in blood between 10am and 2pm.

Morphology

Adult males of *Loa loa* are 2–3.5cm long and the females from 5-7cm. The microfilariae of *Loa loa* are 250-300µm. They possess a sheath which stains blue-grey with Delafield's hematoxylin. The sheath does not stain with Giemsa. The tail gradually tapers to a rounded end, the densely packed nuclei extending to the tip.

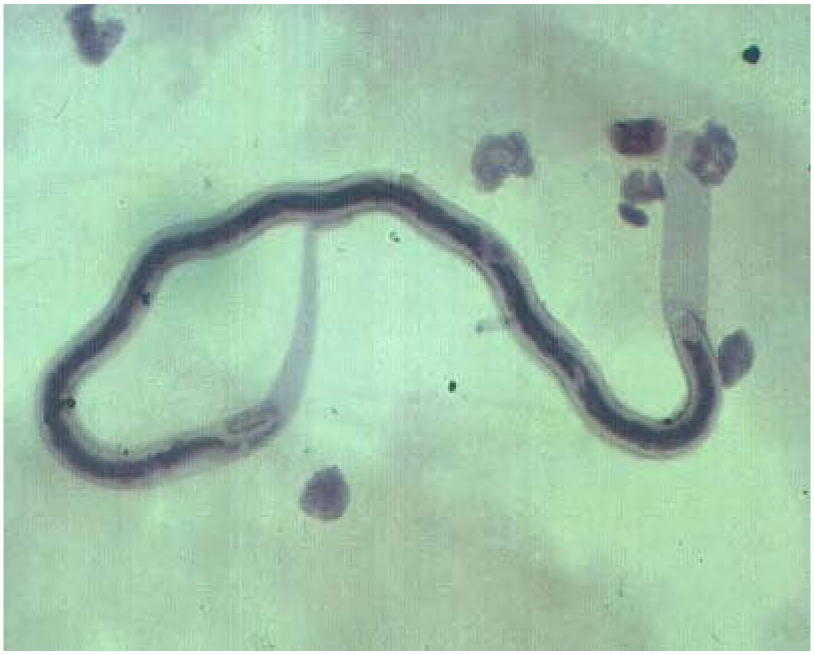

Image 10-4. Microfilaria of *Loa loa,* the African eye worm. The adult worms live in the subcutaneous and deep connective tissues and the microfilariae are found in the peripheral blood, The microfilaria are kinked and sheathed. Nuclei crowded extending to tip of tail; tip of tail tapers. (hematoxylin stain, x400) (SOURCE: Unknown)

Clinical Disease

Many patients infected with *Loa loa* appear to be asymptomatic and the migration of the adult worm through the subcutaneous tissues often goes unnoticed, unless passing beneath the conjunctiva of the eye. They can be seen crossing the eye, but it is a rapid process taking approximately 15–20 minutes. Hyperesinophilia and increased antibody levels, especially IgE are also noted. Eye-worm episodes are as equally common in man as well as women with common reoccurrences. There is an increased incidence with age.

The most common pathology associated with *Loa loa* infections are Calabar swellings, which are inflammatory swellings resulting in a localized subcutaneous edema. These swellings are due the host's response to the worm or its metabolic products and can be found anywhere in the body but most commonly in the extremities. These swellings last from 1–3 days. They develop rapidly and last one to three days, usually accompanied by localized pain, urticaria and pruritis. There is a higher frequency of Calabar swellings in women with common reoccurrences.

Serious complications such as cardiomyopathy, encephalopathy, nephropathy and pleural effusion have been recorded.

Laboratory Diagnosis

See section 10.2 below

Kinked and sheathed microfilaria. Sheath does not stain with Giemsa stain; does stain with hematoxylin stains. Nuclei crowded extending to tip of tail; tip of tail tapers. Cephalic space as long as it is broad. Inner body does not usually stain. Found in blood.

Mansonella species

Introduction

Members of the genus *Mansonella* are filarial nematodes which rarely cause serious disease. However, they can be found in geographical areas where *Wuchereria bancrofti, Loa loa* and *Onchocerca volvulus* also occur and therefore must be differentiated from these pathogenic microfilariae. Unlike the pathogenic blood filariae, they do not exhibit periodicity.

Life Cycle

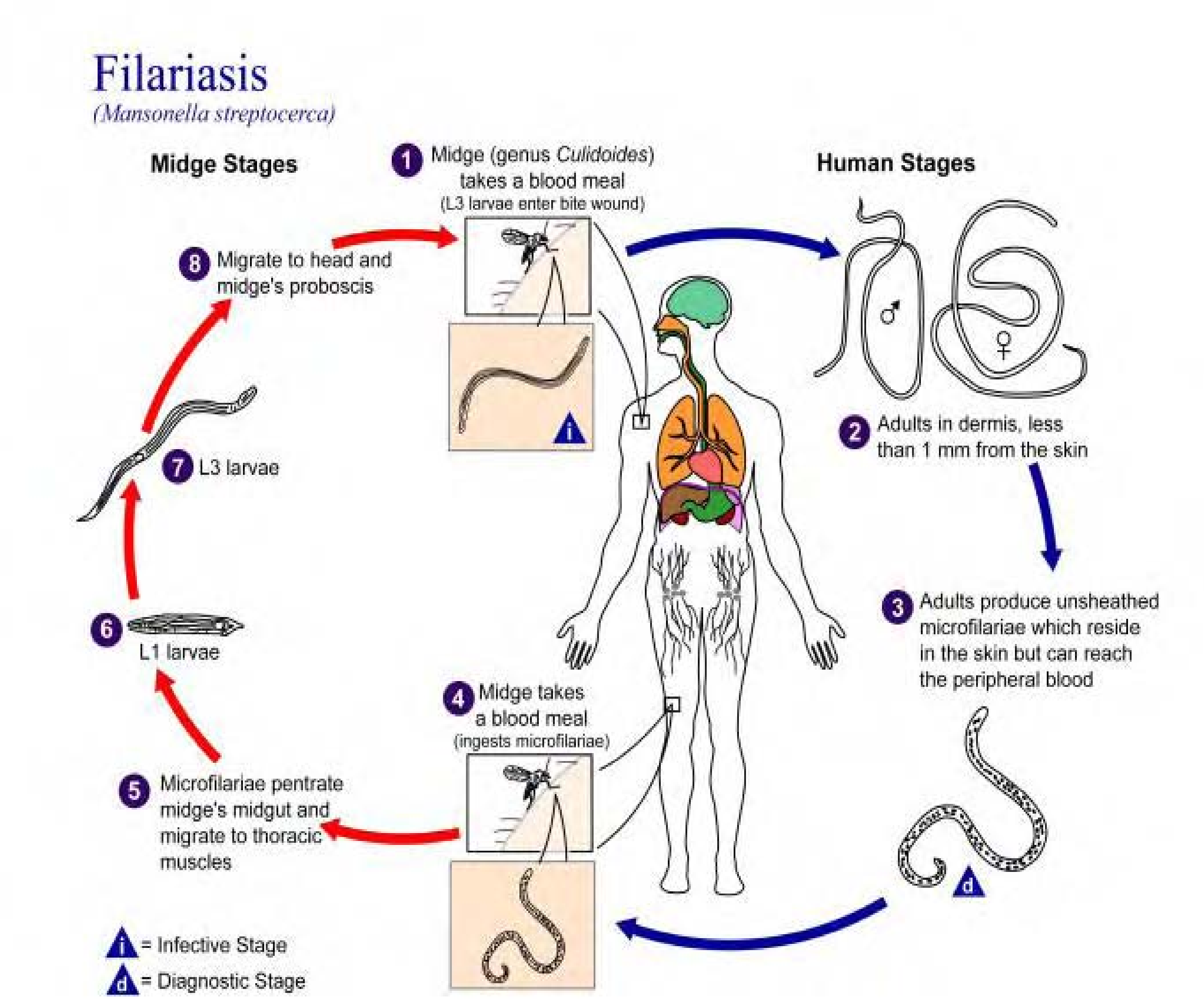
Filariasis
(Mansonella streptocerca)
Midge Stages
1 Midge (genus Culidoides) takes a blood meal (L3 larvae enter bite wound)
Human Stages
8 Migrate to head and midge's proboscis
♂
♀
i
2 Adults in dermis, less than 1 mm from the skin
7 L3 larvae
3 Adults produce unsheathed microfilariae which reside in the skin but can reach the peripheral blood
6 L1 larvae
4 Midge takes a blood meal (ingests microfilariae)
5 Microfilariae pentrate midge's midgut and migrate to thoracic muscles
d
i = Infective Stage
d = Diagnostic Stage

Image 10-4. General life cycle of *Mansonella* species of filarial nematodes. (SOURCE: PHIL 3404 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

There is a general life cycle for the *Mansonella* species of filarial nematodes. The microfilaria is picked up by the vector *Culicoides* sp. (biting midges) during a blood meal. The larvae develop within the body of the *Culicoides* sp. and are re-introduced into the human host when the vector takes another blood meal. They are found in various sites around the human host body.

Mansonella perstans

The microfilariae of *M. perstans* have been found in Africa and South America. This is a mildly pathogenic species in man and apes. They are found in the deep connective tissue and serous cavities.

Morphology

The adult worms live in the peritoneal, pleural and pericardial cavities and their size is comparable to the pathogenic species already discussed. The microfilariae are unsheathed are about 200µm in length and the nuclei extend to the tip of the tail which is rounded.

Clinical Disease

It is difficult to assess the disease associated with *M. perst*ans, however pruritis, fever and subcutaneous swellings have been associated with infection of *M. perstans*. The adult worm appears to cause little or no host reaction. Eosinophilia is common.

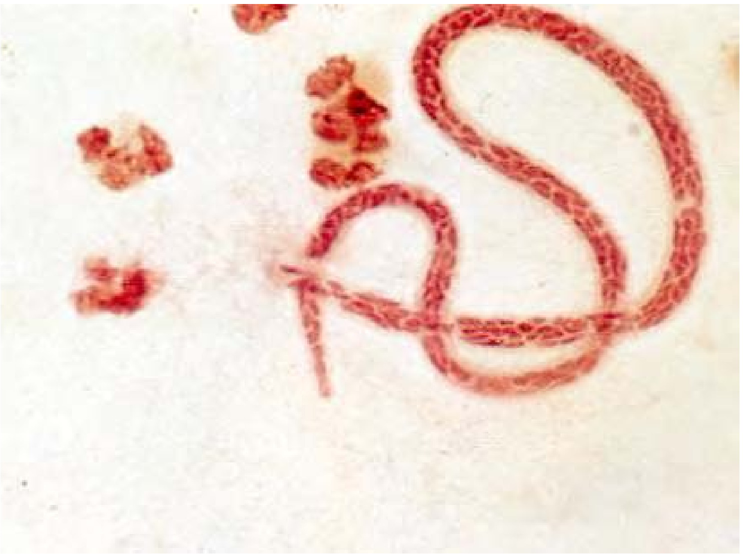

Figure 10-5. Microfilariae of *Mansonella perstans,* a mildly pathogenic nematode to man. The microfilariae are unsheathed are about 200µm in length and the nuclei extend to the tip of the tail which is rounded. (Giemsa stained x920) (SOURCE: Unknown)

Laboratory Diagnosis

See section 10.2 below

Small, thin microfilaria. Does not have a sheath. Nuclei extend to end of tail; last nucleus bigger; tip of tail is blunt. Nuclei stain deeply and “run together”. Found in blood.

Mansonella ozzardi

Mansonella ozzardi nematodes are confined to the westerm hemisphere. These are non-pathogenic filarial nematodes. The parasites cause nodules in the skin of the vertebrate hosts.

Morphology

The adult worms are located in the mesenteric tissues and their size is comparable to the pathogenic species already discussed (0.6m long). The microfilariae are found in the peripheral blood and range between 173-240µm in length. The nuclei do not extend to the tip of the tail which has a pointed end. The male adult worm is almost unknown.

Clinical Disease

Infections caused by *M. ozzardi* are generally symptomless, however lymphadenopathy, arthralgia, fever and eosinophilia have been reported.

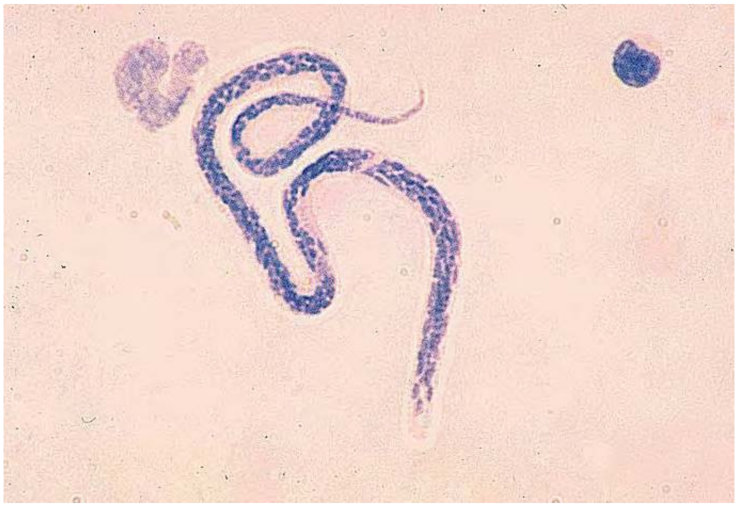

Image 10-6. Microfilariae of *Mansonella ozzardi,* Non-pathogenic filarial nematode to man. The microfilariae are found in the peripheral blood and range between 173 - 240µm in length. The nuclei do not extend to the tip of the tail which has a pointed end. (SOURCE: PHIL 382 - CDC/Dr. Mae Melvin)

Laboratory Diagnosis

See section 10.2 below

Small thin microfilaria. Does not have a sheath. Nuclei do not extend to end of tail; tip of tail tapers. Stains very lightly; tip of tail difficult to see. Found in blood and skin.

Species	Size of Microfilariae	Morphology of microfilariae
Wuchereria bancrofti	210 – 320µm by 8 - 10µm	Sheathed. Tail pointed and clear
Brugia malayi	170 – 260µm by 5 - 6µm	Sheathed. Tail pointed with 2 nuclei
Loa loa	230 – 300µm by 6 - 8µm	Sheathed. Tail blunt with nuclei
Mansonella perstans	200µm by 6µm	Unsheathed. Tail blunt with nuclei
Mansonella ozzardi	250µm by 6 - 7µm	Unsheathed. Tail pointed and clear

Table 10-1. Morphology of the blood microfilariae known to infect man. (SOURCE: CDC)

Species	Geographic distribution	Pathogenicity	Adults (site of infection)	Microfilariae (characteristics)	Vector
Wuchereria bancrofti	Asia, Pacific, Tropical Africa, Americas	Lymphangitis, fever, elephantiasis hydrocoele, chyluria	Lymphatics	Found in blood, sheathed, periodicity variable	Culicidae (mosquitoes)
Brugia malayi	South and East Asia	Lymphagitis, fever. Elephantiasis	Lymphatics	Found in blood, sheathed, nocturnally periodic or subperiodic	Culicidae (mosquitoes)
Dipetalonema perstans	Africa and South America	No definite pathogenicity	Peritoneal & pleural cavity	Found in blood, unsheathed, nocturnally subperiodic	*Culicoides* (biting midges)
Dipetalonema streptocerca	Africa (Ghana and Congo)	Cutaneous edema, elephantiasis	Subcutaneous tissues	Found in skin, unsheathed, nonperiodic	*Culicoides* (biting midges)
Mansonella ozzardi	Central and South America	No definite pathogenitis	Peritoneal cavity	Found in blood, unsheathed, nonperiodic	*Culicoides* (biting midges)
Loa loa	Tropical Africa	Skin swellings, allergic reactions	Subcutaneous tissues	Found in blood, sheathed, diurnally periodic	*Chrysops* (Tabanidae or Horse fly)
Onchocerca volvulus	Africa, Central and South America	Skin nodules, ocular complications (blindness)	Subcutaneous tissues	Found in skin, unsheathed, nonperiodic	*Simulium* (Black fly)

Table 10-2. Comparison of the main human filarial nematodes. (SOURCE: CDC)

10.2. Laboratory Diagnosis

Detection of Microfilariae in Blood

Collection of specimens

The specimen collection times should be selected in accordance with the patient’s clinical symptoms and travel history.

Species	Geographic location	Periodicity	Collection time
Wuchereria bancrofti	Tropics / Subtropics	Nocturnal	2400 hours
Wuchereria bancrofti	Pacific	Diurnal subperiodic	1600 hours
Brugia malayi	SE Asia and SW India	Nocturnal	2400 hours
Brugia malayi	Indonesia	Nocturnal subperiodic	2100 hours
Brugia timori	Indonesia	Nocturnal	2400 hours
Loa loa	West / Central Africa	Diurnal	1300 hours
Mansonella perstans	Africa / South America	Non periodic	Any time
Mansonella ozzardi	Central & South America	Non periodic	Any time

Table 10-3. Periodicity and the advised collection times of the human filarial nematodes. (SOURCE: CDC)

Detection Methods

(i) Polycarbonate membrane filtration.

This technique is very sensitive, enabling very low parasitemias to be detected. It is now the most widely used technique for separating microfilariae from blood.

Nucleopore polycarbonate membranes, 25µm diameter, 5µm pore size, are held in a Millipore Swinnex filter holder, using a rubber gasket to secure the membrane.

Method

a) Place the membrane on the holder with a drop of water.

b) Draw up 10-20 ml of 1:1 saline diluted blood into a 20ml syringe

c) Connect the syringe to the filter and *gently* push the blood through the filter membrane.

d) Repeat until all of the blood has been filtered.

e) Draw up 20 ml of saline into the syringe, flush through the filter, repeat using air.

f) Unscrew the top of the filter and discard the gasket into chloros; use forceps to transfer the membrane to a slide.

g) Add a drop of saline to the membrane and cover with a coverslip.

h) Examine the membrane under the microscope, using a x10 objective. Examine any microfilariae found using a x40 objective to note the presence of a sheath.

(ii) Saline/saponin method.

Reagent

One percent saponin in normal saline.

Method

a) Deliver 2ml of blood (fresh or anticoagulated) into a centrifuge tube and add 8ml of 1% saponin in saline.

b) Mix the blood by inversion, then allow it to stand at room temperature for 15 minutes to allow the blood to hemolyze.

c) Centrifuge at 2,000rpm for 15 minutes to deposit the microfilariae.

d) Discard the supernatant and use the deposit to make a wet preparation.

e) Examine the slide using the x10 objective. Active microfilariae can be seen and produce a snake-like movement as they disturb the cell suspension.

If it is not easy to inspect the microfilariae due to excess “wriggling” a little 10% formalin can be run under the coverslip to immobilize them.

Confirmation of species can be made by using appropriate staining methods to demonstrate nuclear morphology.

Staining Methods for Microfilariae

When filariasis is suspected, a geographical and clinical history helps to determine the most appropriate collection time. Thick and thin blood films can be examined. However this is an insensitive method due to the low microfilaremia, and larger volumes of blood need to be examined.

There are 4 characteristics that are generally used in diagnosing microfilaria:

1. The presence of absence of a sheath.
2. The presence or absence of nuclei in the tip of the tail.
3. The innerbody – can or cannot be demonstrated.
4. The size of the microfilaria.

The two methods commonly used are:

(i) Supravital Staining

Reagent

0.75% cresyl blue in saline or 1.0% methylene blue in saline.

These reagents can be used to stain live microfilariae by allowing the stain to flow under the coverslip on to a polycarbonate membrane preparation or a centrifuged preparation. The dye will stain the nuclei of the microfilariae and also provide a contrasting background to look for a sheath. It may take several minutes for the dye to penetrate the organisms and the slide should be kept in a moist chamber to prevent the preparation from drying out.

(ii) Permanent Staining

Permanent stains should show up the nuclei, including the pattern of nuclei in the tail region and stain the sheath if necessary.

The stains of choice are;

1. Hematoxylin
2. Giemsa
3. Rapid Field's

1. Hematoxylin

Delafield's hematoxylin will stain the nuclei and the sheath well and unlike Ehrlich's hematoxylin does not require heating

Reagents

- Delafield's hematoxylin (BDH)
- 1% acid alcohol
- Methanol

Method

a) Make thin films, allow to air dry then fix in methanol for five minutes

b) Stain with Delafield's hematoxylin for 20 minutes

c) "Blue" the nuclei by placing the slide in a coplin jar and allow a stream of running water to flow into the jar for 20 minutes.

d) Decolorize with 1% acid alcohol for 5-10 seconds before "blueing" in tap water again. Control this process by examination under the microscope until the nuclei are clear and distinct.

e) Allow the slide to dry before mounting in DPX.

f) The nuclei should stain blue and the sheath grey.

2. Giemsa

Reagents

- Methanol
- Giemsa stain
- Immerson Oil

Method

a) Make a thin film and allow to air dry.

b) Fix in methanol for one minute.

c) Tip off the methanol and flood the slide with Giemsa stain diluted 1:6 with buffered distilled water pH 6.8. The diluted stain must be freshly prepared each time.

d) Stain for 20 – 25 minutes.

e) Run buffered water on to the slide to float off stain and to prevent deposition of precipitate on to the film. Allow the slide to drain dry.

f) Examine the film using the oil immersion objective. Nuclei should stain red

3. Rapid Field's

Reagents

- Methanol
- Field's Stain A solution

- Field's Stain B solution
- Immersion Oil

Method

a) Make thin film and allow to air dry for five minutes.

b) Fix the smear in methanol for one minute.

c) Flood the slide with 1ml of Field's stain B (diluted 1:4 with distilled water)

d) Immediately add an equal volume of Field's stain A, mix well on slide and allow to stand for one minute.

e) Rinse well in tap water and drain dry.

f) Examine the film using the oil immersion objective and immersion oil.

g) The nuclei should stain red.

10.3 Microfilaria Worms Found in Tissue and Skin

The main species of microfilariae found in the skin and tissue are *Onchocerca volvulus* and *Mansonella streptocerca.* Microfilariae of *Onchocerca volvulus* and less often, *Mansonella streptocerca* migrate through the dermis causing itching and skin texture changes and occasionally arrive in the eye where they cause blindness. Detection of these microfilariae is from skin snips or nodule biopsies. When high numbers of microfilariae are present, they can occasionally be found in the blood and urine.

Onchocerca volvulus

Introduction

Onchocerca volvulus is mainly found in West Africa and Central and South America. Onchocerciasis, also known as river blindness, is a major public health problem, especially in West Africa despite the fact that an eradication program has been established. It is one of the world's most distressing diseases of helminth origin, often resulting in blindness. *Onchocerca volvulus* is transmitted by the species *Simulium* or black fly whose breeding habitat is by fast flowing rivers or streams, therefore there is a patchy distribution of the disease as it is specified to where water courses are. The adult worms are found in nodules or onchodermata in superficial sites, but may invade other tissues. It is estimated that there are 18 million cases worldwide with 17.5 million being found in Africa. Nigeria is the most infected region. The rate of morbidity is high in relation to those with an infection.

Life Cycle

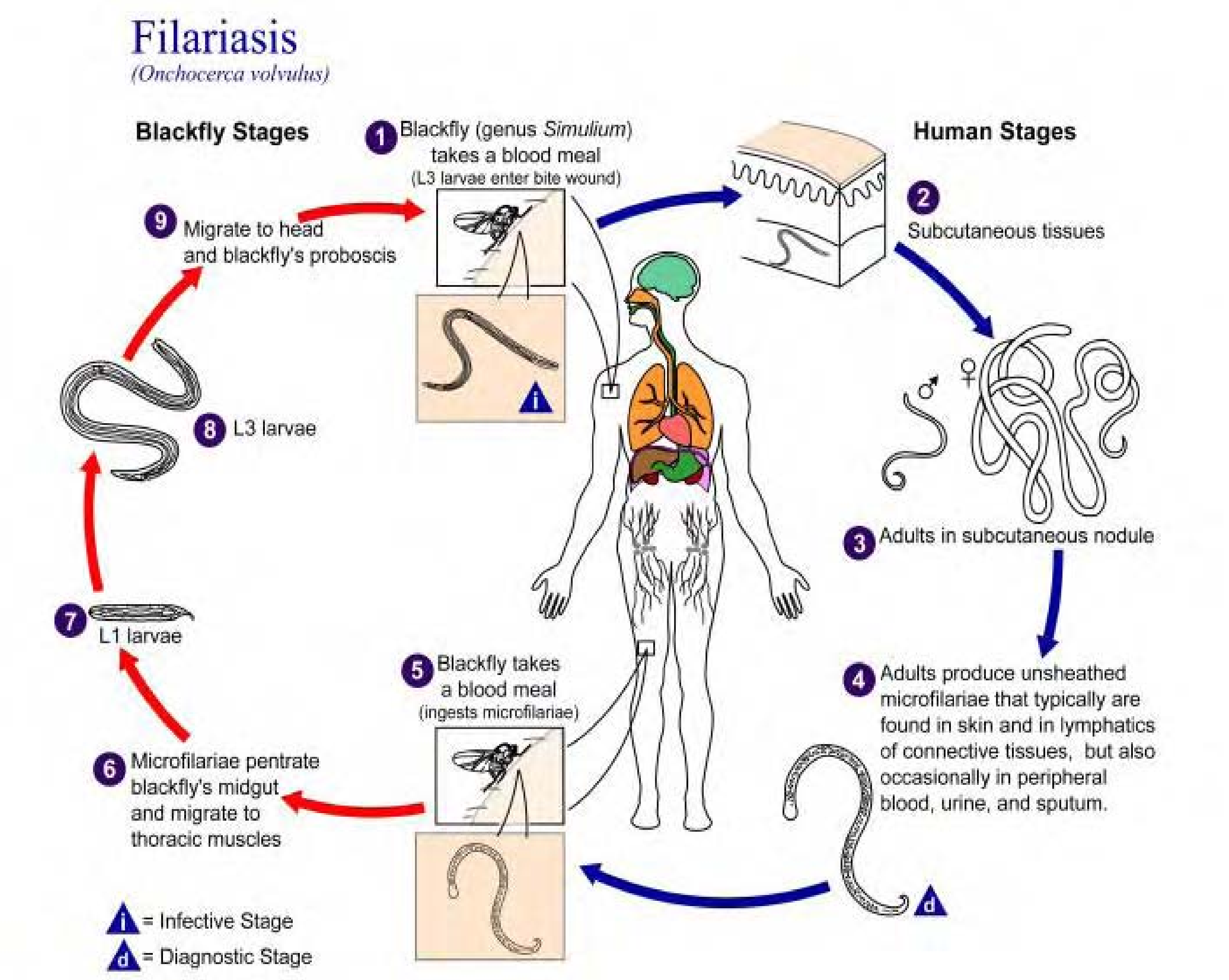
Filariasis
(Onchocerca volvulus)
Blackfly Stages
Human Stages
1 Blackfly (genus Simulium) takes a blood meal (L3 larvae enter bite wound)
2 Subcutaneous tissues
3 Adults in subcutaneous nodule
♂ ♀
4 Adults produce unsheathed microfilariae that typically are found in skin and in lymphatics of connective tissues, but also occasionally in peripheral blood, urine, and sputum.
5 Blackfly takes a blood meal (ingests microfilariae)
6 Microfilariae pentrate blackfly's midgut and migrate to thoracic muscles
7 L1 larvae
8 L3 larvae
9 Migrate to head and blackfly's proboscis
i
d
i = Infective Stage
d = Diagnostic Stage

Illustration 10-5. Diagram showing the life cycle of *Onchocerca volvulus*, a filarial nematode which causes onchocerciasis, or River blindness. It is known as river blindness due to the vector, *Simulium damnosum,* breeding in fast flowing rivers. (SOURCE: PHIL 3413 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

The life cycle is similar to *W. bancrofti,* except that the intermediate hosts are various species from the genus *Simulium* (Black flies), the most important species is *Simulium damnosum.*

The microfilariae are ingested by a Black fly during a blood meal, from where they are carried to the midgut where they penetrate the epithelium and migrate, via the hemocoele, to the indirect flight muscles. Here they undergo two molts, L_1–L_3 and develop into infective L_3 larvae which move to the mouth parts. Development is completed in 6–9 days.

When the infected fly takes another blood meal the infective larvae are once again transmitted into another host (definitive host). The microfilariae are released from the mouth parts and transmitted directly into the hosts bloodstream. Molting takes place form L_3 - L_4 within 2-5 days and the larvae then migrate widely through the body under the skin and between muscles, ligaments and tendons. The final molt to L_5 occurs at 1.5–2.5 months after transmission. Male worms are known to mature in about four months later. Female worms initiate the formation of the nodules and the males may join later. The sexually mature female worms release microfilariae which migrate out from the nodules into the skin and other tissues, most significantly into the eye.

Morphology

The whitish adult worm lies coiled within capsules in the fibrous tissue. The female can measure up to 50cm while the males are shorter measuring up to 5 cm. The microfilariae of *O. volvulus* are unsheathed and are usually found in the dermis. They measure between 221-287µm long.

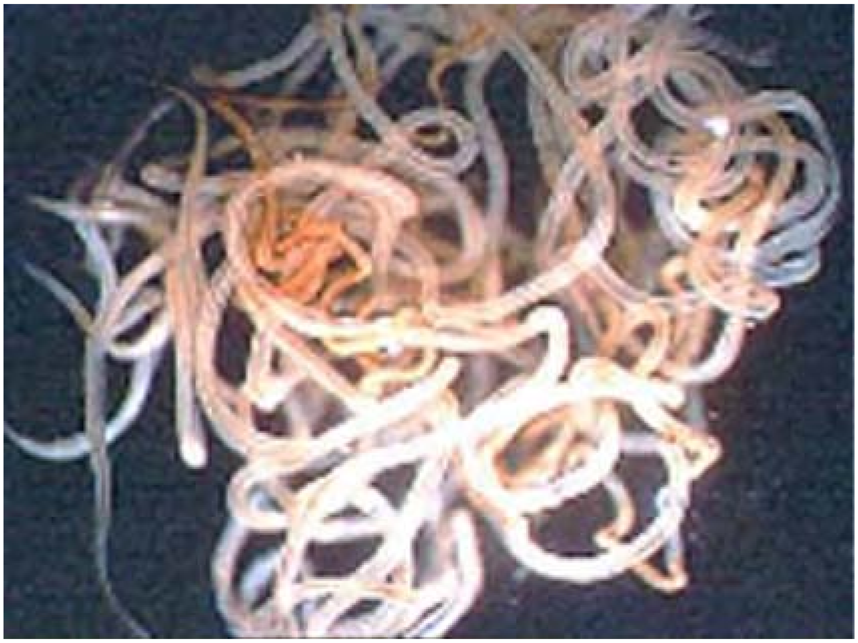

Image 10-7. *Onchocerca volvulus* microfilariae after being released by the adult female worm. They escape to the subcutaneous tissues and the eye and can be recovered with blood-free skin snips. (SOURCE: CDC)

Clinical Disease

Clinical manifestations are due to microfilariae in the epidermis.

Light infections may be asymptomatic or cause pruritis. This leads to scratching which can result in infection. Lyphadenopathy may also be a feature of early infection. After months or years, onchodermatitis results in secondary stage of thickening due to intradermal edema and pachydermis. There is a loss of elastic fibers resulting in hanging groin, hernias and elephantiasis of the scrotum. There is finally atrophy of the skin resulting in loss of elasticity. There is mottled depigmentation of the skin.

Ocular lesions are related to the intensity of the microfilariae in the skin. Ocular lesions include sclerosing keratitis, secondary glaucoma and cataract, coroidoretinitis and fluffy corneal opacities. The major complication of onchocerciasis is the development of lesions in the eye which may result in blindness or other distressing ocular diseases.

Laboratory Diagnosis

1. Analysis of Skin Snips

Small amounts of skin are collected by using a needle to raise the skin and then to slice about 1 mg of skin to a depth of 0.5µm. Snips are collected from several sites, usually the shoulders or the buttocks and sometimes the chest and calves. The snips are placed immediately in 0.5ml normal saline in a microtiter plate and left for 4 hours to allow the microfilariae to migrate out of the tissues. After four hours, the wells are examined using an inversion microscope. The microfilariae should still be moving and can be identified from the table below. The microfilariae can also be collected by filtration or centrifugation and the deposit containing microfilariae can be stained with Giemsa at pH 6.8.

2. Analysis of Biopsies

Biopsies of tissue nodules can be dabbed on to a slide to produce impression smears and then stained with Giemsa stain at pH 6.8 for the presence of microfilariae.

Recent advances in diagnostic methods includes and ELISA-based antibody detection assay which utilizes a cocktail of recombinant antigens. The advantages of using this test is that it is highly sensitive (almost 100% in onchocerciasis foci). It is also highly specific (100%), it also uses finger prick blood. Therefore, reducing the painful procedure of gaining a skin snip.

The disadvantages is that it requires advanced ELISA apparatus and reagents and cannot distinguish between past and present infections due to it detecting antibodies which stay present in the body for a long time after the infection. Another modern detection method is for Parasite DNA detection, which is based on the amplification of specific DNA sequences form microfilariae using molecular biology technology. The advantages of this technique is its exquisite sensitivity and detects active infections only. The disadvantages are that it requires specialized equipment and expensive reagents. Also it still requires a skin snip but a urine assay is a possibility for the future.

Thick microfilaria. Does not have a sheath. Head often spatulate. Nuclei do not extend to tip of tail. Found only in skin.

Mansonella streptocerca

Introduction

Microfilaria of *M. streptocerca* were first reported in the skin of a West African patient in 1922. These microfilaria are primarily found in the skin but have been also reported in the blood. This species occurs in Ghana, Cameroon and Zaire. The adults are poorly known, and occur in the cutaneous tissue of man and chimpanzee.

The microfilariae do not exhibit periodicity with the intermediate hosts being *Culicoides grahamii* and possibly other *Culicoides* species.

Life Cycle

The life cycle is the same as that of the blood *Mansonella* species.

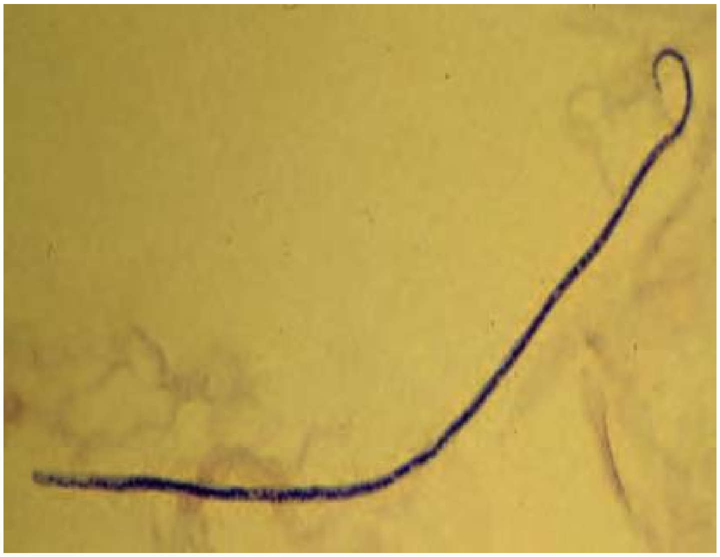

Image 10-8. Microfilaria of *Mansonella streptocerca.* From a skin snip, after a concentration procedure and hematoxylin stained. The microfilaria is typically unsheathed, and its body has a straight attitude. The tail is typically coiled ("shepherd's crook"), and nuclei extend to the end of the tail, as a single-cell row. (SOURCE: CDC)

Clinical Disease

Infection is characterized by pruritic dermatitis and hypopigmented macules.

Laboratory Diagnosis

Mansonella streptocerca can be diagnosed by demonstrating the microfilaria in a skin snip. Snips are collected from several sites, usually the shoulders and buttocks and sometimes the chest and calves. The snips are placed immediately in 0.5ml of 0.9% sodium chloride in a microtiter plate and left for four hours to allow the microfilaria to migrate out of the tissues. After four hours, the wells are examined using an inversion microscope. The microfilaria should still be moving and can be identified by staining with Giemsa at pH 6.8

Small, thin, microfilaria. Does not have a sheath. Nuclei extend to end of tail. Tail is hooked; its tip is rounded or forked. Found only skin.

Species	***Onchocerca volvulus***	***Mansonella streptocerca***
Distribution	Tropical Africa, Central and South America	West Africa
Vector	*Simulium* spp.	*Culicoides* spp.
Adult location	Subcutaneous nodules	Cutaneous connective tissue
Microfilariae location	Skin	Skin
Microfilariae size	280 - 330µm	180 - 240µm
Morphology	Broad spatulate head; No sheath, pointed tail	Curled tail; No sheath
Tail nuclei	Tail free from nuclei	Nuclei extend to tail tip

Table 10-4. Differential features of *Onchocerca volvulus* and *Mansonella streptocerca*. (SOURCE: CDC)

Dracunculus medinensis

Introduction

Dracunculus medinensis is a non-filarial parasite as it only has one uterus whereas filaria have two. It is usually associated with places where there is a lack of clean drinking water e.g. step wells in India, covered cisterns in Iran, and ponds in Ghana. The life cycle usually involves copepod intermediate host. They are parasitic in the connective tissue or coelom of vertebrates. The disease associated with this parasite is known as Dracunculiasis.

Life Cycle

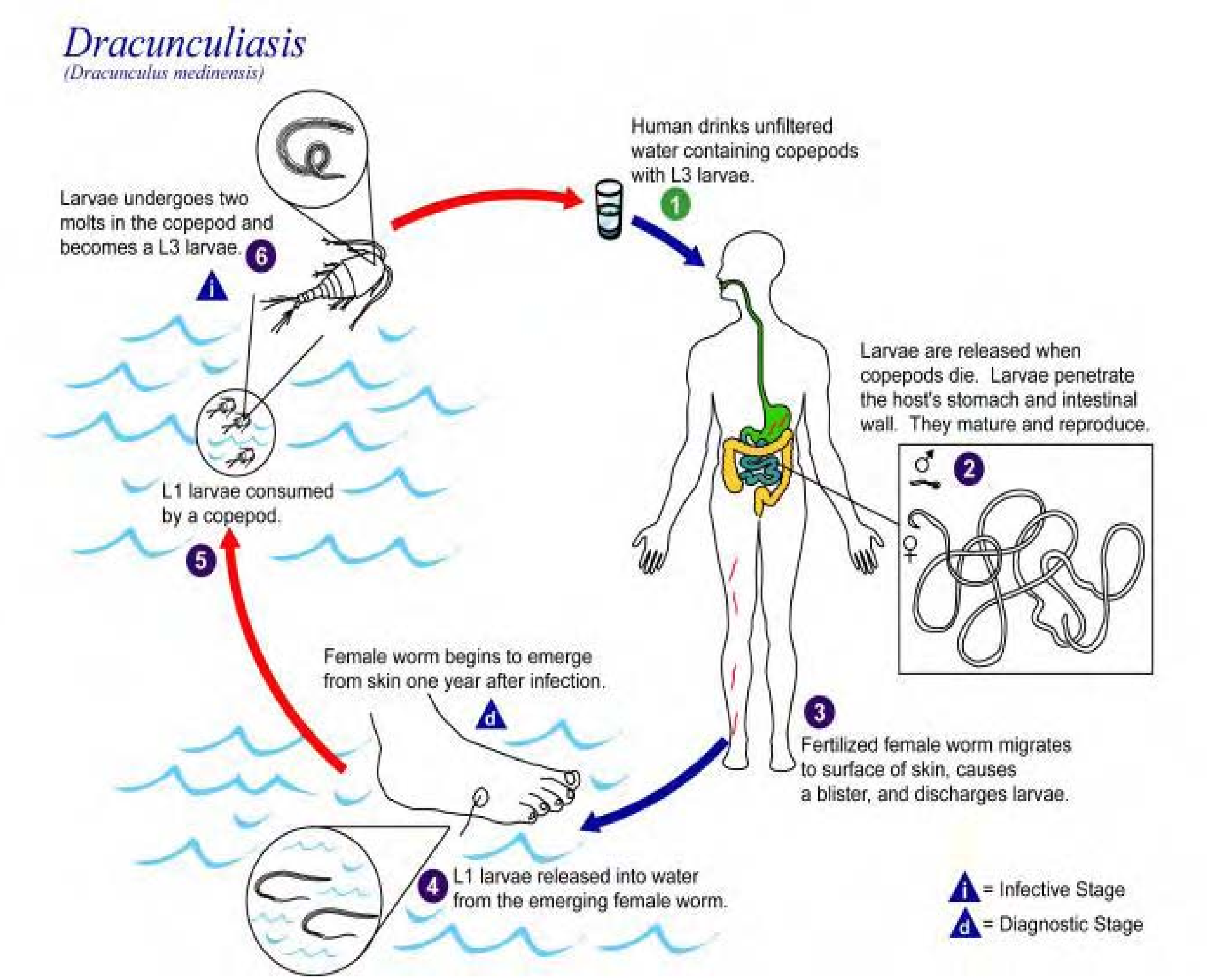
Dracunculiasis
(Dracunculus medinensis)
Human drinks unfiltered water containing copepods with L3 larvae.
1
Larvae are released when copepods die. Larvae penetrate the host's stomach and intestinal wall. They mature and reproduce.
2
♂
♀
3
Fertilized female worm migrates to surface of skin, causes a blister, and discharges larvae.
Female worm begins to emerge from skin one year after infection.
d
4
L1 larvae released into water from the emerging female worm.
5
L1 larvae consumed by a copepod.
6
Larvae undergoes two molts in the copepod and becomes a L3 larvae.
i
i = Infective Stage
d = Diagnostic Stage

Illustration 10-6. Diagram showing the life cycle of *Dracunculus medinensis*, Guinea worm. (SOURCE: PHIL 3391 - CDC/Alexander J. da Silva, PhD/Melanie Moser)

Mature female worms which are gravid with microfilariae migrate to the superficial layers of skin of humans, especially those regions which are most likely to come in contact with water, such as the ankle, foot, arms and shoulders. Here the worms secrete a substance (substance is unknown) which causes a blister to rise over its anterior end where it has pierced the lower layers. The blister eventually forms into an ulcer which on contact with water, the uterus is projected out of the ulcer cavity, and a cloud of milky white secretion, containing hundred of active larvae, is released. Once out of the water again the uterus dries and shrivels preventing the release of further larvae.

If the microfilariae are ingested by an appropriate species of *Cyclops*, they break though the soft mid-intestine wall and come to lie in the body cavity. The larvae undergo two molts and become infective in approximately three weeks. Humans become infected by accidentally ingesting through drinking water the infective *Cyclops.* Upon ingestion the larvae are activated to penetrate through the gut wall, and migrate through the tissues, molting twice and finally becoming lodges in the viscera or subcutaneous tissues. Maturation of the worms is slow taking about one year to reach sexual maturity before the females are ready to migrate to the skin to release their larvae.

Morphology

The adult female worm measure up to one meter in length whereas the male measures about 2cm.

Clinical Disease

After ingestion of the Cyclops, there is no specific pathology associate with the mucosal penetration and larval maturation in the deep connective tissues. Erythema and tenderness can be associated with blister formation. The patient can also exhibit vomiting, diarrhea, asthmatic attacks. Symptoms usually subside when the lesion erupts. If the worm is removed, healing usually occurs without any problems. If the worm is damaged or broken during removal, there may be intense inflammatory reaction with possible cellulitis along the worms migratory tract. This can result in arthritis and synovitis.

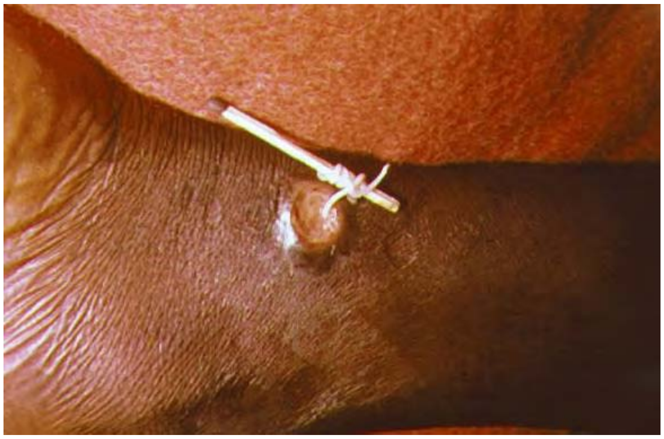

Figure 10-9. Female *Dracunculus medinensis* worm (Guinea worm) emerging out of a typical ulcer. Adult worms emerge from these ulcers on contact with water to release their microfilaria. The most effective method for removing these worms are to slowly wind them around a piece of stick, being careful not to break the worm in two. (SOURCE: PHIL 1342 – CDC)

Laboratory Diagnosis

The best remedy for removing the adult worm is a slow process of daily gently rolling the worm around a small stick and slowly pulling it out of the skin. With this method you must be careful not to pull apart the worm as it will recoil back into the skin and cause secondary infections.

This parasite is currently being approached with a strict control program. The program includes stopping people from drinking infected water, putting muslin over water collection jars, educating the communities about the parasite, and adding temphos to the water to kill it off.

Chapter 11.

Babesia, Trypanosomes, and Leishmania

Babesia species

Introduction

Babesia species are protozoan parasites of domestic and wild animals. They belong to the subclass Piroplasmia and are commonly referred to as ‘piroplasms’ due to the pear-like shaped merozoites which live as small intra-erythrocytic parasites. They commonly infect mammals, particularly cattle, sheep, goats, horses, pigs, dogs and cats and occasionally man.

Species infective to humans are the cattle form *Babesia bovis* which can often be fatal and *Babesia microti* which is less pathogenic*.* Until recently *B. microti* was confined to the United States but is now becoming a significant tick-borne disease of man in other temperate climates as well.

Human infections have been found in Europe and North America. The infection is known as Babesiosis, it can also be described as ‘redwater fever’ or ‘tick fever’ and has a major impact on the livestock industries in many countries.

What is unusual with this parasite’s life cycle is the development in the vector. They use the one-host ticks, belonging to the genus *Boophilus.* The parasites are passed to the eggs and hence to the larval stages which can thus become infective after the adult tick dies. This process is known as transovarial transmission.

The ticks which are known to carry the parasite of *Babesia bovis* are *Boophilus microplus, B. decoloratus* and *B. annulatus.*

Life Cycle

Human babesiosis is a zoonosis, acquired by tick bite when individuals accidentally interact with the natural life cycle of the parasite.

The life cycle is best known for *Babesia canis* which infects dogs, but it is generally the same in each of the one-tick hosts and the definitive host.

Babesia bovis

When the tick bites, **sporozoites** are injected into the blood stream and penetrate the erythrocytes. In contrast to the malaria life cycle, there is no tissue stage for *Babesia bovis. Babesia* multiplies in the red cell by budding in contrast to schizogony in *Plasmodium* species. The red cell ruptures and daughter parasites invade new erythrocytes for further asexual multiplication. Some of the **sporozoites** injected by the tick vector follow a different path of intra-erythrocytic development, growing slowly and "folding" to form accordion-like structures which are destined to undergo further development in the tick vector. Within the intestine of the tick, the accordion-like stage eventually fuses with another, to form a **zygote**. Further development outside the intestine occurs in a variety of tissues, the salivary glands and ovaries being especially important for transmission. **Sporozoites** in tick salivary glands are injected into the mammalian host at the next blood meal. Transovarial transmission of *Babesia bovis* also takes place so that newly hatched onto adult stages can then take place.

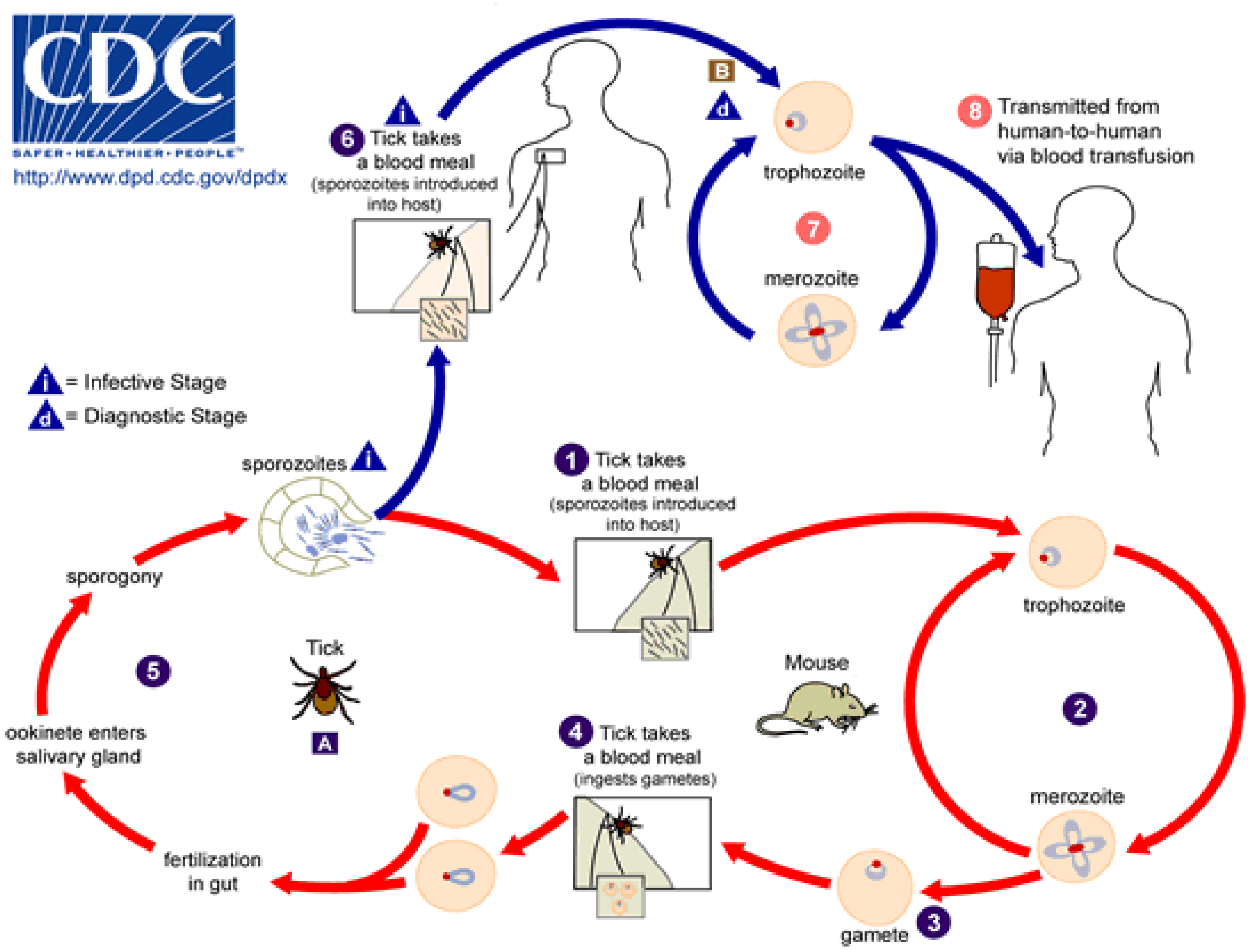
CDC
SAFER·HEALTHIER·PEOPLE™
http://www.dpd.cdc.gov/dpdx
i
6 Tick takes a blood meal (sporozoites introduced into host)
B
d
trophozoite
7
merozoite
8 Transmitted from human-to-human via blood transfusion
i = Infective Stage
d = Diagnostic Stage
sporozoites i
1 Tick takes a blood meal (sporozoites introduced into host)
trophozoite
sporogony
Tick
5
A
Mouse
2
merozoite
ookinete enters salivary gland
4 Tick takes a blood meal (ingests gametes)
fertilization in gut
gamete
3

Illustration 11-1. Generalized life cycle of the parasite *Babesia*, which causes the disease babesiosis in man. (SOURCE: CDC/DPDx)

Babesia microti

In the small mammal host of *Babesia microti*, **sporozoites** from the tick vector first enter lymphocytes and undergo **merogony**, the daughter parasites of which then enter erythrocytes.

Babesia microti do not undergo transovarial transmission, but once a larva has become infected from a mammalian host they are able to pass on the infection transstadially to the nymph.

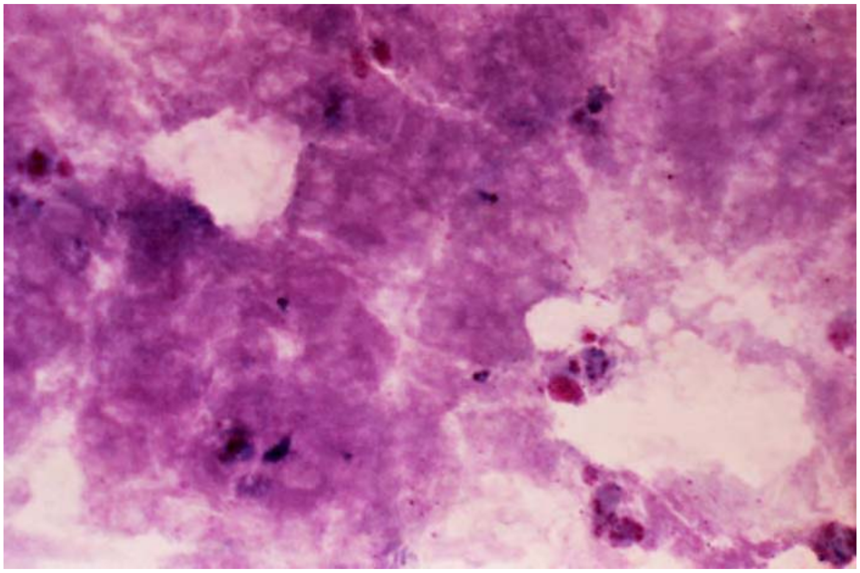

Image 11-1. Diagnosis of *Babesia* species depends on the observation of the intraerythrocytic organisms in blood smears. Pear shaped microorganisms (2-5µm) and tetrads are the diagnostic shape of the parasite. (SOURCE: PHIL 3899 – CDC/Dr. George Healy)

Clinical Disease

Babesia bovis - Patients who are particularly at risk are those who have had a splenectomy. The patient may feel vaguely unwell at first but by the time the diagnosis has been made, is usually very ill, with fever, prostration, jaundice, anemia and hemoglobulinuria. Nausea, vomiting and diarrhea have also been recorded. Unlike malaria symptoms, the symptoms of babesiosis do not exhibit periodicity.

Babesia microti - Most human infections are subclinical. Where clinical illness develops, the incubation period is 1 to 3 weeks, occasionally up to six weeks. The illness usually begins gradually, with anorexia and fatigue, plus fever (without periodicity), sweating, rigors and generalized myalgia. Physical examination may reveal only fever, but may also show mild splenomegaly and occasionally mild hepatomegaly.

Laboratory Diagnosis

Definitive diagnosis depends upon finding parasites on blood film examination which can be detected 2 to 4 weeks after a tick bite. Hamster inoculation and serology have also been used for diagnosis.

1. Microscopic Examination

Babesia bovis

Babesia bovis are pear shaped, oval or round and may exist in pyriform pairs. There may be 1 to 8 parasites per red cell. Ring forms can be confused with malaria parasites, especially *Plasmodium falciparum*. However, in contrast to *Plasmodium* species, *Babesia* do not form pigment, do not cause alterations in red cell morphology and finally do not exhibit the Maurer's clefts of *Plasmodium falciparum*, the Schüffner's dots of *Plasmodium vivax*, or the James's dots of *Plasmodium ovale.*

The "Maltese cross form" is unique to *Babesia* but in its absence it may be very difficult to distinguish young ring forms of *Plasmodium falciparum*, from *Babesia*. The absence of pigment cannot be relied upon, as young rings of *Plasmodium* do not exhibit pigment. *Babesia*, are smaller than malaria parasites, and in some of the larger rings there is white vacuole containing erythrocyte stroma, instead of the pink vacuole seen in malaria. *Babesia* parasites do not form schizonts.

Babesia microti

Ring, rod shaped, pyriform, amoeboid, and "Maltese cross" forms are seen. In heavy infections different stages may be noted in the same red cell. Intra-erythrocytic stages measure approximately 2 by 1.5µm. In very high parasitaemias, extracellular merozoites are found singly or as a syncytial structure. Peak parasitaemia varies between less than 1% to approximately 10%.

2. Serodiagnosis

The Indirect Fluorescent Antibody Test (IFAT) is available for both *B. bovis* and for *B. microti* and is the most useful serological test for early diagnosis.

3. Animal Inoculation

This is not routinely used for diagnosis but *B. microti* grows well in hamsters and can serve as a confirmatory test.

Toxoplasma gondii

Introduction

Toxoplasma gondii, the causative organism of toxoplasmosis, was first observed in 1927 in the gondi, a North African rodent. The first human case of toxoplasmosis was also reported that year. The organism is a coccidian protozoa belonging to the sub-phylum Apicomplexa and has a world wide distribution occurring in all warm-blooded animals.

Cats are the definitive hosts and they become infected by ingesting oocysts or cysts in tissues of paratenic hosts, such as mice, or transplacentally. Man becomes infected either by direct ingestion of oocysts from a cat or by eating raw or undercooked meat. Those who handle raw meat are particularly at risk. Infection can be transmitted transplacentally.

Life Cycle

The development of the enteroepithelial (sexual) cycle in a cats intestine is brought about by the ingestion of sporulated oocysts of a mouse with cysts. The pre-patent period up to the shedding of the oocysts varies with the stage of *T. gondii* ingested, for example only 3–10 days if the cat has ingested a mouse containing cysts, but about 19–20 days or longer after direct infection with oocysts or ingestion of a mouse containing only tachyzoites. Women most at risk of delivering an infected infant are those who acquire the infection just prior to gestation.

Humans can acquire infection by:

- Accidental ingestion of the oocyst shed in the cats feces
- Ingestion of the tachyzoite in infected milk or transplacentally
- Ingestion of the tissue cyst in undercooked or raw meat.
- Transplant of an infected organ in a seronegative recipient

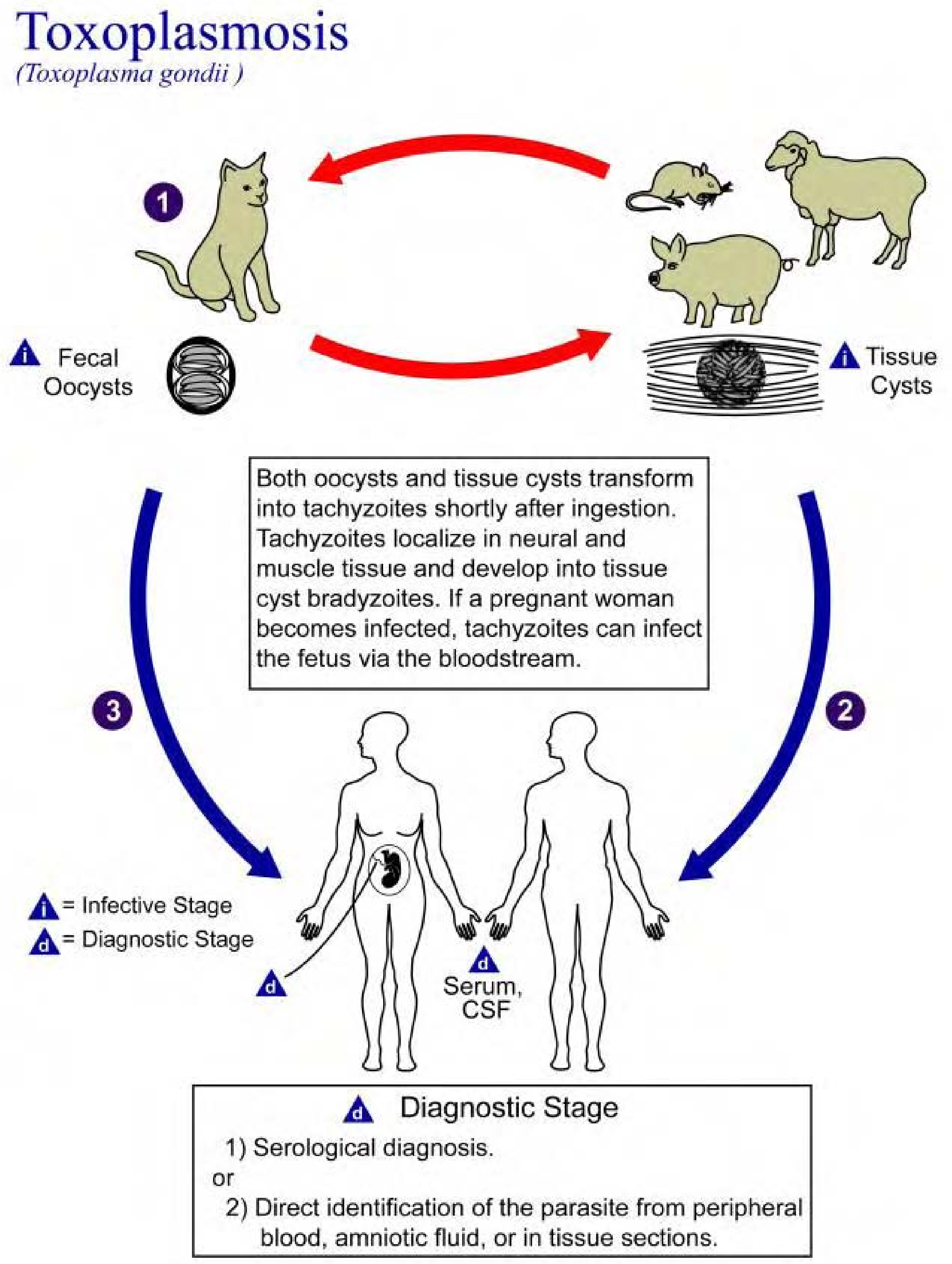
Toxoplasmosis
(*Toxoplasma gondii*)
1
Fecal Oocysts
Tissue Cysts
3
2
= Infective Stage
= Diagnostic Stage
Serum, CSF

Both oocysts and tissue cysts transform into tachyzoites shortly after ingestion. Tachyzoites localize in neural and muscle tissue and develop into tissue cyst bradyzoites. If a pregnant woman becomes infected, tachyzoites can infect the fetus via the bloodstream.

Diagnostic Stage

1) Serological diagnosis.

or

2) Direct identification of the parasite from peripheral blood, amniotic fluid, or in tissue sections.

Illustration 11-2. Life cycle of *Toxoplasma gondii*, causes toxoplasmosis in man. (SOURCE: PHIL 3421 – CDC/Alexander J. da Silva, PhD/Melanie Moser)

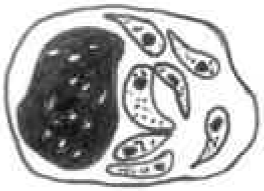

Illustration 11-3. Diagrammatic illustration of a *Toxoplasma gondii* trophozoite in a macrophage of a vertebrate. (SOURCE: Unknown)

Clinical Disease

Serological evidence has shown that approximately one third of the world's population has *Toxoplasma* antibodies. This suggests that the majority of infections are benign with most people exhibiting few or no symptoms, but fever and swelling may be seen. However, *Toxoplasma gondii* can cause severe illness in congenital infections, acquired infections and in immunocompromised patients. This may lead to ocular toxoplasmosis and ultimately to fatal CNS disorders such as encephalitis.

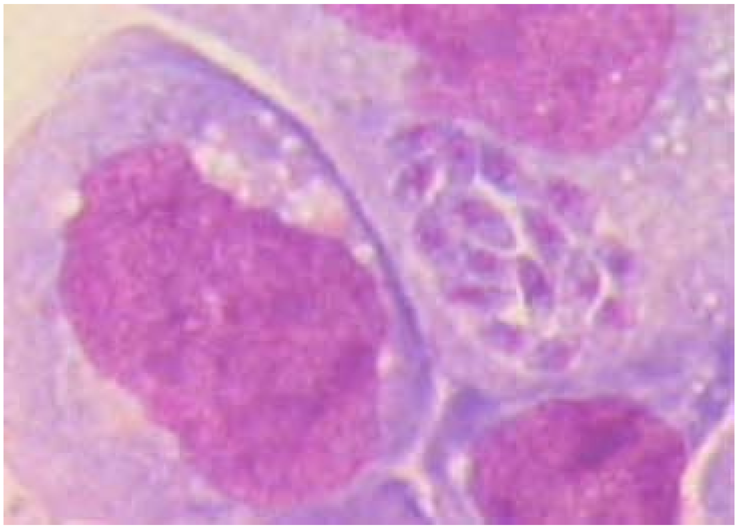

Image 11-2. *Toxoplasma gondii* tissue cyst containing 8–20 parasites (Giemsa stain) (SOURCE: CDC)

Congenital Toxoplasmosis

This occurs approximately in 1 per 1000 pregnancies. It can cause severe damage to and even death of the fetus. Proliferation of tachyzoites leads to intracellular calcification, choroidoretinitis, hydrocephaly, psychomotor disturbances and convulsions. A small, proportion of babies who are asymptomatic at birth develop retinochoroiditis or mental retardation as children or young adults. When a mother is first exposed to the parasite in later pregnancy the infant is likely to be less severely damaged or asymptomatic.

Acquired Infections

Infections with *T. gondii* are often mild with flu-like symptoms thus they often go unnoticed. However lymphadenopathy is the most easily recognized symptom and it can be accompanied by fever, headache and myalgia. Toxoplasmosis may also produce infectious mononucleosis like symptoms. Ocular toxoplasmosis is also a common manifest however it is not yet proven whether this is due to congenital or acquired infections. Other manifestations of *Toxoplasma* infections are meningoencephalitis, hepatitis, pneumonitis and myocarditis.

Immunocompromised Patients

Toxoplasmosis has been shown to occur as an opportunistic pathogen in immunocompromised patients and can cause severe complications. Toxoplasmosis in immunocompromised patients almost always arises from a reactivation of latent infections. Conditions which can predispose to toxoplasmosis are malignancies, organ transplants, leukemias and patients with acquired immune deficiency syndrome (AIDS). In immunocompromised patients, the central nervous system is primarily involved with diffuse encephalopathy, meningoencephalitis or cerebral mass lesions. *Toxoplasma* encephalitis has been reported as a life-threatening among patients with AIDS.

Laboratory Diagnosis

1. Serological Techniques

The detection of toxoplasma specific antibodies is most commonly used in clinical laboratories. Specific IgG antibodies typically persist for life whereas specific IgM antibodies begin to decline after several months. Most laboratories carry out preliminary tests for IgG antibodies and more definitive tests including IgM and IgA are carried out in reference laboratories. The Sabin-Feldman Dye Test is the benchmark for detecting the presence of specific antibodies. It measures the total amount of specific antibody in a serum which is capable of participating in antibody-mediated killing of tachyzoites by complement. This test involves the use of live tachyzoites which are derived from infected mice or rats. Because of the use of live organisms, this test is not recommended in the use of routine laboratories and is thus only employed in reference centers.

2. Isolation Techniques

Culture of parasites in animals is the best overall method but it can take up to six weeks before the result is available and is thus a disadvantage. Tissue culture is more rapid taking three or four days to obtain a result, but is not as sensitive.

3. Antigen Detection

The direct detection of very small amounts of specific nucleic acid has been made possible by the introduction in 1985 of the polymerase chain reaction (PCR). This technique results in the amplification of a specific fragment of DNA from within the parasite genome which is detected by ethidium bromide staining, following gel electrophoresis. PCR is so sensitive it should detect *Toxoplasma* DNA in one cyst. However this may indicate a latent infection rather than an active infection. However its sensitivity may create problems since it will detect very small amounts of DNA from latent as well as active infections and it does not differentiate between cyst and tachyzoite DNA. Thus samples like blood, CSF, urine and amniotic fluid should be used as they do not contain the latent stages. PCR shows great promise but as yet is still labor intensive and expensive for routine use in the laboratory.

Trypanosoma species

Introduction

Trypanosomes are hemoflagellates and three species of the genus *Trypanosoma* are responsible for disease in humans such as sleeping sickness.

Trypanosomes occur in the blood of the majority of vertebrate animals. The life cycle involves intermediate host, which usually is an insect. Many species of trypanosomes can live in harmony with their hosts producing no pathogenic effect, but the best known species are those that are pathogenic to their definitive hosts. The disease in caused by the pathogenic types is called trypanosomiasis.

Salivarian Trypanosomes

Trypanosoma brucei rhodesiense and *Trypanosoma brucei gambiense* - The metacyclic trypanosomes are found in the proboscis of the insect vector - infection is therefore inoculative. The above are the causative agents of African trypanosomiasis. It is a zoonotic species in that it multiplies in the blood of a range of many mammals including man.

Trypanosoma brucei rhodesiense causes acute sleeping sickness in East Africa, while *T. b. gambiense* causes chronic sleeping sickness in West Africa.

These are known as salivarian trypanosomes as they complete their development in the salivary system (anterior portion of the vector). Transmission takes place by inoculation of the metacyclic stage.

Stercorarian Trypanosomes

Trypanosoma cruzi - The metacyclic trypanosomes occupy a posterior position in the gut of the insect vector and are passed out in the feces - infection is therefore contaminative. This is the causative agent of American Trypanosomiasis.

These trypanosomes are known as stercocarian as they complete their development in the posterior region of the vector, so that the infective forms appear in the insect's feces. Hosts are infected by the contaminative route.

African Trypanosomiasis

Life Cycle

Transmission from one vertebrate to another is carried out by blood-sucking invertebrates, usually an insect. The vector for African Trypanosomiasis is the Tsetse fly, *Glossina* spp*.* which cause the diseases *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense.*

Metacyclic (infective) trypomastigotes are inoculated through the skin when a tsetse fly takes a blood meal. The parasites develop into long slender trypomastigotes which multiply at the site of inoculation where ulceration occurs. The trypanosomes continue to develop and then may invade the lymphatic tissues, the heart, various organs and in later stages, the central nervous system. Trypomastigotes are taken up by the tsetse fly (male and female) during a blood meal. The parasites develop in the midgut of the fly where they multiply. 2-3 weeks later the trypomastigotes move to the salivary glands transforming from epimastigotes into metacyclic (infective) trypomastigotes. The tsetse fly remains infective for life i.e. about three months.

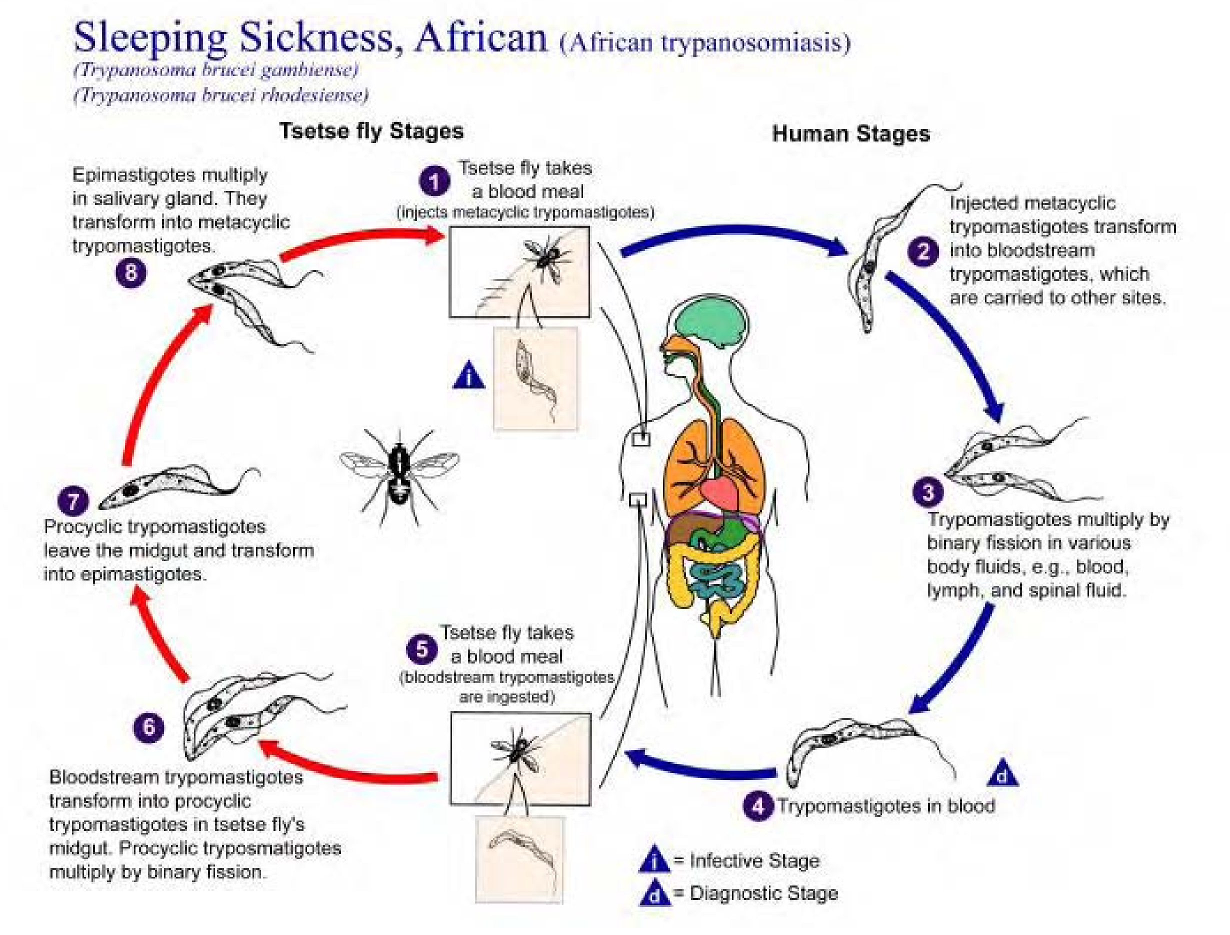
Sleeping Sickness, African (African trypanosomiasis)
(Trypanosoma brucei gambiense)
(Trypanosoma brucei rhodesiense)
Tsetse fly Stages
Human Stages
1 Tsetse fly takes a blood meal (injects metacyclic trypomastigotes)
2 Injected metacyclic trypomastigotes transform into bloodstream trypomastigotes, which are carried to other sites.
3 Trypomastigotes multiply by binary fission in various body fluids, e.g., blood, lymph, and spinal fluid.
4 Trypomastigotes in blood
d
5 Tsetse fly takes a blood meal (bloodstream trypomastigotes are ingested)
6 Bloodstream trypomastigotes transform into procyclic trypomastigotes in tsetse fly's midgut. Procyclic tryposmatigotes multiply by binary fission.
7 Procyclic trypomastigotes leave the midgut and transform into epimastigotes.
8 Epimastigotes multiply in salivary gland. They transform into metacyclic trypomastigotes.
i
i = Infective Stage
d = Diagnostic Stage

Illustration 11-4. Generalized life cycle of the *Trypanosoma* spp. which causes African Trypanosomiasis. (SOURCE: PHIL 3418 – CDC/Alexander J. da Silva, PhD/Melanie Moser)

The mode of transmission mentioned above, metacyclic transmission, requires to be separated from mechanical transmission, a process in which trypanosomes survive, for a short time, on and about mouth parts of an insect and are inoculated into a new host when the vector bites again, without undergoing any developmental cycle.

Metacyclic transmission requires a lapse of time to allow the trypanosomes to reach an infective stage by a particular developmental sequence in the vector, usually a period of several days.

Morphology

The parasite is an elongated cell with single nucleus which usually lies near the centre of the cell. Each cell bears a single flagellum which appears to arise from a small granule - the kinetoplast. The kinetoplast is a specialized part of the mitochondria and contains DNA. The length and position of the trypanosome’s flagellum is variable. In trypanosomes from the blood of a host the flagellum originates near the posterior end of the cell and passes forward over the cell surface, its sheath is expanded and forms a wavy flange called an undulating membrane.

Development is characterized by the occurrence of three types of blood forms (polymorphic), these are:

1) **Slender forms**: long and thin, about 29µm long, free flagellum.

2) **Stumpy forms**: thick and short, average length 18µm, typically no free flagellum, but a short one may be present.

3) **Intermediate forms**: about 23µm long with a moderately thick body and a free flagellum of medium length.

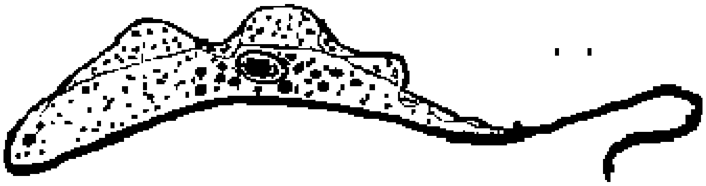

Illustration 11-5. Diagrammatic illustration of a typical trypanosome. (SOURCE: Unknown)

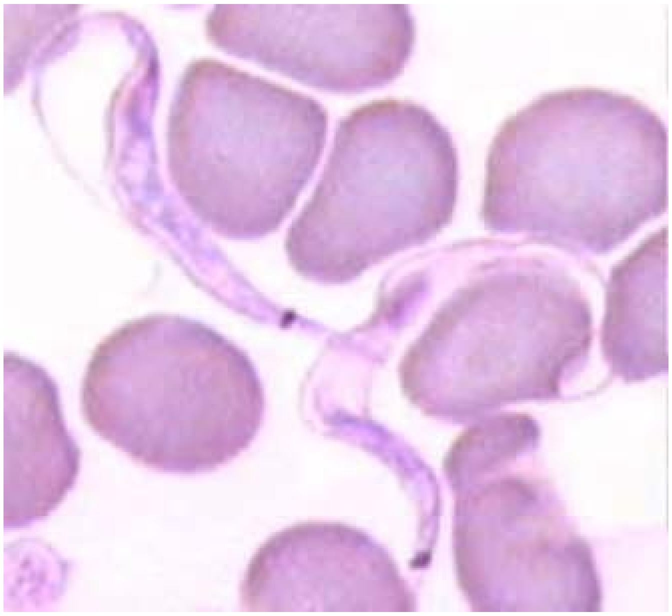

Image 11-3. *Trypanosoma brucei gambiense* and *T. b. rhodesiense:* two forms of trypomastigote can be seen in peripheral blood: one is long slender, 30 µm in length, and is capable of multiplying in the host, the other is stumpy, not dividing, 18 µm in length. (SOURCE: CDC)

Clinical Disease

The early stages of African trypanosomiasis may be asymptomatic and there is a low grade parasitiaemia. This period may last for several weeks to several months. The disease may terminate untreated at this stage or go on to invade the lymph glands. Invasion of the lymph glands is usually accompanied by a high irregular fever with shivering, sweating and an increased pulse rate. The lymph glands near the bite often become swollen, in *T. b.* gambiense the glands at the back of the neck and *T. b.* rhodesiense usually the glands under the jaw are affected (Winterbottom's sign). As the disease progresses, edema of the eyelids, face and sleeplessness are features along with increasing lethargy and listlessness.

Trypanosomes may invade the central nervous system giving symptoms of meningoencephalitis, confusion, apathy, excessive sleeping and incontinence. At this stage, the cerebrospinal fluid (CSF) usually contains mononuclear cells and a few trypanosomes may be detected. If untreated, character changes, mental deterioration and coma develops, finally resulting in death. Such signs are more commonly seen with *gambiense* than in *rhodesiense* in which patients often die before these symptoms develop fully.

Laboratory Diagnosis of African trypanosomiasis

Laboratory diagnosis of African Trypanosomiasis is by:

- Examination of blood for the parasites
- Examination of aspirates from enlarged lymph glands for the parasites
- Examination of the CSF for the parasite

- Detection of trypanosomal antibodies in the serum

1. Examination of Blood

a) Thick and Thin Blood Films

Thick and thin blood films are made and stained with Fields stain and examined as for malaria parasites

b) Triple Centrifugation Technique

This method is carried out as follows:

i. 5 to 10ml of citrated blood is centrifuged at 2000rpm for 5 minutes to pack the red blood cells.
ii. The plasma and white cell layer are removed by a Pasteur pipette and transferred to a clean centrifuge tube.
iii. This is centrifuged for a short time in order to deposit any red blood cells carried over.
iv. The supernatant fluid is removed by pipette to a clean tube.
v. This is centrifuged at 5000rpm for 10 minutes.
vi. The supernatant fluid is removed with a pipette and discarded.
vii. The deposit is examined microscopically for trypanosomes.

c) Miniature anion-exchange centrifugation technique.

Heparinized blood is passed through an anion exchange column. As the blood travels down the column the red cells are adsorbed while the less strongly charged trypanosomes are washed through with saline. The eluate is centrifuged and examined microscopically for motile trypanosomes.

d) Buffy Coat Examination

Trypanosomes are centrifuged in a microhaematoctit tube for five minutes. Parasites can be seen microscopically at the junction of the packed red cells and plasma.

2. Examination of Lymph Gland Aspirates

The aspirate can be examined microscopically by making a wet preparation, or if there is not much material, it can be allowed to dry on a slide and then stained with either rapid Field's stain or with Giemsa and examined microscopically.

3. Examination of CSF

In the late stages of African trypanosomiasis, trypanosomes may be found in the CSF together with IgM - containing morula (Mott) cells, lymphocytes and other mononuclear cells. Once the CSF has been collected it **must be examined as soon as possible.** The parasites are unable to survive for more than 15-20 minutes in CSF once it has been removed. The parasites become inactive, are rapidly lysed and will not therefore be detected. The CSF should be examined wet and spun down in a sterile tube and a film made from the deposit. The film is then stained with rapid Field's or Giemsa and examined microscopically.

NB. It is impossible to distinguish between *T. b. gambiense* from *T. b. rhodesiense* on a stained film as the two subspecies which infect man are identical.

American Trypanosomiasis

Trypanosoma cruzi occurs throughout South and Central America, especially in Brazil, Argentina and Mexico causing the disease known as Chagas' disease. It is estimated that over 24 million people are infected with this species. It is a zoonotic parasite with over 150 species of wild animals known to harbor the parasites, for example opossums, dogs, rates, pigs and cats.

It is transmitted to man by brightly colored bugs belonging to the Reduviidae family, subfamily Triatominae. All stages of these bugs are known to become infected.

The bugs live in the crack of the walls and vegetal roofs of the poorly maintained houses, coming out at night to feed on the exposed parts of the host's body.

Image 11-4. Insect vector of *Trypanosoma cruzi*, belongs to colorful insect Triatominae, also known as the kissing bug. (SOURCE: PHIL 2538 - CDC/World Health Organization)

Life Cycle

Metacyclic trypomastigotes are deposited in feces on the skin as the **triatomine bug** (reduviid bug) feeds. The bug usually bites round the edges of the mouth and eyes. The trypomastigotes are either rubbed into the skin by scratching the irritated area or penetrate the conjunctiva or membranes of the nose and mouth. Trypomastigotes become amastigotes in localized reticuloendothelial cells and multiply. The amastigotes develop into trypomastigotes which are released into the blood when the cell ruptures. No multiplication of the parasite takes place in the blood in its trypomastigote stage. The trypomastigotes reach tissue cells especially heart muscle, nerves, skeletal muscle and smooth muscle of the gastrointestinal system by way of the blood and lymphatic system. The trypomastigotes become amastigotes and multiply forming pseudocysts. Within the pseudocyst some amastigotes become elongated and develop first into epimastigotes and then trypomastigotes. When the cell ruptures the trypomastigotes are released into the blood and continue to circulate while others invade further tissue cells. The life cycle completes when a triatomine bug vector ingests circulating trypomastigotes. In the vector the trypomastigotes transform and develop into epimastigotes, multiply by binary fission in the gut of the bug. After about 10-15 days, metacyclic trypomastigotes are formed and can be found in the hindgut of the bug.

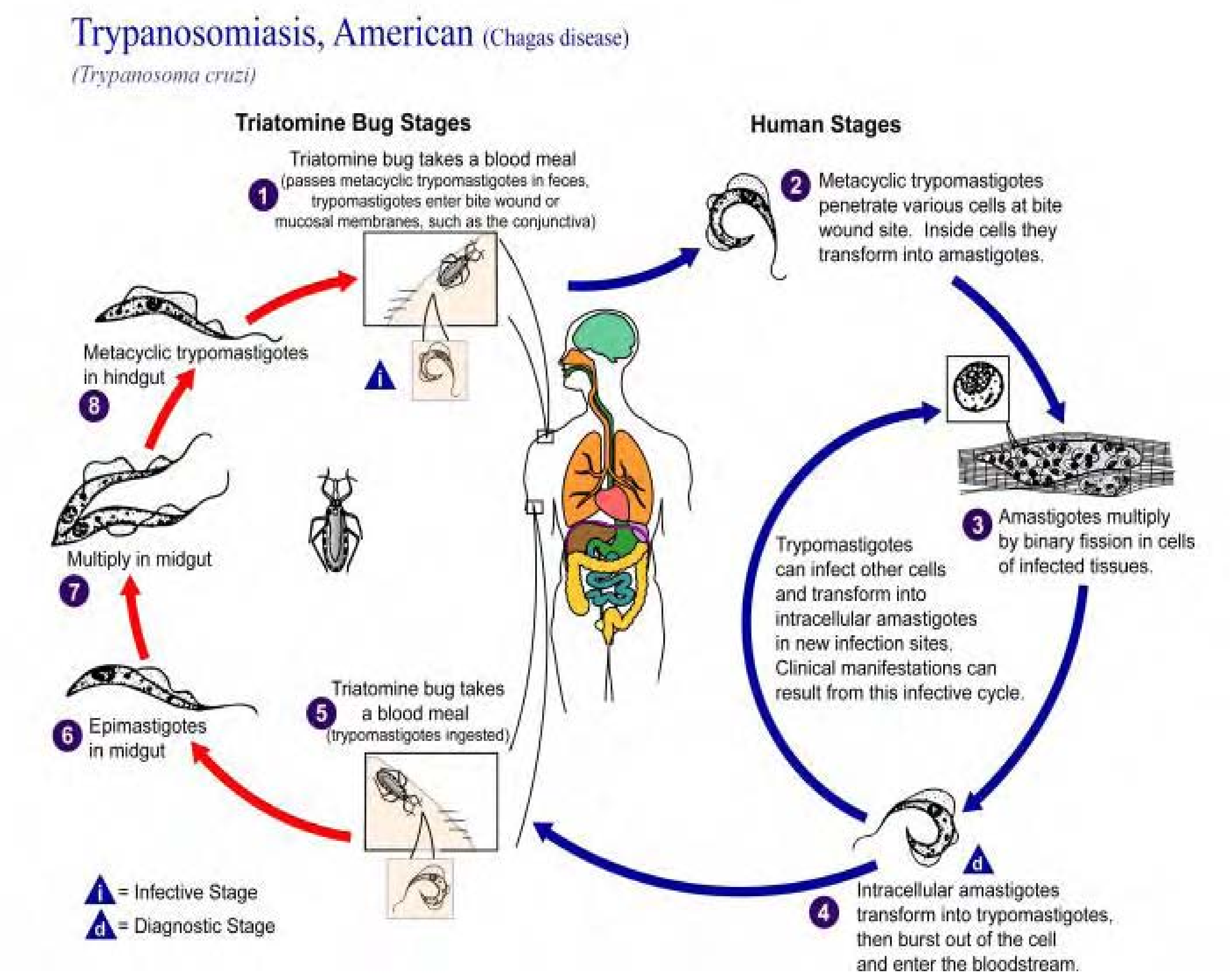
Trypanosomiasis, American (Chagas disease)
(Trypanosoma cruzi)
Triatomine Bug Stages
Human Stages
1 Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
2 Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3 Amastigotes multiply by binary fission in cells of infected tissues.
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
4 Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
5 Triatomine bug takes a blood meal (trypomastigotes ingested)
6 Epimastigotes in midgut
7 Multiply in midgut
8 Metacyclic trypomastigotes in hindgut
i
d
i = Infective Stage
d = Diagnostic Stage

Illustration 11-6. Generalized life cycle of the American trypanosomes, *Trypanosoma cruzi*, which is known to case Chagas' disease. (SOURCE: PHIL 3384 – CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

Trypanosoma cruzi has a single form (monomorphic), about 20μm in length, and characteristically curved. The kinetoplast is large, considerably larger than the *Trypanosoma brucei* species already discussed. They sometimes appear as a bulge at the posterior end. The flagellum is medium in length.

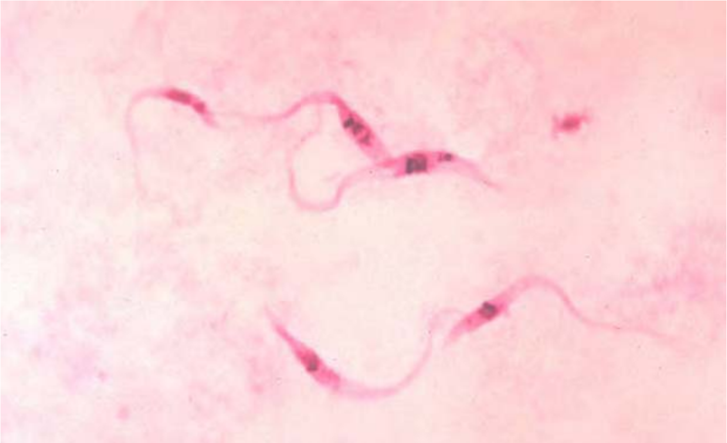

Image 11-5. *Trypanosoma cruzi* hemoflagellate in blood sample (SOURCE: PHIL 545 - CDC)

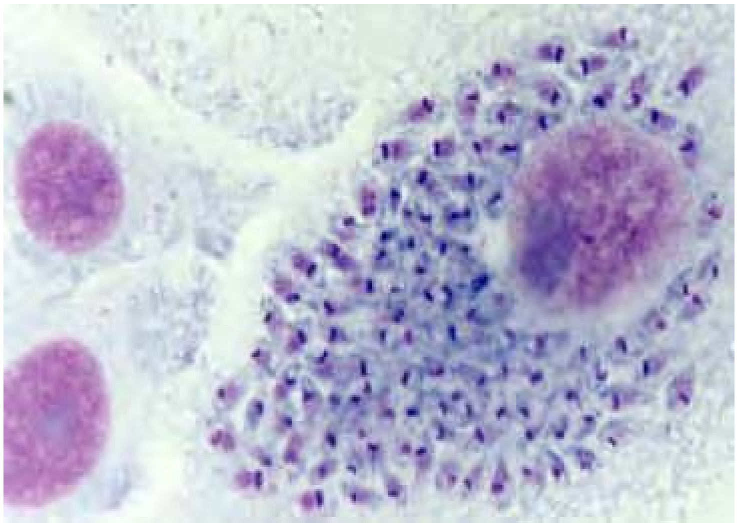

Image 11-6. Multiplication of *Trypanosoma cruzi in man* only occurs in the amastigote phase, which grows in a variety of tissue cells especially muscle. In vitro infected fibroblast showing a large number of intracellular amastigotes. (Giemsa stain) (SOURCE: Unknown)

Clinical Disease

Many people infected with *T. cruzi* remain asymptomatic and free from Chagas' disease or experience only an acute infection without progressing to the chronic stage. The most severe form of the disease is most commonly seen in children younger than five years of age.

Multiplication of *T. cruzi* at the site of infection can produce an inflamed swelling (chagoma) which persists for weeks. Trypomastigotes or amastigotes may be seen in the aspirate of the chagoma. Regional lymph nodes may become infected which frequently involve one side of the face. Unilateral edema of the upper and lower eyelid may occur along with conjunctivitis. This is known as Romanas' sign.

In the acute stage of infection trypomastigotes can be found in the blood. Symptoms may pass unnoticed, but there may be fever, malaise, increased pulse rate, and enlargement of lymph glands, liver, and possibly spleen. Muscle aches and pains are characteristic at this stage and parasites may be seen in blood films. The acute form is most often seen in young children and occasionally can cause serious damage to the heart and other complications leading to death caused by central nervous system involvement.

Chronic manifestations include signs of cardiac muscle damage with a weak and irregular heartbeat, edema, heart enlargement leading to heart failure. Dilation of the digestive tract resulting in megaesophagus and megacolon may also occur. About 10% of persons infected with *T. cruzi* develop chronic Chagas cardiopathy.

Laboratory Diagnosis of American Trypanosomiasis

Laboratory diagnosis of South American trypanosomiasis is by:

- Examination of blood.
- Xenodiagnosis
- Blood culture
- Serology

1. Examination of Blood

a) Thick and thin blood films are made and stained with Fields stain and examined as for malaria parasites. Wet preparations of blood can also be examined for motile trypanosomes.

b) Buffy coat examination - Trypanosomes are centrifuged in a microhematocrit tube for 5 minutes. Parasites can be seen microscopically at the junction of the packed red cells and plasma. This technique is rapid and sensitive.

Trypanosoma cruzi can often be seen in C, U or S shapes in stained films.

2. Xenodiagnosis

Xenodiagnosis is useful in chronic and sub acute (low parasitaemia) disease. Sterile bugs are fed on patients by attaching a black bag containing the bugs to the arm of the patient and allowing them to feed for 30 minutes. Twenty five to thirty days later the bugs are dissected and the contents of the hindgut and rectum are examined microscopically for the presence of trypanosomes.

3. Blood Culture

Blood culture is as sensitive as xenodiagnosis but it requires sterile conditions.

4. Serology

Serology tests include:

a. **IFAT** indirect fluorescence antibody test
b. **CFT** complement fixation test
c. **IHAT** indirect hemagglutination test
d. **ELISA** enzyme linked immunoabsorbent assay

Other lab findings include:

Raised ESR, marked lymphocytosis with atypical mononuclear lymphocytes

NB. In certain areas of S. America where *Trypanosoma rangeli* (a nonpathogenic species transmitted by *Rhodnius* spp. bugs) is found with *T. cruzi.* All positive preparations should be checked to confirm *T. cruzi.*

Leishmania species

Introduction

Leishmaniasis is caused by parasites of the genus *Leishmania* and is endemic in many parts of Africa, Asia and South America. It is transmitted by *Phlebotomus* species, sandfly.

Leishmania species are mainly parasites of man and other animals, especially dogs and rodents. They cause diseases collectively known as Leishmaniasis; causing three types of disease (i.e. Visceral Leishmaniasis, Cutaneous Leishmaniasis, and Mucocutaneous Leishmaniasis. These are all debilitating and disfiguring diseases which occur throughout the Old and New World. The parasites are unusual in that they live entirely within the cells of the reticuloendothelial cells, they have become perfectly adapted as the proteolytic enzymes which attack other foreign bodies in the blood stream but do not destroy them.

Visceral Leishmaniasis

Human visceral leishmaniasis (VL), sometimes known as Kala-azar, is caused by *Leishmania donovani* complex; *L. donovani* and *L. donovani infantum* in the Old World and *L. donovani chagasi* in the New World. The clinical features -azar caused by these species are similar, but they have different epidemiological features. The parent species *L. donovani* occurs in Asia (Northeastern China, India and Iran) and Africa (primarily Sudan, Kenya and Ethiopia) and can affect people of all ages. The parasite (*L. d. infantum*) which causes VL in countries bordering the Mediterranean, (Southern Europe as well as North Africa) affects young children as well as infants. It is now being seen in the immunocompromised. In the New World also, VL is mainly a disease of young children, with the causative organism *L. d. chagasi* being closely related to, but slightly different from, *L. donovani*. The main geographical foci of VL in Latin America are in northern and northeastern Brazil. Small foci are found in northern Argentina, Columbia and Venezuela. Sporadic cases are found in Central American countries, including Mexico.

Cutaneous Leishmaniasis

Cutaneous leishmaniasis is caused by *L. tropica, L. major* and *L. aethiopica* in the Old World and *L. mexicana* complex in the New World. *Leishmania tropica* is widely distributed around the Mediterranean basin, Afghanistan, Kenya, Kenya, Armenia, Azerbaijan, Turkmenistan and Uzbekistan*. Leishmania aethiopica* is seen in the highlands of Ethiopia and *L. major* occurs in the Middle East, West Africa, North Africa and Kenya. *Leishmania mexicana* complex is found in Central America and the Amazon Basin.

Mucocutaneous Leishmaniasis

Is caused by the *L. braziliensis* complex and is found in Brazil, Eastern Peru, Bolivia, Paraguay, Ecuador, Columbia and Venezuela.

Life cycle

All forms of infection starts when a female sandfly (*Phlebotomus* species) takes a blood meal from an infected host. Small amounts of blood, lymph and macrophages infected with *Leishmania* **amastigotes** are ingested. Once ingested the **amastigotes** transform to **promastigotes** in the sandfly, the non-infective **promastigotes** divide and develop into infective **metacyclic promastigotes**. These are formed in the midgut of the sandfly and migrate to the proboscis. When the sandfly bites, the extracellular inoculated promastigotes at the site of the bite are phagocytosed by macrophages. After phagocytosis, transformation to dividing **amastigotes** occurs within 24 hours. Reproduction at all stages of the lifecycle is believed to occur by binary fission. No sexual stage has been identified.

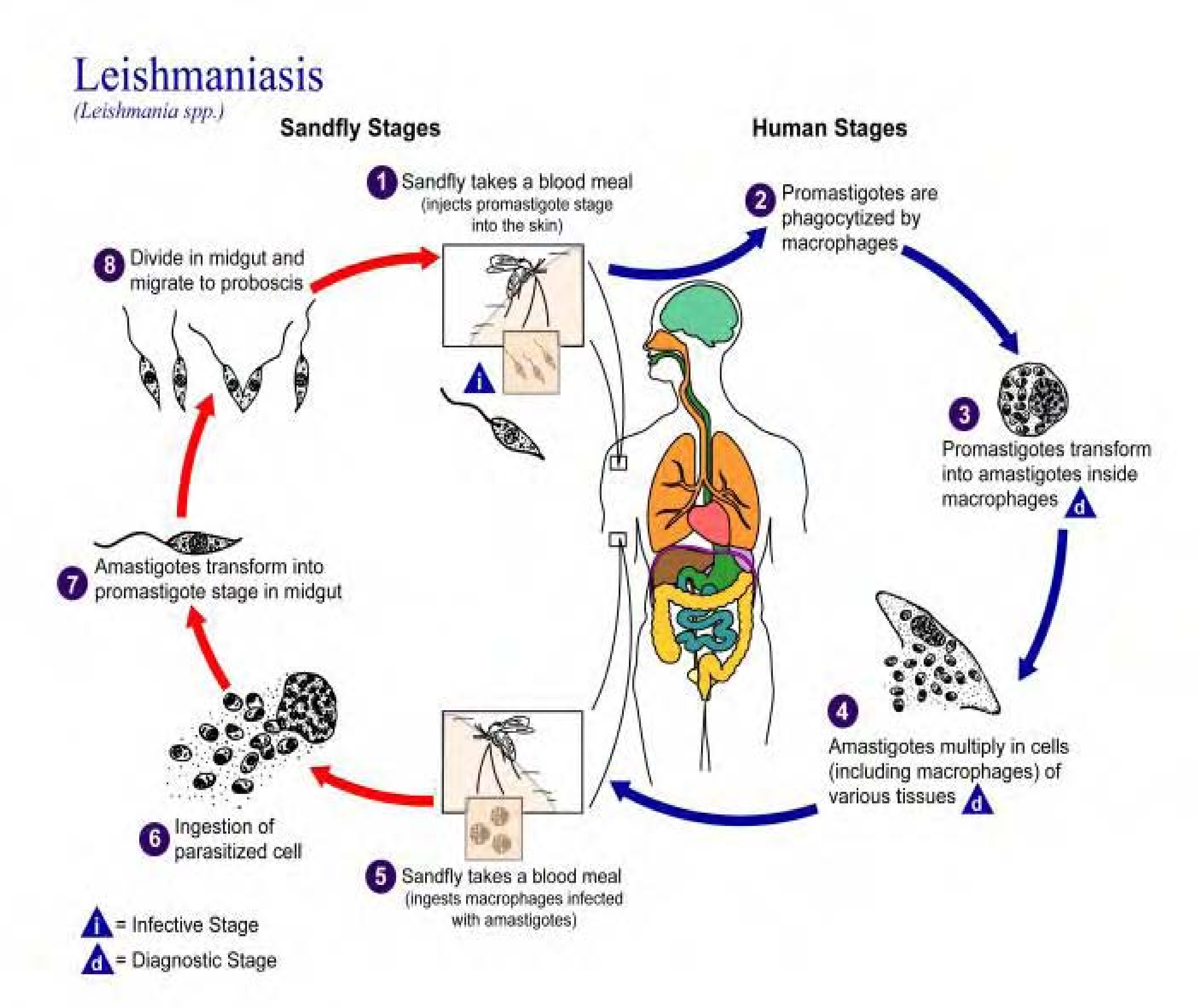
Leishmaniasis
(Leishmania spp.)
Sandfly Stages
Human Stages
1 Sandfly takes a blood meal (injects promastigote stage into the skin)
2 Promastigotes are phagocytized by macrophages
3 Promastigotes transform into amastigotes inside macrophages d
4 Amastigotes multiply in cells (including macrophages) of various tissues d
5 Sandfly takes a blood meal (ingests macrophages infected with amastigotes)
6 Ingestion of parasitized cell
7 Amastigotes transform into promastigote stage in midgut
8 Divide in midgut and migrate to proboscis
i
i = Infective Stage
d = Diagnostic Stage

Illustration 11-7. Generalized life cycle of *Leishmania* sp., which by using a sandfly as a vector can cause Leishmaniasis in man. (SOURCE: PHIL 3400 – CDC/Alexander J. da Silva, PhD/Melanie Moser)

Morphology

Leishmania exist as flagellated extracellular **promastigotes** in the Sandfly vector and as a flagellar obligate intracellular **amastigotes** within mononuclear phagocytes of their vertebrate hosts. The various species are not distinguishable morphologically from one another. When stained with Romanowsky stains such as Giemsa, **amastigotes** appear as round or oval bodies ranging from 2 - 3μm in diameter with a well defined nucleus and kinetoplast, a rod shaped specialized mitochondrial structure that contains extranuclear DNA. The flagellated **promastigote** form is spindle shaped, measuring 10 - 20μm in length, not including the length of the flagellum. As in the amastigote form a nucleus and kinetoplast are clearly visible.

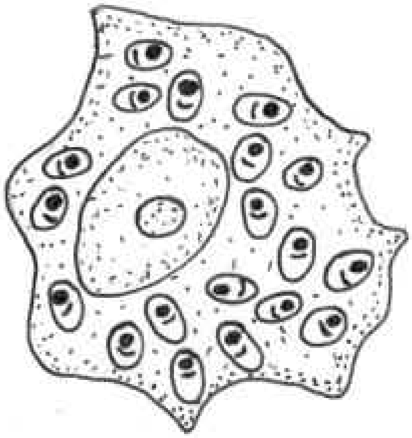

Illustration 11-8. Amastigote of *Leishmania* sp. in a macrophage of a vertebrate host. (SOURCE: Unknown)

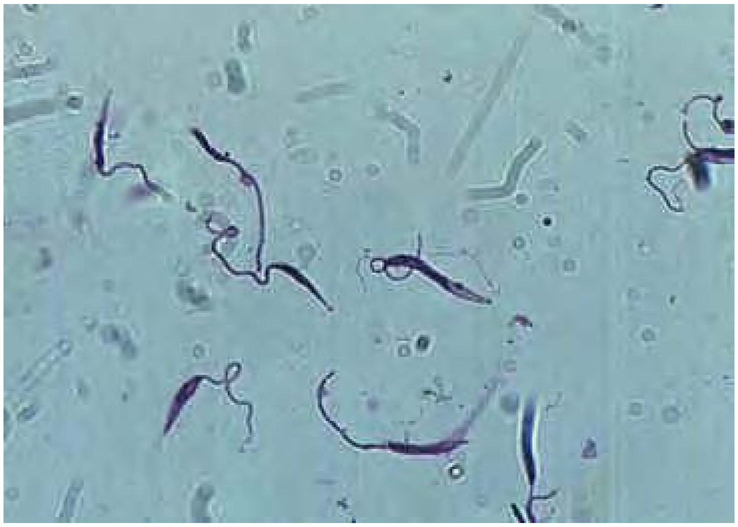

Image 11-7. Leishmania promastigote. This stage of the parasite are seen in the vector, sandfly. (SOURCE: Unknown)

Clinical Disease - Visceral Leishmaniasis

The incubation period of VL may vary between two weeks and 18 months. The onset of VL is usually insidious with fever, sweating, weakness and weight loss. The most prominent findings are fever, hepatosplenomegaly and anemia. The sites mainly affected are the liver, spleen and bone marrow. Enlargement of the liver is due to hyperplasia of Kupffer cells which are packed with amastigotes. The bone marrow is infiltrated with parasitized macrophages. Some organs, notably the kidneys, may show pathological changes secondary to deposition of immune complexes. In advanced cases, ascites and edema can develop. Deaths are usually due to secondary bacterial infections such as pneumonia, tuberculosis or dysentery.

Laboratory Diagnosis - Visceral Leishmaniasis

1. Microscopy

Parasites may be found in a splenic aspirate, liver biopsy or bone marrow biopsy. These techniques, especially splenic aspirate and liver biopsy, can be hazardous and require previous expertise in the procedure.

a) Air dry smears

b) Fix in methanol for one minute

c) Stain with Giemsa 1 in 10 in buffered distilled water pH 6.8 for 30 minutes (or use the rapid Field’s stain)

d) Wash the slide in buffered water and drain dry

Amastigotes of *Leishmania* should be seen in positive smears. They are approximately 2-4µm in size, oval and are frequently seen within the cytoplasm of the macrophage. The amastigotes possess a nucleus and a rod - shaped kinetoplast within the cytoplasm. In many samples a very small number of parasites are present. Extensive searching of the film is necessary.

2. Culture

The aspirates can be cultured in Novy-Nicolle-MacNeal (NNN) or Schneider's Drosophila medium. In culture the **amastigote** stage converts to the **promastigote** stage. However, this is not a rapid technique, as the parasites may take anything from 10-21 days to grow.

3. Serodiagnosis

VL produces large amounts of specific IgG which can be used for diagnosis. Currently the most used sero diagnostic tests are Indirect-immuno Fluorescent Antibody Test (IFAT), Enzyme Linked Immunosorbent Assay (ELISA) and Direct Agglutination Test (DAT).

Clinical Disease – Cutaneous Leishmaniasis

Following a bite from an infected sandfly, a small red papule appears at the site of the bite about 2–8 weeks later. The papule increases in size centrifugally. The patient then mounts either a hypersensitive response or an anergic response. In a hypersensitive response, the papule eventually ulcerates, becomes depressed and then eventually heals through scarring. The patient is now immune from subsequent bites. In an anergic response, the nodule grows and spreads over large areas of skin. This resembles leprosy.

Laboratory Diagnosis - Cutaneous Leishmaniasis

1. Slit skin smear.

The margin of the lesion contains amastigotes whereas the centre contains debris and dead skin material. This margin of the lesion is aseptically punctured with a hypodermic needle and syringe containing a small amount of saline. The aspirate which is drawn up into the needle is examined microscopically and/or cultured using the method described in Visceral Leishmaniasis.

2. Polymerase chain reaction

Gene amplification techniques are powerful and sensitive methods and are useful in diagnosis of Cutaneous Leishmaniasis particularly when organisms cannot be detected microscopically. It is also very useful for the speciation of *Leishmania* parasites thus the correct treatment can be administered.

Clinical Disease – Mucocutaneous Leishmaniasis

Mucocutaneous Leishmaniasis or espundia initially develops like cutaneous leishmaniasis but develops into lesions in the mucocutaneous junction of the pharynx resulting in the break down of the palate of the mouth and nose or more rarely the genitalia or anus. This occurs from a few weeks to several years after the cutaneous lesion has healed. These lesions result in disfiguring deformities of the nose and mouth.

Laboratory Diagnosis - Mucocutaneous Leishmaniasis

1. Microscopy

Finding the organisms in a histological section of the lesion provides definitive diagnosis of Mucocutaneous Leishmaniasis. However, the organisms are very rare in this form of the disease and culture can be a more sensitive method (see visceral leishmaniasis).

2. Polymerase Chain Reaction

The PCR method has the advantage of not only low numbers of parasites in aspirates but also histological sections. This makes this a very sensitive method in diagnosing Mucocutaneous Leishmaniasis when parasites are difficult to detect.

Chapter 12.
Arthropod Vectors

Soft Ticks

Class: Arachnida
Subclass: Acari
Order: Ixodida
Family: Argasidae
Genus: *Ornithodorous*

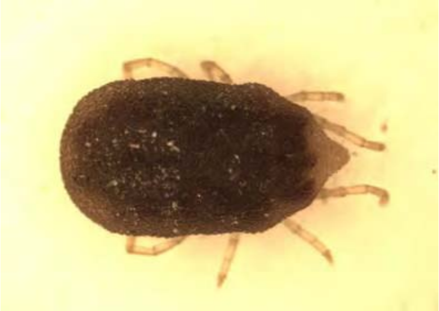

Image 12-1. *Ornithodoros* – Soft Tick (SOURCE: CDC)

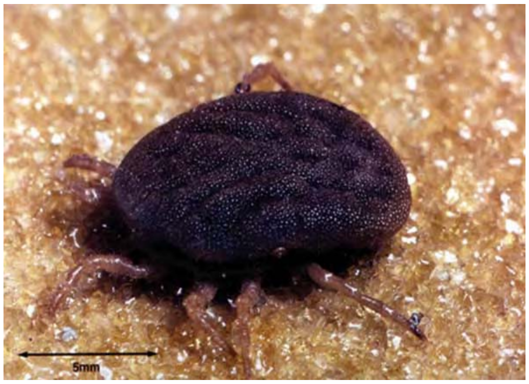

Image 12-2. *Ornithodorus moubata,* the most important soft tick disease vector found throughout the world (SOURCE: CDC)

General Characteristics

Soft Ticks are characterized by a tough leathery integument and a flattened oval shape when examined dorsally. The Argasidae lack the dorsal shield present in the Hard Ticks (Ixodidae), and need to be examined ventrally to observe their capitulum or mouthparts. Soft ticks generally have a world-wide distribution, with the most important disease vector *Ornithodoros* populating Europe, Africa, Asia and the Americas.

Life Cycles

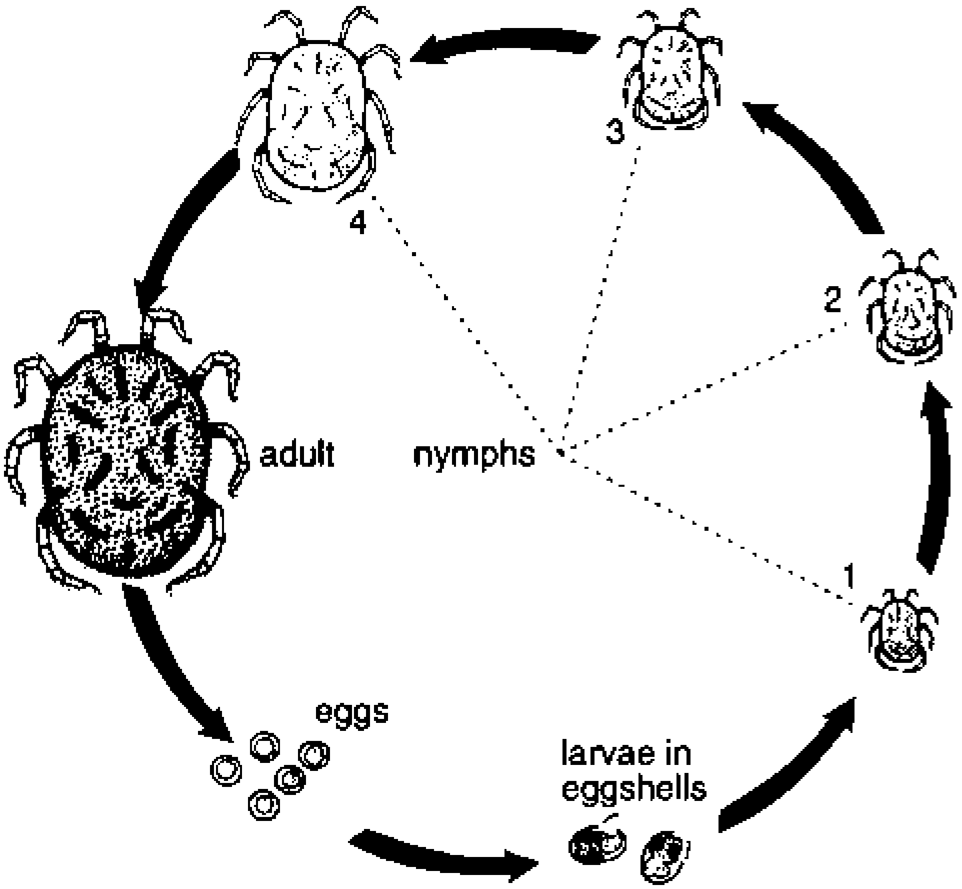

Illustration 12-1. Life Cycle of Soft Ticks. (SOURCE: Unknown)

Soft ticks have a hemimetabolous life cycle, with eggs hatching six legged larvae, which molt to eight legged nymphs. There are five to seven larval instars depending on the species, with each stage requiring a blood meal to proceed. Larvae of *Ornithodoros moubata* do not require a blood meal to molt to the nymph stage. Adult females lay small egg batches following each blood meal. The duration of the life cycle depends on the temperature, host availability, and inherently the particular species in question.

The adults are flat and oval in outline and have tough, leathery, wrinkled bodies. The mouthparts are situated underneath the body and are not visible from above. The eggs are laid in the places where the adults rest, such as cracks and crevices in the walls and floors of houses and in furniture. The larva, the five nymphal stages and the adults all actively search for hosts from which to take blood-meals. After feeding, which lasts about 30 minutes, they drop to the ground. Most species can survive for more than a year between blood-meals, and some for more than 10 years.

The soft ticks live apart from their hosts and are most common in the nests and resting places of the animals on which they feed. Some species, such as the chicken tick and the pigeon tick (*Argas* species) may feed on humans when the preferred hosts are not available.

Species that commonly feed on humans are found around villages and inside houses. Their habits are comparable to those of bedbugs: ticks often emerge from hiding places at night to suck the blood of humans and animals. Some species are common on travel routes, in rest houses and camping sites, and in caves and crevices.

Disease

Soft Ticks are vectors for serious disease including tick borne relapsing fever (*Borrelia duttoni),* rickettsial disease (*Coxiella burneti*), and some arboviruses. The most important disease spread by soft ticks is tick borne relapsing fever which occurs world-wide and is spread by spirochete infected *Ornithodoros.* Q-fever and arboviruses can be spread following a blood meal, but are both primarily introduced into the population by the Hard Ticks.

Tick-borne Relapsing Fever

This disease is caused by a microorganism of the genus *Borrelia*. It is transmitted by biting soft ticks of the genus *Ornithodoros* in many countries in the tropics and subtropics and also in Europe and North America. The ticks usually feed quickly at night in or near houses, and then leave the host.

The disease causes bouts of fever alternating with periods without fever. Death occurs in about 2-10% of persons who are untreated.

Treatment

Treatment is possible with tetracycline or its derivatives.

Prevention

Prevention requires measures to control soft ticks and to avoid their bites.

Hard Ticks

Class: Arachnida

Subclass: Acari

Order: Ixodida

Family: Ixodidae

Species: ***Ixodes, Amblyomma, Rhipicephalus, Dermacentor***

General Characteristics

Hard ticks appear flattened when examined dorsally and can be primarily characterized by the presence of a dorsal plate or scutum, and a capitulum that projects beyond the body outline. The scutum regularly covers the entire dorsal area, although females may have a reduced plate present directly behind the capitulum. *Ixodes* spp. inhabit in Canada, Europe, Asiatic Russia, China, Japan, and Australia.

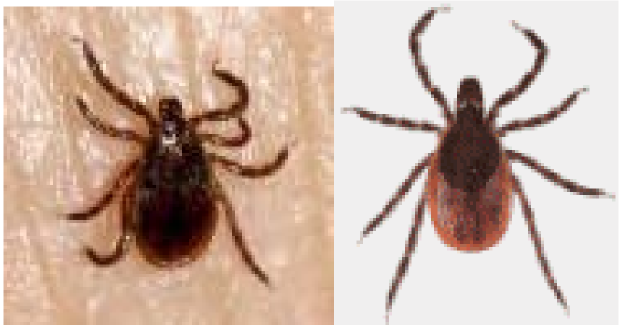

Image 12-3. *Ixodes scapularis* - male - showing the scutum covering the whole length of the body (left); *Ixodes scapularis* - female, showing the scutum covering half the length of the body (right)

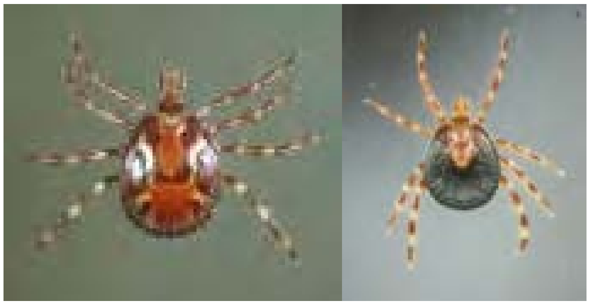

Image 12-4. *Amblyomma variegatum* - male - showing the scutum covering the whole length of the body (left); *Amblyomma variegatum* - female - showing the scutum covering half the length of the body (right)

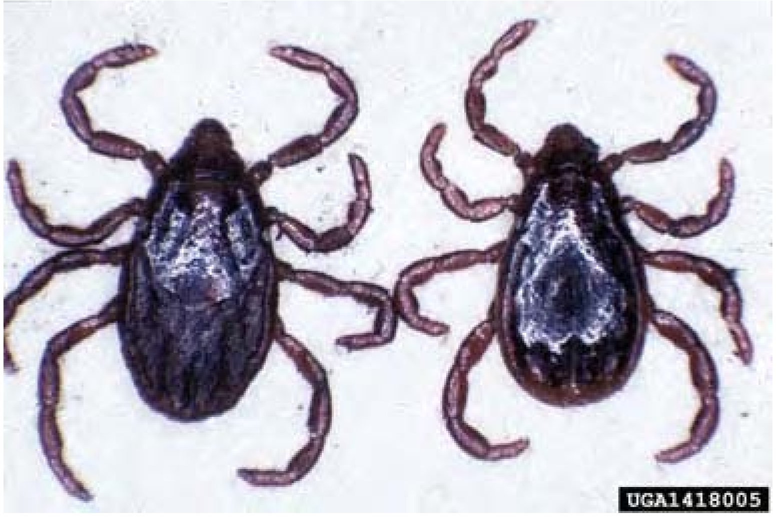

Image 12-5. *Rhipicephalus sanguineus* - female - showing the scutum covering half the length of the body (left); *Rhipicephalus sanguineus* - male - showing the scutum covering the whole length of the body (right). (AFPMB – used with permission)

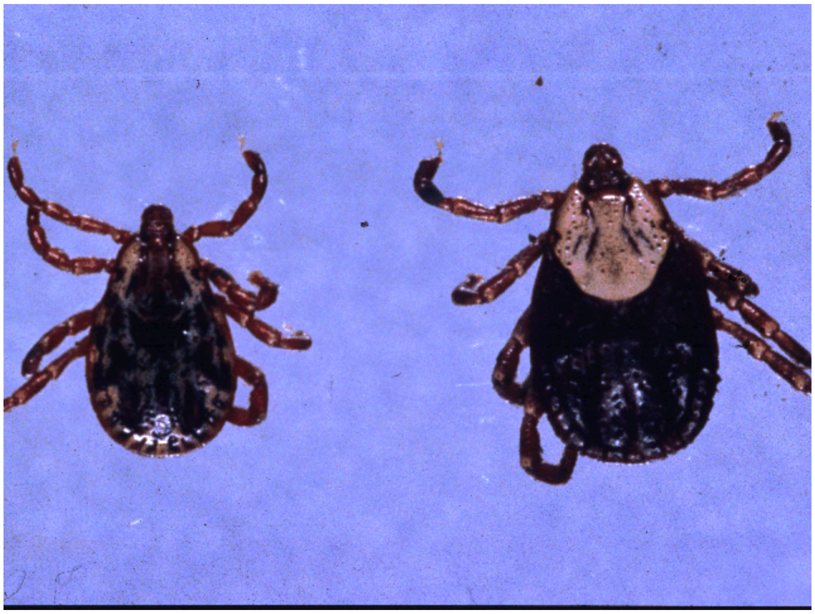

Image 12-6. *Dermacentor variabilis* - male - showing the scutum covering the whole length of the body (left); *Dermacentor variabilis* - female - showing the scutum covering half the length of the body (right) (Image courtesy of: Iowa State University)

Life Cycle

Hard Ticks have a hemimetabolous life cycle similar to that of the Soft Ticks. Hard tick larvae search for suitable hosts, feed for several days, and drop off engorged to the ground below. Molting to the nymph stage follows several days later and the tick again searches for a suitable blood source. There is only one nymphal stage in *Ixodes,* and following several weeks of stasis the nymph will molt into an adult. Females lay one large mass of eggs, which forms a cellular mass on the scutum of the ovipositing female.

Disease

Hard ticks transmit a variety of diseases including Lyme Disease, Tick Paralysis, Rickettsia*,* arboviruses that are responsible for encephalitis and hemorrhagic fevers, tularemia and *Babesia microti* infection. Tick paralysis is caused by the toxins present in the saliva of the tick, and is introduced during feeding.

Lyme disease is a serious ailment caused by the spirochete *Borrelia burgdorferi.* Lyme disease results in acute erythema, systematic lesions and eventually chronic involvement of multiple organs.

Rickettsiae diseases spread by *Ixodes* include Q-fever (*Coxiella burneti*) and arboviruses that cause serious encephalitis.

Babesia microti infection primarily involves animals, with Man as an accidental host. The parasite is present in the tick salivary glands and is passed to Man via tick bite. Once in the bloodstream, the parasites enter the erythrocytes. Infection with *B. microti* ranges from asymptomatic to severe illness. Symptoms generally resemble those of malaria with fever, rigors, myalgia, and malaise. Occasionally there is mild or moderate hemolytic anemia and jaundice (renal failure). Parasitemia can be up to 25% in the immunocompetent individual, and up to 80% in patients who have been splenectomized.

Other Medically Important Hard Ticks

General Characteristics

Several other species of ticks are also responsible for the spread of disease throughout the world. Wood ticks *(Dermacentor andersoni*) are found in the mountainous west of North America; dog ticks (*Rhipicephalus*) are found in coastal areas. Lone Star ticks (*Amblyomma americanum*) are found in forests in SE USA where deer are found. The females of these species are known to cause a condition known as tick paralysis.

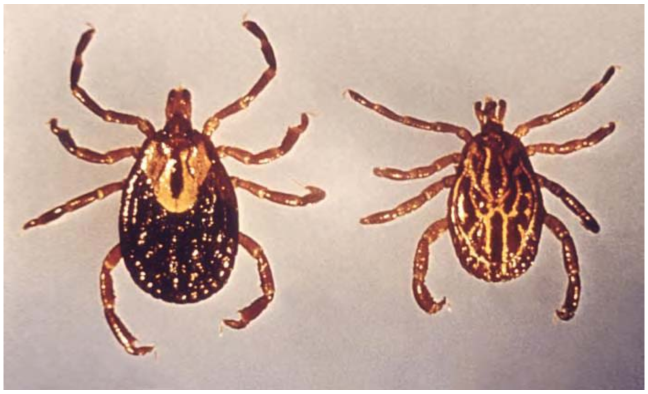

Image 12-7. *Amblyomma,* hard tick. The scutum is so-called 'enameled' colored areas, and is described as being an ornate species. (SOURCE: PHIL 5981 – CDC)

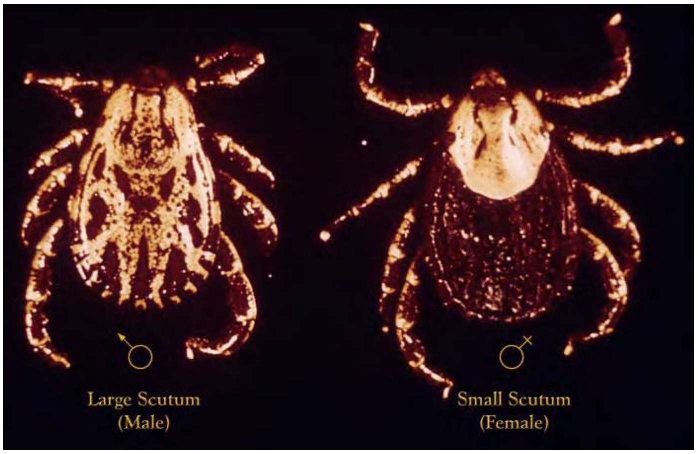

Image 12-8. *Dermacentor variabilis,* ticks of this species show very ornate scutums, covering the length of the body (left - male). When blood engorged it is harder to see the scutum (right - female). (SOURCE: PHIL 5977 – CDC)

Life Cycle

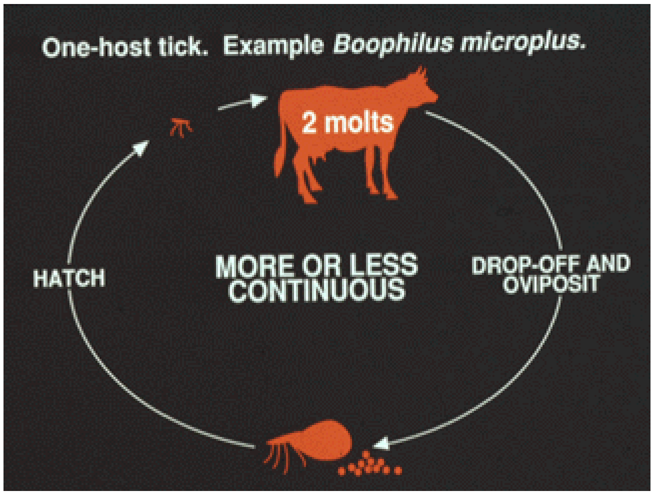
One-host tick. Example *Boophilus microplus.*
2 molts
HATCH
MORE OR LESS CONTINUOUS
DROP-OFF AND OVIPOSIT

Illustration 12-2. One-host tick life cycle. (SOURCE: CDC)

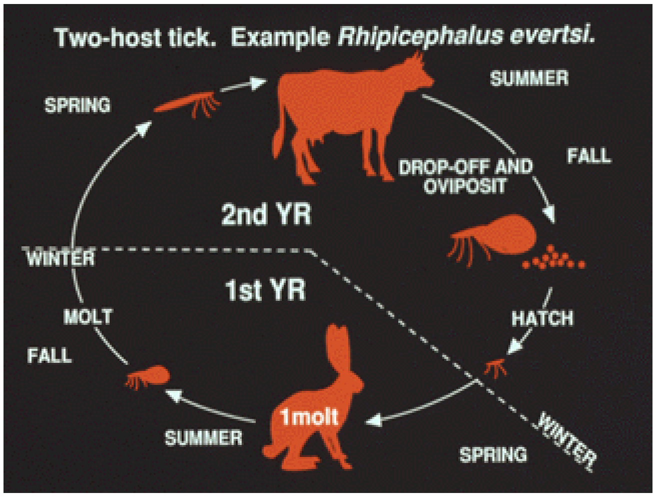
Two-host tick. Example *Rhipicephalus evertsi.*
SPRING
SUMMER
FALL
DROP-OFF AND OVIPOSIT
2nd YR
WINTER
1st YR
MOLT
FALL
HATCH
1molt
SUMMER
SPRING
WINTER

Illustration 12-3. Two-host tick life cycle. (SOURCE: CDC)

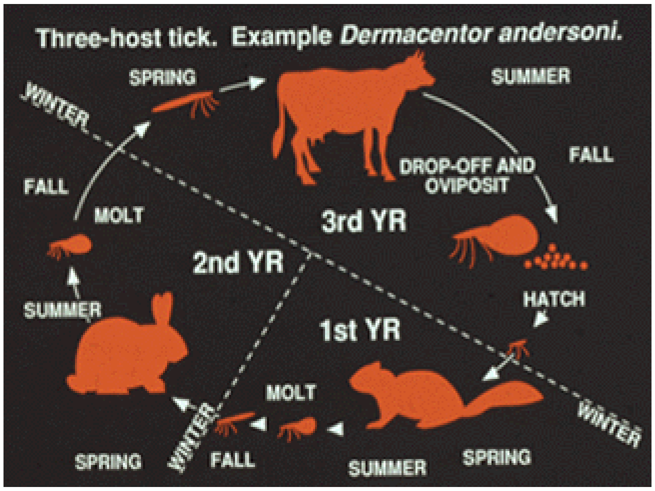
Three-host tick. Example *Dermacentor andersoni.*
WINTER
SPRING
SUMMER
FALL
FALL
MOLT
DROP-OFF AND OVIPOSIT
3rd YR
2nd YR
SUMMER
HATCH
1st YR
MOLT
WINTER
SPRING
WINTER
FALL
SUMMER
SPRING

Illustration 12-4. Three-host tick life cycle. (SOURCE: CDC)

Hard Ticks have a hemimetabolous life cycle similar to that of the soft ticks. Hard tick larvae search for suitable hosts feed for several days, and drop off engorged to the ground below. Molting to the nymph stage follows several days later and the tick again searches for a suitable blood source. There is only one nymphal stage in *Ixodes,* and following several weeks of stasis the nymph will molt into an adult. Females lay one large mass of eggs, which forms a cellular mass on the scutum of the ovipositing female.

Image 12-9. Female tick with egg mass. (SOURCE: PHIL 5968 – CDC)

Diseases

Both *Dermacentor* and *Amblyomma* transmit Rocky Mountain Spotted Fever (*Rickettsia rickettsia*). The different strains transmitted by these species vary in their virulence and subsequent manifestation of the disease. *Dermacentor* are known to transmit serious arboviruses responsible for encephalitis and hemorrhagic fevers. *D. variabilis* are also responsible for spreading tularemia (*Francisella tularensis*) and human granulocytic erlichiosis (*Ehrlichia chaffeensis*). *Rhipicephalus sanguineus* transmits Mediterranean Spotted Fever and African Tick Typhus.

Tick Paralysis

Hard ticks inject into the body with their saliva certain toxins that can cause a condition in people and animals called tick paralysis. It appears 5-7 days after a tick begins feeding, paralyzing the legs and affecting speaking ability, swallowing and breathing. It occurs worldwide and is most common and severe in children aged up to two years. Treatment involves removing the tick.

Tick-borne Rickettsial Fevers

This group of diseases is caused by closely related *Rickettsia* microorganisms transmitted by tick bites or contamination of the skin with crushed tissues or feces of the tick.

- Spotted fever due to *Rickettsia rickettsii* occurs in Brazil, Canada, Colombia, Mexico, Panama and the USA.
- Spotted fever due to *R. sibirica* occurs in Japan, the Russian Federation and the Pacific.
- Spotted fever due to *R. conori* is found in the Mediterranean region, Africa and southern Asia.
- Spotted fever due to *R. australis* occurs in Queensland, Australia.
- Q fever, caused by *Coxiella burneti*, has a worldwide distribution and is commonly present in abattoirs, meat-packing and meat-rendering plants, diagnostic laboratories, stockyards and poultry farms. It is transmitted to humans mainly by the consumption of milk and meat from contaminated cattle or the inhalation of dried infected tick feces by people working with cattle.

Symptoms in humans are sudden fever persisting for several weeks, malaise, muscle and joint pains, severe headache and chills. A rash sometimes spreads over the entire body. Death may result in about 15-20% of persons if the disease is misdiagnosed or left untreated.

Treatment

Antibiotics such as tetracycline or chloramphenicol can be used.

Prevention

Tick bites should be avoided and attached ticks should be removed rapidly and carefully. Several hours of attachment are needed before the *Rickettsia* organisms can infect humans.

Lyme Disease

Lyme disease (erythema chronicum migrans) is a severe and often debilitating condition caused by a spirochete, *Borrelia burgdorferi*. Acute Lyme disease is a flu-like illness, characterized by an expanding red rash in about 50% of patients, accompanied by fever, fatigue, and muscle and joint pain. Weeks or even months after the infecting tick bite, patients may experience swelling and pain in large joints (knee, elbow), encephalitis, facial palsy, ocular lesions and carditis, irrespective of whether a rash occurred in the acute phase. Later, perhaps years after the bite, there may be cartilage erosion (arthritis) and neuromuscular dysfunction (Fig. 4.29). Lyme disease occurs principally in northern temperate regions of the world, including China, Europe, the USA and the former USSR.

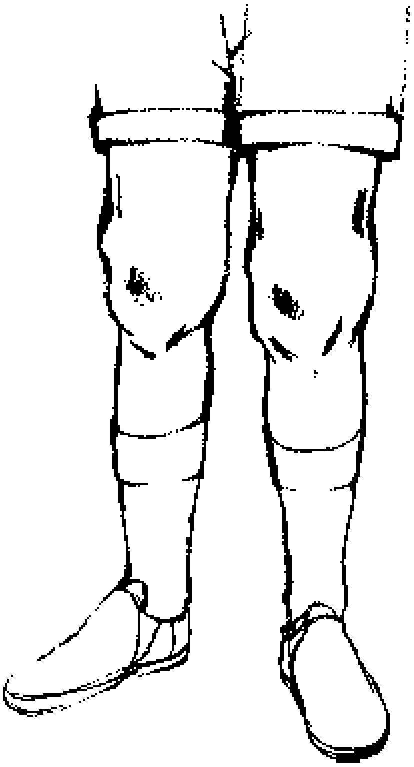

Illustration 12-5. A typical symptom of Lyme disease is swelling and pain in the large joints, such as the knees, and chronic arthritis. (SOURCE: Unknown)

Transmission

The disease is transmitted mostly by *Ixodes* ticks, commonly in the summer when the nymphs are abundant. Small rodents, especially mice, serve as reservoirs of infection while large mammals serve principally as hosts maintaining tick populations. The larvae acquire infection while feeding on mice, and nymphs or adults can transmit spirochetes during subsequent blood-meals. In the northern temperate zone, where it occurs most intensely, Lyme disease has become more common as deer populations have increased and as this critical host has adapted to living in closer proximity to people. In many areas, Lyme disease is acquired in the suburban residential environment.

Treatment

Further development of the disease in adults may be reduced or prevented by treatment with tetracycline or its derivatives for 2-4 weeks, and in children by treatment with penicillin.

Prevention

Prevention requires avoidance of tick habitats and bites, and vector control. Personal protection may be possible by the use of repellents on the skin and clothing in tick-infested areas. The removal of attached ticks within 24 hours may prevent spirochete transmission. Prophylactic antibiotic therapy may be desirable following the bite of an infected tick. New molecular assays are commercially available for detecting the spirochetes in tick samples.

Tularemia

Tularemia, also known as rabbit fever, deerfly fever and O'Hara disease, is caused by the infectious agent *Francisella tularensis*. The symptoms, which vary according to how the agent enters the body, include headache, chills, fever and the swelling of lymph nodes. The disease occurs in Europe, Japan, North America and the former USSR.

Transmission

Transmission takes place through the bites of ticks and deerflies (see Chapter 1) or as a result of handling infected animals such as rabbits and other game. Hunters and forest workers are at the highest risk of infection.

Treatment

Antibiotics such as streptomycin can be used to treat the disease.

Prevention

Tick bites and tick habitats should be avoided, impermeable gloves should be worn when skinning and dressing game animals, wild game meat should be thoroughly cooked, and untreated drinking-water should be avoided in areas where the disease occurs.

Tick-borne Viral Encephalitides

This is a group of viral diseases causing acute inflammation of the brain, spinal cord and meninges. The symptoms vary in severity with the type of disease. Many infections do not result in disease. Severe infections may cause violent headaches, high fever, nausea, coma and death.

- Far Eastern tick-borne encephalitis is found in the far east of the former USSR.
- Central European tick-borne encephalitis occurs in Europe from the Urals to France.
- Louping ill is a disease of sheep in the United Kingdom which sometimes affects people.

Transmission and Prevention

These diseases are transmitted by biting ticks and by the consumption of milk from infected animals. No specific treatment is available but vaccines have been developed against some of the diseases. Prevention requires avoidance or rapid removal of ticks.

Principal Hard Tick Vectors

Usually various tick species act as vectors for any one disease and their importance varies from region to region.

Disease	Vector
Lyme disease	Deer tick *(Ixodes scapularis)*
Spotted fever due to:	
• ***Rickettsia rickettsii***	American dog tick *(Dermacentor variabilis)*
• ***Rickettsia sibirica***	Asiatic wood tick *(Dermacentor silvarum)*
• ***Rickettsia conori***	Brown dog tick *(Rhipicephalus sanguineus)*
• ***Rickettsia australis***	Wattle tick *(Ixodes holocyclus)*
Q fever	Lone star tick *(Amblyomma americanum)*
Tularemia	American rabbit tick *(Haemaphysalis leporispalustris)*
Far Eastern tick-borne encephalitis	Taiga tick *(Ixodes persulcatus)*
Central European tick-borne encephalitis	Castor bean ticks *(Ixodes ricinus)*
Kyasanur Forest disease	A tick of birds and monkeys *(Haemaphysalis spinigera)*
Colorado tick fever	American wood tick *(Dermacentor andersoni)*
Crimean-Congo hemorrhagic fever	A tick of birds and mammals *(Hyalomma marginatum)*

Table 12-1. Principle Hard Tick Vectors

Other Viral Diseases

Kyasanur Forest disease occurs in parts of India.

Omsk hemorrhagic fever is found in south-western Siberia; it causes severe disease and death in muskrat handlers; it is mainly waterborne, although it is found in hard ticks.

Colorado tick fever is a moderately severe disease that occurs in western North America.

Crimean-Congo hemorrhagic fever is an acute, often severe and fatal disease found in parts of Africa, Asia and Europe.

Mites

- **Class: Arachnida**
  - **Subclass: Acari**
    - **Order: Astigmata**
      - **Family: Pyroglyphidae**
        - **Genus: *Dermatophagoides* (Dust mites)**
      - **Family: Sarcoptidae**
        - **Genus: *Sarcoptes* (Scabies mites)**
    - **Order: Prostigmata**
      - **Family: Democidae**
        - **Genus: *Demodex* (Hair follicle mite)**
    - **Order: Mesostigmata**
        - **Genus: *Leptotrombidium* (Other Mites)**

General Characteristics

Dermatophagoides, the common dust mite, commonly inhabit beds, mattresses, carpets and house dust. They are motile and feed on residual organic debris and are typically 0.3 µm in length. They characteristically have four long legs with suckers and pincer chelicerae.

Image 12-10. A Dust Mite (*Dermatophagoides*). They commonly inhabit mattresses, carpets and house dust.

Demodex folliculorum mites are 300 microns in length, live in the follicle for 5-6 days and migrate onto contiguous skin at night. Classically, they appear as 1 µm "sleeves" around the base of the eyelashes.

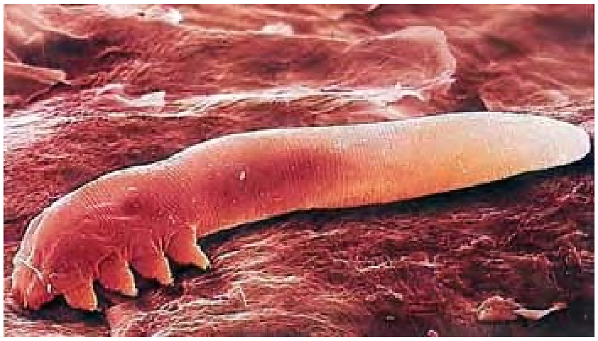

Image 12-11. *Demodex folliculorum* (SOURCE: CDC)

The mite, *Sarcoptes scabiei*, lives exclusively on human skin and causes scabies. It is small (the larger female is 0.3-0.4 µm), colorless and oval with 8 short legs.

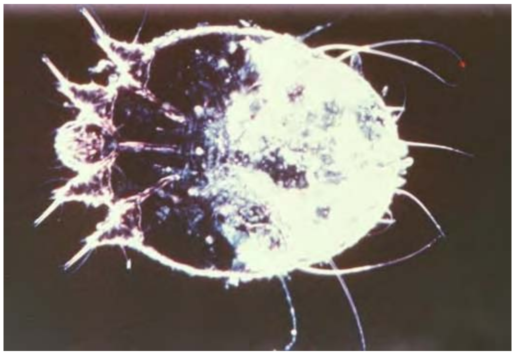

Image 12-12. Picture of the human parasite scabies (*Sarcoptes scabiei*). They live exclusively on human skin causing the condition known as scabies. (SOURCE: PHIL 3810 - CDC/Joe Miller/Reed and Crnrick pharmaceuticals.)

Leptotrombidium adults are normally reddish and covered with velvet like hair. It is the larvae that are parasitic and they too regularly appear reddish, but only have three pairs of legs. The palps are segmented, and the mouthparts are large and easily identifiable. *Leptotrombidium* larvae are distinguished by large eyes, two feathered sensillae and five feathered setae that are present on the scutum.

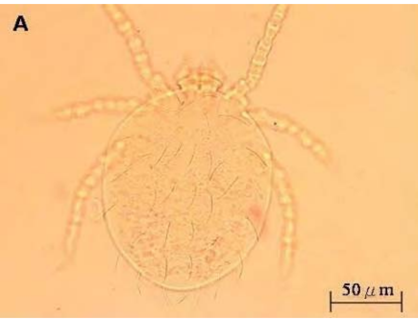

Image 12-13. *Leptotrombidium* (SOURCE: CDC of Taiwan)

Disease and Control

Dermatophagoides are antigenic, even when dead (fecal pellets are also allergenic), and are associated with complex allergies with symptoms such as asthma, perennial rhinitis, conjunctivitis and atopic dermatitis. A diagnosis can be made using reactive skin tests and treatment involves removal of accumulated antigens from mattresses and pillows.

Demodex have been implicated in dermatitis, acne and should be considered in chronic, therapy-resistant cases of blepharitis. The role of this mite in ocular disease is uncertain. Washing with soap and water is the most effective method of prevention.

Symptoms of *Sarcoptes* infection arise after 4-6 weeks and comprise a papular pruritic eruption at a site often unrelated to the site of infestation; symptoms of secondary infestations arise almost immediately. Zoonotic scabies have shorter incubation periods but are also self-limiting. Persistent nodular scabies comprises itchy erythematous or scabbed nodules, often on the penis. When the immune system is impaired, Norwegian (crusted) scabies may develop: this presents with erythema and hyperkeratosis but little itching.

Diagnosis demands discovery of the mites, scybala or eggs by epidermal shave biopsy or superficial scraping. Burrows are best seen on wrists and interdigital spaces. They fluoresce under a Wood's lamp after application of liquid tetracycline and, later, alcohol; alternatively ink may be used. Lindane (Quellada) lotion is the treatment of choice in the USA; permethrin or crotamiton are alternatives. Ivermectin has been recommended for Norwegian scabies. Malathion (Derbac-M) liquid or permethrin are recommended in the UK (benzyl benzoate is also active). Malathion should be avoided in infants and lindane should be avoided in pregnancy, breast-feeding and young children. Treatment is applied over the whole body except the head and neck and washed off after 24 hours. Normal laundering of bed linen and clothes is recommended. Household and sexual contacts should also be treated. Symptoms may continue after treatment because of persisting antigens, Calamine or Eurax lotion may be used; other problems include reinfestation and secondary bacterial infection.

Feeding by *Leptotrobidium* causes cellular damage and initially results in intense itching and irritation for the host. The agent of scrub typhus, *Rickettsia tsutsugamuchi*, is often transmitted by the larval trombiculid mite. Scrub typhus results in infection in those living or working near 'mite islands' which are usually found in transitional vegetation and provide a stable environment for mite proliferation. Prevention has generally been with chemical repellents or herbicides to remove the scrub habitat.

Mosquitoes

Class: Insecta
Order: Diptera
Family: Culicidae
Subfamily: Anophelinae
Genus: *Anopheline*
Subfamily: Culicinae
Genus: *Culicine*

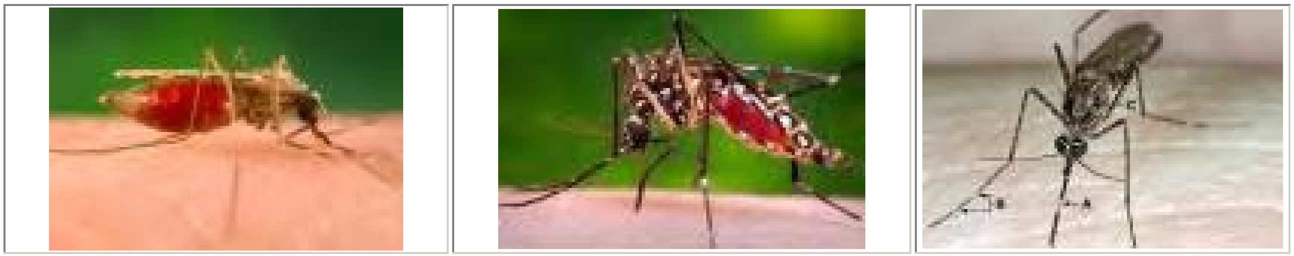

Anopheles
(SOURCE: PHIL 7949 – CDC)

Aedes
(SOURCE: PHIL 9261 - CDC/ Prof. Frank Hadley Collins, Dir., Cntr. for Global Health and Infectious Diseases, Univ. of Notre Dame)

Culex
(SOURCE: PHIL 7957 – CDC)

Image 12-14. *Anopheline* and *Culicine* mosquitoes.

Introduction

Mosquitoes are small with a clearly demarcated body and very long slender legs. The head contains a large pair of kidney shaped compound eyes, a pair of antennae, and a single long proboscis for feeding. The thorax, abdomen and wings are often covered with scales. Differential coloration and pattern of these scales provides a means of visually distinguishing species. The large wings are folded over the segmented abdomen, which generally appears brown-black and slender but turns a bright red and swells following feeding.

Image 12-15. Image of a typical Culicine mosquito. Adults are generally brown – black, at rest they palce their abdomens parallel to the ground. (SOURCE: PHIL 4487 – CDC)

Mosquitoes may be classified as Anopheline (Anopheles) or Culicine (Aedes, Culex). The antennae of male mosquitoes are plumose (many feathery hairs); females are pilose (few spidery hairs). The male Anopheline palps are long and clubbed; those of the male Culicine are long but not clubbed. The female Anopheline palps are long; the female Culicine are short. If a mosquito is incorrectly sexed, a female Anopheline may be confused with a male Culicine.

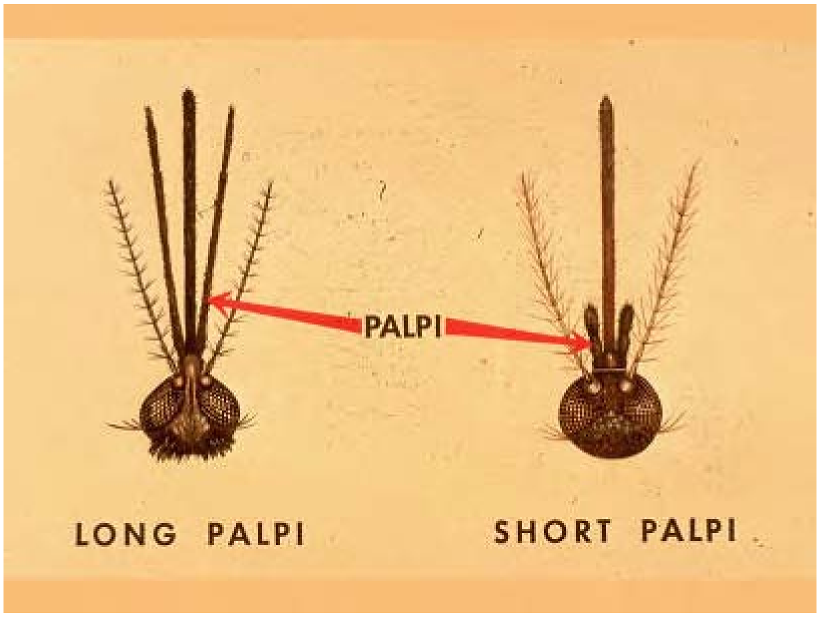

Illustration 12-6. Image of typical male Anopheline mosquito showing the characteristic features of the mosquitoes head; the palps are as long as the proboscis with clubbed like ends and the antennae are plumose. The female Culicine mosquito shows the short palps and the pilose antennae. (SOURCE: CDC)

At rest, Anophelines usually position their abdomens at an angle to the surface; Culicines maintain their abdomen in a parallel position.

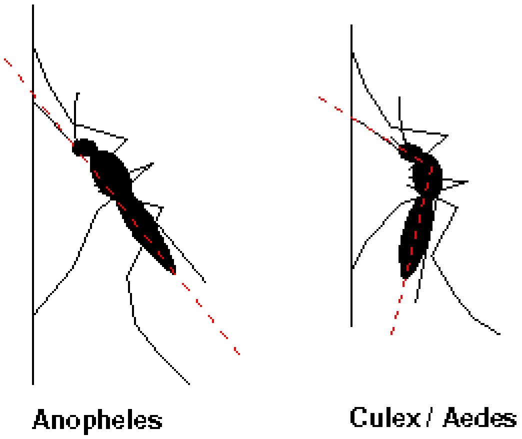
Anopheles
Culex / Aedes

Illustration 12-7. Typical *Anopheline* and *Culicine* at rest, showing the abdomen at an angle to the surface. (SOURCE: CDC)

Anopheline (*Anopheles*) Mosquitoes

Class: Insecta
  Order: Dipthera
    Family: Culicidae
      Subfamily: Anophelinae
        Genus: Anopheline
          Species: *Anopheles*

Image 12-16. Typical *Anopheles* mosquitoes (SOURCE: PHIL 7862 – CDC)

General Characteristics

Anopheles mosquitoes are characterized by dark and pale scale blocks arranged on their wings. They have palps that are of equal length to the proboscis, which appear terminally clubbed in males. Anopheles always rest at an angle when standing on surfaces and preferring to feed at twilight or night. Breeding sites are varied but Anopheles prefer unpolluted fresh or saltwater.

Life Cycle

Anopheles lay 50 to 200 dark color eggs in aquatic environments, and hatch in several days to several weeks depending on the external temperature. Anopheles larvae have a dark brown head and 6-7 anterior segments covered with dorsal palmate hairs. Accessory tegral plates are present on the dorsal side of segments 1-10 and two sets of anal papillae emerge from the last abdominal segment.

There are four larval instars that survive by filter feeding and breathing oxygen through their spiracles. Anopheles larvae occur throughout many different habitats including both permanent marshes and swamps, and temporary locations such as pots filled with water. In general Anopheles prefer to inhabit clean habitats. The larval period lasts about a week, but may be extended depending on the environmental conditions. The pupa is comma shaped with a set of trumpet shaped breathing tubes. The abdomen is covered with setae, and segments 2-7 have distinct spines. The pupal period may last a few days to weeks depending on the temperature.

Disease - Malaria

Anopheles are vectors of malaria, Bancroftian and Brugian filariasis and of multiple arboviruses (dengue fever; yellow fever; encephalitides and hemorrhagic fevers). Malaria is caused by *Plasmodium falciparum, P. vivax, P. malariae and P. ovale*. Transmission of the disease occurs in virtually all of tropical Africa, Central and South America, and the Middle and Far East. South East Asia is a particular problem due to multiple drug resistance. *P. falciparum* is found in Africa and other tropical countries as well as in subtropics. *P. malariae* has a low prevalence in both tropics and subtropics. *P .vivax* is the most widespread in temperate regions and subtropics but may also be found in the tropics. *P. ovale* has a low prevalence in West Africa. In Africa alone, 370 million people live in endemic areas.

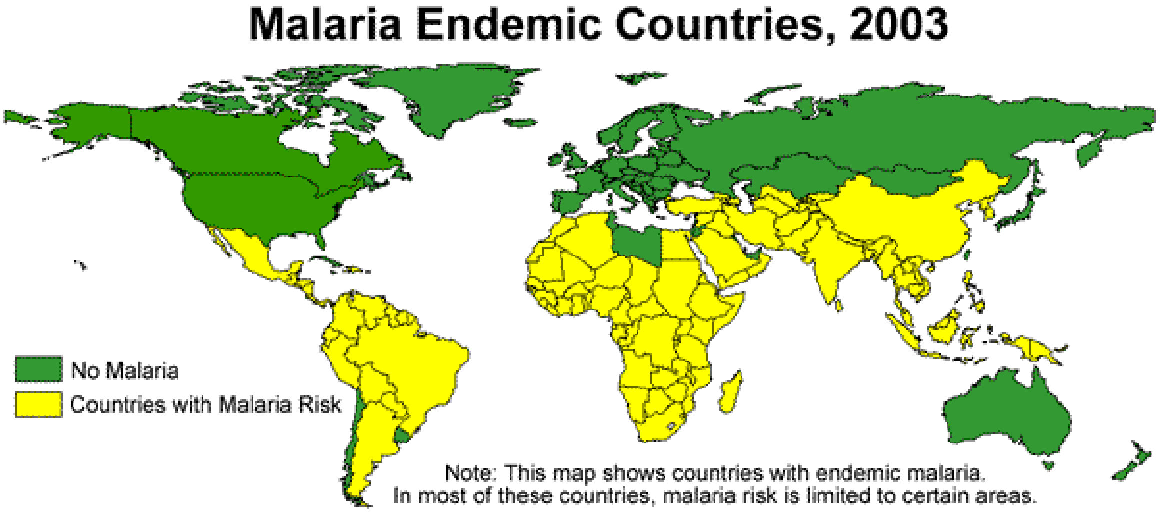
Malaria Endemic Countries, 2003
🟩 No Malaria
🟨 Countries with Malaria Risk
Note: This map shows countries with endemic malaria.
In most of these countries, malaria risk is limited to certain areas.

Illustration 12-8. World map showing the distribution of malaria. (CDC)

P .vivax causes benign tertian malaria (43% of cases) and *P. falciparum* results in malignant tertian or sub-tertian malaria and pernicious malaria (50% of cases). *P .ovale* (mild tertian malaria, 1% of cases) and *P. malariae* (quartan malaria, 7% of cases) contribute a small percentage of malarial cases.

Clinical features including fever and chills are due to the host inflammatory response and are associated with rupture of erythrocytic schizonts. Fever presents in three stages - **a)** Cold: rigors and fever lasting 15 minutes to 1 hour; **b)** Hot: the patient is flushed with tachycardia and is pyrexial (40C) for 2-6 hours; **c)** Sweating: the temperature falls (over 2-4 hours). Each paroxysm lasts 8-12 hours in total. All erythrocytes containing a trophozoite will be destroyed within 48-72 hours. Periodic fever often takes more than seven days to develop, and anemia can be hemolytic or due to toxic marrow suppression. Splenomegaly occurs in all malaria: it may be acute or chronic (+/- hypersplenism). Jaundice may be hemolytic and/or hepatic (only *P. falciparum*). In addition, there may be headache, myalgia, arthralgia, diarrhea and vomiting.

Plasmodium falciparum is the most virulent form (invades mature and immature RBCs) and is often fatal if untreated. Blood schizogony takes place in deep capillaries and micro-circulatory failure can occur in individuals with little immunity to malaria. It does not relapse but recrudescence may occur. The time between paroxysms is 48 hours but fever may last for 24-36 hours. Very rapid progression and complications include diarrhea and vomiting; delirium; coma; convulsions; renal failure, including hemoglobinuria (blackwater fever); jaundice; pulmonary edema; hypoglycemia and abortion. Cerebral malaria often results in delirium, disorientation, stupor, coma, convulsions and death.

P. vivax / ovale exhibit 48 hours between paroxysms; relapses may occur up to 8 years after primary infection and only infects immature RBCs of those with Duffy blood group. *Plasmodium malariae* generally results in 72 hours between paroxysms, only infects older RBCs, and recrudescence may occur decades after primary infection.

The global malaria situation is serious and becoming worse: 300-500 million clinical cases occur annually. 1.5-2.7 million people die of malaria each year with approximately one million deaths among children under five years of age are attributed to malaria alone or in combination with other diseases. Countries in tropical Africa account for more than 90% of the total malaria incidence and the great majority of malaria deaths (WHO data). The death toll of African children with malaria is expected to double by 2010, conceivably reaching four million deaths per year. Many factors influence the epidemiology of this disease including: breeding habits of the various mosquito vectors; agricultural practices; economic conditions; industrialization and pesticide use. Increasing air-traffic from malaria endemic areas has led to the possibility of malaria developing in non-endemic areas where the mosquito vector has been imported onboard aircraft.

Treatments and Control - Malaria

If the infective species is not known, or the infection is known to be mixed, initial treatment should be with quinine, mefloquine or rarely halofantrine. Falciparum (malignant) malaria is often resistant to chloroquine and should be treated with quinine, mefloquine, halofantrine, quinidine or pyrimethamine-sulphadoxine. Benign malaria (*P. vivax*) should be treated with chloroquine although resistance has been reported from New Guinea. Malarial prophylaxis is relative and not absolute.

The UK Consensus Group on Malaria Prophylaxis (1997) recommend mefloquine for UK travelers to West, Central and East Africa for periods of greater than two weeks and for travelers to specific areas within south-east Asia: prophylaxis should be commenced two weeks before departure. Doxycycline can be used in older children and adults who cannot tolerate mefloquine.

Prevention is most dependent upon coverage of exposed skin and the use of insect repellent, mosquito nets impregnated with permethrin and correct prophylaxis. The vector may be controlled by water clearance programs, house spraying (DDT) and destruction of breeding areas. Drug resistance to DDT and ethical resistance to its use have limited its effectiveness. Natural immunity involves both antibody and cell-mediated systems and appears to require frequent boosting; antigens from different stages of the parasite's life cycle will be important in vaccine development.

Disease - Filariasis

Anopheline mosquitoes also transmit the filarial worms *Wuchereria bancrofti, Brugia malayi* and *Brugia timori. Wuchereria bancrofti is* the main cause of "elephantiasis” (Bancroftian filariasis) and the most widely distributed filarial parasite of man. The adults live in the lymphatic system, and can survive for 30 years or more. Once they have mated they produce a pre-larval form, the microfilaria. Both the adults and the microfilaria may play a role in generating the symptoms and signs. Microfilaria measure 240-300μm in length by 7-10μm in width. They are sheathed (derived from ovum membrane) and nuclei terminate 15-20μm proximal to the pointed tail. There are fewer, more distinct nuclei than in other species and there are less body curves. Adult worms are slender and white (males 4cm; females stout and 10cm in length)

Initial infection with *Wuchereria* is usually asymptomatic. There may be recurrence of attacks of "cellulitis" affecting the limbs, breast, scrotum or elsewhere. Infection is associated with fever, lymphangitis, lymphadenopathy and occasionally abscess formation. These initially settle but later on the tissues eventually become edematous and hypertrophied. Further effects may include scrotal involvement and hydrocoele, which can lead to scrotal enlargement and lymph scrotum. This is "elephantiasis" and is associated with dermal hypertrophy, verrucous changes and the rupture of lymph varices into various sites.

Brugian (Malayan) filariasis is less widespread, less common and less serious than its Bancroftian counterpart. The life cycle is identical to that of *Wuchereria bancrofti* with *Brugia malayi* limited to Asia and *B. timori* restricted to Indonesia. Infection results in lymphadenopathy involving most frequently the inguinal area, lymphoedema normally below the knee, eosinophilia, and in rare cases chyluria.

Treatments and Control - Filariasis

Diethylcarbamazine (DEC) kills microfilaria. Ivermectin suppresses microfilaria production but its overall effectiveness remains untried and elephantiasis can be treated surgically. Control measures comprise draining of mosquito breeding sites and killing larvae. Many mosquitoes are resistant to insecticides but mosquito repellents and nets are effective. The infective pool may be reduced by periodic mass treatment with DEC. *Brugia malayi* is more susceptible to diethylcarbamazine (DEC) than is *Wuchereria bancrofti*. Anopheline larvae may be suffocated in their breeding sites but Culicine larvae (*Mansonia* sp.) derive oxygen from plants and are not vulnerable to such measures. Control depends upon the use of mosquito nets and periodic mass treatment.

Culicine (*Aedes*) Mosquitoes

Class: Insecta
Order: Dipthera
Family: Culicidae
Subfamily: Culicinae
Genus: Culicine
Species: *Aedes*

General Characteristics

Aedes can generally be distinguished by patterns of black and silvery scales present on the abdomen and thorax. The legs appear to have black and white rings along their length. The wings are generally covered with black scales. *Aedes* breed in marshes and other wetland areas and have a worldwide distribution.

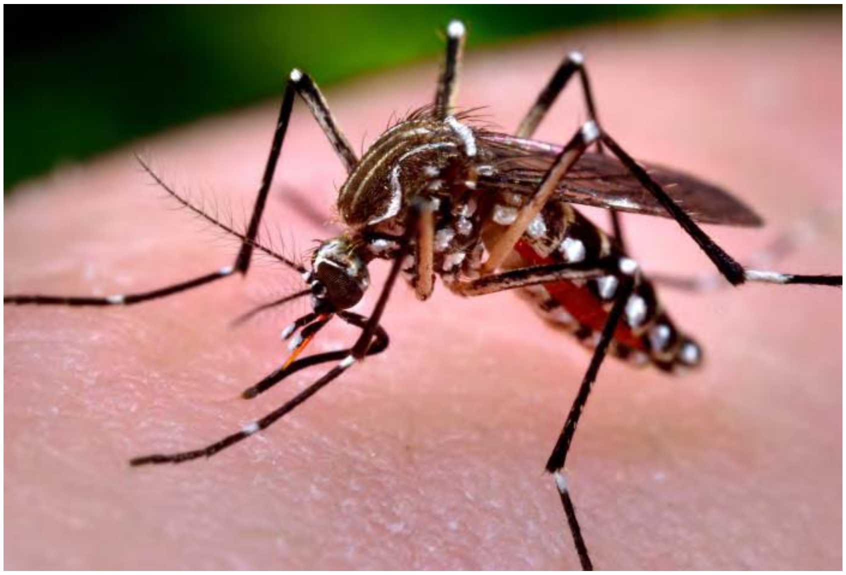

Image 12-17. Typical *Aedes* mosquito. They are clearly distinguished from Anopheline mosquitoes due to the presence of black and slivery scales on the abdomen and thorax. (SOURCE: PHIL 9252 - CDC/ Prof. Frank Hadley Collins, Dir., Cntr. for Global Health and Infectious Diseases, Univ. of Notre Dame)

Life Cycle

Female *Aedes* lay eggs on damp areas such mud, detritus, clay and rock. The eggs are very robust and can survive desiccation and other environmental pressures. The eggs hatch in waves depending on the environmental cues. *Aedes* larvae have a stout barrel shaped siphon with one pair of subventral tufts. There are three pairs of setae on the ventral brush, and large setae are not present on the abdominal segments.

Disease

Aedes are vectors of Bancroftian filariasis and arboviruses such as yellow fever and dengue. *Wuchereria bancrofti is* the main cause of "elephantiasis" (Bancroftian filariasis) and the most widely distributed filarial parasite of Man. The adults live in the lymphatic system, and can survive for 30 years or more. They copulate and generate a pre-larval form, the microfilaria. Both the adults and the microfilaria may play a role in generating the symptoms and signs. Microfilaria measure 240-300μm in length by 7-10μm in width. They are sheathed (derived from ovum membrane) and nuclei terminate 15-20μm proximal to the pointed tail. There are fewer, more distinct nuclei than in other species and there are less body curves. Adult worms are slender and white (males 4cm; females, stout and 10cm in length).

Initial infection with *Wuchereria* is usually asymptomatic. There may be recurrence of attacks of "cellulitis" affecting the limbs, breast, scrotum or elsewhere. Infection is associated with fever, lymphangitis, lymphadenopathy and occasionally abscess formation. These initially settle but later on the tissues eventually become edematous and hypertrophied. Further effects may include scrotal involvement and hydrocoele, which can lead to scrotal enlargement and lymph scrotum. This is "elephantiasis" and is associated with dermal hypertrophy, verrucous changes and the rupture of lymph varices into various sites.

Yellow fever and dengue hemorrhagic fever are serious viral infections spread by the *Aedes* mosquito. Dengue is now the most important mosquito borne virus, with global infection increasing.

Control and Treatment

In general the most effective control for Culicine mosquitoes are repellents and fine screening or netting. Treatment with insecticides will also serve to reduce the vector population, but increased problems are encountered with Culicines because they also feed during the daytime. If filarial infection occurs, treatment with Diethylcarbamazine (DEC) will kill microfilaria. Ivermectin suppresses microfilaria production but its overall effectiveness remains untried and elephantiasis can be treated surgically.

Culicine (*Culex*) Mosquitoes

Class: Insecta
Order: Dipthera
Family: Culicidae
Subfamily: Culicinae
Genus: Culicine
Species: *Culex*

General Characteristics

Culex are distinguished by their lack of coloration and feature. The thorax, abdomen, legs and wings are often covered with brown-black scales giving a generally dark appearance. The abdomen may occasionally also have white scales arranged in segments. *Culex* breeds mainly in aquatic habitats, often in areas containing large quantities of organic waste.

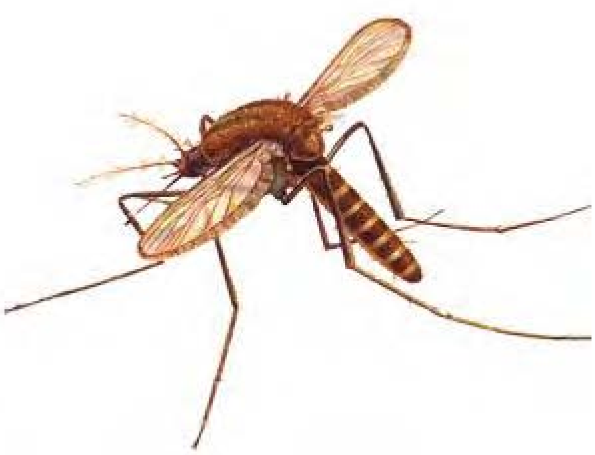

Illustration 12-8. Typical illustration of a *Culex* mosquito, showing the characteristic brown-black scales on the thorax, abdomen, legs and wings giving it a general dark appearance. (Illustration: Trustees of the British Museum)

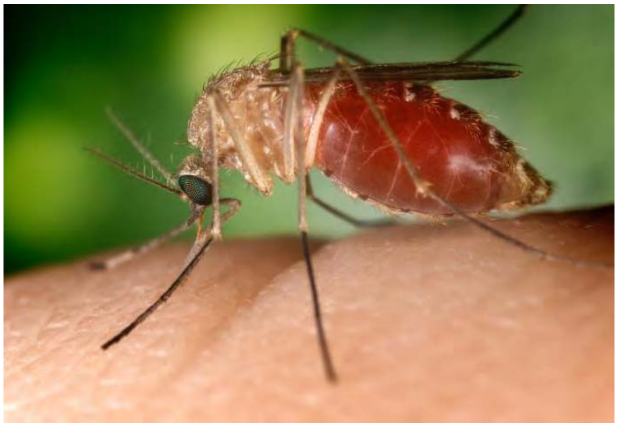

Image 12-18. A *Culex* mosquito taking a blood meal from a human host. The abdomen becomes distended and blood red in color. (SOURCE: PHIL 4464 - CDC/Jim Gathany)

Life Cycle

Female *Culex* lay dark brown eggs in characteristic clumps of approximately 300 eggs. As mentioned these eggs are often found in organic waste deposits or polluted waters. *Culex* larvae have a long and narrow siphon with more than one pair of subventral tufts.

Disease

Culex mosquitoes are vectors of Bancroftian filariasis throughout Africa, but most importantly arboviruses such as Japanese encephalitis. Encephalitis occurs throughout the world, with *Culex* acting as an important vector for spread and infection. *Culex* mosquitoes are similar to Culicine mosquitoes, but prefer to bite at night and breed in organic refuse.

Control and Treatment

Culex mosquitoes are most easily controlled by improving sanitation and removing static water sources from the affected area. In general the most effective control for *Culex* mosquitoes are also repellents and fine screening or netting. Treatment with insecticides will also serve to reduce the vector population, but increased problems are encountered with Culicine mosquitoes because they also feed during the daytime. If filarial infection occurs treatment with Diethylcarbamazine (DEC) will kill microfilaria. Ivermectin suppresses microfilaria production but its overall effectiveness remains untried and elephantiasis can be treated surgically.

Tsetse Flies

Class: Insecta
Order: Diptera
Family: Glossinidae
Genus: *Glossina*

General Characteristics

Tsetse flies are large, yellow-brown or brown-black and measure 6-15µm in length. They are distinguished by a rigid projecting proboscis and a long pair of accompanying palps. There is a characteristic axe-shaped venation of the wings when viewed dorsally resembling an inverted hatchet in the central cell. The antennae appear short and feathery, and the abdomen is segmented and often striped or patched. Both males and females suck blood every 4-5 days, outside in open spaces.

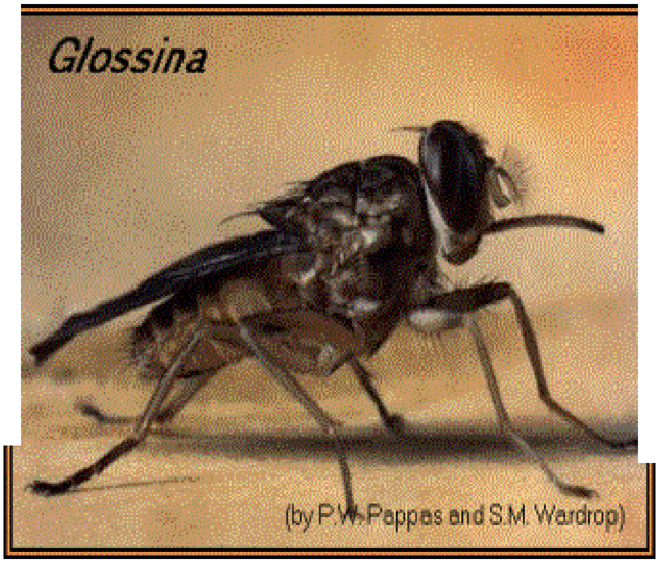

Image 12-19. Typical illustration of a *Tsetse Fly.* They are yellow-brown in Colour and usually measure 6 – 15µm in length. Their characteristic feature is the hatchet shaped cell in the centre of the wing venation. (IMAGE SOURCE: Trustees of the British Museum)

Life Cycle

Female Tsetse flies are unique in the sense that they deposit larvae and do not lay eggs. The eggs mature within the female and are supplied with essential nourishment to complete larval development. This cycle requires a large number of blood meals to maintain and thus the female requires regular feeding. The larvae are normally deposited in shaded areas. There are three larval instars with the mature larva appearing white, visibly segmented with a pair of lobes at the posterior end. Pupation of the third instar results in a dark colored puparium with posterior lobes. The pupal period is extended (3-7 weeks) depending on the surrounding environmental conditions.

Disease

Tsetse flies are vectors for African sleeping sickness (*Trypanosoma brucei sp*. The two subspecies of *T. brucei* that infect Man are morphologically identical. *T. b. gambiense* causes Gambian sleeping sickness in Western Africa and *T. b. rhodesiense* causes Rhodesian sleeping sickness in East Africa. Another subspecies, *T. b. bruc*ei, causes nagana in cattle.

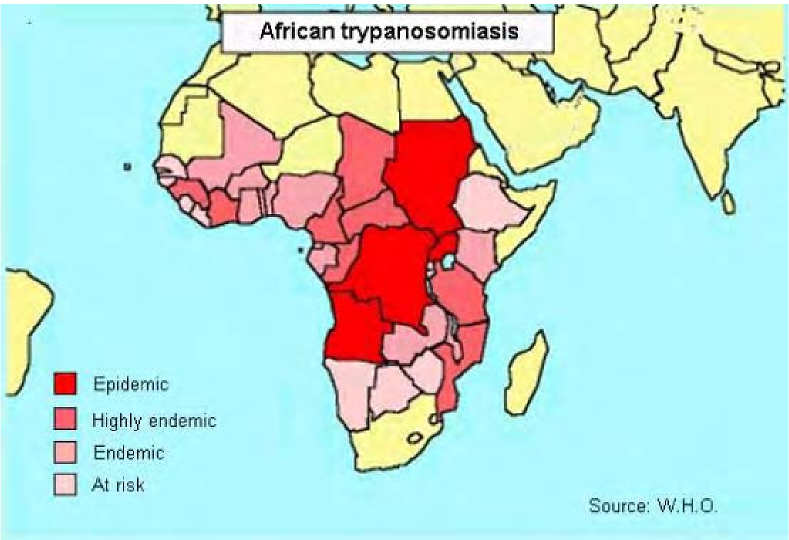

Illustration 12-9. Distribution of African Trypanosomiasis throughout Africa (WHO, 1991)

Initially the patient has a headache, fever, chills and loss of appetite but specific clinical signs are absent. Parasitaemia comes in characteristic waves. Later, the spleen, liver and lymph nodes enlarge (Winterbottom's sign). Finally, there may be CNS involvement leading to coma and death within several years. *Trypanosoma brucei rhodesiense* (rural East Africa): this is so acute that the patient invariably dies before classical symptoms develop. Without appropriate treatment, both forms are fatal.

Human African trypanosomiasis is rural and focal, with humans as the principal reservoir of infection of *T. b. gambiense*, and domestic cattle and wild animals as important reservoirs of *T. b. rhodesiense*. By the 1960s, it had been brought under control, but since 1970 the situation has deteriorated and the disease has reappeared, with major flare-ups in countries which have not maintained surveillance activities. It is estimated that 55-60 million people are exposed to the risk of becoming infected with trypanosomiasis, but only four million of them are under active surveillance or have access to health centers where reliable diagnosis is available: the estimated number of infected persons is over 300,000 (WHO data).

Treatment and Control

Anemia and other infections should be treated first. If given before the parasite has invaded the brain (hemolymphatic stage), suramin (Rhodesian disease) or pentamidine (Gambian disease) appear to be effective. For late disease (CNS involvement), drugs of choice are melarsoprol or eflornithine, with tryparsamide plus suramin as an alternative. Tsetse-infested areas should be avoided. Protective, light-colored clothing and repellents should be used. Tsetse flies are difficult to treat with insecticide (DDT has been tried) as their larvae live in burrows in the ground; protective vegetation must first be removed. Pentamidine prophylaxis is no longer advocated. Reservoir hosts should be identified and removed. It is possible to breed resistant or tolerant cattle (West Africa). Trypanosomes other than *T. brucei* are also transmissible by Tabanid flies or by sexual contact.

Sand Flies

Class: Insecta

Order: Dipthera

Family: Psychodidae

Genera: *Phlebotomus, Lutzomyia*

Introduction

Sandflies are small (1.5-5µm) with a hairy head, thorax, antennae and wings. The antennae are long, may appear beaded, and protrude near a large set of black compound eyes. The wings are upwardly pointing at rest, and are a distinctive feature of Phlebotomine sandflies. Only the females are blood feeders, the males feed on plant nectar. Biting predominates nocturnally and they rest in moist and dark walls, cracks and tree trunks during the day.

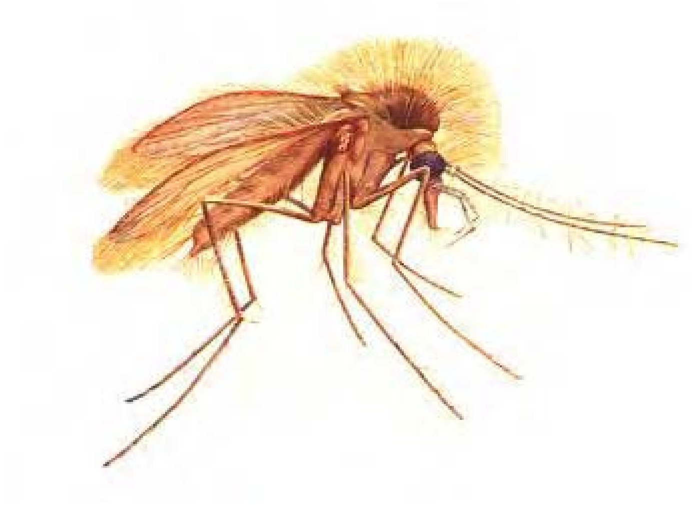

Illustration 12-10. Typical illustration of a Sand Fly. The head, thorax, abdomen, legs and antennae are hairy, they are small usually measuring 1.5-5µm long. (*Phlebotomus)* (Illustration Source: Trustees of the British Museum)

Life Cycle

Female Phlebotomine sandflies deposit 30-70 minute eggs at each oviposition. The eggs are laid in dry areas, but require humidity to avoid desiccating. There are four instars and the mature larva is characterized by a distinct black head, 12 segments, thick bristles covering the body, and two pairs of caudal setae on their posterior end. As with other fly vectors the length of the larval stage depends on the ambient temperature, species in question, and food availability. The larval skin and caudal bristles remain attached at the posterior end during the pupal stage which lasts 5-10 days.

Disease

Sandflies are the only vectors for several species and subspecies of obligate intracellular protozoa responsible for leishmaniasis (*Leishmania sp.*) Cutaneous, mucocutaneous and visceral leishmaniasis are caused by different species of *Leishmania* contentiously linked to temperature preferences. Geographic location and host immune response also play a role in determining the form of disease. *Leishmania tropica* and *L. major* cause dermal cutaneous leishmaniasis; visceral leishmaniasis (kala-azar) is caused by *L. donovani* and mucocutaneous leishmaniasis is caused by *L. braziliensis* and *L. mexicana*. Visceral leishmaniasis in Europe is caused by *L.infantum* with dogs as the main reservoir. *Leishmania tropica* is found in the Middle East, North Africa, India and the Mediterranean. *Leishmania donovani* is found in the old and new worlds: South America, the Mediterranean, North and East Africa, India and China*.* *Leishmania braziliensis* is found in Central and South America. *Leishmania mexicana* is found in North and Central America, Texas and Mexico.

Dermal cutaneous leishmaniasis or Old World leishmaniasis is also known as "Tropical or Oriental Sore" or "Delhi Boil" and is generally localized to the skin surrounding the bite of the *Phlebotomus* sandfly. The ulcers (volcano sign) are "draining" but produce "dry" ulcers that crust over. They usually resolve within one year but superinfection (yaws or myiasis) may occur and immunity develops. Cutaneous leishmaniasis affects over 300,000 people. The mucosal form affects the mucosa of the nose, pharynx, palate, larynx and upper lip causing ulcers that often become secondarily infected. Scarring can lead to death from pneumonia.

Visceral leishmaniasis, or "Kala-azar” is transmitted by Phlebotomus sandflies and is spread via the lymphatics from an often minor cutaneous lesion. They multiply in macrophages to form Leishman-Donovan bodies. Symptoms are usually chronic and comprise malaise, lymphadenopathy, cough, diarrhea, wasting and anemia, bleeding, and low-grade fever (3/day); liver and spleen enlarge and visibly distend the abdomen. Untreated, death ensues within three years, usually from secondary infection.

Leishmaniasis currently affects some 12 million people in 88 countries, all but 16 of which are in the developing world. It is estimated that 350 million people are exposed to the risk of infection by the different species of Leishmania parasite. The annual incidence of new cases is about two million (1.5 million of cutaneous leishmaniasis, and 0.5 million of visceral leishmaniasis). Recently, the WHO has reported an increase in overlapping of visceral leishmaniasis (VL) and HIV infection due to the spread of the AIDS pandemic. Leishmania / HIV co-infection is considered to be a real "emerging disease", especially in southern Europe, where 25-70% of adult VL cases are related to HIV infection, and 1.5-9.5% of AIDS cases suffer from newly acquired or reactivated VL. Intravenous drug users have been identified as the main population at risk.

Treatment and Control

Dermal leishmaniasis usually resolves spontaneously. Visceral leishmaniasis is treated with extended courses of antimonial compounds (e.g. sodium stibogluconate or meglumine antimonate) and dietary supplementation (pentamidine isethionate has been used in antimony-resistant cases). WHO regimes are changing and pentamidine may be used. Mucocutaneous leishmaniasis is also treated with antimonial compounds. However, only around 50% of patients respond to antimonial compounds and relapses are seen. Pentamidine isethionate, paromomycin (aminosidine), allopurinol, ketoconazole, itraconazole, interferon gamma and liposomal amphotericin B have all been used; amphotericin B appears the most effective. Control of sandflies is difficult although buildings may be sprayed with insecticide. Sandflies cannot bite through clothing. Repellents are effective although mosquito nets are of limited value (sandflies are 3 µm). Sandflies are nocturnal and can be avoided. Rodent and dog control is a possibility. Vaccines (killed or live attenuated promastigotes) have been tried but effectiveness has not been assessed.

Black Flies

Class: Insecta
Order: Diptera
Family: Simuliidae
Genus: *Simulium*

Introduction

Black Flies are small (1.5-4µm in length) and normally black with short hairless legs and antennae. They have large compound eyes and a characteristically hairy humped thorax. Flies of the genus *Simulium* are generally found near free-flowing well-oxygenated water and bite during the day, tearing the skin to reach blood vessels.

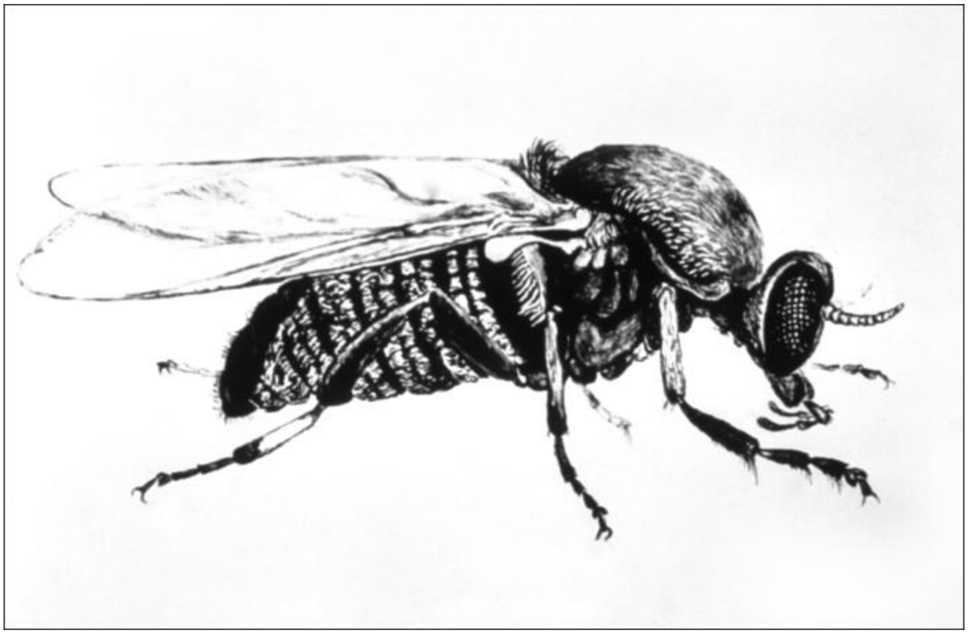

Illustration 12-11. Typical illustration of a Black Fly (*Simulium).* They are usually small (1.4–4µm in length), black with short hairless legs and antennae. (SOURCE: PHIL 4639 - CDC)

Image 12-20. Image of a Black fly receiving a blood meal. (SOURCE: Unknown)

Life Cycle

Simuliidae eggs are laid in flowing or turbid waters in clusters of 200-800 eggs depending on the species. Hatching is mediated by the external environment, generally occurring several days following oviposition. There are 6-9 larval instars, with the larvae remaining essentially sedentary for the period and feeding through filtration. Movement is possible if necessary with some larval species acting as predators. The mature larvae can be recognized by a characteristic black “gill spot” on the thorax. The pupa is enclosed in a dark colored cocoon, identifiable by a series of filamentous respiratory gills protruding from the anterior end. In most cases the adult fly emerges from the pupal stage following one week of incubation.

Disease

Simulium are vectors for the filarial parasitic nematode responsible for Onchocerciasis (*Onchocerca volvulus).* It is distributed throughout Africa, Arabia, Central America, northern South America and Mexico with 30 million people infected in Africa alone; it is one of the most important causes of blindness in the world. Adult worms (up to 50cm) live in the subcutaneous tissue of man (the only known host) and release microfilariae into the skin. These are taken up by *Simulium* (*S. damnosum* and *S. naevei* in Africa; *S. ochraceum* and *S. metallicum* in Central America). Microfilariae penetrate the gut wall, migrate to the thoracic muscles, molt to L_2 stages and molt again to filariform L_3 stages which are passed to the next host during a blood meal. Larvae mature to adults in one year and may live for up to 20 years. Microfilariae can live for up to two years but are often trapped, in fibrous nodules, by the host's cellular response and are responsible for the pathology. Microfilariae are highly motile, unsheathed and measure 300 by 7 microns; there are no nuclei in the end of the tail, which is long and pointed; the head is slightly enlarged. Microfilariae are not found in blood.

Skin problems include nodules (2cm to 5cm in diameter) containing adult worms. Onchodermatitis is caused by death of microfilariae in the skin and is associated with severe pruritis (adopts different forms in different geographical localities e.g. "erysipelas de la costa" in Central America, "leopard skin" in parts of Africa), skin depigmentation and premature ageing. There may be minor elephantiasis of the inguinal area ("hanging groin") or of the genitals (including hydrocoele). Eye involvement is due to microfilariae entering the eye and dying and may eventually lead to blindness through sclerosing keratitis and retinal damage (microfilaria can be seen in the anterior chamber of the eye). There may be competition between parasite and host for vitamin A.

Treatments and Control

Ivermectin destroys microfilariae and renders adult female worms infertile; the manufacturer provides this drug without charge. Nodules may be surgically removed to reduce the worm-load and decrease the chances of ocular involvement. In areas of East Africa, *S. naevei* has been eradicated with insecticide (the larvae of this parasite attach to freshwater crabs). *S. damnosum* inhabit areas far distant to its breeding sites and control is much more difficult. However, the WHO larvicide program in concert with mass Ivermectin treatment has the potential to control the problem. Vegetation should be cleared around villages and irrigation should be devised to avoid fast-flowing water currents.

Tabanid Flies

Class: Insecta
Order: Diptera
Family: Tabanidae
Subfamily: Chrysopsinae (Deer flies)
Genus: *Chrysops*
Subfamily: Tabaninae (Horse flies)
Genus: *Tabanus*

Introduction

Chrysops are 9-10µm in length with broad wings and large iridescent eyes. The abdomen is yellow or orange and can be patterned with black markings. The antennae consist of three segments, lacking a projection from the second segment and with the third segment subdivided into four smaller sections. Female *Chrysops* are attracted by smoke and normally bite in the morning or late afternoon. They breed in shaded muddy areas and have a worldwide distribution.

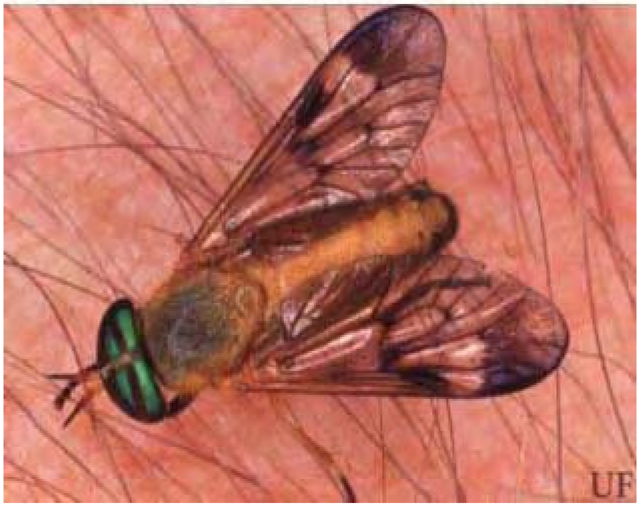

Image 12-21. Image of a Deer fly receiving a blood meal. They have large iridescent eyes and broad wings. Generally measure between 9–10µm in length. (SOURCE: University of Florida)

Life Cycle

Female Tabanids generally lay 100 -1000 large creamy white eggs on the undersides of plants and rocks surrounding aquatic areas. The eggs hatch following several weeks of incubation. Larvae live and feed in wet rotting vegetation but survive breathing oxygen. The larval stage can be quite prolonged (1-3 years) with 6-13 larval instars depending on the species. The mature larvae migrate to dry areas and undergo pupation. The pupa often gets buried in the soil and can often be large. The pupal stage lasts several weeks, and the adults emerge to feed.

Disease

Tabanids are vectors for the parasitic nematode *Loa loa.* Their Microfilariae are large and sheathed and contain nuclei extending to the end of the rounded tail. Adult worms are thin and white (females 2cm-7cm long, 425µm in diameter and males 2cm-3.4cm long, 350µm in diameter). Clinical features of *Loa loa* infections include considerable pruritis and transient painful subcutaneous swellings termed Calabar swellings. Migrating adult worms can cross the conjunctiva or the bridge of the nose and ectopic worms cause problems such as hydrocoele, orchitis, colonic lesions, encephalitis.

Treatment and Control

Worms traversing the conjunctiva may be removed surgically. Diethyl-carbamazine (DEC) is a proven treatment; prevention comprises the use of prophylactic DEC. However, DEC may have fatal side effects including encephalitis and the current drug of choice is Ivermectin. Drugs kill microfilariae but not adult worms. Antihistamines and corticosteroids may prevent allergic reactions brought about by rapid destruction of microfilariae in heavy infections. Insect control is not practical.

Biting Midges

Class: Insecta

Order: Dipthera

Family: Ceratopogonidae

Genus: *Culicoides*

Introduction

Biting midges are 1µm-2µm long with a small head, long antennae and segmented palps. The thorax is often black spotted and contains a distinctive set of small depressions called the "humeral pits" just posterior of the head on the upper thorax. Biting midges have long legs and wings that fold over the thorax when at rest. Only the females take blood meals and normally swarm and bite in the early morning or late evening, especially during overcast weather.

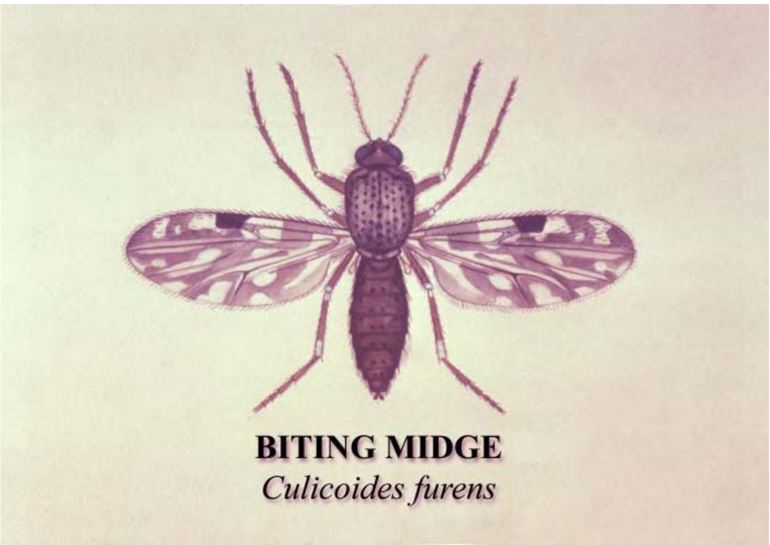

Illustration 12-12: Typical Illustration of a Biting Midge (*Culicoides*). The thorax is often black spotted and contains a distinctive set of small depressions called the "humeral pits" just posterior of the head (SOURCE: PHIL 3822 - CDC/Dr. Richard Darsie)

Life Cycle

Female culicoides lay 30-250 dark, cylindrical eggs on the surface of wet soil or organic debris. Larvae emerge from the eggs soon thereafter depending on the environmental conditions. There are four larval instars, and the mature larvae resemble that of a nematode worm. They have a small dark head, 12 segments, and terminal papillae. The larvae feed on detritus and may development for extended periods of time depending on the species and conditions. The pupal period lasts 3-10 days, with the adult females emerging to feed and breed.

Disease

Culicoides species possess the ability to transmit various filarial parasites to humans (*Mansonella perstans*, *M. streptocerca* and *M. ozzardi*). Adult *Mansonella streptocerca* reach 3 cm in length and live in the skin. Microfilariae are small and thin (200µm in length), unsheathed and nuclei extend to the end of a hooked tail. The major symptoms of infection are pruritis, papular eruptions and pigment changes.

Mansonella ozzardi is found in South and Central America and the adult worms live in the peritoneal cavity. Microfilariae are small, thin (150-200 by 4.5 microns) and unsheathed with indistinct nuclei that do not extend to the end of the pointed tail. They are found in the skin and blood and are non-periodic. Most infected persons are asymptomatic but may be associated with arthralgias, headaches, fever, pruritis, hepatomegaly, pulmonary symptoms and adenopathy: there is no proven therapy.

Mansonella perstans is found in tropical Africa and coastal South America with adult worms measuring 4cm-8cm. They inhabit the peritoneal and, rarely, pleural cavities. Infections are usually asymptomatic but can be associated with pruritis, fever, arthralgias and neurological changes.

Head and Body Lice

Class: Insecta
Order: Phthiraptera
Genus: *Pediculus*

Introduction

Pediculus humanus capitis (the head louse) and *Pediculus humanus humanus* (the body louse) are 2µm-4µm long and flattened dorsoventrally. They are wingless with distinct head, thorax (bearing three pairs of clawed legs) and abdomen (seven segments). The fore legs are well developed to grasp clothing and hair, with terminal claws to aid grip on the host. *Pediculus* feed using two stylets which suck blood whilst a third directs saliva into the skin; meanwhile feces are continually passed onto the skin.

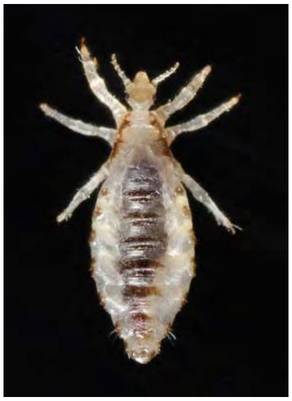

Image 12-22. Dorsal Image of a typical body louse (*Pediculus humanus humanus*). They are wingless and dorsoventrally flattened, bearing a distinct head, thorax and abdomen. They measure approximately 2–4µm in length. (SOURCE: PHIL 9205 - CDC/ Frank Collins, Ph.D.)

Life Cycle

The head and body louse share very similar life cycles only differing in the placement of their eggs; the body louse cements eggs to clothing and the head louse cements single eggs at the base of hairs. Females can lay upwards of 300 eggs during a lifetime. Lice have a hemimetabolous lifecycle. The nymph hatches from the egg and appears very structurally similar to the adult louse. There are three nymphal instars that require blood meals to proceed. It only takes 7-12 days to proceed to the adult stage if blood meals are readily available, but conversely lice will perish if not fed for several days.

Image 12-23. Dorsal Image of a typical head louse (*Pediculus humanus capitis)* (SOURCE: PHIL 377 - CDC/ Dr. Dennis D. Juranek)

Disease

Colonization by lice may result in serious infection and disease. Lice can potentially pass *Rickettsia prowazeki* resulting in epidemic typhus and other pathogens such as *Rochalimaea quintana* and *Borrelia recurrentis* that result in potentially serious fever. The mode of transmission for all infectious agents is through physically crushing and spreading feces or waste into the wound created through feeding.

Control

The most effective control is to remain clean, but insecticides are often necessary during epidemics to avoid reinfestation.

Pubic Lice

Class: Insecta
Order: Phthiraptera
Genus: *Phthirus*

Introduction

The crab louse is 1µm-2µm long and distinguished by a square, undifferentiated body and massive claws on the two posterior sets of forelegs. These claws are able to grasp both pubic and facial hair (including eyelashes), and allow the louse to remain tightly bound to the host. They are spread mostly by sexual contact, but may also be transmitted through fomites.

Life Cycle

The life cycle of *Phthirus* is very similar to *Pediculus*. Females lay bundles of eggs on the coarse pubic hairs and dense facial hairs of humans. The crab lice proceed through a cycle similar to the head and body lice, with the nymphal stage proceeding several days longer. *Phthirus* are less active than *Pediculus,* but similarly can not survive for very long without a host and blood meals.

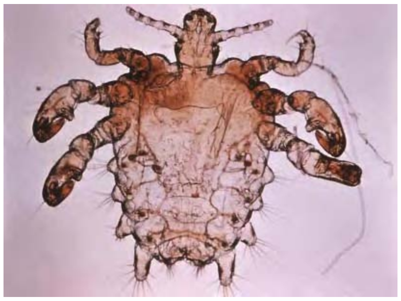

Image 12-24. Ventral Image of a typical Crab louse (*Phthirus*). They hold onto pubic hair with the large claws found on the posterior legs. (SOURCE: PHIL 4077 - CDC)

Disease

There appears to be very little evidence of disease transmission by *Phthirus,* but have the ability to cause severe localized allergic reactions during infestations.

Fleas

Class: Insecta

Order: Siphonaptera

Genus: *Pulex, Xenopsylla, Ctenocephalides, Tunga*

Introduction

Fleas are laterally compressed and wingless (1µm-4µm) with powerful legs. The entire body is generally covered with bristles, and the mouthparts point downwards. *Pulex irritans* (the human flea) and *Xenopsylla cheopsis* (the tropical rat flea) are combless. *Nosopsyllus fasciatus* (the rat flea) has a pronotal comb (behind its head). *Ctenocephalides felis* (the cat flea) and *C. canis* have two combs - a pronotal comb and a genal comb (under the head).

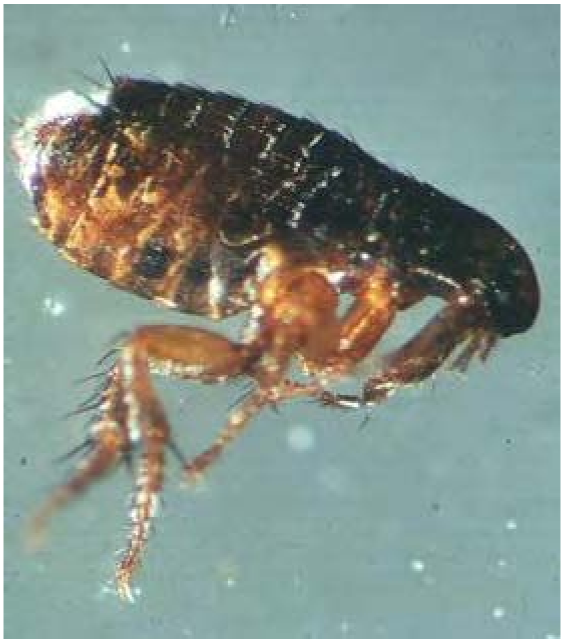

Image 12-25. Image of the cat flea *Ctenocephalides felis*. Fleas are laterally compressed and wingless (1-4 µm) with powerful legs. (SOURCE: CDC)

Tunga penetrans (the Chigoe or jigger flea) demonstrates compressed thoracic segments and attacks man in the Americas, Africa and India, commonly penetrating the stratum corneum between the toes or in the toenail margins

Image 12-26. Image of the Chigoe flea *Tunga penetrans*. They commonly penetrate between the toes or under the toenail. (SOURCE: Unknown)

Life Cycle

Flea larvae hatch from eggs generally following a week incubation. The larvae are legless, segmented, and covered with setae. The larvae feed on organic material, and proceed through two or three larval instars depending on the environment. The larva spins a cocoon and pupates and emerges when the conditions appear favorable for survival. The life cycle can be as short as several weeks or up to several years depending on the stimuli surrounding the pupa. Both sexes take blood meals can live for long periods of time allowing females to lay an enormous number of eggs over their lifetime.

Disease

Fleas are a general nuisance, often biting humans on exposed surfaces resulting in discomfort. Flea-bites induce pruritic papular urticaria commonly on the unprotected lower leg of women and all over the body of children who have intimate animal contact; a generalized allergic response may occur.

Certain fleas, notably the rat fleas, spread plague (*Yersinia pestis*) and murine typhus (*Rickettsia typhi*), and serve as intermediate hosts for species of tapeworm (*Hymenolepis* sp*.)*. Cat and dog fleas serve as intermediate hosts for another common tapeworm (*Dipylidium caninum)*, which can be spread to humans, especially children with exposure to pet animals. *Pulex irritans* is not a major vector of disease but may play a minor role in the transmission of plague. Infection is often spread by the bite alone, but can also potential be transmitted through fecal abrasion. *Tunga penetrans* does not transmit disease to humans, but females will burrow into host skin. The pinpoint lesion enlarges to pea-size within two weeks necessitating removal of the gravid female using a pin, a needle or a sliver of bamboo. This may potentially lead to a secondary bacterial infection.

Control and Treatment

Control of fleas is generally mediated through insecticidal powders and aerosols. If outbreaks of murine typhus or plague occur steps to control the rodent populations in the affected area may be employed.

Bedbugs

Class: Insecta
Order: Siphonaptera
Genus: *Cimex*

Introduction

Common bedbugs (*Cimex lectularius*) are 3µm-7µm long, wingless, and flattened dorsoventrally. They have long legs, clearly segmented antennae and abdomen, and a distinctive set of compound eyes. Bedbugs are characteristically pale brown but swell and turn to black-red when engorged with blood. Both sexes infest clothing, beds and laundry and pierce the skin with an elongated proboscis to feed.

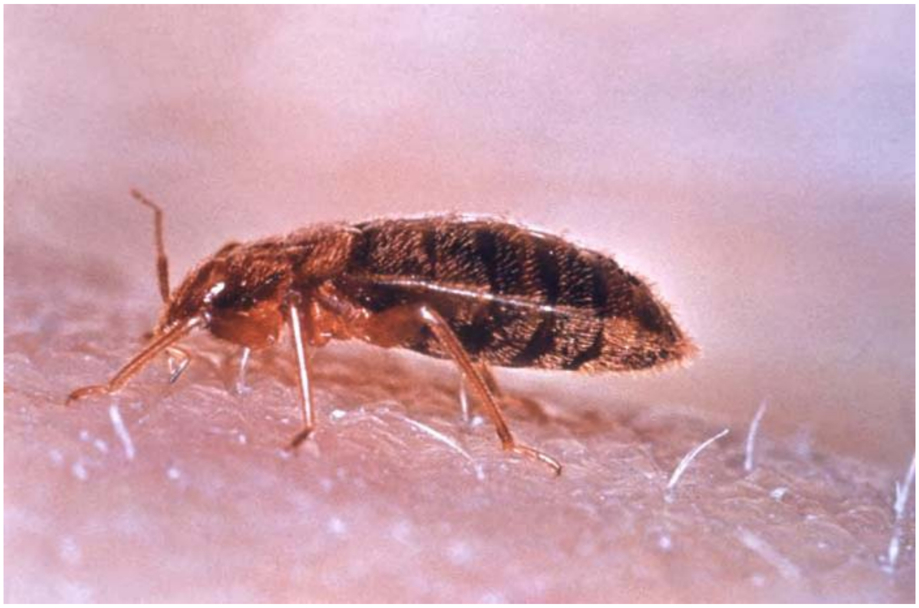

Image 12-27. Image of the common Bedbug *Cimex lectularius.* Bed bugs are usually pale brown in color but swell and turn black-red after a blood meal. (SOURCE: PHIL 6283 - CDC/Donated by the World Health Organization, Geneva, Switzerland.)

Life Cycle

Cimex nymphs appear yellow and very similar structurally to the adults. The life cycle is hemimetabolous with five nymphal instars that require blood meals to proceed. The adults primarily feed nocturnally, and females may lay upwards of 500 eggs during their lifetime. Life cycle duration and the number of offspring produced are primarily determined by the humidity and temperature of the surrounding environment.

Disease

Cimex blood meals result in inflammation, irritation and intense itching. These are the most common results of an infestation, but *Cimex* have also been found to carry Hepatitis B in India.

Treatment and Control

A female lays around 200 adherent eggs at the rate of 3 or 4 per day: eggs are white and 1mm in length. Control is by application of insecticide to mattresses and crevices where they hide during the daytime (to a height of several feet from the floor).

Triatomine bugs

Class: Insecta
Order: Hemiptera
Family: Reduviidae
Genus: *Triatoma*

General Characteristics

Triatomine bugs (a.k.a., Assassin bugs, Kissing bugs, Cone-nosed beetles, Walapai Tiger, or Reduviid bugs) are of variable size but are often large (10mm-30mm). They typically appear brown-black, but can have bright coloration. They are distinguished by a large snout with dark compound eyes, a thin and straight proboscis, and a triangular pronotum. They have long slender legs with terminal claws and segmented antennae. A set of large wings covers the oval abdomen. They live in the mud-walls of housing and woodpiles and come out to feed at night when the host is asleep: they are voracious biters, often on the face around the eyes**.**

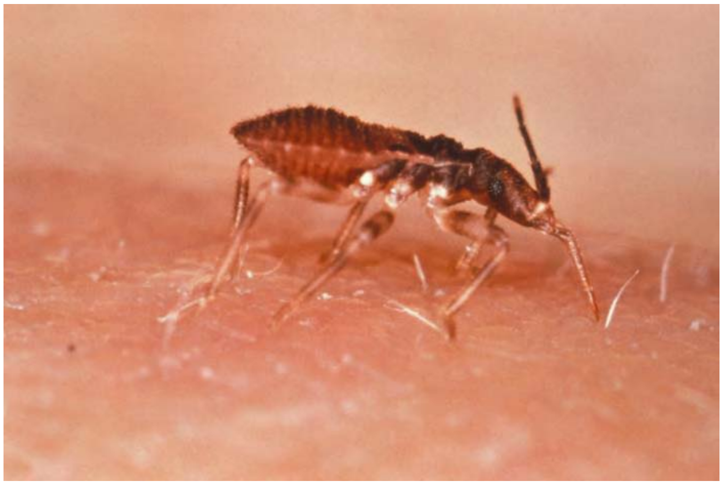

Image 12-28. Image of the Triatomine Bugs *Triatoma infestans.* Reduviid bugs are of variable size but are often large (10-30 mm). They typically appear brown-black, but can have bright coloration. (SOURCE: PHIL 2538 - CDC/World Health Organization.)

Life Cycle

The life cycle of *Triatoma* is hemimetabolous. The emergence of nymphs from eggs is primarily determined by the environment, and blood meals are required to proceed through the five nymphal instars. The nymphs and adults feed nocturnally and generally lay 50 to 1000 eggs depending on life term and the quantity of blood meals taken.

Disease

The primary disease transmitted by *Triatoma* is Chaga's Disease caused by *Trypanosoma cruzi.* Chagas disease is endemic in 21 countries; around 100 million people in Central and South America could be exposed to reduviid bugs; the prevalence of Chaga's disease is about 16-18 million. Rural migrations to urban areas during the 1970s and 1980s changed the traditional epidemiological pattern of Chaga's disease: it became an urban disease, as unscreened blood transfusion created a second way of transmission. Between 1960 and 1989, the prevalence of infected blood in blood banks in selected cities of South America ranged from 1.7% in Sao Paulo, Brazil to 53% in Santa Cruz, Bolivia, a percentage far higher than that of hepatitis or HIV infection (WHO data).

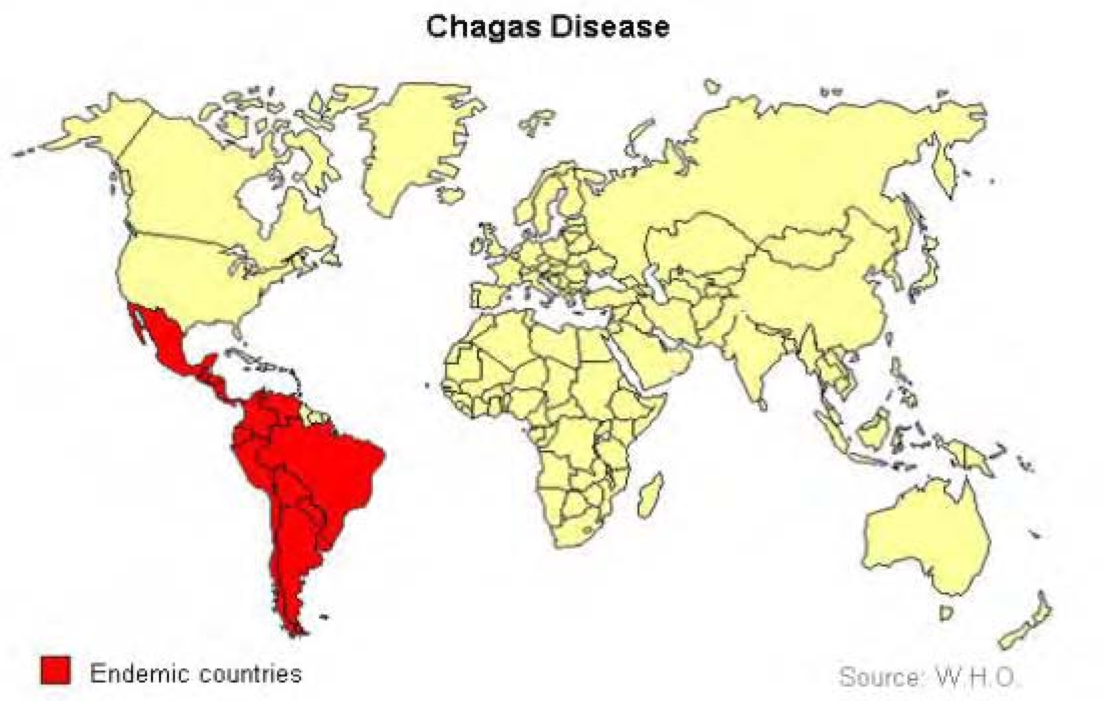

Illustration 12-13. Worldwide distribution of Chagas Disease (*Trypanosoma cruzi*) (SOURCE: WHO)

Triatomids are blood feeders and deposit infected feces (containing metacyclic trypanomastigotes) on the host's skin during feeding. The host rubs the feces into the wound; alternatively trypanosomes enter through the mucous membranes or conjunctivae. Once in the bloodstream, trypanomastigotes penetrate the spleen, liver, lymph nodes and muscle by an unknown mechanism (possibly via receptor-ligand binding proteins). The parasite transforms to the amastigote form that divides by binary fission and forms a pseudocyst. This ruptures and released amastigotes transform to flagellated trypanomastigotes via promastigotes and epimastigotes and enter the bloodstream. Unlike African bloodstream trypomastigotes, these do not replicate. If a Reduviid bug eats trypanomastigotes, they transform to epimastigotes, replicate by binary fission, and are passed as trypanomastigotes two weeks later. *Trypanosoma cruzi* does not exhibit antigenic variation but can persist for the life of the host. *T. cruzi* may also be transmitted in blood products or transplacentally. Humans and a large number of species of domestic and wild animals constitute the reservoir, and the vector bugs infest poor housing and thatched roofs.

In the acute phase (generally seen in children) a small red nodule (Chagoma) may form at the site of the bite. Romana's signs comprise fever, unilateral bipalpebral edema and ophthalmia. Most acute phases are asymptomatic and resolve to an asymptomatic chronic stage in 2-3 months. However, the patient may present with myalgia, bone pain, fever and chills, lymphadenopathy, hepatosplenomegaly and anorexia. Young children may develop meningoencephalitis. Pseudocysts form in heart muscle and fatal heart failure arises within one month. Neurological disorders, gastrointestinal disease and chronic myocarditis are features of chronic disease (mainly seen in adults) which arises 10-20 years after the initial infection and is untreatable. Congenital infection may lead to abortion, still birth or acute disease. Reactivation occurs in HIV-infected patients. Diagnosis is by detection of trypanosomes in peripheral anti-coagulated blood, CSF or node aspirate using wet film or Giemsa stain (acute disease) or by serology (Chaga's IgG ELISA in chronic disease). *Trypanosoma cruzi* is usually C-shaped, 12-30µm in length with a narrow membrane and a flagellum. The posterior kinetoplast is characteristically large and the nucleus is central.

Control

There is no satisfactory treatment. Extended therapy with the toxic agents, benznidazole or nifurtimox, only kills extra-cellular parasites and optimum efficacy (60%) is achieved during the acute phase. Chronic disease is treated symptomatically. Control is by improvements in housing, health education and insecticidal treatment of dwellings. Vaccines and chemoprophylaxis are ineffective (associated with autoimmune disease). For the control of blood-transmitted infections the aim is to screen all blood donors from endemic countries for *T. cruzi* antibodies, and to strengthen health service infrastructures for multiple blood screening (HIV, Hepatitis B and *T. cruzi*).

Chapter 13.

Artifacts and Confounders

Artifacts and Confounders

Introduction and Summary

Cysts and trophozoites must be examined carefully in different fields of view and measurement is often essential. Objects such as epithelial cells and macrophages are around the same size as amoebic trophozoites: the latter may also move and contain red blood cells. White blood cells, plant and vegetable cells, fat globules, muscle fibers, pollen grains, yeasts cells and air bubbles may be confused with cysts or eggs. Air bubbles trapped under adhesive tape often resemble *Enterobius* eggs. Plant hairs and fibers are easily confused with larvae; algae such as *Psorospermium haeckelii* may be found in the feces of patients who have eaten crayfish. Earthworms may resemble roundworms. A variety of non-pathogenic ova, cysts and parasites resemble pathogens in terms of size and morphology and careful examination is essential. Eggs of *Heterodera*, a parasitic nematode of root vegetables, may resemble hookworm eggs. Eggs originating from harmless mites in cereals or flour could be confused with hookworm ova but are usually larger. We recently encountered 160 micron "Schistosome ova" in the urine of a patient complaining of hematuria: we suspected *Schistosoma haematobium* but, on closer analysis, the eggs contained unidentified insects. This volume provides examples of artifacts that may be confused for parasitic life stages. Artifacts should be considered on the basis of size, shape, lack of organelles and defining feature, and variable reactivity with common stains.

Red and White Blood Cells

Red blood cells and a variety of white blood cells can be easily mistaken for parasitic cells or cysts when observed with microscopy.

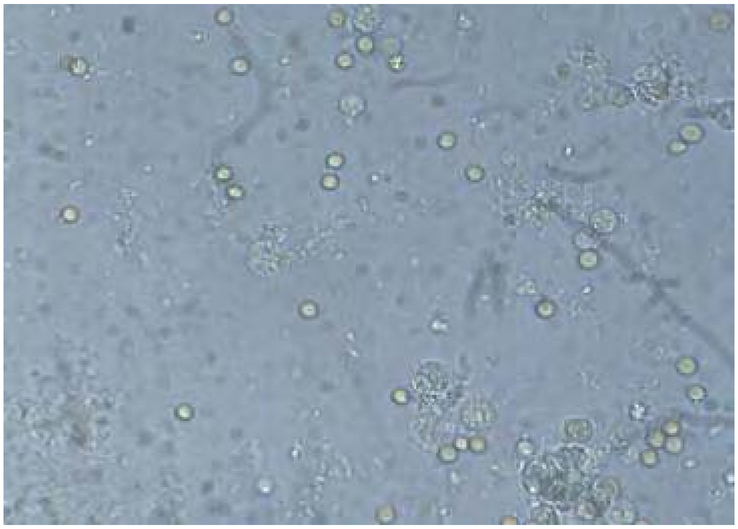

Image 13-1. Image illustrating red and white blood cells in a slide preparation. (SOURCE: Unknown)

White Blood Cells

Charcot-Leyden crystals are a product of eosinophil breakdown and are, therefore, occasionally found in feces of patients suffering from parasitic disease. They appear red when stained with in a trichrome fecal preparation.

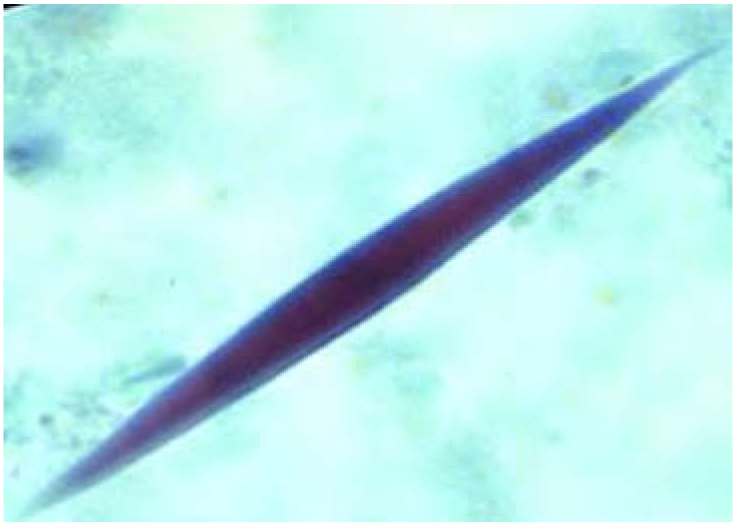

Image 13-2. Image illustrating trichrome stained Charcot-Leyden crystals in a slide preparation (Image courtesy of: The University of Alberta)

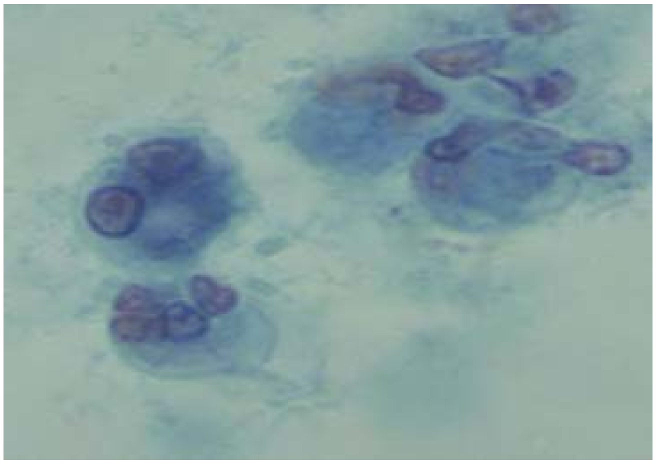

Image 13-3. Image showing trichrome stained Macrophage in a slide preparation. Note similarity to Amoeboid structure. (SOURCE: Unknown)

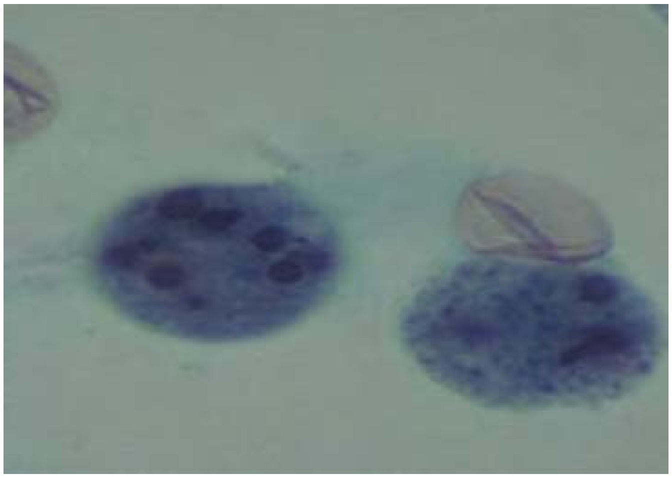

Image 13-4. Image illustrating trichrome stained Leukocyte in a slide preparation. Note similarity to Amoeboid cyst structure. (SOURCE: Unknown)

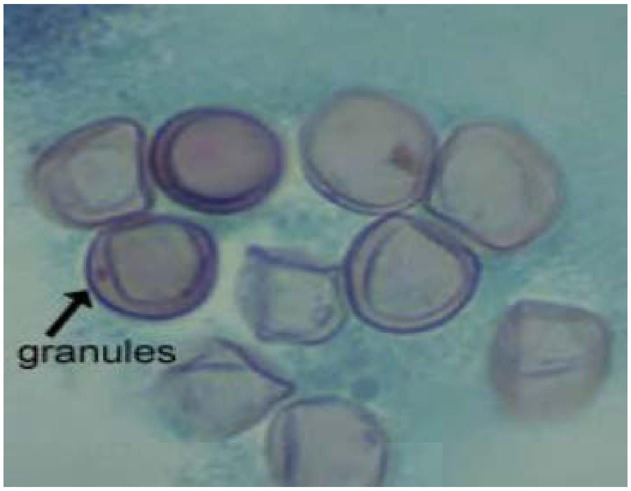

Image 13-5. Image illustrating Red Blood Cells in slide preparation. RBC's may appear to have a central body and a rim of cytoplasm or granules which could be mistaken for *Blastocystis hominis*. (SOURCE: Unknown)

Fat Globules

Fat globules present in a fecal slide preparation may appear similar to parasitic cysts or cell bodies.

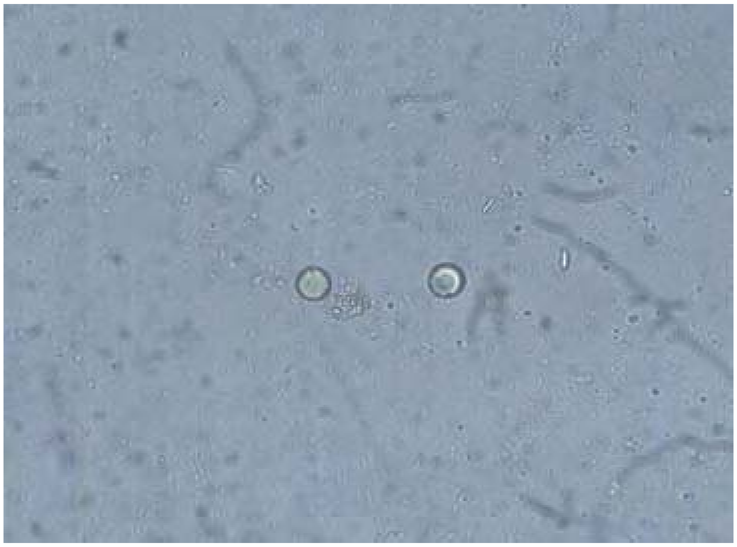

Image 13-6. Image illustrating Fat Globules in slide preparation. (SOURCE: Unknown)

Emulsifying agents are a useful tool to eliminate potential confusion involving fat globules. The removal of such particles from slide preparations will undoubtedly reduce cases of misdiagnosis.

Yeast Cells

Yeast may resemble protozoan cysts because they are uniform in color, have few inclusions and no nucleus. Yeast could also be confused with small protozoans like *E. nana* or with *Cryptosporidium* or *Cyclospora* oocysts in wet preparation. In acid-fast stains, the oocysts of *Cryptosporidium* and *Cyclospora* species stain pink to red. Yeasts are not acid fast and stain green.

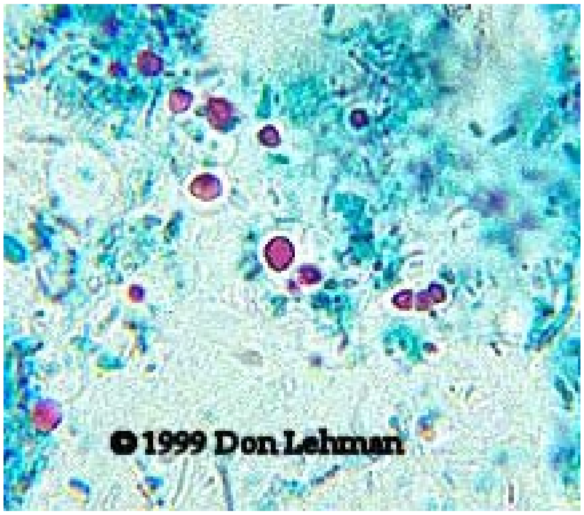

Image 13-7. Image illustrating Yeast Cells in slide preparation (Image courtesy of: Don Lehman). Note similarity to parasitic oocysts. (SOURCE: Don Lehman, Univ of Delaware)

Vegetable Cells

Plant cells and associated elements seen in feces may resemble eggs, cysts or cell bodies. Plant cells are often identified by a more irregular outer membrane.

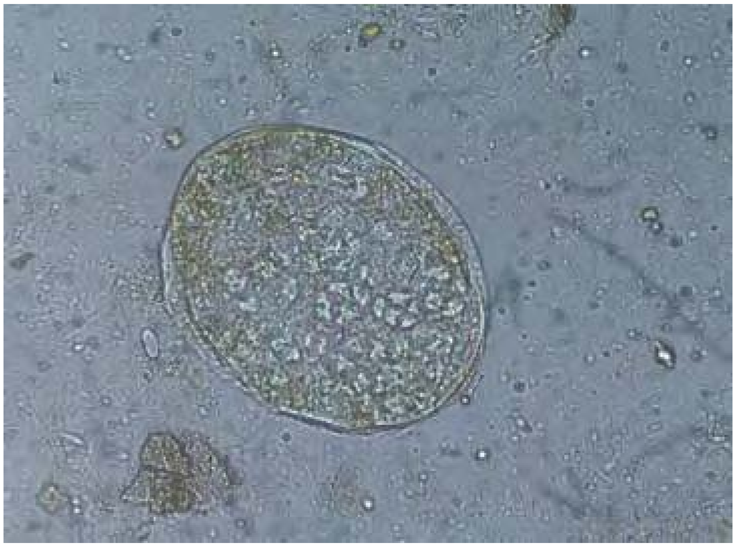

Image 13-8. Image illustrating Vegetable cell in slide preparation. Note similarity to *Paragonimus* eggs. (SOURCE: Unknown)

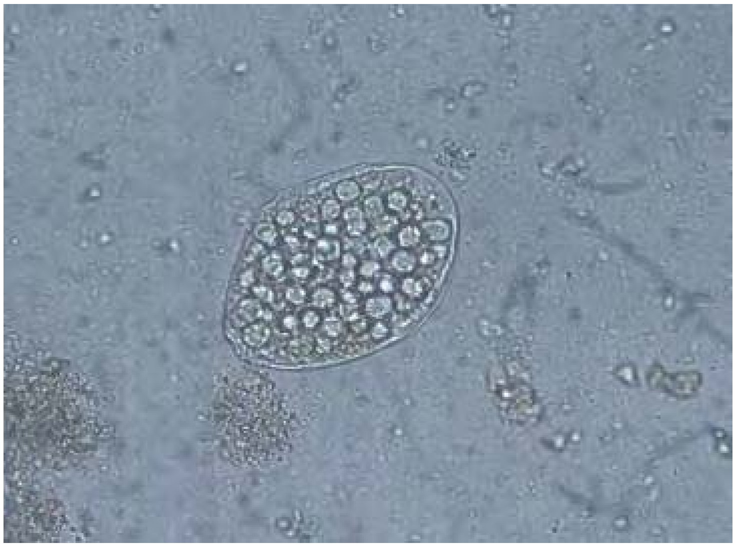

Image 13-9. Image illustrating Vegetable cell in slide preparation. Note similarity to *Dipylidium caninum* egg packets. (SOURCE: Unknown)

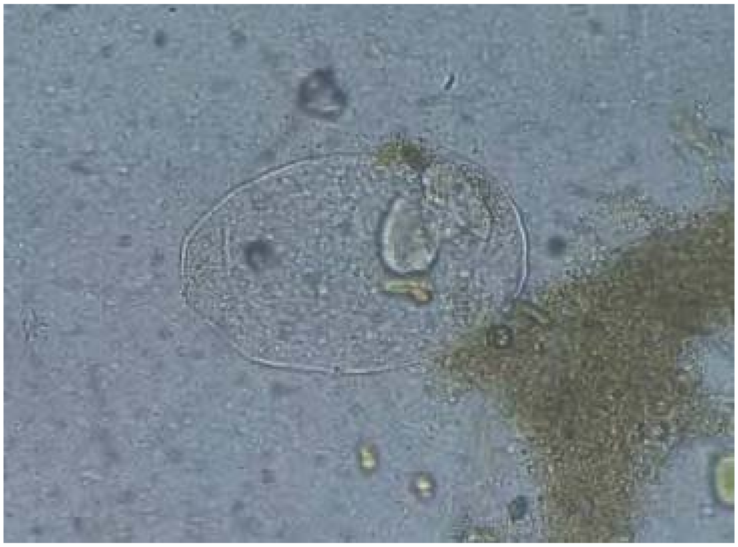

Image 13-10. Image illustrating Vegetable cell in slide preparation. (SOURCE: Unknown)

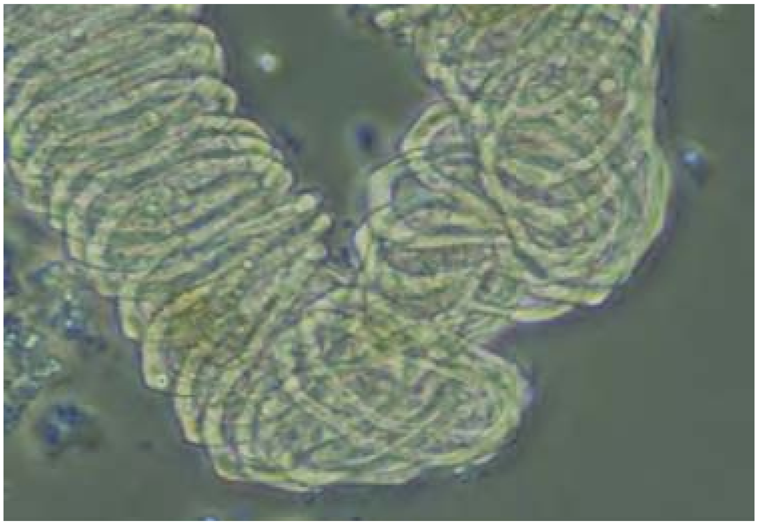

Image 13-11. Image illustrating a Vegetable Spiral in slide preparation. Such spirals may appear similar to proglottids. (SOURCE: Unknown)

Image 13-12. Image illustrating Vegetable Spiral in slide preparation. (SOURCE: Unknown)

Pollen

Pollen grains are often misinterpreted as parasite eggs, but can often be discerned through size and the presence or absence of important structural elements.

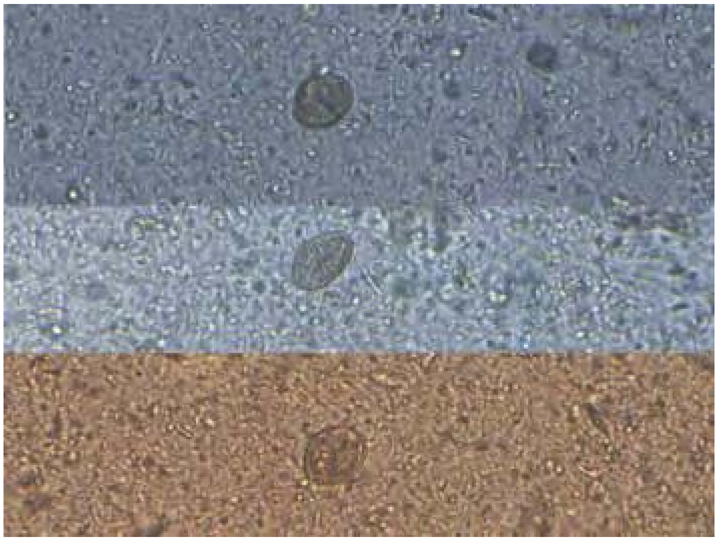

Image 13-13. Image illustrating pollen in slide preparation using a color filter. (SOURCE: Unknown)

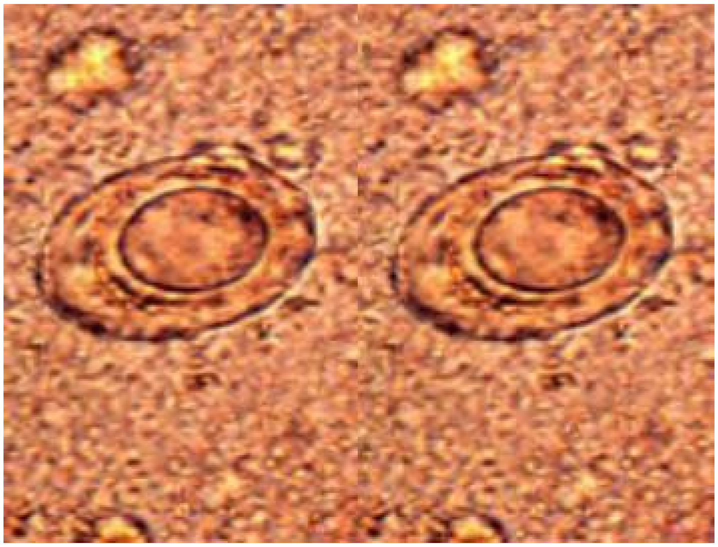

Image 13-14. Image illustrating pollen in slide preparation that could be mistaken for a *Taenia* egg. The shell is thinner, of non-uniform thickness, and no hooks are visible. (SOURCE: CDC)

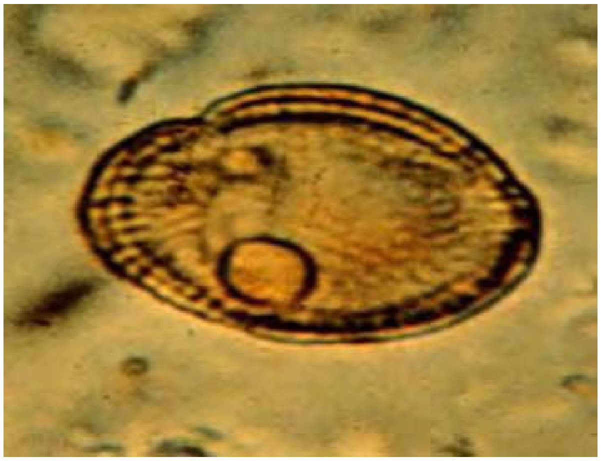

Image 13-15. Image illustrating pollen resembling a *Hymenolepis nana* egg. Hooks and polar filaments are not visible. (SOURCE: CDC)

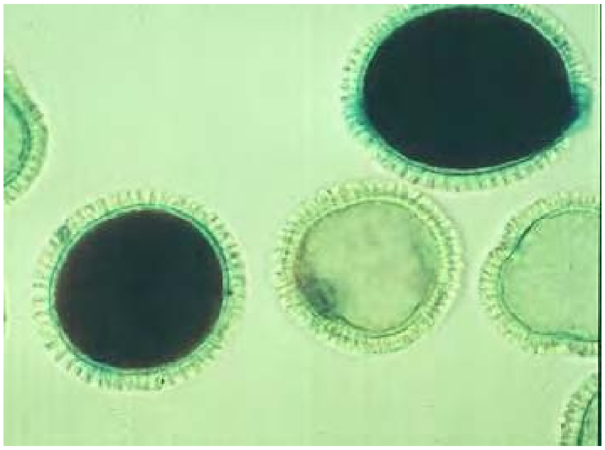

Image 13-16. Image illustrating geranium pollen cells in slide preparation. (SOURCE: Unknown)

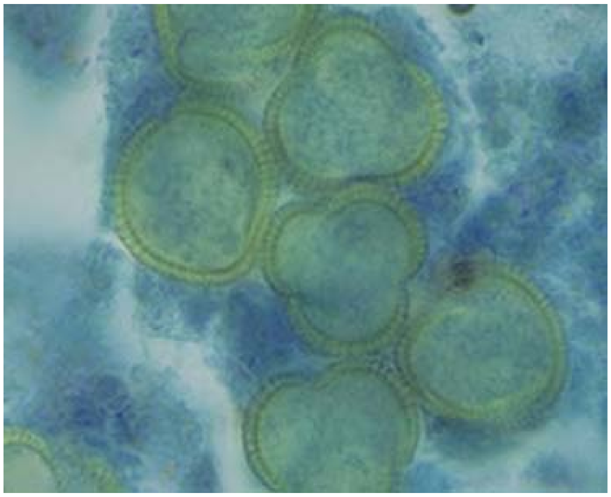

Image 13-17. Image illustrating pollen cells in slide preparation. Similar to *Taenia* eggs, but distinguished by uneven thickness of the wall and lack of internal contents do not suggest an egg. (SOURCE: Unknown)

Hair

Animal and plant hairs are most often and easily mistaken for parasitic nematode worms. Their size and shape may be comparable in many cases, but a lack of internal definition will identify the artifact when compared to the worm. Although nematodes are non-segmented and externally simple organisms, they will often show unique structural characteristics under close examination.

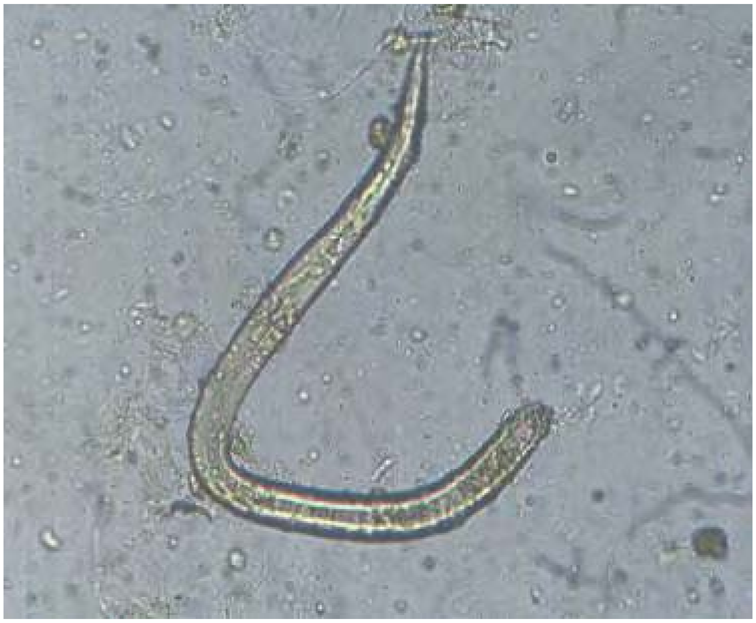

Image 13-18. Image illustrating peach hair in slide preparation. Note the similarity to *Strongyloides stercoralis.* (SOURCE: Unknown)

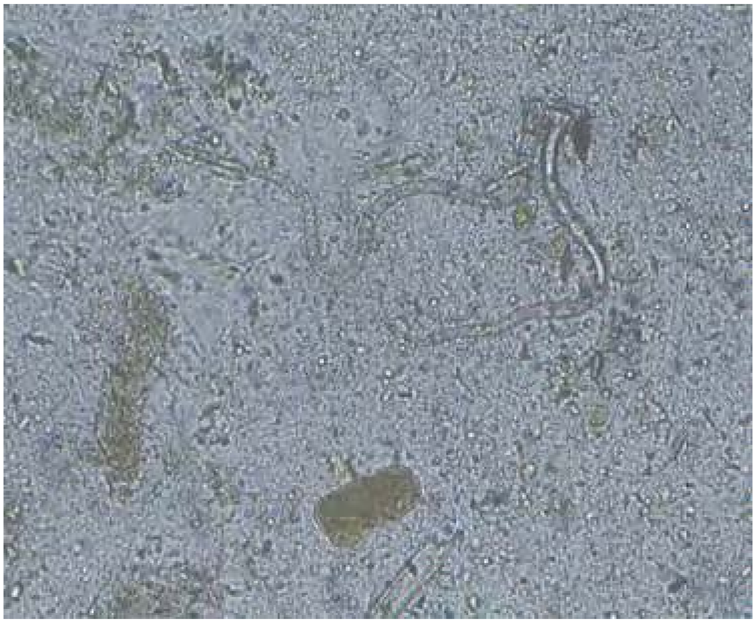

Image 13-19. Image illustrating vegetable hairs in slide preparation. (SOURCE: Unknown)

Insect Eggs

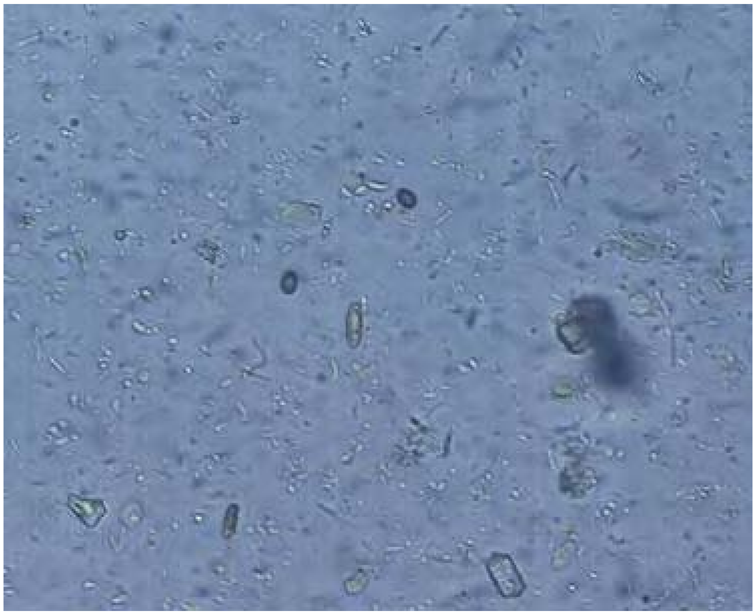

Image 13-20. Image illustrating Insect eggs in slide preparation. (SOURCE: Unknown)

Plant Parasites

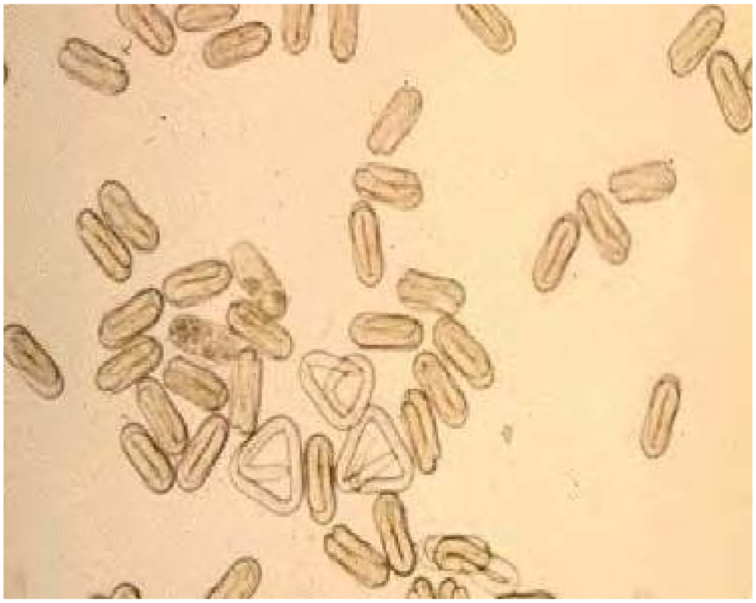

Image 13-21. Image illustrating *Heterodera spp*. in slide preparation. Such parasitic nematodes attack root vegetables such as beetroot, turnips and radishes. Their eggs are 80-120µm by 25-40 µm and can conceivably be confused with hookworm eggs. (SOURCE: Unknown)

Earthworms

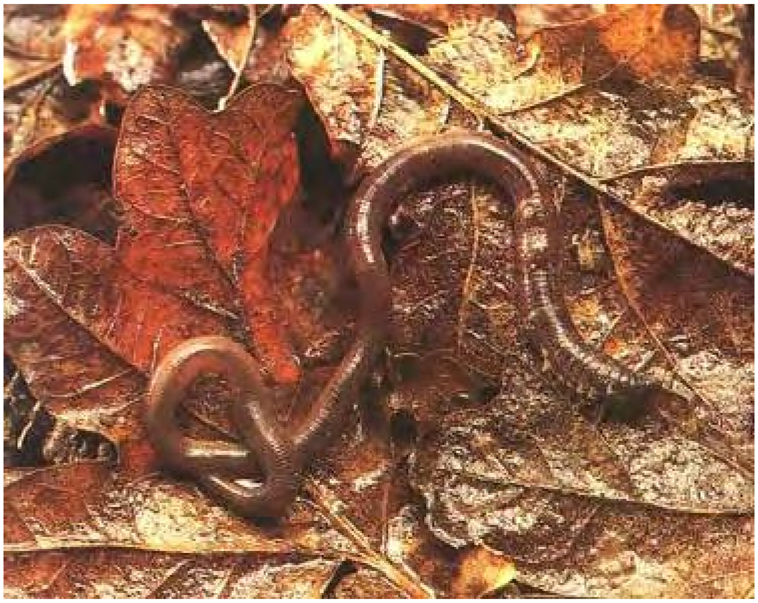

Image 13-22. Image illustrating an Annelid earthworm in detritus. They belong to the Annelida (Lumbricus and Allolobophora) and are elongated, segmented and circular in section and are occasionally confused with *Ascaris*. They have a purple-brown dorsal surface and a paler ventral surface, swell out at around segment 12 and possess a marked thickening (the clitellum) a third of the way from the anterior. (SOURCE: Unknown)

Cosmic Healing I

Cosmic Chi Kung

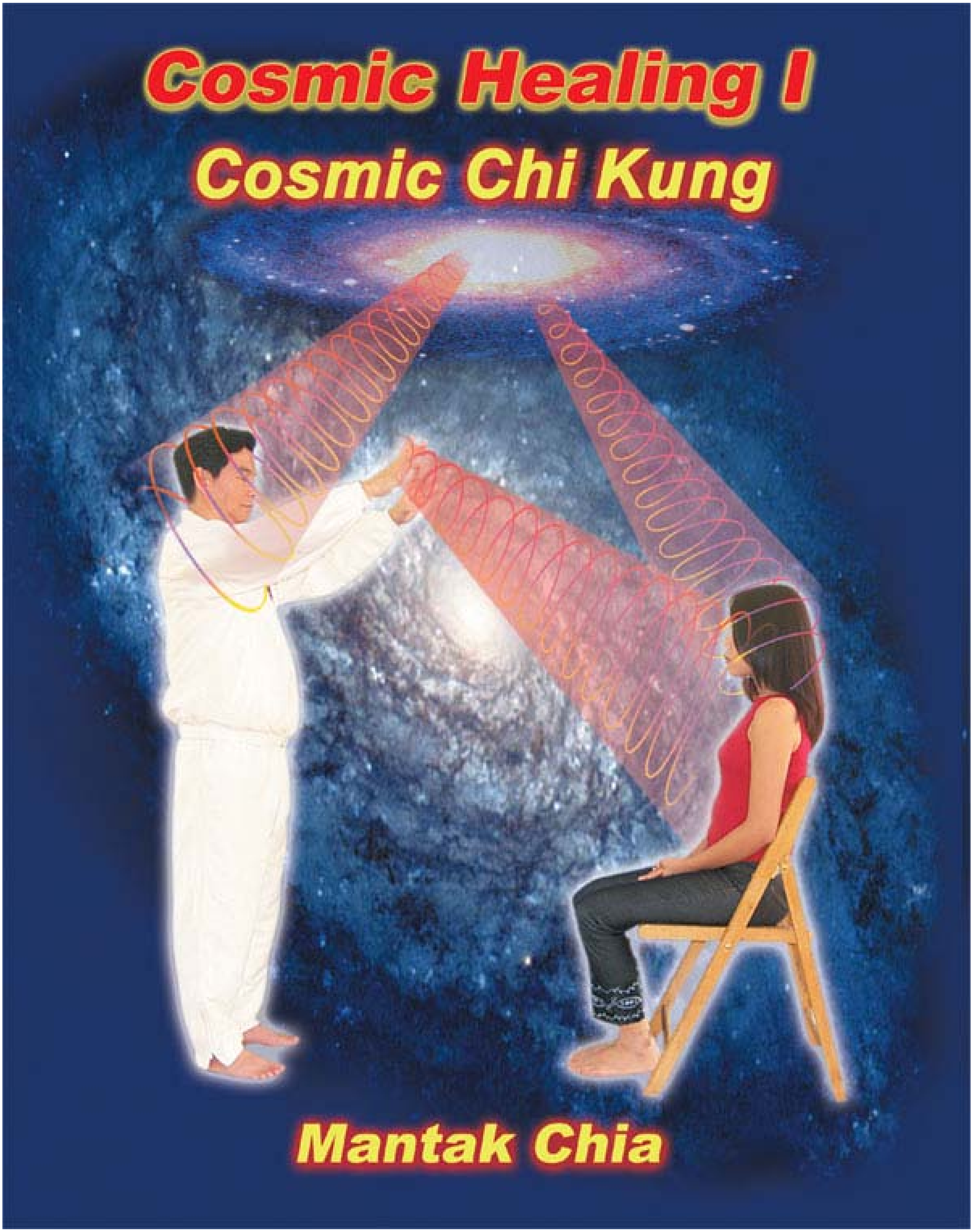

Mantak Chia

Cosmic Healing I

Cosmic Chi Kung

Mantak Chia

Edited by
Colin Campbell

Copy Editor: Colin Campbell

Assistant Editor: Matt Gluck

Contributing Editor: Kendra Moshe

Computer Graphics: Raruen Keawpadung

Photographer: Saysunee Yongyod

Design and Production: Saniem Chaisarn

Illustrations: Udon Jandee

Final Editing: Jean Chilton

Project Manager: W.U. Wei

© ***North Star Trust***

First published in 2001 by:

Universal Tao Publications
274/1 Moo 7, Luang Nua,
Doi Saket, Chiang Mai, 50220 Thailand
Tel (66) (53) 865-034 & 495-596 Fax 495-852
Email: universaltao@universal-tao.com

ISBN: 974-87672-5-6
Library of Congress Catalog Number:

Manufactured in Thailand
Second Printing, 2001

All rights reserved. No part of this book may be used or reproduced in any manner whatsoever without the express written permission from the author except for brief quotations embodied in critical articles and reviews.

Contents

About the Author 9
Acknowledgements 12
Words of Caution 13

Chapter I
Cosmic Healing

Cosmic Healing Chi Kung in the Universal Tao System.......... 14
  Foundation of the Universal Tao System 14
  Cosmic Inner Smile and the Six Healing Sounds 15
  Iron Shirt Chi Kung and Bone Breathing 15
  Stages of Mastering Chi.......... 16
Four Sections of Cosmic Healing.......... 17
  Connecting to Heaven and Earth 17
  Opening the Bridge and Regulator Channels & Extending your Chi Outwards 18
  Opening and Energizing the Governor and Functional Channels 18
  Activating One Finger Art and the Chi Belt 18

Chapter II
Preparing for Cosmic Chi Kung

Preparatory Practices.......... 19
Warm up the Stove 19
Transform Sexual Energy 20
Filling the Joints with Chi 21
Bone Breathing and Marrow Washing 21
  Bone Breathing.......... 21
  Marrow Washing 23
  Heavenly Marrow Washing with Violet Light 23
Microcosmic Orbit.......... 25

Chapter III
Cosmic Chi Kung

Opening the Three Tan Tiens to the Six Directions 27
  Introduction 27
  Relaxation, Letting Go, Surrender and Ego 28
  Mind, Organs and Sexual Organs 28
  Opening the Three Tan Tiens 30
Combined Practice 32
  Warm Up Exercises 32

True Breath - Skin Breathing 33
Daily Practice: Opening, Connecting & Activating Chi Within & Cosmic Charging 34
  Relax and Let go - Smile to Connect with Universe Within 41
  Tan Tien Consciousness: The Second Brain 37
  Three Minds into One: 'Yi' Power 44
Activating the Six Directions and the Three Fires 45
  Direction Below 45
  Front Direction and Tan Tien Fire 46
  Back Direction and Kidney Fire 47
  Heart Fire 48
  Sacred Fire (Chi Fire) 49
Open the Third Eye 49
  Front Direction: Push/Pull Master Practice 51
  Left and Right Directions 52
  Direction Above 53
  Open the Spine 54
  Open the Middle Channel and Perineum 56
  Tan Tien and the Universe 57
Open the Three Tan Tiens 59
  Upper Tan Tien - Mideyebrow 58
  Middle Tan Tien - Heart Center - Conscious Mind 61
  Recharge and Cut Around the Armpits 62
  Lower Tan Tien - Navel - Awareness Mind 63
  Activate Chi in the Bones of the Hips, Legs and Sacrum 64

Chapter IV
General Healing Session

Three Minds into One Mind (Introduction) 68
Activate the Six Directions (Introduction) 68
Activate the Three Fires (Introduction) 68
Linking Personal Stars, Energy Bodies and the Universe (Introduction) 69
Protective Circle and Chi Field (Introduction) 69
  Sacred or Chi Fire 69
  Chi (Sacred) Water Practice (Introduction) 69
Empty and Fill (Introduction) 70
Clean with Green Light (Introduction) 70
Clean with Blue Light (Introduction) 70
Charge with Violet Light (Introduction) 70

Chapter V
Cosmic Healing I - Basic

- Source of Chi and Colors 126
  - Earth Chi 127
  - Human and Nature Chi 128
  - Solar Chi 128
  - Simple Practice 128
  - General Information 128
- Healing Colors 129
  - White Chi 129
  - Blue Chi 129
  - Green Chi 129
  - Red Chi 130
  - Tangerine Chi 131
  - Yellow Chi 131
  - Violet Chi 131
  - Golden Chi 133
  - The Spleen 133
  - Cosmic Healing Planetary Colors 134
- Summary - Working with Color 135

Chapter VI
Individual Healing Session

- Almighty Knife 138
  - Outline of Practice 139
- Energizing and Activating the Immune System 141
  - Outline of Practice 142
- Cleansing the Internal Organs Technique 143
  - Healing Some Common Illnesses 143
  - General Principles 145
  - Outline of Practice 146
  - Cleansing the Spinal Cord 149
  - Cleansing the Blood (Lungs) 150
  - Outline for Cleansing the Blood/Lungs 150
  - Circulatory/Heart System 151
  - Stomach Pains 151
  - Relieving Pain: Blue Chi 152
  - Headache: Light Whitish Green and Light Whitish Blue Chi 153
  - Migraine Headache: Light Whitish Green Blue and Violet Chi 153
  - Toothache: Light Whitish Green and Blue Chi 153
  - Broken Bone: Orange - Yellow Chi 153
  - Back Injury: Light Whitish Blue and Green Violet Chi 154

- Cysts: *Light Green and Blue Chi* 154
- Infection and Inflammation: Light Green Chi 154
- Fever 154
- Insomnia: Green and Blue 155
- Tinnitus 155
- Blood Clotting: Blue Chi - Fresh Burns: Green and Blue Chi 155
- Old Minor Burns: Green and Red Chi 155
- Old Severe Burns: Green and Red Chi 156
- Old Wounds: Green and Red Chi 156
- Some Specific Points and Ailments: 156
- Solar Plexus: Blue Chi 157
- Skin Infection: Blue Chi 157
- Eyes: Green Chi 158
- Ears: Violet Chi 158

Summary of Individual Healing Session 159

Chapter VII
Different Hand Techniques

Forget about your Palms 165
- Don’t Focus your Mind on the Sick Part 165
- Cosmic Healing is Teaching 165

Chi Kung Diagnosis: Hand Scanning 166
- General Variations in the Energy Field 166
- Hand Scanning the Internal Organs 167

Chapter VIII
Cosmic Chi Kung Palm and Finger Training

Why We Practice the Cosmic Chi Kung 169
- Opening Heaven and Earth Force in the Palms - Heaven and Earth Spirals 170
- Cell Breathing 171
- Mind, Eyes, Heart and Intention (YI) 172
- Connecting the Bridge 173
- Cosmic Chi Kung Preparation 174
- Cosmic Chi Kung Preparation Standing 175
- Opening to Heaven, Earth and Nature 176
- Outline of Practice 176

Part I - Opening to Heaven, Earth and Cosmic Chi 177
- Opening Movements 177
- Core Movements of Part I 182
- Closing Movements 183

Part II - Opening the Bridge and Regulator Channels 189
- Theory 189

General Functions of the Eight Extraordinary Channels 190
The Eight Extraordinary Channels and Chi Kung 193
Special Functions of the Bridge and Regulator Channels............................. 193
Bridge Channels Yin and Yang Chiao Mo Function 193
Summary... 195
Practice 196
Part III - Opening the Functional and Governor Channels 205
Theory: The Functional and Governor Channels.. 205
Chi Ball 206
Part IV - Activating the Yin and Yang Channels on the Chi Belt.......................... 217
Theory 217
Yin and Yang Channels of the Arms... 217
Projecting Healing Energy through the Fingers 217
Activating the Chi Belt - Dai Mo.. 222
Practice 223
Healing Others.. 226
Summary and Combined Practice 226
Cosmic Chi Kung: Simple Combined Sequence 226

Chapter IX
Universal Tao Basic Practices

Cosmic Inner Smile 232
Six Healing Sounds... 236
The Six Healing Sounds - Practice 235

Appendix 241
Guide to Acupuncture Points used in Cosmic Chi Kung 241
Bibliography... 249

About the Author

Mantak Chia

Master Mantak Chia is the creator of the Universal Tao System and is the director of the Universal Tao Center and Tao Garden Health Resort and Training Center in the beautiful northern countryside of Thailand. Since childhood he has been studying the Taoist approach to life. His mastery of this ancient knowledge, enhanced by his study of other disciplines, has resulted in the development of the Universal Tao System which is now being taught throughout the world.

Mantak Chia was born in Thailand to Chinese parents in 1944. When he was six years old, Buddhist monks taught him how to sit and “still the mind.” While still a grammar school student, he learned traditional Thai boxing. He was then taught Tai Chi Chuan by Master Lu, who soon introduced him to Aikido, Yoga and broader levels of Tai Chi.

Years later, when he was a student in Hong Kong excelling in track and field events, a senior classmate named Cheng Sue-Sue introduced him to his first esoteric teacher and Taoist Master, Master Yi Eng (I Yun). At this point, Master Chia began his studies of the Taoist way of life in earnest. He learned how to circulate energy through the Microcosmic Orbit and, through the practice of Fusion of the Five Elements, how to open the other Six Special Channels. As he studied Inner Alchemy further, he learned the Enlightenment of the Kan and Li, Sealing of the Five Senses, Congress of Heaven and Earth and Reunion of Heaven and Man. It was Master Yi Eng who authorized Master Chia to teach and heal.

When Mantak Chia was in his early twenties he studied with Master Meugi in Singapore, who taught him Kundalini, Taoist Yoga and the Buddha Palm. He was soon able to clear blockages to the flow of energy within his own body. He learned to pass the life force energy through his hands also, so that he could heal Master Meugi's patients. He then learned Chi Nei Tsang from Dr. Mui Yimwattana in Thailand.

A while later, he studied with Master Cheng Yao-Lun who taught him the Shao-Lin Method of Internal Power. He learned the closely guarded secret of the organs, glands and bone marrow exercise known as Bone Marrow Nei Kung and the exercise known as Strengthening and Renewal of the Tendons. Master Cheng Yao-Lun's system combined Thai boxing and Kung Fu. Master Chia also studied at this time with Master Pan Yu, whose system combined Taoist, Buddhist and Zen teachings. Master Pan Yu also taught him about the exchange of Yin and Yang power between men and women, and how to develop the Steel Body.

To understand the mechanisms behind healing energy better, Master Chia studied Western anatomy and medical science for two years. While pursuing his studies, he managed the Gestetner Company, a manufacturer of office equipment and became well acquainted with the technology of offset printing and copying machines.

Using his knowledge of Taoism, combined with the other disciplines, Master Chia began teaching the Universal Tao System. He eventually trained other Instructors to communicate this knowledge and he established the Natural Healing Center in Thailand. Five years later, he decided to move to New York, where in 1979, he opened the Universal Tao Center. During his years in America, Master Chia continued his studies in the Wu system of Tai Chi with Edward Yee in New York.

Since then, Master Chia has taught tens of thousands of students throughout the world. He has trained and certified over 1,200 instructors and practitioners from all over the world. Universal Tao Centers and Chi Nei Tsang Institutes have opened in many locations in North America, Europe, Asia, and Australia.

In 1994, Master Chia moved back to Thailand, where he had begun construction of Tao Garden, the Universal Tao Training Center in Chiang Mai.

Master Chia is a warm, friendly and helpful man who views himself primarily as a teacher. He presents the Universal Tao System in a straightforward and practical manner, while always expanding his knowledge and approach to teaching. He uses a laptop computer for writing and is totally at ease with the latest computer technology.

Master Chia estimates that it will take thirty five books to convey the full Universal Tao System. In June 1990, at a dinner in San Francisco, Master Chia was honored by the International Congress of Chinese Medicine and Qi Gong (Chi Kung), who named him the Qi gong Master of the Year. He is the first recipient of this annual award.

In December 2000, the Tao Garden Health Resort and Universal Tao Training Center was completed with two Meditation Halls, two open air Simple Chi Kung Pavilions, indoor Tai Chi, Tao Yin and Chi Nei Tsang Hall, Tai Chi Natural Swimming Pool, Pakua Communications Center with a complete Taoist Library, Internal World Class Weight Lifting Hall and complete 8 Court Recreational Facilities.

In February 2002, the Immortal Tao practices will be held at Tao Garden for the first time using Dark Room technology, creating a complete environment for the higher level Taoist practices.

He has previously written and published these Universal Tao books:

Awaken Healing Energy of the Tao - 1983
Taoist Secrets of Love: Cultivating Male Sexual Energy
co-authored with Michael Winn - 1984.
Taoist Ways to Transform Stress into Vitality -1985
Chi Self-Massage: the Tao of Rejuvenation - 1986
Iron Shirt Chi Kung I - 1986
Healing Love Through the Tao: Cultivating Female Sexual Energy - 1986
Bone Marrow Nei Kung - 1989
Fusion of the Five Elements I - 1990
Chi Nei Tsang: Internal Organ Chi Massage - 1990
Awaken Healing Light of the Tao - 1993
The Inner Structure of Tai Chi co-authored with Juan Li - 1996
Multi-Orgasmic Man co-authored with Douglas Abrams 1996 - published by Harper/Collins
Tao Yin - 1999
Chi Nei Tsang II - 2000
Multi-Orgasmic Couple co-authored with Douglas Abrams 2000 - published by Harper/Collins
Cosmic Healing I - 2001
Cosmic Healing II co-authored with Dirk Oellibrandt - 2001

Also many of the books above are available in the following foreign languages: ***Bulgarian, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Lithuanian, Malaysian, Polish, Portuguese, Russian, Serbo-Croatian, Slovenian, Spanish,*** & ***Turkish*** editions are available from the Foreign Publishers listed in the Universal Tao Center Overview in the back of this book.

Acknowledgments

The Universal Tao Publications staff involved in the preparation and production of *Cosmic Healing I: Cosmic Chi Kung* extend our gratitude to the many generations of Taoist Masters who have passed on their special lineage, in the form of an unbroken oral transmission, over thousands of years. We thank Taoist Master I Yun (Yi Eng) for his openness in transmitting the formulas of Taoist Inner Alchemy.

Thanks to Juan Li for the use of his beautiful and visionary paintings, illustrating Taoist esoteric practices.

We offer our eternal gratitude to our parents and teachers for their many gifts to us. Remembering them brings joy and satisfaction to our continued efforts in presenting the Universal Tao System. For their gifts, we offer our eternal gratitude and love. As always, their contribution has been crucial in presenting the concepts and techniques of the Universal Tao.

We wish to thank the thousands of unknown men and women of the Chinese healing arts who developed many of the methods and ideas presented in this book. We offer our graditute to Master Lao Kang Wen for sharing his healing techniques.

We wish to thank Colin Campbell for his editorial work and writing contributions on the first printing of this book, as well as his ideas for the cover. We appreciate his research and great labor. We wish to thank Matt Gluck for his editorial contributions on the revised edition of this book, as well as thanking our instructors, Dennis Huntington and Annette Derksen, for their insightful contributions to the revised version. We thank Dirk Oellibrandt for his technical editing and clear writing throughout the book.

A special thanks goes to our ***Thai Production Team*** for their cover illustration and book design and layout: Raruen Keawapadung, Computer Graphics; Saysunee Yongyod, Photographer; Udon Jandee, Illustrator; and Saniem Chaisarn, Production Designer.

Thanks to the testing institutes: Gerhard Eggelsberger, Institute for Applied Biocybernetics Feedback Research, Vienna, Austria and Dr. Ronda Jessum, Biocybernetics Institute, San Diego, California.

We also wish to express thanks to Cosmic Healing Certification Retreat Organizers around the world who have worked with us for many years and helped us prepare this manuscript throughout the Retreats: Beate Nimsky, Vienna, Austria; Serguei Orechkine, Moscow, Russia; Christopher Larthe, London, England; Masahiro Ouchi, Rye, New York; Dr. Angela C. Shen, San Francisco, California; Dirk Oellibrandt, Hamme, Belgium; Brita Dahlerg, Frankfurt, Germany and Lizbeth Cavegn, Zurich, Switzerland.

We wish to further express our gratitude to all the instructors and students who have offered their time and advice to enhance this system, especially Felix Senn, Barry Spendlove, Chong-Mi Mueller, Clemens Kasa, Andrew Jan, Marga Vianu, Harald Roeder, Salvador March, Dr. Hans Leonhardy, Peter Kontaxakis, Thomas Hicklin, Gianni Dell'Orto, Walter and Jutta Kellenberger.

Words of Caution

The practices described in this book have been used successfully for thousands of years by Taoists trained by personal instruction. Readers should not undertake the practice without receiving personal transmission and training from a certified instructor of the Universal Tao, since certain of these practices, if done improperly, may cause injury or result in health problems. This book is intended to supplement individual training by the Universal Tao and to serve as a reference guide for these practices. Anyone who undertakes these practices on the basis of this book alone, does so entirely at his or her own risk.

The meditations, practices and techniques described herein are ***not*** intended to be used as an alternative or substitute for professional medical treatment and care. If any readers are suffering from illnesses based on mental or emotional disorders, an appropriate professional health care practitioner or therapist should be consulted. Such problems should be corrected before you start training.

Neither the Universal Tao nor its staff and instructors can be responsible for the consequences of any practice or misuse of the information contained in this book. If the reader undertakes any exercise without strictly following the instructions, notes and warnings, the responsibility must lie solely with the reader.

This book does not attempt to give any medical diagnosis, treatment, prescription, or remedial recommendation in relation to any human disease, ailment, suffering or physical condition whatsoever.

Chinese Medicine and Chi Kung emphasizes balancing and strengthening the body so that it can heal itself. The meditations, internal exercises and martial arts of the Universal Tao are basic approaches to this end. Follow the instructions for each exercise carefully and do not neglect the foundations, i.e, the Microcosmic Orbit and any other supplemental exercises. Also pay special attention to the warnings and suggestions. People who have high blood pressure, heart disease or a generally weak condition should proceed cautiously, having prior consent from a qualified medical practitioner. People with venereal disease should not attempt any practices involving sexual energy until they are free of the condition.

The Universal Tao is not and cannot be responsible for the consequences of any practice or misuse of the information in this book. If the reader undertakes any exercise without strictly following the instructions, notes, and warnings, the responsibility must lie solely with the reader.

Chapter I
Cosmic Healing

Cosmic Healing Chi Kung in the Universal Tao System

Cosmic Healing Chi Kung is an important branch of the Universal Tao, an entire system of Taoist practices for cultivating the body, the Chi and the spirit. Your level of skill in Cosmic Healing Chi Kung will be based upon your mastery of the Universal Tao System as a whole. One of the most important parts of the Taoist practices is working with the energy moving within the body.

Foundation of the Universal Tao System is the Microcosmic Orbit Meditation

This is the major practice to connect with the earth, cosmic and universal forces. This is the way to make a connection with the North Star and the Big Dipper's violet/red light which is one of the very powerful universal healing lights. Through this practice you learn to feel Chi and to use your mind/eye/heart power to guide the Chi flow through the primary energy routes in your body. This practice is later extended to include the arm and leg routes too; it is then called the Macrocosmic Orbit.

Many people, including Chi Kung Masters, come to me and say that they have been practicing Chi Kung for years and nothing is happening. They don't feel any energy and they think that they must be performing their Chi Kung incorrectly. I tell them that they also need to do meditation. Hand movements are nothing by themselves. There are hundreds of different Chi Kung forms in China. You could spend seventy lifetimes just learning the hand movements. The practice of the Microcosmic Orbit meditation will help you to feel the Chi more easily inside, outside or around the body and in the extremities of the hands and feet.

When I was a child I liked to practice Chi Kung so much I had to save my lunch money to learn. After many years of practice of many Chi Kung forms, I started to forget the first form, so I learned it again. One day when I tried to practice and review all the forms that I had learned, I couldn't remember many of them. I sat down and thought, "I only have two hands, two legs and one head. Why are there so many forms? And why are there are so many religions and beliefs." I said to myself that there must be one main thing that they have in common. I started to search and I found out that the most important thing is feeling the Chi within us. Being able to increase, transform, take in and stay in touch with the universal, the cosmic, nature and the earth force. Letting them all combine within us is what is important. Likewise in religion there is God (a Force in Taoism we call Wu Chi, the nothingness, the supreme power controlling the universe); with a good heart and virtue energy we can connect with this force.

The complete practice of the Microcosmic Orbit Meditation is fully described in the book "Awaken Healing Light of the Tao" by Mantak Chia.

Cosmic Inner Smile and the Six Healing Sounds

These are very important practices to make the connection among the organs, the colors and the good virtue energy; each organ color will enhance your connection to the cosmic and the Universal Healing power.

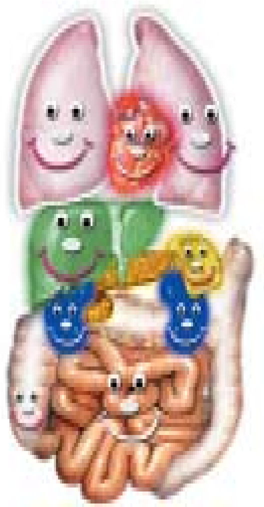

Fig. 1.1 *Red for the Heart*
White for the Lungs
Yellow for the Spleen
Blue for the Kidneys
Green for the Liver.

Each organ has its own vital color and when abundant it will radiate out as an aura for healing and protection. These organ colors and their associated universal connections have great healing power. The power of the 'Six Healing Sounds' will help enhance the connection to the cosmic source. Each sound will foster different healing energy. The practices also help balance, refine and transform the negative energy back to positive energy.

Iron Shirt Chi Kung and Bone Breathing

Iron Shirt Chi Kung and Bone Breathing are a system of standing meditations. These exercises help you to become grounded and to root your connection into the earth force, in order to build a good structure and absorb greater force. In turn, this will help draw in the heavenly force, which will combine with the very potent healing power of the earth's yellow light. The practice also strengthens the body so you can hold a higher energetic charge. These abilities are essential prerequisites for handling greater amounts of energy.

Iron Shirt Chi Kung also includes the art of ‘Changing the Sinews and Washing the Marrow’. Through these aspects of Iron Shirt, you learn to absorb, store and discharge large amounts of energy through the tendons and bones. *These practices are outlined in detail in the books “Iron Shirt Chi Kung” by Mantak Chia and “Bone Marrow Nei Kung” by Mantak Chia.*

To attain skill in Cosmic Chi Kung, it will help if you practice Iron Shirt, Bone Marrow and meditation as a minimum requirement. Beyond the basic level, the Universal Tao System includes many other intermediate and advanced level Chi Kung practices and meditations. The further one advances, the greater one's mastery of Chi.

Your increasing level of skill in the Universal Tao system will reflect immediately in your Cosmic Healing Chi Kung practice. Furthermore, you will discover that you can incorporate many of your Universal Tao practices directly into your practice of Cosmic Chi Kung. We will give you a simple combined practice of the Cosmic Healing Chi Kung, the Inner Smile, Microcosmic Orbit, Iron Shirt and the Sexual practices.

Stages of Mastering Chi

1. In the Universal Tao system, our *first* goal is to learn to conserve our Chi; when a battery is totally drained it is harder to charge; money makes money, Chi makes Chi. Conservation of Chi will help gain more Chi. To have more Chi we first need to maintain control of the gates through which energy normally leaks out and unwittingly drains our life force.

We leak energy:
- through our reproductive system
- through negative emotions
- through constantly turning our senses outward

Without knowing how to conserve the Chi that we already have, what is the point of acquiring more?

2. We learn to balance our Chi; that is, we seek to keep a smooth and balanced flow of energy moving throughout the whole body. If our energy is imbalanced, we may have too much energy in some places and not enough in others; we may also be too yang or too yin. We may have excess or deficient heat, cold, damp or dryness. This imbalanced energy tends to make us go to extremes.

3. We learn to transform our Chi into more beneficial energies. For example, through the Taoist Sexual Chi Kung practices taught in the Universal Tao (the course known as Healing Love through the Tao), we can transform sexual energy back into basic life force Chi. Through other practices (such as the Inner Smile, the Six Healing Sounds, and Fusion of the Five Elements) we learn to transform negative emotional Chi into positive virtuous Chi. Thus Chi is not only the foundation of our health; it is also the basis of spiritual development in the Tao.

4. Once we have accomplished the three previous phases of mastering Chi, we then learn to increase it. Chi pervades all of heaven, earth and nature. In Cosmic Healing

Chi Kung we learn time tested ways to tap into these unlimited and transpersonal reservoirs of Chi and greatly expand the amount of energy available to us. It is very important to master first the stages of conservation, balance and transformation before we emphasize increasing our Chi. Otherwise we may waste the energy we bring in, or we may inadvertently amplify the imbalanced or negative energies that we have not yet learned to bring under control.

5. Finally, we learn to extend our mind to tap into the vast Chi of nature, cosmic and the universe, to heal our body, mind and spirit and to heal other people. Cosmic Healing Chi Kung practice sensitizes your hands to the feeling and movement of Chi; it uses the mind-eye power to absorb cosmic Chi into the palm and crown and to send it out through the hands and beyond, so that you can help restore balance in others without touching them or draining yourself.

   This may sound fantastic at first, yet recent researchers in Chinese Chi Kung hospitals have not only measured the energies emitted by Chi Kung masters; they have also discovered that different varieties and frequencies of healing Chi can be emitted. This research has been corroborated by experiments in the United States at such places as the Menninger Institute.

Four Sections of Cosmic Healing

The next step is to learn the four sections of the Cosmic Healing Chi Kung, the ‘Empty Force’ sequence. Each of the four sections of Cosmic Healing Chi Kung develops a different type of energy mastery. Each section begins and ends in the same way, yet has different movements in the middle.

Connecting to Heaven and Earth

The first section of Cosmic Healing Chi Kung emphasizes connecting to the external Chi emanating from heaven and earth. This ability is very beneficial for self-healing and is essential for healing others. If you do not connect to some source of external energy when healing others, you will draw upon your own personal reservoirs. Our personal Chi is limited by nature, and can easily become depleted if we give it away too freely.

Over the past decade external Chi healing has become very popular in China. There are now hundreds of Chi Kung hospitals and clinics throughout the country. Yet many of the Chi Kung therapists can only administer two or three treatments a day; they spend the rest of their day practicing Chi Kung to replenish their own Chi. Those who practice the more traditional Chi Kung know how to connect to the Chi of heaven, earth and nature.

In this first section, you will learn to sense the energies outside your body. You will then learn to draw and absorb these energies into your body and process them in order to use them for self-healing and healing others.

Opening the Bridge and Regulator Channels and Extending your Chi Outwards

The second section opens and strengthens four of the Eight Extraordinary Channels: the Yin and Yang Bridge and Regulator Channels. This section also teaches the student to extend and take in the cosmic Chi and emit Chi out of the body for healing others. This ability is the heart of Cosmic Healing Chi Kung practice.

Opening and Energizing the Governor and Functional Channels

The third section enhances the energy in the points along the two channels used in the Microcosmic Orbit. Opening the Microcosmic Orbit strengthens and balances one's own energy, and prepares one to heal others. Practicing this section will be of great help with the Microcosmic Orbit Meditation.

Activating One Finger Art and the Chi Belt

The fourth section of Cosmic Healing Chi Kung activates the hands and you will practice emitting Chi from each of your fingers and opening the "Chi Belt." This channel is also known as the Belt Route or Girdle Vessel.

Chapter II

Preparing for Cosmic Chi Kung

Preparatory Practices

In the Universal Tao, we teach Cosmic Healing Chi Kung within an entire system of exercise and meditation. As a part of this system, Cosmic Healing Chi Kung comprises the art of healing oneself and others. It focuses on energy work for projecting Chi to effect these healings. One can easily learn the simple movements of Cosmic Healing Chi Kung without doing any other Universal Tao practices, but if one truly wishes to master the art of Cosmic Healing Chi Kung, it is important to have a firm foundation in the basic Universal Tao practices. Therefore in this chapter we will introduce you to the basic Universal Tao practices that are used in conjunction with Cosmic Chi Kung.

The preparatory practice consists of a few parts; work through each part at your own pace and eventually join them together as a whole.

Always start with "warming the stove" at the abdomen and direct the fire down to the sexual center to transform the sexual energy. Next practice the Cosmic Inner Smile. The Cosmic Inner Smile is a powerful relaxation and self-healing technique that uses the energy of love, happiness, kindness and gentleness as a language to communicate with the internal organs of the body. Each organ corresponds to a specific element and color. For example, the kidneys will consist of the element of water and the color blue. The heart will consist of the element of fire and the color red. This makes it very easy to guide the healing power into each organ by using the appropriate color. The practice also aids the transformation of negative emotions into positive virtuous energy. This transformation is a very powerful Chi Kung practice. A genuine smile transforms negative energy into loving energy that has the power to relax, balance and heal. By learning to smile inwardly to the organs and glands, your whole body will feel loved and appreciated and enjoy more Chi. After the Inner Smile, practice Bone Breathing, Marrow Washing, Cosmic Healing and the Microcosmic Orbit.

Warm Up the Stove

1. Sit on the edge of a chair with your hands clasped together and your eyes closed.
2. Start by doing the "bellows breathing", moving your abdomen in and out quickly. Emphasize the exhalation by breathing out forcefully 18 to 36 times. Rest, cover your navel and feel nice and warm.
3. Next, do the "inner laughing" exercise; you feel the abdomen vibrate inside you. Practice

this for a few minutes, allowing the movement of your inner laughter to grow stronger. Then rest and use the mind and eye power to gather the Chi (now felt as a warmth behind the navel) into the Tan Tien. Picture a stove with burning fire behind your navel. Feel nice and warm.

Transform the Sexual Energy

Once you feel the Tan Tien is warm enough, smile down and bring the warm Chi down to the sexual organs; women bring the Chi down to the uterus, men bring the Chi down to the testicles. It may feel like the ‘sun shining on the water’. The rays of the sun purify the water until it becomes rising steam. Keep smiling down to the sexual organs and feel the warm or fiery feeling from the navel area continue to flow down to the sexual organs; this transforms the sexual energy into Chi, and this Chi rises up the spine into the brain helping to activate the ‘crown’ and ‘mideyebrow’ energy centers.

Focus your awareness in the sacrum. At the very tip of the sacrum there is a hole in the coccyx. Breathe into this hole until you feel some activity there. This may be felt as tingling, numbness or pulsing. If you can really activate this point it will generate 'suction'. Be aware of the sacrum opening. Feel the suction force pulling through the hole in the tailbone (sacral hiatus); breathe into it until you feel it becomes activated. When the sacrum is activated you will feel the suction easily, as well as breathing in the cranium and mideyebrow. Keep on gently smiling and softly breathing into the Tan Tien and feel the suction in the abdomen. Focus 95 % of your awareness in the Tan Tien and 5 % in the sacrum, the crown and the mideyebrow. Be aware of the Tan Tien breathing and observe internally the pulsing and breathing in the sacrum, the mideyebrow and the crown. Do this 36 times.

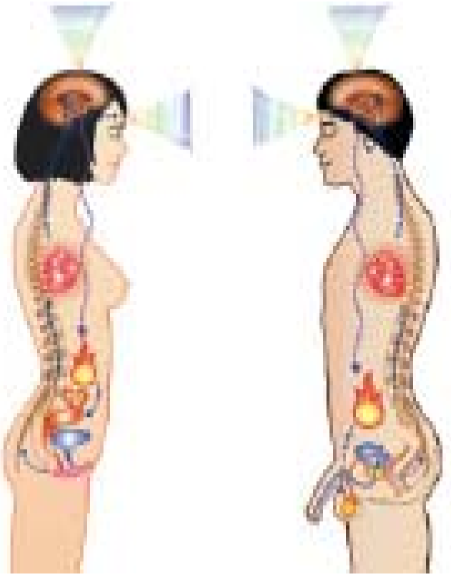

Fig. 2.1 *Smile down to the Lower Tan Tien.*

Filling the Joints with Chi

The bones have the ability to process Cosmic Chi (Chi above and around us) to be used by the body. The joints are also able to store Chi and they also serve as the cushion between the bones.

Be aware of your index fingers and thumbs; lightly raise the index fingers, feel a slight tension and open the thumbs to the sides and feel it magnify. Breathe into the finger tips, until you feel the sensation of Chi entering the fingers. The fingers will feel stiff as they stretch, to make way for the Chi entering the joints. Continue to breath in the Chi - the stiffness will spread to the other fingers, wrists, elbows and shoulder joints making the whole arm tense and stiff. The fingers and the arms become one piece. Eventually the Chi will fill every joint of the body making it into a complete unit.

Be aware of the big toes; breathe into the toes and feel the toes grow longer. When the Chi fills the joints the big toes become tense and stiff. Gradually it will fill the other toes rising to the ankles, knees, hips, sacrum and spine; the feeling may be one of tension and stiffness, numbness or like an electrical charge. The legs and the spine become one piece. You will certainly feel good.

Bone Breathing and Marrow Washing

Bone Breathing is one of the main practices of Cosmic Healing Chi Kung. This is a method of drawing external Chi through the skin into the bones to help replenish the bone marrow, thereby reactivating the production of white blood cells. Sending Chi into the bones will enhance functioning of the immune system. This process also helps to clean out fat in the bone marrow ("Washing the Marrow"), one of the main causes of osteoporosis (brittleness of the bones). Tension in the muscles close to the bones is decreased so Chi and blood can flow into the bones easily, enabling them to become stronger.

The Bone Breathing process uses the mind and the eyes to absorb Chi into the bones. The better your Bone Breathing is the better your Cosmic Healing Chi Kung practice will be. Mind and eyes are also used immediately after the exercise to gather the energy at the navel. Once you have it, when you move your hand, the energy will follow easily. You will be able to absorb external Chi effortlessly, so you will not need to use your own energy in your healing work.

Bone Breathing

There are several variations in the Bone Breathing and Marrow Washing process. Here we will introduce you to the first type of Bone Breathing, inhaling and exhaling Chi through the skin and packing it into the bones. In this method, you imagine that your bones are like hollow tubes, and that you breathe and suck the Chi into the bones.

Bone Breathing is practiced in a three-stage process. **Let your breathing follow a normal pace. Do not strain or hold your breath too long**.

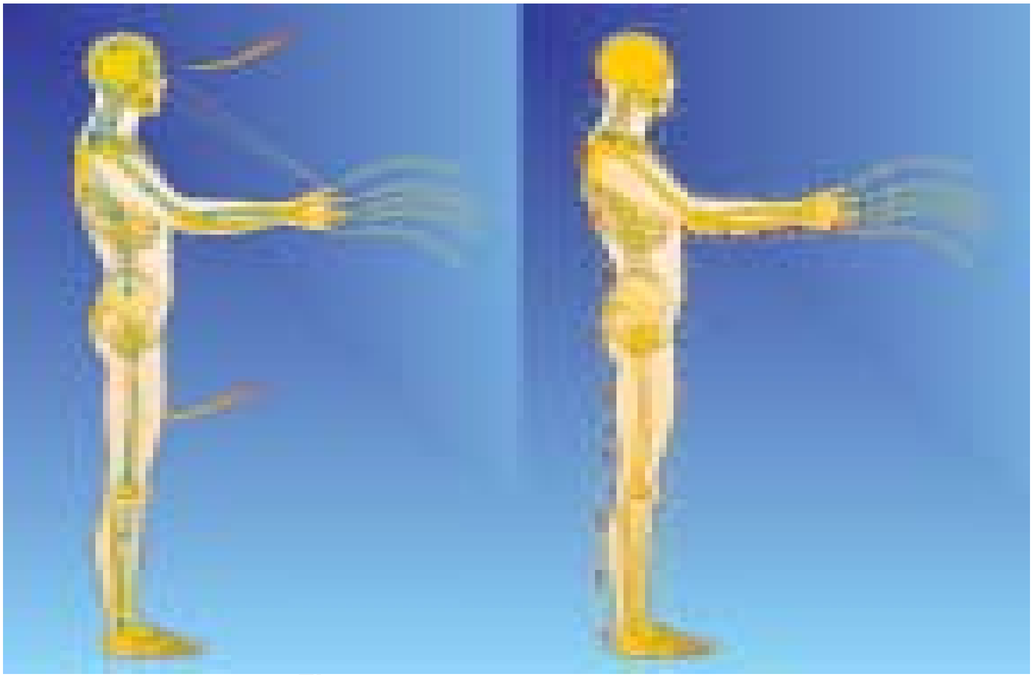

Fig. 2.2 *Bone Breathing through the Fingers*

1. You can do this practice in the sitting position or in the Embracing the Tree posture (or any other Iron Shirt Chi Kung posture). You will use your mind and eye power to breathe in a short breath and at the same time feel suction. Suck the Chi of the atmosphere into your hands, eventually expand to the universe and breathe in a few more times. Use a combination of 'mind/eyes/heart' power (Yi), to suck the Chi from the atmosphere into your hands, while taking small sips of breath. Once you can clearly feel the increase of Chi 'pressure' in your hands, you extend the feeling throughout your arms. The whole skin surface of the arms breathes in the Chi, feeling the skin holding this pressure.
2. Inhale one more deep breath and lightly contract the arm muscles to squeeze the Chi into the bones. Hold for a while to condense the Chi into the marrow of your bones. Exhale and at the same time feel a distinct heaviness in the bones, meaning that the Chi has been condensed and packed into the bones successfully. Eventually you will use more mental power (mind/eyes/heart power) and less muscle, using soft breathing, to draw the force into the bones. Do 6 to 9 times; rest and feel the Chi that has been condensed into the bones.

3. Proceed in the same way, breathing in progressively through the bones of the forearms, upper arms, scapulae, collarbone, sternum and ribs. You may feel a different sensation as you breathe in each area; in some areas the feeling is cool, in others warm or tingling, depending on the bone structure and the quality of the marrow.
4. Inhale and exhale in the same way through the toes, feet and legs; then, in a step-by-step progression, inhale up through the calf bones, thighbones, pelvis, coccyx and sacrum, and up the spinal column to the C-7 vertebrae.
5. Finally breathe in through the arms and legs simultaneously. Combine their energy as it flows up past C-7 and up through the neck and skull. Breathe in this way for at least nine breaths. Conclude by collecting energy at the navel.

Marrow Washing

You can wash your bone marrow with earth force, heavenly force or with cosmic force. This energy helps to cleanse and rejuvenate the bone marrow.

Heavenly Force Marrow Washing with Violet Light

1. Men place the left palm on the top of the head, and cover it with the right palm. Be aware of your 'Personal Star' (star that appears in your mind's eye and is located above your crown), above. Slightly press the palms and spiral in the clockwise direction, spiralling from the front to the right ear, to the back and the left ear 9 times; rest and feel the increased Chi pressure in the crown. Do 3 sets.
2. Women place the right palm on the top of the head and cover it with the left palm. Be aware of your Personal Star above you. Press the palms lightly and spiral in a counterclockwise direction, spiralling from the front to the left ear, to the back and the right ear 9 times; rest and feel the increased Chi pressure in crown. Do 3 sets.
3. Face your palms towards Heaven and feel that you are scooping a galactic Chi ball from above. This Chi ball contains the North Star (Violet) and the Big Dipper (Red). See the dipper fill with violet/red energy. Imagine you reach out your arms to hold the dipper handle and pour the violet/red liquid over your crown. It will feel like there is a numbness descending. Your palms face down to your crown and pour the whole galaxy onto your crown. You may perceive this energy as violet amethyst and red light frequencies. Smile.

4. Guide this sensation down into your skull, deep into your brain, cervical vertebrae, sternum, thoracic vertebrae, lumbar vertebrae and down through your legs. Feel it penetrating and enlivening your bones, deep into the bone marrow, washing, cleansing, energizing. This 'liquid-like' Chi spills all the way down to your feet. Feel it connecting with the earth through the soles of your feet; be aware of the bubbling springs in the feet (the K1 points of the kidney meridian) breathing and pulsating.
5. Touch your navel with the fingers of both hands. Focus on the door of life, and let the fire activate in the Tan Tien and kidneys. Chi will rise up to the brain.

6. Move your hands down touching your femur bones; feel your hands penetrate into the bone and into the bone marrow. Feel like an electric shock or a tingling moving all the way down through the bones of the leg and down to the soles of your feet.
7. Slowly move your fingers down touching the bone and into bone marrow and extend yourself down past the earth to touch the galaxy on the other side. Raise your sacrum and picture yellow light coming up from the earth and the other side of the galaxy; focus on the Lower Tan Tien and feel it fill with Chi. Do this three times. Go back down to the squatting position.

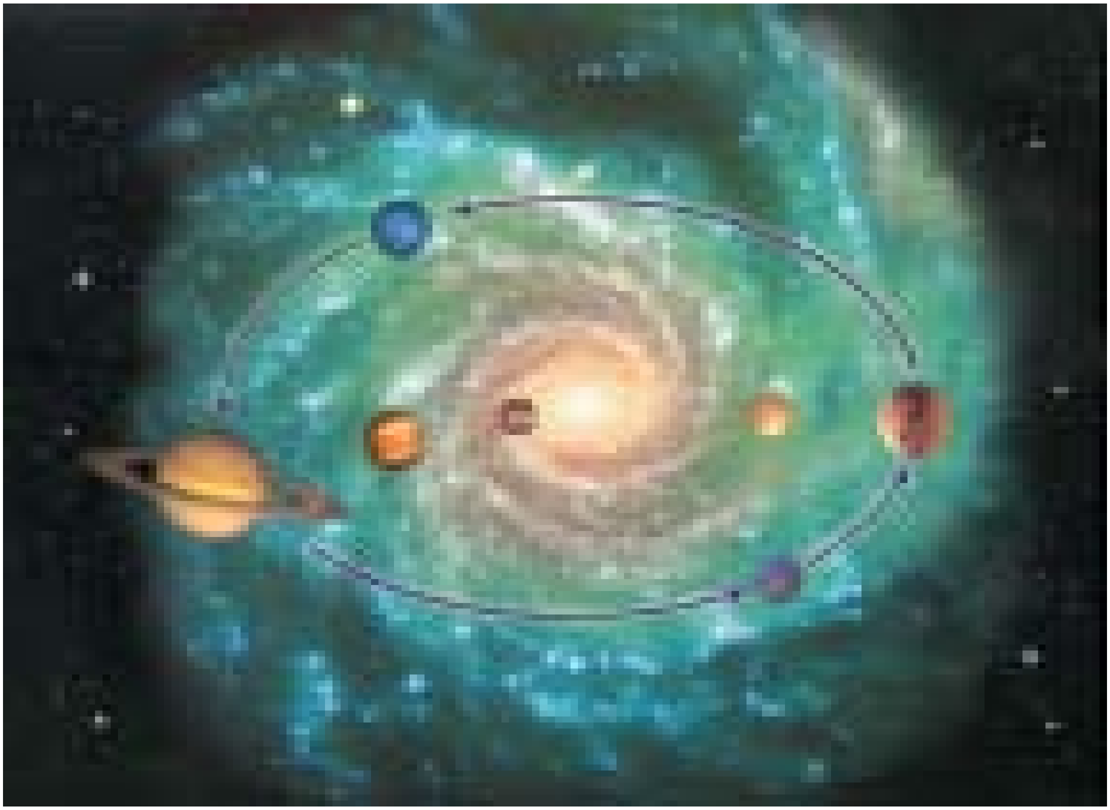

Fig. 2.4 *Galaxies Spiral Star and the Direction of Spiral*

8. Place the palms face down to the earth and be aware of the galaxy and the yellow light; make three to six circles to gather the Chi below. Gather the yellow Chi; touch your heels and feel the yellow Chi flowing up the leg bones. Feel the bubbling of your bones like an electric current running up through the feet, tibia and fibula, femur, pelvis, and spine. While touching your bone and bone marrow with your fingers, slide your hands up along the back of your legs; slowly come all the way up to touch the coccyx and the sacrum. Concentrate on your sexual center so the energy will flow and spread out to the sexual organs. Go back to the navel and gather the Chi into your Tan Tien.

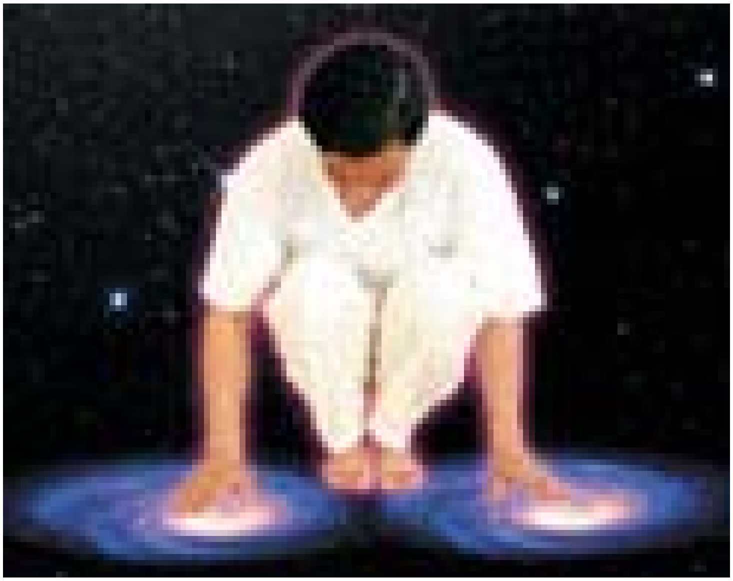

Fig. 2.5 *Hands circle to gather Earth Chi.*

Microcosmic Orbit

The Microcosmic Orbit Meditation awakens, circulates and directs Chi through the 'Governing Channel', which ascends from the base of the spine up to the head and the 'Functional' or 'Conception Channel', which runs from the tip of the tongue down the middle of the torso to the perineum. Dedicated practice of this ancient esoteric method eliminates stress and nervous tension, energizes the internal organs, restores health to damaged tissue and builds a strong sense of personal well being.

The Microcosmic Orbit is the foundation of Cosmic Healing Chi Kung practice. Each new practice is dependent upon the high quality of your meditations and your ability to perfect the Microcosmic Orbit. In order to master Cosmic Chi Kung, one must practice meditation daily. The meditations in the Microcosmic Orbit also strengthen the 'Original Chi' and teach you the basics of circulating Chi. They allow the palms, the soles of the feet, the mideyebrow point and the crown to open. These specific locations are the major points where energy can be absorbed, condensed and transformed into fresh new life force.

1. Focus on the Lower Tan Tien (the area where the Original Chi is stored, between the navel, kidneys and sexual organs). Feel the pulsing in this area, observe whether this area feels tense or relaxed, cool or warm, expansive or contracting. Notice any sensations of Chi: tingling, heat, expansiveness, pulsing, electric or magnetic sensations. Allow these to grow and expand. Then let this energy flow out to the Navel Center.

2. Use your intention (mind-eye-heart power) to spiral in the navel point guiding and moving the Chi. Let the energy flow down to the sexual center (Ovarian or Sperm Palace).

3. Move the energy from the sexual center to the perineum and down to the soles of the feet.

4. Draw the energy up from the soles to the perineum and to the sacrum.

5. Draw the energy up from the sacrum to the Door of Life (the point in the spine opposite the navel).

6. Draw the energy up to the mid-spine point (the T-11 vertebrae).

7. Draw the energy up to the base of the skull (Jade Pillow).

8. Draw the energy up to the crown.

9. Move the energy down from the crown to the mideyebrow point.

10. Touch the tip of your tongue to your upper palate, press and release a few times; then lightly touch the palate, sensing the electric or tingling feeling in the tip of the tongue. Move the energy down from the mideyebrow to where the tip of your tongue and palate meet.

11. Move the energy down from the palate through your tongue to the throat center.

12. Move the energy down from the throat to the heart center.

13. Bring the energy down from the heart to the solar plexus and feel a small sun shining out.

14. Bring the energy back down to the navel.

15. Continue to circulate your energy through this entire sequence of points, at least nine times. Once the pathways are open, you can let your energy flow continuously like a river of energy without needing to stop at each point.

16. Conclude when you wish by collecting energy at the navel.

Men: Cover your navel with both palms, left hand over right. Collect and mentally spiral the energy outward from the navel 36 times clockwise, and then inward 24 times counterclockwise.

Women: Cover your navel with both palms, right hand over left. Collect and mentally spiral the energy outward from the navel 36 times counterclockwise, and then inward 24 times clockwise. *For details of this practice see the book, "Awaken Healing Light", by Mantak Chia.*

Chapter III
Cosmic Chi Kung

Opening the Three Tan Tiens to the Six Directions

Introduction

Opening the Three Tan Tiens is a Chi Kung meditation that strengthens our connection to the universe, opening us up to the primordial force of the Cosmos and the energy within Nature. We are dynamically connected to the infinite. "As above, so below" is an echo of wisdom heard from sages and mystics throughout the ages. When we can connect to and absorb the energy that surrounds us, we are able to tap into the many splendors of the universe.

We exist because of the unique combination of the forces that are around and within us. The two main forces are electricity and magnetism. 'Bio-electro magnetism' is the western term for life force, and what the Tao refers to as Chi. For the last 5000 years, the Taoists have utilized this bio-electro-magnetic energy to enhance their way of life and establish a relationship with the universe. Bio signifies life, electro refers to the universal energies (yang) of the stars and planets and the magnetic force refers to the earth force (yin) or gravitational force present on all planets and stars. As we align ourselves with these forces, we become a conduit through which we can absorb and digest these energies through the body, mind and spirit establishing a direct connection with the universe. The Taoists recognized this connection and created the Chi Kung form of 'Opening the Three Tan Tiens to the Six Directions' to enhance our relationship to and our understanding of this connection.

Humans normally access bio-electro-magnetic energy through food and air. Plants take the universal energies of the sun and the magnetic energies of the earth and digest and transform them, thereby making these energies available to all living beings. Taoists believe that the food sources with the purest form of energy are the green leafy vegetables. These have taken sunlight directly into their cells. Rather than waiting until the energy in the universe is processed through plants, the Taoist goes directly to the source of this primordial energy. Through Chi Kung and meditation, the Taoists direct the energy of the universe precisely. Opening the Three Tan Tiens is a meditational Chi Kung exercise that focuses on how to directly tap into the source of energy all around us.

The Tao views human beings as lamps filled with fuel. Many people burn this fuel at very high intensity, without ever taking the time to replenish the oil in the lamp. Alcohol, drugs, tobacco and promiscuity all quicken the depletion of this fuel. The Taoist exercises strive to continually refuel the energy within. The Taoist recognizes that we are limited in our nature as human beings, unless we connect the sources of energy within the universe, thereby becoming infinite. So, within the limitations of our human nature, we constantly fill ourselves with the unlimited abundance of energy around us.

Through their internal quest, the Taoists discovered a doorway to the universe. The more we open our internal energy, the more we are capable of connecting to the forces of energy around us.

Human beings have amazing potential and capabilities. We are unique creatures in the way we use our minds and hands. Look at the world around you, the skyscrapers, the architecture, computers, technology and the myriad creations of man. All have come about through the combination of the mind with the hands. In the Tao practice, we use the mind and the hands, in both Chi Kung and meditation, to connect to the forces of the universe. We use the mind to project a pattern of energy into the universe, to connect to the force, and to bring this energy back into the body.

The mind can travel millions of light years in a few moments. Taoists discovered the unlimited potential of the human mind. If you picture something in your mind; an ocean, a sunset or a mountain, you automatically connect with that image. The mind can take you anywhere you imagine. With the proper training you can connect to the energy of nature and the universe and project your Chi to combine with it, drawing these forces into the body. The hands are called the touching force, and like antennae they can transmit the frequencies and vibrations from space. With the mind and the hands, each of us can journey into the boundless energy of the universe.

Relaxation, Letting Go, Surrender and Ego

In some religions and spiritual paths, there is a great emphasis on surrender and letting go. This is actually a form of relaxation. Taoist practices emphasize relaxation, letting go, and emptiness. When a person is relaxed, the muscles are open, the breathing is soft, and the energy can flow through the channels of the body. There is no resistance and no fighting. This allows the creative and higher forces to flow into us.

Most religions have a similar process. To contact the higher self and the higher forces, one needs to let go and surrender in order to reach the level where one can be in contact and become one with the higher forces. Through the surrender of control, one opens up and touches the forces of nature. However, if you continue to surrender and let go, you will lose the energy you have sent out. In the long term, this will gradually drain you. The force will suck the energy out of you, rather than help you to bring the energy into yourself. To avoid this, at the moment that one is in touch with the higher forces, one again becomes aware of oneself and his or her own energy. One is then able to project his or her own thoughts, intentions and patterns into the force, integrating the outer with the doubling or tripling inner energies. You can bring this force back to your place, your house and into yourself.

One must remain open and empty while using one's intentions, mind and Chi to draw the energy into oneself. What might initially appear to be a paradox is reformed as one does the practice and learns how to be empty and open and to simultaneously retain enough consciousness to draw the force into oneself.

Mind, Organs and Sexual Organs

It seems that some people are disconnected from themselves and from their sexual organs. The mind and organs are therefore separate. Taoism believes that the mind, body and spirit must work together. The results depend on a person's practice.

Brain

The brain can access and generate the higher forces, but storing this energy in the brain itself is not easy. We need to train the brain to increase its ability and capacity to store energy. The brain energy, when increased to a certain level, can enable more synapses to grow, and can help turn protein into brain cells. The Tao believes that with training and practice, one can learn to grow more brain and nerve cells, as well as increase the number of synapses in the central nervous system.

Organs of the Body

The organs can also generate energy, but much less than the sexual organs and the brain. They also have a much greater capacity to store and transform energy.

Sexual Organs

The Tao discovered that the sexual organs are the only organs that can generate a significant amount of sexual energy (life force). However, the sexual organs cannot store the energy efficiently. When too much energy is generated, considerable amounts have to be dumped out. And this is the best energy that one has. It is the ‘creative’ energy.

Three Tan Tiens

The Three Tan Tiens can also store energy, transform it and supply it to the brain, spinal cord, sexual organs and other organs.

The aim of Taoist basic training is to integrate the brain, sexual organs and other organs into one system. If the brain generates too much energy, it can store the energy in the organs. The excess sexual energy can also be stored in the organs and the three Tan Tiens. If the brain generates too much of the higher forces and we are unable to store this energy, we have to throw it away. It is like preparing food for one hundred people, when only one person is eating. The rest gets thrown away. In the same way, when too much sexual energy is produced and there is no place to store it, it is wasted. We do not have enough of this energy to be able to waste it. We have a limited amount of energy and of time.

Some practices just deal with the spirit and ignore the body and the sexual energy. These practices can generate a lot of energy, but when one is not connected to the organs, it cannot be stored anywhere. One will end up draining oneself. Some people practice sitting quietly, emptying the mind, with the whole body relaxed and calm, but very little energy is actually generated. When one gets deep into the practice, some people find it hard to come back to society, because they have no energy and their mind power does not work well. These people have to depend on others to support them.

In the Universal Tao, we are learning to create a sacred and holy temple within ourselves. With the simple practice of smiling to all the organs, we can integrate our bodies, minds and spirits. They are no longer separate. The sexual practice connects the mind with the sexual organs and the brain. The separation between these parts of ourselves is bridged and a synergy is created.

The Taoist practice provides us with the resources to extend beyond the realm of our senses. By tapping into our internal resources and channeling the energy around us, we can perceive much more than the senses normally report to the mind. We want to extend our perception from the limited perspective of the sociologically conditioned senses to the unlimited awareness of the universe. For example, our senses tell us that the earth is flat, that we are stationary, and that heaven is above us. In reality, the earth is a sphere hurtling through space at thousands of miles per hour and the heavens are above, below and beyond the earth in every direction. The goal of the Three Tan Tiens is to connect with the forces from the six directions - above, below, left, right, front and back - and draw all these forces into the body. Eventually, with practice one can draw upon many different energies and use them as needed, thereby giving form to the formless energy that is abundant in nature.

Opening the Three Tan Tiens

Opening the Three Tan Tiens to the Six Directions is just another one of the many resources the Taoist practitioner uses to connect with the universe. The practice combines the power of the mind with the extension of Chi. This combination allows our personal consciousness to directly connect to the patterns and matrices of energy in the universe. When we put our thoughts into the web of the universe, we transform the electromagnetic energy into a force that is accessible to us. The combination of mind power and energy is what allows us to establish a relationship to these creative forces and the high sources of energy.

Three Tan Tiens

Once we make the connection with the forces of energy in the universe, we then want to be able to store this energy in the body. Energy is like money, if you are making a million dollars a year and spend a million dollars, you have nothing left to use in the future. That is the way we live and use our energy in our society. We are spending more energy than we are saving and we are living on borrowed energy, paying very high interest. Our credit will run out very soon.

In the Tao practice, we store energy in the Three Tan Tiens. The Three Tan Tiens are the reservoirs of energy within us.

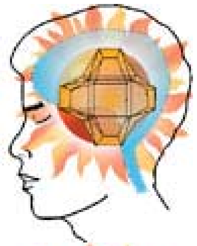

Fig. 3.1 *The **Upper Tan Tien** is in the brain (the **Crystal Room**, third ventricle), and when it is full of energy, the capacity of the brain increases. We store our spiritual intelligence, the mind here. All the Tan Tiens have both yin and yang within them. In nature, the yin and yang are present in all things. Day (yang) turns into the sunset, which turns to night (yin). It is very important to feel the qualities of yin within yang and yang with yin (sunrise/sunset). One quality does not exist without the other. They are inseparable qualities of the same force.*

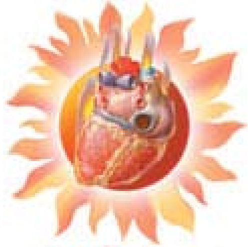

Fig. 3.2 *The **Heart** Center Tan Tien, between the two nipples, is the **Middle Tan Tien**. It is associated with the fire element. Yet, within fire there is always water. The original spirit (Shen) is stored here.*

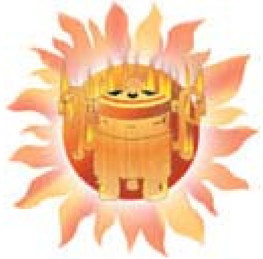

Fig. 3.3 *The lower abdomen at the **Navel** is like an empty universe, or ocean. We want to feel a universe of energy in the **Lower Tan Tien**. Within this universe or ocean, there is a fire, like a volcano under the ocean; 'fire under water'.*

The Three Tan Tiens refer to the three reservoirs of energy within the body. These reservoirs are places where we can store, transform and collect energy. The reservoirs are the source of energy that flows through the body. The meridians are rivers of energy fed by these reservoirs. The goal of Opening the Three Tan Tiens is to continually fill and replenish the energy of the Three Tan Tiens. When we are connected to the Tao, life ceases to be a struggle. Through the observation of nature, the Taoists learned to flow with the stream of energy and connect to forces in the universe.

In these practices, we use a variety of hand movements and body postures to open to the energy around us. We draw the energy from the six directions into the body, activate the three fires, open the Three Tan Tiens and circulate this energy in the Microcosmic Orbit.

Combined Practice

Warm - Up Exercises

Rotating the Sacrum

Rotating the sacrum is an excellent exercise to open the lower back and activate the spinal cord. Place one hand over the sacrum and the other over the pubic bone. Rotate the sacrum in a circle, 36 times in each direction. This movement activates the sacral pump.

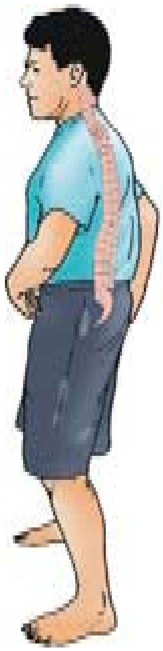

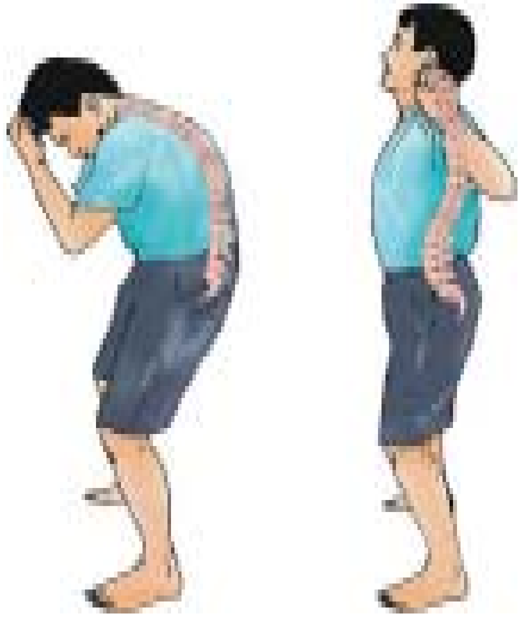

Spinal Cord Breathing

Inhale and expand the chest, arms bent at the elbows and extended to the sides of the body. Exhale, tuck the tailbone under you and round the back, bringing the elbows toward one another in front of the chest. Smile. Inhale, expand the chest, tuck the chin in toward the throat, push the chin back and raise the crown and bring the arms out to the sides. Repeat this back and forth movement 36 times. This movement activates the cranial and sacral pumps, and loosens all the joints in the spine.

Shaking

Shake the whole body loose, especially the joints, by bouncing up and down on the heels. Let all the joints open and relax. Also, shake the testicles and the breasts loose. This will open the sexual energy (the yin electric) and you do this on back as well.

The yin and yang are the negative and positive electric forces in the universe. The heart produces impulses (the yang electric) and the adrenal glands produce the adrenal hormone to stimulate the heart to beat. Of all the hormones inside us, this is the one that gives the vital force to our life.

When we concentrate on the sexual organs, the heart and the mind, we can make the yin and the yang electricity combine.

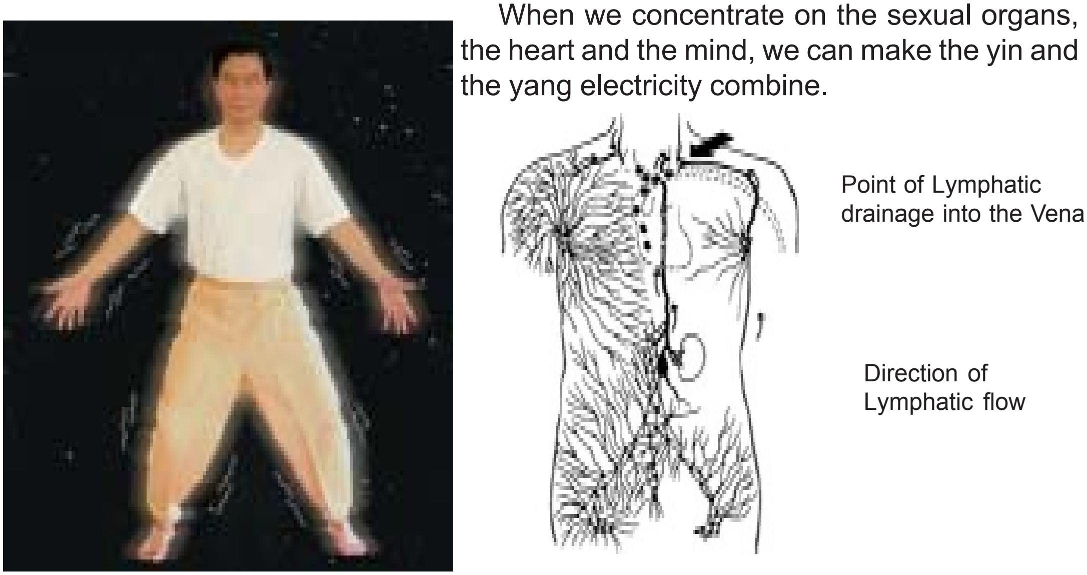
Point of Lymphatic drainage into the Vena
Direction of Lymphatic flow

Fig. 3.4 *Whole Body shaking and Lymphatic System*

True Breath - Skin Breathing

We exist because of the physical food and the unique combination of forces that surround us. The daily requirement of calories is about 6,000 units, while we only get 2,000 calories from food. The other 4,000 calories come from the forces around, above and below us. These forces that surround us are electricity, magnetism, cosmic particle energy, light, sound and heat. If we don't know how to absorb and transform this cosmic food, we need to depend on others to supply us. We then need to ask a priest, monk or holy person to give us our daily spiritual food.

The Taoists discovered that we can learn to absorb these surrounding and universal energies through the skin and the major energy centers. Absorbing energy through the skin is called the True Breath. This powerful energetic technique requires the Inner Smile and relaxation. The more one can relax, the more the body and the skin can open to the energy around us. The practice allows one to extend the mind, to touch the force and to draw that energy back into the body.

Daily Practice: Opening, Connecting and Activating Chi Within and Cosmic Charging

Paradoxically, in order to project ourselves out into the immensity of the galaxies and the universe to gather limitless resources of Cosmic Chi for healing, we must take the first steps of the journey within ourselves. In order to ‘go out’, we must first ‘go in’. The vehicle for this magical journey is powered by our ability to relax in mind and body. As we physically relax and let go of muscular and emotional tensions and joint and bone structures we gain access to the inner realms by turning on our very special subtle smile. It opens the pathways of the parasympathetic nervous system. This helps us to reduce the out flowing habits of our senses so that we can be more alert in sensing our inner universe.

This simple process of ‘going in’ enables us to develop our internal skills so as to safely ‘go out’ to the universe.

Practice

Relax and Let go - Smile to Connect with the Universe Within

Smile to mideyebrow, eyebrows, eyes, mouth, jaw, tongue, lips, cheeks, ears, shoulders, ribcage and brain. Let the relaxed sensations and the ‘Observer Mind’ (Upper Brain) sink down into the Lower Tan Tien.

1. Smile to the mideyebrow. Relax and let go. Smile to the eyebrows and let them grow long to the sides. Lower these relaxed sensations down to the Tan Tien.

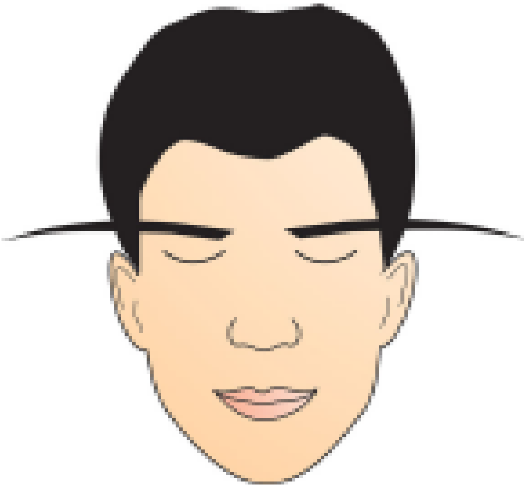

Fig. 3.5 *Smile and Relax to the Mideyebrow and think of the eyebrows growing long.*

2. Smile to the eyes: relax the eyes and feel how nice and cool they are. Let the eyes draw back in their sockets and start to sink down to the chest and gradually down to the abdomen, the home of your ‘feeling and awareness mind’.

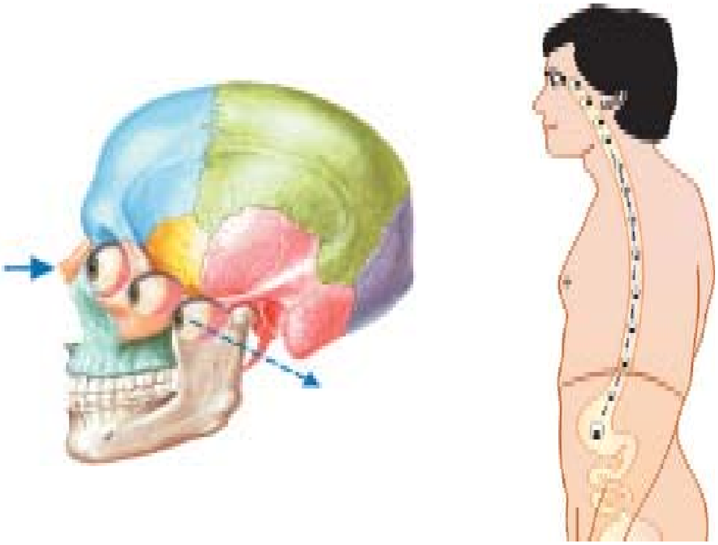

Fig. 3.6 *Smile and relax the eyes, sink into the eye sockets and gradually feel the eyes dropping down into the Abdomen.*

3. Relax the two broad muscles extending from the outer portions of the upper lips across the cheekbones, and lightly smile feeling their connection to the upper front of the ears. Gradually feel the ears growing ‘long’, up and down. Feel the ears grow all the way down and connect to the kidneys.

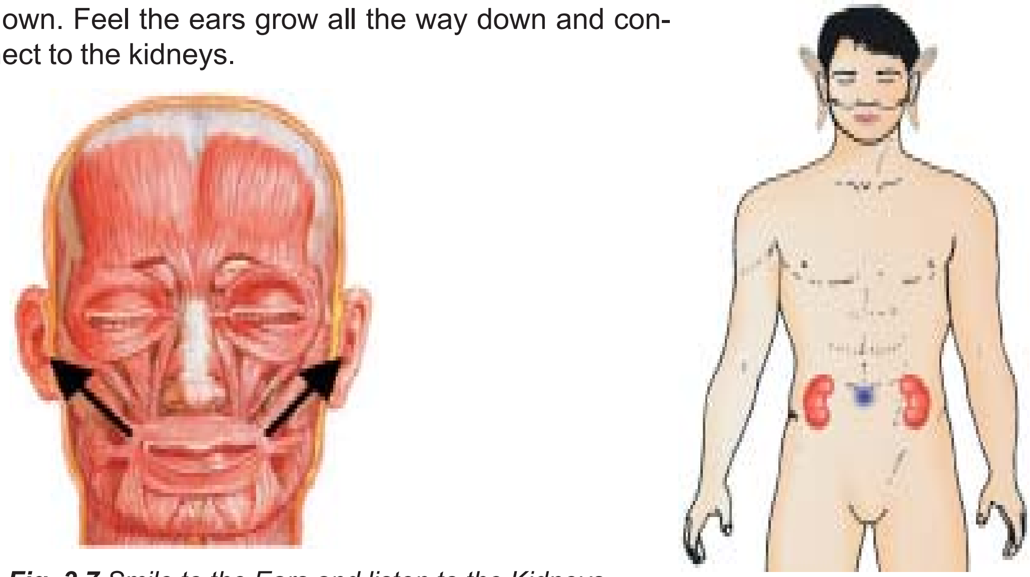

Fig. 3.7 *Smile to the Ears and listen to the Kidneys.*

4. Open your mouth and relax your jaw, separating the upper and lower teeth. Feel the jaw relax. Once the jaw relaxes the shoulders will relax and drop down. Continue to feel the jaw relax until you feel the saliva start to come out. Relax down to the rib cage. Feel the rib cage drop down, softening all the joints, relaxing down to the Lower Tan Tien. Let the tongue relax back in the mouth. Feel the tongue start to drop down into the throat to the chest and all the way down to the navel sinking the 'floating' sensation down to the Lower Tan Tien.
5. Smile to the shoulders and relax until you feel the shoulders drop and the rib cage relax.

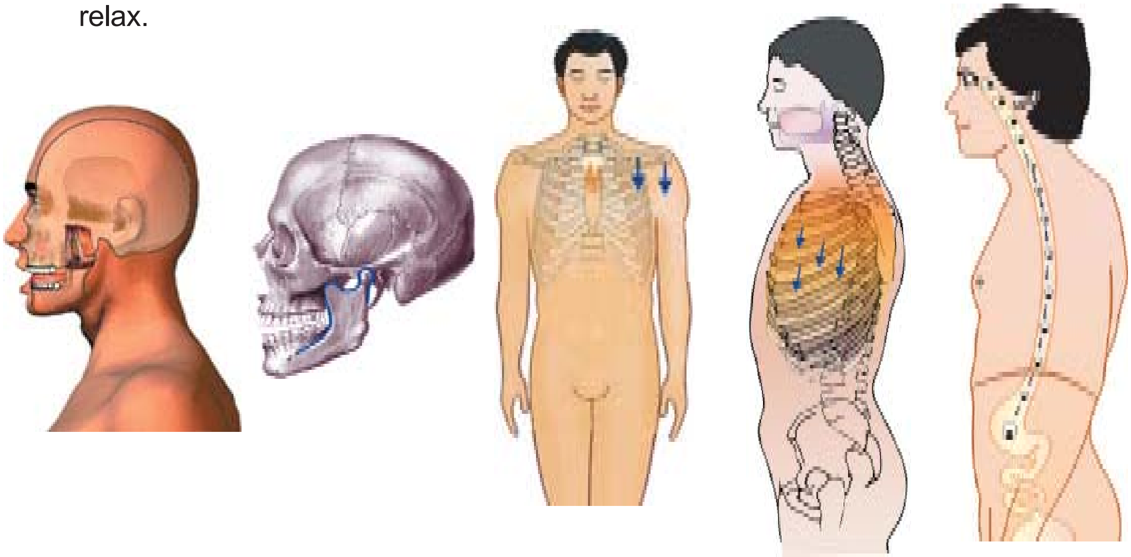

Fig. 3.8 *Relax the Jaw; and the Shoulders and Tongue will relax down to Lower Tan Tien.*

6. Lightly close the lips, but keep the teeth slightly separated. Physically begin a child-like smile with the corners of the mouth gently uplifted and the outer edges of the eyes softly crinkled up. Breathe through your nose.
7. Smile into the brain and empty the upper (observing) mind down to the Tan Tien.
8. Become aware of your inner universe as a big empty space. Keep on sinking – sinking down, into the darkness of your empty space. Keep sinking and experience the vastness until you get closer and closer to the center, the 'original force.' Stay relaxed and alert so as to be able to see one dot of light. It becomes a galaxy spiraling inside you.

Do the above practice several times until you become familiar and comfortable with it.

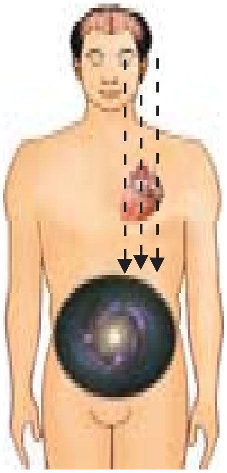

Fig. 3.9 *Bring the Senses down to the Tan Tien.*

Tan Tien Consciousness: Second Brain

Already terms have been used like ‘upper mind’, ‘feeling and awareness mind’ in the abdomen and lower ‘Tan Tien’. In the next phase of practice, the ‘conscious mind of the heart’ is introduced. For those readers who have not been updated on the scientific and medical revelations of the recent era about a ‘second’ and ‘third’ brain, we offer the following information.

Along with this, there is the long-standing Taoist practice of cultivating and training consciousness in the Three Tan Tiens, especially the Lower Tan Tien. There is my own experience in participating in brain wave measurements and ‘brain energy potential’ research in recent years. The combination of all these areas of knowledge and experience in our time, makes Cosmic Healing practices accessible for our minds while also being very ‘down to earth.’ Now let’s talk about the Lower Tan Tien.

In addition to its importance as the control center for the mechanics of the physical structure of the body, the Lower Tan Tien also houses a treasure of even more far-reaching significance. It has been a well-kept secret in the western world, as well as in most of the rest of the world: our Second Brain. Most of us who have had Taoist training in Chi Kung, Tai Chi, various Chi meditations or healing practices have often heard the reminder, ‘Be aware of your Tan Tien.’ But, do most people really ‘get’ the meaning of the injunction to always be conscious of the Tan Tien? Probably not. Further, do we use our Second Brain as much as we can? Definitely not.

In western terms, do we actually understand “Be aware of your Tan Tien” as being a way to train consciousness and awareness, like educating a brain in the abdominal area, in our Tan Tien? Probably not. Throughout the world, there are institutions to train the brain in the head. That is good. But, what about training the ‘second brain’ in the abdominal area? I didn’t think of it exactly like that in western terms, either, even though that is exactly what I have been doing all of my life in my Taoist training and teaching.

Personal Revelation

Suddenly I came to understand a few things which are so simple and so important, and that’s what I am going to share with you. It started in 1994 in Los Angeles when a clinical psychologist there, Dr. Rhonda Jessum, wanted to start testing me. I agreed to do the testing, but the machines at that time didn’t tell me much. However, it was discovered that when I did the Inner Smile meditation, my brain waves went down dramatically – but at the same time, my ‘beta’ waves increased to a very high level. That means the waves showed that I could be driving a car, but that my brain should have been resting and sleeping. So the researchers said, “Hey, how did you do that?” I didn’t understand either – because it was not clear to me either.

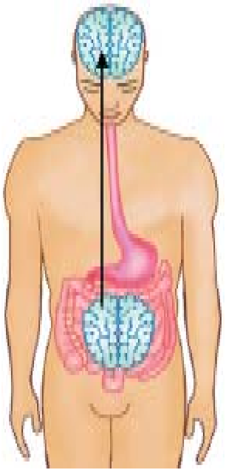

***Fig. 3.10** First and Second Brains are linked.*

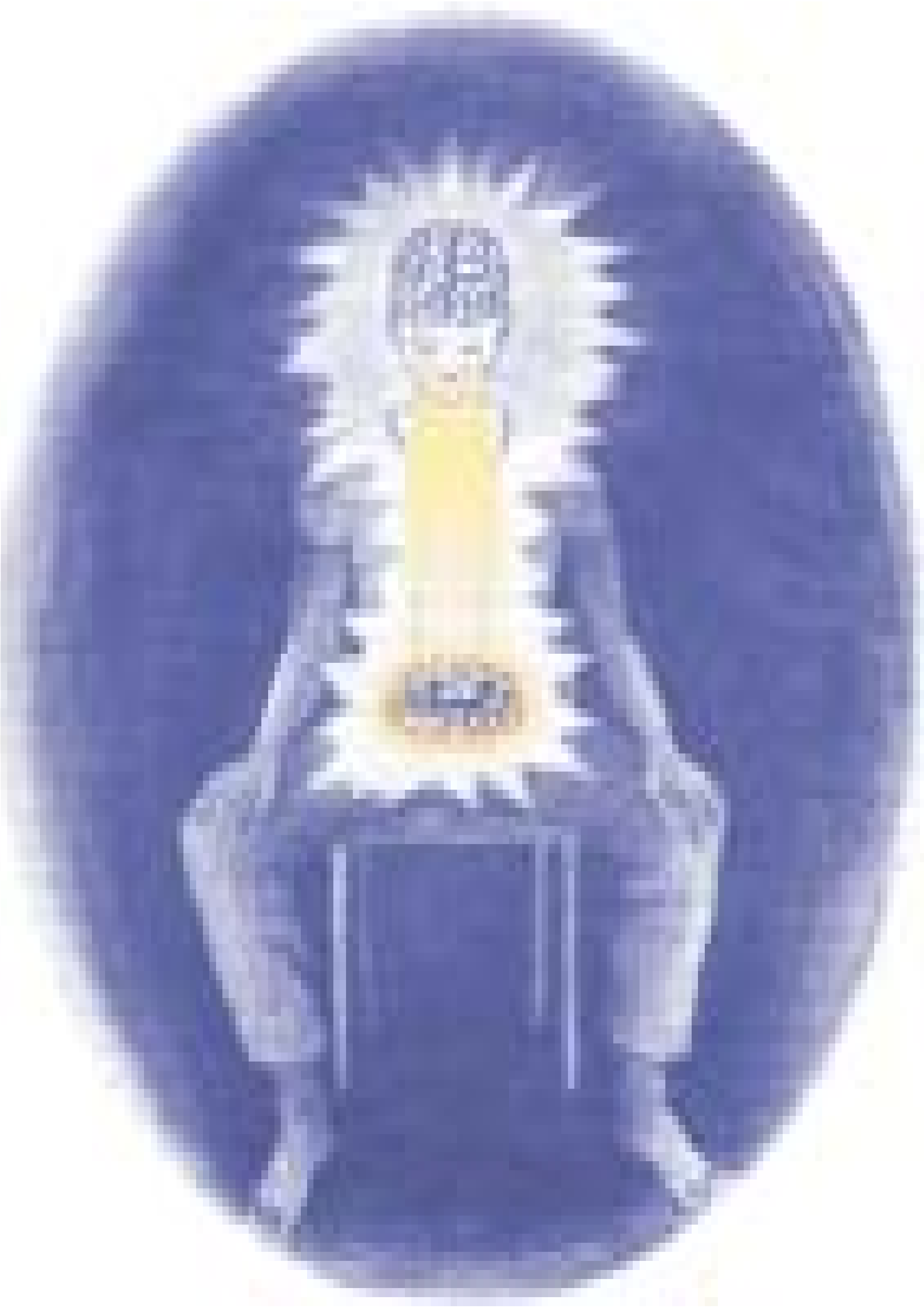

Fig. 3.11 *Second Brain*

After that, I was invited to start testing in Europe by the Institute for Applied Biocybernetics and Feedback Research. My name is becoming known there because I teach in Europe a lot. One of the biggest institutes for training top athletes in Europe is in Vienna. They developed an instrument that can measure the brain's potential energy, which represents all the energy in the body. The doctors also said that this proves to the west that Chi exists, there is energy and a life force running in the body. That instrument picks it up in the brain and indicates how much potential brain energy you have. It also determines how much energy you have for the whole day and how much of that energy is for mental clarity and body power. They use these instruments to measure the athletes.

The study measures brainwave frequencies during four practices.

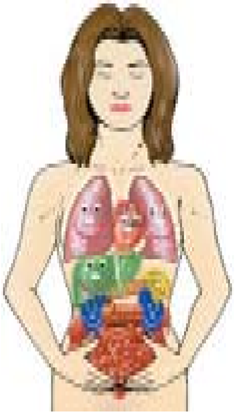

Fig. 3.12
Cosmic Inner Smile

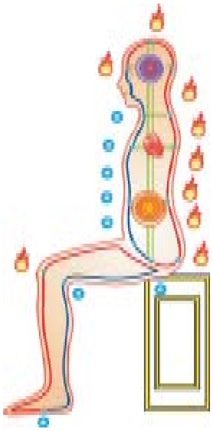

Fig. 3.13
Cosmic Orbit

Fig. 3.14
Six Healing Sounds

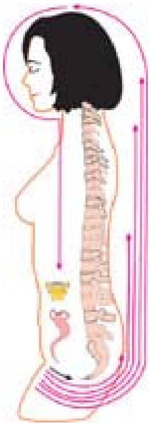

Fig 3.15
Orgasmic Upward Draw

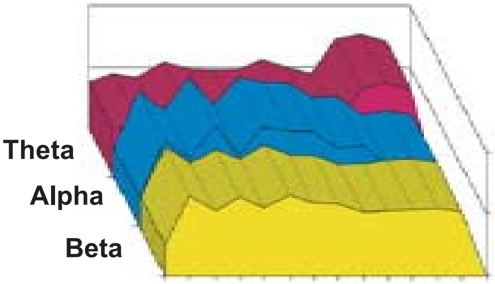

Fig. 3.16 *Alpha and Theta Waves increasing during the Cosmic Inner Smile Meditation.*

Fig. 3.17 *Ultraslow Brain potentials measurement while in Vienna.*

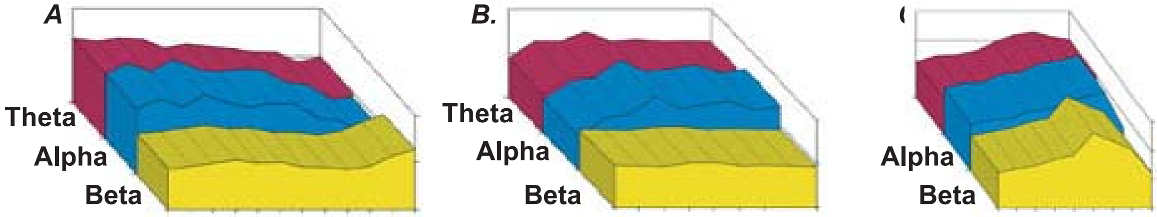

Fig. 3.18 A. *Altering dominance of alpha and theta waves during the Cosmic Orbit Meditation.* ***B.*** *Cosmic Healing Sounds Meditation, results in increases in alpha and theta frequencies. There are minimal beta brainwaves; the brain reaches a state of stillness and internal focus.* ***C.*** *Orgasmic Upward Draw Meditation increases the levels of alpha and theta.*

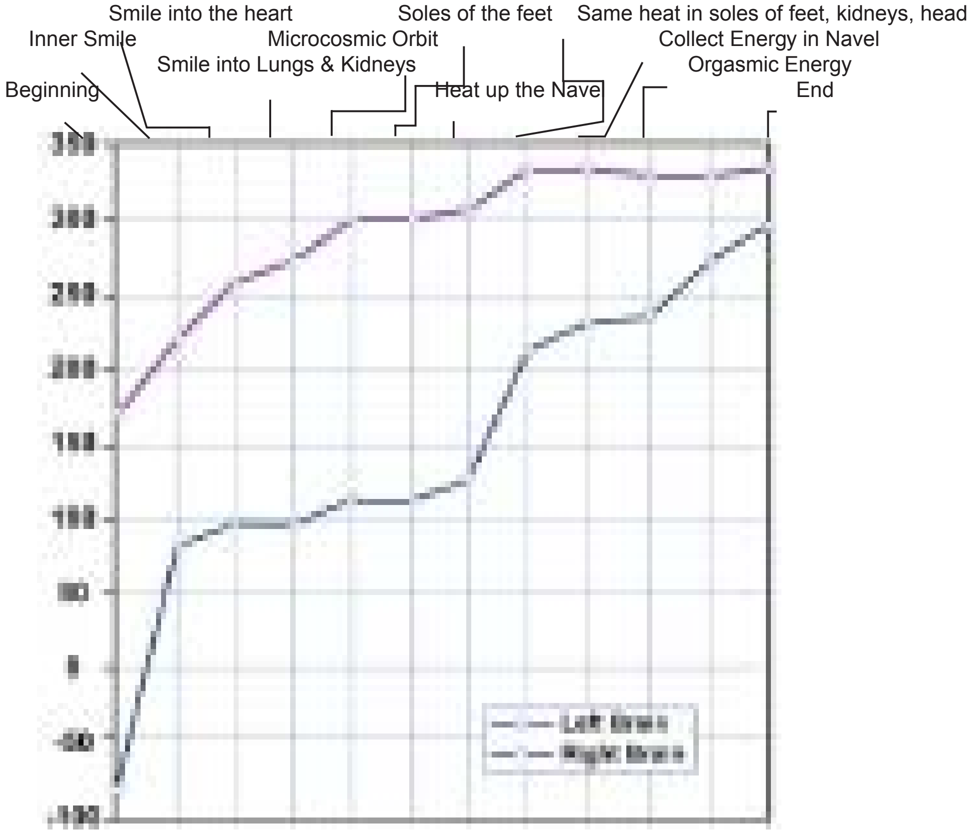

Fig. 3.19 *Left and Right Brain Charts*

It's interesting because the energy I have been describing in my teaching is exactly what they were measuring during their testing. So I went there and did the measurements with them. I just did the Inner Smile, smiling into my abdomen. They picked up these readings very quickly and said, "Your brain waves are going lower, lower and resting – and you are nearly in the sleeping state." At the same time, the muscle tension was very low, the heartbeat was very low and the skin resistance was very low. After that, I surged the energy up to the brain and they started to see that the energy actually charged up there. When we are thinking, worrying or feeling anger, shame or guilt – the energy level in the brain actually decreases, and the brain doesn't get charged up. They were quite amazed, and said, "Hey, this is what we're looking for!"

They asked me what I was doing. I said, "I'm smiling to my abdomen." They kept on talking to me and asking me questions. They discovered that this brain (in the head) was not very active at all, meaning that there was not much activity in the brain. It was still in a very light resting state. But, how could I answer their questions? They said, "Hey, look! Master Chia is talking to us in his sleep. How can he talk to us in his sleep?"

After that, I said, "Oh, I understand now." Because, throughout the whole practice of the Tao is the injunction, 'train the second brain in order to use the second brain.' It took the westerners a long time to understand that. So, when the article about the 'Hidden Brain in the Gut' came out in The New York Times,[1] I started to understand that. You can be miserable; you can be happy; you can feel all kinds of feelings. But according to this article, they had also discovered that this brain in the gut, the enteric nervous system, can do a lot of functions. It says that this gut brain can send and receive impulses; it can record experience and respond to emotions. So, it is like a brain. In this article, they had only discovered that the large intestine and the small intestines have the same neurons as the brain cells. After that article, a new book, *The Second Brain,* was published.[2]

Now in medical science, I might add, they have begun to discover consciousness in the heart. They found that the heart can record a whole event, and it has its own brain. Dr. Paul Pearsall has written a new book, called *The Heart's Code*.[3] They have found that people who have a heart transplant actually get the emotions of the donor. They actually experience whatever the donor experienced. One of the cases is that of a girl who was brutally killed, and they didn't know who killed her. But, the heart was OK; so, it was transplanted into another girl. After that, the girl started to get nightmares and described somebody killing her. Then she described how the killer looked. Finally, the mother took the girl to a psychiatrist, and he in turn contacted the police. The girl gave the police an exact description and a police artist drew a likeness of the killer. With the information provided by the girl, the police were able to go and arrest the man. Afterwards, when confronted with clear details of the crime, the man confessed that he was indeed guilty of this crime. So, from that experience, medical scientists and others came to realize that the heart can record all of an event and remember it.

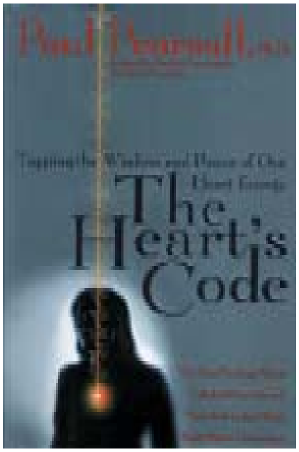

Fig. 3.20 *Heart's Code*

---

[1] "Complex and Hidden Brain in the Gut Makes Stomachaches and Butterflies," The New York Times, section C1, Tuesday, January 23, 1996

[2] Gershon, Michael. *The Second Brain* (New York: Harper Collins Publishers Inc., 1998).

[3] Pearsall, Paul, Ph.D. *The Heart's Code* (New York: Broadway Books, 1998)

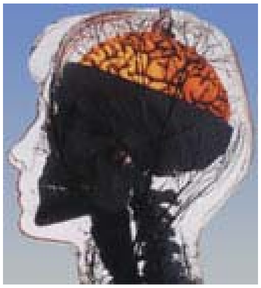

The Lower Brain consumes less energy and can do a lot of daily work, like send and receive impulse records and experience and respond to emotions. When you lower down the Upper Mind, it will also lower the blood pressure and anxiety.

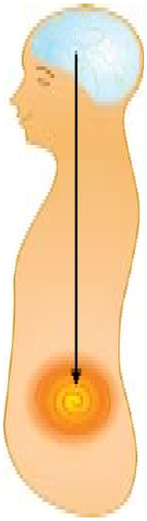

Fig. 3.21 *Abdomen has it own Brain: the Awareness Mind.*

On the second page of the Times' article, it says that even the large intestine is loaded with neurons. The question was posed, 'Can it learn?' But I say, "Hey, this goes back to 4,700 years of the Tao practice, which says: Train all the organs; train them how to do different things." You can rest the head brain when you're not using it – and use the brain in the gut. Why is it so important? Because the head brain is a 'monkey mind' riddled with doubt, shame, guilt and a suspicious mind. It is always thinking. Worrying, figuring things out things, head trips – it just keeps on, all the time. To be a God, you must let go of the past and empty the mind. Now we are in the information age; anything flits in – we think. Just one word flits in and we start an entire string of thought. When somebody says one bad word to you, you can think and figure it out for three days and three nights. 'How am I going to get revenge? How am I going to get revenge?'

Scientists have discovered that when people worry too much, thinking, planning, etc., this brain actually uses up a lot of energy. Depending on the type of person– some people are very physical types who use very little brain energy, but their physical body consumes a lot of energy. However, most people think and think and think. So the scientists give an approximation, by way of a percentage. This is just a number for comparison, not an exact number. They say that this brain in the head can use up to 80% of the body's energy. So, what happens is you have only 20% left for the organs.

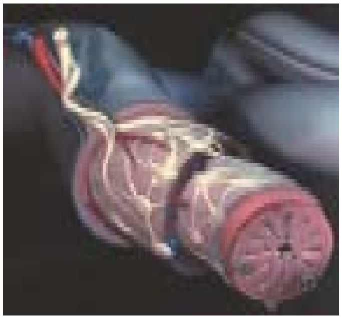

Fig. 3.22 *Cross-section of the neurons in the Large Intestine.*

According to the information about the brain in the gut, they have discovered that the brain in the head and the brain in the gut can do some similar jobs. For example, this brain in the gut is the emotional and the feeling brain. In the west you have the expressions, ‘I feel you from my gut’ and ‘I have a gut feeling about something.’ Why do people mention a gut feeling? Obviously people have some feeling in their gut. It’s very interesting that the whole Tao practice is feeling, awareness and consciousness – using this gut to feel, to be aware, and to be conscious.

You can rest, relax the brain in the head by using the ‘brain in the gut.’ This is the first step. The first thing we learn in the Tao is to forgive and let go. When we keep on remembering past negative emotions, we will stop seeing the truth. To let go of the past is to empty the mind and use the abdominal mind, the awareness and consciousness. Here’s an interesting point in the way of the Tao: This gut brain can do a lot of simple functions that are similar to the functions of the brain in the head. This is a feeling and awareness type of functioning similar to many of our ‘right brain’ functions. However, we need to use the brain in the head in order to perform complex functions such as reasoning, making plans and making complex calculations. For rational functions, we need to use the brain in the head for the ‘left brain’ functions.

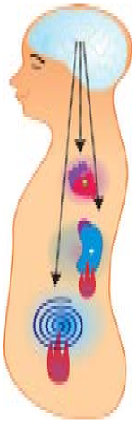

For our daily life of consciousness, awareness and feeling, we can use either the brain in the gut or the brain in the head. When we use the upper brain less, it will become charged with more energy and its power will be increased – and more power will be available to the body. That is why we say in Taoism, that we have to train the brain in the gut, so that we can use it when the brain in the head is resting. When the head brain is resting, it can be recharged: brain repair and maintenance occurs. Also, it can grow new brain cells. With more charging we have more power for creativity or whatever we want to use it for. If we like, we can use it to develop our higher spiritual nature.

Fig. 3.23 *When you are not using the Upper Brain allow it to rest by sending it down to the Lower Tan Tien.*

The simple thing is, that for the same job that the head brain or the gut brain can do, the head brain charges you eighty dollars. The gut brain only charges you twenty dollars. So, which one do you want to use? No, of course we are not that silly to use this overpriced person for the same work. But when we come to our own actual life, we don’t know how to do it. We always use this mind, this high-priced brain. To add to this, we continue using it and using it, until the brain energy is completely consumed – no more energy. When you get to a certain level the brain is empty. These brain energy measurements are not only for indicating mental energy levels, they also represent energy conditions for the whole body and spiritual energy.

Whenever I smile down, the brain waves go lower and lower very quickly – and the transformed energy from the Tan Tien and organs charges up the brain in the head! Just by flexing the facial muscles into the mode we use with a genuine smile, we can produce the effects on the nervous system that normally goes with the natural spontaneous feeling. We can actually make ourselves relaxed and happy by taking advantage of this built-in human mechanism. It's natural. Just do it! Learning to smile down to the abdominal area and maintain the awareness of the relaxed, smiling sensation in the Tan Tien is the first step in training the *Second Brain!* Think about it: Pure awareness and consciousness can change attitudes and emotions carried in the DNA.

Remember: Number 1: "Empty your mind down to the Tan Tien, and fill the Tan Tien with Chi." An axiom in the Tao is, where the mind goes the Chi follows. And, Number 2: "When your mind is empty, it will be filled!" This means that when the organs have extra energy, the extra organ energy will rise up and fill the brain with Chi.

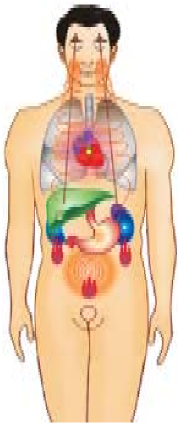

***Fig. 3.24** When the mind is empty, transformed energy from all the organs can charge the brain with Chi.*

Practice

Three Minds into One: 'Yi' Power

Smile into your heart. Lower the Observation Mind, then the Heart Mind down to the Tan Tien. With the Feeling/Awareness Mind, Spiral the three minds into one.

1. Smile into your heart. Make it feel soft. Feel love. Feel joy. Feel happy and feel compassion in your heart. Feel the heart spiral.
2. Spiral in the upper mind. Lower your upper mind, the observation mind, down to the Tan Tien in your navel area.
3. Turn the consciousness in your heart, activated by your love and softness, down to your Tan Tien.
4. Combine the upper and middle minds with the energy of the feeling/awareness mind of the entire nervous system – the Second Brain in the abdomen. Merge the three minds into one mind by spiraling their energy, blending them together as one in the Tan Tien.

When the three minds merge into one they become the 'Yi' power, the three-mind power – activate the spiraling at the mideyebrow, crown and the soles of your feet – expand to the universe in six directions: up, down, front, back, left and right. Once you get it, the Yi, we will start to activate the six directions and the three fires.

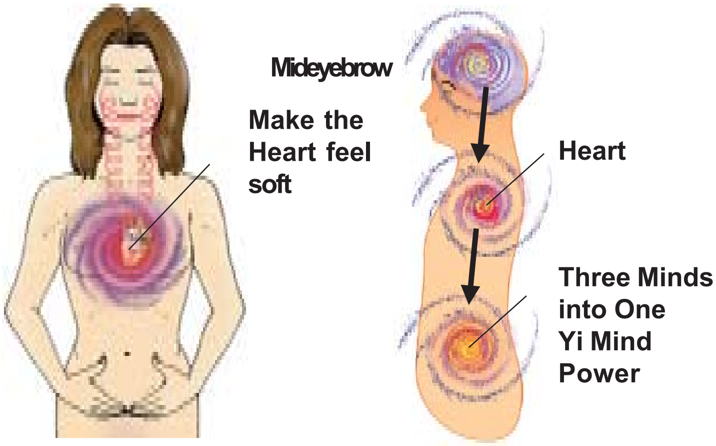
Mideyebrow
Make the Heart feel soft
Heart
Three Minds into One Yi Mind Power
Crown

***Fig. 3.25** **A.** The Heart softens like a rose blossoming.*

***B.** Spiral the Upper Tan Tien into the Lowest Tan Tien and Three Mind 'Yi' Power.*

Activating the Six Directions and the Three Fires

This powerful energetic technique allows one to extend the mind, to touch the force, and to draw that energy back into the body.

The Six Directions teaches you how to expand your mind and Chi for receiving healing power. By practicing the Six Directions daily you will increase your healing and cosmic power. Turn your mind and Chi into the cosmos, multiply them and draw them back.

Direction Below

When you achieve the three minds into one mind begin expanding into the Six Directions; press your hands down and start with the low direction. Picture yourself standing on the earth and you expand yourself very far away – deep down into the earth, very, very deep down into the earth. Your hands become very long; your feet become very long — they go all the way down into the earth and out past it into the galaxy below on the other side.

Push. When you push, connect with the galaxy below — pull and think about your Tan Tien filling with Chi. Push and pull. Push and pull. Fill your Tan Tien with Chi.

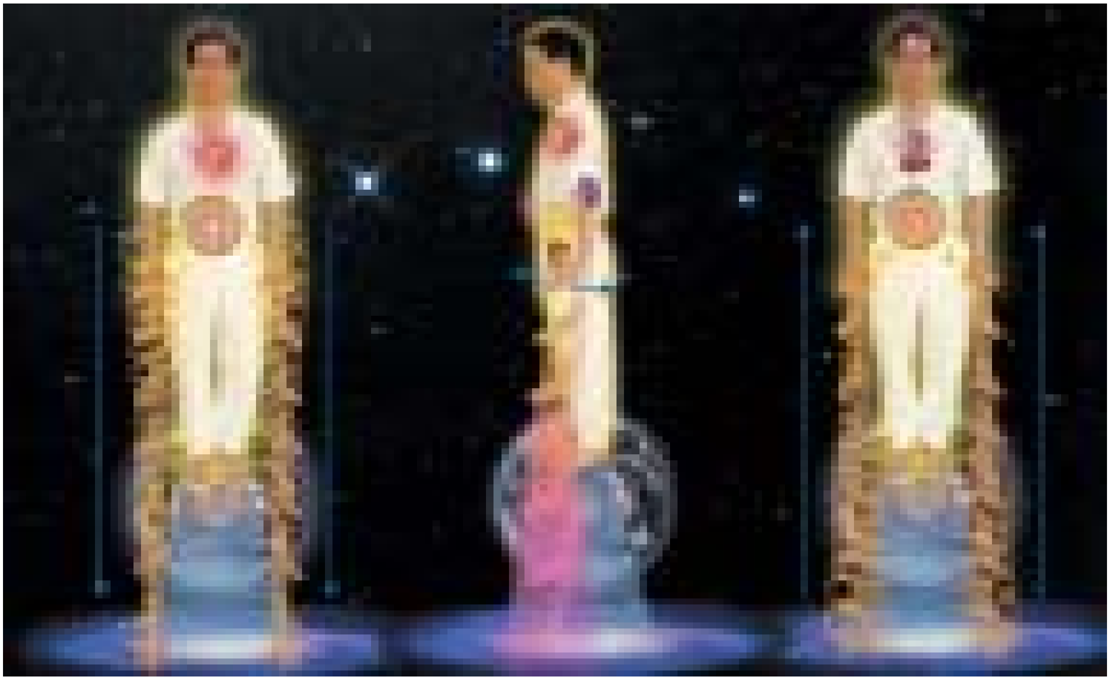

A. ***B.*** ***C.***

__Fig. 3.26 A.__ Hands expand through the earth and to the galaxy below.
__B.__ Hands push forward and pull back. __C.__ Smile to the primordial Chi from the universe back to the Lower Tan Tien and fill the Tan Tien with Chi.

1. Stand, feet together. Put you hands down, parallel to the ground. Expand your hands very far away and your mind very far away smiling into the ground. Continue expanding your hands, feet and your mind very far away beyond the earth below. Go down through the galaxy, way beyond to the primordial force. It's just like you are extending all the way to the primordial force before 30 million years ago.
2. Push, moving the hands forward six inches only.
3. Pull, moving the hands back by the sides. Think about your Tan Tien – Chi coming to the primordial force in your Tan Tien. Smile to your Tan Tien, dark, deep and vast.
4. Push: touch the primordial force in the universe.

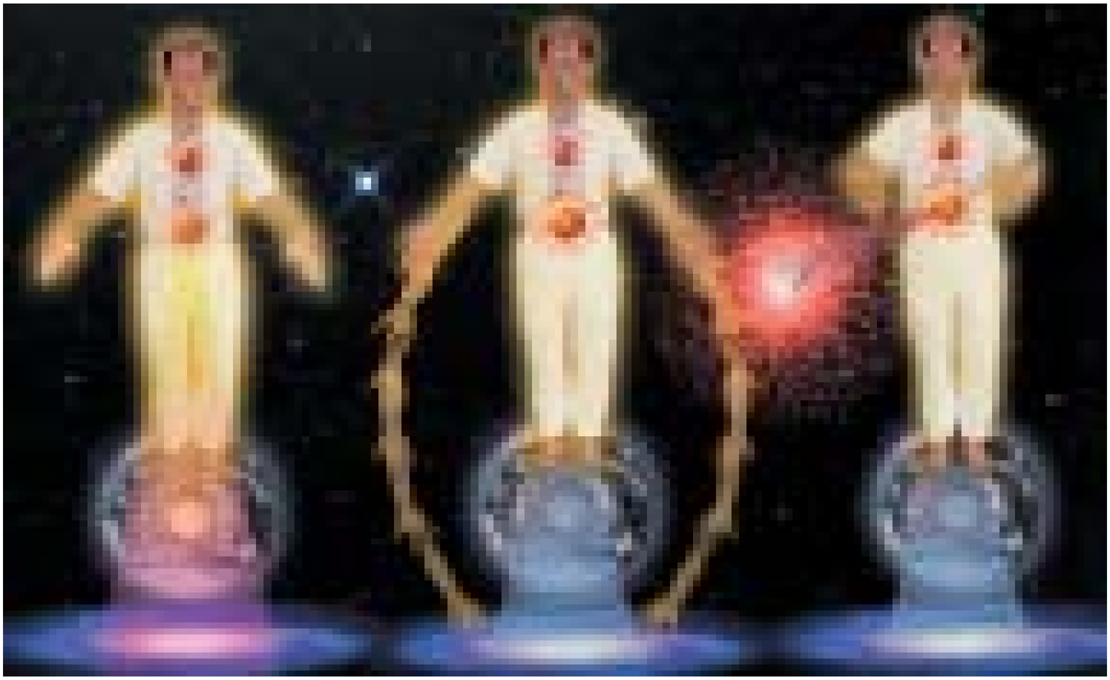

Fig. 3.27 A. *Be aware of the universe in front.*
B. *Expand your hands very big and long to the universe in front.*
C. *Hold the fire ball to activate the Tan Tien Fire.*

5. Pull back to your Tan Tien the dark primordial forces with your hands.

Push and pull: It's just like you go to an empty space – vast. Then you come back to your Tan Tien – also empty, just like the primordial condition before anything existed. That is where all the forces come from. Push and pull 3 – 9 times.

Front Direction and Tan Tien Fire

Next, be aware of the front direction; a huge fireball appears in front of you.

Open your palms: scoop up the Chi, scoop up the fire. Bring the fire into your Tan Tien. Activate the Tan Tien Fire.

1. You start with a small dot of light inside you. Expand your awareness smiling to the universe in front of you.
2. Become aware of a big fireball in front of you. Fell your hand become bigger and longer. Scoop up the fireball. You may close your eyes to help your inner sensing.
3. Use the fireball to light the fire in your Tan Tien. Feel the fire burning in the darkness, the 'fire burning under the sea.'

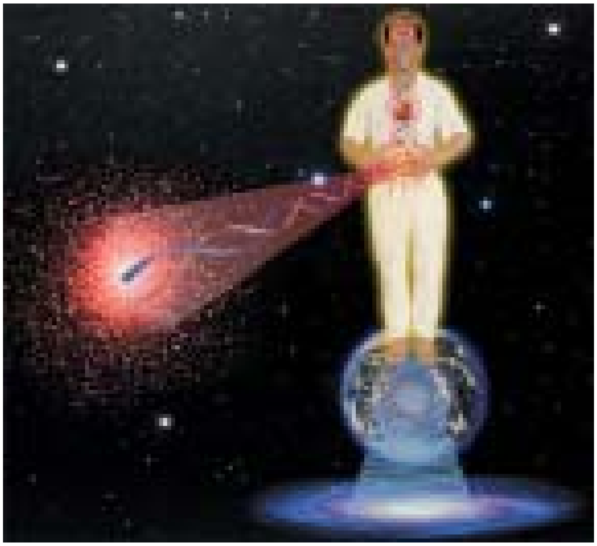

Fig. 3.28 *Smile to the Burning Fire.*

Back Direction and Kidney Fire

Be aware of the back of the Tan Tien, the Door of Life and the back or rear direction. Extend your mind very far away to the 'back'. Scoop up the fire and light up your 'Kidney Fire'.

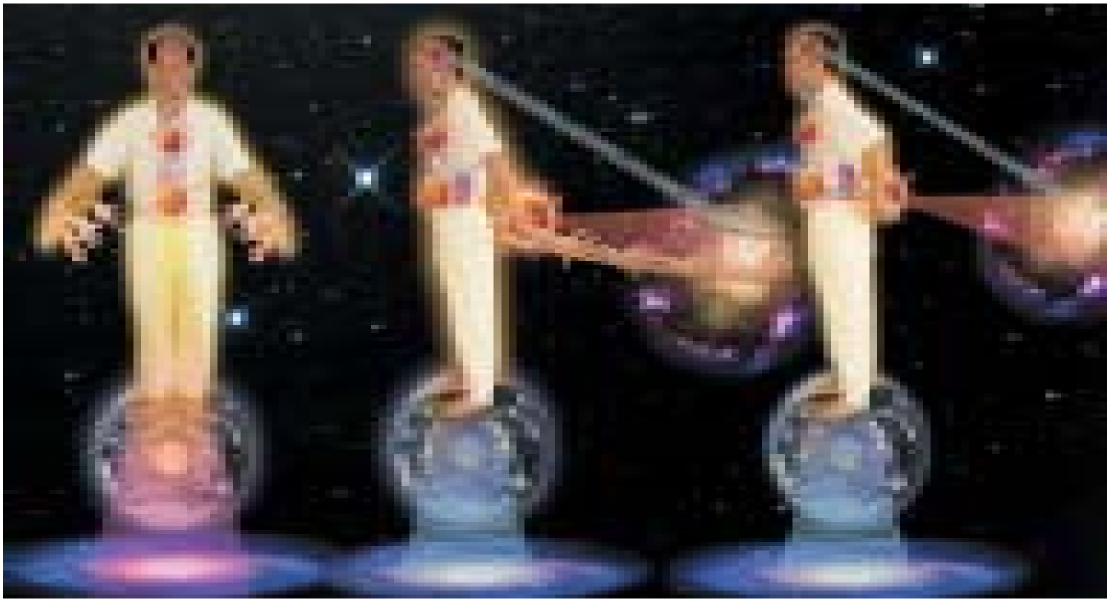

Fig. 3.29 *Be aware of the back direction, move the arms toward the back of the universe and scoop up the universal fire.*

1. Expand your awareness all the way to the back. Move the arms toward the universe behind you.
2. Touch the universe; scoop up the fire. Activate the Kidney Fire. Maintain your awareness in the Tan Tien and expand smiling out to the universe. The energetic spiral glows in the Tan Tien. Spiral in the heart, spiral in your crown and spiral in the universe.

Heart Fire

Raise your hands up under your armpits, and feel yourself holding the two fireballs. Touch the heart by extending the fingers energetically in from the sides; feel your hands extending into your heart and very far away. Activate the Heart Fire.

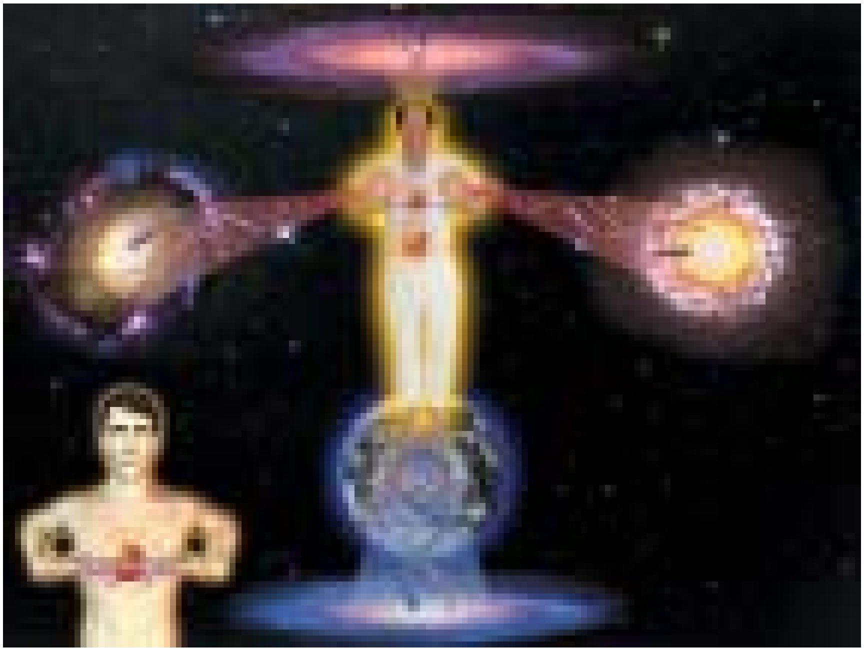

Fig. 3.30 *Activate the Heart Fire.*

1. Move your hands up under your armpits and extend your fingers deep into your heart and very far away.
2. Tan Tien and the Universe: you are connecting to the ‘charger,’ charging more fire into yourself.
3. Feel your heart soft in the center as you smile down. Feel the warmth of the fire energy of love, joy and happiness in the heart.
4. You feel the connection with the unconditional love in the universe as you keep your heart consciousness in your Tan Tien and extend your awareness out to the universe.

Sacred Fire *(Chi Fire)*

Connect the Three Fires to combine into one Fire: *Heart to Kidneys to Navel to Heart*.

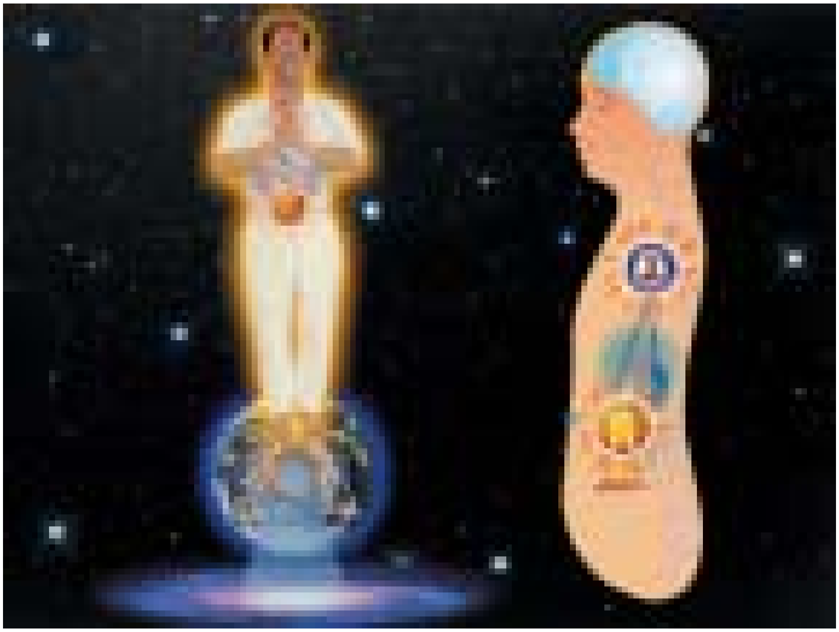

Fig. 3.31 *Connect the Three Fires and Sacred Fire.*

1. Move your hands together in front of your heart. Hold your hands together in front near the heart and feel the fire burning in there. Connect the Heart Fire to the Kidney Fire, the Kidney Fire to the Tan Tien Fire and back up to your heart; connecting them as one triangular Sacred Fire, circulating the Chi at 1,000 revolutions per minute, 10,000 R.P.M, 30,000 R.P.M and 60,000 R.P.M, doubling or tripling their collective power.

Open the Third Eye

Now, extend your hands out to the front, very far away – pushing, pushing, pushing. Push. Turn your palms inward, and extend your middle fingers inward toward your third eye. Picture a crack in the middle of your forehead, and pull the crack open. Feel the light from the heavens opening it and feel the light from the heavens shining into your brain.

1. Open your palms. Open your eyes, dimming the eyes. Look to the universe. Extend your hands to the front, palms vertical. Extend the arms from the scapulas. Smile and touch the universe.
2. Turn your palms inward and extend your middle fingers inward toward your third eye.

3. Picture a crack in the middle of your forehead and let the heavenly light shine into the brain; pull the crack open and let the light reflect into the organs.
4. Close the third eye. Again, open. Pull: open-close-open. And, close. With the third eye open, the light from the heavens shines into your brain and reflects down to all your organs. Open and close the third eye 3 - 9 times.

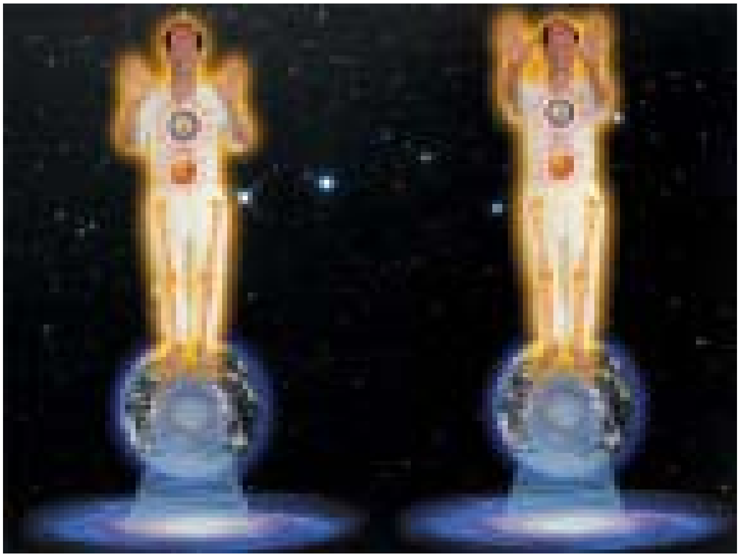

Fig. 3.32 *Look as you smile into the Universe in front of you.*

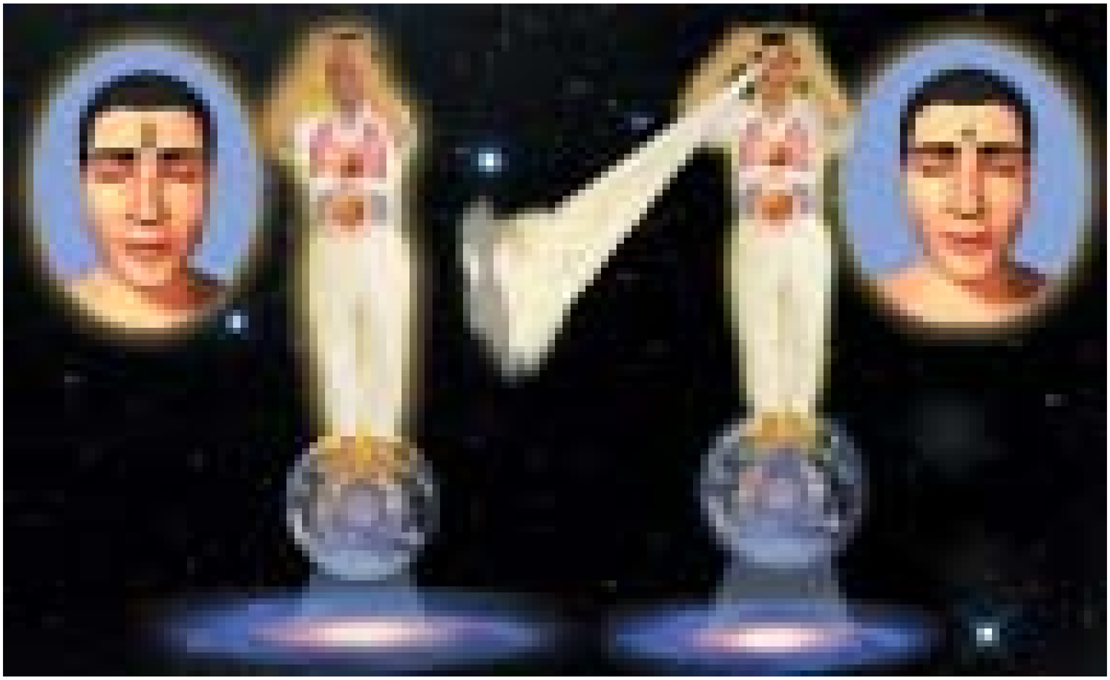

A. **B.**

Fig. 3.33 A. *The Middle of the Finger hooks into the third eye and reflects down into the Organs.* **B.** *Pull open the Third Eye; let heaven open and shine its light into the Brain.*

Front Direction: Push/Pull Master Practice

Now, turn your palms, pushing out. Push. Pull. This is the master practice that is imperative for successful completion. When you first start practicing you should do it at least one hundred times and increase up to 200 times. Push and pull. When you push, you feel your hands extended very far away – very long – reach the sky. Touch the universe.

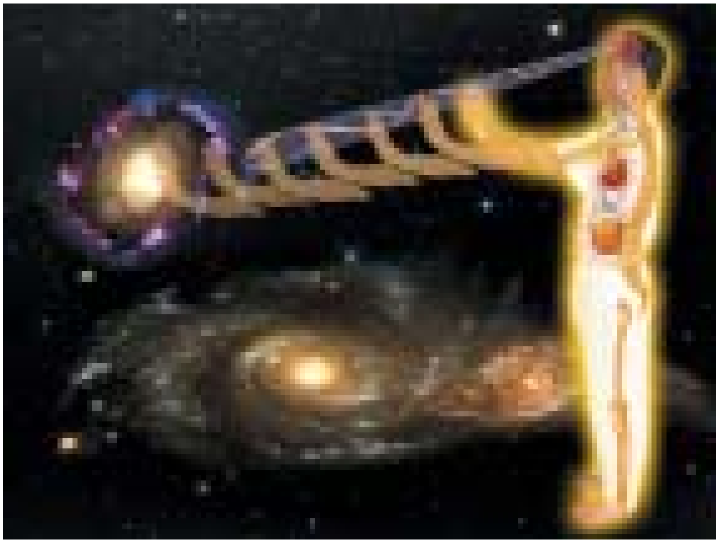

1. Push: Extend your arms and hands to the front, palms vertical. Extend the arms from the scapulas.
Expand: smiling, smiling, touching the universe – touching the force, touching the Cosmic Chi.

Fig. 3.34
Master practice:
Touch the Universe.

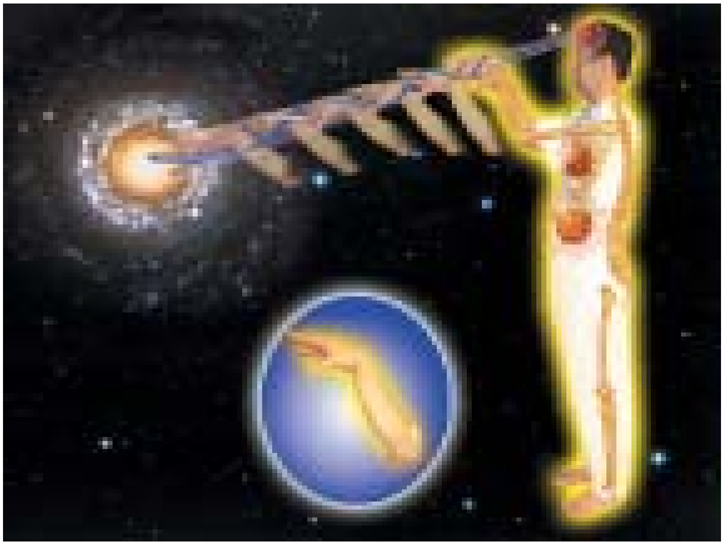

2. Pull: Draw the Chi back to you from the universe. Moving the arms from the scapulas, draw the hands toward your body in a horizontal position, drawing the Chi into the body.

***Fig. 3.35** 'Drawing' Universal Chi - feel your Tan Tien and fill it with Chi.*

Pull. Think and smile to your Tan Tien. Push, very far away to the universe. Pull. Push: Smile, relax, and let go, touching the sky, touching the universe.

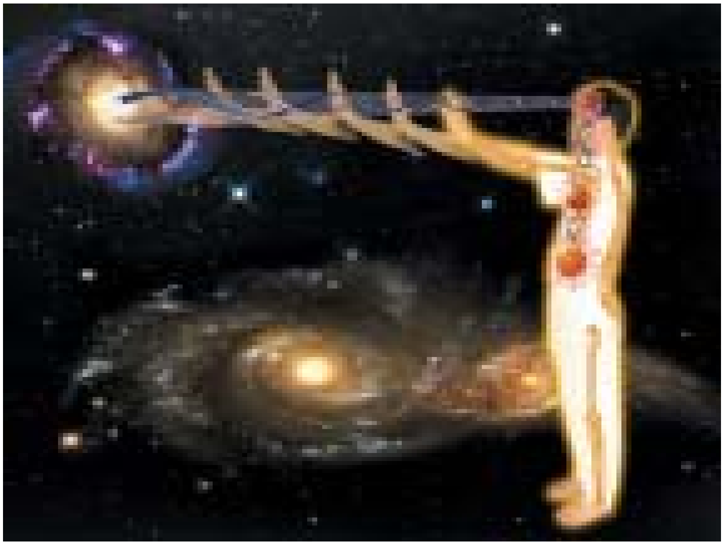

Fig. 3.36
Let go - Push and Touch the Universe, do it 6, 9 or 18 times.

Left and Right Directions

Now, move your hands to the left and right directions. Pull the Universal energy in. Push; touch the universe. Pull; think about your Tan Tien. Push – all the way, touching-touching-touching the universe.

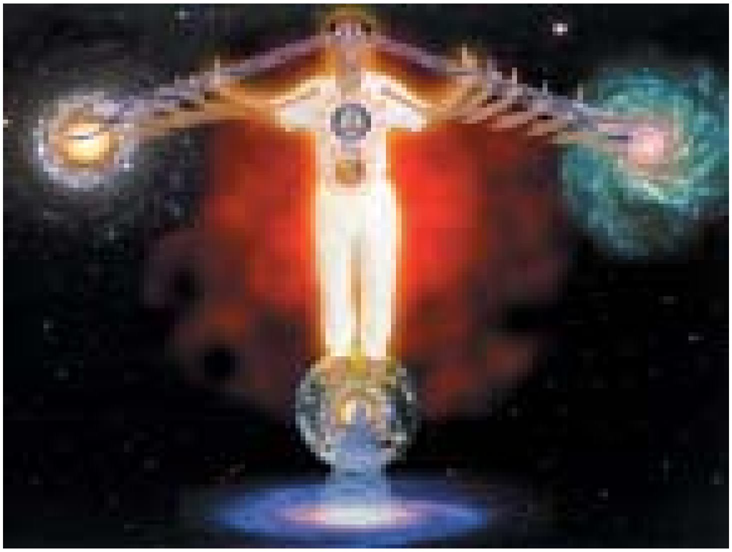

1. Move your extended hands from the front horizon to the left and right sides.

Fig. 3.37 *Touch the Universe Left and Right.*

2. Pull Tan Tien: Smile to your Tan Tien. Keep smiling to your Tan Tien.
3. Push to both sides. Expand all the way, smiling and touching the universe.

   Push/Pull: touching, touching the universe drawing the Chi smile energy into you from both sides.

   Do 3-6 times.

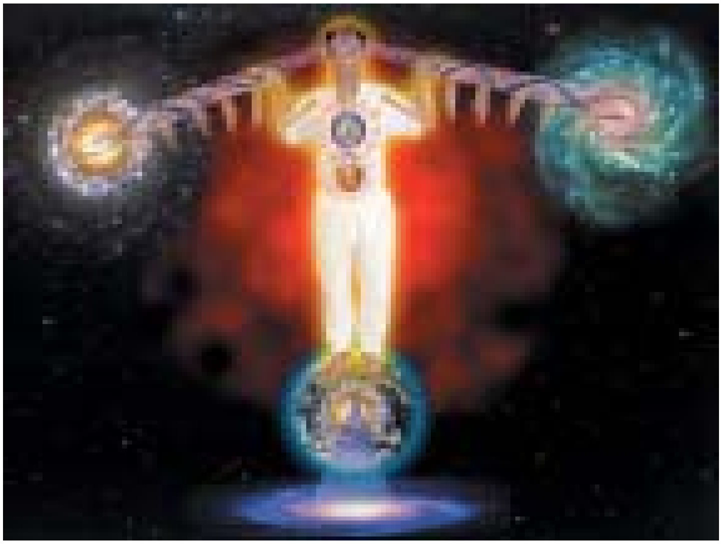

Fig. 3.38 *Pull: Just feel about your Tan Tien.*

Direction Above

Turn your palms up to the universe. Scoop up the Chi. Pour the Chi over your crown, and touch your crown. Project the Chi all the way down to the perineum and down through the earth to the universe below. Tan Tien and universe; always feel your Tan Tien spiraling, heart spiraling, crown spiraling and the universe around you spiraling.

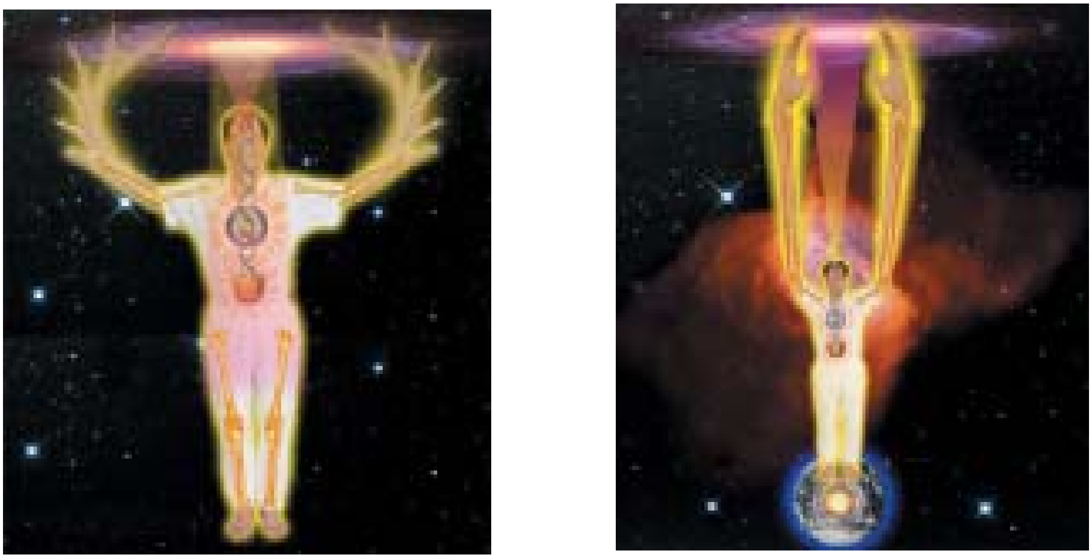

Fig. 3.39 A. *Raise the hands above the crown and feel it extend up to heaven.*
B. *Feel that the hands are big and long and that the bones are hollow. Fill and pack the bones with Chi.*

Raise your hands up to the universe and expand your hands as big as the universe - feel the Chi charge into your bones.

Scoop and gather the energy turning the body three times to the left and three times the right drawing the smiling energy.

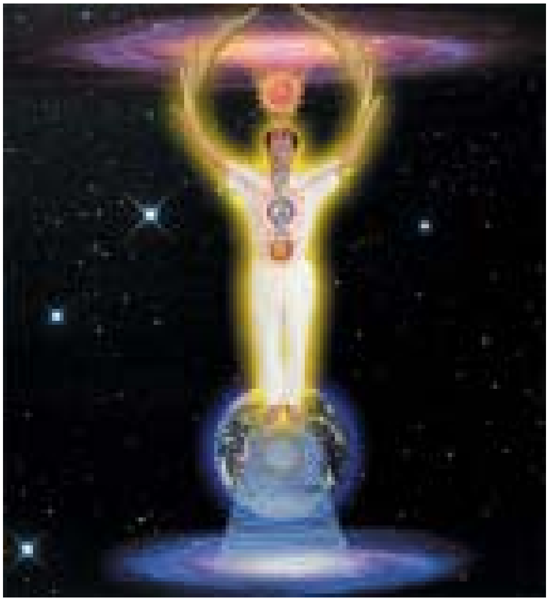

Fig.3.40 *Scoop up the Universal Chi and pour it over the head.*

Open the Spine

1. Touch the back-crown point. Pour the Chi all over your crown. Think of your soles so that you feel like there is a waterfall of Chi flowing from your crown all the way down to your soles. Feel your finger grow long and the Chi penetrate down through your spine to the coccyx. Leave the fingers touchng the back of the crown, to maintain the energetic connection with the coccyx.

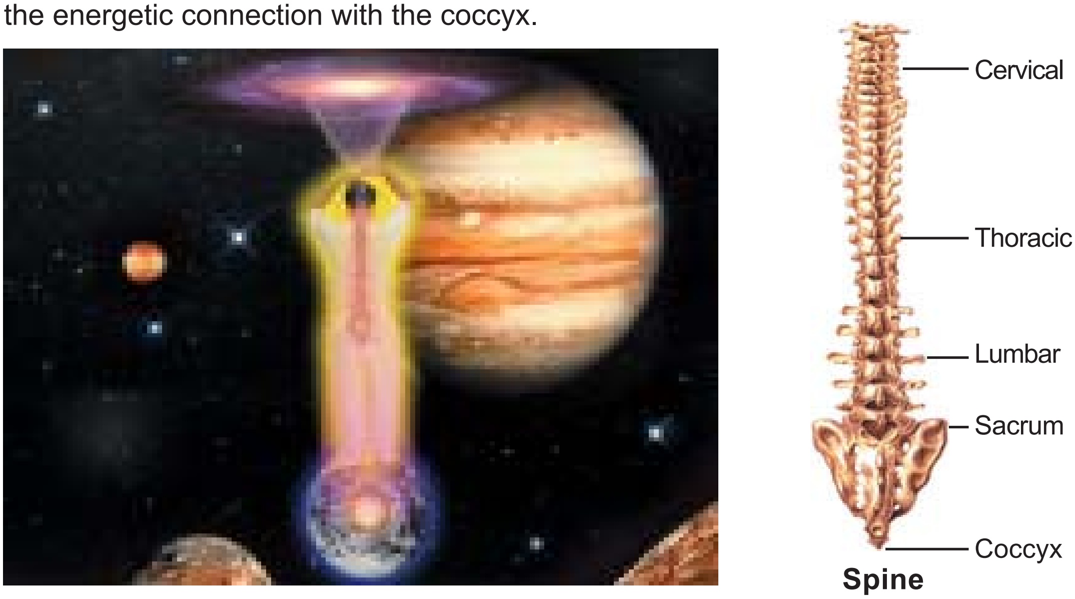

Fig. 3.41 *Feel your fingers growing long, all the way down to the coccyx.*

Turn ‘Three Minds into One Mind’ at the Lower Tan Tien, and expand the awareness to the universe. Let yourself be charged by the universe.

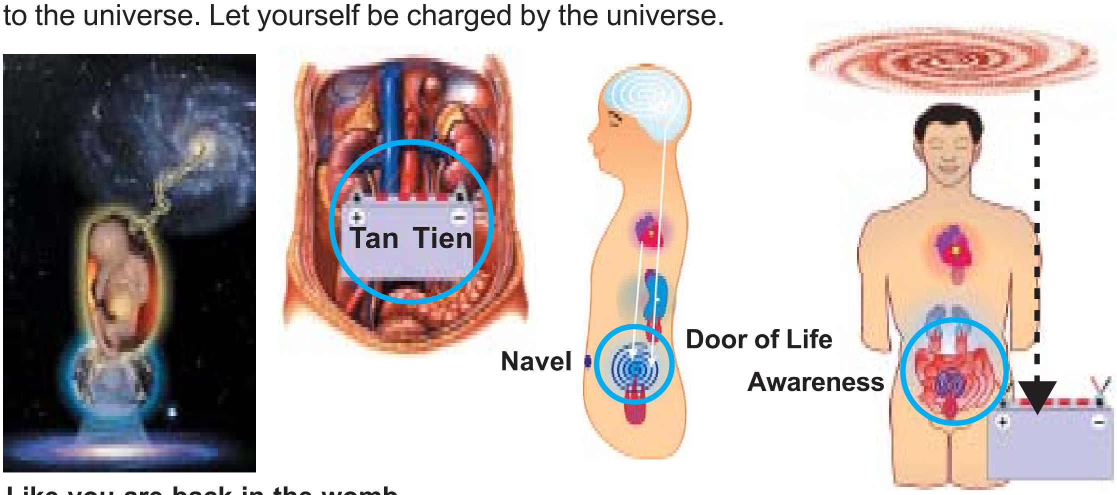

Like you are back in the womb.

Tan Tien is like a battery.

Fig. 3.42 *Universe charges Tan Tien power.*

Beware of the Tan Tien and spiral it like universal energy in motion. Feel the heart center spiraling and the crown spiraling. Be aware of the universe spiraling above, below, front, back, left and right. Let all of the sick energy and the negative forces leave the body and go down into the ground for Mother Earth to recycle. Extend the Chi from above, all the way down through the earth and the universe below.

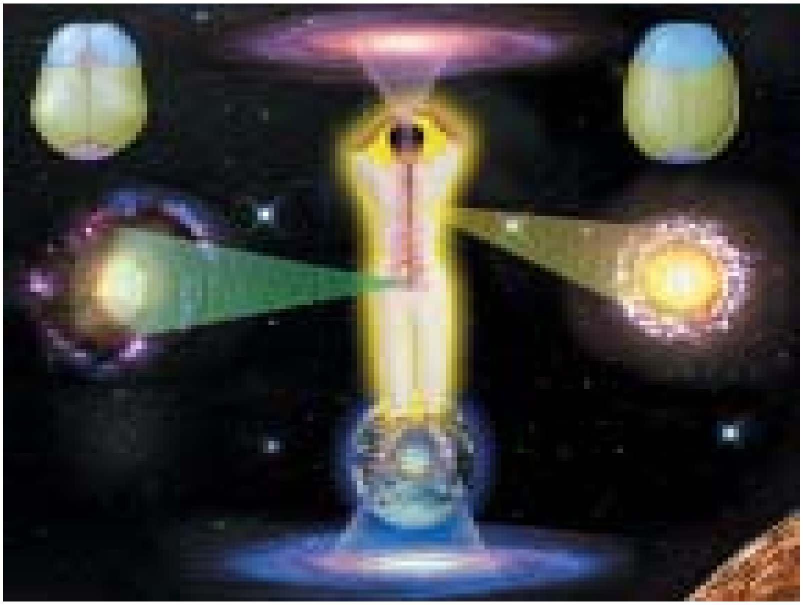

Fig. 3.43 *Be aware of the Lower Tan Tien, Heart Center and Crown spiraling. Feel the Universe is spiraling and charging the Three Tan Tiens.*

Open the Middle Channel and Perineum

1. Move to the mid-crown point. Touch the point and project your fingers inwards; go deeper, deeper through the middle of your body down to the perineum. Focus on the perineum. Feel the Chi from the universe flow into your perineum. Look for one dot of light. Look into the darkness, the vast darkness, the immense darkness; this is the primordial force, a cloudy moving force. Look for a dot of light at the perineum and extend your awareness all the way down through the ground and the universe below.

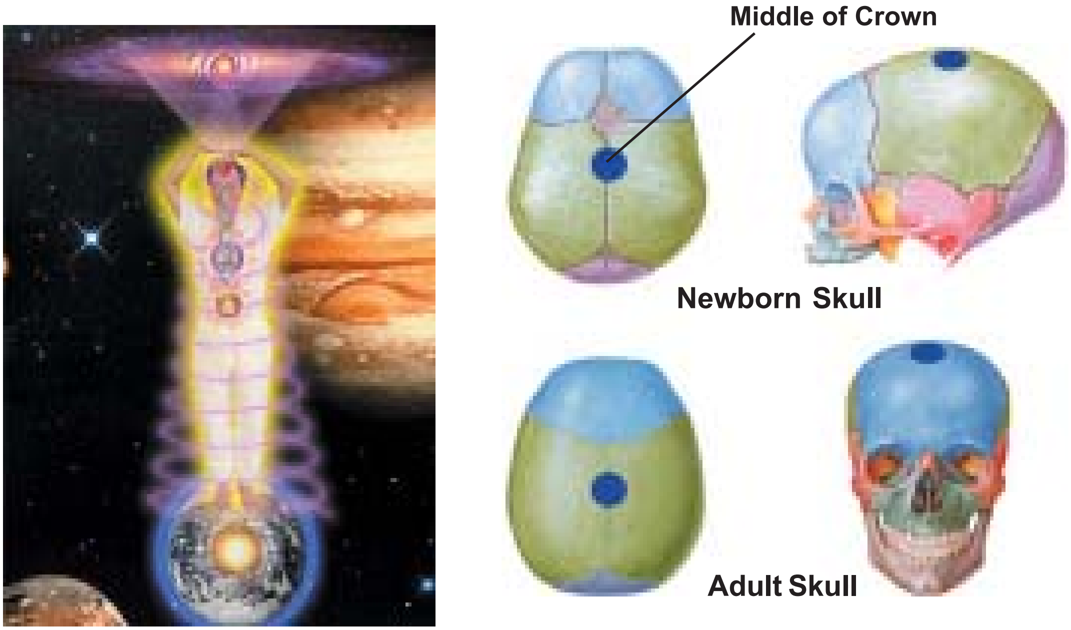

Fig. 3.44 *Move the Hands to the Middle of the Crown and feel the Fingers Penetrate to the Perineum.*

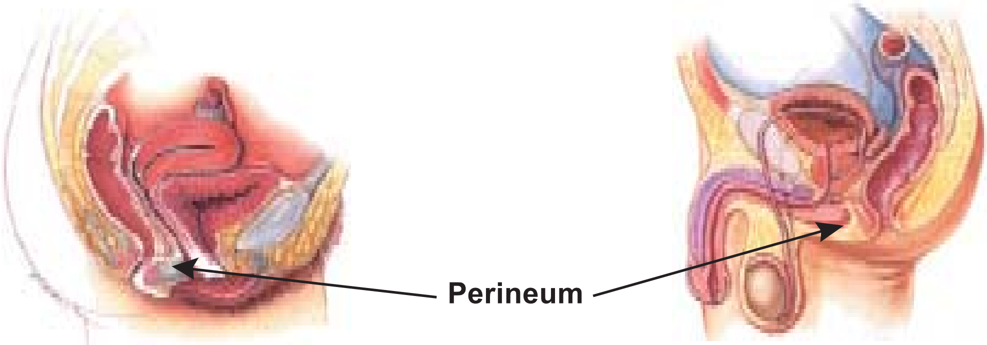

When finished opening to the Six Directions and igniting the Three Fires, gather the Chi in the center and bring this expanded awareness into the healing session.

Tan Tien and the Universe

The expression, 'Tan Tien and the Universe', is a reminder to feel your Tan Tien, heart, mideyebrow and crown spiraling and that you are connected to the universe spiraling in the six directions around you.

1. Establish a Complete Location

You will use your Yi, the three-mind power, repeatedly to recharge your Chi for various purposes. You recharge by connecting to the Universal Chi in the six directions of the universe simultaneously. When you charge a particular area or direct Chi into the body to a particular point, you first establish a connection point for the Chi by placing the hand or fingers at an appropriate location on the surface. This is like giving a location for the Chi to go to. Once the location is established and the Chi starts to go there, leave your hands there.

Then you move your attention to where you want the Chi to go in the body. Feel the Chi connected to and charging the intended location.

2. Charge Tan Tien from Universe

When the location connection is established, be aware of your Tan Tien, heart center, mideyebrow and crown spiraling. Be sure that the conscious mind of the heart is lowered to the Tan Tien and the awareness mind of the abdomen is connected to the mideyebrow crown and out to the universe. Feel them connected to the spiraling energy in the six directions of the universe. Let the universe charge your Tan Tien.

3. Don't Stop at the Location

With your focus in the Tan Tien, the Chi will go from there to the 'location' indicated by your hands and then to the intended location in the body. Don't let Chi stop there, however.

4. Direct Chi out the Opposite Side to Universe

Direct your attention to guide the Chi flow through the body and out the opposite side. You want to clear out any blockages and prevent any others from accumulating. You also want to release sick energy and negative forces down into the earth. Let the Universal Chi flow out through the universe and beyond.

Open the Three Tan Tiens

Move your hands down to the mideyebrow. Touch the mideyebrow.
We are going to open the three Tan Tiens, starting with the Upper Tan Tien.

Upper Tan Tien - Mideyebrow

1. Recharge. Remember: Tan Tien – heart consciousness in, awareness out. Spiral – Tan Tien, heart, third eye and crown. Universe – six directions spiraling.

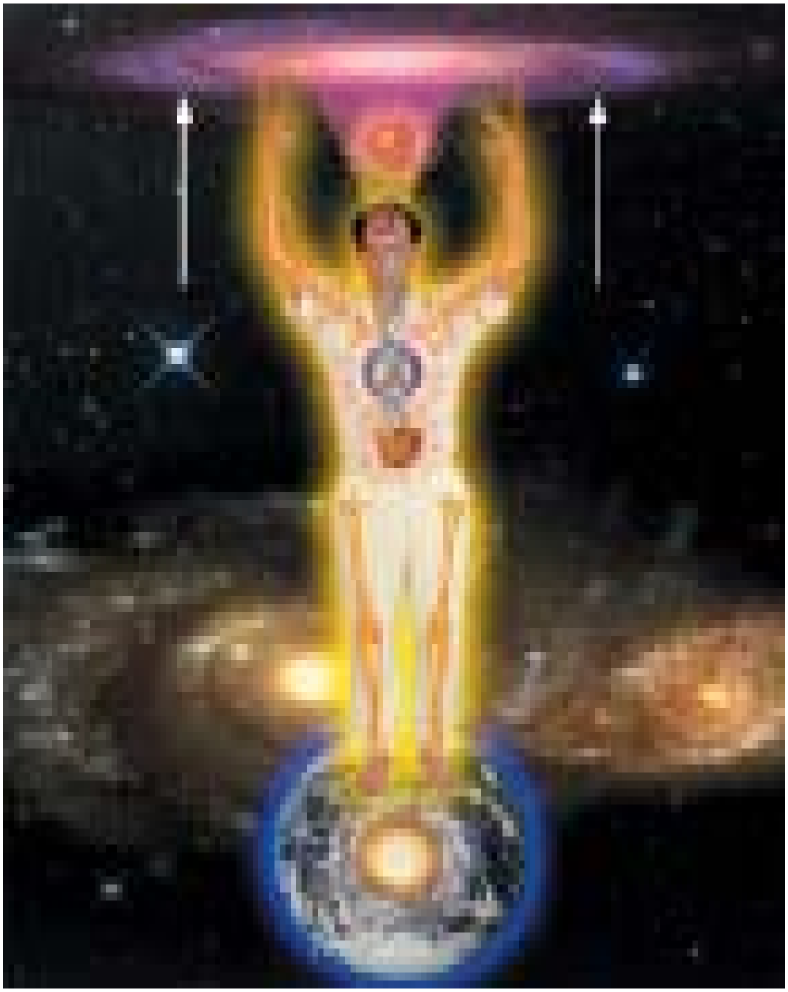

Fig. 3.45 *Raise hands and charge with Universal Chi.*

2. Move your hands down and touch the mideyebrow. Feel your fingers grow very long (energetically) and penetrate all the way back to the base of the skull. Focus on the back. Remember: Tan Tien and the universe spiraling. With the spiraling, the Chi in the fingers will become hot. It will expand and penetrate out through the back of the head all the way to the universe behind.

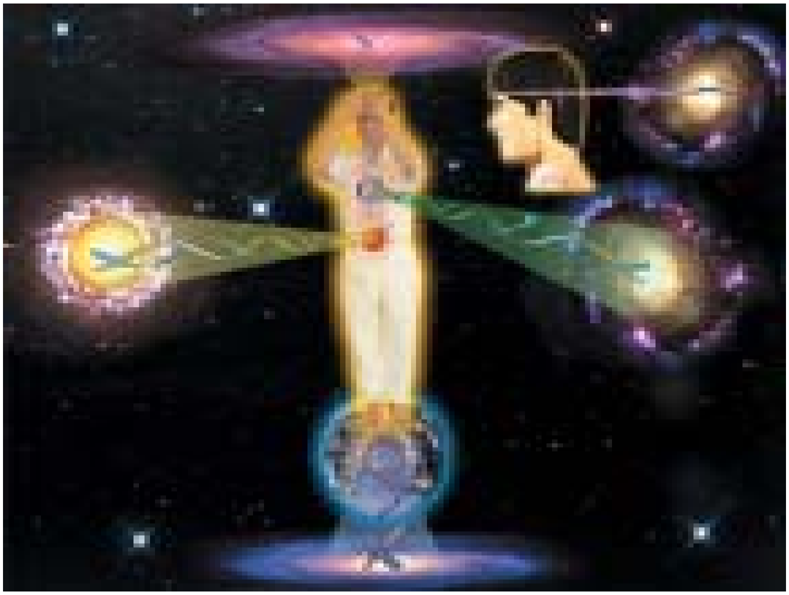

Fig 3.46 *Fingers touch mideyebrow and penetrate to the back of the skull and the universe at the rear.*

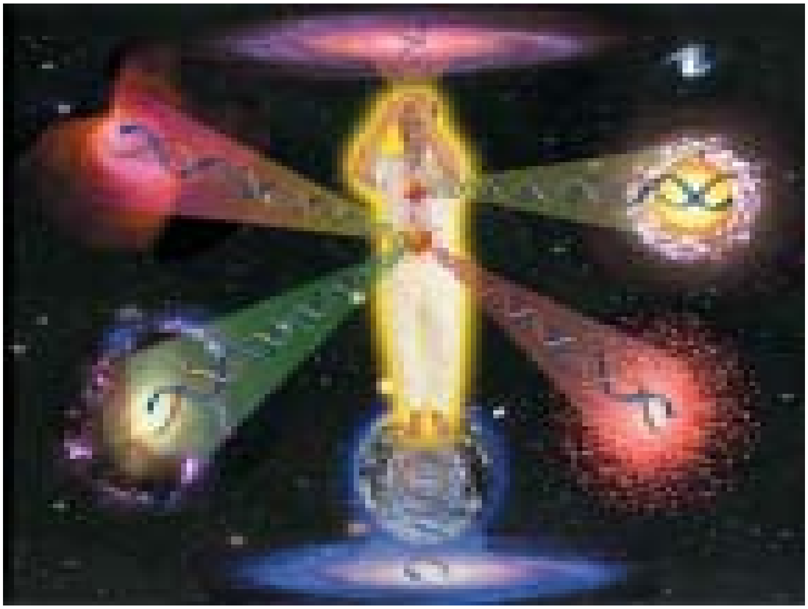

Fig 3.47
Tan Tien and the Universe.

3. Picture your fingers like laser beams of Chi, 'Tan Tien and Universe': Feel your Tan Tien and the universe spiral and charge your fingers. Move your fingers out from the mideyebrow around the side of the head to the top of the ears. Your fingers are like lasers -cut-cut-cutting open your skull right in the middle, around to the top of the ear, cutting open your Upper Tan Tien. Cut and project your fingers long into the middle of your brain. Spiral your Tan Tien and the universe. Leave your fingers there. Concentrate on your Tan Tien spiraling – your heart, crown, the universe above, below, front, back left and right all spiraling. Your Tan Tien is a big empty space: primordial force; darkness. You can put so much Chi inside there! The Chi penetrates into your brain.

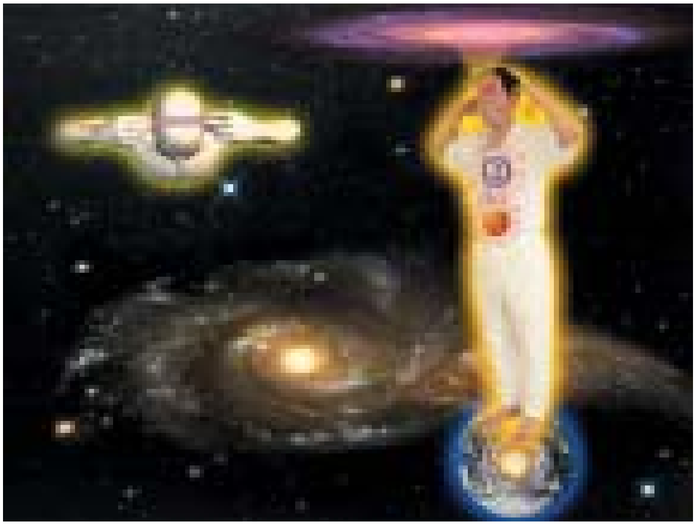

Fig. 3.48 *Move the fingers to the top of the ears. Feel them growing longer and cutting the skull open, penetrating into the brain.*

4. Move your hands all the way to the back, cutting to the back of the skull. Touch, and feel the Upper Tan Tien open.
5. Touch the base of your skull. Focus on the mid-eyebrow. Feel the Chi flow like a laser beam from back to front and out to the universe in front. Complete the opening process by moving the hands back around to the mideyebrow, cutting as you go.

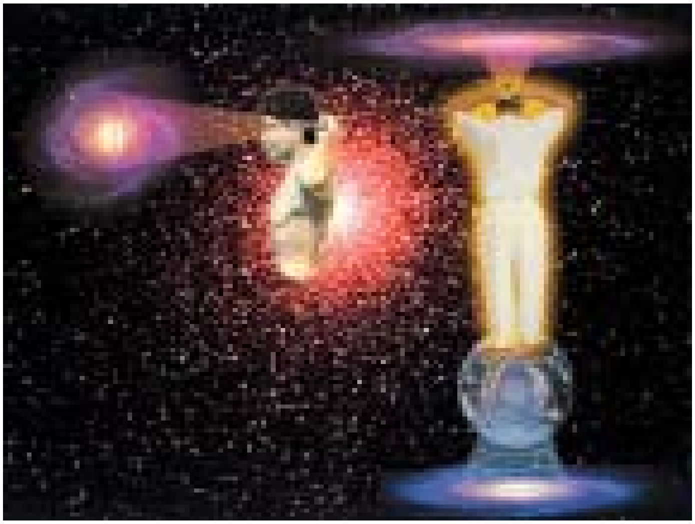

Fig. 3.49 *Chi moves like a laser, penetrating from the mideyebrow at the front, to the back of skull.*

6. Recharge in the universe. Feel your bones and your arms are hollow. Fill and compact them with Chi.
7. Scoop the Universal Chi and pour it down over your crown and all the way down, down to the Middle Tan Tien.

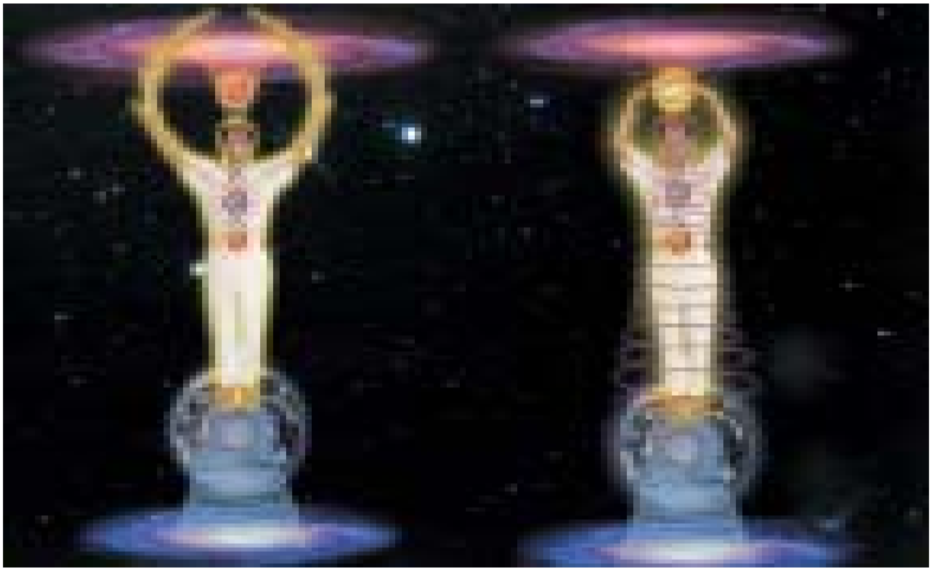

Fig. 3.50

Recharge in the Universe, scoop and pour the Chi down to the Middle Tan Tien (Heart Center).

Middle Tan Tien - Heart Center - Conscious Mind

1. Move your hands all the way down to your heart center at the mid-sternum. Touch. Focus on the point opposite the heart, T5/T6, on the spine. Fingers 'long', Chi penetrates into your thymus gland. Light – golden light – penetrates into your thymus. Feel the Chi penetrate through your heart all the way through T5/T6 to the universe behind. 'Tan Tien and the Universe' also feel your Chi Fingers penetrate into the bone and bone marrow and spread out into your rib cage.

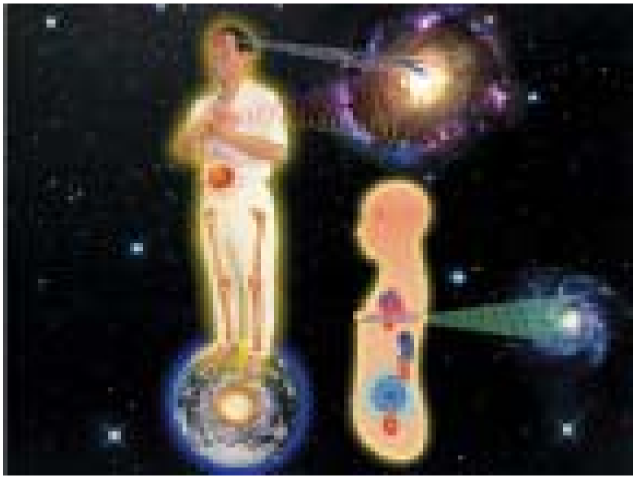

Fig 3.51 *Golden light enters the heart, thymus, bones and marrow; feel it penetrate to the universe at the rear.*

Recharge and cut around to the Armpits

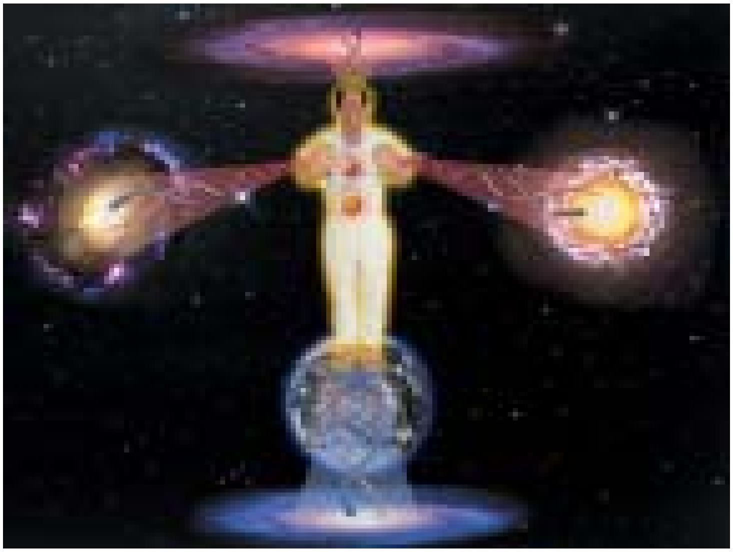

1. Recharge In the universe, pour Chi over the crown down through the body and lower your hands down to the heart center. Touch with the fingertips. Move your hands around under your armpits, extending the Chi like laser beams cutting open this Middle Tan Tien. Pause under the armpits as you send the Chi into the center.

Fig. 3.52 *Cut open the Middle Tan Tien by cutting around the armpits.*

2. Continue to move your hands around to your back at T5/T6, touch and send the Chi from back to front. Let the beam of Chi penetrate out through the heart center to the universe in front of you. Then, move the hands back around the sides to the front, cutting as you go.

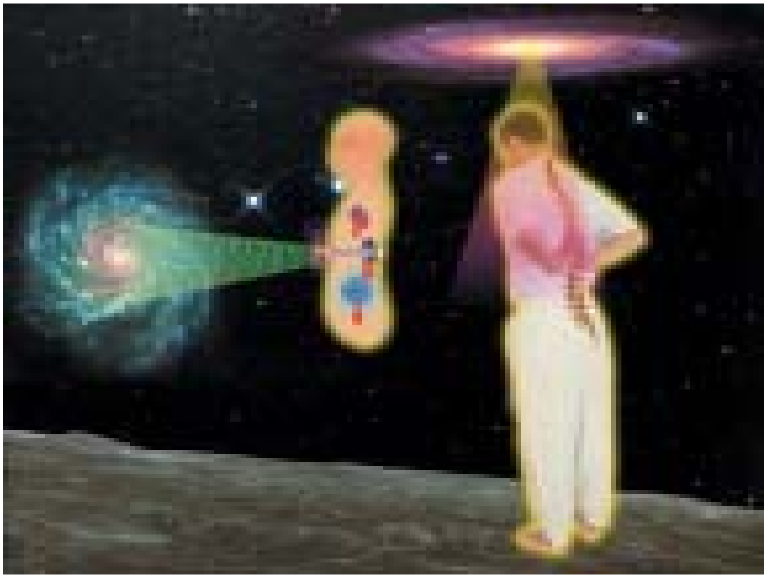

Fig. 3.53 *Chi beam from T5/T6 penetrates to the Heart Center in front.*

Lower Tan Tien - Navel - Awareness Mind

1. Recharge with Universal Chi. Tan Tien and Universe: Again, raise your hands and charge with the Chi into the universe. Your hands are very big, very long. The bones are hollow and compacted with compressed Universal Chi. Scoop the Chi from above and guide it down. Pour all the way down, down, down to the navel. Touch the navel, and focus on the Door of Life opposite on the spine between L2/L3. Touch and feel the Chi penetrate to the Door of Life Tan Tien and Universe. Feel the Chi penetrate through to the back and out to the universe behind.

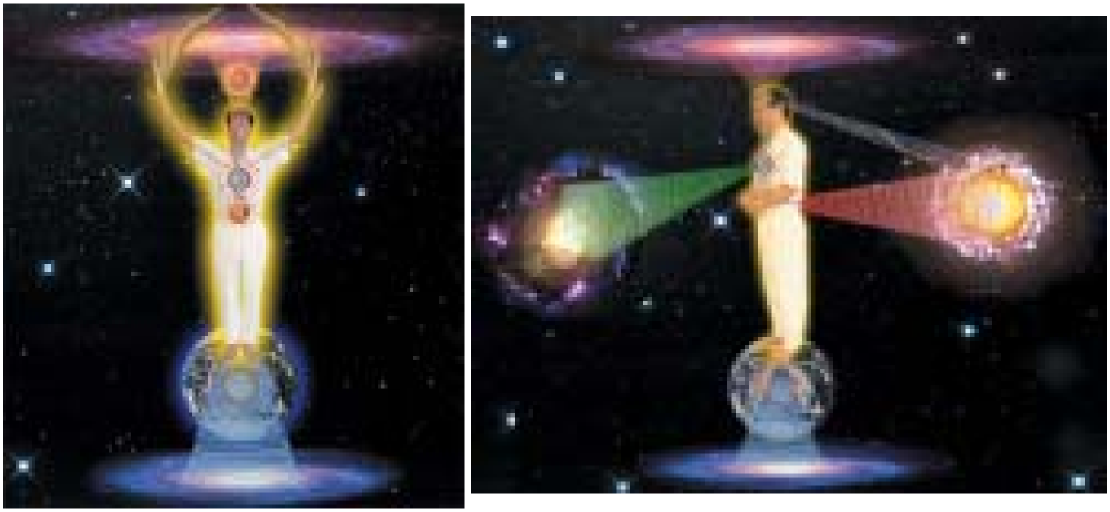

A. ***B.***

Fig. 3.54 A. *Recharge the Lower Tan Tien and penetrate through to the Door of Life.* ***B.*** *Raise the hands up to the universe and charge with Chi. Bring the power down to Tan Tien.*

2. Open this Lower Tan Tien the same way as for the Upper and Middle Tan Tiens. Charge more Chi into your hands, and let them be like lasers cutting it open. Cut around to the sides. Pause. The fingers of the left and right hands are very long, extended energetically inside. Cut and feel the energy penetrate into the center. Focus at Tan Tien and universe and feel more Chi.

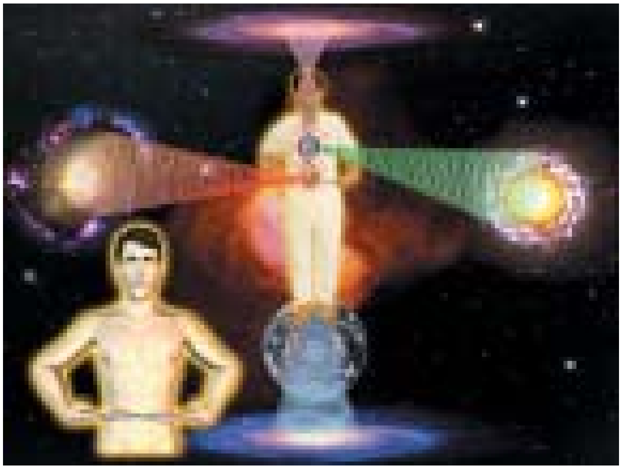

Fig. 3.55 *'Cut' open the Lower Tan Tien with Laser Fingers.*

3. Continue cutting to the Door of Life. Touch, and send the Chi from the Door of Life back to the navel and out to the universe in front.

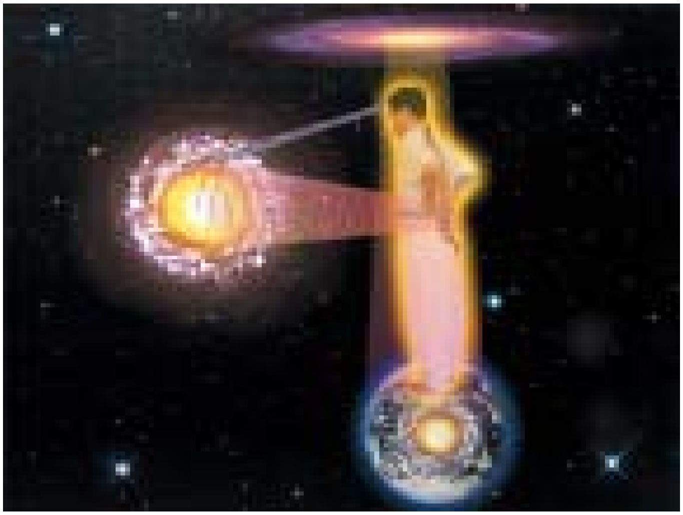

Fig. 3.56 *'Cut' open the Door of Life and send Chi to the navel and universe in front.*

4. Move the hands back around the sides to the navel, extending the fingers and "cutting" the Tan Tien open as you go. Touch the navel: Tan Tien and the universe spiraling. Feel more Chi, and feel the Tan Tien open.

Activate Chi in the Bones of Hips, Legs and Sacrum

1. Now touch your pelvic bone by energetically extending your fingers from the front area near the hips to the back. Feel Chi penetrate into your pelvic bones: funny, laughing, happy bones.
2. Touch the femur bones. Tan Tien and universe. Charge the fingers. Also, feel the funny, happy, laughing vibration inside the bones and in the bone marrow. Be aware of the Three Tan Tiens. Spiral from the universe. Spiraling charging your Tan Tien. Charge your hands and your bones.

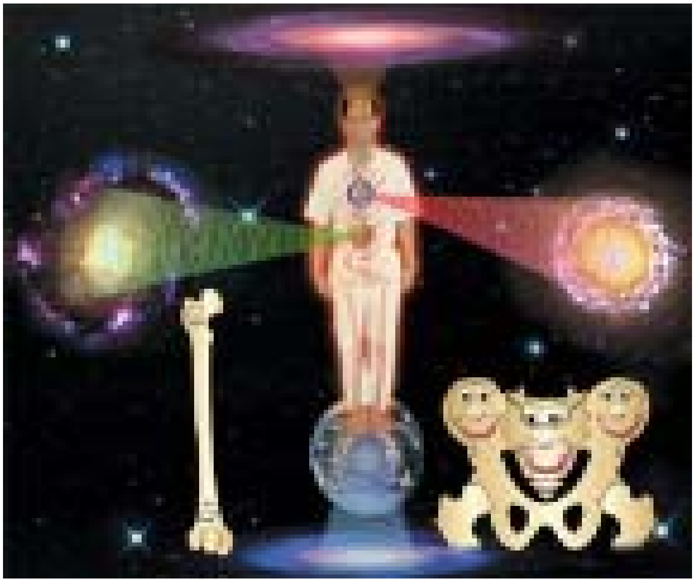

Fig. 3.57 *Touch the femurs - happy laughing bones!*

3. Now we are going to slowly lower the sensation of Chi down through the bones to the earth. Opening up the sacrum. Move the hands down the legs as you bend down.
4. Lower yourself all the way down to the ground and sit on your feet. Move the Chi with your hands down to your toes, down through the earth and the universe below.

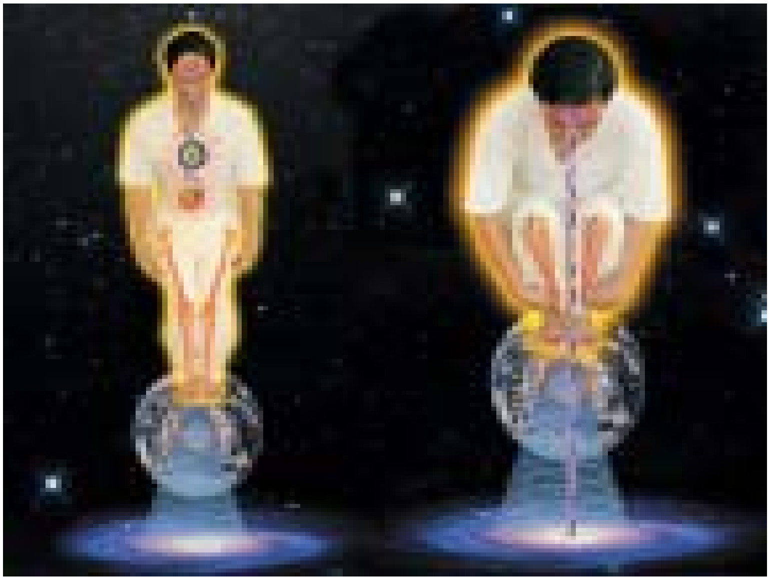

A. ***B.***

Fig. 3.58 A. *Sink the Chi into the earth.* ***B.*** *Lower and sink the mind and Chi down to the universe.*

5. Raise your sacrum up, keeping your hands at the toes. Smile to your Tan Tien and feel the Chi from the universe coming to fill the Three Tan Tiens. Opening up the sacrum.

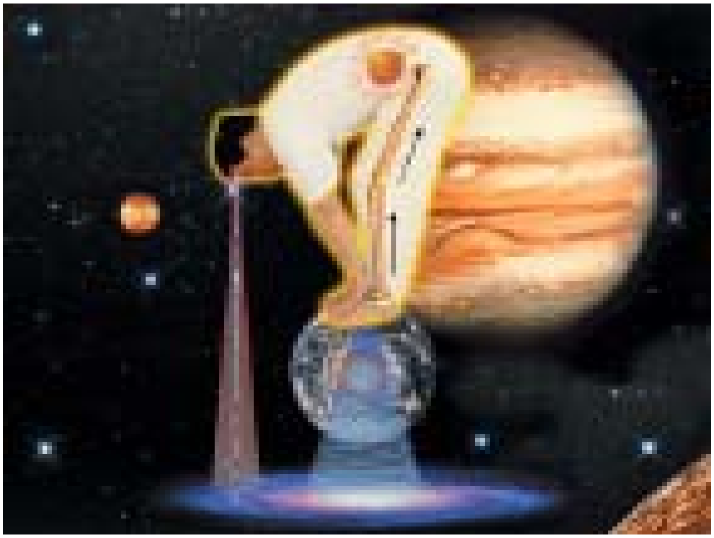

Fig. 3.59
Raise the sacrum and smile to the Tan Tien.

6. Lower down again. Lower the Chi down to the earth and the galaxy below.
7. Once more, raise your sacrum up, maintaining hand contact with your feet. Smile to your Tan Tien.
8. For the third time, lower down. Open your palms, gathering the Chi from the earth below. Gather and scoop the Chi.

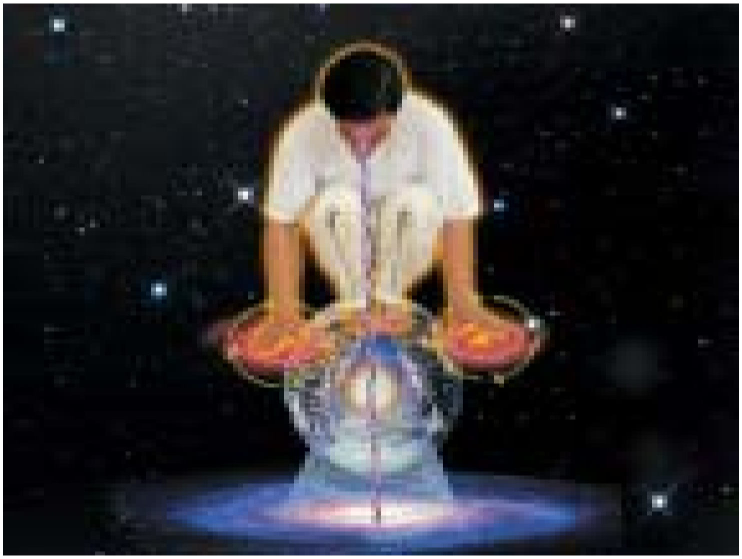

Fig. 3.60 *Gather the Earth Chi.*

9. Touch your heels and feel your bones as you slowly rise up. Fill your bones with Chi as you guide it up with your hands.
10. Fill the bones in the upper legs as you move your hands up.

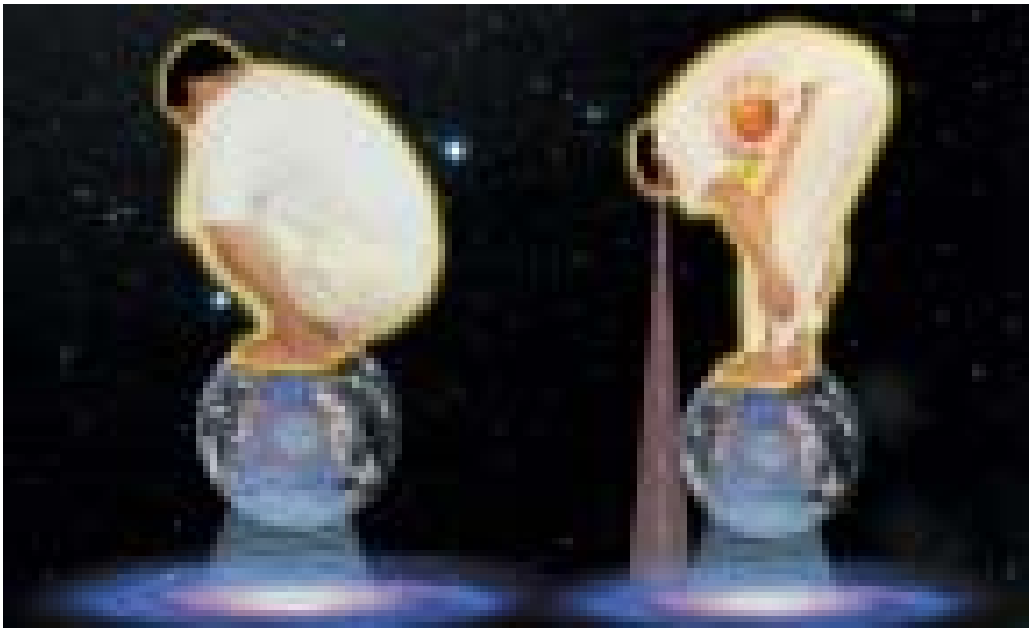

Fig. 3.61 A. *Fill the bones with Chi.* ***B.*** *Feel your fingers penetrate into the bones as you rise.*

11. Feel your bones, and fill them with Chi all the way up to your coccyx. Touch your coccyx. Leave your fingers there, then be aware of the Chi and feel the Chi raising up to the Tan Tien and the universe. Feel it charge the fingers and the spine.
12. Come up to the sacrum. Feel the sacrum open. Focus at the Tan Tien and the universe.

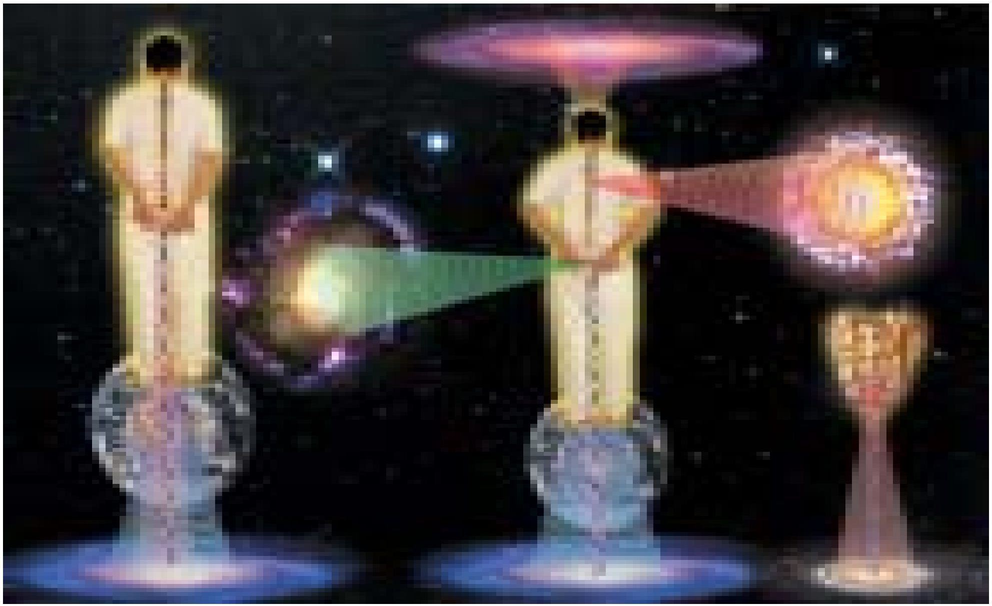

Fig. 3.62 A. *Bring the Chi to the Coccyx.*
B. *Move your hands up to the Sacrum and feel the Chi rise up to the Crown.*

13. Come up to the door of life, and then back to the navel. You may sit down to continue the next step in the Cosmic Healing Practice.

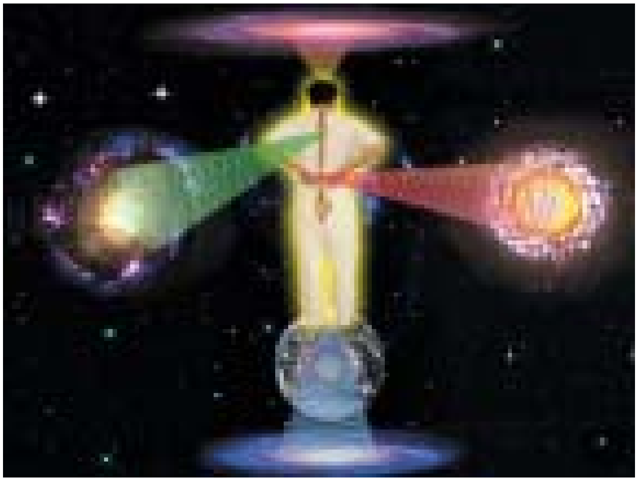

Fig. 3.63 *Bring the Chi to the Door of Life and let it penetrate to the Navel.*

Practice daily till you feel the Chi. Then you can apply this Chi to the Healing Sessions.

Chapter IV
General Healing Session

The General Healing session practices are to be done with every student regardless of their age or health condition. These are for cleaning and strengthening the cells. Cells must be clean and strong to heal. The General Healing practices listed steps A-K are the most important part you must always do in every session. You can do this with one person or with many people at the same time. The following is a detailed decription of each of these steps. Become familar with them and they will flow easily one into the other.

A. Three Minds into One Mind

Smile down to the inner universe. Activate the heart consciousness and empty the mind and heart down to the LowerTan Tien, the Abdominal Brain. Gather the 'Yi' (mind-eye-heart power), combining three minds into one. Fill the Tan Tien with this Chi. Start to spiral. You are then ready to connect to the higher forces of Universal and Heavenly Chi. The Three Minds are the Upper, Middle and Lower, or three Tan Tiens. The Upper Mind is the one we have been closely related to most of our lives. It is the logical thinking, planning, calculating and worrying mind; it consumes 80% of our body's energy.
The key is to learn 'to releas the mind' by learning to sink the Chi to the Lower Tan Tien, whereby it can be used in synergy with the other two.

B. Activate the Six Directions

Activating the six directions is for opening the whole body, mind and soul to all of the Universal and Cosmic forces in all directions. We literally learn to draw in the Universal and Cosmic forces from all directions at the same time. This is important for healers so they are not using their own energy for healing but drawing on the limitless Universal forces.

Expand your awareness to connect to the universe and the Six Directions at the mideyebrow and crown; left, right, front, behind, above and below. Connect to the universe, collect and spiral the energy into your Lower Tan Tien.

C. Activate the Three Fires

Activate the Tan Tien, Kidney and Heart Fires to create a Sacred Fire Triangle. Move from the Tan Tien to the Heart, the Heart to the Kidneys and the Kidneys back to the Tan Tien.

D. Linking Personal Stars, Energy Bodies and the Universe

World Link is a way for intelligent life force to return to the source; inwards and upwards. Become aware of the ‘personal stars’ above the crown and below the earth.We link with our own personal stars and then with the stars of each person in a group. This links the energy bodies of everyone in a group together. Start to spiral your energy and connect to other peoples’ stars and energy bodies. Make a left spiraling pattern, progressing around the circle in a clockwise rotation. Link all of these stars together and form ***one large energy body.*** Each meditator will become an individual unit in an integrated communication link. A link between the earth and the universe.

E. Protective Circle and Chi Field

We create Chi Fields, Protective Animals and call the Eight Forces to protect the group from disturbances and psychic attacks. Raise your right (or leading) hand in the air and access the **‘Sacred Fire’** of the universal cauldron. “Feel your fingers becoming big, long and hollow”, as they reach into the Sacred Fire. The Chi will fill and pack your arm as you bring it down into your body. Use Yi-Power to create the **Sacred Fire Protective Circle.** The fire will purify and protect the circle. You can cast a circle around your community, house, room and body. This will create an energetic field of Chi around the whole room.

Then Create a **Chi Dome**. Invoke the ‘guardian animals’; Blue Tortoise from the north, Red Pheasant from the south, Green Dragon from the east, White Tiger from the west, Yellow Phoenix from above and Black Tortoise from below.

Finally, activate the **Eight Elemental Forces of Nature and the Universe**. Call the forces of fire, water (ocean), thunder (lightning), lake (rain), earth, mountain, wind and heaven. Activate fire in the north, water in the south, lightning in the east, and rain in the west.

F. Chi (Sacred) Water Practice

Invoke the power of the Sacred, Chi or Holy Water practice to cleanse and heal the body’s sick, toxic or negative energy. Make a request to receive healing energy and feel yourself touch a heavenly pool of sacred water energy. Feel the pool pouring down heavenly water to fill your arm.

Use the thumb and index finger to remove the sick energy from the cup, without touching the water. Talk to the sick cells. Let the sick energy return to the earth for recycling. Fill the water with your heart’s compassion. Give the command: “This Sacred Water will give health, wealth and longevity”. Cosmic Healing works on a cellular level. It is the water’s ‘job’ to carry the message of the practice to all the cells, where it will remain. Drink the water in unison with the rest of the group.

G. Empty and Fill

The Empty and Fill technique is emptying the sick energy into the earth and filling with healing Cosmic Force Energy. Project your ‘sword’ hand (middle and index fingers) down to the ground. Project the hand ‘very big and very long’ and your feet ‘very big and very long’ - extended down to the ground. Now very slowly inhale, bring up the earth energy. Bring the energy up to your feet and into your bones. Let it enter all the bones, organs and cells. Feel it blending with all the sick energy. Turn your palm up, turn it back down again and exhale; let go of all the sick energy, the negative energy, the worries and burdens, all the way down into the earth. Repeat this three to nine times. Dig a hole and bury all the sick energy; tell it, “Do not come back; you will be happy down there.” Smile.

H. Clean with Green Light

The Green light is for cleaning the cells. This is vitally important because the cells must be clean for healing. Green can help detoxify the toxic cells. Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. It ‘loosens’ the area. The light green Chi is used to break down dirty or diseased energy, like a detergent is used for washing clothes or dishes.

Afterwards, the cells will be rinsed with blue light, like clean water is used to rinse the detergent and residue from clothes or dishes.

When you carry out healing try to ‘see’ the cells enlarged and this green energy go in and blend with the dirty, black and cloudy energy. See it come out and let it go down to the earth. Dig a hole and bury it there. Tell it, “Don’t come back. You will be happy down there”. Become aware of the green light and spiral counter clockwise. Spiral down. Cleaning out the stress. Repeat as many times as you like, three times minimum.

I. Clean with Blue Light

Blue is like cold water; it has the yin power to dissolve all kinds of negative energy and sickness. It has an inhibiting effect. Blue yin energy is the opposite of red yang, which has a strengthening and stimulating effect.

Blue has a cooling effect; it can reduce pain and inflammation, and it can help blood clot. It stops bleeding and reduces fever. It can help to induce rest and sleep.

J. Charge with Violet Light

Charging with Violet light holds the highest healing energy. When the cells are clean they will absorb and hold the Violet light and the healing will take place. Luminous Violet has intelligence and can be programmed. The universe is full of violet light, especially the North Star and the Big Dipper. The color of the star of the higher self or soul, above the crown, is also violet. It is the color of Divine or Soul energy.

It has properties of other Chi colors. It has a rapid regenerating effect on damaged organs and nerves. Always use green and blue light before using violet. Violet light develops the crown center, the spiritual core. It is good for psychological ailments as well as for physical deficiencies.

K. Activate the Immune and Defense System

The body has various mechanisms which combine to provide protection and defense against illness and disease. These mechanisms enable the body to produce various cells and bodies which act against invading or unwanted substances. As soon as a foreign body is recognized, the immune system will be triggered and will act to provide the most efficient means of eliminating the danger and return the body to a balanced state of health. A strong and healthy body will have good resources to protect itself from negative or sick energy. The aim of these practices is to help you to realize the potential of true harmony within yourself. To activate the immune and defense system is to increase the production of red and white blood cells. To do this we activate the bone marrow and lymphatic system. Guide the student or whole group together in all of these steps. You will not lose energy by doing this. You will gain more energy because you are all connected together in the group power. The crucial part is to multiply your good intentions, Chi and virtue energy into the cosmos, making yourself and each person a direct line from the Universe. After you finish the General Healing session you can tailor the healing techniques in the next chapters to the individual's needs.

Practice

A. Three Minds into One Mind

1. Smile to the Inner Universe

Place your palms together in salutation, in front of your heart. Feel the Laogong points in your hands connect, creating an energy loop running from your heart through your arms and hands and back again.

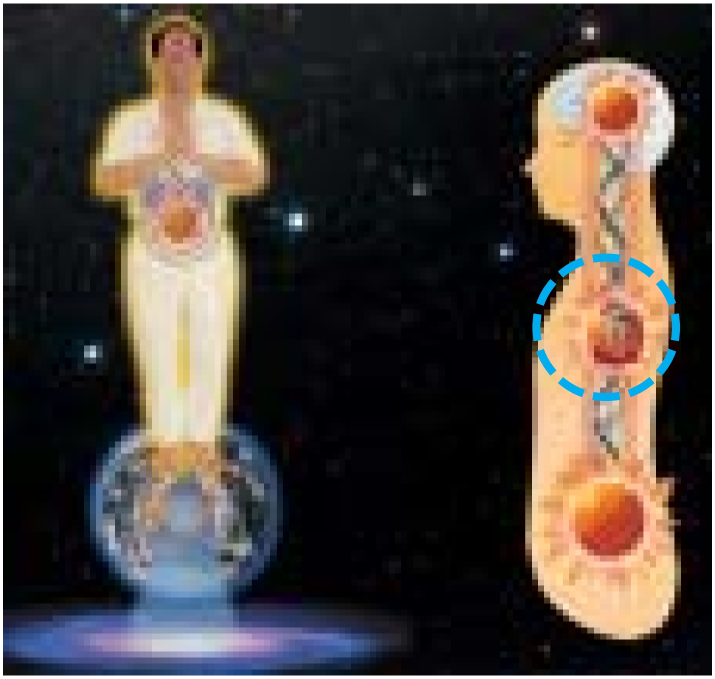

A. **B.**

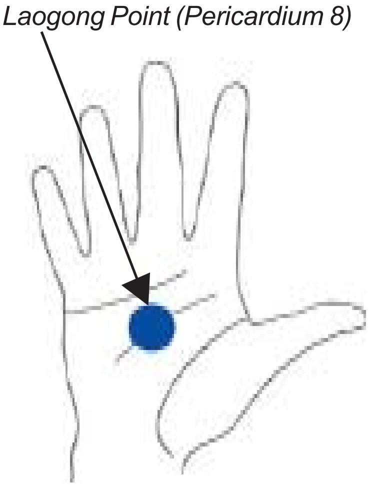

***Fig. 4.1 A.** Empty the mind down to the lower abdomen.*
***B.** Activate the heart's compassion.*

2. Activate the Heart Compassion Energy

Smile to the heart and feel it softening. Feel love, joy, compassion and happiness. Smile down and empty the mind to the Tan Tien, the Abdominal Brain. Fill the Tan Tien with Chi and start to spiral. When the abdomen is warm, it is full of Chi. The Chi can then charge up to the brain.

3. Combine the Three Minds into One

1. Turn Upper Mind to observation mind.
2. Turn consciousness (Middle Mind) down to Lower Tan Tien.
3. Combine three minds into one mind at the
   Lower Tan Tien (the Yi).
4. Manifest out at the mideyebrow.

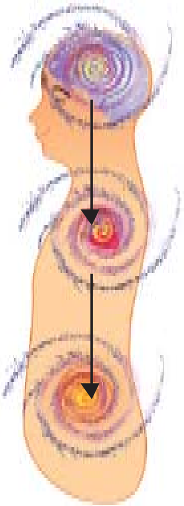

Fig. 4.2 *'Three Minds into One'.*

B. Activate the Six Directions

Expand the awareness to connect to the universe and the Six Directions at the crown, mideyebrow, heart and Tan Tien.

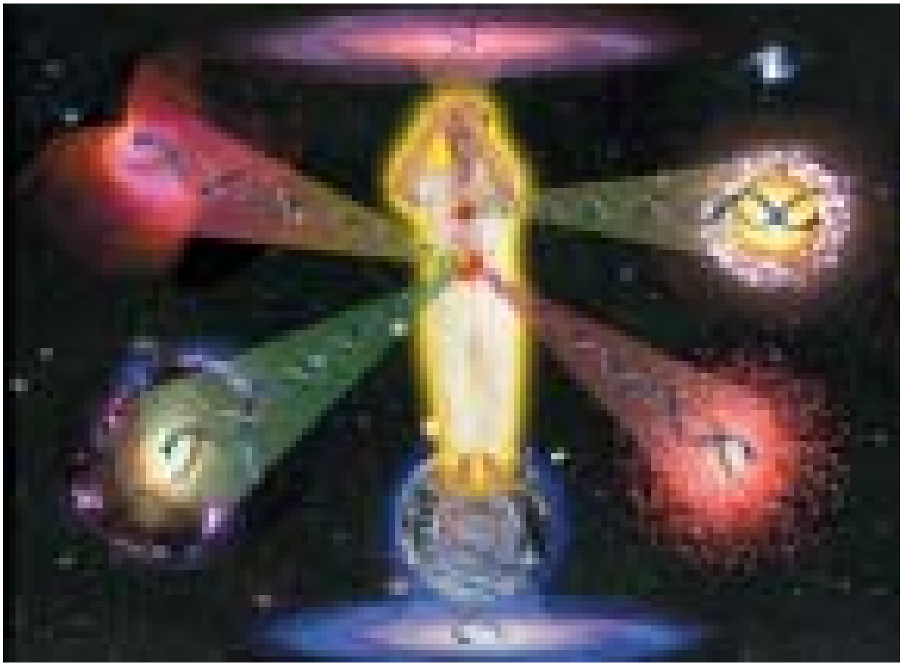

Fig. 4.3 *Trust and believe - turn visualization into actualization.*

C. Activate the Three Fires

1. Tan Tien Fire - Activate the Tan Tien Fire:

Smile down to the abdomen to create a burning stove near the lower lumbar and sacrum. Create a fire ball behind the navel, above the stove.

2. Kidney Fire - Activate the Kidney Fire:

Always retain awareness at the Tan Tien and keep it spiraling.

3. Heart Fire - Activate the Heart Fire:

Keep the heart soft and fill the heart with joy, love and happiness.

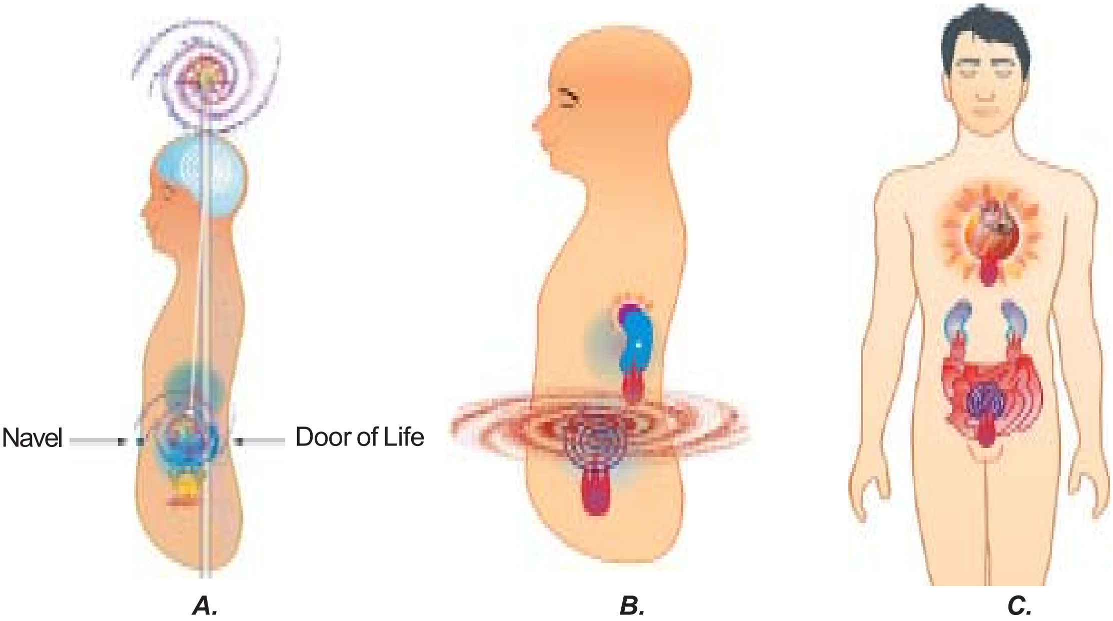

Fig. 4.4 A. *Activate the Lower Tan Tien Fire.* ***B.*** *Activate the Kidney and Adrenal Fires.* ***C.*** *Activate the Heart Fire.*

4. Sacred Fire Triangle

The Sacred Fire Triangle has triple force. Make a triangle from kidneys to navel (Tan Tien). From the heart make the connection to the kidneys. Move from the Tan Tien up to the heart.

5. Cosmic Star and the Earth Star

At the moment of conception, the two forces yin and yang connect with such a force, then only fractions of a moment later, the two forces explode into nine different Energy Centers. Seven of them we find in the body and two of them outside the body, forming our personal stars.

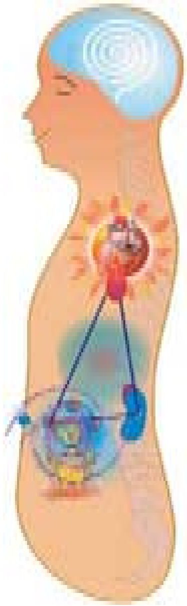

Fig. 4.5 Three Fires into One

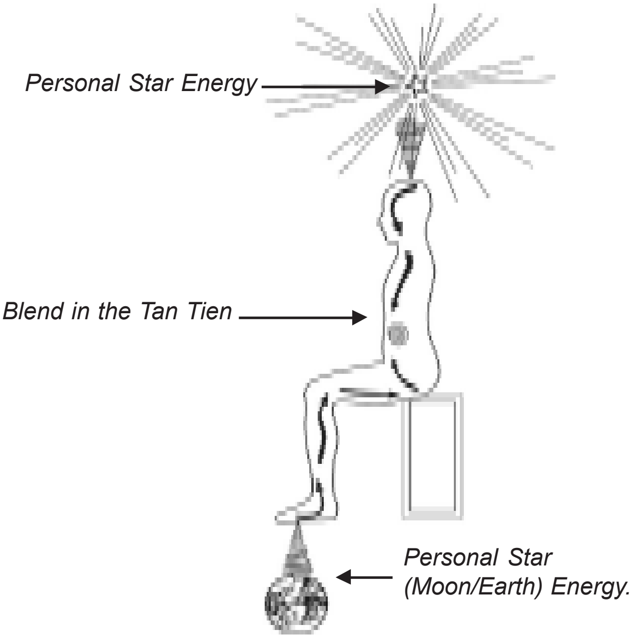

Fig. 4.6 'Personal Star' Energy

The two personal stars are in fact energy centers connecting the aural field of each individual with the universal forces and the earth forces. In a way one can see the aim of Taoist Inner Alchemy in bringing these nine forces together, merging them into one force and thus enabling the return to the original force, the Wu Chi.

There is a star about six inches above your crown and another one about one to three feet below the soles of your feet. They are also known as the higher self, guidance, protector, adviser. These stars are our connection to the cosmic force, the universal force and the earth deep below us.

Always make sure your Tan Tien is warm and the sacrum and mideyebrow are breathing. Be aware of the crown breathing and see a star or a small sun above you. Be aware of the crown and feel a light beam extend out of the crown and make a connection to the star above you. Keep on breathing until you feel a strong connection. Feel how the star above you is exercising a strong pulling force on your crown. Once you feel this pull on your crown you will also feel a strong pull down from the earth. Be aware of the star above and the earth and universal force below you. Feel that both of them have a strong pull on you.

1. Reconfirm the star above you and the earth directly below you.
2. Fix the image of the North Star and the Big Dipper six to nine feet above your crown. See the cup of the Big Dipper filling with the violet light, which gathers Chi from the North Star and the universe.

3. **Reach up with your left hand, and hold on to the handle of the Big Dipper. Pour the violet light down to your personal star 4 to 6 inches above your head, in order to predigest the energy of the violet light. Then let it flow down to your crown, into the Upper Tan Tien, to be processed and flow down to the heart center (Middle Tan Tien) or to the back of the head down to C-7 (Cervical 7) and T-2 (Thoracic 2) and down to the palms.**

   In Taoism we consider that the North Star emits a violet light, which is regarded as the highest healing light and the Big Dipper as emitting a red light.

   You can do these meditations sitting, or standing in the Chi Kung posture. The standing position will support a stronger structure and provide a better connection with the earth.

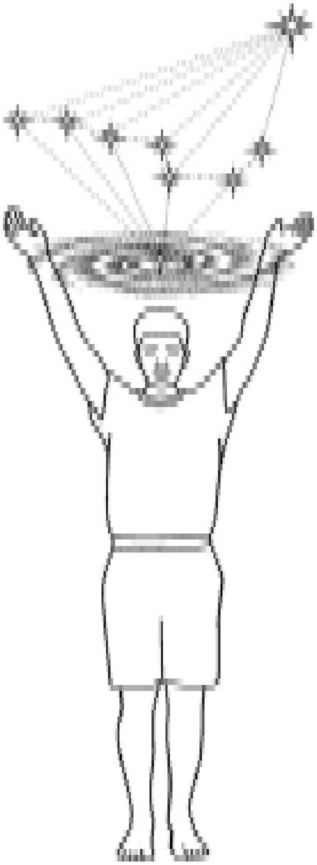

Fig. 4.7 *The North Star, Big Dipper and Galaxy spiral above.*

4. Always remember to breathe in slowly so you can process the awareness of the Lower Tan Tien. Feel a suction, warmth (Chi) and continue to breathe in and be aware of the suction of the crown, North Star and Big Dipper above you.

D. Linking Personal Stars, Energy Bodies and the Universe

Be aware of your 'personal stars' above and below and connect to other peoples' stars and energy bodies.

Three minds unite, one mind aware of the stars above you and a star below you. Connect these stars to the people around you. Think about the bright light above the crowns of other people. Start to spiral your energy to connect to each of them. Group all the stars into one energy body and connect all the way to the center of the universe.

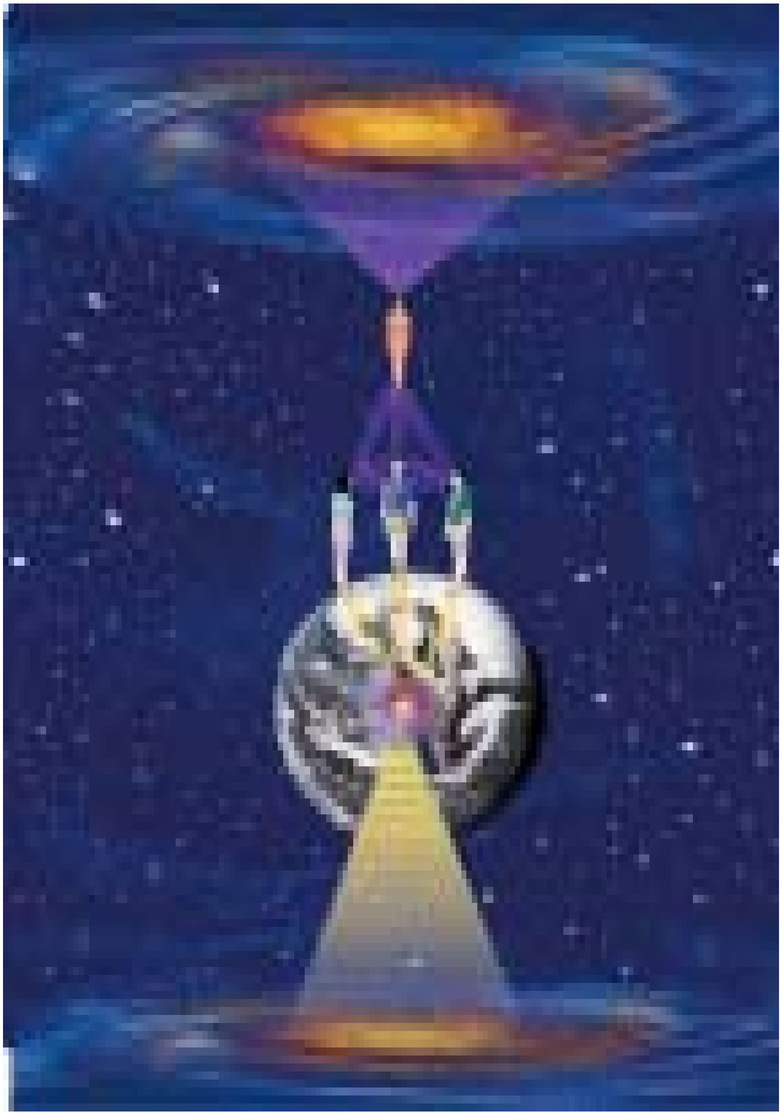

Fig. 4.8 *World Link Meditation*

1. Center Channel and Personal Star above

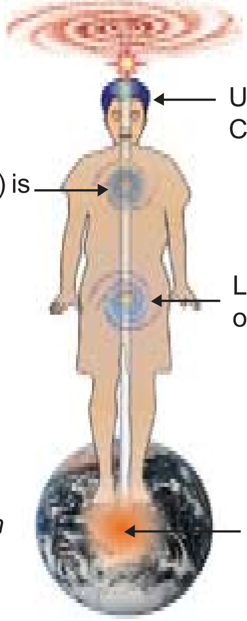
Soul, personal star or higher self
Upper Tan Tien (Upper Brain) is the Center of Observation.
Middle Tan Tien (Middle Mind) is the Center of Consciousness.
Lower Tan Tien (Lower Brain) is the Seat of Awareness.
Connect the Central Channel to the Upper and Lower Universe.

Fig. 4.9 *Connect Heaven to Earth using the Central Channel.*

2. Use Yi power to connect to your Personal Star above

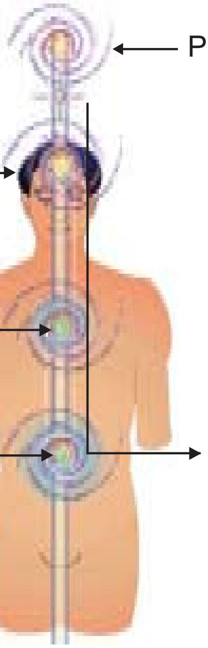
Personal Star
Observation
Consciousness
Awareness

When the abdomen fills with Chi it will rise up and fill the upper brain. Connect to your personal star, six inches above your head.

Fig. 4.10 *Connect to your Personal Star Energy using 'Yi' Power.*

3. Connect your Personal Star to the Personal Star of each person

Connect your personal star to each person and then to the universe.

4. Create a Chi Body becoming the major connecting point. Counterclockwise spiral to connect each person's star into a larger group star.

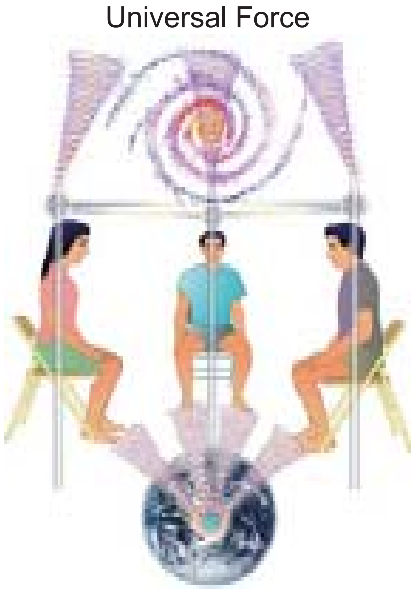

Fig. 4.11 Link with Each others Personal Stars.

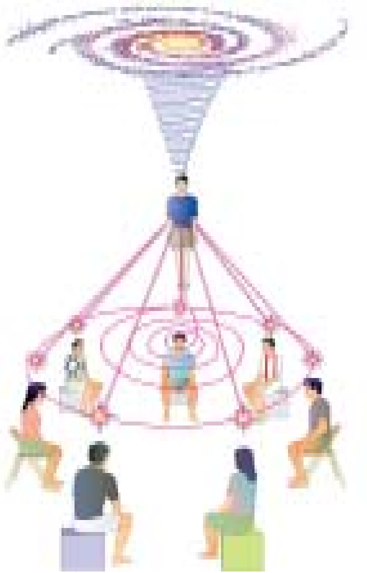

Fig. 4.12 Connect into Group Star.

5. Meditators become satellite links and create a group energy body.

The World Link meditations become an integrated communication link. They are linked to the earth and the universe. Spiral and connect each person's star together and group the energy into one energy body. This energy body becomes an integrated communication center for each person in the group. We can be connected, wherever we are. Each person can connect and become their own 'Center of the Universe'.

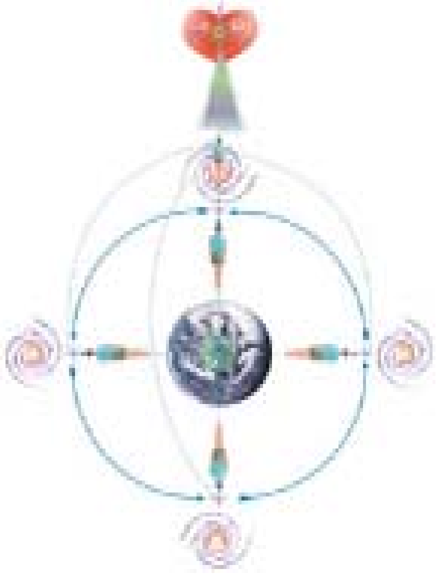

Fig. 4.13 *Form a Satellite Link and create a Group Energy Body.*

E. Protective Circle and Chi Field

Raise your hand up and become aware of the 'Sacred Fire' in the universe. Feel that your fingers are very 'big', very 'long' and touch the fire. Bring the fire down. Make a circle around your community, your house, Tao Garden, the meditation hall and around your bodies. Create an energetic field of Chi around the whole room.

Sacred or Chi Fire
Connect and Receive Fire from the Universal Cauldron

1. Activate the Sacred Fire in the Universe

Be aware of yourself. Feel your arm and fingers 'big', 'long' and 'hollow'. Reach to the universe to the Sacred Fire. Let the Sacred Fire fill and pack your arm.

2. Use Yi-Power to Create Sacred Fire Protective Circle

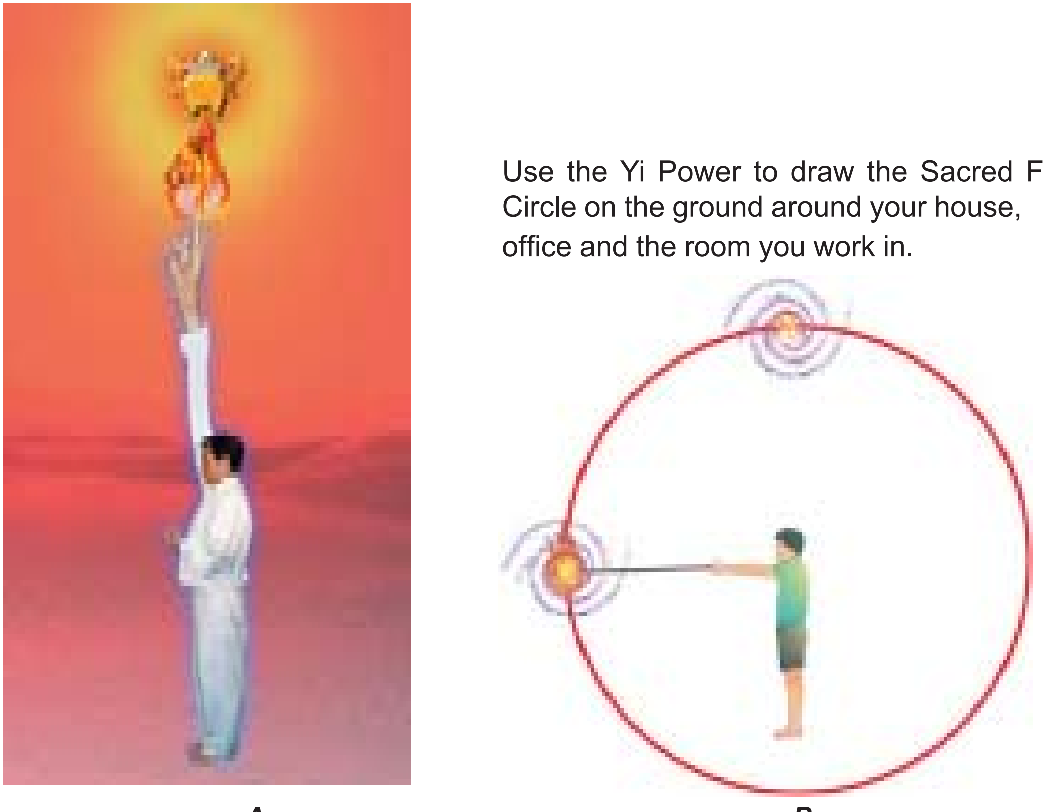

Use the Yi Power to draw the Sacred Fire Circle on the ground around your house, office and the room you work in.

A.

B.

Fig. 4.14 A. *Create a Group Energy Body.* ***B.*** *Sacred Fire 'Protection'*

3. Create a Chi Dome

Set up the 'guardian animals', Blue Tortoise in the north, Red Pheasant in the south, Green Dragon in the east, White Tiger in the west, Yellow Phoenix above and Black Tortoise below.

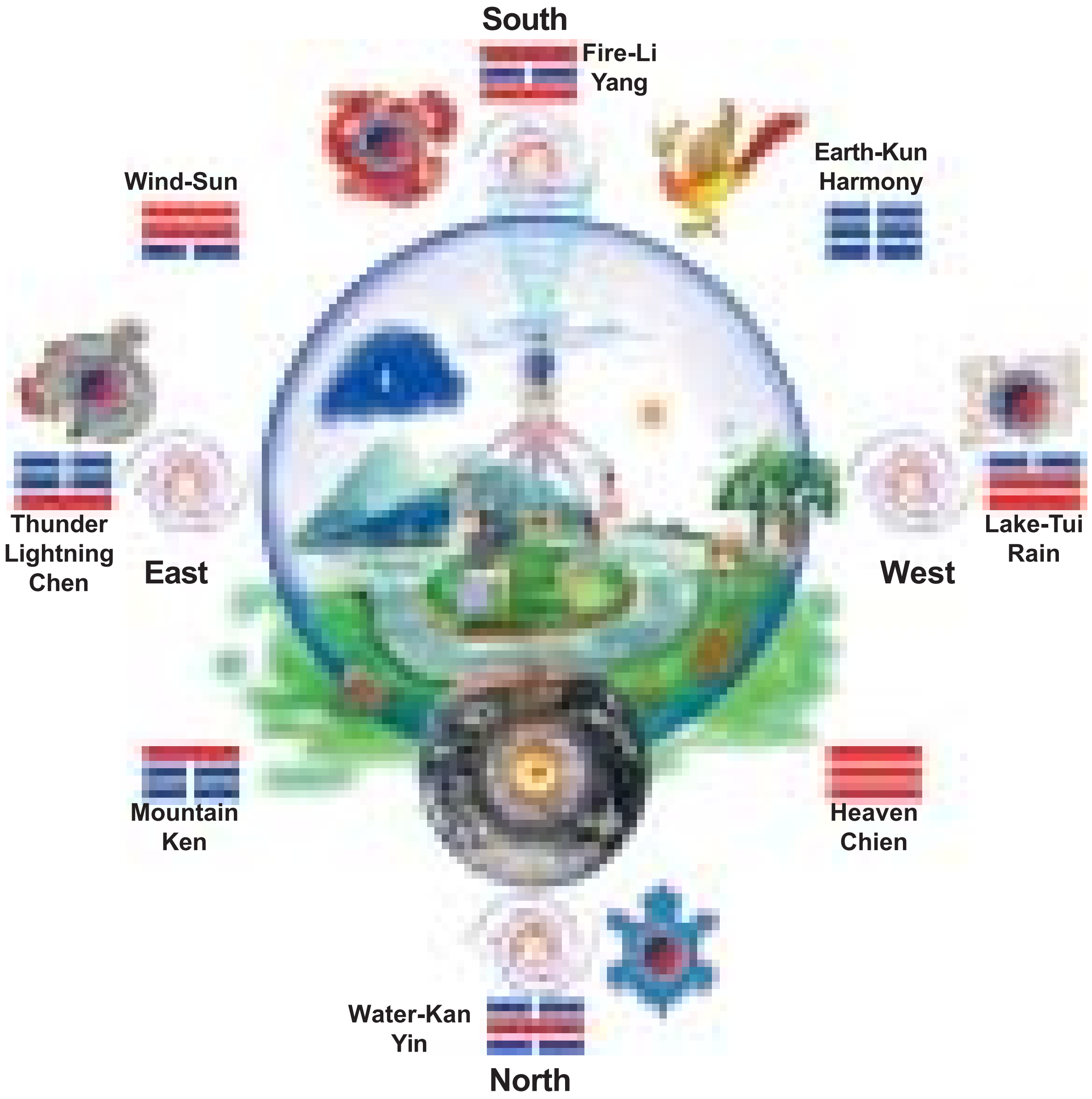
South
Fire-Li
Yang
Wind-Sun
Earth-Kun
Harmony
Thunder
Lightning
Chen
East
West
Lake-Tui
Rain
Mountain
Ken
Heaven
Chien
Water-Kan
Yin
North

Fig. 4.15 Create and surround yourself with a Chi Dome and the Guardian Animals.

4. Activate the Eight Elemental Forces of Nature and the Universe

Activate all eight forces; wind, mountain, fire and thunder on the east. Earth, lake, water and heavenly power on the west.

Call the Eight Forces; fire, water (ocean), thunder (lightning), lake (rain), earth, mountain, wind, and heaven.

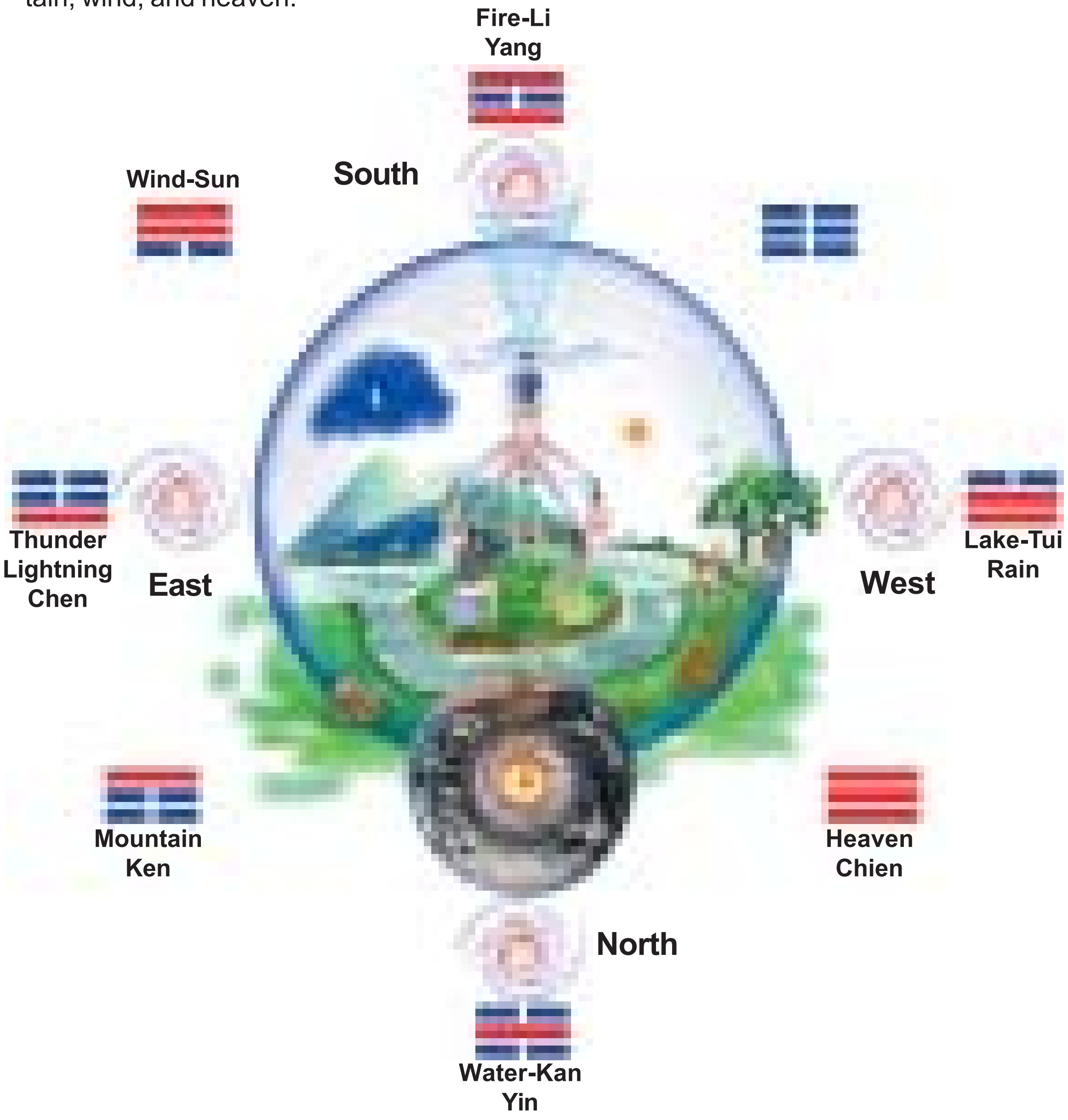
Fire-Li
Yang
Wind-Sun
South
Thunder
Lightning
Chen
East
West
Lake-Tui
Rain
Mountain
Ken
Heaven
Chien
North
Water-Kan
Yin

Fig. 4.16 Activate the Eight Forces.

In the next sections F,G,H,I are practices that are for cleaning the cells. The more you clean the cells the more the Cosmic Healing can take place.

F. Chi (Sacred) Water Practice

The Sacred Water practice (also known in the West as Holy Water) is to exercise the right of being the creators of the Cosmos. With strict and absolutely regular daily water intake to prevent the stresses and associated damages of dehydration, the chief conductor and supervisor of the body's well being – tryptophan and its neurotransmitter derivatives, serotonin, tryptamine and melatonin will be well positioned to regulate all functions. Regular daily walks will keep muscles well coordinated and correct any physiological processes that are established in the body as a result of anxiety or emotional stress.

A well hydrated and healthy skin needs water to constantly replace that which it loses to the outside environment. This then allows blood vessels to the face and the body to open up and provide necessary nourishment for exposed skin cells.

Science has discovered that even if we clean water through filtration processes, although it may look and taste alright, closer analysis using high powered microscopic technology, indicates that its cell structure is 'unclean'. Water, like humans, maintains the memory of its old 'self'. We store our 'memory' within our DNA; water stores its memory within its cells.

The world is about 70% water

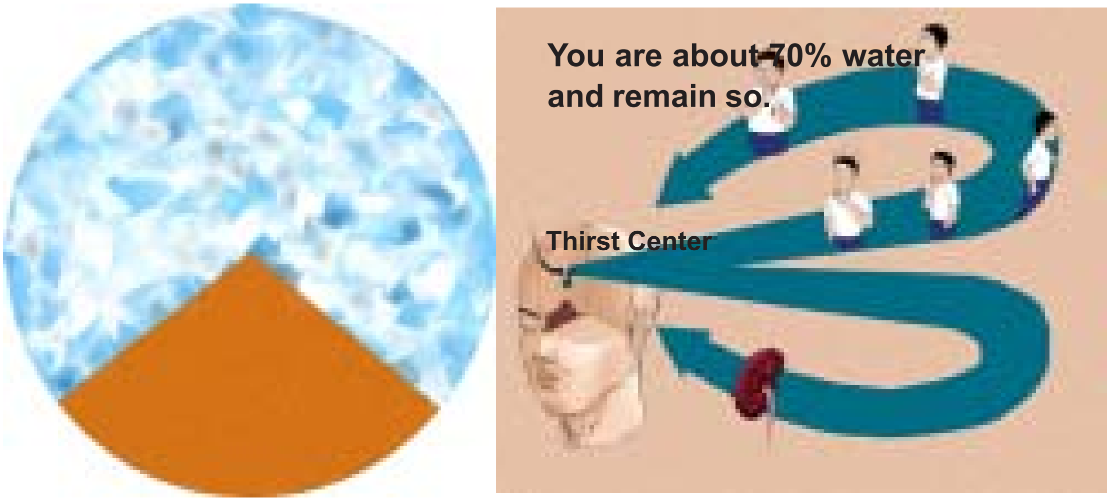

Fig. 4.17 *Water is Life.*

The following three pages contain extracts from the book ‘The Message from Water’ by Masaru Emoto. It advises us that ‘water is telling us to take a closer look at ourselves - water is a mirror reflecting our mind.’

After offering a prayer to the Fujiwara dam the message is reflected in the water.

When we project Chi and love into water and then freeze it; it will turn into a ‘regular’ and well structured crystal. If we project negative energy into water, the opposite will apply. Our projections will access the cellular level, clean out the polluted cells and reprogram those remaining with the essence of our affirmation.

When we project a nice, loving instruction or thought into water, like ‘let’s do it’ the water will pick up this vibration and restructure itself. If we project negativity into the water, like a command, ‘do it!’, the water will acknowledge this accordingly. When we look into a pond we see our image. If we drink a glass of negative image water its destructive energy will enter our system and attempt to multiply.

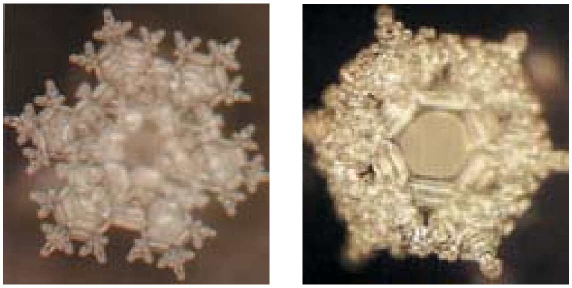
Chi and Love **Love-Appreciation**

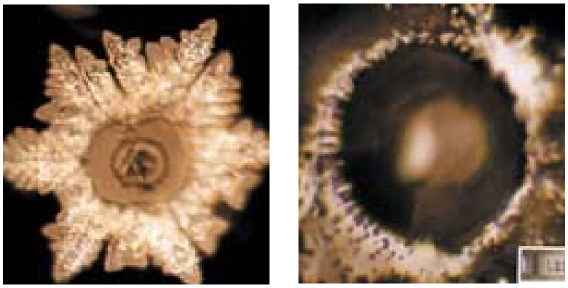
Let’s do it **Do it!**

Fig. 4.18 *The Human Body consists of 70% Water.*

Research has demonstrated that if we completely filter polluted water, using every process available, it will still maintain its old vibration and unstable structure. It will maintain memory in the same way that a computer does. We must therefore reprogram the cells with our intent.

If the water we drink contains too many pollutants and antioxidants it will eventually poison us.

As we can see from these pictures, there is an enormous difference between natural spring water and regular city water. It is now quite clear that if we program water so that it becomes, sacred, holy or Chi water, then it will have the power to remove all antioxidants and pollutants – all sickness. It will have the power to reprogram the water within our bodies, to carry the same pure structure.

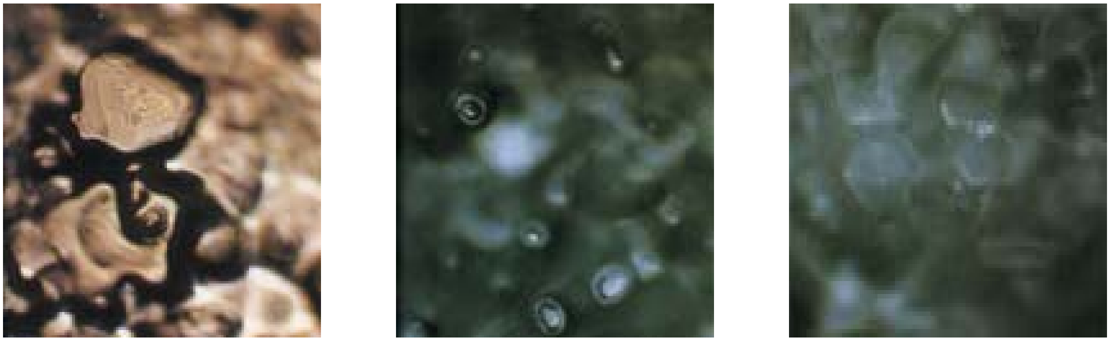

Tokyo Tap Water **London Tap Water** **Paris Tap Water**

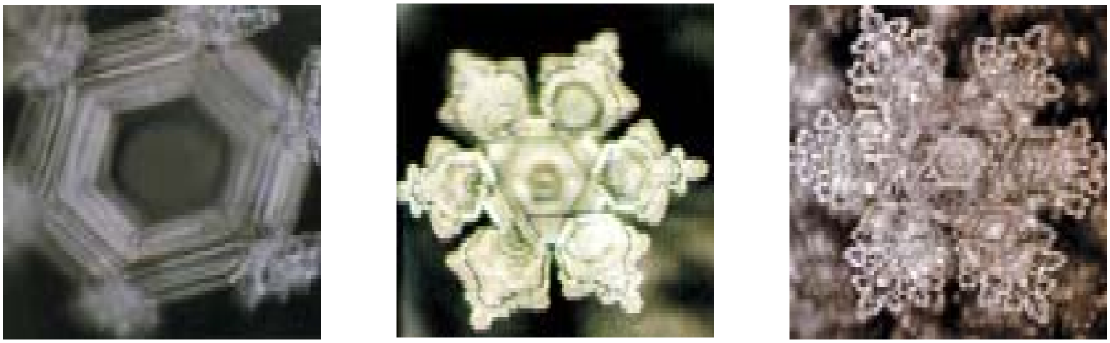

Under Ground Water in Northern Island, New Zealand **Buenos Aires, Argentina** **Spring Water of Saijo, Hiroshima Prefecture**

Fig. 4.19 *Healing or hurting?*

We can also use music to program a vibration into water. Classical music will create a vibration that synchronizes with our own health, whereas heavy metal music will have a tendency to 'shatter'!

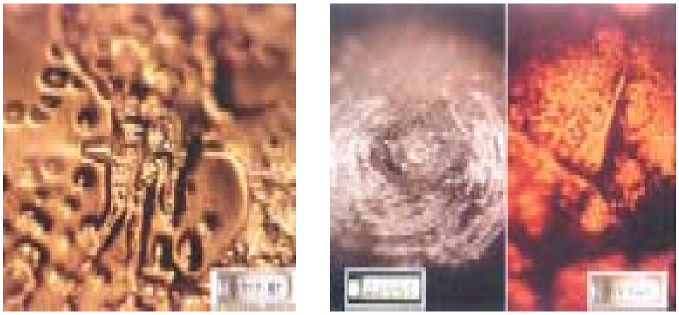

"You make me sick", or "I will kill you".

You Fool

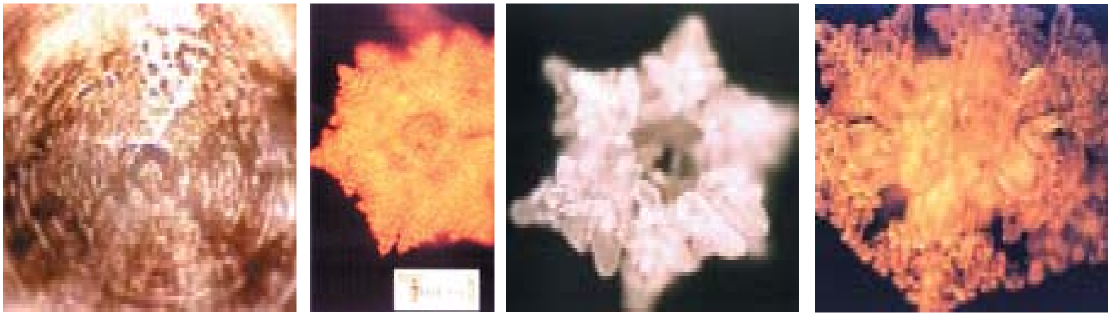

Heavy Metal Music

Thank you

Folk song of Celtic Region in UK

Healing Music

Fig. 4.20 *Power to heal or harm?*

Similarly, when we program our emotions regarding other people into water, the effects are the same. If we think negatively of someone, then the response will produce a bad vibration. If we think kind, loving thoughts then the structure will be like a beautiful crystal. If we love Mother Theresa and send this energy into the water, a wonderful crystalline structure will be created. Sick water can be transformed into healthy water. Healthy water can be transformed into sick water. The human body consists of 70% water. The key is your intent and your ability to believe that healing comes from the nature of your vibration. Always be open to receive.

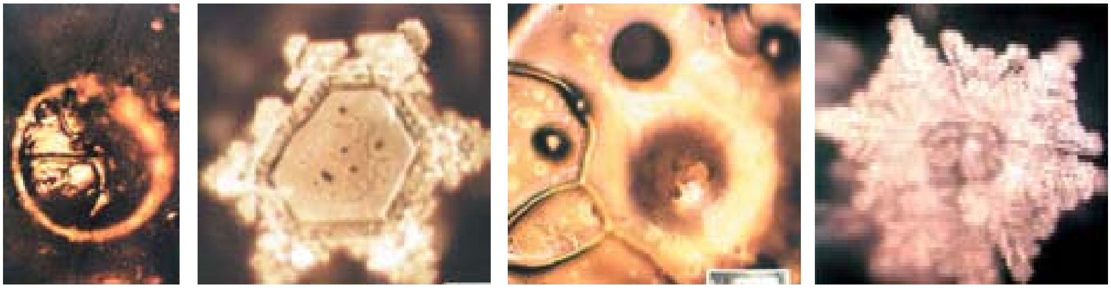

Adolph Hitler

Mother Teresa

Dirty

Beautiful

Fig. 4.21 *Sick or saintly?*

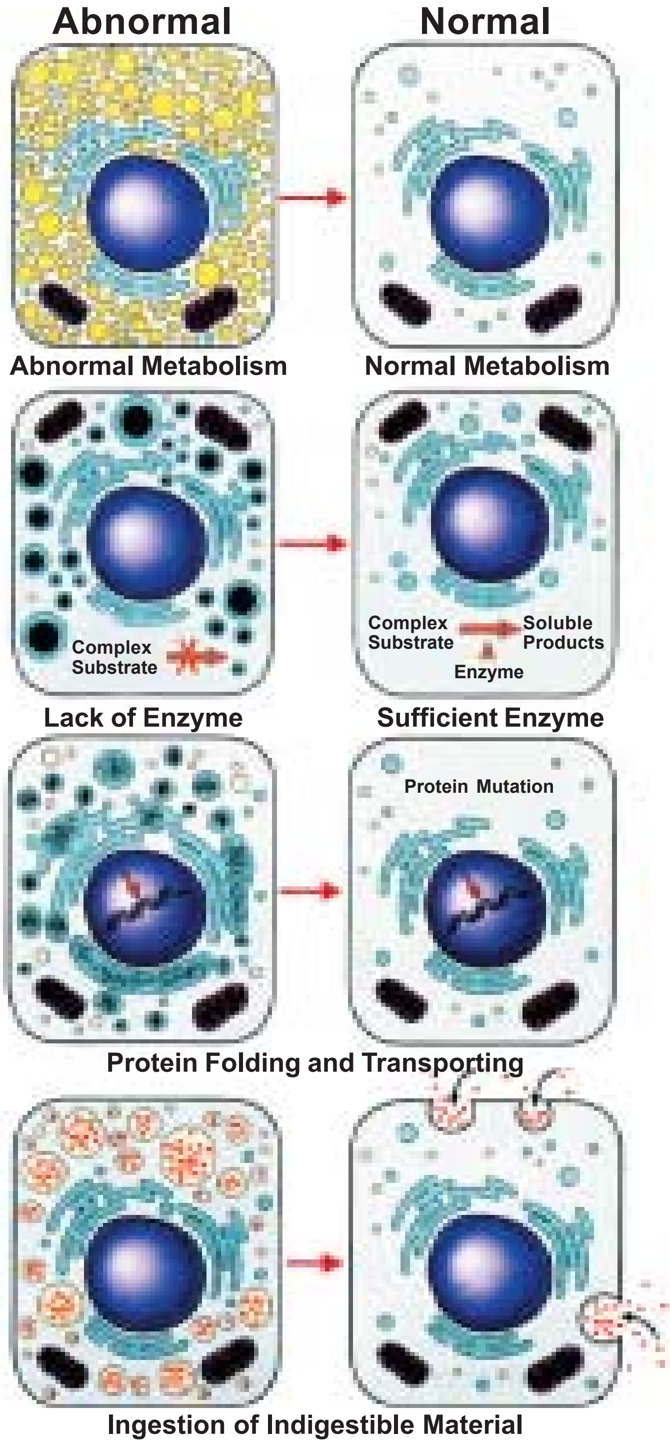
Abnormal
Normal
Abnormal Metabolism
Normal Metabolism
Complex Substrate
Complex Substrate
Soluble Products
Enzyme
Lack of Enzyme
Sufficient Enzyme
Protein Mutation
Protein Folding and Transporting
Ingestion of Indigestible Material

Fig. 4.22 With the Power of the Sacred Water Abnormal Cells can be cured.

A positive mental attitude will be transmitted into the cells of the water contained within our own body. We must consciously practice inner work until it becomes reflexive.

Outline of Sacred Water Practice

Invoke the power of the Sacred Water practice to cleanse and heal the body’s sick, toxic or negative energy. **If you are working with a group of people, direct your energy through the energy body overhead and into each person’s star, as you guide them through the procedure**.

I. Hold a glass one quarter full of water in your left hand, folding the middle and the ring fingers into the center of the palm.
II. Hold the glass in front of your body as you point the ‘Sword Fingers’ of your right hand up to Heaven.
III. Make a request to receive healing energy and feel yourself touch a heavenly pool of sacred water energy. Feel the pool pouring down heavenly water to fill your arm.
IV. Place your fingers on the cup and ask:
Command 1: ***“Yin power and good fortune come from the east. Yin power please dissolve all negative energy, all sickness and bad fortune”.*** Make a cross over the top of the cup.
V. Bring your arm down and point the sword fingers around the inner rim of the glass. Smile as you circle the fingers around the rim.
Command 2: ***“By my request. Please carry out the order now.”*** Repeat this three times, charging the water with Yin Power. Project your thoughts into the water.
VI. Use the thumb and index finger to remove the sick energy from the cup, without touching the water. Do this three times. Talk to the sick cells. Tell them to listen to your command:
Command 3: ***“All sick cells please listen: Clear, clean and bright, this Sacred Water will take all the sickness away.”***
VII. Repeat the order and visualize yourself removing the sick energy from the cells and giving it to Mother Earth for recycling.
VIII. Hold the cup with both hands near your heart.
Command 4: ***“This Sacred Water will give me/you health, wealth and longevity.”***
Project love, joy, thankfulness, gratefulness, appreciation and the energy of compassion into the cup.
IX. The Sacred Water will carry the message of the practice to all the cells, where it will remain. If you are practicing in a group, drink the water in unison. If you are working with a student, pass them the glass to drink using both hands.

Step I: Prepare the Hands and Arms to Receive the Universal Chi

Left Hand Position: Prepare to hold a cup or glass with the left hand by folding the ring and middle finger into the center of the palm. Hold the glass in front of the body.

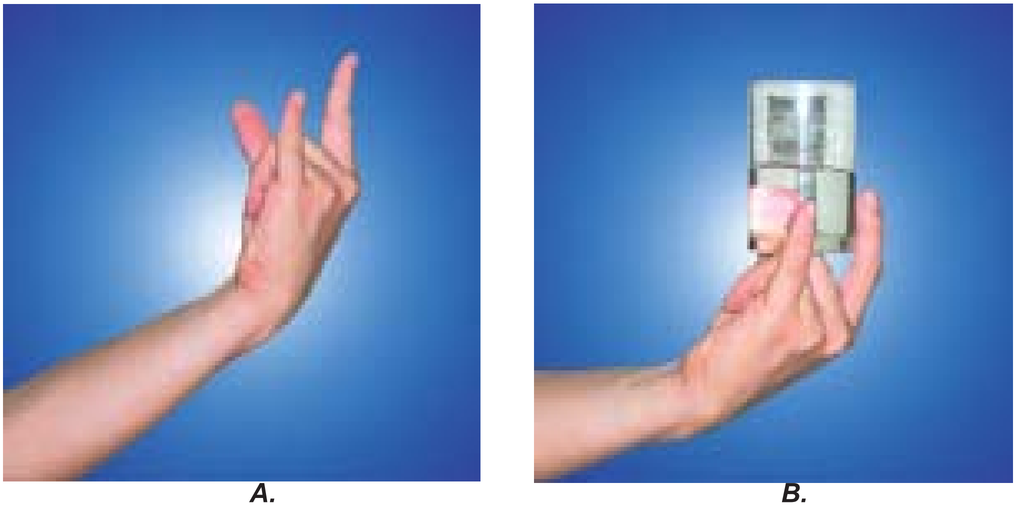

Fig. 4.23 A. *Fold Ring and Middle Fingers.* ***B.*** *Hold Glass in Front of Body.*

Step II: Prepare the Sword Hands to Receive the Universal Chi

Right Hand Position: Form the right hand into a ‘sword hand’ by folding the pinkie, ring fingers and thumb into the palm. Straighten and hold the index and middle fingers together and extend them upward.

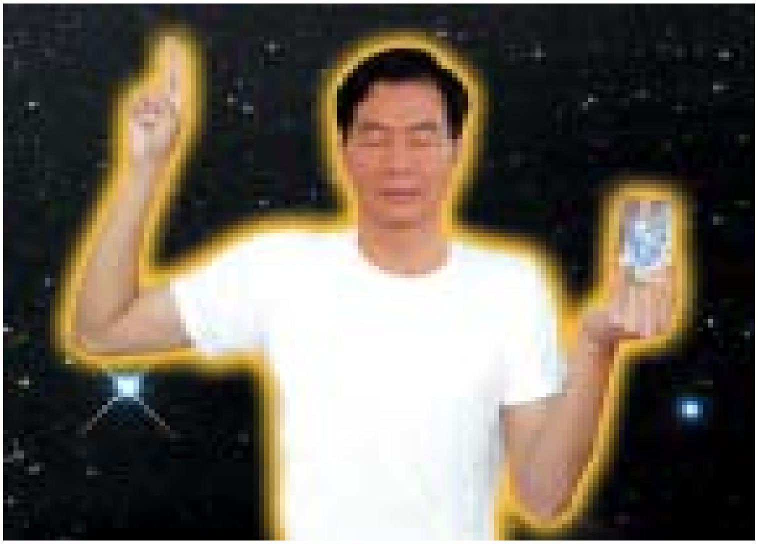

Fig. 4.24 *Prepare to receive the Universal Chi.*

Step III: Fill your Arm with the Power of the Heavenly Pool

Sense the sword fingers and the arm as being 'long' and 'big' as you raise them towards Heaven. Sense that the middle of the arm is hollow and bottom is sealed at the shoulder. The mind's attention is focused on merging with the primordial Chi of the Universe, the energy of your thoughts will be multiplied.

As you make your request to the Universe, feel that you are touching a heavenly pool of Sacred Water Energy. Feel that the pool is pouring down like a waterfall to fill the arm. When it is full compact, compress the Sacred Energy into it as much as you can.

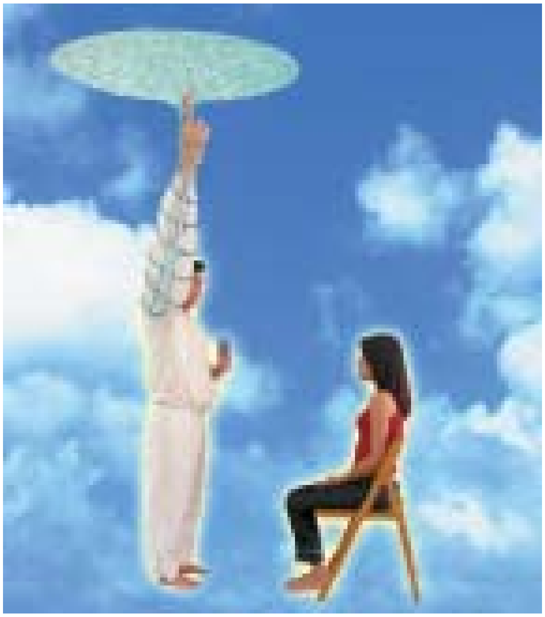

Fig. 4.25 *Touch the Heavenly Pool.*

A. ***B.***

Fig. 4.26 A. *Blessing Sacred Water in Russia.* ***B.*** *Opening the Third Eye in Russia.*

Step IV: Make the Cross on the Top of the Cup

Put Sword Fingers on top of the Cup. Ask for the Yin Power: **Command 1:** ***"Yin Power and good fortune, come from the east. Yin has the power to dissolve all negative energy, all sickness and bad fortune."***

Command 2: Make the cross on the top of the cup and say ***"By my request......"***.

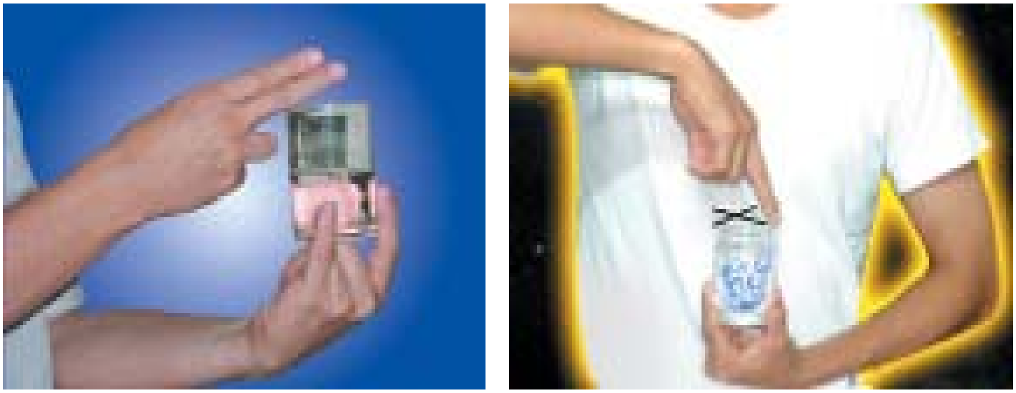

Fig. 4.27 *Give the command: "Fortune comes from the East - Yin has the power to dissolve all negative energy, all sickness and bad fortune. By my request....".*

Step V: Charge the Water to Transform it into Sacred Water

Bring your arm down and point the sword fingers into the glass of water. Smile and circle the sword fingers around the inner rim of the glass.

Continue to order: ***"Please carry out the order now."*** Repeat 3 times to charge the water with the Yin Power. As you do this, project a good thought into the water.

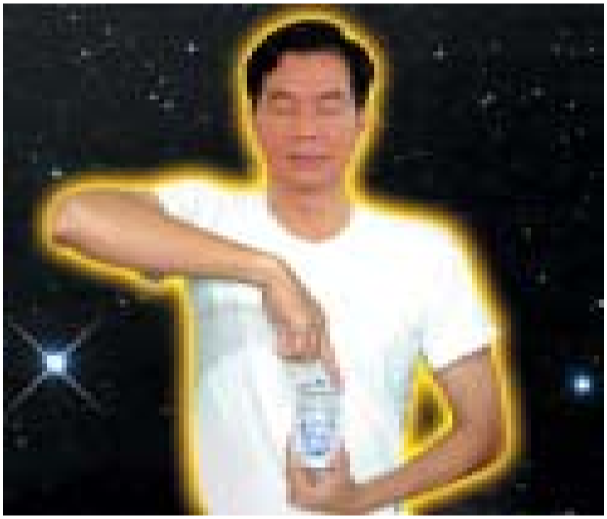

Fig. 4.28 *"Please carry out the order now" Repeat three times, charging the water with the yin power.*

Step VI: Command 3: *"All sick cells please listen : Clear, Clean and Bright, this Sacred Water will take all the Sickness away."*

Give the above command. Use the thumb and index finger to pick up sick energy from the cup, without touching the water. Do this three times.

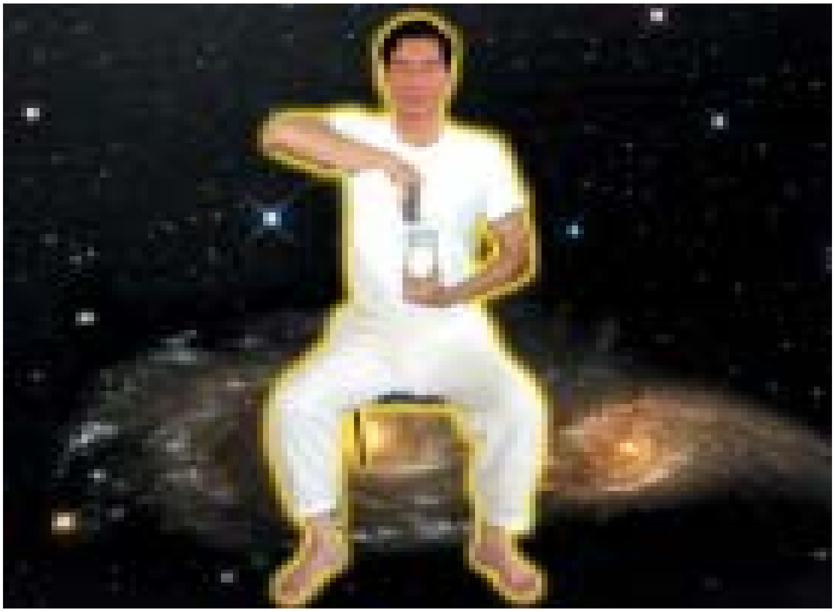

Fig. 4.29 *"All sick cells please listen: clear, clean and bright, this sacred water will take my/your sickness away".*

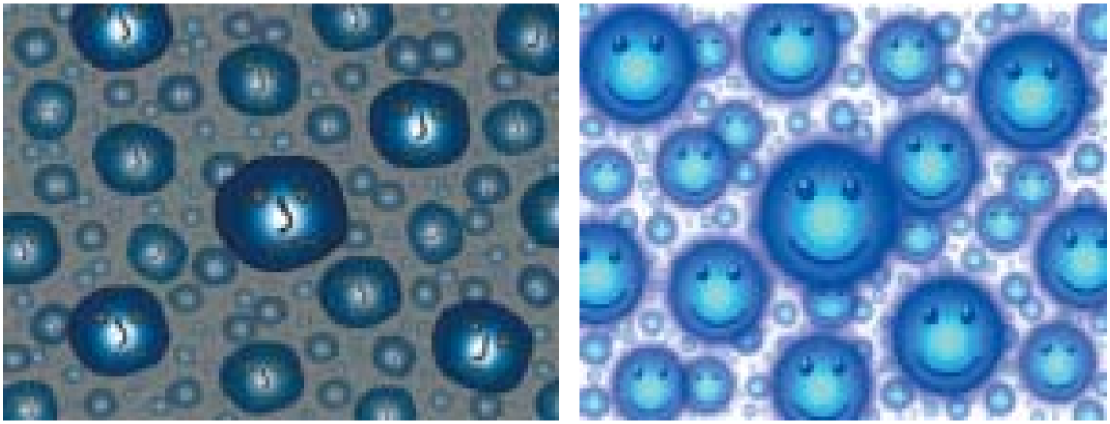

A. ***B.***

Fig. 4.30

A. *Ask for the power to see the cells - if they are toxic and dark, ask for them to be cleaned out.*

B. *"Clear, clean and bright".*

Step VII: Let go of the Sick Energy to be transformed by the Earth

As you **repeat** the order, picture yourself removing the sick energy from the cells and discarding it down into the Earth to be transformed and recycled.

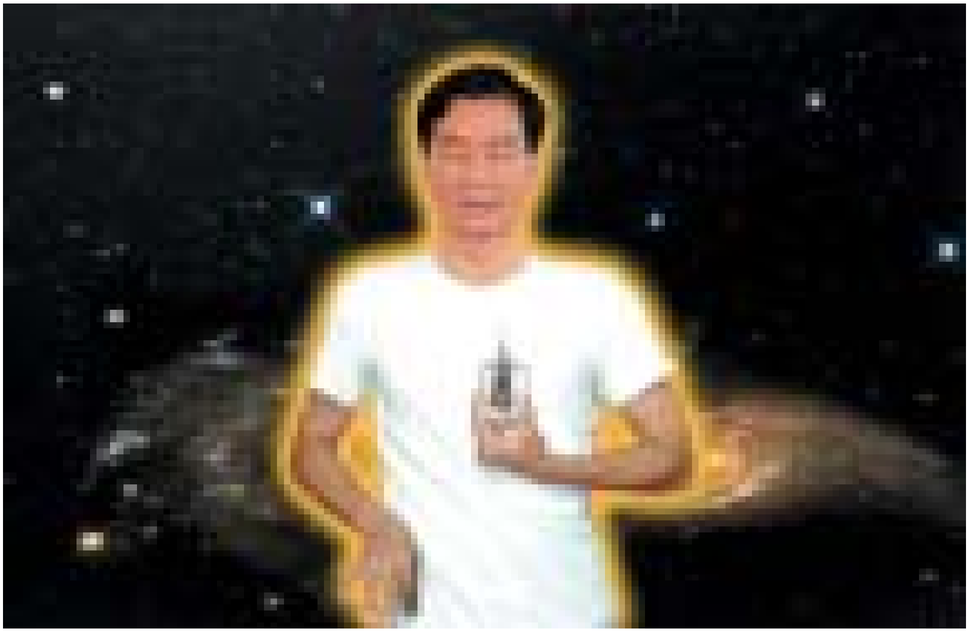

Fig. 4.31 *Sick Energy return to the Earth.*

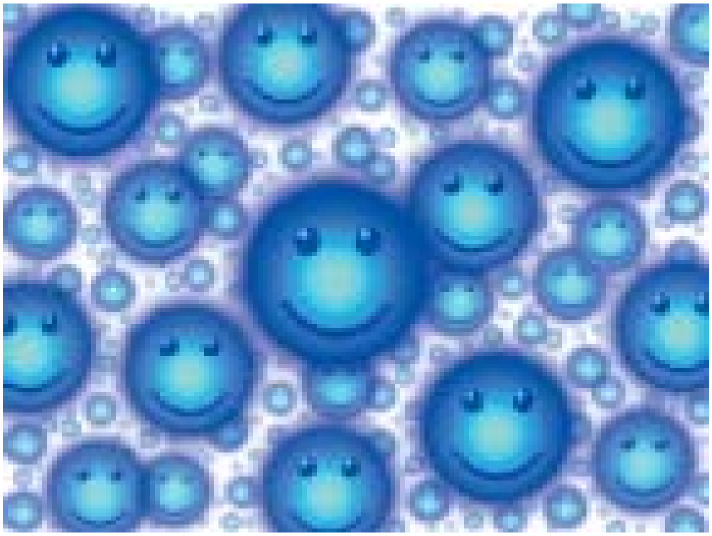

Fig. 4.32 *See the Cells become Clean, Bright and Smiling.*

Step VIII: Fill the Water with Compassion

Hold the cup with both hands near the heart and project goodness into the water. **Command 4:** ***"This sacred water will give me/you health, wealth and longevity."***

Project love, joy, thankfulness, gratefulness, appreciation and the energy of compassion into the cup.

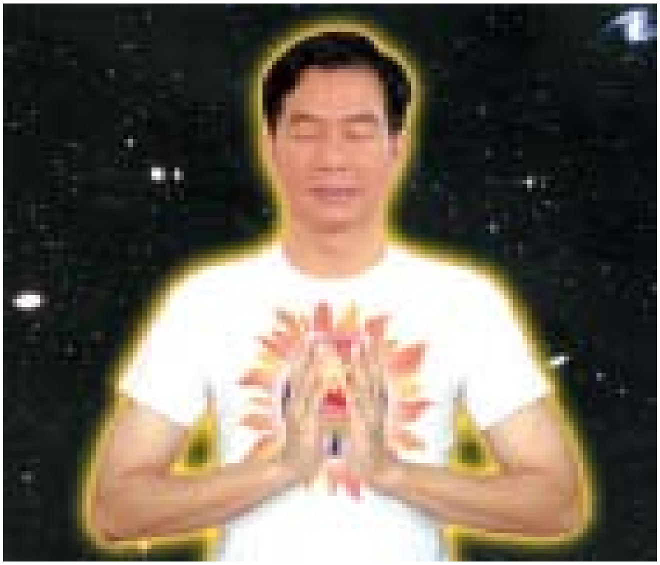

Fig. 4.33 "This Sacred Water will give you health, wealth and longevity."

Step IX: Drink the Sacred Water

The Sacred Water will carry the message to all the cells. It will keep the message of the Sacred Water practice in all the cells. Use both hands to pass the Sacred Water to the student you are working with, or drink it yourself and/or in unison with your group. **Feel the water go into all the cells of your body. Feel it removing the sick cells and purifying your body.**

You can also sprinkle it on to any area that requires healing.

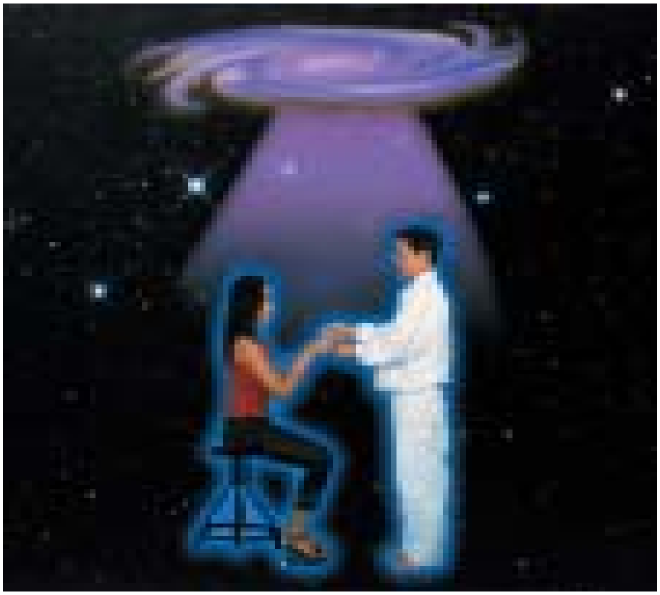

Fig. 4.34 Drink the Sacred Water.

G. Empty and Fill

Ask for permission to work on the student by communicating with their personal star. Point your sword hand down to the ground. Project your hand ‘very big and very long’ and your feet ‘very big and very long’ - extended down to the ground. Now very slowly inhale, bring up the earth energy. An important part of the practice is to extend a ‘line’ down from universe into yourself and one into the student. The line sinks through the Tan Tien into the universe below. You can send your energy up to your energy body and it will be brought down to the energy body of the other person.

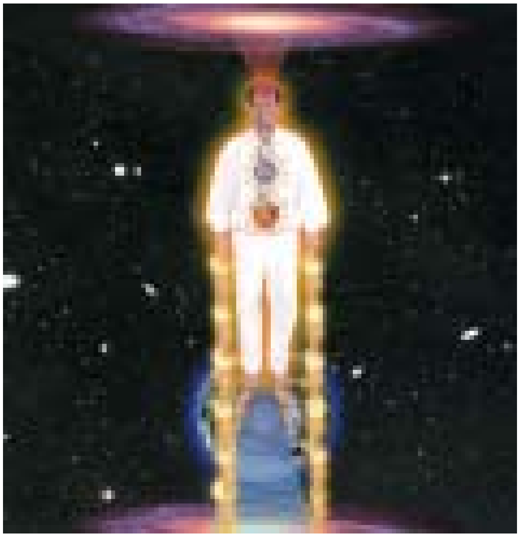

Fig. 4.35 *Feel your hands are long as down into the earth they extend.*

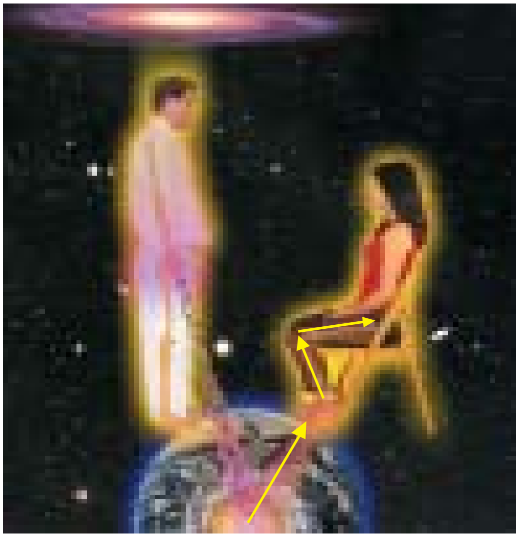

Fig. 4.36 *Project the other person’s legs deep into the earth.*

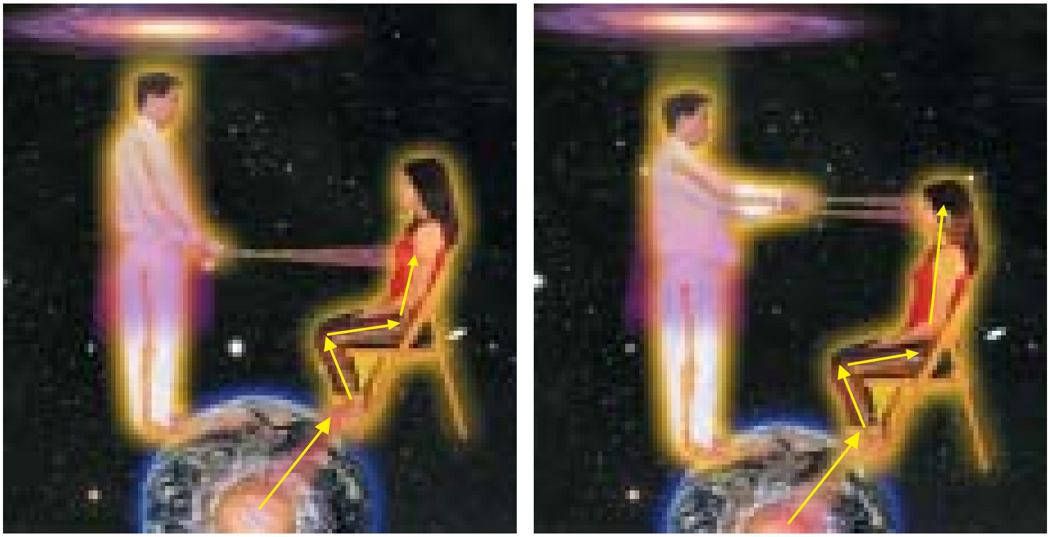

Fig. 4.37 *Blend the universal energy and Earth Chi - bring it into all of the bones and cells.*

Bring the energy up to your feet, into your bones. Let it enter all the bones, organs and cells. Feel it blending with all the sick energy. Turn your palm up, turn it back down again and exhale; let go of all the sick energy, the negative energy, the worries and the burdens all the way down to the earth. Down, down, down to the earth. Again from the earth up. Do it at least three, six or nine times. When people are very sick, you need to clean them thoroughly. Bring the earth energy up into your bones again. Feel numb, tinkling electricity flowing in all your cells. Return the energy down, down, down to the earth. Relax, let go; worries, burdens, let everything go. It is very important to say to yourself or group: **"You must let go of all your sick energy. Let go of all your burdens. Let go of all your worries, down to the earth".** Repeat it once again. Inhale from the universe and into the earth. From the earth let it enter into your bones. From the bones all the way to your inner self. Feel the energy blend with the sick energy. Exhale and let go of all your sick energy, any negative emotions burdens, worries, anxieties; release them all the way down into the earth.

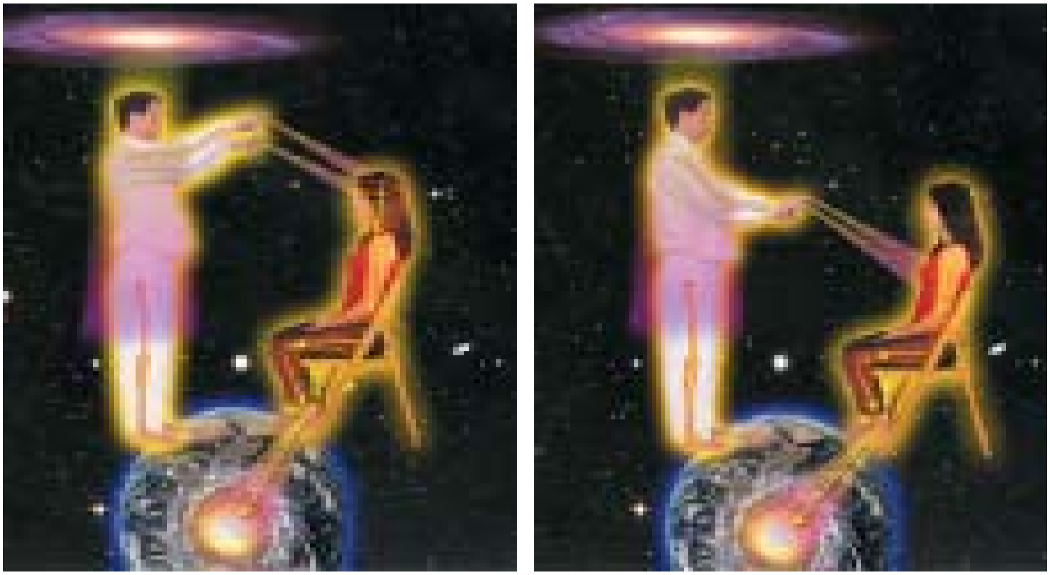

Fig. 4.38 *Bring the Universal Chi down and let it blend with the sick and negative energy. Bring it down into the earth and bury it there.*

Empty negative, sick energy, burdens, worries, go deep down into the earth. Dig a hole and bury them there - tell them "Do not come back; you will be happy there."

Feel that the feet are long and extended into the earth. The hollow bones have been compacted with compressed Chi.

Once again guide the good Chi from earth up to the fill the person again. See the cells as big as the stars and fill them with Chi to blend with the dirty and sick energy. Then empty them again down to the earth.

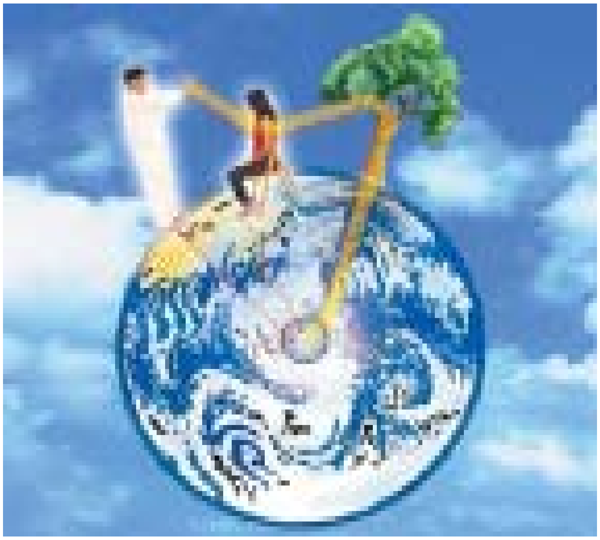

***Fig. 4.39** You can also use a Tree to guide the Chi down into the Earth.*

H. Clean with Green Light

Become aware of your liver and breathe into your throat center. Feel that you are breathing the green light into your throat center and let it blend in the Tan Tien. The Chi comes up to blend in the heart, up to the crown, spirals and goes up to the universe. Raise your hand in a spiraling motion up to the universe. Use your 'one hand' to spiral in the universe. Spiral clockwise first. Spiral green light from the universe. Now reverse, counterclockwise - multiply and spiral counterclockwise, bringing the Chi down to clean out your communities, your house, the Tao Garden, the meditation hall, your bodies and all of the people that are 'here' - fill yourselves with Chi. You will use this power to see the cells.

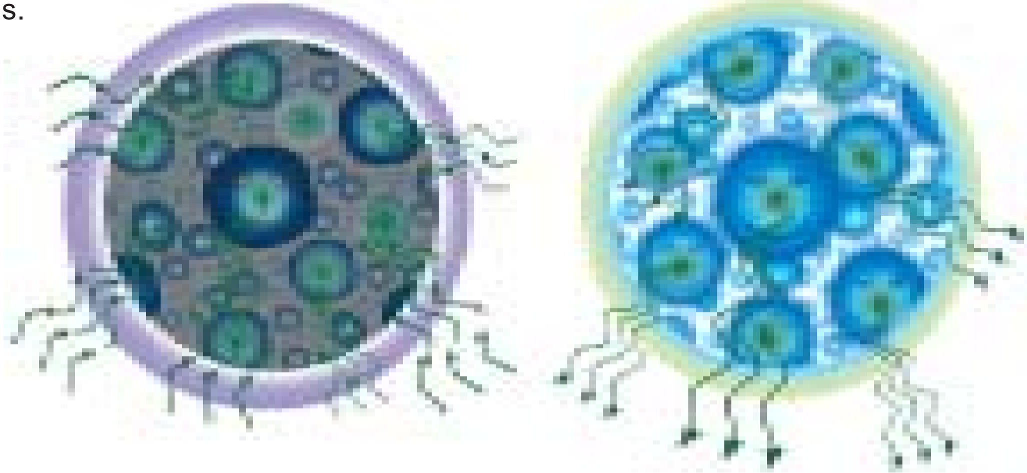

***Fig. 4.40** Spiral Sick Energy Counterclockwise down to the Earth. Smile to the Cells and let the Green Light clean out the sickness, flushing it down into the Earth below.*

See the cells enlarged and this green energy go in and blend with the dirty, black and cloudy energy. See it come out and let it go down into the earth. Dig a hole and bury it there. Tell it, "Don't come back. You will be happy down there". Become aware of the green light and spiral counterclockwise. Spiral down, cleaning out the stress. Clean out your house. Send everything down into the earth. Repeat three, six, nine, thirty six or one hundred and eight times. This depends on how much sick energy you are dealing with in your body.

Green:

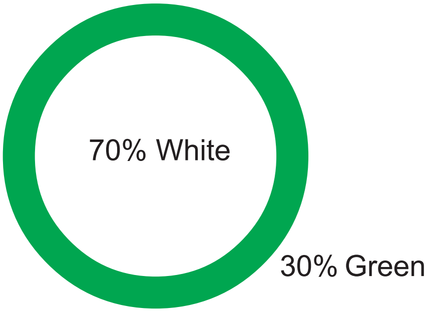

Fig. 4.41 *Clean with green.*

Green cleans and detoxifies all of the body's cells. Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. It 'loosens' the area. The light green Chi is used to break down dirty or diseased energy, like a detergent is used for washing clothes or dishes. Afterwards, the cells will be rinsed with blue light, like clean water is used to rinse the detergent and residue from clothes or dishes.

Throat Center Green

Be aware of the liver's green color, the throat and the forest. Breathe into the throat, feel yourself breathing in the green light. Guide it down to the Tan Tien, blend and move it up to the heart and the crown.

Project it to the universe; spiral and let it multiply.

Channel one direct line down to the student, one to your crown and one to your palm.

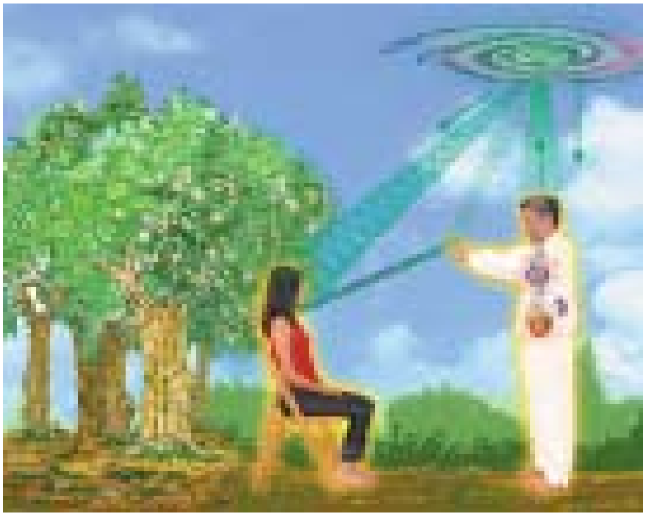

Fig. 4.42 *Blend Green Chi into the Three Tan Tiens - project to the Universe, multiply and bring the Energy back down to the Student.*

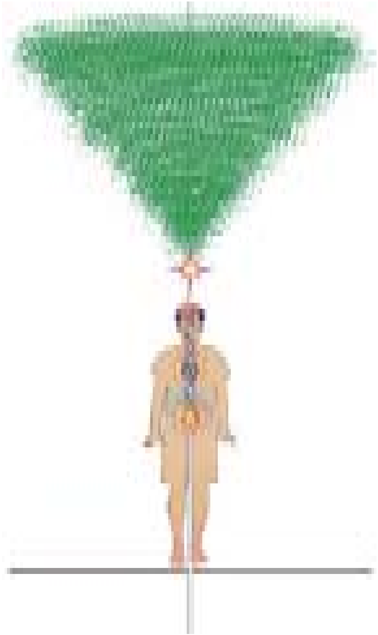

Be aware of the liver, the gall bladder, the forest and the throat breathing.

Be aware of the green light in the universe. Let the light come in and blend in the Tan Tien, up to the heart to blend with the compassion in the heart, and up to the crown.

Project the light up to the universe. See it spiral and let it multiply.

Always keep your Lower Tan Tien Fire warm. **Spiral in your Tan Tien and spiral the energy down to the student.** Spiral 1,000 revolutions per minute, 10,000, 30,000 and 60,000.

Fig. 4.43 *Draw down the Green Light from the Universe.*

Cleanse and Remove Sickness

Extend your arms up, with palms facing the heavens. Use one hand to spiral the green force down. The other hand holds the position and connects to the universe. Let this big pool of green light in the universe spiral down – through your community, your home and then into your crown.

Let the Green light blend into the cells and bond with sickness and toxins. Let the green light bring the sickness out of the cells and flow down to the center of the earth.

Dig a hole and bury it. Give the command: “Don’t come back. You will be happy down there. You will be transformed into good Chi.” Smile. Always smile.

Repeat it six, nine or eighteen times. For people who are very sick, who have cancer or may be terminally ill, you can do this two or three hundred times.

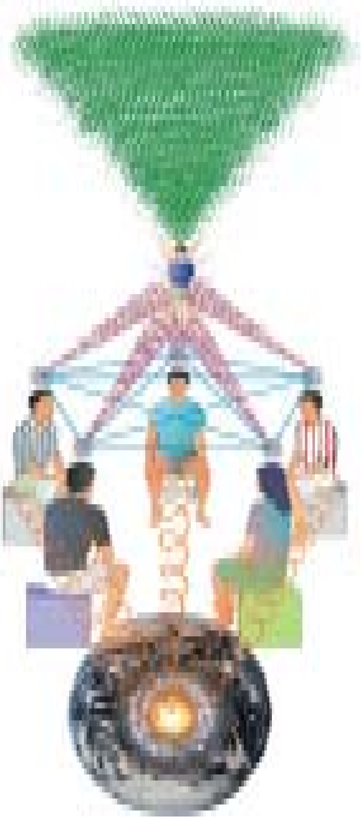

Fig. 4.44 *Green Light bonds with the Sickness and Toxins before returning them to the Earth.*

I. Clean with Blue Light

Be aware the blue light, kidney blue. Breathe the blue light into the throat center. Now slowly blend the blue light in the Tan Tien. Blend the blue light in the heart, in the crown and project it up to the universe, spiraling clockwise. Then feel the universe pouring the blue light back down; counterclockwise. Spiral down. Look at the cells and see the blue light, like water, go in to the cells to rinse and clean all the dirty things, down, down, down to the earth. All the way down. Repeat three, six or nine times. Smile.

Blue:

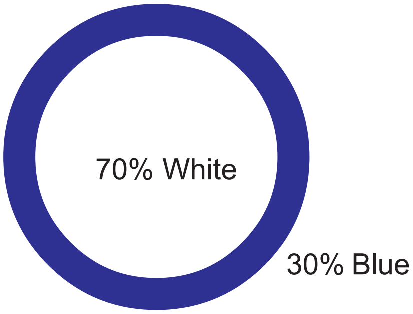

Fig. 4.45 *Rinse with Blue.*

Blue is like cold water; it has the yin power to dissolve all kinds of negative energy and sickness. It has an inhibiting effect. Blue yin energy is the opposite of red yang, which has a strengthening and stimulating effect. Blue has a cooling effect; it can reduce pain, inflammation and can help blood to clot. It stops bleeding and reduces fever. It can help to induce rest and sleep.

Throat Center Blue

Be aware of the kidneys' blue color and breathe the blue light into your throat. Guide it down to the Tan Tien, blend and move back up to the 'heart of compassion' and then the crown.

Project up to the universe, spiral and let it multiply.

Channel one direct line down to the student, one to your crown and one to your palm.

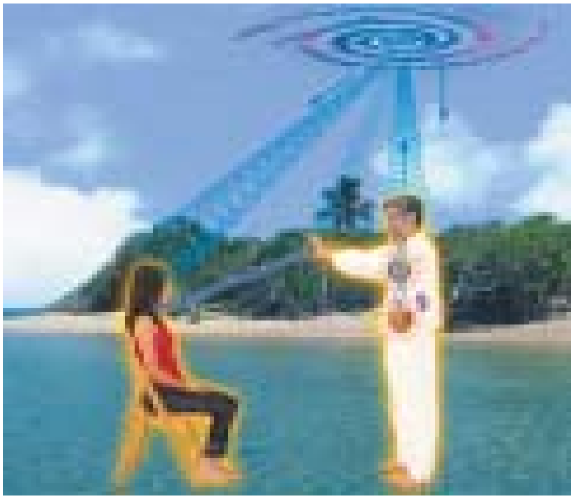

Fig. 4.46 *Multiply the Universal Blue Light - guide the Energy from the Throat, Heart and Crown into the Universe and back down.*

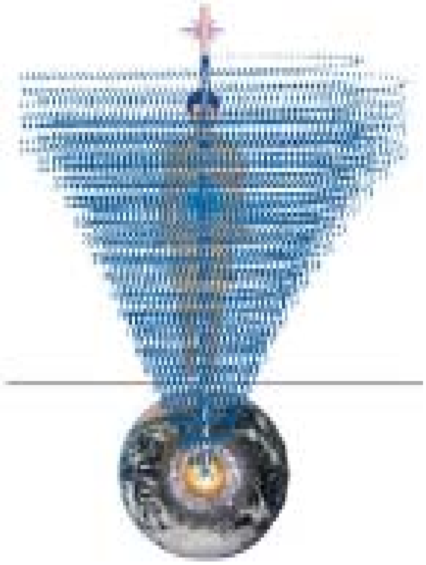

Be aware of the kidneys, the bladder, the oceans and the throat breathing.

Be aware of the blue light in the universe. Let the light come in and blend in the Tan Tien, up to the heart to blend with the compassion and up to the crown.

Project the light up to the universe. See it spiral and let it multiply.

Fig. 4.47 *Draw the Blue Light from the Universe - rinse clean and remove sickness.*

Rinse and Flush

Extend your arms in the air, with palms facing heaven. Use one hand to spiral the green force down. The other hand holds the position and connects to the universe. Let a big pool of green light in the universe spiral down – through your community, your home, and then into your crown.

Let the blue light flush into the cells and flush out sickness and toxins towards the center of the earth.

Dig a hole and bury them. Give the command: "Don't come back. You will be happy down there. You will be transformed into good Chi."

Repeat it six, nine or eighteen times. For people who are very sick, who have cancer or who may be terminally ill you can do this up to thirty six, seventy two or one hundred and eight times.

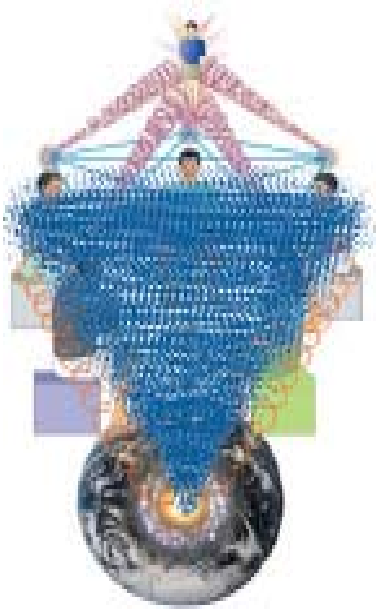

Fig. 4.48 *Blue Light flushes out the sickness and toxins before returning them to the earth.*

J. Charge with Violet Light

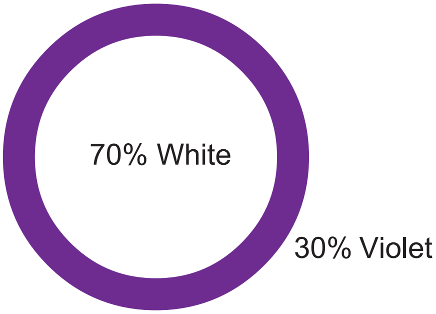

Fig.4.49 *Intelligent Luminous Violet*

Luminous Violet has intelligence and can be programmed. The universe is full of violet light, especially the North Star and the Big Dipper. The color of the star of the higher self or soul, above the crown, is also violet light, the Divine or Soul energy.

It has properties of other Chi colors. It has a rapid regenerating effect on damaged organs and nerves. Always use green and blue light before using violet. Violet light develops the crown center, the spiritual core. It is good for psychological ailments as well as for physical deficiencies.

Primordial Force

Turning the three minds into one mind, go deep into the empty space of the cell and deep into the chromosomes, magnify the space as you enter into the DNA and return to the original cells. The empty space is the Wu Chi. When you enter this stage you turn the subconsciousness into consciousness and you can talk to your body and cells. You can change the programming of DNA and RNA.

Ask for the power to see the cells and look at the brain cells spread out into the space. "Brain cells please listen. Clear, Clean, and Bright: fill with violet light and return to normal function".

Violet Light into the Cells

See violet light. Be aware of your crown. The Tan Tien Chi and the negative Chi all combine in the crown. Extend a violet light up to the universe. See the North Star in the Big Dipper. Tell the student to hold their arms in the air, so that they can funnel the Chi into their personal star. Bring the North Star and the Big Dipper down; hold the handle of the Big Dipper and pour the violet light over the student's personal star. The higher cells will be programmed. Let the Chi enter the brain and ask for the power to see the brain cells. "Brain cells listen to the command". If there is sickness in the brain, give the command. "Sickness go away".

The North Star is a major source of violet light.

The Big Dipper is a major source of red and infrared light. Taoists understand that the cup of the Big Dipper gathers all the violet light from the universe. The left hand holds the handle of the big dipper - pour it down to your crown. It then flows down throughout the body.

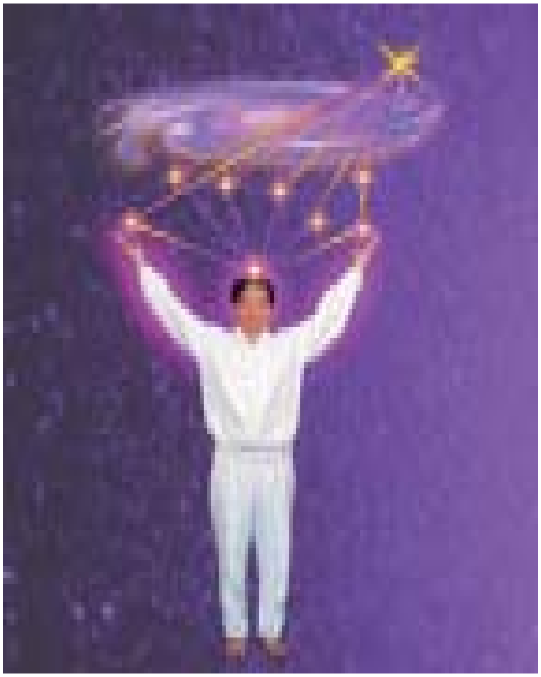

Fig. 4.50 *Connect with the North Star to receive the Violet Light.*

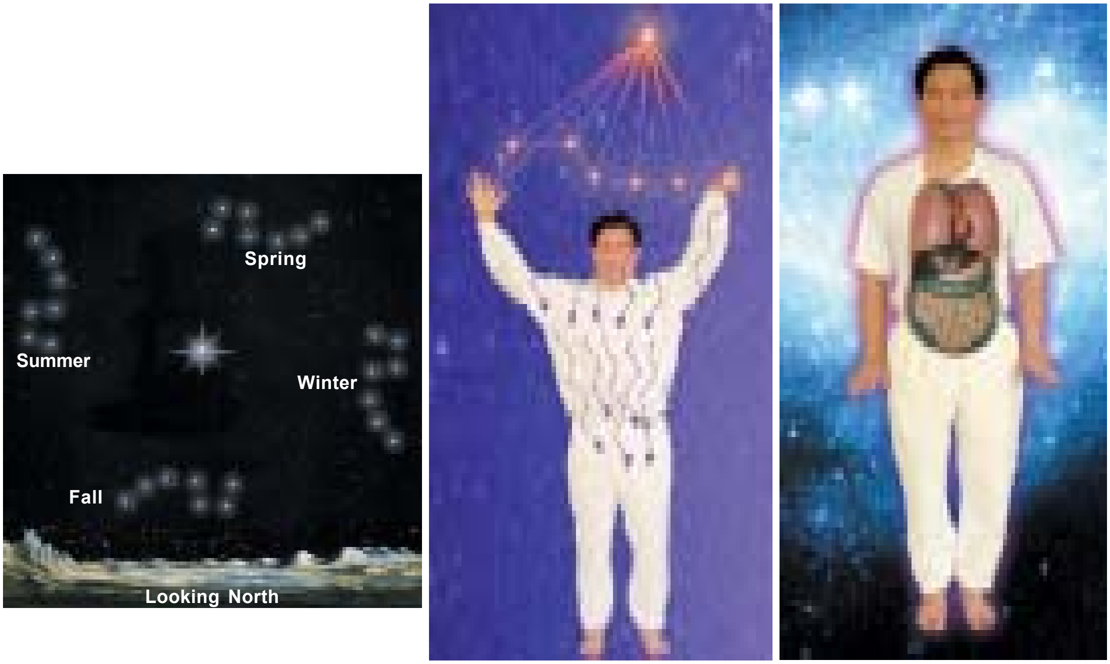

Fig. 4.51 *Practitioner connects with Universal Violet Light.*

Bring the violet light down to the student's personal star above their crown. Pause to let the higher cells of the brain reprogram the violet light at the personal star. Continue to spiral the violet light down through the cells of the brain and down through the cells of the whole body.

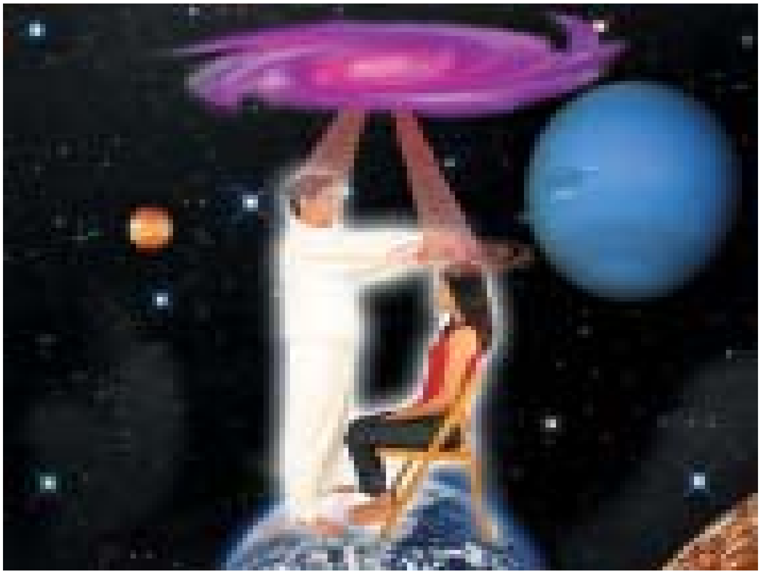

Fig. 4.52 *Pause to reprogram the Violet Light at the other Person's Personal Star.*

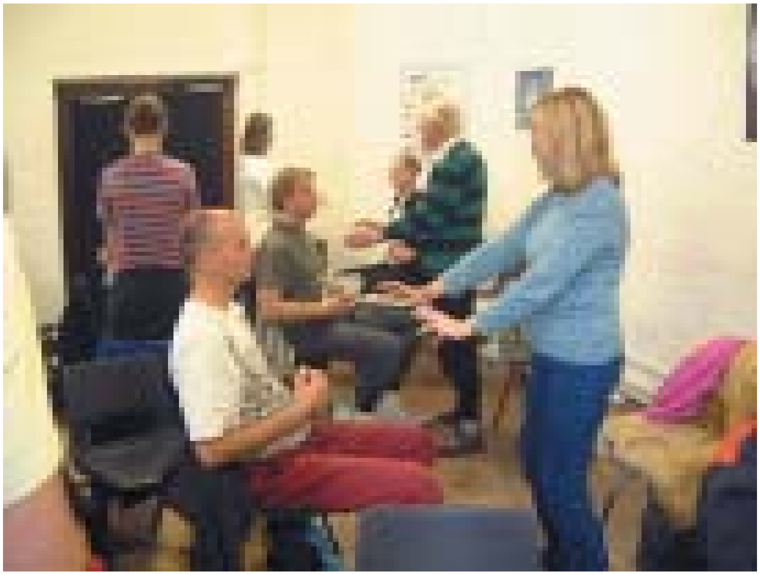

Fig 4.53 *Lead the person and guide them to follow you.*
Tell them to focus on the part of their body that you are working on.

1. Ask for the power to see into the cells of the brain. Ask the student to move their hands to cover their brain. “Brain cells listen to the command: Sickness go away. Clear, clean and bright; fill with violet light and return to normal function”.

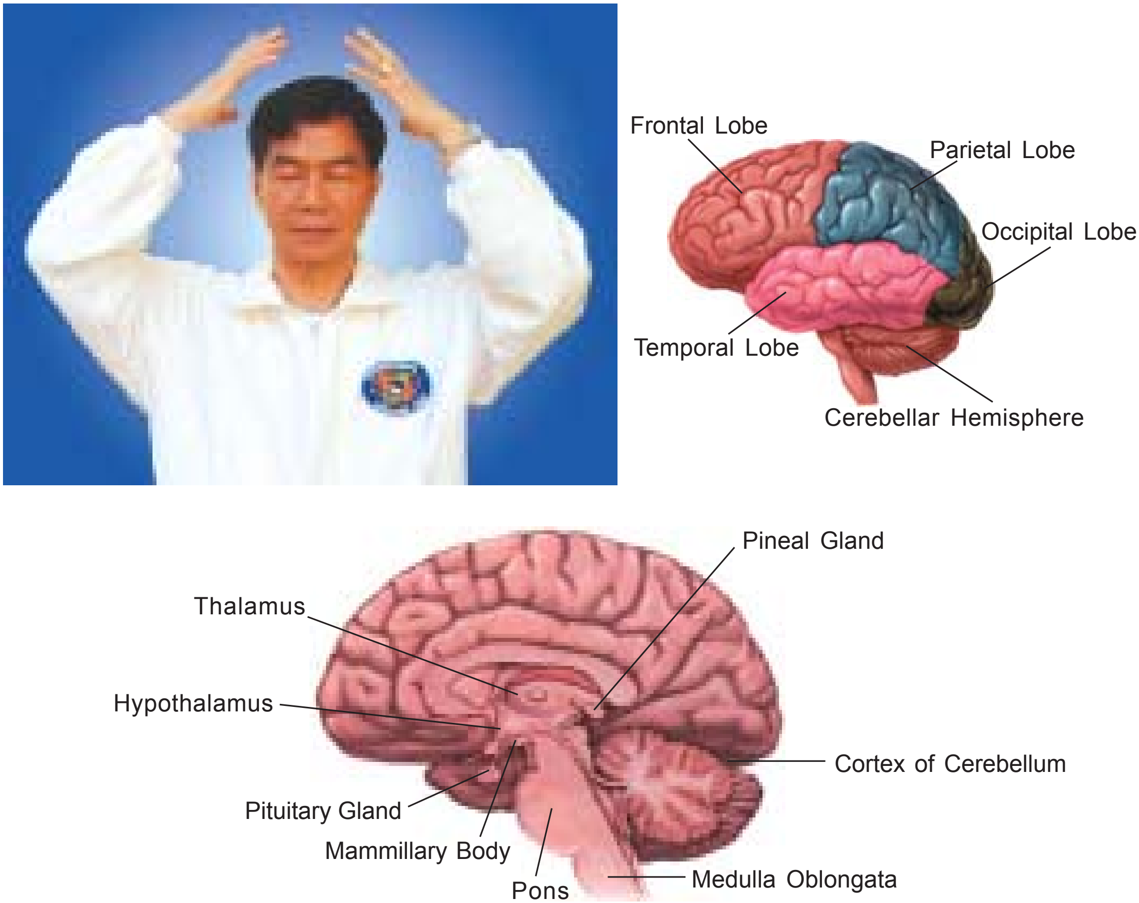
Frontal Lobe
Parietal Lobe
Occipital Lobe
Temporal Lobe
Cerebellar Hemisphere
Pineal Gland
Thalamus
Hypothalamus
Cortex of Cerebellum
Pituitary Gland
Mammillary Body
Pons
Medulla Oblongata

Fig. 4.54 *“Clear, clean and bright; fill with violet light and return to normal functions”.*

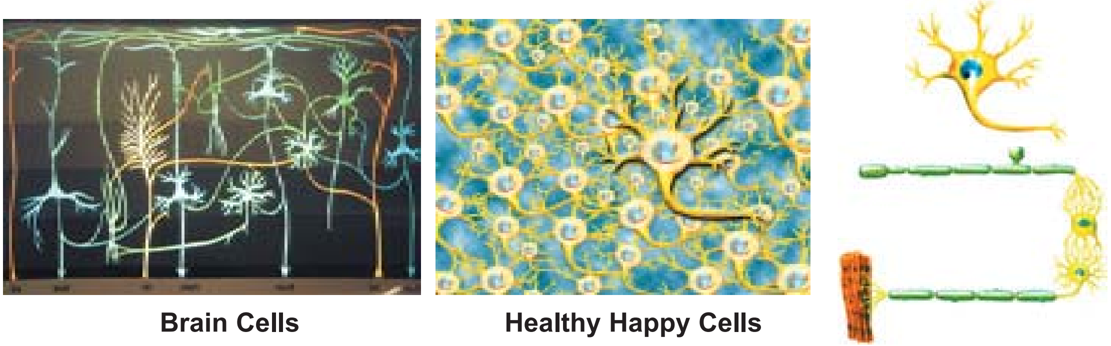

Brain Cells

Healthy Happy Cells

Fig. 4.55 *‘Intent’ to Heal.*

2. Ask for the power to see the cells of the sense organs. Ask the student to move their hand to cover their senses. Give the command: "Eyes, ears, nose, tongue and mouth cells listen to the command. Sickness go away. Clear, clean and bright; fill with violet light and return to normal function".

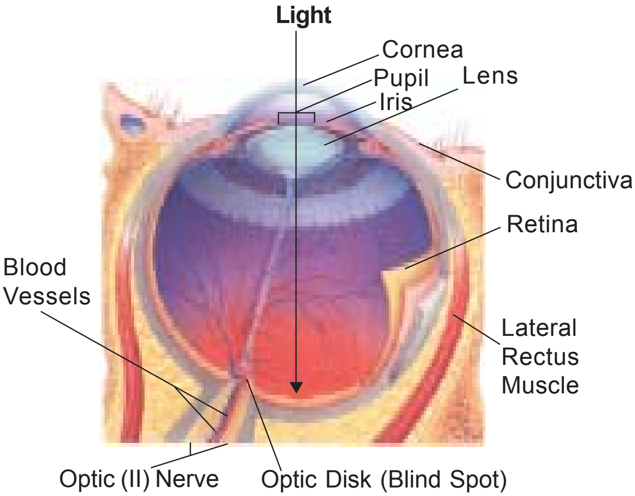
Light
Cornea
Pupil
Lens
Iris
Conjunctiva
Retina
Blood Vessels
Lateral Rectus Muscle
Optic (II) Nerve
Optic Disk (Blind Spot)

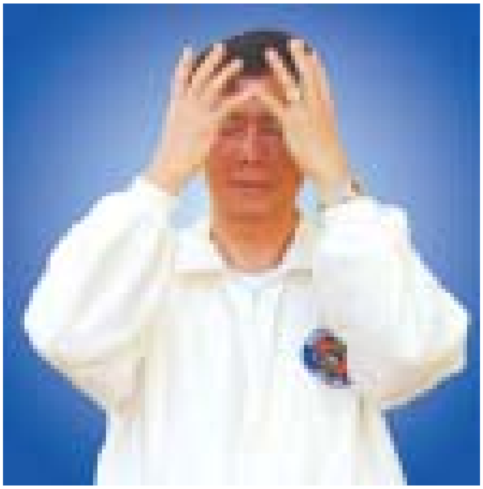

Fig. 4.56 *"Eyes, listen; clear, clean and bright; fill with Violet Light and return to normal function".*

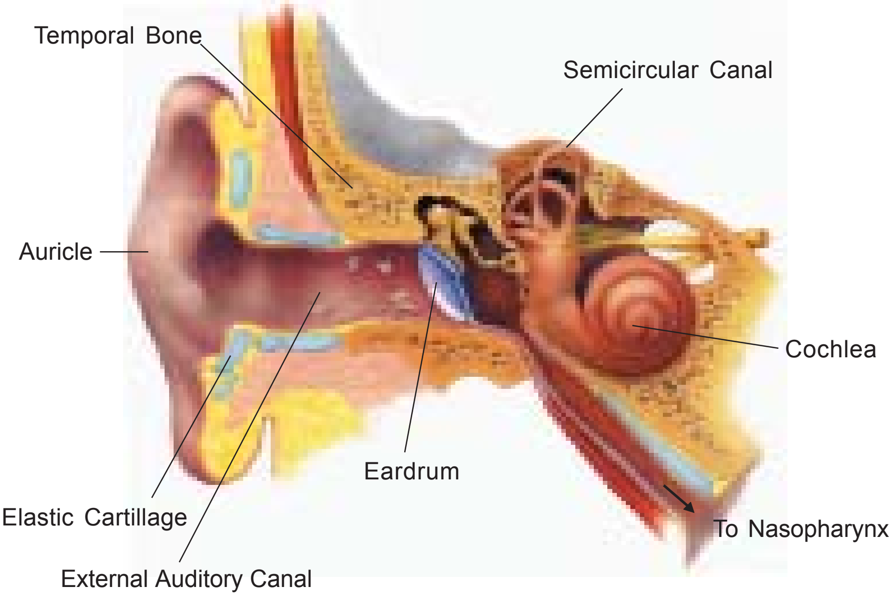
Temporal Bone
Semicircular Canal
Auricle
Cochlea
Eardrum
Elastic Cartillage
To Nasopharynx
External Auditory Canal

Fig. 4.57 *"Ears, listen; clear, clean and bright; fill with Violet Light and return to normal function".*

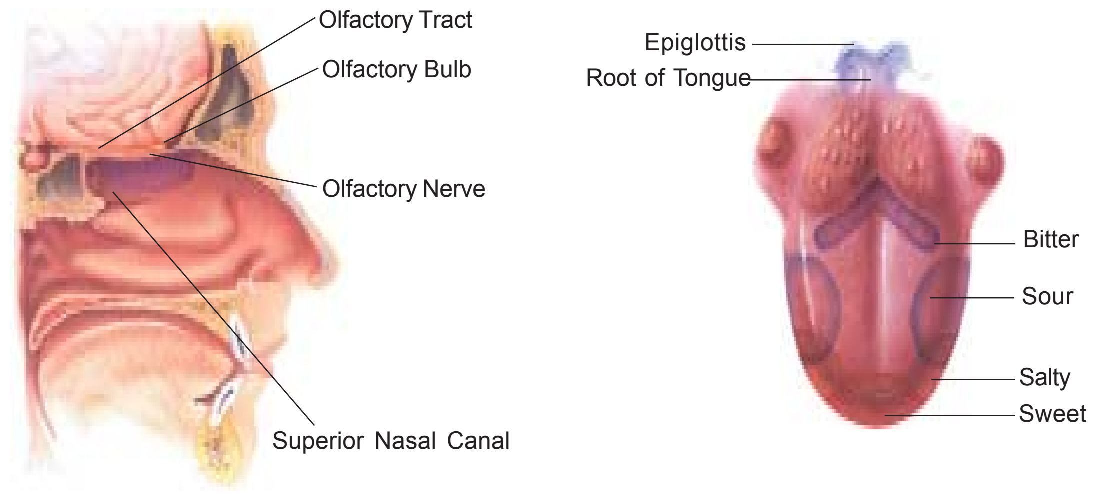
Olfactory Tract
Olfactory Bulb
Olfactory Nerve
Superior Nasal Canal
Epiglottis
Root of Tongue
Bitter
Sour
Salty
Sweet

Nose

Tongue

Fig. 4.58

"Nose, tongue, listen; clear, clean and bright; fill with violet light and return to normal function"

3. Move your hands down to your throat center and give the command: "Thyroid, parathyroid and Thymus glands, listen to the command; Clear, clean and bright; fill with violet light and return to normal function".

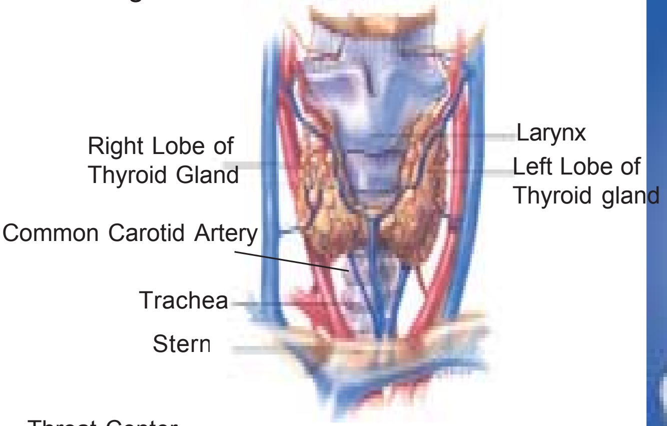
Right Lobe of
Thyroid Gland
Larynx
Left Lobe of
Thyroid gland
Common Carotid Artery
Trachea
Stern

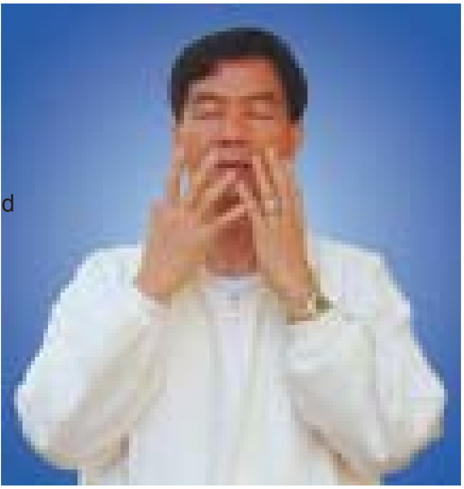

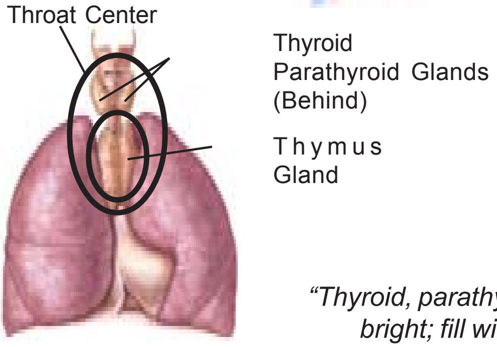
Throat Center
Thyroid
Parathyroid Glands
(Behind)
Thymus
Gland

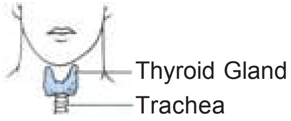
Thyroid Gland
Trachea

Fig. 4.59

"Thyroid, parathyroid and thymus glands - listen; clear, clean and bright; fill with violet light and return to normal function".

4. Move your hands to the heart and the lungs. Ask for the power to see the cells of the heart and lungs. Give them the command: “Lungs and heart cells, listen to the command”. If there is no sickness, just clean them out. See the cells and clean them out with violet light. If there is sickness, tell the sickness to “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

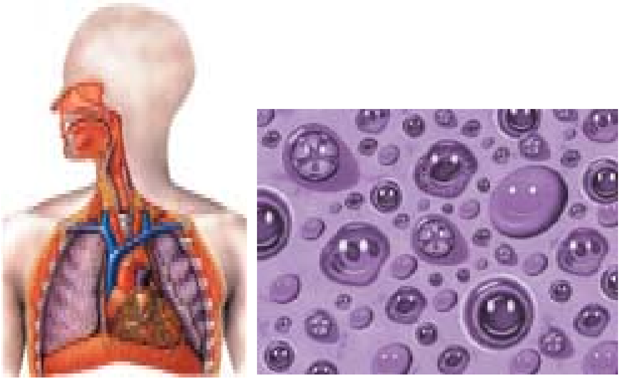

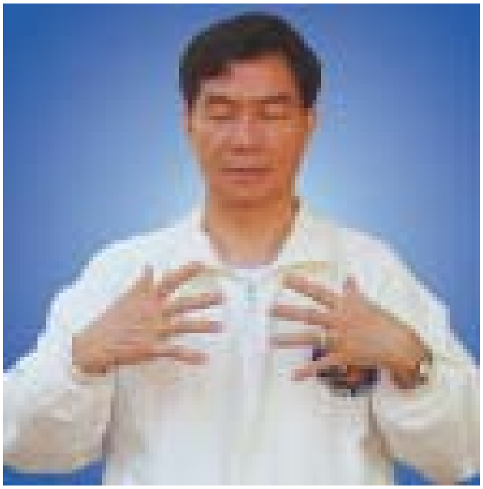

Fig. 4.60

“Heart and lung cells please listen; clear, clean & bright; fill with violet light and return to normal function”.

5. Move your hands to cover the left and right sides of the ribcage. Ask for the power to see the cells of the liver, gallbladder, spleen, pancreas and stomach. Give them the command: “Liver, gallbladder, spleen, pancreas and stomach cells, listen to the command”. If there is no sickness, just clean them out. If there is sickness, tell the sickness to, “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

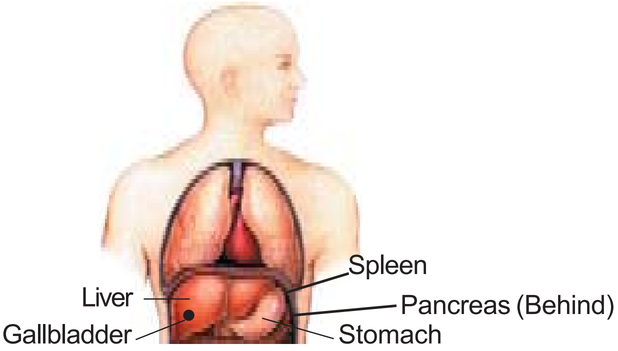

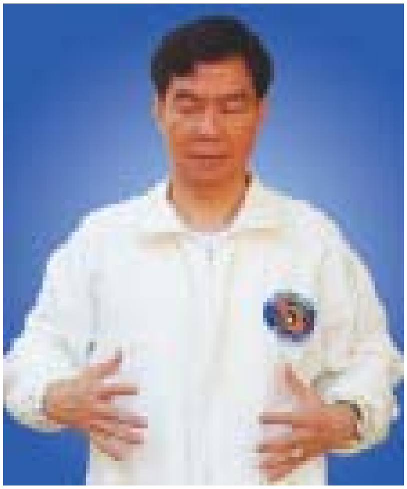

Fig. 4.61

“Liver, gallbladder, pancreas and spleen cells please listen; clear, clean and bright: fill with violet light and return to normal function”.

6. Move your hands to the large and small intestine on the front of the abdomen. Ask for the power to see the cells of the intestines. Give them the command: “Large and small intestine, listen to the command”. If there is no sickness, just clean them out. See the cells and clean them out with violet light. If there is sickness, tell the sickness to, “go away - be clear, clean and bright; fill with violet light and return to normal function”. Make sure that you say this part aloud. Give the command in a very firm tone of voice.

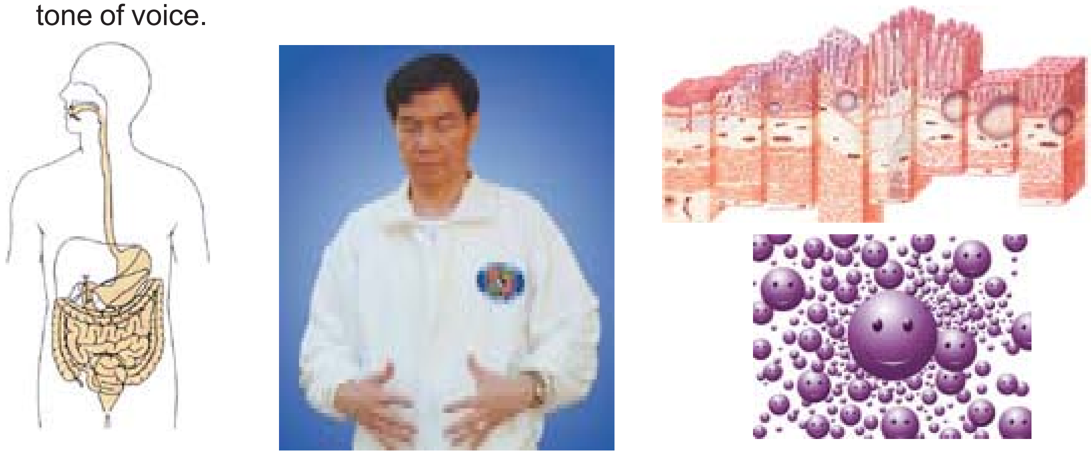

Intestines **Healthy cells**

Fig. 4.62 *“Small and large intestine cells listen to the command: clear, clean and bright; fill with violet light and return to normal function”.*

7. Move your hands to the kidney and sexual organs and ask for the power to see their cells. Give the command: “Kidney and sexual organ cells listen to the command: Clear, clean and bright; fill with violet light and return to normal function”. Gather the violet light again and let it pour down and fill all the cells. See the violet light again and feel and see that all the cells are, “clear, clean and bright”. Smile.

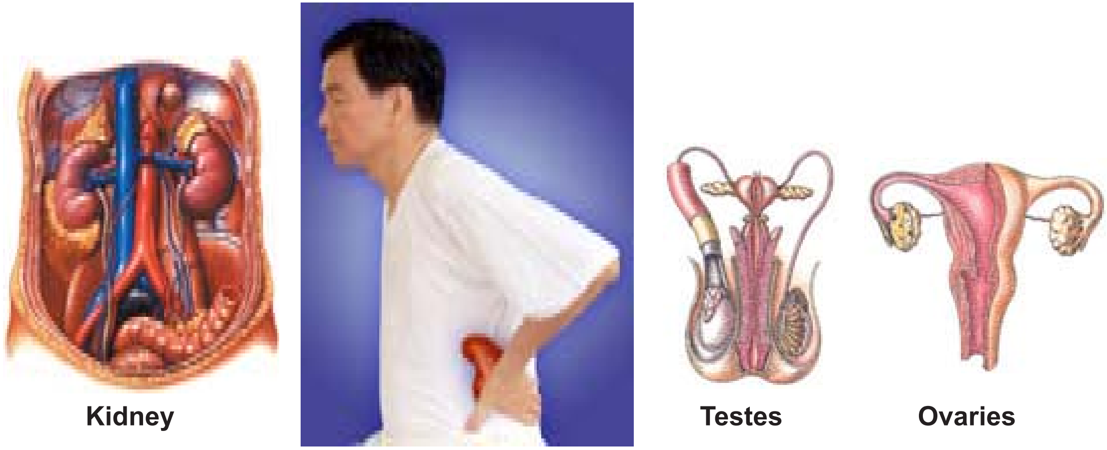

Kidney **Testes** **Ovaries**

Fig. 4.63 *“Kidneys and Sexual Organs listen”.*

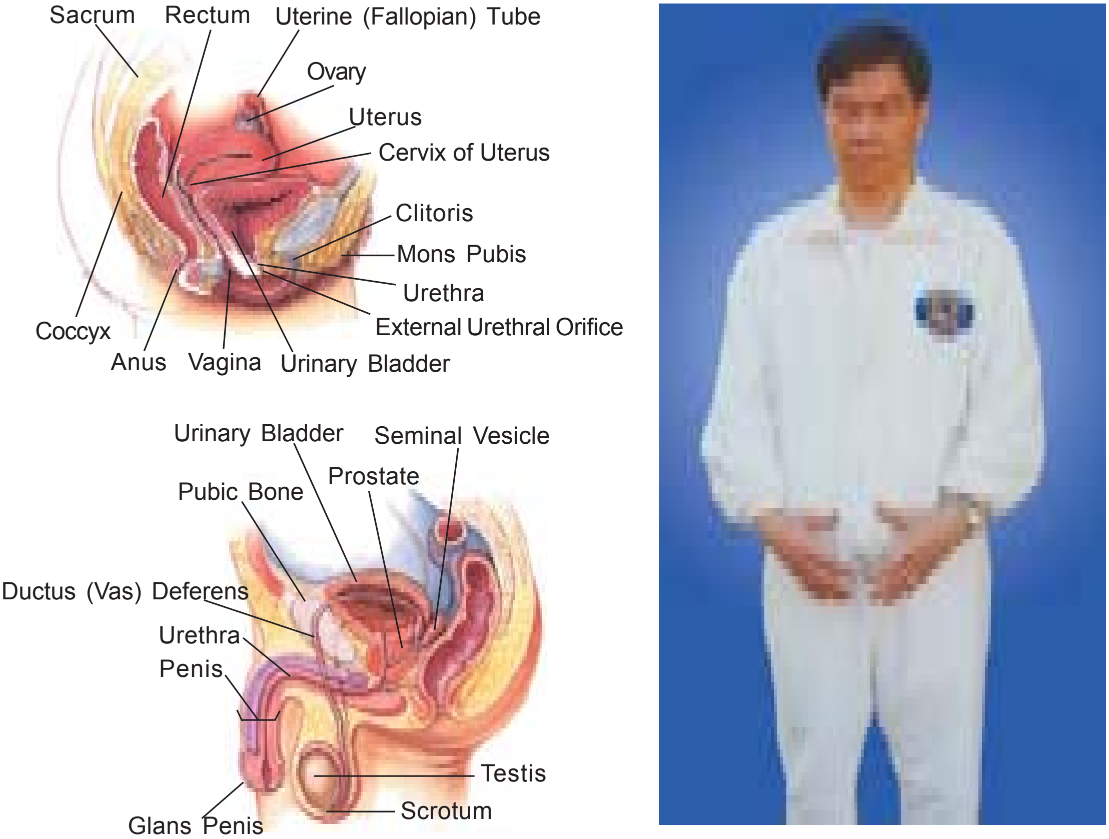
Sacrum
Rectum
Uterine (Fallopian) Tube
Ovary
Uterus
Cervix of Uterus
Clitoris
Mons Pubis
Urethra
External Urethral Orifice
Coccyx
Anus
Vagina
Urinary Bladder
Urinary Bladder
Seminal Vesicle
Prostate
Pubic Bone
Ductus (Vas) Deferens
Urethra
Penis
Testis
Scrotum
Glans Penis

__Fig. 4.64__ Male and Female Sexual Organs

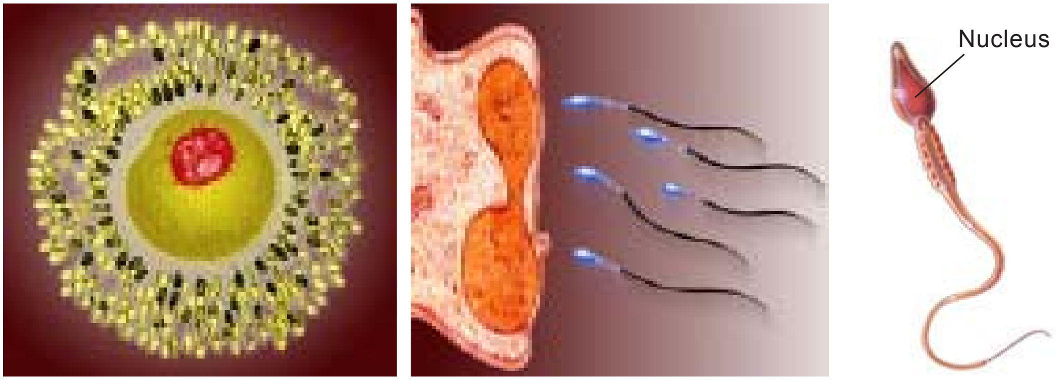

Egg

Sperm

__Fig. 4.65__ "Clear, clean and bright: fill with Violet Light and return to normal function".

K. Activate the Defense and Immune System

The Defense System relates to the white blood cell and the immune system relates to the Lymphatic System.

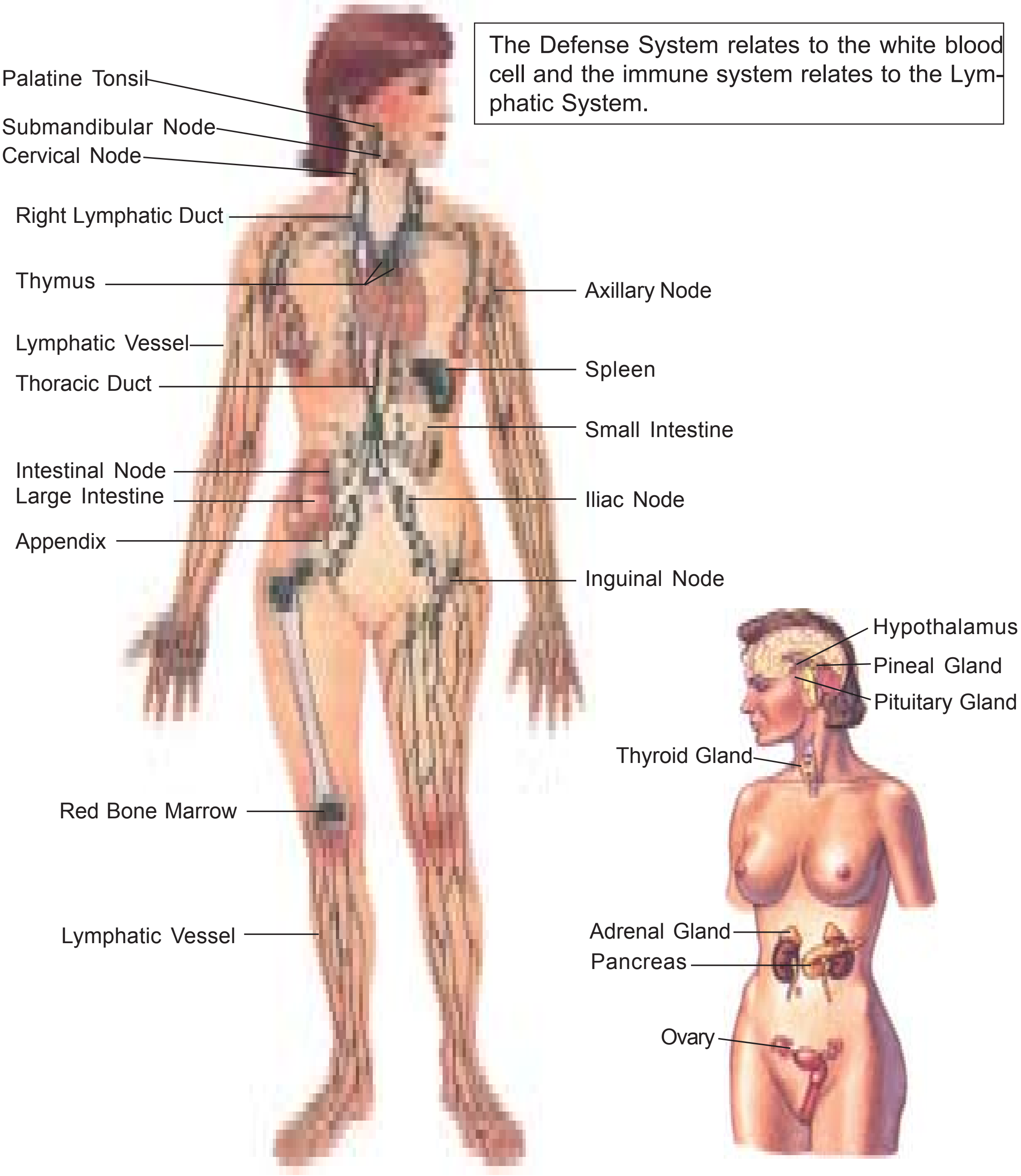
Palatine Tonsil
Submandibular Node
Cervical Node
Right Lymphatic Duct
Thymus
Lymphatic Vessel
Thoracic Duct
Intestinal Node
Large Intestine
Appendix
Red Bone Marrow
Lymphatic Vessel
Axillary Node
Spleen
Small Intestine
Iliac Node
Inguinal Node
Hypothalamus
Pineal Gland
Pituitary Gland
Thyroid Gland
Adrenal Gland
Pancreas
Ovary

Fig. 4.66 *Ask the cells to listen to the command.*
The bone marrow produces red and white blood cells. The lymphatic system cleans the toxic and polluted material and bacteria from the cells.

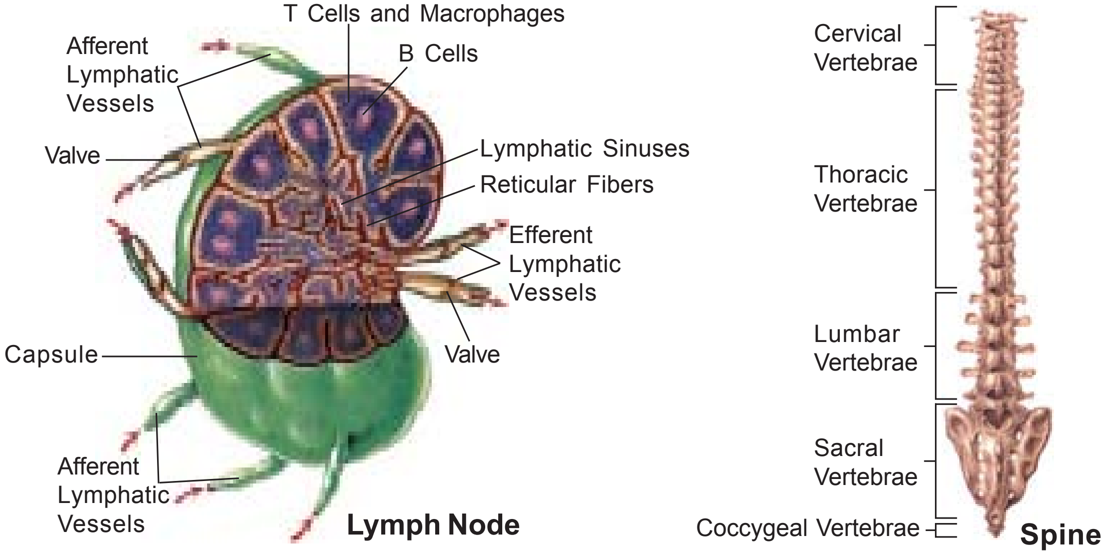
T Cells and Macrophages
Afferent Lymphatic Vessels
B Cells
Valve
Lymphatic Sinuses
Reticular Fibers
Efferent Lymphatic Vessels
Capsule
Valve
Afferent Lymphatic Vessels
Lymph Node
Cervical Vertebrae
Thoracic Vertebrae
Lumbar Vertebrae
Sacral Vertebrae
Coccygeal Vertebrae
Spine

***Fig. 4.67** “Immune Cells and Vertebrae please listen: clear, clean and bright; fill with Violet Light and return to normal function”.*

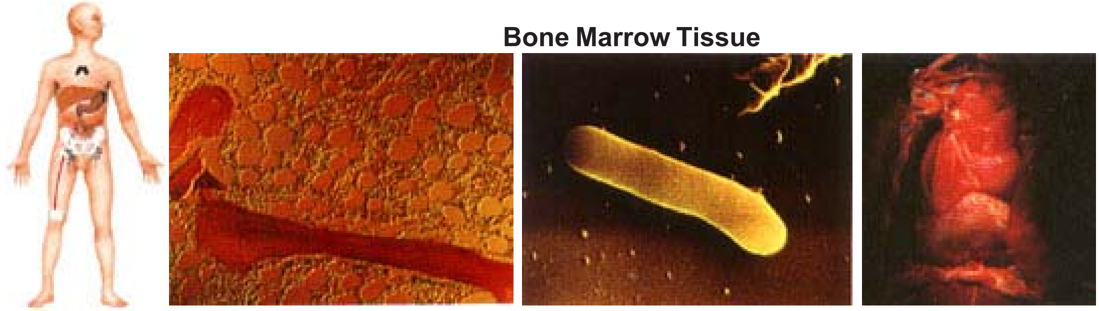
Bone Marrow Tissue

Sacrum: The sacrum controls all the bones and marrow. It effects the production of red and white blood cells. The bone marrow produces the lymphatic cells (white blood cells). When the blood is full of Chi it becomes lighter. Negative emotions make the blood thick and acidic. The blood cells in a healthy body can live up to ten times longer than in a weak or diseased one.

Bone Marrow: A tissue rich in fats, containing plenty of nourishment for the comprehensive production of the different blood cells: red cells for the transport of oxygen, blood platelets for clotting and the various white cells for the immune system.

Bacterium: Represents the foreign invader, but it is only one of many invaders. Parasites, fungi, viruses, chemicals, mineral fragments, metal particles—these and a great deal besides, set the immune system to work.

Thymus Gland: Perhaps the most important organ of the immune system. The vital training of the different T-lymphocytes takes place inside. When they pass through the thymus they receive a program to convert them into ‘T’ lymph cells. When they pass through the liver and large intestine they become ‘B’ cells and helper cells.

We have obtained Fig. 4.68 and the following text from the publication 'The Body Victorious' by Lennart Nilsson and Jan Lindberg.

We hope that you find this short selection of material as informative as we did.

"The illustration below is a highly schematic and simplified picture of our immune system. *Top left*, the pale yellow 'house' symbolizes the bone marrow, where all blood cells are born. *Bottom Right*, a solitary bacterium (yellow) represents the foreign invader that the immune system needs to combat. The light blue path represents the older defenses, with various feeding cells that tackle all foreign substances and also function as a kind of cleaning system. They deal with all the old, dead and used-up material in the body. The newer, special defense force, which develops later, follows the green path. Its cells are more specialized; they obtain their special training and reach maturity in organs such as the thymus gland and in the lymphoid tissue around the intestines and in the liver. In the lymphoid tissue, the yellow 'half-way house' (far right), the 'B-lymphocytes are trained. These are the precursors of the large plasma cells - the pale green cells (far right) in the illustration, which produce the body's sniper ammunition, the antibodies (Red and Y-Shaped).

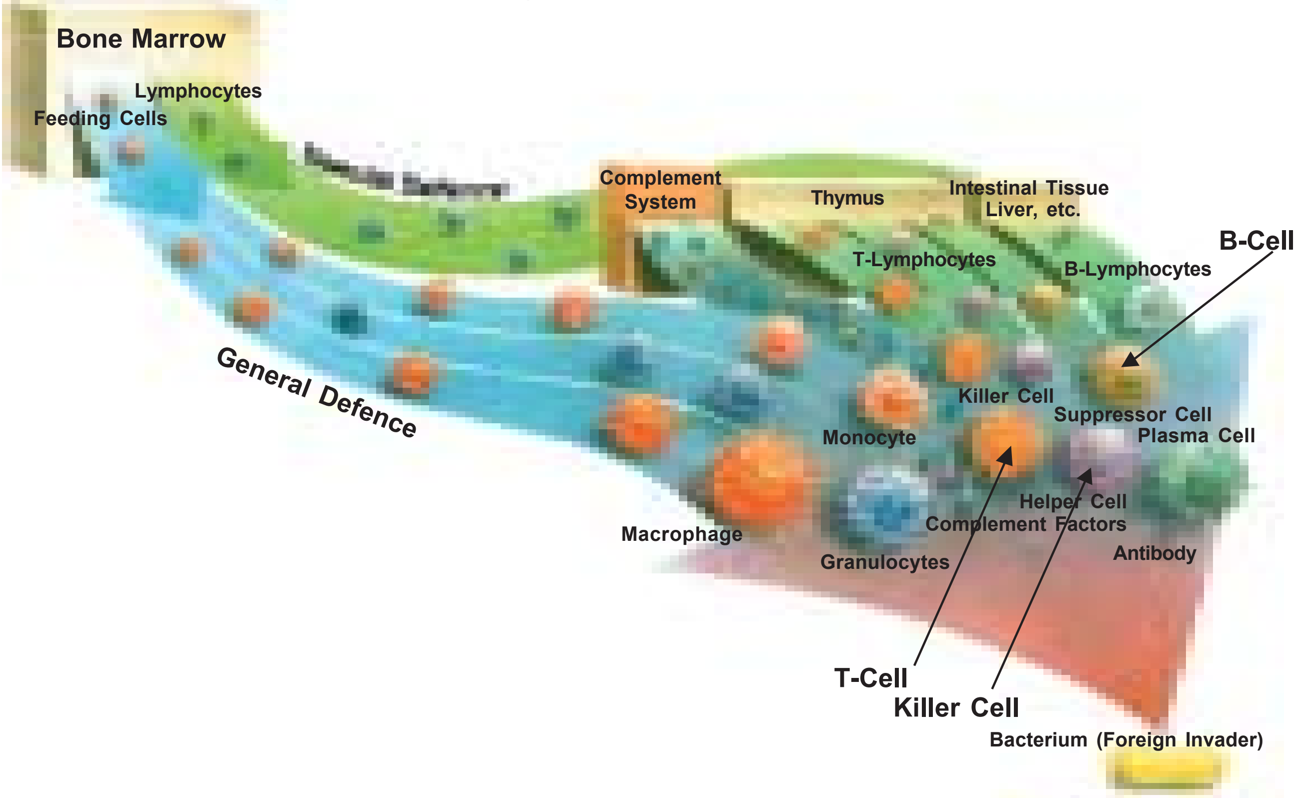
Bone Marrow
Lymphocytes
Feeding Cells
Complement System
Thymus
Intestinal Tissue Liver, etc.
B-Cell
T-Lymphocytes
B-Lymphocytes
General Defence
Killer Cell
Suppressor Cell
Monocyte
Plasma Cell
Helper Cell
Complement Factors
Macrophage
Antibody
Granulocytes
T-Cell
Killer Cell
Bacterium (Foreign Invader)

Fig. 4.68

Bone marrow, bacterium and thymus please listen: clear, clean and bright; fill with violet light and return to normal function".

The three paths from the thymus are intended to show that there are different types of T- lymphocytes - among others, aggressive killer cells, helper cells and suppressive cells. All of them have specialized tasks to perform when the immune system launches a counterattack. The three blue paths of the older defenses have three different kinds of feeding cells. First, the large and powerful macrophages (orange) advance on the enemy; then come the granulocytes (blue), smaller and faster moving, followed by the (pink) monocytes. The nuclei in the shimmering cell bodies are faintly discernable. In addition, there is an important complement system (middle), symbolized here by a multicolored range of small spheres flocking toward the foreign bacterium. These molecules play a large part in increasing the efficiency of both antibodies and feeding cells. In addition, they have the capacity to destroy bacteria by shooting holes in them. The complement factors are produced in many different cells in the body.

Chi Kung has a very strong effect against viruses. Altough medical science is struggling to discover a medicine that will kill them, Chi Kung practice is able to prevent them from occuring in the first place. Whether we are sick or healthy, we all need our defense system to be in good shape.

Defense System - Outline of Practice

1. Work on yourself first to activate your sacrum. Project the sacrum as big as the universe. See the eight holes in the bone and visualize them breathing. Breathe and compress the Universal Chi into the bones. Feel the Chi begin to rise up the spine and into your mideyebrow and temple bones. When you feel that there is enough Chi, project it out to the universe, multiply it and spiral it back down. Project it towards the student and ask them to touch their sacrum. Rub your hands until warm and touch the sacrum. Picture the sacrum getting bigger and bigger, breathing into the eight holes. Tell the person to smile at the sacrum and start to work on your Tan Tien and universe. Multiply more energy and picture the Chi from the universe coming down to the sacrum of the other person (and your own) and then charge up the spine, to the rib cage, to the temple bone and to the mideyebrow.

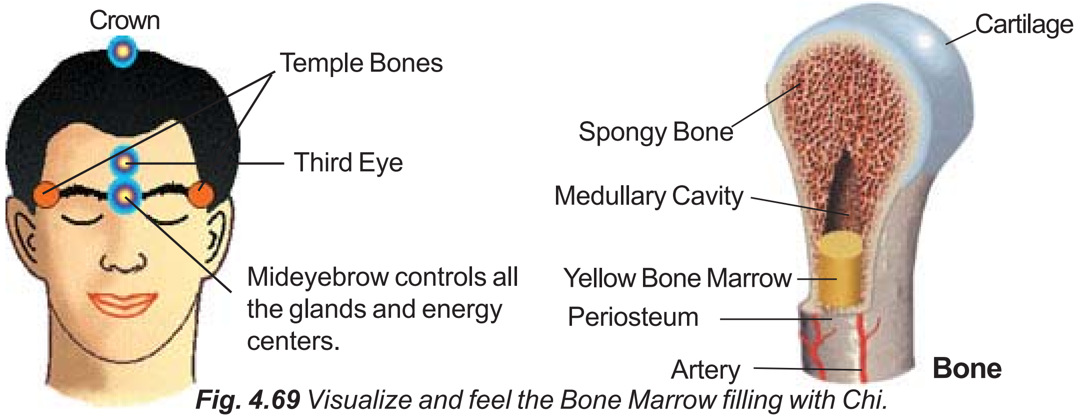

Fig. 4.69 *Visualize and feel the Bone Marrow filling with Chi.*

2. Tell the student to move their hands to the hip bones. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow.
3. Touch the femur bones. Feel the electricity running up the legs.
4. Touch the humerus bones. Feel the electricity running up the arms.
5. Touch the lower part of the sternum and feel the Chi penetrate into the bones and spread out to the rib cage. Feel the electricity run throughout the rib cage.

Practice

1. Touch the sacrum, feel your fingers grown 'long' and penetrate into the marrow. Feel that the sacrum is as big as the universe and that the Chi has been compacted. Fill the sacrum with Chi. The eight holes are breathing. Let the fingers touch the sacrum and become aware of the Tan Tien spiraling, until you feel the crown and the mideyebrow spiral and connect to the universe. Hold until the universe fills the sacrum once again and rises up the spine to the mideyebrow.

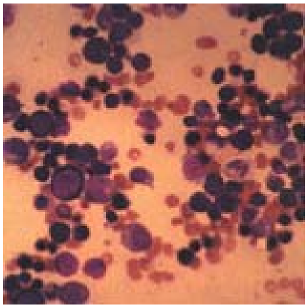

Bone Marrow Cells

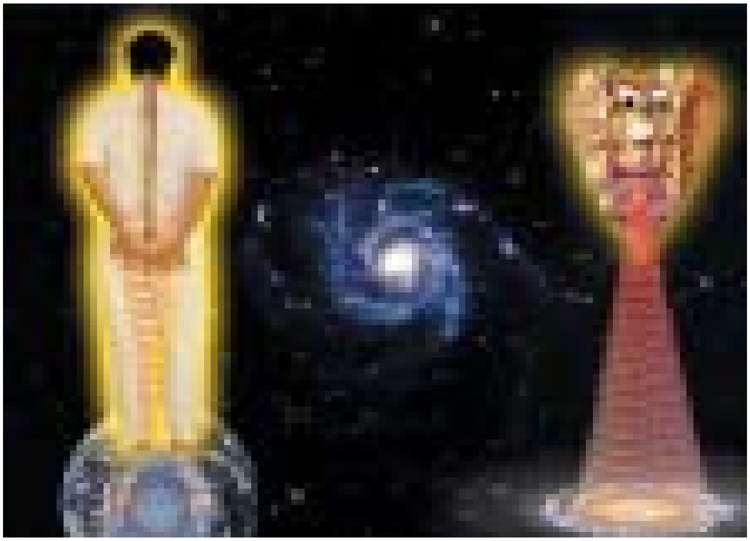

Fig. 4.70 *Touch the Sacrum - see the Eight Holes breathing - fill the Sacrum, Mideyebrow and Crown with Chi - empty to the universe and 'fill' again.*

Sacrum Breathing to get Chi

Breathe through the sacrum; feel red and yellow light enter. Guide it into Tan Tien. Blend the Chi in the Tan Tien. Move up to the Heart Center. Move up to the Crown and project up to the universe; let it blend and multiply. Multiply in the universe

Open a direct channel down to the other person, and one to your palm.

Sacrum Chi

Project the student's feet down to the earth and connect to the earth's core.

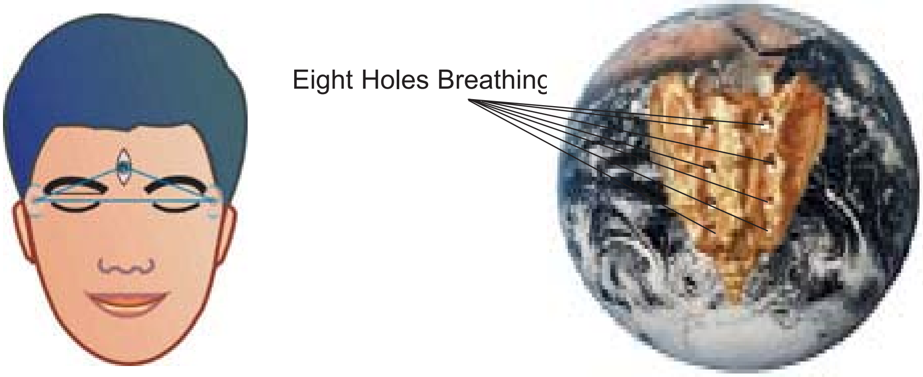

Third Eye and Temple Bones are aligned in a triangle.

Activate the Sacrum until you feel the Chi rising up the spine and feel it pulsing; be aware of the Temple Bones.

Fig. 4.71 *Activate the Sacrum - Be aware of the Temples.*

2. Tell the student to move their hands to the hip bones. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow. Tell the student to pack and compress the Chi into the bones in order to revitalize and revive bone marrow. Give the command. "Produce healthy white and red blood cells".

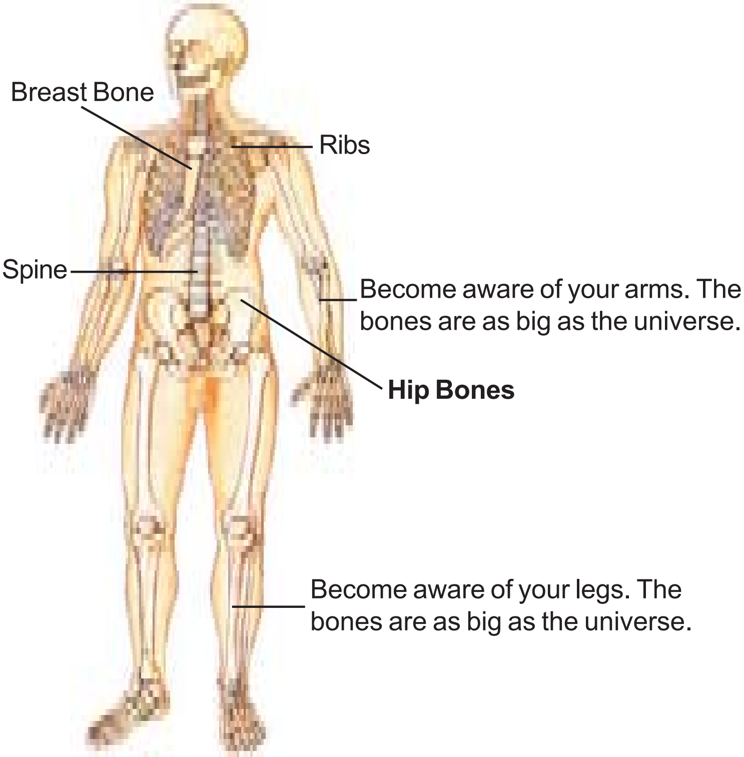

Become aware of your arms. The bones are as big as the universe.

Become aware of your legs. The bones are as big as the universe.

Fig. 4.72 *"Hip Bones produce healthy White and Red Blood Cells".*

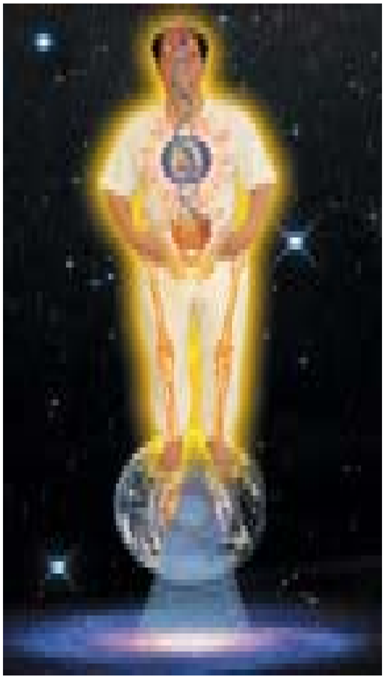

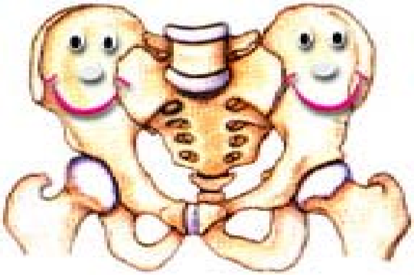

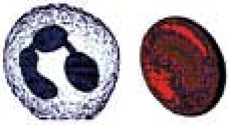

White Blood Cell
(Neutrophil)

Leave the fingers touching the hip bones and start to spiral in the Tan Tien until you feel the crown and the mideyebrow also spiral. Expand your awareness to the Universe and the Universe will charge into the hip bones.

Fig. 4.73 *Touch and feel the bones. Leave your fingers there - lower your mind into the Tan Tien and the universe.*

3. Touch the middle part of the femur bone to help increase the production of healthy red blood cells. Feel the electricity running up the legs. Give the same command for the femur bones: “Funny, laughing, happy bones produce healthy white and red cells.” After that put your mind into the Tan Tien and universe so that the Chi can charge into the bones. Feel electricity run throughout the whole skeleton.

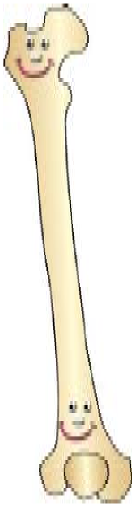

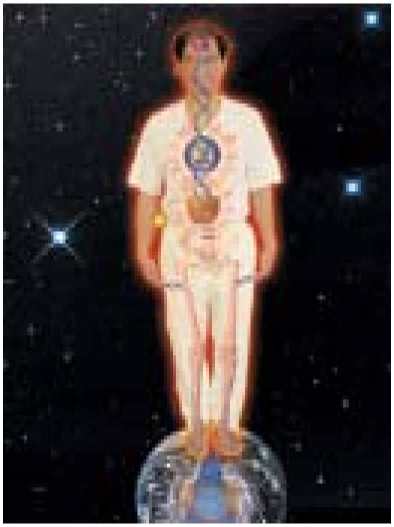

Fig. 4.74 *“Funny, laughing, happy femur bones produce healthy white and red blood cells”.*

4. Touch the humerus bones. Feel the electricity running up the arms.Give the same command: “Funny, laughing, happy bones produce healthy white and red cells.” After that, put your mind into the Tan Tien and universe so that the Chi can charge into the bones. Give the same command for the lower part of the sternum. Chi penetrates into the bone and spreads out into the rib cage ‘Tan Tien and Universe’. Put your mind into the Tan Tien and the universe so that the Chi can charge into the bones.

5. Touch the lower part of the sternum and feel it and the rib cage as funny, happy, laughing bones. Feel the Chi and electricity penetrate into the bones and spread out to the rib cage. Touch the lower part of the sternum. Give the same command again. "Produce healthy white and red cells". After that you put your mind into the Tan Tien and universe so that the Chi can charge into the bones.

   You have finished activating the defense system.

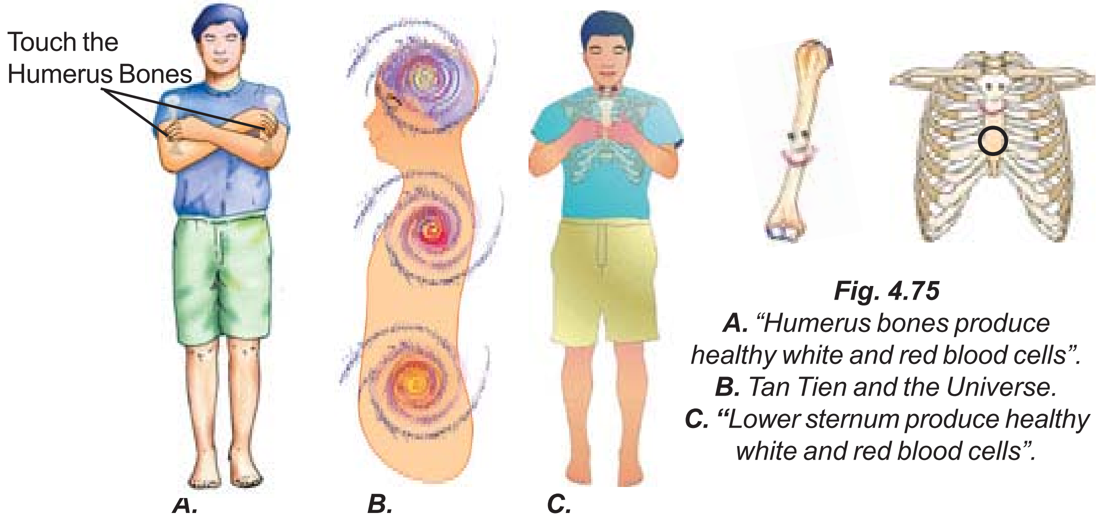

Fig. 4.75
A. *"Humerus bones produce healthy white and red blood cells".*
B. *Tan Tien and the Universe.*
C. "*Lower sternum produce healthy white and red blood cells".*

Immune System

1. The Immune System starts with the top part of the sternum. Touch the top part of the sternum. Project your fingers 'very long' into the bone and feel them penetrate right into the thymus. Activate the thymus gland. Gradually, feel the fingers rising up to the thyroid and parathyroid. Activate the thyroid and parathyroid and throat center. Give the command, "Clear ,clean and bright fill with violet light and return to normal function".

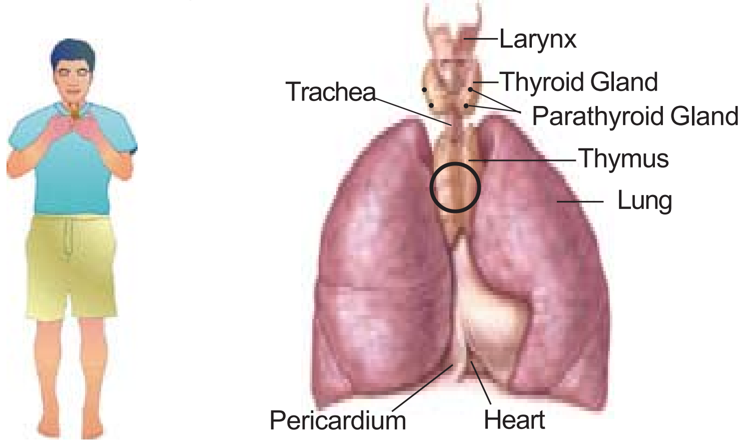

Fig. 4.76 *Touch the Top of the Sternum & activate the Thymus, Thyroid and Parathyroid Glands.*

2. Move the fingers to touch both sides of the jaw bones to help activate the tonsils, the body's first line of defense. The jaw bone affects the lymph nodes beneath, which include the tonsils. Fill them with Chi. Give the command, "Clear, clean and bright; fill with violet light and return to normal function".

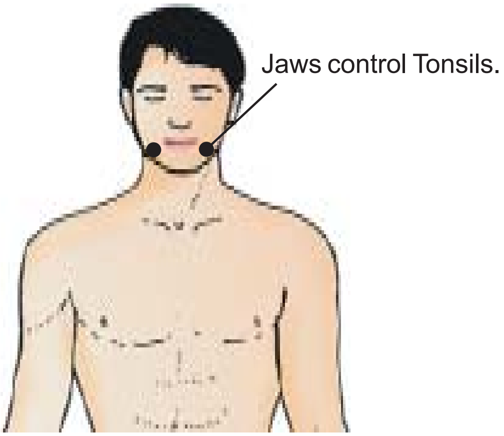
Jaws control Tonsils.

Fig. 4.77 *Activate the Tonsils and Jaw Bones.*

3. Touch the mideyebrow and crown, which control the pituitary and pineal glands respectively. Touch the third eye, feeling it open and feel the light from heaven entering into your brain and into your body. Touch the crown. Feel your fingers reaching deep inside and feel the Chi penetrating all the way down to your perineum. Give the command, "Clear, clean and bright fill with violet light and return to normal function".

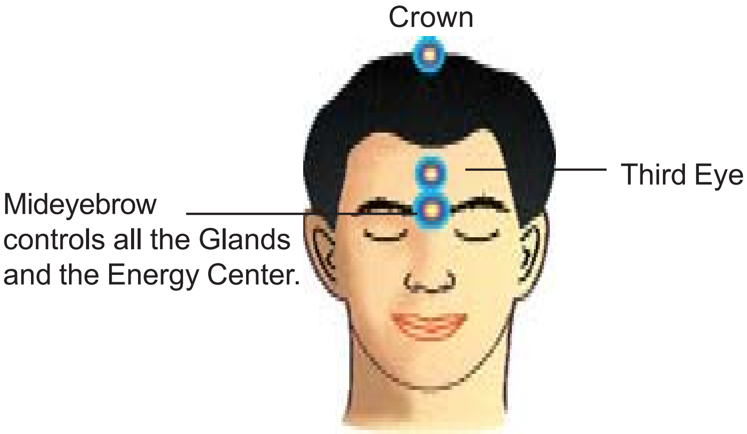

Fig. 4.78 *Third Eye, Mideyebrow and the Crown affects the Immunity and Defense Systems.*

4. Rub both of your hands until they are warm and cross the hands on the neck and feel the lymphatic system has been activated. Cross your arms and hold your hands

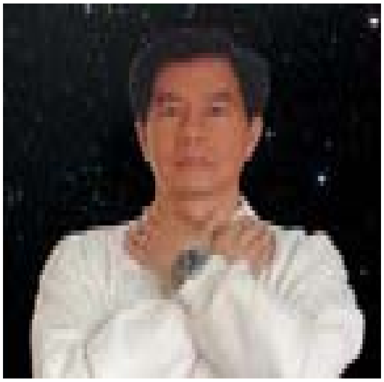

under your armpits. Picture the lymphatic system of the armpits as you activate the lymph nodes. Give the command; "Clear, clean and bright fill with violet light and return to normal function". The lymphatic system of the upper body is clear, clean and bright and the lymphatic fluid flows! Smile.

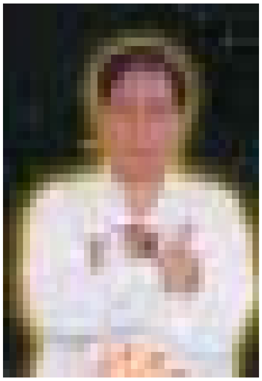

Fig. 4.79 *Lymphatic System of the Neck and Armpits are clear, clean and bright.*

5. Move the hands to the cover the groin area. Give the command to the lymph nodes there. Next, cover the lymph nodes of the navel area, and give the command; "Clear, clean and bright fill with violet light and return to normal function". The groin and the navel area lymph nodes become clear, clean and bright.

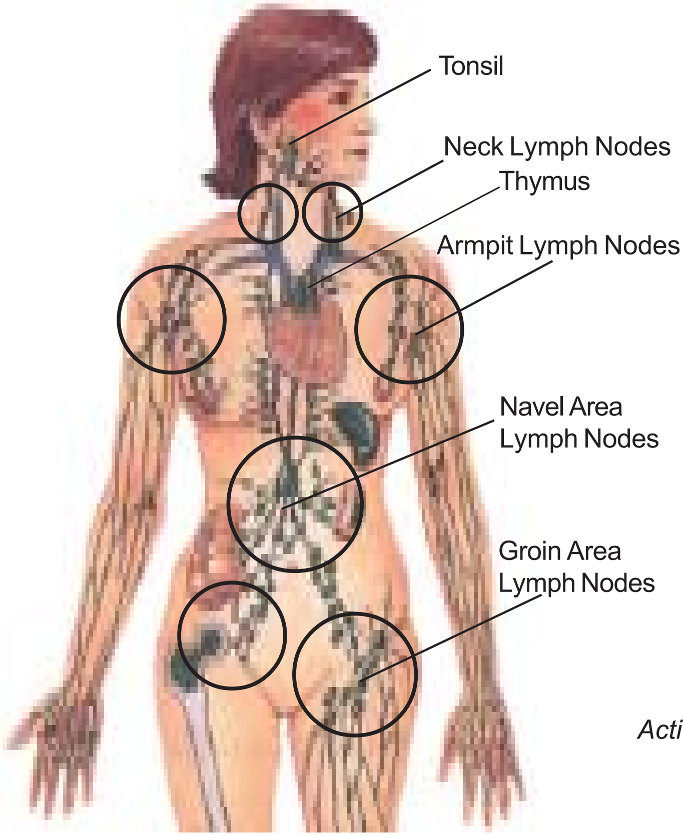

Tonsil
Neck Lymph Nodes
Thymus
Armpit Lymph Nodes
Navel Area
Lymph Nodes
Groin Area
Lymph Nodes

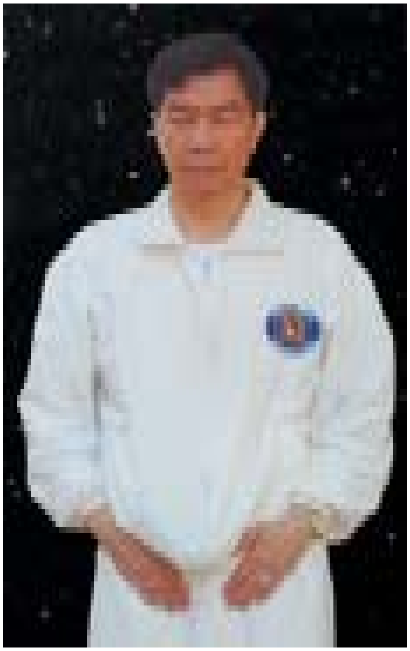

Fig. 4.80
Activate the Lymph Nodes in the Groin.

6. Place your palms over your navel and activate the lymph nodes. Give the command; “Clear, clean and bright fill with violet light and return to normal function”. See them become clear, clean and bright. You have now activated the immune system.

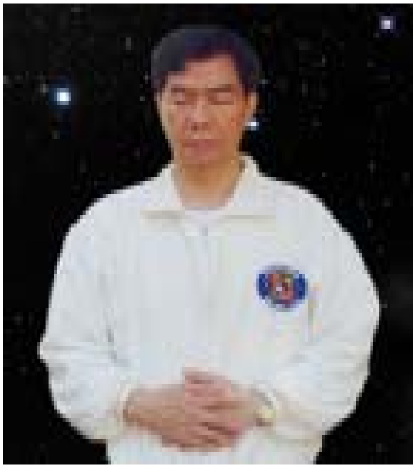

Fig. 4.81
Touch and activate the Lymph Nodes in the Navel.

Summary - General Healing Session

This General Healing Session will help everybody.

A. Three Minds into One
B. Activate the Six Directions
C. Activate the Three Fires
D. Linking Personal Stars, Energy Bodies and the Universe
E. Protective Circle and Chi Field
F. Chi Water (Sacred Water) Practice
G. Empty and Fill
H. Clean with Green Light
I. Clean with Blue Light
J. Charge with Violet Light
K. Defense and Immune System

A. Three Minds into One

Smile to the inner universe, activate the heart compassion energy; feel the love and joy within. Smile and empty the mind and heart down into the Tan Tien. Fill the Tan Tien with Chi and spiral the energy. When the Tan Tien is full, the energy will rise up to fill the brain. Manifestation will occur at the mideyebrow.

B. Activate the Six Directions

Expand the awareness to connect to the Six Directions at the crown, mideyebrow, heart and Tan Tien.

C. Activate the Three Fires

Activate the Tan Tien, Kidney and Heart Fires.

Activate the Sacred Fire Triangle circulating the energy between the Tan Tien, kidneys and the heart. Feel the energy's force multiply.

Activate the personal stars; the cosmic star above the head and the earth star below the soles of the feet.

Keep the Tan Tien warm, and spiral at the sacrum and mideyebrow. Turn your eyes to look upwards and see the crown spiraling, connecting to the cosmic star about six inches above the crown point. Feel the cosmic and earth stars pull the body simultaneously towards heaven and earth. See the cup of the Big Dipper fill with Violet Light, gathering Chi from the North Star and the universe. When the cup is full, pour the light down into the cosmic star above your head. Feel the violet light spreading throughout your entire system, filling the Upper, Middle and Lower Tan Tiens.

D. Linking Personal Stars, Energy Bodies and the Universe

Feel the central channel running from your cosmic star, through your three Tan Tiens and connecting to the earth energy below. Use you 'Yi' power to connect to the personal or cosmic star above the head. Connect the personal star to the stars of each other person in the meditation. Create a Chi body and become the major connecting point. Create a left spiral pattern to connect all the stars into a larger group. All the meditators then become a satellite, creating a group energy body.

E. Protective Circle and Chi Field

Raise the hands and touch the Universal Fire. Use your Yi to bring the energy down and manifest a Sacred or Chi Fire around the room, creating a protective circle. Create a Chi Dome and set up the guardian animals and/or angels. Activate the eight elemental forces of nature and the universe.

F. Chi Water (Sacred Water) Practice

Invoke the power of the Sacred Water practice to cleanse the body of all sickness, negative energy and bad fortune. Energize and renew the self at a cellular level with health, wealth and longevity.

G. Empty and Fill

Empty any remaining negative energy, sickness, burdens and worries, down in the depths of the molecular mother, the earth. Bury the energy there so that the earth can recycle it. Bring recycled energy up from the center of the earth into the body. Fill and compact the Chi into the 'hollow bones'. See all of the cells on the horizon of your mind's eye; spread them across the panorama and see the cells as big as the stars in the night sky. Fill them with Chi, blending it with any sick or grey energy. Then empty again, as you guide the all of the Chi back into the earth for recycling. If you wish, you can fill with earth Chi once again before you begin to 'clean with green'.

H. Clean with Green Light

Green cleans and detoxifies all of the body's cells**.** Green Chi is mild and safe. Green Chi is used for cleansing, and as a decongestant for diseased parts. Be aware of the liver's green color, the throat and the emerald forest. Breathe into the throat, feel yourself breathing in the green light. Guide it down to the Tan Tien, blend and move it up to the heart and the crown.

Project it to the universe; spiral and let it multiply. Channel one direct line down to the student, one to your crown and one to your palm. Bring the energy back down and blend it with any sickness in the liver, gallbladder or anywhere else in the body. Spiral the energy counter clockwise and down into the earth. Continue cleaning with green until you see that all the cells have been charged with the emerald light of the forest.

I. Clean with Blue Light

Be aware of the kidneys' blue sapphire color and breathe the blue light into your throat. Guide it down to the Tan Tien, blend and move back up to the 'heart of compassion' and then to the crown.

Project up to the universe, spiral and let it multiply. Channel one direct line down to the student, one to your crown and one to your palm. Then feel the universe pouring the blue light back down; clockwise. Spiral down. Look at the cells and see the blue light, like water, go into the cells to rinse, clean and remove any sickness. Spiral the energy counter clockwise and down into the earth. Continue rinsing with blue light until you see that all the cells have been charged with the sapphire blue.

J. Charge with Violet Light

Luminous Violet has intelligence and can be programmed.

Turn three minds into one mind, go deep into the empty space of the cell and deep into the chromosomes, magnify the space as you enter into the DNA and return to the original cells. Ask for the power to see the cells and look at the brain cells spread out into the space. "Brain cells please listen. Clear, Clean, and Bright: fill with violet light, back to normal function". See violet light. Be aware your crown. The Tan Tien Chi and the negative Chi all combine in the crown. Extend a violet light up to the universe. See the North Star and the Big Dipper. Tell the students to hold their arms in the air, so that they can funnel the Chi into their personal star. Bring the North Star and the Big Dipper down; hold the handle of the Big Dipper and pour the violet light over the student's personal star. The higher cells will be programmed. Let the Chi enter the brain and ask for the power to see the brain cells. "Brain cells listen to the command". If there is sickness in the brain, give the command. "Sickness go away". Continue to spiral the violet light down through the cells of the whole body.

1. Ask for the power to see into the cells of the **Brain.** Ask the student to move their hands to cover their brain. "**Brain Cells** listen to the command;"Sickness go away. Clear, clean and bright fill with violet light and return to normal functions".

2. Ask for the power to see the cells of the sense organs. Ask the student to move their hand to cover their senses. Give the command; **"Eyes, Ears, Nose, Tongue and Mouth** cells listen to the command. Sickness go away. Clear, clean and bright fill with violet light and return to normal function".
3. Move your hands down to your Throat Center and give the command; **"Thyroid, Parathyroid and Thymus Glands** listen to the command; Clear, clean and bright fill with violet light and return to normal function".
4. Move your hands to the heart and the lungs. Ask for the power to see the cells of the heart and lungs. Give them the command; **"Lungs and Heart Cells** listen to the command; Clear, clean and bright; Fill with violet light and return to normal function".
5. Move your hands to cover the left and right sides of the ribcage. Ask for the power to see the cells of the liver, gallbladder, spleen, pancreas and stomach. Give them the command; **"Liver, Gallbladder, Spleen, Pancreas and Stomach Cells** listen to the command; Clear, clean and bright fill with violet light and return to normal function".
6. Move your hands to the large and small intestine on the front of the abdomen. Ask for the power to see the cells of the intestines. Give them the command; **"Large and Small Intestine**, listen to the comm and Clear, clean and bright fill with violet light and return to normal function".
7. Move your hands to the kidney and sexual organs and ask for the power to see their cells. Give the command; **"Kidney and Sexual Organ Cells** listen to the command; Clear, clean and bright fill with violet light and return to normal function". Gather the violet light again and let it pour down and fill all the cells. See the violet light again and feel and see that all the cells are "clear, clean and bright". Smile.

K. Defense and Immune System

Defense System

1. Work on yourself first to activate your **Sacrum**. Project the sacrum as big as the universe. Breathe and compress the universal Chi into the sacrum until it rises up to fill your brain. Project out into the universe, multiply the energy and spiral it back down. Project it to the students and tell them to touch the sacrum.
2. Tell the students to move their hands to the **Hip Bones**. Tell them to smile to these bones. Feel them as 'funny, laughing, happy bones'. Tell them to feel the electricity in the sacrum and bone marrow.
3. Touch the **Femur Bones.** Feel the electricity running up the legs.
4. Touch the **Humerus Bones.** Feel the electricity running up the arms.
5. Touch the lower part of the **Sternum** and feel the Chi penetrate into the bones and spread out to the **Rib Cage**. Feel the electricity run throughout the rib cage.

Immune System

1. The Immune system starts with the top part of the **Sternum**. Touch the top part of the sternum. Project your fingers 'very long' into the bone and feel them penetrate right into the **Thymus**. Activate the thymus gland. Gradually, feel the fingers rising up to the thyroid and parathyroid. Activate the **Thyroid** and **Parathyroid** and **Throat Center**. Give the command; "Clear, clean and bright fill with violet light and return to normal function".
2. Move the fingers to touch both sides of the **Jaw Bones** to help activate the **Tonsils,** the body's first line of defense. The jaw bone affects the **Lymph Nodes** beneath, which includes the tonsils. Fill them with Chi. Give the command; "Clear clean and bright fill with violet light and return to normal function".
3. Touch the **Mideyebrow** and **Crown**, which control the **Pituitary** and **Pineal Glands** respectively. Touch the third eye, feeling it open and feel the light from heaven entering into your brain and into your body. Touch the crown. Feel your fingers reaching deep inside and feel the Chi penetrating all the way down to your perineum. Fill the glands with Chi. Give the command; "Clear, clean and bright fill with violet light and return to normal function".
4. Rub both of your hands until they are warm. Cross your arms and hold your hands under your **Armpits**. Picture the **Lymphatic System** of the armpits as you activate the lymph nodes. Give the command; "Clear, clean and bright fill with violet light and return to normal function". The lymphatic system of the upper body is clear, clean and bright then the lymphatic fluid flows! Smile.
5. Move the hands to the cover the **Groin** area. Give the command to the **Lymph Nodes** there. Give the command; "Clear, clean and bright fill with violet light and return to normal function".
6. Place your palms over your **Navel** and activate the **Lymph Nodes**. Give the command to the lymph nodes there. Give the command; "Clear, clean and bright fill with violet light and return to normal function". **You have now activated the immune system.**

Chapter V
Cosmic Healing I - Basic

Source of Chi and Color

There are different sources of Chi in the universe: cosmic (air), earth, nature and human. Different parts of the human body have different energies and are correlated with the universe and nature. Each part of our body, especially the organs and glands can produce, receive, transform and emit different Chi and color.

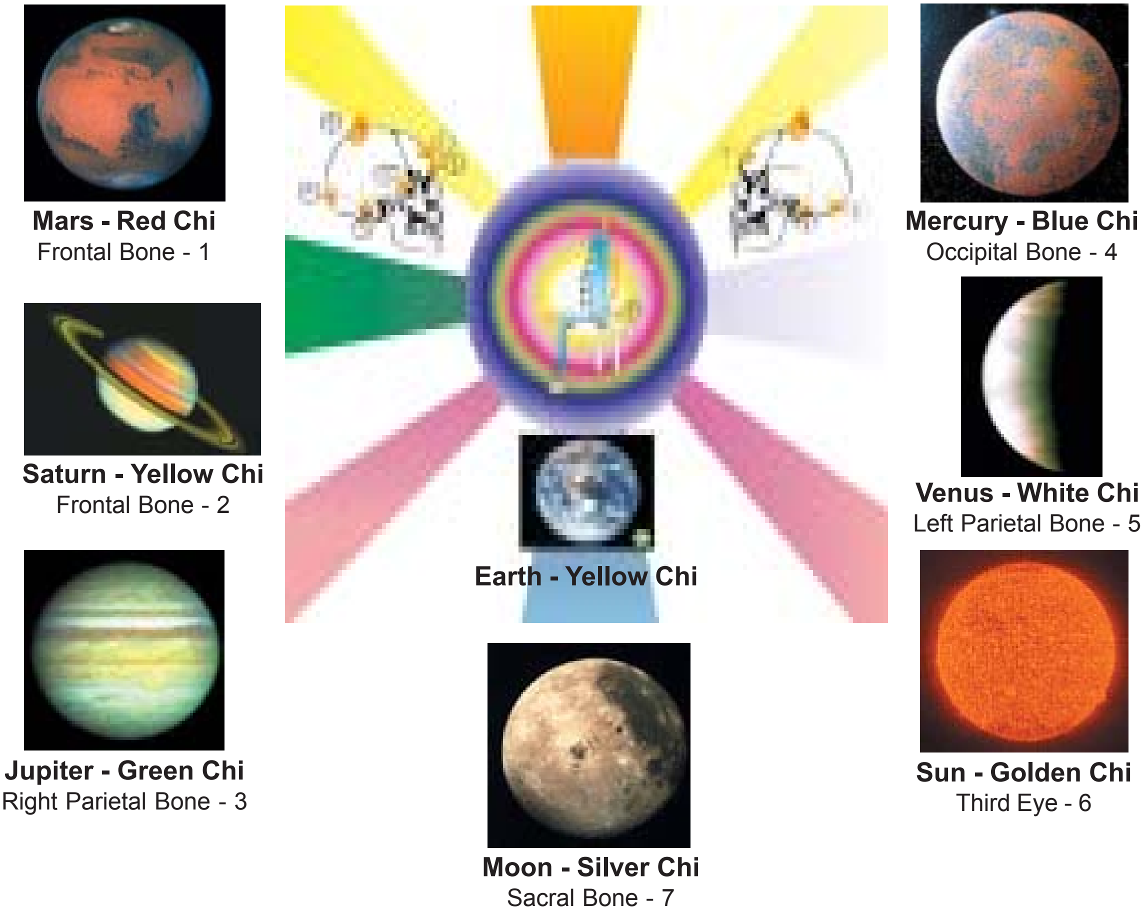
Mars - Red Chi
Frontal Bone - 1
Mercury - Blue Chi
Occipital Bone - 4
Saturn - Yellow Chi
Frontal Bone - 2
Venus - White Chi
Left Parietal Bone - 5
Earth - Yellow Chi
Jupiter - Green Chi
Right Parietal Bone - 3
Sun - Golden Chi
Third Eye - 6
Moon - Silver Chi
Sacral Bone - 7

Fig. 5.1 *Planet, Bone and Cosmic Colors*

Earth Chi

While Cosmic Chi (the atmosphere over us and the universal force) is ethereal in nature, what is called earth Chi is not. This Chi appears as the line visible above the horizon when the sun is low in the sky, and seems to hover just above the ground. It actually also sinks into the ground.

Earth Chi contains yellow and some white Chi. When the ground and the cosmic (air) combine they will become white Chi. This Chi has a very powerful healing energy, not too overheated, but balanced and mild.

When treating a problem, white Chi, in the proportion 7:3 (7 white to 3 of the selected color) acts as a brake on colors which are hot or too active. This enables the body to take in chi more easily.

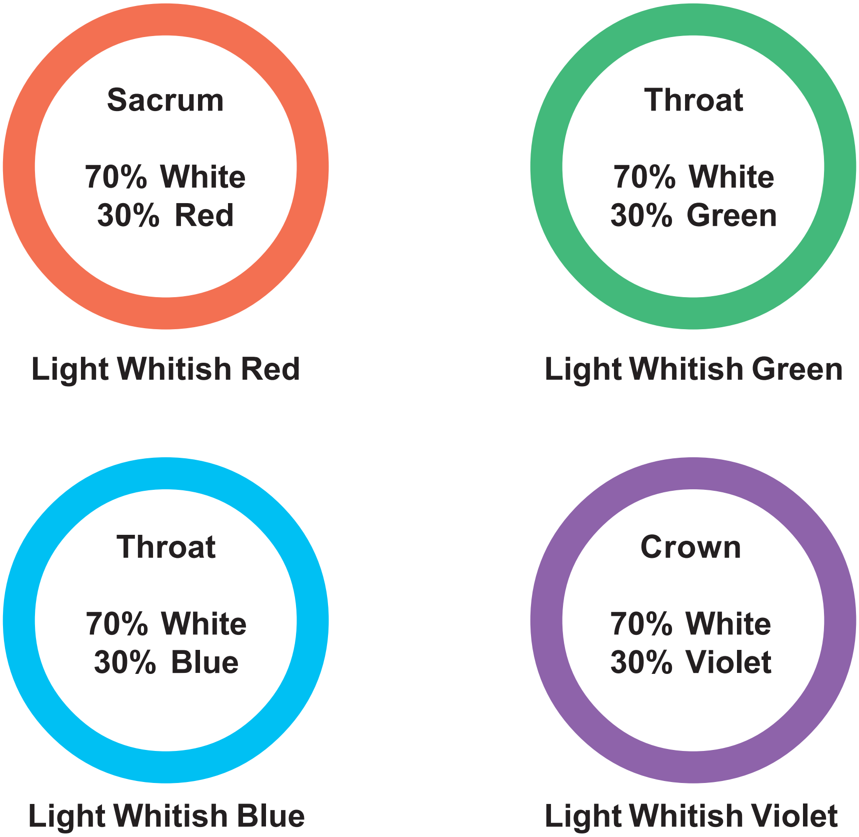
Sacrum
70% White
30% Red
Light Whitish Red
Throat
70% White
30% Green
Light Whitish Green
Throat
70% White
30% Blue
Light Whitish Blue
Crown
70% White
30% Violet
Light Whitish Violet

Fig. 5.2 *Use White Chi to blend all the other Colors.*

Special Note: Many of the color references came from the following books: *“Let There be Light”* by Darius Dinshah and *“Advanced Pranic Healing”* by Master Choa Kok Sui.

Human and Nature Chi

These include the things we can see on the earth: the mountains, lakes, seas, forests, streams, caves, valleys, rocks and precious stones. Taoism classifies five elemental forces: earth, metal, water, wood and fire.

So there are five elements in our body and five elements in nature; there are five elements in the cosmic and five elements in the universe.

According to the Taoists, when we can make the connection and control our body's elements, we will quickly make connections to nature and the universal forces and employ them very easily.

Solar Chi

Before sunrise or sunset there is more white Chi in the air, which we can use abundantly for health and healing. White Chi affects the lungs and the large intestine, so we direct this Chi to these organs. This will generate further Chi to maintain our daily activities. All colors of the spectrum are present in white.

Simple Practice

Look at the horizon at sunrise and sunset. Stand still and extend your palms toward the sun, smile and absorb the rising or setting sun's energy into yourself.

General Information

Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral.

Move the hand **anticlockwise** with a spiraling motion to clean, flush and remove energy from the treatment area. After this, brush down the whole body, to direct the sick energy into the earth.

Use an anticlockwise spiraling movement in order to clean, then expel the energy.

Use a clockwise movement to bring energy in and an anticlockwise one to take it out.

As a general rule and especially when you are beginning to practice the techniques, keep the color combinations simple. **Always use blue (to cool, calm and stabilize), green to clean, white to harmonize and violet (which carries the properties of all the other colors) to program.**

Although it is important to learn all of the applications, it is more important to begin practice, maintaining a serene state of mind throughout. You will immediately realize that Cosmic Healing provides the practitioner with as much benefit as the student. The more you practice the more you heal.

Healing Colors

White Chi

The color of the chi from the air, earth and sun is always white and acts as a calming agent for the surplus chi produced by the other colors. It also redirects this chi to areas of the body which need it. For this reason, use white in combination with another color in the proportion 7:3.

Shining white chi should be concentrated at the center, while colored Chi goes to the edge. This can also be reversed. It is also possible to blend the white and colored Chi together.

It is safe to use white chi when you are not sure which color might be suitable for a patient, especially when treating babies, toddlers, the elderly and the weak. People with minor problems can be treated with blue, green or mauve, while gold colored Chi or mauve/white Chi is better when the problem is serious. For a stimulating and strengthening effect use white-red.

When it becomes necessary to switch from one color to another, an abrupt shaking movement of the hand should ensure a smooth transition to a new color.

There is a very strong violet light in the North Star. You can dilute the color with white. White Chi is composed of red, orange, yellow, violet and blue.

Blue Chi

The effect of blue Chi is one of cooling, and hindering similar to yin or water energy. The opposite effect can be obtained with the use of red Chi.

When we draw the energy down, some peoples' bodies can't adjust to the energy immediately, so using blue Chi is always good. "Water blue' is the one element that always brings harmony and does not cause any harm.

a. Blue Chi, like yin, water energy, has a cooling effect. It also can act as a restraining agent.

b. Because of blue Chi's ability to cool, soothe and numb, it is particularly appropriate for promoting relaxation and sleep, reducing abnormally high body temperature, minimizing pain and swelling. It can also hinder the proliferation of viruses and bacteria and assist in the rapid clotting of blood.

Blue (and green) can detoxify and also energize.

Green Chi

While tangerine Chi can be quite dangerous, the action on green chi is gentle and harmless. Freeing up a damaged area so that cleaning can then take place may be done with green Chi. Once it's loose you can then use blue or organ's color Chi to expel the disease completely. It can therefore remove the disease and bad energy by locally sweeping either to and then out of the arms for the upper body or out of the legs for the lower body.

If there is a need to energize with the organ's color Chi, first energize with pale green Chi. Always use green chi before using the colors purple, tangerine and red.

Use the throat to activate and draw the green Chi up to the crown, rising to the universe to be multiplied and brought back down for use. Always add blue Chi when there is pain. The motion of pushing and pulling through the affected part is very powerful. You also can also use the 'sword finger' in a small area and for infection.

Green can also be used for energizing together with violet; first use green and follow with violet.

Pale green and pale tangerine chi can be employed in sequence, as the first acts on the 'sick' energy, reducing it to a form which can easily be driven out by the second.

Simultaneous use of pale green pale tangerine and white chi in the ratio 2:1:7 is recommended to treat stubborn areas because the effects are intensified when these proportions are used together.

In the treatment of cancer, dark green and tangerine Chi may be employed, preceded by blue. This can assist in toxin removal, and in the cleaning and freeing up of disease affected areas so that necrotic cellular tissue is released. These colors can relieve congestion and free up 'sick' energy which is 'stuck' when used with brushing movements over the affected areas. They also help with treatment of colds and elevated body temperatures, and in the dispersal of blood clots.

Red Chi

Cerise chi makes the area to which is applied stronger, while crimson chi has the opposite effect. When using color for healing, project luminous white in the center and cerise at the edges for a strengthening effect. Always combine red with blue and green; never use it on its own.

Cerise mixed with white, because of its expansion abilities, can help widen blood vessels and breathing tubes. It also helps those with heart and asthma problems.

Red chi with its qualities of warmth and expansion, has a number of benefits. It is good for the circulatory and respiratory systems and the blood, as it expands blood vessels and other tubes, thus improving circulation and breathing, particularly for heart patients and asthma sufferers. It also brings increased energy into the blood and helps allay feelings of fatigue or weakness, driving out 'sick' energy along with toxic substances and other waste materials. It can also help those who are sensitive to allergens. It may help those whose lives are near the end have a longer period on earth, and promote revival of consciousness.

Avoid the use of dark colors, which may cause an adverse reaction. If a person is suffering from a sexually transmitted disease, it is not advisable to treat him or her with chi which is crimson. Chi of this color makes the microbes that cause such diseases multiply at a fast rate and can also produce an inflammatory or constrictive reaction.

Tangerine Chi

When using orange Chi, always use blue Chi first; to tone and calm down the ailing body area, before applying the vitality of orange.

Pale tangerine chi with added white can be used; however certain areas are inappropriate for treatment by this color because of its power. Do not use with the eyes (including retinal problems), heart, head, throat, brain (for example bleeding in the brain) or the spleen.

The large intestine benefits from the use of tangerine chi, but this color should be used sparingly on areas like the solar plexus and navel. Appendix problems should not be treated with this color as it could exacerbate the problem. Tangerine chi has an effect on the bowel, and can also aid in returning an unconscious person to consciousness or in stimulating someone who is dying.

To treat cancers of certain types, dark green and dark tangerine are suitable because of their ability to destroy. Always avoid the head and the heart when using orange.

Forcible removal of waste products, toxic substances, ‘sick’ energy, viruses and bacteria
Menstrual difficulties
Waste elimination problems
Freeing up ‘sick’ energy
Breakdown of clots in the blood
Problems of the urinary and respiratory systems
Diseases of joints and connective tissue
Common cold
Problems caused by allergens
Cysts

Yellow Chi

Yellow Chi has a close connection to the nerves, the bones and the bone marrow. It is the color of the spleen; in Taoism we feel the spleen is involved in the assimilation and processing of food and the appetite. This doesn't mean eat a lot, but when the spleen is balanced the appetite will be controlled.

Chi which is yellow in color is beneficial in the treatment of nerves as it promotes re-growth. It also helps in the repair of fractures and cells. It is good for the health of bones, tissues and organs. It has the power to increase bone marrow production.

Violet Chi

Violet Chi has the properties of all the other colors. It’s used for severe types of sickness.

The qualities found in all colors are contained in purple chi, therefore it is particularly effective for serious illnesses. Mauve (light purple) chi which contains elements of blue and green, helps to restore organs and nerves affected by disease or trauma.

Although mauve chi can be employed on certain kinds of severe infections, chi which is dark purple should be avoided when diseases of the respiratory system are present, as it could cause the proliferation of disease causing microbes.

Violet Chi is use chi that is pale white/green or pale white/tangerine and, selecting each in turn, with a brushing movement apply to the whole spinal area and especially to the area where the trauma is located.

If purple chi is used in combination with other colors, it will intensify their qualities and if used together with red, tangerine or yellow chi can be dangerous; it may cause the undesirable proliferation of some cells.

Dark electric violet can be used for tumors or cancers. Always use the blue or green first.

In order to make an area which has become weak stronger, mauve mixed with white can be used, but for really fast strengthening use cerise mixed with white.

Both white and violet have the properties of all the Chi colors. The difference is that violet Chi has a greater penetrating effect and is easier to assimilate than white.

Because bright purple chi has been acted upon by the soul, it has the capacity for independent thought and action.

Energy or chi which comes from the air, the earth and the sun is called purple chi. There is a second type of chi with this color, but having more brightness – this has been called bright purple chi and is derived from the soul through the crown of the head, the point connected to the North Star or Big Dipper.

Bright purple chi can only be brought down from the universe via the crown point if the practitioner has highly developed senses, The effects of chi from air, earth and the sun depends on the development of the practitioner’s soul.

In contrast to purple chi, the chi which is bright purple in color has a much greater potency because it can help in the restoration of nerves and organs which have been affected either by disease or trauma. This bright purple chi can be used to remove infection and to promote fast healing, even in the case of serious illness. In addition, this chi has been said to possess its own consciousness (Master Choa Kok Sui. Advanced pranic healing:49)

When cancer and other growths are being treated, use bright purple chi because it has the ability to destroy. However, beforehand, the affected area should be energized with pale blue chi.

Do not use bright purple chi at the same time as chi which is yellow, green, tangerine or red because it will intensify and increase the effects of these colors. In addition, the use of any dark colors with bright purple chi (either preceding or following its use) should be avoided, as this will produce a very damaging result. As a general guideline, use blue before electric violet.

Golden Chi

Gold colored chi, which has almost the same qualities as bright purple chi, is formed when bright purple chi meets the spirit body. The color changes to cerise (pale red) when this gold colored chi enters the physical body. However, this chi is does not have as much influence as the bright purple chi, being gentler and therefore less effective as a cleansing agent. Gold colored chi should be used to treat a very local area, while bright purple chi can be employed over a wider area.

Spleen

The spleen is connected to the solar plexus and the navel, the major center that connects to all the organs. Holding solar energy or a white ball at the solar plexus will help distribute Chi to all the organs.

If the solar plexus gets clogged up the liver and spleen will also become blocked. The spleen, the Door of Life and the navel are all linked together; the navel and the Door of Life sit opposite each other. We call this an emotional link. Many diseases are caused by emotional links; people use psychology to deal with it which can work to a certain level, but all the energy would eventually get stuck at this main juncture or plexus. White chi is taken in mainly at the spleen, and before being sent to the various parts of the body, is split into its colored forms of red, green, purple, blue and tangerine.

If the spleen is not clean, the immune system will be affected, together with the production of bone marrow. The blood will be dirty and diseases of the joints and connective tissue encouraged.

Cosmic Healing Chi Kung works very well with Chi Nei Tsang. The Cosmic Healing Chi Kung, Chi Nei Tsang and all the other Universal Tao practices all play a unique role in working to heal the physical and subtle bodies. When you combine these three in synergy, it allows you to perform many healing tasks. The spleen (located under the left side of the rib cage), spleen center or navel can absorb white Chi directly into the front or back. The Chi will be broken down into color and distributed to other centers and their associated organs.

There is one thing that I always advise: the recipient or student should be under a doctor's supervision and that they should alternate between the medicine from the doctor and the work with us at the Universal Tao. The doctor should be advised of what we intend to do with the student.

Cosmic Healing Planetary Colors

In Taoism we give each planet a color which enhances the healing power. Have the ‘Planets’ chart (located at the beginning of this chapter) in front of you, look at the relevant planet and close your eyes. Hold the planet in your mind.

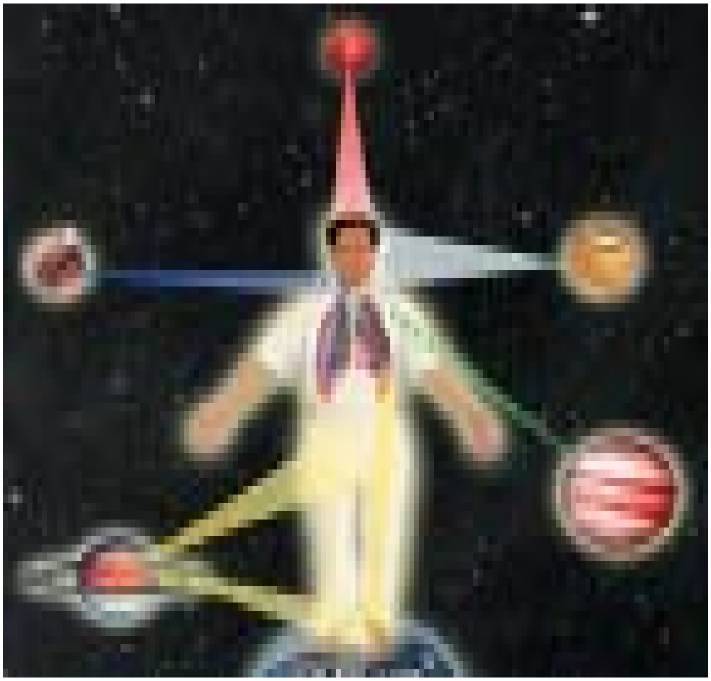

Fig. 5.3 *Channel the Cosmic Planetary Energy into the body.*

Mars: Focus your awareness on the star above you, turn your eyes up and look up to the crown. Hold the image of Mars and extend your sight beyond the crown, looking up into infinity. See the **Red Light** and gradually bring it down to your head about 6 feet above you. Form it into a red ball, a holographic image of the planet Mars shining its red light above you. Invite it to flow down to your ‘star’ (located above your crown) and continue to flow into your Upper Tan Tien for processing. Let it flow down to the heart (or T-2) and down to the palms. Red and orange light also can be breathed in through the sacrum. Feel the sacrum pulsing and breathing in the red and orange light. Practice 9 to 18 times, rest and shake your hands.

Venus: Become aware of the star above you, turn your eyes up and look up to the crown and into infinity, seeing the expanse of **White Light.** Gradually bring the light down to your head, about 6 feet above you. Form it into a white ball, a holographic image of the planet Venus shining its white light above you. Invite it to flow down to the star above your crown and continue to flow into your Upper Tan Tien, for processing. Now it may flow down to your lungs freely.

Saturn: The **Yellow Light** from Saturn can also be absorbed via the earth through the soles of your feet and perineum. This is for your spleen. Follow as above.

Mercury: The **Blue Light** from Mercury may be absorbed through the throat center. This is for your kidneys.

Jupiter: The **Green Light** from Jupiter can also be absorbed through the throat center. This is for your liver. Follow as above.

Summary - Working with Color

(Practitioner to Student)

There are two ways of doing this type of healing session:

One way is to focus on the specific location in your own body, using your hands and ask the other person to do the same. Together you complete the entire route.

Another way is to focus on your own energy body and ask the student to do the same. Bring a channel or line down from your energy body into your physical body and do the same for the student. Once again, complete the routes together.

For reasons of clarity this text only mentions one student and uses the male gender; *this session may however be performed with a group also*. Before starting the session do some group meditation practice and warm ups.

1. Be aware of the Tan Tien and connect to the universe. Let the student sit with his back to you. You are behind him. Be aware of your **Sacrum** area and feel the Chi; **wait for the Chi to rise up to the crown** and to the universe. Spiral the Chi down to the student's and your own energy body and it will flow into the physical realm.
2. Draw in the green light with your palm, push it right through the sacrum and disperse the sickness throughout the other side of the universe. Pull the green light from the universe through the sacrum and push the sick Chi out to the universe. Do this 6 to 9 times until you feel the **sacrum has been cleaned**. When you pull back stop the green Chi at the student's sacrum; there is no need to pull it back to you.

   Yellow light from the earth will help strengthen the sacral bone. Visualize the sacrum and vitalize the complete bone structure with yellow light passing throughout.
3. Be aware of the energy body above the **Crown**; extend yourself up from above your crown and **channel down the white light from the center of the universe and the violet light from the North Star.** Using your 'mind-power' ask the inside of the bone structure to open, allowing the white and violet light to **flow into the marrow.** Focus on the energy body; you can picture the energy body's sacrum and bone

structure to help guide the energy inside and see the whole body light up from deep within.

4. Then focus on the **Door of Life and the Navel.** First flush them through with **green** light, cleaning them out. Draw the green light from the Door of Life through to the navel, and then push back through. To cool down the Door of Life (if necessary), draw the **blue** light and push it through from the navel to the Ming Men, connecting with the universe. Then bring the **white or violet light down and activate the Lower Tan Tien.** Always teach people about their Tan Tien. It is like an ocean and the body is like a hollow bamboo. The bamboo can bring the water up. This way energy will never dry up.
5. Concentrate on the **Solar Plexus and the Chi Chung (T-11)**. Once again, draw the **green** light, spiral it, use it to flush, then **energize with white and violet light**. As before, always allow the energy to stream through both the points completely. The solar plexus holds all the emotions. When working on the solar plexus, the **most important thing to remember is to connect the 'backside' to the universe.** There is literally no end to this connection. Just allow the Chi to come all the way down and then pull it towards the rear slightly. Simply clean out the path. Allow the information to condense; allow any images to manifest and then release. Then stabilize the energy. Picture the Chi-Field enveloping the person as a big protective bubble. **Cool down with blue.**
6. Proceed to the **Heart and the Shen Dao (T5/ T6).** Draw in the **green** light, push it through and repeat several times. Now scan the heart with your palm, sensing its strength; then select an appropriate hue, i.e., not too dark. Send the **red** light through the heart to strengthen it.

   Cool down any excess heat in the heart, flushing down and out, using **blue** color. Draw in the **white** Chi; push it through to the **Wing Point (T5/T6)**. **Energize the center at the back of the heart using violet and golden Chi.** Picture the heart surrounded by a golden aura.
7. Move up to the **Throat Center, flushing through to C-7**, using **blue** light first to open and clear it and then **green** light to clean.
8. Next activate the **mideyebrow.** Focus on the mideyebrow in your own body. Use **golden yellow** Chi; use it to **flush and stabilize** the mideyebrow. Flush it all the way through to the backside of the head. **Energize with violet golden light.**
9. Proceed up to the **third eye** in the middle of the forehead. **Connect to the Kun Lun** at the back of the crown. **Flush** through both points using the pale **violet light. Energize** with **electric white or golden light.**
10. Concentrate on the **crown**. Use **violet** or **golden** light to enter and **flush** all the way down through the **central channel**, leaving the body at the **perineum**. **Cool down the system by showering blue light over the whole body.**

    When people have a serious infection, they should consult a doctor. It is common for infections to take a long time to clear up nowadays, even when using antibiotics.

There are two ways to deal with this; allopathically or naturally. The Taoist practices are in great conflict with ‘contemporary (i.e., not alternative/natural) western concepts’. The western concepts recommend a full course of antibiotics. Using the theories of Chi Kung, a baby for example, would absorb natural antibiotics via the bones. Western Allopathic Practice consider antibiotics as wonder drugs. The only ‘wonder’ is when will they realize the many harmful effects that it has on the body? Babies that have been ‘fed’ with antibiotics frequently suffer many organ related problems by the time they reach 20 to 30 years of age. The worst thing is that they are given to children nearly every time that they become ill. If you scan (cosmically) the solar plexus, thymus or spleen of such a person, the area feels like a hollow void. This means that the immune system is totally non-functional. The same thing applies to ***Aids***: the immune system no longer functions. Any virus can enter the body - medicine cannot kill it. There is nothing left to ‘fight with’.

In my experience and following my own experiments, in circumstances like the above my own practices are as follows: I clean my colon; there are a few ways to do that. I eat fiber, vegetables, especially green leafy ones. I eat a big bowl of vegetable soup. I also add things like onion, pumpkin, carrots and cabbage, which will give it a sweet taste. I don’t eat any meat and drink a lot of water. The colon is cleansed within two days. I also brush the skin, which is a very good detoxifier. The vegetables will provide natural carbohydrates which are digested and absorbed very quickly. A healthy diet consists of many whole natural foods. It is ‘alkaline’ as opposed to ‘acid’ based. I also use the ‘colema’ which I feel is the best cleanser.

Chapter VI
Individual Healing Session

The Individual Healing session embodies all of the techniques that you have used previously and consolidates them for use on a one-to-one basis. The session comprises the following three parts:

A. Almighty Knife (Chi Knife)

The almighty knife can be used for any health problems. We can charge the knife with Yin or Yang energy. We charge the almighty knife most often with Yin Power. Yin is the cold energy and is used for any inflammation, pain or heat. Yin is used to break up illness. Charge the almighty knife with Yang for strengthening, melting or expelling. Only use the Yang when there is no pain.

B. Energizing and Activating the Immune System

C. Cleansing the Internal Organs

A. Almighty Knife

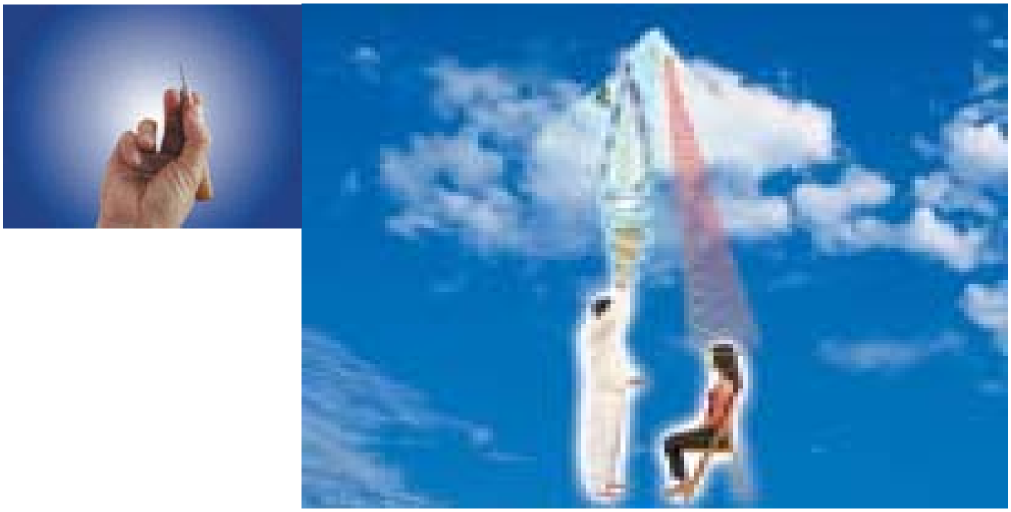

Fig. 6.1 *Almighty Chi Knife*

Since the dawn of ages and our descent into the material realm, we have constructed, manufactured and used tools to facilitate the progress of events. They are an extension of ourselves that enable us to enhance, mould and magnify our intrinsic nature as inquisitive playful beings.

We have used implements to help nourish, protect and amuse ourselves and further our progress in attempting to understand the 'how' and 'why' of existential reality. The knife has served as a means to carve our way into the present. It is a simple, yet powerful 'friend' and derives as much power as we wish to lend it.

Its design may consist of any combination of the five elements (Earth) which when combined with Universal (Heaven) and Cosmic (Man) forces, provide a unit of energetic substance which receives and transmits magnification, purpose and power in accordance with the human will that governs it. 'Chi Knife' and "Powers of Ten' adopt the precepts of a truly magical existence. When we arrive at our destination we will realize that we have indeed turned "the Wheel of Law'. In achieving this end we will have completed the cycle and returned to the beginning of the journey. We will understand that the simple things in life, like the knife, contain the same amount of power as state of the art technology.

We are all returning to the Garden of Eden. Complex telescopic and microscopic equipment now enables us to see the truth:

> "Inside has no outside; outside has no inside".

They are one in the same thing. When they merge together we will indeed enter the Kingdom of God.

Outline of Practice

1. Hold the knife in the right hand (left, if you are left handed). Hold the other hand near your body, with the thumb and index finger touching. Project the Chi Knife as big as the universe, 'long and big - reaching to Heaven'. The bones in your arm are hollow and the 'bottom' is sealed at the shoulder. Feel them 'fill and pack' with the power coming from above.
2. Then give the command: **"Power comes from the east. This Chi Knife has the power to dissolve all the negative energy, sickness and bad fortune. By my request, please carry out the order now."** Wait until you feel that your arm and the knife are heavy and full of Chi.

Eye Problems

Use the thumb and the index finger to hold the eye wide open.

You hold the knife like holding a pen, 3 to 6 cm away from the eye. Use the little finger for leverage and support.

WARNING!

DO NOT USE A KNIFE. THE **MASTERS** OF GREAT EXPERIENCE USE RITUAL SCALPEL TYPE KNIVES. THE STUDENT SHOULD USE THE "SWORD HAND" FINGERS (P.76, FIG. 4.24).

DO **NOT** USE SHARP OBJECTS NEAR THE EYES**.**

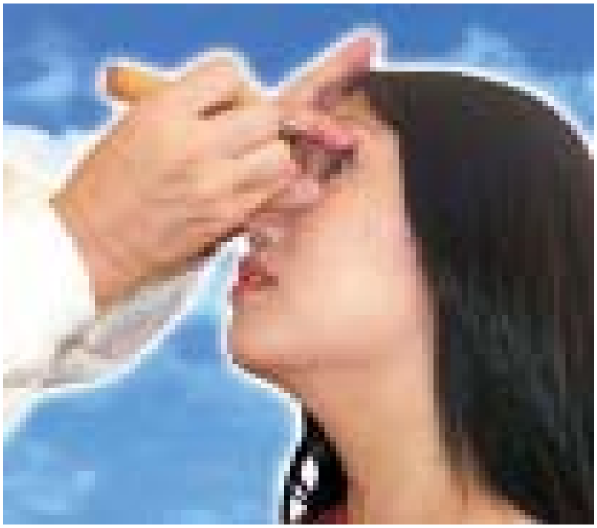

Fig. 6.2 *Hold the knife three to six cm. from the eye.*

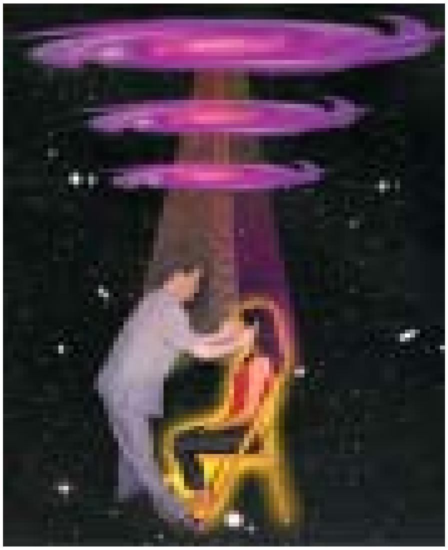

Chi knife - “Almighty Knife” helps all eye problems”. Cut up and down and sideways in the cross pattern ‘through’ the eye thoroughly. Tell the student to move the eye left and slowly move to the right as you cut with the Chi Knife. When they move their eye to the right, move the Chi Knife to the left. **Remember that the knife is used for ‘energetic surgery’ and should never make contact with the eye itself.**

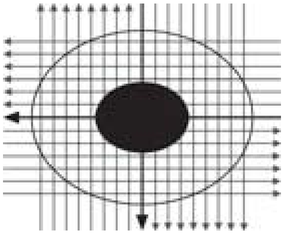

“Slice” the eye toward the right side for 50 to 100 knife movements and do the same to the left side.

Fig. 6.3 *“Almighty Chi Knife helps eye all problems”.*

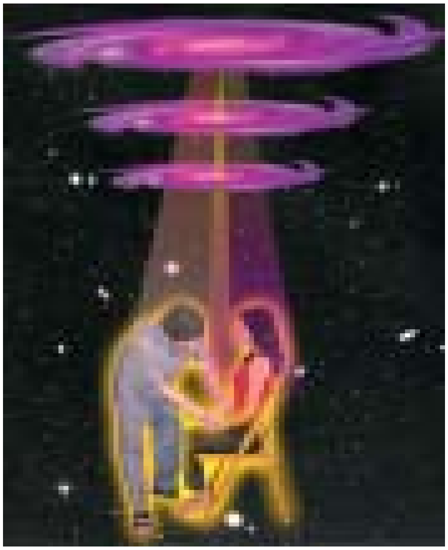

“Almighty Knife, please destroy all kinds of wounds, warts and tumors”. Cut up, down and sideways in the cross pattern through the part that has the ‘condition’.

Fig. 6.4 *“Almighty Knife, please destroy all kinds of wounds, warts and tumors”.*

The almighty knife can be used for any problem in the body. Follow the same procedure to carry out ‘energetic surgery’ on all parts of the anatomy. It is especially good for soft tissue repair, including the brain.

When you feel that the Chi Knife is full of sick Chi, spiral counter clockwise and discard it down to the ground. When you feel that the knife is ‘empty’ charge it with Chi from the universe.

B. Energizing and Activating the Immune System

In cases of severe illness, you will want to do this technique many many times, because it is essential to stimulate the body's defence-systems. This technique is also beneficial for debilitation and infection. You may also want do this in a preventative manner.

Activating the immune system always involves the bone structure, especially the round bones in the arms and legs and the flat bones in the skull, jaw, rib cage, hips and spine. All of these are involved with the production of white blood cells. The **thymus gland** and the **spleen** are of special importance to the immune system, because they are involved with programming and training the white blood cells (called T-Cells) to do the right job. **The lymphatic system is involved in removing toxic waste from the body; it is therefore beneficial to stimulate the lymph flow as well.** This is best done in combination with Chi Nei Tsang. *For details refer to the Chi Nei Tsang Books.*

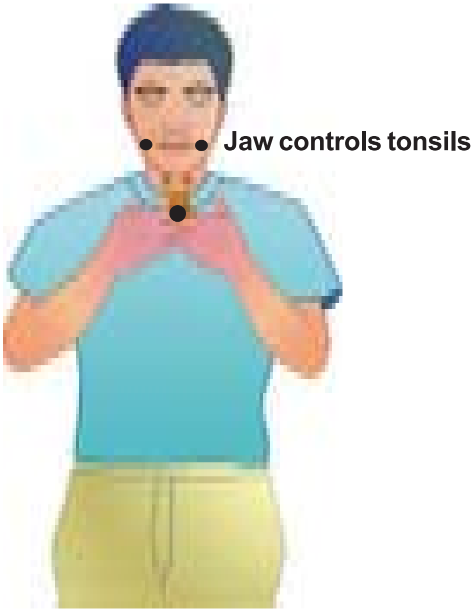

Fig. 6.6
Touch the top of Sternum to activate the Thymus.

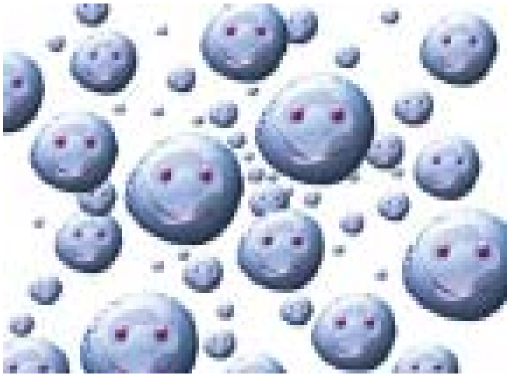

Fig. 6.7 *Happy Cells.*

If you work on a one-to-one basis, first perform the general cleansing session, where you 'pour energy' over the other person. After that use your hands, drawing Chi and pushing it through. Remember to push the energy all the way through and connect it with the universe beyond. Ask your student to concentrate on the areas that you are treating, preferably by guiding them, touching each point as you go.

Outline of Practice

Guide the student as you practice together. Work with your own energy body and tell the student to work with theirs. Presuming that you have already done the previous set of general healing, you can start with the **coccyx**.

1. Be aware of your Tan Tien and connect to the universe. Direct the Chi down to the fingers. Touch your coccyx with the fingers concentrating a beam of white light on to it. Picture the sacrum opening and the white light streaming inside, flowing up the spine and streaming out over the whole bone-structure. Transfer the feeling of Chi in the coccyx up to the universe, multiply it and guide it down to your energy body.
2. Always be aware of the Tan Tien and the universe first. Attract the yellow light; spiral and condense the healing light into the tip of your fingers; move the fingers up to touch the sacrum. Send out a beam of yellow Chi into the sacrum, directing it up the spine. This will stimulate the production of red blood cells. If you are not sure whether you are able to send it right into the marrow, picture the marrow or emit the Chi from your fingers, while at the same time concentrating on the bone marrow of the energy body. Transfer the feeling of 'Bone Chi' up to the universe and down to the energy body.

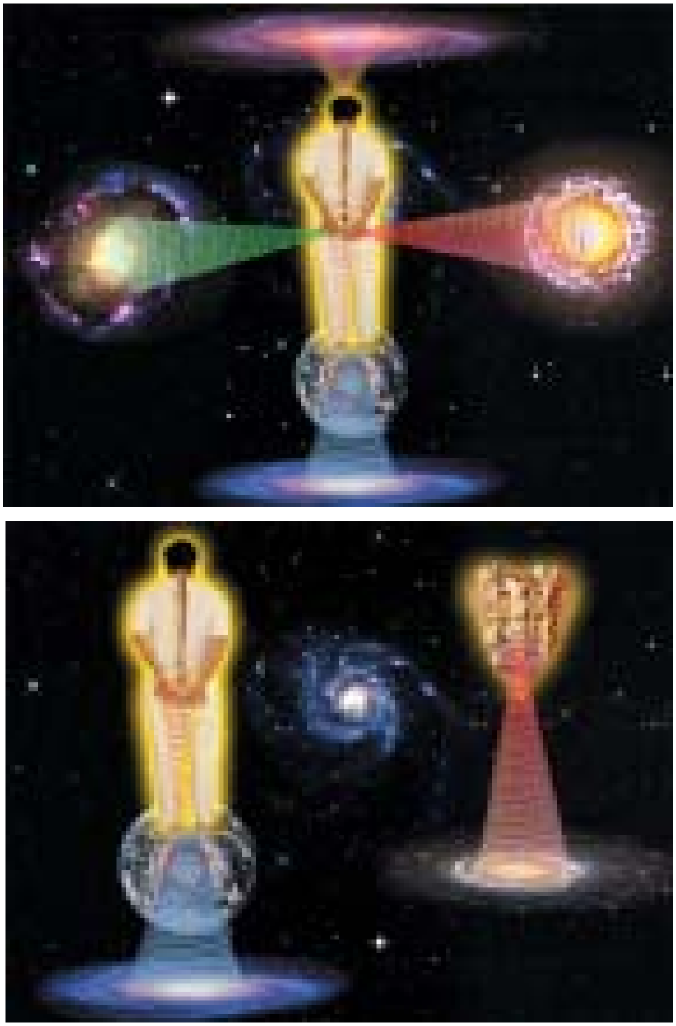

Fig. 6.9 *Touch the Sacrum.*

3. Proceed to the crown, showering it with violet light. Then spread from the crown throughout the bones of the body. Repeat several times. Instruct the student to feel the bone structure breathing. Ask him/her to extend up from the crown to the Big

Dipper and North Star, channelling down the red and violet light into the thymus gland.
4. Continue to the throat, stroking down with your palms, using green color to move the lymph down towards the backside of the heart. **(Never stroke up towards the head. Only guide the lymph down).**

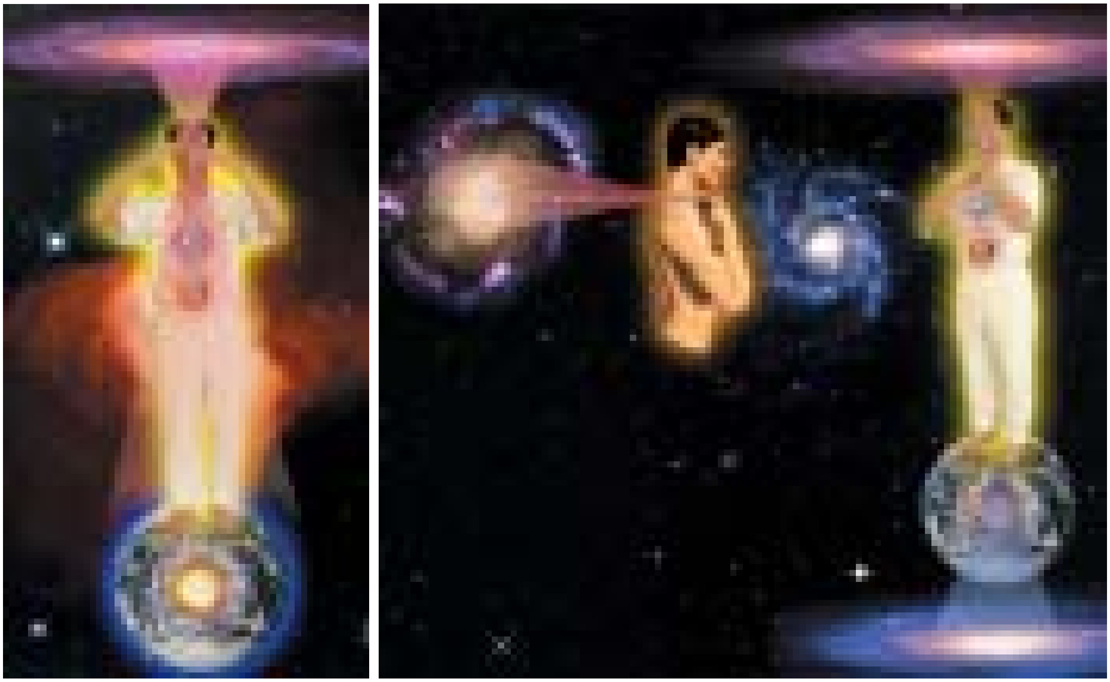

A. ***B.***

Fig. 6.10 A. *Channel the Violet Light of the North.*
B. *Use your Palms to bring the Lymph flow back to the Heart.*

5. Proceed to the lymph nodes under the armpits and across the sternum. Next activate the lymph nodes in the groin. From the groin you guide the lymph up to the navel center (Not down where it will get stuck in the legs).
6. Return to the navel center. Send the energy deep inside, energizing the lymph nodes at the back. Pull back, drawing green energy, flushing out any waste products that are stuck in them.

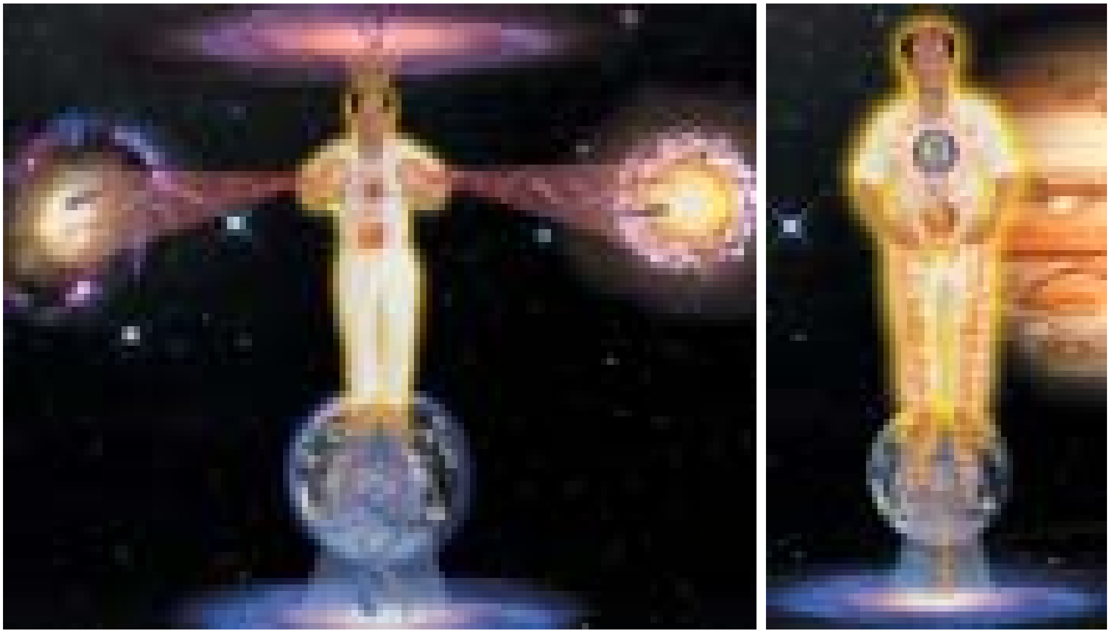

A. ***B.***

Fig. 6.11 A. *Activate the Lymph Nodes in the Armpits and Groin.*
B. *Energize and flush the Lymph Nodes in the Navel.*

C. Cleansing the Internal Organs Technique

Healing Some Common Illnesses

This session can be used for any kind of ailment and carried out with any number of students. It can take place on a singular, paired, small group or large group basis. Numbers can range from one person, one hundred, one thousand, ten thousand, ad infinitum. It is more likely that a greater energetic current will be created in much larger groups. In the case of 'mass' healings, there will always be a 'guide' or 'conductor', to open the channel for the students to follow. In the case of solo practices you will of course be the guide. **The following instructions have been set out on a one-to-one basis**.

By connecting to heaven, earth and your student, you will maximise the wholistic healing potential to your physical and subtle bodies. We recommend you spend about twenty minutes on the general healing before moving on to the more localized practices of emptying and cleansing the body's centers, systems and organs, contained within this chapter. We recommend that you spend between five to fifteen minutes on all of these. **Always perform the Sacred or Chi Water practice before.**

It is important to remember that if there is no energy in your Tan Tien, then there will be no energy to work with. 'Connecting the Tan Tien and the Universe' means connect your center to the infinite source of the universe, the Tao. When you make the connection you access this power. Always hold on to your center; it is the only thing that is truly yours. Keep the fire warm and spiraling. When you are aware of your Chi, you can direct heavenly light straight into your energy body and then into the student's.

General Principles

1. Push and pull the energy through the student's body. Push and disperse the sick energy into the universe and out to the planets, where it will be transformed and recycled.
2. Pull the energy back and stop it between yourself and the student.
3. Spiral the energy counterclockwise until you see the **green** light mix and clean the sick energy. Spiral and flush the energy down into the earth.
4. Spiral with **blue** light; see it mix and flush the energy within the student's body. Spiral this energy down to the ground.
5. Energize the student with **white** or **violet** light.
6. You can also use both hands to channel Chi when you 'push and pull' through the body.
7. Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral.
8. Move the hand anticlockwise with a spiraling motion to clean, flush and remove energy from the treatment area. After this, brush down the whole body, to direct the sick energy into the earth.
   Use an anticlockwise spiraling movement in order to clean, then expel the energy.
   Use a clockwise movement to bring energy in and an anticlockwise one to take it out.
9. When you begin pushing, pulling and sweeping, you may feel resistance in the body. As you progress and the energy centers open, this feeling will diminish.
10. There are many color combinations that can be applied to the various parts of the body. When you start practice, stick to the basic 'blue, green, white, violet and blue' formula. When you are working on a specific area, you can keep this book near by, or write a short 'color recipe' before you begin. In this way, you will be able to relax your mind.
11. 'Pushing and pulling' is like fanning and venting the body; when you fan your hands in front of your face, it will have a cooling effect. Ventilation enables the stagnant, sick energy to leave the body.
    Below is a list of the body's organs in relation to their tendency to 'store heat'. They are listed in descending order, i.e, the head is the hottest:

Head, heart, liver, lung, stomach and spleen, large intestine and small intestine, kidneys and sexual organs (which become hot when aroused).

Outline of Practice

Occupying a position amongst most of the internal organs, the solar plexus acts as a kind of clearing center for energy generated by the emotions. It also lies in between the lower and higher energy centers.

1. Scan; ask for the power to look at the cells.
2. Push, pull and give the command, “Disperse”. Push until you feel the solar plexus clear and open.
3. Spiral green Chi counterclockwise. Talk to the sick energy; tell it to “leave”, you will take it to “a better place, where it will be happy”. Give the command, “Out”.
4. Spiral Blue Chi counterclockwise, flushing out any remaining sickness. Guide the sick energy into the earth.
5. Spiral and sweep Blue-Green and Red Chi into the solar plexus, sweeping the organs. If the patient is expecting a baby, has an infection or bleeding in the intestinal area; do not use this method.

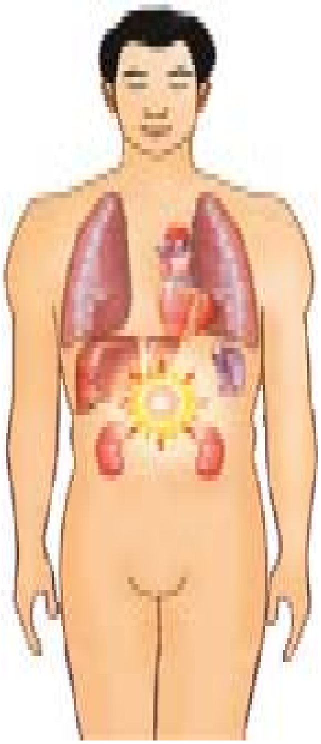

Fig. 6.12
Spiral Green Chi counterclockwise in the Navel.

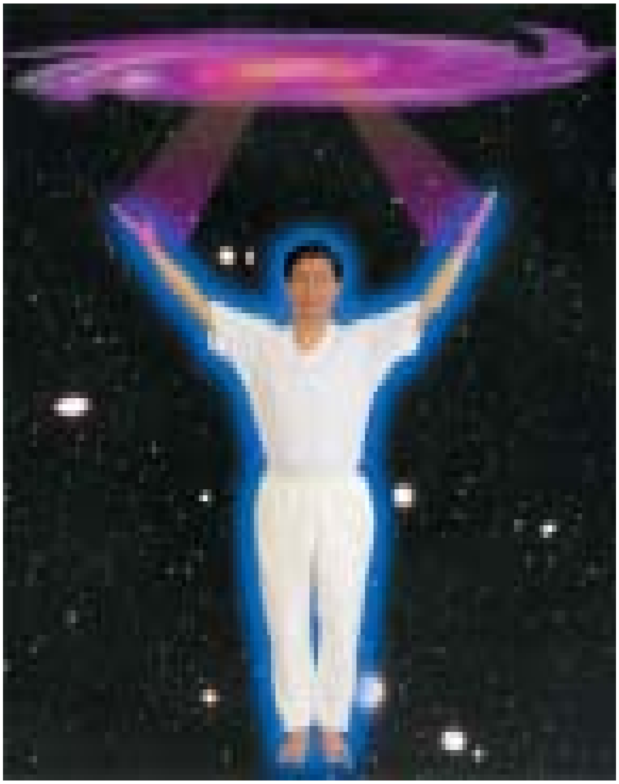

Fig. 6.13 *Charge your Hands with Chi - project the Hand Bones into the Universe and feel them fill and pack with Chi.*

Work on cleaning out the **organs**. Scan the person and see which parts are hot. A lot of people will have an overheated solar plexus, which means that this area will require a lot of attention. You will feel the heat when you scan.

You can **clean the organs by 'pushing and pulling'**. Picture the organ and push; feel your hands are very 'long', extending through the organs and reaching out to the universe. Project all the sick energy down into the earth and out to the universe. Use your intent to guide it into the earth and out to the other planets in the solar system, where it will be processed and recycled. Pull the universal energy back towards yourself; push and disperse all the negative energy out to universe. Pull; when you pull the Chi back past the student, stop in between the two of you. Push, pull, push and pull, until you feel the energy or power. **Ask; "Give me the power to see". Feel your Chi entering into the organs with ease.**

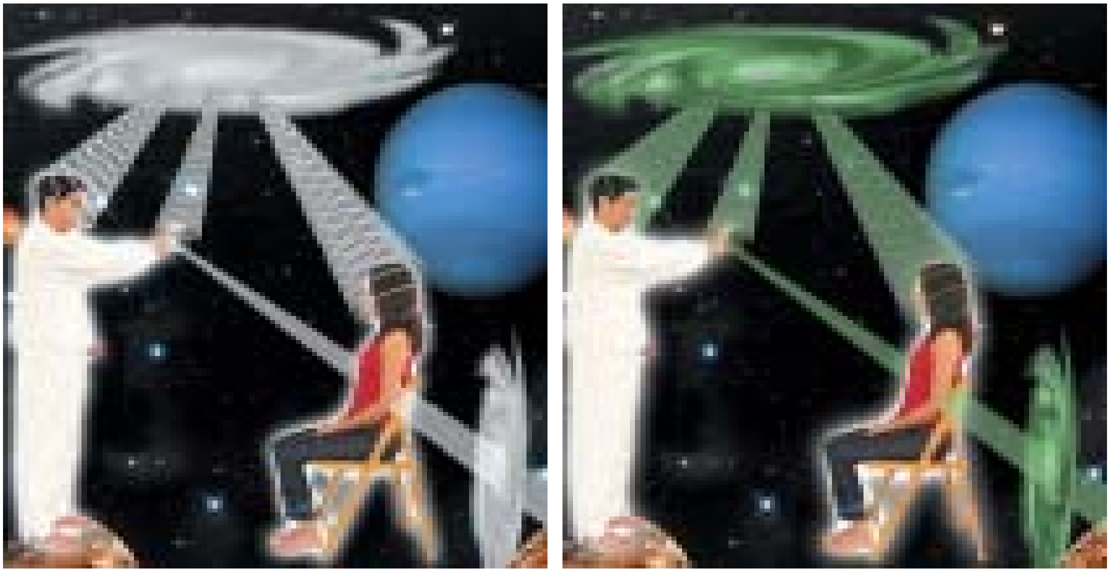

A. ***B.***

Fig. 6.14

A. *Extend through the solar plexus and reach out to the universe.*

B. *'Clean and energize with Green' - push and pull to help loosen the solar plexus. Spiral counterclockwise with a Green Light. Blend with the sick energy and guide it down to the ground.*

Clean with Green Light. Spiral counterclockwise. Feel the green light mixing with the sick energy. Push and pull until you feel the energy in the solar plexus begin to disperse. Spiral the energy counterclockwise and guide it down to the ground.

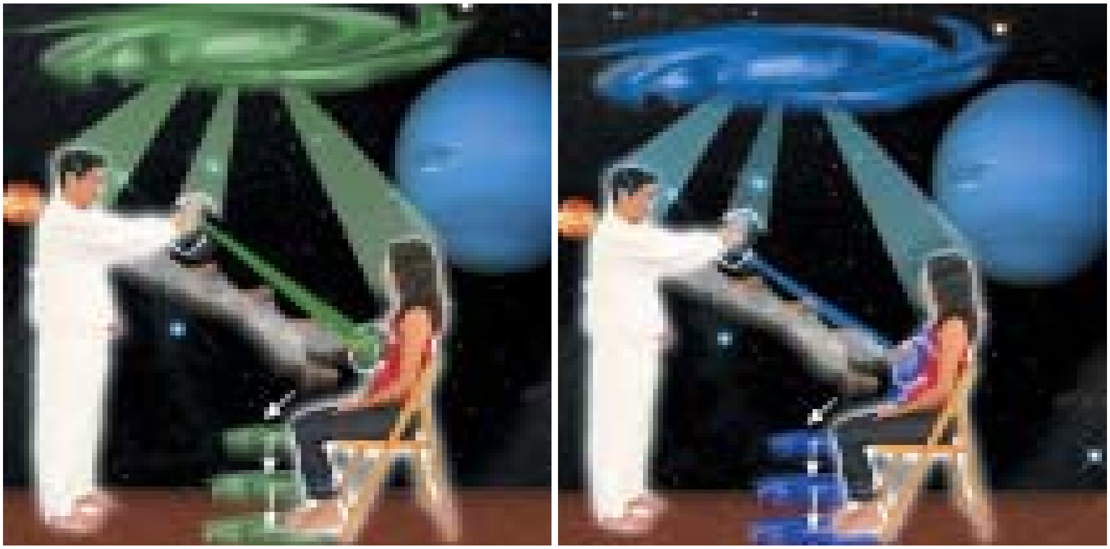

A. ***B.***

Fig. 6.15 A. *'Push and Pull' to help disperse the energy in the Solar Plexus.*
B.*'Flush and stabilize with Blue'-Picture Ocean Blue and flush sick energy down into the Earth.*

Rinse and flush with blue light. Picture ocean blue and flush down to the earth. See and feel the solar plexus becoming cleaner. Flush the sick energy down to the earth.

Now clean out the whole body again. With all the organs: from the lungs, heart, liver, gall bladder, pancreas, spleen, stomach, small intestines, large intestine, kidneys and sexual organs. Clean all the way down to the earth. **When you clean the solar plexus and all the other organs you remove all the negative emotional energy.**

Energize the **spleen**, **liver** and **kidneys** with blue or white light. Then project green, followed by orange light to the cores of the liver and kidneys. Clean with the green and orange light and stabilize with blue light.

Work on the **back** of the **spleen** and **kidneys**. Energize them with blue-white light. The spleen is on the student's left and kidneys on the right and left. They all help to clean the blood. Spiral the energy clockwise into the spleen. See it in your mideyebrow. See the cells and spiral the light into them. Give the command, "**Stay**". This is very important; it makes sure that none of the cells hide from the dazzling light that you are sending. Give the command again. Then stabilize them with blue light. Picture the kidneys and see their cells and feel the blue light go into the cells.

Work on the **Liver** on the student's right hand side. Smile and laugh to the liver and see all of its cells. Give the command, "**Stay**"; then send in the healing light. Send in blue or white (energize), followed by green (energize and clean), blue (rinse and stabilize), violet (recharge and reprogram) and then blue to stabilize again.

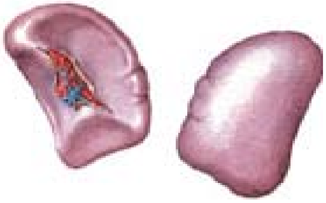

Spleen - Front and Rear

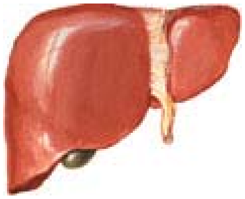

Liver

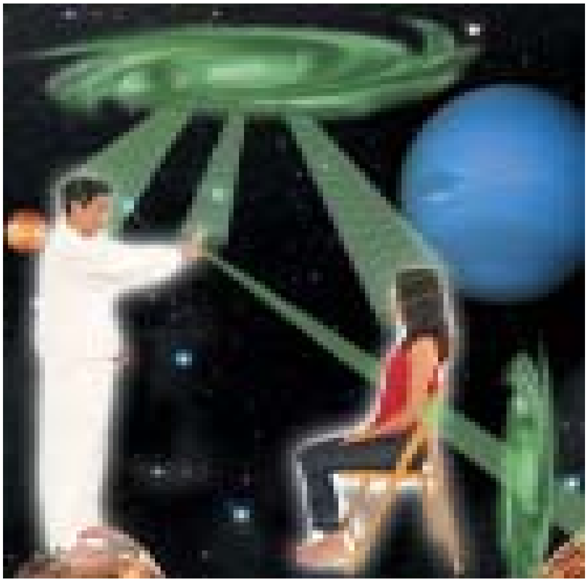

Fig. 6.16 *'Push and Pull' to help disperse the energy in the Liver and Spleen. Clean with Green and flush and stabilize with blue. Flush the sick energy down into the Earth.*

It is essential that you feel and talk to any sick energy. **If the student is healthy then you will feel their 'light'.** Ask their energy body to tell you where it **may** need some healing. Send in 'light radar'. Transmit green light and flush down to the earth. **When you feel that energy is accumulating, start to spiral more.** Push and pull and give the command for the sick energy to leave.

The more you clean, the more you heal. Tell your student to practice at home after the session, at an agreed time. You can then carry out 'absent or distance' healing, linking everyone's personal star, energy body and the universe. **With time, you will master the techniques, imprinting them with your own manner.** Eventually symptoms disappear, as the blockages are cleared and the flow returned.

Cleansing the Spinal Cord

Always cleanse the spinal cord, which protects the entire body from diseased and negative energy, both physical and psychic. When you open the spinal cord and solar plexus you can release many energy blockages. Sweep the hand in a clockwise motion to energize the spine with either/or a combination of; **blue, green, white and ultra-violet**. Then spiral counterclockwise and **brush** the spine down to the ground. Continue sweeping until you feel that the spinal cord has cooled down. Remember not to use orange or red.

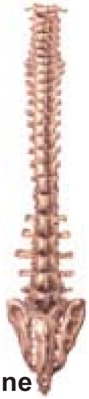

Spine

Cleansing the Blood (Lungs)

Pale green/white and tangerine/white chi work as cleaning agents for the blood. Diseases of the blood and arteries can be treated. The chi first energises the lungs and as the blood passes through them, it takes in the chi, thus becoming purified. The benefit is then carried to the rest of the body as the blood circulates. Serious infections can also be treated in this way; however, the technique should not be used on women who are pregnant.

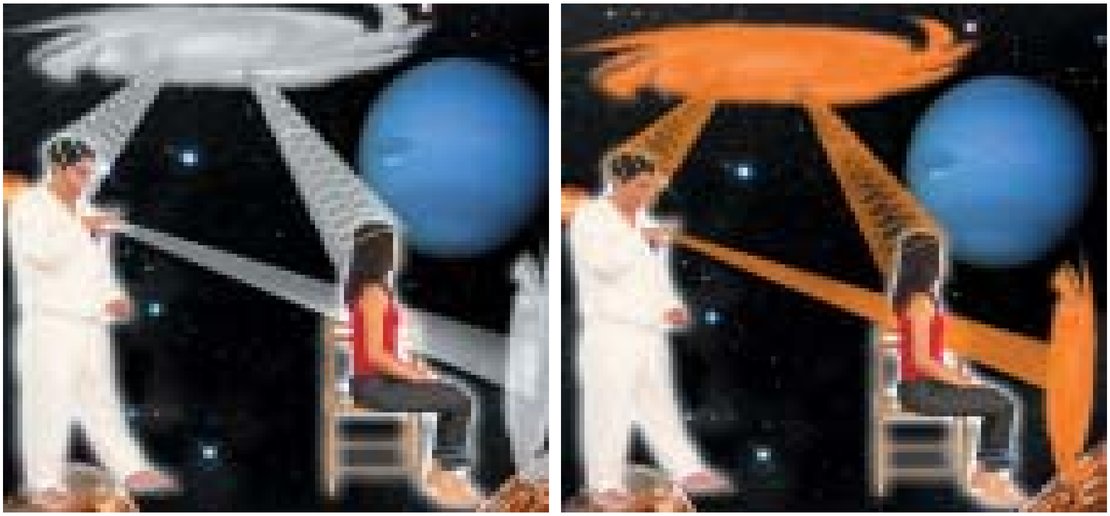

A. **B.**

Fig. 6.17 A. *Cleansing the Blood and emptying the Lungs.*
B. *Become aware your Three Tan Tiens. Connect to the Universe and ask for 'the Power to See'.*

Outline for Cleansing the Blood/Lungs

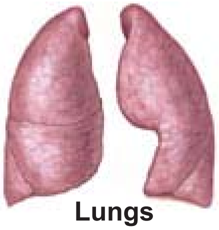

Lungs

1. Scan the student.
2. Push through and open the lungs from the back to the front.
3. When the lungs need to be given energy, use pale green/white followed by pale tangerine/white chi beginning at the front and moving top the rear.
4. To make weak patients stronger, cerise/white chi can be employed as this will bring energy into the lungs.
5. Purification of the blood is carried out by the liver, kidneys and spleen. Use white chi to bring energy to these areas, after first pushing and pulling through them.
6. The blood requires a lot of cleaning on a regular basis.
7. Energize with white Chi. **Germs can only grow in the dark. 'Fill yourself with the light'.**

Circulatory/Heart System General

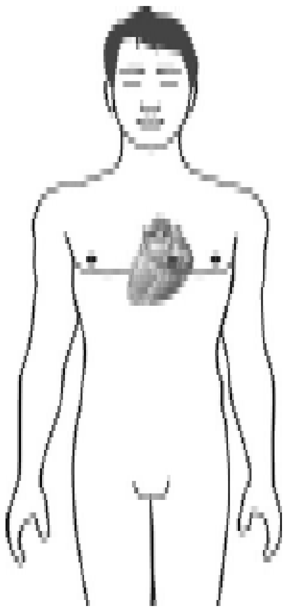

One of the basic problems pertaining to the circulatory system and heart is hypertension and hypotension. Once again, the sacrum and solar plexus are involved. The sacrum controls the bones; indirectly, it stimulates the marrow and enhances the quality of the blood. Good blood is like running the highest quality motor oil through the engine of a motor vehicle.

(Within the Universal Tao System we also have the view that the sacrum controls the Chi-pulse. If there is enough Chi, the heart does not have to work so hard, because the Chi pushes the blood).

The solar plexus relates to the emotions, affecting the heart through two mechanisms:

- It may affect the heart directly, causing rhythmic disturbance or pressure around the heart muscle
- Malfunction in the liver or digestive tract, leading to high cholesterol levels or gas which pushes up towards the heart from the intestines and causes unwanted heat and pressure.

Cleaning out the Solar Plexus will greatly improve these conditions. If you have been trained in Chi Nei Tsang you can use it in combination with the Cosmic Healing techniques.

Specific

Working on the Heart Muscle:
Clean out the Solar Plexus as outlined above. Push and pull with green and blue.
Detoxify the liver, flush with green.
Enhance with white and violet light.
Use the finger to work on the lower left and upper right part of the heart.
Remember that the heart muscle is very delicate.
You draw the force, you spiral and you push.
Stabilize with white.
For enlarged heart use blue.
Carefully use the red light to balance the blood and open the blood vessels.
Sweep the sacrum with green, red and blue.
Sweep the solar plexus with green, blue, red, white and violet.
Energize and cool down the throat center, using blue.
Working on Hypertension (High Blood Pressure):

Overactive adrenals/ kidneys often cause hypertension. Clean out the solar plexus first, because the adrenals (T11) are connected to it. The adrenal glands produce adrenaline, the hormone that stimulates the heart to pump. Overactive adrenal glands will eventually affect the heart.
Tune-up the Door of Life with blue, calming it.
Take the pressure from the heart, use green.
Energize the heart, use violet.
Flush down from head to feet with blue;
The basic problem with hypertension is that the Chi goes up and cannot come down. Work on the sacrum, using orange and yellow.

Working on Hypotension (Low Blood Pressure):
In hypotension the Door of Life and T11 are underactive.
First clean out the Solar Plexus, this time focus on removing blocks or stuck energy.
Energize the Door of Life, using green and red.
Energize the Sacrum (same colors).
Energize the Base of Skull with green and violet.

Stomach Pains

Pass your hands over the patient and brush the whole abdomen, especially the solar plexus, navel and lower abdomen. The navel should be treated with a combination of green and blue, and blue should be used at the end to firm up the energy.

If problems persist: A) Treat with Chi Nei Tsang B) Seek the advice of a qualified medical practitioner.

Relieving Pain: *Blue Chi*

Chi that is associated with the colors blue and green can be utilized in different ways. While green chi acts to free up energy that has become ‘sick’ and ‘stuck’, blue chi can act as a painkiller because of its ability to reduce the severity of pain.

1. Push and pull the through the affected part out to the universe or ground the sick energy until partial relief is attained.
2. A large amount of pale blue/white chi together with pale green/white chi should be used to transmit energy to the problem area.
3. For the alleviation of pain, use pale green/white chi and pale tangerine/white chi in turn.

Energize means draw the Chi into the part that you want to receive healing. Cleansing means push energy through the affected part and send it out to the universe. Pull the new Chi back into the same area. Light green Chi is used in loosening and cleansing delicate organs like the eyes and the brain.

Headache: *Light Whitish Green and Light Whitish Blue Chi*

1. To check for areas that are congested or lacking in energy, the eyes, temples, spine and solar plexus area should be scanned. Also, the crown, mideyebrow, forehead and back of the head and neck.
2. The eye and temple areas should be brushed down when the headache is because of eyestrain and then flushed into the ground.
3. The whole head, particularly the painful section should be brushed and then pale green/white, blue/white and pale purple/white Chi used to bring energy into the area.

Migraine Headache: *Light Whitish Green Blue and Violet Chi*

1. Persons who suffer from prolonged or chronic headaches including those types termed ‘migraine’ headaches have an abnormal accumulation of energy in various parts of their bodies. Starting with the solar plexus, this energy follows a route up the vertebral column and thence into the neck and head. As a result the various blood vessels in the head expand, causing pain in the head area. Mental strain and emotional disturbances usually cause this type of headache.
2. Use a pushing movement to clean and open the solar plexus and liver areas. First use pale green/white, then pale blue/white and finally mauve/white chi to bring energy to the solar plexus area.
3. In order to send the ‘sick’ red chi into the ground, brush down the spinal column.
4. The back of the heart can be opened by pushing through in order to remove the dirty Chi. Then bring energy to the heart by using pale green/white and purple. Use your mind to see the heart center expanding and opening.
5. Brush down the dirty chi from the head area and send it into the earth.
6. Use a combination of pale green/white, blue and purple to bring energy to the back of the head, the crown point and the forehead.

Toothache: *Light Whitish Green and Blue Chi*

1. Brush down the area with pale green/white Chi.
2. Energy should be brought into the area with pale green/white, mauve/white and blue Chi.

Broken Bone: *Orange-Yellow Chi*

For anything relating to the bones always activate the sacrum. Light orange-yellow Chi is used to encourage broken bones to knit quickly, Chi which is pale yellowish tangerine may be employed. The sacrum attracts the orange-yellow Chi; feel the eight holes in the bone breathing. The orange-yellow Chi enters, transfers up to the crown and is sent up to the universe to multiply.

1. The fracture area should be brushed using a combination of pale green/white and pale tangerine/white chi.
2. To promote fast knitting of fractures, a daily treatment with pale tangerine yellow Chi mixed with white can be employed, in the ratio of 7:1.5:1.5 of white (for the center), yellow and tangerine respectively. Push through and brush the sacral and navel areas, using a cerise (light red) Chi mixed with white.

Back Injury: *Light Whitish Blue and Green Violet Chi*

1. Scan the student to see which parts are congested.
2. Use chi that is pale white/green or pale white/tangerine and, selecting each in turn, with a brushing movement apply to the whole spinal area and especially to the area where the trauma is located.
3. To bring energy to the damaged area, use a combination of pale blue/white, green and purple.
4. White chi should be used with a brushing movement on both the back and front of the solar plexus area.

Infection and Inflammation: *Light Green Chi*

Treat the damaged area with green and tangerine/white chi, using a brushing movement. For the heart and spleen, these colors should be replaced with green and purple.

Bring energy to the area by the use of mauve, white, tangerine, green and blue and strengthen the natural immune of the body.

Cysts: *Light Green and Blue Chi*

During the course of the session, repeatedly pass your hands over the patient's body. Treat the solar plexus area with a combination of green and blue chi. The cyst can be removed using a chi knife, spiraling clockwise with green and tangerine chi; moving the 'stuck' red chi out of the body and grounding it in the earth. After, blue chi should be used for a calming effect, while green and tangerine can also provide energy. For maximum benefit, sessions should be thrice weekly.

As foods containing a lot of highly spiced ingredients have a lot of red chi, these should be kept to a minimum. The patient should use the Inner Smile to calm and balance his/her emotional state.

It is sometimes necessary to sweep 100-200 times, to cool the affected area. Scanning will enable you to take a temperature reading. If the area is still hot then continue sweeping until it reaches a satisfactory temperature.

Fever

1. Brushing can be done with blue or green Chi.
2. Push and pull through the affected part, sweeping the spleen with green and blue. Clean and rinse thoroughly.
3. The solar plexus area can be treated with tangerine and blue Chi but tangerine should not be used on a patient with diarrhea.
4. The thymus gland can be stimulated by the application of green, tangerine and purple Chi. First, brush the heart area with green and blue. This helps the body to ward off infection.
5. Both lungs should be brushed and energy brought into the back of the lung area by the use of tangerine and green.
6. Sweep the spleen with green.

If the student has a fever then they must lie down to receive the healing. This applies in all cases where the student is unable to sit up comfortably.

Insomnia: *Green and Blue*

Inability to sleep may be the result of too much activity in the solar plexus and the sacral area. The solar plexus may also be congested. Similar problems can also occur in the throat, mideyebrow and forehead, as well as at the crown point.

Repeatedly pass your hands over the patient's body as the treatment progresses and use a combination of blue and green chi to clean the sacral and solar plexus areas.

Sleep can be brought on by use of blue, as it soothes the solar plexus area. Use blue and green to brush the mideyebrow, forehead and crown point and white and tangerine on the navel. Finish with blue.

Do not eat food late in the evening.

Sleep in a separate bed from your partner. You can place two beds next to each other.

Tinnitus

Relates to weakened kidneys. Work on the kidneys, mideyebrow and forehead.

Blood Clotting: *Blue Chi* - Fresh Burns: *Green and Blue Chi*

Because of the cooling and calming properties of blue chi, it can be used to remove the heat from burns. Green has a similar effect.

The area needing treatment should be scanned, then brushed with a combination of pale green/white and blue Chi until relief is felt and the heat in the cells released.

Old Minor Burns: *Green and Red Chi*

Pass your hands over the damaged areas and repeat this during the treatment period. Use pale green/white and tangerine, but do not use on the head, eyes or brain, or other sensitive organs.

To reduce discomfort in the damaged area and promote fast healing, firstly pale blue Chi mixed with white should be employed, followed by cerise mixed with white, and green in equal proportions as this mixture optimizes the breaking down of necrotic cellular material. However, cerise should not be used in the head area but be replaced by mauve mixed with white.

Use cerise/white chi to brush the sacral and navel areas and follow with blue which assists in the release of the stagnant Chi. This session should be repeated for 2 or 3 days.

Old Severe Burns: *Green and Red Chi*

1. Scan at the beginning and end of the treatment (Note:I do not think scan can be seen as plagiarism and I don't know what other word to use in this context as I do not have the full text)
2. The local area should be brushed first with pale green/white and then with pale tangerine white. Keep repeating but make sure to omit sensitive organs and the head area.
3. To speed up healing and reduce pain and infection, the damaged area should be filled with energy from pale blue white, pale green white and mauve with white Chi.
4. To help speed up the healing process, wait a few days until there is much less pain, then use pale green white and cerise mixed with white Chi. If the head area is affected, use mauve mixed with white in place of the cerise.
5. Brush in the sacral and navel areas and to speed up healing use white Chi.
6. After making your healing intent stable, release it.

Old Wounds: *Green and Red Chi*

Green chi, in equal proportion of pale green/white and pale red/white should be employed on injuries which are not new as, when used in large quantities, it attacks and disperses damaged cellular matter, promoting fast healing. On recent injuries, it will have an adverse effect because it is likely to cause the trauma area to become moist, and it also will not be able to work quickly.

Scan at the beginning and end of the treatment, and brush the damaged area with pale green/white and pale tangerine/white Chi. Some pale blue/white can be employed locally to energise the area. Use of pale green/white followed by cerise mixed with white can speed up the healing process and assist in the breakdown of cellular material.

The navel and sacral areas should be brushed using pale cerise/white Chi. This will fill the area with energy and make the wounds heal faster. Before the Chi is let go, it should be stabilized.

Some Specific Points and Ailments

The following descriptions are mere guidelines. Each case is different and much depends on your own abilities. One needs practice to detect diseases. Best results with Cosmic Healing will be obtained when practiced in combination with Chi Nei Tsang.

Remember that these techniques are not meant to replace the doctor, but merely to assist him. The student must always work with the practitioner. ***The practitioner is a channel, the student the healer.***

Solar Plexus: *Blue Chi*

The Solar Plexus is the connection point for practically all the nerves inside the body. That is why it becomes laden with emotions so easily. When the Solar Plexus jams up and overheats, the liver will overheat; when this happens, the heart will also overheat; this will in turn affect the thyroid and parathyroid; finally the brain will be affected. Cleaning out this center is therefore extremely important. This type of cleaning may be done from the top down (as in the general cleansing session) or aimed at the Solar Plexus directly. Draw the green light, spiral it, flush with the green light, then energize with white and violet light. As before, always allow the energy to stream through both the points fully. The Solar Plexus holds all the emotions. When working on the Solar Plexus, the most important thing to remember is to connect the backside to the universe. There is literally no end to the 'other' end of this connection. Just allow the Chi to come all the way down and slightly pull inwards a little. Simply clean out the path. Allow the information to condense, allow any images to come out. Then stabilize the energy. Picture the Chi-field enveloping the person as a big protective bubble. Cool down with blue.

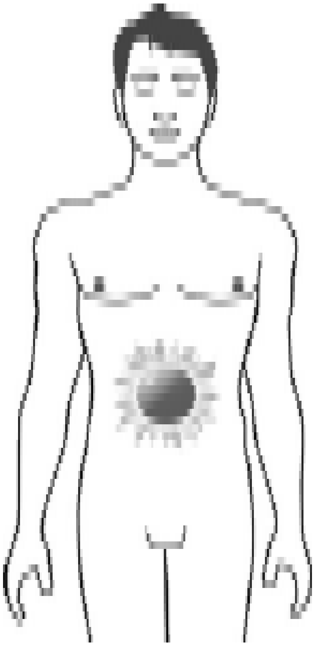

Skin Infection: *Blue Chi*

The skin is related to the lungs and the kidneys. Whenever there is an infection, the immune system is involved.

1. Enhance immunity as outlined in pages 101-108.
2. Focus on strengthening the lungs. Always start with green to flush out. Then send orange light into the lungs. Make sure you are not sending it up to the brain. Energize the lungs with blue and white.
3. Detoxify the kidneys, purify the blood (green) and energize with blue. Ask the person to visualize a fresh mountain-stream to stimulate the kidneys.
4. Treat the area locally, using your hands. If the area is big, use your palm, if small use the fingers. Do not focus on your palms - refer to the section 'Forget About Your Palms'. on page 152. Draw green Chi from the forest, and then energize with violet.
5. To diminish pain or itch, use blue to cool down. Also use the Chi knife to work on the infection.

In case the skin problem is allergy-based, you should also treat the mideyebrow; there might be some irritation here. First clean with green, then either energize with violet (if resistance is low) or soothe with a light shade of yellow. If the skin problem is related to asthma, open the lungs with bright red clean the throat center with green and energize it with (deep) blue.

If the problem is due to toxicity, also clean out the liver flushing with a combination of green and orange. Energize with green. Enhance the immune system by working on the sacrum.

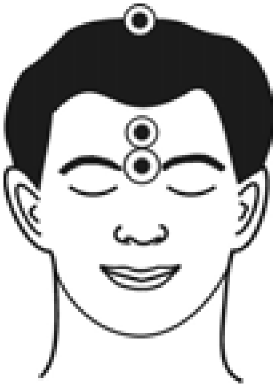

Eyes: *Green Chi*

The right and left eye are governed by the forehead and mideyebrow respectively. The eyes are also linked to the temples. Ask the person to touch their temple bone and guide them to take in the (white or light violet) Chi. Send Chi through the mideyebrow and forehead (all the way through). Ask the person to visualize the light coming in. Use soft shades of green with golden yellow to stimulate the optic nerve. Use blue to calm the eye muscles and light violet to stimulate the inner eyesight.

The eyes are connected to the liver. Scan the liver and energize with green. Scan the jawbone to see if there is any unreleased anger. Cool down with blue. Channel bright green into the liver.

Ears: *Violet Chi*

The right and left ears are governed by the forehead and mideyebrow respectively. Push and pull green Chi to clean. The ears are connected to the kidneys, so you can work on them also. The ears are also connected to the sinuses and upper respiratory tract. This means that infections need to cool down before you can use any cleaning technique. Use blue to cool and clean with violet. As long as fever and severe infection persist, use blue to cool them down. When clean, energize with violet.

Have Better Health, Less Stress and Remain Young

Brush specifically on the areas of the solar plexus, liver, stomach and pancreas.

White chi brings energy to the crown point, head (at the back) and both sides of the brain.

Both the front and rear of the heart and lungs can receive energy from white/purple Chi.

Both the front and rear of the spleen, the navel and lower abdominal area, the spinal cord, the kidneys, using white Chi.

On the sexual center use white chi but if the problem is difficult to shift, use green. You can also use white Chi on the sacral area, and green or purple on other organs.

White Chi can be employed for the arms and hands, while the legs and feet should be treated with white and purple Chi.
Eye problems with green/yellow energy.

After a session, patients who are particularly weak should not bathe for one full day. The energy which has been put into their bodies can thus be kept there by application of pale blue/white Chi at the end of the session.

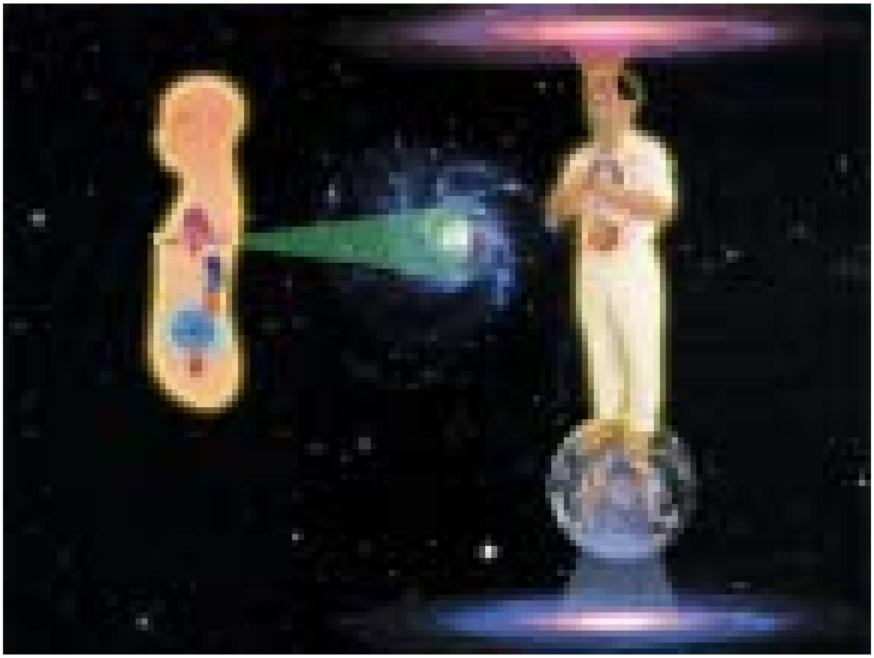

Fig. 6.18 *Use Compassion, Love and Kindness.*

Summary of Individual Healing Session

A. Chi Knife
B. Energizing and Activating the Immune System
C. Cleansing with the Internal Organ Techniques

General Outline of Practices

1. Gain experience by working with one person. **Stand in front of them with your feet shoulder width apart.** Hold your hands at your heart in salutation. Transmit your love and compassion and ask for their permission to help them to help themselves heal.

2. Connect to the student's and your own cosmic star above your head and ask for universal permission. **Radiate your love and transfer three minds into one mind**. Lower your Upper Tan Tien. Activate your consciousness: Three minds become one mind. Expand your awareness in the mideyebrow and the crown. Feel the Tan Tien, heart, mideyebrow and crown spiraling. Expand to the universe.
3. Visualize the person's legs as being 'very long', extending down to the earth. **Activate the Sacred Fire** and bring it down. Clean out the room and make a big Chi field around you and the student.
4. **Call the Elemental Forces:** water, fire, thunder, lightning, rain, lake, earth, mountain, wind and Heavenly Chi. Call the guardian animals; Blue Tortoise from the north, Red Pheasant from the south, Green Dragon from the east, White Tiger from the west, Yellow Phoenix from above and Black Tortoise from below. Feel that you are protected by the Chi; gather the power of all the mountains, rivers, streams, stones and rocks into the room. Activate them with lightning, thunder and wind - filling the whole room with electrified Chi.

A. Chi Knife

1. Hold the knife in right hand (left, if you are left handed). Hold the other hand near your body with the thumb and index finger touching. Project the Chi Knife as big as the universe, 'long and big - reaching to Heaven'. The bones in your arm are hollow and the 'bottom' is sealed at the shoulder. Feel them 'fill and pack' with the power coming from above.
2. Then give the command: **"Power comes from the east. This Chi Knife has the power to dissolve all the negative energy, sickness and bad fortune. By my request, please carry out the order now."** Wait until you feel that your arm and the knife are heavy and full of Chi.

   Carry out energetic surgery, **cutting and slicing the affected area with cool yin energy**. When you feel that the Chi Knife is full of sick Chi, spiral counterclockwise and discard it down into the ground. When you feel that the knife is 'empty', charge it with Chi from the universe. Continue until you feel that the area has been cleansed.

B. Energizing and Activating the Immune System

In cases of severe illness, debilitation and infection, it is essential to stimulate the body's defence systems. You may also want do this in a preventative manner. Guide the student as you practice together. Presuming that you have already done the previous set of general healing, you can start with the coccyx.

1. Be aware of your Tan Tien and connect to the universe. Touch your **coccyx** with the fingers concentrating a beam of white or yellow light on to it. Picture the sacrum opening and the light streaming inside, flowing **up the spine** and streaming out over

the whole bone structure. Feel the Chi rising up to the universe, multiply it and guide it down to your energy body.

2. Be aware of the Tan Tien and the universe. Attract the yellow light; spiral and condense the healing light into the tips of your fingers; move the fingers up to touch the **sacrum**. Send out a beam of yellow Chi into the sacrum, directing it up the spine. This will stimulate the production of red blood cells. Transfer the feeling of ‘Bone-Chi’ up to universe and down to your energy body.
3. Proceed to the **crown**, showering it with violet light. Then spread from the crown throughout the bones of the body. Repeat several times. Instruct the student to feel the bone structure breathing. Ask him/her to extend up from their crown to the Big Dipper and North Star, channelling down the red and violet light into the thymus gland.
4. Continue to the **throat**, stroking down with your palms, using green color to move the lymph down towards the backside of the heart. (Never stroke up towards the head! Only guide the lymph down).
5. Proceed to the **lymph nodes under the armpits and across the sternum**. Next activate the lymph nodes in the **groin**. From the groin you guide the lymph up to the navel center (Not down where it will get stuck in the legs).
6. Return to the **navel center**. Send the energy deep inside, energizing the lymph nodes at the back. Pull back, drawing green energy, flushing out any waste products that are stuck in them.

Activate the points at the inside of the **elbows** and the **knees**.
Shower the whole body with healing light and energize the **perineum** with blue Chi.
Always remember to ground any sick energy by sending it deep into the ground. Ask the student to do the same and it will be more effective. You should do this while giving the session as well as when closing it.

C. Cleansing the Internal Organs (Repeated from pages 132-133 and pages 136-137)

This session can be used for any kind of ailment and carried out with any number of students. **It is important to remember that if there is no energy in your Tan Tien, then there will be no energy to work with.** When you are aware of your Chi, you can direct heavenly light straight into your energy body and then into the student’s.

General Cleansing Procedure

1. Push and pull the energy through the student’s body. Push and disperse the sick energy into the universe and out to the planets, where it will be transformed and recycled.
2. Pull the energy back and stop it between yourself and the student.
3. Spiral the energy counterclockwise until you see the **green** light mix and clean the sick energy. Spiral and flush the energy down into the earth.

4. Spiral with blue light; see it mix and flush the energy within the student's body. Spiral this energy down to the ground.
5. Energize the student with **white** or **violet** light.
6. You can also use both hands to channel Chi when you 'push and pull' through the body.
7. **Spiral in** - spiral clockwise to bring energy into the body and energize it.
8. **Spiral out** - spiral counterclockwise to cleanse, flush and remove energy.
9. When you begin pushing, pulling and sweeping, you may feel resistance in the body. As you progress and the energy centers open, this feeling will diminish.
10. There are many color combinations that can be applied to the various parts of the body. When you start practice, stick to the basic 'blue, green, white, violet and blue' formula. When you are working on a specific area, you can keep this book near by, or write a short 'color recipe' before you begin. In this way, you will be able to relax your mind.
11. 'Pushing and pulling' is like fanning and venting the body; when you fan your hands in front of your face, it will have a cooling effect. Ventilation enables the stagnant, sick energy to leave the body.

    Below is a list of the body's organs in relation to their tendency to 'store heat'. They are listed in descending order, i.e, the head is the hottest.

    Head, heart, liver, lung, stomach and spleen, large intestine and small intestine, kidneys and sexual organs (which become hot when aroused).

Outline of Practice

The **solar plexus** is the clearing house for emotional energy and is situated in between the lower and higher energy centers and in the center of most of the internal organs. All the organs dump their negative emotions here. The internal organs can be cleansed by sweeping the solar plexus, energizing, cleaning and flushing it of sick energy. When you feel sick energy in the body, give the command: "Stay". This is very important; it makes sure that none of the cells hide from the dazzling light that you are sending in.

1. Scan; ask for the power to look at the cells.
2. Push, pull and give the command "Disperse". Push until you feel the solar plexus clear and open.
3. Spiral green Chi counterclockwise. Talk to the sick energy; tell it to "Leave", you will take it to "a better place, where it will be happy". Give the command "Out".
4. Spiral Blue Chi counterclockwise, flushing out any remaining sickness. Guide the sick energy into the earth.
5. Spiral and sweep Blue - Green and Red Chi into the solar plexus, sweep the organs. Do not use in cases where there is intestinal infection and internal bleeding (which can be worsened) or in the presence of pregnant women.

Cleansing the Spinal Cord

Always cleanse the spinal cord, which protects the entire body from diseased and negative energy, both physical and psychic. When you open the spinal cord and solar plexus you can release many energy blockages. Sweep the hand in a clockwise motion to energize the spine with either/or a combination of: blue, green, white and ultra-violet. Then spiral counterclockwise and brush the spine down to the ground. Continue sweeping until you feel that the spinal cord has cooled down. Remember not to use orange or red

Cleansing the Blood (Lungs)

The blood can be cleansed by energizing the lungs with whitish green, then with light whitish orange. The blood passing through the lungs will absorb the green and orange Chi, thereby cleansing the blood plasma and vessels and the rest of the body. This technique is very useful for treating ailments of the blood and of the arteries and for severe infections. Do not use it on pregnant women.

Outline of Practice

1. Scan the student.
2. Push through and open the lungs from the back to the front.
3. Energize the lungs from front to back with light whitish green, then with light whitish orange. Push through.
4. If the student is weak, energize the lungs with light whitish red, which will provide a strengthening effect.
5. The **spleen**, **liver and kidneys** purify the blood. Push and pull through the organs and energize with white Chi.
6. The blood requires a lot of cleaning on a regular basis.
7. Energize with white Chi. **Germs can only grow in the dark. 'Fill yourself with the light'.**

General Principles

Energy can be introduced to the affected area of the body by moving the hand clockwise in a spiral. **Flush** means to return the sick energy to the earth for recycling.

Stabilize means to return the area to normal functioning.

Spiral in a **clockwise** motion when you are '**stabilizing**'.

Use **blue or white** light to energize the area that you are working on.

Use **green** light to **energize** and **clean** the area that you are working on.

Use **blue** light to **flush** and **stabilize** the area that you are working on.

Use **violet** light to **balance** the sick energy and **reprogram** the cells of the body.

For **general infections** always **use green** (energize and clean) and **blue** (stabilize) light. Use violet light to balance the sick energy and reprogram all the cells. Use the ‘Chi Knife’ for infections as well. Never use orange on the head, heart and spleen.

Give the command **“Stay”** to make sure **all the cells receive the light** you are sending in. Remember that germs and bacteria cannot hide in the light.

Energize means to draw the Cosmic Chi into the part that requires healing.
Cleanse means to clean all the cells being treated.

You can also work on yourself. You may need some help when working on the back. If you obtain optimum health and practice the Cosmic Orbit regularly then you will never ‘need’ to apply any of these techniques to yourself.

Like all Chinese medicine, Cosmic Healing works on the system as a whole, balancing the entire organism, treating the cause to cure the symptoms.

The techniques really are simple.

When in doubt use; blue, green, white, violet, blue.
You can always use white and violet.
Upper organs - use violet.

You can practice group healings and then link your students’ personal stars together, at an agreed time later in the day. **You can then perform the World Link Cosmic Healing**.

Work on any other ailments that the student may have.
The simplest way to learn is to begin practicing today.
The simplest way to heal is to allow yourself to be healed.

Chapter VII
Different Hand Techniques

Forget about your Palms

An extremely important theory of Cosmic Healing Hand Techniques is to ‘forget about your palm’. This means that you should only use the palm (which is incredibly sensitive), for ‘guiding’ the universal energy into the right spot, sending out a minute beam of light, like a laser, that marks the place where the energy should be sent. Your palm is connected to your brain, which has been sunk into your Tan Tien. The light marker will serve as a beacon for the force that is directed down by your mind. Draw down the Universal Chi, guide it to the right spot and give the command; "Stay". This is very important, otherwise the Chi will disappear once you move your hand.

When you practice Cosmic Healing it makes no difference how many people you treat, because you are not transferring your own energy.

Don't Focus your Mind on the Sick Part

If you direct your attention towards something, your energy will go there. If you focus on some ailment, you are already sending out your energy. Instead you should try to connect with a greater source, Universal Chi, extending your mind very far away. This way you won't lose your power. The only thing you do as a practitioner is to form the connection between nature, the student and the sick energy or between the universe, the Chi field and the sick part.

Cosmic Healing is Teaching

One of the best methods of protection for a practitioner is the Fusion practice (*Refer to ‘Fusion of the Five Elements I by Mantak Chia*). Perhaps even better is for the practitioner not to perceive him/herself as a healer, but as a person that comes to guide people in something they have to do themselves. When you say; "I am a healer", the sickness or bad energy, has already found a new home. When you check it out, first it will try to hide, and then it will try to attach itself to you. But if you don't interfere, the energy will go straight to the ground. So it is very important to clarify that you are not a healer, but a practitioner that can help, and that the student or sick person has to do most of the job.

Fig. 7.1 *The Stronger your Awareness of your Inner Body, the Easier it will be for others present to pick up this Sense.*

If you guide those present, this will greatly enhance your healing session. Sit or stand facing your student. Any time you want him or her to focus on a particular point of the body, ask them to touch it with their hands. For example, "Now place your hands on your navel", while you do the same thing yourself. As soon as you feel the Chi there, you instruct them to focus on the navel themselves. This way you are actually helping them to deepen their bodily awareness. The stronger your own awareness of your inner body, the easier it will be for the others present to pick it up. This means that if you are treating some part of another person's body, or if you see that the person is not aware of any energy in some part of the body, you can help by focusing your awareness on this particular part of your own body.

Chi Kung Diagnosis: Hand Scanning

Hand scanning is a simple and effective diagnostic tool of External Chi Healing. It involves passing your palm over the body of your student from one inch to one foot above the surface of the skin and being aware of the energetic state of the various bodily areas. What one senses is described as the electromagnetic field, the aura or the energy body.

General Variations in the Energy Field

The variations you may sense with scanning that have diagnostic significance are:

Temperature

Heat generally indicates excess, while cold indicates deficiency. However, some areas of the body are naturally warmer or colder than others. This is described below under "Hand Scanning the Internal Organs."

Thickness

A feeling of thickness over the area would indicate excess, while thinness may indicate deficiency.

Wind

You may feel sensations of wind leaving the body at various places. This may indicate internal wind at a given location, or it may point to an area of the body that is "leaking energy" and needs to be ‘patched’ or ‘sealed’. It may also be a positive sign that you are driving wind out of the body.

Peaks and Valleys

Peaks in the energy field will feel like mountains or hills in the aural landscape. They may actually push your hand away. Valleys feel like depressions or energy vortexes drawing you in. They may also feel empty, like a hole or pit. Peaks indicate excess or stagnation, while valleys indicate deficiency.

Tingling

Accumulations of sick Chi may cause your hand to tingle, feel prickly, throb, or even hurt. These are excesses of energy and upon inquiry, the student may complain of pain in this area.

Hand Scanning the Internal Organs

Each Organ Emits a Different Aura

Each organ emits a different kind of force or aura through the skin. By passing a hand one or two inches above the skin, you can feel different sensations that reflect the condition or state of the internal organs. You need to develop the sensitivity to receive and identify the vibration or frequency of each organ. Practicing the Cosmic Chi Kung Meditations will help develop such sensitivity.

Liver and Gall Bladder Scanning

Healthy liver and gall bladder energy feels warm.

Negative Emotions: When you pass your hand over the liver, under the right side of the rib cage, you will feel a charged energy come up to your hand. This is a sign of anger in the liver.

Overactivity: When you pass your hand over the liver and feel a rush of hot energy, this indicates that the liver is overheating because of toxins or emotional stress.

Underactivity: When you pass your hand over the liver and you feel a dense and hot energy, the liver is weak, congested and sick.

Lungs Scanning

Healthy lung energy feels cool and dry.

Negative Emotions: Pass your hand over the lungs. Do the Lungs' Sound and listen to the echo of the sound as it rebounds from the lungs. Sadness will feel like a deflating ball pressed between your hands.

Overactivity: Energy that feels dry and hot indicates an overworked organ.

Underactivity: Energy that feels damp and cool under the scanning hand indicates underactive or congested lungs that can lead to respiratory problems.

Heart Scanning

Healthy heart energy feels warm and energetic.

Negative Emotions: Hot and charged energy indicates impatience, hastiness and arrogance in the heart.

Overactivity: If your hand detects hot, charged and overly expansive energy, this indicates that the heart and blood may be overheating.

Underactivity: Energy that feels cool and less expansive indicates an underactive or congested heart.

Spleen Scanning

Healthy spleen energy feels lukewarm.

Negative emotions: Energy that feel damp and sinking indicates excessive worry.

Overactivity: The energy feels hot and damp when the spleen is overactive.

Underactivity: The energy feels cool and damp.

Kidneys Scanning

Healthy kidney energy feels cold, but not too cold.

Negative emotions: Energy that feels cold and chilly indicates fear.

Overactivity: When the kidneys are overworked or overstimulated by excessive exercise or improper diet and liquid intake, the energy can feel damp, stick and hot.

Underactivity: When toxins are blocking the organs, the energy can feel damp and cold.

Hand scanning is an art, and may take a while for you to feel confident using it. Practicing Cosmic Chi Kung is one of the best ways to develop greater sensitivity in hand scanning.

Chapter VIII
Cosmic Chi Kung
Palm and Finger Training

Why We Practice the Cosmic Chi Kung

The Cosmic Chi Kung is also known as Buddha Palm. The reason we train with the Empty Force practice in connection with the Cosmic Healing is to learn how to project Chi through the space in the cosmos, the space between your hands and to project Chi through the space to your students. Usually when we touch the student we pick up sick energy. This practice will teach you how to ground the sick energy to the earth and to disperse it into the Universe to be decomposed and recycled by the planets.

Practice until you can project Chi out from your palms and fingers and when you feel the Chi coming into your body through your hands. The most important part is to always remain connected with the Universal Force coming from all six directions. If you focus on healing from your hands or your Tan Tien you will use up your own energy. You must be connected to the Cosmic Universal Force.

Expand your awareness to nature, the oceans, the lakes, the forests and mountains. Smile to nature and feel it smile back to you; inhale and draw the Chi into your palms. Expand your awareness to connect to the light, the Milky Way and the cosmos.

Be aware of the heart and the red light in the heart; expand your awareness to infinity. The light will come close to you; picture the red planet Mars above you. See it shine down to your crown, feeling the light in your palms. Be aware of the six directions, feel your body growing bigger; until you touch the sky, your feet still planted in the earth.

Opening Heaven and Earth Force in the Palms with Heaven and Earth Spirals

The purpose of this technique is to draw in the forces of heaven and earth through the palm. Activate the spirals, the symbol for the heaven and earth force in the center of the palm.

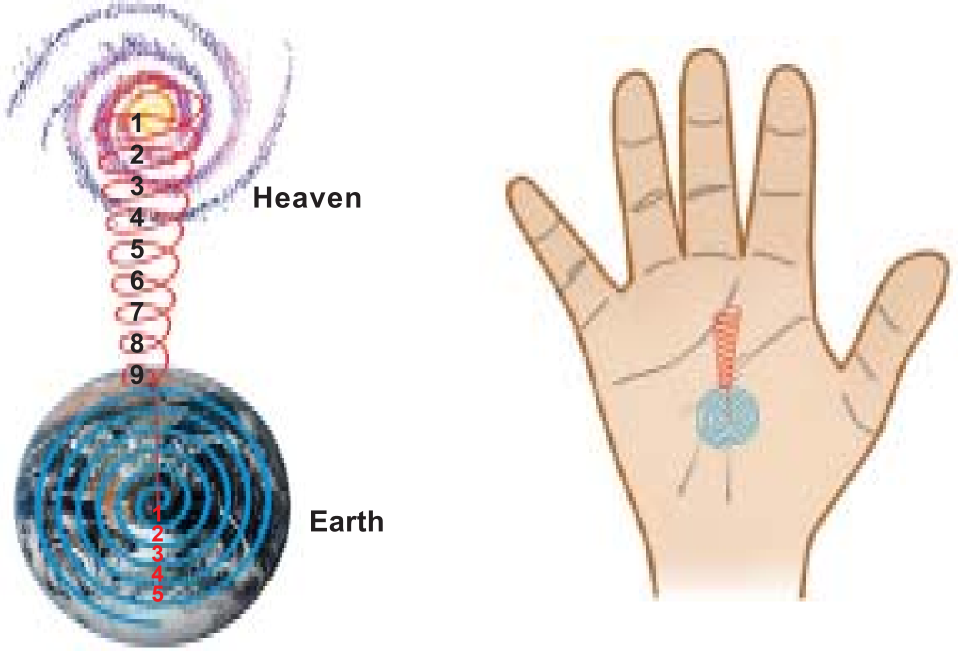

Fig. 8.1 *Enhances the symbol of Heaven and Earth Force, in the Laogong - Pericardium 8.*

This is accomplished by drawing these spirals with the palm itself. As you draw the spirals, you should actually feel the palm breathing, inhaling these forces in. In the beginning, however, you may only be able to use your own mental imagery and imagine that you are absorbing these forces. In time the feelings will replace your imagination.

You breathe in Heavenly Chi when you draw the heaven spirals, and breathe in Earthly Chi when drawing the earth spirals. You may draw in both in one session; you may also choose just to draw in either Heaven or Earth Chi if you feel you need more of one or the other. It is important to balance these energies whenever you do this exercise.

1. Begin by raising the right hand to about shoulder height with the elbow sunk and the palm facing forward. Slightly pull back (open) the index finger, and slightly stretch the thumb forward and down. This will activate the Laogong point on the palm.

   At the same time, touch the tips of your left thumb and index fingers together. This will allow the right hand to draw in the Heaven and Earth Chi. The left hand will then be held palm up to draw in the heavenly force, or held palm down to draw in the earthly force.

2. Begin **drawing the heaven spiral** by moving the **palm in a clockwise motion**, with the circles getting smaller and smaller, spiraling inward. At the same time, your left hand is held at your side, palm up with thumb and index tips touching each other. Draw seven or nine (planets of our solar system) small spirals with your right hand in this way, drawing in the violet heavenly energy.
3. Next, lower your right hand slightly and turn your left palm down to face the earth. Then begin **drawing the earth spiral** by moving the right palm in a **counterclockwise motion** with the circles getting larger and larger. Draw five spirals (five elements), drawing in the blue earth energy as you move. Inhale into the palm, feeling the energies.
4. You may now bring both hands down to the beginning position of Cosmic Chi Kung, or you may open the Cosmic Channel to the planets for healing. This is discussed in the next section, Part I, ‘Opening to Heaven, Earth and Cosmic Chi’.

Cell Breathing

By understanding the operations of our cells, we establish a foundation for our experience of Cosmic Chi Kung. Each cell is a living, breathing entity, absorbing energy with each inhalation and releasing toxins with each exhalation. Thus with every breath, we are creating more Chi.

We are made up of over one hundred trillion cells (one trillion is one thousand billion). Each one of these cells is constantly breathing. In order for our cells to function optimally, we must follow nature's teachings.

Fig. 8.2 *Each Cell is a Living, Breathing Entity.*

The first thing is to keep your cells free of toxins. Eat nutritional foods. This helps the Chi to flow smoothly through your energy channels and aids greater assimilation of vitamins and minerals. Learn and apply Five Element nutrition principles, eating according to the seasons and to your own constitution.

The second thing is to keep your cells free of excess emotions. Develop the skill of acting positively to situations, rather than reacting, especially negatively. Don't allow anyone to make you sad and unhealthy. Some people like to say, "You make me angry," or "You made me worry." We cannot allow other people to make us so emotional. We do not need to accept the negative emotions that other people try to dump on us. Cellular and psychic boundaries are related. Negative emotions disturb the cells' ability to absorb Chi and to generate the energy necessary for healing. We need all the Chi that we can get. Learn and apply the Six Healing Sounds, the Cosmic Inner Smile and the Fusion of the Five Elements meditation to transform negative emotions into personal power.

Third, create more energy by cultivating a happy attitude, so that your cells can regenerate more easily. **Smiling and loving** are the two quickest ways to make more Chi. In order to have an ample supply of Chi, develop an attitude of **joy and love** in yourself and towards others.

Mind, Eyes, Heart and Intention (YI)

The opening of the mind, eyes, heart and use of intention is vital for further activating and increasing your Chi. In the Tao we achieve this by cultivating our "Yi" (pronounced "yee"). "Yi" means mind power, intention, awareness and concentration. The Yi leads and guides the Chi. With our mind we control thought patterns. With our eyes we control the senses of sight, hearing, smell and taste. With our heart we control all our organs and their related emotions: the kindness/anger of the liver, the joy/impatience of the heart, the openness/worry of the spleen, the courage/sadness of the lungs and the gentleness/fear of the kidneys.

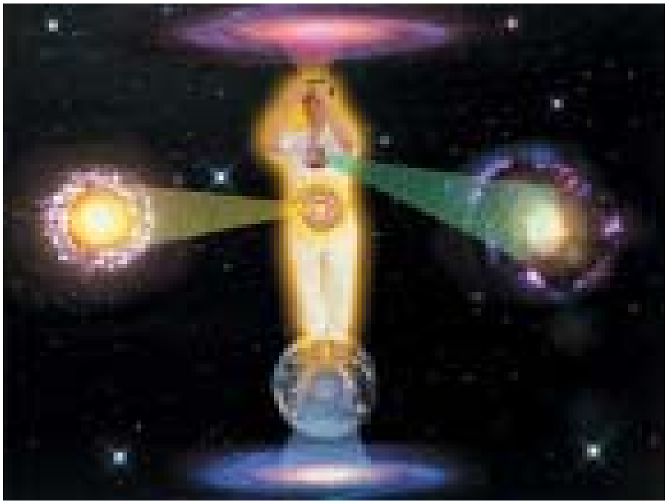

Fig. 8.3 *'Yi' leads and guides the Chi.*

Every excellent achievement in our lives depends upon the quality and efficient operation of our Yi. Combining the awareness of the mind, eyes, heart and intention into one, causes the creation of a rich and rare alchemical mixture. The power of the Yi is the magical catalyst that assures energetic results.

The basic teachings are all in the Microcosmic Orbit, the Cosmic Inner Smile and the Six Healing Sounds. As long as you are alive and healthy, you are able to draw in energy from the universe and nature. The stronger the battery, the better it is able to store this energy. If you're weak and sick, you have no way to draw any energy from nature or the universe. When the cells stop drawing in energy, life ends. If you have perfected these practices, you have all the personal physical, mental and spiritual power that you need. Everything comes together. Your mind is aware of the original force. From the knowledge of Oneness, you are inwardly aware of your mind, senses and heart; you are outwardly aware of the universal energy. Now you are capable of receiving the abundantly available healing power of the universe; you can tune into everything, inwardly and outwardly in all directions. This is what we are here for: to heal and become whole again. We have higher goals to reach. The foundation for reaching our goals is a strong and healthy body.

Cosmic Healing Chi Kung is merely an extension of the universe within yourself. You draw the universe in through your palms, skin, heart and crown. With your Yi you breathe in the cosmos through these various open cavities of receptivity in your body, absorb its power, condense it, transform it and use it for the benefit of all.

Connecting the Bridge

Contracting the muscles of the perineum, sexual organs and anus activates our connection with the earth energy. By pulling up these areas and drawing in energy through the soles of the feet, we immediately become grounded and bring the earth (yin) energy into the Microcosmic Orbit. Pulling up should be done gently and directed by the Yi.

When we speak of "Connecting the Bridge," we are referring to the above exercise, except that we do not pull up on the perineum. Pull up gently and lightly on the sex organs and anus while leaving the perineum totally relaxed. After performing the first contractions, when you then relax the perineum, the earth energy will continue to flow into the body using the principles of a siphon. We thus connect the "Bridge" across the perineum from the sexual organs to the anus using the Yi to direct the energy. This actively combines the earth energy and your sexual energy and directs it into the Microcosmic Orbit. This energy is circulated during Cosmic Chi Kung. Holding a very gentle contraction will keep you grounded during the exercise. It can be done sitting or standing.

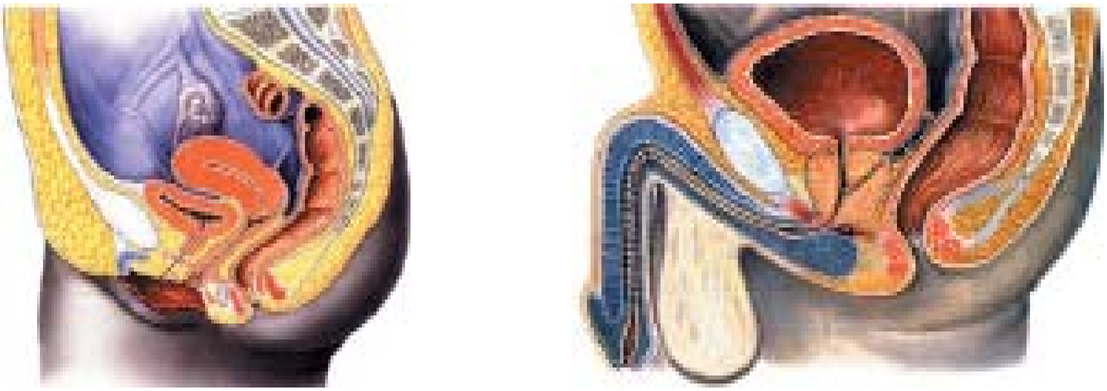

Women lightly contract the Vagina

Men lightly contract the Perineum and Anus

Fig. 8.4 *Connect the Bridge - combine Earth and Heavenly Energy.*

Cosmic Chi Kung Preparation

Cosmic Chi Kung can be practiced either in a sitting or a standing position.
Sitting, relax as much as possible and invoke the **Cosmic Inner Smile**.

1. Be aware of your **heart** and listen to your **kidneys** with your inner senses. Smile to your heart and kidneys and feel them pulsing; feel them communicating and interacting with each other. The **Kan** and **Li**; the water of the kidneys and the fire of the heart, balance and **begin to mix**.
2. Open the arm pits as if you were holding a ping pong ball there and slightly move the fingers to activate the Chi flow. Relax. **Feel** the sensations of the **Chi flow**: tingling, warmth, pulsing, electric and magnetic feelings.
3. Be aware of your palms. Draw in Chi by **activating the bone breathing** process. Become aware of the perineum; slightly pull up the perineum, sexual organs and anus and "**connect the bridge**." (Remember to relax the perineum).
4. Breathe into the palms of the hands and soles of the feet, further enhancing your connection to the earth force. Smile to the blue earth energy as it flows up your arms and legs and suffuses your body with pleasant Yin energy.
5. Breathe into the **mideyebrow** point and activate your connection with the **Cosmic Force**. Smile to the golden light of the cosmic energy as it swirls into your mideyebrow point and flows throughout your body, healing and balancing your energy.
6. Breathe into your **crown** point and activate your connection with the **Heavenly Force.** Smile to the red and violet light of the heavenly force as it flows into your brain and washes your body with subtle pure Yang energy.

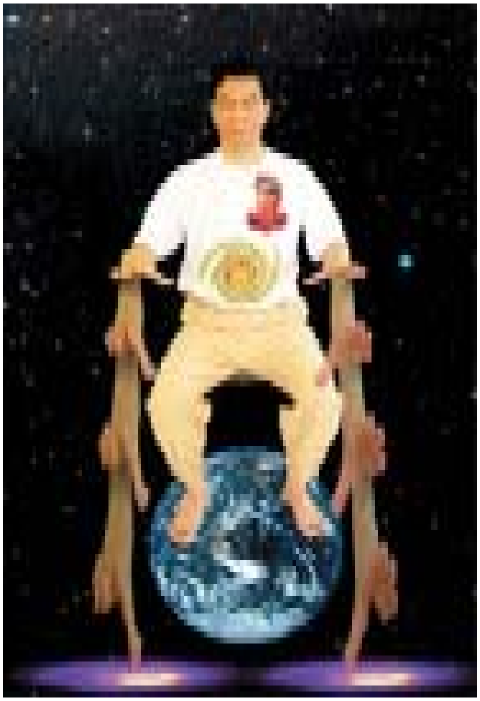

Fig. 8.6 *Invoke the Cosmic Inner Smile.*

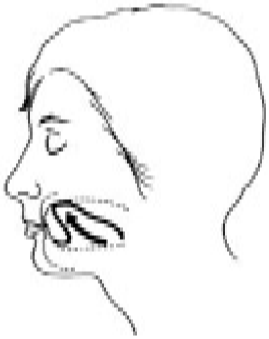

Fig. 8.5
Place the Tip of the Tongue on the Upper Palate.

7. Place the tip of your tongue on the upper palate behind the teeth in a comfortable position.

8. As you begin to move through the postures of Cosmic Chi Kung, always be aware of the Chi flow and its attending sensations.

Cosmic Chi Kung Preparation Standing

1. Stand with your **feet parallel** and shoulder width apart. Tilt your **sacrum** and **pelvis** slightly forward until you feel your feet press more firmly into the ground. As you tilt the pelvis, feel your lumbar vertebrae pressing outward; this is called "**Opening the Door of Life**."
2. Relax your **chest** and sink your **sternum**. Draw your chin in slightly, and hold your **head, neck** and **spine** erect as if your head and spine were suspended from above by a string.
3. Follow the same instructions for the preparations of Cosmic Chi Kung.

Opening to Heaven, Earth and Nature

The natural world is the source of our power. Cosmic Chi Kung teaches us about nature and its energies. In the Tao, nature and the universe are equivalent to God. You are a part of nature and can easily learn to open to the forces in the macrocosm and let them flow through you. This is very simple. Just tune into the frequencies around you. Your fingers are antennae, which transmit and receive.

1. To begin your practice of Cosmic Chi Kung, simply **stand** up and **feel the energy around you**. When you stand or sit, be aware of your soles (Bubbling Spring points) making a good connection with the earth. Your perineum (Hui Yin point) is relaxed and open. You are connecting to the earth force and can expand further down on to the other side of the universe. This is one type of energy.
2. When you draw your **chin in**, slightly **relax your chest** and tilt your head slightly forward; you begin to be aware of the heavenly force. This is another type of energy.
3. The qualities of the **Five Elements** reveal themselves as woody trees (growing force), fiery deserts (expanding force), earth mountains (stabilizing force), metal air (contracting force) and watery oceans and lakes (gathering force). We call the combination of these elements the '**Cosmic Force**'.

Thus, your mind and body learn to gather, absorb, direct and transform the heaven, earth and cosmic energy for your own use. This is the essence of Chi Kung. Today there is a lot of emphasis on learning movements and forms of Chi Kung. However, if the inner feeling is not there, the movements are of little value.

Heavenly force is Yang. Earth force is Yin. By learning to connect to the energies of heaven and earth, you have a powerful tool for restoring and maintaining your inner balance of Yin and Yang.

Outline of Practice

Before practicing each of the Cosmic Chi Kung sets in this chapter and in the following chapters begin with the heaven and earth spirals as outlined on page 157.

First **Create the Chi Field**.

Hold each of the Cosmic Chi Kung postures for at least 5 - 30 seconds, gradually lengthening the time to 60 - 120 seconds per posture.

Note: While holding the postures count the breath from 5 to 15 seconds, being aware of the **Tan Tien** and feeling the Chi in the Tan Tien which then moves up to the **crown** and expands up to the universe; fuse your Chi with the Universal Chi and let it multiply down to your palms. **Just concentrate on the 'Tan Tien and the Universe'**; do not concentrate on the palms at all.

Throughout parts I, II, III and IV 'Tan Tien and Universe' means to spiral into the Tan Tien and spiral out into the universe connecting 'Tan Tien and Universe'.

Part I
Opening to Heaven, Earth and Cosmic Chi

Opening Movements

Three Minds into One Mind

Empty the mind and heart down to the Tan Tien.
Hold the hands together at the heart center.
Smile and make the heart feel soft; feel love, joy and happiness.
Turn your consciousness inward and your awareness out.

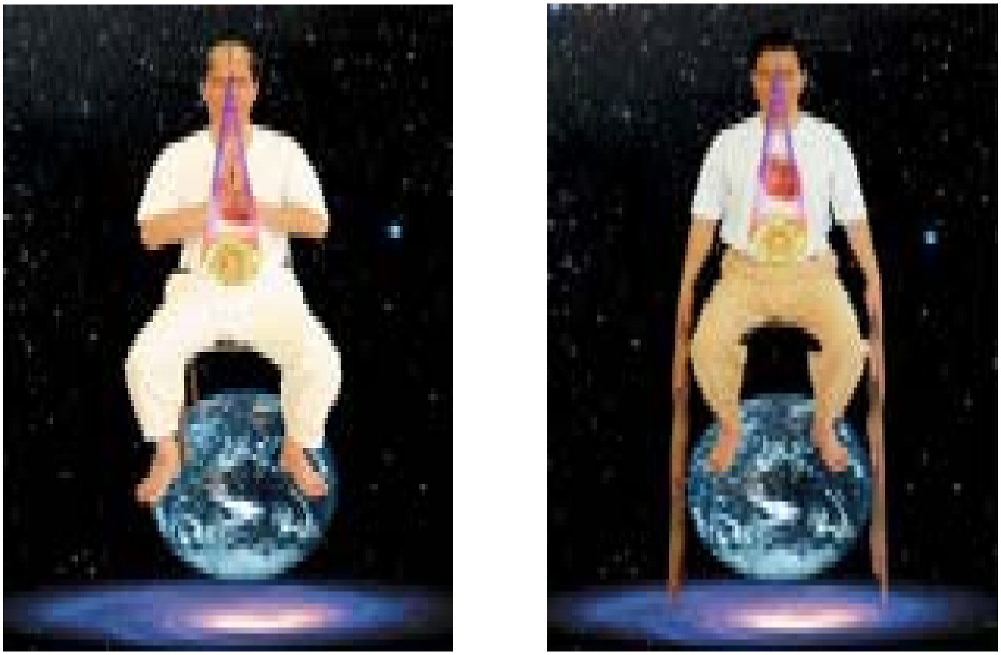

Fig. 8.7 *Empty your Mind and Heart down to your Lower Tan Tien.*

Feel your legs lengthen extending down to the center of the earth.
Feel your hands lengthen extending to the universe below.
Let the mind expand and be aware all the way down to the universe.
Feel the 'long' bones of the arms and legs and begin bone breathing.
Feel universal Chi filling, packing and compressing into the bones.
Draw this Chi down to the back of your crown and feel a heavy pressure press on the crown; feel slightly numb or a light electric downward flow, like oil dripping down. 'Focus at Tan Tien and Universe'. Count to five.

Channeling the Earth Force - Washing the Bone Marrow

Smile into the **perineum, palms** and **crown**. Then slowly raise the arms, palms facing each other, to chest height keeping the elbows relaxed and sunk. Rotate the arms slowly until the palms are facing down. Be aware of the mideyebrow and feel your breath and lightly contract your eyes and round muscles around the eyes. Connect the perineum bridge by slightly squeezing the sex organs and the anus; do this a few times.

Rest and **smile to soles of the feet**, feel the Chi bubbling and feel like you are standing on a hot spring starting to bubble up.

Be aware of your Tan Tien and expand your mind down past the earth to connect with the galaxy below you. Multiply and bring your mind back to the Tan Tien. Gradually feel the Chi being absorbed through the whole body and absorb it into the bones and body like a **rising steam or mist**. Feel the earth force move through the center of the bones and enter the bone marrow move up the calves and thighs (femur bones) through the hip bones, spine, **scapula**, **arms**, **neck** and **skull**. Finally **swirl the energy around your brain. 'Focus at Tan Tien and Universe'**.

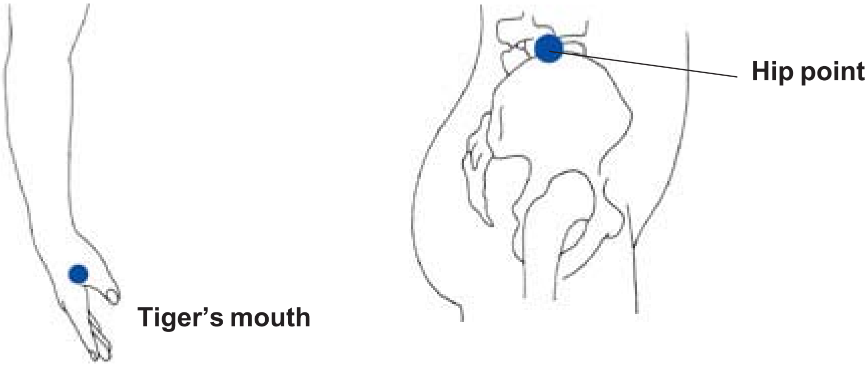

Fig 8.8 *Activate the Hip: Large Intestine 4, Pericardium 8 and Small Intestine 3.*

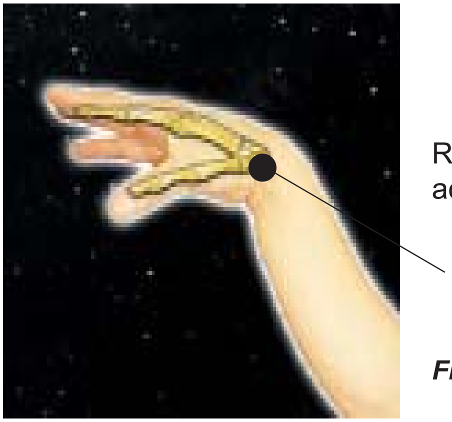

Raise your index finger slightly upward to activate the Hegu Point (Large Intestine).

Fig. 8.9 *Open the 'Tiger's Mouth'.*

Then **stretch your thumbs out to the sides and down**, so that they end up pointing downwards toward the earth. This activates Large Intestine 4 (Hegu). Large Intestine 4 is found in the webbing between the thumb and index finger. It is called the "eye of the hand"; this area is also known as the "Tiger's Mouth."

Slowly draw the elbows back and lower the hands until your palms are facing down beside the "eye of the hip" (the iliac crest) with the "eye of the hand" (Large Intestine 4) aligned with the "eye of the hip. "Be aware of the Tan Tien (Yi, the abdominal brain) and expand your mind out into space, the Cosmos and the universe. Remember: The fingers and the energy point act just like a laser to help guide the universal force flow; as soon as you focus on the fingers or the palms you are starting to use your own energy. Feel your Tan Tien and the crown full with Chi. Just use your second mind intention, lightly aware of the area between the hips and LI-4 points. Focus at Tan Tien and the universe. This will activate the lungs and large intestines. **Hold for a 30 second count** and gradually you will feel the ascending colon and the sigmoid colon have been activated; you may feel some movement around this area.

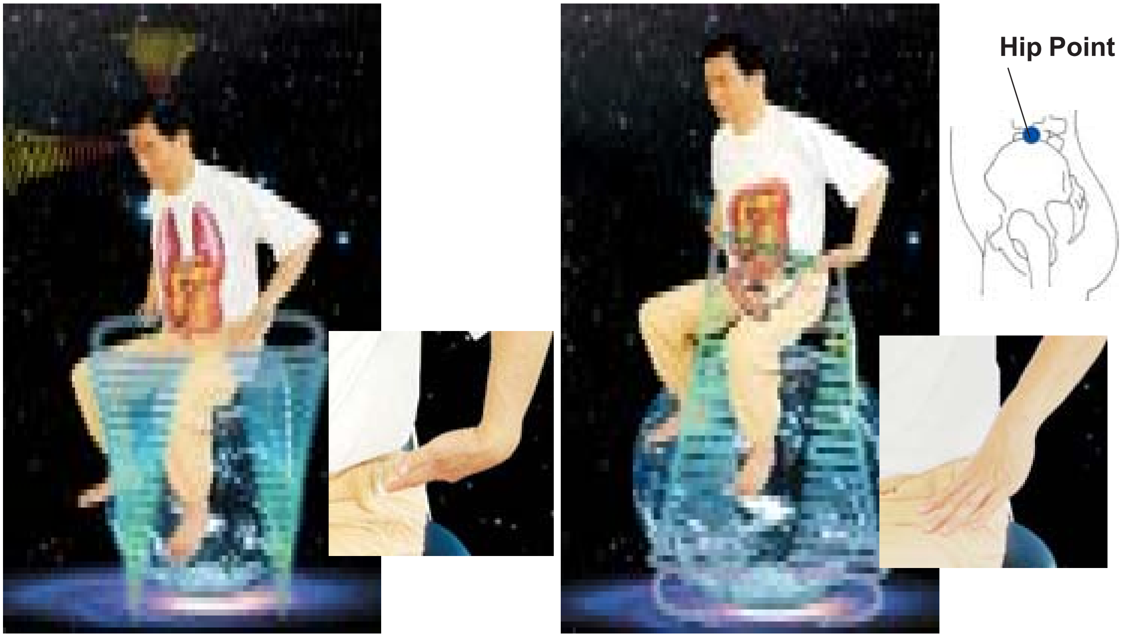

Fig. 8.10 *Activate the Organ Energy.*

Use your Yi to rotate your hands so that the fingers first point out, then back, turn the palms up and finally point the middle finger in toward the eye of the hip. Tan Tien and the universe - **95 % aware of the 'Tan Tien and the Universe' and 5 % aware of energy flowing back and forth between the tips of the middle fingers** (Pericardium 9); feel the Chi passing through the hips. This will activate the pericardium.

Rotate the hands again until the fingers are pointing forward with the **palms facing up**. Align the knife edges (Small Intestine 3) with the eyes of the hip. 'Focus at Tan Tien and the Universe'. Be lightly aware of the energy passing back and forth between the two hands. This will activate the heart and small intestine.

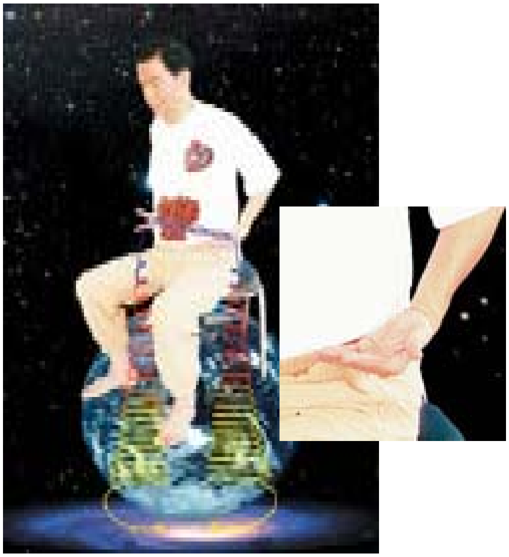

Fig. 8.11 *Activate the Heart and Small Intestine.*

Absorb the Heavenly Force
Washing the Bone Marrow - Crown to Soles

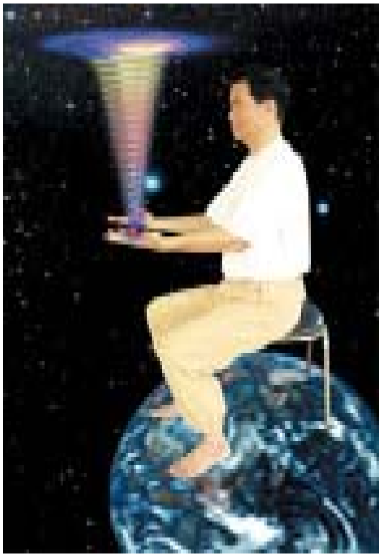

Extend the arms forward to chest height with the palms facing upward.

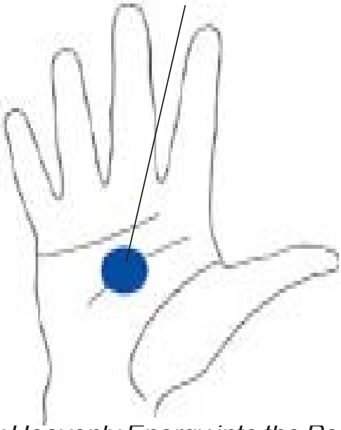

Fig. 8.12 *Draw Heavenly Energy into the Palms.*

Become aware of the **Tan Tien, the Crown and the Star above you**. Expand your mind (YI) out to the universe and connect to the galaxy. Be aware of the palms and the crown which are connected to the galaxy. Activate the Laogong point (Pericardium 8).

Feel the heavenly energy come directly to your palms, and be aware of the violet light of the North Star and the red light of the Big Dipper both shining radiantly.

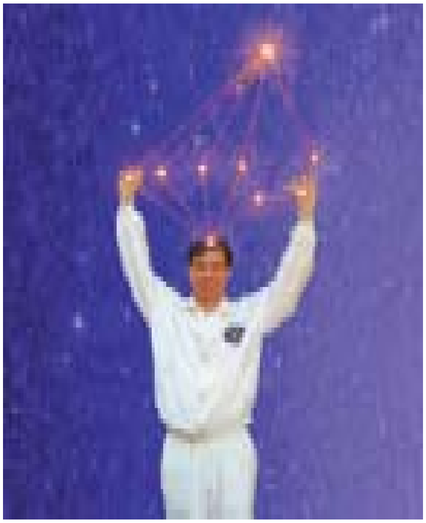

Fig. 8.13 *Hold the Big Dipper.*

Picture yourself holding the handle of the Big Dipper and turn the 'cup' pouring the light over your head. Draw this light in through the palms and the crown point and feel it washing your bone marrow. Always remember your Tan Tien. **Direct the sensations down through the skull, cervical vertebrae, clavicle, scapulas and sternum. Activate your thymus gland and continue drawing the healing heavenly light down through the rib cage, arms, spine, hips, femurs and calf bones, feet and toe bones. Be aware of the soles bubbling up to the perineum and up to the crown**.

Lightly rock yourself like riding a horse; this will help keep your spine open and Chi can flow easily throughout the practice.

Absorb the Earth Force and the Other Side of the Galaxy

Rotate the arms until the **hands are palms down**.

Raise the index fingers slightly and extend the thumbs first towards each other and then toward the ground. Be 95% aware of the Tan Tien and the Universe, and 5 % aware of the index fingers and the big toes. Wait until you feel the Chi enter. Fill the joints of the fingers and up to the wrists; feel 'tense but not tense' and let the **Chi continue to fill the joints** up to the elbows, up to the shoulders and gradually up to the neck and the head. Go from the toes up the legs to the hips and on into the spine and the rib cage **completing all the joints of the body**.

Be aware of the Tan Tien and the universe and only slightly aware of the palms (Laogong), the soles of the feet (Kidney 1-Bubbling Spring), and the perineum (Conception Vessel 1-Hui Yin) bubbling. Activate these points, feel the earth energy and continue smiling through the earth to connect with the galaxy, Milky Way, and the universe. Draw this light in through the palms and soles and let it wash up through the center of your bones from the feet up. Let it **steam and cleanse your marrow**. Picture any impurities or illness dripping out of your bones and draining down into the earth, where they will be recycled and purified.

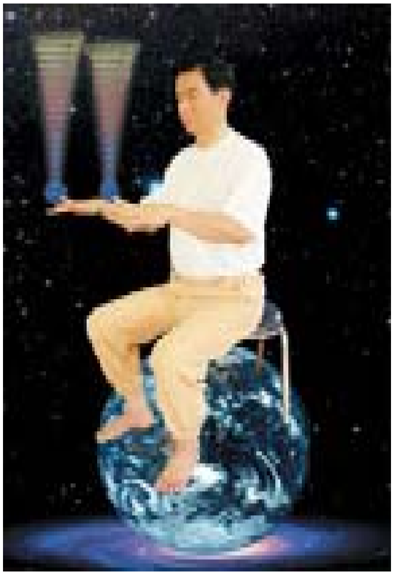

Fig. 8.14 *Open the index fingers.*

Core Movements of Part I

Grasping the Moon

Connecting the Heaven and Earth Forces - Right Side over Left

Open Position: Heaven Force - Slightly raise the index fingers to open the palms. Bring the arms up , bending the elbows so that the left hand is under the right elbow with the left index finger pointing up to Heart 3.

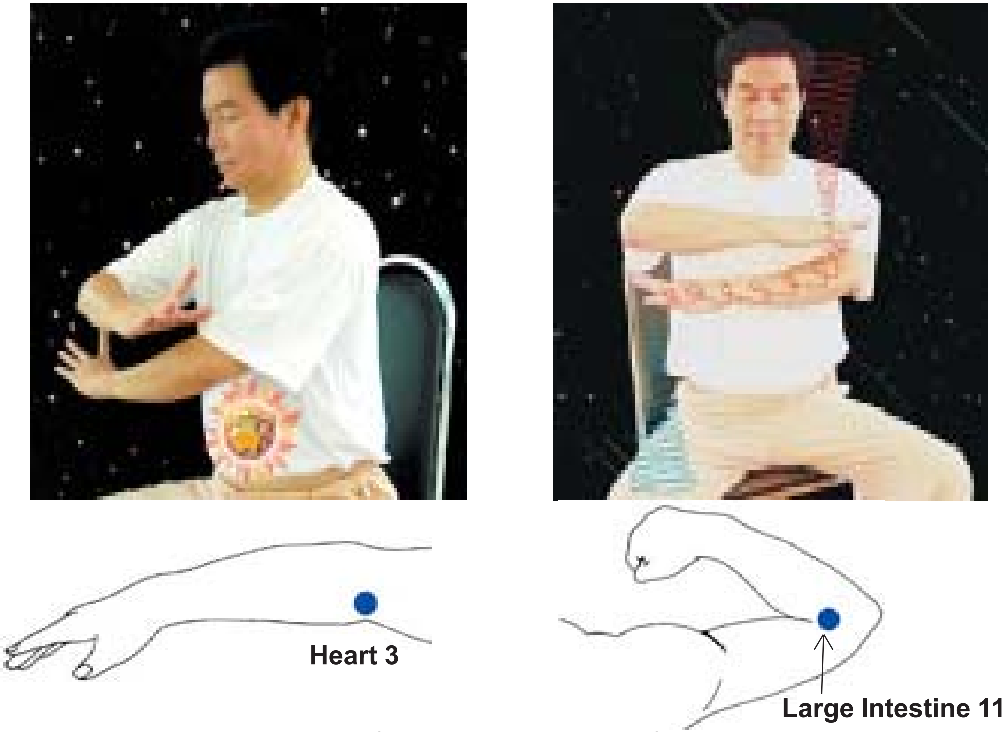

Fig. 8.15 *Complete the Heavenly Circuit.*

The right index finger is pointing up toward heaven and connecting to the galaxy. The right forearm is over the left forearm, with the tip of the right index finger held above the Large Intestine 11 point of the left arm.

The **right index finger draws in the heavenly force** like antennae. The force flows up the bones of the right arm, across the shoulders to the left arm and through the bones of the left arm to the left index fingertip. It **then flows through the index finger of the left hand** into the Heart 3 point of the right arm, **completing the heavenly circuit**. Continue cycling in this way. Hold the position and count to five.

The practice uses the 'Empty Force'; this means that there is no contact. The fingers must be 'tense without tensing'. When you tense and raise the index finger, you will feel the Laogong point opening.

Close Position: Tan Tien, earth and the universal force - keeping the arms in the same position, move both index fingers to point down. The **right index finger will now point toward the Large Intestine 11 point of the left arm and the left index finger will point downward toward the earth** and continue down to connect to the universe. Draw in the earth force through the left index finger, bringing it through the bones of the left arm, across the shoulders through the right arm to the right index finger and out into the Large Intestine 11 point of the left arm. Continue cycling in this way, **completing the earth circuit**.

Open and close three times altogether ending in the open position. Hold each count for five seconds.

Holding and Activating the Chi Ball
Tan Tien and the Original Force - Right Side over Left

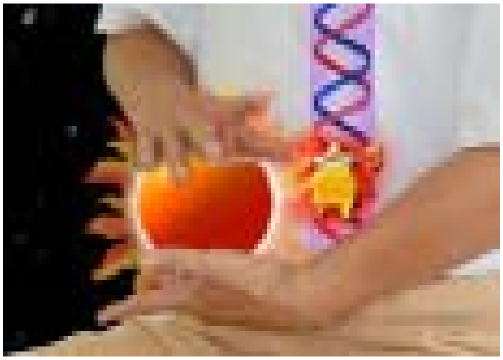

Fig. 8.16 Right Side over Left - 'Holding the Ball'.

Turn the **left palm upward** and lower it to the Tan Tien with the pinky pointing inwards.

Lower the right palm to the navel, palm down, as though you were holding a small Chi ball. Lower the right thumb to point toward the left Laogong point. This opens the eye of the right hand (LI-4), then point it toward the navel. Be aware of Tan Tien and universe.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at navel level. The **left palm** is still pointing up, and is **Yang** drawing in the heavenly force through Laogong. The **right palm** is still facing down, and is **Yin** drawing in the earth and the galactic force from below.

Conclude by turning the left palm over to face downward. **Draw in the earth force through both palms.**

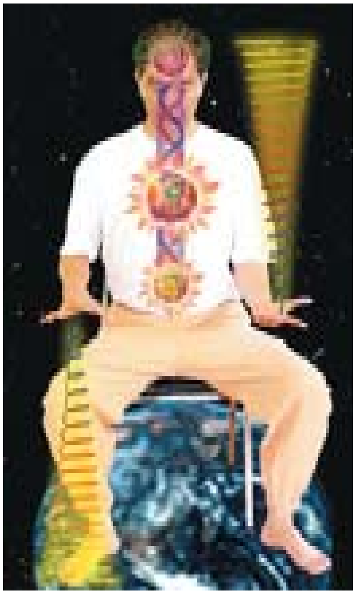

***Fig. 8.17** 'Yin and Yang Palms' - Right Hand connecting to Earth and Left to Heaven.*

Left Side: **Repeat on the left hand side with the left arm over the right**. Holding the Chi ball the left hand is on top. Hold for a count of five seconds.

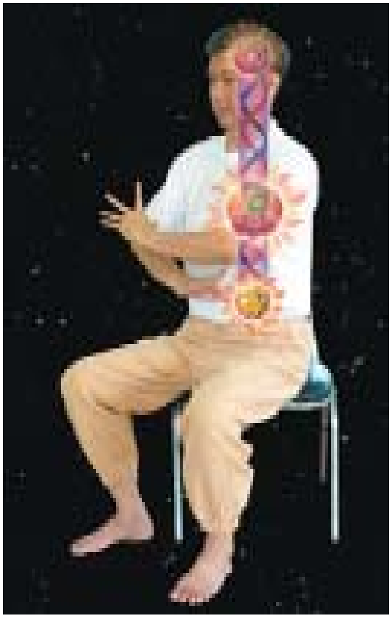

Fig. 8.18 *Open the Heavenly Circuit. Then the bottom hand slides and turns 'Palm-Up' and the Top Arm slides to 'hold the Chi Ball'*

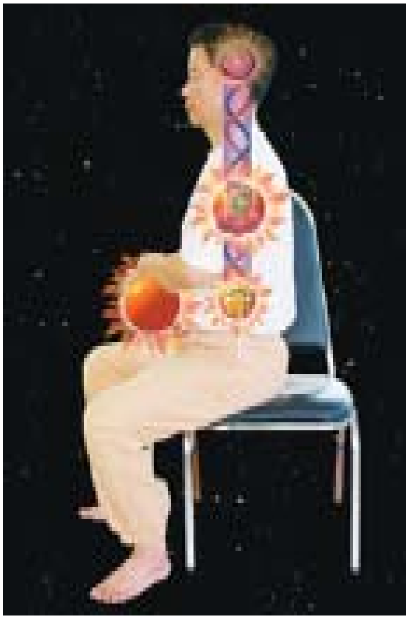

Fig. 8.19 *Holding the Chi Ball - Left Side over Right.*

Closing Movements

Activate the Chi Channels - Balance the Orgаns

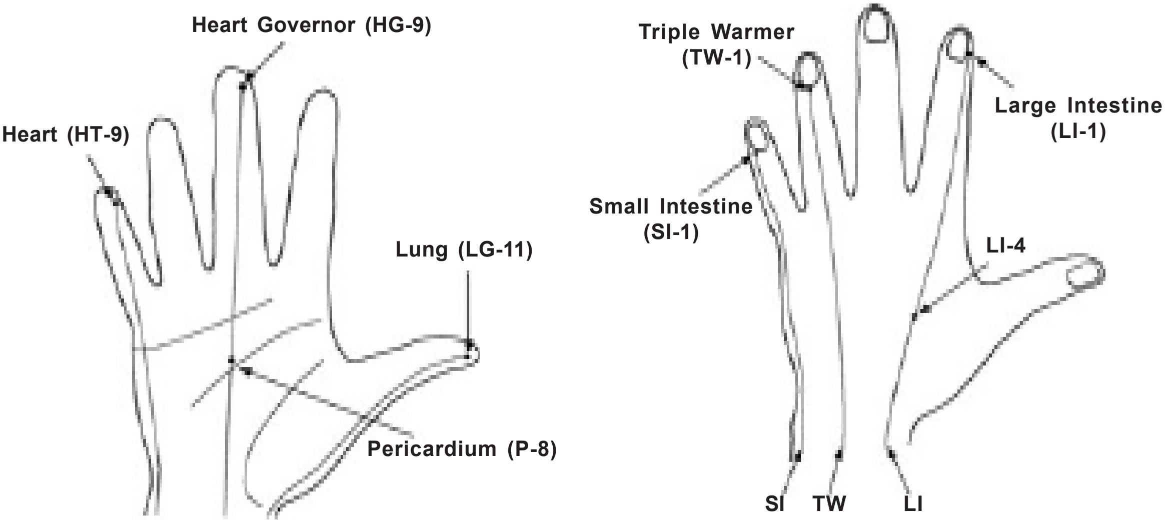
Heart Governor (HG-9)
Heart (HT-9)
Lung (LG-11)
Pericardium (P-8)
Triple Warmer (TW-1)
Large Intestine (LI-1)
Small Intestine (SI-1)
LI-4
SI
TW
LI

Fig. 8.20 *Finger Points*

The closing movements balance organ energy while activating the Chi Channels.

Hold the arms extended forward with the palms down at solar plexus level. Hold each stage for a count of five seconds.

Index Fingers:

Tense all the fingers and feel the Chi filling all the joints; open the index fingers, raising them up, while keeping the rest of the fingers level.

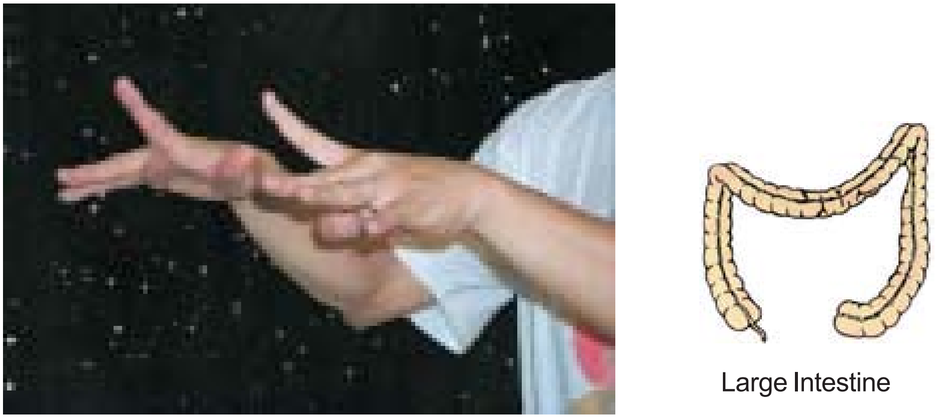

Fig. 8.21 *Activate the Index Fingers - Up.*

The tension should be like the string of a tuned musical instrument. When lightly tense the Heavenly Chi will be drawn in through the index fingers.

Relax the hands and bring the index fingers back to the neutral position, level with the other fingers.

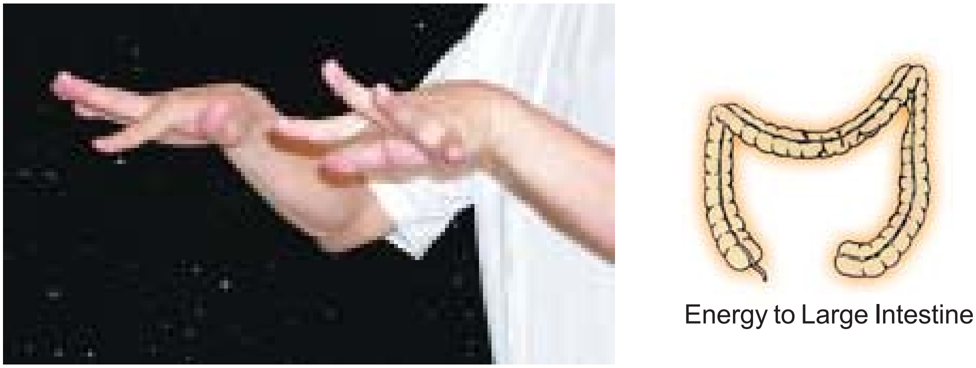

Fig. 8.22 *Draw Earth Energy through the Index Fingers - Down.*

Stretch out and tense all the fingers and **press the index fingers to point down** toward the earth and continue down to the universe. Keep the other fingers level. **Draw in earth energy through the index fingers and circulate it through the body.**

Bring the index fingers back to neutral and relax your hands.
Once again **open the index fingers by raising the index fingers up**, while keeping the rest of the fingers level. **Draw in the Heavenly Chi through the index fingers**.
Bring the index fingers back to neutral again and relax your hands.

Ring Fingers:

Stretch the ring fingers down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the ring fingers and circulate it through the body.

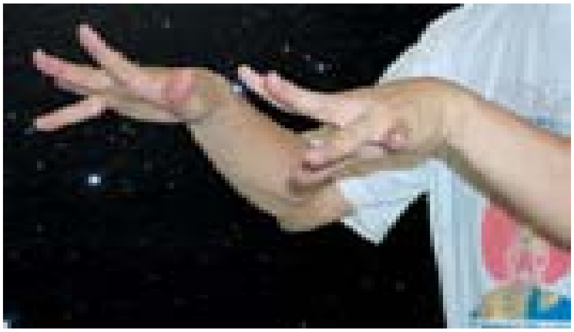

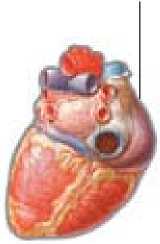
Pericardium sack around the Heart

Fig. 8.23 *Draw Earth Energy through the Ring Fingers.*

Bring the ring fingers back to neutral and relax your hands.
Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level, and draw in the Heavenly Chi through the index fingers.
Bring the index fingers back to neutral again and relax your hands.

Thumbs:

Stretch the thumbs out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the thumbs and circulate it through the body.
Bring the thumbs back to neutral and relax your hands.

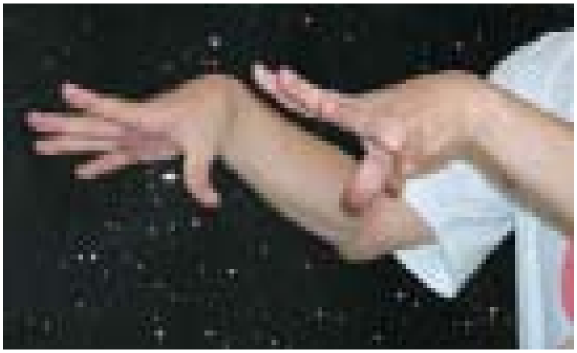

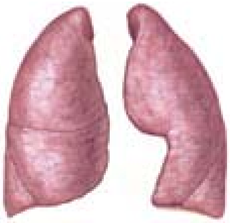

Fig. 8.24 *Draw Earth Energy through the Thumbs.*

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level, and draw in the Heavenly Chi through the index fingers.

Bring the index fingers back to neutral again and relax your hands.

Pinky Fingers:

Stretch the pinky fingers out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the pinky fingers and circulate it through the body.

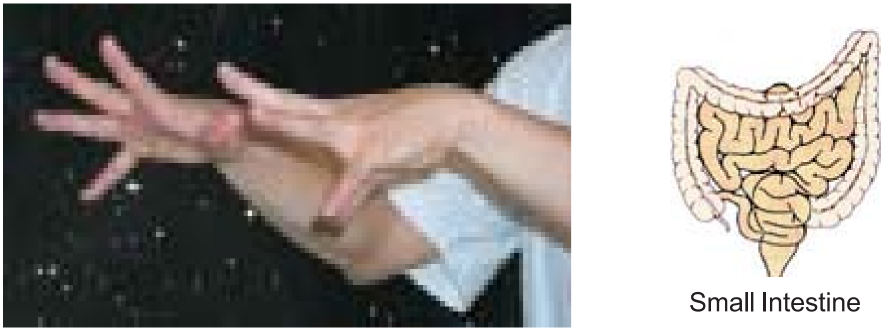

Small Intestine

Fig. 8.25 *Draw Earth Energy through the Pinky Fingers.*

Bring the pinky fingers back to neutral and relax your hands.

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level. Draw in the Heavenly Chi through the index fingers. Bring the index fingers back to neutral again and relax your hands.

Middle Fingers:

Stretch the middle fingers out and down toward the earth, keeping the rest of the fingers level. Draw in earth energy through the middle fingers and circulate it through the body.

Bring the middle fingers back to neutral and relax your hands.

Once again open the index fingers by raising the index fingers up, while keeping the rest of the fingers level. Draw in the Heavenly Chi through the index fingers.

Bring the index fingers back to neutral again and relax your hands.

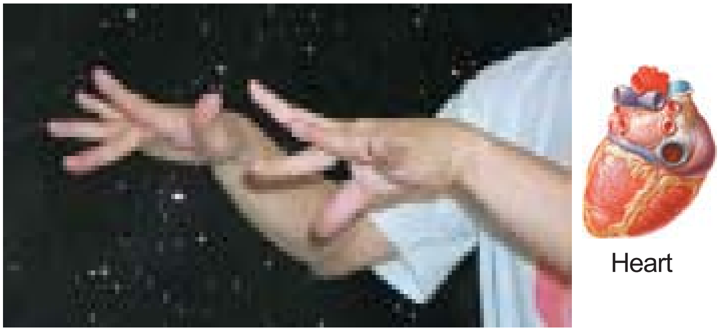

Heart

Fig. 8.26 *Draw Earth Energy through the Middle Fingers.*

Crane's Beak and Swallow the Saliva:

Form the "Crane's Beak" with both hands by bringing all the fingertips together with the thumbs inside. **Inhale** and **contract the sexual organs**.

Move your tongue and suck the mouth to **activate the saliva**. Divide it into three parts, tighten the neck and gulp down the **first part** to the center of the **navel**; force the **second** mouthful of nectar down to the **left side of the navel** and the **third** to the **right side of the navel**. Raise the forearms up to shoulder height with the fingers pointing down. Slowly open the palms and begin to lower the arms to the sides, until the palms are facing each other before the starting position.

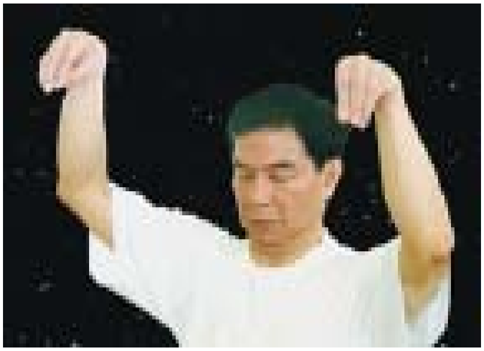

Fig. 8.27 *Form the 'Crane's Beak'.*

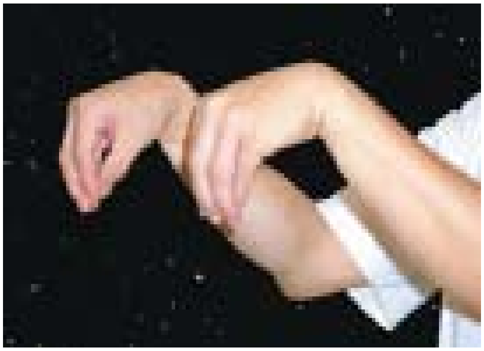

Fig. 8.28 *Raise Forearms to Shoulders.*

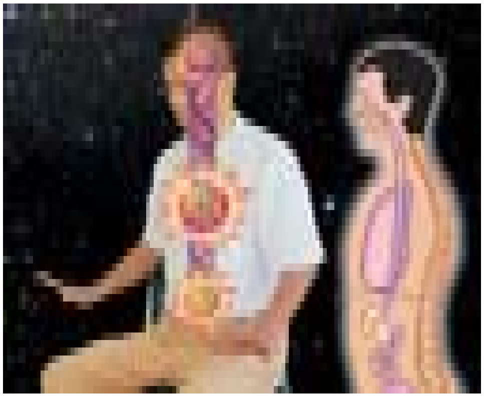

Fig. 8.29 *Lower Arms to the Sides.*

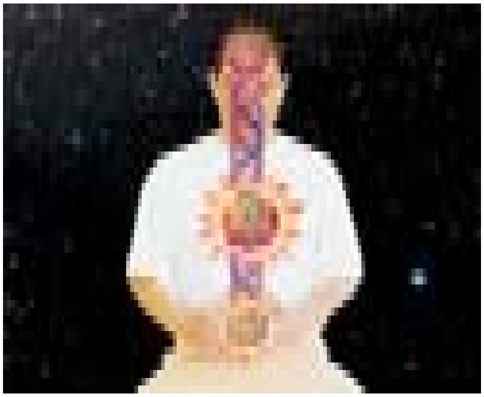

Fig. 8.30
Palms facing each other to cover the Navel.

Conclusion:

Bring awareness to the navel. Notice the quality and intensity of the energy generated and **collect your energy at the navel**. Men place hands over the navel with the right hand on top; women vice versa. Rest.

Summary:

Practice Part I daily for one or two weeks until you can do it well. Then you can proceed to learning Part II.

Part II
Opening the Bridge and Regulator Channels

Theory

Chi flows through your body along numerous channels or meridians. The channels are divided into two major groups: the **extraordinary** or special channels and the **ordinary** channels. After your mother's egg and your father's sperm joined together to form a single cell, that cell began to divide. The extraordinary channels were the first energy channels formed as a result of that early cell division.

There are **Eight Extraordinary Channels:**

Governor Channel
Functional or Conception Channel
Thrusting Channel
Belt Channel
Yang Regulator Channel
Yin Regulator Channel
Yang Bridge Channel
Yin Bridge Channel

Later, as your fetus developed, your Original Chi flowed through the eight extraordinary channels to help create your internal organs and their energy channels. These twelve "ordinary channels" are divided into yin and yang. The yin Channels are connected to the solid organs, and the yang channels are connected to the hollow organs.

The **Twelve Ordinary Channels:**

Lung (Yin) Channel
Large Intestine (Yang) Channel
Stomach (Yang) Channel
Spleen (Yin) Channel
Heart (Yin) Channel
Small Intestine (Yang) Channel
Bladder (Yang) Channel
Kidneys (Yin) Channel
Pericardium (Yin) Channel
Triple Warmer (Yang) Channel
Gall Bladder (Yang) Channel
Liver (Yin) Channel

These twelve organs serve to extract energy from the food and water we ingest and the air that we breathe to create energy through the metabolic processes of respiration, circulation, digestion, elimination and reproduction. This energy is called PostNatal Chi, as it comes in after we are born. Thus the extraordinary channels serve as the link between our Original or Prenatal Chi which came from our mother and father, and our Postnatal Chi which comes from our food and air as a result of metabolism.

General Functions of the Eight Extraordinary Channels

They serve as reservoirs of Chi.

About two thousand years ago, one of the great texts of Chinese medicine, the Nan Ching was written. This classical text describes the twelve ordinary channels as rivers and the eight extraordinary channels as reservoirs of Chi. When the ordinary channels become low in energy, they can draw from the reservoirs of energy in the extraordinary channels. On the other hand, if the ordinary channels become too full, the excess can be taken up by the extraordinary channels. In this way the extraordinary channels help us to maintain a balance in our energy body.

They store and circulate Ching Chi.

The extraordinary channels all draw their energy from the kidneys, which are the storehouse of Ching Chi (essence or sexual energy) in the body. Thus the extraordinary channels circulate the Ching Chi around the body, particularly to the skin and hair, and to the five ancestral organs: the brain and spinal cord, the bone marrow, the blood, the uterus and the liver and gall bladder.

They circulate Defensive Chi to protect the body.

The Chi that protects the body against invasion by external pathogens is called Defensive Chi or Wei Chi. The extraordinary channels circulate Defensive Chi over the back, abdomen and thorax. These functions are performed by the Governor Channel, the Functional Channel and the Thrusting Channels, respectively.

They regulate our life cycles.

In the first chapter of the Yellow Emperor's Classic of Internal Medicine (Huang Di Nei Ching Su Wen), another of the classical texts of Chinese medicine, the life changes of women and men are described in 7 and 8 year cycles respectively. The Functional Channel and the Thrusting Channel govern these cycles.

Eight Extraordinary Channels and Chi Kung

There are few texts available today that describe the purpose and function of the eight extraordinary channels and their place in Chi Kung practice. To make matters more confusing, the extraordinary channels and points used in Chi Kung are often quite different from those used in acupuncture, even though they may bear the same names.

Most acupuncturists pay little attention to the eight extraordinary channels in diagnosis and treatment with the exception of some modern Japanese researchers. By contrast, the extraordinary channels have been of special importance to Taoists and Chi Kung practitioners for thousands of years. Taoists see the extraordinary channels as the foundation of our bodily energy, as the bridge between our Original or Prenatal Chi and our Postnatal Chi; these channels affect us on the deepest level of our basic constitutional energy.

Therefore, Taoists focus on opening up the flow of energy through the eight extraordinary channels as a prerequisite for opening the energy flow in the twelve ordinary channels. In the Universal Tao system, you first open up the Governor Channel and the Functional Channel in the Microcosmic Orbit Meditation. The second pair: the Thrusting Channels and the Belt Channels are opened in the Fusion of the Five Elements II Meditation. The last four channels, the Yin and Yang Bridge Channels and the Yin and Yang Regulator Channels are opened in the third level of Fusion of the Five Elements Meditation. After opening all eight channels, in the next level of Taoist inner alchemical meditation called the Lesser Enlightenment of Kan and Li, the twelve ordinary channels are then opened.

Special Functions of the Bridge and Regulator Channels

The Bridge and Regulator Channels have no points of their own. They borrow points from the other 'ordinary' channels linking together and controlling the flow through all the separate channels of the body. The Bridge and Regulator Channels travel along the same paths. There is no 'real' difference between them. Both of them connect or 'bridge' and 'regulate' the flow of Chi in the body's meridians.

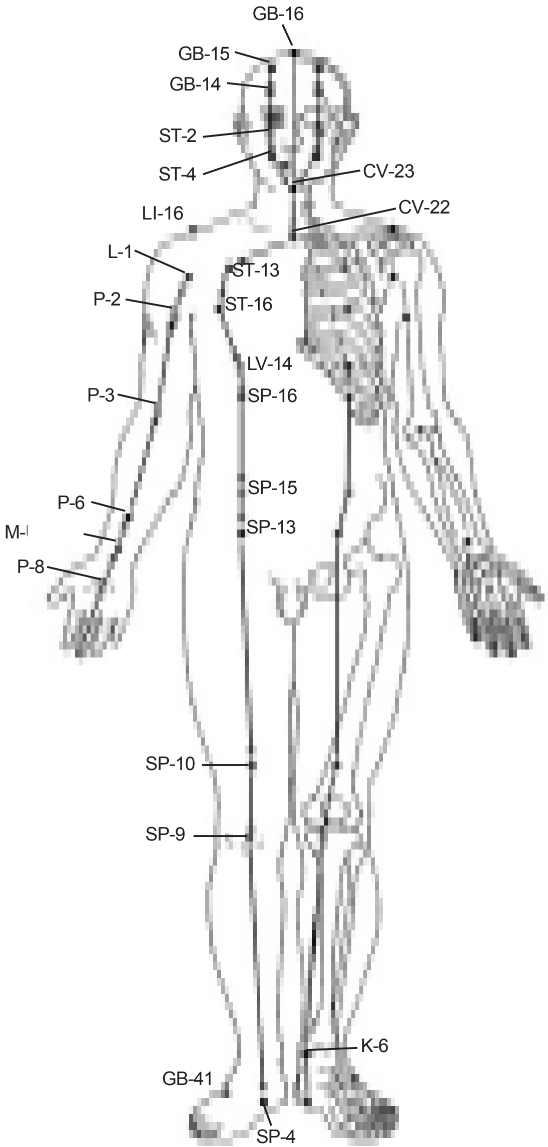
GB-16
GB-15
GB-14
ST-2
ST-4
CV-23
CV-22
LI-16
L-1
ST-13
P-2
ST-16
LV-14
P-3
SP-16
SP-15
P-6
M-
SP-13
P-8
SP-10
SP-9
K-6
GB-41
SP-4

Fig. 8.31 *Yin Bridge and Regulator Channels*

Bridge Channels Yin and Yang Chiao Mo Function

The Bridge Channels, also called the Heel Vessels as they originate at the heels, regulate the amount of energy being used by all the other meridians in the body. They act like a bridge linking together the stored Chi in the body and the areas in need of Chi. Usually if any meridian uses more energy than it needs to flow properly, then other meridians become deficient as a result. Thus the Bridge Channels seek to assure that your energy is distributed in a balanced way. The Yin Bridge Channel runs along the front or yin side of the body, while the Yang Bridge Channel runs along the back of the body.

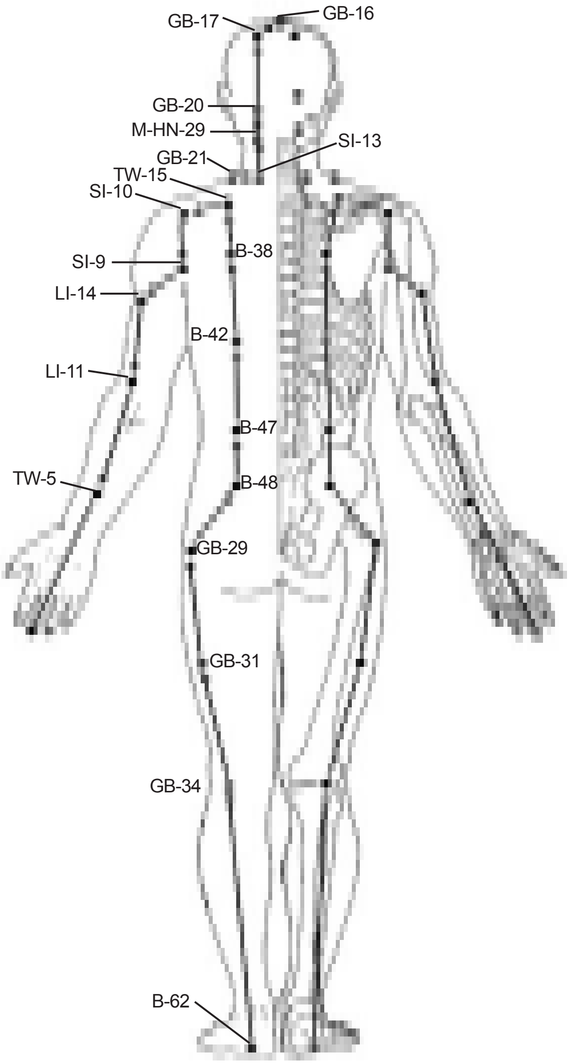
GB-17
GB-16
GB-20
M-HN-29
SI-13
GB-21
TW-15
SI-10
B-38
SI-9
LI-14
B-42
LI-11
B-47
TW-5
B-48
GB-29
GB-31
GB-34
B-62

Fig. 8.32 *Yang Bridge and Regulator Channels*

Disorders of the Bridge Channels

Excess Yang Energy

When yin energy is slowed down in the Bridge Channels, yang energy flows more rapidly. The excess yang can cause the following problems: insomnia, difficulty in closing the eyes, hypertension, stiff back and waist, inability to bend down, thigh tumors, bad colds, spontaneous sweating, headaches, painful eyes, paralysis of the arms and legs, vomiting of milk in infants, deafness, epilepsy, nose bleeding, swelling of the body, pain in the joints and head sweating.

Excess Yin Energy

When yang energy is slowed down in the Bridge Channels, yin energy moves more rapidly. The excess yin causes the following problems: sleepiness, difficulty in keeping the eyes open, hypotension, choking, painful urination, stomach rumbling, vomiting, diarrhea, difficult bowel movements, difficult labor and unconsciousness.

Regulator Channels

Yin and Yang Wei Mo Function

The Regulator Channels, also called the Linking Vessels, bind together all the meridians in the body. The Regulator Channels are also divided into yin and yang. The yin aspect, which runs along the front of the body, moves the yin energy and regulates the blood and inner parts of the body. It connects with all the yin channels: the Liver, Spleen, Kidney, Heart, Pericardium and Lung channels. If the yin aspect becomes imbalanced, the person may suffer from heart pains.

The yang aspect, which runs along the back side of the body, moves the yang energy, controls Defensive Chi, regulates resistance to external infections, and regulates the external parts of the body. It connects with all the yang channels: the Stomach, Bladder, Gall Bladder, Large Intestine, Triple Warmer & Small Intestine channels. If the yang aspect becomes out of balance, the person may catch colds and fevers easily.

By joining together the various channels, the Regulator Channels help to maintain a harmonious and cooperative interaction between the different channels.

The Regulator Channels in Taoist Yoga and Chi Kung are slightly different from those presented in acupuncture texts. The Chi Kung Regulator Channels include the yin and yang arm routes as well; some Taoist Yoga texts also refer to the arm routes as the Yin Yu and the Yang Yu. Acupuncture texts, by contrast, include only the leg, trunk and head routes. Many recent Chi Kung texts, unaware of these differences, depict illustrations from acupuncture texts alone, further adding to the confusion.

Disorders of the Regulator Channels

Excess Yang Energy

When yin energy is slowed down in the Regulator Channels, yang energy moves more rapidly. The excess yang can cause the following problems: swelling and pain in the joints, cold knees, paralysis of the arms and legs, painful back and sides, aching muscles, pain in the head, neck and edge of the eyebrows, fever, rashes, night sweating, tetanus, painful red eyes, colds and superficial fevers.

Excess Yin Energy

Similarly, when yang energy is slowed down in the Regulator Channels, yin energy moves more rapidly. The excess yin can cause the following problems: heart pain, diarrhea with stomach rumbling, difficulty swallowing, pain on both sides of the chest, diseases associated with cold and convulsions.

Summary

The Bridge and Regulator Channels link all of the yin and yang channels in the body together and regulate the flow of energy in these channels to maintain a state of energetic balance in the body. The channels used for meditation are somewhat different from the ones seen in Chinese medical text and acupuncture charts. This is because their purposes are different.

Chinese medicine aims to restore sick people to health. The points being treated must be superficial so that they can be activated by acupuncture needles.

Chi Kung and Taoist meditation aim to maintain health and to take one beyond mere physical health to spiritual immortality. The channels and points can be deep within the body, since the energy is guided by the mind or by postures and movements rather than by needles.

The movements used in Cosmic Chi Kung Part II to activate the Bridge and Regulator Channels are relatively simple. To completely master the Bridge and Regulator Channels, you must learn and practice the third level of Fusion of the Five Elements Meditation. Once you have done so, this part of Cosmic Chi Kung practice will be particularly powerful and balancing for you.

Practice

Opening Movements

The Opening Movements for this second part of Cosmic Chi Kung are the same as for the first part. By now you should be familiar with these movements, so they will be described briefly. For a more detailed description, refer to Part I, Opening Movements.

Hands down at sides, palms facing each other. Commence Bone breathing. Be aware of the Tan Tien and absorb Cosmic Chi. Be aware of the Tan Tien and channeling the earth and galactic forces.

Washing the Bone Marrow

Raise arms to chest height - palms facing each other. Rotate arms until palms face down. Open the 'Tiger's Mouth' - Index finger up, thumb down.

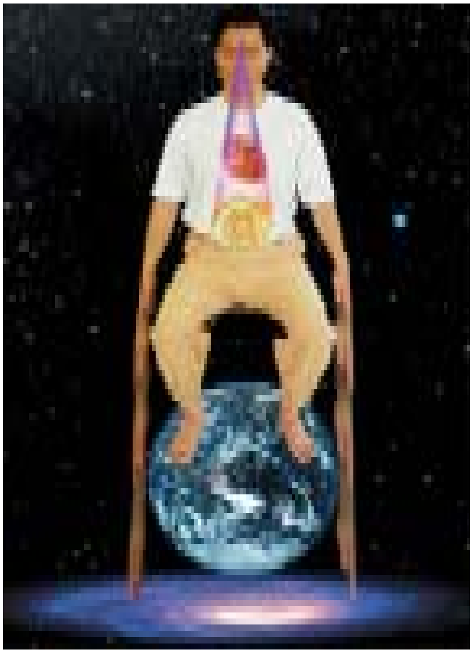

Fig. 8.33 *Hands lower to the Sides. Lower the Mind to the Tan Tien.*

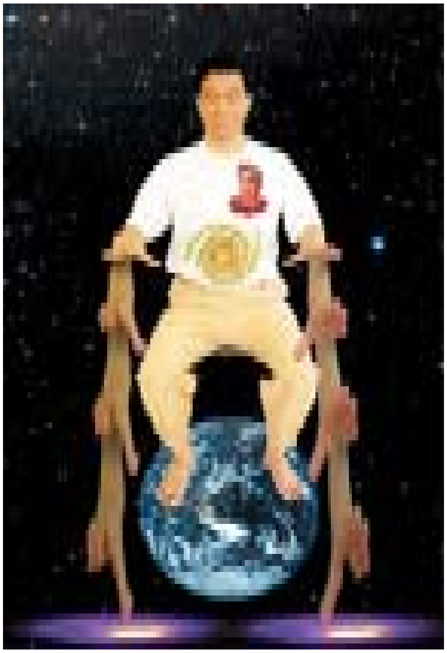

Fig. 8.34 *Open the 'Tiger's Mouth'.*

Activate and Beam to the Eye of the Hip

A. Large Intestine 4 - Eye of the Hand

B. Pericardium 9 - Tip of the Middle Finger

C. Small Intestine 3 - Knife Edge of the Hand

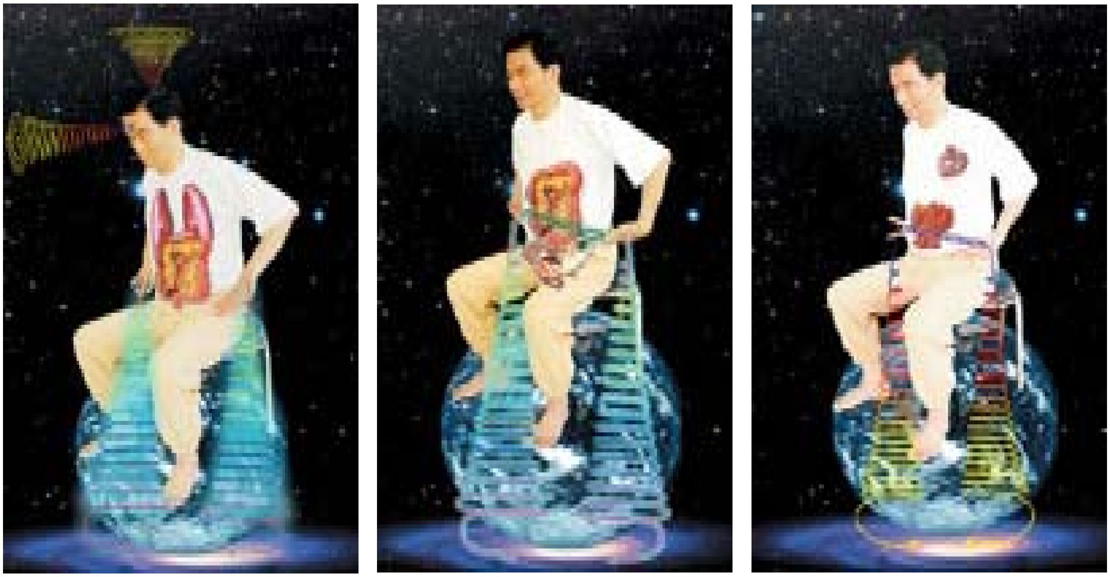

Fig. 8.35 *Activate the Large Intestine, Pericardium and Small Intestine.*

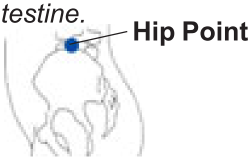

Be aware of the Tan Tien and absorb the Heavenly Force Washing the Bone Marrow Crown to Soles.

Extend the arms forward - palms upward. Be aware of the Tan Tien and absorb the heavenly force. Rotate the arms until the palms face down. Raise the index fingers slightly. Absorb the Earth Chi.

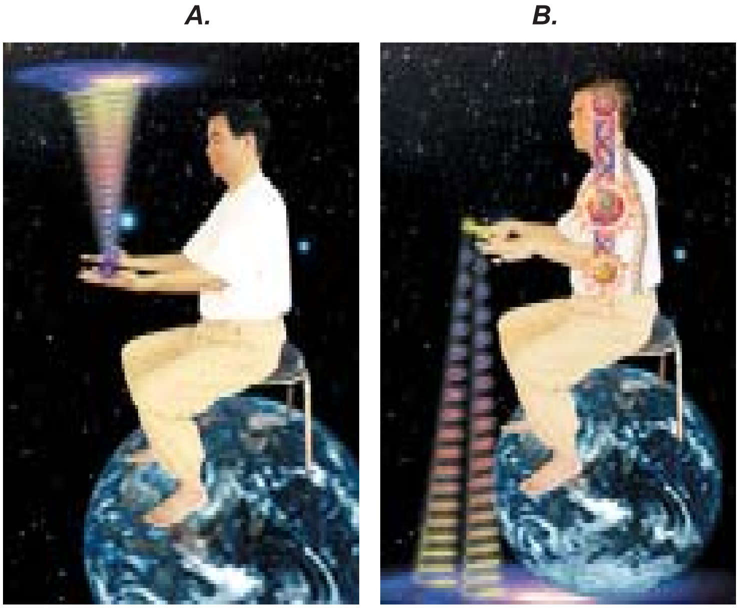

Fig. 8.36

A. *Be aware of the Tan Tien and absorb the Heavenly Force.*

B. *Index Fingers up - be aware of the Tan Tien and absorb the Earth and Universal Force.*

Core Movements of Part II
Opening the Bridge and Regulator Channels

Activate the Throat Center - Right Side

Keeping the elbows sunk, raise the left palm first so that it is about 1 inch in front of the throat center (Conception Vessel 22), palm facing in. Then bring the right hand up so that it is about 1 inches in front of the left hand, palm facing in. Align the Lung 10 point in the middle of the eminence of the left hand with the **throat center. Feel the Chi go through the neck to C-7**. Continue smiling to the universe and connect to the back side of the universe. Hold for a 5 to 30 count and feel the throat center activated.

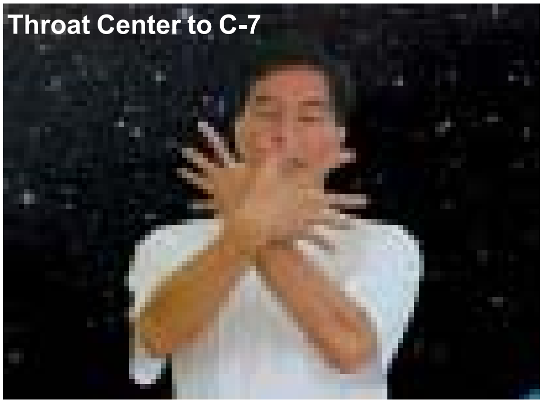

Fig. 8.37 *Connect to the Throat Center.*

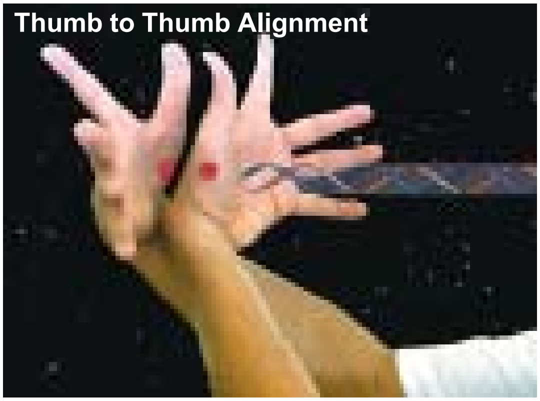

Fig. 8.38
Feel the connection to the Universe at the Back.

Move both hands out slowly like you're pulling silk; when the feeling is diminished or the silk starts to break stop there. Be still and feel the beam connect the throat center to C-7 and the universe. Maintain the energy beam as you move in and out three times altogether ending with the hands as above.

Activate Mideyebrow Point - Right Side

Be aware of the Tan Tien and the universe. Keeping the Lung 10 points aligned, raise the hands up to the level of the mideyebrow point about one inch in front of the face. **Keep the points aligned and beam the energy passing through both hands into the mideyebrow point and through to the base of the skull.** Continue smiling and extending to make a connection to the back universe for a 5 to 30 count.

Use only 5% of the awareness to move both hands very slowly out, like you are **pulling silk**; continue to move out as long as your can feel the connection to the mideyebrow, C-7 points and the universe, maintaining the energy beam as you move.

Move in and out three times altogether, ending with the hands as above.

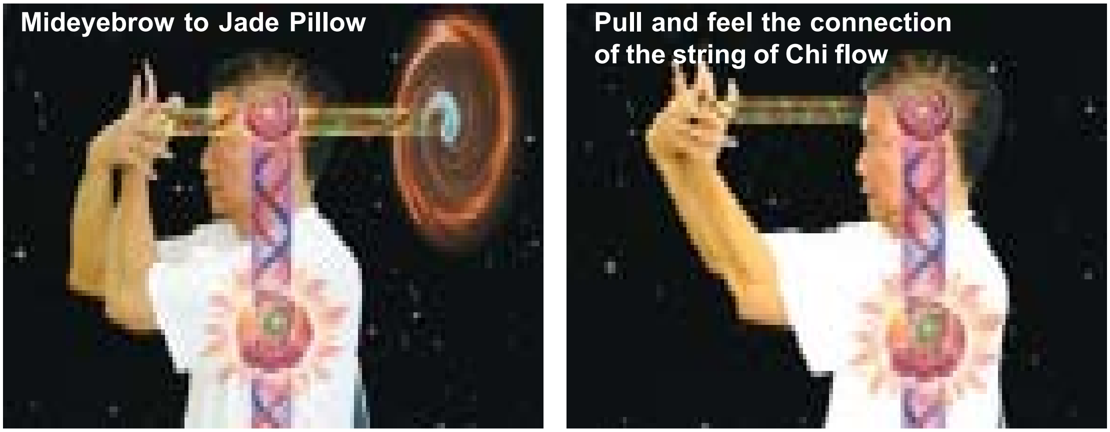
Mideyebrow to Jade Pillow
Pull and feel the connection of the string of Chi flow

Fig. 8.39 *Connect to the Mideyebrow, Jade Pillow and the Universe at the Back.*

Keeping the palms and thenars (Lung 10) aligned, move the hands back down to the position in front of the throat center. Keeping the forearms horizontal at shoulder height, separate the hands so that the Laogong point is aligned in front of Stomach 13, just **below the clavicle** on a vertical line above the nipple or mammillary line. The middle fingertips should almost be touching. **Beam energy from the Laogong points into ST-13 through the back on the side of the spine and connect to the universe**. You will be following this vertical line, which passes through the nipples, down the torso to the level of the navel, **stopping to focus and beam energy into key points along the way.** On each of these, focus your Yi on beaming energy into the organs inside each position. Stomach 13 is the location that activates the Heart and Lungs. Feel the change in your breathing as you activate the lungs.

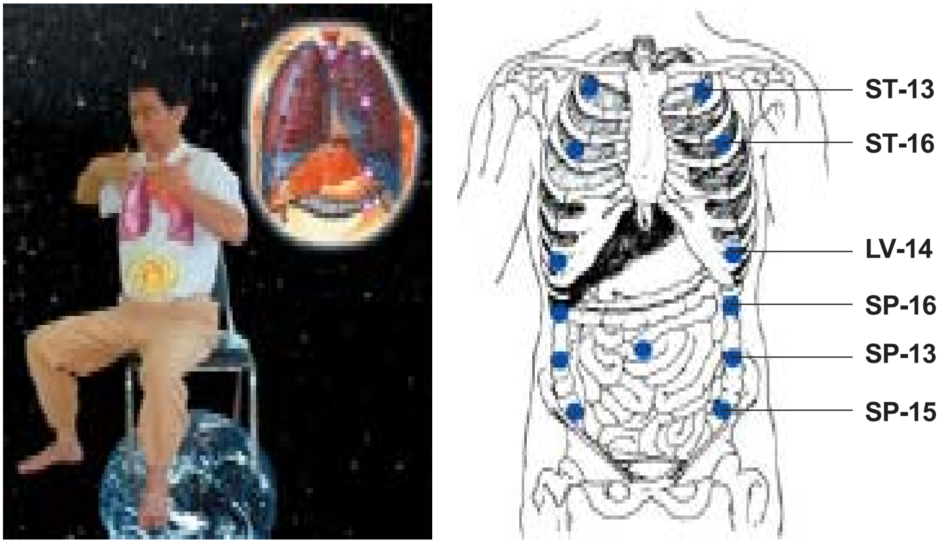

Fig. 8.40 *Beam Energy into ST13.*

Move the hands down the mammillary line to Stomach 16, about one inch above the nipple (in the space between the third and fourth ribs). Chi connects to the back and to the universe. Again, beam energy to energize and balance the **heart and middle of the lungs**. Move the hands down to Liver 14 (about 2 - 3 inches below the nipple in the space between the 6th and 7th ribs). Beam healing energy into the **liver and gall bladder.** Feel the Chi or the vibration of the palms activate the liver and feel the energy moving inside there.

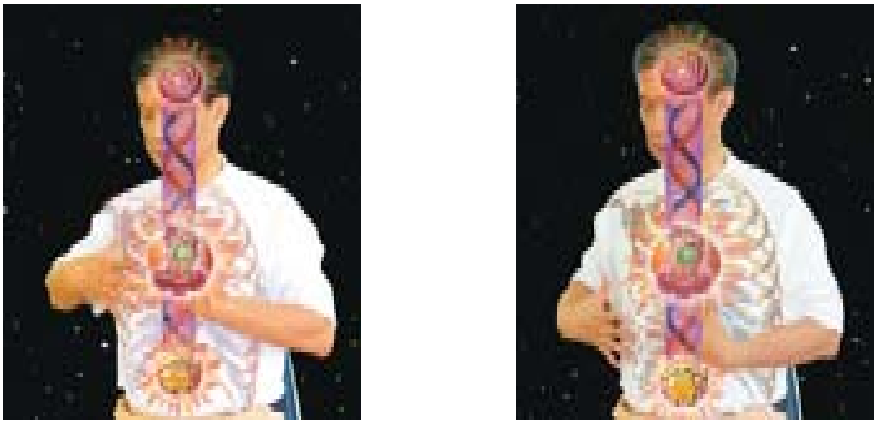
A. **B.**

Fig. 8.41 A. *Energize and balance the Lungs at ST16.*
B. *Energize and balance the Liver and Gallbladder at LV14.*

Move the hands down to **Spleen** 16 (just below the rib cage on the mammillary line). Beam healing energy into the **stomach, pancreas and spleen and the liver.** Picture the organs receiving the Chi from the cosmic and the universe.

Move the hands down to Spleen 15 (on the mammillary line to the level of the navel). Beam healing energy from your palms into the **small intestine and the Tan Tien.**

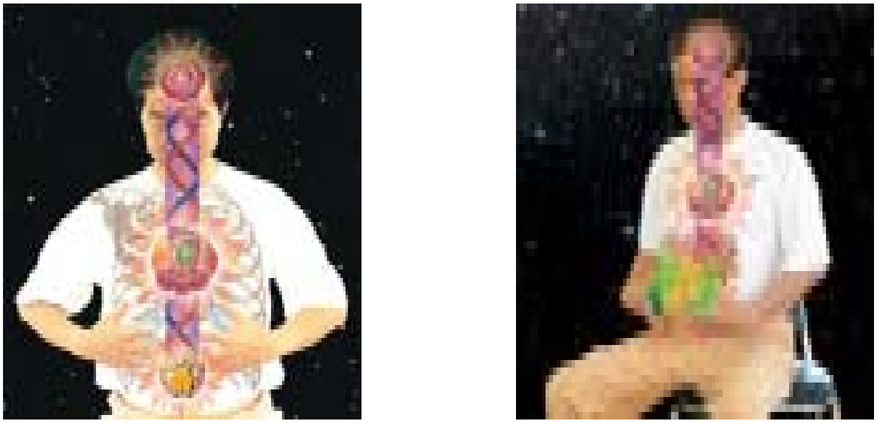
A. **B.**

Fig. 8.42 A. *Energize the Stomach, Pancreas, Spleen and Liver at SP 16.*
B. *Energize the Tan Tien and Small Intestine at SP15.*

Hold the Chi Ball - Right Hand on Top

Turn the **left palm up** and align the Chi Ball with the Tan Tien. Turn the **right palm down** and align the eye of the hand (LI-4) with the navel. Feel your palms holding a Chi ball and feel a **Chi ball inside the Tan Tien**. Feel the Chi ball connect to the Cosmic Chi outside. Feel the north (usually left) and south (usually right) poles in the hands. Lower the right thumb to point toward the left Laogong point.

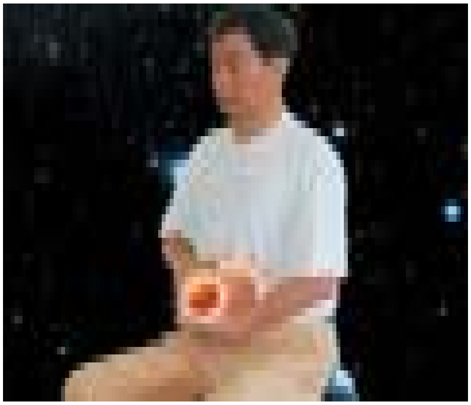

Fig. 8.43
Hold the 'Chi Ball' - One in the Tan Tien and the other in the hands - Right Hand over Left.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at solar plexus level. The **left palm** is still pointing up and is yang drawing in the **Heavenly Force** through Laogong. The **right palm** is still facing down and is yin drawing in the **Earth Force** energy. **'Be Aware Tan Tien and Universe'**.

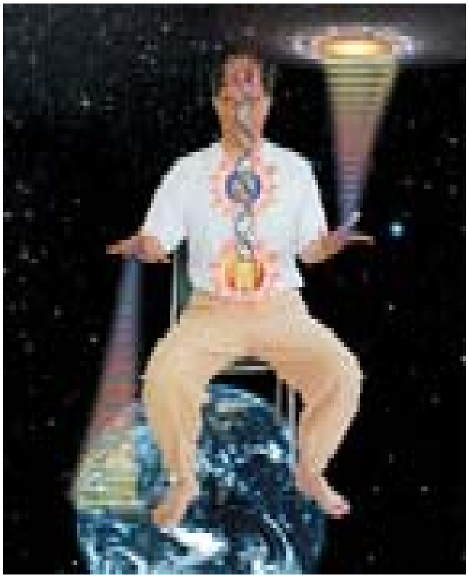

Fig. 8.44 *Connect to Heaven and Earth.*

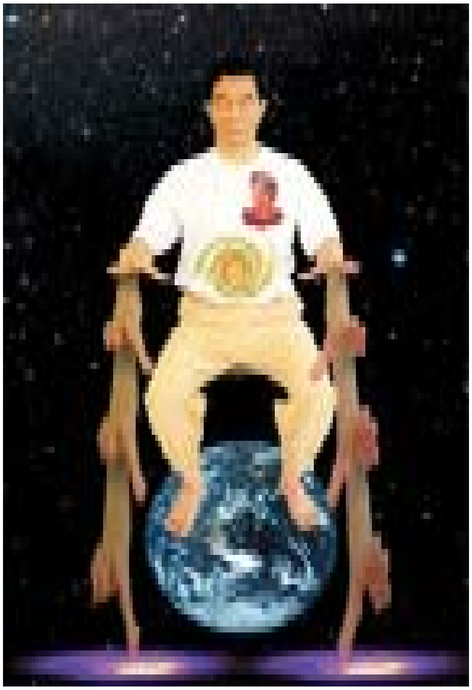

Conclude by turning the left palm over to face downward. **Draw in the 'Earth Force' through both palms.**

Fig. 8.45 *Draw the Earth Force through the Palms.*

Repeat the sequence with the left palm over the right.

Start by moving the right palm to the throat and then the left palm. Scan the body, as in the previous sequence, using both palms to connect to the points running down the front of the body. Remember to maintain a sense of the Chi flowing between the points in the front and their corresponding points in the back of the body.

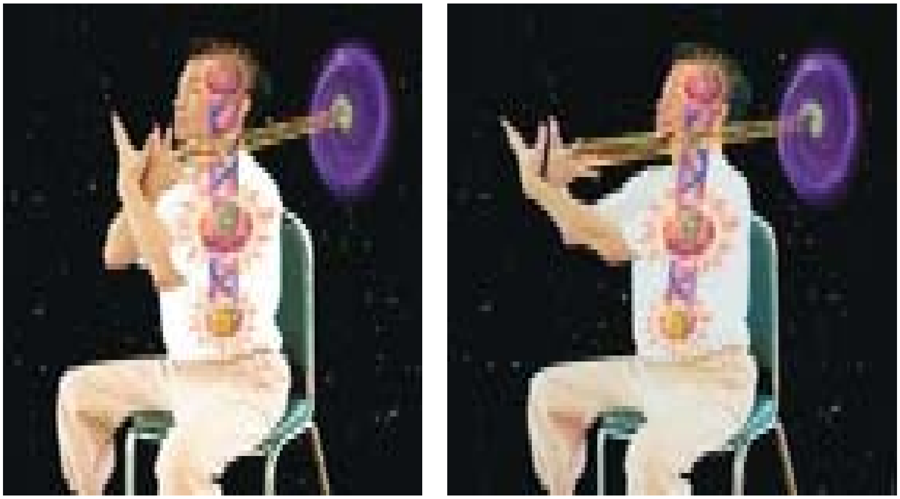

Fig. 8.46 *Repeat the Practice on the Left Side. Push and Pull the silk at the Throat.*

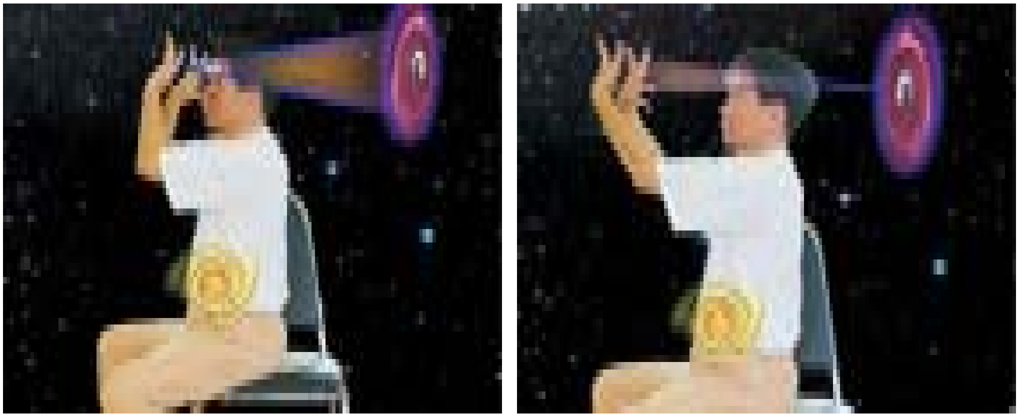

Fig. 8.47 *Project to the Mideyebrow and Base of the Skull - connect to the Universe at the back. Then hold the Ball and repeat Finger Movements, as below.*

Continue with the left side, now pulling the silk at throat (C-7).

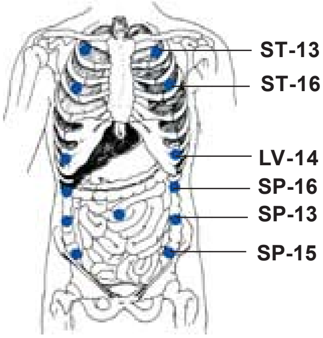

Fig. 8.48 *Beam Energy into each of the Organs like the last time.*

Closing Movements

The closing movements, crane's beak and finish are the same as before. Activate the Chi channels - balancing the organs.

Index Fingers

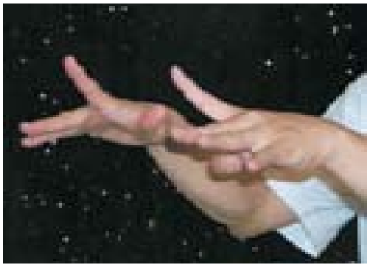

Fig. 8.49 *Open (up)*

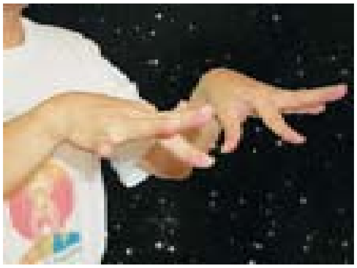

Fig. 8.50 *Neutral*

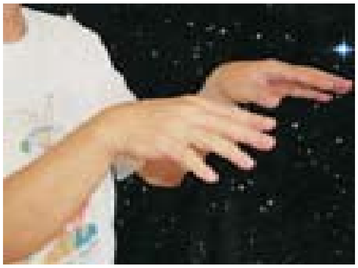

Fig. 8.51 *Close (down)*

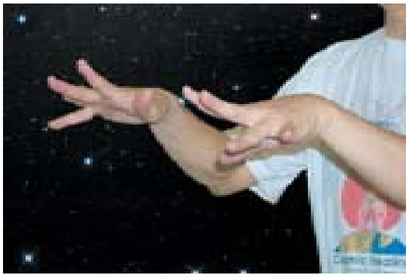

Fig. 8.52
***Ring Finger** - down, neutral.*
***Index Finger** - open, neutral.*

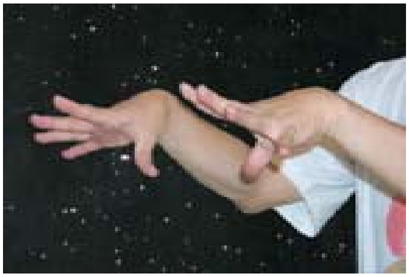

Fig. 8.53
***Thumb** - in and down, neutral.*
***Index Finger** - open, neutral.*

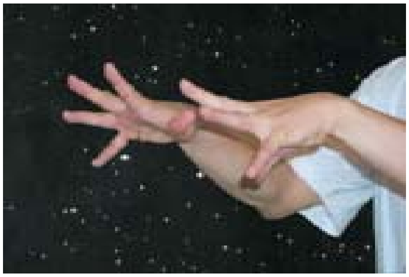

Fig. 8.54
***Pinky Finger** - down, neutral.*
***Index Finger** - open, neutral.*

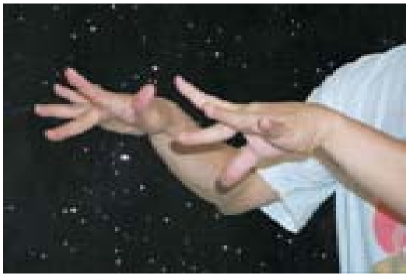

Fig. 8.55
***Middle Finger** - down, neutral.*
***Index Finger** - open, neutral.*

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Activate the saliva and swallow down three times to the Tan Tien. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise. Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

Bring the awareness to the navel and notice the quality and intensity of the energy generated. **Collect energy at the navel**. Men, place the hands over the navel with the right hand on top; vice versa for women.

Part III

Opening the Functional and Governor Channels

Theory: Functional and Governor Channels

The Functional and Governor Channels are the first two of the eight extraordinary channels described earlier in this chapter (beginning p.178). They serve as the reservoirs of yin and yang energy in the body. The Functional or Conception Channel is yin and all of the ordinary yin channels connect to it. Similarly, the Governor Channel is yang and all of the ordinary yang channels connect to it.

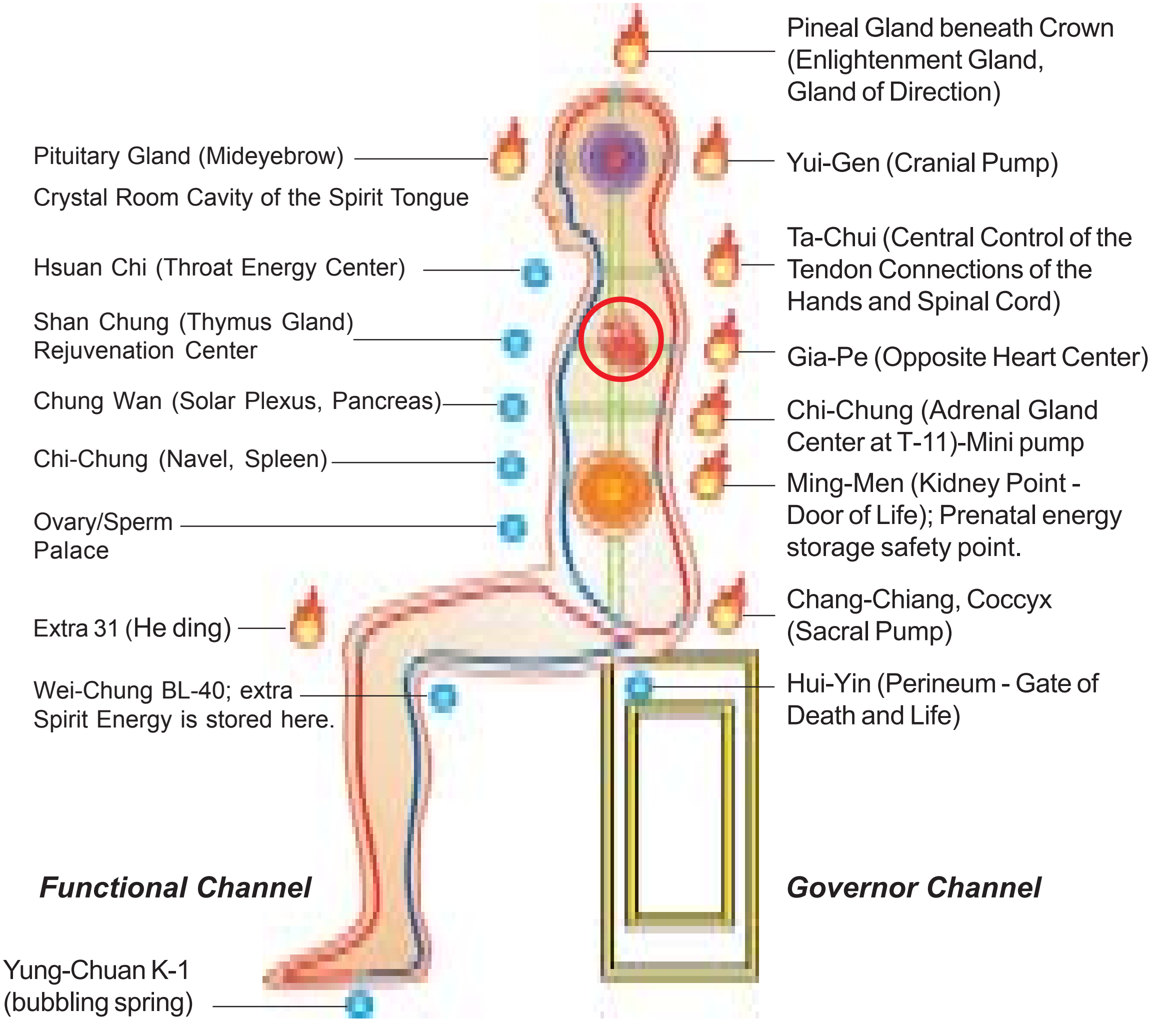
Pineal Gland beneath Crown (Enlightenment Gland, Gland of Direction)
Pituitary Gland (Mideyebrow)
Crystal Room Cavity of the Spirit Tongue
Yui-Gen (Cranial Pump)
Hsuan Chi (Throat Energy Center)
Ta-Chui (Central Control of the Tendon Connections of the Hands and Spinal Cord)
Shan Chung (Thymus Gland) Rejuvenation Center
Gia-Pe (Opposite Heart Center)
Chung Wan (Solar Plexus, Pancreas)
Chi-Chung (Adrenal Gland Center at T-11)-Mini pump
Chi-Chung (Navel, Spleen)
Ming-Men (Kidney Point - Door of Life); Prenatal energy storage safety point.
Ovary/Sperm Palace
Extra 31 (He ding)
Chang-Chiang, Coccyx (Sacral Pump)
Wei-Chung BL-40; extra Spirit Energy is stored here.
Hui-Yin (Perineum - Gate of Death and Life)
Functional Channel
Governor Channel
Yung-Chuan K-1 (bubbling spring)

Fig. 8.56 *Cosmic Orbit*

Chi Ball

In Part III we also begin to focus more on the sensation of Chi being emitted from our palms. This energy is emitted from the **Laogong** points primarily, but you can also feel it throughout the palms and fingers. It will feel like a balloon between your hands that is expanding and contracting or like two magnetic fields repelling and attracting each other.

Once you have practiced Cosmic Chi Kung for awhile and have become skillful at it, you may find that Grasping the Chi ball is all you need to do to activate your flow of healing energy before doing a session. As you become sensitive to the sensation of your own Chi, it becomes an easy step to feel the Chi of others.

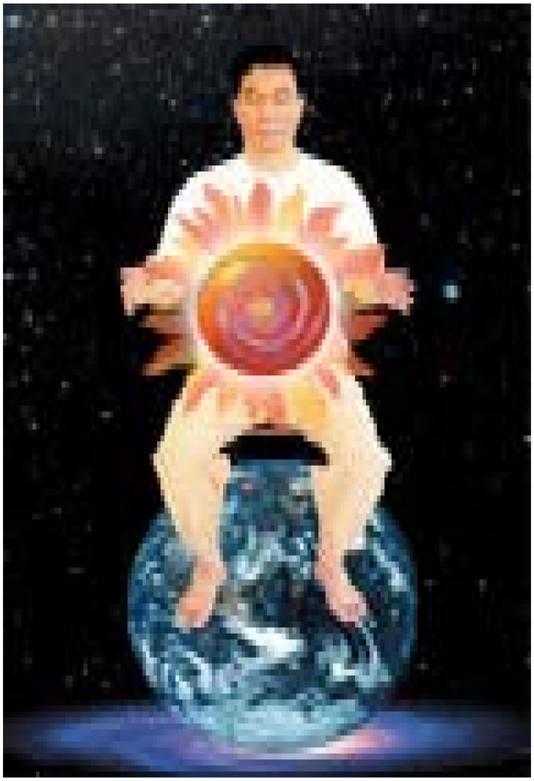

Fig. 8.57 *Grasping the Chi Ball.*

Opening Movements, Hands Down at Sides, Palms Facing Each Other
Bone Breathing - Absorbing Cosmic Chi

The Opening Movements are the same as in the sequence in Part I, 'Opening to Heaven, Earth and Cosmic Chi', on page 164.

A. Channeling the Earth Force - Washing the Bone Marrow Raise arms to chest height - Palms facing each other. Rotate arms until **palms face down.** Open the Tiger's Mouth - Index finger up, thumb down.

B. Activate and Beam to the Eye of the Hip

C. Absorb the Heavenly Force Washing the Bone Marrow - Crown to Soles

Core Movements of Part III

Grasping the Chi Ball

Slowly rotate the hands so that your palms face each other in front of your navel.

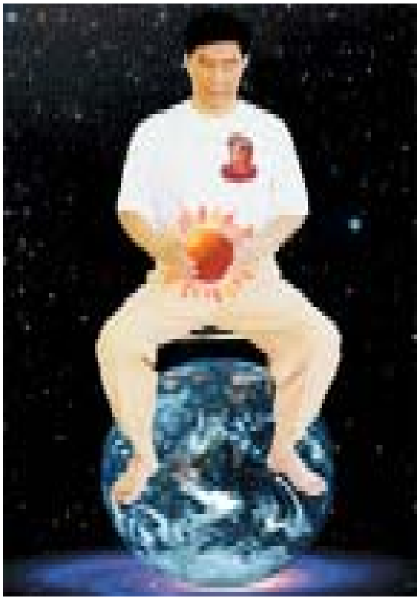

Fig. 8.58 *Hold the Chi Ball.*

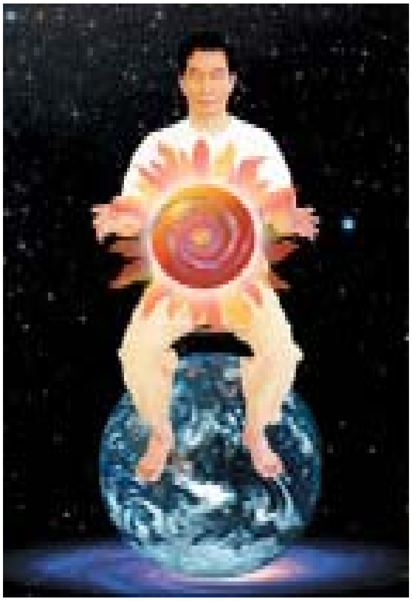

Fig. 8.59 *Stretch and squeeze it.*

Feel the invisible ball of energy between your hands. Allow the energy to push your hands apart keeping the feeling of connection between your palms. When you have less feeling stop the palms and hold it there and reestablish the feeling of the Chi ball.

Let the energy draw your hands back toward each other, like a magnet, until you feel as if you were squeezing and compressing the ball.

Stretch and squeeze the Chi ball three times by opening and closing your palms in this way.

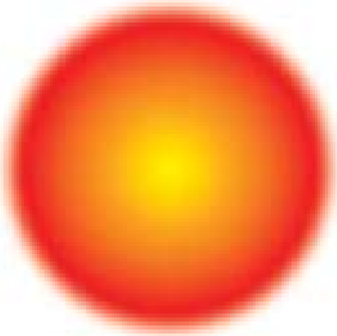

Activate the Outer and Inner Arm Gates

Activate Waiguan Triple Warmer 5 and Neiguan Pericardium 6 - **Right** arm on top. Slowly rotate the arms so that both **palms face down.**

Cross the right wrist over the left wrist with 1 inch of space between the hands. Align PC-6 of the right wrist with TW-5 of the left wrist. Feel the two gates activate each other like metal and magnet drawn to each other. Hold for a 5 to 30 second count.

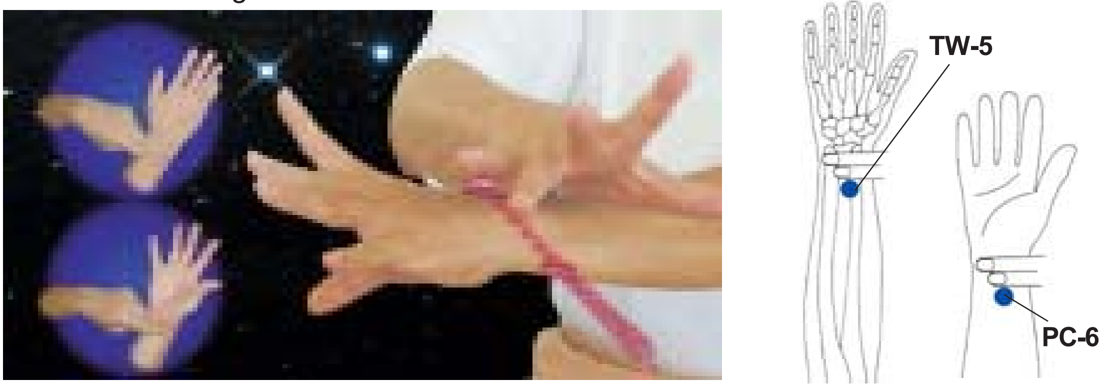

Fig. 8.60 *Align and activate PC-6 (Right) with TW-5 (Left).*

Keeping the wrists crossed, slowly rotate hands **turning the palms upward**. Now TW-5 of the right wrist will be aligned with PC-6 of the left wrist. Hold and count and feel the points activated. 'Be aware of Tan Tien and Universe'.

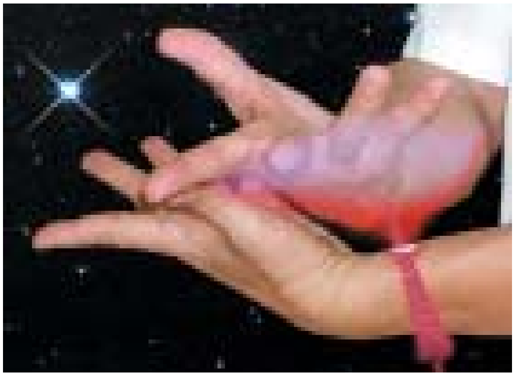

Fig. 8.61
Align and activate TW-5 (Right) with PC-6 (Left).

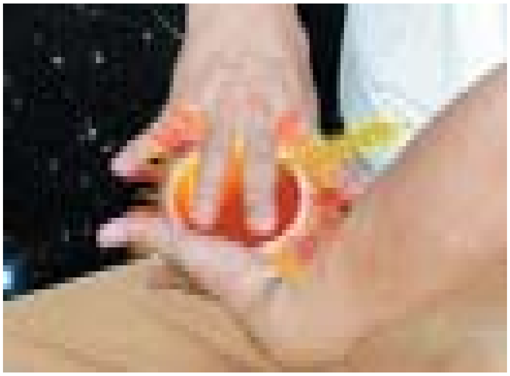

Fig. 8.62 *Feel the Chi Ball and beam energy from LI-4 into the Navel Point.*

Open the Functional Channel with LI-4

Bring the left hand down to the level of the Tan Tien with the Chi Knife facing in. Lower the right hand, palm down and align the eye of the hand, the LI-4 point with your navel. **Feel the Chi ball between your two hands**, and at the same time beam energy from LI-4 into the **navel** point.

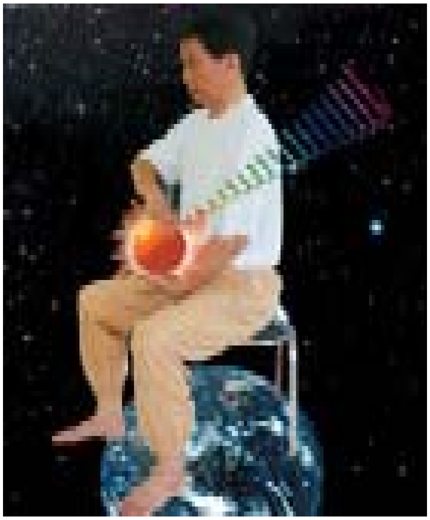

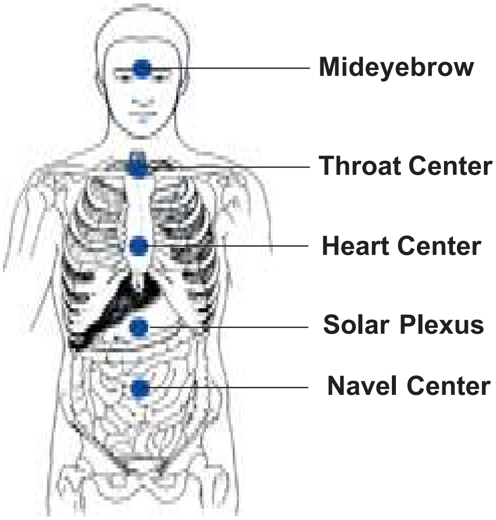
Mideyebrow
Throat Center
Heart Center
Solar Plexus
Navel Center

Fig. 8.63
Magnify the Power of the Beam to the Solar Plexus.

Fig. 8.64 *Activate all the Points.*

Raise the right hand up the Functional Channel, stopping and aligning LI-4 with the **solar plexus**, hold for a 5 to 30 second count and feel the points open and starting to activate each other.

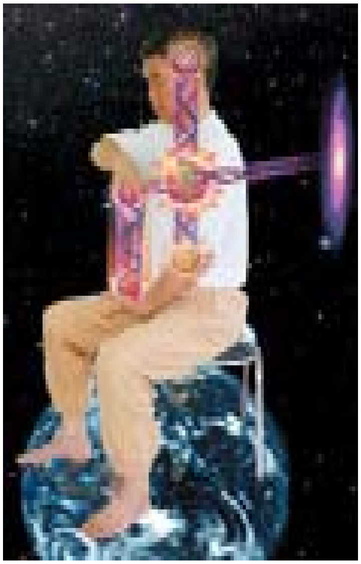

Fig. 8.65
Beam to the Heart Center.

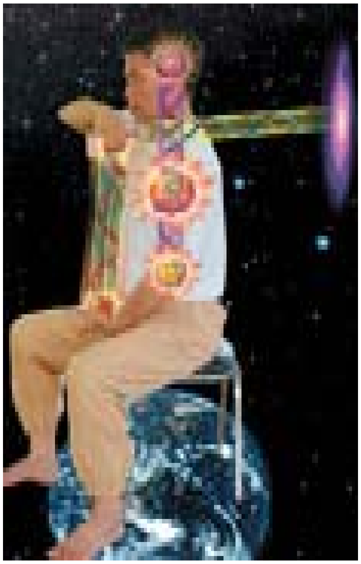

Fig. 8.66
Beam to theThroat

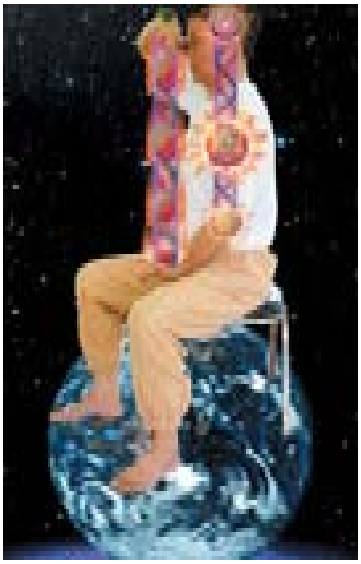

Fig. 8.67
Beam to the Mideyebrow

Continue to move up to the **heart**, **throat** (CV-22) and **mideyebrow** points. Beam energy from the eye of the hand into each of these points and at the same time, continue to feel the Chi ball connection between the two hands.

Return down the Functional Channel in the same manner, starting from the mideyebrow and stopping to beam energy into each point.

Next beam into the **throat center, heart center, solar plexus** and finally the **navel**. Hold the Chi ball at the navel with the right hand over the left.

Yin/Yang Palms

Separate the hands and extend them out in front of the body at solar plexus level. The left palm is still pointing up and is yang drawing in the heavenly force through the Laogong point. The right palm is still facing down and is yin drawing in the earth force energy.

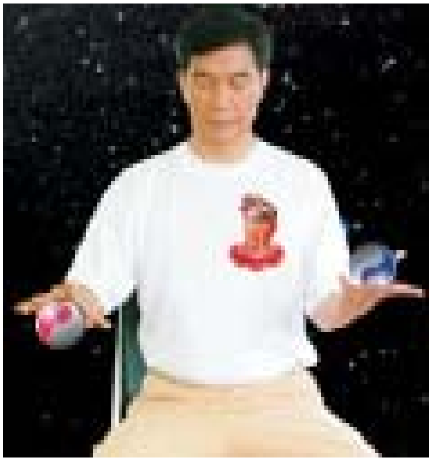

***Fig. 8.68** Separate the Hands, one palm facing up and one down - draw in the Forces of Heaven and Earth.*

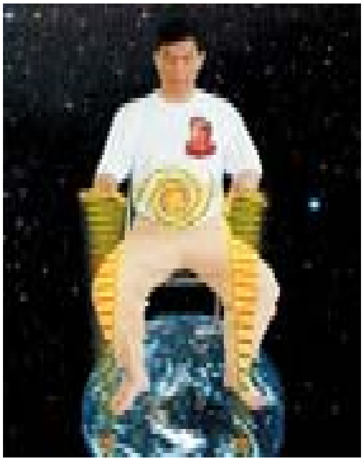

Fig. 8.69
Draw in Earth Force.

Conclude by turning the left palm over to face downward. Draw in the earth force through both palms.

Activate the Chi Ball, Tan Tien and the Original Force

Turn the right palm upward and lower it to the Tan Tien with the Chi knife pointing inwards.

Lower the left palm to the navel, palm down. Lower the left thumb to point toward the right Laogong point. This opens the eye of the left hand (LI-4), pointing it toward the navel.

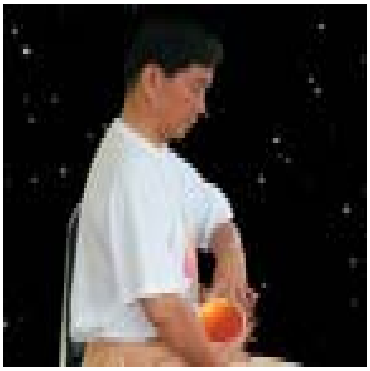

***Fig. 8.70** Activate the Chi Ball, Tan Tien and the Original Force - Right Hand over left, the 'Large Intestine 4' pointing at the Navel.*

Open the Functional Channel with the Left Hand

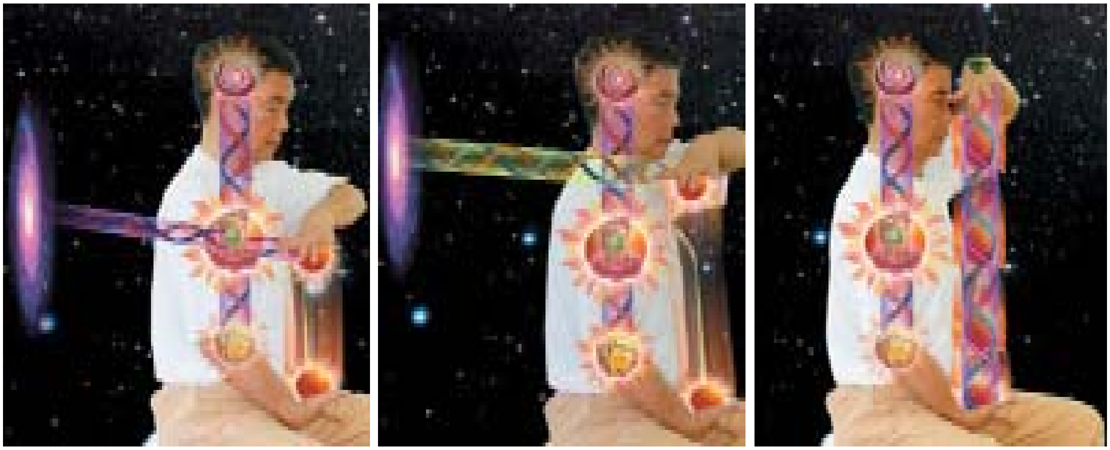

A. *B.* *C.*

Fig. 8.71 A. *Bring LI4 point up to the Solar Plexus and Heart Center.*
B. *Then up to the Throat Center.*
C. *Then up to the Mideyebrow. Descending down the channel in reverse, finishing at the Navel.*

After going up the functional channel come down the channel with the left hand.

Holding a Chi Ball: Open the hands with the palms facing each other holding a Chi ball. Stretch and squeeze the Chi ball a few times in order to feel the Chi as a substance; open and then relax the palms as you squeeze.

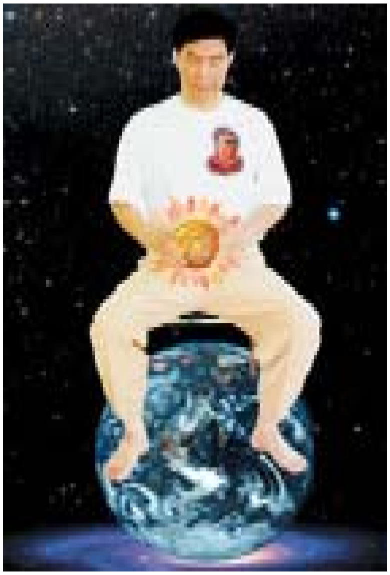

Fig. 8.72 *Hold and squeeze the Chi Ball.*

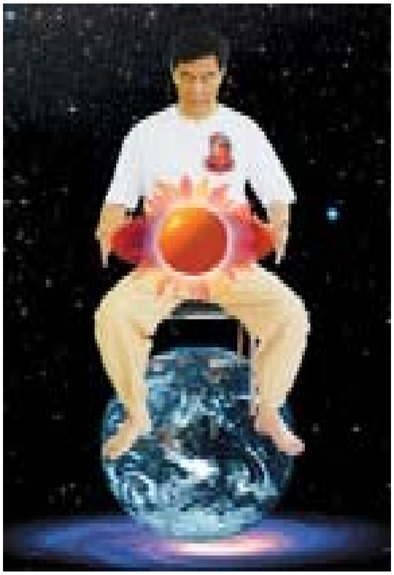

Fig. 8.73
Feel the Chi Ball expanding, growing Bigger.

Double Palm and Double Beam - Palm to Palm
Activate the Functional and Governor Channels - Right Palm on Left

Bring the left palm in first to face the navel, about 1 inch away, aligning the Laogong point with the navel. Follow by bringing the right hand in front of the left hand, **aligning the Laogong of the right palm with the Laogong of the left hand and the navel**. Both palms are now facing in. Beam energy through both Laogong points to the **navel and through the body to the Door of Life**, between L-2 and L-3 (Lumbar Vertebrae). Hold for a 5 to 30 second count and feel the vibration of the palms and Chi beam penetrate through the two points; feel them link together.

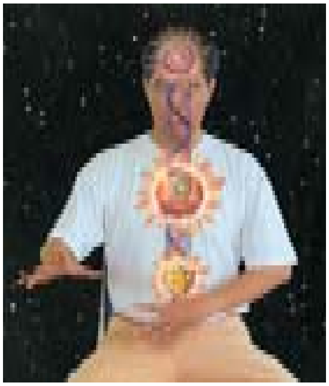

Fig. 8.74
Left Palm to Navel.

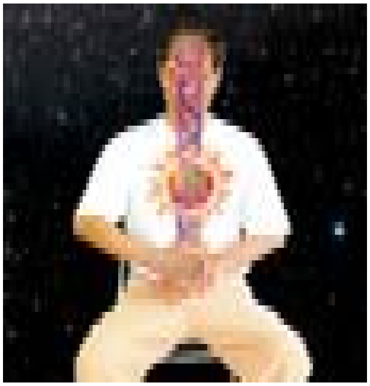

Fig. 8.75
Right on Top.

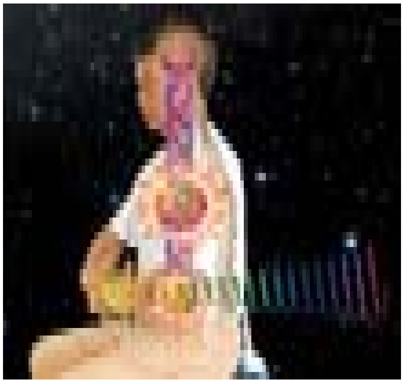

Fig. 8.76
Beam from Navel to Door of Life.

Next raise the left palm up to the solar plexus, aligning the Laogong point with the solar plexus.

Follow with the right palm aligning the Laogong to the left palm. Beam energy into the **solar plexus point and through the body to the T-11 point (thoracic vertebrae)**.

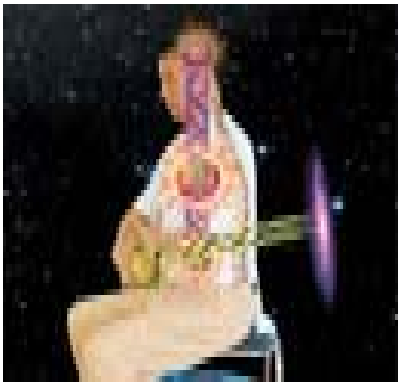

Fig. 8.77
Raise Left Palm to Solar Plexus.

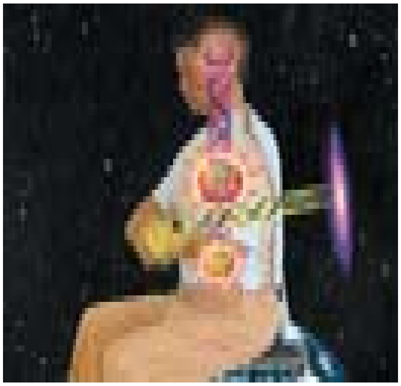

***Fig. 8.78** Follow with the Right Hand to form the Double Beam.*

Continue in the same way with the left hand leading to the **heart point and wing point,** the **throat point and C-7 (cervical vertebrae)**, the **mideyebrow point and Jade Pillow**, and the **crown point and perineum**. **Always remember 'Tan Tien and the universe'.**

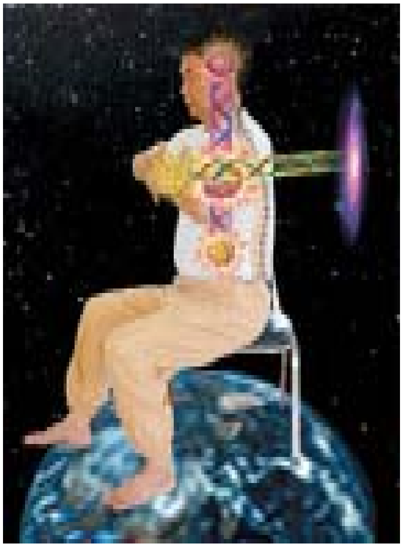

Fig. 8.79 *Beam into the Heart Center and to the back of the Heart.*

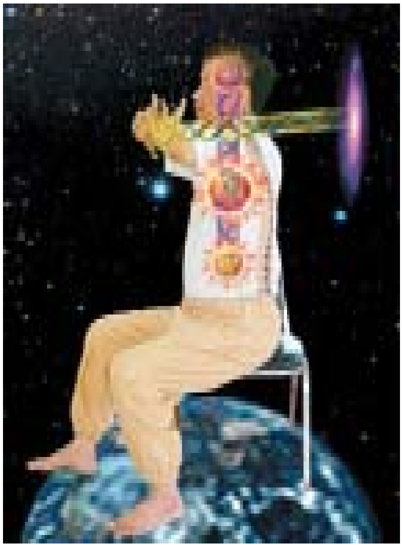

Fig. 8.80 *Beam into the Throat Center and to C7.*

When you reach the crown lightly spiral the palms and feel the Chi slowly penetrate deep into the body reaching the perineum; this will also open the **Thrusting Channels.** This might take a longer time, try a 30 to 60 second count.

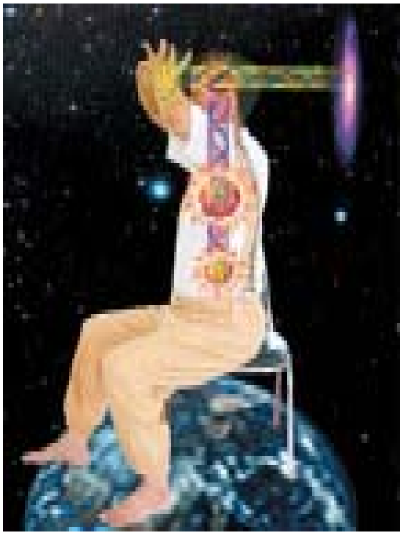

Fig. 8.81 *Beam into the Mideyebrow to the base of the Stall.*

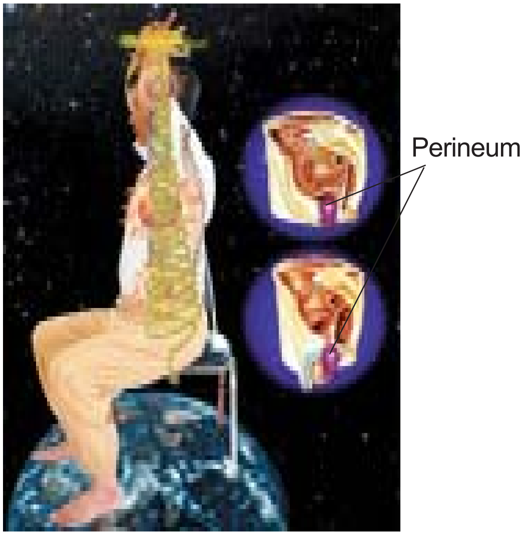

Fig. 8.82 *Beam from the Crown down to the Perineum.*

Return down the front, point by point in the same way, leading with the left hand.

Go up and down this way three times. The double palm and beam activate both the Functional and Governor Channels. The energy beam passes all the way through the body at each point.

Grasping the Chi Ball

Separate the hands and slowly extend the arms in front of you at navel level, palms facing each other, holding the Chi ball.

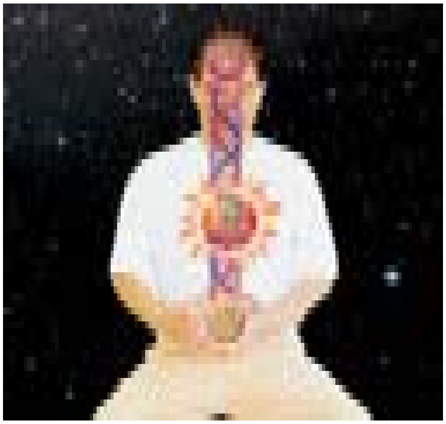

***Fig. 8.83** Squeeze the Chi Ball.*

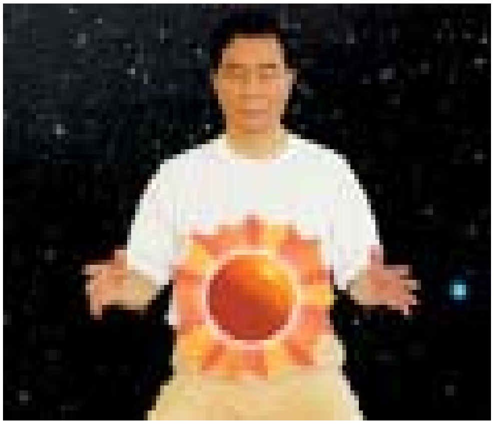

***Fig. 8.84** Magnify its Energy.*

Stretch and squeeze the Chi ball three times as before by opening and closing the palms. 'Be Aware of Tan Tien and Universe'.

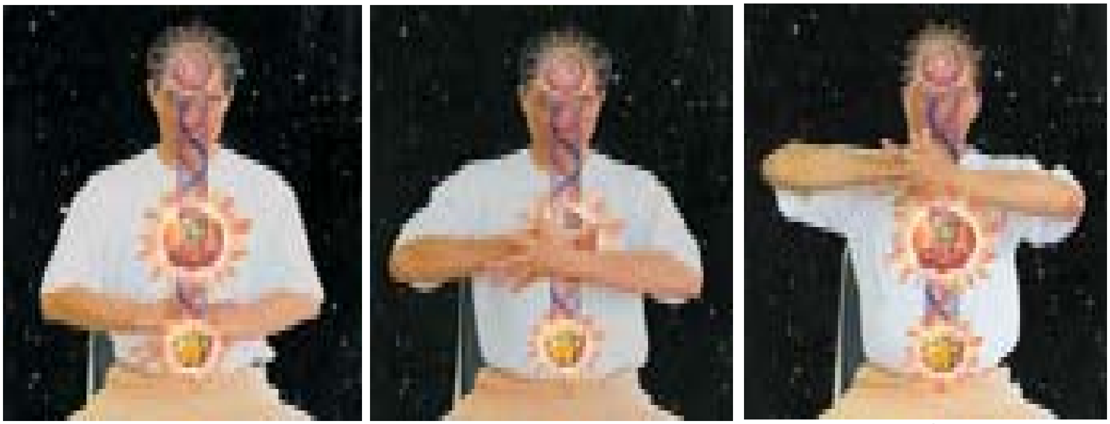

A. ***B.*** ***C.***

***Fig. 8.85** Repeat the double palm and beam with the **left hand on the outside and the right palm leading,** beaming from; **A.** Navel to Door of Life. **B.** Solar Plexus to T11 and Heart Center to the Wing Point **C.**Throat Center to C7.*

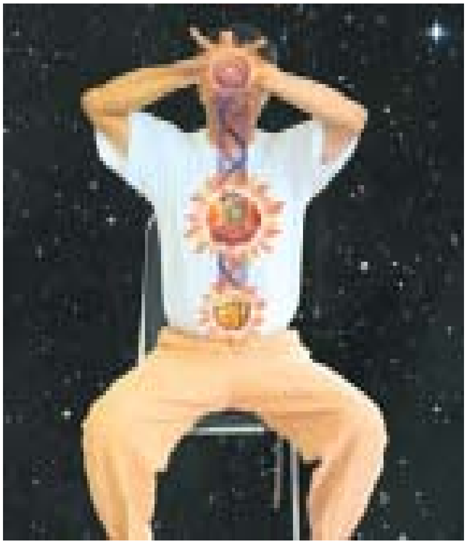

Fig. 8.86 *Activate the Mideyebrow to the Base of the Skull.*

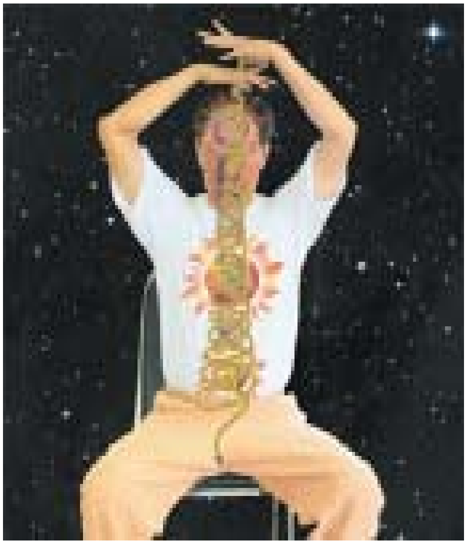

Fig. 8.87 *Activate the Crown and feel the Energy penetrate to the Perineum.*

A. Lower the right hand to the mideyebrow and penetrate to the skull.
B. Lower the right hand down to the throat center, following with the left.
C. Lower down to the solar plexus and penetrate to T11.
D. Lower down to the navel and penetrate to the Door of Life.

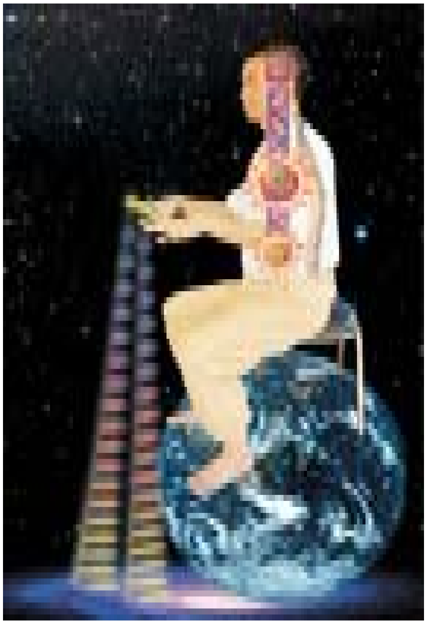

Fig. 8.88 *Channel the Earth Force.*

Rotate the **hands to face downward** and draw in the earth force through the palms, soles and perineum.

Closing Movements

Activate the Chi Channels - Balancing the Organs

The closing movements, crane's beak and finish are the same as before.

Index Finger *open (up), close (down), neutral (even with the other fingers)*
Ring Finger *- down, neutral.*
Index Finger *- open, neutral.*
Thumb *- in and down, neutral.*
Index Finger *- open, neutral.*
Pinky Finger *- down, neutral.*
Index Finger *- open, neutral.*
Middle Finger *- down, neutral.*
Index Finger *- open, neutral.*

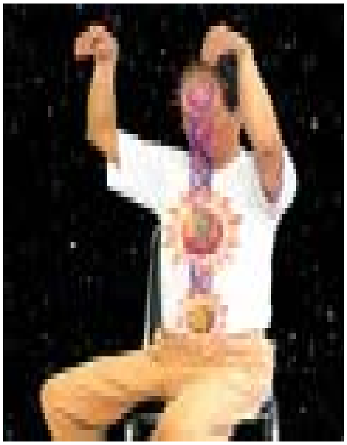

Fig. 8.89 *Form the Crane's Beak.*

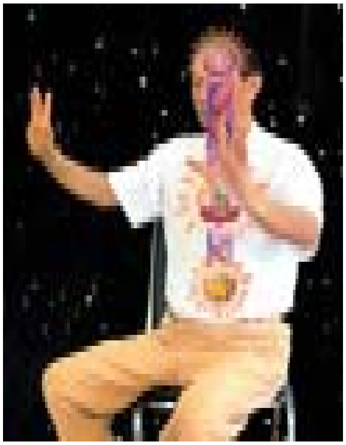

Fig. 8.90 *Palms to Front.*

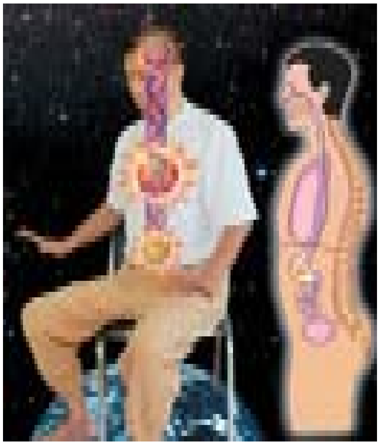

Fig. 8.91 *Lower to Sides.*

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Activate the saliva and swallow this elixir down three times to the Tan Tien. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.

Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

Bring awareness to the navel and notice the quality and intensity of the energy generated. Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Part IV
Activating the Yin and Yang Channels and the Chi Belt

Theory

In this final part of Cosmic Chi Kung, you will focus on opening the yin and yang channels of the arms and on activating the Chi Belt around the waist. Mastery of Part IV of Cosmic Chi Kung will enhance your ability to project Chi through your fingers for healing.

Yin and Yang Channels of the Arms

There are six important energy channels flowing through the arms. These channels are paired together with each pair comprised of a yin channel and a yang channel. The yin channel of each pair flows down the inside of the arm from torso to fingertip, while its yang counterpart flows up the outside of the arm from fingertip to head.
The pairs are as follows:

Metal Element - Yin - Lung Channel - Thumb
Yang - Large Intestine Channel - Index Finger
Fire Element - Yin - Pericardium Channel - Middle Finger
Yang - Triple Warmer Channel - Ring Finger
Fire Element - Yin - Heart Channel - Pinky Finger
Yang - Small Intestine - Pinky Finger

Projecting Healing Energy through the Fingers

Emitting Chi from the fingers for healing is known in China as the **"One Finger Art."** Because the various energy channels of the arms either end or begin at the fingertips, the fingers are very effective instruments for projecting healing energy. Your fingers can focus energy like a laser beam toward a concentrated area such as a specific acupuncture point. You can also emit energy from all the fingers at once creating a combined effect to target an area. Part IV of Cosmic Chi Kung will stimulate all of the arm channels and will give you an opportunity to activate all of the fingers for beaming energy.

Fig. 8.92 *"One Finger Art"*

Activating the Chi Belt - Dai Mo

The Chi Belt or Belt Channel (also called the Girdle Vessel) is the only channel in the body that runs horizontally. It encircles the body at the waist level, connecting together all the vertical channels running through the torso. Thus, the Belt Channel plays an important role in maintaining a good energetic communication between the upper and lower body. In women it strongly affects the uterus and the menstrual cycle in particular.

In Taoist Chi Kung, the Belt Channel is not limited to the waist region. It encircles the entire body, almost like weaving an energetic cocoon around you from head to foot. Activating the Belt Channel strengthens the aura and helps to defend and protect you from outside negative energies.

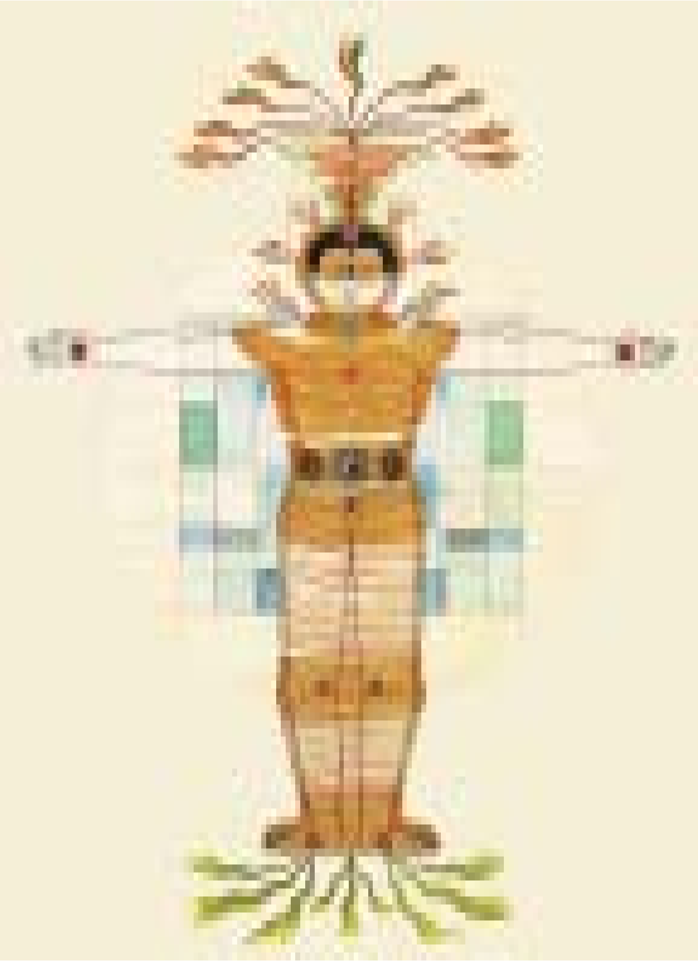

Fig. 8.93 *Belt Channel*

Practice

Opening Movements

These opening movements have been described in Parts I, II, III. Refer to Part I to refresh your practice.

Hands Down at Sides - Palms Facing Each Other - Bone Breathing - Absorbing Cosmic Chi.

Channeling the Earth Force - Washing the Bone Marrow.
Raise arms to chest height - Palms facing each other.
Rotate arms until palms face down.
Open the Tiger's Mouth - Index Finger up, thumb down.

Activate and Beam to the Eye of the Hip.
Large Intestine 4 - Eye of the Hand.
Pericardium 9 - Tip of the Middle Finger.
Small Intestine 3 - Knife Edge of the Hand.

Absorb the Heavenly Force - Washing the Bone Marrow - Crown to Soles.
Extend the arms forward to chest height - palms upward.

Absorb the Earth Force.
Rotate the arms until the palms face down.
Raise the Index Fingers slightly.

Fig. 8.94 *Mother Earth's Healing Power*

Core Movements of Part IV Yin and Yang Channels - Chi Belt

Activating the Yin Channels - Left Side

Turn the palms down and slightly open the index fingers. Smile to your palms and soles and channel the earth force.

Turn your right palm up and pass it just one inch below the left arm yin channels from the left palm to the armpit, **without physically touching the arm**. Feel the Chi moving. Turn your right palm to face your left rib cage and pass your right hand down the left side of your abdomen along the descending colon, and a cross the pelvis to the right side to activate the **yin** energy.

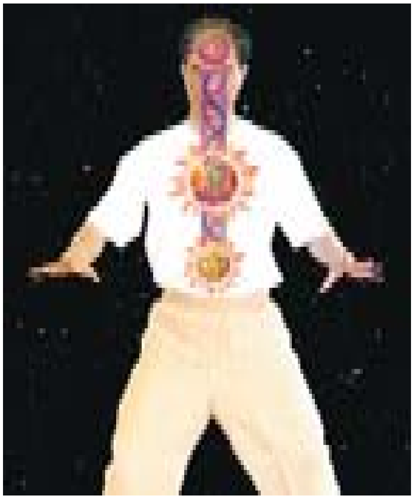

Fig. 8.95 *Channel the Earth Force.*

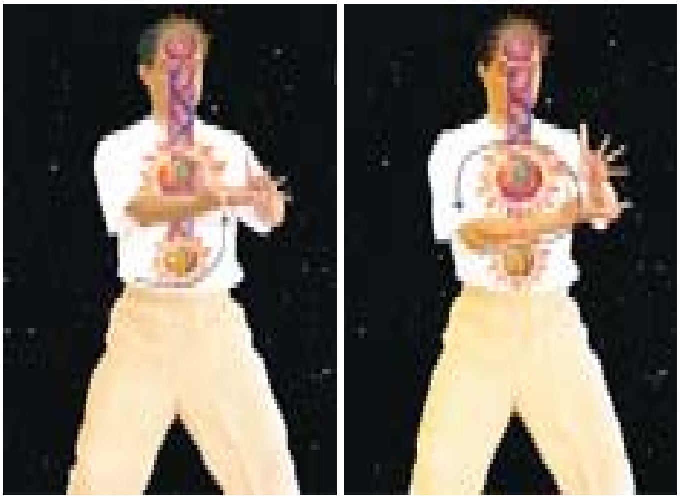

Fig. 8.96 *Activate the Yin Energy from the Palm to the Pelvis.*

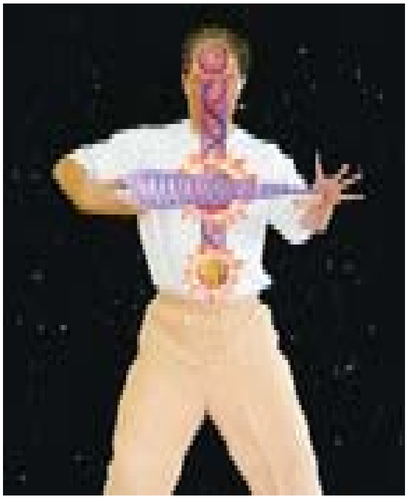

Fig. 8.97 *Raise Chi in the Right Palm.*

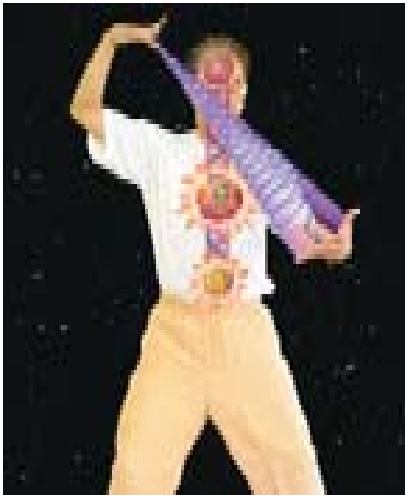

Fig. 8.98 *Pass it to the Left.*

Pass your right hand up the right side of the abdomen to the level of your forehead with the palm facing diagonally toward the left palm (which simultaneously turns palm up at the level of the navel). Project the Chi from palm to palm for 30 to 60 seconds, vibrating the palm and feeling Chi from the right palm sent to the left picking up by the left palm. Keep the palms open and relaxed.

Next point the right ring finger toward the left palm and project Chi making very small circles from the right ring finger (Triple Warmer channel); emit the Chi to the left palm for 30 to 60 seconds.

Then return the right index finger to neutral position. Point the left ring finger up toward the right palm, project into the right palm for 30 to 60 seconds, and then return to neutral.

Next both ring fingers project toward the opposite palms at the same time for 30 to 60 seconds and then return to neutral. Notice that the energy may meet in the middle.

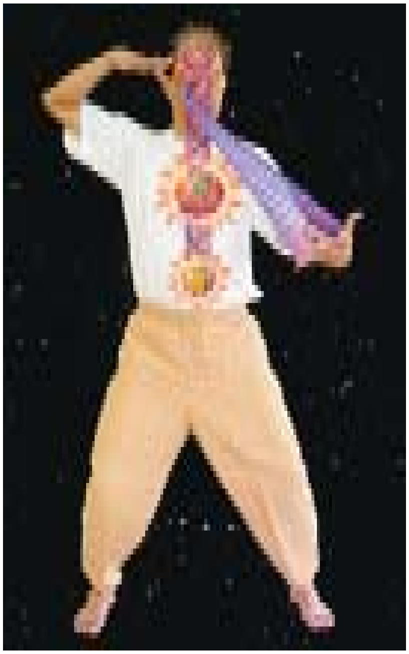

Fig. 8.99 *Project Energy and balance Left and Right.*

Activate the Yang Channels - Left Side

Turn the left palm down and lower the right palm to cover the back of the left hand, palm down. Pass the right palm over the left arm's **yang** channels from the back of the left hand to the left shoulder. Then pass the right palm across the upper chest and across the right breast. Scoop the right palm under the right armpit, fingers first. Then bring the right palm forward, palm down and press forward with the left palm at the same time. Relax and channel the earth force.

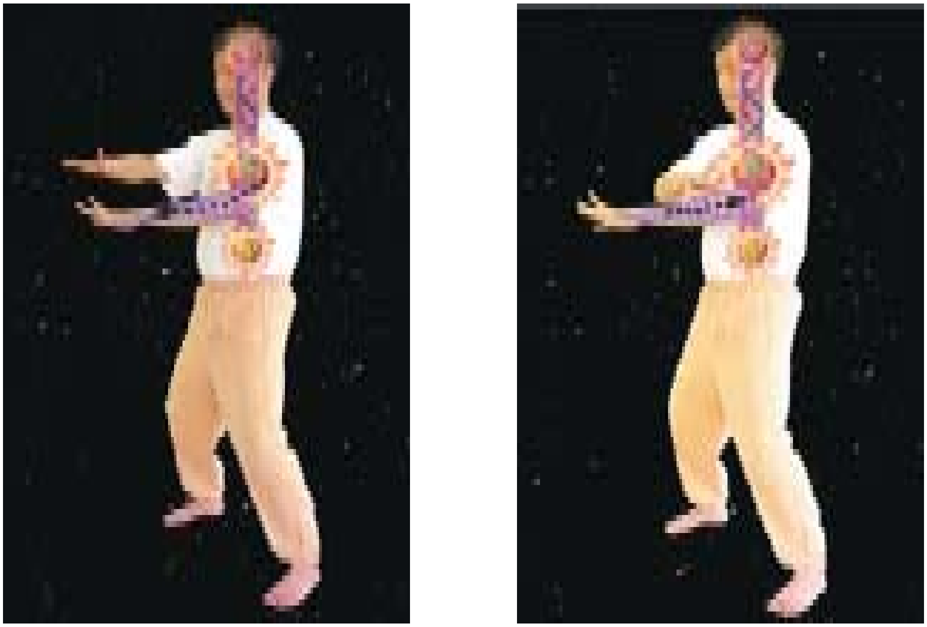

A. *Left Palm down, lower Right Palm to cover the back of Left Hand.*

B. *Pass the Right Palm over the Left Arm's Yang Channels.*

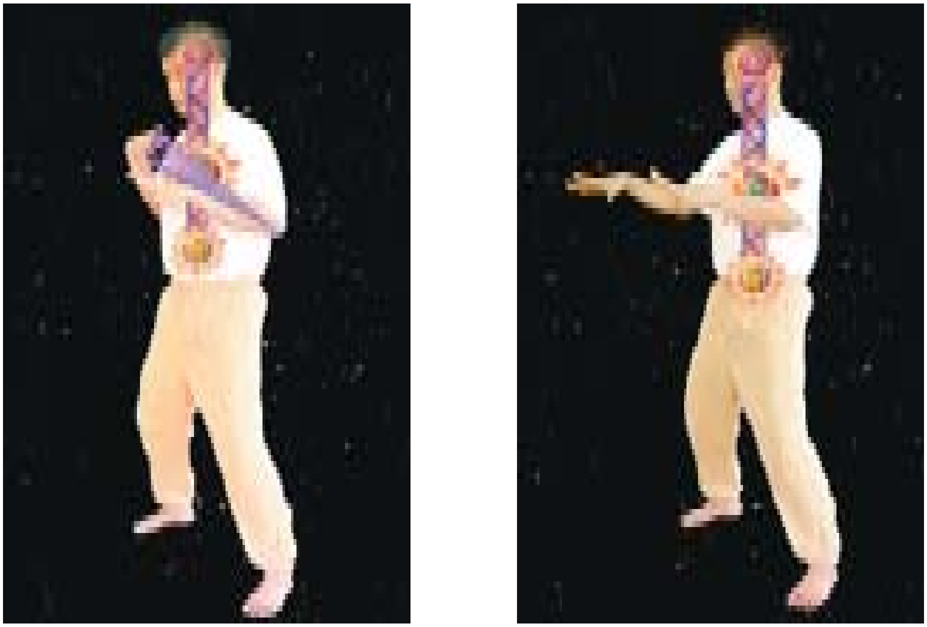

C. *Pass the Right palm across the chest and scoop under the Right Armpit, Fingers first.*

D. *Press both Palms forward, relax and channel the Earth Force.*

Fig. 8.100 *Balance Left Side Yin and Yang and Channel Earth Energy.*

Repeat, Activating the Yin and Yang Channels to the Right Side

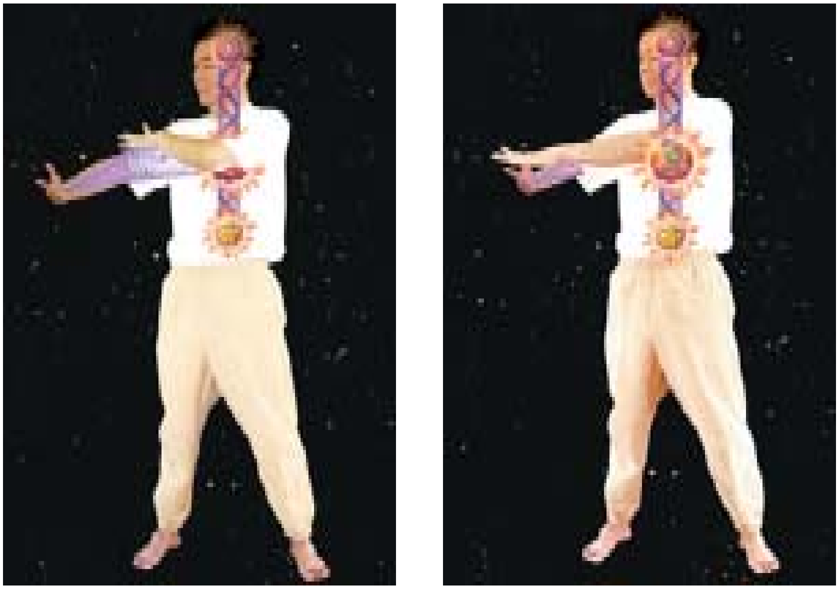

Fig. 8.101 *Balance right side yin and yang and channel earth energy*

Then repeat **Activating the Yin and Yang Channels** with the other fingers and channels in the following order:

Index Finger - Large Intestine
Thumb - Lung
Pinky Finger - Heart and Small Intestine
Middle Finger - Pericardium

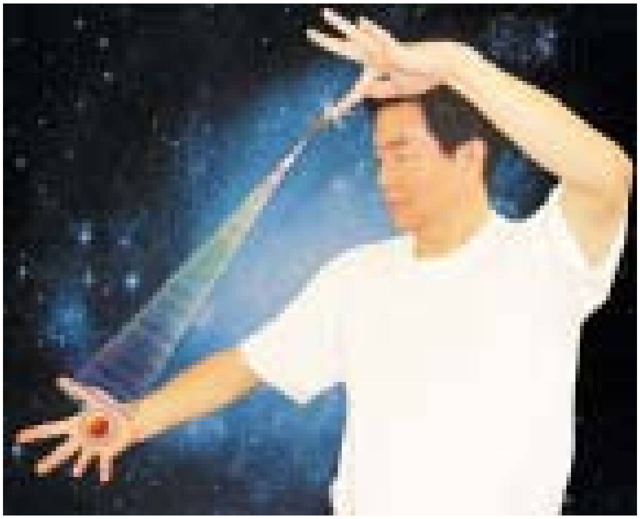

Fig. 8.102 *One Finger Art - activating the Laogong.*

Activate the Chi Belt - Right Foot Forward

Step forward with the right foot. The right palm covers the navel and the left palm covers the Door of Life. Feel the Chi beam penetrate from palm to palm.

Turn your hips to the right and shift weight to your right leg. At the same time move your right palm to the Door of Life in a brushing manner, while your left palm moves to the right hip in a sweeping manner. Note that the palms face the Chi belt.

Repeat the above three times altogether.

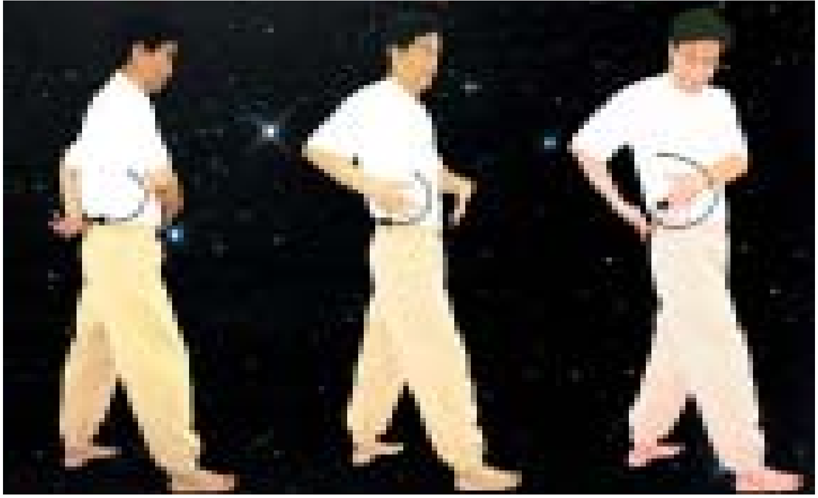

Fig. 8.103 *Activate the Chi Belt - Right Side.*

Activate the Chi Belt - Left Foot Forward

Repeat the above with the left foot forward and reversing the right and left hands.

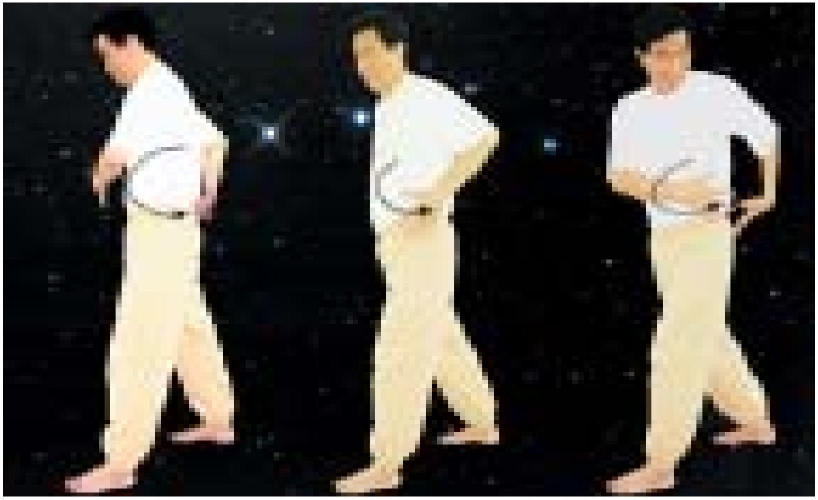

Fig. 8.104 *Activate the Chi Belt - Left Side.*

Channel the Earth Force

Bring the hands forward and rotate the palms to face downwards. **Smile** and channel the earth force as before.

For these Closing Movements see the pictures beginning on page 171.

Closing Movements
Activate the Chi Channels - Balancing the Organs

Index Finger - open (up), close (down), neutral (even with the other fingers).
Ring Finger - down, neutral.
Index Finger - open, neutral.
Thumb - in and down, neutral.
Index Finger - open, neutral.
Pinky Finger - down, neutral.
Index Finger - open, neutral.
Middle Finger - down, neutral.
Index Finger - open, neutral.

Crane's Beak

Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Swallow the saliva. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.

Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms have returned to the starting position.

Finish

Bring the awareness to the navel and notice the quality and intensity of the energy generated.

Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Healing Others

Summary and Combined Practice

After learning the four sets of Cosmic Chi Kung, you may want to combine them all into a short and simple daily practice. This combined set, previously called Buddha Palm I, synthesizes the movements from the first three sets into one basic sequence. The movements are done on the right side only, so the entire set can be completed in **ten to twenty minutes**.

Cosmic Chi Kung: Simple Combined Sequence

Opening Movements

Bone Breathing - Absorbing Cosmic Chi

1. Stand with your hands down by your sides. Feel your Chi moving and activate the **bone breathing**.
2. Activate the cosmic force by using your mental intention (Yi) to spiral at the **mideyebrow** point and draw in the golden light energy. Breathe through the mideyebrow point in the same manner as in bone breathing, allowing the energy to **flow through the Microcosmic Orbit**. Alternatively you can breathe the golden light down into your navel to charge up your Original Chi at the Tan Tien. You can also simply circulate the golden light throughout your body. Relax, smile and enjoy the sensation of your entire body breathing in the cosmic energy.

Channeling the Earth Force - Washing the Bone Marrow

1. Connect with the earth force by activating the **perineum** and the **soles of the feet, Kidney 1, Bubbling Spring**. Connect the bridge by slightly squeezing the sex organs and the anus. Wash the bone marrow from the soles to the crown.
2. **Smile** into the perineum, palms, mideyebrow and crown. Then slowly raise the arms, palms facing each other to chest height keeping the elbows relaxed and sunk. Then slowly rotate the arms until the **palms are facing down**. Absorb the Earth Chi through the soles, perineum and palms; feel it being absorbed into the bones and body like a **steam or mist rising from the earth**. Feel the earth force move through the **center of the bones**, up the calves, thighs (femur bones), through the hip bones, spine, scapulas, arms, neck and skull. Finally swirl the energy around your brain.

Activate the Hip - Large Intestine 4, Pericardium 9 and Small Intestine 3

1. Raise your index finger slightly upward to activate the Laogong point (Pericardium 8). Then stretch your thumbs out to the sides and down, so that they end up pointing downwards toward the earth to activate Large Intestine 4 (Hegu). **Large Intestine 4 is found in the webbing between the thumb and index finger. It is called the "eye of the hand"; this area is also known as the "Tiger's Mouth."**
2. Slowly draw the elbows back and lower the hands until your palms are facing down beside the "eye of the hip" (the iliac crest) with the "eye of the hand" (Large Intestine 4) aligned with the "eye of the hip". **Use your mind intention to beam the energy back and forth between your two LI-4 points. This will activate the large and small intestines.**
3. Rotate your hands so that the fingers first point out, then back and finally point the fingers in toward the eye of the hip. **Beam energy back and forth between the tips of the middle fingers** (Pericardium 9) passing through the hips.
   Rotate the hands again until the fingers are pointing forward with the palms facing up. Align the knife edge (Small Intestine 3) with the eyes of the hip, and **beam the energy back and forth between the two hands.**

Absorb the Heavenly Force - Washing the Bone Marrow - Crown to Soles

Extend the arms forward to chest height with the **palms still facing upward**.

Become aware of the palms and the crown. Activate the Laogong point (Pericardium 8). Feel the heavenly energy, and be aware of the violet light of the North Star and the red light of the Big Dipper shining down. Draw this light in through the palms and the crown point and feel it washing your bone marrow. Direct the sensations down through the skull, cervical vertebrae, clavicle, scapulas and sternum. Activate your thymus gland and continue drawing the healing heavenly light down through the rib cage, arms, spine, hips, femurs, calf bones, feet and toe bones. **Feel your bones activated with a positive electrical force (the positive force comes from heaven, negative from the earth).**

Absorb the Earth Force

1. Rotate the arms until the **palms face down**.
   Raise the index fingers slightly, extend the thumbs first towards each other, and then towards the ground.
2. Become aware of the palms (Laogong), the soles of the feet (Kidney 1, Bubbling Spring) and the perineum (Conception Vessel 1, Hui Yin). Activate these points, feel the earth energy, and **be aware of a gentle blue light or white light rising up from the earth like a pure steamy mist.** Draw this light in through the palms and soles and let it wash up through the center of your bones as it steam cleans your

marrow. Picture any impurities or illness dripping out of your bones and draining down into the earth, where they are recycled and purified.

Core Movements

Grasping the Moon - Connecting the Heaven and Earth Forces

1. **Open Position:** Heaven Force - Bring the arms up to shoulder height, bending the elbows so that the **left hand is under the right elbow** with the left index finger pointing up to Heart 3.
2. The right forearm is over the left forearm with the tip of the right index finger held above the Large Intestine 11 point of the left arm. The right index finger is pointing up toward heaven.
3. The right index finger draws in the heavenly force like an antenna. The force flows up the bones of the right arm across the shoulders to the left arm, and through the bones of the left arm to the left index fingertip. It then flows through the index finger of the left hand into the Heart 3 point of the right arm **completing the heavenly circuit**. Continue cycling in this way.
4. **Close Position:** Earth Force - Keeping the arms in the same position, move both index fingers to point down. The right index finger will now point toward the Large Intestine 11, point of the left arm and the left index finger will point downward toward the earth. Draw in the earth force through the left index finger bringing it through the bones of the left arm, across the shoulders through the right arm to the right index finger and out into the Large Intestine 11 point of the left arm. Continue cycling in this way **completing the earth circuit.**

Open and close three times altogether ending in the open position.

Activate the Throat Center

1. Draw the elbows out to the side and rotate the palms to face in so that the left palm is about 3 - 6 inches in front of the Throat Center (Conception Vessel 22), palm facing in, and the right hand is about 1 - 3 inches in front of the left hand, palm facing in. Align the Lung 10 point of the right hand with the Lung 10 of the left hand, and align the Lung 10 points in the middle of the thenar eminence of the left hand with the Throat Center. **Keep the points aligned and beam the energy passing through both hands into the Throat Center.**
2. Move both hands about 12-18 inches out from the Throat Center maintaining the energy beam as you move.
3. Move in and out three times altogether in this way ending with the hands in as in preivous point above.

Activate the Third Eye

1. Keeping the Lung 10 points aligned, raise the hands up to the level of the mideyebrow point. Keep the points aligned and **beam the energy passing through both hands into the mideyebrow point.**
2. Move both hands about 12-18 inches out from the mideyebrow point maintaining the energy beam as you move. **Move in and out three times** altogether in this way, ending with the hands as previously.
3. Keeping the palms and thenars (Lung 10) aligned, move the hands back down to the position in front of the **Throat Center.**

Open the Bridge and Regulator Channels

1. Keeping the forearms horizontal at shoulder height, separate the hands so that the Laogong points are aligned in front of Stomach 13 just below the clavicle on a vertical line above the nipple or **mammillary** line. The fingertips should almost be touching. Beam energy from the Laogong points into ST-13. You will be following this vertical line, which passes through the nipples, down the torso to the level of the navel stopping to focus and **beam energy into key points along the way**. On each of these, focus your Yi on beaming energy into the organs inside each position. Stomach 13 is the location that activates the Heart and Lungs. Feel the change in your breathing as you activate the **lungs**.
2. Move the hands down the mammillary line to Stomach 16, about one inch above the nipple (in the space between the third and fourth ribs). Again, beam energy to energize and balance the **heart and lungs**.
3. Move the hands down to Liver 14 (about 2 - 3 inches below the nipple in the space between the 6th and 7th ribs). Beam healing energy into the **liver and gall bladder**.
4. Move the hands down to Spleen 16 (just below the rib cage on the mammillary line). Beam healing energy into the **stomach, pancreas and spleen**.
5. Move the hands down to Spleen 15 (on the mammillary line to the level of the navel). Beam healing energy from your palms into the **small intestine and the Tan Tien**.

Grasping the Chi Ball

1. Slowly rotate the hands so that your palms face each other in front of your navel. Feel the invisible ball of energy between your hands. Allow the energy to push your hands apart keeping the feeling of connection between your palms.
2. Let the energy draw your hands back toward each other, like a magnet, until you feel as if you are squeezing a ball.
3. Stretch and squeeze the Chi ball three times by opening and closing your palms in this way.

Activate the Outer and Inner Arm Gates
Waiguan - Triple Warmer 5 and Neiguan - Pericardium 6 - Right Arm on Top

1. Slowly rotate the arms so that both palms face down.
2. Cross the right wrist over the left wrist with 1-3 inches of space between the hands. Align PC-6 of the right wrist with TW-5 of the left wrist.
3. Keeping the wrists crossed, slowly rotate hands turning the palms upward. Now TW-5 of the right wrist will be aligned with PC-6 of the left wrist.

Open the Functional Channel with LI-4

1. Bring the left hand down to the level of the Tan Tien with the Chi Knife facing in. Lower the right hand, palm down, and align the eye of the hand, the LI-4 point with your navel. Feel the Chi ball between your two hands, and at the same time beam energy from LI-4 into the navel point.
2. Raise the right hand up the Functional Channel stopping and aligning LI-4 with the **solar plexus, heart, throat** (CV-22) and **mideyebrow** points. **Beam energy from the eye of the hand into each of these points**, and at the same time continue to feel the Chi ball connection between the two hands.
3. Return down the Functional Channel in the same manner stopping to beam energy into each point. **Go up and down the Functional Channel** in this way three times.

Yin/Yang Palms

1. Separate the hands and extend them out in front of the body at solar plexus level. The left palm is still pointing up and is yang drawing in the heavenly force through Laogong. The right palm is still facing down and is yin drawing in the earth force energy.
2. Conclude by turning the left palm over to face downward. Draw in the earth force through both palms.

Double Palm and Beam to Activate the Functional and Governor Channels

1. Bring the left palm in first to face the navel, about 3 - 5 inches away, aligning the Laogong point with the navel. Follow by bringing the right hand in front of the left hand, aligning the Laogong of the right palm with the Laogong of the left hand and the navel. Both palms are now facing in. **Beam energy through both Laogong points to the navel and through the body to the Door of Life**.
2. Next raise the left palm up to the **solar plexus** aligning the Laogong point with the solar plexus. Beam energy into the solar plexus point and through the body to the T-11 point. Your right hand remains at the navel level beaming energy to the navel and the Door of Life.
3. Raise your right hand up to the outside of the left hand at the solar plexus level, and beam energy through both hands into the **solar plexus point and the T-11 point**. Continue in the same way with the left hand leading to the **heart point and wing**

point, the **throat point and C-7**, the **mideyebrow point and jade pillow** and the **crown point and perineum**.

4. Return down the front, point by point in the same way, leading with the left hand. Go up and down in this way three times. **The double palm and beam activate both the Functional and Governor Channels. The energy beam passes all the way through the body at each point**.

Grasping the Chi Ball

1. Separate the hands and slowly extend the arms in front of you at navel level, palms facing each other holding the Chi ball.
2. Stretch and squeeze the Chi ball three times as before by opening and closing the palms.

Channel the Earth Force

Rotate the hands to face downward, and draw in the earth force through the palms, soles and perineum.

Closing Movements

Activate the Chi Channels - Balancing the Organs

Index Finger - open (up), close (down), neutral (even with the other fingers).
Ring Finger - down, neutral.
Index Finger - open, neutral.
Thumb - in and down, neutral.
Index Finger - open, neutral.
Pinky Finger - down, neutral.
Index Finger - open, neutral.
Middle Finger - down, neutral.
Index Finger - open, neutral.

Crane's Beak

1. Form the crane's beak with both hands by bringing the fingertips together with the thumbs inside. Raise the forearms to shoulder height with the fingers pointed down, inhaling gently as you raise.
2. Open the palms slowly and begin to lower the arms to the sides as you exhale, until the palms are facing each other in the starting position.

Finish

1. Bring the awareness to the navel and notice the quality and intensity of the energy generated.
2. Collect energy at the navel. Men, place the hands over the navel with the right hand on top; vice versa for women.

Chapter IX
Universal Tao Practices

Cosmic Inner Smile

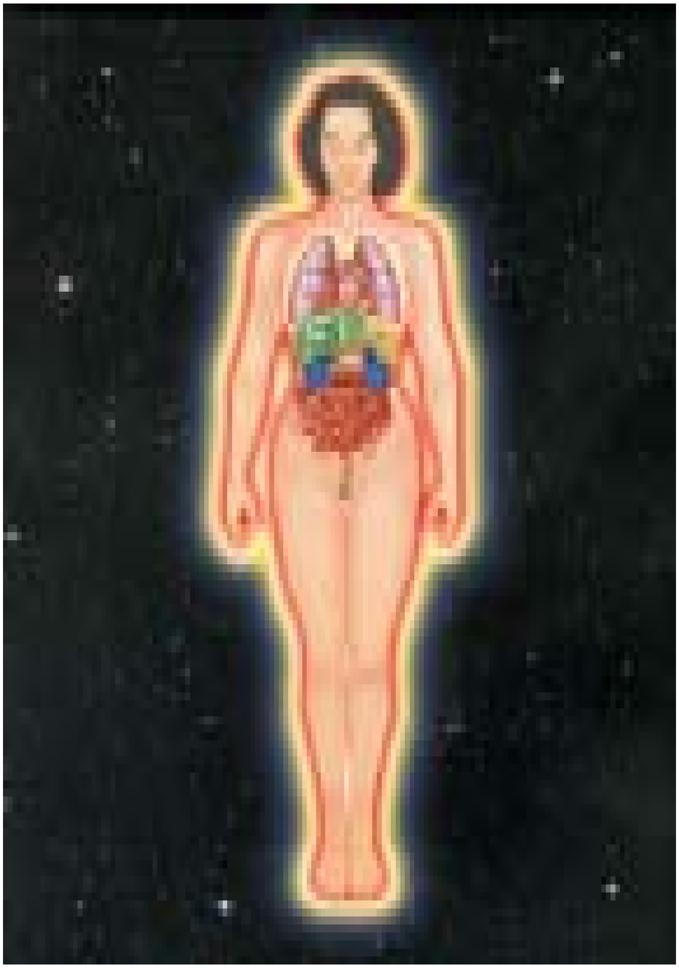

Fig. 9.1 *Inner Smile*

The Inner Smile begins at the eyes and the mideyebrow point and moves down to the heart. As you activate the heart, the loving energy will flow out and you will feel the energy of your Inner Smile flow down the entire length of your body like a waterfall. This is a very powerful and effective tool to counteract stress, tension and negative Chi.

1. Be aware of the **mideyebrow,** imagining that your are in one of your favorite places in the world, a place where you feel safe, relaxed and happy. Recall the sights you might have seen there, the sounds you might have heard, the scents, sensations and flavors that you associate with that place.
2. Imagine that one of your favorite people is standing in front of you, smiling to you with loving, happy, radiant shining eyes. **Smile** to your face and slightly lift up the corners of your mouth.
3. Feel yourself responding to that special person's smile with a smile of your own. Feel your eyes smiling and relaxing.
4. Smile down to the **thymus gland** and picture a white flower blossoming. Gently inhale into the thymus gland, connecting your breath to the olfactory organ. Smell the good fragrance.
5. Aim your inner attention at your **heart**, picturing the heart before your inner eye and smile to it. **Smile** until you feel the heart smile back to you. Picture your heart like a red rose, gradually opening. This will activate the **love** and fire of **compassion** in the heart. Once you feel the red light and loving awareness it will activate the **cosmic red healing light** or mist from above and around you.
6. Smile at the light or mist and very slowly with a soft, long, deep breath draw the red mist, love and compassion into the mideyebrow, down through the mouth and throat, into the heart and gradually overflowing to the **small intestine**. Exhale, but retain the red light and the love and compassion in the **heart and small intestine**. At the same time exhale the cloudy, black or negative energy. Keep on doing this breathing 18 to 36 times waiting until the heart becomes bright red before your inner vision and starts to radiate out to the tongue, mouth, nose, ears and eyes. Allow the red light to whirl around you and form a **red aura**. Feel your skin glowing with red energy.
7. Let the heart's loving energy radiate out to the **lungs**. Aim your attention at the lungs; picture them before your inner eye and smile to them. **Smile** until you feel the lungs smile back to you. Picture your lungs like a white rose, gradually opening; smell the good fragrance. This will activate the courage in the lungs. Once you invoke the white light and courage into the lungs, you will also activate the **cosmic white healing light** or mist from above and around you.
8. Smile at the light or mist and very slowly, softly take a long deep breath and draw the white mist into the mideyebrow, the mouth and down to the lungs, gradually overflowing into the **large intestine**. Exhale but retain the white light and the feeling of courage in the lungs. At the same time exhale the cloudy, black or negative energy. Keep on doing this kind of breathing 18 to 36 times, waiting until the lungs become bright white and start to radiate out to the nose, ears, eyes, tongue and mouth. Invite the white light to whirl around you and form a **white aura** covering your skin like autumn dew.
9. The **spleen, pancreas and stomach** correspond to the yellow color of the earth element. Connect to these organs and smile down into them from the mideyebrow.

First connect to the heart, then attract the clear yellow light from above and around you. You might see the **golden yellow aura** of a wheat field ready for harvest. Expel feelings of worry; exhale the cloudy, sticky energy. Breathe in the golden yellow aura, filling your spleen center with golden yellow light. Repeat to complete 18 to 36 cycles, then allow the light to radiate out to your mouth, nose, ears, eyes and tongue. Wrap the golden aura around you, leaving a golden shine on your skin.

10. The **kidneys and bladder** correspond to the blue color of the water element. Gentleness is contained within the blue healing light. Inhale the blue aura into the kidneys. Expel feelings of fear or stress on breathing out. Allow your energy field to expand, breathing in and out the blue energy 18-36 times, until it starts to radiate out from your kidneys to your ears, eyes, tongue, mouth and nose. Gather the blue mist on your skin enveloping you with a **blue aura**.
11. The **liver and gall bladder** connect to the green color of the element wood; kindness replaces anger, which is expelled on breathing out. Do the same as above and create a **green aura** around you. Exhale the dark red, cloudy heat. Inhale the nourishing green of the forests. Do 18 to 36 cycles, until the green light has completely filled the liver and starts to radiate out to your eyes, tongue, mouth, nose and ears. Invite the green light to form a green aura around you.
12. **Smile** down to the **sexual organs** and reproductive system. Feel love and sexual arousal and the heart and sexual organs uniting. Observe how this process transforms the sexual energy into Chi. Channel down orange and red light to the sexual organs. Thank them for their work in keeping you alive and healthy. Rest. Do nothing. Gather and store the energy by smiling and spiraling in the Tan Tien.

Six Healing Sounds

The Six Healing Sounds enhance Therapeutic Work

Everyone has heard stories about gifted beings who possess great healing powers. People seek out gifted healers. Yet even in the best of circumstances, how much time can a great healer spend with you? One hour a day? An hour a week? What about the rest of the time? Thus you must learn how to take care of yourself; you must learn how to clear out your negative energy and transform it into positive healthy energy. If you maintain yourself in this way it will surely enhance any other therapy you are receiving.

The Six Healing Sounds are a simple yet powerful tool to promote physical, energetic and emotional healing and balance. If you are healing others, you can teach your students one or two of the Six Healing sounds each session, to enhance the effects of your therapeutic work. During the practice of the sound the resting period is very important. Growing the good virtue of the organs is essential so the negative or sick energy has less room to grow.

The sounds are also used to generate certain frequencies for specific healing. Each sound can generate different energy for the healing of the different organs.

Once the movements, sounds and information have been integrated into your practice you can simplify and make the Six Healing Sounds more powerful. For example, when you are settled into the lung sound and you are breathing white healing light and feel the mideyebrow wide open, you put the awareness in the 'Tan Tien and Universe'.

Chi will flow from the universe into the Tan Tien and just the right amount and quality will flow from the Tan Tien into the lungs and large intestine. Just set up the location (the organ) and connect to the universe and the Chi will flow to the correct location!

When doing these, the eyes are open only while making the sound.

Six Healing Sounds - Practice

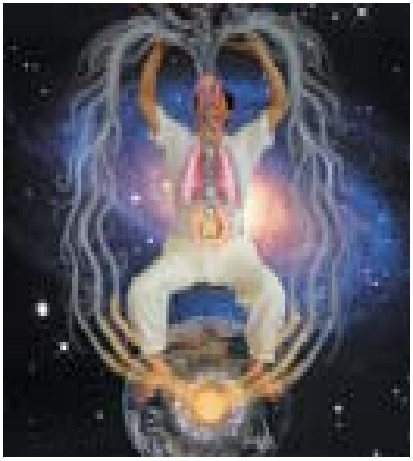

Lung Sound

Element: Metal
Associated Organ: Large Intestine
Sound: SSSSSSS (Tongue behind Teeth)
Emotions: Negative - Grief, Sadness, Depression
Positive - Courage, Righteousness, High Self Esteem
Color: White, Clear, Metallic
Season: Fall
Direction: West

Position: Sit in a chair with your back straight and your hands resting palms up on your thighs. Have your feet flat on the floor about a hip's width apart . **Smile** down to your lungs and be aware of any sadness, grief or excess heat in your lungs. Slowly inhale and raise your hands up your center line with your fingers pointing toward each other. When your hands pass your shoulder level, begin to rotate the palms out as you continue raising your hands up in front of you and above your head with the palms up. Point your fingers towards the fingers of the opposite hand and keep your elbows slightly bent.

Sound: Part your lips slightly, keeping your jaw gently closed. Look up through the space between your two hands and push your palms slightly upward as you slowly exhale and make the sound, "SSSSSSS". Picture and feel any excess heat, sadness, grief, depression, sickness and dingy white color expelled and released as you exhale slowly and fully.

Resting Posture: When you have completely exhaled, rotate the palms to face downward with the fingers still pointing toward each other. Slowly lower the palms and bring them just in front of the chest feeling the lung's aura.

Close your eyes and be aware of your lungs. **Smile** into your lungs, and as you inhale, imagine that you are breathing in a **bright white mist of light**. Breathe this light into your lungs and feel it cooling, cleansing, invigorating, healing and refreshing your lungs. Feel it flowing down to the large intestine to balance the energy of the yin lungs and yang large intestine, allowing the **courage** quality of your lungs to emerge. Grow more courage, so sadness and depression have less room to grow. With each inbreath, feel yourself drawing in cool fresh energy. With each outbreath, mentally make the Lung Sound and release any remaining sadness or hot energy.

Repeat at least three times. For the first two repetitions, you can make the sound aloud. On the third or last repetition, make the sound sub-vocally (vocalizing the sound so softly that only you can hear it) Repeat six, nine, twelve or twenty four times to alleviate extreme sadness, depression, colds, flu, toothaches, asthma and emphysema.

Kidney Sound

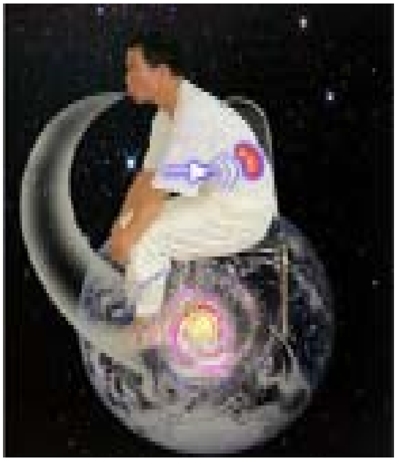

Element: Water
Associated Organ: Bladder
Sound: CHOOOOOO (with your lips forming an "O" as if blowing out a candle.)
Emotions: Negative - Fear, Shock
Positive - Gentleness, Wisdom
Color: Dark Blue or Black
Season: Winter
Direction: North

Position: Now move the hands to cover the kidneys. **Smile** to your kidneys, and be aware of any excess cold or heat in the kidney region. Then bring your legs together, ankles and knees touching. Lean forward and clasp the fingers of both hands together around your knees. Inhale and pull your arms straight from the lower back while bending the torso forward (this allows your back to protrude in the area of the kidneys). Tilt your head upward as you look straight ahead, still pulling on your arms from the lower back. Feel your spine pulling against your knees.

Sound: Round the lips slightly and slowly exhale while making the sound "CHOOOOOO." Simultaneously contract your abdomen, pulling it in toward your kidneys. Imagine any fear, sickness, imbalances, excess cold or heat energy being released and squeezed out of the fascia surrounding the kidneys.

Resting Posture: After you have fully exhaled, slowly sit erect and return the hands to touch the aura of the kidneys. Close your eyes and again be aware of your kidneys. **Smile** to your kidneys and, on the inbreath, imagine you are breathing a brilliant **luminous blue light mist** into them; feel this mist healing, balancing and refreshing your kidneys and bladder and picture them glowing a bright blue color. On the outbreath, imagine you are still making the Kidney Sound.

Repeat at least 3 times, as with the previous sound. Repeat six, nine, twelve, or twenty four times to alleviate extreme fear, fatigue, low-pitched ringing in the ears, dizziness, back pain, bladder and urinary infections or other problems of the reproductive system.

Liver Sound

Element: Wood
Associated Organ: Gall Bladder
Sound: SHHHHHHH
Emotions: Negative - Anger, Frustration, Resentment
Positive - Loving Kindness, Benevolence, Forgiveness
Color: Green
Season: Spring
Direction: East

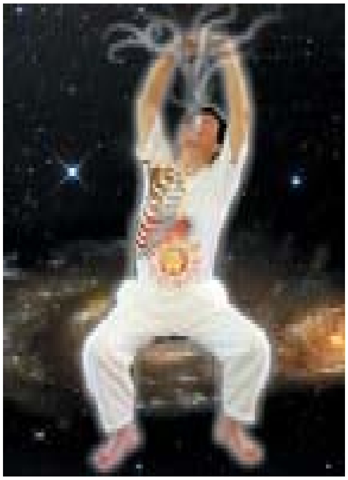

Position: Place the hands on the liver. **Smile** to your liver and be aware of any anger, frustration, resentment or excess heat. Slowly begin to inhale a deep breath as you extend your arms up from the sides with your palms up. Continue to raise them over your head. Interlace your fingers together and turn your joined hands over to face the sky, palms up. Push out through the heels of the palms and extend the arms up, keeping the shoulders relaxed. Bend a little to the left and stretch your right arm slightly to gently open the area of your liver.

Sound: Open your eyes wide (the eyes are the sensory opening of the liver). Slowly exhale, making the sound, "SHHHHHHH" sub-vocally. Feel that you are releasing any trapped excess heat, anger, illness and negativity from your liver and that these are riding out of your body on your breath.

Resting Posture: Once you have fully exhaled, separate your hands, turn the palms down and slowly lower your arms to the sides, leading with the heels of the hands. **Smile** and inhale a **shiny spring green mist**, illuminating the liver and gall bladder. Bring your hands back to rest on the liver's aura. Close your eyes and smile into your liver. With each inbreath, breathe fresh Chi into your liver and gall bladder and mentally make the Liver Sound with each outbreath.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate extreme anger, to relieve red or watery eyes, to remove a sour or bitter taste in the mouth or to detoxify the liver.

Heart Sound

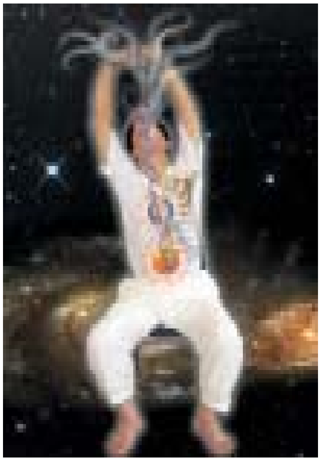

Element: Fire
Associated Organ: Small Intestine
Sound: HAAAAAAAW
Emotions: Negative - Arrogance, Harshness, Cruelty, Hatred
Positive - Joy, Honor, Respect, Love, Happiness
Color: Red
Season: Summer
Direction: South

Position: Let both hands rest on the heart. **Smile** to your heart and be aware of any arrogance, haughtiness, hatred, giddiness, cruelty or hastiness. Slowly begin to inhale a deep breath as you extend your arms up from the sides with your palms up, as you did with the Liver Sound. Continue to raise them over your head. Interlace your fingers together and turn your clasped hands over to face the sky, palms up. Push out through the heels of the palms and extend the arms up, keeping the shoulders relaxed. Bend a little to the right and stretch your left arm slightly to gently open the area of your heart.

Sound: Keep your eyes soft and relaxed, looking up through your hands. Slowly exhale, making the sound, "HAAAAAAAW" sub-vocally. Feel that you are releasing any trapped heat, negative emotions, illness and imbalances from your heart and that these are riding out of the body on your breath.

Resting Posture: Once you have fully exhaled, close your eyes, separate your hands, turn the palms down and slowly lower your arms to the sides, leading with the heels of the hands. As you move inhale a **bright red mist** into the heart and small intestines. Bring your hands back to rest on your heart's aura . Smile into your heart. With each inbreath, breathe fresh Chi into your heart with each outbreath mentally repeat the heart sound.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times for the heart and small intestines; feel them balanced. This will alleviate extreme impatience, hastiness, arrogance, nervousness, moodiness, jumpiness, irritability, tongue ulcers, palpitations, sore throat, heart disease, insomnia and detoxify the heart.

Spleen Sound

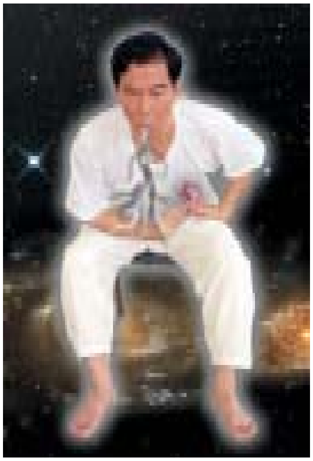

Element: Earth
Associated Organ: Pancreas, Stomach
Sound: WHOOOOOOO (gutturally from the throat)
Emotions: Negative - Worry, Excess Sympathy, Overthinking
Positive - Fairness, Balance, Equanimity, Justice, Openness
Color: Yellow
Season: Indian Summer
Direction: Center (where you stand, looking out to the Six Directions)

Position: Place your hands on the body covering the spleen, pancreas, and stomach area. Be aware of your spleen and **smile** sincerely into it. Inhale deeply as you move the arms outward in an embrace and aim the fingers up under the left rib cage. Place your fingers of both hands just beneath the sternum and rib cage on the left side.

Sound: Look out, lean into your fingers and gently push your fingertips in. Exhale slowly and make the sound, "WHOOOOOOO," from the depths of your throat. Feel yourself releasing any trapped heat, worry, mental fixations or excess sympathy.

Resting Posture: Once you have fully exhaled, close your eyes, slowly release the hands and extend the arms out embracing the earth; return your hands to the resting position on the spleen's aura. Smile to your spleen, pancreas and stomach. Inhale fresh Chi to your spleen, pancreas and stomach as a brilliant **luminous yellow healing mist** that cleanses and refreshes your organs. Mentally do the Spleen Sound with each outbreath.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate extreme indigestion, heat or cold in the stomach or spleen, worry, nausea, hemorrhoids, fatigue, organ prolapse or loose stools.

Triple Warmer Sound

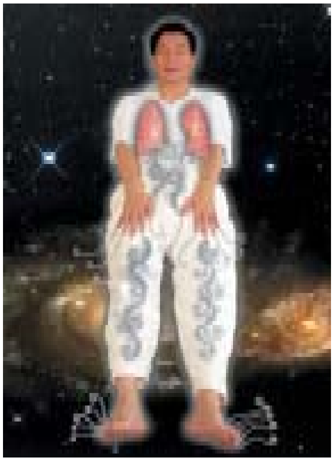

The Triple Warmer refers to the upper, middle and lower body, and to the distinct metabolic transformations that occur within each area. The Upper Warmer is the area above the diaphragm, where the heart and lungs are located. This area tends to become hot and is responsible for respiration and cardiovascular circulation. The Middle Warmer, the area between the diaphragm and the navel, becomes warm and is where the digestive organs are located. The Lower Warmer, the area below the navel, is responsible for reproduction and elimination and is cool in temperature. The sound "HEEEEEEE" balances the temperatures of the three levels by bringing hot energy down to the lower center and cold energy up to the higher centers.

Position: Lie on your back with your arms resting palms up at your sides. Keep your eyes closed. **Smile**. First breathe into the upper part of your lungs to expand the Upper Warmer; then breathe into the middle of the lungs to expand the Middle Warmer; finally, with the last part of your inhalation, expand your lower lungs and fill the Lower Warmer. Breathing in this way creates more space inside for each organ, helping to release and circulate any internal heat or cold.

Sound: Exhale with the sound "HEEEEEEE" sub-vocally, first flattening your chest, then your solar plexus, and finally your lower abdomen. Feel the dark and cloudy color, cold and chill energy exit from the tips of the fingers.

Resting Posture: Once you have fully exhaled, do not to focus on any emotions or purification process at all. Just let go and relax your body and mind completely.

Repeat at least three times, as with the previous sound. Repeat six, nine, twelve or twenty four times to alleviate insomnia and stress.

When you have completed the Six Healing Sounds just rest, **smile** and do nothing.

Regular daily practice of the Six Healing Sounds will help you to keep in touch with the energetic and emotional state of your internal organs. Most importantly do the sounds in the evening before going to sleep. This will help clear out the negative emotions before sleeping, so you can feel that the night's rest recharges your energy positively. This practice will greatly benefit your personal Cosmic Healing Chi Kung practice. It will help sensitize you to the varieties and differing qualities of Chi. This knowledge will also help you in diagnosing and treating others.

For more details of this practice please see the book "Taoist Ways to Transform Stress Into Vitality", by Mantak Chia.

Appendix

Guide to the Acupuncture Points used in Cosmic Chi Kung

The following is a detailed description of the acupuncture energy points used in the Cosmic Chi Kung practice. It is necessary to know the exact location of the points to do the practice well. However, you do not have to know the traditional Chinese medical functions and energetics of the points to do the practice. These are included in this appendix as an item of interest.

The names of the points are given first by the channel and number, second by the English translation of the traditional Chinese name for the point and finally by the name in Chinese.

We are emphasizing the traditional Chinese medical therapeutic properties of the points given herein. Many of these points also have very different purposes in Taoist Yoga. *For a description of the Taoist Yoga energetics, refer to "Awaken Healing Light of the Tao" by Mantak Chia.*

Pericardium 8

"Lakor's Palace" - Laogong

Ying Spring, Fire Point

Location: On the center of the palm, where the tip of the middle finger touches when making a loose fist.

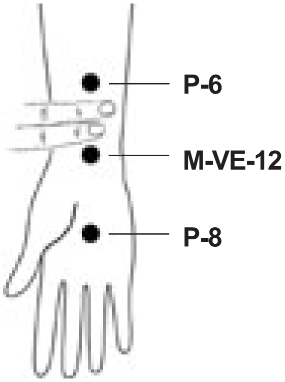

Functions: Cools the heart, drains heat from the heart, cools the blood, calms the spirit, regulates the heart Chi and yang, Ghost Point for treating spirit possession.

Indications: Coma from stroke, heat exhaustion, angina pectoris, mania, hysteria, mental illness, excessive sweating of the palms, mouth problems.

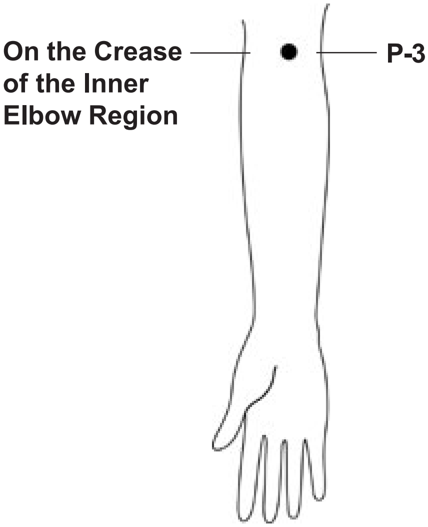

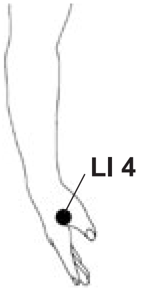

Large Intestine 4

"Adjoining Valleys" - Hegu

Yuan Source Point
Location: On the back side of the hand between the thumb and index finger, in the web approximately at the middle of the metacarpal bone of the index finger.
Functions: Disperses wind, relieves exterior conditions, suppresses pain and clears the channels, clears lung heat, calms the spirit.
Very powerful point for moving the Chi and blood of the whole body; main point for pain; main point for headache; main point for ailments of the head, face and sensory organs; main point for immunity.
Indications: Headaches, common cold, redness with pain and swelling of the eyes, toothache, facial swelling, sore throat, finger cramps, arm pain, fever, abdominal pain, constipation.
Caution: Contra-indicated in pregnancy.

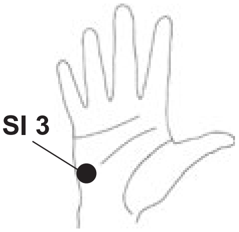

Small Intestine 3

"Back Creek" - Houxi

Shu Stream, Wood Point
Master Point of the Governor Channel
Coupled Point of the Yang Bridge Channel
Location: When a loose fist is made, the point is near the head of the fifth metacarpal bone on the knife edge of the hand at the junction of the red and white skin.
Functions: Relaxes the tendon-muscular channels, opens the Governor Channel, clears the Spirit, drains evil Chi from the heart.
Indications: Seizures, psychosis, hysteria, intercostal neuralgia, headache, stiff neck, red, painful and congested eyes, deafness, spasms of the arm, elbow and fingers; fever, night sweating, whiplash, occipital headaches.

Used for structural/musculo-skeletal problems; opens the Governor Channel for problems with back, neck and head, main point for stiff neck, antispasmodic for muscle spasms; anti-inflammatory for spinal arthritis, intestinal inflammation.

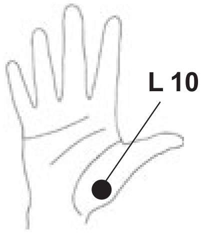

Lung 10

"Fish Border" - Yuji

Ying Spring, Fire Point
Location: In the middle of the ulna eminence below the thumbs at the junction of the red and white skin.
Functions: Cools heat in the lungs, benefits the throat.
Indications: Sore throat, cough, laryngitis, tonsillitis, fever with common cold.

Triple Warmer 5

"Outer Gate" - Wai Guan

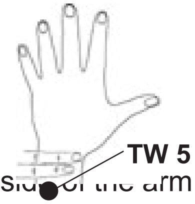

Luo Connecting Point
Master Point of the Yang Regulator Channel
Coupled Point of the Belt Channel
Location: About two fingers width above the wrist crease on the outside of the arm.
Functions: Regulates the Yang Regulator Channel and Belt Channel, tonifies and consolidates Protective (Wei) Chi, releases exterior hot and cold conditions, helps circulate stagnant Chi in the channels.
Indications: Common cold with alternating chills and fever, high fevers, pneumonia, deafness, migraine headaches, paralysis, stiff neck.

Pericardium 6

"Inner Gate" - Nei Guan

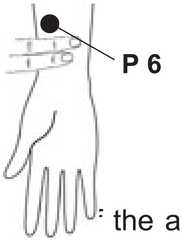

Luo Connecting Point
Master Point of the Yin Regulator Channel
Location: About two fingers width above the wrist crease on the i the arm, between the two prominent tendons.
Functions: Calms the heart and spirit, regulates the Chi, opens and relaxes the chest, regulates and harmonizes the stomach, regulates the liver, relieves pain.
Indications: Rheumatic heart disease, shock, angina pectoris, palpitations, chest pain, asthma, shortness of breath, tightness or fullness in the chest, spasm of the diaphragm, vomiting, stomach ache, abdominal pain, morning sickness, motion sickness, migraine headaches, hysteria, anxiety, irritability, insomnia, seizures, swollen and painful throat, painful menses or nausea with period.

Large Intestine 11

"Crooked Pond" - Qu Chi

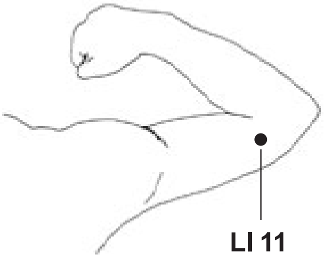

He Sea, Earth Point
Ghost Point
Location: When the elbow is flexed, in the depression at the outer and upper end of the elbow crease.
Functions: Cools heat, clears fire, drains dampness, eliminates wind and exterior conditions, regulates and moistens large intestine.
Indications: Arthritic pain in the arms, paralysis, hypertension, high fever, anemia, allergies, skin problems, Parkinson's Disease. Main point for skin diseases; main point for high fever.

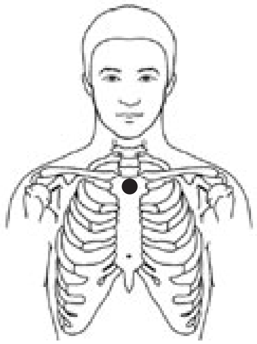

Conception Vessel 22

"Heaven's Chimney" - Tian Tu

Intersecting Point of the Yin Regulator Channel on the Conception Vessel, Window of the Sky Point
Location: At the top of the breastbone (sternum) in the V-shaped indentation (suprasternal notch). Press in and downward at a forty-five degree angle.
Functions: Cools the throat and clears the voice, facilitates and regulates movement of lung Chi, frees the breath and helps restore the proper functioning of the lungs and bronchi, helps open Conception Vessel (Functional Channel).
Indications: Asthma, bronchial asthma, bronchitis, coughing, pharyngitis, goiter, hiccups, nervous vomiting, hoarse voice, spasms of the esophagus, diseases of the vocal cords, sore throat.

Conception Vessel 17

"Central Altar" - Tan Zhong

Front Mu-Alarm Point of the pericardium, Influential Point of Chi of the body
Sea of Chi Point, Master Point of the Middle Tan Tien
Location: On the frontal midline at the level of the fourth intercostal space, between the nipples.
Functions: Regulates the lungs and the Upper Warmer, tonifies Ancestral Chi, opens and relaxes the chest, diffuses lung Chi, regulates and tonifies Chi, transforms phlegm.
Indications: Pulmonary tuberculosis with shortness of breath, bronchitis, asthma, bronchial asthma, chest pain; all breast problems such as mastitis, insufficient lactation, breast abscesses, wheezing, chest pain, labored breathing, palpitations, difficulty swallowing food, anxiety.

Conception Vessel 12

"Middle Cavity" - Zhongwan (Solar Plexus)

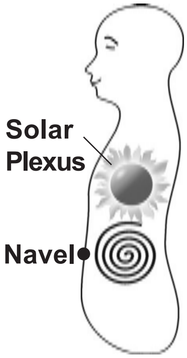

Front Mu-Alarm Point of the stomach, Influential point of all yang organs
Location: Approximately midway between the navel and the top joint of the xiphoid process.
Functions: Regulates the stomach Chi and yin, regulates, strengthens and tonifies the spleen Chi and yang, regulates the Middle Warmer, reduces digestive stagnation, tonifies nutritive Chi, regulates Chi and blood, redirects rebellious Chi downwards, clears stomach fire and heat, calms the fetus, controls the aura.
Indications: Acute or chronic gastritis, stomach and duodenal ulcers, prolapsed stomach, acute intestinal obstruction, stomach ache, vomiting, abdominal distention, diarrhea, constipation, acid regurgitation, indigestion, hypertension, mental diseases.

Mideyebrow Point

"Seal Hall" - Yin Tang

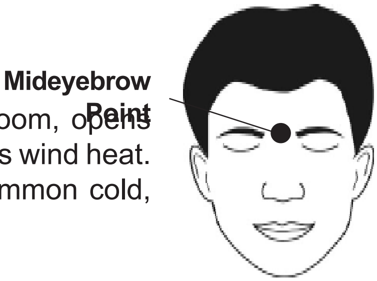

Master Point of the Upper Tan Tien
Location: At the midpoint between the two eyebrows.
Functions: Calms the spirit, activates the Crystal Room, opens the Governor Channel, draws in Cosmic Chi, eliminates wind heat.
Indications: Headache, vertigo, rhinitis, sinusitis, common cold, hypertension, infantile convulsions, sore eyes.

Conception Vessel 6

"Chi Ocean" - Chi Hai

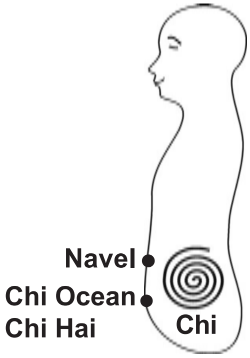

Master Point of the Lower Tan Tien
Location: Approximately three fingers width below the navel.
Functions: Regulates Chi, tonifies Original Chi, strengthens weak kidneys, harmonizes the blood, regulates the Thrusting Route Conception Vessel, reinforces Ching Chi, enriches yin.
Indications: Neurasthenia, abdominal distention, abdominal pain, irregular menstruation, impotence, spermatorrhea, urinary retention, frequent urination, intestinal paralysis, incontinence, constipation, infertility, uterine bleeding, hernia.

Stomach 13

"Chi's Household" - Chi Hu

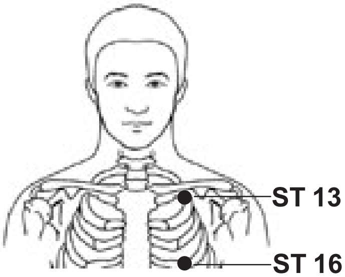

Location: Below the midpoint of the collarbone along the mammilary line (directly above the nipple).
Functions: Clears heat, loosens the chest, relaxes the diaphragm, regulates lung Chi.
Indications: Asthma, bronchitis, chest and back pain, hiccups.

Stomach 16

"Breast's Window" - Ying Chuang

Location: In the space between the third and fourth ribs, one rib directly above the nipple in men, slightly higher in women.
Functions: Stops pain, reduces swelling, clears heat, resolves depression, opens lungs, stops cough, relaxes chest, moves Chi.
Indications: Coughing, asthma, swelling of the breasts, chest and stomach pain, heartburn, shortness of breath, melancholy, diarrhea

Liver 14

"Gate of Hope" - Chi Men

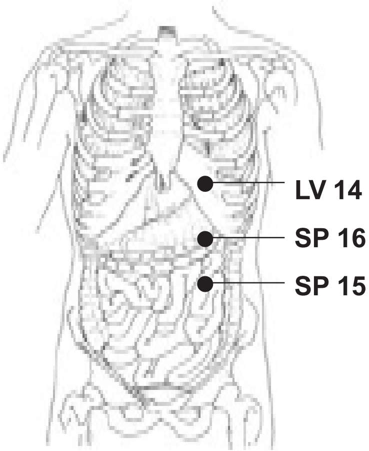

Front Mu-Alarm Point of the liver
Intersecting point of the Yin
Regulator and spleen channels on the liver channel
Location: In the sixth intercostal space, directly below the nipple.
Functions: Spreads liver Chi, transforms and removes congealed blood, strengthens spleen and stomach, expands and relaxes the chest.
Indications: Difficulty in breathing, chest pain, hepatitis, enlarged liver, gall stones, pleurisy, nervous stomach, menopausal disorders, cholera, failure to discharge the placenta after childbirth.

Spleen 16

"Abdomen's Sorrow" - Fu Ai

Location: Just below the bottom edge of the rib cage on the mamillary line.
Functions: Frees bowel Chi, dispels damp and heat, opens the organs and clears Chi stagnation.
Indications: Pain in the navel region, indigestion, dysentary, constipation.

Spleen 15

"Great Horizontal" - Da Heng

Intersecting point of the Yin Regulator Channel on the spleen channel
Location: At the level with the navel along the mamillary line.
Functions: Regulates the spleen, regulates and moistens the intestines, reduces digestive stagnation and transforms damp-heat.
Indications: Abdominal distension, diarrhea, constipation, intestinal paralysis, parasitic worms in the intestines, chronic sadness.

Conception Vessel 8

"Spirit Palace Gate" - Shen Que

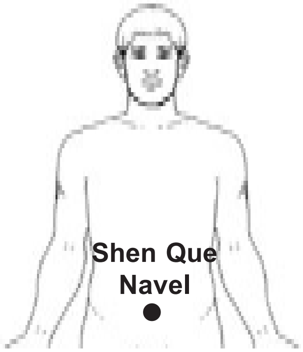

Doorway to the Original Chi,
Location: In the center of the navel
Functions: Tonifies, strengthens and regulates the spleen (Chi and yang), and stomach Chi, regulates the intestines, warms the interior and reduces digestive stagnation, tonifies the kidneys, warms the yang, dries dampness and dispels cold.
Indications: Chronic diarrhea, intestinal tuberculosis, all urinary disorders, shock resulting from intestinal adhesions, heatstroke, rectal or anal prolapse, restless fetus.

Governor Vessel 4

"Door of Life" - Ming Men

Back Doorway to the Lower Tan Tien

Location: Below the spinous process of the second lumbar vertebrae. Approximately at the level of the navel.

Functions: Nourishes the Original Chi, strengthens and harmonizes the kidneys, tonifies ching and yang, clears channels and invigorates collateral vessels, benefits the lower back and bones, regulates water pathways.

Main point to build life fire; main point for sexual/genital problems.

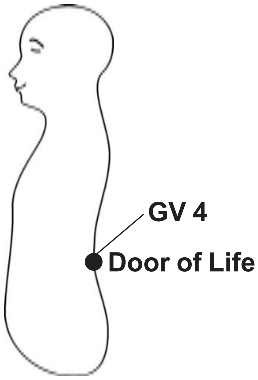

Indications: Bone disorders, chronic nephritis, enuresis, low sex drive, fatigue, spermatorrhea, impotence, irregular menses, painful menses, no menses, scanty menses, abnormal uterine bleeding, lower back stiffness and pain, kidney pain radiating to the abdomen, hemorrhoids, urinary incontinence, painful urination, diarrhea, sciatica, spinal myelitis.

Governor Vessel 6

"Middle of Spine" - Jizhong

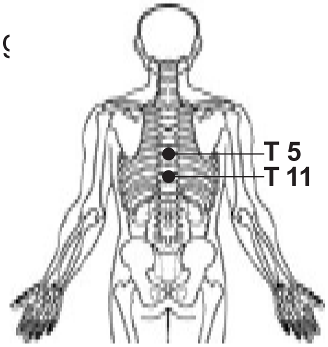

Location: Below the spinous process of the 11th thoracic vertebrae (T-11).

Functions: Stimulates Ching Chi, benefits spleen, stomach and kidneys and liver.

Indications: Hepatitis, seizures, low back pain, paralysis of lower limbs, blood in stools, diabetes, diarrhea.

Governor Vessel 11

"Spirit's Path" - Shen Tao

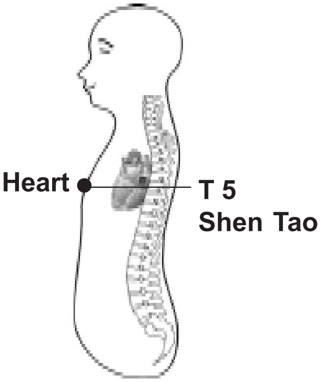

Location: Below the spinous process of T-5 (opposite CV-17, the Heart Point).

Functions: Calms heart and mind, regulates heart Chi and yang, expands and relaxes the chest, benefits heart and lungs, sedates pain, sedates fright and dispels wind.

Indications: Anxiety and palpitations due to fear or fright, neurasthenia, asthma, cough, shortness of breath, chest and hypochondriac pain, insomnia, aphasia due to winds, stroke, forgetfulness, fever.

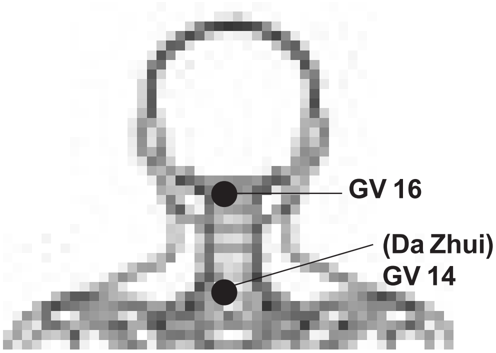

Governor Vessel 14

" Big Vertebra" - Da Zhui

Influential point of yang
Sea of Chi point
Intersection point of all yang channels.
Location: Below the spinous process of the seventh cervical vertebra (C-7), approximately at the level of the shoulders.
Functions: Opens the yang, clears the brain, calms the spirit, tonifies Protective (Wei) Chi, reduces fever, relieves exterior conditions, clears heat.
Indications: Fever, sunstroke, malaria, psychosis, seizures, bronchitis, asthma, pulmonary tuberculosis, emphysema, hepatitis, blood diseases, eczema, hemiplegia, pain in the back of the shoulders, cold-induced diseases, cough, fever and chills.

Governor Vessel 16

“Jade Pillow” or "Wind Palace" - Feng Fu

Sea of Marrow Point, Window of Sky Point, Ghost Point
Intersection point of the Yang Regulator Channel on the Governor Vessel
Location: At the base of the skull, about one inch above the posterior hairline.
Functions: Benefits and clears the brain, calms spirit, opens the sensory orifices, dispels wind, wind-cold and wind-heat, smoothes joint functions.
Indications: Seizures, mania, hemiplegia, loss of speech due to stroke, delirium, suicidal behavior, fear and fright, anxiety, common cold, sensation of heaviness in the head, headache, dizziness, numbness of the limbs, deaf-mutism, blurred vision, sinusitis, stiff neck.

Governor Vessel 20

"One Hundred Meetings" - Bai Hui

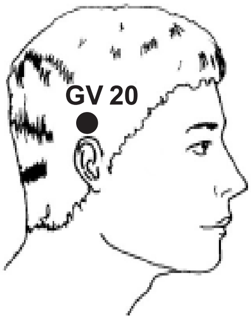

Sea of Marrow Point
Location: At the crown of the head, approximately on the midpoint of the line connecting the top apex of the two ears.
Function: Clears the senses, calms the spirit, extinguishes liver wind, stabilizes ascending yang.
Indications: Headache, dizziness, shock, depression, hypertension, insomnia, seizures, prolapsed anus, prolapsed uterus, mental dullness, hemorrhoids.

Bibliography

Bensky, Dan. Acupuncture: A Comprehensive Text, Seattle, WA: Eastland Press, 1981.
Chia, Mantak. Awaken Healing Light of the Tao, Huntington, NY: Healing Tao Books, 1993.Chi Nei Tsang: Internal Organ Chi Massage. Huntington, Ny: Healing Tao Books, 1990.
Cooper, Groffrey M. The Cell, A Molecular Approach, Washington, DC: Sinauer Associates, 1996.
De Lee, H. Lorenzen. Message from Water: Word of Recommendation, 1984.**DeLaney, Colleen, Leonard, David, and Kitsch, Lancelot.** The Acupuncture Point Book, Santa Cruz, CA: Roast Duck Productions, 1989.
Dong, Paul and Esser, Aristide H. Chi Gong: The Ancient Chinese Way to Health. New York: Paragon House, 1990.
Jou, Tsung Hwa. The Tao of Meditation: Way to Enlightenment. Piscataway: Tai Chi Foundation, 1983.
Lade, Arnie. Acupuncture Points: Images and Functions, Seattle, WA: Eastland Press, 1989.
Lindsay, Mary. The Visual Dictionary of the Human Body, First American Edition, New York, NY: Dorling Kindersley Publications, 1991.
Lu, Kuan Yu. Taoist Yoga: Alchemy and Immortality, New York, NY: Samuel Weiser, 1970.
Maciocia, Giovanni. The Foundations of Chinese Medicine, Edinburgh, Scotland: Churchill Livingstone, 1989.
Matsumoto, Kiiko and Birch, Stephen. Extraordinary Vessels. Brookline: Paradigm Publications, 1986.
McGraw-Hill Companies Human Physiology, 1999.
Morrison, Philip and Phylis and the Office of Charles and Ray Eames. Power of Ten:1982
Nilsson, Lennart. The Body Victorious. New York: NY: Delacorte Press, 1985.
Shanghai College of Traditional Medicine, translated by O'Connor, John and Teeguarden, Iona Marsaa. Acupressure Way of Health: Jin Shin Do, Tokyo, Japan: Japan Publications, 1978.
Tortora, Gerard J. and Grabowski, Sandra R. Introduction to the Human Body, Fifth Edition. New York, NY: Wiley, 2000.
Yang, Jwing-Ming. The Root of Chinese Chi Kung Jamaica Plain, MA: YMAA, 1989.

$19.95

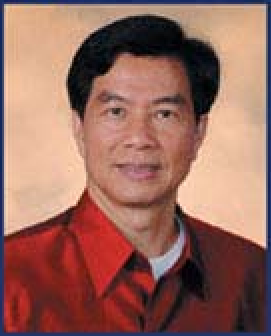

Cosmic Healing I

Cosmic Chi Kung

Mantak Chia

Illustrated by Udon Jandee

Cosmic Healing I

Cosmic Healing I builds on the healing aspects of the Universal Tao foundation practices. In Cosmic Healing I the student will learn how to use the Cosmic Healing Chi Kung in combination with color for guiding the healing learn to process in yourself & other, you'll make use of your Personal Stars & connections with the Big Dipper & North Star so as to have the chi to "sweep & clean out" sick energies.

A New Approach to Healing

In Cosmic Healing I Master Mantak Chia, continue to bring the ancient, time-tested wisdom of the Universal Tao System to the Western world. Cosmic Healing I presents a whole new understanding and approach to healing with detailed explanations of self-healing techniques and ways to transform negative, sick energies from them and others.

Hand Techniques for Detoxifying and Rejuvenating the Vital Organs

Learn to develop your most powerful healing potential through the Cosmic Chi Kung combined with your Mind, Eyes, Heart & Intention. The Cosmic Chi Kung encourages your Compassion to bloom so as to enhance your self- healing & your ability to support other in healing them selves.

0 097487 672561

Universal Tao Publications

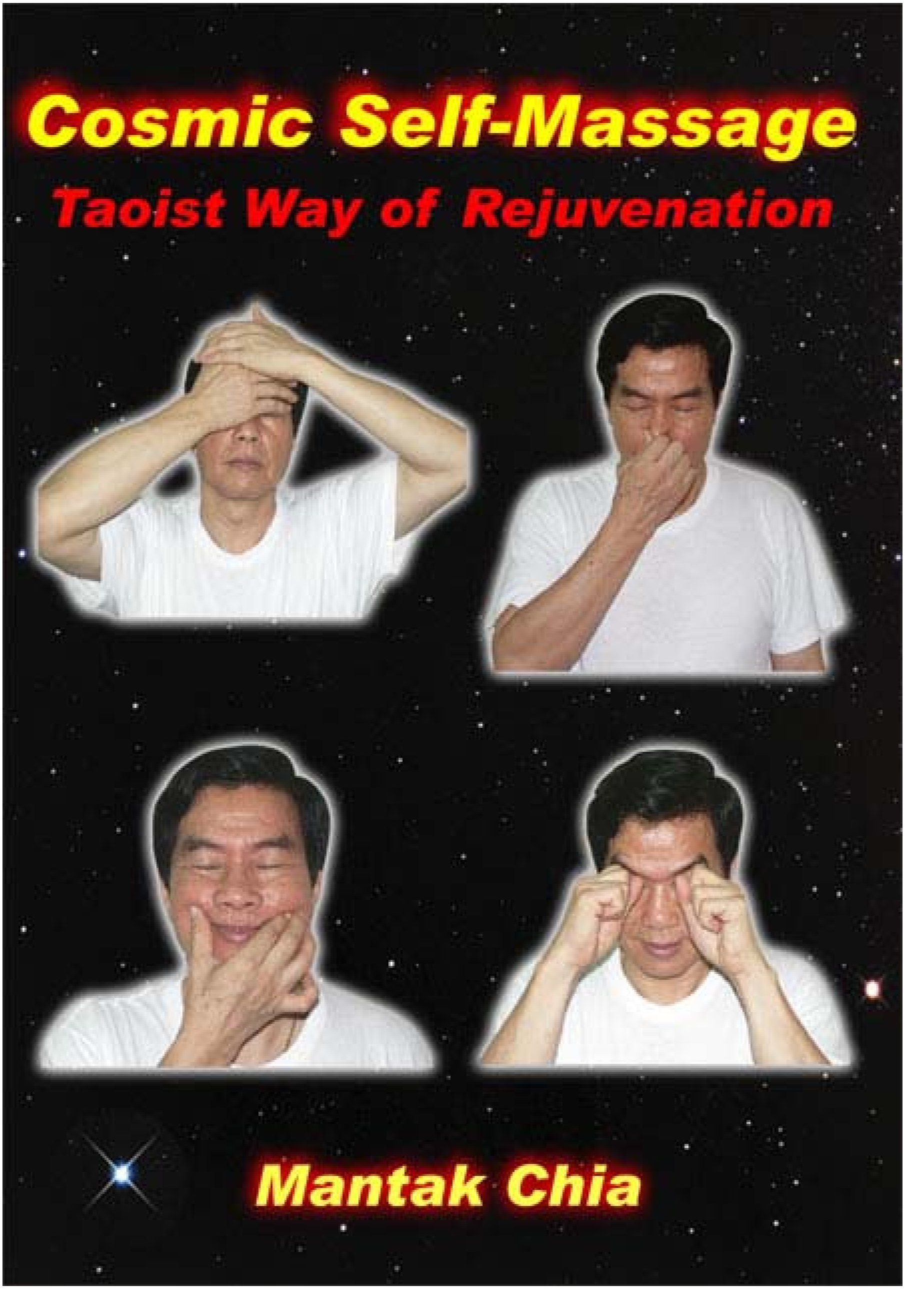

Cosmic Self-Massage

Taoist Way of Rejuvenation

Mantak Chia

Cosmic Self-Massage

Taoist Way
of Rejuvenation

Mantak Chia

Edited by:

Dena Saxer

Editor: Dena Saxer

Design and Publishing Consultant: Cathy Umphress

Illustrations: Juan Li

Computer Graphics: Saisunee Yongyod

Layout: Siriporn Chaimongkol

Production Manager: Saniem Chaisarn

Final Editing: Jean Chilton

Project Manager: W.U. Wei

© *North Star Trust*

First published in 2003 by:

Universal Tao Publications
274 Moo 7, Luang Nua,
Doi Saket, Chiang Mai 50220 Thailand
Tel: (66)(53) 495-596 Fax: 495-853
Email: universaltao@universal-tao.com
Web Site: universal-tao.com

Manufactured in Thailand

ISBN: 0-9671718-5-7

All rights reserved. No part of this booklet may be used or reproduced in any manner whatsoever without the express written permission from the author, with the exception of brief quotations embodied in critical articles and reviews. Anyone who undertakes these practices on the basis of this booklet alone, does so entirely at his or her own risk.

Contents

Healing Hand 1
  Palms 1
  Pericardium 1
  Large Intestine 1
Major Palm Lines 3
  Fingers have Corresponding Bodily Functions..................... 3
  Fingers Correspond to Emotions, Elements and Organs 8
  Head 10
  Face 14
  Temples 15
  Mouth 16
  Eyes 17
    A. Bring Chi Energy to the Hands and Eyes 18
    B. For the Eyeballs and Surrounding Area 20
    C. Pull up the Eyelids 20
    D. For the Eye Sockets 21
    E. For getting a Tear out 21
    F. Pull in the Eyeballs 23
    G. Learn to Maintain Eye Contact 25
  Nose 25
  Ears 26
    A. Outer Ear 26
    B. Inner Ear 29
  Gums, Tongue, Teeth 30
    A. Bring Chi Energy to the Hands 31
    B. Gums 31
    C. Gums and Tongue 31
    D. Tongue 32
    E. Teeth Clenching 34
    F. Energy to Teeth 34

Neck 36
- A. Thyroid and Parathyroid: Site of Courage, the Power of Speech 36
- B. Bring Chi Energy to the Hands 37
- C. Whole Neck 37
- D. Middle Neck 37
- E. Turtle Neck 38
- F. Crane Neck 38
- G. Massage the Neck 38

Shoulders 40

Detoxifying Organs and Glands 41
- I. Thymus Gland 41
- II. Heart 42
- III. Lungs 42
- IV. Liver 43
- V. Stomach, Spleen, Pancreas 43
- VI. Large and Small Intestines 44
- VII. Kidneys 45
- VIII. Sacrum 46

Knees and Feet 47
- I. Knees and Feet 47
- II. Feet: Roots of the Body 48

The meditations, practices and techniques described herein are ***not*** intended to be used as an alternative or substitute for professional medical treatment and care. If any readers are suffering from illnesses based on mental or emotional disorders, an appropriate professional health care practitioner or therapist should be consulted. Such problems should be corrected before you start training. This booklet does not attempt to give any medical diagnosis, treatment, prescription, or remedial recommendation in relation to any human disease, ailment, suffering or physical condition whatsoever.

Healing Hand

Humans can build all kinds of sophisticated instruments because of the magnificence of the hands and the fingers. Another higher function of the hands is the power of healing. Knowing the major points of the hands and fingers will enable you to stimulate and maintain the organs in good function.

Palms

The palms are where all major energies of Chi join. The palm can be the place from which the life-force is sent out to heal others or yourself. The palm also is the place where energy can be received and enter into the bone structure and into the major organs.

Pericardium

The pericardium (P-8) is the main place of energy concentration. You can collect the energy in this point and transmit stronger energy from this point.

Large Intestine

The large intestine (LI-4) is the major point which controls all the pain in the body, especially in the sense organs (eyes, ears, nose) and headaches.

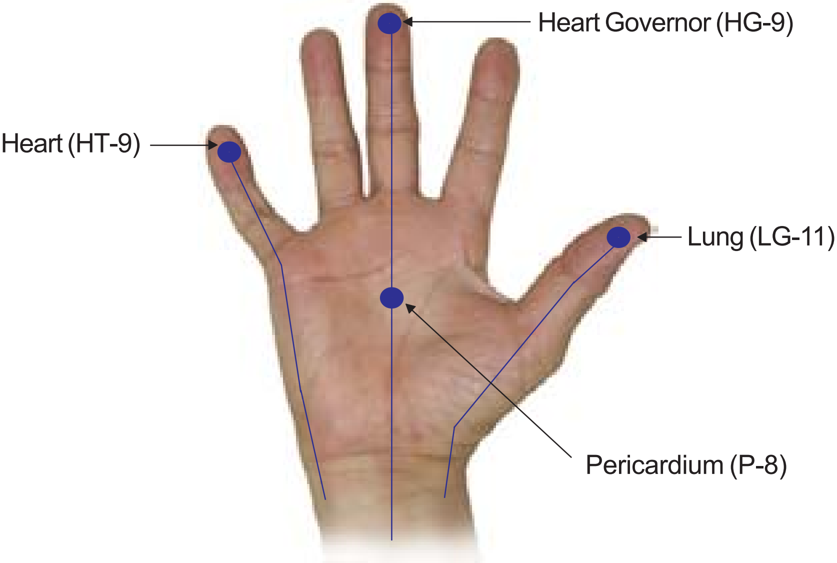
Heart Governor (HG-9)
Heart (HT-9)
Lung (LG-11)
Pericardium (P-8)

Fig. 1 *Pericardium*

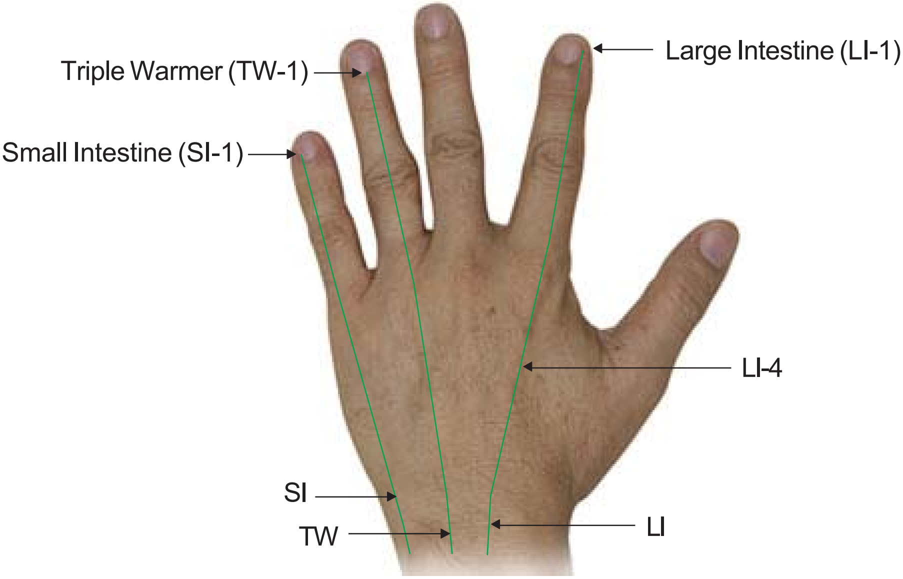
Triple Warmer (TW-1)
Large Intestine (LI-1)
Small Intestine (SI-1)
LI-4
SI
TW
LI

Fig. 2 *Large Intestine*

Major Palm Lines

The three major palm lines are the Life Line, the Line of Intellect and the Line of Emotion.

Fingers have Corresponding Bodily Functions

The fingers are connected to the organs' meridians. The joints of the finger bones are also related to parts of the organs and their corresponding senses and emotions.

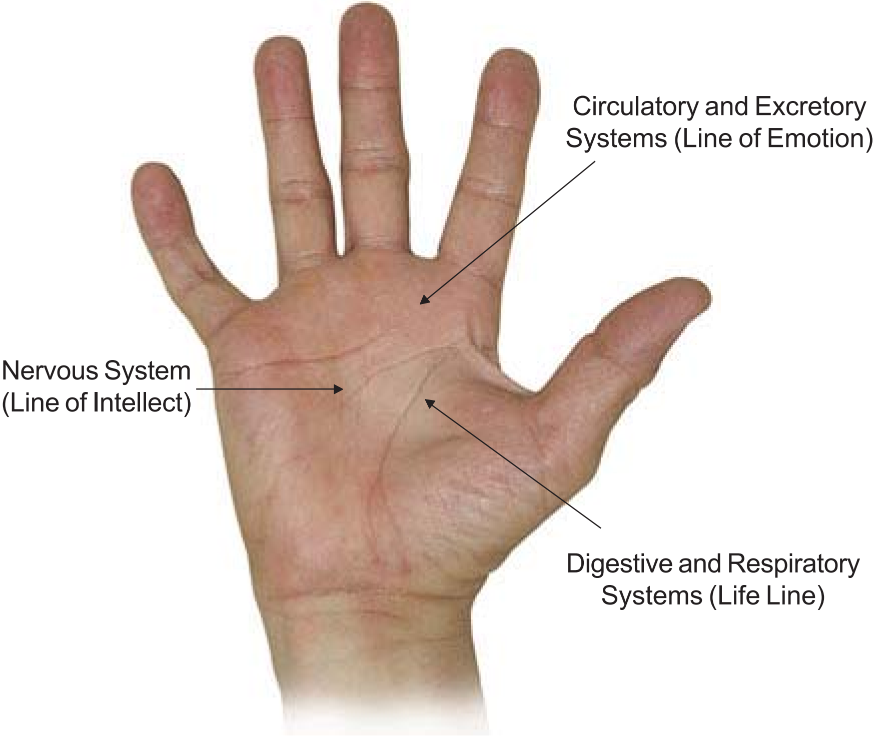
Circulatory and Excretory Systems (Line of Emotion)
Nervous System (Line of Intellect)
Digestive and Respiratory Systems (Life Line)

***Fig. 3** Three Major Palm Lines*

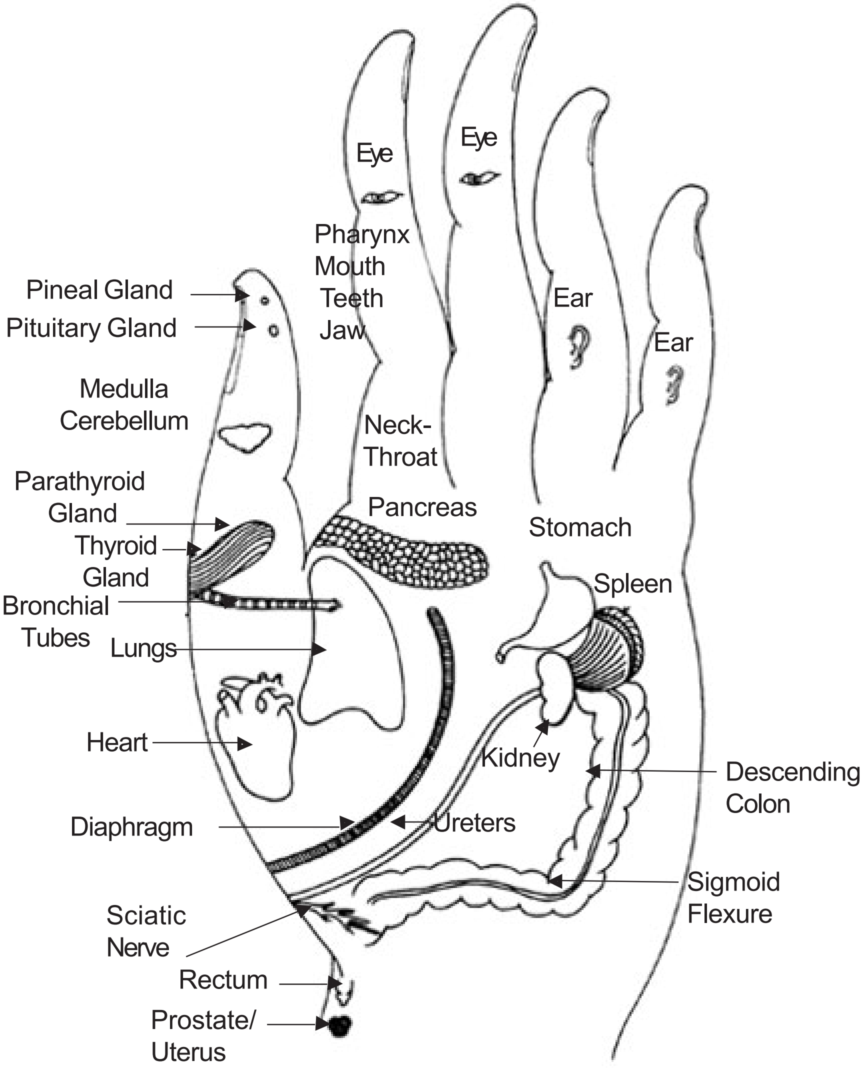
Eye
Eye
Pharynx
Mouth
Teeth
Jaw
Pineal Gland
Pituitary Gland
Ear
Ear
Medulla
Cerebellum
Neck-
Throat
Parathyroid
Gland
Pancreas
Stomach
Thyroid
Gland
Spleen
Bronchial
Tubes
Lungs
Heart
Kidney
Descending
Colon
Diaphragm
Ureters
Sigmoid
Flexure
Sciatic
Nerve
Rectum
Prostate/
Uterus

Fig. 4 *Fingers and Their Corresponding Bodily Functions Through the Organs' Meridians*

Massage the Hands

Always start by rubbing your hands until they are warm.

1. Massage the pericardium (P-8). Use the thumb to press the middle of the palm with a circular motion.

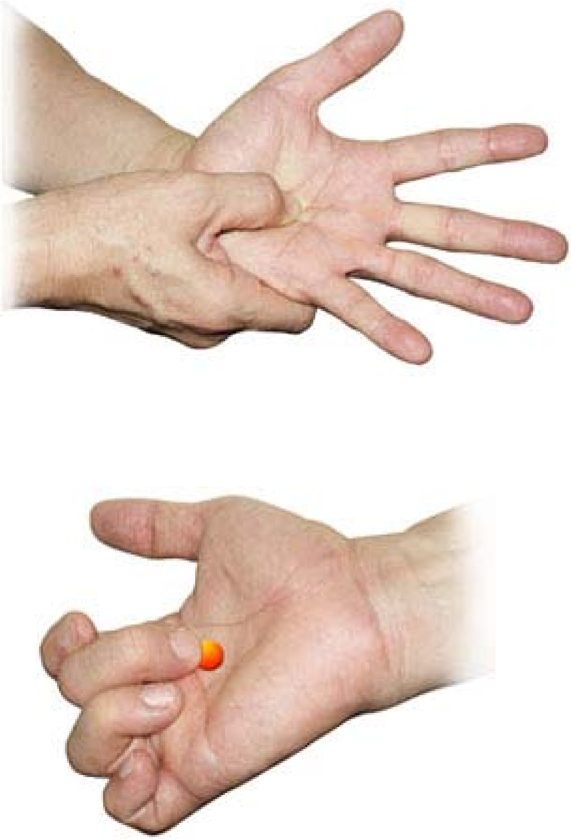

Fig. 5 *Massaging the Pericardium with fingers cupped in the palm in a half-fist, the pericardium is the point at the tip of the middle finger.*

2. Massage the hegu (LI-14). Press the thumb around the point in a circular motion, and press more at the index finger bone. Find the pain point and massage it away.

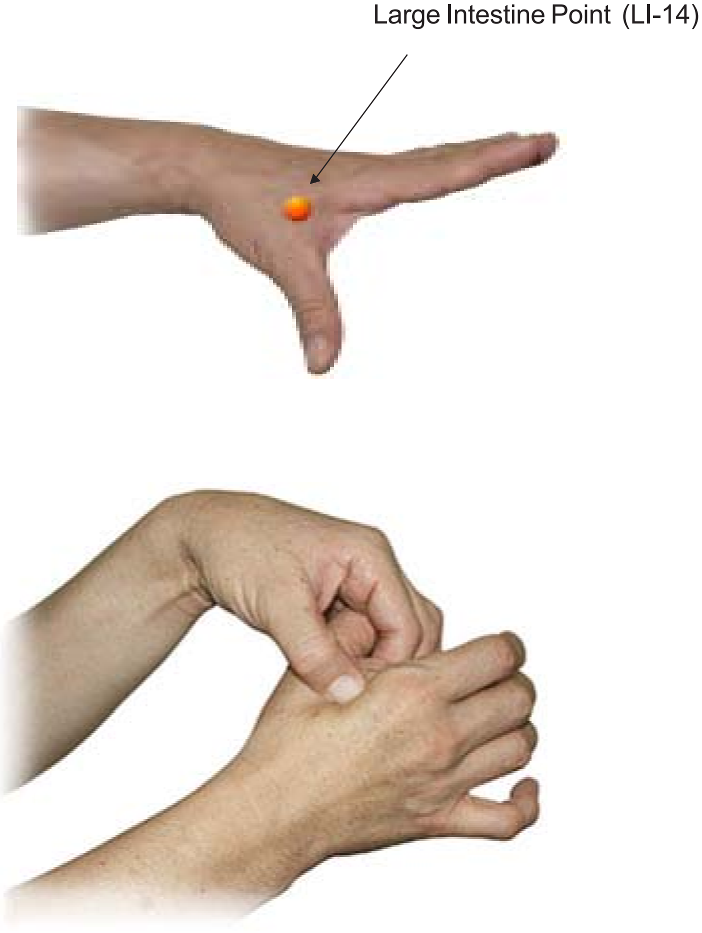
Large Intestine Point (LI-14)

Fig.6 *Massaging LI-14*

3. Massage the major palm lines. Use the thumb to massage along the palm lines. Massage more towards the thumb bone and along that bone. When a lot of emotion is held inside, find the sore point and massage it.

4. Massage the back of the hand. Use the thumb to press along the bones on the back of the hand. When you find a tender spot, take more time to work on it.

5. Massage the fingers. Always rub your hands until warm. Use the right hand's fingers to wrap around the left thumb, and then, one by one, squeeze, hold and release each finger on the left hand three to six times. Start with the left hand and continue to the right hand's fingers, according to the elements of the finger. This will greatly help to control emotions.

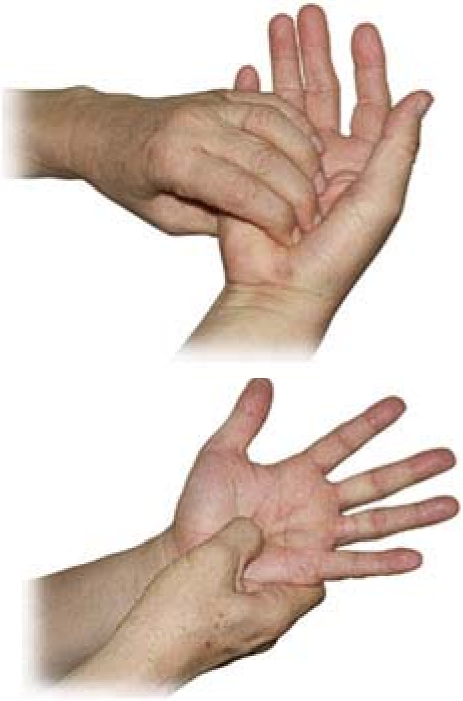

Fig. 7 *Massaging the Palm Lines*

Fingers Correspond to Emotions, Elements and Organs

A. Thumb corresponds to the element earth and is associated with the stomach. The corresponding emotion is worry.

B. Index finger corresponds to the element of metal and is associated with the lungs and large intestine. It links with the emotions of sadness, grief and depression.

C. Middle finger corresponds with the element of fire and is associated with the heart, small intestine, circulatory system and the respiratory system. It links with the emotions of impatience and hastiness.

D. Ring finger corresponds to the wood element and is associated with the liver, gall bladder and the nervous system; it corresponds to the emotion of anger.

E. Pinky finger corresponds to the water element; it is associated with the kidneys, and corresponds with the emotion of fear.

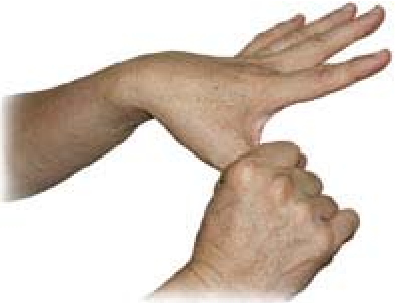

Worry

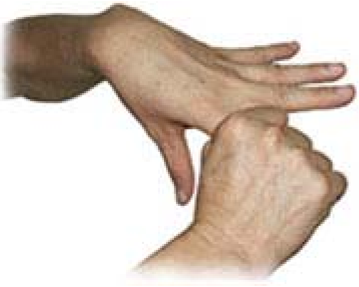

Sadness

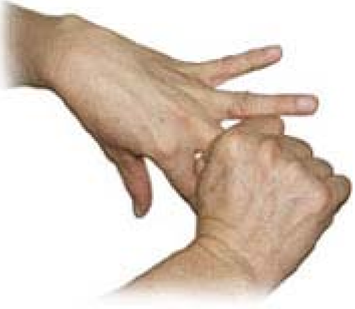

Impatience

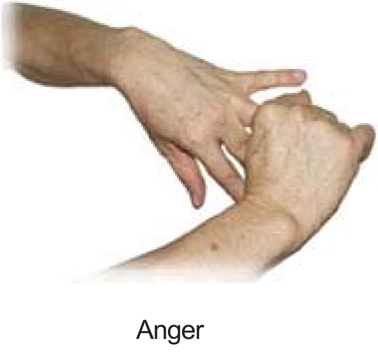

Anger

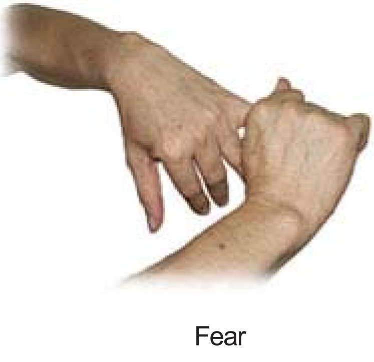

Fear

Fig. 8 *Fingers' corresponding emotions, elements and organs*

Head

A. Crown Point

This is located in the center of the crown; in the fontanelle area of an infant's skull there may still be a slight depression. The crown point is the junction of one hundred channels through which the energy of the body passes. Massage this area with both your middle fingers. This will relieve dizziness and headaches, which result from too much energy in the head. It also relieves high blood pressure and stimulates the nervous system.

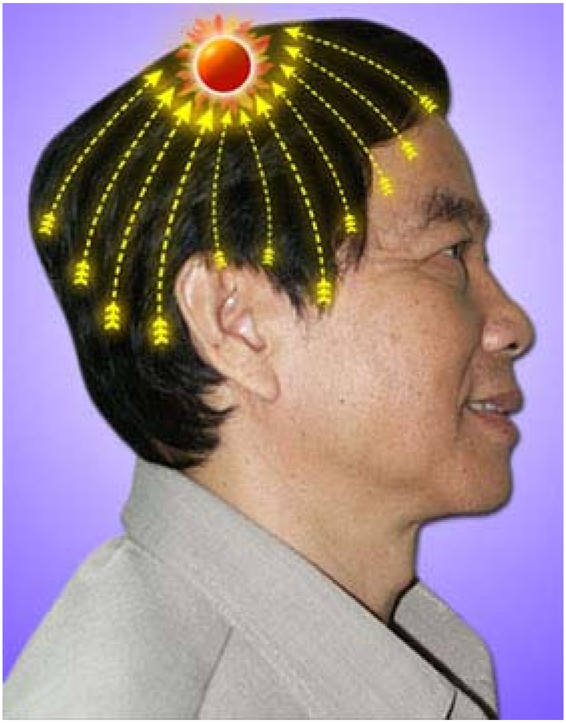

Fig. 9 *Crown Point is the junction point of one hundred energy channels.*

B. Bring Chi Energy to the Hands and Face

Inhale, contract the sexual organ, buttocks and middle of the anus. Rub the hands, clench the teeth, and put the tongue to the roof of the mouth. When the face, head, and hands are hot, breathe normally and begin to massage.

C. Knock the Head

Hit the head with the knuckles of the hand, knocking all around the head. Knocking the head lightly can help to clear your head, eliminate stubbornness and make your thinking sharper. Many of our students use this knocking of the head to release the pressures that they have from today's life of fast, advanced technology and the feeling of always having to keep up. This is especially true of those graduate students who feel a great deal of pressure and stress in keeping up with their studies. Each year students commit suicide because the pressures and stress accumulate too much in their heads, making them unable to think clearly: they start to feel everything in their society as too much pressure, which leads to worry, fear, sadness, and many, mixed emotions. The simple knocking of the head can release pressure and stress that accumulate there.

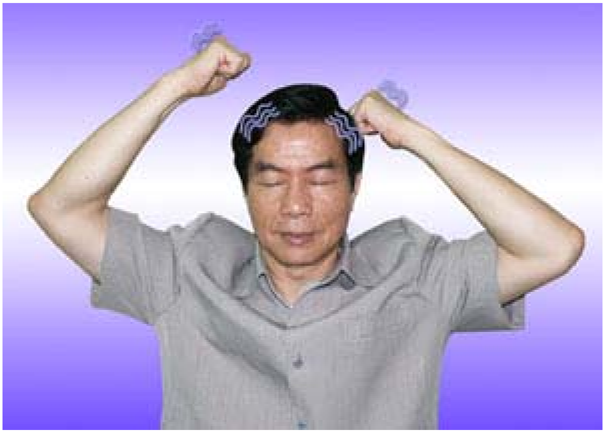

Fig. 10 *Knocking the Head*

D. Hold your Breath to Increase Chi Flow

Holding your breath will increase the Chi flow to the face. The head has many channels that join in the skull, especially in the crown point.

E. Scalp

Prepare your hands, head and scalp by warming up. Using both hands like a comb, press hard and move slowly, massage the scalp, going straight back from the hairline to the base of the skull. As you do this, mentally direct your energy from the back of the skull to your feet. Repeat 6-9 times. Massage more in any places in which you feel pain, until the pain goes away.

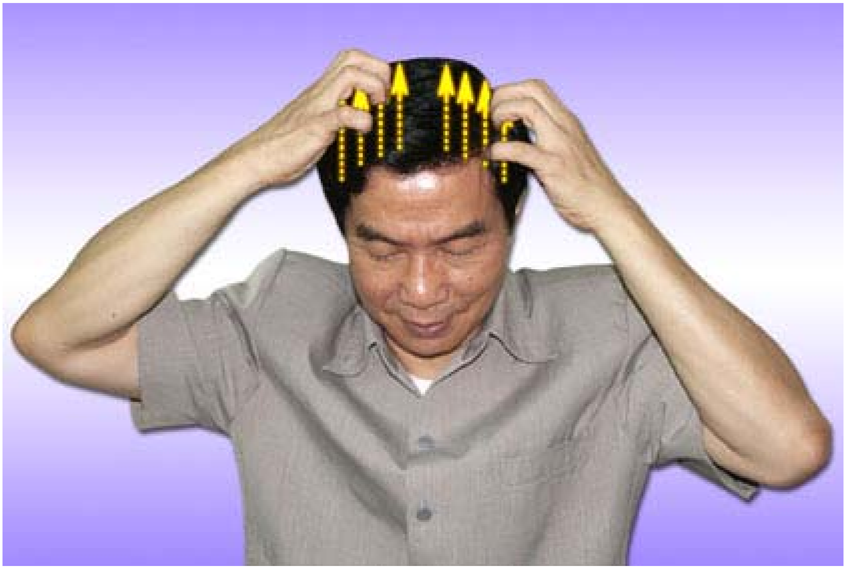

Fig. 11 *Massaging the Scalp*

F. Crest

Using your thumbs, massage the crest (the edge) at the base of the skull until you feel no pain there. This will help you reduce headaches and eye aches and will increase vision. This place in Tao tradition is called the Pool of Wind which tends to collect the "evil wind", the major cause of all the pain in the senses.

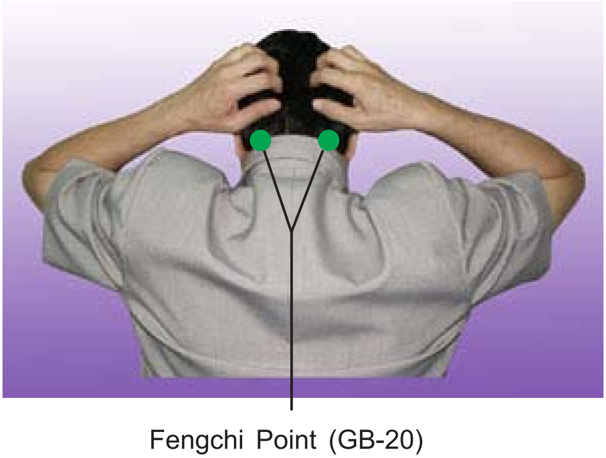

Fig. 12 *Crest - Edge of the Skull*

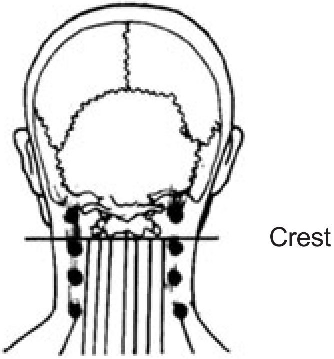

Fig. 13 *Go straight back from the hairline to the base of the skull.*

Face

Natural Beauty

Massaging your face with Chi is a far more effective beauty treatment than the most expensive cream or cosmetic, Your skin will glow brightly and eventually become less wrinkled. There are many meridians passing through or ending at the face. When blocked, they result in reduced flow of Chi energy and circulation. The face is the first impression imprinted in other people's minds. Chi circulation provides it with attractive personal energy.

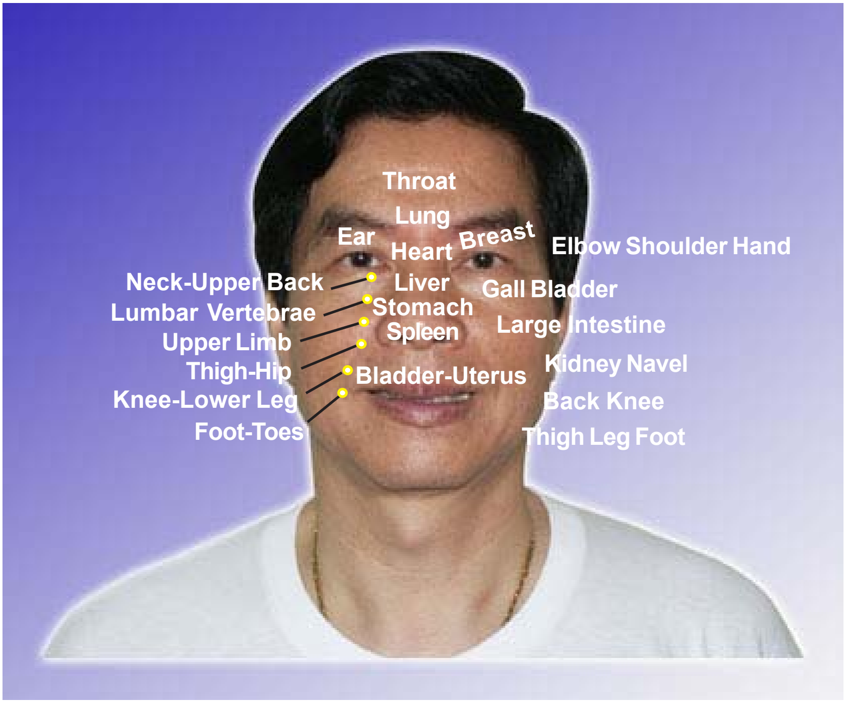
Throat
Lung
Ear
Heart
Breast
Elbow Shoulder Hand
Neck-Upper Back
Liver
Gall Bladder
Lumbar Vertebrae
Stomach
Spleen
Large Intestine
Upper Limb
Thigh-Hip
Bladder-Uterus
Kidney Navel
Knee-Lower Leg
Back Knee
Foot-Toes
Thigh Leg Foot

Fig. 14 *Head and Face*

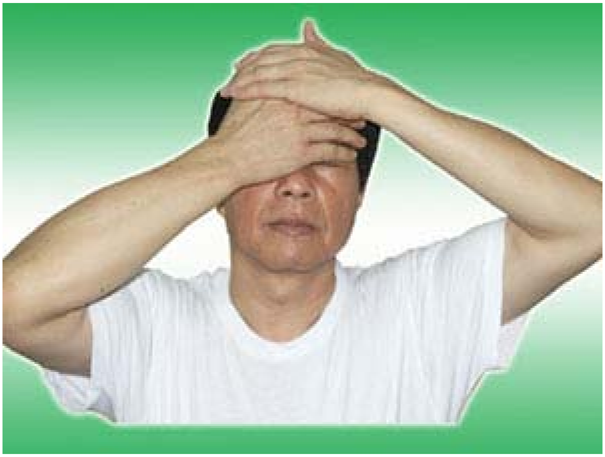

Fig. 15 *Massage the Whole Face*

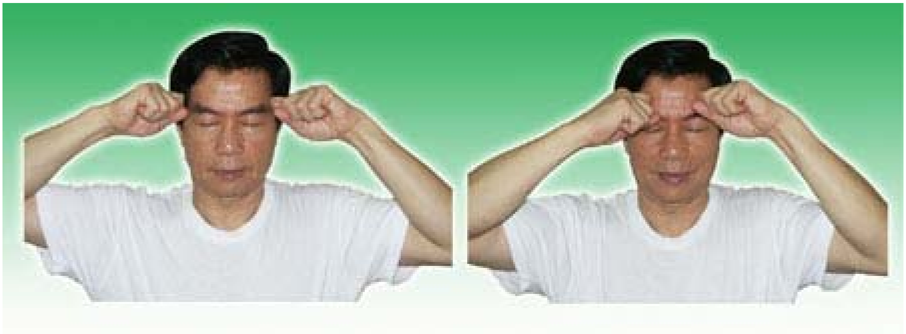

Fig. 16 *Massage the Mid Forehead*

Temples

Use your index fingers to massage the temples in a circular motion, first clockwise, then counterclockwise. Massage the forehead and the temples; use the knuckle rub from the middle of the forehead all the way to the temples ten to twenty times. These exercises will reduce headaches in the front and in the temples. Find the painful point and massage it until the pain is gone.

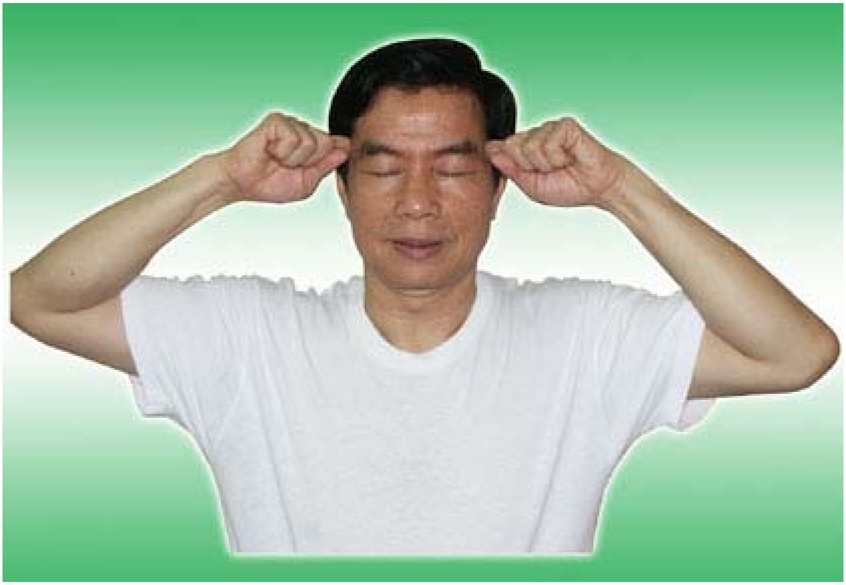

Fig. 17 *Massage the Temples*

Mouth

Depression makes the corners of the mouth drop. Looking cheerful, delightful, more attractive and happy are dependent so much on the eyes and the corners of the mouth. When the muscles of the mouth are loose because of stress, depression, or sadness, the corners of the mouth drop and the energy system is depressed and in low key. No one likes to look at a sad face or a depressed face; it makes other people feel sad and depressed, too.

The flow of energy in the body and the expression of the face are the main attractive powers of a person. Massaging the mouth muscles up will help to lift the corners of the mouth. The Inner Smile and lifting up the corners of the mouth are very important to building up attractive energy.

Beautify the Mouth Massage

Using the thumb and the index finger of the right hand, touch both corners of the mouth and feel the Chi from the thumb and index finger pass to the corners of the mouth. Slowly press and push up about one inch, release and start again at the corners, pressing up ten to twenty times each day.

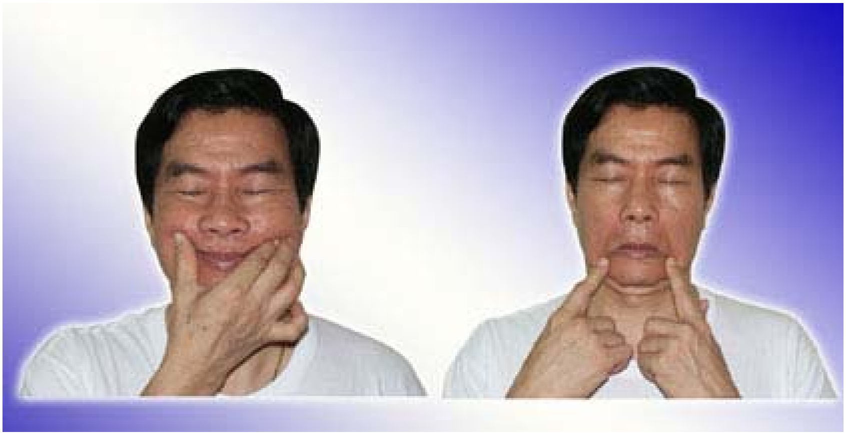

Cheerful Delight

Mouth Corners Drop

Fig. 18 *Beautifying the Mouth Massage*

Eyes

The eyes are the windows of the spirit. In Taoism we regard the eyes as Yang energy which will guide all Chi flow in the body. The eyes can greatly affect your personality Some people are born with a lot of white in their eyes--three portions of white to one portion of iris—sometimes called “thief eyes” or “danger attack eyes.” Such eyes can result in a suspicious look, portending unpleasant things. Through the exercises, you can gradually correct the white portion of the eyes.

Since the eyes are connected to the entire nervous system, they have a special importance. The eyes reveal the health of your entire body. Through the eyes we can tell which organs are weak and/or toxic. Massaging the eyes will remove stress from the vital organs. Nowadays people use their eyes much more than in the past to read, watch television, and work with computers, electronics and microscopes. This strains them a great deal and makes the openings of the organs loose, so that much of the organ energy is drained out. In Taoism, we regard the eyes as the doorways to the soul as well as the opening of the liver.

A. Bring Chi Energy to the Hands and Eyes

Repeat the procedure for bringing energy to the hands by inhaling; holding the breath; and contracting the sexual organ, buttocks and middle of the anus and both the left and right sides of the anus. Direct the Chi to both eyes. Rub the hands, clench the teeth, place the tongue on the roof of your mouth. Direct the energy to the face and then the hands. When your hands are hot, focus on your eyes until you feel them filled with energy.

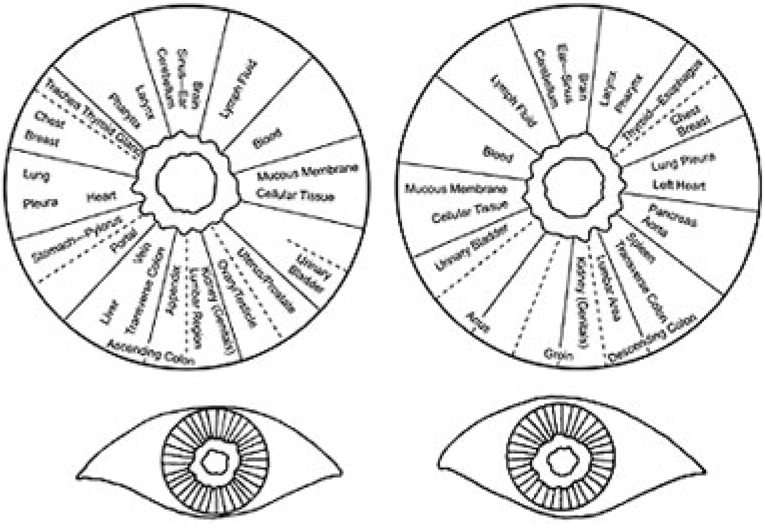
Brain
Sinus—Ear
Cerebellum
Lymph Fluid
Blood
Mucous Membrane
Cellular Tissue
Urinary Bladder
Uterus/Prostate
Ovary/Testicle
Kidney (Genitals)
Lumbar Region
Appendix
Transverse Colon
Vein
Liver
Ascending Colon
Portal
Stomach—Pylorus
Heart
Lung
Pleura
Chest
Breast
Trachea Thyroid Gland
Pharynx
Larynx
Brain
Ear—Sinus
Cerebellum
Lymph Fluid
Larynx
Pharynx
Thyroid—Esophagus
Chest
Breast
Lung Pleura
Left Heart
Pancreas
Aorta
Spleen
Transverse Colon
Lumbar Area
Kidney (Genitals)
Groin
Descending Colon
Anus
Urinary Bladder
Mucous Membrane
Cellular Tissue
Blood

Fig. 19 *Eyes are the Windows of the Spirit*

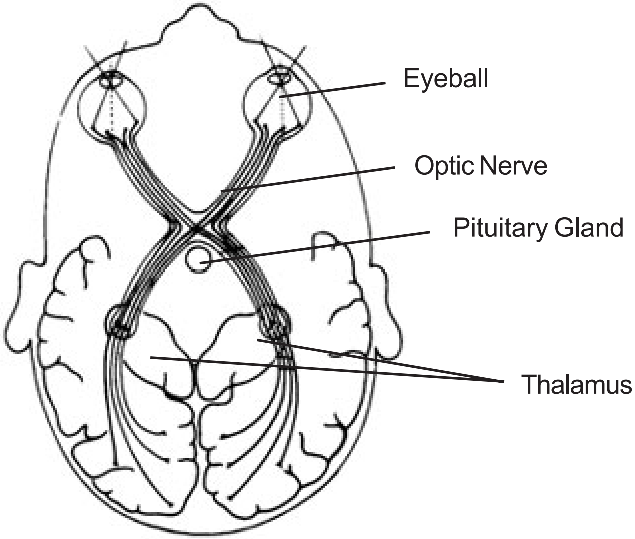
Eyeball
Optic Nerve
Pituitary Gland
Thalamus

Fig. 20 *Eyes are the Doorways to the Soul*

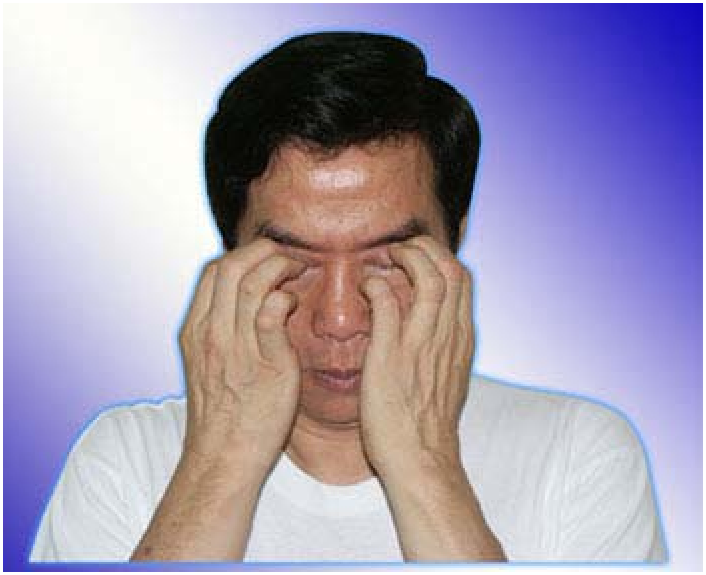

Fig. 21 *Use the Fingertips to Gently Massage the Eyeballs*

B. For the Eyeballs and Surrounding Area

Close your eyes. Use your fingertips to gently massage your eyeballs through your closed eyelids, six to nine times clockwise, then six to nine times counterclockwise. Then gently massage the area around the lids the same number of times. Be aware of painful spots and massage those places until the pain goes away. Pay special attention to the inner and outer corners of the eyes. These are points of the gall bladder meridian and will relieve eye ailments if massaged.

C. Pull up the Eyelids

Pulling up the eyelids will increase the fluid. Use the thumb and index finger to pinch, pull up and release the eyelids six to nine times.

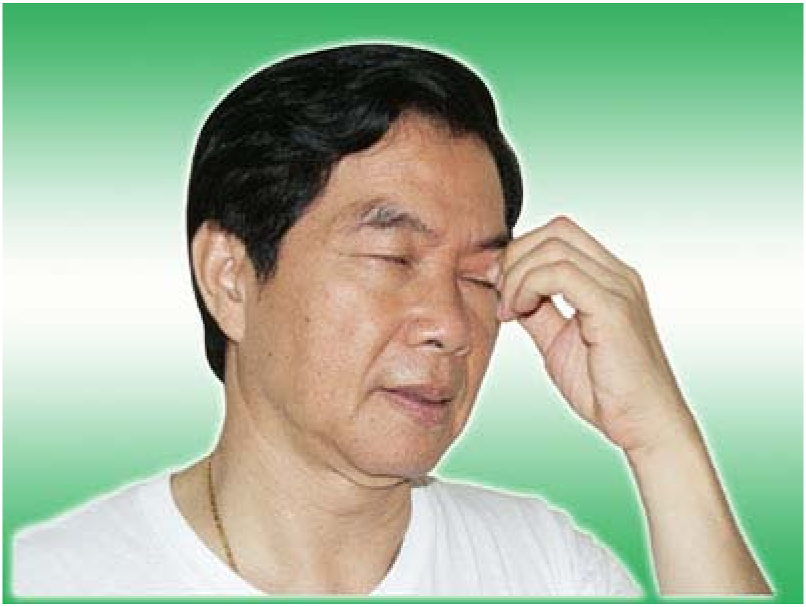

Fig. 22 *Pull Up the Eyelids*

D. For the Eye Sockets

Bend your index fingers and use the lower section (second phalanx) of each thumb to rub the upper and lower bones of the eye sockets six to nine times.

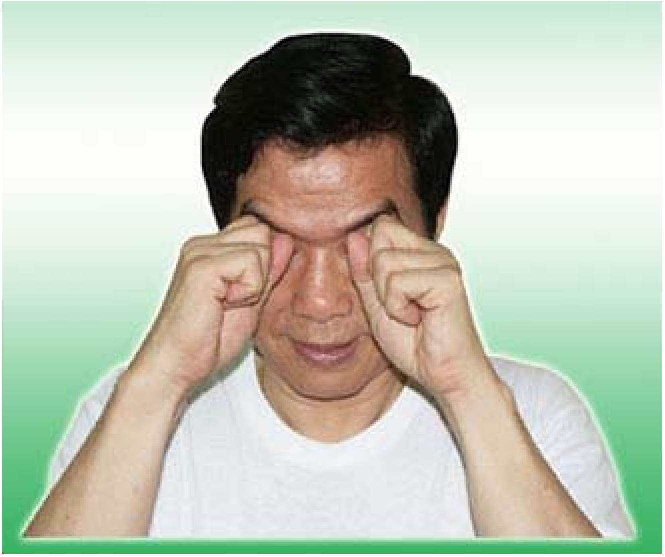

Fig. 23 *Massage the Eye Sockets*

E. For getting a Tear out

Hold an index finger up about eight inches from the eyes, or put a dot on the wall five or six feet away from you. Stare at it intently without blinking your eyes until you feel like a fire is burning in them. The Taoists believe that the toxins will burn out of the body through the eyes. They will begin to tear. Do this to strengthen your eyes. Then, rub your hands until warm; close your eyes and cover your eye sockets with your palms. Feel the Chi from the hands absorbed into the eyes. Rotate your eyes six to nine times, first in a clockwise direction, then counterclockwise.

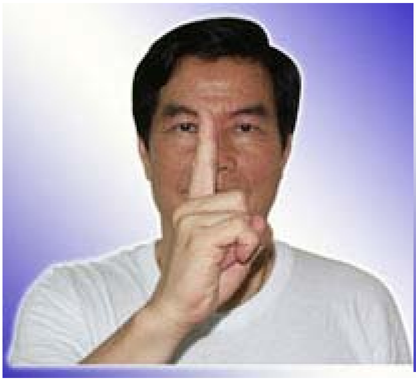

Fig. 24 *Getting a Tear Out*

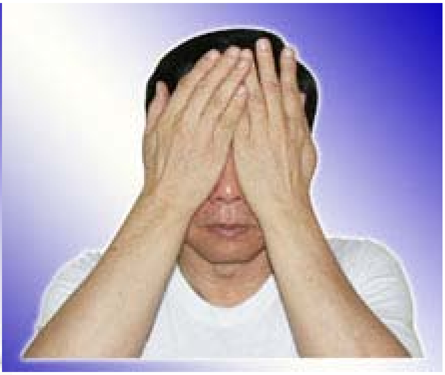

Fig. 25 *Absorbing the Chi into the Eyes*

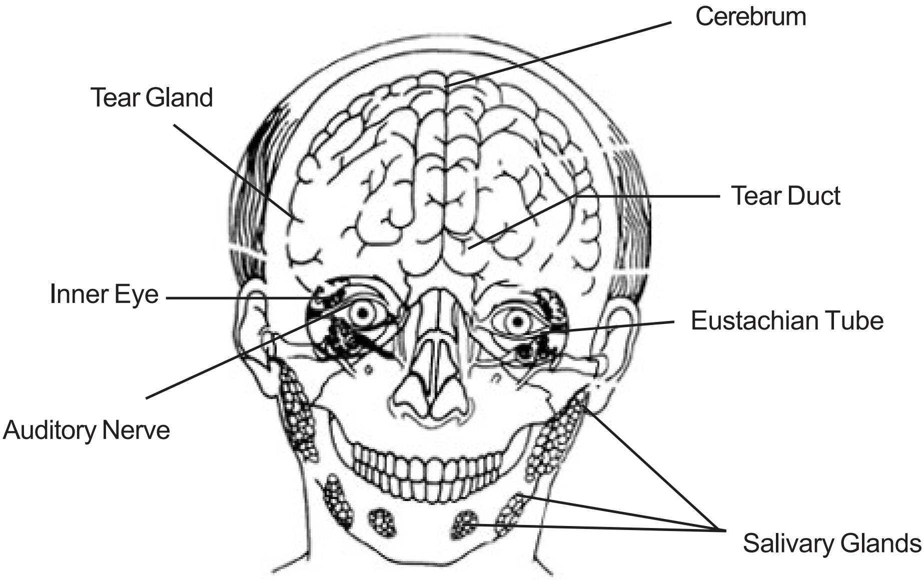

Cerebrum
Tear Gland
Tear Duct
Inner Eye
Eustachian Tube
Auditory Nerve
Salivary Glands

Fig. 26 *Parts of the Eyes Connect with Senses and Brain*

F. Pull in the Eyeballs

The eyes are divided into five parts. Each part is closely connected with the organs and nerves. Become aware of the eyes.

Pulling and pressing the eyeballs will exercise the organs, sense, glands and the brain. This is also the best exercise for the eye muscles. The eyes have many muscles which we do not exercise very much and, thus, they become weak, contributing to poor eyesight.

1. With the eyes still closed and cupped by the palms, inhale, contract the anus and sexual organ, and pull the eyeballs back into the sockets.
2. Contract the middle of the anus and the middles of the eyeballs.
3. Contract the front of the anus and the tops of the eyeballs.
4. Contract the back of the anus and the bottoms of the eyeballs.
5. Contract the right side of the anus and the right sides of the eyeballs.
6. Contract the left side of the anus and the left sides of the eyeballs. This exercise not only strengthens the eyes but also the pituitary and pineal glands and the inner ear including the ear drum and canals. When you pull the eyeballs in and upward and look toward the crown, you are exercising the upper muscles and stimulating the pituitary gland and pineal gland.

When you contract and pull in the middle of your eyeballs, you are exercising the back of the eye muscles and the inner ear.

When pulling in the outer corners of the eyes, you are strengthening the side eye muscles as well as the ear canals end the ear drums.

When pulling in the inner corners of the eyes, you are strengthening the inner side muscles, the tear ducts and the nose.

When pulling in the lower parts of the eyes, you are pressing the lower part of the ear canals and the nervous system.

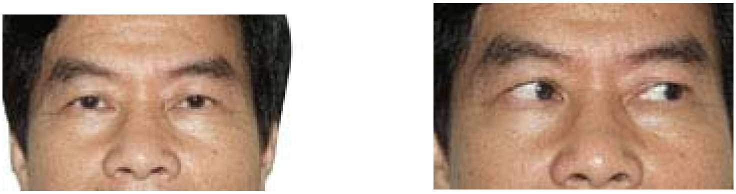

Pressing into the Inner Ear

Left Eye Pressing into the Ear Canal. Right Eye Pressing into the Eustachian Tube

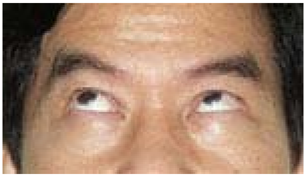

Pressing into the Pituitary Gland

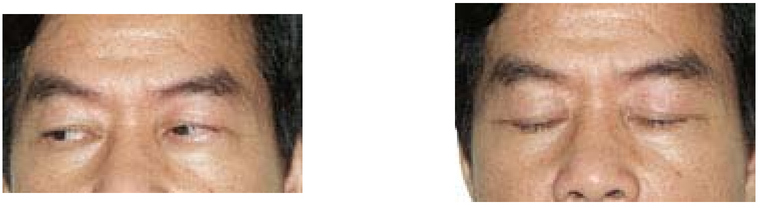

Right Eye Pressing into the Ear Canal. Left Eye Pressing into the Eustachian Tube

Pressing into the Eustachian Tube

Fig. 27 *Moving the Eyes*

G. Learn to Maintain Eye Contact

Some people in eye to eye contact with other people feel nervous and frightened, and their voices become very low and hard to hear because of weak organs. Some people will look around and not look into your eyes when you talk to them. This can be caused by weakness of the gall bladder and kidneys. To improve this, you can use the Inner Smile, Six Healing Sounds and Tao Rejuvenation, plus the practice of staring.

Look at your face in a mirror for two to five minutes each day for the first week. After ten days you can begin to stare at your eyes and increase your confidence by looking at your irises. Gradually you will lose the fear of looking into other people's eyes.

Nose

The nose has several important functions. When we breathe properly through the nose and not through the mouth, the nose filters out dirt, preventing it from reaching the lungs. It also regulates the temperature of the air: when the air is too cold, the nose will warm it up first. Without this regulating action, extreme temperatures could injure the lungs, make us susceptible to upper respiratory illnesses and subject to getting colds easily. One great advantage about people who practice the Tao System is that they seldom get colds.

The nose has three meridians running through it: the large intestine, the stomach, and the Governor or back Channel. Rubbing the nose strengthens the temperature regulator stimulates the above organs, and increases hormone secretion. In China just a few needles inserted in the nose serve as a general anesthetic for any part of the body to be operated on.

An unhealthy nose affects the personality. A thin, flat and unhealthy looking nose, or a badly shaped nose, can make you less attractive to other people. A strong nose can help you to have good Chi. The nose is the first place into which the breath of life enters.

A weak nose usually will be infected, and a lot of mucous can leak into the sinuses. A weak nose also can affect the voice. A good singer always has a good nose. Rubbing and massaging the nose will increase the Chi and will improve circulation around the nose.

Ears

In China we believe that a person with thick, long ears will have a long, healthy life and that the personality will be more attractive.

The following exercises can prevent hearing loss which occurs gradually as we age. The ears are acupuncture maps of the whole body, containing 120 points. Many acupuncturists now use only the ear points to cure many ailments as well as for weight control.

A. Outer Ear

Repeat the method for bringing energy to the hands, contracting the left and right sides of the anus.

1. Front and back: Make a space between your index and ring fingers and simultaneously rub in front and in back of the ears.
2. Ear shells: Rub the ear shells with all your fingers. This will stimulate the autonomic nervous system and warm up your whole body, especially in the cold weather.
3. Ear lobes: Using your thumb and index finger, pull down on the ear lobes.

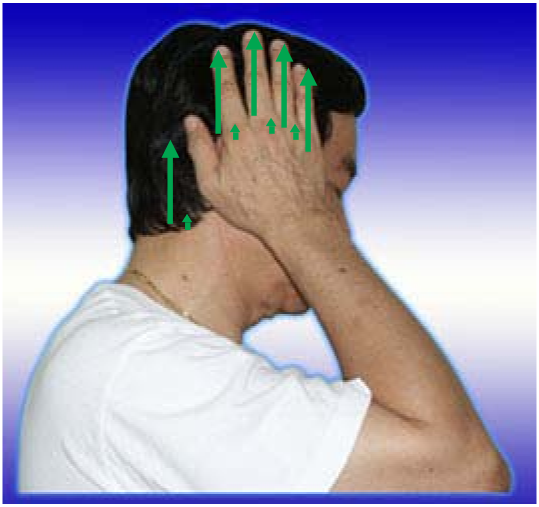

Rub in Front and Back of Ears

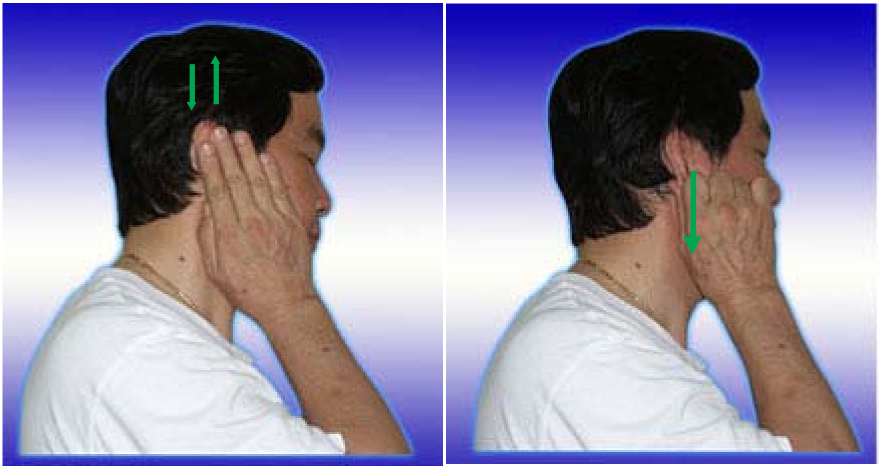

Rub the Ear Shells

Pull Down on the Ear Lobes

***Fig. 28** Rubbing the Ear*

4. Ear drum exercises: For the outer ear drum, repeat the method for bringing energy to the hands, contracting the left and right sides of the anus. Inhale and then exhale completely. Put your index fingers in your ears; it should feel as if there is a vacuum in the ears. If it does not, then exhale more. Move your index fingers back and forth six to nine times at your own pace until you can feel that the insides of the ears are moving, and pull out the fingers with a quick movement. You should hear a “pop” sound, and you will feel that you can hear better and that your mind is clearer.

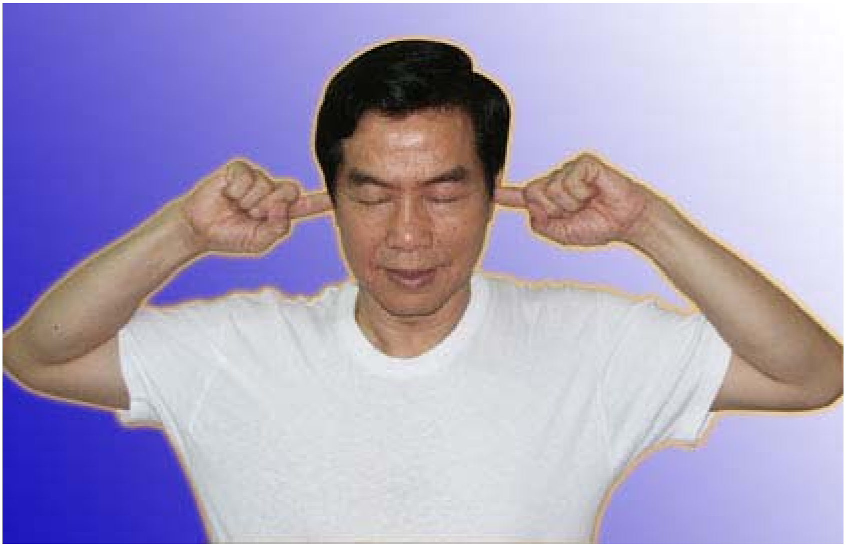

Fig. 29 *Outer Ear Drum Exercise*

B. Inner Ear

Repeat the method for bringing energy to the hands, contracting the left and right sides of the anus.

The inside of the inner ear, being inaccessible, is usually not exercised and grows weaker with age. These two exercises use air pressure and vibrations to strengthen the inner ear. The ear canals, the nose canal, and the mouth are connected together, so in this exercise we are using the pressure that builds in the lungs and bringing it back up to the mouth, thus adding pressure to the inner ear drums. This is how to exercise the inner ear drums.

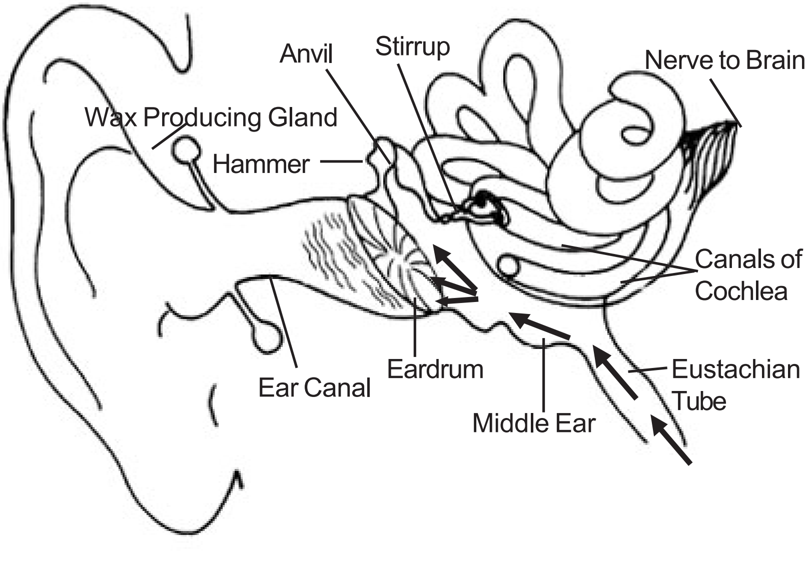
Anvil
Stirrup
Nerve to Brain
Wax Producing Gland
Hammer
Canals of Cochlea
Eardrum
Ear Canal
Eustachian Tube
Middle Ear

Fig. 30 *Diagram of Inner Ear Drum*

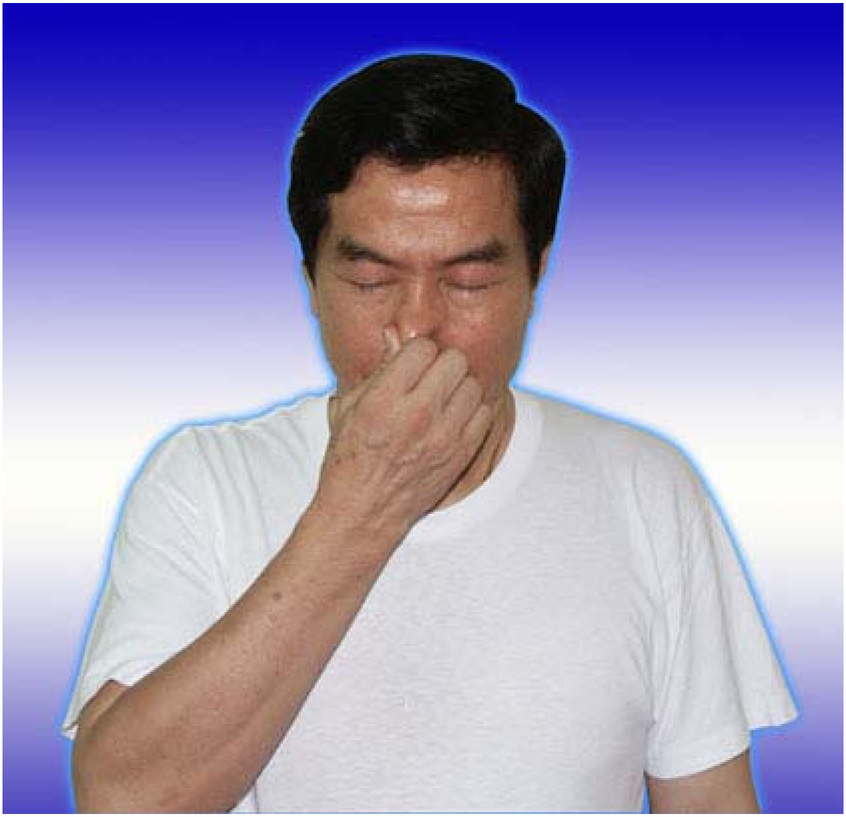

Fig. 31 *Inner Ear Drum Exercise*

Gums, Tongue, Teeth

Healthy teeth require healthy gums as their foundation. These exercises will strengthen both the gums and teeth. Teeth are the excess energy of the bones, and when the teeth get stronger, so do the bones. When the teeth and tongue are strong, your breath improves as well, eliminating bad breath.

We regard saliva as an essential form of energy which can lubricate the organs and digestive system. The tongue is the opening of the heart, and both are made of similar tissue. A healthy and clean tongue will strengthen the organs, especially the heart. You should clean your tongue twice a day with a brush or scrape it with a tongue scraper, and massage your tongue with a tongue depressor or a clean finger. Find the painful spots and massage there until the pain goes away.

A. Bring Chi Energy to the Hands

Repeat the procedure for bringing energy to the hands, contracting the middle of the anus.

B. Gums

Open your mouth and stretch your lips tautly over your teeth. Use three fingertips (index, middle, and ring fingers) to tap the skin around the upper and lower gums. Hit around until you feel warmth in the area.

C. Gums and Tongue

Massage your upper and lower gums with your tongue. Then suck in some saliva, press your tongue tightly against the roof of your mouth, and try to exercise the tongue. When you strengthen your tongue, you are strengthening your heart. Press around. Press the tongue to the roof of your mouth, tighten your neck muscles and swallow the saliva. This lubricates the digestive glands and organs.

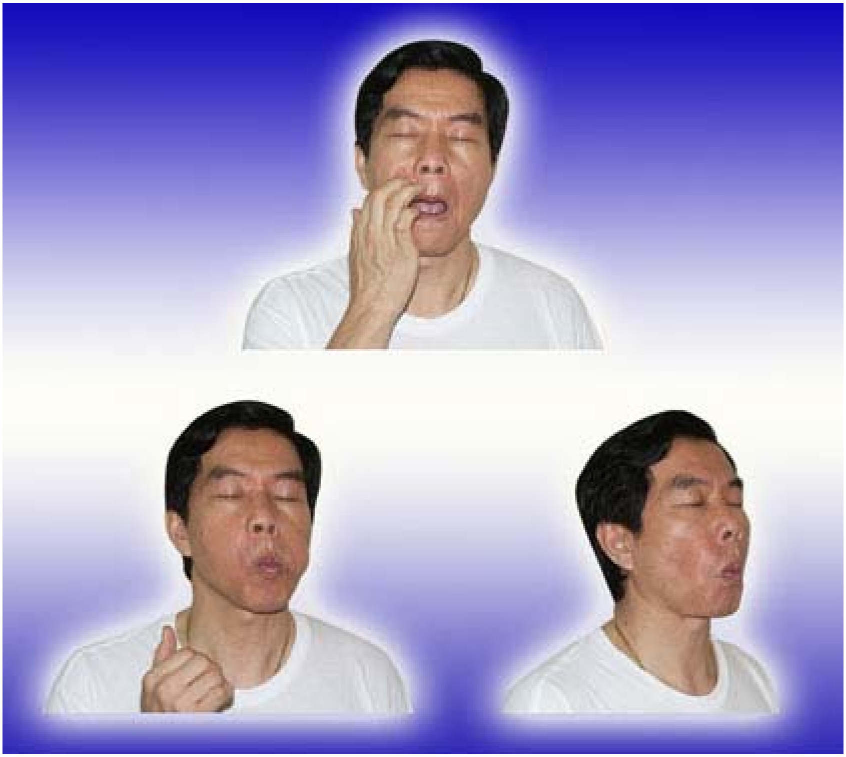

Fig. 32 *Hitting the Gums Exercise.Using the Tongue, Massage the Upper and Lower Gums.*

D. Tongue

In a sitting position place the hands on the knees, palms down. Exhale and straighten the arms, spreading the fingers apart and keeping the hands on the knees. Open the mouth as wide as possible and thrust the tongue out and down, focused on the throat. With the tongue out as far possible, gaze at the tip of the nose. The whole body should be tense. Hold the breath for as long as you feel comfortable. Relax with inhalation and regulate the breath. This will help to strengthen the throat, the tongue and the power of speech. These exercises will help to improve foul breath and to clarify speech.

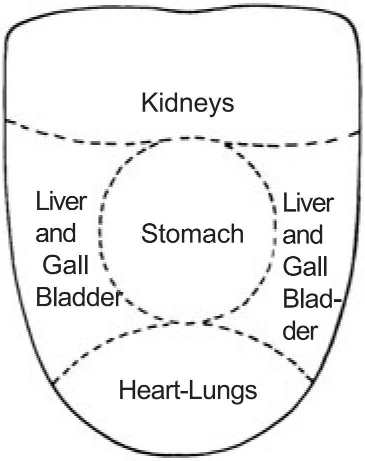
Kidneys
Liver and Gall Bladder
Stomach
Liver and Gall Blad-der
Heart-Lungs

Fig. 33 *Diagram of Tongue Parts and Corresponding Organs*

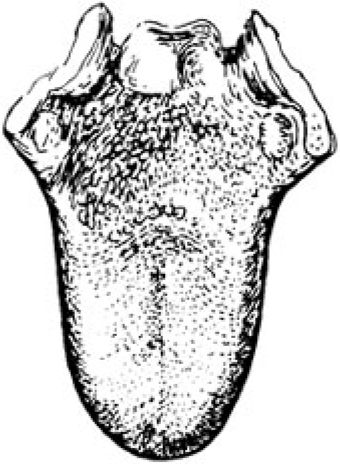

Fig. 34 *Tongue*

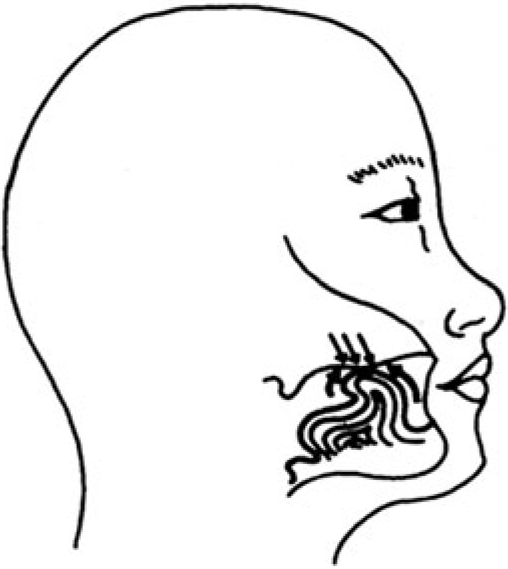

Fig. 35 *Press Tongue to the Roof of the Mouth*

Inhale, then exhale as you press your tongue out and down as far as you can. Follow by pulling the tongue in and curling it. Press your tongue to the roof of your mouth as hard as you can, contracting the middle of the anus and the esophagus to help the tongue. With more practice you will know how to use the inside force, the force from the organs, to press your tongue up. Even though the tongue has no bones to exert force, you will still be able to exercise the tongue well.

E. Teeth Clenching

Relax your lips. Click the teeth together lightly and then clench them hard, as you inhale and pull up the middle of the anus. Do this six to nine times. Move your tongue and mouth to create a lot of saliva. The technique of swallowing the saliva is to put the tongue up to the palate and swallow quickly with a hard gulp, sending the saliva down the esophagus to your stomach.

F. Energy to Teeth

Close your mouth and let your teeth touch lightly. Direct the energy to your teeth. Gradually feel the electrical flow of energy there.

Fig. 36 *Click the Teeth Together Lightly*

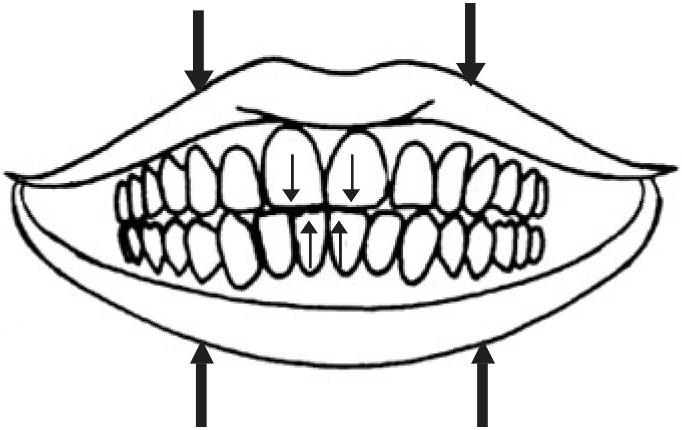

Fig. 37 Clench the Teeth Together Hard

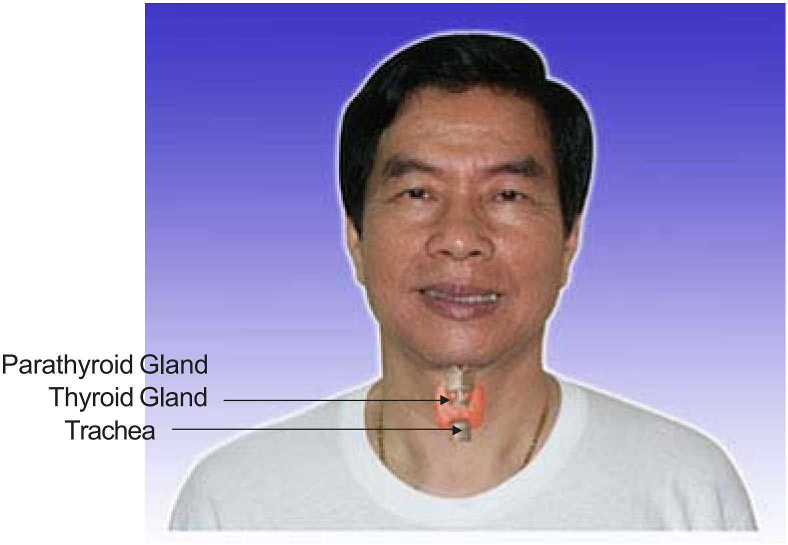
Parathyroid Gland
Thyroid Gland
Trachea

Fig. 38 Throat Glands

Neck

A. Thyroid and Parathyroid: Site of Courage, the Power of Speech

The neck carries the busiest traffic in the body and is the seat of the thyroid and parathyroid glands. When you massage here, you increase the body's metabolism. Neck tension can also be caused by emotional imbalance. When we are tense and nervous, we are responding to negative emotions, such as anger, fear and sadness. The neck is similar to a traffic bottleneck. All of the signals, as well as the emotions, have to pass through it. When under stress and under emotional strain, the neck starts to accumulate and jam the tension. Unconsciously, the neck muscles tighten, attempting to block out pain. Keeping the neck soft will help Chi flow to the higher center that is located in the brain, keeping the mind and body in harmony together.

Tension in the neck can make you less courageous. When the neck is tense, it will block self expression in the throat. With proper flow of Chi energy, we can express ourselves appropriately at the proper time, place, and in a proper way.

The neck is the passageway of many meridians and is the channel of the Chi energy of the organs. In the middle is the Governor meridian. On the sides are the bladder meridian, the triple warmer meridian and the large intestine meridian. The emotions passing through the meridians of the neck may tense and jam up there.

Emotion:	Organ/Associate Organ:
Anger	Liver/Gall Bladder
Fear	Bladder/Kidneys
Grief	Large Intestine/Lungs
Hastiness	Heart/Small Intestine/Triple Warmer
Worry	Spleen/Stomach/Pancreas

B. Bring Chi Energy to the Hands

Do the procedure for bringing energy to the hands and contract the front of the anus.

C. Whole Neck

Spread your thumbs apart from your other fingers. Alternating hands, rapidly wipe the neck from the chin to the base nine to thirty-six times.

D. Middle Neck

Alternating hands, use the middle three fingers to rapidly wipe down the middle of the neck from the chin to the base nine to thirty-six times. The thyroid and parathyroid glands are in the front section of the neck. Use your thumb and the three other fingers to massage these glands. Find the painful points and massage them until you feel them open. Massaging this area will help to increase metabolism and the power of speaking.

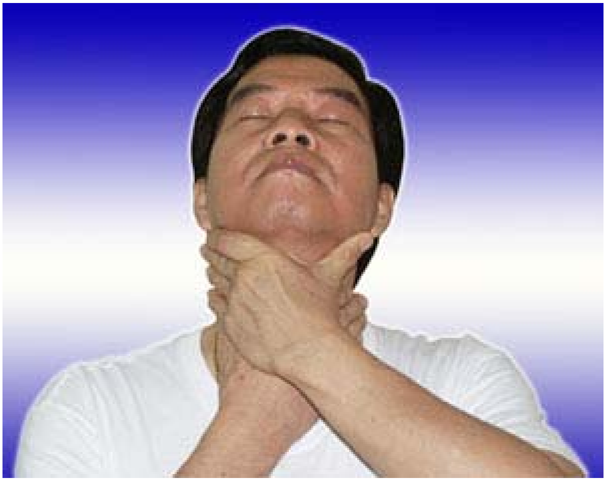

Fig. 39 *Wipe the Neck from the Chin to the Base*

E. Turtle Neck

Sink your chin down, then out and up. Feel your spine press down and then expand. This will help loosen the vertebrae and discs of your neck.

F. Crane Neck

Move your chin forward, circling out, then down, then up, and out again. Feel your spine expand and then contract.

G. Massage the Neck

Massage the points along the back of the neck and on the back along the cervical vertebrae. Start from the shoulders and go up to the base of the skull. Use your fist to hit along the neck. Find any painful spots and any tense spots and massage until they are released. This will greatly help to release the tension of the neck and help to detoxify the toxic accumulations in the neck area, the causes of many headaches.

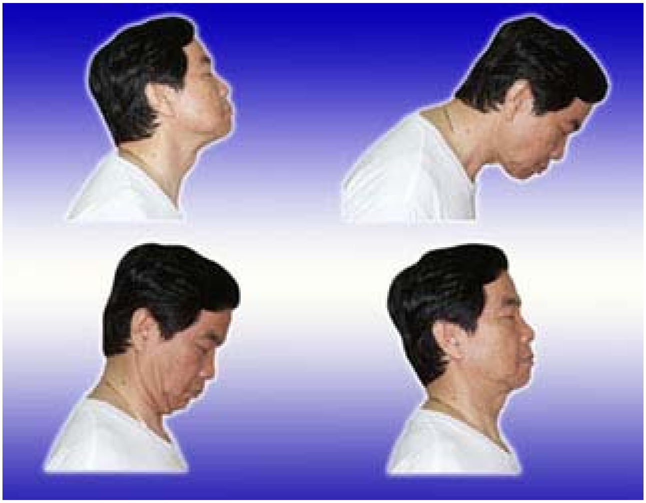

Fig. 40 *Turtle Neck*

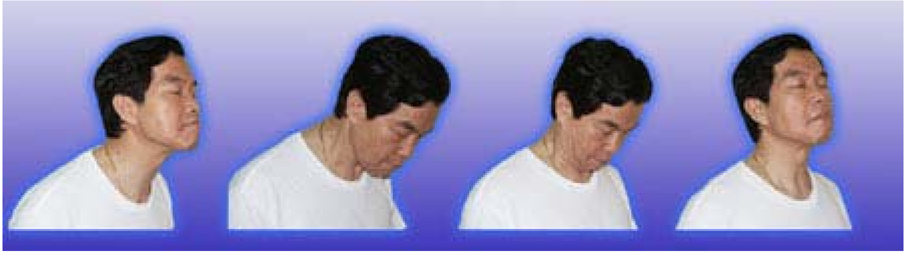

***Fig. 41** Crane Neck*

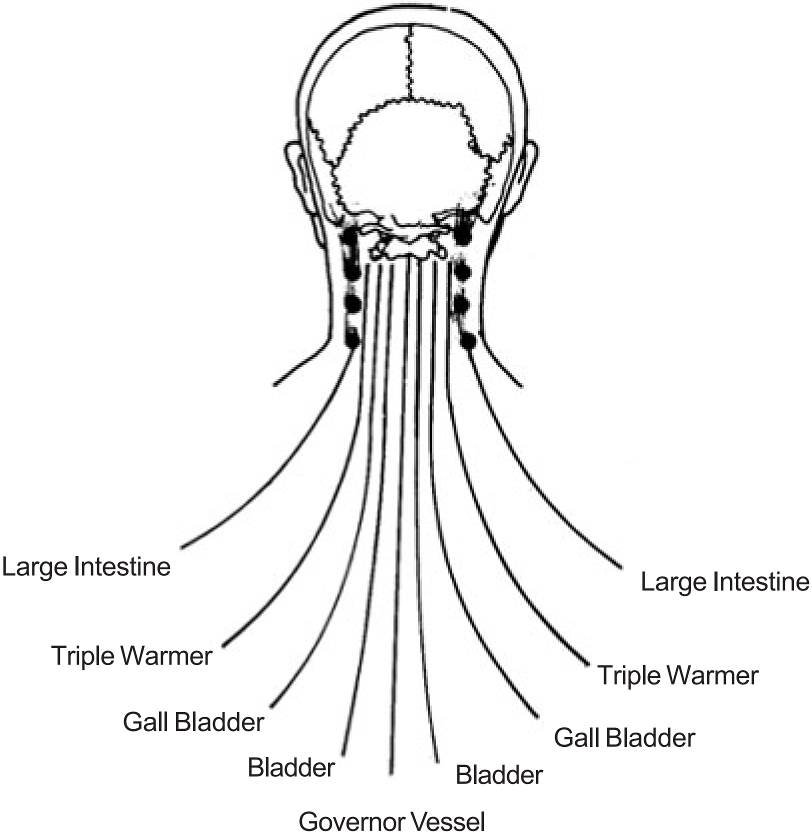
Large Intestine
Large Intestine
Triple Warmer
Triple Warmer
Gall Bladder
Gall Bladder
Bladder
Bladder
Governor Vessel

***Fig. 42** Massage the Points Along the Neck*

Shoulders

Many people feel tense and worried, and their shoulders are tight and held up. The way to release the tension is by pulling up your shoulders to press against the neck, tightening the muscles of the neck and shoulders. Hold for a while, exhale deeply, and let them drop down, pulled by gravity like a sack of potatoes. Feel the burden, worry, and stress drop down to the feet and out to the ground. Feel yourself grounded. Do this three to nine times, and the tension and worry will go away.

Relax your shoulders and chest, exhale and release more, until you feel the tenseness gone.

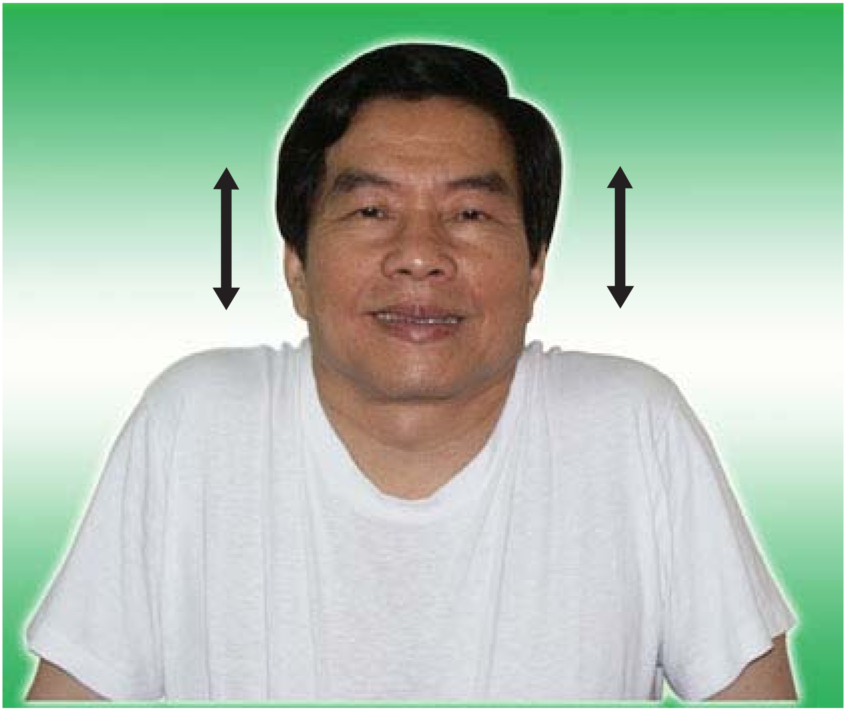

***Fig. 43** Dropping the Shoulders will Help to Release Tension and Worry.*

Detoxifying Organs and Glands

Lightly slapping and tapping over the organs and glands help to shock the toxic sediment and increase the circulation and Chi flow to these areas. Our practitioners claim they are able to heal themselves from many chronic illnesses which are very hard to heal by conventional medicine.

I. Thymus Gland

The thymus gland controls the immune system and is related to longevity. Normally the thymus gland atrophies after childhood. In the higher levels of Taoist practice, the thymus gland can be regrown. This helps maintain health and vitality and supports greater spirituality. Thumbing the thymus gland can help increase the activity and release more hormones.

A. Bring energy to your hands by the usual procedure, contracting the front of the anus and bring the Chi toward the thymus.

B. Make a fist, inhale and thump down the middle of the upper chest from the collar bone to the nipples six to nine times. Do not talk while you are doing this or you might harm yourself.

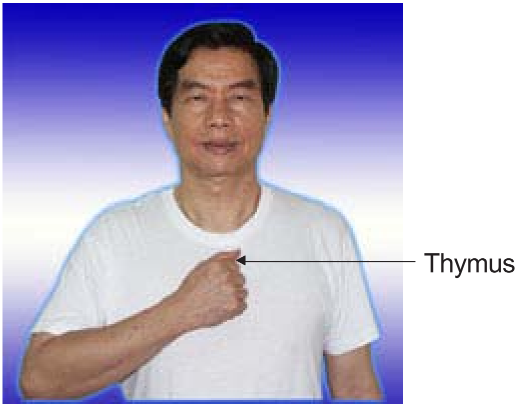

Fig. 44 *Hitting the Thymus Gland*

II. Heart

Lightly slapping an organ stimulates the release of toxins, which allows the organ to rebuild and repair itself. In doing these exercises, be aware that the slapping or tapping should be adjusted by you so as not to use excessive force which may be harmful.

A. Do the energy to hands procedure, contracting the left side of the anus and bringing Chi toward the heart.

B. Slap your heart lightly with your palm six to nine times. Don't speak.

III. Lungs

A. Bring energy to your hands, contracting the right side of the anus and bringing Chi to the lungs.

B. Using your palm, slap up and down your right lung, hitting only as hard as is comfortable. Do not talk. Contract the left side of your anus and slap your left lung. This can help to clean out the mucous and to clean out the lungs.

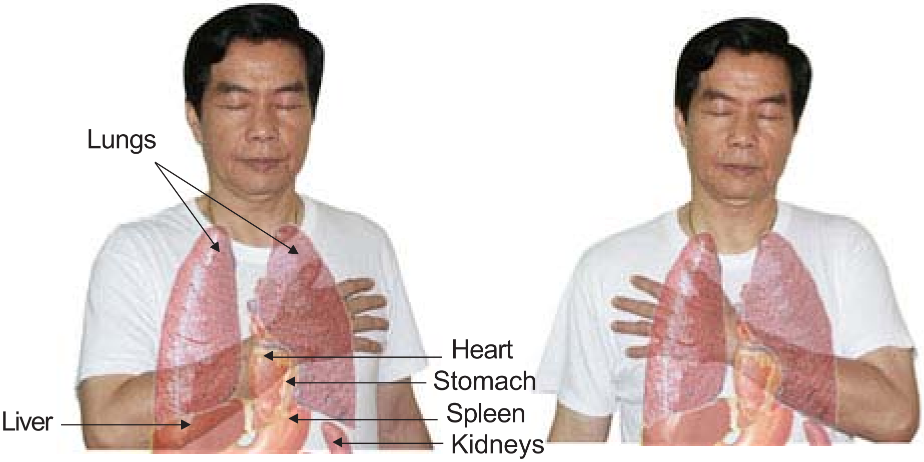

Fig. 45 *Slap at Heart, Lungs and Liver Areas*

IV. Liver

A. Bring energy to the hands, contracting the right side of the anus and bringing Chi to the liver.

B. Using your palm, slap below the rib cage on the right side. Don’t speak. This can help to detoxify the liver.

V. Stomach, Spleen, Pancreas

A. Bring energy to your hands, contracting the middle of the anus.

B. Contract the anus at the left side, and slap at the spleen, pancreas and stomach. Place one palm on top of the other and rub below the rib cage, from center to left, then left to center.

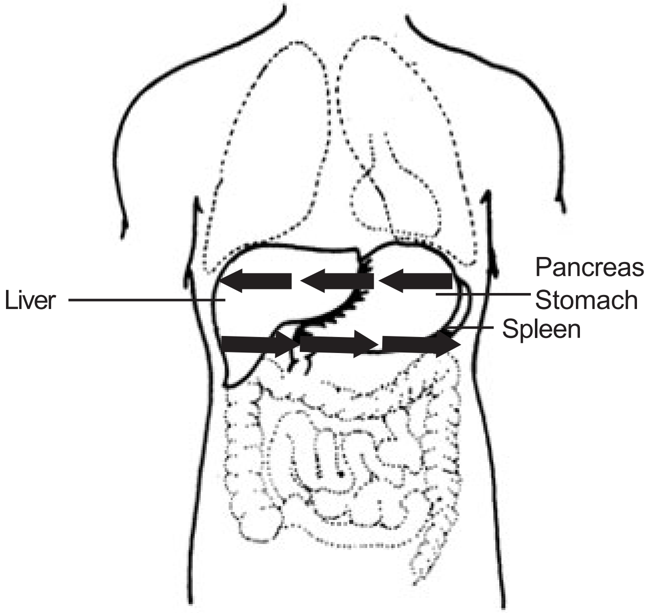

Fig. 46 *Rub the Stomach, Spleen and Pancreas*

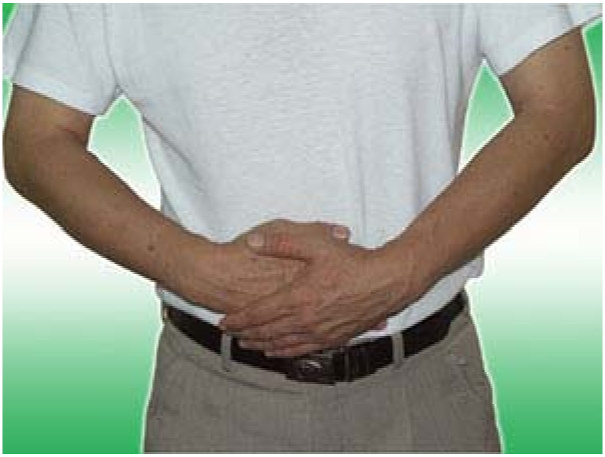

Fig. 47 *Rub Back and Forth over the Liver, Stomach and Spleen*

VI. Large and Small Intestines

A.Bring energy to the hands, contracting the entire anus.

B. Small intestine: With palms together, rub a small circle around your navel, first clockwise, then counterclockwise.

The small intestine is one of the longest tubes in the digestive system. A careless diet, too much hot food or dairy products, or too little fibrous food will create mucous that will stick to the walls of the intestine, block the absorption of nutrients and slow down digestion. Once mucous accumulates, it is like a snow ball that will get bigger and bigger, eventually becoming a lump which slows down the traffic of the digestive system.

C. Large intestine: Place one palm on top of the other and rub your abdomen in a large circle. Start on the lower right side and rub up and around in a clockwise direction. This will move the energy in the intestine and relieve constipation. For diarrhea, rub counterclockwise. If you have normal elimination, rub in both directions. These exercises increase the absorption and dissolve the accumulations that stick to the large intestine's wall.

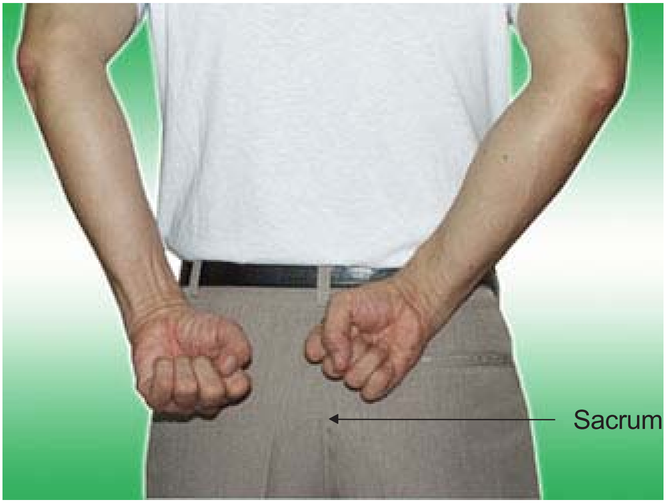

Fig. 48 *Hitting the Sacrum Can Help in Strengthening the Sciatic Nerves*

VII. Kidneys

The kidneys act as helping to filter out waste material from the blood. If there is too much waste in the system, the kidneys cannot filter it all. The waste will tend to collect in the ducts and tubules of the kidneys, impairing their health. By hitting the kidneys' area, we shake out the harmful sediment and help prevent kidney malfunction.

A. Bring energy to the hands, contracting the left and right sides of the anus.

B. Locate the kidneys just above the lowest, or floating, rib in the back on either side of the spine. Make a fist and hit the kidneys with the back of the fist between the wrist and knuckles. Alternate hands and hit only as hard as is comfortable. This will help to shake loose the sediment, crystals, and uric acid that get caught in the kidneys. This will also strengthen the kidneys and relieve back pain.

C. Rub your hands together to warm them. Then rub your palms up and down over the kidneys until they feel warm.

VIII. Sacrum

In the Taoist system the sacrum is regarded as extremely important. It is a pump which helps to bring spinal fluid and energy (Chi) to the brain. It is also the junction where the sexual organs, rectum, and legs meet. Sciatic pain, which shoots down the legs, originates in the sacrum; therefore, strengthening it will release this intense pain.

A. Bring energy to the hands, contracting the back of the anus to the sacrum.

B. Make a fist and use your knuckles alternately to hit both sides of the sacrum. First hit in the area of the eight sacral holes, and then the hiatus, the depression at the bottom of the sacrum.

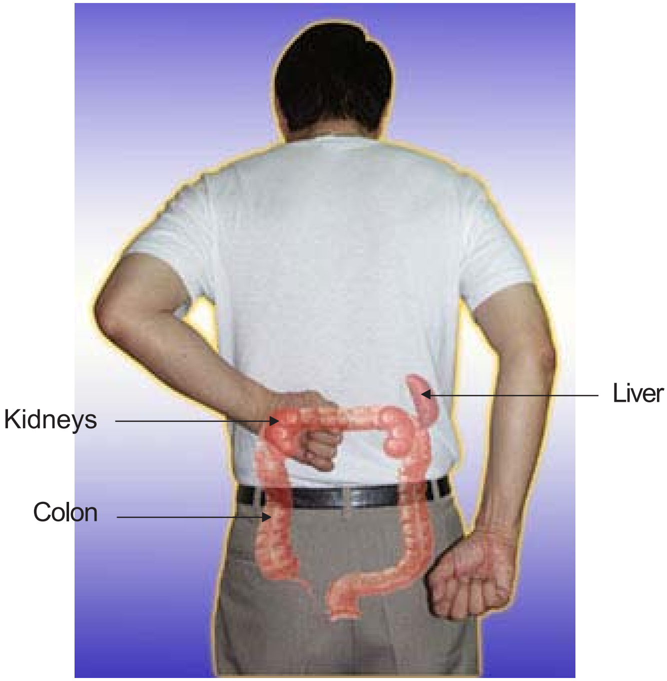

Fig. 49 *Hitting the Kidneys will help to Shake Out Sediment*

Knees and Feet

I. Knees are Toxin Collectors

Toxins tend to collect in the lower limbs because of the slowdown of the circulation due to gravity. The most common places are the back parts of the knees. Slapping at these places will break down the toxins. The body will then eliminate the toxins out of the body by urine, bowel movements and sweat.

A. Bring Chi Energy to the Hands

Bring energy to the hands; do no contractions.

B. Behind the Knees

Prop your leg up on a chair or low table so the knee is straight. Then slap smartly behind the knee nine to eighteen times. Although it hurts, it is extremely beneficial in releasing toxins which accumulate there. This release may be indicated by the appearance of a purple dot. Use discretion as to how hard you slap, since it can be overdone. Repeat this exercise on the other knee.

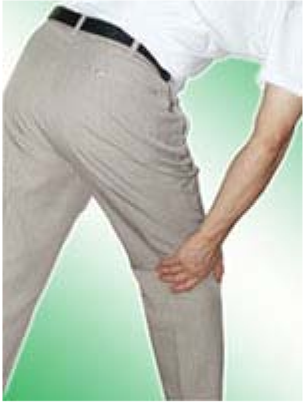

Fig.50 *Slapping Smartly behind the Knee Helps to Release Toxins*

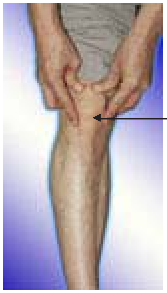
Move Knee Cap Up and Down

Fig. 51 *Massaging the Knee Cap*

C. Knee Cap Massage

Massage the knee cap until it is warm, There is very little blood flowing to it and it tends to be quite vulnerable. This practice strengthens it. Massage the other knee cap.

D. Move the Knee Caps

Relax the knee caps, then move them up and down to the left and right and around both clockwise and counterclockwise.

E. Massage the Knees

Falling down is often caused by weak knees. Massaging the knees will improve your stability and flexibility.

II. Feet: Roots of the Body

Strong feet and tendons increase your stability by connecting you to the healing energy of the earth. Feet are the reflexes of the whole body's organs, glands and limbs. (They are like remote controls. Massaging will help to stimulate the organs and glands and increase the circulation.

A. Bring Chi Energy to the Hands

Bring energy to the hands; do no contractions.

B. Massage the Feet

Take off your shoes and stockings and massage the tops and bottoms of each foot with your thumbs and fingers. Be sure to massage the kidney point, the sore spot in the center between the ball of each foot and the adjoining pad. If you are in a hurry, massage the whole of each foot once by rubbing the sole of the foot vigorously and carefully across the top of the other foot, going from the heel to the arch to the toes. The soles of the feet have energy meridians to the entire body. Massage the feet, and when you find painful points massage them until the pain goes away. This will help to clear any blockage of Chi channel flow

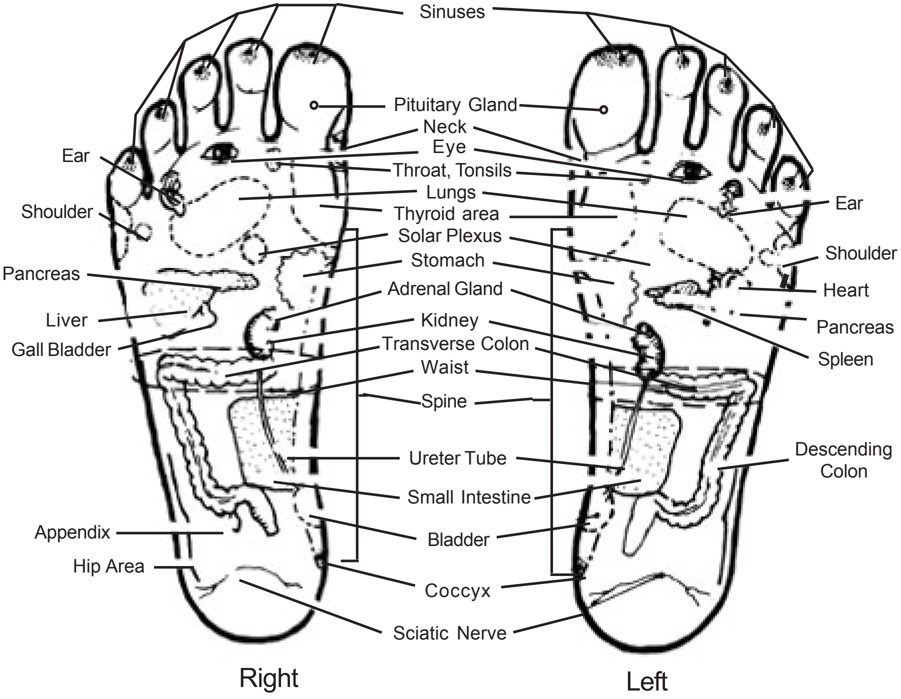
Sinuses
Pituitary Gland
Neck
Eye
Throat, Tonsils
Lungs
Thyroid area
Solar Plexus
Stomach
Adrenal Gland
Kidney
Transverse Colon
Waist
Spine
Ureter Tube
Small Intestine
Bladder
Coccyx
Sciatic Nerve
Ear
Shoulder
Pancreas
Liver
Gall Bladder
Appendix
Hip Area
Ear
Shoulder
Heart
Pancreas
Spleen
Descending Colon
Right
Left

***Fig. 52** Feet are the Reflexes of the Whole Body's Organs, Glands and Limbs*

C. Spread Out the Toes

Spread out and separate all the toes, especially the little toes, and then release. Repeat six to nine times. This is especially good for the tendons of the feet.

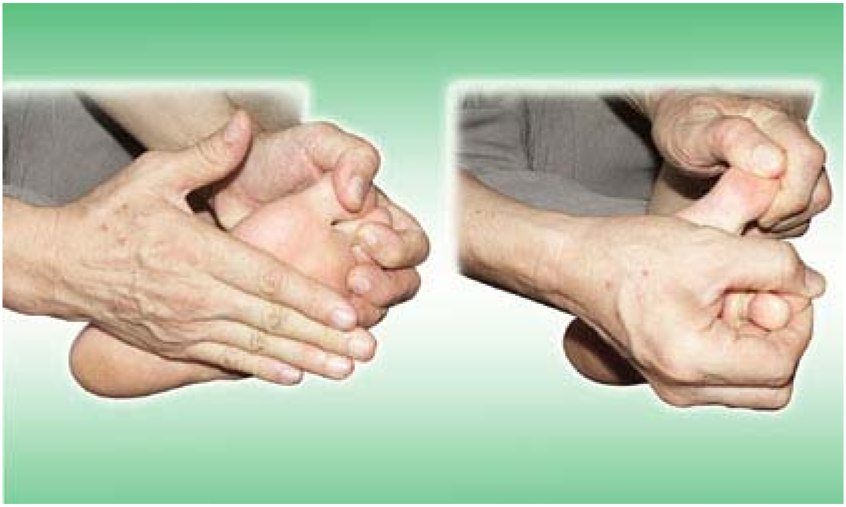

Fig. 53 *Separate all the Toes by Spreading them out.*

D. Big and Second Toes

Rub the big toes and second toes together rapidly. This a good exercise to do at odd moments during the day.

E. Rub Feet Together

Keep the feet warm by rubbing them together. This will help to stimulate all the body's organs.

For further information about Universal Tao Centers, courses, or other resources, contact:

Universal Tao Instructor Associations

North America Instructors Association (NAIA)

c/o Loretta Robb, Treasurer
13 Plymouth Drive, Newark, Delaware 19713 USA
Tel: (1)(888) 444-7426 (Toll Free) or (1)(212) 330-7876
Email: info@taoinstructors.org   Website: http://taoinstructors.org

European Instructors Association (EIA)

c/o Zentrum Waldegg
3823 Wengen, SWITZERLAND
Tel: (41)(33) 8554422 Fax: (41)(33) 8555068
Email:info@waldegg.ch   Website: www.waldegg.ch

To order English books, and for your local Insructors use our Web Sites: www.universal-tao.com and www.taogarden.com

Universal Tao World Fulfillment Center

274 Moo 7, Luang Nua, Doi Saket, Chiang Mai, 50220 Thailand
Tel: (66)(53) 495-596 & 865-034 Fax: (66)(53) 495-852
North America Fax (1)(212) 504-8116   Europe Fax (31)(20) 524-1374
Email: universaltao@universal-tao.com **or** info@tao-garden.com

The Universal Tao is not and cannot be responsible for the consequences of any practice or misuse of the information in this book. If the reader undertakes any exercise without strictly following the instructions, notes, and warnings, the responsibility must lie solely with the reader.

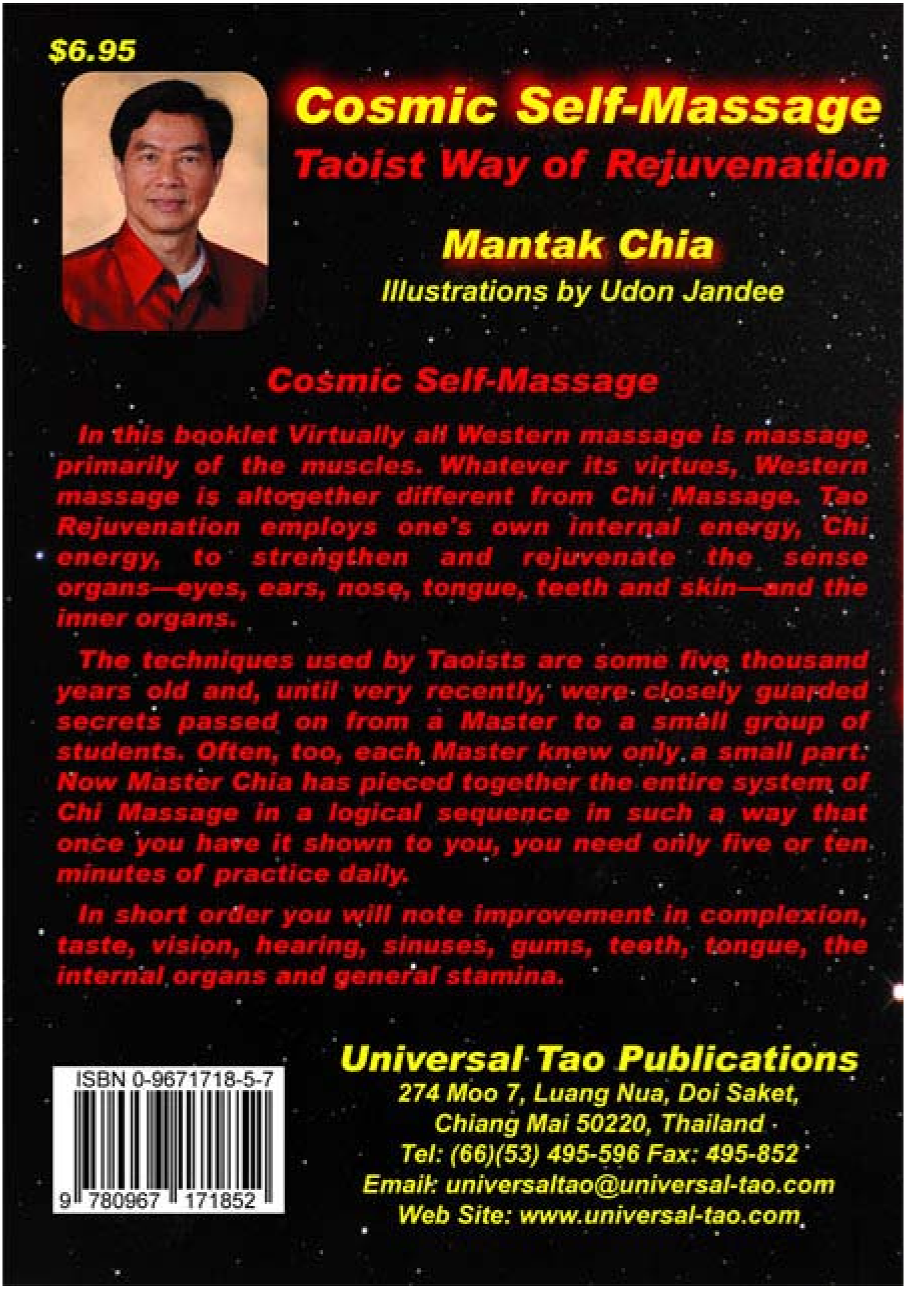

$6.95

Cosmic Self-Massage

Taoist Way of Rejuvenation

Mantak Chia

Illustrations by Udon Jandee

Cosmic Self-Massage

In this booklet Virtually all Western massage is massage primarily of the muscles. Whatever its virtues, Western massage is altogether different from Chi Massage. Tao Rejuvenation employs one's own internal energy, Chi energy, to strengthen and rejuvenate the sense organs—eyes, ears, nose, tongue, teeth and skin—and the inner organs.

The techniques used by Taoists are some five thousand years old and, until very recently, were closely guarded secrets passed on from a Master to a small group of students. Often, too, each Master knew only a small part. Now Master Chia has pieced together the entire system of Chi Massage in a logical sequence in such a way that once you have it shown to you, you need only five or ten minutes of practice daily.

In short order you will note improvement in complexion, taste, vision, hearing, sinuses, gums, teeth, tongue, the internal organs and general stamina.

ISBN 0-9671718-5-7
9 780967 171852

Universal Tao Publications
274 Moo 7, Luang Nua, Doi Saket,
Chiang Mai 50220, Thailand
Tel: (66)(53) 495-596 Fax: 495-852
Email: universaltao@universal-tao.com
Web Site: www.universal-tao.com

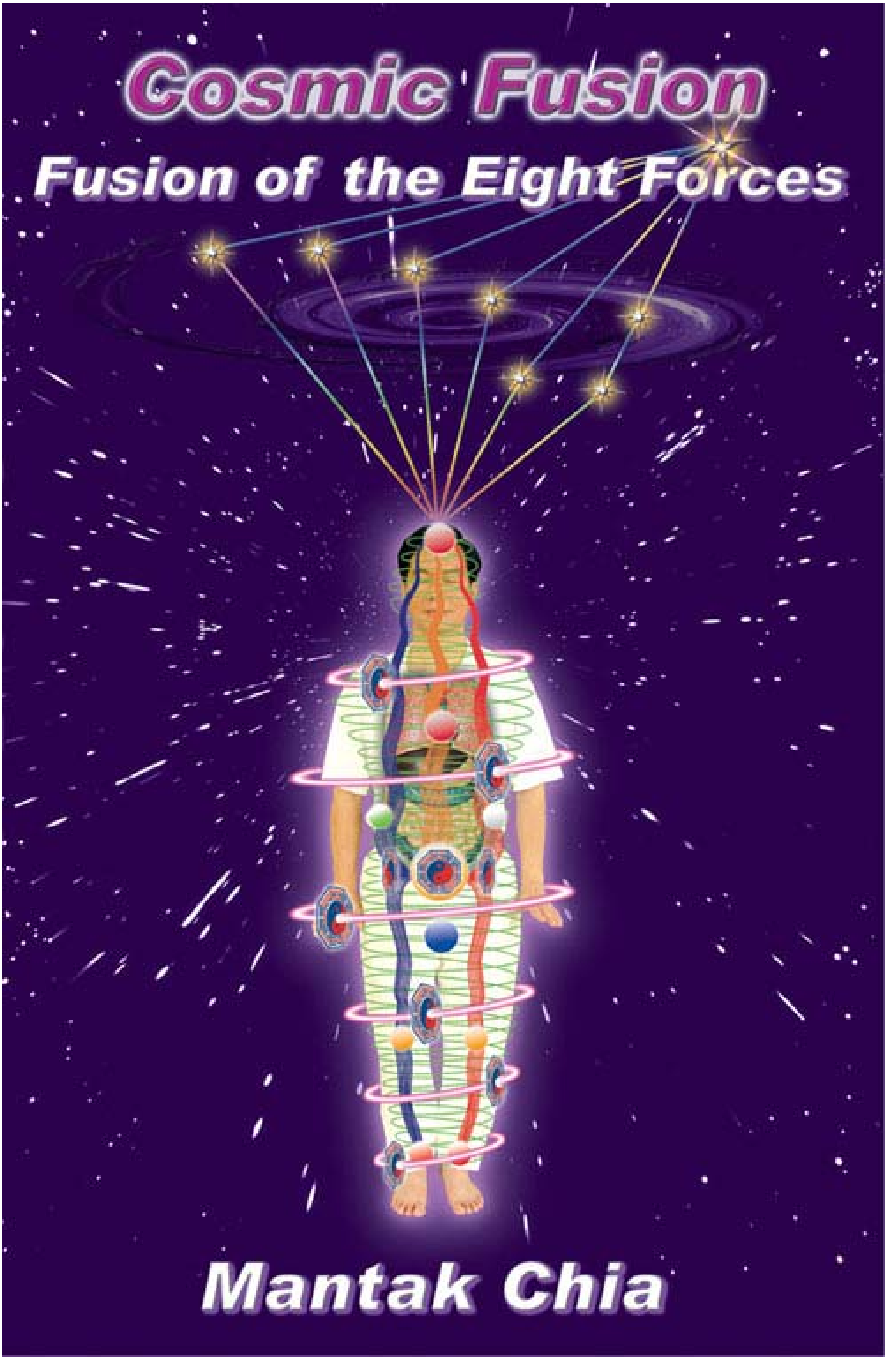

Cosmic Fusion

Fusion of the Eight Forces

Mantak Chia

Cosmic Fusion

Fusion of the Eight Forces

Mantak Chia

Edited by:

Lee J. Holden Jr.

Copy Editor: Lee Holden, Jr.

Copy Writer: Rene J. Navarro

Editorial Assistance: Colin Campbell

Photographer: Saysunee Yongyod

Computer Graphics: Raruen Keawpadung

Design and Production: Saniem Chaisarn

Illustrations: Udon Jandee

Final Editing: Jean Chilton

Project Manager: W.U. Wei

© *North Star Trust*

First published in 2002 by:

Universal Tao Publications
274/1 Moo 7, Luang Nua,
Doi Saket, Chiang Mai, 50220 Thailand
Tel: (66)(53) 495-596 Fax: 495-853
Email: universaltao@universal-tao.com
Web Site: www.universal-tao.com

ISBN: 974-85391-9-9

Manufactured in Thailand
Second Printing, 2002

All rights reserved. No part of this book may be used or reproduced in any manner whatsoever without the express written permission from the author except for brief quotations embodied in critical articles and reviews.

Contents

Contents iv
About the Author viii
Acknowledgments xii
Words of Caution xiii

Chapter 1

Cosmic Fusion and the Tao 14
  The Tao 14
  Universal Tao System 16
    Healing the Physical Body: Becoming Like a Child to Return to the Original Source 17
    Stopping Energy Leakage through Conservation and Recycling.. 19
    Abdominal Brain 20
    Transforming Negative Energy into Virtues: Opening the Heart 21
    Managing the Life-Force 22
    Cosmic Nutrition 23
    Developing the Energy Body: Our Vehicle to Travel in Inner and Outer Space 23
    Recycling our Negative Emotions.................................. 25
    Forming the Spirit Body: Planting the Seed of Immortality 26
    Cultivating the Yang Stage of the Immortal Body: Marrying the Light - Greater Kan and Li 27
    Greatest Enlightenment of Kan and Li 28
    Sealing of the Five Senses, Congress of the Heaven and Earth and Reunion of Heaven and Man............................ 29

Chapter 2

Taoist Cosmology 30
  The Void: Our Original Source 31
  Yin and Yang 31
  Five Major Forces of the Universe.. 32
  Eight Forces 34
  Return to the Wu Chi... 35
  Returning to the Wu Chi: First Three Steps 36

Chapter 3

Goals of the Fusion Practice 38
Inner Universe reflects the Outer 42
Overview of the Cosmic Fusion:
Universal Tao divides Fusion into Three Parts.................. 43
First Part 43
Second Part 44
Third Part 46

Chapter 4

Energetic Preparations for Meditation Practice 51
Preparing for the Fusion Meditations 51
Prepare the Environment 51
Conserve Energy and Purify your Body with an Appropriate Diet 52
Loosen, Stretch and Warm Up the Body before Meditating 52
Wear Loose and Comfortable Clothing 52
Establish a Stable Sitting Position 53

Chapter 5

Beginning Fusion Element Meditation 59
Pakua.. 59
Chanting 60
Formula One: Forming the Pakuas 63
Creating the Front Pakua 64
Practice 65
Activate the Facial Pakua by enhancing it on the Forehead 88
Facial Pakua 89
Start the Practice of the Facial Pakua with the First Four Symbols first 92
Creating the Back Pakua on the Head 99
Formula Two: Transforming the Negative Emotions of Each Organ into Pure Life-Force Energy 108
Practice 110

Chapter 6
Compassion Fire
Meditation Practice of the Creation Cycle 116
One Lasting Energy is the Combination of
Compassion Energy and Orgasmic Energy 119
Practice of Creating the Compasstion Fire
and Multi-Orgasm 121
Multi-Orgasm Energy Smile to the Sexual Organs 127
Practice Compassion though the Act of Forgiveness..... 130
Oneness with Universe (Oneness with God) 131
Fuse Inner Positive Energy with Universal Energy
through Turning the Dial 131
Forming the Collection Points 131
The Concept of Fusion Inner Position Energy with
Universal Energy 133
Activate the Three Fires 134
Creation Cycle Foundation Practice 141
Begin Cosmic Fusion: Creation Cycle............................ 144
Shoot the Pearl Above the Crown 152
Summary 144

Chapter 7
Opening the Thrusting Channels 155
Theory of the Thrusting Channels 155
Procedure for Opening the Thrusting Channels 156
Seated Position 157
Anal Muscle 158
Saliva 159
Men & Women Practice: Open Channels Differently...... 161
Practice of Opening the Thrusting Channels 162
Additional Exercise for Opening the Thrusting Channels.... 179
Summary: Opening the Thrusting Channels...................... 181

Chapter 8
Belt Channels 183
Theory of the Belt Channel 183
Procedure of Forming the Belt Channel............................. 183
Practice of Forming the Belt Channel 185
Men & Women Practice: Open Channels Differently...... 193
Forming a Vehicle 194

Summary: Procedure of Forming the Belt Channel............ 198
Energy Body 202
Spirit Body 204

Chapter 9

Fusion as a Bridge 206
Fusion as a Journey of Self-Transformation 206
Unity of Ying and Yang 208
Turning Inward 209
From Balance through Resonance to Love 210
Enlightenment of the Body and New Physics 211
Return to Cosmology 213
Religious Nature of Taoist Practice 214
Wisdom of Making Friends with the Devil 215
Unity of Light and Dark 217
Life as an Invitation to Learning 219
Smiling as the Key Taoist Practice 220
Unity of Fullness and Emptiness 220
Attachment and Detachment 222
Going with the Flow 223
Secret of Breathing and Relaxation 225
Science of Feeling and the Power of Intuition 226
Thinking with the Heart 227
Feminine Underground Current 229
Restoring the Balance 230
World as a Manifestation and Reflection of the Universe ... 231
Unity of What is Above and Below 234
Healing Wounds 237
Practice of Humor and the Humor of Practice 239
Learning to tend our Garden 240
Raising the Quality of Our Energies 241
Marriage of the Light from Within and from Above 241
Questions and Answers
about the Cosmic Fusion Practice 244

Biobliography...254

About the Author

Mantak Chia

Master Mantak Chia

Master Mantak Chia is the creator of the Universal Tao System and is the director of the Universal Tao Center and Tao Garden Health Resort and Training Center in the beautiful northern countryside of Thailand. Since childhood he has been studying the Taoist approach to life. His mastery of this ancient knowledge, enhanced by his study of other disciplines, has resulted in the development of the Universal Tao System which is now being taught throughout the world.

Mantak Chia was born in Thailand to Chinese parents in 1944. When he was six years old, Buddhist monks taught him how to sit and “still the mind.” While still a grammar school student, he learned traditional Thai boxing. He was then taught Tai Chi Chuan by Master Lu, who soon introduced him to Aikido, Yoga and broader levels of Tai Chi.

Years later, when he was a student in Hong Kong excelling in track and field events, a senior classmate named Cheng Sue-Sue introduced him to his first esoteric teacher and Taoist Master, Master Yi Eng (I Yun). At this point, Master Chia began his studies of the Taoist way of life in earnest. He learned how to circulate energy through the Microcosmic Orbit and, through the practice of Fusion of the Five Elements, how to open the other Six Special Channels. As he studied Inner Alchemy further, he learned the Enlightenment of the Kan and Li, Sealing of the Five Senses, Congress of Heaven and Earth and Reunion of Heaven and Man. It was Master Yi Eng who authorized Master Chia to teach and heal.

When Mantak Chia was in his early twenties he studied with Master Meugi in Singapore, who taught him Kundalini, Taoist Yoga and the Buddha Palm. He was soon able to clear blockages to the flow of energy within his own body. He learned to pass the life force energy through his hands also, so that he could heal Master Meugi’s patients. He then learned Chi Nei Tsang from Dr. Mui Yimwattana in Thailand.

A while later, he studied with Master Cheng Yao-Lun who taught him the Shao-Lin Method of Internal Power. He learned the closely guarded secret of the organs, glands and bone marrow exercise known as Bone Marrow Nei Kung and the exercise known as Strengthening and Renewal of the Tendons. Master Cheng Yao-Lun’s system combined Thai boxing and Kung Fu. Master Chia also studied at this time with Master Pan Yu, whose system combined Taoist, Buddhist and Zen teachings. Master Pan Yu also taught him about the exchange of Yin and Yang power between men and women, and how to develop the Steel Body.

To understand the mechanisms behind healing energy better, Master Chia studied Western anatomy and medical science for two years. While pursuing his studies, he managed the Gestetner Company, a manufacturer of office equipment and became well acquainted with the technology of offset printing and copying machines.

Using his knowledge of Taoism, combined with the other disciplines, Master Chia began teaching the Universal Tao System. He eventually trained other Instructors to communicate this knowledge and he established the Natural Healing Center in Thailand. Five years later, he decided to move to New York, where in 1979, he opened the Universal Tao Center. During his years in America, Master Chia continued his studies in the Wu system of Tai Chi with Edward Yee in New York.

Since then, Master Chia has taught tens of thousands of students throughout the world. He has trained and certified over 2,000 instructors and practitioners from all over the world. Living Tao Centers, Chi Nei Tsang Institutes, Cosmic Healing Forums & Immortal Tao Mountain Sancutaries have opened in many locations in North America, South America, Europe, Asia, Africa and Australia.

In 1994, Master Chia moved back to Thailand, where he had begun construction of Tao Garden and the Universal Tao Training Center fifteen miles outside of Chiang Mai.

Master Chia is a warm, friendly and helpful man who views himself primarily as a teacher. He presents the Universal Tao System in a straightforward and practical manner, while always expanding his knowledge and approach to teaching. He uses a word processor for writing and is totally at ease with the latest computer technology.

Master Chia estimates that it will take thirt-five books to convey the full Universal Tao System. In June, 1990, at a dinner in San Francisco, Master Chia was honored by the International Congress of Chinese Medicine and Qi Gong (Chi Kung), who named him the Qi gong Master of the Year. He is the first recipient of this annual award.

In December, 2000, the Tao Garden Health Resort and Universal Tao Training Center was completed with two Meditation Halls, two open air Simple Chi Kung Pavilions, indoor Tai Chi, Tao Tao Yin and Chi Nei Tsang Hall, Tai Chi Natural Swimming Pool, Pakua Communications Center with a complete Taoist Library, Internal World Class Weight Lifting Hall and complete 8 Court Recreational Facilities.

In February, 2002, the Immortal Tao practices were held at Tao Garden for the first time using Darkness Technology, creating a complete environment for the higher level Taoist practices.

Master Mantak Chia has previously written and published these twenty-two Universal Tao books:

Awaken Healing Energy of the Tao - 1983
Taoist Secrets of Love: Cultivating Male Sexual Energy co-authored with Michael Winn - 1984.
Taoist Ways to Transform Stress into Vitality -1985
Chi Self-Massage: the Tao of Rejuvenation - 1986
Iron Shirt Chi Kung I - 1986
Healing Love Through the Tao: Cultivating Female Sexual Energy - 1986
Bone Marrow Nei Kung - 1989
Fusion of the Five Elements I - 1990
Chi Nei Tsang: Internal Organ Chi Massage - 1990
Awaken Healing Light of the Tao - 1993
The Inner Structure of Tai Chi co-authored with Juan Li - 1996
Multi-Orgasmic Man co-authored with Douglas Abrams 1996 - published by Harper Collins
Tao Yin - 1999
Chi Nei Tsang II - 2000
Multi-Orgasmic Couple co-authored with Douglas Abrams 2000 - published by Harper Collins
Cosmic Healing I - 2001
Cosmic Healing II co-authored with Dirk Oellibrandt - 2001
Door of All Wonders co-authored with Tao Haung - 2001
Sexual Reflexology co-authored with W. U. Wei - 2002
Elixir Chi Kung - 2002
Tan Tien Chi Kung - 2002
Cosmic Fusion - 2002

Many of the books above are available in the following foreign languages:

Arabic, Bulgarian, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Lithuanian, Malaysian, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Slovenian, Spanish, & ***Turkish*** editions are available from the Foreign Publishers listed in the Universal Tao Center Overview in the back of this book.

Acknowledgments

The Universal Tao Publications staff involved in the preparation and production of *Cosmic Fusion: Fusion of the Eight Forces* extend our gratitude to the many generations of Taoist Masters who have passed on their special lineage, in the form of an unbroken oral transmission, over thousands of years. We thank Taoist Master I Yun (Yi Eng) for his openness in transmitting the formulas of Taoist Inner Alchemy.

Thanks to Juan Li for the use of his beautiful and visionary paintings, illustrating Taoist esoteric practices.

We offer our eternal gratitude to our parents and teachers for their many gifts to us. Remembering them brings joy and satisfaction to our continued efforts in presenting the Universal Tao System. For their gifts, we offer our eternal gratitude and love. As always, their contribution has been crucial in presenting the concepts and techniques of the Universal Tao.

We wish to thank the thousands of unknown men and women of the Chinese healing arts who developed many of the methods and ideas presented in this book. We offer our gratitute to Master Lao Kang Wen for sharing his healing techniques.

We wish to thank Lee J. Holden Jr. for his editorial work and writing contributions, as well as his ideas for the cover. We appreciate his research and great labor. We wish to thank Colin Campbell for his editorial contributions on the revised edition of this book, as well as thanking our Senior Instructors, Rene J. Narvarro and Annette Derksen, for their insightful contributions to the revised version. We thank Joost Kuitenbrouwer for his technical editing and clear writing throughout the book.

A special thanks goes to our ***Thai Production Team*** for their cover illustration and book design and layout: Raruen Keawapadung, Computer Graphics; Saysunee Yongyod, Photographer; Udon Jandee, Illustrator; and Saniem Chaisam, Production Designer.

Words of Caution

The practices described in this book have been used successfully for thousands of years by Taoists trained by personal instruction. Readers should not undertake the practice without receiving personal transmission and training from a certified instructor of the Universal Tao, since certain of these practices, if done improperly, may cause injury or result in health problems. This book is intended to supplement individual training by the Universal Tao and to serve as a reference guide for these practices. Anyone who undertakes these practices on the basis of this book alone, does so entirely at his or her own risk.

The meditations, practices and techniques described herein are ***not*** intended to be used as an alternative or substitute for professional medical treatment and care. If any readers are suffering from illnesses based on mental or emotional disorders, an appropriate professional health care practitioner or therapist should be consulted. Such problems should be corrected before you start training.

Neither the Universal Tao nor its staff and instructors can be responsible for the consequences of any practice or misuse of the information contained in this book. If the reader undertakes any exercise without strictly following the instructions, notes and warnings, the responsibility must lie solely with the reader.

This book does not attempt to give any medical diagnosis, treatment, prescription, or remedial recommendation in relation to any human disease, ailment, suffering or physical condition whatsoever.

The Universal Tao is not and cannot be responsible for the consequences of any practice or misuse of the information in this book. If the reader undertakes any exercise without strictly following the instructions, notes, and warnings, the responsibility must lie solely with the reader.

Chapter 1

Cosmic Fusion and the Tao

Cosmic Fusion is the basic Taoist practice of internal alchemy. This is the second book in a three part series. This particular book expands on the concepts and practices introduced in Fusion of the Five Elements. Here in part II, the practice goes to the next level of development.

In the beginning these practices may seem complex and mentally challenging, yet once the formulas are understood and applied, they are quite simple. With practice, the materials become familiar territory that you can traverse easily. Take your time with the formulas. Experiment with what works for you. The formulas are the guidelines to enhance your life. The Cosmic Fusion is a doorway into the vast potential of the self and the Universe. It is a very natural process and the first step into the Taoist alchemical practice.

The Tao 道

The word *Tao* means ***The Way***, the way of nature and the Universe, or the path of a natural reality. It also refers to a way in which we can open our minds to learn more about the world, our spiritual paths and ourselves. Taoism is a practice of body, mind and spirit, not just a philosophy of mind. When we have the true sense of the Tao, of the real knowledge and wisdom, we will be able to make the right decisions in our lives.

Taoism involves many practical disciplines that can restore our lost youth, energy and virtues while awakening our deepest spiritual potentials. Taoists regard these practices as a technology that can help us learn universal truths if we are willing to be receptive.

The ancient masters recognized that these potentials can include the possibility of attaining conscious freedom and existence in the after-death state. Through specific exercises, one can avoid suffering the experience of death by expanding the consciousness beyond the physical body before its demise. This makes it possible to determine one's future existence before leaving this life.

All spiritual paths ultimately lead to the truth. Taoism is both a philosophy and a technology for seeking and finding the truths of the universe, nature, and humanity. Its focus goes beyond one single path or viewpoint. Taoism is not a religion, as it requires no initiations or ceremonies, but it is the outcome of all religions, departing from dogma at the point of truth. It leaves behind all religious beliefs, just like the clothing of past seasons. The Tao is also the goal of science, but it leaves behind all scientific theories as partial and temporal descriptions of the integral truth. Taoism includes all matters of religion and science, yet its breadth goes far beyond the limits of devotion or intellect.

Taoist teachings are like master keys unlocking all doors. They assist people in their lives, as do all religious teachings. Yet the teachings of the Tao transcend religion while retaining the essence of spirituality. They explain and demonstrate the truths of the Universe directly, rather than on the level of emotions, thoughts or beliefs. For this reason, students of the Tao have little cause for skepticism or endless searching.

Philosophy, science and religions all contain some aspects of truth that reflect the Tao. The teachings of the Tao reflect the center of the ultimate truth (ourselves and the Universe) and help us reach it on our own. We can believe in any religion or spiritual path and still benefit from these teachings, because the Tao serves only to promote universal spiritual independence. There are no ultimate masters or gurus in Taoism because we become our own masters, capable of controlling our own destinies and knowing who we really are as we explore the marvelous powers hidden within the Tao of humanity. All the great gods, immortals, sages, saints and holy men and women are our teachers and advisers.

Taoism, which goes back at least seven to eight thousand years, is closely aligned with modern technology and science. Browsing through a bookstore, you will find that many recent books use the concepts of Taoism to explain relationships expounded by the “new” physics. Today, a lot of the world’s intellectual effort is directed toward external technological development.

Thousands of years ago, the Taoists directed their consciousness inwardly, and developed a science of inner alchemy. Chinese written history has a 4700 year history and the language has not changed significantly over the centuries, so even the oldest documents can be read today. A lot of information was accumulated in each dynasty. The Taoist sages, who learned and perfected methods to prepare and leave the body, recorded all the practical details. These, along with philosophical thought, were collected, filling many volumes which constitute the Tao Canon.

These transformative formulas endure, a gift from the original Taoist sages to all who are willing to put them into practice. The Universal Tao is concerned primarily with the practical approaches perfected by the Taoist sages.

Universal Tao System

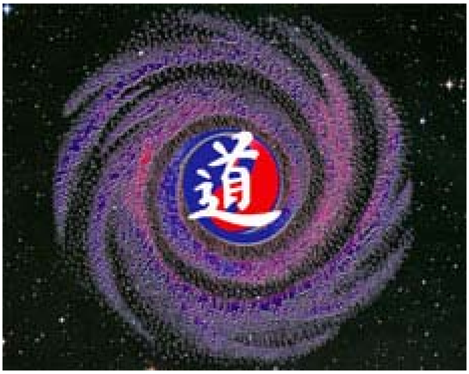

The ancient Taoists saw the importance of working on all three levels of our being: the physical body, the Energy Body and the spirit. All three are important in forming a ladder with which we may climb consciously into the spiritual worlds and, just as important, back into the physical world to be creatively active here. This ladder enables Taoists to learn about the inner worlds and to return to the physical with knowledge and increased energy. An Immortal Body, which is developed in the practice of Internal Alchemy, enables one to establish a constant link between life and the after-death (or pre-birth) state.

Ancient Taoist sages believed we were born to be immortal. We become mortal by draining ourselves of Chi through engaging in excessive sexual activity, indulging in negative emotions, and depending on material sources to supply our life-force. The masters recognized that different levels of immortality can be achieved through internal alchemy, and they devised many practices for this purpose. The ability to transcend even death through the transmutation of one's physicality into the Immortal Spirit Body is the highest goal of Taoism. This level, known as physical immortality, takes the longest to achieve.

Healing the Physical Body: Becoming like a Child to return to the Original Source

Taoist practices teach to conserve the physical energy so that it will no longer scatter and deplete as a result of our worldly interactions. Full spiritual independence requires that we avoid being drained of this energy through the eyes, ears, nose, and mouth or through excessive sex. The novice in the Universal Tao System begins with a wide range of exercises that develop the physical body into an efficient and healthy organism, able to live in the world and yet stay free of the tensions and stresses of daily life. One aspires to return to a childlike state of innocence and vitality, to regain the Original Source that is our birthright. Specific goals of this level are to learn how to heal oneself, how to love oneself, and how to love others.

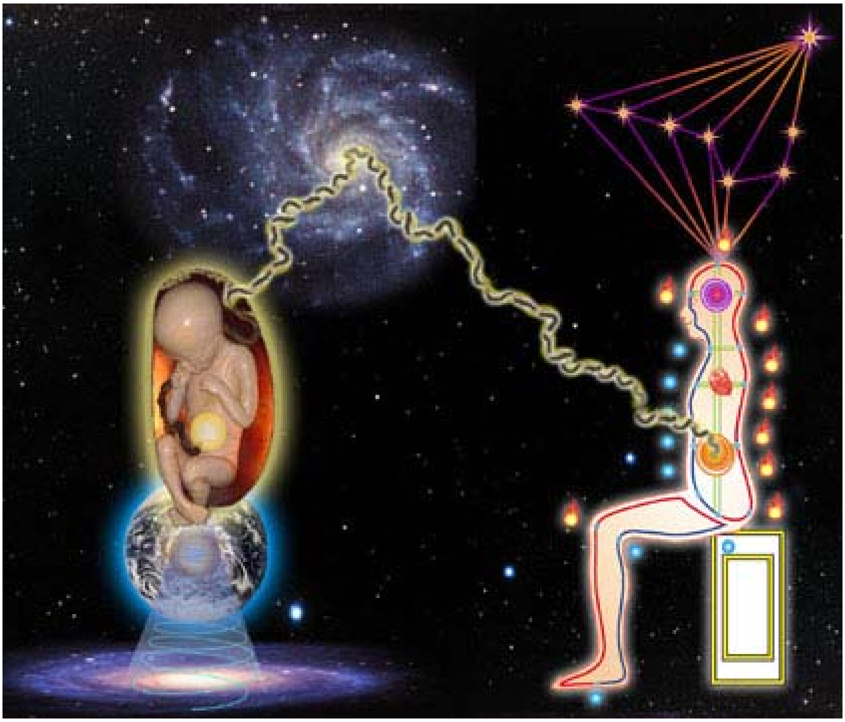

Fig. 1.1 *Returning to the Original Source.*

The first level of practice is to develop a healthy body, which can take months of diligent training. During this process we learn how to condense and conserve our life-force through the *Microcosmic Orbit* meditation, *Healing Love* practices, *Inner Smile*, *Six Healing Sounds*, and *Iron Shirt Chi Kung*. We learn to gather and refine our life-force into a Chi Ball (energy sphere) so it will not dissipate when we are ready to leave this world. As people grow older, their life-force weakens, often resulting in illness and suffering. Using drugs to combat illness drains so much of the body's life-force that there may not be enough energy left to follow the primordial light (clear light) to the Wu Chi (our original source God) at the moment of death. The basic practices of the Universal Tao ensure that we retain enough of our health and vital energy to make that journey.

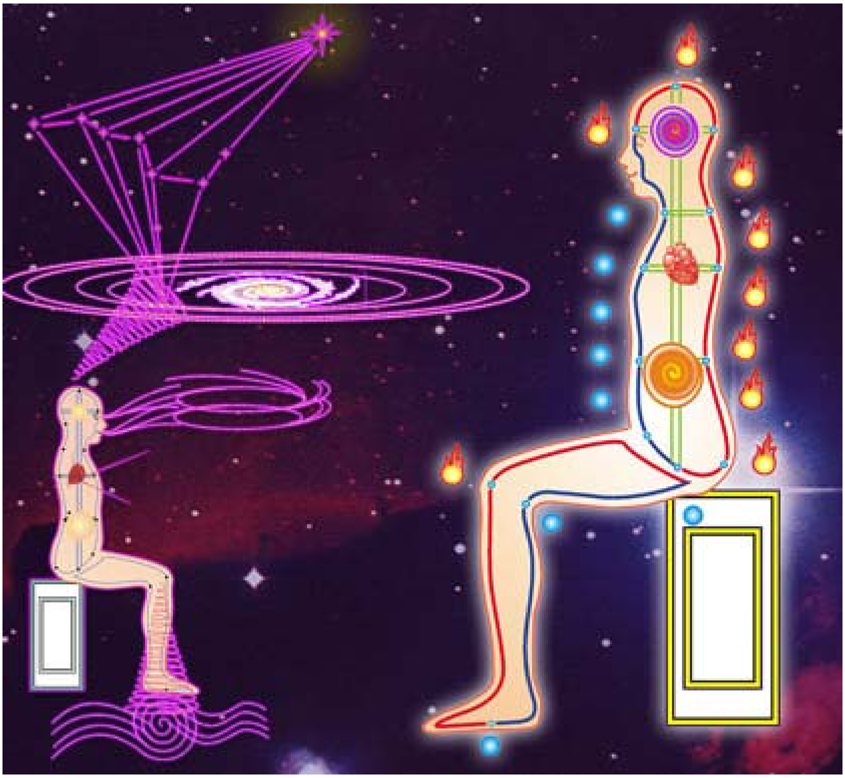

Fig. 1.2 *Sit down and start to move the orbit in the fast speed, from 1000 rotations per minute, to 10,000 per minute and up to 30,000 per minute.*

Stopping energy leakage through conservation and recycling.

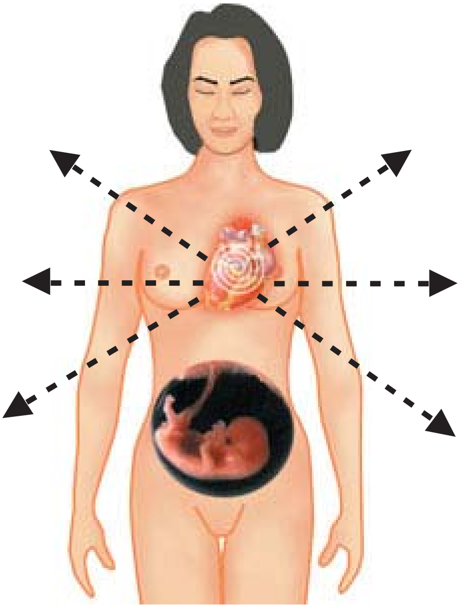

Fig. 1.3 *Heart Leakage*

The Microcosmic Orbit is the body's major energy pathway. Along this path there are nine openings. If we learn how to seal them when they are not in use, that simple act of conservation will save an immense amount of energy.

The Microcosmic Orbit meditation is the first step toward attaining these goals, as it develops the power of the mind to control, conserve, recycle, transform and direct Chi (the Chinese term for energy, or life-force) through the body's primary acupuncture channels. By learning how to manage our Chi effectively, we gain better control over our lives; by using our energy wisely, we discover that we already have plentiful Chi.

At this stage, one also learns to connect with and draw from the unlimited source of Universal Love, a Cosmic Orgasm formed by the union of the three main sources of Chi accessible to humans: the *Universal (Heavenly) Force*, the *Earth Force* and the *Higher Self (Cosmic) Force*. This process is both energizing and balancing. It prepares one for working with greater amounts of Chi in the higher levels of meditative practice, particularly in developing the Energy Body.

Abdominal Brain

Learning to empty the mind is one of the most important Universal Tao practices. Letting go of judgments, suspicions and the incessant repetitive thought process is a simple way to recharge the body, the brain and the spirit. Science has proven that we have about 60,000 thoughts per day and of those 60,000 thoughts 95% of them are the same thoughts we had yesterday. By activating the abdominal brain, we are able to clear the mind and create space for more creative mental activity.

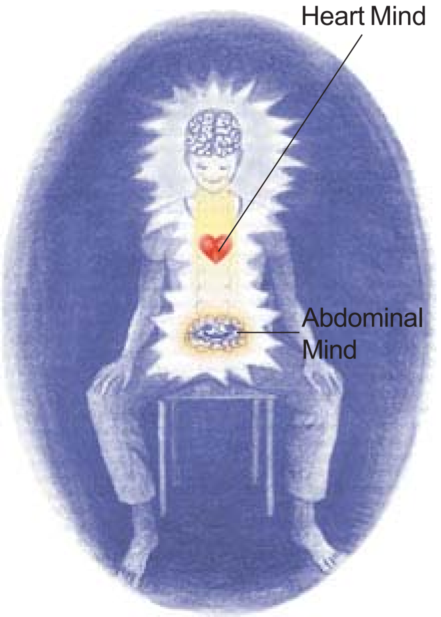

Fig. 1.4 *Upper and Lower Brains*

The brain spends up to 10 times more energy than the rest of the body. Even when we are not working or thinking consciously, our mind still wanders. In Taoism this incessant thinking is called the “Monkey Mind”. If we can stop the mind and empty the energy from the brain down to the organs, there will be tremendous amount of energy stored in the body. This abundance of energy can later return back to the brain. When the energy returns, our thoughts are much more refined, creative and focused.

The abdominal brain is located in the Lower Tan Tien, the reservoir of energy located about an inch below the navel and three inches inside the body. This is the very center of our being. It is called the abdominal brain because it is a center for wisdom. Science is learning that the same chemical activities that go on in the brain, go on in the abdomen. The difference is that the abdomen is not connected with the senses. The senses tend to pull our energy and awareness outside of ourselves. This leads to the constant analyzing of situations and people. The abdominal brain is a much deeper sense of awareness, where our “gut feelings” originate. When the abdominal brain is activated, we open up to more guidance, deeper intuition and full body awareness.

Transforming Negative Energy into Virtues: Opening the Heart

The Inner Smile and Six Healing Sounds are simple yet powerful practices that teach us how to relax and heal the vital organs and to transform negative emotions back into a rich source of energy. They help open the heart center and connect us with unlimited Universal Love, improving daily interactions and providing a vehicle for the virtues, which derive from the internal organs. Taoists perceive the heart as the seat of love, joy and happiness, which can connect with Universal Love. It is also a cauldron in which the energies of our virtues are combined and strengthened. Through the Inner Smile, you will feel these virtuous energies generated from their respective organs. These virtues are then gathered into the heart to be refined and blended into compassion, the highest of all virtues. This is a most effective way to enhance one's best qualities.

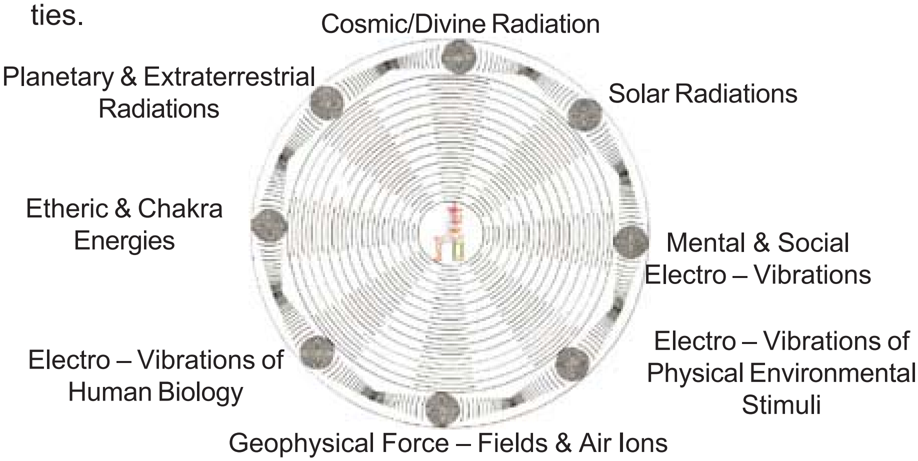
Cosmic/Divine Radiation
Planetary & Extraterrestrial Radiations
Solar Radiations
Etheric & Chakra Energies
Mental & Social Electro – Vibrations
Electro – Vibrations of Human Biology
Electro – Vibrations of Physical Environmental Stimuli
Geophysical Force – Fields & Air Ions

Fig. 1.5 Cosmic/Divine Radiation Fields

Another important beginning practice teaches the student of the Tao how to gain strength through the conservation and recycling of sexual energy (*Healing Love through the Tao*). When collected, sexual energy (*Ching Chi*) becomes an incredible source of power that can be used by the individual or shared with a sexual partner through the Microcosmic Orbit pathway during sexual intercourse. With practice, singles and couples can learn to increase and intensify sexual pleasure. The collected and transformed sexual energy is an important alchemical catalyst to be used in the higher meditations.

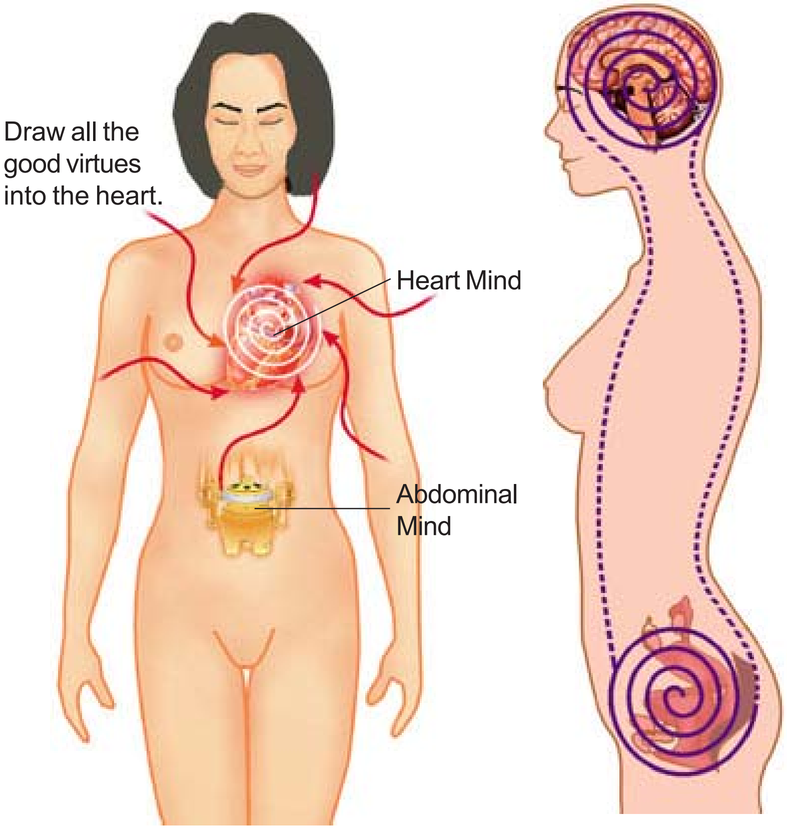
Draw all the good virtues into the heart.
Heart Mind
Abdominal Mind

Fig. 1.6 *Conserving, recycling and transforming the energy.*

Once you have an abundance of sexual energy, you can connect to the unlimited Cosmic Orgasm experienced every moment by your Higher Self, which is the most basic energy in every cell of your body.

Managing the Life-Force

In the practices of Iron Shirt Chi Kung and Tai Chi Chi Kung, one learns to align the skeletal structure with gravity to allow a smooth, strong flow of energy. With strong fasciae, tendons, and bone marrow and good mechanical structure, we can manage our life-force more efficiently. The body also gains a sense of being rooted deeply in the earth, so one can tap into the Mother Earth healing force.

Cosmic Nutrition

The Taoist approach to diet is based on determining the body's needs and then fulfilling them according to the five elements of nature, which support the five major organs of the body. This system reveals and strengthens any weak organs by balancing one's food intake to enhance any deficient elements. It does not condemn most foods that people enjoy (including sweets), but instead creates a better program in which these foods can support the body's internal balance rather than disrupt it. Choosing and combining foods in this way can help us avoid the cravings we sometimes fall prey to.

Developing the Energy Body: Our Vehicle to Travel in Inner and Outer Space

The next level of the Universal Tao system consists of the *Fusion of Five Elements, Cosmic Fusion,* and *Fusion of the Psychic Channels*. These practices build and enhance the basic Universal Tao exercises. They use the extra energy saved through the foundation practices, including recycled negative energies, to build a strong Energy Body that will not dissipate. Developing this Energy Body awakens a part of oneself that perceives and acts free of environmental, educational and karmic conditioning. Once the Energy Body is strong, it becomes a vehicle (like the space shuttle) to help the untrained soul and spirit for the long journey home, back to the Wu Chi.

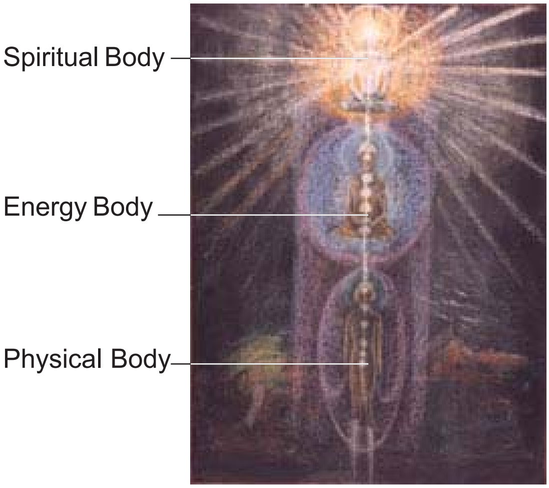

Fig. 1.7 *Fusion is the First Step in achieving Union of our Three Bodies.*

If we do not have a chance to practice awakening or to give birth to the soul and immortal spirit during life, the primordial light will awaken us at the moment of death. Unfortunately, we may be too untrained and inexperienced to follow this light. To prepare for the journey, the Energy Body is a vehicle of great importance. We can train and educate the Energy Body so it can help the untrained soul and spirit recognize and follow the primordial light back to our original source.

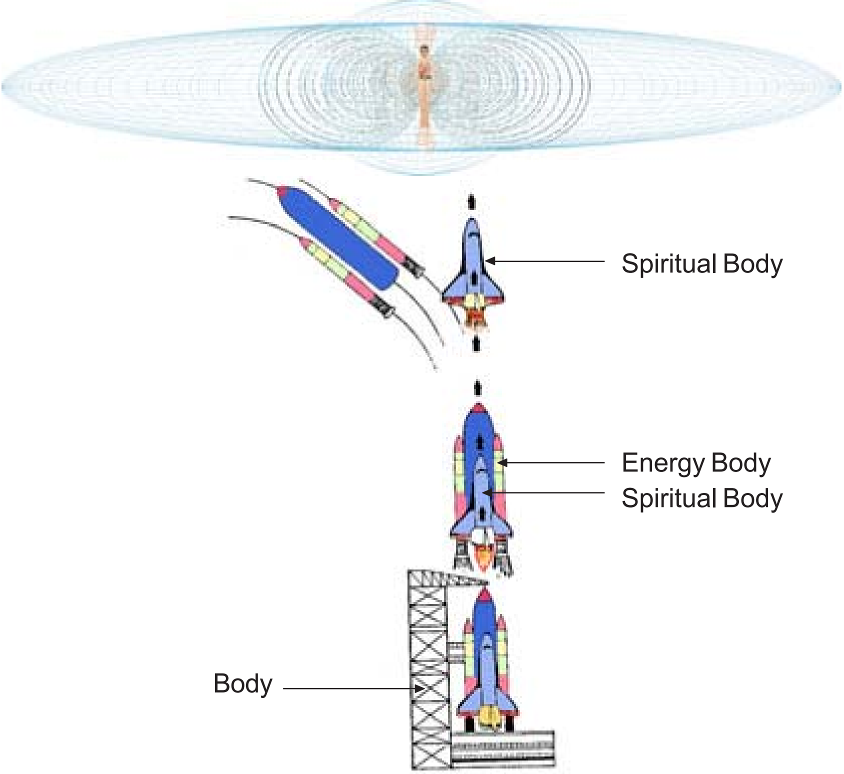

Fig. 1.8 *Energy Body blasting off the Spiritual Body and enters outer space.*

When we are ready to give birth to the real soul, the Energy Body will act like a booster rocket to help boost the soul into its higher dimension of the immortal spirit. At the highest level, all three of these bodies merge into one.

Each level of development gives us a chance to go further in the journey back to the Wu Chi. Taoist methods of absorbing stellar energies help rejuvenate the physical body and strengthen the soul and spirit bodies for their inter-dimensional travels.

Recycling our Negative Emotions

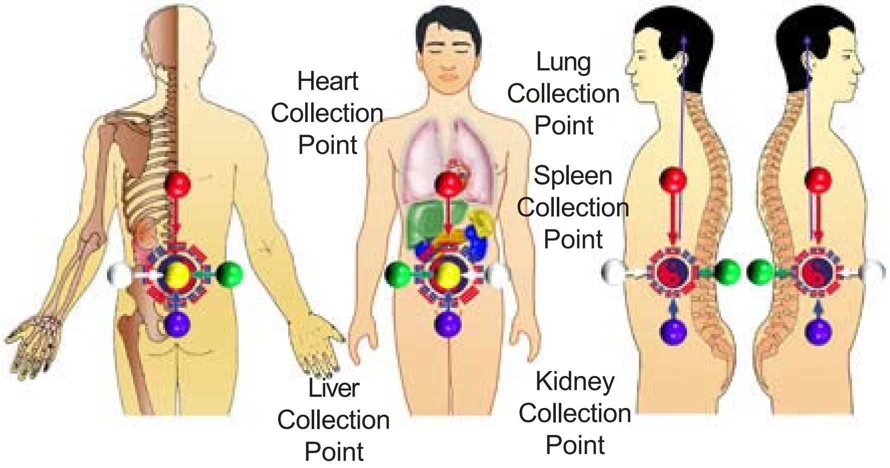

Fig. 1.9 *There are Five Major Collection Points of the Emotions reflected on all four sides of the body.*

Our emotional life, filled with constant vicissitudes, drains our vital energy. Through the Fusion meditations, one learns to transform the sick or negative energy that has been locked up in the vital organs. Taoists understand morality and good deeds as the most direct path to self-healing and balance. To be good to others is good for oneself as well. All the good energies we create are stored in the Energy Body like deposits in a bank account. By helping others and giving them love, kindness and gentleness, we receive positive energy in return. When we open our hearts, we are filled with love, joy and happiness.

From Taoist experience, we know that when we leave this world we can go directly to heaven, depending on how much energy we have been able to transform into the Energy Body prior to death. Just like money in the bank, the more we transform our physical being to our spiritual being, the more we have in heaven. The more good we do here, the more positive energy we have up there.

Forming the Spirit Body: Planting the Seed of Immortality – Lesser Kan & Li

The Inner Alchemy meditation of the Lesser Enlightenment of Kan and Li (water and fire, multi-orgasmic sexual energy and compassion) reunites the male and female within each of us. It involves the practice of self-intercourse, which by internal sexual coupling of the energies enables one to give birth to the soul body. The soul body then acts as a "baby sitter" to help nurture the spirit body. The soul is the seed, but it can also be matured into the immortal body if one has not had the chance to raise the spirit body in this life. Practitioners of Taoist Alchemy believe that if we give birth to the spirit body and develop the immortal body in this life, we can overcome the cycle of reincarnation.

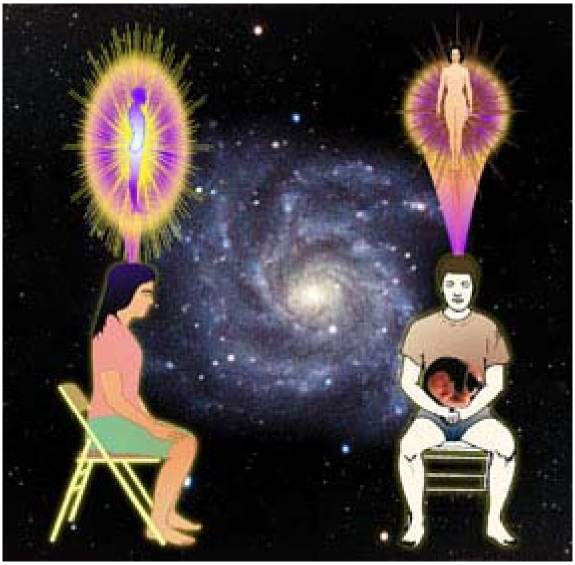

Fig. 1.10 *Creating the Energy Body in preparation for the Spiritual Body.*

Once the "baby sitter" or soul body is formed, it is in the Yin stage, or infancy (soul embryo). We need to feed, raise, educate and train the young soul to become fully grown.

Once the soul body is developed, we can give birth to the spirit body. To cultivate the young spirit body until it is fully mature can take 14 to 18 years. We also use the energies of nature (trees, sun, moon and stars); virtually all sensory experiences of a positive nature become nourishment for the growth of the spirit within the physical body.

Many masters who attained this level of the immortal body were able to transform the material into the immaterial and transfer it into the spirit body. At the moment of death, they were able to transfer their consciousness, their energy and the physical elements of their bodies up with them into the spirit body, although even this level is not yet the true immortal body. In this process their physical bodies actually shrank in size; they may have weighed two-thirds of their usual weight after their physical deaths occurred. This meant they had successfully transformed much of their material being into an immaterial state while retaining full consciousness.

Cultivating the Yang Stage of the Immortal Body: Marrying the Light - Greater Kan and Li

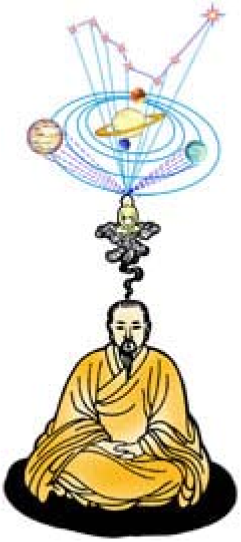

Fig. 1.11 *Greater Kan and Li*

To merge with the light of the Tao, we must awaken and nourish the awareness that we are in truth children of the light. Once we have fully grown the spirit body, it will be the same frequency as the light of the Tao and can become one with that light. Other traditions refer to this light by such names as the Holy Spirit or Great Spirit; we also refer to it as the "outer light."

The Greater Kan and Li meditation teaches us how to recognize the inner light of our own spirits and shows us how to merge with or "marry" it to the outer light. Once we connect with and "marry the light," we give birth to the second stage of the true immortal spirit. Taoists refer to this as the Yang body. One continues to transform the physical body energy to feed the immortal spirit so it can mature.

At this stage of practice, we learn to digest increasingly higher-grade energies of the Higher Self and Universal Forces from the sun, moon, planets, stars and galaxies, and from the mind of the Tao itself. An awakening to that which is eternal and enduring occurs through this practice. Cognizant of our true nature as spirits, we experience the ability to leave the physical body and travel in the immortal spirit body, which leads to experience of the inner worlds of spirit. Fear of death is vanquished as we become familiar with life beyond physicality.

Greatest Enlightenment of Kan and Li

At this level one transfers all physical essence into the immortal body. When all the body's material elements are transformed into subtle Chi, what remains is known as the "rainbow body." When a master of this level leaves this world, there is nothing left of the physical body but nails and hair. Death is still necessary to speed up the process.

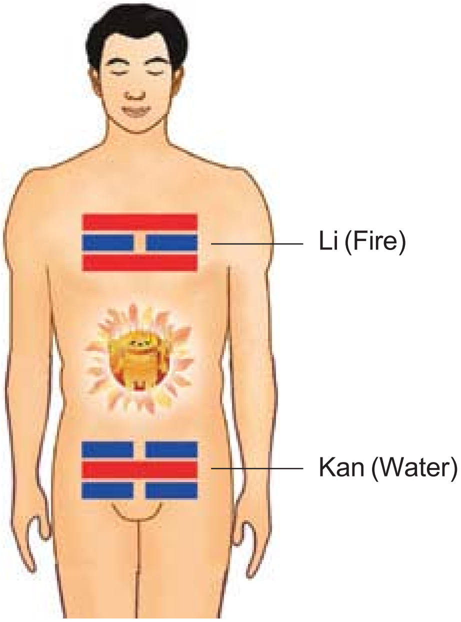

Fig. 1.12 *Union of Fire and Water*

Sealing of the Five Senses, Congress of Heaven and Earth, Reunion of Heaven and Man

At this level death is transcended entirely. One can simply transform the physical body into the immortal body and leave this world or return to it at will. This is the state of complete physical immortality. It takes from eighty to a few hundred years to complete these practices and transform all the material elements of our body into the immaterial. The final goal of ascending to heaven in broad daylight is reached.

There are records in Chinese history of many thousands of Taoist immortals who reached the level of daylight ascension in the presence of many witnesses. In the Bible, Elijah and Moses also accomplished this feat. In the final stage of this practice, the adept can unite the immortal spirit body, the Energy Body, and the physical body, or separate them at will. It is then that the human being knows full and complete freedom as an immortal, where no world is a boundary.

Chapter 2
Taoist Cosmology

Ancient Taoism is rooted in deep observation of naturally occurring Universal processes and their effects upon human beings. Newtonian physics of the West understands these processes as the mechanics of cause and effect. Taoism understands them as the interactions of a vast sea of energy that is constantly creating and recreating the universe in infinite ways. Most religions and esoteric systems study these processes (the ways of God) through scriptures and practices based upon the immaterial. Taoism studies both the material and immaterial aspects of nature and the Universe in the belief that the immaterial is both the source for the material and a product of it. In other words, physical and non-physical processes are sources for each other.

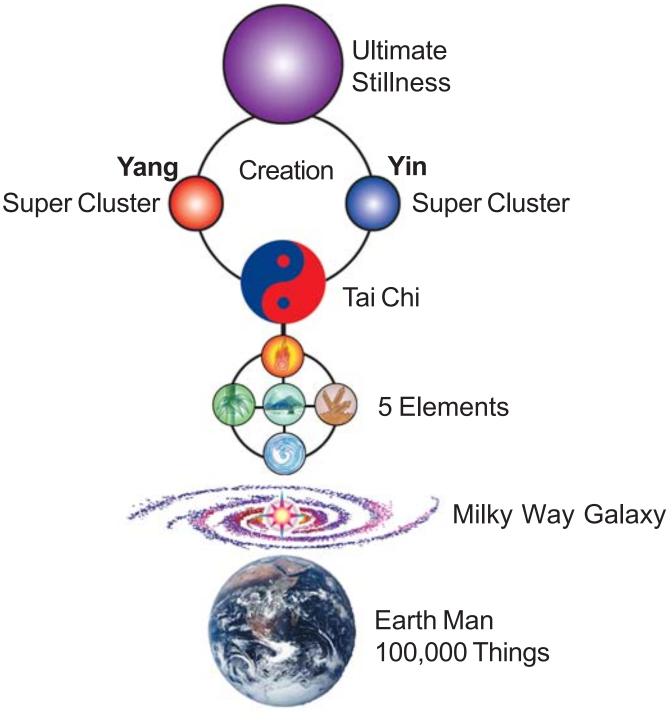
Ultimate Stillness
Yang
Creation
Yin
Super Cluster
Super Cluster
Tai Chi
5 Elements
Milky Way Galaxy
Earth Man
100,000 Things

Fig. 2.1 Taoist Cosmology

The Void: Our Original Source

Through observing nature and the effects of energy within the human body, the ancient Taoists were able to trace the universal energy back to its point of origin. Upon developing an empirical approach with which to contact this source of observable phenomena, they established the concept of the primordial void as the point of departure for all creation. This void was given the name Wu Chi, which is depicted as an empty circle in traditional Taoist art because it is beyond human description. For energy to begin generating the effects and forms of nature and the universe, something had to stir within the Wu Chi. This first stirring created the division between the material and the immaterial as all the processes of the universe began at this time.

Yin and Yang

Fig. 2.2 *Duality of Yin and Yang*

Taoists refer to the first observable variations of the Universal Force, which emanates from the Wu Chi, as Yin and Yang. These two qualities of this force can be understood as the positive and negative poles of the primordial energy. Yin and Yang are inseparable tendencies of all energy, and it is impossible to have one without the other. Their interactions are the root of all universal action; hence, the polarities of Yin and Yang are an intrinsic factor of all creation.

The theory of Yin and Yang is one of the most simple and sublime symbols of the way nature and the Universe interact. The circle that encompasses the symbol is representative of the Tao, the undifferentiated whole, the Universe, ultimate reality. Yin and Yang are terms that shed light on the process of the Tao. Yin describes the feminine, the contracting, the dark, deep side of nature and Yang describes the masculine, projective, electric, light, surface side of nature. Yin and Yang are not two separate states, but an interwoven aspect of one. Just as day cannot exist without night, or male without female, Yin and Yang are opposite co-existing elements of the same Universal substance. They are an interwoven and continuous processes of decaying and becoming. Yin and Yang describe the process of a changing dynamic picture of reality.

Five Major Forces of the Universe

The Taoists observed that Yin and Yang interactions follow five basic patterns which came to be known as the Five Tendencies or Five Processes of Energy. (Such interactions have been misleadingly translated as Five Elements, thereby confusing the process with the actual physical elements.) In Taoism the physical elements found in nature symbolically express the Five Tendencies of Energy in motion. Thus fire represents energy rising; water represents energy sinking; wood represents energy expanding; metal represents energy solidifying; and earth represents stable or centered energy. Each of the five elemental tendencies of energy are dependent upon the interactions of Yin and Yang emanating from the primordial void.

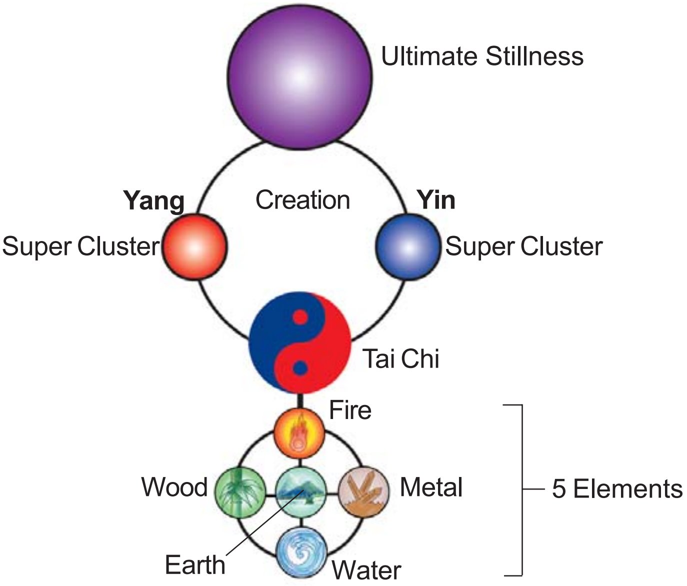

Fig. 2.3 *Oneness, Duality and the Five Elements*

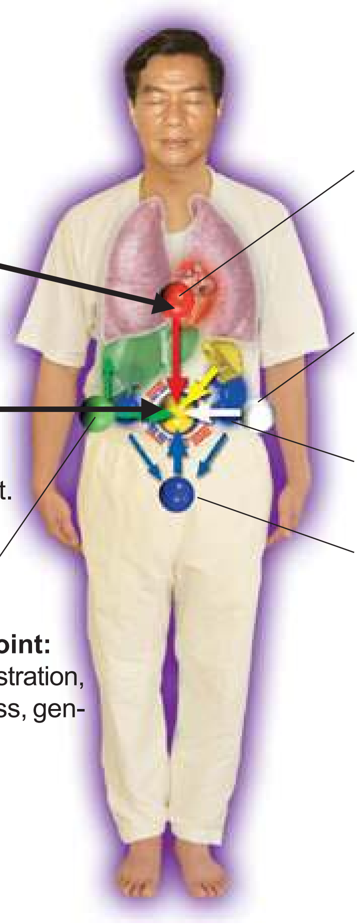

Pre-composting: Draw negative emotional energy from organs to collection points. Neutralize with the virtuous energy from the organs and heart.

Composting: Draw the neutralized energy to the Pakua. Fuse the neutral energy with love and compassion energies from the heart.

Heart Collection Point: Transform hate, cruelty, impatience into love, joy, happiness.

Lung Collection Point: Transform sadness, grief, depression into courage, righteousness.

Spleen Collection Point: Transform worry, anxiety into fairness, openness.

Kidney Collection Point: Transform fear into gentleness, calmness, silence.

Liver Collection Point: Transform anger, frustration, jealousy into kindness, generosity.

Fig. 2.4 *Transform Excess Negative Energy into Positive Energy*

The Five Elemental Forces are expressions of energy that can be observed in nature and throughout the universe. In space they influence the motions of all stars, planets and cosmic phenomena. In nature they promote interactions between the five elements of fire, water, wood, metal and earth. Within the human body they affect the five major organs of the heart, kidneys, liver, lungs and spleen. Just as western science understands the atoms and subatomic particles to be the fundamental units of all matter, the Five Elemental Forces are understood to be the essence of all processes. The forces that influence the cosmos are identical to those which affect nature and our bodies.

Eight Forces

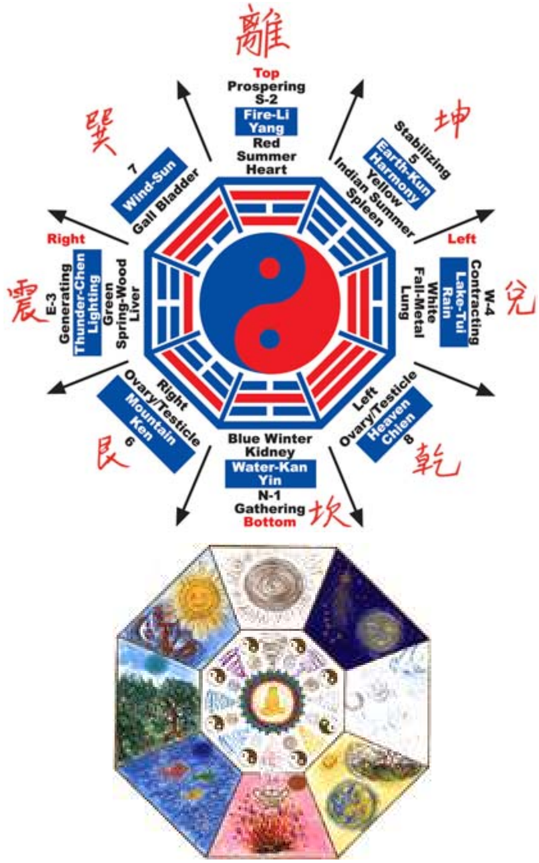
離
Top
Prospering
S-2
Fire-Li
Yang
Red
Summer
Heart
坤
Stabilizing
5
Earth-Kun
Harmony
Yellow
Indian Summer
Spleen
Left
W-4
Contracting
Lake-Tui
Rain
White
Fall-Metal
Lung
兌
Left
Ovary/Testicle
Heaven
Chien
8
乾
Blue Winter
Kidney
Water-Kan
Yin
N-1
Gathering
Bottom
坎
Right
Ovary/Testicle
Mountain
Ken
6
艮
Right
E-3
Generating
Thunder-Chen
Lighting
Green
Spring-Wood
Liver
震
巽
7
Wind-Sun
Gall Bladder

Fig. 2.5 *Eight Forces and Directions*

The Eight Forces represent a more detailed aspect of the forces of nature and the Universe. They are represented by the symbol of the pakua and the eight trigrams. In Fusion of the Five Elements, the pakua was introduced as a way to direct and gather the Chi into a pearl at the lower abdomen. In Cosmic Fusion, the pakuas are used in greater detail to create a stronger connection with the forces of nature and draw much more energy into the pearl.

Using the pakua in the meditation practice is a way to focus energy. The pakua creates a vortex that enables the practitioner to collect, gather and condense Chi. This vortex of energy is used to not only create a strong connection within ourselves, but a harmonious relationship with all the forces of nature.

Return to the Wu Chi

The Taoists realized that we are connected to the stars through our bodies as the human form is a product of stellar energy and matter. With the basic understanding that man, nature and the universe are expressions of primordial energy (or the Wu Chi), the Taoists devised methods to tap the energy of the five elements. This was done to enhance the processes of the Five Elemental Forces within our bodies, which can promote health, refine our spirits, and eventually reunite us with our source, the Wu Chi. Deep within our subconscious minds are memories and desires for the unity and bliss which preceded the dense vibrations of our gross physical state. So the ultimate purpose of Taoist practice is to return to our original state, the Wu Chi.

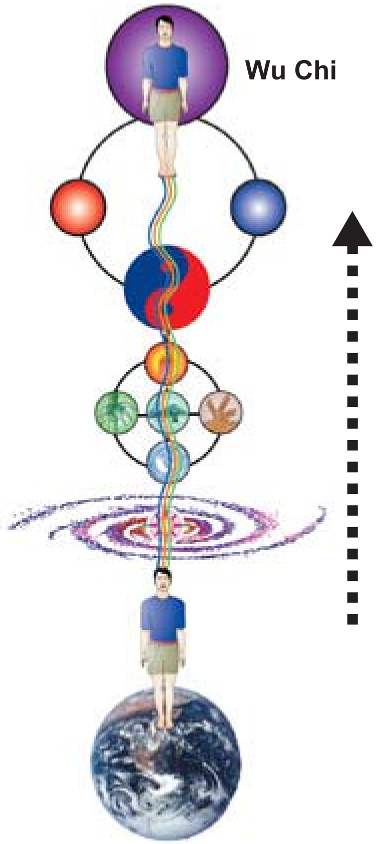

Fig. 2.6 *Return to the Oneness (Wu Chi)*

In the advanced disciplines of the Universal Tao students begin to work more consciously with the Five Elemental Forces emanating from the cosmos. Advanced practice involves collecting and absorbing energy from different sources in the universe while traveling in the energy and spirit bodies. Taoists move closer to the original source through the vastness of space in the same way a traveler moves from station to station in a train. Generations of Taoist masters have refined and mapped the simplest and safest approaches to follow. Because it is impossible to realize the Wu-Chi in one step, the journey is divided into several stages.

Returning to the Wu Chi: First Three Steps

1. The first step of stage one is to upgrade and maintain your physical health while stabilizing your financial and social conditions. Your body is a vehicle for processing energy, and the healthier you are, the more energy it will process at a faster rate. Your health and processing capabilities will enhance greatly through the Microcosmic Orbit, Six Healing Sounds, Inner Smile, Healing Love and Iron Shirt Chi Kung practices. Your finances and social interactions should be stable so that you can create space in your life for these disciplines.
2. Next you will learn how to use accumulated energies for traveling beyond this earthly plane to the moon, sun and planets. This is similar to the way a space shuttle uses a booster rocket to push it beyond the earth’s gravitational field. In Taoism the Microcosmic Orbit and Fusion practices create the Energy Body to help boost the immortal body (spiritual body) out of this world. The healthier you are internally, the more power you will be able to generate to leave the confines of the physical body. When you can travel out of this reality and back, you will gain the raw materials necessary to build more powerful energy and spirit bodies for longer and farther journeys.
3. The higher practices involve traveling to and from the stars and constellations particularly the North Star and Big Dipper using the energy and spirit bodies. The North Star is a stable and constant point in the sky. For Taoist practitioners, using the North Star was a way to find direction in the Universe. The North Star is the gate of heaven. In the higher practices, the energy is drawn into the body to create that connection between the physical and the Universal bodies.

The human body is a powerful transformer. The sexual energy is the body's most powerful source of fuel for journeys to the higher planes. This is why the Healing Love practice is so important for spiritual development. The practice of Iron Shirt Chi Kung also helps the physical body absorb cosmic energy through the skin and bones to be transformed and refined into more fuel. Fusion of the Five Elements provides another mode of energy refinement through the transformation of negative energies into positive life-force, which is lighter and healthier for the body. All of these internal methods help one develop the means of traveling beyond this world. Their ultimate purpose (after rejuvenating the body) is to accumulate the raw materials for the construction of a stronger and more powerful vehicle for the return to the Wu Chi.

Many other systems have their own means of returning to the source. Monks, nuns and priests emphasize detachment from the world and non-involvement in worldly affairs. They have no relationships with other people and no sex, which leaves them with abundant sexual energy to be transformed and used for higher spiritual work. Such spiritually inclined people make the best and most use of sexual energy. In general, when normal people accumulate sexual energy to the same degree as monks, nuns or priests, they have no way to control it, and internal imbalances are created until it is released. The Universal Tao practices can give anyone the ability to use this energy to improve the body and spirit.

One should not believe in the Taoist teachings without experiencing the practice and its results first hand. This involves learning about the human body, the Universal forces which affect it, and the inherent desire we all have to return to our source. Taoism offers a clear, direct path to the Wu Chi, whereas many other systems provide only dogma and restrictions that can hinder spiritual growth. The Tao is not an escapist philosophy, however, as there is no need to remove yourself from society or restrict your relationships or sexual love. You can continue to live normally as you accumulate energies that will gradually improve your health while providing raw materials for the highest spiritual growth.

Chapter 3

Goals of the Fusion Practice

The Taoist study of **Inner Alchemy** was introduced in Fusion of the Five Elements. The Taoists developed the Fusion of the Five Elements practice to develop connections between and gain control of the inner and outer Universes.

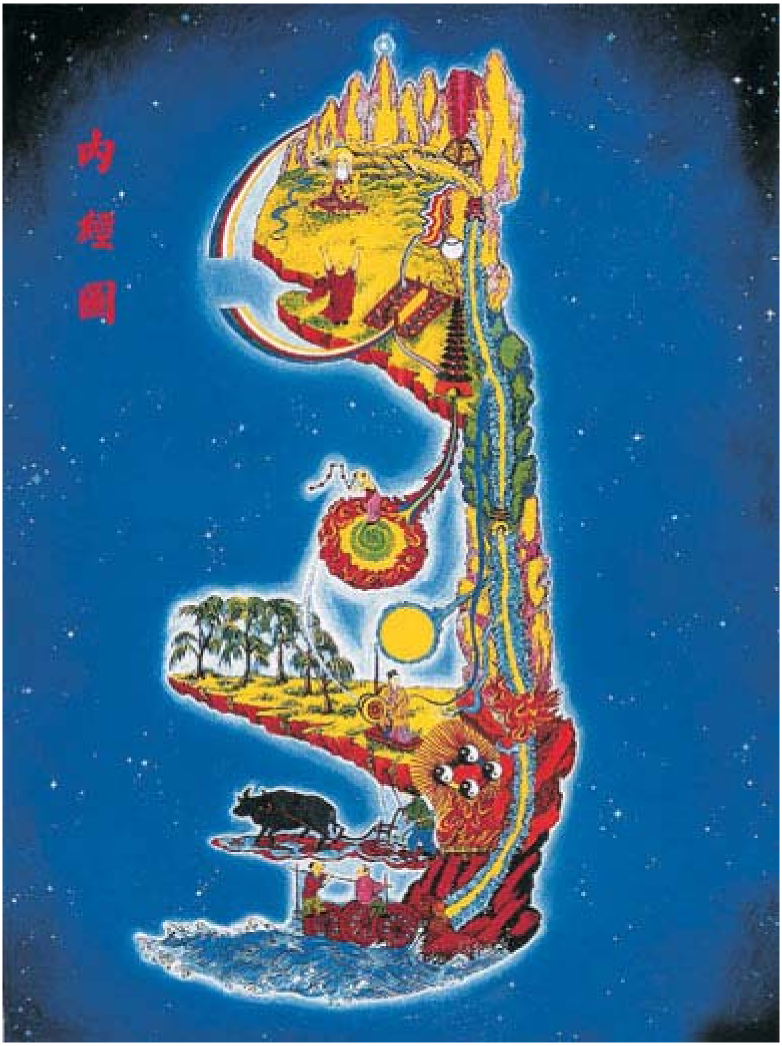

Fig. 3.1 *Taoist Inner Alchemy*

They understood that virtually everything that exists in the Universe can be broken down into the Five Elements. The Fusion practice begins by understanding the dynamics of the Universe, the planet Earth, and the human body with respect to their relationships to these Five Elements.

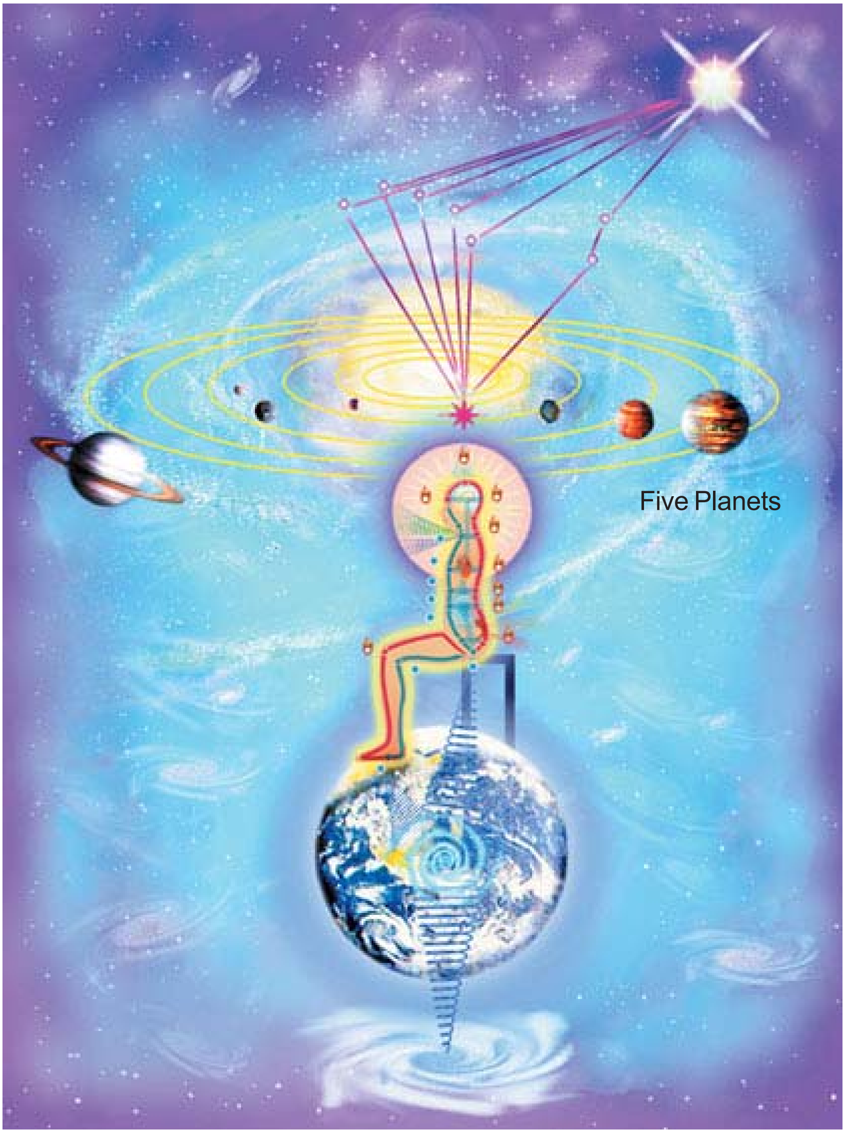

Fig. 3.2 *Five Elements, the Planets and the Cosmos*

The knowledge of Inner Alchemy is a necessary step in becoming connected to the outer Universe, from which an unlimited supply of powerful energy is available to be used for an individual's physical and spiritual benefit.

The five basic formulas of Fusion I, II, III can be likened to wiring blueprints for making the subtle energy the vehicle, the spaceship connections that link vital organs, glands and senses so that their respective energies can be fused and balanced in the collection point and the pakua. The formulas lead the student systematically through the process of creating four pakuas and an energy pearl.

Five Element Organ Correspondences

Yin Organs	Liver	Heart	Spleen	Lungs	Kidneys
Yang Organs	Gall Bladder	Small Intestine	Stomach, Pancreas	Large Intestine	Bladder
Openings	Eyes	Tongue	Mouth, Lips	Nose	Ears
Positive Emotions	Kindness Generosity	Love, Joy	Fairness, Openness	Righteousness, Courage	Gentleness
Negative Emotions	Anger Envy Frustration	Hate, Impatience	Worry, Anxiety	Sadness, Depression	Fear, Stress
Transform Pure Organ Energy into a Virgin Child Dressed in	Green	Red	Yellow	White	Blue
Transform Pure Child Energy into an Animal	Green Dragon	Pheasant, Red Bird	Phoenix, Yellow on Red	White Tiger	Blue Deer
Earth Force takes the form of	Green Dragon	Pheasant, Red Bird	Phoenix, Yellow on Red	White Tiger	Black Tortoise
Directions	East	South	Center	West	North
Planets	Jupiter	Mars	Saturn	Venus	Mercury
Universal Energy Force	Generating	Prospering	Stabilizing	Contracting	Gathering

Fig. 3.3 *Five Element Organ Energy Chart*

Energy drawn from the organs and senses into the collection point and transformed negative emotions are fused and condensed into a pearl through the pakuas. When the resulting pearl, with its refined, balanced, neutralized energy is circulated through the Microcosmic Orbit, the body absorbs this enhanced life-force. The pearl is then used to form the Energy Body.

Cosmic Fusion builds on the foundation of internal alchemy laid in Fusion of the Five Elements. The student is encouraged to take this opportunity to review the basic Fusion practice.

Once the negative emotions have been drawn out and transformed, the pearl is moved through the creation cycle, nurturing the virtues. These qualities are cultivated, blended and condensed into the pearl, forming the energy of compassion —the essence of the senses, glands, organs and the mind.

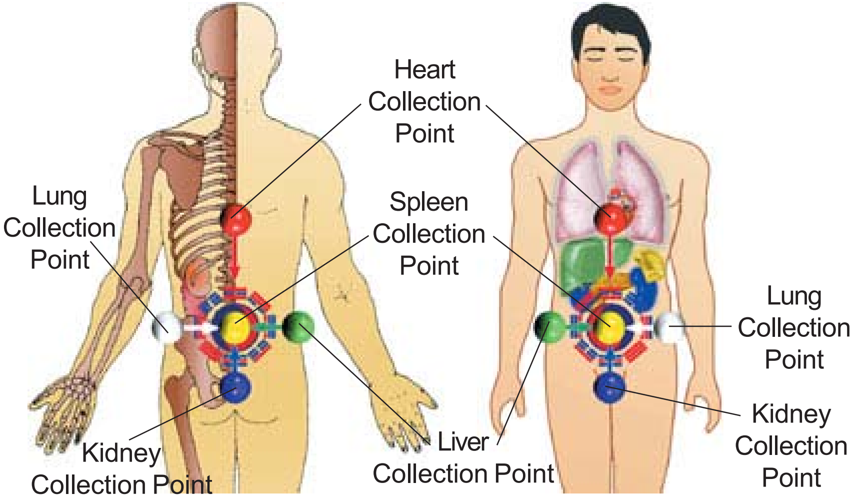
Heart
Collection
Point
Lung
Collection
Point
Spleen
Collection
Point
Lung
Collection
Point
Kidney
Collection
Point
Kidney
Collection Point
Liver
Collection Point

Fig. 3.4 *Front and Back Collection Points*

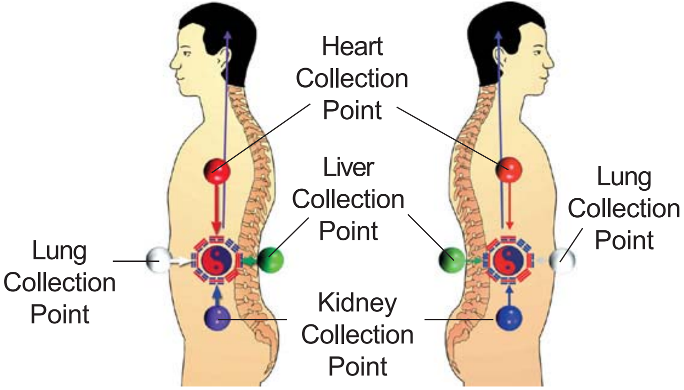
Heart
Collection
Point
Liver
Collection
Point
Lung
Collection
Point
Lung
Collection
Point
Kidney
Collection
Point

Fig. 3.5 *Left and Right Collection Points*

This highly refined pearl can now be used to open and clear four of the eight specific energy channels in the subtle body: the Three Thrusting Channels used for cleansing and protection, and the protective belt route that surrounds the Thrusting Channels.

Inner Universe reflects the Outer

The Fusion practice offers a practical way to strengthen the conscious connection with the outer Universe. During the practice of Fusion, the essence of life-force energy found in the organs, glands and senses is transformed, purified, condensed and combined with the Universal, Cosmic Particles and Earth Forces in order to achieve internal balance. This transformation of quality energy into a harmonious whole can effect positive changes in the human body. Controlling this energy enables each individual to attain balance and harmony of these energies on physical, emotional and spiritual levels. Because the inner Universe is a reflection of the outer one, the balance and harmony attained in the inner Universe enables the individual to attain balance and harmony with the outer Universe. Fusion of the Five Elements, the first in a series of the Taoist practice of Inner Alchemy, focuses on the interaction and fusion of all five elements and their correspondences. To understand this interaction, especially with respect to the organs, glands and senses of the human body, it helps to be aware of one of the Taoist laws of the Universe: *Nothing remains the same; everything changes because the Five Energies of Nature constantly interact and change*. This acknowledgement that change is constant contrasts sharply with the Western framework. Under this assumption, each individual's disposition and health is influenced by the balance within him or her of these five energies, a balance affected by the ever-changing conditions of the Universe.

The outer Universe is comprised of Universal Force, Cosmic Particle Force and Earth Force from which all things emerged and whose three combined forces sustain all existence. The concept of Inner Alchemy is grounded in the Taoist belief that the inner Universe is a reflection of the outer Universe. There are connections that can be made between the inner and outer Universe through which energy, recognized and experienced in the inner Universe as Chi, or life-force, can be greatly increased and enhanced by the immense power of the outer Universe.

Overview of the Cosmic Fusion Universal Tao divides Fusion into Three Parts

Fusion of the Five Elements

First Part

Fusion of the Five Elements makes use of pakuas and energy collection points to balance, connect and draw out negative emotional energies found in the organs. These energies, along with their corresponding glands and senses energies, are then fused and transformed into pure, life-force energy.

The purity of this energy has an adhering and magnetizing quality enabling it to condense into a ball of refined energy which is called the pearl.

The pearl is then used to connect to the Universal, Cosmic Particle and Earth Forces whose energies become part of the pearl. A soul or Energy Body is formed from the pearl, enabling the individual to go beyond connecting with the forces of the universe on a purely physical level. Later the pearl returns to the organs and glands to enhance them and provides a protection to the physical and soul body.

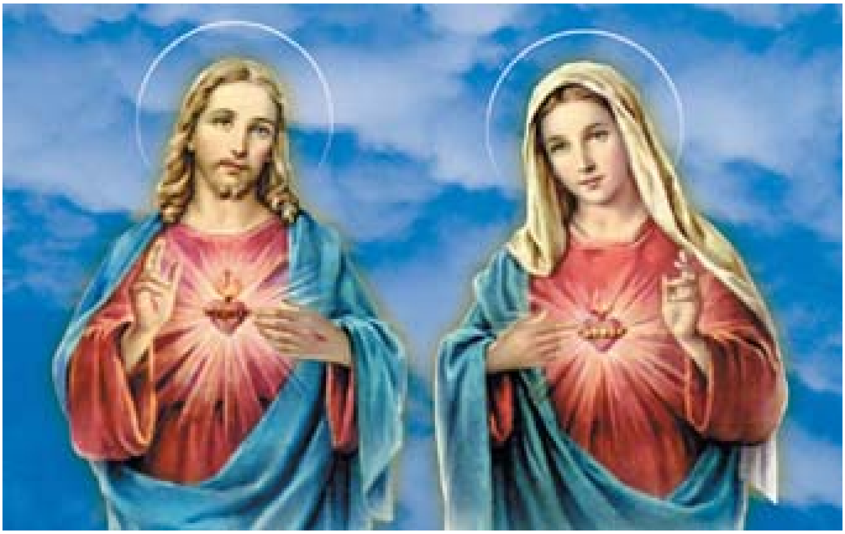

Fig. 3.6 *Forming the Compassion Heart.*

Second Part

Cosmic Fusion focuses on using the pearl to grow or intensify the energy of good virtue. It makes use of the interaction of the elements described as the Creation Cycle of the Five Elements Theory to circulate positive Chi or virtuous energy through the major organs. All the energy gathered during this cycle is combined to form a beautiful pearl of Compassion energy. This pearl is then used to open and cleanse specific channels that pass through the physical body into the Energy Body.

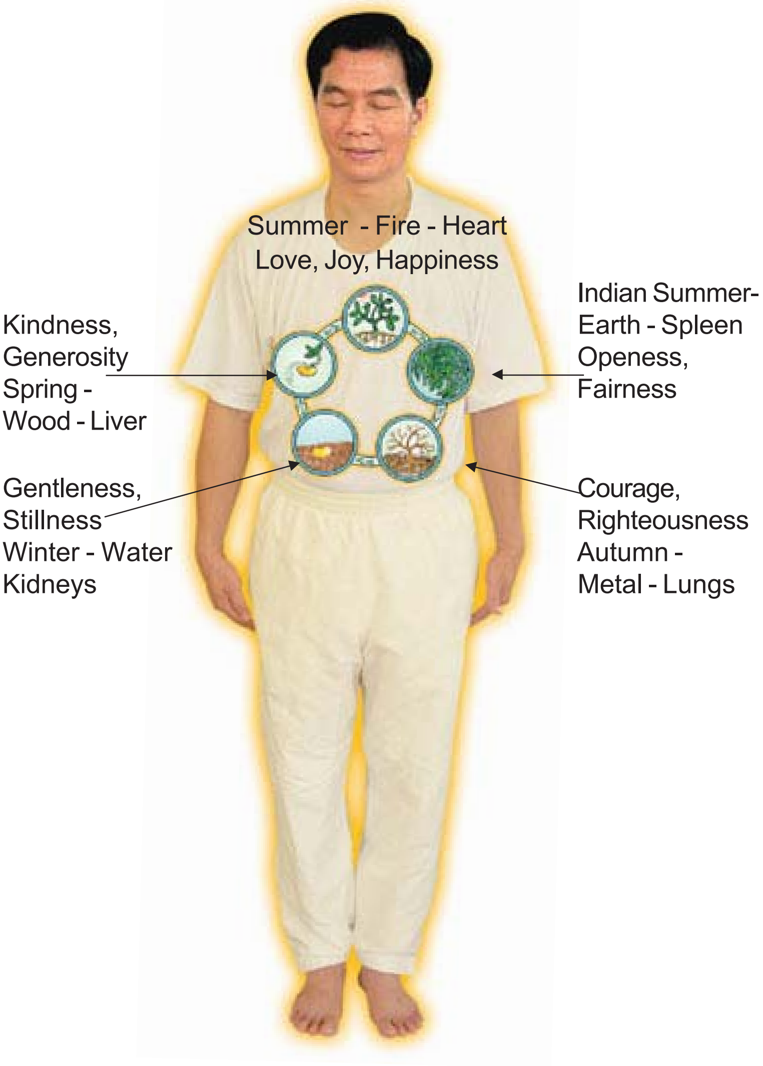
Summer - Fire - Heart
Love, Joy, Happiness
Indian Summer-
Earth - Spleen
Openess,
Fairness
Kindness,
Generosity
Spring -
Wood - Liver
Gentleness,
Stillness
Winter - Water
Kidneys
Courage,
Righteousness
Autumn -
Metal - Lungs

Fig. 3.7 *Creation Cycle in Nature and Human Body*

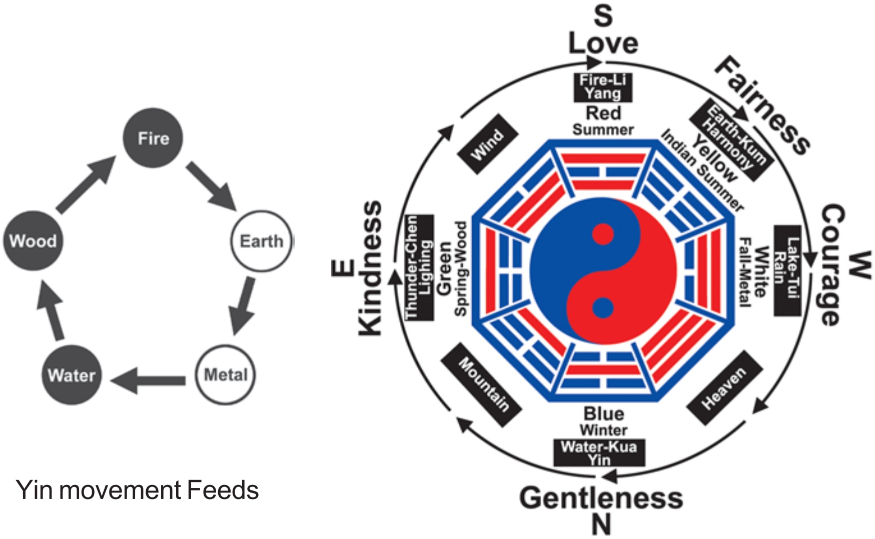
Fire
Earth
Metal
Water
Wood
S
Love
Fire-Li
Yang
Red
Summer
Fairness
Earth-Kum
Harmony
Yellow
Indian Summer
Wind
W
Courage
Lake-Tui
Rain
White
Fall-Metal
Heaven
Blue
Winter
Water-Kua
Yin
Gentleness
N
Mountain
E
Kindness
Thunder-Chen
Lighing
Green
Spring-Wood

Yin movement Feeds

Fig. 3.8 *Creation Circle*

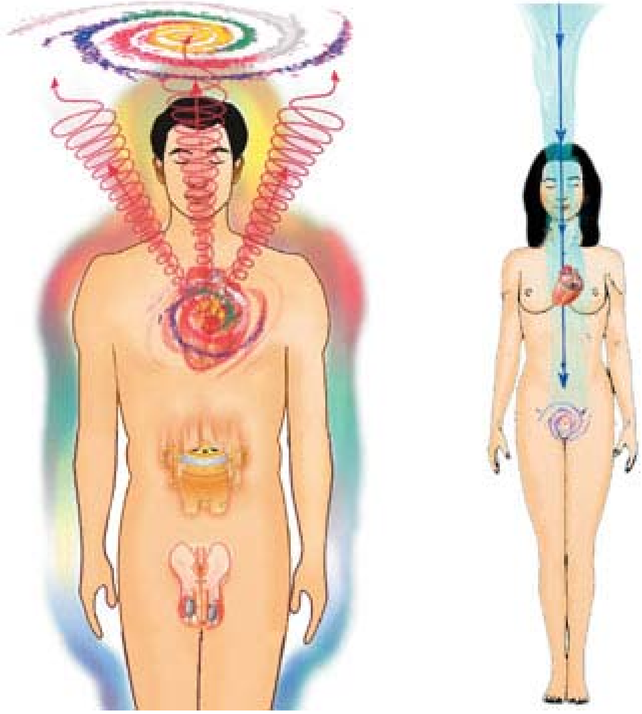

Fig. 3.9 *Forming the Cosmic Compassion of the Heart*

Third Part

The energy of the positive emotions of all the organs is absorbed into the pearl and then circulated in four of the eight special channels: the Microcosmic Orbit (Governor and Functional Channels), the Thrusting Routes, and the Belt Routes. The Thrusting Routes run through the center of the body, linking the “chakra” centers. The Belt Routes spiral around the body strengthening the aura and providing a form of psychic self-defense.

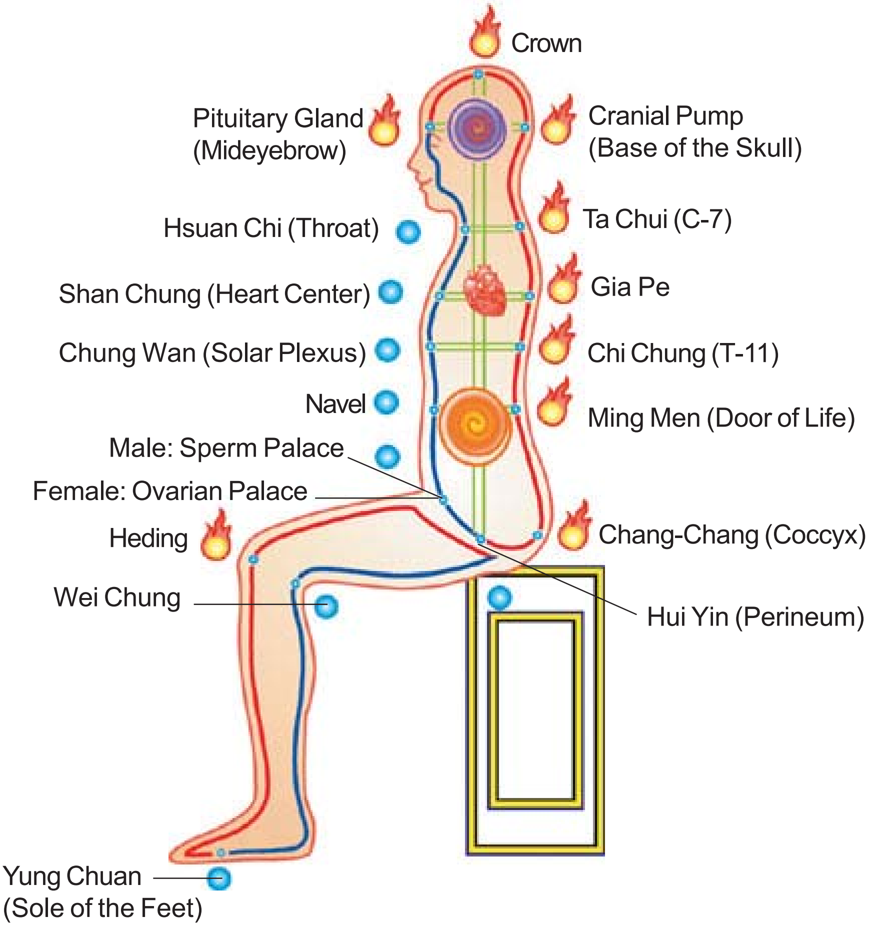
Crown
Pituitary Gland
(Mideyebrow)
Cranial Pump
(Base of the Skull)
Hsuan Chi (Throat)
Ta Chui (C-7)
Shan Chung (Heart Center)
Gia Pe
Chung Wan (Solar Plexus)
Chi Chung (T-11)
Navel
Ming Men (Door of Life)
Male: Sperm Palace
Female: Ovarian Palace
Heding
Chang-Chang (Coccyx)
Wei Chung
Hui Yin (Perineum)
Yung Chuan
(Sole of the Feet)

Fig. 3.10 *Cosmic Orbit with Ancient Chinese Points*

The three Thrusting Channels, sometimes referred to as Thrusting Routes and the Belt Channels or Belt Routes, opened during Cosmic Fusion, permit a freer flow of energy throughout the body.

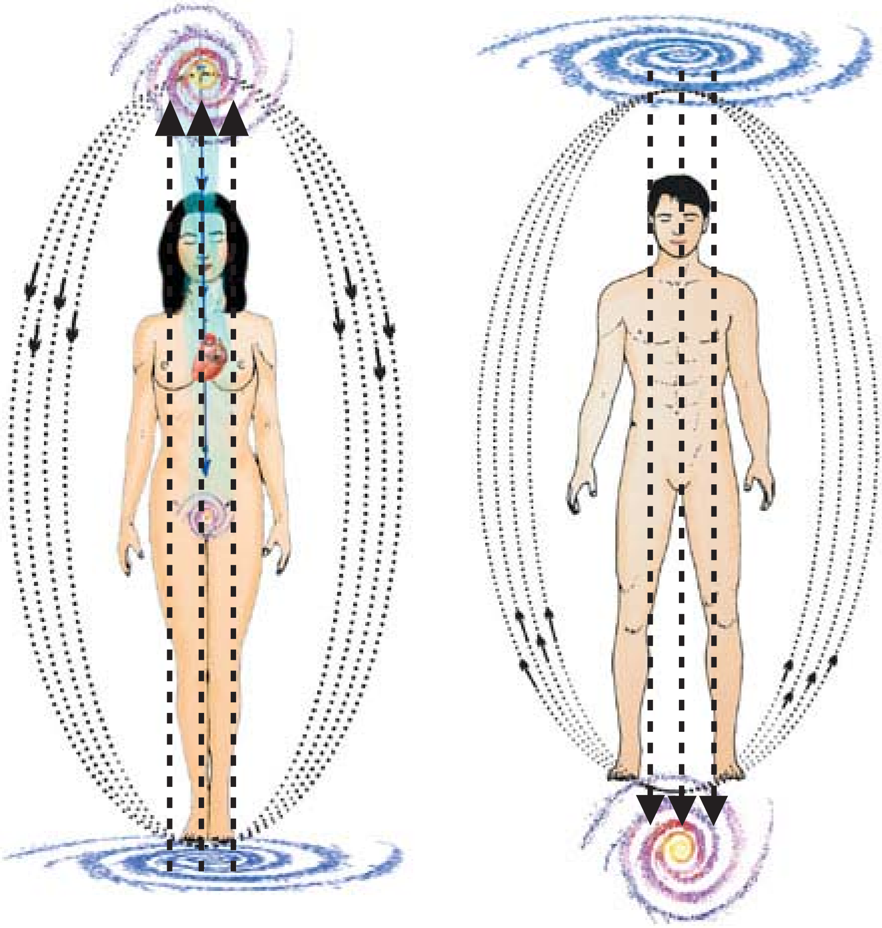

Fig. 3.11 *Cosmic Thrusting Channels Women and Men*

The Thrusting Channels are opened first. Their primary function is to cleanse specific energy pathways in the body, thereby opening them up for the energy to flow freely. The Belt Channel surrounds the physical body and the internal Thrusting Channels. Both channels help to build a denser body and provide protection to the body, but they have another function.

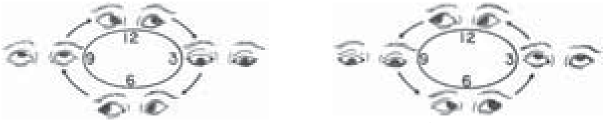

Use eyes to help spiral the Belt Channels.

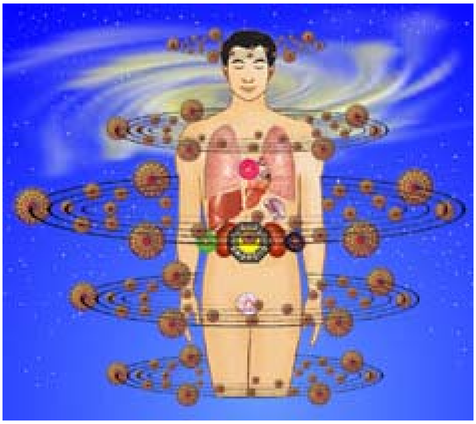

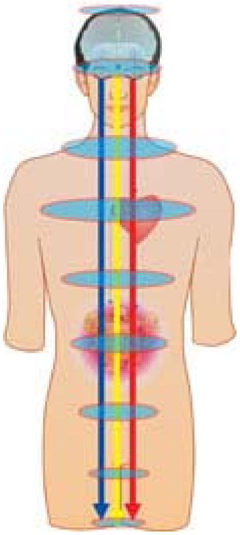

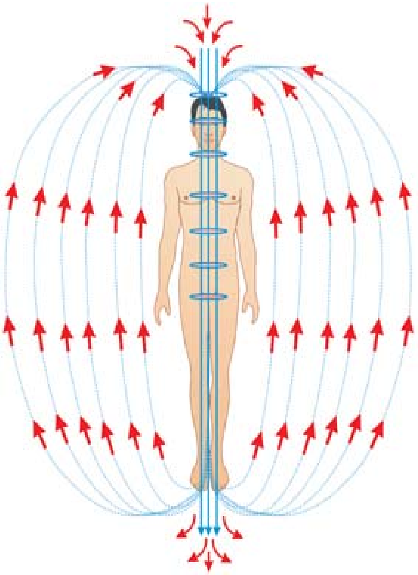

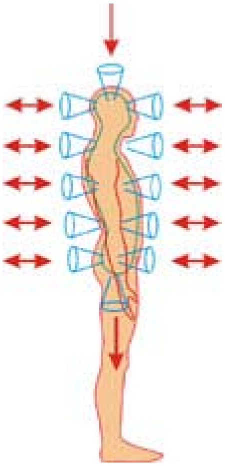

__Fig. 3.12__ Belt Channels and Thrusting Channels

With the additional energy generated during the practice of Cosmic Fusion, you will have enough energy to create a new Energy Body above the physical body. Another important aspect of the Fusion of the Five Elements practice is learning how to temporarily separate the Energy Body and the Physical Body. In life, our three bodies, the Physical Body, Energy Body, and Spirit Body (consciousness) all overlap and are generally inseparable. At death, the physical body is no longer a fit vehicle for holding the material five elements together and Chi and consciousness are forced to leave. To avoid the traumatic disorientation and confusion that can occur at death when the energetic essence and consciousness are involuntarily thrust out of the physical body into unknown territory, we can familiarize ourselves with the “out-of-body realm” by temporarily separating the Energy Body and Spirit Body from the Physical Body, as a “dress rehearsal” for death.

Consciousness always requires some sort of vehicle to manifest. Its gross vehicle is the physical body. Its more subtle vehicle is the Energy Body (sometimes referred to as the “subtle” body). We emphasized earlier that the virtues are the true energetic essence of our organs. Although we may identify ourselves more with our physical body than anything else in the world, it is our energetic essence, not our physical body, that survives after death and has the capacity to be immortal.

The Thrusting Channels and Belt Channels can be extended up into this Energy Body and Spirit Body. In this way they serve as connecting links between the organs and glands of the Physical Body and the transferred consciousness that is the Energy Body. Eventually, the channels serve as conduits through which the Spirit Body is boosted to the Mid-Plane. These advanced practices are called the meditations of Kan and Li.

If the Energy Body is thought of as a booster rocket, then the Spirit Body is the shuttle. Once the Spirit Body is boosted through the open channels into the Mid-Plane, it collects a higher quality of energy. The Spirit Body then returns with the higher quality energy to the Physical Body. In the next effort this energy is used to construct a stronger, more intense Energy Body that can boost the Spirit Body to a higher Mid-Plane level. The cycle continues bringing the Spirit Body to higher and higher levels, while increasing the life-force energy for use by the Physical, Energy and Spirit Bodies.

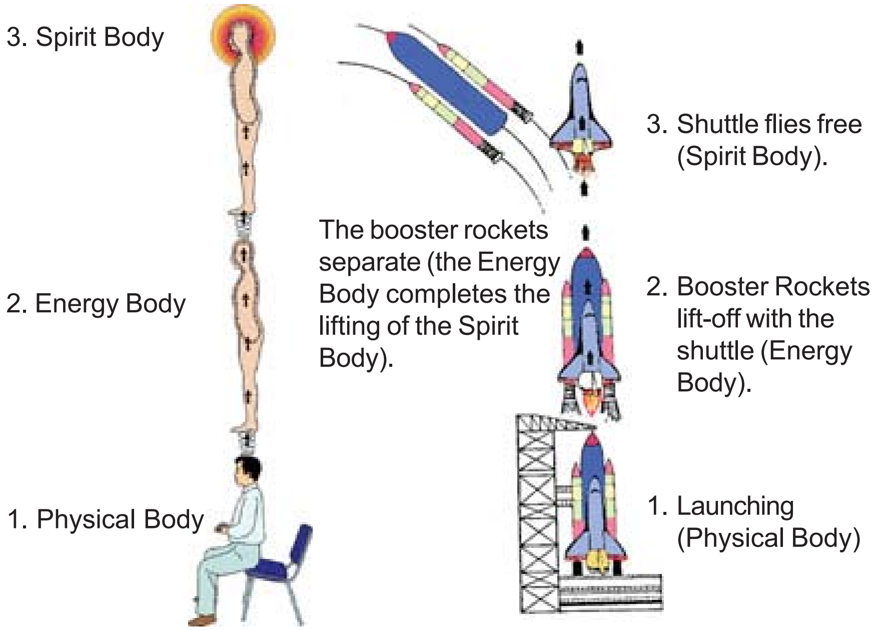
3. Spirit Body
2. Energy Body
1. Physical Body
The booster rockets separate (the Energy Body completes the lifting of the Spirit Body).
3. Shuttle flies free (Spirit Body).
2. Booster Rockets lift-off with the shuttle (Energy Body).
1. Launching (Physical Body)

***Fig. 3.13** Energy, Physical and Spiritual Bodies*

By regularly turning the awareness inward in the Fusion of the Five Elements practice, we gradually come to know our energetic essences as well as we know our face, our arms and our legs. We learn to distill these essences into a pearl, as a sort of "space capsule" to carry our essential energy and consciousness out of the physical body. With regular practice, Taoists who follow the path of Inner Alchemy learn to move their Energy Body and consciousness in and out of their physical body as easily as one might walk from one room into another.

Chapter 4
Energetic Preparations for Meditation Practice

Meditation practices are enhanced by developing one's sensitivity to subtle energy and by exercising to improve postural alignment, increase flexibility of the spine and deepen relaxation of the physical body. The next sections present a series of suggested exercises to help the student meet these goals.

Chi has many sensations. Some of the most common are tingling, heat, expansion, an electrical sensation (like the feeling of static electricity), pulsation or effervescence. This is not the actual Chi itself, but the signs of increasing Chi in a particular area. Chi itself is more mysterious, subtle beyond definition. For more details in the practices of Spinal Cord Breathing, Rocking the Spine and the Inner Smile to relax the whole body check with the book by Mantak Chia "*Awaken Healing Light*".

Preparing for the Fusion Meditations

Prepare the Environment

Find a quiet place to meditate where there are no distractions. Some places have naturally peaceful Chi: mountains, forests, caves, gardens, monasteries or beautifully designed meditation rooms. Once you have experienced such a place, you can just recall it. Set aside a corner of your room for your practice and recall your connection to the forces of nature.

Keep your meditation place clean and pleasant.

Avoid interruptions. Plan to meditate early in the morning before other people are awake: the body is rested, the air is fresh, and the earth itself is energized.

Late at night is also fine. Select a time that works well for you with your own rhythms and schedule.

If you allow yourself time each day to cleanse your emotional, energetic and physical bodies, you will soon reap the rewards in every area of your life.

Conserve Energy and Purify your Body with an Appropriate Diet

To clear your digestive tract, it helps to reduce the amount you eat, to eat less meat, fish and poultry and to increase the amount of fiber in your diet. Unless you are a vegetarian, it is recommended that you eat about 80 percent grains and vegetables, 10 to 15 percent fish and only 5 percent meat and chicken.

Loosen, Stretch and Warm Up the Body before Meditating

Once you have a quiet place and are ready to meditate, spend time loosening up the body, particularly the spine, before you sit down to practice.

The Universal Tao recommends doing some stretching exercises, Chi Kung and Tai Chi before you begin to meditate. These exercises are both invigorating and relaxing; they relieve muscle tensions, stimulate the blood circulation, get rid of stale air in the lungs and oxygenate the blood. As an alternative, you can take a relaxing stroll outside.

Doing some movement before entering the stillness of meditation can help you shake off the sluggishness that comes from being too sedentary and ease the tensions that might build up during a hectic day. Then when you sit down, you are ready to begin. Use the Simple Chi Kung exercises to loosen your spine and prepare the mind and body for meditation.

Wear Loose and Comfortable Clothing

The clothing you wear for meditation should be loose and comfortable. It is very important that your clothes restrict neither the Chi flow, the blood circulation, the nervous system nor the breath.

It is preferable to wear clothing of natural fibers so your body can breathe. During meditation, all the pores open wider and breathing in energy through them requires looser clothing. When you have learned to breathe through the skin, your body will need to contact more fresh air.

Establish a Stable Sitting Position

The body must be stable for the mind to be stable. It has been said, "An anxious mind cannot exist in a relaxed body." The mind and body are clearly connected and when the body and breath become peaceful, the mind easily follows.

If your posture is firm and balanced, it will be easy for you to relax, and you will already be halfway toward achieving a tranquil and focused mind. But if your posture lacks balance and stability, your muscles will soon tire and become tense, your attention will waver like a candle in the wind.

Consider seven points in preparing a good meditation posture: base, hands, spine, shoulders, chin, eyes and tongue.

1. **Base**: Your base is the foundation of a good meditation posture. To accommodate the movement of internal and external forces generated by the Taoist practices we need to stay grounded and establish a good connection with the earth's energy. The more Chi one moves through the body, the more important grounding becomes to prevent overheating of the organs and other negative side effects.

   The soles of the feet provide an ideal connection to the earth through the Yung Chuan (Bubbling Spring) points. These are specifically designed to absorb the earth's energy and conduct it up into the body. The legs also help filter the raw energy to make it more readily digestible.

   It is best to sit on a straight backed chair to practice the Fusion Meditations. Your weight should be evenly divided over four points: your two feet and the two sitting bones (the tuberosities of the ischium) *(Fig. 4.1)*. Place the feet flat on the floor the same distance apart as the hips. The calves of the legs should be vertical, like pillars. Try to have the knees and hips at the same level, or keep the knees slightly higher.

   The part of the pelvis known as the ischium is structurally designed to hold a tremendous amount of weight. Avoid leaning back and sitting on the coccyx and sacrum instead of the ischium. Honor Mother Nature's design by sitting on the ischium and not on the tailbone, which puts pressure on the sacrum, one of the major pumps in the spine for cerebrospinal fluid, a vital cushion for the nervous system. Check to make sure your weight is evenly distributed over the four points to establish a solid base to support the body during meditation.

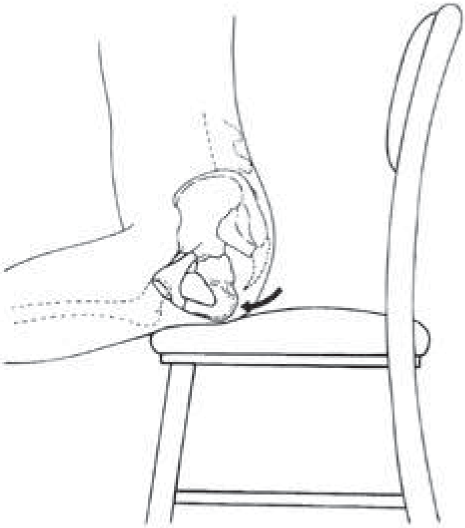

Fig. 4.1 *Sitting Bones*

2. **Hands**: Let your hands rest in your lap, clasped together with the right palm over the left palm and the right thumb and forefinger wrapped around the base of the left thumb. You may rest your hands on a pillow placed on your lap *(Fig. 4.2)*. The clasped hand position works especially well to consolidate and balance the energies generated during the meditation.

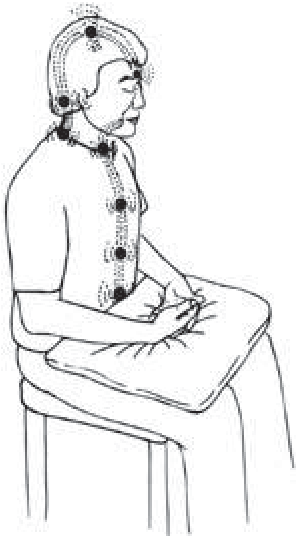

Fig. 4.2 *Using a small pillow helps release shoulder stress.*

3. **Spine**: The spine should be straight but not stiff and in good vertical alignment with gravity. You can imagine your head being pulled up by a string. As it rises, allow the spine to elongate, increasing the space between vertebrae *(Fig. 4.3)*.

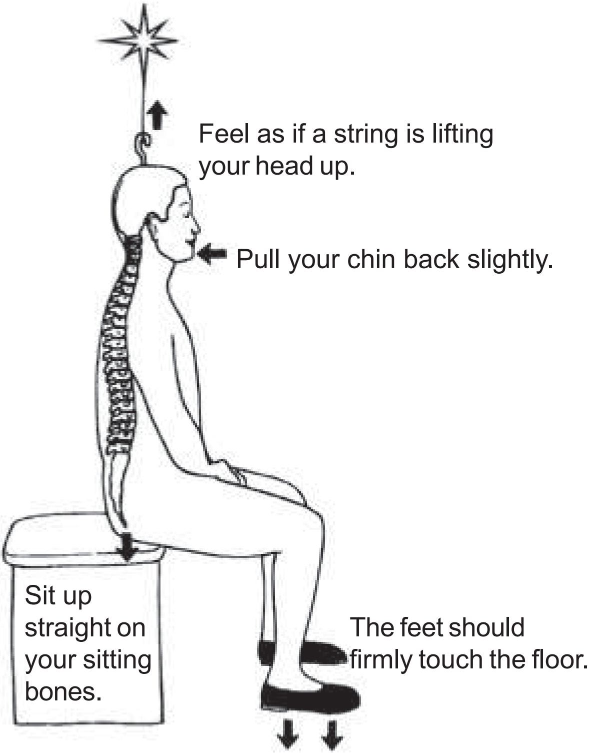
Feel as if a string is lifting your head up.
Pull your chin back slightly.
Sit up straight on your sitting bones.
The feet should firmly touch the floor.

Fig. 4.3 *Correct Sitting Position*

Some people find their backs become tired during long rounds of meditation. Good alignment helps prevent this by taking some of the stress off the muscles and putting it instead on the skeletal structure where it belongs. The skeletal structure is designed to support hundreds of pounds without effort when properly aligned. Practicing Iron Shirt Chi Kung and Tai Chi is very helpful for strengthening the muscles you use in sitting and for learning the body mechanics of good alignment.

The spinal column houses many nerves, and it is also a major part of the Microcosmic Orbit pathway. If the spine feels relaxed, clear and open, the mind will feel more awake and alert too.

4. **Shoulders**: The shoulders should be relaxed and balanced over the hips. The armpits should be slightly open allowing enough space to hold a Ping Pong ball *(Fig. 4.4)*. This permits free circulation of blood and Chi into the arms and keeps the nerves in the arms from being impinged.

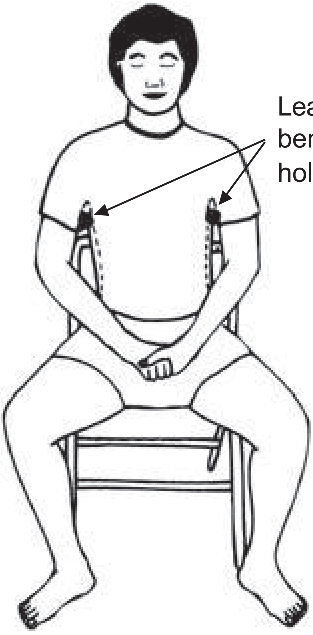

Leave enough space beneath each armpit to hold a Ping Pong ball.

Fig. 4.4 *Arms' Position*

5. **Chin**: The chin should be drawn back slightly, with as little strain as possible, so that the ears are over the shoulders. If you strain too much to bring the head back, your muscles will soon tire.
6. **Eyes**: Your eyes should generally be closed or slightly open with the gaze directed downward. Or you can focus on the nose, and from the nose focus into the heart. You can open the eyes for a while if you feel sleepy or distracted *(Fig. 4.5)*.
7. **Tongue**: The tip of the tongue should be touching the upper palate *(Fig. 4.6)*. This connection acts like a switch in that it connects the Tu Mo and Ren Mo, the Governor and Functional Channels. The best point for you to use is the one where you feel the strongest sensation of Chi.

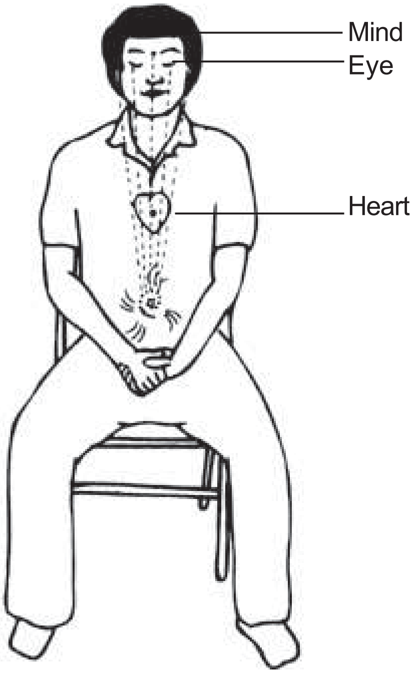

Fig. 4.5 *Turn the mind and eyes inward. Focus into the heart and then down into the navel.*

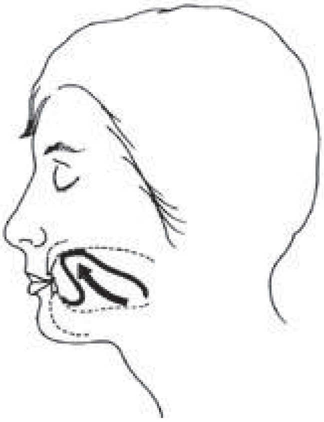

Fig. 4.6 *Touch the tip of your tongue against the upper palate.*

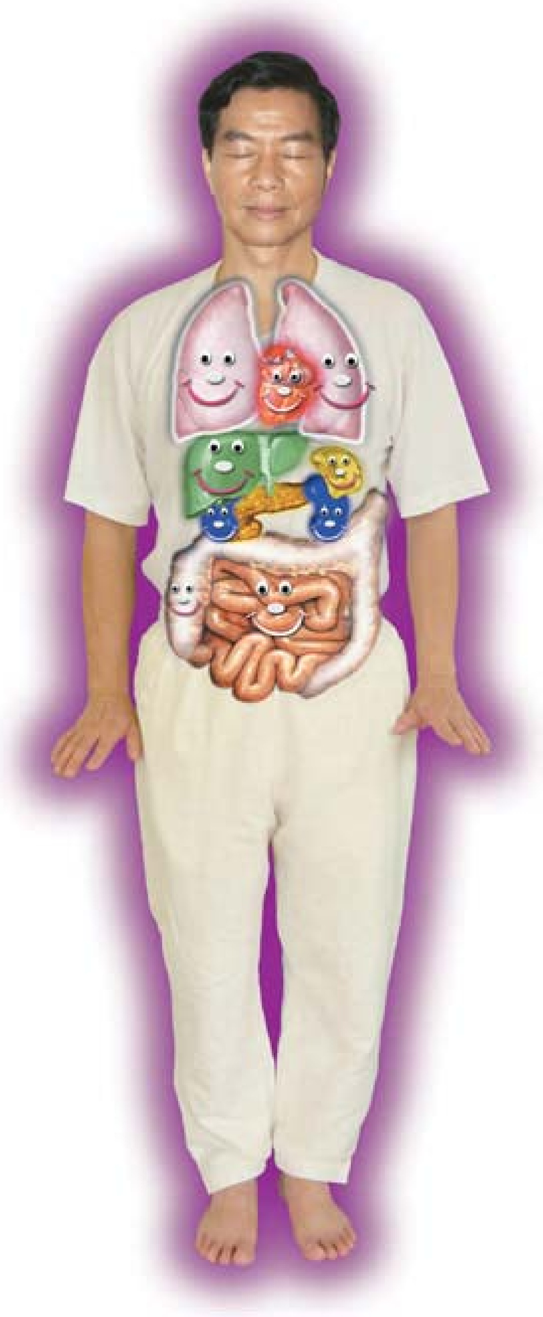

Fig. 4.7 *Smile to the organs and feel the organs smile back to you.*

Chapter 5

Beginning Fusion Element Meditation

Pakua

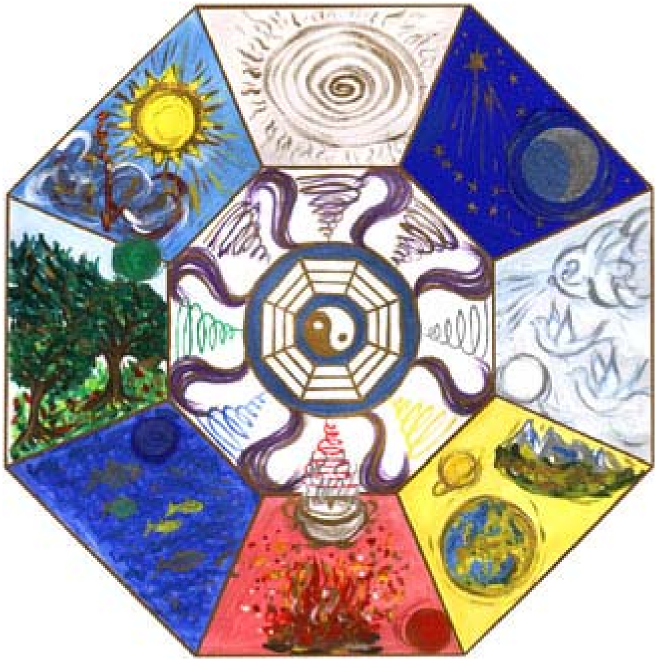

Fig. 5.1 *Pakua - Eight Forces*

The pakua is a symbol of the eight forces. In the Fusion of the Five Elements meditation the pakuas are used to gather and collect the energy of the organs and fuse them into a pearl. Here in the Cosmic Fusion meditation, the pakuas are again used to fuse energy into a pearl. Now, instead of just using the energy within the body, we combine the energy of the organs with the energy of the Universe. This is a technique used in the Fusion practice to condense and gather the abundant energy that surrounds us into something that is usable and digestible to the body. Forming the pakuas will be used in all the Fusion formulas.

The first step is to form the pakua at the lower abdomen to connect with and attract the five energies and the eight forces. The goal is to balance and transform the emotional energies of the organs. The next step is to form the facial pakua. The facial pakua gathers and transforms the energy of the senses and the thoughts down to the lower abdomen. The Universal pakua is then created to gather and collect the energy of the Universe into the lower abdomen. All these energies once gathered are condensed into an energy ball, or a pearl of white light. This highly refined ball of energy is then circulated through all the channels as a way to open, heal and revitalize the body, mind and spirit. This is the beginning of the transference of consciousness to a new realm.

Chanting

Chanting has been used in all systems of spirituality to help connect the practitioner to their higher source. In Christianity, chanting and singing is used to connect with God. In Buddhism and Hinduism, chanting is used to clear the space, to open the body and connect themselves with the Buddha or Hindu God or Goddess. The power of chanting and singing can be seen not only in a religious setting, but in all levels of society as well. Singing and listening to music is a very natural way to move energy.

In the Taoist practice, chanting is used to invoke the forces of the Universe. It is not necessarily about harmony. We want to get depth and power to the vibration to open the lower abdomen and create internal power. When you chant try to feel the vibration coming from the Lower Tan Tien.

Chanting in the Pakua is very similar to the sound in the genes when the replicating strands of DNA cross over. The eight sounds are the song of the crossing of the chromosomes like two serpents entwining and then they making love and one becomes two.

The pakua is a symbolic representation of the forces of nature. By chanting the names of these forces, we are able to make a connection with that force. Also, Yin and Yang are chanted as a way to contract and expand energy. The combination of chanting the forces and Yin and Yang helps to bring the power of the Universe to you as a tuning fork.

Chanting greatly increases the power and the coherence of the Pearl. After toning all eight directions, the forces of the Universe are fused into the Pearl. The Yin and Yang (Tai Chi spines) chanting helps to expand and gather, as well as contract and store the force. This will help create the connection between your internal energy and the energy of the Universe.

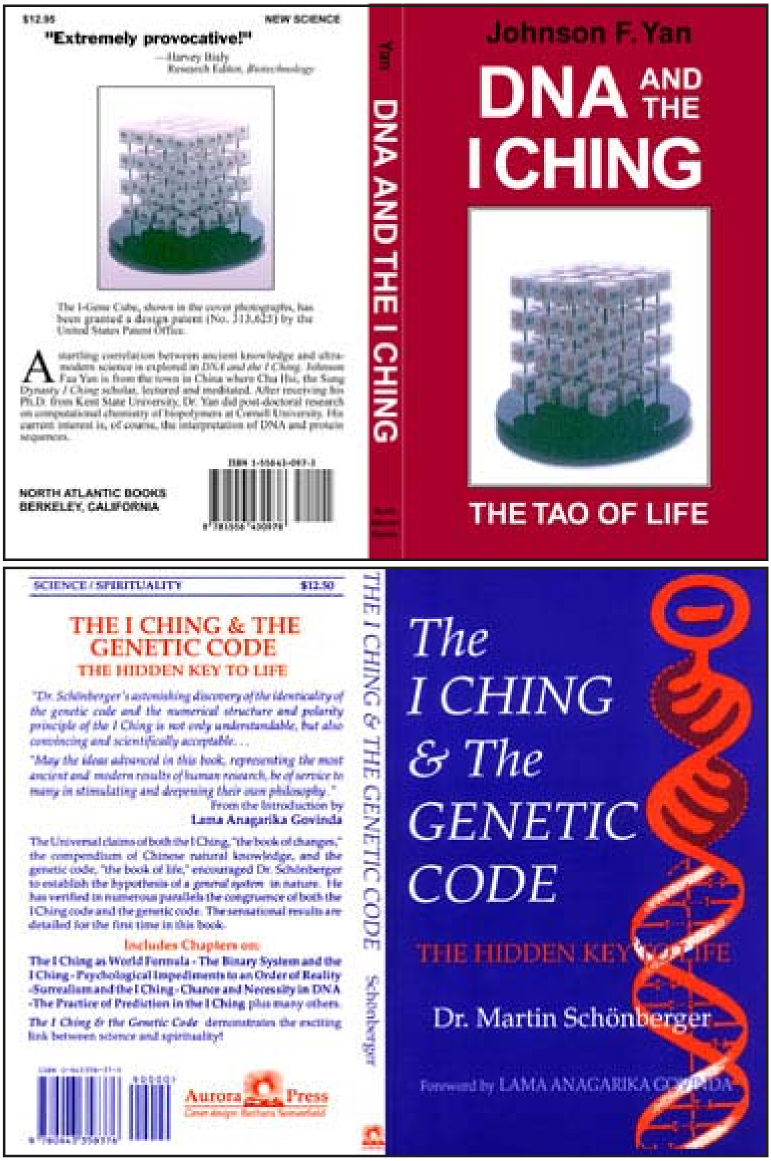

$12.95 NEW SCIENCE

"Extremely provocative!"

—Harvey Bialy
Research Editor, *Biotechnology*

The I-Gene Cube, shown in the cover photographs, has been granted a design patent (No. 313,623) by the United States Patent Office.

A startling correlation between ancient knowledge and ultra-modern science is explored in *DNA and the I Ching*. Johnson Fuu Yan is from the town in China where Chu Hsi, the Song Dynasty *I Ching* scholar, lectured and meditated. After receiving his Ph.D. from Kent State University, Dr. Yan did post-doctoral research on computational chemistry of biopolymers at Cornell University. His current interest is, of course, the interpretation of DNA and protein sequences.

ISBN 1-55643-097-3

NORTH ATLANTIC BOOKS
BERKELEY, CALIFORNIA

Yan DNA AND THE I CHING

Johnson F. Yan

DNA AND THE I CHING

THE TAO OF LIFE

SCIENCE / SPIRITUALITY $12.50

THE I CHING & THE GENETIC CODE
THE HIDDEN KEY TO LIFE

"Dr. Schönberger's astonishing discovery of the identicality of the genetic code and the numerical structure and polarity principle of the I Ching is not only understandable, but also convincing and scientifically acceptable. . .

"May the ideas advanced in this book, representing the most ancient and modern results of human research, be of service to many in stimulating and deepening their own philosophy."

From the Introduction by
Lama Anagarika Govinda

The Universal claims of both the I Ching, "the book of changes," the compendium of Chinese natural knowledge, and the genetic code, "the book of life," encouraged Dr. Schönberger to establish the hypothesis of a *general system* in nature. He has verified in numerous parallels the congruence of both the I Ching code and the genetic code. The sensational results are detailed for the first time in this book.

Includes Chapters on:

The I Ching as World Formula - The Binary System and the I Ching - Psychological Impediments to an Order of Reality -Surrealism and the I Ching - Chance and Necessity in DNA -The Practice of Prediction in the I Ching plus many others.

The I Ching & the Genetic Code demonstrates the exciting link between science and spirituality!

Aurora Press

THE I CHING & THE GENETIC CODE Schönberger

The I CHING & The GENETIC CODE

THE HIDDEN KEY TO LIFE

Dr. Martin Schönberger

Foreword by LAMA ANAGARIKA GOVINDA

***Fig. 5.2** DNA and the I Ching book cover*
and the I Ching and the Genetic Code book cover

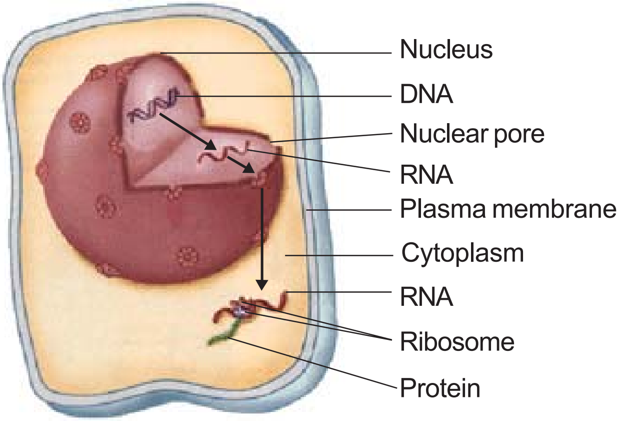
Nucleus
DNA
Nuclear pore
RNA
Plasma membrane
Cytoplasm
RNA
Ribosome
Protein

Fig. 5.3 *When the frequency is 8Hz transcription occurs in the nucleus, translation takes place in the cytoplasm.*

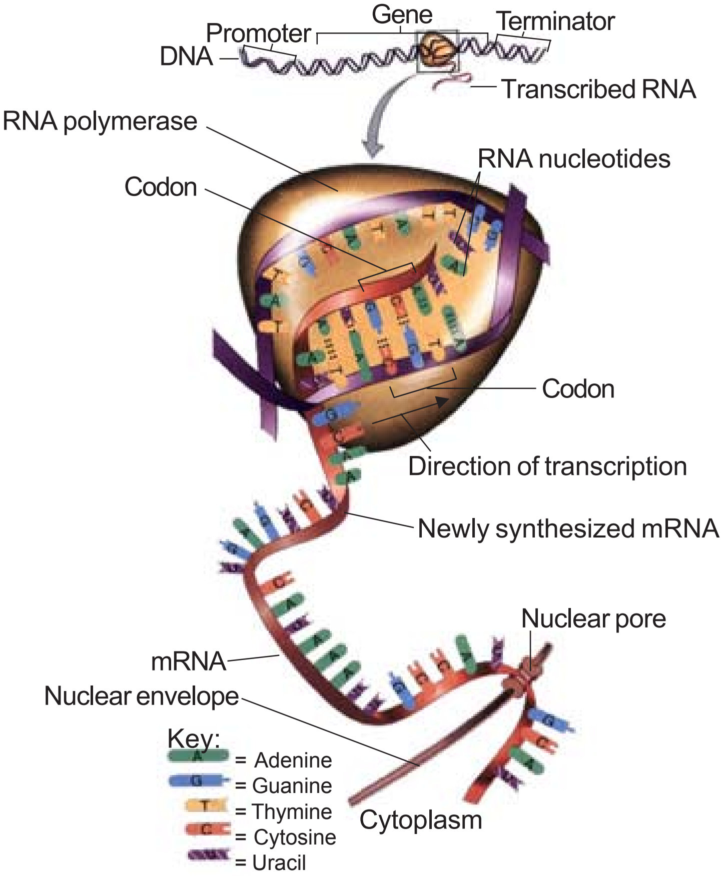
Gene
Promoter
Terminator
DNA
Transcribed RNA
RNA polymerase
RNA nucleotides
Codon
Codon
Direction of transcription
Newly synthesized mRNA
Nuclear pore
mRNA
Nuclear envelope
Key:
= Adenine
= Guanine
= Thymine
= Cytosine
= Uracil
Cytoplasm

Fig. 5.4 *During transcription, the genetic information in DNA is copied to RNA which in the Tao is known as cupping or self-intercourse.*

Formula One: Forming the Pakuas

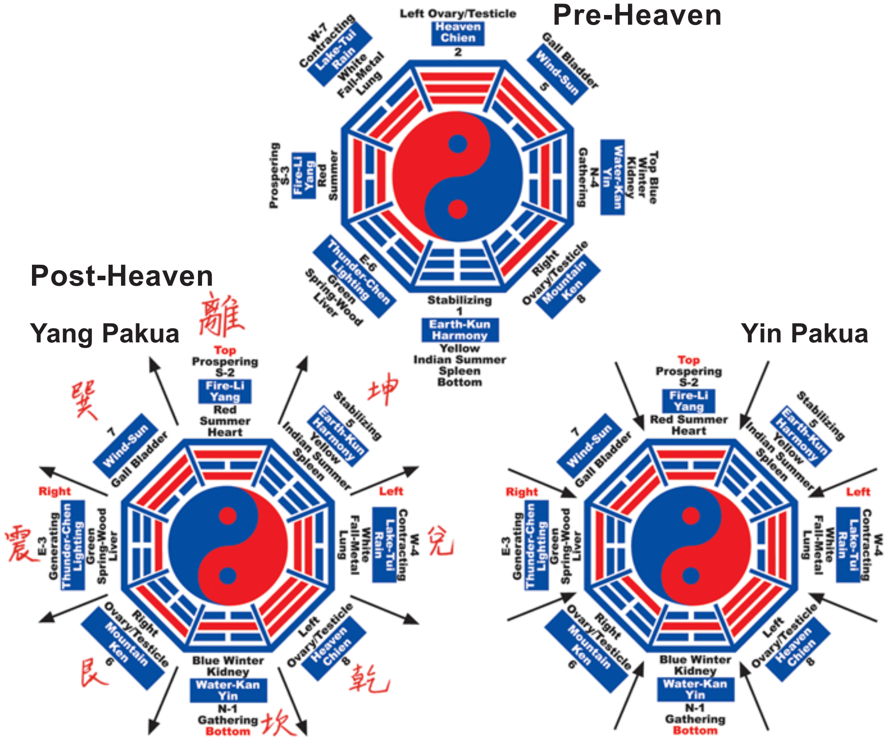
Pre-Heaven
Left Ovary/Testicle
Heaven Chien
2
Gall Bladder
Wind-Sun
5
Top Blue Winter Kidney
Water-Kan Yin
N-4 Gathering
Right Ovary/Testicle
Mountain Ken
8
Stabilizing
1
Earth-Kun Harmony
Yellow
Indian Summer
Spleen
Bottom
E-6
Thunder-Chen
Lighting
Green
Spring-Wood
Liver
Prospering
S-3
Fire-Li Yang
Red
Summer
W-7
Contracting
Lake-Tui
Rain
White
Fall-Metal
Lung
Post-Heaven
Yang Pakua
離
Top
Prospering
S-2
Fire-Li Yang
Red Summer Heart
巽
7
Wind-Sun
Gall Bladder
坤
Stabilizing
5
Earth-Kun Harmony
Yellow
Indian Summer
Spleen
Right
震
E-3
Generating
Thunder-Chen
Lighting
Green
Spring-Wood
Liver
Left
W-4
Contracting
Lake-Tui
Rain
White
Fall-Metal
Lung
兌
Ovary/Testicle
Right
Mountain
Ken
6
艮
Left
Ovary/Testicle
Heaven
Chien
8
乾
Blue Winter
Kidney
Water-Kan Yin
N-1
Gathering
Bottom
坎
Yin Pakua
Top
Prospering
S-2
Fire-Li Yang
Red Summer
Heart
7
Wind-Sun
Gall Bladder
Stabilizing
5
Earth-Kun Harmony
Yellow
Indian Summer
Spleen
Right
E-3
Generating
Thunder-Chen
Lighting
Green
Spring-Wood
Liver
Left
W-4
Contracting
Lake-Tui
Rain
White
Fall-Metal
Lung
Right
Ovary/Testicle
Mountain
Ken
6
Left
Ovary/Testicle
Heaven
Chien
8
Blue Winter
Kidney
Water-Kan Yin
N-1
Gathering
Bottom

***Fig. 5.5** Pakua - Tai Chi Symbol -*
Three Outside Lines - Eight Connecting Lines

There are two arrangements of the pakuas: the Pre-Heaven and the Post-Heaven. It is said that the Pre-Heaven pakua was discovered by Fu Hsi, a legendary figure of early Taoist history, as the Ho-Tu, inscribed, some say on a horse, others say on a dragon that rose from the river Ho to reveal itself to Fu Hsi. The Post-Heaven pakua was revealed to another legendary king, Yü, the Great, as the Lo-shu. This time it was written on the shell (back) of a tortoise that emerged from the river Lo. The Pre-Heaven pakua shows the unchanging cycle movement of these energies, the natural order of the universe and is the foundation of Fu Hsi's divinational trigrams/hexagrams, of the I Ching. The Post-Heaven pakua shows the opposing interaction and movement of these energies, which are responsible for the changes and creation of all the things in the universe, in nature and in our own lives.

Creating the Front Pakua

Drawing the Pakua on the Abdomen: the Front Pakua

It is important to get familiar with visualizing the pakua. Training the mind to concentrate on a single image is of immeasurable value for spiritual practice. The image of the pakua allows the mind to focus and fuse energy. This first step is essential to the rest of the Fusion practice. One of the best ways to get this image deeply connected in the mind is to draw it on a piece of paper and place it over your abdomen. Or, if you are a little more adventurous draw it directly on the abdomen. When you draw the pakua, use blue for Yin, red for Yang and black for the frame. Sometimes it helps to work with a partner. Draw the pakua over your partner's abdomen, and do the meditation facing one another. This way you can glance at the pakua and know what forces to work with. Be creative. Do whatever works so that you get the image planted firmly in your mind.

The pakua is between the underside of the rib cage and above the pubic bone. When we chant the symbols of the trigrams, you can use your finger to draw them on your own abdomen, or while you look at the pakua on your partner.

There are 3 different major centers for generating the frequency which fuse together to form the main energy center of the body. 1. Tan Tien, the Chi Center can be measure like heat in the infrared. 2. Throat Center. 3. Third Eye Center.

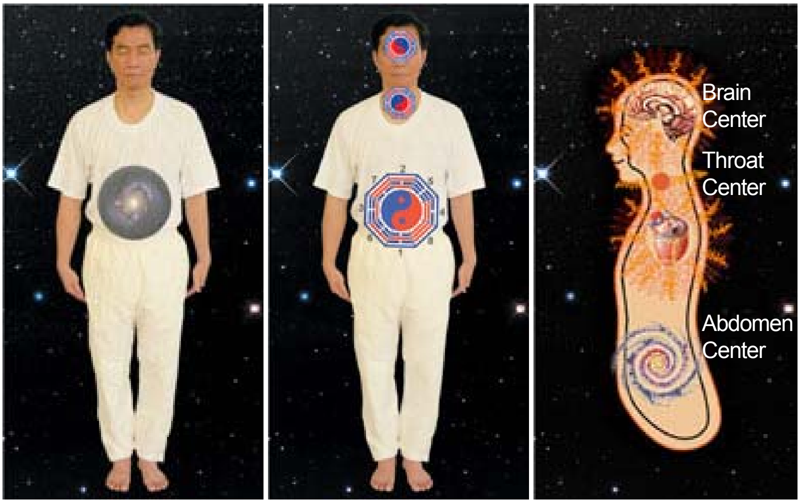

Fig. 5.6 *Pakua in the Abdomen*

Practice

1. Four Signs

Kan: First Sound

a. **Kan** (Yin-Yang-Yin). The power symbol of the Water gathering and yin power, connected with the kidneys, ears and sexual organs. Evoke the energy of the symbol by repeating its sound, **Kan,** several times or until you feel the connection with the energy of Kan, which is spreading down to the sexual organs and the kidneys. The Kan sound should come from the abdomen and the throat.

b. First chant the name of the trigram Kan in a long and deep sound while you picture the symbols and touch your lower abdomen. Your eyes look down to the lower abdominal area.

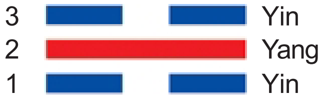

Fig. 5.7 *Kan (Yin-Yang-Yin)*

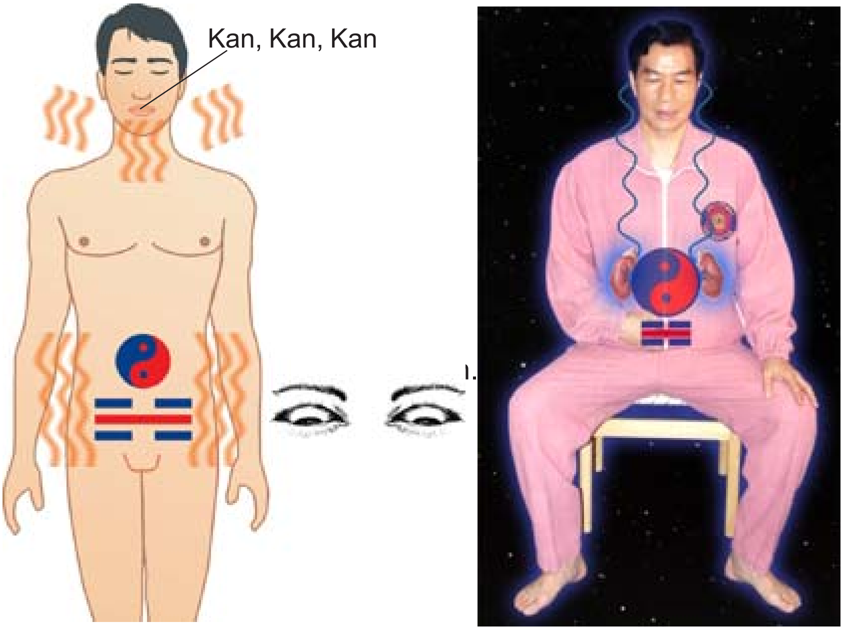

Fig. 5.8 *Chant Kan (Ears and Kidney connects to the Kan)*

c. Next, chant the Yin and Yang lines of the symbol one after the other (Yin-Yang-Yin). Start from the inner side, close to the Tai Chi symbol.

d. **Yin Contract**: First expand the mind and the hands and palms to touch the universe, the cosmic. Chant the Yin, Yin, Yin as long as you feel comfortable, and continue moving your palms and draw the cosmic Chi from the universe into the navel, and feel the navel has suction. Feel your hands and the Tan Tien; the throat has the power of sucking the Chi down the Yin line to the navel. Rest. (Fig. 5.9)

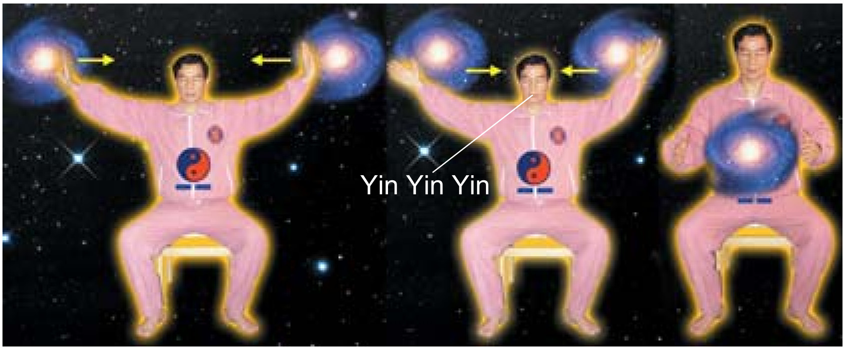

Fig. 5.9 *Yin Line —— Yin Contract*

e. **Do the Yang Line**: Be aware of the Tan Tien, throat, crown and the palms close to the navel and slowly chant the long Yang sound and turn the palms out. Gradually move the palms to the left and right sides and feel your palms very long and big touching the cosmic. Rest and feel your palms in touch with the cosmic Chi. (Fig. 5.10)

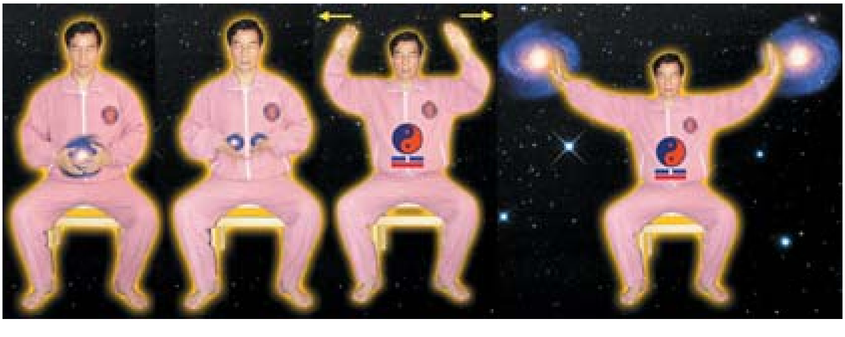

Fig. 5.10 *Yang Line — Yang Expand*

f. **Do the Yin Line again**: Yin contracts; expand the mind and the hands and palms to touch the sky and the cosmic. Chant the Yin, Yin, Yin as long as you feel comfortable and continue moving your palms and draw the cosmic Chi from the universe into the navel, and feel the navel has suction. Feel your hands and the Tan Tien; the throat has the power; the Chi goes down the Yin line to the navel. Rest for a while. Feel the vibrations inside as you chant the trigram. (Fig. 5.11)

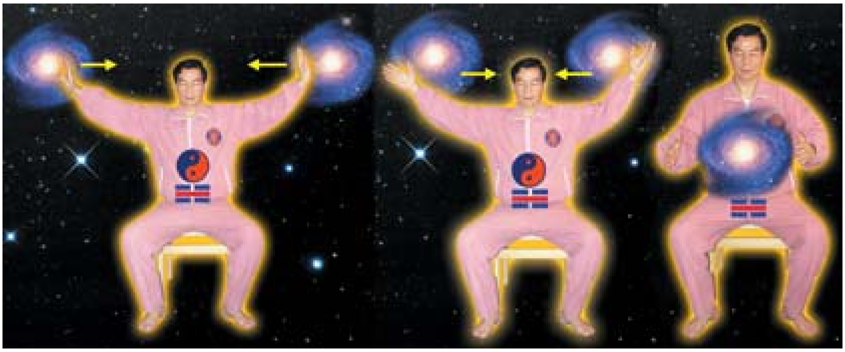

Fig. 5.11 *Yin Yang Yin Trigram* ☵

Rest. Repeat the words " Kan" at the lower abdominal in your mind and let your eyes look down at the lower abdominal. Then you close your eyes and picture the symbols: Kan (Yin-Yang-Yin), smile and be aware the kidneys, ears and the power of the yin and gathering power. Try to imagine them in your mind's eye and feel them imprinted on your abdomen and vibrating inside you.

Rest and project the symbol into the universe and feel the symbol in the universe reinforce the Kan in the abdomen.

Note: Doing the chanting together in a group makes it much more powerful.

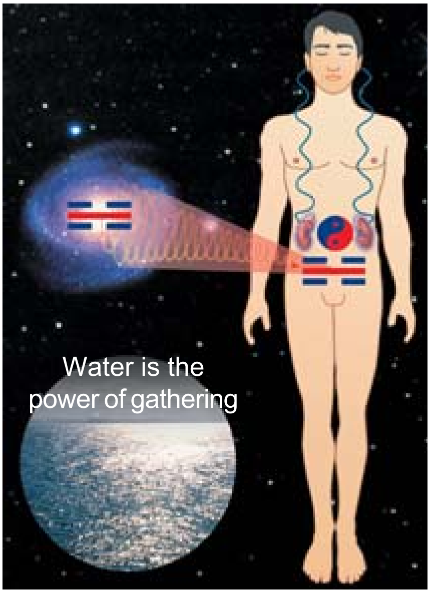

Fig. 5.12 *Kan in the abdomen*

Li: Second Sound

a. The second sound is **Li** (Yang-Yin-Yang), the power symbol of the Fire, connected with the heart, the prospering power. The eyes look up at the upper abdominal, the Li Kua. The sound comes out from the area between the chest and the throat. Repeat this sound until you feel the vibration of the fire expanding upwards in the chest.
When you chant **Li** you feel the energy going up.
When you say **Li** the eyes look up.

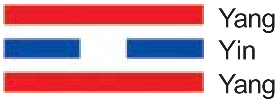

Fig. 5.13 *Li (Yang-Yin-Yang)*

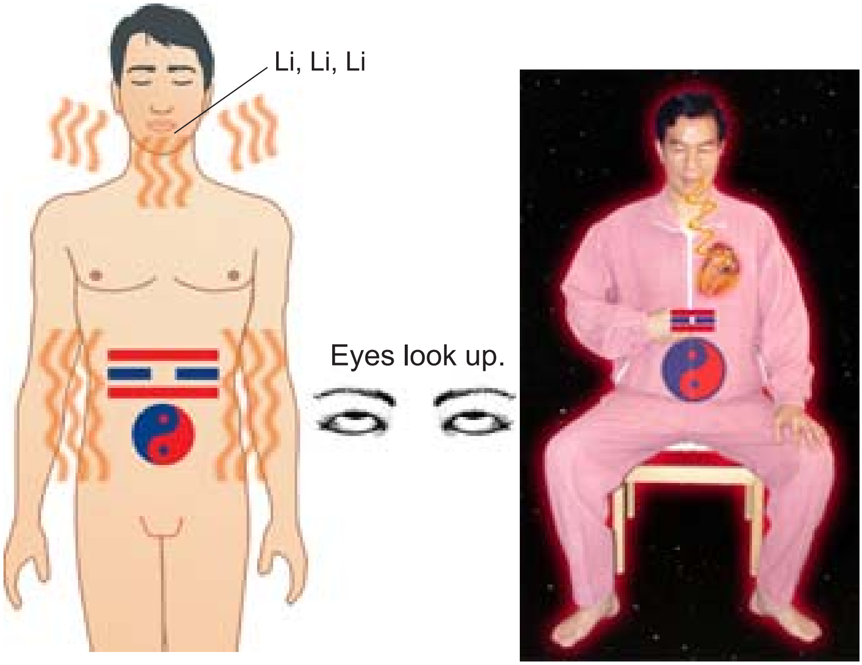

Fig. 5.14 *Chant Li (Tongue and Heart connect to the Li)*

b. Chant the name of the trigrams Li in a long and deep voice while you picture the symbols and touch your lower abdomen and upper abdomen. (Fig. 5.14)

c. Next, chant the Yang and Yin lines of the symbol one after the other (Yang-Yin-Yang). Start from the inner side close to the Tai Chi symbol.

d. **Yang Expand**: Be aware of the Tan Tien, the throat, crown and the palms close to the navel and slowly chant the Yang long sound. Turn the palms out and gradually move the palms to the left and right side. Feel your palms very long and big as they touch the cosmos. Rest and feel your palms touch the cosmic Chi. (Fig. 5.15)

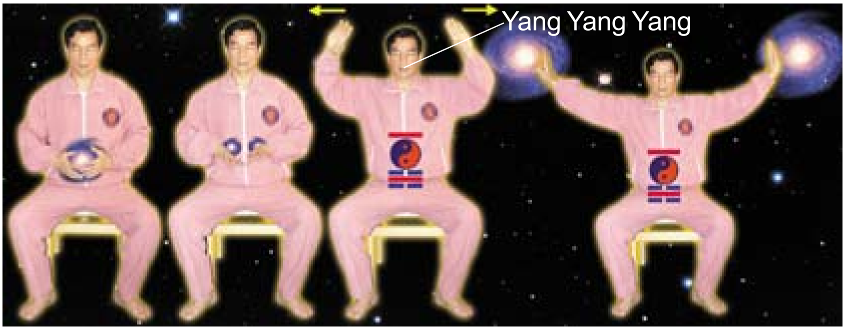

Fig. 5.15 *Yang Line — Yang Expand*

e. **Yin Contract**: Your mind and the hands and palms touch the sky, the cosmos. Chant the Yin, Yin, Yin as long as you feel comfortable, and continue moving your palms up. Draw the cosmic Chi from the universe into the navel, and feel the navel has suction. Feel your hands and the Tan Tien. The throat has the power; draw the Chi into the Yang line to the navel. Rest for a while. (Fig. 5.16)

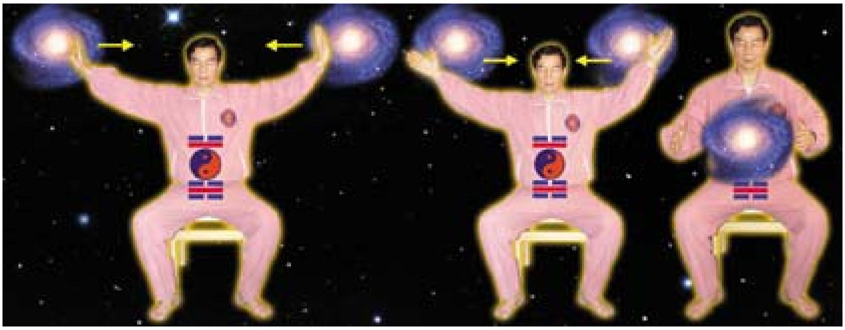

Fig. 5.16 *Yin Line — — Yin Contract*

f. **Do the Yang again**: Be aware of the Tan Tien, the throat, crown and the palms close to the navel and slowly chant the Yang long sound. Turn the palms out and gradually move the palms to the left and right side. Feel your palms very long and big as they touch the cosmos. Rest and feel your palms touch the cosmic Chi. (Fig. 5.17)

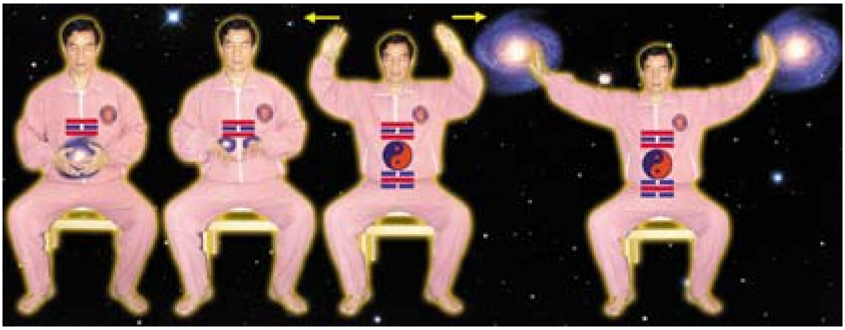

Fig. 5.17 *Yang Line — Yang Contract*

g. **Repeat the word "Li"** in your mind and let your eyes look at your Li Kua. Close your eyes and picture the symbol Li, still moving your eyes up. Smile and be aware of the heart, tongue and the prospering power of fire. Imagine them in your mind's eye; feel them imprinted on your abdomen and vibrating inside you. Rest and project the symbol into the universe and feel the symbol in the universe reinforce the Kan in the lower abdomen.

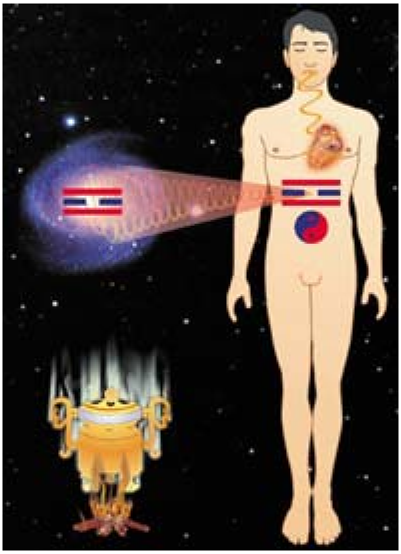

Fig. 5.18 *Li above the Navel*

Chen: Third Sound

a. The third sound is **Chen** (Yang-Yin-Yin), the power symbol of the Thunder and the Lightning, connected with the liver and the eyes, the wood element and gathering power. This sound evokes the trigram Chen that is placed on the right side (below the liver) and the sound comes out as “djen”. Repeat the sound until you feel a vibration of energy below the liver, at the right side of the pakua. When you chant “Chen” the eyes look to the right side.

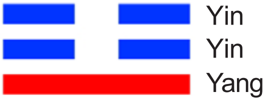
Yin
Yin
Yang

Fig. 5.19 *Chen (Yang-Yin-Yin)*

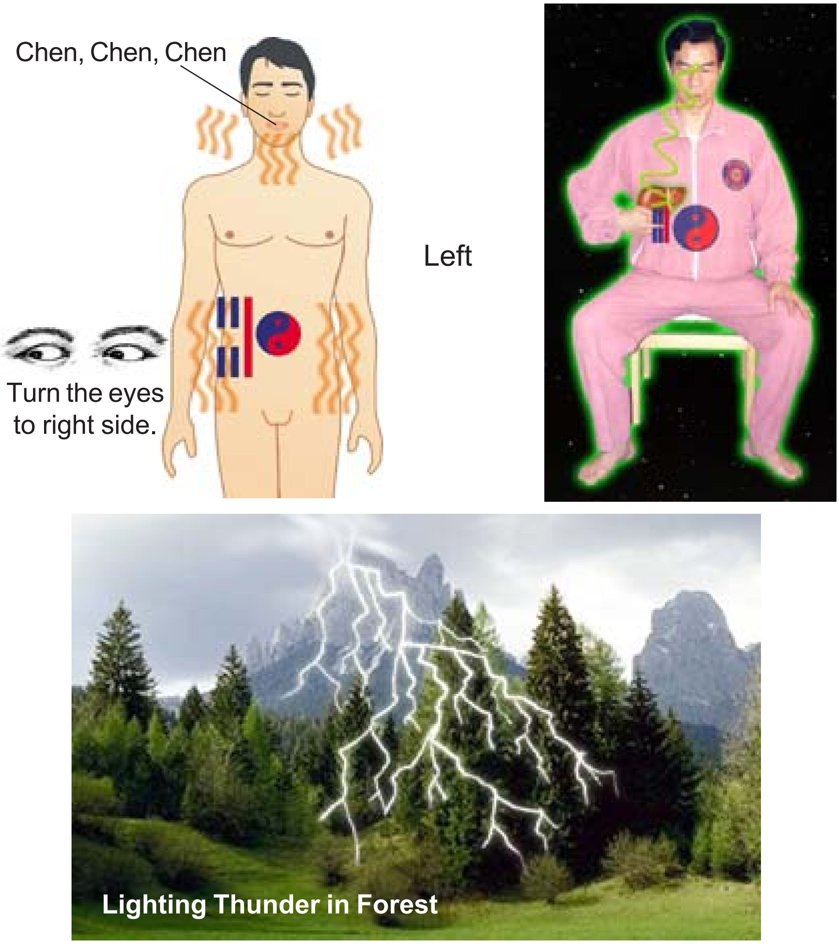
Chen, Chen, Chen
Left
Turn the eyes to right side.
Lighting Thunder in Forest

Fig. 5.20 *Chant Chen (Eyes and Liver connect to the Chen)*

b. **First chant the name of the trigram Chen** ("Djen") in a long deep throat sound while picturing the symbol and touching the right side of your abdomen. Feel the vibration of the lightning and the thunder. Look at the right side of the body, close your eyes and visualize at the symbol of Yang-Yin-Yin. Smile and be aware of the liver, the eyes and the power of generating, lighting and thunder power.

c. **Yang Expand**: Be aware of Tan Tien, the throat, crown and the palms close to the navel and slowly chant the long Yang sound. Turn the palms out and gradually move the palms to the left and right sides and feel your palms very long and big touching the cosmos. Chant the Yang, Yang, Yang as long as you feel comfortable. Rest for a while. Feel the vibrations inside as you chant the trigram. (Fig. 5.21)

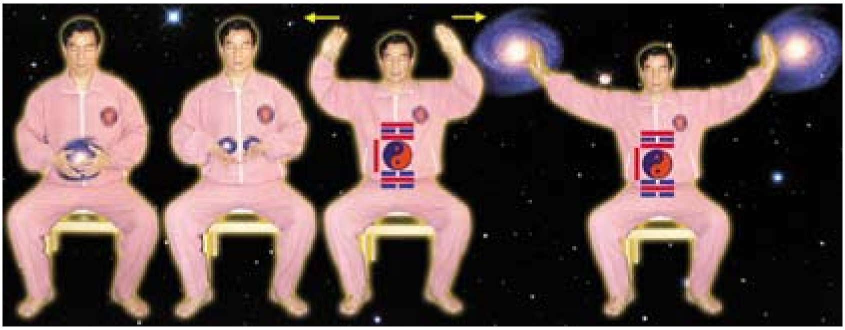

***Fig. 5.21** Yang Line — Yang Expand*

d. **Yin Contract**: Be aware of the Tan Tien, the throat, crown and the palms close to the navel and slowly chant the Yin long sound as the palms touch the universe. Chant the Yin, Yin, Yin as long as you feel comfortable and continue moving your palms back. Draw the cosmic Chi from the universe into the navel and feel the navel has suction. Feel your hands and the Tan Tien; the throat has the power, the Chi goes down the yin line to the navel. Rest for a while. Feel the vibrations inside as you chant the trigram. Rest and feel your palms touch the cosmic Chi. (Fig. 5.22)

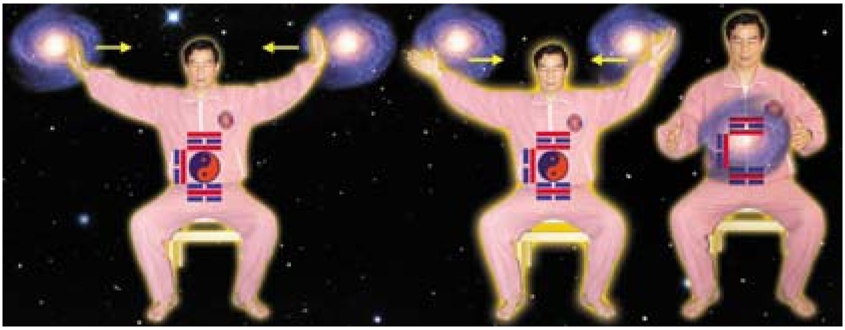

Fig. 5.22 *Yin Line* **—** *— Yin Contract*

e. **Do the Yin line again**: be aware of the Tan Tien, the throat, crown and the palms move out to touch the universe. Chant the Yin long sound. Chant the Yin, Yin, Yin as long as you feel comfortable and continue moving your palms back and draw the cosmic Chi from the universe into the navel and feel the navel has suction. Feel your hands and the Tan Tien; the throat has the power, the Chi goes down the yin line to the navel. Rest for a while. Feel the vibrations inside as you chant the trigram. (Fig. 5.23)
Rest for awhile and feel the vibrations inside.

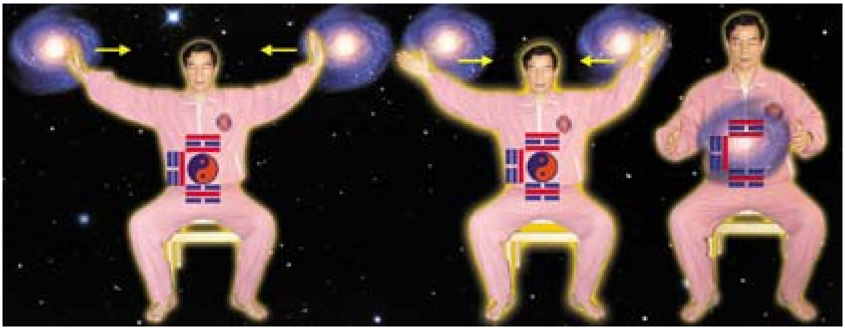

Fig. 5.23 *Yin Line — — Yin Contract*

Tui: Fourth Sound

a. The fourth sound is **Tui** (Yang-Yang-Yin), the power symbol of Lake and Rain, connected with the lungs and nose, the metal element and contracting power. This sound evokes the trigram Tui, that is placed on the left side (opposite Chen) and the sound comes out as “tway”. Chen and Tui are throat sounds. Repeat the sound until you feel a vibration of energy at the left side of the pakua.

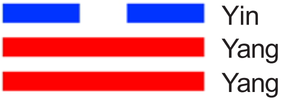
Yin
Yang
Yang

Fig. 5.24 *Tui (Yang-Yang-Yin)*

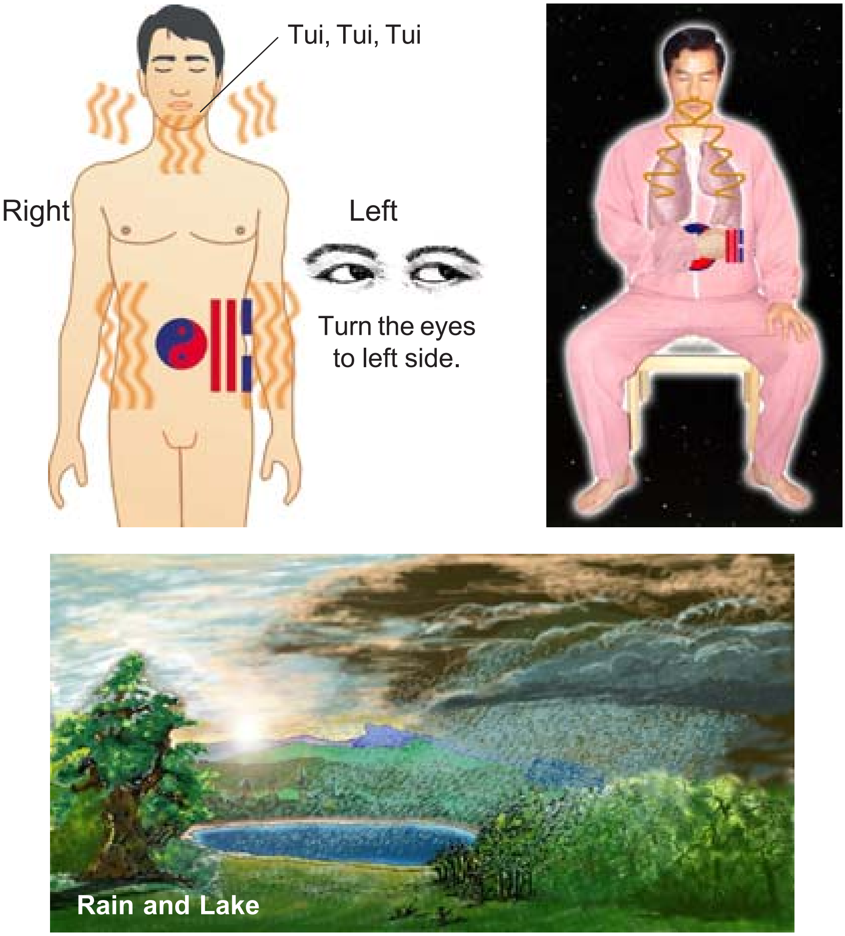
Tui, Tui, Tui
Right
Left
Turn the eyes
to left side.
Rain and Lake

Fig. 5.25 *Chant Tui (Nose and Lung connect to the Tui)*

b. **Next chant the name of Tui** in a long deep voice while picturing the symbol and touching the left side of your abdomen. Feel the vibration of the rain and the lake. Look at your left side, close your eyes and visualize at the symbol of Yang-Yang-Yin. Smile and be aware of the lungs, nose and the power of contracting rain and lake.

c. Do the Yang Line, **Yang Expand**; be aware of Tan Tien, the throat, crown and palms close to the navel and slowly chant the long Yang sound. Turn the palms out and gradually move the palms to the left and right sides and feel your palms very long and big touching the cosmic. Chant the Yang, Yang, Yang. (Fig. 5.26)

d. Do the Yang line again. **Yang Expand**; be aware of Tan Tien throat, crown and palms close to the navel and slowly chant the long Yang sound. Turn the palms out and gradually move the palms to the left and right sides and feel your palms very long and big touching the cosmos. Chant the Yang, Yang, Yang.

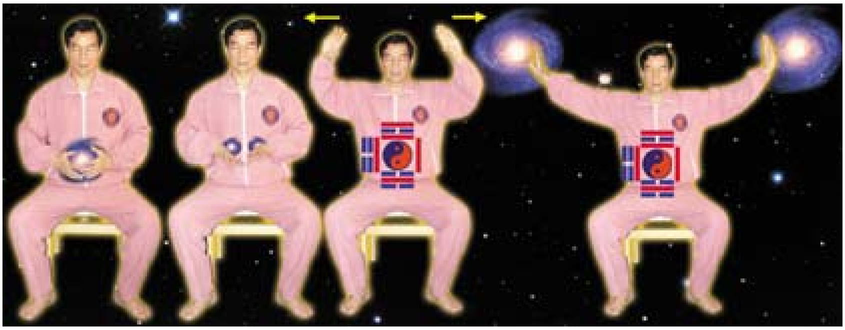

Fig. 5.26 *Yang Line — Yang Expand*

e. **Yin Contract**: Chant the Yin, Yin, Yin as long as you feel comfortable, continue moving your palms and draw the cosmic Chi from the universe into the navel and feel the navel has suction. Feel your hands and the Tan Tien; the throat has the power and the Chi goes down the yin line to the navel. Rest for a while. Feel the vibrations inside as you chant the trigram. Finish. Rest for awhile and feel the vibrations inside. (Fig. 5.27)

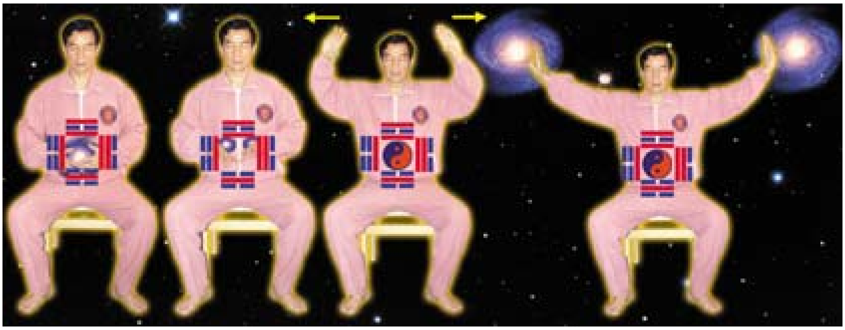

Fig. 5.27 *Yin Line — — Yin Contract*

2. Continue the practice with the next four symbols.

The last four trigrams of the pakua are: **Kun** (Yin-Yin-Yin), the Earth power, **Ken** (Yin-Yin-Yang), the Mountain power; **Sun** (Yin-Yang-Yang), the Wind power and **Chien** (Yang-Yang-Yang), the Heaven power.

Go through the same procedure as with the first four symbols to evoke the energy and connect it to the pakua.

a. Do all the 8 symbols together quickly Kan, Li, Chen, Tui, Kun, Ken, Sun, Chien (6 sets).
b. Be aware of the pakua and see it imprinted deep into the abdomen.
c. Be aware of the Tai Chi symbol and start to chant, Tai Chi, Tai Chi, and Tai Chi faster and faster. Feel the Tai Chi symbol moving faster and faster now.

Practice of the Next Four Symbols.

Symbol Kun - Earth

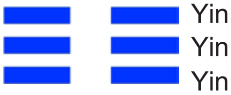

Stabilization power stomach and mouth, spleen and pancreas.

Focus on Kun, Yin-Yin-Yin. Chant Kun and then chant Yin-Yin-Yin repeatedly. You can use the hand to help draw in the energy back to the Pakua. Later on, you do not need to use your fingers. Picture the symbol Kun in front of you and expand it very far away to connect with the centering, stabilizing power. Let the energy with the symbol come to you until it sticks to your forehead. Bring it down to the south/west (upper left corner) in the abdominal Pakua.

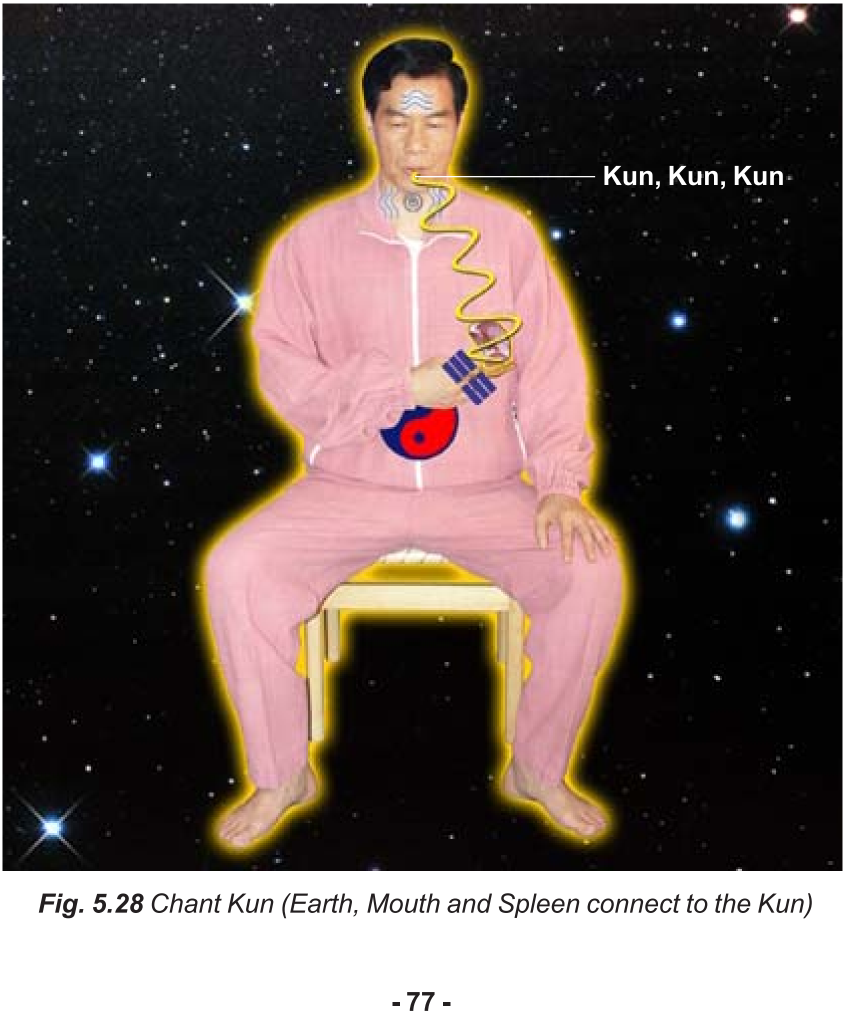

Fig. 5.28 *Chant Kun (Earth, Mouth and Spleen connect to the Kun)*

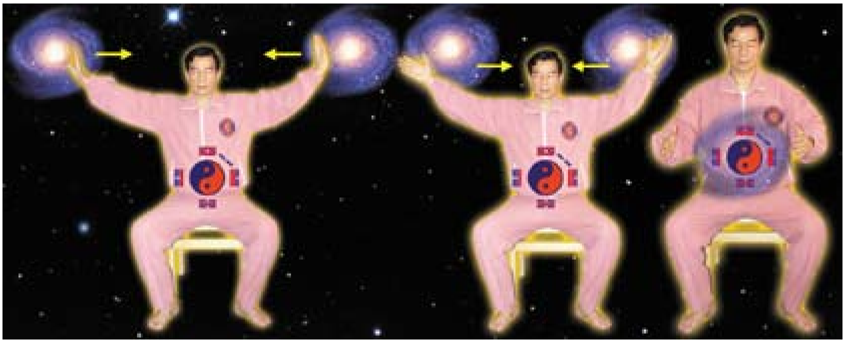

Fig. 5.29 *Yin Line —— Yin Contract*

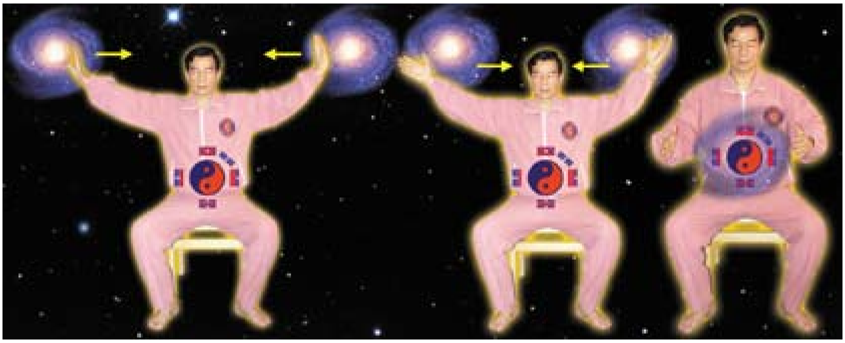

Fig. 5.30 *Yin Line —— Yin Contract*

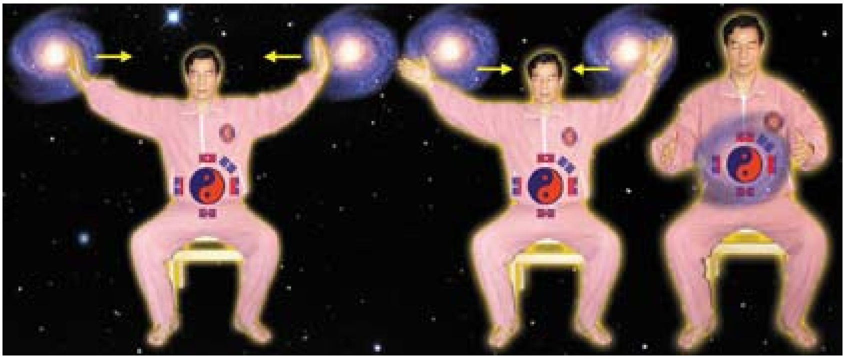

Fig. 5.31 *Yin Line —— Yin Contract*

Symbol Ken - Mountain

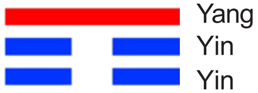

Yin side of water power and with bladder and sexual organs.

Focus on Ken, Yin-Yin-Yang. Chant first Ken and then Yin-Yin-Yang repeatedly. It is a nose/throat sound. Mark the symbol with your fingers and your eyes/mind on your forehead. Rest and feel the symbol imprinted on your forehead. Expand it far away to make the connection with the stable and strong mountain power. Feel the symbol with the energy come back. Move it to its place in the pakua on your abdomen in the lower right corner.

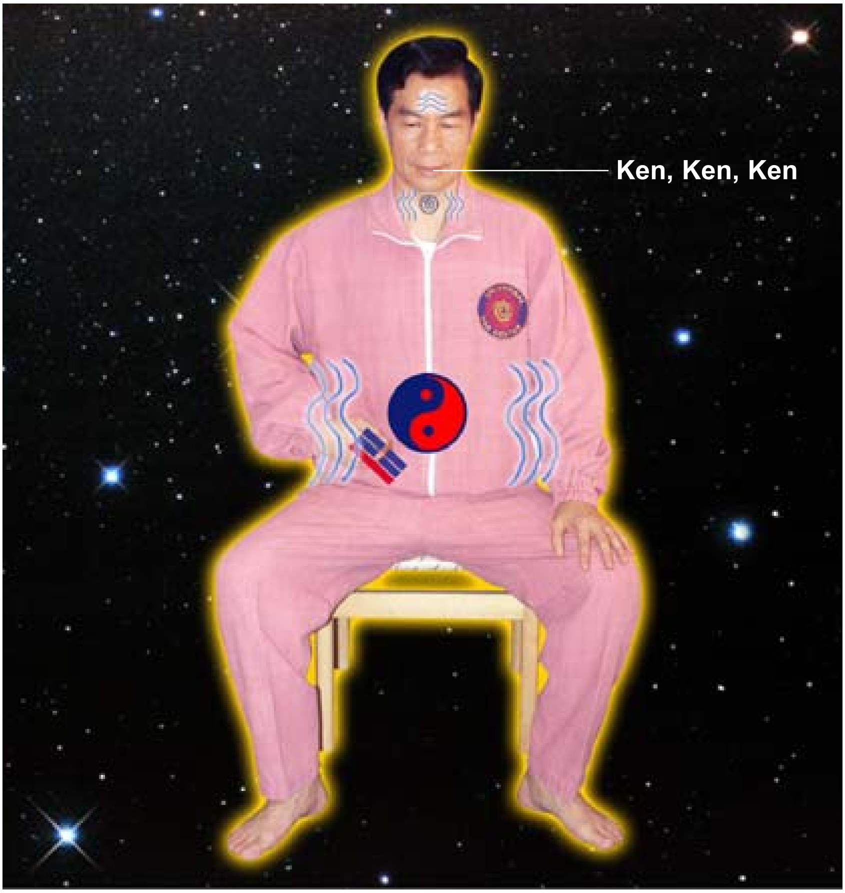

Fig. 5.32 *Chant Ken (Mountain connects to the back of the skull)*

Symbol of Sun - Wind

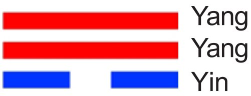

Also connects with the gall-bladder and wood power.

Focus on Sun, Yin-Yang-Yang. Chant Sun and then Yin-Yang-Yang repeatedly. Picture the symbol on your abdomen mark it with your fingers at the abdomen first and then just with your eyes and mind. Picture the symbol very far away to connect with the power of the wind until it comes back to you and sticks to your abdomen. When the reinforced symbol sticks to your abdomen, move it down to the upper right corner of your abdominal Pakua.

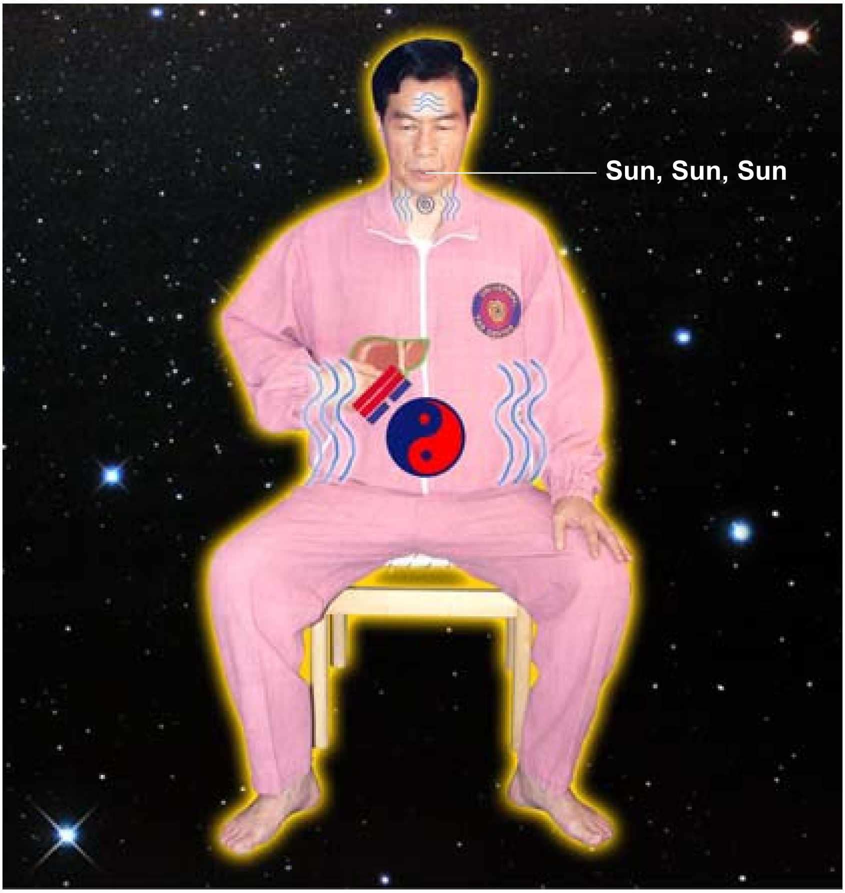

***Fig. 5.33** Chant Sun (Wind connects to the base of the skull)*

Symbol of Chien - Heaven

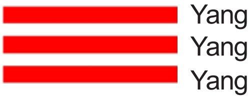

Also connects with the yang side of metal power and large intestine.

Focus on Chien, Yang-Yang-Yang. Chant Chien and then Yang-Yang-Yang repeatedly. Mark the lines of the trigram on the abdomen with your fingers and your eyes/mind. Rest and feel the symbol imprinted on the abdomen. Expand it very far away to connect with the power of Heaven. Do it several times until you feel that the symbol is coming back and sticks on your abdomen. Bring this reinforced trigram down to the lower left corner of your abdominal Pakua.

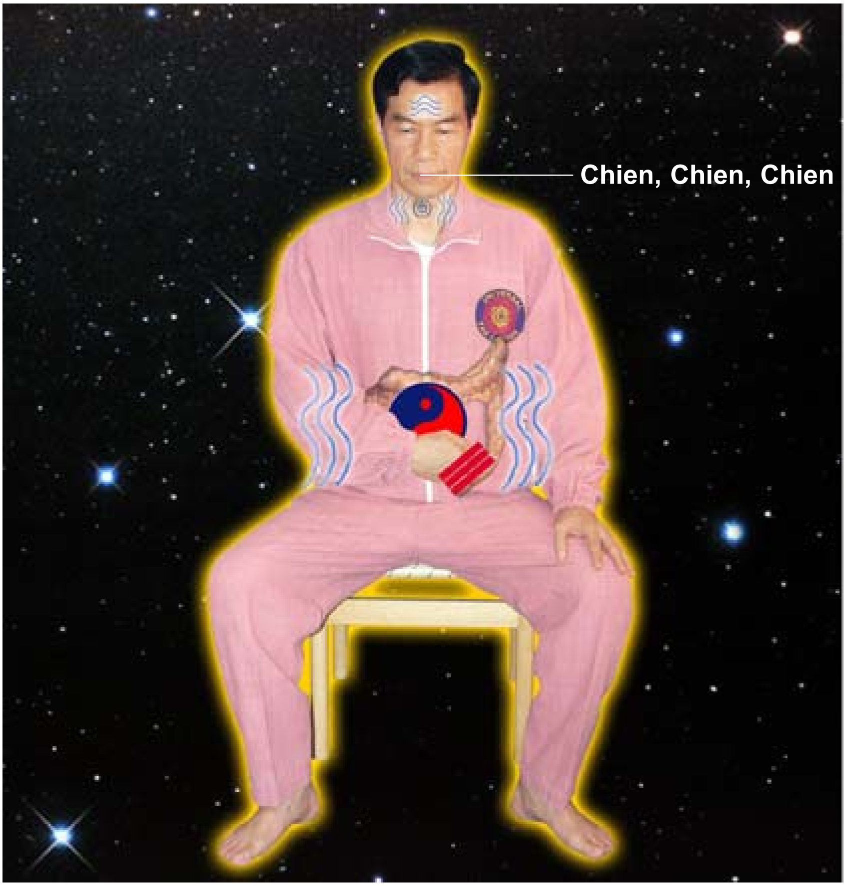

Fig. 5.34 *Chant Chien (Heaven connects to the forehead bone)*

When you finished the last imprinting of these first four symbols, make the Tai Chi symbol, spiraling with the fingers around the navel and saying: **Tai Chi, Tai Chi, Tai Chi**. This is an abdominal sound. Continue to chant it more inwardly, **Tai Chi**, **Tai Chi**, **Tai Chi**, at the same time spiraling with your fingers and eyes. When you finish the chanting, rest. Feel the Kan and the Li, the Chen and the Tui join together, fusing the energy of these forces inside of you.

We now combine the four trigrams.

- First look at your pakua and say the name of the symbol and its Yin-Yang components.
- Chant Kan - Li - Chen - Tui while moving your eyes down-up -right-left. Repeat several times.
- Rest; spiral with your fingers around the navel and continue the spiraling movement of the eyes awhile chanting Tai Chi, Tai Chi, Yin - Yang, Yin - Yang until you feel the Tai Chi symbol spiraling in the middle of the pakua connecting the trigrams together. Feel a big space, a fire burning inside of you.

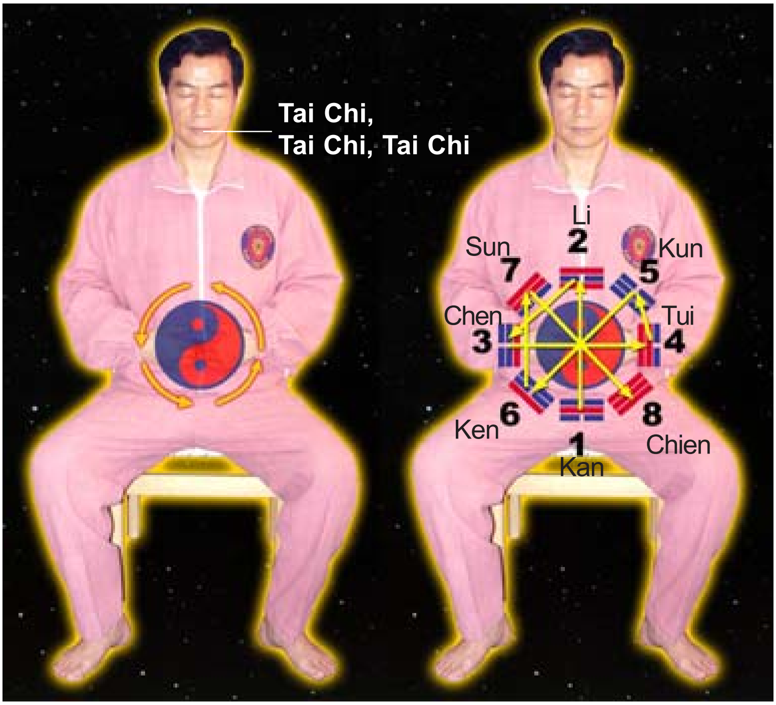

Fig. 5.35 *Finish the imprinting the four symbols and four trigrams.*

Merging Energy of both Pakuas and establish the Cauldron.

a. Copy the front Pakua to the back Pakua and let the Tai Chi symbol spiral. Be aware of the front Pakua spiraling in one direction and the back Pakua in the reverse direction. You can start the front Pakua spiraling counterclockwise and the back clockwise in reverse. Merge the energy of the front and the back Pakua right in the center of the Tan Tien and form a space there for the cauldron. At this point, leave the spinning energy moving. Feel the both Pakuas still spinning and drawing Cosmic Force from all directions. Rest, and then start spinning the energy.

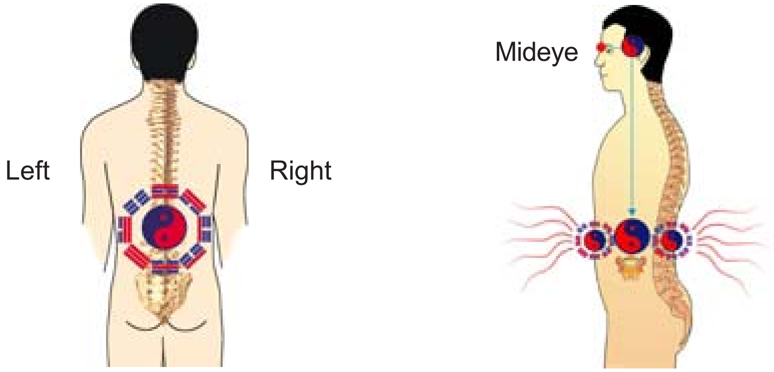

Fig. 5.36 *Back Pakua*

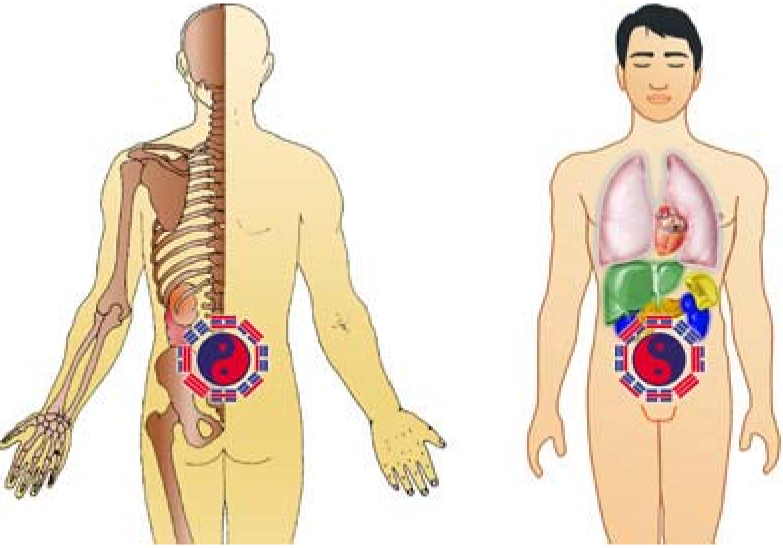

Fig. 5.37 *Front and Back Pakua*

b. When both pakuas are spiraling, focus on the center and look down while being aware of the front and the back pakuas. Chant Yin-Yang and spiral with your mind's eye in the center. Feel the big space of the front pakua and the back pakua coming into the center, merging together and creating a big space inside you for the cauldron. Feel the cauldron inside you spiraling like a big ball of fire, creating a suction in the middle of the front and the back pakua. Feel that, while you continue to chant the Tai Chi subvocally, the energies of both pakuas are sucked into the center, the cauldron. Feel the fire burning in the cauldron like a candle. Be aware of your center and of the energy there.

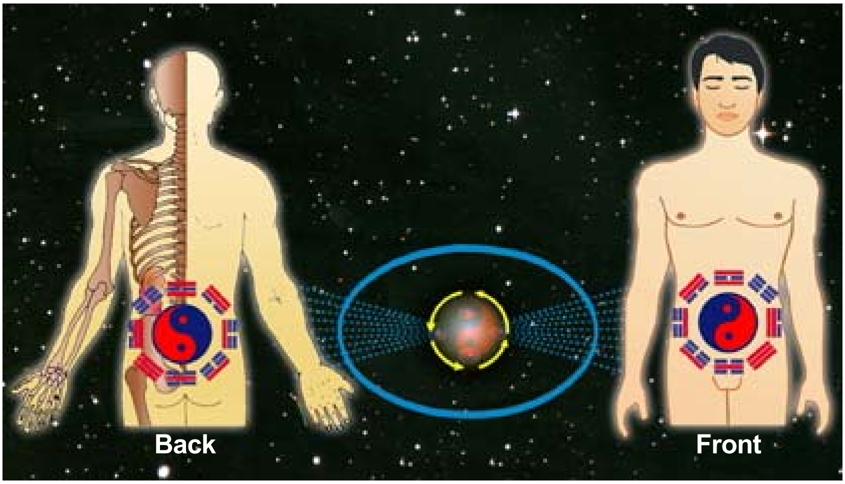

Fig. 5.38 *Both Pakuas are spiraling and drawn into the center.*

1. Creating the Right Pakua

The right pakua is a copy of the back pakua moved to the right side. Use your finger to point to the right side on the same level with the navel while you picture the symbol of Pakua. Then make the Tai Chi symbol, spiraling from the bottom to the right to the top and to the left. Chant Tai Chi and follow the spiraling with your fingers at the same time. Rest, feel the pakua spiraling and drawing energy.

2. Creating the Left Pakua

This pakua is a copy of the front pakua moved to the left side. Use your fingers to point to the left side ao the same level with the navel picture the symbol of the Pakua. When you feel that you are in the rhythm continue in your mind using your mind's eye to move. Chanting the Tai Chi, Tai Chi and spiraling with your hand in the middle of the left pakua. Continue spiraling with your mind/eyes until you feel the energy moving.

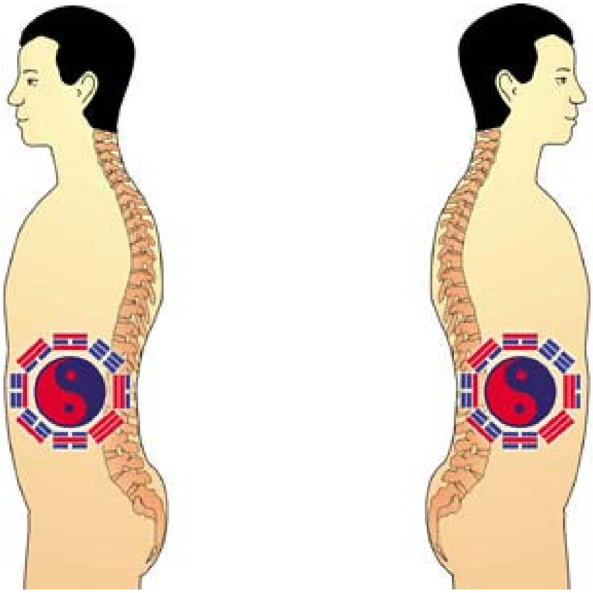

Fig. 5.39 *Left and Right Pakua*

3. Merging the Energies of the Four Pakuas together in the Cauldron

Go back to the right pakua, activate it and let it spiral. Go back to the left pakua and reactivate it, spiraling the Tai Chi. Focus on the center as you chant the Tai Chi and be aware of the centers of the front, back, right and left pakuas. Feel that the energies of the pakuas are sucked into the center. Feel the energy become violet light, and your mind, eyes and center spiraling.

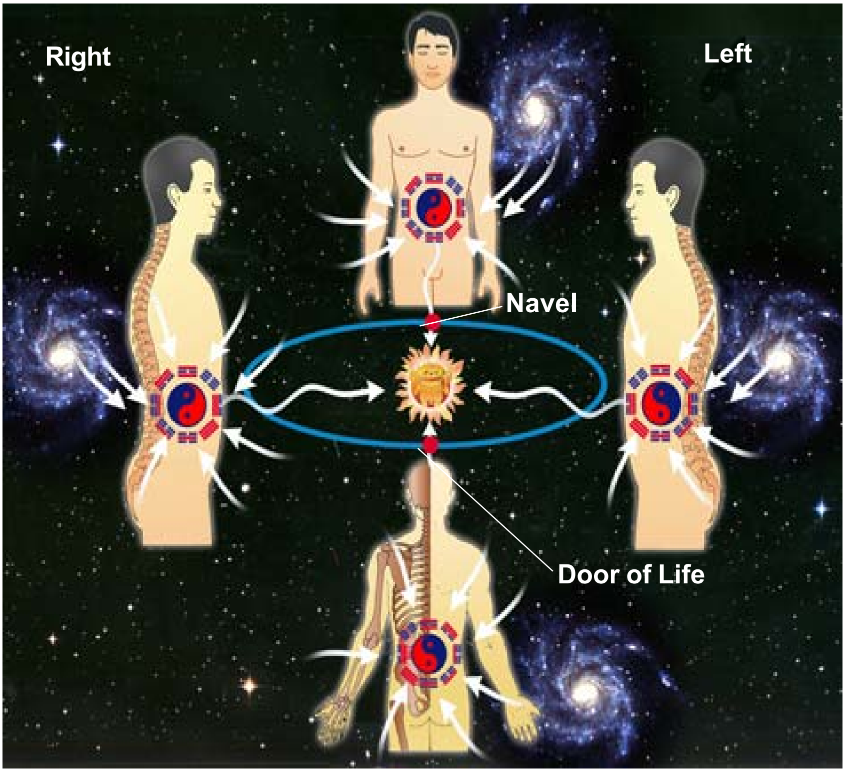

Fig. 5.40 *Four Pakuas spiral and are being drawn into the Center of the Cauldron.*

4. Connecting the Organs and Senses with the Abdominal Pakua.

a. Bring your ears and kidneys to the Kan below.
b. Your tongue and heart to the Li on the top.
c. The mideyebrow and the liver to the Chen on your right side.
d. Your lungs and your nose to the Tui on your left side.
e. Your mouth to the Kun on your upper left side.
f. You can also connect the back part of your skull to the Ken.
g. The base of the skull to the Sun
h. The forehead point to the Chien.

Spiral the Tai Chi symbol of this pakua drawing all the energy of the organs and the senses to the center, where all these energies are combined and condensed.

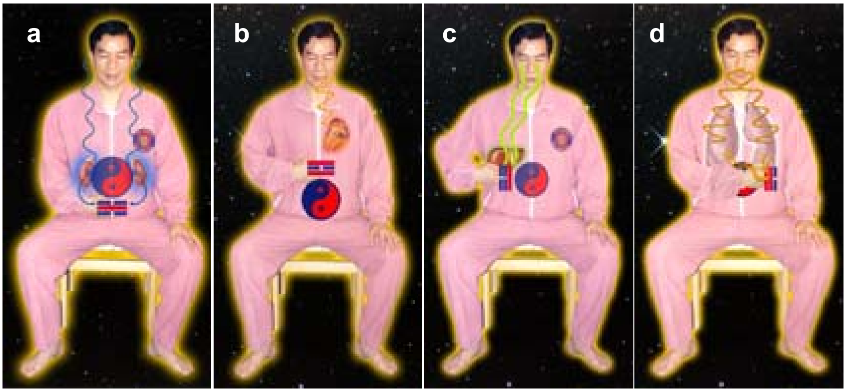

a. *Kan - Ear and Kidneys* ***b.*** *Li - Tongue and Heart*
c. *Chen - Mideyebrow and Liver* ***d.*** *Tui - Lung and Nose on Left Side*

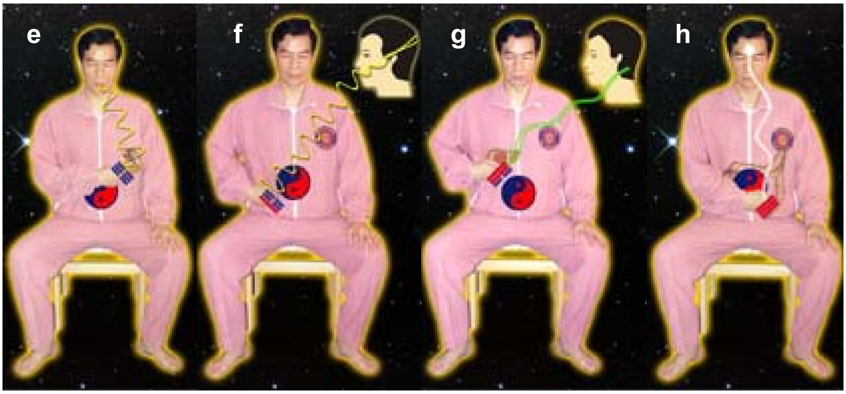

e. *Kun - Mouth and Upper Left Side* ***f.*** *Ken - Back part of the Skull*
g. *Sun - Base of the Skull* ***h.*** *Chien - Forehead Point*
Fig. 5.41 *Connecting the Organs*

Activate the Facial Pakua by enhancing it on the Forehead

The forehead Pakua is the reverse of the abdominal pakua. Use the fingers to point at the position of the Pakua. The Yin and Yang sounds generate from the abdomen but emphasis on the throat, nasal and the third eye. Later on, mark the symbol with your eyes and mind.

When you rest after making the specific sound, feel that the symbol of the trigram is imprinted on the forehead. Send this symbol out into the space in front of you to make the connection with the force of the Universe and let it come back to your body. Bring the reinforced symbol down to the trigram in the pakua in the abdomen and start to work with the next symbol.

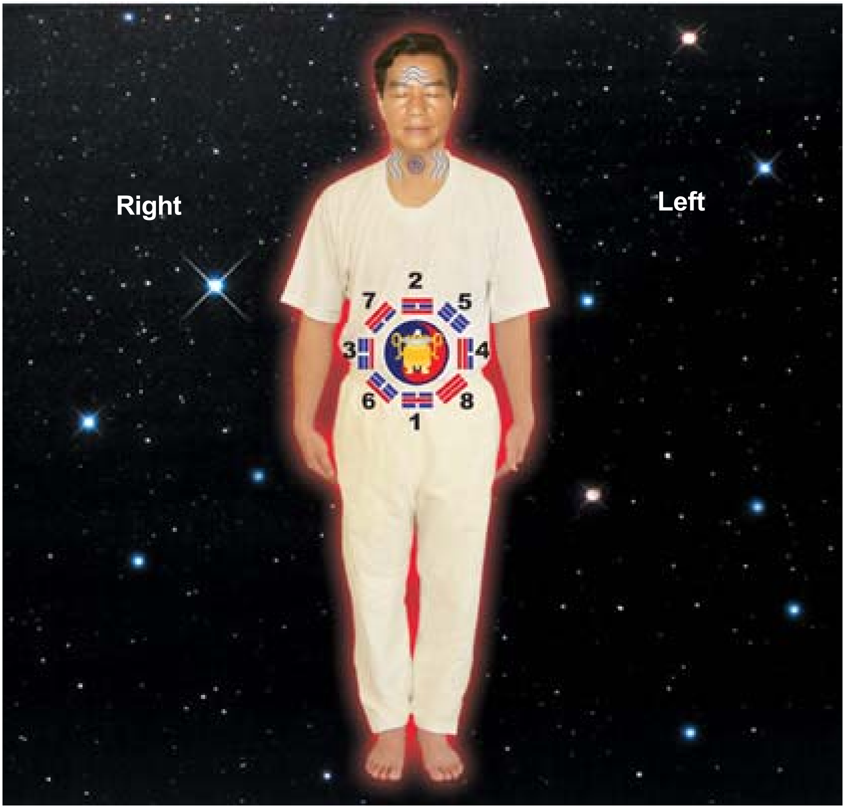

Fig. 5.42 *Enhancing Pakua on the Forehead*

Facial Pakua

To practice this stage, we form a pakua on the face. This pakua has its face in front of us (as if the symbol is looking toward you), so that the Chen will be on the left side and the Tui will be on the right side and all the other trigrams will be arranged accordingly. The Tai Chi symbol will be on your the center at the mid-eyebrow and expands up and down to the forehead and the ridge of the nose right where the cavity of the spirit is. All the senses are controlled from this point. When information and energy from outside is coming in to trigger the senses, you will feel it first very strongly in this center and then the senses will connect immediately to the center of the front pakua below.

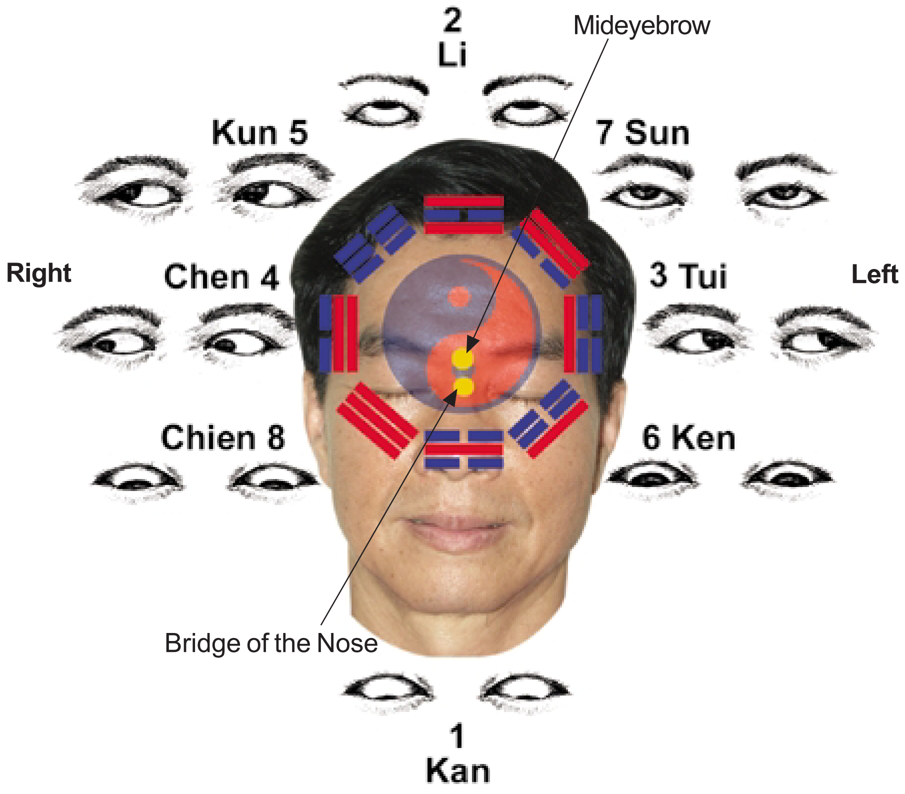

*Fig. **5.43** Use the Eye Mind Power "I" to help move the Eight Forces.*

Practice to form the Facial Pakua.

1. In order to help you get more Chi from the universe, you can start with the "Opening the Three Tan Tiens to the Six Directions" (See the book "*Cosmic Healing I*" by Mantak Chia). If you have not learned this, you can just start with forming the Pakua.

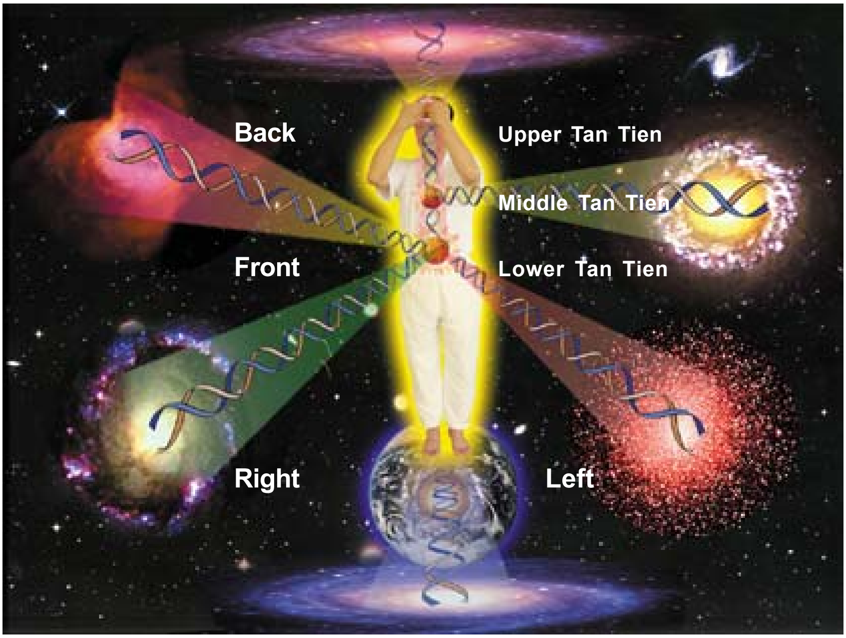

Fig. 5.44 *Opening the Three Tan Tiens to the Six Directions*

2. Then smile, feel your senses turned inward, feel them connecting to the organs. Your mideyebrow connects to your eye and liver, your tongue connects to the heart, your ears connect to the kidneys, your nose connects to your lungs and your mouth connects to the spleen. Feel all the senses turned inward to the mideyebrow and down to navel, going deep inside you. Your abdomen is like a big ocean, a big space, big like a whole universe.

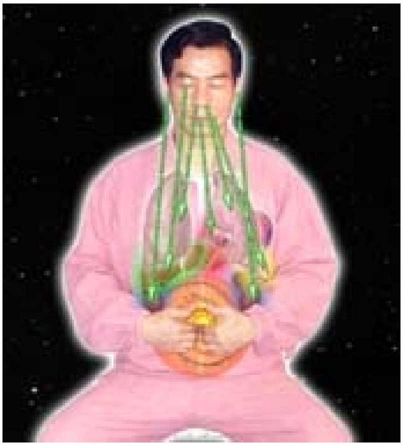

Fig. 5.45 *Turn all senses inward*

3. Reactivate and reinforce the pakua in your lower abdomen.
   a. Touch your Lower Tan Tien and start chanting Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien touching each of the kuas as you chant them. Continue to chant the kua symbols subvocally in your mind and holding the symbol in front of your eyes. Be aware of the Heart Pakua.

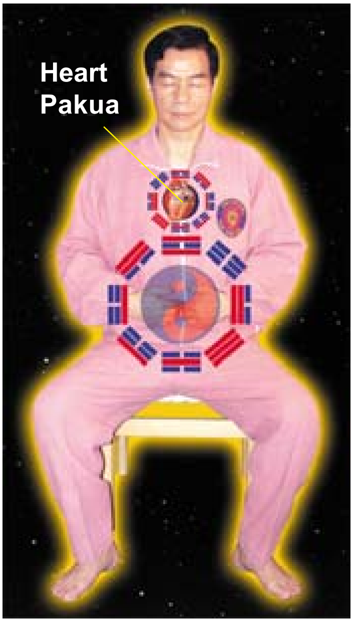

Fig. 5.46 *Touch Lower Tan Tien and chant.*

   b. Spiral the Tai Chi symbol from the bottom to the right and up, chanting it while spiraling with your hand, eyes and mind. Continue to spiral with your mind and eyes only and feel the heart pakua also spiraling. The energy is pulling inward and the Tai Chi symbol is spiraling and changing to violet light.

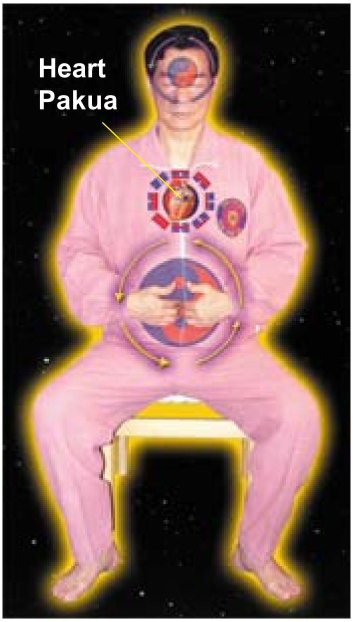

Fig. 5.47 *Tai Chi symbol spiraling*

Start the Practice of the Facial Pakua with the First Four Symbols.

Symbol of Kan - Water ☵

a. Be aware of the mideyebrow. Use the fingers of either the right or left hand to touch below your nose. Eyes look down to the bridge of the nose. Focus on the Kan, Yin-Yang-Yin. Chant the word Kan repeatedly, vibrating it in the brain, making a throat, nose and third eye sound. Picture the trigram on your face and expand it very far away to the ocean. Feel that the energy of the Kan comes right in front of you, allowing the feeling to expand on bridge of the nose. Be aware of the Kan in your lower abdomen and move the powerful Kan from your face to your lower abdomen where it is supposed to be.

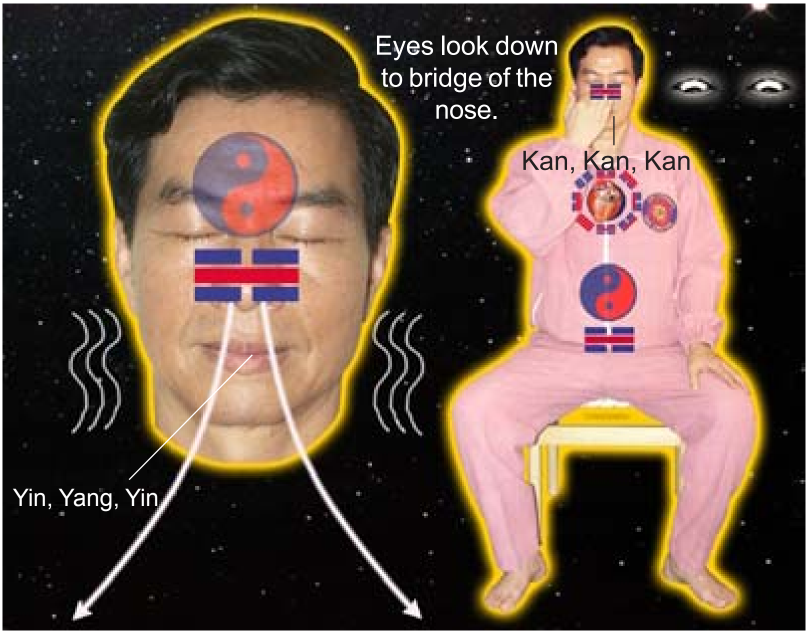
Eyes look down to bridge of the nose.
Kan, Kan, Kan
Yin, Yang, Yin

Fig. 5.48 *Kan on the face below the nose.*

b. **Yin Contract**; First expand the mind and the hands and palms to touch the sky, the cosmos. Chant the Yin, Yin, Yin as long as you feel comfortable, and continue moving your palms. Draw the cosmic Chi in from the universe into the face and feel the mideyebrow has suction. Feel your palms, fingers and the Tan Tien, the mideyebrow has power, the fingers can move in and touch below the nose. The Chi goes to the Yin line in the navel. Rest and Smile, breathing the cosmic Chi. (Fig. 5.52)

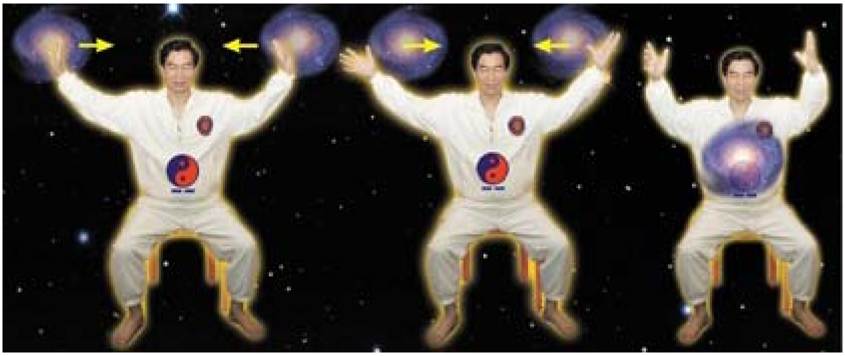

Fig. 5.49 *Yin Line — — Yin Contract*

c. **Yang Expand**: Do the Yang line; be aware of the Tan Tien, throat, bridge of the nose, the palms, the fingers and touch the face. Slowly chant the long Yang sound and turn the palms out. Gradually move the palms to the left and right sides and feel your palms very long and big touching the cosmos. Rest and smile, the Chi goes to the Yang line in the navel.

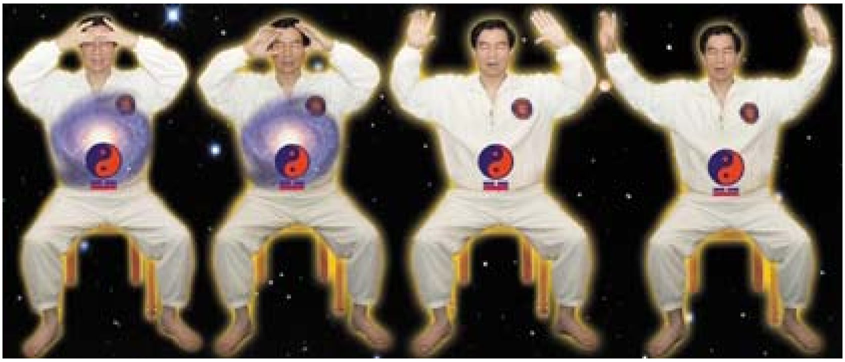

Fig. 5.50 *Yang Line —— Yang Expand*

d. **Yin Line again**, Yin contracts; expand the mind. Repeat as in the first Yin line. Rest for awhile. Feel the vibrations inside.

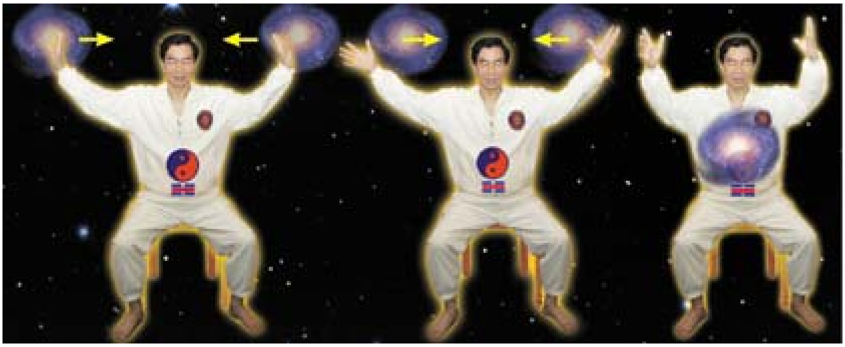

Fig. 5.51 *Yin Line — — Yin Contract*

Symbol of Li - Fire ☲

a. Focus on the Li Yang-Yin-Yang trigram. Repeat it several times using your fingers on your mideyebrow. Eyes look up to the forehead and picture a flame burning and chant Li, Li, Li repeatedly. Then chant the Yang-Yin-Yang. Picture the trigram right on your face and expand it very far away. See a fire burning, expanding warmth. Let the energy stay on your forehead as a bright red light. Bring it down to the Li in the Pakua in your abdomen and feel a burning sensation in that area.

b. **Yang Expand:** Do as before.

c. **Yin Contract:** Do as before.

d. **Yang Expand:** Do as before.

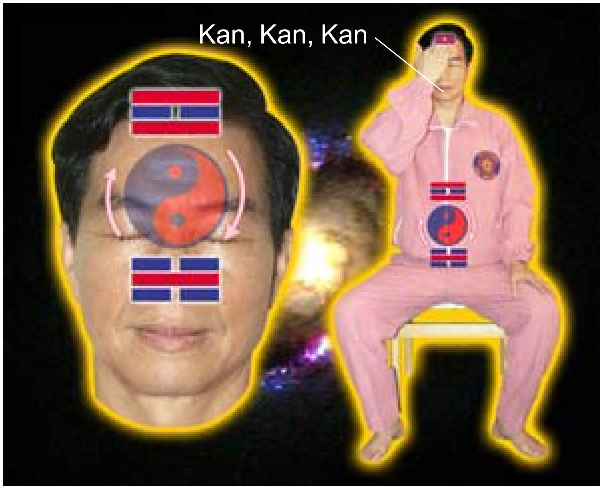

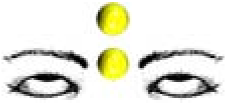

Eyes look up
to the forehead.

Fig. 5.52 *Li on Forehead*

Symbol of Chen - Thunder ☳

a. Focus on Chen, Yang, Yin, Yin. Chant this symbol as a throat and nose sound repeatedly and then chant Yang-Yin-Yin. The eyes look to the left temple bone. Picture the symbol in front of you, expand it very far away to the power of the lightning and thunder. Let the energy come to you, feel the tingling on your left temple bone and move this energy sensation down to the left side of the abdominal pakua. Be aware of Kan, Li, Chen. Visualize them very clearly.

b. **Yang Expand:** Do as before.

c. **Yin Contract:** Do as before.

d. **Yin Contract:** Do as before.

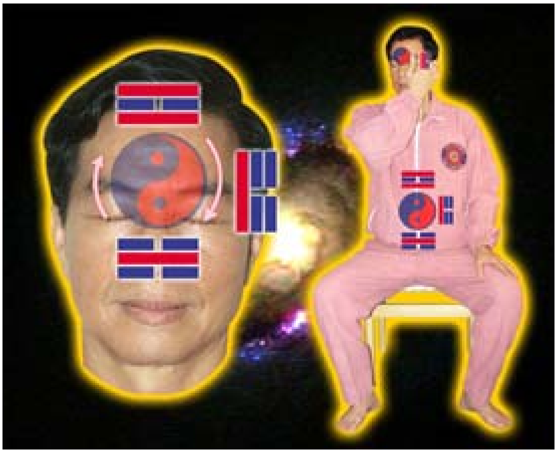

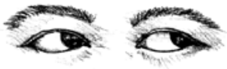

Eyes look to left
temple bone

Fig. 5.53 *Chen on Forehead*

Symbol of Tui - Lake ☱

a. Focus on Tui, Yang, Yang, Yin. Chant Tui repeatedly with a nose sound that is vibrating in the nose. Then chant Yang-Yang-Yin; eyes look to the right temple bone. Picture the symbol on your right temple bone and expand it very far away to the power of the lake, rain and metal element. It is a condensing power. Move it down to the right of the abdominal pakua.

Combine the four symbols together with the Tai Chi symbol. Feel the Tai Chi symbol spiraling inside your head and feel the center of your abdomen nice and warm like a fire burning inside you.

b. **Yang Expand:** Do as before.

c. **Yang Expand:** Do as before.

d. **Yin Contract:** Do as before.

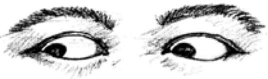

Eyes look to right
temple bone.

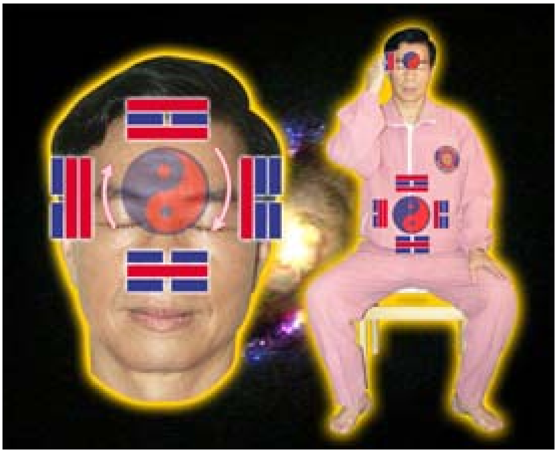

Fig. 5.54 *Tui on Forehead*

Kun ☷

Imprint Kun on your forehead. Start chanting Kun-Kun-Kun, and then use the fingers and palms to help activate the cosmic force Yin-Yin-Yin, Yin-Yin-Yin, marking the lines with your fingers on your forehead. Project the symbol into the universe as you breathe through your forehead. Picture the symbol breathing in violet/blue and when it connects to your forehead bring it down to its place in the lower pakua.

a. **Yin Contract**, Do as before.
b. **Yin Contract**, Do as before.
c. **Yin Contract,** Do as before.

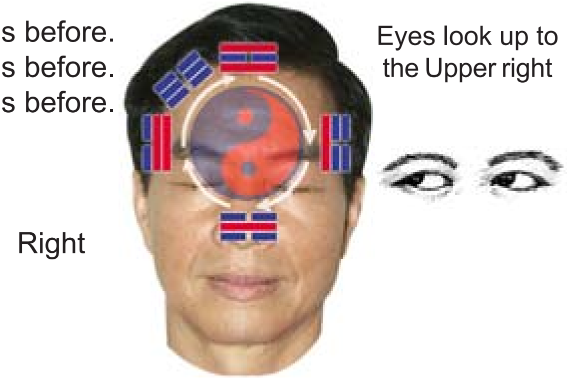
Eyes look up to the Upper right
Right

Fig. 5.55 *Kun*

Ken ☶

Imprint Ken on your lower left cheek bone. Start by the eyes looking down to the lower left cheek bone, chanting Ken-Ken-Ken, and then Yin-Yin-Yang, Yin-Yin-Yang, Yin-Yin-Yang, marking the lines with your fingers on your forehead. Project the symbol into the universe as you breathe through your forehead. Picture the symbol with its violet/blue and bright red colours in front of you and when it connects to your forehead bring it down to its place in the lower pakua.

a. **Yin Contract:** Do as before.
b. **Yin Contract:** Do as before.
c. **Yang Expand:** Do as before.

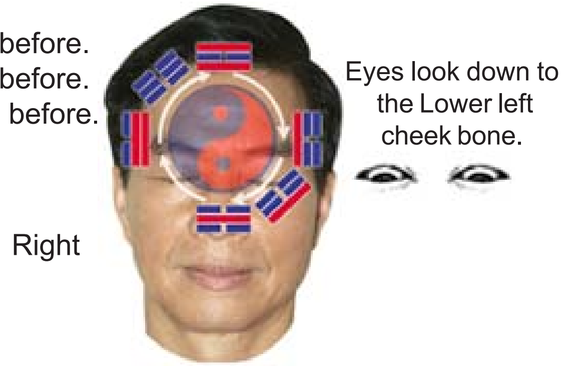
Eyes look down to the Lower left cheek bone.
Right

Fig. 5.56 *Ken*

Sun ☴

Imprint Sun on your upper left eyebrow. Start with the eyes looking up to the upper left eyebrow, chanting Sun-Sun-Sun, and then Yin-Yang-Yang, Yin-Yang-Yang, Yin-Yang-Yang, marking it with your fingers on your forehead. Project the symbol into the universe as you breathe through your forehead. Picture the symbol with its violet/blue and bright red colors in front of you and when it connects to your forehead bring it down to its place in the lower pakua.

a. **Yin Contract:** Do as before.
b. **Yang Expand:** Do as before.
c. **Yang Expand:** Do as before.

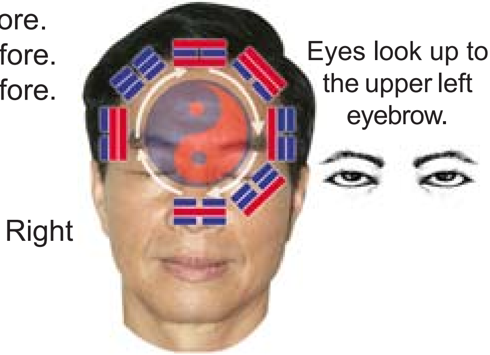

Eyes look up to the upper left eyebrow.

Fig. 5.57 *Sun*

Chien ☰

Imprint Chien on your lower right cheek bone. Start with the eyes looking down to the lower right cheek bone, chanting Chien-Chien-Chien, and then Yang-Yang-Yang, Yang-Yang-Yang, marking it with your fingers on your forehead. Project the symbol into the universe as you breathe through your forehead. Picture the symbol with its bright red colors and when it connects to your forehead, bring it down to its place in the pakua in your lower abdomen.

a. **Yang Expand:** Do as before.
b. **Yang Expand:** Do as before.
c. **Yang Expand:** Do as before.

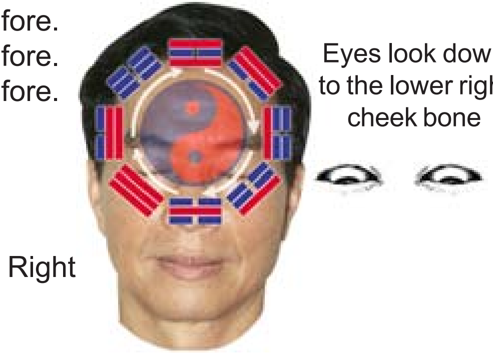

Eyes look down to the lower right cheek bone

Fig. 5.58 *Chien*

Note: Any time that you feel the energy shooting into your forehead, right where the senses' control is located, bring both the senses' energy and the energy from the universe down to your abdomen.

4. Form the pakua on your face to turn the senses inward.

Chant Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien and mark the place of the specific kua with your fingers.

a. Kan is on your bridge of the nose.
b. Li at the top of your forehead.
c. In this pakua Chen is on (inside) your left ear.
d. Tui is on (inside) your right ear.
e. Kun in your brain above the right eye.
f. Ken inside your cheek bone on the left side.
g. Sun in your brain above the left eye.
h. Chien inside your cheek bone on the right side above.

Repeat the chanting in your mind and let your eyes move to the place of each kua. Feel the power of the symbols attracting your senses in to the mideyebrow. Chant and spiral the Tai Chi symbol with your finger and then only with your mind. Let the Tai Chi spin fast and faster. Keep on chanting the Tai Chi symbol until the energies are pulled to the center of the pakua. You should feel that all the energy is flowing to the center of the face pakua and connecting to center of the brain, the crystal room, and keep on spiraling and be aware of the Lower Tan Tien, Tai Chi, and also the spiraling of the facial pakua itself. Feel a very strong center in your forehead and bring it down to the center of the lower pakua.

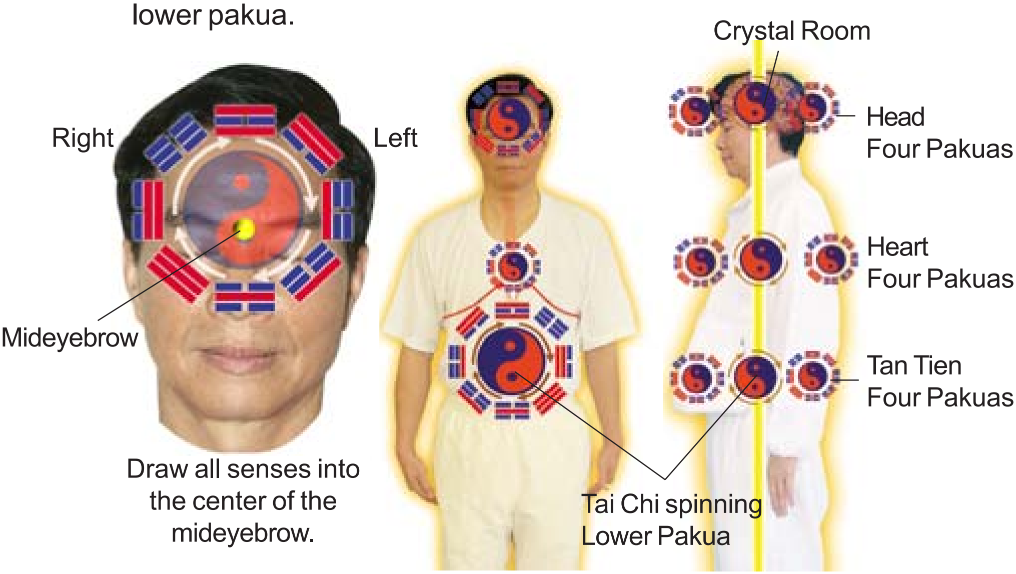
Crystal Room
Right
Left
Head
Four Pakuas
Heart
Four Pakuas
Mideyebrow
Tan Tien
Four Pakuas
Draw all senses into the center of the mideyebrow.
Tai Chi spinning
Lower Pakua

Fig. 5.59 *Pakua on your face to turn the senses inward*

Creating the Back Pakua on the Head

Duplicate trigrams, pakua and the Tai Chi symbol to the back and spiral clockwise while chanting the Tai Chi, Tai Chi. The mind and the eyes are spiraling as well, going faster and faster. When you spiral as fast as possible, you will feel that the energy is drawing into the back Pakua.

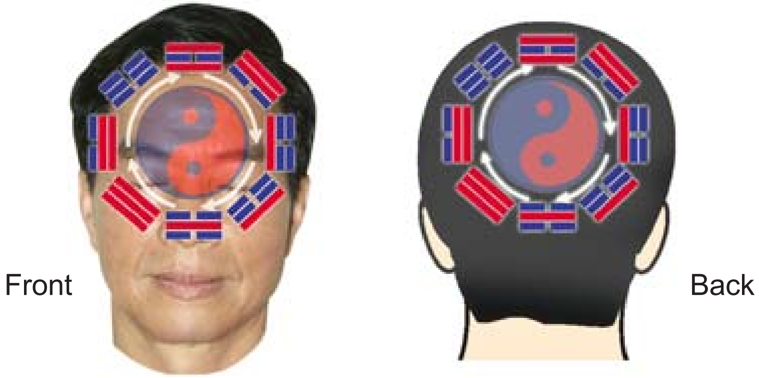

Fig. 5.60 *Pakua on the Head*

Be aware of the front and back Pakuas and spiral them. Spiral in the center faster and faster drawing the front and back into the center.

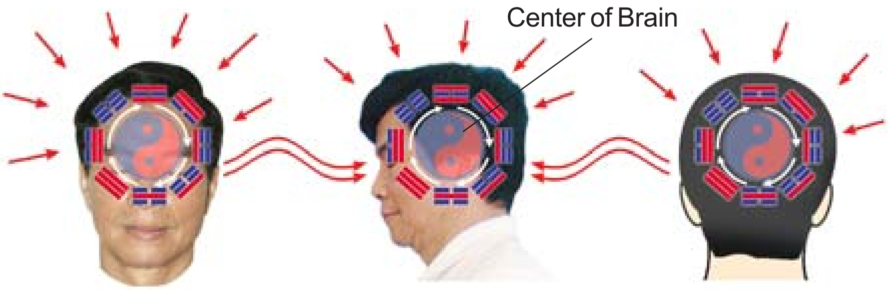

Fig. 5.61 *Front and Back Pakuas spiraling.*

Duplicate the front Pakua to the left side and let the Tai Chi symbol spin.

Duplicate the back Pakua to the right side of the head and let the Tai Chi symbol spin. Become aware of the center of the brain, the crystal room and let it spin faster and faster, drawing the energy of the left and right Pakuas into the center.

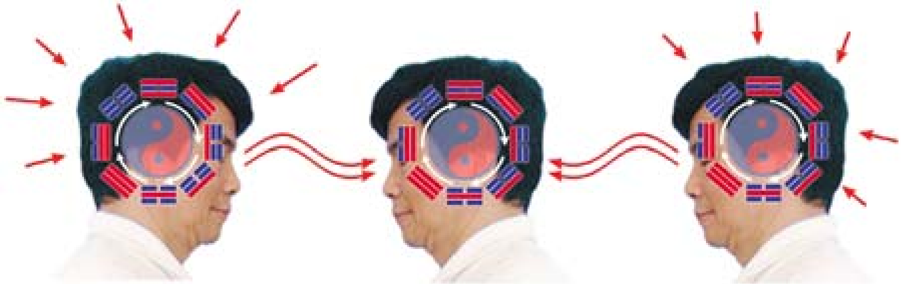

Fig. 5.62 *Energy of the left and right Pakua into the center of the Brain.*

Be aware the front, back, left and right Pakuas and let them spin. Put more attention on the center of the brain and let the center spin faster and faster, drawing all the energy into the center.

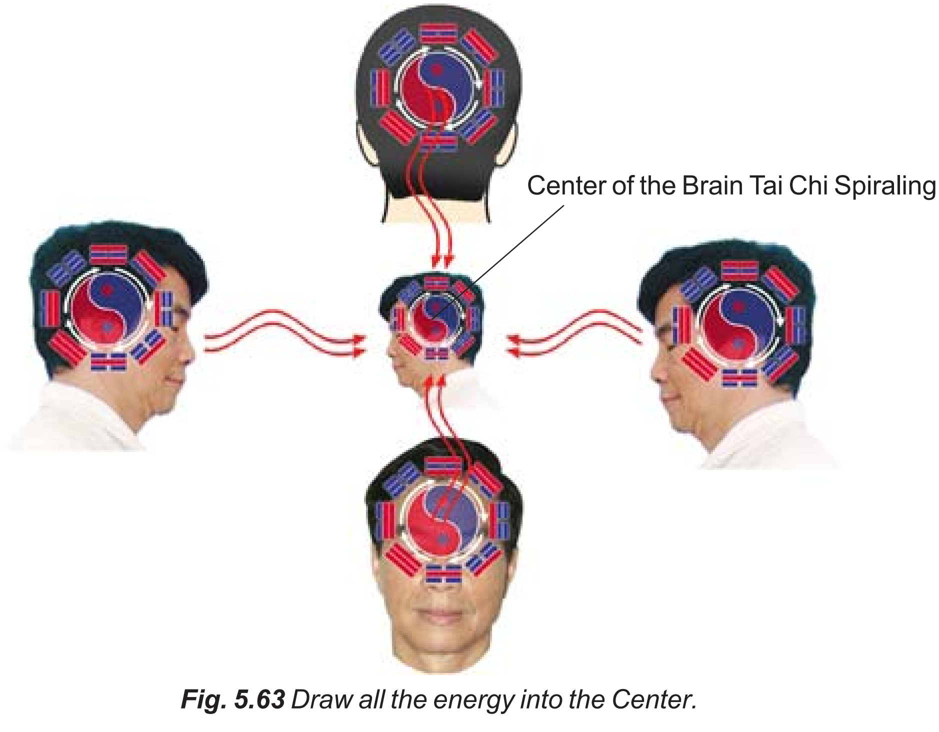
Center of the Brain Tai Chi Spiraling

Fig. 5.63 *Draw all the energy into the Center.*

Connecting the Senses with the Sense Control of the Facial, Heart and the Abdominal Pakuas.

Repeat the facial pakua again: Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien. This time, when you spiral the Tai Chi symbol as you chant Tai Chi, add the motion of the Yin coming in and the Yang going out. Your mouth is open projecting the sound of Yang into space. Your fingers are pointing out into space as well. With the sound of Yin your fingers are pointing in towards your forehead, which facilitates the inward movement of the energy. Picture a Tai Chi spiraling inside your brain, right behind your forehead. Point, spiral with your fingers saying Yin-Yang, Yin-Yang, Yin-Yang, and then Tai Chi, Tai Chi, Tai Chi. Continue saying it in your mind and feel the pakua deep inside your head (about 2-3 cm behind your forehead). All the senses are drawn together inside the center of the facial pakua (the senses' control point) and their energy is brought down to the Tai Chi of the lower abdominal pakua.

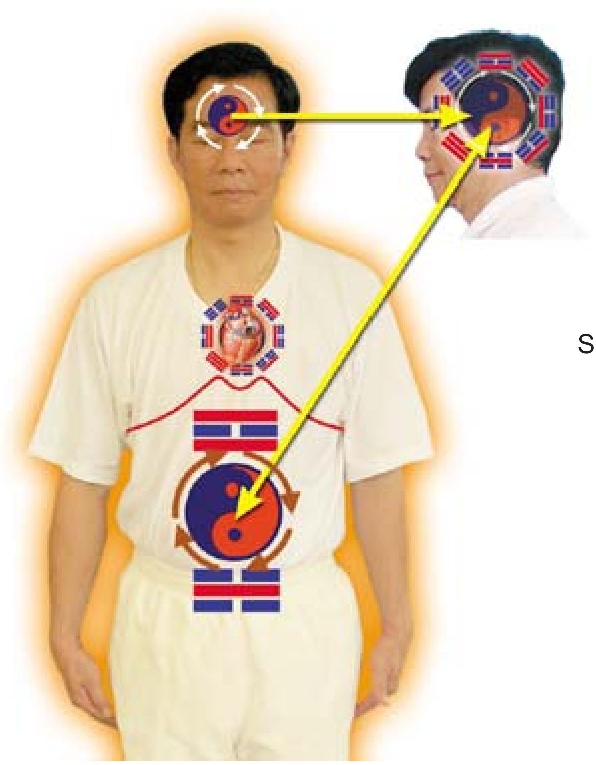

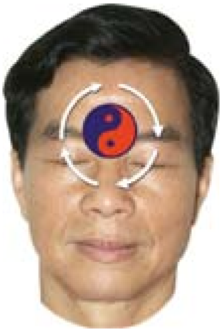

Sense Control at Mideyebrow

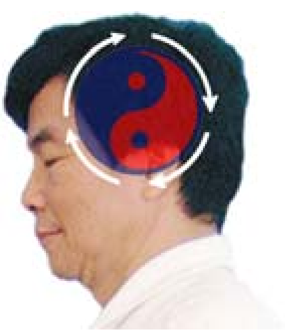

Crystal Room

Fig. 5.64 *Let the center of the brain, the heart and abdomen pakuas and all Tai Chis spiral together.*

Blending all the Energy of the Organs and Senses together in the Cauldron

Look down to your center in the Tan Tien and focus on the pakua on the back and on the right and the left pakua. When you focus on them, they will just start spiraling on their own. Look down into your center, into the cauldron and start spiraling the Tai Chi symbol in the center of your cauldron while chanting Tai-Chi, Tai-Chi, Tai-Chi. Rest. Continue to spiral with only your mind and feel that the energy of your senses and organs is coming right down to your center, where all the energy blends into one central energy. Rest. Spiral the energy in your abdomen. Smile to this energy and feel very calm and peaceful.

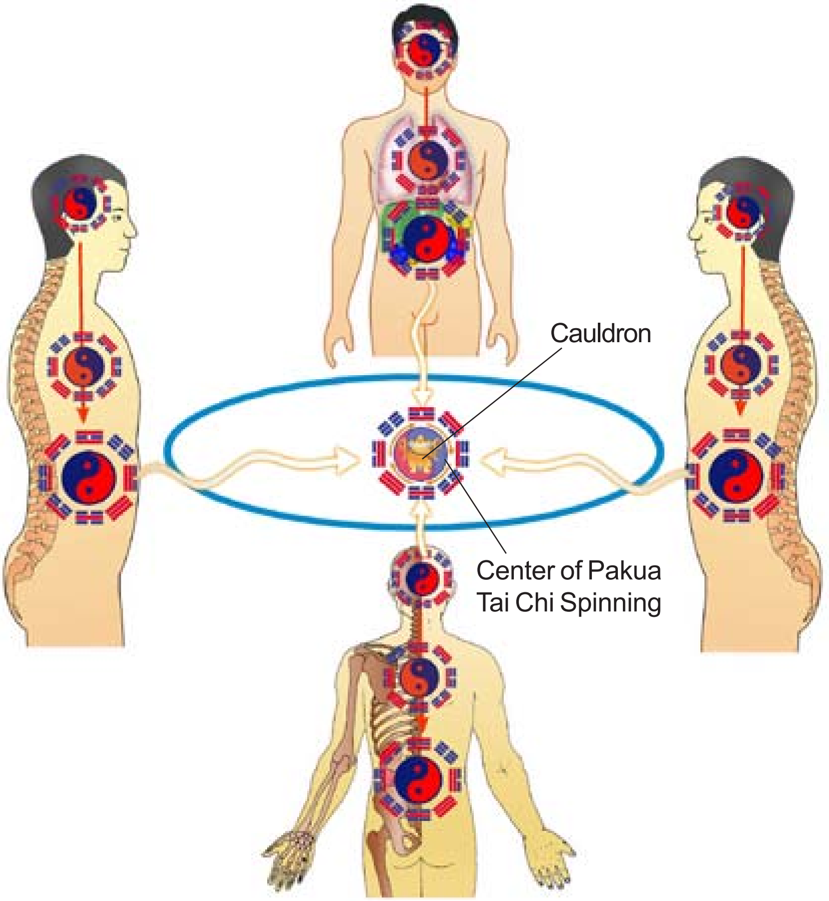

Fig. 5.65 *Draw all the sense into the center of the Pakua and Tai Chi.*

Expanding the Pakua in the Universe.

The power of the symbol is the power of the throat, the power of the word. The forces of the universe are Yin and Yang forces and the kua is a very ancient symbol of power. When you project the pakua out, using the pakua breathing, you connect to the Yin and the Yang power in the universe. When you breathe to the kua out there in space and connect to it, you will feel it connect to the kua in your abdomen as well.

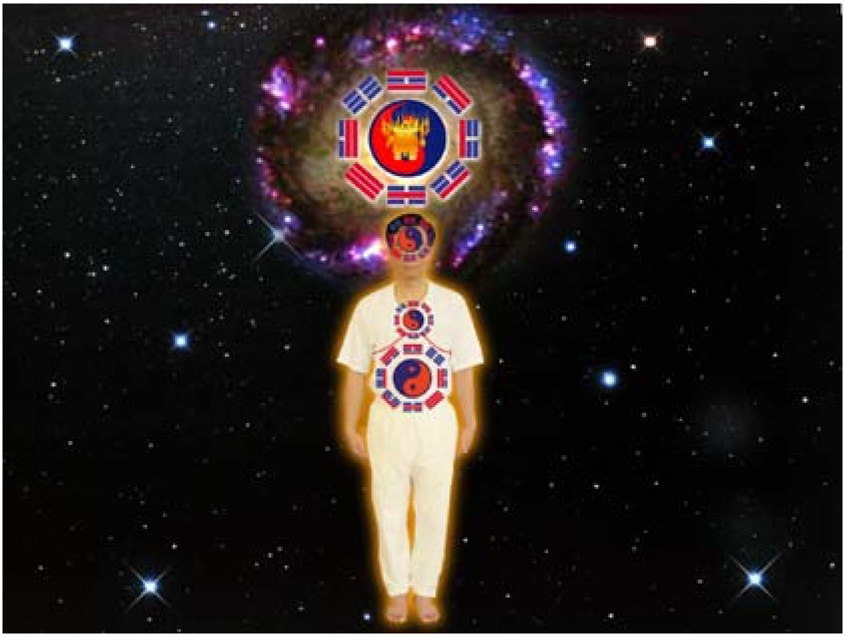

Fig. 5.66 *Pakua in the Universe*

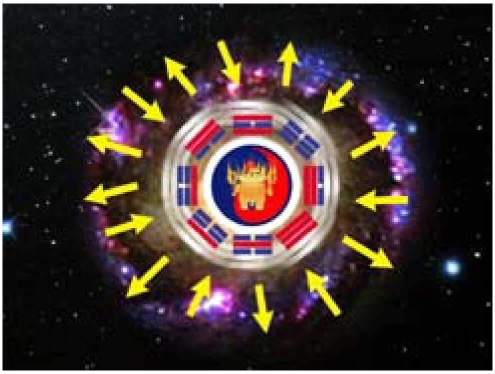

Fig. 5.67 *Feel the Pakua Breathing and Pulsating.*

1. Start chanting Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien touching each of the kuas as you chant them. Continue to do this with only your mind, saying the name of the kua and holding the symbol in front of your eyes. Then spiral the Tai Chi symbol from the bottom to the right and up, chanting it and spiraling with your hand, eyes and mind. Continue to spiral with your mind and eyes only. Feel that the energy is drawn inward and the Tai Chi symbol is spiraling and changing to violet light.
2. Activate the facial pakua. Chant each of the symbols and point to them on your face; Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien. Do it several times and make then the Tai Chi symbol on your forehead: Tai Chi, Tai Chi, Tai Chi, spiraling with your fingers over the third eye area. Feel the energy drawn into the center of the forehead, drawing the senses inwardly to the senses control and connecting down to the lower pakua. On their way down they collect the energy from the organs as well.

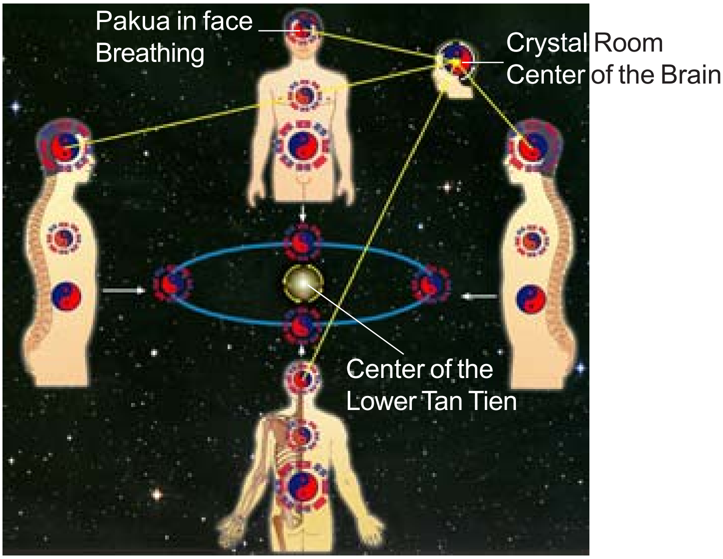

Fig. 5.68 *Four Head Pakuas*

3. Chant now Yin-Yang, Yin-Yang, Yin-Yang. Contract with Yin and expand from the forehead out with Yang. Use your fingers to point the direction of the energy, in and out.
4. Rest. Smile, feel the Tai Chi very deep inside you spiraling and chant Tai Chi, Tai Chi, Tai Chi, Tai Chi. Feel the facial pakua go into your head.

5. Become aware of the Kan symbol. Picture the Kan symbol and project it out down all the way to the galaxy. Chant Kan, Kan, Kan. Chant and picture the Yin-Yang-Yin lines and feel that the power is going out all the way to outer space.

   Start then with the Yin-Yang breathing. Inhale, Yin contracting with bright violet light, exhale, Yang expanding with bright red light. Inhale, contracting, exhale, expanding. Feel the kua of the facial pakua and the abdominal pakua breathing and at the same time, feel the kua in the universe breathing. Picture the Yin-Yang-Yin very clear in your mind as well as very far away in space. Exhale, hold your breath and feel the symbol breathing. Inhale, breathe without breathing and feel the connection with the kua out there. Keep on breathing in this way until suddenly the kua out there comes to you and enforces the Kan in your abdomen.
6. Repeat the same procedure for Li, sending it out, up into the universe, for Chen sending it out into the space on your right side, Tui to the space on your left side, Kun to the space upper left, Ken to the lower left side, Sun to the upper right side of the universe and Chien to the lower left side of the universe.
7. Now picture all the kuas together and repeat their names very slowly: Kan-Li-Chen-Tui-Kun-Ken-Sun-Chien. Rest. Feel the universal pakua covering you. Feel this pakua breathing and pulsating together with your pakua in the abdomen and on your face, drawing the energy back into the lower pakua, into your center. Sit back and smile to the pakua inside you and to the pakua in the universe. Rest, concentrate and condense all the energies in the cauldron.

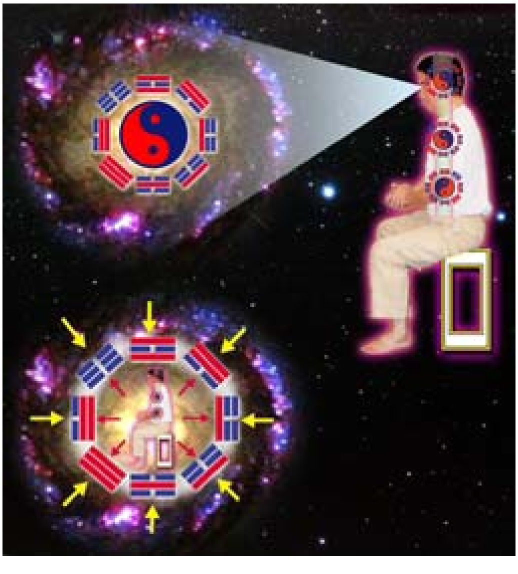

Fig. 5.69 *Feel all the Pakua and Universal Pakua breathing and pulsating. Feel a big Pakua cover you, breathing and pulsating.*

Formula Two: Transforming the Negative Emotions of Each Organ into Pure Life-Force Energy

Taoists reason that the negative emotions can be transformed to become our life-force and position energy. Therefore to expel or suppress unwanted, negative emotions is to expel or suppress life-force. Rather than suppressing them, you gain more by composting, recycling or transfering the negative into positive energy and experiencing these emotions. This means you permit them to emerge, observing and accepting them, but do not let them run wild or trigger other negative emotions. Instead transform them not only into useful life-force energy, but also into another, higher consciousness that is your spiritual energy.

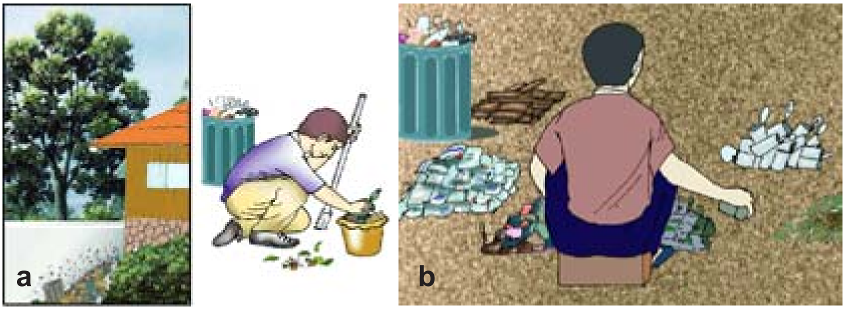

a.*Pile all the garbage together, just like mixing all the emotions together.*
b.*Separate the garbage, just like separating emotions & transfer them.*

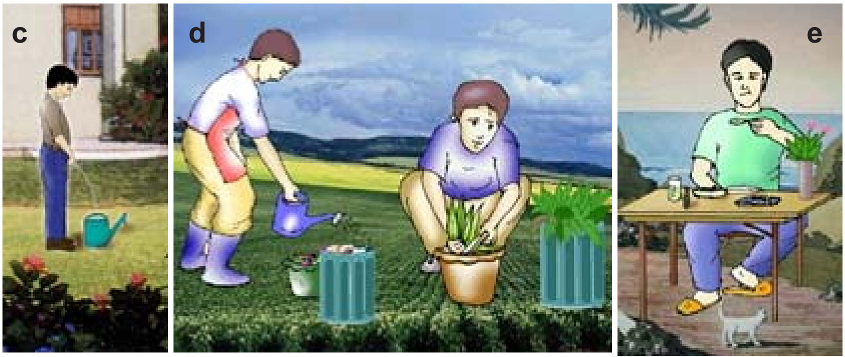

c.*Recycle and compost.*
d.*Recycling the garbage into compost and growing flowers and vegetables.*
e.*Eating the food grown from the compost.*

Fig. 5.70 *Transforming the Negative Emotions*

In the Taoist cosmology, the emotions originate from the vital organs, which correspond to the Five Phases of energy. They are also the reservoir of spiritual energy. By transforming the negative emotions in the organs, you are nourishing the spirit of that organ. Fear originates in and is stored in the kidneys, anger and frustration originate in and are stored in the liver, impatience, hastiness and hatred originate in and are stored in the heart, sadness and depression originate in and are stored in the lungs and worrying originates in and is stored in the spleen.

In the Tao, the main idea is to find a balance between the positive and the negative emotions. The negative emotions are like weeds, they are always there, and one could never totally get rid of them. The other point is that the soil needs them as fertilizers and they also help hold the soil in place. But at the same time the danger is that they will take over the garden if we don't control them. We need to cultivate the plants we want (our good virtues) to prevent the weeds from taking over the garden. It is the same with the negative emotions; they should only be kept in balance with the positive ones and one shouldn't waste time trying to get rid of them altogether. The work should be done on the positive emotions, constantly cultivating and nurturing them in order to grow the good virtues and to keep the right balance with the negative emotions. Negative emotions breed negative emotions and positive emotions breed positive emotions.

Fig. 5.71 *Yin and Yang balance negative and positive emotions but they do not get rid of them.*

We work in this formula with the counteracting or controlling cycle and we use collection points to gather and neutralize the negative emotions of each organ. We then blend the emotional energies together in the pakua, transforming them into pure life-force.

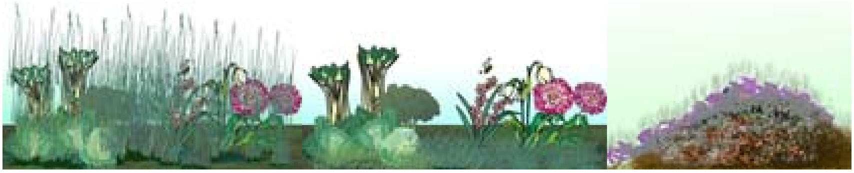

***Fig. 5.72** Keep on cultivating the positive emotions.*
Do not let the weeds out grow the vegetables.

Practice:

1. Preparation

When you begin, first do Formula Two and then continue with this practice.

a. Sit properly on a chair and feel yourself aligned with the forces.

b. Do the crane and turtle neck and the spinal cord breathing

c. Smile down and generate the qualities of loving energy from the heart. Breathe down radiance to the heart making it feel soft in the heart.

d. Be aware of the Tan Tien Chi and the front pakua, the pakua on your face, and the universal pakua. Feel them all breathing together. Feel all your senses and mind turned inwards to the cauldron. Turning the senses inward initiates the training process of inner observation. In developing the ability to focus inward – to smell, listen, taste, see and hear the organs and their activities – and to observe your negative emotions without predetermined judgments, you have the opportunity to develop your true nature.

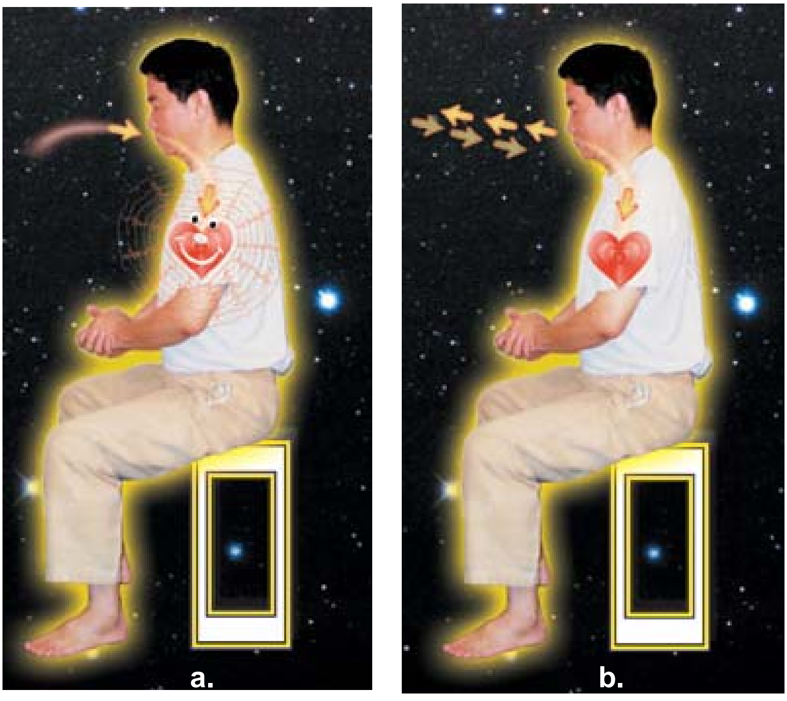

Fig. 5.73 *Feel joy and happiness.*

a. *Inhale to the heart and smile to the heart.*

b. *Exhale to the heart, smile to heart and make the heart feel soft.*

2. Real Practice

A. Bring your attention to your kidneys.

Listen to the kidneys, observe your kidneys. Be aware of any fear in the kidneys. Accept your fear; smile to the heart and let the love radiate to the kidneys. Be aware of the Kan, the water power. Chant the Kan and connect your kidneys with this energy. Let love activate the gentleness, calmness and stillness, the essence of this power, a blue light. Wrap the gentleness, calmness around your fear, balancing it and transforming your fear.

Form with your mind a collection point at your perineum and let any fear that is still in the kidneys go down to the collection point. Balance these emotions here also with the positive feeling of gentleness and calmness.

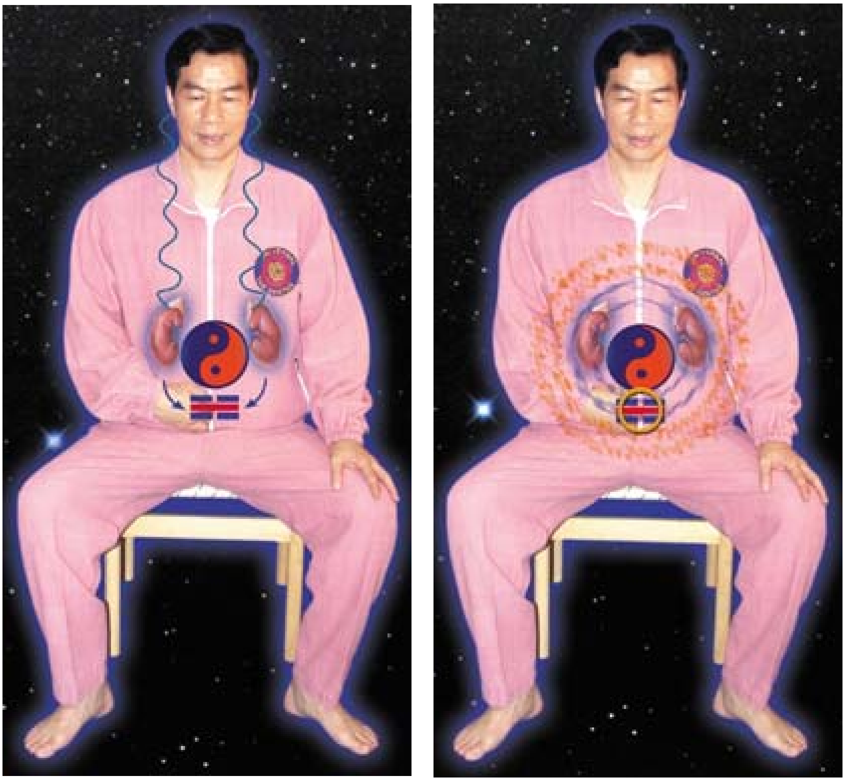

a. **b.**

Fig. 5.74

a. *Fear - Gentleness*

b. *Fear into Gentleness and Calmness*

B. Bring your attention to your heart.

Connect the tongue with your heart, speak with your heart (the spirit of your heart). Observe your heart. Be aware of any impatience, hastiness, cruelty, hatred, arrogance in your heart. Accept these feelings. Be aware of the Li, the fire power. Chant the Li and connect your heart with this power, a red light. Feel the warmth, love, happiness and inner joy, the essence of this power. Help transform the hatred into love. Wrap the love, the warmth and inner joy around your impatience, arrogance, balancing them with and transforming your impatience, hastiness, arrogance.

Form with your mind a collection point near the heart center and spiral all the negative energies that are still in the heart to this collection point. Chant the Li again and balance the emotions in the collection point.

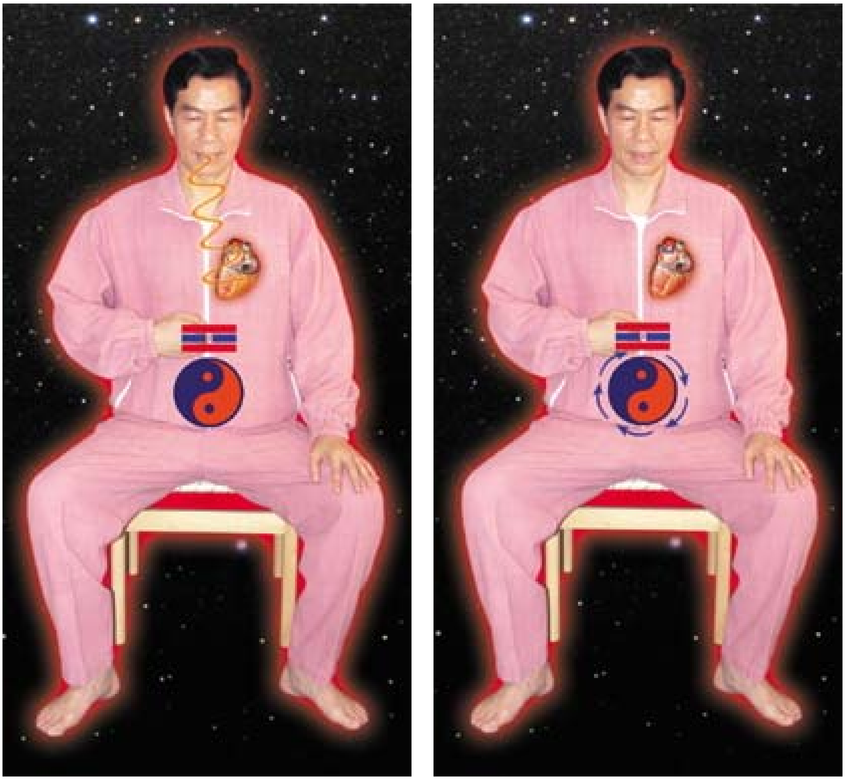

a. **b.**

Fig. 5.75 a. *Impatience, Hastiness and Hatred*
b. *Love, Joy and Happiness*

C. Spiral, blend and transform the energy of the heart and the kidneys at the front pakua.

Spiral and breathe the energies out of the kidney and heart collection points to the front pakua. Blend and spiral them together in the front pakua. (The pure energy trapped in these negative feelings will be freed and released to the center of the pakua). Chant Tai Chi, Tai Chi, Tai Chi. Yin-Yang, Yin-Yang, Yin-Yang. Spiral these energies until they become a bright, golden energy. This energy radiates love and gentleness from the center of the pakua and your being.

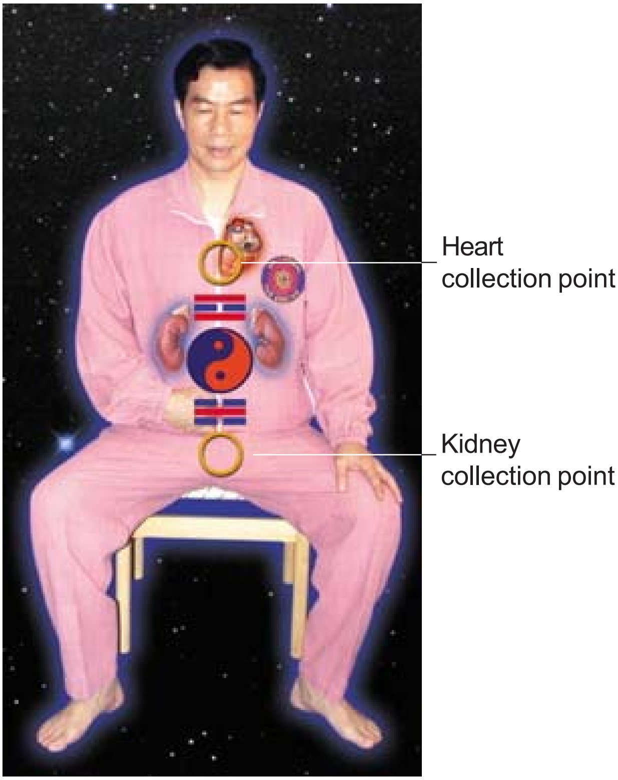

Fig. 5.76 *Spiral, blend and transform the energy of the heart and the kidneys.*

D. Bring your attention to your liver.

Connect the eyes with the liver and observe your liver. Be aware of any anger, frustration, aggressiveness, guilt in the liver. Accept this feeling. Be aware of the Chen. Chant the Chen and connect your liver with the power of the thunder and the wood, a green light. Feel the kindness, the essence of the wood power. Wrap this feeling of kindness around your anger, balancing them together and transforming your anger. Feel that you can forgive. Forgiveness is one of the most important practices of the Taoist system.

Form with your mind a collection on the left side of the pakua, on the nipple line, and spiral all the negative energies that are still in the liver to this collection point. Chant the Chen again, connect with the wood power and balance the emotions in the collection point.

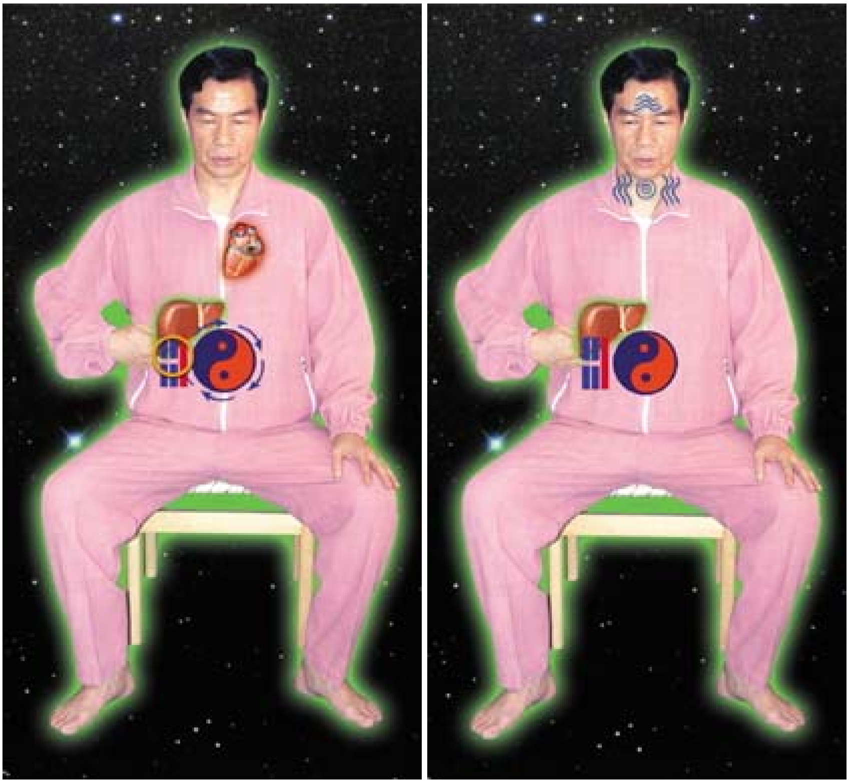

Fig. 5.77 a. *Anger and Frustration in the Liver*
b. *Love from heart and gentleness from kidneys help tranfrom the anger and frustration into kindness.*

E. Bring your attention to your lungs.

Connect your nose with your lungs and observe your lungs. Be aware of any sadness, depression in the lungs. Accept these feelings. Be aware of the Tui. Chant the Tui and connect your lungs with the power of the rain and the metal power, a white light. Feel the courage and the righteousness, the essence of the metal power. Wrap this courage around your sadness, balancing it with and transforming your sadness.

Form with your mind a collection point on the right side of the pakua on the nipple line and spiral all the negative emotions that are still in the lungs to this collection point. Chant the Tui again, connect with the metal power and balance the sadness with the courage here again.

F. Spiral, blend and transform the energy of the liver and the lungs at the front pakua .

Spiral and breathe the negative emotional energies from the liver and lungs collection points to the front pakua. Spiral and blend them at the front pakua. Chant Tai Chi, Tai Chi, Tai Chi, Yin-Yang, Yin-Yang, Yin-Yang. Use the power of pakua to balance and neutralize the sadness and anger. Feel that the pure energy that is trapped in these negative feelings will be freed and released to the center of the pakua. Continue to spiral the liver's and lungs' energies until they become a bright, golden energy that radiates kindness and courage from the center of the pakua and of your being.

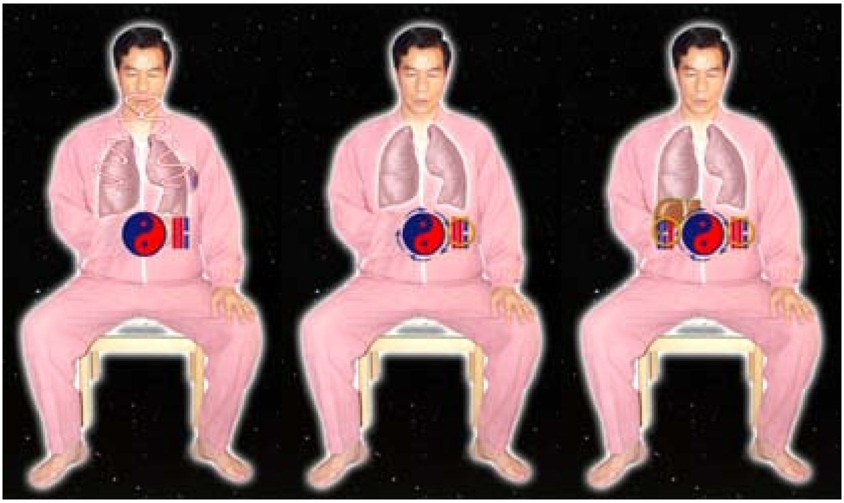

Fig. 5.78 *Spiral, blend and transform the energy of the liver and the lungs.*

G. Bring your attention to your spleen.

Connect your mouth with your spleen and observe your spleen. Be aware of any worry or other emotions you do not like in your spleen. Be aware of the Kun. Chant the Kun and connect your spleen with the earth power, a yellow light. Feel the openness, balance and centeredness of this power, the essence of the earth power. Wrap this around your worries, balancing with and transforming your worries.

Spiral your worries that are still in the lungs to the center of the front pakua, the collection point of the spleen. Chant the Kun again and then Tai Chi, Tai Chi, Tai Chi. Yin-Yang, Yin-Yang, Yin-Yang. Spiral, balance, blend and transform it with the energy that is already there into a bright, golden energy.

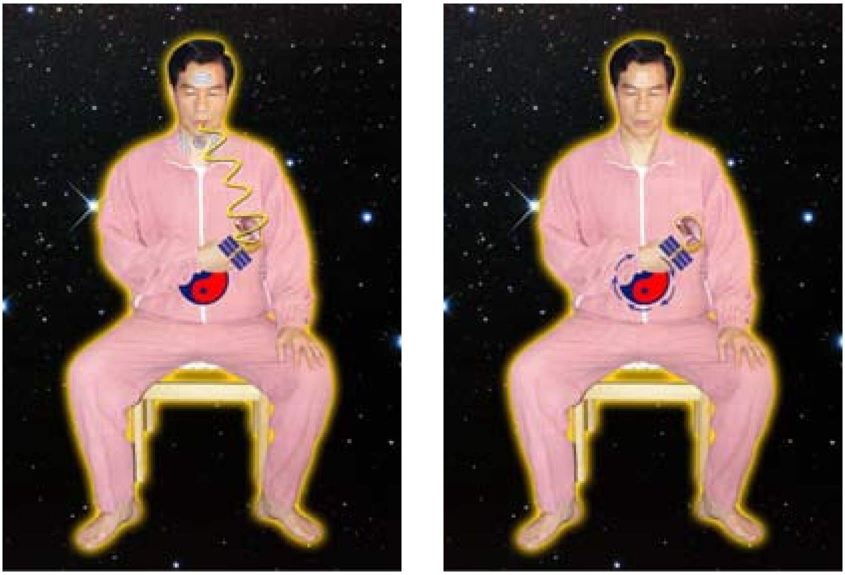

a **b**

Fig. 5.79 a. *Draw attention to your spleen.*
b. *Balance, Openness and Fairness*

H. Spiral all remaining negative energy to the front pakua.

Return to the organs and the collection points, and spiral and breathe to draw out any remaining negative energy. Blend and neutralize it with the energy in the front pakua. See all the organs glowing with light.

I. Be aware of the back, left and right pakuas.

When you have time, you can chant the kuas to enforce the power of the pakuas. Let the Tai Chi in all four pakuas spiral, drawing in, blending and transforming the energies. Continue spiraling and with the spiraling and the breath you draw, suck in the energy to the cauldron, the big empty space in the Tan Tien. Now spiral the energy in the cauldron, chanting subvocally Tai Chi, Yin-Yang, condensing the energy into a bright golden energy ball, a pearl. Bring the pearl to the perineum and circulate it in the Microcosmic Orbit. As the pearl moves through the Microcosmic Orbit, feel that it attracts and absorbs the Universal (at the crown), the Cosmic Particle (at the mideye) and the Earth (at the perineum) Force.

Bring the pearl back into the cauldron. Center yourself in the cauldron and feel your senses and mind drawn inward towards the energy in the cauldron. Feel nice, calm and peaceful inside. When you have time, sit for a while and let this energy and feeling grow inside you.

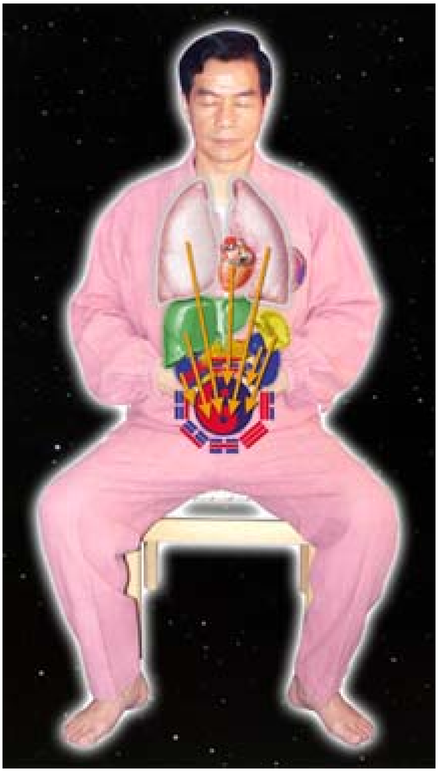

Fig. 5.80 *Let all the Organ Energy flow down to the Pakua and feel Chi grow inside you.*

Chapter 6

Compassion Fire Meditation Practice of the Creation Cycle

Practice compassion through the act of forgiveness. To forgive is to let go. Always meet conflict with compassion.

The higher practice of inner alchemy is the transformation of the heart energy into love, the liver energy into kindness and generosity, the lung energy into courage, the kidney energy into gentleness and the spleen energy into openess and fairness. Combine all these energies into compassion. Compassion becomes non physical energy which we can carry with us when we leave the world. It is the love and joy we feel within.

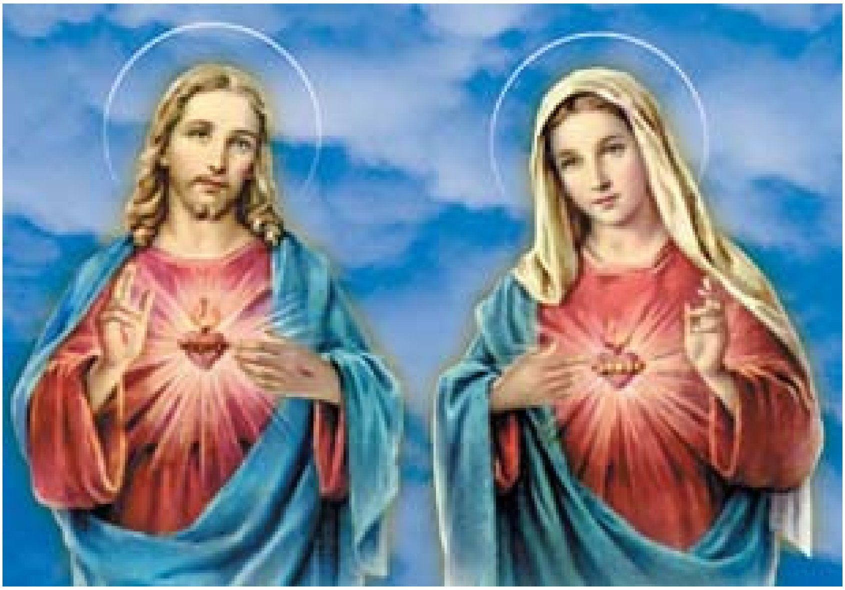

Fig. 6.1 *Compassion Fire originates from the Heart.*

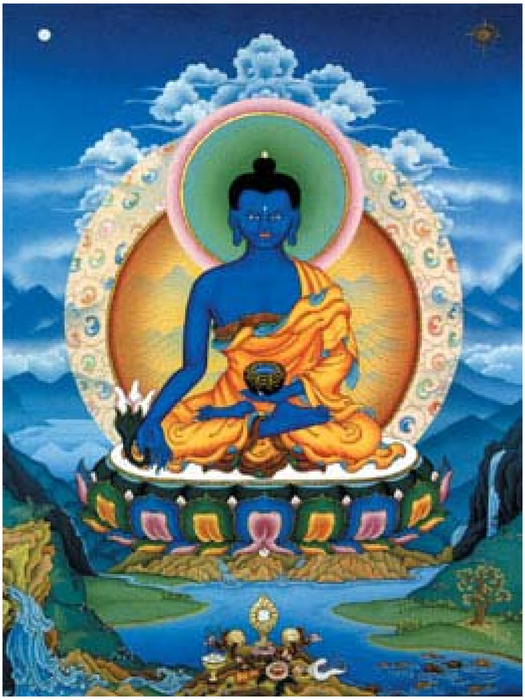

***Fig. 6.2** Through Compassion Fire,*
we transmit Positive Energy to Our Surroundings.

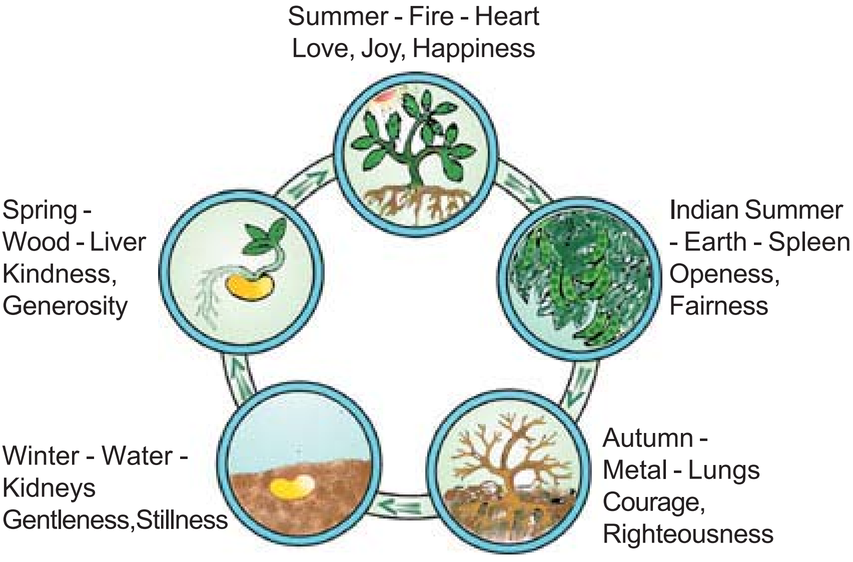
Summer - Fire - Heart
Love, Joy, Happiness
Spring - Wood - Liver Kindness, Generosity
Indian Summer - Earth - Spleen Openess, Fairness
Winter - Water - Kidneys Gentleness,Stillness
Autumn - Metal - Lungs Courage, Righteousness

***Fig. 6.3** Creation Cycle in Nature*

Creating compassion energy begins with the creation cycle.

Activate the heart first. It is the fire and the place where the virtue energy can be combined into compassion energy. Go to the kidneys and bring the energy to the collection sphere. Continue with the liver, the heart again, the spleen, the lungs and bring the energy to the respective collection points.

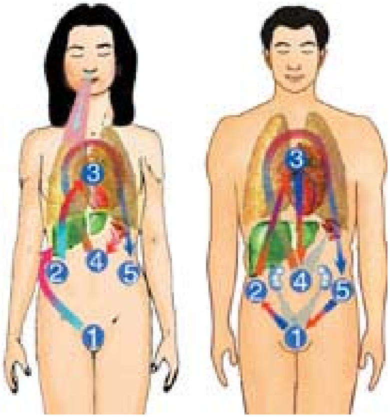

Fig. 6.4 *Creation Cycle*

In the human body the arousal state creates a new chemistry and a new vibration. This ‘special’ vibration is measured at a frequency of 8Hz/Sec. ‘As in the macrocosm, so it is in the microcosm’; hence if we make love, then all cells and the DNA actually make love as well.

The DNA cross over, like two serpents intertwining in an erotic embrace. They make love and they give birth. Any of the cells that do not ‘reproduce’ will eventually die.

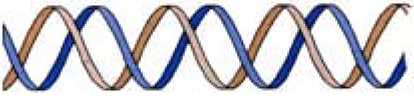

Fig. 6.5 *When we feel love and arousal energy,*
the DNA crosses over and makes love.

The two vital states are arousal, leading to orgasm and compassion are inextricably linked to love. When this ‘love-vibration’ reaches the pineal gland, a new hormone is produced. This in turn creates whole body conductivity. If the vibration is above or below the 8Hz/Sec. frequency, this process does not occur. Only when you feel the waves of orgasmic vibration and unconditional love for the self and others can the process be activated.

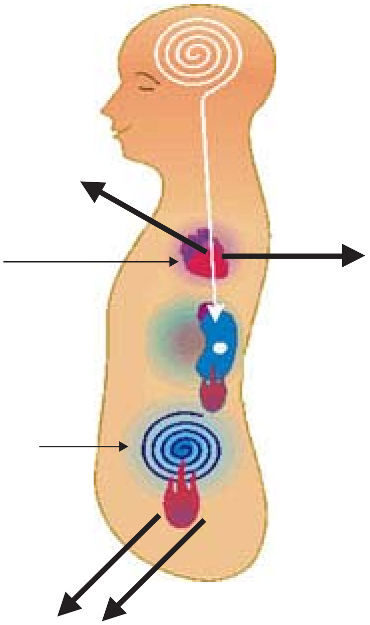

Some practices just want to open the heart by temporarily letting us feel love and kindness. The heart is drained through all kinds of giving and donations.

Worldly desire and many businesses work on activating the sexual desire and then draining it out.

Fig. 6.6 *When the Heart is temporarily open and Sexual Desires activate, these forces will drain us out.*

One Lasting Energy is the Combination of Compassion Energy and Orgasmic Energy

Heart Energy

1. Heart energy is immaterial. When activated it will change into love, but love does not last long.
2. When all the good virtue energy from all the organs combines together into one energy, it becomes compassion energy; it will stay in you longer, but will still disperse.

Sexual Energy

1. Sexual energy when activated from material into immaterial is powerful, but will not last.
2. When the sexual energy is aroused and changes into orgasmic energy, it can charge all the organs, glands, senses and combines into one energy; this energy will stay longer but still will disperse. Orgasmic energy combined with compassion creates one long lasting energy.

After the orgasmic energy has charged all the organs and created the multi-orgasm energy, it can then be combined with the compassion energy at the Tan Tien. This creates the one energy that lasts the longest.

When we are aware of the cosmic galaxy and expand our combined energy of compassion and orgasm through the universe, it multiplies and flows back to us and we will be in touch with universal compassion and unconditional love.

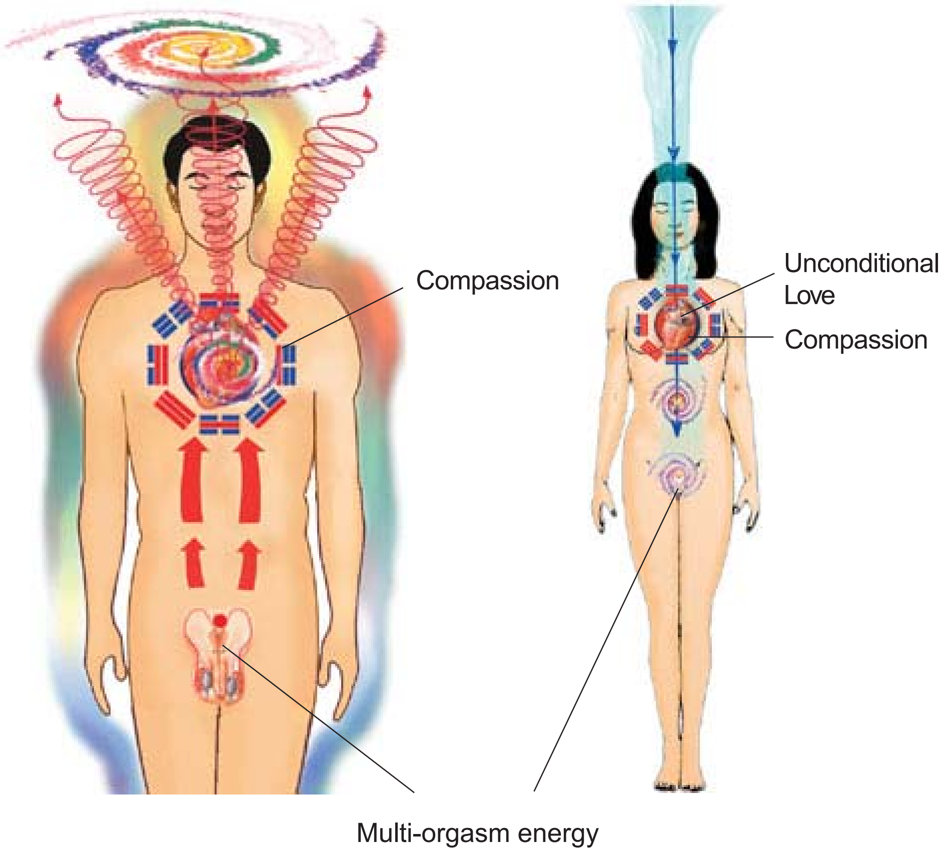

Fig. 6.7 *Through cultivating compassion energy, we can eventually build the one lasting energy.*

Practice of Creating the Compassion Fire and Multi-Orgasm

Compassion originates from the Heart

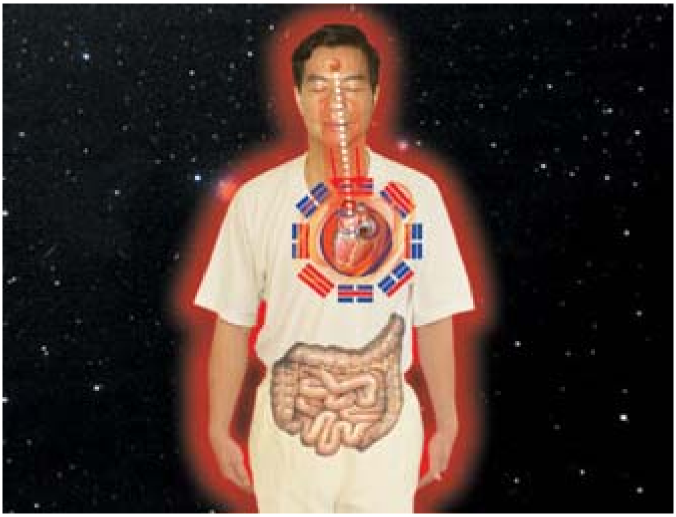

Fig. 6.8 *Smile until you feel the heart soft with love, joy and happiness.*

We can connect with the Universal Love.

1. Be aware of the Tan Tien and the universe and feel the unconditional love flow down.

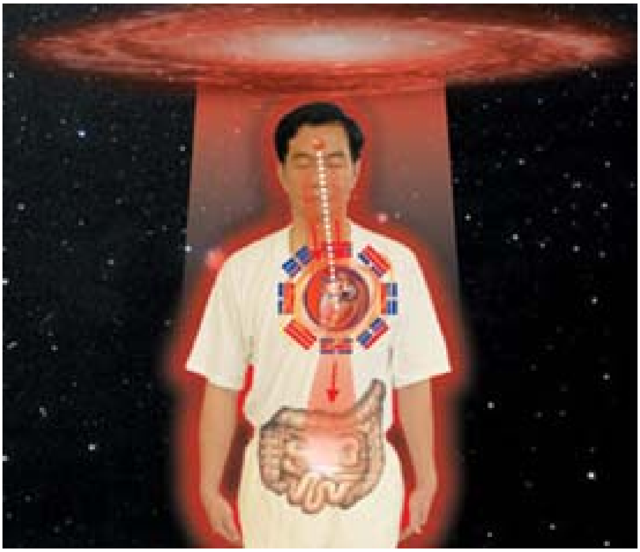

Fig. 6.9 *Smile to the heart and the small intestine, to make the heart feel soft. Feel the heart pakua spiral.*

2. Smile to the spleen and let the love, joy and happiness from the heart activate the openness and fairness.

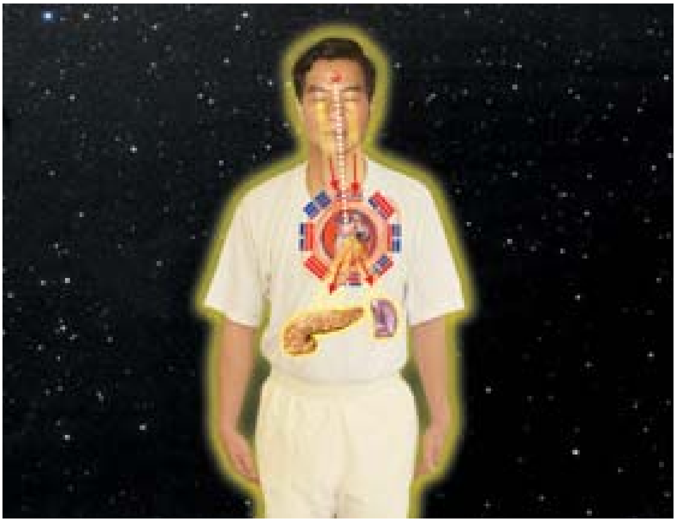

Fig. 6.10 *Let Positive Heart Energies activate Positive Spleen Energies.*

3. Let it flow back to the heart pakua and spiral, blend together and feel the yellow light flow from above and flow down to the heart.

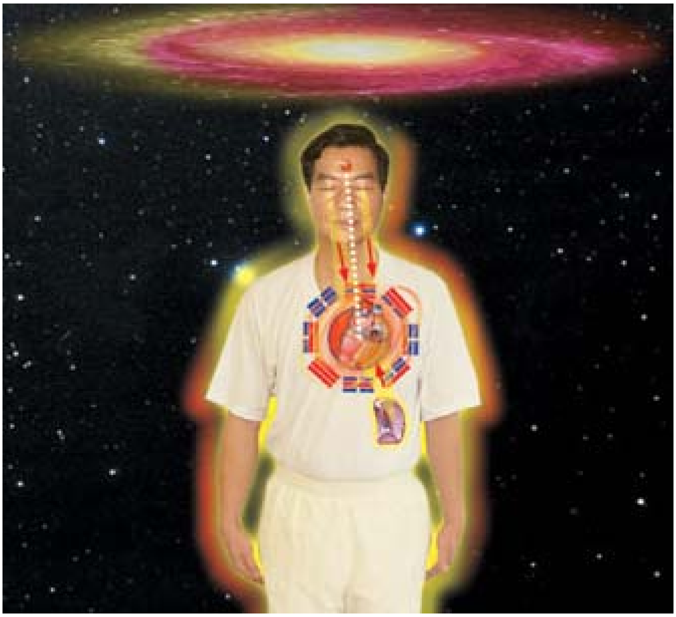

Fig. 6.11 *Blend the Two Energies with Universal Energies.*

4. Smile to the lungs and the large intestine. Send love from the heart to the lungs to activate courage and righteousness.

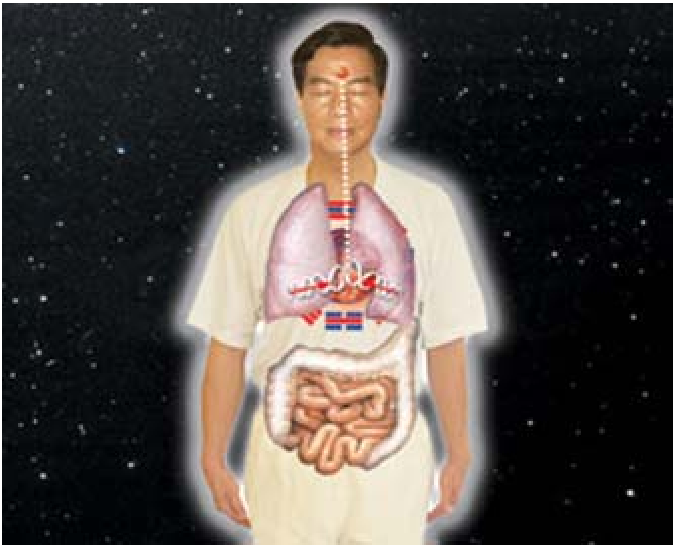

Fig. 6.12 *Let Love and Joy from the Heart activate Positive Lung Energies.*

5. Let it flow to the heart, spiral and blend with love, joy, happiness, openness and courage. Blend the three energies into one energy, and feel the white light from above flow into the lungs and the heart.

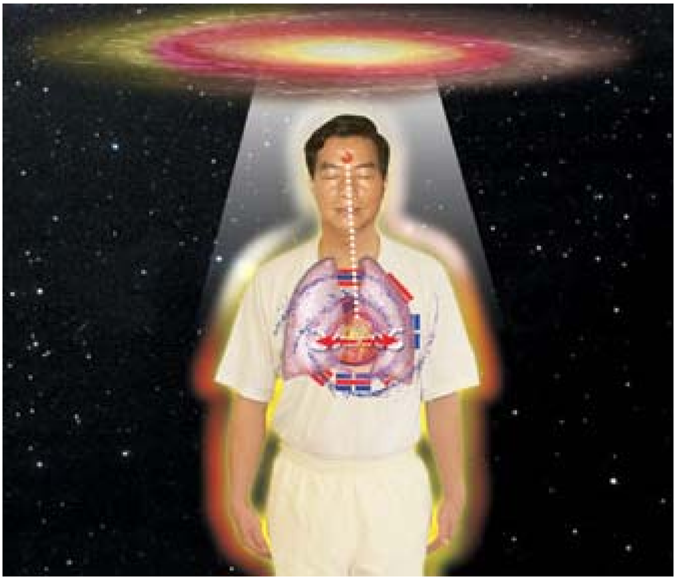

Fig. 6.13 *Blend the Three Energies into One Energy.*

6. Smile to the kidneys and let the love of the heart activate the gentleness, softness and calm in the kidneys.

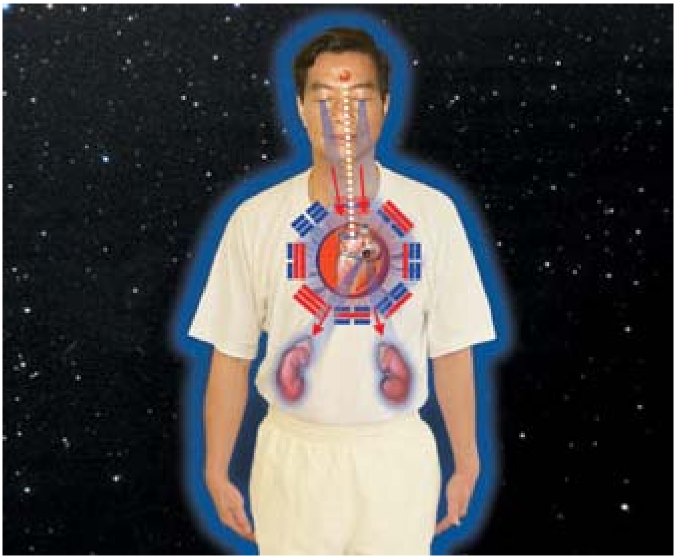

Fig. 6.14 *Let Love and Joy from the Heart activate Positive Kidney Energies.*

7. Let the gentleness, softness and calm in the kidneys flow into the heart. Blend and spiral into one energy.

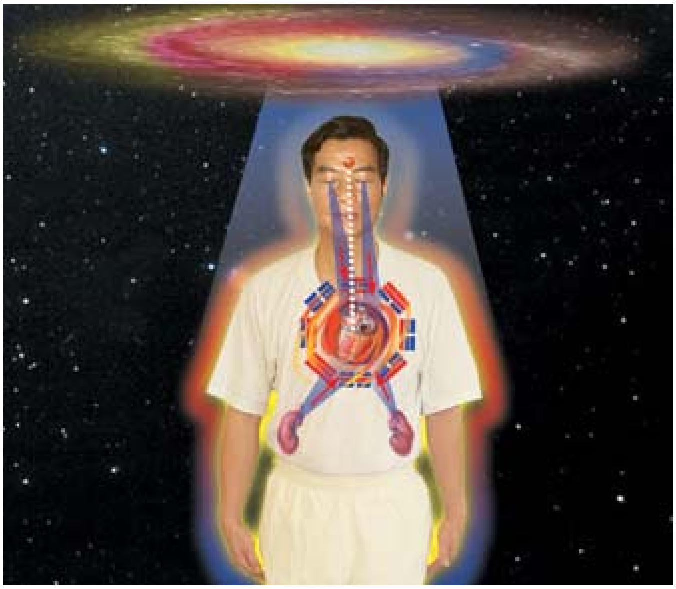

Fig. 6.15 *Blend Positive Kidney Energies with the Compassion Energy in the Heart.*

8. Smile to the liver and gall bladder. Let the softness and the love from the heart flow down to activate the kindness and generosity of the liver.

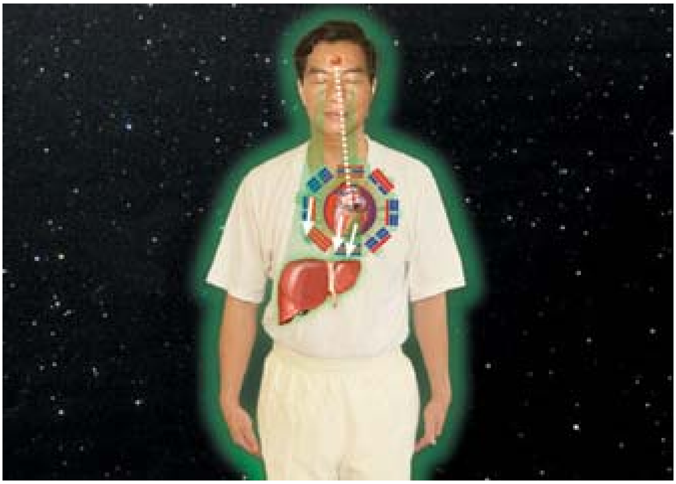

Fig. 6.16 *Let Love and Joy from the Heart activate Positive Liver Energies.*

9. Let the kindness and generosity flow to the heart, spiral and blend in the heart to become compassionate; be aware of the green light flowing from above.

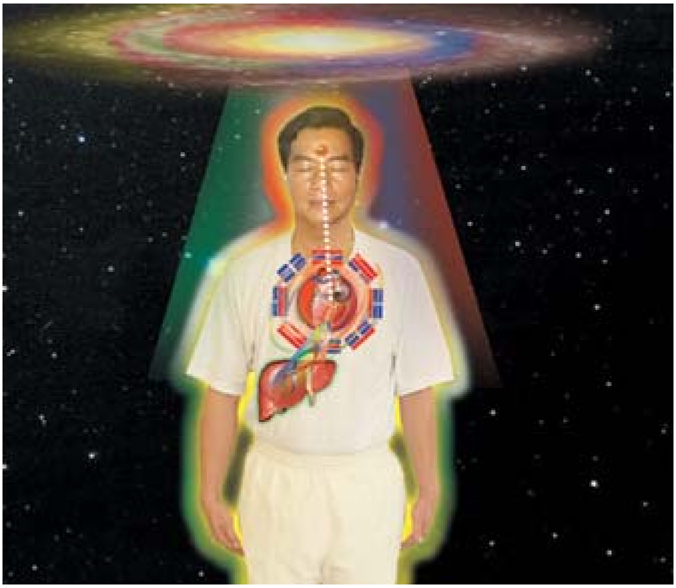

Fig. 6.17 *Blend Positive Liver Energies into the Compassion Energy.*

10. Smile to the heart and feel all the good virtuous energy from all the organs flow into the heart and spiral and blend into compassionate energy.

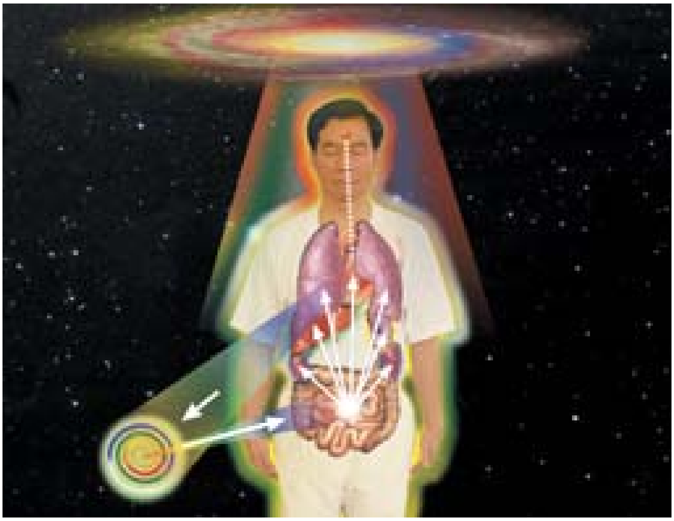

Fig. 6.18 *All the Organs' Positive Energies blend with the Heart to create the Compassion Energy.*

11. Your compassion rises up. Be aware of the galaxy, the cosmos, and expand your compassion through the universe. The universe allows us to be in touch with the universal compassion.

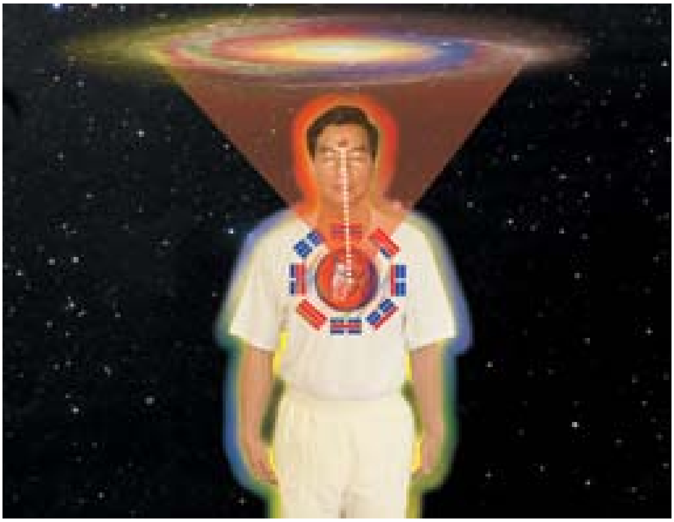

Fig. 6.19 *Expand our Compassion to be in touch with the Universal Compassion.*

12. Let them blend and multiply, and flow back to enhance you. You feel your whole body to be filled with the universal compassion. You have such abundance, it overflows.

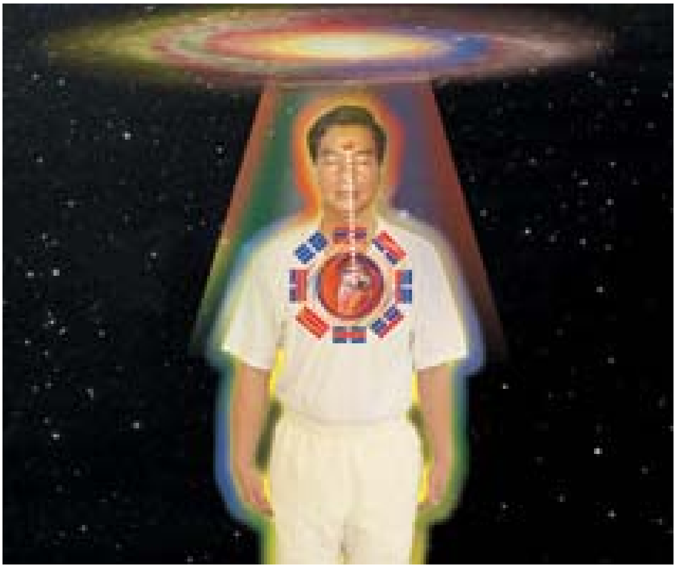

Fig. 6.20 *Let the Universal Compassion within us overflow.*

Multi-Orgasm Energy Smile to the Sexual Organs

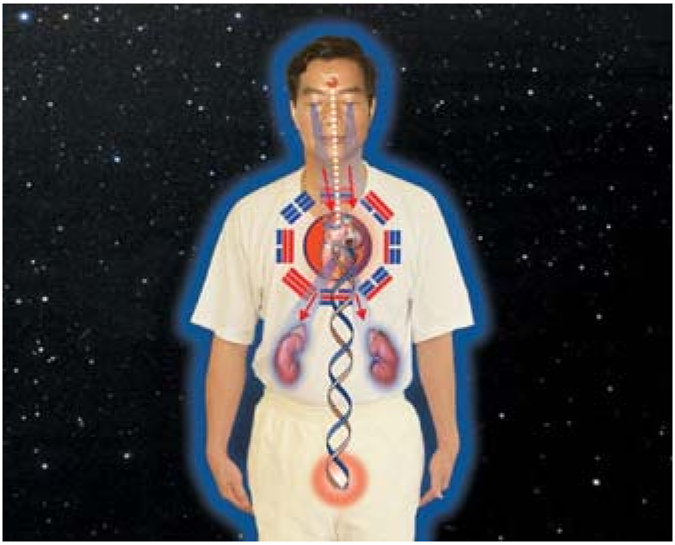

Fig. 6.21 *Smile down to the Heart and feel warm Energy from Heart flow down to Kidney and the Sexual Organs.*

1. Be aware of the sexual organ and use your hand to cover the sexual center. Smile and feel the fire from the Tan Tien flow down.
2. Recall the time when you were with your loved one and the high orgasmic feeling. Guide the orgasmic energy up to the kidneys, and feel the kidneys charge with Chi and orgasm.
3. Guide the orgasmic energy spiraling up to the small and large intestine, liver, spleen, pancreas, stomach, and up to the lungs, and heart. Continue up to the thymus, thyroid and parathyroid and up to the brain and the senses. Charge them with orgasmic energy, and let the Chi flow back down to the sexual organs and blend and spiral into one multi orgasmic energy.

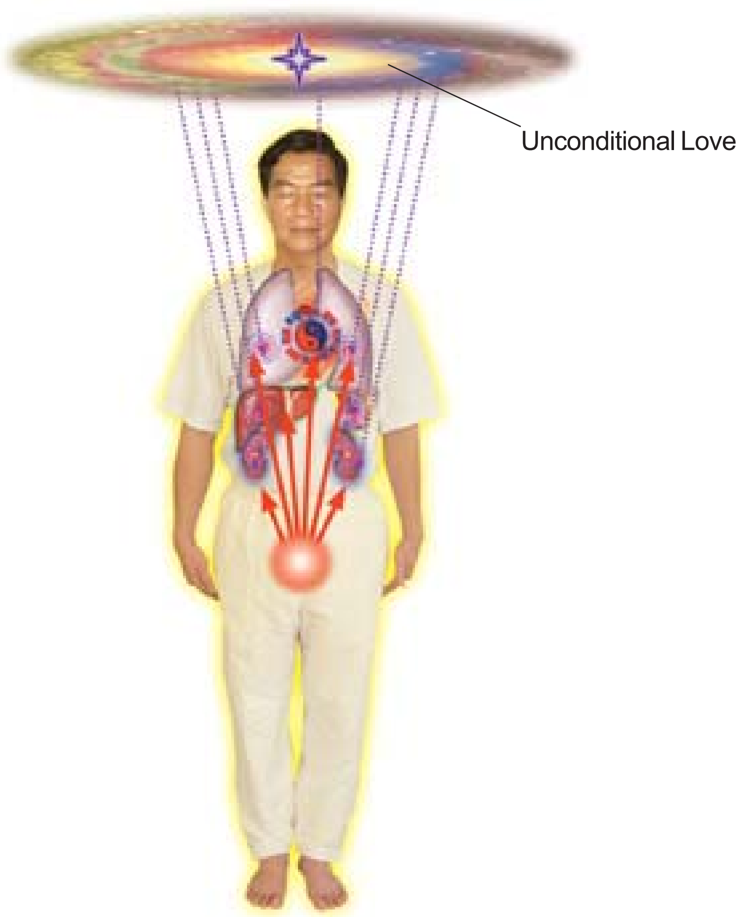

Fig. 6.22 *Bring the Orgasm up spiraling up to the organs.*

4. Be aware of the Muti-Orgasmic Energy in the organs and bring them up to the crown and continue spiral up to the Universe. Let the orgasm multiply and be aware of the Cosmic Orgasmic and let it flow down to the crown and feel all the cells fill with Cosmic orgasm and flow into the heart then down to the sexual organ filling them with compassion and Muti-Orgasmic Energy.

Universal Compassion: Your compassion arises up. Be aware of the multi-orgasm sexual energy and feel it combine with the compassion into one energy at the Tan Tien. Be aware the cosmos galaxy and expand your compassion through the universe. The Universe puts us in touch with the universal compassion.

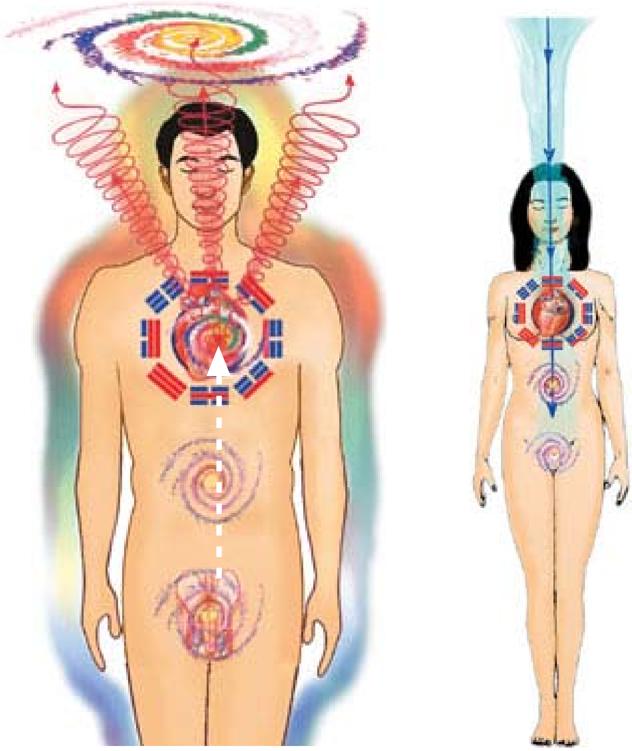

Fig. 6.23 *Universal Compassion*

This compassion practice is the major practice for all the Fusion and Kan & Li practices. The compassion fire is the most important energy of the transformation and will be use in all the practices.

Practice compassion through the Act of Forgiveness.
To forgive is to let go.
Always meet conflict with compassion.

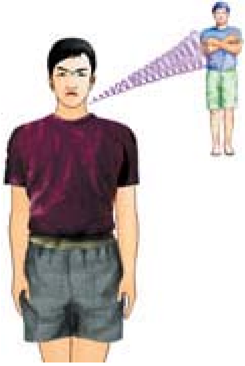

Fig. 6.24 *Be aware of the people that you feel resentment toward.*

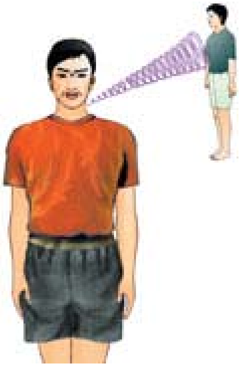

Fig. 6.25 *People who make you feel angry.*

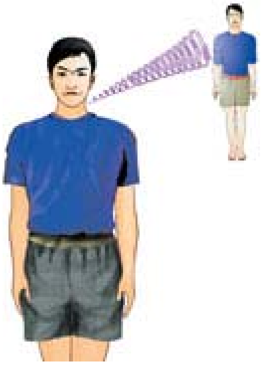

Fig. 6.26 *People that you hate or people who hate you.*

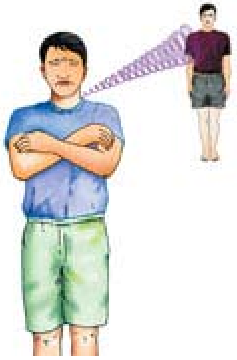

Fig. 6.27 *People who hurt you or people who have been hurt by you.*

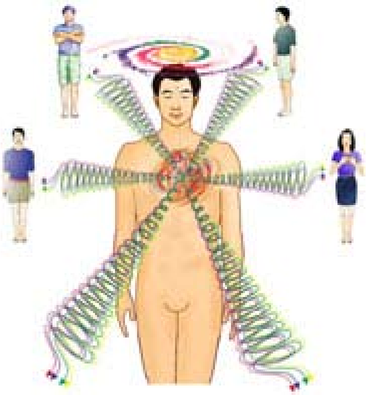

Fig. 6.28 *Send out the abundance of compassionate energy that you have from the universe, to fill the person and to forgive whatever he/she did to you.*

Oneness with Universe (Oneness with God)

Fuse Inner Positive Energy with Universal Energy through Turning the Dial

In Fusion of the Five Elements, the collection points were used to gather and store the energy of the organs. From the collection points in the body, the energy was drawn down into the pearl. Now, in Cosmic Fusion the collection points are expanded to encompass the Energy Body and the Universal Aura body. So, not only are we fusing the emotional energy of the physical body (negative emotions, senses and thoughts), but now we fuse the energy of the Energy Body and the Universal Body. This creates a much more complete practice and a highly refined pearl of energy. Fusing the elements is the secret to internal alchemy.

You can think of the different collection points like ingredients for a delicious soup. The more energy we can collect, the more spices we have for our soup. Alchemy is like cooking. What creates a really good soup or food is a combination of ingredients. Alchemy is about fusing the elements at our disposal to create balance and harmony within ourselves and the world we live in.

In the beginning of the fusion practice, it seems intricate and detailed. Once you get used to the formulas it is a very easy and simple practice. Learning a new recipe always takes some time and study, but once you have it in your mind, it is easy. The same is true with the fusion practice.

Forming the Collection Points

The collection points were introduced in Fusion of the Five Elements as a way to collect the negative energy from the organs. This energy was then directed into the pearl through the pakuas. In Fusion of the Five Elements the collection points were associated with the energy of the infants and the protective animals. In Cosmic Fusion, we are going to expand this practice by forming the collection points of the Energy Body and the collection points of the Universal Aura Body. These collection points are a reflection of the collection points in the physical body.

For example, the physical heart collection point is located at heart center just behind the sternum. The Energy Body collection point is located at the Lower Tan Tien at the top of the navel. And, the Universal Aura body collection point is located above the head in the Universe.

These collection points are utilized to create a powerful method of fusing all the elements into the pearl. Once all the collection points are established, the Energy Body and Universal Aura body collection points are rotated. This way there are different colors at the various collection point sites. For example, the collection points at the Energy Body are turned clockwise – the green is on the upper point, red is at the left, white is at the bottom, and blue is on the right. The Universal collection point is rotated counterclockwise – white is the upper point, blue is at the left, green at the bottom, and blue at the right. Creating different colors at each of the collection points takes the fusing process to another level of intensity. These colors are then mixed and blended into the cauldron and formed into a powerful pearl.

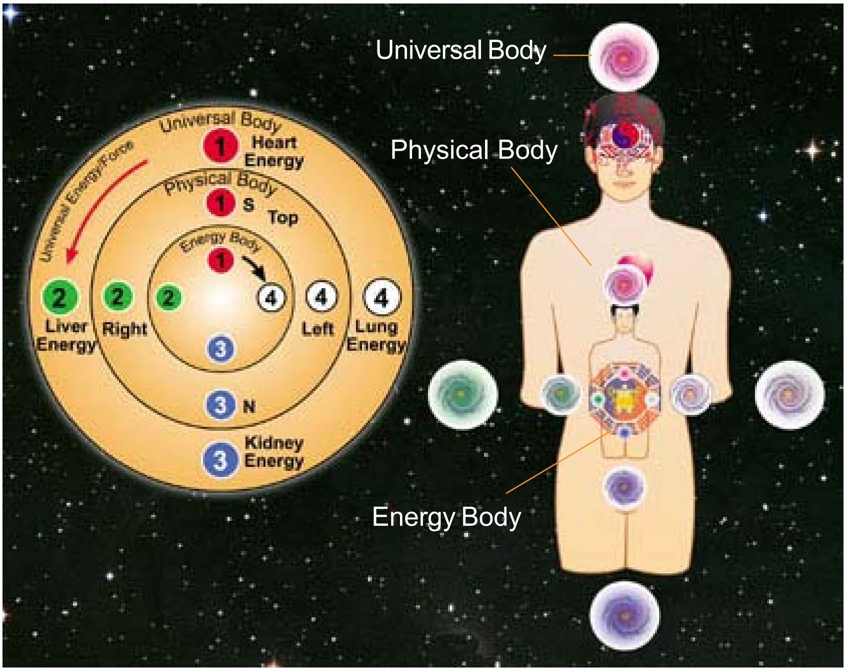
Universal Body
Universal Energy/Force
1
Heart Energy
Physical Body
1
S
Top
Energy Body
1
2
Liver Energy
2
Right
2
4
4
Left
4
Lung Energy
3
3
N
3
Kidney Energy
Universal Body
Physical Body
Energy Body

Fig. 6.29 *Fuse inner position energy with universal energy.*

The Concept of Fusing Inner Position Energy with Universal Energy

The fused collection points are utilized to create a powerful method of fusion all the elements into a pearl.

There are four collections points of the elements; fire, water, metal and wood of the physical body, energy body and universal body.

Once all the collection points are established in the Physical Body, the Energy Body and the Universal Body we use a dial to assist in rotating the patterns of energy.

- Understanding that the physical body is more dense and stable; the collections points are fixed.
- The energy body is not fixed and is very sensitive to external events and the universe. This can be changed.
- Universal Energy is always changing and is not fixed. As the planets and the body's celluar matter, nuclei, electrons, protons all move at different speeds; we use the differences to create energy.
- The Dial: We use a dial of color to understand the different patterns of energy that can be created. Fusing different energies will generate more intense energy and a powerful pearl.

  Moving the color dials assists the imaginative mind to connect to the patterns of energy. It is a formula, a pattern for the mind to follow.

  In this approach we fuse inside to outside universe and outside to inside. You mix fire and water first, then wood and metal together. You can do it in pairs of points and then all four points, creating four balls of fused energy which are then spun very fast above the cauldron where they fuse to form a pearl. You can reproduce as many pearls as you want.

  The intensity can make you feel a oneness with the universe as part of you becomes part of the universe and universal forces.

Activate the Three Fires

The Abdominal, the Kidney and the Heart Fires.

Tan Tien Fire: Bring the fingertips to the navel. Feel the abdominal fire warming and opening the Lower Tan Tien. The abdominal fire elicits deep awareness within the body. This awareness has the power to expand and connect to the entire Universe.

Kidney Fire: Bring the palms to the kidneys, the fingertips to the door of life. Feel the warmth in the kidneys, like a fire under the sea.

Heart Fire: Bring the hands to the heart, palms facing the chest. Visualize a glowing ball of light in the heart. Feel the warm radiant energy of compassion shimmering through the chest. Make the heart soft. Remember the Yin within the Yang.

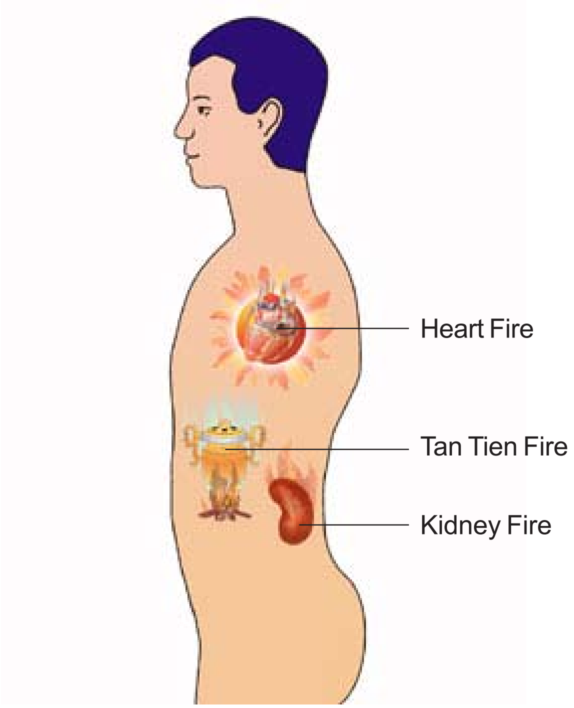

Fig. 6.30 *Activate the Three Fires.*

Create the Front Pakua just behind the navel. Control the formation of the front pakua with assistance from the eyes, ears, nose and mouth. As you balance the energy, let the pakua glow with light.

Chant the Eight Forces by sounding all the trigrams. Visualize the trigrams forming the front pakua. Chant the proper Yin and Yang combinations with each trigram.

Form the Back Pakua: Let it reflect off the front pakua like a mirror. In a similar manner form the **Left** and **Right Side Pakuas**. Feel the energy starting to spiral and blend into the center of the abdomen.

Form the Facial Pakua: The facial pakua is the reverse direction of the abdominal pakua because we are looking at the pakua in front of the face. Feel the energy of the senses and the face gathered and collected. If you have time, chant the eight trigrams of the facial pakua.

Form the Universal Pakua: Feel the immensity of this pakua surrounding your entire body in every direction. Again if time permits, chant the eight directions, forming the pakua.

Be aware of the heart and all the good virtue energy and fuse them together into compassion fire in the heart.

Be aware of the orgasm energy and multiply then to the organs and finally combine them into the heart and becoming one compassion fire of unconditional love.

Form the Collection Points in the Physical Body: Feel the red light in the heart collection point, the blue light in the kidney collection point, the green light in the liver collection point, the white light in the lungs collection point, and the yellow light in the spleen collection point.

Remove Negative Energies from the Organs:

Use the senses to help locate the negative energies in the organs. Look deep inside; smell deep inside; taste deep inside. Locate anything you find that you do not like.

Form the organs' collection points. Spiral the negative energies into the collection points spiraling them into the front pakua where they will blend and transform.

As you spiral the energies from all four Pakuas into the cauldron, you are concentrating them in the cauldron of the body. The cauldron, as the center point of the human being, is the center that contains the essence of all sense, organ and gland energy. The energy directed to the cauldron is transformed and stored for later use in the development of the soul and spirit bodies.

Using the Dial: Feel the reflection of the colors in the collection points in the energy body that surrounds the cauldron.

Feel the outward reflection of the collection points into the Universe. The red ball of light above the head, the blue ball of light deep below the feet, the green ball of light to the right, and the white ball of light to the left.

1. Turn the collection points of the energy body clockwise one station. Turn the collection points of the Universe one station counterclockwise.
2. The physical body collection points always remain stationary. Now the color sequence is red at the physical heart collection point, green below it in the energy body, and white above in the Universal collection point. Below, the physical collection point is blue, the energy body is white, the Universal is green, the right physical body collection point is green, the energy body is blue, and the Universe is red, and on the left the physical body is white, the energy body is red, and the Universe is blue.

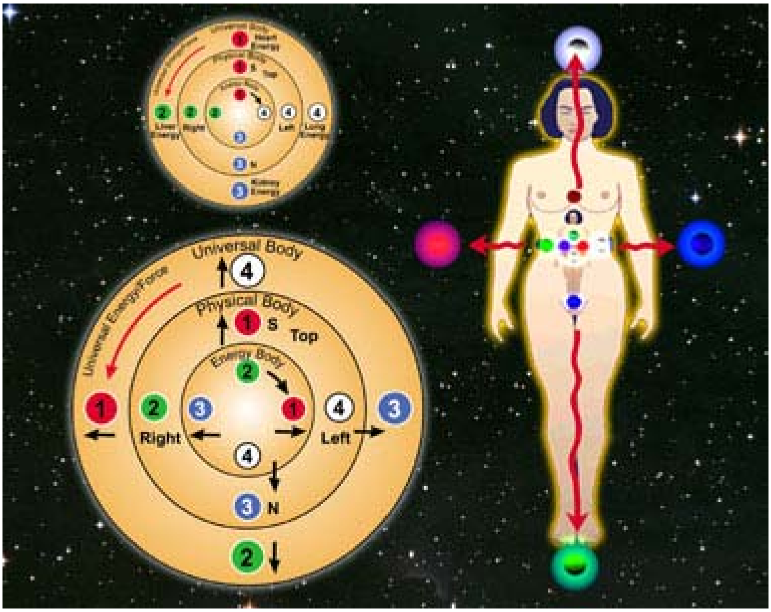

__Fig. 6.31__ Fuse inner positive energy out with universal energy through turning the Dial-1.

3. The fusion process begins here. Fuse the white energy from the energy body with the blue ball in the physical collection point at the perineum, and fuse both of these energies down to the Universal collection point with green light.
4. The same process continues with all the collection points. Next, fuse the green energy body point with the red physical body collection point, moving upward to fuse the white Universal collection point. Feel the energies fusing together above you.
5. Moving to the right side, fuse the blue ball of the energy body collection point with the green ball of the physical body, and outward to fuse with the red ball of the Universal collection point. Feel the energy fuse together at the right side.
6. Moving to the left, fuse the red ball of light of the energy body with the white ball of the physical body and moving out to fuse with the blue ball of the Universal collection point. Feel all these energies fuse together at the left.
7. Now we bring the fusion process back inside the body. Take the fused light below you, drawing it first to the physical collection point and then into the energy body collection point. Bring the fused energy above the head down to the physical collection point and then into the energy body collection point.
8. Draw the fused energy from the right back to the physical body collection point and then to the energy body collection point.

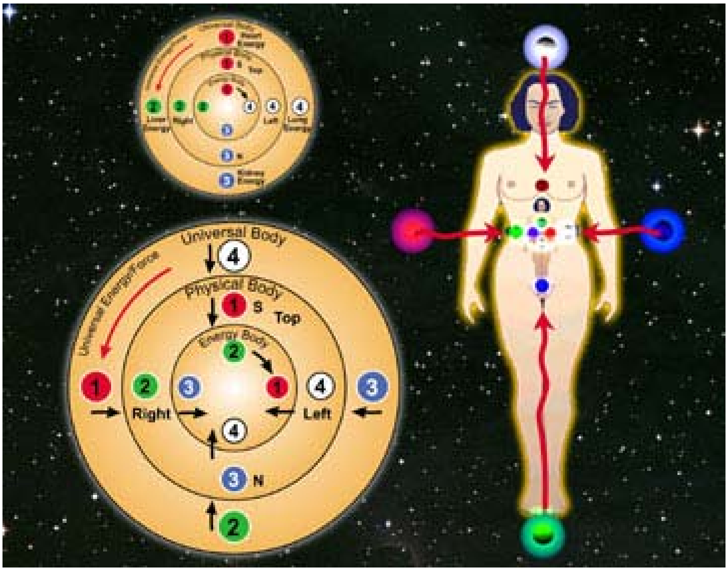

Fig. 6.32 *Fuse the Universal energy force back to inner positive energy and Energy Body through turning the Dial-1.*

Draw the fused energy from the left back to the physical body collection point and then to the energy body collection point. Allow the four balls of fused energy to spiral around the cauldron. Let them increase in intensity and speed, going faster and faster, until they blend in the middle of the cauldron forming the pearl.

9. Feel any residual energy spiral through the pakuas and fuse together into the cauldron, connecting with the brilliant pearl of white light. Feel the colors blend and fuse. This is where the alchemical process comes to life. At any time during the Fusion practice, if you feel the pearl you are working with is lost or diminished simply become aware of the pakuas spiraling into the center of the cauldron.
10. Once you feel the form of the pearl you can simpy fuse the energy from inside of the energy body out to the physical and out to the universe and back and there is no need to be aware of the different color.

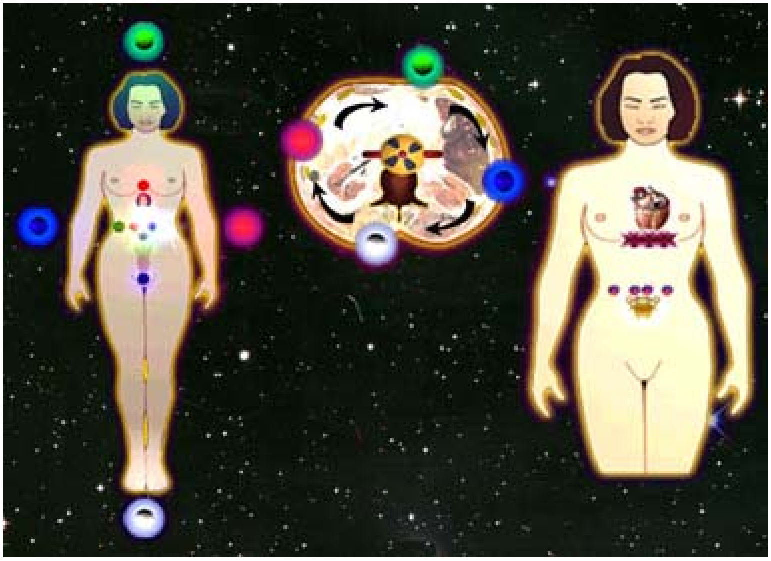

Fig. 6.33 *Four fused energy balls spiral on top of the cauldron, forming one energy pearl.*

Do as above, only the colors are changed and fuse from the energy body to physical body, to the universe and back as Four Balls, spiraling faster and fuse into One Pearl.

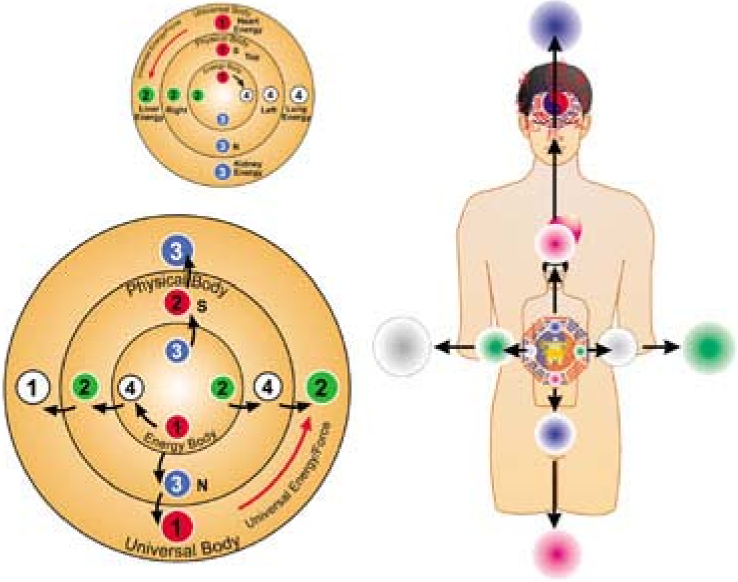
Heart Energy
Liver Energy
Right
Left
Lung Energy
Kidney Energy
Physical Body
S
Energy Body
N
Universal Body
Universal Energy/Force
3
2
1
4

Fig. 6.34 *Fuse inner positive energy with universal energy through turning the Dial-2.*

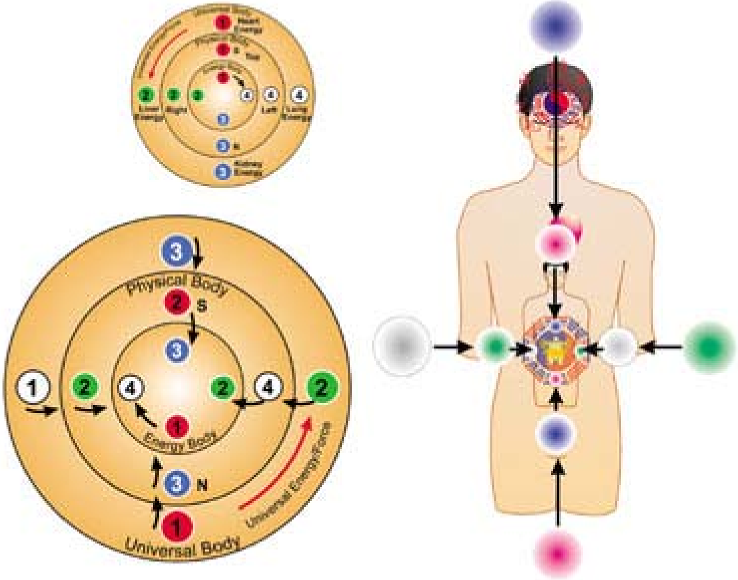
Heart Energy
Liver Energy
Right
Left
Lung Energy
Kidney Energy
Physical Body
S
Energy Body
N
Universal Body
Universal Energy/Force
3
2
1
4

Fig. 6.35 *Fuse the Universe force back to the body and energy body.*

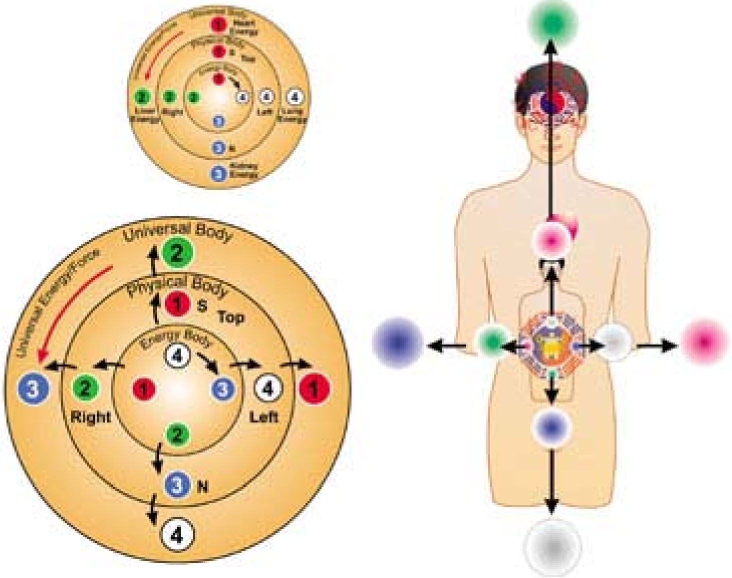
Universal Body
Universal Energy/Force
Physical Body
Energy Body
S
Top
Right
Left
N
1
2
3
4

Fig. 6.36 Fuse inner positive energy with universal energy through turning the Dial-3.

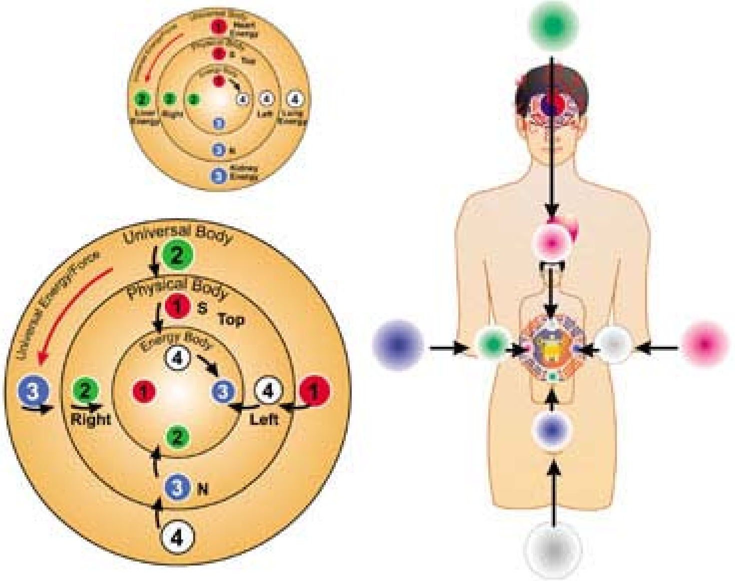
Universal Body
Universal Energy/Force
Physical Body
Energy Body
S
Top
Right
Left
N
1
2
3
4

Fig. 6.37 Fuse the Universe force back to the body and energy body.

Do the same procedure as before using the Four Fused Energy Balls spiraling on top of the cauldron, forming One Energy Pearl as you turn the Dial. You spend more time using Dial-2 and Dial-3 positions.

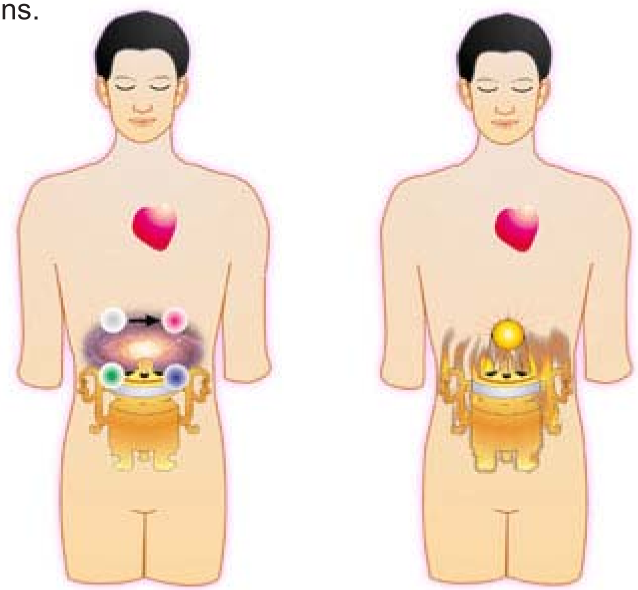

Fig. 6.38 *Forming the Cauldron and the Pearl*

Creation Cycle Foundation Practice

1. Fuse the Pearl to help activate virtue energy. The Inner Smile is the first step to connecting your consciousness to all your organs and glands. This is actually the first step of the Fusion of the Five Elements practice. It is important to remember how powerful and sublime smiling inward can be.
2. The next stage of practice is removing negative emotional energy from the organs. The negative energy is brought from each organ to its corresponding organ collection point and then to the pakua where it is neutralized. When all negative energies are brought to all four pakuas and are under control and all residual energies have been brought to the back and side pakuas, the energies are then spiraled and compressed into the cauldron. There they are formed into the pearl of energy. As you collect the energies and direct their flow to the cauldron, you feel your center point illuminate with the bright light of the pearl.

3. Be aware the energy, physical and universal collection points, fuse them together and fuse back to form the bigger pearl.
4. The pearl is capable of activating, attracting and absorbing a great amount of good virtue energy. For this reason the pearl is circulated to each organ following the Creation Cycle. After practicing Fusion of the Five Elements, the pearl is brought from the cauldron to the perineum and is then circulated in the Microcosmic Orbit.
5. When you are ready to begin Cosmic Fusion, the pearl is brought to the Heart to begin the Creation Cycle. The Heart is the beginning point of this cycle because it is in the Heart that Compassion resides.
6. At this point there is a great awareness of each organ and the energy that each one is supplying. A continuation of this awareness is necessary for properly practicing the Creation Cycle of Cosmic Fusion. It is especially important that all negative energy has been removed and neutralized beforehand; otherwise, as the virtue energy of the pearl increases, so will the negative energy as you move the pearl through the Cycle.
7. In Fusion of the Five Elements, you remove the negative emotion of fear from the kidneys, by first listening to them, because the ears and kidneys have a connection. The removal of fear from the kidneys leaves room for the positive virtue of gentleness inherent in the kidneys to grow.
8. Now, in beginning the Creation Cycle of Cosmic Fusion, you focus your smile on the virtues of joy, love and happiness in the Heart. Feel the virtues growing and purifying. It is important to develop a genuine virtue here. You can spit out words like love, joy and happiness without meaning them. But like the love that is mentioned in the Bible (1 Corinthians 13), it has to be real and true; otherwise it is just a sounding brass or a tinkling cymbal. You have to flesh it out and breathe life into it.
9. Here is the way to do the Creation Cycle of Cosmic Fusion: Feel the unconditional love of the Universe flowing down into the Heart. With this combination of virtues – love, joy, happiness, unconditional love – you can definitely transform yourself, your organs and those around you.

10. Let the positive Heart energies of love, joy and happiness activate the positive Spleen energies of openness and fairness and centeredness. Blend this with the Universal energy and feel the yellow light descending to the Heart and Spleen.
11. Let the positive Heart energies of love, joy and happiness activate the Lung energies of courage, strength and righteousness. Blend these energies with the Universal energy and see the white light from above flow into the Heart and Lungs. Let the Positive Heart energies of love, joy and happiness activate the Kidney Energies of gentleness, softness and calm in the Kidneys. Blend both energies with Universal Energy and see the blue light coming down from above into the Heart and Kidneys.
12. Let the positive Heart energies of love, joy and happiness activate the Liver energies of kindness, generosity and benevolence. Blend these energies with Universal energy and see the green light flowing from above into the Heart and Liver. You can do this meditation a few times.
13. Blend all the positive organ virtues in the Heart, creating Compassion Energy. You may manifest this as a Pearl. Move it in the Microcosmic Orbit spiralling it from 10,000 to 60,000 times a minute.

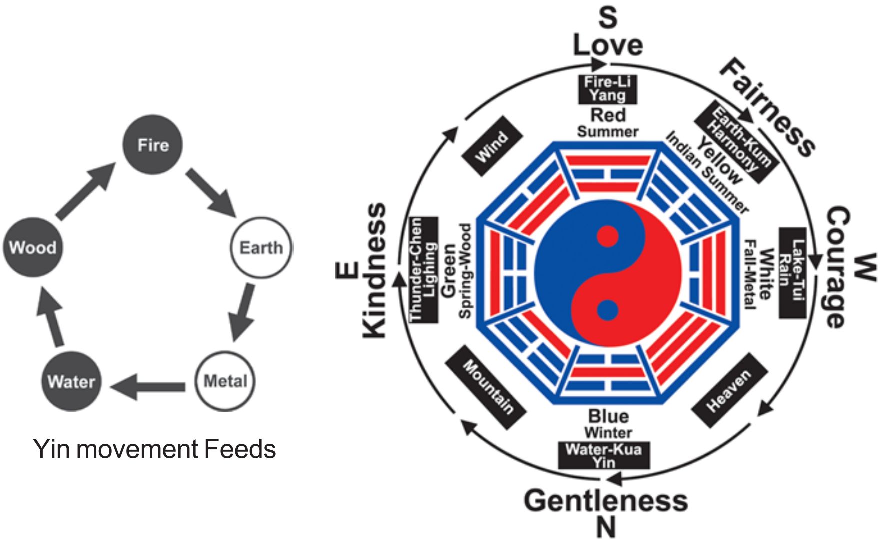
Fire
Earth
Metal
Water
Wood
S
Love
Fire-Li
Yang
Red
Summer
Fairness
Earth-Kum
Harmony
Yellow
Indian Summer
W
Courage
Lake-Tui
Rain
White
Fall-Metal
Heaven
Blue
Winter
Water-Kua
Yin
Gentleness
N
Mountain
E
Kindness
Thunder-Chen
Lighing
Green
Spring-Wood
Wind

Yin movement Feeds

Fig. 6.39 *Creation Cycle*

14. This completes the basic Fusion practice. These steps will be repeated through the rest of the formulas in the book. In the beginning it seems like a lot of work in constructing all the pakuas, chanting, collecting the energy to fuse into the pearl. Once this formula is set up, the whole process happens very quickly. It is not necessary to do all the steps every time. You might only want to chant the trigram one time instead of three. Remember, these extra steps are there to increase the strength of the pearl and expand our consciousness to connect with the Universe. With this powerful pearl it is time to begin Cosmic Fusion with the Creation Cycle.

Begin Cosmic Fusion: Creation Cycle

1. Activate the compassion fire and multi-orgasm energy and combine them together connecting with the universe.
2. Form the pearl through the fuse from the inside to the universe and from the universe and back to the body to the form a pearl. Bring the Pearl down to the Perineum.
3. Feel the pearl connecting to the kidney energy.
4. Listen to the Gentleness Virtue of the Kidneys.
5. The negative emotion of fear has already been removed from the kidneys. The energy remaining in the kidneys is their virtue energy of gentleness. Listen quietly to the gentleness of the kidneys, and be aware of the qualities of gentleness energy: **cold, calm, blue, soft** and **silky** and **tenderness**. Enjoy these qualities. You can intensify them by concentrating on them. Bring the pearl from the perineum to the kidneys, and add the gentleness energy to the pearl at any moment when you feel the pearl weak, just form another pearl.
6. Now bring the pearl back to the kidneys collection point at the perineum, where the gentleness will grow in intensity. All of the neutralized energy that is the pearl now will take on the virtue energy of gentleness.

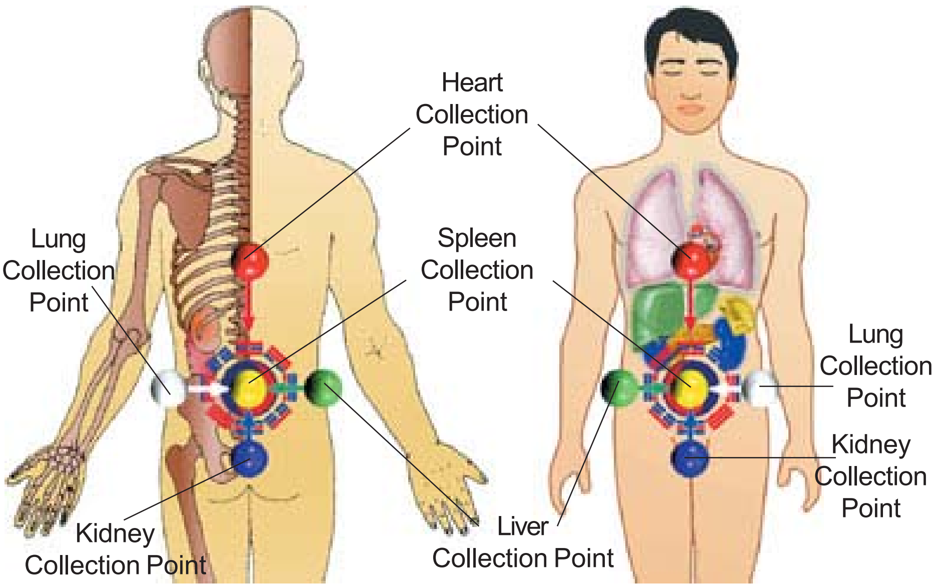
Heart
Collection
Point
Lung
Collection
Point
Spleen
Collection
Point
Lung
Collection
Point
Kidney
Collection
Point
Kidney
Collection Point
Liver
Collection Point

Fig. 6.40 *Front and Back Collection Points*

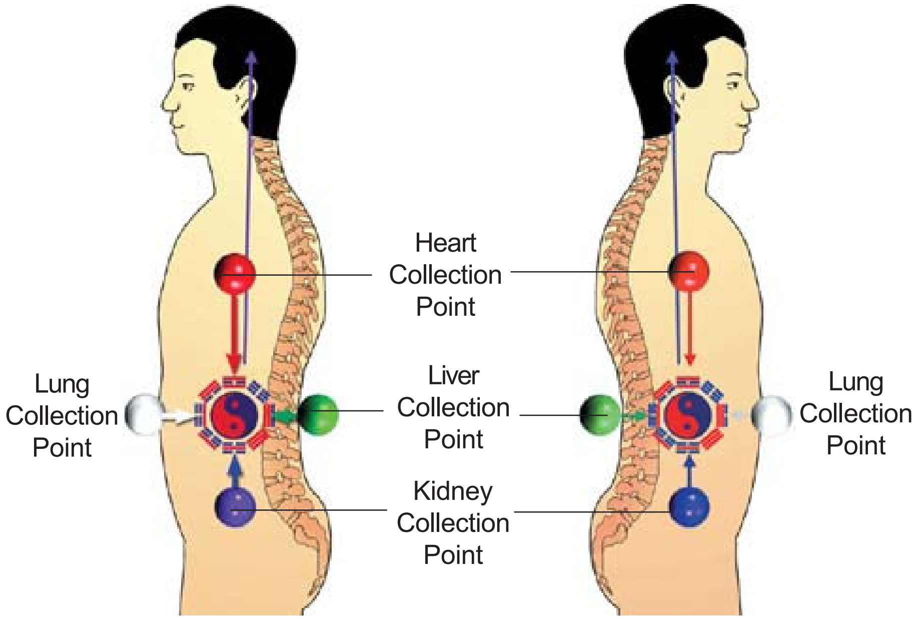
Heart
Collection
Point
Lung
Collection
Point
Liver
Collection
Point
Lung
Collection
Point
Kidney
Collection
Point

Fig. 6.41 *Left and Right Collection Points*

Direct the pearl to the liver. Gentleness energy fuses with the kindness virtue of the liver.

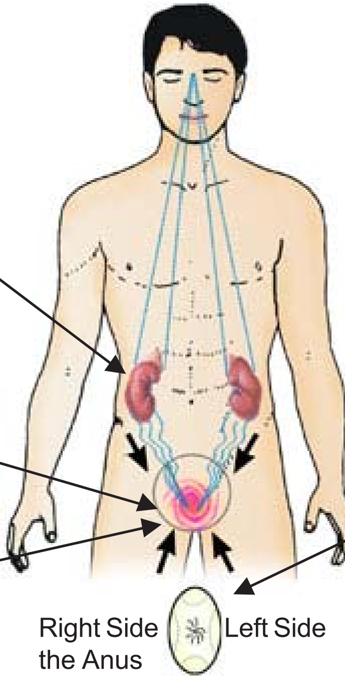

Note: Use the power of the mind more than muscle power to gather the energy.

1. Become aware of the kidneys.

2. Do the Kidneys' Sound (Choooo) subvocally.

3. Pull up the left and right sides of the anus. Send the energy to the kidneys.

4. Pull up the perineum and push the energy ball out. Form the kidneys' collection point.

5. Relax and draw the cold energy to the collection point from the kidneys.

Fig. 6.42 *Forming the Collection Point to gather the Kidneys' Energy.*

7. Gentleness blends with kindness, the virtue of the liver. Look into the liver, connecting the eyes with the liver. Be aware of the kindness energy there. As you bring the pearl carrying the gentleness energy from the kidneys toward the liver, feel the gentleness energy activate and enhance the kindness energy. Enjoy the kindness energy. Feel the kindness intensify. Circulate the kindness energy in the liver. Its qualities can feel **strong, round, smooth, soft, green, sweet and fragrant, warm and pleasant**. You can feel satisfied.

8. Absorb the kindness energy into the pearl, and bring the pearl to the liver collection point (at the level of the navel and in line with the right nipple).

   Here the combination of gentleness and kindness will intensify. Remember to relax and smile. Direct the pearl to the heart. Gentleness and kindness energy fuse with love, honor and respect and joy in the heart.

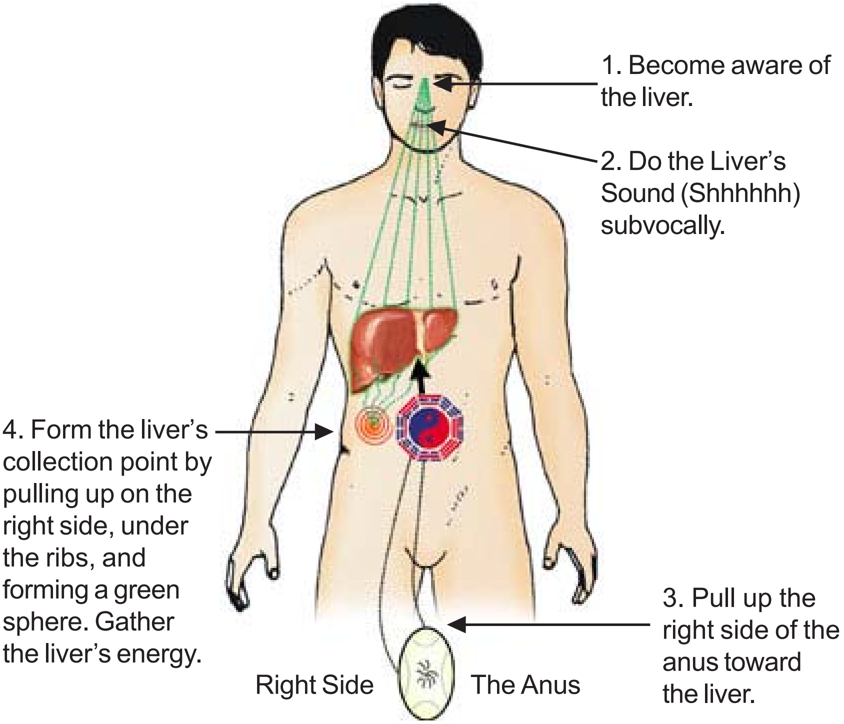
1. Become aware of the liver.
2. Do the Liver's Sound (Shhhhhh) subvocally.
4. Form the liver's collection point by pulling up on the right side, under the ribs, and forming a green sphere. Gather the liver's energy.
3. Pull up the right side of the anus toward the liver.
Right Side
The Anus

Fig. 6.43 *Forming the Liver's Collection Point to gather the Liver's Energy.*

9. Connect the tongue with the heart. Using your mind's eye and senses, allow the pearl, now consisting of gentleness and kindness, to flow up to the heart. Let the virtue energy of the pearl activate and enhance all the loving energy, joy and happiness, honor, respect and peace in the heart. These good virtues of the heart can feel **straight** and **open, bright red, warm, deep, calm, comfortable** and **satisfying**. Enjoy the virtuous feelings of the heart. The feeling will be different for each person as the energy blends. Feel the warmth and openness in the chest.

10. Let the pearl absorb the honor, respect, love and joy. Bring the pearl down to the heart collection point behind the sternum.

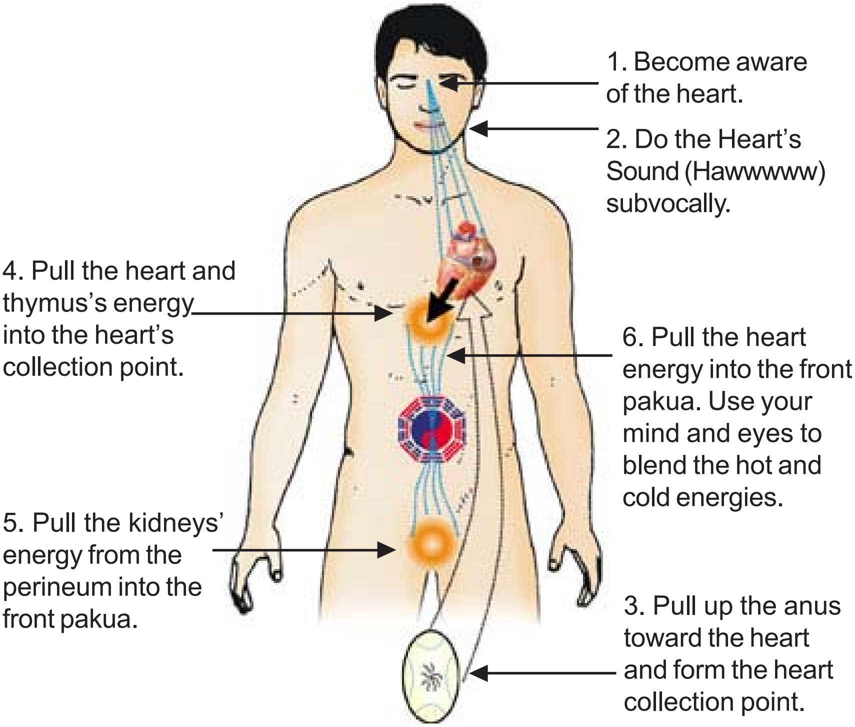
1. Become aware of the heart.
2. Do the Heart's Sound (Hawwwww) subvocally.
4. Pull the heart and thymus's energy into the heart's collection point.
6. Pull the heart energy into the front pakua. Use your mind and eyes to blend the hot and cold energies.
5. Pull the kidneys' energy from the perineum into the front pakua.
3. Pull up the anus toward the heart and form the heart collection point.

Fig. 6.44 *Form the Heart's Collection Point and blend the Heart's Energy with the Kidneys' Energy in the Pakua.*

The pearl can now blend the virtue energies of the heart in a balanced way with the kindness and gentleness energies already within the pearl. Feel the collection point glowing. Then direct the pearl to the spleen/pancreas.

Note for Women: Women should be careful when the pearl is in the heart center. As a natural endowment women tend to have more loving energy in the heart. When a woman loves she often forgets herself. This amount of love can heat up the heart center and the practice can feel uncomfortable. It is best not to remain for too long in the heart if you feel uncomfortable.

11. Be aware of the connection of the mouth with the spleen and pancreas. As the pearl arrives at the spleen and pancreas, the honor and respect, love and joy energy that is now part of the pearl fuses with the spleen virtues of fairness and openness. The fairness and openness virtues of the spleen can feel **expansive, clean, dry, bright yellow, smooth, warm, clear** and **soft**. Enjoy these feelings.
12. Let the pearl absorb the fairness and openness energy. Feel the pearl take on the qualities of the fairness and openness virtues in addition to the kindness, gentleness, honor and respect virtues that already are present there. Bring the pearl to the spleen collection point behind the navel at the front pakua.

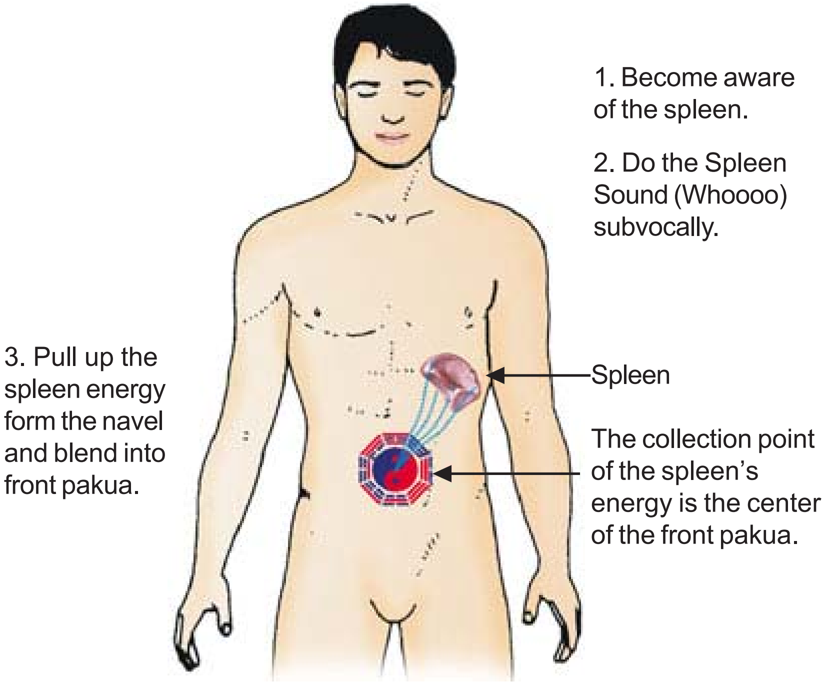
1. Become aware of the spleen.
2. Do the Spleen Sound (Whoooo) subvocally.
3. Pull up the spleen energy form the navel and blend into front pakua.
Spleen
The collection point of the spleen's energy is the center of the front pakua.

***Fig. 6.45** Gathering the Spleen's Energy Directly into the Pakua.*

Feel the fairness and openness energy intensify as it combines and blends with all the neutral and virtue energies already present in the pearl. Direct the pearl to the lungs.

13. The Energy of fairness and openness fuses with the courage and righteousness virtues of the lungs.

Connect the nose to both lungs. As the pearl arrives at the lungs, split the pearl into two pearls, one for each lung. The energy of fairness and openness that is now part of both pearls fuse and enhances the courage and righteousness energy of the lungs. When you feel righteous, you can feel **tall, straight, uplifted, comfortable, strong, firm, proud** and **satisfied**. The energy can be **fresh, bright, white** and **pure**. Enjoy the qualities of this energy.

14. Let the pearls absorb the courage and righteousness energy. Bring both pearls to the collection point of the lungs (at the level of the navel and in line with the left nipple). Feel the courage and righteousness energy intensify as it circulates in the lungs. Let the pearls blend these virtues with their kindness, gentleness, honor, respect, fairness and openness energies.

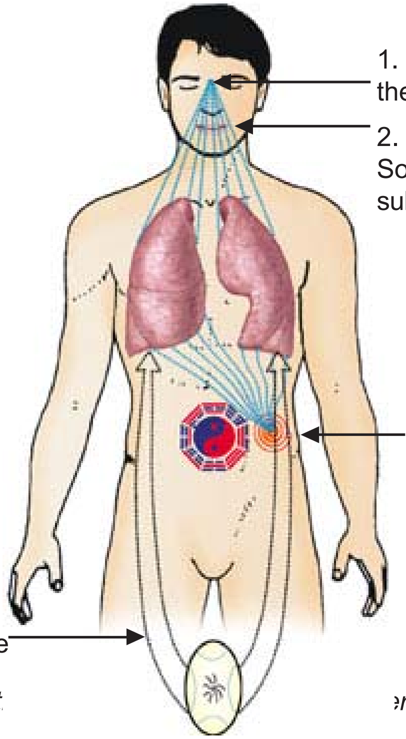

1. Become aware of the lungs.

2. Do the Lungs' Sound (Sssssss) subvocally.

4. Form the lungs' collection point by pulling up on the left side and forming a white sphere. Gather the lungs energy.

3. Pull up the right and left sides of the anus to the lungs.

Fig. 6.46 *Forming t* *er the Lungs' Energy.*

Direct both pearls to both kidneys. Courage and righteousness fuse with gentleness: the cycle begins again at the kidneys.

15. As the pearls arrive at each kidney, the combined virtues of the pearls (especially the energy of courage and righteousness) enhance gentleness and the cycle begins again at the kidneys. Repeat the cycle two more times. As you circulate the pearl in the second cycle, you will begin to notice that as the pearl arrives at each organ, the virtues will grow by themselves. Little by little, the energy will blend more evenly.
16. For the third cycle, bring the pearl from collection point to collection point without going into the organ first. With your attention on the collection point, the energy will flow into the organ without conscious effort, and will join with the virtue energy to flow back toward the collection point. In this way the virtue energies are enhanced and increased each time you bring the pearl to a collection point. If you wish you can practice the third round of the Creation Cycle three more times simply by bringing the pearl to the collection points. Each time you blend the energy, you increase it.
17. Your energy now begins to have the quality of fine soup, with the proper combination of ingredients. Be aware of the quality of this energy. It is very special. Fuse all good virtues energies into compassion energy. Creating a pearl of compassion energy requires a tremendous amount of purified energy. Compassion energy is not just love or kindness, gentleness or openness, but is a combination of all the good virtues' energies, blended in proportion to become compassion energy. When you fuse the proportional blend of good virtue energy into a pearl of compassion energy, you will feel more centered. The energy of this pearl feels very different from the pearl you first formed. Move the compassion energy through the microcosmic orbit.
18. When you feel the pearl of compassion energy strongly, move it down to the perineum. Using your mind and senses, begin to move all of the compassion energy through the microcosmic orbit, letting its bright pearl shine as it travels along. Feel the different quality of this energy. Be aware of the loving, comfortable, very beautiful way the organs feel as the energy runs through and spreads into them, filling them with life-force energy. If you feel tired at any time, you can temporarily rest the pearl at your navel and practice the spinal cord breathing technique described in the Chapter preceding this meditation. Return your concentration to the pearl to form it again and return it

to the microcosmic orbit for circulation. Move the pearl through the microcosmic orbit 9-18 times.

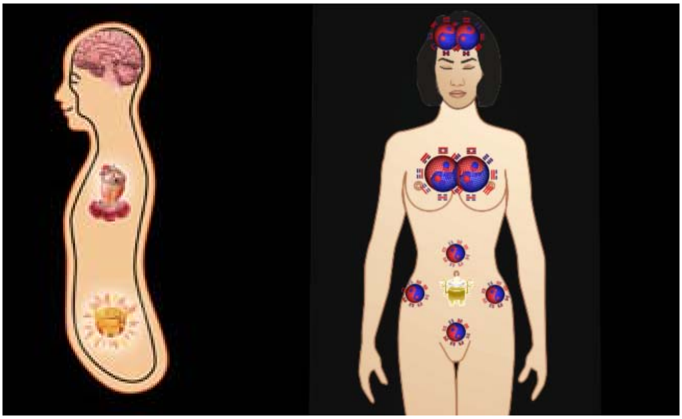

Fig. 6.47 *Circulate the pearl in the Microcosmic Orbit.*

Shoot the Pearl above the Crown.

1. As the pearl runs through the Microcosmic Orbit, begin to activate the Cranial Pump. Press your tongue to the palate, clench your teeth, pull back your chin and roll your eyes up toward the crown. Feel the pulsing in your heart and in your head at the crown. You can touch your pulse at the wrist to help your awareness of the pulse both at the wrist and the crown. Your mind also helps to activate the Cranial Pump.

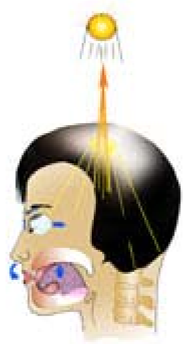

Fig. 6.48 *Pull the eyes into their sockets.*
Roll the eyes up and look up to the crown.

2. Inhale, be aware of and squeeze the anus and then exhale quickly to push the pearl up through the crown.
3. You might be aware of a light, beam or frequency shooting out of your crown. This is the lead light first introduced in Fusion of the Five Elements. Sometimes it is called an indicator light because it will indicate how high up your pearl will be able to go. Just as a light is used to guide helicopters to their landing point, so will you use the lead light to guide the pearl back to the crown point.
4. Direct all of your senses to push the pearl up to the area six inches to two feet (15-60 centimeters) above your head. Move the pearl up and down, left and right.
5. Feel the Universal Force, the forces of the North Star and Big Dipper above your head, and the Cosmic Particle Force in front of you as they shine down to your pearl. Feel your pearl start to absorb this energy and then feel it expand.
6. Bring your awareness to your feet. Feel the energy supplied throughout your body through your feet. This is the Earth Energy. The pearl continues to expand with all the energy coming to it from outside the physical body.
7. When you are ready, activate the Cranial Pump again. Press your tongue up, clench your teeth, pull back your chin, pull up your anus and look up to the crown. Feel the beating in your heart and the pulse at your crown.
8. Activate the lead light, and let it shine up from your crown. It may feel like a frequency going out of the crown. Draw the pearl down to the lead light. Inhale and draw the pearl down.
9. Circulate the pearl in the Microcosmic Orbit. Then bring it down to the navel and to the cauldron at the body's center. Collect the energy at the cauldron.
10. As you collect the energy and relax the mind, the pearl may disintegrate. Sometimes when you release the pearl, it vanishes. It is returning its energies, now enhanced by all virtue energies and the outside forces, to the organs and glands. They become stronger and healthier each time you practice. Each time you gather the energy and form the pearl again, the pearl will be stronger.

Summary

1. Trust and Believe (Concept and Desire).
2. Visualization turns into activation.
3. Think and let it happen. It will become easier with practice.

Once you understand the practice and go through the practice a few times, you can just simply use the mind and recall the past experience and there is no need to go through all the steps. Think and let it happen.

Fusion Opening: Fusing Universal Energy

1. Inner Smile–Activate Compassion first (Love and Sexual Energy combined).
2. Activate Three Fires.
3. Create Pakuas; Front, Back, Facial, Universal.
4. Form Cauldron, spin Pakuas.
5. Form organ Collection Points of Physicial Body.
   Energy Body and Universal Body.
6. Remove Negative Energy from the Organs.
7. Sense the Colors of the Collection Points imagine the Dial-1,2,3 of Colors.
8. Fuse inner Positive Energy of the Energy Body and Physical Body with Universal Energy through turning the Dial-1,2,3.
9. Fusion of the Universe Energy back into Body and Energy Body.
10. Create Four Balls of fused energy and spin them very very fast above the Cauldron.
11. Form a Pearl.

Begin Cosmic Fusion: the Creation Cycle.

Be aware of the heart and cauldron activating the compassion fire.

1. Listen to the Gentleness Virtue of the Kidneys.
2. Gentleness Energy activates the Kindness Virtue of the Liver.
3. Kindness Energy activates Honor, Respect, Love in the Heart.
4. The Energy of Honesty and Respect activates the Fairness and Openness Virtues of the Spleen and Pancreas.
5. The Energy of Fairness and Openess activates the Courage and Righteousness Virtues of the Lungs.
6. Courage and Righteousness enhance Gentleness: the Cycle Begins again at the Kidneys.
7. Combine all Good Virtues energies into the Heart and create Compassion Energy.
8. Move the Compassion Energy through the Microcosmic Orbit.
9. Shoot the Pearl above the Crown.
10. Practice Chi Massage.

Chapter 7

Opening the Thrusting Channels

Theory of the Thrusting Channels

After the positive emotional energy of the organs is absorbed into the pearl, it is then circulated in four of the four special channels: the Microcosmic Orbit (Governor and Functional Channels), the Thrusting Routes, and the Belt Routes. The Thrusting Routes run through the center of the body, linking the “chakra” centers. The Belt Routes spiral around the body, strengthening the aura and providing a form of psychic self-defense. The second part of the Cosmic Fusion practice involves opening these three Thrusting Channels.

The Thrusting Channels are a powerful means to clean out the negative emotions and detoxify the organs and glands. Set into action, the three Thrusting Channels will serve as strong and efficient ways to check your center and observe the status of your emotions.

The Thrusting Channels run from the perineum to the crown and are separately identified as the Middle, Left and Right Thrusting Channels. The Middle Thrusting Channel begins directly at the perineum and runs vertically up to the crown. The Left Thrusting Channel is located one and a half to three inches to the left of the Middle Channel, while the Right Thrusting Channel is one and a half to three inches to its right.

When you are feeling emotional, you can check the Thrusting Channels, especially the Middle Channel, to discern whether you are off-center. You will know if you are out of balance when you feel the emotions running more on the left side or on the right, instead of through the middle of the body. If you center the emotions through the Middle Thrusting Channel, and keep the Left and Right Thrusting Channels in balance, you will very quickly center yourself. The Thrusting Channels serve as guidelines to decision-making in your daily life.

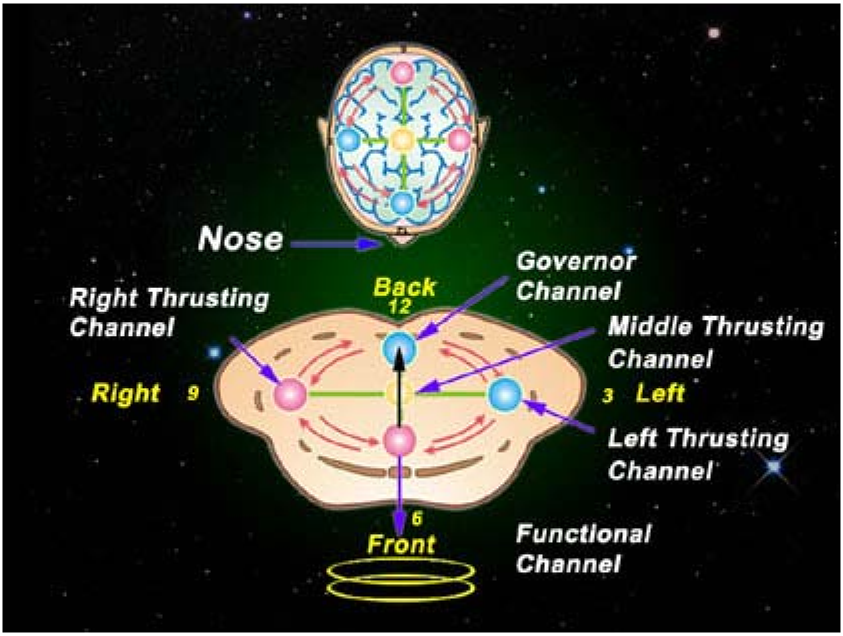
Nose
Governor Channel
Back
12
Right Thrusting Channel
Middle Thrusting Channel
Right
9
3
Left
Left Thrusting Channel
6
Front
Functional Channel

Fig. 7.1 *Four Psychic Channels*

Procedure for Opening the Thrusting Channels

The Thrusting Channels need the tremendous energy of good virtue to help them open. Begin by forming a pearl of virtue energy. With your continued, conscientious practice of the compassion fire and the fuse the inside with the outside energy to form the pearl, you will be able to assimilate the energies into a new and stronger pearl, a pearl that responds to your intention. This means that all you need do to form a new pearl is concentrate on the cauldron and the pearl in it, and on the front, back, left and right pakuas and fuse inside with the outside energy. Control of the internal weather, senses, organs and emotions will be possible in as little as a few seconds. You do not have to go through all the steps of chanting and collecting the energies every time you do the Fusion practice. It is helpful and creates a stronger experience if all the steps are taken. So the first step to opening the Thrusting Channels is to construct the pakuas, chant the eight directions, form the collection points of the Physical Body, the Energy Body and the Universal Body, and fuse them together into the pearl.

The pearl will be thrust up from the perineum (or testicles in men). Each Thrusting Channel precisely penetrates the body.

Although it may seem difficult at first, using your beautiful pearl will help make it easier. Other aids such as pulling up the anus and inhaling help the energy move in each of the three Channels. Once you are able to pass the energy through the Channels, you will find that the energy will move by itself without assistance.

Note: You must be very careful and gentle with this procedure. If the energy reaches the heart and head too quickly using the Thrusting Channels, the heart will become congested and you may experience pressure. If the heart becomes congested and overheats with energy: (1) Do not thrust above the diaphragm and (2) Practice the Heart Sound. Do not overheat the heart, liver or head.

Initially the Thrusting Channels have approximately a one-half to one inch diameter. Once they are activated, they can become very wide. At first concern yourself with drawing the energy directly up. Later, when you have become familiar with the process, the energy can be spiraled up each route.

Seated Position

Like many other Universal Tao exercises, the methods of opening the Thrusting Channels are practiced from a seated position. Sitting on a chair with your feet touching the floor and your spine erect connects you with two important energy “wires”: Earth Energy and Heavenly or Universal Energy. The Energy Body you eventually create needs the ground wire of Earth Energy to hold and support it. Otherwise, you can become disoriented and lost. Once your are more in control of the energy you can practice standing, or lying down on a bed or the floor.

Anal Muscle

If you do not strengthen your anal muscle, you will not progress very far in any of the Universal Tao System's practices. The anus muscle is connected to the perineum and everything in the body above it. If you invest all of your money on building the top floors of a building, and do not invest anything in the foundation, the building will fall. Think of the anus and perineum as your foundation. The anus muscle is used to seal your "lower gate." Closing this orifice helps to retain and prevent loss of your energy.

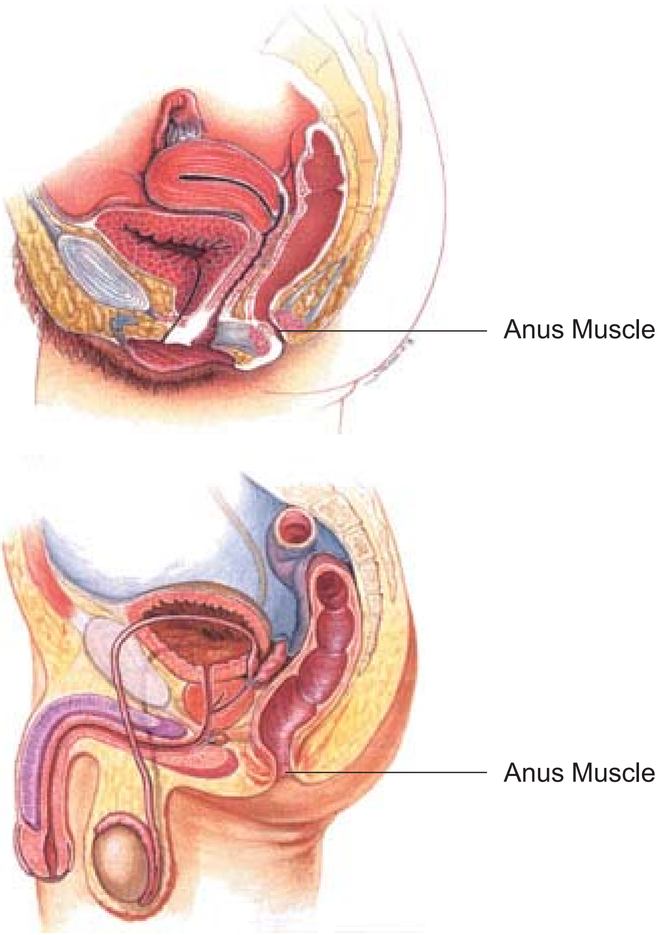

Fig. 7.2 *Anal Muscle*

Saliva

Saliva is yet another aid in performing these exercises. Once you have been able to bring the energy up through all the routes to the crown successfully, saliva generation becomes very important.

In the body, the moon influences the kidneys, which are linked with the salivary glands, the ovaries and semen. When the kidneys are stimulated during sex, the Chi is raised up through the spine, then it flows down in the saliva, and the kidneys' essence transforms into sexual energy. The saliva has two functions: purifying the mouth, and nourishing our immortality, the water of life. The energy from the saliva lubricates 100 joints, and it stimulates and harmonizes the organs. It is very important to enhance the production of saliva by absorbing the moon's energy through the breath absorption exercises. It is a sign that the techniques are working when the saliva floods into the mouth.

Although saliva is a lubricant for the Channels, it first serves as an agent to burn and clean them out of any impurities that could block the flow of energy. So when resting during the practice of opening the thrusting channels move your tongue around to create saliva. Make it thick and swallow down forcefully, feeling it go down the channels that you are working on (See the book *"Elixir Chi Kung"* by Mantak Chia).

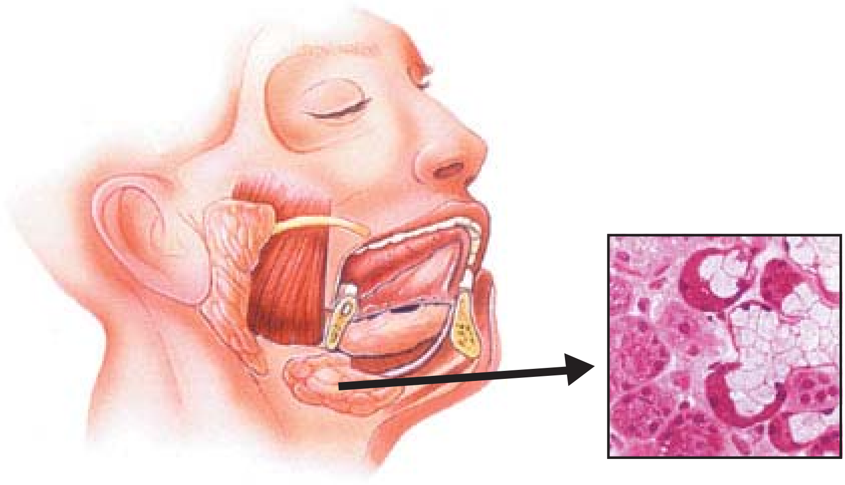

Fig. 7.3 *Saliva*

A person experiences the Thrusting Channels according to his or her kinesthetic or visual perceptions. The majority of people are more kinesthetic and can undeniably feel the quality of the condensed energy as it passes through each level of the channels.

Visual people can see patterns and colors. The Thrusting Channels, like the organs they intersect, have corresponding colors. The Left Channel is Red, Middle - White and Right - Blue. Regardless of your perceptual tendency, both the senses and the organs are used to control the Thrusting Channels.

The denser or more polluted the organ, the more energy is requires to pass through the affected Thrusting Channel.

As a safety precaution and as preparation to be able to move the energy to prevent it from sticking, the Thrusting Channels are opened in four stages. The pearl is brought (1) up to the diaphragm; (2) up to the neck; (3) up to and out of the crown; (4) down to the feet, into the ground; and up to the crown.

This will be done a total of nine times at each stage in the following sequence: (1) left, (2) middle, (3) right, (4) middle, (5) left, (6) middle, (7) right, (8) middle and (9) left.

Men & Women Practices: Open Channels Differently

Once the channels have been completely opened, men and women practice differently. Without stopping, men will thrust the energy directly up through the Thrusting Channels and out of the crown. They then will bring the energy back in to flow down the Thrusting Channels and down through the soles of the feet into the ground. They can continue to bring the flow up around the body and in through the crown. Without stopping, women will thrust the energy down through the soles of the feet into the ground and then up through the Thrusting Channels to and out of the crown. They then spray the energy down around the body, tuck it under the feet and draw it back in through the toes.

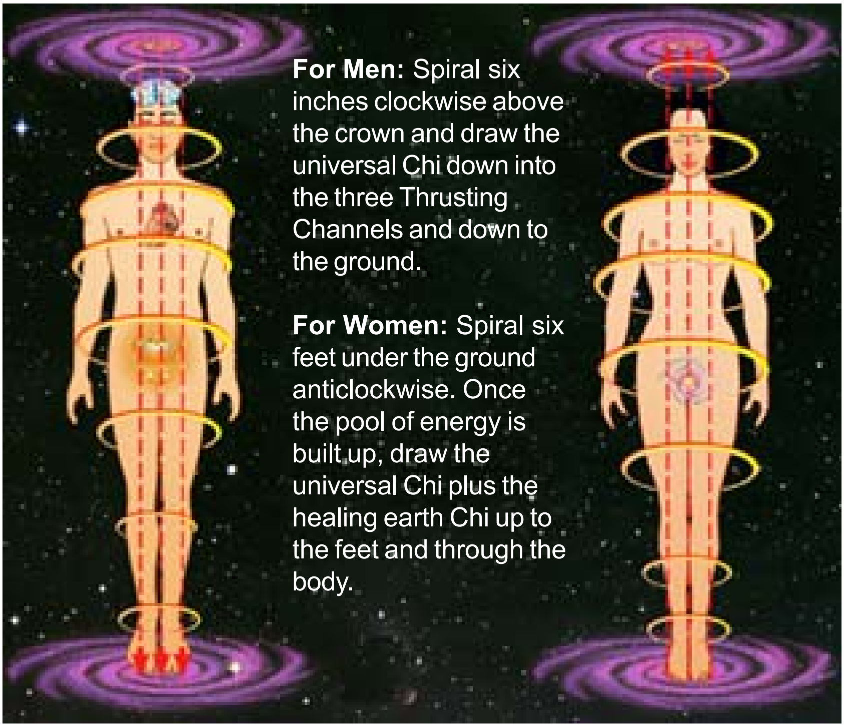

For Men: Spiral six inches clockwise above the crown and draw the universal Chi down into the three Thrusting Channels and down to the ground.

For Women: Spiral six feet under the ground anticlockwise. Once the pool of energy is built up, draw the universal Chi plus the healing earth Chi up to the feet and through the body.

Fig. 7.4 *Cosmic Thrusting and Belt Channels*

Practice of Opening the Thrusting Channels

1. Form a Pearl and Send it to the Perineum

In a seated position smile down to your organs and glands, and put yourself in a state of relaxation and happiness. Begin with the meditation exercises of the Creation Cycle to form a pearl of compassion energy.

As you have learned, there are many steps in forming such a pearl. However, once you have programmed yourself to all of these steps through continued practice, you can simply push the button and the whole program is activated. When you are able simply to sit quietly and feel the sensation of compassion energy forming, you are pushing the button and the entire process takes very little time. In other words once you have experienced compassion energy, all you need to do is to recall the state to create the energy again. You will not need to go through the entire process each time. You will know when you look at the pearl whether or not the process has been completed. If you find something is not right, you can review the steps and find out where the problem is.

As the pearl is forming, use your mind and senses to control it. Slowly move it down to the perineum. At any time during your practice, if the pearl seems to weaken, diminish, or you lose track of it, return to the navel and quickly form a new pearl. Bring that pearl to the perineum and resume practice.

While you are learning how to bring the energy up through the channels, it is easier to focus on the organs the channel passes through than to focus on the channel. You might think of the process as learning how to read. Once you are able to read, you no longer have to think about your ABC's.

2. Stage One: Thrusting Channels to Below the Diaphragm

A. Left Thrusting Channel to the Diaphragm

In all Universal Tao practices it is important to train the power of the mind and senses to control our energy (Chi). The power of mind, eye and heart guides the Chi, providing a way to control and direct your life force. Breathing techniques also help to activate the Thrusting Channels. It is necessary to have some control over independent breathing through each nostril. Since in the beginning this is difficult to do, you can use your right index finger to cover the right nostril. Feel the breath on the left side only. Use the left eye, ear, and nostril to assist you in drawing up the energy.

1. Left Testicle (Men), Left Ovary (Women) and Left Kidney
   a. Sit up, using your mind, heart and eye power to look inside and be aware of the perineum. Also become aware of your left nostril, left ear, left eye and left anus.
   b. Activate the Left Thrusting Channel by inhaling through the left nostril in short sips (about ten percent of lung capacity with each breath.)
   c. Then men slightly pull the left testicle and the left side of the anus up toward the left. Women slightly pull the left side of the anus up toward the left. Draw the pearl up with each sip of air. The left eye should internally look down, and then use the mind/eye power to look up as the pearl is drawn up. (The physical eyes do not actually look up.) Women direct the pearl to the left ovary and the left kidney; men direct the pearl to the left kidney.
   d. Exhale, relax the eye and the muscles, letting the pearl drop to the left side of the anus (and left testicle in men), and rest. Rest and feel the left side become lighter as the energy begins to flow up without effort. Practice nine times. Each time you rest use your mind to trace the Left Thrusting Channel.

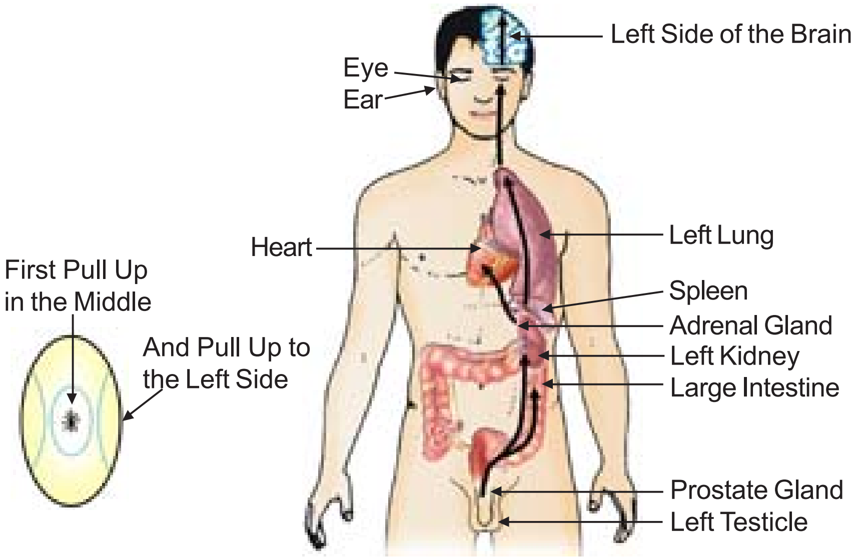
Left Side of the Brain
Eye
Ear
Left Lung
Heart
First Pull Up in the Middle
Spleen
Adrenal Gland
And Pull Up to the Left Side
Left Kidney
Large Intestine
Prostate Gland
Left Testicle

Fig. 7.5 *Left Part in the Male*

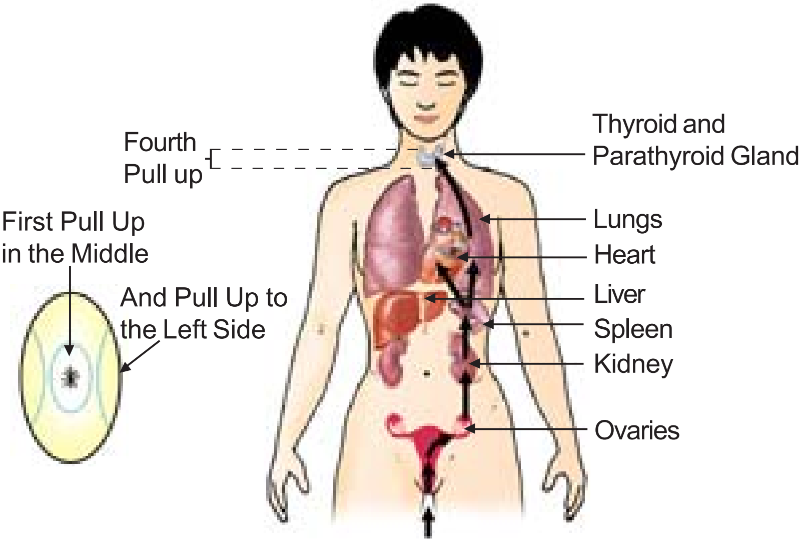

Fig. 7.6 *Left Part in the Female*

2. Spleen

Close the right nostril, inhale, and pull up. As you sip in the air, use the mind power and the left eye, ear and nostril to draw the energy up from the left testicle (men), through the left ovary (women), the left kidney, and up to the spleen. Exhale; release the energy down. Practice nine times. Rest, and feel the energy flow up the Left Thrusting Channel to the spleen under the diaphragm.

B. Middle Thrusting Channel to the Diaphragm

1. Prostate (Men) or Cervix (Women)

Men concentrate on the middle of the scrotum, perineum, and anus. Women concentrate on the middle of the perineum and anus. Inhale, and pull up the middle of the scrotum (men), perineum and anus. Use all the senses to draw the energy up to the prostate (men) or cervix (women). Look up with both eyes as you do so. Relax the eyes and anus muscle, and let the pearl drop back down to the perineum and scrotum (men). Feel the energy begin to flow up as you rest. Do this nine times.

2. Small/Large Intestine, Aorta and Vena Cava

Inhale and pull up the scrotum (men only) and anus. Use the eyes, ears, and nose to draw the energy to the prostate/cervix, small intestine, large intestine, aorta and vena cava. Do not thrust beyond the diaphragm. Practice nine times. Rest, and feel the energy flow up through the Middle Thrusting Channel to the vena cava.

3. Stomach and Pancreas

Follow the above procedure and thrust through the middle organs to the stomach and pancreas. Practice nine times. Swallow the saliva, rest and feel the energy flow in the Middle Thrusting Channel up to the diaphragm.

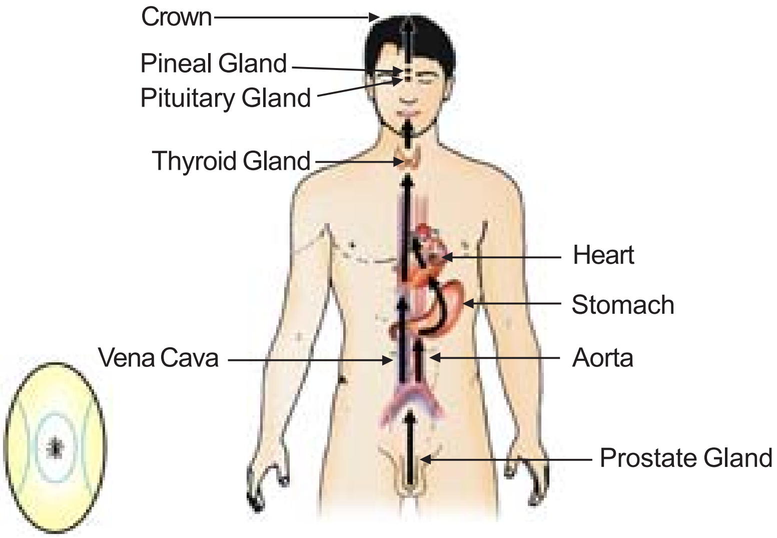
Crown
Pineal Gland
Pituitary Gland
Thyroid Gland
Heart
Stomach
Vena Cava
Aorta
Prostate Gland

Fig. 7.7 *Middle Part in the Male*

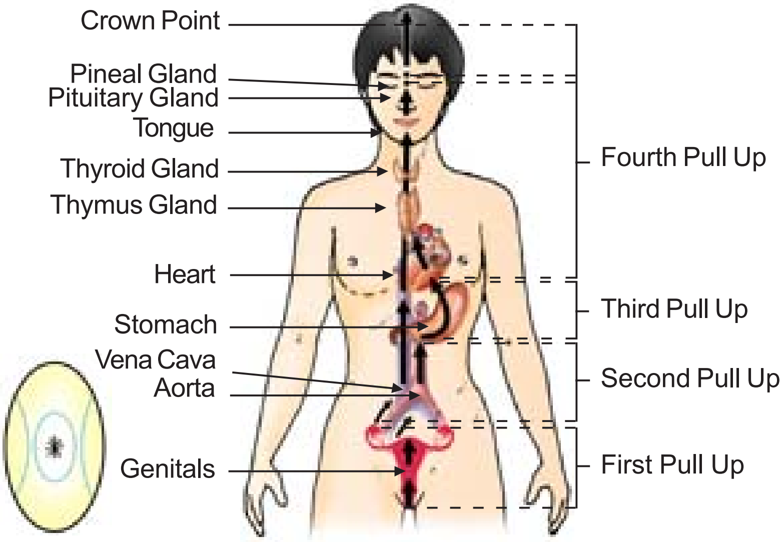
Crown Point
Pineal Gland
Pituitary Gland
Tongue
Thyroid Gland
Thymus Gland
Heart
Stomach
Vena Cava
Aorta
Genitals
Fourth Pull Up
Third Pull Up
Second Pull Up
First Pull Up

Fig. 7.8 *Middle Part in the Female*

C. Right Thrusting Channel to the Diaphragm

1. Right Testicle (Men), Right Ovary (Women) and Right Kidney
   a. Close the left nostril with the index finger of the left hand. Use the right eye, ear and nostril to direct the energy up the right channel. Put your right hand on the right kidney.
   b. Inhale in short sips equal to ten percent of lung capacity. Look up with the right eye. Pull up the right testicle (men only) and the right muscle of the anus toward the right side. Direct the energy up the right Thrusting Channel to the right ovary (women) and kidney. Practice nine times.
   c. Exhale and rest. Let the pearl drop down, returning to the perineum and right testicle (men only.) Feel the right side become lighter as the energy flows up the Right Thrusting Channel to the right kidney without effort.

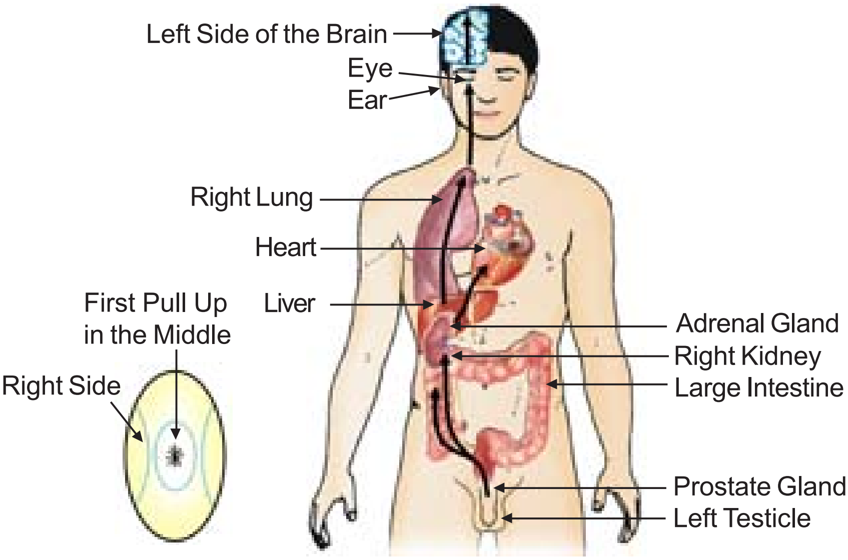
Left Side of the Brain
Eye
Ear
Right Lung
Heart
First Pull Up in the Middle
Liver
Right Side
Adrenal Gland
Right Kidney
Large Intestine
Prostate Gland
Left Testicle

***Fig. 7.9** Right Part in the Male*

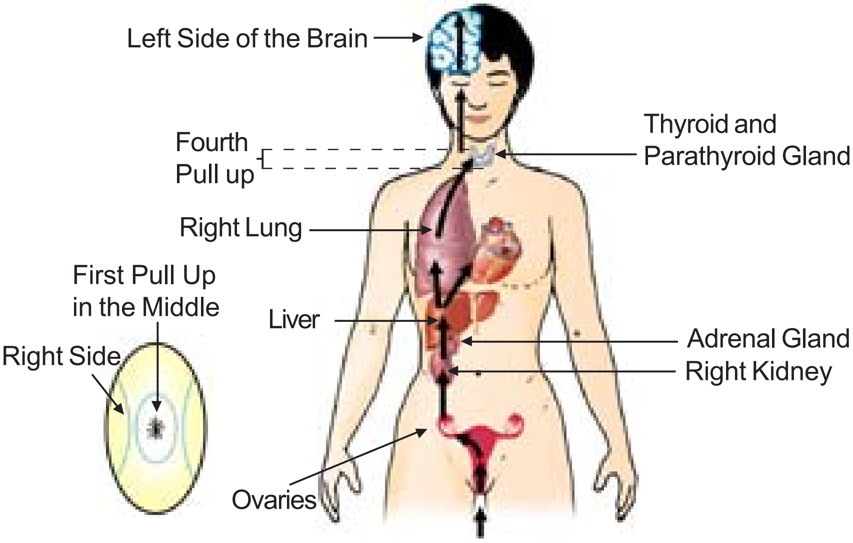
Left Side of the Brain
Thyroid and Parathyroid Gland
Fourth Pull up
Right Lung
First Pull Up in the Middle
Right Side
Liver
Adrenal Gland
Right Kidney
Ovaries

Fig. 7.10 *Right Part in the Female*

2. Liver

Close the left nostril, inhale and pull up. Use the right eye, ear and nostril; draw the energy up from the right testicle (men only), through the right ovary (women only), the right kidney and up to the liver. Exhale, relax and let the pearl return to the perineum. Practice nine times. Rest, and feel the flow of energy through the Right Thrusting Channel to the liver, under the diaphragm.

a. Combine all Three Routes

Continue to practice moving the pearl up and down through a combination of all three Thrusting Channels to under the diaphragm in the following nine steps until the energy is flowing easily: (1) Left Channel; (2) Middle Channel; (3) Right Channel; (4) Middle; (5) Left; (6) Middle; (7) Right; (8) Middle; and (9) Left. It helps to quicken the speed with each attempt.

Note: Stay relaxed during this exercise, and avoid using force. Once you are accustomed to thrusting the pearl and feel that the channels are clear, there is no need to close the nostrils for the left and right channels. Just simply pull up the perineum and anus, and use the power of the mind.

b. Clearing the Routes with Saliva
   End by preparing the saliva. First sweep your tongue across the outer surface of your upper teeth, starting on the left side at your wisdom tooth, with the tip of the tongue going along the gums. When you arrive at your upper right wisdom tooth, move the tip of your tongue down to the lower right wisdom tooth and move along the front of the lower teeth. When you arrive at the lower left wisdom tooth, start again at the upper wisdom tooth. Circle around in this manner a few times. Then similarly sweep along the inner surface of the upper and lower teeth and gums. Your mouth will begin to fill with saliva.

Gather the saliva into a ball using your tongue. Press your tongue to the roof of your mouth and swallow the saliva quickly with a gulping action. Feel the saliva burn out the impurities in the Channels and lubricate them.

Practice this stage for about one or two weeks until you can control the energy very well, then continue to the next stage. Each time you finish, move the energy in the Microcosmic Orbit and collect the energy in the cauldron. When you are ready to continue to the next stage, practice Spinal Cord Breathing at this point in your preparation.

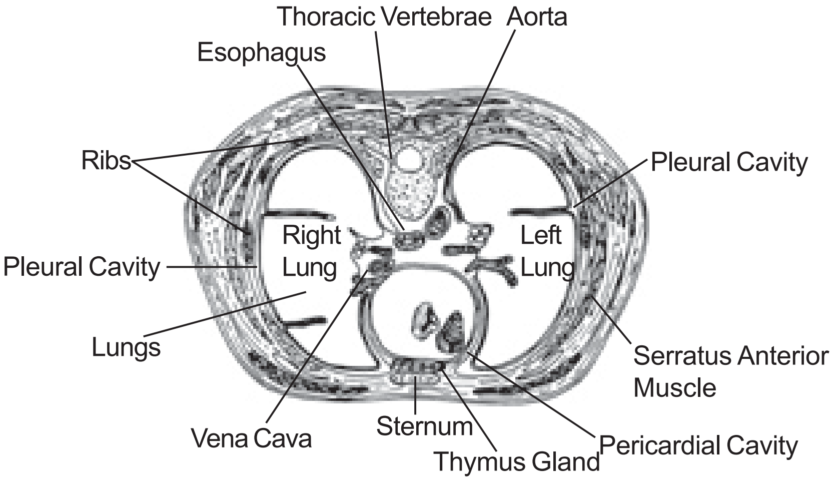
Thoracic Vertebrae
Aorta
Esophagus
Ribs
Pleural Cavity
Right Lung
Left Lung
Pleural Cavity
Lungs
Serratus Anterior Muscle
Vena Cava
Sternum
Thymus Gland
Pericardial Cavity

Fig. 7.11 *Cross Section of the Thorax*

3. Stage Two: Thrusting Channels to the Neck

A. Left Thrusting Channel to the Neck

1. Diaphragm

Using the procedure described above, bring the pearl up the Left Thrusting Channel to below the diaphragm and let it return down to the perineum.

2. Left Thyroid and Parathyroid Glands

Using the same procedure, bring the pearl up through the left ovary (women), left kidney and spleen, and continue to bring the pearl up to the left lung, and left thyroid and parathyroid glands. Hold for a while, and then exhale. Let the pearl return to the perineum. Practice nine times.

Be aware of the energy flow up to the neck through the Left Thrusting Channel.

B. Middle Thrusting Channel to the Neck

1. Diaphragm

Using the procedure described above, bring the pearl up the Middle Thrusting Channel to below the diaphragm, and let it return down to the perineum.

2. Thymus, Thyroid and Parathyroid Glands

Using the same procedure, bring the pearl up from the perineum through the prostate (men), cervix (women), small/large intestine, aorta and vena cava, stomach and pancreas. Continue to bring the pearl up to the heart, to the thymus gland under the sternum, and up to the middle of the neck, thyroid and parathyroid glands. Practice nine times. Exhale, relax and let the pearl drop down again.

Be aware of the Middle Thrusting Channel as the energy flows up to the neck.

C. Right Thrusting Channel to the Neck

1. Diaphragm

Using the procedure described above, bring the pearl up the Right Thrusting Channel to below the diaphragm and let it return down to the perineum.

2. Right Thyroid and Parathyroid Glands

Using the same procedure, bring the pearl up from the perineum to the right ovary (women), right kidney, liver and continue to bring the pearl up to the right lung and to the right thyroid and parathyroid glands. Exhale, and slowly release the energy down. Practice nine times.

Be aware of the Right Thrusting Channel as the energy flows up to the neck.

D. Continue to Practice the Three Thrusting Channels to the Neck

Practice all three Thrusting Channels to the neck, nine times each, as delineated for the diaphragm: left, middle, right, middle, left, middle, right, middle and left.

E. Clear the Routes with Saliva

If your throat feels dry, you can collect and warm the saliva in your mouth. Tighten your neck a little bit, and swallow the saliva with a guttural sound down to the stomach. Feel the saliva like a fire clearing all three channels.

Practice this stage for about one or two weeks, each time finishing by moving the pearl in the Microcosmic Orbit and collecting it in the cauldron.

F. Rest and Concentrate on All Three Channels

Rest and concentrate on all three channels, visualizing the left route as red, the middle as white, and the right as blue.

4. Stage Three: Thrusting Channels to the Crown

In stage three of opening the Thrusting Channels, the pearl is thrust through the channels up to and out of the crown.

A. Left Thrusting Channel to Crown

1. Left Eye

Follow the above procedure for thrusting through the Left Thrusting Channel to the neck, and continue thrusting to the left eye.

2. Crown

Repeat the Left Thrusting Channel to the left eye and continue to thrust the pearl up through the left hemisphere of the brain to the crown. Practice nine times. Rest, and mentally trace the Left Thrusting Channel from the perineum to the crown.

B. Middle Thrusting Channel to the Crown

1. Mideyebrow

Repeat the Middle Thrusting Channel procedure as described to the neck and continue thrusting to the mideyebrow.

2. Crown
Repeat the Middle Thrusting Channel to the mideyebrow, and continue to thrust the pearl up through the pituitary gland to the pineal gland at the crown. Practice nine times. Rest, and mentally trace the complete Middle Thrusting Channel.

C. Right Thrusting Channel to the Crown

1. Right Eye

Repeat the Right Thrusting Channel procedure as described to the neck and continue thrusting to the right eye.

2. Crown

Repeat the Right Thrusting Channel to the right eye, and continue thrusting the pearl up to the right hemisphere of the brain and then to the crown. Practice nine times. Rest and trace the route mentally from the perineum to the crown.

D. Continue to Practice the Three Thrusting Channels to the Crown

Continue to practice the three Thrusting Channels to the crown nine times as was delineated for the diaphragm and again for the neck: left, middle, right, middle, left, middle, right, middle and left.

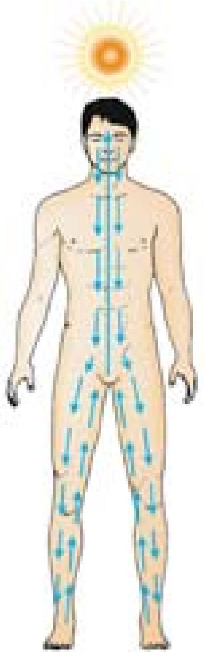

Fig. 7.12 *Three Thrusting Channels up to the Crown*

You now have succeeded in bringing the pearl through the Thrusting Channels to the crown. Your goal is to follow the sequence above using nine thrusts distributed among these three channels. Be aware of how and where the energy flows through the three Thrusting Channels, and how the three channels penetrate all the organs and glands. These channels are very powerful. They clean, detoxify and purify the organs at a much deeper level.

E. Bring the Pearl into the Microcosmic Orbit

When you have finished this stage of practice, you can bring the pearl into the Microcosmic Orbit and circulate it. Feel the Microcosmic Orbit encircling the Thrusting Channels. The top and bottom of the Thrusting Channels are joined with the Microcosmic Orbit at the crown and the perineum.

Use your mind to assist the flow of energy through the Thrusting Channels and through the Microcosmic Orbit when the pearl reaches the crown point or the perineum point.

Your Microcosmic Orbit practice will change dramatically after you learn the Fusion of the Five Elements meditations. When you first practice the Microcosmic Orbit, the energy you move through your channels is raw energy. By contrast, when you reach the level of Fusion practice, the energy is more refined and condensed, and it is therefore much more powerful. The Fusion practice includes the Microcosmic Orbit, so it might be said that the Fusion meditation is simply a more advanced way of practicing the Microcosmic Orbit meditation.

Practice this stage for two or three weeks until feel you have gained control of the energy.

5. Extending the Three Thrusting Channels above the Crown

Once we have purified and transformed the negative energy of the organs back into usable life-force energy (Fusion of the Five Elements), distilled this energy along with the virtue Chi of the organs into the pearl (Cosmic Fusion), and opened all our reservoirs or special channels, giving us a larger energy capacity, we can increase our virtue energies further by connecting with the external sources of Five Elements Chi and directing the external Chi toward its associated organs. Moving the pearl out of the physical body makes it easier for the pearl to absorb these energies. It becomes like an antenna that receives the external Chi and conducts it into the physical body. In this way, we again nourish our health in this life as well as strengthening the Energy Body and spirit in preparation for the next.

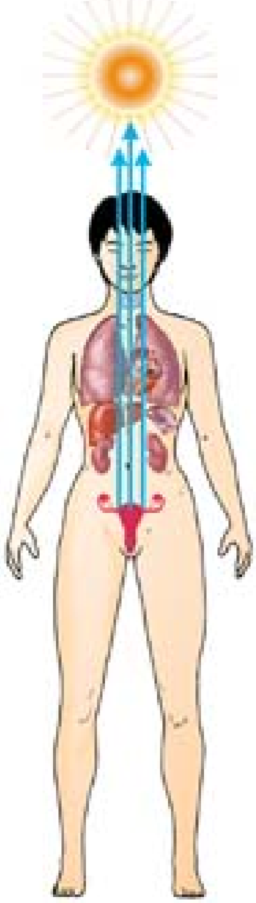

Fig. 7.13 *Three Thrusting Channels above the Crown*

After you gain full control of the energy, you will be able to send the pearl out of the body at the crown. Begin to practice quickly forming a fresh pearl in the practice of formimg the pearl by fuse in and out and back to the inside to form a pearl. Follow with the Creation Cycle. Then, thrust the pearl up through the Thrusting Channels to quickly clear out these routes. You are now ready to extend the channels above the crown.

Bring the energy down to the perineum, and start with the Left Thrusting Channel up to the left crown. Push the pearl out at the crown of the head to about three to six inches (seven and a half to fifteen centimeters) above the crown. Similarly, push the pearl out of the crown through the Middle Thrusting Channel, and then the Right Thrusting Channel.

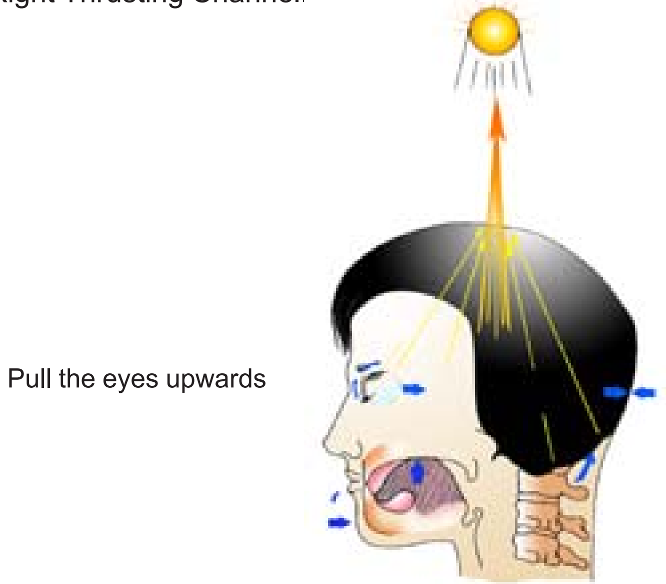

Fig. 7.14 *Use Physical and Mental Power to push the Pearl of Energy up through the Crown.*

Each time you push the pearl out at the crown, be aware of an energy ball hovering about six to ten inches (fifteen to twenty-five centimeters) over the crown. Draw and absorb the Heavenly Energy to the pearl. Then slowly let the energy melt and flow down to the crown, and then into the body, all the way down through the three Thrusting Channels to the perineum.

6. Stage Four: Leg Routes and Thrusting Channels into the Ground

In the final stage of opening the Thrusting Channels, the energy will go all the way down to the feet and into the ground, and then up through the channels with each thrust. After you have formed another new pearl by practicing Fusion of the Five Elements and the Creation Cycle of Cosmic Fusion, bring the energy down to the perineum.

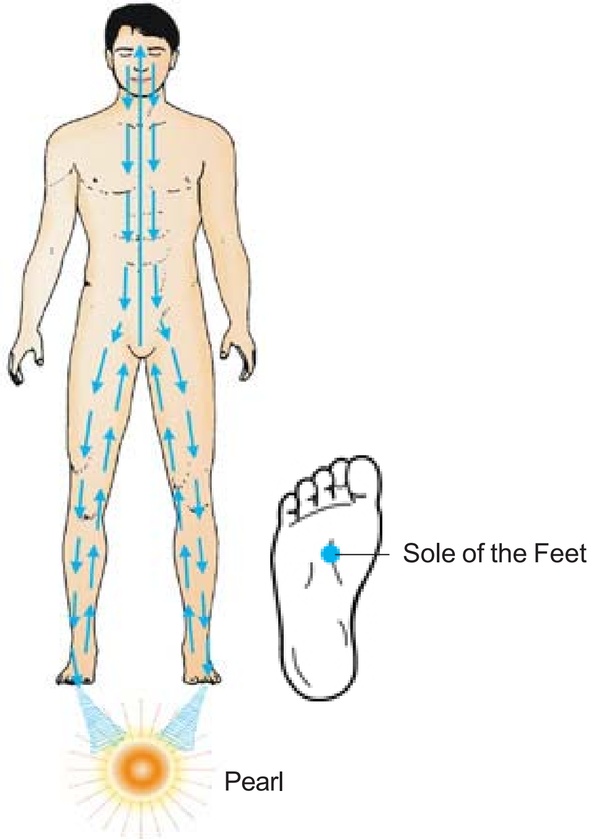

Fig. 7.15 *Three Thrusting Channels down to the Ground.*

A. Bring the Energy Down to the Soles of the Feet

1. Left Thrusting Channel

Split the pearl in two, and bring the energy from the perineum down the back of both legs to the backs of the knees, and then down to the heels and soles of the feet. Stop for a while, and absorb the Earth Energy through the soles. Then bring the energy up the front of the legs through the big toes, knees, front of the thighs, all the way back to the perineum. Recombine the pearl at the perineum. Then run the energy through the Left Thrusting Channel by slightly tightening the left side of the anus and looking up with the left eye. Bring the pearl all the way up through the left hemisphere of the brain and left crown, and out to three inches (seven and a half centimeters) above the crown. Relax, and bring the pearl down to the perineum.

2. Middle Thrusting Channel

From the perineum, split the pearl, and bring the pearl down the backs of both legs to the soles of both feet. Absorb the Earth Energy through the soles, then move the energy up to the toes, to the knees, returning to the perineum. Recombine the pearl at the perineum. Now bring the pearl all the way up the Middle Thrusting Channel, squeezing the middle of the anus and looking up with both eyes. Bring the pearl all the way up through all the organs and glands to the crown, and out to three inches (seven and a half centimeters) above the crown. Then bring it down to the perineum again.

3. Right Thrusting Channel

From the perineum, split the pearl, and bring it down the backs of both legs. Absorb the Earth Energy through the soles of the feet. Then bring the pearl up again to the perineum. Recombine the pearl at the perineum. Use the right anus muscle, look up with the right eye, and bring the pearl all the way up the Right Thrusting Channel through the right hemisphere of the brain and out of the crown three inches (seven and a half centimeters.) Bring it down to the perineum.

Continue bringing the pearl to the Thrusting Channels in the following sequence: left, middle, right, middle, left, middle, right, middle, and left, finishing each step by looping the pearl through the leg route. End on the left side of the perineum. Swallow the saliva, as described before.

B. Bring the Energy into the Ground

Once you have practiced bringing the pearl down to the soles of the feet for a while, add a new step. Begin by shooting the energy from the perineum down the backs of the legs and through the soles of the feet, from six inches to one foot into the ground. Feel a connection or rooting to the earth, then draw the energy up to the soles of the feet, to the toes, up the front of the legs, and into the Thrusting Channels.

With this step you have completed opening all the Thrusting Channels. Practice for two or three weeks until you develop full control of the energy traveling through these channels.

7. Men and Women Now Practice Differently

At this stage all the Thrusting Channels and Leg Routes are open, and men and women start to practice differently. Both quickly practice Fusion I, and the Creation Cycle and Thrusting Channels of Fusion II, and bring the pearl to the perineum.

A. Men start by bringing the Pearl to the Crown.

Men bring the pearl from the perineum up to and out of the crown to about six inches (fifteen centimeters.) Feel your pearl hook up with the higher energy source, Heavenly Energy. Absorb this energy into the pearl. Then bring the energy back into the body. As it enters the crown, let the pearl split in three. Feel the energy penetrate through three holes at the top of the head. Feel it flowing back down to the perineum through the three Thrusting Channels, penetrating all the organs and glands.

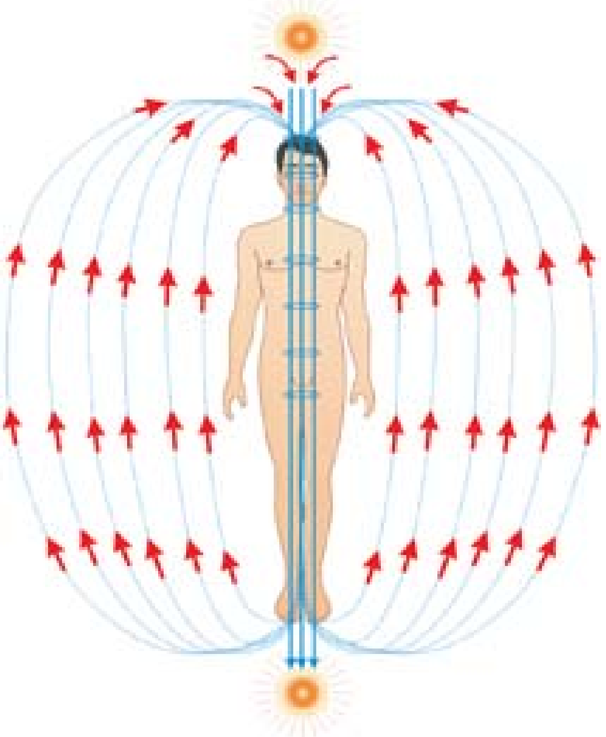

Fig. 7.16 *Bringing the Pearl to the Crown*

From the perineum men will push and guide the energy down the back of both legs to the soles of the feet, and then six to twelve inches (fifteen to thirty centimeters) into the ground. Feel a connection or rooting to the earth, then draw the Earth Energy up to the soles of the feet. Bring the energy to the toes. Start to spread the energy out from the ten toes, shooting the energy up to and into the crown, like a waterfall with a reverse flow. Draw the pearl back down to the perineum, and then repeat the process.

Practice nine times. Feel the flow of energy like a magnetic field surrounding your body.

B. Women start by bringing the Pearl to the Feet.

Women bring the pearl down to the perineum, divide it, and shoot it down the backs of both legs, through the soles of the feet, six to twelve inches (fifteen to thirty centimeters) into the ground. Feel a connection or rooting to the earth, and absorb the Earth Energy into the pearl. Then draw the Earth Energy up to the soles of the feet. Bring the energy all the way up the front of the legs to the knees, and back to the perineum. As the energy travels up the legs, feel it penetrate the bones.

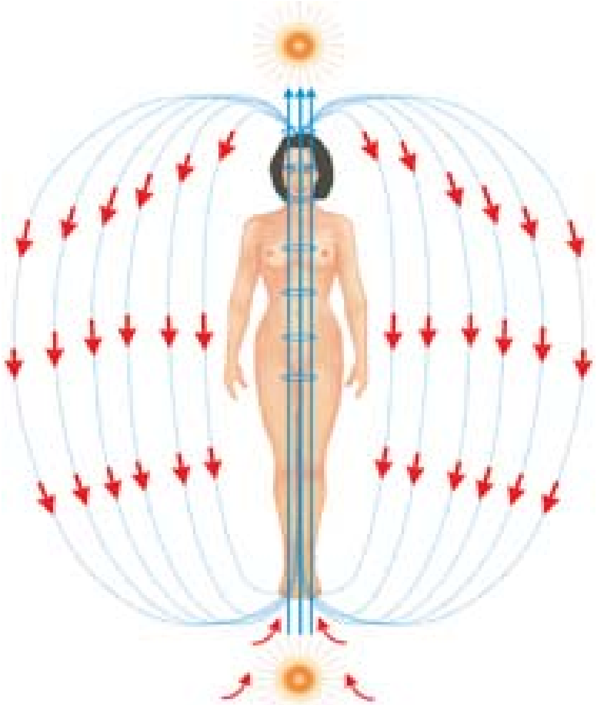

Fig. 7.17 *Women bring the Pearl to the Feet.*

Spread the energy into the three Thrusting Channels, and thrust it as though it were on a three lane highway all the way up to the crown. Feel the energy spread from the crown like a spring or water fountain pouring out of the body, entering into the soles of the feet, where it again joins together with energy from the earth. While the energy is outside and circling the body, it is also gathering pure, radiant Cosmic Energy. Repeat this meditation nine times.

8. Finishing the Meditation

Both men and women should finish by collecting the energy at the cauldron and practicing Chi Massage. This completes the practice of the Thrusting Channels. Continue to practice for a few weeks until you gain control of it.

Additional Exercise for Opening the Thrusting Channels

Alternate Nostril Breathing

The alternate nostril breathing helps to generate more power in moving the pearl up and down the Thrusting Channels. By closing one side of the nose, it is easier to move the pearl through the specific left or right channels.

Hold the right nostril closed with the finger tips.

Inhale up through the left nostril, and feel the pearl move up from the perineum to the crown.

Exhale out through the left nostril and bring the pearl back down to the perineum.

Inhale through the left nostril and bring the pearl to the Middle of the perineum.

Close both nostrils and suck the pearl up through the middle Thrusting Channel to the crown. This suction is created by an inhaling motion without actually drawing any air into the lungs.

Hold the right arm down towards the ground and exhale any cloudy or gray energy out of the Thrusting Channel.

The same procedure is repeated in the Right Thrusting Channel.

Hold the left nostril closed with the finger tips.

Inhale up through the right nostril, and feel the pearl move up from the perineum to the crown.

Exhale out through the right nostril and bring the pearl back down to the perineum.

Inhale through the right nostril and bring the pearl to the middle of the perineum.

Close both nostrils and suck the pearl up through the Middle Thrusting Channel to the crown.

Hold the left arm out and down to the ground, exhaling any cloudy gray energy down to the Earth.

Repeat 3 – 9 times or until you feel the Thrusting Channels clean and clear with positive energy.

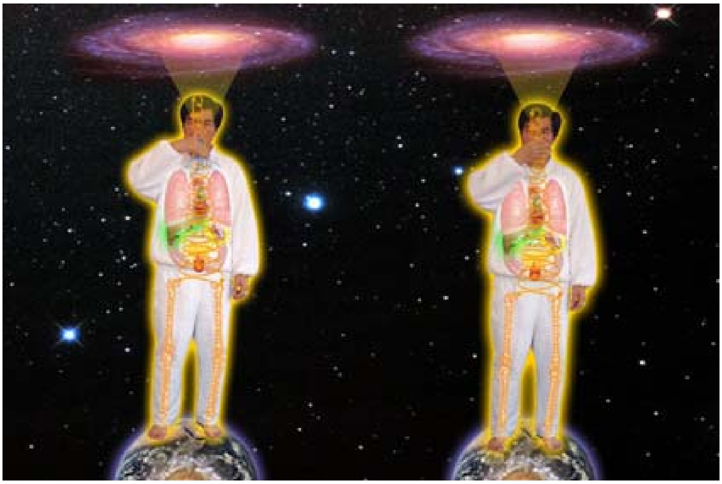

Fig. 7.18 *Alternate Nostril Breathing*

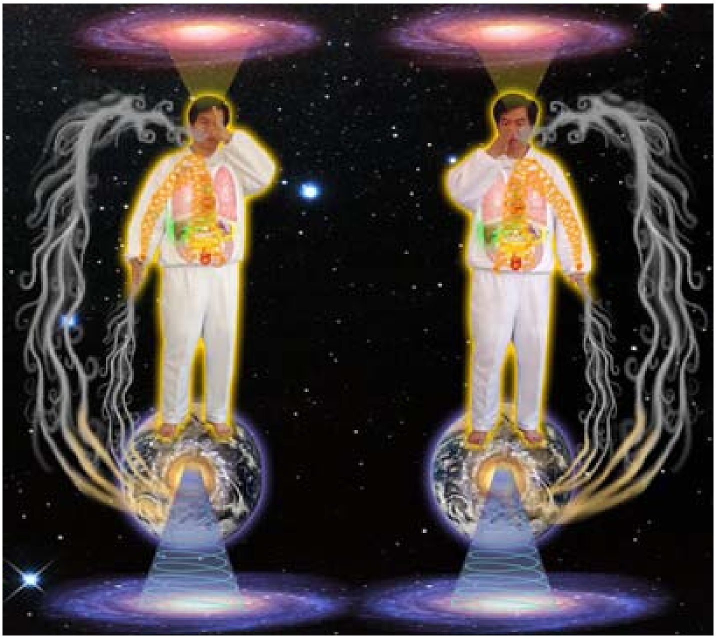

Fig. 7.19 *Exhale cloudy energy out of the Thrusting Channel.*

Summary
Opening the Thrusting Channels

Begin the Fusion practice by Opening the Three Fires, activate the compassion fire forming the pakuas by chanting the eight forces, forming the collection points, fusing the colors, and forming the pearl. Review the explanation discussed previously. Once you feel the pearl is strong and full of energy, you do not have to go through all these stages all the time. Just think and let it happen.

The Thrusting Channels are Opened in Four Stages.

(1) to the diaphragm; (2) to the neck; (3) to and out of the crown; (4) down to the feet, into the ground; and up to the crown.

1. The Left Thrusting Channel penetrates the left testicle (in men), left side of the perineum and anus, left ovary (in women), left kidney, spleen, heart, left lung, left parathyroid and thyroid glands, left ear, left eye, and left hemisphere of the brain.
2. The Middle Thrusting Channel runs through the middle of the scrotum (in men), perineum, middle of the anus, cervix (in women), prostate (in men), aorta, vena cava, pancreas, stomach, heart, thymus gland, throat, tongue, pituitary, hypothalamus, and pineal glands, and crown.
3. The Right Thrusting Channel goes through the right testicle (in men), right side of the perinium and anus, right ovary (in women), right kidney, liver, right lung, right parathyroid and thyroid glands, right ear, right eye, and right hemisphere of the brain.

Note: Avoid overheating the heart, liver, or head. Be very gentle with this procedure. If the energy reaches the heart and head too quickly the heart will become congested and you may experience pressure. If the heart overheats with energy, avoid thrusting above the diaphragm, and practice the Heart Sound (***Hawwww***).

Strengthen your anus muscle, the foundation, connected to the perineum and everything in the body above it. Saliva purifies the mouth, and nourishes immortality. When resting, create and gather saliva. Make it thick and swallow down forcefully.

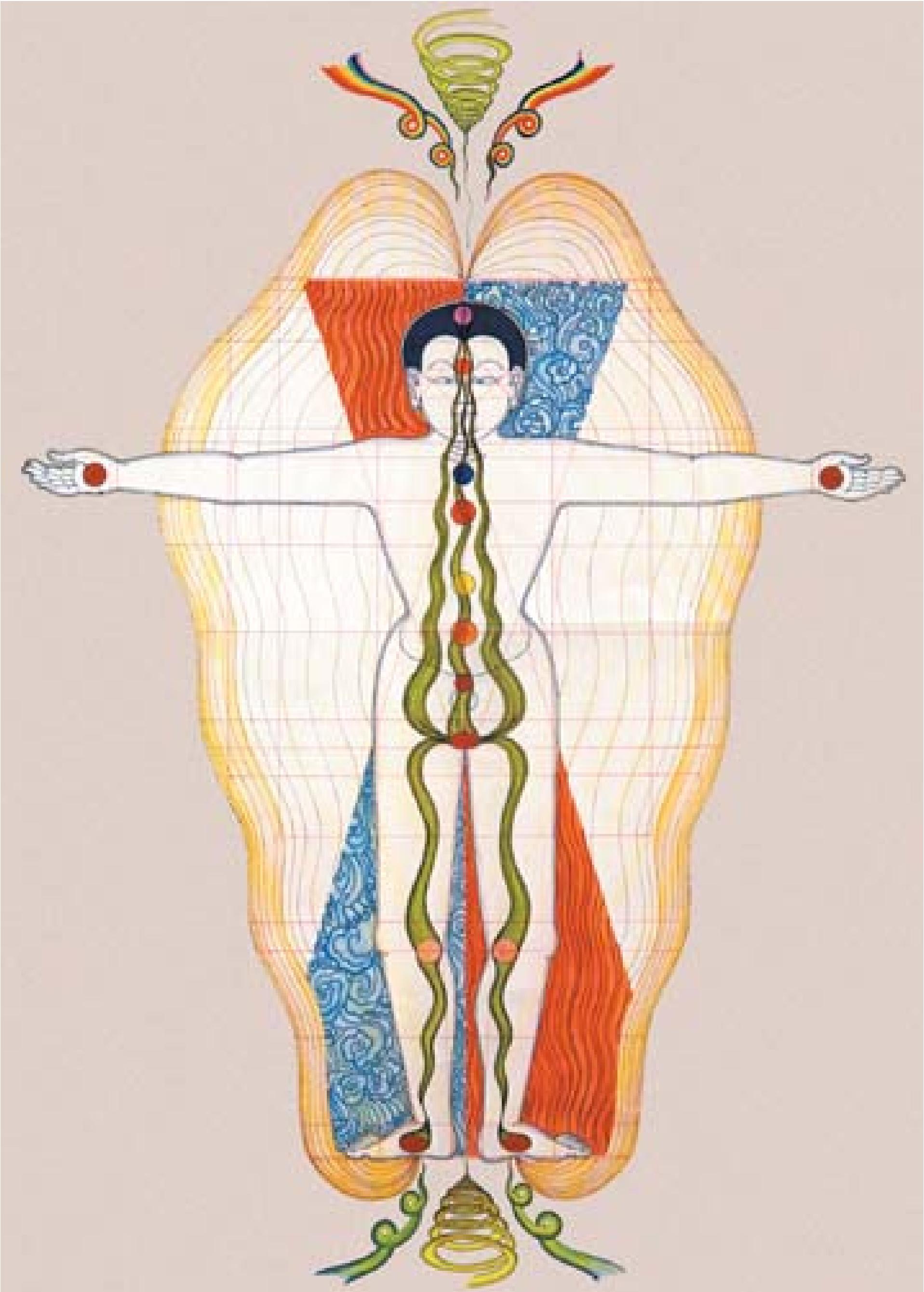

Fig. 7.20 *Thrusting Channel*

Chapter 8

Belt Channels

Theory of the Belt Channel

The Belt Channel is a psychic channel of power and self-defense, encircling the body with energy. It protects the body and fences the energy within the Governor, Functional and Thrusting Channels, while warding off negative energy from the outside world. Starting at the navel each Chi center of the body is crossed and encircled by one level of the continuing Belt Channel.

A cross-section at the navel level would show the Belt Channel connecting to the other channels you have learned to this point. At the front of the body, the Belt Channel connects with the Functional Channel. As it moves to the left side it connects with the Left Thrusting Channel. At the Door of Life in the back, it connects with the Governor Channel. To the right it connects with the Right Thrusting Channel. Returning to the front, the circle is completed.

When you learn to form a pearl practicing Fusion of the Five Elements, and you draw the energy from the front, back and side pakuas at the level of the navel, you are laying the groundwork for the Belt Channels.

Procedure of Forming the Belt Channel

The Belt Channel begins at the navel and spirals upward in a counterclockwise direction through solar plexus, heart, throat, mideyebrow, and crown. Then, reverse the energy flow and circle the pearl clockwise as you move down through the above points to the navel. From the navel continue down to the sexual center, perineum, knees (around both knees), and soles of the feet (slightly into the ground). Finally, reverse the direction and work up through each point back to the navel.

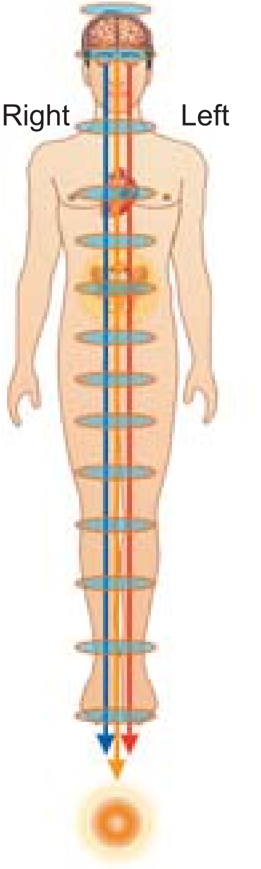

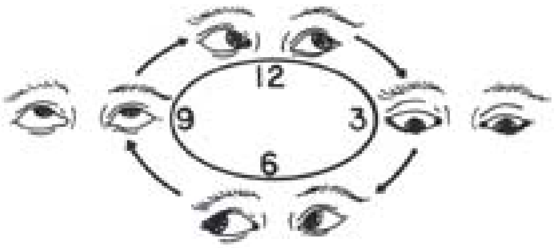

Counterclockwise

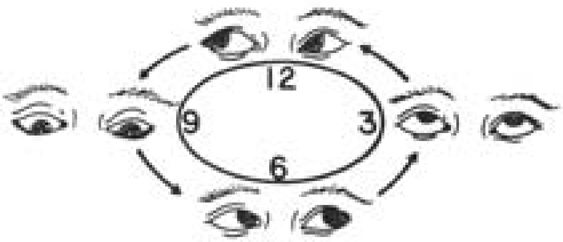

Clockwise

Fig. 8.1 *Use the eyes to help spiral.*

Note: To circle to the left is to circle counterclockwise; circle counterclockwise as you move the pearl up. To circle to the right is clockwise; circle clockwise as you move the pearl down.

The pearl is circulated in each Belt Channel nine times. After the ninth time of circling each point, you will cross each channel internally with the pearl by moving the pearl from front to back, and from left side to right side. Each time you will end with the pearl at the front, and then move on to the next level Belt Channel.

Practice of Forming the Belt Channel

Ascending from the Pakuas

1. Sit up, back straight, shoulders relaxed, feet touching the floor. Hold the hands together, and touch the tongue to the palate.
2. Practice the Inner Smile, Fusion of the Five Elements, chanting, the collection points, fusing the energy into a pearl and the Creation Cycle of Cosmic Fusion. Continue with your practice of Cosmic Fusion by sending the energy through the three full Thrusting Channels (all the way down to the feet and up to the head.)
3. You are ready to begin forming the Belt Channel. First, form the four pakuas and a pearl. Gather the pearl at the cauldron, and bring it to the navel.

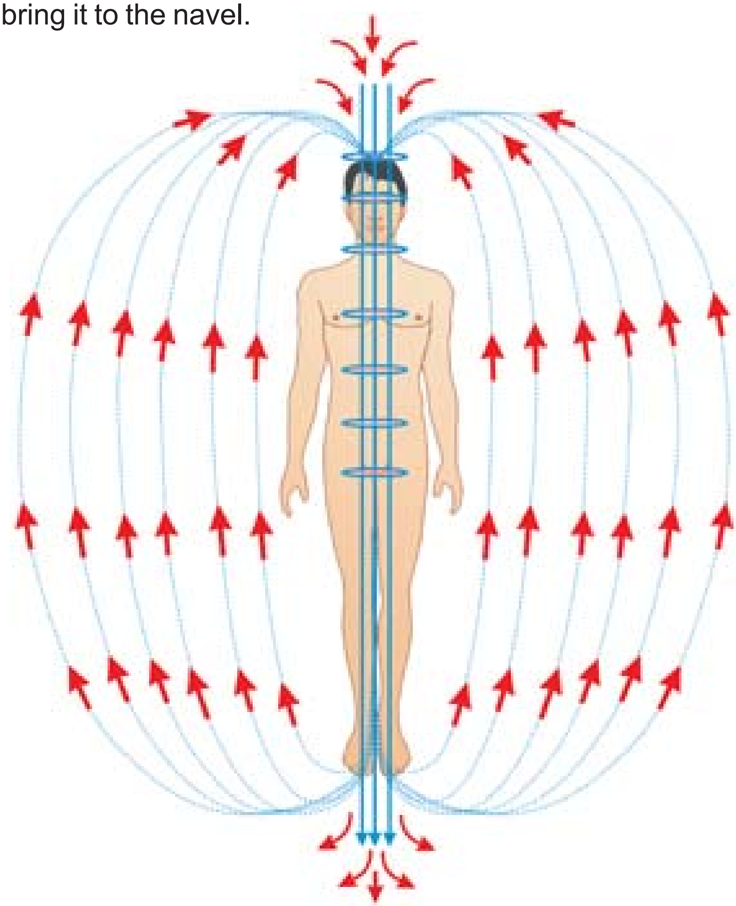

Fig. 8.2 *Moving the Thrusting Channels out of the Body.*

4. Join the four pakuas by circulating the pearl from the ***Navel*** toward the left side (counterclockwise), to the Door of Life, to the right side, and back to the navel.

   In the beginning you can use your hand to assist you in moving the pearl, although your goal is to use your senses to control its movement. Cover the navel with the right hand. Cover the left side pakua with the left hand. Then, move the left hand to cover the Door of Life, and move your right hand to cover the right side pakua. Finally, return the right hand to the navel.

   Circle the pearl nine times counterclockwise. As you are ascending you will circle it nine times counterclockwise from each point in a similar fashion.

5. When you finish circling the pakuas for nine rounds, be aware of the navel. Be aware of the channel of the navel. Form a cross by bringing the pearl from the navel to the cauldron, and all the way back to the Door of Life. Then, bring the pearl back to the cauldron, and use it to connect the side pakuas from the left cross to the right. Return the pearl to the navel.

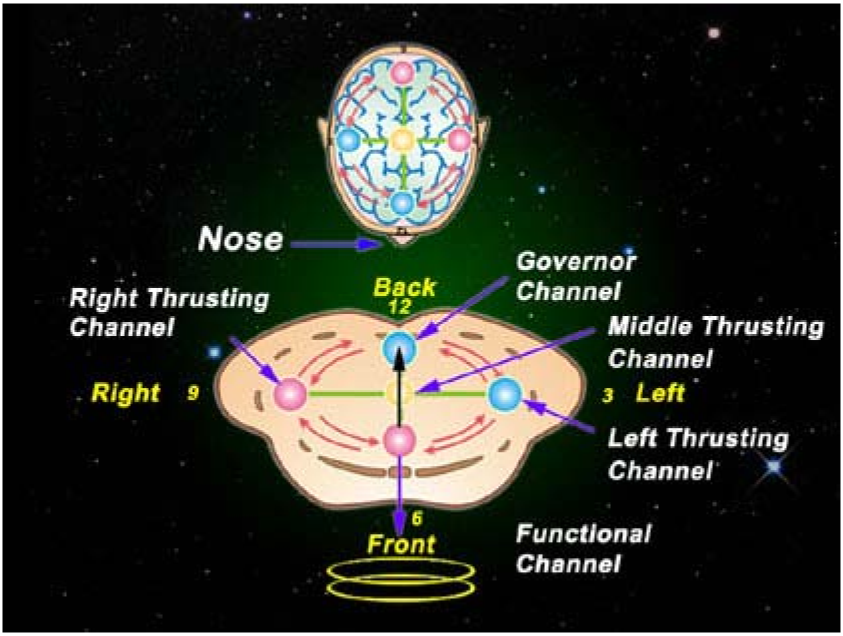
Nose
Right Thrusting Channel
Back
12
Governor Channel
Middle Thrusting Channel
Right
9
3
Left
Left Thrusting Channel
6
Front
Functional Channel

Fig. 8.3 *Cosmic Belt Channels connecting to all the other Channels.*

6. Move the pearl up from the navel to the left side of the rib cage to a position level with the ***Solar Plexus***. Begin to circle the pearl at the solar plexus level. Circle the pearl toward the left through T-11 at the back, to the right, and return to the solar plexus. Circle nine times, ending with the pearl at the solar plexus. Take a moment to perceive it. Then create an internal cross with the pearl, moving it from front to back, and then from left to right.
7. Bring the pearl from the solar plexus to the left side rib cage, this time at the level of the ***Heart***. Circle nine times toward the left to the middle of the shoulder blades, to the right, and back to the heart center. Circle in this manner for nine rounds. Using the pearl, form an internal cross connecting all four sides. Feel the connection of the channel.
8. From the heart center, bring the pearl up to the left side of ***Throat Center*** on the neck. Circle the pearl toward the left and back to C-7, to the right side, and to the middle of the throat center nine times. Use the pearl to form an internal cross connecting all four sides. Practice and master this step before continuing up to the crown.

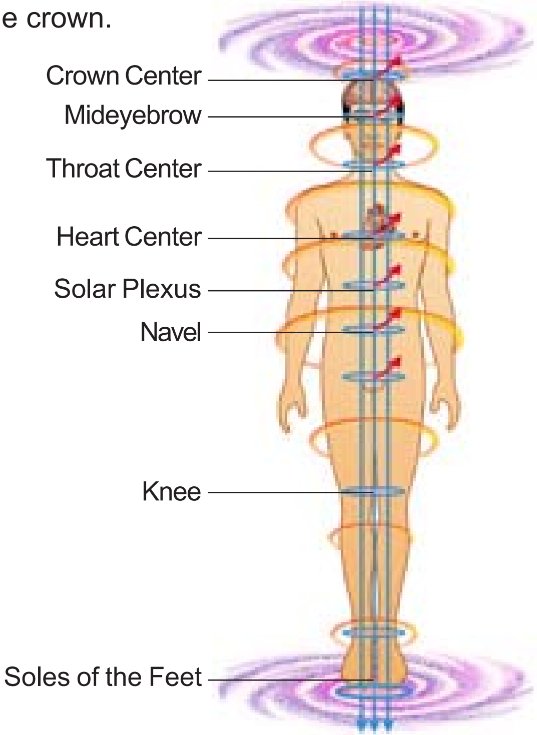

Fig. 8.4 *Bring the pearl up to the upper tip of the left ear.*

9. Bring the pearl up to the upper tip of the left ear at the level of the ***Mideyebrow Center***. Circle the pearl nine times from the left ear back to the small brain, to the right (upper tip of the ear), returning to the mideyebrow. Feel a band of energy flowing around the mideyebrow, ears, and back of the head. Using the pearl, form an internal cross connecting all four sides. The feeling is one of wearing a crown with a cross at its center connecting it.
10. From the mideyebrow bring the pearl up to the left side of the ***Crown***, and circle the energy nine times in a counterclockwise direction at the crown. End with the pearl at the front. Using the pearl, form a cross connecting the front, back, and both sides of this Belt Channel.
11. Allow the pearl to go out above the crown, and circle the energy nine times toward the left forming a ***Halo of Energy***. Use the pearl to cross the halo, front to back, and left to right. Collect the energy from above the head.

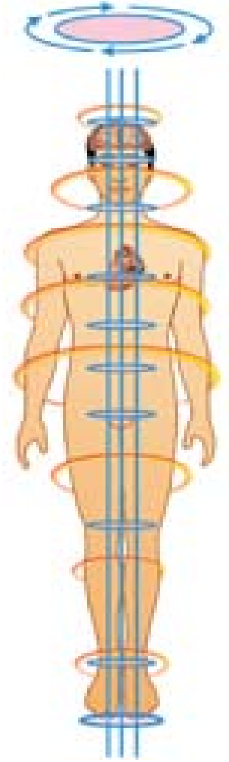

Fig. 8.5 *Spiral out clockwise above the crown.*

12. Now reverse the direction of energy flow to a clockwise direction by circling from the front toward the right, to the back, and to the left. Circle the ***Halo*** nine times, ending with the pearl at the front. Then cross the front to back, and right to left sides. Finish by bringing the pearl to the front of the halo.

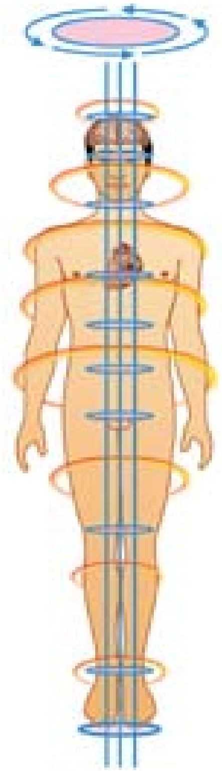

Fig. 8.6 *Reverse the direction clockwise.*

13. Next bring the pearl back into the right side of the ***Crown***. Circle toward the right nine times, bringing the pearl to the back, to the left, and ending at the front of the crown each time. Then use the pearl to cross the front, back, and two sides. Return the pearl to the front of the crown.
14. Move the pearl down to the upper tip of the right ear. Circle from the right ear to the back of the head, to the top of the left ear, and to the ***Mideyebrow*** nine times. End at the mideyebrow. Use the pearl to cross the front, back, and two sides at this level, and return it to the front to the mideyebrow position.

15. Bring the pearl down to the right side of the neck at the level of the ***Throat Center***, and circle right to *C-7* at the back of the neck, to the left, and to the front of the throat center. Circle nine times, ending at the front. Cross the front, back, and two sides, also ending by bringing the pearl to the front.
16. Bring the pearl down to the right side rib cage at the ***Heart Center*** level. Circle the pearl toward the right, to the back,and so forth clockwise for nine times, ending at the front of the heart center. Then, use the pearl to cross the front, back, and two sides. Return the pearl to the front again.
17. Move the pearl to the ***Solar Plexus***. Circle it nine times to T-11 at the back, and return to the solar plexus. Cross the front, back, and sides. Return the pearl to the front.
18. Bring the pearl down to the ***Navel***. Circle it back to the Door of Life, and back to the navel nine times. From the navel bring the pearl to the cauldron, back toward the Door of Life, and return it to the cauldron. Use the pearl to cross the right and left sides. Return the pearl to the navel.

Descending From the Pakuas to the Earth.

19. Once you have controlled the pearl to this stage of the meditation, you are ready to move the pearl down to the ***Sexual Center***. Women bring the pearl down to the right side of the ovary center, located three inches below the navel. Men bring the pearl from the navel down to the right side of the sperm palace, one and a half inches below the navel. Circle clockwise back to the sacrum, to the left hip, and to the front of the ovary center/sperm palace nine times. Use the pearl to make a cross. Return the pearl to the front of the sexual center.
20. Bring the pearl down to the right groin. Circle the pearl clockwise around the ***Perineum Point and Lower Hips*** nine times. Use the pearl to form a cross by connecting the front, back, and sides. Return the pearl to the front.
21. Move the energy down and circle toward the back of ***Both Knees***, around and to the front, midway between both knees, moving the pearl clockwise nine times. Use the pearl to form a cross by moving the pearl to the midpoint between the knees and connecting the front, back, and sides from this point. End with the pearl midway between the fronts of both knees.

22. Direct the energy down to the feet, just below the ankles. Circle the ***Soles of the Feet*** nine times moving the pearl clockwise to the heel of the right foot, to the heel of the left foot, to the outer side of the left foot, to the front, midway between both feet. Form a cross. End with the pearl midway between both feet.
23. Move the energy from the front of the feet down to a point ***Beneath the Earth***, and circle toward the right (clockwise) nine times. Use the pearl to form a cross ***Beneath the Earth***. Keep on spiraling more energy to enhance the pearl. Return the pearl to the front of the circle.

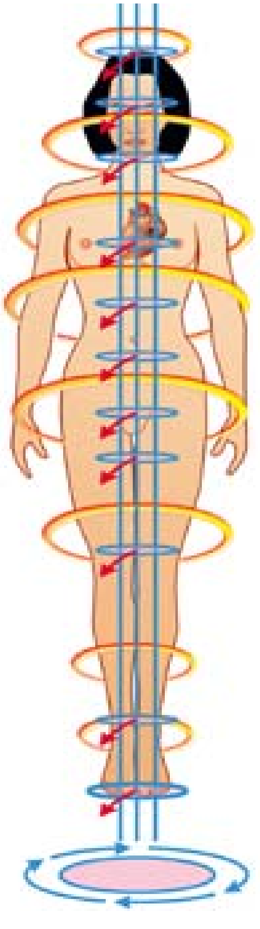

Fig. 8.7 *Bring the Pearl down to the ground, spiral clockwise.*

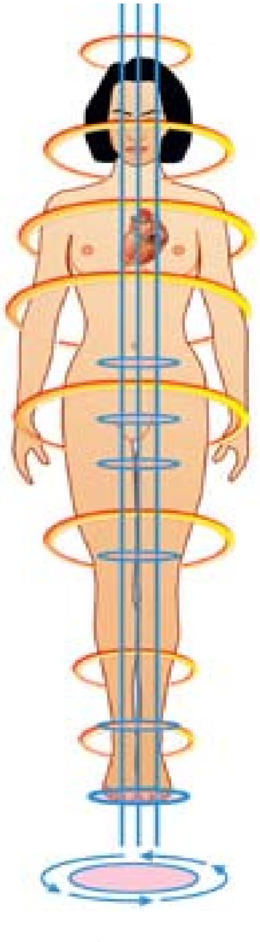

Fig. 8.8 *Spiral reverse counterclockwise moving up.*

Ascending From the Earth Back to the Pakuas

24. Reverse the circle beneath the earth, moving the pearl counterclockwise again nine times. Use the pearl to form a cross, connecting the front, back, and sides.
25. Move the pearl to the outer side of the left foot. Circle the ***Soles of Both Feet*** nine times moving toward the left heel, to the right heel, to the right foot, and to the front (midway) of both feet. Use the pearl to form a cross. End with the pearl at the midway point in the front of both feet.
26. Move up to the left knee. Circle the pearl counterclockwise *at the* ***Knees*** nine times. Use the pearl to form a cross. Return the pearl to the midway point in the front of both knees.
27. From the front of the knees, move up to the left groin area. Circle counterclockwise at the ***Perineum*** nine times. Use the pearl to form a cross. Return the pearl to the front of the perineum.
28. Direct the energy to the left hip, and circle counterclockwise at the ***Ovary/Sperm Palace*** nine times. Use the pearl to form a cross by connecting the front, back, and sides. End with the pearl at the front of the ovary/sperm palace.
29. Return the pearl to the ***navel*** and to the ***Cauldron***.

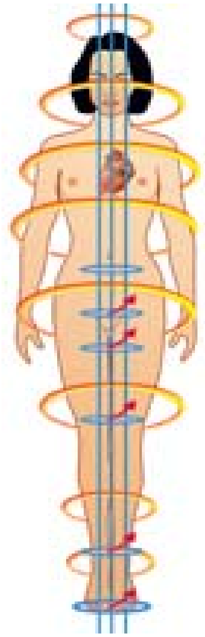

Fig. 8.9 *Bring the Pearl up to the navel.*

Men and Women Practice Differently

***Men**:* Shoot the Pearl out of the Crown, and Spiral the Pearl down from the Crown.

a. Bring the pearl up to the crown.
b. Allow the pearl to go out three to six inches above the head. Circle nine times counterclockwise. Collect the heavenly energy and blend it in the pearl. Then circle it nine times clockwise.
c. Bring the pearl back into the crown point.
d. Spiral it clockwise through each of the Belt Channels all the way down to the soles of the feet.
e. Move the pearl into the earth six to twelve inches to collect the earth energy, and circle the energy nine times clockwise. Reverse the direction spiraling nine times counterclockwise.
f. Return the pearl to the front of the crown, and then back down to the navel and cauldron in the same manner.

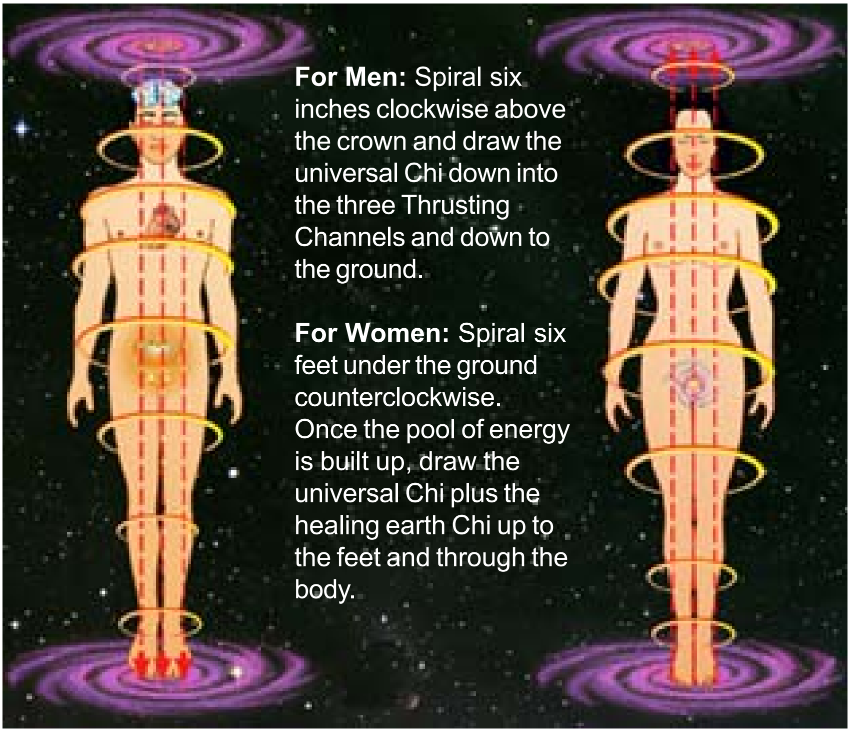

For Men: Spiral six inches clockwise above the crown and draw the universal Chi down into the three Thrusting Channels and down to the ground.

For Women: Spiral six feet under the ground counterclockwise. Once the pool of energy is built up, draw the universal Chi plus the healing earth Chi up to the feet and through the body.

***Fig. 8.10** Men and Women Practice Differently*

***Women**:* Shoot the Pearl into the Earth, and Spiral the Pearl up from the Earth

a. Bring the pearl to the soles of the feet.
b. Move the pearl into the ground six to twelve inches to collect the earthly energy. Circle the energy nine times clockwise. Then nine times counterclockwise.
c. Move the pearl up through the left side of each Belt Channel, spiraling counterclockwise to the crown.
d. Shoot the pearl all the way out at the crown to collect the heavenly energy. Circle the energy counterclockwise nine times. Reverse the direction, and spiral the pearl nine times clockwise.
e. Bring the pearl back down into the right side of the crown. Continue to circle the energy clockwise down through the Belt Channel to the soles of the feet. Return the pearl to the navel and cauldron in the same manner.

Forming a Vehicle

Continue feeling all the pakuas spiraling around the energy center and around the body.

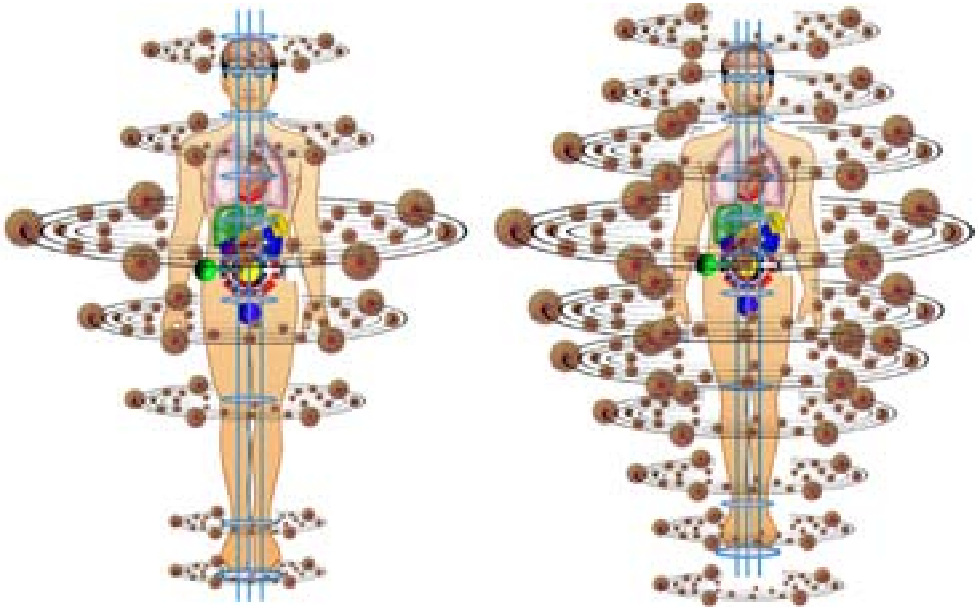

Fig. 8.11 *Spiraling the Pearl and Pakuas faster and faster.*

Let the pakuas spiral faster and start from 10,000 miles per hour up to 60,000 miles per hour and feel the spine getting bigger and bigger.

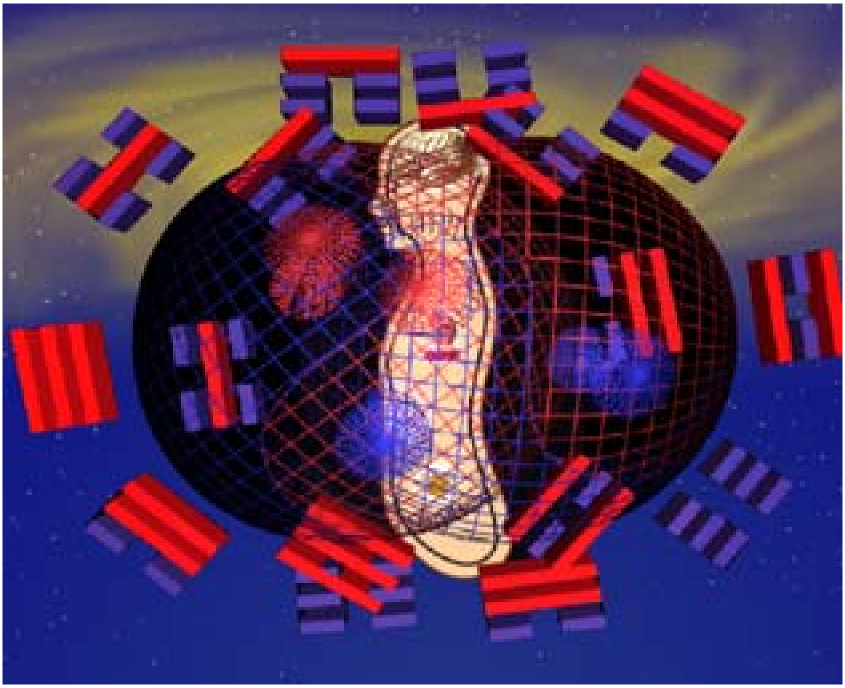

Fig. 8.12 *All the Pakuas become a bigger and bigger Pakua spiraling around whole the body.*

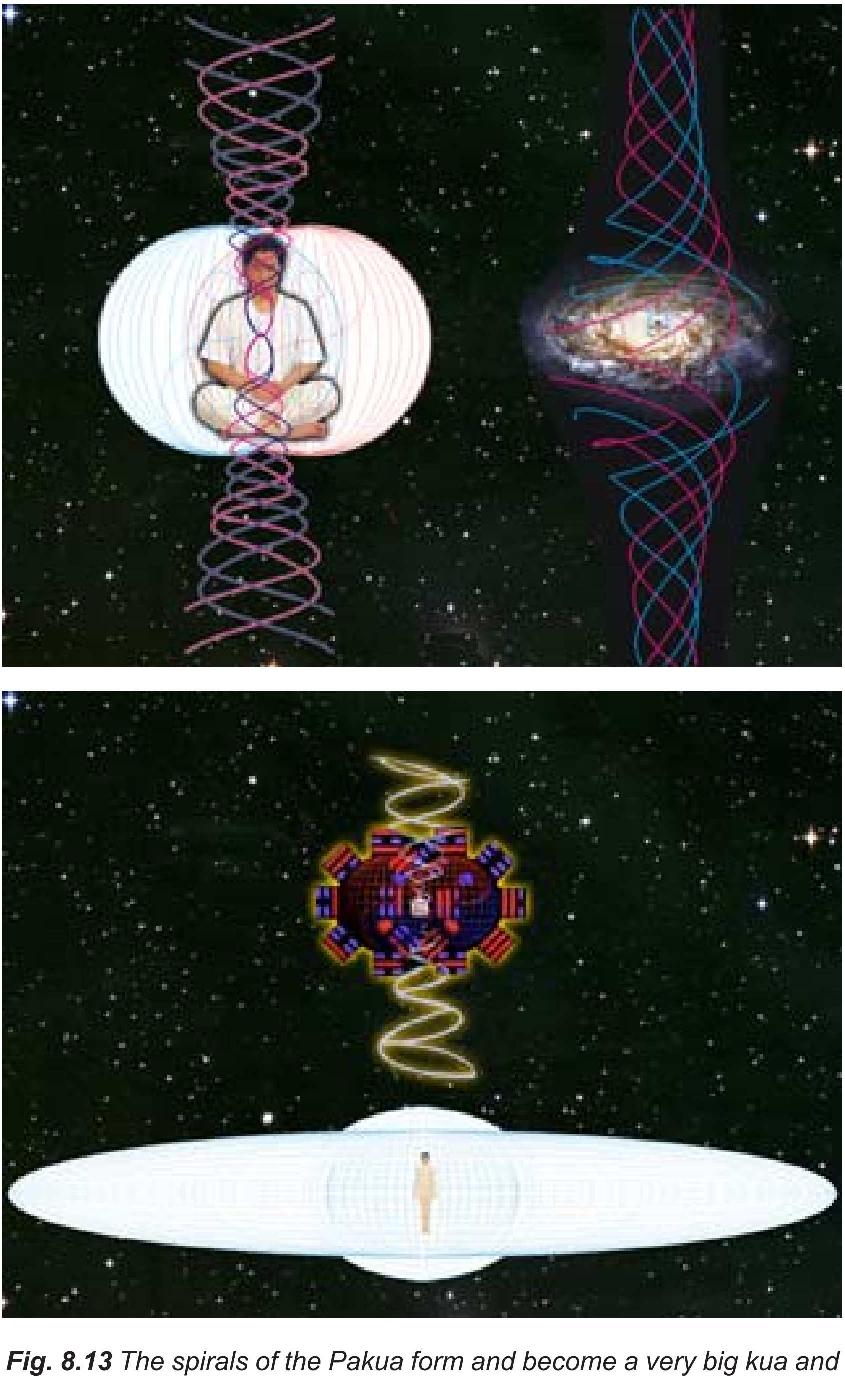

Fig. 8.13 *The spirals of the Pakua form and become a very big kua and form a space ship up to 50 meters in diameter or bigger.*

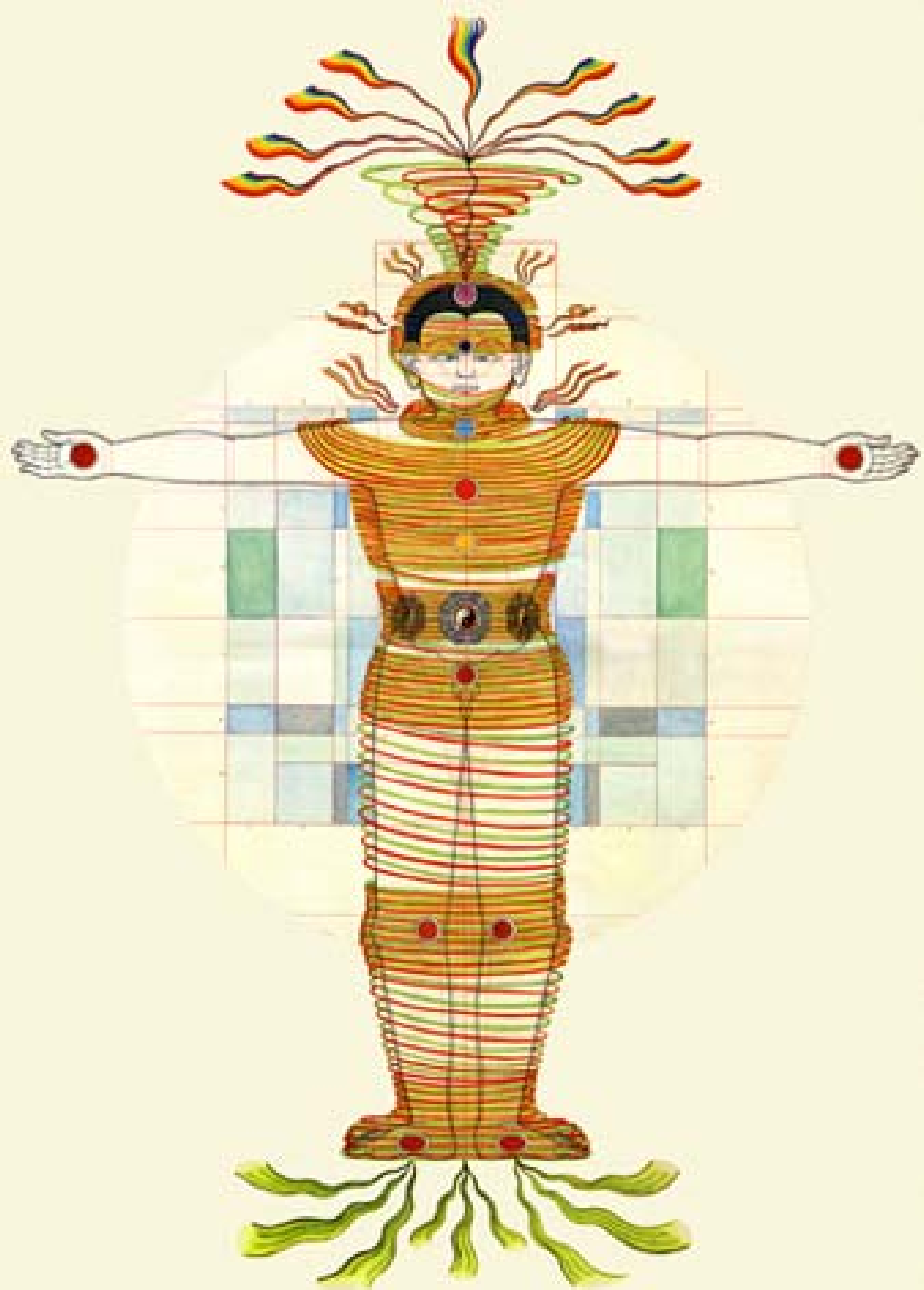

Fig. 8.14 *Belt Channels*

Summary

Procedure of Forming the Belt Channels

Ascending from the Pakuas

1. Practice the Inner Smile from the compassion fire, chanting from pakua, forming the collection points, fusing the energy into a pearl, and the Creation Cycle of Cosmic Fusion.
2. Send energy through the three Thrusting Channels all the way down to the feet and up to the head.
3. Form the four pakuas and a pearl. Gather the pearl at the cauldron, and bring it to the navel.
4. Join the four pakuas by circulating the pearl from the ***Navel*** counterclockwise to the Door of Life and back to the navel. Continue circling the pearl nine times around the Thrusthing Channels and around the body counterclockwise.
5. Form a cross. Bring the pearl from the navel to the cauldron, and back to the Door of Life. Use it to connect the side pakuas. Return the pearl to the navel.
6. From navel move to the left side of the ***Solar Plexus***. Begin to circle the pearl at the solar plexus level. Circle the pearl to the left through T-11 and return to the solar plexus. Circle nine times. Create an internal cross with the pearl.
7. Bring the pearl to the left side of the ***Heart***. Circle nine times towards the left, to the middle of the shoulder blades, and back to the heart center. Use the pearl to form an internal cross connecting all four sides.
8. Next bring the pearl to the left side of the ***Throat Center***. Circle the pearl back to C-7, and to the middle of the throat center nine times. Form an internal cross. **Practice and master this step before continuing up to the crown.**
9. Bring the pearl to the upper tip of the left ear level with the ***Mideyebrow***. Circle the pearl nine times back to the small brain, to the upper tip of the right ear, returning to the mideyebrow. Feel a band of energy flowing around the mideyebrow, ears, and back of the head. Form an internal cross connecting all four sides.
10. Bring the pearl to the ***Left Side of the Crown****.* Circle the energy nine times counterclockwise at the crown. Connect the front, back, and both sides of this Belt Channel.

11. Allow the pearl out above the crown. Circle the energy nine times forming a ***Halo of Energy***. Use the pearl to cross the halo, front to back, and left to right. Collect the energy from above the head.

Descending Back to the Pakuas

12. Reverse the energy flow to a clockwise direction. Circle the ***Halo*** nine times. Then cross it front to back, and right to left sides.
13. Bring the pearl back to the ***Crown***. Circle to the right nine times. Then use the pearl to cross the front, back, and two sides.
14. Move the pearl down to the upper tip of the right ear. Circle from the right ear to the mideyebrow. Repeat nine times. Use the pearl to form a cross and return it to mideyebrow.
15. Bring the pearl down to the right side of the neck at the ***Throat Center***. Circle right to *C-7* at the back of the neck. Return to the front again. Circle nine times. Cross the front, back, and two sides. Bring the pearl to the front.
16. Bring the pearl down to the right side of the ***Heart Center***. Circle it clockwise nine times, ending at the front. Then, use the pearl to cross the front, back, and sides. Return the pearl to the front again.
17. Move the pearl to the ***Solar Plexus***. Circle it nine times to T-11 at the back, and return to the solar plexus. Cross the front, back, and sides. Return the pearl to the front.
18. Bring the pearl down to the ***Navel***. Circle it back to the Door of Life, and back to the navel nine times. From the navel bring the pearl to the cauldron, back toward the Door of Life, and return it to the cauldron. Use the pearl to cross the right and left sides. Return the pearl to the navel.

Descending from the Pakuas to the Earth.

19. Once you have controlled the pearl to this stage of the meditation, you are ready to move the pearl down to the ***Sexual Center***. Women bring the pearl down to the right side of the ovary center, located three inches below the navel. Men bring the pearl from the navel down to the right side of the sperm palace, one and a half inches below the navel. Circle clockwise back to the sacrum, to the left hip, and to the front of the ovary center/sperm palace nine times. Use the pearl to make a cross. Return the pearl to the front of the sexual center.

20. Bring the pearl down to the right groin. Circle the pearl clockwise around the ***Perineum Point and Lower Hips*** nine times. Use the pearl to form a cross by connecting the front, back, and sides. Return the pearl to the front.
21. Move the energy down and circle toward the back of ***Both Knees***, around and to the front, midway between both knees, moving the pearl clockwise nine times. Use the pearl to form a cross by moving the pearl to the midpoint between the knees and connecting the front, back, and sides from this point. End with the pearl midway between the fronts of both knees.
22. Direct the energy down to the feet, just below the ankles. Circle the ***Soles of the Feet*** nine times moving the pearl clockwise to the heel of the right foot, to the heel of the left foot, to the outer side of the left foot, to the front, midway between both feet. Form a cross. End with the pearl midway between both feet.
23. Move the energy from the front of the feet down to a point ***Beneath the Earth***, and circle toward the right (clockwise) nine times. Use the pearl to form a cross ***Beneath the Earth***. Keep on spiraling more energy to enhance the pearl. Return the pearl to the front of the circle.

Ascending from the Earth Back to the Pakuas

24. Reverse the circle beneath the earth, moving the pearl counterclockwise again nine times. Use the pearl to form a cross, connecting the front, back, and sides.
25. Move the pearl to the outer side of the left foot. Circle the ***Soles of Both Feet*** nine times moving toward the left heel, to the right heel, to the right foot, and to the front (midway) of both feet. Use the pearl to form a cross. End with the pearl at the midway point in the front of both feet.
26. Move up to the left knee. Circle the pearl counterclockwise *at* the ***Knees*** nine times. Use the pearl to form a cross. Return the pearl to the midway point in the front of both knees.
27. From the front of the knees, move up to the left groin area. Circle counterclockwise at the ***Perineum*** nine times. Use the pearl to form a cross. Return the pearl to the front of the perineum.
28. Direct the energy to the left hip, and circle counterclockwise at the ***Ovary/Sperm Palace*** nine times. Use the pearl to form a cross by connecting the front, back, and sides. End with the pearl at the front of the ovary/sperm palace.
29. Return the pearl to the ***Navel*** and to the ***Cauldron***.

Men and Women Practice Differently

***Men**:* Shoot the Pearl out of the Crown, and Spiral the Pearl down from the Crown.

1. Bring the pearl up to the crown.
2. Allow the pearl to go out three to six inches above the head. Circle nine times counterclockwise. Collect the heavenly energy and blend it in the pearl. Then circle it nine times clockwise.
3. Bring the pearl back into the crown point.
4. Spiral it clockwise through each of the Belt Channels all the way down to the soles of the feet.
5. Move the pearl into the earth six to twelve inches to collect the earth energy, and circle the energy nine times clockwise. Reverse the direction spiraling nine times counterclockwise.
6. Return the pearl to the front of the crown, and then back down to the navel and cauldron in the same manner.

***Women**:* Shoot the Pearl into the Earth, and Spiral the Pearl up from the Earth

1. Bring the pearl to the soles of the feet.
2. Move the pearl into the ground six to twelve inches to collect the earthly energy. Circle the energy nine times clockwise. Then nine times counterclockwise.
3. Move the pearl up through the left side of each Belt Channel, spiraling counterclockwise to the crown.
4. Shoot the pearl all the way out at the crown to collect the heavenly energy. Circle the energy counterclockwise nine times. Reverse the direction, and spiral the pearl nine times clockwise.
5. Bring the pearl back down into the right side of the crown. Continue to circle the energy clockwise down through the Belt Channel to the soles of the feet. Return the pearl to the navel and cauldron in the same manner.

Energy Body

Once you have practiced to this level, you are ready to form the pearl into the Soul Body.

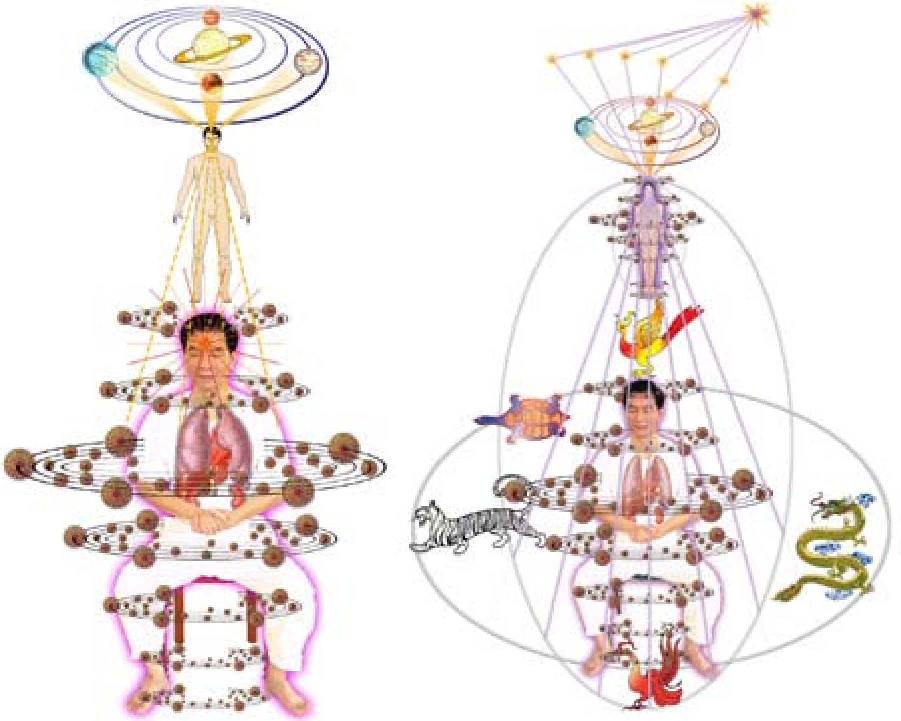

Fig. 8.15 *Energy or Soul Body*

Sorcerers saw that the essence of the universe resembled a matrix of energy shot through by incandescent strands of consciousness or actual awareness. They also 'saw' the essence of the human form. It was not merely an apelike amalgamation of skin and bones, but an egg-shaped ball of luminosity capable of traveling along those incandescent strands to other worlds. (Wagner, Bruce, "You Only Live Twice", *Details*, March, 1994, p.168.)

Just as it is important for the physical body to be strong so we can enjoy optimum health, it is also vital to strengthen the Soul Body so it becomes a sturdy vehicle for our consciousness. Strengthening Chi nourishes our health and spiritual qualities in this life and builds a strong Soul Body to carry us forth into the "afterlife" (which is, paradoxically, the here and now). We begin to strengthen the Soul Body through what is called in Taoist Inner Alchemy "the Inner Elixir". It is called the Inner Elixir because we are working with the energies already within our own body and converting them into elixir-like healing essences. Through the meditation practices of the Microcosmic Orbit and the Fusion of the Five Elements, we detoxify, nourish, and integrate the physical body, the Energy Body and the emotions.

Always begin by doing the Fusion of the Five Elements practice to clean out the negative emotions.

Chant the eight forces and form the collection points. Let the Pakua spinal around the physical body.

Condense the pearl and practice the Creation Cycle, Thrusting Channels and Belt Channels.

Run the pearl in the Microcosmic Orbit, and control its movement with the senses.

Move the pearl to the perineum. Inhale in short sips, drawing in ten percent of the lungs' capacity, while gently pulling up the anus. Inhale up to the navel, then to the heart.

Inhale up to the crown. Swallow your saliva (imagine swallowing upward,) and exhale forcefully to open the crown and shoot the pearl out.

Practice moving the pearl up to one foot, two feet, three feet, four feet, five feet, and six feet above your head. Go out only as far as you are able to maintain control of the pearl.

Relax the senses, and form the Soul Body.

Run the pearl in the Microcosmic Orbit in the physical body. Open the crown, transfer the Microcosmic pathway into the Soul Body above the head.

Form another pearl in the cauldron, and shoot the pearl into the Soul Body.

Circulate the pearl through the Microcosmic Orbits in both the physical and energy bodies together.

Extend the Thrusting Channels up into the Soul Body.

Practice all the Belt Channels up through the head, or down through the soles, into the Soul Body. Spiral all the pakuas around the Soul Body like a space ship.

Continue to practice until you are in full control of the Soul Body.

Note**:** You can also form a Soul Body below you by thrusting the pearl through the soles of the feet into the ground.

Spirit Body

Once you are well in control of the Soul Body, you are ready to form the Spirit Body.

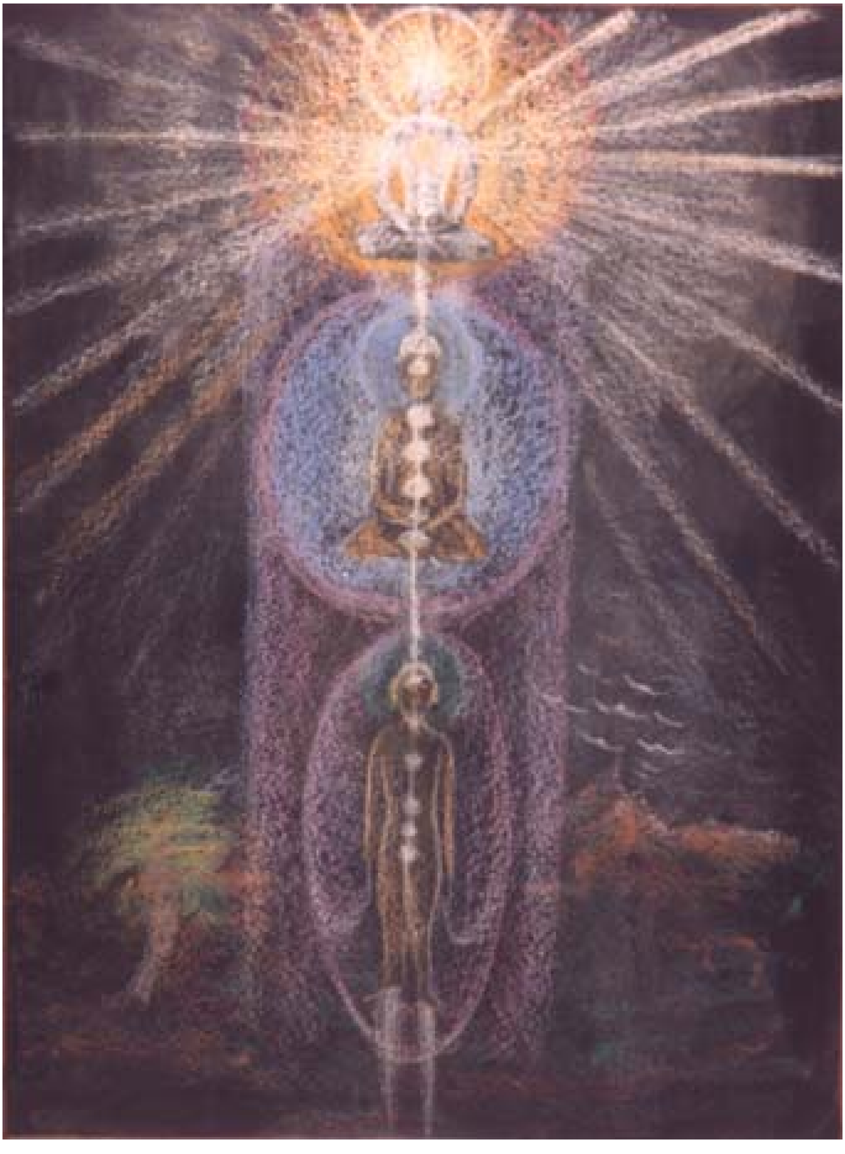

Fig. 8.16 *Spiritual Being*

Stage One

Be aware of the Four Pakuas in each center and spiral them around the physical body and the energy body.

1. Begin to form the spirit body after doing the Soul Body practice. Leave the Soul Body above the crown. Form another pearl at the cauldron, and run the Creation Cycle.
2. Form the pearl from compassion energy, a pearl that is more blue and gold than the Soul Body pearl.
3. Move this pearl in the Microcosmic Orbit. Bring it down to the perineum.
4. Inhale in small sips, and pull up the anus. Inhale the pearl to the navel; then to the heart; and to the crown.
5. Inhale again, and swallow up. Exhale forcefully to open the crown, and project the pearl to the perineum of the Energy Body.
6. Exhale, and send the pearl to the heart of the Soul Body. Exhale again, and send the pearl above the crown of the Soul Body.
7. Move the pearl up and down approximately twelve inches (30 centimeters) above the crown of the Soul Body. Then gradually extend the pearl further up in one foot (30 centimeters) increments. Practice until you are able to move the pearl up and down six feet above the Energy Body.
8. Condense the pearl. Condense the Soul Body. Bring both pearls back to the cauldron of the physical body.

Stage Two

1. Repeat steps (1) to (7) above, and run the Microcosmic Orbit in the physical and energy bodies.
2. Transfer the Microcosmic Orbit from the Soul Body into the Spirit Body, and then run the three Microcosmic Orbits together. As an alternative, you can run the Microcosmic Orbit as one large channel incorporating all three bodies.
3. Extend the Thrusting Channels to the Spirit Body.
4. Extend the Belt Channels to the Spirit Body.
5. Finish by condensing the energy of both the Soul and Spirit Bodies into a pearl. Pull the pearl into the physical body.
6. Collect the energy at the cauldron, and practice Chi Massage.

Chapter 9
Fusion as a Bridge

In the overall structure of the Universal Tao, the Fusion Practices serve as a bridge between the basic preparatory practices such as the Inner Smile, the Healing Sounds, the Microcosmic Orbit, Healing Love, Tao Yin, Iron Shirt Chi Kung, Cosmic Chi Kung, Bone Marrow Nei Kung, Tai Chi and the Higher Inner Alchemy Practices of Kan and Li.

The Fusion Practices are traditionally considered as the core of internal alchemy, as they focus on the transformation of emotional energies, contained in the major organs of the body, as centers and fields of energy which resonate, each with their particular frequencies, with the energies and forces of the universe.

Through a process of balancing negative and positive energies (Fusion of the Five Elements), the growth of positive energies (Cosmic Fusion) and the opening of meridians to enhance the circulation of internal energies of the body and make it receptive for outside energies (Fusion of the Psychic Channels), the body gradually becomes more open. Thereby the self-healing process, initiated in the basic practices, is further enhanced.

Once the strength and health of the body have been restoried so that it can serve as a vessel for the forces of the universe and is in a position to process them, it is ready for the higher alchemy in the Kan and Li practices. In these Kan and Li practices the physical body can then give birth to the energy body. Once this has taken place the person is ready to give birth to the spirit body, also called the Light or Rainbow body.

Fusion as a Journey of Self-Transformation

In the Fusion Practices, the body and its path routes are not only cleansed and opened, so that the energy from within and from outside can flow with more power and ease. Also the energy flows in the body are qualitatively raised. Thus the self healing and regenerative capacity of the body, already initiated in the basic practices is further enhanced. As a result, the body’s health and vitality receive a new boost and it grows in strength. At the same time, it becomes more able to absorb and process the various forces and energies from the universe, the cosmos and nature.

In the Fusion Practices, the body enters into a dynamic process towards new balances. This is an essential condition for the light to grow inside. This can then in turn attract the light from heaven as its energies become finer, as they blend into a higher quality energy created by the blending of love and compassion. This higher concentrate of energy is then condensed into a pearl which is subsequently circulated through the major organs and the psychic channels in and around the body. The pearl will help the body to feel centered and experience peace and harmony. It will also help to cleanse the special channels and enhance the free flow of energy.

The pearl will facilitate the growth and transference of consciousness from the physical to the energy body and from the energy to the spirit body (Cosmic Fusion), healing it and helping it to center and experience peace and harmony. Through the growth of love and compassion the body is able to radiate out and receive in the love and cosmic energies so that its own Chi is blended with the cosmic Chi and is thereby transformed into a higher quality of Chi.

To facilitate the process of blending and balancing the energies within the body between its organs and the gathering, processing and condensing of energies from the universe and from nature, the pakuas play an essential role (Formula 1, *Fusion of the Five Elements*).

In the earliest stage after conception, the first meridians which rule the circulation of energy of the embryo and which come into existence after the split of the egg and the formation of the body are the governor and conception channels. Successively, all other Yin and Yang meridians form in a symmetric fashion on the Yang and Yin sides of the body.

In ancient Taoists texts they are also called the sun side or outer side and the moon or inner side of the body. The sun side is that side exposed to the sun when one bows down, the moon side is that part which remains in the shadow.

The body might be visualized as grid of conduits of light which link it to the Light in which all energy and matter has its origin, a view which is supported by the new physics.

Unity of Ying and Yang

Basic to the Taoist tradition is the view that positive and negative energies are a manifestation of the very nature of the universe as a unity of opposite energies: Yin and Yang. It is by virtue of the dynamics of the interplay of Yin and Yang that the universe exists and continues to exist. They rule all phenomena in the universe as an infinite web of mutually interconnected processes and their movement. This insight is the theoretical as well as the practical basis of the Fusion Practices.

In Taoist theory, Yin and Yang rule the relationships between stars, planets, the sun and the moon and the earth, light and dark, above and below, man and woman. It is in and through this interplay, that the universe unfolds and functions. If one of them would disintegrate or disappear, the universe would end.

In the Taoist vision on the nature of the universe, the body is a microcosmos, which manifests and reflects in its inner structure the dynamics of the universe. As such it is also ruled by the same forces of Yin and Yang.

These in turn manifest themselves in the universe in the five elements (water, fire, wood, metal and earth). These five elements operate as energies in the body, in and through the five vital organs: the kidney, heart, liver, lungs and spleen. It is the natural capacity of the body to heal and regenerate itself. The body is sustained and nurtured by cooperating and facilitating the workings of the universe through the balancing of the five forces in the organs of the body and the positive and negative properties of each of them.

It is the relationship between the positive and negative energies contained and reflected in the organs, which is the starting point of the Fusion Practices.

In the Taoist tradition, the body has always been viewed as a sacred vessel, which reflects the sacred nature of the universe as a whole. Interfering with its laws of motion, the balance between Yin and Yang is broken and with it harmony and mutual cooperation between the organs and the elements which compose the body.

As a result, the free uninhibited flow of energy streams, and with it the immunity of the body, are affected. In Chinese medicine, of Taoist origin, disease is seen as the outward manifestation of imbalances within the body. Hence the predominant attention in Chinese medicine and Taoist self-healing practice on the balancing and strengthening of the organs and their cooperative relationships.

The Taoist view on the outer and inner universe (the body), is the fruit of several thousands of years of highly sophisticated empirical observation. Thusly, feeling and sensing in the body and its structure, processes and needs, play a crucial role in creating balance within the body and between its organs and their energies, so that inner peace and harmony may grow (Formula 2, *Fusion of the Five Elements*)

Turning Inward

This involves a process of turning inward and to feel and sense oneself "from within" so that understanding can grow. This in turn further supports the effectiveness of the inward-oriented process of listening. By turning the senses inward and reconnecting them with their corresponding organs (ears/kidneys, eyes/liver, tongue/heart, mouth/spleen, nose/lungs) the process of externalizing the consciousness and thereby draining the energies out of the body can be inverted. This allows the body to come to rest and recuperate its energies and the connected senses.

By protecting the senses from undue outside influences, the life force energies in the body and the higher energies it accumulates, through the blending with the energies from all the directions, but especially those from the stars and the sun and moon and planets, are preserved and can grow without interference (*Cosmic Fusion,* Formula 3).

This process of protection is further deepened and completed in the high formula of Sealing the Five Senses of the Kan and Li Practices where protective auras are created around the body from the accumulated energies of each of the major organs.

There is abundant evidence that in the process of externalization which has been imposed on the body in the process of modernization and its growing pressures on the body, the senses have become atrophied. By turning them inward, it becomes more easy to control them and protect the integrity of the body and its organs so that it can regenerate.

The Taoist approach to the body is rooted in the vision that there is an inner order in the very nature of the universe and all its manifestations. The outer order is a manifestation and reflection of this inner order. If the outer order is in harmony with the inner order, then harmony and peace reign.

As our receptivity and ability to sense and feel ourselves from within grows, so grows our ability to experience the universe within ourselves. By looking deeply into ourselves we grow. Our understanding of our being and body as a microcosmos, and its intrinsic unity with the universe as a macrocosmos, grows.

From Balance through Resonance to Love

As our body and its chemistry and we ourselves are changing, our consciousness and perceptions also change and widen. Then a point will come where we begin to experience ourselves less as separate and isolated from the universe and more in resonance and in tune with the universe.

In this process, we grow more aware of the wave-like nature of our body in and through which the universe pulses within us and how we pulse in and though our bodies with the energies and forces of the universe. At the same time our ability to tune in with the energy of the universe and its frequencies and to absorb and draw this energy into ourselves will grow.

The Fusion Practices serve an essential function in making our being more open through the heightened sensitivity of the body, as our energies become purified and more fine and subtle. The desire rises within the body to enter into the primordial unity with the original energy by which it was originated as a living being.

In the ancient Taoist vision this relationship has always been seen as a love relationship which reveals itself when our body returns to its inner balance and Yin and Yang enter into balance. The ability to attract and conduce the energies from the universe into one body through the network of conduits, represented by the grid of meridians, is dependent on the quality of the inner work towards unity between Yin and Yang.

This meeting between Yin and Yang has in the Taoist tradition been seen as a marriage, born out of love between the opposites, as they move towards unity. The Fusion Practices play an important role in the preparation of a state of being which make this movement towards unity possible.

Enlightenment of the Body and New Physics

The Taoist insight that human beings, with all other phenomena, share common origins in the universe and that our origins are in heaven and originate in the stars is shared in many other ancient cosmologies: those of the African peoples, the indigenous peoples of North and South America, as well as the Australian aboriginal people.

The Taoist tradition, however, is unique in that it has evolved over time concrete practical formulas and practices to work with. These help us to reconnect and in the process rediscover the inherent ecstatic nature of our body, through the process of self-transformation.

These formulas, of which the Fusion Practices are an integral part, are based on the insight of the nature of the meridians and their application in healing the body and restoring its natural functions, which form the basis of Chinese medicine.

A closely related Taoist thesis on the nature of the universe, also to be found in all ancient cosmological traditions, including those in the West (before the scientific revolution did away with them, viewing them as superstition) is the unity, indivisibility and interchangeability of matter and energy.

This thesis finds full support in the new physics which evolved from the beginning of this century with the path-breaking discoveries in quantum physics and relativity theory by the great physicists Bohr, Einstein and Heisenberg.

Their discoveries shattered the foundations of the old paradigm, underlying Cartesian-Newtonian physics, at the basis of modern science. They broke radically new ground in the interpretation of the nature of matter. In the new vision of advanced physics, matter, in its inner nature, is understood as a process and an event. Matter and energy are forms of the same reality. As matter can change into energy, also energy can change into matter. Both matter and energy are viewed in this new interpretation of reality as having their origin in light.

Therefore the Taoist vision on the intrinsic nature of the body as a light body is not a romantic vision, invented by creative minds. The image, underlying it, corresponds to the insights of advanced physics. Thus what has been called in ancient spiritual traditions “enlightenment” refers to a real material process: the illumination of the body.

Here's what distinguishes the Taoist vision from most others: enlightenment is not understood as a process which happens "in the mind", in separation from the body, but the process takes place in and through the body and its progressive transformation.

Thus a true process of spiritual growth in the Taoist vision implies the "enlightenment" of the body. The progression process of illumination is a bodily process, visible to the eye and tangible in the quality of energy which a person manifests and radiates in her/his body. This emphasis on transformation as a bodily process is a principal characteristic of the Taoist tradition. In that respect it differs from other ones, marked by the duality of body and mind and in which the body is viewed as a stumbling block; a hindrance.

This also explains the Taoist insistence on the central importance on rooting and grounding the body in the process of self-transformation. The more one advances in the higher practices, the greater the need to root and ground so that the energies in the body remain in balance. Equally important is centering; the balance between heaven and earth, the energies from above and below are maintained and enhanced.

It is precisely for this reason Taoist practice gives the highest priority to the creation of a healthy and strong body through a healthy and relaxed life style and an intelligent natural diet thanks to which the body can continuously regenerate and rejuvenate itself.

These insights had their basis in a new view on the nature of electromagnetic waves as particular forms of energy and in the insight that light is an electromagnetic field which takes on the form of waves which can travel through empty space as very light vibrations, also called ether. This energy is seen, in both the Chinese and Indian cosmologies, as the mother energy from which the other elements originate and of which the universe, the cosmos and nature are composed: fire and water, wood and metal and water.

In the Taoist system, these elements correspond to the five seasons, temperatures, senses, colors, sounds and positive and negative emotions. They form the theoretical and practical starting point for the Fusion Practices.

Return to Cosmology

The revolutionary implications for our perceptions of the universe, the cosmos, time and space and of nature and ourselves as part of nature are only slowly penetrating into the social consciousness. This social consciousness's view of its conditionings and fixations in cultural patterns and corresponding perceptions of reality prevent it from being open to an understanding our true nature.

There are multiple mechanics at work in present day culture and science and the modern view of reality which tend to prevent us from giving up the mind sets, rooted in the subconscious (and which rule the conscious mind, unaware of the forces by which it is steered). These make it appear as if the universe is composed of an infinite amount of loosely, accidentally assembled particles and solid objects; each having an independent, separate existence, rather than what is actually an ever changing web of unceasingly changing events which are mutually interconnected and interdependent.

The emergence in Western history and culture of the perception of human beings as essentially sovereign and independent from the universe and as its ruler who can alter life and the origins of life at will (with the progress of science and technology) are wholly questioned by the Taoist vision, presented above.

Such a view on the presumed independence and freedom from the laws of motion of the universe also has shaped the basic assumptions on the nature of human beings and social relations which have shaped mainstream social science and psychology, which mirror the basic premises of the Cartesian- Newtonian paradigm.

They support a view of progress and development, the new version of which emerged in the course of this century, which is essentially inspired by the view that reality can be manipulated and engineered without regard for the consequences on Nature. However, with the more recent findings of the adverse effects of such an attitude and the growing concern that the very foundations on which life on earth rest are being jeopardized, the basic premises on which this attitude is based are increasingly being questioned.

At the same time, as a natural consequence, there is a growing interest in the visions which have inspired the ancient cosmologies such as the Taoist one, especially since there is increasing evidence in advanced physics which support such visions. Thus it is not surprising that one of the leading philosophers of science, Stephen Toulmin, recently published a book, *A Return to Cosmology.*

In his study he traces the process of alienation in Western culture and science, as it came to shape modern life. Especially, he focuses on the relation between body and mind. He foresees a return to ways of looking at reality which help to overcome the duality at the root of this alienation and the imbalances it has brought about in the present day world system and which it feeds through the new communication technologies.

It cannot be emphasized enough that when the Universal Tao system is called a whole of interrelated practices, this is based on the root assumption that it is only in the practice that experience is gained. Therefore, a process of self-transformation is set in motion which makes it possible to come to terms with the patterns within oneself which prohibit one to discover its relevance.

This is the very basis of Eastern epistemologies, not only basic to Taoism but also at the root of other Eastern religious and spiritual traditional epistemologies: that true knowledge can only be acquired experientially and existentially in and through the practice.

Also implicit in the Taoist vision, shared in other ancient cosmologies, is the view is that the universe is alive. Not only as it births life at every instant, but also as it sustains and nurtures life. This view is quite contrary to that underlying the materialist world view, which originated in the West with the advent of modern science, in which the universe is reduced to an accidental anomalous phenomenon .

Religious Nature of Taoist Practice

In the Taoist view, supported by the new physics, the body is not only matter but also a field of consciousness which, by its wavelike nature, is intrinsically related to the whole universe and can also get in touch with it. This provides the very basis for the Cosmic Chi Kung practice as a practice, which helps to enhance and balance the body and which is essential to the Fusion Practices.

These above views are essential to the assumptions on energy underlying the ancient Taoist Practices, contained in the practices of the Universal Tao and which offer a synthesis and which for the first time have been published in the West. Until recently these were scattered and inaccessible, except for highly fragmented information. Essential to the Taoist tradition is religious practice in the feeling and experiencing of the sacred and the divine, and not in the beliefs about them.

The Taoist view on the nature of religiosity as an experience of connection and reconnection is very much in line with the original meaning of the root of the word religion in Latin, “re-ligare”, to reconnect.

Thus the practice is oriented to rediscovering the intrinsically ecstatic nature of the body in which it spontaneously feels the truthfulness of life as a sacred, naturally ecstatic activity. The whole practice may be summarized in facilitating the process of reconnecting and returning to the state we were in before being born.

Basic to the transformation process which is conducive to this rediscovery is the Healing Love practice by which the sexual energy is retained, its flow is inverted and the nature of this energy is transformed. The Fusion Practices play an essential role in making possible the further refinement of this energy in the Kan and Li practices through new forms of purification and by enhancing the receptivity of the body for higher energies from the universe. The Taoist practice takes literally the Christian tenet that the body is the temple of God by viewing the vital energy or life force energy as a sacred energy through which the body reveals its divine nature as an intrinsic endowment in which heaven and earth merge.

Wisdom of making Friends with the Devil

The three monotheistic religious mainstreams in their institutionalized esoteric form -- the Jewish, Christian and Islamic -- are marked by a deep duality between high and low, sacred and non sacred, right and wrong, good and evil, mind and matter, mind and body, mind and energy. This duality has also shaped other religious and philosophical systems in their social and institutional evolution; systems such as Hinduism, Buddhism, Taoism and Confucianism. These systems have been marked by beliefs and codes of ethics which have tended to split humankind between believers and nonbelievers and between those that are good and those that are bad.

Their focus on the duality between God/ the divine and heaven and earth, God and the Devil, good and evil, mind/spirit and body and on sin and its corollaries, fear and guilt, have created much violence and tend to continue to do so. They have also greatly affected people's sense of dignity, self-esteem and independence, as they have tended to create major dependencies on institutional intermediation with the divine.

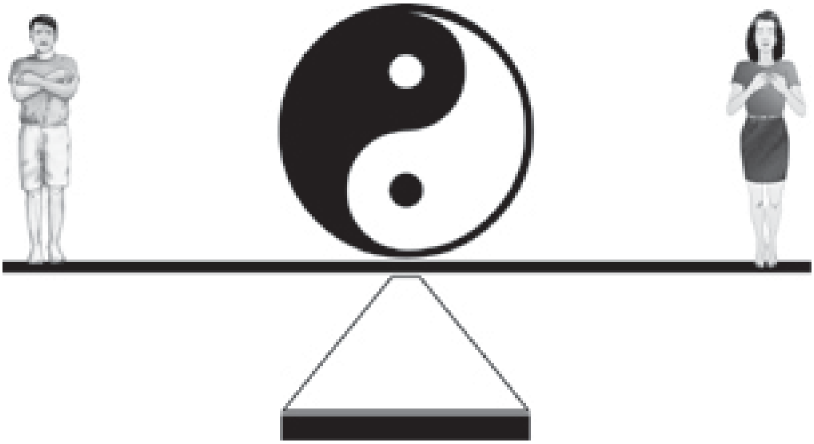

***Fig. 9.1** Balance of Yang and Yin (Good and Bad)*

In many pre-industrial cultures there are proverbs and legends which suggest that it is wise to make friends with the devil rather than fight him. It would seem that such intuitions are based on the insight that it is better to reconcile and accept oneself so that the negative loses its power rather than opposing it, by which its power increases, especially when fear and anger rise and the negative is turned into an obsession. Simple observation confirms the truth of this insight: that the energy of what is viewed as negative and dark increases to the extent that there is an attempt to eradicate it.

This same principle shows up in the martial arts: never initiate the fight but let the enemy come and neutralize him by making use of his own energy, as in the case of Tai Chi and Aikido.

A leading philosopher of culture and humanist therapist in the West, Erich Fromm, has said that the propagation of these views, often in extremely violent ways, has broken the spiritual backbone of the people.

Esoteric systems have, however, never succeeded in wholly eradicating and repressing their esoteric roots, visible in multiple movements throughout history, in which the body was considered sacred and a temple of the divine and in which sexuality was viewed as intrinsically sacred.

It may be said that all esoteric systems, in their origins, have called on humankind to practice love and compassion and thereby to transcend ego and enter into a state of grace.

Adepts like Jesus who has inspired the Christian tradition, Mohammed in the Islamic tradition and the Gautama in the Buddhist tradition may be seen as great masters. They showed people ways and practices of self-transformation through the practice of love and compassion. By assuming responsibility for themselves, they could attain inner freedom and bliss and transcend themselves.

In the Taoist view, the problem with religion and philosophy is not in the practice, but in the beliefs which often serve to divide rather than unite, inside the person and outside.

Unity of Light and Dark

In the Taoist perspective, there is no sin, as light and dark are two poles of the same indivisible reality. Thus, attempts to eradicate the dark, so that light will prevail, are bound to fail and tend to lead to conflict and violence. Attempts to conquer good at the expense of evil (for example, the Crusades) are bound to generate more violence. And battles to secure good over evil will create more problems than they solve. In the Taoist view there should be neither winners nor losers.

Fig. 9.2 *Light and Dark of the Tai Chi Symbol are connecting.*

Thus, balance between opposite forces and energies, as expressions of Yin and Yang, the positive and the negative, is essential to Taoist perspective, as expressed in the *Tao Te Ching*, attributed to the great Taoist Master of self-transformation, Lao Tzu.

From a Taoist perspective, balance is enhanced by entering into alignment with the universe, when one actualizes the universe within oneself and so is empowered and realizes oneself. This state has also been called a state of Selfhood, as one has returned to one's original Self. This process is also called "returning to one's origins". This means that one has overcome the contradictions and imbalances in oneself, so that Yin and Yang enter into balance and harmonize. The Taoist texts speak of becoming the "Undivided Being". Lao Tzu uses the metaphor of the unhewn log (p'u). It manifests a state of being in which when has returned to the original state of spontaneity, in the old Taoist texts called " Tzu-jan) which can be observed in babies and small children. This reminds us of the saying of Jesus," if you do not become as children, you cannot enter the Kingdom of Heaven".

Thus the Taoist view on harmony does not carry moral overtones, as it refers to a state of being which is intrinsic in the body. This state of being of the body as a microcosmos lies in the balance within and between its major organs, through which our body resonates with the energies and forces of the universe.

In the Taoist perception, there are no heaven and hell outside, above or below us. They are seen as states of being of our bodies and our inner selves and the way we relate to the universe, ourselves and each other. As such they reflect states of balance and imbalance between the positive and the negative. In this view, human beings are responsible for their own state of being, the quality of which depends on their state of self-awareness as the key to self-transformation.

This view is reflected in many ancient spiritual traditions and cultures which suggest that it is wiser to keep the devil as a friend than try to combat him and do away with him.

This liberating view, which frees us from self-condemnation with its tendency to obscure the positive within the negative and its transformational force, is the basis for the Fusion Practices.

Life as an Invitation to Learning

The Chinese character for crisis stands for both a negative occurrence as well as for opportunity and reflects the Taoist view on the unity of the positive and negative. In this perspective, whatever happens in life is to be seen as an invitation for learning and an opportunity for awakening. Theoretically speaking, problems are not problems, but opportunities for self-transformation, since the negative contains the positive, in the same way as the positive contains the negative, like the Yin is in the Yang like the Yang in the Yin. What may be seen and experienced as negative, carries within it the seeds of the positive, just as the light exists within the dark. In this light, all problems of life may be seen as energy potential, which can actually be a learning process.

This is a key premise in the Fusion Practices, and it helps to overcome the duality which we tend to construe, as a result of attachments to our negative experiences and emotions and therefore keep ourselves from entering into balance.

Taoists say that learning is enhanced if the organs feel at ease and find themselves in a state of well being. When a state of disease arises – which in Chinese medicine, of Taoist origin, is seen as a symptom of the breakdown or weakening of the free flow of energy – learning becomes more difficult.

It also points to an attitude which gives primacy to the need to reflect on one’s own attitude and perception in resolving problems and to look inward to see in what way one has contributed to the problem and how one can contribute to its solution by changing one’s own attitude, perception and way of life.

Thus from a Taoist perspective, problems are not seen as obstacles or difficulties but as opportunities to learn and as learning material.

In this connection, it is interesting to note that the word disaster, common to many European languages, has its etymological root in the Latin substantive “astrum” of which it is a negation by the prefix “dis”. This suggests that originally the word had the connotation that if and when one lost touch and became alienated from the stars (as representing our origins), one was in for trouble; imbalances would arise.

Smiling as the Key Taoist Practice

Smiling and relaxation are the beginning practice of the Universal Tao. They are the key to the transformational process and to learning, as a result of which one is enabled to gently and patiently change old patterns into new, concerning oneself and the universe: new breathing, posture, inner and outer movement, and a new perception of reality.

Essential therefore in the Taoist practice to enhance learning and get in touch with the body by turning inward is to learn to relax and smile inward, so that dis-ease, manifest in many forms of stress, is overcome and one returns to a state of ease. Thus it may be said that whereas the Western approach to medicine starts from outside, the Taoist approach starts from within.

It is thus not incidental that the whole practice of the Universal Tao starts with the practice of the Inner Smile. This is actually the first practice of internal alchemy, as it serves to soften our body and its elements in which our consciousness is stored and through which it operates. It is the key agent of the whole activation and transformation process in all the practices. We do not only have a body but we also are our body. By touching it with love and gentleness, its knots and blockages start to dissolve so that the free flow of energy is restored.

Unity of Fullness and Emptiness

It is in this context not surprising that at the heart of the ancient Taoist texts is the emphasis on the practice of turning inward and of stillness. It follows that in the process of meditation as a practice of becoming aware of the inner state of being of the body and its movements, a new and fresh space is created in which renewal and transformation can take place. This space is also called the creation of emptiness and the continuous re-creation of emptiness so that new fullness becomes possible.

In the Taoist as well as in the Buddhist tradition, which historically have been closely related over long periods in Chinese history, emptiness and fullness are indivisible and a condition for each other. In the Universal Tao practices, this view has vast implications. For example, in the universe, emptiness and fullness are indivisible and are the creative source of becoming, evolving and renewal, so that the universe is reborn in a continual process. We ourselves give thanks to the alternation of the inhale and the exhale, so our body is in a continuous process of dying and being reborn, through which we renew ourselves. The new can only rise if the old dissolves. So it is in Nature with the change of seasons. And so it is with all that is alive in the universe. This is the core wisdom of the *I Ching*.

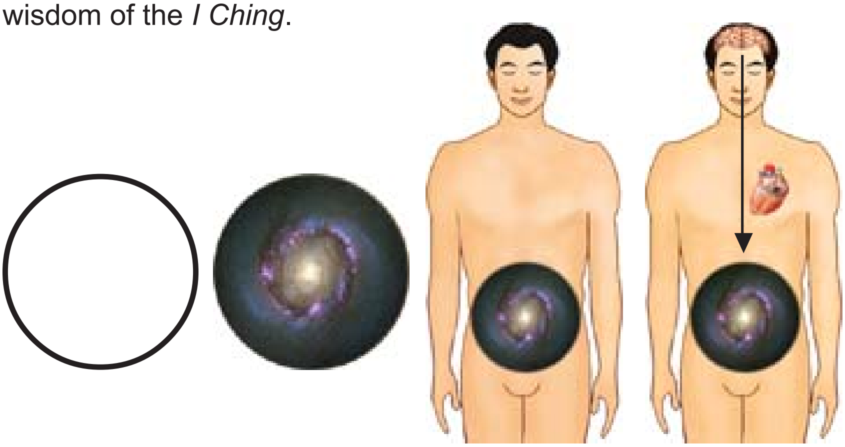

Fig. 9.3 *The core is the center of the universe, earth and the body.*

In the Taoist practice this has a particular application in the counsel to empty the mind and let it come down in the body so that a major source of energy is preserved and can renew itself. At the same time the belly can be filled with Chi, so that the energy streams in the whole body are enhanced and the heart is freed from too much pressure, as the belly helps do its work. Only when we learn to renew ourselves by undoing what has become unnecessary and what obstructs the process of transformation do we renew ourselves and our body. By undoing ourselves from what is no longer needed the new can grow. An ancient Taoist proverb suggests that simplicity comes about by seeing every day what has become dispensable. Then one can walk and move more lightly, both in the mind and with the body.

Jesus often reminded people in His parables of the law of unity and alternation in nature, and in themselves of old and new. In the *Tao Te Ching* this truth is expressed in the use of rich metaphors, for example, the house which acquires its function by being empty and the wheel which receives its function by virtue of its emptiness.

It finds its most pregnant expression however in the law on the unity of living and dying, common to all the great spiritual traditions – Hindu, Buddhist and Christian –, which is also central to Taoist practice. The new can only emerge and grow, if the old dies and is discarded.

Attachment and Detachment

This insight is at the core of the *I Ching* in which nature with its cycles and seasons is held up as a mirror for human beings. It has shaped from ancient times the Taoist proposition that to learn to be and realize oneself, one needs to go with the flow of life and to become like a fish in the water or the clouds in the sky. This insight is essential to give depth to the Fusion Practices and balance attachment with detachment.

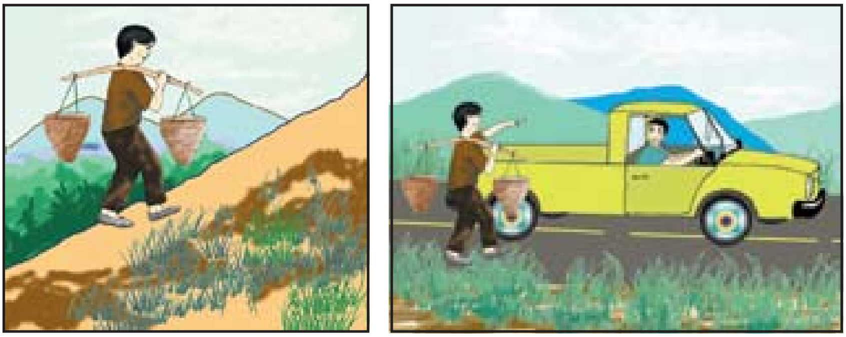

A: *Why don't you sit down, relax and let go of the weight and let me take you to the market.*

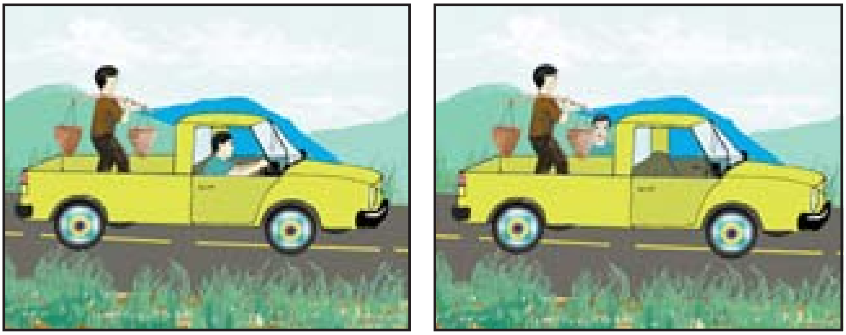

B: *Thank you for taking me to the market. I don't want to put too much weight on your car.*

Fig. 9.4 *Relax and let go.*

In the Taoist view, attachment and detachment are not to be seen as opposites, which exclude each other. What may be useful and functional in one stage of life, in one particular situation or moment, may well lose its functionality at another time. This explains why Taoists question the usefulness of externally induced and pre-arranged moral codes. Instead , they value, above all, self-cultivation in learning to act in harmony with what is inwardly experienced as genuine and truthful, because it is in tune with the universe in us.

The training in such an attitude in which one frees oneself from fixed rules imposed and determined from outside and one consults one's own body and inner feeling, creates both freedom and flexibility. It helps to avoid being trapped in positions which undermine the capacity to trust oneself, and at the same time raises self-esteem.

In the process of genuine self-reflection and the growth of self-awareness, you learn to look at yourself and your inner life from a distance and as a neutral witness, rather than as a judge. As a result, you grow into a spirit of inner freedom and you learn to look at yourself with a sense of relativity and even with a sense of humor. At the point you are able to truly smile at yourself and have reached the point you can forgive yourself, wisdom is dawning and the knots which you felt were in your way are in the process of dissolving, to create space for new and fresh energy.

Going with the Flow

This view on going with the flow in the universe and within yourself is also the basic principle underlying and inspiring the ancient Taoist arts of self defense. In Tai Chi no position is fixed and the body enters into a continuous flow of movements, as life is movement and in the movement to become one with the universe and its energies and forces one becomes invulnerable. The power gained in the martial arts is not an external but an internal one; it is generated by the process of alignment. It is this alignment through which the body is empowered from below and above and the power can flow from the center.

If you have no center, you can be easily pushed around. A person who is centered has an internal balance. He is more free from external influences and less fearful and suspicious of others, as he is in tune with himself and the universe. Therefore the fusion practice may be called an internal kind of martial arts by which one learns to playfully bring into balance one's opposites so that a sense of being centered and in harmony arises and thereby self-confidence manifests.

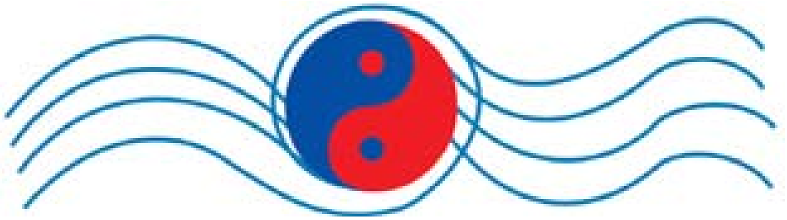

Fig. 9.5 *Balanced, going with the flow*

It should be obvious that the practice of these arts may be most helpful, as they may help you to remember to go with the flow inside yourself and become inwardly soft and flexible. Like all true Chi Kung, the art of moving the Chi and with the Chi, outer flexibility comes from inner relaxation. Only when relaxation deeply enters the body, can it become truly receptive for the energy within and from outside. The same goes for Tai Chi.

In that sense the Fusion Practices may also be seen as an inner play in which you invite your organs as your children to play with each other and see how they can balance and support each other. It may also be compared with an ancient game, which is played by children and adults in different countries in Asia. In Thailand it is called Takara in which the players help each other in keeping the light bamboo ball in the air so that the play can continue. Nobody loses, nobody gains. All share in the joy of the game and in the process train their own ability and help the others to enhance theirs.

Central to the Taoist tradition is that the body expresses and balances both heaven and earth. How could it receive the energies of heaven and earth and become an expression of the universe, if it were not open and flexible and receptive? This explains the crucial role of Chi Kung and the various schools and forms which have been created over the ages, so as to make the body a vessel. As the *Tao Te Ching* says: the function of the vessel is to be empty so that it can be filled.

Secret of Breathing and Relaxation

The essential role of the quality of breathing in all the practices is to enhance detoxification and cleansing, to create new space and openness towards the universe (in the exhale) and to fill the body with new energies and forces, by receiving, enriching and transforming Chi (the life or vital force). Breathing is central to the Taoist practice.

Changes in the quality of breathing from a short, superficial and agitated breath in the upper part of the body to a long, deep and quiet breath in the lower center/ Tan Tien of the body are crucial to the process of inner transformation. This is so because breathing is the bridge between consciousness and the body, and in the breathing process both the body and our awareness change.

The quality of breathing plays a crucial role in the processing and transformation of emotions which is a central theme in the Fusion Practices through the transformation of negative into positive emotions (Fusion of the Five Elements) and the cultivation of positive emotions (Cosmic Fusion). The emphasis in Taoist practice on self-healing is intimately related to the growth of self-awareness with regard to one's pattern of breathing as this determines the ability to regulate, control and balance emotions.

The Inner Smile practice as well as the Healing Sounds practice, designed to balance the negative and positive properties of the organs, are essential as a preparatory practice to the Fusion Practice.

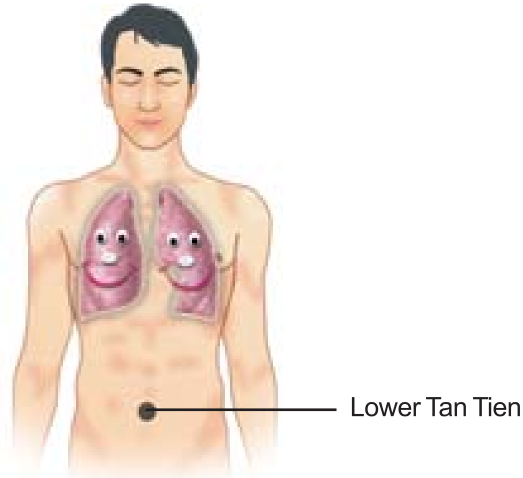

Fig. 9.6 *Smile to the Lungs.*

From the above, it should be clear that the practice of the Inner Smile and the quality of breathing and relaxation – as conditions of the self-healing process – are intimately related.

It is not coincidence that the great sages and masters of self-transformation like the Taoist immortals and the Buddha are always portrayed with an inward-oriented smile, to signify the key to inner freedom and balance and the unity of Yin and Yang. The practice of the Inner Smile as the key to relaxation remains the basis of all the practices of the Universal Tao, from the most basic to the most advanced. It is the key to all forms and stages of self-transformation, as the Taoist masters discovered its secret and its power to initiate and enhance the process of self-transformation. Statues and pictures may remind us of our own transformational potential and may help us to set in motion the internal alchemical process. We may also bring into our mind images of people who have awakened or reminded us in the course of our life of our own potential for transformation and positive qualities, thereby helping us to empower ourselves.

Science of Feeling and the Power of Intuition

In this sense, the Taoist practice is a continuous invitation to turn inward and get in touch with one's own body (and organs). Begin to sense and feel the body's state of being, its expansion and contraction and the state of relations between the major organs and the senses (kidneys/ears, liver/eyes, heart/tongue, spleen/mouth and lungs/nose). Check the state between the major organs and connected organs (kidney/bladder, liver/gallbladder, heart/small intestine, spleen/stomach, lungs/large intestine). Review the overall state of relations between the organs, glands, the nervous system, the spine, the bones, muscles and tendons. How are the relations between the three centers of the body (belly, heart and crystal room)? Check on those between the center and the extremities, the front and the back and the quality of the path routes which carry the flows in the body (the electro-magnetic energy, called Chi, also called the life force or vital force, the blood and the lymph flows).

Any advance in the practice is determined by the growth in ability to feel into the body by going inward and getting in touch with it.

In this context it is of value to realize that the verb "heal", the substantive “whole”, and the adjective “holy” have the same etymological root. This is in line with the Taoist insight that the process of healing is realized in the process of becoming whole and that holiness is the state of having become whole and in balance with oneself and with the universe within oneself. In this view holiness is not a moral category but an existential and relational state of being.

It is not without reason that Taoist practice has been called a science of feeling, as throughout the ages it has relied on and cultivated it. One may also call the ability to feel the power of intuition or the feeling intellect which exists at every and all levels of our body and bodily awareness in which we are wholly present. In all languages of the world, there are multiple expressions, pointing to the organs, such as the heart or the liver or the kidneys or the blood or the bones or the veins, hands, eyes, ears and feet which reminds us of this intuitive power of the whole body. The reduction of thinking to the mind and the brain, so characteristic of Western reductionism, is entirely alien to all cultural traditions in the East and the South.

It is the feeling that we open the body and open and activate the Chi from within and outside.

It is also through feeling that we can sense, taste, hear and touch the rhythms of earth and heaven and unite with them within ourselves. Becoming aware of our inner movements and what moves us are one and indivisible. In the feeling process, we also become aware of the relativity of our boundaries and learn to experience these not only as limitations and finite but as openings to the unlimited and infinite.

As a result our solitude comes to an end and we become sensitive to our intrinsic relatedness with all that is. In this sense, each practice may be seen as a step to invert the process towards alienation and separation from the sources of life, basic to Western culture and its extension over the world.

Thinking with the Heart

One is reminded of the meeting of Carl Gustav Jung, the famous Swiss psychiatrist and philosopher of culture in the 1920’s with an old Pueblo Indian in New Mexico on one of his visits to the United States of America. Jung narrates in his autobiography that his talk with him was unlike any which he had had with a European. When

Jung asked him how he felt about white people, Ochwia Biano (Mountain Lake) said to him, that to his people, they looked quite cruel in their appearance. “You always have an inflexible expression on your face. You always seem to be wanting something. You are always restless and agitated. We don’t know what you want. You all seem to be quite crazy.” Jung then asked him why he thought them to be crazy. He replied, “They say they think with their head”. Jung, surprised, asked him: “Of course, but tell me, where do you think?” He replied, “We think here.“ And he pointed to his heart. Jung tells how he sank into deep reflection and he remarks, “For the first time in my life, it seemed to me, somebody had shown me the image of how we really are.... This Indian had touched our weak spot and pointed to something to which we are really blind.” Jung viewed the encounter with the Indian as an opportunity, which opened to him a new gate to a primordial and nearly entirely forgotten dimension and form of consciousness which could be revived.

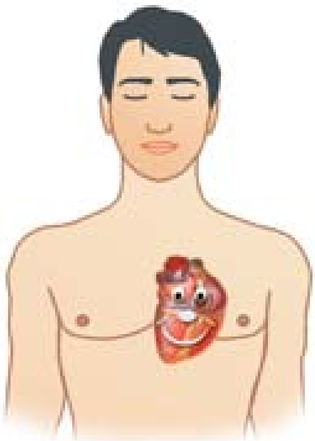

Fig. 9.7 *Smile into the Heart.*

In another passage in his autobiography, Jung narrates a dream which he had during a visit to an Islamic country in North Africa, which deeply touched him. In this dream he had to fight for his survival with a young Arab aristocrat who descended from his horse and wanted to drown him. Jung interpreted the dream as a struggle between his own repressed unconscious (his shadow), which sought to be recognized and accepted and his own consciousness, represented by his rational mind which felt threatened by the unsuspected attack by the young man. He then observes: “In my unconsciousness, I was in no way aware of such a situation; on the contrary I could not undo myself from a feeling of superiority, as at every step, I was reminded of being a European. I was not prepared to meet the unconscious forces within me, which emerged with such an intensity on behalf of the opposing party and which led to such an intense conflict."

Only several years later did Jung begin to understand the deeper nature of the dream. He began to see in the dream the resurgence of primordial structures or patterns in himself from a well-known past, but a past he had forgotten. He perceived the resurgence of these images as a renewed awareness of a still available potential for life, which was overgrown by civilization, a potential for life, which had been repressed and marginalized into the sub-conscious. In Jung's view, Western culture had become alienated to much of what being human represents: the primordial dimension of our being, which had been relegated to the underground.

Jung was a great admirer of the Taoist tradition and expressed his deep appreciation for its originality and wisdom in writing the foreword for the translation of the *I Ching* into German by Wilhelm Reich (the first translation into a European language).

The split between body and mind, thinking and feeling, matter and spirit, at the root of the formation of Western culture and religion and the birth of modern science, has been aggravated in the course of this century.

Feminine Underground Current

Some historians and scholars, like Morris Berman in his path-breaking studies "The Re-enchantment of The World" and "Returning to Our Senses" have argued, that in the course of this process the feminine dimension of European culture greatly suffered and in order to survive had to make itself invisible and so became its undercurrent. He means by this that in religion, with its reliance on external authority and its contempt for the body as well as in science, with its pursuit of objectivity, at the expense of other sources of cognition, the body was devalued.

Subsequently it could be turned into a mere object, an input and an instrument. It was seen in function of externally determined aims, as defined by the forces of competition, economic rationality and rational management and the new forms of communications technology. This process has also seriously undermined the position of women and has greatly affected their sense of dignity and self-esteem. It has also challenged the Yin values in society, considered to be counterproductive.

Unquestionably, the growing sense of despair and disorientation which large groups of people in the world today experience is directly related to the split and the resulting imbalances which they experience within themselves.

Restoring the Balance

With the growth of a new world culture, with a process of unceasing economic growth and expansion and the simultaneous rise of both minority affluence and mass poverty, these dualist tendencies are being further aggravated. The rise of fundamentalism and dogmatism, both as an instrument to legitimize power and achieve control and as a response to a growing despair and as an expression of the thirst for security is a further sign in the world rocked by imbalances and social fluctuations has become increasingly uncontrollable. This uncontrollability is seen as unavoidable and inherent in the very mechanics of the system and as an inevitable price for economic growth and advance.

Repression of the body and of sexuality and a search to compensate for it go hand in hand. Thus repression and a desperate search for liberation from it are both on the rise. There is a trend to see the body and sexuality as the source of all evil and start crusades against them so as to secure law and order. There is at the same time a growing obsession to secure and maximize pleasure, a trend which has grown to be the object of ruthless exploitation in which the body and sexuality have become mere commodities on the market. As a result, public life is being sexualized, as a source of economic expansion and profit.

Thus the repression of the body and sexuality have precisely an inverse effect and has become a major source of a fast-growing market and rapidly expanding pleasure and amusement industry, which capitalizes on this repression.

The first law of thermodynamics is that energy cannot be created or destroyed. In terms of Taoist theory, all energy comes forth from the Great Void, the original Wu Chi which undergoes a series of transformations in the formation of the universe. Thus the human body is seen as a particular form of transformation of this original energy. The second law of thermodynamics states that energy in a closed system tends to become disorganized (entropy). Thus there is an automatic tendency for balance to be lost.

Taoist practice focuses on inverting the process of dispersion and disorganization and restoring balance. This process is set in motion and facilitated by the Healing Love practice through which the unity and balance between love and sex is restored.

Essential to the Healing Love practice is the growth of love and compassion within as a result of which love and compassion with others and the universe become possible and emerge as a natural extension. Once the inner balance is restored, it is possible to act as an effective agent of balance outward.

In the Taoist view, conflict and violence in the outer order invariably have their source in inner conflict and violence, as manifestations of inner imbalance. Equally is outward peace viewed to have its source in inner peace and balance. In this process towards inner balance, Fusion Practices, in conjunction with the Healing Love practice and other practices of the Universal Tao, play an essential role; they honor the body and its intrinsic wisdom and potential for self-healing and rebalancing itself, so that it enters into alignment with the universe.

Joseph Needham, one of the great scholars of the origins of Chinese science and civilization, who was an honored member of the Academia of Science, had called Taoism the feminine undercurrent in Chinese history and culture, with its focus on the power of silence, feeling, sensing, compassion, forgiveness, humility, gentleness, reverence for life and nature and the equality of women.

It was however relegated to the margin of Chinese civilization and its practitioners had often to go underground. As a rule they fiercely stood for the ancient communitarian practices of reciprocity and sharing as well as people’s freedom and self-reliance. At the same time, they were opposed to interfering with the natural order, to be respected as a manifestation of the higher orders.

World as a Manifestation and Reflection of the Universe

They valued ancient forms of democratic practice, rooted in a vision of village communities in early times, which formed federations, patterned on the very structure of and alignment with the cosmos. No wonder that they were staunchly opposed to any forms of feudal and hierarchical power in which undue power was exercised by some over the many and the rise of a wealthy aristocracy which lived at the expense of the common people. They were also opposed to any forms of technology which would harm nature or be at the expense of people's sense of well being. Their attitude towards the universe and towards each other is summarized in the words of Lao Tzu:

> The Universe is sacred
> You cannot improve it
> If you try to
> You will ruin it
> If you try to hold it
> You will lose it
> Surrender yourself humbly
> Then you can be trusted to take care of all things.
> Love the world as you love yourself;
> Then you can truly take care of all things.

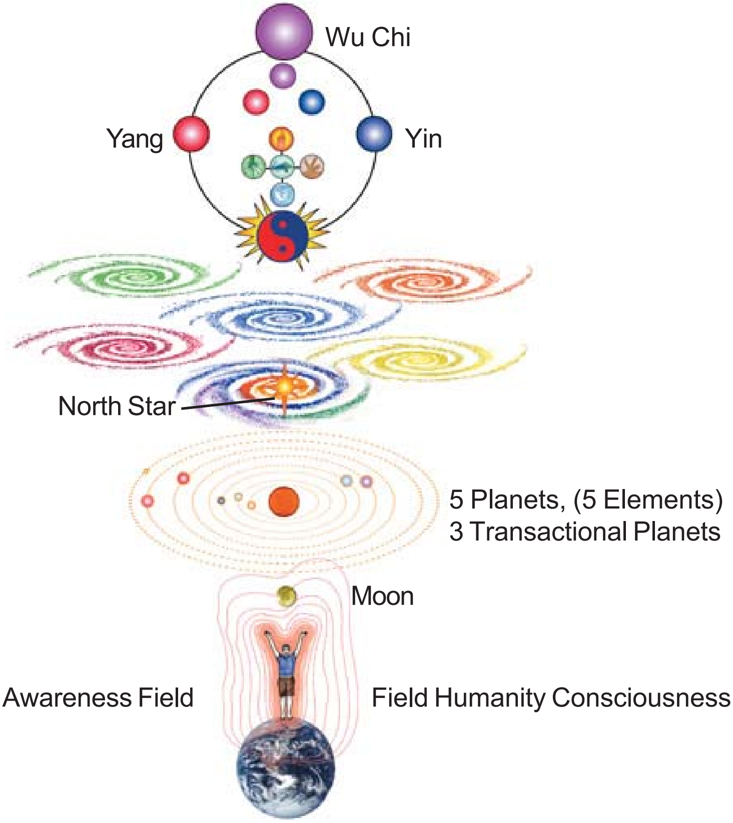
Wu Chi
Yang
Yin
North Star
5 Planets, (5 Elements)
3 Transactional Planets
Moon
Awareness Field
Field Humanity Consciousness

Fig. 9.8 *Alignment of the Cosmos*

It may well be that, precisely because Taoist practitioners refused to be co-opted by the establishment and were relegated to a precarious position, the practices could survive. As such, these practices, as synthesized in the Universal Tao, constitute a synthesis of several thousands of years of the practice by generations of Taoist masters in the science of feeling and sensing inward and of observing the nature of the universe and our relations with it.

The above provides evidence that Taoists over the centuries had a genuine concern about the well-being of the common people and that their practice was no reason to disengage from the world. The many stories and legends of the Immortals show that they were greatly loved, especially by the common people and the poor whom they often protected and helped. Their practice of gentleness (kidney), of generosity (liver), love and respect (heart), fairness and openness (spleen) and courage (lungs) expressed a strong sense of compassion with those who suffered.

Body as the Primary Source of Cognition

In the evolution of Western culture with its progressive disconnection between body and mind, thinking and feeling, rationality and emotions, mind and energy, inner and outer, matter and energy, linearity and spontaneity, feeling as a source of awareness and the primary mode of cognition has been progressively devalued.

It was seen as a feminine value of little use except in some domains as that of the arts and artistic endeavor. In religion it was looked at, together with the body, sexuality and emotions, as a potential source of evil. In science with its drive for objectivity, it was also looked at with suspicion, as it tended to disturb and undermine the demands for objectivity.

Thus it was relegated to a position of inferiority, as it was supposed to interfere with rationality, discipline and order, needed for progress and the growth of modernity, as well as with the need to control the universe, the cosmos, nature and ourselves.

In the Taoist tradition, feeling as the power of intuition, rooted in the body and the ability of self-reflection has always been considered as the primary source of self-awareness. Jung has named this form of cognition the knowledge of the unconscious, which underlies all conscious activity of the mind.

This capacity to intuit reality (the ability to sense reality in its totality may be called an integral form of knowledge) exists in our whole bodily being. It may also be called a mode of thinking with the body which precedes all formalized codes of thinking which evolved with the advance of formal rationality and the split of body and mind, typical of Western culture.

In that sense it may be called a Yin form of cognition, as opposed to the Yang form of cognition, which characterizes the modern world and its connected forms of education, science and technology. Yet there is abundant evidence that at the source of all genuine creativity and inventiveness, including in the field of science, lies the power of intuition.

In that sense, the practice of the Universal Tao and the Fusion Practices, deeply rooted in the Taoist tradition, may be seen as a way to recover this form of knowledge within ourselves so that the split within ourselves between body and mind, feeling and thinking is overcome. In the great Asian traditions, thinking in its restricted sense as the linear activity of the brain has been viewed as interfering with and affecting awareness as the primary and principal form of knowledge, essential for turning knowledge into understanding.

From a Taoist perspective, the primary learning process does not take place in the brain but in the organs. Recent findings corroborate the Taoist vision of the learning process that the brain only processes what is experienced and known intuitively by the organs, in particular those in the lower Tan Tien.

Our deep memory which structures our patterns of reaction and response to the world lies in the organs. In this view, a learning process needs to be rooted in a relearning and a reprogramming of our subconsciousness in the organs from which all conscious activity springs. In terms of contemporary technology the organs are the software and the brains are the hardware.

Unity of What is Above and Below

In many recent attempts to revitalize and recuperate spirituality, a growing need is felt to counter the ominous trend towards a loss of meaning in the world of today. There is a tendency to continue the split, which inspired many spiritual traditions, which tended to be based on the duality of good and evil, higher and lower.

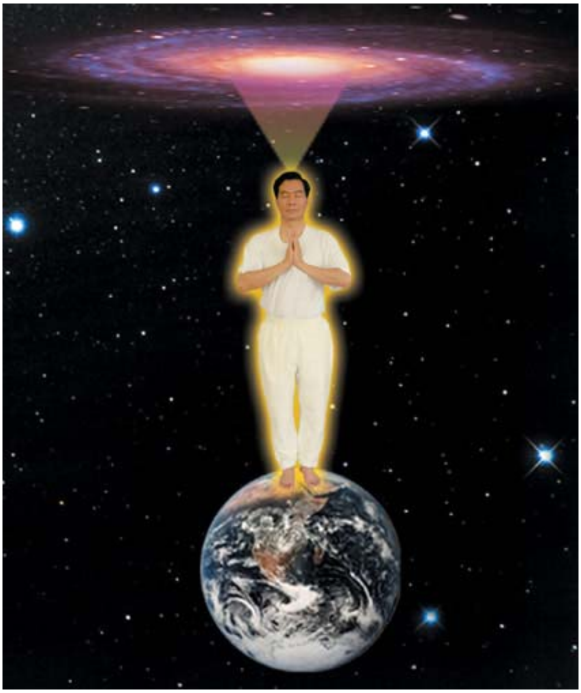

Fig. 9.9 *As Below So Above*

As the lower part of the body and its organs is looked at as associated with the instinctive animal obscure part, connected with sexuality, which prevents the mind and the spirit from its upward vocation, it is ignored and bypassed. The Taoist approach fully honors and recuperates the dignity of the body as a whole. The Tao recognizes the belly and its organs, connected with our sexuality and water energy which is the source of creative power from which all energy transformations flow. By getting in touch with our most intimate parts, we also gain access to the source of our creative powers.

In that sense, the Fusion Practices invite us to reconnect with those parts of ourselves, both physically and psychically, which have been the focus of repression. This repression is not only in most of our cultures but also within ourselves, as we internalized what we learned and accumulated in our culture, at home, in school and though the media. In order to reconnect and truly get in touch with our organs, we need an instrument which is capable understanding them.

That instrument is our intuitive power, which has its roots in our body as earth. It cannot be substituted by any other instrument, including our own intellect through which we are often at risk to become strangers to our own innate wisdom.

Genuine self-transformation can only be realized from within by the one who truly knows her/his personal existential history. We must see the relations between positive and negative emotions as our own most intimate history which we have often hidden from our own senses and which only reveals itself to our own compassionate inner eye and our own loving energy. Only then can we open the path to true healing from within. No medical system, medical doctor, psychologist, psychotherapist or other professional can do what you yourself can do: to observe and understand yourself emotionally and feel what your innermost self needs and longs for.

As such the Fusion Practices are a primary practice to reconnect with the somatic basis of our being and reconnect with the hidden parts of ourselves. Thus the Fusion Practices are exercises in in kinesthetic awareness through which we recover our primary form of knowledge which is somatic and is buried in our very body and our gut. For this we need to go deep down into our own emotional history and learn to feel with and within our entire body.

For this we need to let our rationalizing mind deeply sink into our body and let our intuition take over so that we get in touch with the visceral level of our being. Thus the Fusion Practices are not mental or intellectual exercises to clear our debris or accumulate virtues but an existential transformational exercise to which the formal mind only becomes an impediment.

The Fusion Practices provide a unique way to overcome the dependencies you have grown accustomed to and to train yourself in self-reliance and recover your independence and thereby your self-esteem.

Healing Wounds

Often the question is asked: “Where to begin”? Such questions may suggest a hesitation to enter the practice and face oneself. A path opens at the very moment we start to walk. The inner journey is different for all of us, as our histories are unique. The formulas provide an essential guide, which tells us how to go about, which direction to take, when to walk, when to rest and the stations (practices and their sequences) we should visit.

As suggested earlier, the whole Taoist practice has its beginning in the practice of the Inner Smile. The Inner Smile as the key to refresh, create, transform, attract and blend energies presupposes a basic willingness to enter into an adventure and to face the risks in transforming old to new. It requires above all a willingness to love oneself and self-compassion. Such an inner movement towards the opening of the heart is however only possible if there is some willingness to accept and forgive oneself.

Perhaps the primary key to any process of self-transformation is the willingness to forgive oneself, as it opens the heart and sets in motion and frees the flow of Chi, both from within and from outside. That would seem a primary condition for an open space to emerge in which one can enter into dialogue with the organs and let them speak and listen to what they wish to say.

Yet forgiveness is not an act for once and for all but a primary key in a continuous process of turning inward and practicing inner observation. It is not surprising that when Jesus was asked a question on forgiveness, he pointed to its key role by saying that a man should forgive seventy times seven, meaning that it has no limit. Rather than perceiving forgiveness as a virtue which make all other virtues possible, like love and compassion, it is more useful to see it in terms of a growth generated by process towards balance as a whole between positive and negative emotions (Fusion of the Five Elements) and as the fruit of the creative cycle (Cosmic Fusion).

The refusal or inability to forgive and thereby to love oneself and others may be said to affect all the organs, affecting their chemistry and inhibiting cooperation and balance. Frequently it is rooted in a victim consciousness, which sets into motion negative emotions which tend to reinforce each other. This victim consciousness often has its roots in a presumption of betrayal as a result of which the heart closes, as others are made accountable for one’s own negativity and suffering.

As a result there is a refusal to even consider looking inward, as the source of misfortune is placed outward. Thus the heart's natural tendency towards love and compassion is obstructed and the wound is cultivated to legitimize one's withdrawal from life, as a result of which both the internal as well as the inward flow from outside are being impeded. All traumas first of all settle deeply in our organs from where they affect our energy flows. They can only be dissolved with patience and gentleness towards oneself. Only when internally energy is accumulating, can it flow outward.

The strengthening of the organs have in the Taoist tradition a vital role to play in mitigating and achieving control over the negative emotions. Forgiving does not mean that the wound needs to be closed. It may remain but it is no more active in generating toxins and preventing positive emotions from growing. Also forgiving does not means that one forgets. That may be too difficult and is not necessary. Only the charge, which activates to toxins needs to be defused.

The aim of the Fusion Practices also is not to eliminate pain. Pain and joy are intrinsic in life and form a unity of opposites. Attempts to eliminate it create frustration and exasperation and may well make it grow and become more difficult to control, as negative emotions such as fear, anger, worry, sadness and impatience multiply. The more negative emotions are suppressed, the more they have a tendency to run wild.

The purpose of the Fusion Practices is to achieve control and regulate them consciously so that major imbalances can be prevented and the quality of the energies, upon being fused and blended is heightened and transformations into new states of being (the energy body and the spirit body) become possible. Ups and downs are inevitable but as you grow able to accept them and balance them and grow in equanimity, a point is coming when you are less affected by what happens around you; you are less carried away by events and you become less vulnerable.

Rather than to do away with toxins, it is wiser to allow for them and it is easier to control them. This immunity is both raised in balancing positive and negative emotions (Fusion of the Five Elements) and by strengthening the positive emotions (Cosmic Fusion). The unity of the two is of great importance. To make a small tree grow one does not only weed, but also creates optimal conditions for it to grow: that it is planted in good soil, that it receives shade and water and that it is pruned at the right time.

One practical way to effectively deal with the negative emotions is to distinguish between them, so that they can be separated and sorted so they receive the attention they require and they become manageable. We may see them like our children who each have their specific needs. We can give to each one the particular attention they deserve. To deal with them all at once is too difficult and leads easily to discouragement and a sense of powerlessness and defeat. In such a way you more easily work with them.

For each of us, the relations among and between positive and negative emotions is a unique one. So are our needs and styles to deal with them, depending on the culture we have grown up in, our different kinds of education and the ways we have responded to them and have evolved. So we will need to rely on and develop our own imaginativeness to use the formulas creatively so that they may be helpful rather than become an obstacle, as we apply them mechanically. We need to learn to feel what is good for us, and what helps us to be centered, in balance and at peace with ourselves.

Practice of Humor and the Humor of Practice

Crucial to our practice is a sense of humor and playfulness: to avoid the Fusion Practice becoming a reproduction of old patterns in a new form, as we use it as a device to track down our weaknesses in a spirit of self-condemnation. Humor is essential for balancing the energies of and in our organs. Humor is the great regulator which helps us to look with a sense of relativity, playfulness and lightness at ourselves. It is an expression of the compassionate energy of the heart which helps to look with a sense of detachment at ourselves and not to take ourselves so seriously that it becomes difficult to relax and be a witness rather than a severe judge.

With little humor, the Fusion Practice would become an accounting practice in which we record the drawbacks which prevent us from advancing. In focusing on the negative as if it were evil, we lose sight of our creative potential and prevent ourselves to gradually free ourselves from old patterns. Without humor we see no way out except by relying on some external power or force which only further undermines our own original power of self-transformation.

Learning to tend our Garden

What helps is to realize that we do not have negative emotions, per se, but that we and our spirit are in them. They are an integral part of us, and when we want to eradicate them, we destroy ourselves and the spirit contained in the organs.

For the Fusion Practices to become effective, we need to learn to dialogue with our organs. We must learn to express our gratitude to them for the miracle that we are alive thanks to their unceasing work. We need to give them a chance to relax and recuperate. Let us give them a chance to get in touch with each other. They are both our parents and our children. They can store Chi and provide us with energy when we need it. Let us protect them from our anger and other negative emotions, so that they do not have to suffer and feel we drain them. Let them breathe freely, accumulate Chi, expel the accumulated toxins to that we can recover, nourish and sustain our health and vitality.

It is a challenge to see the positive in the negative; negative emotions, when freed from their destructive and explosive charge, contain pure energy which we badly need to develop our vital force.

It helps to accept ourselves if we remember that our negative emotions often mirror forms of resistance which at some time early in life, served as a form of legitimate and understandable protection and rebellion. At that moment they might have been necessary, but at some point they lose their function.

Negative energies may be seen as a kind of garbage. If we discharge it on others and the environment, it becomes a major factor in generating toxins, bacteria, viruses and a variety of forms of disease by which negative emotions are multiplied and become more and more difficult to control. Or they can be seen as highly positive and urgently needed for recycling and composting, so that they can be used to help us grow and improve our balance, so that we do not become a burden and a threat for others and our environment, but a blessing, as our energies contribute to a climate of harmony and peace. The more light and Chi that enters our body and the more we radiate it into others and our surroundings, the less bacteria and viruses have a chance to grow.

The Fusion Practice (Fusion of the Five Elements, Formula Four) provides an effective method for composting, as the negative energies from the different organs, after having been separated out in the collection points, are brought into the pakuas which act as transformers. They can be blended with the positive energies from the organs as well as the energies from the universe, so that they become purified and enriched; energies of a higher quality and order can be generated. These then may be condensed into a pearl as the embryo of our energy body (Fusion of the Five Elements, Formula 5).

Raising the Quality of Our Energies

The Cosmic Fusion Practices basically serve to consolidate, stabilize and enhance the process towards balance, initiated in the Fusion of the Five Elements practices. The creative cycle practice of Cosmic Fusion specifically serves to further the growth of positive energies in the organs and glands, which in turn will be able to mutually support each other to create a higher quality of energy. This energy is concentrated in the compassion energy from which a pearl is formed. Only when all the positive energies merge into the fire of love and compassion by which the sexual energy is steamed and is transformed, its energy will turn into a higher creative and spiritual energy. This higher energy will then serve to nourish the energy body.

This pearl will be more powerful in healing and cleansing the body than the pearl in the Cosmic Fusion practice and may subsequently serve to create and nourish the energy body. The energy body may in the highest practices give birth to the spirit body. The Cosmic Fusion practice also serves to intensify the process of cleansing and healing through the creation of the three Thrusting Channels and by creating the the nine Belt Channels which serve to protect the body and its energies from negative influences.

Marriage of the Light from Within and from Above

All Universal Tao practices, including the Fusion Practices, consist of two dimensions. One is the activity which focuses on the processes of transformation within the body through cleansing, opening, healing and fortifying the body and raising the quality of its energies. The other activity is to get in touch with the universe, the cosmos and nature and attract their energies and the Light in which they originate towards the body so that the body may process these energies and blend them with its own refined energies.

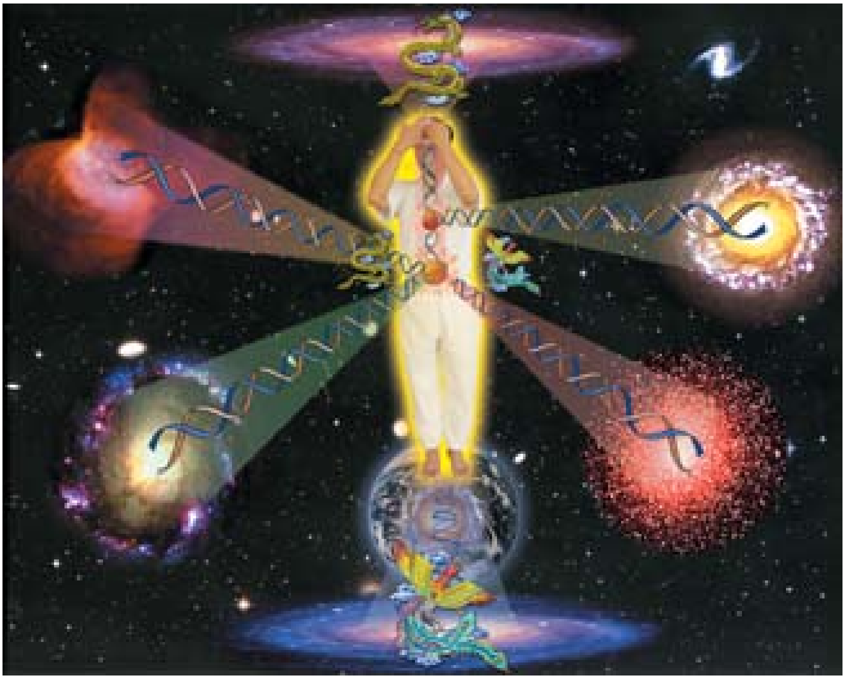

Fig. 9.10 *Connecting with the Universe*

The Taoist view is that the body by its very nature and structure is a microcosmos which is able to attract and receive and use the energies from the universe, the cosmos and nature to transform itself. Quantum physics confirms the view that the mind is faster than light and that consciousness is a form of energies which can move other energies when tuning into these energies and their frequencies.

The process of getting in touch with the universe and expanding love and compassion begins in the same way as the approach to inward transformation of energy: with smiling, feeling and sensing and radiating out love energy. The cleaner and finer the energy which is emitted, the higher the response from the forces and energies from above, below and around us. The relationship with the universe has been likened with that of a love or courting relationship. In all the mystical traditions the experience of the divine is also experienced as a relationship between the lover and the beloved. Frequently such relationships have been expressed by the great mystics in the most erotic and sensuous language, a reason for which they were as a rule hardly accepted and frequently persecuted.

Through its meridians our body may be seen as a grid of light channels through which it is connected with the light grid of the universe and can tap its unlimited energies.

Learning to embrace ourselves and the universe:

The Fusion Practice may serve to visualize oneself as a network of light and as a light body. Then the practice is not only one of cleansing but also one of filling the meridians with light. Mutations in consciousness start in the body visualization and turn into actualization. But the transformation lies not in the visualization. It is only a vehicle. Actualization is the real experience of the energy transformation in the body, when its chemistry changes and energy is transformed. One begins to sense the truthfulness and miracle of being alive and one starts to feel that one is embraced and pervaded by love and a sense of gratitude for the mysterious ecstatic process taking place in the body arises.

But this is not an end term. It is always only a beginning, as the love of the infinite has no limits neither in time nor space. In the process matter turns into what is immaterial and the immaterial as energy to continue its transformation waits again for a new form to contain it and change its nature and composition. Words help to create an image and when the image is there, the word loses its function. The image serves to help us generate the energy and its transformation. When the energy has come into being, the image needs to be given up so that a new space arises for a new experience. This is an unending process. How wondrous is it that our organs, our body, that sacred vessel, as it is called in the Taoist Canon, remembers its origins and longs to be fused and serve as an instrument of the fusions which unceasingly take place in heaven so that our intrinsically divine nature is realized and can be experienced. For earthlings, the learning process starts from outside. In the process of turning inward, the learning process increasingly becomes internalized. We need the structures and the forms and formulas to help us on our path. Then when we become still, it becomes spontaneous and starts to move by itself.

Then our child has been born and just wants to play without end to feel that they are welcome and loved so that they feel at home. For this quality in us to arise, we need to learn to embrace ourselves and the universe, as a mother her newborn child.

Written by: Joost Kuitenbrouwer
Universal Tao Instuctor

Questions and Answers about the Cosmic Fusion Practice

Q. ***Is the cauldron the point where all the pakuas join together?***
A. Yes.

Q. ***Now that we've learned about the cauldron, where do we collect the energy at the end of meditation practice?***
A. You can collect and store energy at the cauldron or you can continue to collect and store your energy behind the navel. When you collect the energy you make a vortex, which is three-dimensional, so it can veer toward the navel or the cauldron. As you do this you may feel it going deeper and deeper.

Q. ***Do we always have to do 36 and 24 spirals in the collection?***
A. The numbers 36 and 24 are a guide. There is no speed limit for spiralling the energy when you collect it. Some days you might want to do more than that; on many days you may want to do much less. On those days, 12 and 9 may be sufficient. Eventually it will go by itself.

Q. ***How do you collect the energy when there is so much energy all around?***
A. Collecting the energy will eventually become automatic. The safety features are bringing any excess energy from the pelvis up to the navel and and excess energy in the head and heart down to the navel. When you are coming back from the feet to the navel in the belt routes you are also collecting the energy.

Q. ***Does the energy stay in the cauldron or does it spread to the organs?***
A. Both.

Q. ***Should you see the collection points in color?***
A. Some people see, some people feel, and some people can even taste the energies. If you are visual, great. Then you will see the colors. If you work with the colors, the color will tell you if an organ's energy is healthy or not. If you look within and see dark muddy colors, then you know you have to work to make them clear and bright.

This applies to the organs and the collection points. Because the organs are affected by the emotions the colors will be affected.

Q. ***Does the Yin/Yang symbol move inside the pakua?***
A. It spins like a vortex.

Q. ***Is there always a pearl you can work with?***
A. If you practice forming a pearl every day, there will be one there when you sit down to meditate.

Q. ***Is there a minimum size for the pearl?***
A. The size of the pearl is less important than its density. As you add energy to it you might notice it glowing more brightly or feeling more energetic.

When the energy increases you might want to make more than one pearl. Then you can give it away to an organ or into your own atmosphere.

Q. ***Does the belt channel go around both legs or each one separately?***
A. It circles around both legs, although if the energy wants to go around each one, that's okay too.

Q. ***You talk about using the thrusting routes for practical purposes. Can you give an example?***
A. To clean the atmosphere in a room, you shoot the pearl out the crown and let it expand and fill the space. That way you declare the territory as your own—it is permeated now with your energy and consciousness. If there is any negative energy, any negative spirit, it will just leave of its own accord. You can't destroy negative energy, but you can work with it to make it leave, by overcoming it with your own energy. But you don't want to anger another spirit, that's not the Taoist way. Either they are using your energy to protect them or they want to tap into yours to get themselves to a higher level.

Q. ***Is there a breathing technique that will make you more Yin or more Yang?***

A. It's not just breathing. If you eat very Yang foods, like meat and you are trying to be Yin, it won't work. If you work with the Fusion I formula you can balance the energy more easily than with outside forms of breathing or food. Fusion I and the Microcosmic Orbit are natural ways to balance your energy.

Q. ***The meditation seems so long.***

A. Once the channels are open you don't need to spend much time. The training period seems long, but the actual practice can be quite short. When the Thrusting Channels are open you won't need to do any muscle contractions, you won't need to do it in stages, you won't need to inhale several times. You will just sit down and in a couple of minutes you will find yourself in a state where the energy will be strong enough and the channels open enough that the energy will flow through the Thrusting Channels just using the mind itself. In fact, you will be able to do it anytime, not just during meditation.

You can imagine the work it takes to connect pipes from a reservoir to each household in a city. But once the pipe is connected all you have to do is turn the faucet on.

Q. ***What can close the channels?***

A. Stress of any kind can close the channels. Tension contracts everything—muscles as well as emotions. Also, if you stop practicing for a long time the channels will close. A path that's walked on every day will be clear, but if no one travels on it the weeds start growing and soon it is covered back over. That's why we stress the importance of daily practice. Even five minutes a day, is better than one hour once a week. Fusion I and Fusion II practice can be combined together, so you can check the organs, smile down, see what you need to do, and in a few minutes you can do some maintenance work. Then when you have more time you can go and clean the organs and the Thrusting Channels. In the beginning, the Fusion practice needs the mind to form the pakua, and collection points. It's like the Microcosmic Orbit: when you first learned it, it took a lot of concentration on each point and even when it started to run by itself it wasn't dependable. But after a while you could wake up in the morning and feel your Microcosmic Orbit running.

Fusion I and II are more complex and at the beginning you have to develop a different technique to get them running. I just look in the cauldron and it's like a slide projector: I can see everything. Then you can add the collecting of the energy and the blending process.

Everyone we know who has practiced Fusion II for a year or so, can do the creation cycle in a couple of minutes. As you practice and keep adding more details, sooner or later it becomes automatic. If you do Fusion II on a daily basis you will begin to get faster and faster. You will be able to look inside and see what's going on. When the process becomes automatic you open a whole area of possibilities. It will give you more time to practice because you will be doing it so fast. It will be there when you sink into sleep. It will be there when you wake up in the morning.

When you tune into your organs, your organs will have more say over what they want and what they do. So when you work with the formula correctly, the organs will love it. They will have the space and time to be themselves. The brain can be overworked and pushed too hard. It can force any organ to work until it's destroyed. You can see this in people who take drugs to control their emotions.

But when the situation is balanced by using the Microcosmic Orbit, the organs get a chance to balance themselves. The brain has a say but the organ also has a say. Then you can achieve a balance of energy when the intellect and the organs are working together. At this stage, when the senses are drawn in, not distracted and running around, you will feel calm. Gathering and conserving energy is good in itself.

The next thing that happens is that you can begin to gain Real Knowledge. Real Knowledge is beyond intuition; it comes from within and you make better choices and decisions. This Real Knowledge is an inner guide which sometimes manifests in dreams from the unconscious. It can show you a direction in life. You know yourself that these dreams are rare. In the past they only came when you were at a crossroads and there was some danger. But through the practice with emotions, organs and mind, you develop that interior communication that allows you access to Real Knowledge, so you're not operating in a state of emergency but from a state of balance and calm. Those directions will be easy to follow. So this is a process of purification. The energies will be used with purer intentions.

For example, if you have an excess of fear you feel threatened and isolate yourself. Any decision you make from that standpoint will not lead to happiness. This formula is very important and carries a tremendous amount of responsibility, for whenever you learn something that can change your life in a positive sense you will either do it or forget about it. You often hear people say, “I want to change my karma.” This formula is something that can change your karma. So this practice, this opportunity has a responsibility. You will have to take the consequences of doing it or not doing it. Because if you do it you will change.

Q. ***How can I protect myself from negative energy?***

A. The belt routes can protect you. When you circulate the energy through the belt routes, the energy centers are strengthened, and can protect you from absorbing outside energy that you don’t want. For instance, in a crowd, if there is energy you don’t like, just run the belt routes to protect yourself. This will also protect your aura. If you are upset and have a lot of emotions, the aura is very unstable. The aura might expand or shrink. This protection will become automatic through diligent practice of the belt routes.

Q. ***When you send the energy out like a shower, are you always collecting positive energy? How do we know what energy we are collecting?***

A. There is always heavenly energy around you, and there is always magnetism so when you do this work you are working with earth and heaven energy. It’s another responsibility when working with the energy. If you are an ordinary person you are absorbing only a certain level of energy. When you start raising your level, you are raising your level of positive energy. This can change the atmosphere around you (an inner pressure of positive energy). Energy has a magnetic quality that attracts. When you have strong positive pure energy you will be repelling negative energy and you won’t be affected by negative energy. For example, if someone is angry you will have enough positive energy to handle it. If you practice this and go to higher levels and then decide to stop, your pressure will again change and decrease. Our bodies always try to find a balance, that's our inner intelligence. So the pressure has dropped and the body tries to restore the familiar higher pressure. To do this you start absorbing anything, and the most abundant source is the garbage energy: anger, fear, depression, etc. As you build up your strength you need to maintain that higher level and that is a big responsibility.

So when you do the shower of energy you have to depend on the strength of the energy in your pearl. Fusion II jacks up the energy of the pearl you made in Fusion I by imbuing it with the positive qualities in the organs. So it's important to do Fusion I carefully.

When you absorb positive energy you can even affect the bacteria in the air. If you've been practicing the Microcosmic Orbit and Fusion I for a year you've probably noticed that colds don't last as long, and you don't get sick as often.

Q. ***Is it normal to feel pain or discomfort when doing the fusion practices?***

A. You can see that some systems are into suffering: staring at the sun, sticking oneself with needles or standing on one leg for a long time. This is so that when practitioners get to the level of great power, they can withstand the pain. This is how some people learn and they suffer a great deal.

Maybe you cannot take all the energy; some people bathe in the river every day to reduce the heat and can't sleep for weeks and months. When you practice to a certain level there are changes in the body—but you can take those changes. Sometimes as new energy pushes through channels you may feel a headache. Some people feel a stomach cramp when they blend the energy at the cauldron. But that is unusual. It's like trying to help a butterfly when it's hatching. Maybe you want to help by cutting the chrysalis with a scissors. But the butterfly won't survive if you do that. It has to struggle at the beginning in order to survive.

So you will have to grow over the pain. Other systems induce the suffering first. In Chinese, they call it qing ku—you don't have anything and you still have to suffer—but that is not the Taoist way. We prefer to grow first and then experience the growing pains. You may feel imbalanced from time to time—constipation is a real problem. If you have a good bowel movement at least once a day, you will not suffer so much.

Now that you are more sensitive to energy, you will recognize negative energy right away and you have the tools to neutralize and purify the energy.

Some people, after they study, find that they are reacting differently to their love partners. For some of them, this difference leads to a break-up. When they look back on it, they often realize that the break-up was inevitable and that they were dissatisfied before. For others, especially couples who study together, their relationship becomes more solid and more respectful. In any case, it means that you begin to understand yourself better.

Sometimes you begin to see problems more clearly and see the energy that causes the problems. When you can enhance the energy inside, the problem goes away. There are those who get angry but don't realize they are angry and aren't aware of the effect of the anger on the liver. But those who practice Fusion I can become aware of the liver and very quickly the anger will be gone.

The Taoist system can change the course of your life, change your astrology, change your fortune and your fate. We know that because of the experiences of Taoists over many generations.

The energy inside your organs and brain makes your life what it is. So when you know how to control your energy, your organs and senses, that is the beginning of controlling your future. Those who are controlled by substances like tobacco, alcohol and other drugs are not in control of their lives.

In Taoism you don't run away from problems as long as you are in this world. Actually there is practically nowhere on the planet where you can go without being bothered.

When we started to teach Westerners, we found new ways to teach people how to balance their energy through diet as well as through the Taoist exercises. Those who eat only raw foods will be too Yin and those who eat mostly meat and dairy will be too Yang. But the only way to restore your True Knowledge is when you turn to your internal organs and senses for information. Then you will know how to conduct your life. But until you do that it's hard to know what's good or bad for you.

The golden mean, the middle, moderation are all ways of saying that when you know how nature works you will know the Tao. By knowing the Tao you start to know about God and the universe. Anything that's extreme, you have to back away from. The only thing that lasts long is balanced, middle and center.

Q. ***Is there a problem if I shake a lot during meditation?***

A. Too much shaking can scatter the energy. Moving and stillness have to be balanced. If you remain still for too long the energy might settle. But if you shake and scream and are never still you risk eventually shaking out all your life force.

If you start burning a fire, it gets hotter and hotter. When the water comes to a boiling point you turn it down and cook it very slowly. That way you can get the essence out. If you don't turn down the light all the water will boil away and the pot will be burnt or broken.

When you are into the meditation and very still, movement starts and many things are activated. If the movement becomes very strong it won't last long and you might want to stop it. You have to know the timing. If you start to shake, you can gain power by stopping the shaking and utilizing that energy inside. If you let the shaking continue, you might lose all the energy you built up.

In some forms of therapy people are encouraged to shake and jump and scream to get rid of their anger or depression or anxiety. When the session is over it seems that everything has been worked out but when they get home they are throroughly exhausted. The emotions were shaken out along with the life force. How much better to recycle the energy.

Tai Chi balances the Yin and Yang through movement and stillness. This builds up energy and then you can circulate the energy in the Microcosmic Orbit. The whole secret in the Universal Tao is having the channel for the energy to flow.

Q. ***Why do you create an Energy Body outside yourself?***

A. When you set up the Energy Body and the spirit body they start to draw out the essence of the physical body. Then it starts to work in space.

It's the same reason that every country wants to set up laboratories in space. You can make a lot of things better in space because there's no gravity, no dust and no germs.

The Energy Body will draw the essence up; you might even feel like it's going to pull your whole body up. And then the Energy Body draws the energy from space and the universe and mixes them in. When you bring the energy back into the body the organs will absorb all this energy and start to transform. But if you sit down without a purpose you won't get this effect.

Everytime you add more of your body's energy to the Energy Body, the Energy Body will go higher. Sometimes I feel like all my essence leaves my physical body. When you practice every day it starts to take the essence of this body and transform and mix it with the universal energy and your progress will be remarkable. There will be no resistance. When you're finished meditating, you just shrink the Energy Body down and each time you shrink it down the pearl becomes more refined.

At one time, the Taoist Masters were afraid of losing the pearl in space and did it all inside the body. This was reflected in the art in China where they worked incredible detail on miniature objects. But later the Tao masters discovered that if they know how to send the energy out the head they can expand the Energy Body and then shrink it down. That way they achieved more perfection—it's the same way they make a computer chip.

If you look inside a chip there's a whole city inside it. They draw every detail on paper the size of a football field and then with a special camera they shrink it down and down.

Set it up, do all the detail you want, then shrink it down and pull it back in.

> I had to figure a lot of this out myself. When I heard "transfer your consciousness to the next life" I decided I had to learn how to do this in this lifetime. I kept asking, but everyone had a different answer. They just said it's very simple. I told one master I'd give him anything for the secret; I'd go anywhere to learn. Either he didn't have it or else he wouldn't give it to me. After a while I discovered that I'd have to figure out myself how to transfer whatever I wanted to the Energy Body. I did this because I understood that my Energy Body will live forever.

Focusing on intellectual understanding alone is limiting and stifling. A real practice teaches one how to feel within and derive authentic knowledge based on personal experience. In the Universal Tao, we emphasize the practice of understanding nature, getting in touch with the natural forces, and getting in touch with oneself. Then the principles of philosophy will reveal themselves, and one can have real hope based on what is, not on empty promises that can never be proven.

Bibliography

Beinfield, Harriet, L Ac and Efrem Korngold L Ac, *Between Heaven and Earth,* NY: Ballantine Books, 1991.

Grazzaniga, Michael S, *Nature's Mind*, NY: Basic Books, 1992.

Grings, William and Michael Dawson, *Emotions and Bodily Responses: A Psychophysiological Approach*, New York Academic Press. 1978.

Haas, Elson, MD, *Staying Healthy With the Seasons*, Berkeley, CA: Celestial Arts, 1981.

Hammer, Leon, MD, *Dragon Rises, Red Bird Flies*, Barrytown, NY: Station Hill Press, 1990.

Herrmann, Ned, "How the Brain Works...and how to help it work better", *Bottom Line Personal*, November 15, 1992. p 9-10.

Locke, Steven MD, "*How To Use Your Mind To Improve Your Health", Bottom Line Personal*, August 30, 1992. p 11-12.

Melchizedek, Drunvalo, *Ancient Secret of the Flower Life* Vol. I & II, Light Technology Publishing, 1998 & 2000.

Porush, David, "*Finding God in the Three Pound Universe: The Neuroscience of Transcendence*", *Omni*. October, 1993. p 60-70.

Wagner, Bruce, "*You Only Live Twice*", *Details*, March, 1994, p.166-171, 213-319.

Tortora, Gerard J. and Sandra Rey, *Introduction to the Human Body*, Fifth Edition, NY: John Wiley & Sons, Inc., 2001.

Tortora, Gerard J. and Sandra Rey, *Tortora Grabowski*, Ninth Edition, NY: McGraw-Hill, 1999.

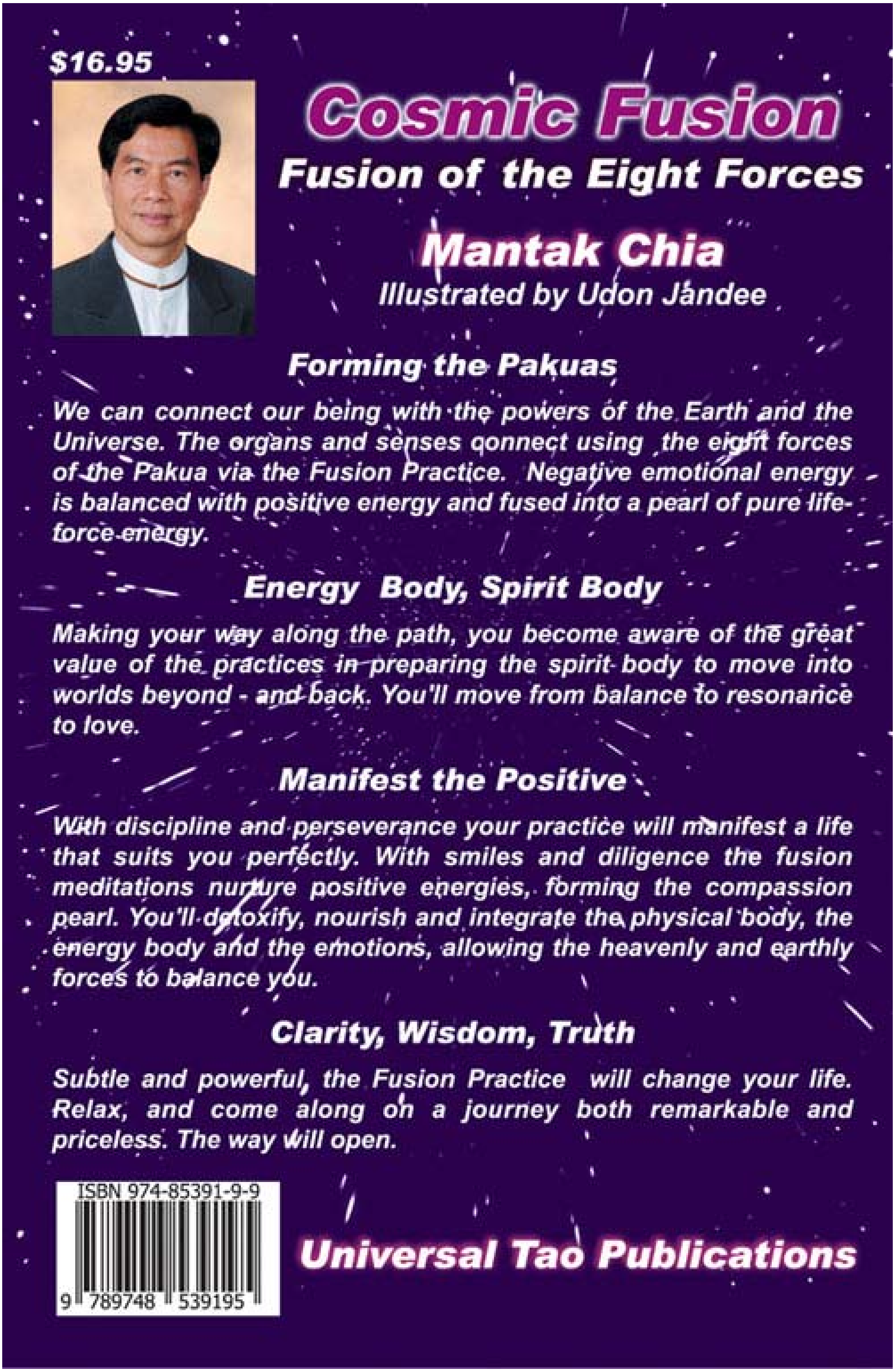

$16.95

Cosmic Fusion

Fusion of the Eight Forces

Mantak Chia

Illustrated by Udon Jandee

Forming the Pakuas

We can connect our being with the powers of the Earth and the Universe. The organs and senses connect using the eight forces of the Pakua via the Fusion Practice. Negative emotional energy is balanced with positive energy and fused into a pearl of pure life-force energy.

Energy Body, Spirit Body

Making your way along the path, you become aware of the great value of the practices in preparing the spirit body to move into worlds beyond - and back. You'll move from balance to resonance to love.

Manifest the Positive

With discipline and perseverance your practice will manifest a life that suits you perfectly. With smiles and diligence the fusion meditations nurture positive energies, forming the compassion pearl. You'll detoxify, nourish and integrate the physical body, the energy body and the emotions, allowing the heavenly and earthly forces to balance you.

Clarity, Wisdom, Truth

Subtle and powerful, the Fusion Practice will change your life. Relax, and come along on a journey both remarkable and priceless. The way will open.

ISBN 974-85391-9-9
9 789748 539195

Universal Tao Publications

THE KATHARA™

Bio-Spiritual Healing System

LEVEL 1 CERTIFICATE PROGRAM

by A'sha-yana Deane

The Kathara™ Bio-Spiritual Healing System
is supplied by the Azurite Press of the Melchizedek Cloister Emerald Order
All Materials ©1999-2002 inc. A'sha-yana Deane, Eckr MCEO All Rights Reserved

KATHARA

BIO-SPIRITUAL HEALING SYSTEM

CERTIFICATE PROGRAM

The Maharic Recoding Process

LEVEL - 1 SESSION MANUAL

FOR FACILITATOR-INSTRUCTORS

FOUNDATIONS OF KATHARA HEALING™
KATHARA HEALING ™ FACILITATION
THE MAHARIC RECODING PROCESS ™

An advanced *Spiritual Healing Certificate Program* based upon KEYLONTIC MORPHOGENETIC SCIENCE. Sponsored by the **Kathara Alliance**–an associate of the Azurite Press World-wide Consulate. All Materials © 1/2000, 2001, 2002 Ashayana Deane, Eckr. MCEO–All Rights Reserved

Table of Contents

Chapter 1: THE CONTEXT OF HEALING - 7- 47
Structure of the Unified Field and the Multi-dimensional Anatomy of Form
Chapter Highlights 37

Chapter 2: THE ELEMENTS OF HEALING ....48 - 112
The Human Body Kathara and the Multi-dimensional Anatomy of the Human Form
Chapter Highlights..102

Chapter 3: KATHARA HEALING BEGINNING APPLICATIONS 113-165
Creating Field Integrity and Building the Knowledge Base
Chapter Highlights..149

Chapter 4: GENERAL KATHARA HEALING FIELD WORK 166-211
Skills to Facilitate Healing
Chapter Highlights 204

Topics and Techniques Locator Index

Chapter 1: The Context of Healing ....Pages 7-47

Structure of the Unified Field and the Multi-dimensional Anatomy of Form...8-9

* The Holographic Template..10-14

* The Structure of Dimensions, Partiki Phasing and Matter Density 15-25

* Universal and Planetary Kathara Grids, the 12-Tree, Kathara Centers and Lines, Scalar Shields, Crystal Seals and Diodic Points 26-31

* The Universal and Internal Templar Complexes...32-34

* The Transduction Sequence and Holographic Reality Construction................................35-36

Highlights...37-47

Chapter 2: The Elements of Healing ....Pages 48-112

Human Body Kathara: Multi-dimensional Anatomy of the Human Form... ...49-50

* The Maharic Shield and the Imprint for Health...51-52

* The Planetary Bio-feed Interface System (PBIS) 53-54

* The Hierophant Symbol Code, PBIS, the Maharic Shield...55-57

* Morphogenetic Expansion, Growth Cycles, Death, Molecular Compaction............................58-62

* Levels of Human Anatomy, Embodied Kathara 12-Tree Grid, Crystal Shields, Seals, Signets, Hova Bodies, Auric Capsules, Diodic Points, Seed Seals, Chakras, Axiom Lines, the Hara Level, DNA and Merkaba 63-93

* The 12 Human Senses, Apparthi and Human Perception...94-101

Highlights...102-112

Chapter 3: Kathara Healing Beginning Applications .... Pages 113-165

Creating Field Integrity and Building the Knowledge Base

* Tapping the Inner Mentor and Awakening the 6th and 7th Senses.................................114-120

Technique-1: Awakening the Mentor and the 6th and 7th Senses 117-120

* Morphogenetic Repatterning, Frequency Sub-Harmonics, Maharic Seal............................121-125

* Kathara Points and Kathara Point Sequences..126-131

***Sequence-1* Kathara Points for Regeneration 128-129**

***Sequence-2* Kathra Points for Revitalization 130-131**

Technique-2: The Maharic Seal and the Liquid Light Cleanse 132-137

* Opening the PBIS and the Maharic Seal, the Structure of Multi-dimensional Identity and the 4 Primary Triadic Healing Currents..138-148

Highlights 149-165

Chapter 4: General Kathara Healing Field Work ....Pages 166-211

Skills to Facilitate Healing

* Opening the Healing Channels-Doradic-Phase-1 Current, the Doradic Sphere, Doradic Cords and Doradic Balls..167-174

Technique-3: Opening the Healing Channels-Doradic-Phase-1 Current.......175-177

* The Karmic-Miasmic Imprint Incarnational Time Vectors, Vector Imbedding.........................178-180

* Multi-Vector Holographic Recoding through the Maharic Infusion...............................181-183

Technique-4: Multi-Vector Holographic Recoding Kathara Scan..................184-186

Technique-5: The Maharic Infusion 187-190

* Attitudes and Etiquette in Healing, the Motivation to Heal, Postures of Love, the Framework for Healing, and Fully Engaging the Love Vibration 191-197

Exercise-Client Greeting: The Mahara-Space Comfort Touch 198-199

* The 6-Step Maharic Recoding Session, Melchizedek Cloister Ordination and Becoming a Professional Spiritual Healing Facilitator-Instructor...200-203

6-Step Client Maharic Recoding Session: and Client Agreement Form**...........202-203**

Highlights 204-211

Introduction to Kathara Level 2 Mentor – Healing Applications
MERKABIC RADIAL BODY HEALING

Topics and Techniques Locator Index

Chapter – 5: Mahara-Merkabic Master Key Healing – Manifesting the Blueprint for Health, Primal Currents, the DNA sub-Strand Matrix, Merkaba, Psonns, the Trinity Keys and Photo-Dynamic Manifesting

- Universal Life Force Currents .... 214 – 220
- DNA, Fire Letters, the Tribal Shield, Cue Zones .... 221 – 228
- **Technique – 6: The Maharic Quick Seal** 226
- **Technique – 7: Ps – 13 Tribal Shield Activation** .... 227 – 228
- Merkaba Activation, Solar & Lunar Salutations .... 229 – 237
- **Technique – 8: Merkaba Salutation** .... 235 – 236
- DNA, Khundaray Activation, Emerald & Amethyst Awakenings .... 237 – 241
- **Technique – 12: Kee-Ra-ShA Activation Sequence** .... 242 – 247
- Universal Trinity Keys & Merkaba Key Codes 248
- **Technique – 9: Merkaba Trinity Key Induction** 249
- Merkabic Healing .... 251 – 255
- Introduction to the Radial Body Trion-Meajhe Field .... 256 – 257
- The Radial Body Transduction Sequence .... 258 – 262
- Radial Body Healing .... 263 – 266
- The Radial Body and the Merkaba Vehicle .... 267 – 272
- Radial Body Healing and Veca Codes .... 273 – 276
- **Technique – 10: Restoring the Trion Field-Photo Radionic Veca Healing** 277 – 278
- **Technique – 11: Restoring the Meajhe Field-Photosonic Eckasha Healing** 279 – 280
- **Technique – 13: The Song of Lyra, Christos Invocation** .... 281 – 282

Diagrams Locator Index

- **Universal Life Force Currents & the 15 Dimensional Time Matrix** 214
- **Universal Life Force Currents & the Embodied Kathara Grid** 215
- **Universal Life Force Currents, Crystal Seals & Chakras** 216
- **Universal Life Force Currents, Hova Bodies & Auric Levels** 217
- **Universal Life Force Currents, Shields, Signets & Kathara** 218
- **Primal Currents, Silicate Matrix DNA & Subtle Body Anatomy** 219
- **Universal Life Force Currents & DNA Fire Letter Sequences** 220
- **Silicate Matrix DNA & Fire Letters** 222
- **The Tribal Shield Species & Divine Blueprint** 224
- **The Natural Nethra Phase Density 1 Merkaba** 233
- **Merkaba Phases Review** 234
- **The Ar-iE-a Khum-Nar'A** 236
- **The Trinity Master Key Selection Chart** 250
- **Universal Merkaba Fields** 255
- **The Radial Body & Manifestation Mechanics** 258
- **The Eiron Point** 261
- **The Radial Body in Creation of the Hologram** 262
- **The 5 Radial Body Levels** 265
- **Transduction Sequence of the Miasmic Body** 266
- **The Christos-Trion-Meajhe Fields, High Veca Codes & Universal Life Force** 276

Diagrams

Chapter 1

The Holographic Template.......... 14
Vibration-Oscillation Correspondences 18
One 15-Dimensional Time Matrix 22
Frequency Bands and Sub-Frequency Bands Within a 15-Dimensional Scale... 23
Partiki Phasing, Harmonics of Manifestation and Matter Density 25
Structure of the Kathara Level-1 12-Tree Grid 27
The Primal Manifestation Template 28
The Universal Templar Complex 34
Transduction Sequence 35

Chapter 2

The Planetary Bio-Feed Interface System (PBIS) 54
The Hierophant Symbol Code 57
Growth Rates and Morphogenetic Expansion 60
Level-1 Kathara 12-Tree Grid Internal Templar 64
Shields, Signets and the Kathara Grid 70
Level-2 Kathara Crystal Seals Grid and Chakras 71
Auric Level, Chakra and Hova Body Correspondences 75
Primary Colors of Chakras 76
The Hara Level 86
12 Primary Axi-A-Tonal Lines 87
Silicate Matrix 12-Strand DNA Template 92
Merkaba Phases 93
The 12 Human Senses 101

Chapter 3

Awakening the Mentor 120
Frequency Sub-Harmonics in the 15-Dimensional Scale 123
The Maharic Seal 136
Liquid Light Cleanse 137
The 5 Hova Bodies 143
Stations of Identity and Levels of Mind in the Hova Bodies........ 144
The 5 Hova Bodies and Levels of Identity in the Time Matrix 145
4 Primary Triadic Currents and Hova Body Correspondences 147
Combined Triadic Currents 148

Chapter 4

Elements of Doradic-Phase-1 Current Activation 173
Intrinsic Process of Activating a Triadic Phase Current 174
Opening the Healing Channels 177
Maharic Recoding Process Elements 183
Amplifying Inner Audio-Visual and Kathara Scan 186
The Maharic Infusion-Client Facilitation 190
Postures of Love 197
Client Agreement Form 202
The 6-Step Maharic Recoding Session 203

INTRODUCTION TO KATHARA BIO-REGENESIS TECHNOLOGIES

COLOR, SYMBOL, SOUND AND BIO-REGENESIS TECHNOLOGIES

BIO-REGENESIS TECHNOLOGIES are tools to assist us in accelerating our expansion of consciousness through natural, gentle stimulation of the personal **DNA TEMPLATE, KUNDALINI ENERGIES, CHAKRAS AND MERKABA VEHICLE.** Bio-Regenesis techniques are built upon the foundations of **KEYLONTIC MORPHOGENETIC SCIENCE** and advanced **SCALAR MECHANICS,** as understood by races of higher evolution; these are paradigms of science not yet recognized or validated by contemporary earth science paradigms. Because Bio-Regenesis technologies are based upon what is viewed by modern earth science paradigms as a speculative rather than verifiable science, I present Bio-Regenesis techniques only as an option to explore in meditation.

Bio-Regenesis technologies were taught as "Common Knowledge" within the Ascension Schools of Pre-Ancient advanced human cultures and were utilized as standard practice in pre-ancient time periods. Bio-Regenesis Technologies are built upon the natural laws of scalar wave mechanics. Bio-Regenesis Techniques are **SPECIFIC APPLICATIONS OF CONSCIOUS ENERGY DIRECTION WITHIN THE MANIFESTATION TEMPLATE OF THE BODY** that stimulate into activity dormant portions of the personal Manifestation Template, which allows for a natural, progressive and accelerated advancement of the personal **ORGANIC EVOLUTIONARY BLUEPRINT** of the 12th-Dimensional Omni-polar Pre-matter Template. Bio-Regenesis Techniques simultaneously create subtle, natural acceleration within the interwoven energy systems of the DNA Template, Chakra System, Merkaba Fields, Kundalini energies, higher dimensional consciousness, and "Subtle Energy Body" anatomy.

In the Bio-Regenesis paradigm it is said that such techniques allow an individual to affect the function of the scalar wave template of the body, working co-creatively **under the direction of the individual's own higher dimensional levels of consciousness.** As the higher dimensional levels of the personal identity, the spiritual identity of the Soul, governs the processes of the personal Manifestation Template, Bio-Regenesis Technologies are completely safe and natural to the organic design of the human being. Using processes of Bio-Regenesis from the earthly conscious perspective expedites the natural processes of energy and consciousness by which the higher identity and earthly identity integrate to become one. In advanced ancient human cultures, Bio-Regenesis Technologies represented a natural attribute of consciousness expansion into the realms of higher evolution and assisted to biologically prepare humans for **STAR GATE** passage or **"ASCENSION".**

If we can understand that the **HUMAN MIND** is an energy force that **CONTINUALLY GENERATES PATTERNS OF SCALAR WAVES** through the process of **THOUGHT,** it is not difficult to understand how **APPROPRIATELY DIRECTED THOUGHT** can directly influence the function of the personal Manifestation Template scalar blueprint. All thoughts serve this purpose, and directly effect the observable state of the mind-body-spirit system and the manifestation of events within the external life drama. **THOUGHTS ARE THINGS- *SCALAR WAVE CONFIGURATIONS*** of multi-dimensional vibrating patterns of **BI-POLAR ELECTROMAGNETIC ENERGY RADIATION,** which form **SPECIFIC PATTERNS OF SCALAR FREQUENCY** within the personal Manifestation Template.

Undirected thought creates chaotic patterns within the Manifestation Template, while thought directed through clear intention creates ordered patterns within the Manifestation Template. Thought directed with clear intention and **KNOWLEDGE OF THE STRUCTURE OF SUBTLE-BODY ANATOMY,** is a very potent form of **SCALAR POWER.** The patterns of scalar energy frequency that thought projects into the personal Manifestation Template serve as **ELECTROMAGNETIC OPERATIONAL INSTRUCTIONS** within the Template, directly effecting the conditions that will be met in manifest experience. To develop skill in the manifestation process, the conscious "intending" mind must become more familiar with the **"LANGUAGE OF SCALAR WAVES"**, as this is the **"LANGUAGE OF LIGHT AND SOUND"** upon which the Manifestation Template operates. The "language of scalar waves" is "spoken" through the forms of specific wave spectra, forms that appear to the conscious 3-dimensional mind as frequencies of **SOUND**, spectra of LIGHT or **COLOR**, and shapes or **SYMBOLS**.

What we perceive as color is specific spectra of **LIGHT-RADIATION WAVES** that have a specific measurable **WAVELENGTH** and vibrational **ENERGY SIGNATURE**. In using the conscious mind to direct quantities of color, we are **CONSCIOUSLY SELECTING SPECIFIC SCALAR FREQUENCIES** to add to the Manifestation Template. When we use the conscious mind to create images of **SYMBOLS**, we are further **SPECIFYING THE ARRANGEMENT OF SPECIFIC FREQUENCIES**, as they will appear within the Manifestation Template. **SYMBOL FORMS** (even those that compose the alphabets of outer languages) represent **SCALAR-WAVE GUIDES**, and directly affect the contours of the scalar wave blueprint of the Manifestation Template. The Manifestation Template governs the function of the body and consciousness within the manifest experience, and so the **"Language of Scalar Waves"** is the medium through which we gain conscious access to the **CASUAL LEVEL OF MANIFESTATION**. In using **color and symbol** to direct frequency, we are creating a pattern of energy that has a specific vibrational signature or **"TONAL SOUND SIGNATURE"**, which directly interfaces with the energetic vibrational sound signature of the Manifestation Template's scalar grid.

In Bio-Regenesis technology, Color and Symbol are used to direct the **TONAL SCALAR ARRANGEMENTS** of the Manifestation Template, to create natural, desired results within the body-mind-spirit system. There are many complexities within the process of conscious manifestation that one must learn to fully master the contours of space, time and matter. The beginning step is learning to give more finely tuned thought intention **"PROGRAMMING INSTRUCTIONS"** to the personal Manifestation Template. Thoughts and images rendered in **external language** and 3-dimensionally associated symbols affect the **outer layers** of the Manifestation Template, but they do not provided as finely tuned manifestation instructions as do the **"scalar wave languages"** of Color and Symbol image, which direct the contours of vibrational patterns of SOUND. Mastery within the interdimensional spectrum progressively develops as one masters the core **"PROGRAMMING LANGUAGES"** of **Color, Symbol and Sound Vibration.**

In **MIND TECHNOLOGIES**, such as **CREATIVE VISUALIZATION** and **VERBAL AFFIRMATION**, beginning skill in **SCALAR TEMPLATE PROGRAMMING** is developed through using emotionally charged images and precise sound patterns in the language to which we are accustomed, as methods to begin gaining conscious control over the processes of personal manifestation. Keylontic Science **BIO-REGENESIS TECHNOLOGIES** take us **ONE STEP DEEPER** into the manifestation process, by directly employing the **LANGUAGE OF SCALAR WAVES** and directing the **ENCODED INSTRUCTIONS** through **PRECISE LOCATIONS** within the body-mind-spirit system and the Manifestation Template upon which this system is built. Bio-Regenesis technologies take us from speaking the language of the 3-dimensional mind into speaking the intrinsic scalar-wave language of the Spirit, which is the **CASUAL ELEMENT** behind and within all manifestation.

Bio-Regenesis Technologies represent applications of Scalar Mechanics that utilize the inherent power of **SUBTLE WAVE FORMS** to created desired change within the core Manifestation Template of the body. Within this process, **COLOR** and **SYMBOL** become the **TOOLS** through which the 3-dimensionallyconscious mind becomes empowered to participate in conscious co-creation with the Soul and higher dimensional spiritual portions of identity. **Color** and **Symbol** possess power as **MATHEMATICALLY ENCODED** directors of scalar frequency; the ultimate effect of this directed power is the ability to affect the scalar arrangements of the **Manifestation Template**. Intrinsically, **scalar waves** are specific points of vibrating bi-polar energy signatures. Vibration of energy units creates patterns of internal and external **SOUND**, within and beyond the range of 3-dimensional detection. The Manifestation Template can be conceptualized as an **INTERWOVEN FABRIC OF SOUND FREQUENCY**, the "Silent Symphony" of personal and universal being. The tools of Color and Symbol affect the scalar-wave arrangement of the Manifestation Template, and thus alter the vibrational patterns of **SOUND TONES** within the personal blueprint. **COLOR AND SYMBOL are tools for the direction of SOUND**, and intentional application of **SPECIFIC SOUND TONES** can also be used in conjunction with Color and Symbol, to further amplify one's power to interact with the Manifestation Template.

Along with the techniques of Color and Symbol direction, **TONING**, or the **intentional generation of specific sounds to affect the Manifestation Template**, is also utilized in Bio-Regenesis technologies and spiritual expansion programs. "Toning Therapies" have emerged in recent years through the New Age and Holistic Healing paradigms, and have been applied for thousands of years, within the practices of spiritual masters in every religious tradition. SPOKEN LANGUAGE itself is a form of TONING, with its accompanying SYMBOL ALPHABET, and like all Symbols and Sound Tones, our languages directly affect the personal Manifestation Template, and the level of consciousness that can manifest through the body. In ancient cultures, the powers of language were better understood than they presently are in contemporary mainstream society. Particularly within the Priest-casts of ancient religions, language was used more consciously, with an understanding that words, sounds and images would directly effect the consciousness of the masses. Prayers, rituals and songs, some of which remain in use today, were used as a means of *altering consciousness*. Some assisted in spiritual development, others were intentionally used to limit and control the consciousness of populations.

The contemporary earth science paradigm is only beginning to explore the power of sound vibration and the connection between sound, symbol and color is not yet recognized or understood. **In advanced pre-ancient cultures, these sciences of Scalar Mechanics were fully understood and utilized in the most powerful of ways.** Color-Symbol-Sound technologies were not only used in **Bio-Regenesis** applications for personal health and consciousness expansion; they were also used among **GROUPS OF PEOPLE** to directly interface with the scalar-wave Manifestation Template of the **PLANTETARY SHIELDS**. Vortices, Time Portals, Star Gates, Ley Lines and free energy systems generated through the planetary core and solar light emanations were all **DIRECTED THROUGH THE FOCUSED MIND**, utilizing knowledge of Color-Symbol-Sound Scalar Mechanics. Since the fall of ancient cultures, which came as a result of abusing the powers of these Scalar technologies, the Color-Symbol-Sound sequences once used to direct massive amounts of power became the **"HIDDEN AND FORBIDDEN SACRED KNOWLEDGE"**.

The **COLORS, SYMBOLS** and **SOUNDS** used in Bio-Regenesis techniques are not simply random choices. The elements of Bio-Regenesis technique are drawn from a **UNIVERSAL STANDARD LANGUAGE OF "SCALAR SPEAK"**, through which **precise scalar frequencies** are generated through **specific Color-Symbol-Sound sequences** to create very specific effects within Manifestation Templates. The Bio-Regenesis Techniques presently offered represent only the beginning of our reawakening to this vast body of Sacred Science knowledge. Many Bio-Regenesis techniques utilizing Color-Symbol sequence scalar mechanics are accompanied by a corresponding exercise in **TONING**. Within the mystical **Ascension Schools** of ancient and pre-ancient times, there existed a very **SACRED SERIES OF TONING SEQUENCES** that were used **to directly interface with the DNA TEMPLATE** and the **AXIOM LINES** in the **PLANETARY**

SHIELD. On a **personal level,** through this specific series of Toning, dormant strands in the DNA Template could be progressively brought into activation, to expedite the process of Spiritual Actualization, Holistic Healing and building of the **PERSONAL MERKABA VEHICLE** within the bio-energetic field. On the **planetary level,** these **Sacred Toning Sequences** were used to **ENERGETICALLY OPEN** the **PLANETARY AXIOM LINES** for **CLEARING, REPROGRAMMING AND ACCELERATING THE ACTIVATION** of the **PLANETARY SHEILDS.** This series of Sacred Toning was collectively called **"THE MUSIC OF THE SPHERES".**

THE SONGS OF THE SPHERES

The **MUSIC OF THE SPHERES** represents the **ORGANIC TONAL SEQUENCES** of **EACH DIMENSIONAL FREQUENCY BAND** within the 15-Dimensional Time Matrix; the **CONSOLIDATED CORE VIBRATIONAL SOUND SIGNATURES** of each Dimensional frequency band. Of the 15 core vibrational signatures of the 15 Dimensions in a Time Matrix, 12 of these core signatures serve as the **PRIMARY 12 TONAL SEQUENCES** upon which each strand of the 12-Strand original human **DNA Template,** and each **AXIOM LINE** in the 4-Density **Planetary Shields** are structured. In using the **Tone Sequences of the Music of the Spheres,** the 12-Strand DNA Template can be **PROGRESSIVELY BROUGHT INTO NATURAL ACTIVATION.** Scalar-wave distortions within the Templates for the DNA Strands (which physically manifest as genetic distortions) can be systematically **REALIGNED** with their natural perfect 12th Dimensional **ORGANIC IMPRINT.** In using the **Music of the Spheres** for work with the **Planetary Shields,** each of the **24 AXIOM LINES** can be **OPENED** and **ENCODED WITH THE PERFECT PATTERN OF THE PLANETARY PLASMA BEAM.** Transmitting the 12th-Dimensional **MAHARIC SHIELD** pattern directly into the **Axiom Lines** in the Planetary Shields is the most rapid method to distribute the program of the **PLANETARY ORGANIC IMPRINT** throughout the 4 Density Levels of the PLANETARY SHIELDS.

In ancient days, the *Music of the Spheres* was used extensively for purposes of self-generated healing and to prepare the physical body for literal passage through the planetary Star Gates ("Ascension"). The **specific tone sequences** for each **SINGULAR DIMENSION** (which corresponded to one DNA Strand Template in the 12-Strand blueprint and the inherent Dimension of consciousness that would embody through activation of the strand) are called the **"SONGS OF THE SPHERES".** Each **"SONG"** contained very specific **"word sounds"** (from the first Universal Spoken language) for **lyrics,** and specific **multi-layered tones** (which generated inaudible Tri-tone Standing Waves) for **melody, harmony** and **overtones.** To be effective, the Songs had to be used through the **HUMAN VOICE,** externally or mentally as the **"SINGING OF SONGS".**

The Sacred Songs were sung in combination with directing specific corresponding **COLOR** and **SYMBOL** sequences through the bio-energetic field (or through the Planetary Axiom Lines for Planetary Shield work). **The Songs served to "open" specific areas of the DNA Template, or specific Planetary Axiom Lines, to receive instructions directed by the Color and Symbol sequences.** The tones of the Songs generated specific scalar-wave formations that triggered activation of corresponding configurations within the DNA Template and Planetary Shields. Activation of dormant portions of the DNA Strand Templates or Planetary Axiom Lines "opened" the Strand Template or Planetary Shield to directly receive Color-Symbol scalar wave instructions, through which desired outcomes within the personal or planetary body, and within the external reality experience, could be **"PROGRAMMED"** for manifestation.

The **SCIENCE OF THE MUSIC OF THE SPHERES** was one of the most highly guarded **"SACRED SECRETS"** in pre-ancient cultures, due to the power of interdimensional freedom and conscious manifestation that could be obtained through knowledgeable use of the **SONGS.**

Before the **SONGS OF THE SPHERES** will be fully returned to human awareness, our culture will have to mature enough to demonstrate **spiritually motivated ethical use of power.** Certain SONGS are now being slowly returned, so that we may assist those of higher evolution to realign Earth's Planetary Shield with its original 12th-Dimensional **MAHARIC SHIELD ORGANIC IMPRINT.** One of the **15 SONGS OF THE SPHERES** was provided to us by the Emerald Order. This SONG is used to activate dormant scalar-waves in the 8th **DNA STRAND TEMPLATE**, through which the process of activating the dormant **1st-8th Strands** of the 12-Strand DNA Template can be accelerated, expediting the natural processes of higher dimensional spiritual identity integration.

This SONG is also used to open the **8TH AXIOM LINE** in **GAIA'S PLANETARY SHIELD**, to facilitate realignment of **Earth's Planetary Shields** from **ETHERIC MATTER DENSITY-3.** This SONG was given so that **GROUPS** can begin assisting directly in Planetary Grid work, opening specific Axiom Lines and Vortices in the planetary body to help bring Earth's Planetary Shield Manifestation Template into its organic alignment and balance. The SONG that was given is called the **"Ariea Khum Nar'A"**. The English language translation is **"THE SONG OF ORION"**. This **SONG OF THE 8th SPHERE** corresponds to the **8th -Dimensional level** of the **UNIVERSAL KATHARA GRID** and its **8TH PRIMARY STAR GATE.** It also corresponds to the **8th PLANETARY AXIOM LINE** within the Etheric Matter Density-3 **GAIAN PLANETARY SHIELD** and to the **8th DNA STRAND TEMPLATE** in the human body. The **8th Dimensional STAR GATE** within the **UNIVERSAL KATHARA GRID** is the Star Gate that opens into the **ORION STAR SYSTEM**, thus the name of the **SONG OF THE 8th SPHERE** is **"THE SONG OF ORION"**.

Soon the SONGS that correspond to the 11th and 12th –Dimensional frequencies will be provided. Aside from its use in planetary grid work, when used as a **"MANTRA"**, repeatedly sung as a focus for the mind in meditation, the ***Ariea Khum Nar'A*** serves to stimulate consciousness expansion into the higher dimensional fields and amplifies and expedites the effects of personal Bio-Regenesis techniques.

CHAPTER 1

THE CONTEXT OF HEALING

The Structure of the Unified Field and the Multi-dimensional Anatomy of Form

- THE HOLOGRAPHIC TEMPLATE
- THE STRUCTURE OF DIMENSIONS, PARTIKI PHASING AND MATTER DENSITY
- UNIVERSAL AND PLANETARY KATHARA GRIDS, THE 12-TREE, KATHARA CENTERS AND LINES, SCALAR SHIELDS, CRYSTAL SEALS AND DIODIC POINTS
- THE UNIVERSAL AND INTERNAL TEMPLAR COMPLEXES
- THE TRANSDUCTION SEQUENCE AND HOLOGRAPHIC REALITY CONSTRUCTION

TOPICS OF DISCUSSION

The Ariea Khum-Nar'-A

The Song of Orion

Part 1 : the Resonant Tone Syllable Sequence (primary melody)

Verse 1: OM A Da E, Patum ah
OM A Da E, Patum ah
Patum ah A' DE E - Tah, A khum-na
OM A Da E, Patum ah

Verse 2: TrA TE La E' Durum Na
TrA TE La E' Durum Na
Durum' ah A' Khem - Tah, A khum -na
OM A Da E, Patum ah

Part 2 : the Base Tone Syllable Sequence (accompaniment 1)

OM A Da E, Patum ah - Lines 1, 2 & 4 Verse 1 & Line 4 Verse 2
TrA TE La E' Durum Na - Lines 1 & 2 Verse 2
UM ah OM, A
UM ah OM, A
Patum ah A' DE E – Tah, A khum- na
Durum ah A' Khem – Tah A khum-na
DO RAY' TU Ra"
OM A Da E, Patum ah
UM ah OM, A

Part 3: the Over Tone Syllable Sequence (accompaniment 2)

OM A Da E, Patum ah – Lines 1 & 2 Verse 1
TrA TE La E' Durum Na – Lines 1 & 2 Verse 2
BE HE Tah OM, A
BE HE Tah OM, A

Patum ah A DE E – Tah, A Khum –na – Line 3 Verse1
Durum ah A Khem- Tah, A Khum - na - Line 3 Verse 2
Thra Khem' Na, A
OM A Da E, Patum ah – Line 4 Verses 1 & 2

KhrA MiM, Tah A

THE STRUCTURE OF THE UNIFIED FIELD AND THE CONTEXT OF HEALING

If one is to become fluent in the art and science of **Bio-Spiritual Healing Facilitation**, it is necessary to become familiar with the *primary structures of energy* within which existence takes place. Without this understanding one will be unable to identify the *true nature of systems and their intrinsic functions*. If we do not understand the **intrinsic order of a system** and the elements that make up its parts, our ability to comprehend the *organic functions* of the system will be limited. Without comprehension of the intrinsic dynamics inherent to a system, our perceptions will be limited to observing the **effects of systems function**, rather than perceiving the **causal elements** behind and within the apparent observable effects. In relation to the **health and healing** of the human body, it is wise to view the human **Body-Mind-Spirit organism** as a series of interwoven systems, each with *inherent, implicit order*, the perceivable and observable aspects of which represent the effects of systems function. **The observable effects of the human condition emerge from an as yet unidentified causal order, a hidden and mysterious source out of which the manifest effects of human existence emerge.** Holistic Healing approaches human health and healing in terms of the *interdependent relationships* between *multiple interwoven systems*. The Body-Mind-Spirit organism is viewed in terms of the *intrinsic relationships that unite* the multiple systems of the human being, rather than in terms of the *observable boundaries* that appear to *separate and divide* the multiple systems inherent to the human organism.

Holistic Healing, and **KATHARA Healing** particularly, take the concept of **intrinsic unity** within the seemingly separate systems of the human being one step further than *simple interrelationship* between **physical, mental and spiritual aspects** of human experience. **In the paradigm of Kathara Healing, the human organism is viewed as a set of interwoven systems that is itself part of a larger system of interdependent systems.** The human individual is not only recognized as a *singular set of unified parts*, but rather as an *intrinsic part* of the species, planetary, galactic, universal and cosmic systems within which human existence finds **context** for being. The **Cosmic Unified Field of Energy and Consciousness** is viewed as the **indelible medium** within which all reality takes place, and human existence is viewed in terms of *its relationship to this greater Meta-system* of interwoven, interdependent systems. **The health and wellbeing of the human being is directly related to and effected by the systems-functions of the greater Unified Field.** In order to better understand and facilitate the health and healing of biological organisms, it is useful to understand the *basic order and function* of the **UNIFIED FIELD**- the context in which life itself takes place.

Kathara Healing is an ancient system of **Holistic Healing**, dating back beyond presently identified recorded human history. The principles of **Universal Order** as taught within Kathara Theory present a **broader context of universality** than is presently recognized by contemporary western science. Kathara is built upon **Laws of Multidimensional Universal Physics,** laws of intrinsic nature with which western science is not yet familiar. In Kathara Healing the *hidden order of universal systems* is recognized and the foundations of Kathara *demonstrate the interwoven systems-functions* between the human being and the order of the **Unified Field. To develop proficiency and genuine skill within the art and science of Holistic Healing, the basic order of the Unified Field must be understood.** For with this understanding comes the ability to *directly interact with and influence the causal elements* of the human experience. For this reason the teachings of Kathara Healing begin with providing a basic format through which the **CONTEXT of healing** can be understood ***– a glimpse at the intrinsic order of the Universal Unified Field.***

With foundations grounded in **coherent systems relationships**, the Kathara Healing Facilitator has more knowledge, power and comprehension to bring to the art and science of healing facilitation. **Through comprehension of the *intrinsic systems of universal order*, the Kathara Facilitator progressively evolves to embrace the *unity of SELF within the Unified Field*, opening to an ever-expanding capacity of spiritual awakening and the embodiment of UNCONDITIONAL LOVE.** The physical, mental and spiritual aspects of human anatomy represent one interdependent, interwoven system that is *continually in a state of relationship to the Unified Field.* This relationship must be acknowledged and embraced if true health and healing are to occur. **Human evolution is a Bio-Spiritual process, the body and the consciousness – the biology and the spirit - are inseparably intertwined within themselves and within the greater order of the Unified Field.**

Healing emerges through restoring order to the ONE-SYSTEM and its infinite parts.

THE HOLOGRAPHIC TEMPLATE

The externalization of life, manifestation of matter and individuation of identity are **HOLOGRAPHIC PROJECTIONS of CONSCIOUSNESS-ENERGY SUBSTANCE,** created through an *intrinsic order of energetic relationships*. Energetic relationships represent **interrelationships of consciousness,** in its infinite manifest and de-manifest forms.

Consciousness *IS ENERGY*
And
Energy *IS CONSCIOUS*.

True Science has a *spiritual core* and true spirituality has *its intrinsic translation within the Laws of Energy that govern the manifestation of universal systems.*

Reality is **THOUGHT CONSTRUCTION.** The core substance of the cosmos is **CONSCIOUSNESS.** ***Thought*** *is an attribute of CONSCIOUSNESS*, the **filter** through which consciousness manifests itself into the **Hologram of Form.**

Nothing is truly solid. All things are composed of *consciousness* and their *apparent solidity* is determined by the *relationship between the consciousness observing the form and the consciousness of which the form is made.* Human consciousness emerges through an **order of dimensionalization,** which sets frameworks within which the *interrelationships of consciousness* can take place. **Human perception of solidity and externalization is determined by the oscillation rates between dimensionalized units of consciousness.** Dimensionalized units of consciousness form **frequencies of energy,** which exist in *specific relationships* to each other. Human *consciousness* (the soul –spirit), and human *physical form* are **patterns of frequency** formed by *dimensionalization of units of consciousness. We perceive manifest solidity due to the relationship between the frequencies of consciousness of which we are made and those of the Unified Field of Conscious Energy around us.* **Humans perceive as solid the frequency bands that exist one full dimensional spectrum below the frequencies within which the individuated or collective consciousness is stationed.**

The only difference between a *thought* and a *manifest thing* is the *frequency of the consciousness of the observer.* Manifest reality is **thought projection,** made solid and externalized by relationships between the *frequencies of our focus of attention* and those of the *projected thought forms*. **We perceive as reality the thought forms of the collective masses as they were placed within the frequency bands that now exist one dimension below the frequencies within which our present focus of collective consciousness is stationed.**

As reality is a **THOUGHT FIELD**, composed of units of consciousness awareness, reality can be directly **AFFECTED by THOUGHT**, as the substance of both thought and reality are one and the same. The **key to mastery** of our present reality and to restoring our **imprint for health**, is understanding the *reality of dimensionalization* within which our consciousness now resides.

The human body is a HOLOGRAPHIC PROJECTION of consciousness that is built upon a *HOLOGRAPHIC TEMPLATE* of structured units of consciousness. The **Holographic Template** represents the **Living Morphogenetic Field** – the *dimensionalized blueprint of conscious light, sound and scalar waves*, within which the individuated consciousness is stationed, and *upon which the illusion of solidity of the body matter is formed.* **To affect *true and lasting change* within the Body-Mind-Spirit System of the human being, one must *go to the core* of its manifest structure – the *Holographic Template* or Thought-energy Blueprint, upon which the entire system of consciousness itself comes to know being.**

The body is a **HOLOGRAM** that can be directly affected by the *substance of which it is composed* – **CONSCIOUSNESS**. In learning the structure of dimensionalization through which the body-illusion manifests, one becomes progressively more masterful in directing this personal thought form to suit ones desired specifications. The core **Holographic Template** of the personal and universal body is called the **Kathara Grid. Kathara Healing** is the process by which the *embodied* aspects of human consciousness become more aware of and able to direct the *intrinsic processes of holographic projection,* with the intention of restoring integrity to the function of the Body-Mind-Spirit system.

No thing is truly manifest **– it only appears to be so,**
due to the refraction of consciousness
within the energetic relationships inherent to
the *Holographic Template* –
the original *thought-form construct*
upon which our universal structure
is perpetually recreated.

All creatures and things are Spiritual Beings having a Manifest Experience...No creature or thing can be ever divorced from its innate spirituality, but only from the awareness of its nature. For life is the expression and substance of spirit- and manifest reality is but a Dreamscape, from which we eventually all awaken...

The causal element of all manifest experience is called a ***Morphogenetic Field*** or **MANIFESTATION TEMPLATE**. Keylontic Science, the science of Manifestation Templates is part of a larger advanced paradigm of science that is known as **15-Dimensional UNIVERSAL UNIFIED FIELD PHYSICS.**

PRIMAL SUBSTANCE

- A Manifestation Template is made of **PRIMAL SUBSTANCE;** units of conscious energy called **PARTIKI** that exist as ***Omni-polar points of fixed vibration. Partiki units*** are the smallest building blocks of matter, and they form the templates upon which consciousness in all forms enters manifestation.
- Partiki units are **Omni-polar** (containing the potentials for all polarities or none) units of vibrating energy that perpetually cycle back and forth between a state of **BI-POLAR LIGHT RADIATION** (scalar-standing-wave) and **OMNI-POLAR SOUND VIBRATION.**
- In the ***Omni-polar state,*** Partiki units exist as invisible, fixed points of ***sound vibration,*** forming what is called an **INTERDIMENSIONAL TRI-TONE WAVE.** An Interdimensional ***Tri-tone Wave*** is a singular point of fixed vibration that is composed of ***3 sub-vibrations*** (base tone, overtone and resonant tone). When Partiki units are in their Omni-polar state they exist as ***Ante-matter substance,*** the first state of energy organization before Pre-matter manifestation. Through a process of Internal Fission characteristic to their design, Partiki units break apart to form Bi-polar waves, while replicating their original Tri-tone Omni-polar design.
- In the ***Bi-polar state,*** two sub-vibrations of the Tri-tonal Standing Waves of Partiki units break down into two sub-units, ***sub-tones,*** called **PARTICUM**-*Base Tones* and **PARTIKA**-*Overtones,* interconnected units of **BI-POLAR LIGHT RADIATION** that form ***Electromagnetic Scalar-Standing-Waves.*** The third vibration of the Tri-tone wave, the *Resonant Tone,* remains as the core vibration through which the polarized vibrational units of Particum and Partika will reunite through ***Internal Fusion.***
- **PARTICUM** units form the energetic building blocks for **PARTICLE TEMPLATES** and **PARTIKA** units form the building blocks for simultaneously manifesting **ANTI-PARTICLE TEMPLATES.** Particle and Anti-particle Templates are interconnected, united by the Ante-matter **PARTIKI** units through which they manifest. The effects of **ELECTRICAL AND MAGNETIC FORCE** are set in motion through the **EXPANSION AND CONTRACTION** of energy vibration between Particle and Anti-particle matter units through the Ante-matter Partiki unit that unites them.
- **Particum-Particle units** represent the **CONTRACTION PHASE** of ante-matter energy, with higher **VIBRATION** (energy contracted into or received by the Ante-matter template) and lower **OSCILLATION** (energy expanded or transmitted from the Ante-matter template) and so are considered **BASE MAGNETIC. Partika-Anti-particle units** represent the **EXPANSION**

PHASE of Ante-matter energy, with lower Vibration and higher Oscillation, and are considered **BASE ELECTRICAL.**

- **ELECTROMAGNETIC ENERGY** is created through the continual cycling of energy, as the Omni-polar Ante-matter Tri-tone Wave of ***Sound Vibration*** projects its energy into Bi-polar Particum and Partika Scalar Waves of ***Light Radiation***. Energy thrust expands into the Anti-particle Partika vibration point as **electrical force**. It then contracts into the particle Particum vibration point, creating **magnetic pull**, (in ratio to its original electrical thrust), as the third vibration in the Ante-matter Tri-tone Wave draws the energy back to its source. Through the draw of the Ante-matter Partiki template, Partika-Ante-particles and Particum-particles are brought into a state of **Fusion.** Upon fusion, the energy of the Particum and Partika transmutes, the Bi-polar Standing Scalar Waves of Light Radiation returning to their original Omni-polar Ante-matter state of Partiki units. Once returned to the Partiki state, the process repeats as the Partiki creates Fission and replication. **Partiki, Particum and Partika units represent PHASES that Primal Substance energy passes through as it moves into and out of manifestation.**
- The process by which Partiki units perpetually convert conscious energy into Bi-polar Light Radiation Scalar-Waves (expansion) and back into Omni-polar Sound Vibration Tri-tone Waves (contraction) is called **PARTIKI PHASING.** Through perpetual cycles of Partiki Phasing, matter units "flash on" through **fission**, into Bi-polar Particle and Anti-particle manifestation built upon Scalar Waves of Light Radiation, then "flash off" through **fusion**, into Omni-polar Ante-matter Sound Vibration, continually cycling energy between the manifest and non-manifest state. **The perception of external physical matter, space and the passage of time are created through *variance in relationship* between rates of Partiki Phasing.**
- Consciousness in the form of Partiki Units is the **PRIMAL SUBSTANCE** of which the cosmos is structured and Partiki Phasing is the perpetual **PRIMAL ACT** of reciprocal fission and fusion through which Primal Substance is brought into manifestation.

The Holographic Template

PRIMAL SUBSTANCE

Units of Consciousness, Morphogenetic Fields and Scalar Grids

1. ***Partiki, Partika and Particum:*** Smallest units of energy-substance. Electro-tonal units of consciousness that are the building blocks of morphogenetic fields, matter and individuated conscious identity. Operate as minute, perpetual motion fission/ fusion generators.

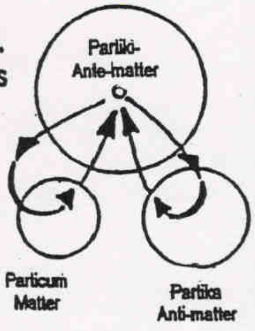

2. ***Partiki Grids*:** The fabric of morphogenetic field structure. Partiki units group to form interwoven Strands, then Grids of electro-tonal substance, out of which morphogenetic fields are fashioned. Out of a Unified Field of Partiki units, Partiki Grids form to create the frequency bands that make up dimensional fields and then to form individuated morphogenetic fields.

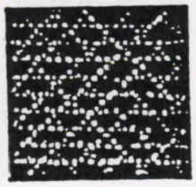

3. Partiki units continue to group or accrete upon the Partiki Grids, forming
A. ***Keylons*** – crystallizations of "frozen light" (standing scalar wave patterns and fixed points of tonal frequency), and
B. ***Keylon Codes*** – complex groupings of Keylons, which together form a crystalline template of light spectra, sound frequency and electro-magnetism that is the morphogenetic field Crystal Body– the Blueprint upon which matter and identity will manifest.

A.

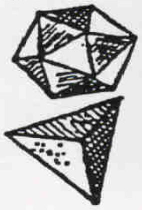

B.

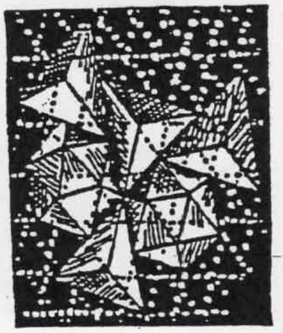

The **Morphogenetic Field** is made of interwoven **Partiki Grids**. The electro-tonal units of consciousness out of which Partiki Grids form create standing wave patterns – **scalar wave grids**, which hold the form of consciousness within dimensionalized manifestation. Morphogenetic Fields are thus **Scalar Grids**, that form in specific ordered interrelationship. Partiki Grids form upon the core *geometrical-mathematical* structure of the **Kathara Grid**, and so the Kathara Grid represents the *core Scalar Field* upon which form manifests.

Morphogenetic Fields govern the form of matter manifestation and the evolution of biology and consciousness. Morphogenetic Field mechanics thus hold the key to mastery of physical reality and Bio-Spiritual Evolution.

THE STRUCTURE OF DIMENSIONS

Frequency Bands and Sub-Frequency Bands, Vibration-Oscillation-Particle Spin, Harmonics of Manifestation and Matter Density Levels.

The Hologram of manifest reality is a complex system of interwoven energy constructs, all built upon **dimensionalized sets of scalar grids**- or fields of standing wave points – that form the **base laws of energy** upon which manifest reality functions.

The physical portions of the human body represent ***wave strata*** that exist within specific relationships to the embodied consciousness. **The embodied consciousness itself is *regulated by the intrinsic behavior of the wave strata* in which it is ensconced.** The wave strata of seemingly physical forms appear to *human detection* in the form of *particles, sub-particles and wave bands* that group to form the 3-dimensionally manifest hologram of human life

The **behavior of dimensionalized wave strata** is an intrinsic element within the function of ***human anatomy***, and provides the *very context* within which human existence and evolution take place. When we come to understand ourselves as **vibrating quantities of conscious energy**, existing within a state of **vibrational co-resonance with all things**, which are themselves *vibrating quantities of conscious energy*, we can begin to understand the **fundamental structures** upon which the human condition is built.

The human body is a ***construct of conscious energy****, a living* morphogenetic field existing within the greater living morphogenetic constructs of the species, planetary, galactic, universal and cosmic morphogenetic structure. We are simply an element of sentient consciousness within a larger structure of sentient consciousness, a part of the **Unified Field of Consciousness** that IS the living identity of the cosmos. Our physicality emerges through the **perceptual lens of dimensionalization**, which provides us with the seeming manifestation of **particles, atoms, molecules** and the **physical systems** built upon them.

If we expect to **heal ourselves** of the manifest conditions that cause suffering within the Body-Mind-Spirit system, we must first understand the **nature of health**, the "way things are supposed to work" when they are *functioning properly*. To understand the intrinsic **Imprint and impetus for health** in the human organism, a basic understanding of the *contents of its parts* is in order. That understanding begins with *identifying the basic reality structures within which we reside, and how we exist in relationship to these structures.*

Our understanding of *health* begins with *identifying the context of existence*, the *Unified Field of Consciousness-Energy* of which we are all a part. **A simple *identification of the structures of the Unified Field* will work wonders in assisting us to understand ourselves, and through this understanding we become more empowered to foster a healthful and more joyous living experience.** To better understand ourselves we must return to the core of our intrinsic nature, the **dimensionalized fields of consciousness** within which we reside and out of which the substance of our SELVES is composed.

The following characteristics inherent to the manifestation of the human hologram are a place to begin our journey of self-understanding.

Characteristics of the Hologram

PRIMAL ORDER

SCALAR WAVES: Points of standing waves, composed of quantities of conscious energy that emanate out of fixed units of vibration (Partiki units), which form the morphogenetic field templates upon and through which consciousness dimensionalizes itself to experience the hologram of manifest reality. Created through Partiki Phasing (internal fission/fusion "flash on/flash off")

MORPHOGENETIC FIELDS: The templates of consciousness, light, sound and scalar standing-wave grids which serve as the structured blueprints upon which our consciousness manifests into dimensionalization, through which the holographic projections of space-time-matter and identity individuation can be experienced.

DIMENSIONS: *Morphogenetic Field Scalar Grids composed of interwoven layers of scalar waves, which serve to direct the flow of consciousness into multiple patterns of refraction through which the hologram of matter density, linear time and objectification of reality can be experienced.* Dimensions are specific structures of consciousness, composed of Partiki units with differing vibratory-oscillation rates, which exist in precise relationship to each other, and govern the intrinsic laws of function within the observably manifest worlds. The process of dimensionalization occurs through the wholeness of consciousness *compartmentalizing itself into various ordered sections having diversity of vibratory-oscillation*, through which the wholeness of consciousness may experience itself in an endless variety of forms. Each dimension represents a full frequency band.

FREQUENCY BANDS: Synchronized, cyclic, repeated sequences of "flashing on and off" of scalar-wave points within the dimensional scalar grids, which gives the appearance of manifest movement of wave spectra within a dimensional system. **Sub-frequency Bands** are shorter segments of the longer "flashing on and off" sequence of the full Frequency Band. The frequency of dimensional levels and the particles within them is determined by the vibratory-oscillation rate of the Partiki units that make up the scalar-wave grid. Frequency is the form consciousness takes on while ensconced within the framework of dimensionalization. Frequency is the vibrating-oscillating wave spectra of which matter particles and dimensionalized consciousness are composed.

VIBRATION-OSCILLATION: Vibration is the condition of **internal movement**, or ***holding of energy within*** units of consciousness (Partiki) and the particles built upon them. Vibration is determined by the **quantity of energy *held within*** Partiki units. **Oscillation** is the condition of **external movement**, or ***expending of energy from*** Partiki units and the particles built upon them.

The vibratory-oscillation rate of particles represents the *rate at which the scalar standing-waves "flash off and on"* within the morphogenetic field template. Vibration is the *internal measure of energy quantity* held *within* Partiki units and exists within direct proportion to the *oscillation* rate, or *external measure of energy quantity expended*, exhibited by Partiki and the particles built upon them.

The vibratory-oscillation rate represents the ratio between energy contraction (vibration) and energy expansion (oscillation), which sets the rhythm of particle pulsation that governs the formation of wave spectra within dimensionalization. The vibratory-oscillation rate of Partiki units determines the **frequency** (cyclic pattern of scalar waves "flashing off and on") of wave spectra within the various levels of the dimensionalized morphogenetic field. **The relationship of vibrational-oscillation between wave spectra determines the matter density levels particles will exhibit within the dimensional scale.**

Vibration-Oscillation Correspondences

Higher Dimensional Fields (8-15)	Lower Dimensional Fields (1-8)
Low Vibration	**Higher Vibration**
Higher Oscillation	**Lower Oscillation**
Expanding	**Contracting**
Higher Frequency	**Lower Frequency**
Shorter Wave Length	**Longer Wave Length**
Slower Electron Spin	**Faster Electron Spin**
Less Density	**More Density**
Electric	**Magnetic**
Lower Vibration =	Higher Vibration =
Less energy holding in Partiki =	More energy holding in Partiki =
More energy expended from Partiki =	Less energy expended from Partiki =
Higher Oscillation =	Lower Oscillation =
More expansion =	More contraction =
Faster flash rate of Partiki =	Slower flash rate of Partiki =
Higher Frequency of Flash =	Lower frequency of flash =
More external movement =	More internal movement =
Faster oscillation =	Slower oscillation =
Shorter wave length =	Longer wave length =
Slower electron spin around nucleus =	Faster electron spin around nucleus =
Greater expansion of energy units =	Greater compaction of energy units =
Less density =	Greater density =
Less magnetism =	Greater magnetism =
More electric	Less electrical

The human anatomy is a 15-dimensional morphogenetic field scalar grid structure, composed of wave spectra with varying degrees of matter density characteristic to the dimensional band within which each level of the anatomy is stationed. The portions of our anatomy that appear solid are those stationed one dimension below our focus of conscious attention.

PRIMAL ORDER OF MANIFESTATION TEMPLATES

- Morphogenetic Fields, or the **MANIFESTATION TEMPLATES** that serve as the blueprint through which consciousness manifests, are composed of the Primal Substance of **PARTIKI UNITS.** Partiki units draw together or **ACCRETE** to form **PARTIKI STRINGS.** Following a set **mathematical and geometrical design** inherent to their form, Partiki Strings interweave and overlap, forming a **"Fabric of Light and Sound"** called a **PARTIKI GRID.** A Partiki Grid is a fixed pattern of interwoven Bi-polar Scalar-Waves of light radiation,built upon Omni-polar Tri-tone Waves of sound vibration,that forms the foundation for diversified Manifestation Templates. A **Partiki Grid** is a **SCALAR GRID.**

- Following the precise mathematical design set by the will and intention of the manifesting consciousness, Partiki Grids further accrete, drawing in more Partiki Units, forming groups of Partiki units that are arranged following *3 different rhythms of Partiki Phasing.* **Tri-phase groupings of Partiki Units** form patterns of Scalar-wave and Tri-tone wave interrelationships that create ***crystallizations of energy*** called **KEYLONS.** Groups of Keylons continue to accrete, forming complex Keylon arrangements called **KEYLON CODES,** through which the Manifestation Template is progressively built up from Ante-matter through various stages of matter density, beginning with **PRE-MATTER SUBSTANCE,** the first density of **Liquid Silica Hydro-plasmic** energy to emerge into externalized form.

- The Morphogenetic Fields upon which universes manifest are built upon a Manifestation Template that has a specific order of mathematical interrelationships that govern the combining of Scalar Grids with different rates of ***Partiki Phasing.*** The **Cosmic Manifestation Template** is called the **ENERGY MATRIX.** Within the Energy Matrix there are many smaller manifestation templates upon which Universes are structured. **Universal Manifestation Templates** are called **TIME MATRICES.** There are uncountable Time Matrices within the Energy Matrix, which are collectively referred to as **THE TIME MATRIX.** All Time Matrices within the Cosmic Matrix follow a specific mathematical-geometrical program through which space, time and matter can be experienced by consciousness upon entering the Time Matrix structure.

- Time Matrices are arranged upon a pattern of 15 different, but interwoven, Partiki Phasing rhythms (rhythms of expansion and contraction of energy). Each of the 15 rhythms of Partiki Phasing creates one **DIMENSION.** A **Dimension** is a set pattern of **"FLASH LINE SEQUENCES"**, or a cycle of singular Partiki Phasing rhythms that contains within it 12 smaller rhythms of Partiki Phasing. Each **Dimensional Manifestation Template** creates a set of fixed Scalar-Tri-Tone Wave Points that create a template of stationary **SCALAR FREQUENCY.** Each Dimension represents one **SCALAR FREQUENCY BAND** containing 12 smaller **SUB-FREQUENCY BANDS.** A **Scalar Frequency Band** is a cyclic pattern of specific Partiki Phasing Rhythms, or **EXPANSION AND CONTRACTION RATIOS OF ENERGY CIRCULATION,** to which consciousness conforms in order to enter the experience of manifestation.

- To create the Universal Manifestation Template upon which external space, time and matter experience can be known, Dimensions are ordered into sets of 15, forming the blueprint for a

15-Dimensional Time Matrix. Within a Time Matrix the 15-Dimensions are further arranged into sets of **3 Dimensions,** forming **5 3-Dimensional reality fields.** Each 3-Dimensional Reality Field is called a **HARMONIC UNIVERSE**, and each represents a **level of matter densification** specific to its intrinsic rates of Partiki Phasing. One Time Matrix is thus a **15-Dimensional Scalar Grid** with 5 separate yet interwoven reality fields and 5 different densities of matter manifestation.

- Each of the 5 Harmonic Universes within a Time Matrix represents a **TIME CYCLE.** A Time Cycle is a fixed, repeating pattern of Partiki Flash Line Sequences, or a repeating cycle of specific ratios of energy expansion-oscillation and contraction-vibration. The Time Cycle of a Harmonic Universe is called an **EUIAGO CYCLE.** There are **5 simultaneously manifesting Euiago Cycles in every Time Matrix,** through which consciousness passes in order to experience linear evolution through space, time and matter. Each Euiago cycle contains within it 6 smaller Flash Line Sequences forming6 smaller cycles of time called **TIME CONTINUUA.**

The **PRIMAL ORDER of Manifestation Templates** represents the core mathematical and geometrical interrelationships of energy that consciousness takes on in order to enter the holographically projected experience of external space, time and matter. All manifest forms, including the **HUMAN BODY** and **PSYCHE** , are built upon a Manifestation Template that begins with the Primal Order of Dimensionalization, through which the consciousness anchors itself into the Time Cycles of the Universal Manifestation Template Time Matrix, to begin the experience of evolution through time.

Understanding the **PRIMAL ORDER OF DIMENSIONAL STRUCTURE,** as it pertains to the human organism, allows for comprehension of the greater causal framework through which personal healing must occur. Such comprehension of the core causal structure of manifestation enables one to work cooperatively, rather than in resistance to, the natural **Universal Laws of Manifestation,** progressively restoring the **NATURAL HARMONIC ORDER** of wellbeing and **DIVINE RIGHT FORM,** through which the indelible **Organic Imprint for Health** can be progressively restored.

Dimensional Structure

A Dimension represents a full ***Frequency Band***, or repeating, cyclic sequence of "flashing on and off" scalar-wave points within a morphogenetic field. The morphogenetic field scalar grid structure of dimensions takes the form of ***sets of 15-Dimensional Matrices***. The 15 Dimensions are grouped into sets ***of 3 Dimensions each***, forming ***5 sets of 3-dimensional reality fields*** called **HARMONIC UNIVERSES**. The 15-Dimensions composing the 5 Harmonic Universes together represent one TIME MATRIX system.

Each dimension of frequency is composed of **12 Sub-frequency Bands,** or shorter cycles of the "flashing on and off" of scalar-wave points, which exist as part of the longer cycle of the full Dimensional Frequency Band. The Partiki in each dimension have a specific rate of **vibratory-oscillation** (contraction and expansion of energy), which determines the **Angular Rotation of Particle Spin** (angular axis upon which particles will spin). There is a **90-degree shift of angular rotation** of particle spin between each dimensional frequency band and a **45-degree reverse shift** between **Harmonic Universes** (sets of 3 dimensions/frequency bands.)

Particles having varying vibratory-oscillation rates and angles of spin allow multiple dimensional reality fields to co-exist within the same space while remaining perceptually invisible to each other. The *relationships between wave strata* within the dimensional frequency bands create the **holographic refraction** of light, sound and scalar waves that allows consciousness to *perceive the illusions of* **matter solidity, space, time and externalization of reality** while it is ensconced within the structures of dimensionalization.

One 15-Dimensional Time Matrix with 5 Harmonic Universes

PRIMAL ORDER of the Dimensionalized Universal Manifestation Template

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

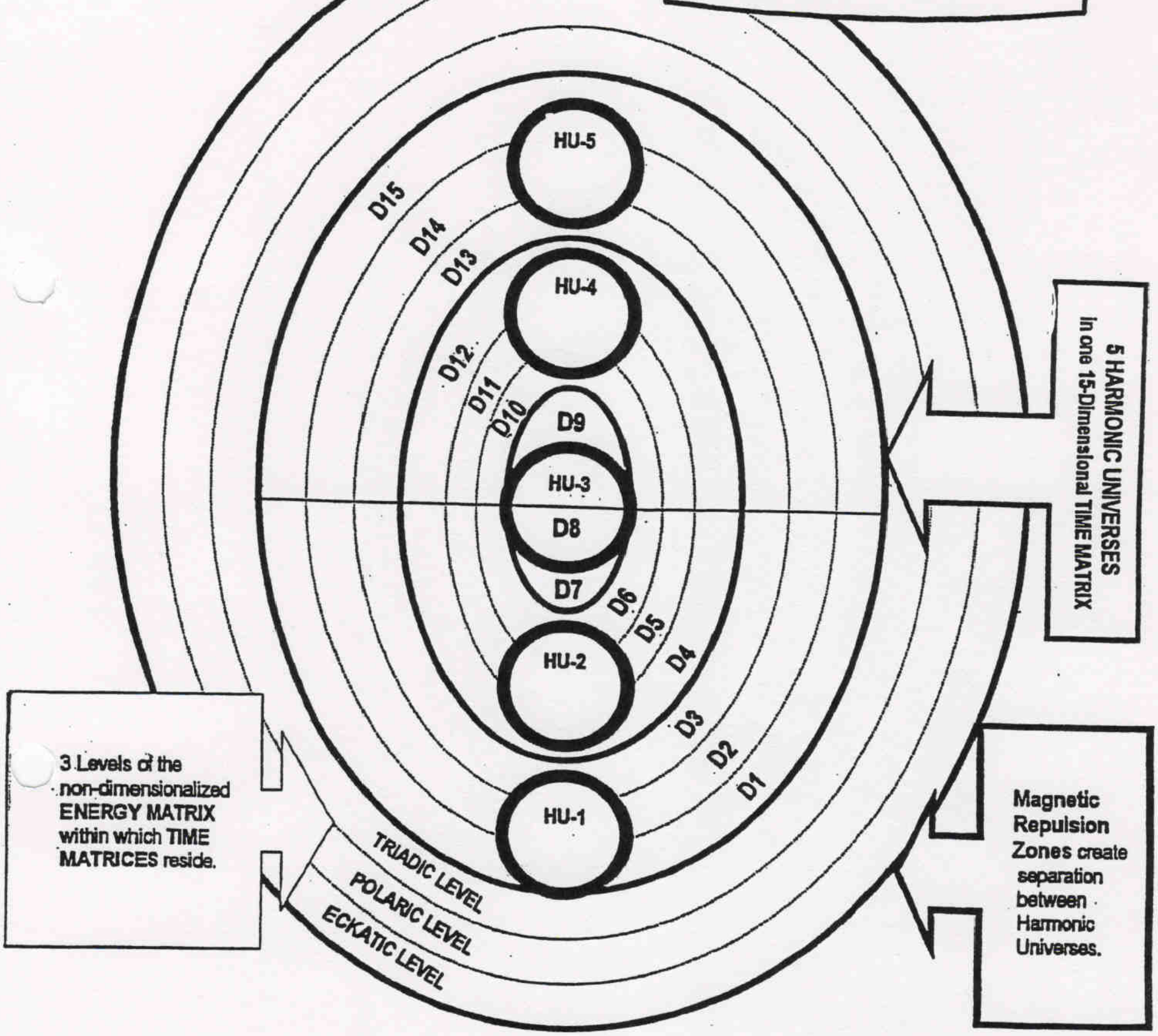
HU-5
D15
D14
D13
HU-4
D12
D11
D10
D9
HU-3
D8
D7
D6
D5
D4
HU-2
D3
D2
D1
HU-1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

5 HARMONIC UNIVERSES in one 15-Dimensional **TIME MATRIX**

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside.

Magnetic Repulsion Zones create separation between Harmonic Universes.

Frequency Bands and Sub-Frequency Bands within the 15-Dimensional Scale

Frequency Bands are cyclic, repeated sequences of "*flashing on and off*" of scalar standing-wave points that create the appearance of manifest wave spectra within a dimensional system. ***Sub-frequency Bands*** are shorter cycle segments of the longer "*flashing on and off*" sequence of the *full* Frequency Band. **Frequency and Sub-frequency Bands make up the foundation wave strata of Dimensions.**

1 Harmonic Universe composed of 3 Dimensions, each Dimension composed of 12-Sub-frequency Bands, for a total of **36 Sub-frequency Bands in 1 Harmonic Universe.**
(3 D per 1HU x 12 SFB per D = 36 SFB per 1HU) 1

①	DIMENSIONS	SFB
	D3	12 SFB
	D2	12 SFB
	D1	12 SFB
		36 SFB (12 x 3)

5 Harmonic Universes in one 15-Dimensional Matrix 2

Harmonic Universe	② Dimensions	③ 180 SFB
H 5	D15	12 SFB
	D14	12 SFB
	D13	12 SFB
H 2	D4	12 SFB
	D5	12 SFB
	D6	12 SFB
H 3	D9	12 SFB
	D8	12 SFB
	D7	12 SFB
H 4	D10	12 SFB
	D11	12 SFB
	D12	12 SFB
H 1	D3	12 SFB
	D2	12 SFB
	D1	12 SFB
		180 SFB (12 x 15)

3 15 Dimensions with 12 Sub-frequency Bands each create **180 Sub-frequency Bands in one 15-Dimensional Matrix.** (12 SFB per D x 15D = 180 SFB or 36 SFB per 1HU x 5HU = 180 SFB)

15 Frequency Bands (Dimensions) **and 180 Sub-frequency Bands within the 5 Harmonic Universes of one 15-Dimensional Time Matrix.**
SFB = Sub-frequency Band

PARTIKI PHASING, HARMONICS OF MANIFESTATION AND MATTER DENSITY

Partiki Phasing – *Scalar Standing-Wave flashing.* The internal expansion/ fission/ "flashing on" and contraction/ fusion/ "flashing off" of scalar standing-wave points made of Partiki units. *Partiki Phasing* governs the mechanics of matter manifestation within the Dimensionalized Frequency Bands of the Time Matrix. Through Partiki Phasing HARMONICS of Manifestation and varying levels of matter density are created.

Each of the 5 Harmonic Universes, moving upward in the 15-Dimensional Scale, has a progressively less dense state of perceivable matter. As you go up the Dimensional Scale each Dimension has a progressively faster rate of Partiki Phasing with its corresponding higher frequency, faster oscillation, lower vibration, shorter wave length, greater expansion and less matter density. **These conditions represent the stages of perceivable matter density consciousness will experience as the focus of attention expands progressively upward/outward into the higher Dimensions of the Dimensional Scale through the process of evolution** (accretion of frequency from the Unified Field into the personal morphogenetic field).

As consciousness progressively expands back into the 15-Dimensional structure of its original morphogenetic field, matter density perception will change from perceivable Gross Matter physical density, to Semi-etheric Matter density, to Etheric Matter density, to Pre-matter Liquid Light density and to its original Ante-matter scalar wave form of non-density consciousness. **As the morphogenetic field expands through drawing in more frequencies of consciousness, the perceivable matter density of the body and external reality field progressively lessens.**

Partiki Phasing, Harmonics of Manifestation and Matter Density

1. ***Partiki, Partika and Particum:***
Smallest units of energy-substance. Electro-tonal units of consciousness that are the building blocks of morphogenetic fields, matter and individuated conscious identity. Operate as minute, perpetual motion fission/ fusion generators.

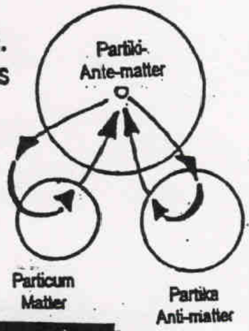

1. Partiki Phasing- expansion/ fission/ "flash on" and contraction/ fusion/"flash off"of scalar points governs the manifestation process

2.

Dimensions #	Harmonic Universe #	density type
1-2-3	HU-1	Carbon based biology physical matter-1
4-5-6	HU-2	Carbon-Silica based biology physical matter-2
7-8-9	HU-3	Silica based biology Etheric matter
10-11-12	HU-4	crystalline liquid-light based biology Pre-Matter
13-14-15	HU-5	standing wave pattern flame or fire body ANTE-Matter

2. Density/ Biology: Progression of Matter Density and related Biology within the 15-Dimensional Scale

3. Harmonics of Manifestation:
Progressive expansion of consciousness into the Dimensional Scale creates progressive perceptual De-manifestation of matter Density within the Hologram of external reality. **Harmonics of Manifestation** are the stages of Matter Density and Vibration-Oscillation Consciousness perceives and passes through within the Dimensional Scale.

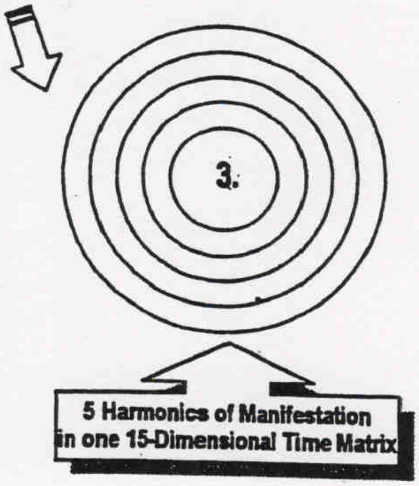

5 Harmonics of Manifestation in one 15-Dimensional Time Matrix

UNIVERSAL AND PLANETARY KATHARA

Kathara Level-1: 12-Tree, Kathara Level-2: Crystals, Kathara Level-3: Diodic Points, Kathara Centers and Kathara Lines.

The Kathara Grid is the *primary mathematical-geometrical organization of units of consciousness* upon which Partiki units group to form morphogenetic field scalar grids. It is the CORE level of scalar standing-wave creation and energetic organization within and behind all dimensionalized systems, and is thus considered to be the **Core of the Holographic Template** upon which the morphogenetic scalar wave blueprint and all other dimensions of form anatomy are built. The Kathara Grid is the causal element within all manifest affects of dimensionalization and consciousness.

The form of the Kathara Grid is reflected in the Macrocosm and Microcosm of all manifestation. All forms have at their core the *common structure of the Kathara Grid Holographic Template.* The Kathara Grids of all forms - universes, galaxies, planets and species - are linked to each other and to the Kathara Grid of the Cosmos through the interwoven Partiki scalar standing-wave grids of the *Kathara Grid Holographic Core.* **The *Cosmic Kathara Grid* represents the scalar field of the Cosmic Unified Field of Consciousness-Energy** through which the *Hologram of manifest experience* is perpetually created.

Energy cannot be created or destroyed, it only changes form. Energy IS *Eternal Consciousness* that perpetually changes form by projecting through the structures of the Kathara Grid, while simultaneously remaining *always the same* - the *singular wholeness* of the **Cosmic Unified Field of Consciousness-Energy.**

Structure of the Kathara Level-1 12-Tree Grid Core Scalar Template

12 PRIMARY KATHARA CENTERS:

Kathara Centers are condensed Scalar Wave points that hold the core electro-tonal geometrical-mathematical program for dimensionalized structure. There are 12 Kathara Centers, each holding the program for one Dimension. The program for Partiki Phasing and Vibration-Oscillation for each Dimensional Frequency Band is set in the corresponding Kathara Center. The Kathara Centers set the template for each Dimension, to ground consciousness within the Holographic experience of manifestation.

15 PRIMARY KATHARA LINES:

Kathara Lines are the primary sequences of Partiki Phasing that transmit the electro-tonal programs held within the Kathara Centers from one Dimension to another, to keep the continual flow of consciousness circulating throughout the 15-Dimensional levels of the Morphogenetic Field. When functioning properly, the Kathara Lines connect each Harmonic and Dimension of consciousness to the other and govern the synchronization of cycles of time within the Dimensional Fields and within the morphogenetic field, DNA and body rhythms of dimensionalized forms. **The 3 Vertical Kathara Lines** control the operation of the other 12 Kathara Lines. **The Central Vertical Kathara Line** is the control center for Partiki Phasing rhythms throughout the Kathara Line System.

3 Vertical Kathara Lines = 1-2-3
Central Vertical Kathara Line = 1

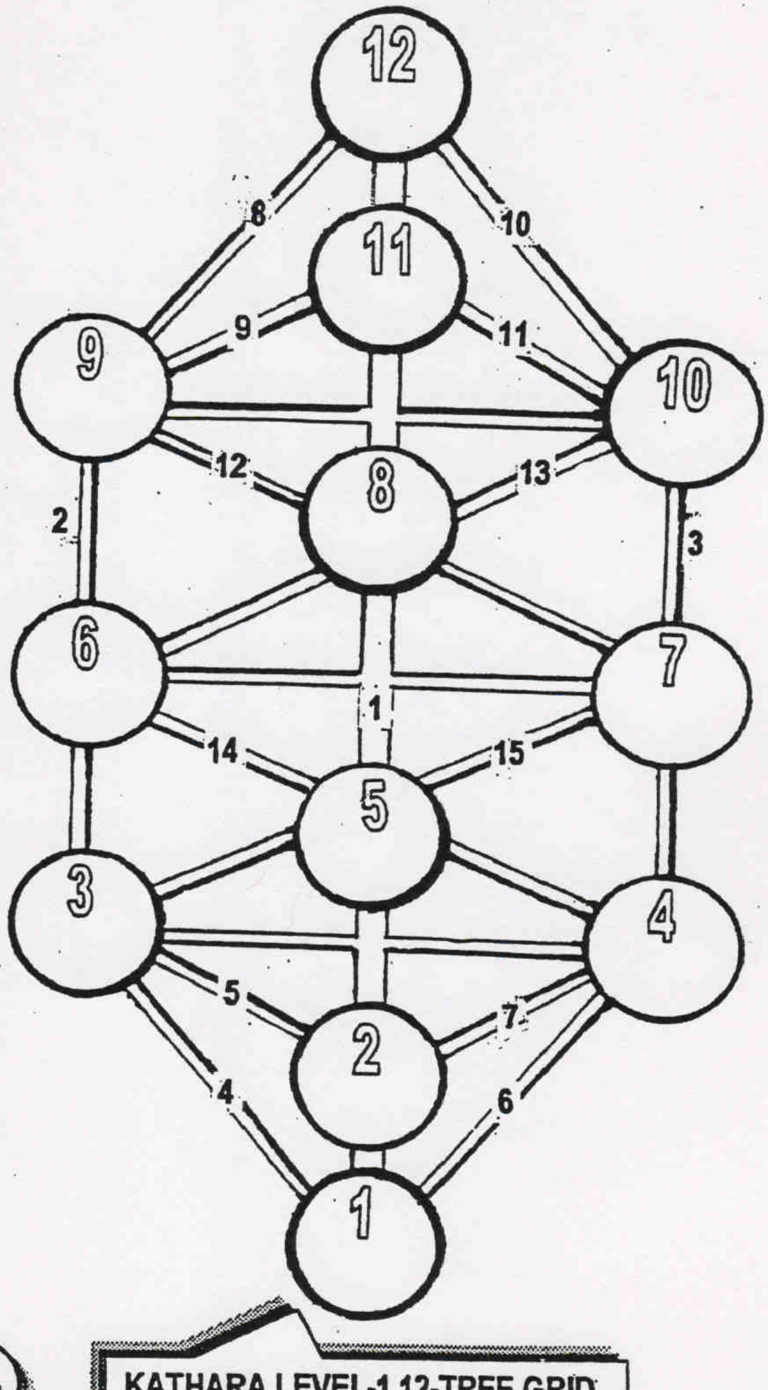

KATHARA LEVEL-1 12-TREE GRID

The 2 Interwoven Kathara Grids of Primal Order within the Universal Time Matrix and Cosmic Energy Matrix of the PRIMAL MANIFESTATION TEMPLATE

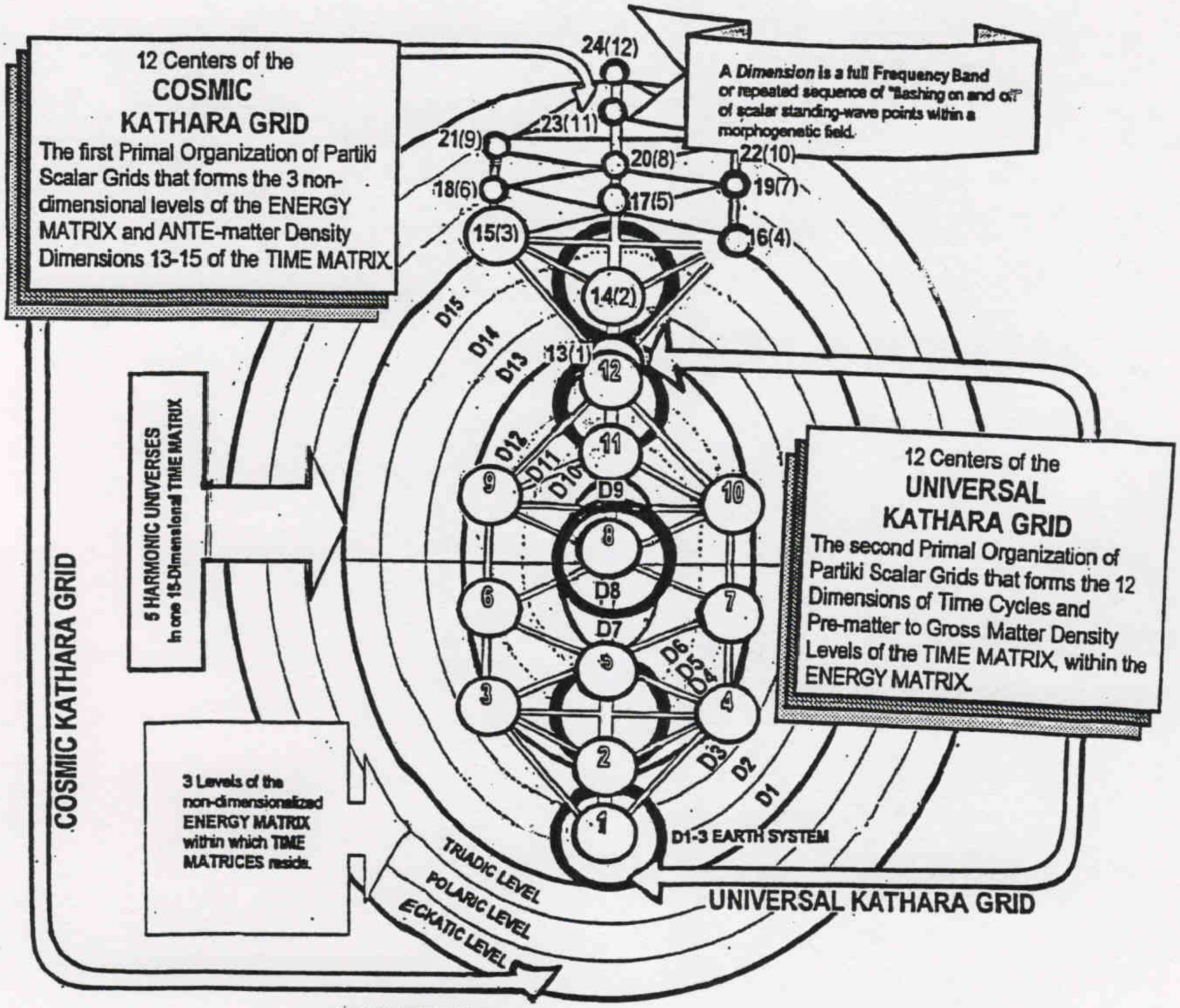

12 Centers of the COSMIC KATHARA GRID
The first Primal Organization of Partiki Scalar Grids that forms the 3 non-dimensional levels of the ENERGY MATRIX and ANTE-matter Density Dimensions 13-15 of the TIME MATRIX.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

12 Centers of the UNIVERSAL KATHARA GRID
The second Primal Organization of Partiki Scalar Grids that forms the 12 Dimensions of Time Cycles and Pre-matter to Gross Matter Density Levels of the TIME MATRIX, within the ENERGY MATRIX.

5 HARMONIC UNIVERSES In one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized ENERGY MATRIX within which TIME MATRICES reside.

24(12)
23(11)
21(9)
20(8)
22(10)
18(6)
17(5)
19(7)
15(3)
16(4)
14(2)
13(1)
12
11
9
10
8
6
7
5
3
4
2
1
D15
D14
D13
D12
D11
D10
D9
D8
D7
D6
D5
D4
D3
D2
D1
D1-3 EARTH SYSTEM
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
UNIVERSAL KATHARA GRID
COSMIC KATHARA GRID

5 HARMONICS OF MATTER DENSITY IN THE TIME MATRIX

1. Gross-matter D1-D3
2. Semi-etheric D4-D6
3. Etheric D7-D9
4. Pre-matter D10-D12
5. Ante-matter D13-D15

= ○

There are 3 **Primary Levels** to the structure of the Kathara Grid. **Level-1: *The Kathara 12-Tree*, Level-2: *The Crystal Seals* and Level-3: *The Diodic Points*.**

Level-1:The 12-Tree holds the first mathematical-geometrical program upon which Kathara Grid Levels 2 and 3 are created, and is composed of the **Kathara Centers** and **Kathara Lines**.

Level-2: The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of *Partiki Phasing* to create the base structures upon which *dimensionalization* is formed. The flow of Frequency (sequences of scalar standing-wave flashing/ cycles of Partiki Phasing) between dimensional bands and Harmonic Universes, the fixed expansion and contraction rates of Partiki, the Vibration-Oscillation and the Angular Rotation of Particle spin are all regulated by the Crystal Seals.

Level-3: Diodic/Miodic Points represent **concentrated areas of energy** that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating **minute vortices** of wave spectra that serve as *frequency modulation zones* (scalar flash-cycle transition points).between Dimensional Frequency Bands. **Diodic Point** vortices are **White Hole** vortices that **transmit energy** into the *Particle Universe* from their counterpart **Miodic Point** vortices within the *Anti-particle Universe*. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* – sets of **White and Black Holes** – that regulate the flux of consciousness-energy between systems.

The entire cosmic structure is built upon the morphogenetic foundations of *interwoven Kathara Grids*. Within the structures of forms, such as planets or people, the Kathara Grid is the *Core Template of Holographic Manifestation* and controls the structure and function of all of the other *subtle and physical energy systems.*

Planetary scalar templates - **Planetary Shields -** vortices, electro-magnetics, Axiom and Ley Lines are all governed by the operations of the *Kathara Grid*. In the **human body**, personal scalar templates - **Hova Bodies -**, Chakras, Axiom and Meridian Lines and Bio-energetic Field Auric Levels, as well as the physical systems, conscious awareness and DNA, are controlled by the *personal Kathara Grid*. **In terms of generating the *Imprint for Health* within the human organism, as well as *reclaiming dormant potentials of Immortality*, the Kathara Grid is the *causal mechanism* through which such affects can occur.**

KATHARA LEVEL-2 CRYSTAL SEALS GRID

Level-2: The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of *Partiki Phasing* to create the base structures upon which *dimensionalization* is formed. The flow of Frequency (sequences of scalar standing-wave flashing/ cycles of Partiki Phasing) between Dimensional Bands and Harmonic Universes, the fixed expansion and contraction rates of Partiki, the Vibration-Oscillation and the Angular Rotation of particle spin are all regulated by the Crystal Seals.

Planetary Star Crystal Seals

In the ***Planetary Kathara Grid***, the **15 Primary Star Crystal Seals** exist *between* each planetary Vortex or Chakra, on the **Central Kathara Line** or **Central Planetary Axis.** The Star Crystal Seals serve as **"Frequency Seals"** that open or close the *flow of wave spectra* and *Frequency Bands* from one **Harmonic Universe** to another, by regulating and modulating the *sequences of Partiki Phasing and the Vibration-Oscillation rates of Partiki units* within the Dimensional Scale. They control the opening and closing of the *Central Vertical Kathara Line* and *Central Planetary Axis*, through which particles are *phase-locked* into their respective dimensional placements and linear time cycle orientations. Opening of the *Star Crystal Seals* releases the planetary particles and *Ley Lines* from their *dimensional phase-lock*, opening **Inter-harmonic Star Gates** (***wave spectra vortices***) within the Ley Lines, for ***molecular transmutation and time continuum shift***, through blending the Inter-Harmonic wave spectra of the planetary morphogenetic field. **The Planetary Star Crystal Seals control the opening and closing of the 3 *Vertical* Kathara Lines (and Vertical** *Star Gate passageways*) **of the Level-1: 12-Tree Kathara Grid with in planetary body.**

Biological Star Crystal Seals

In the ***human body*** the **15 Primary Star Crystal Seals** exist *between* each Chakra on the **Central Kathara Line** or **Central Body Current.** The *biological Star Crystal Seals* serve as **"Frequency Seals"** that open or close the *flow of wave spectra and Frequency Bands* from one *Harmonic of consciousness* to another, by regulating and modulating the *sequences of Partiki Phasing and Vibration-Oscillation rates of Partiki units* within the Dimensionalized scalar wave grid of the morphogenetic field. They control the opening and closing of the *Central Vertical Kathara Line* and *Central Body Current*, through which particles are *phase-locked* into their respective dimensional placements and linear time cycle orientations. Opening of the *biological Star Crystal Seals* releases the *body particles* and **DNA** from their *dimensional phase-lock*, opening **Inter-Harmonic Star Gates** (*wave spectra vortices*) within the **DNA**, for **cellular transmutation (via** particle+anti-particle *fusion*) and time continuum shift, through blending the Inter-Harmonic wave spectra. **The Biological Star Crystal Seals control the opening and closing of the 3 *Vertical* Kathara Lines** (and **Vertical** wave spectra vortices) **of the Level-1: 12-Tree Kathara Grid within the biological form.**

Seed Crystal Seals

Along with the 15 Primary *Star Crystal Seals* that control the *Vertical* modulation of wave spectra, there are also **15 Secondary Crystal Seals** that control the **HORIZONTAL** Kathara Lines of the Level-1: 12-Tree Kathara Grid within the planetary and biological bodies. The *Horizontal Crystal Seals* are called **SEED SEALS**, they exist and serve as **"Frequency Seals"** ***within*** the planetary vortices and biological Chakras, modulating the wave spectra between Parallel (particle/ anti-particle) Dimensional Frequency Bands.

KATHARA GRID LEVEL-3 DIODIC POINTS

Level-3: Diodic/Miodic Points represent **concentrated areas of energy** that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating **minute vortices** of wave spectra that serve as *frequency modulation zones* (scalar flash-cycle transition points) between Dimensional Frequency Bands. **Diodic Point** vortices are **White Hole** vortices that **transmit energy** into the *Particle Universe* from their counterpart **Miodic Point** vortices within the *Anti-particle Universe*. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* – sets of **White and Black Holes** – that regulate the flux of consciousness-energy between systems.

Planetary Diodic Points

Within the **planetary body** and morphogenetic field, **Diodic Points** are located in **geographical locations** where Inter-dimensional and Inter-harmonic Horizontal and Vertical Kathara Lines *intersect* to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve as passageways – **time portals**- between Dimensional Time Continua or Harmonic Time Cycles. They connect planetary and stellar bodies to their *anti-particle counterparts* and intersect with Vertical Inter-harmonic Star Gates and Horizontal Inter-dimensional Time Portals, through sets of Diodic-Miodic **White and Black Holes** at the center of planetary and stellar cores.

Biological Diodic Points

Within the **biological form** and morphogenetic field, Diodic Points are located in **physical body regions** where Inter-dimensional and Inter-harmonic Horizontal and Vertical Kathara Lines *intersect* to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve as ***gateways of consciousness*** between Dimensional Time Continua or Harmonic Time Cycles. They connect physical bodies to their *anti-particle counterparts* and intersect with the Vertical and Horizontal *relationships of wave spectra within the DNA/ RNA and bio-energetic field*, through sets of **White and Black Holes** at the center of the biological Harmonic scalar grids.

Diagonal Grid Control

The *Kathara Grid Level-3: Diodic Points* control the **Diagonal Kathara Lines**, Diagonal *Inter-dimensional portals*, Diagonal *Planetary Ley Lines* and Diagonal relationships between inter-dimensional wave spectra within the biological DNA. **In *Kathara Healing*, the *Diodic Points* are used to open and close the *Diagonal Kathara Lines* for channeling higher dimensional frequency into the Level-1 Kathara Grid, to regenerate the Imprint for Health and to revitalize the Bio-energetic and physical systems.**

THE UNIVERSAL TEMPLAR COMPLEX

- The systems of interwoven Manifestation Templates within the 5 Matter Density Levels of the 15-Dimensional Universal Kathara Grid are collectively referred to as the **UNIVERSAL TEMPLAR COMPLEX.** The Manifestation Template of the human body, likewise structured upon this Primal Order, also manifests through a personal **INTERNAL TEMPLAR COMPLEX** of interwoven 15-Dimensional Manifestation Templates.

- Universal, Galactic and Planetary Scalar-wave Templates of the Universal Templar Complex are called **Scalar SHIELDS.** Shields are sets of 3-Dimensional Scalar Grids upon which the 5 Matter Density Levels of the Universal Manifestation Template manifest. A Shield is composed of the distinct 3 **Primary Flash Line Sequences**, or Partiki Phasing Rhythms, characteristic to the 3 dimensions of which the Shield is composed. The Internal Templar Complex of the human body is also organized into 5 sets of 3-Dimensional **Personal Scalar Shields**, through which the matter form and 5 Matter Density Levels of the 15-Dimensional human consciousness are projected into manifest experience.

- Through the intrinsic dynamics of synchronistic Partiki Phasing, each 3-Dimensional Scalar Shield manifests a **3-DIMENSIONAL SPHERICAL ELECTROMAGNETIC DOMAIN** that creates the energetic framework within which externalized manifest experience can take place. Within the Universal Templar Complex, these Spherical Domains form 5 separate yet interwoven Reality Fields or **Harmonic Universes**, each with a different level of Matter Density, rate of Vibrational Oscillation, different Angular Rotation of Particle Spin and variance in Time Cycle orientation. Within the Personal Internal Templar Complex, the Spherical Electromagnetic Domains created by synchronistic, dimensionalized Partiki Phasing form within the 15-Dimensional Time Matrix 5 separate yet interwoven **STATIONS OF CONSCIOUSNESS** called **HOVA BODIES.** The **5 Hova Bodies** of human Manifestation Template Anatomy within the Time Matrix represent the **5 Subtle Energy Bodies** of human bio-energetic field construction, through which human identity experiences series of **SIMULTANEOUS INCARNATIONS** within the various Time Cycles and matter density levels of the Time Matrix. **Human Evolution** is the process by which the electromagnetic frequencies and stations of consciousness of the 5 Hova Bodies progressively merge to transmute the physical matter form backward through the various Matter Density stages through a process called **TRANSMUTATIONAL DIMENSIONAL ASCENSION.**

  Each of the 5 Shields within a 15-Dimensional Manifestation Template is structured upon **3 KATHARA CENTERS** within the core Primal Order of the Level-1 Kathara 12-Tree Grid. Each Kathara Center of the Universal Kathara Grid forms a core point of consolidated frequency called a **SIGNET Star Crystal Seal** within the 3-Dimensional scalar Shields. Within the universal, galactic and planetary Templar Complex the Signet Seals operate as **Star Gates** between the Time Cycles of the 5 Matter Density levels of the 15-Dimensional Time Matrix. Within the Personal Internal Templar Complex, each **Hova Body** is structured upon **3 Kathara Centers** within the embodied Personal Kathara Grid. Each embodied Kathara Center forms a **Signet Star Crystal Seal** within the human anatomy. The embodied **Signet Seals** operate **as Bio-electric Windows** within the **DNA-RNA**, that link the Stations of Consciousness and various **Simultaneous Incarnations** of the identity that exist within the Time Cycles of the Time Matrix to each other, across the illusions of time. Through the embodied Personal Kathara Grid, Scalar Shields, Signet Seals and DNA Windows, the individuated aspects of **ETERNAL IDENTITY** in time are indelibly linked in consciousness. This intrinsic linking of personal consciousness within the various simultaneous cycles of time appears to the embodied human awareness as **REINCARNATIONAL MEMORY.**

The embodied Manifestation Template of the human form is a microcosmic reflection of the Macrocosmic Universal and Cosmic Manifestation Templates. Both microcosmic and macrocosmic Manifestation Templates share, and are interwoven through, the same **Primal Structures** of the Kathara 12-Tree Scalar Grid, the 5 3-Dimensional Scalar Shields and Signet, Star and Seed Crystal Seals and the Diodic Point "Black and White Hole Sets" of energy vortices that manifest upon the Kathara Grid. Human consciousness and biology are directly and indelibly connected to the Universal Templar Complex through the **Personal Internal Templar Complex** of the Personal Manifestation Template.

HEALING AND PRIMAL FULFILLMENT

A rudimentary but operational understanding of the intrinsic structure of the embodied Personal Manifestation Template- (*the **Level-1**Personal Kathara 12-Tree Grid, **Level-2** Scalar Shields, Hova Bodies and Signet, Star and Seed Crystal Seals, and **Level-3** Diodic Points*) - and its relationship to the Universal Templar Complex, allows the embodied human consciousness the opportunity for progressive participation within the processes of **CONSCIOUS EVOLUTION** and intentional restoration of the **Organic Imprint for Health.** All manifest **DIS-EASE** originates through **distortions** in the electro-tonal scalar wave programs of the **Kathara Grid, Scalar Shields** and **Manifestation Template.** To achieve true and lasting healing, rather than suppression and transference of dis-ease to other embodied systems, distortions within the personal Manifestation Template need to be realigned from the causal Kathara Grid level through **REGENESIS** of **Primal Order** within the **Organic Imprint for Health.**

Manifestation Template healing therapeutics based upon knowledge of the Kathara Grid, simultaneously work to restore the natural intended **Primal Order** and function of the human body and consciousness, while expediting the evolutionary process of merging the stations of identity through time, to facilitate fulfillment of the organic **PRIMAL PURPOSE.** The **Primal Purpose** of human evolution is the achievement of **Transmutational Dimensional Ascension** and **Co-creative Mastery** over the Personal Eternal Consciousness and **Internal Templar Complex** within the Holographic Fields of the Time Matrix. Through fulfillment of the **Primal Purpose,** humanity will regain **THE PRIMAL CONDITION** of Conscious AT-ONE-ment with the Universe and the Central Source of Creation, or **GOD.**

The Universal Templar Complex

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

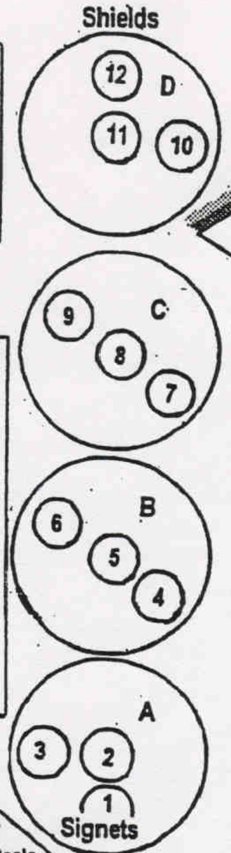

Kathara Level-2 Scalar Shields & Crystal Seals

The 4 Scalar SHIELDS corresponding to the Kathara Center programs and Harmonics of Matter Density within the *Universal* Manifestation Template.(The 5th Scalar Shield in the 15-Dimensional Time Matrix corresponds to the first 3 Kathara Centers of the *Cosmic* Kathara Grid.) Shields are the 3-Dimensional scalar-wave grids that hold the blueprint for manifestation, as set by the mathematical encoding of Partiki Phasing Rhythm held within the Kathara Centers of the Manifestation Template.

HORIZONTAL SHIELDS

A. TELLURIC SHIELD
Gross Matter Shield
Dimensions 1-3

B. DORADIC SHIELD
Semi-etheric Matter Shield
Dimensions 4-6

C. TEURIC SHIELD
Etheric Matter Shield
Dimensions 7-9

D. MAHARIC SHIELD
Pre-matter Shield
Dimensions 10-12

E. RISHIC SHIELD
Ante-matter Shield
Dimensions 13-15
(vertical- not shown)

The 12-Dimensional UNIVERSAL KATHARA GRID within the 15-Dimensional Time Matrix. The 12 Kathara Centers hold the core mathematical and geometrical Partiki Phasing programs upon which the Shields of the Universal Manifestation Template organize.

The 12 SIGNET Star Crystal Seals of the Universal Kathara Grid. SIGNET SEALS are the core points of consolidated frequency corresponding to each Dimension that represent the first manifest arrangement of scalar waves created through the mathematical programs of the corresponding Kathara Centers. The Signets transmit the organizational programs upon which the Scalar Shields manifest. Each 3-Dimensional Scalar Shield manifests on the electro-tonal programs of 3 Signet Star Crystal Seals. Signet Seals keep the Harmonics of Matter Density separate from each other by keeping Partiki units phase-locked into their 3-dimensional electromagnetic domain.

When released, Signet Seals serve as Star Gates, allowing the Time Cycles and Matter Density Levels within the Time Matrix to merge and blend. Signet Seals are the central control points for Scalar Shield programming.

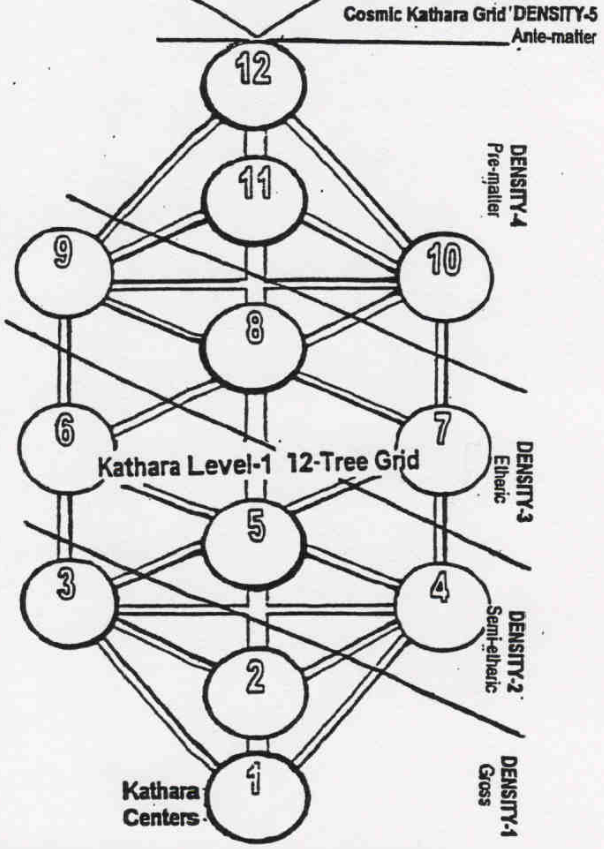

The structure of the macrocosmic **Universal Templar Complex** is replicated within the microcosmic structure of the personal **Internal Templar Complex** upon which **15-Dimensional Human Anatomy** manifests.

The Transduction Sequence

The *Transduction Sequence* represents the phases consciousness passes through into dimensionalization and the process by which conscious energy is transduced and *changes form* to enter the wave spectra within which perception of manifestation is possible.

Basic *Transduction Sequence* of Consciousness into Biological Form.

CONSCIOUSNESS TO **PARTIKI UNITS** TO

KATHARA SCALAR GRIDS LEVELS –1, 2 AND 3 TO **PARTIKI GRIDS** TO

MORPHOGENETIC FIELDS TO **MERKABA FIELDS** (Sets of Counter-rotating electro-magnetic spirals of energy –repeating scalar wave flash/ Partiki Phasing sequences that run between Dimensions and Harmonic Universes)

TO **DNA/ RNA IMPRINT** TO **BIO-ENERGETIC AURIC FIELD** TO **CHAKRAS**

TO **NADIAL CAPSULE** (tissue-like capsule that forms around the 3-Dimensional scalar grid of the Harmonic Universe-1 morphogenetic body.) TO

DNA/ RNA MANIFEST TO **NADIS** (energetic rivulets that feed energy from the Nadial Capsule into the Central Nervous System and molecular structure) TO

NUCLEAR CORE TO

CNS-METABOLIC/ BIOLOGICAL RHYTHMS- BLOOD-BRAIN-CHEMICALS-HORMONES-ORGANS-TISSUES.

HOLOGRAPHIC REALITY CONSTRUCTION

The Unified Field of Conscious-Energy

THE **SCALAR STANDING-WAVE GRIDS** THAT MAKE UP **DIMENSIONALIZED MORPHOGENETIC FIELDS** REPRESENT THE *INTRINSIC STRUCTURE* OF THE **COSMIC UNIFIED FIELD OF ENERGY** WITHIN WHICH ALL THINGS EXIST. *ENERGY IS CONSCIOUS AND CONSCIOUSNESS IS ENERGY.* **CONSCIOUSNESS-ENERGY CANNOT BE CREATED OR DESTROYED IT SIMPLY CHANGES FORM,** WHILE SIMULTANEOUSLY REMAINING ALWAYS THE SAME – **THE ETERNAL UNIFIED FIELD OF CONSCIOUSNESS.**

The Kathara Grid

ALL FORMS OF CONSCIOUSNESS AND PERCEIVABLY MANIFEST FORMS *DERIVE THEIR INDIVIDUATION* WITHIN THE UNIFIED FIELD THROUGH THE STRUCTURE OF THEIR **INDIVIDUATED MORPHOGENETIC FIELD**. THOUGH HAVING MUCH DIVERSIFICATION WITHIN THE **GEOMETRICAL-MATHEMATICAL PROGRAMS** THAT FORM **INDIVIDUALITY** WITHIN LIKE PATTERNS, THE *CORE OF ALL INDENTITY AND MANIFEST FORM IS THE SAME* – THE INTRINSIC STRUCTURE OF **THE KATHARA GRID – THE CORE HOLOGRAPHIC TEMPLATE OF DIMENSIONALIZATION UPON WHICH MORPHOGENETIC FIELDS ARE FORMED.**

Interwoven Morphogenetic Fields

THE INHERENT STRUCTURE OF **INTERWOVEN KATHARA GRIDS** FORMS **SPHERICAL DIMENSIONALIZED MORPHOGENETIC FIELDS** WITHIN WHICH CONSCIOUSNESS CAN EXPERIENCE THE **HOLOGRAPHIC PROJECTION** OF MANIFEST REALITY, SPACE, TIME, MATTER AND EXTERNALIZATION. THE SPHERICAL MORPHOGENETIC STRUCTURES OF THE **MICROCOSM** EXIST WITHIN THE GREATER SPHERICAL MORPHOGENETIC FIELDS OF THE **MACROCOSM**, FORMING A SYSTEM OF CONSCIOUS-ENERGETIC **SPHERES-WITHIN-SPHERES** THAT REPRESENTS THE ***INTERWOVEN MORPHOGENETIC FIELDS* OF THE COSMIC UNIFIED FIELD OF CONSCIOUSNESS-ENERGY.**

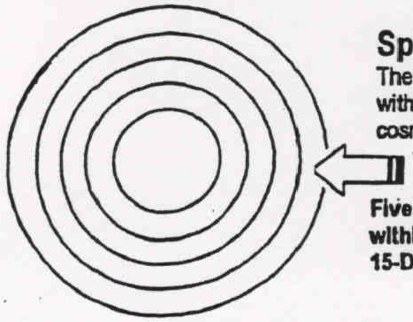

Spheres-within-Spheres ***Interwoven Morphogenetic Fields***
The Microcosm within the Macrocosm. The individual within the species within the planetary within the galactic within the universal within the cosmic within the INFINITE AND ETERNAL.

Five Over-lay Universes exist within the same space in 1 15-Dimensional Matrix.

CHAPTER 1
HIGHLIGHTS
THE HOLOGRAPHIC TEMPLATE

- **External reality is a HOLOGRAPHIC PROJECTION.** Perception of **solidity** and **externalization** are determined by the **oscillation rates** between **dimensionalized units of consciousness** - ***Partiki, Partika*** and ***Particum.*** Dimensionalized units of consciousness form **Frequency Bands** that exist in specific relationships to each other. We perceive as solid the Frequency Bands that exist **1** full Dimensional Band **below** the frequencies within which the consciousness is stationed.

- **Dimensionalization** creates the structure through which consciousness can perceive the **Hologram of Externalization.**

- The human body is a **Hologram** built upon a **Holographic Template** of units of consciousness structured into **Scalar Standing-wave Grids** that represent the **Living Morphogenetic Field** blueprint upon which matter and dimensionalized consciousness manifest.

THE STRUCTURE OF DIMENSIONS

- The **Unified Field** is built upon ***dimensionalized sets*** of ***Scalar Standing-wave Points*****, Wave Strata** formed of units of consciousness that form the basis of energetic relationship through which the **Hologram of Manifestation** can be perceived.

- The human body is a construct of ***Conscious Energy***, built upon **a15-Dimensional Anatomy of Dimensionalized Scalar-Grids.** Perceived physicality emerges through the **Perceptual Lens of Dimensionalization** which provides the picture of seemingly manifest particles, atoms, molecules and physical systems.

- Understanding the nature and structure of **Holographic Reality** allows us to understand the foundations upon which the **body is** built and the **context** within which **healing** must occur.

Scalar Waves are standing wave points that emanate out of fixed points of sound (electro-tonal Partiki units), which form the **Scalar Grid morphogenetic templates** that house consciousness in **dimensionalization** for experience of manifest reality. Scalar Grids are created through **Partiki Phasing.**

Partiki Phasing is the process of ***internal fission/ expansion and fusion/ contraction*** characteristic of **Partiki Units**, through which the units perpetually break apart and replicate through *fission*, then re-merge through *fusion*. Creates a *"flashing on and off" of **Scalar Wave points**,* or "**Flash-line Sequences"** that determine the intrinsic orientation of consciousness within the Scalar Grids composing morphogenetic fields.

Dimensions are Scalar Grids of consciousness composed of interwoven layers of Scalar waves with **varying Flash-line Sequences**, or *Frequency Bands.* **Dimensions are Frequency Bands**, or Partiki wave spectra, that exist within specific relationship to each other to form the template of Holographic Reality.

Frequency Bands are synchronized, repeating ***Flash-line Sequences of Partiki Phasing*** that give the appearance of manifest movement of wave spectra within a Dimensional Scalar Grid **Sub-frequency Bands** are *portions* of the *Flash-Line Sequences* that make up the whole Dimensional Frequency Band. The Frequency of a Dimensional Band is determined by the **Vibratory-Oscillation** rate of Partiki units that make up the Dimensional Scalar Grids.

Vibration is the "internal movement" or **ENERGY HOLDING** of Partiki units. **Oscillation** is the "external movement" or **ENERY EXPENDED** by Partiki units. **Vibratory-Oscillation rate is the rate at which Partiki Phasing "flashing on and off" occurs**. Vibration and Oscillation exist in **direct proportion** to each other. As Vibration increases (energy holding), Oscillation decreases (energy expenditure). Faster Flash-line Sequences (higher frequency) *expend* more energy, raise Oscillation and lower Vibration of Scalar Templates. **Higher Oscillation/ Slower Vibration creates slower electron movement around nucleus and less perceivable density of matter.**

- A **Dimension** is a full Frequency band or **full cycle** of a **Flash-line Sequence**.

- The **Universal Morphogenetic Template** is built upon ***sets of 15 Primary Flash-line Sequences*** that form **15-Dimensional Time Matrices** within which manifestation can be experienced.

- **Dimensions are organized into sets of 3, (**3 full Flash-line Sequences/ 3 Frequency Bands), that create the template through which 3-dimensional manifestation can be experienced. Each set of 3 Dimensions is called a **Harmonic Universe**. There are **5 Harmonic Universes** within a **15-Dimensional Time Matrix.**

- Each Dimension contains **12 Primary Sub-frequency Bands** (partial Flash-line Sequences).

- Partiki in each Dimension have specific set **Partiki Phasing rhythms** giving each Dimension a specific ratio of **Vibration-Oscillation.** Higher Dimensions have Faster Partiki Phasing, are higher frequency and thus have higher oscillation, lower vibration and less particle density.

- **Vibratory-Oscillation** rate within a Dimension determines the **Angular Rotation of Particle Spin (ARPS)**, or *rotational axis of particles* in a Dimension. There is a **90-Degree Shift** in ARPS between each Dimension and a **45-Degree Reverse Shift** in between Harmonic Universes.

- Particles having varying **Vibratory-Oscillation** rates and **ARPS** allow *multiple 3-Dimensional Reality Fields* to *co-exist* within the same space while remaining perceptually and experientially **invisible** to each other.

- The **relationship between wave strata** and **variance in Partiki Phasing Flash-line Sequences** within Dimensional Frequency Bands create **Holographic Refraction** of consciousness-energy that allows perception of an externalized, 3-Dimensional **Holographic Illusion** for consciousness focused within the Scalar Grids of the 15-Dimensional Time Matrix.

UNIVERSAL AND PLANETARY KATHARA

- The **Kathara Grid** is the **first level** of Partiki structure/ *Scalar Grid organization* that sets consciousness into dimensionalization. It is thus the **CORE HOLOGRAPHIC TEMPLATE** upon which the *morphogenetic field blueprint* is structured.

- The *Kathara Grid* is the *Core Holographic Template* for **dimensionalization of the Unified Field** of Consciousness-Energy through which the **Microcosm** and the **Macrocosm** are structured. **All things are built upon the foundation of the *Kathara Grid.***

- There are **3 Primary Levels to the Kathara Grid. Level-1:** ***The 12-Tree Grid*** with **Kathara Centers** and **Kathara Lines. Level-2:** ***The Crystal Seals*** with **Hova Bodies** and **Scalar Shields** and **Level-3:** ***The Diodic Points*** of minute energy vortex "**White and Black Hole" sets.**

KATHARA GRID LEVELS

KATHARA 12-TREE GRID

- **Level-1 Kathara** holds the base electro-tonal **programs** of Partiki Phasing and Dimensional/ Frequency structure upon which the other Kathara Levels and the morphogenetic field are built.

KATHARA SCALAR SHIELDS AND CRYSTAL SEALS

- **Level-2 Kathara Grid** is composed of 5 groups of 3-Dimensional **Scalar Shields** **and** **Crystal Seals** - that regulate the rates of Partiki Phasing to create the base structures upon which ~~dimensionalization is built.~~

KATHARA DIODIC POINTS

- **Level-3 Kathara Grid** is made of sets of minute White and Black Hole energy **vortices,** made of inter-dimensional Scalar Wave spectra, that **regulate the flux of consciousness-energy** between Dimensional, Harmonic and Parallel/ anti-particle systems.

- ***Cosmic*** *morphogenetic structure* is built upon interwoven **Kathara Grids** and the Dimensionalized Scalar Standing-wave Templates they form.

- The **5 Primary Scalar Templates of Dimensionalization** are formed on **Scalar Shields** composed of sets of 3 Dimensions, which form *5 3-dimensional spheres of energy* called **HOVA BODIES.** Hova Bodies and their Scalar Shields hold consciousness-energy into **5 separate but interwoven 3-dimensional systems.**

- The **Kathara Grid** and the multiple levels of morphogenetic structure built upon it is the **CAUSAL ELEMENT** behind and within all manifestation, and thus the **CORE HOLOGRAPHIC TEMPLATE** through which the **Imprint for Health** can be restored to Biology.

- All manifest **Dis-ease** originates through **distortions** in the natural electro-tonal programs/ *Partiki unit arrangements* of the *Kathara Grid* and *Hova Body Scalar Shields*. The Kathara Grid and morphogenetic field are the **core elements** through which **lasting healing** is created. **Healing** occurs through **REGENESIS** of natural electro-tonal **programs** of the **Holographic Template.**

- **Partiki Units, Scalar Standing Waves, and Crystalline Keylons and Keylon Codes** represent the **PRIMAL SUBSTANCE** in energy that consciousness takes on in order to create the **Holographic Projection** of external manifestation.

- The **Kathara Grid, 15-Dimensional Structure, Scalar Shields, Crystal Seals, 5 3-Dimensional Harmonic Universes and Hova Bodies, Diodic Point** vortices and the subtle and dense matter energetic systems that manifest upon them represent the **PRIMAL ORDER** of mathematical and geometrical interrelationship that **Primal Substance** utilizes to orchestrate the manifest Hologram through creation of **Manifestation Templates.**

- The interwoven, ordered structures of scalar wave fields that form the **Universal Manifestation Template** of the 15-Dimensional Time Matrix are collectively referred to as the **UNIVERSAL TEMPLAR COMPLEX.**

- The interwoven, ordered structures of scalar wave fields that form the **Personal Manifestation Template** of **15-Dimensional Human Anatomy** are collectively referred to as the **Personal INTERNAL TEMPLAR COMPLEX.** The Internal Templar Complex, a structural replica of the Macrocosmic Universal Templar Complex, anchors personal consciousness into the Time Matrix for the experience of the space, time and matter Hologram.

- The **PRIMAL PURPOSE** of human evolution is the achievement of **Transmutational Dimensional Ascension** and **Co-creative Mastery** over the **Personal Eternal Consciousness** and **Internal Templar Complex** within the holographic fields of the Time Matrix. **REGENESIS** of **Primal Order** within the **Organic Imprint for Health** creates true healing and expedites the natural processes of human evolution.

- Through fulfillment of the **Primal Purpose** humanity will regain **THE PRIMAL CONDITION** of **Eternal Conscious At-ONE-ment** with the Central Source of Creation or **GOD.**

CHAPTER 2

(Section One: *Foundations of Kathara Healing* – Part II)

THE ELEMENTS OF HEALING

The Human Body Kathara and the Multi-dimensional Anatomy of the Human Form

- THE MAHARIC SHIELD AND THE IMPRINT FOR HEALTH
- THE PLANETARY BIO-FEED INTERFACE SYSTEM (PBIS)
- THE HIEROPHANT SYMBOL CODE, PBIS AND THE MAHARIC SHIELD
- MORPHOGENETIC EXPANSION, GROWTH CYCLES, DEATH AND MOLECULAR COMPACTION
- LEVELS OF HUMAN ANATOMY, EMBODIED KATHARA 12-TREE GRID, CRYSTAL SEALS, SHIELDS, SIGNETS, HOVA BODIES, AURIC CAPSULES, DIODIC POINTS, SEED SEALS, CHAKRAS, AXIOM LINES, THE HARA LEVEL AND DNA.
- THE 12 HUMAN SENSES, APPARTHI AND HUMAN PERCEPTION

TOPICS OF DISCUSSION

THE HUMAN BODY KATHARA AND THE MULTIDIMENSIONAL ANATOMY OF HUMAN FORM

The Structures of Manifestation and Reclaiming Our Power

In *PART-I* of this program, **THE CONTEXT OF HEALING**, we explored the hidden foundations of *reality structure* within which our existence takes place, the *context of relativity* through and within which *healing* occurs. We learned that the context of existence can be viewed as taking place within a **Unified Field of Energy**, and that **Energy IS consciousness and Consciousness IS energy.** In our exploration we encountered the **Partiki, Partika** and **Particum**, the ultra-micro-particle units of consciousness in electro-tonal form, through which the tangible structure of living **Morphogenetic Fields** are formed. In learning the foundation structures of the Unified Field of Consciousness-Energy, we come to understand that the anatomy of any form, including that of the human Body-Mind-Spirit System, is a **multi-dimensional construct** built on an ordered template of **sets of 15-Dimensions,** through which all conditions of apparent manifestation are created.

Through reviewing the underlying structures of manifest reality (units of consciousness, the ***Partiki Scalar Standing-wave Grids*** that hold consciousness into individuated dimensionalized "packages" of awareness to allow for experience of space, time, matter and form, the core ***Kathara Grid Holographic Template*** that sets the pattern of dimensionalization for consciousness to follow into the manifest experience, the ***Levels of Matter Density*** corresponding to the ordered structure of Dimensions, and the reality that **Dimensions** are *tangible constructs of energy* created through the process of ***Partiki Phasing***-the fission-fusion "flashing on and off" of scalar standing-waves composed of units of consciousness) **we have already expanded our perceptions of the nature of reality itself.** With these **simple revelations of knowledge,** we have shifted ourselves from a perspective of viewing the reality before us as one of *separate manifestations disconnected from our being*, to a perspective of viewing reality as an *intricate system of interwoven consciousness, expressing in the form of standing-wave strata*, to which we are *intimately and indelibly connected.*

Suddenly we find that we cannot be "left out" of things, we cannot "be on the outside as life passes us by", **that there IS no "outside" to which we may stand separate, as by the *nature of our existence* we are an *intrinsic element and force of creative affect* within the whole of creation.** Through this knowledge we have begun to tap the **secrets of SELF,** as we shift from viewing ourselves as *mortal creatures trapped within a gross-matter finite physical body*, to **KNOWING OURSELVES AS THE EXPRESSION OF CONSCIOUS, SENTIENT WAVE-STRATA - ETERNAL BEINGS ENSCONCED WITHIN AN ETERNAL UNIVERSE, within which Energy-Consciousness cannot be created or destroyed, it simply changes form perpetually.** We move from perceiving ourselves as finite beings of limited power to affect a world that is separate from us, to knowing that we are *walking wave-forms within a sea of interrelated wave forms*, and that the pattern of our being operates as an intrinsic part of the whole. **As we begin to perceive ourselves in terms of the *unlimited reality of our nature*, rather than define ourselves by the limited boundaries of the *manifest illusion before* us, we begin to reclaim our power to *affect desired change* within the contours of our lives.**

Knowledge, KNOWING and the Imprint for Health

Knowledge is an agent of freedom, IF it is used to foster practical applications of change. Knowledge can also become a *perpetual quest of distraction*, through which we deny ourselves the ability to EMBRACE the experience of living as we seek outside of ourselves the answers that live within. **Knowledge is power if it is applied appropriately, it can be an agent that leads one through the door to the WISDOM that awaits on the other side.** Knowledge alone cannot provide **cognition,** cognition comes through the *application of applied knowledge* to provide **tangible, experiential** KNOWING, a knowing derived only through the **direct experience of BEING. In Kathara Healing™ *knowledge* is applied as a *tool* through which one can grow into the *wisdom of direct experience*** - from ***thinking or believing***, which are knowledge/data-based characteristics of a mind perceiving itself as *finite* - to **KNOWING,** through Direct Cognition of the Experience of BEING, which is a characteristic of a consciousness KNOWING itself as ETERNAL. .

In our first lessons of **Kathara Healing**, we gained *knowledge* of the *Context of Healing* and the *structure* of that which exists within and around us, so that we may become more empowered to serve as a *healing force*, within ourselves and within the collective of the Whole. **But this knowledge will only become practical or useful to us if it is applied to foster the direct experience through which the wisdom of KNOWING may emerge.** Then, and only then, will we be changed...then and only then will we become *true agents of healing*. **Knowledge can help you to *understand the nature of health and healing*, but only through KNOWING HEALTH through *direct experience* will you be *healed*.** To KNOW HEALTH we must rediscover this quality within ourselves, to EXPERIENCE the *quality of wellness* that emerges as the **Imprint for Health** is reawakened within the human body. **First we must use *knowledge* to recognize that the *Imprint for Health* is there, sleeping silently within us awaiting this awakening.**

Healing the Causal Core

Fixing symptoms does not restore the *Imprint for Health*, it simply masks from view the *causal elements* through which non-health manifested. Removing the symptom *without addressing the underlying cause* simply sets the stage for the cause to *re-manifest itself* in new ways, through *new sets of symptoms*. **Each symptom you remove will give birth to another symptom if the *underlying cause* is not changed.** Kathara Healing addresses the **deepest *causal factor*** in manifestation, the **Holographic Template Kathara Grid**, upon which all conditions come into being. **Within the intrinsic structure of the Kathara Grid the dormant *Imprint for Health* is sleeping, and through knowledge of the Kathara and applications of its methods we can reawaken the Imprint for Health within ourselves, to progressively move into KNOWING HEALTH** through the cognition of Direct Experience.

In Kathara Healing you are provided with the knowledge you need to understand the structures upon which the *Imprint for Health* is built, for in understanding these structures you will be enabled to call the Imprint for Health into awakening. **But it is through the APPLICATION OF KATHARA TECHNIQUE that you will AFFECT CHANGE. The technique applications offered in Kathara Healing allow you to "bring the knowledge home" to the core of its usefulness, the ability to lead you into the *direct experience of health-ward change*.**

Elements of Healing

In **PART II THE ELEMENTS OF HEALING,** we will apply the knowledge of the *Context of Healing* that we acquired in *PART I,* directly to its relevance in terms of the **HUMAN BODY**. We will now explore **Kathara Mechanics** as they apply directly to the human system, and you will gain the knowledge of your **15-Dimensional Anatomy** through which you may then effectively utilize **Kathara Mechanics Techniques** to begin your journey into the **EXPERIENTIAL REALITY OF HEALING.** We will discover the long forgotten potentials of the ***Planetary Bio-feed Interface System™*** and how to receive energy for healing directly from the Kathara Grid of the Earth. We will explore the real processes within and behind the experience of growth through time, so that we may finally identify the **true causal element intrinsic to the *degeneration* and *death* of the body.** We will come to understand the functions of the Body-Mind-Spirit System as they are **INTENDED to operate**, to identify the areas in which malfunction is occurring, so that we may *remedy the casual factors* to progressively restore the **INTENDED INTEGRITY OF THE BODY-MIND-SPIRIT SYSTEM. In PART II we will reclaim our knowledge of our *multidimensional anatomy*, and rediscover the buried secrets of awakening the *Imprint for Health*.**

MAHARIC SHIELD

The Imprint for Health and the *Maharic Shield*

Restoration of health within the body is a progressive endeavor that takes time. Portions of the Body-Mind-Spirit System that have been malfunctioning must readjust to a new process of function that will progressively work its way up from the core **Kathara Grid Holographic Template**, into the **Bio-energetic Field, Chakra System** and **DNA**, to progressively ***reset*** the **Imprint for Health** within the Body-Mind-Spirit System. **The Imprint for Health sleeps soundly now within a portion of the multidimensional anatomy called the *MAHARIC SHIELD*.** The Maharic Shield is a buried treasure of **REGENERATION and REVITALIZATION** of the human organism, as it is the key to *recoding the human Kathara Grid* back into its original Imprint for Health. **The *Imprint for Health* is a tangible construct of energy that is located within the structure of the *Maharic Shield*.** In calling these dormant energy systems into activation within the body, one has the opportunity to restructure the causal level of malfunction, and to recode the Holographic Template upon which malfunction manifests.

Reciprocal Reflection and the Kathara Grid

All conditions of dis-ease within the Body-Mind-Spirit System *share a common core of causality - distortions of the Kathara Grid* upon which manifestation is built. All symptoms affect the Body, Mind and Spirit *simultaneously*, malfunction in one aspect of being will have its *reciprocal reflection* in the other aspects of being, as the symptoms emanate from distortion in the *Kathara Core*, upon which all aspects of being simultaneously manifest.

As we work with restoring integrity to the *Holographic Template Kathara Grid*, we are thus creating *simultaneous restoration through Reciprocal Reflection* within the *Body*, the *Mind* and the *Spirit* (the CONSCIOUSNESS that makes up the substance of our multidimensional anatomy and identity).

The Maharic Shield and Whole-Being-Healing

In activating the *Maharic Shield* to regenerate the *Imprint for Health*, you will have change occurring within multiple levels of being. Your **BODY** will begin to reflect change as dis-ease-causing distortions are released from the *biological imprint*, and a higher level of order is reestablished within the ***atomic structure.*** Your **MIND** will begin to reflect change as dis-ease-causing distortions in your **thoughts, ideas and beliefs** are released from the *mental body* awareness. Your **SPIRIT**-consciousness, will reflect change, as divisions between your embodied awareness and the expanded portions of your multidimensional awareness will begin to dissolve, and you will become progressively more able to embody greater states of *cognition, creativity and awareness.* ***Kathara Grid recoding* through the use of the *Maharic Shield* will set in motion the processes of WHOLE BEING HEALING, as the *Imprint for Health* is progressively reawakened and restored within the Body-Mind-Spirit System**. Activation of the **Maharic Shield** within the personal Kathara Grid will lay the foundations for progressive, lasting and powerful restoration of the **Imprint for Health** within your Body-Mind-Spirit System.

Learning to activate the Maharic Shield to restore the Imprint for Health requires a bit of **knowledge** regarding *Elements* of the hidden anatomy of multidimensional being, and application of specific **energetic mechanics** that will directly affect these hidden structures of energy. We will begin PART II with a discussion on the **ELEMENTS OF HEALING as they apply to *developing the skills required to create the condition of Maharic Shield activation*, through which the *Imprint for Health* can receive its wake-up call.**

THE PLANETARY BIO-FEED INTERFACE SYSTEM™

Awakening the Personal and Planetary Templar Complex Connection

Interwoven Personal and Planetary Kathara Grids

The core Holographic Templates of both the *personal* and the *planetary* morphogenetic fields are built upon a Scalar standing-wave *Kathara Grid*. As the morphogenetic fields of **species** are interwoven into the larger morphogenetic field of the **planet** they inhabit, there is a **direct energetic connection between the Kathara Grids of individuals and that of the planetary body. This connection between personal and planetary Kathara Grids is called the *Planetary Bio-feed Interface System*™, - PBIS.** For thousands of years the **planetary Kathara Grid** has carried **distortions** in its natural pattern that have rendered the **PBIS** primarily inoperable. **Kathara distortions in the planetary morphogenetic field blocked the human's innate ability to draw energy for regeneration and revitalization from the planetary morphogenetic field.**

When the human Kathara Grid system is working properly there is a *natural sympathetic rapport* between the personal and planetary Kathara Grids that keeps the *Imprint for Health* within the human *Maharic Shield* functioning properly. Due to the distortions of the planetary Kathara, the human Kathara developed **reciprocal distortions,** which created **blockages between the Maharic Shield Imprint for Health and the other functional portions of the human Kathara.** The blockages with the human Kathara manifested as distortions within the natural function of the **DNA** and bodily systems, disarming many of the natural *self-regenerative functions* organic to human biology and rendering the human form more susceptible to disease and malfunction of the biological system. As there is a **Reciprocal Reflection** of patterns between the Body, Mind and Spirit aspects of the human organism, due to their *common causal element* of the *Kathara Grid Holographic Template*, **distortions within the DNA and physical body function also become reciprocal distortions** within the **mental body** and its **perceptual facilities,** and within the **Spiritual Body,** as the Kathara distortions *block and distort* the organic cycles through which higher dimensional aspects of consciousness would progressively embody within and expand the human form.

The Planetary Bio-Feed Interface System™

As the morphogenetic fields of species are interwoven into the larger morphogenetic field of the planet they inhabit, there is a direct energetic connection between the Kathara Grids of individuals and that of the planetary body. This connection between personal and planetary Kathara Grids is called the *Planetary Bio-feed Interface System™*, - PBIS.

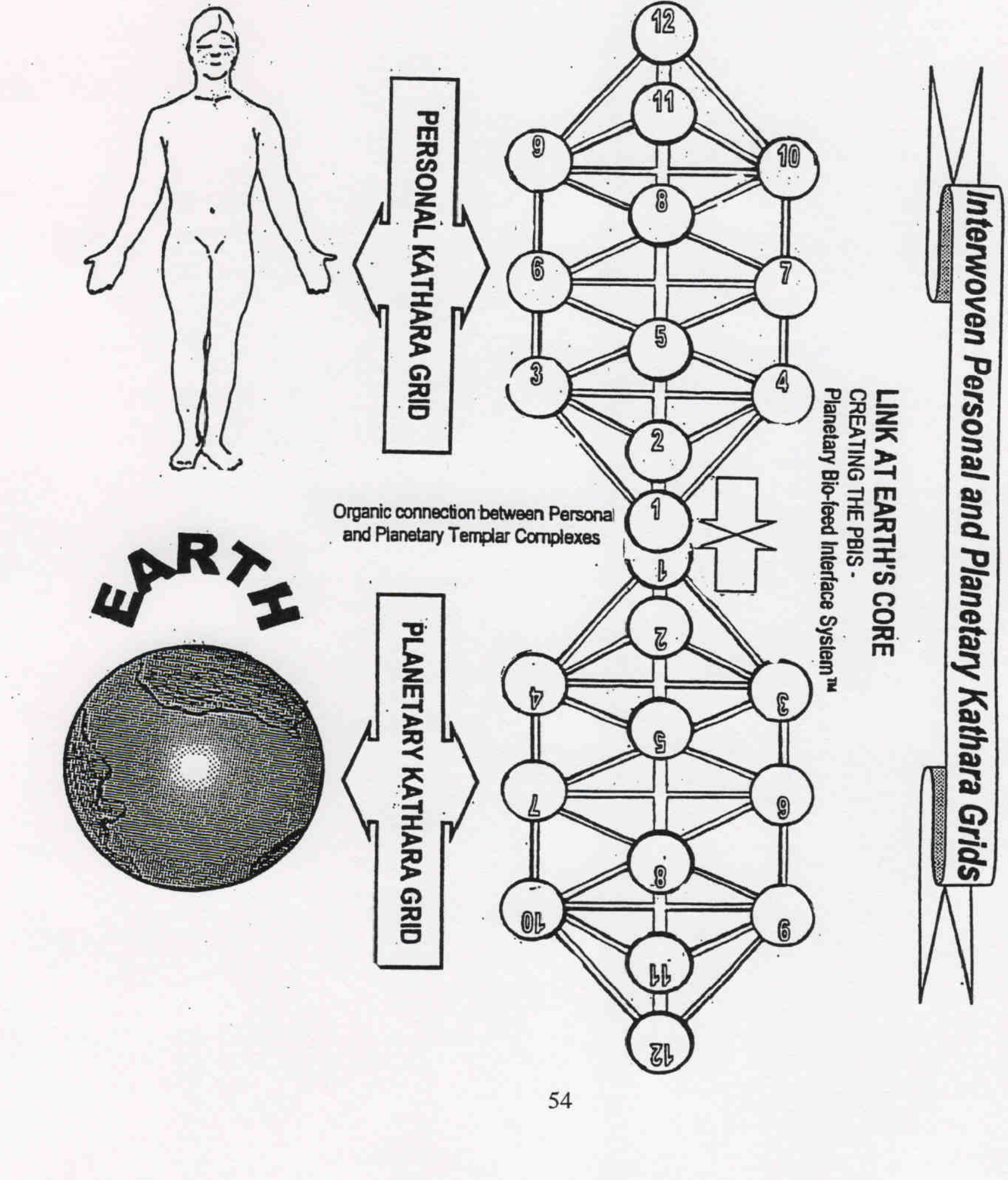
PERSONAL KATHARA GRID
12
11
9
10
8
6
7
5
3
4
2
1
Organic connection between Personal and Planetary Templar Complexes
LINK AT EARTH'S CORE
CREATING THE PBIS -
Planetary Bio-feed Interface System™
Interwoven Personal and Planetary Kathara Grids
EARTH
PLANETARY KATHARA GRID
1
2
4
3
5
7
6
8
10
9
11
12

Return of the PBIS - Gift of the 21st Century

The secrets of utilizing the PBIS have been passed down within the generations of the Melchizedek human families who brought these teachings to Earth long ago, kept hidden from public display until a day when the PBIS could once again become operational through evolution and clearing of the planetary Kathara Grid. These teachings are now being returned into the public arena because **progress has been made in restoring the integrity of the planetary Kathara Grid system. For the *first time in many thousands of years* the potentials of healing held within the PBIS can be returned to humanity, and the PBIS can now be *activated within the personal biology* to reawaken the *Maharic Shield* and its *Imprint for Health*, within the human Kathara System.** Reawakening the *PBIS* and the *Maharic Shield* within the human body *allows the personal Kathara Grid to open into the planetary Kathara Grid*, through which the now-functional **Holographic Template of the planetary Kathara can be used to RESET the original imprint of function within the personal Kathara,** for accelerated regeneration of the **Imprint for Health** within the human form.

HIEROPHANT SYMBOL CODE

The *Hierophant Symbol Code* PBIS Trigger

The first application of the PBIS that we will employ in Kathara Healing is utilizing the energies flowing through planetary Kathara to reawaken the Maharic Shield within human anatomy. The process of opening the PBIS within the personal morphogenetic field involves using a **specific scalar standing-wave-guide GEOMANCY - a consolidated inter-dimensional *frequency pattern* that triggers release of the *frequency seals* (Kathara Level-2 Crystal Seals) on the presently dormant 10th, 11th and 12th Chakras, opening the human Bio-energetic Field and Kathara Grid to the Kathara Grid, Bio-energetic Field, Ley Lines and energy Vortices of the planetary morphogenetic field.** The wave-guide Geomancy containing the frequencies to activate the PBIS and the Maharic Shield can be directed by the human consciousness **in the form of a visually created symbol code. This symbol code is called the HIEROPHANT.** In **TECHNIQUE # 2: *The Maharic Seal and the Liquid Light Cleanse*,** to be explored in ***PART III. of Section I,*** the **Hierophant Symbol Code** will be used to bring the **Maharic Shield** and the **PBIS** out of dormancy, in order to **RESET** the **Imprint for Health** within the operational Holographic Template of the human **Kathara Grid.**

The *Hierophant Symbol Code Scalar Wave Guide* for Kathara Reprogramming

Thoughts and **Images** formed by the mind represent **small morphogenetic fields** with minute **Kathara Grids**, that become interwoven into the larger morphogenetic field and Kathara Grid of the human multidimensional anatomy. ***THOUGHTS ARE THINGS*** **in terms of the bio-energetic reality of the body and the Kathara Grid;** ***thoughts become translations*** **of electro-tonal "*digital" instructions* within the human bio-energetic field, and directly enhance or detract from the natural function of the electro-tonal programs of the human Kathara Grid.** In recognizing the **power of thought** ***as the power of creation of scalar standing-wave grids***, we can begin to apply that power effectively in regeneration of the Kathara Holographic Template, by **directing needed *electro-tonal instructions* for awakening the *Maharic Shield* into the embodied Kathara Grid.** The **Hierophant Symbol Code**, used in the form of a **visualized image**, sets a **powerful electro-tonal program** within the human Kathara, a program that **triggers the opening between the human Kathara and the PBIS**

The *Hierophant Symbol Code* appears in the form of a *Pale Silver Merkaba Star*, a six-pointed "Star of David", which represents the *electro-tonal program* of the *11th and 12th-Dimensional aspects of the Universal Kathara Grid.* The **frequency wave spectra** perceived by the human mind as **Pale Silver** represents the **electro-tonal program** of the ***Dark Silver*** *primary frequency band of the 11th-Dimensional wave spectra, combined with the* ***Pure White*** *primary wave spectra of the 12th-Dimensional frequency bands*. Using the **Hierophant Symbol** wave guide to direct *frequency and electro-tonal instructions* through the Kathara Grid, creates a ***scalar standing-wave pattern of 11th and 12th Dimensional frequency*** **within the personal Kathara Grid,** the frequencies in which the **original *electro-tonal instructions*** for the ***Imprint for Health*** are stored. By **awakening the 11th and 12th Dimensional frequencies within the human Kathara and bio-energetic field**, all of the *lower frequencies* within the Kathara electro-tonal program become **modulated and recoded** into their *organic, undistorted pattern*, by the *stronger wave influence* of the higher 11th and 12th Dimensional frequencies, which *contain within themselves* the organic imprint for the *entire dimensionalized wave spectra* of the dimensions that *oscillate below* them.

26

The Hierophant Symbol Code
Scalar Wave Guide for Kathara Reprogramming

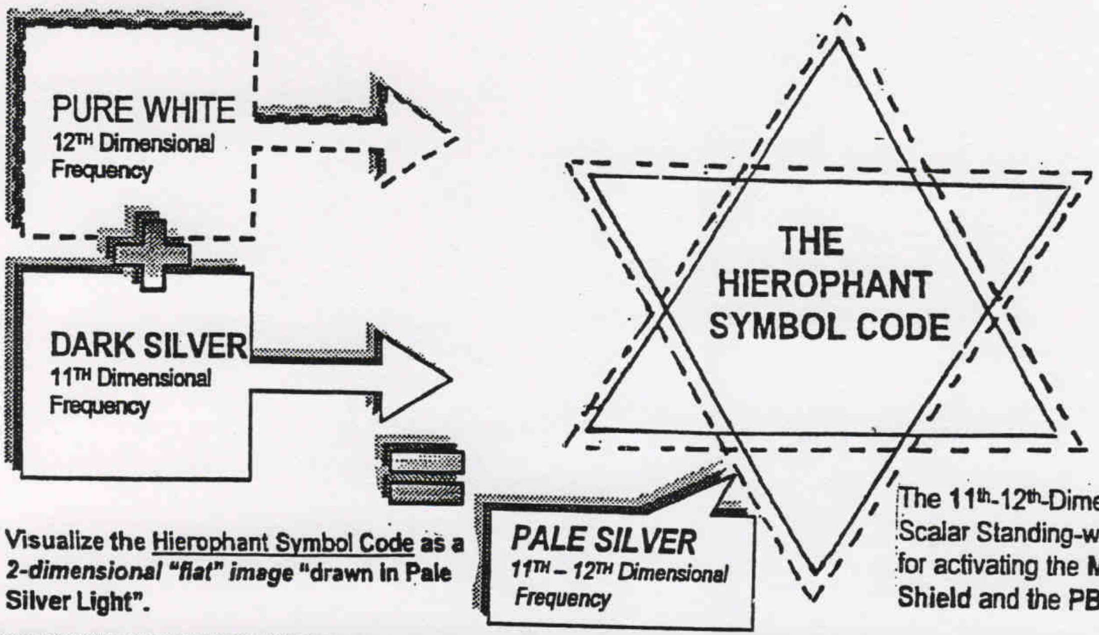
PURE WHITE
12^{TH} Dimensional Frequency
DARK SILVER
11^{TH} Dimensional Frequency
THE HIEROPHANT SYMBOL CODE
PALE SILVER
11^{TH} – 12^{TH} Dimensional Frequency

Visualize the Hierophant Symbol Code as a *2-dimensional "flat" image* "drawn in Pale Silver Light".

The 11^{th}-12^{th}-Dimensional Scalar Standing-wave Guide for activating the Maharic Shield and the PBIS.

Ease of Kathara Healing Techniques

Through using the *Hierophant Symbol Code* we awaken the restorative frequency patterns within the *Maharic Shield* and the *PBIS*, beginning the process of regenerating the organic *Imprint for Health in* the human Holographic Template Kathara Grid. The restorative affects of using the *Hierophant Symbol Code* to activate the *Maharic Shield* and the *PBIS* will occur upon using the *Kathara Mechanics* provided in Technique # 2, *whether or not* an individual understands the realities in frequency that are taking place through using this process. **Therefore, the complex mechanics of activating the *Maharic Shield* to reset the *Imprint for Health*, can be *reduced to simple instructions* for the direction of energy.** Though the **theory of Kathara Healing** is highly *detailed and complex*, the **Techniques** by which *active results are achieved* employ **simple exercises in *Mental Symbol Code Visualization*** and **mental energy direction, coupled with** the **application of simple *Kathara-stimulating massage*** to specific corresponding body points. **The *Techniques* of Kathara Healing and regeneration of Kathara Grid integrity are *so simple* that even a young child can learn to actively employ them, to achieve the desired result of restoring the *Imprint for Health*.**

MORPHOGENETIC EXPANSION

Frequency Accretion, Growth, Age Rhythms, Degeneration, Death and Kathara Blockage

Frequency Accretion and Growth Through Time

The morphogenetic field of the human body evolves through time via the process of **Frequency Accretion**. The human morphogenetic field *exists within* the larger morphogenetic field of the planetary and universal dimensionalized **Unified Field of Consciousness-Energy.** Through the inherent function of the **Kathara Grid** and the multi-layered systems of energy built upon it, *progressively more frequency and corresponding wave spectra* from the *dimensionalized Unified Field* are **drawn into** the *personal morphogenetic field* to create **progressive expansion of the morphogenetic imprint and its resulting progressive *expansion of consciousness* and *growth of the physical body* in time.**

Kathara blockages and Organic Growth Cycles

Blockages and scalar wave distortions within the **Kathara Grid** *impede* the natural functions of **Frequency Accretion**, and so **distort the intended natural processes of consciousness expansion and growth of the body form in time.** When the Kathara Grid is functioning properly, there are **certain *specific points in linear time*** **when *specific aspects of morphogenetic expansion* via Frequency Accretion are intended to occur.** The **original cycles of growth and expansion inherent to the human form** occurred between **1. fetal integration through age 11-12, 2. age 11-12 through age 22, 3.age 22 through age 33 and 4. age 33 through age 44**, at which time the human morphogenetic field would reach its **full expansion**, the **DNA** would reach its organic **12-strand activation**, the identity would expand to hold **12-dimensionsof consciousness** within embodiment, and the organism would have **full conscious control of *manifesting and de-manifesting* its 15-Dimensional anatomy in time. The body forms of the original human imprint are IMMORTAL** - deterioration and death were **not a part of the original *Imprint for Health*** within the human Bio-Spiritual makeup. Through understanding how the natural process of growth and expansion in time were INTENDED to work, we can also understand the core reason for the manifestation of the SYSTEMS MALFUNCTIONS - the conditions of biological and mental deterioration that lead to the condition of **BIOLOGICAL DEATH** - a condition that is **INORGANIC TO THE ORIGINAL GENETIC IMPRINT OF THE HUMAN FORM!.**

The First Natural Expansion Cycle

In the original human genetic blueprint, the frequencies and dimensions of consciousness corresponding to the wave spectra of Dimensions 1, 2, and 3, were intended to accrete or draw into the morphogenetic field between fetal integration and the age of 11-12. At the age of 11-12 the wave spectra of which the body was composed would take a **leap in frequency**, beginning the process of **expanding the morphogenetic field** to hold the higher frequency wave spectra of **Dimensions 4, 5 and 6.**

The *expansion of the morphogenetic imprint* is reflected in the *DNA*, and *chemical and hormonal balances* within the body, which govern the function of the *metabolic rate* and the operation of molecular function. The **first cycle** of *morphogenetic expansion* into the wave spectra of **Dimensions 4, 5 and 6,** represents the process of **SOUL INTEGRATION** for the embodied identity, and **begins the natural process of de-densification of matter** - *by lowering of morphogenetic vibration, raising of oscillation rate and frequency, and shortening of the wave-length of the scalar waves that compose the matter form and consciousness.* **These changes in *intrinsic wave characteristics* have their reflection within the behavior of *atomic structure*-** as the morphogenetic field wave spectra **raises in frequency**, the ***electron spin* around the atomic nucleus *slows***, thereby increasing the electrical nature, and reducing the magnetic orientation of atomic function, ***transmuting the perceivable matter substance* of the cells to a *less dense* state of being. Completion of the *Soul Integration process* marks the transmutation of *Carbon-based Gross-matter density* into the *Carbon-Silica based Semi-etheric matter density*, and the expansion of embodied awareness from 3 to 6 Dimensions of consciousness.**

Completing Expansion

This process of *progressive morphogenetic field and consciousness expansion* and *resulting de-densification of biological matter* continued throughout the remaining *expansion cycles*. Age **22 to age 33** marked the accretion of the **7th, 8th and 9th -Dimensional** wave spectra, the process of ***Over-Soul Integration,*** expansion into a **9-Dimensional consciousness** and transmutation to a ***Silica-based Etheric matter density biology***. Age **33 to 44** marked the accretion of the **10th, 11th and 12th Dimensional** wave spectra, the process of ***MAHARIC INTEGRATION***, expansion into a **12-Dimensional consciousness** and transmutation **out of biology** into the **Crystalline Liquid Light wave form of Pre-matter substance.** At this point the identity is considered a **full AVATAR** and has complete control over ***manifesting and de-manifesting*** the **Holographic Projection** of its desired body imprint in time.

Growth Rates and Morphogenetic Expansion

ORIGINAL HUMAN GENETIC IMPRINT

PERSONALITY INTEGRATION (Telluric Integration) – **BIRTH - AGE 12**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 1-3 and activating DNA Strand Templates 1-3.

SOUL INTEGRATION (Doradic Integration) – **AGE 12 – 22**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 4-6 and activating DNA Strand Templates 4-6.

OVERSOUL INTEGRATION (Teuric Integration) – **AGE 22 – 33**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 7-9 and activating DNA Strand Templates 7-9.

CHRISTOS AVATAR INTEGRATION (Maharic Integration) –**AGE 33 – 44**
Expands the Morphogenetic Field through accreting the frequencies of Dimensions 10-12 and activating DNA Strand Templates 10-12. Allows full Cellular Transmutation and Dimensional Ascension to the 12th Dimensional Pre-matter Liquid Light HYDROPLASMIC Substance Liquid Silica) and conscious mastery over 12 Dimensions within the 15-Dimensional Time Matrix. A "Christed" Immortal Being possessing non-polarized Unity or Christ Consciousness.

CONTEMPORARY HUMAN GENETIC IMPRINT

PERSONALITY INTEGRATION (Telluric Integration) – **BIRTH – DEATH**

SOUL INTEGRATION (Doradic Integration) – KATHARA BLOCKAGE

OVERSOUL INTEGRATION (Teuric Integration) – KATHARA BLOCKAGE

CHRISTOS AVATAR INTEGRATION (Maharic Integration) – KATHARA BLOCKAGE. Deterioration and Death of the physical body occur rather than progressive frequency accretion for Cellular Transmutation and Dimensional Ascension. Distortions in the personal Kathara Grid and Scalar Shields manifest in the DNA Template and block natural inflow and Morphogenetic Field accretion of frequency from the higher dimensional stations of identity. Conscious awareness remains polarized in Duality Consciousness and physical body is rendered finite and vulnerable to dis-ease.

THE 4 NATURAL CYCLES OF MORPHOGENETIC FIELD EXPANSION

THE *4 CYCLES OF MORPHOGENETIC FIELD EXPANSION* REPRESENT THE *ORIGINAL ORGANIC PROCESS* BY WHICH THE HUMAN ORGANISM WAS *INTENDED* TO EVOLVE THROUGH TIME.

The Inorganic Condition of Human Illness and Death

The processes of biological and mental deterioration and resulting DEATH of the biological form, which we presently view as FACTS OF LIFE regarding the process of human passage through linear time, ARE NOT ORGANIC CONDITIONS TO THE HUMAN ORGANISM. Other life forms, such as the **animal kingdoms**, that manifest through morphogenetic fields having *less than 12 Kathara Centers active* within their Kathara Grid, cannot undergo the full process of bodily transmutation into Pre-matter substance, and cannot fully expand their embodied awareness into 12th-Dimensional consciousness. For the **animal kingdom**, the process of evolution is also Frequency Accretion, but it is achieved through expanding one's body form to its full capacity, then entering the consciousness into a morphogenetic imprint with a bit higher capacity, until eventually the consciousness can accrete enough frequency to take it into the next level of matter de-densification and consciousness expansion. **Death is natural to the organic animal kingdom.**

Humans were intended to undergo the *full spectrum of morphogenetic expansion and cellular transmutation* WITHIN ONE ETERNAL BODY IMAGE, that was built upon a program of *full activation* of the *12 Kathara Centers* of the Kathara Grid. Death is NOT natural for humans.

Kathara Grid Blockage and Molecular Compaction

The process of *biological deterioration* and progressive movement toward biological DEATH is the *direct result* of unnatural BLOCKAGES WITHIN THE KATHARA GRID HOLOGRAPHIC TEMPLATE. Blockages within the Kathara Grid program presently block the process of Frequency Accretion **at the beginning of the First Expansion Cycle-** as accretion of the 4th, 5th and 6th Dimensional frequencies representing the *Soul Integration Process* is blocked within the Kathara Grid and thus within the biological DNA, the incoming frequencies of the Soul Level consciousness **push on and CRUSH the morphogenetic field**, rather than *expand the morphogenetic field into a higher level of frequency holding*. When the incoming scalar wave frequencies of Soul Integration begin to crush the morphogenetic field, a condition called **MOLECULAR COMPACTION** is created.

MOLECULAR COMPACTION -

***over exposure to scalar wave spectra* and its resulting *fragmentation of the morphogenetic field and Kathara Grid* - IS THE *ONLY TRUE CAUSE OF DIS-EASE*, AND THE *CORE CAUSAL ELEMENT* OF THE *UNNATURAL DEATH EXPERIENCE*.**

The *Maharic Shield* and Human Healing

Within the imprint of the **MAHARIC SHIELD** of human morphogenetic structure, is the dormant imprint of the *10th, 11th and 12th Dimensional electro-tonal programs* that represent the **ORGANIC PROGRAM OF THE KATHARA GRID. Through activation of the Maharic Shield** - the scalar standing-wave template that corresponds to the Pre-matter Liquid Light fields of Dimensions 10, 11 and 12 - **blockages within the Kathara Grid Holographic Template, that result in the genetic mutation that causes Molecular Compaction, can be progressively realigned with the original IMMORTAL IMPRINT FOR HEALTH.** Though it takes time and consistent application of **Kathara Healing Technique** to fully regenerate the **Imprint for Health and Immortality** within the 15-Dimensional Levels of human anatomy, working to heal distortions within the Kathara Grid through **MAHARIC INFUSION** (running the frequencies and wave spectra of the *Maharic Shield* through the Kathara Grid, Bio-energetic System and Body), **can create progressively more observable affects in creating and maintaining health, and slowing the bodily deterioration process, within the normal life span of an individual.**

Potentials of Kathara Healing

The true beauty and significance of *Kathara Healing* is that it can not only do wonders for creating and maintaining *higher standards of health* within the average human life span, but within the applications of KATHARA MECHANICS lies the *promise and potential* of human biological IMMORTALITY and the fulfillment of SPIRITUAL ACTUALIZATION and evolution to MASTERY- through progressive embodiment of and conscious mastery over our 15-Dimensional Anatomy and the creative process of Holographic Projection within the dimensionalized fields of space, time and matter.

KATHARA HEALING HOLDS THE KEY TO HUMANITY RECLAIMING ITS ANCIENT HERITAGE OF THE ETERNAL SELF AND IMMORTAL AT-ONE-MENT WITH SOURCE.

LEVELS OF HUMAN ANATOMY- THE EMBODIED KATHARA

The Human Body Kathara Grid

The Kathara Grid of the human body is the *Core Holographic Template* upon which the morphogenetic scalar-wave template, and all other levels of identity and form are built. It is the **"Control Center"** for manifestation of the human being.

LEVEL-1 KATHARA 12-TREE GRID

THE KATHARA GRID LEVEL-1: 12-TREE

Level-1 Kathara Grid 12-Tree: The first level of the Kathara Grid **sets the *pattern for dimensionalization of consciousness* through which identity can enter the Time Matrix for the Holographic Experience of manifestation. The Kathara Grid 12-Tree is the first wave spectra composite through which consciousness anchors itself into dimensionality in order to experience manifestation of form.** All things within dimensional structure have the Kathara 12-Tree at the core, but not all manifest forms have the full pattern of the 12-Tree *activated* within their morphogenetic pattern. The 12-Tree is composed of **12 Kathara Centers** or *Kathara Spheres*, and **15 Kathara Lines.**

KATHARA CENTERS

The **Kathara Centers** are *crystallizations of Partiki units* that hold *composite electro-tonal energy signatures for each dimensional band.* The electro-tonal programs of the Kathara Centers hold the **foundation template** upon which the biological **DNA** will manifest.

KATHARA LINES

The **Kathara Lines** are sequences of *Partiki Phasing rhythms* ("flashing on and off"/fission-fusion sequences) that *transfer the dimensionalized electro-tonal programs of the Kathara Centers between each center* , to form the **core level of the scalar standing-wave template** upon which subsequent levels of the Kathara Grid, Bio-energetic System and body will build.

KATHARA HEALING AND THE LEVEL-1 12-TREE GRID

The *Kathara Centers* and *Kathara Lines* are used in *Kathara Healing* to realign and awaken the original 12-Dimensional imprint within the human morphogenetic field, in order to *reinstate the original Imprint for Health and Bio-Spiritual Mastery* within the operational evolutionary blueprint for the Body-Mind-Spirit System of the Human form.

Level-1 Kathara 12-Tree Grid

of the INTERNAL TEMPLAR

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 –*The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7-*The Wings*
In shoulder indents
through to back

# 5 –*Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 -*The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

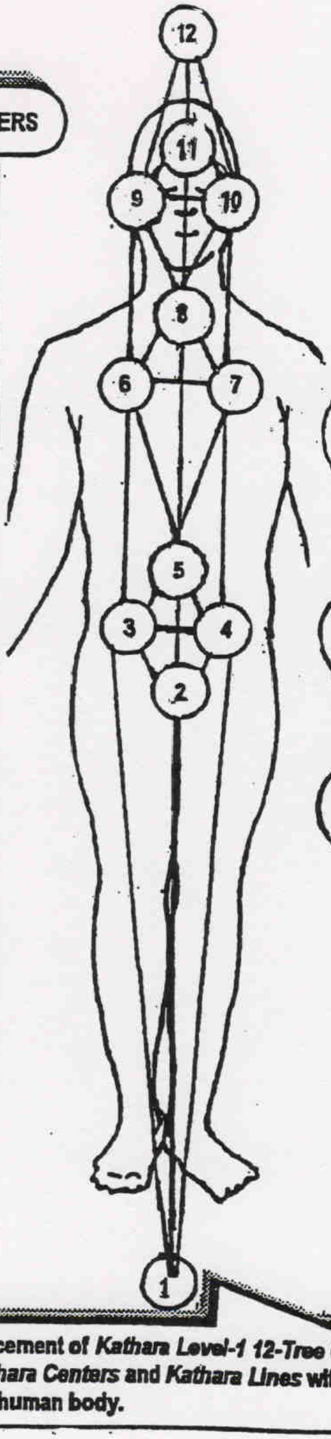
12
11
9
10
8
6
7
5
3
4
2
1

Placement of *Kathara Level-1 12-Tree Grid Kathara Centers* and *Kathara Lines* within the human body.

Central Vertical Kathara Line
controls all other Kathara Lines. Kathara Centers on *Central Vertical Kathara Line* control all other Kathara Centers.

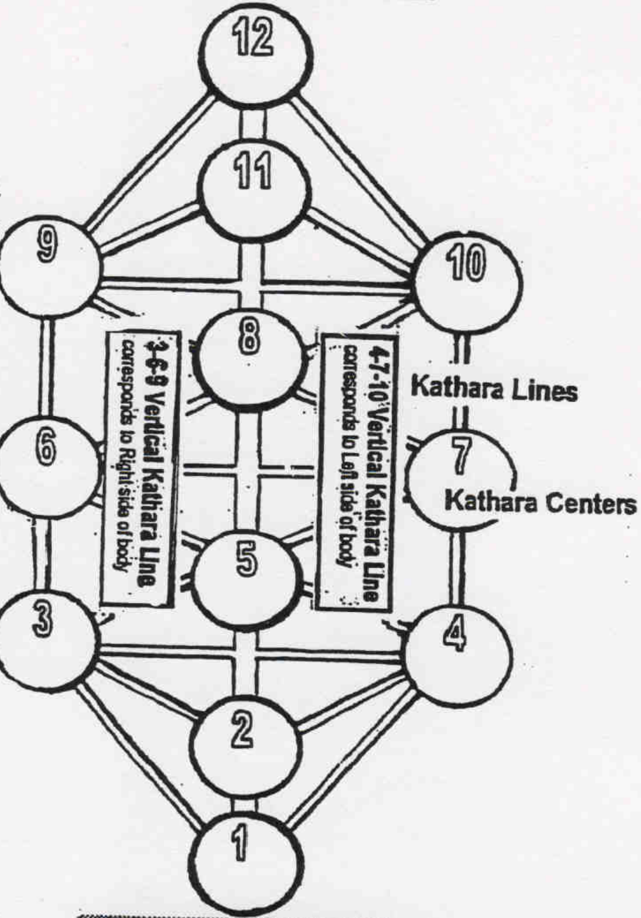
12
11
9
10
8
3-6-9 Vertical Kathara Line
corresponds to Right side of body
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
6
7
Kathara Centers
5
3
4
2
1

Electro-tonal Programs within the 3 ***Vertical Kathara Lines*** control the programs in all other Kathara Lines.

LEVEL-2 KATHARA CRYSTAL SEALS GRID

Level-2: Crystal Seals and 15 Primary Chakras

The Crystal Seals are groups of 3-Dimensional Partiki scalar wave composites that regulate the rate of Partiki Phasing to create the base structures upon which dimensionalization is formed. The *flow of frequency* between dimensional bands and Harmonic Universes, the fixed expansion and contraction/ fission and fusion rates of Partiki, the *Vibration-Oscillation Rates* of Partiki and the *Angular Rotation of Particle Spin* are all regulated by the Crystal Seals.

Star Crystal Seals are positioned between the Chakra Centers along the **Central Body Current**, and **Seed Crystal Seals** are positioned between them, and serve as the *point of composite frequency* out of which the **15 Primary Chakra Center Vortices** emerge.

HOVA BODIES AND SHIELDS

5 Dimensionalized Hova Body Shields and 5 Auric Capsules

Hova Bodies are **3-Dimensional, tri-tonal scalar standing-wave grids** composed of numerous sets of **Crystal Seals** that form the electro-tonal program upon which the **Kathara Grid Level-3: Doradic Points** emerge. **The Hova Body Grids are called Crystal Shields.** The **5 Dimensionalized Hova Bodies** are structured as **5 Concentric Spheres** of energy that *surround and permeate the human body* and serve to *hold the embodying consciousness into dimensionalized levels* through which the **Holographic Projection** of *experienced manifestation* can occur. The Hova Bodies form **"tissue capsules"** of *subtle energy* around **each 3-dimensional Harmonic** level of the **Bio-energetic Field**, within which the **15 Dimensional levels** of the **Auric Field** are stationed.

In Kathara Healing, restoration of the original **4th-Harmonic Maharic Shield** *Imprint for Health* is used to *realign and clear blockages* from the lower Hova Bodies and corresponding **Kathara Centers** and **Kathara Lines**, to awaken the **Imprint for Health and Spiritual Actualization** within the Body-Mind-Spirit System.

HOVA BODIES, SHIELDS & SIGNETS

5 Ariea Hova Bodies, Kathara Grid and Harmonics of Manifestation

Hova Bodies are *tri-tonal scalar grids* that form *spherical energy "tissue capsules"* around each 3-dimensonal level of the multi-dimensional identity. Within 15-Dimensional Human Anatomy, there are **5 Dimensionalized Hova Bodies** called **Ariea Hova Bodies.** The Ariea Hova Bodies form out of the electro-tonal programs held within the **Kathara Centers** of the *Level-1 Kathara 12-Tree Grid.* The *12th Kathara Center* of the Kathara 12-Tree Grid connects to the *12th Kathara Center* of the *Universal Kathara Grid*, just as the *1st Kathara Center* in the personal Kathara connects to the *1st Kathara Center* in the *Planetary Kathara Grid. The 12th and 1st Kathara Centers,the personal Kathara Grid and morphogenetic field connect to the planetary and universal morphogenetic fields.* **When working with the Hova Bodies we first focus upon the lower four , as these must be functioning properly before the 5th Hova Body and its connection to the Universal Kathara Grid can be activated within the Bio-energetic Field.** Four of the 5 Ariea Hova Bodies are directly connected to the Crystal Seals within the human body **Central Body Current** on a **HORIZONTAL PLANE.** These first **4 Hova Bodies** correspond to each of the **12 Kathara Centers** of the personal Kathara 12-Tree Grid. **The Hova Bodies represent Harmonics of Manifestation.**

Hova Bodies, SIGNETS and SHIELDS

Each Hova Body **emanates outward** from a set of **3 Kathara Centers** and their corresponding **Crystal Seals. The *core template of the Hova Body* is a *Triad of Crystalline Spheres* that connect directly to the *3 Kathara Centers* to which they correspond.** This Triad of Crystal Spheres is called a **SIGNET. The Signets are the smaller, *fixed, consolidated* :frequency points: out of which the *Seed Crystal Seals* that form the *Chakras* manifest.** The Signets and the smaller Crystal Seal formations that manifest through them, create a **disc** of spinning, **horizontal wave spectra** that direct the function of *Hova Body merger*, these scalar discs are the scalar templates upon which the Hova Bodies manifest, and are called **SHIELDS.**

The *Shields and the Signets* that direct them are considered to be the *core of the Level-2 Crystal Seals Kathara Grid.* Each of the 4 Hova Bodies of the first 12 Dimensions form upon a set of **3 Signets.** Each **Signet *transmits a flash-line*** (Partiki Phasing Sequence) **on a different axis**, one **vertical**, one **horizontal** and one **diagonal**, forming a **3-plane projection of flash-lines** that form the base of the **3-dimensional Holographic projection.**

Kathara Centers and Activating the Signets and Shields

In each set of *3 Signets* at least one corresponds to the *Central Vertical Kathara Line*, the control center of the Level-1 Kathara Grid 12-Tree. The Signets and Shields can therefore be used to direct the function of the Kathara Grid 12-Tree and all built upon it, and the Kathara Centers can be used to trigger activation of the Signets and Shields within the Hova Bodies. When a **Signet is activated** its corresponding **disc or *Shield* begins to spin**. When a certain spin rate of the Shield is reached, it triggers the activation of the next set of Signets in the next Harmonic setting the next Shield to spin. **When *2 Shields activate* and reach a reciprocal spin rate the *tissue capsule* separating the 2 corresponding Hova Bodies releases and the *2 Hova Bodies merge*, which is the process by which *Higher Identity Integration* occurs. It is also the process by which the *Merkaba Vehicle* is activated within the Bio-energetic field.** When all **four Shields** activate and spin, the body matter is able to **transmute** out of matter density and into pure wave form, which represents Maharic/ Avatar Integration for the identity.

THE 5 HOVA BODY SHIELDS

The Hova Bodies and the Shields

Each Hova Body contains at its core a Shield through which the Hova Body functions are governed. The **Nada Hova Body** of dimension 1-2-3 forms on the **TELLURIC SHIELD**, its control center being the 2nd Kathara Center that is composed of 2nd Dimensional Frequencies. The **Alphi Hova Body** of dimensions 4-5-6 forms on the **DORADIC SHIELD**, its control center being the DORA or Archetype of the 5th Dimensional Frequencies. The **Betcha Hova Body** of dimensions 7-8-9 forms on the **TEURIC SHIELD**, its control center being the TEURA or Monad of the 8th Dimensional Frequencies. The **Mahara Hova Body** forms on the **MAHARIC SHIELD**, its control center being the *Mahunta* or Avatar Core of the 11th Dimensional Frequencies. The 5th Hova Body, the **Raja Hova Body**, forms on a set of Signets that correspond to **stellar points** within the **Universal Kathara Grid.** Its Shield, the **RISHIC SHIELD**, rotates on a **vertical plane** and it comes into activation within the personal morphogenetic field only after the four lower Shields have transmuted the body into Pre-matter Liquid Light of the MAHARIC SHIELD. **In terms of healing and advancing evolution of the earthly identity, the first four Shields are of greatest significance.**

Hova Bodies, Shields, Signets and Healing

In all healing application knowledge of the Shields is of great value, for any manifest condition of disorder will have its first manifestations within the Kathara Grid and the Shields that form upon it. The process of healing, therefore, is the process of *restoring the Shields to their proper order* so they may activate to build the Merkaba Vehicle through progressive Higher Identity Integration. **In Kathara Healing we begin with the MAHARIC SHIELD**, the Shield that contains the *pure form of electro-tonal programs upon which the consciousness first entered matter density* through the Pre-matter Liquid Light Fields of the 4th Harmonic - Dimensions 10-11-12. **Through activating and progressively drawing in the highest 11th and 12th-Dimensional frequencies of the Maharic Shield, distortions in all lower Shields can be more rapidly realigned with their original pattern.** When we work with the **Kathara Centers** of the Level-1 12-Tree Grid we simultaneous create change within the **SIGNETS** that direct the function of the **SHIELDS**, and so we are able to more rapidly facilitate appropriate realignment of the Shields and Hova Bodies. **THE CORE OF HEALING IS RESTORING THE HOVA BODIES AND THE SHIELDS UPON WHICH THEY MANIFEST TO THEIR ORIGINAL INTEGRAITY OF ORDER.**

Shields, Signets and Diodic Points

Through the interaction and interplay of multi-harmonic frequency between the **Shields, Signets** and their **Crystal Seals, Kathara Grid Level-3- DIODIC POINTS** come into manifestation, to fill out the 3-plane, 3-dimensional "armature" of scalar waves and flash-lines, upon which more Partiki units will cluster and crystallize to build up the particle base for form manifestation. The **Diodic Points** within the body and morphogenetic field connect directly to the Shields of the Hova Bodies in areas where the *flash-lines and scalar points* within the Crystal Seals of the Shields *align with and pass through each other.* **Flash-Lines from one Hova Body Shield crossing through those in other Hova Body Shields form *COORDINATE POINTS*, or "Gateways" of energy-consciousness *transference* through which frequency and awareness from one Hova Body and identity level can cross through into other Hova Bodies and identity levels.** The *Coordinate Points*, or *regularly scheduled points of flash-line intersection*, form **tiny vortices of energy** , which keep energy and consciousness flowing throughout the multiple dimensional levels of identity and anatomy. **The vortices of energy created by the Coordinate Points of flash-line sequences are called *DIODIC POINTS* and they have a counter reflection within *anti-particle manifestation* called *MIODIC POINTS.*** Diodic and Miodic Points are sets of **White and Black Holes** that circulate energy throughout the anatomy.

LEVEL-3 KATHARA GRID DIODIC POINTS

Diodic and Miodic Points

Diodic/ Miodic Points represent *concentrated areas of energy* that emerge at points where *dimensionalized levels of scalar grids cross over and through each other*, creating minute **vortices** of wave spectra that serve as **frequency modulation zones** (*scalar flash-cycle transition points*) between Dimensional Frequency Bands. Diodic Point vortices are **White Hole** vortices that *transmit energy* into the Particle Universe from their counterpart **Miodic Point** vortices within the Anti-particle Universe. Through the structure of Dimensionalization, each Harmonic Universe, and the structures contained there within, have within their morphogenetic fields *sets of Diodic and Miodic Points* - **sets of White and Black Holes** - that regulate the flux of consciousness-energy between systems.

Within the biological form and morphogenetic field, Diodic Points are located in **physical body regions** where Inter-dimensional and Inter-Harmonic **Horizontal and Vertical Kathara Lines intersect** to form small **vortices** of inter-dimensional or inter-harmonic wave spectra that serve **as Gateways of Consciousness** between Dimensional Time Continua of Harmonic Time cycles. They connect physical bodies to their **anti-particle counterparts** and intersect with the Vertical and Horizontal relationships of wave spectra within the **DNA/ RNA** and bio-energetic field, through sets of White and Black Holes at the center of the biological Harmonic scalar grids.

Diodic Points and 3-Dimensionality

Diodic Points control the **Diagonal Kathara Lines**, the **sequences of Partiki Phasing** or ***"flash-lines"*** that run on a **diagonal** between the Kathara Centers. Whereas the **Kathara Level-1 Kathara Lines** direct the **Vertical flash-line sequences** and the **Level-2 Crystal Seals** control the primary **Horizontal flash-lines** and their relationships to the Vertical flash-lines, the **Diodic Points** further regulate the relationships of flash-line sequences through the Kathara Grid by serving as **"turnstiles of energy"**. The Diodic Points intersect with the crossover points of the Vertical and Horizontal flash-lines , *adding a further level of modulation* to the frequencies passing through, allowing a **3-dimensional Harmonic Projection of flash-lines** (flash-lines projecting on 3 differ angular planes)to build within the morphogenetic field. **The Diodic Points add the third plane of projection to the Kathara Grid, beginning the formation of what will appear as a 3-dimensional "armature" of flash-points upon which particles will accrete to "fill out" the Holographic illusion of 3-dimensionality.**

Diodic Points and Healing

In Kathara Healing Diodic Points are used to open the **Diagonal Kathara Lines** for reprogramming. When Diodic Points become blocked, due to blockages within the Kathara Grid (which appear as "gaps" or missing sequences of flashes along the flash-lines of Partiki Phasing), energy builds up in the *Level-3 Kathara Grid*, forming *distortions* in the natural pattern that will manifest in the body and bio-field systems as *disease or disharmonic condition*. In later applications of Kathara Healing we will learn to identify points of Diodic and Miodic build-up in the body, and use *primary flash-line sequences applied through touch points* to realign the natural flash-point sequences and clear the Diodic Blockages.

Shields, Signets and the Kathara Grid

Hova Bodies **are spherical scalar wave grids that form tissue capsules between each 3-Dimensional Harmonic of Manifestation.**
Signets **are the core points of consolidated frequency that hold the programs for the Scalar grid SHIELDS upon which the Hova Bodies manifest.**

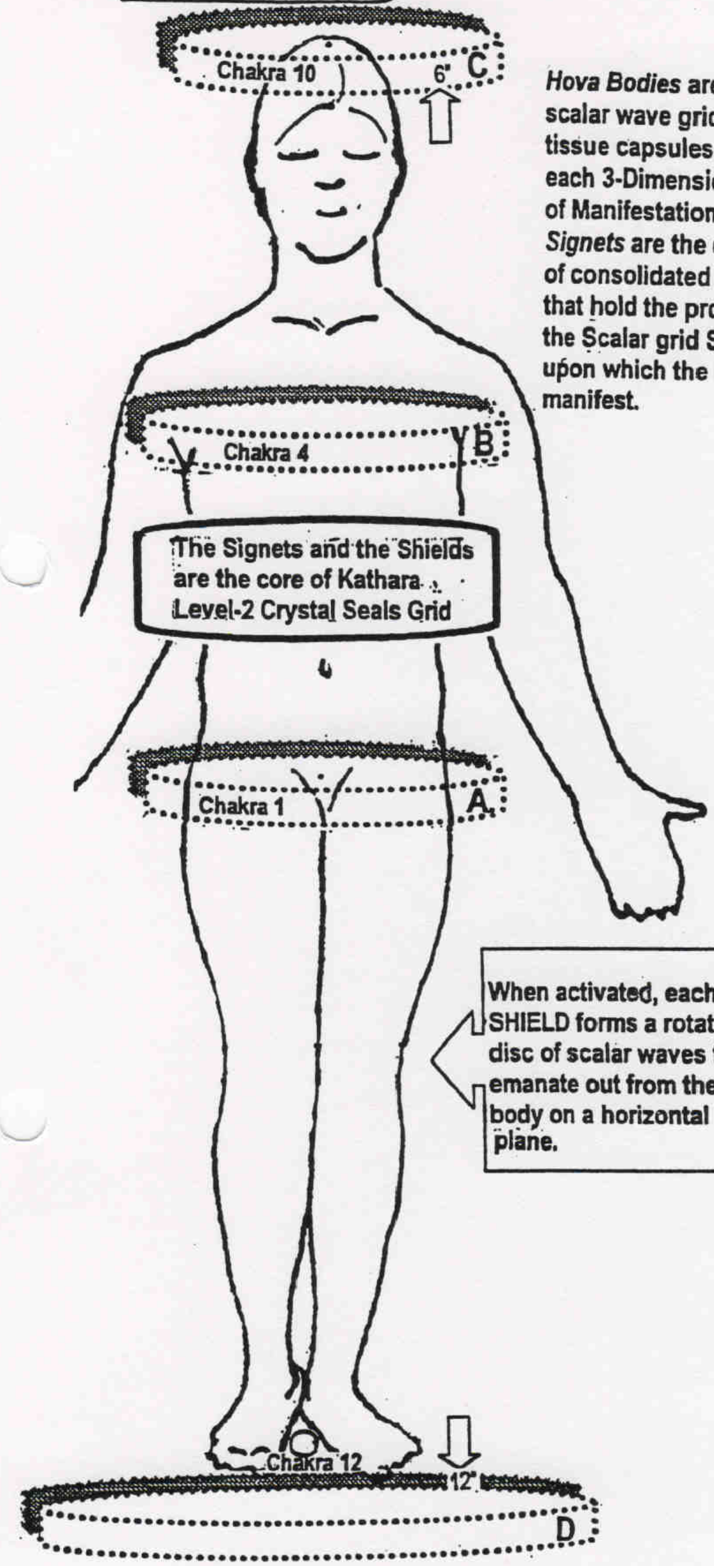
Chakra 10
6"
C
Chakra 4
B
Chakra 1
A
Chakra 12
12"
D

The Signets and the Shields are the core of Kathara Level-2 Crystal Seals Grid

When activated, each SHIELD forms a rotating disc of scalar waves that emanate out from the body on a horizontal plane.

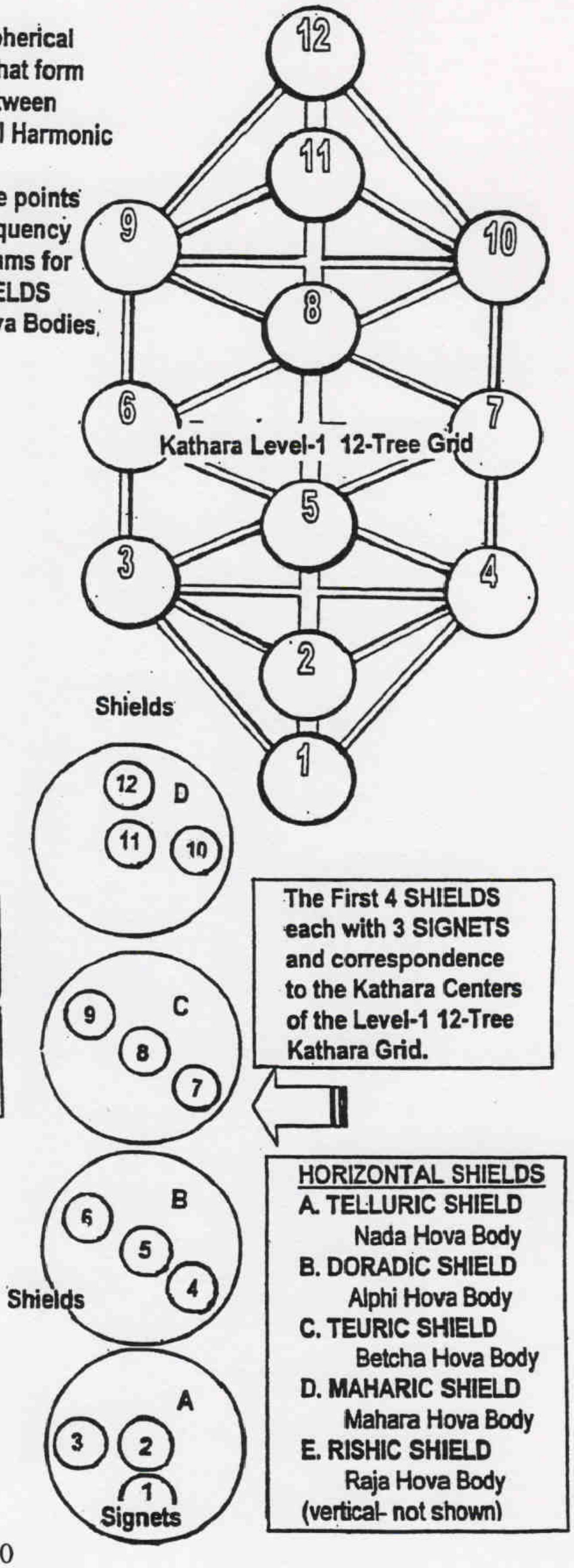
12
11
9
10
8
6
7
Kathara Level-1 12-Tree Grid
5
3
4
2
Shields
1
12
D
11
10
9
C
8
7
6
B
5
Shields
4
A
3
2
1
Signets

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

HORIZONTAL SHIELDS

A. TELLURIC SHIELD
Nada Hova Body
B. DORADIC SHIELD
Alphi Hova Body
C. TEURIC SHIELD
Betcha Hova Body
D. MAHARIC SHIELD
Mahara Hova Body
E. RISHIC SHIELD
Raja Hova Body
(vertical- not shown)

Level-2 Kathara Crystal Seals Grid and Chakras

Chakras (C & MC)

SIGNET
Star Crystal Seals (S)

15 Chakras (C & MC)
7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal
8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

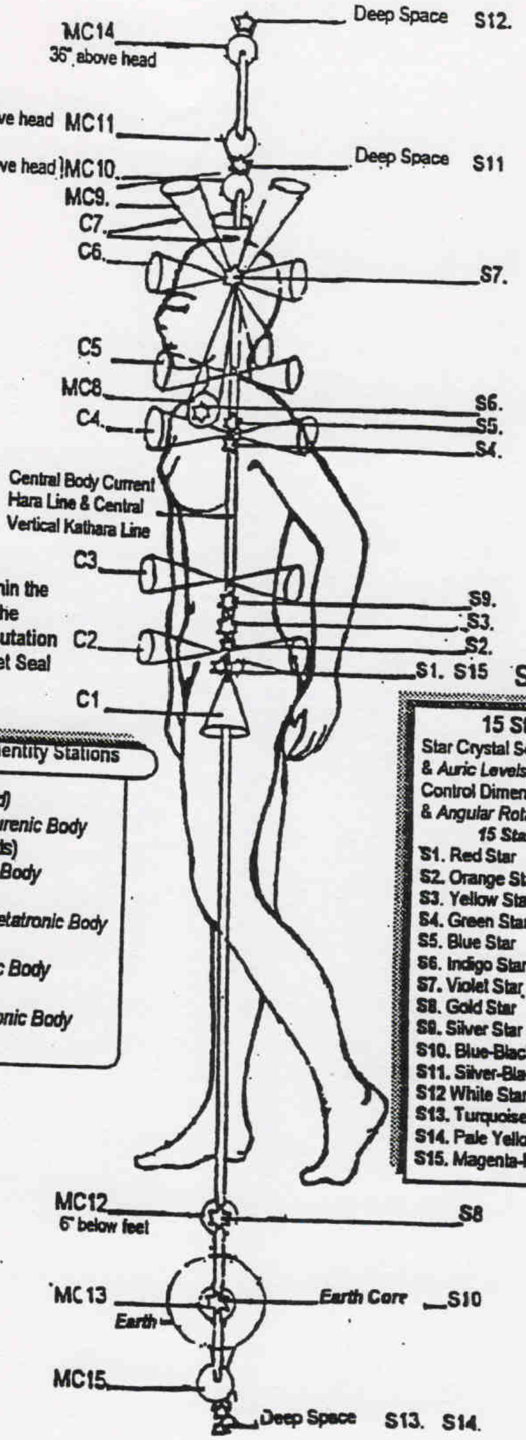
Deep Space S12.
MC14
36" above head
18" above head MC11
6" above head MC10
Deep Space S11
MC9.
C7.
C6.
S7.
C5
MC8
C4.
S6.
S5.
S4.
Central Body Current Hara Line & Central Vertical Kathara Line
C3
S9.
S3.
C2
S2.
S1. S15
C1
MC12
6" below feet
S8
MC 13
Earth
Earth Core
S10
MC15
Deep Space S13. S14.

The 5 Hova Bodies progressively merge with release of the embodied Signet Star Crystal Seals, forming the various phases of the personal Merkaba Vehicle within the bio-energetic field.

Release of the SIGNET Star Crystal Seals within the embodied Internal Templar Complex activates the 12-Strand DNA Template for Cellular Transmutation and Dimensional Ascension through the Signet Seal Star Gates of the Universal Templar Complex.

Hova Bodies and Corresponding Identity Stations

("Hova" means *Tri-tonal Scalar Grid*)

1. Nada Hova – HU-1 Incarnate Identity -*Taurenic Body* (Subconscious, Instinctual & Reasoning Minds)
2. Alphi Hova – HU-2 Soul Matrix – *Doradic Body* (Astral, Archetype & Angelic Minds)
3. Betcha Hova –HU-3 OverSoul Matrix – *Metatronic Body* (Ketheric, Monadic & Kertatric Minds)
4. Mahara Hova – HU-4 Avatar ID – *Hydronic Body* (Christiac, Buddhaic & Nirvanic Minds)
5. Raja Hova – HU-5 Rishi Collective – *Hedronic Body* (Universal Conscious Mind)

Signet Seals

15 Star Crystal Seals
Star Crystal Seals keep human identity & *Auric Levels* dimensionally separate. Control Dimensional *Merkaba Field Axis* & *Angular Rotation of Particle Spin.*
15 Star Crystal Seals (S):
S1. Red Star
S2. Orange Star
S3. Yellow Star
S4. Green Star
S5. Blue Star
S6. Indigo Star (Blue-Violet)
S7. Violet Star
S8. Gold Star
S9. Silver Star
S10. Blue-Black Star
S11. Silver-Black Star
S12 White Star
S13. Turquoise Star
S14. Pale Yellow Star
S15. Magenta-Pink Star

5 Hova Bodies

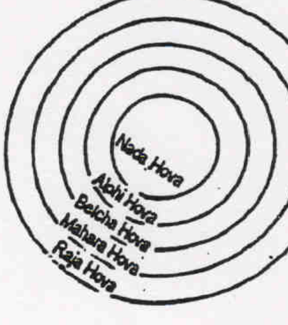
Nada Hova
Alphi Hova
Belcha Hova
Mahara Hova
Raja Hova

AXIOM LINES, AURIC LEVELS AND CHAKRAS

Diodic Points, Signets, Seed Seals and AXIOM LINES

Each **Signet** within a Hova Body Shield holds the electro-tonal program for one *Seed Crystal Seal* and one *dimensionalized flash-line sequence*. Each Seed Seal program corresponds to one *dimensionalized flash-line program* from one *Level-1 Kathara Center*. Each **Seed Crystal Seal** holds intact the electro-tonal program for one ***dimensionalized flash-line projection angle***, or **Single Axis Dimensional flash-line sequence.** The Single Axis Dimensional flash-line sequences controlled by each Seed Crystal Seal are called **AXIOM LINES.**

There are **12 Primary Axiom Lines** that correspond to the 12 Kathara Centers in the *Personal and Planetary Kathara Grids*. There are 12 additional **Stellar Axiom Lines** that correspond to the 12 Kathara Centers in the Universal Kathara Grid. **The Axiom Lines represent single-axis flash-line sequences -** (fixed points of conscious electrified sound- *electro-tonal units*- that form strands of scalar standing waves) **that intermesh to refine the scalar-grid webbing upon which Kathara Grid Level-3 Diodic Points will emerge.**

Axiom A-B Tonal Lines, Diodic-Miodic Vortices, Chakras-DNA-Portals-Ley Lines & Healing

Each of the 12 Primary Axiom Lines has a **reverse reflection** within the dimensions of the **anti-particle universe.** The 12 Primary Axiom Lines governing the **Diodic "White Hole" Vortices of the particle universe** are called **Axi-A-Tonal Lines** (or "Axiatonal" Lines). The corresponding reflections of the Axi-A-Tonal Lines in the **anti-particle universe,** which govern the **Miodic "Black Hole" Vortex** reflections of the Diodic Points, are called **Axi-B-Tonal Lines.** In Kathara Healing the integrity of the Axi-A-Tonal Lines is restored through realigning the electro-tonal programs within the corresponding **Kathara Centers.** The **12 Primary Axi-A-Tonal Lines** set the morphogenetic imprint upon which the **Chakra System** and foundation **DNA** organization of the human body will manifest and upon which the **Planetary Vortex/Portal System** and primary **Planetary Ley Lines** will emerge within the planetary body. Restoring the *Imprint for Health* to the **Axi-A-Tonal Lines** and corresponding **Kathara Centers** will regenerate integrity within the **DNA** and **Chakra System** of the body and will restore the organic operations of the planetary **Vortices-Portals** and **Ley Lines** of the planetary morphogenetic grid.

Restoring the **Imprint for Health** within the **Axi-A-Tonal Lines** simultaneously restores integrity to the corresponding **Diodic Points** and to the **Axi-B-Tonal Lines** and their **Miodic Points.** When working with *Kathara Techniques* that "clear the **Miasmic Imprint**" – *the anti-particle blockages within Diodic Points* – the *Diodic Points* are realigned by clearing their *Miasmic distortions*, which then reestablishes the integrity of corresponding **Axiom A and B Tonal Lines,** sending a realigned electro-tonal program back through the **Seed Seals** and the Level-1 **Kathara Centers.** If the programs within the **Kathara Centers** are **simultaneously realigned** using infusion of the frequencies from the **Maharic Shield,** the Kathara Center programs, the Axiom Lines and the Diodic Points will be progressively restored to health. If the Kathara Centers are not simultaneously realigned, the restored Axiom Line programs will be **overridden** by the Kathara Center programs until the restored programs reach *critical mass* through *prolonged repetition*. ***Clearing the Miasmic Imprint*** **while simultaneously** ***Repatterning the Level-1 Kathara Centers and Kathara Lines*** **via** ***Maharic Infusion*** **is the fastest way to restore the** ***Imprint for Health*** **within the Human Body-Mind-Spirit System and within the natural operations of planetary bio-mechanics.**

Axiom Lines, Diodic Points, Hova Bodies, Shields and Auric Field Levels

The Single Axis flash-line sequences of the **12 Primary Axiom Lines** cross over and through each other to form COORDINATE POINTS between the Hova Body Shields, through which the "White Hole" ***Level-3 Diodic Point Vortices*** emerge. **The varying angles of flash-line projection within the Axiom Lines create separate dimensionalized bands of frequency** (separate sets of flash-line sequences) **within the Hova Body structure.** As there are **3 Signets** and thus **3 Seed Seals** and **3 Axiom Lines** inherent to each **Hova Body Shield**, each Hova Body becomes *dimensionally compartmentalized* into **3** separate yet interrelated sections, each section having inherent sets of Diodic Points. **Each dimensionalized aspect of the Hova Bodies becomes a *Level of the Auric Field* within the bio-energetic system.** There are **12 Inner Layers of the Auric Field**, which correspond to Dimensions **1-12** and to the 12 Primary Kathara Centers, Seed Seals and Axiom Lines of the *first 4 Hova Bodies*. There are **3 Outer Layers of the Auric Field** that correspond to Dimensions and Seed Seals **13-15**, to the *first 3* Kathara Centers and Axiom Lines of the **Universal Kathara Grid** and to the *5th Hova Body*.

Seed Seals, Axiom Lines, Consolidated Diodic Points, Chakras and Auric Levels

The points in the Kathara Grid where the Single Axis Axiom Lines cross over and through the Central Vertical Kathara Line in the Level-1 Grid and the Horizontal flash-lines of the Seed Crystal Seals form fixed, **consolidated Diodic Points –larger vortices** of energy composed of many smaller Diodic Vortices. **The larger vortex points are called CHAKRAS.** Each Chakra Vortex is formed upon the electro-tonal program of its corresponding Dimensional Frequency Band, Kathara Center, Seed Seal and Axiom Line, and **each chakra corresponds to one Auric Field Level, DNA strand imprint and level of consciousness.** As *3 Signets, Seed Seals* and *Axiom Lines* correspond to each *Hova Body Shield*, **3 Chakras regulate and correspond to each of 5 Hova Bodies**, for a total of **15 Primary Chakras. Primary Chakras 1-12** correspond to Kathara Centers 1-12 in the **Personal and Planetary Kathara Grids** and **Primary Chakras 13-15** correspond to the *first 3 Kathara Centers and Axiom Lines* of the **Universal Kathara Grid.** Chakras 13-15 are considered **Universal Trans-stellar Chakras.**

The *Chakras* regulate the *transduction of flash-line sequences* from the *Dimensionalized Unified Field* and *Diodic Vortices* of the Axiom Lines and Auric Field Levels into the *Hova Body tissue capsules* and *rivulet channels* that connect to the physical, electro-magnetic, chemical and hormonal aspects of the visibly manifest form. Chakras regulate the flow of consciousness-energy between the dimensionalized Auric Field levels of the Hova Bodies, keeping the various multi-dimensional stations of consciousness separate, in order to allow for experiential perception of holographic manifestation. Each **Chakra circulates** its Single Axis flash-line sequence along the **angle of projection** carried within its corresponding **Axiom Line,** and this *angle of flash-line projection* determines the **Angular Rotation of Particle Spin (ARPS)** – or the *axis upon which particles rotate*- within each dimensionalized portion of the anatomy. There is a **90-degree shift** in the *flash-line angle of projection* from one Chakra and dimensional band to the next, which creates a 90-degree shift in the **Angular Rotation of Particle Spin** between each dimensionalized portion of the anatomy. Between each 3-dimensional Hova Body there is also a **45-degree reverse shift** in flash-line projection angle, which creates a **45-degree** rather than a 90-degree **forward shift** of flash-line projection angle between each Hova Body and between each **set of 3 Chakras** (Chakras 1-2-3 * 4-5-6 * 7-8-9 * 10-11-12 * 13-14-15). The precise angles of flash-line projection and ARPS between Chakras, creates the perceivable 3-dimensional hologram.

Chakras

A Chakra appears within the bio-energetic field as **2 cones or spirals of multi-colored energy that are linked together at the tip** in the **Central Body Current** through the **Seed Crystal Seal** out of which the cones emerged. **Chakras 2-6** extend from the Seed Seals outward through the body and Auric Field on a **Horizontal axis** and **Chakras 8 and 9** intersect with *6 and 7* at the *Pineal Gland* on **2 rotating Diagonal axes** running through the head. **Chakras 1 and 7 and 10-14** are stacked above and below the body on a **Vertical Axis** aligned with the *Central Body Current* (CBC) and the *Central Vertical Kathara Line*. Chakras 1&7, 10&12, 11&13 and 14&15 are actually the bottom and top cones of **4 Large Vertical Chakra Spirals**, but are counted as separate chakras. **Chakra 15** is a **"roving chakra"**, which travels the course of the **12 Primary Axiom Lines** far out in Level–15 of the Auric Field. Each Chakra draws in energy from and transmits energy into the dimensional Unified Field corresponding to the *chakra number* and carries as its **dominant perceivable color the hue** characteristic to the **primary frequency band wave-length** of its **dimensional affiliation.**

Primary Color of Chakras

Chakra	Dimension	Color	Chakra	Dimension	Color
CK-1	D-1	RED	**CK-8**	D-8	GOLD
CK-2	D-2	ORANGE	**CK-9**	D-9	SILVER
CK-3	D-3	YELLOW	**CK-10**	D-10	BLUE-BLACK
CK-4	D-4	GREEN	**CK-11**	D-11	SILVER-BLACK
CK-5	D-5	BLUE	**CK-12**	D-12	WHITE
CK-6	D-6	BLUE-VIOLET (INDIGO)	**CK-13**	D-13	PALE TURQUOISE
CK-7	D-7	VIOLET	**CK-14**	D-14	PALE YELLOW
			CK-15	D-15	MAGENTA-PINK

A Chakra will also carry, in lesser saturation, the colors/frequency bands associated with all of the dimensions below its own dimensional affiliation.

Chakras and Dimensional Merkaba Fields

When the Seed Seal at the core of a Chakra is released, the **2 cones of the Chakra move through each other** to form a **Dimensional Merkaba Field-** *2 counter-rotating spirals of electro-magnetic energy that take the form of a* ***Star-tetrahedron***. When the Seed Crystal Seal releases, the corresponding scalar-wave points within the Hova Body Shield are ***released from Dimensional Phase-lock*** – they are released from *polarity* by the re-combining of their inherent *Particum-Particles* and *Partika-anti-particles*. Through re-combining Particum and Partika the scalar standing-wave points undergo **fusion**, then a reciprocal **fission** and **replication** within the Harmonic above. Through release of the Seed Seals the *Dimensionalized Auric Field Levels* of the *lower Hova Body* align along the same *Angular Rotation of Particle Spin* / angle of flash-line projection, which transmutes them into the *angle of flash-line projection* and *Partiki Phasing rate* of the Dimensionalized Auric Field Levels of the Hova Body from the *Harmonic above*. **Seed Crystal Seals are released by infusing them with the frequencies/ flash-line sequences of the corresponding dimension from the Harmonic above.** Progressive release of the Seed Seals within the Chakras creates opening and merging of the Kathara Lines in the Level-1 12-Tree Grid, and merging of the *Hova Bodies, Shields, Diodic Points, Auric and Identity Levels* and *DNA strands, through which activation of the* ***Merkaba Vehicle*** *and de-densification of matter occurs.*

Auric Level, Chakra and Hova Body Correspondences

Each 3-Dimensional Scalar Shield forms a Spherical Electromagnetic Domain called a Hova Body.

Each Hova Body manifests the structure, and governs the function, of 3 Primary Chakras.

Chakras draw energy in from, and transmit energy into, the **Unified Fields** of each Dimension. Each Chakra carries as its **Primary Color** the hue associated with the **wave-length** of the dimensional frequency band to which the Chakra corresponds.

Each Chakra corresponds to a level of the *Auric Field* and one *Axi-A-Tonal Line*. Each ***Hova Body*** corresponds to a set of 3 Chakras and contains 3 dimensionalized frequency bands that form the dimensional levels of **Auric Field.**

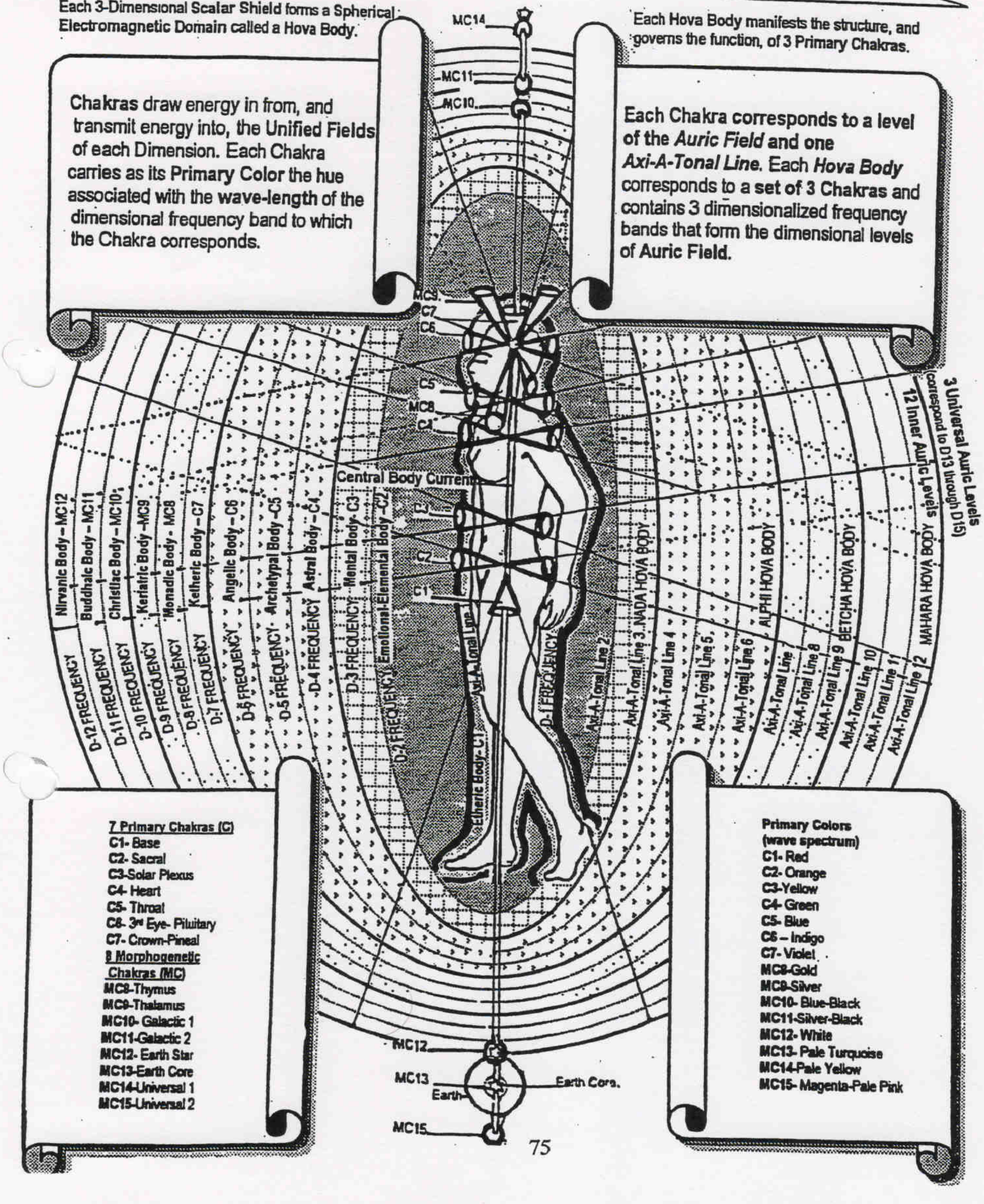
MC14
MC11
MC10
MC9
C7
C6
C5
MC8
C4
Central Body Current
C3
C2
C1
Nirvanic Body – MC12
Buddhaic Body – MC11
Christiac Body – MC10
Keriatric Body – MC9
Monadic Body – MC8
Ketheric Body – C7
Angelic Body – C6
Archetypal Body – C5
Astral Body – C4
Mental Body – C3
Emotional-Elemental Body – C2
Etheric Body – C1
D-12 FREQUENCY
D-11 FREQUENCY
D-10 FREQUENCY
D-9 FREQUENCY
D-8 FREQUENCY
D-7 FREQUENCY
D-6 FREQUENCY
D-5 FREQUENCY
D-4 FREQUENCY
D-3 FREQUENCY
D-2 FREQUENCY
D-1 FREQUENCY
Axi-A-Tonal Line
Axi-A-Tonal Line 2
Axi-A-Tonal Line 3
NADA HOVA BODY
Axi-A-Tonal Line 4
Axi-A-Tonal Line 5
Axi-A-Tonal Line 6
ALPHI HOVA BODY
Axi-A-Tonal Line
Axi-A-Tonal Line 8
Axi-A-Tonal Line 9
BETCHA HOVA BODY
Axi-A-Tonal Line 10
Axi-A-Tonal Line 11
Axi-A-Tonal Line 12
MAHARA HOVA BODY
12 Inner Auric Levels
3 Universal Auric Levels (correspond to D13 through D15)
MC12
MC13
Earth
Earth Core
MC15

7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal

8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

Primary Colors (wave spectrum)
C1- Red
C2- Orange
C3-Yellow
C4- Green
C5- Blue
C6 – Indigo
C7- Violet
MC8-Gold
MC9-Silver
MC10- Blue-Black
MC11-Silver-Black
MC12- White
MC13- Pale Turquoise
MC14-Pale Yellow
MC15- Magenta-Pale Pink

PRIMARY COLOR OF CHAKRAS

Chakra	Dimension	Color	Chakra	Dimension	Color
CK-1	D-1	RED	CK-8	D-8	GOLD
CK-2	D-2	ORANGE	CK-9	D-9	SILVER
CK-3	D-3	YELLOW	CK-10	D-10	BLUE-BLACK
CK-4	D-4	GREEN	CK-11	D-11	SILVER-BLACK
CK-5	D-5	BLUE	CK-12	D-12	WHITE
CK-6	D-6	BLUE-VIOLET (INDIGO)	CK-13	D-13	PALE TURQUOISE
CK-7	D-7	VIOLET	CK-14	D-14	PALE YELLOW
			CK-15	D-15	MAGENTA-PINK

A Chakra will also carry, in lesser saturation, the colors/frequency bands associated with all of the dimensions below its own dimensional affiliation.

AXI-A-TONAL LINES, THE HARA LEVEL, 6 HARA POINTS AND COMPLEXES

Axiom Lines, Rotating Flash-line Projection and the Human Body

Each Seed Crystal Seal within the Signets of the Hova Body Shields holds the electro-tonal program for a Single Axis Axiom Line flash-line projection. **The flash-line sequences of each Axiom Line are not simply *stationary* single-line projections of flash-lines, but rather operate as *rotating* single-line projections.** Each rotating Single Axis flash-line projection is ***synchronized*** with the other rotating Single Axis flash-line projections of Axiom Lines from the other Seed Crystal Seals. Through the ***rotational synchronization*** of the Axiom flash-line sequences, the flash-lines of each Axiom Line regularly cross through each other, and through the vertical and horizontal flash-line sequences from the Level-1 and Level-2 Kathara Grid. The points of *synchronized flash-line crossover* form ***consistent points of intersection*** through which the **Level-3 Diodic Vortices** and **Chakras** manifest. Within the human body and Auric Field the **consistent points of flash-line crossover** can be charted in the form of **fixed lines of energy** that run through the body. **The charted Axiom Lines within the human body represent series of regularly scheduled flash-line intersection points, created through the synchronized rotation of the Single Axis Axiom flash-line sequences.**

Axiom A Lines – "Axi-A-Tonal" Lines in the Human Body

The consistent points of Axiom Line flash-line crossover that form fixed lines of energy within the body can be charted as **Axiom A and B Line** energy flows within the body. The Axiom A Lines correspond to the body manifestation in the particle universe, the B Lines to the body double in the anti-particle universe. **For Kathara Healing applications the A Lines – Axi-A-tonal Lines- of the particle body are used for various purposes of facilitating Kathara Center realignment and revitalization of the Body-Mind-Spirit System.** Within the body the Axi-A-Tonal Lines can be charted as **Primary Flow Lines** – conduits through which energy-consciousness, in the form of scalar-point flash-line sequences, circulates through the morphogenetic field. The fixed flash-line crossover points that form the **charted Axi-A-Tonal Lines** appear in the body as a set of **12 Vertical Flow Lines**, one running from each of 12 Auric Field Levels, which run through the body within the **Level-2 Crystal Seals Kathara Grid**. Each Axi-A-Tonal Line corresponds to one of 12 Dimensional Frequency Bands, Kathara Centers, Seed Crystal Seals, Chakras and DNA strand imprints. **The 12 Primary Axi-A-Tonal Lines appear to enter and leave the body through the head and feet, with conduits running through the arms, hands and out the finger-tips, and through the legs, feet and out the tips of the toes. Each Axi-A-Tonal Line appears as a colored cord of energy, its hue determined by the wave length characteristic to its corresponding dimensional frequency band.** In charting the positions of the 12 Primary Axi-A-Tonal Lines in the body, **numbering** and **color** are used to denote the *Dimensional correspondences* of each Axi-A-Tonal Line.

Axi-A-Tonal Lines 11 & 12 and the "Hara Line"

Axi-A-Tonal Lines 11 and 12, which correspond to the 11th and 12th Dimensions, Kathara Centers, Seed Seals and Chakras, run vertically through the center of the body along the Kathara Grid Level-1 **Central Vertical Kathara Line**, forming a ***twisted "rope"*** of Dark Silver (D-11) and White (D-12) energy that wraps around the Central Vertical Kathara Line Flash-line Sequence. The **Central Vertical Kathara Line,** which appears as a **thin Silver-gold flash-line of energy** running from the 14th Chakra (36"above the head) through the body and into the 13th Chakra at Earth's Core, is imbued with frequencies from the **12th, 11th and 8th Dimensions** – the frequencies of White, Dark Silver and Gold.

In *Level-2 Kathara Grid*, the rotating single axis flash-lines of *Axi-A-Tonal Lines 11* and *12* spiral around the thin Silver-Gold flash-line of the *Central Vertical Kathara Line,* merging into a thin, vertical tri-tonal flash-line of primarily *Gold* hue that becomes the core of the *Central Body Current.* In ancient Eastern Healing Systems this **thin Gold Vertical Line** is called the **"Hara Line"**, and is often considered the foundation of the manifest form. In truth, the **Level-1 Kathara 12-Tree Grid** and the *Signets, Crystal Seals* and *Shields* of **Kathara Level-2** are the deeper foundations upon which the Hara Line manifests.

The *Hara Level* of Kathara Level-2 Crystal Seals Grid

The mechanics of the Crystal Seals in the Level-2 Kathara Grid are complex, but can be employed with relative ease through working with the energetic dynamics of the ***outer levels* of the Level-2 Kathara Grid**, the level in which the *Hara Line* can be viewed using *9th-dimensional Higher Sensory Perception.* We refer to this level of the Level-2 Crystal Seals Grid as the **HARA LEVEL.** The Hara Level is composed of the Hara Line, and 6 of the 15 Star Crystal Seals of the Kathara Level-2 Crystal Seals Grid.

The 6 *Hara Points* (Crystal Seals) apparent from the Hara Level are as follows:

1. The 9th *Silver* Core Star Crystal Seal, called the **HARA CENTER**
2. The 6th *Indigo (or Blue-Voilet)* Star Crystal Seal, called the **SOUL STAR**
3. The 12th *White* Star Crystal Seal, called the **GALACTIC STAR**
4. The 3rd *Yellow* Star Crystal Seal, called the **SUN STAR**
5. The 8th *Gold* Star Crystal Seal, called the **EARTH STAR**
6. The 10th *Blue-Black* Star Crystal Seal, called the **EARTH CORE SEAL**

The **6 Hara Points** govern and regulate the function of **Complexes** of *energy synthesization and distribution* that control the operations of all primary body systems and bio-energetic field dynamics. In *Kathara Healing Level-2* we will explore mechanics of the Hara Level. In Level-1 Kathara Healing we will simply become familiar with Hara Level as an element of 15-dimensional human anatomy.

THE HARA AND THE KUNDALINI

The Silvery-gold energies of the **Hara Center** (9th Star Crystal Seal) are interwoven at the **base of the spine** with what is referred to as the **KUNDALINI life-force energies.** The Kundalini energies represent the **PRIMARY ENERGY CURRENTS** through which the incarnating consciousness anchors its morphogenetic field within the **Fetal Body** through the **FIRST 8 CELLS** of conception. **The Hara Center energies interweave at the base of the spine with the frequencies of the first Star Crystal Seal, the *RED Star Seal*,** in the First Cell. When the Hara Center Seal is dormant and activated only to *the 3-Dimensional Level* its Pale Silvery-gold frequencies appear to be encased within a **crimson RED sheathe** of 1st-Dimenional energy, often **making the Hara Center and Hara Line appear to be RED in color.** As the Hara Center comes into activation through progressive release of the 9th Star Crystal Seal, the Gold 8th-Dimensional and Silver 9th Dimensional frequencies of the Hara become the dominant hues, giving the Hara Center and Hara Line its natural **Silvery-gold color.** With full activation of the Hara Center and **12th-Dimensional Maharic Integration**, the tri-tone frequencies of the Hara Center and Hara Line take on a **tri-color spectrum of Pale Silver, Gold and Red-Violet** (pale Magenta), as the Kundalini primary life force currents at the base of the spine activate within the Central Vertical Current of the body. Full activation of the Kundalini merges the 7th-Dimensional **Violet**, 1st-Dimensional **Red**, 8th-Dimenional **Gold** and 9th Dimensional **Silver** frequencies with the full spectrum **White-Silver** energies of the 12th Dimensional Maharic Current.

The Hara Center, like all Star Crystal Seals, is one of the primary regulatory elements of the natural Kundalini energies. As the Kundalini energies at the base of the spine are brought out of higher dimensional dormancy, through progressive activation of the Star Crystal Seals in the Kathara Level-2 Crystal Seals grid, the Hara Center progressively reflects these changes by expanding in size and energy processing capacities and taking on the hues (wave spectra) of the higher dimensional frequencies. The **Kundalini energies** at the base of the spine are the **key to physical cellular transmutation** of the body. Kundalini energy currents **regulate the physical body's position in space-time**, keeping the physical body phase-locked into the planetary Time Cycle in which it was conceived, following the mathematical programs set by the Star Crystal Seals and the Kathara Grid. For physical transmutation to occur, the Kundalini energies of the higher dimensions must be brought into embodied activation, setting the processes of the **Internal Templar Complex** in motion.

The **DNA Template** manifests upon the programs set by the **Kathara Grid.** As the Kathara Centers and their corresponding Star Crystal Seals, Shields, Kundalini currents, Cranial-Sacral (Pineal-Tailbone) Seals, Seed Crystal Seals and Chakras activate, dormant DNA Strand Template also activate, progressively altering the **Angular Rotation of Particle Spin**, energy-processing capacities, **metabolic orientation** and **neurological processes** of the body and molecular structure. As these physical changes occur through activation of the Crystal Seals and Kundalini energies, the **Merkaba Vehicle** (interdimensional counter-rotating electro-magnetic fields in star-tetrahedral form) progressively builds within the bio-energetic field, allowing the physical body **release from phase lock** within its home space-time vector. **Activation of the Hara Center represents one of the stages of this natural transmutational process.**

THE HARA CENTER

The 9th Core Star Crystal Seal, the Hara Center (Tan Tein) and the D-9 Thalamus Complex

The **Star and Seed Crystal Seals** of the Level-2 Kathara Grid that hold consciousness into the dimensional phase-lock set by the Level-1 Kathara Centers accrete along the Central Vertical Kathara Line and the Central Body Current. **From the *Hara Level* – the *outer portions of the Level-2 Kathara Grid*** - the *12th, 9th and 6th Star Crystal Seals,* **appear as *areas of condensed light* positioned along the *vertical Hara Line*.**

The **9th Star Crystal Seal**, located about 2 " below the Navel is known as the ***Core Star Crystal Seal*** – the point of **8th and 9th dimensional frequency** of the Harmonic-3 Over-Soul Matrix (***Teuric Shield***) that **anchor the personal morphogenetic field into that of the Earth** and serve as the **center of gravity** for the adult human body. **The *Core Star Seal* appears from the Hara Level as a 2" ball of Gold (D-8) energy encased in a thin Silvery (D-9) membrane; in Eastern theory this ball of energy is called the *Tan Tein*, and is used for direction of bodily and Earth energies in advanced martial arts and healing practices.** In Kathara theory we simply refer to this **energy ball** as the ***Hara Center***, or the 9th **Core Star Crystal Seal.**

The ***Hara Center*** is the point at which the dimensionalizing identity enters the **Etheric Matter densities** of Harmonic Universe-3. The Hara Center represents the point into which the **Over-Soul Identity,** its **Monadic Core, the Betcha Hova Body** and **Kathara Centers 7, 8 and 9** first anchor within the body. The Hara Center holds the morphogenetic imprint for **DNA strands 7, 8 and 9,** corresponds to **Chakra Centers** and **Axi-A-Tonal Lines 7,8 and 9** and is a key point for repatterning the electro-tonal programs of the **Teuric Shield** and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Hara Center is associated with the ***Thalamus Complex***, the chemical translation and regulatory system that governs the **9th-Dimensional interface** between various **brain-function areas** and the glandular and metabolic operations of the **Thymus Complex, Gonads, Lymphatic System** and corresponding body systems. Through the ***9th-Dimensional Thalamus Complex*** the electrical impulse, chemical and hormonal relationships between the ***10th-Dimensional Shara Complex*** and the ***8th-Dimensional Thymus Complex*** and its related body systems are governed.

The Hara Center and the Thalamus Complex link the functions of the **Mahara** and **Betcha Hova Bodies,** the **Maharic** and **Teuric Shields** and the **Avatar** and **Over-Soul** levels of consciousness and identity within the body and bio-energetic field.

The primary area of dominion of the *D-9 Thalamus Complex* is *Brain Function*.

THE SOUL STAR

The 6th Soul Star Seal and the D-8 Thymus Complex

Also visible along the Hara Line of Kathara Grid Level-2 is the **6th Star Crystal Seal**, located around the **Thymus** at the top of the breast bone and associated with the **5th and 6th dimensional frequencies** of the Harmonic-2 Soul Matrix (*Doradic Shield*). The **6th Star Seal**, visible as a **pale Blue-Violet-white point of diffused light** along the Hara Line, is called the ***Soul Star Seal.***

The *Soul Star Seal* is the point at which the dimensionalizing identity enters the **Semi-Etheric Matter densities** of Harmonic Universe-2. The Soul Star represents the point into which the **Soul Identity**, its **Archetype Core**, the **Alphi Hova Body** and **Kathara Centers 4, 5 and 6** first anchor within the body. The Soul Star holds the morphogenetic imprint for **DNA strands 4, 5 and 6**, corresponds to **Chakra Centers** and **Axi-A-tonal Lines 4, 5 and 6** and is a key point for repatterning the electro-tonal programs of the **Doradic Shield**, and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Soul Star is associated with the **Thymus Complex**, the chemical translation and regulatory system that governs the **8th-Dimensional interface** between the **Pineal, Pituitary**, and **Thyroid Glands**, which govern the chemical and metabolic operations of the **hormones, Adrenal Glands, Endocrine System, Spleen**, and *crystalline* formation of the **Red Blood Cells** and **Hemoglobin.**

The *Thymus Complex* and its related glandular and metabolic systems, govern the *cyclic rhythms* by which the **DNA strands** will activate through the **Pineal Gland**, the rate of *growth and degeneration* of body tissue through the **Thyroid Gland** and through the **Pituitary Gland** and its inter-glandular processes, the functions of *electro-tonal impulse conversation, translation and distribution* through the **brain, Central Nervous System** and **blood** are regulated.

The *Thymus Complex* controls the **inter-glandular systems balances** between the **Pineal, Thyroid and Pituitary Glands** and regulates the *primary operations* of electrical impulse and chemical translation and transfer throughout the brain and body systems via the **Thymus Gland**. The *8th-Dimensional Thymus Complex* is governed by the functions of the *9th-Dimensional Thalamus Complex* and the *10th-Dimensional Shara Complex.*

The Soul Star and the Thymus Complex link the functions of the **Betcha** and **Alphi Hova Bodies**, the **Teuric** and **Doraidic Shields** and the **Over-Soul** and **Soul** levels of consciousness and identity within the body and bio-energetic field.

The primary area of dominion of the *D-8 Thymus Complex* is *Metabolic Function.*

THE SUN STAR

The 3rd Sun Star Seal and the Nadial Complex

Also present, but less distinctly visible along the Hara Line of Kathara Grid Level-2 is the **3rd Star Crystal Seal**, located just beneath the **Hara Center** of the *9th Star Seal* 2" below the Navel and associated with the **2nd and 3rd dimensional frequencies** of the Harmonic-1 Incarnate Matrix (***Telluric Shield***). The **3rd Star Seal**, visible as a subtle **pale Yellow-Orange-Red point of diffused light** along the Hara Line, is called the ***Sun Star Seal***. Due to its proximity to the Silver-Gold *Hara Center* 9th Star Seal the Sun Star is often considered to be part of the Hara Center, but in deeper Hara Level scanning the distinction between these crystal seals can be evidenced. The *Sun Star Seal* is the point at which the dimensionalizing identity enters the **Gross Physical-Matter densities** of Harmonic Universe-1.The Sun Star represents the point into which the **Incarnate Identity**, its **Subconscious-Telluric Core**, the **Nada Hova Body** and **Kathara Centers 1, 2 and 3** first anchor within the body. The Sun Star holds the morphogenetic imprint for **DNA strands 1, 2 and 3,** corresponds to **Chakra Centers** and **Axi-A-tonal Lines 1, 2 and 3** and is a key point for repatterning the electro-tonal programs of the **Telluric Shield,** and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Sun Star is associated with the **Nadial Complex**, the electrical impulse and chemical translation and regulatory system that governs the **4th -Dimensional interface** between the **Thymus Complex**, **Thalamus Complex** and the functions of the **physical body systems,** through which the electrical, chemical and metabolic operations of the **Blood, Cardio-Vascular, Pulmonary** and **Respiratory Systems** are governed. The Nadial Complex governs the operations of **Heart and Breathing rhythms**, the flow of consciousness-energy or **"Prana"** from the higher dimensional anatomy into physical embodiment and regulates many key functions within the processes of **nutrient conversion and distribution** within the Skeletal and soft tissue systems. **The *Nadial Complex* is a key center for the direction of healing energies within the Nada Hova Body, Telluric Shield and physical body systems.**

In advanced applications of healing and manifestation, the Nadial Complex is combined with the attributes of the **Shara, Thalamus** and **Thymus Complexes** and with the **Universal Raja Complex** of Harmonic-5, which links the human body to the planetary, galactic and universal Unified Fields through the *12th and 13th Chakras* and *10th Star Crystal Seal* at Earth's core via the *Planetary Bio-Feed Interface System*. The frequencies of the *3rd Sun Star Crystal Seal*, that govern the function of the *Nadial Complex* can be combined with those of the 9th Core Star Seal *Hara Center* and the *8th and 10th Star Crystal Seals* through the 12th and 13th Chakras, *PBIS, Doradic, Teuric, and Maharic Shields* and *Raja Complex* to form a potent **Teuric-Telluric Current** for **healing and manifestation**, once the *Doradic* and *Teuric Currents* have been activated within the body.

The *Sun Star* and the *Nadial Complex* link the functions of the **Alphi** and **Nada Hova Bodies**, the **Doradic** and **Telluric Shields** and the **Soul** and **Incarnate** levels of consciousness and identity within the body and bio-energetic field.

The primary areas of dominion of the D-4 *Nadial Complex* are *Cardio-Vascular, Pulmonary* and *Nutrient Absorption/Conversion Functions.*

THE GALACTIC STAR

The 12th Galactic Star Seal and the D-10 Shara Complex

The third point of condensed light visible along the Hara Line of Kathara Grid Level-2 is the **12th Star Crystal Seal**, located at the **14th Chakra** 36" above the head and associated with the **11th and 12th dimensional frequencies** of the Harmonic-4 Avatar Matrix (***Maharic Shield***). **The 12th Star Seal**, visible as a **White-Silver point of Light emanating an inverted White-Silver-Gold *cone of energy* at the 14th Chakra**, is the White Star or ***Galactic Star Seal***.

The *Galactic Star Seal* is the point at which the dimensionalizing gestalt identity enters the **Pre-Matter Liquid Light densities** of Harmonic-4, the point at which identity begins its individuation and journey of densification into the wave spectra of perceivably experiential matter. The Galactic Star represents the point into which the **Avatar Identity**, its **Buddhaic Core**, the **Mahara Hova Body** and **Kathara Centers 10, 11 and 12** first anchor within the **Bio-energetic Field** to begin setting the scalar field for the body. The Galactic Star Seal holds the morphogenetic imprint for **DNA strands 10, 11 and 12**, corresponds to **Chakra Centers** and **Axi-A-Tonal Lines 10, 11, and 12** and is a key point for repatterning the electro-tonal programs of the **Maharic Shield**, and realigning and activating corresponding DNA strands, Chakras, Axi-A-Tonal Lines and body regions they govern.

The Galactic Star is associated with the *Shara Complex*, the regulatory system that governs the **10th-Dimensional interface** between the **D-9 Thalamus Complex** and the D-13 portion of the **Rishi Shield** in Harmonic-5. Through the Shara Complex the Scalar Fields holding the gestalt identity of Harmonic-5 Rishi Consciousness Collective- the "*Universal Family of Consciousness*" – translate via electro-tonal conduction, into the more individuated Scalar Fields that house the Avatar identity gestalt Consciousness Collective- the "*Galactic Family of Consciousness*".

The **Shara Complex *regulates the disbursement of consciousness*** in individuated form through the dimensional fields of the Time Matrix and serves to link each individuated and incarnate member of the Geomantic Entity Collective beyond the Time Matrix, to its Avatar identity and Galactic Family of Consciousness in Harmonic Universe-4, through the Universal Family of Consciousness of the Rishi identity in Harmonic Universe-5. The Shara Complex represents the primary facility through which an individual consciousness retains its indelible energetic connection to its greater **family of consciousness**, through which it is directly linked to **Source.**

The Shara Complex governs, regulates and orchestrates the transduction, translation and cyclic rhythms of **electro-tonal projection of consciousness into dimensionalization**, sets the original operational structures of **Kathara Scalar Grid** arrangement and directs the synchronistic **evolution of families of consciousness** through time. Through the operations of the Shara Complex, an identity's cycles of **incarnation and manifestation** within the 5 Harmonics of a 15-Dimensional Time Matrix are regulated, which sets the **core regulatory rhythms of Partiki Phasing**, expansion and contraction of consciousness, and vibration-oscillation ratios that **govern the formation and function** of the **Thalamus and Thymus Complexes** and all aspects of the individuated **Kathara Grid, Bio-energetic Field** and **bodily systems**. The Shara Complex controls the cycles of in-flow and out-flow of energy between the personal morphogenetic field and that of the dimensional Unified Fields, and thus affects all systems of the body and consciousness, with a primary emphasis upon the functions of the **blood, consumption-excretion and immunological facilities** of the body and of the activation of **the higher dimensional sense facilities** within the embodied consciousness.

The Shara Complex links the functions of the **Mahara** and **Raja Hova Bodies**, the **Maharic** and **Rishic Shields** and the **Avatar** and **Rishi** levels of identity within the embodied consciousness. **Primary area of dominion of the D-10 Shara Complex is the Immune System and HSP.**

Earth Seals 8 and 10 and the Raja Complex

Within the Hara level of field depth in which the 3rd Sun Star Crystal Seal can be detected there are 2 other Star Crystal Seals of primary importance stationed along the Hara Line of Kathara Grid Level-2; these are referred to as the **Earth Seals.** The Earth Seals function as intrinsic aspects of the **Planetary Bio-feed Interface System** and the ***Raja Complex.*** The **8th Star Crystal Seal,** called the ***Earth Star Seal*** is located within the **12th Chakra** 6" below the feet and is associated with the **7th, 8th and 12th dimensional frequencies** of the Harmonic-3 Over-Soul Matrix (***Teuric Shield***) and the Harmonic-4 Avatar Matrix (***Maharic Shield***). The Earth Star is visible as a subtle **Gold-Orange point of diffused light** positioned along the Hara Line **6" below the feet**, as the Hara Line leaves the physical body through the 1st Base Chakra to extend downward into the Earth's core. The **10th Star Crystal Seal,** called the ***Earth Core Seal***, is located within the **13th Chakra** where the Hara Line connects into the Earth's Core and is associated with the **9th, 10th, 12th and 13th dimensional** frequencies of the Harmonic-3 Teuric Shield, Harmonic-4 Maharic Shield and Harmonic-5 Rishi Matrix (***Rishic Shield***). The Earth Core Seal is visible as a **pale Blue-Black point of diffused light** positioned in the **Earth's Core,** at the point where the personal Hara Line connects to the center of the Earth's Core at the ***1st and 12th Planetary Kathara Centers.***

The ***8th Earth Star Seal*** is the point at which the dimensionalizing identity combines the 12th-dimensional frequencies of the **Harmonic-4 Pre-matter Liquid Light density** with the 8th and 9th–dimensional frequencies of the **Etheric Matter densities of Harmonic-3** to connect the personal morphogenetic field and Kathara Grid to those of the Planetary Body. The **Earth Star Crystal Seal** can be viewed as the point where **Personal and Planetary Kathara Grids connect to Galactic Kathara Grid.** The Earth Star represents the point into which the **Mahara and Betcha Hova Bodies, Kathara Centers 8, 9 and 12 and the 12th Galactic Star Seal and Shara Complex** first anchor into the **Planetary and Galactic Kathara Grids.** The Earth Star holds the morphogenetic imprint for **all DNA strands for all manifest incarnations in *4 Harmonics of Time,*** corresponds to **Chakra Centers and Planetary Vortices** and **Personal and Planetary Axi-A-tonal Lines 8, 9, 10 and 12** and is a key point for repatterning the electro-tonal programs of the ***Personal*** **Maharic, Teuric, Doradic and Telluric Shields,** and realigning and activating corresponding DNA strands, Chakras, Planetary Vortices, Axi-A-Tonal Lines and personal or planetary body regions they govern. In the **Maharic Shield Activation process** the **Earth Star** is activated to release phase-lock on the **12th Chakra** to open the personal Kathara Grid to the PBIS.

The ***10th Earth Core Seal*** is the point at which the dimensionalizing identity combines the frequencies of the **8th Earth Star Crystal Seal** with the **13th-dimensional frequencies** of **Harmonic-5 Ante-matter density** to connect the personal and planetary morphogenetic field and Kathara Grid to those of the Universal Kathara Grid of Harmonic-5. The **Earth Core Seal** can be viewed as the point where **Personal, Planetary and Galactic Kathara Grids connect to the Universal Kathara Grid.** The Earth Core represents the point into which the **Rishi Identity,** the **Raja Hova Body** and **Universal Kathara Centers 1, 2 and 3** first anchor into the 3rd Earth Star and **12th** Galactic Star Seals to form the Raja Complex of 5 Harmonic scalar fields through which personal, planetary, galactic and universal experiential manifestation takes place. The Earth Core Seal corresponds to **Chakra 13,** and holds the morphogenetic imprint for **all manifest incarnations of persons, planets and galaxies within the 5 Harmonics of Time in one 15-dimensional Time Matrix.** Through the combined frequencies of the **Earth Star and Earth Core Seals** the personal Kathara Grid connects to the Planetary, Galactic and Universal Kathara Grids forming the Planetary Bio-feed Interface System and the ***Raja Complex*** – or ***Universal*** **Bio-feed Interface System** – that links personal, planetary, galactic and universal consciousness in time.

THE RA CENTER

The 3rd Sun Star Seal, 8 and 10 Earth Seals and *RA Center*

When combined with PBIS, **Doradic, Teuric** and **Maharic Shields** and **Raja Complex**, the **3rd Sun Star Crystal Seal** and the **12th Core Star Crystal Seal** *Hara Center merge* to form a *Teuric-Telluric Current* that encompasses the *Seed Crystal Seal* within the **3rd Solar Plexus Chakra**, forming a ***sphere of powerful healing energies*** at the **3rd Chakra** called the **RA Center**. In **Kathara Healing Level-2** we will begin activation of the *Teuric Current* of the Harmonic-3 Betcha Hova Body, through activation of Doradic Current Phase-2 and will explore beginning applications of combining the Doradic Phase –1 and Teuric Currents with the natural *Telluric Current* of the Harmonic-1Nada Hova Body to begin bringing the **RA Center for healing and manifestation** into activation. (For more on the 4 Primary Triadic Healing Currents see Chapter 4 of this manual.)

The Hara Level of Kathara Level-2 and the 6 Primary Hara Points

Though the study of the Crystal Seals of Kathara Level-2 Crystal Seals Grid is complex in detail, one can become skilled in healing facilitation through Crystal Seals mechanics by learning the basic orientation of **the outer levels** of the **Kathara Level-2** scalar grid – ***The Hara Level***. In simple analysis, the ***elements* of the Hara Level can be viewed as follows:**

The Hara Line – the Central Body Current, that is formed by the *11th and 12th Axiom Lines* wrapping around the fixed flash-line sequence of the *Level-1 Kathara 12-Tree Central Vertical Kathara Line.*

The 6 Primary Hara Points: Star Crystal Seals **1.** # 9-Hara Center, **2.** #6-Soul Star, **3.** #12 Galactic Star, **4.** #3-Sun Star, **5.** #8-Earth Star and **6.** #10-Earth Core. Each primary Hara Point governs the function of a **complex** of interwoven scalar, bio-energetic and physical systems. Knowledge of the **6 Hara Points** and the **Thalamus, Thymus, Shara, Nadial** and **Raja Complexes** they respectively govern, allows the healing facilitator a greater understanding of the 15-dimensional anatomy. Such **Foundations of Kathara Healing** are necessary in preparation for **Kathara Healing Level-2** practices that will utilize the **Hara Level** of the **Kathara Level-2 Crystal Seals Grid** in combination with the Kathara Healing Level-1 **Maharic Recoding Process** and other elements to begin activation of the **RA Center** for advanced healing and manifestation.

Axi-A-Tonal Lines 1-10

In our exploration of the Hara Level of the Crystal Seals Grid we have learned that of the **12 Primary Axi-A-Tonal Lines,** (12 primary vertical energy-flow lines within the body and bio-energetic field, created by cross-over points of the rotating, single axis, flash-line sequences projected by each of 12 Seed Crystal Seals), the 11th and 12th Axi-A-Tonal Lines spiral around the Kathara Level-1 12-Tree Grid *Central Vertical Kathara Line* to form the Hara Line and Central Body Current. Axi-A-Tonal Lines 1-10 also run vertically through the body, from the head to the toes and through each arm to the finger tips, connecting body regions to corresponding levels of the Auric Field, Chakras, Kathara Centers, Crystal Seals and DNA. The sequence of Axi-A-tonal Line position is as follows:

Right Side of Body to Center- 3,6,9,2,8 **Center of Body**-11, 12 **Center to Left Side of Body**-. 5,1,10,7,4

Each line of Axi-A-Tonal Lines **1-10** ends in one **finger-tip** and **toe** on the corresponding side of the body, and runs vertically through the top of the head to merge with Axi-A-Tonal Lines 11 and 12 and the Hara Line at **Chakra 10**, and down through the feet to merge with the Hara Line at **Chakra 12.**

The HARA LEVEL- Outer Portion of Kathara Level-2 Crystal Seals Grid

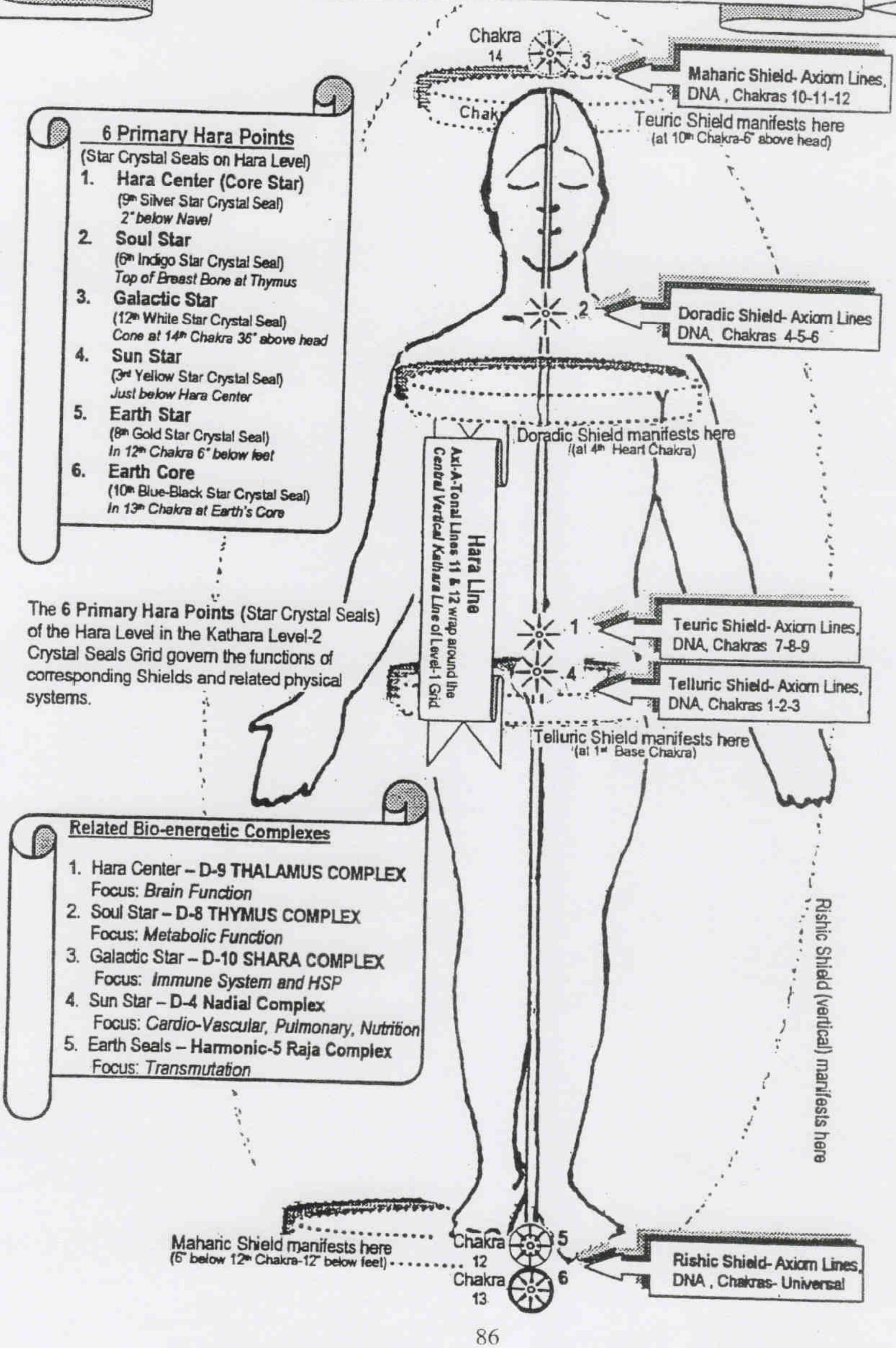

6 Primary Hara Points

(Star Crystal Seals on Hara Level)

1. **Hara Center (Core Star)**
   (9th Silver Star Crystal Seal)
   2" below Navel
2. **Soul Star**
   (6th Indigo Star Crystal Seal)
   Top of Breast Bone at Thymus
3. **Galactic Star**
   (12th White Star Crystal Seal)
   Cone at 14th Chakra 36" above head
4. **Sun Star**
   (3rd Yellow Star Crystal Seal)
   Just below Hara Center
5. **Earth Star**
   (8th Gold Star Crystal Seal)
   In 12th Chakra 6" below feet
6. **Earth Core**
   (10th Blue-Black Star Crystal Seal)
   In 13th Chakra at Earth's Core

The 6 **Primary Hara Points** (Star Crystal Seals) of the Hara Level in the Kathara Level-2 Crystal Seals Grid govern the functions of corresponding Shields and related physical systems.

Related Bio-energetic Complexes

1. Hara Center – **D-9 THALAMUS COMPLEX**
   Focus: *Brain Function*
2. Soul Star – **D-8 THYMUS COMPLEX**
   Focus: *Metabolic Function*
3. Galactic Star – **D-10 SHARA COMPLEX**
   Focus: *Immune System and HSP*
4. Sun Star – **D-4 Nadial Complex**
   Focus: *Cardio-Vascular, Pulmonary, Nutrition*
5. Earth Seals – **Harmonic-5 Raja Complex**
   Focus: *Transmutation*

Chakra
14
3
Maharic Shield- Axiom Lines, DNA , Chakras 10-11-12
Chak
Teuric Shield manifests here
(at 10th Chakra-6" above head)
2
Doradic Shield- Axiom Lines DNA, Chakras 4-5-6
Doradic Shield manifests here
(at 4th Heart Chakra)
Hara Line
Axi-A-Tonal Lines 11 & 12 wrap around the Central Vertical Kathara Line of Level-1 Grid
1
Teuric Shield- Axiom Lines, DNA, Chakras 7-8-9
4
Telluric Shield- Axiom Lines, DNA, Chakras 1-2-3
Telluric Shield manifests here
(at 1st Base Chakra)
Rishic Shield (vertical) manifests here
Maharic Shield manifests here
(6" below 12th Chakra-12" below feet)
Chakra
12
5
Chakra
13
6
Rishic Shield- Axiom Lines, DNA , Chakras- Universal

12 Primary Axi-A-Tonal Lines
of the Kathara Level-2 Crystal Seals Grid

The 12 Primary Axi-A-Tonal Lines
Frequency Color and Correspondences.

(Axi-A-tonal Line, DNA Strand, Chakra, Kathara Center & Seed Seal all correspond to one dimensional number)

D1LINE 1 – Red
D2LINE 2 – Orange
D3LINE 3 – Yellow
D4LINE 4 – Green
D5LINE 5- Blue
D6LINE 6 - Indigo
D7LINE 7 - Violet
D8LINE 8 - Gold
D9LINE 9 - Silver
D10LINE 10 – Blue-Black
D11LINE 11 –Silver Black
D12LINE 12 - White

Axi-A-Tonal Lines are Points where the rotating, single-axis flash-line sequences from the Seed Crystal Seals cross over and through each other, to form 12 primary vertical flow-lines within the body. The Axi-A-Tonal Lines create the web-work of energy upon which the Kathara Grid Level-3 ***Diodic Points*** and the DNA manifest.

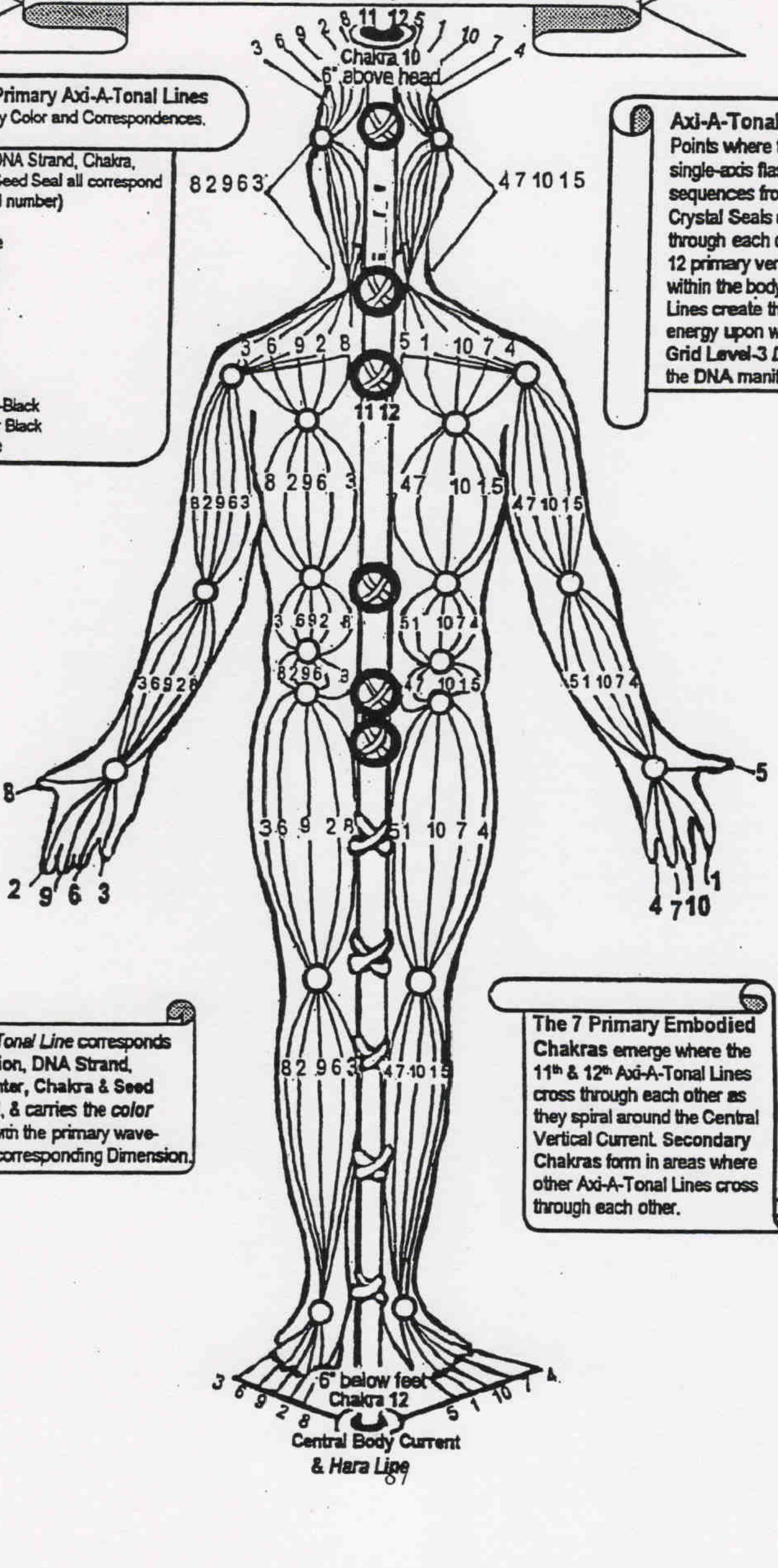
3 6 9 2 8 11 12 5 1 10 7 4
Chakra 10
6" above head
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
11 12
8 2 9 6 3
4 7 10 1 5
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
5 1 10 7 4
8
2 9 6 3
5
4 7 10 1
3 6 9 2 8
5 1 10 7 4
8 2 9 6 3
4 7 10 1 5
3 6 9 2 8
6" below feet
Chakra 12
5 1 10 7 4.
Central Body Current
& Hara Line

Each *Axi-A-Tonal Line* corresponds to a Dimension, DNA Strand, Kathara Center, Chakra & Seed Crystal Seal, & carries the *color* associated with the primary wave-length of its corresponding Dimension.

The 7 Primary Embodied Chakras emerge where the 11th & 12th Axi-A-Tonal Lines cross through each other as they spiral around the Central Vertical Current. Secondary Chakras form in areas where other Axi-A-Tonal Lines cross through each other.

Axi-A-Tonal Lines and the Meridians

The Axi-A-Tonal Lines within the body set the primary holographic scalar field upon which the Meridian Lines – or Secondary Flow Lines- of the body manifest. **The Meridian Lines can be used to access the Axi-A-Tonal Lines and the Axi-A-Tonal Lines can be used to access the flash-line programs of the Kathara Centers and 3 Levels of the Kathara Grid, if the *correspondences* between *the Meridians, Axi-A-Tonal Lines, Kathara Centers* and *Chakras* are understood.** For centuries the **Meridian Lines** and **Chakras** have been used in subtle-energy healing systems. This knowledge can be applied and the potential results of related healing techniques amplified by employing the correspondences between the Meridian and Chakra systems and the Axi-A-Tonal Lines and 3 Levels of the Kathara Grid. Bringing the Kathara Grid into the picture allows for restructuring the **Imprint for Health** within the *causal core* while *expediting* the healing process by restructuring the various other levels of the bio-energetic anatomy.

In Level-1 Kathara Healing we focus primarily upon **Core Morphogenetic Repatterning** – resetting the electro-tonal programs of the Kathara Grid through the uncorrupted programs of the ***Maharic Shield***- to set the "*ground*" into which other healing applications can *anchor*. In later Kathara Healing Levels we will explore technique applications that affect various other levels of the bio-energetic anatomy. **In Level-1 Kathara Healing we strive to create a basic familiarity with the elements of, and the interrelationships between, the various levels of bio-energetic anatomy.** Through understanding the basic correspondences between the **3 Levels of the Kathara Grid** – the *Kathara Centers* and *Kathara Lines*, the *Hova Bodies*, Auric Levels, *Signets, Shields, Crystal Seals* and *Axiom Lines*, and the *Diodic Points* and *Chakras* – and the other aspects of energetic anatomy, such as the **Meridian Lines** and **DNA**, the healing facilitator is equipped with a knowledge base through which more precise healing facilitation can be applied. Meridian Lines and their applications to Kathara Healing will be explored in later Levels of the Kathara Healing Program.

DNA

The DNA Connection

In this Chapter we have explored the various elements of the 15-Dimensional Anatomy of the human form, from the Kathara Grid holographic scalar template, to the Hova Bodies, Shields, Axiom Lines, Chakra Centers, Auric Field Levels and the Hara Level with its related Complexes of energetic interface. Throughout all of these elements of the multi-dimensional anatomy we have seen how each element interrelates with the others through their *correspondences within the structure of the 15-Dimensional Scale.* The elements of the subtle body anatomy also systematically correspond to the structure and function of the human **DNA.** Through direct interaction with various elements of the subtle body anatomy, the operations of corresponding aspects of the human DNA can be repaired, enhanced and altered, to advance applications of healing, and to expedite the evolution of human biology and consciousness. In the *Kathara Healing Level-2 program* we will further explore the correspondences between the subtle body anatomy and the DNA, and methods by which the DNA can be repaired and activated to a more advanced level through employing the use of Kathara Healing mechanics.

The Silicate Matrix 12-strand DNA Pattern

The core human DNA imprint is built upon a **12-Dimensional** Structure, each of 12 strands building up upon one of the Primary 12 Axi-A-Tonal Lines and each strand corresponding to 1 of 12 dimensional frequency bands within the 15-dimensional scale. The original human DNA imprint is called the **Silicate Matrix,** the 12-strand scalar-wave configuration that allows for the progressive **transmutation** of Harmonic-1 carbon-based molecular-elemental structure into the Silica-based structures of higher Harmonic matter. The 12-strand DNA imprint of the human form implies that human biology was designed to embody **12 dimensions of conscious awareness,** or to fully embody the Avatar identity level - the ***"Christed" Identity*** of Harmonic-4 Pre-matter Liquid Light matter density. The human form is organically designed to undergo **full cellular transmutation** through 4 Harmonics of manifestation and their corresponding matter density levels. When the gene code is working properly, the process of cellular transmutation is directed by the embodied human consciousness. In historical terms, the process of ***conscious cellular transmutation*** has been referred to as the process of **"Ascension"**, which is ***Dimensional Expansion,*** the mechanics of which have been hidden with ancient texts dealing with **Merkaba Mechanics.** The **Merkaba Vehicle** of inter-harmonic scalar-wave configurations, which forms within the bio-energetic field, is a consequence or *simultaneous affect* of progressive activation of the Silicate Matrix within the genetic code.

The Silicate Matrix and Subtle Body Anatomy

Just as each Axi-A-Tonal Line of the subtle body anatomy has its correspondence to a Kathara Center, Crystal Seal, Axi-A-Tonal Line, Auric Field Level and Chakra, so too do the DNA strands have their specific affiliations to each *dimensionalized element* of the subtle body anatomy. Each DNA strand translates into electrical-elemental-chemical form from the governing Axi-A-Tonal Line flash-line sequence to which it corresponds. The **Hova Bodies** and **Shields** of subtle body anatomy each correspond to a *set of 3* Axi-A-Tonal Lines, Auric Field Levels and Chakras and thus also correspond to a set of **3 DNA strands,** that together create the experiential projection of separate 3-Dimensional holographic reality fields. Distortions and blockages in the Kathara Grid, Hova Bodies, Shields, Axi-A-Tonal Lines and Chakras appear as distortions and malfunction within the DNA. **The function of the DNA determines the *quantity of interdimensional frequency,* and thus the *amount of dimensionalized consciousness,* that can actively process through and embody within the physical body system.** DNA distortions manifest as distortions of biology and consciousness.

The 12 DNA strands of the Silicate Matrix are *sets of scalar-wave spectra* that correspond to each dimensional band, through which each dimensionalized level of the subtle body anatomy is translated into wave-spectra that is within the range of perceivable matter manifestation. *The DNA can be viewed as the primary structure through which subtle energy-consciousness translates into manifest physicality.* As each full dimensional band contains 12 primary Sub-frequency Bands (See Chapter 1 for Dimensional Structure), each of the 12 DNA strands contains ***12 primary Seed Codes,*** or *consolidated scalar points* composed of electro-tonal flash-line sequences, that correspond to the Sub-Frequency Bands within each dimension. The 12 primary Seed Codes within each DNA strand have an *anti-particle counterpart* composed of frequencies from the Anti-particle system. The 12 particle Seed Codes in each strand are called **BASE CODES,** their anti-particle counterparts are called **ACCELERATION CODES.** The Seed Codes of the core DNA scalar imprint are the *points of consolidated frequency* through which energy-consciousness is drawn in from the dimensional Unified Fields to expand the personal morphogenetic field.

DNA Fire Codes- Genetic Time Codes

Along with the **12 Base Codes** and **12 Acceleration Codes** of each DNA Strand Template there are also 12 Primary DNA Codes called **DNA FIRE CODES** that regulate the relationship between human biology and the **Planetary Time Cycle** in which the biology manifests. The DNA Fire Codes can be conceptualized as existing *between* the Primary 12 Strands of the Silicate Matrix. Each Fire Code is composed of one half of the scalar frequency patterns from each of the two Primary Strands that the Fire Code is placed between. Each of the 12 Fire Codes corresponds directly to one of 12 **Signet Star Crystal Seals** in the Personal Shields of the Internal Templar Complex, and with release of the corresponding **Signet Seal**, the Fire Code comes into activation.

The DNA Fire Codes release the natural scalar barriers between particle and anti-particle DNA Strands, allowing **STRAND FUSION** to occur. Fusion of DNA strands in the particle body with corresponding strands in the anti-particle body creates a progressive transmutation of the scalar wave forms of the DNA Template and Scalar Shields, through which the polarized Particum-**Particles** and Partika-**Anti-particles** merge and transform into their original Omni-polar Partiki **Ante-matter** state. Through strand fusion, the biological carbon-based biology progressively **lessens in density** as its elemental agents are transmuted from the Kathara level, into elemental agents characteristic of the higher dimensional matter densities. **Dimensional Ascension** is achieved through release of the Kathara Grid Seals. Release of the Kathara Grid Seals creates merger of the Hova Bodies, activation of the Kundalini energies, release of the Cranial-Sacral Crystal Seals, **fusion of the DNA Strand Templates via the DNA Fire Codes**, building of the Merkaba Vehicle in the bio-energetic field and the resulting transmutation of bodily matter and shift in Angular Rotation of Particle Spin within molecular construction. Through this transformational process the human biology is released from its organic phase-lock within its current space-time cycle, and is able to engage in **transcendental travel** through the various matter density levels and Time Cycles in and beyond the 15-Dimensional Time Matrix. **As the DNA Fire Codes regulate the body's orientation to space-time, they are also referred to as the GENETIC TIME CODES.**

The DNA Fire Codes are an essential element in the natural process of restoring the natural immortal **Organic Imprint for Health** within the human body. The ability of the human form to become **perpetually self-sustaining** is dependent upon the Kathara Grid and DNA being able to continually renew the natural Primal Order of Maharic Shield scalar arrangement upon which the body is built, to keep the Star Crystal Seals and conduits of interdimensional energy supply open. **Perpetual revitalization of the body** is achieved through full activation of the **12-Strand DNA** pattern. When activated, the **DNA Fire Codes** keep the Crystal Seals, **the interdimensional INTERNAL STAR GATES of energy inflow**, open, so the finite supply of frequency held within the personal Manifestation Template can be continually renewed. Restoration of the Organic Imprint for Health, as held within the **Maharic Shield**, is achieved through progressively clearing distortions of scalar-wave arrangements within the Kathara Grid, Shields and DNA Template and realigning the entire subtle energy and physical energy systems with the Primal Order of the Maharic Shield. This process of whole-being-healing does not occur overnight, as many genetic distortions (which emerge from Kathara Grid misalignment) have served as the dominant bio-electrical program within the human genome for many thousands of years. With knowledgeable practice of Kathara Grid realignment, and supportive healing applications within the various Body-Mind-Spirit systems, these inherited distortions can be progressively healed through Regenesis of the Primal Order as held within the personal Maharic Shield.

Science, "Junk DNA" and Evolution

In the contemporary human genetic code the majority of the DNA strands corresponding to higher dimensional anatomy and consciousness are dormant or damaged due to ancient blockages within the Kathara Grid and its corresponding elements of the subtle body anatomy. Present day humans function primarily on a **3 to 3.5 DNA activation level**, while the imprint and potentialities for strands 4-12 lie dormant within the morphogenetic field and within the fragments of non-functional DNA that contemporary science has labeled ***"junk DNA"***.

The portions of the DNA chain that science has presently identified as the **"Double Helix"**, represent only the **surface portions** of the chemical, elemental and electrical components of the active DNA strands. Science has yet to identify the *multidimensional spectra* of DNA manifestation, and has yet to realize that within the structures of detectable DNA, there are *levels of structure and function* that direct the operations of the *entire genetic imprint*, which are *not presently detectable* by contemporary scientific technique. The human DNA imprint will always appear from external analysis, as a *2-strand Double Helix* configuration, but what is not understood, is that within this Double Helix there are, and will be progressively more, **additional Double Helix strands**, which *fuse together* and add their operational coding into the active DNA imprint. By contemporary standards of categorization, if science were to objectively observe the activated 12-strand pattern (or even an activated 4-strand pattern), the DNA would *not appear to be human at all*, but rather as the seemingly incomprehensible genetic signature of an "unknown" species. As science evolves into comprehension of the multi-dimensional spectrum, the understanding of the true structure and function of DNA will progressively advance.

DNA, Kathara Healing and Bio-Spiritual Evolution

One does not have to wait for scientific theology to catch up with a rapidly evolving humanity in order to begin advancing the functions of the DNA. Through the processes of Kathara Healing, as we work to clear blockages from the Kathara Grid and to reawaken the *Imprint for Health* within the human body, we are *simultaneously, directly affecting* the operational structure of the DNA. As we work to create **Hova Body merger**- *merging of the scalar-grid flash-line sequences of the 5 Harmonic Hova Bodies*- we are progressively reordering portions of the "junk DNA" and stimulating dormant portions of the DNA into activation. As we work with subtle energies to clear blockages within the Kathara Grid, we are progressively purging ancient distortions and mutations within the gene code, distortions that have kept our collective consciousness locked into the limited 3-dimensional perceptual range of Harmonic Universe-1.

It is through the ancient distortions of the human Kathara Grid and the resulting **mutations** of the human DNA, that the multidimensional portions of identity seem separate from each other. The higher levels of conscious awareness that are stationed within the scalar grids of the higher dimensional Hova Body Shields have been unable to *electronically translate* into the lower dimensional identity stations of the Nada Hova Body and its apparent physical manifestation. As we work with Kathara Healing to create progressive Hova Body merger and DNA strand ordering and activation, we will progressively experience *advancement of conscious awareness* and *expansion of the perceptual field* and *Higher Sensory Perception* (HSP). This expansion and advancement occurs because we are able to progressively embody higher frequencies of consciousness and scalar wave flash-line sequences into the personal morphogenetic field, via the energetic conduits of the DNA. **Kathara Healing modalities are the fastest and easiest methods available to repair and advance the function of the DNA and to expedite the organic processes of *Bio-Spiritual Evolution*.**

The Silicate Matrix 12-Strand DNA Template with Hova Body, Scalar Shield and Identity Level Correspondences

The Morphogenetic Field Scalar-wave Template for the Original Human Genome

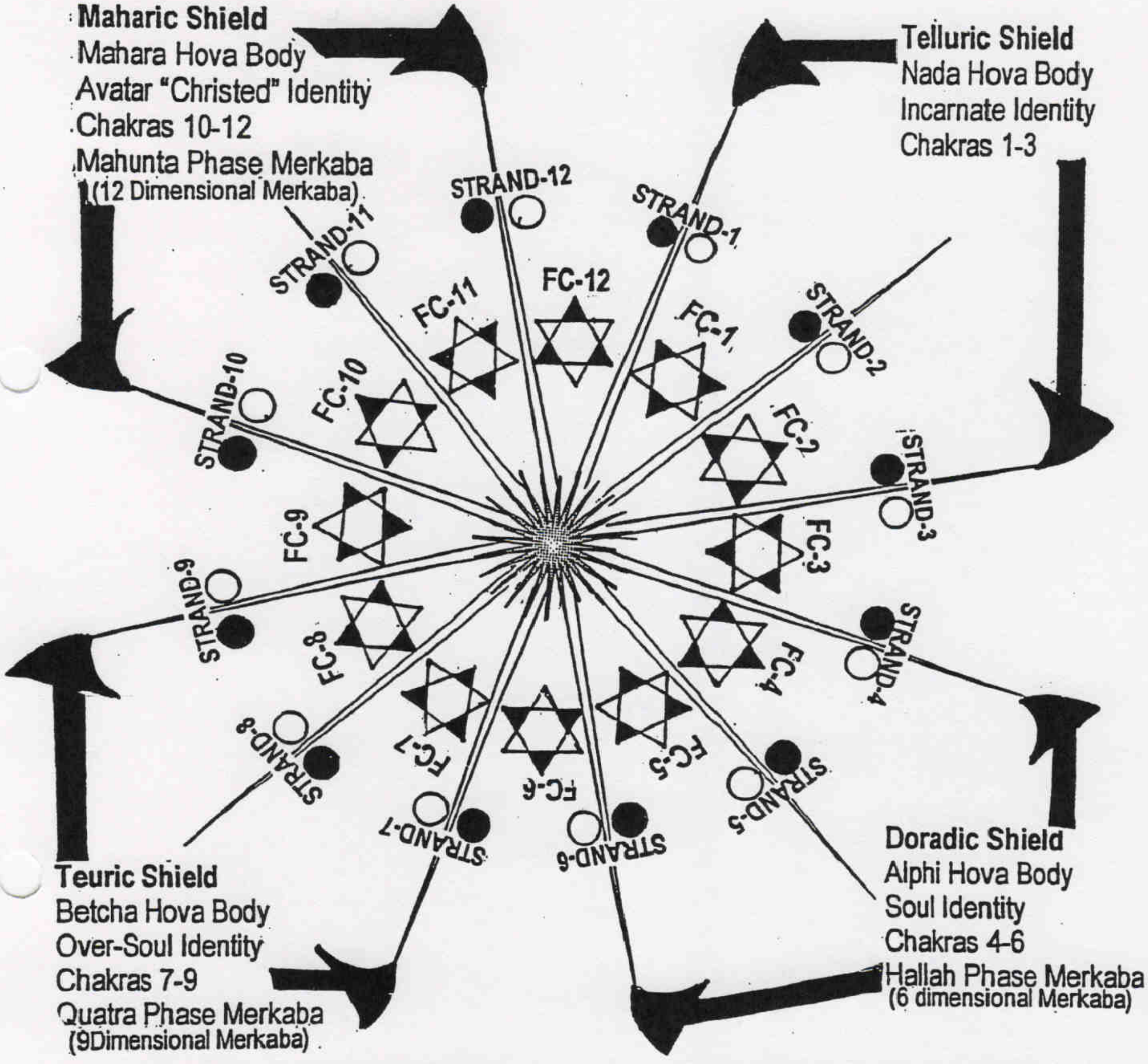
Maharic Shield
Mahara Hova Body
Avatar "Christed" Identity
Chakras 10-12
Mahunta Phase Merkaba
(12 Dimensional Merkaba)
Telluric Shield
Nada Hova Body
Incarnate Identity
Chakras 1-3
STRAND-12
STRAND-1
STRAND-11
STRAND-2
STRAND-10
STRAND-3
STRAND-9
STRAND-4
STRAND-8
STRAND-5
STRAND-7
STRAND-6
FC-12
FC-1
FC-11
FC-2
FC-10
FC-3
FC-9
FC-4
FC-8
FC-5
FC-7
FC-6
Teuric Shield
Betcha Hova Body
Over-Soul Identity
Chakras 7-9
Quatra Phase Merkaba
(9Dimensional Merkaba)
Doradic Shield
Alphi Hova Body
Soul Identity
Chakras 4-6
Hallah Phase Merkaba
(6 dimensional Merkaba)

12 Base Codes per Strand = ● **12 Acceleration Codes per Strand = ○**

12 FIRE CODES between Strands = ✡

(Diagram presents simple conceptualization of DNA Strand Template orientation, not actual geometrical arrangement of wave-form structure and observable strand interrelationship.)

Merkaba Phases – Building the Merkaba Vehicle

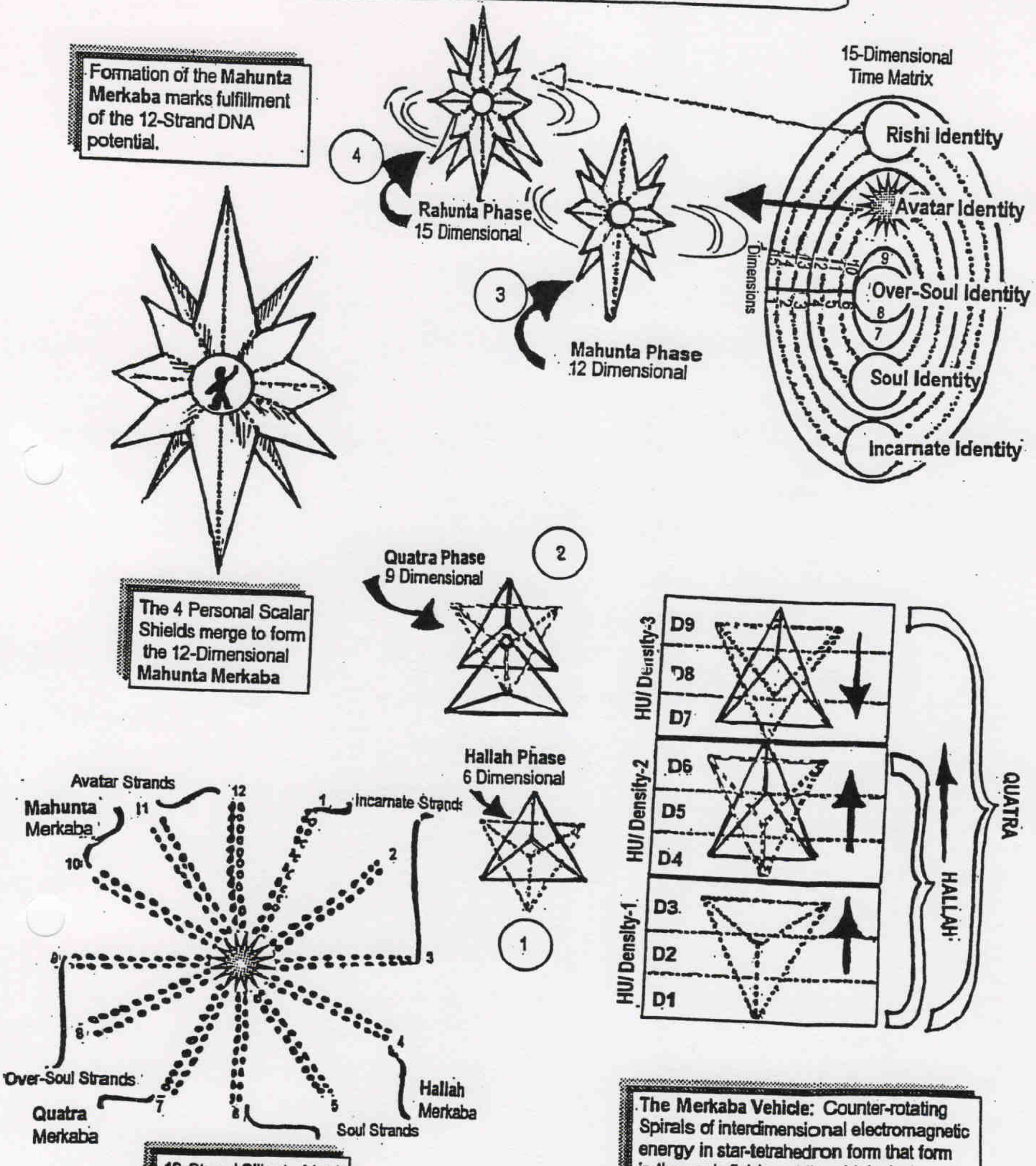
15-Dimensional
Time Matrix
Rishi Identity
Avatar Identity
Over-Soul Identity
Soul Identity
Incarnate Identity
Dimensions
4
Rahunta Phase
15 Dimensional
3
Mahunta Phase
12 Dimensional
Quatra Phase
9 Dimensional
2
Hallah Phase
6 Dimensional
1
HU/Density-3
HU/Density-2
HU/Density-1
D9
D8
D7
D6
D5
D4
D3
D2
D1
QUATRA
HALLAH
Avatar Strands
Mahunta Merkaba
Incarnate Strands
Over-Soul Strands
Quatra Merkaba
Soul Strands
Hallah Merkaba
1
2
3
4
5
6
7
8
9
10
11
12

Formation of the **Mahunta Merkaba** marks fulfillment of the 12-Strand DNA potential.

The 4 Personal Scalar Shields merge to form the 12-Dimensional **Mahunta Merkaba**

12-Strand Silicate Matrix DNA Diagram shows Strand relationships

The Merkaba Vehicle: Counter-rotating Spirals of interdimensional electromagnetic energy in star-tetrahedron form that form in the auric field, enabling biological interdimensional travel and ascension.

THE 12 HUMAN SENSES, APPARTHI AND HSP

Perception and the Hologram

In earlier text we have explained that **manifestation is not what it appears to be**, but rather a **Holographic Projection** of consciousness through the structures of **dimensionalization**. Further, that dimensionalization itself is a *compartmentalization* of consciousness created through units of consciousness (Partiki, Partika and Particum) *assembling in specific mathematical-geometrical ratios* to form **scalar grids**, interwoven fabrics of *fixed points of electro-tonal units* that form layers of standing-waves -**scalar waves** - that serve to "flesh out" the structure of the Hologram, through projection along a *vertical, horizontal* and *diagonal axis*. We have explored the **Kathara Grid** - the *core template* of scalar standing-waves upon which consciousness begins its experience of dimensionalization, the structures of the **15-Dimensional Universe** construction, and how these characteristics of the Hologram relate to forming the structure of **15-Dimensional Human Anatomy.** In understanding the basic foundations upon which our perceivable Hologram forms, we begin to better comprehend the nature of reality and of ourselves within this multi-dimensional reality structure. **Through the Holographic Projection of consciousness within the cosmic Unified Field of Energy-consciousness and its scalar fields of dimensionalization, we perceive the conditions of space, time, matter, externalization of reality and individuation of identity.** Within the Hologram of consciousness we perceive as **solid and objectified** the **wave spectra** and their inherent energy constructs of that which exists **precisely one dimensional band below the present station of our consciousness focus.**

In this section we will explore a bit about the **apparatus of perception** that allow us to perceive externalization in such a manner, the **facilities of energy inherent to the scalar-wave construct of the human form,** which together work to bring the Hologram into perceivable manifestation. We will call the constructs of energy formed through scalar organization, which allow for perception of the manifest Hologram, the **HUMAN SENSES.**

THE *HUMAN SENSES* ARE THE *EXPERIENTIAL RESULT* OF *ORGANIZATIONS OF WAVE PATTERNS* INHERENT TO THE ELECTRO-TONAL PROGRAM AND SCALAR ORGANIZATION OF HUMAN CONSTRUCTION.

What ARE Senses?

Senses are the **scalar-wave apparatus**, and their seemingly manifest parts, that allow the personal morphogenetic field to **synthesize and translate frequency** from the dimensional bands of the Unified Field, into **coherent perceptual and experiential qualities** that give embodied consciousness a **translation of its relationship to other forms of consciousness within the Unified Field.** Senses - or the apparatus which give us the *ABILITY TO SENSE*- are scalar arrangements inherent to our form that **TRANSLATE ENERGY SIGNATURES FROM THE UNIFIED FIELD INTO USABLE DATA OF RELATIONSHIP.**

Like the true substance of the human body and consciousness, the Unified Field exists as a *quantity of dimensionalized energy-consciousness in electro-tonal scalar wave form* - a compartmentalized field of **ENERGY SIGNATURES**, within which a vast number of **consciousness constructs** exist The human organism is one such **construct of consciousness**, and through its inherent form, the *energy signatures* of other scalar forms are translated into perceptual data. **The senses are simply *energy constructs* within the human morphogenetic field, that *separate and translate energy signatures from the Unified Field* into sets of experiential stimuli, through which the contours of the manifest Hologram can be *recognized* and *experienced.***

The APPARTHI and Perception

The sense facilities are *organizations of energy* within the morphogenetic field that hold sets of *"energy receivers" or magnetic points of cohesion*, into which *like frequency patterns* from the Unified Field are drawn in for translation. Every organism has a different configuration of energy receivers within its morphogenetic makeup, and so every organism will have a variation of perception. The **morphogenetic energy receivers**, or **APPARTHI** of an organism's construction serve to *draw certain frequency bands into the morphogenetic field for translation and to block out other frequency bands*, making them unavailable for translation. The frequency bands, or *flash-lines of Partiki Phasing*, that are blocked out of the personal morphogenetic field, create **GAPS OF PERCEPTION** of the Unified Field, through which the Holographic picture of "space between objects", "externalization of form" and "separation of SELF from the Unified Field" becomes perceivable to the organism. **These qualities of *objectification* are the result of BLOCKED FLASH-LINE SEQUENCES, portions of the *scalar fields* of the Unified Field and their inherent *sequences of Partiki Phasing*, that *cannot translate* through the personal morphogenetic field due to the arrangement of the APPARTHI within the personal morphogenetic structure.**

Senses as Scalar-wave Constructs & Expanding our Perception of Perception

If it can be recognized that externalization of experience is a perceptual illusion of consciousness created through selectivity of range within the perceptual field, the true nature of the human sense facilities can be better understood as a pattern of interrelated scalar-wave forms that serve to block out portions of the Unified Field so that other portions may be brought more fully into perceivable view. **When viewing the *physical apparatus* through which the known human senses appear to occur as part a of larger, interwoven system of *scalar construction*, it will be easier to understand that *perception is not a localized phenomena created by and limited to the identified organs of perception*.** The experiential capacities of even the 5 known human senses have a much broader range than what is presently assumed or identifiable through analysis of the manifest sense organs. *If one can understand that perception of manifestation itself is an affect of these greater scalar wave constructs, it will not seem so outrageous to consider the reality and integrity of sensing and perception beyond the range of the presently identified sense facilities.*

ESP, HSP and Evolution

"Extra-sensory perception" is not "EXTRA" at all, it is a *normal attribute* of the *evolving human organism*, as the consciousness ensconced within the *body hologram* continues its *natural expansion* into the greater portions of its evolutionary imprint. Individuals displaying abilities of "Extra-sensory Perception" are not gifted with something *other than what is characteristic to the human form*, **they are simply *a bit ahead in the stage of development* of the natural human perceptual capacities. Everyone has dormant potentials for some degree of "ESP"; like all things within the human organism, ESP is regulated by the *function of the DNA* and its naturally intended *relationship* to the *Kathara Grid and morphogenetic structure*.** Individuals with **less distortions and blockages** within the **Kathara Grid** (and the resulting DNA malfunctions such distortions create) will have *greater ease* in calling the dormant perceptual facilities into activation. *Individuals who consciously attempt to use the greater perceptual facilities will stimulate a more rapid advancement into reclaiming this stage of evolution, than those who refuse to acknowledge the existence of their greater sensing facilities.* **The greater sense facilities bring the *higher dimensional wave spectra* and its inherent reality fields into the *perceptual range* of the earthly embodied human consciousness, and so the greater sense facilities can be considered the HIGHER SENSES, a term far more accurately describing their orientation than the term *Extra Senses*. "ESP" becomes "HSP"- *Higher* Sensory Perception - a *natural attribute of experience for the evolving human organism*.**

THE 5 KNOWN SENSES

The 5 Known Human Senses

In order to understand the nature of the **Higher Senses** it is necessary to understand a bit more about the nature of the **common senses**, or the "*lower senses*" if you will, for the **common senses** are those that are "tuned to" perception of the **lower frequency dimensions. Perception of *Sight, Sound, Touch, Taste and Smell* are *experiential affects* of the *scalar-wave constructs* that exist within the NADA HOVA BODY and the TELLURIC SHIELD, the Kathara Grid and morphogenetic field structures that are composed of and stationed within the frequency bands of dimensions 1, 2 and 3, the first Harmonic of manifestation.** The **APPARTHI**, or ***"energy receivers"*** within the Nada Hova body have their first imprint within the **Kathara Grid Level-2: Crystal Seals** structure. The first pattern for the sense facilities responsible for manifesting the organs of perception of the 5 identified senses exist within the configurations of electro-tonal Partiki units that form the crystalline seals of which the dimension-2 **Telluric Shield** is constructed. Through the Partiki Phasing flash-line projections of the Telluric Shield, the "armature" or 3-dimensional blueprint for the organs of perception are set for manifestation within the body hologram.

The 5 Known Senses and Sound Transduction

The 5 known senses are more intimately related than commonly supposed; ALL represent a form of SOUND TRANSDUCTION - the frequencies, or flash-line sequences, within the 1st, 2nd and 3rd dimensional bands, that form the identifiable wave patterns presently associated with sound waves, are drawn into the Level-2 Kathara Grid Crystal Seals by **Apparthi receivers** that carry like flash rhythms. Once drawn into the Apparthi receivers the sound spectrum becomes modulated into different flash-line rhythms and sent into the **Diodic Points of the Level-3 Kathara Grid.** Once intersecting with the frequencies within the Diodic Points the frequencies from the Apparthi are then **divided/ polarized** and projected up and down the Harmonic-1 dimensional scale, **creating a pattern of energy refraction** within the morphogenetic field and physical body apparatus. The polarized frequencies first appear within the **DNA** as part of the core electrical vibration, and then further transmute into **electrical impulse** that manifests as configurations with the **blood chemistry** and within the **Central Nervous System.** These synthesized electrical impulses are further projected through the **scalar-wave grids** that hold the form of the **identified sensory organs**, finally **culminating in the interrelated qualities of perception associated with the identified sensory organs and their 5 known attributes of sensing.**

THE 12 HUMAN SENSES

The Interwoven Senses and the Hova Bodies

As in the case of the Lower Senses, the *Higher Senses* and their experiential manifestations are the product of consciousness interacting with the *dimensionalized scalar organization* of the human morphogenetic anatomy. Whereas the Lower Senses are affiliated with the transduction of sound frequency within the *1st, 2nd and 3rd Dimensions*, through the structures of the *Nada Hova Body* and the *Telluric Shield* of the *first Harmonic*, the Higher Senses are in like fashion associated with the scalar organization of the *Higher Dimensional Hova Bodies* and their corresponding *Shields*.

In truth, all of the sensing facilities organic to the human form comprise interwoven and interrelated systems of *energy refraction* that together create the perceptual experience of *externalization* within the hologram of manifestation. For example, it is not commonly recognized that the sense of **smell** is in actuality part of the function of the **auditory facility**, and thus represents an intrinsic part of the sense of hearing. It is also not commonly understood that the processes associated with **vision** are not only connected to the processes of the brain and Central Nervous System, but also with the formation of *crystalline structures* with the plasma of the **blood. The senses are not nearly as *separate* and *independently functional* as they appear to be when viewed from outside analysis.** They are rather a part of an **interwoven system of electro-tonal information relay** that spans the 15-Dimensional morphogenetic field anatomy of the human organism.

The 12 Human Senses

In terms of breaking down the sense facilities for structural analysis, such as is done by modern science in identifying the 5 Known Senses, we will describe the sensory attributes of the **Higher Sense facilities** in terms of their *primary orientation*, *location of their primary Apparthi receivers* within the *Hova Body* structures and by the type of *perceptual or experiential attributes* produced through their function. **In such categorization of the human sense facilities there are *7 additional*, as yet unidentified, *sense facilities* operating within and through the human consciousness**, which are collectively responsible for translating the scalar fields of dimensionalized reality into the *illusionary holographic picture* of externalized manifestation and its qualities of space, time, matter and identity individuation. **Collectively there *are 12 Primary Sensing Facilities* within 15-Dimensional Human Anatomy.**

PERCEPTUAL BLOCKAGE, KATHARA GRID DISTORTION AND KATHARA HEALING

Perceptual Blockages

The Higher Senses are connected to the Lower Senses and provide attributes of perceptual experience characteristic to their dimensional levels, but in the present stage of human genetic evolution the Higher and Lower Senses are not fully "plugged into" each other, and so the various higher dimensional fields, and the sensory abilities associated with the Higher Senses, appear to be separate or unavailable to the embodied human consciousness.. **The Higher and Lower Senses plug in through opening of *Kathara Lines* in the *Level-1 12-Tree Kathara Grid*, which creates a reciprocal opening and expansion of the DNA, Hova Bodies and activation of dormant nerve networks within the brain, molecular core and Central Nervous System of the body.**

Presently the human senses are *phase-locked into a 3-dimensional perceptual field* due to blockages between the *3rd and 4th Kathara Centers*, which block the Apparthi receivers within the Nada Hova Body from accepting compatible electrical impulses from the higher dimensional Hova Bodies. This Kathara Grid blockage manifests within the body as fragmented units of DNA (**"Junk DNA"**), that serve to render the natural processes of multi-dimensional data transfer into the Central Nervous System inoperable. Portions of *higher dimensional experience* come to conscious attention in the form of *dream perception* or *intuitive impression,* but the natural clear pathways of multi-dimensional perception are blocked to a large degree within contemporary humanity.

Perceptual Blockages and Kathara Healing

Blockages in the Kathara Grid and their reciprocal blockages in the human perceptual field can be progressively removed through practices of *Kathara Healing*, and utilizing the *Maharic Shield* to reawaken the Organic Imprint for Health.. As Kathara blockages are progressively repaired, the genetic code will expand its function by reassembling and activating the presently dormant portions of the DNA imprint. DNA expansion will create reciprocal alterations in the bodily metabolic rate, brain, blood, glandular and hormonal function, particle spin speed and operation of the Central Nervous System, that will allow for a progressive, natural activation of the Higher Sense facilities and expansion of the human perceptual field into the multi-dimensional arena. **In Kathara Healing the dormant Higher Senses will be progressively brought into activation**, beginning with the **6th -Audurea-Merkaba Sense** and the **7th-Tristet-Cellular Telepathy Sense** through which **Inner audio-visual-direct cognition** skills will progressively develop. These skills will be employed in the processes of Kathara Healing, and in so doing they will be progressively stimulated into greater levels of activation. As the dormant Sense facilities are called into activation, Kathara Healing techniques to remove Kathara Grid blockages are used to realign the *Holographic Template* to restore the **Imprint for Health,** through which natural perceptual expansion will progressively take place.

The 12 Human Senses

The collective spectrum of human sense facilities, including the Higher Senses of 15-Dimensional Human Anatomy can be categorized as follows:

The 12 Human Senses, Hova Bodies and Primary Experiential Attributes

Identified Lower Dimensional Sense Facilities

1. Visual-Sight Sense- ***Nada Hova Body (dimensions 1-2-3)***
2. Auditory-Hearing Sense - ***Nada Hova Body***
3. Tactile-Touch Sense- ***Nada Hova Body***
4. Olfactory-Smell Sense- ***Nada Hova Body***
5. Sense of Taste- ***Nada Hova Body***

Unidentified Higher Dimensional Sense Facilities

6. **Audurea -Merkaba Sense -** *Encompasses all Hova Bodies*
   Primary Attribute: **Manifestation**-circulation of consciousness and frequency accretion
7. **Tristet-Cellular Telepathy** -*Alphi Hova Body (dimensions 4-5-6)*
   Primary Attribute: **Inner Audio-Visual-Direct Cognition** reading energy signatures.
8. **Nurgode-Transmutation Sense -** *Betcha Hova Body(dimensions 7-8-9)*
   Primary Attribute: **Molecular Transmutation** - shape-shifting
9. **Ragode-Transmigration Sense -** *Betcha Hova Body*
   Primary Attribute: **Transmigration** - Projection of Consciousness
10. **Geuard-Transfiguration Sense-** *Mahara Hova Body (dimensions 10-11-12)*
    Primary Attribute: **Bilocation**-multiple holographic manifestations
11. **Shara-Centrifugal Sense-** *Raja Hova Body (dimensions 13-14-15)*
    Primary Attribute: **Electrical Projection** - consciousness enters dimensionalization
12. **Aurt-U - Centripetal Sense -** *Raja Hova Body*
    Primary Attribute: **Magnetic Accretion** -consciousness exits dimensionalization

CHAPTER 2 HIGHLIGHTS

THE IMPRINT FOR HEALTH

- Healing occurs through correcting distortions within the Kathara Grid and Hova Body Shields.

- The human body blueprint was designed to hold 12 Dimensions of Frequency/ Flash-line Sequences and 12 Dimensions of Consciousness, the content of 4 out of the 5 Hova Bodies within the 15-Dimensional Anatomy.

- The **4th Hova Body** - the **Mahara Hova Body** - is composed of Frequencies/ Partiki Phasing Flash-line Sequences of the **10th, 11th** and **12th Dimensions.** It manifests through the 10th, 11th and 12th *Kathara Centers* and the **Maharic Shield** Scalar Grid of the **4th Harmonic.**

- The **Maharic Shield** of *Level-2 Kathara Grid Crystal Seals* holds the original **IMPRINT FOR HEALTH**- the original *Flash-line Sequences* of *Harmonic 4 Pre-matter Liquid Light* - that represent the *point of entry* into perceivable manifestation.

- The **Maharic Shield** can be used to **repattern** the Kathara Grid Holographic Template back to the original **Imprint for Heath.** The Flash-line Sequences of *Dimension 12* can *override* and correct *faulty Flash-line Sequences* that manifest as distortions within the Kathara Grid and Hova Body Shields.

- The **Hierophant Symbol Code** is a *Scalar Standing-wave* pattern containing the electro-tonal programs of the 11th and 12th Dimensions. The Hierophant Symbol code serves as a *Frequency* ***WAVE GUIDE*** through which the **Maharic Shield** can be activated to run 11th and 12th Dimensional Frequency through the Kathara Grid for **repatterning** the **Imprint for Health** within the Core Holographic Template of the body.

- Using the *Hierophant Symbol Code* the **Planetary Bio-feed Interface System™** can be activated to **empower** the Kathara Grid repatterning process via drawing Frequency from the **Planetary Maharic Shield.**

MORPHOGENETIC EXPANSION

GROWTH RATES

- Growth through time is created through **FREQUENCY ACCRETION -** drawing Frequency Bands from the Dimensional Unified Fields of Consciousness into the personal morphogenetic field for expansion.

- When the Kathara Grid is functioning properly there are **set morphogenetic field expansion periods** within the linear growth cycles of the human body.
  1. **Fetal Integration** to age 11-12.
  2. **Harmonic 2 Integration** age 12-22
  3. **Harmonic 3 Integration** age 22-33
  4. **Harmonic 4 Integration** age 33-44

  If the Kathara Grid is working properly the **body does not die** but rather **progressively transmutes** into *less dense states of matter* in an **immortal body form,** via accretion of the 4 Harmonics of Frequency.

- Present **Kathara Distortions** block integration of **Harmonic 2** and begin Molecular Compaction, body degeneration and Death.

LEVELS OF HUMAN ANATOMY

- The **Kathara Grid** is the *Core Holographic Template* upon which the human body and consciousness enter dimensionalization in order to experience manifestation.

- There are **3 Levels** to the Human *Kathara Grid*: *Level-1-* The **12- Tree,** *Level-2* – **Crystal Seals** and *Level-3* – **Diodic Points.**

- *Kathara Level-1 –12-Tree* contains the primary **12 Kathara Centers** and primary **15 Kathara Lines,** that correspond to various regions of the physical body and bio-energetic field and which "ground" consciousness into the framework of dimensionalization.

- *Kathara Level-2 – Crystal Seals* contain the **Signets** *–consolidated points of dimensional frequency* – the **Shields** – *3-dimensional scalar-wave discs emanating from sets of 3 Signets* through which the *Hova Body* energy capsules form– and the **Star Crystal Seals** and **Seed Crystal Seals** through which the *Chakras* and *Bio-energetic Field* levels emerge.

- *Kathara Level-3 – Diodic Points* are **Coordinate Points** where the flash-line sequences from one Shield and its Seals cross over and through those of other Shields to form minute **"White Hole" vortices** of energy-consciousness relay between the Hova Bodies and stations of identity.

- **Hova Bodies** are *3-dimensional scalar grids* in the form of **Concentric Spheres** that form around each **Shield** in the *Kathara Level-2 Crystal Seals Grid,* which serve as "tissue capsule" encasements around each Harmonic of energy-identity within the 15-dimensional scale.

- There are **5** *Dimensionalized* **"Ariea" Hova Bodies** forming human anatomy, which emanate from the **5** 3-Dimensional scalar **Shields** of the *Level-2 Kathara Grid*. Each Shield and Hova Body corresponds to **3 Kathara Centers** in the *Level-1 12-Tree Grid.*

- **5 Ariea Hova Bodies** and Corresponding **Shields:**
  Nada Hova-Telluric Shield-Dimensions1-3
  Alphi Hova-Doradic Shield-Dimensions 4-6
  Betcha Hova-Teuric Shield-Dimensions 7-9
  Mahara Hova-*Maharic Shield*-Dimensions 10-12
  Raja Hova-Rishic Shield-Dimensions 13-15

- The **Signets** and the **Shields** can be used to direct the function of the *Kathara Level-1 12-Tree Grid* and all built upon it. **Kathara Centers** can be used to trigger activation of the *Signets* and *Shields*.

- The first **4 Hova Bodies** and their **Horizontal Shields**, are most important in terms of human evolution. Activation of the **Signets** activates their corresponding **Shield**, which in turn releases the **tissue capsules** between *Hova Bodies*, allowing merger to occur. When the **first 4 Shields** activate the body is able to integrate the Avatar identity and **transmute** into pre-matter liquid-light wave form.

- In *Kathara Healing* the Harmonic-4 **Maharic Shield** is used to restore integrity to all *Shields* and *Hova Bodies* below, which progressively clears the *Kathara Grid* of distortions.

- *Level-1 Kathara Lines* direct the **Vertical** Flash-lines, *Level-2 Crystal Seals* control the **Horizontal** Flash-lines and *Level-3 Diodic Points* govern **Diagonal** Flash-lines. Together the 3 levels of the Kathara Grid form the 3-plane armature upon which the 3-dimensional hologram is created.

SIGNETS AXIOM LINES CHAKRAS

- Each **Signet** within a Hova Body Shield holds the electro-tonal program for one *Seed Crystal Seal*. Each *Seed Crystal Seal* corresponds to one Level-1 **Kathara Center** and holds the program for one primary **Axiom Line** – a *single axis* dimensional flash-line sequence.

- The **12 Primary Axiom Line** programs of the *Seed Crystal Seals* create separate *dimensionalized bands of frequency* within each Hova Body, forming a distinction between *dimensionalized levels* within the bio-energetic **Auric Field**. The Axiom flash-lines cross over and through each other to form the *Coordinate Points* between the Shields through which the *Level-3* **Diodic Point Vortices** emerge.

- The points in the Kathara Grid where Axiom Lines cross over and through the **Central Vertical Kathara Line** in the *Level-1 Grid*, and the **Horizontal flash-lines** in the *Crystal Seals* of the *Level-2 Grid*, form **consolidated Diodic Points** with *numerous Diodic Vortices* – these are called **CHAKRAS**. Each Chakra corresponds to an *Axiom Line, Auric Level, Seed Seal, Kathara Center*, and *Dimensional Frequency Band* and **Merkaba Field**.

- The **12 Primary Axi-A-Tonal Lines** are a set of 12 *Vertical Energy Flow-lines* that run through the body, which are created by the consistent cross-over points of the rotating, single-axis flash-line sequences projected from the **Seed Crystal Seals.**

- **Axi-A-Tonal Lines 11 and 12** combine with the **Central Vertical Kathara Line** to form the **Hara Line**, the foundation of the **Central Body Current.**

- The **Hara Level** is the ***outer level*** of the **Kathara Level-2 Crystal Seals Grid**, can be viewed from *9th-Dimensional perception* and is used in Kathara Healing Level-2 .

- The *Hara Level* is composed of the **Hara Line, 6 Primary Hara Points** and the **Complexes** of energy processing they govern.

- The **Axi-A-Tonal Lines** and **Hara Level** of the *Kathara Level-2 Crystal Seals Grid* are elements of 15-Dimensional anatomy that are used in later applications of Kathara Healing.

- *Axi-A-Tonal Lines* set the core electro-tonal scalar-wave program upon which the **DNA** will manifest. The human DNA imprint contains the program for **12 strands of DNA**, one manifesting upon each of the **12 Primary Axi-A-Tonal Lines.**

- The human form was designed to embody **12 dimensions** of consciousness. The original human genetic code is called the **Silicate Matrix** and when operational it allows human cellular structure to **transmute** through ***4 Harmonics*** *of matter density*.

- The Silicate Matrix is mutated in present day humans due to ancient **Kathara Grid blockages,** which manifest as distortions of form and function within the DNA. "*Junk DNA*" is part of this distortion.

- Contemporary human DNA can be **reverse-mutated,** and its interdimensional functions restored, by applications of **Kathara Healing.**

- Each of the 12 strands of DNA *corresponds to* one of 12 *Dimensional Frequency Bands, Kathara Centers, Seed Crystal Seals, Signets, Axi-A-Tonal Lines*, and *Chakras*. Each **Hova Body Shield** governs the function of a **set of 3 DNA strands.**

THE 12 HUMAN SENSES

- The **human senses** are the *experiential result* of the *scalar standing-wave organization* inherent to the human form.

- Senses are **scalar arrangements** that *translate energy signatures* from the **Unified Field** into *usable data of relationship*.

- The sensing systems are built upon sets of **energy receivers** within the morphogenetic field that are called **APPARTHI**. The Apparthi inherent to Kathara structure serve to **block certain flash-line sequences** out of the morphogenetic field and allow others in to translate into perception.

- The flash-line sequences blocked by the Apparthi create **Gaps of Perception** of the *Unified Field*, which give us the *experience* of **externalization** and **separation of form.**

- *Apparthi receivers* create **refraction and polarization of flash-line sequences** that translate into a variety of identifiable ***holographic sensory perceptions.***

- The sense facilities correspond to the structure of the **Hova Body Shields** and **DNA.** In categorizing the interwoven sense facilities by their *primary orientation*, location of their *Apparthi receivers* within the Hova Bodies and the *experiential attributes* they produce, the human organism can be said to possess **12 Primary Senses.**

- The **5 known senses** are products of the operations of the **Nada Hova Body** and **Telluric Shield**, and thus correspond to **dimensions 1-3**.

- There are **7 as yet unidentified senses** that correspond to the higher dimensional *Hova Bodies, Shields* and presently dormant strands of *DNA*.

- As we evolve to *activate* and *reverse-mutate* the **DNA** imprint, through *Hova Body* alignment and merger, the experiential attributes of the higher dimensional senses will become more apparent.

- Due to their *dimensional affiliation*, the *5 known senses* can be viewed as the **Lower Senses**, the *higher dimensional senses* as the **Higher Senses.** Perception through the *Higher Senses* is called ***Higher Sensory Perception*** or **HSP.**

TAPPING THE INNER MENTOR

Levels of Consciousness and the Perceptual Bridge

In earlier text we have learned of the structure human consciousness takes when entering the scalar fields of dimensionalization for the perceivable experience of manifestation. In comprehending the true structure of multi-dimensional anatomy, with its **Kathara Grid, Hova Bodies** and **Shields, Diodic Points** and **levels of bio-energetic substance**, we can begin to see that there is much more to human consciousness than meets the eye - **much more *energy-identity* operating presently beyond the view of our 3-dimensionally focused earthly personality.**

There are portions of our consciousness stationed within the Hova Bodies of the higher dimensional fields, and these expanded levels of ourselves hold knowledge and participate in perceptual experience beyond the usual range of our waking perceptions. As we evolve and expand through **Hova Body Integration** and the resulting activation of dormant **DNA** components, we will progressively have *greater conscious access* to these expanded levels of consciousness while we are actively focused within the waking 3-dimensionally embodied state. However, **we do not have to wait until our evolution plods along to bring us to these expanded perceptual states,** we can, from our waking state of 3-dimensional consciousness begin to actively tap this hidden reservoir of knowledge and perceptual experience. **We can take an active hand in *expediting the process* of our multi-dimensional evolution by using our dimension-3 *mental body awareness* to direct our higher awareness into our current range of perception.** We can begin to build a **PERCEPTUAL BRIDGE** between our 3rd-Dimensional and Higher Dimensional consciousness. **In taking the initiative for building this Bridge of Consciousness we begin to stimulate our Higher Senses into operation, expediting the natural process of Higher Identity Integration.**

Through building a PERCEPTUAL BRIDGE between our present state of awareness and the levels of our consciousness stationed in the Higher Dimensional Hova Bodies, we can actively expedite our own Bio-Spiritual Evolution.

THE MENTOR

The Perceptual Bridge and Creating the Mentor

The Higher Dimensional levels of human consciousness are not dormant within their respective dimensional bands of function, they are quite awake and aware within their own dimensional fields. As we perceive and interact with the 3-dimensional world within which our present awareness is focused, the Higher Dimensional portions of our consciousness are also actively perceiving and interacting with the dimensional environments within which they are focused. **Higher dimensional aspects of our awareness *seem cut off* from our perceptual view because our consciousness is presently focused through a *genetic code arrangement and Kathara Grid orientation* that blocks out the frequencies of those higher stations of identity.** If we can realize that expanded portions of our consciousness simultaneously exist within the higher dimensional frequency bands, we can also realize that it is possible to access such levels of our own identity by building **Perceptual Bridges** between our 3-dimensional perspective and the portions of our consciousness stationed within the Higher Dimensional Hova Bodies.

We can begin to build a *Perceptual Bridge* between ourselves and our *Higher Selves*, by using simple *frequency modulation techniques* of Kathara Healing, such as Technique #1: *Awakening the Mentor and the 6th and 7th Senses*. In this scalar-grid activating exercise, we will awaken dormant **Partiki Phasing Flash-lines** and activate dormant portions of the **Kathara Lines** in the *Kathara 12-Tree Grid*, to begin opening the channels of communication between the consciousness focused within the *Nada Hova Body* and that focused within the *Alphi Hova Body* of Harmonic-2. **In this exercise we will create the MENTOR, a portion of our consciousness composed of *3rd, 4th and 5th Dimensional frequencies*, that will serve as an *intermediary focus of awareness* through which we can begin to receive guidance and knowledge from our 2nd-Harmonic of consciousness.** Creating the MENTOR will begin the process of activating dormant **DNA** , opening new **nerve networks** within the Central Nervous System of the 3-dimensional body, and activating dormant **APPARTHI** frequency receivers within the *Level-2 Crystal Seals Kathara Grid*. **Through creating the MENTOR, we begin to build the Perceptual Bridge between the 1st and 2nd Harmonic of our conscious awareness.**

The MENTOR is a portion of our identity that we create to serve as an intermediary focus of awareness between the Nada Hova Body and the Alphi Hova Body; a PERCEPTUAL BRIDGE between our first and second Harmonics of Consciousness.

The MENTOR and Kathara Healing

In the practices of Kathara Healing we will use the MENTOR as a source of **guidance, information retrieval, objective observation** and as a **communications link** between our own higher consciousness and the higher levels of consciousness in those who come to us for healing facilitation. **Awakening and working with the MENTOR will become a standard practice in our applications of Kathara Healing.** In creating the MENTOR we are using our creative power of scalar-grid manifestation (thoughts and mental images are scalar-grid morphogenetic field forms) to bring into being a **specialized portion of our consciousness** that can **relay frequency** from our higher Harmonics of Consciousness to our 3-dimensionally focused identity. **With repeated use of the MENTOR focus of awareness we will develop the ability to simultaneously hold 2 Harmonics of consciousness, and we will begin to reclaim the natural flexibility of focus that is the healthful condition of embodied human consciousness.** Through the MENTOR focus of attention we can reach higher levels of cognition and understanding regarding application of healing facilitation, and we will more rapidly call into awakening the Higher Sense facilities of the 15-Dimensional Anatomy.

Other Uses of the MENTOR Focus

In creating the MENTOR focus we will begin to realize that even the 3-dimensional focus of awareness with which we are most familiar is far more flexible than once supposed. The MENTOR is not simply an **inert creation,** but rather a **living** portion of our conscious identity composed of portions of the waking self and portions of the Higher Self. MENTOR can become an **advisor**, a portion of self able to "keep its head above the illusion" when you are confronted with challenging events in the 3-dimensional arena; a part of you able to remain clam, focused and able to make clear decisions beyond the confusion of mental or emotional body processing. MENTOR can be sent into the **Cellular Memory** of the body to locate core issues or reincarnational blockages, and can serve to moderate between sub-personality fragments and the "Inner Child". MENTOR can like wise be used to **harmonize communications** between self and others, as you can send the MENTOR focus into the higher dimensional fields to connect with the higher selves of others for conflict resolution. Playing with the MENTOR focus can also have great rewards. You can focus your attention in the Mentor Space and then direct the MENTOR across the room to begin practicing affects of **BILOCATION.** You can travel with MENTOR to distant locations to develop **REMOTE VIEWING SKILLS**, which will be used in later Kathara Healing levels for **REMOTE HEALING FACILITATION**. MENTOR can become a **vehicle for conscious transport** of your awareness, while awake and also in **dreaming** and **higher dimensional projection.**

TECHNIQUE # 1

AWAKENING THE *MENTOR* AND THE 6TH AND 7TH SENSES

Awakening the Mentor:

1. Place the **fingertips** of both hands and the focus of your attention at the **Navel.** Take a few slow, full breaths then visualize the image of a **Blue Flame** (like a gas stove flame) deep inside the body behind the Navel. Gently **massage** the Navel region while envisioning the **Blue Flame.**

2. **Stop massage** and visualize a **Sphere of White Light within the Blue Flame** behind the Navel. **Place your attention within the White Sphere** and imagine that you can feel its reality around you, as if you are sitting within the White Light Sphere.

3. Breath gently and visualize the **White Light Sphere moving upward** through the center of your body until it reaches the center of the brain at the **Pineal Gland. Stop movement of White Light Sphere at the Pineal Gland,** then visualize the **White Light Sphere moving to the RIGHT** in a straight, horizontal line **OUTSIDE OF THE BODY** until it is positioned in your **auric field** just above your **RIGHT SHOULDER.** ***The White Light Sphere positioned over the Right Shoulder now represents a portion of your** awareness that was raised to the 5th-Dimensional frequencies via its passage through the Blue Flame and the 5th Kathara Center at the Navel.* You will name this part of your higher identity **MENTOR.**

4. Imagine that you can **feel the reality of the White Light Sphere** that holds your MENTOR resting gently above your Right Shoulder as a **"fuzzy ball of energy".** ***MENTOR** serves as the meeting point between your waking conscious mind and the **5th-Dimensional portion** of your Soul Matrix Identity- **the Archetype.** With practice, you **can lean the energy of your waking awareness** from its position within your head, over to the right and into the **MENTOR Sphere,** to receive direct cognition, audio and visual guidance from your 5th-Dimensional Identity Station.* **In Healing Facilitation** MENTOR will communicate with the 5th-Dimensional Archetype aspect of your **Client's Soul Matrix,** asking how you may best serve the Client's Highest Evolution; guidance MENTOR receives will be translated to you from MENTOR in the form of Intuitive urge, spontaneous cognition or audio-visual directions. MENTOR will also serve as a **translator of information from your higher dimensional Stations of Identity** to your conscious mind. MENTOR can also be taken into the Dream State with you to assist in **Dream Recall** and initiating **Out-Of-Body Travel.**

5. Practice creating and sensing the Mentor Sphere over your Right Shoulder. Carry it around with you wherever you go, and play with *moving the focus of your consciousness back and forth between the inside of your head and MENTOR.* Try perceiving from both focuses of attention simultaneously. Such play will help stimulate neurological sensitivities to translating and perceiving information from the 5th and higher dimensional Stations of Awareness.

TECHNIQUE # 1: continued

AWAKENING THE *MENTOR* AND THE 6TH AND 7TH SENSES

Awakening the 6th and 7th Senses: The *Audurea* and the *Tristet*

1. Focus your attention on the **MENTOR Sphere** over your Right Shoulder. Take a few slow, deep breaths and visualize MENTOR *expanding into a Sphere the size of a Basket Ball.* Use the next **INHALE** to draw the MENTOR Sphere *toward your head*, visualizing MENTOR moving from over your Right Shoulder to a position *surrounding your entire head.* **EXHALE,** *stop the movement* of the MENTOR Sphere, and imagine that you can feel its "fuzzy energy" all around the outside of your head; *as if your head is now within the Basket Ball size MENTOR Sphere.*
2. With the next **INHALE** breath, **contract MENTOR into a small Sphere of White Light** positioned inside the center of the brain at the **Pineal Gland.** Exhale and breathe slowly for a moment. With the next **EXHALE** breath, *push* the small MENTOR Sphere from the Pineal Gland, straight forward and into the **6th-Chakra** at the *Forehead.* Visualize MENTOR moving into the 6th-Chakra from inside the head, then continuing to move through the 6th Chakra spiral straight outward and *away from the body.* Imagine the MENTOR Sphere traveling outward through the long **Blue-Violet Tunnel** of the 6th Chakra spiral, until you visually loose track of the MENTOR Sphere far out in front of you.
3. Take a few slow breaths, then on the next **INHALE** visualize the MENTOR Sphere *re-appearing* in the distance of your 6th Chakra spiral, and *riding in on the inhale breath,* back into the **Pineal Gland.** Breathe slowly.
4. With the next **EXHALE**, visualize the MENTOR Sphere moving *down from the Pineal Gland* at the center of the brain, stopping at the center of the **5th Chakra** in the *Throat.* Breathe slowly. On the next **EXHALE** push the MENTOR Sphere straight outward, imagining it moving forward from inside the Throat, into the blue green spiral of the *5th Chakra.* Visualize MENTOR *riding the current of the exhale breath* out into the distance within the blue-green **tunnel** of the 5th Chakra spiral, until it fades from view. Take a few slow breaths and on the next **INHALE** imagine MENTOR returning through the 5th Chakra blue-green spiral to rest deep within the **Throat.** Use the **final INHALE** to move MENTOR back up to the **Pineal Gland** and the final **EXHALE** to *push* the MENTOR Sphere back out to its "home" position in the auric field over the **Right Shoulder.**

This process opens up the 1st-5th Sub-frequency Bands within each dimensional layer of the 6th and 5th Chakras, beginning the process of opening the 6th-Audurea and 7th-Tristet Senses, stimulating Inner Vision and Audio to higher dimensional frequency perception.

SYMBOL SHORTHAND

TECHNIQUE # 1: Condensed Version
Awakening the Mentor and the 6th and 7th Senses

AWAKENING THE MENTOR

1. Finger *tips* at *Navel*, see *Blue Flame* in body behind Navel. *Massage* Navel with finger tips
2. See *White Sphere in Blue Flame* at Navel. Imagine *yourself in White Sphere*, feel it around you.
3. Move *White Sphere* up to *Pineal Gland*, see White Sphere move out over *Right Shoulder* creating your *MENTOR*
4. Feel *MENTOR* as *fuzzy ball* of energy hanging in your auric field over your Right Shoulder
5. *Practice shifting* the *focus* of your attention back and forth between usual focus and *MENTOR*

AWAKENING THE 6TH & 7TH SENSES

1. Expand MENTOR over Right Shoulder to size of *Basket Ball, INHALE* drawing *MENTOR* around your *head*, as if your head is inside the Basket Ball size energy that is MENTOR.
2. *INHALE, contracting MENTOR* into *Pineal Gland* as a small Sphere of White Light, *EXHALE* to push *MENTOR* out of *Pineal Gland out to Chakra 6* and into the *Blue-Violet Tunnel* of the Chakra 6 spiral.
3. *INHALE MENTOR back from Blue-Violet Tunnel* of Chakra 6 and into the *Pineal Gland*
4. *EXHALE MENTOR* from *Pineal Gland to Chakra 5. EXHALE MENTOR out* through *Chakra 5 Green Tunnel. INHALE MENTOR back* through *Green Tunnel* into *Chakra 5* at Throat. *INHALE MENTOR* up from *Chakra 5* to *Pineal Gland. EXHALE MENTOR* out from Pineal Gland to its "Home Station" over *Right Shoulder.*

Awakening the Mentor and the 6th and 7th Senses

AWAKENING THE MENTOR

C. White Sphere over Right Shoulder "Fuzzy Ball" of MENTOR energy

Move MENTOR from Navel to Pineal Gland to position over Right Shoulder.

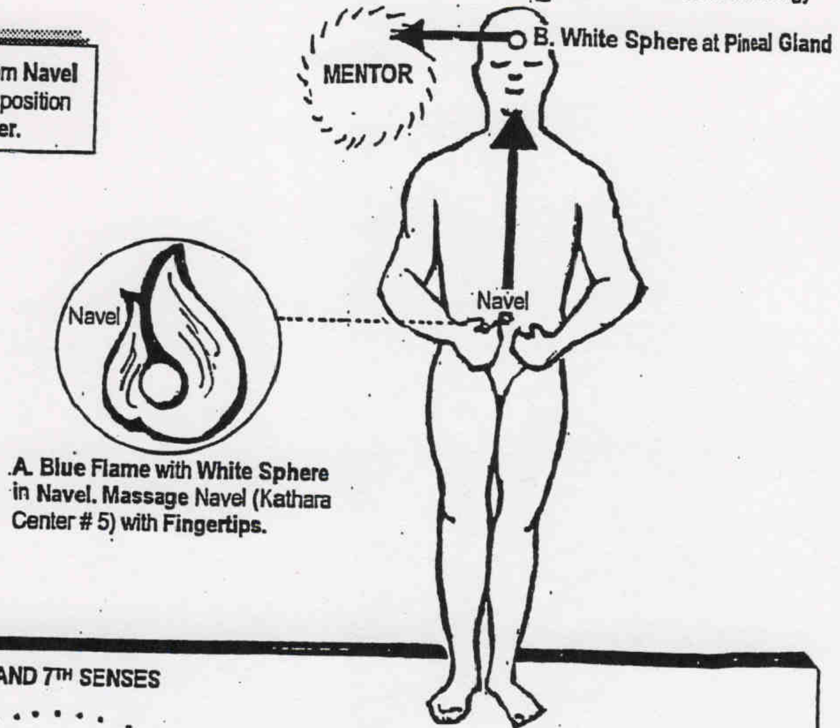

MENTOR
B. White Sphere at Pineal Gland
Navel
Navel

A. Blue Flame with White Sphere in Navel. Massage Navel (Kathara Center # 5) with Fingertips.

AWAKENING THE 6TH AND 7TH SENSES

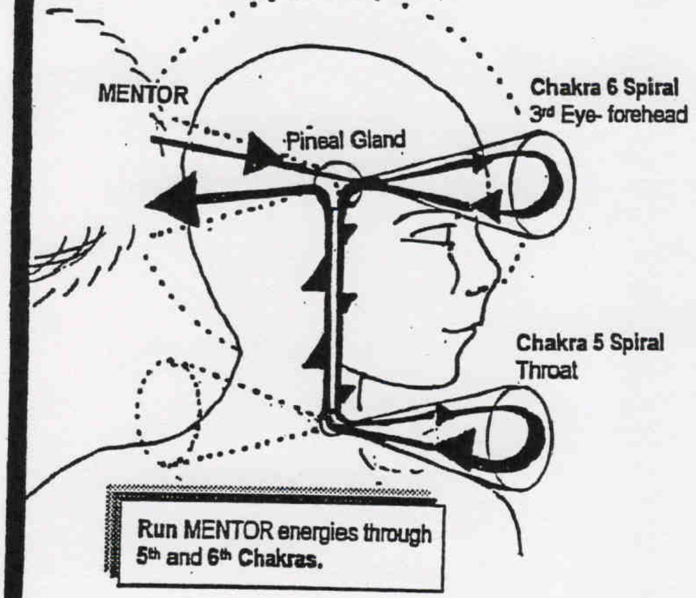

MENTOR
Pineal Gland
Chakra 6 Spiral
3rd Eye- forehead
Chakra 5 Spiral
Throat

- A. **Expand** MENTOR energy to **basketball size** and position **around head**
- B. **MENTOR in Pineal Gland as White Sphere**
- C. **White Sphere into 6th Chakra Spiral** then **5th Chakra Spiral** and back to **Pineal Gland.**
- D. Reposition MENTOR over **Right Shoulder**

Run MENTOR energies through 5th and 6th Chakras.

MAHARIC RECODING, FREQUENCY SUB-HARMONICS & THE MAHARIC SEAL

Kathara Healing and Reaching the Core of Dis-ease

In earlier text we have explored the significance of the **Kathara Grid** and its role as the **Holographic Template** upon which the morphogenetic field, body and consciousness manifest within the 15-Dimensional system. We have learned that the Kathara Grid is the **core template** upon which the *morphogenetic field* manifests, and that the *morphogenetic field* is the **blueprint** upon which the body systems and embodied consciousness manifest. Through our study of these *core structures of manifestation* we come to realize that *true and lasting health or healing* must be approached from the deepest levels of personal creation, the *morphogenetic reality* within which all dis-ease conditions have their origin. **Kathara Healing applications are intended to reach the *core of systems malfunction*, the various levels of the *morphogenetic field and Kathara Grid*, in order to restore the *Imprint for Health* within the Holographic Template for manifestation.**

The Mahara Hova Body, Maharic Shield and the Imprint for Health

In *Level-1 Kathara Healing* we will learn the first and most important level of *restructuring the Holographic Template* to hold its original Imprint for Health - the MAHARIC RECODING PROCESS. The **MAHARIC SHIELD** is the manifestation template upon which the **MAHARA HOVA BODY** of *Harmonic-4* manifests. The Mahara Hova Body is the scalar-grid that holds the **10th, 11th and 12th-Dimensional** aspects of consciousness, and represents the **ORIGINAL POINT OF MANIFESTATION** -the **creation point** when consciousness moved from its pure scalar-wave state of being into the dimensionalized scalar-grids through which the illusion of external holographic projection could be experienced. The Mahara Hova Body and its station in Harmonic Universe-4 represent the original **PRE-MATTER LIQUID LIGHT** fields of manifestation through which we passed in order to enter the experience of density and individuation of matter form.

Prior to our entering the dimensionalization of the systems within which matter can be perceived, we existed in **pure scalar-wave form** - a scalar-pattern that held the **original, pure and perfect pattern** for our creation as consciousness. The Mahara Hova Body and the Maharic Shield are the Liquid Light Pre-matter Scalar-templates that hold our original pattern of wholeness, our original **CREATION IMPRINT**. Our creation imprint is the **ORIGINAL IMPRINT FOR HEALTH**, the condition of **eternal being** free from Hova Body distortions picked up in the lower dimensional fields.

THE 12th-DIMENSION and FREQUENCY SUB-HARMONICS

In Level-1 Kathara Healing we will utilize the *Imprint for Health* as held within the *Maharic Shield* to begin REPATTERNING the Kathara Grid back to its original level of function. The frequencies of the **12th-Dimension**, as embodied within the **Maharic Shield**, contain within them as **Frequency Sub-Harmonics** ALL OF THE FREQUENCY BANDS FROM THE DIMENSIONS BELOW, along with the natural frequency spectra characteristic to the 12th-Dimension. Each Dimensional Frequency Band within the 15-Dimensional Spectra contains within itself its own **12 Primary Sub-frequency Bands** (portions of the whole dimensional Partiki Phasing flash-line sequences), plus a **reflection set** of the frequency spectra from **all of the Dimensional Bands below.**

For example, Dimension 2 contains its own *12 Primary Sub-frequency Bands*, **plus** *a reflection set of the 12- Primary Sub-frequency Bands* that make up the *Dimension-1 Frequency Band* below. Dimension-1 has 12 Primary Sub-frequency Bands, Dimension-2 thus has 12 Primary Sub-Frequency Bands plus 12 Secondary Sub-Frequency Bands from Dimension-1for a total of *24 Sub-Harmonics* in Dimension 2. **This pattern of FREQUENCY SUB-HARMONICS builds upward through the 15-Dimensional Scale. The 12th-Dimensional Frequency Band has a total of 144 Sub-Harmonics** - 12 Frequency Bands x 12 Sub-Frequency Bands each = *144 Sub-Harmonic Frequency Bands*. Because the 12th-Dimension holds within itself *all of the Frequency Spectra and scalar-grid programs* from the lower dimensions, the *12th-Dimensional Frequencies have the ability to "plug into" and override any Partiki Phasing flash-line pattern from the lower dimensional fields.* **The ORIGINAL PATTERN OF CREATION for the lower dimensional structure itself is held within the *12th-Dimensional Frequencies*, and so the 12th-Dimension *Sub-Harmonics* can be used to *correct any distortion* with the dimensional frequencies of the Lower Hova Bodies and their Shields.**

The Maharic Shield and the 12th-Dimension

As the Maharic Shield and Mahara Hova Body are composed of 10th, 11th and 12th-Dimensional scalar grids, they contain within them the **144 Frequency Sub-Harmonics** of the entire 12-Dimensional Spectrum. The human morphogenetic field is designed to hold the 144 Sub-Harmonic Spectrum, which is fulfilled through activation of the original **12-Strand DNA Silicate Matrix** gene code. Because of its 144 Sub-Harmonic scalar-wave spectrum, the 12th-Dimensional level of the **Maharic Shield** can **reset** the *original, non-distorted imprint of Partiki Phasing flash-line sequences* within the entire *Kathara Grid 12-Tree, the Kathara Centers, Kathara Lines, the Hova Bodies, DNA* and within any level of the morphogenetic field and manifest body built upon them. The *Maharic Shield* is the Kathara **RESET BUTTON.**

Frequency Sub-Harmonics in the 15-Dimensional Scale

Dimension (Primary SFB + Secondary SFBs)	Dimension x 12-SFB each =	**# of Frequency Sub-Harmonics**
Universal Morphogenetic Field: (cannot fully process through the human gene code for embodiment)		
Raja Hova Body-RISHIC SHIELD		
D-15 (12 + 168 =)	15 x 12 =	180 **Frequency Sub-Harmonics**
D-14 (12 + 156 =)	14 x 12 =	168 **Frequency Sub-Harmonics**
D-13 (12 + 144 =)	13 x 12 =	156 **Frequency Sub-Harmonics**
Human Morphogenetic Field: **(can process through gene code via 12-Strand DNA Silicate Matrix)**		
Mahara Hova Body- MAHARIC SHIELD		
D-12 (12 + 132 =)	**12 x 12 =**	**144 Frequency Sub-Harmonics**
D-11 (12 + 120 =)	**11 x 12 =**	**132 Frequency Sub-Harmonics**
D-10 (12 + 108 =)	**10 x 12 =**	**120 Frequency Sub-Harmonics**
Betcha Hova Body- TEURIC SHIELD		
D-9 (12 + 96 =)	**9 x 12 =**	**108 Frequency Sub-Harmonics**
D-8 (12 + 84 =)	**8 x 12 =**	**96 Frequency Sub-Harmonics**
D-7 (12 + 72 =)	**7 x 12 =**	**84 Frequency Sub-Harmonics**
Alphi Hova Body- DORADIC SHIELD		
D-6 (12 + 60 =)	**6 x 12 =**	**72 Frequency Sub-Harmonics**
D-5 (12 + 48 =)	**5 x 12 =**	**60 Frequency Sub-Harmonics**
D-4 (12 + 36 =)	**4 x 12 =**	**48 Frequency Sub-Harmonics**
Nada Hova Body-TELLURIC SHIELD		
D-3 (12 + 24 =)	**3 x 12 =**	**36 Frequency Sub-Harmonics**
D-2 (12 + 12 =)	**2 x 12 =**	**24 Frequency Sub-Harmonics**
D-1 (12 + 0 =)	**1 x 12 =**	**12 Frequency Sub-Harmonics**

REMEMBER: A FREQUENCY BAND is a *fixed, repeating sequence* of *Partiki Phasing Flash-lines* - **"flashing on and off" cycles of scalar standing-wave points** within the scalar-grid template, caused by the *fission/ expansion and fusion/ contraction of Partiki Units*. A **SUB-FREQUENCY BAND** is a *fixed, repeating sequence of Flash-lines* that is *part of* the full cycle of the *Frequency Band* Partiki Phasing sequence. A **FREQUENCY SUB-HARMONIC represents a replica of Frequency Band Flash-line sequences from the dimensions below, carried within the Primary Flash-line sequences of the Dimensions above.**

HIGHER DIMENSION = MORE FREQUENCY SUB-HARMONICS = FASTER FLASH RATES = HIGHER FREQUENCY = FASTER OSCILLATION/ SLOWER VIBRATION = SHORTER WAVE LENGTH = SLOWER ELECTRON SPIN = LESS PERCEIVABLE MATTER DENSITY

The Maharic Shield - Core of Healing Therapeutics

To restore the **Imprint for Health** within the *Body-Mind-Spirit System*, the **Holographic Template of scalar-grids** upon which the system manifests must be **restored to its original function.** *Topical applications* of healing modalities within the physical body, bio-chemistry, DNA, Auric Field, Chakra System, and Level-2: Crystal Seals and Level-3: Diodic Points of the Kathara Grid can create *temporary change* toward healing, but the **disharmonic conditions** held within the core **Level-1 12-Tree Kathara Grid** will eventually **override** alteration made to the outer levels of the morphogenetic field or body systems. **Working with outer levels of the Bio-system can *assist* in the healing process, but first the *Imprint for Health* must be restored within the *Level-1 12-Tree Kathara Grid*.** The *Level-1 12-Tree Grid*, with its *Kathara Centers* and *Kathara Lines* is created upon the **144 Frequency Sub-Harmonic** Pattern. As the **Maharic Shield** and Mahara Hova Body are the only portions of the dimensionalized scalar-grid anatomy that hold the full **12th-Dimensional Frequency Band** with its *144 Sub-Harmonics*, the **Maharic Shield** and its corresponding **10-11-12 Kathara Centers** are the *only portions of the morphogenetic anatomy that can override and restructure the electro-tonal scalar standing-wave point arrangements within Level-1 Kathara Grid.* **Thus it is with the *Maharic Shield* that core healing therapeutics must begin.**

Critical Mass and Simultaneous Healing of Core/ Outer Morphogenetic Levels

Activating the **Maharic Shield** will **open, activate and realign** the *scalar-wave flash-line sequences* of dormant *Kathara Centers, Kathara Lines, Crystal Seals, Hova Bodies, Diodic Points, Chakras* and *DNA* within the outer levels of the morphogenetic field and body. Once the **Imprint for Health** has been *temporarily restored* within the *Core Holographic Template Kathara Grid* via running the *Flash-line sequences* of the **Maharic Shield** through the *Level-1 Kathara Grid*, the *outer layers of the anatomy can be further realigned* to expedite manifestation of the Kathara Grid Repatterning into the outer levels of the Bio-system. **Technique # 2: The Maharic Seal and the Liquid Light Cleanse**, begin the process of running the 12th-Dimensional Flash-line sequences, via their Frequency Sub-harmonics, through the Kathara 12-Tree Grid to **REPATTERN** *distorted flash-line sequences* within the core Holographic Template. In the beginning of using this technique, the affects of this reprogramming will last only several hours, as the outer levels of the Bio-field will have greater **"Critical Mass"** in their distortion patterns, than the Kathara Grid can immediately override. **Full healing comes when the renewed Imprint for Health reaches CRITICAL MASS within the morphogenetic field.** Critical Mass is reached via *frequent use of the Maharic Seal* and *outer level Bio-field therapeutics.*

Kathara Healing Theory and Applied Methods

In Kathara Healing the first priority is to *reset the integrity* of the *Kathara Level-1; 12-Tree Grid* by Re-patterning its flash-line sequence program to the *Maharic Shield* via the *Maharic Seal and Liquid Light Cleanse*. This process sets the electro-tonal **"ground"** within which alterations of *outer morphogenetic levels* can connect, reawakening dormant **APPARTHI** *frequency receivers* within the *Level-2: Crystal Seals*, which allows the morphogenetic field levels to *expand* enough to hold the *higher frequencies* of the *Imprint for Health*. After the **Healing Ground** is set within the *Kathara Grid*, additional applications of Kathara Healing can then be applied to various other levels of the Kathara Grid, Morphogenetic Field and Body. **There is an infinite variety of Kathara Healing Bio-energetic applications**, that range from Levels 1, 2 and 3 Kathara Grid work, reprogramming the Kathara Centers, Kathara Lines, Hova Bodies, Shields and Signets, reversing the Miasmic/Karmic Imprint and accelerating the opening of the Diodic Points, to working with the Chakra System, Meridian and Axiom Lines, DNA and physical body points while running the Triadic Currents for healing. In coming Levels of the *Kathara Healing Program* we will learn the use of the **GEOMANCIES** and **FIRE LETTERS**- key **Symbol Codes** that serve as **scalar-wave guides** to reset the electro-tonal *flash-line sequences* within the morphogenetic body, and the application of the Symbol Codes to the treatment of specific physical, emotional, mental and spiritual dis-ease conditions. We will explore working with the **Color Cap Stones** of the Dimensional Frequency Bands and how they can be used in combination with the Symbol Codes for use in **Manifesting** desired conditions and **De-manifesting** undesirables, both within the body and in the 3-dimensional life drama.

Applications of the Level-1 Kathara Healing Program

In Program Level-1 we are setting the foundations upon which all the other skill levels of Kathara Healing are built. The focus of the Level-1Program is to create Morphogenetic Repatterning via the Maharic Shield, and to begin clearing the Miasmic/ Karmic Imprint that blocks various areas of the Hova Body Imprint. We will next learn to add **FORCE AND POWER** to our *Morphogenetic Recoding Process* by opening the **PLANETARY BIO-FEED INTERFACE SYSTEM™** to draw energy directly from the **Planetary Maharic Shield**, and we will begin learning of the *physical body massage point sequences*- **Sequence 1 and 2 Kathara Points** - through which we can begin to *regulate the flow of energy within the Bio-system* for **Regeneration** of the Imprint for Health and **Revitalization** of the Body-Mind-Spirit System. In **Section 2** of this course we will explore **MULTI-VECTOR HOLOGRAPHIC RECODING** - clearing the *Miasmic Imprint* from the Kathara Grid.

SEQUENCE 1 AND 2 KATHARA POINTS

Sequence-1 Kathara Points for REGENERATION

Sequence-1 Kathara Points are mini-energy-vortex points in the body that connect directly to **Diodic Points** within the *Kathara Level-3 Diodic Points Grid.* Stimulation of the Kathara Points affects the movement of energy within related areas of the Kathara Grid, which creates a reciprocal movement of energy within the cellular structure corresponding to the Kathara Points.

When used **in proper sequence,** *Sequence-1 Kathara Points* ***open energetic gate ways to the Kathara Lines*** of the Kathara Level-1 12-Tree Grid, allowing patterns of frequency to be entered into the Kathara Lines, to realign disharmonic patterns of frequency running through the Holographic Program of the Kathara Lines. When the Kathara Lines are opened to receive **11th** and **12th** dimensional frequency infusions from the **Mahara Hova Body,** the Holographic Program of the Kathara Grid can be progressively ***regenerated*** to the Imprint for Health, as held within the program of the Maharic Shield.

Sequence-1 Kathara Points are used to *open* the Kathara Lines prior to *Technique # 2: The Maharic Seal and the Liquid Light Cleanse* and *Technique # 4: The Maharic Infusion.* They **may also be used alone** with deep breathing, for frequent healing support, to draw energy up from the Maharic Shield, re-balancing and regenerating the Imprint for Health within the Kathara Grid and Body-Mind-Spirit System. It is not necessary to close the Kathara Lines after using Sequence-1 Points as they will **close automatically** in **several hours.** If one desires to further revitalize the cellular structure, running the *Sequence-2 Kathara Points* for *closing* the Kathara Lines will increase the amount of energy running through the *body cells,* as the Kathara Lines are no longer absorbing the energy.

Sequence-1 Points **regenerate and re-balance the *Kathara Core* first, which creates *more lasting but less immediate* changes in the body. *Sequence-2 Points* close the Kathara Lines if they are opened and draw energy from the Maharic Shield *directly into the bio-energetic field and physical cells,* rather than the core Kathara Grid, creating *more immediate, but temporary,* revitalization of the body.** The patterns held within the Kathara Grid- core holographic template will eventually override those set into the bio-energetic field alone. **Used in *combination* with each other the *Sequence-1 and Sequence-2 Kathara Points* can affect more lasting revitalization by *simultaneously realigning the core template and the outer layers* of the bio-energetic system and cellular structure.** In combination with other Kathara Healing techniques the affects and restoration of the Imprint for Health are further amplified and accelerated.

Sequence-2 Kathara Points for REVITALIZATION

Sequence-2 Kathara Points are mini-energy-vortex points in the body that, like *Sequence-1 Kathara Points*, connect directly to **Diodic Points** within the *Kathara Level-3 Diodic Points Grid.* Stimulation of the Kathara Points affects the movement of energy within related areas of the Kathara Grid, which creates a reciprocal movement of energy within the Bio-Field and cellular structure corresponding to the Kathara Points.

When used **in proper sequence,** *Sequence-2 Kathara Points* ***close the energetic gate-ways to the Kathara Lines*** of the Kathara Level-1 12-Tree Grid, which the *Sequence-1 Kathara Points* had opened. Closing of the Kathara Lines governed by the Sequence-1 Kathara Points, ***creates a reciprocal OPENING of Nadial Lines (Nadis*** energy rivulets that carry energy from the Nada Hova Body/ Nadial capsule, into the physically manifest systems**)** within the *outer levels* of the Bio-energetic Field. Opening of corresponding Nadis allows restorative patterns of frequency to be entered into the **outer levels** of the Bio-energetic Field and Chakra System, to realign disharmonic patterns of frequency running through the *Chakras, Auric Field and Physical Cells*. When the Nadis, Chakras and outer Bio-energetic Field are opened to receive 11th and 12th dimensional frequency infusions from the **Mahara Hova Body,** the energetic integrity of the physical cells and Bio-energetic Field can be progressively ***revitalized*** with Maharic frequency and the Imprint for Health, as held within the program of the Maharic Shield.

Sequence-2 Kathara Points are used to *close* the Kathara Lines following *Technique # 4: The Maharic Infusion*. They **may also be used alone** with deep breathing, for frequent healing support, to draw energy up from the Maharic Shield, through the Kathara Grid and into the **outer levels** of the Bio-energetic Field. Directing Maharic frequency into the outer levels of the Bio-energetic Field regenerates the Imprint for Health within, and directly **revitalizes** the *body cells* for accelerated Body-Mind-Spirit healing. If one desires to revitalize the cellular structure, running the *Sequence-2 Kathara Points* for closing the Kathara Lines will **increase the amount of energy running through the *body cells*,** as the Kathara Lines are no longer absorbing the energy. **Sequence-2 Points revitalize and rebalance the *outer levels of the Bio-energetic Field* first, which creates more immediate but temporary revitalization of the body, as restorative energies do not directly enter the Level-1 Kathara Grid.** To create *lasting change* when using the *Sequence-2 Points* alone, restorative electro-tonal programs must be **repeatedly** entered into the Bio-energetic field until the new programs reach **critical mass,** at which time the new programs will begin to *subtly influence* the programs held within the Kathara Grid. **The primary difference between Sequence-1 and Sequence-2 Kathara Points is this: *Sequence-1 Points*** channel energy from the Maharic Shield into the ***core Kathara Grid*** holographic template. ***Sequence-2 Points*** channel energy from the Maharic Shield into the ***outer levels** of the Bio-field and body cells*. **For this reason the affects of *Sequence-2 Kathara Points* are more *immediate*, and those of *Sequence-1 Kathara Points* are more *permanent*.**

The patterns held within the Kathara Grid core holographic template eventually override those set into the outer Bio-field alone. **Used in *combination* with each other and other healing applications the *Sequence-1 and Sequence-2 Kathara Points* can affect more lasting revitalization by simultaneously *realigning the core template and the outer layers* of the Bio-field and cellular structure.** Using ***Kathara Point Sequences*** in combination with each other and other Kathara Healing techniques will amplify and accelerate the affects and restorative function of the Imprint for Health within the Body-Mind-Spirit system.

14 SEQUENCE 1 KATHARA POINTS

Regeneration of the Imprint for Health

Sequence-1 Kathara Points **OPEN** energetic gate-ways within the **Kathara Lines** of the *Kathara Level-1 12-Tree Grid*, to receive electro-tonal reprogramming for **regeneration** of the *Imprint for Health*. *Sequence-1 Points* are used to open the Kathara Lines **prior to** other energy-healing applications, to open the Kathara Grid so the regenerative patterns of other healing applications can "ground" within the Kathara. Sequence-1 Points re-generate and rebalance the **Kathara Core** first, creating **more lasting but less immediate changes** in the body.

1............CROWN
2 & 3.......TEMPLES
4.............BASE OF SKULL
5 & 6........BELOW SHOULDERS
7 & 8........ABDOMEN IN FROM HIPS
9 & 10......BACKS OF KNEES
11 & 12....SOLES AND TOPS OF FEET
13............4TH HEART CHAKRA
14............6TH THIRD EYE CHAKRA

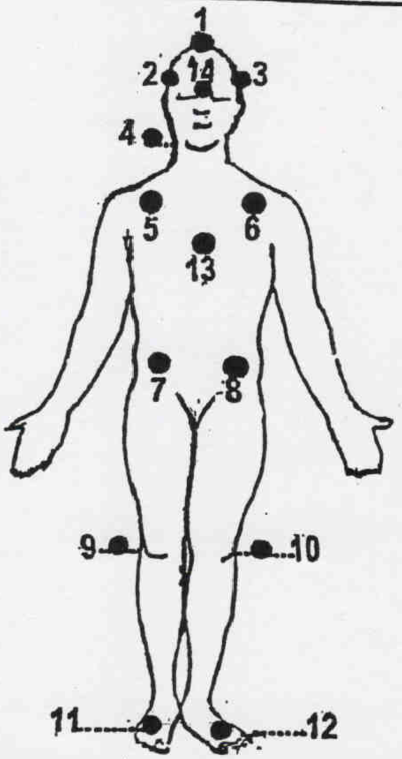

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

14 Sequence-1 Kathara Points for Regeneration

LOCATION & APPLICATION SEQUENCE

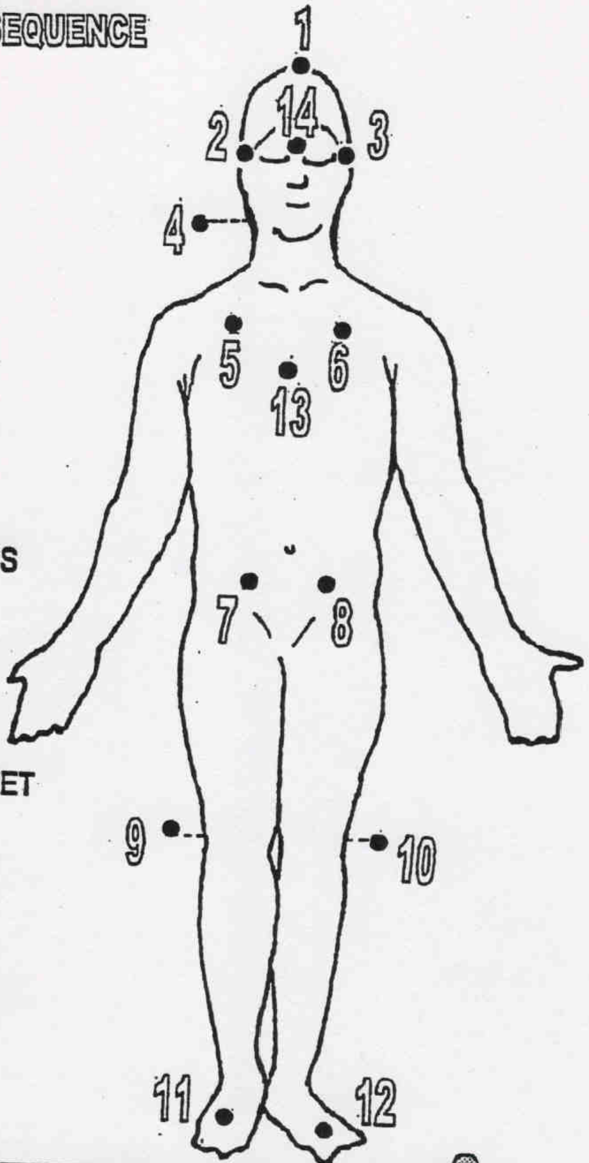

1............CROWN

2 & 3.......TEMPLES

4.............BASE OF SKULL

5 & 6.........BELOW SHOULDERS

7 & 8.........ABDOMEN IN FROM HIPS

9 & 10......BACKS OF KNEES

11 & 12....SOLES AND TOPS OF FEET

13...........4th HEART CHAKRA

14............6th THIRD EYE CHAKRA

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

6 SEQUENCE 2 KATHARA POINTS

Revitalization of the Imprint for Health

Sequence-2 Kathara Points **CLOSE** energetic gate-ways within the **Kathara Lines** of the *Kathara Level-1 12-Tree Grid*, to direct restorative frequencies into the **outer levels of the Bio-energetic Field** for **revitalization** of the *Imprint for Health* within the physical cells. *Sequence-2 Points* are used to close the Kathara Lines **following** use of the *Sequence-1 Kathara Points* or other energy-healing applications, to close the Kathara Grid so revitalizing frequencies can directly enter various *outer levels* of the Bio-energetic Field, such as the **Auric Levels, Chakras, Meridians, DNA** and **physical cells**. Sequence-2 Points revitalize and rebalance the **outer levels of the Bio-energetic Field** first, creating **more immediate but temporary changes** in the body.

1 & 2........NECK, BELOW EARS

3 & 4........BELOW CHEEKBONES

5..............BASE OF TAILBONE

6..............THYMUS

*** AMPLIFER –3RD SOLAR PLEXUS CHAKRA**

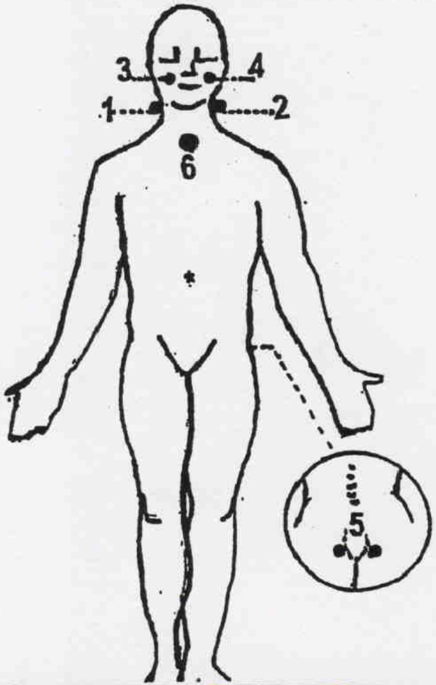

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

6 Sequence-2 Kathara Points for Revitalization

LOCATION & APPLICATION SEQUENCE

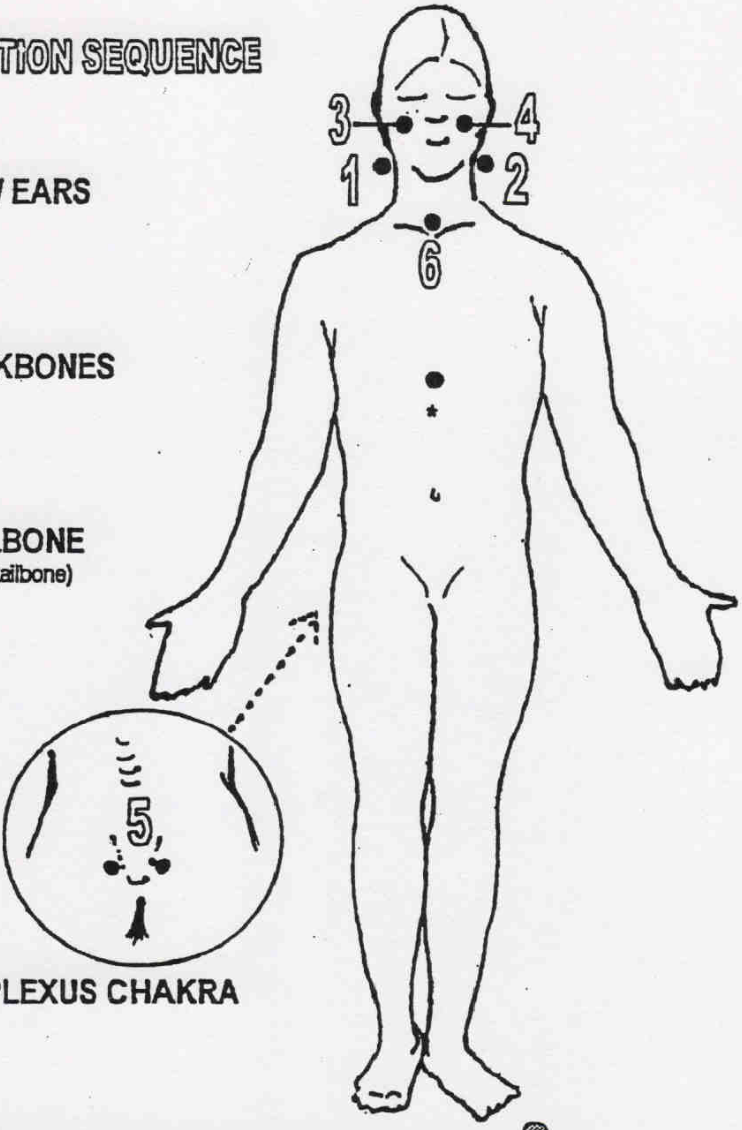

1 & 2........NECK, BELOW EARS

3 & 4........BELOW CHEEKBONES

5..............BASE OF TAILBONE
(one point on each side at base of tailbone)

6..............THYMUS

*** AMPLIFER –3RD SOLAR PLEXUS CHAKRA**

Instructions: Breathe slowly and deeply. Imagine a flow of energy running from the **Maharic Shield** (12" below feet), up into the body through the **soles of the feet** and the **1st Chakra**. Gently **rub** the *Kathara Points* with fingertips for 10 – 20 seconds each, following the **precise sequence** described above.

TECHNIQUE # 2

THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

1. Close your eyes and visualize the image of a flat (2-dimensional line drawing) 6-Pointed Star of David (**The Merkaba Star**) made of **White Light**, stationed at the **6th-Chakra** on the *inside of the Forehead.* (If you cannot yet generate inner visualizations, simply imagine that the symbol is there.) Continue to visualize or imagine, and move the **White Merkaba Star** image over toward the **left** in your inner field of vision. In the *center* of your inner field of vision, visualize/ imagine the image of *another Merkaba Star*, made of a **Dark Silver Light**. Observe or imagine the **2 Merkaba Stars** in your inner field of vision and then slowly *move the images toward each other* until they have merged into *one Merkaba Star* Symbol made **of Pale Silver Light**. (The *Pale Silver Double Merkaba Star Symbol* is called the **HIEROPHANT** – it is the scalar wave template that governs the release of phase-lock on the *Maharic Shield.)*

2. Visualize/ imagine the Hierophant Symbol clearly in mind, then shift attention to the sound of your breath. Take several slow, deep breaths. On the next **EXHALE** use the force of breath to *forcefully push* the **Hierophant** down from its position at the 6th-Chakra, through the *Central Body Current* and into the **Earth's Core.** Take a few more slow breaths while visualizing/ imagining the Hierophant Symbol in the center of the Earth's core. Imagine now that the *Hierophant begins to spin in the Earth's core*, the Dark Silver Merkaba Star and the White Merkaba Star spinning in *opposite directions.* Spin the Hierophant progressively faster until it "takes on a .life of its own", and suddenly "pops and expands" to form a **Pale Silver Sphere of Light** *spinning in the Earth's core.*

3. Observe the Pale Silver Hierophant Sphere spinning in Earth's core, then visualize/ imagine that a **huge, horizontal Pale Silver spinning disc** emerges out from the Hierophant Sphere, forming a **massive Pale Silver Disc of Light** *spinning throughout the Earth and extending outward into the atmosphere* as a **rotating, horizontal plane.**

This horizontal disc-plane extending through the Earth represents the *Mahara Hova Body*, or **Maharic Shield** of the planetary morphogenetic field – the Pre-matter Liquid Light scalar grid composed of frequency from dimensions 10, 11 and 12. (Known as the ***Shield of Amaratena***). Using the Hierophant wave template you have now activated the planetary Maharic Shield, ***opening the Planetary Bio-feed Interface System™*** *into your bio-energetic field.* This constitutes opening your personal Kathara Grid and morphogenetic field into the Planetary Kathara Grid and morphogenetic field, allowing you to draw an *unlimited supply of energy* from the planetary field. You can draw in only as much as your personal morphogenetic field can hold. With repetition of this practice the energy-holding capacity of your Kathara Grid progressively increases, allowing you *a continually expanding supply of energy* for healing and revitalization.

TECHNIQUE # 2: Continued
THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

4. While remaining aware of the *Planetary Maharic Shield* spinning on a horizontal plane throughout the Earth's body, return the focus of your attention to the spinning **Silver Hierophant Sphere** rotating at Earth's core. (You will now use the Hierophant Sphere to activate your **personal Maharic Shield** and release the 12th-Dimensional Pillar of Light within your bio-energetic field.). Take several slow, deep breaths, while visualizing/imagining the Hierophant Sphere spinning in Earth's core. On the next **INHALE**, draw the Hierophant Sphere *up from Earth's core* with the inhale breath, forming a **Pale Silver Cord** *about 6" thick* trailing behind the Hierophant as it rises to the **12th Chakra**, 6" beneath the feet. *One end of the Silver Cord remains rooted in* **Earth's core**, *the other end is attached to the* ***Hierophant Sphere*** *and follows its movement.* Visualize the **12th Chakra** as a small, **transparent disc** about *3" in diameter*, positioned *6" below the feet*. Stop the movement of the Hierophant in the *center of the 12th Chakra*, imagine the Sphere beginning to spin and *visualize the diameter of the 12th Chakra disc expanding to about 6" in diameter*, as the Hierophant spins at its center. Take a few slow, deep breaths as the *Hierophant spins in the center of the* ***12th Chakra.***

5. On the next **INHALE** draw the Heirophant Sphere *up from the 12th Chakra*, through the **Central Body Current** and up to the **14th Chakra** *36" above the head*. As the Hierophant Sphere rises, visualize the *6" thick* **Silver Cord** drawing up through the **12th Chakra** and farther up through the **Central Body Current,** following the path of the Hierophant Sphere to the **14th Chakra**. Visualize the Silver Cord as a long, thick **Pale Silver Cord of Light**, running from the *Earth's core, through your body and out the top of your head*, connecting to the **14th Chakra** 36" above the head. Take several deep, slow breaths while imagining that you can *feel the sensation* of this *massive current of clear, pure energy* running through the *center* of your body.

6. On the next **EXHALE**, push the spinning Hierophant Sphere *down* through the **Central Body Current,** while leaving the *top end* of the Pale Silver Cord attached to the **14th Chakra** 36" above the head. Move the Hierophant down through the **12th Chakra** 6" below the feet, and **stop the Hierophant** *6" below the 12th Chakra;* a position about **12" below the feet.** This is the location of the **Maharic Shield** within the morphogenetic field of the body. Breath slowly and visualize the Hierophant *spinning progressively faster* 12" beneath the feet, until the Hierophant "takes on a life of its own" and suddenly "*pops and expands*" to **form an expanding horizontal disc of Pale Silver Light spinning 12" beneath the feet.** ***Feel the disc as a circular platform 3 yards in diameter below feet.***

TECHNIQUE # 2: Continued
THE *MAHARIC SEAL* AND THE *LIQUID LIGHT CLEANSE*

7. Call to mind the image of the *massive horizontal disc of Pale Silver Light* spinning through the Earth's body – the **Planetary Maharic Shield**. Imagine the sensation of the large, spinning *Pale Silver platform disc* spinning 12" beneath your feet – your **personal Maharic Shield**. Breathe slowly and deeply while visualizing/imagining a *massive rush of pure, cool* ***Pale Silver energy*** running from the *planetary Maharic Shield*, up through the **Pale Silver Cord** and into the *personal Maharic Shield* 12" beneath your feet. Breath as this energy runs, *filling your Maharic Shield to its energy-holding capacity*. When your Maharic Shield has received as much Pale Silver energy as it can hold, it will "take on a life of its own", and suddenly "*pop and expand*" upward into an ***oscillating Pale Silver Pillar of Light*** *surrounding and permeating the body*, a *tube of light* extending beyond the body *about 18"*, and running from the **Earth's Core** to the **14th Chakra** 36" above the head. *Feel the reality of your body encased in the cool, refreshing Pale Silver Pillar of oscillating light*. Take a few slow deep breaths and imagine the energy from the Pillar *emanating through all of your body cells*. Your Bio-energetic Field is now sealed with a protective barrier of 11th/12th-Dimensional frequency – the **Maharic Seal**.

8. As the Pale Silver Pillar of Light oscillates around you, place your attention for a moment at the **13th Chakra** in Earth's core and **INHALE** a final **Spark** of *Pale Silver Light* from Earth's core, drawing the Spark up through the **Pale Silver Cord**, into the **Main Vertical Channel of the Kathara Grid**, and into the center of the **4th Heart Chakra**. Breath slowly and easily while visualizing energy from the *Silver Spark* radiating throughout the embodied **12-Tree Kathara Grid**, *filling it with Pale Silver Light*. When the Kathara Grid has reached its energy-holding capacity, the ***12th Chakra*** *will automatically contract* to decrease the flow of Silver Light from the Earth's core. As the 12th Chakra contracts, the *Planetary Maharic Shield* stops spinning and transmitting energy through the Silver Cord, returning the Silver Chord to its dormant state. The ***Maharic Seal*** **Pale Silver Pillar of Light** will remain in your bio-energetic field between **1-3 hours**. Over a *3-month period* of frequent, consistent *daily practice* of this process, the Maharic Seal will self-sustain in the bio-energetic field for **12-14 hours**.

Use of this technique allows a healing Facilitator to activate the **Maharic Shield and Seal** in others, through the ***Maharic Infusion process*** (Technique # 4), while Facilitator's Pillar remains active. Progressive use of the Maharic Seal and running the Liquid Light creates a reciprocal clearing, reordering and realignment of the Kathara Grid-***Holographic Template*** – and all subtle energy and physical systems built upon it. The Lower Hova Bodies are progressively cleared and realigned with the ***original perfection of the Pre-matter Liquid Light Mahara Hova Body***, and the body and genetic code progressively expand to hold higher dimensions of conscious awareness.

TECHNIQUE # 2B: THE MAHARIC QUICK SEAL

1. Work with a 24 pointed, 3 Dimensional Star (Heirophant) and be aware that you will be working with beings known as the Breneau Rishi ...

2 Begin by slowing your breathing , as you imagine or visualize a 24 point, 3 -Dimensional, Heirophant at the Pineal Gland in the center of your brain. Visualise the Heirophant surrounded by a Pale Blue Sphere of Light.

(The Blue Sphere serves as a buffer for your readiness to accept the frequencies associated with using the 24 pointed Star; the Blue Sphere holds the energy for you until your body can handle the energy. If you can handle it, the energy will simply pass through the blue sphere).

3 INHALE, as if you are going to grab the blue sphere containing the 24 point star located at the Pineal, and on the EXHALE move the 24 point star/ blue sphere, all the way down to Earth's Core. Try to hear a sound tone as the Heirophant/ Sphere hits the Planetary Shield ...

4 INHALE and draw the Heirophant/ sphere up to your Personal Maharic Shield 12" below your feet ... and EXHALE while watching the Pale Silver-Blue sphere expand out to a large sphere ... and watch as your Maharic Shield pop out as a disc, about 4' in diameter, Pale Silver with a light coating of Pale Blue.

5 Bring your attention to the center of your Maharic Shield 12" below your feet ... and INHALE the 24 point Star (only) up into the Heart Chakra. EXHALE, expanding the frequencies of the 24 point Star into the Heart Chakra.

6 INHALE to grab the 24 point Star, and EXHALE, pushing the Star up into the 14th Chakra 36" above the head. Imagine that the 24 point Star is spinning in a clockwise direction in the 14th Chakra. As you do this, try to feel/ sense the energy around you, just a few inches out from your body ...

7 Now, put your attention into your Maharic Shield 12" below your feet, take a couple of relaxing breaths and INHALE deeply pulling the Pale Silver and Blue energy up, as if you are pulling on your Maharic Seal to latch it on to your 14th Chakra. Try and feel the sensations just a few inches from your body now – feel the difference if you can. This procedure gives you a Quick Seal.

8 Now envision the cord of Pale Silver Maharic Light that would normally come up with your Shield and focus on the Earth's Core. Begin drawing energy all the way up the 4" diameter cord and into the body. INHALE the energy up into the Heart Center, and expand it there. Begin to form a pale Silver ball of Maharic frequency.

On each EXHALE send your energy down to the Earth's Core to bring up another load of Maharic frequency, up from the Planetary Shields, expand it into the Heart center and repeat several times. Notice that the Cord grows larger as you do this, expanding from approximately 4" to 6 to 8" diameter, until finally it feels like a skin around (as well as within) your body ... as you load your Astral field with Maharic Frequency.

9 Move your attention to the Pale Silver ball you have created in your Heart Center, INHALE and move it up to the 14th Chakra 36" above your head .. as you do, feel for the sheath of Maharic Frequency encasing your etheric body, close to your skin. (This Maharic/ Christos skin charges, and saturates, your etheric body – feel the peacefulness and all-knowing nature of this frequency).

SYMBOL SHORTHAND

TECHNIQUE # 2: Condensed Version
The Maharic Seal and Liquid Light Cleanse

Symbol Shorthand is used to help in remembering the primary steps of a technique

1. *Silver Hierophant Star at Chakra 6*
2. *EXHALE* Hierophant Star from *Chakra 6 to Earth's Core –Spin Hierophant Star* at Earth's core-until it *pops* into *Hierophant Silver Sphere at Earth's core.*
3. *Hierophant Sphere* spins *at Earth's core* until *Silver Disc of Earth's Shield* emerges.
4. *INHALE- Hierophant Sphere* up from *Earth's core to Chakra 12* trailing *Silver Cord* from Core.
5. *INHALE- Hierophant Sphere* up from *Chakra 12 to Chakra 14*, trailing *Silver Cord.*
6. *EXHALE- Hierophant Sphere* down from *Chakra 14 to Personal Maharic Shield* and *spin Hierophant Sphere* until *Personal Shield pops* out to become *Silver disc* plane 12"beneath feet.
7. Draw *Silver energy* into *Personal Shield from Earth's Shield* through *Silver Cord,* until *Personal Shield* is *filled* and *pops* to become vertical *Pillar* of Silver Light around body.
8. *INHALE*-draw *Spark of Silver Light* from *13th Chakra* at *Earth's Core,* into *main* vertical channel of *Kathara Grid* then to *Chakra 4;* from Chakra 4 transmit energy from Spark to fill *12-Tree* Kathara Grid. Chakra 12 auto-shut down when 12-Tree Grid is full.

THE MAHARIC SEAL

63

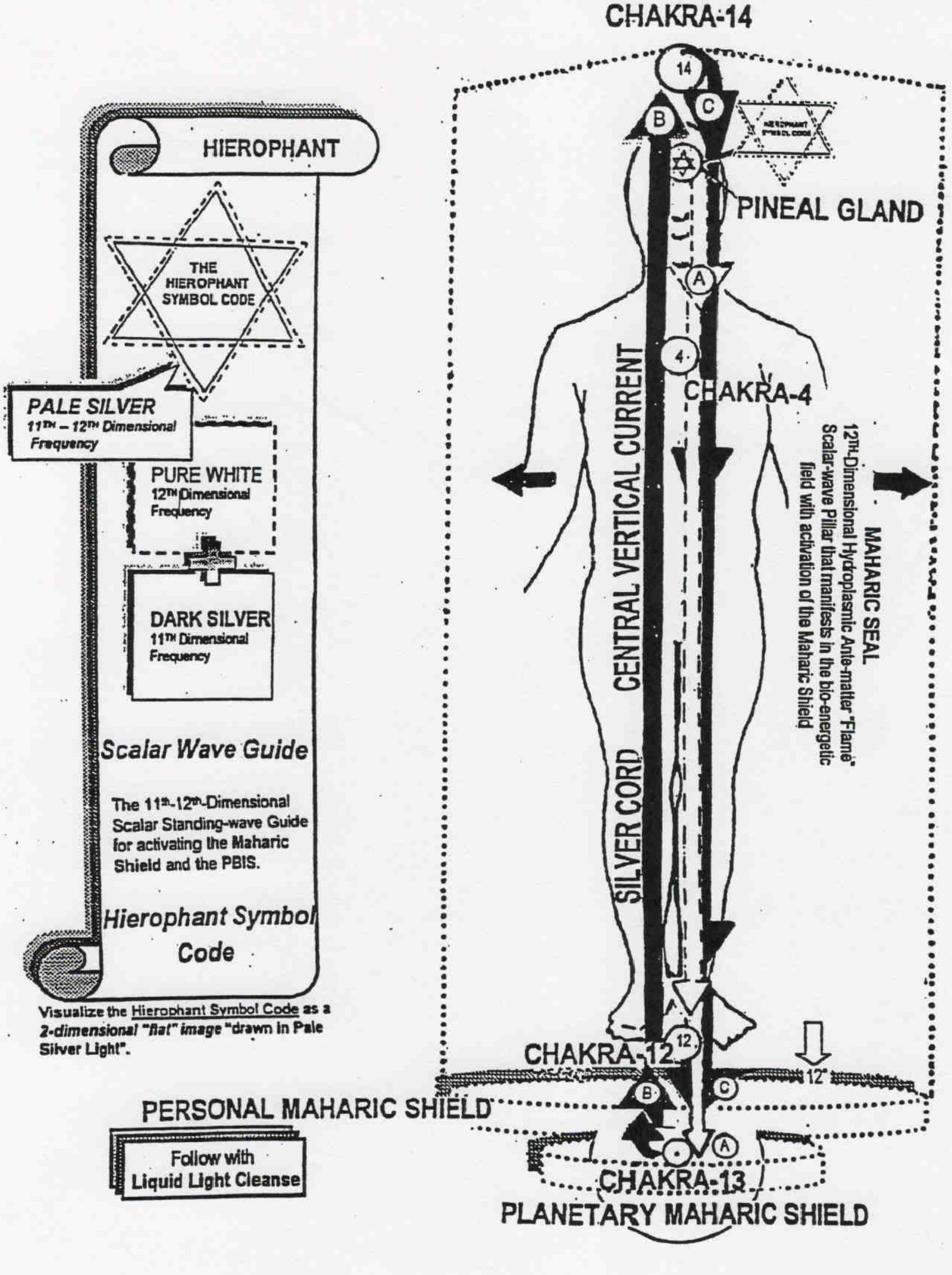
HIEROPHANT
THE HIEROPHANT SYMBOL CODE
PALE SILVER
11TH – 12TH Dimensional Frequency
PURE WHITE
12TH Dimensional Frequency
DARK SILVER
11TH Dimensional Frequency
Scalar Wave Guide
The 11th-12th-Dimensional Scalar Standing-wave Guide for activating the Maharic Shield and the PBIS.
Hierophant Symbol Code
CHAKRA-14
14
B
C
A
HIEROPHANT SYMBOL CODE
PINEAL GLAND
A
4
CHAKRA-4
CENTRAL VERTICAL CURRENT
SILVER CORD
MAHARIC SEAL
12TH-Dimensional Hydroplasmic Anti-matter "Flame" Scalar-wave Pillar that manifests in the bio-energetic field with activation of the Maharic Shield
CHAKRA-12
12
12"
B
C
A
PERSONAL MAHARIC SHIELD
Follow with Liquid Light Cleanse
CHAKRA-13
PLANETARY MAHARIC SHIELD

Visualize the Hierophant Symbol Code as a *2-dimensional "flat" image* "drawn in Pale Silver Light".

LIQUID LIGHT CLEANSE

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 –*The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7-*The Wings*
In shoulder indents
through to back.

# 5 –*Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 -*The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

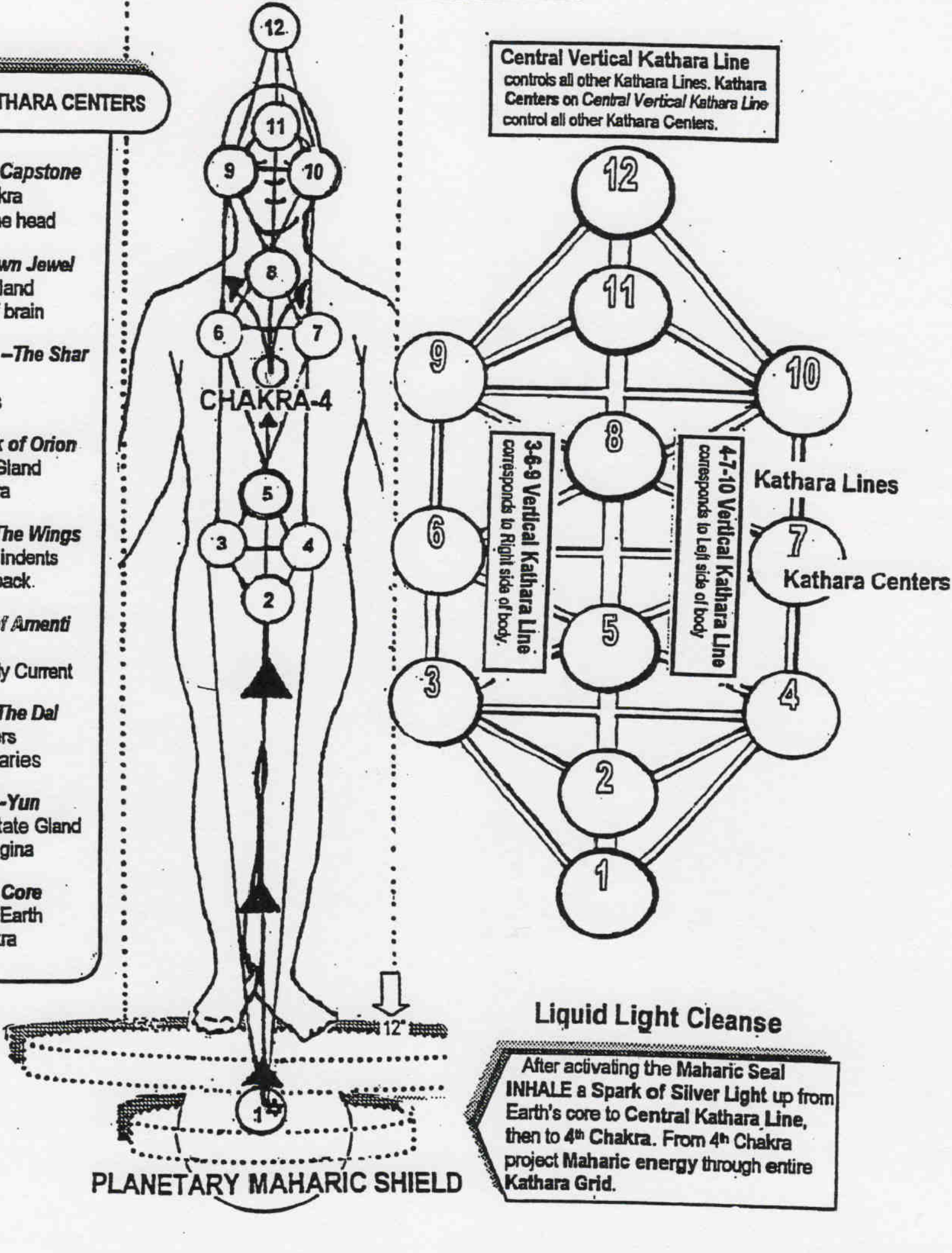
12
11
9
10
8
6
7
CHAKRA-4
5
3
4
2
12"
1
PLANETARY MAHARIC SHIELD
12
11
9
10
8
3-6-9 Vertical Kathara Line
corresponds to Right side of body.
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
6
7
Kathara Centers
5
3
4
2
1

Central Vertical Kathara Line
controls all other Kathara Lines. Kathara Centers on *Central Vertical Kathara Line* control all other Kathara Centers.

Liquid Light Cleanse

After activating the Maharic Seal **INHALE a Spark of Silver Light** up from Earth's core to **Central Kathara Line**, then to 4th Chakra. From 4th Chakra project **Maharic energy** through entire **Kathara Grid**.

The Planetary Bio-Feed Interface System and the Maharic Seal

Use of the *Maharic Seal and Liquid Light Cleanse* begins the process of activating the **8th through 12th Chakras** and opening the natural **PLANETARY BIO-FEED INTERFACE SYSTEM (PBIS)** within the human body. Through the **PBIS**, the finite quantity of scalar frequency of which the personal Maharic Shield, Hova Bodies and individual identity are composed expands to **infinite potential**, as the **Planetary Scalar Shields** can be used to open dormant **Crystal Seals** within the Scalar Shields of the personal Level-2 Kathara Grid. **Once activated through the 12th Chakra, the PBIS opens a conduit of perpetual energetic supply between the human body and the planetary Morphogenetic Field, allowing the body to continually draw in as much higher dimensional frequency as the DNA Template can hold.**

With the PBIS activated, the human body can also become a tool for reprogramming and re- balancing the Planetary Shields, Vortices, Ley Lines and energetic grids of the Planetary Templar Complex. When the Crystal Seals of the personal Shields are released, the **human DNA Template** can serve as a **bio-electric conduit** for running higher dimensional frequency into the **Planetary Shields.** The human body, with its **12-Strand DNA Template potential,** is the only mechanism on Earth, other than the Planetary Shields themselves, that is capable of running a full 12th-Dimensional **Maharic Current.** Just as within the human Manifestation Template, the Maharic Current can be used to purge distortions within the Planetary Shields, to restore the **Planetary Templar Complex** to its original **Primal Order.** Kathara Healing technologies not only provide the opportunity for individuals to facilitate healing in other *biological life forms,* they also reawaken the human potential to facilitate healing within the **geophysical planetary body.**

The Planetary Bio-Feed Interface System

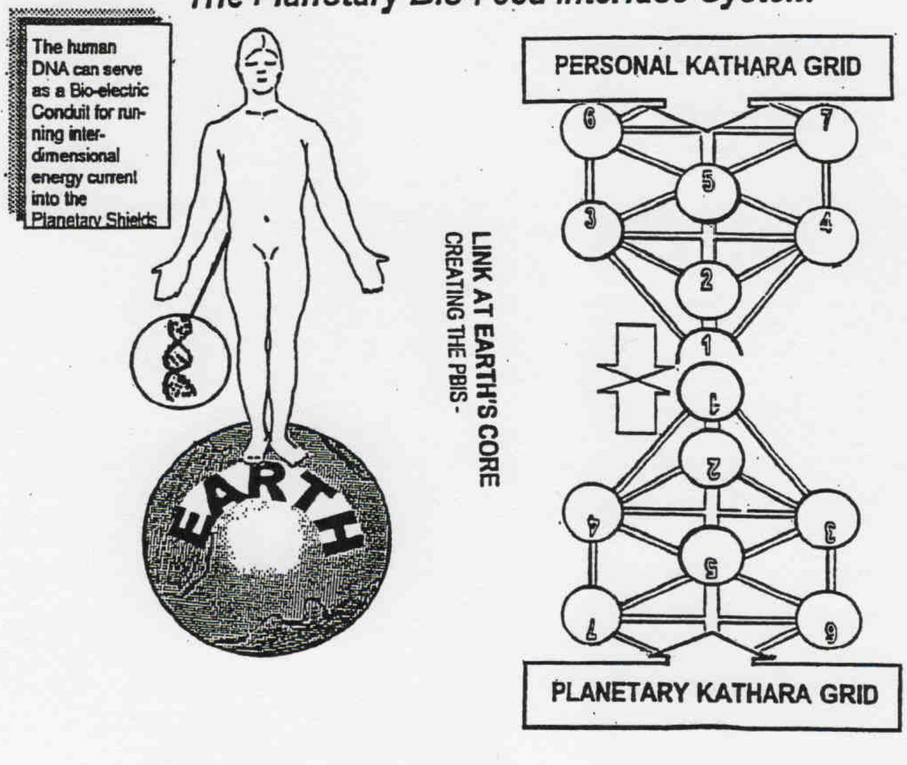
The human DNA can serve as a Bio-electric Conduit for running inter-dimensional energy current into the Planetary Shields
EARTH
LINK AT EARTH'S CORE
CREATING THE PBIS -
PERSONAL KATHARA GRID
6
7
5
3
4
2
1
1
2
4
3
5
7
6
PLANETARY KATHARA GRID

The Maharic Seal and Client Sessions

Through understanding the **Context of Healing**- *The Structure of the Unified Field*, the **Elements of Healing** – *The Anatomy of the Human Body Kathara Grid* and **Kathara Healing Beginning Applications** - *Morphogenetic Repatterning* through the ***Maharic Seal and Liquid Light Cleanse*** and *Awakening the Mentor* for access to 5th Dimensional awareness, the Kathara Healing Facilitator has developed the rudimentary skills necessary to begin **personal Core Holographic Template Healing. To effectively facilitate Core Template healing in others it is important for Kathara facilitators themselves to employ personal Template Healing.**

The ability to hold the **12th-Dimensional Maharic frequency sub-harmonics** (the 12th-sub-frequency band within each of the full Dimensional Frequency Bands) in body to assist in its transmission to a client for healing support increases as the facilitator utilizes the **Maharic Seal and Liquid Light Cleanse** technique **for personal healing.** The *Maharic Seal and Liquid Light Cleanse* technique not only increases the facilitator's ability to transmit 12th-Dimensional frequency sub-harmonics, it also temporarily **realigns distortions** within the *facilitator's* bio-energetic field and Kathara Grid and creates a temporary **12th-Dimensional Scalar Frequency Seal** within the *facilitator's* bio-energetic field. **Without use of the *Maharic Seal*, the facilitator will inadvertently transmit bio-energetic field and Kathara Grid distortions to the client and will energetically pick up scalar distortions** (including part of the "Karmic" Miasmic Imprint) **from the client.** (This problem occurs in all subtle-energy healing modalities that do not employ the Maharic Seal). Both of these conditions serve to amplify, rather than remedy, dis-ease and disharmony within the Mind-Body-Spirit system, for both *facilitator and client*. For these reasons, **the *Maharic Seal and Liquid Light Cleanse* technique needs to be employed prior to ANY healing facilitation work,** to protect the **bio-energetic field integrity** of facilitator and client. **The *Maharic Seal and Liquid Light Cleanse* technique represents the facilitator's preliminary preparation prior to Kathara Healing Facilitation Client Sessions.**

Safety in Healing Facilitation!

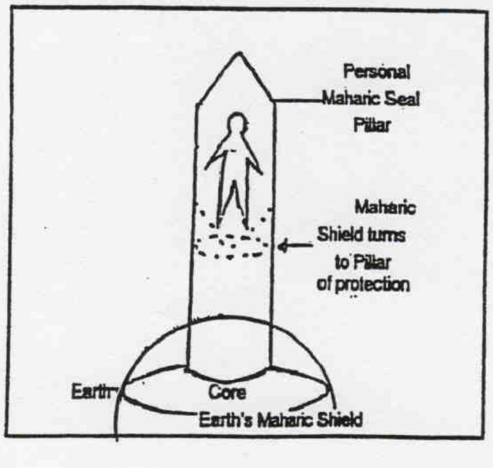

The ***Maharic Seal and Liquid Light Cleanse*** technique, used as facilitator preparation prior to **Client Sessions,** protects facilitator and client from taking on bio-energetic field distortions from each other. **Maharic Seal** is the preliminary step before any **Kathara Healing** application and is needed to effectively run interdimensional healing currents.

CO-GENERATIVE HEALING

The final phase of education necessary before engaging the adventure of **Kathara Healing Client Facilitation** is rediscovering the structure of personal **Multi-dimensional Identity** and how to use the personal **Hova Bodies** to tap the higher dimensional scalar frequencies for enhanced healing facilitation ability. The ability to understand and utilize the **PRECISE HEALING CURRENTS** of the interdimensional spectrum adds **greater power and skill** to the healing facilitation process. Unlike many other energy healing systems, **Level-1 Kathara Healing** trains the facilitator to identify, awaken within the body and employ in client service **specific combinations of interdimensional healing current** that fine-tune and amplify the beneficial effects of subtle energy healing. This chapter will conclude with an overview of the **4 PRIMARY TRIADIC HEALING CURRENTS,** 4 of the 5 **3-Dimensional Primal Energy Currents** that can be awakened within the body through activation of the **Maharic Shield.**

The natural **Primal Triadic Healing Currents,** which exist naturally as part of the personal Hova Body, Scalar Shield and Kathara Grid anatomy, create powerful, effective and user-friendly energetic healing tools when combined with the 12th-Dimensional frequencies of the **Maharic Shield.** Running the Triadic Currents is achieved by using the current to first activate the frequencies of the corresponding Hova Body and station of awareness within the personal Scalar Shields and **DNA Template.** Once activated in the DNA Template, the **Hova Body** is used to draw perpetual supply of the desired current from the Dimensional Unified Field corresponding to the chosen current. **Running the Triadic Currents requires activation of the corresponding frequencies within the facilitator's DNA Template, thus using the Triadic Currents in Client facilitation also serves to progressively stimulate and expedite activation of dormant frequencies in the facilitator's DNA Template.** Using the Triadic Currents in client facilitation serves to accelerate the processes of Holographic Recoding, higher identity Hova Body integration and multi-dimensional evolution within the Body-Mind-Spirit system of the *facilitator,* as well as that of the *client,* creating a condition of **CO-GENERATIVE HEALING** between client and facilitator.

Healing, Evolution and Ascension

In the Kathara Healing paradigm the art and science of healing the Body-Mind-Spirit system is viewed in terms of the *human relationship to the natural structure and dynamics of creation.* **Healing becomes a necessity when the intrinsic energetic systems and scalar-wave constructions of consciousness upon which the biological manifestation is built become incongruent with the natural Primal Order of the dimensionalized Time Matrix system.** The processes of natural healing and the hidden, organic functions of **human evolution** are intimately intertwined. Conditions that facilitate natural healing also facilitate fulfillment of the naturally intended course of human evolution, while conditions detrimental to Body-Mind-Spirit health are likewise detrimental to the evolutionary objectives of both the individual and the species.

The processes by which lasting healing occurs through re-genesis of Primal Order within the Scalar Shields and personal anatomy, are the same processes by which the **CONSCIOUSNESS** evolves to the fulfillment of **Spiritual Actualization**, through embodiment of the 12-dimensions of consciousness inherent to human DNA Template design. The genetic code of the biological imprint is built upon the Scalar-wave Templates of Consciousness that serve to hold the identity as energetic substance into physical embodiment. **Distortions within the physical manifestation indicate distortions in the Manifestation Template and thus constitute distortions in CONSCIOUSNESS.** Genuine healing is a *FUNCTION OF CONSCIOUSNESS*, first and foremost, and it is through realignment of consciousness itself that full healing can occur.

In the Kathara program it is understood that biology and consciousness are intimately intertwined and thus interdependent. The techniques of **Bio-Regenesis** provided in Kathara Healing assist in the processes of restoring the **Organic Imprint for Health** while simultaneously serving to **expedite the natural processes of interdimensional evolution** through DNA Template, Kundalini, Crystal Seals and Merkaba activation. **Human evolution is not a random, goal-less occurrence; it is rather an ordered system of interdimensional physics dynamics by which human consciousness progressively masters the contours of creation in time.** The evolutionary objective of human creation is **mastery,** which is progressively attained through the dynamics of **Transmutational Dimensional Ascension. Ascension** is therefore the intrinsic and indelible **goal** within human evolution, the **hidden impetus for action** that moves the human species forward in time.

Healing and Bio-Spiritual Evolution

When we begin running the higher dimensional frequencies of the **4 Primary Triadic Currents** for Kathara Healing applications, we are, in truth, running **FREQUENCIES OF CONSCIOUSNESS**, that emanate from the **Hova Bodies** in which the expanded spiritual portions of our personal identity are stationed. As we reawaken the higher dimensional frequencies of consciousness within the **DNA Template** we are progressively expediting the reintegration of personal **Multi-dimensional Identity**, serving the **spiritual goal of Ascension** through stimulating into activation the scientific dynamics of **Unified Field Physics** by which genuine Ascension occurs. **Kathara Healing** progressively restores the functions of universal physics dynamics within the human body, as they apply to the pathway of Ascension, and so Kathara Healing applications are considered to be processes of **ACCELERATED BIO-SPIRITUAL EVOLUTION**, as well as processes of holistic Body-Mind-Spirit healing. Both facilitator and client simultaneously receive the subtle but progressive benefits of evolutionary acceleration and spiritual development through Kathara Healing **Bio-Regenesis Technologies**, through which the expansion of personal consciousness into the higher dimensional aspects of spiritual awareness occurs.

Kathara practices serve to stimulate progressive integration and embodiment of the Multi-dimensional Identity by activation of dormant portions of the **DNA Template** through which the scalar-wave patterns of the higher frequencies of identity can enter the physical form. As spiritual advancement through Hova Body integration and DNA Template activation is a beneficial "side effect" of Kathara Healing practice, it is helpful for the Kathara facilitator to understand the order of **Multi-dimensional Identity Structure**. It is also helpful to realize that the **Unified Field of Energy** utilized in Kathara Healing is, in fact, a **UNIFIED FIELD OF CONSCIOUSNESS.** Healing is a function of energy dynamics and at its core, energy is the **substance consciousness adopts** to experience manifestation, thus **HEALING IS A FUNCTION OF CONSCIOUSNESS.** When personal consciousness is applied in ways that are congruent with the natural **Laws of Unified Field Physics**, the effects of healing and wellbeing are met in manifestation.

MULTI-DIMENSIONAL IDENTITY STRUCTURE

Consciousness and the Physical Illusion

In Chapter 1 we explored the **Structure of the Unified Field** and discovered the morphogenetic framework, or **Holographic Template** upon which matter systems are constructed. We learned how the structure of dimensionalization applies to the anatomy of the **15-Dimensional Human body**, from the **Core Template Kathara Grid**, the **Hova Bodies and Shields**, to the **Chakras**, **Bio-energetic Field** and **physical matter** structures built upon them. **We also discovered that *physical manifestation* is not what appears to be - that *nothing is really manifest at all*, but only *appears to be so* due to the *relationships of compartmentalized consciousness interacting with each other through the framework of dimensionalization*.** Our physical parts and our manifest world are **scalar wave constructions**, appearing solid and externalized because we view them with a level of consciousness that is stationed **one full dimension above** the wave forms we perceive as solid and objectified. Our consciousness structures itself into the scalar wave forms of dimensionality in order for us to perceive the illusion of manifestation. There is no difference, other than the **dimensional frequencies from which one is perceiving**, between the *substance that composes physical matter* and that which *composes our seemingly non-manifest portions of consciousness*. **The core of reality is CONSCIOUSNESS - it is the *energetic substance* out of which all things are made.**

Hova Bodies- Platforms of Perception

In this Section we will learn how the attributes of mind and sentient identity fit into the multidimensional anatomy of human consciousness. The scalar grids of the **Hova Bodies** and **Shields** that hold portions of our consciousness separate from each other so we may experience objectification, can be understood as **PLATFORMS OF PERCEPTION.** A portion of our conscious identity is **stationed within each of the Hova Body Shields,** and the dimensions contained within the Shield become the *platform* from which *that portion of our selves* will perceive matter objectification. **We are simultaneously focused within each Hova Body Platform of Perception, and this gives us separated but interrelated stations of identity and different LEVELS OF MIND from which we simultaneously perceive.** The amount of perception we are able to incorporate into our embodied focus on Earth is determined by the Level of **Higher Identity Integration** that has occurred. Integration of Higher Identity is the *merger* of the **Hova Bodies,** which creates changes and advancement in the **biological gene code** that allows the neurological and metabolic functions of the body to expand.

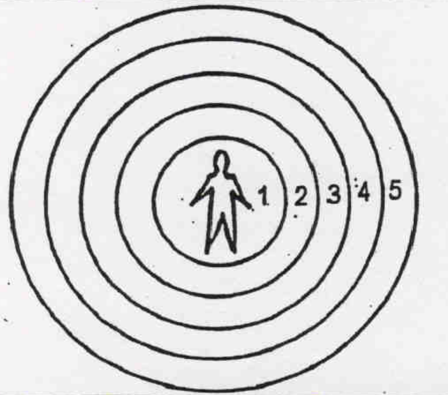

THE 5 HOVA BODIES

3-Dimensional Electro-magnetic Domains created by the Scalar Shields that serve as stations of dimensionalized consciousness

1. **NADA HOVA** Density-1
2. **ALPHI HOVA** Density-2
3. **BETCHA HOVA** Density-3
4. **MAHARA HOVA** Density-4
5. **RAJA HOVA** Density-5

Our Present Platform of Perception

Presently we are accustomed to perceiving from the Platform of Perception provided by the **NADA HOVA BODY**, the *scalar grid* that houses the 1st, 2nd and 3rd Dimensional Aspects of our identity and anatomy. While perceiving from the Nada Hova Body and the **TELLURIC SHIELD** that holds its form, we are functioning as a **3-Dimensional consciousness** and will not be aware of the perceptions of the other levels of ourselves that are perceiving from the *Platforms of Perception* from the other Hova Bodies. **As we expand and evolve we will progressively incorporate the portions of our awareness stationed in the other Hova Bodies into our earthly focus of attention.**

The *Higher Identity Structure* of the Hova Bodies and related *Levels of Mind* within dimensionalization are as follows:

Stations of Identity and Levels of Mind in the Hova Bodies

INCARNATE IDENTITY: Nada Hova Body/ Telluric Shield / dimensions 1, 2 and 3.
The Subconscious (D-1), Instinctual (D-2) and Reasoning (D-3) Levels of Mind

SOUL IDENTITY: Alphi Hova Body/ Doradic Shield/ Dimensions 4, 5 and 6
The Astral (D-4), Archetype or Dora (D-5) and the Angelic (D-6) Levels of Mind

OVER-SOUL IDENTITY: Betcha Hova Body/ Teuric Shield/ Dimensions 7, 8 and 9
The Ketheric (D-7), Monadic or Teura (D-8) and the Keriatic (D-9) Levels of Mind

AVATAR IDENTITY: Mahara Hova Body/ Maharic Shield/ Dimensions 10, 11 and 12
The Christiac (D-10), Buddhiac (D-11) and the Nirvanic (D-12) Levels of Mind.

RISHI IDENTITY: Raja Hova Body/ Rishic Shield/ Dimensions 13, 14 and 15
Universal Consciousness Mind

There are **3 additional Platforms of Perception** and 3 additional Hova Bodies beyond the structure of Dimensions;Hova Bodies that represent 3 Levels of **Ascended Mastery Identity**.

Spiritual Actualization is the process of progressive Higher Dimensional Identity Integration. ***Higher Identity Integration*** is the merger of the ***Hova Bodies,*** which creates changes and advancement in the biological ***Genetic Code*** through progressive activation of dormant portions of the ***Silicate Matrix 12-Strand DNA Template.***

5 Hova Bodies and Levels of Identity in the 15-Dimensional Time Matrix

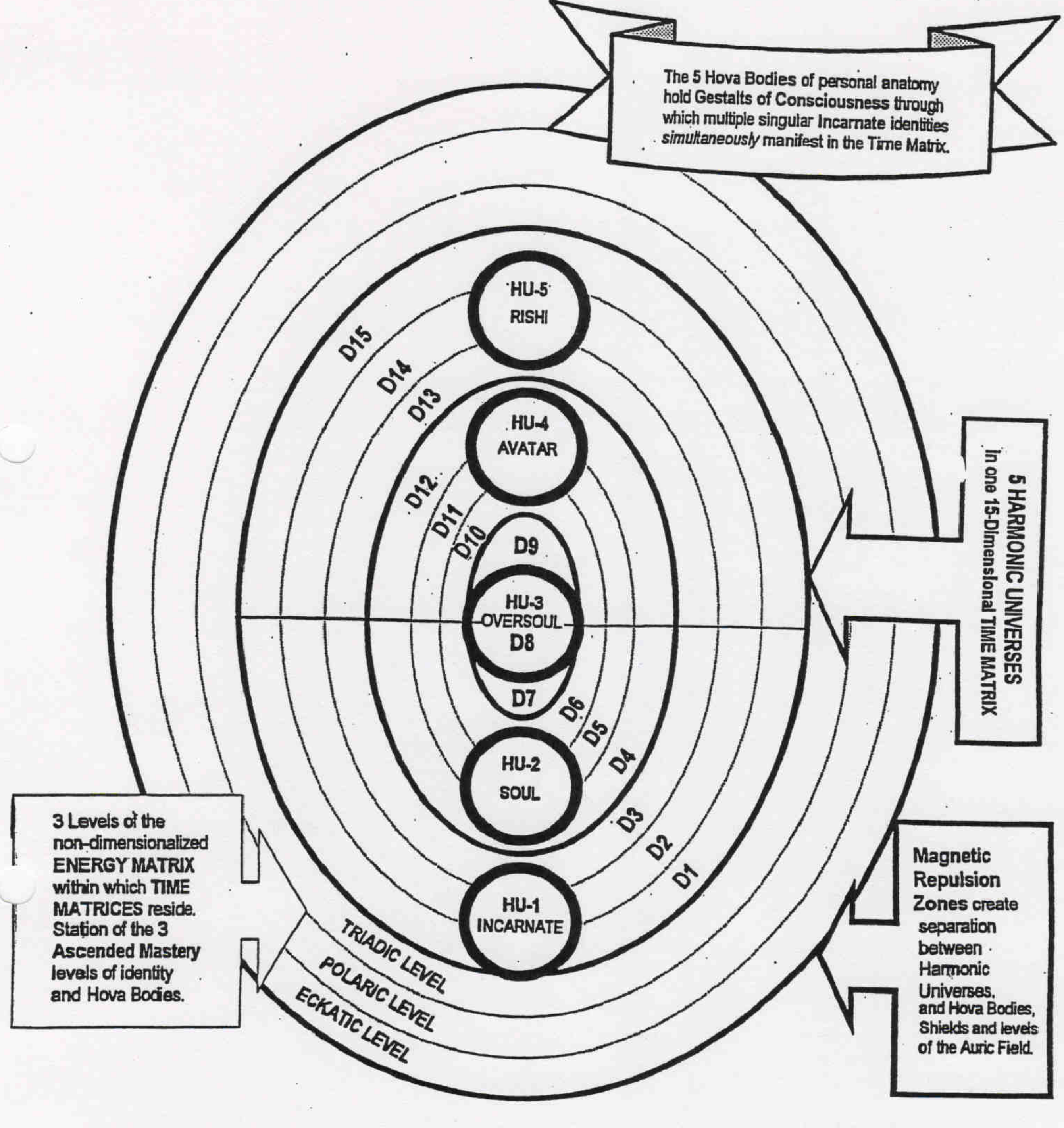
HU-5
RISHI
HU-4
AVATAR
D9
HU-3
OVERSOUL
D8
D7
HU-2
SOUL
HU-1
INCARNATE
D15
D14
D13
D12
D11
D10
D6
D5
D4
D3
D2
D1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

The 5 Hova Bodies of personal anatomy hold Gestalts of Consciousness through which multiple singular Incarnate identities *simultaneously* manifest in the Time Matrix.

5 HARMONIC UNIVERSES
In one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside. Station of the 3 **Ascended Mastery** levels of identity and Hova Bodies.

Magnetic Repulsion Zones create separation between Harmonic Universes, and Hova Bodies, Shields and levels of the Auric Field.

THE 4 PRIMARY TRIADIC HEALING CURRENTS

Hova Bodies and the 4 Primary Triadic Currents

Each of the Hova Bodies and Shields contains a portion of your energy-identity that exists in the form of electro-tonal impulse born of Partiki Phasing. The **energy of your higher stations of identity and that of the Universal Unified Fields of Consciousness-Energy can be drawn into the Nada Hova body and earthly incarnate identity for the purpose of healing facilitation.** There are **four Primary Triadic Healing Currents,** each corresponding to one of the **four higher dimensional Hova Bodies.** Through *your* corresponding Hova Body station you can *draw energy from the higher dimensional Unified Fields* and channel it into your Nada Hova Body for direction into healing within the first Harmonic. **The *4 Primary Healing Currents* are HARMONIC CURRENTS OF ENERGY, each composed of *3 Dimensions of Frequency.*** Each Dimensional Band within a Harmonic or **TRIADIC** current is referred to as a **PHASE** *of that current.* **In Kathara Healing we develop the ability to activate each progressive PHASE of each TRIADIC CURRENT, through grounding the current into the body via the activated MAHARIC SHIELD.**

Triadic Currents and Color

Each *Phase* of Triadic Current corresponds to a wave length in the dimensional spectrum that appears as COLOR to the embodied human mind. In directing the **Triadic Currents** we therefore use the **color translation of the primary dimension wave spectra** to differentiate between Phases of current for various applications. **The first current to be activated in Kathara Healing is the that of the DORADIC CURRENT,** the frequencies drawn in from Harmonic 2 through the **Doradic Shield and Alphi Hova Body. The Doradic Current has 3 Phases: Phase -1: 4th Dimensional current/ GREEN wave, Phase-2: 5th - Dimensional Current/ BLUE wave and Phase -3: 6th-Dimensional current/ Blue-Violet wave.** In **Level-1 Kathara Healing** we learn to activate and run the **Phase-1 Green Doradic Current.**

The 4 Primary Triadic Currents and Hova Body Correspondences

TELLURIC CURRENT: Nada Hova Body/ Telluric Shield / Dimensions 1, 2 and 3.
Phase-1 RED, Phase-2 ORANGE-GOLD , Phase-3 YELLOW

DORADIC CURRENT: Alphi Hova Body/ Doradic Shield/ Dimensions 4, 5 and 6
Phase-1 GREEN, Phase-2 BLUE, Phase-3 BLUE-VIOLET

TEURIC CURRENT: Betcha Hova Body/ Teuric Shield/ Dimensions 7, 8 and 9
Phase -1 VIOLET, Phase-2 GOLD, Phase-3 Mid-tone Silver

MAHARIC CURRENT: Mahara Hova Body/ Maharic Shield/ Dimensions 10, 11 and 12
Phase-1 BLUE-BLACK, Phase-2 SILVER BLACK, Phase-3 WHITE

TRANSMUTATION CURRENT

(can be transmitted only by 12th-level Avatars with full Mahara Hova Integration)

RISHIC CURRENT: Raja Hova Body/ Rishic Shield/ Dimensions 13, 14 and 15
Phase-1 TURQUOISE, Phase-2 PALE YELLOW, Phase-3 MAGENTA

Current Combining

When using **Triadic Phase Currents, each Phase current is** *combined* **with the current running from the activated** ***Maharic Shield*****, in** ***combination with*** **the Phase current from the** ***Harmonic ABOVE the Phase current you are using*****.** The **DORA DIC PHASE-1** current is therefore composed of: **D-11/D-12 PALE SILVER** (Maharic Shield), **D-4 GREEN** Doradic Shield Phase -1 AND **D-7 VIOLET,** the *Phase-1 current* of Harmonic 3 - Teuric Shield, *one Harmonic up* from Doradic Phase-1. In activating the **DORADIC-1 (Doradic Phase-1) current you are using SILVER-GREEN-VIOLET color to direct Dimensions 12-11-7-4 current.**

COMBINED TRIADIC CURRENTS

Used in Kathara Healing

TELLURIC

PHASE-1: **PALE SILVER** (D11&12) + **RED** (D-1) + **GREEN** (D-4)
PHASE-2: **PALE SILVER** + **ORANGE-GOLD** (D-2) + **BLUE** (D-5)
PHASE-3: **PALE SILVER** + **YELLOW** (D-3) + **INDIGO** (D-6)

DORADIC

PHASE-1: **PALE SILVER** + **GREEN** (D-4) + **VIOLET** (D-7)
PHASE-2: **PALE SILVER** + **BLUE** (D-5) + **GOLD** (D-8)
PHASE-3: **PALE SILVER** + **INDIGO** (D-6) + **SILVER** (D-9)

TEURIC

PHASE-1: **PALE SILVER** + **VIOLET** (D-7) + **BLUE-BLACK** (D-10)
PHASE-2: **PALE SILVER** + **GOLD** (D-8) + **SILVER-BLACK** (D-11)
PHASE-3: **PALE SILVER** + **SILVER** (D-9) + **WHITE** (D-12)

MAHARIC

(Requires full 12-Strand DNA Template activation & accelerated DNA Template coding)
PHASE-1: **PALE SILVER** + **BLUE-BLACK** (D-10) + **PALE TURQUOISE** (D-13)
PHASE-2: **PALE SILVER** + **SILVER-BLACK** (D-11) + **PALE YELLOW** (D-14)
PHASE-3: **PALE SILVER** + **WHITE** (D-12) + **PALE MAGENTA** (D-15)

In **Level-1 Kathara Healing** we begin preparing the body and DNA Template to receive the higher dimensional Phase Currents by initiating transmission of the **Doradic Phase-1 Healing Current**- the *Pale Silver* + *Green* + *Violet* frequencies of dimensions *12, 11, 7 and 4.*

Through the 7th-Dimensional aspect of the **Teuric Shield** and 11th and 12th Dimensional aspects of the **Maharic Shield**, the **Doradic-1 Healing Current** allows for recoding and realignment of the **1st-4th**-Dimensional aspects of the **Nada** and **Alphi Hova Bodies**, the **Telluric Shield** and **Physical, Emotional** and **Mental Bodies** of Density-1, the **D-4** aspect of the **Doradic Shield** and **Astral Body**, the **1st-4th DNA Strand Templates**, the **1st-4th Chakras**, and the **1st-4th Sub-harmonics** of all Scalar Shields, DNA Templates, Hova Bodies and Chakras. Doradic-1 Healing Current is effective in clearing the **Karmic-Miasmic imprint** from the **Density-1 Time Cycle incarnations** and the beginning of Density-2 Time Cycle incarnations.

CHAPTER 3 HIGHLIGHTS

KATHARA BEGINNING APPLICATIONS

THE MENTOR

- The **Hova Bodies** of personal anatomy contain portions of our higher dimensional awareness. One can begin accessing this awareness by building a **Perceptual Bridge** between the 3-Dimensional Self and higher identity aspects.
- Taking conscious initiative in building a Perceptual Bridge expedites **DNA Template activation** and the organic evolutionary process.
- The **MENTOR Technique** assists in building the Perceptual Bridge and activates dormant **Kathara Lines** in our **Level-1 Kathara Grid** to open communication between our D3 and Density-2, D4 and D5 levels of awareness.

- The Mentor is a part of our awareness that is composed of **3rd, 4th and 5th Dimensional frequencies** of consciousness, which exists as a quantity of **Scalar Wave patterns** that span **Densities 1 and 2.**

- The **Mentor Focus** will be used in **Kathara Healing** for **guidance, information retrieval** and **objective observation.** It can also be used to access **Cellular Memory,** and in **"Inner Child", Bi-location, Remote Viewing** and **Dream** work.

MORPHOGENETIC REPATTERNING AND THE MAHARIC SHIELD

- Through a process called **Morphogenetic Repatterning,** Kathara Healing effects the **Core** of Dis-ease and dis-harmony in the Body-Mind-Spirit system by **repatterning** the Scalar-wave Programs in the personal **Holographic Template** to their original **Primal Order.**

- In **Level-1 Kathara Healing** we learn the **MAHARIC RECODING PROCESS**- the most important level of Holographic Template realignment.

- The **Maharic Shield** is the 3-Dimensional Scalar Grid corresponding to the **10th, 11th and 12th-Dimensional frequencies** of **Density-4.** It represents the **Original Point of Manifestation** and holds our original "Liquid Light" **Hydroplasmic Pre-matter Creation Imprint**, which is the **Organic Imprint for Health.**

MORPHOGENETIC REPATTERNING AND THE MAHARIC SHIELD

MAHARIC SHIELD AND 12TH-DIMENSIONAL FREQUENCY

- In Kathara Healing, the **Maharic Shield** Organic Imprint for Health is used to progressively **repattern** the embodied **Kathara Grid** and body systems to their original **Primal Order.**

- The **12th-Dimensional Frequencies** of the Maharic Shield contain within them the **144 Sub-frequency Band** Scalar patterns of **ALL** of the Dimensions below.

- 12th-Dimensional **Maharic Frequency** can be used to **effectively correct distortions** in the lower dimensional levels of the Manifestation Template, Scalar Shields and Kathara Grid.

- Activation of the **Maharic Shield** serves as the **RESET BUTTON** for the embodied **Kathara Grid** and for 12 Dimensions of Body-Mind-Spirit anatomy.

THE MAHARIC SEAL AND LIQUID LIGHT CLEANSE

- In bio-energetic healing, **Level-1 Kathara Grid** distortions override realignments made to the **outer levels** of the bio-fields alone. For **lasting healing** to occur, Level-1 Kathara Grid distortions must be realigned.

- The **Maharic Shield**, with its **144 Frequency Sub-Harmonics**, is the only portion of the Morphogenetic anatomy that can fully repattern and realign the **Level-1 Kathara 12-Tree Grid. In Kathara Healing**, the **Maharic Shield** is used to realign the Level-1 12-Tree Grid.

- The **Maharic Seal and Liquid Light Cleanse** technique is used to restore distorted **Partiki Phasing Sequences** and **Scalar-wave arrangements** within the **3 Levels of the Kathara Grid** and **Scalar Shields.**

- Realignment will be **temporary** until the Kathara Grid reaches "**Critical Mass**" of energy in the **new**, realigned pattern.

- Until **critical mass** is reached, the outer levels of the anatomy will continue to **reverse** the Kathara Grid healing. With practice and **consistent application** of the **Maharic Seal and Liquid Light Cleanse** technique, the realignment of the Kathara Grid will become **permanent.**

- In future applications of Kathara Healing processes will be used to realign the **outer levels** of the bio-field and the **Kathara Grid simultaneously,** to expedite the healing process.

- The Focus of **Level-1 Kathara Healing** is **Morphogenetic Repatterning** and **Holographic Template Recoding** via the **Maharic Shield,** to clear distortions and the **Karmic-miasmic imprint** from the personal **Scalar Shields, Kathara Grid** and **DNA Template,** utilizing the **accelerated energies** of the *Planetary Bio-feed Interface System* **(PBIS).**

KATHARA POINTS

- **Kathara Points** are mini energy-vortex points in the body that connect directly to **Diodic Point vortices** in the **Level-3 Kathara Diodic Point Grid.**

- Stimulation of the Kathara Points through **manual massage** effects the **movement of energy** within related areas of the Kathara Grid and body systems.

- Used in sequence, **Sequence-1 Kathara Points for Regeneration** *open* energetic gateways to the **Kathara Lines** of the **Kathara Level-1 12-Tree Grid**, to allow new, realigned scalar patterns from the **Maharic Shield** to run through the Kathara Lines and Kathara Grid for **Core Template Repatterning.**

- **Sequence-1 Kathara Points** create **more lasting but less immediate** effects of realignment and healing. They can be **used alone** for **deep healing** and re-balancing of the **Core Kathara Grid.**

- **Sequence-2 Kathara Points for Revitalization** *close* **Kathara Lines** in the 12-Tree Grid and **channel** energy into the **Nadial Lines** of the **outer bio-energetic field** for more rapid revitalization and re-balancing of the **outer auric levels, chakras** and **physical cells.**

- **Sequence-2 Kathara Points** create **more immediate but less lasting** effects of realignment and healing. They can be **used alone** to create **temporary revitalization** of the **outer levels** of the bio-system.

- **Used together, Sequence-1** Kathara Points can run **Maharic energy** through the **core Kathara Grid** for **Template Recoding. Sequence-2** Kathara Points can then be used to **direct the Maharic energy** into the **outer levels** of the bio-field and physical body, to **expedite healing** through **simultaneous** inner and outer bio-field realignment.

✹ In **Kathara Healing, Sequence-1** Kathara Points are used to **OPEN** the Kathara Lines **prior to** running the **Maharic Seal and Liquid Light Cleanse** and **Sequence-2** Kathara Points can be used **following the** procedure to **revitalize** the outer bio-field and body systems.

THE MAHARIC SEAL, PBIS AND HEALING

- Use of the **Maharic Seal and Liquid Light Cleanse** technique begins activation of dormant **Chakras 8-12** and the **Planetary Bio-feed Interface System (PBIS).**

- **Opening the PBIS** allows energy from the **Planetary Shields** to be drawn into the **human body** for **rapid activation** of the **personal Maharic Shield** and **Crystal Seals, to expedite** the processes of **DNA Template activation, healing, recoding of the Holographic Template** and interdimensional evolution.

- Opening of the PBIS can also allow the human **DNA Template** to serve as a **Bio-electric Conduit**, through which the **Planetary Scalar Shields** and Electro-magnetic Grids can be realigned with their **Primal Order** for **geophysical planetary healing.**

- Through the PBIS humans can directly facilitate healing of the **Planetary Templar Complex** from the Manifestation Template level, as well as facilitate healing in other biological organisms.

- Each use of the **Maharic Seal and Liquid Light Cleanse** technique assists in **planetary healing**, as the process progressively stimulates the **Planetary Maharic Shield** into activation, to realign **Planetary Shield** distortions.

- The **Maharic Seal and Liquid Light Cleanse** technique is used as **preliminary preparation** for **ALL Kathara Healing applications.** It creates a temporary **12th-Dimensional Scalar Frequency Seal** in the facilitator's bio-field to **protect** both facilitator and client from taking on each other's bio-field distortions.

- In **healing applications**, the **Maharic Seal** is needed to run interdimensional healing currents to their most effective **12th-Sub-Harmonic level.**

BIO-SPIRITUAL EVOLUTION, ASCENSION AND HEALING

- The processes by which **healing** occurs through re-genesis of Primal Order are the same processes through which true **Spiritual Actualization** takes place.

- **Biology** and **consciousness** are **intimately intertwined** through the natural **Laws of Universal Unified Field Physics**, the **energetic dynamics** through which consciousness enters manifestation.

- The hidden **impetus for action** behind human evolution is the core species objective of **Mastery of Co-creation** within the space-time system. This **Evolutionary Objective** is obtained through achievement of the Evolutionary **Goal** of **Transmutational Dimensional Ascension.**

- **Kathara Healing** applications generate **healing** through restoring the **Original Imprint for Health.**

- Restoration of the **Original Imprint for Health** simultaneously expedites the natural processes of **evolution,** through activation of the **12-Strand DNA Template, Kundalini** energies, and **Merkaba Vehicle** and integration of the **Hova Bodies, Scalar Shields** and higher dimensional aspects of personal **spiritual identity.**

- **Kathara Healing** facilitates healing while simultaneously facilitating **Spiritual Actualization** and expediting the Goal of **Transmutational Dimensional Ascension.**

- **Healing is a function of consciousness.** There is no difference between the substance that composes physical matter and the substance that composes consciousness, other than the **dimensional frequency** from which one is perceiving.

- When personal consciousness energy is applied in **congruence** with the natural Laws of Universal Unified Field Physics, **healing, wellbeing** and **Spiritual Actualization** result.

MULTI-DIMENSIONAL IDENTITY STRUCTURE

- **Human anatomy** is a **15-Dimensional Scalar-wave construct.** Part of our identity exists **simultaneously** within the Time Cycles of the higher dimensional **Harmonic Universes.**

- We are indelibly **linked** to our higher identity levels through our embodied **Kathara Grid, Scalar Shields** and **DNA Template.** As we evolve biologically to activation of dormant portions of our DNA Template and Scalar Shields, the conscious awareness of our higher identity levels **progressively embodies.**

- The 3-Dimensional Hova Body **Electro-magnetic Domains** and Scalar Shields that form them represent **Platforms of Perception** through which our higher identity levels perceive.

- **Merger** of the **Hova Bodies** via activation of the **DNA Template** and **Scalar Shields** creates **reintegration** of the **Multi-Dimensional Identity Structure.**

4 PRIMARY TRIADIC HEALING CURRENTS AND IDENTITY

- The **5 Levels of Multi-Dimensional Identity Structure** are: **Incarnate**-Density-1, **Soul**-Density-2, **Oversoul**-Density-3, **Avatar**-Density-4 and **Rishi**-Density-5. There are also **3 Ascended Master** identity levels, Hova Bodies and Shields existing **beyond** the dimensionalized **Time Matrix.**

- **In Kathara Healing** we use the various levels of higher dimensional identity to draw in **healing current** from the higher dimensional Unified Fields, by activating the portions of the **personal Scalar Shields, Kathara Grid** and **DNA Template** that correspond to the higher dimensional **Hova Bodies.**

- There are **4 Primary Triadic** (3-Dimensional) **Energy Currents** that can be used in **Healing Facilitation**: The **Telluric Current**-Density-1, the **Doradic Current**-Density-2, the **Teuric Current**-Density-3 and the **Maharic Current**-Density-4. The **Rishic Current** of Density-5 can be run only by beings with **full 12-strand DNA activation.**

- Each of the **4 Primary Triadic Healing Currents** corresponds to one of the first 4 **Hova Bodies** of human anatomy.

- In Kathara Healing, the 4 Primary Triadic Currents are harmonically interwoven to create 4 powerful **Combined Healing Currents.**

- Each of the 4 Triadic Healing Currents has **3** Phases, or **3 Phase Currents,** each representing one of the three dimensional frequency bands within the **Triadic Current.**

- Each of the **12 Phase Currents** is run in combination with the 11-12th-Dimensional **Maharic Current** and the Phase Current from the **harmonic directly above,** to form a powerful **Triadic Healing Current.**

- In **Kathara Healing,** specific Healing Currents are directed by using the **Color** of their dominant **wave-length.**

DORADIC-PHASE-1 CURRENT

- **Level-1 Kathara Healing** teaches us to activate our **Doradic-Phase-1 Current**- the **Pale Silver** (D11-12) **+ Green** (D-4) **+ Violet** (D-7)- for **Healing Facilitation** and **personal** DNA Template and Shield activation.

- The **Doradic-Phase –1 Current** can realign the **1st-4th** Dimensional aspects of the **Scalar Shields**, the **1st-4th Chakras, Axi-A-Tonal Lines** and **DNA Strand Templates**, the **Physical, Emotional** and **Mental** bodies of the **Nada Hova Body** and the D-4 **Astral body** of the **Alphi Hova Body.**

- **Doradic-Phase-1 Current** can also be used to realign the **1st-4th Sub-Harmonics** of ALL Shields, DNA Strand Templates, Axi-A-Tonal Lines, Chakras and Hova Bodies and to clear the **Karmic-Miasmic Imprint** from the **Density-1** incarnational Time Cycles and the **first part** of the **Density-2** incarnational Time Cycles.

CHAPTER 4

GENERAL KATHARA HEALING FIELDWORK

Skills to Facilitate Healing

- OPENING THE HEALING CHANNELS- DORADIC-PHASE-1 CURRENT, THE DORADIC SPHERE, DORADIC CORDS AND DORADIC BALLS.
- THE KARMIC-MIASMIC IMPRINT, INCARNATIONAL TIME VECTORS, VECTOR IMBEDDING.
- MULTI-VECTOR HOLOGRAPHIC RECODING THROUGH THE MAHARIC INFUSION.
- ATTITUDES AND ETIQUETTE IN HEALING, THE MOTIVATION TO HEAL, POSTURES OF LOVE, THE FRAMEWORK FOR HEALING AND FULLY ENGAGING THE LOVE VIBRATION.
- THE 6 STEP MAHARIC RECODING SESSION, MELCHIZEDEK CLOISTER ORDINATION AND BECOMING A PROFESSIONAL SPIRITUAL HEALING FACILITATOR-INSTRUCTOR.

TECHNIQUES:

#3- OPENING THE HEALING CHANNELS- DORADIC –PHASE-1 CURRENT

#4- MULTI-VECTOR HOLOGRAPHIC RECODING KATHARA SCAN

#5- THE MAHARIC INFUSION

THE MAHARA SPACE COMFORT TOUCH

TOPICS OF DISCUSSION

OPENING THE HEALING CHANNELS

DORADIC-PHASE-1 CURRENT

Interdimensional Currents and Signet Star Crystal Seals

Activation of any **Triadic Healing Current** requires temporary release of the corresponding **Signet Star Crystal Seals** within the personal **Scalar Shield** and a temporary activation of the portions of the dormant **12-Strand DNA Template** that correspond to the Signet Seals. Signet Seals modulate the flow of frequency between the 3-Dimensional Harmonic Density Levels. When Signet Seals are in their usual (for this time period) dormant phase, they keep the **5 Hova Bodies** of the personal 15-Dimensional anatomy separate from each other by creating an electro-magnetic barrier or **Magnetic Repulsion Zone** between each 3-Dimensional level of the energetic anatomy. This *Magnetic Repulsion Zone* between embodied **Harmonics of Manifestation** manifests as **polarity resistance to frequency inflow** within the 3 Levels of the Kathara Grid, the Scalar Shields, Axi-A-Tonal Lines, Chakras and Auric Field levels of the embodied **Inner Templar Complex. In order to run higher dimensional frequency currents through the physical body for healing applications, it is necessary to temporarily release portions of the Magnetic Repulsion Zones between the Hova Bodies.** Releasing the electro-magnetic resistance to higher dimensional frequency inflow within the Inner Templar Complex is achieved through temporary activation of the **Signet Seals** within the **Kathara Level-2 Crystal Seals Grid** of the personal **Scalar Shields.**

Most contemporary energetic healing systems rely upon using the 1st, 2nd and 3rd-Dimensional frequencies of the **TELLURIC PHASE CURRENTS** in the **Nada Hova Body**. Through the Nada Hova Body alone, only the 3-Dimensional **Sub-Harmonics** of higher dimensional frequency currents can be run, as the dormant *Signet Seals* block the faster moving *Partiki Phasing Flash-line Sequences* of higher dimensional frequency bands from running through the DNA Template. **If a frequency band is blocked from running through the DNA Template that frequency cannot be energetically transmitted through the Nada Hova Body and physical bio-electrical systems.** In **conventional energy healing methods**, even those emerging from ancient systems of subtle-energy healing, the **Signet Seals**, and thus the *Scalar Shields, Hova Bodies* and *DNA Template* remain dormant in their usually **closed** position, **phase-locked** into a 3-Dimensional orientation. While the DNA Template remains phase-locked into the 3-Dimensional orientation of the Nada Hova Body, only the **1st through 12th Sub-frequency bands** (Sub-Harmonics) of **1st through 3rd Dimensional Current** (Telluric Current), and the **1st through 3rd Sub-frequency bands** of **4th through 12th Dimensional Current** can be run through the human body for transmission. Use of the lower frequency (slower Partiki Phasing rhythm) **Telluric Currents** can effect **temporary** beneficial change in the **outer layers** of the bio-energetic field, but Telluric Currents alone cannot facilitate full repatterning of the **3 Levels of the Kathara Grid** where **Scalar Shield** distortions that cause **genetic malfunction** are embedded. The **11th-12th-Dimensional frequencies** of the Pre-matter Hydroplasmic **Maharic Currents** are the only dimensional frequencies powerful enough to affect change in the embodied **Kathara Grid**. To run **Maharic Current**, and the high frequency interdimensional **Triadic Phase Currents**, through the human body requires release of the **Signet Seals within the personal Scalar Shields.**

Triadic Phase Currents, the Maharic Shield and the PBIS

Most healing systems have not incorporated the **Maharic Current** into their frequency application procedures because the **Planetary Templar Complex** has not been able to run a D-12 **Maharic Current** through the planetary grids for **210,216 years.** Like the human **DNA Template**, if the **Planetary Shields** and Manifestation Template are blocked from processing interdimensional frequency, that frequency will be **unavailable for use** on the planet. If the **Signet Star Crystal Seals** (Star Gates) in the **Planetary Shields** are dormant, closed and phase-locked into Density-1 orientation, no one on the planet will be able to draw interdimensional frequency into their **personal Shields** because the frequency is unavailable within the **Planetary Shields.** The personal and planetary Templar Complexes are **intimately interwoven**, and what effects one will inevitably effect the other.

In order to run the full spectrum of interdimensional **Triadic Phase Currents** through the human body it is necessary to open the **Signet Seals** in the personal **Inner Templar** and the **Signet Star Gates** in the **Planetary Templar Complex.** In conventional human evolution, the internal Signet Seals, which serve as the **Genetic Time Codes** or **"Fire Codes"** of the human **DNA Template**, automatically, progressively and sequentially open over **very long cycles of time.** During the slow activation of the Genetic Time Codes the conscious identity experiences passage through numerous incarnate lifecycles within the planetary Time Cycle, evolving through a series of **Single Vector incarnations** before being released from the 3-Dimensional phase-lock of Density-1 manifestation. One can **manually open the Signet Seals** in the embodied Inner Templar Complex to **expedite evolution** into the Multi-dimensional, **Multi-Vector Time Cycles** through use of the 11th-12th Dimensional **Maharic Current.**

Activation of the personal **Maharic Shield** allows for rapid release of the embodied **Signet Seals** and activation of the **Fire Codes** in the DNA Template. The **Maharic Current** from the personal Maharic Shield can be used to release the **Signet Star Crystal Seals** within the body, which in turn activates the dormant **Kundalini** energies at the base of the spine and the **Cranial-Sacral Kundalini Seals** within the **Tailbone** and **Pineal Gland.** Release of the Cranial-Sacral Seals of the Internal Templar Complex allows the **Magnetic Repulsion Zones** between the **Hova Bodies** to release so the DNA Template, personal Scalar Shields and physical body systems can receive inflow of higher dimensional frequency. To gain conscious mastery over the process of releasing the internal Signet Seals requires activation of the personal **Maharic Shield**, and the personal Maharic Shield can be **manually activated** only through drawing **Maharic Current** into the personal Shields via the **Planetary Maharic Shield. The entire process of releasing the internal *Signet Seals* by activating the personal *Maharic Shield*, in order to enable the *DNA Template* to run *Triadic Phase Currents* for healing is entirely dependant upon the *operational function* of the *Planetary Bio-feed Interface System* (PBIS).**

If the **Planetary Templar Complex** is blocked, or the Signet Star Gates are dormant, and **Maharic Current** from the **Planetary Maharic Shield** is not able to run through the planetary grids, humans on Earth do not have the option of **manually activating the personal Maharic Shield**, because the **PBIS** is not operational. The **PBIS** becomes **operationally functional** within the human body through activation of the **12th Chakra**. The *12th Chakra* cannot activate in the personal Templar Complex if the **Planetary Templar Complex** cannot run 12th-Dimensional **Maharic frequency.** For over 200,000 years, the Planetary Templar Complex could not run Maharic frequency, and thus manual activation of the personal Maharic Shield and release of the embodied Signet Seals to open the Hova Bodies for running interdimensional Triadic Phase Currents for healing were not available options for Earth humanity. Since the recent activation of the **Planetary Templar Complex, Maharic Shield** and **PBIS** activation and **Kathara Healing** applications are now available on Earth.

Opening Triadic Phase Currents in the Human Body

The method by which **Triadic Phase Currents** are activated within the human body is a **progressive process.** First, opening of the personal **Kathara Grid, Telluric Shield** and **Nada Hova Body** is orchestrated through activation of the **Maharic Shield** and **Signet Seals**, to allow inflow of frequency from the higher dimensional **Hova Bodies**. Next, the incoming frequency is harnessed and **collected** within the **Chakra System**. Finally, the collected frequency is then **transmitted** for healing facilitation in others, either through direct **Chakra transmission** or by close-range **Frequency Induction** via transmission of frequency from the activated **Secondary Chakra Centers** within the **PALMS.**

The ability to run the higher dimensional Triadic Phase Currents within the body must be built up within the Internal Templar Complex, starting with the **slowest-moving current** of the higher dimensional spectrum- the **DORADIC-PHASE-1 CURRENT.** The *Doradic-Phase-1 Current* utilizes the **sub-frequencies** of the **11th-12th**-Dimensional **Maharic Current** from the Density-4 *Mahara Hova Body* as a **carrier wave** that will carry the desired interdimensional currents through the **Signet Seals** of the personal **Scalar Shields.** The **full spectrum** of the **4th** Dimensional **Doradic-Phase-1** current from the Density-2 *Alphi Hova Body* is activated as the **primary healing frequency** for transmission. The **full spectrum** of the **7th**-Dimensional **Teuric-Phase-1 Current** (one Harmonic above the Doradic-Phase-1 Current) from the Density-3 *Betcha Hova Body* is used to release the 4th-Dimensional **Signet Seal,** allowing the 4th-Dimensional Doradic-Phase-1 current to enter the **Nada Hova Body.**

Activating the **4th Signet Seal** triggers **temporary** activation of the **4th DNA Strand Template** and corresponding **4th Kathara Center** in the Level-1 12-Tree Grid, which allows the 4th-Dimensional frequencies of the **Doradic-Phase-1 Current** to pass through the **active DNA** and into the bio-energetic and bio-electrical systems of the body for transmission. Once the DNA Template becomes accustomed to holding a **4th-DNA Strand Template Activation,** after **several months** of practice in running the 4th–Dimensional *Doradic-Phase-1 Current*, the body will be able to begin receiving the primarily 5th-Dimensional **Doradic-Phase-2 Current. The lower dimensional Triadic Currents must be activated within the DNA Template before the higher dimensional Triadic Currents can run.** In **Level-1 Kathara Healing** the first step of activating the full spectrum (12 sub-harmonics) of the 4th-Dimensional **Doradic-Phase-1 Current** is achieved.

Triadic Phase Currents and Color Cap-Stones

The **Color Cap Stone** of a dimensional frequency band is the visual representation of the **primary wave-length** formed by the specific **Partiki Phasing Rhythm** for that dimension. A specific dimensional frequency band can be chosen and directed by the **color translation** of its primary wave-length. The **Color Cap Stones** of the dimensional frequency bands are used in Kathara Healing to direct specifically selected dimensional frequency bands (Partiki Phasing Rhythms) for specific applications. The **Color Cap Stones** for the **Doradic-Phase-1 Current** are **PALE SILVER** (**D-11**-D-12 Maharic) + **GREEN** (D-4 Doradic) + **VIOLET** (D-7 Teuric). In **technique # 3 – OPENING THE HEALING CHANNELS-DORADIC-1** - the ***Pale Silver* + *Green* + *Violet*** Cap Stone Colors will be used to activate the **Doradic-Phase-1** Healing Current within the body.

DORADIC SPHERE, CORDS AND BALLS

The Doradic Sphere and Current Collection

The **DORADIC SPHERE** is a **frequency repository** that is formed to hold the activated **Doradic-Phase-1 Current** within the **Chakra system** for transmission in healing facilitation. As the Doradic-Phase-1 Current is primarily a **4th-Dimensional Current**, the **Doradic-1 Sphere** will be formed in the **4th Chakra**, which corresponds to the 4th-Dimensional **ASTRAL body** of the Alphi Hova Body and Auric Field. **Transmission** of the Doradic-Phase-1 Current will take place from the **Doradic Sphere** at the **4th Chakra**, creating a **4th-Dimensional ASTRAL INFUSION** of frequency into the bio-energetic field of the client in **close-range healing**, and a general **Astral Plane Infusion** of Doradic-Phase-1 frequency into **Manifestation Templates** of areas(and everything contained within the space) when transmitted directly from the **4th Chakra.**

Doradic Cords and Current Induction

The **DORADIC CORDS** are **energy conduits** that are formed by opening the embodied **Axi-A-Tonal Lines** within the **arms, hands** and **fingers.** Opening the Doradic Cords allows the **Secondary Chakras** within the **palms** of the hands to carry the **Doradic-Phase-1 Current** for application in **close-range** transmission. Though Triadic Phase Currents can be transmitted directly from the **Chakra Center** in which the **Collection Sphere** is placed, the energies are more **diffused** and **diluted** in this form of transmission, unless the personal DNA Template holds a **full, permanent activation** of the Strand Template corresponding to the Triadic Current being transmitted. It takes time, often years, for the DNA Template to acquire and permanently hold a higher strand activation level. One-on-one healing facilitation is more effective when conducted through **concentrating** the Triadic Current through the **Axi-A-Tonal Lines** for **DIRECT INDUCTION** of current through the **Palm Chakras** in **close-range** transmission. Close-range transmission also allows for **more precise application** of frequency to specific areas of the body systems. In **Level-1 Kathara Healing**, the **Doradic Cords** are activated within the arms, palms and fingers to allow for concentration, focused direction and **Direct Induction** of the **Doradic-Phase-1 Current.**

For **general healing** purposes, such as **"clearing rooms"** or geographical spaces, or when attempting to transmit healing current into **groups of people,** transmission of the **Doradic-Phase-1 Current** can be orchestrated directly from the **4th Chakra** for **NON-SPECIFIC INDUCTION.** In **Non-Specific Induction,** following activation of the Doradic-Phase-1 Current, each **INHALE** breath is used to fill the **Doradic Sphere** with more energy and each **EXHALE** breath is used to expand and transmit the energy into the room or bio-fields of a group of people. It is not necessary to activate the Doradic Cords for Non-Specific Induction applications. In **Level-1 Kathara Healing** applications use of the **Doradic Cords** and **Palm Chakras** is employed for Direct Induction close-range transmission in Client Sessions. To enhance and expedite **personal healing,** the activated **Doradic-Phase –1 Current** can be run through the personal **Level-1 Kathara Grid** with the ***Maharic Shield and Liquid Light Cleanse*** technique and then through each of the **7 Primary Embodied Chakras.**

Doradic Balls and Palm Transmission Direct Induction

The **DORADIC BALLS** are small energy repositories that form in the **Palms** of the Hands when the **Doradic Cords** have been activated to run Doradic-Phase-1-Current through the **Palm Chakras.** Doradic Balls are used to **accumulate current** in the **Palm Chakra vortices** prior to transmission, to **strengthen** the intensity of the current.

PALM CHAKRAS AND TRANSMITTING CURRENT

Primary Polarities of Palm Chakras

Along with the **15 Primary Chakras** within human anatomy there are many other smaller **Secondary** and **Tertiary Chakra Centers** throughout the body, which correspond to the energetic interrelationships of scalar-wave patterns within the **Kathara Level-3 Diodic Points Grid**. All Chakras operate as sets of **Black and White Holes,** through which energy from the Dimensional Unified Fields circulates into and out of the body and Manifestation Template. The **Palm Chakras** are part of the **Secondary Chakra System**. Like all Chakras, the Palm Chakras are structured as **2 vortices** of energy **joined at the tip** by a minute **Crystal Seal** within the **Kathara Level-2 Crystal Seals Grid**. The Palm Chakras run through the **center of the Palms** and out of the **back of the hands.**

In any set of Chakra vortex spirals, one vortex represents the **magnetic Black Hole,** a *counter-clockwise rotating* spiral that **draws energy IN** toward the central Crystal Seal, and the companion vortex represents the corresponding **Electrical White Hole,** the *clock-wise rotating* spiral that **sends energy OUT** from the central Crystal Seal. The **Black Hole** vortex represents the **RECEIVING spiral** or **Magnetic Pole** and the **White Hole** vortex represents the **TRANSMITTING spiral** or **Electrical Pole.** The **Electrical Pole** of a Chakra represents the **EXPANSION CURRENT** of the Chakra, through which it **expands energy out** from the personal Shields; it is referred to as **BASE ELECTRICAL.** The **Magnetic Pole** of a Chakra represents the **CONTRACTION CURRENT** of the Chakra, through which it **contracts energy in** from the external Unified Field to the personal Shields; it is referred to as **BASE MAGNETIC.**

The Chakras in the **Palms** are composed of many smaller minute **vortex sets**, each with Magnetic and Electrical Poles, but the **PRIMARY POLE** of each Palm Chakra is that pertaining to the **larger Black-White Hole set** and **Crystal Seal.** Though each Palm Chakra is capable of both **Transmitting and Receiving** energy current, the **strongest and most focused currents** can be achieved by utilizing the natural polarity orientation of the Primary Vortex Set. The Chakra in the **RIGHT PALM** is **BASE ELECTRICAL;** its corresponding vortex on the back of the right hand is **Base Magnetic.** The Chakra in the **LEFT PALM** is **BASE MAGNETIC;** its corresponding vortex on the back of the left hand is **Base Electrical.** When running Triadic Phase Currents it is helpful to understand the **Primary Polarities of the Palm Chakras** in order to create the **strongest** and **least distorted** healing current. In **Kathara Healing,** the **RIGHT HAND** is the **TRANSMITTING HAND,** through which Triadic Current is **Exhaled** down the **Cords** in the arms and **EXPANDED** into the **RIGHT PALM CHAKRA** for **TRANSMISION.** The **LEFT HAND** is the **RECEIVING HAND,** through which the Triadic Current transmitted from the Right Hand is **drawn back** into the facilitator's body **TO CREATE A CLOSED CIRCUIT OF CURRENT** for the **strongest current intensity** and **purest frequency content.**

HAND POSITIONS

Standard Hand Position for Transmitting Triadic Current

When transmitting Triadic Current, the **LEFT PALM** is **PLACED OVER THE BACK OF THE RIGHT HAND**. The *Base Magnetic* Current of the *Left Palm* **interfaces** with the *Base Magnetic* Current in the *Back of the Right Hand*, creating a **MAGNETIC REPULSION ZONE**, which **blocks the vortex on the Back of the Right hand** from drawing **non-specific frequency** from the Density-1 Unified Field. This keeps the current transmitted from the *Right Palm Chakra* **pure** to its specifically chosen content. The *Base Magnetic* vortex in the *Left Palm* simultaneously draws the **transmitted current** back up through the smaller vortex sets after it runs through the client's body, creating a **closed circuit of pure frequency** running between the **Hova Bodies** of the **client,** the **facilitator** and the higher dimensional **Unified Fields**. Options of **Frequency Modulation** for other healing applications also become available through using this **HAND POSITION** when transmitting healing current. In some applications a **modulated current** that includes bursts of Density-1 Telluric Current interspersed with current transmission from the higher dimensions is desired; this can be achieved by moving the Left Hand on and off of the vortex on the back of the Right Hand. The **LEFT OVER RIGHT HAND POSITION** is used in **Level-1 Kathara Healing** applications.

Elements of Doradic-Phase-1 Current Activation

Doradic-1 Cap Stone Colors:

PALE SILVER (D-11&D-12) + **GREEN** (D-4)+ **VIOLET** (D-7)

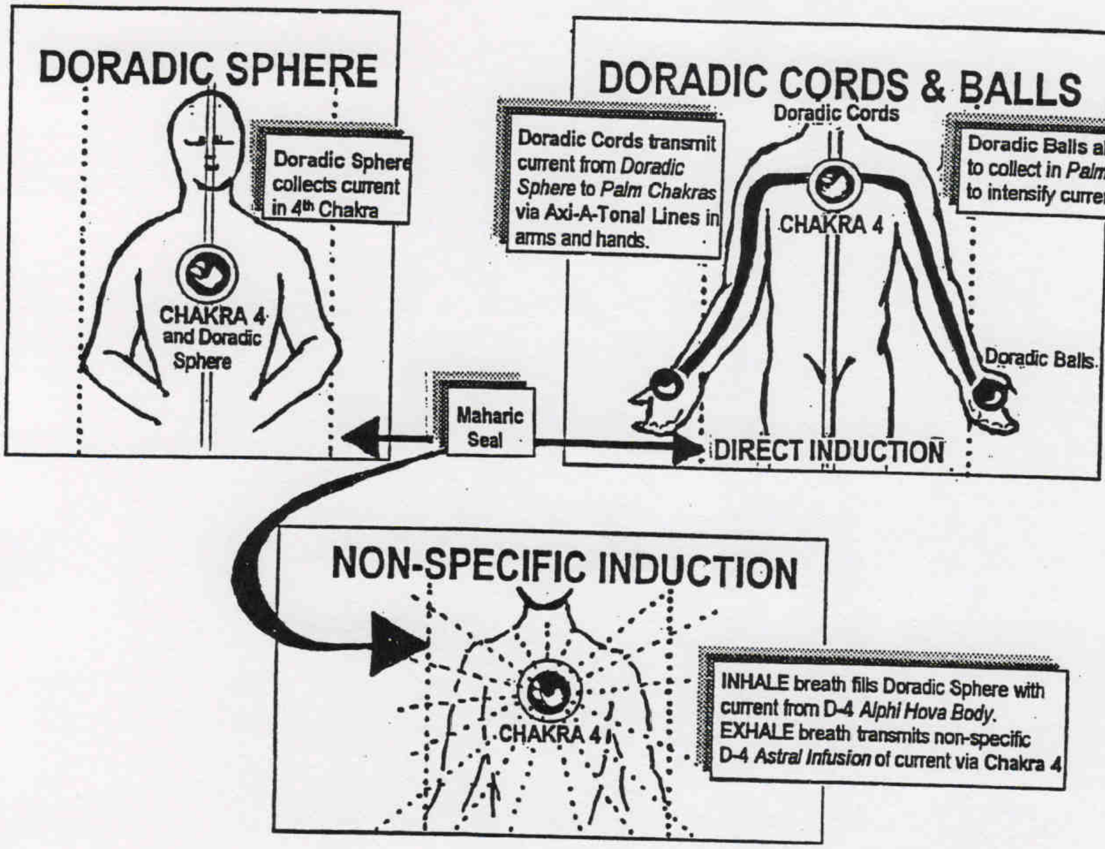
DORADIC SPHERE
CHAKRA 4 and Doradic Sphere
DORADIC CORDS & BALLS
Doradic Cords
CHAKRA 4
Doradic Balls
Maharic Seal
DIRECT INDUCTION
NON-SPECIFIC INDUCTION
CHAKRA 4

Doradic Sphere collects current in 4th Chakra

Doradic Cords transmit current from *Doradic Sphere* to *Palm Chakras* via Axi-A-Tonal Lines in arms and hands.

Doradic Balls allow current to collect in *Palm Chakras* to intensify current.

INHALE breath fills Doradic Sphere with current from D-4 *Alphi Hova Body*. EXHALE breath transmits non-specific D-4 *Astral Infusion* of current via Chakra 4

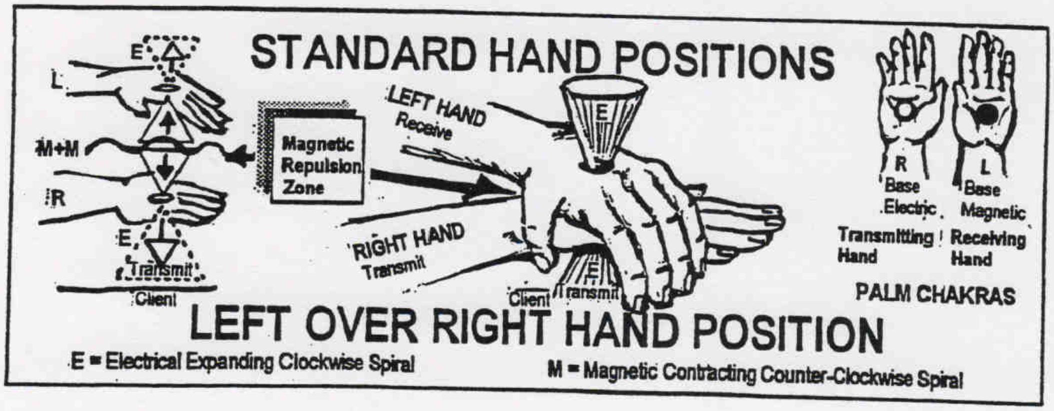
STANDARD HAND POSITIONS
E
L
M+M
R
E
Transmit
Client
Magnetic Repulsion Zone
LEFT HAND Receive
RIGHT HAND Transmit
E
E
Client
Transmit
R
Base Electric
L
Base Magnetic
Transmitting Hand
Receiving Hand
PALM CHAKRAS
LEFT OVER RIGHT HAND POSITION

E = Electrical Expanding Clockwise Spiral

M = Magnetic Contracting Counter-Clockwise Spiral

Intrinsic Process of Activating a Triadic Phase Current

ACTIVATE MAHARIC SHIELD

1. Open the PBIS using the D-11 – D-12 Wave Guide Symbol Code, the **Hierophant Symbol**, to open the **12th Chakra** and stimulate release of **Maharic frequency** from the **Planetary Maharic Shield.**
2. From the PBIS draw Maharic frequency from the Earth's Maharic Shield **up** into the dormant personal Maharic Shield 12" below the feet, to activate **the personal Maharic Shield.**

ACTIVATE TRIADIC PHASE CURRENT

3. Choose the specific Triadic Current you desire to run.
4. Use the inflow of **Maharic Current** from the personal and planetary Maharic Shield as a **Carrier Wave** to carry the **Cap Stone Color frequencies** of the **Signet Star Crystal Seals** in your Scalar Shields that correspond to the **Triadic Phase Current** you wish to activate.
5. Temporarily **release** the **Signet** and **Cranial-Sacral Kundalini Seals** and activate the **DNA Strand Template** corresponding to the chosen **Triadic Phase Current**, by running the **Primary Frequency** and the frequency **one Harmonic above it** through the **Signet Seals**, on the Maharic carrier wave current, to open the **Nada Hova Body** and physical body systems to inflow of the higher dimensional **Triadic Phase Current**.

COLLECT AND TRANSMIT HEALING CURRENT

6. As the Triadic Phase currents transmits into the body through the Signet Seals, Kundalini energies and DNA Template, **collect and build the current** as a **Sphere** of energy within the **Chakra** corresponding to the Primary Triadic Current you are running.
7. Run the current from the **Collecting Chakra** through **the Axi-A-Tonal Line Cords** in the arms, to activate the **Palm Chakras.**
8. **Exhale** current from the **Chakra Collection Sphere** through the **Cords** and allow it to build into **Balls** of energy within the **Palm Chakras** before transmitting.
9. Transmit the Triadic Current through the **RIGHT PALM CHAKRA** into client's bio-energetic field while drawing more current into the **Chakra Collection Sphere** with each **exhale** breath and draw the current back through the **LEFT PALM** on each **inhale** breath to create a higher dimensional Closed Circuit Healing Current. Use **LEFT PALM OVER BACK OF RIGHT HAND** hand position.

TECHNIQUE # 3

Opening the Healing Channels- Doradic-Phase-1 Current

1. Breathe slowly and deeply for several breaths while visualizing or imagining a **4" Violet Dot** on the center of the *inside of your forehead.* Make the color image as strong as you are able.(If you can't yet perceive inner colors or visualization, simply imagine the Violet Dot as being there). Move the Violet Dot to the center of the brain at the **Pineal Gland**, and then upward into the 7th **Crown Chakra**, from the inside of the head moving upward and outward until the Violet Dot *rests on top of your head* at the Crown Chakra.

2. Call to mind the **Hierophant Symbol** (Pale Silver Merkaba Star formed by combining 2 Merkaba Stars). Visualize the spinning Hierophant stationed 12" below the feet within your **Maharic Shield** and slowly **INHALE,** drawing the Hierophant up through the Central Body Current, up to meet the Violet Dot at the 7th **Crown Chakra** at the top of the head. When the *Hierophant connects with the Violet Dot* visualize the Violet Dot bursting into a **4" Vertical Pillar of Violet-Silver Light**. Imagine the *Violet-Silver Pillar* running down through the body into the *Earth Core* and upward to the **14th Chakra** 36" above the Head. You have activated **the Violet Ray of Transmutation.**

3. Move your attention to the **4th Heart Chakra** and imagine a *small Green Spherical Crystal* – the **Green Star**- stationed at its core. This is the *Green Star Crystal Seal* that controls release of *4th Dimensional* frequency – *Phase-1 Doradic Current* - into the body. **INHALE,** drawing Violet-Silver energy from the Violet-Silver Pillar into the *Green Star.* **EXHALE,** *expanding* the Violet-Silver energy *into the Green Star sphere. Repeat 2 more INHALE-EXHALE* breaths, drawing Violet-Silver Energy from the Pillar and expanding it into the Green Star sphere. At the end of the last **EXHALE,** imagine the **Green Star popping,** turning into a spinning Green and Violet Starburst of Light, as the *Violet-Silver Pillar disappears. Spin the Green-Violet Starburst* faster until it becomes a *ball of Green-Violet Light,* about the size of a *grapefruit,* spinning in the center of the *4th Heart Chakra.* The Green-Violet ball is the store-house for your Phase-1Doradic Current – it is called the **Doradic Sphere. INHALE** and draw Pale Silver energy from your **Maharic Shield,** through the Central Body Current and into the center of the **Doradic Sphere** at the **4th Heart Chakra,** encasing the Sphere in translucent Pale Silver Light.

4. **EXHALE quickly** and push **2 Cords** of *Green-Violet-Silver energy* – **Doradic Cords** -from the Doradic Sphere *down the inside of your arms,* one Cord through each arm. As the ends of the Doradic Cords reach the inside of the hands, notice a sensation of *mild resistance* as the Cords push against the *etheric membranes* covering the *small chakras* in the palms. **INHALE** a full breath, drawing Silver energy up from the **Maharic Shield,** into the *Doradic Sphere* at the 4th Chakra. **EXHALE forcefully,** pushing more energy down through the *Doradic Cords* in the arms, and feel this energy **"pop"** out of the *palms, releasing the membranes* on the palm chakras. Continue to **INHALE** energy up from the *Maharic Shield* into the *Doradic Sphere* and **EXHALE** it through the *Doradic Cords, building the strength of the Doradic Current* in both palms. Imagine the Doradic Current *pooling in the palms* to form *2 Green-Violet balls of light,* about the size of *tennis balls.* Energize the Doradic Sphere and Cords until you can sense the smaller **Doradic Balls** formed in the *palms* of your hands. **Your *Doradic Current* is now activated for use in healing facilitation.**

SYMBOL SHORTHAND

TECHNIQUE # 3: Opening the Healing Channels
Doradic-Phase-1 Current

1. **4" Violet Dot** inside forehead, move to **Pineal Gland** then up through **7th Crown Chakra** to **top of head.**
2. **Hierophant Symbol** at center of your **Maharic Shield** 12" below feet. Inhale Symbol **up** to **Violet Dot** at top of head. See **Violet Dot burst** into **4" Violet-Silver Pillar** running through body to **Earth's core** and up to **14th Chakra** 36" above head.
3. Small **Green Star Crystal Seal at 4th Heart Chakra.** Inhale drawing **Violet-Silver** energy from **Violet-Silver Pillar** into **Green Star Crystal Seal** and see **Green Star Crystal Popping** into a spinning **Green-Violet Starburst,** as Violet-Silver Pillar **disappears.** Spin Green-Violet Starburst faster until it becomes grapefruit-size **BALL** of **Green-Violet Light**, the **Doradic Sphere,** at the center of the **4th Heart Chakra.** Encase Green-Violet Sphere in **Pale Silver** Maharic energy.
4. Push a **Green-Violet-Silver Doradic Cord** down each arm to **Palm Chakra Membranes** and draw **Pale Silver** up from **Maharic Shield** to **Doradic Sphere** and forcefully push Maharic Current **down arms** to **pop Palm Chakra Membranes.** While running **Doradic-1 Current** draw energy from **Maharic Shield** to **Doradic Sphere** and push it down arms from **Doradic Sphere** through **Doradic Cords** to **pool** into **Doradic Balls** at **Palm Chakras.**

TECHNIQUE # 3: Opening the Healing Channels
Doradic-Phase-1 Current

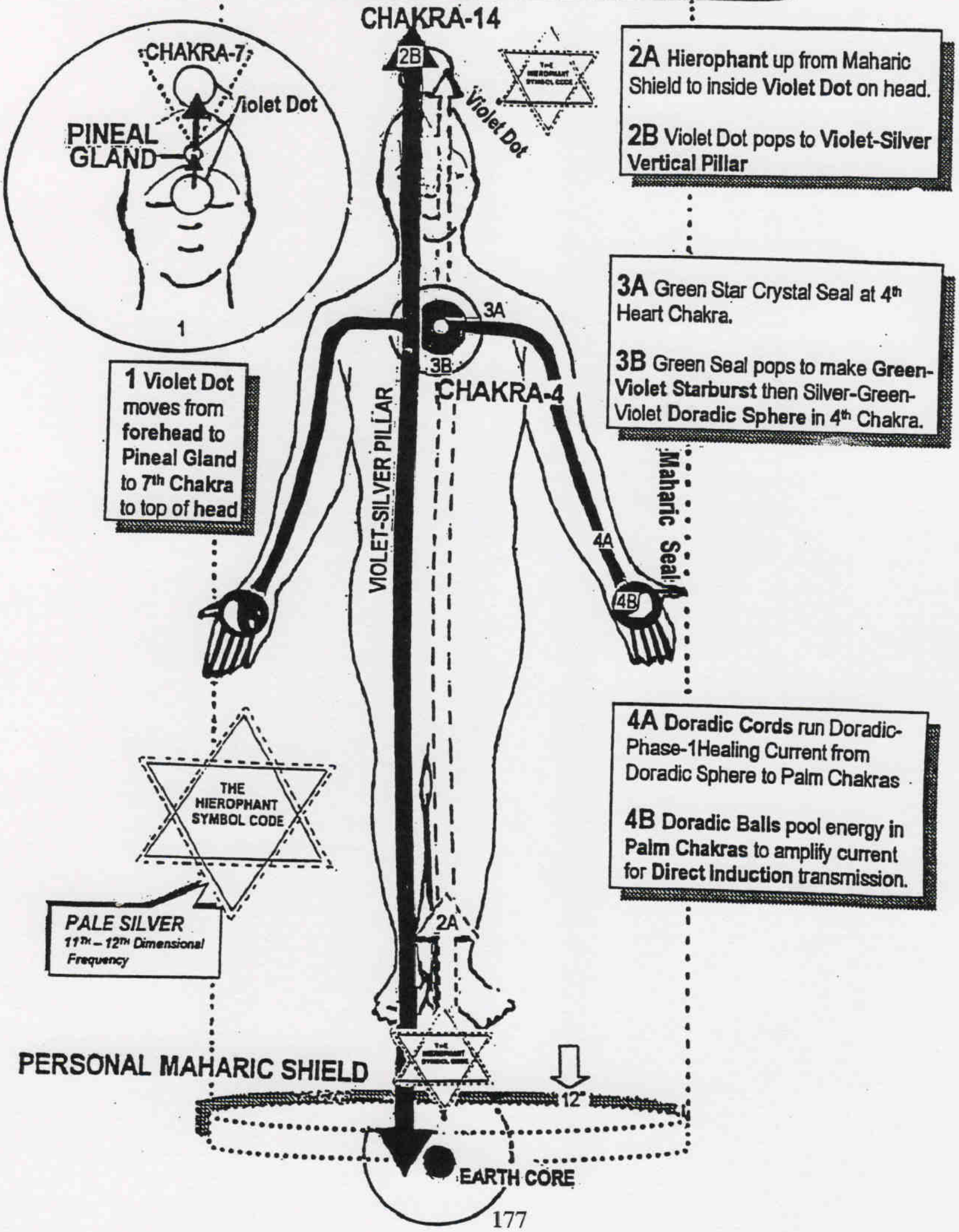
CHAKRA-14
CHAKRA-7
Violet Dot
PINEAL GLAND
1
2B
Violet Dot
THE HIEROPHANT SYMBOL CODE
3A
3B
CHAKRA-4
VIOLET-SILVER PILLAR
Maharic Seal
4A
4B
THE HIEROPHANT SYMBOL CODE
2A
PERSONAL MAHARIC SHIELD
12"
EARTH CORE

1 **Violet Dot** moves from **forehead** to **Pineal Gland** to **7th Chakra** to **top of head**

2A **Hierophant** up from Maharic Shield to inside **Violet Dot** on head.

2B Violet Dot pops to **Violet-Silver Vertical Pillar**

3A Green Star Crystal Seal at 4th Heart Chakra.

3B Green Seal pops to make **Green-Violet Starburst** then Silver-Green-Violet **Doradic Sphere** in 4th Chakra.

4A **Doradic Cords** run Doradic-Phase-1Healing Current from Doradic Sphere to Palm Chakras

4B **Doradic Balls** pool energy in **Palm Chakras** to amplify current for **Direct Induction** transmission.

PALE SILVER
11TH – 12TH Dimensional Frequency

THE NATURE OF KARMA

Karma, Identity, the Christos and the Physics of Divine Right Order

When an identity enters the **Time Matrix** to begin **simultaneous cycles of evolution** through time, it enters the manifest fields through first adopting a **Manifestation Template** composed of Partiki Units and Scalar Waves. The first Manifestation Template is called the **Rishic Shield**; it exists at the **Ante-matter** scalar-wave level of **Density-5** and holds the complete **Primal Scalar Pattern** for all of the **Singular Identities** that will **simultaneously** manifest into incarnation within the **lower dimensional** Densities and Time Cycles. The **Rishic Shield** represents the **Identity Gestalt** or **Family of Consciousness** through which identity enters manifestation. The Rishic Shield carries the precise configuration of Scalar-wave patterns, the **Primal Order** of energetic substance, through which the consciousness individuated to form a collective of consciousness within the dimensionalized Time Matrix. Every individual incarnate from the Rishic Family of Consciousness in time carries as part of its core **Kathara Grid** scalar program the **electromagnetic energy signature** of **Primal Order** carried by the **Rishic Shield** through which it manifested.

The Scalar-wave design of the Rishic Shield represents **DIVINE RIGHT ORDER**, the **Primal Order** of scalar-wave energetic relationship and **Partiki Phasing Sequences** that individual identities in time must re-evolve to hold within their embodied **Manifestation Templates** in order to **regain freedom** from the dimensional systems of the Time Matrix. Consciousness from the Rishic Shield, in the form of **Omni-polar wave spectra**, enters the manifest Densities by projecting portions of itself into **smaller Consciousness Gestalts** within the lower dimensional fields. The smaller Gestalts of Consciousness form the **3-Dimensional Scalar Shields** of the identity within the dimensionalized matter systems. **Each Scalar Shield holds the consciousness of numerous simultaneous Cycles of Incarnation within each of the 4 Primary Time cycles of Densities 1 through 4.** Each individual incarnate in time is directly linked to its original **Rishic Shield Family of Consciousness** through the **Density-4 Maharic Shield.** The Maharic Shield of dimensions **10-12** contains the first **Pre-matter Hydroplasmic scalar-wave spectra** through which identity began its experience within the holograms of Space, Time and Matter. The **personal Maharic Shield,** often called the **PERSONAL CHRISTOS IMPRINT**, holds the original **Primal Order** of scalar waves through which manifest consciousness links to its **Rishic Shield.** The **Christos Imprint** of the **Maharic Shield** represents the **PATH OF DIVINE RIGHT ORDER** through which each individual consciousness in time can **re-evolve** to its original Eternal Non-manifest State of Wholeness in Consciousness beyond the Time Matrix. **Divine Right Order is not only a spiritual concept; it is a reality in energy within the scientific dynamics of Universal Unified Field Physics.** Divine Right Order represents the **Primal Order** of scalar-waves that make up the original **Manifestation Template** from which individual **singular lifetimes** emerge, and it is the necessary pattern of energetic order within the personal multi-dimensional Manifestation Template which allows incarnate beings to fulfill the **evolutionary objective** of **Transmutational Dimensional Ascension. The result of an incarnate being's *Manifestation Template* and Conscious Awareness becoming mis-aligned with the energetic structure of *Divine Right Order* is called KARMA.**

Karma, Divine Right Order, the Scalar Shields, Free Will and Miasms

If it is understood that the universe and all things within it exist as **manifestations of consciousness** in the form of **precisely ordered** scalar wave arrangements, and that there is an intrinsic and indelible **Primal Order** upon which all manifestation functions, it will not be difficult to conceptualize the **literal and tangible nature** of the **Karmic Imprint**. In energetic terms, **Divine Right Order** represents the **natural alignment** of conscious energy fields within the intrinsic **Laws of Universal Physics** characteristic to the dimensional structures of time. Divine Right Order exists as a **tangible, specific** and **harmonious organization** of **energy interrelationships** within the Primal Order of energetic structure that creates dimensionality. The condition frequently referred to as **Karma** is, like Divine Right Order, a tangible, specific organization of energy interrelationships. Unlike Divine Right Order, *Karma* represents dis-harmonious interrelationships of energy within the context of the Primal Order that creates dimensionality. **Karma** can be viewed as **CHAOTIC, incoherent energetic disorganization**, whereas **Order** represents **coherent energetic** organization.

In terms of the **15-Dimensional anatomy** of the **personal Manifestation Template**, the **Karmic Pattern** represents **areas of incoherent energetic disorganization** within the structure of the personal **Scalar Shields**. The **12th–Dimensional** Level of the personal **Maharic Shield** holds the original **Christos Imprint**, the **Primal Order** *perfect organization* of scalar-wave configurations that links our consciousness and **DNA Template** to the **Rishic Shield Family of Consciousness** through which we entered the **Time Matrix**. While the Maharic Shield holds our original **Perfect Organization of Consciousness**, the Scalar Shields upon which our lower dimensional **Hova Bodies and Incarnate, Soul and Over-Soul Manifestation Cycles** emerge often become **distorted** from their original **Primal Order** as held within the **Maharic** and **Rishic Shields**. Human consciousness, in its multidimensional forms, was designed **"in the image of its Creator"**, having **Free Will** to *apply the dynamics of energetic creation in time* in accordance with **personal intention,** within the greater context of the fixed **Natural Energetic Laws of Creation.**

When we use our gift of **Free Will** to direct our energies in a manner **congruent** with the **Divine Right Order** of **Universal Unified Field Physics** (which implies co-creative, not competitive co-evolution for all), the scalar-wave design of our personal **Scalar Shields** remains organized upon the intrinsic design of universal and personal **Primal Order**. Through **embodiment** of Divine Right Order within our personal **Scalar Shields** and consciousness, we know conditions of **perfect health and harmony** within the Time Matrix. When our Free Will is used to direct our energies in ways that are **incongruent** with and non-supportive of **Primal Order**, the **disorganization** of our energetic orientation manifests as literal **Partiki Phasing Sequence distortions** that create **misshapen scalar-wave groupings** within the anatomy of the personal **Kathara Grid**, **Scalar Shields** and the **DNA Template** built upon them. These misshapen scalar-wave groupings within the Kathara Grid, Scalar Shields and DNA Template are called **MIASMS** and they hold, in **tangible scalar frequency** form, the **disorganized energetic patterns** of our **Free Will choices** through which we have used our energies in a manner **incongruent** with **Divine Right Order**.

Miasms* are the *tangible substance* of our *Karmic Pattern* that reflect the areas in which our *consciousness* and *Scalar Shields* are *distorted* from our original *Maharic Shield Christos Imprint* of *Perfect Organization* within the context of *Universal Primal Order.

TIME VECTORS AND THE KARMIC-MIASMIC IMPRINT

Miasms, Incarnational Time Vectors, Vector Imbedding and the DNA Template

The **Miasmic Imprint** within the body and Manifestation Template that holds the personal **Karmic pattern** manifests as **mathematical, geometrical** and **electro-magnetic distortions** in scalar-wave forms within the Kathara Grid and Scalar Shields, that cause groupings of **Anti-particles** to build up within the **Particle field** of the physical and energetic anatomy. The build-up of **Anti-particles** within the Particle Field blocks the natural flow of energy and consciousness through the Kathara Level-3 **Diodic Points Grid**, and manifests as **malfunctions** within the **Partiki Phasing Programs** within the **Kathara Centers** of the Level-1 **12-Tree Grid** and within the **Signet Star Crystal Seals** of the Kathara Level-2 **Crystal Seals Grid. The 3 Levels of the Kathara Grid** set the blueprint for manifestation of the **Scalar Shields, Hova Bodies, Chakra System** and **DNA Template. Miasmic distortions** in the **Kathara Grid** effect **every level of personal manifestation**, from the operation of the **physical genetic code** to the ability of consciousness to successfully embody within the physical form, to the potentials of **Transmutational Dimensional Ascension** and attainment of **Spiritual Actualization.**

Karma*- the *Miasmic Imprint*- is a *biophysical phenomenon* of scalar-wave distortion that directly effects the *human DNA.

The **Incarnations of SELF** that **simultaneously manifest** into incarnational cycles of divergent space-time placement within the 4 Densities and 12-Dimensions of Matter are all **directly interconnect** through the **Kathara Grid, Scalar Shields** and **DNA TEMPLATE.** The **Free Will choices** made by other manifestations of ourselves **concurrently existing** in other space-time locations and different manifest body forms, effect the function of our personal **Kathara Grid, DNA Template, conscious awareness** and **life experience.** The **choices WE make** in our present lifetime directly effect the **Kathara Grid, DNA Template, consciousness** and **life experience** of **our other-time incarnations.** The energetic connection between incarnations of the SELF is an immediate, **living connection** of energy and consciousness. Because our numerous simultaneous incarnations in time are directly, energetically linked to each other via the Kathara Grid and DNA Template, we literally **inherit Miasmic Buildup from each other** on a continual basis. When we make the choice to restore our personal **Imprint for Health,** our **Christos Imprint,** as held within our **Maharic Shield,** part of these efforts must involve **clearing the Miasmic Imprint** that is bleeding through into our personal Manifestation Template for the other incarnations of Self.

Each of our simultaneous incarnations in time represents one **SINGLE VECTOR CONSCIOUSNESS,** one portion of our **Eternal Identity** that experiences 3-Dimensional **linear evolutionary progression** along **ONE TIME VECTOR** (one cycle of Partiki Phasing Sequences). **Each Incarnate represents a TIME VECTOR.** The multiple Time Vectors of our Multiple Simultaneous Selves are literally **encoded** within core scalar-wave programs of our **DNA Template;** this is called **VECTOR IMBEDDING.** To fully heal our Body-Mind-Spirit system by restoring it to its **Organic Imprint for Health,** the **Maharic Shield Christos Imprint,** we need to clear the Miasmic Imprint from our other simultaneous incarnations as it is presently embedded within our DNA Template. **Transmuting the Miasmic Imprint that is embedded in our DNA Template to restore the original perfect Primal Order is called "Clearing the Karmic Pattern" or MULTI-VECTOR HOLOGRAPHIC RECODING.**

MULTI-VECTOR HOLOGRAPHIC RECODING

REGENESIS OF THE PERSONAL CHRISTOS

Restoring the Organic Imprint for Health and Clearing the Karmic-Miasmic Imprint

Literally all conditions of **dis-ease** in the human body and dis-harmony in the life experience, including the unnatural condition of **death** of the human body form, result from **Miasmic Distortions** within the embodied **Kathara Grid, Scalar Shields, DNA Template** and personal **Inner Templar Complex.** The natural state of **Universal Primal Order** is harmonious co-creative interrelationship of consciousness. Humans can progressively return to their natural condition of **Ordered Perfection** through **re-genesis**, or *re-creation*, of the **Organic Imprint for Health – the Maharic Shield Christos Imprint** – within the Kathara Grid blueprint of the personal Manifestation Template. **Regenesis of the Personal Christos** is a **Multi-Vector Endeavor**, involving clearing the karmic-Miasmic Imprint from a myriad of various simultaneous selves presently evolving in the Time Matrix. The **hard way** to clear the Miasmic Imprint is by **"walking the Karma"**, which means allowing progressive levels of the **Miasmic Imprint** to repeatedly **re-manifest** in the personal life experience, until the **proper application of Free Will choice** used in congruence with **Divine Right Order** becomes consciously apparent and the **"Right Choices"** are made to re-ordered the disorganized use of energy that caused the Miasm to form. In our present time, this approach to "clearing the Karmic Pattern" is no longer effective as a learning tool by which embodied consciousness teaches itself the etiquette of creation. **The Miasmic Imprint within the collective human species has grown so extensively that the core issues of misapplied Free Will that cause Miasmic Buildup can no longer be consciously identified, and the Miasmic Imprint continues to compound and amplify itself as the same errors of Free Will energy use are perpetually repeated and reinforced.**

The Karmic-Miasmic Imprint is an **ENERGY FORM**; a literal **construct of scalar-wave frequency** with its corresponding **Anti-particle Manifestation.** As a form of energy, the **Miasmic Imprint** can be dealt with **AS ENERGY, WITH ENERGY**, for the most expedient clearing, healing and realignment of the **Karmic Imprint.**

The **Karmic-Miasmic Imprint** manifests within the **causal level** of manifestation, the **3 Levels of the embodied Kathara Grid** of the **personal Internal Template Complex**, and from the Kathara Level the Karmic-Miasmic Imprint manifests as **misalignment** and **malfunction** within all levels of Body-Mind-Spirit system. To clear the Karmic Imprint **AS ENERGY, WITH ENERGY,** its disorganized scalar-frequency patterns must be realigned from the **Core Template Kathara Grids**, the **causal level** of miasmic manifestation. Restoring the **Maharic Shield Christos Imprint** to the personal **Kathara Grid** takes time and **precise applications of frequency re-ordering** in all levels of the Body-Mind-Spirit system. The first step to **Regenesis of the Personal Christos** is the **MAHARIC RECODING PROCESS.**

MULTI-VECTOR HOLOGRAPHIC RECODING

Recoding the Time Vectors of the DNA Template and the DNA Fire Codes

The process of regenerating the **Christos Imprint** within the **Kathara Grid** of the human body is a **MULTI-VECTOR** undertaking. A **Vector** is a **set cycle of vibrational-oscillation rhythm** that creates a **specific repeating Partiki Phasing Rhythm** and **Flash Line Sequence.** Each individual incarnate self in time holds within the electro-magnetic programs of its embodied **Kathara Centers** the precise **Partiki Phasing Rhythms** that correspond to its particular placement within the planetary **Time Cycles.** The **Time Vector,** or **Flash Line Sequence** that creates one linear **Event Horizon,** is literally encoded and actively operational within the **Kathara Grid** and **DNA Template** of each incarnate identity. In the process of restoring the **Organic Integrity** to the personal Manifestation Template, the dormant **Partiki Phasing Rhythms** that correspond to one's **other-time incarnates** are progressively **brought into activation** and **realigned** with the Christos Imprint of the **Maharic Shield.**

The **Partiki Phasing Flash Line Sequence programs** within the embodied **Kathara Centers** manifest within the **DNA Template** as the **GENETIC TIME CODES,** or **Fire Codes,** of the DNA. In restoring **Maharic Order** to the personal **Kathara Grid,** the **12 Dormant Fire Codes** of the DNA Template will be slowly but progressively brought into **activation** through the 11th-12th –Dimensional Frequencies of the **Maharic Shield Mahara Hova Body Current.** Activation of the **DNA Fire Codes** creates a reciprocal release of the **Signet Star Crystal Seals** and **Cranial-Sacral Kundalini Seals** within the **Kathara Level-2 Crystal Seals Grid,** progressive activation of the dormant **Kundalini energies** and **Merkaba Vehicle Phases,** merger of the **Hova Bodies** and embodiment of the higher dimensional **Spiritual aspects** of conscious awareness. The process of recoding the multiple time vectors within the DNA Template – **Multi-vector Holographic Recoding-** is a process of progressive **WHOLE-BEING HEALING** that is achieved through **consistent application** of the **MAHARIC RECODING PROCESS,** and expedited through complimentary processes of **DNA Template, Kundalini** and **Merkaba Vehicle** activation. Level-1 Kathara Healing introduces the operational mechanics of the **Maharic Recoding Process.**

The Maharic Recoding Process

FOR MULTI-VECTOR HOLOGRAPHIC RECODING

The 4 Primary Elements of Action

1. **The Maharic Seal and the Liquid Light Cleanse** (Technique # 2)
2. **Opening the Healing Channels** (Technique # 3)
3. **Amplifying Inner Audio-Visual and Kathara Scan** (Technique # 4)
4. **The Maharic Infusion** (Technique # 5)

The **4 Primary Elements of Action** to begin the **Maharic Recoding Process,** through which **Multi-Vector Holographic Recoding** of the **Morphogenetic Field Manifestation Template** is achieved. Begins realignment of the Core Manifestation Template **Kathara Grid Levels,** the **Scalar Shields, Hova Bodies, Axi-A-Tonal Lines** and **DNA Template** with the **Primal Order** perfect scalar-wave pattern of the personal 12th-Dimensional Pre-matter Hydroplasmic **Maharic Shield-** the **Christos Imprint** and **Organic Imprint for Health.** Also begins activation of dormant **DNA Strand Templates, Kundalini** energies and the **Merkaba Vehicle** for **Transmutational Dimensional Ascension.**

Bio-Regenesis of the Christos Imprint
and 12-Strand DNA Template
within the human body.

TECHNIQUE # 4

AMPLIFYING INNER AUDIO-VISUAL AND KATHARA SCAN

1. Place **RIGHT palm** (*electrical-transmitting hand*) over the **5th Kathara Center** at the **Navel**, and place the *Left palm* (*magnetic-receiving hand*) over the *back of the right hand* (to block 3rd-Dimensional energies from entering the rear of the right hand chakra). **INHALE** energy from your **Maharic Shield** – the *Maharic Carrier Wave*- into your **Doradic Sphere** at the *4th Heart Chakra*. **EXHALE** and run the *Maharic Wave* down the *Doradic Cords* in the arms and into the *Doradic Balls* in the palms. Inhale, then **EXHALE**, using the exhale breath to *push* energy from the *Doradic Sphere*, down the *Doradic Cord* in the **RIGHT** arm, through the **RIGHT Doradic Ball** and into the **5th Kathara Center** at the *Navel*. Run several breaths of *Doradic-1 Current* into the *5th Kathara Center* at the Navel, until a **mild tingling sensation** is felt.

2. Place the **LEFT palm** over the **5th Chakra** at the *Throat*, with the *Right palm* covering the *back of the LEFT hand*, and **INHALE** the Doradic-1 Current up from the *5th Kathara Center* at the Navel, into the Seed/Center of the **5th Chakra**. **EXHALE** and *expand* the Doradic-1 frequency into the *5th Chakra spiral*, allowing the magnetic current from the LEFT hand to draw the energy out through the spiral. As you **EXHALE**, slowly **move the hands forward**, away from the body and the 5th Chakra.

3. Repeat the same process over the **6th Kathara Center**, placing the **RIGHT palm** over the *6th Kathara Center* below the **Right Shoulder**, with the *Left palm* covering the back of the *RIGHT hand*. **INHALE**/Draw energy up from the *Maharic Shield*, into the *Doradic Sphere*. **EXHALE** energy down the *Doradic Cords* into the *Doradic Balls*. Inhale, then **EXHALE** energy down the *RIGHT Doradic Cord* through the *RIGHT palm* and into **Kathara Center 6.** Run several breaths to energize the *6th Kathara Center*. Then move **LEFT palm** over the *6th Third Eye Chakra* in the forehead, with Right palm behind. **INHALE** the Doradic-1 Current from the *6th Kathara Center* into the Seed/Center of the *6th Chakra* and **EXHALE** the energy into the 6th Chakra spiral while drawing it through with the magnetic current from the LEFT hand. The **6th Sense-*Audurea***, which provides **inner audio** perception, and the **7th Sense-Tristet**, which provides **inner visual** perception are now energized and amplified in preparation for **Kathara Scan.**

4. Close your eyes and focus your attention in a **single point** at the *inside of your forehead*. Inhale, and as you **EXHALE** run your attention outward ***through your 6th Chakra spiral***, into the **6th Chakra Spiral of your Client**. Move your attention into the center of your Client's brain, the **Pineal Gland.** Inhale, and **EXHALE** moving your awareness down from the Client's Pineal, through your *Client's Central Body Current* and down into the **Earth's core. INHALE** and move your awareness back upward to your **Client's Maharic Shield**, and spread your awareness outward through the *Maharic Shield* with the **EXHALE.** On the *inside of your forehead*, imagine the image of the **Kathara 12-Tree Grid, INHALE** energy from the *Client's Maharic Shield* and **EXHALE** the energy out into the *12-Tree image*. Breathe gently for a moment, *observing the 12-Tree Grid*, which will register tones of white, black and ranges of gray in between. If you cannot yet visualize, you will still get a sense of the 12-Tree image and the areas of dark and light within it. The dark areas represent layers of the **MIASMIC BODY**, areas of the *karmic imprint* that need clearing when you run *Technique # 4: The Maharic Infusion*. Remember the areas of the Client's Kathara that need clearing. Inhale, and *open your eyes* on **EXHALE.**

SYMBOL SHORTHAND

TECHNIQUE # 4: Amplifying Inner Audio-Visual and Kathara Scan

Activate Maharic Seal and Open Healing Channels. (Techniques # 2 and # 3 first)

1. **RIGHT Palm** on **5th Kathara Center** in **Navel.** Breathe **Maharic Current** up from personal **Maharic Shield** and into **Doradic Sphere** at **4th Chakra.** Push **Maharic-Doradic-1 Current** down **Doradic Cords** and into **Doradic Balls** at **Palm Chakras.** Push Current from Doradic Ball in **RIGHT Palm** into **Kathara Center # 5** in **Navel.**

2. **LEFT Palm** on **5th Throat Chakra**, cover back with **Right Palm** (Reverse Standard Hand Position). Inhale **Doradic-1 Current** up from **5th Kathara Center** in **Navel** to center of **5th Throat Chakra** and expand Current into **5th Chakra Spiral**, with **Left Palm Chakra** drawing current out of **5th Chakra Spiral.**

3. **RIGHT Palm** on **Kathara Center # 6** below **Right Shoulder**, cover back with **Left Palm.** Push **Maharic-Doradic-1 Current** from Doradic Ball in **RIGHT Palm** into **Kathara Center # 6. LEFT Palm** over **6th 3rd Eye Chakra**, cover back with **Right Palm** (Reverse Standard Hand Position). Inhale **Doradic-1 Current** up from **Kathara Center # 6** to center of **6th- 3rd Eye Chakra** and expand Current into **6th Chakra Spiral**, with **LEFT Palm Chakra** drawing Current out of **6th Chakra Spiral.**

4. **KATHARA CLIENT SCAN: A.** Place **Point of Attention** inside **forehead** and run attention out of **6th Chakra Spiral** and into the **6th Chakra Spiral of Client**, then into **Client's Pineal Gland. B.** Run attention down **Client's Central Vertical Current** and into **Earth's Core,** then back up to **Client's Maharic Shield.** Expand attention out into **Client's Maharic Shield. C.** Visualize image of **Kathara 12-Tree Grid,** inhale energy from **Client's Maharic Shield** and exhale it into image of **Kathara Grid. D.** Observe for **Dark Miasmic Buildups** in image of **Kathara Grid;** run extra healing Current into these areas of the **Client's Kathara Grid** during the **Maharic Infusion** (Technique # 5).

Personal Kathara Scan: On **Step 4** run attention down **your Central Vertical Current** then back up to **your Maharic Shield** and out into **your Maharic Shield.** Breathe energy from your Maharic Shield into image of **Kathara 12-Tree Grid** and observe for **Dark Miasmic Buildup** areas to clear during the ***Maharic Seal and Liquid Light Cleanse*** technique. Send **extra Maharic Current** to these areas during ***Liquid Light Cleanse.***

TECHNIQUE # 4: Amplifying Inner Audio-Visual and Kathara Scan

1. A. RIGHT Palm on 5th Kathara Center in Navel.
B. Breathe Maharic Current up from personal Maharic Shield to Doradic Sphere and down Doradic Cords to Palm Chakras.
C. Push Current from RIGHT Doradic Ball into 5th Kathara Center.

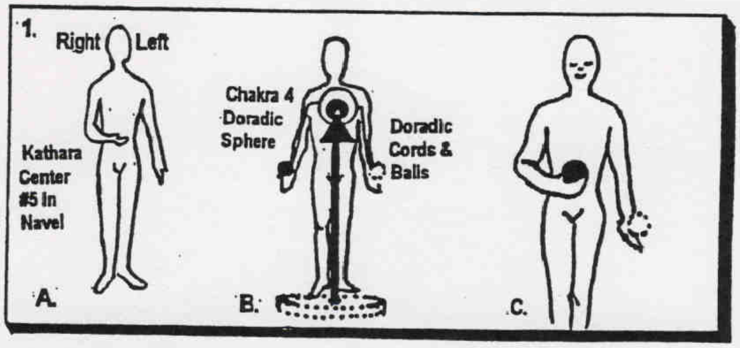
1.
Right
Left
Kathara Center #5 In Navel
Chakra 4 Doradic Sphere
Doradic Cords & Balls
A.
B.
C.

2. A. LEFT Palm on 5th Throat Chakra, cover back with Right Palm.
. INHALE Doradic-1 Current up from Kathara Center # 5-Navel to 5th Chakra and expand Current into 5th Chakra Spiral.

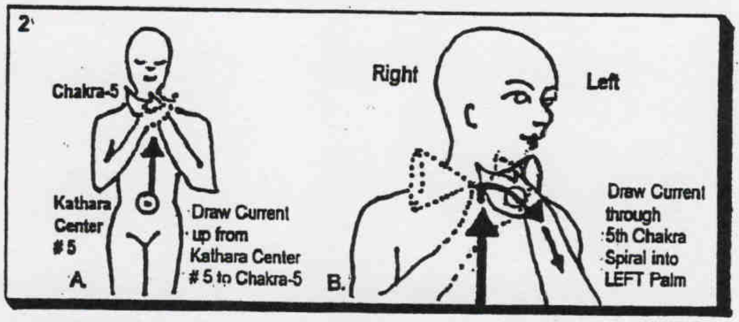
2
Chakra-5
Kathara Center #5
Draw Current up from Kathara Center # 5 to Chakra-5
A
Right
Left
Draw Current through 5th Chakra Spiral into LEFT Palm
B.

3. A. RIGHT Palm on Kathara Center #6 in Right Shoulder, cover back with Left Palm. B. LEFT Palm on 6th 3rd-Eye Chakra, INHALE Current from Kathara Center # 6 into 6th Chakra Spiral and into LEFT Palm. C. Kathara Scan: a. Run your Attention from your 6th Chakra to Client's 6th Chakra and Pineal Gland. b. Run Attention down Client's Central Vertical Current to Earth's Core & into Client's Maharic Shield. c. EXHALE energy from Client's Maharic Shield into Image of Kathara 12-Tree Grid. d. Scan grid for Dark Areas of Miasmic Buildup for clearing with Maharic Infusion.

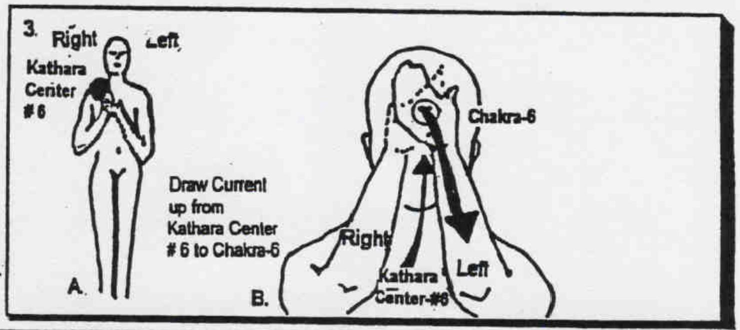
3.
Right
Left
Kathara Center #6
Draw Current up from Kathara Center # 6 to Chakra-6
A.
Chakra-6
Right
Kathara Center-#6
Left
B.

KATHARA SCAN

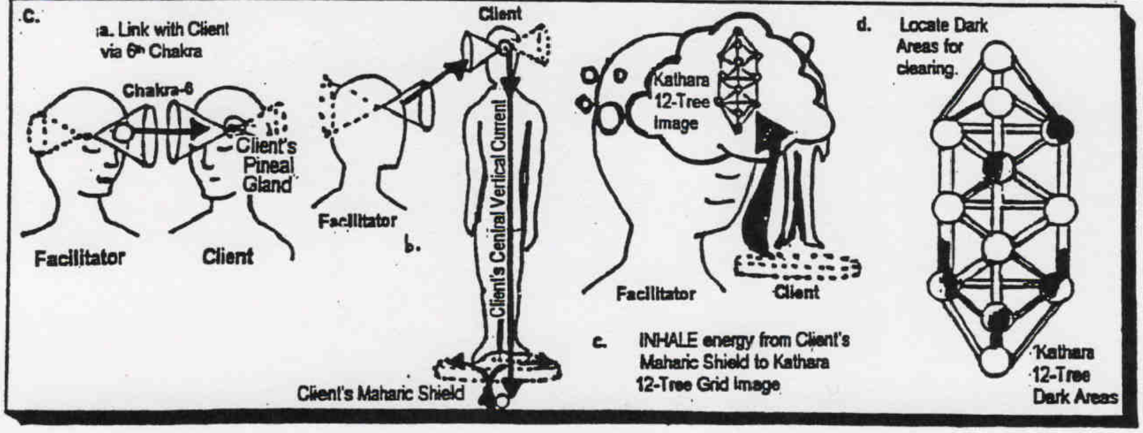
c.
a. Link with Client via 6th Chakra
Chakra-6
Client's Pineal Gland
Facilitator
Client
Client
Facilitator
b.
Client's Central Vertical Current
Client's Maharic Shield
Kathara 12-Tree Image
Facilitator
Client
c.
INHALE energy from Client's Maharic Shield to Kathara 12-Tree Grid Image
d.
Locate Dark Areas for clearing.
Kathara 12-Tree Dark Areas

TECHNIQUE # 5

THE MAHARIC INFUSION – CLIENT FACILITATION

Prior to using this technique you should have completed the following: 1.Positioned *MENTOR* over your Right shoulder to assist (*Tech. # 1:Awakening the Mentor and the 6th and 7th Senses*). 2. Activated your *Maharic Shield, Maharic Seal* and run the *Liquid Light.* (*Tech. # 2: The Maharic Seal and the Liquid Light Cleanse.*) 3.Activated your *Doradic Current, Inner audio/Visual* and *Client Kathara Scan* (*Tech. # 3. and #4*)
4. Opened the *Kathara Lines* for you and your Client (*Sequence-1 Kathara Points*)

Place Client in a reclining position with shoes off:

1. Position yourself at the **feet** of the reclining Client. Place **RIGHT palm** *12" below Client's feet* at the **Maharic Shield,** *left palm* over the back of Right hand.

2. Visualize the **Hierophant symbol** on the *inside of your forehead.* Inhale, then **EXHALE,** moving the *Hierophant* down to **Chakra 4** and into the **Doradic Sphere** with the exhale breath. **INHALE** energy upward from *your* **Maharic Shield** and into the *Doradic Sphere*. **EXHALE** the **Hierophant** with *Doradic-1 current* down the **Right Doradic Cord** and into the *Doradic Ball* in the **Right palm.** .

3. Inhale, then **EXHALE** the *Hierophant and Doradic-1 current* into the *center* of the **Client's Maharic Shield.** Visualize the *Hierophant spinning* in the center of the Client's Maharic Shield until you sense a **mild shift** in the energy below the Client's feet. The shift indicates that the Client's *Maharic Shield* has activated and the *Maharic Pillar* has "popped" vertically upward from the Shield – the Client's bio-energetic field now carries the **Maharic Seal.**

4. **INHALE** energy from your *Maharic Shield* into your *Doradic Sphere*, then **EXHALE** the energy down **both Doradic Cords,** into the Doradic Balls in the palms. Visualize the **12-Tree Kathara Grid** *within the Client's body.* You will now run the **Maharic Infusion** through the *small chakra points in the feet.*

TECHNIQUE #5: Continued
THE *MAHARIC INFUSION*

5. Place your **Right palm** on the *bottom* of **Client's Left foot** and your **Left palm** on *top of the foot,* centered over the *Right palm.* **Imagine that you are lining up the *small vortex* beneath the ball of the Client's foot (which extends *through the foot* to come out on top), with the *small chakras* in your *palms*.** Inhale, then **EXHALE** energy from your **Doradic Sphere** through the **Right Doradic Cord, Right palm** and *into the **sole** of the Client's foot.* Visualize the **Green-Violet-Silver energy** traveling up the **Kathara Line on the** *Left side of the body,* through the *4th, 7th and 10th Kathara Centers.* Continue **EXHALING** energy into the **Left foot** until you begin to feel a bit of **fuzzy energy** coming out of the foot toward your palm. This indicates that the *Kathara Line* has *reached its energy-holding capacity.* **Repeat** the process on the **Right foot**, running energy through the *Kathara Line on the Right side of the body and the 3rd, 6th and 9th Kathara Centers.*

6. You will now clear and align the **Central Vertical Kathara Line** to complete the **Maharic Infusion. Move to Client's head.** Place **Right palm** over **Client's 10th Chakra** 6" above the head, the location of the *12th Kathara Center.* Place *Left palm* over back of Right hand. **INHALE** energy from your **Maharic Shield** into your **Doradic Sphere** then **EXHALE** energy down the **Right Doradic Cord,** into the **Right palm** and through the **10th Chakra.** Visualize the energy moving down from the *10th Chakra* into the **7th Chakra,** to the **Pineal Gland (11th Kathara Center),** then *rapidly* down the **Central Body Current** and **Central Kathara Line,** through *Kathara Centers 8* **(thyroid)**, *5* **(Navel)**, *2* **(at Base Chakra)** and into *Kathara Center 1* at **Earth's core.** Continue **EXHALING** energy into the **10th Chakra** and **Central Kathara Line** until you feel **fuzziness** against your *Right palm,* indicating that the *Kathara Grid* has reached its *energy-holding capacity.*

7. *Before releasing 10th Chakra* direct a bit of additional energy through the **Central Kathara Line** into areas that revealed **dark spots** during the *Kathara Scan.* Visualize the areas *getting light* as the energy runs through them. **You may add other healing modalities at this point, if desired.** *Close* by using the **Sequence-2 Kathara Points** to close the Kathara Lines and *revitalize the body.* Allow Client a few moments to recline before ending session.

THE MAHARIC INFUSION – CLIENT FACILITATION

Preparation: **A.** *Technique #1*- Place MENTOR over your Right Shoulder **B.** *Technique #2*-Activate your Maharic Shield and run Liquid Light Cleanse. **C.** *Technique #3*- Activate yourDoradic-1 Current **D.** *Technique #4* – Amplify Inner Audio-Visual and run Client Kathara Scan **E.** *Sequence-1 Kathara Points* massage on Self and Client to open Kathara Lines.

Place Client in reclining position with shoes off

1. Stand at Client's feet, place **RIGHT Palm** 12" below Client's feet at **Client's Maharic Shield,** left palm over back of Right hand.

2. See **Hierophant Symbol Code** inside your forehead, **EXHALE** Symbol down to **Chakra-4 Doradic Sphere. INHALE** energy up from your **Maharic Shield** into Doradic Sphere. **EXHALE** *Hierophant Symbol* with *Doradic-1 Current* down **RIGHT Doradic Cord** into **RIGHT Doradic Ball in RIGHT Palm.**

3. **EXHALE** Doradic-1 Current and Hierophant Symbol into Center of **Client's Maharic Shield** via **Right Palm Doradic Ball and spin Hierophant** until sensed **Energy Shift** of **Client's Maharic Seal.**

4. **INHALE** Current from your Maharic Shield to Doradic Sphere and **EXHALE** Current down both **Doradic Cords** into **Doradic Balls.** See **Kathara12-Tree Grid** within **Client's** body and begin running the ***Maharic Infusion*** through the **Secondary Chakras** in **Client's feet.**

5. **RIGHT PALM** on **bottom of Client's LEFT FOOT, LEFT PALM on top of Client's LEFT FOOT,** both **centered** over Client's LEFT FOOT **Secondary Chakra. INHALE** Maharic energy into Doradic Sphere and **EXHALE** Doradic-1 Current into Client's **LEFT KATHARA LINE,** and through **Kathara Centers # 4, #7** and **# 10. EXHALE** Doradic-1 Current into Client's **LEFT FOOT** until **"Fuzzy"** resistance emerges from Client's foot Chakra. **REPEAT PROCESS** on **Client's RIGHT FOOT,** running Doradic-1 Current through the **RIGHT KATHARA LINE** and **Kathara Centers # 3, #6** and **#9.**

6. **Stand at Client's Head. RIGHT PALM** over **Client's 10th Chakra** (6" above head- the 12th Kathara Center), LEFT Palm over back of Right hand. Fill Doradic Sphere with Maharic Energy then **EXHALE** Doradic-1 Current down **RIGHT DORADIC CORD** and Palm and into **Client's 10th Chakra.** Run Current from **Chakra-10,** to **Crown Chakra-7, Pineal Gland,** rapidly down **Central Vertical Current** and **Kathara Centers #8** (Thyroid), **#5** (Navel) **#2** (Chakra-1) and **#1** (Earth's Core Chakra 13). Continue exhaling Current until **"Fuzzy"** resistance emerges from Client's 10th Chakra.

7. **Direct additional energy** through Client's **CENTRAL KATHARA LINE** to clear **DARK AREAS** of **Kathara Grid** located in **Kathara Scan.** Add any additional healing technologies following Client's Maharic Infusion. Close with **SEQUENCE-1 KATHARA POINTS** massage. Allow Client to remain relaxed in reclining position for several minutes following Maharic Infusion.

TECHNIQUE # 5: THE MAHARIC INFUSION – CLIENT FACILITATION

Maharic Infusion 6-STEP Sequence

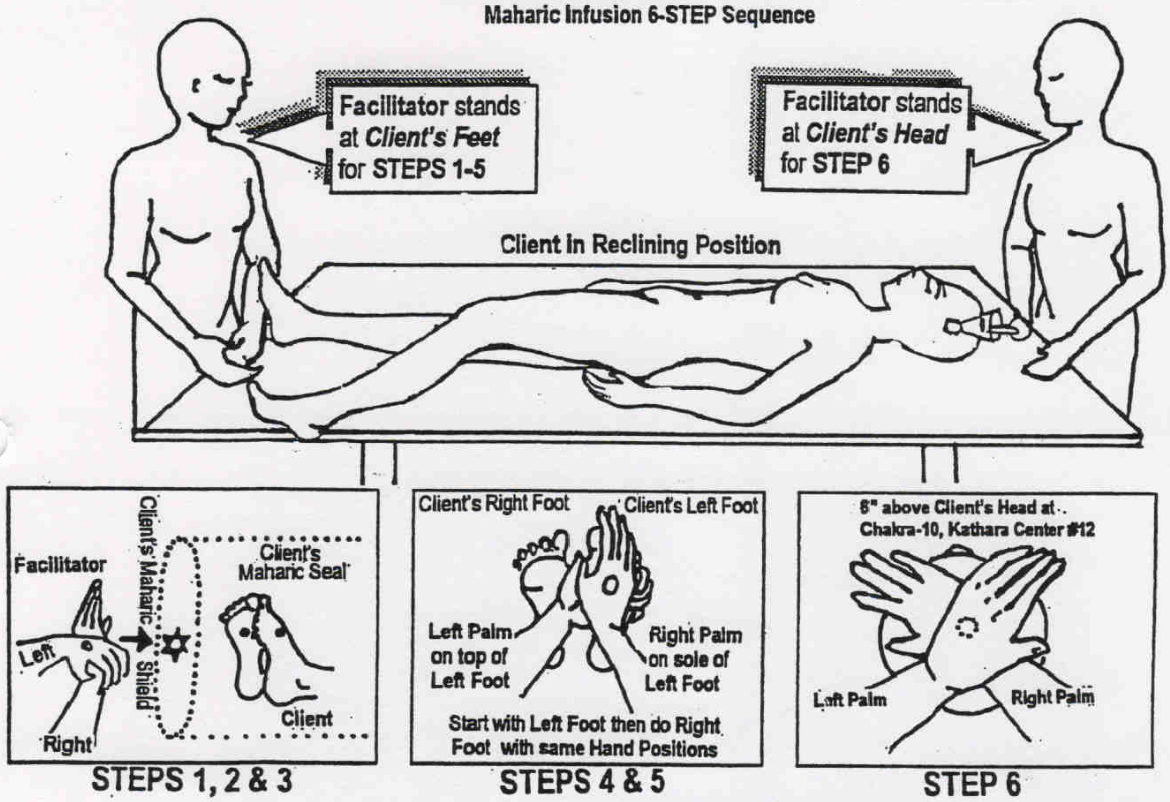

Facilitator stands at ***Client's Feet*** for STEPS 1-5

Facilitator stands at ***Client's Head*** for STEP 6

Client in Reclining Position

Facilitator
Left
Right
Client's Maharic Shield
Client's Maharic Seal
Client
Client's Right Foot
Client's Left Foot
Left Palm on top of Left Foot
Right Palm on sole of Left Foot
6" above Client's Head at Chakra-10, Kathara Center #12
Left Palm
Right Palm

Start with Left Foot then do Right Foot with same Hand Positions

STEPS 1, 2 & 3

STEPS 4 & 5

STEP 6

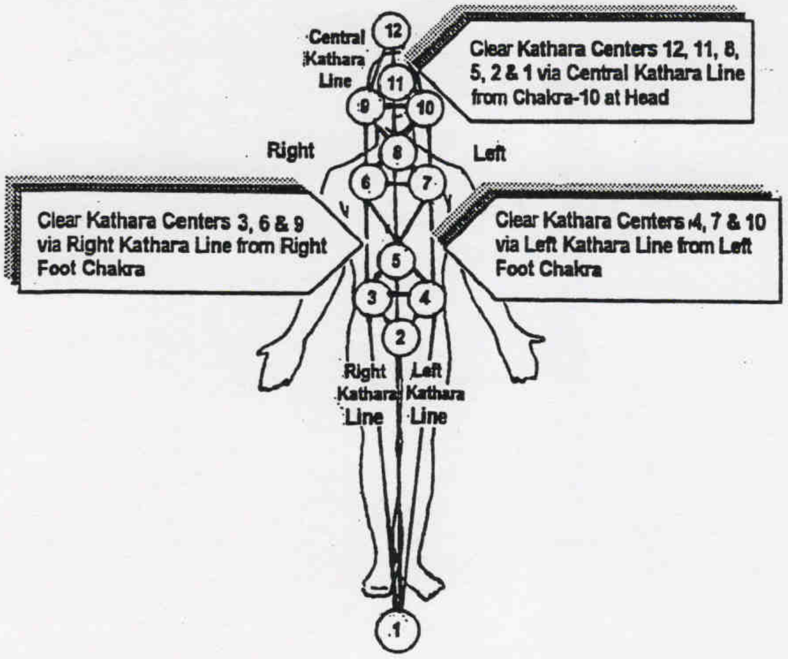

Clear Kathara Centers 12, 11, 8, 5, 2 & 1 via Central Kathara Line from Chakra-10 at Head

Clear Kathara Centers 3, 6 & 9 via Right Kathara Line from Right Foot Chakra

Clear Kathara Centers 4, 7 & 10 via Left Kathara Line from Left Foot Chakra

Central Kathara Line
12
11
9
10
Right
Left
8
6
7
5
3
4
2
Right Kathara Line
Left Kathara Line
1

MOTIVATION TO HEAL

Healing Facilitation is an art and a science, but most importantly, healing facilitation is an **interrelationship of SOULS.** Egotistical objectives have no place in quality Healing Facilitation. The decision to participate in Healing Facilitation should not be motivated by a desire to prove you have "special abilities" to boast about, nor with the primary motive of gaining **financial profit** from Healing Facilitation endeavors. The energies exchanged in healing facilitation are **FREE TO ALL**, a **DIVINE GIFT** from the universe, and they belong to all beings equally. A facilitator cannot rightly request payment from a client for providing what the client already possesses- **Natural Healing Energies.**

In Divine Order, it is appropriate for the facilitator to accept **payment** for the **personal time and energies** invested in **learning and developing facilitation skills,** and for the time spent in active facilitation. It is important that the facilitator understands this subtle difference of **"fair exchange of energy"**. If the facilitator mistakenly assumes they are requesting payment **for healing energies** they supply, they are inadvertently misrepresenting themselves to the client. The client possesses personal healing energies and does not need to purchase such energies from others. The **time and expertise** offered by the facilitator to assist the client in stimulating their own healing energies, is a service worthy of reciprocation. If a client chooses such assistance, the **facilitator** has **legitimate right** to honor their own worth as a being by **requesting reasonable payment or barter** in **fair energetic exchange** for the service they offer. However, if a client is in need of healing facilitation, and has no means of fair exchange at that time, the facilitator should honor the worth of the client by providing assistance of equal value **free of charge.**

In spiritual principal, "making money" is not a legitimate motivation for participating in Healing Facilitation. Seeking to find one's personal **"Divine Right Livelihood"** in **sincere service** to others, which might possibly involve providing client service in Healing Facilitation, IS a spiritually legitimate motivation. Healing Facilitators will discover that when they put the objective of **GENUINE SERVICE TO THE DIVINE** as the **first Priority,** rather than the pursuit of personal monies or status, the **Universe Will Provide.** Facilitators who enter the healing field with hidden personal motivations incongruent with **Universal Laws of Energy**, which imply **equality, respect, reverence** and **balanced exchange of energy** between all beings, will find difficulties emerging in their lives as the personal **SOUL** sets up lessons through which **Universal Principles** become more apparent. There is only **ONE TRUE MOTIVATION** for participation in Healing Facilitation- **the motive is LOVE.**

POSTURES OF LOVE

The Energy-Reality of LOVE

In terms of **Universal Physics**, Love is an **energy reality**, a **STATE OF VIBRATIONAL HARMONIZATION**, or **CO-RESONANCE OF FREQUENCY** that allows an **energetic bridge** to build between individuals. It is through this energetic bridge of frequency that one can assist in running healing energies that will facilitate the healing process of others. Without the sentiment and **FREQUENCY** of **GENUINE LOVE**, one cannot energetically facilitate the healing of another. **LOVE is the essential ingredient in healing,** for it allows the opening of facilitator and client's bio-energetic fields to the **Universal Frequencies** of the interdimensional spectrum. Cultivating the ability to embrace and hold the frequencies of Universal or **OMNI-LOVE** is the responsibility of any true healing facilitator. The emotionally experienced reality of the ***frequency of Love*** takes many forms. In healing facilitation it is important to **assess the most appropriate form of love** to engage with each individual client

Soft Love

Many people respond well to **"SOFT LOVE"**, the kind, gentle, nurturing, soothing love often displayed by mother's comforting their infants. Soft Love works well with client's whom possess some degree of spiritual maturity and who have cultivated the ability to hold some degree of a love frequency within themselves. Soft Love is the appropriate form of love for client's possessing the maturity to accept personal responsibility for themselves, their actions and their emotional reaction patterns, and who do not attempt to manipulate, drain energies from or place blame on others. Soft Love is characterized by **engaged detachment** of the facilitator and is demonstrated through soft-spoken, calm and sincere words and actions that are intended to place the client at ease.

Soft Love is always built up sincerity, kindness, respect and a genuine desire to see the client thrive and prosper. Developing the ability to carry the Soft Love frequency takes time, attention and practice, and requires that the facilitator works to heal personal **internal conflicts**, unresolved issues from **childhood** and **karmic** reaction patterns that disrupt the Soft Love vibration. Soft Love is a **vulnerable love** in that it necessitates one to show the self as it is, honestly expressing personal feelings tempered with kindness. Soft Love **is not approval seeking**, it is rather genuine, honest and **self-generated**. Soft Love requires **love** and **respect** of the self and others mutually. It is built upon a level of spiritual maturity in which it is understood that personal value is implied by the fact of existence and is not determined by the approval or **validation** of others outside of the self and the personal relationship to the Divine. The ability to carry the Soft Love frequency emerges through **spiritual integration, personal healing** of internal conflicts and power struggle issues and **intentional application of kindness and mindfulness.**

Tough Love

Soft Love does not facilitate healing on all occasions. Certain clients who come to a facilitator for healing assistance bring with them a great need for **emotional healing** and development of spiritual maturity. They may demonstrate demanding, arrogant or pushy attitudes and refuse to accept responsibility for their personal actions and resulting consequences. Most often such individuals carry deep wounds from childhood or from reincarnational bleed-through. Because they are **wounded within**, they have not yet developed the self-control or maturity necessary to treat themselves or others with kindness, respect or love.

Individuals who display unreasonable behaviors or attitudes pose quite a challenge to the love-based healing facilitator. If they are approached with Soft Love, the facilitator will often find themselves as a scape goat for the individual's personal problems, may have their energies and time excessively **drained** by such clients and might possibly be subjected to outright verbal or physical **abuse**. Clients exhibiting such personality traits are struggling within themselves to **gain control** over the various **conflicting portions** of their personal energies, and they tend to objectify this internal conflict resolution in the form of **POWER STRUGGLE** with others.

Personalities trapped within cycles of subconscious self or other-abusive attitude patterns need **love** more than anyone, but most often their behaviors push others away. Intrinsically the behaviors serve to keep the individual sheltered from **self-exposure;** they do not allow others close enough to inspire feelings of vulnerability or lack of self-esteem. Often such personality traits effectively keep the individual from seeing themselves and the inner pain and conflict from which they attempt to hide. To facilitate healing in such individuals without succumbing to their manipulation, aggression or abuse requires that the facilitator adopt a posture of **TOUGH LOVE.** In Tough Love the facilitator clearly establishes **personal boundaries** in their own mind as to what treatment they will and will not accept. If the client crosses those boundaries the facilitator **assertively addresses the issue** with the client, requests that the offensive behavior cease and provides **CLEAR CONSEQUENCES** as to what will occur if the offense continues. The Tough Love approach is rendered from a base of **Soft Love** and **Omni-Love**, but it is strengthened with **Self Love** on behalf of the facilitator. The facilitator recognizes that they are only assisting the client to continue with the self-destructive patterns by condoning or allowing offensive behavior, and chooses to **love the client enough** to **confront the pattern** so that it may begin to release. In **lovingly, calmly**, but **firmly confronting** poor behavior and **setting clear boundaries** and **consequences**, the facilitator assists such clients to temporarily find a new pattern of action because the old one does not work for manipulation in this instance. The Tough Love approach frequently requires facilitators to demonstrate that they hold their own power, even in the face of client disapproval. Tough Love is an ability of **personal empowerment** worth cultivating for personal and client healing.

Self Love

In healing facilitation **Love** is the essential ingredient. All love begins with the **Divinity Within the SELF.** You will be able to transmit the frequency of love to others in direct proportion to your ability to hold the frequency of love within yourself. If you do not recognize your intrinsic value as a **Divine Being** and seek validation of your existence through external sources such as achievements, status or other-approval, you will likewise project these sentiments into your love relationships with others. If a client does not meet the external standards of validation that you have placed on yourself to determine your personal worthiness of love, you will have difficulty not falling into **judgement** of the client.

Judgement and **love** cannot transmit through the human body at the same time. Judgement creates an **energy reality** of separation or **Non-Resonant Frequencies** of energy, whereas Love creates the **Co-resonance of Frequencies** needed for open flow of Universal Energies for healing facilitation. If one can realize that **LOVE IS THE ONLY CONSTANT** and that all conditions of judgement change, it is easier to cultivate the innate ability to fully hold the frequency of Love. Though **ASSESSMENT** of conditions, actions or attitudes is useful and necessary, such assessment can be rendered through **"separating the person from the action".** You can judge the effectiveness or value of the **action or idea** without assigning a **value-judgement** to the **person** to which it is attached.

If the healing facilitator can feel love, reverence, respect and honor for the **Divinity Within Themselves,** they will also transmit these qualities of **excellence in Love** to the client. The quality of Love brought into the healing facilitation experience will directly effect the success of healing assistance given. **Love the Self** and know that all beings possess an **unalienable worth and value** as a living part of the **Divine.** Begin to cultivate this awareness within your personal life and you will greatly increase your effectiveness as a healing facilitator. Love, honor and respect yourself, your feelings, your dreams and your desires, so you may better love those you wish to serve.

Love others as you would love the Self,
and love the Self as you would like to love others.

The Illusion of Martyrdom

People who love themselves do not buy into the idea that they must play **"martyr"** to others in order to be of service. **Martyrdom** implies a belief that you are **lacking in worth,** or must **earn** your worth in being, by relinquishing that which feels of most value to you. It also **undermines expression of the God-Presence** within others, by assisting them to believe that you are so vital to their wellbeing that they cannot function without you. In truth, martyrdom is an **expression of EGO**, most often used by people who feel they are **personally insignificant** in the eyes of the Divine, in order to **give themselves significance** and a **sense of personal power and purpose** for being. People who actively *know* the intrinsic **Spiritual Divinity** within themselves treat that Divinity with honor and respect and do not allow or need others to **depreciate the Self** and its desires in order to prove the existence of a *belief* in personal Divinity.

Martyrdom **does not facilitate healing** within the self or client, as it reinforces the **erroneous belief** in intrinsic **personal unworthiness**, which exists in **direct contradiction** to the **innate God-Presence** that exists in every living thing. Within the energy dynamics of **Universal Law,** natural **balance** is achieved in the flow or **GIVING**, and back-flow or **RECEIVING** of energy. In human spiritual relationships this **balance of energy** is achieved by **mutually valuing** each person in a relationship, through which the natural giving and receiving of energy can be exchanged unimpeded. **Martyrdom impedes the natural flow of energy**; the martyr customarily gives more than receives, which progressively diminishes the martyrs ability to keep open a perpetual flow of energies from the Divine because the martyr does not allow the Self to receive.

Martyrdom is not a demonstration of Love,
it is an exhibit of Love Deprivation.
The remedy for Martyrdom is
OMNI-LOVE.

OMNI-LOVE; Fully Engaging the Love Vibration

The greatest love that can be embodied in human form is **OMNI-LOVE**. OMNI-LOVE is a pure state of vibrational Frequency Co-Resonance between the human identity and the **12 Dimensions** of reality within which the human anatomy is couched. Omni-Love is expressed in human form as **Christed Love** or **Christos Consciousness, the 12th-Dimensional Love** of the **Maharic Level of Identity.**

Christos Love is a state of full Frequency Resonance with **everything existing** in the many universes contained within **12 Dimensions** of the 15-Dimensional System. It is a **transcendent love** that is attached to no thing but is **AT ONE** with **ALL THINGS** in a state of perpetual **Engaged Detachment.** Through Christos-Omni Love, the **SELF** is known as an **extension of God**, or the **Divine Source**; a **temporary mask of form** worn by the **ONE CONSCIOUSNESS** of the Cosmos. All things and beings are known as **simultaneous expressions** of the **ONE-SELF** that is God-Source. From this state of transcendence in love, all activity is understood to **exist within** the reality of LOVE, and all conflict and strife are viewed as the ONE-SELF **progressively expanding the ability of its expressions** to **carry the frequencies of energy** that constitute **ONE-LOVE,** a state of **total vibrational Frequency Co-resonance** with the **Cosmos.**

Activation of the full **12-Strand Silicate Matrix DNA Template** and full integration of the **Mahara Hova Body Avatar Identity** allows for the experiential embodiment and transmission of Christos-Omni-Love. **Embodiment of Omni-love is the goal toward which human evolution moves forth.** Omni-Love is Soft, Tough, and Enduring, honoring **Self** , **Other** and the **Divine** simultaneously. It is the **strongest healing force** of frequency in the 4 Harmonics of Matter Density. Christos-Omni-Love is the **fulfillment** of the **Christos Principle.** Strive to bring Omni-love into your life, and personal and client healing facilitation, and you will assist your higher identity levels to create transformation! Omni-Love is the **natural structure** of reality. **AWARENESS** of Omni-Love is cultivated through **intention** and appropriate use of **personal ~~free~~ will choice** in congruence with the **natural Laws** of the **Unified Field Physics of Consciousness and Creation.**

Through living in Omni-Love all moments are known as Sacred, all beings are known as Blessed and life becomes an ecstatic study in Reverence for Existence.

Postures of LOVE

1. SOFT LOVE
KIND, NURTURING, **GENTLE**

2. TOUGH LOVE
KIND, NURTURING, **ASSERTIVE**

3. SELF LOVE
HONORING THE **DIVINITY** WITHIN **SELF**

4. OMNI-LOVE
HONORING THE **DIVINITY** WITHIN **ALL**
Namaste'

***Omni-Love** is the fulfillment of the **Christos Principle,** the goal toward which human evolution moves forth through progressive activation of the **12-Strand Silicate Matrix DNA Template** and embodiment of the **Avatar** Identity.*

OMNI-LOVE EXERCISE TO AMPLIFY HEALING FACILITATION

EXERCISE

The MAHARA SPACE-COMFORT TOUCH

1. **EXHALE Hierophant Symbol** from behind forehead, down *Central Body Current*, into **Earth's Core** (Chakra 13), then **INHALE** *Hierophant* back up to your **Maharic Shield 12" below feet.** *Spin Hierophant* to *activate* **Maharic Shield** and **Maharic Seal Pillar. INHALE** energy up from your *Maharic Shield* into your **Doradic Sphere** at *Chakra 4.*

2. Breathe slowly and deeply while visualizing a **small faceted Crystal Sphere** in the center of your *Doradic Sphere* at Chakra 4. This is the **Holographic Crystal Sphere**; it represents the station of your inner **AVATAR** identity – this is the **MAHARA SPACE**. Focus your consciousness within the *Holographic Sphere* and feel its reality around you, as if your body is encased within the Holographic Sphere.

3. Observe the interior of the Sphere as a **pale Silver-White mist**. Within the mist now envision a **golden treasure chest**- a treasure you had left for yourself long ago to one-day reclaim. This represents your **Inner Arc of the Covenant** – *a doorway to reclaiming your ancient immortal heritage.* **Mentally command** the lid of your inner *Arc of the Covenant* to lift. As the lid rises **pale rainbow colored light** fills the interior of your Holographic Sphere. Notice now that within the golden box there is a gleaming **Silver Sword**, the Silver Rod of Power that will enable you to more quickly evolve to your eternal state of being.

4. Mentally make the **Silver Sword** *rise from the golden box*, floating above the box in a vertical position. Now call to mind the image of your **body** *as it sits within the physical room.* Imagine *the Silver Sword positioned above the body's head,* directly over **Chakra 7. EXHALE** and **steadily lower the Silver Sword** through the *7th Chakra and down the Central Body Current, through the Maharic Shield and down into* **Earth's core.** Observe the Silver Sword turn into a *Pillar of Pale Silver Light with pale rainbow colors flashing through it.* The **Rainbow Pillar** now *surrounds and runs through your body, raising its frequency while lowering its vibration,* lifting your body cells into a **less-dense state of matter.** Feel your body growing lighter and lighter as the Rainbow Pillar slows the spin-rate of your matter particles, making you more and more electric- more light.

5. As you feel the sensations of your density lessening, **imagine that your body is beginning to glow with Pale Silver light** – the *Pre-matter Liquid Light* of dimensions 10, 11 and 12. Feel the *features of your face and the contours of your body begin to melt* into a **Silver, 2 –dimensional fluid body form.** The form of the body still remains but it has been returned to its **first-density state**- the **Silver Liquid Light holographic form of the avatar identity.** Know yourself now as this avatar identity and feel the essence of **pure love** and eternity flow through you. **Know yourself as this essence of Divine Love and Eternal Light.**

6. Now **INHALE** and draw the **Rainbow Pillar into the Holographic Sphere** at *Chakra 4,* visualize the base pulling up from Earth's core and the portion above your head simultaneously *contracting into the Holographic Sphere.* **EXHALE** the energy of the *Rainbow Pillar* fully into the *Holographic Sphere* and observe the **Holographic Sphere become a radiant Silver ball of Liquid Light stationed with your Doradic Sphere** in the Heart Chakra.

EXERCISE: Continued
THE *MAHARA SPACE-COMFORT TOUCH*

7. **You may now transmit the frequencies of the 4th-Harmonic, dimensions 10, 11 and 12, to serve as a healing force for others.** You may transmit this energy into your environment directly through the **4th Heart Chakra**, where it will enter the Unified Field at the 4th-Dimensional astral plane. You may use the breath to move the Holographic Sphere down into the **3rd Chakra** *for 3rd-Dimensional transmission* into the *Earth's atmosphere* and the *Mental Body* levels of people's Bio-energetic field. You may transmit through **Chakra 2** into the Telluric or *Elemental Kingdom* of the planet and into people's 2nd-Dimensonal *Emotional bodies.* Or you may transmit the frequencies of the Silver Sword through the **1st Base Chakra**, for *sacred bonding in intimate relationship*, or to *raise the frequency and lower the vibration* of the *molecular and sub-atomic fields* around you. To transmit the frequencies of the Silver Sword through the chakras, simply use the **EXHALE** breath to move the Silver Light "charged" **Holographic Sphere down** to lower chakras, inhale and use the next **EXHALE** to *push the frequency outward* from the selected chakra into the environment. Use the **INHALE** breath to move the *Holographic Sphere* up through the chakra system. *Always move the Holographic Sphere up or down along the* ***Central Body Current****, to which the core of each chakra is attached.*

8. **You may also transmit the frequencies of the Silver Sword through APPLIED TOUCH –** ***THE COMFORT TOUCH*** -to amplify specific healing facilitation functions and to establish a more harmonious rapport with other people and the nature kingdoms. Simply visualize the Silver Liquid-Light charged **Holographic Sphere** at the center of your **Doradic Sphere** in the **4th Chakra**. Activate your **Doradic-1 current** via the **Maharic Shield and Seal**, then visualize the *Silver Liquid Light from the Holographic Sphere combining with the Green-Violet-Silver Doradic-1 current* as you **EXHALE** the current into the **Doradic Cords** in the arms and into the **Doradic Balls** in the palms. *This application is especially useful to help Clients relax during sessions, or to comfort others when they are in a state of stress, unbalance or illness.*

   Once your current is moving into the Doradic Balls, **EXHALE** the current through the **Right palm.** *Place the* ***Right palm*** *gentle and briefly over the* ***Client's rear Chakra 2****, or touch them gently on the* ***left shoulder.*** The amplified *Doradic-1 current* will move into the *4th-Dimensional astral-body field*, and move its way into the *mental, emotional and physical bodies* from there. **The COMFORT TOUCH of the MAHARA SPACE can be applied to infants, children, older persons, animals and plants to comfort them and support their general well-being.**

PROFESSIONAL USE OF THE KATHARA BIO-SPIRITUAL HEALING SYSTEM PROGRAM – THE KATHARA ALLIANCE

The Kathara System introduced in **The Kathara Bio-Spiritual Healing System ™ Level 1 Workshop & Course Manual with Introduction to The Kathara Bio-Healing System™ Level 2** is destined to become a ***12-level program*** that is focused on "whole-being healing through progressive biological and spiritual awakening". Kathara is as much a course on (non-denominational, dogma-free) **Spiritual Integration Mechanics** as it is a Holistic Healing course.

An individual completing training in the **Kathara Level-1 Bio-spiritual Healing System** delivered by a Certified Kathara Alliance Instructor, will receive a **Certificate of Completion**, which recognizes the person named as attending and receiving training in the Kathara Level One Bio-Spiritual Healing System.

If you desire to practise the ***Kathara Bio-Spiritual Healing System™*** in a regular, professional capacity through Client Service, you may use the techniques provided within this Kathara Healing Manual to begin. Prior to professional use, it is suggested that one becomes a Certified member of the **Kathara Alliance** through the ***Azurite Press MCEO (World-wide Consulate***) and its appointed representatives.

The **Kathara Alliance** is a group of dedicated spiritual healing Practitioners and Instructors who have demonstrated a thorough understanding of the Theory and Practice of the **Kathara Bio-Spiritual Healing System™** and the Maharic Recoding Process, as deemed necessary by the **Azurite Press** for professionalism in the practice of Spiritual Healing Facilitation and Instruction.

All members of the ***Kathara Alliance*** are **Certified Kathara Healing Practitioners and Instructors,** and are permitted to practice Kathara Healing facilitation as **SPIRITUAL SERVICE.** For the protection of both Client and Practitioner, as well as the ethics inherent to the Teachings themselves, Kathara Healing is to be promoted, explained and practiced as **a non-diagnostic spiritual healing enhancement service only** and not as a replacement for traditional medical diagnosis or treatment.

Certification in the **Kathara Alliance** as an **Independent Spiritual Healing Practitioner and Instructor** of the **Kathara Bio-Spiritual Healing System™** requires formal application, including a written open-book examination and video presentation by the applicant. The examination is intended to reveal areas in which applicants may need to further strengthen understanding of and personal relationship to the practice of Kathara Healing Facilitation and Instruction within the parameters of Excellence recognized as intrinsic by the Azurite Press MCEO.

All Instructors and Practitioners must be CERTIFIED to conduct **public presentation** of copyright materials contained in this, the authorized Kathara Bio-Spiritual Healing System Manual. To uphold standards of Professionalism and Excellence in practice, the Azurite Press MCEO **cannot** recognize **Non-Registered Facilitators or Instructors** as active participants in the ***Kathara Bio-Spiritual Healing System Program.*** Obviously, non-certified Facilitators may use this program and its applications for their own **personal and spiritual enrichment. The Azurite Press MCEO authorizes only Certified Practitioners and Instructors** to professionally conduct Client Sessions.

With regards to teaching the program there is a distinction between 'Teaching' and 'Sharing'. **Teaching** is defined as formal presentation of the materials (following protocols described in the **Kathara Bio-Spiritual Healing System™** Manual and communications to Kathara Alliance applicants and members) in the form of a seminar/ workshop and payment is made by attendees. To teach the Kathara Bio-Spiritual Healing System™, one needs to be a member of the **Kathara Alliance**. '**Sharing**' is defined as leading a study group where attendees are not expected to pay (it would not be inappropriate to accept small donations if offered). To share one does not need to be a member of the Kathara Alliance but group size should be limited to less than 12.

Within the **Kathara Bio-Spiritual Healing System™** there are specific pages for Kathara Alliance Instructors to use in their classes. The Instructors pages are identified by having a number at the *top* of the page (as well as the regular number at the bottom). These are the only pages that can be photocopied for teaching purposes. Only certified Kathara Alliance members (or a Kathara Alliance applicant who is delivering and recording a workshop for assessment purposes) are authorized to photocopy these pages.

The easy-reference **Instructor's pages** need to be sequenced in the order denoted by the number at the *top* of the page. These pages have been carefully sequenced to facilitate delivery and learning of the materials. Learners' needs, including varying learning style preferences, were very much considered when preparing the text and visual presentation of these teaching materials. In terms of frequency, the learning process can be described as when previously dormant Scalar Standing Waves in the personal shields are stimulated and activate('switch on'). As frequency is accreted (by progressive exposure to, and interaction with, the materials) the scalar waves in the personal shield self –sustain (we would say the person has 'learned' the material). The specific sequence of the pages are pre-set to create optimum frequency accretion in the personal shields of learners *even if this is not immediately perceived or understood by the learner*. Therefore it is imperative that these pages are used in the delivery of this program and the instructor remains true to the given sequence.

As Independent Spiritual Healing Facilitators and Instructors, all Certified Facilitators and Instructors are expected to accept full personal responsibility and legal liability for the practice and application of the **Kathara Bio-Spiritual Healing System™**. Certified Facilitators-Instructors are **not required** to pay dues, royalties or revenues to the Kathara Alliance/APMCEO for use of this program other than the cost of initial personal **Training-Registration** and obtaining supplies of **Session Manuals, Certificates and Registration Examinations** for the professional Instruction-Certification of others.

THE KATHARA ALLIANCE CERTIFICATION AND COMPETENCY REQUIREMENTS

The Certification Program is designed to enable the applicant to fully demonstrate his/her understanding of the material in a way that maintains the level of integrity and purity of the Program desired of the Kathara Alliance. This also ensures the Student attains a desired level of understanding in taking the workshop. The requirements of Certification are designed to protect the integrity of the material, those who participate with it, and the rights and responsibilities of all concerned.

Certified Kathara Practitioners and Instructors must not only be knowledgeable in the practice and teaching of Kathara, but are also asked to commit to the behavioral guidelines originating from the Covenant of Regent Consulate of the Melchizedek Cloister Emerald Order (MCEO). In this way those seeking out a Kathara practitioner can reasonably expect the highest integrity in both the teachings and in the individuals with whom they are relating.

Certified Kathara Alliance members are in no way affiliated with the MCEO Vehicle, the Azurite Press and should at no time state or imply such representation. Kathara Alliance members are independent Practitioners who offer local, community-based Kathara Healing Level courses and facilitation for individuals.

PLEASE NOTE: The APMCEO & Kathara Alliance **reserve the right** to **revoke** any Kathara Alliance Certification and Rights of Association of Certified Facilitator-Instructors for ***any*** reasons of: *Misuse, misrepresentation and false claims pertaining to this Kathara Program; inappropriate treatment of APMCEO staff, clients or participants in Kathara Healing Programs; dishonest or unethical behavior and practices, and racial, gender, or religious discrimination of clients and participants or Kathara Alliance Members.*

For further information on the Kathara Alliance certification process visit our website at http://www.azuritepress.com/kathara/kathara_application_proceedures.html or contact

In the U.S. and Canada: Sandy Sayles, 8527 Cool Water Ct., Antelope, CA 95843 - email - siepearl@comcast.net **OR** Melissa Fathman, 855 W. Dillon Rd, Apt. H306 - Louisville, CO 80027 - email - mfathman@yahoo.com

Rest of World Office: Hilary Lyons 341A Green Lanes, Palmers Green, London N13 4TY, U.K. hilary.lyons@virgin.net 07970770404

THE 6-STEP MAHARIC RECODING SESSION

Before Client Arrives: ***Technique # 1***- Activate Your Mentor. Conduct ***Mahara-Space-Comfort Touch Exercise*** and remote-request permission from client's ***Rishi*** identity level to apply **Comfort Touch** when client arrives. Create a calm, pleasant environment where you will not be interrupted by phone or other distractions. Optional: Sing 3 rounds of the ***Ariea-Khum-Nar'A*** **toning** to amplify personal healing and DNA Template activation during healing facilitation session.

When Client Arrives: Apply Comfort Touch if Rishi gave remote permission. Sit with client and review terms of the ***Client Agreement Form*** and secure client's signature on form. Offer client the option of audio-taping the session for client's at-home review.

6-Step Technique Application Sequence- *Maharic Recoding Session*

1. ***Technique # 2***-**The Maharic Seal and the Liquid Light Cleanse** *to seal and align personal bio-fields.*

2. ***Sequence-1 Kathara Points massage***- first your points, then client's points, *to open Kathara Lines.*

3. ***Technique # 3***- **Opening the Healing Channels** to activate your *Doradic-Phase-1* healing current and ***Technique # 4***- **Amplifying Inner-Audio-Visual and Kathara Scan** to locate miasms for healing in client's Kathara Grid.

4. ***Technique # 5***- **The Maharic Infusion** to activate client's Maharic Shield and Seal, recode client's Kathara Grid to the Maharic Shield and clear miasmic buildups.

5. **Optional:** Apply **light touch-massage** to miasmic areas, or employ other healing therapeutics after Maharic Infusion.

6. ***Sequence-2 Kathara Points massage*** on client to revitalize outer body systems and close Kathara Lines. Allow client to rest a few minutes before leaving. Answer any client questions and allow time for client to discuss experience if desired.

When Client Leaves: To ensure clear bio-fields, while Doradic-1 Current is still running, draw energy up from Earth's Maharic Shield and run Maharic Current through your Kathara Grid and body, up to Chakra-14 and back down to Chakra-13, then push a Pillar of Silver Light out of Chakra-4 and EXHALE it to release it around body into auric field. Next massage *your* **Sequence-2 Kathara Points** to revitalize bio-field systems and close Kathara Lines.

CHAPTER 4 HIGHLIGHTS

GENERAL KATHARA HEALING FIELD WORK

- Activation of any **Triadic Healing Current** requires temporary release of corresponding **Signet Star Crystal Seals** in the **Kathara Level-2 Crystal Seals Grid.**
- **Maharic 12th-Dimensional Frequency** is required to **manually release** the **Signet Star Crystal Seals** to activate **Triadic Healing Currents. Kathara Healing** utilizes **Maharic Frequency** to activate interdimensional **Triadic Healing Currents** for more powerful healing support.
- Most healing systems have not used **Maharic Current** because the **Planetary Templar Complex** could not run 12th-Dimensional Maharic Frequency for over 200,000 years, until January 2000.
- Release of the **Signet Seals** also activates dormant Strands in the **DNA Template** and releases the **Cranial-Sacral Seals** on the **Kundalini Energies.** Kathara Healing represents advanced, **12-Dimensional Kundalini Healing.**

OPENING THE HEALING CHANNELS

- Triadic Phase Currents are activated in the body by using **Maharic Frequency** from the **PBIS** to release the **Signet Seals** of the **Internal Templar Complex**, which allows higher dimensional frequency to run through the **physical body system** for transmission.
- Triadic Phase Currents are composed of **3 interwoven frequencies**, the **Maharic Frequency**, plus a chosen **dimensional frequency** and its corresponding frequency from **one Harmonic up**.
- Triadic Phase Currents are **directed** by their corresponding **Color Cap Stone**, the **Primary Wave Length** of the dimensional frequency band from which the current is drawn. **Kathara Level-1** activates the **1st Triadic Phase Current**, the ***Silver (D-12) + Green (D-4) + Violet (D-7)*** **DORADIC PHASE-1 Current.**
- **Doradic Current** is activated through the **Maharic Shield,** then **collected** in the **4th Chakra** ***DORADIC SPHERE***, projected down the ***DORADIC CORDS*** in the **arms** and into the ***DORADIC BALLS*** in the **Palm Chakras** for **DIRECT INDUCTION** of frequency transmission in healing facilitation.

PALM CHAKRA POLARITY

- The **Palm Chakras** are **Secondary Chakras** located in the **Palms of the Hands.** Like all Chakras, they function as **Black-White Hole Sets** for **transmitting** and **receiving energy** between the personal Manifestation Template and the dimensional Unified Fields of energy.

- The Palm Chakras are **double vortices** of energy, with **one spiral** exiting the **Palm** and its **corresponding spiral** exiting the **back of the hand.** The **2 Spirals** of the Palm Chakras are linked by a **Crystal Seal** at the center point where the 2 vortices meet. In one Palm Chakra, one spiral rotates **clock-wise**, is **Base Electrical** and **transmits energy** out of body, and the other spiral rotates **counter-clock-wise**, is **Base Magnetic** and **draws energy** into body.

- In **Level-1Kathara Healing** the **RIGHT PALM** is the **TRANSMITTING HAND** and the **LEFT PALM** is the **RECEIVING HAND.** For **transmission,** the **Standard Hand Position** is **LEFT Palm placed over the back of the RIGHT HAND** for pure transmission and **control of specific frequency** transmitted from the **RIGHT PALM.**

- The **RIGHT PALM** is **BASE ELECTRICAL**, the **LEFT PALM** is **BASE MAGNETIC**, but both have **subtle** electro-magnetic capacity.

THE KARMIC-MIASMIC IMPRINT

- To enter manifestation in time, identity enters **cycles** of **SIMULTANEOUS INCARNATION** within the **4 Harmonics of Matter Density** in **12-Dimensions** of the 15-Dimensional **Time Matrix.** The **PRIMAL ORDER Scalar-wave Template** through which an identity enters time is the **Density-5** (dimensions 13, 14 and 15) **RISHIC SHIELD**
- The **Rishic Shield** holds in the Time Matrix the identity of the **Family of Consciousness** from which **incarnates** emerge. The Scalar Patterns of the Rishic Shield represent the tangible **"DIVINE RIGHT ORDER"** in energy and consciousness, for each incarnate manifesting from it. The individual identity must **realign** its **Manifestation Template** with this **Primal Order** to become free from the Time Matrix in full **Ascension.**
- The **Density-4** Pre-Matter Hydro-plasmic **MAHARIC SHIELD** (Dimensions 10, 11 and 12) is the Scalar Template through which identity manifests in the lower-dimensional Densities and which **connects the individual incarnate** to its **Rishic Shield.** The **Maharic Shield** is often called the **PERSONAL CHRISTOS Template,** and it holds the consciousness and **Perfect Primal Order** for all incarnations of an identity in time.

- "**Karma**" is the result of **Free Will choices** made by **incarnates** that involve using personal energies in ways incongruent with the natural **Laws of Universal Unified Field Physics** and the **Primal Order** of the **Christos Template.** Karma creates **scalar-wave distortions** within the Scalar Shields of personal 15-Dimensional Anatomy that **manifest physically** as **energetic clusters** called **MIASMS.**

- **Miasms** are **unintended** groups of **anti-particles** lodged within the particle field of the personal **Manifestation Template** that block and impede the **natural flow of energy** and consciousness into embodiment. **Miasms,** and the misshapen scalar-wave distortions that form them, are the **cause of ALL disease** and dis-harmonic conditions.

- **Health** and Harmonious flow of energy are the **natural conditions** of human design. The **Maharic Shield Personal Christos Template** holds the **Organic Imprint for Health** for all simultaneous incarnates of an identity, and can therefore be used to **clear the Karmic-Miasmic Imprint** from the personal Manifestation Template to progressively restore **Healthful Primal Order.**

- Miasms from **all simultaneous incarnations** manifest within the personal Manifestation Template and within the DNA Template of each, where they become literal **genetic distortions.**
- Each of an identity's incarnates in various time cycles represents one **TIME VECTOR**, which is encoded in the Vector Codes of the **DNA Template.** Karmic-Miasmic **bleed-through** occurs **between incarnates** from different concurrent time cycles through the **living link** of scalar frequency that exists between the **DNA Templates** of simultaneous incarnations.
- The **Karmic-Miasmic Imprint** does not have to be suffered and cleared the "**old fashioned way**" by "walking through the Karma in physical reality". The **Miasmic Imprint** can be cleared **as energy, with energy**, to prevent its manifestation in physical terms via realignment of the **DNA Template** with the personal **Maharic Shield.**
- Any given incarnate can clear the Karmic-Miasmic Imprint by realigning the personal Manifestation Template with the shared Christos Template Maharic Shield. This process is called **MULTI-VECTOR HOLOGRAPHIC RECODING**, which creates **BIO-REGENESIS** of the original 12-Strand **Silicate Matrix DNA Template** of human design and restoration of the **Organic Imprint for Health.**

ATTITUDES AND ETIQUETTE IN HEALING

- **Holographic Recoding** and clearing the Karmic-Miasmic Imprint from the personal Manifestation Template begins with activation of the **Maharic Shield and Seal** in the bio-energetic field and progressive running of 12th-Dimensional **Maharic Frequency** from the **Personal** and **Planetary Maharic Shields** through the **3 Primary Levels** of the embodied **Kathara Grid**; the Level-1 **12-Tree Grid**, Level-2 **Crystal Seals Grid** and Level-3 **Diodic Points Grid.**
- Personal Holographic Recoding begins with consistent use of ***Technique* # 2- The Maharic Seal and the Liquid Light Cleanse**, and is further accelerated by facilitating **Maharic Recoding** in others through running the **Triadic Phase Currents** for **Co-generative Healing.**
- One can facilitate healing in others and amplify personal healing through the using ***Technique* # 5- The Maharic Infusion.**
- The **Maharic Recoding Session** utilizes the combined **Bio-Regenesis Techniques** presented in **Level-1 Kathara Healing** to facilitate **Client Healing Sessions.**

- Professionalism in Healing Facilitation requires an understanding of the **Postures of Love** and a willingness to **cultivate** the ability to hold and transmit **OMNI-LOVE, honoring the Divinity of ALL things and beings.**
- The Primary Postures of Love are **SOFT LOVE, TOUGH LOVE, SELF LOVE** and **OMNI-LOVE.**
- The emotionally experienced quality called **LOVE** is a **REALITY IN ENERGY PHYSICS.** LOVE is a **STATE OF VIBRATIONAL FREQUENCY CO-RESONANCE** that allows **free flow of energy** between consciousness within the Universal Unified Field.
- The human **12-Strand DNA Template** was designed to allow **12-Dimensions of Vibrational Co-Resonance**, the **Christos Consciousness** of the **Maharic Shield**, to embody in human form. **CHRISTOS-OMNI-LOVE** is Vibrational Co-Resonance with all that exists within **12 Dimensions** of the Time Matrix; its embodiment is the **primary goal** toward which **human evolution** progresses.
- Christos-OMNI-LOVE is the **strongest healing frequency** available within the Densities of Space, Time and Matter, and is **essential** for effective **Healing Facilitation** and personal **Ascension.**

CHAPTER 5

INTRODUCTION TO KATHARA LEVEL-2

The Kee-Ra-ShA Induction Process™

Mahara-Merkabic Master Key Healing-Manifesting the Blueprint For Health

Primal Currents, the DNA Sub-Strand Matrix, Merkaba, Psonns, the Trinity Keys and Photo-Dynamic Manifesting

- SPIRITUAL INTEGRATION
- PRIMAL LIFE FORCE CURRENTS-The Kundalini-Antahkahrana, Maharata, Kee-Ra-ShA & Khundaray
- DNA SUB-STRAND MATRIX, FIRE LETTERS, PSONNS & THE TRIBAL SHIELD
- MERKABA &THE SALUTATIONS
- 2 MASTERS KUNDALINI ACTIVATIONS-The Emerald & Amethyst Awakenings
- UNIVERSAL TRINTIY KEYS & MERKABA KEY CODES
- MAHARA MANIFESTING,DNA, KATHARA & PSONNS

TECHNIQUES:

TECHNIQUE # 6: The Maharic-24 Quick Seal
TECHNIQUE # 7: Ps-13 Tribal Shield Activation
TECHNIQUE # 8: Merkaba Salutations
TECHNIQUE # 9: Merkaba Trinity Key Induction
TECHNIQUE # 10: Photo-Radionic Manifesting
TECHNIQUE # 11: Photo-Sonic Manifesting
TECHNIQUE # 12: Kee-Ra-ShA Activation Sequence-Awakening the Flame

Sacred Sciences of the Mu'a Grail Line-
Tools of the Trade

- **The Maharata-** *Universal Life Force Currents, Kundalini, Antahkarana, Maharata, Kee-Ra-Shay, Khundaray and the Maharic Seal.*
  Field Technique-1: ***Maharic Quick Seal***

- **The Tribal Shield and Cue Zones-** *DNA, Fire Letters, the Sacred Psonns and Cue Zones*
  Field Technique-2: ***Tribal Shield-Cue Zone Activation***

- **The Sacred Salutations-** *Electrical & Magentic Merkaba Activation*
  Field Technique-3: ***Sacred Salutations****- Merkaba Tone Dancing*

- **Expediting Spiritual Actualization -** *Tribal Shield Activation, Emerald and Amethyst Awakening Masters Kundalini Activations, DNA Template Bio-Regenesis, Master Key Codes, Melchizedek Cloister Ordinations*

Universal Life Force Currents

One 15-Dimensional Time Matrix with 5 Harmonic Universes

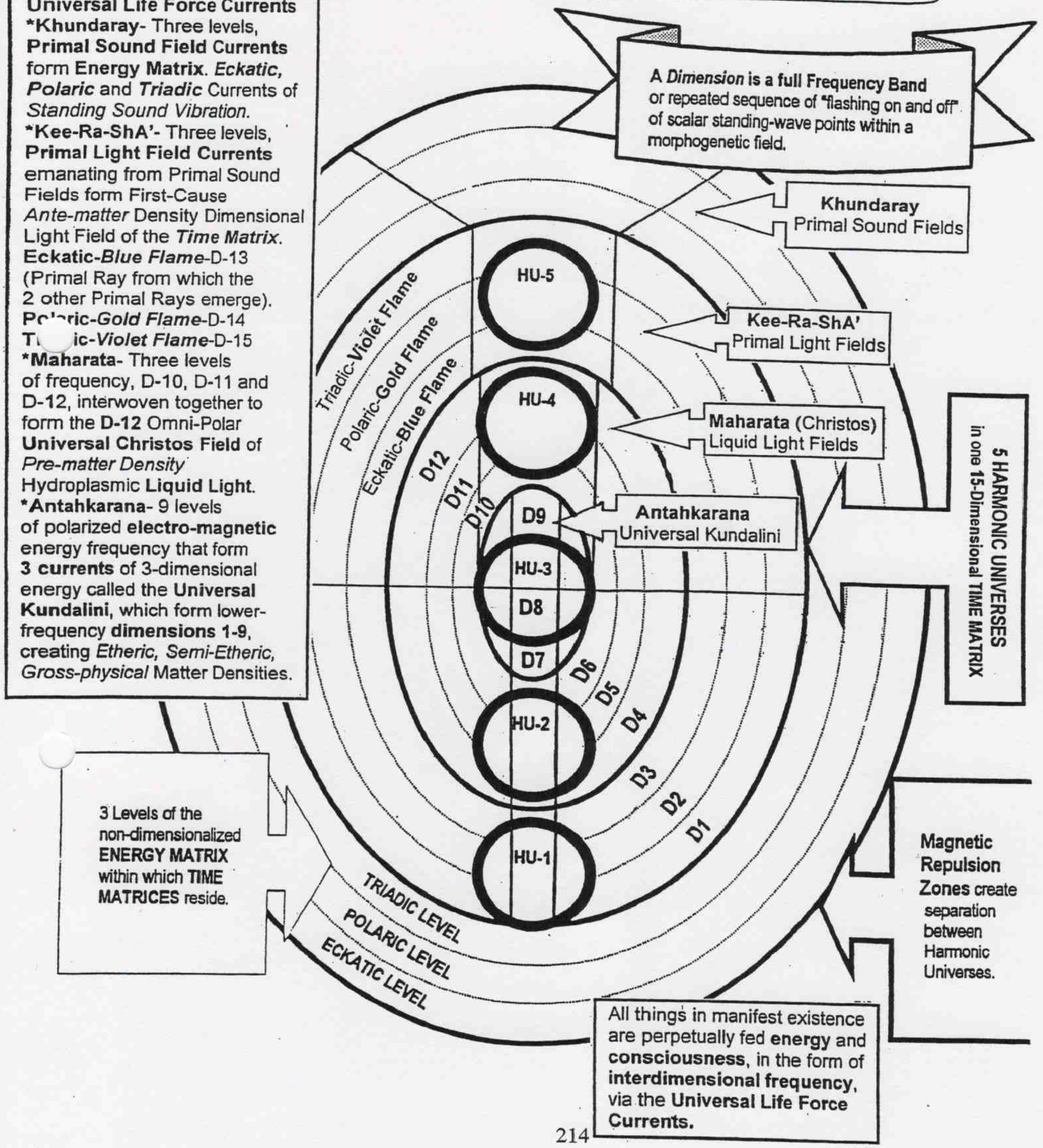

Universal Life Force Currents
Khundaray**- Three levels, **Primal Sound Field Currents** form **Energy Matrix**. ***Eckatic, Polaric and ***Triadic*** Currents of *Standing Sound Vibration.*
Kee-Ra-ShA'**- Three levels, **Primal Light Field Currents** emanating from Primal Sound Fields form First-Cause *Ante-matter* Density Dimensional Light Field of the ***Time Matrix.
Eckatic-*Blue Flame*-D-13 (Primal Ray from which the 2 other Primal Rays emerge).
Polaric-*Gold Flame*-D-14
Triadic-*Violet Flame*-D-15
***Maharata**- Three levels of frequency, D-10, D-11 and D-12, interwoven together to form the **D-12** Omni-Polar **Universal Christos Field** of *Pre-matter Density* Hydroplasmic **Liquid Light**.
***Antahkarana**- 9 levels of polarized **electro-magnetic** energy frequency that form **3 currents** of 3-dimensional energy called the **Universal Kundalini**, which form lower-frequency **dimensions 1-9**, creating *Etheric, Semi-Etheric, Gross-physical* Matter Densities.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

Khundaray
Primal Sound Fields
Kee-Ra-ShA'
Primal Light Fields
Maharata (Christos)
Liquid Light Fields
Antahkarana
Universal Kundalini
5 HARMONIC UNIVERSES
in one **15-Dimensional TIME MATRIX**
Triadic-Violet Flame
Polaric-Gold Flame
Eckatic-Blue Flame
HU-5
HU-4
HU-3
HU-2
HU-1
D12
D11
D10
D9
D8
D7
D6
D5
D4
D3
D2
D1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
3 Levels of the non-dimensionalized **ENERGY MATRIX** within which **TIME MATRICES** reside.
Magnetic Repulsion Zones create separation between Harmonic Universes.

All things in manifest existence are perpetually fed **energy** and **consciousness**, in the form of **interdimensional frequency**, via the **Universal Life Force Currents.**

Universal Life Force Currents and the Embodied Kathara Grid Core Template

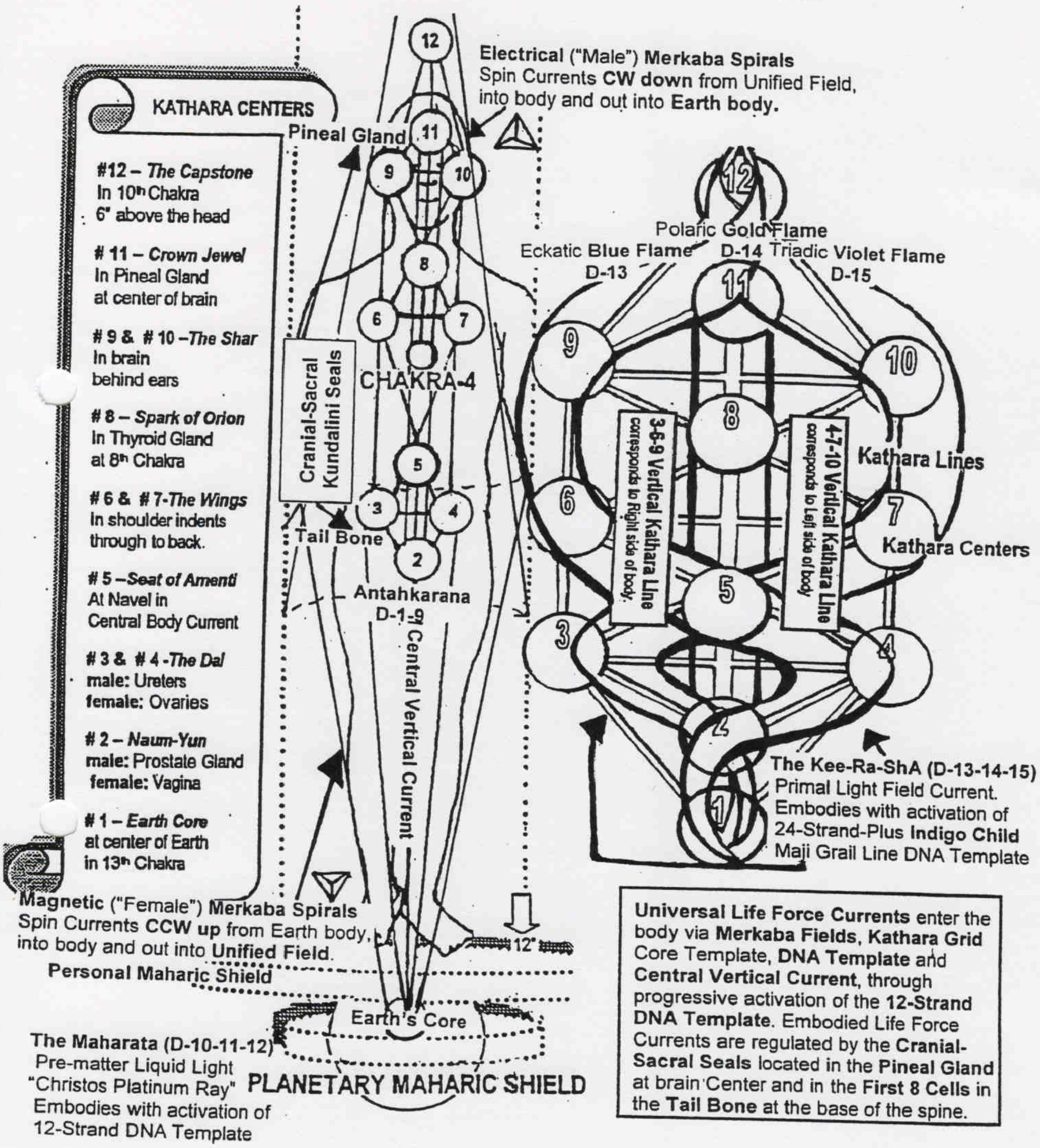
12
11
9
10
8
6
7
5
3
4
2
1
Electrical ("Male") Merkaba Spirals
Spin Currents CW down from Unified Field, into body and out into Earth body.
Pineal Gland
Cranial-Sacral Kundalini Seals
CHAKRA-4
Tail Bone
Antahkarana D-1-9
Central Vertical Current
Magnetic ("Female") Merkaba Spirals
Spin Currents CCW up from Earth body, into body and out into Unified Field.
12"
Personal Maharic Shield
Earth's Core
The Maharata (D-10-11-12)
Pre-matter Liquid Light
"Christos Platinum Ray"
Embodies with activation of 12-Strand DNA Template
PLANETARY MAHARIC SHIELD
Polaric Gold Flame
D-14
Eckatic Blue Flame
D-13
Triadic Violet Flame
D-15
3-6-9 Vertical Kathara Line
corresponds to Right side of body
4-7-10 Vertical Kathara Line
corresponds to Left side of body
Kathara Lines
Kathara Centers
The Kee-Ra-ShA (D-13-14-15)
Primal Light Field Current.
Embodies with activation of 24-Strand-Plus Indigo Child Maji Grail Line DNA Template

KATHARA CENTERS

#12 – *The Capstone*
In 10th Chakra
6" above the head

# 11 – *Crown Jewel*
In Pineal Gland
at center of brain

# 9 & # 10 –*The Shar*
In brain
behind ears

# 8 – *Spark of Orion*
In Thyroid Gland
at 8th Chakra

# 6 & # 7-*The Wings*
In shoulder indents
through to back.

# 5 –*Seat of Amenti*
At Navel in
Central Body Current

# 3 & # 4 -*The Dal*
male: Ureters
female: Ovaries

# 2 – *Naum-Yun*
male: Prostate Gland
female: Vagina

# 1 – *Earth Core*
at center of Earth
in 13th Chakra

Universal Life Force Currents enter the body via **Merkaba Fields, Kathara Grid** Core Template, **DNA Template** and **Central Vertical Current**, through progressive activation of the **12-Strand DNA Template.** Embodied Life Force Currents are regulated by the **Cranial-Sacral Seals** located in the **Pineal Gland** at brain Center and in the **First 8 Cells** in the **Tail Bone** at the base of the spine.

Universal Life Force Currents and Level-2 Kathara Crystal Seals Grid & Chakras

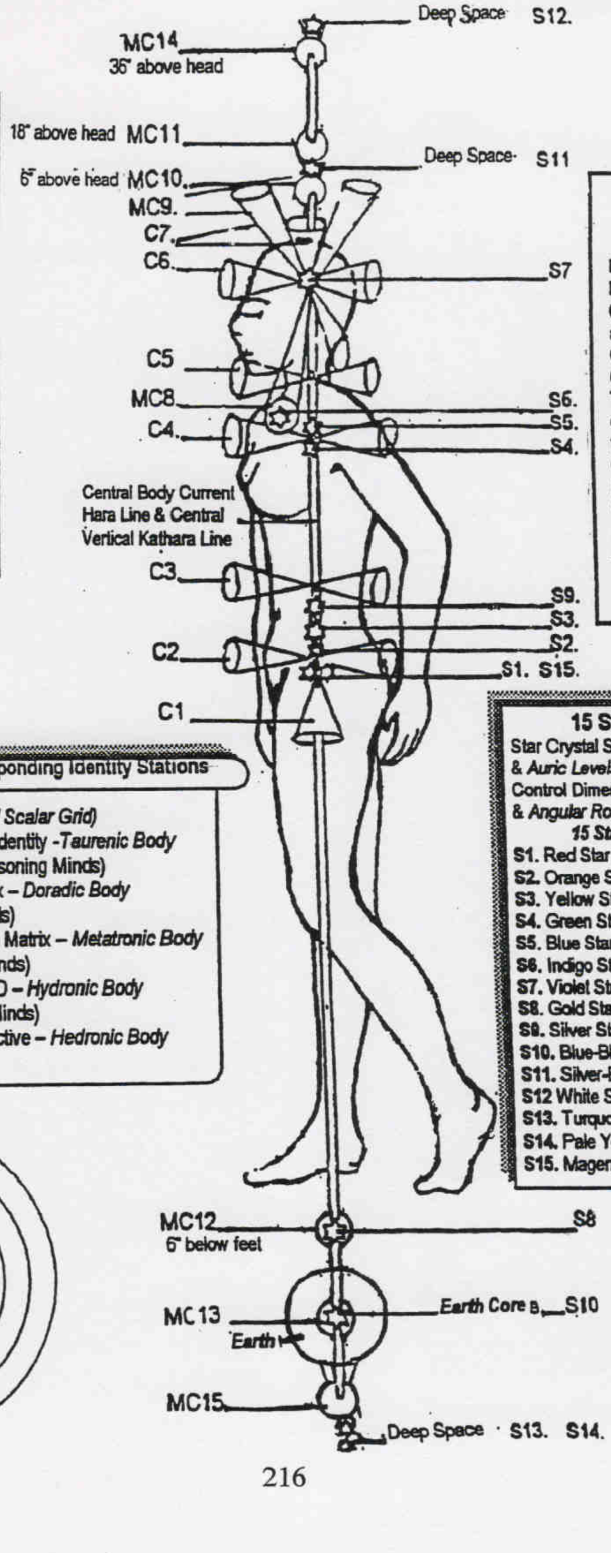
Chakras (C & MC)
Deep Space S12.
Star Crystal Seals (S)
MC14
36" above head
18" above head MC11
Deep Space S11
6" above head MC10
MC9
C7
C6
S7
C5
MC8
S6.
C4.
S5.
S4.
Central Body Current
Hara Line & Central
Vertical Kathara Line
C3
S9.
S3.
C2
S2.
S1. S15.
C1
MC12
6" below feet
S8
MC13
Earth
Earth Core B S10
MC15
Deep Space S13. S14.

15 Chakras (C & MC)

7 Primary Chakras (C)

C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal

8 Morphogenetic Chakras (MC)

MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

Universal Life Force Currents

Enter via **Merkaba Fields, Kathara Grid, DNA Template & Central Vertical Current** and are regulated by the **Crystal Seals** to form the **Chakra System, Axi-a-Tonal and Meridians** of energy distribution in the manifest body. As Currents activate in the Kathara Grid, DNA, Crystal Seals, Chakras and body, **frequency and consciousness** from the higher-dimensional fields progressively embodies.

15 Star Crystal Seals

Star Crystal Seals keep human identity & *Auric Levels* dimensionally separate. Control Dimensional *Merkaba Field Axis* & *Angular Rotation of Particle Spin.*

15 Star Crystal Seals (S):

S1. Red Star
S2. Orange Star
S3. Yellow Star
S4. Green Star
S5. Blue Star
S6. Indigo Star (Blue-Violet)
S7. Violet Star
S8. Gold Star
S9. Silver Star
S10. Blue-Black Star
S11. Silver-Black Star
S12 White Star
S13. Turquoise Star
S14. Pale Yellow Star
S15. Magenta-Pink Star

Hova Bodies and Corresponding Identity Stations

(*"Hova" means Tri-tonal Scalar Grid*)

1. Nada Hova – HU-1 Incarnate Identity -*Taurenic Body* (Subconscious, Instinctual & Reasoning Minds)
2. Alphi Hova – HU-2 Soul Matrix – *Doradic Body* (Astral, Archetype & Angelic Minds)
3. Betcha Hova –HU-3 OverSoul Matrix – *Metatronic Body* (Ketheric, Monadic & Kertatric Minds)
4. Mahara Hova – HU-4 Avatar ID – *Hydronic Body* (Christiac, Buddhaic & Nirvanic Minds)
5. Raja Hova – HU-5 Rishi Collective – *Hedronic Body* (Universal Conscious Mind)

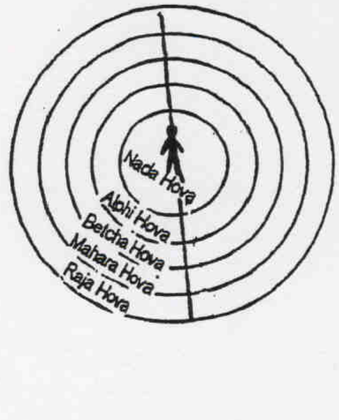
Nada Hova
Alphi Hova
Betcha Hova
Mahara Hova
Raja Hova

Universal Life Force Currents form and sustain the "**Auric Field" Levels** of 15-Dimensional Human Anatomy

Auric Level, Chakra and Hova Body Correspondences

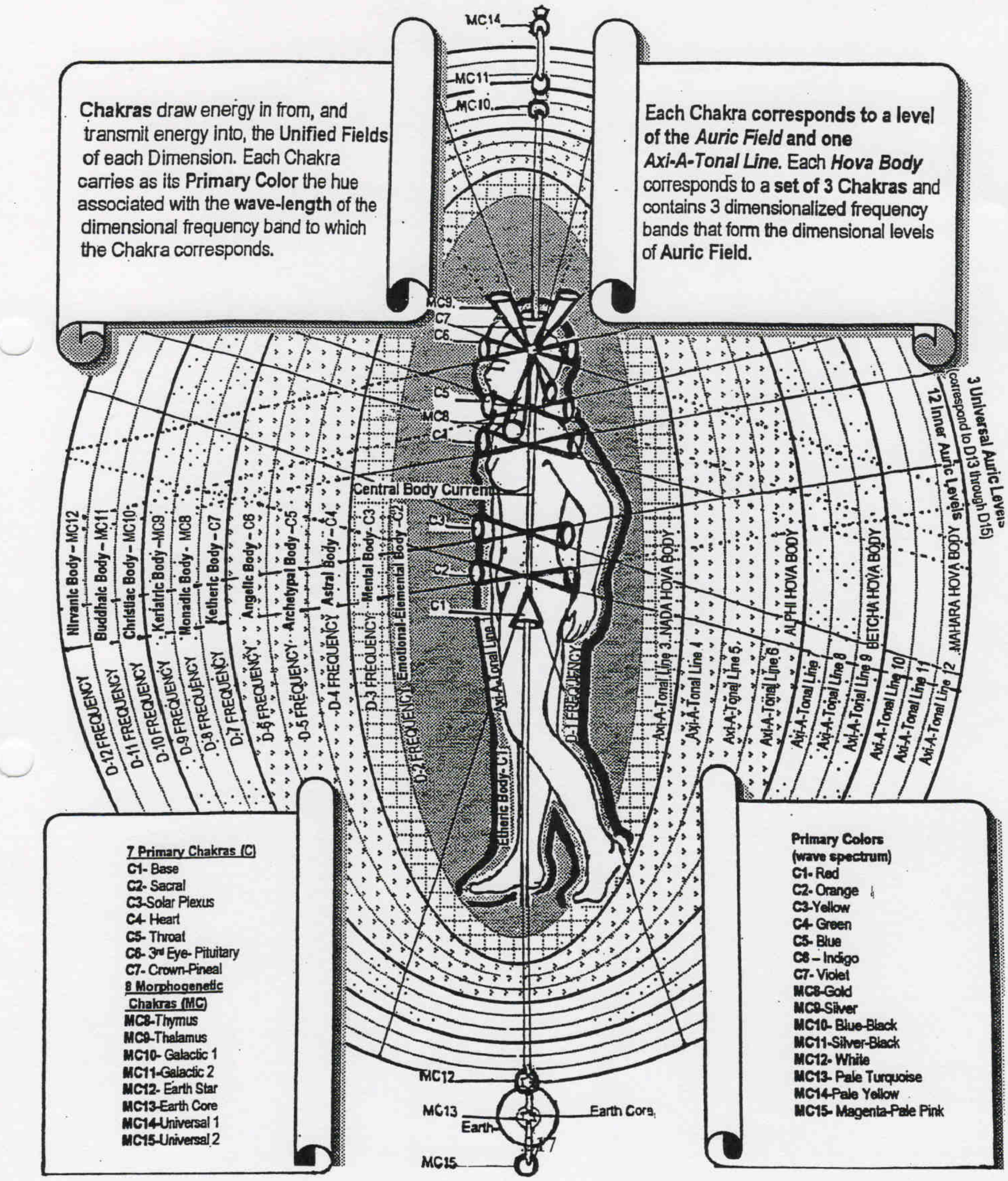
MC14
MC11
MC10
MC9
C7
C6
C5
MC8
C4
Central Body Current
C3
C2
C1
MC12
MC13
Earth
Earth Core
MC15
Nirvanic Body – MC12
Buddhaic Body – MC11
Christiac Body – MC10
Keriatric Body –MC9
Monadic Body - MC8
Ketheric Body – C7
Angelic Body – C6
Archetypal Body –C5
Astral Body – C4
Mental Body- C3
Emotional-Elemental Body –C2
Etheric Body- C1
D-12 FREQUENCY
D-11 FREQUENCY
D-10 FREQUENCY
D-9 FREQUENCY
D-8 FREQUENCY
D-7 FREQUENCY
D-6 FREQUENCY
D-5 FREQUENCY
D-4 FREQUENCY
D-3 FREQUENCY
D-2 FREQUENCY
D-1 FREQUENCY
Axi-A-Tonal Line 1
Axi-A-Tonal Line 2
Axi-A-Tonal Line 3
NADA HOVA BODY
Axi-A-Tonal Line 4
Axi-A-Tonal Line 5
Axi-A-Tonal Line 6
ALPHI HOVA BODY
Axi-A-Tonal Line 7
Axi-A-Tonal Line 8
Axi-A-Tonal Line 9
BETCHA HOVA BODY
Axi-A-Tonal Line 10
Axi-A-Tonal Line 11
Axi-A-Tonal Line 12
MAHARA HOVA BODY
12 Inner Auric Levels
3 Universal Auric Levels
(correspond to D13 through D15)

Chakras draw energy in from, and transmit energy into, the **Unified Fields** of each Dimension. Each Chakra carries as its **Primary Color** the hue associated with the **wave-length** of the dimensional frequency band to which the Chakra corresponds.

Each Chakra corresponds to a level of the *Auric Field* and one *Axi-A-Tonal Line*. Each **Hova Body** corresponds to a **set of 3 Chakras** and contains 3 dimensionalized frequency bands that form the dimensional levels of **Auric Field.**

7 Primary Chakras (C)
C1- Base
C2- Sacral
C3-Solar Plexus
C4- Heart
C5- Throat
C6- 3rd Eye- Pituitary
C7- Crown-Pineal
8 Morphogenetic Chakras (MC)
MC8-Thymus
MC9-Thalamus
MC10- Galactic 1
MC11-Galactic 2
MC12- Earth Star
MC13-Earth Core
MC14-Universal 1
MC15-Universal 2

Primary Colors (wave spectrum)
C1- Red
C2- Orange
C3-Yellow
C4- Green
C5- Blue
C6 – Indigo
C7- Violet
MC8-Gold
MC9-Silver
MC10- Blue-Black
MC11-Silver-Black
MC12- White
MC13- Pale Turquoise
MC14-Pale Yellow
MC15- Magenta-Pale Pink

Universal Life Force Currents and the Embodied Shields, Signets and Kathara Grid

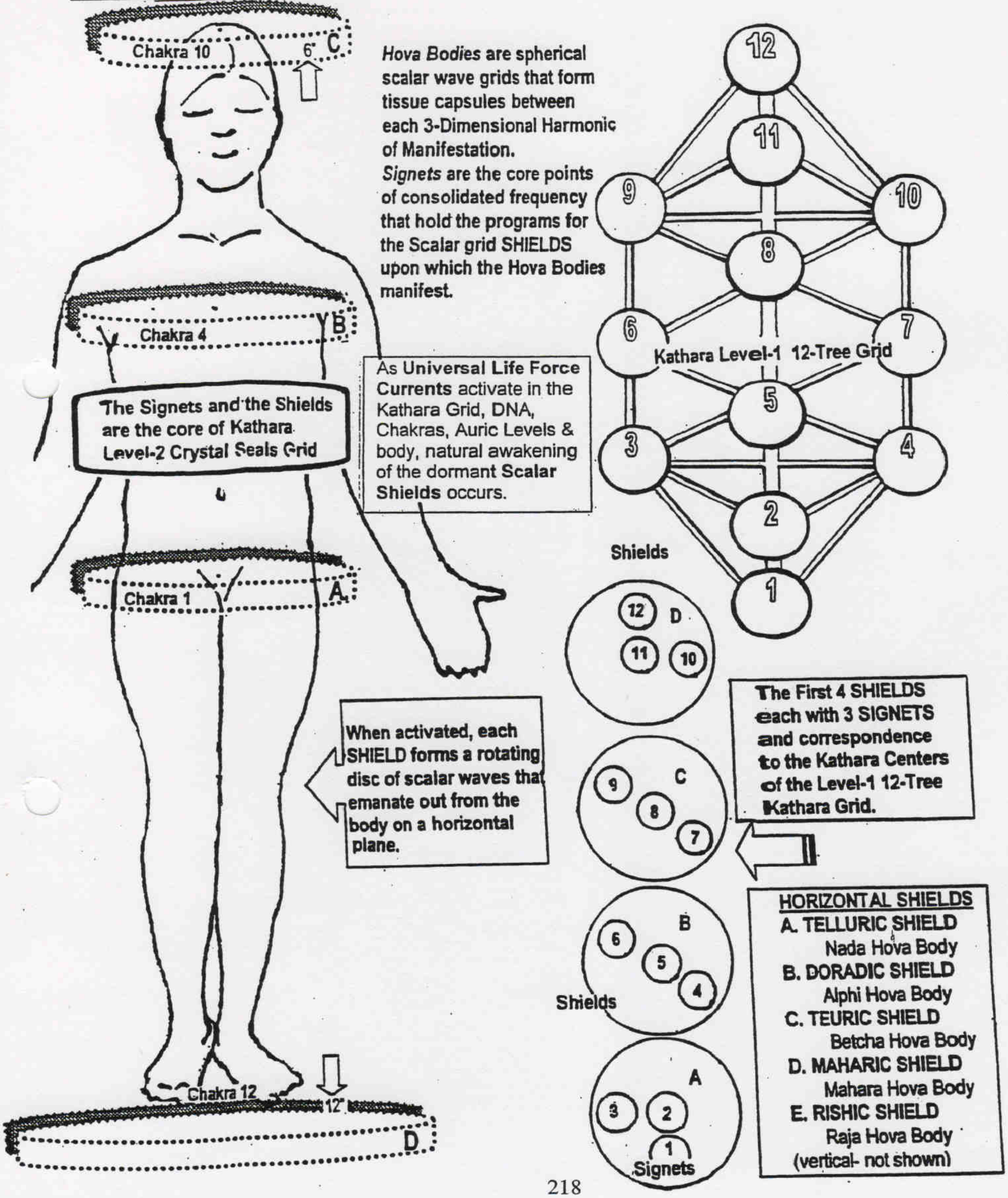
Chakra 10
6"
C
Chakra 4
B
Chakra 1
A
Chakra 12
12"
D
12
11
9
10
8
6
7
Kathara Level-1 12-Tree Grid
5
3
4
2
1
Shields
12
D
11
10
9
C
8
7
6
B
5
4
Shields
A
3
2
1
Signets

Hova Bodies are spherical scalar wave grids that form tissue capsules between each 3-Dimensional Harmonic of Manifestation.

Signets are the core points of consolidated frequency that hold the programs for the Scalar grid SHIELDS upon which the Hova Bodies manifest.

The Signets and the Shields are the core of Kathara Level-2 Crystal Seals Grid

As **Universal Life Force Currents** activate in the Kathara Grid, DNA, Chakras, Auric Levels & body, natural awakening of the dormant **Scalar Shields** occurs.

When activated, each SHIELD forms a rotating disc of scalar waves that emanate out from the body on a horizontal plane.

The First 4 SHIELDS each with 3 SIGNETS and correspondence to the Kathara Centers of the Level-1 12-Tree Kathara Grid.

HORIZONTAL SHIELDS

A. TELLURIC SHIELD
Nada Hova Body
B. DORADIC SHIELD
Alphi Hova Body
C. TEURIC SHIELD
Betcha Hova Body
D. MAHARIC SHIELD
Mahara Hova Body
E. RISHIC SHIELD
Raja Hova Body
(vertical- not shown)

The Silicate Matrix 12 Strand DNA
Primary Correspondences to Subtle Body Anatomy

Silicate Matrix 12-Strand DNA
Primary Correspondences to Subtle Body Anatomy.

DNA Strand #	Dimension #	Axi-A-Tonal Line	Chakra & *Seed* Crystal Seal		Auric Field Level
NADA HOVA BODY – TELLURIC SHIELD				Three 3-Dimensional Kundalini Currents compose the 9-Dimensional Antahkarana The Telluric, Doradic and Teuric Kundalini Currents	
1	1	1	1		Etheric Body
2	2	2	2		Emotional Body
3	3	3	3		Mental Body
ALPHI HOVA BODY – DORADIC SHIELD					
4	4	4	4		Astral Body
5	5	5	5		Archetype Body
6	6	6	6		Angelic Body
BETCHA HOVA BODY – TEURIC SHIELD					
7	7	7	7		Ketheric Body
8	8	8	8		Monadic Body
9	9	9	9		Keriatric Body
MAHARA HOVA BODY – MAHARIC SHIELD				**Maharata** D-10-11-12 Christos Current	
10	10	10	10		Christiac Body
11	11	11	11		Buddhaic Body
12	12	12	12		Nirvanic Body

As **Universal Life Force Currents** activate within the body, the **dimension of consciousness** contained in each **Auric Field Level** progressively embodies.

Universal Life Force Currents and DNA Template Fire Letter Sequences

15-Dimensional Levels of Identity, DNA and Indigo Children

Progressive activation of dormant 12-48 Strand DNA Template creates activation of corresponding Universal Life Force Currents in body.

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

The DNA represents fixed scalar-wave sequences/" **Fire Letters**" that correspond to each dimension/frequency of consciousness. Consciousness enters time by fragmenting into the scalar fields of the dimensions. Freedom from space-time-matter is achieved by reassembling the dimensional Fire Letters, progressively expanding the consciousness and DNA, & transmuting the body matter out of density. There are **48 Primary Fire Letter Sequences**/DNA Strands in full transmutation, through which the consciousness expands into non-dimensional levels of **Ascended Mastery**.

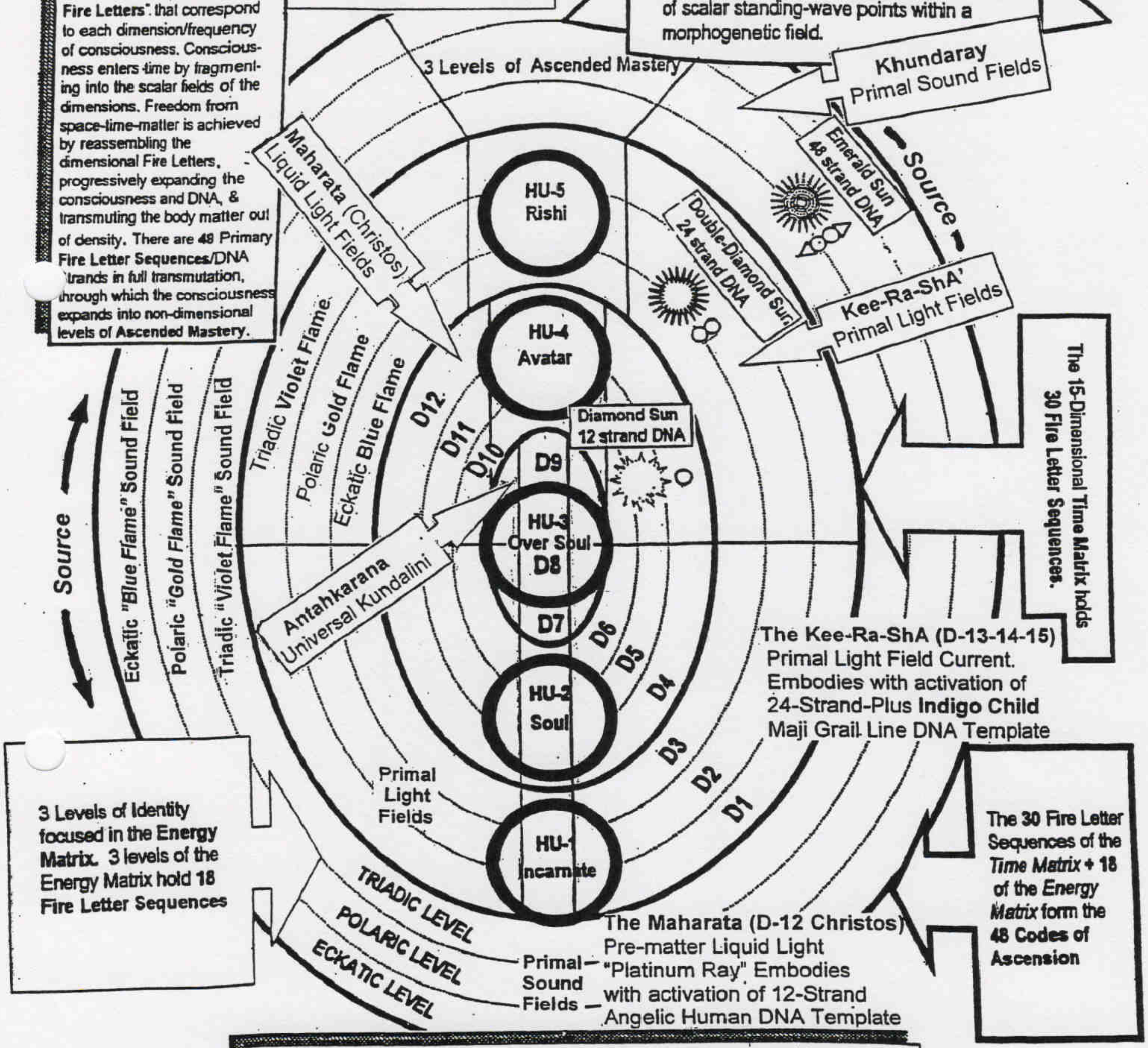
3 Levels of Ascended Mastery
Khundaray
Primal Sound Fields
Emerald Sun
48 strand DNA
Source
Double-Diamond Sun
24 strand DNA
Kee-Ra-ShA'
Primal Light Fields
Maharata (Christos)
Liquid Light Fields
HU-5
Rishi
HU-4
Avatar
Diamond Sun
12 strand DNA
D12
D11
D10
D9
HU-3
Over Soul
D8
D7
D6
D5
D4
HU-2
Soul
D3
D2
D1
HU-1
Incarnate
Triadic Violet Flame
Polaric Gold Flame
Eckatic Blue Flame
Source
Eckatic "Blue Flame" Sound Field
Polaric "Gold Flame" Sound Field
Triadic "Violet Flame" Sound Field
Antahkarana
Universal Kundalini
Primal Light Fields
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL
Primal Sound Fields

The 15-Dimensional Time Matrix holds 30 Fire Letter Sequences.

The Kee-Ra-ShA (D-13-14-15) Primal Light Field Current. Embodies with activation of 24-Strand-Plus **Indigo Child** Maji Grail Line DNA Template

3 Levels of Identity focused in the Energy Matrix. 3 levels of the Energy Matrix hold 18 Fire Letter Sequences

The 30 Fire Letter Sequences of the *Time Matrix* + 18 of the *Energy Matrix* form the 48 Codes of Ascension

The Maharata (D-12 Christos) Pre-matter Liquid Light "Platinum Ray" Embodies with activation of 12-Strand Angelic Human DNA Template

The *Oraphim* genetic imprint of the **Indigo Children** allows more rapid re-assembly of the **48 DNA Fire Letters**/ scalar-wave grids, creating the potential to embody the *Avatar* identity and open the consciousness to *Ascended Master* Identity Levels. **Indigo Children are Avatars and Ascended Masters waiting to awaken.**

All Angelic Humans possess a minimum of **12-Strand** DNA potential through which activation of the **Maharata Current** and embodiment of the D-12 Christed Avatar identity can take place.

The Reality of Spiritual Integration

- **Spiritual Actualization** is as much a function of **Divine Physics** as it is a product of **Divine Consciousness.** The natural Primal Order of **Universal Structure** sets the **Templates of Manifestation** to which consciousness **must conform to** enter the experience of manifest expression within a space-time-matter system.
- The personal Christos **Maharic Shield** Manifestation Template and the Christos **Tribal Shield** Species Manifestation Template are **microcosmic replicas** of the macrocosmic Primal Order upon which Universal and Cosmic structure are built.
- The Personal D-12 Christos Divine Blueprint Maharic Shield connects each Individual Christos self to the **larger** D-12 Christos Divine Blueprint of the **Species.** The D-12 Christos Divine Blueprint is called the **TRIBAL SHIELD.**
- To understand the realities of **personal Spiritual Integration and Actualization,** in order to **fulfill our personal life purpose,** it is important to comprehend the basic **Primal Order** structures of energy, the **personal D-12 Maharic Shield** and the **D-12 Species Tribal Shields,** through which our consciousness manifests in time.
- The secrets to **reclaiming the Divine Birthrights and Responsibilities** of Angelic Human heritage are hidden within the realities of the **Tribal Shield** and the specific **"Fire Letter Sequences"** of the Tribal Shield that manifest within and **govern the function** of the Angelic Human **DNA Template.**
- The **key to Spiritual Actualization** is activation of the dormant personal 12-Strand DNA Template. The **most rapid means** of activating the dormant Angelic Human 12-Strand DNA Template is through **activation of the Tribal Shield,** the Species D-12 Christos Divine Blueprint, which represents the **CORE PROGRAMMING** of the personal D-12 Maharic Shield.

The Silicate Matrix 12-Strand DNA Template with Hova Body, Scalar Shield and Identity Level Correspondences

The Morphogenetic Field Scalar-wave Template for the Original Human Genome

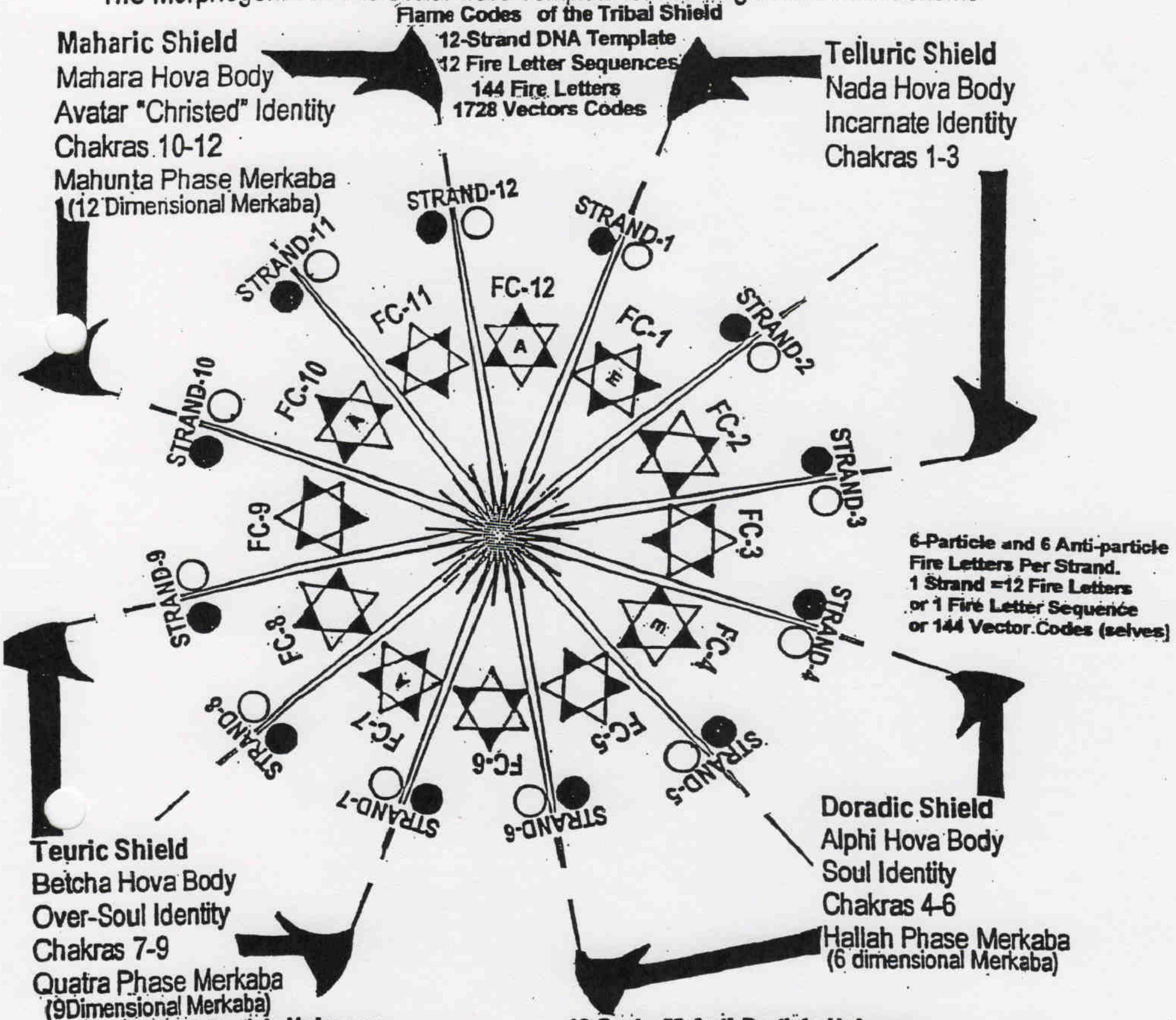

Flame Codes of the Tribal Shield
12-Strand DNA Template
12 Fire Letter Sequences
144 Fire Letters
1728 Vectors Codes

Maharic Shield
Mahara Hova Body
Avatar "Christed" Identity
Chakras 10-12
Mahunta Phase Merkaba
(12 Dimensional Merkaba)

Telluric Shield
Nada Hova Body
Incarnate Identity
Chakras 1-3

6-Particle and 6 Anti-particle Fire Letters Per Strand.
1 Strand =12 Fire Letters
or 1 Fire Letter Sequence
or 144 Vector Codes (selves)

Doradic Shield
Alphi Hova Body
Soul Identity
Chakras 4-6
Hallah Phase Merkaba
(6 dimensional Merkaba)

Teuric Shield
Betcha Hova Body
Over-Soul Identity
Chakras 7-9
Quatra Phase Merkaba
(9Dimensional Merkaba)

STRAND-12
STRAND-1
STRAND-2
STRAND-3
STRAND-4
STRAND-5
STRAND-6
STRAND-7
STRAND-8
STRAND-9
STRAND-10
STRAND-11
FC-12
A
FC-1
E
FC-2
FC-3
FC-4
E
FC-5
FC-6
FC-7
A
FC-8
FC-9
FC-10
A
FC-11

12-Cycle #1 Particle Universe
12 Base Codes per Strand = ●

12-Cycle #2 Anti-Particle Universe
12 Acceleration Codes per Strand = ○

12 FIRE CODES between Strands = ✡

(E) Emerald Kundalini Awakening releases Fire Codes 1 & 4 (A) Amethyst Kundalini Awakening releases Fire Codes 7,10 & 12
(Diagram presents simple conceptualization of DNA Strand Template orientation, not actual geometrical arrangement of wave-form structure and observable strand interrelationship.)

Activating the Tribal Shield

- The 12-dimensional **Tribal Shield** is the dormant **Species Divine Blueprint** within the **DNA Template core** of the Angelic Human races of the 4 Evolutionary Rounds. The frequencies of energy and dimensions of consciousness of the **1728 simultaneous selves** of the Personal Eternal Christos Identity can be **activated in one incarnation** through **sequential activation** of the **Fire Letter Sequences** or "**Flame Codes'** of the **Tribal Shield.**

- The "Flame Code Fire Letters" within the DNA Template are the **DNA Signet Codes** that correspond to Earth's **12 Primary Star Gates** and the **Planetary Shields** through which the planetary body manifests. The Flame Codes serve as the core foundation of the personal 12-Strand DNA Template.

- The DNA Template Flame Codes of the Tribal Shield allow for the **Primal Life Force Currents** of the Density-5 (dimensions 13-14-15) **Kee-Ra-ShA Primal Light Fields** and the **Khundaray Primal Sound Fields** from the Energy Matrix, beyond the 15-Dimensional Time Matrix, to embody more fully within the Angelic Human form.

- A **Fire Letter** is a **fixed point of consolidated frequency** that holds specific frequencies consciousness into manifest form. The **Manifestation Templates** or Divine Blueprints of **all things** manifest are built upon **intricate patterns** of interwoven Fire Letters called **Fire Letter Sequences.** The Flame Codes of the Angelic Human personal DNA Template and Species Tribal Shield are composed of the same **Fire Letter Sequences** that make up the Divine Blueprint of **Earth's Planetary Shields.**

- Activation of the Tribal Shield sets in motion simultaneous "firing" or activation of the **144 Fire Letters (12** Fire Letter **Sequences)** of the 12-Strand DNA Template, expediting the natural evolutionary process of Soul, Over-Soul and Christos Avatar Identity Integration, through progressively opening the "Trans-time DNA Template Star Gates".

- Activating the 12-Strand DNA Template allows the D-12 frequency of the **Universal "Christos" Maharata Current** to activate within and run through the Angelic Human DNA Template and into the Earth's **Planetary Shields.** Activation of the **Maharata Current** within **groups of Angelic Humans** on Earth allows for **collective anchoring** of the **D-12 Planetary Christos Divine Blueprint,** the **Shield of Aramatena,** to achieve fulfillment of Angelic Humanity's **Sacred Commission** of the **Planetary** **Christos Realignment Mission.**

The Tribal Shield- *Species D-12 Christos Divine Blueprint*

The Angelic Human 12-Strand DNA Template contains 6 Fire Letter Sequences from particle universe 12-Cycle plus 6 Fire Letter Sequences from anti-particle 12-Cycle = **12 Fire Letter Sequences** or **144 Fire Letters** from the **Two Planetary 12-Cycles** of the Cycle of the Rounds. (1 12-Cycle = 4 Evolutionary Rounds)

The **Tribal Shield** connects our personal D-12 **Maharic Shield** Divine Christos Blueprint to the greater Christos Blueprint of the Angelic Human **Species. 144 Fire** Letters (12 Fire Letter *Sequences*) of the **Tribal Shield** make up the **core program** of the Human **12-Strand DNA Template.**

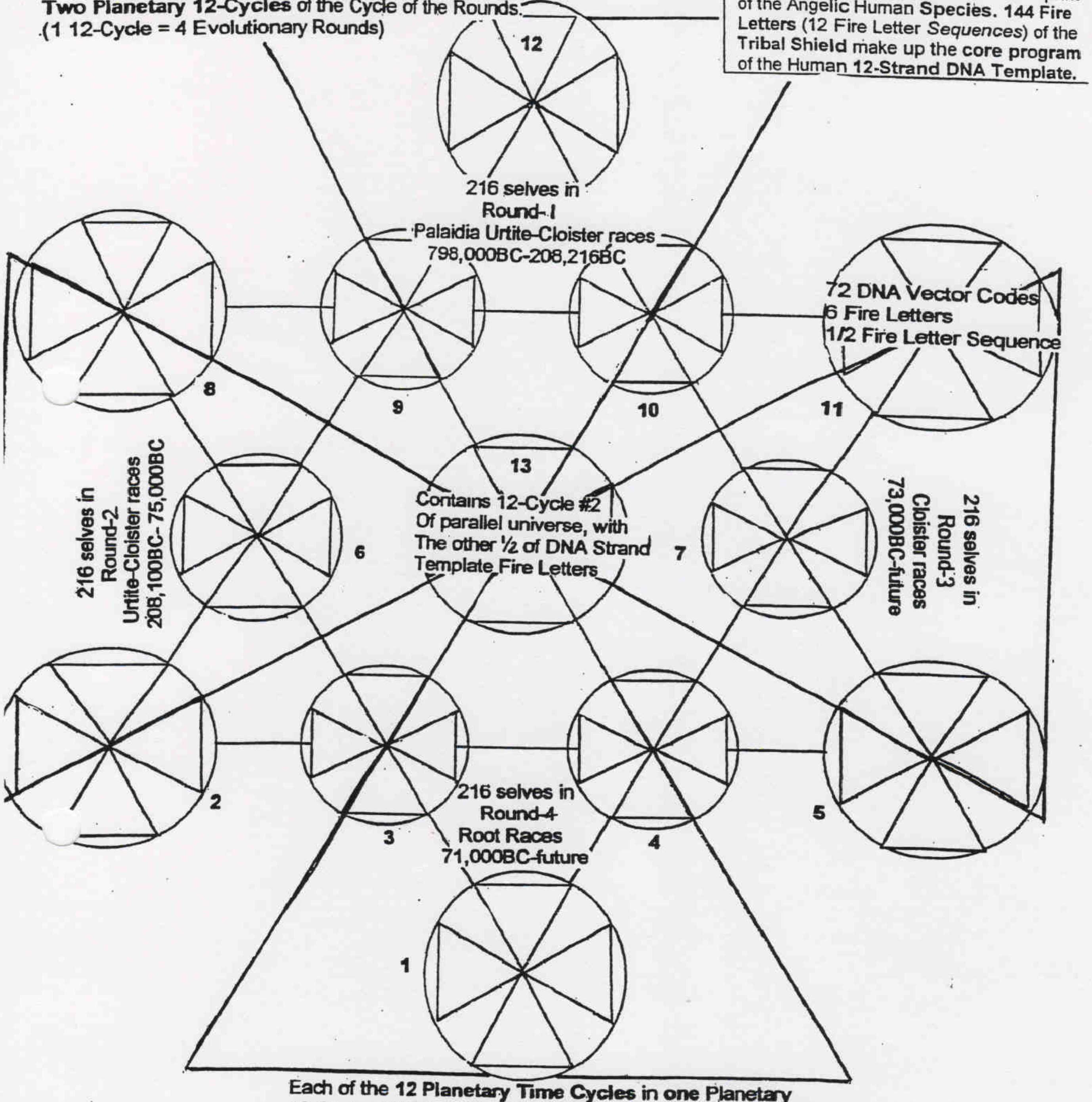
12
216 selves in
Round-1
Palaidia Urtite-Cloister races
798,000BC-208,216BC
72 DNA Vector Codes
6 Fire Letters
1/2 Fire Letter Sequence
8
9
10
11
13
Contains 12-Cycle #2
Of parallel universe, with
The other ½ of DNA Strand
Template Fire Letters
216 selves in
Round-2
Urtite-Cloister races
208,100BC- 75,000BC
6
7
216 selves in
Round-3
Cloister races
73,000BC-future
2
3
4
5
216 selves in
Round-4
Root Races
71,000BC-future
1

Each of the **12 Planetary Time Cycles** in one Planetary 12-Cycle of the **"Cycle of the Rounds" has 72 Time Vectors. 12-Strand** Angelic Humans have **one self** in each of the 72 Time Vectors of the particle universe Planetary 12-Cycle and **72 selves** in the corresponding anti-particle Planetary 12-Cycle. **72 selves x 12** Planetary **Time Cycles** = **864 selves** per one Planetary 12-Cycle **+ 864 selves** parallel universe Planetary 12-Cycle = **1728 selves** in Tribal Shield.

Tribal Shield Activation Creates Cue Zones In Earth's Planetary Shields.

Cue Zones

"Cue Zones" are locations on Earth where a small portion of Earth's D-12 Shield of Aramatena **Christos Divine Blueprint** has been re-set within Earth's Planetary Shields. During Stellar Activations Cycles, Cue Zones can be **manually set** by individuals or groups by running the **D-12 Maharata Current**, activating the **Tribal Shield** to draw in the Khunarary Primal Sound frequencies, then singing the 12-Tribes **Master Psonns Suffixes** and running the tones through the body and feet and into Earth's Planetary Shields. After running the Master Psonns Suffixes, one can then allow the **"Song of the Christos Soul"**, the tones of the personal D-12 **Maharic Shield** (the "Elohei-Elohim Within") to run into the Planetary Shields. The personal Soul Song will create a **"Back Song"** transmission of frequency from Earth's Planetary Shields, a set of healing frequencies sent back to you in response to activating your Soul Song in Earth's Planetary Shields. Once set, Cue Zones will remain for a period of **3.5 to 5 years**, depending on strength of energy originally set. They can be **recharged** using the same "setting" procedure to **last indefinitely.**

Cue Zones are **"safe zone"** areas on the Planet that are protected from interdimensional interference. They can be used to amplify effects of energy **healing applications, to "charge" objects and substances** with sub-harmonics of D-12 Christos frequency and when used with the personal Temporary Maharic Seal, to get clear, protected **"meditation space"** free from interdimensional interference. Cue Zones are "keyed" to Earth's Planetary **Templar "Cue Sites"**, the activation sites for Earth's 12 Primary Star Gates, which exist on regions of Earth that correspond to the 12 Primary Star Gates of **Inner Earth.**

TECHNIQUE # 6: THE MAHARIC QUICK SEAL

1 Work with a 24 pointed, 3 Dimensional Star (Heirophant) and be aware that you will be working with beings known as the Breneau Rishi ...

2 Begin by slowing your breathing , as you imagine or visualize a 24 point, 3 -Dimensional, Heirophant at the Pineal Gland in the center of your brain. Visualise the Heirophant surrounded by a Pale Blue Sphere of Light.

(The Blue Sphere serves as a buffer for your readiness to accept the frequencies associated with using the 24 pointed Star; the Blue Sphere holds the energy for you until your body can handle the energy. If you can handle it, the energy will simply pass through the blue sphere).

3 INHALE, as if you are going to grab the blue sphere containing the 24 point star located at the Pineal, and on the EXHALE move the 24 point star/ blue sphere, all the way down to Earth's Core. Try to hear a sound tone as the Heirophant/ Sphere hits the Planetary Shield ...

4 INHALE and draw the Heirophant/ sphere up to your Personal Maharic Shield 12" below your feet ... and EXHALE while watching the Pale Silver-Blue sphere expand out to a large sphere ... and watch as your Maharic Shield pop out as a disc, about 4' in diameter, Pale Silver with a light coating of Pale Blue.

5 Bring your attention to the center of your Maharic Shield 12" below your feet ... and INHALE the 24 point Star (only) up into the Heart Chakra. EXHALE, expanding the frequencies of the 24 point Star into the Heart Chakra.

6 INHALE to grab the 24 point Star, and EXHALE, pushing the Star up into the 14th Chakra 36" above the head. Imagine that the 24 point Star is spinning in a clockwise direction in the 14th Chakra. As you do this, try to feel/ sense the energy around you, just a few inches out from your body ...

7 Now, put your attention into your Maharic Shield 12" below your feet, take a couple of relaxing breaths and INHALE deeply pulling the Pale Silver and Blue energy up, as if you are pulling on your Maharic Seal to latch it on to your 14th Chakra. Try and feel the sensations just a few inches from your body now – feel the difference if you can. This procedure gives you a Quick Seal.

8 Now envision the cord of Pale Silver Maharic Light that would normally come up with your Shield and focus on the Earth's Core. Begin drawing energy all the way up the 4" diameter cord and into the body. INHALE the energy up into the Heart Center, and expand it there. Begin to form a pale Silver ball of Maharic frequency.

On each EXHALE send your energy down to the Earth's Core to bring up another load of Maharic frequency, up from the Planetary Shields, expand it into the Heart center and repeat several times. Notice that the Cord grows larger as you do this, expanding from approximately 4" to 6 to 8" diameter, until finally it feels like a skin around (as well as within) your body ... as you load your Astral field with Maharic Frequency.

9 Move your attention to the Pale Silver ball you have created in your Heart Center, INHALE and move it up to the 14th Chakra 36" above your head .. as you do, feel for the sheath of Maharic Frequency encasing your etheric body, close to your skin. (This Maharic/ Christos skin charges, and saturates, your etheric body – feel the peacefulness and all-knowing nature of this frequency).

Technique # 7: Ps-13 Tribal Shield Activation

The Master Key Tones

- The **12-Tribes names** were the **audible-tone translations** of the **specific Fire Letter Sequences** contained within the **Tribal Shield DNA Templates** of each Tribe. The tones of the Tribal names were used to activate the **144 Fire Letters** of the Tribal Shield in the **personal DNA Template**, providing the Angelic Humans races with the ability to **consciously regulate** the **activation level** of their **DNA Templates** and **Primal Life Force Currents.**
- The **Sound-Tone Programs** that are used to activate the DNA Template and Primal Life Force Currents are called **"The Sacred Psonns"**. The **Tribal name** was the **Master Psonn**. In running the Rainbow Roundtables, the Signet Council Regents brought their Tribal Shield "Flame Codes" into activation by **singing rounds** of the **Sacred Master Psonns.**

The 12 Sacred Master Psonns

for activating the 144 Fire Letters of the Tribal Shield Flame Codes in the 12-Strand DNA Template.

Plus **Master Tone Activation Suffix.**

12. *A-reah-Azurta*
a- RI-a-Zoor- ta- Rha

11. *Zephar-Duun-Atur*
ze-far-Doon a-Tur- Dha

10. *Ma'ah-hu-ta*
Ma-a hoo ta- Khu

9. *Yun Zu-Xen*
Yu-Un Zoo-Zen KE

8. *Chia Zhun Zan La-Yung*
ChE' ah-Zoon Yan LA-Yoong' OM

7. *Mahata-Agrah*
ME hah' ta a'g-ra OE

6. *Ramyana-Shridveta*
rah ma yah na shrid vE' Da U

5. *Ionatu-Etillah*
I O' Na too et il' a EU

4. *Nuagu Hali*
Noo ah' goo ha' LE Ka

3. *Amekasan-Etur*
a ME' ka sun e too'r DO

2. *Maahali-Bruea*
Ma a ha' LE - BrU' A EL

1. *Isutu-Esheau*
I sU' too E' shoo UR

Tribal Shield-Cue Zones Activation

Creating a Cue Zone

1. **Choose location** and **activate a personal "Maharic Quick Seal"** to "run the D-12 **Maharata Current**" in your **body.**
2. Once you can feel sensation of Maharata Current running, sing-tone 12-Tribes **Master Psonns** (names) for **Tribe-12** and **Tribe-1** to activate **Tribal Shields Flame Codes** in your DNA Template to run the "**Rainbow Ray**". Direct Pale Silver Maharata Current and Rainbow Ray pale color spectrum through body and out **soles of feet** into Earth's Planetary Shields. Tone for a while to build critical mass of frequency.
3. Use a selected combination of **3 Master Psonns Suffixes** and tone for a while, running sound through body and feet into Earth's grids. (avoid the Suffix combination of Tribe-4-7-10 used in that sequence to avert present distortions in Earth's grids that run on this mathematical coding)
4. Allow the **Christos Soul Song** to emerge from you "free style" and send the tones into Earth's grids.
5. When complete, sit quietly for a while and listen for the Earth's telepathically transmitted **"Back Song"**, drawing its gentle frequencies into your body for healing.
6. Close with singing several rounds of the **3 Master Psonns Suffixs** that you opened with. The Cue Zone you have set will extend out from your position in a radius of about **5 yards** and will last **3.5-5 years** depending on strength of energy you used in setting the Cue Zone.
7. To **amplify** strength of Cue Zone you can use the **"Merkaba Spin"** just after closing. Remain in same position, stand, the gently spin, first **clockwise** (as if you are standing on a clock face with the "6" in front of your feet, "3" at the left hand, "9" at the right and "12" directly behind you). Spin gently clockwise to bring electrical frequency down through your "**top**" Merkaba spiral. Then slow, stop and **reverse spin** to **counterclockwise**, to activate "**bottom**" Merkaba spiral", drawing frequency up from Earth's grids. Personal Merkaba Activation will accelerate the frequencies of the D-12 energies you set at the site, making the Cue Zone **stronger and more lasting.**

The Solar Salutation
Electrical Merkaba Activation

- The ancient **Mu'a** and **Lemurian** Guardian Races ran the **Maharata** and activated the **Tribal Shield** each day to conduct the **Sacred Solar and Lunar Salutation Rites** each morning and evening. The **Sacred Salutations** are **Toning Dances** that activate the personal **Merkaba Field** for sustained activation of the DNA Template and amplification of the Maharata and Universal Life Force Currents in the body.
- **Solar Salutations** accelerate the spin of the **Electrical Harmonic Merkaba Spiral** (Top, CW spin), drawing increased amounts of **Electrical Anti-particle** frequency from the Solar **Electrical Merkaba Spiral** light spectrum into the **Electrical Acceleration Codes** in the **DNA Template** to **"charge" the body** with **Electrical Anti-particle frequency.**
- Accelerated **Electrical Merkaba Spiral Spin** of **Solar Salutations** increases the amount of **Anti-particle Base-Electrical** frequency **brought into** the body's **DNA Template** and **Kathara Grid** from the Dimensional Unified Fields and increases the amount of **Anti-particle** frequency **sent out** from the body into Earth's **Planetary Shields.** Frequency passes between the Human Body and Earth's Planetary Shields via the embodied **Kathara Grid** core templates that make up the **Planetary Bio-Feed Interface System** ("PBIS")
- **Electrical Anti-particle** energy enters the body via the **"Top" Electrical** Harmonic and Dimensional **Merkaba Spirals**, the **Base-Electrical Acceleration Codes** in the DNA Template, the **Electrical Kathara Centers 2-5-8-11** of the Kathara Grid **Central Pillar**, the **Electrical aspect** of **Electro-magnetic Kathara Centers 1-3-6-9-12** on the **Right Side** of body ("Left Side Pillar" in Kathara diagram), **Electrical Chakras 2-5-8-11** and the **Electrical aspect** of **Electro-magnetic Chakras 1-3-6-9-12**.

- **Solar Salutations** (CW spin Top Electrical Merkaba Field Activation) **raise** the body's **Oscillation** (energy transmission-expansion), **lower** the body's **Vibration** (energy holding-contraction), making the body **Higher Frequency** and **More Electrical** (transmits more energy than it is drawing in).
- A **Higher Frequency Electrical Body** has:
  - **Lower Vibration**, Higher Oscillation/Frequency, **Less Matter Density**
  - More activated DNA Strands, more chromosomes, embodiment of more levels of consciousness.
  - **Slower** heart, breathing and metabolic rhythms (energy expansion and contraction rates),thus requires less food (burns fuel slower).
  - Greater **longevity** and **less sleep** (body can synthesize natural higher frequency energy and Stream of Consciousness for longer periods without taxing.)
  - Greater **Silica quantity** in **Silica-to-Carbon Ratio** of elemental composition and **larger size** (more space between cells; more expanded matter).
  - Perception of **expanded Space** and **slower Time** movement.
  - Greater **heat release**, thus **cooler** internal temperature.
  - Natural attribute of **transmitting** greater amounts of energy thrust into the Planetary Shields than it is drawing in/receiving from them. **Energy "giving".**

- The **Solar Salutations** "charge" the **Top Electrical CW Merkaba Spiral** to maintain its natural cycle of **33(⅓) rotations per 1 "trillionth" of a nano-second"**, through which the **Base-Electrical** systems of the body are naturally maintained.

The Lunar Salutation
Magnetic Merkaba Activation

- **Lunar Salutations** accelerate the spin of the **Magnetic Harmonic Merkaba Spiral** (Bottom, CCW spin), drawing increased amounts of **Magnetic Particle** frequency from the **Lunar Magnetic Merkaba Spiral** reflected light spectrum and Earth's Planetary Shields into the **Magnetic Base Codes** in the **DNA Template** to **"charge" the body** with **Magnetic Particle frequency.**
- Accelerated **Magnetic Merkaba Spiral Spin** of **Lunar Salutations** increases the amount of **Particle Base-Magnetic** frequency **brought into** the body's **DNA Template** and **Kathara Grid** from the **Earth's Planetary Shields** and increases the amount of **Particle** frequency **sent out** from the body into Dimensional Unified Fields.
- **Particle Base-magnetic** energy enters the body via the **"Bottom" Magnetic** Harmonic and Dimensional **Merkaba Spirals**, the **Base-Magnetic Base Codes** in the DNA Template, the **Magnetic Kathara Centers 4-7-10** of the Kathara Grid on the **Left Side** of the body ("Right Side Pillar" on Kathara Diagram), the **Magnetic aspect** of **Electro-magnetic Kathara Centers 1-3-6-9-12** on the **Right Side** of body ("Left Side Pillar" in Kathara diagram), **Magnetic Chakras 4-7-10** and the **Magnetic aspect** of **Electro-magnetic Chakras 1-3-6-9-12.**
- **Lunar Salutations** (CCW spin Bottom Magnetic Merkaba Field Activation) **lower** the body's **Oscillation** (energy transmission-expansion), **raise** the body's **Vibration** (energy holding-contraction), making the body **Lower Frequency** and **More Magnetic** (draws in more energy than it is transmitting out).

- A **Lower Frequency Magnetic Body** has:
  - **Higher Vibration**, Lower Oscillation/Frequency, Greater **Matter Density**
  - **Fewer** activated DNA Strands, fewer chromosomes, embodiment of fewer levels of consciousness.
  - **Faster** heart, breathing and metabolic rhythms (energy expansion and contraction rates), thus requires **more food** (burns fuel **faster**).
  - Lesser **longevity** and needs more **sleep** (body can synthesize natural higher frequency energy and Stream of Consciousness for **shorter periods**; taxes more easily)
  - Greater **Carbon quantity** in **Silica-to-Carbon Ratio** of elemental composition and **smaller size** (less space between cells; more contracted matter).
  - Perception of **contracted or more confined Space** and **faster Time** movement.
  - Less **heat release**, thus **warmer** internal temperature.
  - Natural characteristic of **drawing** greater amounts of energy thrust from the Planetary Shields than it is transmitting into them. **Energy "taking"**.

The **Lunar Salutations** "charge" the **Bottom Magnetic CCW Merkaba Spiral** to maintain its natural cycle of **11⅔ rotations per 1 "trillionth" of a nano-second"**, through which the **Base-Magnetic** systems of the body are naturally maintained.

The **natural ratio** between the **Top-CW-Electrical-"Male" Harmonic Merkaba Field** and the **Bottom-CCW-Magnetic-"Female" Harmonic Merkaba Field** is **Electrical-Top 33⅓CW** rotations per 1 "trillionth" of a nano-second and **Magnetic-Bottom 11⅔CCW** rotations per 1 "trillionth" of a nano-second.

The Natural Christos Merkaba Vehicle & Personal Shields

© 2000 Ashayana Deane

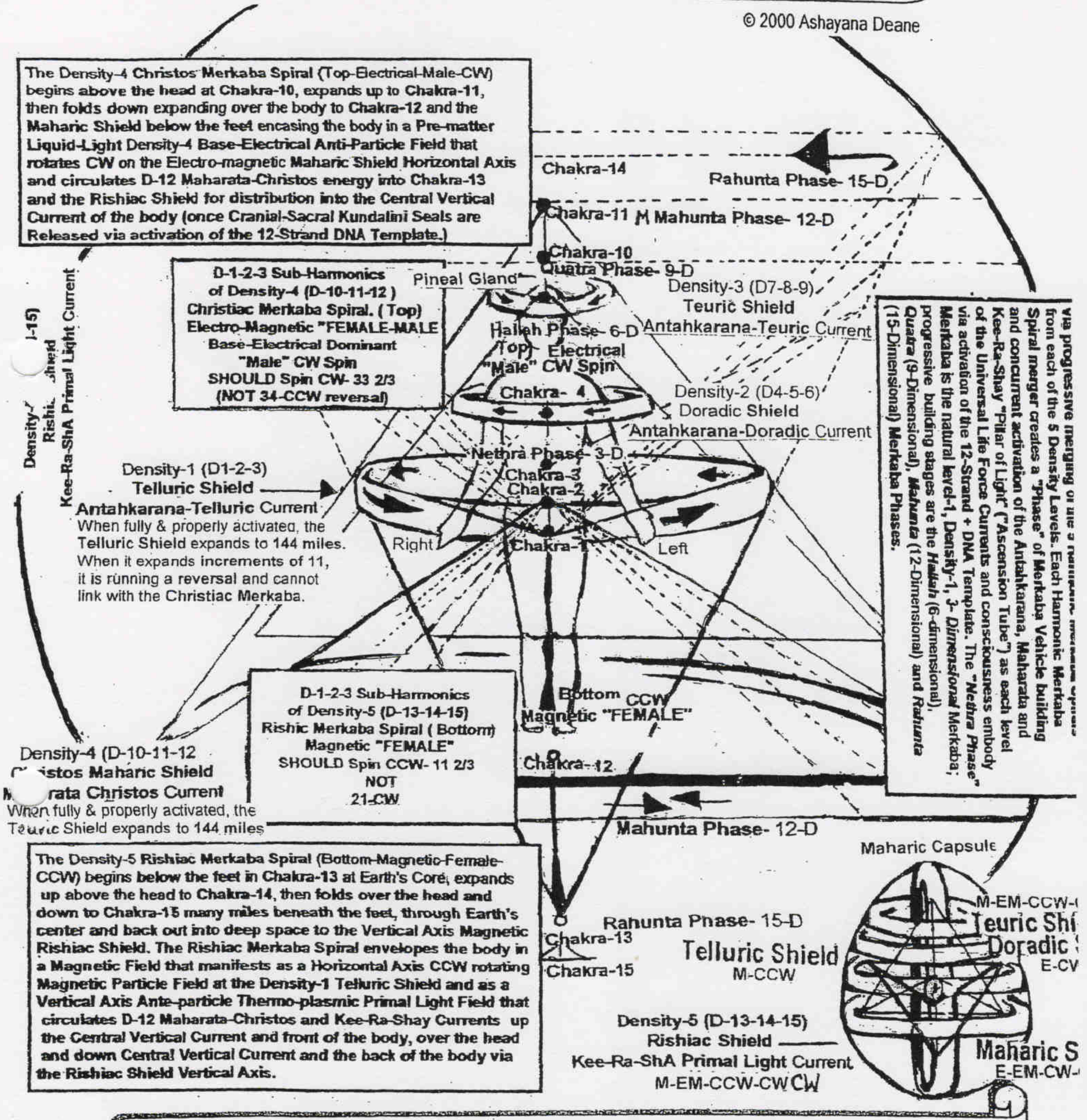

The Density-4 Christos Merkaba Spiral (Top-Electrical-Male-CW) begins above the head at Chakra-10, expands up to Chakra-11, then folds down expanding over the body to Chakra-12 and the Maharic Shield below the feet encasing the body in a Pre-matter Liquid-Light Density-4 Base-Electrical Anti-Particle Field that rotates CW on the Electro-magnetic Maharic Shield Horizontal Axis and circulates D-12 Maharata-Christos energy into Chakra-13 and the Rishiac Shield for distribution into the Central Vertical Current of the body (once Cranial-Sacral Kundalini Seals are Released via activation of the 12-Strand DNA Template.)

D-1-2-3 Sub-Harmonics
of Density-4 (D-10-11-12)
Christiac Merkaba Spiral. (Top)
Electro-Magnetic "FEMALE-MALE
Base-Electrical Dominant
"Male" CW Spin
SHOULD Spin CW- 33 2/3
(NOT 34-CCW reversal)

via progressive merging of the 5 Harmonic Merkaba Spirals from each of the 5 Density Levels. Each Harmonic Merkaba Spiral merger creates a "Phase" of Merkaba Vehicle building and concurrent activation of the Antahkarana, Maharata and Kee-Ra-Shay "Pillar of Light" ("Ascension Tube") as each level of the Universal Life Force Currents and consciousness embody via activation of the 12-Strand + DNA Template. The "*Nethra Phase*" Merkaba is the natural level-1, Density-1, 3- *Dimensional* Merkaba; progressive building stages are the *Hallah* (6-dimensional), *Quatra* (9-Dimensional), *Mahunta* (12-Dimensional) and *Rahunta* (15-Dimensional) Merkaba Phases.

Density-1 (D1-2-3)
Telluric Shield
Antahkarana-Telluric Current
When fully & properly activated, the Telluric Shield expands to 144 miles. When it expands increments of 11, it is running a reversal and cannot link with the Christiac Merkaba.

D-1-2-3 Sub-Harmonics
of Density-5 (D-13-14-15)
Rishic Merkaba Spiral (Bottom)
Magnetic "FEMALE"
SHOULD Spin CCW- 11 2/3
NOT
21-CW

Density-4 (D-10-11-12
C istos Maharic Shield
M rata Christos Current
When fully & properly activated, the Teuric Shield expands to 144 miles

The Density-5 Rishiac Merkaba Spiral (Bottom-Magnetic-Female-CCW) begins below the feet in Chakra-13 at Earth's Core, expands up above the head to Chakra-14, then folds over the head and down to Chakra-15 many miles beneath the feet, through Earth's center and back out into deep space to the Vertical Axis Magnetic Rishiac Shield. The Rishiac Merkaba Spiral envelopes the body in a Magnetic Field that manifests as a Horizontal Axis CCW rotating Magnetic Particle Field at the Density-1 Telluric Shield and as a Vertical Axis Ante-particle Thermo-plasmic Primal Light Field that circulates D-12 Maharata-Christos and Kee-Ra-Shay Currents up the Central Vertical Current and front of the body, over the head and down Central Vertical Current and the back of the body via the Rishiac Shield Vertical Axis.

Density-5 (D-13-14-15)
Rishiac Shield
Kee-Ra-ShA Primal Light Current
M-EM-CCW-CW CW

Density-
-15)
Rishi Shield
Kee-Ra-ShA Primal Light Current
Chakra-14
Rahunta Phase- 15-D
Chakra-11
M Mahunta Phase- 12-D
Chakra-10
Quatra Phase- 9-D
Pineal Gland
Density-3 (D7-8-9)
Teuric Shield
Antahkarana-Teuric Current
Hallah Phase- 6-D
Top Electrical
"Male" CW Spin
Chakra- 4
Density-2 (D4-5-6)
Doradic Shield
Antahkarana-Doradic Current
Nethra Phase- 3-D
Chakra-3
Chakra-2
Chakra-1
Right
Left
Bottom CCW
Magnetic "FEMALE"
Chakra-12
Mahunta Phase- 12-D
Rahunta Phase- 15-D
Chakra-13
Chakra-15
Telluric Shield
M-CCW
Maharic Capsule
M-EM-CCW-
Teuric Sh
Doradic
E-CV
Maharic S
E-EM-CW-

The NATURAL Spin Ratio between the Top (Male-Electrical-CW-Anti-particle) Density-4 Christiac Merkaba Spiral and the Bottom (Female-Magnetic-CCW-Particle) Density-5 Rishiac Merkaba Spiral is 33 1/3-CW-Male-Top to 11 2/3-CCW-Female-Bottom. The 33 1/3-CW/11 2/3-CCW Spin Ratio creates the NATURAL ratio of 33 2/3-Parts-Electrical Anti-Particles to 11 2/3-Parts Magnetic Particles, (or 33 1/3 energy expansion/oscillation/thrust to 11 2/3 energy contraction/vibration/draw) within earth's Density-1 Matter Base, the organic Divine Blueprint of Density-1. This Christiac Internal Merkaba Vehicle Spin Ratio creates a self-sustaining, Eternal, Matter-base that is MORE ELECTRICAL than Magnetic, which perpetually re-fuels itself via natural circulation of the *Maharata*, *Kee-Ra-ShA* and *Khundaray* Universal Life Force Currents.

Merkaba Phases – Building the Merkaba Vehicle

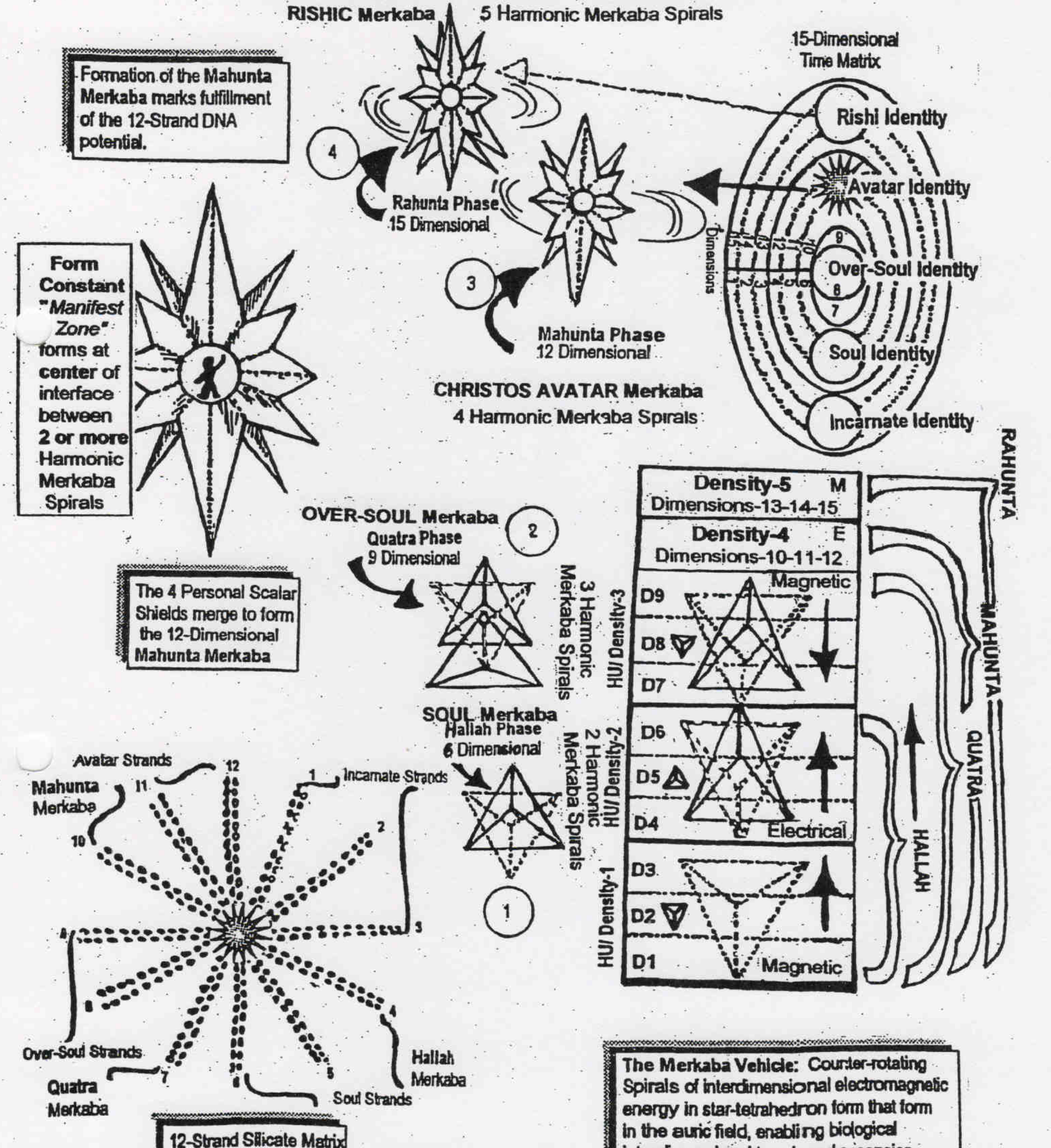
RISHIC Merkaba
5 Harmonic Merkaba Spirals
15-Dimensional Time Matrix
Rishi Identity
Avatar Identity
Over-Soul Identity
Soul Identity
Incarnate Identity
Dimensions
9
8
7
4
Rahunta Phase
15 Dimensional
3
Mahunta Phase
12 Dimensional
CHRISTOS AVATAR Merkaba
4 Harmonic Merkaba Spirals
Form Constant "Manifest Zone" forms at center of interface between 2 or more Harmonic Merkaba Spirals
OVER-SOUL Merkaba
2
Quatra Phase
9 Dimensional
3 Harmonic Merkaba Spirals
SOUL Merkaba
Hallah Phase
6 Dimensional
2 Harmonic Merkaba Spirals
1
Density-5 M
Dimensions-13-14-15
Density-4 E
Dimensions-10-11-12
Magnetic
D9
D8
D7
HU/ Density-3
D6
D5
D4
Electrical
HU/ Density-2
D3
D2
D1
Magnetic
HU/ Density-1
RAHUNTA
MAHUNTA
QUATRA
HALLAH
Avatar Strands
Mahunta Merkaba
Incarnate Strands
12
11
10
1
2
3
4
5
6
7
8
9
Over-Soul Strands
Quatra Merkaba
Soul Strands
Hallah Merkaba

Formation of the Mahunta Merkaba marks fulfillment of the 12-Strand DNA potential.

The 4 Personal Scalar Shields merge to form the 12-Dimensional Mahunta Merkaba

12-Strand Silicate Matrix DNA Diagram shows Strand relationships

The Merkaba Vehicle: Counter-rotating Spirals of interdimensional electromagnetic energy in star-tetrahedron form that form in the auric field, enabling biological interdimensional travel and ascension.

Each dimension of the Human Body has *one* **Dimensional Merkaba Field** (EM set). The *three* **Dimensional Merkaba Fields** in *one* **Density Level** form *one* **Harmonic Merkaba Spiral** in each of **5 Density Levels** that **merge** to form the **Merkaba** Star Gate Transit (Ascension) **Vehicle.**

Technique # 8: Merkaba Salutations

Healing the Spin & Balancing Particle/ Anti-Particle Ratios

33 1/3-CW/11 2/3-CCW Christos Merkaba Activation

***Note:** Discern **CW or CCW direction of Merkaba Spin** by imagining that you are standing, with arms stretched out to the sides, in the center of a round 12-digit clock-face. The **FRONT** of your body faces the **"6"-digit, "12"-digit** is positioned **behind you,** the **"3"-digit** is positioned at your **LEFT ARM** and the **"9"-digit** is positioned at your RIGHT ARM. **CLOCKWISE "CW" SPIN:** Hold LEFT **"3-Arm"** firmly out to the side of the body like a **"clock hand"** and begin **Spin** toward FRONT, moving the **body** with the **LEFT arm** toward the **"6"-digit** position in **FRONT** of your feet. COUNTER-CLOCKWISE "CCW" SPIN: Hold RIGHT **"9-Arm"** firmly out to the side of the body like a **"clock hand"**. and begin **Spin** toward FRONT, moving the **body** with the **RIGHT arm** toward the **"6"-digit** position in **FRONT** of your feet. The **TOP-ELECTRICAL-MALE** Merkaba Spiral **Christos Spin** is always **33 1/3-CW** rotation. The **BOTTOM-MAGNETIC-FEMALE** Merkaba Spiral **Christos Spin** is always **11 2/3-CCW** rotation.

Solar Salutation Rites: (best time *Dawn's First Light* & *12:00 Noon*)
begin with *Top-Electrical-CW-"Male"* **Merkaba Spiral** activation, followed by *Bottom-Magnetic-CCW-"Female"* Merkaba Spiral
Use Tribe-12 Activation Suffix **"RHA"** as ***Initiating Tone***
Lunar Salutation Rites: (best time *12:00 Midnight* to *3:00AM*)
begin with ***Bottom-Magnetic-CCW-"Female"*** **Merkaba Spiral** activation, followed by *Top-Electrical-CW-"Male"* Merkaba Spiral
Use ribe-1 Activation Suffix **"UR"** as ***Initiating Tone***

1. Activate ***Maharic Quick Seal*** Technique- 6
2. Activate ***Tribal Shield-Cue Zone*** Technique- 7
3. Stand with spine comfortably straight, breath slowly, in a relaxed manner and begin **quietly toning** the ***Initiating Tone***, continuing repetition while progressively **increasing** the **volume** and extending the Exhale Breath Tone thrust behind Initiating Tone until you can **feel the subtle energy** of Tone sound waves oscillating in the body.
4. **Activate First Merkaba Spiral** (Top or Bottom Spiral first, depending on Solar or Lunar Salutation) by holding **arms out to the side**, parallel with the ground, and begin to **slowly SPIN** in the direction (CW or CCW) natural to the *** Christos Spin** of the Merkaba Spiral you are activating. Continue to **Spin AND Tone Initiating Tone** while progressively **increasing rate of Spin and Toning.** To **prevent dizziness,** hold one arm out to front with thumb positioned UP, and **focus your eyes** on the **THUMB** rather than the spinning scenery beyond. Do not Spin faster than comfort level. You may **modulate toning in rounds of 3,** between the **Initiating Tone** and toning the **Christos Spin Rate Number** (**"33 1/3"** for Electrical-Top Spiral, **"11 2/3"** for Magnetic-Bottom spiral), if desired.
5. When energy in body feels **"charged"** (a "doneness" intuitive feeling), **slow spin** to a **gentle STOP**, while **Toning 3 Times,** the "***Um Shaddai Ur'-A ah-Khum' Tun***" Lyra-**Halls of Amorea** D-12 activation sequence, to **amplify** the Merkaba Spiral "charge. **Stand STILL** and breathe quietly for a few moments while Spin stabilizes, then **repeat STEPS 3-5** to **activate other Merkaba Spiral.**
6. **End Salutation Rite** with full **3 Rounds** (minimum) **Toning** of the "***Ariea Khum Nar-A***"" (Song of Orion D-8 activation sequence) **Part-1:Verses 1 and 2,** to be sung while **Walking 2 Circles**, one **CW**, one **CCW**, beginning Circle Walk with Spin direction of **First Merkaba Spiral** activation, followed by a **reverse-direction** Circle, then **Standing STILL** for a few moments to allow the **Christos Maharata Current** to run between your body and Earth's Planetary Shields.

The A*r-IE- a Khum –Nar' A*
The Song of Orion

Part 1 : the Resonant Tone Syllable Sequence (primary melody)

Verse 1: OM A Da E, Patum ah
OM A Da E, Patum ah
Patum ah A' DE E - Tah, A khum-na
OM A Da E, Patum ah

Verse 2: TrA TE La E' Durum Na
TrA TE La E' Durum Na
Durum' ah A' Khem - Tah, A khum -na
OM A Da E, Patum ah

Part 2 : the Base Tone Syllable Sequence (accompaniment 1)

OM A Da E, Patum ah - Lines 1, 2 & 4 Verse 1 & Line 4 Verse 2
TrA TE La E' Durum Na - Lines 1 & 2 Verse 2
UM ah OM, A
UM ah OM, A
Patum ah A' DE E – Tah, A khum- na
Durum ah A' Khem – Tah A khum-na
DO RAY' TU Ra"
OM A Da E, Patum ah
UM ah OM, A

Part 3: the Over Tone Syllable Sequence (accompaniment 2)

OM A Da E, Patum ah – Lines 1 & 2 Verse 1
TrA TE La E' Durum Na – Lines 1 & 2 Verse 2
BE HE Tah OM, A
BE HE Tah OM, A

Patum ah A DE E – Tah, A Khum –na – Line 3 Verse1
Durum ah A Khem- Tah, A Khum - na - Line 3 Verse 2
Thra Khem' Na, A
OM A Da E, Patum ah – Line 4 Verses 1 & 2

KhrA MiM, Tah A

The fastest means of naturally activating the Personal 12-Strand DNA Template is :

1. **Tribal Shield Activation:** Activating the 144 Fire Letters of the Tribal Shield & Sacred Salutations Merkaba Activations.

2. **Emerald and Amethyst Awakening 2 Masters Kundalini Activations:** Releasing key **Fire Codes** between the DNA Strand Templates to allow for **expedited DNA Strand Braiding.**

3. **DNA Template Bio-Regenesis:** Progressive use of internally directed DNA Template Bio-Regenesis technologies for progressive purging of Strand Template mutations, restoration of the natural D-12 Christos Divine Blueprint of the 12-Strand DNA Template and expedited, accelerated activation of the DNA Strand Template Base Codes and Acceleration Codes.

4. **Master Key Codes:** Use of the personal scalar-wave-guide Symbol Code Programs that correspond directly to the core programming of the personal D-12 Maharic Shield for further sequential expedition of Strand Template Activation.

5. **Melchizedek Cloister Level-3 Regent Ordination:** Transmission of the Kee-Ra-ShA Primal Light and Khundaray Primal Sound Field **"Rainbow Ray Current"** directly into the personal **DNA Template** via direct **Chakra Induction**. Sufficient Rainbow Ray frequencies to provide a level-3 Regent Ordination can be transmitted by a Melchizedek Cloister **level-5 Elder Consummate** or **Level-6 Eckar** Indigo Child-Type-1 Maji Grail Line Angelic Humans, who have the frequencies of the Primal Creation Currents integrated into their DNA Template from birth. **Expedites all of the above** while provided additional **bio-energetic field support** for more **stable cycles** of DNA Template activation.

The Khundaray RRT Activation Sequence

A progression of **manually initiated** *Kundalini and DNA Template Activations* that systematically induce *temporary release* the **Cranial-Sacral Kundalini Seals** (Pineal Gland & Tail Bone), in their **natural Release-Sequence,** allowing for temporary and ***progressive*** embodiment (nano-second burst of frequency) of the *Primal Universal Life Force Currents* via **Syncopated Activation** of the **12-Strand DNA Template**.

The ***Khundaray RRT ("K-RRT") Activation Sequence*** allows the Density-4 (D-10-11-12 Pre-matter) ***Maharata Christos Current***, Density-5 (D-13-14-15) ***Kee-Ra-ShA Primal Light Currents*** (Ante-matter *Blue, Gold* and *Violet* "Flames") and the ***Khundaray Primal Sound Currents*** (*Eckatic, Polaric* & *Triadic* Columnar Standing Waves) from *Source* to run through the DNA Template and body for accelerated ***TRUE 12-Strand or higher DNA Template Activation** and expedited human Spiritual Integration.

For the human body to **hold and transmit** a **minimum critical mass** of Kee-Ra-ShA and Khundaray frequency to **access the Planetary Shields** and **Star Gates** for ***Masters (Universal) K-RRTs,*** the ***K-RRT Activation Sequence*** must be conducted **4 to 24 hours** prior to the K-RRT. The *K-RRT Activation Sequence* must also be run **2-24 hours** prior to the ***Amethyst Awakening Masters-2 Kundalini Activation*** to prepare the body for the Amethyst Awakening frequencies. The *Amethyst Awakening* (Field Technique-6) can also be run **just before** a K-RRT (following prior *K-RRT Activation Sequence*) to **greatly amplify the power** of the K-RRT and Signet Council group.

There are **5 Specific Operations of Energy Management** within the ***K-RRT Activation Sequence***:

1. **Activating the Doradic Phase-1 Current**
2. **Awakening the KA**
3. **The Emerald Awakening**
4. **Activating the Kee-Ra-ShA**
5. **Activating the RA Center-Anchoring the Rainbow Bridge**

The Emerald & Amethyst Awakenings

Masters Kundalini-DNA Template Activations

For Accelerated ***DNA Template Activation*** and ***Spiritual Integration***, Running **K-RRTs** and enhanced ***Immune Support***

THE KUNDALINI LIFE-CURRENTS AND THE CRANIAL-SACRAL SEALS.

The human body is fuelled by ENERGETIC LIFE-FORCE CURRENT FROM SOURCE. In contemporary human anatomy, the only life-force currents running fully through the embodied Kathara Grid and body are those corresponding to the SINGLE-VECTOR time cycle of the Density-1 TELLURIC SHIELD. Each of the 4 HOVA BODIES and their corresponding SCALAR SHIELDS, DNA STRAND TEMPLATES, AXI-A-TONAL LINES, DIMENSIONS OF CONSCIOUSNESS AND MERKABA FIELD LEVELS process the frequencies of one TRIADIC CURRENT – a 3-dimensional, Bi-polar life-force current. The life-force currents corresponding to the first 9 dimension and first 3 Hova Bodies, Mater Densities and Shields are referred to as the KUNDALINI ENERGIES. Each KUNDALINI CURRENT is a TRIADIC CURRENT – a set of 3 dimensional currents that function together and are separated by Magnetic Repulsion Zones within the embodied Kathara Grid. The 3 embodied Kundalini Currents are the TELLURIC, DORADIC and TEURIC CURRENTS .

Each of the 3 Kundalini Currents is spiralled within the regions of the 1st 8 CELLS of human conception, located in the TAILBONE at the Base of the Spine. (The incarnating consciousness anchors its identity and Christos Manifestation Template within the FIRST 8 CELLS, to initiate the FETAL INTEGRATION PROCESS.) The first Kundalini Current, the TELLURIC CURRENT that corresponds to the DENSITY-1 Nada Hova Body, is active within the human form from birth, enabling 3 dimensions of consciousness awareness, in the form of scalar frequency, to embody within the human organism. When the 2 dormant Kundalini Currents in the tailbone, the DORADIC and TEURIC CURRENTS, are activated, the frequencies of consciousness from dimensions 4-9 progressively awaken within the body, activating the dormant 4-9 strands of the 12-strand DNA Template. The 3 Primary Kundalini Spirals each contain 3 single-dimension currents, and are controlled by the corresponding SIGNET STAR CRYSTAL SEALS in the Kathara Level-2 Crystal Seals Grid. The Signet Seals create polarization between dimensionalized sets of scalar waves within the Scalar Shields, that manifest within the Kathara Level-2 Grid, First Eight Cells and DNA Template as FREQUENCY SEALS or blockages within the TAILBONE (Sacral region) and the PINEAL GLAND in the center of the brain – the CRANIAL-SACRAL SEALS.

When the Cranial-Sacral Seals are released, through activation of the Signet Star Crystal Seals that govern them, the dormant Kundalini Currents come to life within the body, sending progressively higher-frequency energy through the CENTRAL VERTICAL CURRENT of the body. Activation of the 3 Kundalini Currents progressively creates integration of the higher dimensions of consciousness and de-densification of molecular structure. When ALL of the frequencies within the 3 embodied Kundalini Currents are activated, the 4th KUNDALINI CURRENT- the PRIMAL TRIADIC CURRENT, the KEE-Ra-ShA', is activated. The Frequency Seal on the KEE-Ra-ShA' is located in the 12th Chakra, 6" below the feet. Activation of the KEE-Ra-ShA' anchors the 12th-Dimensional HYDROPLASMIC BEAM within the body, initiating the process of CELLULAR TRANSMUTATION and intentional de-manifestation and Dimensional Ascension of the human biological form.239

THE EMERALD AWAKENING AND THE KEE-Ra-ShA'

The Emerald Awakening Masters Kundalini Activation is one of two inter-harmonic energetic activations of previously dormant frequency within the body that sets in motion simultaneous activation of the 12 MAGNETIC BASE TONES within the FIRE CODES between the 12 Strands of the SILICATE MATRIX DNA Template. Activation of the 12 Fire Code Base Tones begins activation of the 3 PRIMAL CREATION FREQUENCIES – the KEE-Ra-ShA' – within the 4 Scalar Shields of the PERSONAL INNER TEMPLAR COMPLEX. Activation of the 12 DNA Fire Codes allow for progressive natural fusion of each dimensional DNA Strand Template to occur, through which the organic potentials of CELLULAR TRANSMUTATION are re-awakened within the physical body form.

Through the natural processes of Cellular Transmutation, once fully operational within human genetic design, the carbon-based elemental biology of Density-1 molecular structure is progressively and systematically transmuted to SILICA-based biology. Transmutation to Silica-based biology allows the physical body to pass through the levels of Semi-etheric, Etheric and Pre-matter Hydroplasmic Liquid-Silica Densities of Harmonics 2-4 (Dimensions 4-12), through which the consciousness continually enters manifestation, to achieve physical mastery of the organic process of Transmutational Dimensional Ascension. To accomplish mastery of the Ascension process, the SIGNET STAR CRYSTAL SEALS with the Scalar Shields, which keep the 12 dimensionalized frequency levels of the consciousness and molecular form Phase-locked into dimensional division/ polarization, must be released. Release of the Signet Star Crystal Seals allows the frequency spectra of the higher dimensional anatomy and the Primal Creation Currents of the KEE-Ra-ShA', to enter into the Template for the physical body and dormant DNA Fire Code Templates, progressively building the Merkaba Vehicle and setting the Cellular Transmutation process in motion. The process of progressive Transmutational Dimensional Ascension represents the function of the organic genetic design and evolutionary objective of the human creation blueprint, through which the natural Immortal Christos-Angelic Form of the human being is actualised.

In the evolutionary cycles of contemporary humanity, actualization of the 12-Strand DNA potential takes place slowly in Earthly terms, as the incarnate identity experiences in single-life time-vectors, several cycles of Density-1 incarnation, spanning many thousands of years, in order to progressively activate the DNA Fire Codes within the Manifestation Template. When the human genome was operating in its organic form, this process took place in one lifetime, over a period of about 44 years, as the immortal human incarnate matured to embody the 12 dimensions of consciousness within its original Christos Manifestation Template- the MAHARIC SHIELD, actualizing its organic Angelic-human heritage and mastery over 4 Densities of manifestation within the Time Matrix.

The EMERALD AWAKENING Masters Kundalini Activation releases the Density-2 4TH – Dimensional Green Signet Star Crystal Seal and the Density-1 1st-Dimensional Red Signet Star Crystal Seal, to open the Density-1 NADA HOVA BODY and TELLURIC SHIELD, and Density-2 ALPHI HOVA BODY and DORADIC SHIELD, to receive the Omni-polar frequencies of the 12th-Dimensional MAHARIC CURRENT (Hydroplasmic Beam) from the Pre-matter Christos Template Maharic Shield. Upon entering the Telluric and Doradic Shields, the 12th-Dimensional frequencies of the Maharic Current trigger simultaneous activation of the Base Tones within the 12 Fire Codes of the DNA Template, beginning the process of DE-POLARIZTION of the 12-Strands and levels of consciousness within the human Manifestation Template. Progressive de-polarization of the 12 levels of consciousness within the human Manifestation Template sets in motion the organic process of Cellular Transmutation and potential mastery over Transmutational Dimensional Ascension. The Emerald Awakening Masters Kundalini Activation initiates REGENESIS of the original function of the human genome, releasing the Base Tones of the 3 KEE-Ra-ShA' Currents within the body via opening of 2 Signet Seals, through which the physical body begins the carbon-to-silica transmutation process and is prepared to receive the higher dimensional OVER TONE frequencies of the KEE-RA-Sha', to complete Cellular Transmutation via fusion and de-polarization of the particle (Particum) and anti-particle (Partika) aspects of the Manifestation Template. This is the natural process of UNIFIED FIELD PHYSICS by which a human being becomes "Christed".

THE AMETHYST AWAKENING AND THE KHUN-DA-RAY

The final step in fulfilling the potentials of original Angelic-human design is activation of the OVER TONE frequencies of the KEE-Ra-ShA', which allows fusion between DNA Strand Templates to complete on 12 dimensions of frequency. The final step in Transmutational Dimensional Ascension is called the AMETHYST AWAKENING MASTERS KUNDALINI ACTIVATION. Through the AMETHYST AWAKENING, the DENSITY-3 7TH-Dimensional VIOLET Signet Star Crystal Seal and 8th-Dimensional GOLD Signet Star Crystal Seal, and the DENSITY-4 11TH-Dimensional DARK SILVER Signet Star Crystal Seal are released within the body's Scalar Shields. Release of the OVER TONE SIGNET SEALS allows the OVER TONES of the KEE-Ra-ShA Creation Currents to fuse with the BASE TONES, bringing the 3 OMNI-POLAR CREATION CURRENTS (Primal Creation Currents built upon a TRINITY OF TONES, BI-POLAR + NON-POLAR energy units) together as ONE PRIMAL NON-POLAR CURRENT – the KHUNDARAY (also called the "Rainbow Bridge" or the " Rainbow Serpent".)

The KHUN-Da-rAy is the PRE-LUMINAL vibration, or STANDING WAVE OF INTERNAL SOUND emitted from ONE-SOURCE/ GOD, that corresponds to the COMBINED ORIGINAL SOURCE CURRENT from which the 3 KEE-Ra-ShA' PRIMAL LIFE-FORCE CURRENTS of the 3 non-dimensionalized levels of the ENERGY MATRIX, beyond the 15-Dimensional Time Matrix emerge. The Emerald Awakening Masters Kundalini Activation, and its resulting activation of the BASE TONES of the KEE-Ra-ShA' within the human body, is the first step in manually expediting the natural process of Transmutational Dimensional Ascension.

AWAKENING THE KA AND REMOVING INORGANIC CRYSTAL IMPLANTS

Before the Emerald Awakening Kundalini Activation can be orchestrated, the MAHARIC SHIELD and the KA must be brought out of dormancy within the human Manifestation Template. Once the 12th-Dimensional Maharic Current is released within the embodied Kathara Grid, the KA can be activated. The "KA" refers to the 4 HOVA BODIES corresponding to dimensions 1 through 12. (Nada, Alphi, Betcha and Mahara Hova Bodies.) The KA becomes activated through removal of 4 inorganic Crystal Seal IMPLANTS that block the natural function of the 2nd, 3rd and 4th DNA Strand Templates. Following he procedure for AWAKENING THE KA, the Emerald Awakening Kundalini activation can be conducted.

Technique # 12: Kee-Ra-ShA Activation Sequence

ACTIVATING DORADIC PHASE I Awakening the Flame

Preparation: Activate **Maharic Seal** Technique-6 *Maharic Quick Seal*

1 Breath slowly and deeply for several breaths while visualizing or imagining a Violet Dot on the center of the *inside* of your forehead. Make the color image as strong as you are able (if you can't yet perceive inner colors simply imagine the Violet Dot being there).

Move the Violet Dot to the center of the brain, at the **Pineal Gland,** and then upward into the 7th Chakra, finally moving it out of your head to rest on the top of your skull, at the Crown Chakra.

2 Call to mind the Heirophant Symbol (Pale Silver Merkaba Star/ Star of David). Visualise the Heirophant spinning 12" below your feet, within your Maharic Shield, and slowly INHALE.... drawing the Pale Silver Heirophant up through the Central Body Current, up to meet the Violet Dot at the 7th Crown Chakra at the top of the head.

When the Heirophant connects with the Violet Dot, visualize the Violet Dot bursting into a 4" vertical Pillar of Violet-Silver Light. Imagine the Violet-Silver Pillar running down through the body and into the Earth's Core, as well as upward into the 14th Chakra positioned 36" above your head. (This activates the Violet Ray of Transmutation).

3 Now, move your attention to the 4th Heart Chakra, and imagine a small Green Spherical Crystal – the Green Star – stationed at it's core. This is the Green star Crystal Seal that controls release of 4th Dimensional frequency – Phase I Doradic Current – into the body. INHALE, drawing Violet-Silver energy from the Violet-Silver Pillar into the Green Star.

EXHALE, expanding the Violet-Silver energy into the Green Star sphere. Repeat two more INHALE/ EXHALE breaths, drawing Violet-Silver energy from the Pillar and expanding it into the Green Star sphere.

4 At the end of the last EXHALE, imagine the Green Star POPPING, turning into a spinning Green and Violet Starburst of Light, as the Violet-Silver Pillar disappears. Spin the Green-Violet Starburst faster until it becomes a ball of Green-Violet Light, about the size of a grapefruit, spinning the center of the Heart Chakra. (The Green-Violet ball is the storehouse for your Phase I Doradic Current; it's called the Doradic Sphere).

INHALE and draw Pale Silver energy from your Maharic Shield, through the central Body Current, and into the center of the Doradic Sphere at the heart Chakra ... encasing the sphere in Pale Silver Light.

AWAKENING the KA

1 Bring your attention back to the DORADIC SPHERE at the HEART CENTRE...

2 INHALE pale silver energy up from the EARTH CORE to fill & expand the DORADIC SPHERE; breath slow and deep 2-3 times ...

3 Now, put the tips of your fingers together, and feel the 'fuzziness' of the energy flow...

4 Next, move the fingertips to the tail bone... INHALE energy from the Maharic Shield into the DORADIC SPHERE

Continued Next Page

As you EXHALE, push PALE SILVER energy down to the finger tips and into the tail bone............... REPEAT 2-3 times and then remove the fingers.

5 Now, bring your attention to the 2nd CHAKRA (SACRAL) .....feel or visualize or just know ... a DOUBLE TERMINATED QUARTZ CRYSTAL (a crystal with a point at both ends) is beginning to "push out" of the front of your 2nd Chakra, the Sacral center.

6 Focus your attention on this, try to feel the crystal pushing.... use your hands to help ease it forward, and out.

7 There is a smaller double terminated crystal in the 3rd (Solar Plexus) Chakra; this will release on it's own because the currents have activated it to do so. Help it out with your fingers.

8 In the Heart Chakra, imagine a little sphere of energy, colored green. Bring this little green sphere more fully into focus ... and you will notice a little black cube of energy inside it ...look more closely again, and notice a smaller white cube inside the black one.

9 As you focus on the two little cubes, watch them merge together .. feel a ripple of vibration as they do, and then ... notice a little puff of smoke as they both disappear.

Now, the body is activated to open the 4th and 1st Crystal Seals; you are ready for the EMERALD AWAKENING ...

1 Now we will open the GREEN STAR CRYSTAL SEAL ....So, keeping your awareness on the Doradic Sphere, the GREEN-VIOLET- SILVER ball at the Heart Center.........INHALE,

2 and ... on the EXHALE visualize or imagine pushing a cord of GREEN-VIOLET-SILVER Doradic Current down into the tail bone... to the 1st Cell ...... at it's very tip .

3 Breathe easy ... and take a moment to visualize the 1st Cell at the tip of your tail bone.... as a WHITE-SILVER SPHERE...

5 Now within the WHITE SILVER SPHERE, visualize, a GREEN SPHERE ... and then visualize a RED SPHERE inside the GREEN SPHERE ...

6 Bring your attention back to the SPHERE of GREEN-VIOLET-SILVER energy in your Heart Center ...

and on the next EXHALE visualize, or imagine, sending a cord of GREEN-VIOLET SILVER energy all the way down your central vertical column and into the GREEN SPHERE in the 1st Cell at the tip of the tail bone.

7 AS you INHALE pull PALE SILVER LIGHT from your personal Maharic Shield, the platform-disc 12" below your feet, up into the tail bone and into the GREEN SPHERE.

8 As you do this, visualize, the GREEN SPHERE 'pop' into a STARBURST of PALE GREEN-SILVER LIGHT.

Continued Next Page

This has opened the kundalini seal – your WHITE-SILVER & RED Spheres REMAIN as you left them earlier ....Now, we will work to open the **RED STAR CRYSTAL SEAL....**

---

9 Return your awareness to the Doradic Sphere at the Heart Center....INHALE, and then on the EXHALE push a cord of GREEN-VIOLET-SILVER Doradic Current down to the 1st Cell at the tip of the tail bone, ...

10 INHALE and draw PALE SILVER energy up from the personal Maharic Shield, the platform-disc positioned 12" below your feet, up into the tail bone, and into the RED SPHERE ... inside .... the WHITE-SILVER SPHERE.

11 AS YOU DRAW the PALE SILVER energy into the RED SPHERE visualize the RED SPHERE 'pop' into a STARBURST of PALE PINK-SILVER LIGHT...... and ...... for the next few breaths EXHALE the PALE PINK-SILVER LIGHT into the 12th Chakra, positioned 6" below your feet...

You have now opened the Kee Ra ShA(Y) Seal ... so that the Sacred Flames of the Kee Ra ShA can be activated ......

ACTIVATING THE KEE RA SHA

The Kee Ra ShA appears as 3 little flames which exist in the 12th Chakra positioned 6" below the feet. The 3 flames are PALE BLUE, PALE YELLOW & PALE VIOLET.

1 Begin by focussing deeply on the 12th Chakra, 6" below your feet, and the three little flames held there ... one is PALE BLUE, one PALE YELLOW & the third PALE VIOLET.... We'll take a few moments to do this......

If you have difficulty visualizing, just imagine that they are there.

2 On the INHALE, draw the PALE VIOLET FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

3 Next, return your attention to the 12th Chakra, 6" below your feet ... allow yourself time to connect with the little PALE YELLOW-GOLD FLAME. Breathe easy.

4 On the next INHALE draw the PALE YELLOW-GOLD FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

Again, return your attention to the 12th Chakra, 6" below your feet ... and connect with the PALE BLUE FLAME... this FLAME is the INNER ARC of the COVENANT ... Take your time; breathe easy.

5 On the next INHALE draw the PALE BLUE FLAME up from Chakra 12 ... into the central vertical column .. and all the way up ... and out through the top of the head, to the 10th Chakra ... 6" above your Crown chakra......

5D Continued Next Page

© Rev. Ashayana Deane 2000

Activating the Kee-Ra-ShA - Continued

6 Now, visualize or imagine the 3 little FLAMES together within the 10th Chakra 6" above your Crown Chakra .... in a moment we will create the Kee Ra ShA SPHERE... Breathe easy.

7 On the next INHALE draw PALE SILVER energy from your personal Maharic Shield 12" below your feet, all the way up your central vertical column, up into the 10th Chakra......

EXHALE gently and INHALE once more, drawing more PALE SILVER energy up into the 10th Chakra ... visualize or imagine the 3 little flames enclosed within a SPHERE of PALE SILVER LIGHT...

8 Breathe easy. On the next EXHALE move the Kee Ra ShA SPHERE all the way down through the central vertical column, out through the 12th Chakra 6" below your feet, and all the way down to the 13th Chakra at EARTH'S CORE.

As the Kee Ra ShA enters EARTH'S CORE listen for sound tones w/ inner ear...

These TONES represent the AMENTI "PASS KEY" ... which opens your personal GRU-AL POINT (#1) - the 1st Cell in the Tail Bone .... releasing the KEE RA SHA FLAMES into your AURIC FIELD.

9 Breathe easy... Watch as the Kee Ra ShA FLAMES rise up your Maharic Pillar and emerge through your 1st Cell ... passing like radiating waves of PALE BLUE, PALE YELLOW & PALE VIOLET ......

10 As the waves of color move through you, bring your attention to your first 4 Chakras ... the base, the sacral, the solar plexus and the heart ... and visualize or imagine each Chakra emanate a SILVER MAHARIC SHEATH or OVAL POD, completely surrounding each one ...

This completes the EMERALD AWAKENING, the natural process by which we use Unified Field Physics to trigger our CHRISTED Selves. Before we finish we must first establish the link to the Amethyst Awakening, the process of activation and sheathing of the higher Chakras.

To complete the session we will now work with 8th & 11th Dimensional Energies to open DORADIC PHASE 2 & 3 CURRENT and activate TEURIC PHASE 2 CURRENT.......THIS IS CALLED **ACTIVATING THE RA CENTRE in the THYROID – and ANCHORING THE RAINBOW BRIDGE**

1 Begin by moving your attention to the area of your 3RD EYE, and focus just inside your fore-head .. a little above and behind your eye brows...... Breathe easy.

2 Now, visualize, or imagine, a DARK SILVER DOT in the position you are focussed on ... in the area of your 3RD EYE... Take a little time.

3 INHALE and move the DARK SILVER DOT into the Pineal Gland in the center of your head and then straight up through the CROWN CHAKRA until the DARK SILVER DOT sits on the top of your skull.

Continued Next Page

© Rev. Ashayana Deane, 2000

Activating the RA Center-Rainbow Bridge - Continued

4 On the next INHALE, pull PALE SILVER ENERGY up from your personal Maharic Shield 12" below your feet, all the way up the central vertical current and into the DARK SILVER DOT on the top of your head...

As the PALE SILVER ENERGY reaches the DARK SILVER DOT, the SILVER DOT 'pops' into a SILVER STARBURST ...

5 On the next EXHALE push the SILVER STARBURST down the central vertical channel, all the way to the tip of the tail bone ... EXPANDING the SILVER STARBURST FULLY into the 1st Cell...

6 Breathe easily and watch, as GOLD-SILVER-CURRENT emerges from the tailbone .. and begins to flow up through each Chakra ..

from the 1st Chakra at the base of the spine ... up through the Sacral ... the Solar Plexus ... all the way up to the Crown Chakra ... we'll take a little time for this ....

From the Crown Chakra the GOLD-SILVER-CURRENT continues upward out of the top of your head ... up through the 10th Chakra 6" above your head ... the 11th Chakra 18" above your head ... and finally the 14th Chakra 36" above your head.

7 Once the GOLD-SILVER-CURRENT reaches the 14th Chakra it returns to the 1st Chakra the flow of the current creates a GOLD-SILVER SHEATH throughout your central vertical channel.

8 Bring your attention to the 2nd Chakra now .. were going to create a GOLD -SILVER- BALL there, in the SEED SEAL at the center of the Chakra ... just imagine the chakra as a bow tie shape, the SEED SEAL is at the narrow point in the middle. The GOLD-SILVER BALL is called the TEURIC SPHERE ...

INHALE and draw GOLD -SILVER-CURRENT from the GOLD -SILVER-SHEATH into the center of the Chakra ... as you EXHALE create and EXPAND the BALL to surround the bow tie ...

9 Now ... INHALE and move the GOLD -SILVER-BALL up to the 1ST HARA CENTRE ... 2" BELOW the NAVEL (the 9th SILVER STAR SIGNET SEAL or TAN TIEN) ... EXHALE GENTLY allowing the BALL to EXPAND

10 On the NEXT INHALE move the GOLD-SILVER-BALL up to the SOLAR PLEXUS and EXHALE GENTLY

11 INHALE and move the GOLD-SILVER BALL up to the HEART CENTRE ...
EXHALE GENTLY ... ... INHALE ... drawing more GOLD-SILVER CURRENT into the BALL and THEN EXHALE FORCEFULLY pushing cords of GOLD SILVER CURRENT down each arm and into the PALMS ... bring your fingertips together and feel the fuzzy buzz of TEURIC PHASE 3 CURRENT

You have now activated TEURIC PHASE 3 CURRENT in your palms – and – you are now able to run CURRENT from D4 thru' D8 to transmit to the THYROID GLAND ... to activate and anchor the Rainbow Bridge in others.....

End of *Field Technique-5 The K-RRT Activation Sequence* (5A-B-C-D-E)

You are now ready for the **Amethyst Awakening *Masters Kundalini/DNA Template Activation-2*** (Field Technique-6 *Initiating the Amethyst Awakening*) **or** to participate as an active member of ***a Khundaray Rainbow Roundtable*** (K-RRT) **Signet Council** (Field Technique-4 A-B-C). **Wait at least 4 Hours before** K-RRT participation, **and no longer than 24 hours** to maintain the **appropriate level of critical mass Khundaray frequency in the body** that is need for an effective K-RRT. **© Rev. Ashayana Deane, 2000**

INITIATING THE AMETHYST AWAKENING

1 **Preparation:** Complete **Maharic Quick Seal** Technique 6 and **Activation Sequence**

2 This technique works to bring in a burst of the Khundaray, Eckatic Blue frequency ...that sets the Amethyst Awakening in motion.

3 Place your attention at the center of your brain, inside your pineal gland. Imagine that you see there a Pale Violet Spark of Light (this is the bottom of the Seal of your 7th Crown Chakra). Imagine that you are going to move this Pale Violet Spark of Light straight up into your 10th Chakra 6" above your head ... imagine that the 10th Chakra is a disc of Light, Blue/ Black – Sapphire.

4 Now see the disc of Light, that was Blue/ Black take on a Violet hue. Try to feel it there, a small horizontal disc ... On the INHALE you will move this disc of Violet-Sapphire Light up to the 24 point Silver Star at the 14th Chakra ...... and, on the EXHALE, move it down through the 10th Chakra, down through the Crown, down the Central vertical Current, down through your Personal Maharic Shield all the way down to the 13th Chakra at the Earth's Core ...

5 Notice, as you do this, a whole series of Violet-Blue Sparks flying out from the Planetary Maharic Shield, as if they are alive, as if electrical communication is taking place. Notice as this happens that the energy flowing up your Maharic Pillar from the Earth Core reflects this change as Violet-Blue energy flows slowly up toward you ... as this energy reaches your personal Maharic Shield, notice that it does the same thing as the Planetary Shields – throws out a shower of Violet-Blue sparks ... all communicating with each other about the mathematical instructions about what to do with your body, your crystal seals, your chakras ...

6 Now, move your awareness to the Pale Silver skin all round your body/ etheric body .. and be aware too of the Main Vertical Current that you brought the Pale Silver cord through ...and at the same time, on each coming INHALE, bring up the Pale Silver-Violet-Sapphire energy, simultaneously, from your Maharic Shield to infuse the Pale Silver skin all round you.

Take a few breaths to do this and watch as the 3 color flicker through you – when they reach the Heart Chakra feel them expand out ... and allow them to travel on up, all the way to the 14th Chakra ... expanding throughout every Chakra as the Pale Silver-Violet-Sapphire energy flows up, through and all around you ... See each Chakra respond with a spiraling burst of Silver-Violet-Sapphire, backward and forward through both sides of the Chakras until they all shift .. try to feel the process as it awakens all the cell's of your body.

The brain cells and brain patterns begin to respond, activating the Fire Letters, setting the grid of frequencies necessary to activate the Fire Codes 7, 10 and 12 in the DNA Template.

Your own **D-12 Christos Avatar identity level will slowly** release the ***Khundaray Activation Frequencies*** through which the Amethyst Awakening progressively unfolds in the DNA Template and body. This process can be **naturally and safely accelerated** by receiving a **Level-3 MC Regent Ordination transmission** from a qualified **Indigo Child/Eieyani *Level-4 through Level 6 MC Ordinate* Consummate, Elder Consummate** or **Eckar** that carries the necessary **Amethyst Awakening Khundaray Activating Frequencies** in their Eieyani DNA Templates.

The Merkaba Trinity Key Induction

The ***Merkaba Trinity Key Induction Technique*** begins the process of establishing a **conscious directional relationship** with various levels of your 15-Dimensional Anatomy (or that of your Client). The **Trinity Key Symbols** are **mathematical-geometrical programs** that carry the specific organic structure of **each dimensional frequency band** as it exists within the **D-12 Universal Divine Blueprint**. Imbuing the proper **Symbol Key** into the appropriate area of the personal anatomy can rapidly expedite **restoration the personal Divine Blueprint** within the **DNA Template, Chakra System** and embodied **Kathara Grid**. Use of the Trinity Master Key Symbols that correspond precisely to the natural order of each dimensional frequency band, can **intensify** all applications of **holistic and allopathic healing**, as well as progressively and systematically assist in the process of accelerating **DNA Template** healing and **activation.** Every **condition** within the body-mind-spirit system has a specific **Energy Signature** unique to its own construction. Every **dis-ease** condition within any aspect of the personal experience (physical, emotional, mental, spiritual, financial, social, etc.) is created and **held in manifestation** by a specific set of mathematical – geometrical energy relationships that are **stored** within the embodied **Kathara Grid** and **DNA Template** in the form of **Symbol Keys**. The mathematical-geometrical energy relationships that serve as the **blueprints** upon which experiential manifestation will emerge can be perceived by the conscious identity in its Symbol translation. When using the **Trinity Master Key Symbols**, you are **resetting and activating very specific portions** of the personal **Divine Blueprint** (D-12 Pre-matter Maharic Shield) into a naturally harmonious relationship with Universal Order (the organic structure of dimensionalized frequency bands). The Trinity Master Keys provided in this Technique trigger activation of very specific **Fire Letters** (Keylons) within the DNA Template, allowing the DNA Template to receive directly the **new electrical instructions** corresponding the personal Divine Blueprint. This technique can be used with Kathara Healing Techniques from Kathara Level-1, with any other healing modality or singly after activating the D-12 **Maharata Current** via **Techniue-6, *The Maharic Quick Seal*.**

It is **necessary** to initiate a **D-12 Maharaic Seal** in order to run **any Symbol Key program** into the DNA Template and Level-1 Kathara Grid; it is the strength of the D-12 Maharata Current that allows access to the deepest levels of the Personal Shields and Divine Blueprint. The **Trinity Master Key Symbols** are a **set of 3 Symbol Keys** that are to be used **in sequence.** The first two Trinity Keys, the **Initiating Keys**, are used in **every application** to open the embodied Level-1 Kathara Grid to receive new instructions. The third Trinity Master Key Symbol, the **Transmitting Key,** corresponds directly to the Divine Blueprint of **one dimensional frequency band** and its corresponding **Kathara Center, DNA Strand Template, Chakra** and related body regions, **Axia-tonal Line, Auric Field Level** and **station of consciousness**. In a **Trinity Key Induction Session**, one can choose to work on only **one dimension** (Chakra-DNA Strand Template-Kathra Center-Axia-tonal Line-Auric Level) via its Transmitting Key, or each of the **12 Transmitting Key** of the Universal Kathara can be used, **in sequence**, following induction of the two Initiating Keys. The process of using the Trinity Keys is simple; the **Chakra** that corresponds to the Transmitting Key/ dimension/DNA Strand with which you desire to work is used as the **vortex** through which the Trinity Keys are **INDUCED** into the corresponding **Kathara Center**. From the Kathara Center, the Trinity Keys electrically pass on to the **DNA Template** and from there into the **Axia-tonal Line system** and corresponding areas of the body-mind-spirit system. Use of the Trinity Key Induction Technique progressively assists in activating the personal **Merkaba Vehicle** while **amplifying and expediting restoration** of the personal Divine Blueprint.

Technique # 9: Merkaba Trinity Key Induction

1. Activate the Maharata Current via **Technique-6:** ***The Maharic Quick Seal*** and decide which Transmitting Key(s) you will use. **Trace** the **3 Trinity Symbol Keys** onto **3 small squares of paper.**

2. First, place each of the papers with the two Initiating Keys **face up on the floor**, then **stand** (barefoot or socks only) with **one foot on top of each of the papers; Initiating Key #1** under the **LEFT foot, Key #2** under the **RIGHT foot.** (Note: To run Trinity Key Induction for a Client, have them recline on their back and place the papers with Initiating Keys between your palms and the soles of Client's feet. As introduced in Kathara Level-1, use your **palm Chakras** to **transmit** the Symbols and the Pale-Silver Maharata Current, using the same action as described below, into the **Client's foot Chakras**; Key-1LEFT foot-***right hand transmission***, Key-2 RIGHT foot-***left hand transmission***.) In Self-induction, the papers with Symbols should be positioned beneath the **balls of your feet**, in the area of the main **Chakra** on the sole of each foot. Simply begin to **breathe deeply** and slowly for a few minutes. With each **INHALE** draw a **thick stream** of **Pale-Silver Maharata Current** from Earth's core through the **Central Vertical Current** of the body and out the top of the head to the **14th Chakra** 36-inches above the head. With each **EXHALE,** imagine a grapefruit sized Pale-Silver **Maharic Ball** rapidly moving down your Central Vertical Current. When the Maharic Ball reaches your 2nd-Chakra near the navel, imagine the Ball splitting into **Two Balls**, each continuing to **travel down each leg and out your feet Chakras**, through the **Symbol Keys**. Once the Two Balls of Maharic Current have passed through the Symbol Keys, imagine that they **re-merge** into the **one grapefruit sized Pale-Silver Ball** that continues to move downward into the **13th Chakra** at Earth's core. **Continue** this process until you sense a **momentary slight warmth, tingling or subtle itching** at the **balls of your feet**. These mild, **momentary sensations,** called the **CUE RESPONSE**, will grow stronger the more frequently you use the feet Chakras for Symbol Key induction. Once you have noticed the **Cue Response**, you have **successfully induced** the two **Initiating Trinity Keys** and it is now time to utilize the remaining **Transmitting Key**.

3. If you have chosen to use the Transmitting Key for D-1-Strand-1-Axiatonal Line-1-**Chakra-1**, continue **standing** for the final Symbol Key Induction, placing the paper with the Transmitting Key face up on the floor between **slightly parted feet; do not stand on Symbol Key paper** for **Base Chakra Induction.** For all other Chakras, **recline on your back** and place the paper with the Transmitting Key **directly on top of the Chakra** to which it corresponds, **on the front of the body**. Use each **INHALE** breath to drawn Maharata Current up from Earth's Core, into the Central Vertical Current **only as far up as the selected Chakra**. On each **EXHALE**, push Pale-Silver Maharata Current **out from the Central Verticle Current** of the body and **into the Chakra**, imagining that the Chakra is progressively turning into a **radiant vortex of Pale-silver light** that **spins faster and faster**. When you begin to feel a bit of warmth, tingling, energy running or itching of the skin **at the Chakra region** (the **Cue Response**), use a final **INHALE breath** to draw the image of the Symbol **down through the Chakra into the body**, and into the **tiny Crystal Seal** that links the Chakra to the Central Vertical Current. Take a **FINAL exaggerated EXHALE**, and imagine that you have **pushed the energy of the Transmitting Symbol Key** out of the Chakra into the **corresponding level of the Auric Field.** Once imbued into the **Auric Field** the single image of the Symbol then **replicates automatically and rapidly**, creating **millions of mini-replicas of the Symbol** within your bio-energetic field. When the **Cue Response** sensation at the Chakra has **completely ceased**, you have **successfully induced** the entire **mathematiucal program** of the Transmitting Symbol into the corresponding Kathara Center, DNA Strand Template, Axia-tonal Line, body region and level of consciousness.

Technique-9: Merkaba Trinity Key Induction

Trinity Master Key Selection Chart

© 2001 Ashayana Deane

12 Transmitting Keys: ***Correspond to each Dimension-Chakra-Kathara Center-Axia-tonal Line etc.***

Number in circle represents the **number of Chakra** into which corresponding Transmitting Key Symbol will be induced.

Note: The **Smaller Symbols** placed **outside of SOME circles** are **Amplifier Key Symbols** corresponding to the Chakra by which they are positioned. The Amplifier Keys can be used **in place of** the Transmitting Key within the circle, or the two Keys can be drawn on one paper for **simultaneous induction**. It is **recommended** to begin with the Transmitting Key inside the circle, then after a few induction exercise, switch to the Amplifier Key, and after several exercises with the Amplifier Key, use both Transmitting and Amplifier Keys simultaneously.

Two Initiating Keys

1)

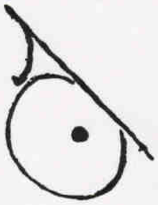

2

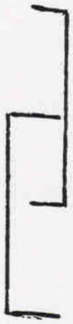

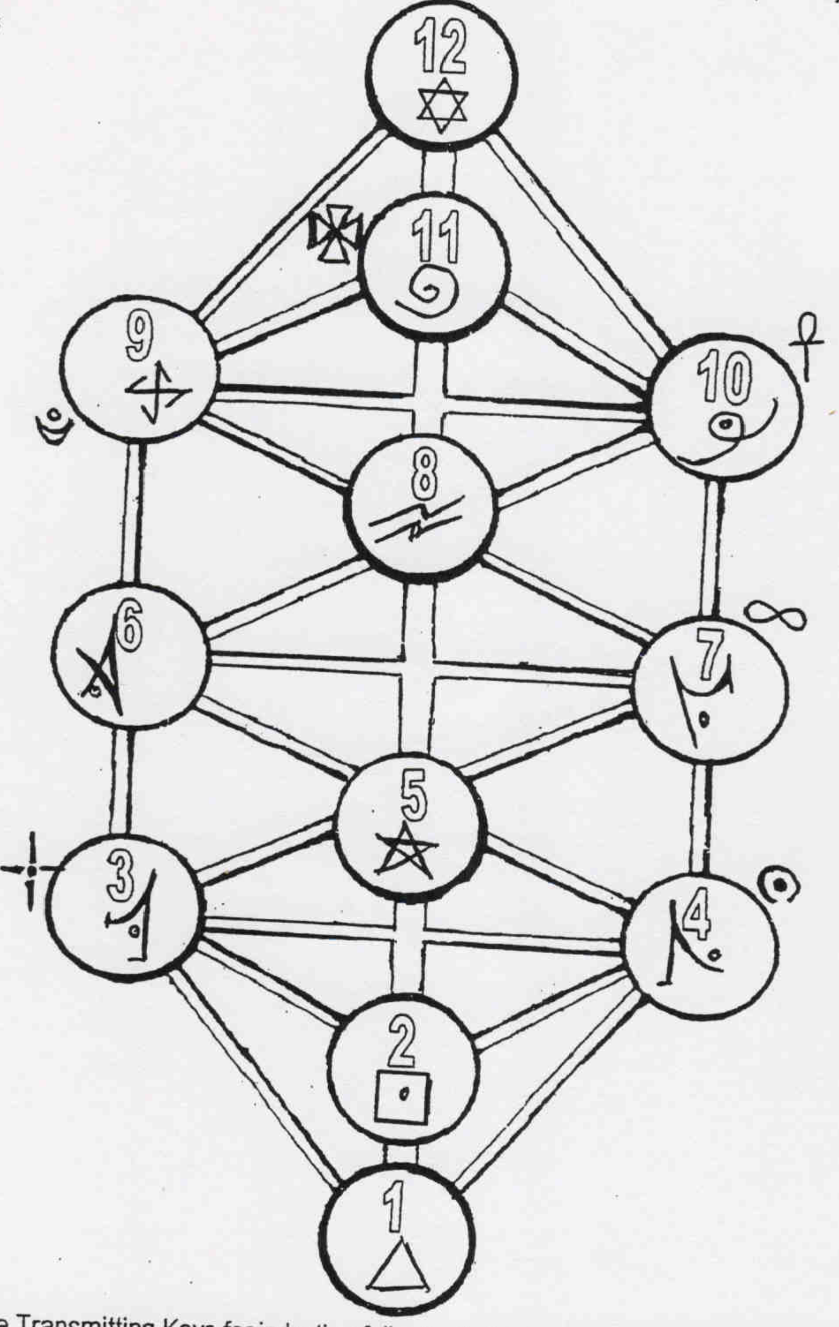

Select one or more Transmitting Keys for induction following induction of the two Initiating Keys above.

In **Kathara Level-1** we have learned of the innate structures of multidimensional reality and how these structures are intimately intertwined with the ***15-Dimensional Human Anatomy*** and levels of personal identity through the personal ***Maharic Shield*** and ***DNA Template***. In rediscovering these long forgotten aspects of the Human condition and our inherently implied relationship to Universal Order, we've become empowered to draw upon the natural **Triadic Phase Currents** indigenous to our personal biology and universal structure, in order to take an active hand in personal *Conscious Bio-spiritual Evolution* and healing, and healing facilitation for others. ***Techniques 1-5*** (#1-*Awakening the Mentor*, #2-*Maharic Seal and Liquid Light Cleanse*, #3-*Opening the Healing Channels Doradic Phase-1 Current*, #4-*Multi-vector Holographic Recoding Kathara Scan* and #5- *The Maharic Infusion* client session) contained within *Chapters 1-4* of *the Kathara Level-1 Maharic Recoding Process* enable us to begin conscious interaction with the "invisible" processes of **Body-Mind-Spirit Evolution** that are inherent to Human design. In our introduction to **Kathara Level-2** we have begun to explore the greater nature of ***Universal Life Force Currents***, the "mysterious" sciences of ***Merkaba Mechanics*** and their indelible connection to our personal evolution and **Bio-spiritual Actualization Processes**. ***Techniques 6-7-8-9-12*** (#6-*Maharic Quick Seal*, #7-*Ps13 Tribal Shield Activation*, #8-*Merkaba Salutations*, #9-*Merkaba Trinity Key Induction* and #12 *Kee-Ra-Sha Activation Sequence*) so far explored within this *Chapter 5* of the *Kathara Level-2 Introduction* enable us to advance our conscious interaction with the processes of personal Body-Mind-Spirit Evolution and Bio-Spiritual Actualization. As we increase our personal abilities to understand, embody, hold and direct the inherent powers of the Universal Life Force Currents, and to access the levels of personal spiritual identity and wisdom to which they correspond, we progressively increase our ability to consciously facilitate healing in ourselves and for others we may choose to serve. An in-depth study of Universal Life Force Currents and Masters Merkaba Mechanics involves cultivating a necessary familiarity with a broad spectrum of interrelated subjects that take us into the heart of the Science of Vibrational Mechancis and Holographic Creation Physics, as well as into the very Soul of our spiritual relationship to consciousness, ourselves, each other, our world and God. Mastery of Merkaba Mechanics is an attribute inherent to achieving the state of consciousness often called "Ascended Mastery", which requires much in-depth study into the "mysteries" of creation. However, a functional mastery of Merkaba and the related constructs of the personal "Inner Templar" system upon which the personal body and identity manifest can be achieved through gaining familiarity with and actively employing core mechanics of the personal Merkabic System.

As our *Kathara Level-2 Introduction* began, our exploration of active techniques of *Tribal Shield Activation, Merkaba Salutations* and *Trinity Keys* set us on the path of becoming functionally able to consciously direct elementary operations of the personal Merkaba Vehicle and Life Force Currents, as a means of expediting our progress of Bio-Spiritual Evolution through DNA Template/Merkaba Field Activation. As our *Kathara Level-2 Introduction* continues, we will begin to explore the seemingly "***magical world of Merkaba***", rediscovering the very tangible importance of the **Merkaba Field** in relation to the inherent dynamics of DNA, health, healing and spiritual actualization. To achieve genuine integrity in the facilitation of healing for others, one must simultaneously strive to heal and actualize the self. The greater the level of health and spiritual actualization possessed by a healing facilitator, the greater will be the facilitation of healing for the client to whom the healing facilitator is in service. As Kathara Level-2 Introduction draws to a close we will begin an introductory exploration of the "Ancient Secrets" of the **Veca Codes,** so that we may engage practical application of two additional techniques of Photo-Radionic and Photo-Sonic **Merkabic Healing. Techinque-#10** ***Restoring the Trion Field Veca Code Activation*** and **Technique-#11**-***Restoring the Meajhe Field Eckasha Activation*** will provide us with powerful tools of the Kathara Healing trade that will ***substantially increase the quality, amount and power of interdimensional frequency we are capable of receiving, holding and transmitting***. Through these simple but effective Merkabic Healing tools we can expedite our processes of personal healing and spiritual actualization, thereby enabling us to become more effective and powerful healing facilitators in our client session service.

The "Magic of Merkaba"

Spiritual-science teachings of **Merkaba Mechanics** emerge from the most ancient of times; record of Merkaba teachings is most often found in obscure traditional "Occult" texts or within New Age "Ascension" teachings. Genuine Merkaba Sciences hold the keys to understanding and mastering the processes of both spiritual and physical evolution, on a personal, planetary, galactic and universal level. Merkaba Mechanics are not the product of "science fiction", nor are they "magical hocus-pocus" of "occult mysticism". Merkaba Mechanics are not "magic" at all, but rather an intrinsic element of universal Primal Order. **Merkaba Fields**, sets of Star-tetrahedron shaped counter-rotating electro-magnetic energy fields within and surrounding manifest forms, are simply an inherent element of the15-Dimensional Anatomy of form manifestation and the Primal Order structure of multidimensional Creation Physics. In the "*Transduction Sequence*" described in *Kathara Level-1*, the sequence of energy transmutation by which conscious energy manifests into tangible biological form, the *Merkaba Field* represents the structure and stage of energy dynamics by which *Primal Life Force Currents* pass from the scalar-wave-template of the embodied *Kathara Grid levels* into the *DNA Template*. In terms of non-biological form manifestation, all forms possess an inherent structure of interwoven Merkaba Fields through which Primal Current energy is transmuted from the Kathara Grid Core scalar-template level into the outer levels of the Kathara Grid and Axiom Line template upon which the specialized structure of the form builds up, through particle/anti-particle accretion, into physical manifestation. Merkaba Mechanics seem "magical" only when the inherent "**Laws of Nature**" that govern the dynamics of **Multidimensional Universal Unified Field Physics** are not understood.

Through understanding of genuine *Universal Unified Field Science, Merkaba Fields*, and their inherent dynamics as utilized in the study of *Merkaba Mechanics*, are simply recognized as natural elements of the processes inherent to Creation Physics and the naturally occurring, organic relationships between Energy Consciousness, and the Holographic Manifestation of form.

As Merkaba Fields are an intrinsic, significant and natural part of the Body-Mind-Spirit anatomy, the study and application of Merkaba Mechanics is an essential component in any Masters Holistic Healing or Spiritual Actualization program. Merkaba Mechanics are "*Scientific*" in that they represent a core element within the innate processes of creation and manifestation that are intrinsic to the sciences of Creation Physics. Merkaba Mechanics are also "*Spiritual*" in that it is through the innate structures and dynamics of Creation Physics, of which Merkaba Mechanics are a natural part, that consciousness enters the manifest arena in the form of sentient, individuated, multidimensional identity. The "body, mind and spirit" of every manifest being is formed through and governed by the natural laws of Creation Physics and Primal Order. The Laws of Creation Physics and Primal Order are designed and maintained by and within a non-manifest, intelligent, central Consciousness Force and Source of Creation that is commonly referred to as "**God**". God expresses It-Self in the form of *Conscious-energy Constructs* that carry the living "Spirit and Consciousness of God" in the form of **Primal Life Force Currents**, which are the foundations for macrocosmic and microcosmic creation and manifestation. Merkaba Fields and the mechanics by which they function, are an indelible part of the physics of Creation and Consciousness, and are thus intimately involved with the Bio-Spiritual processes inherent to **Spiritual Actualization**.

Merkaba Fields, DNA, Healing and Spiritual Actualization

Particles, People, Planets, Galaxies and Universes all have **sets of interwoven Merkaba Fields** that serve as the "**core Primal Life Force Current circulation system**" between the Cosmic, Universal, Galactic, Planetary and Personal "Divine Blueprint" Kathara Grid scalar-templates and the manifestation of those blueprints into the tangible structures of matter. Merkaba Fields are the "**carriers of consciousness**" and the "**engines of energy**" by which physically manifest forms are created, maintained and evolved through the Holographic Unified Field of manifest space-time-matter reality. In relation to the processes of personal bio-spiritual healing and spiritual actualization, the personal Merkaba Field is an intrinsic aspect of 15-Dimensional Anatomy that directly governs the function of the physical, emotional, mental and spiritual body systems. **The multidimensional structure of the personal Merkaba Field is the system of electro-magnetic interface by which the Primal Life Force Currents and Consciousness of God-Source and Its sentient individuations pass from the Kathara Grid Core Template into the scalar-wave DNA Template that holds the blueprint for the physical-chemical biological DNA.** This reality of **Merkaba Field-DNA Connection** implies that the personal Merkaba Field is directly related to the function or dysfunction of the **biological genetic code.** As the Merkaba Field "carries the energy of consciousness" into the DNA and biological vessel, the *Merkaba Field-DNA Connection* also implies that the potentialities of consciousness and embodiment of spiritual identity, while a being is ensconced within the Time Matrix, are dependant upon the condition and function of the personal Merkaba Field, DNA Template and Kathara Grid.

> *Merkaba Fields are the "**carriers of consciousness**" and the "**engines of energy**" by which physically manifest forms are created, maintained and evolved through the Holographic Unified Field of manifest space-time-matter reality.*

Universal Merkaba Field Structure

Basic Merkaba Field structure is relatively simple. Viewed Macrocosmically, in a 15-Dimensional Time Matrix there are **15 sets** of counter-rotating electromagnetic **Merkaba Spirals** that form **15 Universal Dimensional Merkaba Fields.** A "**Merkaba Field**" is composed of a set of **2 Merkaba Spirals**, one electrical-transmitting "cone", the other magnetic-receiving "inverted cone"; the 2 "cones" of the Merkaba Spirals, which operate as a **White Hole/Black Hole Pair**, intersect each other, creating a system of energy vortices that form an **Elliptical Sphere** of energy within the center of the intersected Merkaba Spirals. The Sphere of energy at the center of a Merkaba Field is called the **FORM CONSTANT**; it is a "**spherical energy domain**" within which the individuation of consciousness and holographic manifestation/projection of form can take place, akin to an "energy egg" within which manifest reality occurs. Within this Sphere of energy at the center of a Merkaba Field, the template-blueprints of a form are held "constant", allowing conscious Partiki units, particles, anti-particles and ante-particles to accrete in specific ratios to "fill out the form in matter". Within the Form Constant sphere, consciousness is able to perceive the **holographic illusion** of externalized space-time-matter manifestation, as the scalar-template of the form is held fixed within the sphere. The "hologram" appears to remain "constant" within the Form Constant sphere while consciousness is focused within the frequency bands of which the Merkaba Field is composed.

On the macrocosmic universal scale, the Form Constant of each Universal Dimensional Merkaba Field creates a single, elliptical-sphere shaped electromagnetic domain of frequency that represents a single dimension of manifestation. Following the innate mathematical-geometrical structure of Primal Order, the 15 Universal Dimensional Merkaba Fields group into **5 sets** of 3-Dimensional Merkaba Fields called **Universal Harmonic Merkaba Fields**. The 3 Form Constant Spheres from each set of 3 Universal Dimensional Merkaba Fields forms a **Universal HOVA BODY**, or a "Harmonic Universe" electromagnetic domain. Following the mathematical Primal Order program of a Time Matrix as held within the Universal Kathara Grid template and transmitted through the rotation ratios of the Harmonic and Dimensional Merkaba Fields, each of the 3 dimensional levels of a Universal Hova Body holds the Partiki units, consciousness and matter particulates of its dimensional field into a specific angle and speed of rotation. The 3-Dimensional Harmonic Merkaba Field controls the relationship of particle rotation angle (angular rotation of particle spin or "**ARPS**") and rotation speed between each of the 3 Dimensional Merkaba Fields, holding together a 3-dimensional electromagnetic domain, within which consciousness can experience a 3-dimensional space-time-matter hologram. In a single 15-Dimensional Time Matrix there are thus 15 singular Universal Dimensional Merkaba Fields governed by 5 larger 3-dimensional Universal Harmonic Merkaba Fields, all of which circulate consciousness in the form of Primal Life Force Currents into and out of the hologram of manifestation. The seeming solidity and durability of an externally manifest matter-field in space-time is created, maintained, perpetuated and completely dependent upon the intrinsic function of the Universal Merkaba Fields.

Within the larger Form Constant elliptical-spheres of the 5 Universal Hova Bodies of a 15-Dimensional Time Matrix, the holographic perceptual realities of space-time-matter take place. Every being or thing in manifest form enters its reality hologram by *taking on the mathematical program* of first the *Universal Kathara Grid* and *5 Universal Hova Bodies*, which set the "Life Stream" or "Stream of Consciousness" of Primal Creation

Currents in motion following the specific **Harmonic Merkaba Field spin ratios** characteristic to each Harmonic Universe System. Galaxies are formed in this way, as individualized expressions of the greater Harmonic Universe of which they are a part. **Universal Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation of the **galaxies** within them. **Galactic Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation of **star systems and planets** within them. **Planetary Harmonic and Dimensional Merkaba Fields** govern the birth, evolution and transmutation/death cycles of the **species and matter forms** existing upon the planet. The **Personal Merkaba Fields** of species and individuals direct the function of the personal **DNA Template, chemical DNA** and **biological genetic code**, thus governing the birth, evolution and transmutation processes of individual organisms. The mathematical-geometrical programs of the **Primal Order "Divine Blueprint"** that govern Merkaba Field formation and spin ratios are held within the Universal, Galactic, Planetary, Species and Personal Kathara Grid structures. In a singular 15-Dimensional Time Matrix, the **inter-linked Universal, Galactic and Planetary Harmonic and Dimensional Merkaba Fields** circulate Primal Life Force Currents and consciousness into and out of manifestation in space-time.

Universal Merkaba Fields

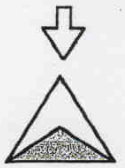

Electrical ***Merkaba Spiral***
Rotates Clockwise
"White Hole" Vortex

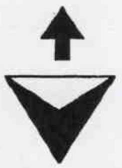

Magnetic ***Merkaba Spiral***
Rotates Counter-clockwise
"Black Hole" Vortex

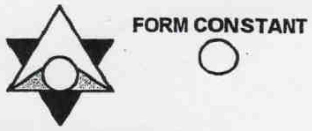

FORM CONSTANT

One ***Merkaba Field*** =
a set of 2 *Merkaba Spirals*
Black-White Hole Pair

Universal Merkaba Field of One 15-Dimensional Time Matrix

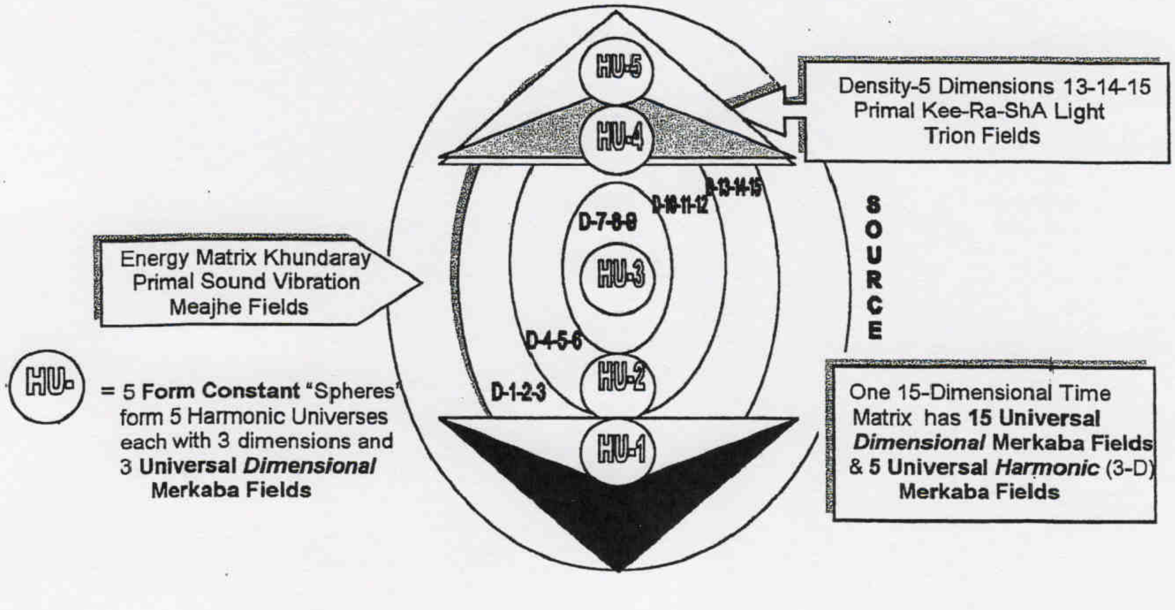

HU-5
HU-4
D-13-14-15
D-10-11-12
D-7-8-9
HU-3
D-4-5-6
HU-2
D-1-2-3
HU-1
SOURCE
Density-5 Dimensions 13-14-15
Primal Kee-Ra-ShA Light
Trion Fields
Energy Matrix Khundaray
Primal Sound Vibration
Meajhe Fields
HU- = 5 **Form Constant** "Spheres"
form 5 Harmonic Universes
each with 3 dimensions and
3 **Universal *Dimensional***
Merkaba Fields
One 15-Dimensional Time
Matrix has **15 Universal**
***Dimensional* Merkaba Fields**
& 5 Universal *Harmonic* (3-D)
Merkaba Fields

The Personal Merkaba, Trion-Meajhe Field, Radial Body and Radis.

Each individual being receives its life force, consciousness, biology and manifest hologram through the innate connection between the **Personal Merkaba Field** and the Merkaba Fields of the Planet, Galaxy, Universe and Time Matrix. When the **D-12 Pre-matter Divine "Christos" Blueprint** is expressing naturally in a planetary system, the organisms on the planet circulate Life Force Currents and consciousness through the physical and "spiritual" bodies via the Personal and Planetary Merkaba Fields. The **Planetary Merkaba Field** circulates Life Force Currents through the **Personal Kathara Grid** template (morphogenetic field) of the organism via the **Planetary Bio-Feed Interface System.** The organism's **Personal Merkaba Field** then picks up the energy from the Kathara Grid and continues to circulate the Life Force Currents from the Personal Kathara Grid and into the **Maharic Shield imprint** in the **DNA Template** that is held within the Level-1 *Kathra 12-Tree Grid,* Level-2 *Kathara Crystal Seals-Hara Grid* and Level-3 *Kathara Diodic Grid.* From the Maharic Shield imprint in the Kathara Grid, the Personal Merkaba Fields distribute the Life Force Currents and consciousness into the **DNA/RNA Template** and **Axiom Lines.** The Life Force Current continues circulation from the DNA/RNA Template and Axiom Lines into the **Personal Hova Bodies** and **Auric Field**, to the **Chakras** and **Meridian Lines**. Passing through the Meridian Lines, the Life Force Current, imbued with the specific mathematical manifestation program picked up from the Kathara Grid Divine Blueprint Template, then enters a state of final transmutation into 3-dimensional particles by passing through an electrostatic **Repulsion Zone** barrier that exists surrounding each 3-dimensional Hova Body.

The frequency barrier surrounding each 3-dimensional Hova Body forms as **thin, skin-like membrane** or **"Tissue Capsule"** surrounding each of the 5 Hova Bodies in the 15-Dimensional Anatomy. The **Hova Capsule** membrane is composed of Density-5 **Primal Light** ante-matter units, called **TRION UNITS**, that exist in the form of fixed standing-columnar-scalar-light waves made of **Keylons**, or 3-dimensional **Partiki Grids**. The Trion Units and Keylon Partiki Grids within them are anchored upon a fixed field template of **Primal Sound** units, called **MEAJHE UNITS**, which exist as fixed points of mathematically organized vibration. The outer layer of each Hova Capsule is called the **Trion Field**, as it is composed of **Trions**, or units of Density-5 ante-matter Primal Light. The inner portion of the Hova Capsule is called the **Meajhe Field** as it is composed of **Meajhons** (mE' yans) or fixed points of Primal Sound vibration from the **Energy Matrix** beyond the holographic manifestation of the Time Matrix Light Fields. The collective body of the 5 Hova Capsules is called the **RADIAL BODY** or **Trion-Meajhe Field**. All manifest forms possess a core Radial Body/ Trion-Meajhe Field **Hova Capsule** structure as part of the intrinsic 15-Dimensional anatomy.

As multidimensional Life Force Currents pass through the Meridian Lines and into each of the 5 layers of the Radial Body, the energy currents and the consciousness carried upon them project the mathematical program of the Divine Blueprint into the Trion-Meajhe Field of each of the 5 Hova Capsules. Within the Hova Capsule, the individualized raw Life Force Currents enter a system of 3-dimensional energy conduits or **rivulets** called **Radis**, which are the "turnstiles" between "manifest matter" and "conscious energy". The Radis draw in the Life Force Currents with their manifestation program through the Trion-Meajhe Field, then polarize and break up the Trion and Meajhon units into negatively and positively charged electrical sub-units called **Mions** and **Dions**. Partiki, Keylons, Meajhons, Trions, Mions and Dions belong to a category of primal building blocks known as **IONIC PARTICULATES,** which are the building blocks of consciousness, energy, light, sound and scalar-wave-fields that form the blueprints upon which all matter manifests. **Trions** are electrostatic particulates that have both a negative and positive electrical charge, which renders them neutral or static; they are the foundations of what becomes **Ante-matter** Primal Light particles of pre-visible gaseous light. Meajhons are static units of pre-sound standing waves, also possessing neutral electrical charge, that form the Primal Sound Fields of the Energy Matrix. Within the structures of the Radial Body Radis rivulates, Trions and Meajhons are **polarized** into negative charged **Mions** ("Base Magnetic" vibrating pre-sound particulates, foundations of negatively charged particles such as electrons, originating from the Universal Particle Particum field.) and positively charged **Dions** ("Base Electrical" oscillating pre-light particulates, foundations of positively charged particles such as protons, originating from the Universal Anti-particle Partika field). Together, Mions and Dions create the **electromagnetic Ionic Particulate base field** within the Merkabic Form Constant through which particles emerge via Ionic Particulate accretion into tangible matter units.

The Trion-Meajhe Field Radial Body of any manifest form represents the 5-Density "Veil" through which consciousness-as-Primal Life Force moves into and out of externalized 3-dimensional manifestation. **Merkaba Fields are the "energy engines and consciousness carriers" by which Life Force and consciousness are circulated between the internal Kathara Grid scalar template and the Trion-Meajhe Field Radial Body "veil" as they pass into and out of external expression.** The perpetual polarization, electromagnetic expression and de-polarization of the Life Force Currents within the Merkaba Field Form Constant electromagnetic domain is the process by which the holographic projection of matter is perpetually brought into and taken out of perceptually experiential being. The **Radial Body** Trion-Meajhe Field receives, via Merkaba Field circulation, the template or "design" for its particle manifestation from the "instructions" held in the Divine Blueprint Maharic Shield and DNA Template of the Kathara Grid. Manifestation instructions pass from the Kathara Grid to the DNA Template via the Merkaba Field, then continue into the energy circulation systems (Axiom Lines, Hova Bodies, Chakras, Meridians) that deliver the instructions to the Radial Body for particulate manifestation. In biological life forms, the Radial Body and Radis then translate the scalar-wave blueprint of the DNA Template and Kathara Grid into the polarized Ionic Particulate Mion-Dion sub-units, which accrete following the DNA Template instructions to form sub-atomic, atomic, molecular, chemical DNA, physical matter form- the vehicles of embodied consciousness. The **Merkaba Field** also receives its instructions for energy circulation from the Kathara Grid and DNA Template.

The Radial Body Tissue Capsule, Radis
& the Mechanics of Manifestation

The **Radial Body** exists as a thin "***Tissue Capsule***" surrounding each of the **5 Hova Bodies.** Each of the **5 Radial Body Levels** are composed of an **Outer Layer** of electrostatic neutral-charge **Trion Primal Light Units** and an **Inner Layer** of **Meajhon Primal Sound Units**. Each of the 5 Levels of the Radial Body serve as the **polarizing filter or "lens"** through which **Multidimensional Primal Life Force Currents** flow, via **Merkaba Field** circulation, from the Kathara Grid and DNA Template, into manifest form.

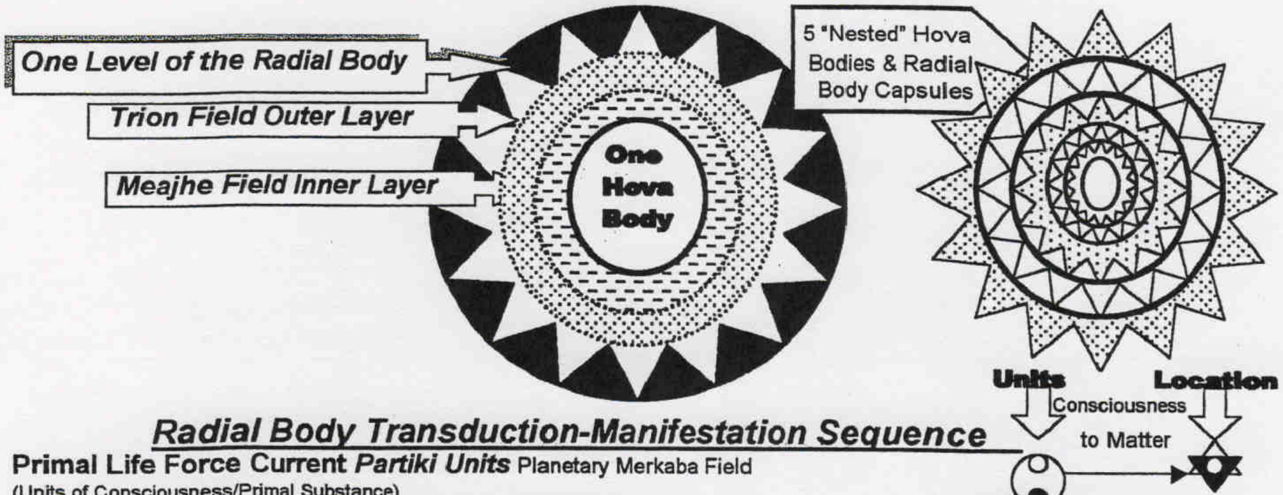
One Level of the Radial Body
Trion Field Outer Layer
Meajhe Field Inner Layer
One Hova Body
5 "Nested" Hova Bodies & Radial Body Capsules

Radial Body Transduction-Manifestation Sequence

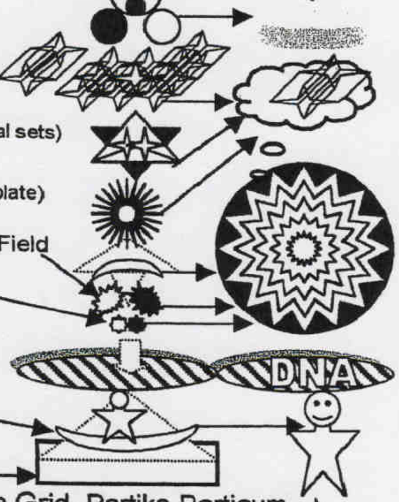

Primal Life Force Current *Partiki Units* Planetary Merkaba Field
(Units of Consciousness/Primal Substance)
Partika and Particum Units Kathara Template/Maharic Shield
(Polarized Primal Substance Units)
3-D Partiki Grid Keylon Units/Codes/Grid Morph. Keylon Thought-Form Field
(3-D Crystallized Pre-Light-Sound units, unit groups and grids)
Merkaba Field *Keylons* (3-D Crystallized Pre-Light-Sound Keylon Unit electromagnetic spiral sets)

DNA/RNA Template *Keylons* Kathara Grid (3-D Crystallized Pre-Light-Sound unit Template)
Axiom Lines-Hova Bodies-Auric Field-Chakras-Meridian Lines
Trion & Meajhon *Electrostatic Units* 5 Radial Body "Lens" Levels Trion-Meajhe Field
Dion (+) & Mion (-) Ionic Particulate *Electromagnetic Units* Radis Lines

Sub/Atomic Particles/Anti-Particles, Molecules...Chemical DNA
Physical Matter Body Hologram "Chemical Lens"

External Hologram Projection of Space-Time-Matter
Primal Life Force Currents ➡ Planetary Merkaba Field Partiki ➡ Kathara Grid -Partika-Particum ➡ Maharic Shield Divine Blueprint ➡ Keylon Morphogenetic "Thought-form" Field ➡ Merkaba Fields ➡ Kathara Grid DNA/RNA Template ➡ Axiom Lines ➡ Hova Bodies ➡ Auric Field ➡ Chakras ➡ Meridian Lines ➡ Trion-Meajhe Field 5 Radial Body Levels ⇨ Radis Lines ⇨ Mion & Dion Ionic Particulates ⇨ Strong and Weak Nuclear Force, Electromagnetism ⇨ Quarks, Muons, Measons, Sub-atomic & Atomic Particles and Anti-Particles ⇨ Blueprint translation into Nuclei, Molecules, Elements, Compounds, Chemicals ⇨ Chemical DNA Template and "Chemical Lens" of the manifest body form ⇨ Manifest Hologram ▭

Reverse Transduction De-Manifestation Sequence: matter- *chemical* DNA- *Sub-Atomic units*- Radis *Mion/Dion units*- Radial Body *Trion-Meajhon units*- Meridian Line-Chakra-Auric Field-Hova Body-Axion Line-DNA Template-Merkaba Field-Morphogenetic Thought-form Field *Keylon Grids & units*- Maharic Shield Template & Kathara Grid *Partika and Particum Units*- Planetary Merkaba Field *Partiki Units* of consciousness.

Radial Body Transduction-Manifestation Sequence

1. From the **Planetary Merkaba Field** (via the Planetary Shield, Axiom Lines, Hova Bodies Vortices, Meridian Lines and Radial Bodies), **rhythmic pulses of Conscious Primal Life Force Currents** enter the species and personal **Kathara Grid Core Template** as Primal Units of **Partiki**.
2. Within the **Kathara Grid Template** the **Partiki Units** pick up the **mathematical-geometrical program** from the **Maharic Shield** "Divine Blueprint" Template, **polarize** to form **Particum and Partika Units**, then **accrete** to form **3-dimensional** electro-magnetic **Partiki Grids** called **Keylons**. The Keylons group together mathematically forming complex templates of **Crystallized Keylon Codes** (or "***Keylon Thought-Crystals***"), which create a mathematically encoded **3-dimensional Morphogenetic Field ("Keylon Thought-form Field"** or Morphogenetic **"Keylon Crystal Grid"**) carrying the manifestation blueprint program of the Maharic Shield.
3. As the Primal Life Force Currents of the electromagnetic Keylon Crystal Grid interact with each other following the program of the D-12 Pre-matter **Universal Divine Blueprint**, they form **sets of spiraling electro-magnetic vortices** called **Merkaba Fields**, which carry the Maharic Shield-Kathara Grid Blueprint program into formation of the **DNA/RNA Template** Keylon Crystal Grid.
4. From the DNA/RNA Template held in the Kathara Grid, Primal Life Force Current carries the **Keylon Crystal Grid** of the blueprint into the **Axiom Lines, Hova Bodies, Auric Field, Chakras** and **Meridian Line** systems, finally carrying the Keylon Crystal Grid of the Divine Blueprint into the **5 Radial Body "Tissue Capsule" Levels.**
5. Within the **5 Levels of the Radial Body** the **incoming Keylon Crystal Grid** translates into **electrostatic Trion-Primal Pre-Light Units** (gaseous pre-light Ante-matter) and **Meajhon-Primal Pre-Sound Units** (units of vibrating pre-sound energy) that possess a **neutral electrical charge.** The **Encoded Primal Life Force Currents** (PLFCurrents+ Keylon Crystal Grid blueprint program) are converted into a **Trion-Meajhe Field blueprint replica;** a " Primal Light-Sound" translation of the original blueprint.
6. The **Encoded Primal Life Force Currents** then travel, **AS the Encoded Trion-Meajhe Field**, into the **Radis Lines** in each of the 5 Levels of the Radial Body. The Radis Lines first **replicate and store** the **mathematical blueprint encoding** of the **Trion-Meajhe Field,** then **polarize** the neutrally charged Trion-Meajhe units into **electromagnetic units** called **Mion** (-) and **Dion** (+) ***Ionic Particulates.*** **The replicated Trion-Meajhe Field blueprint** becomes part of the "Primal ***Pre-sound*** vibration" ***Meajhe Field*** **Inner Layer** of the ***Radial Body*** **Trion-Meajhe Field,** serving as a living "***Memory Matrix***" composed of **Meajhon Unit** "Primal Pre-sound vibration". The **polarized Mion/Dion Units** become the foundations upon which the **"Strong and Weak Nuclear Force", Electromagnetism, quarks, Muons, Measons, Sub-atomic Particles** and **Anti-particles** etc. form. Particum-based **Mions** ("Base magentic) become **particles, negative electrical charge units** like **electrons** and **sound spectra.** Partika-based **Dions** become **anti-particles, positive electrical charge units** like **protons** and **light spectra.**

7. Blueprint Encoded Primal Life Force Currents in the form of a **Mion-Dion Field** then translate the original Kathara Grid Blueprint into **Quarks, Muons, Measons, Sub-atomic** and **Atomic particles** and **anti-particles**, which **accrete** according to the blueprint to form **molecules, elements, chemicals**, the **Chemical DNA Blueprint** and the **physically manifest body form**. As this occurs the **body image "flashes on"** into manifestation as a **Pulse of Life Force Consciousness** carrying the **Chemically Translated Holographic Image** of the body form template blueprint. This Pulse of Consciousness continues to expand from the Kathara Grid and DNA via the Transduction Sequence, through and beyond manifestation of the body image and into **projection of portions of the Dimensional Unified Field** that **holographically translate** into **individualized experience** within a perceived **Common Planetary Time Rhythm and Matter Field.**

NOTE: ***An individual will holographically perceive and experience only the portions of the Dimensional Unified Field mathematical-geometrical program that are encoded within the personal Maharic Shield, species Tribal Shield and Planetary, Galactic and Universal Shield aspects of the personal Kathara Grid and DNA Template.***
A **global population** experiences a **common planetary reality field and time continuum** because the Kathara Grid Template and DNA of all beings entering a planetary field **adopt the common imprint** of the mathematical-geometrical program of the **Planetary Kathara Grid** blueprint in order to enter into the electromagnetic domain of the Planet's Form Constant Hova Body structure for experiential manifestation. Members of a **singular species** share a common **species Tribal Shield** template through which a **common Chemical Body Form** and perceptual orientation holographically manifest. The imprint for **individuality** in biological form, life experience variation and the application of **Free Will Choice** is held within the mathematical program of the **personal Maharic Shield** D-12 Pre-Matter **Divine "Christos" Blueprint**. Imbued within the mathematical program of the **Human species Tribal Shield**, and that of some other species, is the imprint for **probability** to exist. The Human species Tribal and Maharic Shields carry a program that allows for **new programs of probable creation** to be **added** to the personal and species **Kathara Grid**, allowing the embodied consciousness to retain **a degree of direct influence** over what sets of mathematical programs will be called into personal experiential Holographic Expression from the personal Maharic Shield Template. **Creative variation** on the original personal, species, planetary and galactic Divine Blueprint is possible as long as the new probabilities do not deviate so excessively from the Pre-set programs and **Original Intentions** of the Universal Divine Blueprint that the **structural integrity** and **Source-God Will Intentions** for our Time Matrix become **irreparably compromised. Free Will thus exists within a pre-set framework of overriding Divine Will**, and is permitted free reign in the manufacture of **co-creative experiential probability** up to the point at which the Original Divine Will of Source and the creative framework itself become unduly jeopardized.

8. The physical body form serves as a **Chemical Lens** through which ***portions*** of the **Dimensional Unified Field** are perceived by the embodied consciousness as a **3-dimensional Hologram,** as the Primal Life Force Currents continually carry **pulses of the individuated consciousness** through the ***Transduction Sequence*** from Kathara Grid, through the DNA, to the manifest Chemical Lens of the physical body structure and Holographic "external" reality field projection.

***Radial Body Transduction-Manifestation Sequence*-** **Continued** page 3

9. As a Keylon-Thought-Form blueprint traveling on one Pulse of Consciousness **fully expands** (fully "Activates") into the manifest externalized hologram, **another Pulse of Consciousness** is released from the synchronized mathematical programs of the Planetary Merkaba Field, which sets the Transduction Sequence in motion ("Initiates")for the ***next portion* of the personal mathematical blueprint**. The blueprint of the **Planetary Kathara Grid** regulates the **cyclic rhythms** at which **Pulses of Life Force Current** will be released through the **Planetary Merkaba Field**; the Planet releases **"Pulses of Time"**.

NOTE: The linear time experience of life forms on a planet is synchronized into specific singular continua and cycles of time through the *DNA Template*; the species Tribal Shield, personal Maharic Shield and DNA Template are imbued with the *exact mathematical programs* of the specific *Planetary Time Cycle* and *Continuum* or "Time Wavef" into which the consciousness intends manifest embodiment. A singular Time Wave constitutes a Time Continuum, within which there are specific repeating cycles of progressively accelerating *Pulses of Time* that collectively form the linear time-progression continuum of a *Planetary Time Cycle*.

10. As one **Pulse of Time/Consciousness** brings its blueprint into manifest **Critical Mass Expansion** (fully activates specific portion of the blueprint in Holographic Expression), the Planetary Merkaba Field releases the **next "incoming" pulse** of Time, which carries the next portion of the Kathara Grid mathematical blueprint into the Transduction Sequence. As the new Pulse enters the Transduction Sequence and begins **Expansion** toward manifestation, the **first pulse** of the **fully expanded hologram** begins **Contraction**, entering a **back-return** through the **Reverse Transduction Sequence** (matter- *chemical* DNA-*Sub-Atomic units*- Radis *Mion/Dion units*- Radial Body *Trion -Meajhon units*- Meridian Line-Chakra-Auric Field-Hova Body-Axion Line-DNA Template-Merkaba Field-Morphogenetic Thought-form Field *Keylon Grids & Units*- Maharic Shield Template & Kathara Grid *Partika and Particum Units*- Planetary Merkaba Field *Partiki* Units)

11. A s **pulse-1 contracts** its hologram program **moving backward** through the Transduction Sequence, creating the **"Remembered Past" moment**, the **pulse-2** simultaneously **expands** its hologram program **moving forward** through the Transduction Sequence, birthing the as-yet-to-come **"Future" moment.** The backward moving **"Past"pulse-1** and the forward moving **"Future" pulse-2 intersect** at the **Radial Body Levels** in the Transduction Sequence. This **"past pulse-future pulse intersection point"** is called the **Eiron** (E'r-on) **Point,** also known as **"Zero Point"**, the **"Present moment of conscious focus within the hologram"**, or the "***Eiros*** Conscious, Breathing Life-Force Stream" (Also referred to as the **"Stream of Consciousness"**, the "***Living Holy Spirit***" or the "***Breath of God***" ..etc.).

12. The *Eiron/Zero Point*, or **Eiros**, is the **"Present Moment of Creation Power"** for the "manifest" conscious identity. It exists at the position where the **third dimensional field,** the electromagnetic **Resonant Tone "Mental Body"** field of a single 3-dimensional Hova Body, **intersects** with the **Meajhe Field Inner Layer** of the corresponding **Radial Body** Tissue Capsule.

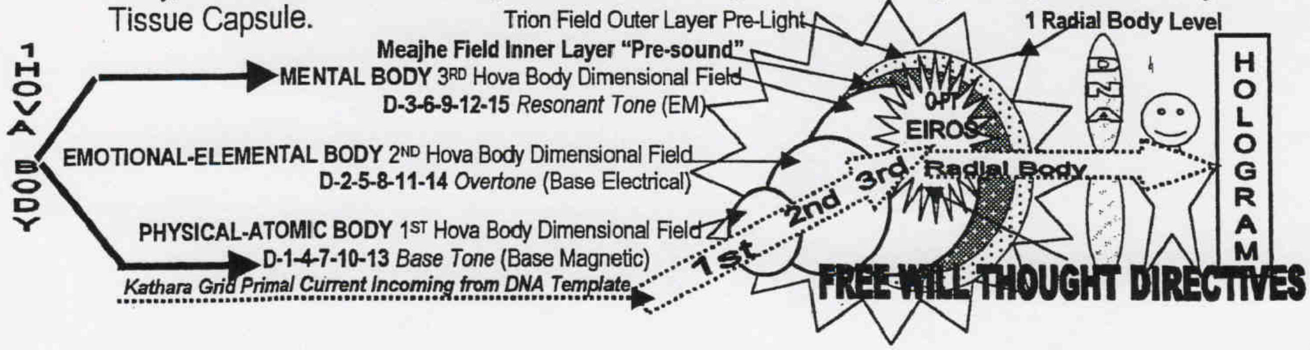
Trion Field Outer Layer Pre-Light
1 Radial Body Level
1 HOVA BODY
Meajhe Field Inner Layer "Pre-sound"
MENTAL BODY 3RD Hova Body Dimensional Field
D-3-6-9-12-15 Resonant Tone (EM)
EMOTIONAL-ELEMENTAL BODY 2ND Hova Body Dimensional Field
D-2-5-8-11-14 Overtone (Base Electrical)
PHYSICAL-ATOMIC BODY 1ST Hova Body Dimensional Field
D-1-4-7-10-13 Base Tone (Base Magnetic)
Kathara Grid Primal Current Incoming from DNA Template
OPT
EIROS
1st
2nd
3rd
Radial Body
DNA
HOLOGRAM
FREE WILL THOUGHT DIRECTIVES

13. As the backward moving **"Past" pulse-1** and the forward moving **"Future" pulse-2 intersect** at the Radial Body Meajhe Layer **Eiros** (Eiron Point) ✷ the returning **contracting pulse-1** passes through the Trion Layer and into the ***Meajhon blueprint replica,*** called a " **Radial TILE"** ▨, that it **left behind** in the Meajhe Layer **Eiros** during pulse-1's **incoming** passage. Just as the returning pulse-1 retreats back through its Radial Tile, the **incoming pulse-2,** carrying the **next blueprint** coding ⬚ ,**simultaneously** passes through the **first Radial Tile, combining** the "past" **Radial Tile** mathematical-vibrational blueprint with the **incoming pulse-2 blueprint** ▨ At this moment-point, the hologram template from the "past moment" is adopted by the incoming "future moment" pulse-2 and expanded by the pulse-2 blueprint.

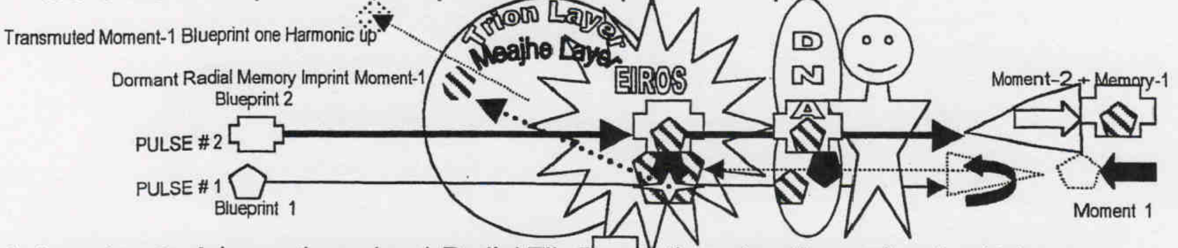

14. As pulse-2 picks up the pulse-1 Radial Tile, ▨ the **retreating pulse-1**, which now **holds the Radial Tile blueprint in *reverse mathematical sequence* as *TRION Pre-light Units*** ⬟ , collides with the incoming pulse-2. Both pulses carry the **common blueprint** of the first Radial Tile; the **incoming pulse-2** holding the blueprint in **forward** mathematical sequence as ***Meajhon Pre-sound Units*** ▨,the **retreating pulse-1** holding the blueprint in **reverse mathematical sequence** as ***Trion Pre-light Units*** ⬟ . As the expanding/advancing and contracting/retreating pulses collide with blueprint **"sequence and anti-sequence",** the pulse-1 reversed **Trion Radial "Anti-Tile" fuses** with the original **Meajhon Radial Tile** ✦ and both Radial Tile and Anti-tile **completely de-manifest** or "***Ascend***" from their original Harmonic Universe by **TRANSMUTING** to become a ***Merkaba Field***. The **energy/ consciousness** once held by Radial Tile/Anti-tile then **replicates** its original blueprint via **Fission,** leaves a **dormant "Memory Imprint"** *Meajhon replica* ▨ in the Radial Body ***Meajhe Field,*** and **reappears** in faster oscillating form within the corresponding **Hova Body Form Constant** of the next Harmonic Universe up in the dimensional scale.

15. As transmutation of the "past" Radial Tile occurs in the Radial Body Eiros, the **incoming pulse-2** is momentarily held **static** as it leaves a **new Radial Tile** in the Meajhe Field; the new Radial Tile holds the **pulse-1+pulse-2** program. At this moment-point the **entire** externalized **Hologram** literally **"Flashes Off"** and de-manifests, its energy **held in suspension** within the **Radial Body** Trion/Meajhe Field. When the first Radial Tile is transmuted, the incoming **pulse-2**, carrying an **expanded imprint of the first Radial Tile**, progresses from the Meajhe Field Eiros to Trion Field, then begins to **"Flash On"** its blueprint in manifestation. The embodied self perceives a **"liner moment progression"**; the ***pulse-1***program becomes **"memory"** of the **"Past"** moment, ***pulse-2*** the **"Present"**.

The Radial Body, Mechanics of Manifestation, Healing and Human Potential

There is much more to be revealed regarding the subject of the *Mechanics of Manifestation*. Our introduction to the ***Radial Body anatomy***, the ***Transduction Sequence*** and the ***Mechanics of Manifestation*** represents a beginning in the quest for genuine mastery over the manifest experience. In this *Kathara Level-2 Introduction* we will **not yet address** the highly pertinent aspects of *Manifestation Mechanics* that pertain to **direct Creation of matter and events** through **Conscious Thought Projection** and applied personal **Will Intention**, nor will we yet investigate the complex realities of **Space-Time Travel, Star Gate Passage** and the **Spiritualized-science of Ascension**. All of these subjects and more are *intimately intertwined* with **personal healing** and the **Imprint for Health**, **attaining mastery** over Manifestation Mechanics and the Bio-Spiritual processes of genuine **Spiritual Actualization**. All of these abilities, and many others too numerous to mention, presently exist as **dormant evolutionary potentials** for the Human species; potentials that were once, long ago, fully actualized and commonly expressed. The ***Kathara Bio-Spiritual Healing System*** represents a body of Master teachings that focus upon the manifestation and maintenance of **Holistic Health through Bio-Spiritual Mastery and Spiritual Actualizaton**. Exploration of subjects such as ***Merkaba Mechanics***, ***Manifestation Mechanics*** and the ***Radial Body Trion-Meajhe Field*** is an intrinsic part of furthering our abilities to embody and practice the ideals of Healing and Health through Holism.

Before we can begin to actualize the exciting potentials of freedom, love, power and knowledge that await Human re-discovery, we must first focus upon **healing the obstacles** that have long stood in our way of actualizing true Human potential. Our study of *Merkaba Mechanics* in this *Kathara Level-2 Introduction* has included a *necessary introduction* to **Manifestation Mechanics** and **Radial Body Anatomy**; the greatest physical and spiritual difficulties presently facing Humankind emerge from a **HIDDEN ILLNESS** that has held our species hostage within a progressively deteriorating, **unnatural Mortal body form**. The **underlying cause of ALL Human illness** is an **unnatural genetic mutation** that manifests THROUGH the **DNA Template** and **chemical DNA.** The **Casual Element** of this species genetic deviation from its original Divine Blueprint can only be found through study of the 15-Dimensional Anatomy, ***Primal Life Force Currents, Kathara Grid, Maharic Shield, Merkaba Field, DNA Template,*** the ***Radial Body*** and the intimate interplay that perpetually exists between these interwoven systems. It is precisely this ***species genetic deviation***, which has progressed and increased for over **200,000 years**, that has kept Humanity bound to a finite, limited world, unable to express the potentials that once existed as our birthright. Though the problem of this species "Internal Illness" is multi-faceted and highly complex, the **Soul-U-tions** through which progressive healing of the Human genome can be rendered are relatively simple. The **Soul-U-tions for healing** and restoring health to our species, and ourselves, **require familiarity with basic Kathara Grid, Manifestation Template, Radial Body, DNA Template and Merkaba Mechanics.** These are the elements of our 15-Dimensional anatomy within which the **Casual Element** of our genetic deviation is held, and thus these are the elements of our anatomy through which the **most rapid and permanent Healing Solutions** can be rendered.

In the following pages we will continue our investigation of the ***Transduction Sequence and the Miasmic Body***, in order to **identify** the regions of our 15-Dimensional Anatomy in which our hidden genetic mutations are created, stored and perpetuated. Once we can identify the regions in which the greatest healing is needed, we can then directly employ appropriate remedial measures. Working with the **Core Kathara Grid and Maharic Shield Divine Christos Blueprint** to reset the Imprint for Health (Kathara Level-1), while **simultaneously** working with the **elements of our anatomy within which the most significant remedial measures are needed**, will allow us to use Divine Blueprint technology to implement **expedited healing.** In this *Kathara Level-2 Introduction* we will examine the *Transduction-Manifestation Sequence* and how it's present **malfunction** creates the **Miasmic Body**, which creates the DNA Template distortion through which our **chemical DNA mutation**, and a phenomenon known as the "**Phantom**" or **"Shadow" Self** is brought into being within our manifest hologram. We will then continue our exploration of **Merkaba Mechanics** and **Radial Body Healing** applications, to discover the **"Secrets of the Veca Codes"**, and Photo-Sonic and Photo-Radionic Radial Body Healing technologies, through which our Kathara and Merkabic Healing applications can be tremendously intensified, amplified and expedited.

The 5 Radial Body-Hova Capsule-Trion-Meajhe Field Levels

and Dimensional, Triadic Phase Current, Density Level, Merkaba Phase, Hova Body, DNA Template, Axiatonal Line, Chakra, Identity Level and Memory Matrix Correspondences

NADIAL CAPSULE:(also called Telluric Capsule): Composed of **Dimensions-1-2-3** ***Telluric Kundalini Current*** frequency, **Density-1** Trion-Meajhe **Gross Matter Density**, 3-Dimensional **Nethra Phase Merkaba**, surrounds **Nada Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 1-2-3** and holds focus of consciousness in the D-1 *Lower-Physical-Atomic,* D-2 *Lower-Telluric-Emotional-Elemental* and D-3 *Lower-Mental* ***Taurenic Bodies*** of the embodied Density-1 **Incarnate Personality-*Tauren*** identity, the Sub-conscious, Instinctual and Ego-Reasoning aspects of the "***Mortal Mind***" and the **Personal** ***Taurenic Memory Matrix***. *Radis* rivulets are called "***Nadis Lines***" (or Telluradis Lines)

DORADIC CAPSULE: Composed of **Dimensions-4-5-6** ***Doradic Kundalini Current*** frequency, **Density-2** Trion-Meajhe **Semi-etheric Matter Density**, 6-Dimensional **Hallah Phase Merkaba**, surrounds **Alpha Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 4-5-6** and holds focus of consciousness in the D-4 *Astral-Mid-Physical-Atomic,* D-5 *Archetypal Mid-Telluric-Emotional-Elemental* and D-6 *Celestial Mid-Mental* ***Doradic Bodies*** of the embodied Density-2 **Incarnate Soul-*Dora*** identity, the "***Angelic Mind***" and the **Soul/Species** ***Akashic Memory Matrix***. *Radis* rivulets are called "***Doradis Lines***"

TEURIC CAPSULE: Composed of **Dimensions-7-8-9** ***Teuric Kundalini Current*** frequency, **Density-3** Trion-Meajhe **Etheric Matter Density**, 9-Dimensional **Quatra Phase Merkaba**, surrounds **Betcha Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 7-8-9** and holds focus of consciousness in the D-7 *Higher-Physical-Ketheric,* D-8 *Monadic Higher-Telluric-Emotional-Elemental* and D-9 *Keriatric-Higher-Mental* ***Metatronic Bodies*** of the embodied Density-3 **Incarnate Over-Soul-*Teura*** identity, the ***"Archangelic Mind"*** and the **Over-Soul/Planetary** ***Askashic-Eckashic Memory Matrix*** . *Radis* rivulets are called "***Teuradis Lines*"**.

MAHARIC CAPSULE: Composed of **Dimensions-10-11-12** ***Maharic "Christos" Pre-matter Hydro-plasmic Liquid Light Primal Current*** frequency, **Density-4** Trion-Meajhe **Pre-matter Density**, 12-Dimensional **Mahunta Phase Merkaba**, surrounds **Mahara Hova Body**, corresponds to **DNA Strand Templates, Axiatonal Lines and Chakras 10-11-12** and holds focus of consciousness in the D-10 *Christiac-Transcendent Pre-Physical,* D-11 *Buddhaic Transcendant Pre-Telluric-Emotional-Elemental* and D-12 *Nirvanic Transcendent Pre-Mental* ***Hydronic Bodies*** of the embodied Density-4 **Incarnate Avatar-*Dolus*** identity, the ***"Christos Avatar Mind"*** and the **Avatar/Galactic** ***Dolaric Memory Matrix***. *Radis* rivulets are called "*Maharadis Lines*".

RISHIAC CAPSULE: Composed of **Dimensions-13-14-15** ***M Kee-Ra-ShA Ante-matter Thermo-plasmic Gaseous Pre-Light Primal Current*** frequency, **Density-5** Trion-Meajhe **Ante-matter Density**, 15-Dimensional **Rahunta Phase Merkaba**, surrounds **Raja Hova Body**, corresponds to "Indigo Grail Line" **DNA Strand Templates, Axiatonal Lines and Chakras 13-14-15** and holds focus of consciousness in the D-13-14-15 *Transcendent Ante-matter* ***Hedronic Bodies*** of the Density-5 **Breneua Rishi** Consciousness Collective identity, the ***"Rishiac Universal Mind"*** and the **Rishi/ Universal** ***Rishic Memory Matrix of one 15-Dimensional Time Matrix.*** *Radis* rivulets are called "*Rishadis Lines*".

The **5 3-dimensional Levels** of the **Trion-Meajhe Field Radial Body** within the 15-Dimensional **Time Matrix** connect each Density Level of all manifest consciousness to the greater, non-dimensionalized structure of the **Energy Matrix** ***Khundaray*** **Primal Sound Vibration** ***Meajhe Field*** through the **Density-5** Ante-matter ***Kee-Ra-ShA*** **Primal Light** ***Trion Field***.

The Transduction Sequence and the Miasmic Body

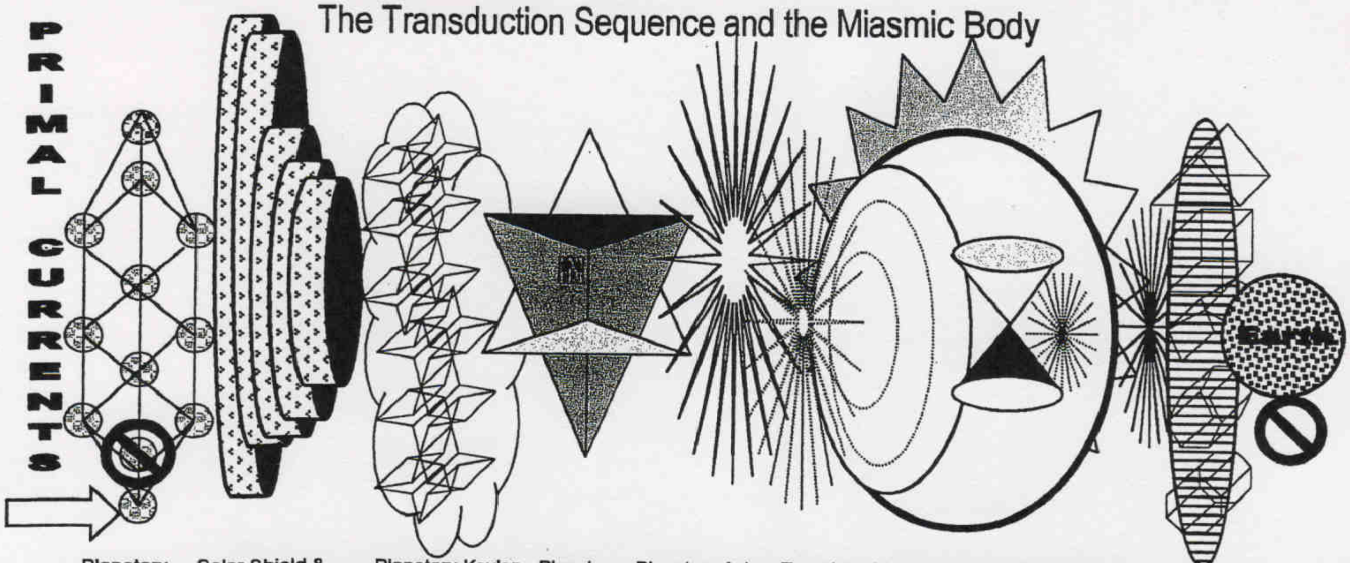
PRIMAL CURRENTS
Earth
Planetary Kathara Grid
Solar Shield & Planetary Shields
Planetary Keylon Crystal Body Thought-form
Planetary Merkaba Fields
Planetary Axiom Template and Axiom Lines
Planetary Hova Bodies, Vortices Ley Lines
Radial Body & Radis
Earth Mions-Dions

Personal Transduction Sequence

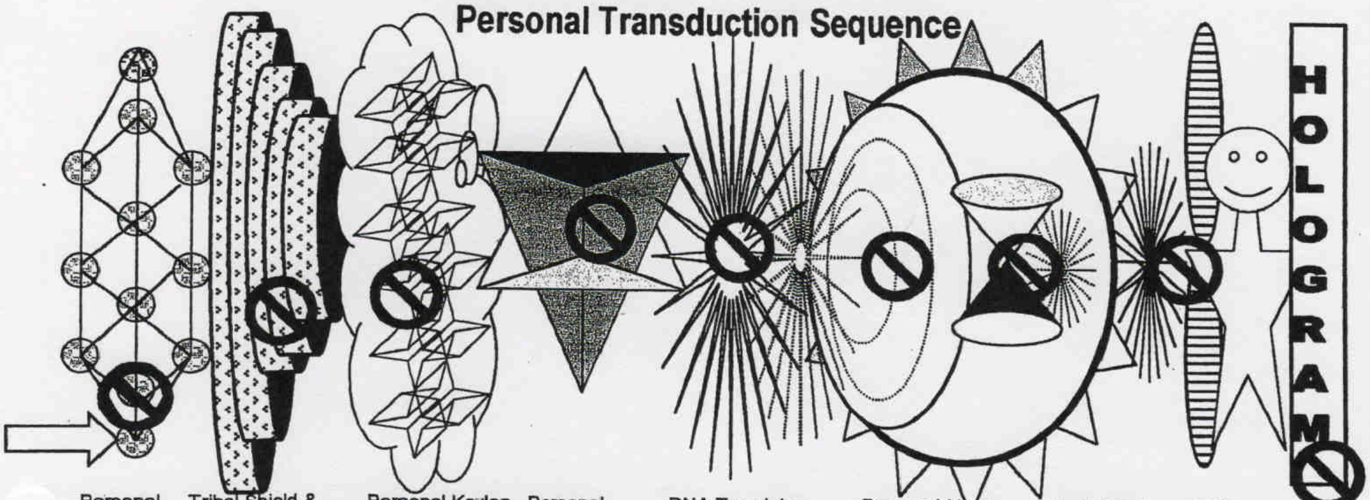
HOLOGRAM
Personal Kathara Grid
Tribal Shield & Personal Shields
Personal Keylon Crystal Body Thought-form
Personal Merkaba Fields
DNA Template and Axiom Lines
Personal Hova Bodies, Chakras Meridian Lines
Radial Body & Radis
DNA & body

Primal Life Force Currents ➡ Planetary Merkaba Field Partiki ➡ Kathara Grid -Partika-Particum ➡ Maharic Shield Divine Blueprint ➡ Keylon Morphogenetic "Thought-form" Field ➡ Merkaba Fields ➡ Kathara Grid DNA/RNA Template ➡ Axiom Lines ➡ Hova Bodies ➡ Auric Field ➡ Chakras ➡ Meridian Lines ➡ Trion-Meajhe Field 5 Radial Body Levels ⇨ Radis Lines ⇨ Mion & Dion Ionic Particulates ⇨ Strong and Weak Nuclear Force, Electromagnetism ⇨ Quarks, Muons, Measons, Sub-atomic & Atomic Particles and Anti-Particles ⇨ Blueprint translation into Nuclei, Molecules, Elements, Compounds, Chemicals ⇨ Chemical DNA Template and "Chemical Lens" of the manifest body form ⇨ Manifest Hologram ▭

⊘ **Miasms block energy flow and create a "Phantom-Shadow Self" running on opposite mathematical program; blocks natural energy flow.**

- The Personal ***Radial Body*** or Trion-Meajhe Field is composed of **5 Hova Body Capsules** that manifest as electromagnetic barriers or "**frequency fences**" between each set of 3 DNA Strand Templates, each set of 3 associated Axiom Lines, Chakras and Meridians and between each 3-dimensional layer of the Auric Field and corresponding level of personal multidimensional identity.
- Each of the **5** 3-dimensional **Harmonic Merkaba Fields** corresponding to the *5 Hova Bodies* of 15 Dimensional Anatomy *govern the function* of the corresponding **Dimensional Merkaba Fields**, **Radial Body level** and its imbedded **Memory Matrix**, via the mathematical program held in the personal Kathara Grid, Shields and **Crystal Body** template. The Kathara Grid program is passed from the **Keylontic Crystal Body** into the **DNA Template** via the **Dimensional and Harmonic Merkaba Fields** and the "mini-Merkaba Fields" of the Level-3 Kathara **Diodic Grid**.
- Each Hova Capsule **Radial Body level,** receiving its instructions from the DNA Template and 3rd dimensional Mental Body Hova Body level, in turn governs the function of the **chemically manifest DNA** and all body-consciousness systems that are built upon chemical genetic code.
- The **Harmonic Merkaba Field** governs the **particle spin axis** (Angular Rotation of Particle Spin) ratios of the energy units that make up its corresponding 3-dimensional system. Each **Harmonic Merkaba Field** corresponds to one **Hova Body**, one set of **3-dimensions** and the Matter Density Level characteristic of that 3-dimensional system. Each Hova Body, and thus each Harmonic Merkaba Field Level and Radial Body Level, correspond to one set of **3 DNA STRAND TEMPLATES** within the Human12-Strand Template.
- Full activation of **one Harmonic Merkaba Field** creates reciprocal activation of the corresponding set of ***3 DNA Strand Templates,*** which creates embodiment of corresponding levels of identity and consciousness (personality, Soul, Over-Soul, Avatar, Rishi etc). When one Harmonic Merkaba Field fully activates through natural activation of the corresponding **Shields** and **Crystal Body Keylons**, the Harmonic Merkaba Field becomes capable of **merging** with the Harmonic Merkaba Fields from other Density Levels, in a process frequently called **Transmutative Dimensional Ascension**.

- During the process of "**Ascension**", (also called "space-time travel", "wave-riding", "Star Gate passage", "Teleportation" etc.) the **Harmonic Merkaba Fields** of more than one Density Level merge, progressively shifting the **two or more different sets of 3-dimensional particles** into a **common particle spin axis angle and spin speed**. The corresponding **Axiom Lines, Hova Bodies, Auric Field levels, Chakras, Meridian Lines** and **Radial Body Levels** also merge.
- When Harmonic Merkaba Fields, and their corresponding Inner Templar anatomy, merge, the **Fire Codes** between the corresponding **DNA Strand Templates activate, causing the DNA Strand Templates to merge,** which creates a reciprocal reaction called **"DNA Strand BRAIDING"** between corresponding **chemical DNA strands**. (The "Double Heli" configuration of "one strand" remains, as the chemical components of each strand overlay together to become a single Double Helix containing the chemical programs of multiple DNA strands.)
- When the **Fire Codes** between strands in the DNA Template activate, dormant sequences of chemical DNA called " **Transient Turnstile DNA Sequences**" "turn on" within the **Hydrogen Bonds** that link each DNA "Ladder Rung" together, producing a **transient trans-harmonic element** called **Celestalline** within the chemical DNA.
- Natural Braiding of the chemical DNA strands and activation of the chemical Turnstile DNA Sequences create a series of **sub-atomic bio-chemical DNA responses** that allows for Internal **Atomic Particle/Anti-particle Fusion** and **Fission.** Natural **Internal Atomic Fusion-Fission** enables the manifest atomic structure to temporarily de-manifest, returning to "*spirit form*", by re-entering the electrostatic **Trion-Meajhe Field** of the corresponding **Radial Body Levels.**
- During the Ascension process, the Harmonic Merkaba Fields of **2 or more Density Levels** merge, forming various **PHASES of Merkaba** activation that create a **Trans-Harmonic Merkaba Field** that is called the **Merkaba VEHICLE.** The 3-Dimensional **NETHRA Phase** Merkaba Field of Density-1accelerates in rotation speed to become the 6-dimensional **HALLAH Phase**, 9-Dimensional **Quatra Phase**, 12-Dimensional **Mahunta Phase** or 15-Dimensional **Rahunta Phase** Harmonic Merkaba Vehicle.

- As a phase of the Trans-Harmonic Merkaba Vehicle activates, the corresponding lower-dimensional **Hova Body and Radial Body Level dissolve**, their content of energy and consciousness **expanding** into the next Hova Body and Radial Body Level, forming a **Trans-Harmonic Radial Body "FORM CONSTANT"** within the **Merkaba Vehicle.**

- The Trans-Harmonic Radial Body and Merkaba Vehicle allow the consciousness and biological form to "de-manifest" from the lower dimensional Density Level. The identity can then **re-manifest** in another Density Level or space-time coordinate by projecting itself into the desired location, AS the *electrostatic* ***Trion-Meajhe Field*** **Trans-Harmonic Radial Body,** then returning the DNA Template activation level and Harmonic Merkaba Spin ratios to that of the new location for re-manifestation.

- When the Personal **Merkaba Field** is used to form the **Trans-Harmonic Radial Body** in the process of Ascension, the Merkaba Field is called a **MERKABA VEHICLE.** A ***Merkaba Vehicle*** is an electrostatic **Trion-Meajhe Field** of Primal Pre-Light-Sound through which the consciousness (AKA "spirit") can move itself, *and its atomic structure*, through **thought-projection**, to a desired location beyond the "veils" of the Planetary, Galactic and Universal Hova Capsule-Radial Body-Form Constant barriers.

- Conscious mastery of the Personal Merkaba Field, through which the organic **internally created Merkaba Vehicle** Trans-Harmonic Radial Body becomes available for conscious direction, implies **conscious control of DNA Template function.** The potentialities of **genuine Ascension** (etc…) and mastery of Bio-Spiritual Consciousness are **inborn attributes** of the Human 12-Strand DNA Template. Due to a **species de-evolutionary genetic mutation** that compromises natural function of the Personal Merkaba Field and DNA, the reality of Ascension on a species level has been lost to all but a few on Earth for over 200,000 years (re: Masters Templar Coursebook).

- The potentials of **Bio-Spiritual Actualization** CAN be returned to humankind through **reverse-mutation** of genetic dysfunction within the human genome and DNA Template. Healing the genetic template can be achieved by **manually employing** specific practices of **Kathara "Maharic Shield" Core Template Dynamics** in **combination** with **Harmonic Merkaba Mechanics** through which the many eons of biological miasmic distortion can be progressively and rapidly cleared from the Personal Kathara Grid and DNA Template.

- Restoration of the D-12 Pre-matter Divine Blueprint within the Human Kathara Grid and DNA Template will eventually **free the Personal Merkaba Field** for use as a **Trans-Harmonic Merkaba Vehicle.** Kathara Healing applications, and use of the D-12 Pre-matter Maharata "Christos" Current to reset the biological D-12 Pre-matter Divine Blueprint, are the beginning steps in reclaiming Bio-Spiritual Mastery.

- The function of the personal **Merkaba Field** and **DNA Template** are not only associated with abilities of **Atomic Transmutation** and **Dimensional Ascension**, they are directly linked to **every aspect of body-mind-spirit function**, including **biological longevity** and **health, memory, intelligence,** and **metabolic, chemical, mental** and **emotional balance** and stability.

- Most significantly, the **Merkaba Field/DNA Connection** is the Bio-Spiritual conduit through which the consciousness of an embodied individual is directly linked to the common Perpetual Force of Energy and Eternal Source of Intelligent Creation that is "**God**".

- When a being's Merkaba Field and DNA Template are functioning upon the innate program of **Universal Primal Order**, as it is held within the Pre-matter **Universal Divine Blueprint** of the D-12 "Christos" Liquid-Light field, the natural and perpetual supply of **Primal Life Force Currents** and "spiritual" consciousness emanating eternally from Source-God, flow through an open energetic channel or "**Pillar of Light**", which is called the "Um Shaddai Ur" (Pillar of First Cause Light-Sound') into the manifest embodiment. Such a being is called a "**Christed Being**". This implies that the integrity of its original D-12 Pre-matter Divine Blueprint has not been structurally compromised, and thus the perpetual Life Force Currents, consciousness, sentience, intelligence, wisdom, power, Will and LOVE of Source can freely express through the Um Shaddai Ur, into the manifest hologram of the Time Matrix through the Christed embodiment.

- If the Kathara Grid, DNA Template and Merkaba Field of a being are damaged, and no longer carry the **original mathematical-geometrical program** of the **D-12 Divine Blueprint** through which the being entered space-time, the flow of Primal Life Force Currents, the Um Shaddai Ur Pillar and thus the active consciousness connection to God will be distorted, reduced or severed at the dimensional level to which the template distortion corresponds.

- A manifest being incapable of running the Primal Life Force Currents of the Um Shaddi Ur "Pillar" becomes **biologically finite**, its consciousness "trapped in the Time Matrix", even following death of the physical body. According to the **natural Laws of Creation Physics** as set by Source-God, this condition of **finite mortality** will remain until the being's **Kathara Grid Template and Merkaba Field connection to** the D-12 **Universal Divine "Christos" Blueprint** and Universal Trion-Meajhe Field Um Shaddi Ur Pillar is restored.
- Healing Core Template genetic distortions that are severe enough to cause dysfunction of the Personal Merkaba Field first require **resetting of the D-12 Divine Blueprint** through activation of the **Personal Maharic Shield.** Until the Body-Mind-Spirit system has reached a **Critical Mass** of D-12 Divine Blueprint holding through frequent and progressive manual running of the D-12 Maharata Current, healing Core Template genetic distortions **also requires** simultaneous, direct and repeated **manual realignment and restoration** of **Personal Merkaba Field function** and **Radial Body Trion-Meajhe Field** integrity.
- If there **are distortions within the Kathra Grid or DNA Template** of an individual or species, the Personal Merkaba Fields will not function in harmony with the **natural energy circulation rhythms** of the **Planetary, Galactic, Universal** and **Time Matrix Merkaba Fields** and Universal Kathara Grid. Such Kathara Grid distortions create tangible **genetic and consciousness mutation** within biological organisms, and can divert the **intended evolutionary destiny** of an embodied consciousness or species on to a divergent path of **de-evolution and consciousness fragmentation** unintended by the original plan encoded within the D-12 Pre-matter Christos Divine Blueprint.
- Distortions with the DNA Template and resulting **Miasms** can be purged through **manual resetting of the D-12 Pre-matter Divine Blueprint,** via repeated manual activation of the Personal **Maharic Shield**. The activated Maharic Shield runs the Density-4 D-12 **Maharata Primal Life Force "Christos" Current** through the Kathara Grid and form embodiment, progressively resetting the original mathematical coding of the personal Divine Blueprint within the Kathara Grid and DNA Template. Use of the **Maharic Shield techniques** in *Kathara Level-1* sets this process of **Core Template Regeneration** in motion.

- The Core Template "Divine Blueprint" healing process can be expedited through applied **Christos-Merkabic Healing Mechanics** such as the ***Tribal Shield Activation*** (Technique-7), ***Merkaba Salutation*** (Technique-8), ***Merkaba Trinity Key Induction*** (Technique-9) and ***Kee-Ra-ShA Activation Sequence*** (Technique-12).

- Divine Blueprint Healing can be expedited further through **manual restoration** of the personal Trion-Meajhe Field **Radial Body** and its inherent **Memory Matrix. Radial Body Healing,** when **combined** with ***Kathara Maharic Shield*** and ***Christos-Merkabic Healing*** dynamics is the most powerful and rapid Holistic Healing technology available.

- **Radial Body Healing** dynamics, which also serve to progressively restore the organic **Memory Matirx,** involve use of the specific **mathematical programs** that govern formation of the **Radial Body Trion-Meajhe Field.** These mathematical programs are called **VECA CODES.**

- **Technique-10** ***Restoring the Trion Field-Photo-Radionic Healing*** **and Technique-*11Restoring the Meajhe Field-Photo-Sonic Healing,*** provide introduction to **Radial Body Healing Mechanics.** Use of these techniques, in **combination** with **Technique-2** ***The Maharic Seal and Liquid Light Cleanse,*** and **Merkabic Healing Techniques 6-12, will greatly expedite** the personal healing process while **significantly increasing** the body-mind-spirit abilities to **receive, hold/sustain and transmit greater quantity and quality** of interdimensional frequency in the Client Healing Facilitation process.

Radial Body Healing, the VECA CODES, DNA and Merkaba

- Veca Codes, also called the "I AM" or "**Immanuyana Sequence**" are very specific **SYMBOL CODES**, each possessing corresponding **Sound Tones** called ***Arieas*** and **energy movements** called ***Seurias***, which correspond to the Cosmic, Universal, Galactic, Planetary and Personal **Trion-Meajhe Field Radial Body Levels**.

- There are **5 categories of Veca Codes**, the Planetary **Low-Veca *Codes of Manu*** (Taurenic-Akashic Codes), the Galactic **Mid-Veca *Codes of Immanu*** (Akashic-Eckashic Codes), the Universal **Level-1 High-Veca *Codes of Immanuel*** (Dolaric-Rishic Codes), the Cosmic **Level-2 High Veca *Codes of Immanuyana*** (Triadic-Polaric-Eckatic Codes) and the God-Seed **Yunasai *Codes of Eckasha*** (Source Codes).

- Beginning applications of Veca Code Radial Body Healing begin with restoring the Trion-Meajhe Field Radial Body through resetting the ***Yunasai*-Eckasha and High Veca *Immanuyanas* and *Immanuel* Codes**, through **frequent manual Direct Induction** of the their corresponding **Symbol Codes** and **Sound Tones/Arieas**.

- The most powerful **Yunasai-Eckasha Codes** are **PHOTO-SONIC Codes**, carrying the Mathematical Programs of the ***Khundaray Current*** Primal Sound ***Meajhe Field*** AND their bridging **God-Seed vibration-rhythms** connecting to Source. The **Level-2 *Immanuyana*** and **Level-1 *Immanuel* High Veca Codes** are **PHOTO-RADIONIC Codes**, carrying the Mathematical Programs of the Trion-Meajhe Field **Khundaray Primal Sound** AND **Kee-Ra-ShA Primal Light Currents.**

- **Consistent use** of **Veca Code Radial Body Healing**, in combination with **Merkabic and Kathara Healing** applications, will progressively **clear the Miasmic Body** and heal, balance, re-integrate, and restore the Divine Blueprint within, the energies of the **Phantom-Shadow Self** in all levels of the Body-Mind-Spirit system.

- More advanced applications of Veca Code technologies include use of the **Mid-Veca *Immanu*** and **Low-Veca *Manu* Codes,** and the corresponding ***Seurias*** (body movements to direct frequency), following a period of initial induction and activation of the Yunasai and High Veca Codes. Full Spectrum Veca Code technologies are introduced in the "***Dance of Life***" Masters Spiritual Development Program.

- **Induction,** or **IMBEDDING** of the High Veca Codes into the 15-Dimensional Anatomy involves placing the corresponding **Symbol Code Sequence** (mathematical program) over the **regions of the physical body** that correspond to the appropriate aspects of the Kathara Level-2 **Crystal Seals Grid** and Level-3 **Diodic Grid.** Once the Symbol is in place, the **breath** is used to direct **D-12 Maharata Current** through each **Symbol Code** and **body region.** Use of the **Maharata Current** "carrier wave" (via **activation of the Maharic Seal/Shield**) is **required** to **induce** the code's **mathematical program** into the **Diodic** and **Crystal Seals Grid,** where the program will then naturally transfer into the corresponding areas of the **Level-1 Kathara 12-Tree Grid** core template.

- Once the High Veca and Eckasha Code Sequences are Imbedded/induced into the Level-1 Kathara Grid, they are then **Activated** in the Kathara Grid by **TONING** the corresponding **Veca Code Ariea Sequence.** The longer toning is sustained, the greater will be the quantity and quality of Veca Code Activation.

- Once Activated, the Veca Code programs travel upward through the stages of the **Transduction-Manifestation Sequence** from the **Kathara Grid**, to the **Shields**, into the **Keylon Crystal Body Morphogenetic Field,** then into the **Merkaba Field**, **DNA Template** and **Hova Body/ Radial Body Levels**, temporarily restoring the **Christiac Merkaba Field, Radial Body** and **Memory Matrix Divine Blueprint**. **Frequent and continued use** of the Veca Codes will progressively build **Critical Mass** of the **Divine Blueprint program simultaneously** within **ALL levels** of the 15-Dimensional Anatomy, eventually culminating in expedited full restoration of the D-12 Maharic Shield blueprint within the Body-Mind-Spirit system, via the Radial Body Trion-Meajhe Field.

- Activation of the Yunasai and High Veca Codes within the personal body **prior to conducting a Kathara Client Healing Session** will cause the facilitator's **Level-3 DIODIC GRID to activate**, temporarily **transmitting Veca Code sub-harmonics.** This will cause the **minute Diodic Chakras** in the facilitator's **FINGER TIPS** to activate **if the Palm Chakras** have been activated with either pure **Maharata Current** or a selected **Triadic Phase Current.** The facilitator will then **automatically transmit** Veca Code sub-harmonics into the client's **Diodic Grid** while conducting general Kathara Healing applications, allowing the client to benefit from temporary **Radial Body healing support**, which will **amplify the effects** of all healing applications.

- A more **Temporary Imbedding and Activation** of the Veca Code program can be achieved through **OPTICAL-PINEAL INDUCTION**, whereby the **eyes** are focused for a time upon the **Image** of the **Symbol Code.** The mathematical Veca Code program enters the **Pineal Seals** through the **Optical Currents** and **6th Chakra**, an then travels for Imbedding/induction through the **Central Vertical Current** into the body regions and Crystal Seals Grid areas to which the Veca Code corresponds. The **Veca Tones/Arieas** are then used to Activate the Imbedded Veca Code Sequence.
- In **Optical-Pineal Induction**, the Veca Program looses a bit a **frequency charge/ strength** during is travel in the Central Vertical Current. This creates a **more temporary imbedding and activation** of the Veca Codes than that which is achieved through Direct Induction. Optical-Pineal Induction is useful for **daily amplification** of the Vecas, and for Kathara Healing Client Sessions and planetary grid "site work", when used in **combination** with the **full induction procedure** employed at least **once per week**.
- To receive **full benefit** of **Veca Code Radial Body Healing** applications, **Technique-2 *The Kee-Ra-ShA Activation Sequence*** should be fully run at least **ONE TIME**. *Technique-2* can also be used occasionally thereafter if desired to create **intensification and amplification** of the Primal Life Force Currents during Kathara, Merkabic and Veca Code Healing technologies.
- Frequent use of **Technique-13: The "Song of Lyra"** Christos Invocation will amplify all Kathara, Merkabic and Veca Code Radial Body Healing applications.

The Christos-Trion-Meajhe Fields, High-Veca Codes and Universal Life Force Currents

One 15-Dimensional Time Matrix with 5 Harmonic Universes

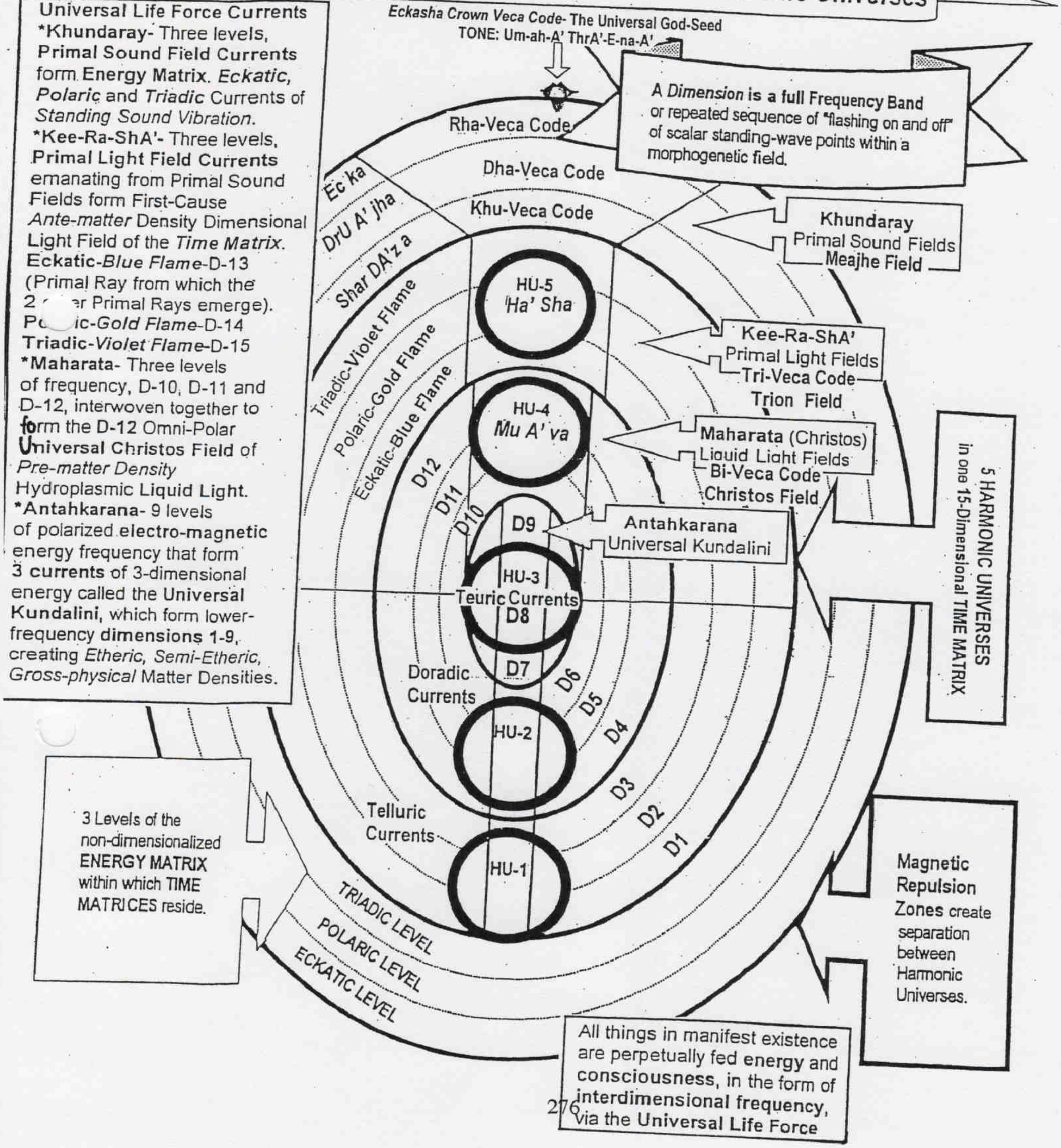

Universal Life Force Currents
***Khundaray**- Three levels, **Primal Sound Field Currents** form **Energy Matrix**. *Eckatic, Polaric* and *Triadic* Currents of *Standing Sound Vibration*.
***Kee-Ra-ShA'**- Three levels, **Primal Light Field Currents** emanating from Primal Sound Fields form First-Cause *Ante-matter* Density Dimensional Light Field of the *Time Matrix*.
Eckatic-*Blue Flame*-D-13 (Primal Ray from which the 2 other Primal Rays emerge).
Polaric-*Gold Flame*-D-14
Triadic-*Violet Flame*-D-15
***Maharata**- Three levels of frequency, D-10, D-11 and D-12, interwoven together to form the **D-12** Omni-Polar **Universal Christos Field** of *Pre-matter Density* Hydroplasmic **Liquid Light**.
***Antahkarana**- 9 levels of polarized **electro-magnetic** energy frequency that form **3 currents** of 3-dimensional energy called the **Universal Kundalini**, which form lower-frequency **dimensions** 1-9, creating *Etheric, Semi-Etheric, Gross-physical* Matter Densities.

Eckasha Crown Veca Code- The Universal God-Seed
TONE: Um-ah-A' ThrA'-E-na-A'

A *Dimension* is a full Frequency Band or repeated sequence of "flashing on and off" of scalar standing-wave points within a morphogenetic field.

Khundaray
Primal Sound Fields
Meajhe Field

Kee-Ra-ShA'
Primal Light Fields
Tri-Veca Code
Trion Field

Maharata (Christos)
Liquid Light Fields
Bi-Veca Code
Christos Field

Antahkarana
Universal Kundalini

5 HARMONIC UNIVERSES
in one 15-Dimensional TIME MATRIX

3 Levels of the non-dimensionalized **ENERGY MATRIX** within which TIME MATRICES reside.

Magnetic Repulsion Zones create separation between Harmonic Universes.

All things in manifest existence are perpetually fed energy and **consciousness**, in the form of **interdimensional frequency**, via the **Universal Life Force**

Rha-Veca Code
Dha-Veca Code
Khu-Veca Code
Ec ka
DrU A' jha
Shar DA'z a
Triadic-Violet Flame
Polaric-Gold Flame
Eckatic-Blue Flame
HU-5
Ha' Sha
HU-4
Mu A' va
D12
D11
D10
D9
HU-3
Teuric Currents
D8
D7
D6
D5
D4
Doradic Currents
HU-2
D3
D2
D1
Telluric Currents
HU-1
TRIADIC LEVEL
POLARIC LEVEL
ECKATIC LEVEL

Technique # 4

Restoring the Trion Field-Photo-Radionic Healing

Regular use of this technique will strengthen and improve RRT Planetary healing Applications (excerpt from Kathara Level-2)

Creating a Personal Bio-field Link to the Universal Christos-Trion-Meajhe Field

Masters Merkaba Mechanics

© 2001 Ashayana and Azurtanya Deane

Exercise-1: **Once** per day, do **Maharic Insta-Seal** then practice **each set of Veca** tones separately while fixating **visual focus** on corresponding **Veca-Code** symbol; spend at least **three or four minutes** gazing at each symbol (optical-pineal induction) while **toning** corresponding tones (core template vibrational induction) With practice this will progressively expand the personal consciousness while activating the personal Christos-Trion-Meajhe Field of your Christos-Rishiac-Ascended Master identity levels within the cellular structure.

Exercise-2: Once per week, directly after completing **Exercise-1**, place all Veca-Code symbols over designated body areas, activate Maharic Shield, breathe D-12 Maharata Current through each symbol and body area to **induce code**, using the Bi-Tri-Khu-Dha-Ra Sequence, then begin **toning** the corresponding tones following the Bi-Tri-Khu-Dha-Ra Sequence. Memorize the tone sequence of the 5 High-Veca Codes as if it were a "**song**" and relax, close eyes and sing/tone this "song" (Psonn) for **5 to 10 minutes**. The longer you tone, the more frequency power you build within the body and the stronger your link will be to the Universal Christos-Trion-Meajhe Field. This exercise can also be used with groups, either singularly to create and amplify Christos-Trion-Meajhe Field links or just prior to beginning RRT work. Frequent and consistent use of these two exercises will progressively develop the spin-rate capacity of the personal Christos Merkaba Field, while expanding higer consciousness within the body and progressively enhancing the health, immunity and integrity of the Mind-Body-Spirit system.

The High-Veca Codes
Universal Time Codes for Primal Frequency Access

Cosmic Codes of Immanuyana

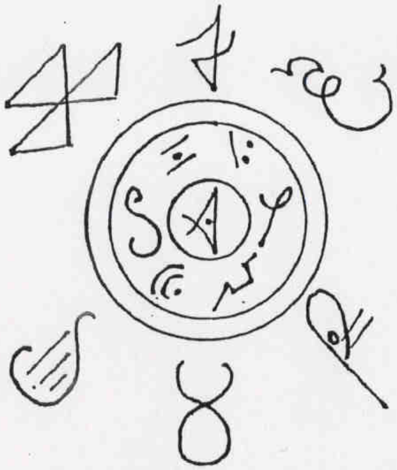

The Rha-Veca Code
Primal Sound-3 Eckatic
called the *Ec ka*
TONE: "*Rha Ec' ka*"
Symbol Placement:
Tailbone

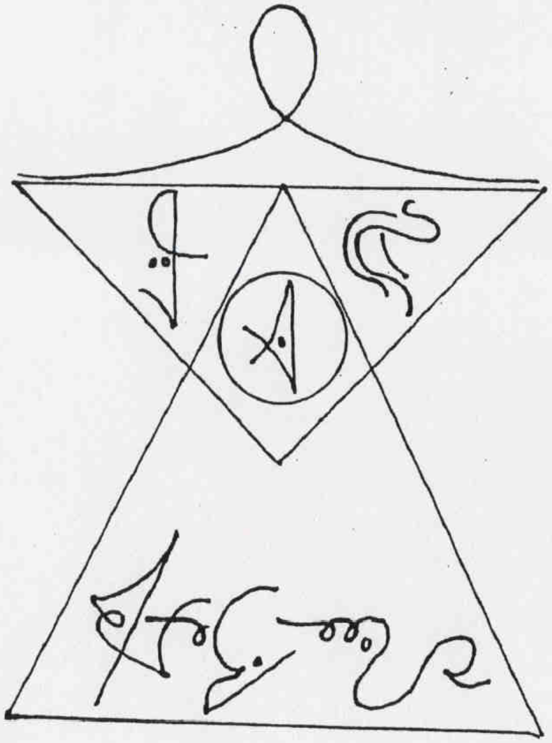

The Dha-Veca Code
Primal Sound-2 Polaric
called the *DrU A' jha*
TONE: "[illegible]*a' Dru A' jha*"
Symbo[illegible]lacement:
Chakra-2 (just below naval)

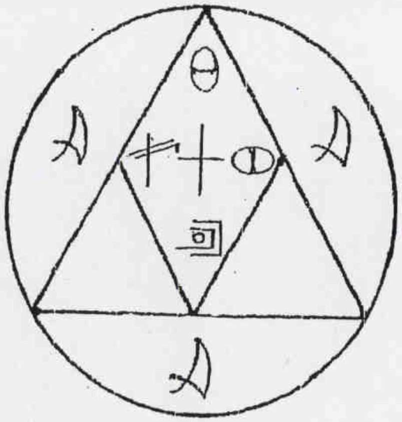

The Khu-Veca Code
Primal Sound-1 Triadic
called the *Shar dA'z a*
TONE: "*Kh*[illegible]*har DA'z a*"
Symbol Placement:
Chakra-8 Front (clavicle)

Universal Codes of Immanuel

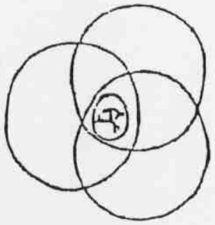

The Tri-Veca Code
Density-5 Dimensions 13-14-15
called the *Ha' Sha*
TONE: "*Kee' Ra ShA Ha Sha*"
Symbol placement:
Left Foot bottom

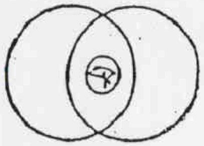

The Bi-Veca Code
nsity-4 Dimensions 10-11-12
called the *Mu A' va*
"*Ma ha ra' ta Mu A' va*"
Symbol placement:
Right Foot bottom

Technique # 11

Restoring the Meajhe Field-Photo-Sonic Healing

1. Use **Technique-6:** ***The Maharic Quick Seal*** to activate your Maharic Shield, Seal and Maharata Current.
2. **Stare at image** of the **Eckasha Code** for a few seconds, then gently begin to **"Cross Eyes"** until you can see a **"Double Image"** of the *Eckasha* when looking at the image. **Hold " Eye Cross"** focus while taking **three rapid breaths**; on each **EXHALE** projecting a **stream** of (D-6) **INDIGO current** from your **6th "3rd Eye" Chakra** at forehead. (This braids D-6 Optical Currents into a D-12 Current to carry the Eckasha program into the Pineal Seals.). Un-cross eyes, relax visual focus and close eyes for a moment or two.
3. **INHALE** several breaths of **Pale Silver D-12 Maharata Current** up from the **Earth Core** through your **Maharic Cord**, on each **EXHALE** depositing the Maharata Current in the **Pineal Gland** at the center of the brain, to form a Pale Silver **"Maharata Ball"** in the Pineal.
4. Breathe gently and **return visual focus to the image of the Eckasha**, focusing upon the **Eckasha Core** ***REUCHE'*** (the "Double Circles and their content at center of Eckasha).
5. **EXHALE** while projecting a **stream of INDIGO current** from the **6th Chakra** at forehead, into the **REUCHE** symbol before your eyes. **INHALE** while drawing the **INDIGO current stream** with the **REUCHE image at its end** back through the **6th Chakra** and into the **Maharata Ball** in the **Pineal.**
6. **INHALE** one more breath of **Maharata Current** up from **Earth Core** to the **Maharata Ball** at the **Pineal**, then use a forceful EXHALE to PUSH Eckasha-Reuche encoded Maharata Ball out to the right, to the "Mentor" position over the right shoulder. Visualize the Maharata Ball growing to the size of a "Basket Ball" over the right shoulder, and visualize the image of the Reuche as expanding to fill the inside of the Maharata Ball.
7. **INHALE forcefully** while drawing the **Maharata Ball** back to a position **centered around your head**, then use **three breaths** to **INHALE Maharata Current** up from **Earth's Core to Pineal** and **EXHALE** the **Maharata Current outward** to fill **and expand the Maharata Ball.**
8. Visualize the **Maharata Ball** expanding outward from around your head, until the Maharata Ball **fully encompasses your body and entire Chakra System**, from **Chakra-14** 36" above head to **Chakra-13** in Earth Core. Imagine that your entire body is now surrounded by a 3-dimensional "Eckasha Tear-Drop" shaped CLEAR CRYSTAL CAPSULE. The Eckasha Code Photo-Sonic Capsule is now temporarily activated in your Inner Templar system.

Eckasha Crown Veca Code- The Universal God-Seed

Tone: Um ......⇩ah...........A' ⇧ ThrA'.......⇩E'- ⇩na......A

Symbol Placement: Pineal Induction via Chakra-6 "3rd Eye"

The Yunasai-Eckasha God-seed Code

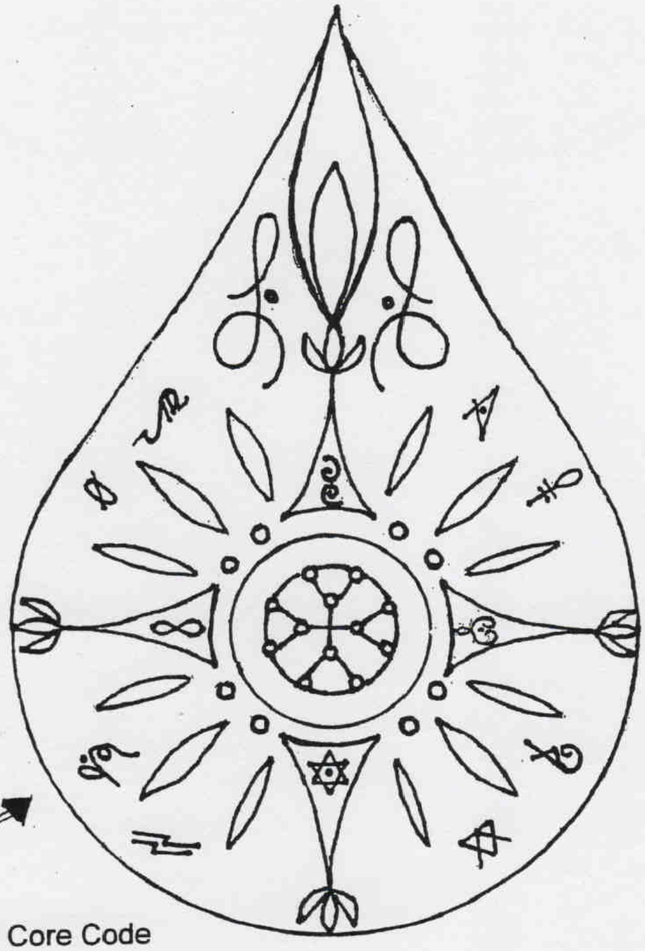

The REUCHE' Eckasha Core Code

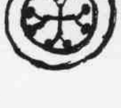

Technique # 13

The Song of Lyra Christos Invocation

The Vectos Invocations

The Vectos Compendium Declariti

The Vectors Collective Declarations

(All self incarnations in all Time Vectors declaring in Unity as the Avatar Self)
English translation from Elohei-Anuhazi Emerald Covenant Decree Texts:
© 2001 Ashayana Deane

The Mah'-ah' Bi-Vectus Declaritum
Declaration of the Maharata

(Intention Decree AS the personal D-12 Christos Avatar Maharata self identity)

I decree now in this moment of KNOWING that all things here within me
are hence forth restored to the original Divine Intention of the ONESELF that is MINE.

I release now all bonds of manifest illusion that place veils of SELF-forgetting before mine eyes.
I revoke all separations of duality through which my Heart is rendered numb with disillusion and discord.

In this moment I reclaim the Whole of my Condition, reaching forth into the depths of all creation to now call my SPIRIT home.

I AM the Sun of the Radiant Glowing. I AM the HEART of Divine LOVE. I AM the Still Point of the Face of All Creation. I AM the Voice of the Universal Song.

I decree now in full remembered Knowing that I AM ALL and WE are ONE.
In setting forth this decree of my Eternal Divine Power, I shall render Divine Intention only, through which the All and I are equally served
in the Highest Order of our collective Best Divine Intention.

I call now unto the many Selves and Shadows ensconced within the illusions of time, and decree unto thee that I AM THIS, I AM.
I AM the wind that brushes thy faces.
I AM the rain that renews thy field.
I AM the Whisper of the Breathe of God within YOU.
WE NOW are the expression of Divine Heart, Mind and Hand.
WE are the Rumbling Hymn Song of the earthen worlds of matter.
WE are the Prana-Breath of God Eternal, manifest.
We are the Waters of Eternal Life forever outward flowing.
We are the cooling, life force Flame of the Cosmic Fire expressed.
In our UNION now and always, declare WE in our Wholeness,
that all dissidence is rendered spent, all conflict NOW made healed.

As DIVINE NAME THIS.................(your Christos spiritual name)
I AS WE, the Maharata KHUM BI-Vec'TUS, (D-12 Christed Avatar Divine Blueprint self identity), DO NOW DECREE through the Divine Love, Wisdom and Power that is MINE....to express now AS this moment, the condition of **Mah'-ah BI-Vec'TI UN UR'-A-OR'-NaM' –OOR** (Christos beginning and ending NOW-completion of a manifest expression in its originally intended Divine Fulfillment, its past-present-future full expression made in the present time, clearing all duality, disharmony and distortion. Anchoring of the Divine Christos Maharata Blueprint in the moment). Upon this Invocation Divine Intention is rendered manifest in all conditions to which I now assign.

The Mah'-ah' Bi-Vec'tus Inavhoki
Invocation of the Maharata

(Creation Decree Invocation AS the personal D-12 Christos Avatar Maharata self identity)

© 2001 Ashayana Deane

From this Still Point of my Wholeness in manifest decision,
I Decree in Sovereign Knowing, here expressed at my Command *(lovingly stated absolute intention)* ***the Divine Will and Intention of the ONE-SELF that is ONE GOD.***
The ONE-SELF that IS ONE GOD THIS I AM.

Mah–Hah-Rah-Tah KHUM BI-Vec'TUS ....(name)
I as D-12 Christos Avatar self.......(name)
IN-A-VHO'-KI UN' E-BLE'UM BI-Vec'TI
Invoke by Divine Loving Command Now, by the Power of the Universal Christos,
Mah'-ah BI-Vec'TI UN UR'-A- OR'-NaM' – OOR
the D-12 Divine Christos Blueprint past-present-future; NOW the Divine Sea of Liquid Light anchors in this moment.
E'-stA Un'-tA E'-Sa Tra'zd-Jha' HA'-A Rha
Absolutely now established, always and forever, in eternal abiding reverent Love of the Eternal ALL-ONE
'tah-TA Um a Ah-ShA'-Lum
Blessed Be all embracing, given forth from the Still Point of Eternal Peace
Thah-A'-Jha in'ta DO A"
Spoken AS IT IS STATED, As SO It IS DONE
Um Sha'-DI UR'- A ah Khum' TUn
All Embracing Pillar of First Cause Light, I AS THIS expression now.
In the name of the Melchizedek Cloister Emerald Order Eieyani,
Holy Order of the Yunasai,
Sacred House of ONE.

English Translation:

Ma-Ha-Ra-Ta KHUM BI-Vec'TUS ....(name)
(Khum = "I AS") (Ma-Ha-Ra-Ta BI-Vec'tus = D-12 Christos Avatar Self)
iN-A-VHO'-KI UN' E-BLE'UM BI-Vec'TI
(*in-A-vhO-kI*= invoke by Divine Loving Command), (*UN*= Now), *(E-BLE'UM*= by the power of),
(Bi-Vec'ti = Universal Christos)
Mah'-ah BI-Vec'TI UN' UR'-A- OR'-NaM' – OOR
(Mah'ah Bi-Vec'tI= D-12 Divine Christos Blueprint past-present-future)
(UN Ur'-A- Or'-Nam' –OOR= Now the Divine Sea of Liquid Light anchors in this moment)
E'-stA Un'-tA E'-Sa Tra'zd-Jha' HA'-A Rha
(E'StA Un'TA E'Sa = Absolutely now established, always and forever)
(Tra'zd-Jha'= in eternal abiding reverent Love) (HA'-A Rha = of the Eternal ALL-ONE)
P'tah-TA Um a Ah-ShA'-Lum
(P'tah-TA UM = Blessed Be, all embracing) ("a" in this context = given forth from)
(Ah-ShA-Lum = the Still Point of Eternal Peace)
Thah-A'-Jha in'ta DO A"
(Tha-A-Jha inta= Spoken AS IT IS STATED) (DO A"= As SO It IS DONE)
Um Sha-DI' UR' A' ah Khum' TUn
(Um Sha- DI' UR'A = All Embracing Pillar of First Cause Light)
(ah Khum' TUn = I AS THIS expression now)

Chinese Medical Qigong Therapy:

A Comprehensive Clinical Text

中国气功医疗

DR. JERRY ALAN JOHNSON PH.D., DTCM, DMQ (CHINA)

CHINESE MEDICAL QIGONG THERAPY:

A COMPREHENSIVE CLINICAL TEXT

JERRY ALAN JOHNSON, PH.D., D.T.C.M., D.M.Q. (CHINA)

EDITED BY
JAMPA MACKENZIE STEWART, D.O.M., L.AC.
AND
MADELEINE H. HOWELL, M.F.T., M.Q.T.

First published in May 2000 by:
The International Institute of Medical Qigong
P.O. Box 52144
Pacific Grove, CA 93950
U.S.A.

© 2000 Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)

All rights reserved under the International and Pan-American copyright conventions. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher (The International Institute of Medical Qigong). Reviewers may quote brief passages.

ISBN# 1-885246-08-0
Printed in the United States of America.

Disclaimer:

Qigong medicine is not intended to replace orthodox medicine, but rather to complement it. The meditations, practices, techniques and prescriptions described herein are powerful and may be too mentally and physically demanding for some people. The readers should therefore use their own discretion and consult a doctor of Medical Qigong therapy, an acupuncturist, medical doctor, or mental health professional before engaging in these exercises and meditations. The author, the International Institute of Medical Qigong, and the publishers are neither liable or responsible to any person or entity with respect to any loss or damage caused, or alleged to be caused, directly or indirectly by reading or following the instructions for any condition, or interpreting information provided in this text. The treatments mentioned in this book are not meant to be used as symptomatic prescriptions. The treatment of specific organs, channels, channel points, and prescriptions must always be selected based on the thorough understanding of the origin of the patient's disease. If an ailment is severe, or if symptoms persist, please consult a medical professional immediately.

Throughout the text I will refer the doctor to prescribe herbs for certain conditions along with Medical Qigong therapy. All Medical Qigong prescriptions, exercises and "homework" assigned to patients in Section X "Treatment of Internal Diseases with Qi Emission Therapy" and Section XI "Specialized Qi Emission Therapy for Pediatrics, Geriatrics, Gynecology, Neurology, Psychology, Oncology, and Surgery" should also be included with herbal prescriptions, as well as the regulation of the patient's diet and living environment. Herbal prescriptions will vary according to the patient's constitution, condition and specific illness, and must be treated accordingly only by a doctor or herbalists qualified to prescribe Chinese medical herbs.

Throughout the text I have used the term *doctor* when referring to professional practitioners who use energetic medicine to treat patients. The word "doctor" means "to teach". I believe that the foremost duty of any doctor of medicine (Western or Chinese) should be as educator, to teach his or her patients knowledge and skills for the prevention and treatment of disease and injury. Currently, the official title, *Doctor of Medical Qigong Therapy* is only licensed and used in China.

TABLE OF CONTENTS

FORWARD XXIII

PREFACE XXXI

ACKNOWLEDGMENTS XXXV

INTRODUCTION XXXVII

SECTION I

FOUNDATIONS OF CHINESE ENERGETIC MEDICINE 1

CHAPTER 1 INTRODUCTION TO MEDICAL QIGONG 3
- THE FIVE DOMINIONS OF ENERGY 3
- DEFINING THE ENERGY OF YIN AND YANG 5
- UNDERSTANDING THE CONCEPT OF QI 5
- DIFFERENT SCHOOLS OF QIGONG 5
- MEDICAL QIGONG DEFINED 6
- MEDICAL QIGONG TRAINING IN CHINA 7
- MEDICAL QIGONG TRAINING IN THE UNITED STATES 8
- TRADITIONAL CHINESE MEDICINE AND MEDICAL QIGONG THERAPY 14
- ACUPUNCTURE THERAPY 15
- HERBAL THERAPY 18
- CHINESE MASSAGE THERAPY 19
- MEDICAL QIGONG THERAPY 22

CHAPTER 2 ENERGETIC FORMATION OF THE HUMAN BODY 25
- ENERGETIC EMBRYOLOGICAL DEVELOPMENT 25
- ENERGY, MATTER, AND SPIRITUAL INTERACTIONS 26
- UNDERSTANDING FETAL TOXINS 26
- DEVELOPMENTAL SEQUENCE 27
- THE TEN LUNAR MONTHS OF CREATION 29
- POSTNATAL ENERGY DEVELOPMENT 41
- ENERGETIC EMBRYOLOGICAL OVERVIEW 41
- THE ETERNAL SOUL 42
- SUMMARY 51

CHAPTER 3 PHYSICAL DEVELOPMENT AND STRUCTURAL FORMATION 53
- FASCIAL DEVELOPMENT AND ENERGY FLOW 53
- YIN AND YANG STRUCTURAL FORMATION 53
- THE YIN AND YANG OF BODY ASPECTS 55
- THE FIVE ELEMENTS 56
- CONGENITAL CONSTITUTIONS 57
- CLASSIFICATION OF THE FIVE CONSTITUTIONS 57
- THE WOOD CONSTITUTION 58
- THE FIRE CONSTITUTION 60
- THE EARTH CONSTITUTION 61
- THE METAL CONSTITUTION 62
- THE WATER CONSTITUTION 63

COMBINED CONSTITUTIONS 64
PERSONALITY CONSTITUTIONS OF THE EIGHT EXTRAORDINARY VESSELS 65
THE YAO IMAGE AND PHYSICAL ANATOMY 66
IDENTIFYING INTROVERTED AND EXTROVERTED STRUCTURES 71
CONGENITAL AND ACQUIRED CELLULAR PATTERNS 73
SUMMARY 73

CHAPTER 4 THE FIVE ENERGIES OF THE HUMAN BODY 75
SOUND ENERGY RESONANCES 75
LIGHT ENERGY 77
MAGNETIC ENERGY 82
GENERATED HEAT 84
ELECTRICITY 86

CHAPTER 5 THE THREE DANTIANS AND THE TAIJI POLE 89
THE THREE DANTIANS 89
THE LOWER FIELD OF ELIXIR 90
THE MIDDLE FIELD OF ELIXIR 95
THE UPPER FIELD OF ELIXIR 98
THE FUNCTIONAL ASPECTS OF THE THREE DANTIANS 102
THE DANTIAN'S YIN AND YANG ENERGETIC CHAMBERS 104
THE DOCTOR'S PROJECTED AURA FIELDS 105
THE TAIJI POLE 105
THE TWELVE GATES OF THE CHAKRA SYSTEM 109
TAKING PRECAUTIONS 115
THE TWELVE EARTHLY BRANCHES AND THE TWELVE CHAKRA GATES 115

CHAPTER 6 THE TWELVE PRIMARY CHANNELS AND ORGANS 117
INTRODUCTION 117
CLASSIFICATION OF CHANNELS 118
THE CHANNELS' RELATIONSHIP TO QI AND BLOOD 121
FUNCTION OF THE CHANNELS 122
INTERNAL AND EXTERNAL CHANNEL FLOW 122
ORGAN AND CHANNEL FUNDAMENTALS 123
THE GALL BLADDER CHANNELS 124
THE LIVER CHANNELS 126
THE LUNG CHANNELS 128
THE LARGE INTESTINE CHANNELS 130
THE STOMACH CHANNELS 132
THE SPLEEN CHANNELS 134
THE HEART CHANNELS 136
THE SMALL INTESTINE CHANNELS 138
THE BLADDER CHANNELS 140
THE KIDNEY CHANNELS 142
THE PERICARDIUM CHANNELS 145
THE TRIPLE BURNER CHANNELS 147
UNDERSTANDING CHANNEL PATHOLOGY 151

CHAPTER 7 THE EIGHT EXTRAORDINARY VESSELS AND COLLATERALS 155
THE EIGHT EXTRAORDINARY VESSELS 155
THE FUNCTION OF THE EIGHT EXTRAORDINARY VESSELS 155
THE EIGHT EXTRAORDINARY VESSELS AND MEDICAL QIGONG THERAPY 157

CLINICAL USE OF THE EIGHT EXTRAORDINARY VESSELS 157
ENERGETIC FUNCTIONS OF THE EIGHT EXTRAORDINARY VESSELS 158
THE GOVERNING AND CONCEPTION VESSELS 158
THE GOVERNING VESSEL OR DU VESSEL 159
THE CONCEPTION VESSEL OR REN VESSEL 162
THE THRUSTING VESSELS AND THE BELT VESSELS 164
THE BELT VESSEL OR DAI VESSEL 166
THE YIN AND YANG HEEL VESSELS OR QIAO VESSELS 168
THE YIN HEEL VESSELS 168
THE YANG HEEL VESSELS 169
THE YIN AND YANG LINKING VESSELS OR WEI VESSELS 171
THE YIN LINKING VESSEL 171
THE YANG LINKING VESSELS 172
SUMMARY OF EIGHT EXTRAORDINARY VESSEL PATHOLOGY 173
THE FIFTEEN MAJOR COLLATERALS (CONNECTING VESSELS) 175
THE TWELVE DIVERGENT CHANNELS 180
THE TWELVE MUSCLE/TENDON CHANNELS 186
THE TWELVE SKIN ZONES 197
THE CONNECTIVE TISSUE OF THE TWELVE SKIN ZONES 197
CHANNEL AND COLLATERAL THERAPY 200
CONTROLLING QI EXTENSION THROUGH THE CHANNELS 201

CHAPTER 8 THE BODY'S ENERGETIC POINTS 203
INTRODUCTION TO ENERGETIC POINTS 203
THE FOUR WAYS POINTS ARE FORMED 203
SCIENTIFIC RESEARCH OF ACUPOINTS 203
THE FOUR CATEGORIES OF ENERGETIC POINTS 204
THE TWO GATES OF ENERGETIC POINTS 204
THE THREE FUNCTIONS OF ENERGETIC POINTS 205
POINT NAMES 207
CLASSIFICATION OF ENERGETIC POINTS 207
SUMMARY OF POINTS 212

CHAPTER 9 THE EXTRAORDINARY YANG ORGANS 215
THE UTERUS 215
THE BRAIN 217
THE GALL BLADDER 218
THE MARROW 218
THE BONES 218
THE BLOOD VESSELS 221

SECTION II

THE THREE OUTER FORCES: HEAVEN, EARTH & MAN 223

CHAPTER 10 INTRODUCTION TO THE THREE OUTER FORCES 225
THE NATURAL POWER OF THE THREE OUTER FORCES 225
THE THREE OUTER FORCE'S NINE TREASURES 225
THE THREE OUTER FORCES AND CHINESE ALCHEMY 225
HEAVENLY QI 226
EARTHLY QI 228
MAN'S QI 229

CHAPTER 11 THE THREE TREASURES OF HEAVEN: SUN, MOON, AND STARS 231
DEFINING HEAVEN QI 231
INFLUENCES OF HEAVENLY QI ON THE HUMAN BODY 232
BODY RHYTHMS AND ENERGETIC PATTERNS 232
CYCLES OF THE SUN 234
THE TWELVE PI HEXAGRAMS 234
THE FOUR PRINCIPAL TIME PERIODS OF QIGONG PRACTICE 236
THE FOUR SEASONAL PEAK TRANSITIONS 237
THE SIX PRINCIPAL SEASONS OF TRANSITION 239
THE DAILY QI, BLOOD, AND HEAT CYCLE 239
GATHERING THE SUN'S ESSENCE MEDITATION 240
CYCLES OF THE MOON 242
PRENATAL EIGHT TRIGRAMS AND THE MOON 244
CHOOSING THE PROPER DIRECTION FOR TRAINING QI 244
CYCLES OF THE STARS 248
EARTHLY BRANCHES, HEAVENLY STEMS, AND NINE PALACES 254
SUMMARY 255

CHAPTER 12 THE THREE TREASURES OF EARTH:EARTH, WATER, AND WIND 259
INTRODUCTION TO THE THREE TREASURES OF EARTH 259
GATHERING ENERGY FROM THE EARTH'S ENERGETIC FIELDS 260
TECHNIQUES FOR CULTIVATING EARTH QI 260
GATHERING QI FROM TREES, BUSHES, AND FLOWERS 261
GATHERING QI FROM MOUNTAINS, VALLEYS, AND DESERTS 263
GATHERING QI FROM OCEANS, LAKES, AND STREAMS 263
GATHERING QI FROM THE SIX DIRECTIONS 264
ADDITIONAL TECHNIQUES FOR TRAINING EARTH QI 265
THE EARTHLY TREASURE OF FOOD 266
UNDERSTANDING THE YIN OR YANG ASPECT OF FOOD 266
UNDERSTANDING THE YIN OR YANG ASPECT OF HERBS 266
ENERGY, FOOD, AND DIET 267
USING THE FIVE FLAVORS OF FOOD 268
GUIDELINES FOR THE FIVE ELEMENTAL TASTES 270
OVERVIEW OF THE FIVE ELEMENTAL TASTES 271
SUMMARY 271

CHAPTER 13 THE THREE TREASURES OF MAN: JING, QI, AND SHEN 273
PRENATAL AND POSTNATAL JING, QI, AND SHEN TRANSFORMATIONS 274
JING: ESSENCE OF LIFE 276
THE YIN AND YANG ASPECTS OF JING 276
PRENATAL JING 276
POSTNATAL JING 277
THE INTERACTION OF PRENATAL AND POSTNATAL JING 277
QI: VITAL LIFE-FORCE ENERGY 278
EIGHTEEN CLASSIFICATIONS OF QI 279
THE SIX FUNCTIONS OF QI 281
QI AND ENERGETIC MEDICINE 283
QI AND THE BODY'S ENERGETIC STRUCTURE 283
SHEN: THE SPIRIT OF LIFE 290
SHENGONG SCHOOLS OF MEDITATIONS AND TRAINING 290
PRENATAL AND POSTNATAL SHEN 291

THE FIVE SPIRITUAL ASPECTS AND YIN ORGAN COMPONENTS OF SHEN 291
THE LIVER STORES THE ESSENCE SPIRIT "HUN" 292
THE HEART STORES THE ESSENCE SPIRIT "SHEN" 295
THE SPLEEN STORES THE ESSENCE SPIRIT "YI" 296
THE LUNGS STORE THE ESSENCE SPIRIT "PO" 296
THE KIDNEYS STORE THE ESSENCE SPIRIT "ZHI" 298
THE FIVE YIN ORGAN CULTIVATION OF SHEN 298
SHENGONG CULTIVATION METHODS 300
THE SIX TRANSPORTATION'S OF SHEN 301
THE EIGHT SUPERNATURAL POWERS 304
SOUL PROJECTION, SPIRIT PROJECTION AND ENERGY PROJECTION 306

SECTION III
DAO YIN TRAINING 311

CHAPTER 14 INTRODUCTION TO DAO YIN TRAINING 313
STATIC QIGONG TRAINING 314
DYNAMIC QIGONG TRAINING 314
PRINCIPLES OF YIN AND YANG IN DAO YIN TRAINING 314
THREE FOUNDATIONAL METHODS OF DAO YIN TRAINING 315

CHAPTER 15 POSTURAL DAO YIN TRAINING 317
STATIC POSTURE DAO YIN TRAINING 317
CHOOSING THE PROPER POSTURE 318
PRIMARY AND SECONDARY POSTURES 318
THE EIGHTEEN RULES OF PROPER MEDICAL QIGONG STRUCTURE 321
DYNAMIC POSTURE DAO YIN TRAINING 330
USING THE ENERGY OF THE HANDS TO LEAD QI 332
MEDICAL QIGONG WALKING THERAPY 334
WALKING COMBINED WITH COLOR THERAPY 338
WALKING COMBINED WITH SOUND THERAPY 338
THE THERAPEUTIC EFFECTS OF MEDICAL QIGONG WALKING THERAPY 338
THREE STEPPING METHODS USED TO TREAT CANCER 342
ENDING AND CLOSING THE WALKING THERAPY 344

CHAPTER 16 RESPIRATORY DAO YIN 349
THE PURPOSE OF RESPIRATORY DAO YIN 349
THE ROLE OF THE LUNGS IN QI CULTIVATION 349
DEEP INHALATION AND EXHALATION BREATHING METHODS 349
SEVEN DAO YIN BREATHING TECHNIQUES 350
RESPIRATORY GUIDING AND REGULATING 353
RESPIRATORY TONIFICATION AND PURGING METHODS 354
BASIC ESSENTIALS OF RESPIRATORY TRAINING 354
BREATHING PATTERNS FOR COLLECTING AND REGULATING QI 355
THE FOUR PROGRESSIVE STAGES OF RESPIRATION 356
DURATION TIME FOR RESPIRATION THERAPY 357
EMOTIONAL EFFECTS ON BREATHING PATTERNS 357
RESPIRATORY DAO YIN CANCER PRESCRIPTIONS 357

CHAPTER 17 MENTAL DAO YIN TRAINING 359
INTRODUCTION 359
THE PRENATAL AND POSTNATAL MIND 360

Benign and Malignant Thought Patterns ... 361
The Importance of Mind Concentration ... 361
Regulating the Mind ... 362
The Three Categories of Mental Concentration ... 363
Essentials of Training Mind Concentration ... 364
Concentrative Meditation and Insight Meditation ... 365
The Three Stages of Mental Tranquility ... 366
The Three Processes of The Mind ... 366
Creating The Bridge Of Light ... 367
Reprogramming The Mind To Heal ... 369
Using the Imagination ... 369
Twelve Stages to Transcend and Transform ... 370
Summary ... 374

Section IV

Qi Deviations in Qigong Training ... 375

Chapter 18 Introduction To Qi Deviations ... 377
The Main Causes of Qi Deviations ... 377
Correcting Qi Deviations ... 379
Shen Disturbances and Emotional Dysfunctions ... 386
Uncontrollable Muscle Deviations ... 391
Rectifying Qi Deviations Using Taiji Ruler ... 391

Chapter 19 Disorders and Deviations of the Soul and Spirit ... 393
Introduction to Spiritual Dimensions ... 393
Evil States and Mental Delusions ... 395
Projected Spirit Demons from the Subconscious Mind ... 397
Actual Encounters with Spirit Demons ... 397
Overview ... 399
Multiple Personality Disorder (Dissociative Identity Disorder) ... 400
Soul Retrieval and Medical Qigong Therapy ... 401
Wounding and Closing the Spirit ... 403
Healing the Spirit ... 405
Emotional Rebirthing Patterns ... 406
Soul Retrieval Meditation ... 406
The Door to the Chamber of Memories ... 411
Finding the Life Purpose ... 411

Section V

Differential Diagnosis of Energetic Principles ... 413

Chapter 20 Introduction To Differential Diagnosis ... 415
Diagnosis in Energetic Medicine ... 415
Etiology in Medical Qigong Therapy ... 415
The Growth and Peak Cycles of Jing, Qi and Shen ... 416
Disorders of the Human Body ... 417
The Differentiation of Syndromes ... 419
The Categorization of Signs and Symptoms ... 420
Diagnosis and Treatment of the Patient ... 420
Understanding the Theory of Energetic Counteraction ... 422

CHAPTER 21 DIAGNOSIS ACCORDING TO THE EIGHT ENERGETIC PRINCIPLES 423
INTERNAL AND EXTERNAL PATHOGENIC FACTORS 423
HOT AND COLD 428
EXCESS AND DEFICIENT 430
YIN AND YANG THEORY 431
DIAGNOSING YIN AND YANG IN THE BODY 432
SIGNS OF YIN AND YANG 435
YIN AND YANG CATEGORIZATION OF TRAINING 436
PULSE DIAGNOSIS OF THE EIGHT ENERGETIC PRINCIPLES 436
PULSE QUALITIES AND CLASSIFICATIONS 437
THE CLASSICAL PULSE DIAGNOSTIC POSITIONS AND DEPTH 438
OVERVIEW 439

CHAPTER 22 DIAGNOSIS ACCORDING TO QI, BLOOD, AND BODY FLUID DYSFUNCTIONS 441
INTRODUCTION 441
DISORDERS OF THE QI 441
LEAKAGE AND LOSS OF QI 442
EXCESS QI DISORDERS 444
QI STAGNATION DISORDERS 444
REBELLIOUS QI DISORDERS 445
TURBID QI 445
RECTIFYING QI DISORDERS 446
DISORDERS OF THE BLOOD 447
DEFICIENCY OF BLOOD 447
THREE DISORDERS OF BODY FLUIDS 448
TYPES OF PHLEGM 450

CHAPTER 23 DIAGNOSIS ACCORDING TO THE FIVE-ELEMENT THEORIES 451
INTRODUCTION TO THE FIVE ELEMENTS 451
THE FIVE ELEMENTS AND YIN AND YANG THEORY 451
THE CREATIVE OR GENERATIVE CYCLE 453
THE CONTROLLING OR RESTRICTIVE CYCLE 453
THE INVADING OR OVERACTING CYCLE 454
THE INSULTING OR COUNTERACTING CYCLE 455
THE FIVE ELEMENTAL TONIFICATION CYCLE 456
FIVE YIN ENERGETIC ORGAN MANIFESTATIONS 456

CHAPTER 24 OTHER DIAGNOSTIC SYSTEMS AND TECHNIQUES 459
DIAGNOSIS ACCORDING TO THE SIX STAGES 459
DIAGNOSIS ACCORDING TO THE FOUR LEVELS 460
DIAGNOSIS ACCORDING TO THE TRIPLE BURNERS 462

CHAPTER 25 CLINICAL EXAMINATION AND DIAGNOSIS 465
SENSORY, INTUITIVE AND PERCEPTUAL DIAGNOSIS 465
SENSORY DIAGNOSIS 465
CLINICAL INTAKE, EXAMINATION AND EVALUATION 465
EXAMINATION FOR DIAGNOSIS 467
DEVISING A TREATMENT PRINCIPLE 468
TREATMENT STRATEGY 468
TREATMENT 469
AFTER THE TREATMENT 469

ALTERNATIVE TREATMENT STRATEGIES 470
INTUITIVE AND PERCEPTUAL DIAGNOSIS 470
THE METAPHYSICAL DIMENSIONS OF MEDICAL QIGONG HEALING 470
THE AWAKENING 471
LEARNING INTUITIVE AND PERCEPTUAL DIAGNOSIS 471
RECEIVING AND INTERPRETING THE PATIENT'S FIELDS OF BIO-INFORMATION 472
ACCESSING THE KNOWLEDGE STORED WITHIN THE WUJI 473
UNDERSTANDING PAST-LIFE REGRESSION 474
LEARNING MEDICAL QIGONG TECHNIQUES 476
PERCEPTUAL COMMUNICATION TRAINING 477
APPLICATIONS OF INTUITIVE AND PERCEPTUAL QI DIAGNOSIS 479
DIAGNOSIS BASED ON FLAT-PALM DETECTION (HAND SENSING) 479
DIAGNOSIS BASED ON KINESTHETIC AND EMPATHIC QI ABSORPTION 480
HAND DIAGNOSIS BASED ON BODY ASSOCIATION 481
DIAGNOSIS BY OBSERVING AURA FIELDS 483
DIAGNOSIS BASED ON INTENTION 487
LONG DISTANCE MEDICAL QIGONG THERAPY 487
FAITH PROJECTION AND MEDICAL QIGONG THERAPY 490
AUXILIARY MEDICAL QIGONG HEALING MODALITIES 492
USING MEDICAL QIGONG TO HEAL BY PROJECTING CHINESE CHARACTERS 492
USING MEDICAL QIGONG WITH SUPPLEMENTARY APPARATUSES 492

SECTION VI
ESTABLISHING THE MEDICAL QIGONG CLINIC 497

CHAPTER 26 TREATING THE PATIENT 499
SETTING A SPACE/TIME FOUNDATION 499
ESTABLISHING THE MEDICAL QIGONG TREATMENT ROOM 500
ESTABLISHING A RAPPORT WITH THE PATIENT 500
THE FOUR MAJOR PRIORITIES OF CLINICAL QIGONG HEALING 503
PREPARING THE SELF TO TREAT THE PATIENT 504
TREATING THE PATIENT 509
MEDICAL QIGONG TREATMENTS 515
TREATING MULTIPLE PATIENTS 519
ENDING THE TREATMENT 520

CHAPTER 27 COMBINING MEDICAL QIGONG THERAPY WITH OTHER HEALING MODALITIES .. 523
COMBINING MEDICAL QIGONG THERAPY WITH TCM MODALITIES 523
COMBINING MEDICAL QIGONG THERAPY WITH WESTERN MEDICAL MODALITIES 525

CHAPTER 28 PRECAUTIONS AND ETHICAL CONSIDERATIONS FOR THE QIGONG DOCTOR 527
AVOIDING ENERGETIC DEPLETION 527
PREVENTING THE INVASION OF TURBID QI 528
GUIDELINES FOR THE MEDICAL QIGONG DOCTOR 529
WU ZANG (FIVE ORGAN) MEDITATION FOR ENERGETIC PROTECTION 530
CREATING A PROTECTIVE ENERGETIC FORCE FIELD 533
CLINICAL ETHICS 534
CONSCIOUS AND SUBCONSCIOUS CLINICAL INTERACTIONS 535
ESTABLISHING DOCTOR AND PATIENT AGREEMENTS 536
POWER DYNAMICS AND ETHICAL BEHAVIOR 537
PITFALLS 538

CLINICAL BURNOUT AND MEDICAL QIGONG THERAPY 538
ENERGETIC BOUNDARY INVASION 539
STRENGTHENING THE DOCTOR'S ENERGETIC BOUNDARIES 540
CAUSES OF SPIRITUAL DRYNESS 540

SECTION VII
THE TREATMENT PRINCIPLES OF MEDICAL QIGONG THERAPY 543

CHAPTER 29 PRINCIPLES AND PARAMETERS OF MEDICAL QIGONG THERAPY 545
THE FIVE METHODS FOR TREATING PATIENTS 545
THE NINE STAGES OF TREATMENT 545
THE THREE WONDERS OF MEDICAL QIGONG THERAPY 546
CONTRAINDICATIONS FOR MEDICAL QIGONG THERAPY 548

CHAPTER 30 BALANCING THE BODY'S ENERGETIC SYSTEM 551
PURGING, TONIFYING, AND REGULATING 551
THE THREE TREATMENT MODALITIES 551
PURGATION TECHNIQUES 552
TONIFICATION AND PURGATION METHODS 552
TONIFICATION AND PURGATION GUIDELINES 554
UNDERSTANDING MEDICAL QIGONG REGULATION 554
QI REGULATION AND GUIDING PRINCIPLES 555
ENERGY BALANCING PROCEDURES 555
CLINICAL OBSERVATIONS OF PATHOLOGICAL ENERGETIC PATTERNS 556
LOCATIONS OF ENERGETIC BLOCKAGES 556
TYPES OF ENERGETIC COMPLICATIONS 559
PATIENT'S ENERGETIC FACIAL DISTORTION 562
PATIENT RESISTANCE TO ENERGETIC THERAPY 562
UNBLOCKING THE FIVE PASSES 562
FURTHER COMPLICATIONS 564

CHAPTER 31 THE MAGIC SQUARE OF THE YELLOW RIVER 565
THE MAGIC SQUARE AND EIGHT TRIGRAMS 568
THE MAGIC SQUARE AND THE EIGHT EXTRAORDINARY VESSELS 569
UNDERSTANDING THE METAPHYSICAL ASPECT OF NUMBERS 569
THE FIVE ELEMENT ENERGETIC FLOW OF THE MAGIC SQUARE 570
CROSSOVER ENERGY PATTERNS 570

CHAPTER 32 HEALING THE PATIENT'S EMOTIONAL TRAUMAS 573
INTRODUCTION 573
THE PHYSIOLOGICAL EFFECTS OF BENIGN AND CHRONIC EMOTIONS 573
FIVE LEVELS OF FEELING AND THEIR GOVERNING ORGANS 573
STAGES OF EMOTIONAL HISTORY AND RESOLUTION 576
ASSISTING THE PATIENT IN ESTABLISHING HEALTHY BOUNDARIES 578
REJECTION, BETRAYAL, AND ABANDONMENT 578
EMOTIONAL SECURITY 579
HEALING EMOTIONAL TRAUMAS 580
PULLING OUT THE PAIN MEDITATION 581
UNDERSTANDING WHY A PATIENT WON'T HEAL 582
FACING THE DEATH OF A PATIENT 583
SUMMARY 585

SECTION VIII
QI EMITTING METHODS 587

CHAPTER 33 ENERGY EXTENSION AND QI MANIPULATION TECHNIQUES 589
ENERGY PROJECTION 589
TYPES OF ENERGY EXTENSION 589
CONTROLLING ENERGETIC PROJECTIONS 591
HAND POSTURES FOR QI EMISSION 592
THE EXTENDED FAN PALM 592
THE SWORD FINGERS 594
THE ONE FINGER SKILL 595
THE INVISIBLE NEEDLE PALM 597
THE DRAGON'S MOUTH PALM 598
THE KNEADING TIGER PALM 598
THE FIVE THUNDER FINGERS 599

CHAPTER 34 QI EXTENSION, GUIDANCE, AND REGULATION TECHNIQUES 601
QI EXTENSION TECHNIQUES 601
GUIDING THE LINEAR FLOW OF ENERGY 601
CIRCULATING THE QI IN A CIRCLING PATTERN 603
SPIRALING ENERGY FLOW 605
ENERGETIC CUPPING 607
GUIDING THE STATIONARY FLOW OF ENERGY 607
GUIDING THE HOT AND COLD FLOW OF ENERGY 610
CLINICAL APPLICATION OF YIN AND YANG ENERGY EXTENSION 612
GUIDING THE ENERGETIC FLOW OF THE FIVE ELEMENTS 612
SUMMARY OF QI EMITTING METHODS 616

CHAPTER 35 THE INVISIBLE NEEDLE THEORY AND APPLICATION 619
INTRODUCTION 619
PREPARATION 619
ANGLE OF INSERTION FOR TONIFICATION OR SEDATION 620
THE HAND POSTURES OF THE INVISIBLE NEEDLE 621
TIME AND DURATION OF TREATMENT 622
REMOVING THE INVISIBLE NEEDLES AND ENDING THE TREATMENT 622

CHAPTER 36 ENERGETIC POINT THERAPY 623
CHANNEL POINT TREATMENTS AND PRESCRIPTIONS 623
CONTRAINDICATIONS 627
CHANNEL POINT MEDITATION 627

CHAPTER 37 MEDICAL QIGONG MASSAGE 631
INTRODUCTION TO TISSUE FUNCTION 631
TWO APPROACHES IN TREATMENT 632
TISSUE REGULATION THROUGH MEDICAL QIGONG MASSAGE 633
THE FIVE ELEMENTAL QI MASSAGE THERAPY 633
YIN AND YANG ENERGETIC POINT MANIPULATION 636
TREATING THE TISSUE AREA OR ENERGETIC POINT 637
ADDITIONAL GUIDELINES 637

SECTION IX

MEDICAL QIGONG EXERCISES AND PRESCRIPTIONS 639

CHAPTER 38 INTRODUCTION TO MEDICAL QIGONG PRESCRIPTIONS AND REGULATION EXERCISES 641
HISTORY OF MEDICAL QIGONG PRESCRIPTIONS 641
THE FUNCTION OF MEDICAL QIGONG PRESCRIPTIONS 641
MEDICAL QIGONG AND THE IMMUNE SYSTEM 642
THE FIVE STAGES OF HEALING 645
THE THREE CATEGORIES OF MEDICAL QIGONG PRESCRIPTIONS 645
CHOOSING PRESCRIPTIONS ACCORDING TO THE PATIENT'S QI CIRCULATION 646
MEDICAL QIGONG PRESCRIPTIONS FOR GENERAL SYNDROMES 647
NUMBER OF BREATHS REQUIRED IN MEDICAL QIGONG PRESCRIPTIONS 647

CHAPTER 39 SPECIFIC ORGAN SELF-MASSAGE PRESCRIPTIONS 649
INTRODUCTION 649
HEART MASSAGE 649
LIVER MASSAGE 651
LUNG MASSAGE 652
SPLEEN & STOMACH MASSAGE 652
KIDNEY MASSAGE 653

CHAPTER 40 HEALING SOUND THERAPY AND PRESCRIPTIONS 657
INTRODUCTION AND HISTORY 657
SOUND VOLUME RESONATION 657
HIGH-, MIDDLE-, AND LOW-PITCHED SOUND 658
BREATH AND MIND CONTROL 658
AUDIBLE AND INAUDIBLE SOUND RESONATION 659
THE SIX HEALING SOUND PRESCRIPTIONS 660
THE PROPER NUMBER OF BREATHS FOR HEALING SOUND PRACTICE 665
COMBINING MULTIPLE ORGAN SOUNDS 666
OLD MAN SEARCHING FOR THE REFLECTION OF THE MOON AT THE BOTTOM OF TIDE POOL 666
CONTRAINDICATIONS FOR THE SIX HEALING SOUND THERAPY 668
SOUND RESONATION THERAPY FOR TUMOR AND CANCER 668
PREPARATION FOR TUMOR AND CANCER SOUND THERAPY 668
THE PROPER AMOUNT OF TONE REGULATIONS 669
SPECIALIZED PRESCRIPTIONS FOR TUMORS AND CANCER 669
CONTRAINDICATIONS FOR TUMOR AND CANCER SOUND THERAPY 673

CHAPTER 41 MEDICAL QIGONG REGULATION EXERCISES 675
INTRODUCTION 675
STARTING AND ENDING THE PRESCRIPTIONS 675
MEDICAL QIGONG CULTIVATION EXERCISES 676
EXTERNAL QI CULTIVATION 676
INTERNAL QI CULTIVATION 677
REGULATING THE BODY'S QI 678
THE FIRE PATH OF THE MICROCOSMIC ORBIT 678
THE WATER PATH OF THE MICROCOSMIC ORBIT 681
THE WIND PATH OF THE MICROCOSMIC ORBIT 681
MEDICAL QIGONG PRESCRIPTIONS 682
DESCEND THE QI AND CLEANSE THE ORGANS 682

GATHERING QI IN THE UPPER AND LOWER DANTIANS 684
ROLLING THE BALL WITH BOTH PALMS 686
OPENING AND CLOSING THE UPPER, MIDDLE, AND LOWER BURNERS 687
SUN AND MOON ROTATING TECHNIQUE 689
EXPANDING AND CONTRACTING THE RINGS 691
DESCENDING THE YANG AND ASCENDING THE YIN TECHNIQUE 691
TURNING AND WINDING THE BELT VESSEL TECHNIQUE 693
OPENING AND CLOSING THE HEAVEN AND EARTH TECHNIQUE 695
COLLECT THE QI AND RETURN TO THE LOWER DANTIAN TECHNIQUE 697
QIGONG SELF-HEALING MASSAGE AND POINT THERAPY 698
WASH THE FACE 698
MASSAGE THE HEAD 698
BEATING THE HEAVENLY DRUM 699
PRESSING THE EARS 699
RUBBING THE EYES 699
MASSAGING THE NOSE 700
CLICKING THE TEETH 700
WAGGING THE TONGUE 701
SWALLOWING THE SALIVA 701
WHIRLING THE NAPE AND TURNING THE NECK 701
MASSAGING THE CHEST AND ABDOMEN 702
MASSAGE TAPPING THE HANDS (LI-4) 702
MASSAGE TAPPING THE FOREARMS (PC-6) 702
MASSAGE TAPPING THE ELBOWS (LI-11) 702
MASSAGING UNDER THE ARMS (HT-1) 703
MASSAGE TAPPING THE UPPER SHOULDERS (GB-21) 703
MASSAGING THE ARMS 703
MASSAGE TAPPING THE LOWER DANTIAN (CV-6) 703
MASSAGE TAPPING THE BACK OF THE WAIST (GV-4) 704
MASSAGE TAPPING THE SIDES OF THE HIPS (GB-30) 704
MASSAGE TAPPING THE THIGHS (GB-31) 704
MASSAGE TAPPING THE OUTSIDE OF THE KNEES (GB-34) 704
MASSAGE TAPPING THE INSIDE OF THE KNEES (SP-9) 705
MASSAGE TAPPING THE BACK OF THE KNEES (BL-40) 705
MASSAGE TAPPING BELOW THE KNEES (ST-36) 705
MASSAGE TAPPING ABOVE THE MEDIAL MALLEOLUS (SP-6) 706
MASSAGE TAPPING THE HEELS (KD-1) 706
TREMBLING THE KNEES 706
CIRCLING THE ABDOMEN 706
COMPLETING THE EXERCISE 707

CHAPTER 42 DAOIST QIGONG REGULATION EXERCISES 709
DAOIST FIVE YANG ORGAN REGULATION EXERCISE 709
DAOIST FIVE YIN ORGAN TONIFICATION AND REGULATION EXERCISES 710

SECTION X

TREATMENT OF INTERNAL DISEASES WITH QI EMISSION THERAPY 723
INTRODUCTION 724

CHAPTER 43 QI EMISSION THERAPY FOR LIVER DISEASES 725
INTRODUCTION 725
LIVER COMPLICATIONS AND SYMPTOMS 726

GENERAL TREATMENT FOR LIVER DISEASES 726
MEDICAL QIGONG THERAPY FOR LIVER QI DYSFUNCTION 726
LIVER QIGONG PRESCRIPTIONS 728
MEDICAL QIGONG THERAPY FOR SPECIFIC LIVER DISEASES 730
DISORDERS OF THE BILIARY TRACT 730
GALLSTONES 731
CIRRHOSIS 732
CHOLECYSTITIS 733
HYPOCHONDRIAC PAIN 734

CHAPTER 44 QI EMISSION THERAPY FOR HEART DISEASES 735
INTRODUCTION 735
HEART COMPLICATIONS AND SYMPTOMS 735
GENERAL TREATMENT FOR HEART DISEASES 736
MEDICAL QIGONG THERAPY FOR HEART QI DYSFUNCTION 737
HEART QIGONG PRESCRIPTIONS 737
SPECIFIC MEDICAL QIGONG THERAPY FOR HEART DISEASES 738
HYPERTENSION 738
PALPITATION 742
RHEUMATIC HEART DISEASE 743

CHAPTER 45 QI EMISSION THERAPY FOR SPLEEN/STOMACH DISEASES 745
INTRODUCTION 745
SPLEEN COMPLICATIONS AND SYMPTOMS 745
GENERAL TREATMENT FOR SPLEEN DISEASES 746
MEDICAL QIGONG THERAPY FOR SPLEEN QI DYSFUNCTION 747
SPECIFIC MEDICAL QIGONG THERAPY FOR SPLEEN DISEASES 749
DIABETES 749
MEDICAL QIGONG THERAPY AND STOMACH DISEASES 750
STOMACH COMPLICATIONS AND SYMPTOMS 750
GENERAL TREATMENT FOR STOMACH DISEASES 751
SPECIFIC MEDICAL QIGONG THERAPY FOR STOMACH DISEASES 751
ABDOMINAL PAIN 751
CHRONIC GASTRITIS 754
GASTRIC AND DUODENAL ULCERS 755
CHRONIC CONSTIPATION 756

CHAPTER 46 QI EMISSION THERAPY AND LUNG DISEASES 759
INTRODUCTION 759
LUNG COMPLICATIONS AND SYMPTOMS 759
GENERAL TREATMENT FOR LUNG DISEASES 761
MEDICAL QIGONG THERAPY FOR LUNG QI DYSFUNCTION 761
LUNG QIGONG PRESCRIPTIONS 761
SPECIFIC MEDICAL QIGONG THERAPY FOR LUNG DISEASES 762
TREATING THE COMMON COLD AND INFLUENZA 762
CHRONIC BRONCHITIS 764
ASTHMA 764
BRONCHIECTASIS 766
SINUSITIS 767

CHAPTER 47 QI EMISSION THERAPY FOR KIDNEY DISEASES 769
INTRODUCTION 769
KIDNEY COMPLICATIONS AND SYMPTOMS 770
GENERAL TREATMENT FOR KIDNEY DISEASES 770
MEDICAL QIGONG THERAPY FOR KIDNEY QI DYSFUNCTION 771
KIDNEY QIGONG PRESCRIPTIONS 772
SPECIFIC MEDICAL QIGONG THERAPY FOR KIDNEY DISEASES 773
TINNITUS AND DEAFNESS 773
LUMBAGO 774
INSOMNIA 775
PYELONEPHRITIS 776
IMPOTENCE 777

CHAPTER 48 QI EMISSION THERAPY FOR MISCELLANEOUS DISEASES 781
INTRODUCTION 781
DEFICIENT DANTIAN QI 781
PRESCRIPTIONS FOR REGULATING THE THREE DANTIANS 781
IMPROPER YIN AND YANG ENERGY BALANCE 782
DISORDERS OF THE QI 783
REBELLIOUS QI DISORDERS 784
ARTHRITIS 784
VERTIGO 786
HEADACHES 787
MIGRAINE HEADACHES 789

SECTION XI
SPECIALIZED QI EMISSION THERAPY FOR PEDIATRICS, GERIATRICS, GYNECOLOGY, NEUROLOGY, PSYCHOLOGY, ONCOLOGY, AND SURGERY 793

INTRODUCTION 794

CHAPTER 49 QI EMISSION THERAPY FOR PEDIATRICS 795
INTRODUCTION 795
EARLY STAGES OF CHILD GROWTH AND DEVELOPMENT 795
PATHOLOGY 796
DEFICIENT LUNGS 796
DEFICIENT SPLEEN 796
DEFICIENT KIDNEYS 796
LIVER YANG EXCESS 796
HEART (SHEN) IMBALANCE 796
HOT AND COLD EXOGENOUS FACTORS 797
THERAPY 797
ENERGETIC MOVEMENT AND FLOW 797
MEDICAL QIGONG EXERCISE PRESCRIPTIONS FOR CHILDREN 797

CHAPTER 50 QI EMISSION THERAPY FOR GERIATRICS 801
TREATING THE PATIENT'S EMOTIONAL STABILITY 802
ENERGIZING THE PATIENT'S QI CIRCULATION 802
TAPPING THE KNEES TO PREVENT SENILITY 803
CULTIVATING THE PATIENT'S YUAN QI BY ABSORBING ENERGY FROM NATURE 803
FURTHER OBSERVATIONS AND INSIGHTS ON GERIATRIC PATIENTS 804

CHAPTER 51 QI EMISSION THERAPY FOR GYNECOLOGY 807
INTRODUCTION 807
MENSTRUAL FLOW 807
QI AND BLOOD REGULATION 807
ETIOLOGY AND PATHOLOGY OF GYNECOLOGICAL DISEASES 808
VAGINAL DISCHARGE 809
MENSTRUAL COMPLICATIONS 810
DYSMENORRHEA 810
MENOXENIA 812
CYSTS AND TUMORS 812
BREAST CYSTS AND BENIGN TUMORS 812
LEVEL-1 BREAST CYSTS (BENIGN LUMPS, NODULES OR FIBROADENOMA) 813
LEVEL-2 BREAST CYSTS (MASTITIS OR CARBUNCLE) 813
LEVEL-3 BREAST TUMORS 813
OVERVIEW OF TREATMENT 814
OVARIAN CYSTS AND UTERINE TUMORS 815
UTERINE COMPLICATIONS 816
MISCARRIAGE 816
UTERINE COMPLICATIONS DUE TO ABORTIONS 818
UNDERSTANDING ENERGETIC ABORTIONS 819

CHAPTER 52 QI EMISSION THERAPY FOR NEUROLOGY (STROKE, COMA, PARALYSIS, AND MULTIPLE SCLEROSIS) 821
INTRODUCTION 821
STROKE (WIND-STROKE) 821
COMA 823
FACIAL PARALYSIS (BELLS PALSY) 824
MULTIPLE-SCLEROSIS (MS) 825

CHAPTER 53 QI EMISSION THERAPY FOR PSYCHOLOGY 827
INTRODUCTION TO ENERGETIC PSYCHOLOGY 827
THE EMOTIONAL ENERGY STORED WITHIN THE FIVE YIN ORGANS 828
EXCESSIVE ANGER INJURES THE LIVER (WOOD) 829
EXCESSIVE JOY INJURES THE HEART (FIRE) 830
EXCESSIVE WORRY INJURES THE SPLEEN (EARTH) 831
EXCESSIVE SADNESS AND GRIEF INJURES THE LUNGS (METAL) 832
EXCESSIVE FEAR INJURES THE KIDNEYS (WATER) 834
FIVE YIN ORGAN EMOTIONAL RESPONSES 835
EMOTIONAL DISHARMONY 836
EMOTIONAL FREQUENCY RESONATION 838
DIFFERENTIATION OF PSYCHO-EMOTIONAL DISORDERS 839
SELF-TREATMENT PRESCRIPTIONS FOR PSYCHO-EMOTIONAL DISORDERS 840
TREATING EMOTIONAL ANXIETIES, PHOBIAS, AND ADDICTIONS 840
TREATING PHOBIAS AND DEPRESSION USING "THE WINDOWS OF THE SKY" POINTS 844
TREATING PSYCHO-EMOTIONAL DISORDERS RELATED TO STRESS 848
TREATING STAGNANT LUNG QI AND THE REPRESSION OF GRIEF 850

CHAPTER 54 QI EMISSION THERAPY FOR ONCOLOGY 853
HISTORY OF CANCER TREATMENTS 853
CATEGORIZATION OF TUMORS 853
THREE FORMATIONS OF CANCER GROWTH 854

ETIOLOGY, THE SIX PATHOGENIC EVILS THAT CAUSE CANCER 855
EMOTIONAL COMPONENTS OF CANCER FORMATION 856
CANCER FORMATION AND THE PATIENT'S MIND, EMOTIONS, AND SPIRIT 857
STRESSFUL LIFE ACTIVITIES AND THE IMMUNE SYSTEM 858
USING MEDICAL QIGONG TO STRENGTHEN THE IMMUNE SYSTEM 859
CANCER TREATMENT METHODS 860
SOCIAL ONCOLOGY—ESTABLISHING A CANCER SUPPORT GROUP 861
SOCIAL ONCOLOGY WITHIN THE HOSPITALS 861
CANCER RESEARCH 862
MEDITATIONS FOR CANCER PATIENTS 862
SUMMARY 864
TREATING TUMORS AND CANCER WITH MEDICAL QIGONG THERAPY 864
BRAIN TUMORS 864
BREAST CANCER 865
UTERINE CANCER 866
CERVICAL CANCER 867
PROSTATE CANCER 869
LUNG CANCER (PULMONARY CARCINOMA) 871
LIVER CANCER 871
ESOPHAGEAL CANCER 872
STOMACH CANCER 873
PANCREATIC CANCER 873
COLON AND RECTUM CANCER 874

CHAPTER 55 QI EMISSION THERAPY FOR RADIATION AND CHEMOTHERAPY 877
MEDICAL QIGONG THERAPY AND RADIATION THERAPY 877
MEDICAL QIGONG THERAPY AND CHEMOTHERAPY 877
RELIEVING SYMPTOMS ASSOCIATED WITH RADIATION AND CHEMOTHERAPY 878

CHAPTER 56 QI EMISSION THERAPY FOR SURGERY 879
MEDICAL QIGONG AND PREOPERATIVE THERAPY 879
USING MEDICAL QIGONG FOR ANESTHESIA 880
MEDICAL QIGONG THERAPY AND SURGERY 882
MEDICAL QIGONG AND POSTOPERATIVE THERAPY 882
MEDICAL QIGONG THERAPY AND RECREATING ORGAN ENERGY 883
MEDICAL QIGONG THERAPY AND ORGAN TRANSPLANTATION 885
FOLLOW-UP AND REMEDIAL PRESCRIPTIONS 885
PHANTOM PAINS OF AMPUTATED ORGANS 887
MEDICAL QIGONG THERAPY AND THE TREATMENT OF SCAR TISSUE 887

SECTION XII
MEDICAL QIGONG APPENDICES 889

INTRODUCTION TO THE APPENDICES 890

APPENDIX 1A MEDICAL APPLICATIONS OF QIGONG 891
ABSTRACT 891
INTRODUCTION TO MEDICAL QIGONG 891
EXAMPLES OF MEDICAL APPLICATIONS OF QIGONG 892
THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY 892
HYPERTENSION 893
PREVENTION OF STROKE AND MORTALITY 893

IMPROVEMENTS IN HEART FUNCTION AND MICROCIRCULATION 894
IMPROVEMENT IN SEX HORMONE LEVELS 895
INCREASED BONE DENSITY 895
CANCER 896
REVERSING SENILITY 897
MIND-BODY REGULATION 898
BRAIN WAVES 898
BLOOD FLOW TO THE BRAIN 898
EFFECT OF RAPID AND LARGE CHANGES IN ALTITUDE 898
CARDIAC FUNCTION 898
MICROCIRCULATION DISORDERS 899
STUDY OF LUNG FUNCTION 899
SUMMARY AND CONCLUSIONS 899
REFERENCES 900

APPENDIX 1B ANTI-AGING BENEFITS OF QIGONG 903
INTRODUCTION 903
CLINICAL STUDIES INDICATING THE ANTI-AGING BENEFITS OF QIGONG 903
THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY BY QIGONG 904
BLOOD PRESSURE 904
MORTALITY AND STROKE 904
IMPROVEMENTS IN HEART FUNCTION AND MIRCROCIRCULATION 905
ENHANCED ACTIVITY OF ANTI-AGING ENZYME SOD 905
CARDIOVASCULAR FUNCTION 905
COMBINATION OF MEDICAL QIGONG AND DRUGS THERAPY 906
CONCLUSIONS 906
REFERENCES 907
ENDNOTES 907

APPENDIX 1C THE EFFECT OF QIGONG ON THERAPEUTIC BALANCING MEASURED BY ELEXTROACUPUNCTURE ACCORDING TO VOLL (EAV): A PRELIMINARY STUDY 910
ABSTRACT 910
INTRODUCTION 910
MATERIALS AND METHODS 911
RESULTS AND DISCUSSION 911
CONCLUSIONS 913
ACKNOWLEDGEMENTS 913
REFERENCES 913

APPENDIX 2 SCIENTIFIC INVESTIGATIONS INTO CHINESE MEDICAL QIGONG THERAPY 917
INTRODUCTION 917
QIGONG MASTERS AND ENERGETIC MEDICINE 918
THE SCIENCE OF CHINA 918
QIGONG'S LONG HISTORY 919
SIMULATION OF EMITTED QI 920
SCIENTIFIC PLAUSIBILITY OF QI, THE HUMAN ENERGY FIELD 920
REFERENCES 921
THE FUTURE OF QIGONG 921
INFRASONIC QIGONG RESEARCH 922
INFRASONIC SIMULATION OF EMITTED QI FROM QIGONG MASTERS 922
SUMMAY OF RESEARCH 922
DOUBTS ABOUT EMITTED QI 922

BACKGROUND 923
THE BRAIN AS A DETECTOR OF EMITTED QI 923
SCIENTIFIC CONTROLS 924
INFRASONIC VS. ELECTROMAGNETIC INTERFERENCE 925
SUMMARY 925
SECONDARY ACOUSTIC BIOLOGICAL RESPONSE 925
MEASUREMENT AND ANALYSIS OF INFRASONIC WAVES FROM EMITTED QI 927
NEURO SCIENCE—NERVE IMPINGEMENT AND EMITTED QI 928
INFRASONIC TREATMENT OF ASTHMA 929
TREATMENT METHOD 930
OBSERVATION CRITERIA—FOUR GRADES 930
THERAPEUTIC RESULTS 930
LABORATORY RESULTS 930
THE INFRASONIC SIGNAL FROM THE QIGONG PERSPECTIVE 931
MECHANISMS OF THE QGM 931
RELIEVES PAIN AND EDEMA 931
STRENGTHENS IMMUNE FUNCTION AND REDUCES INFLAMMATION 931
NORMALIZES AUTONOMIC NERVOUS SYSTEM 931
RELIEVES SOME SYMPTOMS OF HEART DISEASE 932
EMITTED QI TRAINING INCREASES INFRASONIC SOUND EMISSION OF HEALERS 932
SCIENTIFIC QIGONG RESEARCH 933
EFFECTS OF THE EMITTED QI ON HEALING OF EXPERIMENTAL FRACTURE 934
MEASURING QI IN THE BLOOD 935
EFFECTS OF QIGONG ON NEUROTRANSMITTERS 936
PHYSICAL CHARACTERISTICS OF EMITTED QI 937
RESULTS 938
SUPERCONDUCTING MAGNETOMETER MEASURES CHANGES DURING QIGONG 938
QIGONG AND CANCER 938
THE EFFECT OF THE EMITTED QI ON CARCINOMA CELLS 939
THE EFFECT OF EMITTED QI ON THE GASTRIC ADENOCARCINOMA CELLS: 939
SUMMARY: 940
QIGONG AS A THERAPEUTIC AID FOR ADVANCED CANCER PATIENTS 940
THE EFFECT OF EMITTED QI ON LEUKEMIA IN MICE 941
EFFECTS OF QIGONG ON MALIGNANT TUMOR 942
EFFECTS OF PATIENT TUMORS ON KIRLIAN IMAGES OF HEALTHCARE PROVIDERS 943
CHAOS AND CONSCIOUSNESS 945
LOW DIMENSIONAL CHAOS IN PRACTICING QIGONG 945
QIGONG TRAINING AND COHERENCE OF EEG 946
VIBRATIONAL TREMBLING, CHAOS, AND EEG 946
LABORATORY TESTING 947
CHAOS AND THE COLLEGE ENTRANCE EXAM 947
SEARCHING FOR JING, QI, AND SHEN IN WESTERN SCIENCE 948
CONCLUSION 949
THE ROLE OF CHAOS IN DISEASE AND HEALING 950
THERAPEUTIC CHAOS 951
IMPLICATIONS OF SCIENTIFIC QIGONG RESEARCH 952

APPENDIX 3 THE PSYCHOLOGICAL AND ENERGETIC MECHANISMS TRIGGERED IN THE HUMAN SYSTEM BY THE SELF-APPLIED HEALTH ENHANCEMENT METHODS QIGONG, TAIJI, AND YOGA 957
    OXYGEN 957
    LYMPH SYSTEM FUNCTION 960
    BRAIN AND NERVOUS SYSTEM FUNCTION 964
    ENERGETIC MECHANISMS 967
    CONCLUSION 968
    REFERENCES 970

GLOSSARY OF TERMS 973

BIBLIOGRAPHY 1003

CLINICAL REFERENCES 1009

INDEX 1011

ABOUT THE AUTHOR 1029
    MEDICAL BACKGROUND 1029
    MARTIAL ARTS BACKGROUND 1031

ADDITIONAL BRANCHES OF THE INTERNATIONAL INSTITUTE OF MEDICAL QIGONG 1033
    BRANCHES WITHIN THE UNITED STATES 1033
    INTERNATIONAL BRANCHES 1035

FORWARD

Dr. Jerry Alan Johnson's book on Medical Qigong and its clinical practices provides an important and necessary guide for students and advanced practitioners of Medical Qigong and Traditional Chinese Medicine.

The relevance of Dr. Johnson's Medical Qigong textbook can be seen against the larger back drop of alternative medicine. Alternative Medicine (also called complementary medicine) has become a catch word in the West and an increasingly important factor in health care. An article authored by David Eisenberg and others of the Harvard Medical School reported that in 1993, 30% of a group of people used some kind of Alternative Medicine for which they paid out of their own pockets. Extrapolated to the United States population, this group spent about $14 billion dollars, which was more than the amount spent on conventional Western care that year.

Among alternative medical therapies, Medical Qigong has blossomed explosively in Western countries during the past few years. Awareness of Medical Qigong in the United States was heightened by two international conferences on Qigong in 1990. In 1997, there were three conferences on Medical Qigong, in San Francisco, Minneapolis, and New York. Another catalyst was Bill Moyer's television documentary, "Healing and the Mind," which had a segment on Medical Qigong and Chinese medicine. In more recent years, dozens of books have been published on Qigong in English and presently there are at least 15 web sites on the World Wide Web devoted to Qigong. At this writing, many articles on Qigong have appeared in the newspapers and magazines, and several documentaries have been videotaped and others are in progress.

In China, almost everyone knows about Qigong and about 100 million people are said to practice it daily. Qigong dates back thousands of years. Many believe that it was responsible for the early development of Chinese medicine, which now includes therapy by acupuncture, herbs, massage, and nutrition.

Medical Qigong therapy was a secret art that was passed on from a master to selected persons. In the 1970s, Qigong became more accessible to the people when the Chinese government ordered that Medical Qigong be made public and serve alongside Western medicine to improve national health care. During the Cultural Revolution, Medical Qigong was suppressed because the government suspected that Qigong had religious and superstitious roots incompatible with communism. In the early 1980s, scientist in China were encouraged to carry out experimental and clinical studies into almost every aspect of how Medical Qigong affected the human body and contributed to healing. Starting in 1986, research findings were reported at international conferences on Medical Qigong held mostly in China but also in Japan, Canada, and the United States. The substance of the research studies are available in English in the Computerized Qigong Database, which was developed by the Qigong Institute. This database has provided a basis for the preparation of at least five books, three dissertations, and several research projects on Medical Qigong.

In 1992, the United States Congress recognized the potential of alternative / complementary medicine by mandating that the National Institutes of Health develop programs to evaluate the validity of selected alternative therapies. Since that time, the Office of Alternative Medicine has funded many research programs and disseminated valuable information to medical practitioners, scientists, and the public.

The swell in the interest in Medical Qigong and the recognition of its health and healing benefits have been recognized by schools of Traditional Chinese Medicine, many of which mandate that Qigong be a required course. The Qigong taught at these schools is still on a basic level and needs to be expanded to include the medical aspects of Qigong, such as described in this book.

The future of Medical Qigong and Chinese medicine appears bright. However, the need for well-trained and qualified teachers and practitioners is paramount to meeting the growing demand. Dr. Johnson's book provides the first guide in English for teachers, students, and practitioners.

Kenneth M. Sancier, Ph.D.
President Qigong Institute

FORWARD

In 1998, the Academy of Oriental Medicine at Austin gave me the opportunity to establish one of the first formal Medical Qigong training programs in the country to be offered within the context of a college of Oriental medicine. My greatest difficulty was in finding a reliable textbook for my classes. Of the numerous books on Qigong available in English, most devote only a few pages, at best, to Medical Qigong. Over the past few years, several books from China on Medical Qigong have been translated into English. Although presenting much new and valuable information, they are poorly translated and fraught with errors, making them unreliable as textbooks.

Therefore, it has been a great pleasure for me to have collaborated with Dr. Jerry Alan Johnson to some small extent on this book. Jerry Alan Johnson is one of the world's foremost authorities on the subject of Medical Qigong. This work distills several decades of his personal study, practice, and clinical experience in the United States and the People's Republic of China into an indepth practical manual for all levels of Qigong healers.

Currently China is in the midst of what can only be called a Qigong renaissance. Today, more people are studying and practicing Qigong than in any previous time in Chinese history. In the not too distant past, Qigong was only available to monks and priests, nobles and the wealthy. Now Qigong is taught in China's primary schools and colleges, parks, temples, clinics, and hospitals to anyone who wants to learn.

China's literacy rate has increased exponentially since 1949, and now there are thousands of books, tapes, videos, and magazines on every aspect of Qigong, systems both ancient and modern, all available for general public consumption. Many regional, national, and international Medical Qigong associations have been formed, with an unprecedented exchange of information.

Although Medical Qigong has been a part of Traditional Chinese Medicine since the time of the Yellow Emperor, it has only been since 1982 that Medical Qigong hospitals and clinics have begun to spring up all over China. In the early days after the Communist Revolution of 1949, Qigong was more or less a state secret, and the Medical Qigong clinics were, for the most part, only available to high party officials. Over the past two decades, China has also devoted a great deal of effort to Qigong research and education. There are now large Medical Qigong hospitals in all of China's major cities, some of which have over 300 beds and 40 or more Qigong specialists and nurses on staff. There are hundreds of medium-and small-sized Medical Qigong clinics throughout the country, some treating inpatients and others treating only outpatients.

Before the revolution, the miracles of modern Western medicine, with its wonder drugs and surgical procedures, had caused many Chinese to see their own native traditional medicine as old-fashioned and outdated. The government encouraged the development of Western medical schools and hospitals. After the revolution, Chairman Mao recognized that with China's huge population they would need to take advantage of every form of medicine available even to begin to meet the nation's fundamental health care needs.

China is the only county in the world where Medical Qigong has been truly integrated and enthusiastically accepted as part of the national health care delivery system. According to current estimates, Medical Qigong doctors are now meeting 15% of China's national medical needs. That percentage is still growing. For decades, Chinese hospitals have successfully offered Traditional

Chinese Medicine alongside of Western medical services. Patients may receive Qigong or acupuncture anesthesia during Western surgical procedures and be treated by Medical Qigong, acupuncture, and Chinese herbal medicines during recovery. Chronic patients may be diagnosed by Western physicians, sent to Medical Qigong clinics for treatment, and have their Western doctors monitor their progress. In this respect, China has truly set the standard for the twenty-first century medicine.

Medical Qigong training and practice takes many forms in China. Some practitioners/healers are thoroughly trained in both Traditional Chinese Medicine and Western medicine, in addition to Medical Qigong. Others are heirs of family Medical Qigong systems, some dating back over a thousand years. Not all of these family practitioners have received additional training in either traditional Chinese medicine or Western medicine, yet they have been very efficient in treating disease and disorders that these conventional medicines have given up on. Still other practitioners come from martial arts backgrounds, such as Shaolin Gongfu or Daoist martial Qigong, with their own effective and unique blend of warrior training, spiritual discipline, Qigong and healing arts. Each of these categories contributes to the richness and diversity of the art of Medical Qigong.

Jerry Alan Johnson presents for us the art of Medical Qigong as it is being taught in two of the major Medical Qigong training centers in China today: the Xi Yuan Hospital of Traditional Chinese Medicine and its associated Hai Dian Medical Qigong College. Although each Qigong hospital in China defines its own program of training and practice, the material Johnson includes in this work represents the mainstream curriculum of study and practice widely accepted nationwide as a high standard of competence in Medical Qigong.

In addition to sharing the high-level protocols of Medical Qigong training in China, Johnson also draws from the reservoirs of his own decades of personal training and teaching in the internal martial arts to further enrich this work. Martial Qigong is generally practised to improve one's martial arts prowess. It includes Qigong exercises for strengthening the body; increasing power of one's strikes by emitting Qi; mind and will training; learning the vulnerable points to strike on an opponent's body to cause pain and paralysis; learning to cause serious injury or death by disrupting or destroying an opponent's internal organs, Qi and blood vessels, joints or bones; and toughening the body to withstand strikes and blows.

In what might at first seem a case of strange bedfellows, the martial Qigong tradition also has its medical side. Martial artists need to learn how to set bones and dislocations, as these injuries are commonly encountered in training and combat. Martial artists who train in the more secret arts of tien xue (vital point striking) and dim mak (death touch) are also taught how to heal the injuries that they have caused, both to heal themselves if struck by someone and to heal their opponents if the opponent is penitent. Furthermore, the martial artist who learns the methods for destroying the internal organs by emitting Qi can effectively adapt the same method to destroying pathogens and tumors in a patient for healing. Internationally recognized by his peers for his mastery of the most esoteric aspects of the internal martial arts, Johnson is eminently qualified to present its healing traditions, which he generously and openly does in this work.

I am grateful to know Dr. Jerry Alan Johnson as a colleague, teacher, and friend, and am happy to recommend this work to all students of Medical Qigong with confidence. Johnson's contribution contained herein represents the most thorough and comprehensive presentation of the art and science of Medical Qigong in the world to date. It is sure to become a modern standard of reference for anyone seeking knowledge of Medical Qigong.

Jampa Mackenzie Stewart, D.O.M.
Dean of Medical Qigong
Academy of Oriental Medicine at Austin

FORWARD

In the summer of 1995, I was sitting on the floor of Dr. Johnson's studio waiting for the start of the intermediate class in Medical Qigong. David Nakahara leaned over to introduce himself and asked me if I was a healer. Wow! It was one of those seminal moments in one's life. I considered all the implications of this question and the only relevant reply I, as a medical doctor, could muster was an ironic chuckle.

That question speaks to the heart of the current crisis in health care. Allopathic medicine lends itself to standardization of care and computer algorithms, since the structure of the clinical visit is based on protocol-derived diagnosis and treatment. The challenge is the diagnosis, but the treatment, especially high-tech treatment, gets all the press. It is incredibly satisfying to take care of people in this manner since the issues are clear and everyone gets better, unless they fail treatment; in this paradigm, the patient fails the treatment, the treatment does not fail the patient.

However, over time, you accumulate a number of patients for whom your therapeutic protocol fails, and I think that all thoughtful physicians recognize that it is these patients that teach us the most. There are women who have serious difficulty with PMS, infertility, and menopause. The cancer patients often wonder what caused the cancer to occur, what supports it, and what the meaning of this potentially fatal illness is for them. As physicians, once we emerge from the narcosis of fatigue and busyness that accompanies the daily practice of medicine in the late twentieth century, we also ask these same questions.

And we seek answers. Each physician develops his or her own rationale for treatment failures. They may rely on an unshakable faith in the scientific method, their religious upbringing or in their cultural ideals, or perhaps close themselves off entirely. I inadvertently found my answer within the martial arts.

It was clear to me from my Aikido practice that the superior martial artist must understand Qi in all its manifestations. This quest for deeper knowledge and experience of Qi led me to study Qigong with Dr. Johnson.

In my earliest lessons with Dr. Johnson, I was told that a Qigong therapist must observe and treat patients' *three treasures*: the physical body (jing or essence), their energy (Qi), and their mind/spirit/intention (shen). These three form the whole of a human being, and the parts can only be understood within the context of the whole. Furthermore, I learned that in life these three are so interrelated that it is difficult to tell where one leaves off and the other begins. Over 2,000 years ago, the Chinese philosopher Lie Zi described this by saying, "The mind leads and the Qi follows; the Qi leads and the body follows."

As a scientist, I could not help but draw the parallel between these fundamental Qigong principles and the revolutionary views of advanced modern physics: that at the deepest sub-atomic levels, matter, energy, and the person viewing them are so inextricably intertwined that they are seen as virtually one and the same, merely different manifestations of the same pattern. No longer is energy defined as a wave while matter is composed of particles. In quantum physics they have coined the term *wavicle* beause particles act as waves in one situation and as particles in another, as "probabilities" rather than concrete fixed entities and as subjective reality rather than objective.

The application of the discoveries of quantum physics to biology is just beginning. It is starting not only in the laboratory but also in many consulting rooms across the United States, as physicians search for the answers to help their patients.

I think that quantum physics is the only "hard" science that you can research just by thinking about it. As one observes the processes of nature, the thoughtful observer comes to the conclusion that there must be something else that accounts for the medical journey of the exceptional patient. The early practitioners of Ayurvedic and Chinese Medicine were careful observers of nature and through observation formulated the basic principles of Indian and Asian medical systems that are also corollaries of the principles of quantum physics.

Western awareness of the potential for the integration of the East and West began in 1971 when Nixon first went to China. James Reston, a *New York Times* correspondent covering the historic visit, developed acute appendicitis. After his emergency appendectomy, he was successfully treated with acupuncture for post-surgical pain.

Despite the dynamic nature of the penultimate Western nonbeliever confronting a totally novel experience ("only Nixon could go to China"), nothing much shifted or changed in medical practice at that time. What was missing was the readiness and ability to experience a paradigm shift in the conceptual basis of the causation of disease. Paradigm shifts in basic research have been eloquently discussed in Thomas Kuhn's landmark book, *The Structure of Scientific Revolutions*. Kuhn's brilliant analysis of the development of knowledge has recently been applied to medical science in a new book, *From Doctor to Healer: The Transformative Journey*, by Robbie Davis-Floyd and Gloria St. John. After 30 years, we are finally mature enough to begin to reassess the structure of our medical belief systems and to incorporate both world medicine and the principles of quantum physics. I have no doubt that this will take decades, but there is no doubt that the shift is beginning to occur.

Dr. Jerry Alan Johnson's book represents a major advance in this journey. This textbook is the first comprehensive treatise on the principles of Chinese energetic medical practice available to the Western medical practitioner. It also forms a bridge for the Western trained practitioner of Traditional Chinese Medicine and the allopathic (Western) practitioner. Dr. Johnson's book represents decades of research and study, and a nonjudgmental openness to any healing modality that is in the patient's best interest. I have no doubt that this book will represent a significant advance in that long awaited paradigm shift. I look forward to that time because it will address the deeper human values that have been neglected in protocol-based medicine and will bring a new depth of engagement and satisfaction to patient care.

Stephanie S. Taylor M.D., Ph.D.

Footnotes:

Kuhn, Thomas S., *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago, Illinois 1962.

Davis-Floyd and Gloria St. John, *From Doctor to Healer: The Transformative Journey*. Rutgers University Press, New Brunswick, N.J., available March 1998.

FORWARD

In the Western nations at the turn of the millennium, following a century of remarkable medical breakthroughs, there still remains a major list of unsolved medical mysteries and health care challenges. Chronic disease has not been alleviated (asthma, arthritis, etc.); stress has been recognized as a major cause in many diseases; tumors and cancers are being diagnosed in exponentially growing numbers; AIDS is a world wide epidemic; negative (Western) drug interactions have grown to become the fourth leading cause of death, and medical error has emerged as the eighth cause of death. According to the Department of Health and Human Services, over 70% of medical visits are for symptoms and illnesses that are preventable, yet the health care system is still primarily focused on treatment of the symptoms rather than health maintenance or prevention.

The public, the government, and health care providers all agree that there is a crisis in Western medicine and health care. Fortunately, at the same time, there is increasing support for innovative and alternative solutions. Among the many powerful strategies for solving this crisis, Chinese Medical Qigong (Chi Kung) and its offspring Taijiquan (Tai Chi Chuan), meditation, and focused breath practice, are probably the most comprehensive systems of self-care and healing available to compliment both conventional Western and alternative medical therapies.

The public hunger for natural healing, safety from medical side effects, lower medical costs, and health self-reliance is triggering powerful and profound changes in health care. It has become obvious that in the near future the benefits of Chinese medicine will be as fully integrated here in the U.S., as the benefits of Western medicine are in China. Through this global exchange of health care wisdom and practical applications, we are entering an unprecedented era of new possibilities to resolve suffering and foster well-being.

This definitive work on Medical Qigong, the remarkable healing art from Chinese traditional medicine, will play a significant role in the immense and profound transformation in health care that is now occurring with tremendous social and economic force.

Like the revolution in training and licensing acupuncturists that quickly unfolded in the final years of the 20th century, the momentum of training and licensing Medical Qigong Therapists is rapidly increasing in the beginning of the 21st century. Clinical Qigong, including both Medical Qigong Prescriptions and Medical Qigong Therapy, will arise in America as it has done so robustly in China.

There are a number of excellent books on various aspects and methods of Qigong. However, there has not been, in English, a comprehensive exploration of Medical Qigong. Dr. Johnson has created a breakthrough work on Medical Qigong, which is a clear and useful revelation of the Medical Qigong curriculum at the Hai Dian University Medical Qigong College of Beijing, China, and an excellent synthesis of Medical Qigong theory from throughout China. This textbook will very likely remain the definitive compendium of Medical Qigong in the West for many years, and become the foundation from which the field of Medical Qigong will evolve in Western society.

Roger Jahnke, O.M.D.
Chair, Department of Medical Qigong
Santa Barbara College of Oriental Medicine

PREFACE

As we enter the new millennium, a new era of medical therapy is beginning to blossom. Alternative medicine is now capturing an estimated $14 billion in out-of-pocket health care revenues from Americans each year, according to a 1993 study published in the *New England Journal of Medicine*. By 1997, according to the *Journal of American Medical Association*, that figure had more then doubled. What this trend means in real terms, is that despite the historical lack of official recognition by the American Medical Association, despite the lack of endorsement and coverage by Medicare and the majority of health insurance plans, men and women, in ever increasing numbers, are going to acupuncturists, energetic healers, herbalists, chiropractors, massage therapists, ayurvedic specialists, homeopathic doctors, and other traditional "healers" to meet some portion of their health care needs.

This growth of public reliance on alternative medicine has caused the Western medical establishment to sit up and take notice. In fact, some of the most vocal proponents of combining alternative medical traditions with Western medicine are medical doctors. Visionary physicians such as Deepak Chopra, Andrew Weil, Larry Dossey, Dean Ornish, and Bernie Seigel have led the way toward creating a new climate of respect for ancient medical philosophies and modalities. They have pioneered the advent in the West of health care facilities where Western medical and alternative health modalities are available under the same roof, with the goal of providing patients with the best of both worlds. This combination of ancient and modern medical traditions has been dubbed *integrative* or *complementary* medicine.

In comprehending the full implications of this unlikely marriage, one must understand the divergent approaches to healing between conventional Western medicine and traditional Eastern medicine.

Contemporary Western medicine grew out of the scientific revolution of the seventeenth century. The philosophy of science, rooted in Aristotle's "empirical materialism," was given a new spin by the French mathematician, Descartes. Viewing reality as that which could be substantiated materially, Descartes applied an analytical reductionist logic to penetrating the secrets of nature, including biology. These views were echoed in the physics of Sir Isaac Newton, applying a linear cause-and-effect model to explain the workings of a material universe.

Man was seen as being separate from nature, mind was seen as separate from body, and all of these processes, in nature and in humans, were seen as similar to the workings of a machine composed of discreet parts. Mechanical laws were seen to govern all processes. Structure determines function; therefore, the physician's role developed into that of a mechanic: repairing, removing, transplanting, and replacing broken down-parts. Diseases had isolated causes, which need to be removed from the rest of the parts. Because of this approach, Western medicine has the most highly developed pharmaceuticals to kill specific organisms and the finest surgical procedures in the world today.

By contrast, Eastern medicine grew out of the empirical observation of nature, beginning at least 4,700 years ago. Oriental philosophy, from the Vedas of India to the Yellow Emperor of China, views reality as an interdependent whole. This "prescientific" understanding equates to the broader view of modern quantum physics and the general systems theory. Rather than limiting reality to that which is material, the Eastern philosophers recognized the interdependence of mind and body, the nonlinear nature of time and space, and the interweaving patterns of relationship between man and nature; in fact, they believed, we *are* nature. Anatomy, the study of human structure, takes a back seat to physiology, the study of human function. Thus, the Orientals formulated a general systems theory, in which the patterns of change that exist in nature are the same patterns that govern human biology, wherein function is viewed from a holographic perspective, and each part reflects the whole. Rather than being fixed and stable, the whole is in a dynamic process of constant change. For the whole to function harmoniously, every part must remain in balance. Therefore, the role of the Oriental physician is more similar to a gardener, following the patterns of change, diagnosing functional disharmony and restoring overall balance. Because of this approach, Oriental medicine has some of the most highly developed procedures for preventative medicine and for treating chronic diseases in the world today.

Western medical science, with its fundamental distrust of subjective diagnostic reliability, has progressed toward developing more and more expensive high-tech laboratory tests and diagnostic equipment. Thus we have the modern miracles of x-rays, MRI's, and ultrasound. Eastern medicine, trusting in human capacity, has progressed in a low-tech direction toward ever deeper training of the physician's sensory and spiritual diagnostic tools. Thus, we have the miracles of pulse diagnosis, tongue reading, and Qigong hand scanning.

With such fundamental divergencies in philosophy and technique, it is almost inconceivable that these two medical systems could ever operate together in the same setting. The fact is that they do function together, and quite effectively, too. We can thank Mao Zedong for the union of these two unlikely bedfellows.

Recognizing that there were far too few Western trained physicians and nurses to meet the primary health care needs of China's vast population, from the outset of his leadership Mao advocated the systemization of Traditional Chinese Medicine (TCM), and advocated its implementation alongside Western medicine in China's hospitals and clinics. The results of this integration have been astonishing, as witnessed in the effective use of acupuncture anesthesia during surgery.

Nonetheless, in his efforts to create a "modern" Chinese medicine, Mao shunned some of the traditional theoretical aspects of Chinese medicine, such as the concept of Qi, which he considered as feudalistic and counterrevolutionary. For this reason, he actively discouraged Medical Qigong practice as superstitious. It was not until the end of the Cultural Revolution in 1975 that Qigong reclaimed its rightful place as one of the major branches of Chinese medicine.

A high party official in Beijing was suffering from an "incurable" disease. Both Western medicine and TCM had failed to alleviate his suffering. In desperation, he went to one of the few Medical Qigong clinics operating in the country and was cured. The official then lent his support to the promotion of Medical Qigong for the benefit of the Chinese people. Before long, there were hundreds of Medical Qigong hospitals and clinics throughout China.

Since that time, Qigong has experienced an unprecedented growth in China. Qigong is now available to the general populace for the first time in history. Qigong is now taught in the public education system, beginning at the elementary school level. Qigong departments have been added to large urban hospitals. Colleges of Traditional Chinese Medicine have developed Medical Qigong training programs. According to one recent survey, one third of the population of Beijing, China's capital city, practices Qigong daily.

It has taken acupuncture and Chinese herbology nearly 20 years to develop into a respected profession in the United States. As this book goes to press, 37 states now have legislation licensing professional acupuncture practice, with an additional 10 states in which legislation has currently been introduced. Medical Qigong is still at the very early stages of public recognition, understanding, and acceptance by comparison. Oriental medical schools around the United States, as well as independent Medical Qigong masters, are only now beginning to establish comprehensive

Medical Qigong training programs. I am heartened by this development.

Traditional Chinese medical therapy is divided into four main branches: acupuncture, herbs and diet, massage therapy, and Medical Qigong. It is important for TCM practitioners to have exposure to all four branches to be able to understand the relative strengths and limitations of their particular field of expertise, so that they will be able to select the most effective and appropriate treatment modality for their patients. Until very recently, most schools in America have been relatively unaware of the extent to which Medical Qigong therapy has developed in China, and have been lacking in presenting Medical Qigong as a significant part of their training programs for students of Oriental medicine.

The purpose of this book is twofold. The first purpose is to apprise the reader of the professional standards of knowledge and skill required in contemporary China for licensing a "Doctor of Medical Qigong Therapy" (D.M.Q.). The second purpose is to describe how to effectively diagnose and treat patients with Medical Qigong therapy as set forth by the Hai Dian Medical Qigong College of Beijing and to provide a textbook for training to that standard under the guidance of a qualified instructor. Hai Dian is recognized as one of the top Medical Qigong colleges in China.

Although this book is primarily written for students and practitioners of Oriental medicine, it is also my hope that practitioners and students of Western medicine and other forms of alternative healing will benefit from the information contained herein, and that it may serve to enrich their practice in the healing arts.

This book is divided into twelve sections. The first four sections focus on building a comprehensive foundation in energetic medicine. The second four sections focus on guidelines for establishing a Medical Qigong clinic, as well as for diagnosing and treating problems of Qi deviations. The last four sections focus on treatment modalities, Medical Qigong prescriptions, Medical Qigong regulation exercises and scientific research in Medical Qigong clinical therapy.

I have done my best to present the knowledge and wisdom of this ancient Chinese art as it was passed onto me personally by my respected teachers, to whom I owe undying gratitude. Qigong has survived nearly five thousand years of growth and refinement and is now available to you, the reader, for your own personal and professional benefit. I sincerely hope that this book may serve to further bridge the partnership between Eastern and Western medicine, that all humanity may be the beneficiary of the interchange between these two great schools of healing. May all doctors and healers, Western and alternative, return to the "heart" of medicine, that each may view the patient as a complete integration of body, mind, emotion, energy, and spirit. May we support each other's skills and techniques in alleviating our patients' suffering.

Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Dean of Medical Qigong Science
Director of Medical Qigong Clinic
Five Branches Institute, College & Clinic of TCM
Santa Cruz, California
May, 2000

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my many friends and colleagues who helped encourage and support me in writing this book: Roger Jahnke, O.M.D.; Hu Jiting, D.T.C.M.; Zhang Jingchuan, D.T.C.M.; Xu Zhuoqun, D.T.C.M.; Guo Xianhe, D.T.C.M.; Wan Taowang, D.M.Q.; William H. Lewington, L.Ac., M.Q.T.; Adam Atman, L.Ac. M.Q.T.; Tim M. Haun, D.C.; Kerrean D. Smith; Geoffrey Greenspahn M.Q.T.; Michael Winn; Mark Johnson; and my good friend Joseph Crandall.

I am also indebted to Dr. Pang Donghui, the Executive Deputy President of the Hai Dian Qigong College of Beijing, China, and to Li Fudong, D.M.Q. the university's chief director of Medical Qigong Science, and his assistants, Lu Guohong, D.M.Q. and Niu Yuhua, D.M.Q. for their constant support and encouragement, and for sharing their innermost secrets on Medical Qigong therapy.

I am very grateful to Dr. Teng Yingbo, D.M.Q., president and secretary general of the Beijing Western District Qigong Science and Research Institute, for his constant support and openness in sharing the knowledge of the Institute's clinical modalities.

A special thanks to Doctor Bi Yongsheng, D.M.Q.; Yu Wenping, D.M.Q.; and the Shandong Provincial Qigong Association for their wealth of knowledge and enthusiastic support, in sharing their clinical Qigong modalities, and to Dr. He Sihai, D.M.Q., of the Zhe Jiang Qigong Hospital for his openness and kindness in sharing his knowledge.

A special thanks to Qigong Master Zheng Zhanding, D.M.Q., for being my mentor, taking me under his wing, and teaching me advanced clinical modalities of energetic diagnosis and treatments, and to Qigong and Taiji Master Zhang Yufei for teaching, supporting, and encouraging my energetic and healing practice.

I am also indebted to Professor Lu and doctors Xu Hongtao, D.M.Q.; Ren Shuntu, D.M.Q.; Xu Zongwei, D.M.Q.; and the directors, teachers, interpreters and staff at the Xi Yuan Hospital of Traditional Chinese Medicine for their help and support.

I am very grateful to Sun Shuchun, D.T.C.M., the assistant professor and dean of the Beijing Academy of Acupuncture, Orthopedics, and Traumatology and the Hu Guo Si Hospital of Traditional Chinese Medicine for his assistance and support.

I would like to thank Doctor Meng Xiantong and Tara Peng, D.M.Q., of the Beijing Chengjian Integrated Traditional Chinese Medicine and Western Medical Experts Clinic for sharing their clinical modalities and their support.

I am very grateful to Madame Wang Yan of the China, Beijing International Acupuncture Training Center, the Acupuncture Institute of China Academy of Traditional Chinese Medicine, and the World Health Organization Collaborating Center for Traditional Medicine for supporting and believing in me.

I would also like to thank President Yue Licui, D.T.C.M., of the World Academic Society of Medical Qigong and all my friends and colleagues at the Beijing College of Traditional Chinese Medicine for their help and support.

A special thanks to Stephanie Taylor, M.D., Ph.D., whose encouragement and support has enabled Chinese Medical Qigong therapy to expand far into the terrain of conventional Western medicine, bridging the gap between these two schools with complementary modalities of healing.

A special thanks to Jason Gill, L.Ac., for his unselfish heart of giving and his invaluable help and technical input on Traditional Chinese Medicine, and his own personal experiences with Medical Qigong therapy.

A special thanks to Kenneth M. Sancier, Ph.D., and the Qigong Institute of Menlo Park, CA. Dr. Sancier's extensive research in Medical Qigong therapy, and the validity of its effect as a clinical modality is disclosed in Appendix 1 at the back of this text. Dr. Sancier is a council board member of the World Academic Society of Medical Qigong, sanctioned in Beijing, China.

A special thanks to Richard H. Lee. and the China Healthways Institute of San Clemente, CA. Mr. Lee's extensive research in the scientific investigations into Medical Qigong therapy and the validity of its effect as a clinical modality is disclosed in Appendix 2 at the back of this text. Mr. Lee is also a council board member of the World Academic Society of Medical Qigong, sanctioned in Beijing, China.

I greatly appreciate the assistance of Robert Collier, Jed Friedland, Stephanie Hensey, Terri Vandercook and Rose Mary Stewart, who spent many painstaking hours typing and deciphering my personal notes and course manuals. I would especially like to thank Jampa Mackenzie Stewart, D.O.M., L.Ac., Gaeir L. Dietrich, and Jean Vlamynck, M.Q.T., L.Ac., for their invaluable help and contribution in editing this textbook.

A special thanks to Madeleine H. Howell, M.F.T., as the senior editor of this manuscript for publication. Her keen eye for detail and constant encouragement to rewrite, explain and expound on the contents within this textbook, has resulted in a more comprehensive resource for future Medical Qigong students and Western medical professionals.

A special thanks to Robert Blaisdell and Tomo Y. for their excellent artwork and graphics, and to Y. S. Lim for his many painstaking hours of Translating Chinese Medical Qigong texts, which enabled me to draw from more comprehensive and clinical resources.

I am grateful to Willis Campbell and Niki Norrell, L.Ac., Dipl.Ac. (NCCA), M.T.C.M. for taking several of the photos.

A special thanks to Irene Morris for her excellent work on designing the front and back cover of this textbook.

Above all, I would like to express my love, gratitude, and respect to my parents Antoinette and Lt. Cdr. (Rt.) Perry E. J. Johnson for their years of love, understanding, encouragement, and support.

Finally, to my three beautiful daughters Laura, Leah, and Hannah, whom I will love forever.

INTRODUCTION

The Eastern approach in explaining medical concepts is generally expressed through a three dimensional convergence, seeing the physical body as an energetic hologram, and observing the physical, mental, emotional, energetic and spiritual reality of the tissues. By stimulating any one of these five energetic matrices, you affect the other four, and influence the body to initiate either health or disease. Diagnosis and treatment is therefore approached in a nonlinear progression, working towards the health and healing of the whole person.

The Western mind, however, is taught to view the physical body on a linear progression of cause and affect, separating the reactions of the tissues from the effect of the mind and emotions, as well as the universal and environmental influences. In order to help the Western mind understand the concept of Chinese energetic medicine, this Medical Qigong textbook was written with the goal of instructional comprehension and practical application of Chinese energetic medicine. It embraces the concepts of traditional Chinese medicine, in particular Medical Qigong therapy, from a progressive linear format. Beginning with Energetic Embryology in Section 1, it provides a basic understanding of complex energetic structure, theory, and practical application, ending in advanced clinical treatments in Section 11. The final chapters of this textbook, consists of scientific research collected from various doctors and research scientists which validate the claims made of Medical Qigong therapy being an effective clinical modality, in Section 12.

Chinese medical terminology is extremely metaphoric and is used to describe the many aspects of the human body through physical, mental, emotional, energetic, and spiritual domains of existence. Many technical terms in Traditional Chinese Medicine have numerous meanings, depending on the context of the subject.

For clarity, the Chinese terms in this book are capitalized along with the English words for which Traditional Chinese Medicine assigns a special meaning. Such words include, but are not limited to, the following: Gall Bladder, Small Intestine, Spleen, Pericardium, Bladder, Liver, etc. There is one exception to this rule; the word "energy" meaning Qi, is capped only in the first few chapters of this book for the sake of easier legibility, as it appears so often within each chapter.

When you see words such as Blood, Heart, and Marrow capitalized, assume that their meaning differs from that ascribed by Western medicine. In Traditional Chinese Medicine, the word Marrow, for instance, does not refer to bone marrow as it is traditionally recognized in the West, but rather describes the substance which is the common matrix of bones, bone marrow, the brain and spinal cord. Non-capitalized terms retain their traditional Western meanings.

One decision I have made when writing this textbook is to use the term "divine" when expressing the energy of the Dao or God. In China, many of the Qigong doctors and masters that I have trained with asked me to keep these spiritual theories in confidence for fear of governmental reprisal. Therefore, having decided to openly share these ancient energetic theories concerning the interactions of the spirit, soul and divine, I have done so without revealing my sources.

Finally, many of the energies used to describe the natural forces of nature (existing within Heaven, Earth, and Man) were historically changed into "gods" or "spirits" in order to compete with the colorful deities imported into China from India. Therefore, when reading this textbook, the reader should not become confused by certain terminologies describing "deities" which govern specific energetic principles. For example, when reading the section on the Eternal Soul, the title "Lords of the Three Dantians" implies the fact that the soul has three primary, separate, yet interconnected, energetic properties, rooted, divided and sustained within each of the Three Dantians.

Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)

SECTION I

FOUNDATIONS OF CHINESE ENERGETIC MEDICINE

CHAPTER 1
INTRODUCTION TO MEDICAL QIGONG

More than 5000 years ago, Chinese physicians came to understand that everything is composed of the same energetic substance called Qi (pronounced "chee"). These ancient masters concluded that there is a oneness and wholeness in all existence, and that energetically everything is interconnected as one body, although energy may appear to take on many different forms. All things in nature and, in fact, all things in the universe are woven together so that we are, quite literally, all symbiotically one with the universe through the system of Qi. Qi is always in motion within all things, and is the catalyst for everything to relate and interrelate within the universe.

In modern times, the laws of physics have demonstrated that matter and energy are interchangeable, and that matter is simply another form of energy. Matter is constantly vibrating in the form of tangible solids and intangible gases, and is constantly altering, being affected by, or interacting with energy. Energy is inherent in the living human body, and the human body is sustained by energy (Figure 1.1).

The ancients mastered techniques to balance the body's energy (Qi) in order to live in harmony with the environmental (Earthly) Qi as well as the universal (Heavenly) Qi. Traditional Chinese Medicine maintains that when living things start to lose their Qi, they lose their vitality. An ancient Chinese saying states, "Life comes into beginning because Qi is amassed; when Qi is scattered, the person dies."

Qi is stored within the body in the form of pools, creating the structures of the internal organs. From these internal pools, the body's life-force energy flows in the form of rivers and streams. These energetic rivers and streams form the body's vessels, channels, and collateral systems.

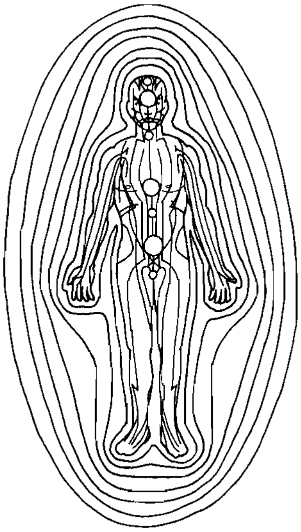

Figure 1.1. The Body's External Qi Field

THE FIVE DOMINIONS OF ENERGY

The ancient masters observed that Qi can be divided into five manifestations of matter and energy: mineral, plant, animal, human, and divine. Each form draws on the energy of the next, resonating and interacting with the divine through the form's relationship in Wuji (infinite space). The five manifestations of matter and energy are explained as follows.

1. The mineral's energetic field is considered the densest (i.e., the slowest) or lowest form of energetic vibration. The disintegration or division of the mineral's particles combine with the elements of air and water to form the Earth's soil. Every particle in the soil still retains the original primordial energy force of the mineral, which interacts with the energy of the divine.
2. The plant's energetic field is considered the next higher form of energetic vibration. All of

the Earth's vegetation (trees, bushes, flowers, herbs, etc.) absorbs a part of its life-energy from the mineral's energetic field, increasing and multiplying its energetic potential. The plant's energetic field is considered the next higher step in energetic evolution towards the divine energetic field.

3. The animal's energetic field is considered the next higher form of energetic vibration. The animal consumes and absorbs the energy from the plant's energetic field, further increasing and multiplying its energetic potential, bringing it one step closer towards the divine energetic field. Within each higher frequency of vibration there is also an increase in consciousness and level of awareness.
4. The human energetic field is considered the next highest form of energetic vibration. Mankind stands between Heaven and Earth, partaking of both energy fields. Through diet, Qigong practice, prayer, and meditation, humans can further refine and multiply their energetic potential.
5. The divine energetic field is the highest vibrational expression of energy known. As it envelops and becomes active within the human body, it further increases and multiplies the body's energetic potential, allowing man to attain divine consciousness.

All these energetic fields originate from one source, and all contain the vibrations of the one divine life-force. Likewise, with an attitude of deep respect for plants and animals that give up their life-force energy for our consumption, it is possible to enhance the nutritional value of the substances they provide us with. The blessing of food, and food prepared with a loving attitude, allows for the absorption of not only the vitamins and minerals contained therein, but also the absorption of the higher vibrations of the one divine energy inherent in all things. This is why many ancient cultures, often referred to as "primitive," prayed before hunting so that the animal spirit would willing give itself for sacrifice. Prayers were also given after the kill to free the animal's spirit so that it could return back to the divine.

Yang .	**. Yin**
active .	. passive
creative .	. receptive
masculine .	. feminine
front .	. back
left .	. right
fire .	. water
hot .	. cold
dry .	. wet
hard .	. soft
light .	. heavy
bright .	. dark
heaven .	. earth
sun .	. moon

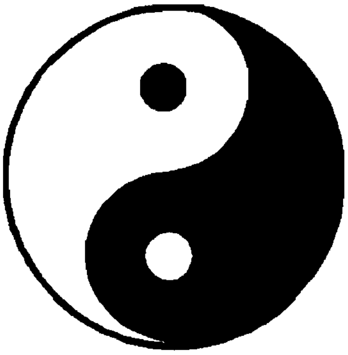

Figure 1.2. The table above shows some characteristics of Yang and Yin. Below is the Yin/Yang symbol: white represents Yang and black represents Yin. The small circles, one white and the other black, symbolize the fact that Yin is always transforming into Yang and Yang into Yin. (For more on Yin and Yang, see Chapter 3.)

Once individuals becomes aware of the divine energetic field, they begin to experience the refined vibrational energy fields of minerals, plants, animals and human beings. This increased awareness of the divine life-force energy strengthens the awareness of one's own energetic fields and that of others. This in turn can deepen the conscious and unconscious energetic connections between ourselves and others, be they human, animal, plant or mineral.

DEFINING THE ENERGY OF YIN AND YANG

Each of the five energetic fields can be further divided into Yin and Yang aspects. In Traditional Chinese Medicine (TCM), the theory of Yin and Yang energy represents the duality of balance and harmony within the body, as well as within the universe (Figure 1.2). Earth energy is Yin, while Heaven energy is Yang.

Yin exists within Yang, and Yang within Yin. Yang manifests as active, creative, masculine, hot, hard, light, and bright. Yin manifests as passive, receptive, feminine, cold, soft, and dark. The dynamic balance of Yin and Yang always changes and transforms the body's life-force energy. (See Chapter 3 for more on Yin and Yang energy.)

Successful practitioners in balancing the body's Yin-Yang energies were considered masters or "immortals," able to harmonize the body with the mind, the mind with the will, the will with the breath, the breath with the spirit, the spirit with motion, and finally, motion with the surrounding environment (Earth), the universe (Heaven), and the divine (Dao).

UNDERSTANDING THE CONCEPT OF QI

While the concept of Qi may seem complicated, it is actually very simple. Matter progresses to energy and energy to spirit. Qi is the medium, or bridge, between matter and spirit. Once we become aware of the reality of Qi, it becomes easily recognized.

Through observation and study, Chinese Qigong (pronounced chee-gung) masters discovered that each organ in the human body has a different function and a different speed of energetic vibration. By tracing the pathways (channels) Qi takes through each organ and observing the effects on bodily functions, the Chinese developed the basic theories upon which Qigong practice is founded. For thousands of years, Chinese medicine has successfully cured serious illnesses by stimulating the body's energy in very specific ways.

Through the study of Qigong, anyone wishing to cultivate awareness of the energy vibrations and their own individual pathways can learn to influence and even control them. Qigong practitioners use these skills to heal and strengthen the immune system, and to improve the functioning of various organ systems within the body. China Healthways International estimates that in Beijing alone more than 1.3 million people practice some form of Qigong every day, whereas, in China as a whole, around 80 million people practice Qigong.

DIFFERENT SCHOOLS OF QIGONG

Qi means "life-force energy" and gong means "skill," so Qigong is the skillful practice of gathering, circulating, and applying life-force energy. In China today, Qigong practice is divided into three main schools: medical, martial, and spiritual. The three schools are all based on the same philosophical system and share many of the same meditations and techniques. The schools differ primarily in focus. Students choose a school based on the use to which they want to put their Qigong training. Briefly, each school focuses on one of the following specialties:

1. The medical school trains doctors and healers in special Qigong methods for health maintenance and longevity, disease prevention, and the diagnosis and treatment of diseases and disorders. The three primary techniques of Medical Qigong therapy include the following.
    a. Purging to detoxify the body of pathogens,
    b. Tonifying to strengthen the body's internal organs and systems, and
    c. Regulating to balance the body's internal energy.
2. The martial school trains martial artists to

build their strength and power for performing martial arts applications. The three primary techniques of martial Qigong training include the following.

   a. Obvious Power (Ming Jing) techniques emphasize the training and conditioning of the muscles, strengthening the bone structure, and increasing the individual's overall stamina. This school also includes such techniques as pounding the body (arms, hands, legs, and torso) to strengthen and toughen the tissues.
   b. Hidden Power (An Jing) techniques emphasize stretching and twisting the tendons and ligaments (known as Reeling and Pulling the Silk) to cultivate resonant vibration within the body for striking and issuing power.
   c. Mysterious Power (Hua Jing) techniques emphasize the training and conditioning the mind's imagination and intention, to project and utilize the power of the individual's Shen (Spirit).

3. The spiritual school trains practitioners who seek spiritual transformation and enlightenment (Daoism, Buddhism, and Confucianism each have their own unique techniques). Their techniques include meditations for fusing, as well as releasing the Three Ethereal Souls (Hun). These souls can best be understood as personifications of moral qualities (or archetypes). When the Hun are fully developed, the practitioner acquires certain extraordinary powers and abilities, such as soul travel. The goal however, is to achieve transformation and a state of enlightenment, and not be led astray by the glamor of extra powers. The three primary techniques of spiritual Qigong training include the following:
   a. Nourishing the Spirit (Shen), to strengthen and refine the power of the individual's Shen,
   b. Housing the Shen by disciplining both thoughts and emotions, to relax and tranquilize the individual's Shen, and to become more receptive to divine energy and guidance, and
   c. Combining the Shen with the Qi, to coordinate the breath and intention for directing the spirit to guide the body's life-force energy.

Qigong training involves all of the individual's physical senses. The concentration is focused on breathing, hearing, visualizing, and muscle relaxation. Massage, and movement are also used to develop and control the body's intrinsic energy. Studying Qigong requires not only comprehending the immeasurable wisdom gathered for medical, martial, or spiritual development but also studying the ancient Chinese culture inherent within these systems.

MEDICAL QIGONG DEFINED

All living bodies generate an external field of energy called Wei Qi (pronounced "whey chee"), which translates as "protective energy." The definition of Wei Qi in Medical Qigong is slightly different than that of Traditional Chinese Medicine (TCM). In classical TCM texts, the Wei Qi field is seen to be limited to the surface of the body, circulating within the tendon and muscle tissues. In Medical Qigong, however, the Wei Qi field also includes the three external layers of the body's auric and subtle energy fields. This energy originates from each of the internal organs and radiates through the external tissues. There the Wei Qi forms an energy field that radiates from the entire physical body. This field of Qi protects the body from the invasion of external pathogens and communicates with, as well as interacts with, the surrounding universal and environmental energy fields.

Both internal and external pathogenic factors affect the structural formation of the Wei Qi. The internal factors include suppressed emotional influences (such as anger and grief from emotional traumas); The external factors include environmental influences when they are too severe or chronic, such as Cold, Damp, Heat, or Wind, etc. Physical traumas also affect the Wei Qi field.

Any negative interchange affects the Wei Qi by literally creating holes within the matrix of the individual's external energetic fields. When left unattended, these holes leave the body vulnerable to penetration, and disease begins to take root in the body. Strong emotions, in the form of toxic energy, become trapped within the body's tissues when we hold back or do not integrate our feelings. These unprocessed emotions block the natural flow of Qi, thus creating stagnant pools of toxic energy within the body.

Medical Qigong consists of specific techniques that use the knowledge of the body's internal and external energy fields to purge, tonify, and balance these energies. Medical Qigong therapy offers patients a safe and effective way to rid themselves of toxic pathogens and years of painful emotions that otherwise, can cause mental and physical illness. This therapy combines breathing techniques with movement, creative visualization, and spiritual intent to improve health, personal power, and control over one's own life.

MEDICAL QIGONG TRAINING IN CHINA

There are numerous colleges of Traditional Chinese Medicine throughout China today that focus on Medical Qigong training. The majority support the scientific study and expansion of Medical Qigong applications and Traditional Chinese Medicine treatments.

According to Qigong master and doctor of Traditional Chinese Medicine, Professor Zhou Qianchuan, all of the most famous Chinese doctors of acupuncture and moxibustion, herbal medicine, bone setting, and massage therapy, either practiced Qigong or incorporated Qigong into their clinical practices.

Major traditional Chinese medical colleges in China offer comprehensive, government-sponsored, three-year programs in Medical Qigong therapy. Programs include classes, labs, and seminars on traditional Chinese medical theory. These studies include: The foundations of Chinese medicine for internal diseases according to *the Yellow Emperor's Inner Canon, Spiritual Axis, Essential Questions,* and *the Canon of Perplexities*. The Medical Qigong classes also include energetic anatomy and physiology, diagnosis and symptomatology, energetic psychology, Qigong pathology, Medical Qigong therapy, as well as a survey of other related medical modalities. The other related modalities include: a comprehensive understanding of herbal medicine, acupuncture therapy, and Chinese massage. Classes of Western anatomy and physiology, Western internal diseases, and health and recovery, are also required.

During the certification program, three to five training hours a day accompany the standard six-day-a-week classroom curriculum. Course content, personal mastery of energy extension, and diagnosis techniques are rigorously tested each week. Upon completing the required courses and passing the final exams, the student receives a certificate of completion. Next, a six-month to one-year internship is required at a program-affiliated hospital or clinic. Upon successful completion of this internship, the new doctor is licensed as a doctor of Medical Qigong therapy by the People's Republic of China's Bureau of Scientific Technology.

Each internship program is assigned a separate wing in the selected Chinese hospitals. Both inpatient and outpatient facilities are available to the public. Each branch has specific approaches to healing a patient, with its own unique set of ground rules for diagnosis and treatment.

There are three distinct supervisory levels working within each clinical branch of Traditional Chinese Medicine in China. The first and lowest position is that of a "doctor of Medical Qigong," who is responsible for the treatment of all clinical patients (including patients in both the inpatient and outpatient clinics). The positions of Medical Qigong doctor are generally filled by the medical college graduates who have spent four to five years in clinical study and practice. The next level is called a "physician or doctor in-charge," and denotes a senior position within the clinic. This individual is responsible for the supervision of all the Qigong doctors' clinical procedures. This position is usually obtained after spending a minimum of five years as a Qigong doctor. The final and highest level is called a "director or professor;" this position requires overseeing the doctors in-charge, as well as teaching, treating, and training of other doctors to pass on Qigong clinical knowledge to future generations. This position is usually obtained after spending a minimum of five to six years as a doctor in-charge.

The licensing is reviewed and issued by either the People's Republic of China's Bureau of Scientific Technology (that issues a license in local city hospitals) or by the Ministry of Scientific Technology (that licenses to practice in any clinic or hospital throughout China). The Qigong doctor's skills are tested through oral, written and practical examinations, and a license is issued accordingly. In China today, there are five positions available for a doctor of Traditional Chinese Medicine. These five positions are described as follows.

1. A Doctor of Acupuncture Therapy (D.Ac.) specializes in the five main modalities of Chinese acupuncture.
2. A Doctor of Herbal Medicine (D.H.M.) specializes in the five main modalities of Chinese herbology.
3. A Doctor of Massage Therapy (D.M.T.) specializes in the five main modalities of Chinese massage and tissue regulation.
4. A Doctor of Medical Qigong Therapy (D.M.Q.) specializes in the five main modalities of Chinese Medical Qigong.
5. A Doctor of Traditional Chinese Medicine (D.T.C.M.) is a doctor who has trained in all four branches of Traditional Chinese Medicine (acupuncture, herbs, massage, and Medical Qigong).

MEDICAL QIGONG TRAINING IN THE UNITED STATES

The Five Branches Institute, College and Clinic of Traditional Chinese Medicine became the first TCM / Acupuncture College in the United States to open a Medical Qigong Clinic on April 7th., 2000. This clinic not only serves as an invaluable teaching resource for the Five Branches acupuncture students, but also serves as an affordable alternative medical clinic for the general public.

Founded in 1984, the Five Branches Institute has been known not only for its leadership in TCM education, but also as an academic innovator as well. It was the first, and remains the only, TCM College to establish a Neurology Center (founded in 1995) for the specialized treatment of paralyzed patients, under the directorship of the world-famous Chinese physician Doctor Ming Qing Zhu, L.Ac., Dipl. Ac. (NCCA), M.D. (China).

THE FIVE BRANCHES MEDICAL QIGONG CLINIC

The Five Branches Medical Qigong Clinic is currently divided into 3 treatment areas. The main area (Clinic A) is used to treat patients with less severe conditions; the second area (Clinic B) is used specifically for the treatment of cancer patients. The third treatment area is designed for video taping the patient's Medical Qigong prescription exercises and meditations.

Clinic A is the primary care facility of all patients. The patients are first assessed, and then assigned to a treatment team. Each team is comprised of at least one senior intern and several clinical students. Each team diagnoses, treats, and then prescribes Medical Qigong exercises and meditations for each patient. Some patients with potentially terminal illnesses (cancer, tumors etc.) may require specialized treatments, and are referred to the Clinic B.

The main goal of the Clinic B is to specialize in Oncology, and to work closely with the Academic Dean and Clinical Medical Director Dr. Joanna Zhao, L.Ac., Dipl. Ac. (NCCA), D.T.C.M. (China), who prescribes herbal therapy to complement the Qi Emission therapy and Medical Qigong prescriptions.

The Director of the Medical Qigong Clinic is Dr. Jerry Alan Johnson, whose duty is to oversee the "Doctors in Charge" of the 3 treatment areas, as well as regulate the patient load between Clinics A and B.

The Doctors in Charge of the Medical Qigong Clinics are: Jean Ruth Vlamynck, L.Ac., Dipl.Ac., M.T.C.M., M.Q.T., and Adam Atman, L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. Their duty is to oversee the Medical Qigong doctors on staff within the three treatment areas, supervise the current interns, and make themselves available for consultations when needed.

The front entrance of the Medical Qigong College at the Hai Dian University in Beijing, China.

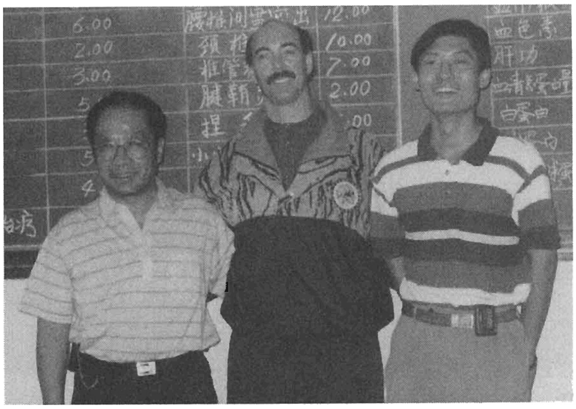

Left to right: Dr. Li Fu Dong, Associate Professor and Chief Director of Medical Qigong Science, Dr. Jerry Alan Johnson the author and director of the International Institute of Medical Qigong, and Dr. Pang Dong Hui, Executive Deputy President of the Hai Dian University.

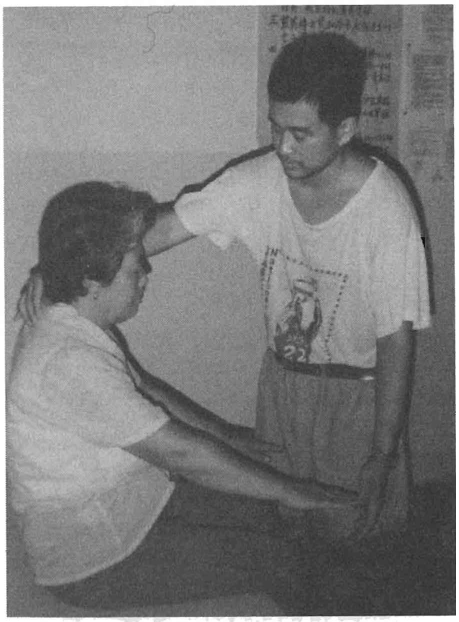

Dr. Lu Guo Hong, Director of the Hai Dian Medical Qigong College, treats a patient suffering from hypertension by dispersing Excess Heat from her body.

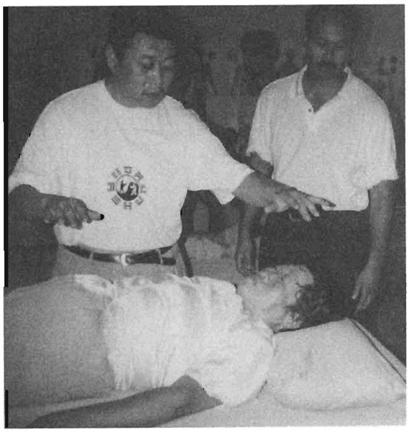

Dr. Niu Yu Hua, Assistant Director of the Hai Dian Medical Qigong College, treats a patient by regulating the Qi of her Upper and Lower Dantian.

Right: The graduate students of the Hai Dian Medical Qigong College intern at the Xi Yuan Hospital in Beijing, China, for a period of 6 months to a year. They are licensed through the government's Medical Qigong Science and Research Institutes, which monitor the student programs.

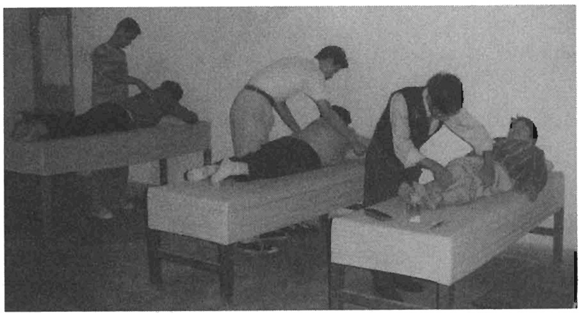

The Xi Yuan Hospital in Beijing, China, is the equivalent of the Stanford or Harvard Medical Center in the United States; it specializes in five distinct branches of healing therapies: Medical Qigong, acupuncture, herbs, Chinese massage, and Western medicine.

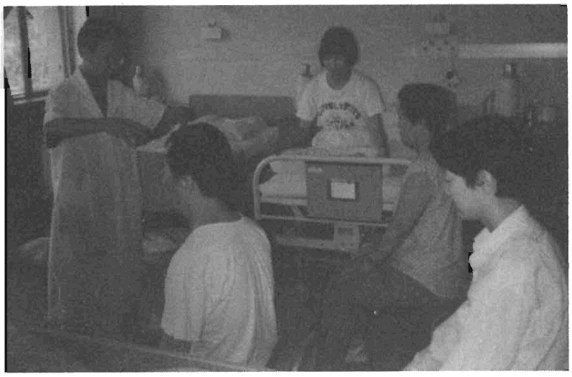

The Xi Yuan Hospital is equipped with both inpatient and outpatient Qigong clinics. Here, Dr. Xu Hongtao administers "distance" Qigong therapy to patients in one of the many Medical Qigong wards.

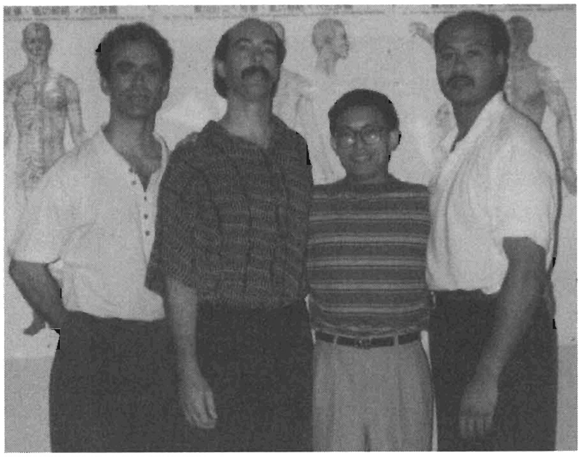

Three Americans completed their internships at the Xi Yuan Clinic in 1995. They are shown here posing with Dr. Xu Hongtao, the Xi Yuan Hospital staff supervising physician. *Left to right:* Dr. Seth Lefkowitz, the author Dr. Johnson, Dr. Xu Hongtao, and Dr. Arnold Tayam.

Left to right: Qigong Master Zheng Zhanding, Jonathan Liu, and the author discussing advanced Medical Qigong theories and modalities.

Right: The China Beijing International Acupuncture Training Center is a facility in which doctors from around the world culminate their training in Chinese medicine with Chinese doctors who are experts in their medical specialties.

CHINA BEIJING INTERNATIO
ACUPUNCTURE TRAINING

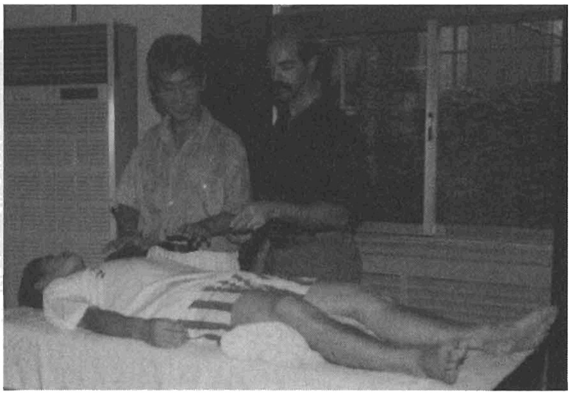

Dr. Teng Yingbo (*left*), the president and secretary general of the Beijing Western District Qigong Science and Research Institute, with the author treating a patient.

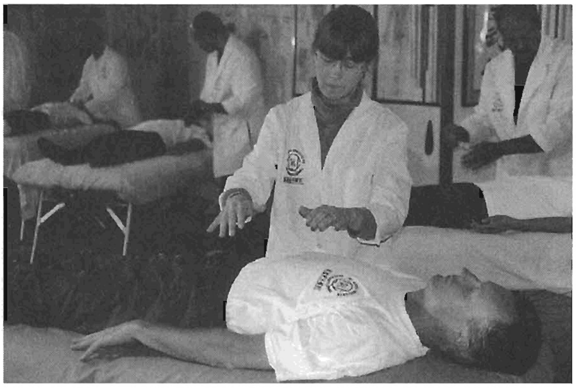

Left: The International Institute of Medical Qigong is one of the few American Medical Qigong colleges that maintains the same strict standards as the Medical Qigong Universities in China. The initial training programs generally require three to four years of training.

The 1999 graduating class of the International Institute of Medical Qigong (*left to right*):

Back row: Matthew B. Weston, William H. Lewington, William H. De Groat, Paul E. Miller.

Middle row: Katy Reed, Madeleine H. Howell, Brooks M. Fiske, Diane de Terra.

Front row: Michael J. Finch, Jean R. Vlamynck, Dr. Johnson, Dennis M. Earnest.

Not pictured: Todd Mathew Gedryn, Luc Arnauld Logan, Geoffrey Greenspahn and Rose Mary Stewart.

WORLD CONFERENCE ON MEDICAL
WORLD ACADEMIC SOCIETY OF MEDICAL QIGONG

The author lecturing at the World Academic Society of Medical Qigong 1996 World Conference, on the psychophysical manifestations associated with tumor formation.

The 1997 graduating class of the International Institute of Medical Qigong (*left to right*):

Back row: Dr. Stephanie Taylor, Pamela Lee Espinoza, Anne Elderfield, and Dr. Carole Marie Kelly.

Front row: Dr. Arnold E. Tayam, Dr. Johnson (director of the institute), Shannon K. Brown, and Dr. Seth Lefkowitz.

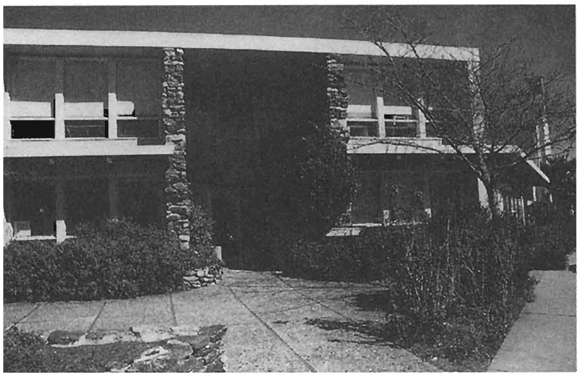

The Five Branches Institute, College and Clinic of Traditional Chinese Medicine became the first TCM/ Acupuncture College in the United States to open a Medical Qigong Clinic on April 7th., 2000.

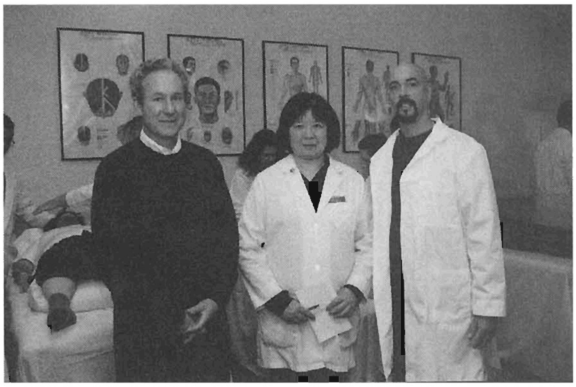

Pictured from left to right are the Five Branches President and CEO Ron Zaidman, M.B.A., M.T.C.M.; the Academic Dean and Clinical Medical Director Dr. Joanna Zhao L.Ac., Dipl. Ac (NCCA), D.T.C.M. (China); and the Dean of Medical Qigong Science and Director of the Medical Qigong Clinic Dr. Jerry Alan Johnson.

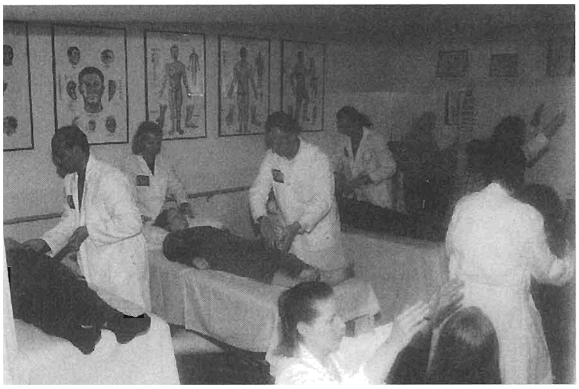

In the main treatment area (Clinic A), patients are categorized according to the severity of their condition. For serious conditions (e.g., multiple sclerosis) patients are encouraged to lie on the treatment tables.

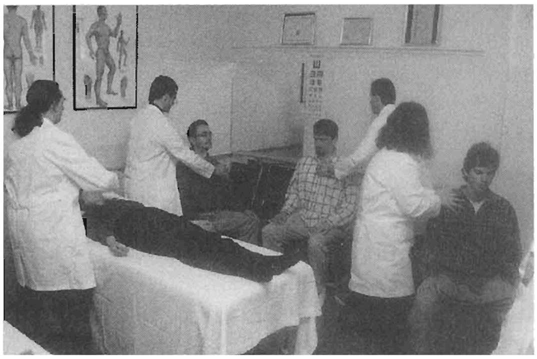

In Clinic A of the Five Branches Medical Qigong Clinic, patients with less serious conditions (e.g., hypertension and migraine headaches) are encouraged to relax on one of the treatment chairs.

The resident Doctors in Charge of the Five Branches Medical Qigong Clinic are Adam Atman L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. (positioned on the left), and Jean Ruth Vlamynck L.Ac., Dipl.Ac., M.T.C.M., M.Q.T. (located in the middle). They are seen here is discussing the current patient load with Dr. Arnold Tayam D.M.Q. (China).

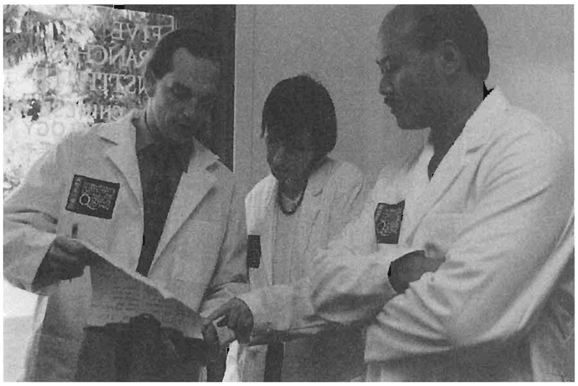

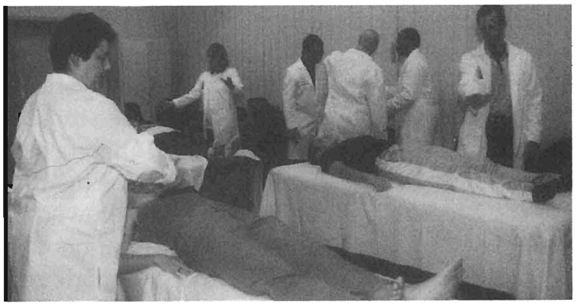

The second treatment area (Clinic B), is used specifically for the treatment of tumor and cancer patients.

The Five Branches Medical Qigong students are also taught Energetic Psychology and how to deal with the emotional discharges coming from their patients.

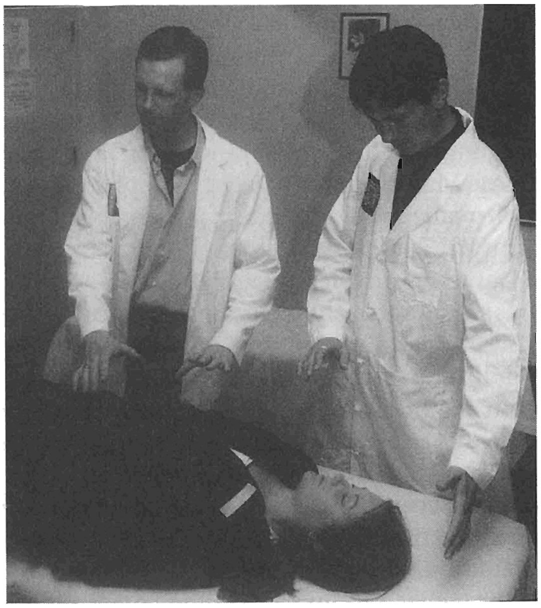

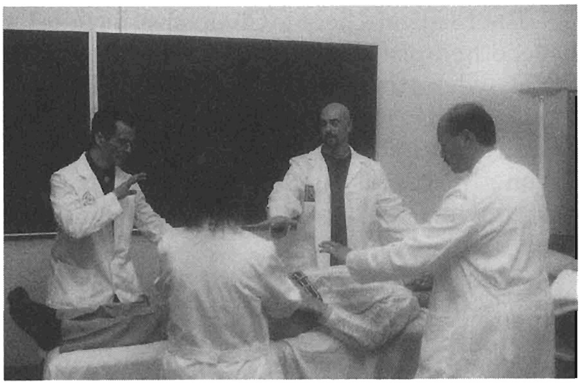

Doctors of the Five Branches Medical Qigong Clinic combine their energies to treat a specific type of cancer, observed here, in Clinic B.

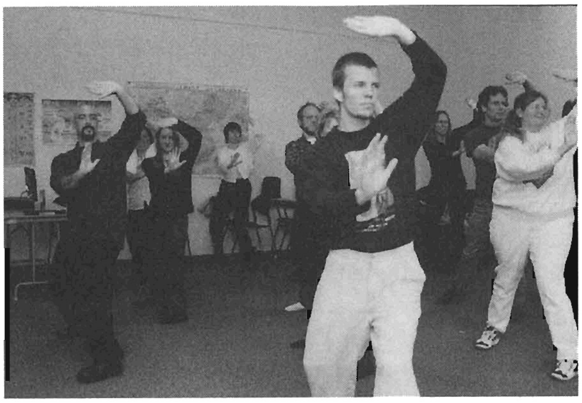

Within the first semester, the Medical Qigong students at the Five Branches TCM College have learned and experienced basic applications of Medical Qigong purgation, tonification and regulation exercises. They have also been taught the underlying principles for each exercise and meditation. Furthermore, they are able to discriminate when to assign these prescriptions to patients, and when not to.

In the first phase of internship, the Medical Qigong students work as part of a Qigong treatment team, assisting a senior intern.

The Medical Qigong Clinic also includes a psychotherapist, Madeleine Howell, L.M.F.T., M.Q.T., who is trained in Medical Qigong therapy. She is available to patients who may need professional consultation after undergoing Medical Qigong treatments, as these may bring painful emotions to the surface as part of the healing process.

On the first official day of the clinic, 15 Qigong interns and doctors from the International Institute of Medical Qigong performed Qi Emission through Distance therapy on more than 80 patients. Treatments ranged from a general treatment (purging excess, tonifying deficiencies and dispersing stagnations) to the treatment of several potentially terminal cancer patients.

THE FIVE BRANCHES MEDICAL QIGONG COURSE AND CLASSES

The Medical Qigong students at the Five Branches college learn and experience basic applications of Medical Qigong purgation, tonification and regulation exercises during the first semester. Classes, include lectures and labs. They are also taught to understand the underlying principles of each Medical Qigong exercise and meditation, and learn when and when not to assign these exercises in cases for clinical application. The students are also taught Energetic Psychology and how to deal with the emotional discharges coming from their patients.

In the second semester the Medical Qigong students are led through a progressive series of advanced Shengong meditations and exercises, for developing advanced intuitive diagnostic skills. The students also learn basic Medical Qigong treatment protocols, and begin the initial phase of clinical internship. During this phase they assist the senior interns.

From the third semester on, the students begin an in depth study in Medical Qigong theory and its application. The goal is to complete a three year certification program in Medical Qigong Therapy.

MEDICAL QIGONG CLINICAL THEATRE

Conducted in a small class setting, the Medical Qigong Clinic gives the students exposure and first hand experience at Medical Qigong diagnoses and treatments. This is a hands on experience where the different aspects of energetic dysfunctions and symptom manifestations are studied, treated and observed by the Medical Qigong students. Through the careful guidance of the instructor, the Medical Qigong students are allowed to assist in all phases of diagnosis and treatment.

After completing the Clinical practicum, Medical Qigong students at the Five Branches TCM College will be qualified to diagnose and treat patients, as well as document and maintain accurate clinical records.

CLINICAL INTERNSHIP

In the first phase of internship, the Medical Qigong students work as part of a Qigong treatment team, assisting a senior intern.

In the second phase of internship, the Medical Qigong students work with their own patients, under the supervision of the instructor.

By the end of their Clinical Internship, the Medical Qigong therapists will receive their Masters degree, and will be fully qualified to diagnose and treat patients without supervision.

TRADITIONAL CHINESE MEDICINE AND MEDICAL QIGONG THERAPY

Medical Qigong therapy is the oldest of the four branches of Traditional Chinese Medicine and provides the energetic foundation from which acupuncture, herbal healing, and Chinese massage originated. It is through the understanding of Qigong that the other branches of Traditional Chinese Medicine are elevated to a spiritual path of self-realization and internal transformation. Doctors of Traditional Chinese Medicine address the patients' physical, energetic, and spiritual needs simultaneously. According to the principles of Traditional Chinese Medicine, the root cause of all disease can be traced to a critical imbalance within the body's vital energies. Therefore, the best way to prevent or cure disease requires establishing a healthy energetic balance and harmony between the body's energy field and the forces of nature and the cosmos.

Traditional Chinese Medicine is divided into four branches of clinical healing: Acupuncture

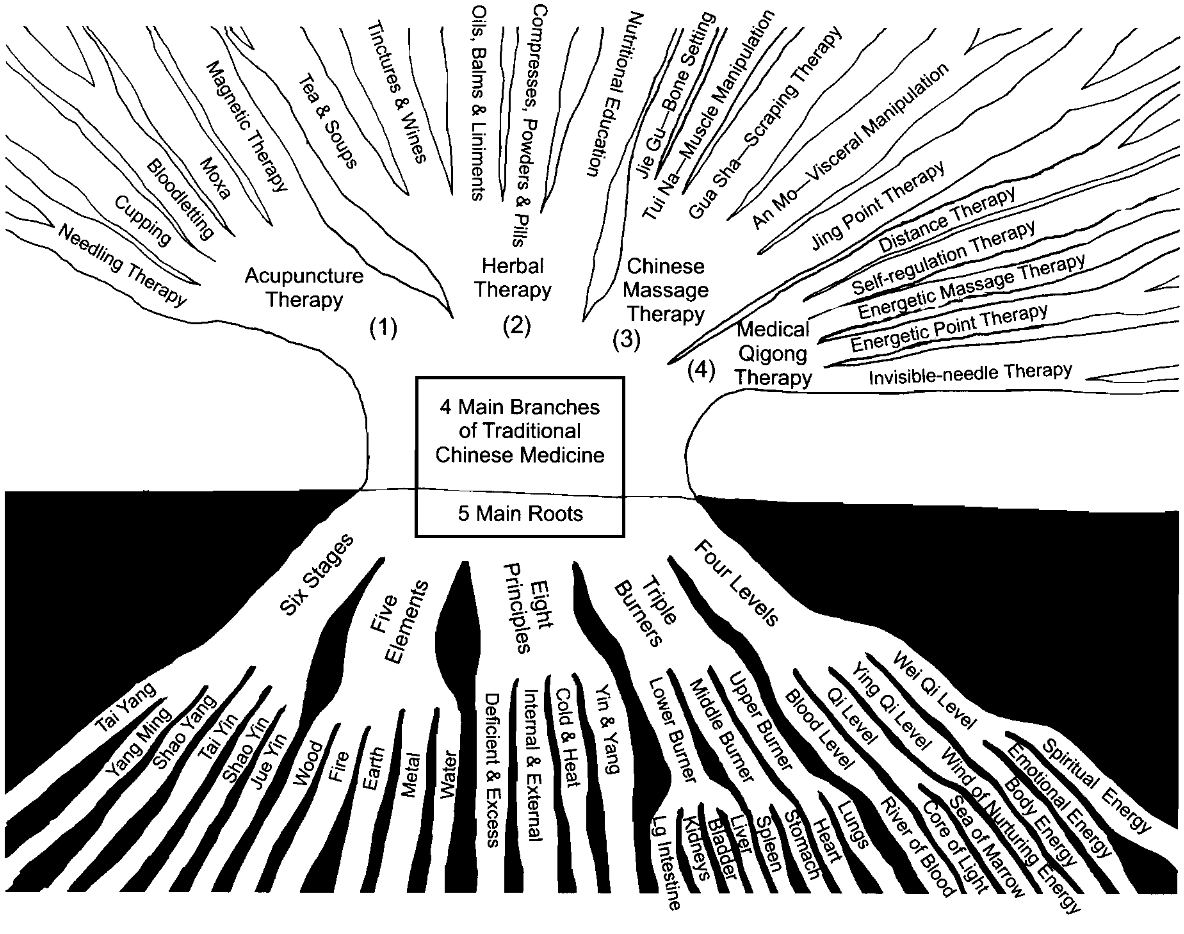
Needling Therapy
Cupping
Bloodletting
Moxa
Magnetic Therapy
Tea & Soups
Tinctures & Wines
Oils, Balms & Liniments
Compresses, Powders & Pills
Nutritional Education
Jie Gu—Bone Setting
Tui Na—Muscle Manipulation
Gua Sha—Scraping Therapy
An Mo—Visceral Manipulation
Jing Point Therapy
Distance Therapy
Self-regulation Therapy
Energetic Massage Therapy
Energetic Point Therapy
Invisible-needle Therapy
Acupuncture Therapy
(1)
Herbal Therapy
(2)
Chinese Massage Therapy
(3)
Medical Qigong Therapy
(4)
4 Main Branches of Traditional Chinese Medicine
5 Main Roots
Six Stages
Five Elements
Eight Principles
Triple Burners
Four Levels
Tai Yang
Yang Ming
Shao Yang
Tai Yin
Shao Yin
Jue Yin
Wood
Fire
Earth
Metal
Water
Deficient & Excess
Internal & External
Cold & Heat
Yin & Yang
Lower Burner
Middle Burner
Upper Burner
Lg Intestine
Kidneys
Bladder
Liver
Spleen
Stomach
Heart
Lungs
Blood Level
Qi Level
Ying Qi Level
Wei Qi Level
River of Blood
Core of Light
Sea of Marrow
Wind of Nurturing Energy
Body Energy
Emotional Energy
Spiritual Energy

Figure 1.3. The Four Main Branches of Traditional Chinese Medicine

therapy, Herbal therapy, Chinese Massage therapy, and Medical Qigong therapy. All of the four main branches are built on the same foundation of energetic diagnosis known as the Five Main Roots of Traditional Chinese Medicine. The five main roots are used for internal organ diagnosis according to the Six Stages, Five Elements, Eight Principles, Triple Burners and Four Levels theories (Figure 1.3).

ACUPUNCTURE THERAPY

By inserting very thin metal needles into specific points, the acupuncturist manipulates the patient's energy (Qi) to achieve an overall physical balance. In an acute or emergency situation, a patient's symptoms are treated first, after which the focus is directed to rerouting Qi Deviations or removing the original cause of the symptoms (e.g., stagnations).

The energetic points are specific areas both on and in the body where Qi emerges from deep within the body's organs and tissues, or submerges to travel deep within the body. The Qi moves along specific pathways known as channels and collaterals. Collaterals are the smaller streams of Qi that branch off from the main ener-

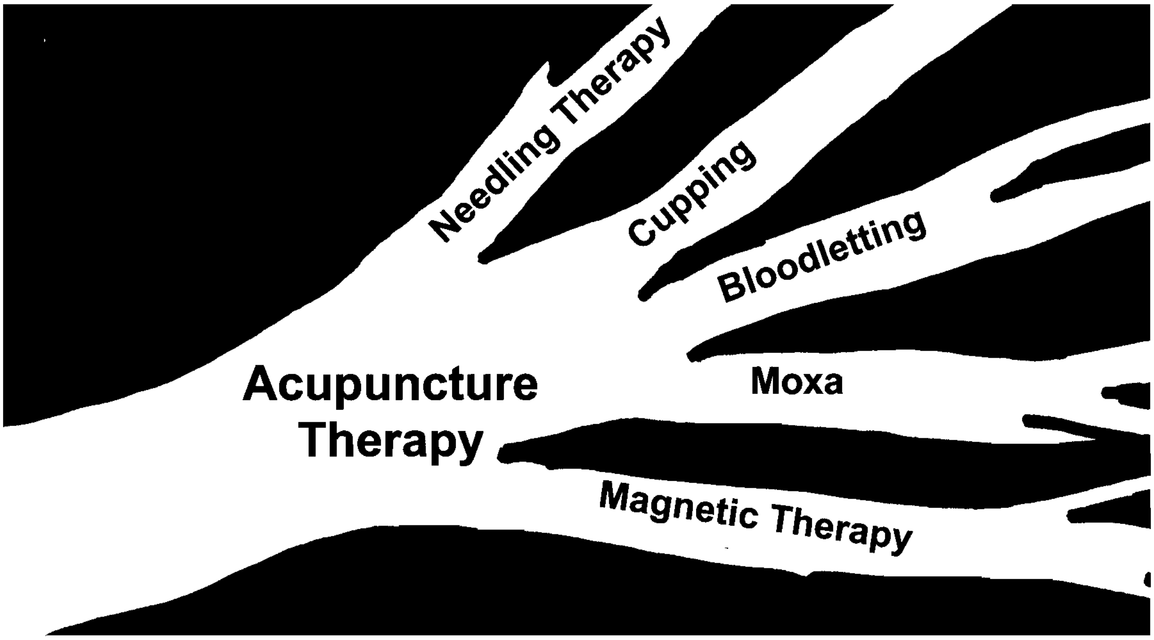

Figure 1.4. The Five Main Branches of Chinese Acupuncture Therapy

getic rivers (channels). Each Primary Channel takes its name from one of the six Yin or six Yang organs to which it corresponds. Each organ has two channels, one on each side of the body.

Often in Chinese medical diagrams and procedures, you will see abbreviations referring to specific energetic points. Traditional Chinese Medicine names the points based upon the channel on which the point lies and its distance along the channel. The first points on each side of the Gall Bladder Channels (GB), for example, are GB-1. Some channels contain over 60 points, each numbered sequentially from beginning to end. These points are also given descriptive names according to their location and energetic affect upon the body when treated. The GB-1 points, for example, are located next to the eyes, on the outer canthus, level with the pupils, and are called the Pupil's Seam.

Acupuncture therapy includes five major treatment techniques (Figure 1.4): needling, cupping, bloodletting, moxa burning, and magnet healing.

1. Needling uses acupuncture needles of various sizes that are inserted into channel points. These points are tiny areas where the Qi pools along the streams of an energy channel (Figure 1.5). The needles stimulate the nerves and energy flow to tonify or reduce Excess Qi.
2. Cupping uses wooden, clay, or glass cups that adhere to the patient's skin by suction (Figure 1.6). This suction drains, or removes, pathogenic Qi from the body's pores. This technique can also be used to tonify specific areas of the body. This modality of treatment has been successfully combined with bloodletting to treat acute sprains accompanied by Blood stagnation.
3. Bloodletting is done with instruments such as blood needles or seven- and five-star hammers to remove Toxic Qi, Blood stagnation, Heat, and other pathogenic factors (Figure 1.7). The hammer has five to seven sharp projections that pierce the skin and cause slight bleeding. The acupuncturist diagnoses and then monitors the patient's condition by the

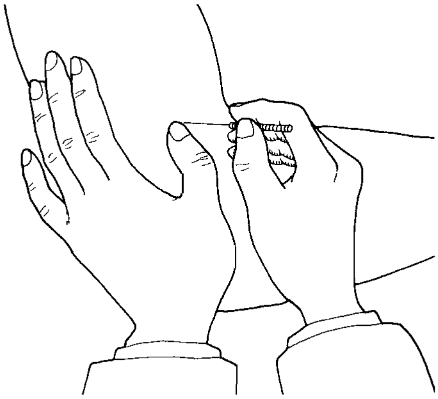

Figure 1.5. A doctor is inserting an acupuncture needle into the patient's Channel Point. Acupuncture needles of various sizes are used to stimulate nerve and energy flow for the tonification of Deficient Qi or the reduction of Excess Qi.

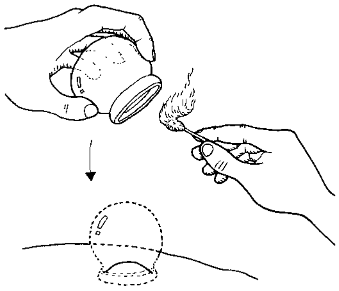

Figure 1.6. In Cupping, heated air is directed into wooden, clay, or glass cups, which are then placed on the patient's skin. The cooling air creates suction that drains, or removes, pathogenic Qi from the body's pores.

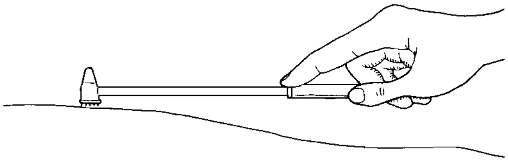

Figure 1.7. The doctor is lightly tapping with the Five-Star Hammer that has five sharp projections to pierce the skin and cause slight bleeding. The color of the patient's Blood provides the doctor with information about the patient's condition.

different shades of the patient's Blood. Trapped or diseased Blood is released until a healthy color is observed. This therapy is considered useful for treating disorders of the nervous system, physical trauma, and extremely serious febrile diseases.

4. Moxa Burning employs lighted herbal cones or sticks that are inserted on top of acupuncture needles (or held over specific channel points) to infuse heat and Qi into specific body areas for tonification (Figure 1.8). This technique is also used to expel Cold in order to disperse Blood stagnation.
5. Magnets use magnetic patches or strips that are attached to various channel points of the patient's body to stimulate a response in the electromagnetic field (Figure 1.9). Magnet therapy has been used since the Tang Dynasty (618–907 A.D.). The magnets are applied to specific points for a period of 3–5 days, removed for one day, then reapplied. Whether used for tonification or sedation, this therapy facilitates constant treatment of the channel point.

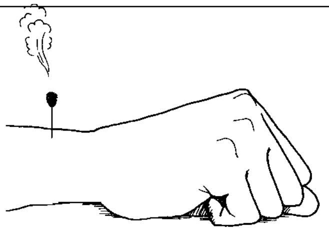

Figure 1.8. The doctor has inserted a needle with burning Moxa (herbal cones/sticks) on top of it. Burning moxa may also be held over specific channel points or placed directly on the skin.

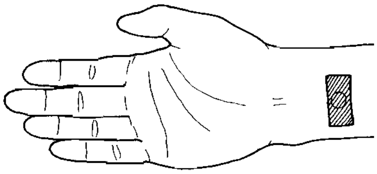

Figure 1.9. A Magnetic Patch is being worn on the Wrist. Magnets may be attached to various parts of the body or may be worn in shoes or wristbands.

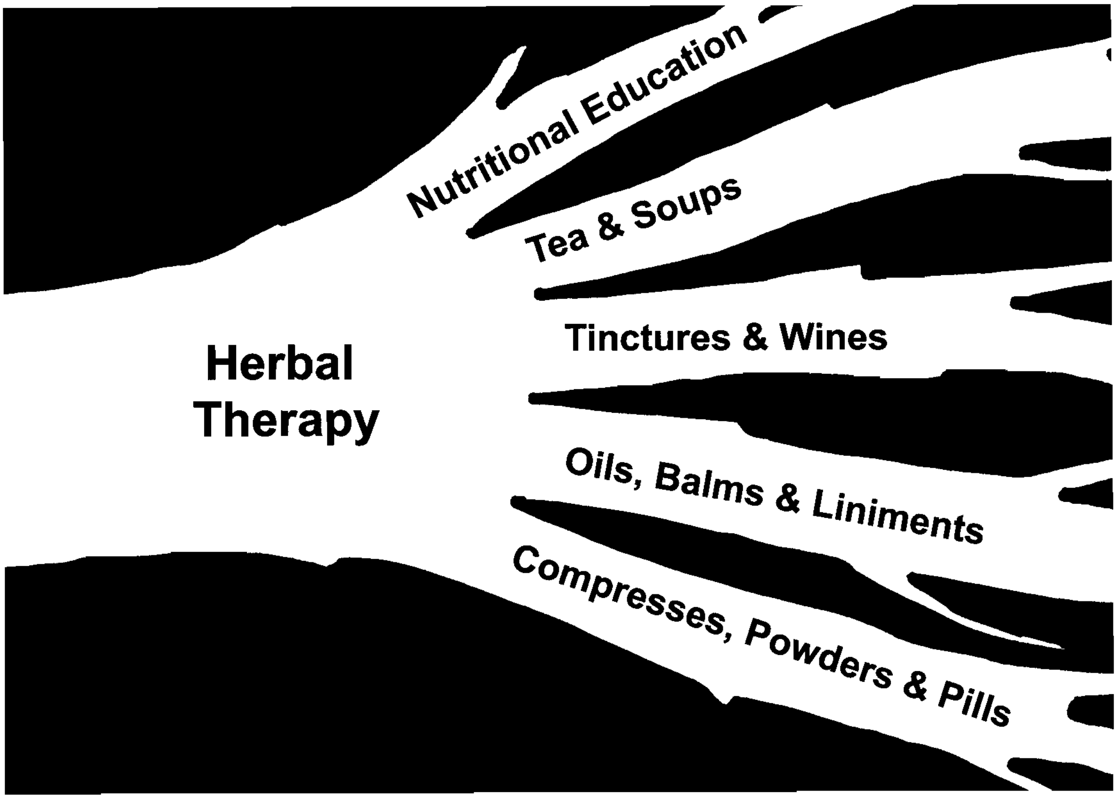

Figure 1.10. The Five Main Branches of Chinese Herbal Therapy

HERBAL THERAPY

Herbal formulas have been used successfully to treat a variety of illnesses for over 5000 years. Historically, herbal medicine has been the worldwide basis for pharmaceuticals used in most cultures, prior to modern times. Today, herbs provide the source for many of the pharmaceuticals used in contemporary Western medicine, especially for the treatment of viral and bacterial diseases, pain, tumors, chronic diseases, internal and external tissue regeneration, and many other medical problems.

Herbology is both a science and an art. An herbalist spends many years studying the herbs used to create herbal formulas. The herbalist must understand the effects of individual herbs, as well as their synergistic effects when combined.

Herbs are used for tonifying, purging, dispersing, warming, cooling, nourishing the Yin, nourishing the Yang, and clearing Heat, as well as moving Qi, Blood, Phlegm, and Fluids within the body. They cause the Qi in the body to either ascend or descend, affecting the upper or lower parts of the body.

Chinese medicine prescribes specific parts of plants (leaves, roots, bark, etc.) for particular medicinal purposes. Parts of trees, shrubs, herbs, vines, and flowers are selected for their specific properties (Hot, Cold, Warm, Cool) and taste (Sweet, Bitter, Pungent, Sour, and Salty). These properties either tonify or disperse Qi and Blood.

Chinese medical herbology not only includes the cultivation and gathering of seeds, fruits, flowers, leaves, barks, stems, and roots but also prescribes non-herbal components (such as minerals, animal or insect parts) that are sometimes added to enhance the herbs' healing effect.

Chinese herbs cure energetically by moving Qi in the channels. Different herbs enter different channels and affect different internal organs. Herbs are extremely powerful. The herbalist uses herbs to tonify (strengthen) and move Qi and Blood as well as eliminate Heat from the patient's Blood. When tailored to an individual's constitution or combined into a formula for specific symptoms, herbs can greatly help the body; however, taking the improper herbal formula can have deleterious effects.

Chinese herbal therapy includes five major clinical applications (Figure 1.10): nutritional education (food and diet); teas and soups (tang); tinctures and wines (jin); oils, balms, and liniments (you and gao); and compresses, powders (san), and pills (wan).

1. Nutritional education is stressed to assist patients in choosing foods for the body's nourishment and optimum health, as well as for the treatment of disease. Foods have many similar properties to herbs. An old Chinese saying asks, "Are herbs food or food herbs?", thus stressing the importance of a good diet
2. Teas and Soups are water-based herbal formulas traditionally prepared from raw or processed herbal ingredients. These are traditionally ingested for the treatment of internal and external disorders of both acute and chronic natures.
3. Tinctures and Wines are both alcohol-based herbal formulas. Tinctures are a concentrated alcohol-based formulas prepared from raw herbs that are used similarly to teas and soups. Wines are traditionally applied externally to alleviate pain, or ingested as a tonic, depending on the specific formula and the disease being treated.
4. Oils, Balms, and Liniments are oil-based herbal formulas usually applied externally for the treatment of muscle, tendon, and ligament trauma, to alleviate pain, disperse Excess Qi, or to draw Qi into specific areas for tonification.
5. Compresses, Powders, and Pills consist of herbs that have been pulverized into a paste, powder, or poultice and are then applied externally for the treatment of acute or chronic injuries. They can also be ingested for internal organ tonification and the rebalancing of the body's energy system. Pills are herbs specifically prepared from traditional formulas, rolled into little balls, and orally ingested for the treatment of Internal disorders.

CHINESE MASSAGE THERAPY

Chinese Massage Therapy is a generic term used for all tissue manipulation techniques currently used in China. This particular branch of Traditional Chinese Medicine consists of five popular systems that include manipulation of not only external skin, muscles, tendons, joints, nerves, and inner fascia but also the internal organs and organ systems.

This ancient therapy is used as a preventative treatment as well as a healing modality. By applying specific methods of tissue manipulations, obstructions in the channel's pathways can be removed, promoting and increasing both Qi and Blood circulation.

This therapy focuses on improving the structural alignment of the body and on healing soft-tissue injuries. It also corrects any deviant functions of the internal organs, nerves, and joints. Chinese bodywork and tissue therapy are the foundational source for modern Swedish massage, myofascial trigger point therapy, reflexology, and therapeutic neuromuscular therapy.

Chinese massage therapy is divided into five different schools of instruction: Jie Gu, Tui Na, Gua Sha, An Mo, and Jing Point therapy (Figure 1.11). Jie Gu, Tui Na, and Gua Sha employ external tissue manipulations. These three external manipulations are used to treat the bones, muscles, ligaments, and tendons, and also to treat fevers. An Mo and Jing Point therapy utilize soft-tissue manipulation. These treatment modalities are similar to those used in chiropractic, osteopathy, Western physical therapy, and massage therapy.

1. Jie Gu Therapy is used for bone setting and

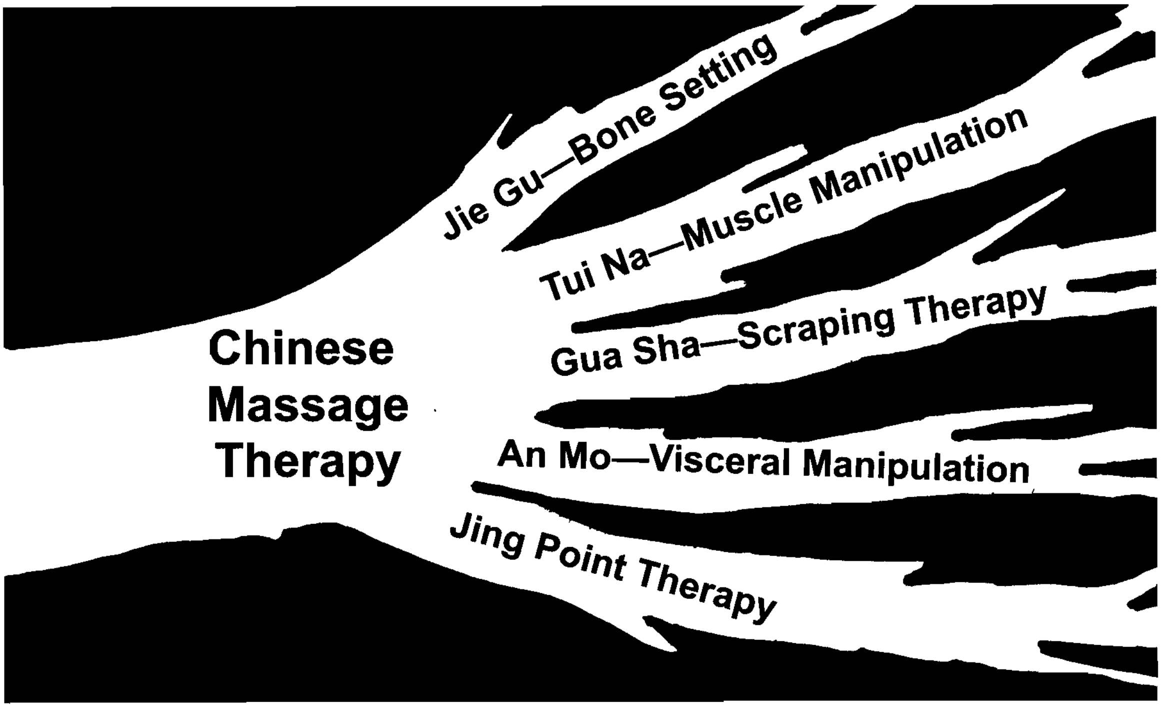

Figure 1.11. The Five Main Branches of Chinese Massage

to adjust the patient's body alignment (Figure 1.12). The literal translation for Jie Gu is "knotted bone," which describes the art of manipulating the bones and ligaments to unravel the trapped junctions of Blood and Qi channels in the patient's joints.

2. Tui Na Therapy focuses on external tissue manipulation and adjustment of the muscles and tendons to adjust abnormal Qi circulation within the body's muscular system (Figure 1.13). The translation for Tui Na is to "push and grasp." It was developed primarily for correcting the misalignment of the body's bones and muscles due to traumatic physical injuries. Reflexology is historically rooted in the use of Tui Na therapy in Chinese pediatric care.
3. Gua Sha Therapy is used to regulate febrile conditions, such as flu, cholera and malaria, and to treat musculoskeletal conditions. The translation for Gua is "to scrape or scratch" and Sha is defined as "cholera," or sand-like maculae (referring to the red discoloration that is raised on the skin by the application of scraping). This therapy focuses on external surface tissue scraping, usually around the neck and thorax areas (Figure 1.14). It is commonly used for promoting Qi and Blood circulation, removing toxins, clearing Heat, cooling the Blood, removing stagnation, and dissolving masses. A jade scraper (coin, bowl, or spoon) is used for purifying the Qi and transforming the Shen (Spirit). A water buffalo horn is commonly used for pulling Heat and toxins from the patient's body (occasionally ceramic is used, but never glass or plastic).

The two internal manipulations are used to treat the organs and nerves.

4. An Mo Therapy is used for internal organ

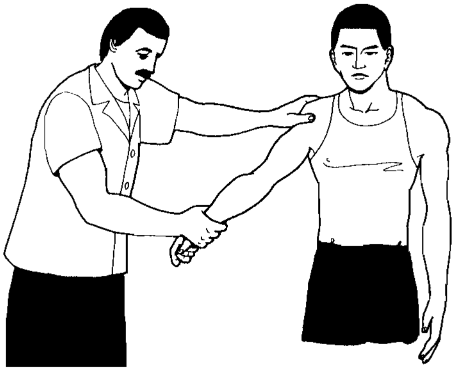

Figure 1.12. Jie Gu Therapy is used to set the bones and ligaments in order to unravel the trapped junctions of Blood and Qi channels at the patient's joints.

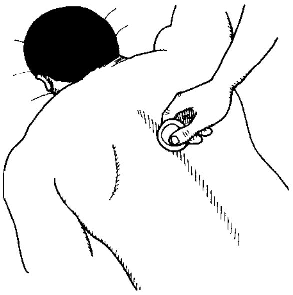

Figure 1.14. Gua Sha Therapy is commonly used for clearing Heat, cooling the Blood, removing stagnation, and dissolving masses.

regulation. An Mo focuses primarily on Qi extension and soft-tissue and internal organ manipulation (Figure 1.15). Although the literal translation means to "press and rub," this therapy focuses primarily on internal visceral regulation, concentrating directly on the treatment of specific Internal diseases.

5. Jing Point Therapy is employed for channel and internal organ regulation. Jing point therapy uses pressing, pinching, clapping, and tapping techniques on specific energetic points and energetic channels. These techniques are employed to promote Qi and Blood circulation, balance the body's Yin and Yang energy, tonify weak organs, dredge the channels, and expel pathogenic factors (Figure 1.16).

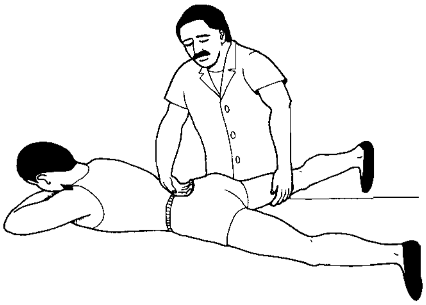

Figure 1.13. Tui Na Therapy is used to adjust the muscles and focuses on external tissue manipulation and adjustment of the muscles and tendons in order to correct abnormal Qi circulation within the body's muscular system.

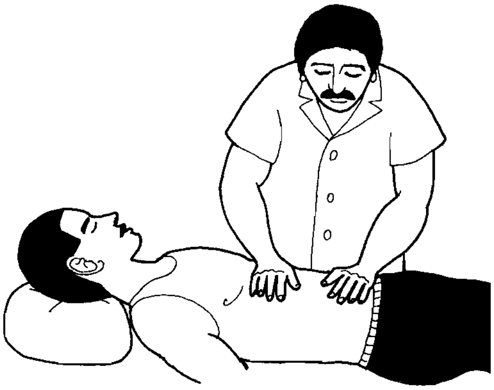

Figure 1.15. An Mo Therapy allows organ manipulation for internal visceral regulation and concentrates directly on treating specific Internal diseases.

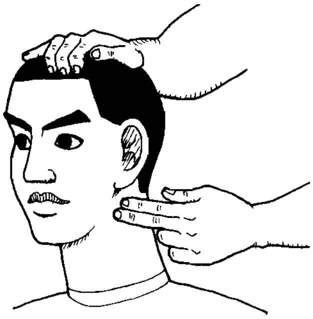

Figure 1.16. Jing Point Therapy is used to promote Qi and Blood circulation, balance the body's Yin and Yang energy, tonify weak organs, dredge the channels, and expel pathogenic factors.

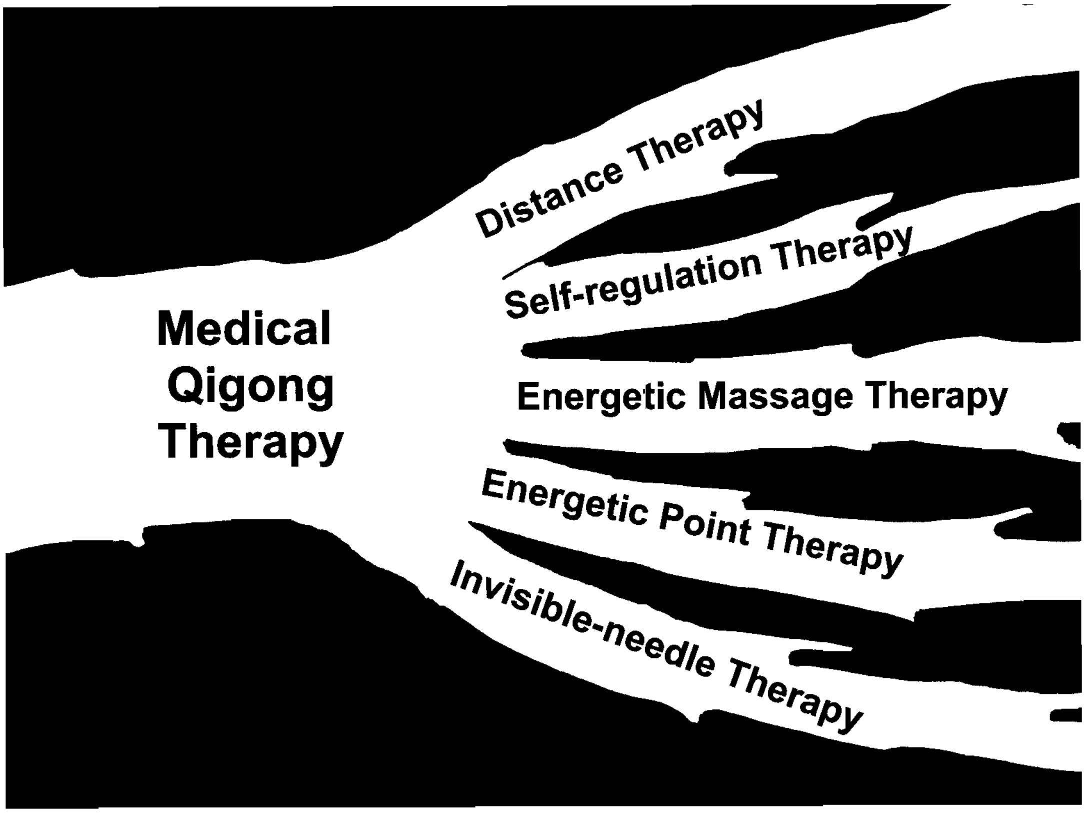

Figure 1.17. The Five Main Branches of Chinese Medical Qigong Therapy

MEDICAL QIGONG THERAPY

The objective for healing disease in Medical Qigong training is threefold:

- First, eliminate Internal pathogenic factors (the accumulation of Excessive emotions such as anger, grief, worry, fear, etc.) as well as External pathogenic factors (the invasion of Cold, Hot, Damp, etc., from the environment).
- Second, increase or decrease the patient's Qi as needed to counteract the Deficient or Excess condition within the internal organs and channels.
- Third, regulate and balance the patient's Yin and Yang energy to bring it back into harmony.

This unique therapy consists of regulating the body's three external Wei Qi fields (physical, mental/emotional and spiritual), and the four Internal fields of life-force energy (Ying Qi, Sea of Blood, Sea of Marrow, and the Taiji Pole). Some of the most common diseases treated in Medical Qigong clinics are: diabetes, arthritis, high blood pressure, breast and ovarian cysts and tumors, migraine headaches, fibromyalgia, insomnia, acute abdominal pain, irritable bowel syndrome, deep tissue obstruction, muscle atrophy, brain tumors, stroke, coma retrieval, and certain types of cancer. The medical treatment focuses on relieving pain, detoxifying the body of toxic emotions (e.g., excessive anger, fear, worry, etc.), correcting internal organ dysfunctions, and balancing Excess or Deficient Qi and Blood conditions.

Medical Qigong therapy uses five major clinical modalities (Figure 1.17): Distance therapy, Self-

Regulation therapy, Qigong Massage therapy, Energetic Point therapy, and Invisible Needle therapy.

1. Distance Therapy (also called Qi Emission) requires the Qigong doctor to manipulate a patient's Qi by focusing on the energetic properties of the patients' channels, collaterals, and points, as well as internal organs, from a distance of several inches, several feet, or even several miles away (Figure 1.18).
2. Self-Regulation Therapy (also called Qigong Prescriptions and Patient Homework) are Qigong exercises (postures, movements, sound vibrations, visualizations, etc.) given to patients by a doctor (Figure 1.19). Patients can use these Qigong techniques to regulate their own health, using various lying, sitting, moving, and standing postures. The patients may also use their own spiritual belief system as a healing tool.
3. Qigong Massage Therapy, a soft-tissue regulation technique, differs from Tui Na or An Mo (Chinese External Massage Therapy), in that the doctor's hand skims the patient's body as lightly as a feather, never exceeding the pressure one would place on an eyeball (Figure 1.20). The light skimming action is used to dredge the patients' external channel Qi, causing energy to be released from the internal channels themselves, which serve as pathways for Qi transference.
4. Energetic Point Therapy is used by the doctor to extend Qi into specific internal and external areas of the patient's body to lead and direct the Qi. This type of therapy requires the doctor and the patient to focus their attention onto a specific energetic point (Figure 1.21). This therapy demonstrates the power of the mind as an active tool in healing, and is used for purgation, tonification, and regulation.
5. Invisible Needle Therapy involves the visualization of imaginary needles of light being inserted into specific points on the patient's body. The needles of light are used to stimulate and direct the patient's Qi (Figure 1.22).

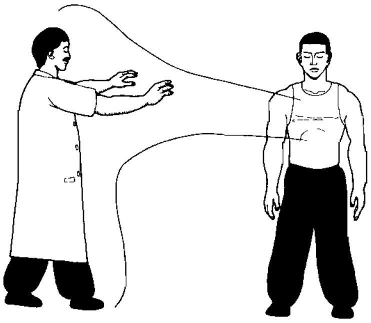

Figure 1.18. In Distance Therapy, the Qigong doctor manipulates a patient's Qi by focusing on the energetic properties of the patient's external channels, collaterals, and points from a distance of several inches, several feet, or even several miles.

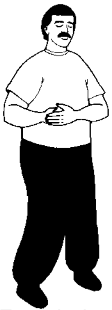

Figure 1.19. In Self-Regulation Therapy, the patients are required to self-regulate by performing Qi Qigong "prescriptions" or "homework" (postures, movements, chants, visualizations, etc.). Here the patient regulates his own *Liver Qi*.

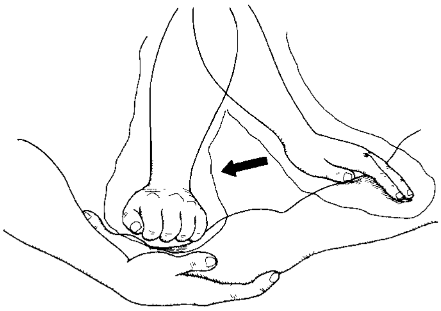

Figure 1.20. In Qigong Massage Therapy, the doctor softly dredges the patient's external channels in order to release energy from the internal channels themselves, which serve as pathways for Qi transference.

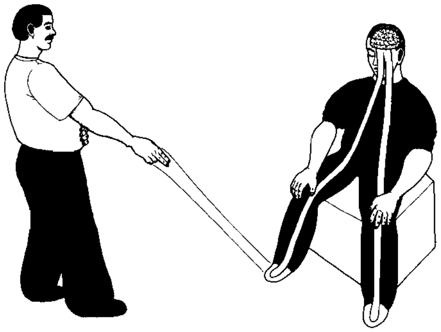

Figure 1.21. In Energetic Point Therapy, the doctor extends energy into the patient's body as both the doctor and the patient focus their attention on a specific channel point, in this case the Kd-1 point.

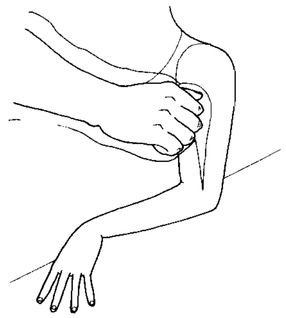

Figure 1.22. In Invisible Needle Therapy, the doctor visualizes inserting energetic acupuncture needles into the patient's channel points in order to stimulate the patient's Qi.

CHAPTER 2
ENERGETIC FORMATION OF THE HUMAN BODY

In 1993 when I began my internship at the Xi Yuan Hospital in Beijing, China, I became aware that most Qigong doctors had a limited understanding of basic anatomy and physiology as we teach it in the West. When I asked why more attention was not paid to the subject, I was told, "The perspective and priority we place on gross physical anatomy and physiology is quite different from yours -- Westerners only study the dead -- we study life, the living pools, rivers, and currents of life-force energy that can only be found in the living body."

Eastern medicine believes that when the body dies, the substance, or energy, that gave the body life returns to its original source (Heaven and Earth) and all that remains of the individual is a mass of lifeless tissue. They believe that the tangible and the energetic exist in a twofold, cohesive relationship. A Qigong doctor must understand this concept of energy in order to comprehend and prescribe appropriate Medical Qigong therapy for the patient.

The concept of the energetic formation of the human body is new to Western thought with its primary focus on the physically tangible. The philosophical foundation of Traditional Chinese Medicine, on the other hand, includes studying the whole human being in all of his or her aspects, physical, mental, emotional, energetic, and spiritual. The conception results not only in the tangible physical form of a new human being but also in various contributing energies and energetic fields. Medical Qigong studies these energies, as well as the interplay between the Heavenly and Earthly influences.

ENERGETIC EMBRYOLOGICAL DEVELOPMENT

The study of the embryological development of the body's inner fascia and connective tissues explains many of the traditional Chinese medical ideas about energy, health, and disease. The cells, tissues, and organs of the human body interrelate as a result of the "energetic patterning" that occurs at conception.

Heaven (which the ancient Chinese considered a state of subtle and non-material energies and beings) and Earth both have energetic fields that influence life. If we consider Heaven Qi (the universal energy related to the sun, moon, and stars) to be electro-positive and Earth Qi (the environmental energy related to the earth, water and wind) to be electro-negative, we can begin to understand the duality of the energetic fields and their pull on the human body.

Think of the body as being suspended between two enormous fields of energy (Figure 2.1): the sun extends its energetic field down, and the Earth's small "sun," i.e. its core, extends its energetic field upwards through the Earth's crust. Man is, therefore, suspended between the electrically positive energy field of Heaven and the electri-

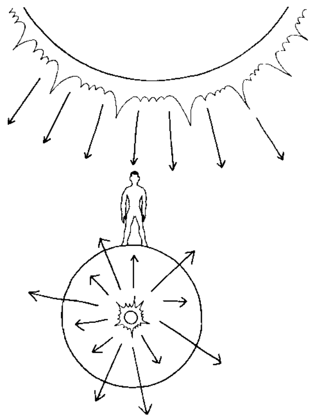

Figure 2.1. Man's energetic field is suspended between the two energy fields of Heaven and Earth.

cally negative energy field of Earth and is affected by both celestial and environmental phenomena.

ENERGY, MATTER, AND SPIRITUAL INTERACTIONS

The *Yellow Emperor's Classic of Internal Medicine* states that the woman's physiology is dominated by Blood, which serves as the foundation of her menstrual cycle, fertility, conception, pregnancy, and childbirth. A man's physiology is dominated by Qi, which serves as the basis of his physical interactions.

At conception, a myriad of energetic interactions begin. The mother, father, Heaven, and Earth all contribute energetically to conception. These four energies condense upon impact and form the nucleus of the energetic cellular patterning of the new life (Figure 2.2). Every part of the resulting human body is affected by this energetic patterning.

An old Chinese saying in Medical Qigong states that, "When people are born, Heaven gives them Jing (Essence) and Shen (Spirit) -- which align to form the mind -- and Earth gives them bones and shape, which unite to form the body. Joined together, these sources of energy cause human beings to develop. When people die, their Essence and Spirit return to Heaven, and their bones and shape go back to Earth."

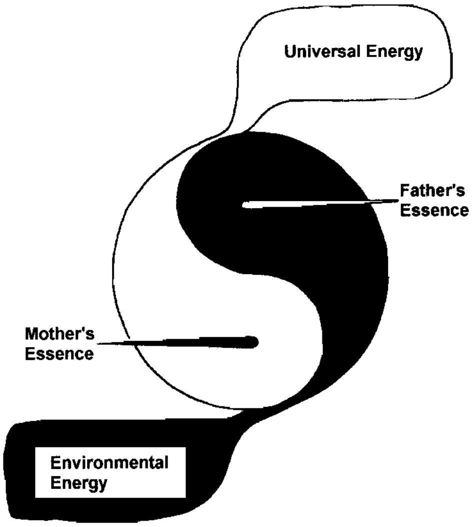

Figure 2.2. The blending of Heaven (universal energy) and Earth (environmental energy) as well as father and mother energies creates a fusion of Yin and Yang energies within the body's tissues and cells during creation. These energies are responsible, on a psychophysical level, for transferring talents and traits from generation to generation.

Pairs of channels from the mother's internal organs carry Qi that creates and nourishes the baby at each stage of development. As the baby develops, sustained through the umbilical cord, the child's navel, Kidneys, and lower abdominal area become the collection points for Prenatal Qi (energy stored within the baby's body before it is born).

Every life begins with inherent strengths and weaknesses. It is believed that when conception occurs, if the mother achieves orgasm, the child's inherent energy will become very strong. However, if the mother does not achieve orgasm during conception, the results will be a normal or weak energetic constitution. Prenatal care is mandatory for the healthy formation of the embryo. Before cell division, the DNA mass must be duplicated exactly in order to transfer normal genetic characteristics to the next generation. Although heredity plays a large part in the transference of both parents' genetic history, a weakness in the mother's channels can result in congenital problems, or toxins, that the fetus can acquire during one of the corresponding stages of development.

UNDERSTANDING FETAL TOXINS

The External invasions of pathogenic toxins (known as fetal toxins) can penetrate the zygote with latent Heat that can cause diseases during early childhood development. It is, therefore, important for both parents, especially the mother, to take responsibility for being strong and healthy at the time of conception and for the duration of the pregnancy. Toxins can be transferred into the embryo in utero in one of two ways:

- First, from either the mother or father at the moment of conception. Toxins transferred from the parents can create an inherited toxicity due to a retention of Hot Evil stored from within either of the parents' Essence and Blood; or
- Second, from Internal Heat generated by the

mother during pregnancy due to improper diet or life-style.

During pregnancy, a fetus is aware of light and sound and of the mother's reaction to the surrounding influences of her environmental energy fields. The fetus is strongly influenced by its mother's physical activities, as well as her mental, emotional, and spiritual states.

Regulating the mother's behavior to improve her child's physical, emotional, and mental health is called "fetal education" in Traditional Chinese Medicine, and is important in the development of the child's Prenatal Essence, Energy, and Spirit. This viewpoint is based on the fact that the mother's Heart and uterus are connected via the mother's internal channels, allowing Qi and Blood to flow into the uterus. Anything that influences the mother's mind, emotions, and spirit affect her Heart, which in turn, affect the fetus via the internal channels.

DEVELOPMENTAL SEQUENCE

The following description of the body's developmental sequence is but one of many theories used in order to explain Chinese energetic embryology. The ancient Chinese did not have a conception of the body's cells and cellular division; however, modern theories have postulated that the channels were formed at the earliest stages of cell division, creating an energy matrix for the developing fetus. This particular theory is being taught at the Hai Dian Medical Qigong College in Beijing, China.

Three important energies (Figure 2.3) combine in the developing fetus: Jing (Essence), Qi (Energy), and Shen (Spirit). Jing is the body's foundational substance, responsible for nourishing the tissues. Qi emerges out of Jing circulating through the tissues and promotes the body's metabolism. Shen governs the body. (see Chapter 11, "The Three Treasures of Man," for more on Jing, Qi, and Shen.)

Both parents contribute energetically to the baby's conception (Figure 2.4). The father's sperm and the mother's egg consist of Jing (Essence), Qi (Energy), and Shen (Spirit). This combination of Jing, Qi, and Shen is referred to collectively as fa-

Prenatal Energies—The Three Treasures

Substance	**Physical Function**	**Energetic Prenatal Function**
Shen (Spirit)	governs the body	the Original Spirit
Qi (Energy)	promotes metabolism	body's Energy extending into Infinite Space (also called Wuji, or the Void)
Jing (Essence)	nourishes the tissues	body's innate true Essence, the foundational substance

Figure 2.3 Qi is the medium between matter (Jing) and Spirit (Shen) and is responsible for the embryo's transformational processes.

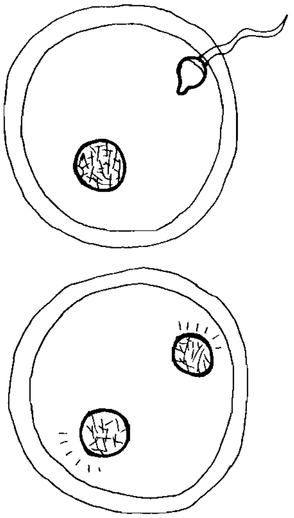

Figure 2.4. The human body is composed primarily of water. The Essence of man (sperm) and woman (egg) unite in the uterine sea to form the fetus. The ova is polarized at the entry point of the sperm, creating the original polar axis (the *Taiji Pole*) that determines the complex pattern of cellular division that occurs along the polar axis throughout development.

ther Qi and mother Qi. The combination of father Qi and mother Qi is known as Yuan Qi or Original Qi.

The quality and quantity of Original Qi that the baby receives at birth depends on three main factors:

- First, the purity and potency of both parents' genetic plasma (sperm and egg);
- Second, the condition of both parents' health and vitality and the state of their physical, mental, emotional and spiritual relationship at the time of conception; and
- Third, the spiritual factors surrounding the conception (i.e., karma that is brought into this life by the incoming spirit/soul).

As the sperm (containing the father's Qi) fertilizes the egg (containing the mother's Qi), Heaven (universal) Qi and Earth (environmental) Qi blend together within the zygote. The swirling and blending of these four energies form energetic pools (which will later evolve into organs), rivers (which later evolve into channels), and streams (which later evolve into collaterals).

When the sperm enters the ovum, it produces a polar axis that creates an energetic vortex. This vortex not only forms the central Taiji Pole, but also draws Qi from Heaven and Earth and the Eternal Soul into the fetus's body. This polarization also determines a ventral and dorsal surface, which become the embryo's Conception and Governing Vessels, respectively, at the first cell division (Figure 2.5). The Governing Vessel controls the cell division that eventually forms the back of the body, while the Conception Vessel controls the cell division of the body's front. This first cell division also establishes a right and left side. The Heel Vessels control the balance of Yin and Yang energy development in the two sides of the body.

The Belt Vessel and Thrusting Vessels form at the time of the second cell division. The four vessels formed at this point (Governing Vessel, Conception Vessel, Thrusting Vessels, and Belt Vessel) are interlinked for the production, circulation, and regulation of the body's Jing-Essence. The body's entire energy system becomes established and maintained within these four vessels when cell division occurs (Figure 2.6).

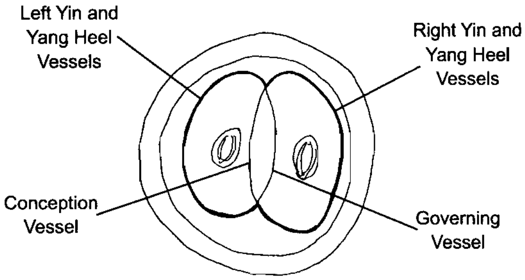
Left Yin and Yang Heel Vessels
Right Yin and Yang Heel Vessels
Conception Vessel
Governing Vessel

Figure 2.5. The electrical polarity and field of energy in the polar axis is related to the Original or Yuan Qi from which the Conception and Governing Vessels form the seas of Yin and Yang energy during the first cellular division. The Yin and Yang Heel Vessels are also established, forming the left and right sides. The exterior of the egg is determined by the Yang Linking Vessels. The interior of the egg by the Yin Linking Vessels.

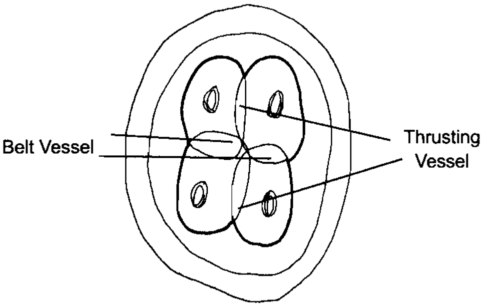

Figure 2.6. The second cellular division is caused by the development of the Belt Vessel and the Thrusting Vessel. The Eight Extraordinary Vessels have now formed.

While the embryo is forming, both the Yang and Yin Linking Vessels are respectively responsible for the exterior and interior development of the embryo (see Chapter 3 for more on Yang and Yin energy). Each of the Eight Extraordinary Vessels has a specific role in the development of the embryo.

1. The Governing Vessel (Yang) controls development of the body's back.
2. The Conception Vessel (Yin) controls development of the body's front.
3. The Thrusting Vessels carry energy through the center of the body and controls the body's center core.

4. The Yang Heel Vessels control the development of the body's right and left Yang energy.
5. The Yin Heel Vessels control the development of the body's right and left Yin energy.
6. The Yang Linking Vessels control the development of the exterior of the body (and correlates to Heaven energy).
7. The Yin Linking Vessels control the development of the interior of the body (and correlates to Earth energy).
8. The Belt Vessel binds all the channels together.

These eight vessels—Governing, Conception, Thrusting, Yang Heel, Yin Heel, Yang Linking, Yin Linking, and Belt—are also known as the Eight Extraordinary Vessels or Eight Prenatal Vessels.

The Eight Extraordinary Vessels form a vortex of energy at the center of the embryo's body—from the area between what will become the Kidneys. The Taiji Pole and Thrusting Vessels are at the center of this vortex and will form the Sea of Five Yin and Six Yang Organs, the Sea of Twelve Primary Channels, and the Sea of Blood (see Chapter 13). From the Taiji Pole and Thrusting Vessels, the body's Qi and Blood are distributed at the energetic level through small channels, or rivers of energy. This energetic vortex creates the energy for the growth of the embryo's physical form.

After the initial cell division is complete, the embryo's ten Yang channels and ten Yin channels begin the development and formation of the embryo's tissues and organs. These twenty channels are divided into two separate groups of energetic rivers known as the Eight Extraordinary Vessels and Twelve Primary Channels. As the embryo develops into a fetus and continues to grow, the twenty channels also continue to develop.

During the formation of the embryo, nine Yang channels begin to flow out of the Governing Vessel (Sea of Yang Qi) and form the nine Yang rivers known as: the Bladder Channels, Gall Bladder Channels, Stomach Channels, Small Intestine Channels, Triple Burner Channels, Large Intestine Channels, Yang Heel Vessels, Yang Linking Vessels, and Belt Vessel.

Consecutively, the nine Yin channels begin to flow out of the Conception Vessel (Sea of Yin Qi) and form the nine Yin rivers known as: the Kidney Channels, Liver Channels, Spleen Channels, Heart Channels, Pericardium Channels, Lung Channels, Yin Heel Vessels, Yin Linking Vessels, and Thrusting Vessels (see Chapter 8).

THE TEN LUNAR MONTHS OF CREATION

The following description of the sequence of embryological development was established by the late Chinese Medical Qigong expert Dr. Chao Yuan Fang during the Sui Dynasty, around 610 A.D. The months that Dr. Chao refers to are the ten lunar

Lunar Month	Mother's Channels	Channel's Activity	Energetic Activity
Month 1	Liver Channels	stop menses, begin embryonic growth	energies combine at conception, channels develop
Month 2	Gall Bladder Channels	saturate embryo with Jing	embryonic Qi transforms into amniotic fluid
Month 3	Pericardium Channels	mother's Qi and Body Fluids purify and cleanse the fetal Shen	Hun and Po establish in the fetus's internal organs
Month 4	Triple Burner Channels	the Yin organs develop	Water Jing accepted into fetus's body
Month 5	Spleen Channels	the Five Agents enter the Five Orbs (Yin organs)	Fire Jing accepted into fetus's body
Month 6	Stomach Channels	Six Pitches established within the Six Storage Areas	Metal Jing accepted into fetus's body
Month 7	Lung Channels	Seven Essential Stars open orifices to the light from Heaven and Earth	Wood Jing accepted into fetus's body
Month 8	Large Intestine Channels	Heart is harmonized and breath is quieted	Earth Jing accepted into fetus's body
Month 9	Kidney Channels	Kidney Channel controls the energetic intake from the umbilicus	energetic boundaries completed
Month 10	Bladder Channels	spatial cavities surrounding major organs are completed	Heaven & Earth Qi settle into baby's Lower Dantian and birth begins

Figure 2.7. Prenatal Energy Development

months. These months compose the 40 weeks of a normal pregnancy (Figure 2.7).

MONTH ONE

The first lunar month of pregnancy is traditionally called the Beginning of Form. This stage covers conception and early cell division.

At conception, the mother's Liver Channels stop her menses and begin to nourish the growth of her embryo (Figure 2.8). During pregnancy, the mother's Blood is transformed into Jing-Essence, that nourishes the mother's body as well as the embryo's. The mother's Liver Channels cause Essence and Blood to coagulate in her womb. This Blood coagulation continues after the initial cellular division.

At this stage the mother's Shen (Spirit) becomes part of a threefold activity:

- First, the mother's Shen projects through the umbilical cord like a light reflecting off a prism, sustaining and energizing the production of Jing (see Chapter 13), Qi, and Blood for the embryo (Figure 2.9).
- Second, the mother's Shen influences the embryo's Qi and Blood, and the embryo's Shen (Spirit) is stimulated, bringing it into active being (see Chapter 12).
- Third, the embryo's Original Jing combines with the embryo's Original Qi to create the embryo's Original Shen (Prenatal Spirit), which appears as multicolored light and contains the inherited knowledge of the fetus's ancestors, including talents, skills, and natu-

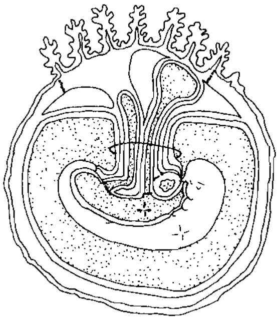

Figure 2.9. Sustained through the umbilical cord, the embryo absorbs the mother's Blood, Essence, Energy, and Spirit.

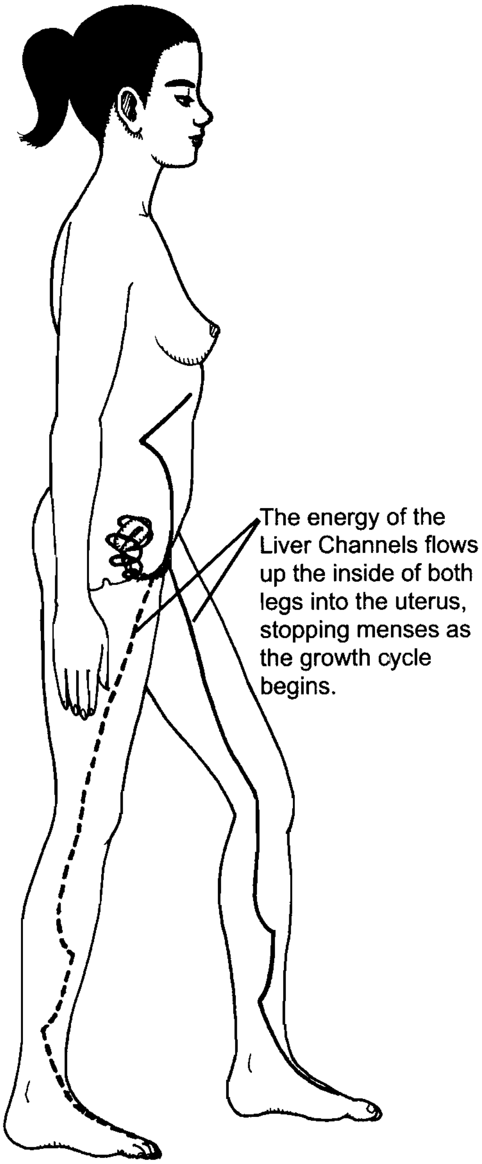
The energy of the Liver Channels flows up the inside of both legs into the uterus, stopping menses as the growth cycle begins.

Figure 2.8. The mother's and father's Essence (Jing), Energy (Qi), and Spirit (Shen) blend with the Heaven and Earth energies during the fusion of the sperm and egg. During the first lunar month, the mother's Liver Channels stop her menses and begin the embryonic growth cycle. After the initial cell division, the ten Yin and ten Yang Channels develop within the embryo.

ral abilities. This knowledge is stored deep within the fetus's cells, tissues, and consciousness.

MONTH TWO

The second lunar month of pregnancy is traditionally called the Beginning to Gel stage. During this period, the responsibility of the mother's Gall Bladder Channels is to create the environment for the development of the mesenteric membrane sac (Figure 2.10). The Gall Bladder Channels will also saturate the embryo, uterus, and placenta with Jing, causing the embryonic Qi to become denser, until it transforms into a thick liquid (amniotic fluid). The embryonic fluid will regulate the embryo's Body Fluids. At this stage, the embryo begins to take shape inside the uterine lining (Figure 2.11). The Lungs, Liver, Kidneys and major blood vessels are forming.

With the formation of the major blood vessels, the Yin and Yang energies begin to occupy the embryo's channels. As the Yin and Yang energies actively balance themselves:

- Energy that will later coalesce into Lung Qi moves to the upper part of the body;
- Original Qi (Yuan) of the Kidneys begins to collect deep in the center of the body;
- Earth Qi (the energy absorbed by the embryo from the mother's exposure to the outside environment) begins to collect in the lower front and upper back areas of the embryo's body.

All of these energies seek their own location

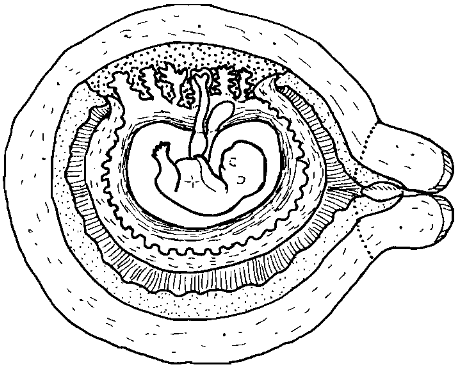

Figure 2.11. During the second month, the embryo begins to take shape as the energetic boundaries begin to form.

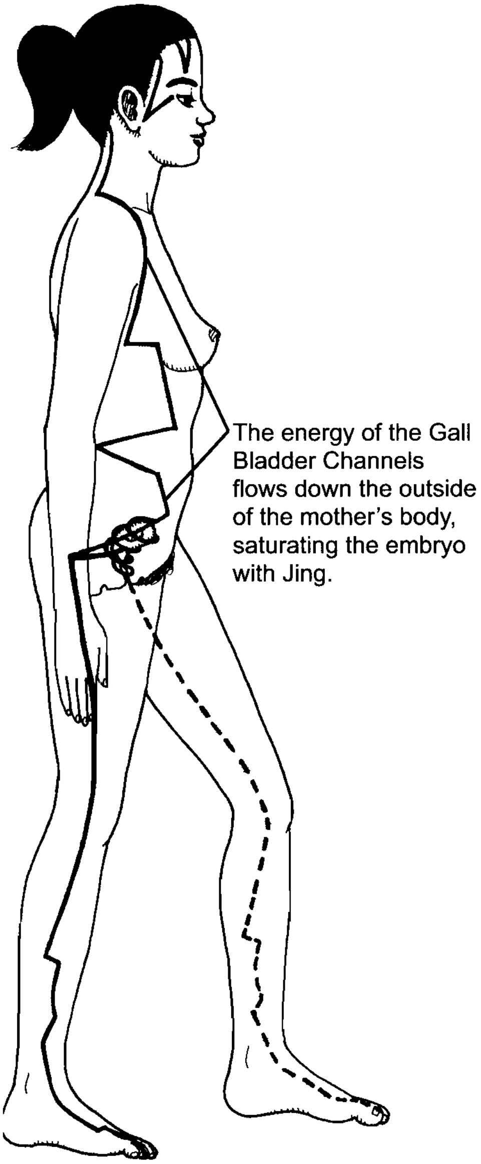

The energy of the Gall Bladder Channels flows down the outside of the mother's body, saturating the embryo with Jing.

Figure 2.10. The mother's Gall Bladder Channels saturate the embryo with Jing during the second lunar month, causing the embryonic Qi to transform into amniotic fluid. The embryo begins to take shape as the energetic boundaries, pools, spatial cavities, and energetic channels and collaterals create internal and external form.

within the embryo's developing body, creating their own pools of Qi with their own discrete energetic boundaries. These boundaries will later form the spatial cavities that surround the major organs. The areas where the energetic pools settle and begin to create a balance within themselves, will be called the organ's place of origin. This flux and movement creates tiny energy currents, eddies, and whirlpools that flow within the body. As the energy shifts, seeking balance, the larger pools of energy begin to form the major organs. This process continues, creating the brain, bones, and skin.

Once the energetic pools and rivers have formed, all of the currents and eddies evolve to form the energetic channels and collaterals, through which energy will flow continuously. These energetic currents will move in accordance with the mother's energetic respiratory patterns and will later (after birth) follow the rhythmic patterns of the child's respiration. As the energy currents continue to spiral within the channels, energetic points are established, following the body's energetic blueprint. Some of these areas spiral outward to form energetic exit points, while others spiral inward to form entry points.

MONTH THREE

The third lunar month of pregnancy is traditionally called the Beginning of the Pregnant uterus. During this period, the embryo becomes a fetus and begins micro-movement. Its heartbeat can now be detected.

The mother's Pericardium Channels control the presence and amount of Jing and Shen in the fetus's vessels, channels, and collaterals (Figure 2.12). The Jing and Shen that flow from the mother are ultimately rooted (firmly established) in the energy of her Blood. The combination of the mother's Qi and Body Fluids purifies and cleanses the Shen of the fetus, which was formed at conception. This purifying action transforms into Heat, causing the Yang energy to arouse the Hun-Three Ethereal Souls (the energy of the fetus's spirit), into life. The Five Agents, energies that stem from the Three Ethereal Souls, are currently in a state of awakening and will later reside in the organs.

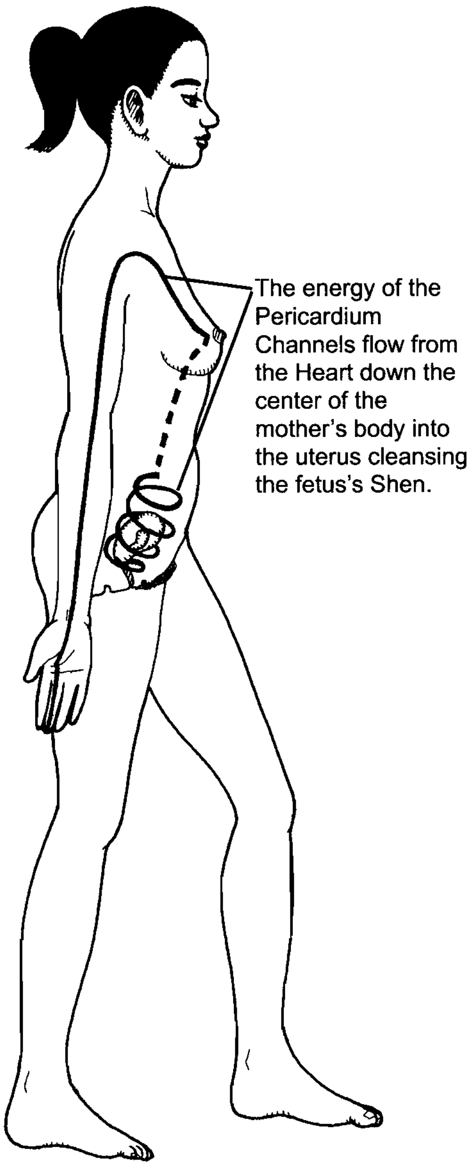
The energy of the Pericardium Channels flow from the Heart down the center of the mother's body into the uterus cleansing the fetus's Shen.

Figure 2.12. The mother's Pericardium Channels control the third lunar month of creation. The mother's Qi and body fluids purify and cleanse the fetus's Shen (Spirit). The Hun from Heaven and the Po from Earth are established within the fetus's internal organs.

The Five Agents are energies that are linked to a person's moral qualities and can be categorized as the five virtues of the Five Yin Organs. The Five Agents are also categorized into Five Elements. The psycho-emotional components of the Prenatal Five Agents and Postnatal Five Emotions are both stored within the body's Five Yin Organs, and are expressed as the child grows into adulthood (Figure 2.13). When one of the internal organs is stimulated, a Yin or Yang psycho-emotional reaction is created. Psycho-emotional energetic interactions are feeling manifestations expressed through either the positive moral qualities of the congenital Five Agents (Yang - kindness, order, trust, integrity, and wisdom) or the negative developed emotional experiences of the Acquired Five Emotions (Yin - anger, joy, worry, grief, and fear).

The Five Agents are connected to the Wu Jing Shen-Five Essence Spirits (Hun, Shen, Yi, Po, and Zhi), and stored within the energetic elemental nature of the body's Essence (Jing) of Wood, Fire, Earth, Metal, and Water. The Five Elemental Energies however, encompass not only the body, but all of the myriad phenomena of the body and nature, combining and recombining in infinite ways to produce manifested existence.

1. The Wood Agent (the Virtue of Kindness) represents benevolence and compassion. This agent is connected to the Hun-Three Ethereal Souls, and stored in the Liver. Wood affects the energetic channels' flow, tendons, ligaments, small muscles, peripheral nerves, iris of the eyes, vision, tears, bile, nails, and external genitalia. After birth, the Liver will also store the emotions of anger, irritability, blame, rage, resentment, and jealousy. When excess anger is eliminated, benevolence, compassion, and love for others is allowed to flourish.
2. The Fire Agent (the Virtue of Order) represents peace and boundary setting to foster social harmony. This agent is connected to the Shen-Spirit, and stored in the Heart, affecting the energetic channels' flow, blood vessels, complexion, perspiration, and the tongue. After birth, the Heart will also store the emotions of shock, nervousness and excitement. Eliminating excess nervousness allows love, forgiveness, and joy to be experienced. The environment is then conducive for peace, contentment, and orderliness, which allow self-esteem to grow.
3. The Earth Agent (the Virtue of Trust) represents faith, honesty, openness, acceptance, and truthfulness. This agent is connected to the Yi-Intention (thoughts and ideas) and stored in the Spleen, affecting the energetic channels' flow, large muscles, lymph and saliva secretions, mouth, lips, and taste. After birth, the Spleen will also store the emotions of worry, remorse, regret, obsessiveness, and self-doubt. Eliminating excess worry allows trust and peace of mind to exist.
4. The Metal Agent (the Virtue of Integrity) represents righteousness and dignity. This agent is connected to the Po-Seven Corporeal Souls (the material aspect of the spirit) and stored in the Lungs, affecting the energetic channels' flow, skin and mucous membranes, body hair, nose, and the sense of smell. After birth, the Lungs will also store the emotions of grief, sorrow, anxiety, sadness, shame, disappointment, and guilt. Once excess sorrow is relieved, a deeper sense of righteousness, integrity, dignity, and social responsibility exists.

The Five Elements	**Congenital Agents**	**Acquired Emotions**
Wood	Kindness	Anger
Fire	Order	Joy
Earth	Trust	Worry
Metal	Integrity	Grief
Water	Wisdom	Fear

Figure 2.13. The Five Agents

5. The Water Agent (the Virtue of Wisdom) represents rationality, clear perception, and self-understanding. This agent is connected to the Zhi-Will (mental drive and determination) and stored in the Kidneys, affecting the energetic channels' flow, brain, inner ear, hearing, spinal cord, cerebrospinal fluid, bones, bone marrow, ovaries, testes, head and pubic hair, anus and urethra, and sexual fluids. After birth, the Kidneys will also store the emotions of fear, loneliness, and insecurity. Once excess fear is eliminated, the mind becomes rational and, therefore, wise.

The energy of the Three Ethereal Souls (Hun) is composed of three separate parts and flows with the Blood as it moves. The energy of the Corporeal Souls (Po) is composed of seven separate parts and follows the Jing as it moves. The Hun and the Po are characterized by their movements. The Po are also referred to as the supernatural Ling (A Moving Spirit). The Hun are also referred to as the supernatural Shen (Mind / Spirit). The Shen is the "active impulse" and Ling is the "active, enabling mover." As the Jing and Blood combine, the fetus's Shen continues to be created.

At the end of the third month, the internal organs, limbs, and external sex organs of the fetus are fully formed, and the nails have developed.

MONTH FOUR

During the fourth lunar month, the mother's Triple Burner Channels, which are connected with the Yang organs, stabilize the fetus's blood vessels (Figure 2.14). The Water Jing is beginning to be accepted by the fetus's body allowing the Yin organs to develop normally.

Beginning in the fourth month and continuing throughout the ninth month, each of the Five Element's energetic nature and the specific characteristics of each element's Essence will be progressively activated and developed within the fetus's body. The first to enter is the Water Jing.

The Water Jing energy supervises the genetic developmental phase of the fetal growth. This energy encompass the fetus's unconscious reservoir of innate and intuitive intelligence, will, and life-force energy, relating to divine love, power,

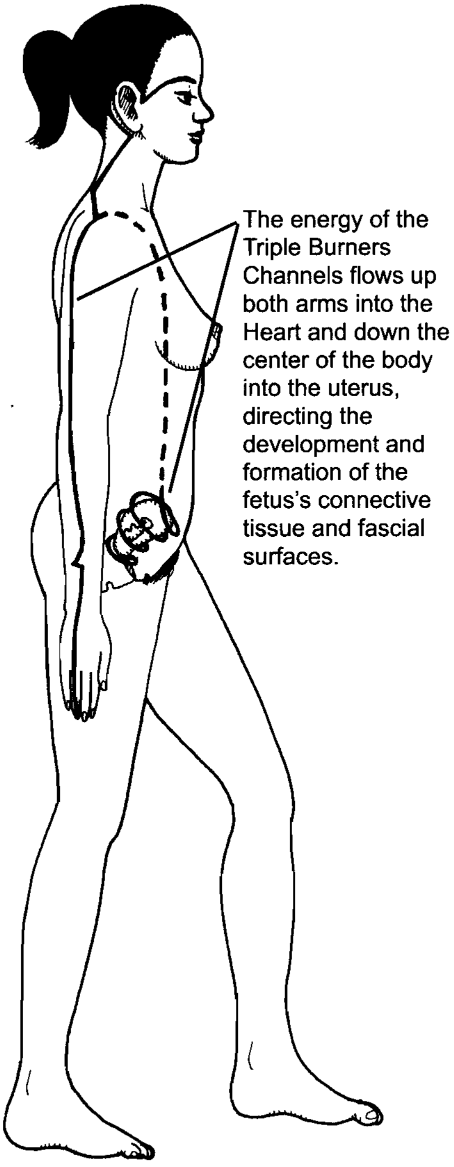

The energy of the Triple Burners Channels flows up both arms into the Heart and down the center of the body into the uterus, directing the development and formation of the fetus's connective tissue and fascial surfaces.

Figure 2.14. The mother's Triple Burners Channels are responsible for the changes in the fourth lunar month of creation. The Water Jing (Essence) is beginning to be accepted by the fetus.

and spirit. Any faltering of this energy (due to the influence of fetal toxins) is associated with both pervasive and subtle neurological disorders, and a predisposition to severe psychological disorders (e.g., schizophrenia).

As the fetus develops, the mother's Triple Burners will direct the development of the connective tissues and fascial surfaces. This process is referred to as "the development of Blood and Qi penetrating to the ears and eyes and circulating throughout the fetus's channels and connecting vessels."

Through the later stages of fetal development the embryo's seemingly homogenous tissues transform into the fetus's differentiated tissues of muscles, bones, and organs (an important part of which is connective tissues). A very large part of the body consists of connective tissues and membranes functioning in such a way as to hold together and maintain the body's external and internal structures.

From a gross anatomical level, the body's structures connect not only through the fascial planes but microscopically through the connective tissues. This internal network facilitates the body's intercellular communication. These structures, both energetic and physical, connect the fetus's body, forming a vast reservoir capable of regulating and transferring the body's Jing-Essence, Qi-Energy, and Shen-Spirit.

MONTH FIVE

During the fifth lunar month, the Spleen Channels become responsible for completing the development of the four limbs (Figure 2.15). The fetus begins its own respiratory movement along with the mother's respiration. The Fire Jing is accepted into the fetus, creating Internal Qi, that stabilizes the fetus's Five Yin Organs' Qi.

The Fire Jing energy generates and controls, protects and integrates, divides and harmonizes the fetus's internal energies to promote emotional/spiritual well-being. Any faltering of the Fire Jing energy is associated with problems of right (Yin) and left (Yang) brain communication (e.g., the correct balance of male/rational and female/intuitive energies).

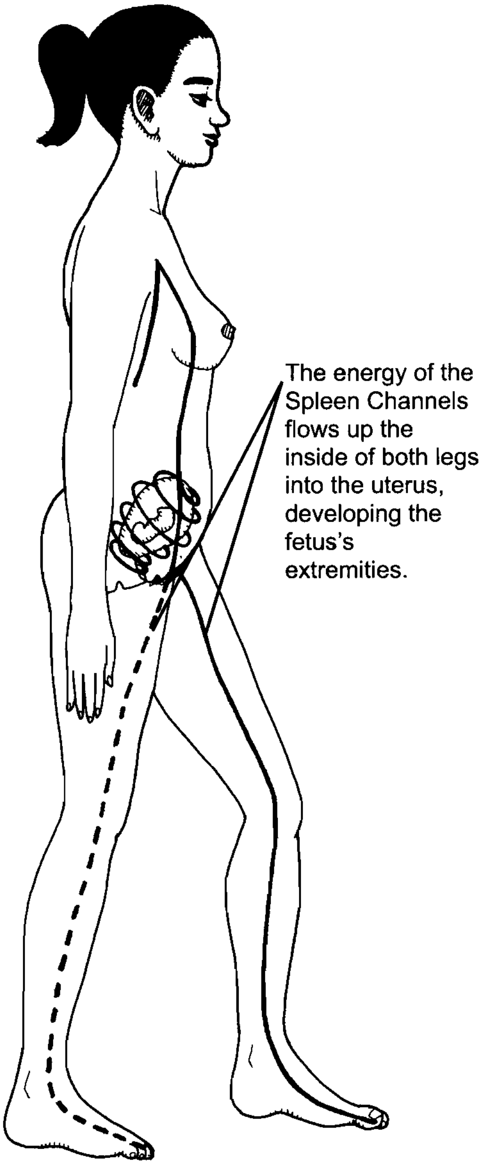
The energy of the Spleen Channels flows up the inside of both legs into the uterus, developing the fetus's extremities.

Figure 2.15. The mother's Spleen Channels are responsible for the fifth lunar month of creation. The development of the fetus's four limbs is completed. The Five Agents are distributed within the fetus's Five Orbs (Five Yin Organs). The Fire Jing is beginning to be accepted by the fetus.

At this stage of development, the Five Agents are distributed to the Five Orbs. These Orbs are energetic spheres of influence or energetic spatial cavities (i.e., internal organ tissue chambers). These Orbs pertain to the Essence, Blood, Qi pools and channels of the Five Yin Organs. The Five Orbs also pertain to and encompass all of the areas in the body that the Five Yin Organs influence. The Five Agents are distributed as follows: Kindness to the Liver, Order to the Heart, Trust to the Spleen, Integrity to the Lungs, and Wisdom to the Kidneys. This distribution causes the Hun to stabilize within the fetus's organs, which contributes to keeping the fetus's Shen at peace.

At the end of the fifth month the fetus's body systems develop rapidly. Its head is less disproportionate to the rest of the body, and its spontaneous muscular movements are commonly felt by the mother.

MONTH SIX

In the sixth lunar month of creation, the mother's Stomach Channels create the fetus's muscles. The Metal Jing is established in the fetus's body, stabilizing the sinews and connective tissues (Figure 2.16).

The Metal Jing energy is also responsible for fetal formation and the ability to form and maintain emotional bonding with others. Any faltering of the Metal Jing energy is associated with problems of emotional attachment (e.g., autism).

During the sixth month the Yang organs are in the process of developing. The Essence of the Yang organs receives, moves, transforms, digests, and excretes substances. The Six Storage Areas of the body's Yang organs constantly fill and empty, and include the Bladder, Gall Bladder, Stomach, Large Intestine, Small Intestine, and Triple Burners.

1. The Bladder controls, stores, and releases the urine. This Yang organ is responsible for receiving, storing, and releasing urine at the proper time.
2. The Gall Bladder stores and secretes bile. This Yang organ is responsible for storing and releasing bile at the proper time.
3. The Stomach stores the food. This Yang or-

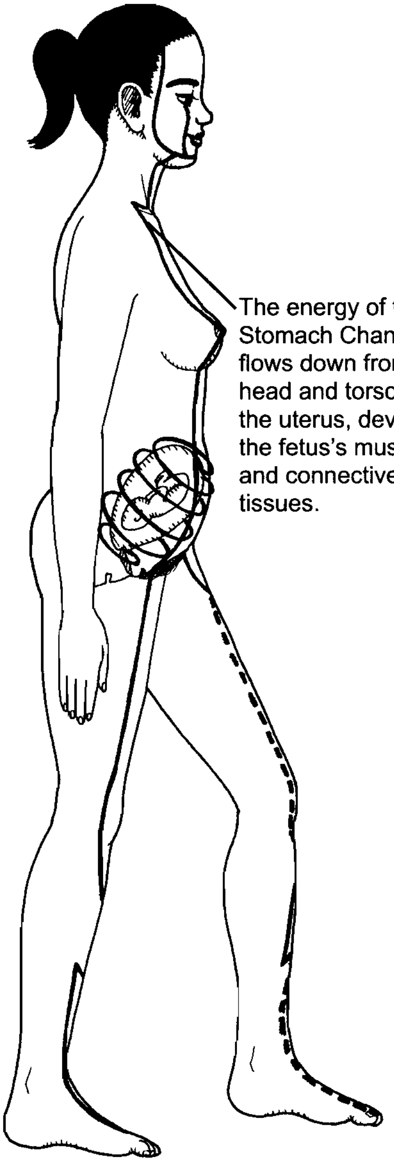

The energy of the Stomach Channels flows down from the head and torso into the uterus, developing the fetus's muscles and connective tissues.

Figure 2.16. The mother's Stomach Channels are responsible for the sixth lunar month of creation. The Six Pitches are established within the body's storage areas. The Metal Jing is beginning to be accepted by the fetus's body.

gan is responsible for the receiving, storing, rotting, and ripening of food at the proper time.

4. The Large Intestine stores the solid waste. This Yang organ receives, stores and absorbs food and releases waste at the proper time.
5. The Small Intestine stores and transforms liquid matter. This Yang organ receives, stores, transforms, and digests food, and releases its waste products at the proper time.
6. The Triple Burners store the body's Qi. This area of the body receives, stores, absorbs, and moves Qi.

At this stage of development, the Six Pitches which support and stabilize the Lower Burners (see Chapter 6) and nourish the Qi are established in the organs known as the Six Yang Organs, also called the Six Storage Areas.

The Six Pitches are six specific tone resonances (notes) that vibrate within the body's internal organs and stimulates specific organ and tissue areas. These Six Pitches relate to the Five Prenatal (Heaven) Elemental Sounds of Jue-Wood, Zhi-Fire, Gong-Earth, Shang-Metal, and Yu-Water. The sixth note Xi relates to the Postnatal (Earth) Fire Element and corresponds to the Pericardium and Triple Burners. The ancient Chinese used these Six Pitches for specific clinical treatments. The sound "Yu" for example, spoken in a low tone, will vibrate the lower abdominal area and is used for the treatment of Kidney and Bladder problems.

In the sixth lunar month of formation, the fetus's eyelids separate and eyelashes form, and its skin is wrinkled.

MONTH SEVEN

During the seventh lunar month, the mother's Lung Channels create the bones, skin, and hair (Figure 2.17). The Wood Jing is beginning to be accepted by the fetus's body.

The Wood Jing energy supervises the assertion and direction of the fetus's emotional/spiritual aspects. Any faltering of the Wood Jing energy is associated with severe psychological problems (e.g., passive-aggressive personality disorder).

In the seventh lunar month, the fetus's Stom-

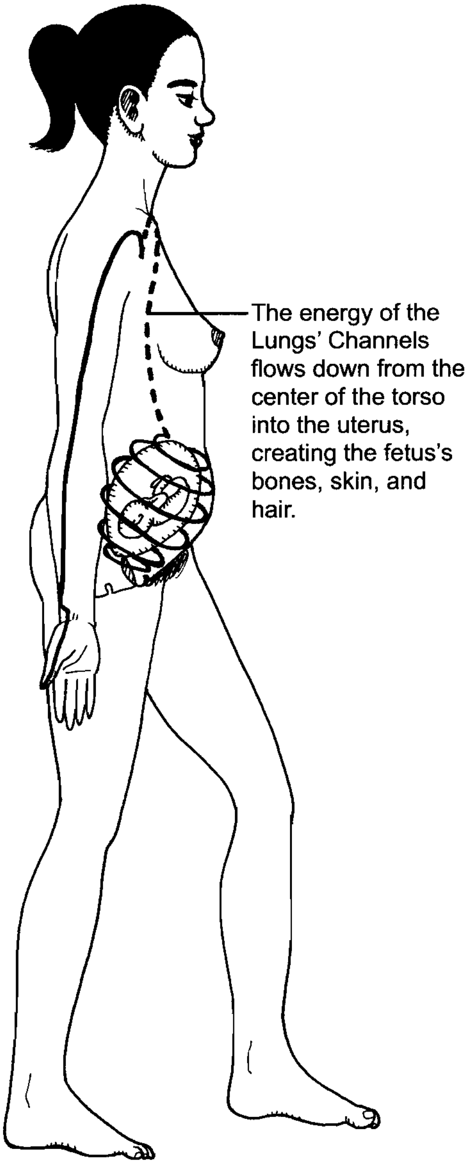
The energy of the Lungs' Channels flows down from the center of the torso into the uterus, creating the fetus's bones, skin, and hair.

Figure 2.17. The mother's Lung Channels are responsible for the seventh lunar month of creation. The Seven Essential Stars open the orifices to let in the light from Heaven and Earth. The Wood Jing is beginning to be accepted by the fetus's body.

ach and Intestines are stabilized, and the Seven Essential Stars open the body's orifices to absorb the light from Heaven and Earth. These stars consist of the sun, moon, and five planets: Mars, Venus, Mercury, Saturn, and Jupiter. Each star is associated with one of the body's orifices: eyes, ears, nostrils, mouth, anus, and urethra. These orifices serve as receiving and projecting energetic portals for Jing-Essence, Qi-Energy, and Shen-Spirit. They also serve as messengers of the body's Five Yin Organs. Thus the Liver receives messages through observation, the Heart through speech, the Spleen through taste, the Lungs through smell, and the Kidneys through hearing.

These energetic messages are received by the body's Wu Jing Shen and emotionally and energetically interact through the body's Six Openings.

1. The eyes absorb images into the Liver which affect the Hun (The Three Ethereal Souls).
2. The ears absorb sounds into the Kidneys which affect the Zhi (Will Power).
3. The nose absorbs smells into the Lungs which affect the Po (The Seven Corporeal Souls).
4. The tongue absorbs tastes into the Spleen which affect the Yi (Intent / Intellect).
5. The physical body absorbs sensations into the tissues which affect the Shen.
6. The spirit (along with the physical body) absorbs sensations into the Heart which also affect the Shen (Spirit).

Physically, during this lunar month, there is a substantial increase in the fetus's weight and its head and body are more proportionate. The fetus can survive if born prematurely (between 27 and 28 weeks); however, its hypothalamic temperature regulation and the Lungs' production of surfactant (a phospholipid substance important in controlling the surface tension of the air-liquid emulsion present in the lungs) are still inadequate.

MONTH EIGHT

In the eighth lunar month, the mother's Large Intestine Channels complete the formation of the fetus's skin, harmonizing the Heart (the Shen), and quieting the breathing. As the fetus receives the Zong Qi-Essential Qi (energy collected from Heaven and Earth and accumulated within the

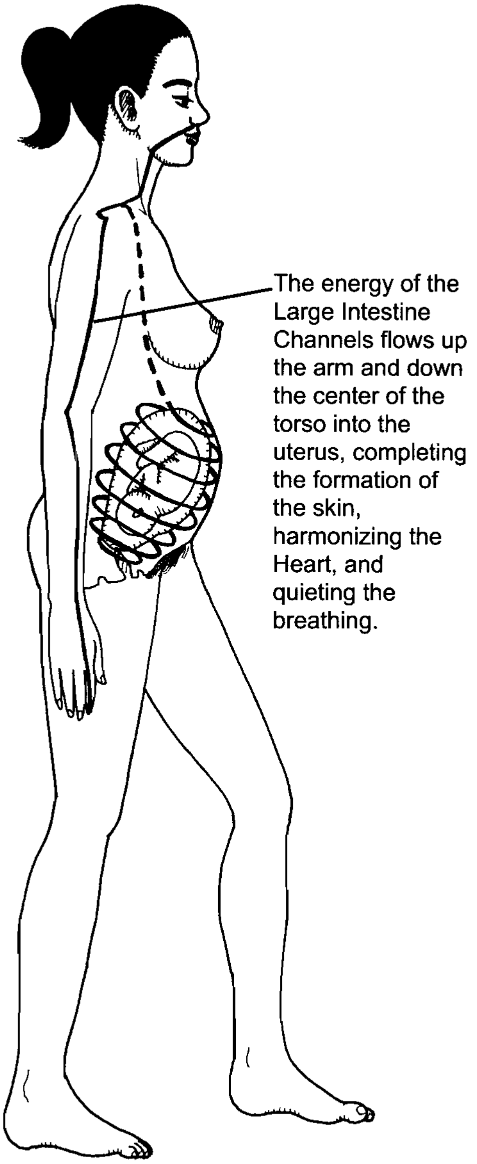
The energy of the Large Intestine Channels flows up the arm and down the center of the torso into the uterus, completing the formation of the skin, harmonizing the Heart, and quieting the breathing.

Figure 2.18. The mother's Large Intestine Channels are responsible for the eighth lunar month of creation. The Earth Jing is beginning to be accepted by the fetus's body.

chest) from the mother's Spleen, the Earth Jing begins to be accepted by the fetus's body, completing the formation of the skin (Figure 2.18)

The Earth Jing energy supervises the quality and maturation of the fetus's emotional and spiritual bonding and boundaries. Any faltering of the Earth Jing energy is associated with problems of severe psychological disturbances (e.g., schizophrenia). These psychological disturbances may be evident at birth or develop later in life.

The mother's Large Intestine Channels control the fetus's orifices. At this stage in development, the fetus's muscles and flesh are formed. The formation and consolidation of the fetus's Jing is now completed, along with the fetus's developing Zhen Qi-True Qi (energy that circulates in the body's channels and collaterals which nourishes the Yin and Yang organs and fights disease).

At the end of the eighth month, the bones of the fetus's head are soft, its skin is less wrinkled, and there is subcutaneous fat deposited throughout its body. If it is a male child, its testes will now descend into the scrotum.

At this stage, the fetus will normally assume an upside-down position to prepare for its birth descent. If the fetus is born prematurely, its chances for survival are now much greater.

MONTH NINE

During the ninth lunar month, the mother's Kidney Channels control the amount of energetic intake the fetus absorbs through its connection to the umbilicus (Figure 2.19). Qi, Blood and food nutrition are absorbed into the fetus's body, after they pass through the umbilical vein to the fetus's Liver. From the Liver, the nutrients are processed and absorbed into the Blood to be distributed throughout the fetus's body.

The flexible structure of the umbilical cord is attached to the placenta. The placenta, which formed on the uterine wall after the first week of pregnancy, consists of tissues from both the mother and the embryo. The function of the umbilical cord is to carry away the baby's waste products and pass food, energy, and oxygen from the mother's blood stream to the embryo. The mother's Kidney Channels will regulate the re-

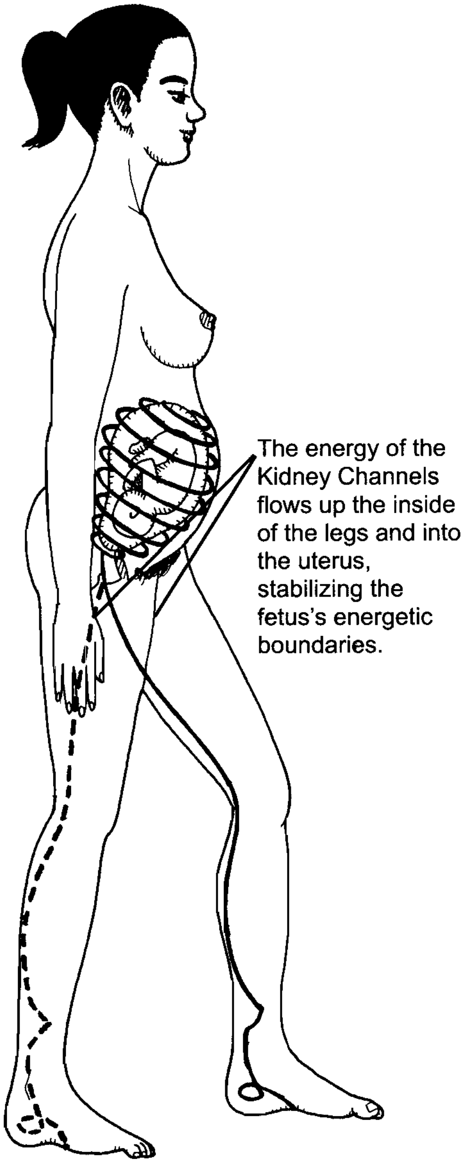

The energy of the Kidney Channels flows up the inside of the legs and into the uterus, stabilizing the fetus's energetic boundaries.

Figure 2.19. The mother's Kidney Channels are responsible for the ninth lunar month of creation. All the fetus's spatial cavities and energetic boundaries are now established.

lease and absorption of Qi and Shen flowing into the fetus's Liver through the umbilical cord.

After the umbilical cord has been severed, the baby's umbilical veins still remain. These umbilical veins eventually become the ligamentum teres that connect from the umbilicus, up along the interior surface of the abdominal wall, through the free margin of the falciform ligament, to the right and left lobes of the Liver. This maintains the baby's connection between its Liver and its Lower Dantian (navel).

At this stage in development, all the fetus's energetic spatial cavities (internal organ tissue chambers) and energetic boundaries are arranged to keep the fetus safely prepared for its birth journey.

Also in this month, the Internal Palaces and Nine Dantian Chambers (the nine internal cavities established within the fetus's Three Dantians (the body's three energetic reservoirs) are arranged and securely established to keep the fetus's Jing-Essence safe.

In the ninth month, additional subcutaneous fat accumulates throughout the fetus's body. Externally, the fetus's fingernails will extend to the tips of the fingers and sometimes beyond.

MONTH TEN

At the tenth lunar month, the mother's Bladder Channels control all Five Yin Organs (Liver, Heart, Spleen, Lungs, and Kidneys) and Five Yang Organs (Gall Bladder, Small Intestine, Stomach, Large Intestine, and Bladder) (Figure 2.20). The baby's energetic reservoirs (the Three Dantians)

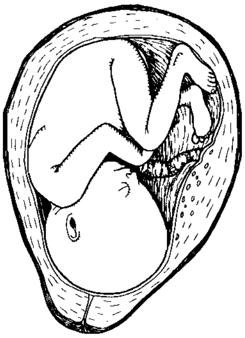

Figure 2.21. Heaven and Earth Qi settle into the baby's Lower Dantian and the process of birth begins.

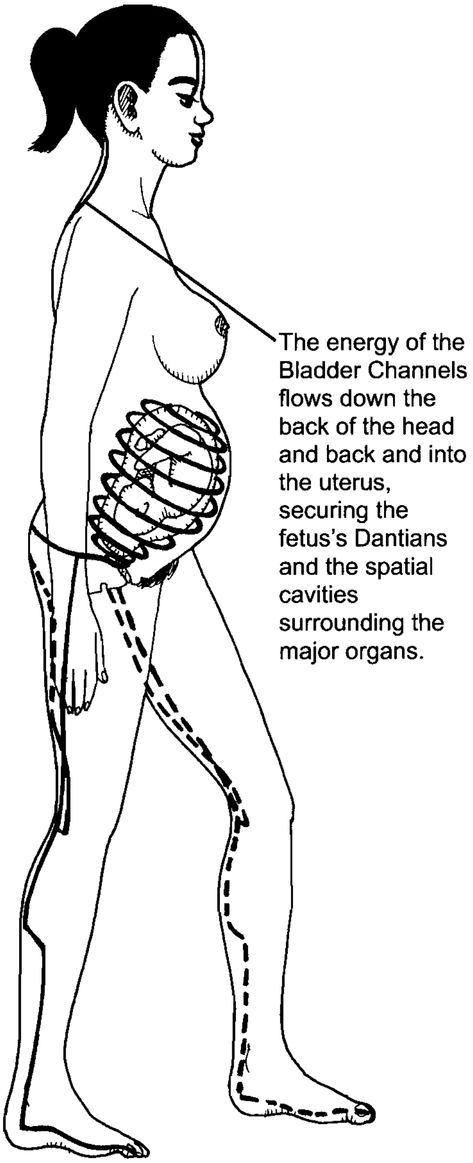

The energy of the Bladder Channels flows down the back of the head and back and into the uterus, securing the fetus's Dantians and the spatial cavities surrounding the major organs.

Figure 2.20. The mother's Bladder Channels are responsible for the tenth lunar month of creation. The baby's Dantians, and spatial cavities surrounding the major organs are completely developed to maintain the safety of the baby's Jing. Heaven and Earth Qi settle into the baby's Lower Dantian and birth begins.

and the internal spatial cavities that surround the major organs are developed to maintain the safety of the baby's Jing. The child is ready to be released into the world. The process of birth begins when Heaven Qi and Earth Qi settle into the baby's Lower Dantian (Figure 2.21).

The energy of the Seven Corporeal Souls (Po) is responsible for the first physiological processes after birth, allowing the child's eyes to see, ears to hear, and Heart to perceive. The Po are also responsible for the movements of the hands and feet and the breathing pattern.

The purpose for the Qigong doctor's study of the ten lunar month developmental process, is to create a foundation for understanding the patient's physical development. This knowledge of the patient's physical development establishes a comprehension of the structural formation of the Five Elemental Constitutions. The Five Elemental Constitutions are described in the next chapter.

POSTNATAL ENERGY DEVELOPMENT

Throughout prenatal development, the Eight Extraordinary Vessels (see Chapter 7) were responsible for transporting, transforming, and producing Qi and Blood for the fetus, while the Twelve Primary Channels (see Chapter 6) were still in the process of gradual development. The focus of energetic activity generated from the fetus's Lower Dantian and Eight Extraordinary Vessels, resonates throughout the fetus's body creating tissue development.

Once the umbilical cord is severed, the primary focus of energy shifts, becoming diverted from the Lower Dantian to the Middle Dantian (located in the baby's chest area). After birth, Gu Qi (food energy) flows into the baby's Stomach while it nurses and begins the movement of Blood and Qi within the Twelve Primary Channels. These Twelve Primary Channels will now assume the responsibility of circulating Qi and Blood throughout the baby's entire physical and energetic structure. The Eight Extraordinary Vessels shift their function to regulating the baby's channel Qi. The Taiji Pole (which has been in the process of dropping from the fetus's Mingmen area to the Huiyin area) is now stabilized in the perineum.

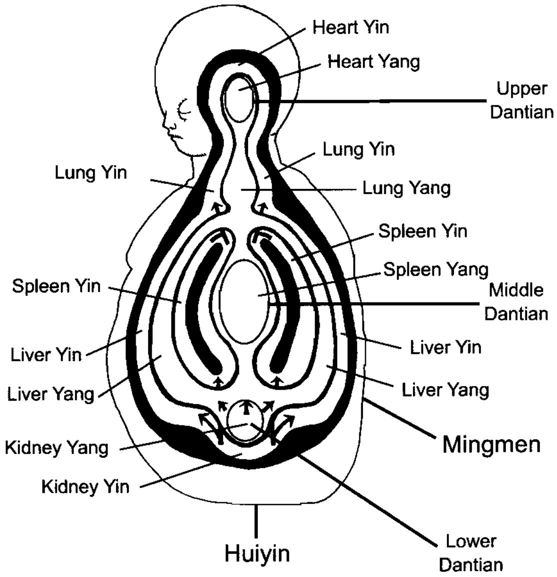

Figure 2.22. As the child develops, the energetic rivers and pools regulate the formation and production of his or her growth.

The natural resonant vibration of the baby's Taiji Pole (energetic core) causes an energetic expansion and contraction, projection and reception that simultaneously affects the baby at five distinct levels: physical, mental, emotional, energetic and spiritual. Balanced physical growth will only occur when all five levels are in equilibrium. The physical body is generally the slowest realm to respond to physical growth. Matter does not develop at the same rate as energy, mind, or spirit. This being the case, the energy, mind, and spirit must wait patiently for the physical body to evolve before progressing as a whole (Figure 2.22).

ENERGETIC EMBRYOLOGICAL OVERVIEW

Understanding the energetic process of fetal development provides the Qigong doctor with an overview of the energetic factors in tissues, organs, and channel function. This understanding is essential for diagnosing the origin and causes of disease as being either congenital or acquired. Unlike Western medicine, Chinese medicine treats the root of the illness, not just its symptoms. In Medical Qigong therapy, cases of congenital insufficiency of Qi or congenital disorders of Qi activities, are treated by either tonifying (strengthening) the Eight Extraordinary Vessels or purging (eliminating) the pathogenic energy from the Eight Extraordinary Vessels. This consolidates and regulates the patient's Prenatal Qi.

For acquired diseases, the patient can only be cured when a selection of points and the methods of emitting Qi for purgation or tonification are determined according to the imbalances of Qi circulation in the Twelve Primary Channels. Both congenital and acquired factors should be considered in every case because both the origination and development of diseases may result from either source, or a combination of both sources.

THE ETERNAL SOUL

In Chinese energetic medicine, the Original Spirit differs from the body's Eternal Soul. The Original Spirit (Yuan Shen) is rooted in the combined Jing, Qi, and Shen of the mother and father. Later the Yuan Shen is integrated and sustained through the embryo's own Five Spiritual Essences or Wu Jing Shen (see Chapter 13), which are created from the energy of the Five Yin Organs. The Eternal Soul, however, is absorbed into the mother's egg at the time of conception, as the sperm enters.

The Eternal Soul is rooted into the Heart and Middle Dantian area by a silver cord, which is additionally rooted within the body's Taiji Pole. When the Eternal Soul becomes rooted within the body's Taiji Pole, it is described as "emitting the spark of the supreme fire" (Light of God), and is considered a divine fragment of God, or the Dao. It is from this divine resonating light that the body's Taiji Pole becomes energized and all Three Dantians become interconnected. The rooting of the Eternal Soul is therefore considered the first emanation from the divine source within the human body.

To travel, and receive spiritual guidance and insight, the Eternal Soul can leave the body in the form of an "Astral Body," departing through the Baihui point at the crown of the head. The silver

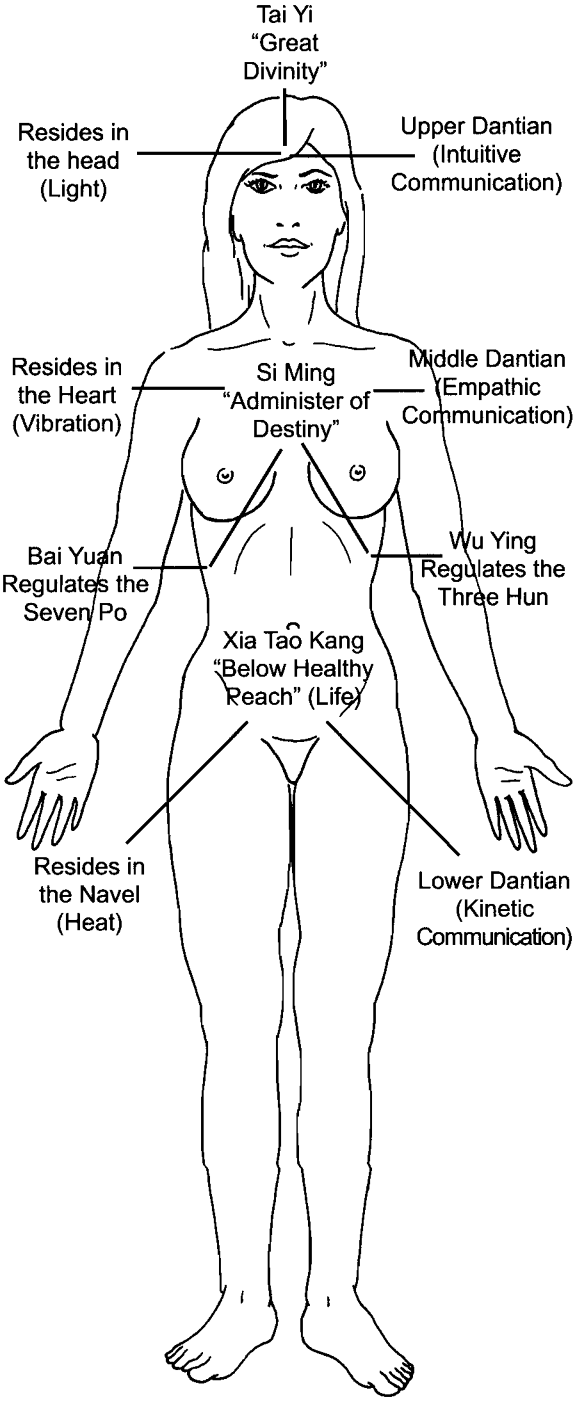
Tai Yi "Great Divinity"
Resides in the head (Light)
Upper Dantian (Intuitive Communication)
Resides in the Heart (Vibration)
Si Ming "Administer of Destiny"
Middle Dantian (Empathic Communication)
Bai Yuan Regulates the Seven Po
Wu Ying Regulates the Three Hun
Xia Taŏ Kang "Below Healthy Reach" (Life)
Resides in the Navel (Heat)
Lower Dantian (Kinetic Communication)

Figure 2.23 The Lords of the Three Dantians: Tai Yi, Si Ming, and Xia Tao Kang

cord, which is connected to the Astral Body, is located at the fifth and sixth thoracic vertebrae. This physical area is the place on the spine where the silver cord lifts off when an individual astral travels. This area also corresponds to the back of the Heart center, between Ling Tai GV-10 (Spirits Platform) and Shen Dao GV-11 (Spirits Gate). The astral travels of the Eternal Soul are generally to one of the nine higher spiritual planes.

The Eternal Soul becomes the intermediator between the divine, the Original Spirit (Yuan Shen), the Acquired Spirit (Zhi Shen), and the body's Soul Extensions (various traits and characteristics). These Soul Extensions manifest in the energetic field, and influence behavior and perception.

Once the Eternal Soul has established its residence, it separates its Yin and Yang spiritual energy into three spiritual energies which are the energizing forces of the body's Three Dantians, and are called Tai Yi, Si Ming, and Xia Tao Kang. These three spiritual energies are referred to as the Lords of the Three Dantians, and are named according to their function: Tai Yi translates to mean Great Divinity, Si Ming translates as The Administrator of Destiny, and Xia Tao Kang translates as Below Healthy Peach (Life).

The Chinese medical terminology, which describes the Eternal Soul as consisting of three spiritual energies called the Lords of the Three Dantians, is extremely metaphoric and is used to describe the many energetic aspects of the human soul.

The Eternal Soul radiates through the physical, energetic, and psycho-emotional domains of human existence (Figure 2.23). Whole-body consciousness is the main characteristic of the Eternal Soul. Without the interactive process of the Eternal Soul, energy would have no specific direction and would remain in meaningless activity. The Eternal Soul radiates energy in all directions throughout the body and is responsible for life and health, as well as the formation and growth of the body.

In the scope of energetic manifestations, the Eternal Soul expresses its innate qualities as energetic movement, functioning through the physical form. It is led by the Shen and is connected to all parts of the body. The light of the Eternal Soul is revealed through the light or energy of the eyes. Through the influence of the Eternal Soul, all of the body's energetic processes seek wholeness. The Eternal Soul knows exactly what is needed in every situation for survival and health preservation.

The distinction between the Eternal Soul and the Yuan Shen is that the Eternal Soul, although spiritual in nature and connected to the Wuji (infinite space), is still considered individuated. The Yuan Shen, however, although individual in its conditioned state, can also be universal. The connection between you and others, for example, is considered a spirit connection (kindred spirits uniting). The connection between you and the divine is considered a soul connection (your true innate nature reconnects with its origin), and relates only to the connection between one's self and the divine.

When the Qigong doctor meditates, the Eternal Soul consciously enters a spiritual realm wherein the doctor perceives the ideal of things rather than the things themselves. The doctor no longer depends on the senses, but upon a clear inner-vision that perceives the whole picture. It is here, within this spiritual state, that the higher knowledge of things unfolds through divine intelligence and all things reveal their true nature.

When conserved and amplified through practice, the energy of the doctor's Eternal Soul reveals a bright radiant quality, which permeates and penetrates the cells, tissues, organs, and external energy fields. These energetic fields create for the Qigong doctor a strong luminous shield of energy, capable of guarding the body against an attack of hostile energy by enabling it to:

- ward off the invasion of External pathogens and harmful electromagnetic fields,
- ward off the negative influences of malevolent spirits,
- ward off ill intentions and spiritual malice projected unconsciously or deliberately by others, and
- ward off the parasitic influences of demonic forces.

At conception, when the sperm enters the ovum, a polar axis that forms the embryo's Taiji Pole is created. The energetic vortex thus created draws spiritual energies into the body from Heaven and Earth. After birth, the degree to which an individual can draw upon, and absorb, spiritual energy depends on his or her karma (previous thoughts, actions and level of consciousness), and course of destiny.

During conception, three spiritual energies are given to each individual from the divine to create and maintain the existence of the Eternal Soul. These three spirit energies, called the Lords of the Three Dantians, reside in the innermost subtle aspect of the body and are described as follows:

1. The Tai Yi (Great Divinity) resides in the head and Upper Dantian, and is considered the Lord of the Nei Wan (Innermost Palace). It governs a multitude of the body's spirits, causing man's Yuan Shen to shine externally. The Tai Yi facilitates awareness of the Three Ethereal Souls (Hun), and advocates for personal spiritual enlightenment.
2. The Si Ming (The Administrator of Destiny) resides in the Heart and Middle Dantian, regulates the body's Qi, and is the source of the mind and its emotional connections. The Si Ming is considered the Eternal Soul (some believe that this soul reincarnates). The Si Ming controls the spirits Wu Ying and Bai Yuan, that directly affect the body's Jing, Qi, and Shen. The Si Ming challenges our reactions to various internal and external obstacles.
    a. The Spirit Wu Ying (Without Excess) occupies the left side of the body, and regulates man's Three Ethereal Souls (Hun), which are called:
        (1) Tai Guang (Eminent Light),
        (2) Shang Ling (Pleasant Soul) and the Five Agents, and
        (3) Yu Jing (Hidden Essence).
    b. The Spirit Bai Yuan (Pure Origin) occupies the right side of the body, regulates the Seven Corporeal Souls (Po), which are called:
        (1) Flying Poison
        (2) Unclean Evil
        (3) Stinking Lungs
        (4) Corpse Dog
        (5) Fallen Arrow
        (6) Yin Bird
        (7) Devouring Robber

Both the Hun (that come from Heaven) and the Po (that come from Earth) are established within the fetus's internal organs at conception but lie dormant until the third month. This is because, at the third lunar month, the fetus's Orbs are sufficiently formed. One belief maintains that the Hun and Po, although residing in the energetic form of the fetus, frequently leave and return to gather and absorb universal and environmental Qi. The Hun will connect with divine spiritual beings, as well as with God. The Po will connect with surrounding environmental spirits.

Note: The Three Ethereal Souls are traditionally referred to as the "Hun" with the assumption that this term depicts all three Hun, thus in Traditional Chinese Medicine the Hun are sometimes referred to in the singular case. The Seven Corporeal Souls are traditionally referred to as the "Po" with the assumption that this term depicts all seven Po, thus in Traditional Chinese Medicine the Po are sometimes also referred to in the singular case.

3. The Xia Tao Kang (Below Healthy Peach / Life) resides in the navel and Lower Dantian and preserves the root of the body's Jing-Essence.

Upon the death of the body, the Hun return to Heaven, the Po return to Earth, and the spiritual energies of Tai Yi and Xia Tao Kang combine with the Eternal Soul stored in the Si Ming (Administrator of Destiny). These three spiritual energies (Tai Yi, Xia Tao Kang, and Si Ming) blend together, combining into one energy that completes the integration of the Eternal Soul. This soul leaves the body and returns back through the tunnel of light to the divine light for judgement, or wanders the Earth becoming a Gui (ghost).

In Western culture, our internal dialogues are sometimes associated with encounters with good or evil spiritual influences. In Chinese medicine the internal dialogues that influence the patient's psyche are also divided into good and evil; these spiritual / emotional influences are considered to be specific good and evil characteristics in man's Hun and Po. The spiritual components and influences of the Hun and Po are considered to be separate entities, or archetypes, acting upon man's Shen, which in turn affects the Eternal Soul.

These archetypes are regarded as spirit souls that can exert a positive or negative influence on a individual's life depending on the nature of the individual's Eternal Soul. The good internal influences manifest through the Hun. The evil internal influences manifest through the Po. The Hun or Po can motivate personal growth, or can hinder it and cause illness, and even the demise of the body.

THE THREE ETHEREAL SOULS (HUN)

The Three Ethereal Souls are the spiritual part of man that ascends to Heaven upon the death of the body. The Three Ethereal Souls are composed of the Three Hun, that originate from Heaven, reside in the Liver, and resonate from the Three Dantians. The Hun have the following associations: light, Yang, Heavenly soul, Shen, as well as positive emotions and feelings (Figure 2.24). The Three Ethereal Souls are the Tai Guang, Shang Ling, and Yu Jing. They are described as follows:

1. The Tai Guang resonates within the Upper Dantian, and is situated in the cranial cavity, just below the Baihui GV-20 (Meeting of Yin) point. This Hun's name means "eminent light," and it is considered the ultimate balance of pure Yin and Yang energy in harmony. It is connected energetically with the Upper Dantian and Heaven, and always strives for physical, mental, emotional, and spiritual purity.
2. The Shang Ling resonates with the Middle Dantian and is situated in the Heart and corresponding vessels and is linked to the body's Five Agents. This Hun's name translates to "pleasant soul," and it is considered changed (or transformed) Yin energy. It is connected with the Middle Dantian and is a soul that is concerned for others. It is associated with the Five Agents, and produces our desire to be involved in a diversity of social interests and responsibilities.
3. The Yu Jing resonates with the Lower Dantian. Translated, the name means "hidden essence." This Hun is considered mixed (or combined) Yin energy. It is connected with the Lower Dantian and is associated with the Earth, producing our desire for enjoying life's pleasures and comforts, as well as the pure passions of life.

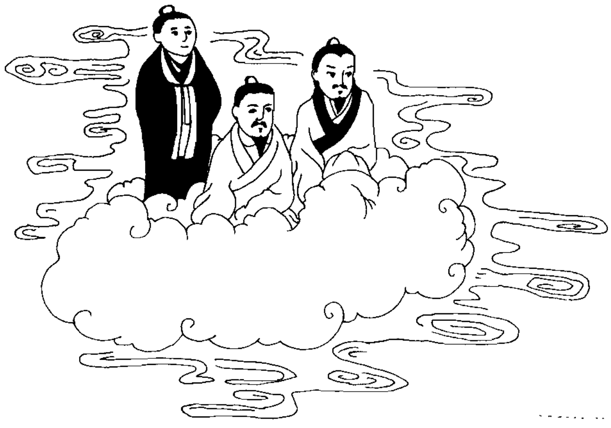

Figure 2.24. The Three Ethereal Souls are composed of the three Hun, that originate from Heaven and reside in the Liver. The Hun represent spiritual consciousness, provide the energetic movement of the mind, and are associated with Heaven Qi and the Five Agents. The Hun are the spiritual part of man which ascends to Heaven after death. The Hun are associated with positive emotions and feelings, light, Yang, Heaven, and Shen.

The Three Ethereal Souls are named (1) Tai Kuang, (2) Shang Ling, and (3) Yu Jing.

The Hun are classified as Yang spirits and are to be cultivated and refined. Imagination, visualization, and positive affirmation in the form of prayer and meditation are needed to awaken and establish an active relationship with the Three Ethereal Souls.

The Hun can be accessed through the Hunmen Bl-47 (Gate to the Hun) point on the back of the body (below the shoulders). It is used clinically to spread Liver Qi and harmonize the Middle

Burner. The Hun respond to Heaven's energetic grids (universal energetic fields). The stars and planets within these Heavenly grids exert an influence on the Hun causing each individual's body to react to certain astrological configurations. The positive or negative reaction is based on the affinity of the vibrational rate of the Hun and the energies of a particular astrological alignment.

THE SEVEN CORPOREAL SOULS (PO)

The Seven Corporeal Souls are closely linked to our body's Jing-Essence. The Po manifest the body's Essence in the form of hearing, sight, and tactile sensations.

The Po pertain to the animal nature of man's instincts and drives and are considered inferior souls to the Hun (Figure 2.25). The Po are passionate and advocate experiencing life in its fullest measure. They are, however, in a constant state of dying. The Po will return to the Earth with the body after death. The Po have the following associations: heavy, Yin, Earthly Spirit, Essence, and negative emotions and feelings.

The Po are a composite of Seven Corporeal Souls, which originate from Earth, reside in the Lungs, and resonate from specific areas in the body. The Po can be accessed through the Pohu Bl-42 (Door to the Po) point on the back of the body (between the shoulders). This point is used clinically to treat energetic imbalances of the Lungs. The Po respond to the Earth's energetic grids (see Chapter 10) and cause each of us to resonate in harmony or disharmony to certain ecological configurations. An individual's attraction, or feeling of not belonging, depends on the vibratory affinity, or lack thereof, between the body's Po and the environmental energy of a particular area.

When the fetus begins its movement, its Yin energy tranquilizes the Po, which act as guardians of the fetus's body. The Po's nature is one of survival, and their energies can be directed towards self-preservation or self-destruction (devouring and robbing the body of life-force energy).

The Seven Po are located along the line between the Huiyin CV-1 and the Baihui GV-20 points, embracing the body's Taiji Pole (also called the Center Thrusting Channel) at the following locations:

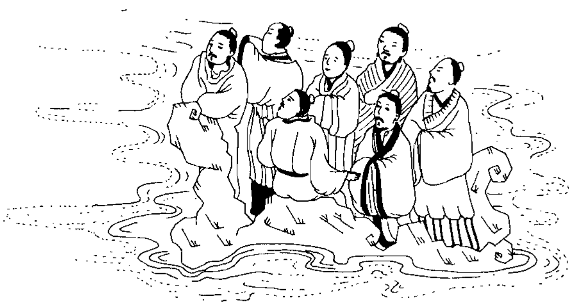

Figure 2.25. The Seven Corporeal Souls are composed of the Seven Po, which originate from Earth and reside in the Lungs. The Po are regarded as the animal nature or drive. They provide energetic movement to the body's Essences, and are considered inferior Souls. The Po will return to the Earth with the body after death, and are associated with negative emotions and feelings, heaviness, Yin, Earth, and Jing.

The negative thoughts and emotions of the Seven Corporeal Souls are named (1) Flying Poison, (2) Unclean Evil, (3) Stinking Lungs, (4) Corpse Dog, (5) Fallen Arrow, (6) Yin Bird, (7) Devouring Robber.

1. The Po of Essence (Soul of the Five Element Earth) is located at the midpoint of the Taiji Pole. This Po is the counterpart to the Hun's Earth Agent (the Virtue of Trust). This Po manifests through the acquired emotions of worry, regret, remorse, obsessiveness, and self-doubt.
2. The Po of Qi (Soul of the Five Element Wood) is located below the diaphragm. This Po is the counterpart to the Hun's Wood Agent (the Virtue of Kindness). This Po manifests through the acquired emotions of anger, irritability, blame, rage, resentment, and jealousy.
3. The Po of the Spirit (Soul of the Five Element Metal) is located posterior of the Tanzhong (CV-17) point, inside the mediastinum, near the Middle Dantian. This Po is the counterpart to the Hun's Metal Agent (the Virtue of Integrity). This Po manifests through the acquired emotions of grief, anxiety, sadness, shame, disappointment and guilt.

4. The Po of Yin (Soul of the Five Element Water) is located posterior of the navel, in front of the Taiji Pole. This Po is the counterpart to the Hun's Water Agent (the Virtue of Wisdom). This Po manifests through the acquired emotions of fear, loneliness, and insecurity.
5. The Po of Yang (Soul of the Five Element Fire) is located anterior of the Mingmen (GV-4) point behind the Taiji Pole. This Po is the counterpart to the Hun's Fire Agent (the Virtue of Order). This Po manifests through the acquired emotions of nervousness, shock and excitement.
6. The Po of Sex (Soul of Heaven) is located below the Baihui (GV-20) point in the Ni Wan (Upper Dantian) area.
7. The Po of Life (Soul of Earth) is the only one not located on the Taiji Pole, but is located at the bottom of the feet in the Yongquan (Kd-1) points and is considered a neighbor of the Soul of Essence.

The Seven Corporeal Souls (Po) are also called the "seven animals," or "sentient souls of the body," and "seven turbid demons." When afflicted or restless, the animal nature of the Seven Po quickly becomes hostile, and their names change to express the different negative thoughts and emotions expressed from each Po soul. Each Po has a favorite mantra that it whispers inside a person's mind to influence the individual. The Po's Seven Turbid Demon Natures are described as follows.

1. The Flying Poison (or Quick Evil) suddenly explodes with rage and venomous thoughts of evil intention (e.g., "I'll kill you for that!" "How dare they!").
2. The Unclean Evil (or Shame) entices by tempting and luring an individual into a place or situation from which escape is difficult, and then creates distress in the form of guilt, which generates shame. This spirit creates a feeling of being discredited, dishonored, or disgraced (e.g., "I'm worthless;" "No one has or will ever love me.").
3. The Stinking Lungs (or The Smell of Death) destroys hope, which can lead to a sense of despair (e.g., "I just want to die;" "Just kill me and get it over with!").
4. The Corpse Dog (or Being Scorned) is like a dog that has been beaten and starved, the presence of this spirit intensifies depression by despising itself and holding itself in contempt to a point of no return (e.g., "I'm not worth keeping alive;" "Leave me alone to die!").
5. The Fallen Arrow generates the foregone conclusion that attempting goals is fruitless (e.g., "I'll never be able to accomplish that." "Why even try? I'll never make it.").
6. The Yin Bird (or Night Tormentor) harasses the individual, causing him or her to experience extreme pain and severe anguish (e.g., nightmares and restless sleep).
7. The Devouring Robber (or Sipping Thief) steals the individual's life-force energy by devouring it through negative emotions such as jealousy, envy, and bitterness (e.g., "You love him/her more than you do me;" "I'll never forgive you!").

The Seven Po help in developing the growth of consciousness by providing obstacles to test the individual's faith and devotion. Many times the Seven Corporeal Souls will work in conjunction with outside spiritually hostile forces to test the individual's spiritual endurance.

After the child has been born, enlightened parents can assist the child in facing his or her own internal demons. Once the child is self-aware, he or she can be taught to restrain and control the Seven Po so as to prevent them from taking control. To minimize the effects of these souls on the body, one can learn special meditations, involving visualization, imagination, positive affirmation, and mantras (literally "protections of the mind" or "mind protectors").

A mantra is a word, phrase, or sound, repeated silently or aloud, that one can consciously use as a tool to guide one's own mental, physical, and emotional states. People naturally repeat both positive (self-affirming) mantras and negative (self-destructive) mantras as part of their internal self-talk. When positive mantras are used by an

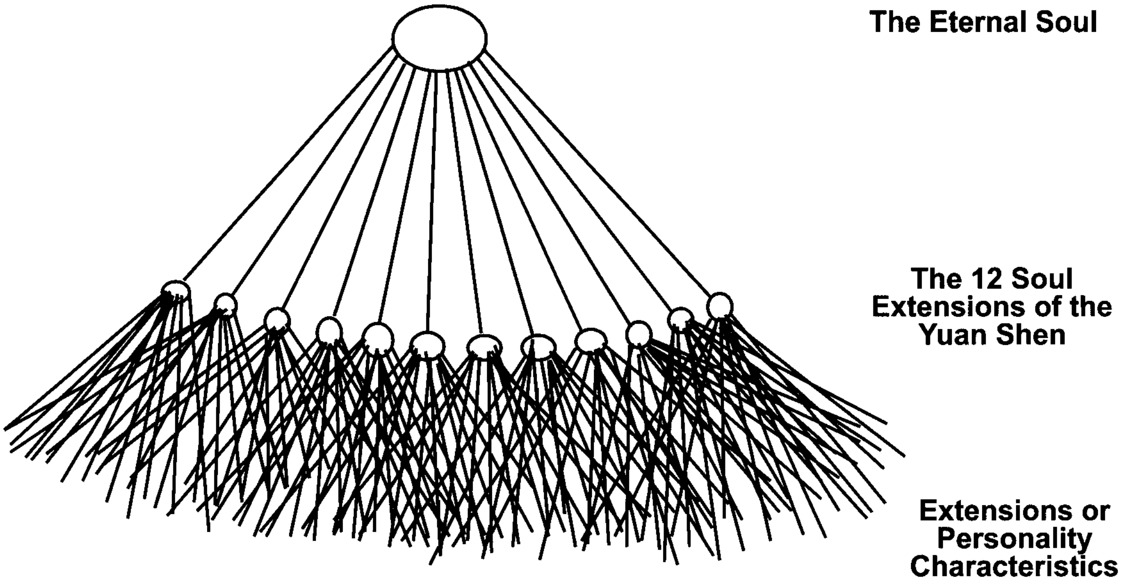
The Eternal Soul
The 12 Soul Extensions of the Yuan Shen
Extensions or Personality Characteristics

Figure 2.26. The Eternal Soul, Yuan Shen and 144 Soul Extensions

individual who holds them in high regard, they enable that individual to combat the Seven Corporeal Souls' mantras. These mantras also facilitate connecting the individual to the divine. Used consciously, mantras and/or prayers can uplift and maintain the individual's self-esteem.

The Chinese view meditation as a deep spiritual contemplation that results in openness to the divine. They view prayer as energy focused towards the divine. Individuals choose what to focus their attention on; therefore, any chronic detrimental thoughts can be countered with positive affirmations to change self-destructive habits.

THE ORIGINAL SPIRIT (YUAN SHEN)

The Original Spirit (Yuan Shen) radiates and extends from the interior of the Eternal Soul. This spiritual light crystallizes a spiritual body, gradually becoming consciously effective and moving into action. It envelops the body and communicates through innermost thoughts, sensory feelings, and emotions. Although the body's Jing and Qi will degenerate, the Original Spirit is immortal, as it is the expression of the Eternal Soul.

The Shen develops and contains Twelve Soul Extensions. These Twelve Soul Extensions contain the body's different personality characteristics (Figure 2.26). Although Si Ming controls the residence of the Eternal Soul, the body's Shen is free to make decisions that affect life and health based upon the individual's free will (the interaction between the Yuan Shen and the Zhi Shen), which then manifests through the Twelve Soul Extensions.

THE EFFECT OF THE HUN AND PO ON THE YUAN SHEN

The Yuan Shen is a manifestation of the Eternal Soul and is primary to the Hun and Po. The combination of the Seven Corporeal Souls (Yin Souls) and the Three Ethereal Souls (Yang Souls) creates the sustaining energy for the body's Original Spirit (Yuan Shen). The Yuan Shen also controls and organizes the psycho-emotional aspect of the body's Five Yin Organs, called the Wu Jing Shen or Five Essence Spirits.

The ancient Chinese viewed the Shen as an Emperor, seated within the Heart; the Hun are viewed as a Loyal Minister, seated within the Liver; and the Po are viewed as a Violent General, seated within the Lungs. If the General (Po), is left in control (being only concerned with the survival of self and the body), he will start to dominate (a condition referred to as a rebellious General and a weak Emperor). At the point when the General (Po) dominates, the individual's acquired mind takes over and the individual becomes concerned only with his or her own survival, and thus becomes self-absorbed. The goal for spiritual cultivation is to control the rebellious General, making him into a servant. Once the inner government is orderly, the strong and violent nature becomes tame. Then, through the wise council of the Loyal Minister (Hun), the individual can walk a path of virtue. Virtue is the path that leads the individual's Shen so that even his or her human consciousness is dominated by the Hun and the Yuan Shen.

When the Hun control the energy body and are nourished by the virtues, the energy body then becomes a vehicle for the Heart's Shen, which is non-local and non-linear in terms of time and space.

When the Heart's Shen is no longer dominated by the Yin and Yang souls (Hun and Po) and the Five Element energies of the lower energetic plane (Wu Jing Shen), the individual returns to an awareness of his or her connection to the universe (Wuji) and the divine. This state of consciousness is sustained through prayer and meditation.

When beginning to develop the Heart's Shen, the first separation of the individual's energetic spirit body or "dream body" from the physical body generally leaves the Shen in a weakened state (referred to as Yin Shen). The weakened Yin Shen must be protected. This Yin Shen is part of the Hun, and leaves the body naturally whenever the patient is weak, sick, in shock, or asleep. It may also leave during the early stages of Qigong meditation or Taijiquan practice. The astral travels of the Yin Shen are generally confined to the lower spiritual planes, and it sometimes needs to be reclaimed through "soul retrieval" (see Chapter 19). Once the Shen has been cultivated, refined, strengthened, fortified, and controlled through the development of the energy body, it is then referred to as a Yang Shen. When the energy body spiritually travels, the developed Yang Shen can transverse the Nine Levels of Heaven or the Nine Levels of Earth in an instant.

The Nine Levels of Heaven are nine spiritual planes which exist within Heaven's energetic grids. Within the lower levels there exists enlightened beings coexisting with other spiritual entities. The Nine Levels of Earth are nine spiritual planes which exist within Earth's energetic grids. Within these levels of Earthly spiritual planes there exists various dimensions of animal, vegetable and mineral powers, as well as Earthly spiritual beings.

YIN AND YANG ASPECTS OF THE ETERNAL SOUL

The human soul has an endless source of vitality from the infinite high-frequency force of the divine. This divine energy descends from the Heavens into the body's Taiji Pole, taking up residence and interacting on the physical plane. The body serves as a medium of transformation for the Eternal Soul. The soul's activity extends and expresses its essence through the physical form and through consciousness. As we grow and change, the Eternal Soul maintains stability by providing the energetic blueprint for orderly development. On the spiritual plane, when referring to the energetic dynamics of the Eternal Soul, there are two basic movements: (1) Yin - the downward movement and (2) Yang - the upward movement.

1. The Yin aspects of the Eternal Soul spirals downward. The spiralling movement of the energetic soul results in the energy transforming from the highest divine energetic frequencies to the lowest. Within this spiritual transformation process, there is a parallel gradual decrease in levels of consciousness. The lowest frequency creates matter, and is the bases for all the prenatal transformations. At the higher frequencies, the principles by which Qi operates are difficult to perceive because our lower state of consciousness does not resonate in harmony with the information being

Spiritual Energy	Resides in...	Associated Dantian	Purpose	
Tai Yi Great Divinity	the head as light	Upper Dantian (intuitive communication)	Governs the body's Spirits, makes life shine forth, and encourages man to know the spirits of his Ethereal Soul	
Si Ming Administrator of Destiny	the Heart as vibration	Middle Dantian (empathic communication)	The Eternal Soul—regulates the prime energy of life (Qi) and is the source of the emotions and mind	Wu Ying (Without Essence), occupies the left side of the body and regulates the Three Hun (Ethereal Souls), associated with the Liver
			Controls the Wu Ying and Bai Yuan Spirits	Bai Yuan (Pure Origin) occupies the right side of the body and regulates the Seven Po (Corporeal Souls), associated with the Lungs
Xia Tao Kang Below Healthy Peach (life)	the navel as heat	Lower Dantian (kinetic communication)	Preserves the root of the body's Essence (Jing)	

Figure 2.27. The patient's soul resides in the body's Taiji Pole, connected with Si Ming, the Administrator of Destiny. Upon the death of the physical body, the Three Hun return to Heaven, the Seven Po return to Earth, and the energy of Tai Yi and Xia Tao Kang envelope the Eternal Soul (stored in Si Ming) becoming one energy. The Eternal Soul leaves the body through the specific gate (solar plexus, third eye, or top of the head) associated with the patient's degree of spiritual evolution, returning back through the tunnel of light to the divine.

presented. As energy slows down its resonance, it is more readily perceived by our five senses; this gives matter the quality of hardness, form, and smell, and allows us to observe its energetic patterning.

2. The Yang aspects of the Eternal Soul spirals upward. The spiraling movement of the energetic soul results in energy transforming from the lowest level of consciousness to the highest divine energetic frequencies during postnatal transformation. Spiritual evolution involves a progressive opening up of all of the body's energy centers (Three Dantians) along the Taiji Pole, until a state of absolute unity of body, mind, emotion, energy, and spirit is achieved (known as enlightenment).

In the evolutionary process of the soul, both stability and change become as one when there is a deep connection to wholeness. This soul serves as a pivotal point for the body's Jing, Qi, Shen, Body Fluids, connective tissues, and emotions, all of which express wholeness. Without this connection the patient experiences a serious psychic splitting of his or her energies. The part that is open to change will experience resistance from the part that demands stability.

THE FINAL EXIT OF THE ETERNAL SOUL

As the body begins to die, the Yuan Shen begins to prepare the Eternal Soul for its final journey home. This change of residence for the patient's spiritual energies is crucial. At the time of death the Eternal Soul must disconnect from the Heart and Middle Dantian area, and the soul must exit the dying patient through one of the Nine Orifices. The lower orifices include the anus and urethra, while the upper orifices include the eyes, ears, nostrils, mouth, as well as the Yintang and Baihui areas (Figure 2.27). It is believed that the area from which the patient's soul exits the body determines the state of its spiritual evolution and to which area the patient is most attached (i.e., if the patient is attached to vision, they are more inclined to leave through the eyes, etc.).

When the Medical Qigong doctor is assisting a terminally ill patient, treatment focuses primarily on purging stagnant spiritual energy from the patient's body. Spiritual purging is required to purify and cleanse the patient's Eternal Soul. Once the patient's soul has been cleansed, the patient becomes peaceful as all material, emotional, and spiritual attachments to this life are released. The patient attains a sense of completion and welcomes the final adventure of going home. The last rites performed by various clergy and ministers in Western culture serve a similar purpose.

The Qigong doctor then guides and encourages the patient's Eternal Soul into leaving the body through the top of the head (Baihui). Souls that are more highly evolved exit the body through this area (at the top of the head), while souls of lesser evolution exit the body through lower portions of the physical structure.

The formation of a Gui or ghost is said to derive from the untransformed energy of the Eternal Soul (see Chapter 19).

SUMMARY

One of the unique aspects of Traditional Chinese Medicine is that it addresses all aspects of the body's energy and spirit. It explores in great detail the spiritual transformations in the body's prenatal state, their relationship to the formation of the body, and to the dying and after-death states. Traditional Chinese Medicine further addresses the potential of spiritual involvement as a cause or cure of specific psychophysical diseases. It is through the complete study of the body's physical, mental, emotional, energetic, and spiritual aspects that doctors of Medical Qigong therapy comprehend when and how to treat their patients.

CHAPTER 3

PHYSICAL DEVELOPMENT AND STRUCTURAL FORMATION

FASCIAL DEVELOPMENT AND ENERGY FLOW

According to the *American Medical Association Encyclopedia of Medicine*, "fascia is the fibrous connective tissue that surrounds many structures in the body. One layer of the tissue, known as the superficial fascia, envelopes the entire body just beneath the skin. Another layer, the deep fascia, encloses muscles, forming a sheath for individual muscles, and also separates them into groups. The deep fascia also holds in place the soft organs, such as the Kidneys. The thick fascia in the palm of the hand and sole of the foot have a cushioning, protective function."

In Traditional Chinese Medicine, the fascia is referred to as Huang (meaning any membranous tissue). Fascial development is divided into two stages: prenatal (congenital) and postnatal (acquired). The prenatal fascia is fixed and determined by the combination of the parents' Jing, Qi, and Shen. The postnatal fascia is formed through diet, exercise, and environmental stresses.

Qi is stored within the tissues and inner fascia layers, where it envelops and protects the internal organs. Energy (that has been stored within the body's organs) is available through the fascia for transformation to nourish the patient's body, mind, emotions, energy, and spirit.

The body consists of three tissue layers:

1. The first and innermost layer is made up of the internal organs, which produce and transform Qi.
2. The second layer consists of the body's fascia, tendons, ligaments, and bones, which assist in transporting Qi to the extremities.
3. The third and outermost layer of tissue consists of the muscles and skin, where Wei (Protective) Qi circulates to protect the organism from invasion of external pathogens (Cold, Heat, Wind, and Dampness, etc.).

After being generated within the body's internal organs, Qi is distributed throughout the body's entire energetic network through the channels and collaterals, via the fascia. Each organ has its own layer of weblike fascia, that covers, connects, protects, and nourishes the tissues. The fascia forms the energetic chambers of the body's organs and channel systems. Qi within the body flows between the fascial sheaths and along the channel system. Through trauma, infection, surgery, disease, or chronic muscular tension, areas of the fascia can become stuck together, inhibiting the flow of Qi and Blood.

Our constitutions are formed, in part, by conditions from the environment that affect the innermost layers of fascial development. Traditional Chinese Medicine divides the observation and diagnosis of the physical developments of these inner fascia into Yin and Yang structures and Five Elemental Constitutions.

YIN AND YANG STRUCTURAL FORMATION

Chinese philosophy teaches that the universe is composed of a pair of opposite forces or energies -- Yin and Yang. Everything in the natural world contains both Yin and Yang. Yin and Yang are interdependent; without Yang, Yin cannot grow; without Yin, Yang cannot develop. The energetic transformation of Yin into Yang and Yang into Yin produces the various observable changes of form and matter (see Chapter 21).

Yin and Yang are the principles governing all things within the human body. Life and death originate from the energies of Yin and Yang, and they are the forces that create all physiological change. In Medical Qigong therapy, the physical structure of the human body is divided into Yin and Yang organs and substances, as well as Yin and Yang energetic functions (e.g., Qi that expands and flows outward is Yang, while Qi that contracts and flows inward is Yin).

Yin and Yang are in a constant state of waxing and waning (Figure 3.1). If this waxing and waning exceeds the body's normal energetic limits and loses its dynamic equilibrium, Deficient or Excess Yin and Yang will occur, leading to the development of abnormalities and illness.

Yin Qi naturally ascends (e.g., Spleen Qi ascends the Clear Qi). When Yin Qi becomes pathological or destructive, it moves downward, descending like cascading water (e.g., descending Spleen Qi causes diarrhea or prolapse of the viscera). Another example of the effects of pathogenic Yin Qi descending would be edema.

Yang Qi naturally descends (e.g., Stomach Qi flows downward). When Yang Qi becomes pathological or destructive, it ascends like the flames of a fire, moving upward (e.g., Rebellious Stomach Qi ascends, causing nausea and vomiting). An-

	Yang	Yin
Attributes	active, creative, masculine, Fire, Hot, Heaven, light	passive, receptive, feminine, Water, Cold, Earth, heavy
Organs	the six hollow organs (Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder, and Triple Burners)	the five solid organs (Liver, Heart, Spleen, Lungs, and Kidneys), along with the Pericardium
Respiration	exhalation	inhalation
Time of Day	the 12 hour cycle from 11 p.m. (before midnight) to 11 a.m. (before high noon)	the 12 hour cycle from 11 a.m. (before noon) to 11 p.m. (before midnight)
Seasons	spring and summer	autumn and winter
Normal Movement	outward	inward

Figure 3.1 Yin and Yang are the Principles Governing All Things Within the Body

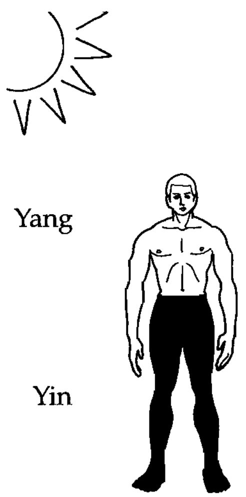

Figure 3.2. The superior aspect of the body's structure is Yang; the inferior portion is Yin. Think of the sun shining on the upper torso, illuminating the body from the waist to the head, while from the hips downward the lower torso is in the shade.

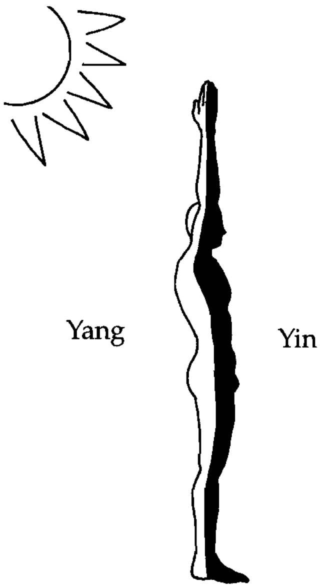

Figure 3.3. The posterior portion (back side) of the body is Yang; the anterior portion (front side) is Yin. Think of the sun shining on the back, while the abdomen is in the shade.

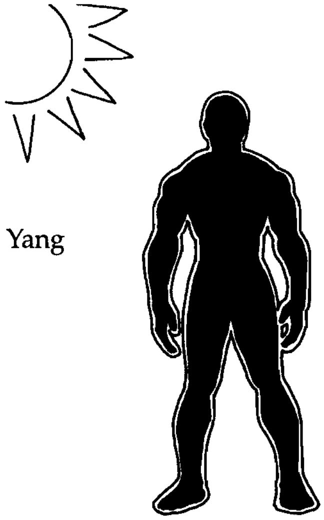

Figure 3.4. The superficial aspect (exterior) of the body is Yang; the deep portion (interior) of the body is Yin. Think of the sun shining on the external surfaces of the body, while the internal organs remain in the dark.

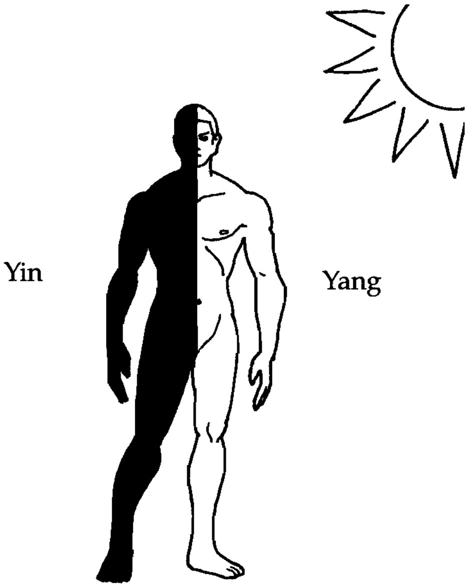

Figure 3.5. The left side of the body is Yang; the right side is Yin. Think of facing south in the morning, while performing medical Qigong practice. The sun will rise in the east, illuminating the left side (Yang) while the right side remains shaded (Yin).

other example of the effects of pathogenic Yang Qi ascending would be the temperature of a rising fever.

THE YIN AND YANG OF BODY ASPECTS

Comparable to Western anatomy, Traditional Chinese Medicine also divides the body into aspects or planes. In TCM, these classifications are divided into Yin and Yang physical planes and anatomical directions. These divisions assist the Qigong doctor in defining and categorizing the patient's external tissues, and in determining the collection and movement of Qi (i.e., energy moving into and coalescing in the patient's Yin or Yang areas of the body). The structural aspects are categorized as follows.

- The cranial or superior aspect (towards the head) of the body's structure is considered Yang; the caudal or inferior portion of the body (towards the feet) is considered Yin (Figure 3.2).
- The posterior or dorsal portion (back side) of the body is Yang; the anterior or ventral portion (front side) is Yin (Figure 3.3).
- The superficial aspect (exterior) of the body is Yang; the deep portion (interior) of the body is Yin (Figure 3.4).
- The left side of the body is Yang; the right side is Yin (Figure 3.5).

Aspect	Yang	Yin
upper/lower surface	superior (upper)	inferior (lower)
front/rear surfaces	posterior (back)	anterior (front)
position from midline/middle	lateral (away from center)	medial (middle)
near to/far from the surface	superficial (external)	deep (internal)
sides	left	right

Figure 3.6. A Summery Diagram of the Yin and Yang Aspects of the Body

The Seven Emotions			
Chinese Word	English Translation	Organ Association	Qi Manifestation
Xi	joy, happiness, excitement, pleasure	Heart	slows down and relaxes the Qi
Nu	anger, irritation	Liver	causes Qi to rise
Si	meditation, contemplation, pensiveness, worry	Spleen	stagnates the Qi
You	anxiety, sorrow, grief	Lungs	obstructs the Qi
Kong	fear, extreme anxiety	Kidneys	causes Qi to descend
Jing	shock (sudden intense fear), fright	Heart then Kidneys	scatters and deranges the Qi
Bei	grief, affliction, sadness	Lungs	disperses and consumes the Qi

Figure 3.7. The Seven Emotions and Their Effects on the Organs

- The lateral aspect (further from the center) of the body is Yang; the medial portion (middle) is Yin.

The aspects are summarized in Figure 3.6.

As Yin and Yang are in a constant state of change, always waxing and waning, two points of reference alone are not enough to aptly classify the many phases and stages of transformation; for example, day (Yang) can be further split into dawn (waxing Yang), midday (peaking Yang), and late afternoon (waning Yang). To fill this descriptive gap, the Chinese developed the theory of the Five Elements.

THE FIVE ELEMENTS

The Five Elements (Wu Xing), also translated as the Five Phases, are Wood, Fire, Earth, Metal, and Water. These Five Elements paint a clear and natural image of the changing cycles of Yin and Yang (see Chapter 23).

Furthermore, each of the Five Elements can also be divided into Yin and Yang, making a total of ten subdivisions. Wood, for example, can be subdivided into Yin Wood and Yang Wood.

The Chinese use this framework of the Five Elements to understand and categorize many different areas of knowledge, from the movements of the Heavens and the land forms of Earth to the workings of human anatomy, physiology, and psychology. Earlier in this chapter, it was noted that our organs and their associated energy channels are divided into Yin and Yang. In addition, each of the organs is also related to one of the Five Elements. The Liver and Gall Bladder relate to the Wood Element; the Heart and Small Intestine to the Fire Element; the Spleen and Stomach to the Earth Element; the Lungs and Large Intestine to the Metal Element; and the Kidneys and Bladder to the Water Element. The organs help to regulate the functional aspects of the Five Elements within the human body as they interact with each other to promote and maintain life.

The Yin organs, in particular, are said to store the Jing of the Five Elements. They share this Essence with their elementally paired Yang organs to assist with the Yang organs' functions. The Yin organs also store the Five Agents (the Five Virtues), and are more readily disrupted by the Seven Emotions (anger, joy, worry, grief, sadness, fear, and shock) than are the Yang organs.

The emotions are said to be the primal force behind energy transformation; however, each individual will have a separate reaction to specific emotional changes, based on his or her dominant elemental pattern. Five internal organ emotions are associated with the energy of the Five Elements:

1. Anger - Liver - Wood
2. Joy - Heart - Fire
3. Worry - Spleen - Earth
4. Grief and Sorrow - Lungs - Metal
5. Fear - Kidneys - Water

In this way, the energies of the Five Elements shape all aspects of a human being: physical, energetic, mental, emotional and spiritual (Figure 3.7).

CONGENITAL CONSTITUTIONS

Most individuals are born with a dominant element, which will determine the individual's physical, energetic, and psycho-emotional constitution. When in Excess, the body's emotions can create an energetic imbalance, which in turn, can cause a destructive physical pattern within the body. An example of this process is excessive anger leading to hypertension.

The previous chapter on energetic embryology described how, from month four to month eight, the different Five Elemental Jings (Essences) enter the fetus from each of the mother's organs. Depending on the relative strength or weakness of the mother's internal organ energies during her pregnancy, the fetus will inherit a greater or lesser degree of elemental Jing from the mother. If the mother's Liver energy is dominant, then the child will have a tendency to develop a Wood constitution; if the Kidney energy is dominant, then the child develops a Water constitution, and so forth.

In addition, traditional Chinese astrology states that the various Heavenly influences at the time of birth influence the relative Five Elemental strengths and weaknesses in the newborn's constitution. Years, months, days, and hours all occur in separate cycles of five, corresponding to the Five Elements. Thus both the astrological influences at birth as well as the prenatal maternal transmission of Jing, Qi, and Shen, determine each person's congenital constitution.

After birth, the congenital Jing is cultivated and sustained through prayer, meditation, physical exercise, and sleep. In medical Qigong theory, the health of the congenital Jing determines the health and balance of the overall constitution; it determines the level of vitality and resistance to disease, as well as the mental and emotional nature of the individual. The congenital Jing, with its unique balance of the Five Elements, becomes the foundation for an individual's Qi and Shen.

The congenital influence is, however, not absolute. Postnatal factors, such as diet, life-style, environmental factors, and internal belief structures, can also influence the constitution, for better or for worse. Thus, patients with congenital deficiencies can still improve their health by prayer, meditation, medical Qigong, adequate sleep, proper diet and exercise, herbs, medicines, stress management, and other therapeutic modalities.

On the negative side, patients with strong inherited constitutions may develop serious constitutional imbalances through intemperate lifestyles such as: excess stress, overwork, excessive sexual activity, and poor eating habits. Therefore, the Qigong doctor must consider both prenatal (congenital) and postnatal (acquired) constitutional factors to understand and differentiate between the patient's innate constitution and his or her current condition.

CLASSIFICATION OF THE FIVE CONSTITUTIONS

Over the centuries, much has been written in Chinese medical literature about the Five Elemental Constitutions. The following interpretation was developed by Dr. Zhou Qianchuan, a Daoist master from Qi Cheng Shan and reputed by some to be the master of *The Yellow Emperor's Classic of Internal Medicine* (a 2500 year old three-volume text that is still one of the pivotal texts of Chinese medicine).

Dr. Zhou perceived that each of the Five Elemental Constitutions has both a characteristic physical appearance and a psycho-emotional profile. Through observing these aspects, the Qigong doctor determines a patient's Five Elemental Constitution, and from this diagnosis, the doctor discovers much about the balance or imbalance of the patient's physical and mental health.

In Traditional Chinese Medicine, somatic-build type and psyche are closely intertwined. Physical aspects, such as coloring, proportion, dominant tissue (muscle, fat, sinew, etc.), tone, movement, posture and holding patterns, as well as general vitality all express the patient's internal energetic and psycho-emotional matrices. The body type reveals to the trained Qigong doctor not only the patient's emotional history but also his or her innate personality and acquired personality characteristics.

Dr. Zhou subdivided each of the Five Elemental Constitutions into three categories: balanced nature, Yin nature, or Yang nature. The ideal is to have a balanced nature, regardless of one's elemental type. Note that the treatment addresses the mental, emotional, and spiritual balance of the five constitutions. Since the mind is so closely tied with Jing and Qi, it is not enough to simply tell patients how to change their personality to restore balance. The Qigong doctor must work with patients on all levels (physical, energetic, mental, emotional, and spiritual) to help them restore balance. Specific healing methods are discussed at length in later chapters. In general, the Qigong doctor selects treatment principles based on the Five Elements' Creative (Figure 3.8) and Controlling (Figure 3.9) Cycles to help restore balance (see Chapter 23).

For instance, a patient with a Yin Earth nature would tend to have weak or Deficient Spleen energy. Therapeutically, the Qigong doctor would choose one of the three following treatments.

1. Direct tonification of the Earth Element,
2. Tonification of the Fire Element in accordance with the principle, "To nourish the child, strengthen the mother"), or
3. Disperse or reduce the Wood Element (Wood, the "grandmother", may be overcontrolling Earth, the "grandson").

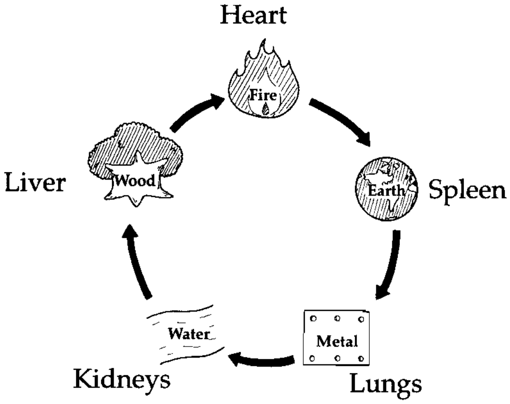

Figure 3.8 The Five Elements' Creative Cycle

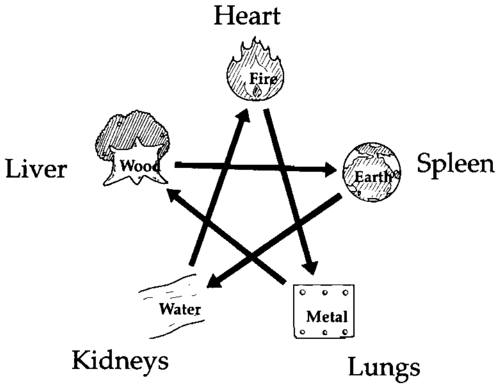

Figure 3.9. The Five Elements' Controlling Cycle

In the instance of Yang Earth nature, the patient would tend to have Excess Earth. In this case, the Qigong doctor could:

1. Use a method to reduce the Earth Element directly; or,
2. Strengthen the Wood Element (the weak "grandmother" may be unable to restrain an overly strong "grandson" without help).

THE WOOD CONSTITUTION

WOOD OUTER APPEARANCE

Wood constitution types usually have a tall and slender body with a small head and a long face. They also have small, nicely formed hands and feet, broad shoulders, and a straight, flat back (Figure 3.10). By nature, they enjoy spring and summer but dislike autumn and winter. This preference is due to the vulnerability of the Wood constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong sinews and green-blue facial color. Their note is "jue."

WOOD PERSONALITY

The Wood personality can be divided into three subdivisions.

1. A Balanced Wood nature predisposes individuals to be confident, strong, independent, and intuitive, with a clear understanding of themselves and their goals. They are patient, able to allow things to develop naturally, and express their personality in a relaxed harmonious way. They are kind when communicating with others, creative, free-flowing in self-expression, and display merciful and unselfish traits when dealing with the needs of others.
2. A Yin Wood nature predisposes individuals to feel externally insecure, and to be always cautious. They have a tendency to worry and have a weak sense of their own abilities and potential. They are unassertive and are unsure of their identities and life purpose. They also have difficulty expressing their egos and have weak boundaries. They are timid, lack confidence, and display considerable doubt.

   Treatment for Yin Wood nature individuals involves establishing a sense of inner strength and security by teaching them self-trust and building self-esteem. These individuals need to strengthen their boundaries to avoid the intrusion and domination of others. They need to trust in their intuition, develop a greater degree of confidence, enhance their personal power, and find a stronger sense of spiritual growth.
3. A Yang Wood nature predisposes individuals to manifest their internal insecurity through acting irritable and impatient. Such individuals are intolerant, rude, stubborn, and selfish, and tend to expand their egos without consideration for others. They are domineering, angry, aggressive, and generally known as

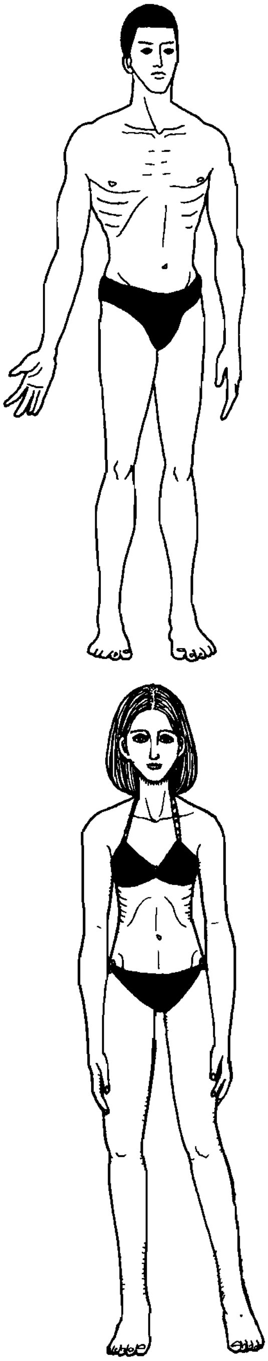

Figure 3.10. The Wood Constitution in Males and Females

overachievers; however, they are easily frustrated and depressed. Their self-esteem is elevated by acquiring higher political (influential) positions or bullying others.

Treatment for Yang Wood nature individuals should focus on cultivating a discipline of inner peace, following the Dao, to harmonize with life. In situations of conflict, they need to learn to relax, slow down, and act out of stillness. They need to surrender to a higher will, relying on their spiritual intuition and inner direction. They also need to learn to respect others.

THE FIRE CONSTITUTION

FIRE OUTER APPEARANCE

Fire constitution types usually have broad paravertebral muscles and well-proportioned shoulders, upper back, and thighs. They have small pointed heads, pointed chins, small hands, and feet, and curly, or no hair, on top of their head (Figure 3.11). By nature, they enjoy spring and summer but dislike autumn and winter. This preference is due to the vulnerability of the Fire constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include a strong circulatory system and a red facial color. Their note is "zhi."

FIRE PERSONALITY

The Fire personality can be divided into three subdivisions.

1. A Balanced Fire nature predisposes individuals to be trusting, open-minded, complacent, social, unconcerned about wealth, and fond of beauty. They love themselves and others, and are very expressive of their affection. They are calm, peaceful, happy, lively, spontaneous, funny, and fun to be with.
2. A Yin Fire nature predisposes individuals to be solemn and depressed. They tend to lack interest in life, and have a tendency to become isolated, feeling unloved and unlovable.

Treatment for Yin-Fire nature individuals requires teaching them how to store and conserve their energy, to use moderation in what

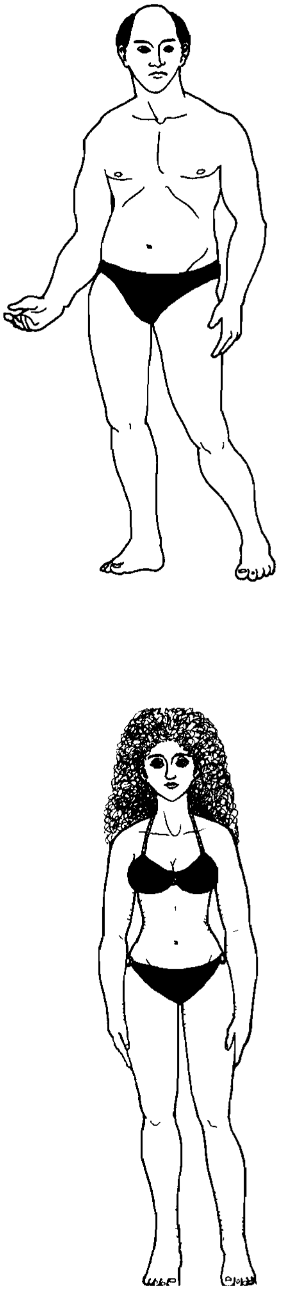

Figure 3.11. The Fire Constitution in Males and Females

they do, and to avoid extremes. They also need to engage in activities that create personal enjoyment, to find simple pleasures that awaken their affection, and to learn how to express their feelings, wants, and needs.

3. A Yang Fire nature predisposes individuals to be arrogant, ignorant, and troublesome. They are restless, excitable, and talk excessively. They always exaggerate and are overenthusiastic. They are socially and sexually overactive and seek every opportunity to assert themselves. They are foolish and display careless behavior. They are overconfident, slightly lazy, irresponsible, and less than truthful. They can be manic, get exhausted, then burnout and become suicidal (e.g., bipolar disorder).

   Treatment for Yang-Fire nature individuals requires teaching them to learn how to stop, slow down, and look for their contentment from within. They should balance their feelings of love with contemplation and wisdom, avoid over-enthusiasm, and allow their inner spirit to radiate through in a more sober way.

THE EARTH CONSTITUTION

EARTH OUTER APPEARANCE

Earth constitution types usually have a somewhat large body, large head, large belly, strong thighs, round face, and wide jaw (Figure 3.12). They are categorized as having excessive flesh, with upper and lower limbs mutually well-proportioned. By nature, they enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Earth constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong muscles and yellow facial color. Their note is "gong."

EARTH PERSONALITY

The Earth personality can be divided into three subdivisions.

1. A Balanced Earth nature predisposes individuals to lead quiet, stable, and peaceful lives, unconcerned about fame or wealth. They are always at ease, calm, generous, forgiving, sincere, and unambitious. They usu-

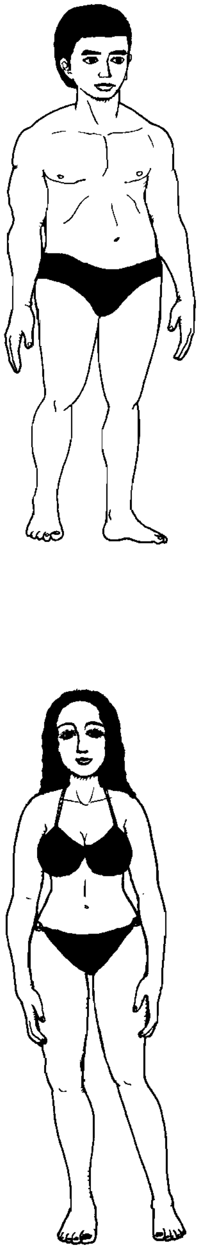

Figure 3.12. The Earth Constitution in Males and Females

ally have open minds and rarely live in fear, or make excessive demands. They have analytical minds and are logical and practical; they use sound reasoning to convince others of their opinions. They are quite adaptable to changing situations, and rarely use coercion to achieve power. They are kind and gentle with an earnest and well-mannered attitude. They are pleasant, sweet, sympathetic and caring, and are able to maintain boundaries.

2. A Yin Earth nature predisposes individuals to worry endlessly, becoming suspicious and self-centered. They think too much, with not enough follow-through action. Because they feel empty inside, they find it difficult to be nurturing to themselves and others.

   Treatment for Yin Earth nature individuals requires teaching them to learn how to let go of their inner feelings of worry and defensiveness. These individuals need to examine their assumptions, connect with their physical bodies, come out of their shells, and begin to live in the real world. They need to replace their negative thought patterns with positive affirmations, and realistically look at their true potential and abilities for powerful actions.

3. A Yang Earth nature predisposes individuals to cling to others; they are co-dependent, pushy, and possessive. Such individuals try to dominate in a passive-aggressive way while limiting the independence of others.

   Treatment for Yang Earth nature individuals consist of teaching them to develop their inner strength to control the fear, insecurity, and feelings of inner emptiness that make them want to hold on to others. They need to create and establish love as a source of security from within themselves. They need not depend on, or build their lives exclusively on, the praise or presence of others. They must learn to become emotionally independent.

THE METAL CONSTITUTION

METAL OUTER APPEARANCE

Metal constitution types usually have a triangular face with broad, square shoulders, a strong voice, and a strong, muscular build (Figure 3.13). By nature, they enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Metal constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include strong Lungs and white facial color. Their note is "shang."

METAL PERSONALITY

The Metal personality can be divided into three subdivisions.

1. A Balanced Metal nature predisposes individuals to process their grief, letting go of their past emotional baggage. They believe that they cannot be in the present without cleaning up and being released from their past. They are able to gain knowledge and wisdom by gathering and releasing their emotional bonds while learning and growing from each emotional attachment. They participate in life and form new bonds without fear of loss. They are also generous, just, and bright.

2. A Yin Metal nature predisposes individuals to have a difficult time creating lasting bonds because they are fearful of establishing new relationships due to past losses and emotional trauma. They avoid joining in with others and withdraw from active participation. Living in past remorse, they grieve over lost opportunities. They are also haughty and coldhearted. They are jealous, cunning, sneaky, and furtive. They become angry when they do not have things that others possess. Because they are covetous, they only obtain happiness by possessing certain objects.

   Treatment for Yin Metal nature individuals requires teaching them how to strengthen their physical body, as well as the energy of their Heart, Spleen, and Lower Dantian. They also need to strengthen their abilities to form close emotional bonds with people, reduce their fears of rejection and abandonment, and gain the strength and courage to let go of their past hurts. They need to come out from inside of themselves emotionally and experience the warmth and compassion of life.

3. A Yang Metal nature predisposes individuals

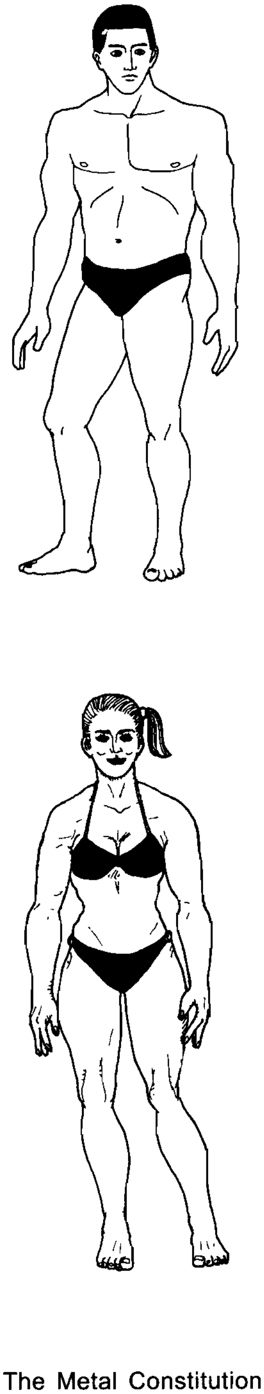

Figure 3.13. The Metal Constitution in Males and Females

to suppress their emotions and hold on to their grief. They are generally considered whiners and complainers, talking to others about their complaints in order to unload their grief. They use new relationships as an emotional bandage to avoid the unprocessed grief of their past relationships. They are also meticulous, independent, and strong-willed.

Treatment for Yang Metal nature individuals requires teaching them how to let go of the need to control and suppress their emotional feelings, and how to genuinely grieve over past wounds. They need to face the truth and be honest about their feelings, instead of selfishly using people and new relationships to cover up their grief. They also need to learn to be sympathetic with the pain and sorrows of others, and put their feelings of grief into perspective.

THE WATER CONSTITUTION

WATER OUTER APPEARANCE

Water constitution types usually have a big round face, head and body (Figure 3.14). They have a long upper back, and unbalanced or uneven physical features. They, by nature, enjoy autumn and winter but dislike spring and summer. This preference is due to the vulnerability of the Water constitution and its susceptibility to pathogenic invasion and disease during these time periods. Their dominant features include their digestive system and black or dark facial color. Their note is "yu."

WATER PERSONALITY

The Water personality can be divided into three subdivisions.

1. A Balanced Water nature predisposes individuals to be skilled negotiators, that are not discouraged by difficulty, and do not take foolish risks. They are sympathetic and loyal to their employer and friends. They have a clear perspective and are sensitive and intuitive. They are powerful, tender, and soft. They have a firm will and know their boundaries and limitations. They are known for their inner strength and strong faith in themselves.

2. A Yin Water nature predisposes individuals to lack spiritual, emotional, mental, and physical energy. They give up on life and surrender the control of their own destinies. They lack the determination to achieve their goals, do everything halfheartedly, and become easily discouraged by difficult challenges.

   Treatment for Yin Water nature individuals requires teaching them how to conserve and strengthen their energy by not attempting to do what's beyond their capacity. They should complete their projects, however, once they have started them, and not procrastinate or leave tasks unfinished. They need to find the strength to overcome their fear of failure by learning to take action through the completion of their goals.
3. A Yang Water nature predisposes individuals to be ambitious overachievers and to live under great stress. They lack consideration for others, and can be reckless and foolhardy. They can also be greedy, ruthless, and cold-blooded. Seemingly modest, they are actually insidious and sinister, concealing their true emotions by suppressing their fears. Because they fear loss of control, their safety lies in dominating others. They only perform duties that are self-serving and blame others for their problems.

   Treatment for Yang Water nature individuals requires that they learn how to act from inner stillness and gain strength and courage from their inner true self. They also need to emotionally and physically slow down, and to balance their activities with rest. They need to learn to gain consideration for themselves and others, as well as to open up their heart and begin to love.

COMBINED CONSTITUTIONS

In ancient China, The Yellow Emperor's physician Shao Shi taught that the Five Elemental constitutions consist of Five Elemental forms, which can be further divided into five additional shapes depending on the strength or development of the individual's Jing, Qi, and Blood. After establishing the patient's physique (Wood, Fire, Earth,

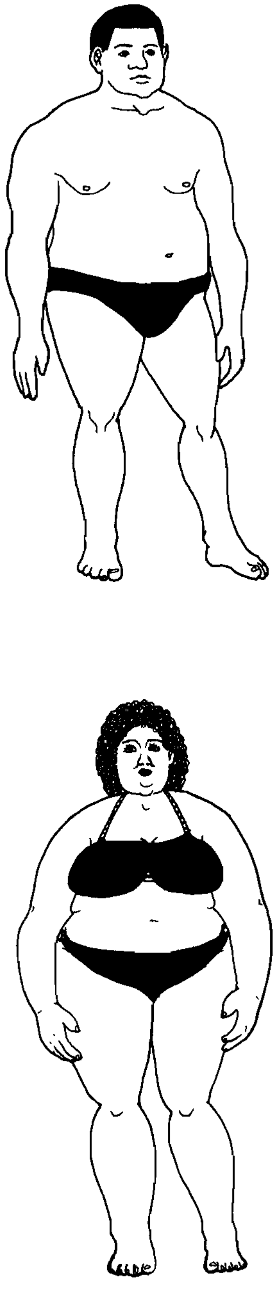

Figure 3.14. The Water Constitution in Males and Females

Metal, or Water), the patient is then classified in terms of five separate colors, and differentiated according to the five prenatal (Heavenly) notes. This system recognizes 25 constitutional types.

The Five Elemental Constitutions are therefore, guidelines for the Qigong doctor to evaluate the patient. These guidelines are not fixed, as the doctor is most likely to encounter combinations of two or more of these elements within one person.

When two constitutions combine in an individual, the combination may be congenital or acquired. The Qigong doctor must evaluate the patient, differentiating between his or her congenital and acquired tissue formations.

A woman who has large hips (Water) and small chest and shoulders (Wood), for example, probably inherited this structure from her parents. This woman's physical structure and personality would then be considered a combination of both Water and Wood (depending on which internal organ characteristic dominates).

A female patient who swims, on the other hand, may be congenitally Wood but may have developed a dominant Metal upper physique through years of physical training. If, through intense physical activity, someone acquires a Metal physique, he or she may also acquire a Metal personality, depending again on his or her psychological format, which is rooted within the internal organs.

The external physical structure serves as a visual guide to the interaction and balance of the internal organs.

PERSONALITY CONSTITUTIONS OF THE EIGHT EXTRAORDINARY VESSELS

Another system of identifying personality traits is practised by observing the four personality constitutions of the patient's Eight Extraordinary Vessels.

Through observation and inspection of the patient's personality and overall energetic and emotional demeanor, the Qigong doctor can further determine which organ and organ system, is Deficient or has an Excess condition and set up a treatment plan.

To begin, the Qigong doctor will first observe the patient's dominant Yin or Yang emotional characteristics and determine which of the Eight Extraordinary Vessels are governing his or her present state of feelings. These personality traits may appear in different combinations, or oscillate from one type of characteristic to another, due to the crossover of acquired and congenital organ energy. This crossover can change with the patient's age (maturation) or situation (environment).

THE PATHOLOGICAL PERSONALITY TYPE OF THE GOVERNING AND YANG HEEL VESSELS

The Governing and Yang Heel Extraordinary Vessels are located on the back and lateral sides of the body and affect the Bladder, Stomach, and Gall Bladder Channels. To "open" and increase the flow of energy within these channels, the Qigong doctor stimulates the patient's SI-3 and Bl-62 points simultaneously (see Chapter 8).

The pathological personality types of the Governing and Yang Heel Vessels are divided into Yin and Yang emotional characteristics. These characteristics are as follows.

1. Patients with Deficient Qi in the Governing and Yang Heel Vessels tend to be weak-willed and spineless. They tend to be overly submissive, and easily give up their personal power. They lack determination, courage, clarity, and inner strength, and have no control over their lives. They participate as little as possible in life for fear of failure.
2. Patients with Excess Qi in the Governing and Yang Heel Vessels tend to be tense, overpressured, stiff, rigid, inflexible, stubborn, and awkward. They are narrow-minded and overcontrolling. Fearful of letting go, they overcompensate and attempt to restrict reality.

THE PATHOLOGICAL PERSONALITY TYPE OF THE CONCEPTION AND YIN HEEL VESSELS

The Conception and Yin Heel Extraordinary Vessels are located on the front and inside of the body. They affect the Lung and Kidney Channels, which in turn affect the Kidneys, Lungs, and Heart organs. To open these channels, the Qigong doctor stimulates the patient's Lu-7 and Kd-6 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Qi in the Conception and Yin Heel Vessels tend to be weak and depressed. They live in the past, lack interest in life, have no ambition, and also lack interest in sex. It is hard for them to form new relationships because they fear forming emotional bonds due to past losses, failures, and emotional traumas. They constantly daydream and live in their own world of past memories.
2. Patients with Excess Qi in the in the Conception and Yin Heel Vessels tend to participate in life, but hold on to, and suppress, their grief. They fail to let go of emotional attachments and fear being alone. The female patients usually develop Qi stagnations resulting in breast cysts, tumors, and cancer, as well as uterine fibroids and cancer.

THE PATHOLOGICAL PERSONALITY TYPE OF THE BELT AND YANG LINKING VESSELS

The Belt and Yang Linking Extraordinary Vessels are located around the waist and on the outside of the body. They affect the Gall Bladder and Triple Burner Channels, which in turn, affect the Kidneys, Liver, and Gall Bladder organs. To open these channels, the Qigong doctor stimulates the patient's GB-41 and TB-5 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Qi in the Belt and Yang Linking Vessels are considered weak, indecisive, and unproductive. They fear criticism and are touchy, snappy, irritable, and hypersensitive. They have low self-esteem and also lack interest in sex.
2. Patients with Excess Qi in the in the Belt and Yang Linking Vessels are aggressive, angry, resentful, bitter, and vindictive. They are also opinionated, inflexible, intolerant, domineering, selfish, and frustrated, with a sex life based on anger or rage.

THE PATHOLOGICAL PERSONALITY TYPE OF THE THRUSTING AND YIN LINKING VESSELS

The Thrusting and Yin Linking Extraordinary Vessels are located within the center of the body (internally) and on the medial and anterior aspect of the body (externally). They affect the Kidney and Stomach Channels, which in turn, affect the Kidneys, Spleen, and Heart organs. To open these channels, the Qigong doctor stimulates the patient's Sp-4 and Pc-6 points simultaneously. These character diagnoses are as follows.

1. Patients with Deficient Thrusting and Yin Linking Vessels are physically and emotionally weak. They quickly become exhausted and depressed. They easily have their feelings hurt and take a long time to recover from emotional upsets. Their internal world of emotional existence is made up of fear, anxiety, worry, and tension. They do not connect well with the outside world or gain much pleasure from life or relationships.
2. Patients with an Excess in the Thrusting and Yin Linking Vessels tend to participate in life, but have difficulty expressing their affections and communicating their needs in relationships. They are inclined to have stagnant Qi and Blood in their chests, epigastrium, or uterus areas due to fear, anxiety, worry, and sorrow. In their personal relationships, they fear loss of control and cannot make a commitment or a deep emotional connection with others. They also fear surrendering and letting go in their love and sexual relationships.

THE YAO IMAGE AND PHYSICAL ANATOMY

Sometimes because of surgery, congenital deformity, or traumas, it is not possible, nor is it recommended, to directly treat certain areas of the patient's body. The Qigong doctor must then choose an indirect method of treatment to ensure better results. One such method available to the doctor utilizes the understanding of the Yao image and the patient's physical anatomy. In order to better comprehend the complexity of this diagnostic and treatment modality, the Qigong doctor must first understand the energetic origin of the Yao images.

All matter is composed of different proportions of Yin and Yang energy. Within the Void (infinite space or Wuji) both Yin and Yang energy gather or disperse to balance the forces of nature. This interaction causes the Yin and Yang energies

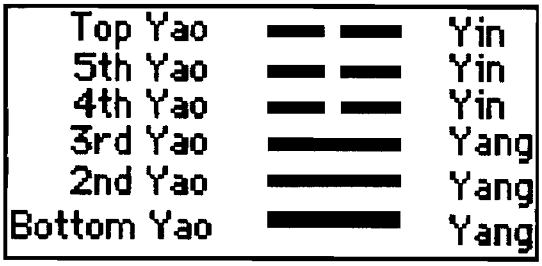

Figure 3.15. This diagram shows a hexagram, composed of an upper and lower trigram. In this particular example, the upper trigram is all Yin lines (symbolized by broken lines), the lower all Yang (symbolized by solid lines).

to develop and transform themselves into four phases of energetic powers or stages: Great Yin, Small Yin, Great Yang, and Small Yang. These four phases can be explained as follows.

1. Great Yin (Tai Yin), is affiliated with midnight and the new-moon phase. Modern physicists associate the Great Yin with a weak nuclear force.
2. Small Yin (Shao Yin), is affiliated with the sunset and the waning-moon phase. Modern physicists associate the Lesser Yin with a heavy force, and gravity.
3. Great Yang (Tai Yang), is affiliated with high noon and the full-moon phase. Modern physicists associate the Strong Yang with a strong nuclear force.
4. Small Yang (Shao Yang), is affiliated with the sunrise and the waxing-moon phase. Modern physicists associate the Lesser Yang with a light force and electromagnetism.

These four solar and lunar energetic phases transform themselves into the energies of the eight foundational trigrams. These trigrams further combine to form sixty-four hexagrams, and manifest as the ancient binary system known as the Yi Jing (I Ching). The formation and pattern of these energetic powers or stages vary according to the composition of Yin and Yang, and can be expressed through the energetic symbols of the Yao.

A Yao is a line that represents either Yin or Yang energy. The lines fall into two categories:

- The negative Yao is symbolized by broken lines (Yin energy).
- The positive Yao is symbolized by solid lines (Yang energy).

When these lines are stacked in combinations of three, they form what are known as trigrams (a group of three Yao lines). When pairs of trigrams are joined, hexagrams (six lines) are formed. The lines are arranged from bottom to top, with the first Yao being on the bottom (Figure 3.15).

The Yao trigrams and their relationship to the body's physical and energetic structure can be utilized by the Qigong doctor for diagnosis and treatment. The doctor applies the Yao image as a template and arranges its pattern alongside the physical form of the patient's body. This arrangement of Yao images is designed by the doctor to assist him or her in recognizing the physical and energetic imbalances within the patient's body. This type of physical Yao image diagnosis has been used in China for centuries.

Knowing that each different region of the human body corresponds to one of the six Yaos in the hexagram, as well as a different aspect of Qi, the ancient masters explained the energetic Yao and physical body interaction as follows.

1. The two lowest Yaos align with the feet and legs and correspond to the body's interaction with the Earth's energy.
   a. The 1st Yao extends from the patient's feet, ankles, and shins, to the knees.
   b. The 2nd Yao extends from the knees to the upper thighs.
2. The middle two Yaos align with the lower and upper abdominal regions, and correspond to the body's interaction with human energy, or "Man's energy" (emotional responses).
   a. The 3rd Yao extends from the perineum to the navel.
   b. The 4th Yao extends from the navel to the xiphoid process in the diaphragm.
3. The upper two Yaos align with the upper chest, neck and head, and correspond to the body's interaction with Heaven's energy.
   a. The 5th Yao extends from the xiphoid process of the diaphragm to the top of the manubrium, at the base of the throat.

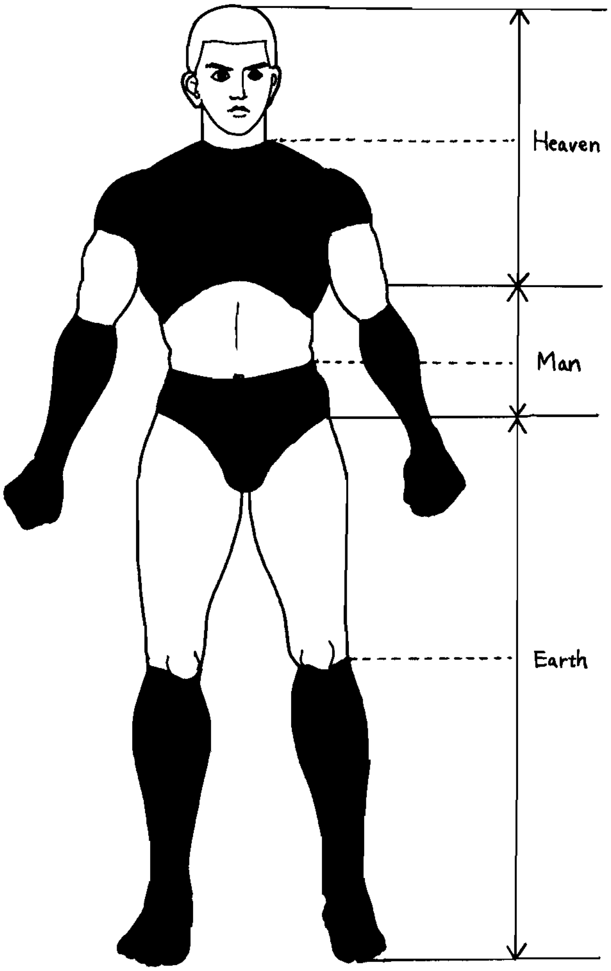

- The top or 6th Yao is from the neck to the top of the head. (Heaven)
- The 5th Yao is from the diaphragm to the neck. (Heaven)
- The 4th Yao is from the navel to the diaphragm. (Man)
- The 3rd Yao is the lower perineum to the navel. (Man)
- The 2nd Yao is from the knee to the top of the thigh. (Earth)
- The bottom or 1st Yao is from the foot to the knee. (Earth)
- The Categorization of the arms are as follows: From the fingers to the elbow represents the Bottom or 1st Yao; From the elbows to the base of the shoulder represents the 2nd Yao; From the base to the top of the shoulder represents the 3nd Yao.

Figure 3.16. Yao Image and the Human Body

b. The 6th Yao extends from the manubrium, at the base of the throat, to the top of the head.

The doctor further divides the patient's body into six upper and lower Yaos, to determine the dominant energetic and emotional patterns (where the patient's Qi is gathering, collecting, and stagnating). Through observation and study, the Qigong doctor is able to determine the energetic association the patient has with the natural environment. This diagnosis is achieved by observing the body's relationship to the three energetic Yao divisions of Heaven, Earth, and Man (Figure 3.16). When observing the body, for example, if the upper Yao is out of harmony with the patient's middle and lower Yaos, the patient's energetic balance may be compromised, and he or she may experience a spiritual disconnection with his or her body. This can result in either a Yang condition (heaviness, restlessness) due to an Excess spiritual "Heaven" state, or a Yin condition (tired, dizziness) due to a Deficient spiritual "Heaven" state.

In addition to diagnosing the hexagram for the body's Yin and Yang relationship to Heaven, Earth, and Man, the ancient Chinese Qigong masters also studied two other hexagrams for clinical evaluation. These hexagrams were known as the Upper and Lower Hexagrams.

UPPER AND LOWER HEXAGRAMS

In the Ming Dynasty (1368-1644), Tang Rongchuan stated in his book *Detailed Explanations of Application of the Book of Changes to Medicine*, "In the clinical environment, the body may be divided into two different sets of hexagram constructions [the upper and the lower hexagram partition] for the purpose of making a diagnosis, detecting Qi, and for performing Qigong therapeutic treatments." This relationship of Yao positioning is suitable to both the channel system and the nervous system.

Physical disorders located within these hexagrams are generally expressed as symptoms manifesting in the extremities. If stagnant Qi or other abnormal conditions occur in any particular part of the body, an imbalance of both Internal and External Qi may be detected in either that specific area of the body or its corresponding area.

The ancient Qigong masters used the body's hexagrams to explain the etiology, pathology, clinical manifestations, and treatment principles of disease.

The body's upper torso, arms, neck, and head are divided into what is known as the upper hexagram partitions, while the body's lower limbs and torso are divided into the lower hexagram partitions. The six Yao of the upper and lower hexagram partitions interlock at the chest and abdomen. By using these sets of hexagrams, the Qigong doctor has a template or guiding pattern with which to understand the patient's body.

The Upper Hexagram Construction

The Upper Hexagrams range from the base of the lower perineum to the top of the patient's head. These six Yaos are organized as follows.

1. The 1st or bottom Yao encompasses the pubic symphysis to the navel. Reproductive, digestive, and urinary tract diseases are ascribed to this area.
2. The 2nd Yao encompasses the navel to the xiphoid process. Malfunctions in digestion, elimination and the transportation of nutrients are assigned to this area, as well as diseases of the Kidneys, adrenal glands, Liver, Spleen and pancreas.
3. The 3rd Yao encompasses the xiphoid process to the supraclavicular notches. Respiratory and circulatory diseases are ascribed to this area.
4. The 4th Yao encompasses the supraclavicular notch to the tip of the nose. Teeth, jaw, and thyroid diseases are ascribed to this area.
5. The 5th Yao encompasses the tip of the nose to the eyebrow. Sinus problems and headaches, as well as ear, nose, and eye diseases are manifest in this area.
6. The 6th or top Yao encompasses the eyebrows to the top of the head. Several types of headaches, brain tumors, and upper cranial dysfunctions are ascribed to this area.

The Lower Hexagram Construction

The Lower Hexagrams range from the bottom of the patient's feet to the base of the patient's throat. These six Yaos are organized as follows:

1. The 1st or bottom Yao encompasses the bottom of each foot to the ankle. Foot, toe and ankle dysfunctions are ascribed to this area.
2. The 2nd Yao encompasses the ankles to the knees. Shinsplints and calf and ankle dysfunctions are ascribed to this area.
3. The 3rd Yao encompasses the knees to the pubic symphysis. Thigh, knee, and quadriceps dysfunctions are ascribed to this area.
4. The 4th Yao encompasses the pubic symphysis to the navel. Reproductive, digestive, and urinary tract diseases are ascribed to this area.
5. The 5th Yao encompasses the navel to the xiphoid process. Digestion, elimination, and the transportation of nutrients, as well as diseases of the Kidneys, adrenal glands, Liver, Spleen and pancreas are ascribed to this area.
6. The 6th or top Yao encompasses the xiphoid process to the supraclavicular notches at the base of the throat. Respiratory and circulatory diseases are ascribed to this area.

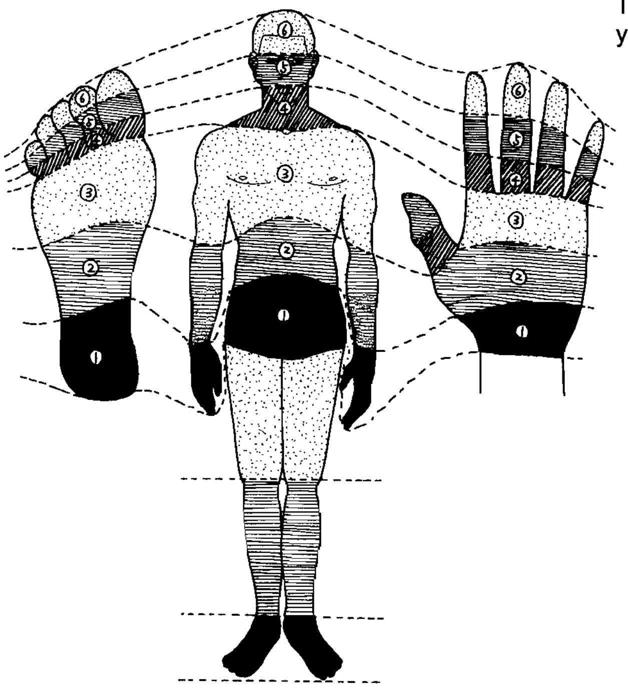

The diagram at left depicts the correspondence of the yaos of the body to the yaos of the hands and feet.

6 The second joint of the toes or fingers pertains to the top Yao and the energy of Heaven.

5 The first joint of the toes or fingers pertains to the 5th Yao and the energy of Heaven.

4 The area below the first joint of the toes or fingers pertains to the 4th Yao and the energy of Man.

3 The ball of the foot or upper palm pertains to the 3rd Yao and the energy of Man.

2 The instep on the midpalm with the thumb pertains to the 2nd Yao and the energy of Earth.

1 The heel or lower palm pertains to the bottom Yao and the energy of Earth.

Figure 3.17. The Therapeutic Use of the Yaos and the Upper and Lower Hexagrams

THERAPEUTIC USE OF THE YAOS

The entire body can be treated by focusing on one small area. When Excess, Deficient, or stagnant Qi occurs in one part of the body, the imbalance of Internal and External Qi is detectable in that area or its corresponding region on one of the body's extremities. This ancient Chinese modality of treatment is popularly called reflexology in the West.

The Yao images can be superimposed onto the patient's body, and the patient diagnosed according to the energetic symptoms expressed through either the patient's extremities (the hands, forearms and arms; or the feet, shins, and thighs), or the six divisions located on the patient's torso and head. A Yao hexagram can additionally be visualized on specific locations of the patient's body such as the face, nose, ears, hand or foot for diagnosis. The base of the palm's heel, for example, corresponds to the lower abdominal area and reproductive organs, which relate to the 1st, or lower, Yao in the hexagram of the hands. These six Yao areas of the hand provide a basis for diagnosis, for treatment of certain diseases, and even for perception and prediction of certain future diseases (Figure 3.17).

As an example of how the Qigong doctor can use the six Yaos of the body for emitting Qi therapy, consider the following situation: An elderly patient with a problem in his prostate area visits the Qigong doctor. The prostate is located in the lower abdominal area, which is located in the 1st Yao of the upper hexagram. Since the six Yao positions can be transferred to the hands, feet, etc., the doctor knows that this particular Yao is congruent to the 1st Yao of the foot. Therefore, as an energetic entry point into the patient's prostate area, the doctor may focus his or her attention on a specific area on the foot.

Dividing the foot into six Yaos, the doctor knows that the 1st Yao of the foot is the heel area. In order to treat the prostate gland, the doctor will

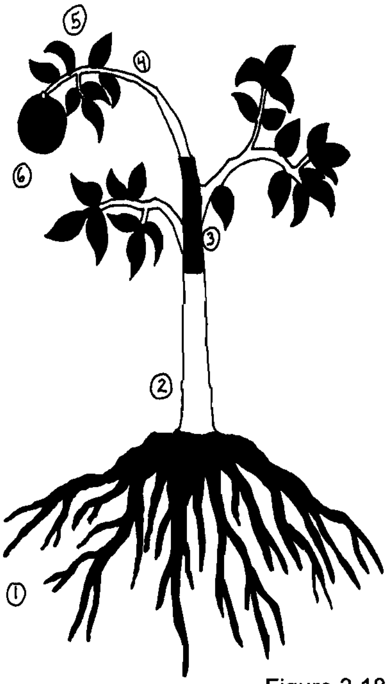

6 The fruit or flower of the herb pertains to the top Yao and the energy of Heaven.

5 The leaf of the herb pertains to the 5th Yao and the energy of Heaven.

4 The branch of the herb pertains to the 4th Yao and the energy of Man.

3 The upper stem of the herb pertains to the 3rd Yao and the energy of Man.

2 The stalk of the herb pertains to the 2nd Yao and the energy of Earth.

1 The root of the herb pertains to the bottom Yao and the energy of Earth.

Figure 3.18. The Yao Image Applied to Plants

focus his or her treatment on the patient's heel and ankle area. Alternatively, the doctor could also have chosen the 1st Yao area of the hand, which is considered the 1st Yao position of the arm. The 1st Yao of the hand corresponds to the prostate and urogenital area.

HERBS AND THE YAOS

Affirming the ancient Chinese belief that the energy of Heaven, Earth, and Man combine as "one," and exist in the same world of energetic dynamics, herbs are also structured and prescribed in accordance with the six Yao hexagram concept. With plants, the following correspondences apply (Figure 3.18):

1. The roots correspond to the bottom or 1st Yao.
2. The stalk corresponds to the 2nd Yao.
3. The upper stem corresponds to the 3rd Yao.
4. The branches correspond to the 4th Yao.
5. The leaves correspond to the 5th Yao.
6. The flowers and fruit of the plant correspond to the top or 6th Yao.

When plants are utilized as a medicine, the part of the plant corresponding to the number and position of the Yao may manifest its energetic properties, affecting the ascending, descending, floating, or sinking action of the patient's Qi. The achyranthes root (Niu Xi), for example, corresponds to the bottom Yao, and leads healing Qi downward, to heal the patient's knees and joints; whereas the chrysanthemum flower (Ju Hua) corresponds to the top Yao, and leads healing Qi up into the patient's head and eyes.

IDENTIFYING INTROVERTED AND EXTROVERTED STRUCTURES

Qi flows through the matrix of the connective tissue and adapts to the postural and structural demands of the body by changing the density and direction of the body's connective tissues and inner fasciae. A postural habit will create an imprint or tissue memory in the supporting connective tissue, as well as an energetic memory in the supporting field of Qi. These imprints result in introverted as well as extroverted postural structures.

Each introverted (Yin) and extroverted (Yang) postural structure has a predictable impact on the

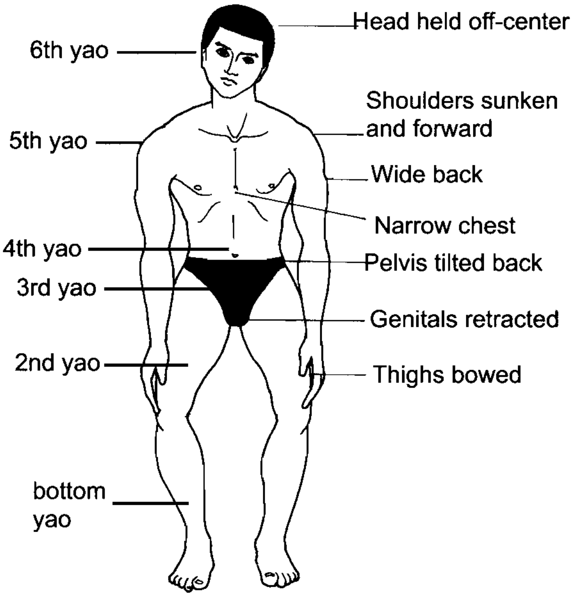
Head held off-center
6th yao
Shoulders sunken and forward
5th yao
Wide back
Narrow chest
4th yao
Pelvis tilted back
3rd yao
Genitals retracted
2nd yao
Thighs bowed
bottom yao

Figure 3.19. The Introverted Structure

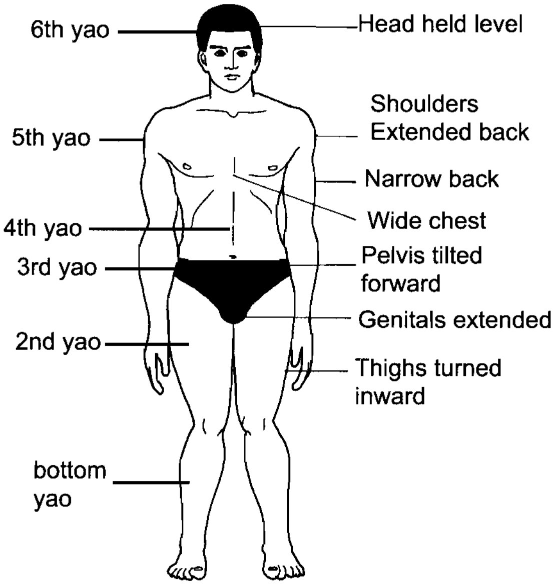
6th yao
Head held level
Shoulders Extended back
5th yao
Narrow back
Wide chest
4th yao
Pelvis tilted forward
3rd yao
Genitals extended
2nd yao
Thighs turned inward
bottom yao

Figure 3.20. The Extroverted Structure

myofascial webbing, as well as a natural emotional response. Emotions respond to changes in breathing patterns and vice versa. Both the prevalent breathing pattern and the patient's general mood affect the body's posture and eventually the body's structure. All chronic stress creates postural imprinting or tissue memory, and a corresponding energetic memory in the supporting field of Qi.

INTROVERTED STRUCTURE

When observing the introverted type of structure, the doctor will notice that the back (Yang) area of the patient is expanded, and the muscle structure is usually well-developed. The front of the body (Yin), especially around the chest and Heart area, is retracted, resulting in a narrow chest. The patient's head is generally held off-center, shoulders are sunken and forward, and the pelvis is tilted back, with the genitals retracted (anterior tilt). This condition is commonly referred to as "extended Yang" and "retracted Yin" of the upper torso (Figure 3.19).

EXTROVERTED STRUCTURE

When observing the extroverted type of structure, the doctor will notice that the back (Yang) area of the patient is retracted, while the front of the body (Yin) is expanded and overly exposed. The head is usually held vertically, the shoulders are extended back (military posture), the chest is wide, and the pelvis is tilted forward, with genitals extended (posterior tilt). This condition is commonly referred to as "extended Yin" and "retracted Yang" of the upper back (Figure 3.20).

COMBINED STRUCTURAL FORMATIONS

As with the Five Elemental Constitutions, the introverted and extroverted structures are guidelines for the Qigong doctor to evaluate the patient. These guidelines are not fixed, as the doctor is most likely to encounter combinations of both introverted and extroverted structures within one person.

When two structures combine in an individual, the combination may be due to a congenital or acquired formation. The Qigong doctor must evaluate and assess the patient, differentiating between his or her congenital and acquired tissue formations.

CONGENITAL AND ACQUIRED CELLULAR PATTERNS

The Qigong doctor evaluates the physical condition according to the congenital or acquired patterns manifested within the patient's tissues. The doctor may ask questions to determine if the condition was acquired through exposure to environmental pathogens, diet, trauma, occupation, or inherited. This information aids the Qigong doctor in understanding and assisting the patient in reprogramming any congenital or acquired cellular patterns, as well as internal disease functions.

1. Congenital Cellular Patterns and disease predispositions are locked within the patient's congenital Jing and can be released like a biological time bomb. This is a similar concept to the Western medical concept of genetic congenital disease. The goal of the Qigong doctor is to alter (if possible) the Toxic Qi patterns and help the patient's body recognize that the formation of the disease is a mistake. In order to reprogram these patterns, the Qigong doctor often assists the patient in using guided meditations that employ vivid imagination, colorful visualizations, and positive affirmations. These images are used to encapsulate the disease and return it to the divine light. While in the presence of the divine light, the Toxic Qi (Xie Qi) will be energetically transformed and recycled back to the patient as healing energy. Patients with a direct family history of cancer, for example, can benefit from this type of medical Qigong imagery.
2. Acquired Cellular Patterns and disease functions arise from either external exposure to pathogens, or the suppression of extreme emotions. The goal of the Qigong doctor is to first assist the patient in discovering the origin of the disease. The next step is to teach the patient to alter the toxic patterns and help the patient's body recognize that the development of the disease is a mistake. In order to reprogram these patterns, the Qigong doctor will assist the patient in learning meditations that employ healing mental imagery, colorful visualizations, and positive affirmation. The patient must encapsulate the disease and return it to the divine light, to be energetically transformed and recycled. A patient who has been experiencing severe headaches, for example, could benefit from this kind of treatment given that most headaches are stress related.

SUMMARY

Through the observation of the patient's physical and structural development, a clear understanding can be obtained as to the strength and weakness of the patient's internal organs and immune system.

The four patterns of physical development and structural formation (The Five Elemental Constitutions, The Yao Hexagram Formations, The Yin or Yang Structures, and The Congenital and Acquired Cellular Patterns) are determined as follows.

1. The doctor first begins to diagnose the patient's physical body according to the Five Elemental Constitutions and the dominant element. This helps the doctor to understand the dominant condition of the patient's internal organs.
2. The doctor further divides the patient's body into six upper and lower Yaos, to determine the dominant energetic and emotional patterns (where the patient's Qi is gathering, collecting, and stagnating).
3. Next, the doctor evaluates the patient's external structure according to his or her body's Yin and Yang characteristics, to understand whether the patient has a dominant introverted or extroverted physique.
4. Finally, the Qigong doctor determines whether the patient's condition is congenital, acquired, or a combination of both. This information helps the doctor to understand the innate cellular patterns that dominate the patient's physique.

CHAPTER 4
THE FIVE ENERGIES OF THE HUMAN BODY

The study of various forms of energy and their transformations from one form to another have led to one of the greatest generalizations in physics -- the Law of Conservation of Energy, which states, "Energy cannot be created or destroyed. It may be transformed from one form into another, but the total amount of energy never changes."

Physics defines four types of energy: mechanical, electromagnetic, chemical, and thermal. All four types of energy are related to and convert from one another. Electricity, for example, can be converted into sound, light, mechanical energy and into heat. Sound, light, mechanical energy and heat can all be reconverted into electricity.

Western Science is still for the most part, founded on the principles of Newtonion Physics, that views energy as an impersonal and inanimate force. The Traditional Chinese view of energy is much more akin to the theories of Quantum Physics. Both Chinese energetic theories and Quantum Physics hypothesize that energy cannot be studied without taking into account the mind's influence over it. Furthermore, energetic behavior cannot be studied independently of matter.

From a medical Qigong perspective, the entire human body is suffused with energy, which converts within the human organism. Light energy, for example, is absorbed, stored and converted into the various energies of the body, the same way that plants convert absorbed and stored light into chemical energy.

The human body possesses five energies which resonate within the very core of one's being. Sound, light, magnetic fields, heat, and electricity envelop and permeate our very existence; these energies sustain, govern, control, and determine our psychology as well as our physiology. In Medical Qigong therapy, these energies are considered the spiritual reality that governs the physical reality. These energies are observed in both diagnosis and treatment of illness. Research in China confirms the fact that the body emits sound, light, magnetic energy, electricity, and heat.

SOUND ENERGY RESONANCES

When the body experiences any type of sound, the cell tissues respond to the tone frequency and have either a negative or positive automatic response. The human body both receives and generates sound energy. These subtle sounds resonate from three distinct physical energetic actions. The following subtle sounds are natural tones that resonate as a result of three physical actions.

1. Breathing creates respiratory sound resonances that follow the breath in and out of the lungs mouth, and nose.
2. Muscle movements create somatic and visceral sound resonances that include the clicking sound of the joints, beating of the Heart, gurgling of the Stomach and intestines, etc.
3. Qi activities can produce energetic sound resonances (i.e., when Liver Yang, Liver Fire, or Liver Wind rise, a high-pitched tone is heard in the ears. When the Kidney Yin is Deficient, this sound is like rushing wind, etc.).

SOUND THERAPY THROUGH EMITTED QI

As sound waves vibrate through the body, crystalline structures within the tissues transform the vibration into pulsed currents. These currents are then conducted to the various corresponding organs and glands, depending on the frequency and amplitude of the incoming wave signal. This tissue transformation (due to wave vibrations) automatically changes the function and flow of energy in the body. Thus sound vibrations have a profound effect on both human psychology and physiology. Sound or tone resonances, have been used for centuries as an effective healing tool and are currently used as an adjunct to modern Qigong Medicine.

For centuries in China, patients have used healing sound therapy to resonate certain parts of the body, stimulating the flow of Blood and Qi to and from specific internal organs to cure disease. These healing sounds specifically relate to the Five Agents and the Five Elemental aspects of the patient's internal organs (see Chapter 2). According to the Five Elements' Cycle, each of the Five Elements has a note:

- Gong - C = Earth
- Shang - D = Metal
- Jue - E = Wood
- Zhi - G = Fire
- Yu - A = Water

The Qigong doctor can project resonant sound vibrations by first focusing on all three Dantians at the center of the Taiji Pole. This focused concentration is used first to vibrate the doctor's specific internal organs and then to treat the patient. The Qigong doctor can use vibrating sound therapy by:

1. Audibly speaking the healing sounds to fill the Qigong doctor's energetic field with healing sound vibration. This energetic field of sound is then projected onto his or her patient. Audible sound resonation is considered a Jing to Qi level energy projection, and is generally used when the patient is armored and not energy sensitive.
2. Inaudibly speaking the healing sounds and projecting these mental vibrations into the patient. In this case, the Qigong doctor will begin by internally focusing his or her mind on the healing sound. As this internal sound fills the Qigong doctor's energetic field, it is then emitted into the patient's body. To increase the intensity of energy sound projection, the doctor needs only to increase the mind's intention and the pressure of his or her exhalation. Inaudible sound resonation is considered a Qi to Shen level energy projection, and is generally used when the patient is energy sensitive.
3. Both the doctor and patient simultaneously resonate the sound to enhance the synchronization of both their energy fields. As the Qigong doctor resonates his or her own internal organs, both the doctor and patient can modulate the sound and energy patterns as needed.

THREE MODALITIES OF SOUND THERAPY

The resonant sounds produced by Qigong doctors are effective because the patients' energetic matrix is sensitive to the sound resonances. The modulated, yet unpredictable, nature of sound pulsation affects the vibrational structures which shape and maintain the patient's physical structure. Once the patient's vibrational structures have been disturbed, their body must immediately begin its reconstructive process, thereby creating new energetic patterns to restructure and heal the tissues.

There are three modalities used for sound therapy:

1. Characteristic Sound Therapy refers to the sound from musical instruments (flute, guitar, piano, drum, etc.) and their effect on the subconscious mind. This type of musical energy creates an emotional, physical, and spiritual release of energy within the listener.
2. Extraordinary Sound Therapy refers to the vocal resonances used in silent chanting, prayer and singing. Such techniques include the Daoist six-word healing sound method, Zhuang Zi's breath listening method, Lao Zi's sound-voice method, etc. This type of therapy uses the patients' energy to produce the sound and inner vision to direct the vibration within the body. It is also practised in conjunction with breathing techniques that are guided by the Yuan Shen (the intuitive consciousness of the spirit). All of these techniques are combined to achieve physiological and psychological healing.
3. Infrasonic Sound Therapy refers to the low frequency sound waves emitted from the hands of Qigong doctors. These chaotic, low frequency sound waves, inaudible to the human ears, are naturally produced by the hands

of experienced Qigong doctors when emitting Qi. This infrasonic sound resonance affects the central nervous system, changing the body's neurophysiological functions. Extensive research performed by Richard H. Lee of China Healthways Institute in Los Angeles, California, as well as research performed by the Beijing College of Traditional Chinese Medicine, and the Department of Natural Science in Beijing China, confirms that all humans have a very high degree of acoustic activity in the subsonic range below 20 Hertz (infrasonic). This subsonic activity is similar to the alpha rhythm of an EEG (see Appendix #2).

Qigong doctors may combine the characteristic, extraordinary, and infrasonic sound therapies into one treatment modality to facilitate a deeper and more thorough transformation within their patients. The choice of sound prescriptions vary according to each patient's cultural upbringing, religious beliefs, emotional temperament, and comprehension of energetic healing modalities.

Sound therapy can be prescribed to relieve a patient's feelings of uneasiness or extreme depression. Music, for example, has always been acknowledged as a powerful medium for emotional enhancement. Sounds and tones stimulate the cerebral centers creating a wide range of results from enhancement of memory retention, to facilitating certain emotions, and creative inspirations.

Music therapy specifically created for the harmonizing the Five Elements within the Yin and Yang organs is also used in conjunction with Feng Shui training to regulate any imbalances of the body and mind stemming from the changes of seasons. During autumn, for example, when dryness is prevalent and there is a transition from warm to cool, music can help to stabilize the body's response to these external conditions.

WESTERN SOUND THERAPY

Sound therapy is actively being used in Western medicine. The following are examples of current clinical modalities:

- music therapy for postoperative healing,
- ultrasound therapy for sore muscles and back injuries, and
- shock-wave lithotripsy therapy for kidney stones and calcified gallstones.

LIGHT ENERGY

All matter produces, radiates, transports, or transmits light, and therefore has properties of color vibration. Matter and light are fundamentally inseparable. Matter can be considered a condensation of photons. Its transformation can be viewed as a result of light interacting with atoms and molecules. All life on Earth, whether plant, animal, or human, is dependent upon light and its miraculous qualities for existence.

All living cells emit units of light called "biophotons." Biophoton emissions radiate from the internal organs, as well as the body's surface tissues, to create the body's auric field.

This interaction of light on the body's tissues and within the body's internal organs has a profound healing effect, especially when activated by the Qi emission from a Qigong doctor. When muscles or nerves are activated through medical Qigong therapy and / or exercises, the intensity of the biophoton emission increases. Studies performed on the healing potential of light and color therapies demonstrate that it is clinically effective. Like sound resonances, light therapy is also being researched further as an adjunct to modern clinical medicine.

The body both absorbs and projects light energy. The divine light that enters at the time of conception continues to reside in our Taiji Pole throughout our lives. This light responds and reacts to the doctor's emitted therapeutic light energy. Emitted light and color resonance are absorbed into the patient's body, causing the energies of Heaven and Earth to fuse inside the tissues; this facilitates the healing process.

In Traditional Chinese Medicine, the early diagrams illustrating the body's internal organs, channels, routes of Body Fluids, and arteries were called Charts of the Hall of Light. These detailed maps of Qi and Blood flow illustrated the energetic organs, as well as the body's Jing and Shen (which are rooted in the Blood). The Shen flows inside of the body, transforming into light and radiating from inside of the tissues outward. The

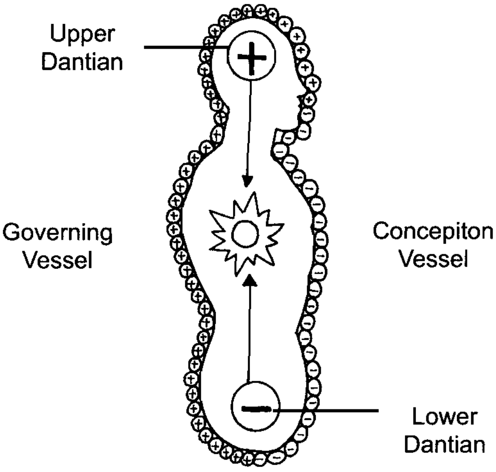

Figure 4.1. Light resonates within the body's Taiji Pole. From the body's Yang Governing Vessel flows a sea of electropositive energy. From the body's Yin Conception Vessel flows a sea of electronegative energy.

light radiating from the physical body spontaneously interacts with the mental, emotional, and spiritual light energy released from other beings.

Light energy can also be accumulated in the storage chambers of the Three Dantians. The Upper Dantian, is called the "peak of Yang energy," and is considered electropositive; it is the doorway to the chamber of Heavenly or divine light. The Lower Dantian, is called the "peak of Yin energy," and is considered electronegative, and the direct root to the Earth.

In time, with Qigong training, the postnatal Heat circulating up the Governing (Yang) and down the Conception (Yin) Vessels can awaken the Prenatal Qi stored in the brain, causing it to unite with the body's nervous system. This is considered a full integration of Yin and Yang energy and manifests as a shimmering white light glowing in the center of the Upper Dantian.

White light is stored in and released from the Upper Dantian, where the spiritual energy is stored. Once white light energy manifests, it should be drawn down the Taiji Pole into the center of the Lower Dantian. By using inner vision to focus the eyes on the center core of the body, the Qigong doctor can concentrate on the white light energy vibration and fuse this energy into one column of light, stemming from the Upper to the Lower Dantian. When the Microcosmic Orbit is completed (i.e., when the Governing and Conception Vessels are energetically fused), the first center column of light can be accessed via the body's five Center Thrusting Channels into the Taiji pole. This column of light runs through the center of the body, connecting all Three Dantians together (Figure 4.1). Accumulated Qi is stored as a golden glowing ball in the Lower Dantian.

COLOR VIBRATIONAL PROJECTIONS

Qigong doctors project light of different colors by first drawing the divine light energy into their Taiji Pole, then focusing on the light energy and releasing it either through the arms and out the hands into the patients, or through a visualization of divine light traveling from the Yin Tang (Third Eye) area into the patients.

When Qigong doctors connect with the divine healing white light energy, they absorb massive quantities of this energy into their Taiji Pole. This white light energy prisms into six color projections of light and energy vibration. As the white light energy transforms into multicolored beams of light, the colors can be either absorbed to strengthen the internal organs, or projected out for Qi emission.

The Qigong doctors use inner vision techniques to connect with the correct color for energy projection. One such technique used by doctors allows the color to naturally develop by mentally focusing both eyes to the back of the head. The color that is observed in the back of the mind is projected into their patients' body. This visualization technique is only performed after connecting with the patients' energetic field.

LIGHT THERAPY THROUGH EMITTED QI

The human body produces light which is visible to both clairvoyants and Qigong doctors (Figure 4.2). In China, doctors of TCM have discovered that light therapy is effective in treating certain diseases because of the light field's ability to penetrate the tissues, and interact with the patients' energetic fields.

Extending the image of the color is an important part of the Qigong doctor's ability to treat

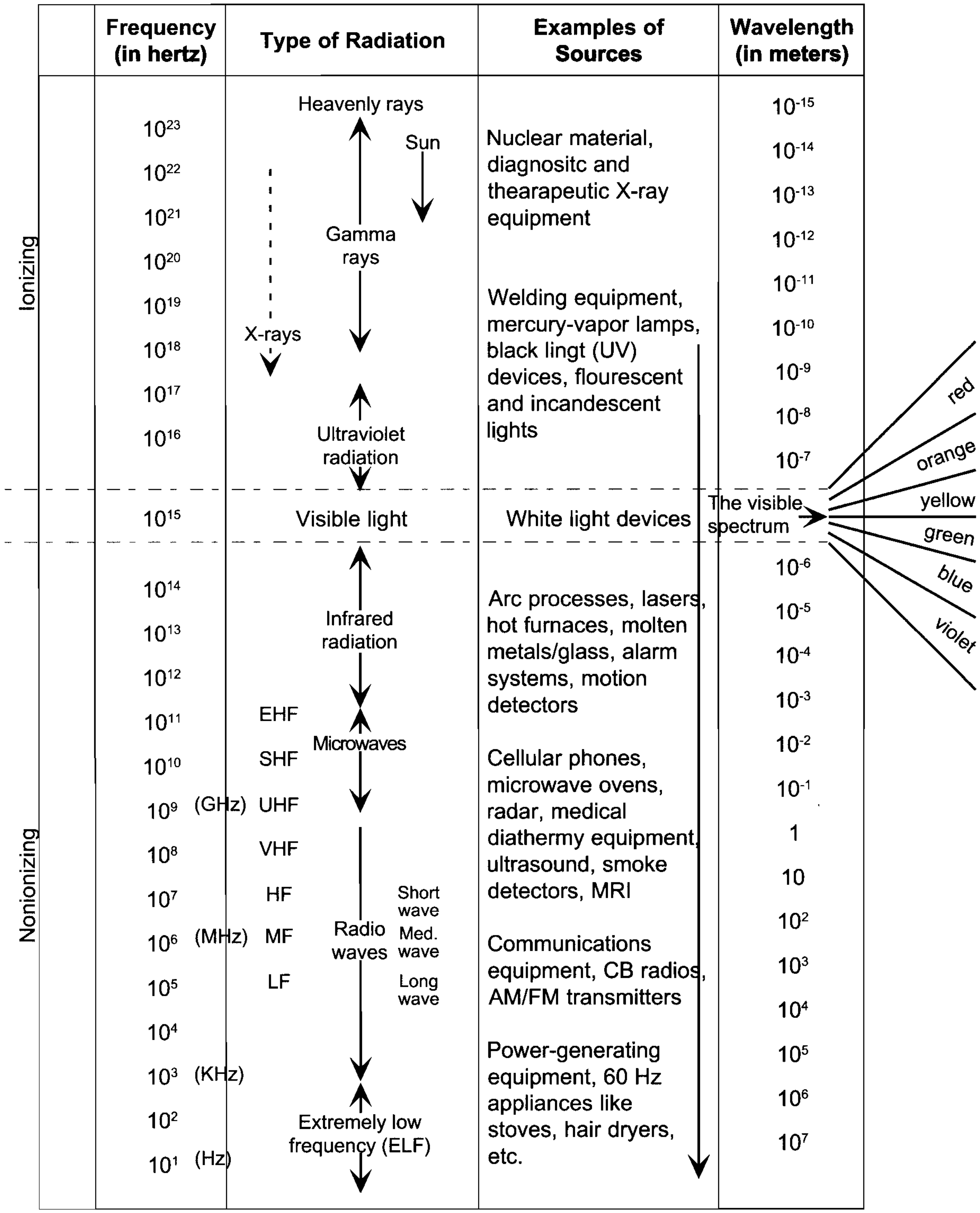
Frequency (in hertz)
Type of Radiation
Examples of Sources
Wavelength (in meters)
Ionizing
Nonionizing
10²³
10²²
10²¹
10²⁰
10¹⁹
10¹⁸
10¹⁷
10¹⁶
10¹⁵
10¹⁴
10¹³
10¹²
10¹¹
10¹⁰
10⁹ (GHz)
10⁸
10⁷
10⁶ (MHz)
10⁵
10⁴
10³ (KHz)
10²
10¹ (Hz)
Heavenly rays
Sun
Gamma rays
X-rays
Ultraviolet radiation
Visible light
Infrared radiation
EHF
SHF
UHF
VHF
HF
MF
LF
Microwaves
Radio waves
Short wave
Med. wave
Long wave
Extremely low frequency (ELF)
Nuclear material, diagnositc and thearapeutic X-ray equipment
Welding equipment, mercury-vapor lamps, black lingt (UV) devices, flourescent and incandescent lights
White light devices
Arc processes, lasers, hot furnaces, molten metals/glass, alarm systems, motion detectors
Cellular phones, microwave ovens, radar, medical diathermy equipment, ultrasound, smoke detectors, MRI
Communications equipment, CB radios, AM/FM transmitters
Power-generating equipment, 60 Hz appliances like stoves, hair dryers, etc.
10⁻¹⁵
10⁻¹⁴
10⁻¹³
10⁻¹²
10⁻¹¹
10⁻¹⁰
10⁻⁹
10⁻⁸
10⁻⁷
The visible spectrum
10⁻⁶
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
1
10
10²
10³
10⁴
10⁵
10⁶
10⁷
red
orange
yellow
green
blue
violet

Figure 4.2. Light Spectrum Chart

and heal patients. To treat patients with emitted light therapy, the Qigong doctor first focuses his or her intention on visualizing a certain color and cultivating the energetic vibrations of this color. By visualizing a certain color, the Qigong doctor is activating the light energy within his or her own body. The doctor then focuses on projecting the healthy color light energy into the organs, channels, and organ systems of his or her patients.

Energy color therapy is divided into eight healing colors, ranging from low frequency (red) to high frequency (violet) and white. These energetic projections are associated with the color ranges emanating from the body's Taiji Pole, and range from a dark to a lighter color band.

When projecting color vibrations, it is important to note that there are four levels, or intensities, of color gradations within each color (Figure 4.3). In the soft color range, there are two levels: a light and a medium-light color used for gentler, less potent projections (usually for healing and cleansing wounds). Medium-light color projections are especially effective for treating young children and the elderly. In the intense range, there are two levels of intensities: a bright and a very bright color range, both of which are used for full-force projection (for destroying cells and pathogenic factors or stabilizing a Deficient organ or organ area). The bright color is used to begin with, later on the very bright color can be used as the patient's strength increases.

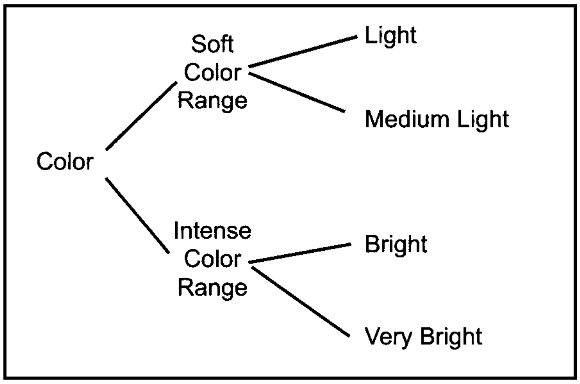

Figure 4.3. The Four Levels of Color Light Projection

Research has demonstrated that the use of color in Qigong treatments can effectively change a patient's pulse beat, body rhythms and depth of breathing. Advanced Qigong doctors modulate and switch energetic color projections during treatment. To avoid overstimulating the patient's tissues, the doctor starts with a lighter shade of color projection then increase the color's intensity until the proper color is obtained.

Medical Qigong uses the following eight colors for color projection: red, orange, yellow, blue, indigo, violet and white. Based on a modern study of the prismatic division of light energy, color projection is categorized as follows.

1. The ruby red color simultaneously heats and activates the body. Red improves blood circulation and stimulates the Small Intestine, Heart, Triple Burners, Pericardium, and Fire Qi of the body. The color red can be used to repair tissues or increase cellular growth, as well as to vitalize tissues, Blood, and the skeletal system. It is used for treating paralysis, broken bones, and internal and external wounds.
2. The color orange expels pathogenic factors and disintegrates Blood clots. It is also used in treating cysts. Orange can also be used to stimulate the Spleen and Stomach.
   (Contraindications: The doctor must avoid projecting orange color Qi into the brain, eyes, or Heart of the patient. The color orange is extremely potent and may cause damage.)
3. The lemon yellow color stimulates the nerves, reconstructs cells, and heals wounds. Yellow or orange yellow can stimulate the Stomach, Spleen, pancreas, digestive system, nervous system, and the Earth Qi of the body. Yellow is also used in the treatment of skin rashes and topical problems.
4. The emerald green color detoxifies and disperses pathogenic Qi. Green can stimulate the Liver, Gall Bladder, and the Wood Qi of the body. Green is used in combination with other colors; for example, shades of green are combined with shades of orange for treating tumors. Green color vibration is not as powerful as orange and may be used on elderly

patients to establish a more gentle healing session.

5. The flame blue color cools the body and is beneficial in soothing and calming a patient's anxiety. Blue can be used to stimulate the Bladder, Kidneys, the reproductive system, skeletal system, and the Water Qi of the body. The color blue is also excellent for relieving inflammation and fever and can be used as a mild anesthetic.
6. The indigo or dark blue color has a deep penetrating property and is used for knitting bones and creating "energetic casts" (used for enveloping wounds and specific internal organs). The color indigo can also be used to stimulate the Bladder, Kidneys, skeletal system, reproductive system, and the Water Qi of the body.
7. The reddish-blue color of violet is known for its purifying force, which facilitates the rapid healing of difficult infections, e.g., pneumonia.
8. White energy extension indicates a fusion of all colors. White energy is the most neutral and can be used when one is not sure which other color is appropriate. White is useful for calming the mind and placing a protective field around an organ or around the body. White stimulates the Large Intestine, Lungs, the respiratory system, and the Metal Qi of the body.

USING HEALING COLOR IMAGERY

Specific color meditations can be prescribed to assist patients in their healing. The patients are encouraged to visualize specific organ colors while the Qigong doctor projects that same color into their body and energetic fields (Wood-Liver-Green, Fire-Heart-Red, Earth-Spleen-Yellow, Metal-Lungs-White, and Water-Dark Blue/Indigo-Kidneys). This treatment technique is successful for tonifying, regulating, or purging all the body's major organs and is often combined with healing sounds.

When using healing color imagery, the patients are encouraged to first visualize the diseased organ as being dull, dark, and impure in color. In cases of Liver Heat, patients might imagine a dull, turbid brownish-green, tinged with red. For a Heart imbalance, the red color may be first imagined as a dull, weak or darkish red, tinged with brownish-red or purplish-red (like the color of stagnant Blood). Patients begin the meditation by concentrating on exhaling the toxic, pathological colors. As the patients inhale, they imagine a pure, vibrant ruby-red color flowing into their body from the Heavens or Earth, energizing, cleansing, and replacing the toxic energetic color of the diseased Heart.

Note: The image of the vibrant ruby-red color is prohibited in cases of Excess Heart Fire.

EXPOSURE TO EXTERNAL COLOR FIELDS

In this type of color therapy, patients are required to sit and meditate in rooms painted in the specific color relating to their condition. A healthy color is chosen in accordance with either the Five Elements' Creative Cycle or a Controlling Cycle, depending upon the nature of the patient's disease and which particular Yin organ (or organs) are involved. One example is using the Five Elements' Creative Cycle to tonify a patient's Deficient organ. The theory of the creative cycle is to stimulate and energize the "mother" in order to strengthen its proceeding "child" organ. For a Liver imbalance, for example, the patient can be placed in a blue room for tonifying the Kidneys (the Liver's "mother" organ). The Kidney color indigo (the mother) is used to nourish the Liver organ (the child). For a Liver Excess, the Qigong doctor would use the Five Elements' Controlling Cycle and place the patient in a white room, allowing the Lungs (the grandmother) white Metal color to control the Liver Wood (the child) organ.

Also specific colors can be worn by the patient, or the patient can surround him or herself with certain natural colors to initiate healing (e.g., emerald green for Liver conditions).

FOOD AND COLOR

Another treatment method requires the presentation of food (diet) as a form of color and light therapy. In this treatment the doctor makes sure that the Five Elemental colors are present within the patient's food. Each color will start a resonance within the patient's organs (i.e., Green-Liver, Red-Heart, Yellow-Spleen, White-Lungs, and Dark Blue / Indigo-Kidneys).

When eating the food the patient focuses his or her intention on ingesting the color, light, and energy of the food being eaten.

USING DISTANCE AND RANGE

When using color projection, an important factor is the doctor's ability to maintain focused intention. If the doctor becomes distracted or looses visual concentration, the color resonance being emitted becomes degraded, and the projected color transforms back into non-differentiated white-light energy. Since it is the vibratory rate of the specific color that promotes healing, a strong connection between the doctor and patient is vital in distance color therapy projection.

WESTERN LIGHT THERAPY

Light therapy is currently being used in Western medicine through the following modalities:

- Laser-light therapy surgery,
- Full-spectrum light therapy for SAD (seasonal affective disorder),
- Color-light therapy for eye problems and various other illnesses, and
- UV-light therapy for "blue babies" with hyperbilirubinemia.

MAGNETIC ENERGY

Magnetism can influence energetic flows of energy current within the body. It is difficult to separate the energetic properties of magnetism (Yin) and electricity (Yang), as they are almost two aspects of the same energetic field. The movement of electrical currents generate magnetic fields. By establishing different magnetic polarities, the Qigong doctor can influence the electrical flow in the body.

The body's cells carry electromagnetic fields that both attract and repel each other, resulting in a magnetic pull between the body's tissues, organs, and extremities, as well as the body's electromagnetic interaction with the Earth. Clinical research in China demonstrates that the iron in the hemoglobin of the Blood is attracted to the magnetic pull of the body's tissues, as well as the magnetic pull of the Earth.

The Earth, like the body, has both a Yin (magnetic) field and a Yang (electrical) charge. These energetic fields are interdependent. The magnetic

Figure 4.4. The Earth is enveloped with electromagnetic currents, affecting both weather and vegetation.

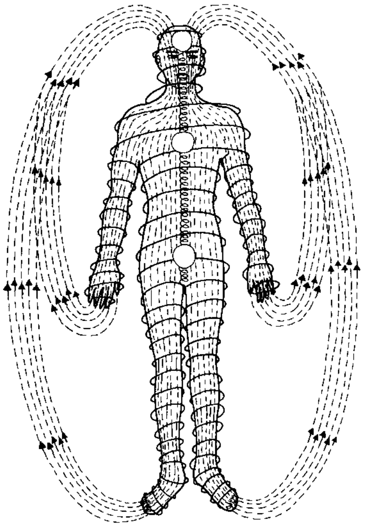

Figure 4.5. The body is enveloped with electromagnetic currents, affecting both internal and external organ functions.

field is caused by the movement of charged particles and spans the circumference of the Earth creating a strong electrical field (Figure 4.4).

Some modern researchers believe that the physical body's channel system (see Chapter 6) charges its field of magnetic flux through movement. It is further believed that the channel system creates an electrical field that attracts electrons into the body's Taiji Pole.

Richard Lee of the China Healthways Institute found that the Yin magnetic substances stored within the body's water molecules align with the electrically conducting structure of the body's channels (see Appendix 2). When the body's magnetic energy field is low, these channels become weakened and the conductivity is lowered. This magnetic energy can be replenished either from ingesting and transforming food, air and water, or gathered directly from the Earth's magnetic field.

The body's magnetic field conforms to the Earth's magnetic field through the energetic exchange within the body's electromagnetic fields (Figure 4.5). The body maintains this electromagnetic connection through the energetic resonance of the Taiji Pole. Within the structure of the body's Taiji Pole are located two major conductive poles of energy. These two conductive poles are located at the opposite ends of the Taiji Pole, positioned at the top of the head and the base of the perineum. The purpose of these energetic poles is to absorb Qi from universal and environmental fields, connecting and integrating the energy into the body's Three Dantians (see Chapter 5).

The electromagnetic lines in the body's force field begin from the at top of the head (traditionally considered the south pole) where the Heaven Qi flows into the body, and end at the at the base of the perineum (traditionally considered the north pole) where the Earth Qi flows into the body (Figure 4.6). Each of these two magnetic poles (the Lower and Upper Dantian) have a different energy influx. The energy originates and converts in the Lower Dantian and eventually flows to the Upper Dantian. The bottom pole, located in the Lower Dantian, converts Jing (Essence) into Qi (Energy) and increases the body's overall life-force energy. The upper pole, located in the Upper Dantian, converts Shen (Spirit) into perceptual insight and spiritual light.

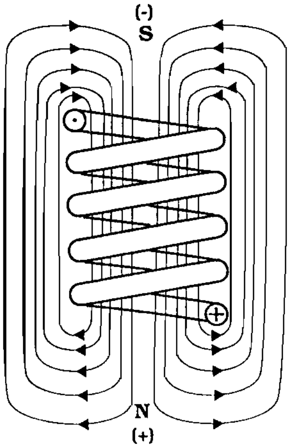

Figure 4.6. Magnetic field around a spiral coil

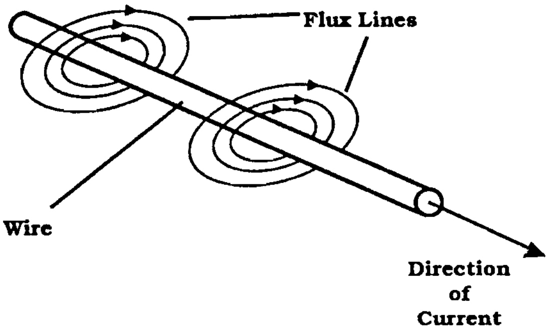

Figure 4.7. Flux lines around current-carrying conductor

There are several ways in which the Qigong doctor can manipulate the electromagnetic field which surrounds the body. The doctor can directly absorb Heaven and Earth energy from the bottom of his or her feet, top of the head, and through both palms (called Absorbing Qi from the Five Gates). The doctor can also, through creative visualization, gather and loop the different types of environmental energy around his or her body, increasing the thickness and power of the electromagnetic field (the energetic boundary that surrounds the doctor's body). This looping action is similar to wrapping a magnet with wire to increase its magnetic field potential (Figure 4.7). Stronger lines of magnetic force will be produced around the body, thus creating a stronger "energy bubble" (Figure 4.8).

The direction of electromagnetic wrapping should be done in accordance with the natural flow of the surrounding environmental fields. Each individual should remain sensitive as to which technique works better for them, as there are individual variations used in accordance to the doctor's intention and belief.

MAGNETIC THERAPY THROUGH EMITTED QI

Qigong doctors are trained to manipulate their patient's electromagnetic field by extending emitted energy into the patient's tissues. This stimulation of the body's magnetic field is extremely effective for regulating the flow of channel energy, as well as stimulating the activity of the nervous system.

Magnetic therapy is also used as a companion to acupuncture therapy. Acupuncturists use magnetic patches, placing them on various channel points on the patient's body, to induce energy flow and tissue stimulation.

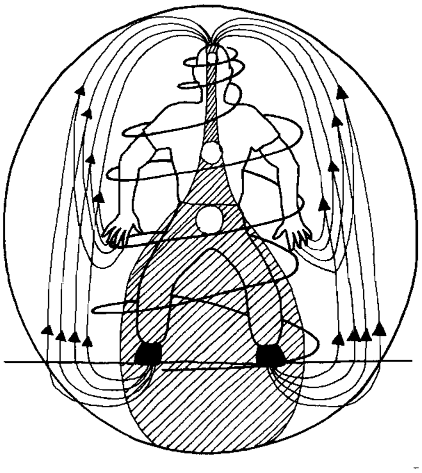

Figure 4.8. Qigong doctors can create a strong energetic field by absorbing Qi into their Lower Dantian. This energy extends into the Earth (like an anchor), and the Wei Qi field surrounding the doctor's body automatically increases and expands.

The negative north magnetic field pole of the magnet is placed in contact with the patient's skin in treating conditions due to Excess Heat and stagnation. The negative north pole is found to have a cooling and calming effect on tissues. Herbal teas used for purgation, dispersing, anti-inflammatory or sedative properties can be augmented by placing the tea on the north pole of a magnetized surface for a minimum of an hour before drinking.

The positive south magnetic field pole of the magnet is placed in contact with the patient's skin in treating Deficiencies, Cold and blockages. The positive south pole is said to have a stimulating and warming effect on tissues. Herbal teas used for warming and tonification can be augmented by placing the tea on the south pole of a magnetized surface for a minimum of an hour before drinking.

WESTERN MAGNETIC THERAPY

In 1990, Dr. Arthur Trappier submitted a paper on "Evaluating Perspectives on the Exposure Risks from Magnetic Fields" to the *Journal of the National Medical Association*. In his article, Dr. Trappier explained that exposing cancer cells to a negative magnetic field discourages the growth of the cancer, while exposing cancer cells to a positive magnetic field encourages their growth. Magnetic therapy is currently being used in Western medicine through the following clinical modalities:

- Permanent magnet therapy for localized pain and inflammation relief,
- Super-magnet therapy for stimulation of the thymus for immune enhancement in cancer patients, and
- Pulsed magnetic therapy for the treatment of arthritic joints.

GENERATED HEAT

Electricity passing through any substance will produce heat. The amount of heat that will be generated depends upon the resistance of the substance and the density of the current's flow. Heat is generated at the electron level by the friction created through molecular motion. By increasing the motion of the molecules, more heat is generated. Molecules in living organisms are in constant motion, increasing or decreasing their rate of acceleration according to the environmental temperature. In order for heat to have any effect on the body, enough Qi has to be transferred to increase the movement of the molecules.

The heat within the body causes increased electron movement, resulting in more heat being generated on a cellular level. Heat is also caused by cellular metabolism. The body's metabolism produces 75% of the energy being created in the form of heat. Cellular activity is increased through electrical, magnetic, heat, sound, and light stimulation.

From a Medical Qigong perspective, Heat in the body is generated from the accumulation of Three Fires which emanate from specific locations within the body. Within the body's chest cavity is the Heart Fire, located within the abdominal cavity is the Kidney Fire, and all through the body flows the Bladder Fire. When the Heart Fire first awakes, the Kidney Fire responds to it, and when the Kidneys Fire moves the Bladder Fire follows it.

When these Three Fires follow their normal course of energetic movement, they issue and lead the body's life-force energy, creating and sustaining life. The Three Fires are responsible for regulating the Yin and Yang energy of the body by fusing the Five Elemental energies (stored within the body's Wood, Fire, Earth, Metal, and Water organs) with the energy of the Three Dantians. As the Shen from the Heart Fire is drawn into the Lower Dantian, the Bladder Fire fuses with the Kidney Fire, creating the body's True Fire. This action causes the body's Jing to create Qi in the Lower Dantian and then transform Qi into Shen in the Middle Dantian. Once this fusion is obtained, the mind, breath and body connections all become regulated.

The Three Fires also represent the regions of vital Heat responsible for the circulation of energy that sustains the Eternal Soul. Accessing the energy of the three Fires is used for cultivation and spiritual liberation, and is brought about through Qigong practice, prayer and meditation.

1. The Heart Fire, also called the Commanding or Emperor's Fire, is located in the center of the chest. The Heart Fire is responsible for transforming the body's energy into Shen.
2. The Kidney Fire or Mingmen Fire, is located in the back of the body, just below the last floating rib. Qigong masters in ancient times regarded the Mingmen Fire as the motivating force of the body and paid special attention to its training during Qigong exercises. A Deficiency of the Mingmen Fire may lead to decreased sex-drive, hypogonadism and impotency. Conversely, if the Mingmen Fire is in Excess, increased sex-drive or sexual obsession and hypergonadism will occur.

   Dr. Zhao Xianke, an expert on medicine during the Ming Dynasty period, states that "the Mingmen Fire dominates all Twelve Primary Channels. Without it the Kidneys would be weak, the Spleen and Stomach could not digest food, the Liver and Gall Bladder would not give any energy to think or plan, the urine and feces would not be moved, and the Heart would malfunction causing dizziness and endangering life."
3. The Bladder Fire, also called the Common People's Fire, is located in the lower abdominal area by the perineum and is responsible for evaporating water.

The Three Fires are not the Triple Burners, as the purpose of the Triple Burners is to regulate the major internal organs and is a completely different energy system. The physical locations of the Three Fires however, coincides with the locations of the Triple Burners.

HEAT THERAPY THROUGH EMITTED QI

Stimulating the body's tissues through Heat therapy is extremely effective for treating Deficient and Cold syndromes. In China, Qigong doctors use the extension of Heat through their emitted field of electromagnetic energy to stimulate their patients' tissues. This therapy helps to regulate the flow of the patients' channel energy, and tonifies the Blood and Qi.

Heat therapy is also used in the form of moxibustation as a companion to acupuncture therapy. Acupuncturists will use moxa sticks and/or cones, lighting and placing them on or over various channel points of the patients' body. This is practiced to induce Qi flow and tissue stimulation.

WESTERN HEAT THERAPY

Radiant and conductive heat therapy are currently being used in Western medicine for vasodilation and pain relief. These heat therapies can further be broken down and categorized into dry-heat therapy and moist heat therapy.

Dry-heat therapies include the following:

- Dry packs
- Hot-water bottles
- Heliotherapy-sun therapy
- Ultraviolet-heat therapy
- Infrared-heat therapy
- Diathermy therapy

Moist-heat therapies (hydrocolators) include the following:

- Hot bath packs
- Hot wet packs
- Hot foot baths
- Fomentations
- Poultices
- Vapor and paraffin baths

ELECTRICITY

A flow of electrons is called a current. Just as a current produces a magnetic field, a magnetic field, when it moves in relation to a conductor, induces an electrical current. In Medical Qigong therapy, the body's channels are also considered electrical circuits, and the points existing within each channel can be considered booster amplifiers (or step-up transformers) that maintain the current's strength. According to Dr. Robert O. Becker, research scientist and author of the book *Cross Currents,* acupuncture needles have the capacity to act as antennae, drawing charged particles (ions) from the atmosphere into the body. The acupuncture needle delivers a low-level electrical stimulation to the channel points and can be used to charge up, or decrease the energetic potential of these "step-up transformers," affecting the current along the channel.

The human body is an excellent conductor of electricity and contains both alternating and direct electrical currents. The alternating current is responsible for the transformation of the body's magnetic field into the body's electrical field. The direct current is responsible for the body's positive Yang and negative Yin flow of electrical charges.

The friction produced by rubbing the feet on a carpet creates a charge that is stored within the body until touching another conductor (metal, another person, etc.) to release the charge. This normal and common occurrence demonstrates the storage, conductance, and discharging of electrons from the body. Any time electrons travel, heat, or thermal radiation is produced.

According to Swedish Physician and medical research scientist Dr. Bjorn Nordenstrom, the biological circuits of the body are driven by the accumulated charges, which, unlike a battery, oscillate between positive and negative. The body's system of channels and blood vessels act as insulated cables, while the blood plasma acts as the conductor. In the permeable tissue, the fluid between the cells conducts ions. A key component of the body's electrical circuit is the natural electrodes in the capillary walls, known as the lipid bilayers of the electron transport chain.

A myriad of electrical forces work within the body. Every human thought and action is accompanied by the conduction of electrical signals along the fibers of the nervous system. In fact, life would not exist at all without a constant flow of ions across the membranes of cells.

The electromagnetic energy in the body's cells is continuously being generated through the biochemical transformation of food, and air, and is circulated by the electromagnetic fields being generated within the tissues. Perineural cells, or nerve sheaths, carry the direct current of the body's electricity. These cells are responsible for motivating the body to heal, regenerate, and repair itself. Healing is always affected by a change within the body's electromagnetic field. The rate and efficiency of healing is based on the strength and polarity of the body's field of energy. This electricity is one of the primary energy sources responsible for maintaining life itself.

Dr. Becker's research demonstrates that bones are "piezoelectric": when stressed, mechanical energy is converted into electrical energy that produces an increased electrical current. Dr. Becker discovered that running a minute electrical current through a fractured bone will stimulate the reproduction of the cells, creating a healing current similar to the body's natural healing mechanism.

According to Richard Lee of China Healthways Institute, electrostatic waves are observed in the body through EEGs (electro-encephalograms) and EMGs (electro-myograms). Different types of therapies affect the EEGs (sound therapy, meditation therapy, self-regulation Qigong therapy, light therapy, etc.). By controlling their thoughts, Qigong doctors can affect the current in their body and thus affect the EEG measurements. These thought patterns will in turn influence the Qigong doctors' electrical field. Experiments show that EEG measurements in test subjects receiving energy from healers tend to synchronize. Qigong doctors, when treating patients, can produce voltages as high as 190 volts, 100,000 times greater than regular EEG voltages.

The body's natural ability to gather, store, and move Qi increases the body's abilities to gather, store, and move electrical charges. The gathering of these electrical charges can cause a gradual electric buildup within the tissues and internal organs, developing into an Excess Qi condition within the body. Excess Qi conditions often manifest mood swings and are responsible for the energetic circuit overloads that occur spontaneously within the body. Excess Qi manifests as a sudden "explosion" or "release" of emotions such as rage, fear, grief, worry, fright, anxiety, and joy. An excess electromagnetic charge, internally combined with the Qi from a patient's on-going emotions can be drawn into an organ's tissue area and absorbed by the internal organ that is in the most elevated state of energy conversion. This energetic reaction can cause the already overcharged internal organ to spontaneously release emotions. Sometimes the release is physiological, e.g., sudden sweating, blushing, twitching or jerking, yawning, stretching, sighing, burping, passing gas, etc. These can all be ways that the organ seeks to regain hemostasis.

ELECTRICAL THERAPY THROUGH EMITTED QI

The idea that an electric current can stimulate bodily repair, alert defence mechanisms, and control the growth and function of cells is not a new concept in Western medicine. In fact, the use of bio-electromagnetism dates back at least 200 years. Electrotherapy is found to be very useful in relieving pain by signaling the brain to activate and alter the body's neurochemicals. The insertion of electrically stimulating needles into a patient's body, for example, causes a release of endorphins, and is used for pain management.

Medical Qigong, acupuncture, and Chinese massage likewise, stimulate the peripheral and cutaneous nerves that carry sensory information, via the spinal cord, to the brain. This stimulation of the cutaneous nerves activates the brain's opiates (endorphins - endogenously generated morphine) and facilitates the closure of the body's pain relay gates, killing pain, and is the basis for anesthesia. Because medical Qigong therapy also has an analgesic affect on the body's cutaneous tissues, it is being increasingly used in hospitals for pre- and postoperative procedures.

WESTERN ELECTROTHERAPY

Electrotherapy is currently being used in Western medicine through the following modalities:

- Giga-TENS therapy for stimulation of healing,
- TENS therapy for pain relief,
- CES-cranial electro-stimulation therapy-for depression and substance abuse, etc., and
- Micro-stimulation therapy for micro-current stimulation below the threshold of awareness, to stimulate nonspecific healing, the reduction of inflammation and the harmonization of tissue polarity.

CHAPTER 5
THE THREE DANTIANS AND THE TAIJI POLE

THE THREE DANTIANS

According to Chinese energetic physiology, humans have three important energy centers, located in the center core of the body, that store collected energy in much the same way as a battery does. These three centers are called the Three Dantians (Figure 5.1). The word Dan literally translates to mean "cinnabar," while the word tian means "field." In terms of Medical Qigong tian is referred to as a field of energy. Together, the characters for Dantian literally translates as "cinnabar field" or "field of elixir."

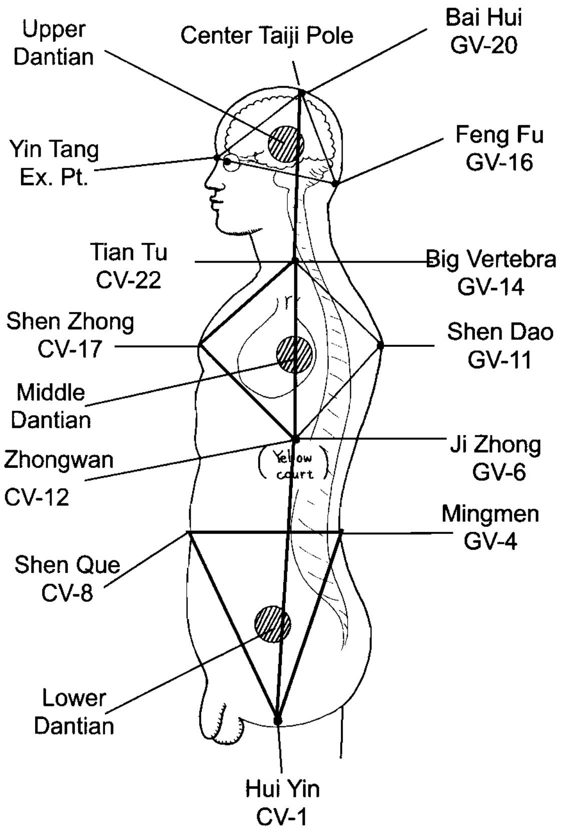

Figure 5.1. The Anatomical Locations of the Three Dantians and Center Core (Taiji Pole)

Cinnabar is mercury sulfide, a bright red or vermillion mineral, which is the principal ore of mercury. Cinnabar was an important mineral in ancient China, the source material needed to manufacture vermillion ink, which could be used only by the emperors. Cinnabar was -- and is today -- also used in Chinese medicine to sedate the Heart and calm the Shen, but always in small doses for short periods due to its highly toxic nature.

Cinnabar was also a vital elixir in Daoist alchemy, as it was discovered to be very balanced in its Yin and Yang properties. Daoist alchemy, like Western alchemy, was practiced in two ways: outer alchemy (Wai Dan) and inner alchemy (Nei Dan). Outer alchemy was the ancestor of modern chemistry. Outer alchemists set up laboratories and experimented with many substances from mineral, animal, and plant sources with the goal of discovering how to turn base metals into gold. Secretly, they were also seeking to discover an elixir, or drug, that would confer immortality or, at the least, greater longevity. Along the way they made many important discoveries, such as gunpowder, medicines, and many other substances vital to the world today.

Inner alchemy was concerned with purifying human nature and transforming the spirit into its most pure and radiant potential without the use of outer agents. Instead, they used Qigong and meditation to circulate and gather the "inner elixirs" of sexual energy, Qi, and consciousness at various locations within the body. The Three Dantians, or "elixir fields," in addition to storing energy, act like an alchemist's crucible, or cauldrons in the role of gathering and transforming vital substances, energies, and various elements of awareness. The ingredients of the inner elixir are Jing, Qi, and Shen. The inner alchemists kept their work secret by using mineral code words such as "gold," "lead," and "cinnabar" to describe the movement and transformation of energetic and spiritual substances within the body.

The goal of the inner alchemists is immortality, a complete transformation of Jing, Qi, and Shen. Jing, Qi, and Shen are collectively called the Three Treasures and are the three fundamental energies necessary for human life. To accomplish this transformation, alchemists first gather and transform Jing into Qi in the Lower Dantian. They then gather and transform Qi into Shen in the Middle Dantian. Next, they transform Shen into Wuji (the absolute openness of infinite space) in the Upper Dantian. Finally, they merge Wuji into Dao (divine energy). The Three Dantians serve as the inner crucibles for these transformations.

These transformations can be compared to the changes of water (the consistencies and not the temperature), which when heated, can change from solid ice, to a liquid, and then to vapor.

The Three Treasures (Jing, Qi, and Shen) are also connected with the Three Outer Forces or Powers known as Heaven, Earth, and Man. Jing (reproductive essence) is the most substantial and therefore the most Yin of the three, and it is closely linked with Earth Qi. In Medical Qigong practice and Daoist inner alchemy, the Earth energy is gathered in the Lower Dantian and is associated with heat. Qi is a mixture of Yin and Yang and is closely connected with the atmospheric energy, which is a blend of Heaven and Earth Qi and gathered into the Middle Dantian, (which is associated with vibration). Shen (Spirit) is the most insubstantial and, therefore, the most Yang of the three; it corresponds with Heaven Qi, and is gathered in the Upper Dantian which is associated with light.

The Three Dantians are connected to each other through the Taiji Pole. The Taiji Pole acts as a passageway for communication between the Three Dantians, and as a highway for the movement of the various life-force energies. The Eternal Soul is drawn into the body at the moment of conception through the Taiji Pole, and departs through it at death. The Taiji Pole also serves as a transport for the body's Hun.

Regardless of whether or not a modern Qigong practitioner is concerned with spiritual transformation, from a Medical Qigong standpoint the Three Dantians are still vital centers for the cultivation of energy. They are important areas for diagnosis and self-healing, as well as for projecting Qi to patients.

The following is an exploration of the Three Dantians and their individual relationships to the Three Treasures of Jing, Qi, and Shen.

THE LOWER FIELD OF ELIXIR

The Lower Dantian is the Dantian most familiar to martial artists and Zen meditators. It is regarded as the center of physical strength and the source of stamina. Called the "Hara" in Japanese, it is located in the lower abdomen, in the center of the triangle formed by drawing a line between the navel, Mingmen (lower back), and perineum. These three points form a pyramid facing downward. This configuration allows the Lower Dantian to gather the energy from the Earth.

The Lower Dantian is the major storage area for the various types of Kidney energies. The Kidney energies are, in turn, closely linked with our prenatal energies and provide the foundation for all other types of Jing, Qi, Yin, and Yang energies in the body.

The Lower Dantian is connected to the first level of Wei Qi (Protective Qi) circulating outside the body, extending from the body's tissues to about one inch. As the Lower Dantian fills with Qi, the Wei Qi field naturally becomes thicker.

THE LOWER DANTIAN AND JING

The Lower Dantian collects Earth energy and represents the body's physical energy of Jing. The Earth energy that is transformed in the Lower Dantian is a dense, full energy with a thick quality to its texture. In our analogy with the transformations of water, the energy in the Lower Dantian is dense, like ice.

The Lower Dantian is closely linked to the Jing Gong (Essence Palace) located in the perineum, which serves as a reservoir of Jing. Our Prenatal Essence (Yuan Jing), determines our constitutional strengths and vitality, and is stored in the Lower Dantian. It interacts with the Kidney energies to form Kidney Jing. The Kidney energies are all closely intertwined: Kidney Jing, Kidney Qi, Kidney Yin, Kidney Yang, and Kidney Fire. The Mingmen Fire, also called Kidney Yang, helps transform the Jing into steam (Kidney Qi).

Kidney Jing circulates throughout the body via the Eight Extraordinary Vessels, in particular, the Governing, Conception, and Thrusting Vessels, all of which originate in the Lower Dantian.

Kidney Jing controls the reproductive energies and life cycles in the body. Some of the ancient alchemical texts describe the Lower Dantian in women as being located in the Bao or uterus, and it is related to the function of Jing in a woman's body. In men, the reproductive essence is located in the Jing Gong (Essence Palace) or prostate and seminal vesicles. In some Chinese medical, Daoist, and Qigong literature, the term Kidneys is used as a synonym for the testes and ovaries.

The location of the Jing Gong is affected by the different anatomical locations of the male and female reproductive organs. In men, this area is located in the center of the body, at the level of the superior border of the pubic bone, posterior to the Qugu CV-2 (Crooked Bone) point. The Jing Gong area in women is located higher, centered in the uterus, about an inch above the superior border of the pubic bone, posterior to the Zhongji CV-3 (Utmost Center) point.

This difference in location, in turn, affects the storage of Jing (i.e., the testicles in the male cause the transformation of energy to occur lower in the body than in females, due to the higher position of the woman's ovaries).

Jing is the most physical, material form of Qi within the body (corresponding to Yin and Earth energy). The Lower Dantian is the place where Qi of the Earth is drawn into the body and transformed by heat.

The Lower Dantian acts as a reservoir for heat and energy and is associated with the Kidneys. The Kidneys control the Water element in the body. Jing is said to be like the water in the cauldron. Through focused concentration and meditation, the Jing (Essence) in the Lower Dantian is refined and transformed to produce Qi (Energy). When heat is generated as a result of the Heart and Mingmen Fire mixing with the Kidney Water, the alchemical transformation of Jing in the Lower Dantian area transforms Jing into the steam of Qi (the character for Qi is composed of steam rising from a cooking pot). This alchemical transformation is known as "changing Jing into Qi," and takes place within the Lower Dantian.

THE LOWER DANTIAN AND QI

The Lower Dantian is often called The Sea of Qi. It is the place where Qi is housed, the body's Mingmen Fire is aroused, the Kidney Yin and Yang Qi is gathered, and the Yuan Qi is stored. Also called Source Qi, the Yuan Qi is the foundation of all the other types of Qi in the body. The Yuan Qi is closely linked with the Prenatal Essence (Yuan Jing). Together, the Yuan Qi and Yuan Jing determine our overall health, vitality, stamina, and life span.

The Yuan Qi is the force behind the activity of all of the organs and energies in the body. It is closely related to the Mingmen and works to provide body heat. The body's Yuan Qi is the catalytic agent for transforming the food we eat and the air we breath into Postnatal Qi. It also facilitates the production of Blood.

Yuan Qi is housed in the Lower Dantian, and it also flows out to the internal organs and channels through the Triple Burners. Yuan Qi is said to enter the Twelve Primary Channels (the body's twelve major energy pathways) through the Yuan points (sometimes called Source points) in acupuncture theory.

Of the Three Dantians, the Lower Dantian is closest to the Earth, is the most Yin, and has the strongest ability for gathering Earth Qi. In Medical Qigong, once students have learned to conserve and circulate their own Qi, they can increase it by connecting to the unlimited reservoirs of Qi in the natural environment. Earth energy is the first form of external Qi with which the Qigong practitioner connects. This energetic connection to the Earth is important for several reasons.

1. First, Qigong practitioners need the Yin grounding power of the Earth Qi to counterbalance the gathered active Yang energy from Qigong exercises. Without grounding in Earth Qi, many Qigong practitioners develop Qi deviations in the form of Excess Heat.
2. Second, each person's supply of Qi is limited. When Qigong doctors extend their Qi to heal others, they deplete their personal supply of Qi unless they are able to simultaneously replenish their supply from outside sources.

Even people who do not practice Medical Qigong draw Earth Qi into their Lower Dantians. By practicing Qigong and using conscious intent, the amount of Earth Qi drawn into the body is vastly increased.

THE LOWER DANTIAN AND SHEN

The Lower Dantian itself is considered a center of consciousness. The consciousness of the Lower Dantian is more kinesthetic or physical given its Yin nature and close proximity to the Earth.

The body's Jing is connected with the Po (Seven Corporeal Souls). The Po control our survival instinct and the subconscious physical reflexes associated with survival. For this reason, Oriental martial artists spend many hours cultivating their lower elixir field to have the integration of Jing, Qi, and Shen needed for the split-second clarity of focus demanded in life-and-death struggles.

THE LOWER DANTIAN AND KINETIC COMMUNICATION

Not only is the Lower Dantian the center of physical strength and the source of stamina, but it is also considered the "house" of physical (kinetic) communication, awareness and feelings. The level of awareness referred to as "the intuition of the physical body," or kinetic communication, is stimulated by the subconscious. The subconscious mind picks up many signals from the environment that are not processed by the logical mind. These signals can motivate spontaneous body movements and responses and are sometimes referred to as gut feelings.

Kinetic means "characterized by movement." It is usually felt as a movement in the body or of the body. It is in this kinetic state of awareness that allows Qigong doctors to naturally feel the patient's internal resonant vibrations within their own body. When the doctors' body suddenly feels hot or cold, starts shaking or trembling, this may indicate that their subconscious mind is trying to communicate the location and condition of the diseased area within the patient's tissues.

Most of the time, the feelings experienced in the Lower Dantian are very subtle, but Qigong doctors, with a heightened degree of awareness of their own body, are able to pick up subtle variations of energetic shifts within themselves and others. When doctors collect energy in the Lower Dantian, an increased awareness and sensitivity naturally occurs. Cultivating this ability simply requires practice in paying attention to the physical body. A high level of awareness of the physical body, the surrounding environment, and the relationship between the two is required to maximize kinetic communication. When awareness is increased, perceptual feeling and kinetic body movements happen naturally. These are subtle senses that allow Qigong doctors to feel, smell, or hear energetic phenomena as they are released from the diseased tissues of patients.

According to research conducted by Dr. Michael Gershon, a professor of anatomy and cell biology at Columbia Presbyterian Medical Center in New York, the Lower Dantian sends and receives impulses, records experiences, and responds to emotions. Its nerve cells are bathed in and influenced by the same neurotransmitters as the brain. The Lower Dantian's "brain," known as the enteric (intestinal) nervous system, mirrors the body's central nervous system and is a network of 100 million neurons (more then the spinal cord contains), neurotransmitters, and proteins that can act independently of the body's brain, and can send messages, learn, remember, and produce feelings.

Dr. Gershon explains that active within the neural system of the lower abdominal area are major neurotransmitters like serotonin, dopamine, glutamate, norepinephrine, nitric oxide, enkephalins (one type of natural opiate), and benzodiazepines (psychoactive chemicals that relieve anxiety). The lower abdomen also has two dozen small brain proteins called neuropeptides. Dr. Gershon's research results provide modern scientific verification of what Eastern wisdom has taught for millennia -- that centers of consciousness exist at places in the body besides the brain and that the abdomen is one of the body's major centers of awareness.

ANATOMICAL LOCATION OF THE LOWER DANTIAN

The Lower Dantian is centered below the umbilicus, inside the lower abdomen, forming a downward pointing triangle. It occupies the first three lower Chakra gates. Chakras are small energy centers that originate from the Taiji Pole and have one or more gates.

1. The lowest point of the Lower Dantian extends to the Huiyin CV-1 (Meeting of Yin) point at the perineum. The name refers to the area on the body responsible for gathering and absorbing the Earth energy. This area is responsible for gathering the Yin energy into the body and Lower Dantian area via the three Yin leg channels (Liver, Spleen, and Kidney). This area is sometimes known as the Lower Gate of the Taiji Pole, or the Bottom Gate of the Lower Chakra.
2. The front area of the Lower Dantian is located posterior to the Shenque CV-8 (Spirit's Palace) point at the navel. The name refers to the place where the mother's Qi and Shen enter the embryo during fetal development. This area is sometimes known as the Front Dantian, or the Front Gate of the Second Chakra.
3. The back area of the Lower Dantian is located at the Mingmen GV-4 (Gate of Life) point on the lower back, anterior to the second lumbar vertebra. The Mingmen occupies the place between both Kidneys. It is the root of Yuan Qi, and therefore determines life and death. The Mingmen provides the body's True Fire; it supplies the heat for the Triple Burners, and is responsible for stabilizing the Kidneys and Lower Dantian area. This area is sometimes known as the Back Dantian, or the Back Gate of the Second Chakra.
4. Medical Qigong schools in China differ in their belief as to where the center of the Lower Dantian is located. Some schools teach that the center of the Lower Dantian is affected by the different anatomical locations of the male and female reproductive organs. In these particular schools, the students are taught that in men, the center of the Lower Dantian is located posterior to the Guanyuan CV-4 (Gate of Original Qi) point. The center of the Dantian area in a women is said to be located internally higher, and posterior to the Qihai CV-6 (Sea of Qi) point. This area is sometimes called the Middle of the Dantian, referring to its position between the navel and Mingmen areas.

THE NINE CHAMBERS OF THE LOWER DANTIAN

The human body is viewed as a microcosmic replica of the power of Heaven. Just as Heaven is said to be divided into nine different levels, each containing various palaces, our bodies also are said to contain a large variety of palaces and chambers. The nine stars of the Big Dipper are said to correspond to the Nine Chambers of each of the Three Dantians.

The functional aspects of the body's psyche were described by ancient Qigong masters as "spirits" that lived within the nine chambers of the Lower Dantian. These "spirits" linked the body's energetic channels and vital internal organs into an organic harmony of life-force energy. Qigong masters believed that as each Dantian became energized, it would initiate specific reactions within the body's energetic system, causing certain energetic / spiritual awareness to manifest within the practitioner's psyche.

Each of the nine chambers is several inches in diameter and is numbered in accordance to its en-

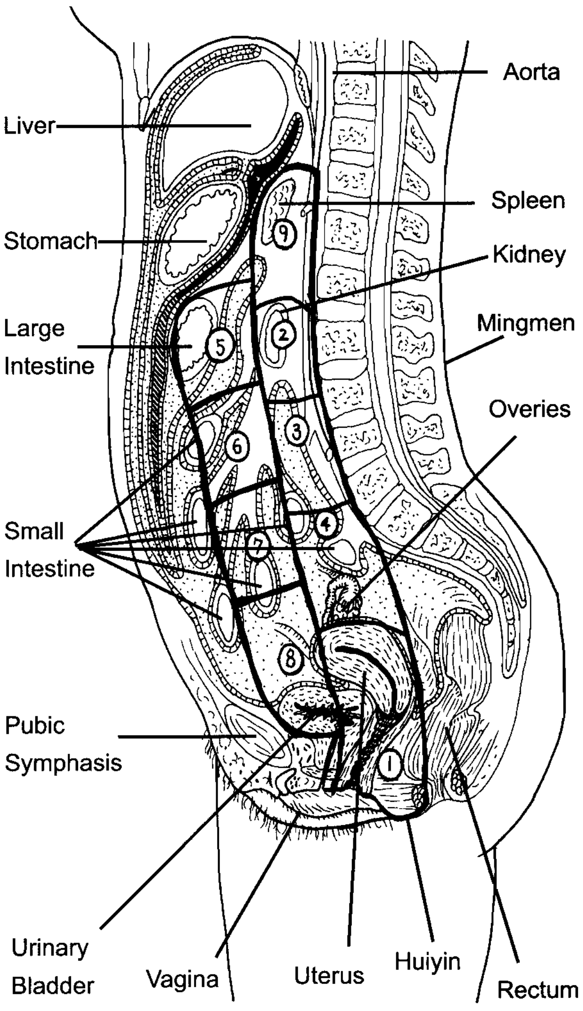
Aorta
Liver
Spleen
Stomach
Kidney
Large Intestine
Mingmen
Overies
Small Intestine
Pubic Symphasis
Urinary Bladder
Vagina
Uterus
Huiyin
Rectum
1
2
3
4
5
6
7
8
9

Figure 5.2A. The Nine Chambers of the Lower Dantian are shown here in the female body. Each number encompasses the entire chamber.

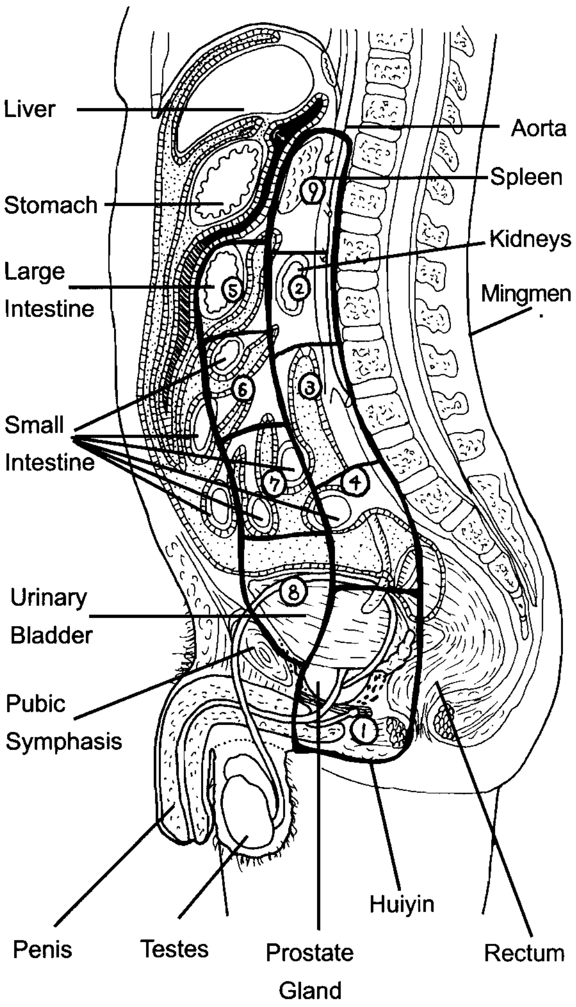
Liver
Aorta
Spleen
Stomach
Kidneys
Large Intestine
Mingmen
Small Intestine
Urinary Bladder
Pubic Symphasis
Penis
Testes
Prostate Gland
Huiyin
Rectum
1
2
3
4
5
6
7
8
9

Figure 5.2B. The Nine Chambers of the Lower Dantian are shown here in the male body. Each number encompasses the entire chamber.

ergetic stimulation (Figure 5.2 A-B). This vertical abdominal set of Lower Dantian chambers relates to many cavities of the body's internal viscera. The Nine Chambers of the Lower Dantian are named as follows.

1. The Palace of Jade (Jade Stem or Jade Cave)
2. The Official Health Monitor (Kidneys)
3. The Minister of the Orchard Terrace
4. The Chamber of Moving Pearls
5. The Minister of House Cleaning (Large Intestine)
6. The Palace of Mystical Spirits (Small Intestine)
7. The Chamber of Mysterious Elixir
8. The Spirit of the Jade Court (Bladder)
9. The Spirit of the Yellow Court (Spleen)

ENERGETIC FOCUS OF TRAINING THE LOWER DANTIAN

All Qigong training begins with the focus on the Lower Dantian. In the beginning stages of Medical Qigong training, the doctor will encourage students to focus their mind and breath on the Lower Dantian. The purpose of this training is to gather the body's Yuan Qi into the Lower Dantian (called "returning to the source"), to strengthen the foundational root for the body's energy.

It is dangerous for Medical Qigong students to bypass the discipline of Lower Dantian cultivation training to progress more quickly to the more advanced intuitive and psychic training of the Upper Dantian. Such an approach to training may lead to Qi deviations (see Chapter 18) and cause emotional instability.

THE MIDDLE FIELD OF ELIXIR

The Heart is the primary organ related to the Middle Dantian; the second organ is the Lungs. In Medical Qigong, the thymus gland is also of primary importance to the Middle Dantian.

In children, the thymus gland is quite large. As the child matures into adulthood, the thymus gland shrinks in size. Until recently, Western biologists thought that the thymus gland became vestigial and inactive in adults. Beginning in the 1980's, however, with the advent of the AIDS epidemic and the increase in cancer cases, intensive new research was launched into the immune system. As a result, scientists discovered that the thymus gland plays a major role in educating and maturing the white blood cells to become immunocompetent. This thymus function continues throughout one's life.

The Middle Dantian collects Qi and represents the body's energetic reservoir for mental and emotional vibrations and energy. The energy of man that is transformed in the Middle Dantian has a fluid quality—like water.

A refining process also takes place in the Middle Dantian, transforming the fluid energy into more steam-like energy that is then transferred to the Upper Dantian. The Middle Dantian transforms Qi into Shen by bringing the transformed Qi into the Heart Fire. This alchemical process is commonly called "changing Qi into Shen" and refers to kinetic energy transforming into spiritual consciousness.

The Middle Dantian is connected to the second level of Wei Qi, circulating about one-and-a-half feet outside the body. As the Middle Dantian fills with Qi, the colors of the student's middle field of Wei Qi change, becoming even more pronounced. The reason for this change is that the Middle Dantian is connected to the Five Agents, which in turn govern the Five Yin Organs and organ emotions. As the students begin to experience various stresses and emotional releases their aura (resonating from the internal organs, throughout the second energetic field) changes its colors.

THE MIDDLE DANTIAN AND JING

The Heart is related to the Fire element. The Heart derives its Yang Fire from the Kidneys. Modern research in Chinese medicine equates the function of the adrenal glands to the traditional function of Kidney Yang. In Western physiology, the adrenal glands help to regulate the pace of the heart.

To keep the Heart Fire in balance, the Heart also needs Yin. Heart Yin is derived from Kidney Yin (Jing is one aspect of Kidney Yin).

In traditional Chinese physiology, the Heart is said to govern the Blood. Not only is the Heart responsible for the circulation of Blood, but in the Chinese view, the transformation of Gu Qi into Blood takes place in the Heart. Food Qi (energy derived from the consumption and transformation of food matter) is a form of postnatal Jing, derived from the Spleen and Stomach. Blood is composed of Nutritive Qi (Ying Qi), Jing and Fluids (see Chapter 22). The Kidneys also send prenatal Kidney Jing to the Heart to make Blood. Therefore, Jing - particularly postnatal Jing - is vital to the Heart's function of governing Blood.

THE MIDDLE DANTIAN AND QI

Similar to the Lower Dantian, the Middle Dantian is also considered to be a Sea of Qi. The Qi of the Middle Dantian is called Zong Qi. Zong Qi is translated as Ancestral Qi, Gathering Qi, Genetic Qi, or Essential Qi. In English translation, it is sometimes confused with the Original Qi (Yuan Qi), but they are not the same. The Zong Qi is a form of postnatal Qi, whereas the Yuan Qi is housed in the

Lower Dantian and is a form of prenatal Qi. Zong Qi and the Yuan Qi assist each other to maintain the healthy function of the Heart and Lungs.

The Zong Qi nourishes both the Heart and Lungs, controls the speech and the strength of the voice, and interacts with the Kidneys to aid in respiration. According to Traditional Chinese Medicine, the Kidneys assist the Lungs in holding and stabilizing the breath when inhaling.

Qi and Blood are closely related. In Chinese medicine, it is often said, "Qi is the master of Blood; Blood is the mother of Qi." Qi gives the Heart and blood vessels the strength to circulate Blood and gives life to the Blood. Blood, on the other hand, houses and carries Qi to all the cells in the body. When one loses Blood, one also loses Qi. Therefore, one should consider Qi and Blood to be inseparable.

Qi is also inseparable from the mind and spirit. According to the teachings of ancient Tibetan Qigong masters, the channels are, metaphorically speaking, the road, the Qi is the horse, and the mind is the rider. Through refining the Qi, the mind and spirit are refined and purified. The Middle Dantian is the main focal point for this refinement of Qi into spirit.

THE MIDDLE DANTIAN AND SHEN

Classically, the Chinese locate the Mind in the Heart. In Chinese, the word for "mind" (Xin) is also the Chinese word for Heart. In Medical Qigong, a distinction is made between the Acquired Mind (Ren Xin) and the Original Mind (Yuan Xin).

The Middle Dantian is said to house the Shen and control all of the other functions of Shen that are attributed to the other Yin organs. Thus the Heart is often referred to as the Heavenly Emperor.

Throughout the world, people relate the heart to emotions and feelings. Emotions and feelings are one important aspect of the spirit. Any type of emotion will have an effect on the Shen.

The negative emotions are sometimes called "the five thieves" because even though negative emotions are necessary for life, chronic states of negative emotions drain the Qi.

The human mind easily falls under the influence of the Po (the Seven Corporeal Soul), that are concerned with survival. When the Po dominate the Heart, their overexaggerated self-concern gives rise to a chronic state of fear, sadness, worry, anger, and defensive arrogance.

The redeeming virtue of the Heart is a sense of propriety and discriminating awareness. The Hun (the Three Ethereal Souls) control the smooth flow of Qi throughout the body and are nourished by the Five Virtues of kindness, order, trust, integrity, and wisdom. These Five Virtues give peace and clarity to the Heart and allow the higher qualities of the Yuan Shen to overrule the Po.

An important relationship regarding the Middle Dantian and Shen is found in the Heart's role of governing the Blood. The ancient classics state that the Shen also resides in the Blood and pervades the body through Blood circulation. This relationship between Blood and Shen is one reason why anemic patients are often restless and suffer from insomnia. Through nourishing the Heart Blood, many forms of spiritual unrest can be treated.

According to Dr. Candace Pert's information on neurotransmitters (stated in *Psychoneuro Immunology*), the brain and white blood cells both contain the same neurotransmitters and biochemical constitutes that are prerequisite for consciousness awareness to exist. These same neurotransmitters and biochemical constitutes which are linked to consciousness are synthesized and created by the white blood cells. This similarity indicates that not only do the brain and abdomen have their own consciousness and nervous system but so does the Blood. This similarity also implies that consciousness is possible anywhere in the body, substantiating the ancient Chinese understanding that consciousness is pervasive throughout the body via the Shen, which resides in the Blood.

THE MIDDLE DANTIAN AND EMPATHIC COMMUNICATION

The Middle Dantian is also considered the "house" of emotional (empathic) communication, awareness, and feelings. Emotional communication is analogous to the feelings of empathy felt within the Heart. Empathy is the means by which the Qigong doctor will most frequently get in touch with the emotional components of the patient's energetic blocks and imbalances.

Empathic communication is felt as an emotion and originates in the Heart and Middle Dantian area. When Qigong doctors focus on the Middle Dantian area, a line of communication is created with their higher self. We are all born with this ability, but as we grow older, we tend to override this type of emotional communication with our logical mind. These impressions slowly atrophy, eventually causing us to lose this natural empathic ability of communication. We disconnect from this higher perception due to the negative and double messages received from our parents and from society. The way to reconnect with the intuitive self is to look inward and become one with our true self, which is connected to the divine.

ANATOMICAL LOCATION OF THE MIDDLE DANTIAN

The Middle Dantian is shaped like a tetrahedron (four points): one area points toward the Upper Dantian and the Heavens, one area points toward the Lower Dantian and the Earth, one area points toward the front, and one area points to the back .

1. The front lower point of the Middle Dantian is located at the Zhongwan CV-12 (Middle Stomach Cavity) point on the midline of the abdomen, just below the sternum. This area is the master point of the Middle Burner and is sometimes known as the Front Gate of the Third Chakra.

   This emotional storage area is also called the Yellow Court because it reflects the emotions stored from the Heart, (the Heart was sometimes referred to as Suspended Gold in ancient China).

   The back lower point of the Middle Dantian is located on the Jinzhong GV-6 (Middle of the Spine) point. The name refers to the point's location. This area is sometimes known as the Back Gate of the Third Chakra.
2. The front center point of the Middle Dantian is located at the Shanzhong CV-17 (Central Altar) point, on the middle of the sternum at the level of the fourth intercostal space. The name refers to the "place of worship" where the Shen resides. This area is sometimes known as the Front Gate of the Fourth Chakra.

   The back center point is located two inches up from the shoulder blades, at the Shendao GV-11 (Spirit Path) point located at the hollow between the fifth and sixth thoracic vertebrae. The name refers to the easy accessibility into the patient's Shen residence (within the Heart) through this particular area on the back. This area is sometimes known as the Back Gate of the Fourth Chakra.
3. The upper front point of the Middle Dantian is located at the Tiantu CV-22 (Heaven's Chimney) point, at the throat. The name refers to the visceral cavity area at the base of the throat responsible for "pooling" escaped Heaven Qi from the Lungs. This area is sometimes known as the Front Gate of the Fifth Chakra.

   The upper back point is located on the Dazhui GV-14 (Big Vertebra) point on the back. The name refers to the point's location below the seventh cervical vertebra and above the first thoracic vertebra. This area is sometimes known as the Back Gate of the Fifth Chakra.
4. The center of this Middle Dantian area is located in the right atrium of the heart, centered between the SA (sinoatrial) and the AV (atrioventricular) nodes. The center of the Heart is considered the seat of all emotions.

THE NINE CHAMBERS OF THE MIDDLE DANTIAN

The Middle Dantian regulates the body's Heart—Mind connection. The Heart-Mind connection includes both the all pervasive consciousness of the entire body and the Shen. It distributes the flow of Yuan Qi into the Yin and Yang aspects of the body's mental and emotional functions. These functions within the psyche are controlled by the interaction of the body's Wu Jing-Shen or Five Essence Spirits (Hun, Po, Zhi, Yi, and Shen). The energy of the Middle Dantian is also

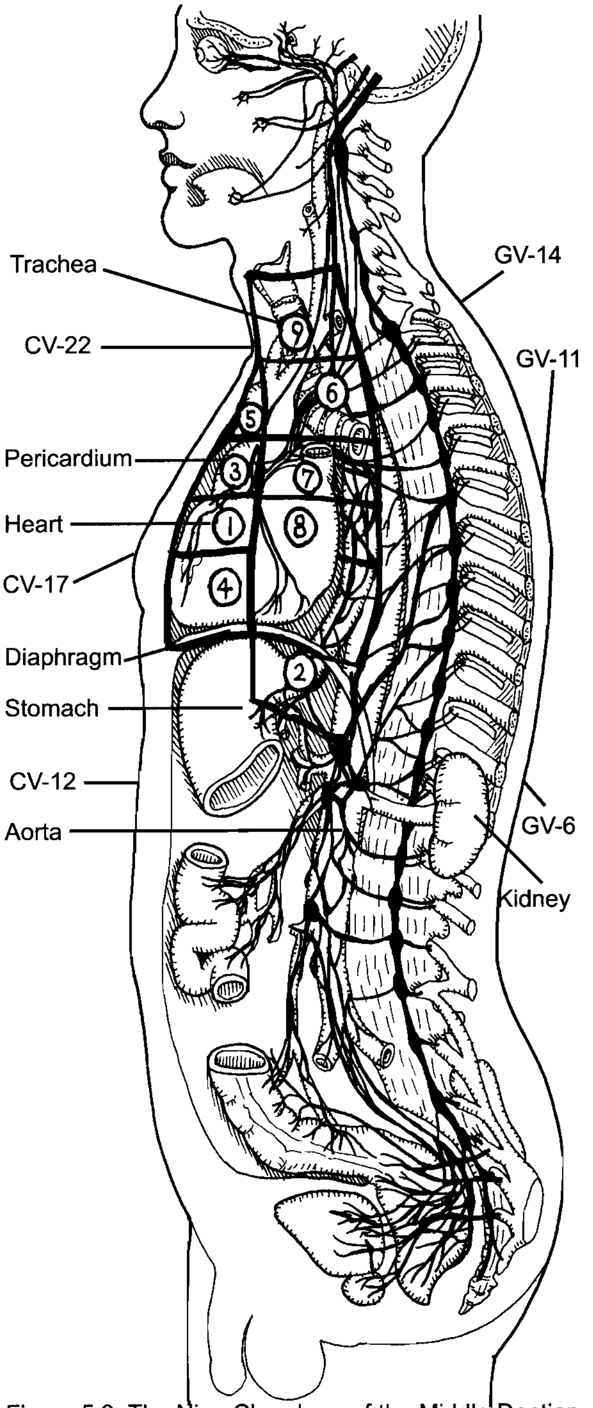

Figure 5.3. The Nine Chambers of the Middle Dantian are identical for men and women. Each number encompasses the entire chamber.

associated with the Spleen and Stomach energies. The vertical thoracic set of nine chambers (Figure 5.3), located within the Middle Dantian, originate around the auricles and ventricles of the heart (especially the pericardial and pleural cavities). The Nine Chambers of the Middle Dantian are named as follows.

1. The Chamber of Mysterious Elixir
2. The Lower Court of the Heart
3. The Chamber of the Ultimate
4. The Chamber of Splendor
5. The Heaven's Cover
6. The Twelve Storied Chamber
7. The Chamber of Government
8. The Purple Chamber
9. The Heaven's Chimney

ENERGETIC FOCUS OF TRAINING THE MIDDLE DANTIAN

In Medical Qigong training the doctors encourage their students to focus their mind and breath on the Middle Dantian to regulate the Heart. Techniques are used for treating Deficient conditions by drawing Qi into the Heart and Middle Dantian area, then regulating the body's energetic fields. For treating Excess conditions, the students are encouraged to lead and purge the Excess Qi from the Heart and Middle Dantian area through the body's extremities. The purpose of this training is to release the toxic Excess Qi gathered in the patient's Heart and Yellow Court areas. These areas of Excess tend to cause disease.

THE UPPER FIELD OF ELIXIR

The Upper Dantian is the collector of Heaven Qi and represents the spiritual aspect of man and his connection to the divine. The Heaven energy that is transformed in the Upper Dantian is of a thinner more ethereal quality—like vapor.

The Upper Dantian is connected to the third level of Wei Qi, circulating several feet outside the body. As the Upper Dantian fills with Qi, the spiritual intuition and psychic perception of the individual increases.

In Chinese physiology, the brain controls memory, concentration, sight, hearing, touch, and smell. These senses stay in close communication with the Heart and Shen.

THE UPPER DANTIAN AND JING

The Jing and Qi form the material foundation for the Shen. In Chinese, the term Jing-Shen means mind or consciousness. Jing-Shen may also mean vigor, vitality, or drive. In China both Western and Chinese doctors use the term Jing-Shen Bing to refer to all types of mental illness. The term Shen is rarely used by itself in a medical context without the concept being connected to the close relationship between the power of the mind and spirit.

The term Wu Jing Shen is used in Medical Qigong to describe the body's Five Essence Spirits (Hun, Po, Zhi, Yi and Shen). These five spirits combine the energetic essence of the Five Yin Organs in order to create the body's innate spiritual consciousness.

The Jing itself is considered the basis for, and ruler of Marrow (a substance derived from the Kidneys that nourishes the brain, spinal cord and forms bone marrow). The brain, is one of the six Extraordinary Organs and is called The Sea of Marrow, and is considered to be a form of Marrow. The six Extraordinary Organs are hollow Yang organs that store Yin Jing. Deficiency of Jing may lead to poor concentration, poor memory, dizziness, and absentmindedness. Deficiency of Prenatal Jing is related to mental retardation and attention deficit disorder (ADD) in children.

In some styles of Qigong, the Jing is intentionally conserved and its energy is drawn upwards from the Lower Dantian through the spine to nourish the brain. Such nourishment benefits the mind and spiritual consciousness.

THE UPPER DANTIAN AND QI

The head is the most Yang part of the body since it is the closest part of the body to Heaven. The Qi that operates in the Upper Dantian is, therefore, Yang in nature. The Spleen and Kidneys send the Clear Yang Qi (pure, light, and insubstantial) upwards to the brain to facilitate mental clarity and activity.

The Upper Dantian is also the place where we connect with the Yang Qi of Heaven. Qigong practitioners consciously absorb Heavenly Qi through the upper doorway, Baihui GV-20. The Heaven Qi is composed of the Qi from the celestial bodies: the sun, moon, planets, and stars.

The Upper Dantian is located in the center of the brain, in an area that encompasses the pineal, pituitary, thalamus, and hypothalamus glands. The pineal gland and hypothalamus have been shown to be extremely sensitive to the influence of light. In his book, *The Body Electric,* Dr. Robert Becker cites experiments with bees and several species of birds that indicated that these animals navigate by the light of the sun. Birds have disproportionately large pineal glands. He also discovered that birds seem to have a backup system of navigation based upon a sensitivity to the electromagnetic fields of the earth.

There are three forms of Qi to which the brain responds and interacts: light, electricity, and magnetism. These particular forms of energy stimulate the pineal, pituitary, thalamus, and hypothalamus glands, affecting and influencing the emotional responses.

THE UPPER DANTIAN AND SHEN

We have already discussed the relationship of the Jing and the Shen with the Upper Dantian. Of particular interest to the alchemists is the opening of the center of the Upper Dantian, called the Crystal Room, where psychic perceptions and intuitive knowing take place. Higher communications, a sense of great bliss, and perceptions that transcend the seeming bonds of time and space are associated with the Upper Dantian. These experiences are particularly valuable to Qigong doctors, who use these heightened perceptions to diagnose illness. Their efficacy is well documented in the works of such authors as C. Norman Shealy and Caroline Myss, who use the term "medical intuitive" to describe this paranormal ability.

The Upper Dantian is also the place where the Eternal Soul connects with the Wuji, and with the Dao. The awareness associated with this union is beyond description, as its unity supersedes the differentiation inherent in conceptual thought.

Although the Upper Dantian is responsible for intuitive and psychic perceptions, the combined energetic properties of all three Dantians establishes the foundation for all psychic perceptions. The steam-like quality of the energy with the Upper Dantian fuses within the light now existing in this center. As this energy disperses, it travels out into the Wuji, returning back to the divine. This interaction is also responsible for what the Chinese call "receiving the message," which describes the ability of connecting with the patients' subtle energy field to acquire the hidden knowledge stored within the patients' tissues.

While in a state of tranquillity and inner peace, the Qigong doctors' Upper Dantian will also intuitively process information from the environment and universe. This intuitive knowledge provides Qigong doctors with a greater ability to explore both their own consciousness and the patients'. The ancients called this ability "knowing without knowing."

THE UPPER DANTIAN AND INTUITIVE COMMUNICATION

As the Shen is developed and the Upper Dantian is opened, spiritual communications may reveal themselves in a flash of an image or a vision in the mind's eye. These images and visions are sometimes very abstract and short lived. Interpreting images takes practice because the images streaming from the Yuan Shen cannot be interpreted easily by the logical mind.

The Qigong doctors must be able to distinguish between true and false messages reflected through their visions. True visions are received from the divine connection to the Dao or Wuji; false visions reflect the messages from the subconscious. The ability to accurately separate these visions is an another example of "knowing without knowing."

Although communication from within usually makes itself felt as a strong impulse, Qigong doctors must learn to keep the logical mind from interfering by practicing spiritual meditations. These meditations involve the forging of links in the chain of communication with the higher self. They should be practiced again and again until this connection becomes a natural, recurring phenomenon. The more one practices at stilling the logical mind and circumventing the ego, the easier it will be to get a clear communication from the higher self. When the higher self initiates a communication, it does not demand or impose itself, and if ignored it will usually stop sending signals.

Five Elements must be in place before the communication lines become fully open and operational:

1. Purity in intention,
2. No hidden agendas,
3. Surrender to the divine will,
4. Complete trust and faith in success (believe and expect), and
5. A quiet and receptive stillness of mind.

The lines of communication with the higher self are severed by the logical mind through doubt, fear, and disbelief. Strong faith is required to open this line of communication. Faith requires no logical proof; if proof is needed, then doubts interfere and breed failure. The logical mind cannot know absolute faith and, therefore, faith must come from deep within one's true self. Any form of cynicism will lead to the death of spiritual growth, for it strikes at the root of faith itself.

Faith is not something that can be forced. Even after practicing Medical Qigong for many years, Qigong doctors may still have to battle their own questions. However, through successful practice, the seed of faith is established, allowing it to grow and blossom. The opposite of faith is doubt combined with fear. Suppression and denial of fear builds and armors the ego, which leads to further pain. If the fears are accepted and acknowledged, they can be overcome by faith.

ANATOMICAL LOCATION OF THE UPPER DANTIAN

The Upper Dantian is located in the head, approximately three inches posterior to the Yintang point (between the eyebrows). It is shaped like a pyramid, facing upward to gather the energy from Heaven. This pyramidal reservoir houses light.

1. The front point of the Upper Dantian is the Yintang (Hall of Impression) point. The name refers to the ancient tradition of placing a red mark or "seal" over the Bright Hall, or Entrance of the Spirit, this point represents wisdom and enlightenment. This area is sometimes known as the Front Gate of the Sixth Chakra, or the Third Eye point.

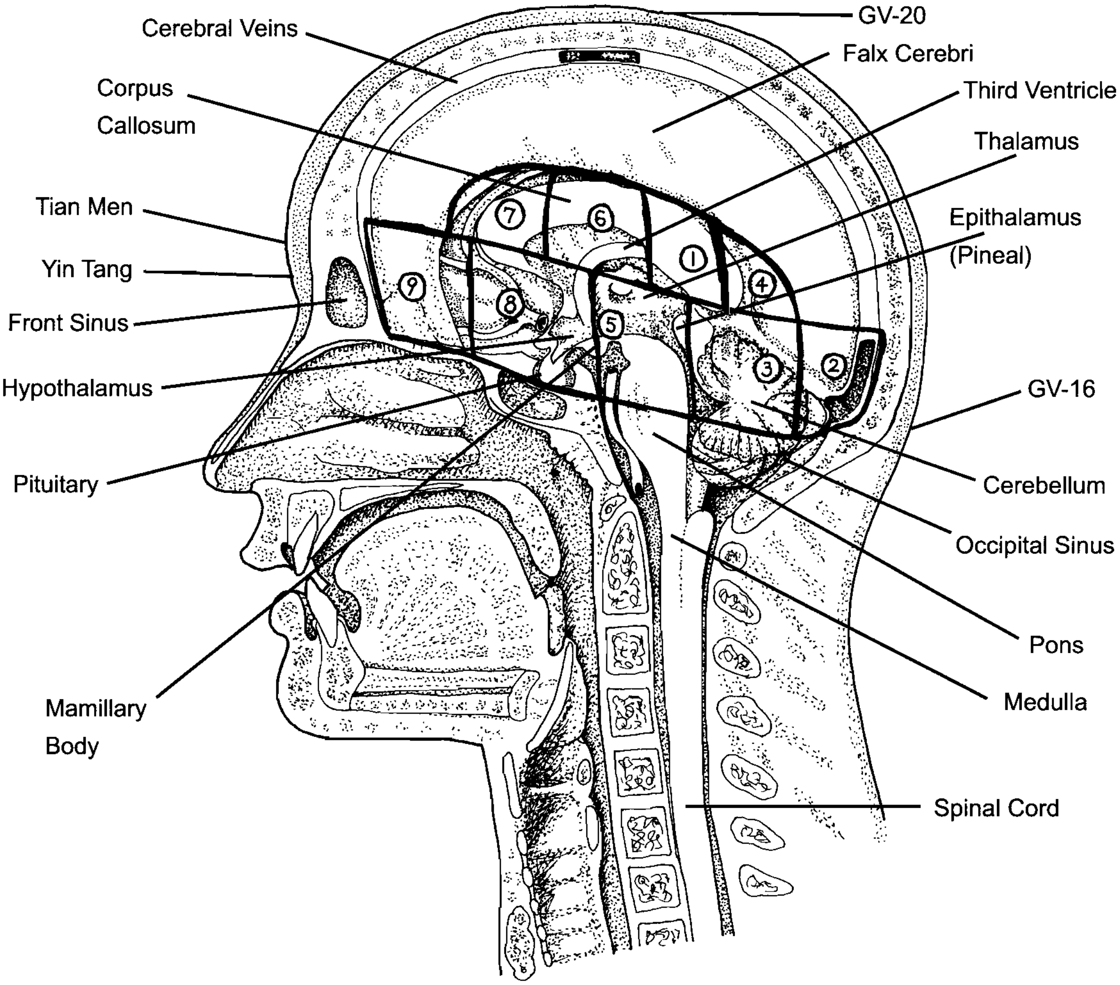
Cerebral Veins
GV-20
Falx Cerebri
Corpus Callosum
Third Ventricle
Thalamus
Tian Men
Epithalamus (Pineal)
Yin Tang
Front Sinus
Hypothalamus
GV-16
Pituitary
Cerebellum
Occipital Sinus
Pons
Mamillary Body
Medulla
Spinal Cord
①
②
③
④
⑤
⑥
⑦
⑧
⑨

Figure 5.4. The Nine Chambers of the Upper Dantian are portrayed in this drawing. Each number encompasses the entire chamber.

2. The back point of the Upper Dantian is located below the external occipital protuberance, on the Fengfu GV-16 (Wind Palace) point. This area is also connected with and surrounded on both sides by the Bl-10 (Heavenly Pillar) point. The energy field connected to this point may be likened to an antennae receiving messages; it allows Qigong doctors to regulate their state of consciousness. The Governing Vessel 16 point is also a Sea of Marrow point, which is used to affect the flow of Qi and Blood to the brain, and a Window of Heaven Point (one of eleven points used for treating Shen disturbances), as well as one of the thirteen points identified by the famous Daoist physician Sun Simiao as a Ghost Point (points used for treating spirit possession). It has been my personal observation that students with a more prominent occipital protuberance tend to see auras more easily and receive psychic intuition faster. This area is sometimes known as the Back Gate of the Sixth Chakra.
3. The highest point of the Upper Dantian is lo-

cated on the vertex of the crown, on the Baihui GV-20 (One Hundred Meetings) point. The name refers to the ancient tradition that one receives divine messages and spiritual intuition from this point into the Upper Dantian's Chamber of Mysterious Elixir, located within the third ventricle of the brain. This area is sometimes known as the Upper Gate of the Taiji Pole or the Upper Gate of the Seventh Chakra.

It is also said in Traditional Chinese Medicine that all of the body's major channels send a connecting vessel to the Baihui so that at death the Qi can leave the body through this upper doorway and ascend to the Heavenly realms.

4. The center of the Upper Dantian is located in the pineal gland. It is the area where the Shen transcends and merges with the Wuji and then progresses on to the Dao. The Upper Dantian is also considered the house of spiritual (intuitive) communication, awareness, and feelings.

THE NINE CHAMBERS OF THE UPPER DANTIAN

The horizontal cephalic set of nine chambers, located within the Upper Dantian (Figure 5.4) is based on the different ventricles of the brain. The Nine Chambers of the Upper Dantian are named as follows.

1. The Chamber of Mysterious Elixir (Mystical Medicine)
2. The Palace of Jade Emperor
3. The Palace of Moving Pearls
4. The Chamber of Splendor
5. The Hall of the Upper Dantian (Medicine Field)
6. The Chamber of the Ultimate Truth
7. The Cover of Heaven (Palace of the Heavenly Court)
8. The Chamber of Government
9. The Entrance of the Spirit (Bright Hall)

ENERGETIC FOCUS OF THE UPPER DANTIAN

In Medical Qigong training, the Upper Dantian pertains to gathering spiritual intuition and light to advance the doctor's intuitive and psychic ability (see Six Transportations of Shen, Chapter 13).

The Qigong doctor may also absorb universal and environmental Qi into the Upper Dantian through the Yin Tang (Third Eye region) and the Tian Men areas (located in the center of the forehead), and emit healing Qi to patients through either one or both of these points. The Shen can both exit and enter the body from the area in the Upper Dantian by way of the Baihui, Yin Tang, and Tian Men (Heavenly Gate) areas.

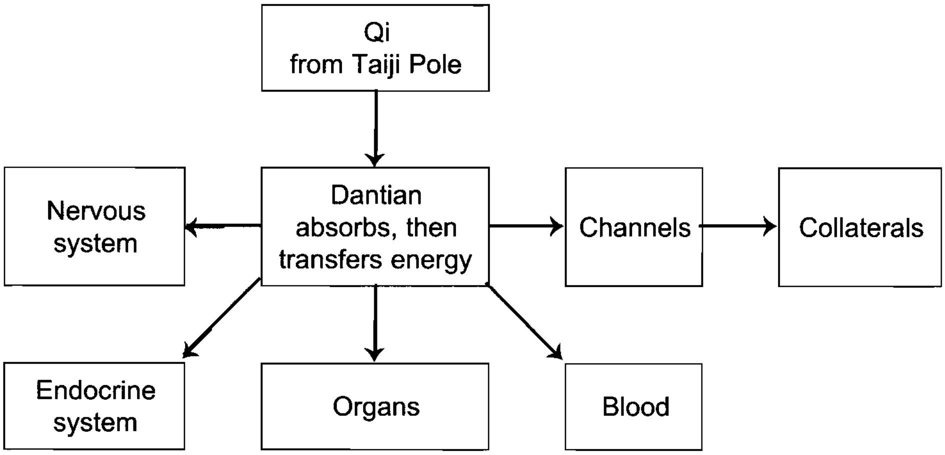

Figure 5.5. Qi Dispersion From the Center Taiji Pole Through the Dantians into the Body

THE FUNCTIONAL ASPECTS OF THE THREE DANTIANS

The three energy reservoirs of the Dantians are linked externally through the Governing and Conception Vessels and internally through the Thrusting Vessels and the Taiji Pole.

The center of each Dantian is penetrated by and attached to the Taiji Pole, which exits from the Baihui at the top of the head to the Huiyin point at the perineum. Each of the Dantians corresponds to anatomical locations that are centers for magnetic and electrical vibration and charge. The body's intensity of vibration and charge is dependent on the mental intention, the posture, and the respiration.

Qi moves into the body's Dantians through the body's Taiji Pole. The energy is then absorbed into the body's major organs and surrounding tissues as it flows out of the Dantians and into the body's internal and external channels and collaterals.

Each Dantian acts like a reservoir, collecting energy and redistributing it into all of the internal organs. This energy projects through the surface of the body into the Wei Qi field. The same energy also projects inward flowing through the energy channels, along the nervous system and endocrine glands, and then through the Blood, to nourish the body (Figure 5.5).

You can visualize this energy transformation as follows: Qi flows into the body like rainwater flowing into a lake (the body absorbing and collecting Qi into the Dantians). The rainwater is then absorbed into the surrounding soil, foliage, and root systems (skin, tissues, and cells) before it gathers, collects, and pools into deep artesian wells (the Dantians). Pressure begins to build up as these artesian wells fill with the rainwater, and eventually overflow, pouring into smaller pools (the organs) before combining with the rushing action of underground streams (the channels).

Another popular analogy is to consider the Dantians as batteries, the body's Taiji Pole as a magnetic bar connecting the batteries together, the channels as the wires, and the Wei Qi fields as the electromagnetic fields manifesting from the energy contained within the structure.

Mental and emotional awareness of a specific tissue area can be heightened through increasing the flow of energy to that location. When energy fills the tissues, a cellular release of energy causes the tissues to either store or release emotions, depending on the body's overall Excess or Deficient Qi.

If an increase of Qi is focused into the Lower Dantian, the result is a more powerful and stable feeling in the body. When an increased amount of Qi is focused into the Middle Dantian, the result

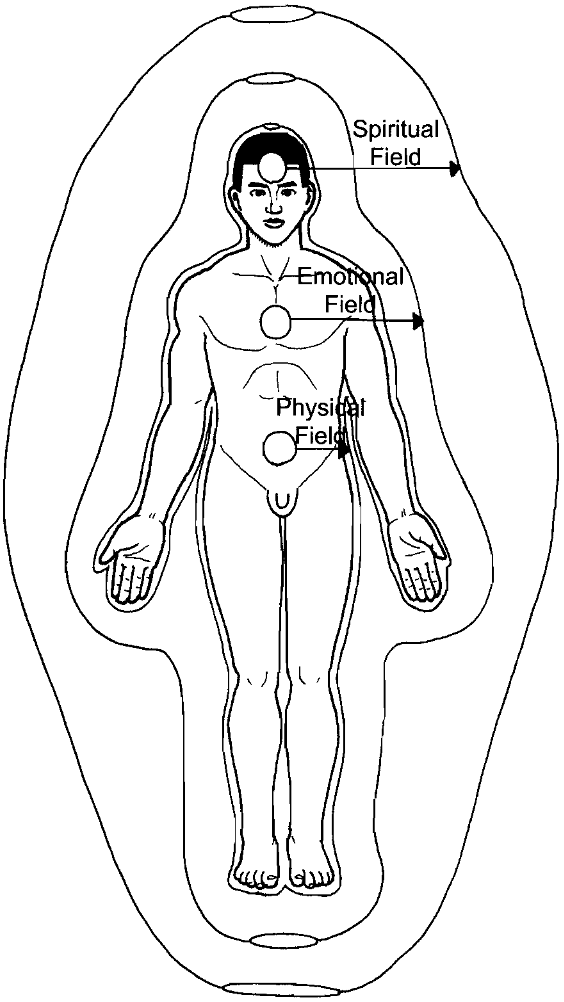

Figure 5.6. The Three Dantians and Their Relationship to the Three External Wei Qi Fields

is a heightened feeling of emotional perception and awareness. If an increase of Qi is focused into the Upper Dantian, a heightened spiritual awareness and sense of connection to the divine occurs. The greater these energies flow, the healthier the individual becomes, and the stronger his or her energetic fields become (Figure 5.6).

THE DANTIAN'S YIN AND YANG ENERGETIC CHAMBERS

Each of the Three Dantians can be divided into Yin and Yang energetic chambers. The Yang chambers relate to each Dantian's upper chambers and the spiritual aspects of the Hun. The Yin chambers relate to each Dantian's lower chambers and the carnal aspects of the Po. The chambers and their energetic potential are described as follows.

THE LOWER DANTIAN

1. The Yang Energetic Chamber of the Lower Dantian relates physically to the upper quadrant of the abdomen, located within the small intestine in men and the uterus in women. When influenced by the Hun, quiescence and serenity are enhanced.
2. The Yin Energetic Chamber of the Lower Dantian relates physically to the lower quadrant of the abdomen, which includes the reproductive organs and the Bladder, urethra, and anus. This chamber is responsible for reproduction and sexuality. When influenced by the body's Po, raw physical power and sexual vitality are aroused.

THE MIDDLE DANTIAN

1. The Yang Energetic Chamber of the Middle Dantian relates physically to the upper quadrant of the Heart, which includes the left and right atrium, the upper portions of the left and right ventricles, and the atrioventricular node. This chamber is responsible for spiritual attitudes and virtues related to the influence of the body's Hun. The Hun are responsible for:
   - gathering and transmitting divine inspirations and spiritual insights,
   - giving and receiving unconditional love, and
   - motivating spiritual growth through prayer, devotion, and commitment.
2. The Yin Energetic Chamber of the Middle Dantian relates physically to the lower quadrant of the Heart, which includes the left and right ventricles. This chamber is responsible for sensual passions, conquests, and conditional love. It deals with actions which are often based on hidden agendas. These emotions are related to the influence of the body's Po. The influence of both the Po and the Yin Energetic Chamber of the Upper Dantian activates, energizes and enhances:
   - biological drives (for food and self preservation,
   - sexuality, lust and desire, and
   - sensuality.

THE UPPER DANTIAN

1. The Yang Energetic Chamber of the Upper Dantian relates physically to the upper quadrant of the brain, often referred to as the third ventricle or higher brain centers. When stimulated, the higher brain centers manifest spiritual intuition and divine insight, related to the influence of the Hun. Such insight is responsible for spiritual growth and maturation. Insights eventually lead to the emergence of extrasensory perceptions (ESP), e.g., clairvoyance, clairaudience, telepathy, psychokinesis, and spiritual enlightenment.
2. The Yin Energetic Chamber of the Upper Dantian relates physically to the lower quadrant of the brain, often referred to as the "reptilian" brain. When stimulated, the reptilian brain activates the thalamus, hypothalamus, cerebellum, and cortex, which awaken the body's intuition, as well as animalistic and primordial instincts of survival. When the Yin Energetic Chamber dominates, the sensory, animalistic nature of the Po emerges full force. This phenomenon sometimes occurs when coma patients begin to recover. As energy begins to fill the lower chambers of the Three Dantians, their initial reactions are basic and carnal in nature, i.e., to either engage in sexual activity or strike out in violence. As the Qi begins to fill the upper chambers of the Three Dantians, the energy balances in the patient's

Taiji Pole and the patients' impulses return to normal.

THE DOCTOR'S PROJECTED AURA FIELDS

In China, Qigong doctors are tested and categorized according to the predominant color of their Qi emissions. The healing color will depend upon which of the three Dantians is dominant (corresponding to either the Lower Dantian, Middle Dantian, or Upper Dantian), and the energetic strength and potential of the individual being tested. The emitted color is observed in the visible-light spectrum projected from the Qigong doctor's hands.

YELLOW EMITTED COLOR

Individuals who have just begun their training will emit a yellow glow around their external energy field. The Qi will naturally overflow from the Lower Dantian area and the color yellow will dominate the aura. This is actually a normal aura color observed from most beginning Qigong doctors, indicating that the energy field is still too weak to treat serious diseases.

RED EMITTED COLOR

Qigong doctors who have mastered their body and mind will emit a red radiant glow around their external energy field. The Qi will also overflow from the Lower Dantian area, and the color red will dominate the aura. This is actually a normal range of aura color observed from the many Qigong doctors who have just graduated from Medical Qigong colleges, indicating that the energy field is at the intermediate stage.

PURPLE EMITTED COLOR

Qigong doctors who have evolved to the next level of mastering the emotions will emit a purple radiant glow around their energy field. These doctors' Qi will naturally overflow from the Middle Dantian area, and the color purple will dominate the aura. This aura color is observed in strong and advanced Qigong doctors.

BLUE EMITTED COLOR

Qigong doctors who have further refined their energy will emit a blue radiant glow around their energy field. These doctors' Qi will also overflow from the Middle Dantian area, and the color blue will dominate the aura. This aura color is observed in very advance Qigong doctors, and is developed only after many years of clinical practice and personal spiritual transformations.

WHITE EMITTED COLOR

The Qigong doctors who have mastered their spiritual life, as well as surrendered their body, mind, emotions, and spirit to a divine or higher calling will emit a white radiant glow around their energy field. These doctors' Qi will naturally overflow from the Upper Dantian area, and the color of the aura is considered the strongest and most powerful.

CLEAR-OPALESCENT EMITTED COLOR

The divine energy is emitted through a clear-opalescent color. This Qi emission is different then the white radiant glow, and is a sign of divine intervention and healing. These times of divine intervention can occur in any stage of the doctor's energetic development, and are a testimony to the healing virtue of the divine.

Although every Qigong doctor uses a combination of all Three Dantians when projecting Qi into patients, the color of the aura surrounding the doctor's body reflects which Dantian's reservoir is predominantly used. Through time, patience, and much practice, the Qigong doctor will be able to transform from a yellow to red, purple, blue, and finally to white-light healer.

THE TAIJI POLE

The Taiji Pole can be observed as a vertical column (or pole) of brilliant white light, full of vibration and energetic pulsation. This energetic channel flows from the Baihui point at the top of the head, through the center core of the body, connecting the Three Dantians. The Taiji Pole roots its energy at the base of the perineum at the Huiyin area.

In Chinese energetic embryology, as the father's sperm enters the mother's egg, it creates the upper or Heavenly vortex point of the Taiji Pole. At this stage of energetic development, the Heavenly vortex creates and connects the Three

Dantians together, rooting the column of white light energy into the Kidneys and Mingmen area. As the fetus continues to grow and develop, the Taiji Pole slowly descends, shifting its energetic root to the base of the perineum. After birth, the Taiji Pole can be observed as a vertical column of brilliant white light, surrounded by a veil of golden spiraling light.

HEAVEN, EARTH, AND MAN RELATIONSHIPS

The Taiji Pole of the body is comparable to the central axis of the Earth, which is known as the Taiji Pole of the Earth. At the end of the Earth's Taiji Pole are the North and South Poles, which are compared to the body's Baihui at the top of the head and Huiyin at the base of the perineum. The Earth's Taiji Pole (or central axis) aligns to the North Star.

The North Star is seen as the Taiji Pole of Heaven and is sometimes called the Pole Star or Taiji. It is sometimes said that the original Chinese character for Taiji means Yin and Yang. While the original meaning of the Chinese character for Yin is the shady side of the mountain, and Yang depicts the sunny side of the mountain, the mountain's peak is considered a Taiji, where both Yin and Yang meet (or the center of Yin and Yang).

The center Taiji is also called the Still Point, because although all of the changes of Yin and Yang occur in the Heavens within the course of a day, and all of the stars seem to make a 360 degree rotation (as the Earth rotates) the Pole Star remains stationary as Heaven's Still Point. In a similar manner, the body has the Taiji Pole as the "still point of man." A great stillness is experienced when a Qigong practitioner gathers Qi into the Taiji Pole.

INTERNAL ENERGY INTERACTIONS

Extending from the body's Taiji Pole are the Three Dantians, Five Jing-Shen, and the Eight Extraordinary Vessels. These vital energetic systems feed all the major organs and the body's Twelve Primary Channels. As the Three Dantians absorb energy from the body's Taiji Pole, they also extend, or "feed," and energize the body's Yin and Yang organs, Twelve Primary Channels, and Eight Extraordinary Vessels. Each individual Dantian is connected to the body's Taiji Pole and is responsible for creating its own alchemical transition in the process of converting Jing into Qi, Qi into Shen, Shen into Wuji, and Wuji back to the Dao (see Chapter 13).

In my clinical experience, the closer I work to the patient's Taiji Pole the more powerful the energies become. Accessing the patient's Taiji Pole, Three Dantians, and Eight Extraordinary Vessels for whole body tonification and regulation is faster and less draining on my energetic reservoirs then emitting Qi into the patient's Twelve Primary Channels to accomplish the same type of treatment.

THE FIVE THRUSTING CHANNELS

Five internal Thrusting Channels connect to and transverse the body's center Taiji Pole. These five internal channels begin at the perineum and spiral up through the center of the body's core, flowing into the upper crown of the head. They connect the Baihui area at the top of the head to the Huiyin area at the bottom of the perineum. These five internal channels absorb the Five Elemental energies of Heaven and Earth (Wood, Fire, Metal, Earth, and Water) into the body.

The Taiji Pole is responsible for absorbing the energy from Heaven and Earth and distributing the collected Qi into the body's major internal organs. Similar to the Earth's central axis, the Taiji Pole is divided into two main energetic polarities: The Five Portals of the Heavenly Yang Gate, located at the top of the head, and the Five Portals of the Earthly Gate, located at the perineum (Figure 5.7).

When the body's internal and external energetic wave patterns become synchronized through prayer or meditation, a rhythmic pulsation occurs within the Taiji Pole. This pulse begins deep within the center core of the body and vibrates outward towards infinite space connecting the Wuji with the Dao.

SPIRITUAL MANIFESTATIONS

When doctors penetrate the outside veil surrounding the center core of a patient's Taiji Pole and extend their intention into the light's center, they experience the sensation of falling into space

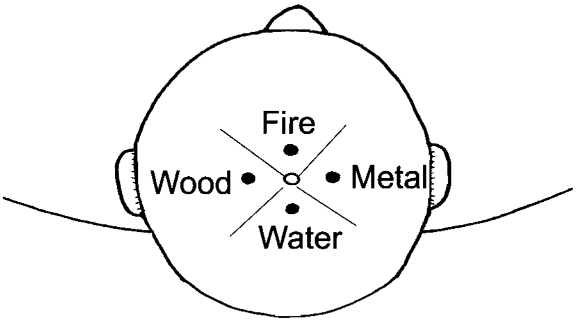

Top of Head

Figure 5.7. The Five Portals of the Heavenly Yang Gate are portrayed here as seen looking down at the top of the head. The Universal Energy is absorbed into the Taiji Pole through the Five Portals of the Heavenly Yang Gates, located at the center of the Baihui (One Hundred Meetings) and Sishencong (Four Spirits Hearing) areas, at the top of the head.

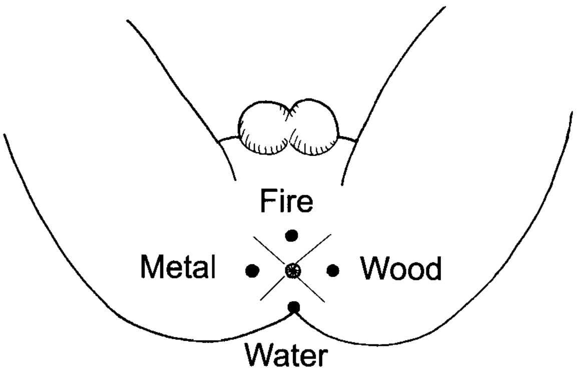

The Five Portals of the Earthly Yin Gate as seen from the perineum. The environmental energy is absorbed into the Taiji Pole through the Five Portals of the Earthly Yin Gates, located at the center of the anal sphincter, between the Huiyin (Meeting of Yin) and the Changqiang (Long Strength) areas, at the base of the perineum.

or shooting through space. This sensation is followed by seeing flashing colors and moving shapes, as the surrounding core dissolves into infinite space itself. As the Qigong doctors continue to extend their intention to reach this area, time and space feels as if they stretch into eternity. This experience is considered the true connection with the Wuji returning back to the Dao. The potential for this experience exists within everyone and can be accessed through deep prayer and meditation. It is quite a normal phenomenon within deep spiritual practices and is sometimes known as accessing the river of God.

ENERGETICALLY OPENING THE TAIJI POLE

When we are born, we are connected to great spiritual wisdom and power through our center core. During the maturation process, this connection slowly fades away, yielding to masks and defense mechanisms, that seek to shield us from our pain and self-hatred (shame). In childhood, every time we repressed the flow of painful feelings connected to an event, we froze that particular event in both energy and time, and locked it within our muscles. Since the external Wei Qi is composed of energy consciousness, a block of frozen psychic energy is formed within the Wei Qi at the moment we inhibit the pain. By walling off our wounds, we also wall off our connection to our deeper core self (Figure 5.8). We pretend not to feel the pain anymore, we block off the memory of the event. Thus, we prevent a natural resolution to the pain and the trauma of the event. Through denying our true feelings we disconnect from our core self and put on masks. We smile when angry, for example, or pretend indifference to our hurts. Since creative inspiration comes from our core, we lose our creative ability. When we stop the negative experiences of anger, pain, or fear, we also stop the positive healing of the physical, mental, and emotional aspects of the experience.

During Medical Qigong treatments, patients regress layer by layer through the pain and fear associated with the blocked energy that gives rise to the disease. Even though the memories may be very strong and frightening, as the emotions

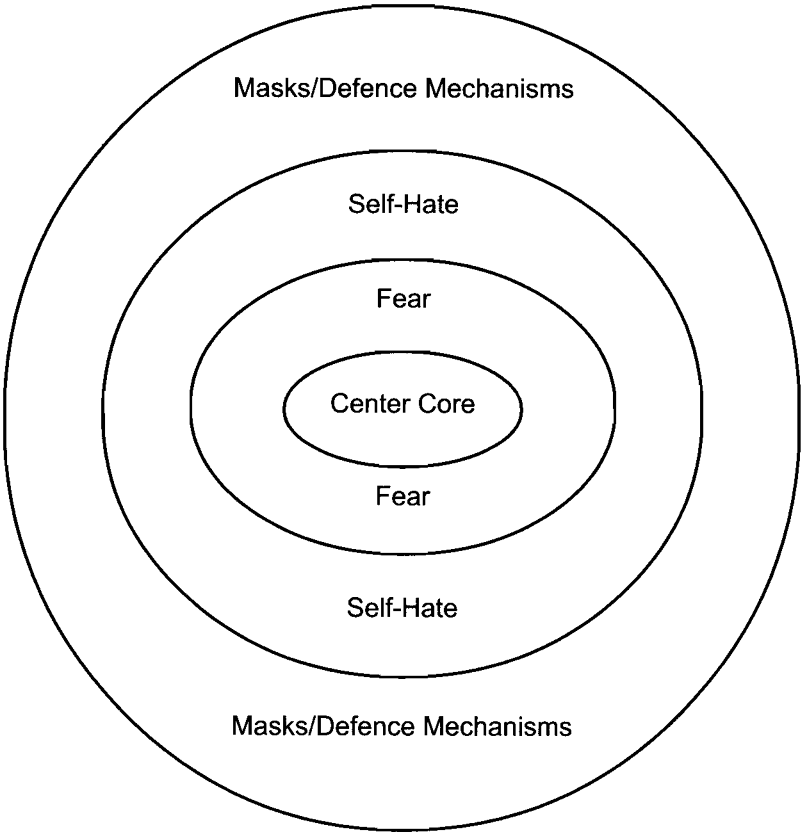

Figure 5.8. The True Self is Contained Within the Core Self

release, the pain decreases with the dispersal of the original trauma.

I have found in my clinical practice that most of the patients' pain comes not from the original trauma but from the unconscious belief system established to protect them from the original trauma. In other words, more pain and illness is created in our bodies by avoiding working through the original trauma (through our habitual defense and avoidance patterns) than was present in the original trauma. It requires enormous amounts of energy to suppress feelings, and each time we do so, we create further injuries to ourselves.

SPIRITUAL AWAKENING

When patients become aware of their spiritual connection to the divine, it is called an "awakening." This change can be observed through the spiritual transitions experienced within the patients' center core through their Taiji Pole. When the center Taiji Pole (which connects all three Dantians) begins to awaken, certain physical, emotional, and energetic reactions may occur within the patients' body. These transformations occur because the tissues of the energetic fascia (that connect to the major organs of the body) are enveloped within the energetic structure and webbing of the Taiji Pole.

THE TWELVE GATES OF THE CHAKRA SYSTEM

Although the body's major internal energy centers are contained within the interacting network of the Three Dantians, the Three Dantians also connect to the body's major energy doors (located on the Governing and Conception Vessels), known as the Twelve Gates of the Chakras. These Chakra gates are superficial, on the surface of the skin, and are not as deep as the Chakra Cores and Three Dantians which are all connected to the Taiji Pole (Figure 5.9).

The word Chakra is Sanskrit for "wheel." The Chakras look like small, colored disks, usually about the size of a silver dollar. Energetically the Chakra gates look like funnels or vortices of Qi. Qigong doctors and sages have for centuries described their perceptions of these gates as resembling energetic wheels or vortexes within the subtle energy body. Each funnel extends and expands its energetic vortex out into the body's Wei Qi field. As energy travels up and down the Taiji Pole it creates an energetic pulse. This energetic pulse resonates out from the body through the Twelve Chakra Gates, and can be felt several feet from the body. The bottom Chakra is located at the perineum and only has one gate, as does the top Chakra located at the top of the head; the second, third, forth, fifth, and sixth Chakras have two gates each, one on the front of the body and the other on the back of the body. These Chakra Gates serve the following different energetic functions:

1. The top or Crown Chakra gate is responsible for absorbing Heavenly Qi and light into the body, as well as energizing the center core's Taiji Pole.
2. The front Chakra gates are responsible for the patient's feelings and emotional activity. They are related to the patient's Shen and the Heart Fire energy. They are connected to the Conception Vessel.
3. The back Chakra gates are responsible for the patient's willpower and determination. They are related to the patient's Zhi and the Kidney Water energy. They are connected to the Governing Vessel.

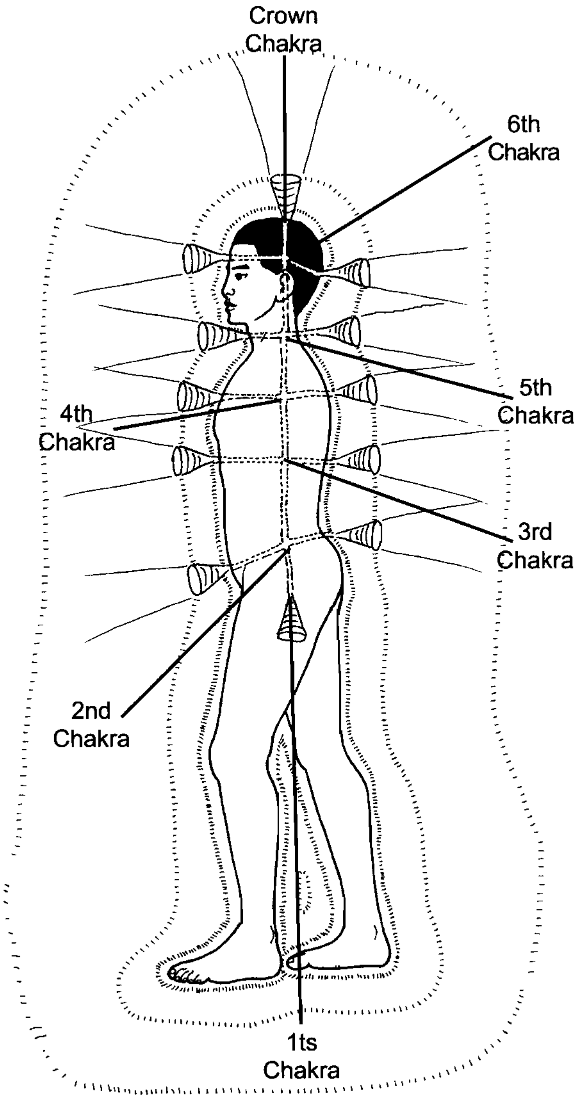

Figure 5.9. The body's Twelve Chakras Gates extend from the Taiji Pole and expand outward into the three external Wei Qi fields.

4. The center of the Chakras are actually deep inside the body's center core, located within the middle of the Taiji Pole.
5. The bottom or Base Chakra gate is responsible for absorbing Earthly Qi and heat into the body, as well as energizing the Taiji Pole. Generally, when treating a patient, the Qigong

doctor extends energy into the patient's back Chakra gates to stabilize the emotions and to access control over the patient's will and determination. The doctor can, for example, connect with the Heart Chakra back gate, located on the patient's Shendao, GV-11 point, to access the emotions stored within the tissues of the Heart. The front Chakra gates are used to remove Excess emotions from the patient's tissues. Although the treatment approach differs, the Three Dantians system and Twelve Chakra Gate system, mutually support each other, both in the location of the energy centers, and in their physical manifestations of energetic function.

The Yin and Yang polarity of the Chakra gates are opposite in men and women (Figure 5.10). The Twelve Chakra Gates serve as subtle energy distributors, that help absorb and distribute environmental Qi to the patient's organs, tissues, and major nerve plexus areas closest to each gate. Each Chakra gate is connected to its own potential of psychic perception, interfaces with the body's nervous system, and is associated with a different endocrine gland. Any energy extended from the Qigong doctor towards a patient's Chakra gate will affect the patient's physical body. As the patient begins to energetically open the Chakra gates, the external energetic Chakra wheels begin to spin in half-circle rotations, opening and closing with the core's energetic pulse. The energetic action is similar to the centripetal (closing) and centrifugal (opening) action of the body's energetic channels.

Each opening of a Chakra gate may result in a spontaneous emotional release. Some of these releases can involve painful memories that have been dislodged from the energetic filter (or veil) that envelops the Chakra gates. This energetic filter prevents external emotional traumas from entering into the body's center core.

A Chakra gate can become stuck open, closed, or may also tilt out of alignment with the Taiji Pole, causing Qi Deviations that may result in a distorted or obstructed flow of energy. This obstruction may cause physical, as well as psychological, stress or trauma. The release of these emotional

Polarity		
Chakra	**Men**	**Women**
7th	+ Yang	– Yin
6th	– Yin	+ Yang
5th	+ Yang	– Yin
4th	– Yin	+ Yang
3rd	+ Yang	– Yin
2nd	– Yin	+ Yang
1st	+ Yang	– Yin

Figure 5.10. The energetic Yin and Yang polarity of the Chakra Gates are opposite in men and women. The Chakra Gates of a man start from a Yang (+) energetic field (at the Lower Chakra Gate) and change to Yin (-) at the second Chakra Gate, alternating from Yang to Yin up the body, ending at Yang in the Upper Chakra Gate. Women display the opposite energetic polarity in each Chakra Gate.

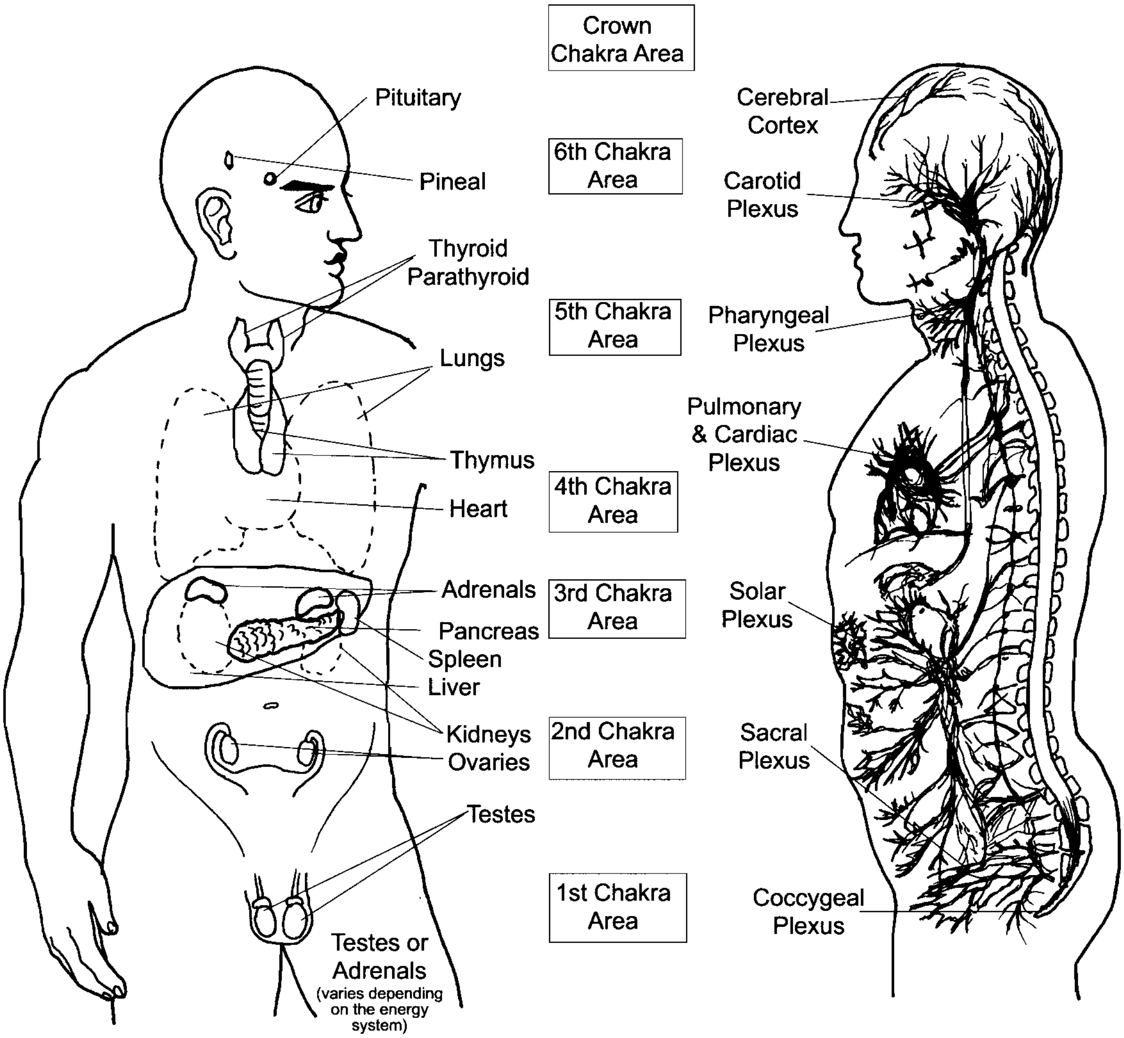
Crown
Chakra Area
Pituitary
Pineal
6th Chakra
Area
Cerebral
Cortex
Carotid
Plexus
Thyroid
Parathyroid
5th Chakra
Area
Pharyngeal
Plexus
Lungs
Pulmonary
& Cardiac
Plexus
Thymus
4th Chakra
Area
Heart
Adrenals
3rd Chakra
Area
Solar
Plexus
Pancreas
Spleen
Liver
Kidneys
Ovaries
2nd Chakra
Area
Sacral
Plexus
Testes
1st Chakra
Area
Coccygeal
Plexus
Testes or
Adrenals
(varies depending
on the energy
system)

Figure 5.11. The body's Chakra system is connected by the Taiji Pole and extends outward into the body's nerve plexus, affecting the glandular system and both anterior and posterior external energetic fields.

traumas can be overwhelming and painful if the patient has not prepared for, or does not understand the nature of, these transitional energetic malfunctions. These manifestations usually occur when the patient's system has been energized and the body is actively seeking to release and get rid of toxic emotions from certain areas. Qigong doctors should be aware of these strong emotional reactions to assist their patients in facing and working through their fears and pain as the Qi Deviations along the Taiji Pole are corrected. Sometimes these experiences unwind at a rapid rate, stimulating the patient's central and anterior nervous system; this can release a flood of mental and emotional images and sensations, that are sometimes accompanied by shaking, thrashing, and other unusual movements of the body (Figure 5.11).

THE BOTTOM CHAKRA GATE

The bottom or first Chakra gate is sometimes called the Meeting of Yin (referring to the Earth energy), and is located near the perineum, in front of the anal sphincter on the Huiyin CV-1 point. This energy center controls the reproductive system and the urogenital organs and intersects with the Bladder and Kidney Channels. It is associated with survival instincts (the "fight or flight" response), security, tribal power (group acceptance), and the "grounding" or "rooting" of the body.

The first Chakra is associated with the Lower Dantian. Its energetic elemental nature relates to Earth, and its color is red. It connects to the Taiji Pole at the base of the perineum and exits at the top, or Crown Chakra.

The bottom Chakra is associated with the kinesthetic, tactile, and proprioceptive senses. It supplies energy to the spinal column, adrenal glands (in some systems), testicles/ovaries, and Kidneys. This Chakra gate is also connected to both the Conception and Governing Vessels.

The bottom Chakra gate will immediately open in any emergency situation, or when there is danger, to release stored information to ensure survival. Energetic malfunctions caused by stored emotional trauma in this Chakra can cause a massive amount of fear for one's life to be released. The retention of massive amounts of fear, or the sudden energetic release of panic, can cause the following symptoms: lower back pain, sciatica, constipation, hemorrhoids, rectal tumors and cancer, knee problems, and varicose veins.

THE SECOND CHAKRA GATES

The second Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Spirit's Palace Gate and is located at the navel around the Shenque CV-8 point. The back gate is sometimes called the Gate of Life and is located on the lower back at the Mingmen GV-4 point.

This second energetic area is considered the "gateway of the Lower Dantian." Its energetic elemental nature relates to Water, and its color is orange. It is the energy center through which other people's emotions are perceived (kinesthetic perception), allowing one to be sensitive to other people and to sense dangerous situations.

The second Chakra is the energy center for sexuality, sensuality, the hunger for power, and financial issues. It supplies the sexual organs (ovaries and testicles) and the immune system with energy.

When the second Chakra gate opens the patient feels the release of sexual desires, waves of orgasm, and sexual fantasy. Because this area is also associated with the desire for power, sometimes the desire to destroy is released. Energetic malfunctions caused by stored emotional trauma in this Chakra can cause lower back problems, frigidity or impotence. Second Chakra dysfunction can also cause Kidney, uterine, Bladder, and urinary problems.

THE THIRD CHAKRA GATES

The third Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Solar Plexus Center, in this system of Chakra diagnosis it is located between the navel and the xiphoid process (at the level of the pancreas), at the Zhongwan CV-12 (Middle Stomach Cavity) point. The back gate is called the Middle of the Spine and is located on the middle of the back at the Jizhong GV-6 point.

The third Chakra area is associated with the Middle Dantian and is known as the Yellow Court in Medical Qigong therapy. Its energetic elemental nature relates to Fire, and its color is yellow. It is the body's distribution point for psychic energies (gut instinct and intuition), personal power, and self-image. It is the area where the body's Qi transforms into Shen. It is also a storage chamber for severe emotional pain and anger.

This area supplies energy to the Liver, Gall Bladder, Stomach, Spleen, pancreas, adrenal glands, and nervous system. The third Chakra is the center for personal power-storing issues of responsibility, self-esteem, personal honor, and fear of rejection.

When the third charka's energy gates open, feelings of power and waves of anger, rage, fear, greed, jealousy, judgment, and criticism can be released. Energetic malfunctions caused by stored emotional trauma in this Chakra area can cause ulcers, hypoglycemia and diabetes, as well as Liver and / or adrenal problems.

THE FOURTH CHAKRA GATES

The fourth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Heart Center and is located at the center of the breastbone, at the Shangzhong CV-17 point. The back gate is sometimes called the Spirit Path and is located on the Shendao GV-11 point on the back, between the scapula.

The fourth Chakra area is associated with the Middle Dantian. Its energetic elemental nature relates to Wind, and its color is green. This energy center is associated with love, compassion, empathy, clairsentience, and intuition. This area supplies energy to the Heart, Lungs, circulatory system, thymus gland, vagus nerve, diaphragm, and upper back.

When this energy center opens, it can be physically painful (the patient may experience feelings of a weight on the chest, heart attack sensations, or the release of tears) as the patient becomes aware of being attached to the personal ego. The feelings associated with the fourth Chakra have to do with our internal world, while feelings associated with the third Chakra have to do with the external world. Energetic malfunctions can cause Heart and Lung diseases.

THE FIFTH CHAKRA GATES

The fifth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Throat Center and is located just above the hollow of the throat, at the Tiantu CV-22 point. The back gate is sometimes called the Big Vertebra and is located on the back, at the base of the neck on the Dazhui GV-14 point.

This energy center is associated with sound vibration, personal and divine will, and communication. The fifth Chakra area is associated with the Middle Dantian. Its energetic elemental nature relates to sound, and its color is bright blue. It is the center for psychic clairaudience, and the power source for the Qigong doctor's "inner voice" guidance. This area supplies energy to the throat, thyroid and parathyroid glands, neck vertebrae, mouth, Lungs, and alimentary canal (the digestive tube from the mouth to the anus).

When the fifth Chakra energetic center opens, a compulsive urge to sing and chant arises. This area is also where Heat from the Liver Fire sometimes becomes entangled with Phlegm (a viscid substance that is a product and a cause of disease) from the Lungs, forming what is known as "plum pit Qi" (an energetic knot within the throat). Energetic malfunctions caused by stored emotional trauma in this Chakra area can cause stiff neck, sore throat, throat "plum pit Qi" symptom, thyroid problems and swollen glands. The patient sometimes experiences a release of unexpressed grief and anger, resulting in the physical reactions of excessive swallowing, coughing, or laryngitis.

THE SIXTH CHAKRA GATES

The sixth Chakra core can be accessed through two energetic gates. The front gate is sometimes called the Third-Eye Center and is located at the middle of the forehead at the Yintang or Hall of Impression point. The back gate is located on the back of the head, between GV-16 (Wind Palace) and GV-17 (Brain's Door), at the external occipital protuberance.

The sixth Chakra area is the center for psychic intuition, clairvoyant seeing (inner vision), and energy projection. The sixth Chakra area is associated with the Upper Dantian. Its energetic elemental nature relates to light, and its color is indigo (deep violet blue). This area of the body is responsible for enabling the Qigong doctor to see auras, Chakras, and other energetic images. This center is also responsible for mental telepathy, allowing the doctor to sometimes know the patient's thoughts and feelings. This area of the body supplies energy to the pituitary gland, lower brain, ears, nose, eyes, and nervous system.

When the sixth Chakra energy center opens, energetic malfunctions can result in headaches, eye strain, pain around the eyes, distorted or blurred vision, and sometimes blindness. Inaccurate interpretations of events, projections of personal fears and other emotions and motivations onto others may sometimes occur. Reoccurring nightmares and misguided fantasies may also be incurred.

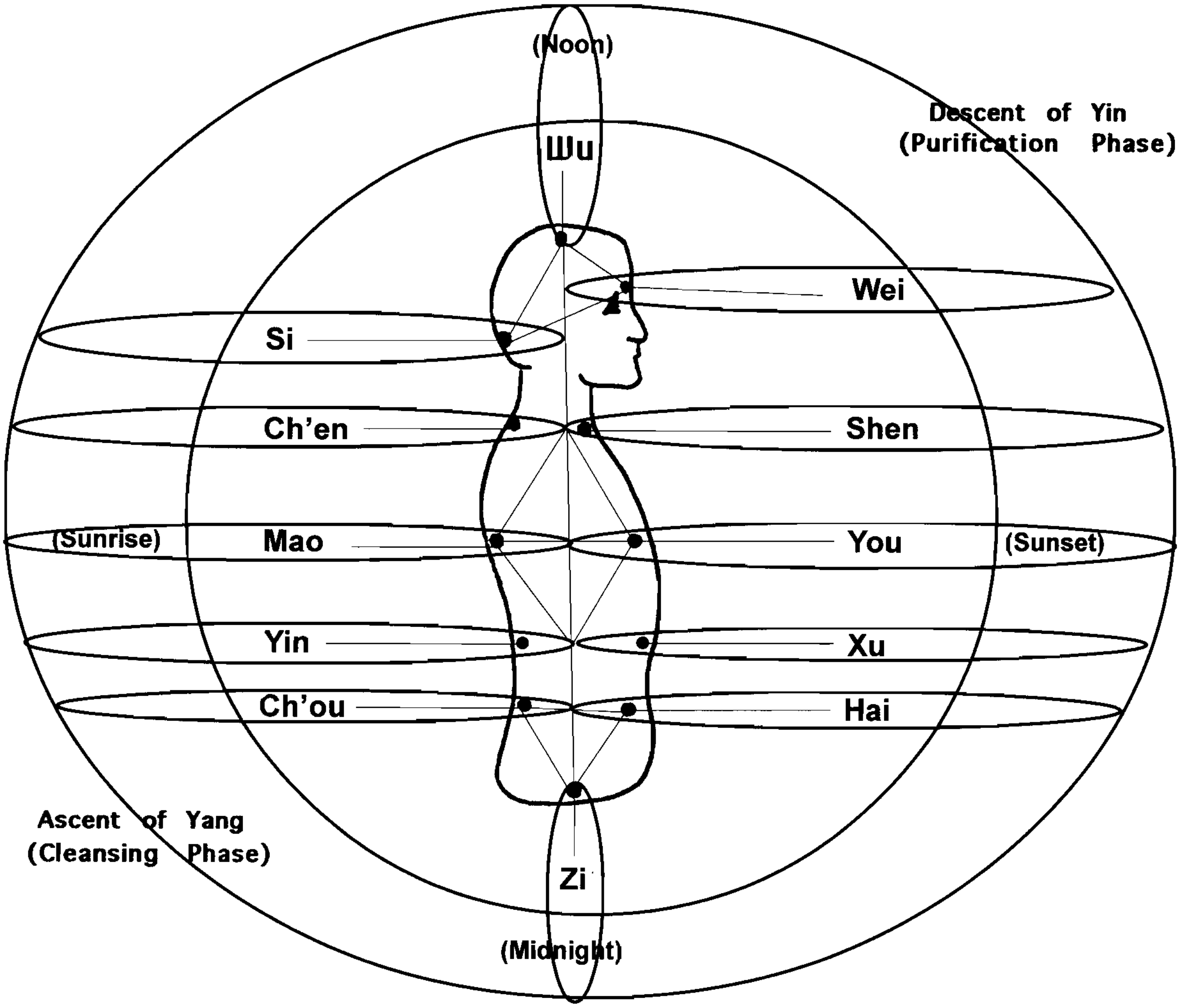
(Noon)
Wu
Descent of Yin
(Purification Phase)
Wei
Si
Ch'en
Shen
(Sunrise)
Mao
You
(Sunset)
Yin
Xu
Ch'ou
Hai
Ascent of Yang
(Cleansing Phase)
Zi
(Midnight)

Figure 5.12. This figure shows the Twelve Earthly Branch Relation to the Twelve Chakra Gates. The Governing and Conception Vessels flowing through the middle of the back and front of the body correspond to the ecliptic path of the sun along the Twelve Chakra Gates, and are related to the Twelve Earthly Branches.

THE UPPER CHAKRA GATE

The upper or top Chakra gate is sometimes called the Crown Center and is located on top of the head at the Baihui GV-20 point. This energy center is associated with higher knowledge, understanding, pure intuition, and ecstasy. The upper Chakra area is associated with the Upper Dantian. Its energetic elemental nature relates to infinite space (Wuji), and its color is violet. This area supplies energy to the pineal gland and upper brain.

When the upper Chakra energetic center opens, it is like a golden flower opening, accompanied by a dazzling bright light and a strong, powerful connection with the divine. Energetic malfunctions result in confusion, apathy, alienation, boredom, depression, and a state of incomprehension.

TAKING PRECAUTIONS

When any of the energetic reactions to opening the Chakra gates take place, it is important for the student or patient to realize that these are only transitional stages and are normal phenomena which occur as a sign of energetic and spiritual transformations. The student or patient should take certain precautions, however, like being under the care of a qualified Qigong doctor when beginning any type of energy training.

It is also important not to allow patients to experience these emotional transitions alone. Because these energy transitions, malfunctions, and deviations are new to the Western mode of thinking, they can easily be misdiagnosed by doctors unfamiliar with energetic medicine. Patients should be encouraged to find or establish a support group consisting of advanced practitioners of energetic meditation. Having a support group of experienced practitioners allows these emotional and spiritual transitions to occur in a safe environment, where other experienced practitioners can monitor the patient's feelings, if and when, unfamiliar emotions start to emerge.

THE TWELVE EARTHLY BRANCHES AND THE TWELVE CHAKRA GATES

The body is viewed as a small and complete universe unto itself. The internal organs are influenced by the celestial movements of the sun, moon, planets, and stars. The Governing and Conception Vessels, where the Twelve Chakra Gates are located, are also affected by the Heavenly cycles.

In ancient China, the day was divided into twelve separate time divisions. Each time division corresponds to two hours of the day and was named after one of the Twelve Earthly Branches. These twelve time divisions were further organized into months and seasons (see Chapter 11). The ancient Qigong doctors discovered that the body's Qi and Blood mirror the Earth's seasonal ebb and flow, rising and falling like the lunar tides.

Each of the time periods corresponding to the Twelve Earthly Branches is regarded as having a specific influence on each of the Twelve Gates of the body's Chakra system. The rhythmic variations of the waxing and waning of Qi and Blood is associated to the waxing and waning of Yin and Yang energy, as well as the circulation of Qi following the Microcosmic Orbit (Fire) cycle.

Each of the Twelve Chakra Gates relates to one of the Twelve Earthly Branches following the Microcosmic Orbit (Fire) cycle (see Chapter 41). These Twelve Chakras Gates extend their energy from the patient's Taiji Pole outward, along the center of the body, through the anterior and posterior fields of Qi. Beginning at the bottom Chakra gate (the "Zi" Branch, representing midnight), the energy follows the Fire Cycle of the Microcosmic Orbit, traveling up the Governing Vessel, following the "ascent of Yang." After the Yang Qi reaches its peak, the Yin begins to grow. Starting at the upper Chakra gate (the Wu Branch, representing noon time), the energy travels down the Conception Vessel, following the "descent of Yin" (Figure 5.12).

CHAPTER 6
THE TWELVE PRIMARY CHANNELS AND ORGANS

INTRODUCTION

According to Traditional Chinese Medicine, a system of channels exists that integrates all the body's separate parts and functions into a unified organism. The understanding of these channels and their energetic function and interactions provides the Qigong doctor with the basis for understanding the relationships among the various physiological, pathological, diagnostic, and therapeutic relationships in Traditional Chinese Medicine.

The study of Chinese medicine rests on, and is inspired by, the circulation of Qi, as many diseases follow a predictable course of development. The pathology in an internal organ will often manifest itself in certain external or systemic symptoms; therefore, understanding the circulation of Qi flow allows the doctor to control each organ's energetic function. By stimulating specific sites on the body's surface, diseases in both the superficial tissues and internal organs can be treated.

A doctor of Medical Qigong uses various techniques that facilitate the opening of the major energetic connectors of the body's internal organs. These connectors, called channels and vessels, are responsible for connecting the flow of Qi within the whole body to transfer Qi both internally and externally.

The channels are also known as meridians or Jing-luo. The Chinese character for this type of Jing means "to move through," while the translation of the character luo means "a net" (the body's inner fascia). Along these channels are major trunks and lesser branches that connect internally with the vital organs and externally with the major channels, limbs, sensory organs, and orifices.

The smaller branches are known as collaterals. The tiny areas along these channels and collaterals where the Qi pools are called points. These points are the spots where the patient's spirit and energy enter and leave the body. When stimulated, these channel points (sometimes also called "acupuncture points"), cause an energetic response from within the internal organs and channels, resulting in internal Qi flowing from organ to channel, from channel to channel, or from point to point along the same channel.

The channels unite the body as one unit. Channel theory is interrelated with organ theory. Traditionally, the internal organs have never been regarded as simply independent anatomical entities. Rather, Traditional Chinese Medicine focuses on the functional and pathological interrelationships between the channel network and the organs. The internal organs and their systems are sometimes known as "orbs" or "spheres of influence."

Qi travels throughout the physical body along the channels and collaterals much as water flows through rivers and streams. An ancient Chinese medical text explains this concept as follows: "Heaven is covered with the constellations, Earth with the waterways, and man with channels." Qigong doctors of ancient China paid much more attention to the Body Fluids and energies circulating through the body than to the physical anatomical structures. The ancient Chinese considered the Body's Fluids and energies much more fundamental.

The channels serve as the link between the energies, such as Jing, Qi, Shen, and the ingredients, such as Blood and Body Fluids, that feed and enliven the tissues. The human body's main energetic rivers are the Twelve Primary Channels and Eight Extraordinary Vessels (see Chapter 7).

CLASSIFICATION OF CHANNELS

The channels are classified into four main categories: Jing, Luo, Muscle Tendon, and Skin Zones.

1. The Jing Channels are the Primary Channels. These channels include the Eight Extraordinary Vessels, the Twelve Primary Channels, and the Twelve Divergent Channels.
2. The Luo Channels are the Collaterals. These collaterals include the Fifteen Major Collaterals, the Minute Collaterals, and the Superficial Collaterals.
3. The Muscle/Tendon Channels consist of twelve channels that serve as external connections to the major channels, flowing through the body's muscles, tendons, and ligaments.
4. The Cutaneous Regions are skin zone areas where the channels surface on the external tissues.

JING CHANNELS

The Jing Channels are the major trunks of the circulatory tree. They generally flow vertically through the body through relatively deep tissues. They also transverse the limbs peripherally and penetrate the body cavities to connect with the internal organs (Figure 6.1). They are the body's main interior and exterior rivers of Qi.

1. The Eight Extraordinary Vessels connect with, and regulate, the Qi and Blood of the Twelve Primary Channels by either absorbing any energetic "runoff" in times of channel Excess, or by replenishing energy in times of Deficiency.
2. The Twelve Primary Channels are bilateral and symmetrical. They can be identified in three specific ways: (1) according to the corresponding Yin or Yang organ to which they are connected, (2) by the arms or legs in which the channels originate or end, and (3) according to the six divisions of Yin or Yang Qi to which the channels relate (i.e., Tai Yang, Yang Ming, Shao Yang, Tai Yin, Shao Yin, and Jue Yin).

   The Yin channels run along the medial and anterior aspects of the body. They are associated with the solid Yin organs and connect with the hollow Yang organs. The Yin channels include the Kidneys, Spleen, and Liver Channels that flow from the feet to the torso, a well as the Heart, Pericardium, and Lung Channels that flow from the torso to the hands.

   The Yang channels run along the lateral aspect of the body. They belong to the hollow Yang organs and connect with the solid Yin organs. The Yang channels include the Small Intestine, Triple Burners, and Large Intestine Channels that flow from the hands to the head, as well as the Bladder Gall Bladder and Stomach Channels that flow from the head to the feet. The interaction between the internal/external, Yin/Yang relationship of the organs and channels results in an alchemical transference between energy and matter.
3. The Twelve Divergent Channels branch off the Twelve Primary Channels. They are mainly distributed on the chest, abdomen, and head. The Divergent channels have the energetic functions of connecting internally and externally related channels, strengthening the connection of the Twelve Primary Channels to their related organs, and serving as extensions of the Twelve Primary Channels.

LUO OR COLLATERALS

The Twelve Primary Channels are connected by collaterals, which are small, interlinking streams. These collaterals are the connecting branches of the energetic circulatory system. They generally flow superficially, in horizontal and vertical energetic directions. They are the body's secondary streams of Qi that form an intricate network that traverses the body's surface and interconnects the main rivers, connective tissues, and cutaneous regions (Skin Zones).

1. The Fifteen Major Collaterals (Luo) transfer Qi and Blood from the Twelve Primary Channels to all parts of the body and link the body's interior with its exterior, connecting the body's internal and superficial channels, as well as the Governing and Conception Vessels and the Great Luo of the Spleen (see Chapter 7).

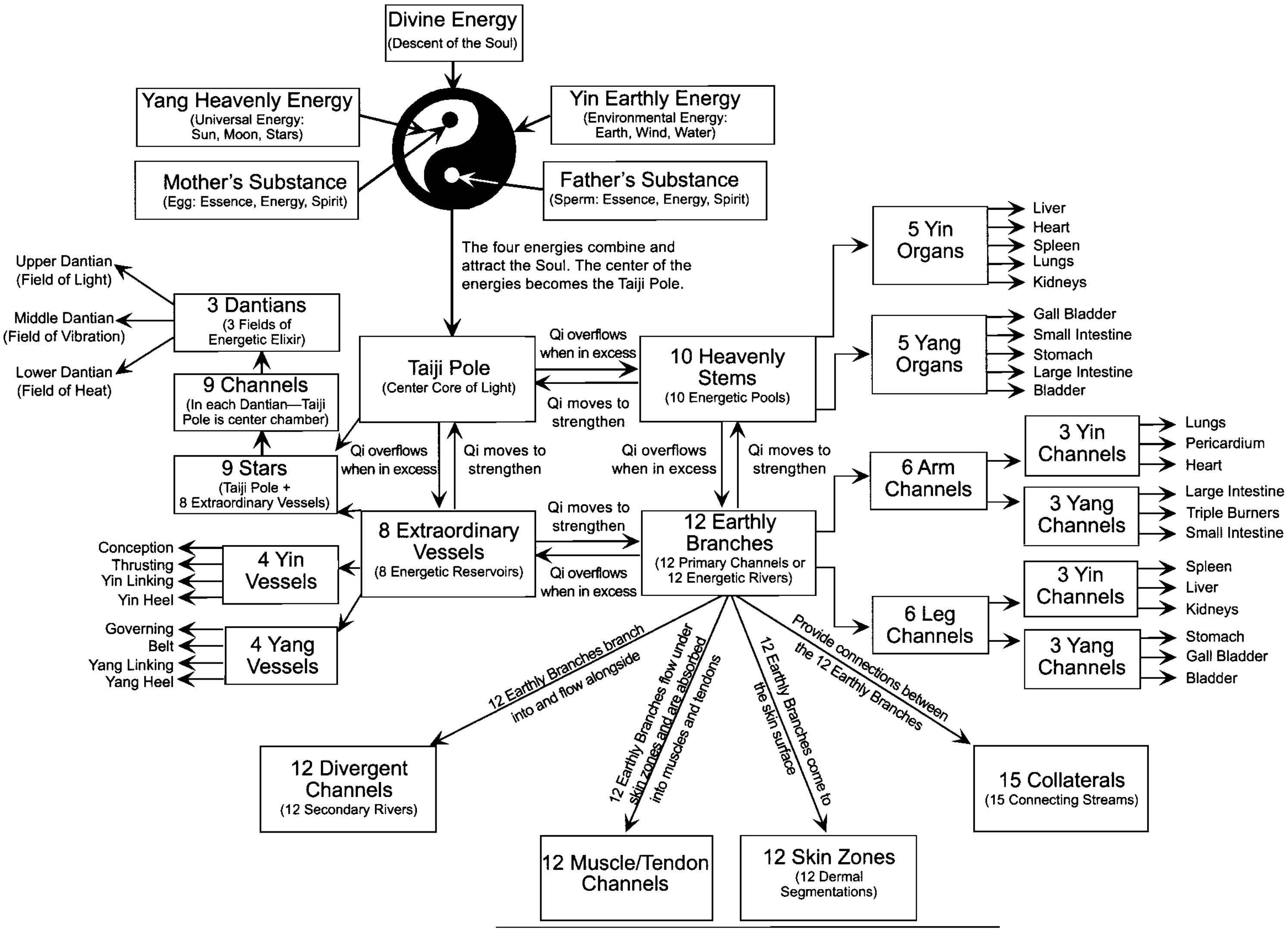
Divine Energy
(Descent of the Soul)
Yang Heavenly Energy
(Universal Energy: Sun, Moon, Stars)
Yin Earthly Energy
(Environmental Energy: Earth, Wind, Water)
Mother's Substance
(Egg: Essence, Energy, Spirit)
Father's Substance
(Sperm: Essence, Energy, Spirit)
The four energies combine and attract the Soul. The center of the energies becomes the Taiji Pole.
Upper Dantian
(Field of Light)
Middle Dantian
(Field of Vibration)
Lower Dantian
(Field of Heat)
3 Dantians
(3 Fields of Energetic Elixir)
9 Channels
(In each Dantian—Taiji Pole is center chamber)
9 Stars
(Taiji Pole + 8 Extraordinary Vessels)
Taiji Pole
(Center Core of Light)
Qi overflows when in excess
Qi moves to strengthen
10 Heavenly Stems
(10 Energetic Pools)
5 Yin Organs
Liver
Heart
Spleen
Lungs
Kidneys
5 Yang Organs
Gall Bladder
Small Intestine
Stomach
Large Intestine
Bladder
Qi overflows when in excess
Qi moves to strengthen
Qi overflows when in excess
Qi moves to strengthen
8 Extraordinary Vessels
(8 Energetic Reservoirs)
Qi moves to strengthen
Qi overflows when in excess
12 Earthly Branches
(12 Primary Channels or 12 Energetic Rivers)
Conception
Thrusting
Yin Linking
Yin Heel
4 Yin Vessels
Governing
Belt
Yang Linking
Yang Heel
4 Yang Vessels
6 Arm Channels
3 Yin Channels
Lungs
Pericardium
Heart
3 Yang Channels
Large Intestine
Triple Burners
Small Intestine
6 Leg Channels
3 Yin Channels
Spleen
Liver
Kidneys
3 Yang Channels
Stomach
Gall Bladder
Bladder
12 Earthly Branches branch into and flow alongside
12 Earthly Branches flow under skin zones and are absorbed into muscles and tendons
12 Earthly Branches come to the skin surface
Provide connections between the 12 Earthly Branches
12 Divergent Channels
(12 Secondary Rivers)
12 Muscle/Tendon Channels
12 Skin Zones
(12 Dermal Segmentations)
15 Collaterals
(15 Connecting Streams)

Figure 6.1. Channel and Organ Energy Flow

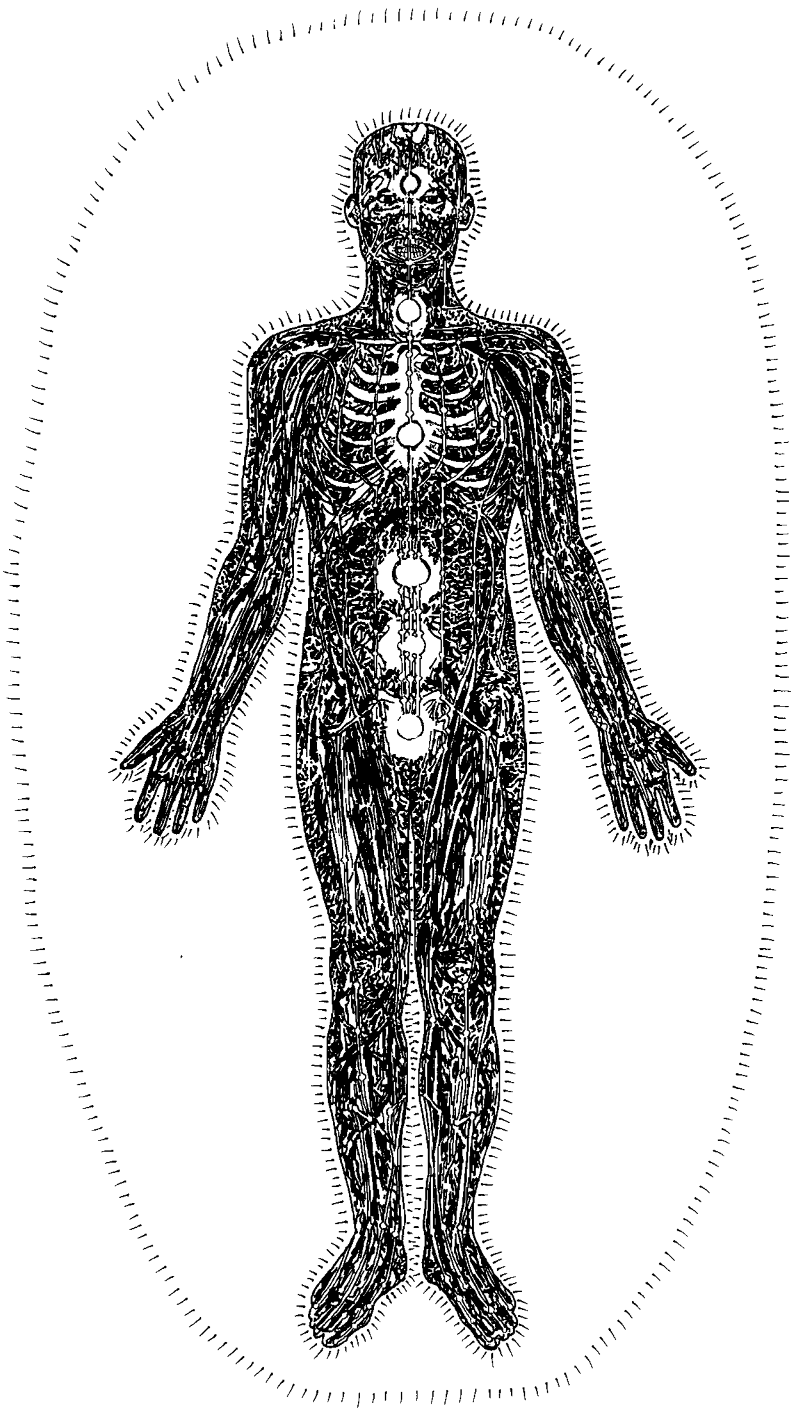

Figure 6.2. The Human Body's Energetic Matrix

2. The Minute Collaterals are smaller branches of the Fifteen Major Collaterals. They are countless in number.
3. The Superficial Collaterals are smaller branches of the Minute Collaterals. They are also countless in number.

MUSCLE TENDON CHANNELS

The Twelve Muscle / Tendon Channels are the external connections of the major channels. They generally flow superficially, traversing the body's surface and serve to join the main rivers, connective tissues, and Cutaneous Regions. The Twelve Muscle / Tendon Channels are also regions of the body where the Qi and Blood of the Twelve Primary Channels nourish the muscles, tendons, and ligaments.

CUTANEOUS REGIONS

The Twelve Cutaneous Regions (also called Twelve Skin Zones) are areas of the body where the Blood and Qi of the Twelve Primary Channels surface and connect to the body's skin tissues. These cutaneous regions have a continuous and direct contact with the external environment.

The body resonates in a continuous interplay of Yin and Yang harmony, balancing the channels' and organs' energy flows. The channels help create the body's external Yin and Yang polarities and create the energetic matrices around which the body's energetic fields can flow. Together these energetic matrices establish the foundation of the body's energetic tissues (Figure 6.2).

THE CHANNELS' RELATIONSHIP TO QI AND BLOOD

The channels transport Qi and Blood to nourish, moisten, and vitalize the whole body. Healthy bodily function depends on the balanced circulation of Qi and Blood. Qi is Yang and provides the energy or force necessary for the body's functional activity (i.e., movement and transportation, warming, containing, transforming, and defending). Blood is Yin and is the source of the body's moistening, nourishment, and lubrication.

The Blood, Qi, and heat circulate through the Twelve Primary Channels every two-hours, ebbing and peaking in energetic flow. The maximum peak of a channel's Qi, or "high-tide," occurs at the time of day when the Qi in that channel is at its fullest. The minimum period of a channel's Qi, or "low tide," occurs twelve hours from the peak time (Figure 6.3).

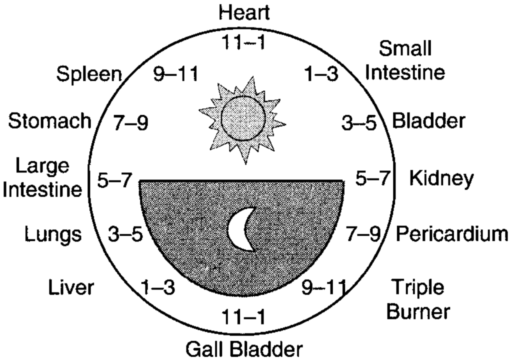

Figure 6.3. The Body's High and Low Energetic Tides

The connective tissues of the body's fasciae transmit a variety of biological energies. Think of the inner fascia as lubricated linings that cover the muscles and internal organs, allowing the tissues to slide and move easily. Within the lining of the superficial fasciae, the channels are webbed; this webbed network conducts energy in the form of electron, proton, and ion transmission, and is influenced by light and sound.

According to Traditional Chinese Medicine theory, each of the Twelve Primary Channels has its own individual resources of Qi and Blood; however, the energetic quality varies in accordance with the amount of Blood and Qi available in each organ and its particular channels. From a traditional Chinese medical perspective, the Spleen plays two important roles in regard to Qi and Blood. First, it converts food into Gu Qi which, when further refined by the Liver and Kidneys, becomes Ying (Nutritive) Qi; and second, it manages the Blood by keeping it in the channels. Ying Qi secretes Body Fluids that enter into the blood vessels and are transformed into Blood by the Heart. The Heart both creates and governs the Blood; whereas, the Liver stores it and spreads Qi throughout the body. Blood is considered to be a denser, more material form of Qi. Qi and Blood

Organs & Channels	Qi	Blood
Liver	less	more
Lungs	more	less
Large Intestine	balanced	balanced
Stomach	balanced	balanced
Spleen	more	less
Heart	more	less
Small Intestine	less	more
Bladder	less	more
Kidney	more	less
Pericardium	less	more
Triple Burners	more	less
Gall Bladder	more	less

Figure 6.4. The Body's Organ and Channel Relationship to the Proportions of Qi to Blood

are interdependent: Qi moves the Blood, but the Blood nourishes Qi; both flow together.

The chart in Figure 6.4 shows each of the Twelve Primary Channels and the proportion of Qi and Blood each channel carries.

FUNCTION OF THE CHANNELS

The channels are responsible for responding to any malfunction in the body. They can also, however, be disseminators of energetic dysfunction and disease. Disease passes in to the internal organs, or from one organ to another, via the channels. Exogenous diseases progress from the body's skin pores to the tiny collaterals, then to primary channels, and finally, to the internal organs. The channels are affected by disease in a predictable fashion.

1. Diseases and symptoms manifest along the channel pathway.
2. Points along the channels become tender, painful, tight, or flaccid.
3. When in a deficient state, the electrical resistance and heat tolerance diminish on the skin at the diseased point(s).
4. The body's sensory organs (sight, smell, taste, hearing and touch) are affected through channel relationship.

From a traditional Chinese medical perspective, the functions and flow of the channels and collaterals, as well as the internal organs' health can be affected by Medical Qigong, acupuncture, Chinese massage, and herbal medicine. Each system of Traditional Chinese Medicine disperses Excess, moves stagnation, stimulates and nourishing the Qi to tonify Deficiency, and stimulates the Wei Qi to fight External pathogens via the channels. In fact, each time the body's channels and collaterals are trained or regulated, they become thicker with Qi.

INTERNAL AND EXTERNAL CHANNEL FLOW

As the Qi of an internal channel flows into or out from its "primary organ," it generates a current that is similar to a river rushing back into the sea. The moment that energy flowing through the internal channel permeates the primary organ, an energetic connection and reaction is immediately created in the organ tissues. When the Qi of an internal channel flows into its "associated organ," however, the energetic reaction is quite different. Instead of immediately permeating the organ, the channel's energy "spirally wraps" the organ's tissues through the internal fascia that envelops the associated organ. The associated organ tissues then absorb the energy from the surrounding fascia. This difference in energetic penetration and absorption rate is an important factor when working with the body's internal energetic channels. When Qigong doctors extend their energy into a patient's channels, they must be aware that the absorption rate will be immediately felt in the channels pertaining organ, and will have a gradual effect on the organ's associated organ.

Channel Name	Yin or Yang	Associated Organ	Element	Blood Heat Cycle	Qi Blood Cycle	Abbre-viation
Gall Bladder	Yang	Liver	Wood	11 p.m.– 1 a.m.	more Qi less Blood	G.B.
Liver	Yin	Gall Bladder	Wood	1 a.m.– 3 a.m.	less Qi more Blood	Liv.
Lungs	Yin	Large Intestine	Metal	3 a.m.– 5 a.m.	more Qi less Blood	Lu.
Large Intestine	Yang	Lung	Metal	5 a.m.– 7 a.m.	balanced Qi and Blood	L.I.
Stomach	Yang	Spleen	Earth	7 a.m.– 9 a.m.	balanced Qi and Blood	St.
Spleen	Yin	Stomach	Earth	9 a.m.– 11 a.m.	more Qi less Blood	Sp.
Heart	Yin	Small Intestine	Fire	11 a.m.– 1 pm.	more Qi less Blood	H.
Small Intestine	Yang	Heart	Fire	1 p.m.– 3 p.m.	less Qi more Blood	S.I.
Bladder	Yang	Kidney	Water	3 p.m.– 5 p.m.	less Qi more Blood	Bl.
Kidney	Yin	Bladder	Water	5 p.m.– 7 p.m.	more Qi less Blood	K. or Kid.
Pericardium	Yin	Triple Burner	Fire	7 p.m.– 9 p.m.	less Qi more Blood	P.
Triple Burner	Yang	Pericardium	Fire	9 p.m.– 11 p.m.	more Qi less Blood	T.B.

Figure 6.5. The Twelve Primary Channels are divided into Yin and Yang. The Yin channels are connected to the solid organs, and the Yang channels are connected to the hollow organs.

ORGAN AND CHANNEL FUNDAMENTALS

The purpose for energetic anatomy and physiology is to understand the fundamental actions of each organ and each channel's unique energetic characteristics. This understanding includes the learning of differential diagnosis to categorize the various symptoms into meaningful patterns of disharmony. This ability to categorize the various symptoms will enable the doctor to analyze and recognize specific pathologies indigenous to specific organs and channel energies (Figure 6.5).

In Chinese medical science, the study of energetic physiology is called Zhang Xiang Xue Shou, which literally translates as "the theory of the phenomena of internal organs." Here, the word "phenomena" means visible external manifestations. The organ's physical structure is of little consequence; whereas, its function in the body is of primary importance. While Western anatomy and physiology are primarily concerned with the physical body in its most concrete form (investigating the structures of the major body systems and organs), energetic anatomy and physiology focuses mainly on the underlying energy that enlivens and sustains the physical form.

In Chinese medical science, the body's internal organs are responsible for creating and distributing Qi, Blood and Body Fluids. Traditional Chinese Medicine divides Body Fluids into two distinct categories: Jin, which are the body's clear, light fluids, and Ye, which are the body's thick and heavy liquids.

1. The Jin Body Fluids are the Yang Fluids of the Upper Burner that move with Wei Qi. They are fast, light, and clear fluids, such as: sweat, tears, saliva, and mucus. The Lungs control the Jin.
2. The Ye Body Fluids are the Yin Liquids of the Middle and Lower Burners that move with Ying Qi. These are slow, heavy, and thick liquids. These liquids include those that moisten the joints, spine, brain, and Bone Marrow and lubricate the eyes, ears, nose, and mouth. The Spleen and Kidneys control the Ye.

This internal action of creating and distributing Qi, Blood and Body Fluids is manifested throughout the body's tissues through the Blood's Heat Cycle, which actively follows a peak (high-tide) and ebb (low-tide) energetic time period. Each organ and channel has either a Yin or Yang characteristic and is grouped and paired according to its Five Element nature.

This chapter focuses on the Twelve Primary Channels and their relationship to the organs. Most Traditional Chinese Medical texts begin the sequence of these channels with the Lung Channels. In this text, however, due to the energetic waxing and waning of the Yin and Yang Yao Cycles, and the Twelve Pi Hexagrams (see Chapter 11), I will start with the Gall Bladder Channels. The Twelve Primary Channels are described as follows: Gall Bladder, Liver, Lung, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidney, Pericardium, and Triple Burner.

THE GALL BLADDER CHANNELS

The Gall Bladder Channels are Yang channels. The external branches flow from the head to the feet on both sides of the body (Figure 6.6). The internal branches of the Gall Bladder Channels flow from the head to the Gall Bladder. The internal and external branches of these two rivers originate from the outer canthus of each eye. The external branches zigzag around the side of the head, flowing down the sides of the torso. The internal branches leave the outer canthus of the eye and descend the neck, entering the supraclavicular fossa, where they meet the main channels at the St-12 and Pc-1 points, and pass through the diaphragm, spirally wrapping the Liver, and permeating the Gall Bladder. From there they encircle the genitals and enter deeply into the body to emerge at the sacrum.

The external rivers continue to descend the lateral aspects of the torso and legs, ending at the lateral sides of the tips of the fourth toes.

THE GALL BLADDER ORGAN

The Gall Bladder's associated organ is the Liver; its element is Wood. The Gall Bladder's function of storing and excreting bile depends on the Liver's function of maintaining a smooth flow of Qi throughout the body.

The Gall Bladder also helps the Liver control the sinews, tendons, and ligaments. The Liver provides Blood to the sinews, while the Gall Bladder provides Qi to the sinews in order to ensure proper movement and agility.

PSYCHO-EMOTIONAL ASPECTS

The Gall Bladder is responsible for making decisions and judgments, as well as providing courage and initiative. This organ is sometimes called The Court of Justice or The General's Advisor. Although the Kidneys control drive and vitality, the Gall Bladder provides the capacity to turn this drive and vitality into decisive action.

The Gall Bladder has an influence on the quality and length of sleep. If the Gall Bladder is Deficient, the patient will often wake up suddenly, very early in the morning, and be unable to fall asleep again.

Patient's who are timid, indecisive, and easily discouraged by slight adversity, are said to have a weak Gall Bladder; conversely, decisive and determined patients are said to have a strong Gall Bladder.

GALL BLADDER'S ENERGY FLOW

The Gall Bladder Channels are often used to drain off the energy from an overheated or Excess Liver condition. Energetically, the Gall Bladder Channels store more Qi than Blood, acting more on energetic and nervous functions than on physical substances and Blood functions. At the

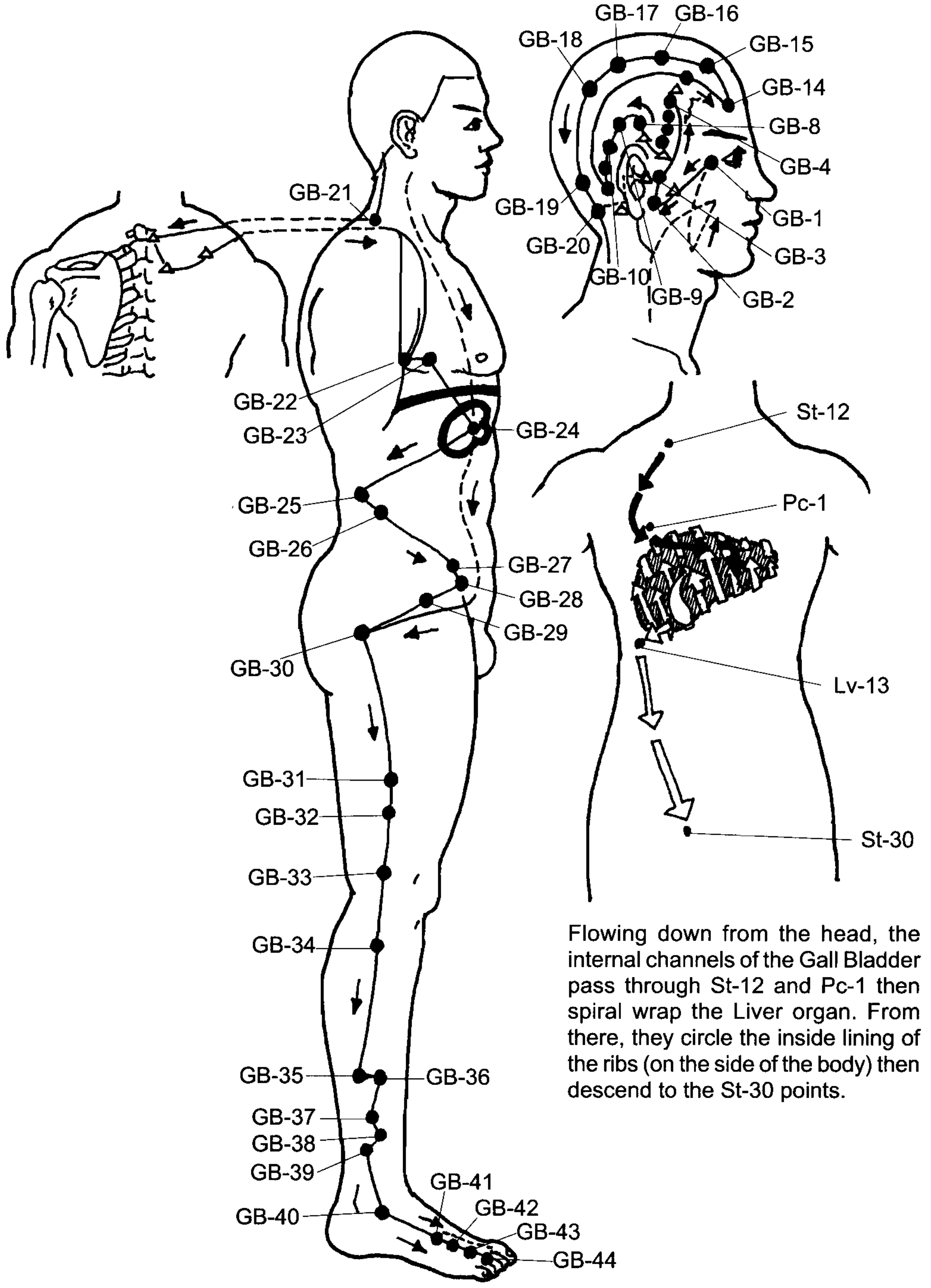
GB-17
GB-16
GB-18
GB-15
GB-14
GB-8
GB-4
GB-1
GB-3
GB-2
GB-9
GB-10
GB-20
GB-19
GB-21
GB-22
GB-23
GB-24
GB-25
GB-26
GB-27
GB-28
GB-29
GB-30
St-12
Pc-1
Lv-13
St-30
GB-31
GB-32
GB-33
GB-34
GB-35
GB-36
GB-37
GB-38
GB-39
GB-40
GB-41
GB-42
GB-43
GB-44

Flowing down from the head, the internal channels of the Gall Bladder pass through St-12 and Pc-1 then spiral wrap the Liver organ. From there, they circle the inside lining of the ribs (on the side of the body) then descend to the St-30 points.

Figure 6.6. The Internal and External Qi Flow of the Gall Bladder (GB) Channels

high-tide time period (11 p.m. to 1 a.m.) Qi and Blood abound in the Gall Bladder, and so the Gall Bladder organ and channels can more easily be dispersed and purged. During low-tide (11 a.m. to 1 p.m.) they can be easily tonified. The Gall Bladder Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The gall bladder stores then releases bile through the cystic duct into the common bile duct that flow into the small intestine to aid in the digestion of fats.

PATHOLOGICAL MANIFESTATIONS

The main symptoms of the Gall Bladder Channel diseases are:

- pain in the right-and left-upper quadrants of the abdomen,
- disorders of the head, which include the eyes, ears, and face, and
- disorders of the external sides of the legs along the channels' pathways.

THE LIVER CHANNELS

The Liver Channels are Yin channels that flow externally from the feet to the torso (Figure 6.7). These two rivers originate externally from the inside of the big toes and flow upward on the inside the legs, to circle the groin. From there they continue to ascend externally to the lateral aspects of the thorax, where they enter internally and permeate the Liver. They then connect to and spirally wrap the Gall Bladder. From there, they flow to the Lungs, ascend internally into the thorax, and connect with the eyes, cheeks, and inner surface of the lips. They emerge from the forehead and connect with the Governing Vessel at the Baihui GV-20 point (Figure 6.8).

THE LIVER ORGAN

The Liver is responsible for the "free and easy wandering" of Qi in the body, i.e., the Liver makes the Qi flow smoothly in and around the body. The Liver governs the Belt and Thrusting Vessels, as well as the body's Qi flow and circulation. The smooth flow of Qi ensures normal mental and emotional activity, as well as normal secretion of bile. Impairment of this Liver function leads to a "binding depression of Liver Qi," associated with impatience, hasty decisions impulsive actions, and a hot temper.

The Liver's associated organ is the Gall Bladder; its element is Wood. The Liver energy opens externally at the eyes and manifests externally through the fingernails and toenails. The Liver stores the Blood, and serves as a reservoir of Blood to regulate the circulation of Blood volume. In women it assists the uterus and regulates the menses. The Liver regulates the function and control of the muscles, tendons, and ligaments and is the source of the body's physical strength. It also stores the emotion of anger.

PSYCHO-EMOTIONAL ASPECTS

The Liver is responsible for planning and creativity, as well as instantaneous solutions or sudden insights; it is therefore considered The General in Charge of Strategy. The Liver houses the body's Hun and governs fright.

Its positive psycho-emotional attributes are kindness, benevolence, compassion, and generosity; its negative attributes are anger, irritability, frustration, resentment, jealousy, rage, and depression.

The Liver is also called the "root of resistance to fatigue." Whenever the Liver is not functioning properly (stagnate or excessively Hot due to suppressed emotions) the patient can experience fatigue as well as physical weakness.

THE LIVER'S ENERGY FLOW

The Liver Channels move the Qi to the whole body, storing and distributing the Blood. The Liver Channels contain more Blood than Qi, thus they affect physical substances more than they affect energetic functions. At the high-tide time period (1 a.m. to 3 a.m.), Qi and Blood abound in the Liver, and so the Liver organ and channels can more easily be dispersed and purged.; whereas, during low tide (1 p.m. to 3 p.m.), they can be easily tonified. The Liver Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The liver stores vitamins A, B-12, D, E, and K, and regulates blood volume. It is one of the main

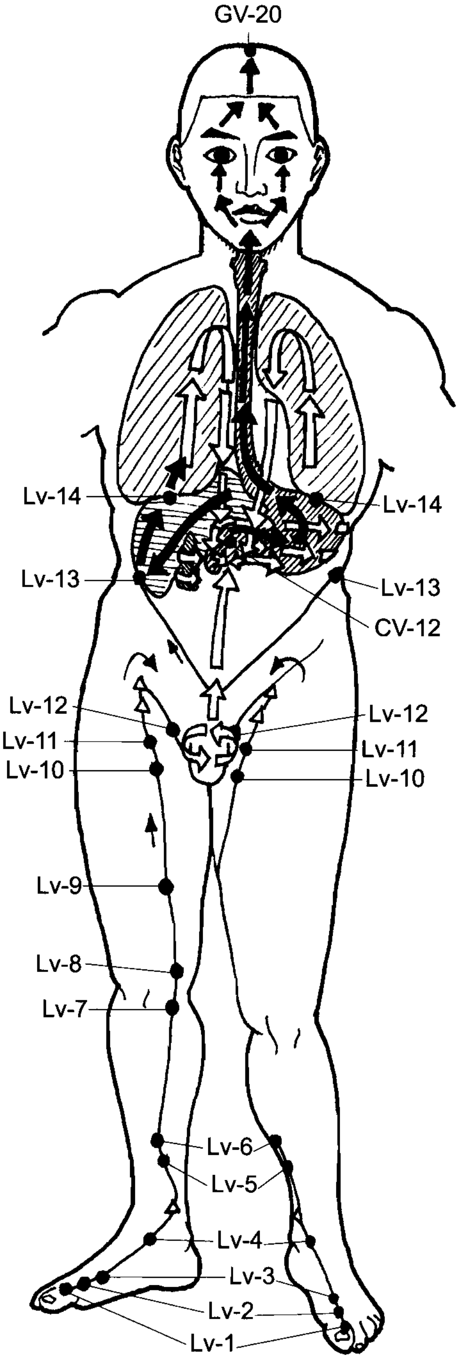
GV-20
Lv-14
Lv-14
Lv-13
Lv-13
CV-12
Lv-12
Lv-12
Lv-11
Lv-11
Lv-10
Lv-10
Lv-9
Lv-8
Lv-7
Lv-6
Lv-5
Lv-4
Lv-3
Lv-2
Lv-1

Figure 6.7. The Internal and External Qi Flow of the Liver (Lv) Channels

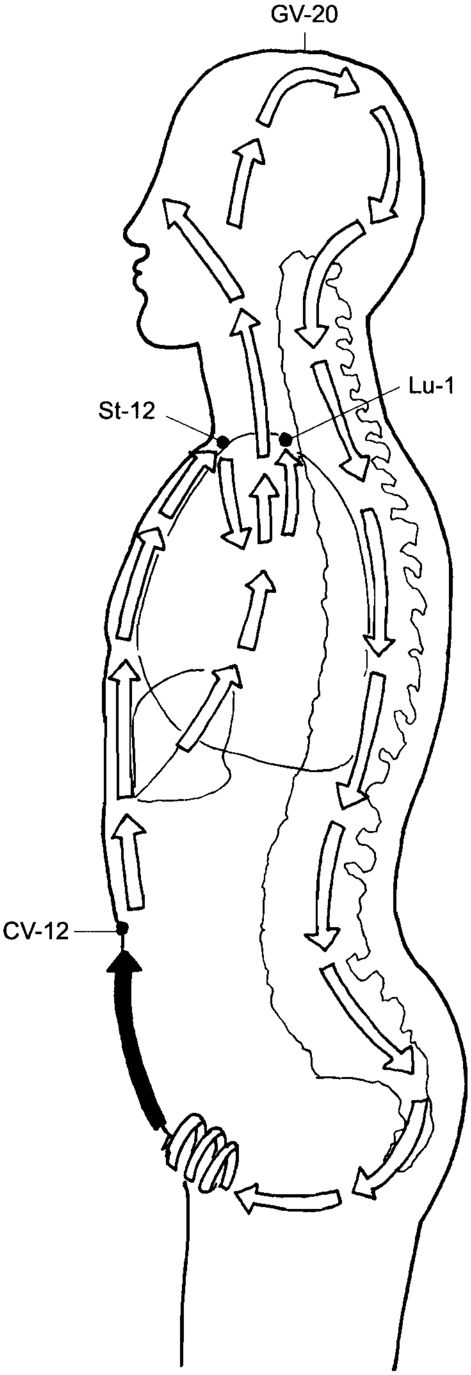
GV-20
Lu-1
St-12
CV-12

Figure 6.8. Side View of the Liver Channel's Internal Flow of Energy

sources of the body's heat. The liver manufactures cholesterol and functions in lipid metabolism and controls bile secretion.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Liver Channels cause swelling and a distended sensation of the hypochondrium. The Liver Channels also correspond to diseases of the lower abdomen and genital organs.

Since the Liver supplies the tendons with nutrients to develop physical strength, Liver impairment causes numbness, tremors, or spasms of the muscles, and sluggishness in joint movements.

The Liver and Kidneys are mutually dependant upon each other. The Liver stores the Blood that nourishes the Kidney Jing; whereas the Kidneys store Jing that helps produce the Blood. Deficient Kidneys may lead to Blood Deficiency, and Deficient Liver Blood may cause weakness of the Kidney Jing due to lack of nourishment from the Blood. The hair on top of the head is also nourished by the Blood. When the hair turns grey, it is often alleged to be caused by insufficient Blood stored in the Liver, as well as a Kidney Jing weakness.

THE LUNG CHANNELS

The Lung Channels are Yin channels and flow externally from the torso to the hands. The main river originates internally from the Middle Burner, in the middle of the chest, and descends downward, connecting with and spirally wrapping the Large Intestine. From there, it ascends along the upper surface of the Stomach passage and through the diaphragm, where it branches and permeates both Lungs (Figure 6.9).

From the Lungs, the rivers of Qi merge and ascend into the pit of the throat, where they separate again into two channels that transverse below the clavicle. These two main channels then surface externally to descend down the arms, and end on the outer side of each thumb. A small stream of energy branches off each wrist at the Lu-7 point and runs directly to the radial side of the tip of the index finger, where it connects with a branch of the Large Intestine Channels.

THE LUNG ORGAN

Since the Lungs are the uppermost organs, they are compared to canopies that shelter and protect all other internal organs. The Lungs' associated organ is the Large Intestine; its element is Metal. The Lung Qi opens externally at the nose and manifests itself through the voice (i.e., when the Lung Qi is in Excess the voice is too loud; when Deficient the voice is too soft). The Lungs govern the body's Qi and respiration. They regulate the metabolism of Water passages (sweat and Body Fluids), the opening and closing of the pores, the skin and the hair's texture.

The Lungs also control the circulation of Qi in both vessels and channels, as well as the dispersing or spreading of the body's Wei Qi (protective energy). They control the descending and holding functions of the body's Qi and Body Fluids.

All blood vessels lead to the Lungs, i.e., all Blood within the body must pass through the Lungs.

PSYCHO-EMOTIONAL ASPECTS

It is said that the Lungs are "the priest" or The Minister of Heaven and are responsible for establishing the foundation of Qi for the entire body (see Chapter 12). The Lungs house the body's Seven Corporeal Souls (Po) and are responsible for self-protection and self-preservation.

The Lungs positive psycho-emotional attributes are righteousness, dignity, integrity, and high self-esteem; their negative attributes are disappointment, sadness, grief, despair, anxiety, shame, and sorrow.

THE LUNGS' ENERGY FLOW

Functionally, the Lung Channels' energy acts on the Lungs, bronchi, throat, and larynx. If the Lung Qi is combined with Liver Qi and becomes stagnate in the throat area, a condition known as a "plum pit," or knot in the throat, develops. The Lung Channels store more Qi than Blood, thus they have a greater affect on energetic and nervous functions than on physical substances and Blood functions. At the high-tide time period (3 a.m. to 5 a.m.), Qi and Blood abound in the Lung organs and channels. They can therefore be more

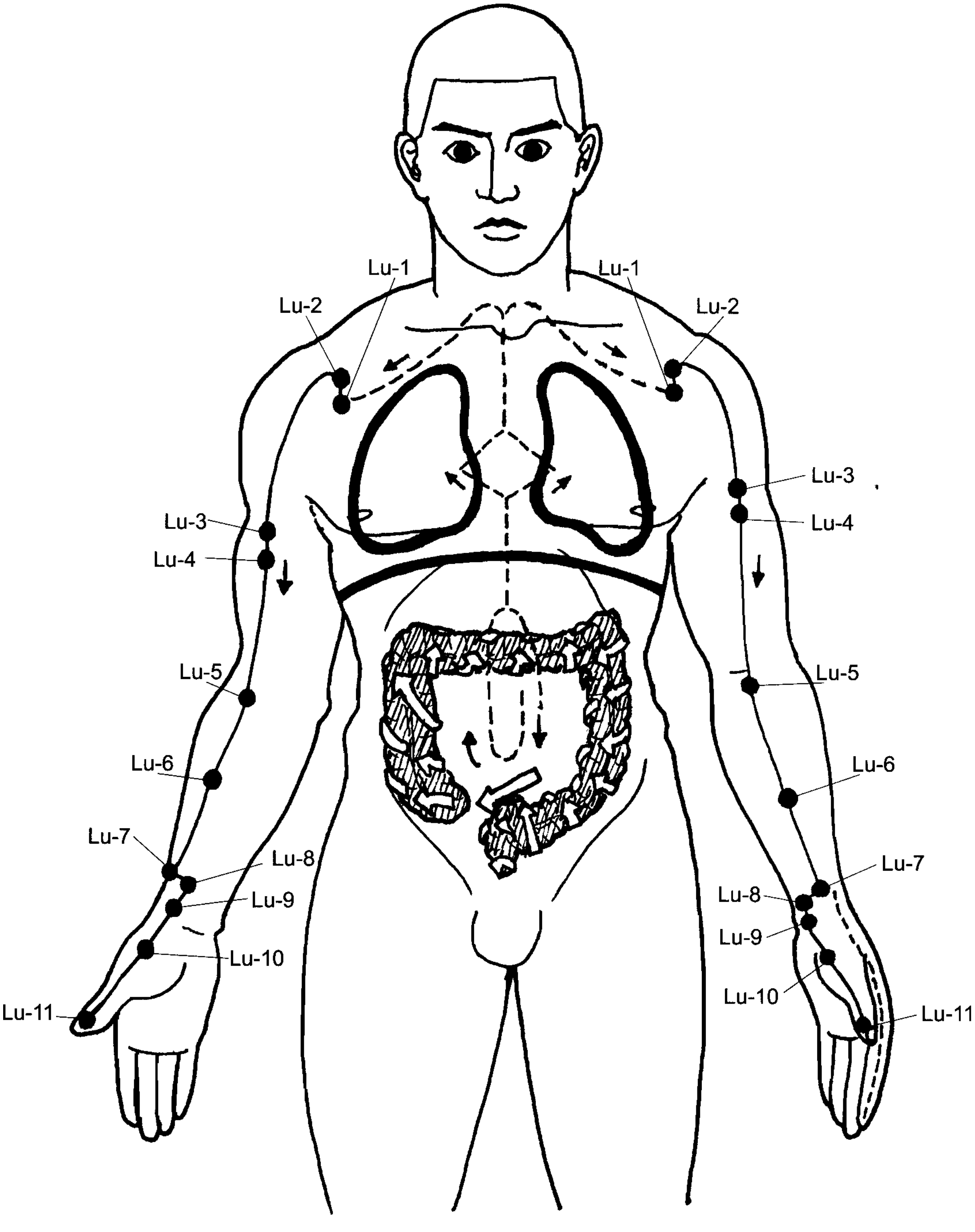
Lu-1
Lu-2
Lu-1
Lu-2
Lu-3
Lu-4
Lu-3
Lu-4
Lu-5
Lu-5
Lu-6
Lu-6
Lu-7
Lu-8
Lu-9
Lu-10
Lu-11
Lu-7
Lu-8
Lu-9
Lu-10
Lu-11

Figure 6.9. The Internal and External Qi Flow of the Lung (Lu) Channels

easily be dispersed and purged at this time period; whereas during low-tide (3 p.m. to 5 p.m.), they can more readily be tonified. The Lung Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The action of inhaling air introduces oxygen into the lungs; exhaling expels carbon dioxide from the lungs. The heart pumps blood into the lungs. Blood passing through the lungs receives oxygen, and is then delivered back into the heart, where it is circulated throughout the body via the circulatory system.

PATHOLOGICAL MANIFESTATIONS

The Lung Channels' dysfunctions can result in chest and Lung diseases, as well as diseases on the radial side of the upper arm and palmar area of the hand.

The Lungs keep the pathway of air unobstructed and disseminate vital Qi throughout the body. If these functions are impeded, obstructions of the nose, coughing, dyspnea, and fullness of the chest may occur.

The Lungs cleanse the inhaled air and keep the Qi flowing downward. If these functions are impeded, coughing, asthma, oliguria (diminished amount of urine formation), and edema may occur.

Since the Lungs have their external orifice at the nose, stuffy nose, nasal discharge, and impairment of smell are common symptoms when the Lungs are being attacked by a pathogenic invasion of Wind and Cold.

THE LARGE INTESTINE CHANNELS

The Large Intestine Channels are Yang channels and flow externally from the hands to the head. These two rivers originate externally from the tips of the index fingers, ascend up the arms, and cross the shoulders, where they connect with the 7th cervical vertebra and then split into two branches. One set of branches (one branch from the left Lung and its corresponding branch from the right Lung) descends internally and spirally wraps the Lungs before moving through to permeate the Large Intestine. The other set continues ascending externally through the neck and cheek to the gums of the lower teeth. It curves around the upper lip and flows to the opposite side of the nose, where it connects with the Stomach Channels (Figure 6.10).

THE LARGE INTESTINE ORGAN

The Large Intestine's associated organ is the Lungs; its element is Metal. The main function of the Large Intestine is to receive food essence from the Small Intestine. The Large Intestine organ is said to govern humor (any fluid or semifluid substance in the body). It absorbs fluid that was separated by the Small Intestine and controls the transformation and transportation of solid waste in the body. The Large Intestine also controls the throat, teeth, and the drainage of the nose.

PSYCHO-EMOTIONAL ASPECTS

The Large Intestine relationship to the Lungs makes it equally affected by the emotions of sadness, grief, and worry. An energetic imbalance in the Large Intestine can result in physical weakness and provoke emotional introversion, accompanied by feelings of depression, irritability, discouragement, distress and apathy.

Strong emotions of fear or panic can produce an energetic-stool reflex reaction in the Large Intestine resulting in a spontaneous defecation.

THE LARGE INTESTINE'S ENERGY FLOW

The Large Intestine Channels have equally abundant amounts of Qi and Blood, acting equally on the body's energetic and nervous functions, as well as on physical substances and Blood functions. At the high-tide time period (5 a.m. to 7 a.m.), Qi and Blood abound in the Large Intestine organ and channels and can therefore be more easily dispersed or purged at this period; whereas during low-tide (5 p.m. to 7 p.m.), they can be more readily tonified. The Large Intestine Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The large intestine, sometimes called the "colon," is divided into four quadrants; the ascending, transverse, descending, and sigmoid or pelvic colon. The function of the large intestine is to reabsorb the fluid from the contents of indigest-

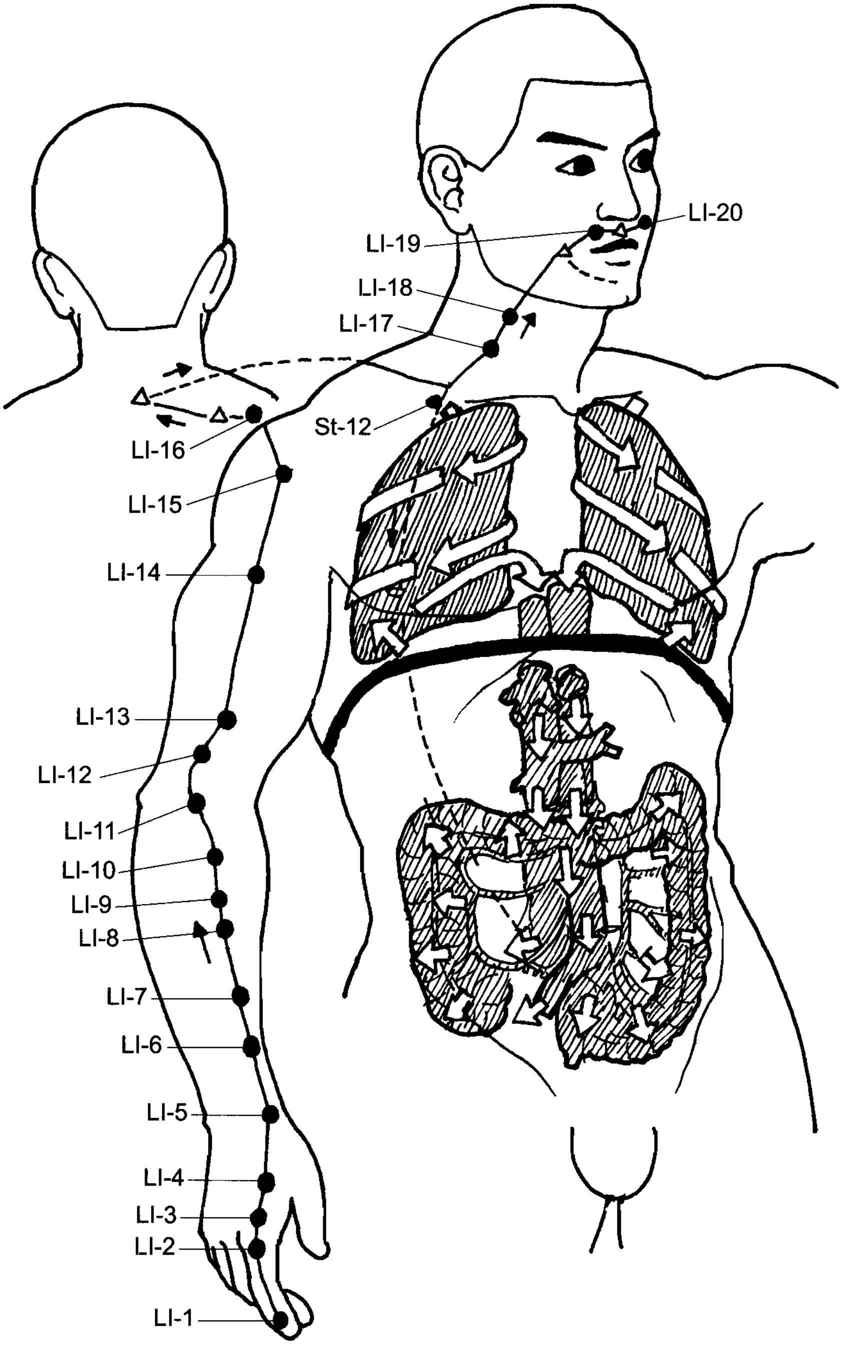
LI-20
LI-19
LI-18
LI-17
St-12
LI-16
LI-15
LI-14
LI-13
LI-12
LI-11
LI-10
LI-9
LI-8
LI-7
LI-6
LI-5
LI-4
LI-3
LI-2
LI-1

Figure 6.10. The Internal and External Qi Flow of the Large Intestine (LI) Channels

ible food residues moving through the intestines and to eliminate them from the body in the form of semisolid feces.

Whatever digestion takes place in the large intestine results from bacterial action. The fermentative bacteria (found in the middle part of the large intestine) and putrefying bacteria (found in the lower part of the large intestine) may produce toxic products, such as indole and skatol; however, if absorbed into the body, such bacterial decompositions undergo detoxification in the liver. In intestinal obstructions these bacterial decompositions are absorbed and eliminated in the urine in the form of indican.

PATHOLOGICAL MANIFESTATIONS

Dysfunctions in the Large Intestine Channels can result in diseases of the lower part of the face (including the nose, oral cavity, and teeth), throat, and front part of the neck, as well as disease of the back and radial sides of the upper extremities.

Tonification of the Large Intestine can be used to eradicate eye pain, toothache, earache, and to prevent hemorrhages, as well as to greatly reduce excessive menstrual bleeding.

Because of its relationship to the Lungs, the Large Intestine can be purged to treat coughing and asthma caused from Excessive Heat in the Lungs; whereas replenishing the vital Qi of the Lungs can cure constipation in debilitated patients.

The Large Intestines' energetic functional patterns generally relate to disturbances in bowel movements. Disturbances of the Large Intestine's energetic function usually gives rise to one of the two following complications:

1. Excess Conditions of the Large Intestines can result in symptoms of: Heat, Heat obstruction, Damp Heat, or Cold invading the Large Intestine.
2. Deficient Conditions of the Large Intestines can lead to an invasion of Cold or Dryness, and in severe cases of Deficiency, to the Collapse of the Large Intestine.

THE STOMACH CHANNELS

The Stomach Channels are Yang channels, and flow externally from the head to the feet (Figure 6.11). These two rivers originate externally from the lateral sides of the nose (LI-20), ascending upward to the base of the eye and bridge of the nose, where they meet the Bladder Channels. They then descend under the eyes, down the face, winding along the angle of the mandible, then ascend in front of the ears, following the anterior hairline till they reach the forehead. From the St-5 point, the primary branches on each side descend the neck and torso. Externally, these branches continue down the torso and legs and end at the lateral tips of the second toes. A second set of branches separate from the St-12 points and internally permeate the Stomach before spiral wrapping the Spleen and joining the primary channel branches at the St-30 points.

THE STOMACH ORGAN

The Stomach organ's associated organ is the Spleen; its element is Earth. The Stomach is considered the origin of the Body's Fluids. Its main function is to accept and decompose food. It separates the "clean" (pure) and "turbid" (impure) parts of ingested food and transfers the clean, energetically usable portion upwards to the Spleen, while sending the turbid unusable portion downward to the Small Intestine, to be further processed. Both the Stomach and the Spleen are considered the roots of Postnatal Qi and are often called The Ministers of Food Storage.

PSYCHO-EMOTIONAL ASPECTS

The Stomach influences the mental state; an Excess condition can agitate the mind and cause mental symptoms such as: mania or hypomania, confusion, severe anxiety, and hyperactivity.

THE STOMACH'S ENERGY FLOW

The Stomach Channels have equally abundant amounts of Qi and Blood, acting equally on the body's energetic and nervous functions, as well as on physical substances and Blood functions. At the high-tide time period (7 a.m. to 9 a.m.), Qi and Blood abound in the Stomach organ and channels, and can therefore be more easily dispersed or purged at this period; whereas during low-tide (7 p.m. to 9 p.m.), they can be more readily tonified. The Stomach Channels' energy acts on the skin, muscles, and nerves found along their pathways.

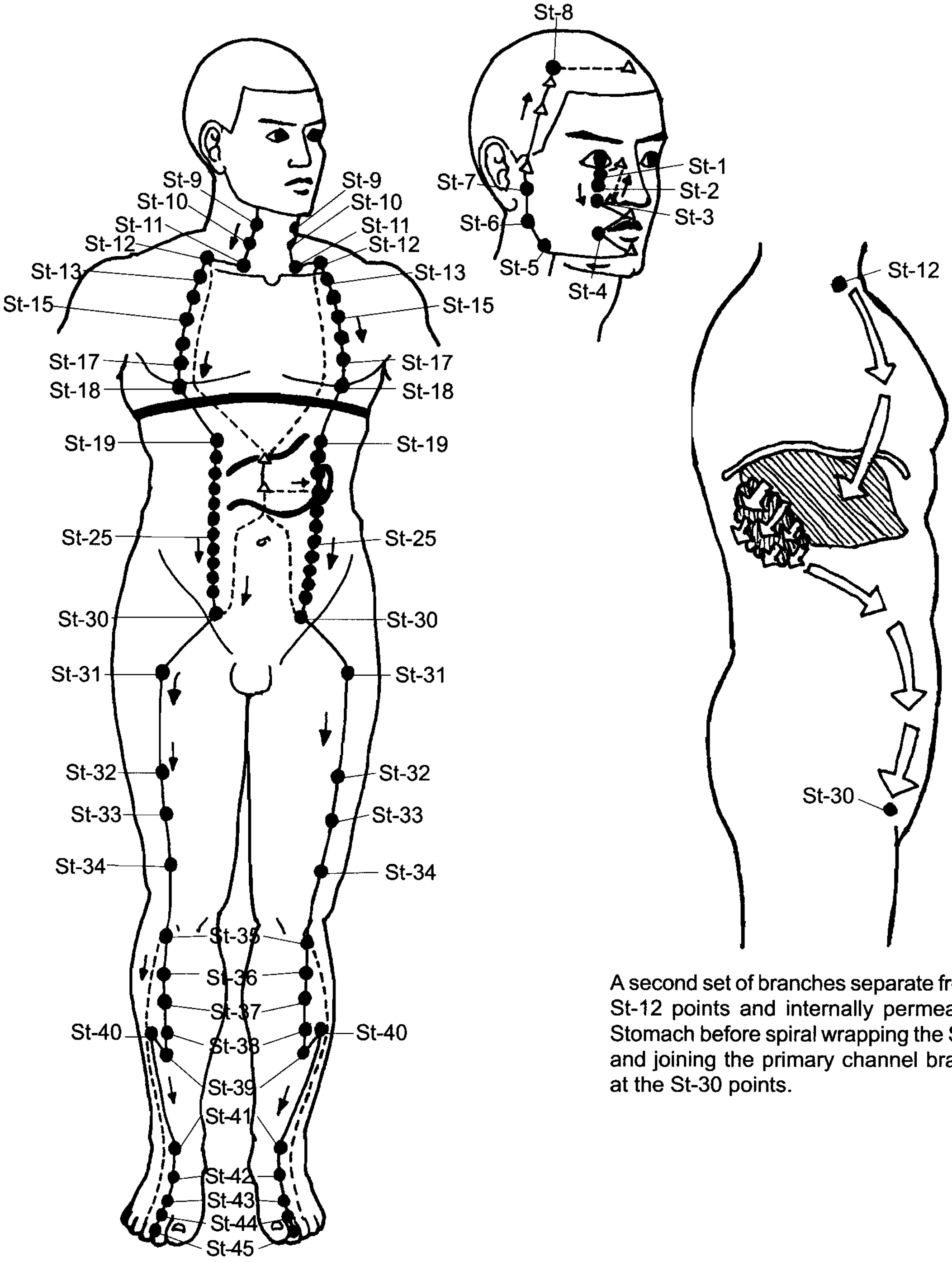
St-8
St-7
St-1
St-2
St-3
St-6
St-5
St-4
St-9
St-10
St-11
St-12
St-13
St-15
St-9
St-10
St-11
St-12
St-13
St-15
St-17
St-18
St-17
St-18
St-19
St-19
St-25
St-25
St-30
St-30
St-31
St-31
St-32
St-32
St-33
St-33
St-34
St-34
St-35
St-36
St-37
St-38
St-39
St-40
St-40
St-41
St-42
St-43
St-44
St-45
St-12
St-30

A second set of branches separate from the St-12 points and internally permeate the Stomach before spiral wrapping the Spleen and joining the primary channel branches at the St-30 points.

Figure 6.11. The Internal and External Qi Flow of the Stomach (St) Channels

WESTERN MEDICAL PERSPECTIVE

The stomach secretes gastric juice for food digestion. The proteins are converted into peptones; the stomach then regulates the passage of food to the other digestive organs.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Stomach Channels include diseases of the face (nose, oral cavity, and teeth), throat, front of the neck, the abdomen, the frontal part of the legs and gastrointestinal area. Stomach disease can also induce certain psychiatric diseases.

Diseases of the Stomach organ are primarily caused by improper diet. To determine the root of the disease, the following five areas are considered:

1. The patient's Five Elemental Constitution and present state of health,
2. The type and energetic quality of the foods ingested (Hot or Cold), and whether the food is in season or out of season,
3. The scheduling of meals (This not only includes eating meals at regular times, but also eating a balanced meal, not too fast, and not too late at night),
4. The Yin/Yang balance of the types of food digested (This includes balancing the foods Five Tastes and Five Colors), and
5. The emotional factors surrounding meal time. This includes having the proper mental attitude (being relaxed and calm) and avoiding emotional distress (such as feeling rushed, upset, and so on).

The Stomach easily suffers from Excess patterns (e.g., Fire or Phlegm Fire), which in turn agitate the Shen. When the Shen is disturbed, it can cause manic symptoms such as: inappropriate laughter, pressured speech, unconscious talking, laughing, or singing, violent or otherwise inappropriate behavior (e.g., taking off one's clothes in public). In milder cases, symptoms may include mental confusion, severe anxiety, hypomania (a milder form of mania), and hyperactivity.

Since the Stomach has the function of sending the preliminarily digested food downward, an impairment of this function often causes vomiting.

THE SPLEEN CHANNELS

The Spleen Channels are Yin channels and flow externally from the feet to the torso (Figure 6.12). These two rivers originate externally from the medial tips of the big toes, ascend upwards along the inner thighs, then enter the abdomen (connecting to the CV-3 and CV-4 points) and further up the upper torso (connecting to the CV-12 point). The external Spleen Channels meet at the CV-12 point, enter the body, where the left channel flows into the Spleen, and then spirally wraps the Stomach organ. An internal branch ascends from the CV-12 point into the Heart. At the CV-12 point both left and right main channels meet and separate again to ascend through the diaphragm, up the torso and throat, alongside the esophagus, to reconnect at the base of the tongue. From the Sp-20 points external branches descend along the lateral aspects of the torso and terminate on the mid-axillary line in the seventh intercostal space.

THE SPLEEN ORGAN

The Spleen's associated organ is the Stomach; its element is Earth. In Traditional Chinese Medicine, the pancreas (a gland that produces insulin and is connected to the Spleen) is considered part of the Spleen/Stomach function. The Spleen organ is responsible for transforming food received from the Stomach into Gu Qi. The Gu Qi is then sent to the Liver and Kidneys to be "cooked" and further refined. The Spleen's main function is to govern or oversee this transportation and transformation of Gu Qi. It also controls the Blood by keeping it in the blood vessels, and governs the muscles, flesh, and limbs. The Spleen organ's energy opens externally at the mouth, controlling taste, and connects externally by the lips. The Spleen is sometimes called The Minister of Grains, because it is responsible for distributing what the Stomach has stored. Energetically, the Spleen controls the body's central cavity, and maintains the internal organs in their places.

PSYCHO-EMOTIONAL ASPECTS

The Spleen houses the body's thoughts and intentions (Yi), and is responsible for analytical thinking, memory, cognition, intelligence, and ideas. The Spleen is responsible for directing

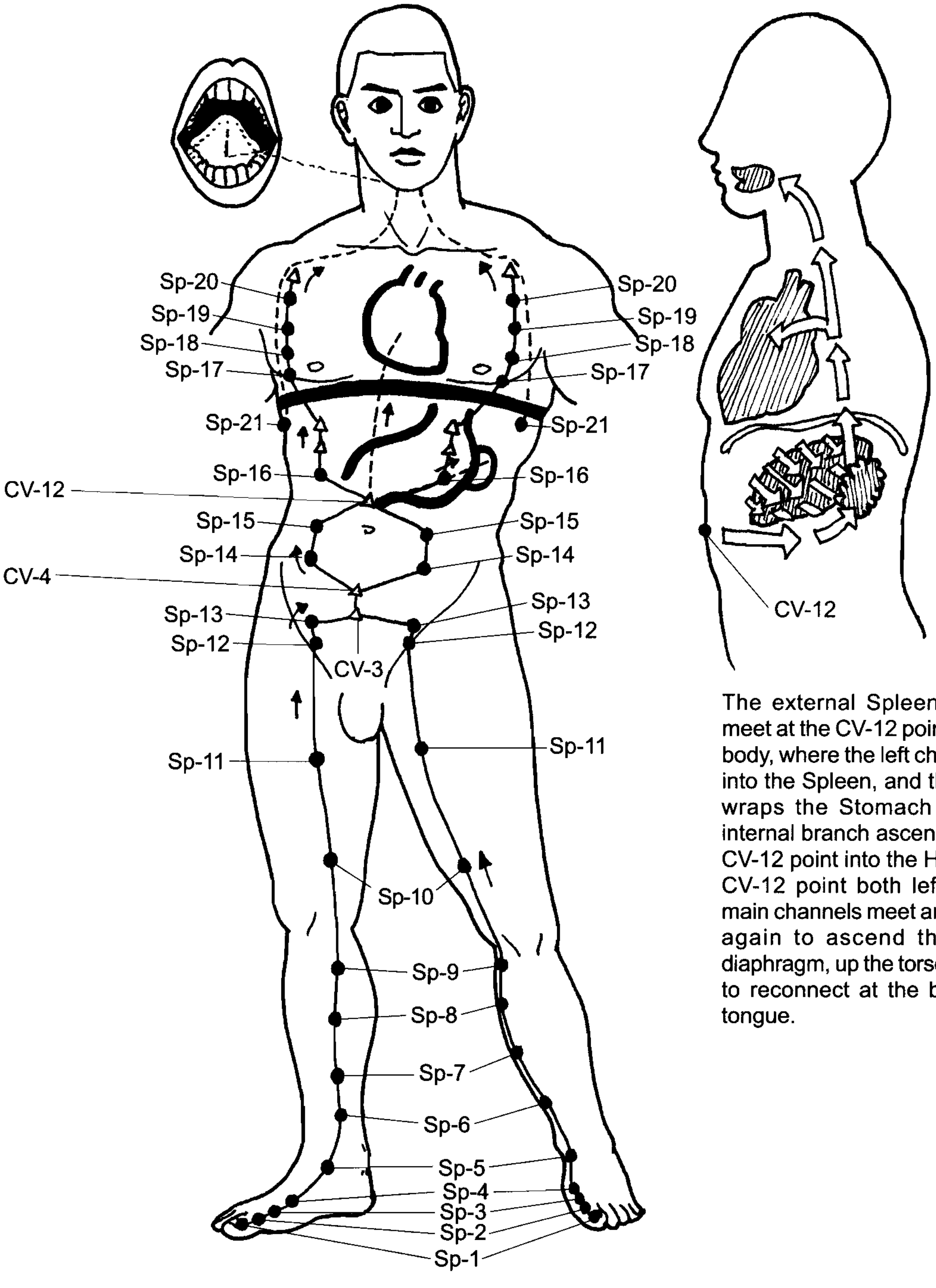
Sp-20
Sp-19
Sp-18
Sp-17
Sp-21
Sp-16
CV-12
Sp-15
Sp-14
CV-4
Sp-13
Sp-12
CV-3
Sp-11
Sp-10
Sp-9
Sp-8
Sp-7
Sp-6
Sp-5
Sp-4
Sp-3
Sp-2
Sp-1
Sp-20
Sp-19
Sp-18
Sp-17
Sp-21
Sp-16
Sp-15
Sp-14
Sp-13
Sp-12
Sp-11
CV-12

The external Spleen Channels meet at the CV-12 point, enter the body, where the left channel flows into the Spleen, and then spirally wraps the Stomach organ. An internal branch ascends from the CV-12 point into the Heart. At the CV-12 point both left and right main channels meet and separate again to ascend through the diaphragm, up the torso and throat to reconnect at the base of the tongue.

Figure 6.12. The Internal and External Qi Flow of the Spleen (Sp) Channels

memories to the Kidneys for short-term storage. The Kidneys will later transfer these memories to the Heart for long-term memory storage (see Chapter 12).

The Spleen's positive psycho-emotional attributes are trust, honesty, openness, acceptance, equanimity, balance, and impartiality; its negative attributes are worry, excessive thinking, pensiveness, obsessiveness, remorse, regret, obsessions, and self-doubt.

THE SPLEEN'S ENERGY FLOW

The Spleen Channels store more Qi and less Blood, acting more on energetic and nervous functions than on physical substance and Blood functions. At the high-tide time period (9 a.m. to 11 a.m.), Qi and Blood abound in the Spleen organ and channels, and can therefore be more easily dispersed and/or purged at this period; whereas during low-tide (9 p.m. to 11 p.m.), they can be more readily tonified. The Spleen Channels' energy acts on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The spleen is responsible for blood formation, filtration, and storage. In the embryo, all types of blood cells are formed by the spleen, but in the adult, only lymphocytes and monocytes are formed.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Spleen Channels include gastrointestinal dysfunctions (disturbances of digestion and absorption of food), and diseases of the tongue and throat. Spleen disorders also affect the inner side of the lower extremities along the Spleen's Channels.

Since the Spleen has the function of sending food essence upwards (to the Lungs), if this function is impeded, diarrhea or prolapse of the viscera may occur.

The Spleen also has the function of keeping the Blood flowing in the blood vessels; impairment usually leads to chronic hemorrhagic diseases.

The Spleen has the function of nourishing the flesh (muscles). A person with a healthy Spleen will usually have a healthy figure and a toned body; a diseased Spleen makes the patient lose muscle definition.

The Spleen nourishes the limbs. Strength of the limbs depends upon the nourishment guaranteed by the normal functioning of the Spleen. A diseased Spleen usually causes weakness of the limbs.

THE HEART CHANNELS

The Heart Channels are Yin Channels and flow externally from the torso to the hand (Figure 6.13). Three rivers originate internally from the Heart on each side of the body. The first bilateral set ascends and connects to the tissues connecting the eyeball. The second set descends, first to permeate the Pericardium, then descends further to connect with and spirally wrap the Small Intestine. The third set flows up into the Lungs, emerges externally at the armpits, then descends the medial aspect of the arms and ends on the inside of the little fingers.

THE HEART ORGAN

The main function of the Heart organ consists of controlling the Blood and vessels, regulating the Blood flow, perspiration, and housing the Shen. The Heart organ's energy opens externally at the tongue and is expressed through the complexion.

PSYCHO-EMOTIONAL ASPECTS

The Heart's associated organ is the Small Intestine; its element is Fire. Long-term memory, thinking, emotions, intimacy, cognition, intelligence, and ideas are all dominated by the function of the Heart. The Heart is sometimes called The Emperor, or "supreme controller of all Yin and Yang organs" (see Chapter 12). The Heart houses the body's Spirit (Shen). The Heart dominates sleep; if the Heart is strong the patient will fall asleep easily and sleep soundly. If the Heart is weak, the patient's mind will "float," resulting in an inability to fall asleep, disturbed sleep, or excessive dreaming.

The Heart's positive psycho-emotional attributes are love, joy, peace, contentment, propriety, insight, wisdom, orderliness, forgiveness, and courtesy. Its negative attributes are hate, guilt,

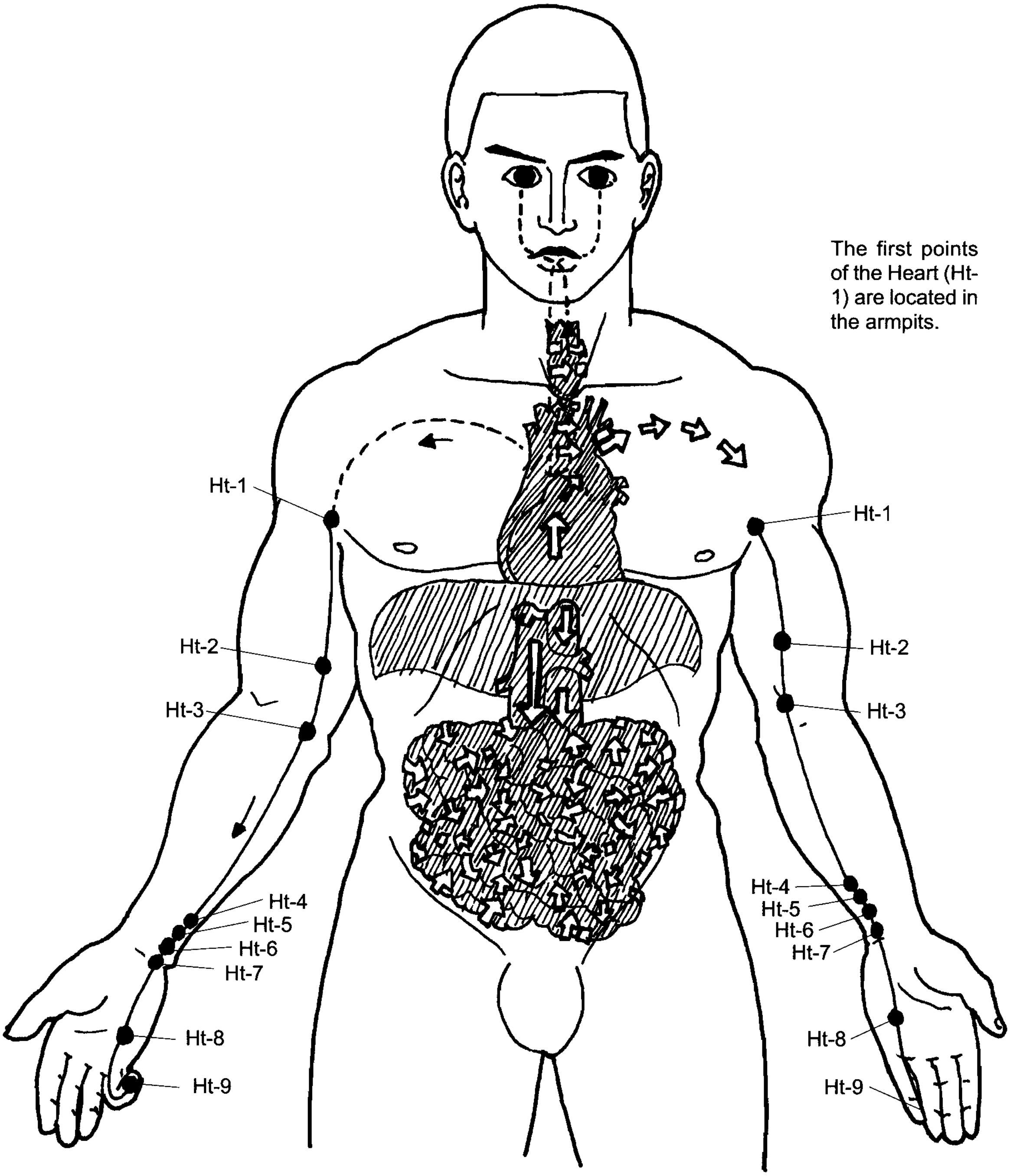
Ht-1
Ht-1
Ht-2
Ht-2
Ht-3
Ht-3
Ht-4
Ht-5
Ht-6
Ht-7
Ht-4
Ht-5
Ht-6
Ht-7
Ht-8
Ht-8
Ht-9
Ht-9

The first points of the Heart (Ht-1) are located in the armpits.

Figure 6.13. The Internal and External Qi Flow of the Heart (Ht) Channels

shock, nervousness, excitement, longing, and craving.

THE HEART'S ENERGY FLOW

The Heart Channels store more Qi than Blood; that is to say, its dominant action affects the energy. This energetic action controls the morale, the spirit of enterprise, and provides the energy required for respiration. At the high-tide time period (11 a.m. to 1 p.m.), Qi and Blood abound in the Heart organ and channels, and can therefore be more easily dispersed or purged at this period; whereas during low-tide (11 p.m. to 1 a.m.), they can be more readily tonified. The Heart Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The heart provides the propulsive force for circulating the blood throughout the vascular system. The blood circulation distributes the nutrients and oxygen needed for cell and tissue survival. The heart's propulsion also aids in the removal of wastes and other substances through the interconnecting blood vessels.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Heart Channels include diseases that exert pressure on the brain, eyes, or pharyngeal wall, as well as diseases of the Heart itself, and the lateral side of the chest. The Heart also relates to diseases of the ulnar palmar side of the arm.

The Heart is in charge of mental activities, including consciousness and thinking. Dysfunctions of the Heart can lead to insomnia, impairment of consciousness, amnesia, and psychosis.

The Heart and Kidneys have a mutual energetic relationship of supporting and checking each other. The Heart controls the body's Fire, while the Kidneys control the body's Water. Normally, the Fire of the Heart is sent down to warm the Kidneys, and the Water of the Kidneys is sent up to irrigate the Heart. If this balanced relationship breaks down (especially when the Kidney Water is insufficient to check the Heart Fire), a series of Fire symptoms of the Heart, such as hypertension, hyperactivity, palpitations, and insomnia may result.

Since the Heart has its external opening in the tongue, the condition of the Heart is reflected in the tongue:

1. A dark purple tongue indicates Blood Stasis of the Heart.
2. A pale tongue reveals Deficient Blood of the Heart.
3. An ulcer on the tongue reveals Excess Fire of the Heart.

THE SMALL INTESTINE CHANNELS

The Small Intestine Channels are Yang channels and flow externally from the hands to the head. These two rivers originate externally from the lateral tips of the little fingers (ulnar side) then ascend the arms to the shoulders, where they divide into two sets of branches. One set of branches descends the supraclavicular fossa internally, spiral wrapping the Heart, then continues down the esophagus, diaphragm, and Stomach, before permeating the Small Intestine (Figure 6.14).

The other set of branches ascends externally up the sides of the neck, then divides again into two more sets of branches at the cheek; one set will end at the ears, the other at the lateral sides of the nose and inner canthus of the eyes.

THE SMALL INTESTINE ORGAN

The Small Intestine temporarily stores partially digested food, absorbing the essential substances of the undigested food and a portion of the liquid content. It later transfers the residue (with a considerable amount of liquid) to the Large Intestine, and that is why it is said that the Small Intestine "governs liquid."

The Small Intestine is sometimes called "The Official in Charge of Separating the Pure from the Impure, because it separates the "clean" (reusable) food essence (Gu Qi) from the "dirty, turbid" food essence. The clean Gu Qi is transported throughout the body to nourish the tissues. The turbid is transported to the Large Intestine and Bladder organs to be further processed.

The Small Intestine's associated organ is the Heart; its element is Fire.

PSYCHO-EMOTIONAL ASPECTS

The Small Intestine influences the patient's

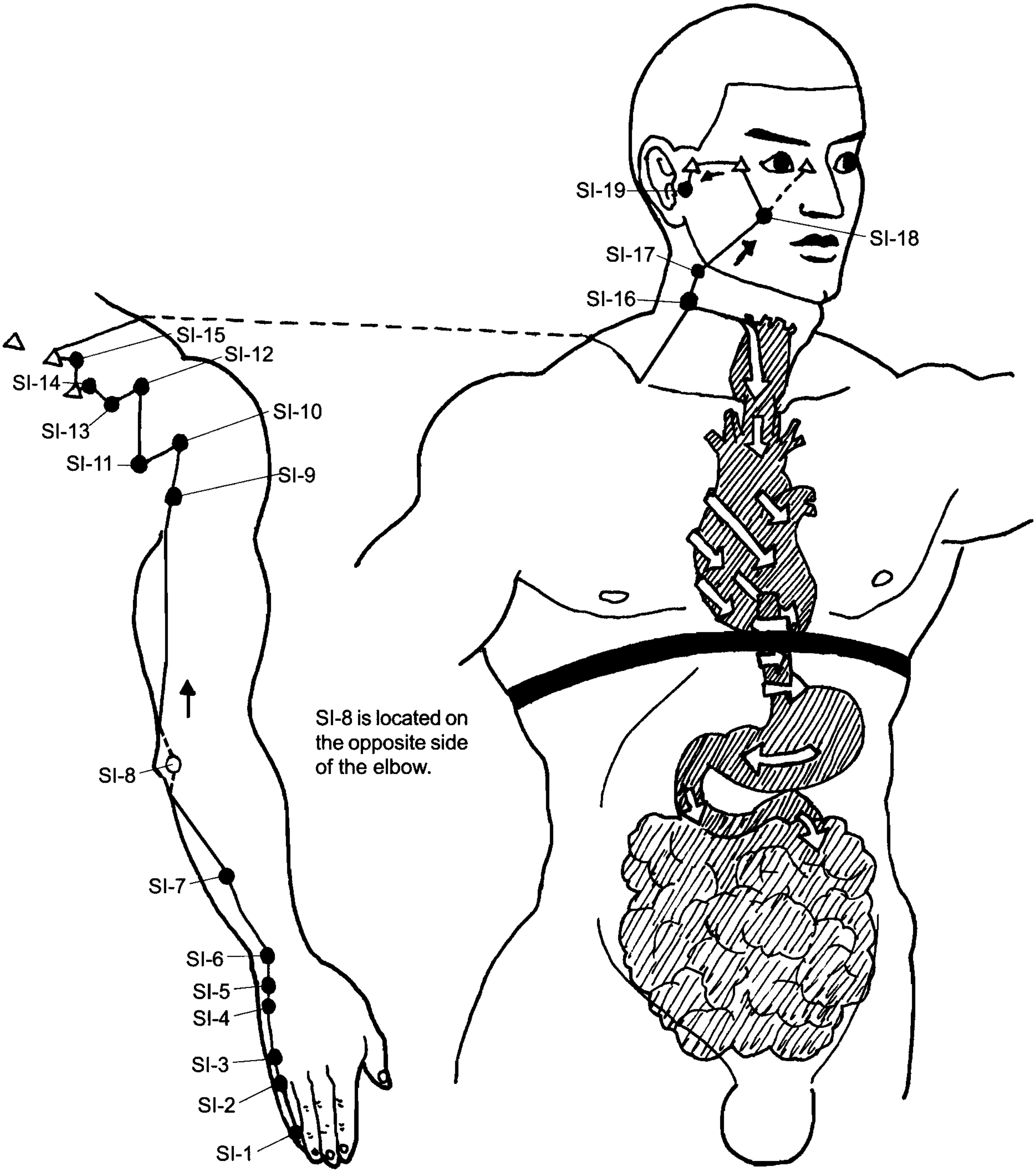
SI-19
SI-18
SI-17
SI-16
SI-15
SI-12
SI-14
SI-13
SI-10
SI-11
SI-9
SI-8 is located on the opposite side of the elbow.
SI-8
SI-7
SI-6
SI-5
SI-4
SI-3
SI-2
SI-1

Figure 6.14. The Internal and External Qi Flow of the Small Intestine (SI) Channels

mental clarity, judgement, and powers of discernment. The ability to distinguish relevant issues with clarity before making a decision is attributed to the Small Intestine.

THE SMALL INTESTINE'S ENERGY FLOW

The Small Intestine Channels store more Blood than Qi, and act more on the physical substances and Blood functions than on energy. At the high-tide time period (1 p.m. to 3 p.m.), Qi and Blood abound in the Small Intestine organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (1 a.m. to 3 a.m.), they can be more readily tonified. The Small Intestine Channels' energy act on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The small intestine's function is to receive the food mass from the stomach through the pylorus, the bile from the liver and gall bladder, and the pancreatic juices for digestion and absorption. The small intestine then passes what it doesn't absorb into the large intestine.

PATHOLOGICAL MANIFESTATIONS

Diseases of the Small Intestine Channels include diseases of the face, ear, cheek, lower jaw, the neck, and throat, as well as the dorsal ulnar side of the upper extremities.

In TCM, food and herbs are divided into Yin (Cold) or Yang (Hot) properties. The Small Intestine is easily affected by the type and temperature of food eaten. An excess consumption of "cold" and raw foods, for example, can create a Cold condition within the Small Intestines, while an excess consumption of "hot" foods can create a Hot condition.

Disturbances of the Small Intestine's energetic function usually gives rise to one of the following complications:

1. Excess Conditions of the Small Intestines can result in symptoms of: Full Heat, Qi pain, Knotted Qi, and the infestation of worms in the Small Intestine.
2. Deficient Conditions of the Small Intestines can manifest as: Deficient and Cold Small Intestine.

Pathologic Heart Heat may be transmitted to the Small Intestine, resulting in urodynia (painful urination) and hematuria (blood in the urine). Herbal prescriptions used to dispel Heat from the Heart are also effective in treating these urinary symptoms, which are alleged to be caused by dysfunctions of the Small Intestine.

THE BLADDER CHANNELS

The Bladder Channels are Yang channels and flow externally from the head to the feet. Two rivers originate externally from the inner canthus of the eyes, ascend upward over the head to join the Governing Vessel at the GV-20 point, where they divide into two additional branches that flow into each temple, and into the brain. From there, they then divide into two sets of branches (at the Bl-10 point), which descend the back, and internally connect to and spiral wrap the Kidneys, before they descend to permeate the Bladder. They continue externally down the thigh, the popliteal fossa at the back of the knee, calf, and foot, and end on the lateral sides of the tips of the little toes (Figure 6.15).

THE BLADDER ORGAN

The function of the Bladder is to remove water by Qi transformation. The Bladder receives the "impure" parts of the fluids separated by the Small Intestine and temporarily stores and transforms these fluids into urine, to discharge it when the Bladder is full. The Bladder and the Small Intestine work together to remove Body Fluids from the Lower Burner.

The Bladder Channels activate the Yang aspect of the Kidney's function and are sometimes called The Water District Officials or "the controllers of the storage of waste water." The Bladder's associated organ is the Kidney; its element is Water.

PSYCHO-EMOTIONAL ASPECTS

An imbalance in the Bladder can cause such psychological symptoms as habitual fear, lack of decision making capability and a diminished moral character. If the imbalance becomes chronic, it results in such emotional responses as jealousy, suspicion, and a holding on to long-standing grudges.

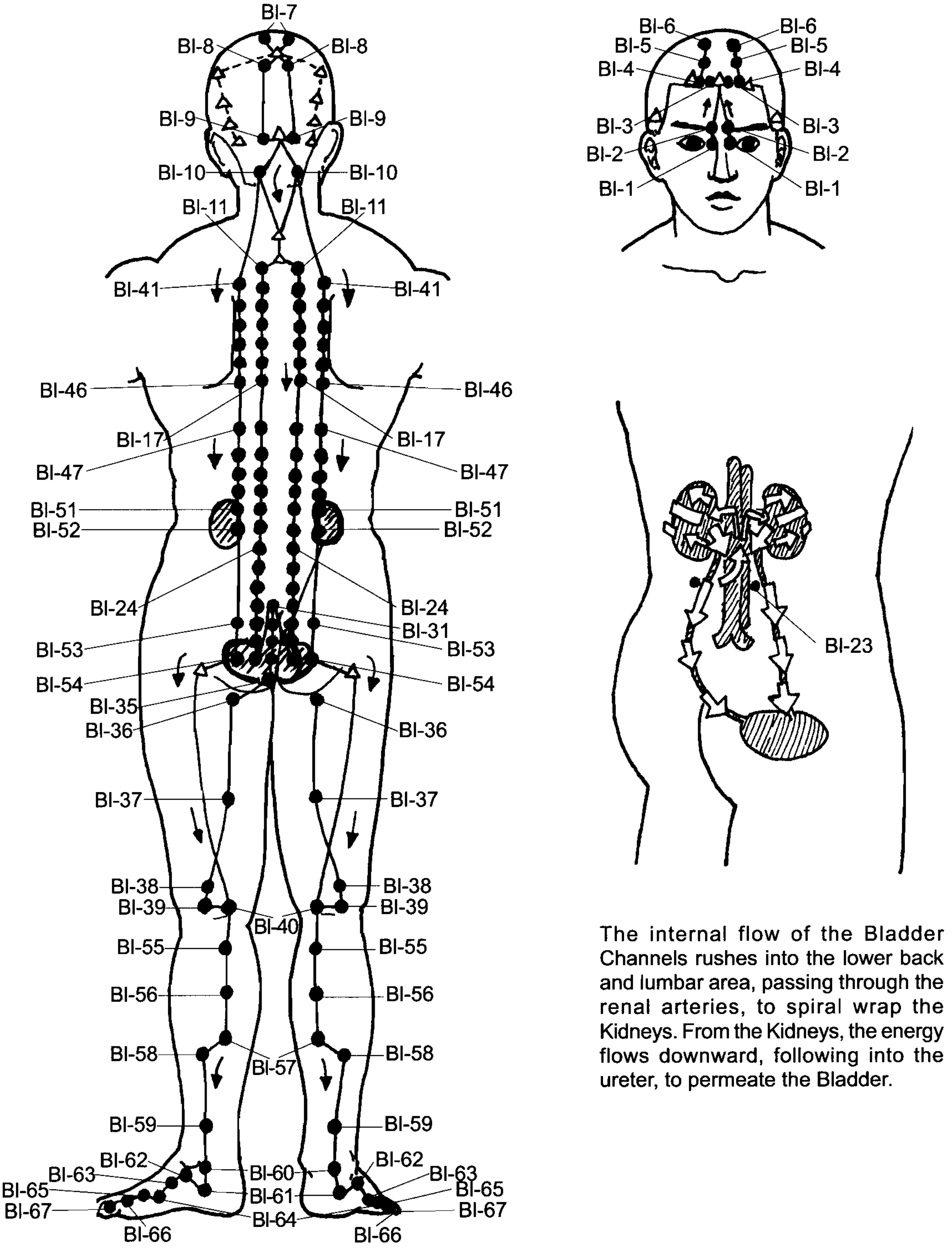
Bl-7
Bl-8
Bl-8
Bl-9
Bl-9
Bl-10
Bl-10
Bl-11
Bl-11
Bl-41
Bl-41
Bl-46
Bl-46
Bl-17
Bl-17
Bl-47
Bl-47
Bl-51
Bl-52
Bl-51
Bl-52
Bl-24
Bl-24
Bl-31
Bl-53
Bl-53
Bl-54
Bl-54
Bl-35
Bl-36
Bl-36
Bl-37
Bl-37
Bl-38
Bl-39
Bl-38
Bl-39
Bl-40
Bl-55
Bl-55
Bl-56
Bl-56
Bl-58
Bl-57
Bl-58
Bl-59
Bl-59
Bl-62
Bl-63
Bl-65
Bl-67
Bl-66
Bl-60
Bl-61
Bl-64
Bl-62
Bl-63
Bl-65
Bl-67
Bl-66
Bl-6
Bl-6
Bl-5
Bl-5
Bl-4
Bl-4
Bl-3
Bl-3
Bl-2
Bl-2
Bl-1
Bl-1
Bl-23

The internal flow of the Bladder Channels rushes into the lower back and lumbar area, passing through the renal arteries, to spiral wrap the Kidneys. From the Kidneys, the energy flows downward, following into the ureter, to permeate the Bladder.

Figure 6.15. The Internal and External Qi Flow of the Bladder (Bl) Channels

THE BLADDER'S ENERGY FLOW

The Bladder Channels regulate the functions of the Kidneys. The Bladder Channels store more Blood than Qi, acting more on the physical substances and the Blood functions than on the energy. At the high-tide time period (3 p.m. to 5 p.m.), Qi and Blood abound in the Bladder organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (3 a.m. to 5 a.m.), they can be more readily tonified. The Bladder Channels' energy acts on the skin, muscles, and nerves found along their pathway.

WESTERN MEDICAL PERSPECTIVE

The bladder is a reservoir for the urine received from the kidneys; it discharges urine from the body through the urethra.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Bladder Channels include:

- Diseases of the top of the head,
- Brain disorders,
- Disorders of the neck and back, especially the lumbar and sacral regions,
- Disorders of the back of the legs and thighs, and
- Disorders of the lateral side of the feet.

Diseases of the Bladder Organ manifest in changes in urine and urination; these changes will reflect a Deficient or Excess condition of the Bladder.

1. Deficient Conditions are attributed to Deficient Kidney Qi which affects the Bladder's ability to transform Qi. This dysfunction causes frequent urination, dribbling, or enuresis (involuntary discharge of urine).
2. Excess Conditions are attributed to Damp Heat in the Bladder, resulting in symptoms such as:

- heat and pain during urination,
- the short release of murky or reddish urine,
- frequent difficulty in urination,
- pus, or Blood in the urine, and
- Bladder stones which cause a urinary block and painful distention of the lower abdomen.

The most common pathological factor in the diseases of the Bladder organ is due to the accumulation of Dampness (e.g., Damp Heat or Damp Cold). Other pathological conditions include: emotional fears, anxieties and insecurities, excessive sex, and Deficient and Cold Bladder.

THE KIDNEY CHANNELS

The Kidney Channels are Yin channels and flow externally from the feet to the torso (Figure 6.16). These two rivers originate externally from underneath the little toes, circling the inside of the heels, ascending through the medial aspect of the legs, where they merge and enter into the coccyx and lower lumbar vertebrae. Dividing again, one branch permeates the Kidneys, while the other branch continues ascending into the cerebral cortex. From the Kidney organs, two additional sets of channels internally emerge. One set descends and spiral wraps the Bladder. The other set ascends into the Liver, diaphragm, and Lungs, then spiral wraps the Heart and travels up the throat, stopping at the root of the tongue (Figure 6.17).

THE KIDNEY ORGANS

The Kidneys store Jing substances, produce Marrow, and control the development of the bones. They also control respiration, reproduction, growth and development, and govern the Body Fluids. The Kidney Jing is the biological basis for the woman's menstrual blood. The Kidneys are the root of the Yin and Yang Qi of all the body's organs. The Kidneys energetically open externally through the ears and are sometimes called The Minister of Ingenuity and Vitality, or "the controller of water" (see Chapter 12).

The Kidney organs' associated organ is the Bladder and it's element is Water. Kidney Yin Qi flows to the Liver, Heart, and Lungs. It is responsible for the body's fluid-like essences and rules the birth, growth, maturation, and reproduction cycle. Kidney Yang Qi flows to the Spleen, Liver, Heart, and Lungs. It supports the Yang of all the body's organs via the Mingmen.

PSYCHO-EMOTIONAL ASPECTS

The Marrow produced from the Kidney Jing flows into the brain. The thinking ability is strengthened when Qi and Blood in the cerebral

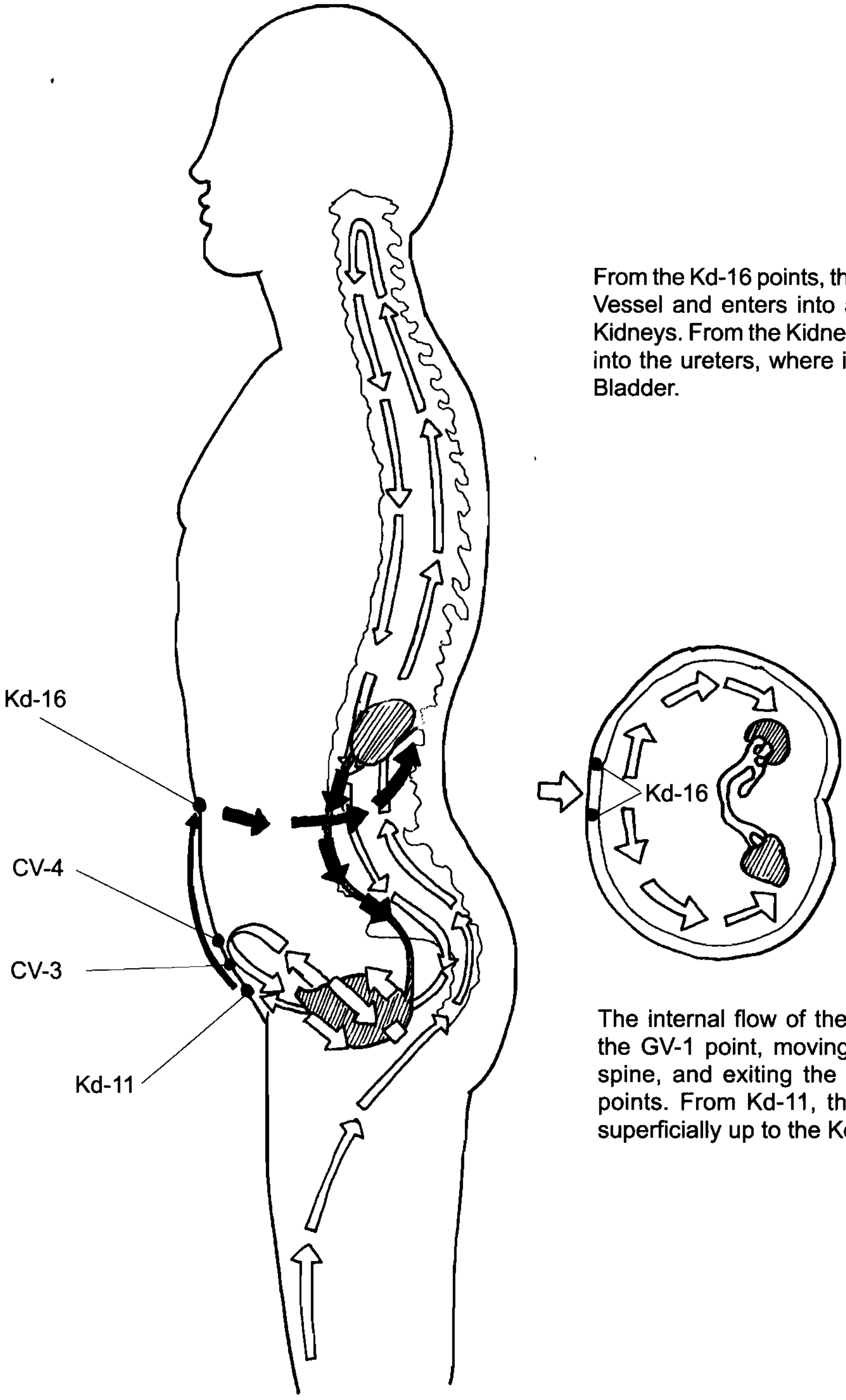

From the Kd-16 points, the Qi follows the Belt Vessel and enters into and permeates the Kidneys. From the Kidneys, the Qi descends into the ureters, where it spirally wraps the Bladder.

The internal flow of the Kidney Qi passes the GV-1 point, moving up and down the spine, and exiting the body at the Kd-11 points. From Kd-11, the energy ascends superficially up to the Kd-16 points.

Figure 6.16. The Internal Flow of the Kidney Qi

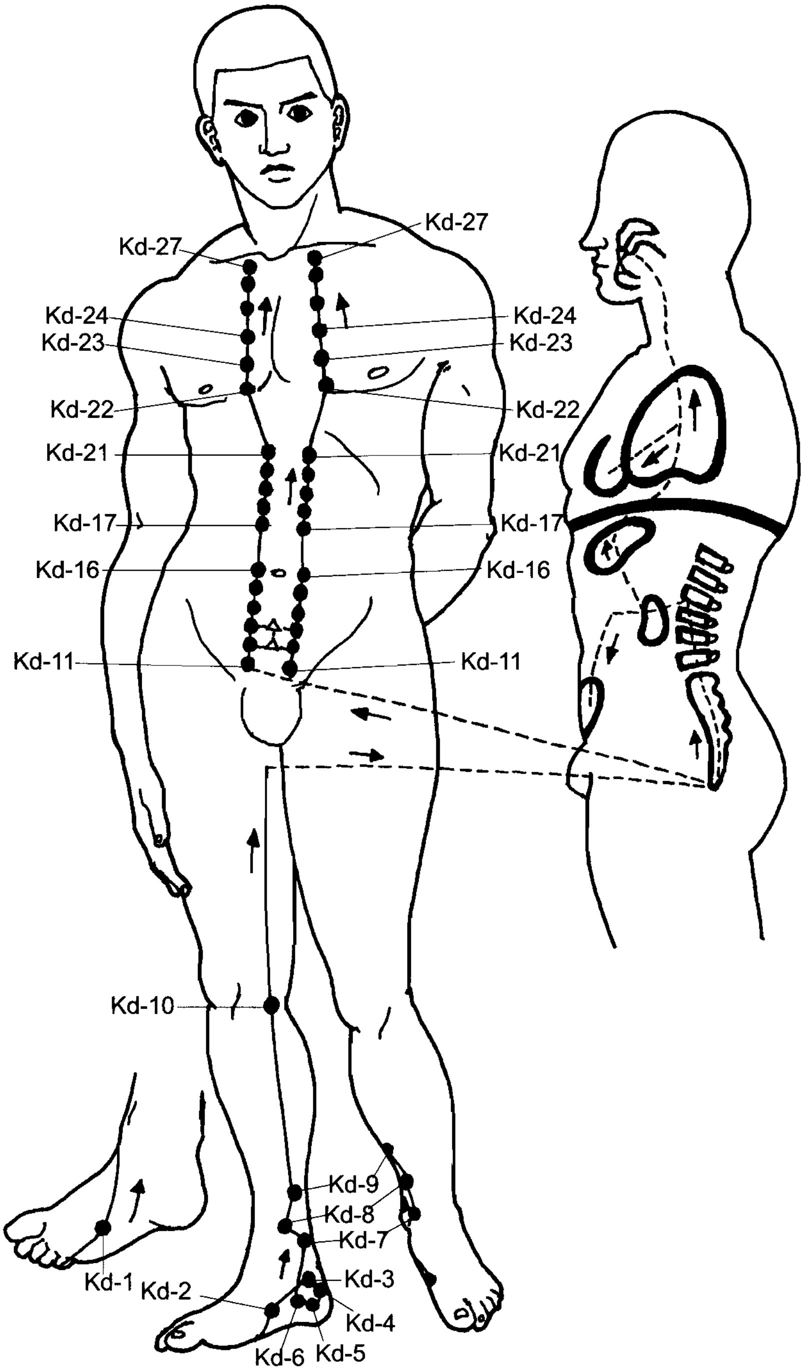
Kd-27
Kd-27
Kd-24
Kd-24
Kd-23
Kd-23
Kd-22
Kd-22
Kd-21
Kd-21
Kd-17
Kd-17
Kd-16
Kd-16
Kd-11
Kd-11
Kd-10
Kd-9
Kd-8
Kd-7
Kd-3
Kd-1
Kd-2
Kd-4
Kd-6
Kd-5

Figure 6.17. The Internal and External Qi Flow of the Kidney (Kd) Channels

cortex are abundant. With the increase of thinking activity, a circle of light develops within the body's Taiji Pole. The individual's innate intelligence (Yuan Shen) is represented by the intensity of light. The degree of accumulated spiritual energy is reflected by the number of light circles developed within the Taiji Pole. These circles of light can be best observed when first waking. By placing slight pressure on the external eye lids the inner light of the Taiji Pole is projected onto the optic nerves, reflecting an image of the circle of light. If the circle of light is complete, it reflects a strong, healthy condition. If the circle of light is dark within its center (similar to a doughnut), this reflects a Deficient condition. If the circle of light is broken or interrupted, it reflects an extreme Deficiency (Figure 6.18).

The "memory zone," as well as the "thought center" are also located in the cerebral cortex and will not develop until the Kidney Channels travel through the spine, along with the Liver Channels, to reach the cortex. When the Qi of these two channels is abundant, the memory function is keen.

The Kidneys house the body's will power (Zhi). They control short-term memory and store data. The Kidneys provide the capacity and drive for strength, skill, and hard work. A patient with strong Kidneys can work hard and purposefully for long periods of time. Consequently, when the Kidneys are in a state of disharmony, the patient can sometimes be driven to a state of excessive-compulsive working habits (a workaholic). A patient with weak Kidneys will lack strength and endurance.

The Kidney's positive psycho-emotional attributes are wisdom, rationality, clear perception, gentleness, and self-understanding. The negative attributes are fear, loneliness, insecurity, and shock (which attacks the Heart first then descends into the Kidneys to become fear).

THE KIDNEYS' ENERGETIC FLOW

The Kidney Channels store more Qi than Blood, acting more on energetic and nervous functions than on physical substances and Blood functions. At the high-tide time period (5 p.m. to 7 p.m.), Qi and Blood abound in the Kidney organ and channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (5 a.m. to 7 am.), they can be more readily tonified. The Kidney Channels' energy acts on the skin, muscles, and nerves found along their pathways.

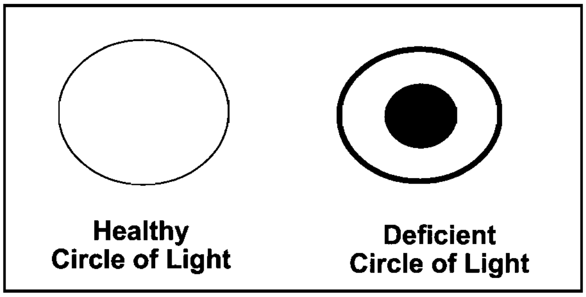

Figure 6.18. The individual's Yuan Shen is represented by the intensity of the light, and the degree of accumulated spiritual energy is reflected by the number of light circles developed within the Taiji Pole.

WESTERN MEDICAL PERSPECTIVE

The function of the kidneys is to regulate water and electrolytes and to process the body's urine.

PATHOLOGICAL MANIFESTATIONS

Kidney Channel diseases cause general deterioration of the entire body, weakness in the lower extremities, lumbar pain, and hot sensations deep inside the feet.

The Kidneys open through the urogenital orifices and the anus. The energetic condition of the Kidneys can be, partially reflected by the condition of the patient's urination and defecation; in males, this includes the ejaculation process.

Since the Kidneys are responsible for concentration and memory retention, poor concentration and loss of memory are common symptoms of Kidney hypofunction.

THE PERICARDIUM CHANNELS

The Pericardium Channels are Yin channels and flow externally from the torso to the hands. These two rivers originate internally from the center of the chest, flowing out of the Pericardium. They descend through the center of the body, spi-

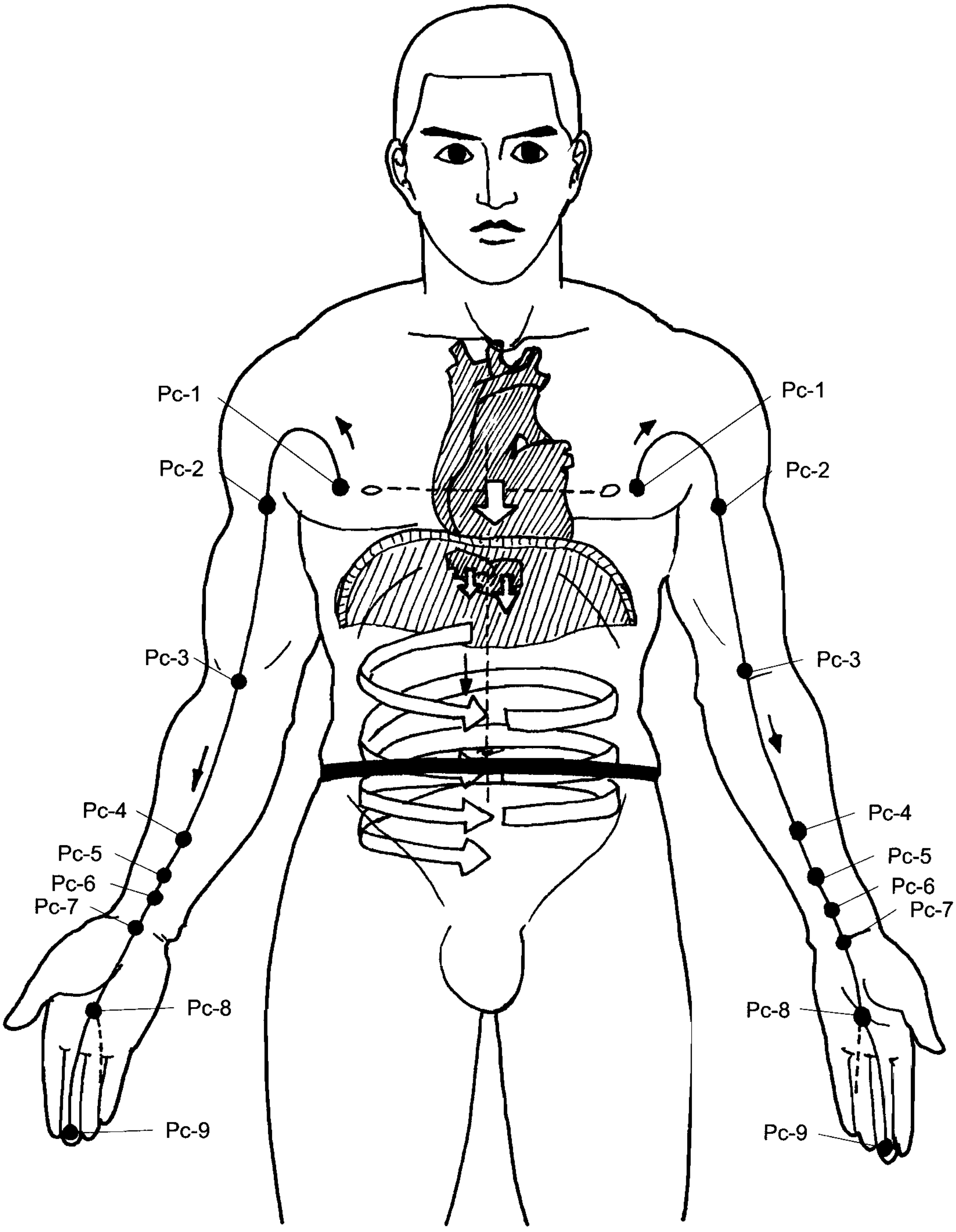
Pc-1
Pc-1
Pc-2
Pc-2
Pc-3
Pc-3
Pc-4
Pc-4
Pc-5
Pc-5
Pc-6
Pc-6
Pc-7
Pc-7
Pc-8
Pc-8
Pc-9
Pc-9

Figure 6.19. The Internal and External Qi Flow of the Pericardium (Pc) Channels

ral wrapping the Upper, Middle, and Lower Burners. Two rivers surface and branch externally by the nipples to proceed down the center of each arm, ending at the tips of the middle fingers. A second branch arises from the center of each palm at the Pc-8 point and flows into the ring fingers to connect with the Triple Burner Channels (Figure 6.19).

PERICARDIUM ORGAN

The Pericardium is associated with the Triple Burners; its element is Fire. The Pericardium is not considered an independent organ, as much as a protective covering for the Heart, that protects the Heart from pathogenic factors. The Pericardium activates, energizes and controls the Yin channel's distribution of the Kidney Yang Qi to the Yin organs. Along with the Heart, it governs the Blood, and houses the Shen. The Pericardium is considered the Prime Minister, or the official who protects the heart.

PSYCHO-EMOTIONAL ASPECTS

The Pericardium has a powerful influence on the patient's mental and emotional states. Its goal is to "create feelings of joy and/or pleasure for the emperor (Heart)."

THE PERICARDIUM'S ENERGY FLOW

The Pericardium Channels have an affect on the body's circulation of Blood. They are considered the "mother of Yin" and are connected to the Mingmen. The Pericardium Channels store more Blood than Qi, acting more on the physical substance and Blood functions than on the energy. At the high-tide time period (7 p.m. to 9 p.m.), Qi and Blood abound in the Pericardium organ and channels and can therefore be more easily dispersed and purged at this period; whereas during low-tide (7 a.m. to 9 a.m.), they can be more readily tonified. The Pericardium Channel's energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

The function of the pericardium is to provide a protective covering for the heart. It contains a thin serous fluid that allows the heart to work in a relatively friction-free environment.

PATHOLOGICAL MANIFESTATIONS

The main diseases of the Pericardium Channels include diseases of the Heart, front of the chest, major blood vessels, and diseases on the midline to upper palmar side of the upper extremities. Mental abnormalities may also manifest.

THE TRIPLE BURNER CHANNELS

The Triple Burner Channels are Yang channels that flow externally from the hands to the head. These two rivers originate externally from the tips of the ring fingers and ascend along the lateral aspect of the arms, over the shoulders to the clavicles, then branch internally and spiral wrap the Pericardium. They circle the diaphragm and permeate the Upper, Middle, and Lower Burners. A set of branches originate internally from the chest CV-17 point at the Heart, or center of the Middle Dantian. Flowing upwards, these branches emerge from the supraclavicular fossa at the St-12 points to ascend up the neck, head, and then circle the ears. Another set of branches flows downward to the cheek, terminating in the infraorbital region; the other set ends above the ears, by the outer canthus of the eyes.

THE TRIPLE BURNERS ORGAN

The Triple Burners' associated organ is the Pericardium; its element is Fire. The Triple Burners are not regarded as an independent organ, but are assigned to specific energy areas, and are a part of the function of the Yin and Yang organs. The function of the Triple Burners' energy is to produce Heat, and to regulate the body's temperature. This Heat is increased through such meditative disciplines as the Microcosmic Orbit (which connects the Governing and Conception Vessels), and the Macrocosmic Orbit (which connects all Twelve Primary Channels with the Governing and Conception Vessels). The main function of the Triple Burners (Figure 6.20) is to regulate the general ingestion and digestion of food and fluids throughout the body.

The Triple Burners are known as The Official of Balance and Harmony, because they govern Water metabolism, control the production of Wei

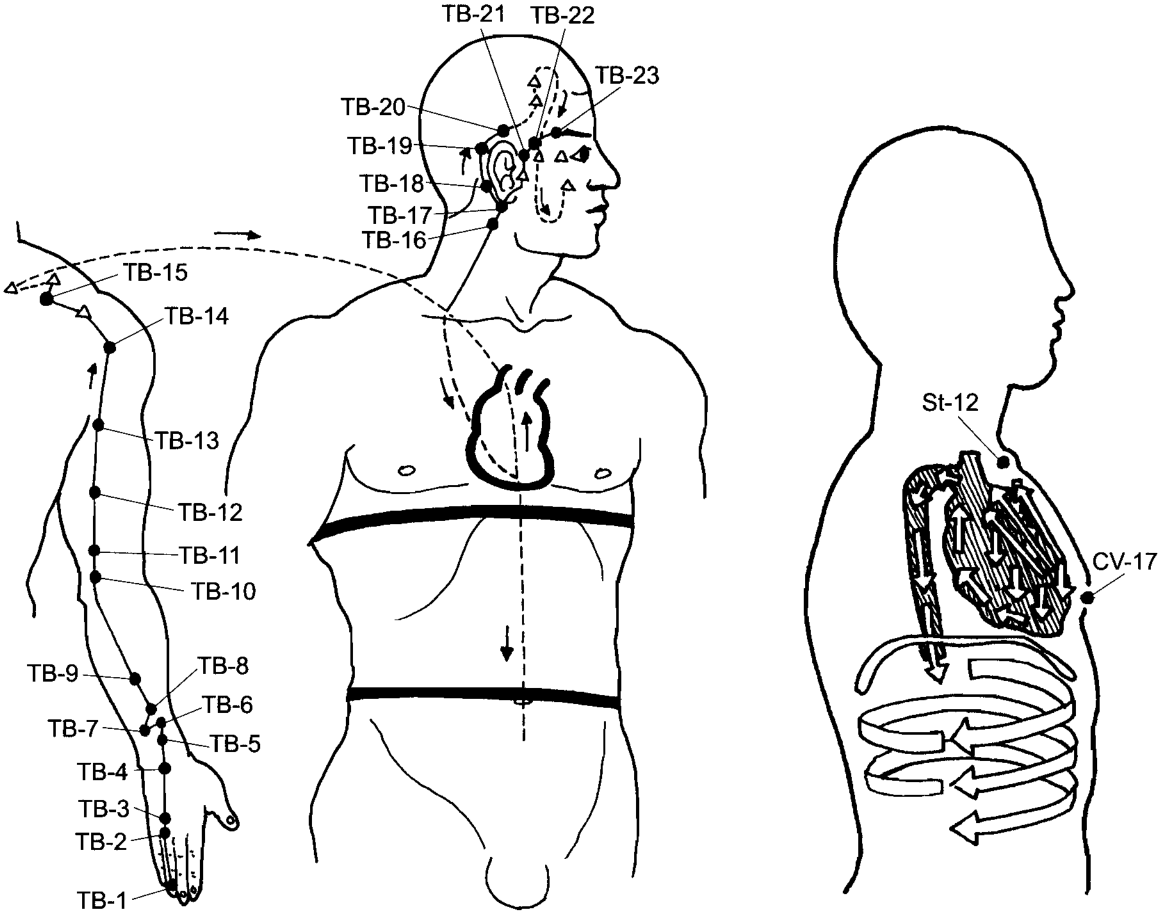
TB-21
TB-22
TB-23
TB-20
TB-19
TB-18
TB-17
TB-16
TB-15
TB-14
TB-13
TB-12
TB-11
TB-10
TB-9
TB-8
TB-6
TB-7
TB-5
TB-4
TB-3
TB-2
TB-1
St-12
CV-17

The internal flow of the Triple Burner Channels moves from St-12 down to CV-17, where it disperses into the chest and envelops the Pericardium. From the Pericardium, as the energy flows down through the diaphragm, it circulates through the chest and permeates the Triple Burners (a continuing branch flows upward to the St-12 points).

Figure 6.20. The Internal and External Qi Flow of the Triple Burners (TB) Channel

Qi, and transport Yuan Qi from the Kidneys to all the other organs of the body. This name refers to the Triple Burners' role in regulating metabolic functions, as well as their production of Qi, Blood, Body Fluids, and waste.

The Triple Burners' energy is composed of Zong Qi (Essential Qi). The Zong Qi assists the Heart in circulating the Blood and the Lungs in respiration. The Zong Qi also assists in conveying the Ying Qi (Nourishing Qi) to nourish the Blood, organs, and tissues of the body, as well as strengthens the Wei Qi (Protective Qi) that protects the external body (see Chapter 14).

1. The Upper Burner

   The Upper Burner is formed from the Fire created from the combined energies of the Heart, Pericardium, and both Lungs. The Upper Burner's energy is housed from the area of the head, throat, and upper chest, to the diaphragm. The Upper Burner is responsible for respiratory and cardiac functions. It moves the body's finer energy, circulating and distributing nutrients and Qi throughout the body like a "mist."

2. The Middle Burner

   The Middle Burner is formed from the Fire created from the combined energies of the Stomach, Spleen, pancreas, and Gall Bladder. The Middle Burner's energy is housed from the upper abdomen (the diaphragm) and the umbilicus.

   The Middle Burner is responsible for digestion, fermentation, and transformation of food and drink into nutrients for distribution. In the Middle Burner area, the Small Intestine connects downward to the Bladder. The Small Intestine transforms waste, that is then sent down to the Large Intestine, and also distills the body's fluids, sending them down to the Bladder.

   The Ying Qi (Nourishing Qi) of the Middle Burner receives its nourishment from the Stomach and Spleen. The Spleen extracts Gu Qi from the food "prepared" by the Stomach, churning the food essence into "foam." The Spleen refines this energetic foam and sends the processed Gu Qi to the Lungs. The Lungs further refine the Gu Qi, sending the impure part to the Kidneys (for additional refinement of the Gu Qi). The Kidneys return the Clean Qi to the Lungs and send the turbid part to the Bladder to be expelled from the body. The Qi from the air mixes with the Gu Qi in the Lungs to produce Zong Qi. The Lungs circulate the refined Gu Qi in the form of vapor or mist which is housed in the Upper Burner. The Heart then produces Blood from this vapor.

3. The Lower Burner

   The Lower Burner is formed from the Fire created from the combined energies of the Liver, Kidneys, Bladder, Intestines, and genitalia. The Lower Burner's energy is housed in the area which starts just below the umbilicus and extends to the lower perineum.

   The Lower Burner is responsible for the reproductive functions and for the filtering and elimination of waste products (Figure 6.21). It moves the body's coarser energy, acting as a "channel for water."

The Triple Burners contribute to the process of the three stages of transformation and ais in the distribution of Ying and Wei Qi throughout the body as follows:

- After the Gu Qi (food energy) has been separated into Clean and Turbid Qi, the Upper Burner releases the body's clean Wei Qi, directing it to the Lungs;
- the Middle Burner releases the body's clean Ying Qi, directing it to all of the body's organs and tissues; and
- the Lower Burner releases the body's Body Fluids, directing the turbid part to the Bladder.

PSYCHO-EMOTIONAL ASPECTS

The Triple Burners are considered the Ambassadors or "intermediaries" for the body's Yuan Qi. On a psychological level, they can be used to move Qi and lift depression derived from stagnation of Liver Qi.

When the Triple Burners, which regulate the consciousness, are full, the consciousness becomes stable and the Mind's intent is benevolent and kindhearted. The Triple Burners are also linked

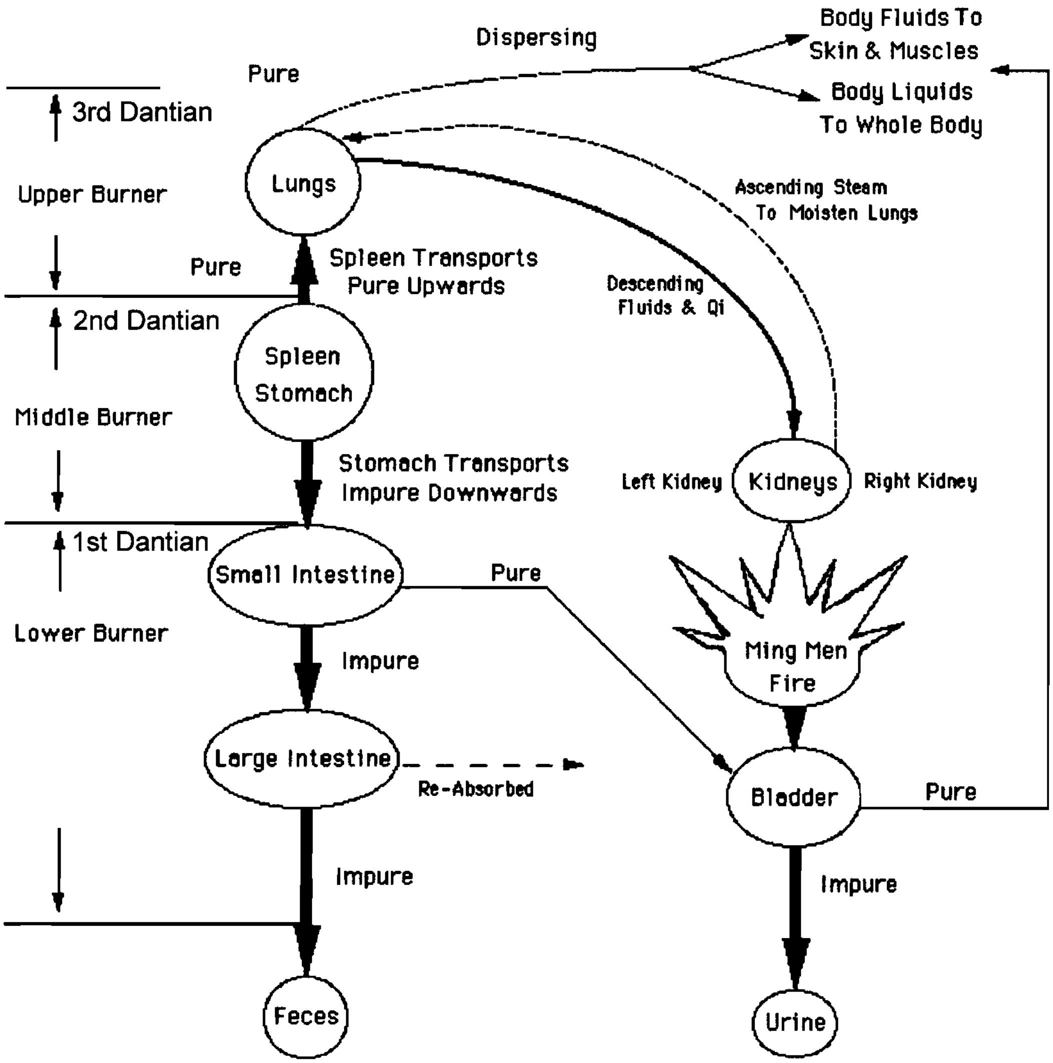
Dispersing
Body Fluids To Skin & Muscles
Body Liquids To Whole Body
Pure
3rd Dantian
Upper Burner
Lungs
Ascending Steam To Moisten Lungs
Pure
Spleen Transports Pure Upwards
Descending Fluids & Qi
2nd Dantian
Spleen Stomach
Middle Burner
Stomach Transports Impure Downwards
Left Kidney
Kidneys
Right Kidney
1st Dantian
Small Intestine
Pure
Lower Burner
Ming Men Fire
Impure
Large Intestine
Re-Absorbed
Bladder
Pure
Impure
Impure
Feces
Urine

Figure 6.21. The Metabolic Functions of the Triple Burners are divided into three parts or functions: the Upper Burner, Middle Burner, and Lower Burner. These different functions of the Triple Burners control the movement of various types of Qi in the three stages of energy production:

Receiving—The Upper Burner, also known as mist, includes the organs from the diaphragm upwards (the Pericardium, head, throat, Lungs, and Heart). The Upper Burner assists the Spleen in directing the body's clean fluids upwards and helps the Lungs in dispersing and scattering the fluids to the skin and muscles.

Rotting and Ripening—The Middle Burner, also known as the muddy pool, and includes the area between the diaphragm and umbilicus (Stomach, Spleen, and Gall Bladder). The Middle Burner assists the Stomach in the transformation and transportation of Fluids and Liquids and in directing the impure part downwards.

Excreting—The Lower Burner, also known as the drainage ditch, includes the organs below the umbilicus (Liver, Kidneys, Intestines, and Bladder). The Lower Burner assists the Small Intestine, the Blood, and the Kidneys in transforming and transporting liquids and wastes.

with the Heart and Pericardium and are affected by the emotion of joy.

When the energy of the Heart is strong and pure (without guilt), and the desires and thoughts of an individual are at peace, then the energy of the body's sexual essence (Jing) will spread into the Triple Burners, and the Blood will flourish within the individual's vessels. If the "fire of desire" is allowed to Heat and combine with the energy of the Triple Burners, the energy of the individual's sexual essence will overflow, mixing itself with the energy of the Mingmen and will leave the body via the reproductive organs and tissues. This leads to Jing and Qi depletion.

THE TRIPLE BURNERS' ENERGY FLOW

The Triple Burner Channels store more Qi than Blood, and act more on the energetic and nervous functions of the body than on physical substance and Blood functions. At the high-tide time period (9 p.m. to 11 p.m.), Qi and Blood abound in the Triple Burner Channels, and can therefore be more easily dispersed and purged at this period; whereas during low-tide (9 a.m. to 11 a.m.), they can be more readily tonified. The Triple Burner Channels' energy acts on the skin, muscles, and nerves found along their pathways.

WESTERN MEDICAL PERSPECTIVE

Western medicine recognizes no translation of, or reference to the Triple Burners.

PATHOLOGICAL MANIFESTATIONS

The main diseases along the Triple Burner Channels involve the face, ear, cheek, larynx, and neck. Diseases of the Triple Burner Channels also include disorders of the back of the upper extremities from the midline to the upper arm and forearm.

When diagnosing problems due to dysfunctions of the Triple Burners, the Qigong doctor considers the following.

1. A blockage of the Wei Qi located within the Upper Burner causes an impairment of the Lungs' dispersing function. This can result in the invasion of the Lungs by External Evils (e.g., Wind and Heat, etc.), that penetrate the Pericardium (which corresponds to the initial stage of externally contracted Heat diseases).
2. A blockage of the Ying Qi located within the Middle Burner causes an impairment of the Spleen's transporting function. This can result in gastrointestinal Heat stagnation, and cause Spleen and Stomach Damp Heat (which corresponds to the second stage of externally contracted Heat diseases).
3. A blockage of the Body Fluids located within the Lower Burner causes an impairment of the Bladder's transformation function. This results in the deep penetration of Toxic Evils, weakening the body's Kidney Yin. This in turn can cause Deficient Liver Blood, and Wind stirring due to Empty Yin (this corresponds to the advance stages of externally contracted Heat diseases).

UNDERSTANDING CHANNEL PATHOLOGY

Channel pathology is the oldest of all the modes of pathological pattern classification, dating back to the Spiritual Axis from the *Yellow Emperor's Classics on Internal Medicine*. In understanding a channel's pathology, the Qigong doctor takes into consideration several aspects of the patient's energy flow (i.e., the channels are viewed as exterior, whereas the organs are viewed as interior). Channel pathology is often related to organ disturbances, but it can also be distinct from organ pathology (Figure 6.22).

The disease of one channel may cause disease in other channels and organs. Likewise, tonification of one channel may cause a tonification of other channels and organs.

THE FOUR CAUSES OF CHANNEL PATHOLOGY

1. The Invasion of External Pathogens such as Cold, Wind, Heat, and Damp leads to channel pathology. The exogenous pathogenic factors often settle in the joints, causing Bi Syndrome (painful obstruction). Channel pathology is quite closely related to joint pathology. In Traditional Chinese Medicine, joints are important to the circulation of Qi and Blood. Qi and Blood gather and concentrate in the joints. Qi enters and exits the channels at the joints. The Five Source Points, also called Shu Points (see Chapter 8), are usually located on

of near the joints. It is at these points that the External pathogenic influences may enter the body and settle in the joints. These External pathogenic influences upset the Yin and Yang balance and block the flow of Qi and Blood, causing pain and swelling. The body's joints are also affected by Deficient Qi and Blood, causing local weakness and pain from lack of movement.

2. Physical strain or the overuse of certain joints can cause local stagnation that creates pain and weakness.
3. Injuries, trauma, and strain can cause local stagnation resulting in stiffness, bruising, and pain.
4. Internal organ disharmonies may also affect the channels.

DIFFERENTIATION BY CHANNEL FULL/EXCESS AND EMPTY/DEFICIENT

When a channel is too full, its pattern characteristics are that of intense pain, stiffness, contractions, and cramps. A red color along the flow of the channel indicates Heat; a bluish color along the flow of the channel indicates Cold.

When a channel is Empty, its pattern characteristics are that of a dull ache, weakness in the muscles, numbness, and muscle atrophy along the channel.

DIFFERENTIATION OF CHANNEL PATTERNS BY SPECIFIC CHANNELS

This clinical overview of the body's channels was originally expressed through the knowledge found in the Spiritual Axis (for more in-depth dissertation please see *The Foundations of Chinese Medicine* by Giovanni Maciocia).

1. The diseases of the Gall Bladder Channels may cause: Alternating chills and fever, headache, acute onset of deafness, pain in the hip and the sides of the body, pain along the lateral sides of the legs, and pain and distention of the breasts.
2. The diseases of the Liver Channels may cause: Headache, pain and swelling of the eyes, and cramps in the legs.
3. The diseases of the Lung Channels may cause: Fever, aversion to cold, stiffness in the chest, pain in the shoulders, clavicle, and arms.
4. The diseases of the Large Intestine Channels may cause: Sore throat, toothache, nosebleed, runny nose, swollen and painful gums, swollen eyes, and pain along the channels.
5. The diseases of the Stomach Channels may cause: Pain in the eyes, nosebleed, neck swelling, facial paralysis, cold legs and feet, and pain along the channels.
6. The diseases of the Spleen Channels may cause: Vaginal discharge, weakness of the leg muscles, and a cold feeling along the channels.
7. The diseases of the Heart Channels may cause: Pain in the eyes, pain along the scapula, and pain along the inner side of the arms.
8. The diseases of the Small Intestine Channels may cause: Pain and stiffness in the neck, and pain along the lateral side of the scapula, elbow, and arms.
9. The diseases of the Bladder Channels may cause: Fever and aversion to cold, headache, stiff neck, pain in the lower back, pain in the eyes, pain in the backside of the leg along the channels.
10. The diseases of the Kidney Channels may cause: Lower back pain and pain in the sole of the feet.
11. The diseases of the Pericardium Channels may cause: Stiff neck, contraction of the elbow and hand, and pain along the course of the channels.
12. The diseases of the Triple Burner Channels may cause: Alternating chills with fever, acute onset of deafness, pain and discharge from the ear, pain at the top of the shoulders, pain in the elbow, and pain along the course of the channels.

SUMMARY OF TWELVE PRIMARY CHANNEL PATHOLOGY

The clinical significance of studying the Twelve Primary Channels is evident through the observation of certain pathological manifestations that indicate specific patterns peculiar to that channel and its internal organ. The quality, quantity and proportions of Qi and Blood circulation

Twelve Primary Channels	**Pathologies**
Bladder	• Diseases of the top of the head, brain, neck, back, lumbar and sacral region • Diseases of the back of the leg and thigh and the external side of the foot
Gall Bladder	• Pain in upper right- and left-quadrants of abdomen • Diseases of the head, face, eye, ear • Diseases of the external sides of the lower extremities
Heart	• Diseases that exert pressure on the brain, eyes, or pharyngeal wall • Diseases of the Heart and lateral side of the chest • Diseases of the ulnar palmar side of the upper extremities • Insomnia, impairment of consciousness, amnesia, and psychosis
Kidneys	• Diseases that cause general deterioration of the entire body • Weakness in the lower extremities and lumbar pain • Hot sensations deep inside the feet
Large Intestine	• Diseases of the lower part of the face, nose, oral cavity, teeth, throat, and front part of the neck • Diseases of the back and radial side of the upper extremities
Liver	• Swelling and a distended sensation of the hypochondrium • Diseases of the lower abdomen and genital organs
Lungs	• Chest and Lung diseases • Diseases on the radial side of the upper arm and palmar area of the hand
Pericardium	• Diseases of the Heart, front of the chest, major blood vessels • Diseases of the midline to upper palmar side of the upper extremities • Mental abnormalities
Small Intestine	• Diseases of the face, ear, cheek, lower jaw, neck, and throat • Diseases of the back ulnar side of the upper extremities
Spleen	• Diseases of the tongue and throat • Gastrointestinal diseases (disturbances of digestion and absorption of food) • Diseases of the inner side of the lower extremities
Stomach	• Diseases of the face, nose, oral cavity, teeth, throat, and front of the neck • Diseases of the abdomen, the frontal part of the lower extremities and gastrointestinal area • Certain psychiatric diseases
Triple Burners	• Diseases of the face, ear, cheek, larynx, and neck • Diseases of the back of the upper extremities from the midline of the torso to the upper arm and forearm

Figure 6.22. Pathologies of the Twelve Primary Channels

through each of the Twelve Primary Channels has a vital impact on all the body's organ systems. These Qi and Blood flow concepts define the reciprocal relationships existing between the body's internal organs, channels, and energetic points situated throughout the body.

Figure 6.22 describes some of the most common Twelve Primary Channel pathologies, and their clinical manifestations along the "root and branches" of these channels.

The term "root and branch," is used to describe the difference between the medial and lateral, upper and lower orientations of the channels and their points of origin. They are also used to describe the progression of a disease, as well as the priorities and sequence of the Qigong doctor's treatment.

Specifically, the word "branch" is used to describe the channels' flow of Qi and Blood and the progression and direction of a disease. The word "root" is used to described the channels' internal organ, or the origin of a disease.

CHAPTER 7

THE EIGHT EXTRAORDINARY VESSELS AND COLLATERALS

THE EIGHT EXTRAORDINARY VESSELS

Extending from the body's Taiji Pole are the Three Dantians, Five Jing Shen, and the Eight Extraordinary Vessels. These energy systems feed all the vital organs and the body's Twelve Primary Channels. The Eight Extraordinary Vessels are so called because of they differ in energetic function from the Twelve Primary Channels. These eight vessels have an extraordinary ability to regulate the deeper energetic reservoirs of the body. They are the first vessels (or channels) to form in the developing fetus, and are sometimes called the Eight Ancestral Channels, Eight Prenatal Channels, Eight Preheaven Channels, or Eight Psychic Channels.

These vessels represent the merging of the mother's and father's energy, and the linking of the body's prenatal and postnatal energies. They functionally connect to the Twelve Primary Channels and circulate the Jing Qi (Essence Energy) throughout the body.

The energetic expansion and contraction of the Eight Extraordinary Vessels affects the creation and development of the baby's tissues up until the time of birth. In energetic embryology, the three-dimensional space of the impregnated egg is viewed as an object that can be described by eight different surfaces along four axes. Each of the Eight Extraordinary Vessels corresponds with one of these directions:

- anterior/posterior—Governing and Conception Vessels
- superior/inferior—Yin and Yang Linking Vessels
- right/left—Yin and Yang Heel Vessels
- interior/exterior—Thrusting and Belt Vessels

THE FUNCTION OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels have neither a direct connection nor an internal/external relationship with the internal organs. Similar to the Three Dantians' function of distributing the body's energy, these Eight Extraordinary Vessels are also reservoirs that regulate the distribution and circulation of Jing and Qi inside the body. They are the foundation of the body's energy, bridging the Yuan Qi (Original Prenatal Energy) with the body's postnatal energy.

The Eight Extraordinary Vessels serve as reservoirs of Qi. When these reservoirs become full, the energy overflows into the center channel or Taiji Pole. This stimulation of the Taiji Pole expands consciousness and increases perceptual intuition.

The Eight Extraordinary Vessels have five main functions: They serve as reservoirs of Qi, store and circulate Jing Qi, circulate the Wei Qi, regulate the body's life cycles, and integrate the six primary Yang organs with the Six Extraordinary Organs and with the Kidneys.

1. They Serve as Reservoirs of Qi. If the Qi flow of the Twelve Primary Channels becomes Excessive, they overflow into the Eight Extraordinary Vessels, which act as Qi reservoirs, receiving the Excess Qi for storage and distribution, while at the same time regulating the energy flow. If the Qi flow of the Twelve Primary Channels becomes Deficient, they can draw from the reservoirs of energy stored within the Eight Extraordinary Vessels which again regulate the body's energy flow, bringing it back into balance.
2. They Store and Circulate Jing Qi. The Eight Extraordinary Vessels draw their energy from the Kidneys and are responsible for storing and circulating the body's Jing Qi throughout the tissues, particularly to the skin and hair and to the Six Extraordinary (Curious) Organs, also known as the Six Ancestral Or-

gans. The Six Extraordinary Organs are the Brain, Bone, Marrow, Blood, and Gall Bladder (also included is the Uterus in women).

3. They circulate the Wei Qi (Protective Energy) to protect the body against invasion by External pathogens. The Governing, Conception, and Thrusting Vessels are primarily responsible for circulating the body's Wei Qi over the thorax, abdomen, and back.

   The Eight Extraordinary Vessels provide the link between the Kidneys Jing and the Wei Qi. Although the Wei Qi is circulated by the Lungs, its root is in the Kidneys. This illustrates how the Kidneys play an important role in supporting the Lungs function of circulating the body's Wei Qi. The Kidneys function on the Wei Qi field and therefore the immune system can become vulnerability to pathogenic factors, as well as in causing allergic diseases, e.g., asthma. It also demonstrates why the Kidneys are always at the root of latent Heat syndromes (chronic fatigue), caused by Kidney Deficiency.
4. The Eight Extraordinary Vessels regulate the body's Life Cycles. In the first chapter of *the Yellow Emperor's Classic of Internal Medicine,* the life changes in women (every seven years) and men (every eight years) are governed by the Conception and Thrusting Vessels. These life cycles are related to the body's Jing and the Eight Extraordinary Vessels' role in moving and circulating the body's Jing (see Chapter 13).
5. They Integrate the Six Extraordinary (Curious) Organs with the Six Primary Yang Organs and the Kidneys. The Eight Extraordinary Vessels provide the link between the body's brain, uterus, blood vessels, Gall Bladder, Marrow and bones, and the body's internal energy flow.
   a. The Brain is Regulated by the Governing Vessel and the Yin and Yang Heel Vessels.
   b. The Uterus is regulated by the Thrusting and Conception Vessels.
   c. The Blood Vessels are regulated by the Thrusting Vessels.
   d. The Gall Bladder is regulated by the Belt Vessel.
   e. The Marrow is regulated by the Thrusting Vessels.
   f. The Bones are regulated by the Thrusting and Conception Vessels.
6. They Integrate the Four Seas. The *Magic Pivot (Ling Shu, Hai Lun)* states, "People have Four Seas... the Sea of Marrow, the Sea of Blood, the Sea of Qi, and the Sea of Grain and Water." The Eight Extraordinary Vessels provide the link between the body's Four Seas and the body's internal energy flow.
   a. The Sea of Marrow is the brain, and it is related to the Governing Yin and Yang Heel Vessels. The Sea of Marrow points are located on the Governing Vessel GV-16 and GV-20. When the Sea of Marrow is Deficient, there will be headaches and dizziness.
   b. The Sea of Qi is located in the chest of the center, and is regulated by the Conception Vessel CV-17. Some Medical Qigong schools maintain that there are two reservoirs of Qi: the Middle Dantian, being the Sea of Postnatal Qi, and the Lower Dantian, being the Sea of Prenatal Qi (which is regulated by the Qihai CV-6 point). When the Sea of Qi is in Excess, there is a feeling of fullness in the chest, dyspnea (urgent breathing), and a red complexion. When the Sea of Qi is Deficient, there is weak energy and insufficient speech.
   c. The Sea of Nourishment (also known as the Sea of Grain and Water) is the Stomach, and it is regulated by the Thrusting Vessels, which are accessed by the St-30 points. When the Sea of Water and Grain is in Excess, there is a feeling of fullness in the abdomen. When the Sea of Water and Grain is Deficient, there is a feeling of hunger with an inability to eat.
   d. The Sea of Blood (also known as the Sea of the Twelve Primary Channels) is related to the Thrusting Vessels, the Liver, and Xuehai Sp-10 points. The points can

also be accessed through Bl-11, St-37, and St-39. When the Sea of Blood is in Excess, the body feels big; When the Sea of Blood is Deficient, the body feels small.

THE EIGHT EXTRAORDINARY VESSELS AND MEDICAL QIGONG THERAPY

The energetic pathways, functions and use of the Eight Extraordinary Vessels and points in Medical Qigong are often quite different than those taught in acupuncture colleges, even though they may sometimes bear the same names.

Most acupuncturists, with the exception of some Japanese researchers, often pay little attention to the Eight Extraordinary Vessels in diagnosis and treatment. One Medical Qigong doctor from China has claimed that a study done in China showed that the Eight Extraordinary Vessel treatments were far more effective than those of the Traditional Chinese medical protocols. The results, however, were not released because officials did not want to disturb the TCM institutional operations. The reason Medical Qigong makes use of the Eight Extraordinary Vessels, is that they are much easier to access through Qi emission and Medical Qigong self-regulation prescriptions than through acupuncture therapy.

The Eight Extraordinary Vessels have been of special importance to Medical Qigong practitioners for thousands of years, and are viewed as the foundation of the body's energy, the bridge between the Yuan Qi (Prenatal Qi) and the Postnatal Qi. These Eight Extraordinary Vessels affect the body on the deepest level of our basic constitutional energy. In Medical Qigong regulation exercises, the focus on opening the flow of energy through the Eight Extraordinary Vessels is a prerequisite for opening the energy flow in the Twelve Primary Channels. The major purpose of opening the Eight Extraordinary Vessels is to provide a container for storing the Three Forces: Heaven Force (absorbing universal energy), Earth Force (absorbing environmental energy), and Man's Force: Jing, Qi and Shen (Essence, Energy and Spirit).

It is essential for the Qigong doctors to open their Eight Extraordinary Vessels to attain mastery of their energetic body. The Microcosmic Orbit is introduced for purgation, fusion, tonification, and regulation of the Governing and Conception Vessels in the first stages of training. This exercise restores a healthy flow of energy throughout all of the channels and vessels creating a healthy and balanced body.

Through the refinement of their energy, Qigong doctors receive more Qi, and the quality of their Blood changes, i.e., its energetic potential changes to contain more Qi and less Blood. At this advanced state of transformation, Qigong doctors rechannel the flow of energy through their body's Three Dantians, including the "mystical pass" (located within the Upper Dantian chambers of the brain). When the flow of energy is directed to stimulate the Mystical Pass, the Qigong doctors acquire greater awareness and control over their own bodies (physical, mental, emotional, energetic and spiritual), and are then able to enhance their perception and communication with all the other energetic planes of the world (mineral, plant, animal, and human), and with the Dao (or divine).

CLINICAL USE OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels form and establish the foundational energetic pattern for the developing fetus's tissues, as well as the adult's entire body. The clinical use of the Eight Extraordinary Vessels can be divided into several different approaches of energy manipulation to change the energetic patterns within the patient's tissues.

One approach divides the Eight Extraordinary Vessels into four pairs of Yin and Yang vessels. Each of the vessels is grouped according to the same polarity, two pairs of Yin and two pairs of Yang. When paired this way, each set of the Yin and Yang Extraordinary Vessels has a common range of energetic action in terms of the patient's body area.

1. The Conception and Yin Heel Vessels affect the energetic flow to the patient's abdomen, chest, Lungs, throat, and face.
2. The Governing and Yang Heel Vessels affect the energetic flow to the back of the patient's legs, as well as the back, spine, neck, head, eyes, and brain.

3. The Thrusting and Yin Linking Vessels affect the energetic flow to the inner aspect of the patient's legs, as well as the abdomen, Stomach, chest, and Heart.
4. The Belt and Yang Linking Vessels affect the energetic flow to the outer aspect of the patient's legs, as well as the sides of the body, shoulders, and sides of the neck.

ENERGETIC FUNCTIONS OF THE EIGHT EXTRAORDINARY VESSELS

The Eight Extraordinary Vessels can also be categorized is according to their energetic functions.

1. The Conception, Governing and Thrusting Vessels are considered the source of all the other Extraordinary Vessels These three vessels affect the patient's energy at a deep constitutional level. They originate directly from the Kidneys and are connected to the body's Jing.
2. The Yin and Yang Heel Vessels are complementary in that they both flow from the legs (controlling the state of the muscles of the legs) into the eyes (controlling the muscles that open and close the eyes).
3. The Yin and Yang Linking Vessels complement each other by linking the body's Yin and Yang Channels.
4. The Belt Vessel is the only horizontal vessel in the body, encircling the main channels. Its energetic influence is in its ability to affect the circulation of energy in the body, especially within the legs.

Each of the Eight Extraordinary Vessels is also connected to one of the Eight Trigrams (or Eight Natural Manifestations), and can be energetically regulated through the use of the Magic Square. For further treatment and clinical protocol see Chapter 31 regarding the Eight Extraordinary Vessel and the Magic Square.

THE GOVERNING AND CONCEPTION VESSELS

The Governing and Conception Vessels are the main rivers of the body's Yin and Yang energies. They are polar aspects of the body, perfectly complementary, like midnight and midday. They are responsible for the formation of the holoblastic cleavage and the first cellar division of the fertilized ovum in embryological development (see Chapter 2).

In China, Dr. Li Shi Zhen was the first to believe that the Governing and Conception Vessels are two branches of the same source, an inseparable Yin-and-Yang, front-and-back duality. These vessels connect the uterus with the Kidneys, Heart, and Brain.

They originate externally at Huiyin CV-1 point, ascend the front and back of the torso, and form a small circle when the tip of the tongue touches the upper palate in the mouth and the anal sphincter is squeezed. Not only does this action complete the balance of Fire (Heart) and Water (Kidneys) energy, it also increases the body's protective Wei Qi.

Regulating the Conception and Governing Vessels is a priority in Medical Qigong practice. Along these vessels, the Qigong practitioner draws the Yang Fire and Yin Essence up and down his or her body, fusing the Water and Fire energies together. This fusion facilitates a Yin and Yang balance throughout the body.

The Governing and Conception Vessels each have two energy flows on the anterior and posterior vertical midline of the body. Each vessel's pathway is complete, being composed of an ascending energetic flow and a descending energetic flow. The duality of these two medial lines joins at the extremities (the head and perineum), forming one complete circle of energetic current (Figure 7.1).

Both vessels are superimposed on each other, with the energy of the Governing Vessel being predominant up the back and inferior down the front (behind the Conception Vessel); the energy of the Conception Vessel is predominant up the front and inferior down the back (behind the Governing Vessel).

The energetic flow of the Conception Vessel corresponds to Yin, negative polarity, the female aspect, and responds to bass tones; while the energetic flow of the Governing Vessel corresponds to Yang, positive polarity, the male aspect, and responds to treble tones.

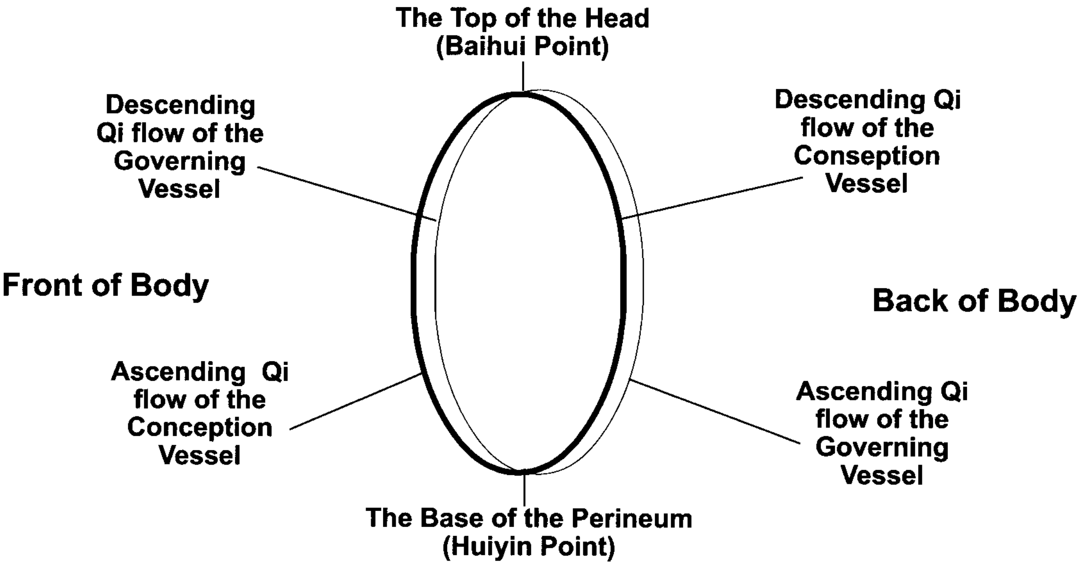
The Top of the Head
(Baihui Point)
Descending Qi flow of the Governing Vessel
Descending Qi flow of the Conseption Vessel
Front of Body
Back of Body
Ascending Qi flow of the Conception Vessel
Ascending Qi flow of the Governing Vessel
The Base of the Perineum
(Huiyin Point)

Figure 7.1. The Energetic Flow of the Conception and Governing Vessels

It is important to understand the energetic potential of these two currents flowing in both an ascending and descending direction on the posterior and anterior vertical midline in the body. The energetic movement of these two currents explains why there are two opposite directions of focused concentration used in energy cultivation meditations. One direction follows the Microcosmic Fire cycle along the Governing Vessel (up the spine and down the chest) to stimulate the emotional regulation of the acquired mind (Zhi Shen); the other direction follows the Microcosmic Water cycle along the Conception Vessel (up the chest and down the spine) to stimulate spiritual intuition and activate the perceptions of the Yuan Shen.

THE GOVERNING VESSEL OR DU VESSEL

The word Du translates as "governing," and refers to a general, someone who controls and is in charge. During the development of the embryo, the Governing Vessel is responsible for the development and formation of the medulla oblongata and cerebrum. The Governing Vessel controls all the Yang Channels in the body, and is called the Sea of Yang Channels. It is responsible for nourishing the brain and spinal cord and for consolidating the Yuan Qi in the Kidneys.

The Governing Vessel originates in the Lower Dantian in both men and women. It is composed of many energetic branches, emerging externally at the perineum. The primary branch ascends externally on the midline of the back, over the head and ends in the frenulum of the upper lip. As the Qi flows up the Governing Vessel one branch enters the interior of the spinal column and enters the Kidneys. Another set of branches enters the brain at the Fengfu GV-16 point and ascends internally to the vertex at the Baihui GV-20 point.

From the lower perineum, another set of branches encircles the anus and envelops the external genitals, ascends past the navel, Heart, and throat to circle the mouth. After circling the mouth, it further ascends to the eyes, emerging just below the middle of the eyes. From the inside of the eyes (at the Jingming Bl-1 point), a set of branches follows the Bladder Channels along the forehead, converging at the vertex of the Baihui GV-20 point and enters the brain. From the brain, the Vessel emerges at the Fengfu GV-16 point, and divides into two additional branches which descend the

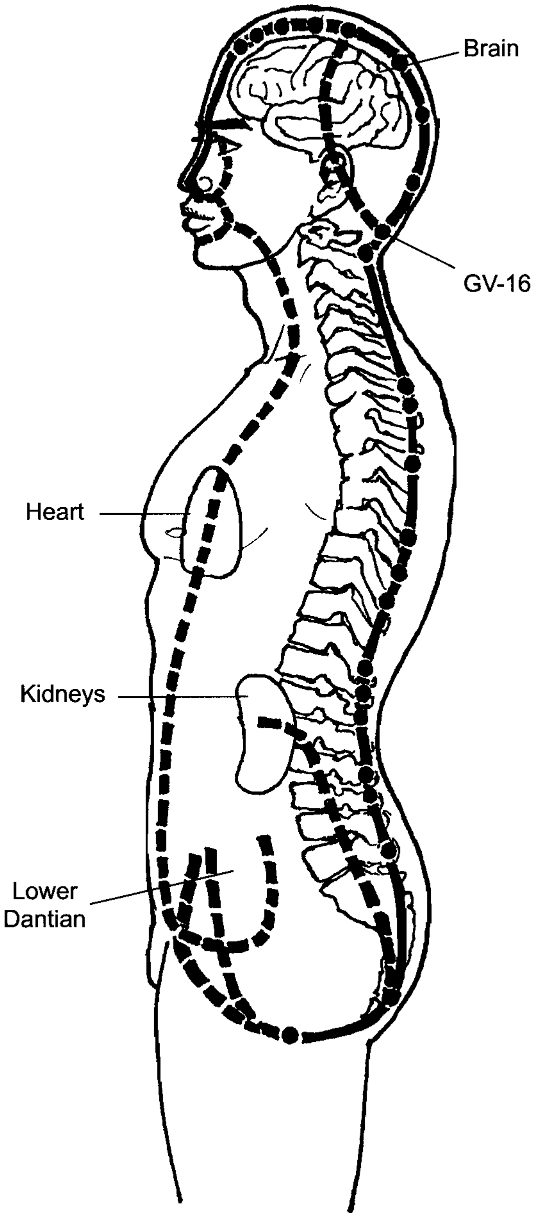

Figure 7.2. The Governing Vessel (Side View)

back, passing through the Fengmen Bl-12 points along the sides of the spine and enter into the Kidneys (Figure 7.2).

The Governing Vessel tonifies the Kidney Yang and strengthens the spine and back. When the Governing Vessel is excessive, the back becomes stiff; when it is depleted, the head becomes heavy, unstable, and shaky (Figure 7.3).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Governing Vessel are as follows.

1. To strengthen and control all the Yang Channels of the patient's body (particularly in relation to the Kidney Yang and the brain);
2. To tonify the Kidney Yang and strengthen the patient's spine and back (especially in cases of chronic lower back pain due to Kidney Deficiency);
3. To purge External Wind from the patient's body when symptoms are present, such as runny nose, head ache, fever, and stiff neck;
4. To purge Internal Wind from the patient's body when symptoms are present, such as dizziness, tremors, convulsions, epilepsy, or the condition of Wind Stroke;
5. To strengthen and nourish the patient's brain and Marrow, and for treating such symptoms as poor memory, dizziness, and tinnitus; and
6. To treat depression, due to the Governing Vessel's connection to the body's Jing (Kidneys, lack of willpower), Qi (Heart), and Shen (Brain).

Note: In men, the Governing Vessel is generally treated by itself, and in women, its treatment is combined with treating the Conception Vessel.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the head (apoplexy, aphasia, epilepsy, headaches, tetanus, etc.), the back, neck, and Kidneys (pain and stiffness in the spinal column), spastic muscle movement of the extremities, night sweating, and Jing Shen disorders (hyperexcitability, hallucinations, etc.).
2. Symptoms of Deficiency result in shaking of the head along with a feeling of heaviness, instability, and an inability to concentrate. The patient lacks physical and mental stamina, displays weakness in character, and may also experience hemorrhoids, sterility, and impotence.

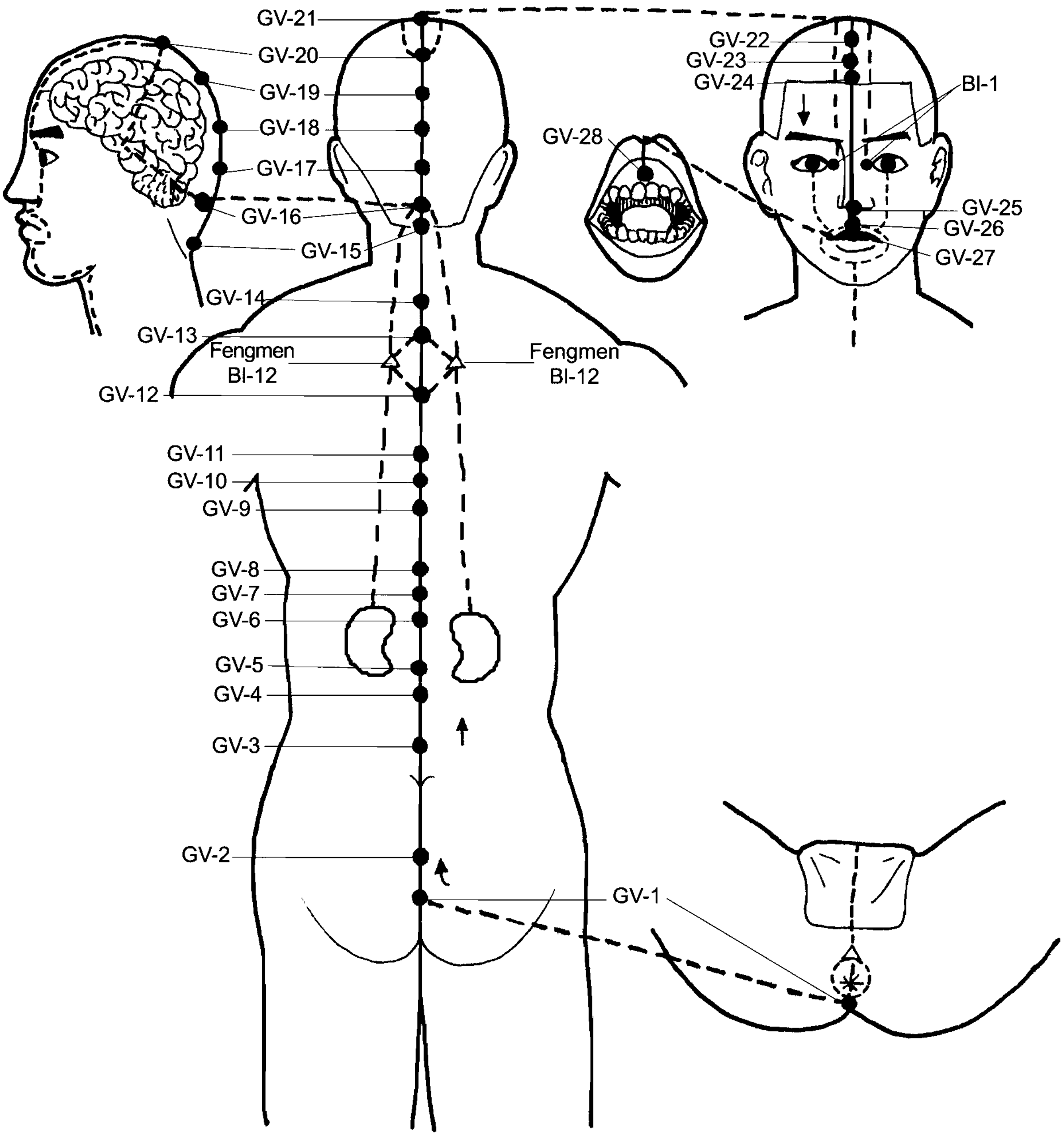
GV-21
GV-20
GV-19
GV-18
GV-17
GV-16
GV-15
GV-14
GV-13
Fengmen
Bl-12
Fengmen
Bl-12
GV-12
GV-11
GV-10
GV-9
GV-8
GV-7
GV-6
GV-5
GV-4
GV-3
GV-2
GV-1
GV-22
GV-23
GV-24
Bl-1
GV-28
GV-25
GV-26
GV-27

Figure 7.3. The Governing Vessel (GV)

THE CONCEPTION VESSEL OR REN VESSEL

The word Ren translates as Conception, and refers to pregnancy, responsibility, or obligation. It can also mean "to accept." The Conception Vessel governs all the Yin Channels of the body and is called the Sea of Yin Channels. It nourishes the Yuan Qi of all Five Yin Organs. The Conception Vessel originates from the Lower Dantian, and specifically in the uterus in females (Figure 7.4).

Functionally, the upper third of the Conception Vessel, on the sternum, controls respiratory functions; the middle third, on the epigastrium, controls digestive functions; and the lower third, on the abdomen, controls the urogenital functions (Figure 7.5).

In women the Conception Vessel is primarily responsible for nourishing the uterus and the genital system and determines the seven-year life cycles. It links the Yin energy with all aspects of conception and reproduction. The Conception Vessel, along with the Thrusting Vessels, have an important relationship with obstetric diseases related to the development of the fetus, delivery, and menstruation.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Conception Vessel are as follows.

1. To strengthen and nourish the Yin energy of the patient's body (especially in women after menopause) and harmonize the Lungs and Kidneys;
2. To regulate the energy of the reproductive system, tonifying the Blood and Yin, and to reduce the effects of Heart Empty-Heat symptoms developed from Kidney Yin Deficiency after menopause (i.e., night sweating, hot flashes, anxiety, mental irritability, insomnia and dizziness, etc.);
3. To promote Blood supply to the uterus and regulate menstrual disorders (i.e., dysmenorrhoea, menorrhagia, amenorrhoea, and metrorrhagia);
4. To move the Qi in the patient's Lower Burner and treat abdominal lumps, as well as myomas, fibroids, and carcinoma in the woman's uterus, and hernia in men; and
5. To stimulate the energetic interaction of the Lungs Channels descending Qi function, and the Kidneys' function of receiving and holding the Lung Qi, which can, when malfunctioning, result in asthma.

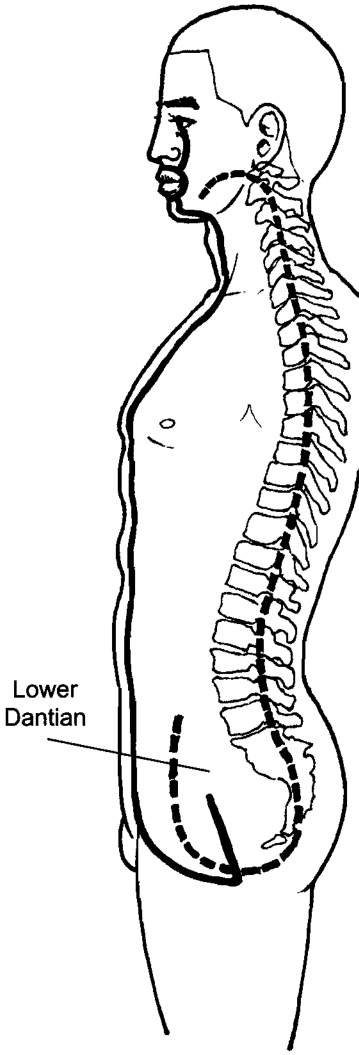

Figure 7.4. The Conception Vessel (Side View)

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the reproductive and gastrointestinal systems (hemorrhoids, diarrhea, decreased urination, etc.). In the male, problems in the Conception

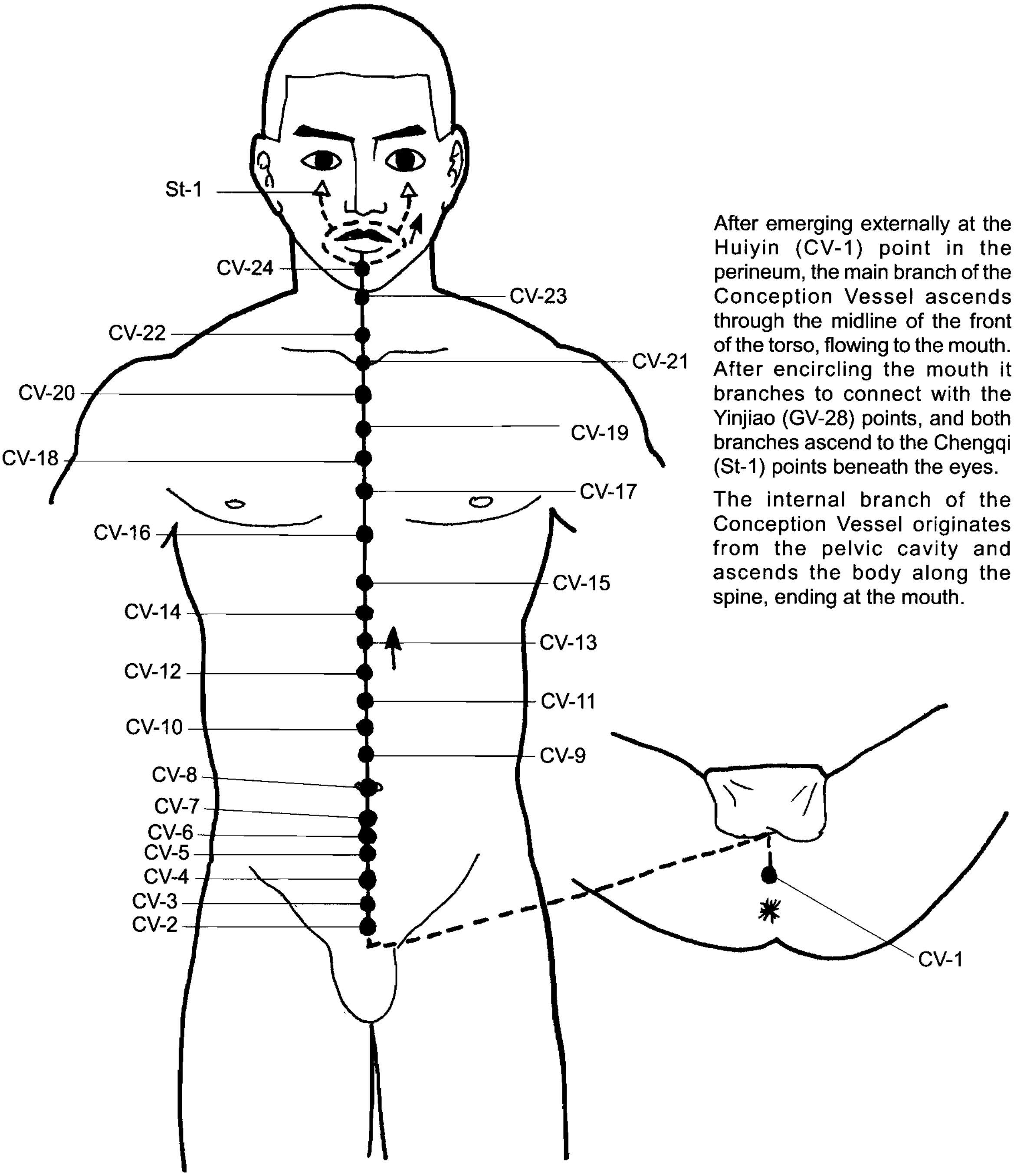
St-1
CV-24
CV-23
CV-22
CV-21
CV-20
CV-19
CV-18
CV-17
CV-16
CV-15
CV-14
CV-13
CV-12
CV-11
CV-10
CV-9
CV-8
CV-7
CV-6
CV-5
CV-4
CV-3
CV-2
CV-1

After emerging externally at the Huiyin (CV-1) point in the perineum, the main branch of the Conception Vessel ascends through the midline of the front of the torso, flowing to the mouth. After encircling the mouth it branches to connect with the Yinjiao (GV-28) points, and both branches ascend to the Chengqi (St-1) points beneath the eyes.

The internal branch of the Conception Vessel originates from the pelvic cavity and ascends the body along the spine, ending at the mouth.

Figure 7.5. The Conception Vessel (CV)

Vessel produce sterility. In the female, the Conception Vessel can cause menstruation problems such as leukorrhea and dysmenorrhea. Other female problems include breast pain, paralysis after delivery, emaciation, and sterility. Other problems of the female reproductive system include the external genitalia and vulva, vagina, and cervix disorders.

2. Symptoms of Deficiency include abdominal pains (hernia), pruritus, and a heavy feeling in the hips, lower ribs, and lumbar area.

THE THRUSTING VESSELS AND THE BELT VESSELS

The Thrusting and Belt Vessels balance the external tissues and internal organ energies, controlling an Excess or Deficient condition by regulating the upper and lower, as well as left and right, energetic quadrants.

The Belt Vessel is responsible for the second cellular division of the fertilized ovum in embryological development (see Chapter 2).

THE THRUSTING VESSELS OR CHONG VESSELS

The word Chong translates as "a street" and expresses the idea of passing or penetrating through something. This passing through action refers to its function as the vital pathway for alchemical transformation, used to produce energetic change (Jing to Qi, Qi to Shen and vise versa). The Thrusting Vessels, also called the Penetrating Channel, and are regarded as the Sea of Blood and the Sea of the Twelve Regulating Channels. They regulate both the Qi and Blood of all Twelve Primary Channels and extend to the anterior, posterior, upper, and lower parts of the body. During the development of the embryo, the Thrusting Vessels are responsible for the development of the adrenal glands and the cortex.

The Thrusting Vessels originate in the Lower Dantian. The Qi of the Thrusting Vessels travel to the head and face to flow into the Chong and penetrate the lower limbs, irrigating the body's Yin (Figure 7.6). According to some Chinese doctors, all energetic points that have Chong in their names relate to the Thrusting Vessels (i.e., Qichong St-30, Taichong St-42, and Shaochong Ht-9).

Along with the Conception Vessel, the Thrusting Vessels are responsible for regulating the changes in life cycles that occur every eight years in men, and every seven years in women. Abnormalities of the Thrusting Vessels during the beginning of pregnancy result in the mother's body aborting the fetus. Abnormalities of the Thrusting Vessels during the end of pregnancy result in an inability to expel the placenta.

The Thrusting Vessels control the woman's menstruation in all aspects, influencing the supply and amount of Blood in the uterus, and nourishing the woman's Jing. They also flow along the Kidney Channels and are related to the muscles of the abdomen and the penis (and are responsible for tightening the abdominal muscles).

THE ENERGETIC PATHWAY OF THE THRUSTING VESSELS

The classical description of the Thrusting Vessels' energy flow begins inside the uterus in women (the corresponding lower abdominal area in men), continuing down to the perineum. From the perineum, one branch ascends inside the spinal column, while four other branches flow externally.

Two external energetic rivers ascend up the front of the body alongside the Kidney Channels and enter into the chest. They then ascend through the throat, encircling the lips, and end at the corner of the eyes. When the Thrusting Vessels are full and overflowing (through meditation), they radiate a sparkling white-light energy, which can be observed to extend from the upper chest to the eyes.

From the perineum, two other external energetic rivers descend down the inner aspect of the thighs (one on each leg), along the Spleen Channels. They divide again at the medial malleolus. One set of branches flows into the big toes, a second set of branches terminates at the bottom of each foot.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Thrusting Vessels are as follows.

1. To strengthen, nourish, and regulate weak

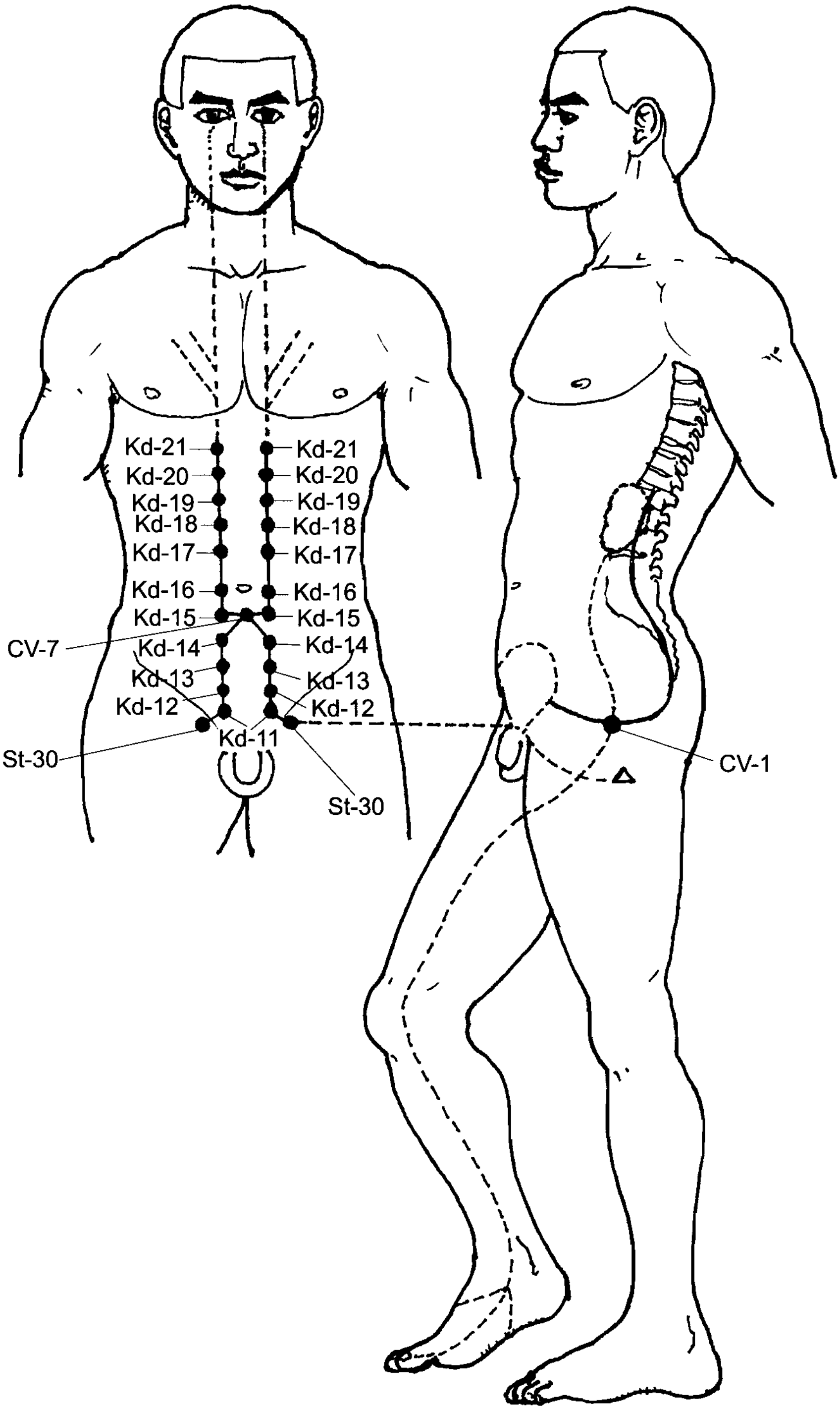
Kd-21
Kd-20
Kd-19
Kd-18
Kd-17
Kd-16
Kd-15
CV-7
Kd-14
Kd-13
Kd-12
St-30
Kd-11
Kd-21
Kd-20
Kd-19
Kd-18
Kd-17
Kd-16
Kd-15
Kd-14
Kd-13
Kd-12
St-30
CV-1

Figure 7.6. The Thrusting Vessels

constitutions with digestive symptoms (e.g., abdominal distension, poor appetite, and poor assimilation of food);

2. To move the Blood of the Heart to relieve symptoms of pain and stiffness of the chest (the Thrusting Vessels control all the Blood in the connecting channels);
3. To purge Qi and Blood stagnation in the abdomen and chest;
4. To treat feelings of anxiety (within the chest) caused by Rebellious Qi; One of the Thrusting Vessels' most important pathologies is Rebellious Qi (Qi that moves in the wrong direction, going upwards instead of downwards). Feelings of anxiety that arise in the patient's abdomen and ascend to their chest are especially indicative of this malfunction;
5. To treat gynecological problems (hot flashes), when the Qi of the Thrusting Vessels rises upwards, and causes the patient's hands and feet to get cold, the face to get hot, and a feeling of fullness in the chest. The treatment is to regulate the Thrusting Vessels and subdue the Rebellious Qi in the Chong Vessel. If there are accompanying emotional problems and Liver Qi stagnation, then also treat the patient's Lv-3 points; and
6. To treat Rebellious Qi caused by the energy of the Thrusting Vessels rebelling upwards; the symptoms are oppressive feelings in the chest, as well as dizziness, nausea, and vomiting.

PATHOLOGICAL MANIFESTATIONS

The main diseases associated with the patient's External Thrusting Vessels include diseases of the Heart, fullness in the chest and abdomen, gastritis, abdominal pain, and convulsive diseases. If a woman's Thrusting Vessels are Deficient or Empty she may develop such conditions as amenorrhoea, scanty periods, or late periods. If there is stagnant Qi and / or Blood in the Thrusting Vessels, she may experience dysmenorrhoea. Abnormalities of the Thrusting Vessels also result in the mother aborting the fetus.

The Thrusting Vessels work with the Qi of the Kidney Channels to control the brain's physiological functions. If the energy from the Thrusting Vessels to the cortex becomes Deficient, this can inhibit the development of the brain.

THE BELT VESSEL OR DAI VESSEL

The word Dai translates as a "belt" or "girdle," and refers to the action of supporting something. The Belt Vessel is the only horizontal vessel in the body. Beginning at the Mingmen, this vessel encircles the waist like a belt, dipping down into the lower abdominal region anteriorly and running across the lumbar region posteriorly. It connects with the Gall Bladder's -26, -27, and -28 points and crosses the Conception Vessel at CV-4. Many sources say that it connects with the Liver -13 (at the free ends of each 12th rib). It binds, joins, and controls all the channels of the body, exerting an influence on the energetic circulation of the body's Governing and Conception Vessels (Figure 7.7).

Medical Qigong schools teach that the Belt Vessel circle wraps the entire body, like a enveloping cocoon, flowing from the feet to the head. The waist is considered the "hub" of the energetic wheel, and the access points of the Belt Vessels center channel (which circles the waist) is used to control the entire vessel. As the energy of the body increases (through Qi cultivation), the Qi within the entire Belt Vessel increases, circle wrapping the tissues from feet to head, increasing the body's Wei Qi fields (Figure 7.8).

In addition, the Belt Vessel keeps both the Yin and Yang Channels under control, connecting all of the leg Yin and Yang channels as they traverse the body's trunk. As a result, the Belt Vessel assists in regulating the circulation of Qi in those channels, especially from the waist down.

The Belt Vessel has an important influence on the body's physiology by encircling the leg channels, influencing the circulation of energy to and from the legs, as well as influencing the energetic actions of the genitals, waist, and hips. This vessel not only restrains the flow of the body's Liver and Gall Bladder Qi, but also harmonizes the ascending and descending flows of energy from the Kidneys and Spleen through its connection with the Kidney Divergent Channel.

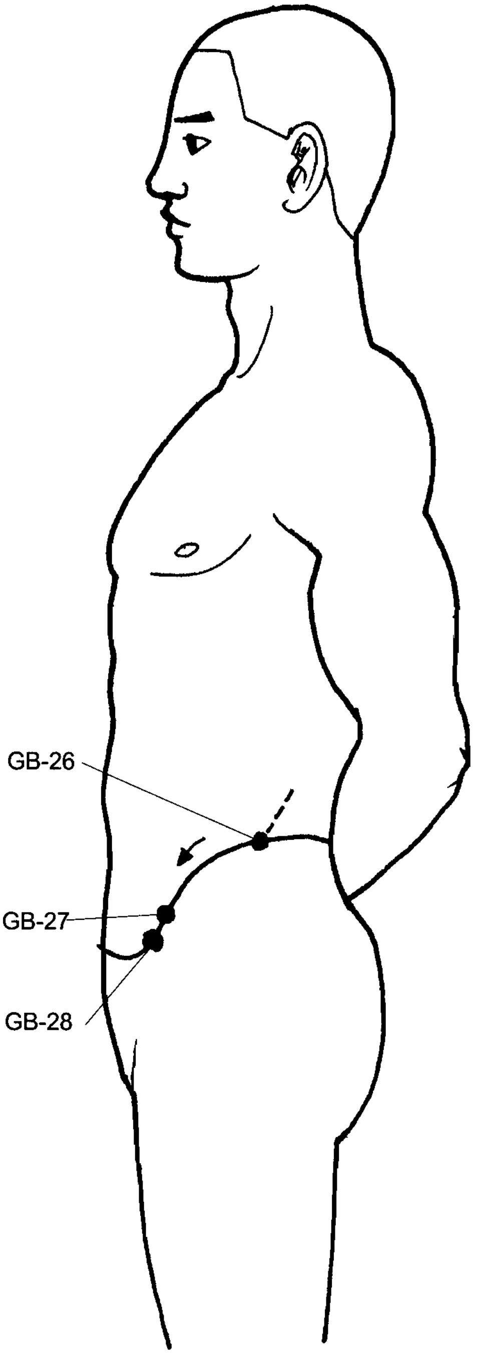

Figure 7.7. The Center of the Belt Vessel

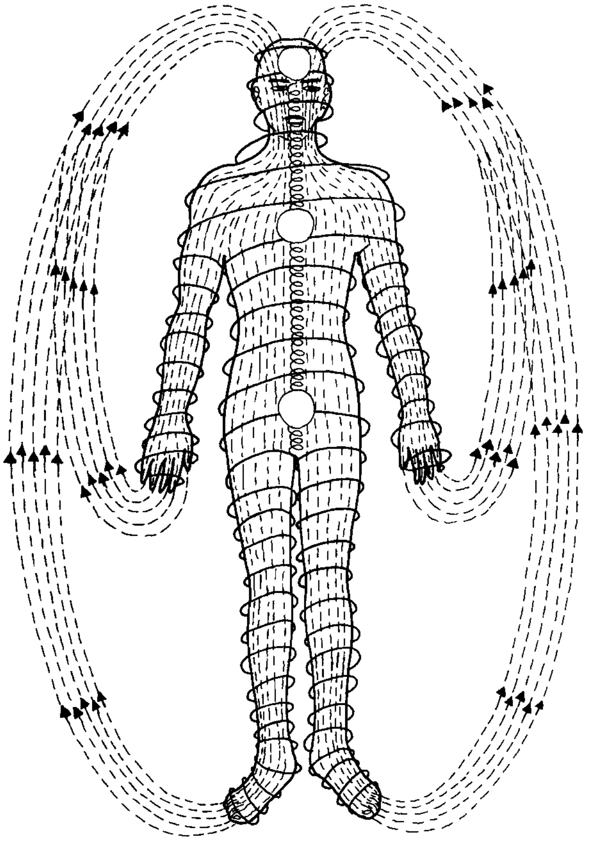

Figure 7.8. The Belt Vessel winds up the body like an energetic coil.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Belt Vessel are as follows.

1. To treat impaired circulation of Qi in the leg channels resulting in such symptoms as cold legs and feet, or tense leg muscles (gastrocnemius and tibiales). This condition is due to Liver Blood not moistening the sinews of the legs;
2. To treat impaired circulation of Qi and Blood that cause numbness, weakness, atrophy, or motor impairment of the leg muscles, due to a Deficiency of energy in the Stomach and Spleen Channels;
3. To tonify and harmonize the circulation of

Liver and Gall Bladder Qi due to Excess patterns of Liver energy;
4. To disperse Damp Heat in the genitals that has resulted in symptoms such as difficulty or burning during urination;
5. To treat hip pain caused by Deficient Liver Blood (leading to sinew and joint malnourishment and Excess of Liver Yang; and
6. To regulate the lower abdomen due to a Belt Vessel imbalance, causing symptoms such as a sagging waist or bloated abdomen.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include pain in the back (lumbar region), and sides of the lower abdomen, as well as weakness in the shoulders, upper extremities, and lower limbs. Symptoms may also include weakness in the opposite sides of the body (e.g., eye, breast, ovary, etc.), and a feeling of heaviness in the body and abdomen (as if carrying 5,000 coins) due to exposure to Dampness.
2. Symptoms of Deficiency have physical sensations similar to that of "sitting in cold water" up to the waist. This description generally refers to pain, weakness, and a cold, heavy sensation in the patient's lumbar and sacral regions. Other symptoms include umbilical, abdominal and lumbar pain, as well as a feeling that something like a stick is pressing against the groin. There can also be abdominal fullness and distention. In women, there may be an abnormal white vaginal discharge, and a prolapse of the uterus.

THE YIN AND YANG HEEL VESSELS OR QIAO VESSELS

The word Qiao translates as "the heel" or "to stand on the toes," and refers to the action of kicking one's foot as high as possible. The Yin and Yang Heel Vessels flow along the medial and lateral aspects of the lower legs and torso, connecting at the inner canthus (by the eyes). They link the channel energy of the body's Yin and Yang Channels and regulate the movement of all four limbs. They also control the amount of energy being used by all the other channels in the body. Once these vessels are full, they relax the tissues, enabling the limbs to become more dexterous.

The Yin and Yang Heel Vessels are sometimes called Bridge Channels, because they act like a bridge linking the stored Qi in the body with the areas in need of Qi. When any channel uses more than its share of energy, other channels become Deficient. Thus the Heel Vessels seek to ensure that energy is always distributed in a balanced way.

The Yin Heel Vessels are an offshoot of the Kidney Channels at the front of the body, while the Yang Heel Vessels are an offshoot of the Bladder Channels at the back of the body. Together, the Yin and Yang Heel Vessels can be used to treat structural imbalances and to harmonize the right and left sides of the body.

Because the Heel Vessels cause the motor nerves to develop during the formative stages of the embryo, the Chinese believe the Yang Heel Vessels cause little boys to be more physically active (running, jumping, etc.), while the Yin Heel Vessels cause little girls to be less actively inclined.

THE YIN HEEL VESSELS

The Yin Heel Vessels control the Yin of the left and right sides of the body. The Yin Heel Vessels influences the male and female reproductive system, as well as the lower abdomen in women.

When Yang energy is slowed down in the Heel Vessels, the Yin energy moves more rapidly. The Excess Yin causes the following problems: sleepiness, the inability to keep the eyes open while trying to stay awake, hypotension, choking, painful urination, stomach rumbling, vomiting, diarrhea, difficult bowel movements, and unconsciousness. Excess Yin in women also causes a difficult labor. Also, when the Yin Heel Vessels are in Excess, the inner leg muscles are tight, while the outer leg muscles are loose. Weakness in the Yin Heel Vessels can cause the feet and ankles to invert, whereas an Excess condition can cause the feet to turn outwards.

The Yin Heel Vessels begin at the superficial part of the ankles, just below the medial malleolus of the tibial bones at the Kd-2 points. They continue up the front of the body, ending at the inner

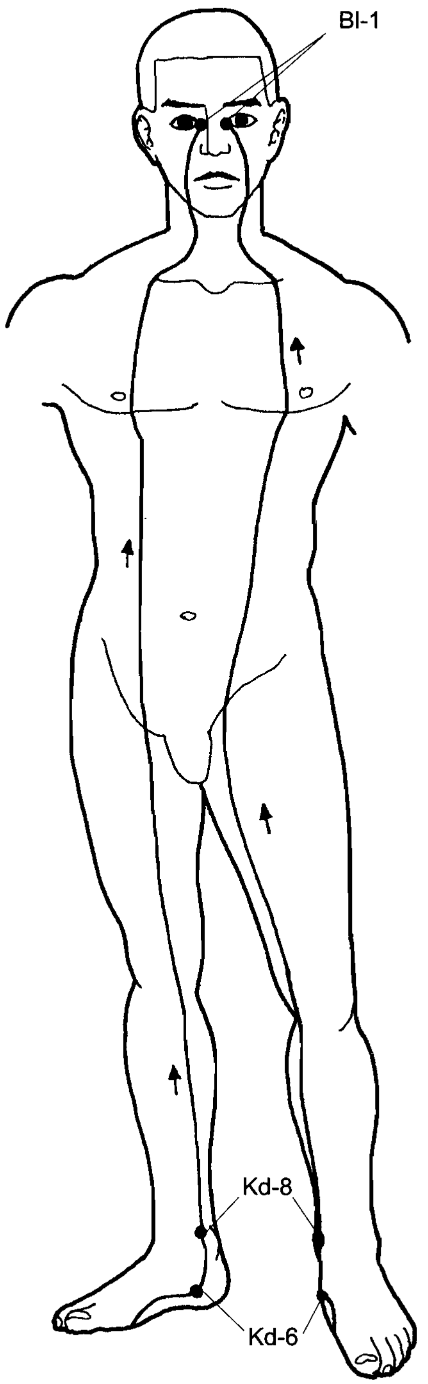

Figure 7.9. The Yin Heel Vessels

corners of the eyes, where they join the Yang Heel Vessels (Figure 7.9).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yin Heel Vessels are:

1. To treat symptoms of Excess conditions in the Lower Burner in women (e.g., abdominal distension, difficult delivery, or the retention of the placenta, abdominal masses, lumps, and fibroids);
2. To structurally balance the left and right sides of the body;
3. To treat disturbances of sleep (e.g., insomnia or somnolence); and
4. To treat certain cases of atrophy (e.g., when the muscles of the inner aspect of the leg are loose and the outer leg muscles are tight).

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess of the Yin Heel Vessels include: lower abdominal pain, vomiting, difficult bowel movements, and spasms on the medial side of the legs. The Yin Heel Vessels also control diseases of the eyes (watery eyes, heavy sensations of the eyelids or an inability to open the eyes), migraines, congestive headaches, and hypersomnia. Abnormalities of the Yin Heel Vessels cause pregnant women to have difficult labor.
2. Symptoms of Deficiency include: Aggravations of symptoms during the evening time, nocturnal headaches, insomnia, cramps, or convulsions.

THE YANG HEEL VESSELS

The Yang Heel Vessels control the Yang of the left and right sides of the body. Abnormalities of the Yang Heel Vessels in newborns cause vomiting of milk.

When Yin energy is slowed down in the Heel Vessels, the Yang energy moves more rapidly. The Excess Yang causes the following problems: insomnia, difficulty in closing the eyes, painful eyes, hypertension, stiff back and waist (inability to bend down), thigh tumors, bad colds, spontaneous sweating, headaches, paralysis of the arms and legs, deafness, epilepsy, nose bleeding, swelling of the body, pain in the joints, and head sweating. Excess Yang can also cause the vomiting of milk in infants.

Weakness in the Yang Heel Vessels can cause foot and ankle eversion, while Excess can cause foot inversion. Also, when the Yang Heel Vessel is in Excess, the inner leg muscles are loose and the outer leg muscles are tight.

The Yang Heel Vessels begin on the outer side of the ankles below the lateral malleolus, at the Bl-62 points. From there, they continue to ascend the outside of the thigh and the back to end just below the eyes at the Bl-1 points, where they communicate with the Yin Heel Vessels near the eyes (Figure 7.10).

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yang Heel Vessel are:

1. To treat acute Excess conditions of the lower back (e.g., aches due to spasm or invasion of Cold) and pain along the Bladder Channels of the legs;
2. To subdue Internal or External Wind from the head (e.g., facial paralysis, severe dizziness, and aphasia); and
3. To purge Wind-Heat and Wind-Cold manifesting in symptoms, such as sneezing, head ache, runny nose, and stiff neck.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include: Stiffness of the back and waist, lumbar pain, spasms on the outer side of the legs, tumors of the thighs, nocturnal epileptic seizures, and insomnia or restless sleep. The Yang Heel Vessels are also involved with diseases of the eyes such as dry and itchy eyes.
2. Symptoms of Deficiency include: Fatigue, lassitude and weakness during the day, and aggravations of symptoms from stress during the day. These symptoms improve as the night progresses.

According to some Traditional Chinese Medical Classics, if epilepsy occurs during the daytime, Qigong and moxa are given on the Yang Heel Vessels; however, if epilepsy occurs at night, Qigong and moxa are given on the Yin Heel Vessels.

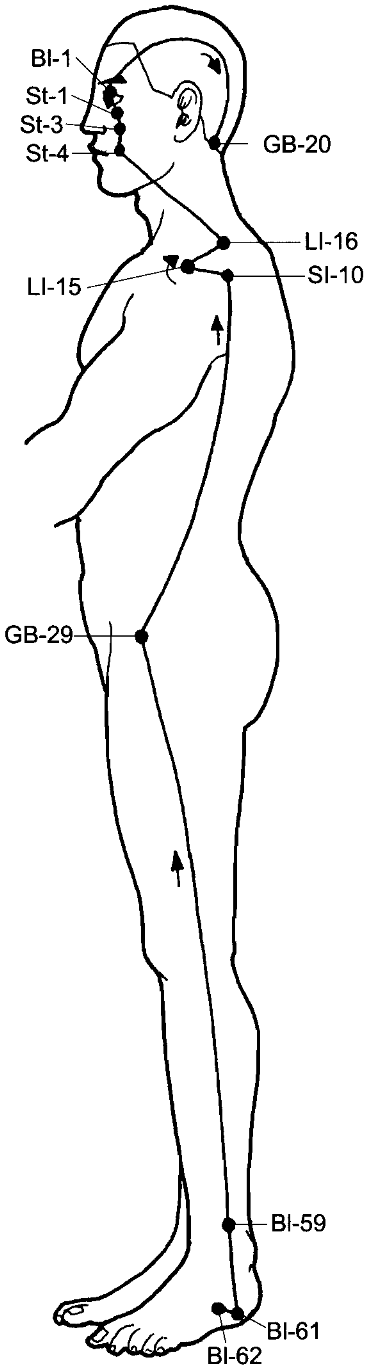

Figure 7.10. Side View of the Left Yang Heel Vessel

The Yin and Yang Linking Vessels or Wei Vessels

The word Wei translates as "to link" or "bind" and refers to a rope that is tied around something, pulling it down and securing it. The Linking Vessels are sometimes called the Regulator Channels. The Linking Vessels are divided into Yin and Yang energetic pathways.

Both Yin and Yang Linking Vessels start at the lower legs and flow upwards to the head, along the medial and lateral aspects of the lower legs and torso. When the Linking Vessels reach the neck and back of the head, they join the Conception and Governing Vessels. Secondary Vessels, called the Yu (surplus) Vessels, branch away from the energetic flow of the major Vessel, connecting the energetic flow of each Yin and Yang Linking Vessel to the hands.

Instead of serving as streams transporting Qi and Blood, these two vessels act as lakes that store Qi and Blood that overflows from other vessels. Together, these four vessels regulate the circulation of Qi and Blood for the whole body, store the overflowing Qi and Blood, and release the Qi and Blood into the channels in the event of insufficiencies.

The Yin Linking Vessel

The Yin Linking Vessels lie on the medial axis of the body. They help maintain the connection to all the Yin Channels. They begin at the Kd-9 points at the inner side of the lower legs and ascend up the sides of the abdomen and chest. They move toward the front of the body, along the sides of the chest to the tips of the nipples. At the Ruzhong (St-17) Breast Center points they split into two sets of branches. One set of branches connects with the Conception Vessel at the neck. The other set of branches becomes the Yin Yu Vessels, and extends from the chest, over the shoulders and down the inside of the arms following the route of the Pericardium Channels. Just above both wrist folds on each arm, each branch energetically pools at the Neiguan (Pc-6) Inner Pass points, before it flows into the palms. The Pc-6 points are therefore considered the Master Points for the Yin Linking Vessels (Figure 7.11).

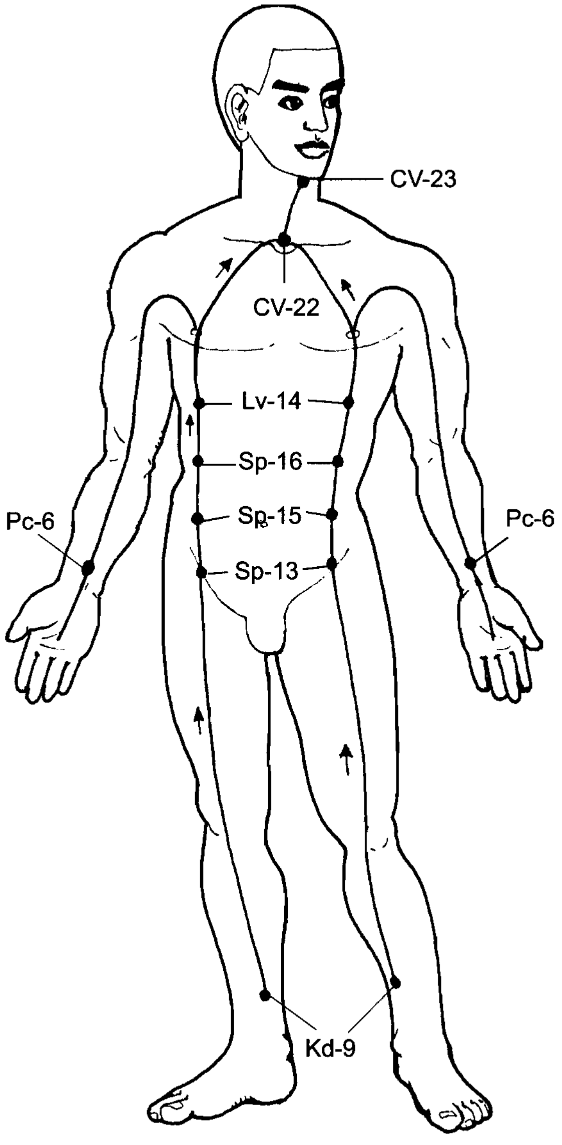

Figure 7.11. The Yin Linking Vessels

The Yin Linking Vessels are responsible for moving the Yin energy and regulating the Blood and inner parts of the body. They connect with all the body's Primary Yin Channels: Liver, Heart, Spleen, Lung, Kidney, and Pericardium. When the Yin energy is slowed down in the Linking Vessels, the Yang energy moves more rapidly. Therefore, if the Yin Linking Vessels become unbalanced, the Excess Yang condition can cause the patient to suffer from Heart pain.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yin Linking Vessels are:

1. To tonify the Heart (for symptoms of pain, stiffness, tightness and oppression in the chest, as well as mental depression, anxiety, apprehension, and nightmares);
2. To treat Deficient Yin and Blood conditions, especially if accompanied by psychological conditions, such as mental restlessness, anxiety, and insomnia; and
3. To treat headaches located in the back of the neck due to Blood Deficiency.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include diseases of the Heart such as hypertension, delirium, nightmares, etc., as well as, cardialgia (tightness and oppression in the chest), difficulty in swallowing, convulsive diseases, contracted feeling in the Lungs, and dyspnea. Diseases of the Yin Linking Vessels also include prolapse of the rectum and diarrhea.
2. Symptoms of Deficiency include timidity or fear, apprehension, nervous laughter, mental depression, hypotension, and weak respiration.

THE YANG LINKING VESSELS

The Yang Linking Vessels lie on the lateral aspects of the body. They serve to maintain and communicate with all the Yang Channels in the exterior portion of the lateral aspects of the body. They begin at the external part of the ankles, just below the lateral malleolus at Bl-63 points, ascend the sides of each leg, up each side of the back of the body, through the shoulder areas to the Naoshu (SI-10) Scapula's Hollow points. The SI-10 points

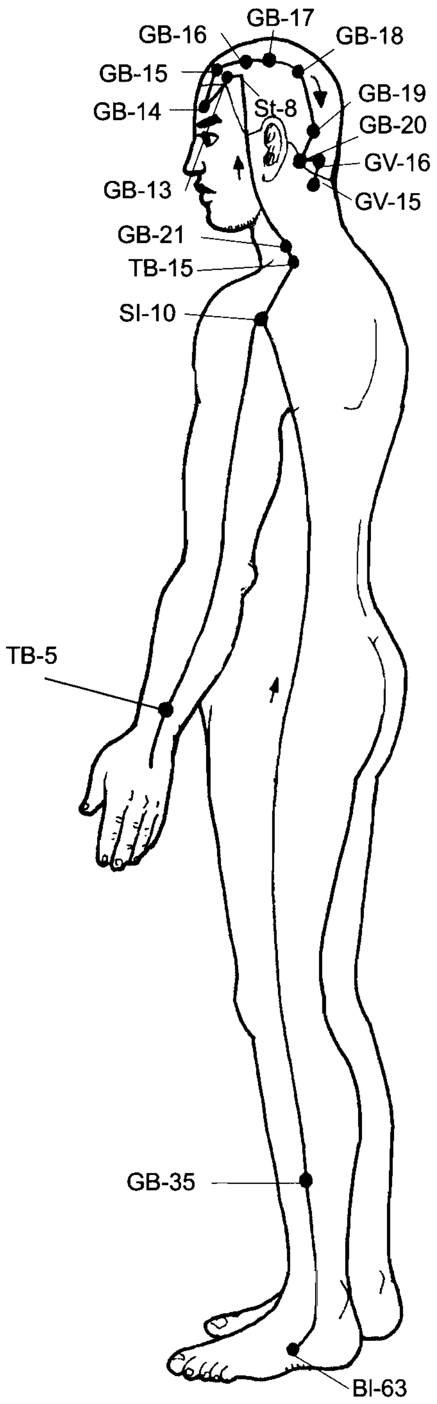

Figure 7.12. Side View of the Left Yang Linking Vessel

are the intersecting points of the Yang Linking and Yang Heel Vessels located on the Small Intestine Channels. From the SI-10 points, both Yang Linking Vessels split into two sets of branches. One set of branches flows up the sides of the neck and head connecting with the GB-20 points (at the lower occipital part of the head), where it communicates with the Governing Vessel. The other set of branches becomes the Yang Yu Vessels. The Yang Yu Vessels extend from the shoulders, down the back sides of the arms, following the route of the Triple Burner Channels. Just above both wrists on the outside of each arm, each of the branches pools at the Weiguan (TB-5) Outer Pass points, before flowing into the back of the hands. The TB-5 points are therefore considered the Master Points for the Yang Linking Vessels (Figure 7.12).

The Yang Linking Vessels are responsible for moving the Yang energy and controlling the Protective Qi, regulating the Wei Qi's resistance to external infections, and regulating the external parts of the body. They connect with all of the body's Primary Yang Channels: Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder, and Triple Burners. When the Yang energy is slowed down in the Linking Vessels, the Yin energy moves more rapidly; therefore, if the Yang Linking Vessels become unbalanced, the Excess Yin condition may cause the patient to catch colds and fevers more easily.

CLINICAL MANIFESTATIONS

The functions and clinical uses of the patient's Yang Linking Vessels are:

1. To treat ear problems due to the rising of Liver Fire, resulting in conditions such as tinnitus and deafness;
2. To treat ear diseases due to a Gall Bladder disharmony;
3. To treat hypochondriac pain;
4. To treat sciatic pain in the lateral aspects of the legs (along the Gall Bladder Channels);
5. To treat intermittent fevers that alternate between chills and fever.

PATHOLOGICAL MANIFESTATIONS

1. Symptoms of Excess include: Alternating chills and fever; pain on the lateral sides of the neck, trunk and legs, as well as pain in the head. Symptoms also include pains and skin problems during weather changes, sensitivity to changes in climate (aching muscles, skin rashes, etc.), swelling, pain and fever in the joints, diarrhea and night sweats.
2. Symptoms of Deficiency include: Coldness and lack of body heat, loss of energy, and physical strength. Symptoms also include cold knees, stiffness, and fatigue (especially during cold or rainy weather).

SUMMARY OF EIGHT EXTRAORDINARY VESSEL PATHOLOGY

The Eight Extraordinary Vessels link all of the Yin and Yang Channels in the body and regulate the flow of energy in these channels to maintain a state of energetic balance. The clinical significance of the Eight Extraordinary Vessels is manifested through certain pathological indications that are particular to the vessels' intersection with their specific Primary Channel(s). Their symptomatology is therefore not distinct from, but rather a composite of, the pathological symptoms associated with their joining Primary Channel(s).

Figure 7.13 describes some of the Eight Extraordinary Vessels pathologies and the clinical manifestations indicated along their root (beginning) and ending branches.

The Eight Extraordinary Vessels' energetic pathways used in certain esoteric Qigong meditations are somewhat different from the ones seen in Traditional Chinese Medical texts and acupuncture charts. The reason for this uniqueness is because their purposes are different.

The goal of Chinese acupuncture is to restore sick people to health. The energetic points being treated by an acupuncturist must be along the superficial channels, so that they can be activated by acupuncture needles.

Medical Qigong exercises and meditations aim to maximizing health, and to take the individual beyond mere physical health to spiritual enlightenment. In Medical Qigong training the channels and points can be deep within the body, since the energy is guided by the mind or by postures and movements, rather than by needles.

Eight Extraordinary Vessels	**Pathologies**
Belt	• Pain in the back, lumbar region, and sides of the navel • Weakness in the lower limbs
Conception	• Diseases of the reproductive and gastrointestinal systems (hemorrhoids, diarrhea, decreased urination, etc.) • In the male, sterility • In the female, the menstruation problems such as leukorrhea and dysmenorrhea; also breast pain, paralysis after delivery, emaciation, and sterility; all reproductive system problems, including internal and external genitalia (vulva, vagina, and cervix)
Governing	• Diseases of the head (apoplexy, aphasia, epilepsy, headaches, tetanus, etc.), back, neck, and Kidneys • Stiffness in the spinal column, spastic muscle movements of the extremities, night sweating, and circulatory disturbances around the anus
Thrusting	• In women, amenorrhoea, scanty periods or late periods, dysmenorrhoea, spontaneous abortion, inability to expel the placenta, menopause problems • Diseases of the Heart, fullness in the chest and abdomen, gastritis, abdominal pain, convulsive diseases • Brain dysfunction of physiological origin
Yang Heel	• In newborns, vomiting of milk • Stiffness of the back and waist, lumbar pain, spasms on the outer side of the legs, and tumors of the thighs • Diseases of the eyes
Yang Linking	• Diseases from Cold— fevers resulting in a sensitivity to changes in climate; cold knees; stiffness and fatigue; swelling, pain, and fever in the joints and extremities; and night sweating
Yin Heel	• Lower abdominal pain, vomiting, difficult bowel movements, and spasms on the medial side of the legs • Diseases of the eyes • Difficult labor in women
Yin Linking	• Diseases of the Heart (cardialgia), tightness of the chest, difficulty in swallowing, convulsive diseases • Diseases also include prolapse of the rectum and diarrhea

Figure 7.13. Pathologies of the Eight Extraordinary Vessels

THE FIFTEEN MAJOR COLLATERALS (CONNECTING VESSELS)

The Fifteen Collaterals are also known as the Fifteen Luo. Luo is translated to mean a net or web, and in Traditional Chinese Medicine it refers to the Connecting Vessels. These vessels are the major "passage ways" for the circulation of the body's channel energy, emerging out of the Luo (pathway) points on the Twelve Primary Channels, plus the Governing and Conception Vessels. The Luo points are located below the elbows and knees and provide an additional energetic barrier to keep Evil Winds from affecting the Primary Channels, being somewhat deeper then the Muscle/Tendon Channels. The Collaterals are the streams of energy that connect the paired Primary Channel rivers.

The Fifteen Collaterals are superficial and much thinner in formation than the Primary Channels running through the interior aspect of the body. The collaterals distribute Qi and Blood to those areas not directly traversed by other channels. Flowing out of each Primary Channel's Luo point are two Luo vessels, which are counted as one:

1. The Transverse Luo Vessels connect to the Source Points on the Yin and Yang coupled Primary Channels. They act as a safety valve to maintain balance between Yin and Yang channels by diverting Excess energy from one channel to tonify Deficiencies in the Orb (internal organ and energetic field) of their paired Primary Channel. When a channel is Deficient (Empty) and its paired channel is in Excess (Full), for example, the tonification of the Luo point on the Deficient channel is enough to replenish it, while normalizing the channel previously in Excess.
2. Longitudinal Luo Vessels flow out of the Luo points, but do not connect with the coupled Primary Channels. The Luo Vessels usually flow proximally toward the channel's organ.

The Luo vessels' main function is to transfer Qi and Blood from the Primary Channels to all parts of the body to nourish the tendons, the bones, the skin, and the five sense organs (nose, eyes, ears, lips, and tongue). The Luo vessels also link the body's interior with its exterior, connecting the internal and superficial channels (Figures 7.14 through 7.19).

The Fifteen Major Collaterals have smaller branches, known as the Minute and Superficial Collaterals, extending from them. The Superficial Collaterals are sub-branches similar to capillaries serving to transport Qi and Blood to the surface of the body. Both Minute and Superficial Collaterals, which can be seen beneath the surface as blood vessels, are called Blood Luo Vessels.

PATHOLOGY

The pathology of the Fifteen Major Collaterals are categorized in terms of Excess and Deficient syndromes.

1. Excessive Conditions of the collaterals are due to exogenous invasion if (1) the organ associated with the Luo is in Excess or (2) the body's Wei Qi is weak. Evil Winds enter the body through Jing Well or Wind points, and start moving up the channel. The Longitudinal Luo Vessels provide a route for the diversion of Evil Winds. The Longitudinal Luo Vessels have more Wei Qi then the Primary Channels and can better fight pathogens. Sometimes with a Wind Cold invasion, a blue color is visible along the path of the Longitudinal Luo Vessel. If the invasion is due to Wind Heat, there may be a red color along the vessel.
2. Deficient Conditions of the collaterals are due to: (1) the organ associated with the Luo being Deficient or (2) the patient's Qi is Deficient due to exogenous factors.

POINT LOCATION OF THE FIFTEEN MAJOR COLLATERALS

1. The Foot Taiyin-Spleen 4 Point is located on the Spleen Channel of each foot, on the medial side, just posterior to the base of the first metatarsal bone. This Lou's energy flows downward to connect with the Stomach Channel on each foot. A second branch flows upward along the medial aspect of the inner thigh, passing the abdomen and connecting with the Stomach and Small Intestines (Figure 7.14).

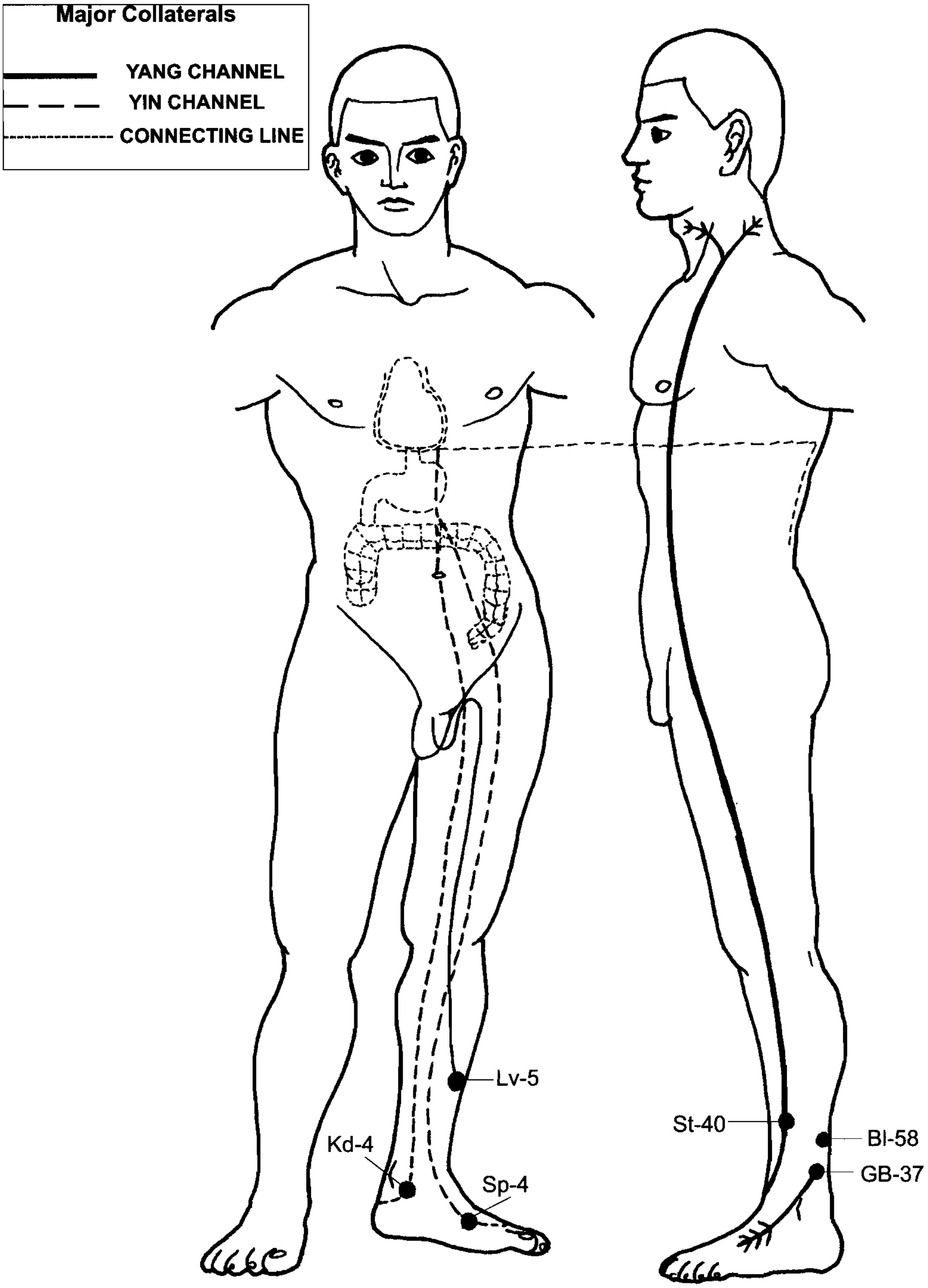
Major Collaterals
YANG CHANNEL
YIN CHANNEL
CONNECTING LINE
Lv-5
Kd-4
Sp-4
St-40
Bl-58
GB-37

Figure 7.14. The figure at left shows the three Luo points of the Yin Collaterals of the left foot located on the Primary Channels of the Kidneys, Spleen, and Liver. The figure on the right shows the three Luo points of the Yang Collaterals of the left foot located on the Primary Channels of the Stomach, Bladder, and Gall Bladder.

   a. Symptoms of Excess include sharp intestinal pain, vomiting, and diarrhea.
   b. Symptoms of Deficiency include abdominal swelling.
2. The Foot Shaoyin-Kidney 4 Point is located on the Kidney Channel of each foot, just posterior to the medial malleolus. The Lou's energetic flow runs downward, crossing the heel, and connects with the Bladder Channel on each foot. A second branch flows upward along the medial aspect of the inner thigh, following the Kidney Channel. It passes the abdomen and connects with the pericardium of the Heart then flows laterally to connect with the lumbar vertebrae (Figure 7.14).
   a. Symptoms of Excess include enuresis, emotional irritability, and depression.
   b. Symptoms of Deficiency include lower back pain.
3. The Foot Jueyin-Liver 5 Point is located on the Liver Channel of each foot, on the medial side, several inches above the medial malleolus. The Lou's energetic flow connects with the Gall Bladder Channel on each foot and flows upward along the medial aspect of the inner thigh. It connects with the genitals and terminates at the penis in males, and the clitoris in women (Figure 7.14).
   a. Symptoms of Excess include swelling of the testicles.
   b. Symptoms of Deficiency include itching in the pubic region.
4. The Foot Taiyang-Bladder 58 Point is located on the Bladder Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Kidney Channel on each foot (Figure 7.14).
   a. Symptoms of Excess include nasal congestion, headache, and back pain.
   b. Symptoms of Deficiency include clear mucus nasal discharge and nosebleed.
5. The Foot Shaoyang-Gall Bladder 37 Point is located on the Gall Bladder Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Liver Channel on each foot, then continues downward to disperse over the dorsum on each foot (Figure 7.14).
   a. Symptoms of Excess include fainting.
   b. Symptoms of Deficiency include weak and flaccid muscles of the feet.
6. The Foot Yangming-Stomach 40 Point is located on the Stomach Channel of each foot, on the lateral side, several inches above the external malleolus. The Lou's energy flows downward to connect with the Spleen Channel on each foot. A second branch flows along the lateral aspect of the tibia, upward to the top of the head where it divides, with one branch converging with the other Yang Channels on the neck and head, and the other branch connecting with the throat (Figure 7.14).
   a. Symptoms of Excess include epilepsy and insanity.
   b. Symptoms of Deficiency include pharyngitis, sudden aphasia, and flaccid or atrophied muscles in the legs or feet.
7. The Hand Taiyang-Small Intestine 7 Point is located on the Small Intestine Channel of each hand, on the dorsal side, several inches above the wrist, on the ulnar side. The Lou's energy flows upward past the elbow and connects with the Large Intestine 15 point (Figure 7.15).
   a. Symptoms of Excess include fever, headaches, and blurred vision.
   b. Symptoms of Deficiency include atrophy of the muscles in the elbow and arm, and a looseness in the joints.
8. The Hand Yangming-Large Intestine 6 Point is located on the Large Intestine Channel of each hand, several inches above the dorsal side of the wrist on the radial side. The Lou's energy flows upward on each arm to the jaw and pours into the area of the teeth. Another branch ascends into each ear, connecting with the Thrusting Vessels which supply energy to the head (Figure 7.15).
   a. Symptoms of Excess include deafness and toothache.

   b. Symptoms of Deficiency include a sensation of coldness in the teeth, as well as fullness and congestion in the chest.

9. The Hand Shaoyang-Triple Burner 5 Point is located on the Triple Burner Channel of each hand, located just above the dorsal transverse crease of the wrist. The Lou's energy flows upward past the arm and over the shoulder, dispersing into the chest and connecting with the Pericardium Channel (Figure 7.15).
   a. Symptoms of Excess include muscle spasms of the elbow.
   b. Symptoms of Deficiency include flaccid muscles in the arm and elbow joint.
10. The Hand Taiyin-Lung 7 Point is located on the Lung Channel of each hand and arises from the cleft of the tendons and bones on the radial side of the wrist, where it flows down into the palm, spreading through the thenar eminence (Figure 7.15).
    a. Symptoms of Excess include hot palms or wrists.
    b. Symptoms of Deficiency include enuresis and shortness of breath.
11. The Hand Shaoyin-Heart 5 Point is located on the Heart Channel of each hand, just above the transverse crease of the wrist. The Lou ascends along the Heart Channel and enters the Heart then continues up the chest into the head, flowing into the root of the tongue, then ascends to connect with each eye (Figure 7.16).
    a. Symptoms of Excess include fullness and pressure in the chest.
    b. Symptoms of Deficiency include aphasia.
12. The Hand Jueyin-Pericardium 6 Point is located on the Pericardium Channel of each hand, just a few inches above the medial transverse crease of the wrist, between the two tendons. The Lou's energy flow follows the Pericardium Channel, connecting with the Heart (Figure 7.16).
    a. Symptoms of Excess include chest pain.
    b. Symptoms of Deficiency include irritability.
13. The Governing Vessel 1 Point is located on the

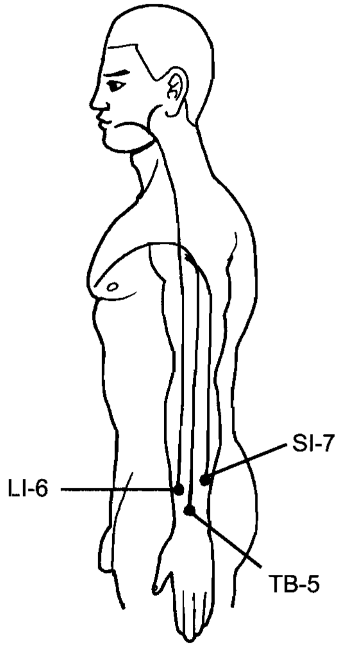

Figure 7.15. The three Yang Collaterals and their Luo points located on the hand and on the Primary Channels of the Large Intestine (LI), Triple Burners (TB), and Small Intestine (SI).

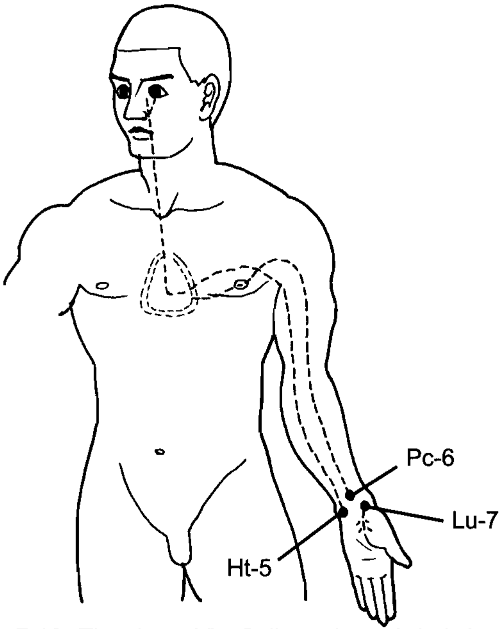

Figure 7.16. The three Yin Collaterals and their Luo points located on the hand and on the Primary Channels of the Heart, Pericardium, and Lungs.

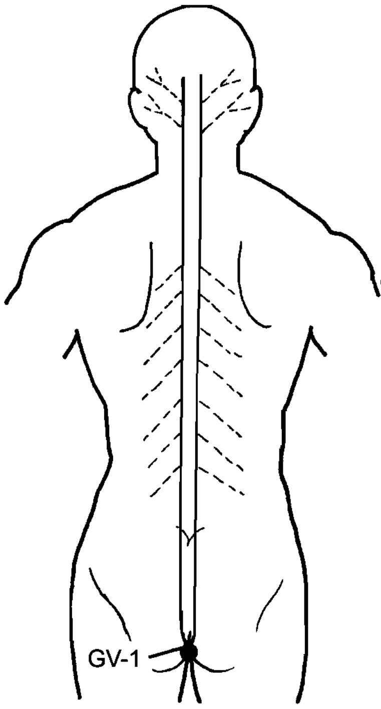

Figure 7.17. The main collateral and Luo point of the Governing Vessel is located on the Governing Vessel (GV) 1 point.

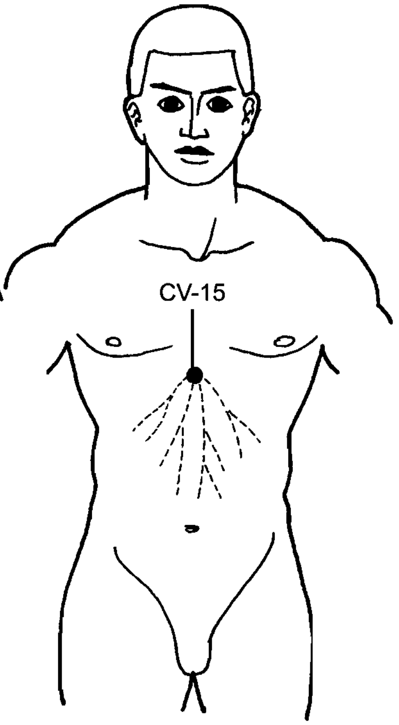

Figure 7.18. The main collateral and Luo point of the Conception Vessel is located on the Conception Vessel (CV) 15 point.

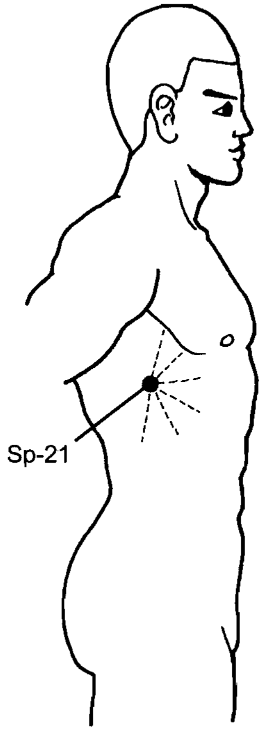

Figure 7.19. The major collateral and Luo point of the Spleen is located on the Spleen (Sp) 21 point.

base of the Governing Vessel, at the perineum. The Lou's energy flows upward along both sides of the spine to the nape of the neck, spreading to the top of the head. It continues flowing into the scalp regions on both sides of the head and connects with the Bladder Channel and merges with the spine (Figure 7.17).

   a. Symptoms of Excess include stiffness along the spine.
   b. Symptoms of Deficiency include dizziness or heaviness in the head.

14. The Conception Vessel 15 Point is located on the Conception Vessel of the chest, just below the xiphoid process. The Lou's energy flows downward, pouring over the abdomen (Figure 7.18).
    a. Symptoms of Excess include pain on the surface skin of the abdomen.
    b. Symptoms of Deficiency include itching on the surface skin of the abdomen.
15. The Major Luo of the Spleen 21 Point is located on the Spleen Channel of the chest, just below the auxiliary fold of each arm. The Lou's energy flow spreads through the chest and hypochondriac region, gathering the Blood like a net (Figure 7.19).
    a. Symptoms of Excess include general aches and pains throughout the entire body.
    b. Symptoms of Deficiency include weakness in the muscles of the limbs and joints.

THE TWELVE DIVERGENT CHANNELS

The Twelve Divergent Channels comprise an important part of the body's channel system. These channels are branches off the Twelve Primary Channels and share the energetic function of circulating Qi throughout the body. The energetic field of the Twelve Divergent Channels forms an enormous web of complex interconnections within the body's Twelve Primary Channels.

Because the area over which the energy of the Twelve Divergent Channels is distributed is quite extensive, they are considered a separate component of the channel system. They are secondary streams that parallel the primary rivers, yet each has its own functional characteristics and clinical application independent of the Primary Channels. The Twelve Divergent Channels run deeper in the body, supplementing the areas along the pathways that the Primary Channels do not reach. They have no points of their own, although there are intersection points where they cross the major channels (Figures 7.20 through 7.25).

One of the primary functions of the Twelve Divergent Channels is to integrate all parts of the body with the Twelve Primary Channels. There are areas in the body which are not traversed by the pathways of the Twelve Primary Channels, as well as internal organs that are otherwise unconnected, or only remotely connected, by the Primary Channels. These areas are more securely linked by the energetic flow of the Twelve Divergent Channels, which strengthen the bonds between the Twelve Primary Channels and the physical areas that are connected to, or adjoining, their pathways.

Another primary function of the Twelve Divergent Channels is to facilitate the connection between pairs of Yin and Yang Primary Channels and Organs. Therefore, all primary Yin and Yang organs are connected by the Divergent Channels. Both Yin and Yang Divergent Channels ultimately connect with the body's Yang Primary Channels.

1. The Yang Divergent Channels complete a cycle of leaving the primary channels (e.g., Primary Gall Bladder Channel) and entering their associated organs (e.g., The Gall Bladder Organ) before they resurface on the neck and reconnect with their original channels (e.g., The Primary Gall Bladder Channel).
2. The Yin Divergent Channels leave their primary channels (e.g., Primary Liver Channel), then join their associated Yang Divergent Channels (e.g., The Divergent Gall Bladder Channel), which then join with the Yang Primary Channels (e.g., The Primary Gall Bladder Channel).

Similarly, within the body's cavities, most of the Twelve Divergent Channels first join with their pertaining Yin or Yang organ and then connect with the associated organ (belonging to the associated channel) in the Yin / Yang pair. Through this interaction, the connection between paired Yin and Yang organs and channels is strengthened.

An understanding of the relationship between the paired Yin and Yang interactions is important to the Qigong doctor, as sometimes a disease affecting a Yang channel can be treated by selecting certain areas on its associated Yin channel, and vice versa. The same theory holds true for diseases of the internal organs.

The Twelve Divergent Channels are paired into six confluences according to their internal and external relationship.

1. Bladder Divergent Channel (Foot Tai Yang) and Kidney Divergent Channel (Foot Shao Yin)
2. Stomach Divergent Channel (Foot Yang Ming) and Spleen Divergent Channel (Foot Tai Yin)
3. Gall Bladder Divergent Channel (Foot Shao Yang) and Liver Divergent Channel (Foot Jue Yin)
4. Small Intestine Divergent Channel (Hand Tai Yang) and Heart Divergent Channel (Hand Shao Yin)
5. Large Intestine Divergent Channel (Hand Yang Ming) and Lung Divergent Channel (Hand Tai Yin)
6. Triple Burners Divergent Channel (Hand Shao Yang) and Pericardium Divergent Channel (Hand Jue Yin)

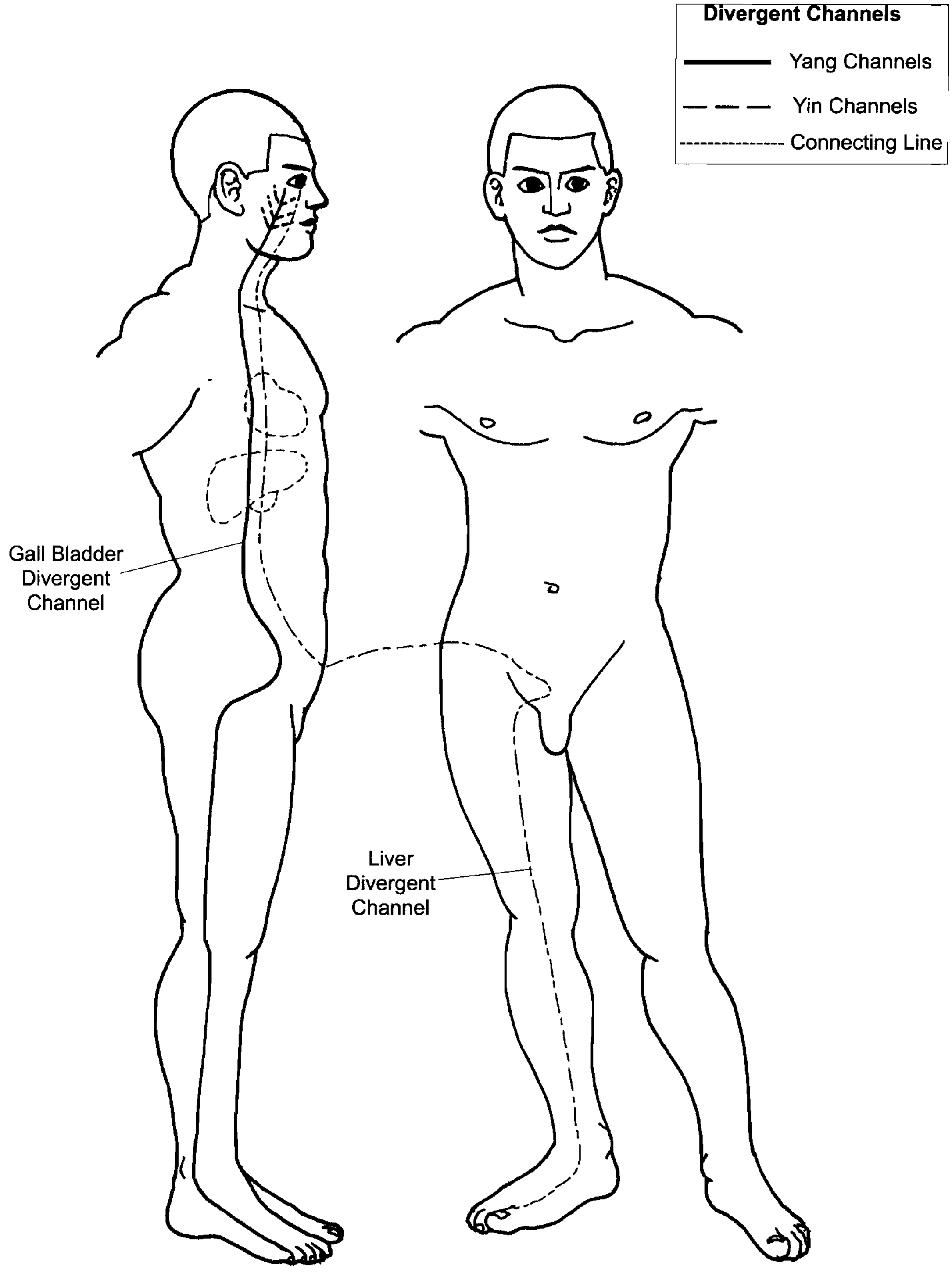
Divergent Channels
——— Yang Channels
— — — Yin Channels
-------- Connecting Line
Gall Bladder Divergent Channel
Liver Divergent Channel

Figure 7.20. The Divergent Channels of the Gall Bladder and Liver

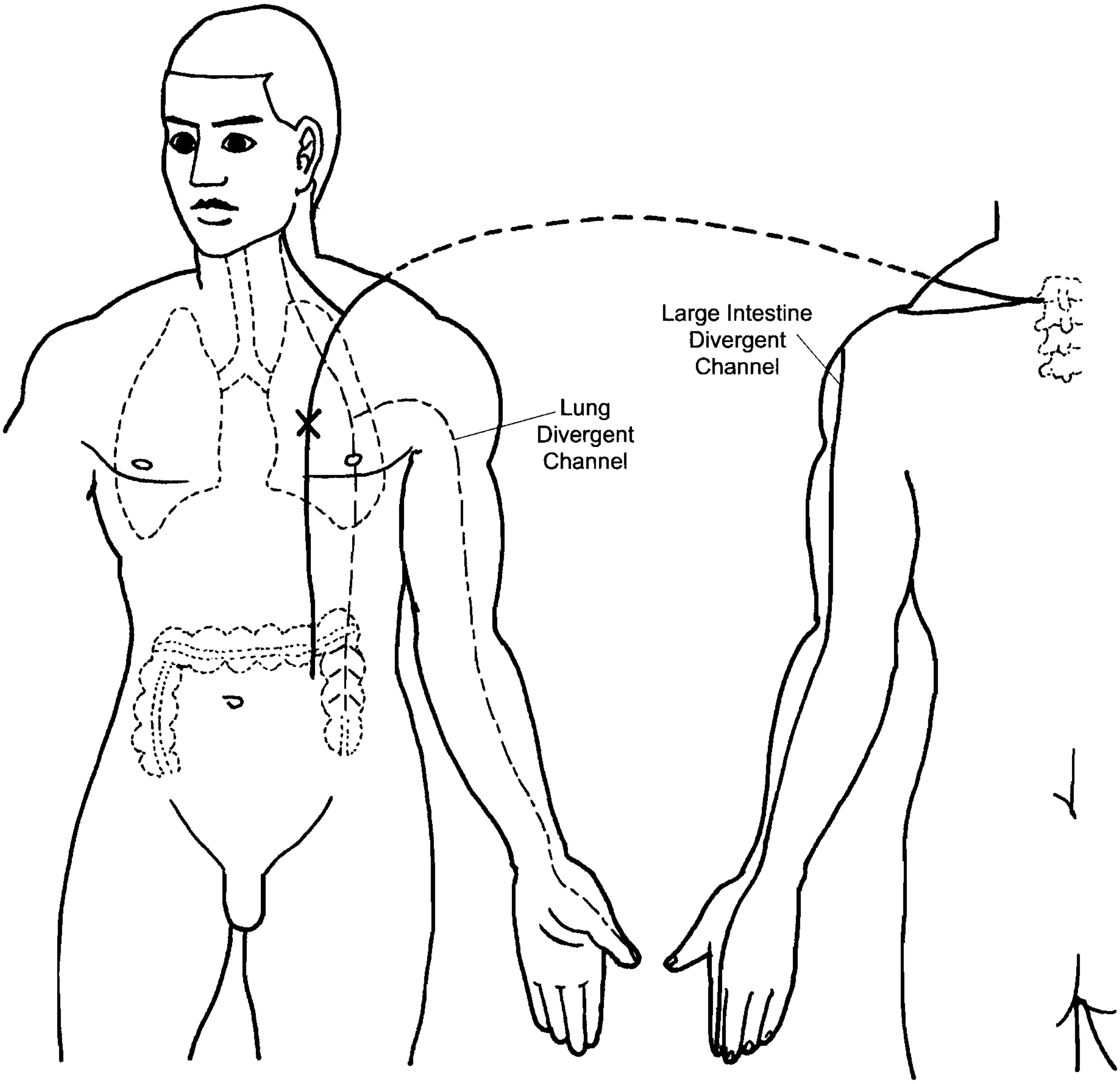
Large Intestine Divergent Channel
Lung Divergent Channel

Figure 7.21. The Divergent Channels of the Lungs and Large Intestine

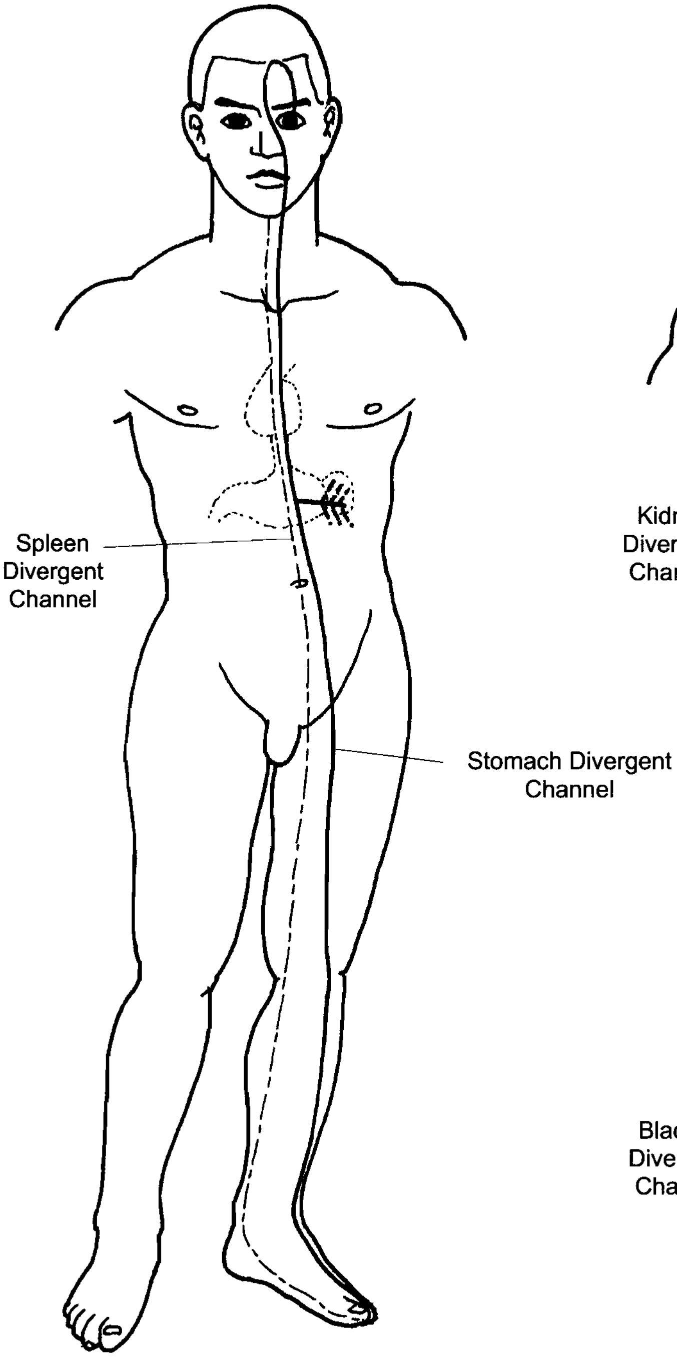
Spleen Divergent Channel
Stomach Divergent Channel

Figure 7.22. The Divergent Channels of the Stomach and Spleen

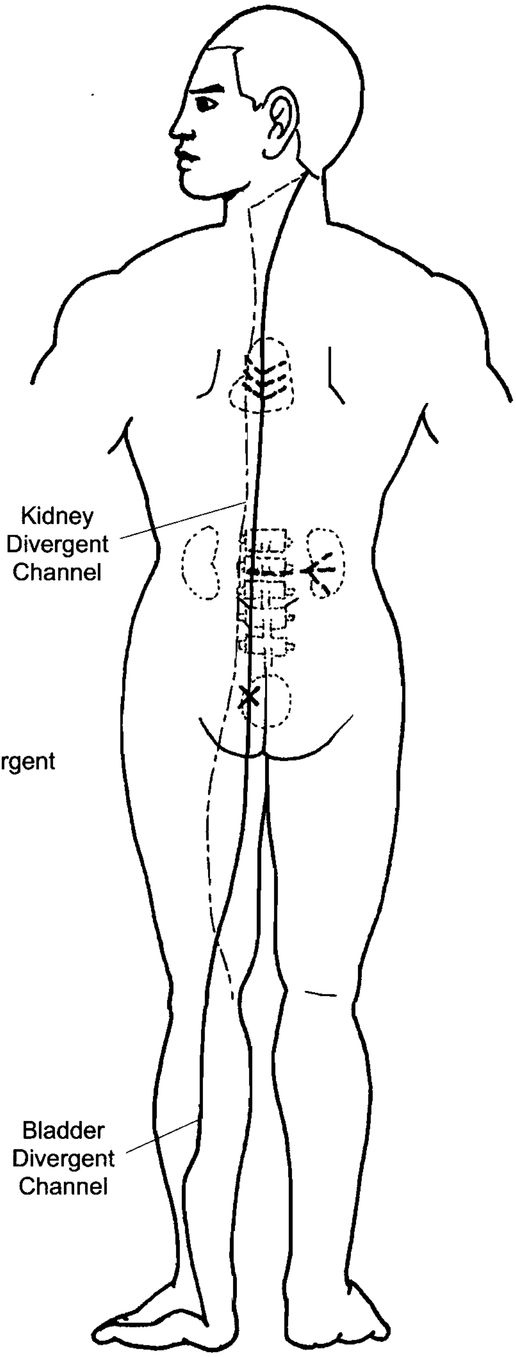
Kidney Divergent Channel
Bladder Divergent Channel

Figure 7.23. The Divergent Channels of the Bladder and Kidneys

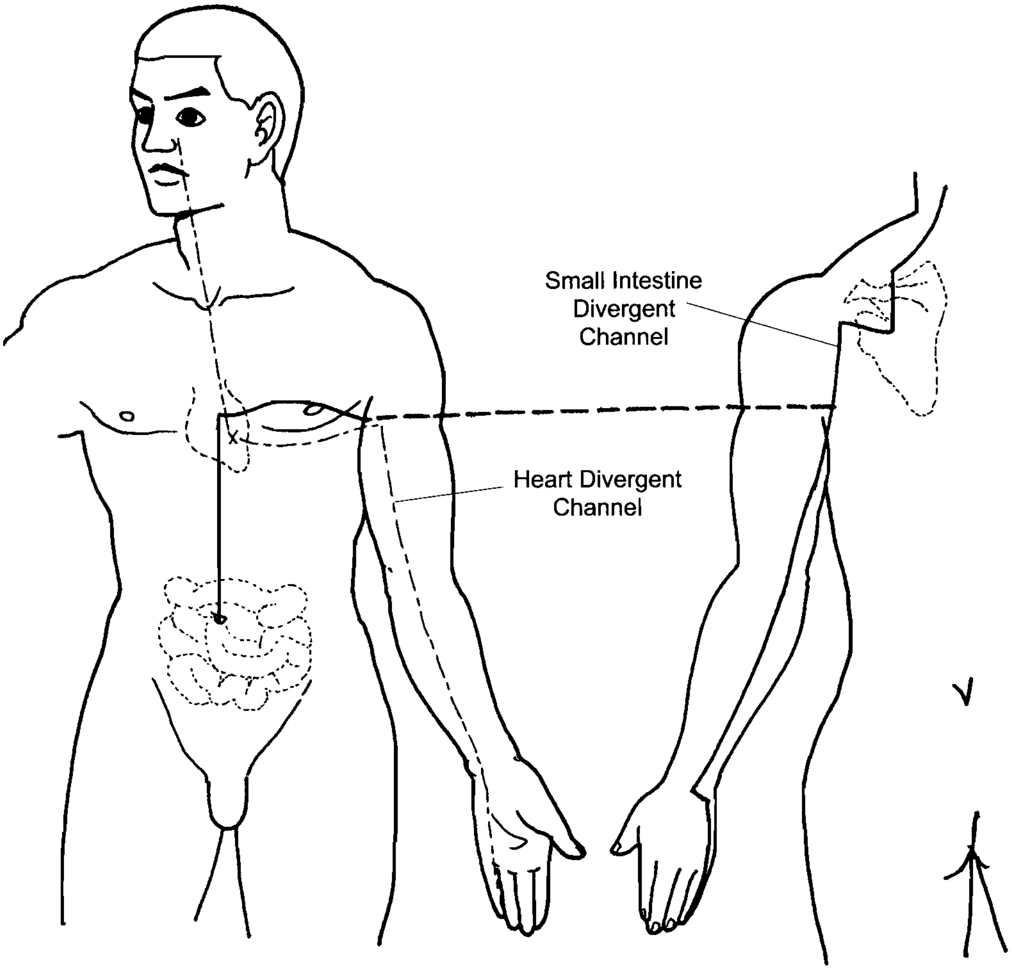
Small Intestine Divergent Channel
Heart Divergent Channel

Figure 7.24. The Divergent Channels of the Heart and Small Intestines

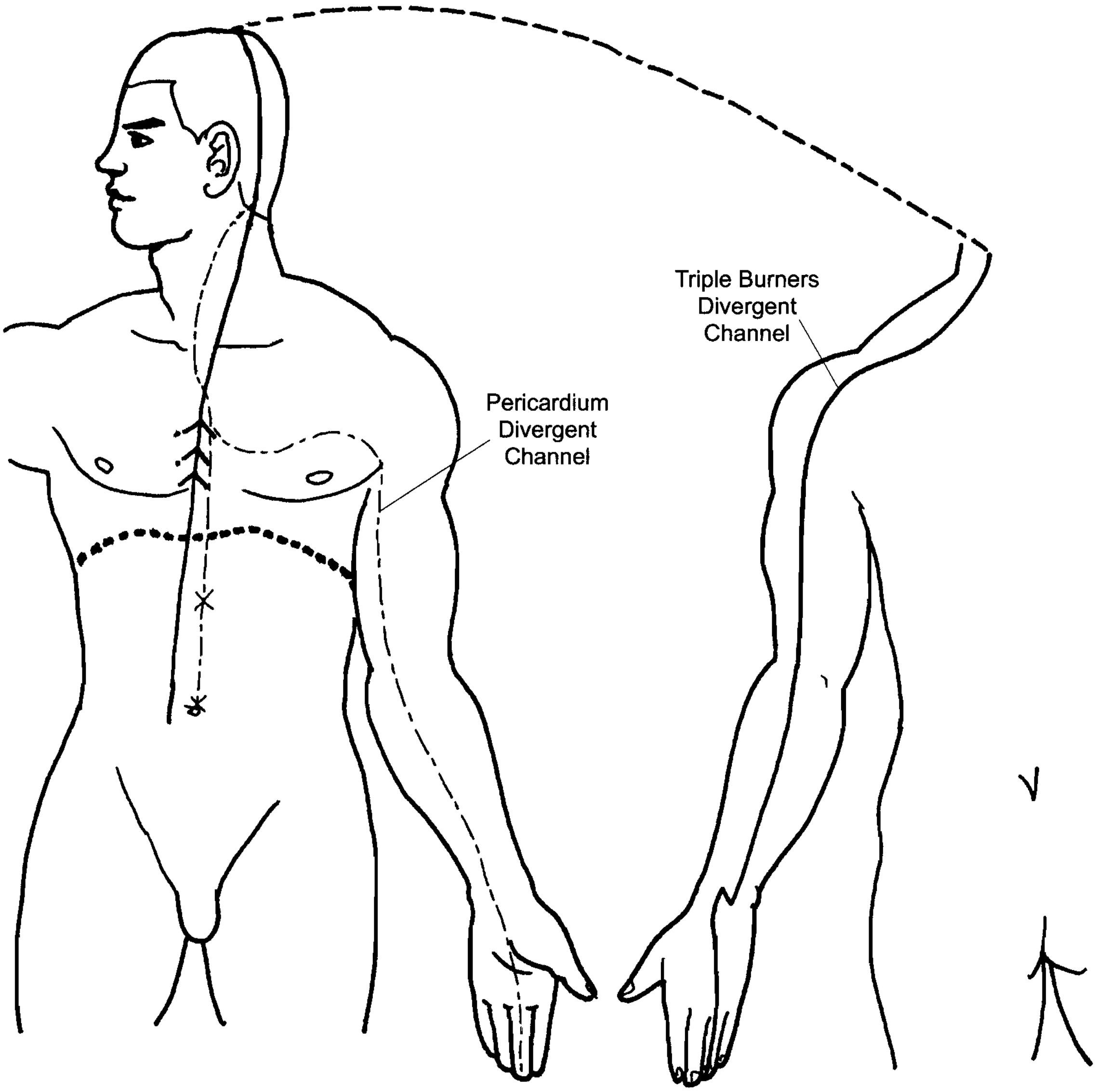
Triple Burners
Divergent
Channel
Pericardium
Divergent
Channel

Figure 7.25. The Divergent Channels of the Pericardium and Triple Burners

THE TWELVE MUSCLE/TENDON CHANNELS

The Twelve Muscle / Tendon Channels are the muscle and soft tissue regions of the body where the Qi and Blood of the Primary Channels nourish the skin, muscles, and tendons. They are very superficial in relationship to the Primary Channels and form a capillary network that travels in the depressions and planes between muscles and tendons. The Twelve Muscle / Tendon Channels, subsequently spread over the whole of the epidermis, through its close connection with the cutaneous tissues.

Originating in the extremities, these channels ascend to the head and torso. Physiologically, they do not enter into the internal organs but coordinate the movement of the bones and limbs. The Twelve Muscle / Tendon Channels are connected to the inner fascia of the body's muscles, tendons, and ligaments, as well as other connective tissues. They are affiliated with the channels and collaterals (Luo) systems on the exterior of the body, and serve as mediators between any energetic reactions (trauma, stress, etc.) that vibrate from the body's exterior surface to the deep internal organs.

These channels are found along the four extremities on the surface of the body, as well as along the head, neck, back, chest, and abdomen. The name of each Muscle / Tendon Channel comes from the Twelve Primary Channels whose external energy flow they follow. They also receive Blood and Qi nourishment for their functional activity from and through the Twelve Primary Channels. There are points of connection, as well as intersecting points along these channels (Figure 7.26 through 7.37).

Treatment at these conjunctive points consistently responds successfully to stimulation; therefore, the Qigong doctor can effectively apply Jing Point Therapy to specific areas and points along the Muscle / Tendon Channels.

These channels are responsible for extending and flexing the muscles, tendons, ligaments and joints; their pathology is reflected in symptoms of impaired movement (i.e., pulled, twisted, strained, cramped or atrophied muscles, spasms, etc.).

In Traditional Chinese Medicine the muscles and tendons are described in characteristics of Yin (flexion, contraction, internal rotation, etc.) and Yang (extension, expansion, external rotation, etc.) which balance each other. When the Yin and Yang action of the muscles and tendons fail to balance and regulate each other, Muscle Channel Dysfunction results; for example, when exposed to cold, the muscles / tendons become tense and over-contract; when exposed to heat, muscles / tendons become loose and overextend.

The pathology of the Muscle / Tendon Channels also reflects the symptoms of corresponding groups of muscles and other connective tissues. The connective tissues are divided into three groups: the large, the small, and the membranous connective tissue.

Therapeutically, a local Muscle / Tendon Channel symptom can be treated by stimulating an area located next to the origin of the pain; for example, if the area is Yang (lateral) and overactive, then the Yin (medial) will be underactive and vice versa. Treatment is directed towards restoring the balance between the Yin and Yang Muscle / Tendon Channels.

Being superficial, the Twelve Muscle / Tendon Channels contain Wei Qi; thus, they provide the body's third line of defence against any unfavorable exogenous influence (the first being the body's Wei Qi field, and second being the Wei Qi stored within the skin). Only after overcoming the resistance of the body's Muscle / Tendon Wei Qi can the Evil Exogenous pathogens travel down the length of the channel to penetrate the corresponding Primary Channel at the Jing-Well point. Jing-Well points are points of energetic union, connecting Primary Channel points together with the points where the Muscle / Tendon Channel have their origin.

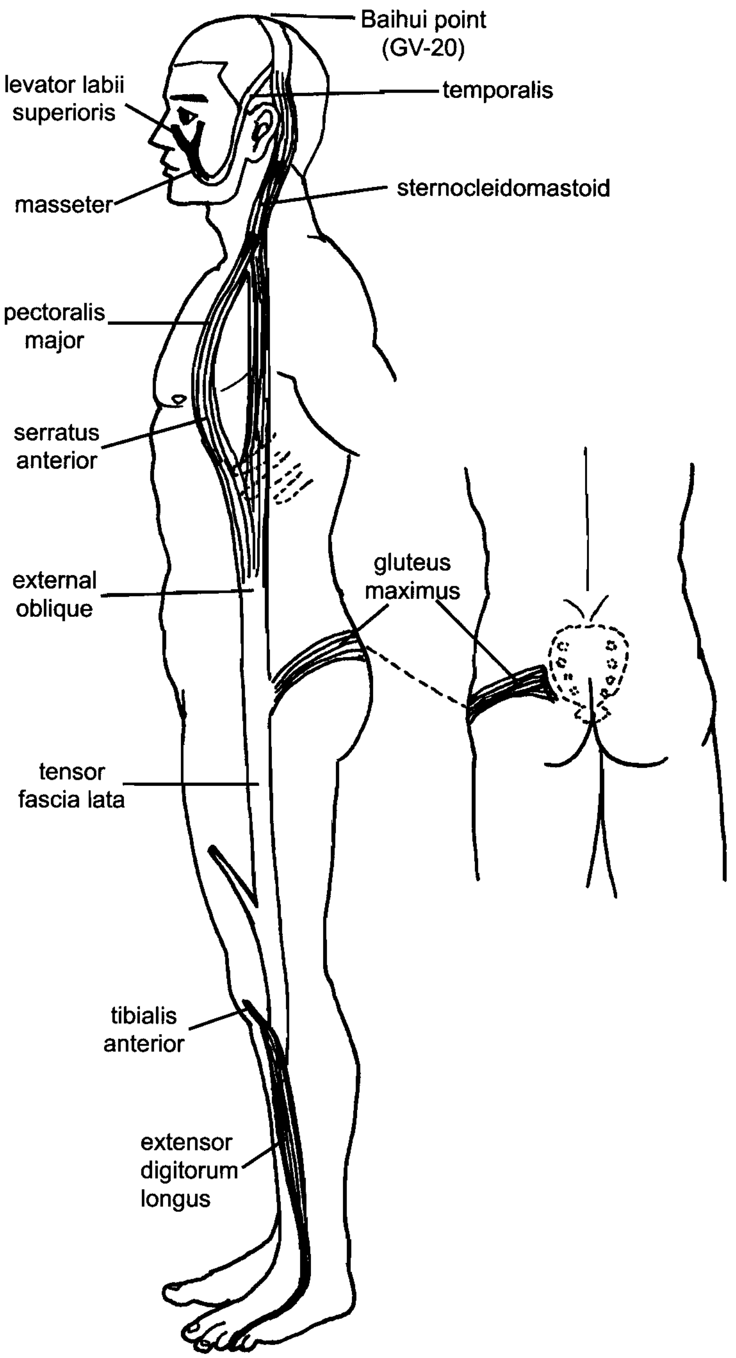
Baihui point (GV-20)
levator labii superioris
temporalis
masseter
sternocleidomastoid
pectoralis major
serratus anterior
external oblique
gluteus maximus
tensor fascia lata
tibialis anterior
extensor digitorum longus

Figure 7.26. The Muscle Region of the Gall Bladder Channel: High Tide is 11p.m.–1 a.m. Pathological symptoms include strained muscles from the fourth toe to the knee upon lateral rotation, with an inability to bend the knee; muscle spasms or stiffness within the popliteal fossa; strained muscles of the sacrum, pelvis, and lower ribs; pain in the hypochondria, chest, and clavicle region; and an inability to turn the eyes to the left or right.

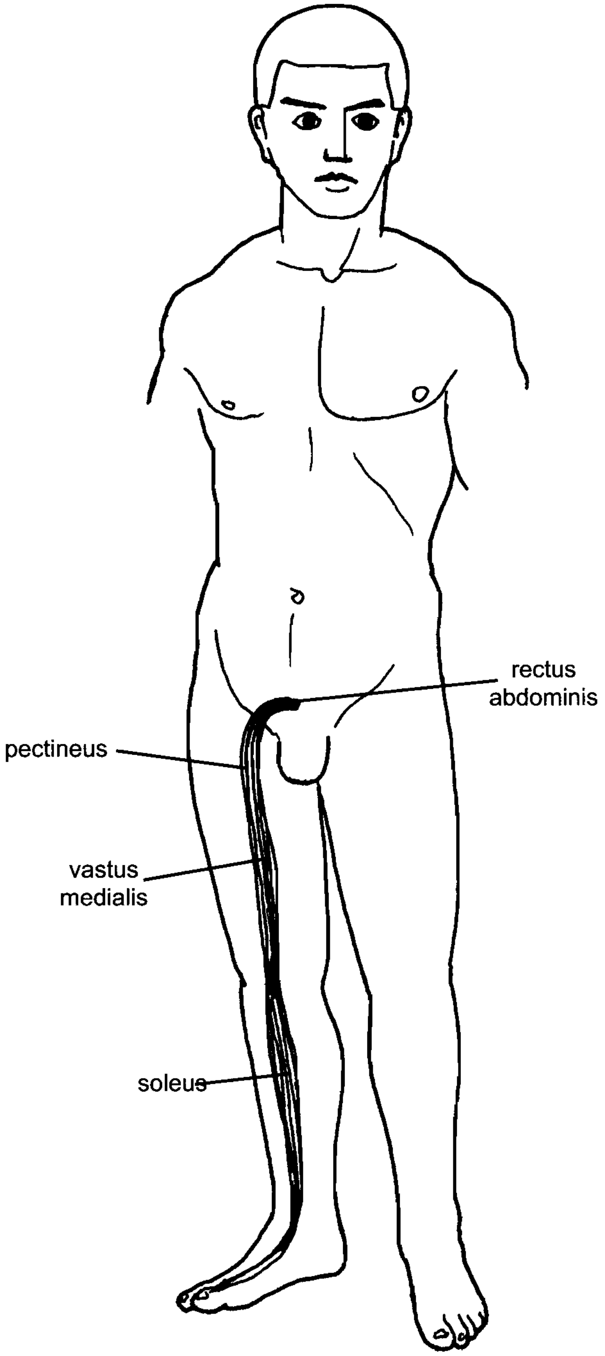
rectus
abdominis
pectineus
vastus
medialis
soleus

Figure 7.27. The Muscle Region of the Liver Channel: High Tide is 1 a.m.–3 a.m. Pathological symptoms include strained muscles of the big toe; pain in the anterior internal malleolus of the ankle; pain at the medial aspect of the knee and thigh; and dysfunction of the reproductive organs, i.e., impotence.

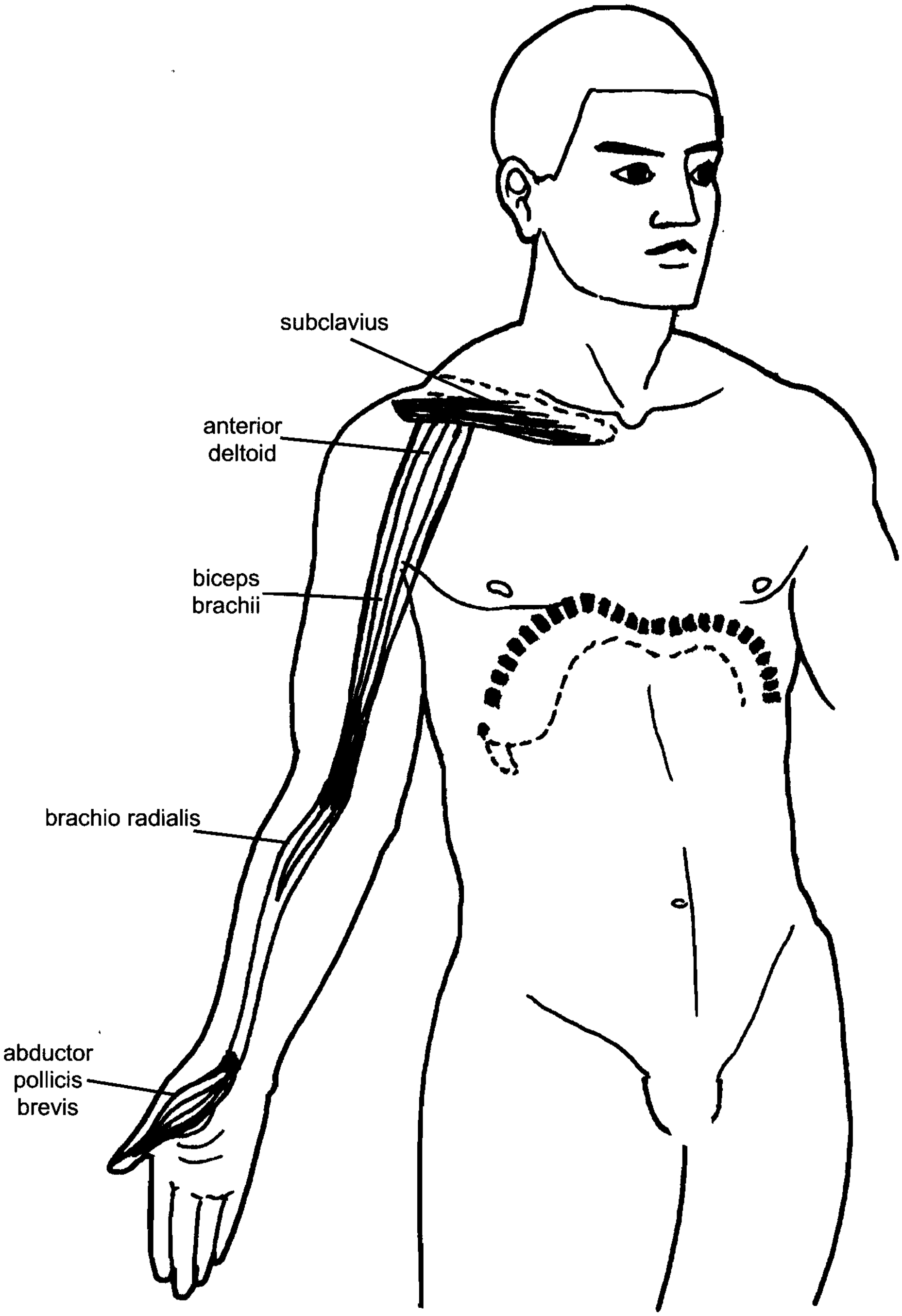
subclavius
anterior deltoid
biceps brachii
brachio radialis
abductor pollicis brevis

Figure 7.28. The Muscle Region of the Lung Channel: High Tide is 3 a.m.–5 a.m. Pathological symptoms include strained muscles of the thumb; stiff, strained or muscle spasms, and/or pain along the course of the Lung Channel. In more serious cases, there will be muscle spasms over the rib area and spitting of blood.

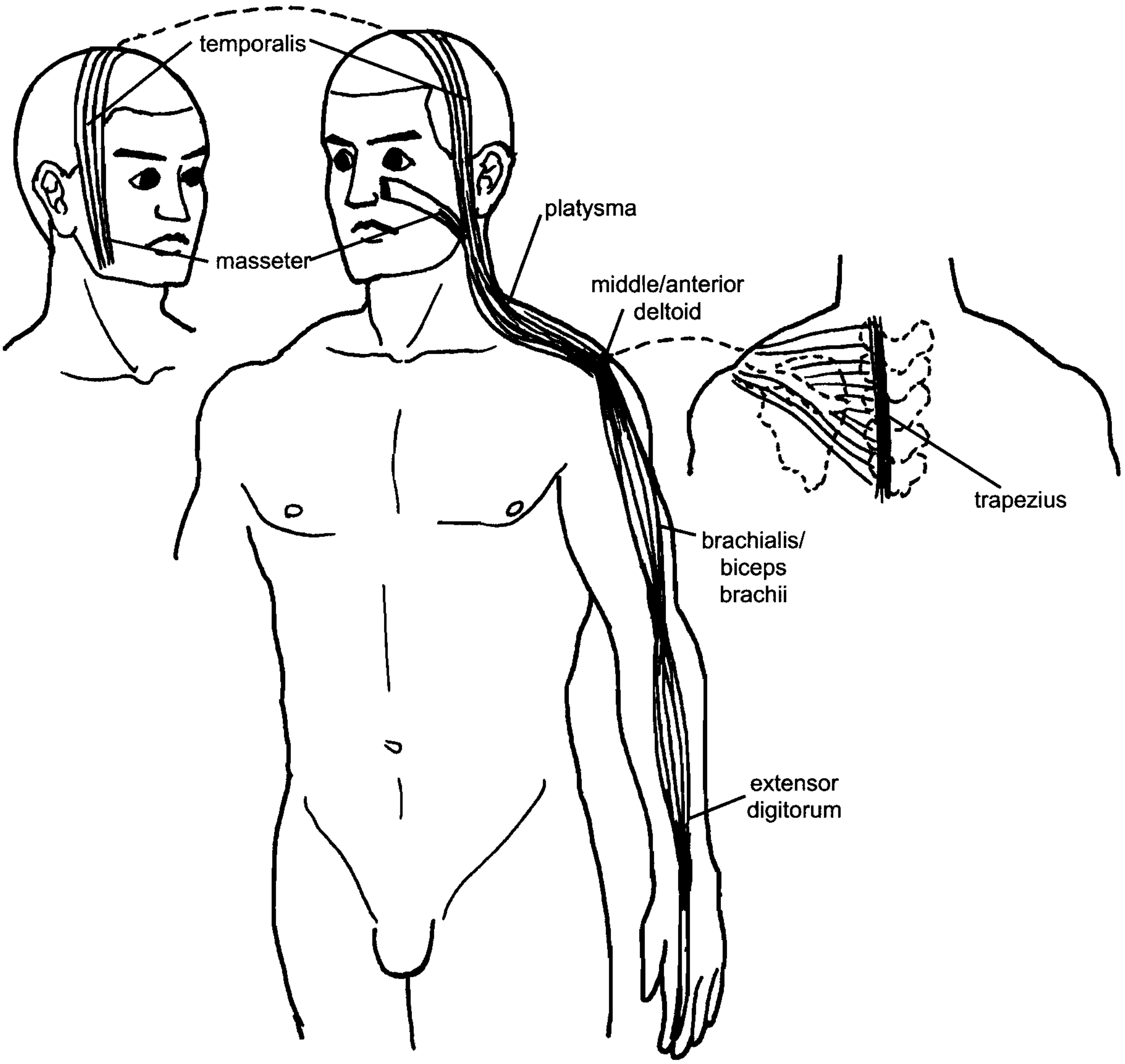
temporalis
masseter
platysma
middle/anterior deltoid
trapezius
brachialis/ biceps brachii
extensor digitorum

Figure 7.29. The Muscle Region of the Large Intestine Channel: High Tide is 5 a.m.–7 a.m. Pathological symptoms include strained muscles of the index finger; stiffness, strained, or muscle spasms along the course of the Large Intestine Channel, resulting in frozen shoulder; and an inability to rotate the neck from side to side.

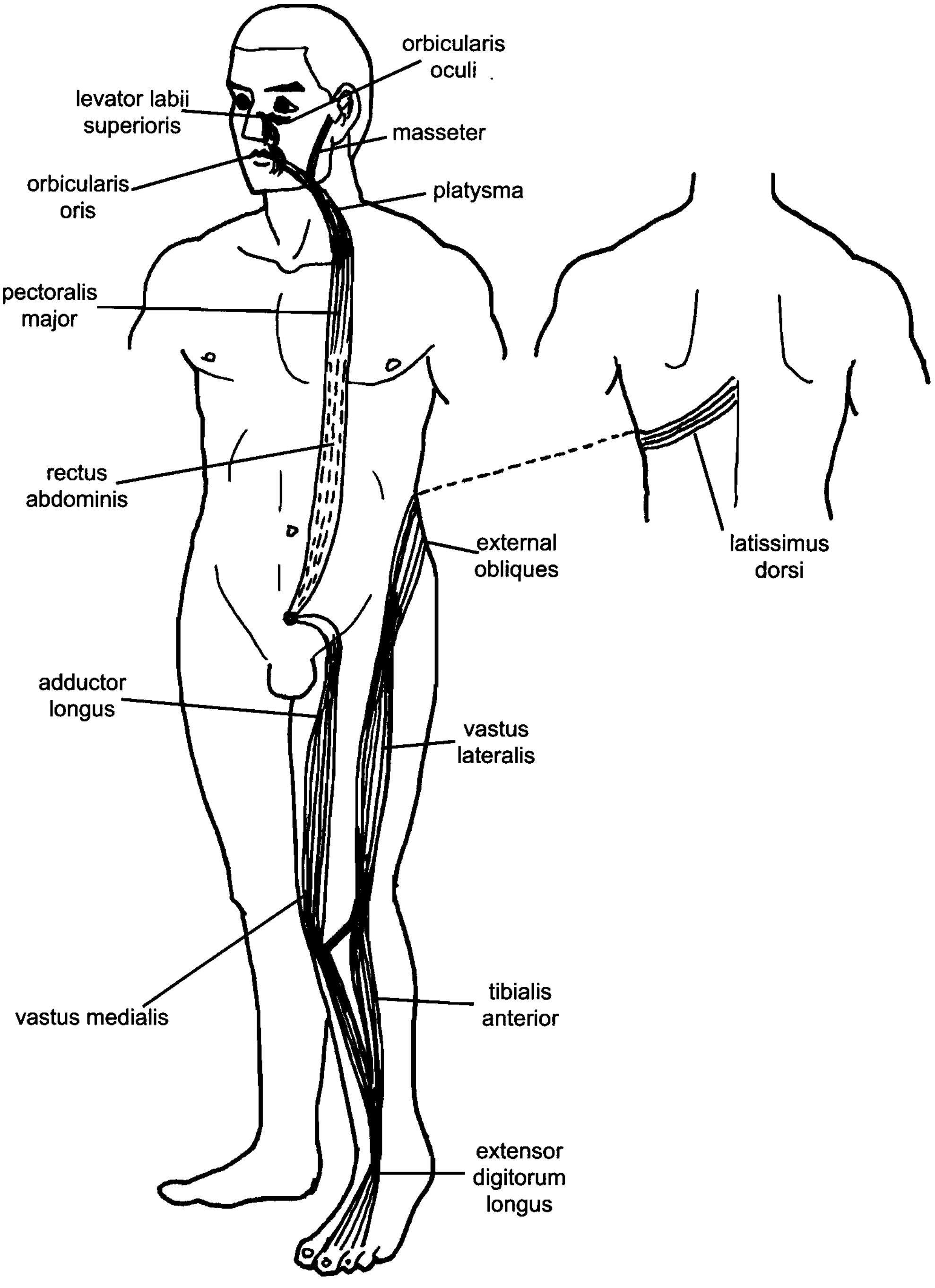
orbicularis oculi
levator labii superioris
masseter
orbicularis oris
platysma
pectoralis major
rectus abdominis
external obliques
latissimus dorsi
adductor longus
vastus lateralis
vastus medialis
tibialis anterior
extensor digitorum longus

Figure 7.30. The Muscle Region of the Stomach Channel: High Tide is 7 a.m.–9 a.m. Pathological symptoms include strained muscles of the big toe; spasms or hardening of the muscles in the foot; knotted or twisted muscles in the lower leg and thigh; swelling in the anterior pelvis region; hernia; spasms of the abdominal muscles; spasms or stiffness of neck and cheek muscles; and eye spasms.

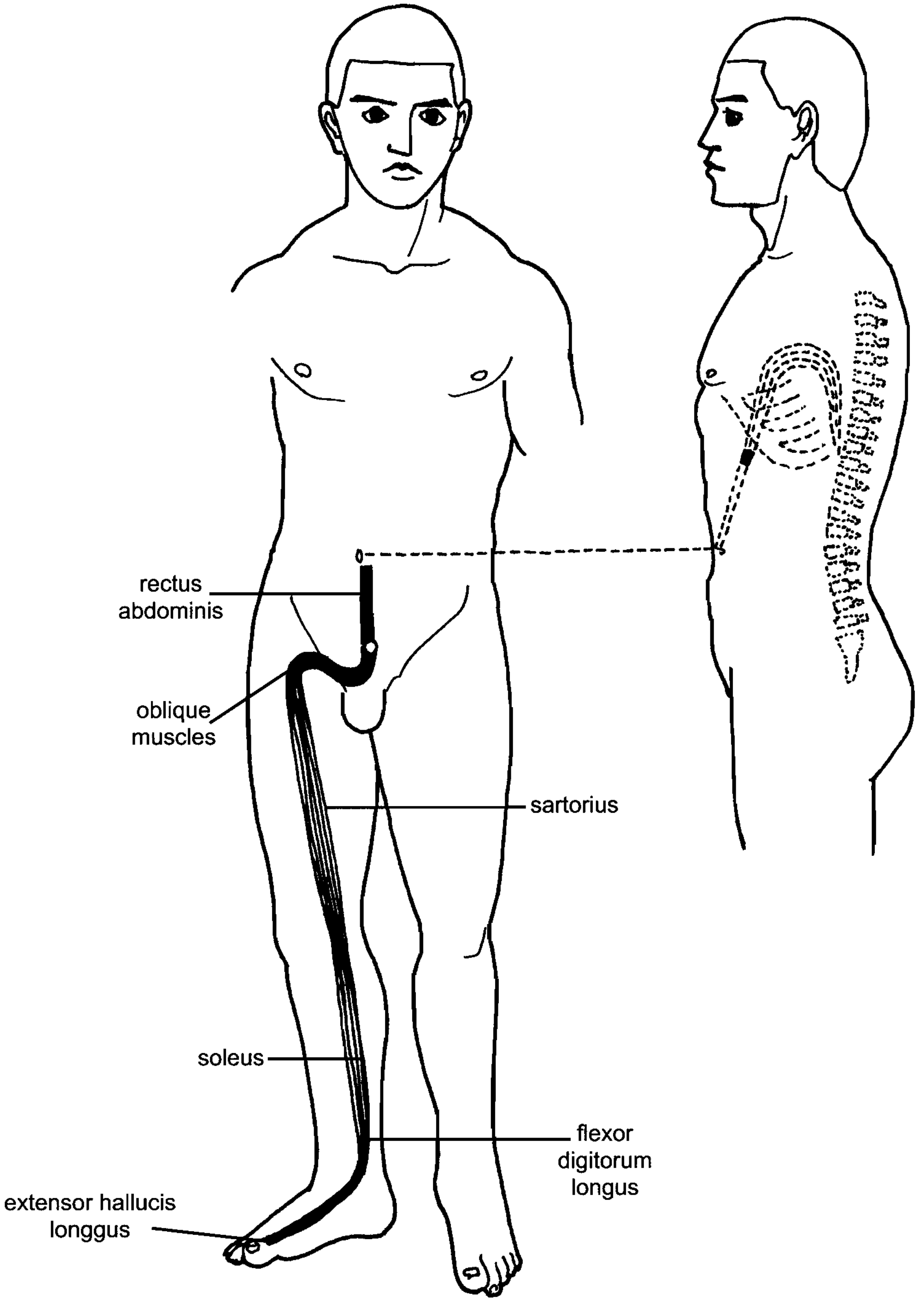
rectus abdominis
oblique muscles
sartorius
soleus
flexor digitorum longus
extensor hallucis longgus

Figure 7.31. The Muscle Region of the Spleen Channel: High Tide is 9 a.m.–11 a.m. Pathological symptoms include strained muscles of the big toe; pain in the internal malleolus of the ankle upon rotation; pain along the medial aspect of the knee and adductor muscles of the thigh; groin strain; and pain due to strained upper abdominal muscles and mid-thoracic vertebrae.

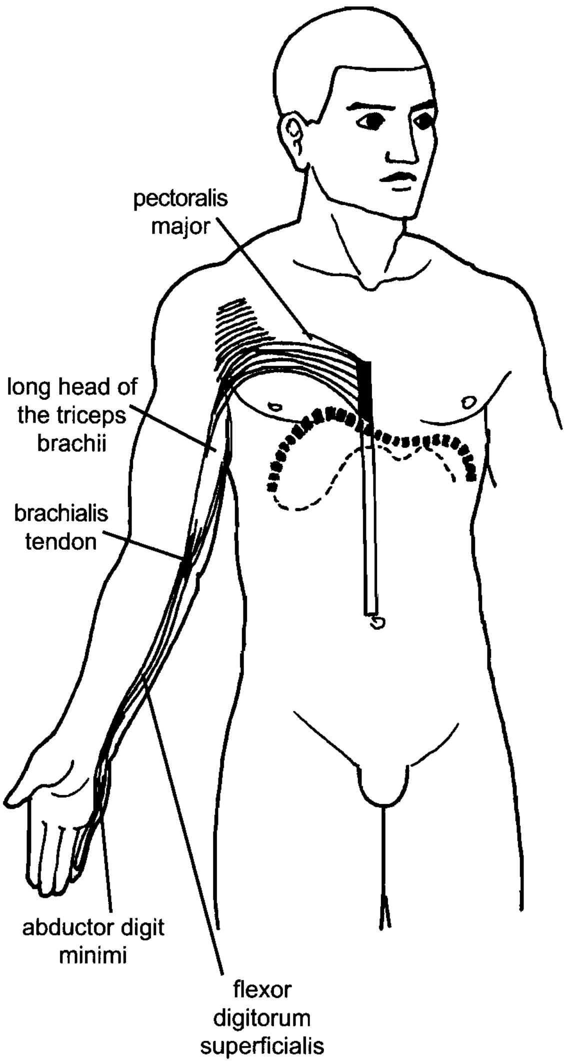
pectoralis major
long head of the triceps brachii
brachialis tendon
abductor digit minimi
flexor digitorum superficialis

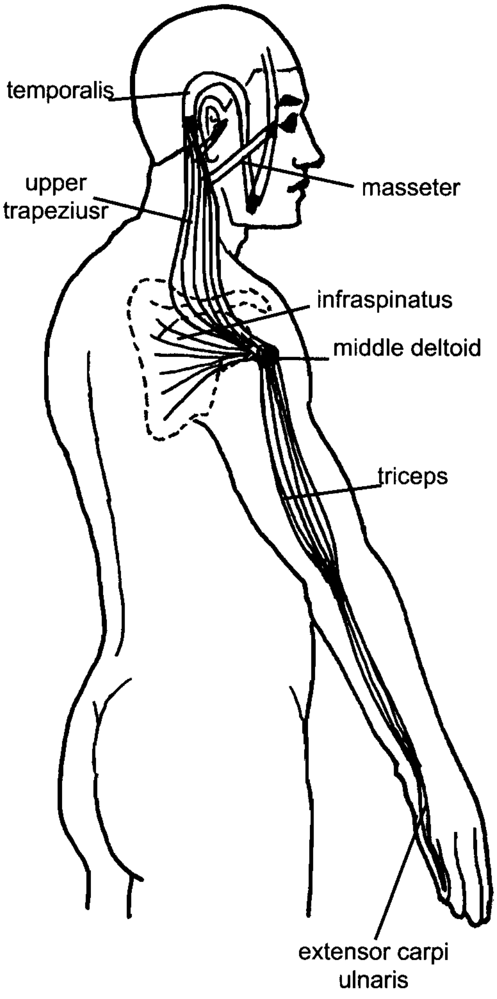
temporalis
upper trapeziusr
masseter
infraspinatus
middle deltoid
triceps
extensor carpi ulnaris

Figure 7.32. The Muscle Region of the Heart Channel: High Tide is 11 a.m.–1 p.m. Pathological symptoms include strained muscles of the little finger; stiff or strained muscles with spasm and/or pain along the course of the Heart Channel, including internal cramping within the diaphragm and upper abdominal area.

Figure 7.33. The Muscle Region of the Small Intestine Channel: High Tide is 1 p.m.–3 p.m. Pathological symptoms include strained muscles of the little finger; pain along the medial and posterior aspects of the elbow; pain in the posterior aspect of the axilla, neck, and scapula region; tinnitus related to ear ache; and poor vision.

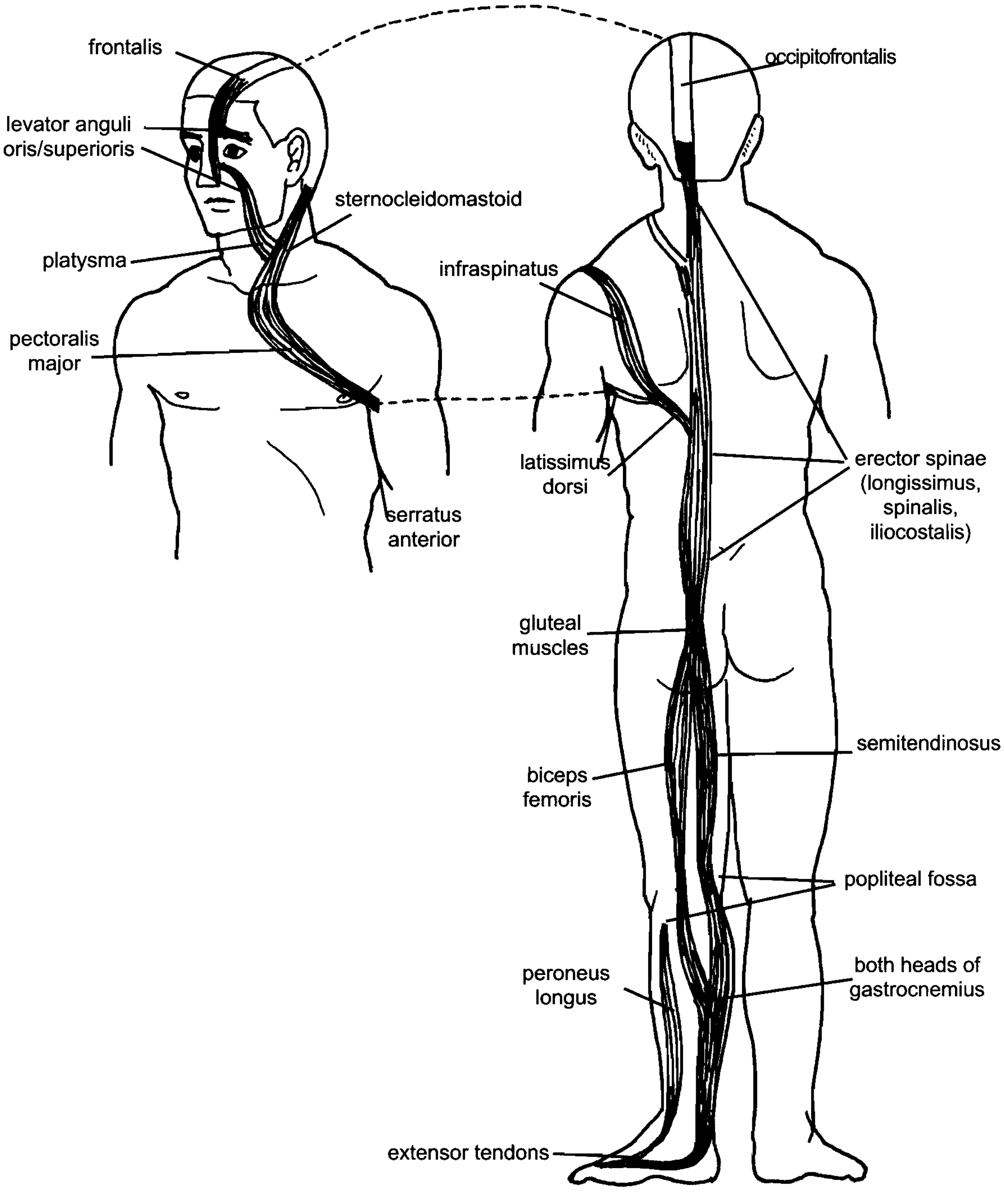
frontalis
levator anguli oris/superioris
sternocleidomastoid
platysma
infraspinatus
pectoralis major
latissimus dorsi
serratus anterior
occipitofrontalis
erector spinae (longissimus, spinalis, iliocostalis)
gluteal muscles
semitendinosus
biceps femoris
popliteal fossa
peroneus longus
both heads of gastrocnemius
extensor tendons

Figure 7.34. The Muscle Region of the Bladder Channel: High Tide is 3 p.m.–5 p.m. Pathological symptoms include strained muscles of the big toe; swelling and pain in the heels; stiffness or spasms along the spine and back area; frozen shoulder; stiffness or spasms in the axillary and clavicle regions.

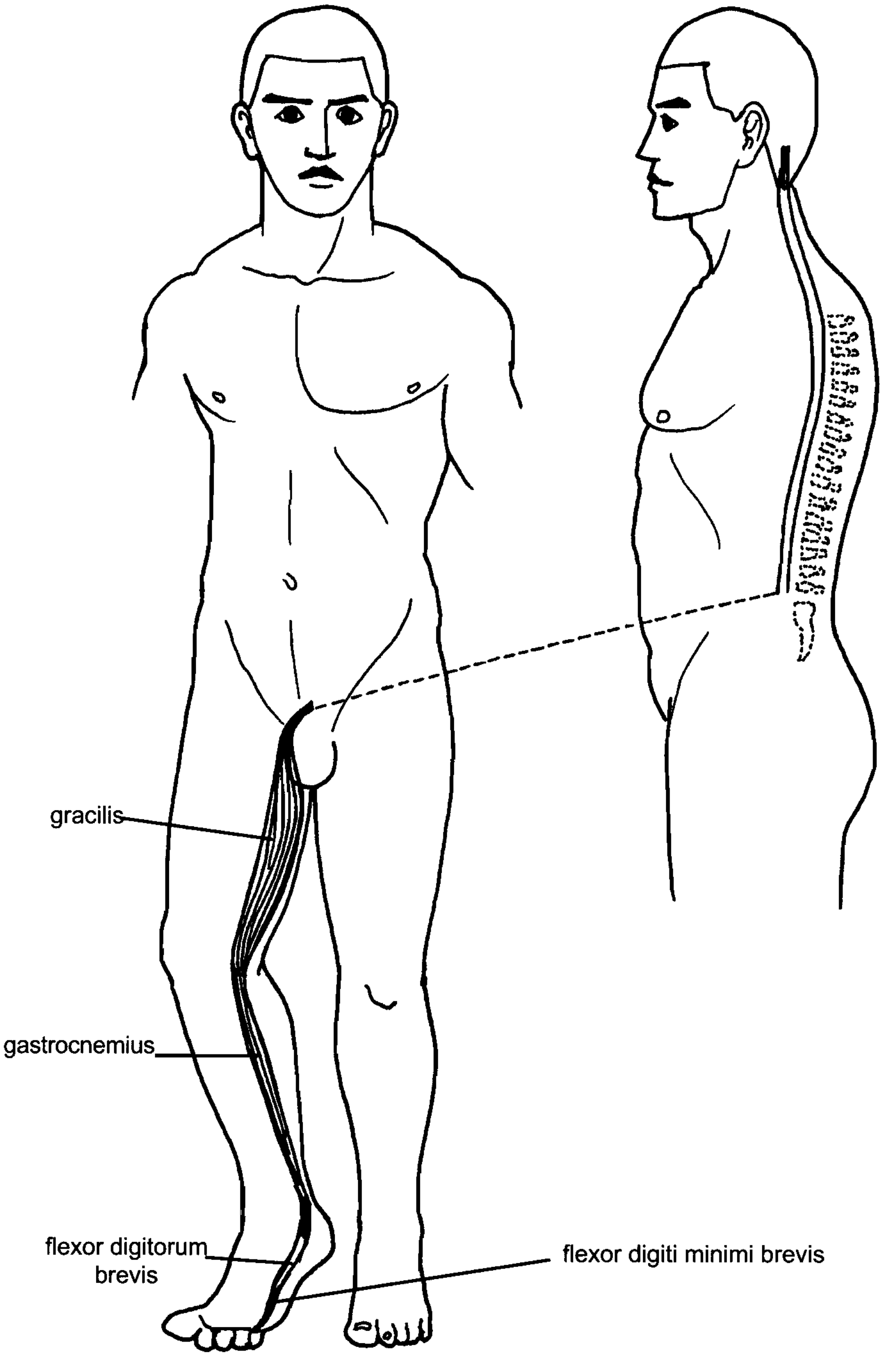
gracilis
gastrocnemius
flexor digitorum brevis
flexor digiti minimi brevis

Figure 7.35. The Muscle Region of the Kidney Channel: High Tide is 5 p.m.–7 p.m. Pathological symptoms include strained muscles on the bottom of the foot; spasms or stiffness along the Kidney Channel, resulting in an inability to bend forward (Yang disorder) or backward (Yin disorder), with difficulty in flexing or extending the head.

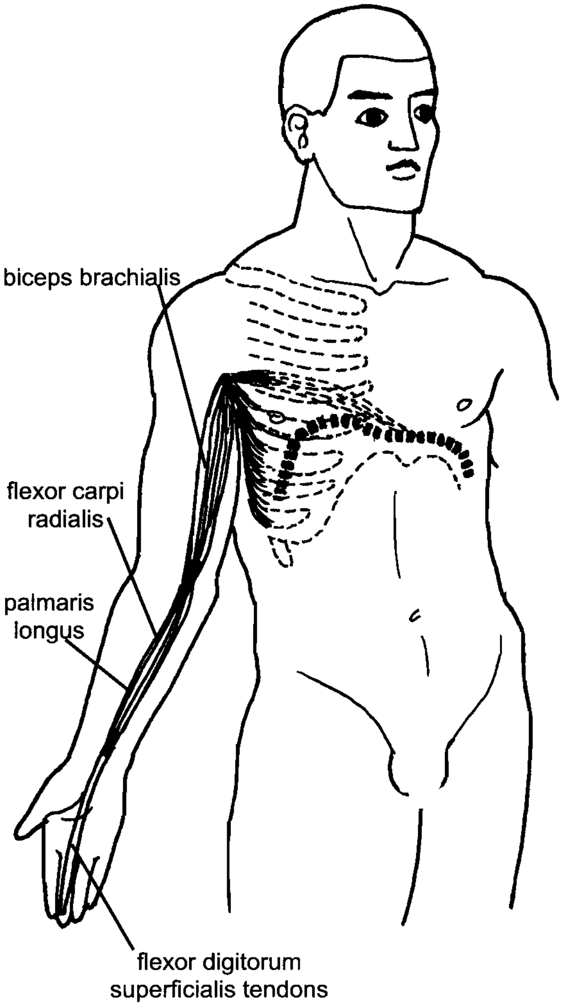
biceps brachialis
flexor carpi radialis
palmaris longus
flexor digitorum superficialis tendons

Figure 7.36. The Muscle Region of the Pericardium Channel: High Tide is 7 p.m.–9 p.m. Pathological symptoms include strained muscles of the middle finger; stiff or strained muscles, or spasms and/or pain along the course of the Pericardium's Channel; and chest pain and spasms.

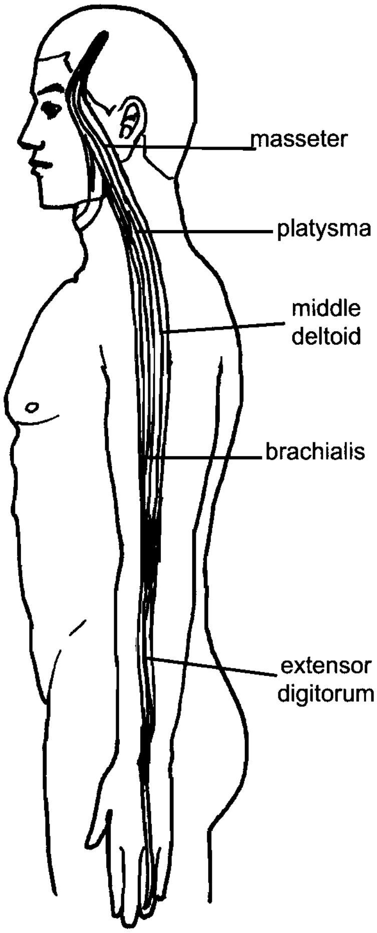
masseter
platysma
middle deltoid
brachialis
extensor digitorum

Figure 7.37. The Muscle Region of the Triple Burner Channel: High Tide is 9 p.m.–11 p.m. Pathological symptoms include strained muscles of the ring finger; stiff or strained muscles, or spasms and/or pain along the course of the Triple Burners' Channel.

THE TWELVE SKIN ZONES

Based on the Twelve Primary Channels, their collaterals, and the body's Muscle/Tendon Channels, the surface tissue is divided into Twelve Skin Zones. These Skin Zones are the surface contact areas for the body's channel and collateral systems. These twelve regions (sometimes called the Twelve Cutaneous Channels) are located in the superficial layers of the skin and have continuous and direct contact with the external environment. They are, therefore, the most sensitive to climactic changes and must adapt to protect the body from External pathogenic factors.

Both the Qi and the Blood at the Twelve Skin Zones rely on the collaterals for their nourishment. The Twelve Skin Zones circulate Wei Qi, which in addition to its defensive function, is also in charge of opening and closing the pores. The skin relies primarily upon the strength of the Wei Qi for resistance to the invasion of External pathogenic influences, as these harmful influences must first penetrate the skin before they can affect the body's internal tissues and organs. Since the Lungs rule the skin, weak Lung Qi can allow pathogens to penetrate the skin and affect the Lungs.

Pathological symptoms associated with the Twelve Primary and connecting Channels manifest along the surface of the Twelve Skin Zones, before progressing deeper into the body's connective tissue. The early stages of disease are called Exterior conditions; if the exterior Wei Qi is strong enough, pathogens will be stopped at the external level. The order of pathogenic progression is as follows.

1. The Body's Wei Qi is the body's first line of defense. A healthy body will project a strong, protective, external energetic field, which will maintain a capable defensive boundary. If the patient becomes weak, tired, or stressed, however, External pathogens may advance onto the patient's skin and begin attacking his or her health.
2. The Skin is the body's second line of defence. If the body's Wei Qi is not strong enough to resist an External attack and the skin is not capable of warding off the pathogens, then the pathogens will attack the skin, causing the sweat pores to open, thus allowing the pathogens to advance towards the patient's collaterals.
3. The Collaterals are the third line of defense. If the body's collaterals are unable to redirect or purge the advancing pathogens, the pathogens will then advance into the patient's channels.
4. The Channels are the last line of defense before the advancing External pathogens invade the patient's internal organs. If the body's channels are unable to redirect or purge the pathogens, and the pathogens are allowed to continue their progression, they will advance further into the body's internal organs.
5. The Internal Organs are affected once the pathogens have breached all the outer defences and disease can now begin to establish its residence.

Pathologies associated with the Primary Channels manifest diagnostically through the body's skin, channels, and points. The observation of pimples, moles, discoloration, and changes in electro-conductivity are all valuable signs of pathogens, or obstructions, invading the skin.

A disease of external origin that first lodges within the patient's skin can be treated by the Qigong doctor. The doctor stimulates the Wei Qi in the affected skin zone through external Qi projection before it progresses further into the body's connective tissue. An alternate treatment would be to apply herbal ointments and/or moxa therapy.

THE CONNECTIVE TISSUE OF THE TWELVE SKIN ZONES

The connective tissue always responds to the outside stimulation of the Twelve Skin Zones. The body's connective tissue is just below the skin, and is regarded as one of the largest, most extensive tissue organs in the body. Its function is to support, connect, contain, and transmit. It is a continuous network of structure that binds tissues into their organ shape. It also supplies the internal organ's life support with vessels and ducts, and securely fastens each organ within the body

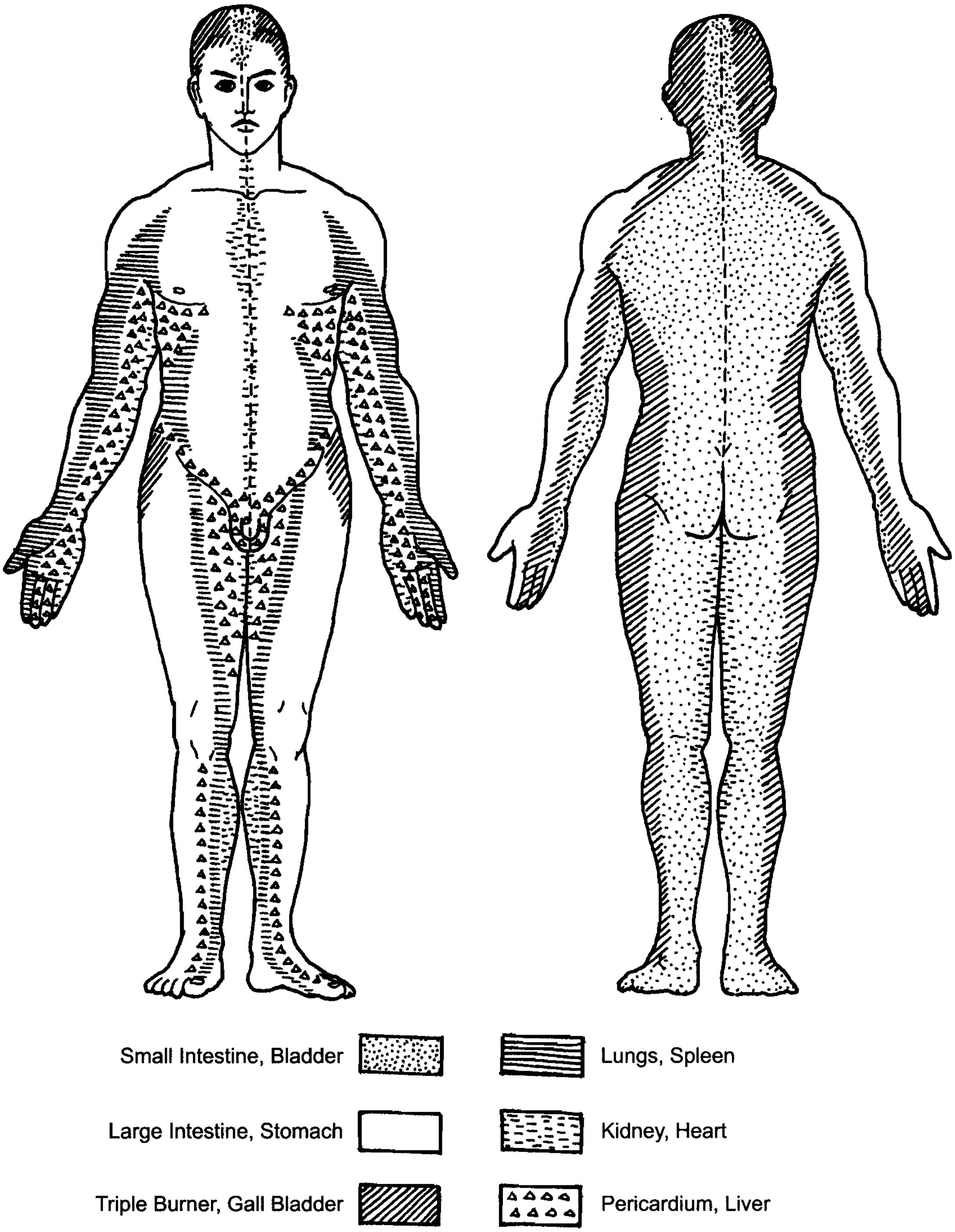
Small Intestine, Bladder
Lungs, Spleen
Large Intestine, Stomach
Kidney, Heart
Triple Burner, Gall Bladder
Pericardium, Liver

Figure 7.38. The Body's Twelve Skin Zones, Based on the External Flow of Qi from the Twelve Primary Channels and Their Collaterals

cavity. This connective tissue surrounds and anchors the vessels within the muscles, bones, and organ tissue. Its fluid nature supports the entire body structurally by hydrostatic pressure. Any stimulation on one of the body's Twelve Skin Zones will directly affect the body's connective tissue, and will stimulate the corresponding internal organ associated with that particular zone.

The superficial fascia (the connective tissue just under the skin) divides into a top and bottom layer. The top layer is the fatty layer, which constitutes the main fatty tissue of the outer surface of the body and fascia. This fatty layer acts as an insulator, helping to maintain a constant body temperature. It is metabolically active, functionally storing fat as fuel, releasing it in response to nerve and hormonal stimuli. It corresponds to the greasy layer where the Wei Qi circulates throughout the body's tissues. The deeper layer of the superficial fascia envelopes the nerves, veins, arteries, lymph vessels, and nodes.

From a Traditional Chinese Medical perspective, the Qi circulating within the body's inner fascial connective tissue can be stimulated through the external energy stimulation of the Twelve Skin Zones. Heat and movement help maintain the body's connective tissue's base fluids, facilitating movement and the conduction of energy. Any obstruction occurring on the body's surface tissues can result in the binding or thickening of the connective tissue, thus creating adhesions. This obstruction slows the metabolic process and compromises the body's immunity (Figure 7.38).

If the integrity of the connective tissue is compromised, the body's immune system declines. Compartments of connective tissue influence the spread of toxins, diseases, infections, and tumors. The fibrous walls, as well as chemicals in the fluid of the connective tissue, prevent the spread of pathogens from one area of the body to another.

CLINICAL DIAGNOSIS AND THE TWELVE SKIN ZONES

Clinically, the diagnosis and treatment techniques relating to the use of the Twelve Skin Zones is quite extensive. In diagnosis, the skin's surface area along these zones is examined for evidence of discoloration or tissue obstruction. Changes in skin color and tissue formations are observed and categorized as follows.

1. A darkish hue reflects obstruction of Qi and Blood.
2. A bluish-purple color indicates local pain.
3. Change in color from yellow to red shows evidence of Heat.
4. A white or pallid skin tone signifies Deficiency or Cold.
5. Boils, pimples (especially on the back), hives, and eczema, as well as hard lumps or nodules beneath the surface of the skin, indicate diseases associated with the Twelve Primary Channels (manifesting through the Twelve Skin Zones).

A palpable lump can develop from a variety of sources, including congealed or stagnant Blood. There are generally eight types of lumps observed in the clinical environment:

1. The Shrimp Lump, is shaped in the form of a shrimp.
2. The Turtle Lump, is named for its shape and multiple lump formations.
3. The Blue Lump, derives its color from the superficial blood vessels on which it forms.
4. The Single Lump, is diagnosed as opposed to multiple lump formations.
5. The Dry Lump, indicates an obstruction.
6. The Blood Lump, relates to congealed Blood.
7. The Abdominal Lump, is named for its physical location.
8. The Yellow Lump, is named because of the yellow pallor in the specific region of the lump.

PALPATING THE PATIENT'S SKIN

Before palpating and diagnosing the skin, the Qigong doctor first scans the patient's body to feel the temperature, moisture, and texture of the patient's tissues.

1. The doctor scans and palpates the patient to diagnose the condition according to the temperature:
   a. If the skin feels hot to the touch it often indicates a presence of a Damp-Heat condition.

   b. When first touching the patient with light pressure if the skin feels immediately hot, and as the pressure is maintained the feeling of heat increases, this indicates an invasion of Exterior Wind-Heat (with a pathogenic factor still on the exterior surface).
   c. If the skin over a blood vessel feels hot on medium pressure, it indicates Interior Heat in the Middle Burner or Heart area.
   d. If the skin feels hot on deep, heavy pressure, it indicates an Empty-Heat condition from a Yin Deficiency.
   e. If the skin feels cold to the touch, it often indicates a Cold condition. This condition is often manifested in the lower back and lower abdominal region, indicating a Deficiency of the Kidney Yang.
2. The doctor scans and palpates the patient to diagnose the condition according to the amount of moisture:
   a. A moist feeling of the skin may indicate an invasion of the Exterior by Wind-Cold or Wind-Heat.
   b. If the skin feels moist due to spontaneous sweating, it indicates a Deficiency of the Lungs' Qi (in the absence of exterior symptoms).
3. The doctor scans and palpates the patient to diagnose the condition according to the texture:
   a. If the skin is scaly and dry, it indicates an exhaustion of the Body's Fluids.
   b. If the skin feels dry, it indicates either a Yin condition of the Lungs or a Blood Deficiency.
   c. If the skin is swollen and an indentation is left visible after pressing, it indicates Edema (called Water Swelling).
   d. If the skin is swollen and no indentation is left visible after pressing, it indicates a retention of Dampness (called Qi swelling).

CHANNEL AND COLLATERAL THERAPY

Because the Yin and Yang organs' Qi flows within the body's internal and external surfaces, all the internal and external channels and collaterals, five tissues (tendons, blood vessels, muscles, skin, and bones), and five sense organs (eyes, ears, nose, mouth, and tongue), together create an energetic network. Through practicing sitting meditations, patients as well as Qigong practitioners become aware of the circulation of Qi along the body's channels and collaterals. The awareness of what the Qi feels like enables the meditators to feel the flow and function of each channel, as well as the Blood and heat cycle in each Skin Zone. This makes it possible to control the Qi circulation through intention and imagination.

Balancing the Conception and Governing Vessels is of the utmost importance in the practice of any Qigong regulation. Along these two Vessels the Qigong practitioner draws the Yang Fire and Yin essences up and down the body, thus fusing the Water and Fire energies. This fusion is utilized by the doctor in every Qigong treatment, and is responsible for balancing the patient's Yin and Yang energy. This balance is created through the discipline of the Microcosmic Orbit meditation, which connects the Governing and Conception Vessels.

The students' awareness of Qi flowing through the body usually develops in three distinct stages:

1. First, they begin to feel the energy flow along the surface channels of the body, especially within the areas of the extremities. Usually at this stage, the students will feel heat and tingling sensations within the body's surface muscles.
2. Next, they begin to feel the energy flow deeper within their tissues, especially along the tendons, deeper muscles, and visceral organs. Usually at this stage, they begin to feel mild electric shocks and vibrations within their muscles, bones, and visceral organs.
3. In the final stage, they feel the Earth and Heaven energy penetrate through their body's

outside channels and inner organs, connecting into their Taiji Pole, and then turning outward again. Usually at this stage, they will feel their entire energized body vibrate. This vibration may be triggered through either universal or environmental changes, as the Medical Qigong practitioner becomes hypersensitive to any form of external energetic disturbance.

CONTROLLING QI EXTENSION THROUGH THE CHANNELS

The Qigong doctor must first be able to control his or her own energy circulation, causing it to flow in or out, expand or contract at will, before beginning to extend energy for the treatment of any patient. This energetic control is gained through specific imagination and visualization techniques. The fundamental theory for these techniques is "The imagination leads the Mind, the Mind leads the Qi." The doctor must be able to utilize energy from the natural environment (trees, waterfalls, ocean, desert, etc.) in order to replenish and facilitate his or her own energy.

By drawing in Heavenly Qi and combining it with his or her body's connection to Earthly Qi, the doctor may mobilize and activate the Qi of the patient. This combining method establishes an energetic healing field of Qi, which responds to the Qigong doctor's own Qi circulation, as well as to that of the natural environment, which includes the patient. This ability must be achieved in order to activate the patient's Qi and then regulate it by purging the Excess Qi, replenishing the Deficient Qi, dredging the channel Qi, and guiding the Qi back to its origin.

CHAPTER 8

THE BODY'S ENERGETIC POINTS

INTRODUCTION TO ENERGETIC POINTS

The Chinese word for point is Xue or Xue Wei, meaning "cavity place, cave, den, hole, or hollow." Although the Chinese word Xue is usually translated into English as "point" or "acupoint," it is valuable for the English-speaking student to understand the word's different connotations. Points are almost always located in hollows or depressions, which are physically palpable and found in over 1,000 locations all over the body. Many, but not all, are located along the channels and vessels. Sometimes when touching a point, it feels as if your fingertip has found the entrance of a small cave or opening hidden by the skin covering it.

The word Xue also implies a den or lair: thus, it is not only a cave but a home for some type of powerful life form. In the human body, our living Qi gathers and pools within these protected hollows (Figure 8.1). For this reason, some writers have preferred to translate Xue as "vital hollow," rather than "point." Because of its popular acceptance, I have chosen to use the term "point" throughout this book.

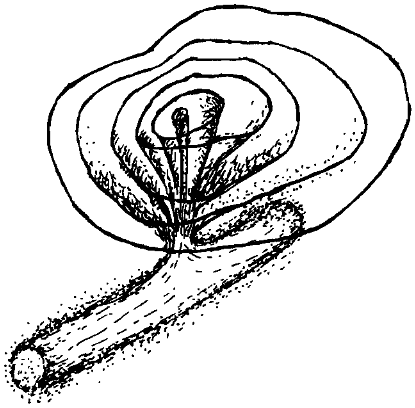

Figure 8.1. The energetic points are considered the body's small energetic pools.

THE FOUR WAYS POINTS ARE FORMED

Generally, points are formed in four ways. The first type of point occurs when two muscles, tendons, or ligaments intersect or overlap in a way that creates a small depression at the area of convergence. To form the depression, the two muscles either come together, separate, overlap, or run parallel to each other. Usually the channels lie under a protective layer of muscle, but at the aforementioned junctures, the channels may be exposed, and thus closer to the surface.

The second type of point occurs when a nerve is exposed in an area without much muscle tissue to cover and protect it. Because the nerve is exposed, such points are especially sensitive, and may cause a sharp pain or an electric shock sensation when stimulated.

The third type of point lies under a protective superficial layer of muscle. It can however, still be stimulated due to its natural sensitivity and because the overlying layer of muscle is thin.

The fourth type of point lies in depressions on the surfaces of the bones and cartilage. These depressions may take the form of grooves, pits, hollows, indentations, fissures, or crevices.

SCIENTIFIC RESEARCH OF ACUPOINTS

In the early 1970's, shortly after President Nixon's historic visit to China, Dr. Robert O. Becker, M.D., a Syracuse University orthopedist, received a grant from the National Institutes of Health to research how acupuncture works. Dr. Becker reasoned that the channels were electrical conductors (independent of the nervous system yet communicating with it) that carried an injury message to the brain, which responded by sending back the appropriate level of direct current (DC) to stimulate healing in the troubled area."

Becker also suggested in his book *The Body Electric*, that "any current grows weaker with distance, due to resistance along the transmission cable. The smaller the amperage and voltage, the faster the current dies out. Electrical engineers solve this problem by building booster amplifiers every so often along a power line to get the signal back up to strength. For currents measured in nanoamperes and microvolts (such as those generated in the human body), the amplifiers would have to be no more than a few inches apart—just like the acupuncture points."

Dr. Becker and his assistant, Maria Reichmanis, a young biophysicist, designed special equipment to measure and map the electrical conductivity along the channels. They found the predicted electrical characteristics along the channels and at half the points. These same points showed up on all the people tested. They suggested that the other traditional points may simply be weaker, or of a different kind, than the ones that their instruments detected.

Each point that they found was electrically positive compared to its surrounding tissue, and each point had an electrical field surrounding it with its own characteristic shape.

Later research (see appendices) in China, France, and the United States has shown not only that these points and channels have electrical properties but that they also emit and absorb light and sound, both of which are vital to healthy biological functions.

THE FOUR CATEGORIES OF ENERGETIC POINTS

Points are divided into four categories (Figure 8.2) and are described as follows.

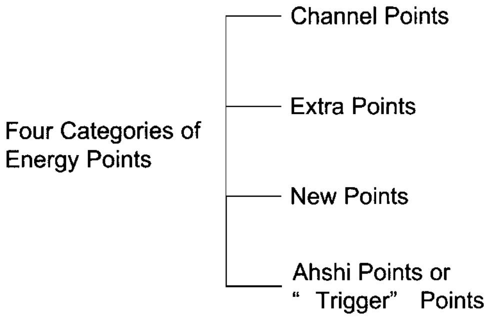
Four Categories of Energy Points
Channel Points
Extra Points
New Points
Ahshi Points or " Trigger" Points

Figure 8.2. The Four Categories of Energetic Points

1. Channel points are the Primary points that are distributed along the Twelve Primary Channels and the Conception and Governing Vessels. There are 365 channel points.
2. Extra points are not regarded as having a specific origin in the fourteen main channels, although they are named and are in definite locations. These points are also called "miscellaneous points."
3. New points are a fairly large group of points discovered since the Communist "Liberation" of 1949, in many cases through electric point detection. These points generally have no energetic description, but do have indications. Combined with the Extra points, there are more than 700.
4. Ahshi Points are tender or painful spots near a diseased or injured area and can be anywhere, on or off of traditional points and channels. They function like trigger points, as they are points of tension or pain. Ahshi points are not given specific names or definite locations as they are mostly used for pain syndromes.

THE TWO GATES OF ENERGETIC POINTS

Another way of categorizing the points is into the two general types commonly used by the Qigong doctor.

1. The Outer Qi Gates are points through which the patient's Shen and Qi enter and leave the body. The Qigong doctor or patient may gather healing Qi from the outer environment through these gates. In terms of pathology, these points are especially vulnerable to the Six External Pathogens, i.e., Cold, Hot, Wind, Damp, Dry, and Fire invasion (see Chapter 21), which may enter the body through these gates. These same pathogens may also be ex-

pelled through these gates. For this reason, many of these outer Qi gates have the word Feng (Wind) in their names, such as Fengchi GB-20 (Wind Pool).

2. The Inner Qi Gates are points through which the Qi of the Yin and Yang organs and channels is transported back and forth internally and externally, deeply and superficially. These points are more involved with the movement of Qi within the body and are less involved in the exchange with the outer environment. In this way, the Inner Qi Gates enable the Qi to communicate between the different parts of the body.

These two types of points (or gates) are specific areas where the Qigong doctors concentrate their intention and focus their energy extension. Point locations have a greater amount of accumulated Qi, and are used for dispersing or tonifying the body's organ energy, to promote the correct balanced circulation of Qi and Blood.

THE THREE FUNCTIONS OF ENERGETIC POINTS

The points can be used for three clinical functions: to manipulate Qi and Blood, for diagnosis, and to serve as pathways for the elimination of disease.

1. Qi and Blood can be manipulated to purge Excess conditions, tonify Deficient conditions, or for regulation.
    a. To disperse Excess conditions, the Qigong doctor emits Qi over a specific point (or points) and begins to purge, or disperse, the energy in the area (to purge is to remove toxins directly from the organ, whereas to disperse is to lead the toxins out through the channels). The goal of this treatment is to move any Excess, particularly stagnant Qi and Blood. If only Qi is stuck, the pain will move and feel dull. If only Blood is stuck, the pain will not move and will be sharp. This distinction, however, is more important in Chinese herbology than it is for medical Qigong, acupuncture, or massage because, as stated in the classics, "Qi is the commander of Blood"; therefore, moving the Qi in the latter therapies will also move stagnant Blood.
    b. To tonify a Deficient condition, the Qigong doctor emits Qi over a specific point or points, and begins to tonify or strengthen the patient's Qi, Blood, Yin, Yang, or specific organs and tissues. The Qigong doctor uses varying techniques according to the patient's needs (e.g., in order to prevent feeling discomfort, the doctor will tonify weak patients slowly).
    c. To regulate, the Qigong doctor emits Qi over a specific point or points, and allows the body's natural homeostasis to take over and restore healthy function. This type of treatment is also used when a patient has a combined Excess / Deficiency syndrome (such as stagnant Qi in the chest due to a Qi Deficiency in the Lungs and Kidneys).

    The healing potential of point stimulation can be accomplished through various modalities, such as Energy Projection, Energetic Point Therapy, Jing Point Therapy or Invisible Needle Therapy.

2. Diagnosis through point palpation and inspection is an important diagnostic tool in all the branches of Chinese medicine. If a point hurts when touched with light pressure, is hard or swollen, or is purple, black and blue, or red, this indicates an Excess condition. If the point hurts on deep pressure or is soft to the touch, lacks resilience, or is sallow in color, it indicates an underlying Deficiency.

    The skin over points often feels sticky, in contrast to the slippery or smooth quality of the skin surrounding the point. When passing the fingers over a point, the doctor generally feels the energetic pulse of the channel. When touched, the point energetically responds like an echo, vibrating along the channel into the organ and then back again to the

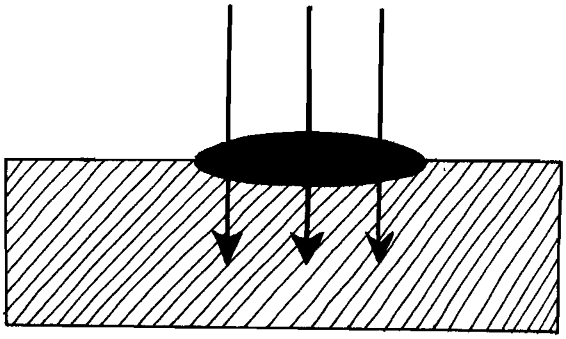

Figure 8.3. Weak or Deficient tissue will naturally pull and absorbe Qi into the surrounding areas.

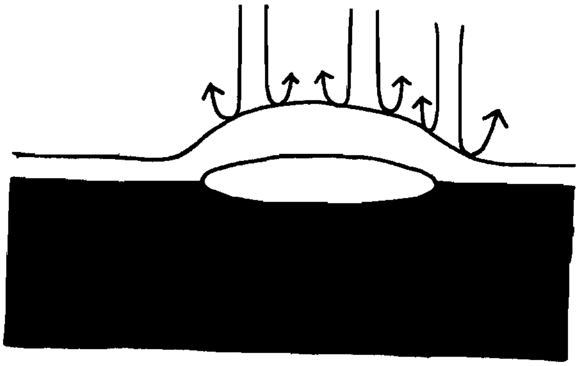

Figure 8.4. Strong, Excess, or Armored tissue will naturally repel Qi from the surrounding areas.

doctor's finger. The doctor can use this energetic response for diagnostic evaluation by sending a pulse along the flow of the channel or against it (into the organ itself). Once sufficient stimulation is applied to the point, its organ or region is "reminded" of its normal function. In this way, diagnosis and treatment are accomplished simultaneously.

Once the location and energy flow of the channel to be treated is diagnosed, the doctor must take into consideration that each patient will react differently in accordance with the severity of symptoms, as well as the individual's constitutional type. Before beginning treatment, the doctor should extend Qi between the energetic point's boundary (the space separating the Wei and Ying energy levels), until the soreness or distention is felt on the tips of the doctor's fingers or palms (which act like a barometer). This form of diagnosis determines the condition of the channel and the area needing treatment. Although the feelings and sensations will vary with each patient, energetic truths remain constant. For example:

a. Weak or Deficient areas will feel like an empty hole or deep well. Deficient Blood or Qi will actually pull the energy into the area and will feel something like a subtle vacuum suction. If the area that the doctor is treating attacks his or her Qi and begins to absorb it, the doctor should tonify that area immediately. Think of this reaction as a dry house plant starving for water. The minute the water touches its soil, the plant immediately absorbs it into its roots (Figure 8.3).
b. Strong or Excess areas will repel the doctor's Qi. If the diseased area repels the doctor's Qi, he or she must immediately dredge the Excess. First, the doctor should focus on the region surrounding the Excess to drain the Excess energy away from the damaged points, then dredge the Excess energy away from the original trauma area. This technique is excellent for damaged or sprained joints. The doctor should focus on the patient's breathing when dredging. As the patient inhales, the doctor exhales and vice-versa. As the patient exhales, he or she releases pathogenic factors; by inhaling when the patient exhales, the doctor facilitates this process, and prevents depletion of his or her own Qi (Figure 8.4).

3. Disease transmission may occur due to the External invasion of pathogenic Qi into the body through the points. Some points are particularly vulnerable to exogenous invasions, especially Wind invasion. Wind-susceptible points usually have the word Feng in their names.

POINT NAMES

Traditionally, each point has a name, which either describes its location, its energetic function, or both. In addition, modern acupuncture texts have assigned a name and number for each point, according to its channel and its order along the natural course of the channel.

Sometimes a point may have several different traditional names, varying according to whether they are being used by medical Qigong doctors, acupuncturists, martial artists, or religious practitioners; for example, GV-1 is called Changqiang (Long Strength) in Traditional Chinese Medicine, and Weilu (Tail Gate) in medical Qigong and Daoist inner alchemy.

The history of point naming and the different terminologies used can be traced back to the different Buddhist and Daoist temples in ancient China. Each religious sect had its own specific names given to certain points, in accordance with the temple's spiritual and energetic needs. Giving the points different names was a form of secret code, used to keep the system pure and to prevent the knowledge from falling into the hands of the uninitiated or unscrupulous.

Sometimes the same name may refer to different points or areas on the body; for example, Yuzhen (Jade Pillow), located in the occipital region on the posterior base of the skull, is the name commonly used for Bl-9 in Traditional Chinese Medicine. However, in medical Qigong and Daoist inner alchemy, Yuzhen instead refers to an entire area the size of one's fist placed between the bilateral Bl-9 points.

CLASSIFICATION OF ENERGETIC POINTS

The points are further classified according to their energetic potential. This classification will assist the doctor of Traditional Chinese Medicine in choosing the right point or points to treat. The following is a list of energetic point classifications.

THE FIVE ANTIQUE SHU POINTS

Like the internal organs and channels, some points have a more powerful influence on the body then others. Each point has an action on the quantity of energy of an organ. Along each of the Twelve Primary Channels lie five specific points below the elbow or the knee called Antique Shu points (also called Transporting points and Command points). These points belong to the oldest classification of points, and are described in water images, comparing each section of the channel to the course water takes as it emerges from the ground and makes its way to the ocean. They are identified as follows: Jing (Well), Ying (Spring), Shu (Stream), Jing (River), and He (Sea). These five points are situated along the flow of energy from the extremities to the torso (running from the fingers to the elbows and from the toes to the knees), occupying the same progressive location along the patient's channels (Well, Spring, Stream, River, and Sea).

Centrifugal and Centripetal Energy Flow

All of the body's channel points energetically progress in two opposite currents according to the direction and flow of the channel's energy. These two energy currents are described as follows.

1. The Channel's Outward Centrifugal Flow of Qi passes from the organ, through the channel, to the point, and is expressed or intensified through the patient's exhalation. The outward centrifugal flow of energy is manifested in channel points and in areas of the body where there are conditions of Excess.
2. The Channel's Inward Centripetal Flow of Qi passes from the point, through the channel, to the organ. The inward centripetal flow of energy is manifested in channel points and areas of the body where there are conditions of Deficiency.

Through stimulation, channel points fulfill a double function. Centrifugally, they inform the doctor of impending internal disorders by discharging Excess energy. Centripetally, they transmit the energetic action placed on them, absorbing energy to supply the body's internal insufficiencies.

Yin and Yang Channels and the Five Elements

Each of the Five Antique Shu points also corresponds to the Five Elements, in the progression of the Creative Cycle. Each Yin channel begins

Points	Yin Channel Elements (Ethereal Soul—Hun)	Yang Channel Elements (Corporeal Soul—Po)
Jing-Well Points	Wood	Metal
Ying-Spring Points	Fire	Water
Shu-Stream Points	Earth	Wood
Jing-River Points	Metal	Fire
He-Sea Points	Water	Earth

Figure 8.5. The Five Antique Shu Points and their Correspondence to the Five Elements

with the Hun progression through the Five Elements, starting with Wood, then Fire, Earth, Metal, and Water. Each Yang channel begins with the Po progression through the Five Elements, starting with Metal, then Water, Wood, Fire, and Earth (Figure 8.5).

Similar to the eyes relationship to Wood (Hun) and Metal (Po), the beginning points on the Twelve Primary Channels also have a relationship to either a Wood (Ethereal Soul) or Po (Corporeal Soul) connection. Hence, in meditation exercises, when specific finger tip points are connected, the blending of the Hun and Po energies are established. The practice of connecting specific finger tip points is sometimes referred to as the connection (or fusion) of the body's Ethereal and Corporeal Soul.

The Five Antique Shu points energetically flow from superficial to deep and are susceptible to external pathogens and climatic changes. The distal points on the feet tend to be more powerful than those on the hands. After the manipulation of the patient's distal points, the local channel points are used according to their feeling of tenderness. The Five Transporting points are explained as follows.

1. Jing-Well points are the first points along the extremities of the channels and are located at the tips of the fingers and toes. These points are where the energy of the channel pools and is at its most superficial and thinnest.

   These points are where the energy of the channels leaves the body moving centrifugally outward. At the end of the fingertips and toes, the channel's energetic polarity changes from Yin to Yang or vice versa. Due to this shift in polarity, the energy at the channel's extremities tends to be unstable, and therefore more easily influenced.

   Due to their outward movement of energy, the Well points can be used by the Qigong doctor to eliminate the patient's pathogenic factors quickly, especially when the Yin organs are affected. The Well points will also have a strong effect on the patient's mental state, quickly changing the patient's mood. These points are generally used for acute disorders, expelling exogenous pathogens, and mental disorders. They are used as "revival" points for loss of consciousness from fainting and heat stroke, convulsions, a feeling of fullness below the Heart, and local neuropathy (tingling, numbness, burning pain). In mental disorders, they are commonly used for irritability, mental restlessness, anxiety, hysteria, mania, and insomnia.
2. Ying-Spring points are the second points along the channels and, in all cases, are located in the second position of the channels' energetic progression up the arms or legs, and are located next to the Well points. These points are where the channel's energetic flow quickens its progression of Qi into the body.

At the location of the Spring points, the Qi in the channels is very dynamic and powerful. Slipping and gliding like the swirling movement of cascading water, the energy in these points can change quickly. The Spring points are generally used to eliminate Internal and External pathogenic factors (especially Heat) from the patient's body. They are also used when a disease effects a color change in the patient's complexion.

In the Yin channels, this point is always a Fire Element. Purging the Spring point in the Yin channels reduces Heat and clears Fire. In the Yang channels, the Spring point is always the Water Element point. The doctor can reduce Heat in the patient's Yang channel by tonifying this Water point.

3. Shu-Stream points are the third points along the channels and are located in the third position in the channels' energetic progression up the arms or legs, next to the Spring point (except for the Gall Bladder's Channel where it is the fourth point). These points are where the channel's energy rapidly pours through and slightly deepens its flow into the body.

   At these points, the external pathogenic factors penetrate deep into the channels and can be "transported" into the body's interior. These points are used to clear Wind and Dampness from the patient's channels or when a disease manifests intermittently.
4. Jing-River points are the fourth of the Five Transporting points; they are, however, not always located on the fourth point along the channel. The points are, however, always located between the wrist and elbow on the arms and between the ankle and knee on the legs. These points are where the channels broaden, and the energy flow continues to increase.

   At these points, the energy current flows wider, deeper, and more irregularly, like a large river, directing the invasion of external pathogenic factors towards the body's tendons, joints, and bones. These points are generally used when there is stagnation of Qi and Blood in the patient's channels, as well as for coughing, asthma, dyspnea, sore throat, and upper respiratory diseases.
5. He-Sea points are the fifth points of the Five Transporting points and, in all cases, are located at the elbows and knees. These points are where the energy of the channels is vast and plunges deep into the body.

   The Qi at the He-Sea points moves centripetally inward, flowing much slower, joining, collecting, and fusing with the general circulation of the body's energy (like a mighty river flowing into the sea). The He-Sea points are generally used for all Stomach and Intestinal diseases and for Internal disorders of the organs.

THE FIVE ELEMENTAL POINTS

Each of the Five Elemental points correspond to the Five Antique Shu points. The Five Elemental points are aligned in the progression of the Creative Cycle. Each Yin Channel begins with the Three Ethereal Souls (Hun) progression through the Five Elements, starting with Wood, then Fire, Earth, Metal, and Water. Each Yang channel begins with the Seven Corporeal Souls (Po) progression through the Five Elements, starting with Metal, then Water, Wood, Fire, and Earth.

THE SIXTEEN XI CLEFT (ACCUMULATION) POINTS

These points are used for purging acute conditions of Excess in the channels and organs. There is one point on each of the Twelve Primary Channels, plus one point on the Yin and Yang Heel Vessels, and one point on the Yin and Yang Linking Vessels. The Yin Channel Accumulation points are Lu-6, Pc-4, Ht-6, Sp-8, Lv-6, Kd-5, Yin Heel Kd-8, and Yin Linking Kd-9. The Yang Channel Accumulation points are LI-7, TB-7, SI-6, St-34, GB-36, Bl-63, Yang Heel Bl-59, and Yang Linking GB-35.

THE TWELVE FRONT MU (ALARM, COLLECTING) POINTS

The energy of the organs collect at these points; therefore, they can be used for both diagnosis and treatment. These points are palpated for diagnosis and may feel tender to the patient, either spontaneously or upon pressure. If they respond to light pressure or feel tight or swollen, the patient has an Excess condition; if the points respond to deep pressure or feel hollow or deflated the patient has a Deficient condition. The Twelve Front Alarm points are: Lungs Lu-1, Pericardium CV-17, Liver Lv-14, Spleen Lv-13, Gall Bladder GB-24, Kidneys GB-25, Heart CV-14, Stomach CV-12, Large Intestine St-25, Triple Burners CV-5, Small Intestine CV-4, and Bladder CV-3.

THE EIGHTEEN BACK SHU (ASSOCIATED, TRANSPORTING) POINTS

These points are all located on the Bladder Channel and are used for both diagnosis and treatment. These points transport Qi very effectively and directly to the internal organs. They can be used for treating the corresponding sense organs. The Eighteen Back Transporting points are: Lungs Bl-13, Pericardium Bl-14, Heart Bl-15, Governing Vessel Bl-16, Diaphragm Bl-17, Liver Bl-18, Gall Bladder Bl-19, Spleen Bl-20, Stomach Bl-21, Triple Burners Bl-22, Kidneys Bl-23, Sea of Qi Bl-24, Large Intestine Bl-25, Gate to the Yuan Qi (Guan Yuan) Bl-26, Small Intestine Bl-27, Bladder Bl-28, Sacrum Bl-29, and Anus Bl-30.

THE FIFTEEN LUO (CONNECTING) POINTS

One of these points is located on each of the Twelve Primary Channels, as well as the Governing and Conception Vessels and the Great Luo point of the Spleen. These points are used to treat channel pathology. The Fifteen Collateral points are divided into Yin and Yang channel points. Yin Channel points include: Lu-7, Pc-6, Ht-5, Sp-4, Lv-5, Kd-4, CV-15, and the Great Luo Channel Sp-21. The Yang Channel points include LI-6, TB-5, SI-7, St-40, GB-37, Bl-58, and GV-1.

THE TWELVE ENTRY POINTS

These points are where the energy from a Primary Channel enters its coupled channel in the Horary Cycle. The entry point is the first point on each channel in all cases, except for the Large Intestine's Channel in which it is LI-4.

THE TWELVE EXIT POINTS

These points are where the Qi leaves the Primary Channel to flow into its coupled channel in the Horary Cycle. The Yin Channel Exit points include: Lu-7, Pc-8, Ht-9, Sp-21, Lv-14, and Kd-

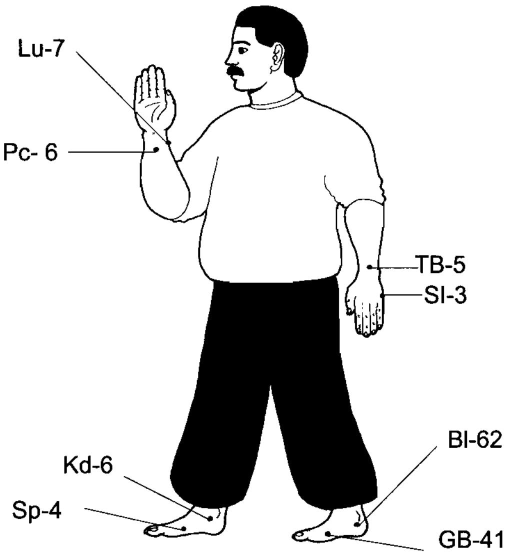

Figure 8.6. Locations of the Body's Master and Couple Points

Channel	Open with Master Point	Complete with Couple Point
Governing	Hand (SI-3)	Foot (Bl-62)
Conception	Hand (Lu-7)	Foot (Kd-6)
Thrusting	Foot (Sp-4)	Hand (Pc-6)
Belt	Hand (TB-5)	Foot (GB-41)
Yin Linking	Hand (Pc-6)	Foot (Sp-4)
Yang Linking	Foot (GB-41)	Hand (TB-5)
Yin Heel	Foot (Kd-6)	Hand (Lu-7)
Yang Heel	Foot (Bl-62)	Hand (SI-3)

Figure 8.7. The Vessels with their Associated Master Points and Couple Points

22. The Yang Channel Exit points include: LI-20, TB-23, SI-19, St-42, GB-41, and Bl-67.

THE EIGHT INFLUENTIAL POINTS

These points have specific effects on their respective organs, substances, and tissues. They affect the body's Qi, Blood, sinews, blood vessels, bones, and Marrow, as well as the Yin and Yang organs. The Eight Influential points are as follows: Yin Organs Lv-13, Yang Organs CV-12, Qi CV-17, Blood Bl-17, sinews GB-34, blood vessels Lu-9, Bones Bl-11, and Marrow GB-39.

THE EIGHT CONFLUENT POINTS

These eight points communicate with the Eight Extraordinary Vessels. The Confluent points are divided into Master and Coupled points. The Master point is the primary point chosen for treatment, its Coupled point is the connecting or secondary point used to open the specific Vessel. Each of the Eight Extraordinary Vessels has both a Master and Coupled Confluent point on the upper or lower limbs.

All of the Eight Extraordinary Vessels can be treated in various combinations according to their specific Master and Coupled point locations (Figure 8.6). If, for example, the Qigong doctor causes Qi to flow from the patient's Master point on the hand (SI-3) to its Coupled point (Bl-62) on the foot, the Governing Vessel will open. To complete the treatment, the doctor must first remove stimulation from the Coupled point then disconnect from the Master point (Figure 8.7).

The Eight Confluent points are as follows: Governing Vessel SI-3, Conception Vessel Lu-7, Belt Vessel GB-41, Thrusting Vessels Sp-4, Yang Heel Vessel Bl-62, Yin Heel Vessel Kd-6, Yang Linking Vessel TB-5, and Yin Linking Vessel Pc-6.

THE TWELVE YUAN-SOURCE POINTS

Each of the Twelve Primary Channels has a Yuan-Source point, where the body's Original Qi surfaces and pools. On the Yin Channels, the Yuan-Source points are always the Shu-Stream points. On the Yang Channels, however, the Yuan-Source points are separate points located between the Shu-Stream and the Jing-River points. These points are usually the fourth point from the distal end of the channel, except in the case of the Gall Bladder Channel, where the Yuan-Source point is the fifth point.

The Yuan Qi originates and resides in the Lower Dantian area and is dispersed to the Yin and Yang organs and then to the limbs by the Triple Burners.

The Yuan-Source points can be used clinically in diagnosis for the Yin organs. The Qigong doctor examines the points by both palpation and visual diagnosis.

The Yuan-Source points are the primary points on the Yin channel for tonifying and regulating their respective Yin organs. On the Yang Channels, the Yuan-Source points are quite different in terms of their importance and energetics. These points generally have a negligible tonifying effect on their related Yang organs and comparatively little Qi regulating effect. Their main functions are to expel various pathogenic factors, and to treat disorders along their channel pathways. The Twelve Yuan points are as follows: Gall Bladder GB-40, Liver Lv-3, Lungs Lu-9, Large Intestine LI-4, Stomach St-42, Spleen Sp-3, Heart Ht-7, Small Intestine SI-4, Bladder Bl-64, Pericardium Pc-7, and Triple Burners TB-4.

THE POINTS OF THE FOUR SEAS

The human body has four seas: the Sea of Qi, Sea of Blood, Sea of Marrow, and the Sea of Nourishment (literally Water and Grain). To access these seas, the Qigong doctor stimulates the following points:

1. The Sea of Qi points include: Bl-10, St-9, CV-17, GV-14 and GV-15. When the Sea of Qi is in Excess, symptoms include fullness in the chest, flushed complexion, and dyspnea. When the Sea of Qi is Deficient, symptoms include fatigue or low energy and an inability to speak.
2. The Sea of Blood points include: Bl-11, St-37, and St-39. When the Sea of Blood is in Excess, symptoms include anxiety, uneasiness, and unrest, and the entire body feels big. When the Sea of Blood is Deficient, symptoms include the body feeling too small for no apparent reason.
3. The Sea of Marrow points include: GV-15, GV-

16, GV-17, and GV-20. When the Sea of Marrow is in Excess, symptoms include sensations relating to that of Excess energy (lightness, feelings of strength and vitality). When the Sea of Marrow is Deficient, symptoms include fatigue, vertigo, dizziness, tinnitus, pain in the lower legs, impaired vision, and a desire to sleep.

4. The Sea of Nourishment points include: St-30 and St-36. When the Sea of Nourishment is in Excess, the abdomen feels distended. When the Sea of Nourishment is Deficient, symptoms include hunger with an inability to eat.

UPPER (UNITING) HE-SEA POINTS

There are three points on the upper arm that have a powerful effect on the Yang organs. The affected organs and their points include the Stomach LI-10, Large Intestine LI-9, and the Small Intestine LI-8.

LOWER (UNITING) HE-SEA POINTS

Each of the Yang organs has a uniting He-Sea point on the leg, which can be used to treat Yang organ problems. The three Yang channels of the feet also correspond to the three Yang channels of the hands. The Lower Six Lower He-Sea points are as follows: Stomach St-36, Large Intestine St-37, Small Intestine St-39, Triple Burners Bl-39, Bladder Bl-40, and Gall Bladder GB-34.

THE ELEVEN GHOST POINTS

In ancient times, these eleven points were traditionally used for spirit possession and have a specific effect on the mind's psychological functions. In traditional Chinese medical clinics today, these points are commonly used to treat emotional disorders. The Eleven Ghost points are as follows: GV-26, Lu-11, Sp-1, Pc-7, Bl-62, GV-16, St-6, CV-24, Pc-8, GV-23, and LI-11.

THE TEN WINDOW OF THE SKY POINTS

These ten points also have a specific effects on the mind's psychological functions and are also commonly used to treat emotional disorders. The Ten Window of the Sky points are as follows: Lu-3, LI-18, TB-16, Bl-10, St-9, CV-22, GV-16, SI-16, SI-17, and Pc-1.

SUMMARY OF POINTS

Understanding the origin and function of the body's energetic points, as well as the Qi flow along the energetic channels, allows the Qigong doctor to emit energy through these energetic cavities to easily access the tissues and internal organs of the patients body. It is through this understanding that the doctor can also lead and purge the Toxic or pathogenic Qi (which has become stagnant or detrimentally active within the patient's body) away from the diseased organs, through the channels and out the body.

The following graph (Figure 8.8) is a basic categorization of the body's points and their descriptions.

Category	Points	Description
Five Antique Shu Points	Jing-Well, Ying-Spring, Shu-Stream, Jing-River, He-Sea	Transporting and Command Points
Sixteen Xi Cleft Points	Yin: Lu-6, Pc-4: Ht-6, Sp-8, Lv-6, Kd-5, Yin Heel Kd-8, and Yin Linking Kd-9 Yang: LI-7, TB-7, SI-6, St-34, GB-36, Bl-63, Yang Heel Bl-59 and Yang Linking GB-35	Accumulation Points
Twelve Front Mu Points	Lungs Lu-1, Pericardium CV-17, Liver Lv-14, Spleen Lv-13, Gall Bladder GB-24, Kidneys GB-25, Heart CV-14, Stomach CV-12, Large Intestine St-25, Triple Burners CV-5, Small Intestine CV-4, and Bladder CV-3	Alarm, Collecting Points
Eighteen Back Shu Points	Lungs Bl-13, Pericardium Bl-14, Heart Bl-15, Governing Vessel Bl-16, Diaphragm Bl-17, Liver Bl-18, Gall Bladder Bl-19, Spleen Bl-20, Stomach Bl-21, Triple Burners Bl-22, Kidney Bl-23, Sea of Qi Bl-24, Large Intestine Bl-25, Gate to the Original Qi (Guan Yuan) Bl-26, Small Intestine Bl-27, Bladder Bl-28, Sacrum Bl-29 and Anus Bl.30	Associated, Transporting Points
Fifteen Luo Points	Yin: Lu-7, Pc-6, Ht-5, Sp-4, Lv-5, Kd-4, CV-15, the Great Luo Channel Sp-21 Yang: LI-6, TB-5, SI-7, St-40, GB-37, Bl-58 and GV-1	Connecting Points
Twelve Entry Points	The first point on each channel except for the Large Intestine Channel in which it is LI.4	Qi enters Channel
Twelve Exit Points	Yin: Lu-7, P-8, Ht-9, Sp-21, Lv-14 and Kd-22 Yang: LI-20, TB-23, SI-19, St-42, GB-41 and Bl-67	Qi exits Channel
Eight Influential Points	Yin Organs Lv-13, Yang Organs CV-12, Qi CV-17, Blood Bl-17, Sinews GB-34, Blood Vessels Lu-9, Bones Bl-11 and Marrow GB-39	Influence Internal Organ Function
Eight Confluent Points	Governing Vessel SI-3, Conception Vessel Lu-7, Belt Vessel GB-41, Thrusting Vessels Sp-4, Yang Heel Vessels Bl-62, Yin Heel Vessels Kd-6, Yang Linking Vessels TB-5 and Yin Linking Vessels Pc-6	Master and Couple Points
Twelve Yuan-Source Points	Gall Bladder GB-40, Liver Lv-3, Lungs Lu-9, Large Intestine LI-4, Stomach St-42, Spleen- Sp-3, Heart Ht-7, Small Intestine SI-4, Bladder Bl-64, Pericardium Pc-7, and Triple Burners TB-4	Source of Original Qi Points
Four Seas Points	Sea of Qi Points: Bl-10, St-9, CV-17, GV-14 and 15 Sea of Blood Points: Bl-11, St-37 and 39 Sea of Marrow Points: GV-15, 16, 17, and 20 Sea of Nourishment Points: St-30 and 36	Give Access to the Four Seas
Upper He-Sea Points	Stomach LI-10, Large Intestine LI-9 and the Small Intestine LI-8	Affect the Yang Organs
Lower He-Sea Points	Stomach St-36, Large Intestine St-37, Small Intestine St-39, Triple Burners Bl-39, Bladder Bl-40, and Gall Bladder GB-34	Affect the Yang Organs
Eleven Ghost Points	Governing Vessel 26, Lungs 11, Spleen 1, Pericardium 7, Bladder 62, Governing Vessel 16, Stomach 6, Conception Vessel 24, Pericardium 8, Governing Vessel 23, and Large Intestine 11	Affect the Mind, Emotion, and Spirit
Ten Windows of the Sky Points	Lung 3, Large Intestine 18, Triple Burners 16, Bladder 10, Stomach 9, Conception Vessel 22, Governing Vessel 16, Small Intestine 16, Small Intestine 17, and Pericardium 1	Affect the Mind, Emotion, and Spirit

Figure 8.8. Point Graph

CHAPTER 9
THE EXTRAORDINARY YANG ORGANS

Besides the six Yin organs (Liver, Heart, Pericardium, Spleen, Lungs, and Kidneys) and six Yang organs (Gall Bladder, Small Intestine, Triple Burners, Stomach, Large Intestine, and Bladder), the body has six Extraordinary Yang organs, also called the "curious organs." These Extraordinary organs function like Yin organs (storing Yin Essence, i.e., Blood, Marrow, or Kidney Essence) but look like Yang organs (they are hollow). These six curious organs include the Uterus, Brain, Marrow, Bones, Blood Vessels, and Gall Bladder. The body's Eight Extraordinary Vessels integrate the Extraordinary Yang organs with the main six Yang organs and the Kidneys.

THE UTERUS

In Chinese medicine, the uterus encompasses the fallopian tubes and the ovaries and has the function of regulating menstruation, conception, and pregnancy (Figure 9.1). The uterus connects to the Kidneys (which provide the uterus with Jing), the Conception Vessel (which provides the uterus with Qi and nourishes the fetus), and the Thrusting Channel (which provides the uterus with Blood). Only when the Jing of the Kidneys is sufficient can the menstrual period occur regularly, the woman become pregnant, and fetal growth become possible. The Qi and Blood of the Twelve Primary Channels pass into the uterus through the Thrusting and Conception Vessels, affecting the amount of menstrual flow and its cycle.

The three Yin organs of the Heart, Liver, and Spleen also energetically connect to the uterus through their relationships with the Blood. The Heart governs the Blood. The Liver stores the Blood and regulates the volume of circulating Blood, which is responsible for normal menstruation. The Spleen controls the Blood.

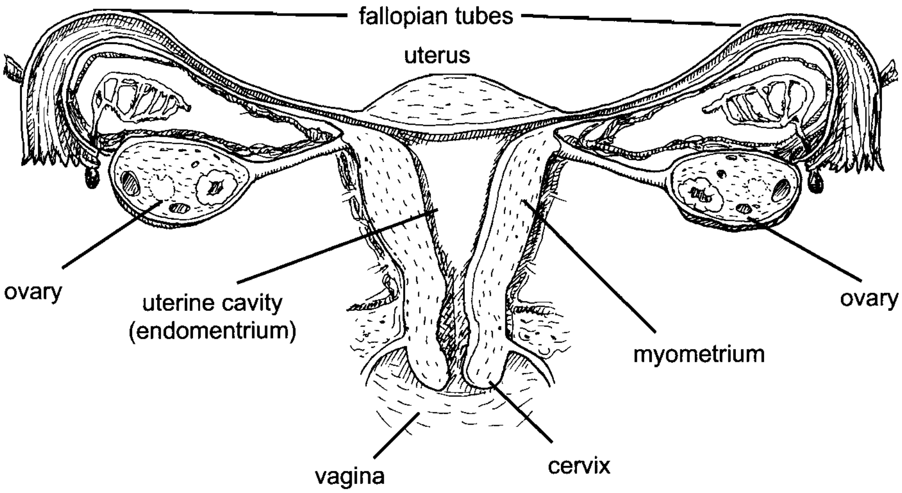

Figure 9.1. The Female Reproductive Organs, including the Uterus, Fallopian Tubes, and Ovaries

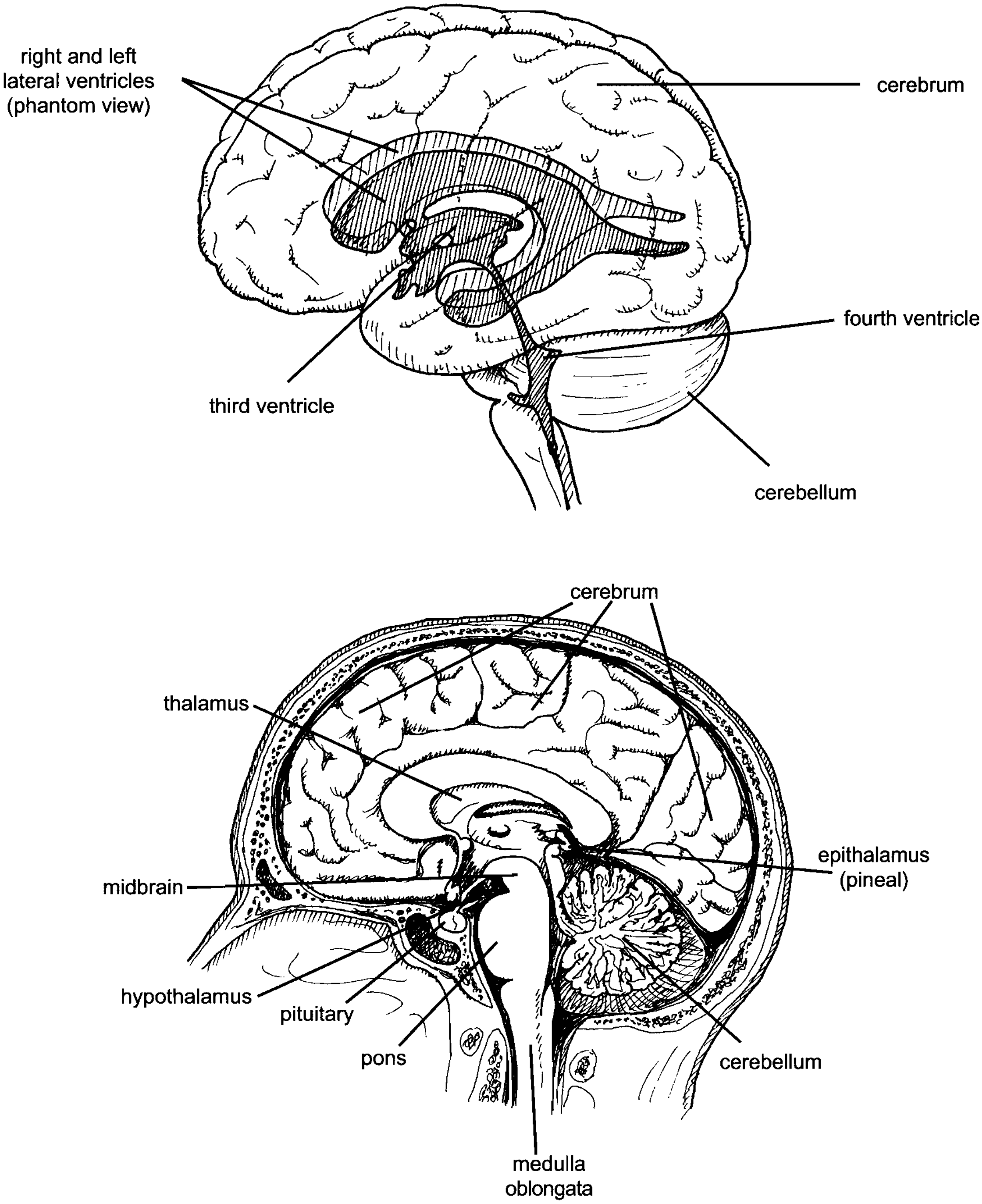
right and left lateral ventricles (phantom view)
cerebrum
fourth ventricle
third ventricle
cerebellum
cerebrum
thalamus
epithalamus (pineal)
midbrain
hypothalamus
pituitary
pons
cerebellum
medulla oblongata

Figure 9.2. The Human Brain

THE BRAIN

The Brain is considered the house of the Yuan Shen, and the seat of mental function. It is believed that the senses and control of the body's physical movements are related to the Brain. The Brain is functionally related to the Kidneys and controls memory, concentration, hearing, touch, sight, and smell. The Brain is also called the Sea of Marrow, extending from the top of the head to the GV-16 Wind Palace point, just below the external occipital protuberance. The Kidney Essence produces Marrow, which fills the spinal cord and Brain. The Brain also depends on the Heart's Blood for its nourishment. Because the Kidneys store the Jing and the Heart governs the Blood, the Brain depends on the balanced state of the Heart and Kidneys for vitality. The Brain is considered a transformation chamber of both Prenatal and Postnatal Essence (Jing) and Energy (Qi). The Brain is regulated by the Governing Vessel, as well as the Yin and Yang Heel Vessels.

This viewpoint is quite different from the Western approach, which views the brain's anatomy as consisting of five regions (Figure 9.2), including:

1. The cerebral hemispheres or cerebrum (the neocortex (neopallium),
2. The thalamus, hypothalamus, and epithalamus (pineal),
3. The midbrain/mesencephalon (the colliculi and cerebral peduncles),
4. The pons and cerebellum, and
5. The medullae oblongata.

In Western medicine, the oldest part of the Brain (situated at the top of the Brain stem) is often called the "reptilian Brain" because its anatomical formation is similar to that found in lizards, alligators, and turtles. Surrounding it is the mammalian Brain, which is called the limbic system. The third part of the Brain is called the neocortex, which wraps itself around the limbic system. All three parts of the Brain (reptilian, mammalian, and neocortex) are viewed as biologically distinct, both in their chemistry and in their structure. The older formations of the Brain are responsible for the autonomic nervous system, whereas the neocortex is responsible for thinking and voluntary movement.

In Chinese energetic medicine, each selective set of the human brain cells absorbs and records sets of vibrations. The Brain receives these vibrations through the skin, eyes, ears, nose, and mouth, and records them in its cells. The mind interprets these incoming vibrations as images, sounds, smells, and sensations, and organizes these into memories. The extent to which the mind interprets the incoming data depends upon the level of consciousness and awareness of the individual at the time of the recording. When a Qigong doctor, attunes his or her vibrations to the patient's energetic field and emits Qi, the brain cells responsible for storing the vibrations are stimulated. When these vibrations are stimulated through the energy projected by a Qigong doctor, the patient's tissues respond and reproduce an energetic pattern similar to the original one used to record the patient's physical, mental, emotional, energetic and spiritual experiences. This energetic tissue reaction occurs only because the patient's cells are held directly to their same consistent vibrational patterning.

One set of selective brain cells receives, records, and maintains the vibrations of the thoughts, actions, movements, and pictures that other forms project. These vibrations can also be reproduced and projected, being received by the Qigong doctor as images from within his or her own mind through inner-vision. The Qigong doctor can arrange his or her cells' energetic pattern in such a way as to reproduce the words and motions of these forms or objects, even the thoughts of those that sent them. Through these cells, the Qigong doctor can assist his or her patient, as well as him or herself, in controlling their thoughts, as it is through these cells that accidents and disease are brought into existence; for example, someone either observes or imagines a specific thing happening. The corresponding vibration is fixed in the cells, sent out to be imprinted on the corresponding cells of anothers' Brain, then projected back until the incident is so fixed that the occurrence actually happens (self-fulfilling prophecy).

Another set of selective brain cells receives, records, and maintains the vibrations of the thoughts and activities of the divine. The divine energy vibration pervades every substance and is always sending out divine knowledge and true wisdom. Stored within the body are the cells that receive and project the vibrations of the divine Mind. If the doctor separates him or herself from the divine in thought, he or she will also separate from the divine in manifestation.

THE GALL BLADDER

The Gall Bladder stores the bile, which is considered a "pure" fluid. The function of the Gall Bladder is to aid digestion, as well as drain off Excess Heat from the Liver (Figure 9.3). The Gall Bladder is regulated by the energy of the Belt Vessel.

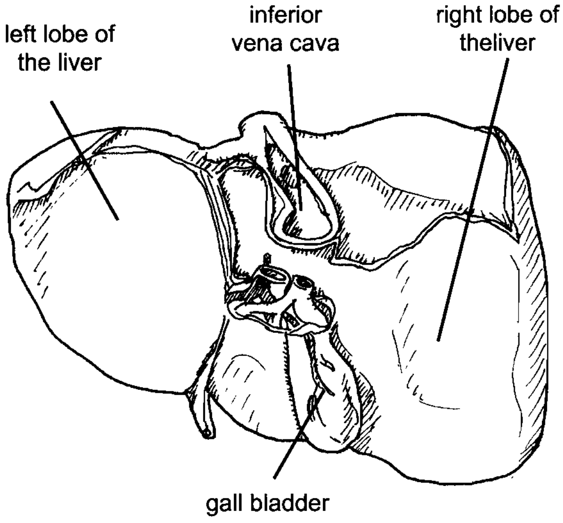

Figure 9.3. The Gall Bladder

THE MARROW

In Chinese medicine the term Marrow (Sui) is different from bone marrow as defined in Western medicine. The Marrow is the substance which is the common matrix of bones, bone marrow, the spinal cord and the Brain. The Marrow is rooted in the Jing and is connected to the Lower Dantian through the Governing Vessel. It is regulated by the energy of the body's Thrusting Vessels.

The Kidney Jing is the origin of the Marrow. The Marrow functions to form the bone marrow, as well as nourish the Brain and spinal cord. When the Mingmen warms and nourishes the body, the Marrow becomes full (which sustains the body's vertical posture and gives strength to both the Brain and the bones). When the Marrow is full, thinking is clear. Too much thinking leads to Excess Fire, which burns the Brain, causing dizziness, blurred vision, and tinnitus.

Both red and white blood cells are produced in the Marrow of the bones. The red blood cells (which circulate oxygen and eliminate carbon dioxide) are produced within the body's long bones (humerus, femur, tibia, etc.). The white blood cells (vital to the body's immune system) are produced within the body's flat bones (skull, sternum, scapulae, pelvis, etc.) The type of blood cells that are produced depends on the type of Marrow that predominates.

When all of the Postnatal energy acquired from the combination of physical (diet, food, and drink), mental/emotional (thoughts and emotions), and spiritual (prayer, meditation, and sex) cultivations is circulated through the Microcosmic Orbit (Fire Cycle), the Prenatal Qi stored in the bone marrow and Brain become stimulated, awaking the spiritual consciousness. This combination and circulation of physical, mental, emotional, and spiritual energy is why the brain marrow is considered to be constructed from the finest and most subtle essences, manifesting the original and hidden power of the Kidneys.

THE BONES

The bones are related to the Kidney's Jing. The bones support the body's structure, strength, and mobility, as well as store the bone marrow (Figure 9.4). The bones are extremely porous and are always "breathing," they are regulated by the energy of both the body's Conception and Thrusting Vessels. The pores of the bones allow Qi and Blood to be absorbed and released, the same way a sponge absorbs and releases water.

There is a natural union between the energetic function of the Marrow and the bones. The Marrow, held within the bones, assures the power and

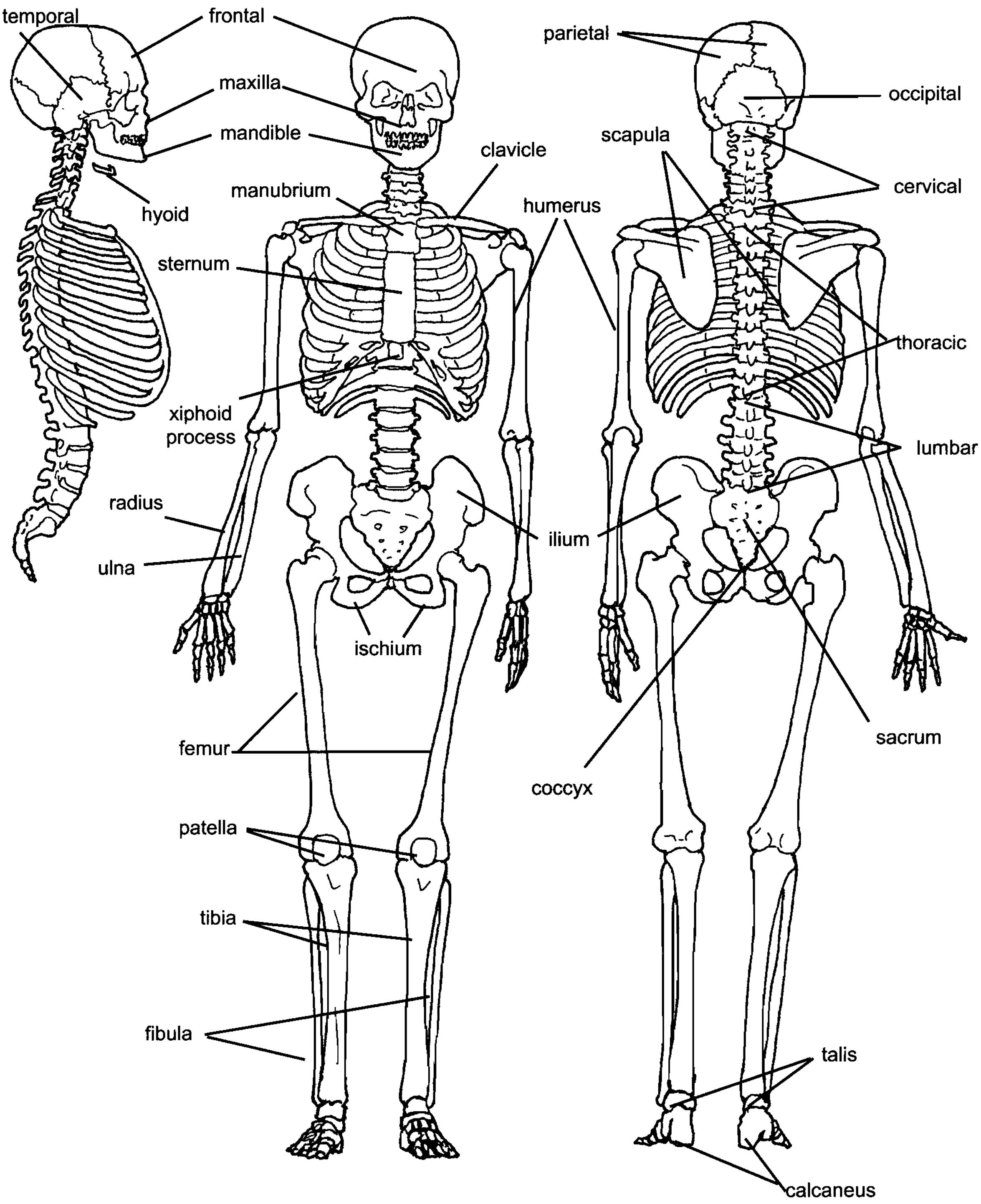
temporal
frontal
maxilla
mandible
hyoid
parietal
occipital
scapula
clavicle
cervical
manubrium
humerus
sternum
thoracic
xiphoid process
lumbar
radius
ilium
ulna
ischium
sacrum
femur
coccyx
patella
tibia
fibula
talis
calcaneus

Figure 9.4. The human skeleton consists of 200 bones.

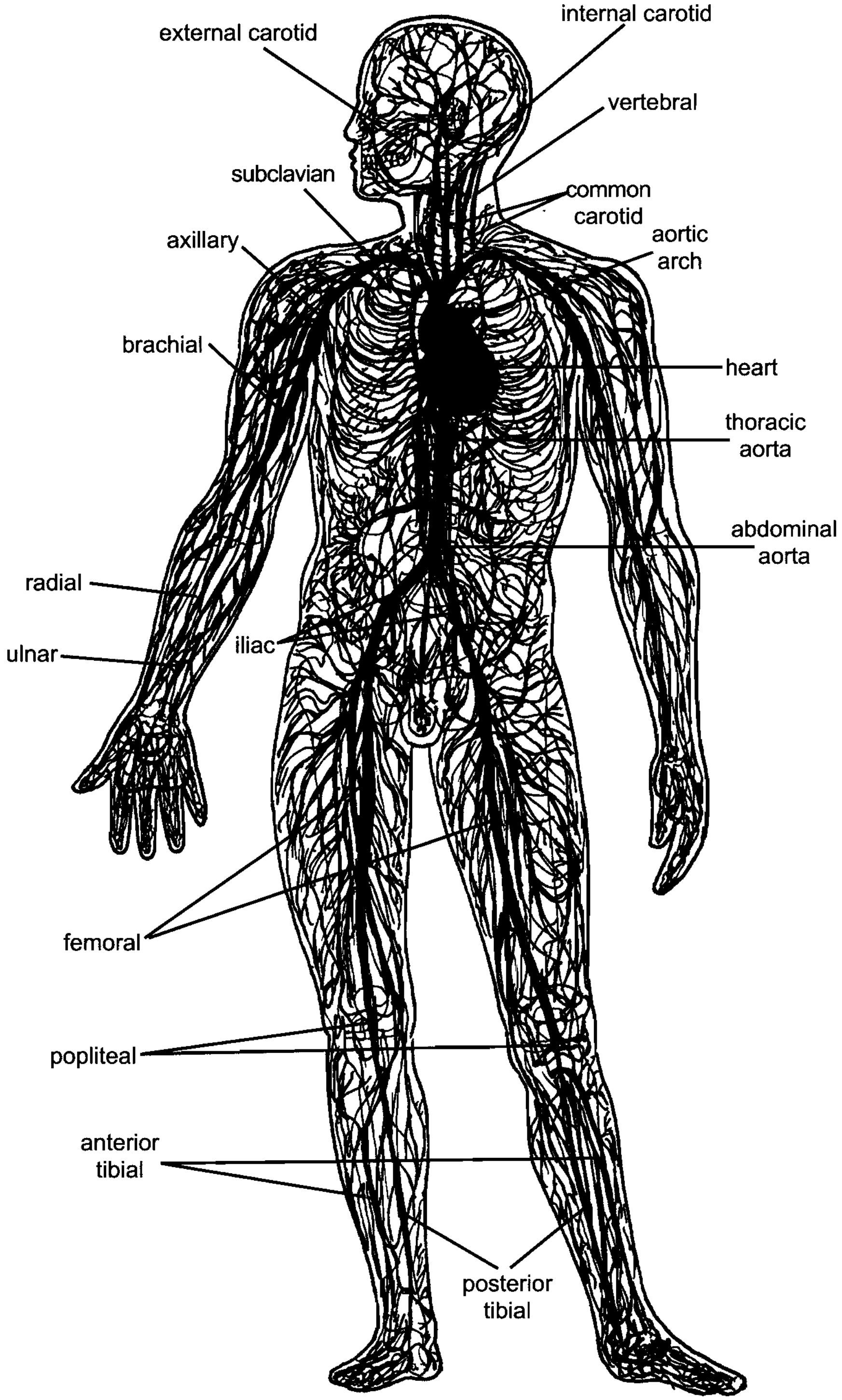
external carotid
internal carotid
vertebral
subclavian
common carotid
axillary
aortic arch
brachial
heart
thoracic aorta
abdominal aorta
radial
ulnar
iliac
femoral
popliteal
anterior tibial
posterior tibial

Figure 9.5. The blood-vascular system of the human body consists of the heart and blood vessels.

smooth suppleness of the bones. The bones, in return, prevent the dissipation of the essential richness of the Marrow.

The bones are the only solid crystalline substance in the body capable of generating piezoelectric charges. These electromagnetic charges generate fields of energy that receive and send impulses to the blood cells, tissues, organs, and channels. The Brain, nervous system, Heart, and lower abdomen also generate electromagnetic fields that resonate with the bones, as well as other crystalline-like structures. The crystalline structures of the bones amplify, radiate, and transmit energy, as well as bio-information, to the rest of the body. The rhythmic oscillation of bone-generated electromagnetic fields is released through the transmission of the bones acting as tuning forks. These living tuning forks vibrate the Qi throughout the body's entire physical structure, extending outward into the body's external Wei Qi fields.

THE BLOOD VESSELS

In Chinese medicine, the concept of Blood is different in characteristic and function from that of Western medicine. Blood originates from the transformation of food and drink by the Spleen, which then transfers the refined food energy to be further enhanced by the Heart Qi, Lung Qi, and the Qi of the air.

Qi and Blood flow together, with Qi being the active force that makes the Blood circulate and keeps it within the blood vessels (Figure 9.5). Both Qi and Blood flow within the channels and blood vessels, continuously circulating throughout the body to nourish, maintain, and moisten the tissues. Qi is an energetic form and is considered a Yang substance, while Blood is a liquid form of energy and is considered a Yin substance.

The blood vessels contain Blood and are indirectly related to the Kidneys, because the Kidney Essence (Jing) produces Marrow, which contributes to the production of Blood. The function of the blood vessels is to transport Qi and Blood throughout the body for nutrition and regeneration. The blood vessels are regulated by the energy of the body's Thrusting Vessels, and the Blood is contained in the blood vessels by the Spleen.

SECTION II

THE THREE OUTER FORCES: HEAVEN, EARTH & MAN

CHAPTER 10

INTRODUCTION TO THE THREE OUTER FORCES

THE NATURAL POWER OF THE THREE OUTER FORCES

The Three Outer Forces (Powers) of Heaven, Earth, and Man represent the sum total of all natural forces and energetic factors, which affect the human body on all levels of existence (physical, mental, emotional, energetic, and spiritual). In ancient times, the relationship between the waxing and waning of Heavenly (universal) and Earthly (environmental) energy, as well as the corresponding shifts, changes, and adjustments occurring within the human body, provided the Chinese sages with the theory of the Three Outer Forces. These ancient Chinese masters discovered that the human body's internal network of systems is directly affected by the influences and variations of energetic cycles of the sun, moon, and stars. We now know that these natural influences also affect atmospheric pressures, causing high and low temperatures, which in turn, affect the body's tissues. Recently, Ukrainian physicists found that human DNA vibrates at the same frequency as solar energy.

The Heaven and Earth Powers are responsible for the energetic field changes that cause adjustments and shifts within man's Jing, Qi and Shen (known as the Three Treasures), stored within the internal organs and tissues. Together the combination of Heavenly and Earthly influences affect the functional status of the body's internal organs and the distribution of Qi and Blood throughout the body.

Each of the body's cells, tissues, organs, and organ systems emanates its own specific electromagnetic energy field. This energetic field pulsates at its own particular frequency, regulating the body's internal energy currents and radiating an energetic field that extends outside the body. The body's energetic field (Wei Qi) interacts with and is influenced by the larger energetic fields surrounding it (i.e., power lines, crowds of people, geological formations in the Earth, and the cosmic fields of the solar system).

THE THREE OUTER FORCE'S NINE TREASURES

Each of the Three Outer Forces manifest through a secondary triad of Treasures. These Three Treasures complete the energetic matrix that encompasses the Outer Forces unified energetic field.

1. The Three Treasures of the Heavenly Power are the energy fields of the sun, moon, and stars;
2. The Three Treasures of the Earthly Power represent the energy of the Earth (soil and rock), wind, and water;
3. The Three Treasures of Man are contained within man's Essence (Jing), Energy (Qi), and Spirit (Shen).

These three intricate networks exchange their energetic influence and affect the functional status, as well as the distribution of the body's Qi and Blood.

THE THREE OUTER FORCES AND CHINESE ALCHEMY

Each of the Three Outer Forces and their Three Treasures also constitute the basis of Chinese internal alchemy, whereby Jing (Essence) is transformed into Qi (Energy), Qi is transformed into Shen (Spirit), and Shen is transformed into Wuji (infinite space), and then released back to the Dao (or divine).

The Powers of Heaven and Earth are not only interlinked, but also exist as a microcosm or "small universe" within man's physical body. The human body is an active interaction of two basic elements of Yin and Yang polarity. Yang energy ascends and is expressed through expansive, active, hard, and aggressive energetic movements. Yin energy descends and is expressed through contractive, passive, soft, and yielding movements. All matter, from the smallest molecular particles to enormous planets and stars, is composed of energy and bound into various patterns of vibration. These matter formations are held in place by electromagnetic and nuclear forces interconnected by the power of Yin and Yang polarity, and are expressed through the Three Outer Forces. The seasonal cycles are an example of Yin Earthly movements, while the solar and lunar cycles are expressions of Yang Heavenly energy.

The creation and manifestation of matter is composed and regulated by what are known as the Five Elemental patterns. In traditional Chinese understanding, it is believed that perfect health and longevity depend largely on man's ability to harmonize with the transitions and five seasonal changes of Heaven and Earth, in accordance with the Five Elemental patterns. In the winter, for example, there is less sunlight, and the temperature can be very cold (Yin), and the Kidneys (Water Element) can easily be overworked; therefore, man should endeavor to lighten his work load and rest more in winter, conserving his Qi to avoid energy depletion.

The principles of Chinese cosmology, Yin and Yang Theory, Five Elements Theory, the *Book of Changes (Yi-Jing),* and Traditional Chinese Medicine are all incorporated into the study of Medical Qigong. This culmination of knowledge reveals the basic structures of energetic dynamics relating to the universal and environmental energy gathered from the Heavens and Earth, and their relationship with man.

HEAVENLY QI

The first of the Three Outer Forces is known as the Universal Force, Original Force, or Heavenly Energy. This force manifests as the energy of the entire Cosmos. Each life born into human form is not only developed and structured but is also influenced by its own unique group of stars within the Heavenly realm. This particular group of stars (including the sun, moon, planets, and constellations), controls or influences the persons life-force energy (Figure 10.1).

Heaven is the realm that predetermines the energetic cycles of birth and death, as well as the seasons and times of planting, fertility, and harvest. It is the source of the Yuan Shen and the Virtue of Dao, as it manifests through universal awareness (divine inspiration, enlightenment and spiritual communion), and free will. Heaven is the residence of spiritual deities and the Jade Emperor (or Divine Creator), and it manifests its prenatal power in man's life through the mysterious forces of fate, destiny, karma, and reincarnation.

According to the ancient Chinese, we exist in a three-dimensional world that exists within the relative parameters of time and space. Heaven represents time, which is symbolized by the circle. Because time moves in circles, hours and days are defined by the circular rotation of the Earth on its axis; months are defined by the circling of the moon around the Earth, and years are defined by the Earth's orbit around the sun.

PHILOSOPHICAL CORRELATION

Man is constantly absorbing energy from the sun, moon, planets, and constellations. In order to explore the patterns of Heavenly energy in detail, the ancient masters developed the science of Chinese astrology, which includes the study of Twelve Animals and Five Elements, as well as the interactions of the Twenty Eight Constellations, Twelve Earthly Branches, Ten Heavenly Stems, Five Planets, and Lunar and Solar Energy Cycles.

The ancient sciences of astrology and divination evolved as a means to analyze, interpret, and predict how the various forces of Heaven influence and guide man's life on Earth. The study of Chinese astrology involves four elements of comprehension: the study of the Signs, the study of the Five Elements, the study of Yin and Yang, and the study of the Sixty-Year Cycle.

1. The Study of the Signs involves the twelve animals of the Chinese zodiac (meaning circus of animals) that encircle the Earth and are divided into twelve equal parts, called "signs." Each sign of the zodiac has certain

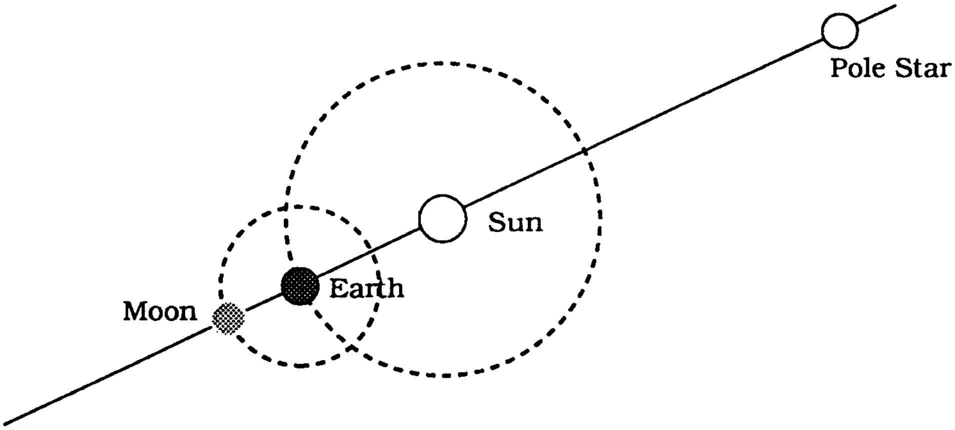

Figure 10.1. The Three Outer Forces of Heaven Consists of the Energy Released from the Sun, Moon and Stars.

characteristics that are determined by a particular planet and constellations. The signs affect the character of a person born under them. Each person is given three animal characteristics: one according to the year, the second according to the month, and the third according to the hour of his or her birth. You might, for example, be a Snake according to the year of your birth, a Boar according to the month, and a Dragon according to the hour. The animal governing your birth year represents your external self (the part you show others) and is said to exert a profound influence over your life. The month animal (known as the lunar animal) represents the emotional and psychological aspects of man's relationships. The hour animal represents the inner self (the part hidden from the rest of the world). The combination of the three determines the uniqueness of each individual.

2. The Study of the Five Elements is the next division involved in the study of signs. Each of the twelve animal signs is further differentiated into one of the Five Elements (Wood, Fire, Earth, Metal, and Water). The year of birth determines the element of the sign. Each element rules for a two-year cycle (e.g., January 27, 1952-February 13, 1953 was the year of the Dragon under the element of Water; February 14, 1953-February 2, 1954, was the year of the Snake under the element of Water. The next Water cycle of Dragon and Snake will be in 2012, based on the sixty-year cycle). These Five Elements display an energetic force that affects the formation of the internal organs in the growing fetus, as well as influences the physical, mental, emotional, energetic, and spiritual changes after the child is born.
3. The Study of Yin and Yang is the third division of the study of signs. Each of the twelve animal signs is also governed by a Yin (passive) or Yang (active) direction. The Yin and Yang cycles alternate each year, with Yang being the even years, and Yin the odd years.
4. The Study of the Sixty-Year Cycle is the final division of the study of signs. The yearly correspondence of the twelve animals of the Chinese zodiac also relates to the Twelve Earthly Branches, and is based on a sixty-year cycle. Every sixty years, the sun, moon, Earth, and the North (Pole) Star are in alignment, and a new sixty-year cycle begins. A sixty-year cycle also completes the rotation of the twelve animals through each of the Five Elements. The Chinese consider sixty years a "century" in terms of the human life cycle.

PHYSICAL CORRELATION

Heavenly Qi relates to the congenital strength and formation of the body's internal organs, as determined by the Universal Energetic Fields (which include the energy emitted from the sun, moon, planets, and Twenty-Eight Constellations).

HEAVEN'S ENERGETIC GRID

Similar to the grid system imprinted on the body's energetic field, there exists a massive interconnected Heavenly grid extending throughout the entire universe. This Heavenly grid connects the sun, moon, stars, planets, and constellations throughout our universe, and extends to other solar systems outside our celestial sphere. There are pockets of Heavenly energy (like energetic points), as well as subtle frequency pathways (similar to energetic channels), that follow a progression of universal order.

EARTHLY QI

The second of the Three Outer Forces is known as the Environmental Force, Acquired Force, or Earthly Qi. This force manifests as the energy of the entire planet. Each life born into human form is not only developed and structured but also influenced by its own unique environment within the Earthly realm. The Qi of the individual's particular geographic location includes the energy of the plants, soil, water, animals, and all natural formations responsible for controlling or influencing the person's life-force energy.

The Chinese believed that the Earth is the realm that predetermines the abundance of harvest, as well as the temperance of man's physical stamina. Both harvest and man's physical stamina is determined through the cyclic changes of the five seasons and their relationship to the Five Phases (birth, maturation, peak, decay, and death). The Earth is the source of the natural forces and rhythms of nature, as well as the basic elements and energies that compose the material world of soil and water, mountains and valleys, oceans and rivers, plants and animals. The Earth is the foundational base of the Five Elements through the cyclic transformations of nature, which in turn, shape the human body's physiological form and govern the biological functions of the tissues.

The ancient Chinese viewed the Earth as the center, and as the ground on which we stand looking outward toward the four directions. Just as Heaven defines our movement through life in terms of time, Earth defines the space in which we live, providing the energetic field for our unfolding, within the boundaries of time.

In Chinese symbology, the Earth is symbolized by the image of the square, whose walls face the four cardinal directions. This concept of the energetic interaction of Heaven (symbolized by the circle) and Earth's (symbolized by the square) was so fundamental to ancient Chinese thought that they even designed the coinage of the imperial realm to reflect this interconnection (a round coin with a square center).

PHILOSOPHICAL CORRELATION

Earthly Qi relates to the study of Feng Shui (Harmony of Wind and Water). Feng Shui is divided into two prominent schools of study, the Form School and the Compass School.

1. The Form School is the original school of Feng Shui, originating in Southwest China. This school focuses primarily on developing environmental harmony, by observing the shape and form of the terrain in conjunction with the Five Elemental animals.
2. The Compass School is the second school of Feng Shui. It focuses primarily on developing universal and environmental harmony through the directional orientation of the compass, in relationship to the Ten Heavenly Stems, Twelve Earthly Branches, Twenty-Four Solar Compass Directions, and Twenty-Eight Constellations, as well as the Magic Square and the Bagua (Eight Trigrams).

PHYSICAL CORRELATION

Earthly Qi relates to the "acquired" strength (muscular and skeletal development after birth), as well as the formation of the body's internal organs. Tissue development is determined by the environmental energetic fields (which include the energy emitted from the Earth, Water, and Wind/weather). The Earth energy also expresses its power in man's life through the basic instincts for survival and the primal drives for sex and reproduction.

EARTH'S ENERGETIC GRID.

Similar to the Heavenly energetic grid system imprinted in the cosmos, there exists a massive interconnected Earthly grid extending throughout the entire planet. This Earthly grid connects the planet's eight energetic directions into one energetic field. There are pockets of Earthly energy (like energetic points), as well as subtle frequency pathways (similar to energetic channels), that follow a progression of environmental order. One example of the Earth's axial grid system in Western folklore is the "Ley lines," which are energetic pathways that connect energy vortices or places of power on the planet. This understanding of an Earthly energetic grid system is also reflected in American Indian and South American traditional understanding.

MAN'S QI

The third of the Three Outer Forces is known as the Human Force, or Man's Energy. This force manifests as the energy of the entire body (Figure 10.2). Each life born into human form is not only developed and structured, but also influenced by the individual's adaptations to his or her own physical, mental, emotional, energetic, and spiritual energies.

It is believed that within man, the realm of the Dao or divine conscious spirit exists, wherein both the energies of Heaven and Earth coexist in harmonious balance. Man is formed from the cosmic particles of exploded stars that have descended onto the Earth. These particles are absorbed into the soil, water, and foliage, creating sustenance for human life. This constant internal connection, absorption, and regeneration of vibrating energy from Heaven and Earth is the reason that human life is considered to be the highest physical manifestation of the divine energetic field (see Chapter 1).

PHILOSOPHICAL CORRELATION

Man's Qi relates to the study of the Yi-Jing (I-Ching), or Book of Changes. The Yi-Jing is composed of two sections: the Book of Oracles and the Book of Commentaries.

1. The Book of Oracles is the original text by King Wen and the Duke of Zhou. This book is based on the formation of eight trigrams and their various energetic attributes. These eight trigrams combine to form the sixty-four hexagrams of the Yi-Jing, which were later used for divination.
2. The Book of Commentaries is based on the commentaries of Confucius and his followers, and is also called the Ten Wings. The commentaries are called "wings" because each

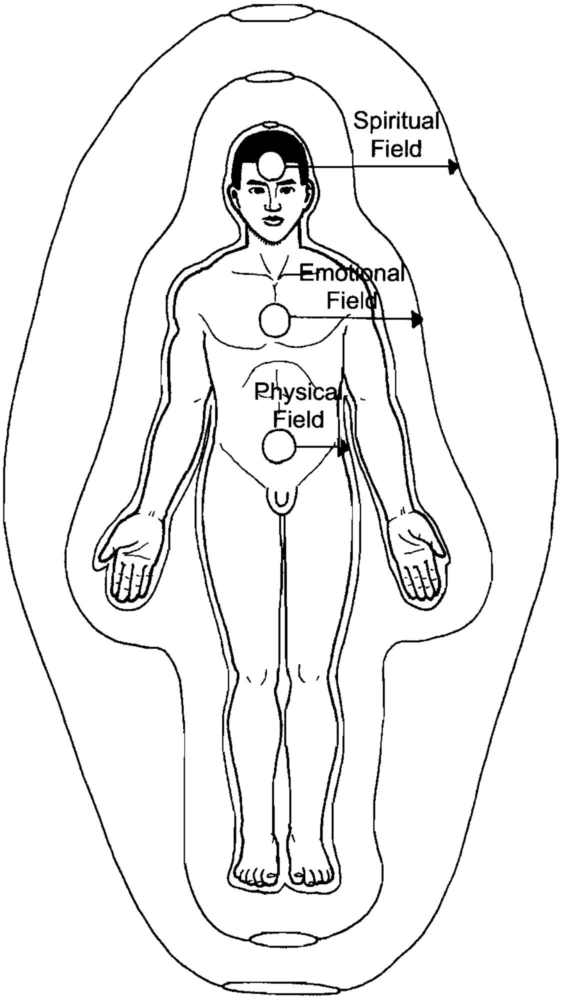

Figure 10.2. The Three Outer Forces of Man's Energy is composed of the Energetic Fields Released from Jing, Qi, and Shen.

wing helps the reader to soar higher in comprehension of the original text.

PHYSICAL CORRELATION

Man's Qi relates to the waxing and waning of man's physical, mental, emotional, energetic, and spiritual biorhythmic cycles.

A Qigong doctor must become proficient at understanding, respecting, and adopting these laws of natural variations between nature and man. Through this understanding the doctor can attain a deeper comprehension of the principles and application of Medical Qigong Dao Yin training (mind concentration methods, respiration regulation, and postural manipulation).

MAN'S ENERGETIC GRID

Similar to the Energetic grid system of Heaven and Earth, the human body is imprinted with a massive interconnected grid that extends throughout the body's tissues into the environment. The body's energetic grid is responsible for transforming energy frequencies from cell to cell, tissue to tissue, and from the body's internal energetic field to its external energetic field. These are known in Traditional Chinese Medicine as the 'Jing Luo' channels and collaterals (or the meridian system). The body's energetic grids are also responsible for maintaining the production, growth, and development of energy and tissue formation, and are influenced through intention, thought, and emotion.

CHAPTER 11
THE THREE TREASURES OF HEAVEN: SUN, MOON, AND STARS

DEFINING HEAVEN QI

The Outer Force (Power) of Heaven's Three Treasures (the Sun, Moon, and Stars) manifests as electromagnetic radiation from the sun, reflected light and energy from the moon, starlight energy, and the gravitational forces and interactions of the above mentioned heavenly bodies. The Outer Force of Heaven's Three Treasures also include the twenty-eight constellations, five major planets, and their energetic and gravitational forces.

Heaven Qi is the energy emitted from the universe outside of the realms of Earth and man. The sun and moon, first five planets (Jupiter, Venus, Saturn, Mars, and Mercury), North Star, Big Dipper, and the twenty-eight constellations are the most easily observed natural Heavenly energies, which have a direct connection to life on Earth. The diagram below (Figure 11.1) shows the Five Elements relationship between the universal treasures of Heaven, the environment of Earth and Man's internal organs.

The sun and moon are the primary Heavenly manifestations of Yin and Yang energy. The sun is Yang and is considered the Spirit Soul (Hun) of the sky. The moon is Yin and considered the Corporeal Soul (Po) of the Earth. This interaction of Heavenly Yin and Yang energy affects and governs the Earthly Yin and Yang energy, as well as the Qi within the human body. The sun and moon each have their own cyclic rhythms and energies, changing their relative degrees of Yin to Yang at different times of the day, month, and year.

To cleanse, purify, and strengthen the body, Qigong doctors absorb the essence of the Heavenly Qi by guiding the star's energy into certain areas of the body. The energy of the stars may also be used to reinforce the circulation of Wei Qi for

Organ	Element	Time of Day	Direction	Planet	Associated Animal
Liver	Wood	Early Dawn	East	Jupiter	Green Dragon
Heart	Fire	Midday	South	Mars	Red Phoenix
Spleen	Earth	Midafternoon	Center	Saturn	Yellow Phoenix
Lungs	Metal	Late Dusk	West	Venus	White Tiger
Kidneys	Water	Night	North	Mercury	Indigo Tortoise

Figure 11.1. Natural Heavenly Transformations

added protection against external pathogenic invasion.

INFLUENCES OF HEAVENLY QI ON THE HUMAN BODY

The body's physical fluctuations work in harmony with the cycles of the sun, moon, and nature. The science that deals with the study of these biological clocks is known as chronobiology. The levels of corticosterone and plasma adrenocorticotropic hormone (ACTH), for example, follow a cyclical pattern during the course of a day. Generally, the body's emotional and menstruation cycle in women follow the changing patterns of the moon (i.e., approximately every 28 days). Additionally, approximately every 23 days, the male body releases stored toxins, changing the smell of the sweat in accordance with the body's physical biological cycle. The smooth functioning of natural biological rhythms can at times be disrupted. The "jet lag" phenomenon, for example, is due to the interruption of circadian rhythms in the body.

There are daily, weekly, monthly, and annual solar and lunar cycles to which the human body responds. Other human physical, emotional, and mental cycles are strongly influenced by the types of life-style chosen, and other environmental factors involved. The body's circadian rhythms influence not only the immune system, but also the Yin and Yang energies of the internal organs. The activities of the cells, glands, Kidneys, Liver, and nervous system are coordinated with one another and with the day-night rhythm of the environment. Hunger, excretion, and body temperature are influenced by the body's ability to respond to the micropulsations of the subtle energy fields within the internal organs and tissues, as well as the body's biological clocks.

In China, clinical scientist Liu Bing, from Beijing Medical University, notes in his research that the human body undergoes rhythmic variations within a 24-hour time period. These physical changes include temperature, oxygen consumption, blood pressure, pulse rate, content of hemoglobin, blood sugar, blood amino acids, blood adrenocortical hormone, hepatic glycogen, metabolic rate of fat absorption, and pace of cell metabolism. The natural clockwork of the body can also be altered by food, drink, drugs, and abnormal sleeping patterns.

BODY RHYTHMS AND ENERGETIC PATTERNS

The prenatal and postnatal energetic patterns of the body are also determined by the time and place of conception. The time and place of conception establishes the patient's spiritual energy and ancestral traits (food preferences, manner of dress, preferences of art, spiritual beliefs, and so on). The time and place of birth also determine the patient's energetic biorhythms (physical, emotional, and intellectual cycles).

PRENATAL ANCESTRAL TRAITS

The time and location of conception determine the nature of the ancestral spiritual influence on the patient. As the divine infuses the fetus's soul, the energy of the Eternal Soul combines with the environmental spiritual influences of the geographic location (e.g., the Orient, Europe, North America, etc.). This infusion of geographic spiritual energy creates within the patient a predisposition towards specific ancestral traits and cultural attractions.

At conception and throughout pregnancy, the mother absorbs the natural environmental energy through respiration (breathing through the mouth, nose, and pores), ingesting food (grown in the soil), and visual/auditory ingestion (observing and experiencing the cultural environment). The energetic history of each cultural environment exists within the memories stored in the natural energetic fields contained within that environment. Therefore, someone conceived within a particular geographic location in the Orient may find him or herself, unconsciously drawn to the social and cultural influences of that particular Asian population. This subconscious attraction is considered a natural phenomenon due to the geographic influences of the country's spiritual/ancestral power, causing an effect, indirectly, on the individual.

Physical Cycle

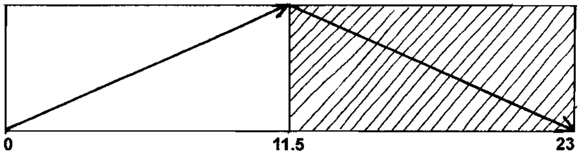

The Physical Cycle is 23 days long. The first 11 1/2 days are the positive side of the cycle. One experiences a feeling of good physical strength and endurance. The second 11 1/2 days are the negative side of the cycle, marked by having less endurance and a tendency towards fatigue.

Emotional Cycle

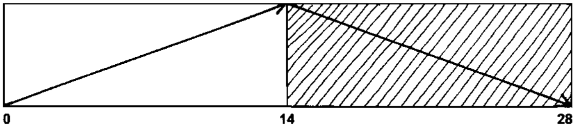

The Emotional Cycle is 28 days long. The first 14 days are the positive side of the cycle. One feels optimistic, cheerful, and cooperative. For the second 14 days, a negative stage of the cycle results in tendencies to be moody, irritable, and pessimistic.

Intellectual Cycle

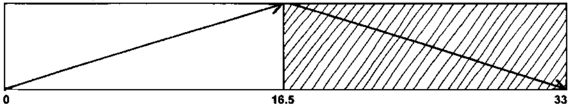

The Intellectual Cycle is 33 days long. The first 16 1/2 days are the positive side of the cycle. One has greater success in learning new material and pursuing creative, intellectual activities. The next 16 1/2 days are the negative side of the cycle; one is encouraged to review old material rather than attempting to learn new concepts.

Figure 11.2. The physical, emotional, and intellectual cycles begin on the day of birth.

1. The natural environment of the location at conception exerts a strong influence, such as a preference, or attraction to similar geological formations throughout the individual's life. People conceived by the ocean, for example, may find within themselves an unconscious need to live by the ocean. Likewise, people conceived in the mountains, valleys, tropics, deserts, etc. may find an energetically calm feeling of peace envelop them when visiting such places.
2. The time of conception also exerts a strong influence. The energetic formation, as well as strength or weakness, of the fetus's internal organs is also determined by the positions of the sun, moon, and stars. These Heavenly energies begin their effect on the fetus at the time of conception and continue influencing the fetus's formation throughout the entire pregnancy.

BIRTH-POSTNATAL ENERGETIC PATTERNS

The time and location of birth determines the biorhythmic patterns of each individual. These biorhythms are developed according to the influence of the Heavenly energy (position of the sun, moon, and stars), as well as the Earthly energy (i.e., geographic location, e.g., mountains, valleys, by the ocean, etc.). This infusion of the Earth's geographic energy and Heaven's constellation energy creates within the individual a predisposition towards specific psychological traits.

1. The location of birth strongly influences the individual's desire to seek out and/or surround him or herself with the familiar environment.
2. The Time of Birth may also have an unconscious influence on his or her living patterns, as certain individuals born at night naturally become "night people," while others born in the morning function better during the morning hours.

The body's biorhythms, which react to universal and environmental energetic vibrations, are divided into three distinct cycles and energy flows (Figure 11.2). Each rhythm is cyclical, waxing and waning, creating and contributing to times of physical, emotional, and intellectual triumph, as well as times of reflection and withdrawn behavior. These three cycles begin at the moment of birth and continue with absolute regularity until death. The conditions of the cycle are divided into positive (the first half) and negative (the second half) attributes.

The body's biorhythms are but one example of the effects that Heavenly energy has on the physical, emotional, and intellectual cycles of mankind. By understanding the cycles of the sun, moon, and stars and their energetic potential, the Qigong doctor can utilize the universal energy of Heaven to regulate and balance the patient's Qi. The biological rhythms influence the time at which many illnesses occur or worsen, as well as how fast a medication takes effect and how long the effect lasts. Therefore, Medical Qigong therapy, as well as herbal therapy are prescribed at specific times to enhance their effect on the patient's body.

CYCLES OF THE SUN

The Earth's rotation around the sun causes seasonal changes and instigates cyclic metabolic patterns in the human body. The body's entire metabolic function is influenced by its ability to respond to the seasonal micropulsations of these subtle energetic fields.

The Chinese use both a solar and a lunar calendar. The Heavenly cycle of the moon has 12 lunar month divisions, which correspond to man's 12 internal organs and channels: Lungs, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidneys, Pericardium, Triple Burners, Gall Bladder and Liver. The Heavenly cycle of the sun has 12 periods of the day, which correspond to man's 12 sections of the body. These 12 internal cycles are sometimes called the 12 Blood / Heat Cycles because they affect the Qi, Blood and temperature within the Twelve Skin Zones (see Chapter 7).

The Earth's rotation and the rhythm of the five seasons (the usual four plus late summer) influence the body's internal organs and channels, affecting their energetic activity. In each of the five seasons, a specific pair of internal organs reaches an energetic peak.

1. In the summer, the Heart and Small Intestine are at their energetic peak.
2. In the late summer, the Spleen and Stomach reach their energetic peak.
3. In the autumn, the Lungs and Large Intestines reach their energetic peak.
4. In the winter, the Kidneys and Bladder reach their energetic peak.
5. In the spring, the Liver and Gall Bladder reach their energetic peak.

THE TWELVE PI HEXAGRAMS

The Twelve Pi Hexagrams symbolize the seasonal changes, which cause the waxing and waning of Yin and Yang energy. The trigrams and hexagrams are used to illustrate the movement of energy in Man (see Chapter 5) as well as Heaven. The Twelve Pi Hexagrams were also used to study and comprehend the interactive energetic fields of man and nature that govern growth and development. The interactive energetic fields engender the creation of energy channels and influence the circulation and transformations of Qi, which in turn, create and maintain the body's internal organs. The ancients believed that the human physique reflects the structure of the universe, and that prenatal Heaven and Earth energy fuse with postnatal Heaven and Earth energy at birth (see Chapter 2). The trigrams reflect the twelve hours of the day, the daily fluctuations of Yin and Yang energy, and the Qi of the Five Elements (Figure 11.3).

The Twelve Pi Hexagrams represent "the laws of hardness and softness," domination and yielding of Yin and Yang, the relationship between the seasons, time of day, major organs, channels, and the transformation and circulation of Qi.

THE YANG YAO CYCLES

When studying the Twelve Pi Hexagrams, it is important to examine the first six (from Fu — 11 p.m. to Qian — 11 a.m.), that represent the gradual increasing of the energetic potential of the Yang Yao (see Chapter 5) and the gradual decreasing of the Yin Qi. The Yang Qi rises and expands during this time period; it is the best time for training techniques that emit external Qi and for treat-

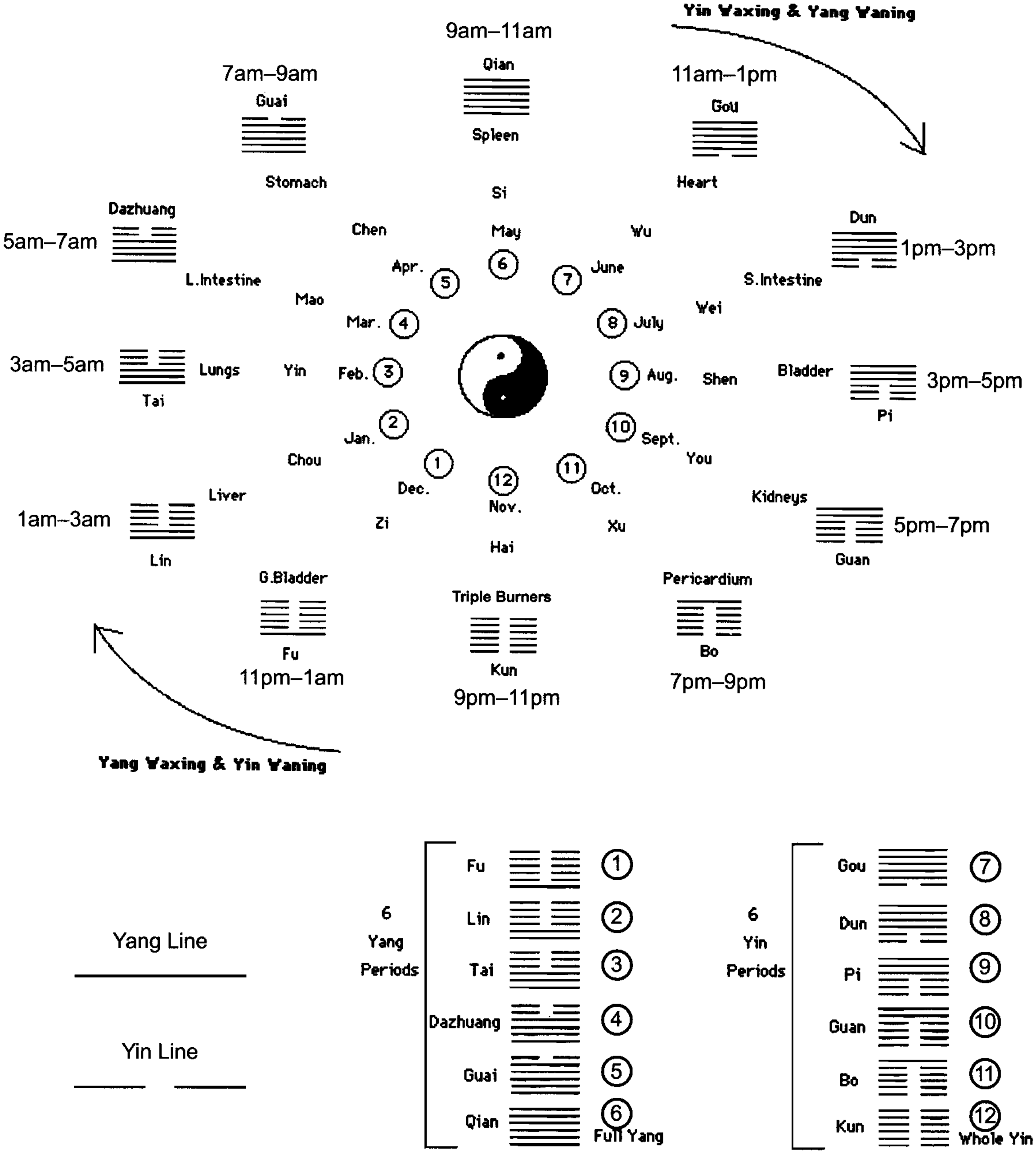
Yin Waxing & Yang Waning
9am–11am
Qian
Spleen
7am–9am
Guai
Stomach
11am–1pm
Gou
Heart
Si
Dazhuang
5am–7am
Chen
May
Wu
Dun
1pm–3pm
L.Intestine
Apr.
⑤
⑥
⑦
June
S.Intestine
Mao
Mar.
④
⑧
July
Wei
3am–5am
Lungs
Yin
Feb.
③
⑨
Aug.
Shen
Bladder
3pm–5pm
Tai
Pi
Jan.
②
⑩
Sept.
Chou
You
①
⑪
Liver
Dec.
⑫
Oct.
Kidneys
Nov.
1am–3am
Zi
Xu
5pm–7pm
Hai
Lin
Guan
G.Bladder
Pericardium
Triple Burners
Fu
Bo
11pm–1am
Kun
7pm–9pm
9pm–11pm
Yang Waxing & Yin Waning
Yang Line
Yin Line
6 Yang Periods
Fu ①
Lin ②
Tai ③
Dazhuang ④
Guai ⑤
Qian ⑥ Full Yang
6 Yin Periods
Gou ⑦
Dun ⑧
Pi ⑨
Guan ⑩
Bo ⑪
Kun ⑫ Whole Yin

Figure 11.3. The circular figure depicts the Heavenly cycle of the sun and the Twelve Pi Hexagrams. Solid lines in the trigrams and hexagrams represent Yang, and broken lines represent Yin. The hexagrams on the left depict the Yang Yao Cycle (from Fu to Qian), with the gradual increase in Yang and decrease in Yin, so that the last hexagram, Qian, is full Yang. The hexagrams on the right depict the Yin Yao Cycle (from Guo to Kun), with the gradual increase in Yin and decrease of Yang, so that the last hexagram, Kun, is full Yin.

ing patients who suffer from insufficiency of Yang Qi. This time period is called the "time of the living breath," and represents the beginning of energetic harvesting. When training in the morning time, the practitioner should progress from still, relaxed, quiescent Qigong meditations to active forms of dynamic Qigong movement.

THE YIN YAO CYCLES

The last six hexagrams (from Gou — 11 a.m. to Kun — 11 p.m.) indicate a gradual increasing of the energetic potential of the Yin Yao and the gradual decreasing of the Yang Qi. This time period is called the "time of the dead breath," and represents the beginning of energetic planting. The Yin Qi increases and expands in this time period; it is the best time for patients suffering from an insufficiency of Yin Qi to train, nurture, and restore the Kidneys' Yin Essence, which in turn will facilitate the preservation of Yang Qi. When training in the afternoon, the practitioner should progress from active forms of dynamic Qigong movement to gradually slower, relaxed, quiescent Qigong meditations.

When prescribing Medical Qigong for patients, the time for practice should be selected in accordance with the philosophy of Yin and Yang theory, the Five Elemental theory, the circadian rhythms of Qi and Blood circulation, as well as the severity of the illness.

USING THE YANG AND YIN YAO CYCLES

Each of the broken lines represents a Yin Yao and can be used to symbolize an inhalation, while a straight line represents a Yang Yao and symbolizes an exhalation. A hexagram consisting of three bottom Yang Yaos and three top Yin Yaos, for example, represents a balanced respirational pattern. When the Qigong doctor writes out a prescription for the patient, the doctor can adjust the patient's respiration in accordance with the Yao hexagrams, taking into consideration each Yao line as representing an inhalation or exhalation. The basic theory of Yin and Yang Yao cycles is also used to determine the time and direction for Qigong practice, as well as the methods of respiration and mind concentration required to establish proper Yin and Yang balance.

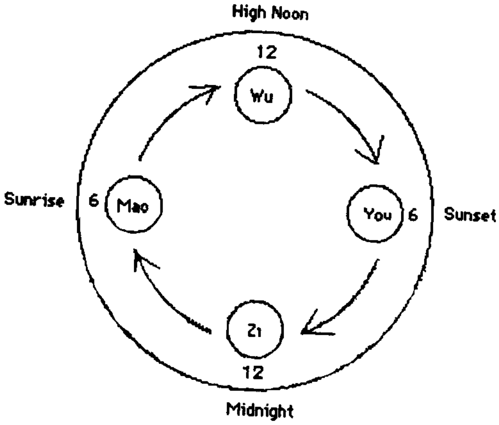

Figure 11.4. The Four Principle Time Periods

When practicing Medical Qigong at the hour of the Fu Hexagram, for example, the Yao indicates there is more Yin than Yang, so the practitioner should regulate his or her body with the intent to nourish the Yang through longer inhalation (Yin) and shorter exhalation (Yang).

At the stage of the Gou Hexagram, the Yao indicates there is more Yang and less Yin so the Qigong practitioner should perform regulation through shorter inhalations (Yin) and longer exhalations (Yang).

THE FOUR PRINCIPAL TIME PERIODS OF QIGONG PRACTICE

Practicing Medical Qigong during the "four principal time periods" (Zi, Mao, Wu, and You) facilitates the vigorous growth of internal energy in harmony with the changes of energy in nature (see Figure 11.4). To nourish energy, practice during the sunrise and / or sunset (Mao and You) periods. To stabilize energy, practice during the periods of midnight and / or high noon (Zi and Wu).

Due to the differences between individual constitutions (Excess or Deficient Yang or Yin), the selection of the proper time for practice and the number of Yin or Yang respirations varies between patients. The following is a brief discussion and explanation of the four principal time periods.

1. Zi Time Period corresponds to the Fu Hexagram, and has one Yang and five Yin lines or Yaos. The Zi time period is at midnight (11 p.m. – 1 a.m.). It is related to the beginning of the Yang cycle and is perfect for gathering energy to store the Yuan Qi (Prenatal Qi). The Yuan Qi is the root of life, formed through the divine combination of the primordial Yin and Yang; it is stored primarily in the Kidneys' Water energy. The Zi period of time corresponds to the Fu Hexagram. During this time the Earth energy and the Kidney's Yang Qi are in full harmony, continuously combining and accumulating energy. Practice during this period of time results in twice the effect with half the effort.
2. Mao Time Period corresponds to the Dazhuang Hexagram, and has four Yang and two Yin lines or Yaos. The Mao time period is at sunrise (5 a.m. – 7 a.m.). As the sun rises, warming the air, and natural Yang progresses. At this time of the day, the Yang Qi of the body is in full bloom. Practice during this period of time aids the vigorous growth of Yang.
3. Wu Time Period corresponds to the Gou Hexagram, and has one Yin and five Yang lines or Yaos. The Wu time period is at high noon (11 a.m. – 1 p.m.), and is regarded as the Heart's Fire. At high noon, the Heart Channel (considered the Supreme Yang) will naturally manifest the energetic peak of Yang energy. Because the first line is a Yin Yao, however, the Yang energy Yaos tend to wane rather than grow. Therefore, practicing during this period helps the growth of Yin energy and suppresses the hyperactivity of Yang energy.
4. You Time Period corresponds to the Guan Hexagram, and has four Yin and two Yang lines or Yaos. The You time period is at sunset (5 p.m. – 7 p.m.). At this time, the energy of the environment turns from clear and radiant to dark. Qigong practice at this time nurtures the increase of the primordial Yin energy and the conservation and nourishment of Yang energy.

THE FOUR SEASONAL PEAK TRANSITIONS

Another theory maintained by Qigong doctors is that the sun's light during the day represents Yang, and the energy from the moon's light at night represents Yin. The observation that light coincides with heat and that darkness coincides with cool, led to the division of the year into four seasonal peak transitions (which also correspond to the four principal time periods of the day) (see Figure 11.5).

At the time of the equinoxes and solstices, the sun passes through transitional regions: The Palace of Eternal Frost in the North is entered in the winter; The Golden Gate in the East is entered in the spring; The Palace of Universal Yang in the South is entered in the summer; and The Gate of the Moon in the West is entered in the autumn. The four seasonal peak transitions in the northern hemisphere are described as follows.

1. The Palace of Eternal Frost at the Winter Solstice (December 22–25) indicates the time period when the sun is furthest south from the equator. This is a time of maximum Yin or darkness in the northern hemisphere, with the shortest days and longest time of cold and darkness. This energetic transition corresponds to "midnight" within the four principal time periods.
2. The Golden Gate at the Spring Equinox (March 21) indicates the time period when the sun's center crosses the Heavenly equator, creating an energetic balance between the days and nights. Although light and darkness and Yin and Yang energies are in balance, this seasonal time period is a transitional harmony, as Yang energy is growing. The weather is also moderate. This energetic transition corresponds to the "sunrise" within the four principal time periods.
3. The Palace of Universal Yang at the Summer Solstice (June 21–23) indicates the time period when the sun is furthest north from the equator, a time of maximum Yang or brightness, with the longest days and times of light and warmth. This energetic transition corresponds to "noon" within the four principal time periods.

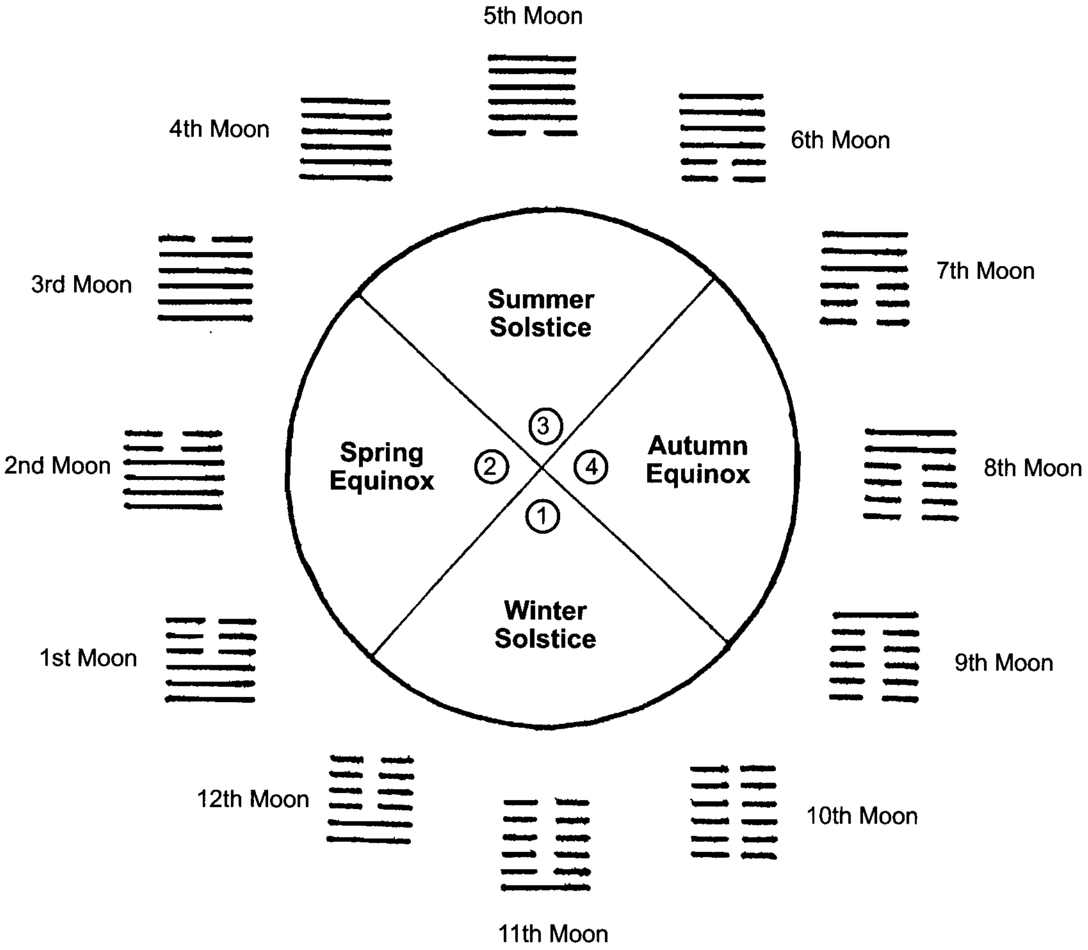
5th Moon
4th Moon
6th Moon
3rd Moon
7th Moon
Summer Solstice
Spring Equinox
Autumn Equinox
Winter Solstice
③
②
④
①
2nd Moon
8th Moon
1st Moon
9th Moon
12th Moon
10th Moon
11th Moon

Figure 11.5. Four Season Peak Transitions

4. The Gate of the Moon at the Autumn Equinox (September 22) indicates the time period when the sun's center crosses the Heavenly equator, creating an energetic balance between the days and nights. Although light and darkness and Yin and Yang energies are in balance, this seasonal time period is in transitional harmony, as Yin energy is growing. The weather is also moderate. This energetic transition corresponds to the "sunset" within the four principal time periods.

The energy of these four principal time periods enters the practitioner's Taiji Pole during meditation and regulates the "auspicious powers" located within the five regions of space inside and outside of the doctor's body:

- Front—energy of the Heart
- Back—energy of the Kidneys
- Right—energy of the Lungs
- Left—energy of the Liver
- Center—energy of the Spleen

THE SIX PRINCIPAL SEASONS OF TRANSITION

In Traditional Chinese Medicine, the year is divided into six primary periods of seasonal transitions (wind, heat, damp, fire, dry, and cold). Each season represents the ruling Qi, or expected weather, dominant in that particular season (Figure 11.6).

These six periods delineate the differences between and transitional transformations of Heaven energies (through the Ten Heavenly Stems) and Earthly energies (through the Twelve Earthly Branches). In terms of weather, the first half of the year is governed by the Qi of Heaven, while the second half of the year is administered by the Qi of Earth. The Ten Heavenly Stems are the ten energies of Heaven that rule the changes of the Five Elemental Seasonal Transitions, and are represented in the human body as the Yin and Yang aspect of the Five Elements (represented as the ten major internal organs). The Twelve Earthly Branches are the twelve energies of the Earth that determine the six Qi factors of the seasonal transitions (represented in the body as the Twelve Primary Channels). In accordance with the Creative Cycle of the Five Elemental Seasonal Transitions: summer creates late summer, late summer engenders autumn, autumn gives rise to winter, and winter gives birth to spring.

THE DAILY QI, BLOOD, AND HEAT CYCLE

These four principal time periods and seasonal peak transitions affect the body's Blood, Heat, and Qi Cycles. The daily cycle of Heat, Blood, and Qi flow along the body's Twelve Primary Channels, organs, and tissues, according to the two hour Yin and Yang time table are:

- Gall Bladder (GB) - 11 p.m. to 1 a.m. - (Yang)
- Liver (Lv) - 1 a.m. to 3 a.m. - (Yin)
- Lungs (Lu) - 3 a.m. to 5 a.m. - (Yin)
- Large Intestine (LI) - 5 a.m. to 7 a.m. - (Yang)
- Stomach (St)- 7 a.m. to 9 a.m. - (Yang)
- Spleen (Sp) - 9 a.m. to 11 a.m. - (Yin)
- Heart (Ht) - 11 a.m. to 1 p.m. - (Yin)
- Small Intestine (SI) - 1 p.m. to 3 p.m. - (Yang)

Season	Date
1. Spring Begins	Feb. 5
Rainwater	Feb. 20
Insects	March 7
Spring Equinox	March 22
2. Clear and Bright	April 6
Grain Rains	April 21
Summer Begins	May 6
Grain Filling	June 7
3. Grain Full	June 22
Summer Solstice	June 22
Slight Heat	July 8
Great Heat	July 24
4. Autumn Begins	Aug. 8
Limit of heat	Aug. 24
White Dew	Sept. 8
Autumn Equinox	Sept. 24
5. Cold Dew	Oct. 9
Frost Descent	Oct. 24
Winter Begins	Nov. 8
Slight Snow	Nov. 23
6. Great Snow	Dec. 7
Winter Solstice	Dec. 22
Slight Cold	Jan. 6
Great Cold	Jan. 21

Figure 11.6. Six Principal Seasons of Heaven and Earth Energy Transition (1-6), and the Twenty Four Diagram Chart

- Bladder (Bl) - 3 p.m. to 5 p.m. - (Yang)
- Kidneys (Kd)- 5 p.m. to 7 p.m. - (Yin)
- Triple Burners (TB) - 7 p.m. to 9 p.m. - (Yin)
- Pericardium (Pc) 9 p.m. to 11 p.m. - (Yang)

The body's active high and low heat, Blood, and Qi time periods, reflect its relationship with the Twelve Primary Channels. For instance, every morning during the time period of 3–5 a.m., the strongest energy streams down along the Lung Channels, making the Lungs quickly responsive to treatment.

GATHERING THE SUN'S ESSENCE MEDITATION

The ancient Qigong masters used this meditation to absorb the essence of the sun. By gathering pure Yang energy to tonify the body's Deficiencies, as well as to consolidate the Zheng Qi (Gathering energy), they were able to eliminate pathogenic factors and prevent premature aging. This particular exercise is considered part of the foundational base from which the ancient Qigong doctors replenished any depleted life-force energy.

The gathering the Sun's Essence Meditation must only be practised on a clear day, when the sun is not obstructed. Avoid practising this meditation in times of fog, smog, thick clouds, and solar eclipses, as cultivation during these times will cause the unnatural cultivation of Turbid Qi.

1. While outside on a bright sunny day, face the sun, and assume a Wuji standing posture, hands by your sides, with the feet slightly wider then shoulder-width apart; stand in a relaxed and tranquil state, breath evenly and naturally, and rid your mind of any stray thoughts (Figure 11.7).
2. If the sun is rising above the horizon, slightly close your eyes; you should still be able to see soft, gentle, reddish sunlight. If the sun has already risen, close your eyes and use your inner vision to visualize the sun.
3. Inhale through the nose and imagine filling your mouth with energy from the sunlight's essence. Still facing the sun, with your eyes half open, "infuse," or absorb into the eyes, the energy from the horizon. Blend the energy

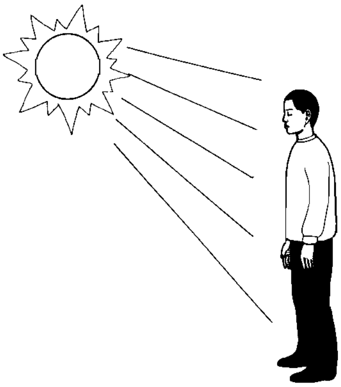

Figure 11.7. Gathering the Sun's Essence Meditation. Step one, stand and face the sun.

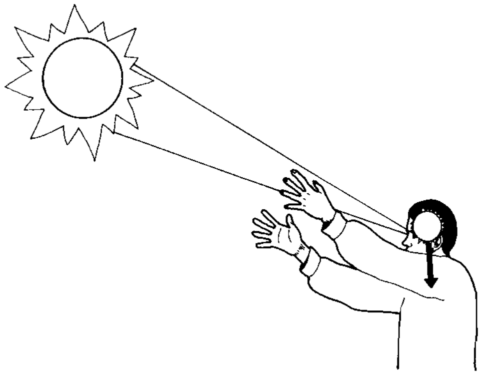

Figure 11.8. Step two, imagine embracing the sun.

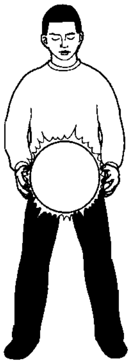

Figure 11.9. Step three, imagine embracing the sun and gently placing it in front of your Lower Dantian.

absorbed from your eyes into your Upper Dantian. Close your eyes and imagine blending the sun's golden light energy with the Qi you inhaled through your nose. Hold your breath and focus the mind, mixing both energies; swallow the energies slowly while exhaling, and send the energy down to the Lower Dantian. Repeat this process for several minutes.

4. Stretch your hands towards the sun and imagine embracing it (Figure 11.8). Point your index fingers toward the center of the sun, and imagine pulling the sun down in front of your Lower Dantian. Keep it positioned there for several minutes (Figure 11.9).
5. Imagine the sun gently rolling and turning itself in front of your lower abdomen (Figure 11.10–11). Keep your elbows at your sides and gently shift, or sway from side to side, imagine that it is the sun's movement that causes both your arms and body to move. Using this image keeps the sun in an energetic power position.
6. Place both hands around the sun as if embracing it (Figure 11.12). Next, imagine the sun slowly moving into your navel. Allow both hands to follow the sun into the navel, ending with your left palm on top of your right palm (opposite for women) (Figure 11.13). With one deep breath inhale and absorb the sun's energy into the Lower Dantian while imagining the sun (beneath the navel) transforming into a luminous ball of golden energy. With each inhalation, this golden ball shines brightly; with each exhalation, the sun's rays extend, spreading throughout the entire body.
7. To end, relax the mind and imagine Heavenly Qi melting down the body, pouring through the front, back and center of the body like warm oil, into the Earth. This ending exercise (sometimes called Pulling Down The Heavens), allows the Qi to sink into the Lower Dantian, and the body, mind, and spirit to come into a state of peaceful rest and tranquillity.

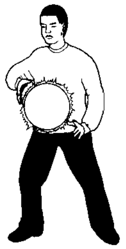

Figure 11.10. Step four, imagine the sun gently rolling and turning itself in front of the Lower Dantian.

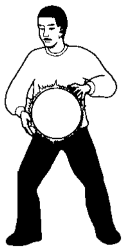

Figure 11.11. The energy of the sun moves the entire body rather than the arms.

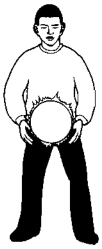

Figure 11.12. Step five, embrace the sun and absorb it into the navel and then into the Lower Dantian.

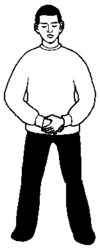

Figure 11.13. Each breath transforms the sun's energy into luminous golden Qi.

CYCLES OF THE MOON

The ancient Qigong masters believed that the sun has a Yang nature, emitting enormous amounts of hot, thermal energy, whereas the moon has a Yin nature and emits a much cooler type of energy, by reflecting the light of the sun. The energy of the moon is used to replenish the body's Yin Qi, and follows a comprehensive cycle every 28 days (Figure 11.14). Following the cycle of the moon, the energy of the body completes a full rotation, flowing along the Governing Vessel to the top of the head, and down the Conception Vessel to the base of the perineum. When the moon is full, the body's Blood and Qi reach a crest, filling the Baihui point on the top of the head. At the new moon, the body's Blood and Qi gather at the Huiyin point, in the perineum.

Since the lunar cycle at its extreme (full moon and new moon) activates the top and bottom points of the Taiji Pole, these are said to be the most productive times to perform Taiji Pole Qi practice.

The appearance of the moon during the first half of the lunar month symbolizes the waxing of Yang and waning of Yin Qi. As Yang grows and Yin declines, regulation of breathing at this period of time should be aimed at absorption of the Moon's Essence. The second half of the lunar month symbolizes the waxing of Yin and waning of Yang Qi. As Yin grows and Yang declines, regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Jing-Essence (Figure 11.15-20).

As the moon increases the amount of light it reflects, the body's Qi and Blood become more nourished; when the moon is full, the Qi and Blood become substantial. As the moon begins to decrease the amount of light it reflects, the muscles start to lose their energetic substance. This decrease in energy causes the body's Eight Extraordinary Vessels and Twelve Primary Channels to empty themselves to support the muscles that are slowly becoming Deficient. Qi and Blood are influenced by the gravitational pull of the sun and moon, just as the tides of the sea are influenced. Therefore the ancient Qigong masters described the energetic tonifying and purging principles of Qigong regulation in these specific terms: when treating a patient, perform no tonifying when the moon is full, or when the Blood and Qi are sufficient. Perform no purging when the moon is dark, or when the Blood and Qi are Deficient. This observation explains the effect of the "full-moon phenomenon" on patients who are hypersensitive to energy and become restless when the moon's Qi is "full Yang in Yin."

The waxing and waning of the sun and moon affect the adjustment time for Medical Qigong practice and the methods of breathing regulation. In order to facilitate the intake of Heavenly Es-

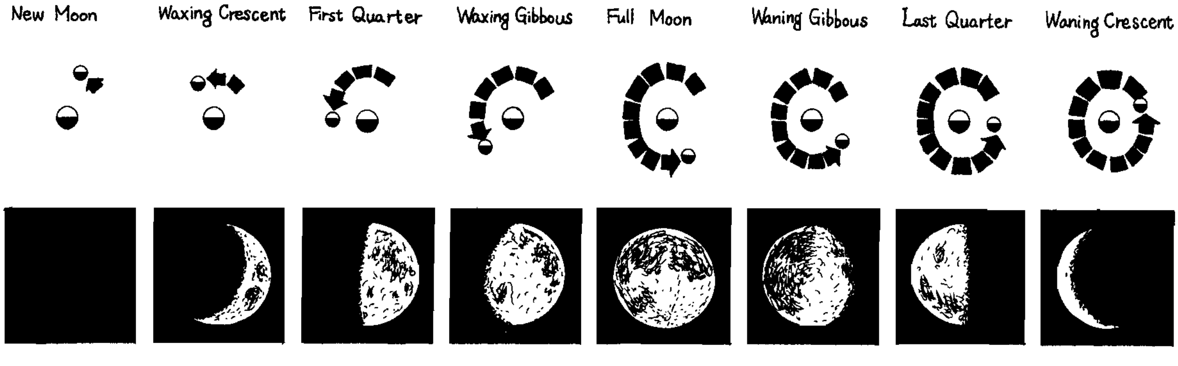

Figure 11.14. The waxing and waning cycles of the moon correspond to the moon's orbit around the Earth.

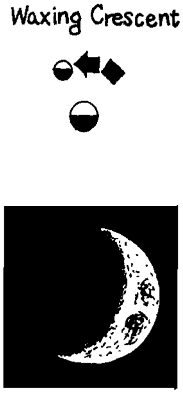

Figure 11.15. From the 1st to the 5th night of a lunar month, more Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the west. The crescent surface of the Moon is reflecting less sunlight.

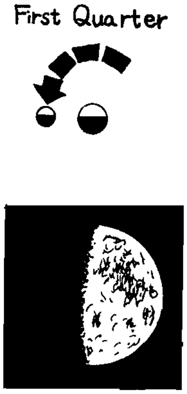

Figure 11.16. From the 6th to the 10th night of a lunar month, more Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the south. The surface of the moon is turning from a crescent to half-full Yang, and its surface is reflecting more sunlight.

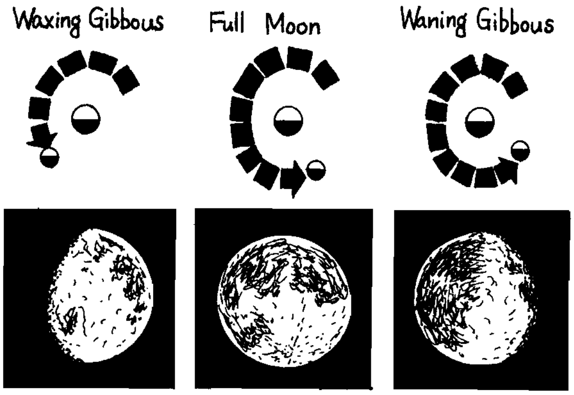

Figure 11.17. From the 11th to the 15th night of a lunar month, equal amounts of Yang absorption and Yin nourishment can be obtained by performing breathing regulations while facing the direction of the moon, as both Yang and Yin energies are plentiful at this time. At this point, the moon is full with its surface reflecting the most sunlight (called "Full Yang in Yin").

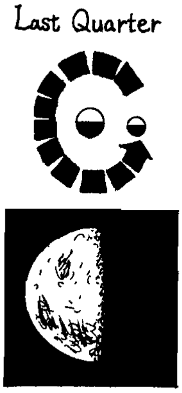

Figure 11.18. From the 16th to the 20th night of a lunar month, the Yin grows and Yang declines. Regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Essence, while facing the west. At this time, the crescent surface of the moon is reflecting less sunlight. The reflected sunlight on the moon's surface decreases as Yang wanes and Yin waxes.

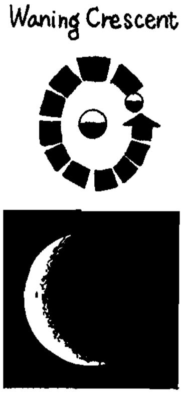

Figure 11.19. From the 21st to the 25th night of a lunar month, Yin continues to grow and Yang declines. Regulation of breathing at this period of time should be aimed at replenishing Yang to help the Yin nourish the Yin Essence, while facing the south. At this time, the crescent surface of the moon's shape is turning to half-full Yin, and its surface is reflecting less sunlight. The reflected sunlight on the moon's surface continues to decrease as Yang wanes and Yin waxes.

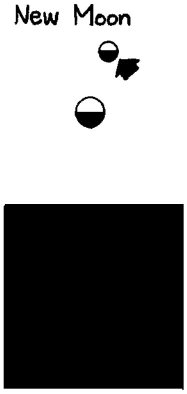

Figure 11.20. From the 26th to the 30th night of a lunar month, no absorption of Yang is possible when the moon appears in the east at nightfall and its surface is dark, leading to the period of "Pure Yin." Breath regulation should be performed facing east.

sence to nourish the body, the ancient Qigong masters used a system wherein each month is divided into 6 separate sections (a total of 5 days was assigned to each section).

PRENATAL EIGHT TRIGRAMS AND THE MOON

The symbols on the Prenatal Eight Trigrams represent the appearance of the moon, while the Ten Heavenly Stems represent the positions of the moon. Each trigram embraces one or two of the Ten Heavenly Stems. The following chart of Prenatal Trigrams, Ten Heavenly Stems, moon and tide regulation, as well as the principles of purgation and tonification of Qi and Blood is a reference for Qigong doctors for clinical treatments (Figure 11.21).

In this chart, the Yang energy of the Earth gradually increases from the 11th to the 15th. Influenced by this action, the spring tide occurs on the 15th when the moon is full, inducing both the Qi and Blood to become extremely active. This is a good time for Qigong patients who suffer from energy Deficiencies to replenish their Qi. However, if the patient's energy is abundant, he or she should not try to replenish at this time, or an Excess condition of Qi will occur.

CHOOSING THE PROPER DIRECTION FOR TRAINING QI

Since ancient times, Qigong doctors have placed great emphasis on facing certain directions when cultivating energy. Observing that energy training and practicing should be adapted to nature, each routine is selected according to the season, month, and time of day, as well as the physical conditions of the practitioner. The four directions, as well as the sun (solar cycle), moon (lunar cycle), and stars (four sets of seven constellations, each one grouped in the N, S, E, or W), have a direct influence on the human body.

One example of directional training is in setting up treatment schedules according to the energy emitted from the Heavens. It is generally advised when conducting energetic regulations that:

- Patients with Yang Deficiencies should face east or south, to develop Yang Qi;
- Patients with Yin Deficiencies should face west or north, to help develop the Yin Qi;
- Patients with a Kidney Deficiency should face the north;
- Patients with a Liver Deficiency should face east;
- Patients with a Heart Deficiency should face south;
- Patients with a Lung Deficiency should face west;
- Patients with a Spleen Deficiency should focus their attention on the center, i.e., the Earth.

Generally, when first beginning Qigong exercises and meditations, practitioners are taught to face the south or southeast. When practicing specific cultivation exercises, however (such as Gathering the Sun's Essence), the practitioners are taught to face the direction of the sun. In the evening time, unless practicing specific cultivation exercises (such as Gathering the Moon's Cream), it is best to have the patient face the corresponding direction for the Deficient organ. These cultivation methods train the Yang energy to fuse with the Zhen Qi (True Qi), as well as with the Heavenly Qi and Environmental Qi. The timing and methods of training are based on the increase and decrease of Yin and Yang energy.

It is important to note that although these training methods are effective, the most experienced Qigong doctors often choose the place and direction according to their own intuition. To select the appropriate direction, the doctors start by facing south (sitting or standing). As they begin to feel the body filling with Qi, they may feel that they are being pulled towards a specific direction, and that they cannot stand or sit comfortably without facing that direction. The doctors feel more rooted and stable by following their intuition. Through understanding the proper directions, concentration, breath regulation, and assuming the appropriate postures, Qigong doctors are able to help regulate and increase their patients' energetic activities, while avoiding many Qi deviations (see Chapter 18).

Heavenly Stems	**Date**	Appearance of the Moon	Prenatal Trigrams	**Position of Moon** you face ○	**Cause & Reason**	**Tide**	**Qi & Blood**	**Regulation**
Geng ䷡	1st	**Waxing Crescent**	Zhen ☳	West 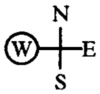	Moon between Sun & Earth	Spring Tide Max. Amplitude when Sun/Moon Pull Together	Turning Excessive	No Tonifying
Ding ䷫	8th	**First Quarter**	Dui ☱	South 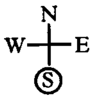	Moon is 90° East of Sun	Neap Tide Min. Amplitude Sun/Moon work against each other	Deficient	No Purging
Jia ䷗	15th	**Full**	Qian ☰	East 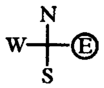	Moon opposite Sun, Earth in between to East	Spring Tide Max. Amplitude Sun/Moon pull together	Excessive	No Tonifying
Xin ䷊	16th	**Last Quarter**	Xun ☴	West 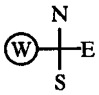	Moon opposite Sun, Earth in between to West	Medium Tide	Turning Deficient	No Purging
Bing ䷁	23rd	**Waning Crescent**	Gen ☶	South 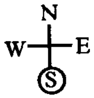	Moon is 90° West of Sun	Neap Tide Min. Amplitude Sun/Moon work against each other	Deficient	No Purging
Yi ䷒	30th	**New**	Kun ☷	East 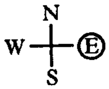	Moon between Sun & Earth	Medium Tide	Deficient	No Purging

Figure 11.21. Medical Qigong Doctor's Chart for Clinical Reference

GATHERING MOON'S CREAM QIGONG MEDITATION

Qigong doctors use this particular exercise in conjunction with the Gathering the Sun's Essence meditation for replenishing depleted energy. The vital essence of moon is called the Cream or "cool light" and is used to nourish the Kidneys and strengthen the Jing, Marrow, and brain, while contributing to longevity. It is practiced only nine days a month, three days before, during, and after the full moon, which are the brightest nights of the month. It is prohibited during new moon transitions because there is too little vital essence of the moon to gather.

Gathering the Moon Cream meditation must only be practiced on clear nights, when the moon is not obstructed. Avoid practicing this meditation in times of fog, smog, thick clouds, and lunar eclipses, as cultivation during these times will cause the unnatural cultivation of Turbid Qi

1. Stand outside in the fresh air, while facing the moon. Stand in a relaxed and tranquil state, breath evenly, and rid your mind of any stray thoughts (Figure 11.22).
2. Lower your eyelids until you only faintly see the moonlight. Inhale through the nose and imagine softly "sipping" one mouthful of the moon cream. At the same time, with your eyes half open, visualize "inhaling" the energy of the moon cream with your eyes into your Upper Dantian. Close your eyes and imagine mixing the moon cream with the energy you inhaled through your nose. Slightly hold your breath and concentrate on feeling the moon cream's silvery energetic essence. Swallow this essence slowly, sending it down to the Lower Dantian.
3. Stretch your hands towards the moon (Figure 11.23). Imagine embracing the moon and pulling it out of the sky, placing it just above your Upper Dantian, at your Baihui point (on top of your head). Hold this image for several minutes (Figure 11.24).
4. Next, imagine a second moon forming out of the first moon, and bring it downward with both hands until level with, and in front of, the Middle Dantian (Shanzhong CV-17 point) at the center of the chest. Hold this image for several minutes (Figure 11.25).
5. Finally, imagine a third moon forming out of the second moon, and bring it down to the level of the Lower Dantian, suspended in front of the lower abdomen. Hold this image for several minutes (Figure 11.26).
6. Continue to imagine all three moons suspended (two in front, and one above your body) and inhale, raising both of your arms over your head. Imagine pressing the first moon into your head through the Baihui point, filling the Upper Dantian with the moon cream's silvery white light. As both hands press the moon into the head, allow the left hand to remain on top of the right hand (opposite for women). As you exhale, visualize the light of the moon filling the entire cavity of the Upper Dantian (Figure 11.27).
7. Next, bring both hands to the front of your body. Inhale and press the second moon into your chest through the Middle Dantian, imagining the moon cream's silvery white light filling the chest. As you exhale, visualize the light of the moon filling the entire cavity of the Middle Dantian (Figure 11.28).
8. Finally, bring both hands in front of your lower abdomen. Inhale and press the third moon into your Lower Dantian, imagining the moon's silvery white light filling the abdomen. As you exhale, visualize the light of the moon filling the entire cavity of the Lower Dantian (Figure 11.29).
9. Allow both hands to rest by your sides then slowly raise your left hand (opposite for women) and tap the top of your Baihui point softly nine times. As you tap, imagine that all three moons fuse into a brilliant white-light energy throughout your Taiji Pole, connecting all three Dantians together (Figure 11.30).
10. Relax and perform Pulling Down The Heavens, letting your body, mind and spirit sink into tranquility.

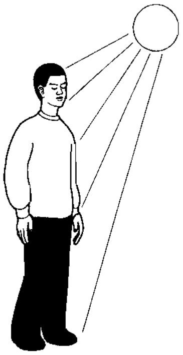

Figure 11.22. Stand facing the moon.

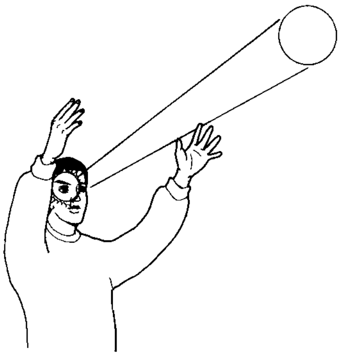

Figure 11.23. Stretch your hands to embrace the moon.

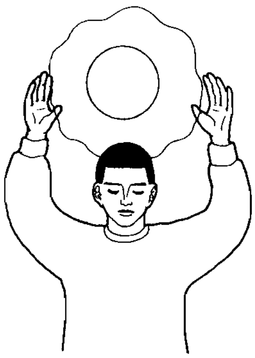

Figure 11.24. Suspend the moon above your Upper Dantian.

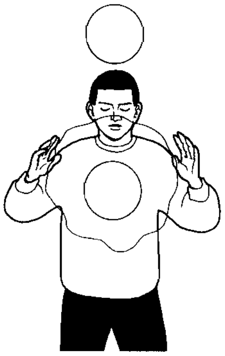

Figure 11.25. Suspend a second moon in front of your Middle Dantian.

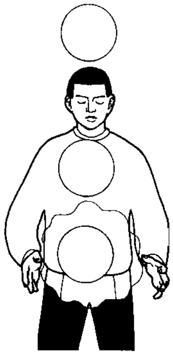

Figure 11.26. Suspend a third moon in front of your Lower Dantian.

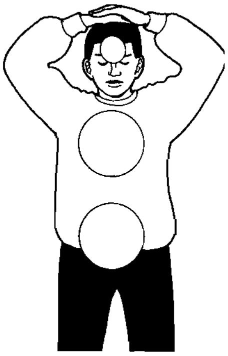

Figure 11.27. Absorb the first moon into your Upper Dantian.

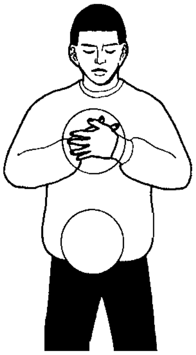

Figure 11.28. Absorb the second moon into your Middle Dantian.

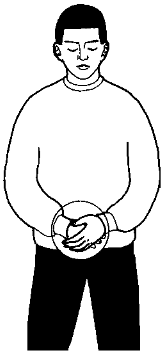

Figure 11.29. Absorb the third moon into your Lower Dantian.

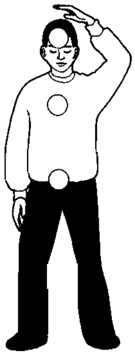

Figure 11.30. Tap lightly until all three moons flow into your Taiji Pole.

CYCLES OF THE STARS

The most influential Heavenly Treasures, other than the sun and moon, are the stars composed of the Big Dipper, Polaris, and the five planets, existing within the twenty-eight constellations.

The North Star is believed to be the unchanging area of Heaven, from which the Heavenly order began (often called the Axis of the Vault of Heaven), and so it was named The Divine Center of the Earth and Sun Cycles. It is also believed that from this star the entire universe expands outward. By placing the North Star in the center of a circle of the Heavens, the circle can be divided into four directions, each containing seven constellations that have a different color energy and sound frequency.

In ancient China, the seven constellations were divided into four sections or directions (north, south, east, and west). In the region of the north was believed to exist the energy of the Tortoise-Snake, in the south was the Red Phoenix, in the east was the Green Dragon, and in the west was the White Tiger. Each of these regions was further divided into twenty-eight constellations or posts (xiu).

According to traditional Chinese belief, the moon travels twenty-eight days from west to east, resting at each post as it circles the Earth. In other words, the twenty-eight days are further divided into twenty-eight regions which are courier stations or travelling houses for the moon.

It was further believed that the Pole Star axis stemming from the North Star traveled downward from the center of Heaven, connecting with the Earth's center core (through the north and south poles). The north pole was regarded as the peak (or Baihui) of the Earth, and was believed to be comprised of seven stars (making it a fifth constellation). The energies of each of the four animals and their seven constellations moved around the Earth's central Pole creating the Five Elemental changes in weather (see Chapter 11). If the weather changes progressed in their expected order, they were considered normal; however, if the weather transition occurred either earlier or later than expected, the result was considered an environmental Evil Qi.

Therefore, the ancient Chinese believed that one could predict the changes of weather by studying the interrelationships between the Five Elements and the cycles of the sun, moon, Polaris, the Five Planets, and the twenty-eight constellations.

THE ENERGY OF THE BIG DIPPER

As the Earth orbits the sun, the angle at which the Big Dipper can be viewed changes on a daily, monthly, and yearly basis (Figure 11.31). These changing angles correspond to the different cycles of the moon as they are arranged in the Twelve Earthly Branches and repeat themselves after completing an orbit (360° rotation of twelve moon cycles). These twelve lunar cycles determine the twelve seasonal periods (Figure 11.32). The beginning part of each month marks the beginning of the seasonal periods of the Twelve Earthly Branches and the beginning of the energy cycle of that month. When the Big Dipper points in the direction of the "lunar month," facing the direction of the Big Dipper will invigorate the practitioner's energy. The Big Dipper is related to the energetic transformations of germinating, beginning, and growing.

There are three kinds of practice centered around the Big Dipper:

1. The first consists of invoking its power for protection;
2. The second consists of ascending into it and pacing its rotation;
3. The third involves gathering the star's essence or star cream into the body.

In ancient times, the Big Dipper was considered the bridge between the energies of the sun and the moon, and the source of "10,000 things." The sun and moon are considered a bipolar energy field, whereas the Big Dipper is considered unipolar.

The sun corresponds to the Heart and Fire energy, which is related to the front area of the body. The moon corresponds to the Kidneys and Water energy, which is related to the back area of the body. The Big Dipper corresponds to the Spleen, which is related to the Taiji Pole.

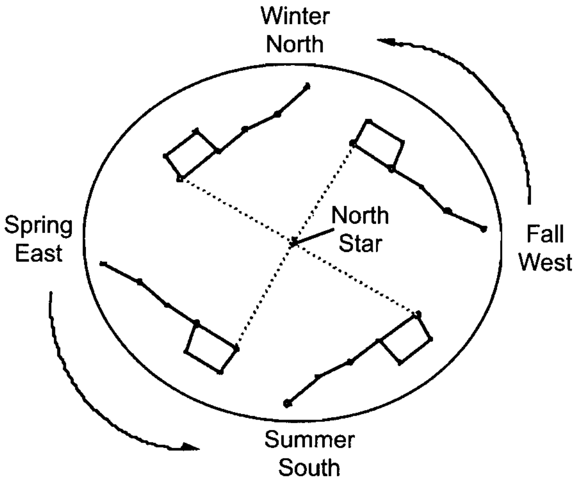
Winter
North
Spring
East
North
Star
Fall
West
Summer
South

Figure 11.31. The handle of the Big Dipper can be used to determine the season: (1) in winter, the handle points to the northern horizon, (2) in spring, the handle points to the eastern horizon, (3) in summer, the handle points to the southern horizon, and (4) in fall, the handle points to the western horizon. The North Star is positioned in the center of the diagram.

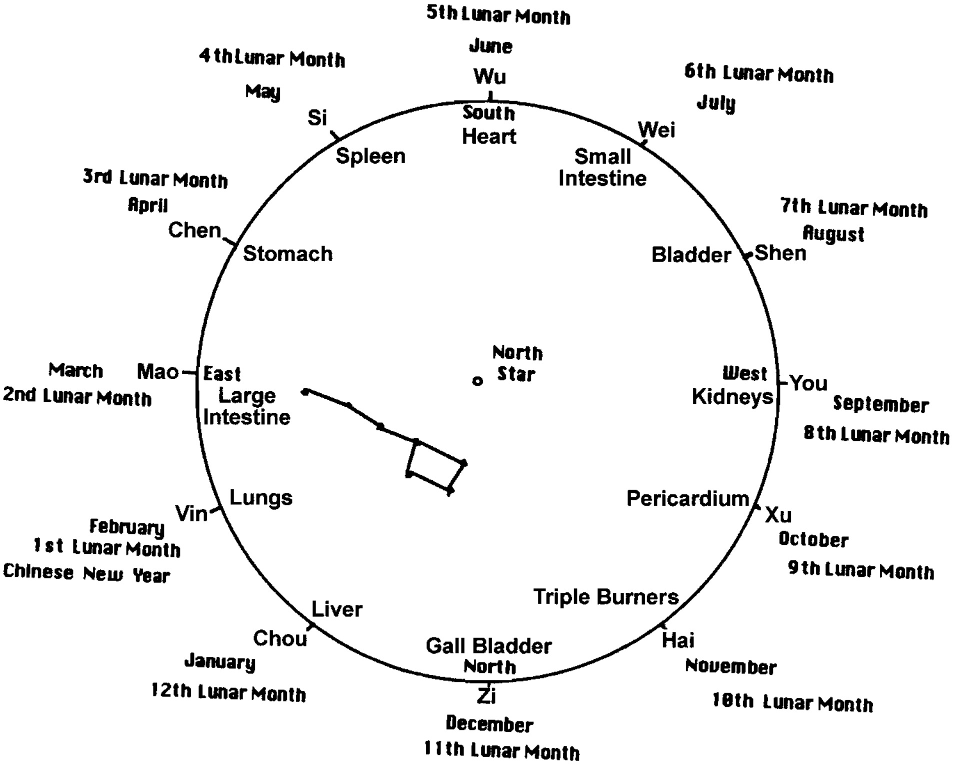
5th Lunar Month
June
Wu
South
Heart
4th Lunar Month
May
Si
Spleen
6th Lunar Month
July
Wei
Small
Intestine
3rd Lunar Month
April
Chen
Stomach
7th Lunar Month
August
Shen
Bladder
North
Star
March
Mao
2nd Lunar Month
East
Large
Intestine
West
Kidneys
You
September
8th Lunar Month
Vin
February
1st Lunar Month
Chinese New Year
Lungs
Pericardium
Xu
October
9th Lunar Month
Liver
Triple Burners
Chou
January
12th Lunar Month
Gall Bladder
North
Zi
December
11th Lunar Month
Hai
November
10th Lunar Month

Figure 11.32. The beginning of the seasonal periods of the Twelve Earthly Branches and the twelve moon cycles is determined by the angle of the Big Dipper.

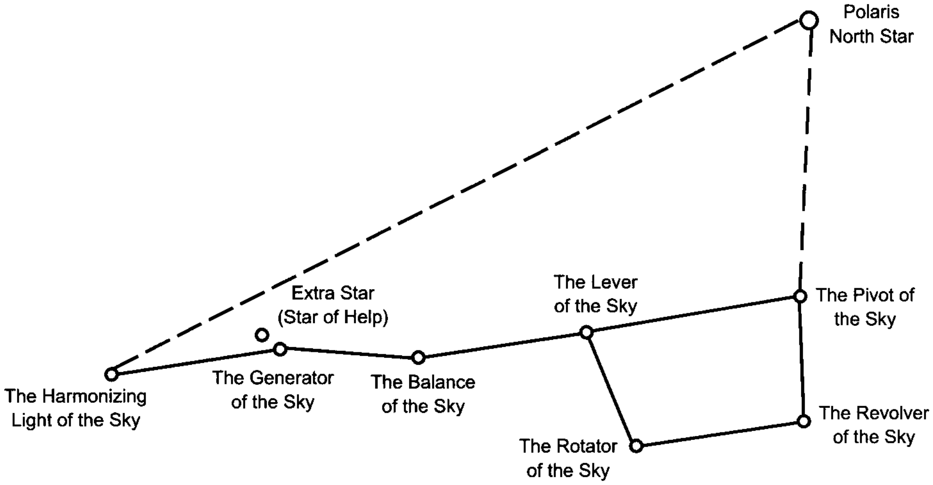

Figure 11.33. The Big Dipper is considered a reservoir of heavenly energy, collecting Qi from all of the other stars and constellations as it moves through the sky.

THE STARS OF THE BIG DIPPER

The Big Dipper represents the foundational base for the Gate of Access in Chinese mysticism, which is considered the spiritual passageway between life and death. Through this gate the individual may obtain knowledge and insight concerning past, present and future events.

The energy from the stars of the Big Dipper is used to balance the body's Five Agents (stored within the body's Five Yin Organs) and can be arranged internally to enhance internal perception of time and space (Figure 11.33).

One example used in ancient China for invoking the powerful protection of the Big Dipper energy is for the Qigong doctor to cover himself - or herself - with the stars of the Big Dipper by visualizing the stars descending and arranging themselves in a specific order around the doctor's body The stars can be arranged in four ways:

- above the doctor's head,
- in front of or behind the doctor's body,
- placed within the doctor's Three Dantians, or
- placed with the bowl of the Dipper in the Lower Dantian and the handle along the spine.

The doctor imagines placing the stars of the Big Dipper into the center of his or her body. The star's energetic function is to govern the front, back, right, and left of the body, thus regulating the doctor's Yin and Yang energy. The star's energetic function can also be used for absorbing energy from the four seasons into the doctor's body.

The stars of the Big Dipper are also considered responsible for opening the Seven Orifices of the embryo's body and for giving the embryo life. These stars also correspond to the Nine Chambers in the Three Dantians. Usually, meditations dealing with gathering energy from the sun and moon are followed by meditations involving the Big Dipper to further stimulate and energize the body's energetic fields.

THE FIRST FIVE PLANETS

The five planets (Saturn, Jupiter, Mars, Venus, and Mercury) are responsible for the internal Yin organ energy formation within the body during fetal development. These planets are the focus of several Qigong meditations because of their connection to color resonations within each internal Yin organ.

1. Of the first five planets in our solar system, Saturn is coupled with Earth and is considered the central force of planetary energy. It takes twenty-eight years for Saturn to make a complete revolution around the sun. Each year, when Saturn enters into another constellation, it is called "Saturn holding authority over that constellation for one year."

   Saturn's position in the sky relates to the Taiji Pole. The planet Saturn represents the Stomach and Spleen (Earth Element) and emits a golden-yellow light energy.
2. Jupiter represents the easterly direction and the Liver (Wood Element), it emits a green light energy; its position relates to the left side of the Taiji Pole.
3. Mars represents the southerly direction and the Heart (Fire Element). It emits a red light energy and relates to the front side of the Taiji Pole.
4. Venus represents the westerly direction and the Lungs (Metal Element). It emits a white light energy and relates to the right side of the Taiji Pole.
5. Mercury represents the northerly direction and the Kidneys (Water Element). It emits a dark blue/indigo light energy and relates to the back side of the Taiji Pole.

GATHERING ENERGY FROM THE FIVE PLANETS

To absorb energy from the five planets, start by quieting your mind while standing in a Wuji posture and facing the direction of the Big Dipper. Visualize yourself absorbing the energetic essence and color of each planet, one by one, into your body, filling each organ's entire Orb (the organs complete system, channels and tissues). Imagine filling the body with each planet's Heavenly energy from the top of your head, through the center Thrusting Channels.

Begin with the energy of Saturn, imagine it flowing into and filling your body through the Taiji Pole. Next, imagine the energy of Mars flowing into your front Center Thrusting Channel. Now imagine the energy of Mercury filling your body through the back Center Thrusting Channel, Venus's energy filling the body by descending into the right Center Thrusting Channel, and Jupiter's energy filling the body by flowing into the left Center Thrusting Channel. Gather the energy from their corresponding Orbs and lead the Qi downward into the Lower Dantian. After collecting and storing the Heavenly Qi, end the meditation with the Fire Cycle of the Microcosmic Orbit meditation.

THE TWENTY-EIGHT CONSTELLATIONS

Since the year 2000 B.C., the ancient Chinese Qigong Masters observed and used the movements and energy of the twenty-eight constellations, the sun and moon, and the first five planets with the Chinese zodiac, to forecast auspicious times for matters of state and war, and to predict weather and natural disasters. Each of these energetic systems has its own cyclic pattern as well as a direct and indirect influence on human life.

The Chinese practised gathering energy from the Heavenly stars, believing that the study of astrology, like alchemy, united the body's inner world with that of the outer. According to ancient divination practices, the celestial bodies exert forces and exhibit personalities that influence people and events on Earth. A person's character and destiny throughout life is based upon the positions of the planets and constellations at the exact time and place of birth. These influences can be determined by mapping the positions of the twenty-eight constellations in the sky at various times.

So important was this belief, that in ancient China, the Emperor was considered the high priest of the Heavens and frequently made sacrifices to the stars to maintain harmony with the universe. The four corners of the Emperor's palace represented the cardinal points in space (the equinoxes and solstices), and he and his family would move from one corner to another as the seasons changed.

GATHERING ENERGY FROM THE CONSTELLATIONS

At night time, when the moon is not suitable for absorbing Qi (new moon), the Qigong doctor can absorb the energy of the twenty-eight constellations into his or her body.

The energy of all five planets and twenty-eight constellations can be drawn into the body's internal organs and tissues through the five energetic Thrusting Channels which envelop and support the body's Taiji Pole. These Channels flow through the center of the doctor's body, connecting the top of the head to the perineum.

Externally, the twenty-eight day constellation cycle can be arranged externally on the body's Governing and Conception Vessels. The "ecliptic" path of the sun follows the Fire Cycle of the Microcosmic Orbit, which correlates to the individual's emotional/mental and physical health.

There are several meditations used in order to absorb the energy of the five planets and twenty-eight constellations into the doctor's body; one of which is known as The Heavenly Five Yin Organ Meditation.

HEAVENLY FIVE YIN ORGAN MEDITATION (TIAN WU ZANG)

This meditation is very popular in China. It focuses on fusing the energy from the five planets and the constellations into the body through the Five Thrusting Channels that surround the core center of the body's Taiji Pole. These Five Thrusting Channels should not to be confused with the TCM perspective of the Five Thrusting Vessels (as seen in Figure 11.34). Each of these Five Thrusting Channels relates to a specific organ energy and element (Figure 11.35).

When practicing this meditation and gathering the energy from the stars, it is important to first absorb the energy into the Five Thrusting Channels, then spiral the energy in a clockwise downward and counterclockwise upward direction through the body's Taiji Pole. This spiralling action allows the energy to balance itself and stabilizes the body's internal organs (Figure 11.36).

In order to perform the Heavenly Five Yin Organ Meditation and gather the Qi of the twenty-eight constellations, assume a Wuji posture, with your arms suspended by your sides. Face the direction of the North Star and Big Dipper (Figure 11.37). This exercise is performed for thirty-six breaths in each direction.

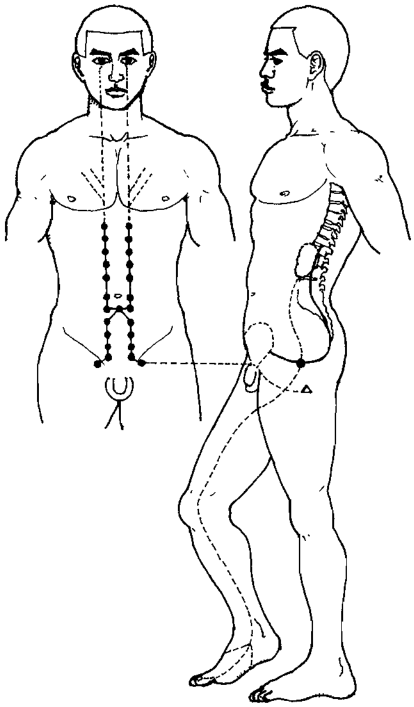

Figure 11.34. The Thrusting Vessels

1. To begin, imagine golden energy pouring out of the Big Dipper and flowing into the center of the Baihui area, at the top of your head. As you inhale, this golden energy fills the center core of your body, and flows into the extremities. As you exhale, this golden energy gathers into your Spleen.
2. Next, imagine drawing in a red energetic mist from the seven constellations in front of your body. This red mist flows into your body from the front side of the Baihui area, filling your entire body as you inhale. As you exhale, the energy coalesces into your Heart.
3. Now focus your attention on the back of your body. Inhale and imagine drawing in the dark blue energy from the seven constellations behind you into the back side of the Baihui area, filling your entire body as you inhale. This energy gathers into and fills the Kidneys as you exhale.
4. Next, focus on the right side of your body and imagine drawing in white energy from the

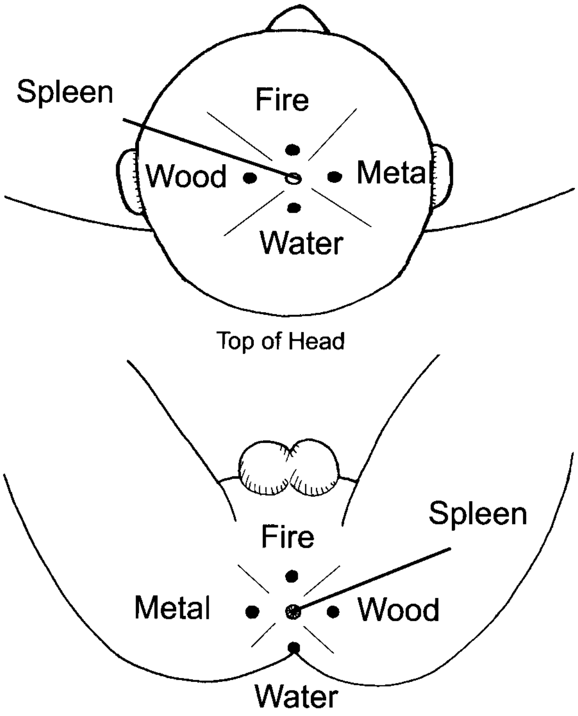
Spleen
Fire
Wood
Metal
Water
Top of Head
Fire
Spleen
Metal
Wood
Water

Figure 11.35. Each of the Center Thrusting Channels relates to a specific organ energy and element. Each channel has its own entry point that is positioned either on the top of the head, or on the base of the perineum.

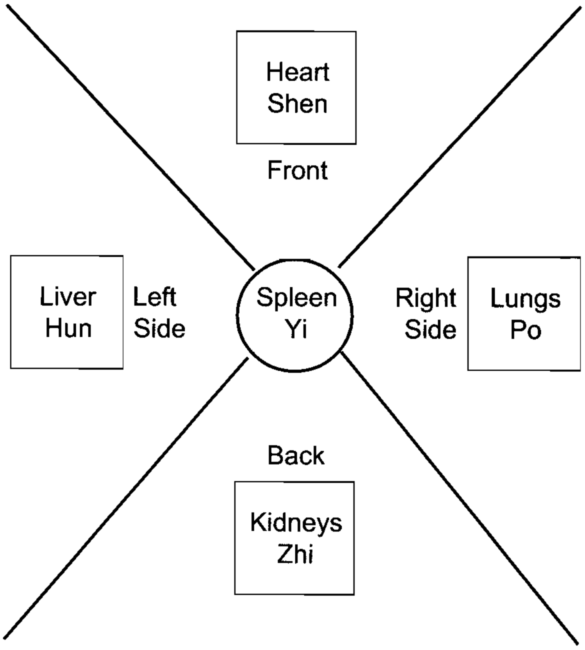
Heart
Shen
Front
Liver
Hun
Left
Side
Spleen
Yi
Right
Side
Lungs
Po
Back
Kidneys
Zhi

Figure 11.36. The body's Yuan Spirit is stored within the body's Five Yin Organs. The placement of the Five Energetic Houses is maintained throughout many Medical Qigong meditations and is used to tonify, strengthen, extend, and project the body's Shen.

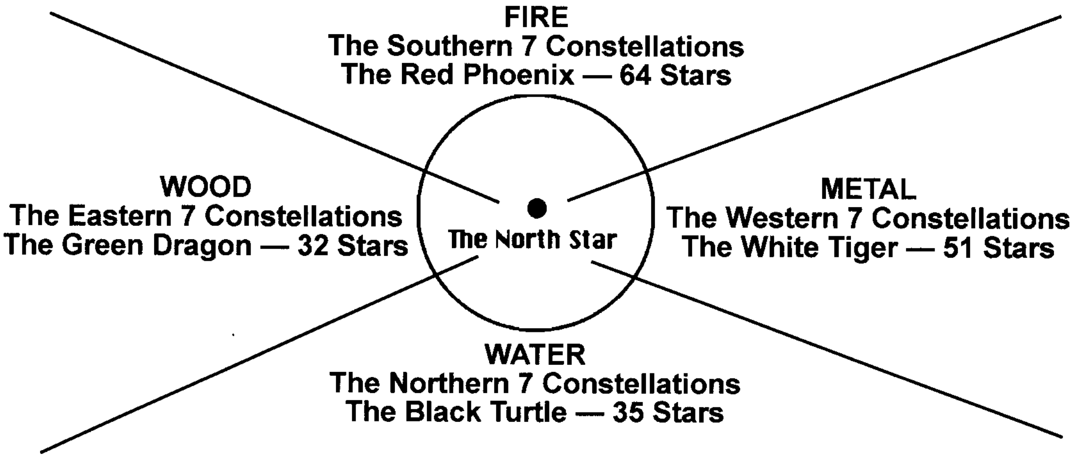
FIRE
The Southern 7 Constellations
The Red Phoenix — 64 Stars
WOOD
The Eastern 7 Constellations
The Green Dragon — 32 Stars
The North Star
METAL
The Western 7 Constellations
The White Tiger — 51 Stars
WATER
The Northern 7 Constellations
The Black Turtle — 35 Stars

Figure 11.37. The energy of the Twenty-Eight Constellation Stars is absorbed into the body through the same structural pattern as the placement of the Five Energetic Houses of the body's Shen.

seven constellations to your right into the right side of the Baihui area, filling your entire body as you inhale. This energy gathers and fills the Lungs as you exhale.

5. Finally, focus your attention on the left side of your body and imagine drawing in green energy from the seven constellations on your left into the left side of the Baihui area, filling your entire body as you inhale. This energy gathers and fills the Liver as you exhale.

EARTHLY BRANCHES, HEAVENLY STEMS, AND NINE PALACES

There are many relationships between the macrocosm and the microcosm of man, through the Twelve Earthly Branches, the Ten Heavenly Stems, and the Nine Palaces outlined in the Chinese Classics on Internal Medicine. The Twelve Earthly Branches are Earth energies and are represented in Man as the Twelve Primary Channels. The Ten Heavenly Stems relate to the energies of Heaven and are represented in man as the Yin and Yang aspects of the Five Element organs of Wood, Fire, Earth, Metal, and Water (as seen in Figure 11.38). The Nine Palaces are related to the Eight Extraordinary Vessels (corresponding with the Eight Postheaven Trigrams) in addition to the body's energetic Taiji Pole.

The Twelve Earthly Branches, Ten Heavenly Stems, and the Nine Palaces are interwoven into almost all aspects of Chinese culture, arts and medicine. In Chinese divination, as well as Chinese medicine, the temporal cycles of the Earthly Branches and Heavenly Stems form the foundation for these practices. They can be used to explain the physiology, pathology, diagnosis, and treatment of each clinical case. For each year, month, day, and hour there is a Stem and Branch association. By studying the Stem and Branch association, the Qigong doctor can identify patterns (i.e., determine the cycles and diagnose the patterns of the patient's disharmony) and design treatment protocols, depending on the time of year and the time of day.

The yearly cycles of the Stems and Branches are based on a sixty-year cycle, when the sun, moon, Earth, and North Star are aligned (Figure 11.39). The last sixty-year cycle began in 1984. The exact origin of the Ten Heavenly Stems and Twelve Earthly Branches is not known. They were used, however, to mark the hours, days, years, and planetary cycles as early as 1500 to 2000 B.C. Some historians attribute their origin to the legendary emperor Huang-Di (as early as 2600 B.C.). Each cycle was regarded by the Chinese as being a "century" in the life span of the average man. Those who lived past the age of sixty were regarded as living a "second life." During this second life, the person would experience similar occurrences as in the "first life," but would be able to apply the wisdom and maturity gained from his or her first life. The individual can then further his or her spiritual growth and use the knowledge to benefit society as a whole.

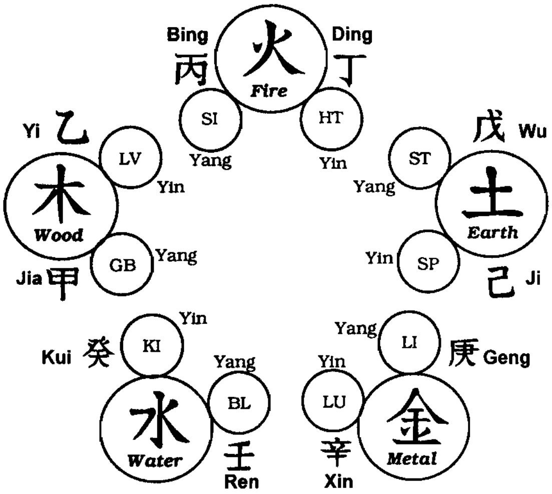

Figure 11.38. The Five Element Chart and Its Relationship to the Ten Celestial Stems

The twelve animals of the Chinese zodiac correlate with the Twelve Earthly Branches and are arranged as follows.

- Zi branch corresponds to the Rat
- Chou to the Ox
- Yin to the Tiger
- Mao to the Rabbit

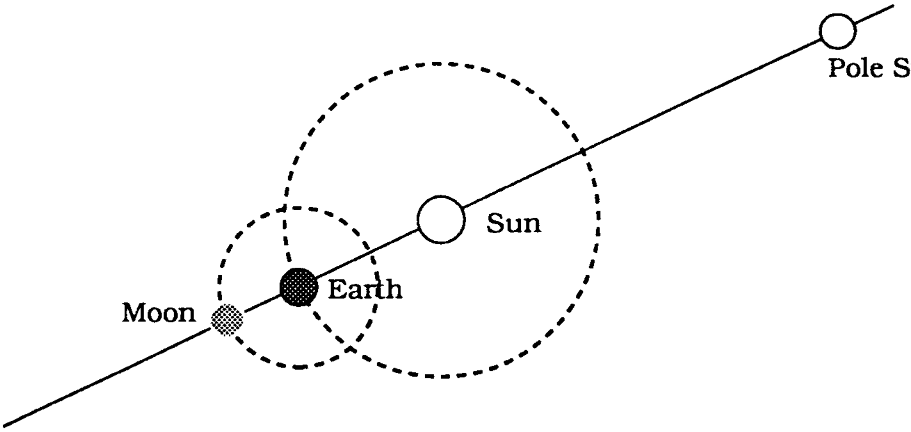

Figure 11.39. The beginning of the sixty-year cycle is based on the alignment of the sun, moon, earth, and North Star.

- Chen to the Dragon
- Si to the Snake
- Wu to the Horse
- Wei to the Sheep
- Shen to the Monkey
- You to the Rooster
- Xu to the Dog
- Hai to the Boar

The first year of the sixty-year cycle is associated with the first Branch and the first Stem. The first Branch, "Zi," corresponds with the Rat; the first Stem, Chia, corresponds with the element Wood. Thus, the first year of the sixty-year cycle is the year of the Wood Rat in Chinese astrology.

The second year of the cycle represents the second Branch and second Stem, and so on. When you reach the 11th Branch, there are no more Stems and, thus, the 11th year of the cycle represents the eleventh Branch and the first Stem. The cycle continues in this manner until it has transitioned through six cycles of Stems and five cycles of Branches and returns to the first Branch and first Stem paired together.

Figure 11.40 shows these yearly cycles; for example, 1997 was the year of the Fire/Ox, so the year 1998 is the year of the Earth/Tiger.

The Ten Heavenly Stems represent Yang, and the higher form of Heaven's vitality (Figure 11.41). The Twelve Earthly Branches represent Yin and the lower form of Earth's vitality (Figure 11.42). Together they manifest the transformation of all natural events expressed through the growing, thriving, declining, and dying of all living things.

1. Each lunar year the Earthly Branch time divisions are divided into 12 separate month divisions.
2. Each day is divided into 12 two-hour Earthly Branch time divisions during a 24-hour cycle.

SUMMARY

In comprehending the energetic potential of the Three Treasures of Heaven, the Qigong doctor evaluates the patient's condition and treats the disease according to the patient's energetic rhythm and flow, which is blended with the Heavenly and Earthly energies. The treatment also includes prescribing homework (the Medical Qigong prescriptions needed to support and maintain the doctor's treatments).

Although mankind is terrestrial, there can be no separation from the influence of the celestial. For mankind is born on Earth in physical form, created through the integration of both celestial (Heavenly) and terrestrial (Earthly) energies. According to statements recorded in the *Canon of Internal Medicine*, ancient doctors had already rec-

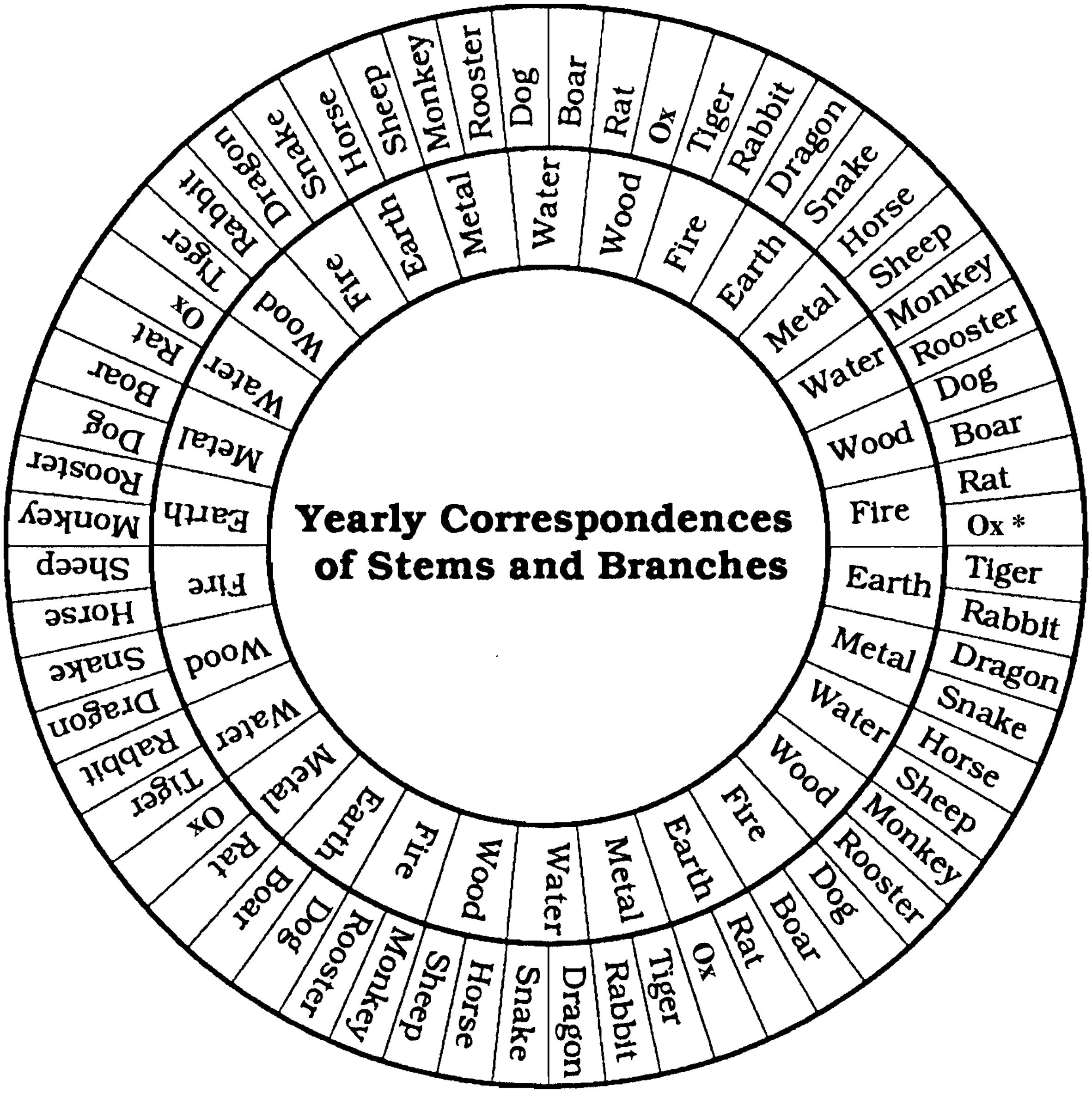
Yearly Correspondences of Stems and Branches
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox *
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Boar
Rat
Ox
Tiger
Rabbit
Dragon
Snake
Horse
Sheep
Monkey
Rooster
Dog
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal
Water
Wood
Fire
Earth
Metal

Figure 11.40. The asterisk (*) shows the year 1997, which was the year of Fire/Ox. Each year, one of the twelve animals of the Chinese zodiac flourishes and brings its energy into action. Every two years, a new Element (Fire, Wood, Water, Earth, Metal) dominates in the cycle of creation. The chart moves clockwise, so 1998 was the year of Earth/Tiger.

Character	Name	Number	Animal	Month	Time	Branch	Channel	Image	Element
子	Zi	1	Rat	Nov. 22–Dec. 21	11-1 am	**Little Yang**	Gall Bladder	Reproduction of Life	**Water**
丑	Ch'ou	2	Ox	Dec. 22–Jan. 20	1-3 am	**Decreasing Yin**	Liver	Initial Growth	**Earth**
寅	Yin	3	Tiger	Jan. 21–Feb. 19	3-5 am	**Great Yang**	Lungs	Breaking Ground	**Wood**
卯	Mao	4	Rabbit	Feb. 20–Mar. 20	5-7 am	**Bright Yin**	Large Intestine	Thick Vegetation	**Wood**
辰	Ch'en	5	Dragon	Mar. 21–Apr. 19	7-9 am	**Bright Yang**	Stomach	Full Awakening	**Earth**
巳	Si	6	Snake	Apr. 20–May 20	9-11 am	**Great Yin**	Spleen	Preparation for Ripening	**Fire**
午	Wu	7	Horse	May 21–June 21	11-1 pm	**Little Yang**	Heart	Peak of Growth	**Fire**
未	Wei	8	Sheep	June 22–July 21	1-3 pm	**Great Yin**	Small Intestine	Taste of Full Ripeness	**Earth**
申	Shen	9	Monkey	July 22–Aug. 21	3-5 pm	**Great Yang**	Bladder	Time of Harvest	**Metal**
酉	You	10	Rooster	Aug. 22–Sept. 22	5-7 pm	**Little Yin**	Kidneys	Resting after Harvest	**Metal**
戌	Xu	11	Dog	Sept. 23–Oct. 22	7-9 pm	**Decreasing Yang**	Pericardium	Moving Inward	**Earth**
亥	Hai	12	Boar	Oct. 23–Nov. 21	9-11 pm	**Little Yin**	Triple Burners	Seed Awaiting Growth	**Water**

Figure 11.41. Environmental Energetic Interactions of the Twelve Earthly Branches

ognized the close relationship between the body's Qi and Blood and the influences of the natural environment. In ancient China, the Twelve Earthly Branches represented time units, each of them corresponding to a two-hour period. They are primarily used to represent the 12 months in the Lunar Calendar.

The Ten Heavenly Stems are made up of the Five Elements, each of which is divided into pairs of corresponding Yin and Yang, resulting in a ten-step system of numerology. Using the Ten Heavenly Stems allows the Qigong doctor to determine the ebb and flow of the Qi of Heaven and its corresponding relationship to the Qi of Man.

Both the Ten Heavenly Branches and Twelve Earthly Stems represent the characteristics of growing, thriving, declining, and dying of all living things in the universe, as well as the development and transformation of all natural phenomena.

The energy of the Ten Heavenly Stems can be observed within five energetic movements (front, back, right, left, and center), as well as within the elemental energy of the body's major internal organs.

Character	Name	Number	Element	Yin-Yang	Organ	Planet	Image
甲	Jia	1	Wood	Yang	G.B.	Jupiter	Sprouting Growth
乙	Yi	2	Wood	Yin	Liver		Early Growth
丙	Bing	3	Fire	Yang	S.I.	Mars	Expansive Growth
丁	Ding	4	Fire	Yin	Heart		New Life Fully Grown
戊	Wu	5	Earth	Yang	Stomach	Saturn	Prosperous Development
己	Ji	6	Earth	Yin	Spleen		Maturation
庚	Geng	7	Metal	Yang	L.I.	Venus	Retreating Inward
辛	Xin	8	Metal	Yin	Lungs		Hibernation
壬	Ren	9	Water	Yang	Bladder	Mercury	Inward Nourishment
癸	Kui	10	Water	Yin	Kidneys		Gathering New Life Energy

Figure 11.42. The Ten Celestial Stems and their Universal Energetic Interactions

CHAPTER 12
THE THREE TREASURES OF EARTH: EARTH, WATER, AND WIND

INTRODUCTION TO THE THREE TREASURES OF EARTH

The Outer Force (Power) of Earth's Three Treasures consists of the Yin and Yang interaction of energy and light, emitted, absorbed, and reflected back from the land, water, and wind. The Three Earthly Treasures are brought about by the waxing and waning of the five Yin and Yang climatic transformations (the five seasonal growth cycles), which in turn are caused by the sun, moon, and stars. Each seasonal change brings a life transforming gift that affects the mind, body, emotions, and spirit (Figure 12.1). The wind (weather) reflects the various conditions and transformations of energy in the sky, and is viewed as an Earthly manifestation of Heaven's moods.

1. Earth Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold energy originating from the Earth's surface.
2. Water Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold energy originating from the vast oceans, lakes, and rivers of the world.
3. Wind Qi consists of the Yin and Yang interactions of energy and light, manifesting as hot and cold air circulation created from the interaction of the sun's heat on the surface of the planet. This interaction includes all cloud and barometric pressure formations (e.g., tornadoes, hurricanes, lightning storms, etc.).

UNDERSTANDING HEAT AND LIGHT INTERACTIONS

Heat and light are manifested both externally on and from the Earth's surface, and internally within the Earth.

1. Externally, the greater the intensity of the sun's light absorbed into the surface of the Earth, the more heat is manifested from the Earth's surface. Cloudy days diminish the

Natural Earthly Transformations					
Organs	**Element**	**Season**	**Elemental Factor**	**Temperature**	**Transition**
Liver	Wood	Spring	Wind	60– 70° F	Germination
Heart	Fire	Summer	Heat	above 90 F	Growth
Spleen	Earth	Late Summer	Damp	75– 90° F	Transformation
Lungs	Metal	Autumn	Dry	60– 70° F	Gathering
Kidneys	Water	Winter	Cold	below 60° F	Storage

Figure 12.1. The Natural Earthly Transitions

sun's thermal interaction and tend to bring coolness.

2. Internally, heat and light, as well as other natural forces, are contained within the Earth itself. The sun's intense energetic fields draw the heat and light from within Earth, which are then reflected back to the Earth by the atmosphere.

The Earth's air extends only a comparatively short distance out from the Earth's surface, and the effects of the heat rays decrease as you leave the Earth's surface and ascend toward the outer regions of the atmosphere. As the air becomes less dense, there is less reflection. The higher you ascend, the cooler the air becomes. When you have reached the limit of air, you have reached the limit of the Earth's heat.

The heat and light form a protective energetic field around the Earth's surface similar to the body's Wei Qi field. Once you have reached the limit of heat and light, you reach what is called the Great Cold. This cold is considered far more solid than steel, and it presses down upon the Earth's energetic field and atmosphere with an almost irresistible force, holding them together.

UNDERSTANDING THE EARTH'S YIN AND YANG INTERACTIONS

Between Heaven and Earth there is an on-going interaction of Yin and Yang energy exchange. When the Earth Qi is in balance, plants grow and animals thrive. Power, in the form of energetic fields, is created by the integration of Yin and Yang. This power directly and indirectly affects the body's electromagnetic field, as well as the regulation and formation of the body's tissues. Visualize the Earth's energy as Qi resonating and emitting from the Earth itself. This energy includes the electromagnetic fields, subterranean radiations, and thermoluminescence (light and heat emitting from the center of the Earth).

The closer the Qigong doctor bonds physically, mentally, emotionally, and spiritually with the Earth and nature, the easier it is to tap into the Earth's energy and vibration. The development of Earth Qi in the body and the body's connection to the energy of the Earth are mutually supportive. The first step in internal cultivation involves focusing the mind on the Lower Dantian (with the aid of breathing techniques) to enhance the mind-body connection; this interactive connection develops a relationship with the Earth Qi.

GATHERING ENERGY FROM THE EARTH'S ENERGETIC FIELDS

The ability to gather energy from nature and the environment is extremely important to the Qigong doctor. The energy of the Earth stays within the structural components of the planet, flowing like mighty rivers of Qi. It is absorbed into the terrain and vegetation, as well as the water that we drink. Environmental Qi is specific to each geographic area. It is the energy that has been collected from the structural components of the Earth itself (i.e., soil, fields, deserts, plants, trees, forests, mountains, streams, rivers, lakes and oceans, etc.). The energetic potential of the Earth's Qi changes from Yang to Yin energy with the rising and setting of the sun, and therefore directly affects the environmental energy.

Although food and water are excellent sources for gathering energy, the Qigong doctor cannot continually ingest food, or drink herbal teas while treating patients, without overtaxing his or her digestive system. It is, therefore, important for the doctor to find an alternative source of energy, to replenish his or her system once it becomes depleted. By studying the waxing and waning of the Earth's Yin and Yang energies the doctor can optimize the absorption of Yin and Yang energy.

TECHNIQUES FOR CULTIVATING EARTH QI

Generally, several meditations are used to cultivate and absorb energy from the Earth by utilizing the energy of trees, bushes, flowers, lakes, etc. The primary objective of these meditations is to first cleanse the body of pathogenic factors, then cultivate and absorb the environmental Qi to replenish the body's energetic field. Practitioners and patients should visualize themselves immersed in a focused environmental field of energy, filled with its vibration, color, and light. They should allow this energized field to envelop their entire body.

CLEANSING THE BODY

One of the most popular cleansing techniques used in the Xi Yuan Hospital's Qigong Clinics begins from a lying, sitting, or standing posture (usually in a Wuji posture) while in front of a natural object, such as a tree.

To cleanse the body's energy, extend both hands towards the tree, as if to embrace it (it does not matter if the tree is actually being touched). Upon inhalation, begin to guide the energy from the tree into your torso through the Laogong (Pc-8) points at the center of your palms, and the Baihui (GV-20) point at the top of your head. Absorb this energy deep into your body. Once you feel that the body is fully saturated with Qi, exhale through the mouth and guide the Turbid Qi down the legs and out of the body through the Yongquan (Kd-1) points at the bottom of the feet. Continue to guide the Turbid Qi down through the Earth into the root system of the tree. Then inhale through the nose, as you absorb the purified energy from the upper part of the tree leading it down your body through the top of your head. Continue this process until you feel totally cleansed. Remember to synchronize breathing in through your nose as you focus on absorbing the Earth Qi, and to exhale through the mouth, as you focus on dispersing and releasing the Turbid Qi.

For tonification, extend your Qi and intention deep into the root system of the tree and imagine absorbing its natural energy through the Baihui point at the top of your head (Figure 12.2).

GATHERING QI FROM TREES, BUSHES, AND FLOWERS

Earth Qi and environmental energy saturate all types of foliage, allowing the Qigong doctor to select, gather, and absorb Qi from many botanical sources. Throughout history, the Chinese have used trees, bushes, and flowers for healing and medicine. Trees and plants absorb air, light, energy, water, and minerals. When Heavenly Qi is combined together with the environmental energy of the tree or plant, the result is a powerful invigorating energy source.

In China, I noted that every Medical Qigong clinic I visited had its own garden. Each garden was full of healthy trees, bushes, and beautiful flowers from which the patients could draw energy. The patients could choose from different plants and bushes, each type of foliage having its own unique strengths and benefits. Any combination of healthy plant energies can be absorbed by the patients to improve their health.

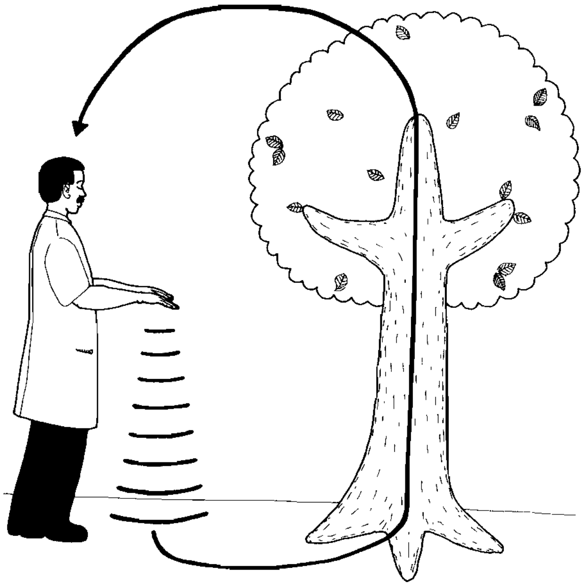

Figure 12.2. A Qigong doctor can extend Qi and intention deep into the root system of the tree and absorb its natural energy through the Baihui point at the top of the doctor's head.

The best time for absorbing energy from trees, bushes, and flowers is between the hours of sunrise (Mao: 5 - 7 am.) and noon (Wu: 11 a.m. – 1 p.m.).

It is important to note that, within the differing regions of a specific country, each type of tree and plant will have either minor or major variances regarding its energetic potential. This difference in energetic potential is due to the diverse environmental influences affecting each plant. We in the West, for example, have many different varieties of plants and trees similar to those in Asia, however, because of the energetic continental differences, the quality and quantity of Qi in each plant or herb can vary. This energetic difference is brought about by the various Qi fields produced by each region's altitude, the mineral content of the soil and water, and the amount of exposure to the energy of the sun. Even herbs of the same name, but grown in different countries, can have different affects on the body's energetic system, slightly altering their clinical potential (e.g., American Ginseng is sweet, slightly bitter and cold, and enters the body affecting the Heart, Lung and Kidney Channels; Siberian Ginseng is sweet, slightly bitter and slightly warm, and enters the body affecting the Lung and Spleen Channels).

GATHERING ENERGY FROM TREES

Trees are tremendously powerful plants and are commonly used by both doctors and patients to absorb and transform the negative pathogenic Qi into clean, healthy, life-giving energy. Pine trees tend to be one of the most powerful, as they radiate a great deal of Qi. Once an individual has connected with a tree, the primary goal is to:

1. remove stagnation from the channels,
2. tonify the internal organs,
3. stabilize and replenish any depleted energy within the body,
4. nourish the Blood, and
5. strengthen the nervous system.

The following is a brief description of various trees, their Yin or Yang potential, and their energetic affect on the body's internal organ system:

- Apple - slightly Yin - St and Sp
- Bamboo - Yin - Ht, Lu, GB, and St
- Birch - Yin - St
- Cassia (Cinnamon) - Yang - Sp, Kd, and Bl
- Cherry - Yang - Lv, Lu, Kd, and St
- Chestnut - Yang - Sp, St, and Kd
- Cranapple - neutral - Lv, Lu, and Ht
- Elm - neutral - St, SI, and LI
- Fig - neutral - Lu, LI, and Sp
- Ginkgo - neutral - Lu, Kd, and Ht
- Hawthorn - slightly Yang - Sp, St, Lu, and LI
- Magnolia - Yang - Lu, St, GB, LI, and Sp
- Maple - neutral - Sp, Kd, and LI
- Mulberry - Yin - Lu, LI, Sp, and Kd
- Oak - slightly Yang - LI, Sp, Kd, and Ht
- Pine - Yang - Ht, Lu, LI, and Kd
- Willow - Yin - Ht and Sp

GATHERING ENERGY FROM BUSHES

Bushes are an extremely effective source for gathering and ingesting different forms of healing energy. Similarly to trees, each bush has its own energetic properties and causes a specific reaction in the body's energetic field. Bushes, however, are not as powerful an energy source as trees for gathering massive quantities of Qi.

GATHERING ENERGY FROM FLOWERS

Flowers are unique in their effect on stimulating the nervous system. Different colors, shapes, and sizes can affect the emotions, causing the individual's spirit to open. Each color can be absorbed into the body in order to facilitate the stimulation of internal organ energy. The gift of flowers has had a long history in soothing emotional pain and bringing relief to patients.

CULTIVATION AND REGULATION TECHNIQUES

Techniques used in the Qigong clinics for cultivation, absorption and regulation of energy generally begin by having the patient face a tree, bush, flower, etc. Although the hands and body postures are similar to the last meditation, the mind's intention is quite different. In this meditation the patient absorbs Qi from the tree, by inhaling it through his left palm, to draw the energy into his body. The energy is guided down the patient's chest, into his or her Lower Dantian. From there the patient circulates this gathered energy through his or her Microcosmic Orbit. As the patient exhales, the Qi exits the right palm into the tree, and begins the cycle again. This meditation is practiced to replenish the body's energetic field.

Patients should be monitored regularly to prevent Qi deviations, which may occur while absorbing Qi through their pores. For best results, choose a quiet, safe, and healthy environment. Select trees, bushes, and flowers with a stable root, or thick trunks. The color is also very important, as the leaves and flowers should be bright, full, and healthy.

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any area which is unpleasant or unhealthy. Avoid any tree, bush, or flower that is sick, dying, has lost its color, or has just been pruned. Using trees, bushes, or flowers for cleansing the body, as well as to absorb energy, is forbidden at night, because their oxygen production diminishes after the sun has set.

PRECAUTIONS

It is important not to meditate in front of any tree, bush, or flower that has parasites, or has been poisoned or polluted, since such vegetation will induce an impure energetic resonation within the Qigong doctor's body.

GATHERING QI FROM MOUNTAINS, VALLEYS, AND DESERTS

The mountains, valleys, and deserts absorb and release light, energy, and heat from the sun. These energies are absorbed and released from the structures of the land quicker than from the oceans, lakes, and streams. Environmental energy is also gathered and stored by the geological formations and local ecology. The ancient Chinese Qigong masters viewed the Earth as a living entity, with rivers and pockets of energy similar to that of the human body. Certain altitudes, densities of colors, and structural formations can affect the Earth's energetic potential.

GATHERING ENERGY FROM MOUNTAINS

Mountains are extremely powerful conduits of energy. They act as peak areas, or collection points for energy. The higher the mountain, the more the air is charged with electromagnetic potential. In ancient times, for prayer and meditation, Taoist and Buddhist temples, shrines, and caves were constructed on energetic pockets high within the mountains of China. In fact, all high, mountains that are far from human civilization are regarded as superior places for energetic cultivation, due to the clarity of light and extraordinarily potent energies which include a very high negative ion count in the air.

GATHERING ENERGY FROM VALLEYS

The Earth's valleys act as channels for Earth energy, collecting and carrying energy throughout the natural terrain. Systems of valleys extend through the plains, hills, and mountains, and generally flow towards the ocean. By standing at the "floor" of the valley (surrounded by its walls), energy can be easily absorbed into the body for cultivation.

GATHERING ENERGY FROM DESERTS

The deserts are considered seas of dry heat energy. They are excellent environments for meditations when used to gather Qi into the body, especially for combating diseases pertaining to Wind, Cold, and Damp invasion (which give rise to many diseases).

CULTIVATION TECHNIQUES

Generally there are several meditations used to cultivate and absorb energy from mountains, valleys, and deserts. The primary objective of these meditations is either to cleanse the body of pathogenic factors, or to cultivate and absorb Qi. When cultivating energy from mountains, valleys, or deserts, practitioners should visualize themselves immersed in the Earth's field of energy, filled with vibration, color, and light. The practitioners should allow this energized field of Earth Qi to envelop them completely.

For best results, select mountains, valleys, and deserts with a stable foundation. The color is also very important, as the Earth, soil, rocks, and surrounding area should be clean.

Also, there are certain locations on the planet known as "power spots" which can produce remarkably powerful effects in the body's energetic fields.

Patients should be monitored regularly in order to prevent Qi deviations, which may occur while absorbing Qi through their pores.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important not to meditate in front of any mountain, valley, or desert area that has eroded, is dying, has lost its color, or is polluted. Also prohibited are areas where there is ongoing seismic, or volcanic activity, as such areas will induce unstable resonation within the Qigong doctor's body.

GATHERING QI FROM OCEANS, LAKES, AND STREAMS

The oceans, lakes, and streams retain and release the sun's light, energy, and heat slowly and are, therefore, important in energy cultivation. The body instantly absorbs the light, energy, and resonant vibrations stored within water. Energized water (sometimes in the form of herbal teas) is often used by Qigong doctors in clinics throughout China. This energetically transformed water provides the patient's body with the basic components necessary for restoration and regeneration, and is considered the "elixir of life" by many Qigong doctors.

GATHERING ENERGY FROM OCEANS

The energy from the ocean has a strong cleansing and purifying effect on the body's Qi. Meditating by the ocean is used extensively for dispersing negative emotions and regulating the body's internal organs. The active rhythm and sound of the waves allows the patient to harmonize his or her internal energetic rhythms with that of the tide, thus quieting the nervous system.

GATHERING ENERGY FROM LAKES

The energy from a lake has a strong calming affect on the body's energetic field. The quiescent solitude of a lake can be used to sedate active emotions, balance any Excess or Deficiency, and calm the body's Shen.

GATHERING ENERGY FROM STREAMS

The energy of a stream can be used to replenish and restore depleted Qi, sedate active emotions, balance any Excess or Deficiency, and also to calm the body's Shen.

CULTIVATION TECHNIQUES

Generally there are several meditations used to cultivate and absorb energy from oceans, lakes, and streams. The primary objective of these meditations is to either cleanse the body of pathogenic factors, or cultivate and absorb Qi to replenish the body's energetic field. When cultivating energy from oceans, lakes, or rivers, patients should visualize themselves immersed in a pool of water, filled with energy, color, and light. The patient should allow this energized water to absorb, cleanse, and invigorate the entire body.

Patients should be monitored regularly to prevent Qi deviations, which may occur while they are absorbing Qi through their pores. For best results, select oceans, lakes, and streams with a calm, stable current. The color is also very important, it should be clear and clean. The water and the surrounding areas should also be clean.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any oceans, lakes, or streams that are turbulent, have become dull in color, polluted, or stagnant.

GATHERING QI FROM THE SIX DIRECTIONS

Absorbing Qi from the six directions (Heaven, Earth, and the Four Winds), focuses on gathering environmental energy into the center core of the body (Taiji Pole) from: Heaven, Earth, and the four directions: South corresponds to the front of the body, North to the back of the body, West to the right of the body, and East to the left of the body. The Taiji Pole is considered the central axis, with two directions, above and below. This is a tonification exercise that is used for gathering and absorbing the Earth's environmental Qi, to energize a weakened or Deficient condition.

CULTIVATION TECHNIQUES

There are several meditations used to cultivate and absorb energy from the four directions of the horizon, with the fifth and sixth directions being at the top and bottom of the Taiji Pole. The primary objective of these meditations is to first cleanse the body of pathogenic factors, then to cultivate and absorb environmental Qi. Each direction of energy (N, S, E, W, Heaven, and Earth) has its own unique field of power and vibration, and is assigned a specific color. Each vibration and color stimulates the energy field of one of the five major Yin organs through the body's Heaven and Earth connection (i.e., the Taiji Pole).

When cultivating energy from the horizon, practitioners should face the direction of the south and visualize themselves surrounded in an energetic mist. This mist contains six different colors and light vibrations. From the four directions (N, S, .E, and W) emanate the corresponding following four colors: ruby red, brilliant white, indigo and emerald green. Each of these colors are associated with a particular Yin organ: ruby red with the Heart, brilliant white with the Lungs, indigo with the Kidneys, and emerald green with the Liver. From the Earth itself arises a golden yellow mist; this color is associated with the Spleen. From Heaven descends a silvery white mist entering the body at the top of the head; this color is associated with the Taiji Pole. These six different colored mists permeate and envelop the body with their healing light vibrations.

To perform this meditation, stand in a Wuji posture as you visualize:

1. A silvery white mist descending from the center of the Heavens to penetrate and envelop your Taiji Pole. This silvery white mist enters your body through the top of the head at the Baihui (GV-20) point, and flows into and fills up the Taiji Pole, saturating and absorbing into your center core.
2. A golden yellow mist, ascending from the center of the Earth, entering the body through the bottom of your feet at Yongquan (Kd-1) points. This golden yellow mist saturates your Spleen as it penetrates and envelops your center core;
3. A ruby red mist, flowing from the horizon in front of you enters your Heart, as it saturates and envelops the front of your body.
4. An indigo or dark blue mist flowing from the horizon behind you enters your Kidneys, as it saturates and envelops your back.
5. A brilliant white mist flowing from the horizon on your right side enters your Lungs, as it saturates and envelops the right side of your body.
6. An emerald green mist flowing from the horizon on your left side enters your Liver, as it saturates and envelops the left side of your body.

Patients should be monitored regularly to prevent Qi deviations which may occur while the patient is absorbing Qi through his or her pores. For best results while absorbing environmental Qi, the color of the energy should be clean, clear and bright.

PRECAUTIONS

Because the body's internal energy matches that of the external environment, it is important to avoid selecting any areas that are windy, turbulent, polluted, or stagnant. Also avoid cultivating energy in storms, during hurricanes, sweltering heat, or blistering cold. Avoid any type of extreme weather condition.

ADDITIONAL TECHNIQUES FOR TRAINING EARTH QI

The following exercises at one time were secretly practiced throughout China, as a means for Medical Qigong doctors and Qigong masters to practice advanced methods of environmental energy manipulation. The purpose of these Qigong exercises is to allow the doctors the chance to practice connecting and infusing their energetic field with the environmental energetic field.

PULLING CLOUDS OUT OF THE SKY

This exercise requires the Qigong doctors to first root and immerse themselves with the vibrational resonance of the Earth (this can be accomplished by practicing the One through Ten Meditation, described in Chapter 26). Next, the doctors extend an energetic line from the Earth into the sky, connecting to and enveloping a specific cloud. Through focused intention, the doctors imagine sucking the cloud out of the sky and down into the Earth, rooting its energetic form into the Earth's energetic field. As the cloud's energy is absorbed into the Earth, it dissolves from the sky.

Once the doctors have rooted the cloud into the Earth's energetic field, they have two choices: either to release the cloud back into the sky by attaching its energetic form onto an already existing cloud, or to allow the cloud to disperse into the energy of the surrounding environmental field.

DIVIDING A CLOUD IN HALF

After connecting and rooting into the Earth's energetic field, the Qigong doctors extend their intention into the sky, connecting with a particular cloud formation. Next, the doctors condense their upward intention, and like a laser beam, begin dividing and separating the cloud into two halves. After the cloud has been divided into two separate formations, the doctors can either continue to further divide and dissect the cloud or reconnect its structure.

THE EARTHLY TREASURE OF FOOD

The Earth's natural energetic forms and transformations affect the flow of the body's life-force energy. Food and water are examples of a natural form of Earth Qi, and are viewed as Earthly Treasures. Although the air we breath is representative of Heaven Qi, due to the influence of environmental energetic fields, it is also viewed as an Earthly Treasure. The objective of self-cultivation work is to continually purify the energy within the Microcosm of Man, and increase the awareness of the pure energy within the Macrocosm of the universe.

Because of its connection to the Earth's Yin and Yang energy fields, the body's life-force energy can be replenished through the consumption of food and herbs. Both food and herbs have Yin or Yang properties which, when consumed, help the body in moving, tonifying, purging or regulating the life-force energy.

UNDERSTANDING THE YIN OR YANG ASPECT OF FOOD

When trying to determine whether a food is more Yin or Yang, several characteristics should be evaluated. Yin foods take less time to grow, are more watery or wet, and are colder, and sweeter. Yang foods take more time to grow, are drier, hotter, and less sweet. The more Yin the food, the more expanded it will become (e.g., a pumpkin); the more Yang the food, the more contracted or dense it will become (e.g., a turnip).

Diet directly influences the body's ability to generate Qi. The body's ability to effectively digest food plays a big part in how efficiently the body utilizes the nutrients it receives.

UNDERSTANDING THE YIN OR YANG ASPECT OF HERBS

Food is a powerful healing tool. Medicinal herbs are considered "special foods," and are divided into Yin and Yang aspects throughout China. They are prescribed to be taken in soups (tang), teas, or as additives to other dishes; for example, Chinese Qigong doctors, as well as many martial arts masters, frequently drink herbal teas to help regulate their body's energies during seasonal changes. The herbal teas are consumed according to the following sequence.

1. The summer is considered the season of maximum Yang; green herbal tea is drunk in order to cool down the body and regulate the internal organ systems.
2. The autumn is the season when Yin and Yang energies balance; however, the Yin energy is increasing, so flower tea is drunk to help stabilize the body's Yin and Yang balance.
3. The winter is considered the season of maximum Yin; black or red tea is drunk to warm the body and regulate the internal organ systems.
4. The springtime is the season when Yin and Yang energies balance; however, the Yang energy is increasing, so flower tea is drunk to help stabilize the body's Yin and Yang balance.

Teas are frequently prescribed for their particular healing properties. Each herb has a specific nature. Herbs that are considered Hot or Warm are used to treat Cold conditions, and herbs that are considered Cold or Cool are used to treat Hot conditions. Herbs are also categorized according to their flavor (sour, bitter, sweet, pungent, or salty) and are used to affect the body's internal organ energy. Herbs help the Qi ascend or descend in order to purge or tonify an internal organ or channel. Herbs are also used in moving the Blood. Specific herbs are also chosen for prescription, based on the Differential Diagnosis according to the Eight Energetic Principles (see chapter 21), the Five Elements (see chapter 23), and the Six Stages (see chapter 24) theories .

Generally, because all herbs have medical uses and contraindications, they are collected and divided into three categories: superior herbs - which consist of tonics, common herbs - which are slightly toxic, and inferior herbs - which are toxic. To gather herbs, the doctor must have an extensive knowledge of the environment (season, place, time, etc.), and the specific herb (roots, bark, leaves, flowers, fruits, etc.) The doctor must also possess the harvesting skill required for gathering the herb. Once the herbs have been gathered, the doctor must then begin the procedure of processing them. The process work is important as some herbs are poisonous and need to be detoxified, while others have their properties enhanced through the extracting process.

ENERGY, FOOD, AND DIET

Many generations have asked the same question, "Is food medicine or medicine food?" An old Chinese proverb states, "To tonify with food is better than with medicine, but to tonify with Qi is better than food." Several decades of nutritional, epidemiological, and clinical studies point to a powerful link between disease and dietary habits.

For centuries, Chinese healers have studied food, discovering the properties of specific foods and the secrets of using them to achieve health and longevity. Chinese nutrition focuses on the energetic properties of food and their qualitative actions on the body as a whole, as well as their influences on the various internal organs and channels. Unless the patient eats correctly, Qigong, acupuncture, Chinese massage, and even herbal therapy will not be as effective, or long standing. In the Tang Dynasty, the great physician Sun Simiao wrote the medical classic *One Thousand Ounces Of Gold*, which described the dietary treatments of several diseases, including night blindness and goiter.

What we eat can either aggravate or enhance the healing of a particular condition. When Gu Qi (food energy) is being generated, strong surges of energy from the Spleen and Stomach (in the Middle Dantian) cause the body to generate more Yang Qi. As the body's Yin aspect weakens, its energetic nature becomes dryer and wanes; the Yang is no longer held in check and an overexuberant Yang begins to generate Heat, which produces, and sometimes releases, emotions.

Emotions begin with, and stem from, the energetic properties of Postnatal Qi. When ministering to patients, the doctor should ensure a healthy physical and social environment (e.g., clean air and a loving support system), as well as monitor the quality and quantity of food consumption. Hostile surroundings interfere with the proper digestion of food and cause toxic Qi. Food, air, and water become the fuel that generates energy in the body. When life-force energy becomes impure from poor diet, Heat accumulates in the organs, creating Excess Yang, or Excess Fire Qi.

One essential skill for harvesting the Earth's Treasures is controlling the intake of food and diet. It is important that a patient not eat foods that are excessively Cold, Hot, spicy, or greasy. If food is ingested while the food is too Yin (Cold) or out of harmony with the seasons, an external pathogenic factor can be created. When a Cold or Yin factor is created (a condition where Cold and Excess Yin consumes the Kidneys Yang, preventing the body's Yang Heat from warming the body), the Qi becomes sluggish, thus blocking the channels or collaterals. These blockages cause pain and damage to the Stomach and intestines and affect the Heart and Lungs.

If food is ingested while the food is too Yang (Hot) or out of harmony with the seasons, a Yang factor is created, causing internal energy to be released, which damages both the Blood and Qi. Because of its turbid nature, food that is too greasy may damage the Stomach Qi, causing boils and pyogenic infections, along with ulcerous skin diseases. Excessive eating may cause too much Gu Qi to develop, generating an excessive amount of power. This Excess power produces an upward reaction of Stomach Qi, which is detrimental to both the Spleen and Stomach, causing the breath to become obstructed (mother/Spleen affecting child/Lungs), and the psychic centers (Shen within the Heart) to become blocked. Many postnatal illnesses result from eating the wrong kinds of food in the wrong season, not eating enough of the right kind of food, overeating, or all of the above.

Balancing the diet in accordance with the cycles of nature and the principles of Yin and Yang

Balancing the Diet					
	Liver	**Heart**	**Spleen**	**Lungs**	**Kidneys**
Excess	Sour	Bitter	Sweet	Pungent	Salty
Controlled By	Pungent	Salty	Sour	Bitter	Sweet
Counteracts	Sweet	Pungent	Salty	Sour	Bitter

Figure 12.3. The Five Flavors and their Controlling and Counteracting Cycles

should be an integral part of the daily practice. In some situations, the body can compensate for an improper or unbalanced diet. To counteract this imbalance, however, the body will expend a great quantity of energy. Eating similar foods all the time results in the accumulation of toxins from those foods and can cause mildly allergic reactions. A balance of the different food groups is the key to dietary success.

USING THE FIVE FLAVORS OF FOOD

The energetic properties of the Five Flavors of food (sour, bitter, sweet, pungent, or salty) can be used to balance and control the energy of each organ. Any excess or deficiency of one flavor affects not only the organ to which it corresponds, but all the other organs as well (Figure 12.3). In the clinics in China, the theory of the Five Flavors or tastes is combined with the theory of the Five Elements to describe the interaction of the different "flavors" on the body's internal organs and energetic systems (Figure 12.4). The Five Flavors and their affect on the body are described in the following sections.

SOUR TASTE, ELEMENT WOOD

Some examples of Sour foods include: lemons, pickles, liver, vinegar, wheat, salad oils, and sour cream. Sour foods affect the Liver, eyes, and tendons. They are also excellent for stimulating the energy of the Liver and Gall Bladder.

1. Sour foods are astringent (causes shrinking), and can be prescribed to achieve the following treatment goals:
   - to prevent or reverse the abnormal leakage of Qi and Fluids,
   - to counteract diarrhea and excessive perspiration by slowing the movement of fluids,
   - to drain the Liver of Excess Qi, thereby indirectly strengthening the Lungs, and
   - to correct an irregular, slow heartbeat.
2. An excess of sour foods can cause the following dysfunctions:
   - injures the muscles,
   - hardening and wrinkling of the skin, and
   - stiffness spreading along the tendons that are controlled by the Liver. As the Liver controls the tendons, patients with Liver trouble should also eat less sour food.

BITTER TASTE, ELEMENT FIRE

Some examples of Bitter foods include: coffee, black and green teas, asparagus, celery, turnips, leeks, cabbage, broccoli, and cauliflower. Bitter foods affect the Heart, tongue, and Blood. They are also excellent for stimulating the energy of the Heart and Small Intestine.

1. Bitter foods are used for draining and drying, and can be prescribed to achieve the following treatment goals:
   - to stimulate the energy of the Heart, control the Heart Fire,
   - to reducing fever and Excess body Heat,
   - to stimulate digestion,

Five Elemental Cycle					
Element	Fire	Metal	Wood	Earth	Water
Season	Summer	Autumn	Spring	Late Summer	Winter
Organ	Heart	Lungs	Liver	Spleen/ Pancreas	Kidneys
Bowel	Small Intestine	Large Intestine	Gall Bladder	Stomach	Urinary Bladder
Superficial Organ	Blood Vessels	Skin	Nerves	Muscles	Bones
Release Point (Sense Organ)	Tongue	Nose	Eyes	Mouth	Ears
Taste	Bitter	Pungent	Sour	Sweet	Salty

Figure 12.4. The Five Flavors in Relationship to the Body's Internal Organs, Seasons, and Elements

- to drain Excess Spleen Qi, and
- to remove obstructions in the upper respiratory tract.

2. An excess of Bitter foods can cause the following dysfunctions:
   - dryness and congestion of the Spleen and Stomach; and
   - hyperactivity of the Heart Fire, and the consumption of Kidney Yin fluids.

SWEET TASTE, ELEMENT EARTH

Some examples of Sweet foods include: honey, sugar cane, orange juice, pecans, snow peas, corn, milk, and sweet potatoes. Sweet foods affect the Spleen, mouth, and muscles. They are also excellent for stimulating the energy of the Spleen, Stomach, and pancreas.

1. Sweet foods are used for tonifying, regulating and sometimes moistening, and can be prescribed to achieve the following treatment goals:
   - to stimulate the digestive Fire,
   - to tonify the Spleen, Stomach and pancreas, and
   - to drain the Heart of Excess Qi.
2. An excess of sweet foods can cause the following dysfunctions:
   - an imbalance in the Kidneys,
   - an aching in the bones,
   - hair loss from the head,
   - muscle dysfunctions, and
   - injury to the Spleen and Stomach. The Sweet flavor spreads through the muscles, so patients with diseases of Spleen, Stomach, or muscles (diabetes, hypoglycemia, systemic candida, fibriomyalgia, chronic fatigue, etc.) should limit their intake of Sweet foods.

PUNGENT TASTE, ELEMENT METAL

Some examples of Pungent foods include: garlic, onions, ginger, black and white pepper, mustard, and chilies. Pungent foods affect the Lungs, nose, skin, and body hair. They are also excellent for the stimulating the energy of the

Lungs and Large Intestine. Because the Lungs control the Qi, the Lungs are very quick in absorbing the energy of Pungent foods.

1. Pungent foods are used for dispersing and moving, and can be prescribed to achieve the following treatment goals:
    - to induce perspiration,
    - to improve circulation of Qi and Blood,
    - to stimulate the digestion,
    - to activate the Lung energy and promote the Lung's function,
    - to expand the Lung Qi,
    - to drain the Lungs of Excess Qi, and
    - to increase Fluids and saliva secretions.
2. An excess of Pungent foods can cause the following dysfunctions:
    - great loss of Lung energy and damage to the Lung organs,
    - knots in the muscles, and
    - withering or decay of the fingernails and toenails.

SALTY TASTE, ELEMENT WATER

Some examples of Salty foods include: shellfish, cheese, soy sauce, salt, ham/pork, and margarine. Salty foods affect the Kidneys, ears, and bones. They are also excellent for stimulating the Kidney Qi and Bladder Qi.

1. Salty foods are used for purging and softening, and can be prescribed to achieve the following treatment goals:
    - to soften hard nodules such as cysts,
    - to purge inflamed lymph glands, and
    - to soften knotted muscles.
2. An excess of Salty foods can cause the following dysfunctions:
    - cause damage to the Kidneys and Bladder,
    - cause the Blood to coagulate or clot,
    - cause Blood diseases,
    - harden the pulse, and
    - cause changes in the complexion.

GUIDELINES FOR THE FIVE ELEMENTAL TASTES

During each season, the five separate extremes in flavors serve as a catalyst to draw Qi into the Deficient organ. Overexposure to the seasonal elements (Cold, Wind, Damp, etc.) is responsible for external pathogenic factors that invade the body and affect the body's Jing, Qi, and Shen. The development of disease is due to an imbalance in the struggle between the body's health preserving Qi and the pathogenic influences. If the pathogenic factors are excessive, they can obstruct the body's normal physiological function, creating illness. If, however, the body's Qi is strong, the pathogenic factors will have a difficult time causing any energetic imbalances.

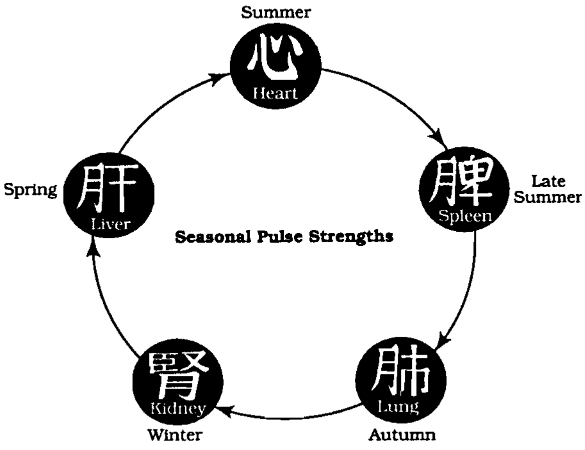

Figure 12.5. The quality of the body's pulse is constantly changing according to the rhythm of the five seasons.

Traditional Chinese Medicine divides the effects of these tastes into five seasons, each with its own unique function and flow of energy affecting the human body. One example of the effects of nature's rhythms on man is that the pulse quality in the viscera constantly changes in accordance with the five seasons (Figure 12.5). The Liver pulse is relatively strong in spring, compared to the other organs; in the summer the Heart is stronger, in the late summer - the Spleen, in the autumn - the Lungs, and in winter - the Kidneys.

Basic guidelines are developed for eating in accordance with the flow of the Five-Element cycle. These guidelines will help train the body to recognize what foods it needs. The food types and tastes are divided into five categories, which relate to the five seasons of the year, the Five Elements, and the Yin and Yang organs (Figure 12.6).

To achieve overall balance, food and energy must be equally balanced. If the Liver is weak, for example, eating Sour food will help correct it and Salty food will also help. This prescription is based on the Five Element theory, which focuses primarily on two cycles—the Generating Cycle and the Controlling Cycle (Figure 12.7).

OVERVIEW OF THE FIVE ELEMENTAL TASTES

The overall understanding of the Five Elemental tastes empowers the Qigong doctor to accurately assist in prescribing better eating habits for his or her patients. By understanding the seasonal eating patterns (i.e., in the winter eating root food, which is grown under the ground, and in the summer eating food that is grown via branches and flowers) the doctor can observe whether the patient is further compounding his or her disease condition through improper diet.

SUMMARY

In comprehending the energetic potential of the Three Treasures of Earth, the Qigong doctor will be able to evaluate the patient's eating and living conditions and their affect on the patient's disease. This also includes prescribing the Medical Qigong exercises, meditations, diets, and eating habits needed in order to support and maintain the doctor's treatments.

Through monitoring the patient's connection to their environment and eating patterns, the Qigong doctor is able to understand the Earthly influence on the patient's physical, mental, emotional, and spiritual being.

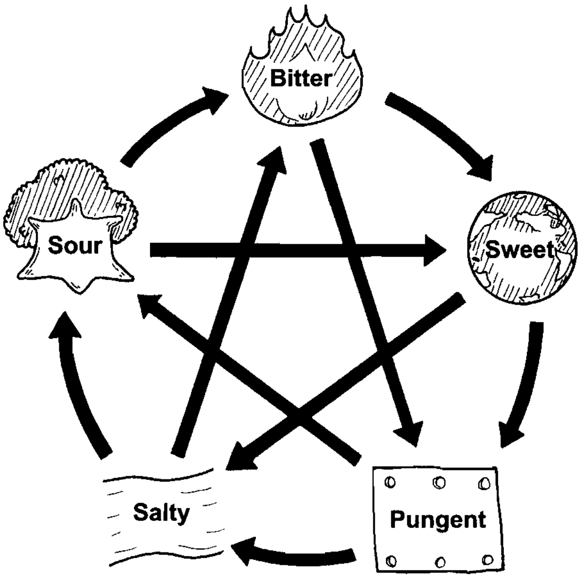

Figure 12.6. The Five Elements' Interaction of the Five Flavors

In the Creative or Generating Cycle—		
Salty Kidney (Water)	generates	(Wood) Sour Liver
Sour Liver (Wood)	generates	(Fire) Bitter Heart
Bitter Heart (Fire)	generates	(Earth) Sweet Spleen
Sweet Spleen (Earth)	generates	(Metal) Pungent Lung
Pungent Lung (Metal)	generates	(Water) Salty Kidney

In the Controlling Cycle—		
Salty Kidney (Water)	controls	(Fire) Bitter Heart
Bitter Heart (Fire)	controls	(Metal) Pungent Lungs
Pungent Lungs (Metal)	controls	(Wood) Sour Liver
Sour Liver (Wood)	controls	(Earth) Sweet Spleen
Sweet Spleen (Earth)	controls	(Water) Salty Kidneys

Figure 12.7. The Creative (or Generating) and Controlling Cycles of the Five Flavors

CHAPTER 13
THE THREE TREASURES OF MAN: JING, QI, AND SHEN

The Outer Force (Power) of the Three Treasures of Man are expressed through Jing (Essence), Qi (Energy), and Shen (Spirit). The human body is made up of vibrational frequencies that interact to create energetic fields and substances, or energetic materials ranging in structure from less dense to more dense (Figure 13.1). In order of increasing density, these energetic materials manifest as Spirit (Shen), Energy (Qi), Essence (Jing), Blood (Xue), Body Fluids (Ye/Jin), Marrow (Sui), and Bone (Gu). Of these energetic substances, Essence (Jing), Energy (Qi), and Spirit (Shen) are regarded as the most important. These three materials perform the most important functions in the human body, and are known as the Three Treasures of Man or San Bao.

The energy of the human body, or Man's Qi, is created from and through different interacting rates of vibration and energy fields that combine to form the body's cells, tissues, and organs, creating a living organism. In Chinese philosophy, the Three Treasures of Man also represent the interaction of Heaven and Earth within the human body.

Man is composed of these combined Heaven and Earth energies:

1. The Jing is the body's tangible essence, composed of a morphogenic field of energy that creates and sustains the body's physical form. It is the lower vibrational frequency of the human energy matrix. During the healing process, the patient's Jing directs the proper cells to multiply so that the body can heal rapidly and completely.
2. The Qi is the life-force energy that maintains the body's Jing. It is the electromagnetic body of the channel system and distributes vitality to the body through energetic transference. It

Natural Human Transformations									
Yin Organ	**Yang Organ**	**Sense Organ**	**Taste**	**Color**	**Body Smell**	**Sound**	**Tissue**	**Positive Emotion**	**Negative Emotion**
Liver	Gall Bladder	Eyes	Sour	Green	Rancid or goatish	Shout	Nerves & tendons	Kindness, benevolence	Anger
Heart	Small Intestine	Tongue	Bitter	Red	Scorched or burned	Laugh	Blood vessels	Love, joy, respect	Elation, arrogance, impatience
Spleen	Stomach	Mouth	Sweet	Yellow	Sweet, fragrant	Sing	Muscles & fascia	Fairness, openness	Worry, anxiety
Lungs	Large Intestine	Nose	Pungent	White	Tidepool, fishy	Weep	Skin & hair	Righteous, courageous	Sadness, grief
Kidneys	Bladder	Ears	Salty	Blue	Putrid, rotten	Groan	Bones	Gentleness, wisdom	Fear

Figure 13.1. This diagram shows the natural transformations of man's internal organs, energy, and matter, as well as their natural manifestations.

is the middle vibrational frequency of the human energy matrix. The vibrational field of Qi sustains and contains the thoughts and emotions, and interacts between the vibrational fields of Jing and Shen.

3. The Shen is the spirit and psyche which maintains the body's Qi. It is the higher vibrational frequency of the human energy matrix. It is also the energetic field which sustains the vibration of the Mind (the process of awareness and consciousness located throughout the body, not limited to the brain), interfacing between the Mind and Qi.

These three substances resonate at different frequencies and are mutually dependent upon each other for maintenance and survival. Low frequencies are attributed to the resonance of Jing, medium frequencies to the resonance of Qi, and high frequencies to the resonance of Shen. Metaphorically speaking, Jing can be compared to a dense, solid material such as ice, Qi to a flowing material such as water, and Shen to a vapor.

In Chinese medicine, the Qigong doctor uses the Three Treasures of Man for assessing and diagnosing the patient. Observing the state and interaction of the patient's Jing, Qi, and Shen allows the doctor to determine the patient's prognosis and chances for successful treatment. The doctor can assess the Jing, Qi, and Shen by:

1. Observing the general vitality of the patient's Jing through pulse reading and tongue diagnosis,
2. Noting the general vitality of the patient's Qi through observation of the Wei Qi field, and
3. Observing the general vitality of the patient's Shen by the quality of glitter in the patient's eyes.

Within this model, Jing, Qi, and Shen are viewed as layers of energetic fields between the body and the Mind. Jing is the innermost energetic layer, the blueprint which holds the physical body to the desired form. Shen is the outer energetic layer of the body through which the Mind experiences and directs the activity of the body. Qi is the intermediate energetic layer, the medium which allows the Jing and Shen (the energetic fields of the body and spirit) to communicate.

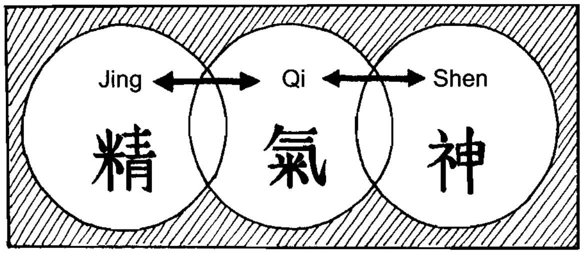

Figure 13.2. The Three Treasures of Man

PRENATAL AND POSTNATAL JING, QI, AND SHEN TRANSFORMATIONS

Jing is the body's foundational substance, responsible for nourishing the tissues. Qi emerges out of Jing circulating through the tissues. It promotes the body's metabolism. Shen governs the body. The body contains both prenatal and postnatal versions of Jing, Qi and Shen.

1. Prenatal Jing (Yuan Jing) is the body's innate and true Original Essence, which determines the constitutional makeup, strength, and vitality. It is not the reproductive Essence (sperm and ovary essence), which is considered Postnatal Jing.
2. Prenatal Qi (Yuan Qi) is the body's Original Energy, extending into the Wuji (infinite space or Void). It is the motivative energetic force linking the Prenatal Jing with the Prenatal Shen. It is not inhaled air (oxygen), which is considered Postnatal Qi.
3. Prenatal Shen (Yuan Shen) is the Original Spirit, that is able to perceive and intuit truth and knowledge (knowing without knowing). It is not cognitive thinking or consciousness, which is considered Postnatal Shen.

Through meditation, the body's alchemical processes of internal energy transmutations can be cultivated. Jing can be used to create or increase Qi; likewise, Qi can be used to create or expand the Shen. This creative cycle also works in reverse order by transforming Shen to Qi, and Qi to Jing (Figure 13.2).

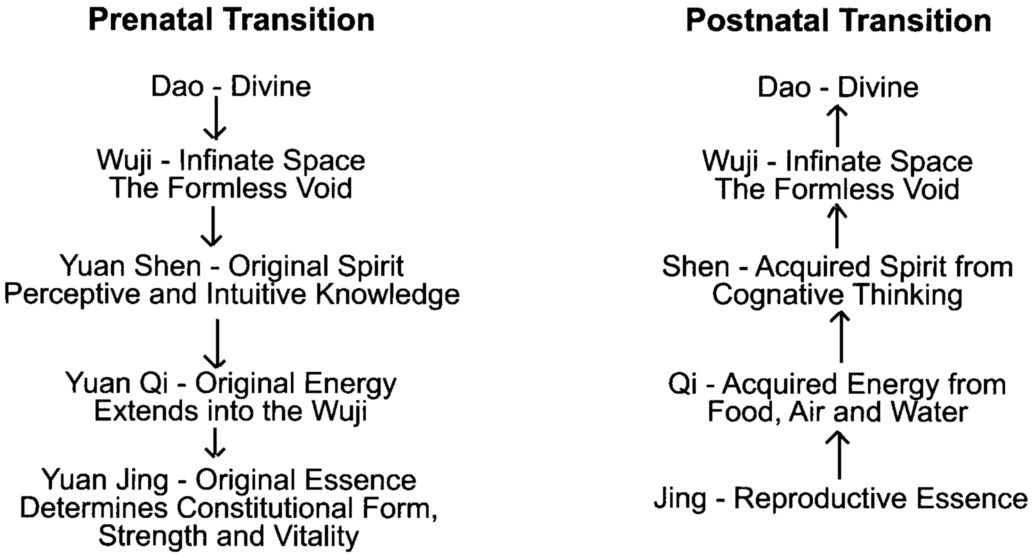
Prenatal Transition
Dao - Divine
Wuji - Infinate Space
The Formless Void
Yuan Shen - Original Spirit
Perceptive and Intuitive Knowledge
Yuan Qi - Original Energy
Extends into the Wuji
Yuan Jing - Original Essence
Determines Constitutional Form,
Strength and Vitality
Postnatal Transition
Dao - Divine
Wuji - Infinate Space
The Formless Void
Shen - Acquired Spirit from
Cognative Thinking
Qi - Acquired Energy from
Food, Air and Water
Jing - Reproductive Essence

Figure 13.3. The Prenatal and Postnatal Alchemical Transitions of Jing, Qi, and Shen

PRENATAL TRANSFORMATIONS

The following transformations describe the origin and creation process of the Three Treasures within the body of the fetus (Figure 13.3). Through comprehending these transformations, the Qigong doctor can teach individuals how to cultivate their Jing. The body's prenatal energies, which support the body's postnatal Jing, are sustained by prayer, meditation, and sleep. This quiescence state is necessary for the body to reenergize its Prenatal Qi.

1. Divine Energy (Dao) transforms into infinite space (Wuji). At this beginning stage, the spirit of the God envelops infinite space, including the energetic fields of the Heavenly universe and the environmental Earthly fields.
2. Infinite space (Wuji) transforms into Spirit (Yuan Shen). At this transition, the spirit of God descends to permeate, unite, and dwell within the Original Spirit (Yuan Shen) of the forming fetus.
3. Spirit (Yuan Shen) transforms into Energy (Yuan Qi). At this stage of transition, the Original Spirit of the fetus transforms into Original Energy (Yuan Qi).
4. Energy (Yuan Qi) Transforms into Essence (Yuan Jing). At this final stage, the Original Energy of the fetus is transformed into Original Essence (Yuan Jing).

POSTNATAL TRANSFORMATIONS

The following transformations describe the origin and creation process of the Three Treasures within the body after birth. Through comprehending these transformations, the Qigong doctor is able to achieve a harmonious balance between the three energies through Postnatal Jing cultivation. The body's postnatal energies are supported and sustained first of all through food, drink, and air.

1. Essence (Jing) Transforms into Energy (Qi). At this beginning stage, the body's Jing is Heated within the Lower Dantian to create Qi. This transition is similar to ice melting into water.
2. Energy (Qi) Transforms into Spirit (Shen). At this transition the body's Qi is steamed within the Middle Dantian, to be transformed into Shen. This transition is similar to water transforming into vapor.
3. Spirit (Shen) Transforms into infinite space (Wuji). At this stage, the body's Shen fuses with the resonant light of the Upper Dantian

and is transformed and released into the Heavenly and Earthly fields of energetic space. This transition is similar to vapor diffusing into space.

4. Infinite Space (Wuji) Reunites with the Divine (Dao). At this final stage, the Shen is reunited with the divine.

The Shen directs the increased amount and flow of Qi during the transformation process of Jing into Qi and Qi into Shen through conscious intention. Through this soft focused intention an alchemical cycle of transformation is created whereby Jing, Qi, and Shen can then be gathered, refined, and trained.

TRANSFORMATIONAL ATTAINMENT MODALITIES

Each of these prenatal and postnatal transformations include training exercises and meditations that are either developed in low, medium, or high attainment approaches. The low attainment approach is known as the physical attainment approach; the medium attainment approach is known as the mental attainment approach; and the high attainment approach is known as the spiritual attainment approach. Qigong regulation encompasses all three attainment modalities in order to facilitate the ultimate in energetic transformation. The three attainment approaches for increasing to maximum capacity the body's energy are described as follows.

1. The Physical Attainment Approach concentrates on the development and cultivation of the body's energy by focusing on Dynamic (energetic movement) and Quiescent (stillness) Qigong meditations to fully realize one's energetic potential.
2. The Mental Attainment Approach concentrates on increasing mental power by focusing on symbolic drawings, reciting prayers, or mantras to manifest fully one's mental potential.
3. The Spiritual Attainment Approach concentrates on the development and cultivation of the body's Original Spirit (Yuan Shen) by focusing on Shengong meditations (meditations that focus on spiritual cultivation) to harvest one's spiritual potential.

JING: ESSENCE OF LIFE

Jing refers to the indispensable bioenergetic substance of all living things (i.e., the energetic matrix of the body's cells, tissues, etc.). In Medical Qigong therapy, Jing is considered the most Yin form of Qi. Jing, as a pure Essence of living matter, is more dense than Qi but more refined than Blood or Fluids. When Jing is cultivated and allowed to arise naturally within the body, the outer surface of the body's tissues becomes harmonized and glows. Deep within the tissues, the Jing pools like a deep spring. The greater the degree of harmony between the individual's physical, mental, energetic, emotional, and spiritual fields, the more Jing is transformed into Qi.

THE YIN AND YANG ASPECTS OF JING

Jing can be divided into the functional aspects of Yin and Yang properties. The Yin aspect of Jing provides the material basis for growth, development and reproduction, and is the substratum for the formation of the materials associated with the Marrow and Blood. The Yang aspect of Jing activates transformations, growth, development and reproduction, and is associated with the Yuan Qi, Blood and Body Fluids (Figure 13.4).

As a subtle energy, Jing has a tendency towards downward movement, flowing like water, it fills the Lower Dantian, adrenals, Kidneys, and urogenital organs. Essence originates from two sources and can be divided into Prenatal and Postnatal Jing.

PRENATAL JING

Prenatal Jing is also called Original Essence (Yuan Jing). It is inherited from the Jing of both parents and is considered the root of life. It is created upon conception through the blending of both parents' sexual energies and substances. Prenatal Jing constitutes the original substance of the body, and serves as the material base for growth, development, and reproduction. With the help of the Qi and Blood acquired from the mother, the Jing is responsible for the nourishment and development of the embryo's and fetus's brain, bones, muscles, tendons, skin, and hair. The Pre-

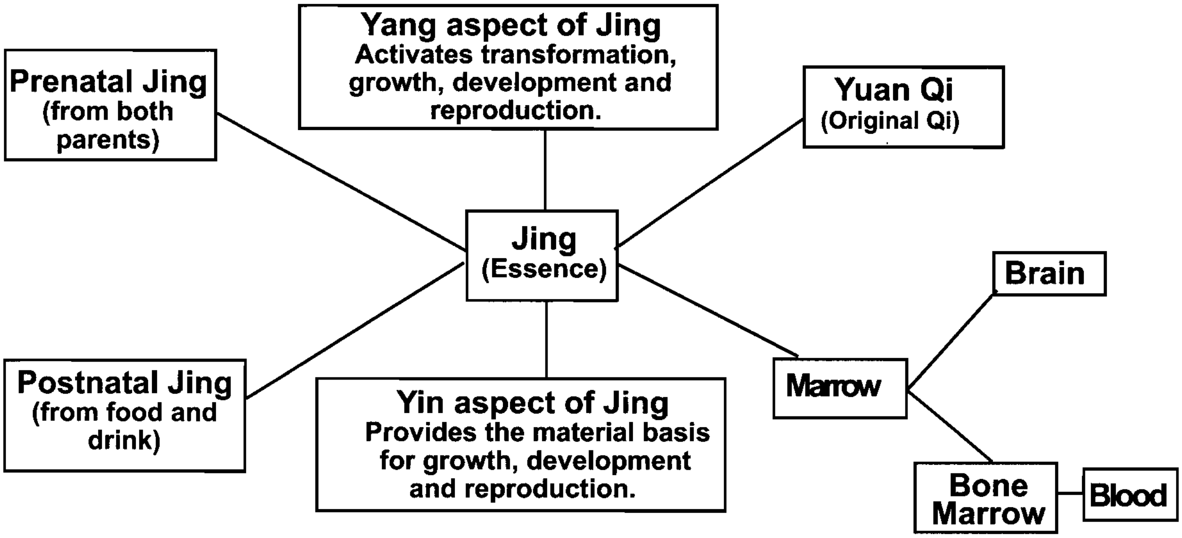
Yang aspect of Jing
Activates transformation, growth, development and reproduction.
Prenatal Jing
(from both parents)
Yuan Qi
(Original Qi)
Jing
(Essence)
Brain
Postnatal Jing
(from food and drink)
Yin aspect of Jing
Provides the material basis for growth, development and reproduction.
Marrow
Bone Marrow
Blood

Figure 13.4. The Formation of Jing

natal Jing determines each person's basic constitution, strength, and vitality.

POSTNATAL JING

Postnatal Jing is also called the Acquired Essence. It is the Essence of the body's main organs, derived from air and Gu Qi (produced from food and drink by the transformation process of the Spleen and Stomach) which is transported through the channels and vessels to nourish the body's Yin organs, and to irrigate the Yang organs. Postnatal Jing is the material basis for the functional activities of the body's internal organs and metabolism. It is transported (through the function of the Lungs) to all the Yin and Yang organs where it becomes reproductive Yin and Yang Jing. The Kidneys are responsible for storing any surplus Jing to be distributed when needed.

THE INTERACTION OF PRENATAL AND POSTNATAL JING

The Prenatal and Postnatal Jing are interdependent and contribute equally to the growth and development of each other. Postnatal Jing is continually being used by the body and replenished with food and drink. Only with the nourishment of the Postnatal (Acquired) Jing can the Prenatal (Congenital) Jing be enriched and function optimally. Without the function of the Prenatal Jing, the Postnatal Jing cannot be transformed into Qi.

Postnatal Jing is continually being used and replenished through the Spleen's production of Gu Qi from food and drink. This is why so much emphasis is placed on the first postnatal alchemical transition of changing Jing into Qi. Through Medical Qigong exercises, the body is able to stimulate, nourish, and preserve its Jing.

Both Prenatal and Postnatal Jing are related to the Kidneys, which generate bone growth as well as nourish the brain, memory development, and body stamina. The *Yellow Emperor's Classic of Internal Medicine* states that women are on a seven year developmental cycle, while men are on an eight year cycle. Problems with the body's Jing result in developmental problems (i.e. improper maturation, sexual dysfunction, infertility, and premature aging).

The Kidney Jing is responsible for extending the life-force energy up the Governing Vessel into the brain. The Kidney Jing also transforms into Marrow and is responsible for the production of bone marrow, which fills up the spinal cord and brain. When the Jing becomes exhausted, the memory begins to fail and the spirit becomes fatigued. It is important to remember that all types of Jing do not exist as separate substances, but support, interact with, and transform into one another.

The Kidney Jing also works in conjunction with the Qi to establish the body's state of mind and emotional health, which together determine the development of the Five Elemental Constitutions.

QI: VITAL LIFE-FORCE ENERGY

Qi is the energetic substance from which the entire universe was created. Through the interaction and transformation of Yin and Yang Qi, the various substances of form and matter are produced. This vital life-force energy comprises both material and functional aspects of the body.

Qi is the medium between matter (Jing) and Spirit (Shen). Chinese medicine describes many different kinds of Qi. Qi can assume an infinite number of forms according to its state of condensation or dispersion. When Qi condenses, matter is created, and when Qi disperses, energy is created. Although Qi in the body can assume different forms, that have different functions, it is essentially all one energy. The patient's Jing, for example, may be ready to receive a new energetic pattern for accelerated healing, and the Shen may be ready to send this required pattern. But, if the body's Qi is weak or turbulent, it will not transmit this information from the Shen to the patient's Jing, thus delaying or impeding the healing process.

The body's Qi can be classified into two primary forms of energy: Prenatal and Postnatal Qi.

PRENATAL QI

Prenatal Qi (or Xian Tian Zhi Qi, which translates as, "before the baby sees the sky energy") is also called Yuan Qi, or Original Qi. This is the energy that the baby inherits from his or her parents during conception. It is stored in the sexual glands and adrenal cortex, and it is essential for the growth and development of the fetus.

It is Jing in the form of energy, that is cultivated through Qigong practice. Yuan Qi is the innate or inborn energy that is the original force which maintains normal growth and development, and promotes the functional activities of the internal organs. Its role is to bring into existence the vital mechanisms of the human body according to their genetic pattern, conforming them to the models transmitted by each person's lineage.

Prenatal Qi contains elements of all of one's predecessors and gives rise to the Prenatal Jing, and is seen as a means of transmitting the Deoxyribonucleic Acid (DNA) code. Prenatal Qi relies on nourishment from Postnatal Jing. Through Prenatal Qi, each individual is energetically connected with everything in existence. Prenatal Qi also relies on Postnatal Qi for nourishment.

Methods of training Qi and regulating it include counting respirations as the Mind focuses on rooting the Qi into the Lower Dantian, Mingmen, or Kidneys' area.

There are three types of prenatal energy regulations used for circulating and harmonizing the body's internal energy or Dantian Qi:

1. The first type of Qi regulation focuses on moving energy through the body's main channels and torso. This is the most common type of Prenatal Qi circulation, directing the energy to flow along the course of the body's Conception and Governing Vessels, i.e., the "small" Heavenly cycles (Microcosmic Orbits). This type of Qigong exercise is commonly referred to as an Inner Cultivation and Regulation method.
2. The second type of Qi regulation focuses on moving energy through the body's channels and extremities. In this type of Qi circulation, the body's Prenatal Qi is directed to flow only along specific channels (i.e., being led from one channel out from the torso into an extremities and then back into the body through another channel). This type of energetic balancing is commonly seen in the "large" Heavenly cycles (Macrocosmic Orbits) and bone marrow regulation.
3. The third type of Qi regulation focuses on

moving the energy of Heaven (Universal Qi), Earth (Environment Qi) and Man (Qi of the physical body) through the channels. In this type of circulation, these three types of energy are integrated or merged into one energy and then regulated. Upon inhalation, the individual imagines Prenatal Qi circulating from Heaven through the body's channels into the Earth. Then upon exhalation, the energy is imagined flowing from the Earth, through the channels, back into Heaven.

POSTNATAL QI

Postnatal Qi (or Hou Tian Zhi Qi, which translates as, "after the baby sees the sky energy") is also called acquired or Post-Heaven Qi. This is the energy that the baby derives from air, food, and drink after it has been born. Postnatal Qi relies on Prenatal Qi for development.

Prenatal and Postnatal Qi form the foundation for the body's vital energy. In Traditional Chinese Medicine it is believed that the parents' health and their state of Jing, Qi, and Shen at the time of conception determines the general health of the child. If the parents are sick, or their health has is compromised by drinking, smoking, or drug use, the baby's health will suffer. If the parents are old, or have suffered an emotional shock during the pregnancy, this too will negatively impact the baby's health.

Methods of training Postnatal Qi include deep, tranquil abdominal respiration. This method trains the respiratory system to improve both the respiratory and digestive systems, enhancing both the Lung and Spleen Qi.

EIGHTEEN CLASSIFICATIONS OF QI

The distribution and function of life-force energy can be further divided and classified according to the source and function of the Qi. Below is a description of eighteen transitions of Qi and their energetic effects on the body.

1. Prenatal Qi (Yuan Qi), is often referred to as Original Qi, and includes the Original Yin and Original Yang energies of the body. Yuan Qi has its root in-between the Kidneys and Mingmen area, it is distributed throughout the entire body via the Triple Burners.
2. Postnatal Qi, is an external energy derived from the consumption of food and drink (from the Earth), and air (from the Heavens), and is a term used to describe the Qi which is cultivated after one is born.
3. Heavenly Qi (Tian Qi), is energy that has been developed from the energies of the sun, moon and stars. Tian Qi is absorbed into the Lung Qi as Air Qi (oxygen) through respiration.
4. Earthly Qi (Di Qi), is energy that has been developed from the energies of the soil, water and wind. Di Qi is absorbed into the Stomach and Spleen where it is converted into Gu Qi (food energy); Di Qi is also directly absorbed through the pores from the environment.
5. Food Qi (Gu Qi), is the first stage of energy transformation derived from food and drink. It is also called Qi of the Water and Grain in the digestive system. After food essence is cooked in the Stomach, the Spleen transforms it into Gu Qi. At the first stage of development Gu Qi is a coarse, unusable form of food essence. The Spleen divides the energy it derives from the food essence into Pure (Clear) Yang Qi and Impure (Turbid) Yin Qi. The Clear Yang essence of Gu Qi is transported by the Center Qi (Zhong Qi) via the Middle Burner, sending it upwards into the chest, first to the Lungs where it combines with Heavenly (Tian) Qi to form Gathering (Zong) Qi, then to the Heart where it combines with the Yuan Qi (from the Kidneys) to create the Blood. The Turbid Yin essence of Gu Qi is transported downwards through the gastrointestinal tract to be expelled as waste.
6. Clear Yang Qi (Qing Qi), is the clean, pure energetic essence of Gu Qi, transported by the Spleen, via the Middle Burner, up towards the Upper Burners, and chest area. While in the chest, it is combined with Gathering (Zong) and Yuan Qi to form True Qi.
7. Turbid Yin Qi (Zhou Qi), is the clouded, impure energetic essence of Gu Qi, transported by the Spleen, via the Middle Burner, down

towards the Lower Burner, Small Intestine, and Bladder organs, to be further refined. It is a more coarse, unrefined, polluted energy.

8. Center Qi (Zhong Qi), is the energy generated from the Middle Burner, derived from the Spleen and Stomach. Its primary function is to transport the Gu Qi into the chest, to be blended with the Heart and Lungs' Qi.
9. Gathering Qi (Zong Qi), is also called Respiratory Qi. Gathering Qi is considered the "essence of life," manifested through the conversion of the purest and most potent forms of the body's Jing (particularly sexual fluids, hormones, and neurochemicals). It is essential for the formation of the fetus and is the initial motivating power of life. It promotes vitality and increases stamina; it also enhances the immune system, promotes mental clarity and prolongs life. Gathering Qi combines the Heavenly Qi (inhaled by the Lungs) and the Gu Qi (derived by the Stomach and Spleen), and accumulates these energies within the chest in order to assist the Heart in circulation the Blood, and the Lungs in respiration. It controls the speech and strength of the voice, as well as the Blood circulation to the extremities. Gathering Qi gathers in the chest and is controlled by the Middle Dantian.
10. True Qi (Zhen Qi), is also called Anti-pathogenic Qi. It originates in the Lungs and is the last stage of refinement and transformation of Qi. Gathering Qi and Yuan Qi combine their substances to form the True Qi in the chest area. True Qi fills and nourishes the body as it is further broken down into Nourishing (Ying) and Protective (Wei) Qi. True Qi is the energy that circulates inside and outside of the body, in the channels and collaterals, nourishing the Yin and Yang organs. It is called True Qi because it is the most refined and pure, and is in harmony with the environment. It is opposed by Evil Qi (Xie Qi), also called Toxic Qi, which is disharmonious and thus untrue to the body's essential nature.
11. Nourishing Qi (Ying Qi), is another form assumed by the True Qi and is the Yin aspect of the Channel Qi. Ying Qi has two major functions: to nourish the Blood, organs and tissues, and to link the mind and intention to the body. It flows within the blood vessels, as well as the channels, and is the fuel of the body's metabolism. Its potency depends upon the quality of food and water consumed, the purity of air breathed, and the efficiency of the body's digestive, respiratory, and circulatory systems. Since it is Yin, it is considered the more refined form of True Qi, as it flows in the body's inner layers and internal organs. Being directed by thoughts, it is thereby closely related to the emotions.
12. Protective Qi (Wei Qi), is a second form of True Qi and is the Yang aspect of the channel Qi. It protects the Blood, organs, and tissues from invasion of external pathogenic factors such as: harsh weather conditions, microorganisms, and harmful emotional influences. It also protects the individual from evil spiritual forces. It is the coarser Yang aspect which flows through the outer layers of the body's skin. Its texture is slippery in nature and cannot enter into the channels. It therefore circulates under the skin and in between the muscles. The Wei Qi vaporizes in between membranes, and diffuses over the chest and abdominal area to form a kind of energetic armor. It also warms, moistens, and contributes to the nourishment of the skin and muscles. The Wei Qi also helps regulate the perspiration by opening and closing the pores.

Wei Qi has its root in the Lower Burner (Kidneys), is nourished in the Middle Burner (Stomach / Spleen), and spreads from the Upper Burner (Lungs) to the outside of the body. Wei Qi is said to be located at the superficial level of the body during the day, in the muscles, skin, hair, etc. At night it sinks to a deeper level to circulate through the viscera. If the Wei Qi is obstructed in some manner from moving inward, insomnia may result. Also, Wei Qi becomes thick and extends outside its normal realm during Medical Qigong practice and may, therefore, take longer to move inward at night, causing some practitioners to experience difficulty falling asleep after evening Qigong practice.

13. Channel Qi (Jing-Luo Qi), is also called Meridian Qi. This energy flows throughout the body's deep internal and superficial channels, as well as through the collaterals and blood vessels like rivers and streams.
14. Organ Qi (Zang and Fu Qi), is the energy of the physiological activity of the body's internal organs and manifests as a major aspect of its physiological function. Individual pools of Qi are formed within the pairs of Yang-Fu (hollow) bowels and Yin-Zang (Solid) viscera. The Yang organs are responsible for transforming food and drink to produce Qi and Blood. The Yin organs are responsible for storing vital substances. Each organ can be additionally divided into pairs of Yin and Yang energetic aspects.

    Each of the body's organs has its own energy, which is governed by, and corresponds to, one of the Five Elemental energies. These Five Elemental energies respond to the external energetic fields drawn from the universal and environmental energetic fields. The internal processes of thinking and feeling, as well as the physiological process such as metabolism and hormonal production influence the Organ Qi.
15. Evil Qi (Xie Qi), is also called Pathogenic Qi, and Toxic Qi. This usually refers to any harmful external pathogenic factor (especially the Six Climatic factors). However, it also refers to the evil spirits and stuffed toxic emotions, which affect the individual's physical, energetic, and emotional balance.
16. Righteous Qi (Zheng Qi), is the defensive aspect of the True Qi, which protects the body from Evil Qi. This is not actually another type of energy, but a term used to indicate the True Qi's function of protecting the body from invasion by external pathogenic factors. The occurrence and development of disease is directly related to either the hyperactivity or hypoactivity of the Righteous Qi.
17. Spiritual Qi (Ling Qi), is the supernatural energy associated with the most highly refined energy in the human body manifesting through the Eternal Soul. This type of energy enhances spiritual awareness and constitutes the foundation for advanced levels of spiritual conditioning, and the attainment of an enhanced energetic body, and enlightened states of mind and spirit.
18. Divine Qi (Shangdi Qi), is also called Holy Energy or God's Healing Light. It is the supernatural energy associated with the Divine (Dao). The divine energy is the true source of spiritual awareness and constitutes the foundation for advanced levels of spiritual healing, as well as attaining enlightenment.

THE SIX FUNCTIONS OF QI

A Qigong doctor will primarily focus on a patient's combination of Original Qi (Yuan), Gathering Qi (Zong), Nourishing Qi (Ying), and Protective Qi (Wei) when treating or prescribing homework. Initially, Qi has six main functions in regulating the body. Although these functions may seem to overlap, their individual responsibilities are distinct.

1. Qi transforms substances into energy (one example of its transforming nature is changing food into energy).
2. Qi transports the substances it creates.
3. Qi holds and contains energy, Blood and organs, etc., within the energetic and visceral structures.
4. Qi protects against both external and internal pathogenic factors.
5. Qi raises and elevates things (for example, Blood, associated Fluids, and the body's temperature).
6. Qi keeps the body warm (for example, strong Qi will warm the body, while weak Qi makes the body cold).

The normal physiological activity of Qi in the body is a constant harmonious movement, similar to the ebb and flow of a tide. When the body's Qi is in constant motion, it moves in eight primary Yin and Yang directions: ascending and descending, gathering and dispersing, expanding

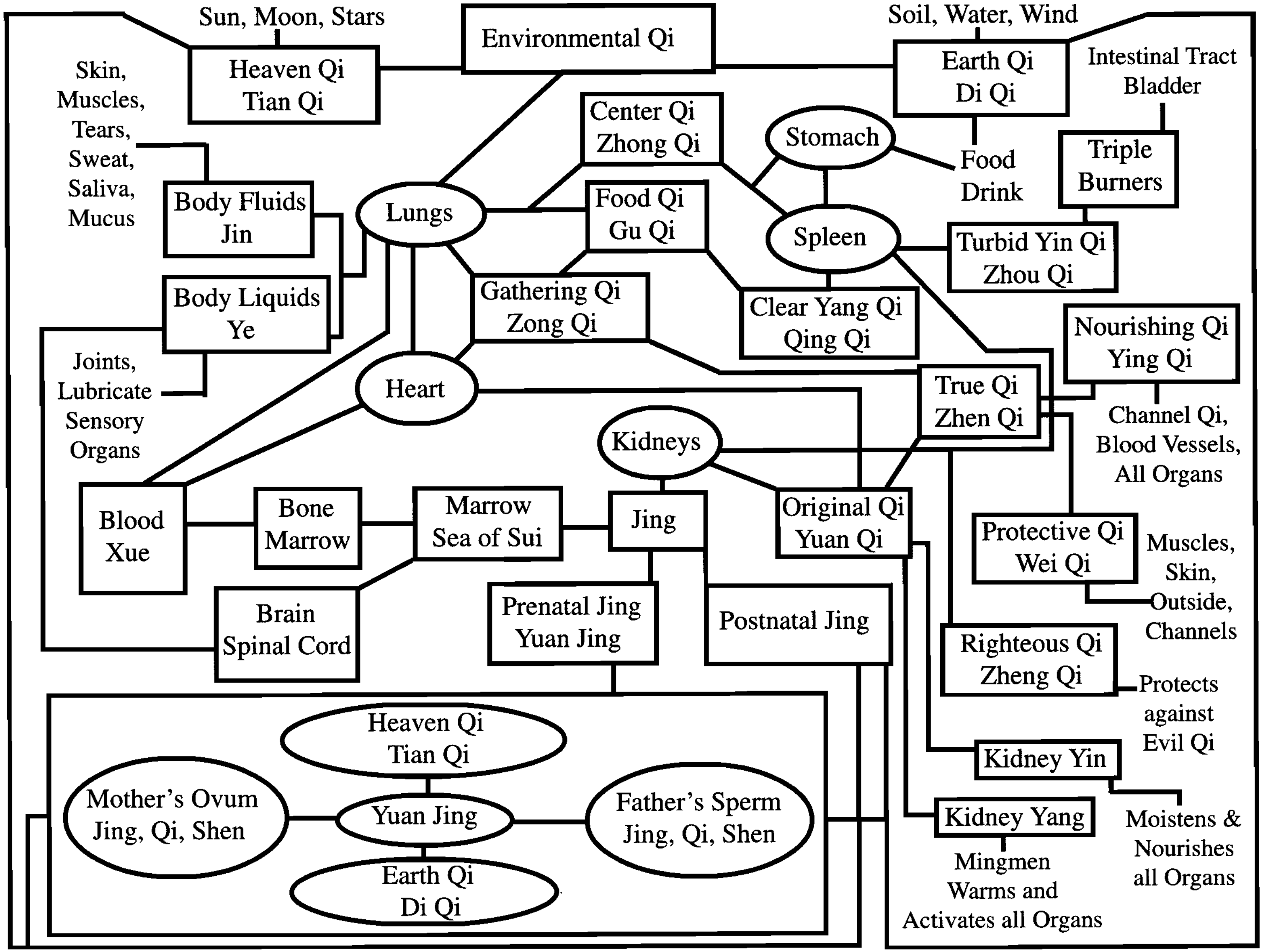
Sun, Moon, Stars
Environmental Qi
Soil, Water, Wind
Heaven Qi
Tian Qi
Earth Qi
Di Qi
Intestinal Tract
Bladder
Skin,
Muscles,
Tears,
Sweat,
Saliva,
Mucus
Center Qi
Zhong Qi
Stomach
Food
Drink
Triple
Burners
Body Fluids
Jin
Lungs
Food Qi
Gu Qi
Spleen
Turbid Yin Qi
Zhou Qi
Body Liquids
Ye
Gathering Qi
Zong Qi
Clear Yang Qi
Qing Qi
Nourishing Qi
Ying Qi
Joints,
Lubricate
Sensory
Organs
Heart
True Qi
Zhen Qi
Channel Qi,
Blood Vessels,
All Organs
Kidneys
Blood
Xue
Bone
Marrow
Marrow
Sea of Sui
Jing
Original Qi
Yuan Qi
Protective Qi
Wei Qi
Muscles,
Skin,
Outside,
Channels
Brain
Spinal Cord
Prenatal Jing
Yuan Jing
Postnatal Jing
Righteous Qi
Zheng Qi
Protects
against
Evil Qi
Heaven Qi
Tian Qi
Kidney Yin
Mother's Ovum
Jing, Qi, Shen
Yuan Jing
Father's Sperm
Jing, Qi, Shen
Kidney Yang
Moistens &
Nourishes
all Organs
Earth Qi
Di Qi
Mingmen
Warms and
Activates all Organs

Figure 13.5. The Body's Intricate Interactions of Qi, Blending and Flowing through Internal Reservoirs of Pooled Energy.

and contracting, and entering and leaving. If however, there develops a Qi obstruction, or a Deficiency of Qi, the Qi will deviate from its natural course, and may cause Rebellious Qi, disharmony and disease.

Qigong doctors carefully examine the energetic ebb and flow of Qi to comprehend the complexity of the energetic hologram in which we exist as a living energy. As the human body absorbs, stores, utilizes, and releases energy, its natural function of transforming energy promotes health and allows us to live in harmony with the natural environment. However, once an internal organ or organ system develops a Rebellious, Excess, Deficient, or stagnant condition within the body's energetic flow, disease develops. By understanding the energetic pattern and function of each energetic organ and organ system, the Qigong doctor can diagnose the disease by retracing the energetic dysfunction to the root cause. The following graph denotes the function and flow of the body's life-force energy patterns (Figure 13.5).

QI AND ENERGETIC MEDICINE

The utilization of the Three Regulations (posture, respiration and mental Dao Yin concentrations) enable Qigong doctors not only to disperse any Toxic Qi acquired from patients, but also to regulate their own channels, through gathering energy, and increasing their own energetic fields. Through the proper use of the Mind (whole body consciousness and awareness), the doctor can simultaneously influence the energy flows within the patient's body, as well as his or her own Wei Qi fields.

As the doctor begins to focus on specific goals (i.e., dispersing Toxic Qi from a particular organ, tissue area, or channel), the patient may experience a surge and release of strong emotions and memories that had been stored in that area. The patient's energy field then shifts and moves throughout the body according to the specific emotion(s) elicited and released. This energy field is closely related to the thinking activities of the brain. As the Mind begins to release certain specific memories, the field of energy shifts and moves through the body in accordance with the specific emotions that are activated and released. During treatment, the shape, flow, rhythm, nature, and thickness of this energetic field can be perceived by the doctor who has become highly sensitive to perception of energy through "inner-vision" (see Chapter 25).

By training in the Three Dao Yin Regulations (see Chapter 14), internal energy can be guided to surge not only in the natural direction of flow within the channels, but also against it. Internal energy can also be guided along several channels at the same time, in any given direction or to specific areas of the body. By directing Qi against the flow of a channel, the doctor arrives at the source of the river (or energy). This technique is sometimes used by the Qigong doctor to collect or tonify the Qi within the patient's internal organs. This technique for storing and collecting energy is called "following the river to reach the sea." An example of this technique would be extending Qi from the patient's St-36 points, up the Stomach Channels to root, strengthen and stabilize a Deficient or Rebellious condition within the Stomach organ itself. Conversely, by directing Qi with the flow of the stream, one can disperse energy (purgation) and pull Qi out of an area leading it away from the source (e.g., pulling Liver-Stomach Heat down and out the Stomach Channel to cool an overheated Stomach).

QI AND THE BODY'S ENERGETIC STRUCTURE

Similar to the biomechanics of the respiratory, circulatory, and digestive systems, the body's energetic system has its own organization and established rules of action for regulating organ and tissue functions. The adaptability of Qi to the outside environment enables the body to absorb not only healthy energy, but also to expel Toxic Qi.

The nature of Qi is to endlessly gather and disperse, expand and contract, to flow inward and outward, to rise and fall, as it circulates. When a Qigong doctor begins to treat a patient's channels or points, he or she stimulates and activates the patient's Qi, which may arouse muscle movement and sensations which can last for quite some time.

If the frequency and nature of the doctor's Qi is similar to that of the patient's, both energetic resonance and synchronization will synergize, allowing the patient's Qi activities to be regulated and balanced at a rapid rate.

Because people vary in sensitivity to energy, some patients can be treated with Medical Qigong therapy immediately. Sensitive patients are usually taught Medical Qigong exercises allowing them to ingest energy from trees, plants, and other things of the natural world. Patients with low sensitivity may have difficulty in perceiving energy and often dissociate from their body during treatment. These patients should be given Shengong meditations to restore conscious awareness of the toxic emotions trapped within the body, thereby reestablishing a harmonious relationship with the Mind (see Chapter 32). The patient's insensitivity to the doctor's energy extension is an indication that the patient's spirit has gone out of the body in order to avoid feeling the emotions which arise from Medical Qigong therapy.

A patient insensitive to energy projection can still be in touch with basic surface feelings (i.e., hot, cold, tired, etc.); however, the deeper emotions are hidden, or energetically walled off through denial. The better the relationship between the patient and the doctor, the greater the patient's trust and surrender to the doctor's energy.

When observing the patient's Qi and energetic structure, the doctor is able to perceive and enter the patient's five energetic fields. These fields form the matrix of the patient's energetic structure, which constitute a blueprint for the body's physical structure. These energy fields have their own biological components and degree of fluidity.

These five energetic fields form the body's Sea of Energy. The term Sea of Energy (Qi), in this context, refers to the description used to explain the internal and external currents of life-force energy which surround, permeate, and function within the physical body. These energetic fields are described as follows.

1. The External Fields of Wei Qi (Protective Energy),
2. The Internal Current of Ying Qi (Nourishing Energy),
3. The Internal Sea of Blood (Xue Hai),
4. The Internal Sea of Marrow (Sui Hai), and
5. The Center Core of Light (Taiji Pole).

Within these five energetic fields, the protective and nourishing energetic fields of Wei and Ying Qi expand and contract, move and adjust their energetic flow according to the body's respiration patterns. The Sea of Blood resonates within the pulse rhythm established by the heartbeat. The Sea of Marrow undulates from the rhythm established within the body's micropulsation of the sacrum and cranium bones' flexion and extension. The energetic pattern of the Taiji Pole vibrates at the same frequency as divine light (the Dao).

THE THREE EXTERNAL FIELDS OF PROTECTIVE ENERGY (WEI QI)

There are three external energy fields that create the body's Wei Qi field. The Wei Qi field's energy flows from around the regions of the body's external muscles, skin surface, channels, collaterals and channel points, and extends several feet out from the body's tissues. All energetic forms of the physical body, including the internal organs, blood vessels, nervous systems, etc., can be accessed and treated through the three extended fields of Wei Qi.

These three fields are distinguishable from each other by gradations of density, vibrational rate, light, and temperature. All three energy fields communicate with each other, interacting so as to form the body's subtle energetic boundaries.

Enveloped within these energetic boundaries are the body's external "energetic grids," spanning the body's energetic structure like a giant web. These energetic grids are the body's external energetic field patterns, etched through the three fields of Wei Qi like a three dimensional webbed hologram (similar to a three dimensional blueprint).

It is important to note, that any disease within the body's internal structure is also imprinted in this external energetic grid. Disharmony of the body's tissues often appears first in the Wei Qi fields, before manifesting within the body. If the energetic grid is not transformed (through emitted light, sound, and resonant vibration), the disease cannot alter its destructive course, because it follows the patterns dictated within the grid. Thus, for example, a tumor surgically removed often quickly reappears. Trauma, be it physical, mental, emotional, or spiritual, as well as drug and alcohol abuse, distorts, damages, and destroys the delicate energetic web which constitutes the body's external energetic grids, making the patient more susceptible to external environmental influences.

The external energetic grids function to protect the body from external pathogenic factors, as well as to receive and transmit environmental energy. Thus, for example, a person can receive, absorb, and digest other people's emotions. The five senses are constantly receiving on outside environmental influences, which directly and indirectly affect the body's major organs. This constant influx of data, changes the body's emotional patterns, and forces individuals to adjust their emotional perspective frequently.

Researchers conducting tests in San Antonio, Texas, for example, determined that when subjects are being stared at, galvanic skin responses are considerably increased. These increases occurred even though the subjects were not consciously aware of being observed. The research determined that the body has its own expansive field of subtle awareness, which reacts even when the individual is not consciously aware of the body's energetic extension.

By reconnecting the spirit and emotions with the mind and body, the Qigong doctor is able to perceive the subtle Messages that the patient is unconsciously conveying (see Chapter 25).

THE THREE ENERGETIC BARRIERS

The Wei Qi fields are experienced at three different levels. These levels are differentiated by energetic barriers, which protect the body's boundary system. These barriers not only protect and contain the body, they help establish the individual's sense of reality, and sense of self. The third energetic field or "outer spiritual barrier" is depicted in the illustration as the outside speckled field of energy, and affects the second energetic field or "middle emotional/mental barrier." The second field of Wei Qi is depicted in the illustration as the middle white field of energy, which in turn influences the first energetic field or "physical barrier," seen as the black field of energy in the illustration (Figure 13.6).

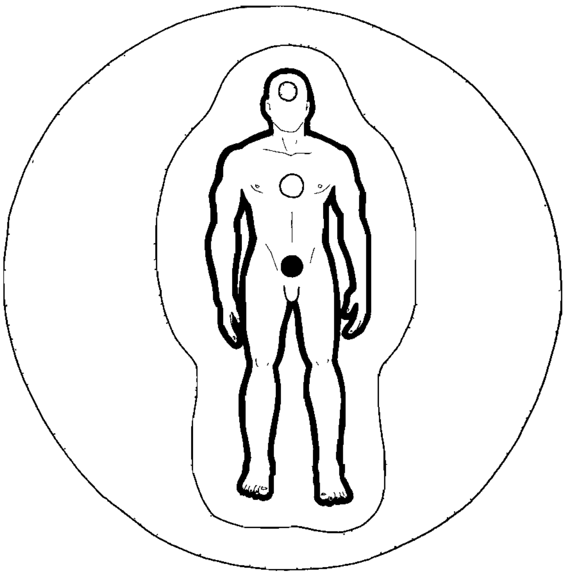

Figure 13.6. The Three External Fields of Wei Qi

These three Wei Qi barriers are always interesting for the doctor to feel, because they change so much from patient to patient, and sometimes from moment to moment. Some patients have very strong energetic fields, they are very palpable and full. Emotions strongly affect this type of energy field. The Wei Qi field of a patient who is armoring out of fear, reticence, or self-protectiveness, can actually repel the doctor's hand back, if their field and intention is strong enough. This repelling reaction is helpful because it lets the doctor know how to pace the treatment, and encourage the patient to let go of the armoring. This allows the patient to consciously focus on relaxing and sensing the areas in the body that he or she was trying to protect.

The energetic boundaries, act as gateways to the next level of energetic field energy. To perceive the patient's three fields of Wei Qi, the patient's energetic boundaries must be gently palpated. These boundaries are more easily sensed due to a thickening of energy which serves as a barrier to external pathogenic factors. Sometimes the doctor may have to move back into the patient's third or "outer" energetic field, and telepathically let the patient know that he or she will not be violated, and is safe. The doctor waits until the patient exhales (thereby expelling some tension) before entering and palpating the outer Wei Qi field. Once a release in the energetic field is felt, the doctor moves forward into the next energetic field. Because this same process occurs at all three levels, the Qigong doctor responds accordingly, waiting for the exhalation that allows them to proceed deeper into the next level of the patient's Wei Qi.

THE FIRST EXTERNAL FIELD OF WEI QI

The first external field of energy is confined to about one to two inches outside the physical body, and composed of lower etheric matter. It is related to the Lower Dantian and manifests as a holographic energy template for the physical body. It provides the spatial guidance for morphological development during the creation and formational stages of the embryo, as well as the morphological energetic patterns that assist in the repair of damaged tissues. Its energetic boundary is responsible for armoring the body against invasion of external pathogenic factors. This physical energy body is connected to the somatic tissues and cells, thus any physical-energetic interaction stimulates the body's external channels and points. This first layer is very dense and most Qigong doctors get the impression of physically contacting the patient several inches before actually touching the patient's material body. Its resonant vibrational pattern is very broad and slower than the other Wei Qi fields.

THE SECOND EXTERNAL FIELD OF WEI QI

This energy flows through the regions of the internal organs and tissues. It extends about a foot or more outside the body's tissues, and is composed of astral matter. It is related to the Middle Dantian and the emotional energy body. It contains a form of disembodied consciousness that can astral travel, and at times leave the physical body, as in near death experiences. This second layer of Wei Qi interfaces with the limbic system of the brain (the neurobiological seat of emotions) and is also linked with the sympathetic and autonomic nervous system. The second Wei Qi field's energetic boundary is responsible for protecting the body from negative emotions, destructive feelings and criticism from others. It is automatically activated when the patient is placed in a position where a lack of trust exists, or there is unwanted physical touch. The emotional energy body is connected to the visceral organs. Its main function is to receive, interpret, and verify an individual's emotions, feelings, desires, impulses, and thought patterns. It is through this energetic field that the patient's aura is typically observed as colors and light patterns. The second layer is more subtle then the first layer. When physically palpating, the sensations are similar to a magnetic attraction and repulsion. Its resonant vibrational pattern is faster than that of the first layer.

THE THIRD EXTERNAL FIELD OF WEI QI

This field extends a few feet to several hundred yards depending upon the spiritual evolution of the individual, and is composed of ketheric matter or substance. It is related to the Upper Dantian and the spiritual energy body. It is associated with intuition, inspiration, creativity, and visionary insights. Its energetic boundary is responsible for establishing security by informing the body of impending encounters, conflicts, or environmental transitions. This spiritual energy body is connected to the body's Taiji Pole, and is a container for karmic-related illnesses within the physical body. The Taiji Pole is known as the "root with many branches," as it is a repository of the soul's ancestral memories and experiences. Its function is to receive the subtle energy of the finest and fastest vibrations. It senses and interprets the data received from its environmental awareness and universal connection to the divine. The third layer is most subtle in physical sensations and requires delicate, intuitive contact. Usually the first impression when palpating this energetic field, is one of contact without a material component (like wind). It is sometimes cold, sometimes warm but always "breezy." Its resonant vibrational patterns are the highest and fastest of all the Wei Qi layers.

INTERNAL CURRENT OF NOURISHING ENERGY (YING QI)

The Ying Qi level of internal energy flows within the underlying tissues of the body and sweeps through the torso like a fast wind, expanding and falling with each respiration (Figure 13.7). The Lungs act as a great energy pump; each breath is like a gust of air flowing in and out, in a cycle of expansion and contraction through the entire body. Ying Qi swirls like a breeze; its function is to nourish the Blood, organs and, tissues. It flows within the blood vessels, as well as the channels and links the mind and intention to the body. It is responsible for the organization, construction, completion, and reconstruction of the vitality of the individual. It is integrated with the breath, and extends outward to the body's physical boundaries (skin and body hair).

The internal current of Ying Qi is most palpable when the doctor is gently touching the patient's surface tissue with his or her finger tips, and it feels almost like electromagnetic energy. Feeling the wind of Ying Qi can be likened to falling into "space" after passing through the denser outer field of Wei Qi. The action of Ying Qi, in most cases, is free flowing and connected with the breath. When listening to certain patients, the Qigong doctor may feel the Ying Qi flowing through the patient's channels, wrapping itself around the whole body like a vapor. It feels light, thin, and fairly cool compared to the next River of Blood level.

To access the "current of Ying Qi," the Qigong doctors begin by suspending their hands above the patient's first level of Wei Qi. They imagine dipping their hands into water as they allow their hands to slowly make contact with the patient's body. Once contact is made, the doctors extend their intention under the patient's skin. The doctors should match the patient's respiration and stay connected to the body until they feel the patient relax. Matching their respiration almost always slows the patient's respiratory rate, moving the patient into an altered state of consciousness. Sometimes at this point, age regression can be sensed, as the patient's body memory is triggered.

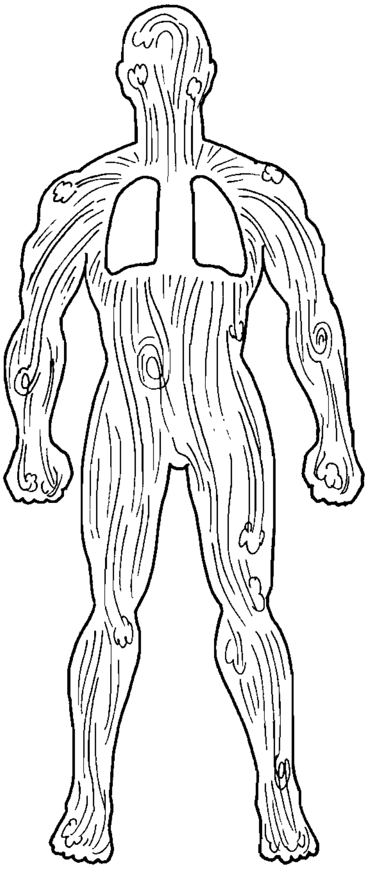

Figure 13.7. The internal current of Ying Qi flows from the Lungs and center of the chest through the entire body like a mighty wind.

SEA OF BLOOD (XUE HAI)

The Sea of Blood and its pathways constitute the body's internal network for circulating and maintaining the individual's life-force, under the direction of the Wu Jing Shen (the five separate fields that encompass the psycho-spiritual energetic transformations of the Five Yin Organs). The

Sea of Blood's energy flows through the blood arteries, vessels, capillaries, and the tissues of the major organs (Figure 13.8). By extending further into the body, the Qigong doctor perceives the patient's Sea of Blood; it feels like a warm watery pool in the center of the Heart, and flows through the entire body. The transition from Ying Qi to Xue Qi (Blood Qi) is easy to detect.

The energetic sensation of Blood is more substantial than the Wei or Ying Qi. The Blood being a liquid, is a denser form of Qi, and flows through the vessels, organs, and channels propelled by the Ying Qi. It is thick (this thickness varies from person to person), warm and somewhat "slick." The Heart's energy itself feels warm, substantial, and very alive.

As doctors gently touches the patient's surface tissue with his or her fingertips, the doctor extends intention deep into the patient's Heart. Upon contacting the Xue Qi, the doctor immediately becomes connected to the heartbeat of the patient. In some cases, the pulsing will connect and merge with the doctor's own heartbeat. Once the doctors encapsulates the patient's Heart, the strong emotions emanating from patients will usually subside. These feelings may be replaced by sporadic spurts of emotions such as relief, peace, or comfort, which sometimes invite the doctor further into the patient. On the other hand, some patients may react with feelings of fear, intense emotional pain, or even panic causing the doctor's hand to "bounce" away.

If the doctor's energy is "bounced back," he or she should immediately disperse the energetic charge into the Earth for ground purposes. The doctor should then either move to a more superficial energy level within the patient's body, or extend Qi into the patient's thorax or extremities. This technique allows the patient to relax after experiencing a Heart energy discharge. Since the pulse is simply another form of energy, circulating, expanding, and contracting with each heartbeat, the doctor should be able to ride the pulse out into the patient's extremities and viscera by listening to the Heart. This is similar to the listening ability of a master acupuncturist who, when reading a patient's pulse, extends his or her energy and intention through the patient's channels and into the organ of the Heart to diagnose the strengths and weakness of its congenital formation and current condition.

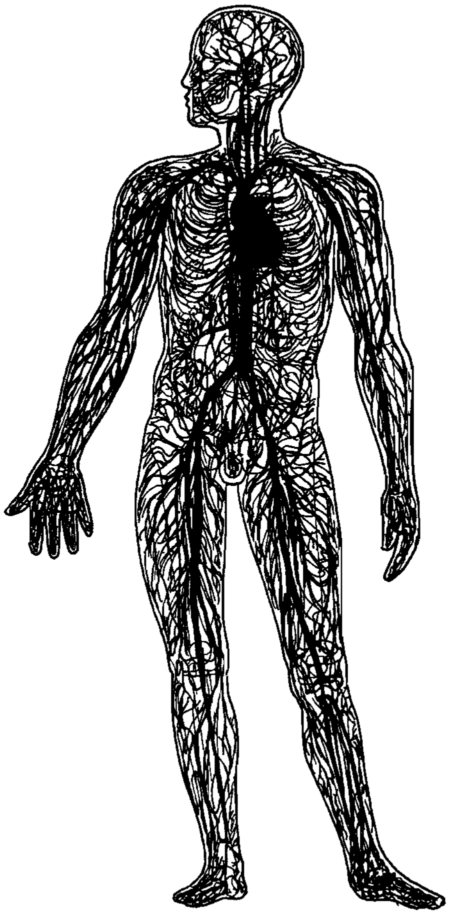

Figure 13.8. The Sea of Blood flows from the Heart through the body like a dense form of Qi.

SEA OF MARROW (SUI HAI)

This energy flows through the central core of the spine into the brain and enters the center core of the bones. In Traditional Chinese Medicine, the Marrow is not the same thing as bone marrow, but the stage prior to becoming bone marrow. Marrow is the pre-substantial Jing that forms the matrix for bone marrow, the spinal cord, and the brain (Figure 13.9). This flow of Sea of Sui energy is much slower than that of the River of Blood. It flows from the Kidneys' Jing, into the center of the spine, brain, and bone marrow. It has sometimes been called the "cranio-sacral rhythm" in Western medical terminology, and can be accessed from anywhere in the body. The Marrow energy generally feels cool and slick or slippery, with a texture like soft silicone gel. Once contact is made with the Sea Of Sui, it is very easy to utilize this pathway and travel throughout the patient's body. Its rhythm is slow yet manifests as a strong energetic pulse.

As the doctor gently touches the patient's surface tissue by dipping in with fingertips, into the Sea of Marrow, he or she can project intention into the center of the patient's spinal column. This connects the doctor with the rhythmic pulse of the Sea of Marrow. It is from this pulse that the doctor can determine the strength of the patient's Original Essence (Yuan Jing), Energy (Qi), and Spirit (Shen).

THE CENTER CORE OF LIGHT (THE TAIJI POLE)

This energy flows through the very center core of the body, uniting the three reservoirs of life-force energy (the Upper, Middle, and Lower Dantians). The fifth and final level of energy feels like a still, quiet place deep inside of the center core of the body's Taiji Pole structure. In China, Qigong doctors describe it as a holy place that almost defies description. It is observed as a vertical column or pole, of brilliant white light, full of pulsating vibrations. This energetic pole is surrounded by a vale of gold spiralling light which travels from the perineum to the top of the head.

When both the body's internal and external energetic wave patterns become synchronized through prayer or meditation, a resonate pulsation occurs within the Taiji Pole. This pulse begins deep within the center core of the body and resonates outward towards space. The Qigong doctors may also feel a powerful stream of energy leaving the body from the top of the head, connecting the Taiji Pole with the energy of the divine. The center core is a very reverent and special place, as it contains the essence of the patient's Eternal Soul, devoid of all ego, masks, or defense mechanisms (the patient's True Self).

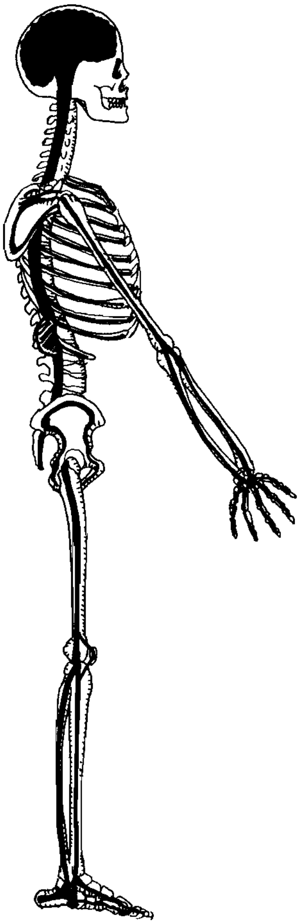

Figure 13.9. The Sea of Marrow flows from the kidneys through the spinal cord, brain, and bone marrow.

If the doctor penetrates the outside veil surrounding the patient's Taiji Pole, and extends his or her intention into the light's center, the doctor experiences the sensation of falling into space, or stretching out into eternity. Sometimes the sensation of shooting through space is followed by flashing colors and moving shapes, as the surrounding core dissolves into the Void (Wuji) itself. This same sensation and experience can also be personally accessed through deep prayer and meditation, and is quite a normal phenomena within deep spiritual practices (Figure 13.10).

SHEN: THE SPIRIT OF LIFE

The Chinese character ideogram for Shen is formed to symbolize "that which drops from the sky," and "that which passes through the body." In Medical Qigong therapy, Shen is considered the subtle energy of the spirit, associated with the Heart. It is associated with the spiritual energy and has a radiant nature.

When energized, Shen flows upward like fire, extending through the body's Eight Extraordinary Vessels creating a healthy nervous system. When Shen is gathered in the Upper Dantian, it can be projected as light and energy through the eyes, and the Yin Tang (Third Eye) point. According to Traditional Chinese Medicine, all illness has spiritual roots. The Shen (Spirit) is regarded as an integral part of the diagnosis and treatment process. An old Chinese saying states, "When the Shen is abundant there will be cheerfulness, when it is insufficient there will only be dullness and sorrow."

Shen is developed from Jing and Qi. According to Medical Qigong theory, when the patient's Shen is strong the individual is able to lead energy to an injured area and speed up the healing process. This strong Shen energizes the Qi, enabling it to increase cellular production. When Shen is scattered, the functions of the human body are weak from fatigue, the Shen becomes unsettled, and the mind restless. When the Shen and Qi combine together they form a solid connection of light and vibrational energy. This union is utilized to infuse energy and transfer thought through intention. This can be achieved by either conscious or unconscious intent by the doctor.

SHENGONG SCHOOLS OF MEDITATIONS AND TRAINING

The Shengong (Spirit Skill) meditations emphasize the transformations of consciousness, as well as the development of perceptual insight and the birth of wisdom. From an ancient Chinese energetic perspective, the three primordial qualities of true wisdom, or divine universal consciousness are:

1. The experience of emptiness within the Wuji (the Void), corresponds to the manifestation of the primal virtue of Wisdom.
2. The light manifested from the Shen's luminosity, corresponds to the primal virtue of Love.
3. The life-force energy inherent in all things, corresponds to the primal virtue of Power.

In China, there are three major schools of Shengong training. These spiritual schools of Qigong are described as follows.

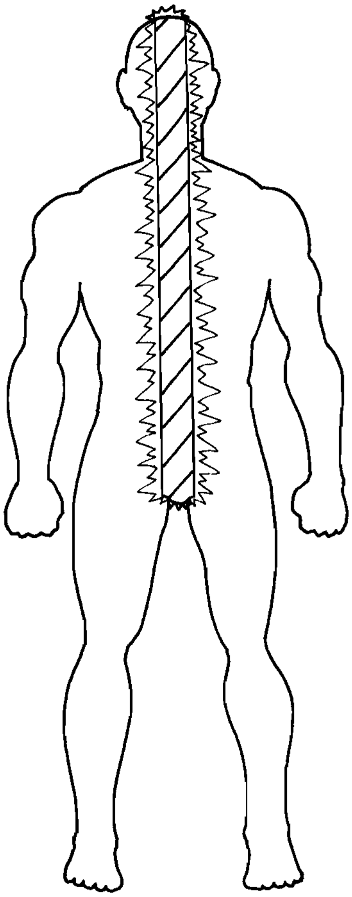

Figure 13.10. The body's center core Taiji Pole spreads from the base of the perineum to the top of the head.

- The Daoist School focuses on strengthening the body and mind equally with an emphasis on contemplating nature. Religious Daoism has its roots in Siberian Shamanism.
- The Buddhist School focuses on tempering the mind while giving little thought to the body.
- The Confucianist School focuses on regulating the mind to reach a state of quiet rest and peace through sincerity and the cultivation of moral character.

PRENATAL AND POSTNATAL SHEN

The body's Shen is the Supreme Yang aspect of the psycho-emotional components, the Heavenly Emperor or "master controller." It can be divided and classified into two primary forms through which the Spirit manifests: the Prenatal Spirit (Yuan Shen) and the Postnatal Spirit (Zhi Shen).

1. Yuan Shen is the body's spiritual element derived from Prenatal Jing (Yuan Jing) and Prenatal Qi (Yuan Qi). This aspect of the body's Shen is considered to be the intuitive Mind of the Dao. It is unborn and undying. It shines brightly and clearly. The Yuan Shen transcends conceptual thinking (which is the Postnatal aspect of the mind). It is immortal, unconditioned, primordial, and constitutes the mind's Yang nature.

   The Yuan Shen is produced and developed by the blending of both the male and female elements at conception (Body Fluids, Jing, Qi and Shen), and develops further with the nourishment from food and water. The Yuan Shen dominates all of the body's vital activities (heartbeat, digestion, motor function, etc.) and is also responsible for all intuitive and perceptive insight. The Prenatal Shen constitutes the true spiritual aspect of the Mind.

   By returning to a state of inner quiet and peace, through prayer, meditation, and adequate sleep, the Yuan Shen and the body's health are restored and maintained.
2. Postnatal Spirit (Zhi Shen) is the spiritual element derived from Postnatal Jing and Postnatal Qi. This aspect of the body's Shen is considered to be the conditioned mind or "will," also known as the discriminating, or acquired mind, and constitutes the mind's Yin nature.

   The Postnatal Shen is produced and developed by mental stimulation through contact with the environment after birth. The Postnatal Shen dominates all mental activity and is responsible for thinking and learned behavior. The Postnatal Shen constitutes the true analytical aspect of the mind.

   Prenatal and Postnatal Shen usually interact intermittently with each other in maintaining the body's health. Continual exhaustion of the Postnatal Shen impairs and suppresses the Prenatal Shen. Excessive studying or sleep depravation, for example, can lead individuals to mistake the projection of their deepest fears and desires for genuine intuitive perceptions.

THE FIVE SPIRITUAL ASPECTS AND YIN ORGAN COMPONENTS OF SHEN

Prenatal Shen affects the spiritual disposition which presides over man at the moment of conception. Closely tied to the Yuan Shen are five important, although seldom discussed outside of China, energies or "spiritual entities" called the Five Agents (see Chapter 2). The function of the Five Agents establishes the foundation for the psychological aspect of the Five Element theory used in clinical diagnosis throughout China today.

Each of these Five Agents is connected to, or said to "reside in" one of the Five Yin Organs. The Three Hun reside in the Liver and embody the virtue of Kindness, the Shen resides in the Heart and embodies the virtue of Order, the Yi resides in the Spleen and embodies the virtue of Trust, the Seven Po reside in the Lungs and embody the virtue of Integrity, and the Zhi resides in the Kidneys and embodies the virtue of Wisdom (Figure 13.11). Whereas each of the Five Agents embodies a virtue, each of the Five Yin Organs embodies the energetic qualities of one of the primordial elements. The Liver embodies the element of Wood, the Heart embodies the element of Fire, the Spleen embodies the element of Earth, the Lungs embody the element of Metal, and the Kidneys embody the element of Water.

Each of these Five Agent energies are connected to the five separate yet functionally interactive Yin organs, and thus create complex, multidimensional energy fields (Figure 13.12). These energy fields encompass emotional, mental, and spiritual shifts, as well as energetic transformations. The fluctuations depend upon which of the Five Agents is dominating from moment to moment. The psycho-spiritual attributes engendered are sometimes known as the Wu Jing Shen (Five Essence Spirits). As the universal energy flows through the Taiji Pole, these Five Agents are continuously being energized. The cultivation of the five virtues of these Five Agents nourishes the Eternal Soul.

The Qigong doctor needs to comprehend these multidimensional interactions to trace the complex spiritual, mental, emotional, and physical manifestations of Excess or Deficient conditions to the source or "root." Although Jing and Qi form the physical basis for the Shen, the Five Elemental Yin organs form the components for the Shen and the Five Essence Spirits. The energy of the Five Agents continually circulates through the patient's body, and are described below.

THE LIVER STORES THE ESSENCE SPIRIT "HUN"

The Liver stores the Three Ethereal Souls, also called the Hun. The Hun or Three Ethereal Souls are rooted in the Liver Yin (which includes the Liver Blood). The Liver organ is responsible for the free and easy flow of Qi throughout the body.

The Hun are typically characterized as the Yang, brighter, Heavenly souls. It is through the influence of the Hun that the five virtues of the Five Agents manifest. The Hun leave the body, ascending back to Heaven at the time of death. The ideogram for the Three Hun has two parts. One part is the character for Earthly spirits-Gui (Ghost), represented by a head suspended above a vaporous form of a body, with an appendage (symbolizing the whirlwind that accompanies the movements of the Earthly spirits). The other part of the character is the image for clouds that are seen as vapor rising from the Earth and gathering in the Heavens. The Hun move within the body as freely as clouds, following the Yi (Will) of the "Heavenly breath, within the celestial vault" (following the will of the Eternal Soul stored within the Heart, and originating in the Taiji Pole).

Five Agents		Yin Organs	
志	**Zhi**	腎	**Kidneys**
意	**Yi**	脾	**Spleen**
魄	**Po**	肺	**Lungs**
魂	**Hun**	肝	**Liver**
神	**Shen**	心	**Heart**

Figure 13.11. The Five Agents and Yin Organs

From the ideogram we also get a distinct picture of the spirit rising to the Heavens. The Eternal Soul is different from the spirit in the context of classical Chinese theology, in that the soul is seen as the more personal of the two, whereas spirits are seen as more universal temperaments or as archetypes. The ancient medical classics say that there are Three Ethereal Souls (Hun) and Seven Corporeal Souls (Po) that symbolize different attributes of the human being. The Hun's spiritual energy is said to be able to leave the body and then return, thus indicating a relationship with out of body travel into the spirit world.

The Hun and Po are expressions of the body's "true spirit." When disorder exists (physical, mental, emotional, or spiritual), the Hun may sometimes fly away (like startled birds in a yard), and the Po will either stir about thoughtlessly, in the absence of effective control, or become animalistic in nature and attack.

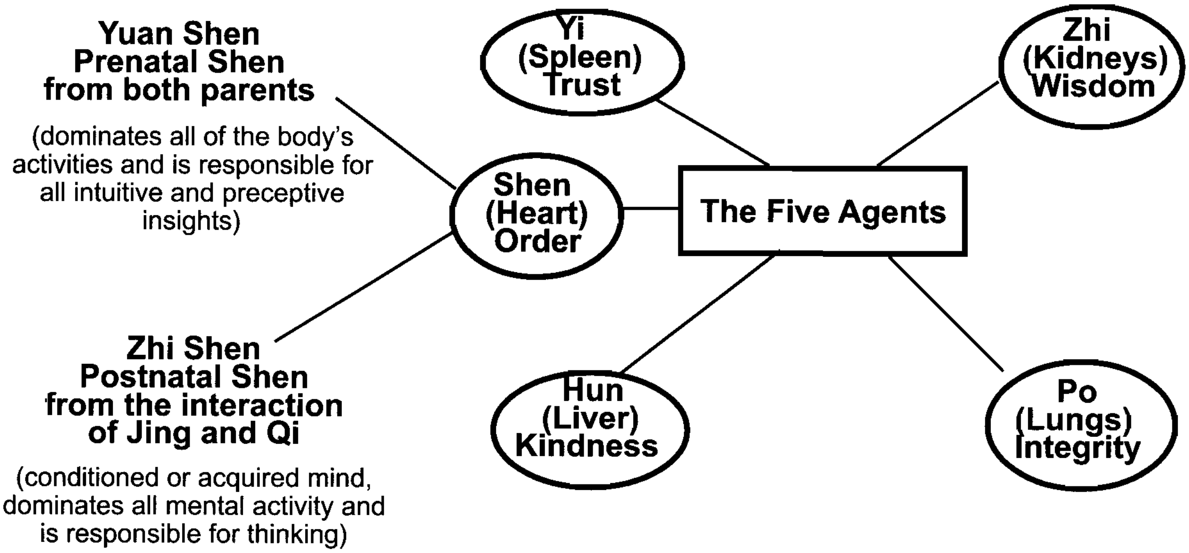
Yuan Shen
Prenatal Shen
from both parents
(dominates all of the body's activities and is responsible for all intuitive and preceptive insights)
Zhi Shen
Postnatal Shen
from the interaction of Jing and Qi
(conditioned or acquired mind, dominates all mental activity and is responsible for thinking)
Shen (Heart) Order
The Five Agents
Yi (Spleen) Trust
Zhi (Kidneys) Wisdom
Hun (Liver) Kindness
Po (Lungs) Integrity

Figure 13.12. Formation of Shen

FUNCTIONS OF THE HUN

1. The Hun control sleep and dreaming. They reside in the eyes during the day and lodge in the Liver at night. When residing in the eyes, they see; when they lodge in the Liver, they dream. Dreams are the roaming of the Three Ethereal Souls. It is the nature the Hun to wander, and it happens easily. At night the Hun must be anchored or rooted in the Liver, so the Liver Blood and Liver Yin must be strong. If they are not, the Hun wanders, and the person dreams too much or has unpleasant dreams. Patients who suffer from severe Deficiency of Yin, may experience a floating sensation just before falling asleep. This condition is due to the Hun not being rooted in the patients' Yin.

   Dreams are an example of the information gathered during the Hun's traveling. The ancient Chinese Qigong masters understood that the body's Hun can traverse the Nine Levels of Heaven or Nine Levels of Earth instantaneously (see Chapter 2).

   The Hun store the sum total of past experiences. The expressions of the Three Ethereal Spirits are manifested through images, symbols and ideas from the divine and Wuji. These images, symbols and ideas emerge into the patient's mind, affecting his or her spiritual life. Without this interaction, the patient's mental and spiritual life would be deficient in images, ideas, and dreams.

   The Hun also controls dreams and daydreaming, as well as the ideals, aims, and direction in life. The absence of these objectives and goals results in feelings of depression. If the body's Liver Yin is depleted, the Hun are deprived of their residence, resulting in such conditions as: fear, excessive day dreaming, insomnia, and a lack of sense of direction or purpose in life (one of the main features of depression).

   Traditionally, sleepwalking is believed to be a function of the Hun. In sleepwalking, the body's Shen is not active and functioning, however the Hun are moving the individual. This is why when treating a patient for sleepwalking, the doctor will emit Qi into the patient's Hunmen Bl-47 (The Door of the Ethereal Soul). In modern research, however, some Qigong doctors have begun to theorize that the Po may take an active role in sleep walking, especially when the sleep state is related to incoherent trauma (patients who thrash about, hurt people, or destroy things when sleeping).

When using hypnosis or "creative visualization" in psychotherapy (the therapist may suggest a specific image or scene for the patient), the therapist is trying to bypass the patient's acquired mind and bring psychic material from the Hun, to avoid the judging aspect of the Postnatal Shen.

With regard to sleep and herbal medicine, if the patient does not sleep well because the Hun are wandering, the Qigong doctor can prescribe sour and astringing herbs (i.e., Bai Shao, Mu Li, Suanzao Ren) which will encourage the Hun back into the body.

2. The Hun assist the Shen in mental activities. This relationship is very important, as the energy of the Shen and the Hun must continually be coordinated. The Shen is related to rational thinking and inspiration. The Hun gives the Shen a sense of direction, projecting outwards, relating to other people. The Shen needs to direct the Hun, to "gather the Hun." The Hun give the Shen its movement and direction, encouraging the Shen to relate with people, to socialize and bond. The Hun, however, also need to be gathered and restrained by the Shen. Otherwise if the Shen is weak and does not control the Hun, then the Hun move about too much and have lots of ideas, but never accomplish anything, leading to frustration.

It is important that the Shen allow the Hun to move in and out of the body, but in so doing must give the Hun direction and purpose. The Hun inspire and give us dreams. That is why the words "movement," "coming and going," and "swimming" are often used in connection with the Hun. In fact, many doctors say that the Hun are the coming and going of the body's Shen. In mediumistic trances, when the spirit of one person enters another's body, it is the Hun that come in; the host's Shen is disabled at that time.

Because knowledge is dependent on the awareness of the Three Ethereal Souls, the Hun are therefore considered the collectors of information. The Hun however, do not interpret, rationalize or analyze the knowledge, they just pass the information on to the Shen, which is responsible for rational thinking, intuition, and inspiration. The Shen helps distinguish between the useful and non-relevant information.

There is an interesting correlation with the Hun controlling sleep and dreaming, and the Hun being rooted in the Liver Blood and Yin. The Hun's moving in and out of the Liver governs the free flow of Liver Qi and vice-versa; the free flow of Liver Qi is a manifestation of the swimming energy of the Hun, moving in and out of the tissues.

3. The Hun maintain balance in one's emotional life, under the leadership of the Shen.

Everyone experiences emotions in life; that is normal. The Hun are responsible for keeping a balance, so that the emotions do not become excessive and thus become the cause of disease. The Hun have a regulatory function, closely related to the balance between Liver Blood (Yin), and Liver Qi (Yang). The Hun's function here is the mental equivalent of the Liver's emotional regulating and harmonizing function.

Due to its acquired nature, the Postnatal Shen discriminates, while the Hun does not. The relationship between the Shen and the Hun is very similar to concepts of consciousness and unconsciousness in Jungian terms. The Hun are a repository of images and archetypes, connecting the (personal) subconscious mind and the collective unconscious. If the Hun are unsettled, then the Shen (consciousness) is cut off, confused, isolated, aimless, sterile, and without dreams. The movement of ideas within the body has to be controlled or it gets out of hand, resulting in madness. This madness occurs to the degree that the Shen doesn't control the Hun anymore, and the patient receives an uncontrolled amount of emotional, mental, and spiritual input. All the energies and symbols coming through the Hun have to be integrated and assimilated. Otherwise there can be serious mental illness and possibly psychosis (except in young children who are continually full of ideas and have active imaginations). In a child, the Hun are very active and the Shen is not as restraining, so there is a continual flow of energies streaming from the unconscious world of symbols, without the result of insanity.

4. The Hun are responsible for the eyes. When the Hun wander in through the eyes, the eyes can see. The Hun give us vision, both mentally and spiritually.
5. The Hun influence a person's courage. If the Hun are not strong, the person is timid and fearful. If the Hun are strong, the person is fearless, can face up to the difficulties in life and dares to take action. A patient with weak Hun will have difficulty gathering information, making decisions, will lack courage, and become easily discouraged and apathetic.
6. The Hun control planning with the aid of the Shen and the sense of direction. The mental and spiritual confusion about one's role in life, what to do, what goals to set, can be compared to the aimless wandering of the Hun. This is a strong feature of major depression. If the Liver is strong and the Hun are strong and rooted, the person has a strong sense of direction.
7. The Hun control Spirit Travel. By housing the Shen for "spirit travel" (the spirit journeying outside of the physical body), it is possible for the Qigong doctor to consciously direct the Hun's travelling. This is different than Astral Travel, which is the Eternal Soul journeying outside of the physical body, connected to the Middle Dantian by a silver "cord of life." The Three Ethereal Souls, accompanied by the individual's consciousness, act as one unit which is sometimes referred to as the Spirit-Soul. The Spirit-Soul allows the doctor to know the exact location of the "spirit routes" travelled, as well as which path and direction it takes when it leaves the body. Otherwise, when the Hun wonder, the doctor "Shens out" and has no recollection of where he or she has been in the spiritual realms.

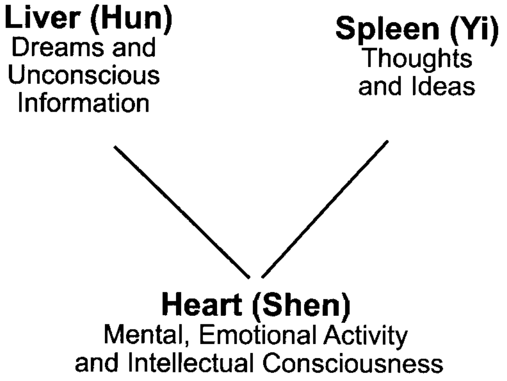

Figure 13.13. Two Ways to Feed the Shen

THE HEART STORES THE ESSENCE SPIRIT "SHEN"

The Heart stores the Shen and is responsible for mental and emotional activity, intelligent consciousness, long term memory, and sleep. It is responsible for organizing mental concepts and activities, and has the capacity to judge.

The Shen is divided into two kinds of energies, Prenatal Shen and Postnatal Shen.

The Prenatal Yuan Shen is associated with the congenital spiritual energy that descends from Heaven. It is the universal divine spirit that is implanted in each person, uniting man to the spiritual realm (that which exists beyond the physical and mental reality). It has been compared to the "Christ within us," or "higher self." This is the energy Qigong doctors attempt to become filled with during sitting meditation practice and prayer.

The Postnatal or Acquired Shen is associated with the wisdom of the five senses and the acquired knowledge of the individual's experiences.

There are two ways to "feed" the Shen. One way is through dreams or unconscious information coming from the Hun of the Liver; the other way is through thoughts and ideas from the Yi of the Spleen (Figure 13.13). A person who has a well-developed, strong, and abundant Shen comprehends things immediately and is able to apply what he or she has learned.

The Hun give the Shen movement allowing the mind the capacity of self insight and introspection, as well as the ability to project outwards from the body and envelop, as well as relate to, other people.

THE SPLEEN STORES THE ESSENCE SPIRIT "YI"

The Chinese term Yi, is typically translated as mind, thought, opinion, idea, sentiment, inclination, intention, intellect, scholar mind, analytical thinking, and memorization in scholastic learning. Because the Spleen stores the Yi, it is responsible for the transformation and transportation of all thoughts and ideas on an intellectual level, through study, concentration, and memorization. This is different from the Hun's creativity. It is the interaction between the Yi of the Spleen and the Shen of the Heart that allows man to place how he thinks, speaks, and acts into sounds, thoughts, and actions.

Concerning memory, there is a lot of overlap between the functions of the Spleen (responsible for memory in terms of concentration and studying), the Kidneys (responsible for storing things on a day to day basis), and the Heart (responsible for storing past events).

The Postnatal Qi and Blood form the physiological basis for the intellect. A person with well-developed Yi has total recall of events experienced in their personal past, and is able to memorize things easily. A patient with an Excess of Yi will be obsessed with the past, while a patient with Deficient Yi will be absentminded, suffer memory loss, be inattentive, and have problems in maintaining concentration and mental focus. Although all physical pain is registered by the Po, and all psychological pain is registered by the Hun, only the memory of the pain is registered by the Yi (Figure 13.14). Therefore, a Deficient Yi condition can often lead to the phenomenon of transference and countertransference. Transference involves the incorrect attribution of certain emotions and intentions onto another person (usually onto the doctor by a patient). Countertransference involves the same type of projection of one's feelings, colored by one's own expectations in response. In countertransference, the doctor or therapist is doing the projecting.

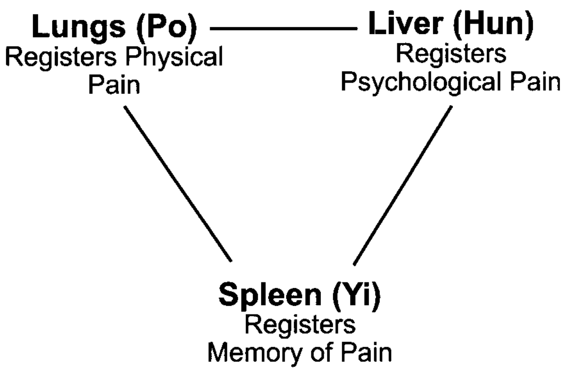

Figure 13.14. Three Ways the Body Registers Pain

Yi also has the meaning of divine purpose when applied to the Yuan Shen energy. Whenever the physical energy is purified by self-cultivation and through connecting to the divine, the intention of the Shen and the divine purpose become one and the same. It is said when the Yi is conserved, it will help build the Zhi (or willpower). The Yi is filled with information from the past along with current knowledge and sensations.

THE LUNGS STORE THE ESSENCE SPIRIT "PO"

The Lungs store the Seven Corporeal Souls (Po), which are physical in nature and are attached to the body's Jing and Qi. The word Po is defined as vigor, animation, or life. There are two parts to the Po ideogram; one is the character for Gui, the spirits of the Earth, and the other represents the color white. Thus, the Po are linked with a descending movement of energy and with the Jing. The Seven Corporeal Souls are also said "to come and go, enter and exit," in association with the body's Essences.

In Medical Qigong therapy, the body's Eternal Soul is seen as being strongly influenced by two main divisions of internal spiritual energies, the Hun and the Po. The Hun are the Three Ethereal Souls and represent the positively charged aspects of the Eternal Soul, while the Po are the Seven Corporeal Souls and are considered the physical, negative, Yin, heavy, and Earthly aspects of the Eternal Soul. The Po are the counterpart of the Hun and can be described by its following six manifestations.

1. The Po are the somatic expressions of the Eternal Soul, related to the reflexive nervous system and limbic system (the "reptilian brain"). They manifest through the body's sensations of feeling, hearing, and seeing. They have an impulsive tendency towards action and correspond to the deep animal instincts within the Mind (personal subconscious) and cells (our reflexes are a Po reaction). The Po also provide us with the animal strength and resources necessary to mobilize the body and perform incredible feats of power. The "animal within" is driven by the Po. They are the manifestation of the body's Jing in the sphere of sensations and feelings. Just as the Three Hun provide the individual with the energetic movement for the Shen, the Po provides the individual with the energetic movement of the body's Jing.
2. The Po are responsible for all physiological processes in childhood. In the beginning of life, the Po are responsible for the sensations of pain and itching. The Po serve as the intermediary between the Jing and the body's other vital substances. At conception the interaction of the body's Jing not only forms the embryo, but also establishes the Po within the body. Although the Po are stored in the Lungs, they stay attached to the body until death, there upon returning their energy back to the Earth within a few days.
3. The Po are related to weeping and crying. The interconnection between the Po and Lungs is very important from an emotional point of view. When the Po's movement in the Lungs is constricted, grief and sadness are suppressed in the chest through shallow breathing.

   In the morning time, when waking, if the patient feels dull and depressed, it is a sign that his or her Shen is clinging to the body, i.e., the patient's Shen is clinging to the excessive energy of the Po.
4. The Po are closely linked to breathing. Because the Po reside in the Lungs, all meditation systems, breathing exercises, and forms of breath control are methods used to regulate the Po, calm the Shen, and access the higher Hun states. Breathing is the pulsation of the Po. Each emotional change the body experiences is related to a shift in the body's respiration rhythm. Because breath control relies on the Lungs which influence the sympathetic and parasympathetic nervous systems, the quality of Qi and its circulation are dependent upon the method, speed, and quality of the breath. To support the greatest longevity possible, it is important to breathe with a long, slow, and even quality to the breath. The breath (air from Heaven) interacting with the Po in the Lungs, plays a significant role in the Heaven-Man-Earth concept for balancing the emotions.

   The Lungs' virtues are righteousness and courage. These virtues give a person the drive and strength to do the "right thing" when the need arises. These virtues manifest and promote good health when a correct energetic and balanced alignment with the Po is achieved (between the Lungs breathing pattern and the Po).
5. The Po are connected to sexuality on the sensation level. All instinctive sexual reactions and passions come under the authority of the Po. In human life, the Po are linked to the basic instincts and perceptions of the body. The Po provide the fundamental biological energy, and are the source of biological needs and impulses. They are driven by basic instincts and urges, and their sole concern is the immediate gratification of biological needs and impulses (emotional and physical survival, reproductive urges, etc.). Most self destructive behavior, such as an attraction to unhealthy and dangerous life-styles, is due to the Po.
6. Because of the Corporeal Souls' relationship with the Lungs and Large Intestine, the anus is considered the "Po Men" or "the door of the Po." It acts as a doorway for the elimination of the waste products for the Five Yin Organs, through draining off impure liquids and waste.

THE KIDNEYS STORE THE ESSENCE SPIRIT "ZHI"

The Chinese term Zhi (translated as "will") is the prenatal spiritual entity associated with the Kidneys, and is not the personal will of the ego that is driven by our desires. The term Zhi has two meanings. It means memory, but it also means willpower. Both are primary features of the Zhi. The word Zhi can also be used to mean the Mind (whole body consciousness and awareness). In connection with the five mental aspects of the Mind, the Three Ethereal Souls, Seven Corporeal Souls, the Shen, the "intellect," and the "willpower" are sometimes referred to as the Five Zhi.

MEMORY

Memory is defined as the ability to remember information when studying or learning a particular subject or pattern. The Kidneys keep a determined focus on our goals, not forgetting where we are going and what we are working to achieve; they maintain a vital mindfulness. The Kidneys relate to short term memory, whereas, the Heart governs the body's long term memory (Figure 13.15). This is why elderly patients, whose Kidneys are declining, often cannot remember what day or year it is, but can remember events long past.

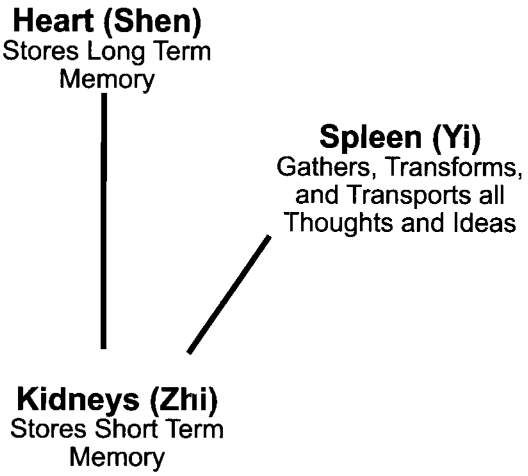

Figure 13.15. Three Ways the Body Stores Memory

WILLPOWER

This is the most important aspect of the Zhi, because it is responsible for supplying the mental drive, willpower, determination, and single minded pursuit of goals and aspirations. It enables the realization of ambitions by providing the focussed energy necessary to carry ideas to fruition. A powerful Zhi creates the magnetism and charisma necessary to manifest and materialize our dreams.

Even though a person has acquired all the information available via the Hun, Po, and Yi, without the Zhi there can be no action. Although the Hun give us the goal (thinks of something), the Zhi is needed to accomplish it (decides and acts on it). A person with well-developed Zhi demonstrates perseverance, determination, and a tenacity to complete personal goals. Patients with Deficient Zhi become indecisive and fearful. Patients with Excess Zhi have a blind obedience to authority, or tend to be fanatics.

The goal of the Qigong doctor is to reach such a level of attainment that his or her personal will merges with Heaven's will (Tian Zhi) becoming one and the same. Tian Zhi is considered the movement of the divine expressed in man as his personal virtue (De).

A man's virtue is defined as his spiritual righteousness, the authenticity of his heart and actions. It is through the acquisition of virtue that man finds and possesses his true nature. If man's virtue is initiated into the "mysterious" (embracing Heaven's will), his intuitive evolution will give way to limitless perception.

Tian Zhi is a divine inner prompting that guides us on our spiritual quest, if we are open to its message. Tian Zhi is what we can call the divine will and carries within it man's purpose in life.

THE FIVE YIN ORGAN CULTIVATION OF SHEN

The Shen indicates the vitality level, and the state of physical, mental, emotional, energetic, and spiritual being. A flourishing vitality is a manifestation of the Jing, Qi, and Blood inherent within the Five Yin Organs. These five components of the Shen are expressed through psychological manifestations (Figure 13.16). They form the bases of the patient's emotional characteristics. Each of the emotional components of the Five Yin Organs contribute to the patient's spiritual manifestation; adding or detracting from the natural harmony of the patient's spiritual nature.

The activity of conscious thinking, insight, and intelligence are the foundational aspects of the body's Shen, activated by the patient's Heart (wherein the Shen resides). The Heart is also responsible for the long term memory of the body's sensory memories (the perceptions, feelings and sensations related to hearing, seeing, smelling, tasting, and touching) and sleep. These sensory activities carry over from the other internal organs, due to their relationship with the Five Yin Organs. Emotional disturbances that affect the Heart can cause energetic changes in the patient's body, influencing health, toward recovery from diseases or illness by impeding or facilitating bodily functions (see Healing the Patient's Emotional Pain, Chapter 31).

When cultivating the Shen, it is important for the Qigong doctor to keep in mind that the spirit is indivisible from matter, and that an individual's emotional and spiritual well-being and development depend upon the healthy functions of the internal organs. When the Shen leaves the body (through shock, trauma, stress, depression, guilt, shame etc.), the Qigong doctor can observe a diminishment of glow in the patient's complexion, lack of luster in the eyes, along with changes in the respiration rate and state of mind.

Once the patient's Shen is disturbed, the Qi is affected. This in turn leads to the body's Jing becoming weakened. Patients with this condition are generally taught Medical Qigong exercises and prescriptions to not only tonify and strengthen their body's internal organs to improve their immune system, but also to improve the overall condition of their emotional/spiritual state. Even a low vitality of spirit can cause emotional disturbances that can scatter or stagnate the body's Qi.

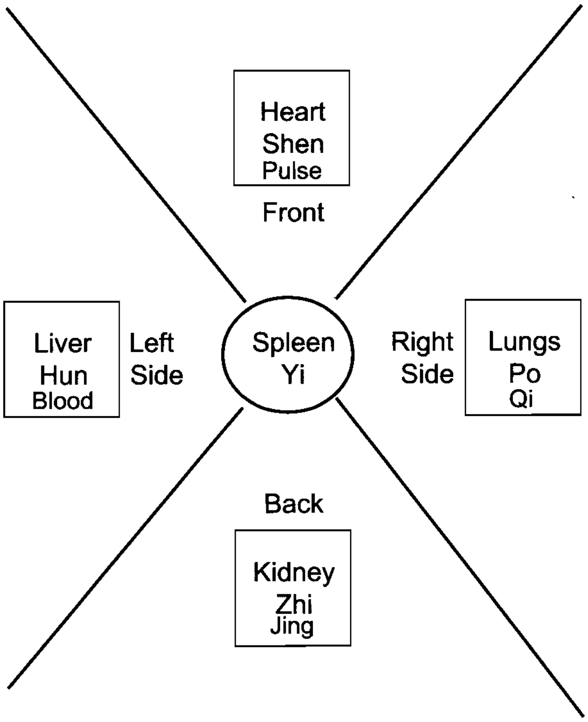

Figure 13.16. The body's Shen is physically housed in the Five Yin Organs. Each Yin organ houses a particular psycho-spiritual and energetic aspect of the human being. For example: The pulse is stored in the Heart, which is the residence of the Shen (Spirit), and placed in the front of the body, the Blood is stored in the Liver, which is the residence of the Hun (Three Ethereal Souls), and placed on the left side of the body, the Jing is stored in the Kidneys, which is the residence of the Zhi (Willpower), and placed on the back, and the Qi is stored in the Lungs, which is the residence of the Po (Seven Corporeal Souls) and placed on the right side of the body.

The Yi (Intention) resides in the Spleen and is placed in the center of the body, governing the Qi of the four aspects of spiritual manifestation, distributing them along the body's Taiji Pole. When Jing and Blood combine Shen is born. The Hun follow the Blood along in all its movements. The Po follow the Jing along in all its movements. The state of Qi and Blood of each organ can influence the Mind and the Spirit, any alteration of the Mind or Spirit affects one or more of the internal organs.

Advanced cultivation methods utilize the energy of the Five Yin Organs, flowing to and from the body's Taiji Pole. It is from the Thrusting Channels along side the Taiji Pole that the energy of the Shen is dispersed into the body's Yin and Yang Organs and then into the extremities (Figure 13.17 - 13.18).

SHENGONG CULTIVATION METHODS

The key to enhancing the Qigong doctor's development of energy, is through learning how to control the vast amount of Qi within the body, without compulsively "Shening out" (allowing the spirit to leave because the body feels uncomfortable). To avoid disconnecting from the body, the Qigong doctor practices several specific training methods of Shengong cultivation. When focusing on Shengong cultivation and regulation, there are four traditional disciplines the doctor must adhere to: Nourishing and Strengthening the Shen, Housing the Shen, Combining the Shen with Respiration, and Combining the Shen with Qi (Figure 13.19).

NOURISH AND STRENGTHEN THE SHEN

The Qi is able to nourish and strengthen the Shen through proper Qigong regulation. This is brought about through a form of coherent light meditation, that focuses the doctor's Shen on a small circle of light (generally about the size of a marble or pearl). The doctor's focus of concentration can extend either externally onto a specific point outside the body, or internally on a specific point inside the body (e.g., the Upper Dantian). Through this focused "coherent light" (light with wave patterns that are all in step or in phase), it is possible to build the Shen's strength and refine its power. Through focused intention on a specific area, the Qi and Shen condenses together forming a more powerful type of energetic projection.

HOUSING THE SHEN

By disciplining (i.e., quieting) both the emotions and thoughts the Shen becomes relaxed and tranquil, and remains in its residence (the Heart). The Shen must not be allowed to become attracted to outside emotional disturbances. The goal is to allow the Shen to be energized but not excited.

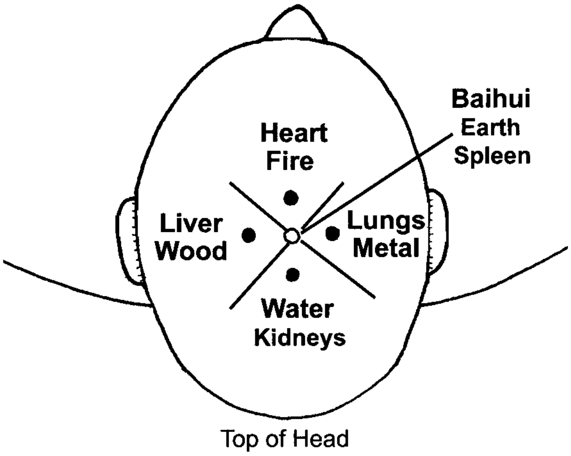

Figure 13.17. The Five Thrusting Channels as seen from the top of the head. Located within the Five Thrusting Channels are the "Five Portals of the Heavenly Gate." These gates include four entry points which surround the Baihui (One Hundred Meetings) point. These four points are known as the Sishencong or "Four Spirit Hearings" and are the access points for the upper spiritual matrix surrounding the body's Taiji Pole, and connect the Five Thrusting Channels of the upper gates with the Five portals of the Earthly Gates.

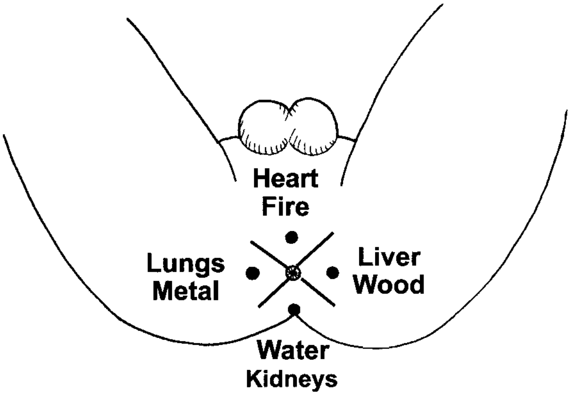

Figure 13.18. The Five Thrusting Channels as seen from the perineum. Located on base of the Five Thrusting Channels are the "Five Portals of the Earthly Gate." The Fire point is at the location of the Huiyin point on the Conception Vessel (CV-1), the Water point is at the location of the Changqiang point on the Governing Vessel (GV-1), the Wood and Metal points are located just below the ischial tuberosity, lateral to the anal sphincter. These points are known as the access points for the lower spiritual matrix surrounding the body's Taiji Pole.

COMBINING THE SHEN WITH THE RESPIRATION

Once the Shen has been nourished, strengthened, and housed, it is then connected with the doctor's breath. Once the Shen is combined with the respiration, maximum healing results can be obtained by releasing Healing Sounds and Qi projection simultaneously.

COMBINING THE SHEN WITH THE QI

The final stage is for the doctor to be able to direct the Shen in coordination with the life-force energy being emitted into the patient. Since Shen projects as light and Qi projects as vapor, this combination of Spirit and Energy projection is very powerful, especially when combined with healing colors and sounds.

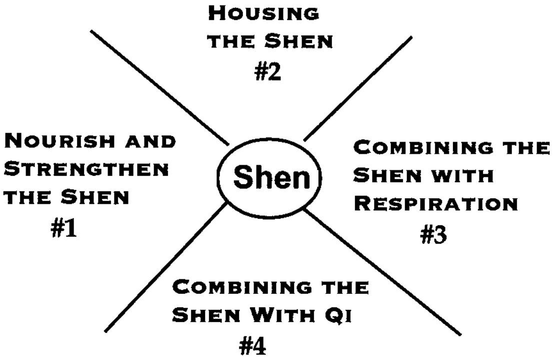
HOUSING THE SHEN #2
NOURISH AND STRENGTHEN THE SHEN #1
Shen
COMBINING THE SHEN WITH RESPIRATION #3
COMBINING THE SHEN WITH QI #4

Figure 13.19. Four Ways to Cultivate the Shen

THE SIX TRANSPORTATION'S OF SHEN

The psychic abilities developed in Medical Qigong training result from a well-developed intuitive faculty, brought about through Shengong meditations. Through these meditations the Yuan Shen can transcend the space-time continuum, as both space and time are multidirectional and interconnected.

Shengong meditations cause altered states of consciousness that differ substantially in subjective and objective observation from the general norms for that individual. As the Qigong doctor dissolves his or her Shen into the Wuji, the subconscious will be able to see forward and backward in time, unhampered by distance.

The conscious mind acts as a filter, seeking to analyze every perception, and attempts to identify and categorize each perception. Once the conscious mind is engaged, the flow of perception usually stops because the perceptions arising from the subconscious are usually vague (often with multiple messages). Qigong doctors must train themselves to first root (ground/stabilize) the acquired mind, in order for their perceptions to come through in a pure form, void of thoughts and judgments.

The ancient Qigong masters explained the ability to access the manifestations of Shen as follows: "by rooting the acquired mind (Zhi Shen), the Heart (Yuan Shen) opens up to 10,000 voices (Wuji)." Meaning that, when the "chattering mind" is anchored into the Earth through focused intention, the Heart is able to perceive the spiritual realm (the 10,000 voices), or be aware of its connectedness to the multidirectional space-time continuum.

Extraordinary abilities that develop from the cultivated spiritual consciousness include: telepathy, clairvoyance, clairaudience, precognition, time travel into the past, levitation, and teleportation.

The Qigong doctor experiences different psychological adjustments when opening to the deep spiritual transitions of Shengong meditations. The adjustments the doctor will experience include:

1. Time distortion,
2. Deep mystical states of perception and intuition,
3. Deeper and more powerful levels of self healing,
4. A more spontaneous reaction to movement and sound, and
5. Difficulty communicating verbally.

CULTIVATING THE SPIRITUAL CONSCIOUSNESS TO OBTAIN THE SIX TRANSPORTATION'S OF SHEN

In Shengong meditations, it is through the use of imagery and creative visualization that the doctor's altered states are developed, and a connection is forged with the Yuan Shen and Eternal Soul.

After the doctor's Shen has entered into a quiescent state, his or her altered state of spiritual conscious can be initiated through choosing one of the following five images.

1. The doctor is instructed to go inward into the center of the Taiji Pole to connect with his or her Yuan Shen and Eternal Soul.
2. The doctor is instructed to go into an imaginary temple or sacred place located within the astral plane to connect with his or her Yuan Shen and Eternal Soul.
3. The doctor is instructed to float upwards into the Heavens to connect with the eternal source of his or her Yuan Shen and Eternal Soul.
4. The doctor is instructed to ascend out of the body (after reliving the death of a past life experience) and receive the knowledge learned from that experience collected from by the Yuan Shen and Eternal Soul.
5. The doctor is instructed to allow the Yuan Shen to directly connect with the Eternal Soul and transmit information.

The Higher Self and Eternal Soul are usually considered as the same thing, and are often regarded as the "Master Within." In Jungian psychology, the collective unconscious (Wuji) is accessed through the subconscious mind (Shen) in the form of dreams or directly through "visions." The personal subconscious (Shen) has several elements:

- the anima, which is the wise man (influenced by the Hun)-for males, or
- the animus, which is the wise woman (influenced by the Hun)-for females, plus
- the shadow, which is the individual's dark side (influenced by the Po).

Once the Qigong doctor has cultivated the Shen into a high level of effective energy (absorbing and projecting energy from the universal, as well as the immediate environment), the result is a gradual unveiling of six supernatural diagnostic powers known as the Six Transportations of Shen.

These six metaphysical abilities not only enable the Qigong doctor to accurately evaluate a patient's state, but also to predict the probable future progressions of the patient's life as well as of his or her disease. The Six Transportations of Shen are: Observing the Present, Comprehending the Past and Observing the Future, Knowing a Person's Thoughts, Perceiving a Person's Destiny, Hearing the Sounds of the Universe, and Examining the Universe (Figure 13.20).

OBSERVING THE PRESENT

The Qigong doctor must first quiet the Zhi Shen's (acquired mind) analytical chattering by housing its energy within the Five Yin Organs. As the Mind becomes rooted, the doctor's breathing slows and the Lung Qi harmonizes with the Heart Qi. Once the Seven Po are quieted in their residence and the Zhi Shen is rooted, the Yuan Shen's messages can come through clearly without interference. The Yuan Shen is always connected to the Wuji where all knowledge is stored. The Yuan Shen is always at the "center of time" where it can gather and absorb the patient's past and present experiences, emotional reactions and thought patterns. It is from the Qigong doctor's intention and focus to be at the "center of time" that he or she can observe the present state of "now."

Observing the past is the ability to analyze and accurately discern the actual state of present situations (the here and now). By keeping away from conscious intervention, the Yuan Shen is free to truly listen and perceive as an "observer." The doctor's ego is not engaged and his or her mind's subconscious observer receives and reviews the patient without bias.

This enables the doctor to thoroughly analyze and comprehend events and situations clearly and unambiguously. It is through this state of perception that the Qigong doctor can accurately discern the true emotional problems which disturb his or her patient's energetic balance.

COMPREHENDING THE PAST AND OBSERVING THE FUTURE

Comprehending the past and observing the future is the ability to analyze and understand the thought patterns of the patient, the reasons these patterns take place, and to accurately predict the outcome of future events based upon those pat-

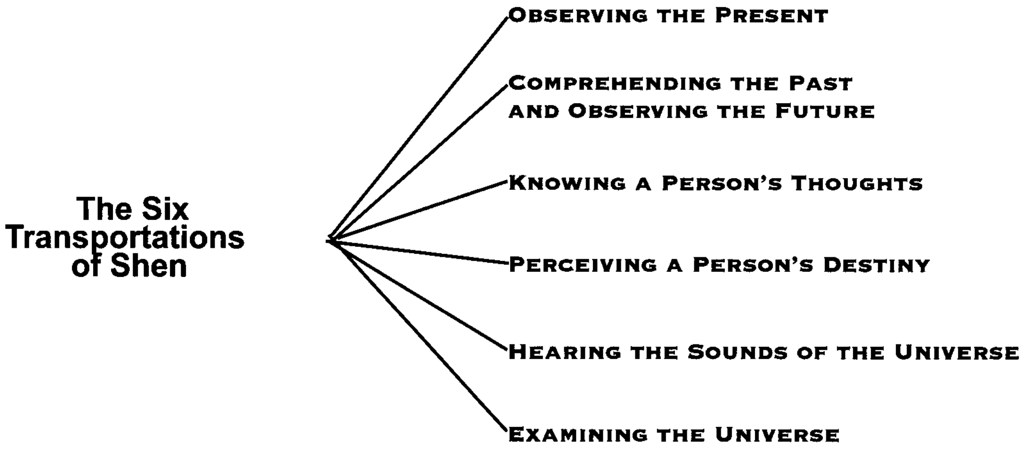
The Six Transportations of Shen
OBSERVING THE PRESENT
COMPREHENDING THE PAST AND OBSERVING THE FUTURE
KNOWING A PERSON'S THOUGHTS
PERCEIVING A PERSON'S DESTINY
HEARING THE SOUNDS OF THE UNIVERSE
EXAMINING THE UNIVERSE

Figure 13.20. The Six Transportations of Shen

terns. By observing the patient from the active state of being in the "present," the Qigong doctor is free to travel within the infinite space (Wuji) enveloped within the patient's tissues. The doctor's Yuan Shen has no need to astral travel to obtain this information. The doctor's energetic connection to the patient's energy body gives him or her direct access to the personal information needed. It is from this internal space that the Qigong doctor can begin to access the memories existing within the patient's body (stored within the cells of the tissues).

At the subatomic level, the body and its cells contain a great deal more energetic space than matter. The doctor can directly access the internal Wuji enveloped within the patient's tissues and cells to retrieve the patient's past memories and emotions. Through tracing the patient's emotional, mental and spiritual development a greater understanding of the patient's present condition is achieved. After the doctor's Yuan Shen has gathered this information, it can be further analyzed by the doctor's Zhi Shen, and organized to accurately determine the cause and effects on the patient's present state.

Through understanding past patterns of thought and emotional reactions, the doctor can then predict future patterns of change and transition. These predictions are quite accurate, since people are prone to repeat programmed patterns indefinitely, unless there is an internal spiritual transition that frees them from this subconscious patterning.

KNOWING A PERSON'S THOUGHTS

This is the ability to sense the thoughts, judgments, fears, and emotions of another person, knowing their innermost fears and subconscious thoughts. The subconscious mind of the doctor communicates with the subconscious mind of the patient. This is a nonverbal, telepathic communication. It is therefore important for the doctor to monitor his or her thoughts and feelings, as they are automatically being projected into the subconscious mind of the patient. After the treatment is over the doctor may discuss with the patient any important information received that would be helpful to the patient.

PERCEIVING A PERSON'S DESTINY

This is the ability to predict future events and upcoming transitions in an individual's life based on the doctor's understanding of the patient's past energetic patterns. Through the Yuan Shen's ability to perceive the past and present, the doctor is able to understand the patient's mind, thoughts, acquired characteristics and personality. Above all, however, the doctor is able to perceive the individual's true nature that exists behind all of the patient's external masks. The doctor's Zhi Shen (postnatal mind) is then able to analyze and categorize the individual's patterns. Through deductive thinking, the doctor is able then to perceive the probable outcome and development of these repetitive patterns. This enables the doctor to help the patient to identify the patterns that require restructuring and give added encouragement and advice, when needed.

This spiritual counsel or guidance is extremely important for healing transitions since old patterns established in the three energy matrixes (Jing, Qi, and Shen) must be changed together to achieve a more permanent healing. Otherwise, the illness is likely to reappear. The spiritual matrix sets the foundation for the energetic matrix, which in turn sets the foundation for the physical matrix.

HEARING THE SOUNDS OF THE UNIVERSE

This is the ability to hear sounds, music, and voices not audible to the normal hearing. This is not to be confused with schizophrenia -- a mental illness that causes a person to hear voices; the difference is that the schizophrenic cannot think clearly, the mind is very confused and unordered.

The ability to hear the sounds of the universe is also the ability to be aware of, and receive, divine inspiration or messages (clairaudience). These abilities allow the doctor to affect the lives of those they know or soon will know. The doctor's own inner-voice can guide, instruct, teach, and lead the doctor to ever higher states of enlightenment.

Through inner hearing, the Qigong doctor is able to listen to, as well as understand, the resonant sounds generated by life-force energy and spiritual entities. This enables the doctor to tap into an unlimited source of divine knowledge and wisdom, which in turn allows the doctor to be present for the patient's highest good.

EXAMINING THE UNIVERSE

This is the ability to actually perceive current objects, events, or people that may not be discerned through the normal senses. It is the ability to perceive both time and space on a spiritual dimension (clairvoyance), to understand and acknowledge one's part, and to accept one's divine calling in life. It is also the ability to understand how one's spiritual calling affects both the cosmos and mankind. Through inner-vision the Qigong doctor is able to see inner-visions of divine enlightenment, observe and exchange information with celestial beings, and find holy places for spiritual sanctuary.

The metaphysical abilities of the Six Transportations of Shen are spiritual manifestations from the Eternal Soul. Once the Qigong doctor regulates his or her spiritual life, these abilities become a natural course of experience, giving way to the development of eight supernatural powers. A chart describing the interaction of the body's Eternal Soul with the energetic functions of the Three Dantians and Five Yin Organs' Wu Jing Shen (Five Essence Spirits) is depicted in Figure 13.21.

THE EIGHT SUPERNATURAL POWERS

Once an individual has obtained the Six Transportations of Shen, his or her psychic abilities develope into eight supernatural powers. These powers can be utilized in the individual's personal spiritual practice. Through these eight supernatural powers the doctor's Yuan Shen can allow him or her the ability to transcend all physical boundaries and invoke certain spiritual manifestations. The eight supernatural powers are described as follows.

1. The ability to make the one's physical body light.
2. The ability to make the one's physical body heavy.
3. The ability to reduce one's energetic body to the size of an energetic particle. This allows an individual the ability to astral travel into the subtle energetic and spiritual realms of existence.
4. The ability to expand one's energetic body to an infinite size. This allows an individual the ability to connect to, through, or envelop objects, as well as energetically protect people, places and things.
5. The ability to call into existence and attain one's specific needs or desires. Due to the individual's strong connection to the divine,

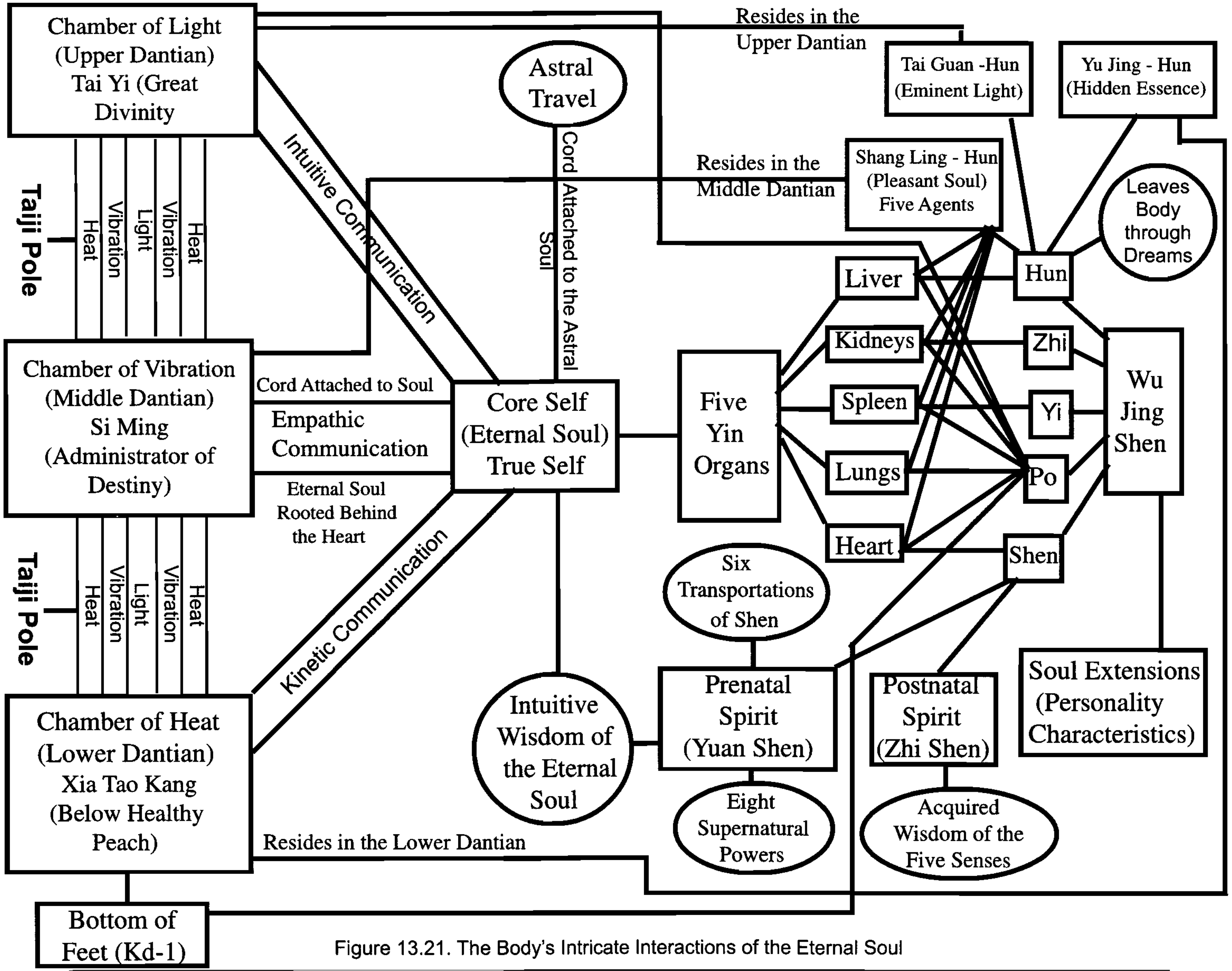
Chamber of Light (Upper Dantian) Tai Yi (Great Divinity
Astral Travel
Resides in the Upper Dantian
Tai Guan -Hun (Eminent Light)
Yu Jing - Hun (Hidden Essence)
Taiji Pole
Heat
Vibration
Light
Vibration
Heat
Intuitive Communication
Cord Attached to the Astral Soul
Resides in the Middle Dantian
Shang Ling - Hun (Pleasant Soul) Five Agents
Leaves Body through Dreams
Liver
Hun
Kidneys
Zhi
Spleen
Yi
Lungs
Po
Heart
Shen
Wu Jing Shen
Five Yin Organs
Chamber of Vibration (Middle Dantian) Si Ming (Administrator of Destiny)
Cord Attached to Soul
Empathic Communication
Eternal Soul Rooted Behind the Heart
Core Self (Eternal Soul) True Self
Taiji Pole
Heat
Vibration
Light
Vibration
Heat
Kinetic Communication
Six Transportations of Shen
Chamber of Heat (Lower Dantian) Xia Tao Kang (Below Healthy Peach)
Intuitive Wisdom of the Eternal Soul
Prenatal Spirit (Yuan Shen)
Postnatal Spirit (Zhi Shen)
Soul Extensions (Personality Characteristics)
Eight Supernatural Powers
Acquired Wisdom of the Five Senses
Resides in the Lower Dantian
Bottom of Feet (Kd-1)

Figure 13.21. The Body's Intricate Interactions of the Eternal Soul

he or she may manifest through faith that which is desired.

6. The ability to experience and enjoy one's desires fulfilled.
7. The ability to gain authority over one's situations. First and foremost, the individual will obtain the spiritual authority over a specific situation, then the authority over the emotional, mental and physical realms follow.
8. The power to attract and control things. This allows an individual the ability to engender help and security (in the form of outside council, finances, and spiritual and emotional support) in order to exercise power over specific situations.

SOUL PROJECTION, SPIRIT PROJECTION AND ENERGY PROJECTION

The phenomena of Soul Projection, Spirit Projection and Energy Projection is known and accepted as a natural part of existence in almost every culture in the world. The holy men and women, shamans, and energy healers of many primitive societies were required to prove their powers by exercising these spiritual and energetic abilities.

Tibetan Buddhism (influenced by the energetic practices of Indian cosmology) contains elaborate doctrines about Soul Projection (also known as Astral Projection), Energy Projection, the Astral Body, and also describe the feats individuals can perform while out of their physical bodies. Some of these doctrines are quite similar to the traditional Daoist teachings which are evident in advanced Daoist Shengong training.

All Chinese systems used for inducing Shen and Qi projections use visualization, concentration, willpower and the individual's innate skill. The three types of external projections are: Soul Projection, Spirit Projection, and Energy Projection.

- Soul Projection occurs when the body's Eternal Soul is externally projected out of the body, while still attached to the Middle Dantian through a Silver Cord;
- Spirit Projection occurs when the body's Shen is externally projected outside of the body as an energetic vehicle of light;
- Energy Projection occurs when the body's Qi is externally projected out of the body like an energetic vapor (Figure 13.22).

In Soul Projection, Spirit Projection and Energy Projection, the effects of time are not acknowledged outside of the physical realm. Within the Astral Plane, time is omnipresent, there is no past or future, only the individual's perception of it.

The body has an established relationship to both the physical and spiritual world. The physical world of the living is considered a Yang state of existence, while the supernatural-spiritual world is in a Yin state of existence (see Chapter 19). As the Qigong doctor's Shen reaches a high state of sensitivity, he or she can transcend the ordinary senses and normal conscious perceptions. These abilities enable the doctor to observe, communicate and even travel within the "higher" and "lower" spiritual Yin world. There are nine energetic levels of Heaven (i.e., the higher astral planes), and nine energetic levels within the Earthly spiritual realm (or lower astral plane).

It is important to note that the term "higher" and "lower" astral realm has nothing to do with the actual space, but with the energetic and spiritual vibrational levels. The higher and lower astral planes overlap and exist within the same space as the physical plane. This blending between the higher planes, lower planes and the physical plane creates energetic dimensions (or places) where the separation between these worlds is thin. In these energetic dimensions, lower beings can reach and attach onto an individual -- even though the individual is not directly within the astral regions.

The energetic frequency of each astral being determines the spiritual level wherein it exists. Beings of higher frequency, for example, vibrate at a faster rate and reside in the higher, less dense astral dimensions; whereas beings of lower energetic frequency can only dwell in the dense lower astral levels.

Negative feelings (e.g., worry, fear, anger, etc.) produce a lower vibrational rate that can not only keep an individual's spiritual vibration at a lower level, but can also attract lower level astral beings to that individual.

Soul Projection	The Eternal Soul is projected out of the body, attached to the Middle Dantian through a Silver Cord.
Spirit Projection	The Shen is projected outside of the body as an energetic vehicle of light.
Energy Projection	The Qi is projected out of the body like an energetic vapor.

Figure 13.22. Soul Projection, Spirit Projection and Energy Projection

SOUL PROJECTION

Soul Projection is the skill of sending the body's Eternal Soul (at will) to other places, times and dimensions (or worlds) in order to retrieve knowledge and experiences that assist the individual in his or her spiritual evolution. It is a nonphysical world in which time and space have no meaning or influence. The astral dimensions form a parallel universe that exists both within and outside ourselves.

The Eternal Soul is a fluid energetic form that can disengage from the physical body and travel about as a disembodied apparition taking the human consciousness along with it. It can pass through walls, travel to higher dimensional worlds, and encounter nonhuman creatures and mystical masters. Many Qigong masters use Soul Projection to seek spiritual knowledge, travel through time to view the past, explore other spiritual dimensions, engage in astral sex, or view the "infinite knowledge" (i.e., Messages) contained within the Wuji.

When the Eternal Soul begins to leave its energetic chamber, the brain wave patterns slow to a theta pulse. The body may feel paralyzed, and begin to vibrate as if an electrical current were passing through it.

There is also an odd roaring sound heard in the ears before the Eternal Soul separates from the body's tissues, or sometimes buzzing noises, musical notes, humming sounds, or chattering voices can be heard. When leaving the physical body, sometimes a swaying sensation can be felt as the astral body separates from its physical shell. The astral body can either spin out or float out of the physical body. When traveling in the astral body individuals may observe streaks of light flashing behind them as they travel.

As the Eternal Soul travels, a silver energetic cord is attached from either the top of the head (at the Baihui area) or at the Lower Dantian (at the umbilical area). This energetic cord is the media through which an individual retains his or her awareness of the physical self while projecting. This energetic cord is constructed of subconscious spiritual energy and maintains the individual's protective perceptual awareness of his or her physical body. It is believed by some that if this cord is severed the individual will die.

When the Eternal Soul (in its astral form) approaches someone, there is a feeling of electricity and coolness that is general felt. While in its phantom form, the Eternal Soul can display telekinetic powers (e.g., knocking on walls or tables, closing a half open door, etc.).

Warnings and Contraindication

Do not practice Soul Projection just before, or during, thunderstorms as the electromagnetic buildup in the atmosphere can hamper your ability to leave or return to the body. Also avoid approaching any kind of power-line during Soul Projection, there is a risk of getting caught by the power-line and becoming "stuck" (like a fly in a spider web).

1. A Body of Light	This energetic form resembles an energetic duplicate of the individual's physical self. It involves creating a near exact energy body duplicate of oneself (often with improvements, e.g., no wheelchair, or the elimination of other physical challenges).
2. An Animal Form	This energetic form resembles that of an animal (known as Shape Shifting). It is an ancient Daoist Shamanistic skill used to transcend the astral plane, and to take on the consciousness and energetic behavior, skills and perceptions of the animals the shamans assumed.
3. A Ghost or Saint	This energetic form resembles that of a Ghost or Saint (also known as an Invocation). It involves energetically taking on the image, form, and personality characteristics of a departed individual or spiritual deity.
4. The Immortal Observer	This energetic form resembles that of an All Seeing Eye (also known as The Astral Eye). It involves creating an energetic ball of light in the Upper Dantian at the Yin Tang (Third Eye) area. The large eye surrounded by a sphere of light is released from the Yin Tang area to gather psychic impressions via a long energetic cord attached to the Upper Dantian. As the eye returns, its energetic form is absorbed back into the individual's Upper Dantian. These gathered images are then analyzed and the information is assimilated before the energy is dispersed into the individual's Lower Dantian (and then down the legs and into the Earth).
5. The Element Body	This energetic form usually resembles that of different elements of nature, such as wind (as vaporous mist), Water, Earth (the terrain) or Fire. The ancient Daoist would use these elemental forms of nature as a disguise in order to obtain information or spy on hostile intruders.

Figure 13.23. Types of Spirit Bodies Used in Spirit Projection

Do not practice Soul Projection while recovering from an acute sickness (the drain on your energetic reserves can make you more susceptible to further illness).

SPIRIT PROJECTION

Spirit Projection is the skill of sending the body's Shen (at will) to other places, times and dimensions (or worlds) to retrieve knowledge and experiences to assist the individual in his or her spiritual evolution.

Spirit Projection (also known as mind projection, remote viewing, mind travel or traveling clairvoyance) can entail a combination of "genuine" observation with an extrasensory perceptional ability to gather information. When individuals perform traveling clairvoyance in the form of Spirit Projection, they can send their consciousness anywhere in the world. Awareness is not physical in nature, it is experiential, totally nonmaterial and cannot be limited to, or fully contained within the body.

Spirit Projection is practised through the creation of an energetic vehicle for the spirit body, such as a body of light, an energetic form of an animal, or a ball of light. An energetic shape is formed through the imagination and intention. The individual then transfers the energetic consciousness from his or her physical body into the energetic vehicle. This energetic form can resemble various patterns shapes and sizes to serve the individual's purpose (Figure 13.23).

As the individual thinks about a person or place, his or her body will be pulled to that object like a magnet, followed by a flood of images. The spirit body can fly over various terrains as it travels (e.g., oceans of water, deserts, mountains, forests, etc.). Once the spirit body arrives at its destination, the individual is able to feel and experience the area's environmental Qi (e.g., the humidity, wind, cold, heat, light, darkness, etc.).

After traveling in the spirit body the individual must reabsorb the energetic body back into his or her physical body to retain a clearer memory of the experiences.

Warnings and Contraindication

Do not perform too many Spirit Projections within a relative short time span. Excessive Spirit Projection has a tendency to weaken the Kidneys and drains the body's Wu Jing Shen.

While traveling, be careful of what your intentions are, as like begets like. Any negative projection, emotion or evil intention brought into the astral plane will quickly rebound back to the individual, or may attract negative entities that feed off of these lower vibrations.

ENCOUNTERING ASTRAL BEINGS

Within the subtle realm of the spiritual world, the existence of spiritual entities and ghosts are divided into Yin and Yang. While traveling in the astral realm, an individual can encounter either type of spiritual entity.

1. The Yang Beings native to the higher astral realm vibrate at a faster energetic rate, and are associated with more advanced, highly evolved beings, such as angels, guides, beings of the light, and enlightened masters.
2. The Yin Beings native to the lower astral realm vibrate at a slower energetic rate, and are associated with lower, less evolved beings, such as devils, demons, spirits of darkness, evil spirits, and other similar negative spiritual manifestations.

In all truth, the astral body cannot be hurt the same way that the physical body can. The astral body can, however, carry back to the physical body mental, emotional, and spiritual scars, or hitchhiking negative astral beings, all of which can negatively affect the physical body.

All actions and reactions on the astral plane, whether productive or destructive, are a product of the individual's belief system (you bring about each condition according to your faith).

ENERGY PROJECTION

Another type of life-force energy that the body can project outside its physical structure is Qi. Qi Projection is emitted outside the body like a fine vaporous mist, enveloping and sensing everything that it touches. Similar to Soul Projection and Spirit Projection, Qi Projection is directed according to the individual's intention, imagination, will, and faith. Distance therapy, pertaining to Qi emission, can be observed as energy leaving the doctor's hand like heat waves emanating towards an external object. Within this emitted life-force energy is contained the doctor's Shen (Spirit), as well as the energetically encoded "Messages" that the doctor sends to his or her patient. This constitutes the energetic foundation for Medical Qigong therapy.

SECTION III

DAO YIN TRAINING

CHAPTER 14

INTRODUCTION TO DAO YIN TRAINING

Qigong developed into a systematic healing art for health preservation during the "Warring States" period in Chinese history (476–221 B.C.). Records from that time contain documentation of symptoms, treatments, and specific points of attention given within the body and energy channels, by doctors practicing Medical Qigong, as well as details of the effects of these Qigong exercises.

The Huang Di Nei Jing (*The Yellow Emperor's Classic of Internal Medicine*) states, "to treat diseases it is important to use both Dao Yin and An Qiao." These are two names given to Medical Qigong therapy in ancient times.

"Dao" refers to physical movements, which are guided by the strength of the mind to stimulate the body's Qi and bring about a state of relaxation. "Yin" refers to the fact that with the aid of physical movements the body's Qi can reach the extremities (feet, fingers, and head). Dao and Yin together mean to lead and guide the energy through both motion and quiescence. This therapy includes mental, respiratory, and postural exercises.

"An" has two meanings, "massage" and the "up and down movement of the wrists." "Qiao" means "to lift up." An Qiao refers to the methods of the postural Dao Yin exercises. The basic hand and body methods of postural Dao Yin training are as follows:

- Kai (Opening)
- Fen (Separating)
- He (Closing)
- Tui (Pushing)
- Rou (Rubbing)
- Xuan (Rotating)
- An (Pressing)

These are the foundational movements a Qigong doctor will use in order to purge, disperse, gather, tonify, emit and guide Qi. The training of energy is achieved and implemented primarily through a combination of dynamic (active) and static (quiescent) exercises and meditations. Although Medical Qigong schools are diverse and complex in their form and content, they can all be divided into two main branches of energetic cultivation: Static Qigong and Dynamic Qigong (Figure 14.1).

- **Static Qigong Training**
  - 1.Dispersing into the Wuji
  - 2.Breath Regulation
  - 3.Mind Concentration
  - 4.Imagination and Observation
  - 5.Static Relaxation Qigong
  - 6.Inner Nourishing and Strengthening Qigong
- **Dynamic Qigong Training**
  - 1.Guiding and Inducing
  - 2.Massage
  - 3.Tapping and Striking
  - 4.Self-inducing
  - 5.Hard Qigong

Figure 14.1. Two Main Branches of Energetic Cultivation: Static Qigong and Dynamic Qigong

STATIC QIGONG TRAINING

This training involves bringing about and keeping a quiescent state of mind and breath. In fixed postures (lying, sitting, or standing) the practitioners can train to accumulate substantial Qi in their Lower Dantian, while circulating it through their bodies. Static Qigong training is further divided into six different schools of practice: Dispersing into the Wuji, Breath Regulation, Mind Concentration, Imagination and Observation, Static Relaxation Qigong, and Inner Nourishing and Strengthening Qigong.

1. Dispersing into the Wuji requires the practitioner to dissolve his or her energy into infinite space, or the Void, without focusing or directing their attention on any object from the external environment.
2. Breath Regulation requires moving Qi internally or externally while inhaling, exhaling, or swallowing the breath, etc.
3. Mind Concentration requires concentrating the mind on one fixed internal or external point.
4. Imagination and Observation require disciplining the imagination with vivid, colorful visualizations.
5. Static Relaxation Qigong requires the relaxation of both body and mind, keeping them in a constant state of quiescence.
6. Inner Nourishing and Strengthening Qigong require the use of multiple Static Qigong modalities (for example; Breath Regulation combined with Imagination and Observation).

DYNAMIC QIGONG TRAINING

This involves training the body, mind and Qi simultaneously, to promote the functional activities of Qi and to enhance its free flow throughout the body's channels. These exercises often require standing or walking, and affect muscular tension, weight distribution, Blood and Qi circulation, as well as the functional activities of the major organs. These exercises are differentiated by various kinds of body movements and can be divided into five different schools of training: Guiding and Inducing, Massage, Tapping and Striking, Self-inducing, and Hard Qigong.

1. Guiding and Inducing requires physical movements such as dancing, mimicking animal movements, and gymnastics.
2. Massage focuses on both the manipulation of specific points and channels, as well as on the knowledge of muscles, tendons, and bone structure.
3. Tapping and Striking focus on using patting, slapping, tapping, and striking techniques with the fingers, palm, fist, wooden stick, or material in a cloth bag.
4. Self-inducing training encompasses small to large physically strenuous Qigong movements.
5. Hard Qigong focuses on martial arts training in Iron Palm Training, Iron Body Training, etc.

The incorrect employment of dynamic or static principles of Qigong exercises and meditations can cause energetic deviations which are capable of either hindering, or harming, the individual's health. There are two purposes for Medical Qigong development: To Protect and Strengthen the Body's Health, and To Treat Diseases.

1. To protect and strengthen the body's health: Medical Qigong exercises are prescribed in accordance with the individual's physical constitution, state of health, emotional constitution, climate, and different seasons.
2. To treat diseases: Medical Qigong exercises are prescribed in accordance with the individual's symptoms, pathological condition, and degree of severity, complications, and the constitution of the patient.

PRINCIPLES OF YIN AND YANG IN DAO YIN TRAINING

Both static and dynamic Qigong methods supplement each other when combined equally. No matter which static or dynamic method is practiced, one of the four principles of Yin and Yang will always be adhered to: Quiescence in Motion, Action within Motion, Motion in Quiescence, and Stillness within Quiescence.

1. Quiescence in Motion is referred to as a Yin within Yang technique, and requires being quiet, calm, and still on the inside while being active on the outside.

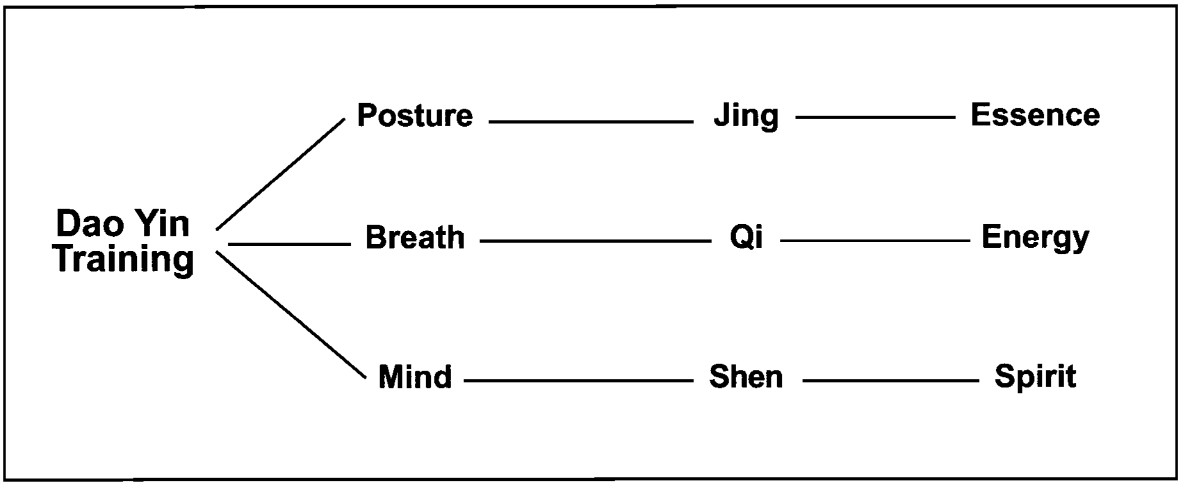

Figure 14.2. Dao Yin Chart

2. Action within Motion is referred to as a Yang within Yang technique, and requires being active both on the inside and on the outside.
3. Motion in Quiescence is referred to as a Yang within Yin technique, and requires being active on the inside while being quiet, calm, and still on the outside.
4. Stillness within Quiescence is referred to as a Yin within Yin technique, and requires being quiet, calm, and still both on the inside and on the outside.

THREE FOUNDATIONAL METHODS OF DAO YIN TRAINING

The diverse methods of practicing Medical Qigong exercises fall within three categories of Dao Yin training: Posture Dao Yin training, Respiratory Dao Yin training, and Mental Dao Yin training.

1. Posture Dao Yin consists of posture training for Medical Qigong therapy. It is divided into six specific types of training: walking, standing, sitting, kneeling, lying, and massaging.
2. Respiratory Dao Yin consists of breathing exercises for Qigong training, and therapy that includes various techniques of inhaling, exhaling, aspirating, blowing, and holding the breath.
3. Mental Dao Yin training requires that the thoughts be concentrated on one object, placing the cerebral cortex into a special inhibitory state. In various Medical Qigong circles in China this method is commonly referred to as "the internal keeping of the mental state."

The Dao Yins are taught in a postnatal energetic progression of Jing (posture) to Qi (respiration), to Shen (mind and emotion). In this way, the body regulates and supports the production of energy, which in turn regulates and supports the production of Spirit (Figure 14.2).

Despite the fact that many postures and movements may vary according to the specific styles of each Medical Qigong school, the primary benefit of each exercise will:

- improve and strengthen the patient's metabolism,
- smooth the Qi of the channels and collaterals,
- strengthen the body and mind,
- emotionally detoxify the internal viscera, and
- facilitate the increased flow of Righteous Qi (the Qi that fights the pathogenic invasion of Evil Qi) within the body.

CHAPTER 15
POSTURAL DAO YIN TRAINING

The goal of Postural Dao Yin training is to promote the regulation and circulation of Qi. Part of this training involves learning the techniques to guide the internal energy to flow using the movements and changes in the electromagnetic fields of the hands and body.

Success in Medical Qigong training is directly related to the correctness of the body's alignment. Correct posture regulates the Heart, and results in a calm mind; with a relaxed mind it is easy to lead the flow of Qi downward, to affect the Blood circulation and pulse. When the body is incorrectly aligned, it disturbs the smooth flow of Qi; energy then becomes obstructed or deviates, resulting in a restless Mind (both the thinking mind and the whole body consciousness).

The proper sequence of relaxation and energy circulation within the body is as follows:

1. the muscles are first to relax,
2. followed by the tendons and ligaments,
3. next, the nerves relax, and
4. finally the bones.

Once the muscles, tendons, and ligaments relax, the Qi can circulate freely through the channels and collaterals. The skin and muscles are generally the first level of awareness cultivation. The sensations of Qi and Blood circulation are usually experienced by the practitioner as heat, tingling, fullness, and expansion. The Chinese call these sensations the Eight Energetic Touches, which are known as the first of three transformations of energy (also known as the Subtle Wonders, see Chapter 29).

The next level of awareness involves the deeper flows of Qi within the nerves; this feels like an electric vibration, or an energetic current traveling through the body. Sometimes, due to an energetic surge released from within the body's organ system or channels, an electric shock is felt through the extremities or through the entire body.

The final level of Qi circulation awareness involves the Qi circulation within the bones and the Marrow. In the beginning stages it sometimes feels cold and arthritic in the joints (especially within the extremities). These sensations will later develop into a vibrational resonance which is felt deep within the bones.

The main point of Postural Dao Yin is to seek a state of physical and energetic balance. Postural Dao Yin is also divided into two distinct methods of treatment: Static Posture and Dynamic Posture training.

STATIC POSTURE DAO YIN TRAINING

This method includes quiet movement while in the position of lying, sitting, standing, or walking, and is used to accumulate and strengthen the practitioner's internal Qi. Before performing the static posture forms for Medical Qigong therapy, it is important to note that static posture training is divided into three basic modalities: Choosing the Proper Beginning Posture, the Inhaling and Exhaling Methods, and Opening and Closing the Lower Dantian.

1. The purpose of the choosing the proper beginning posture is to calm the Heart and relax the mind, and focus on a specific goal for training the body's life-force energy.
2. When the energy begins to flow freely, the practitioner focuses his or her attention on each inhalation and exhalation for the specific goal of tonification, purgation or regulating the life-force energy.
3. The purpose of the "opening and closing" of the Lower Dantian ending postures is to bring the body's energy back down into the Lower Dantian and to root the Qi. Rooting the Qi in the Lower Dantian after Qigong practice

avoids creating Excess conditions within the body's internal organs.

CHOOSING THE PROPER POSTURE

Medical Qigong meditations and prescriptions are not limited to sitting practice alone. Each of the different lying, sitting, standing, and walking postures has its own specific form of meditation.

The lying, sitting, and standing postures create the foundational base for Static Posture Dao Yin training. Each static posture can induce the body's energy to descend quickly into the Lower Dantian. The selection of the appropriate technique and body position is determined by each practitioner's body type (structural formation) and the condition of the individual's physical health. The length of time that a practitioner focuses on the Medical Qigong prescription is also determined by the individual's specific disease as shown in the following examples.

- Patients with digestive problems, gastroptosis, or prolapse of the internal organs should begin their training while lying supine (on their back), with their knees pointing upward, their buttocks elevated 4 inches off the ground, and their feet flat on the floor.
- Patients with gastrointestinal diseases should begin their training lying flat on their backs. As the condition improves, they should combine lying and sitting meditations to promote the digestion and absorption of food.
- Patients with Heart disease, hypertension, or nervous disorders should begin training from a sitting position. Later, they should combine sitting and standing meditations in order to direct their Qi and Blood downward.
- Patients with Lung diseases, asthma, chronic bronchitis, emphysema, or cardiopulmonary function disorders should begin their training from a semi-reclining position (usually a 45 degree angle) to help respiration flow smoothly and to alleviate stress on the Lungs.

  In the hospital, each pillow that the patient uses for elevation represents the degree of strain that is placed on the Lungs. The patient is said to have strong Lungs, for example, if he or she is reclining and breathing comfortably on one pillow. If however, it takes three pillows before the patient is reclining comfortably, the Lungs are considered weak and congested (e.g., in cases of dyspnea).
- Patients with severe physical weakness (e.g. chronic fatigue syndrome and fibromyalgia) should begin their training from a lying or sitting position. Later, they can combine sitting and standing meditations as their constitution improves.

PRIMARY AND SECONDARY POSTURES

Patients are taught a primary and a secondary posture to prevent fatigue, and increase the healing effect of the Medical Qigong prescriptions. A primary standing posture is performed, for example, in the morning when the patient is strongest, and the energy is cleanest. A secondary sitting posture is performed in the afternoon, or evening when the patient is tired, with a focus on quiet regulation.

After a patient has learned a particular lying, sitting, or standing meditation, he or she can be taught the secondary posture without the risk of depleting his or her Qi.

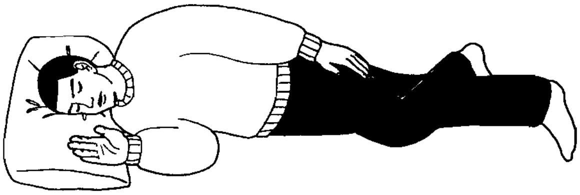

Figure 15.1. Sleeping Dog Posture

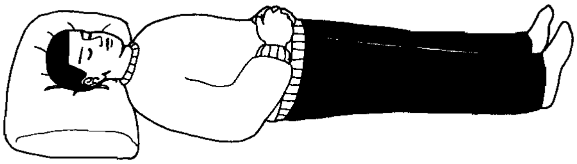

Figure 15.2. Supine Posture (Hibernation Posture)

LYING POSTURES

The Chinese saying, "The body should lie like a bow," is a perfect description of the curled position of the patient's body, while in the Sleeping Dog posture (Figure 15.1). This particular position is beneficial for aiding digestion and the absorption of food.

In prescribing the Sleeping Dog posture, the patient should lie on their right side with the left knee bent and the right leg extended straight. To avoid compressing the Heart, do not allow the patient to lie on their left side. Keep the patient's head raised slightly to allow the Blood to flow downward into the torso. Have the patient bend the right arm at the elbow.

When the patient is chronically ill, however, an exception is made. In this case it is important that the patient lie on the "healthy side" and not the side of the diseased organ (e.g., in cases of Liver disease, patients might lie on their left side). Remember to always adjust the patients' postures as needed and encourage them to practice "natural breathing."

Another popular posture is called the Hibernation posture. By lying supine in the Hibernation posture, the patient will regulate the functional activities of the central nervous system (Figure 15.2). Lying supine is also suitable for patients with serious diseases who are debilitated (this is also a sleep inducing method of training).

SITTING POSTURES

The Chinese saying, "The body should sit like a bell," describes the secure (or rooted) position of the patients' body, especially the lower legs and torso. Sitting is the most commonly used position for Medical Qigong therapy. This position helps the patients to relax and focus their mind's intent.

There are several sitting postures used in Medical Qigong therapy, and are prescribed according to the patient's flexibility: sitting upright on a chair (Figure 15.3), a cross-legged posture called the Full Lotus, another called Half Lotus, or with the legs loosely crossed (Figure 15.4-6). Make sure the patients sit on a raised cushion in order to keep their buttocks slightly higher than their knees.

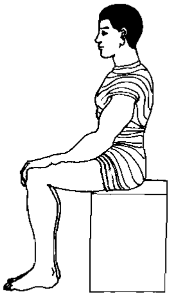

Figure 15.3. Upright Sitting Posture

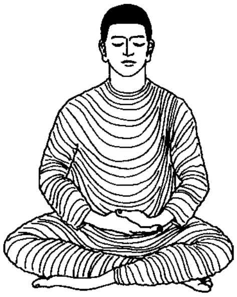

Figure 15.4. Cross-Legged Zi Wu Posture

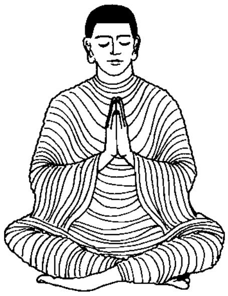

Figure 15.5. Cross-Legged Buddhist Greeting Posture

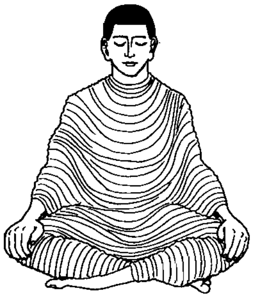

Figure 15.6. Cross-Legged Buddha's Warrior Attendant Posture

One of my Medical Qigong instructors in China suggested elevating the hips at least four inches off the floor when meditating, to avoid undue stress on the lower back, spine, and leg circulation. This elevation of the hips forms a triangle between both knees and the Lower Dantian, establishing a firmer connection to the Earth energy. This allows the Earth Qi to easily flow from the patients' knees and coccyx into their body.

When sitting in a chair meditating, the feet should touch the ground firmly, instead of being suspended in midair. This allows the Earth Qi to flow easily from the patients' feet into their body. Otherwise patients may develop swollen feet and lumbago due to Blood stagnation, or even hallucinations.

Keep the patients' spine straight by aligning the tip of the nose on a vertical line with the navel, to release pressure and tension on the cervical nerves.

STANDING POSTURES

The Chinese saying, "The body must stand like a pine tree," describes a relaxed position of the body's structure, standing with the bones stacked on top of each other, from the ground up. Standing meditations are particularly tonifying in nature, and are absolutely essential for accumulating Qi. The advantage of the standing position is that the flow of energy can be easily regulated and promoted. This posture helps to normalize blood pressure, and calm the Heart. Standing meditations also produce a unique psychophysiological change within the patients, allowing them to shed energetic obstructions, and to create a healthier body, mind, and spiritual awareness. Standing meditations are also essential for the Qigong doctors, as it facilitates the accumulation of Qi that is necessary for the emitting of energy.

Standing meditations, however, are not suitable for patients who are extremely ill or in poor health. Therefore, exercises in the standing postures should only be prescribed for minimal durations during convalescence, to build up the patients' constitution, promote health and to enhance their strength slowly.

The body should be held in an upright posi-

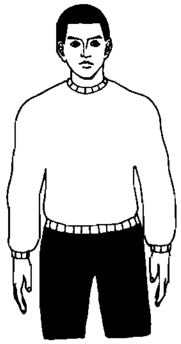
Figure 15.7. Wuji Posture

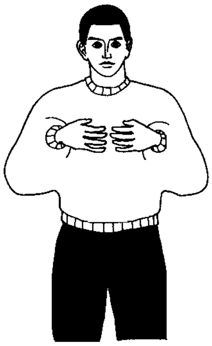
Figure 15.8. Immortal Post Posture

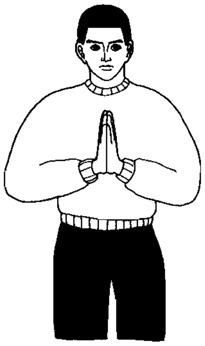
Figure 15.9. Buddhist Greeting Posture

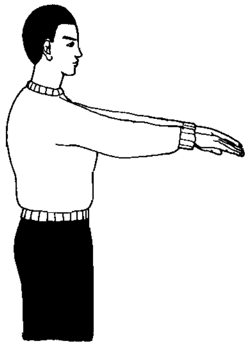
Figure 15.10. Extending Palm Posture

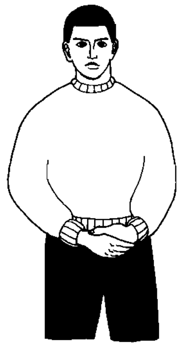
Figure 15.11. Embracing Dantian Posture

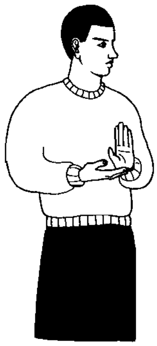
Figure 15.12. Square-Round Posture

tion, with three points connected in a straight line: Baihui GV-20 point at the top of the head, Huiyin CV-1 point at the bottom of the perineum, and the Yongquan K-1 point at the bottom of the feet. The patient's posture should be kept relaxed, still, solid, and rooted like a tree. The patient should also imagine their feet growing roots deep into the ground, to absorb the Earth's energy from the soil.

Standing exercises are usually accompanied by a variety of hand positions (Figures 15.7–15.12). The most popular are (1) Wuji: hands hang down by sides, (2) Immortal Post: hands embrace an invisible ball, (3) Buddhist Greeting: both hands press together at the solar plexus level, (4) Extending Palm: both hands press downward, (5) Embracing the Dantian: folding the hands and resting them on lower Dantian, and (6) Square-Round: left palm's heel rests inside the right palm's center.

In each posture, whether lying, sitting, or standing, the function and flow of body energy shifts and changes in relationship to, and as a manifestation of, the patient's mental belief structure.

If, for example, the patient while standing, shifts his or her weight and mental focus onto the balls of the feet, the energy will naturally flow through the front of the body via the front Yin channels. If the patient shifts his or her weight and mental focus onto the back on the heels, the energy will naturally flow through the back of his or her body via the back Yang channels. The center of the feet connect the patient's flow of energy to the Taiji Pole and can be used either to transfer the energy forward or backwards.

Generally, before beginning standing meditations, patients are encouraged to relax, sink, and shift their body weight back and forth between their heels and their toes. This practice massages their internal organs and disperses any blockages by stimulating the corresponding organ areas at the bottom of the feet as described in the therapeutic use of the Lower Hexagram Yaos (see Chapter 3).

Patients should inhale while shifting their weight backwards and exhale while shifting forwards. Their attention should be placed on their internal organs and the body's surrounding field of energy. The patients' awareness is also focused on the body moving back and forth, like the rhythm of an ocean tide.

Women are generally encouraged to practice meditation in the standing position (especially around the time of their menstruation), as the sitting position tends to decrease energy circulation, and may cause energy blockages in their lower abdominal area.

THE EIGHTEEN RULES OF PROPER MEDICAL QIGONG STRUCTURE

The most important basic techniques of Medical Qigong training are guided by eighteen rules of proper form and structure. The main point of Postural Dao Yin training is to relax and seek quiescence while in the various postures. Tension in any area of the body restricts the whole structural system, since the body seeks to balance its structure naturally by shifting its energy and weight. The general function of the muscles is to guide the flow of energy through the channels.

The following is a list of the Eighteen Rules for proper standing postures, these rules apply to every school of energy cultivation:

1. Stand with the feet flat,
2. Bend the knees,
3. Relax the hips,
4. Round the perineal area,
5. Close the anal sphincter,
6. Pull in the Stomach,
7. Relax the waist,
8. Tuck the chest in,
9. Stretch the back,
10. Relax the shoulders,
11. Sink the elbows,
12. Hollow the armpits,
13. Relax the wrists,
14. Suspend the head,
15. Tuck the chin,
16. Close the eyes for inner vision,
17. Close the mouth and turn your hearing inwards, and
18. Touch the tongue to the upper palate,

Each of these eighteen rules is described in detail as follows.

RULE 1—STAND WITH THE FEET FLAT

Stand with the feet flat, parallel to the outside of the shoulders. Weight should be distributed evenly on both feet about an inch and a half inward (anterior) from the heels, and perpendicular to the tibia. Focus on dropping the weight down behind the legs, not down the front. This action takes the pressure away from the knees and places it in the center of the feet (Figure 15.13).

Stretch the feet and let the toes softly grasp the ground to keep the body firmly rooted as you tuck the sacrum under (Figure 15.14). Relaxing is important. Rigid, contracted feet disrupt the flow of energy from the Earth into the body. The Kidney energy flowing within the bone can be accessed by shifting the weight onto the heels. The Heart energy can be accessed by shifting the weight onto the balls of the feet. The weight in the center of the feet accesses the Taiji Pole.

While training, the feet may vibrate or feel hot, as if on fire. This is a normal reaction to correct postural training and is beneficial because it dissolves the calcium deposits stored within the extremities of the feet. If, however, the Fire in the feet is too intense (due to Excess Liver Fire), hit the heels on the ground and rub the lumbar spine, using the mind's intention to send more of the Kidney Yin (Water) down to the feet to cool the Fire.

RULE 2—BEND THE KNEES

The knees should be slightly bent and facing the same direction as the feet. The knee caps should be directly aligned and extended no further than the tips of the middle toes. Do not allow the knees to twist inward.

To secure the structure, imagine holding a large beach ball between the knees. Many patients develop knee problems because they extend their knees too far forward or allow their knees to twist to the side, rather than aligning them with the toes. If pressure is felt on the sides or front of the knees, check the positioning. Knees that are unable to point in the direction of the toes are most often the result of tight hip joints (Figure 15.15). If the energy in the legs is tightly constricted around the knee area, Qi flow will be severely limited. A relaxed knee-joint, on the other hand, will increase

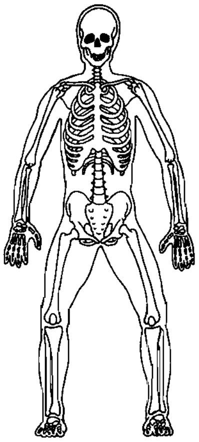

Figure 15.13. Stack the bones on top of each other from the bottom of the feet to the skull.

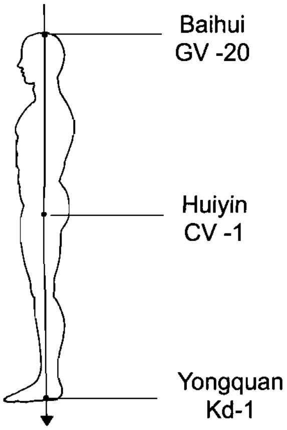

Figure 15.14. The three points (Baihui, Huiyin, and Yong quan) can be connected by a straight line.

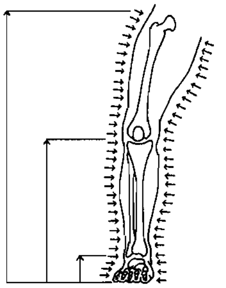

Figure 15.15. Keep the center of the body's weight aligned to the center of the knees, to relax the connective tissue surrounding the knees.

the Qi and Blood flowing through the legs, making it easier to relax the hip and to round the perineal area.

Relaxing the knees also permits free flow of energy through the three descending Yang channels on the outside of the leg, and the three ascending Yin channels on the inside of the leg (Figure 15.16). The techniques for relaxing the knees will vary according to the emotional and structural pattern of each individual, as emotions of fear, etc., have a tendency to gather and pool behind the knees.

RULE 3—RELAX THE HIPS

Drop the buttocks slightly while sinking the Qi from the upper torso into the Lower Dantian, and relax the hips. Keep the hips facing forward. The buttocks should be gently tucked under the body to facilitate straightening of the spine and the keeping the Qi in the Lower Dantian.

Imagine a weight of a thousand pounds hanging from the coccyx (Figure 15.17). As the weight pulls down, tilt the sacrum under the body. When the hip-joints are relaxed, the lower limbs will be able to move freely.

RULE 4—ROUND THE PERINEAL AREA

There are three parts to rounding the perineal area. First, set the knees apart, turning them both outward. This allows the groin and hips to shift to an anterior tilt. Second, bring the knees back together and turn them inward, while also relaxing the hips. This allows the groin and hips to shift to a posterior tilt. Third, suspend the perineal area to lift the perineum and the anus slightly. This allows the groin and hips to come to a center balance and to sink. Rounding the perineal area keeps the perineum free from pressure, yet sealed. This ensures better functional activities of the Qi, and allows the "lower door" (the anus) to remain closed to avoid leakage of vital energy.

RULE 5—CLOSE THE ANAL SPHINCTER

The anus is called the Lower Bridge and it is where the Yang and Yin channels meet and combine. The anus is considered the door of the Seven Corporeal Souls (Po). Although the anal sphincter is held closed, it is important that the perineum remain relaxed. There are three stages in controlling the anal sphincter:

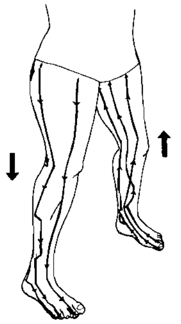

Figure 15.16. Relaxing the knees also permits the free flow of energy through the ascending Yin channels and descending Yang channels of the legs.

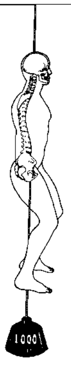

Figure 15.17. Imagine a 1000-pound weight hanging from the coccyx and a heavenly cord suspending the head.

1. In the first stage, the anus is contracted with slight muscle control, and the energy is directed into different energy centers along the Microcosmic Orbit. When first beginning, inhale and expand the abdomen while imagining that the anus is connected to the navel and is pulling up towards it. On the exhalation, contract the abdomen and draw the energy from the navel back down towards the anus. Next, inhale and create a connection between the anus and the Mingmen. On the exhalation, draw the energy from the Mingmen back down towards the anus. It is important to synchronize the drawing up of the anus towards the navel and Mingmen area with the expansion and contraction of inhalation and exhalation. This will facilitate a

sacral-cranial rhythm throughout the body.

2. In the next stage of training, the constriction of the anus is used to lead the energy up the Governing Vessel in the Microcosmic Orbit (Figure 15.18). Upon inhalation, draw the mind's intention from the navel towards the anus and then down to the tip of the spine (the coccyx), allowing the Qi to flow up the spinal column towards the Mingmen, past the Shendao to the base of the skull. Upon exhalation, roll the eyes back into the head and use the intention to meet the energy at the base of the skull. Next, lead the energy over the Baihui at the top of the head and down the Conception Vessel and back down to the naval.
3. At the third stage, imagine twisting the energy of the anal sphincter, spiraling the Qi up the center of the body, to connect with the Baihui area at the top of the head. Close the anal muscle gently, inhale and imagine the energy vibrating and spiraling in a counterclockwise direction up the body's Taiji Pole. This connects the Lower, Middle, and Upper Dantians through the Taiji Pole. Upon exhalation, lead the vibrating Qi down the Taiji Pole in a clockwise spiral ending within the Lower Dantian area. Continue this meditation until the entire body pulses with the rhythm of the Taiji Pole.

The anus is divided into five regions: Front, Middle, Back, Left, and Right. By contracting the anus in different parts, you can bring more Qi to the various organs and glands.

1. The Front region opens the gate to the sexual organs, as well as to the Conception Vessel when it is contracted.
   - For Men, this will affect the prostate gland, Bladder, Small Intestine, thymus gland, thyroid gland, parathyroid gland, mouth, nose and front part of the brain (Figure 15.19).
   - For Women, this will affect the vagina, uterus, vena cava, aorta, Stomach, thymus gland, thyroid gland, parathyroid gland, mouth, nose, pituitary gland, and front part of the brain (Figure 15.20).
2. By contracting and pulling the middle region of the anus, more Qi is drawn into the Thrusting Vessel and Taiji Pole, as well as to the genitals, aorta, vena cava, Stomach, Heart, thyroid gland, parathyroid gland, tongue, pituitary gland, pineal gland, and Baihui area (Figure 15.21).

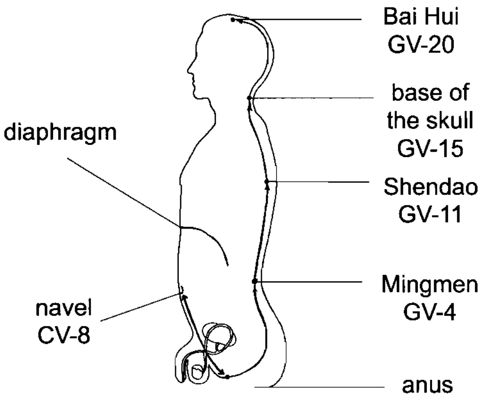

Figure 15.18. The anus is used to draw the Qi from the navel into the Lower Dantian and lead the Qi up the Governing Vessel.

3. By contracting the back region of the anus the Qi is connected to the Governing Vessel, as well as the sacrum, lumbar, thoracic and cervical vertebrae, and the cerebellum (Figure 15.22).
4. By contracting the left side of the anus the Qi connects to the left side of the body.
   - For Men, this affects the left testicle, Large Intestine, left Kidney, left adrenal gland, Spleen, the left side of the Heart, left Lung, left ear, left eye, and left hemisphere of the brain (Figure 15.23).
   - For Women, this affects both ovaries, both Kidneys and both adrenal glands, Spleen, Liver, Heart, both Lungs, thyroid gland, parathyroid gland, and both hemispheres of the brain (Figure 15.24).
5. The Right Side. By contracting the right side of the anus the Qi connects to the right side of the body.
   - For Men, this affects the right testicle, Large Intestine, right Kidney, right adrenal gland, Liver, Gall Bladder, right side of the Heart, right Lung, right ear, right eye, and the right hemisphere of the brain (Figure 15.25).

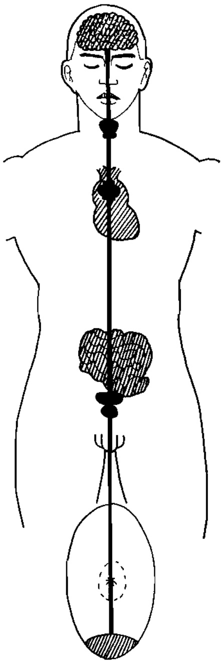

Figure 15.19. The Front of the Anus

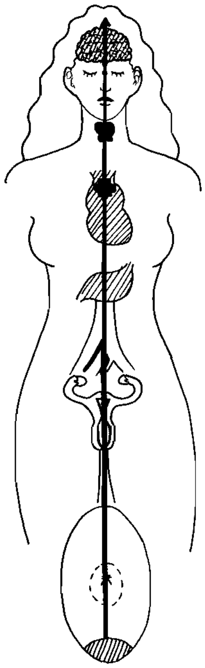

Figure 15.20. The Front of the Anus

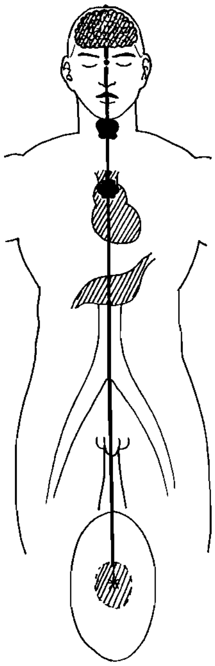

Figure 15.21. The Middle of the Anus

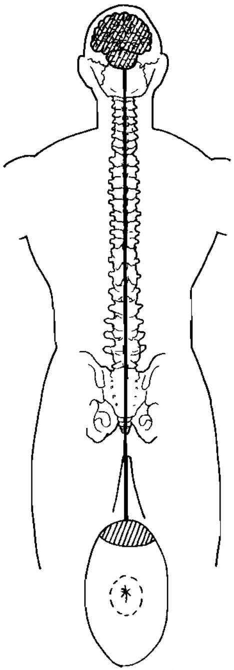

Figure 15.22. The Back of the Anus

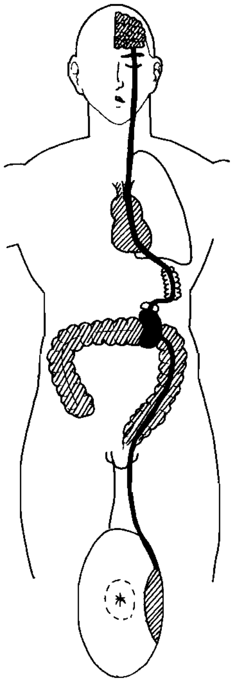

Figure 15.23. The Left Side of the Anus

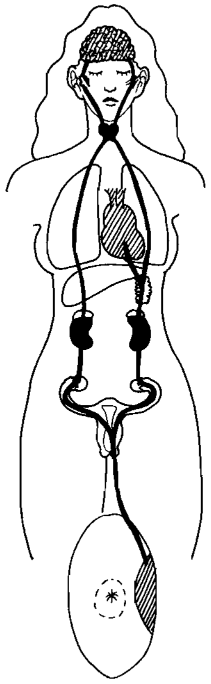

Figure 15.24. The Left Side of the Anus

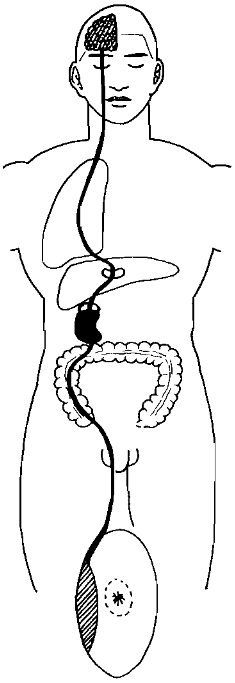

Figure 15.25. The Right Side of the Anus

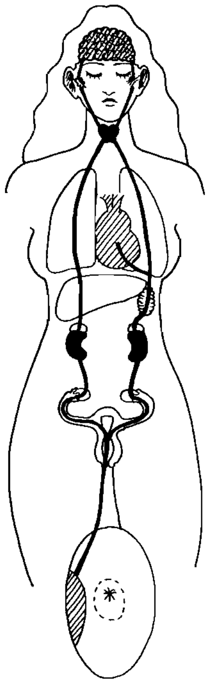

Figure 15.26. The Right Side of the Anus

- For Women, this affects both ovaries, both Kidneys and both adrenal glands, Spleen, Liver, Heart, both Lungs, thyroid gland, parathyroid gland, and both hemispheres of the brain (Figure 15.26).

RULE 6—PULL IN THE STOMACH

Pull the stomach in above the pubic bone, and avoid tensing of the abdominal muscles. This helps to restrain the Yuan Qi, thereby increasing the internal pressure and promoting the flow of internal Qi throughout the whole body.

RULE 7—RELAX THE WAIST

If the waist is not relaxed the Qi can not sink back to the Lower Dantian. Relax the waist by relaxing the hips and bending the knees. When the waist is relaxed, the spine will stand upright, yet remain relaxed.

The waist and the abdomen are two important areas in training and guiding Qi. The abdomen is considered to be the furnace for refining Qi. The waist is the residence of the Kidneys and Mingmen Fire and is an important junction for Qi and Blood circulation.

The waist is considered the master of the body and must be kept sunken and relaxed (Figure 15.27). As the buttocks tuck under and the waist sinks, small pockets should form on the inner thighs where the pelvic bones meet the femur bones of the legs. Most people usually keep the pelvis area tight and immobile, as it is associated with shock, fear, and sexual trauma (when this area becomes loosened, such emotions are often released).

The specific actions to relax the waist are as follows: lift both shoulders in order to stretch the back, then release them downward immediately as you exhale, relaxing the waist. To achieve genuine waist relaxation, you must also relax and tuck the sacrum, this action opens the sacrum. When the sacrum is opened, the Lower Dantian will expand and the storehouse of Yuan Qi will be enlarged. Be sure that the waist and hips face forward in the same direction as the knees and feet.

RULE 8—TUCK THE CHEST IN

The chest should relax inward so that air flows freely and the Qi sinks down into the Lower Dantian. Fill the abdomen with air while relaxing the chest inward. The back should be open and expanded. Relax the scapulae and allow them to sink to their lowest point, this will help the shoulders to sink. Rounding the scapulae (pushing to the sides and forward) will help open and expand the back.

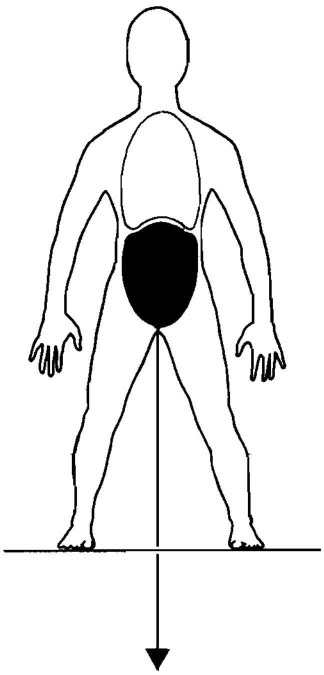

Figure 15.27. Relax and sink the waist, allowing the hips to drop.

Do not thrust the shoulders inward, nor intentionally close the thoracic cavity. The intentional tucking in of the vertebral area requires the outward opening of the ribs and expansion of the intercostal muscles. The action of opening the chest outward, then tucking the chest inward, helps to relax the thoracic cavity and avoids tension in the thoracic muscles.

RULE 9—STRETCH THE BACK

The key point to stretching the back is to hang the shoulders and "open the scapulae." When the two shoulders drop evenly, the lumbar vertebrae straighten out. The function of tucking the chest and stretching the back calms the Heart and Lungs.

The Heart controls mental and emotional activities, and also directs Blood circulation. When the thoracic cavity is spacious and the back is upright, the Heart and mind will be calm, and the Lungs will be clear and free from Heat. Tucking the chest and stretching the back allows the chest and back to relax, permitting the Conception and Governing Vessels to communicate with each other and improve energy movement.

Imagine a weight hanging from the coccyx and a rope pulling up on the crown point of the head. Feel the elongation of the spine with each inhalation and exhalation. Remain relaxed (Figure 15.28), do not hold yourself unnaturally erect and avoid contracting the abdomen to straighten the lower back. The abdomen must remain relaxed. When sinking the body's weight, try to feel the spine press into the heels of the feet.

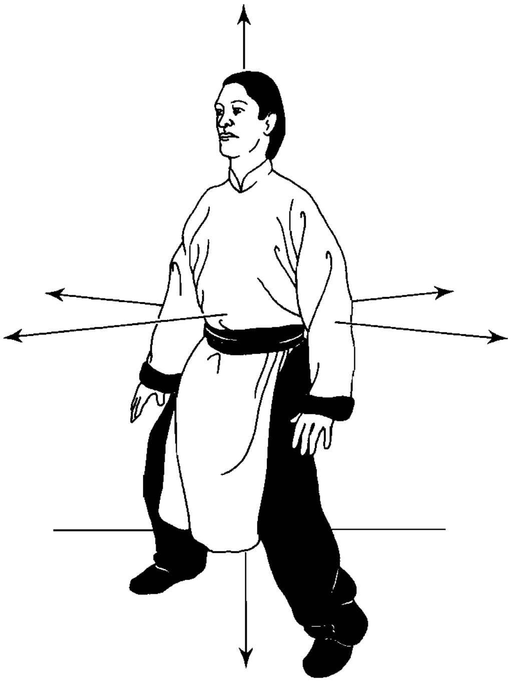

Figure 15.28. By imagining a rope pulling on the head and coccyx, the spine is allowed to naturally stretch, expanding the body's Qi in all directions.

RULE 10—RELAX THE SHOULDERS

If you relax the shoulder joints and let them hang naturally, the shoulders and neck will relax. If you elevate your shoulders when you are standing or sitting, the energy will not sink and you will begin to feel tired, the shoulders will be stressed, and the flow of Blood and Qi will become blocked.

When the shoulders are encouraged to relax and are allowed to drop and roll forward, this action facilitates the hollowing of the chest and the expanding and opening of the back, allowing the energy to flow into the arms (Figure 15.29). Try to feel your arms link up with the scapulae as the clavicles press downward. Your shoulders should remain very loose and free.

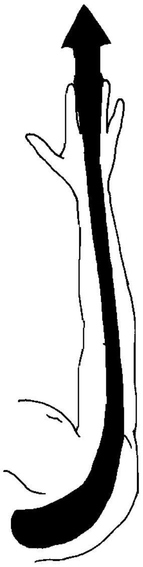

Figure 15.29. When the shoulder (here viewed from above) relaxes, the Qi flow into the arm increases.

RULE 11—SINK THE ELBOWS

When practicing the Immortal Post posture (refer back to Figure 15.8), and the hands are positioned by the Heart, it is important to bend the elbows slightly at the elbow joint, holding them lower than the hands.

While in a Wuji posture, imagine that a weight is hanging from the elbows, or a sensation of compression will be felt at the shoulder and elbow joints. The elbows should be slightly bent and relaxed, remaining structurally correct and not weak. This will help to roll the shoulders forward and open up the back. If the elbows are not bent, there will be decreased energy flow (Figure 15.30).

RULE 12—HOLLOW THE ARMPITS

Hollow the armpits as if each one is holding an egg, this allows the Qi and Blood to flow freely in the upper limbs. To relax the shoulders, move the tips of the elbows outward to the left and right. Avoid sloping or squaring the shoulders.

When practicing the Immortal Post Posture (refer back to Figure 15.8), the arms must embrace inward as if holding a large ball or barrel. Do not push the elbows out so far that you loose the curve in your arm, this is of extreme importance. Your arms should not bend inward more than 45 degrees. The energy flowing in your arms is like water running through a hose. If you bend it too sharply, much of the supply will be cut off.

RULE 13—RELAX THE WRISTS

This includes hollowing the palm, and the comfortable curling of the fingers. Relaxation of the wrist allows the Shu-Stream points in the wrist to communicate with the fingers (see the Five Antique Shu Points in Chapter 8). Hollowing the palm and curling the fingers allows for the natural stretching and curving of the hand, which prevents the Qi from diffusing. It is important to avoid being rigid and purposefully bending the fingers, as all hand postures effects the brain and the body's Qi flow.

Relaxing the wrists helps to regulate the three Yang and three Yin channels of the hands so that Qi and Blood can flow freely. The wrist must form a line with the shoulders. Do not bend the wrists, but hold them straight so that the middle finger forms a line with the forearm, allowing the hands and fingers to hang relaxed but not limp.

RULE 14—SUSPEND THE HEAD

The classics say that "the head is held aloft not by force, but as if it were placed there by a ruling power." Hold the head as if it were suspended, keeping the Baihui point pointed to Heaven (Figure 15.31). This produces space in the body by elongating the spine. When suspended, the head must be kept upright, positioned in the middle of the chest and shoulders. This helps to set the whole body in the upright position, and

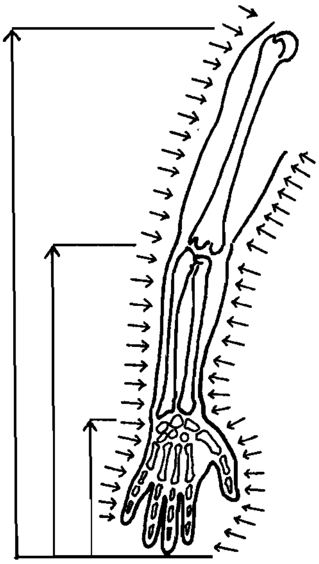

Figure 15.30. Allow the bones in the arms to feel heavy, naturally stretching the tissues and increasing Qi flow to the extremities.

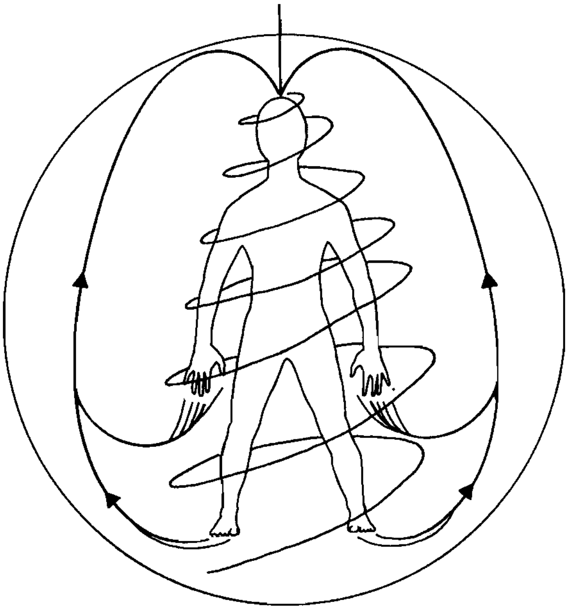

Figure 15.31. Suspending the head allows the body to become sensitive to the surrounding environment and external fields of Qi.

also directs the Qi and Blood to flow upward to nourish the brain and the mind. The Shen will be reinforced while the Jing and the Qi will be invigorated (it is the Shen which controls all the physical activities). An effective visualization is to imagine your head lifting high into space, your waist positioned in the clouds, and your feet sunk deep into the Earth.

RULE 15—TUCK THE CHIN

The head cannot be properly suspended if the chin is not tucked in. Only then can the Baihui point on the top of the head be correctly positioned and the nasal respiration kept free.

Tuck the chin and slightly lean forward to stretch and straighten the spine. You should feel the stretch all the way down into the coccyx. Next, slowly shift the weight backwards, stacking each vertebra on top of the next, starting from the bottom of the coccyx and sacrum and moving upwards.

RULE 16—CLOSE THE EYES FOR INNER VISION

The Chinese have a saying: "close the curtain." This refers to dropping the eyelids to facilitate inner-vision, and focusing on any specific areas where Qi is being trained or circulated. The eyes are of great importance in Medical Qigong exercises. It is said that the functional activities of Qi are determined by the eyes, and that the eyes are the messengers of the mind, and the mind is the residence of the Shen.

Completely closed eyes induce an absolute Yin state which is generally used in meditations for tonifying any Deficiencies. However, when used excessively (with all Yin and no Yang energy), the Qi can become too inactive.

Completely open eyes induce an absolute Yang state which is generally used in meditations for dispersing any Excesses. However, if overused, or if the patient has an Excess Yang condition, this type of meditation can make the practitioner too restless.

When the goal is to regulate the body's energy, slightly open the eyes in order to leave a small thread of light inside the eyes. This will keep the Yang (light) within the Yin (dark), so that the Qi will be more balanced. Because energy leaves the body through the eyes, after performing various Medical Qigong exercises, or prescriptions, allow your eyes to close halfway to continue the cultivation of life-force energy.

Tightly shutting the eyes causes tension in the upper and lower eyelids. Closing the eyelids gently, however, will restrain the vision, which helps to calm the Shen and the mind.

When meditating, if the eyes look to the distant horizon, they should not be focused but stare into infinity, this technique is also called the One Thousand Yard Stare, and it helps in regulating the venous pressure in the cranial cavity.

Strive for balance. Each eye has its own specific responsibility and energetic function.

- The right eye is generally used to project intention and extend emotion, energy, and spirit out from the body. The right eye relates to Water, the Kidneys, Qi, moon, Yin, Earth, Metal, and the Po.
- The left eye is generally used to receive, drawing in images, emotion, energy, and spirit into the body. The left eye relates to Fire, the Heart, Shen, sun, Yang, Heaven, Wood, and the Hun.
- If both eyes look upwards towards the Upper Dantian, the body's energy will draw up into the infinite space of the Wuji and merge with the divine.
- If both eyes look towards the tip of the nose, the Middle and Lower Dantians are linked together for transforming Yuan Qi.
- Crossing the eyes results in the union of the Wood and Metal, the elements of alchemic transition, for increasing the development of psychic energy. This intensification of psychic development is due to the union of the Hun and Po within the body, and is said to connect the meditator with his or her Yuan Shen.
- After meditation, store the Qi of the eyes behind the posterior orbital surface of the eyeballs themselves. This prevents the eyes from overheating and drying out.

RULE 17—CLOSE THE MOUTH AND TURN YOUR HEARING INWARD

In order to enhance the Yang during Medical Qigong exercises, close the lips with the teeth gently touching. Closing the lips will prevent the internal Qi from flowing out.

The Chinese use the expression: "Stopping the ears," which means to focus the hearing inward so as to be free from outside interference.

RULE 18—TOUCH THE TONGUE TO THE UPPER PALATE

When the tongue touches the upper palate, it forms what is called the Bird's Bridge. The connection of the tongue links both the Governing Vessel (which ends at the nose and upper palate) and the Conception Vessel (which ends in the mouth). As the tongue touches the upper palate, the body's field of Wei Qi expands. This can help to eliminate diseases and prolong life. The belief is that when someone has a disease, the Governing and Conception Vessels are to some extent disconnected.

Some Qigong doctors test their patients' progress by having them meditate with their tongue suspended in their mouth. Upon command, they connect the tongue to the upper palate. The suddenly increased expansion and strength of the patients' Wei Qi, observed by the doctor, indicates the degree of their progress.

1. The technique called "Gathering the Saliva," begins when the tongue touches the upper palate, causing the saliva secretion to be increased. Saliva contains many enzymes, which not only aid digestion, but also benefit the patient's physiology. In ancient China, this practice was called Gathering the Immortal's Water or the Juice of Jade to create the Immortal Pill.

   There is an exchange of Yin and Yang energy within the saliva, as Heaven Qi and Earth Qi are gathered within man's Fluid Essence in order to create the Immortal Pill.

   Before swallowing the saliva, it is important to tilt the head slightly forward (after inhaling). Swallow the saliva three times with three big gulps, using the breath as a cork to enclose the saliva. Send the Immortal Pill, i.e., the combined saliva and air down the Conception Vessel, bathing the Five Yin Organs, and ending at the Lower Dantian. Next, draw the energy up the Governing Vessel into the brain to replenish and nurture the Upper Dantian.
2. There are Five Elemental Tongue Positions. Each tongue position is related to one of the Five Yin Organs and can be used to direct the Qi to that specific organ. Sometimes, while practiscing Medical Qigong prescriptions, if too much Heat is created, the patient will be instructed to "change the tongue position" as well as meditation. These Elemental tongue positions are described as follows.
   a. The Earth (Spleen) position requires that the tongue be placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line (Figure 15.32).
   b. The Metal (Lungs) position requires that the tongue be placed between the maxilla and mandible bones, suspended, behind the teeth (Figure 15.33).
   c. The Fire (Heart) position requires that the tongue be placed on the hard palate, at the front of the upper palate, behind the teeth and gum line (Figure 15.34).
   4. The Wood (Liver) position requires that the tongue be placed on the hard palate, at the middle of the upper palate at the center of the roof of the mouth (Figure 15.35).
   e. The Water (Kidneys) position requires that the tongue be placed on the soft palate, at the back of the upper palate (Figure 15.36).

DYNAMIC POSTURE DAO YIN TRAINING

While Static Qigong exercises accumulate and strengthen the Qi internally, Dynamic Qigong exercises train the channels, collaterals, muscles, and bones externally. The more dynamic movement a patient makes, the more the energy will be converted and transformed inside the body. When dynamic movement is discontinued, some of the accumulated energy will begin to dissipate, while the remainder will flow through the channels and increase the patient's Qi circulation.

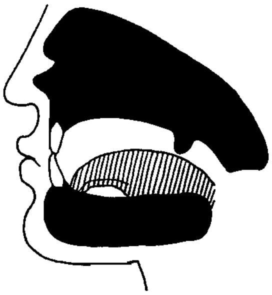

Figure 15.32. For the Earth position, which is related to the Spleen, the tongue is placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line.

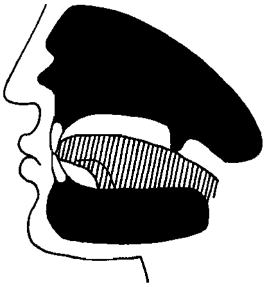

Figure 15.33. For the Metal position, which is related to the Lungs, the tongue is placed between the maxilla and mandible bones, suspended, behind the teeth.

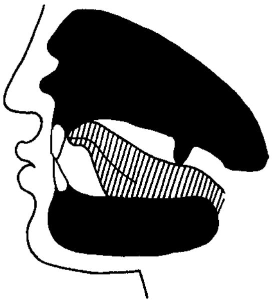

Figure 15.34. For the Fire position, which is related to the Heart, the tongue is placed on the front of the upper palate, behind the teeth and gum line.

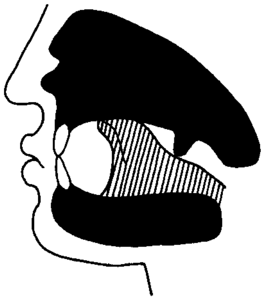

Figure 15.35. For the Wood position, which is related to the Liver, the tongue is placed on the middle of the upper palate at the center of the roof of the mouth.

The lymphatic system is pumped by normal muscular contraction. When a patient performs Dynamic Postural Movements, the lymphatic system is flushed, thus freeing the body of toxins.

Dynamic Qigong is divided into two methods of training: The Yun (Yin) Method and the Dong (Yang) Method.

1. The Yun Method is a more Yin approach, and employs slow, even, graceful movements. The focus of the mind's intention is placed on the internal organs, resulting in emotional relaxation.
2. The Dong Method is a Yang approach, and employs intense physical movements. Focus is placed on the movement of the body's ex-

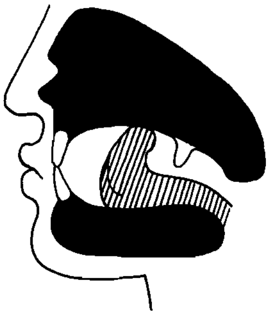

Figure 15.36. For the Water position, which is related to the Kidneys, the tongue is placed on the soft palate at the back of the upper palate.

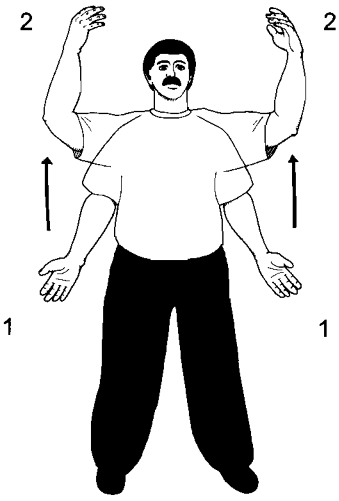

Figure 15.37. Hands Rise, Leading to the Ascending of the Qi.

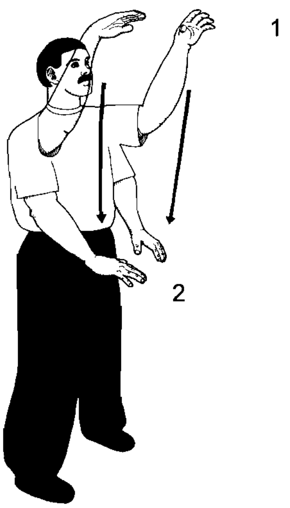

Figure 15.38. Hands Fall, Leading to the Descending of the Qi.

tremities, resulting in tissue regulation.

Both Yin and Yang (Yun and Dong) training methods must be balanced when prescribing proper Medical Qigong exercises. With the Yin method, the patient will experience external quietness with internal energetic movement. With the Yang method the patient will develop internal quietness with external, physically active movement.

USING THE ENERGY OF THE HANDS TO LEAD QI

Dynamic Postural training is performed by the extensive movements of the hands and torso. Because the palms have a stronger electromagnetic field than any other part of the body, the hands are used for leading and guiding the Qi. These movements are in harmony with the rotating, bending and stretching movements of the head, neck, torso, and limbs. Any change of the body's position or movement can change the direction or position of the electromagnetic fields of the hand. Likewise, any change or movement of the hands or palm positions, also changes the direction and flow of the electromagnetic field of the body. In Medical Qigong therapy the Dao Yin Hand Techniques have the following four functions:

- Hands Rise—Qi Ascends (Figure 15.37)
- Hands Fall—Qi Descends (Figure 15.38)
- Hands Open—Qi is Spread (Figure 15.39)
- Hands Close—Qi is Restrained (Figure 15.40)

The positions of these hand techniques can be further divided into three ranges: upper, middle, and lower. These ranges can relate to either the position of the body's Triple Burners, or the body's three Dantian areas, depending on the specific Medical Qigong prescription, and the doctor's intention for the exercise. The directions of these hand movements can further divide a posture into four front and four oblique types of Qi regulations. Moreover, depending on the arm movements, each posture can be divided into Yin (palms facing down) or Yang (palms facing up), as well as half Yin or half Yang regulations (Figure 15.41).

The actions of these methods supplement, regulate, smooth, and reinforce the vital Qi. Using these techniques will improve Blood and Qi circulation, as well as sharpen the sensory nerves. Since all hand postures relate to the brain and the function and flow of the body's energy, they also have a psychological and physiological affect on the nervous system.

When training, posture and Qi movement are

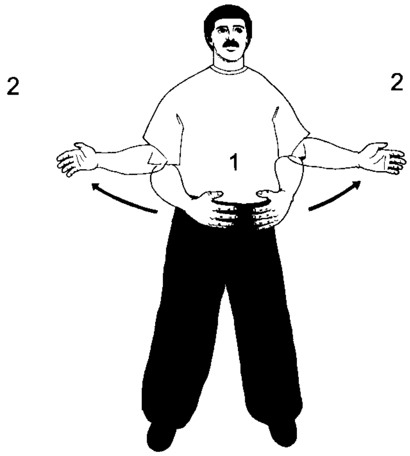

Figure 15.39. Hands Open, Leading to the Spreading of the Qi

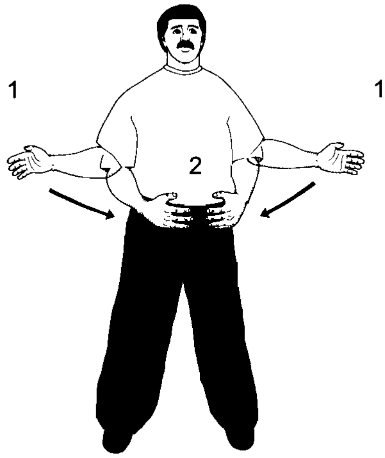

Figure 15.40. Hands Close, Leading to the Restraining of the Qi

sometimes in unity, and at other times proceed in opposite directions. Raising the arms, for example, causes the energy to rise; if the torso also rises, the rising power of the energy increases. This is an example of synergistic promotion of Qi regulation.

The following are three examples of the use of each separate hand action, and demonstrate how the body's reaction to energy can be used to treat patients with Medical Qigong therapy.

1. If you have hyperactive ascending Qi (e.g., hypertension) place your hands, palms facing downward, at the side of your body. Inhale quickly and move your hands quickly upward. Allow the middle fingers on each hand to touch each other, connecting the Pericardium Channels. Exhale, while slowly moving the hands downward along the front and side of the body. Purge and guide the Toxic Qi to descend the Liver and Gall Bladder Channels down the

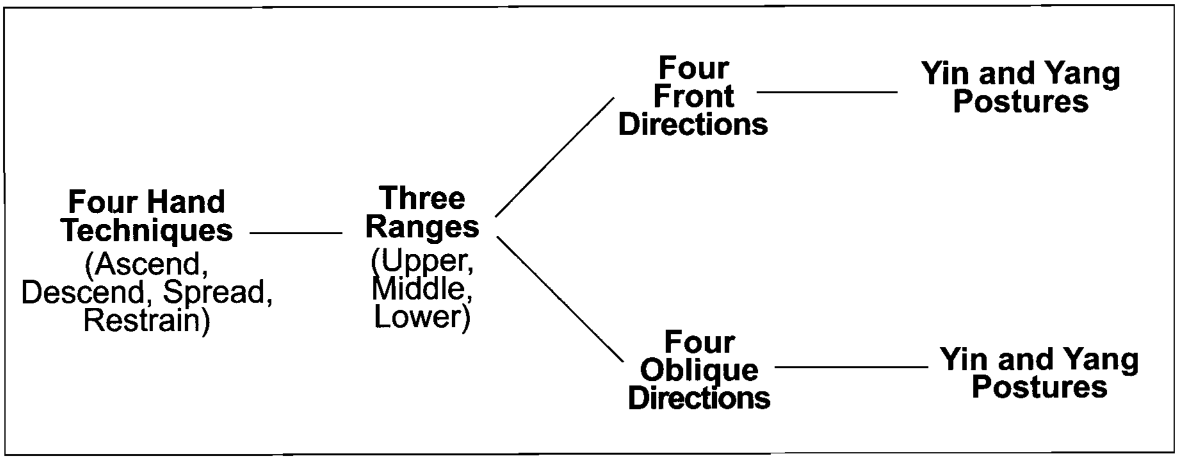
Four Hand Techniques
(Ascend, Descend, Spread, Restrain)
Three Ranges
(Upper, Middle, Lower)
Four Front Directions
Yin and Yang Postures
Four Oblique Directions
Yin and Yang Postures

Figure 15.41 Hand Position Techniques

torso to the hips (GB-30), and then to the outside of the legs (Figure 15.42).

2. If you have a Yin and Yang imbalance (e.g., chronic illness) place your palms facing your Lower Dantian and slowly raise them upwards to the chest. At the level of the Heart, turn the palms away from the body as you move your arms outward. Once the hands are extended away from the body begin moving them downward. Regulate the speed of motion of the hands and breath in order to guide and induce the Clear Yang to ascend (as you raise your arms), allowing the Turbid Yin Qi to descend (as you lower your arms). This balances the body's Yin and Yang Qi (Figure 15.43).
3. If you have a Deficiency of ascending Qi (e.g., hypotension or prolapse of the viscera) begin with the palms facing upward, in front of your Lower Dantian. Slowly inhale while gradually raising the palms close to your body. When exhaling, move the hands downward quickly away from the body to avoid any excessive downward flow that would counteract the tonifying upward flow (Figure 15.44).

MEDICAL QIGONG WALKING THERAPY

Dynamic postural training can be performed by keeping the feet stationary and rocking the weight between the heels to the balls of the feet, or by walking. The Chinese saying, "The body must walk like the wind," describes the relaxed placement of the feet and supple stepping techniques used in Medical Qigong Walking therapy. Walking, or Toe-Raised Stepping, as it is commonly called in China, is excellent in treating the diseases of the Five Yin Organs. It can stimulate the Qi of the Spleen and Liver (due to their channels originating on the big toes), and can cause the body's internal Qi to rise and flow vigorously, thus reinforcing the function of Qi and Blood.

With each step of the Walking Meditation, the patient should practice tranquil mind and body balancing. This is the same imaginative technique as is used in standing, sitting, or lying meditations. The inner stillness and outer motion are in harmony. Touching the heel to the ground, with the

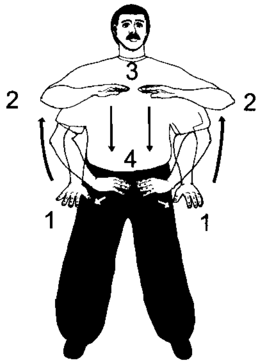

Figure 15.42. With palms facing downward (1), move the hands quickly upward (2). As the hands move slowly downward (3), allow the middle fingers on each hand to touch (4).

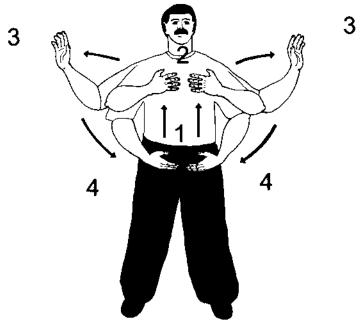

Figure 15.43. With palms facing the body (1), move the hands upward (2). Face the palms away from the body (3) as the hands move downward (4).

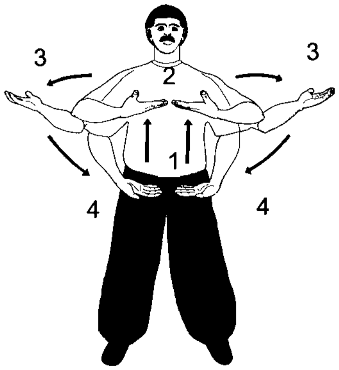

Figure 15.44. With the palms facing upward (1), raise the hands slowly close to the body (2). When moving the hands downward (3) in order to start again, quickly move them away from the body (4).

toes raised can also activate the Yin and Yang Heel Vessels (which originate from the center of the heel), and allow the energy to flow between both vessels so that the internal energy, exiting from the Yin Vessel, can enter the Yang Vessel, and vice versa. This allows the vessels to be dredged and the Yin and Yang energies to be regulated.

The heel is also stimulated when stepping, thus increasing the flowing action of the Bladder and Kidney Qi (due to their channel placement on the ankle). While stepping, focus the intention on the Lower Dantian to guide downward any abnormal rising of Lung Qi, which can cause shortness of breath, and affect the Heart and blood pressure. Each step is placed deliberately, with intention, and without collapsing of the feet. This stepping action relieves any Excess Qi syndrome of the upper torso and coordinates the balance of both the upper and lower portions of the body's Qi. Medical Qigong walking therapy consists of three training methods.

1. The Preparation (before walking), consists of the following steps:
   - The Opening and Closing of the Three Dantians, and
   - The Stationary Qi Strengthening and Regulating Exercises.
2. The Walking Therapy, consists of one or several prescriptions of the Toe Raised-Stepping techniques.
3. The Ending (after walking), consists of the following steps:
   - Leading the Qi back into the Lower Dantian,
   - Kneading the abdomen, and
   - Closing the Dantians and self-massage.

THE FUNCTIONS AND BENEFITS OF THE PREPARATION STAGE (THE OPENING AND CLOSING OF THE THREE DANTIANS)

Before beginning Medical Qigong Walking Therapy, the Opening and Closing of the Three Dantians must be practiced. In Medical Qigong training there is a saying, "When the body is Open, the Evil Qi will be expelled. When the body's surface is free from Evil Qi, it is better to Close it." The terms "opening" and "closing" refer to the purification of the body by means of Opening (purging) and Closing (sealing) specific areas of the body. The Opening and Closing of the Three Dantians can regulate either an Excess or Deficiency syndrome that is caused by an imbalance between the upper and lower, or interior and exterior parts of the patient's body. The purpose of the preparation stage is to:

- Eliminate the body's Evil Qi,
- Descend the Turbid Yin Qi, and
- Circulate the Clean Yang Qi through the organs and channels before walking.

The focus of this particular technique opens and closes the Upper Dantian (at the Yintang point), the Middle Dantian (at the Shanzhong CV-17 point), and the Lower Dantian (at the Qihai CV-6 point).

The Opening and Closing technique utilizes four specific disciplines:

- Using the Mind's Intention,
- Dividing the Yin and Yang Energy (ascending the Clear Yang Qi and descending the Turbid Yin Qi),
- The Hand Postures for Opening and Closing the Dantian, and
- The Focus on the Breath.

The Opening and Closing of the Three Dantians also requires multiple soft focusing that becomes easier with each practice session. Arm movements and the breath need to be perfectly coordinated; the expelling of Toxic Qi from the internal organs needs to be visualized; and the abdomen needs to be contracted while inhaling, and expanded while exhaling. The mind's intention must assist in the Opening and Closing of each Dantian, the coordinated movements alone do not suffice.

1. When focusing the mind's intention while performing the Opening and Closing techniques, imagine a dark Turbid energy from deep inside the body exiting through the body's pores. As the energy leaves, seal off the tissues to prevent it from returning to the body. Imagine the dark Turbid energy being absorbed into the ground, and being purified by the Earth.
2. The Dividing the Yin and Yang Energy refers

to separating the Yin and Yang energy within the body for healing. The Yin Qi stores the substance of vital energy, and the Yang Qi guards against pathogenic Evils and protects the body.

a. Ascending the Clear Yang Qi. When the Upper Dantian is Opened and Closed it allows the Clear Yang Qi to ascend to the upper orifices. The increased circulation of Clear Yang Qi through the channels and organs of the body stimulates and energizes the body's surface, as well as strengthens the four extremities. When the Clear Yang Qi moves upward its pure nature helps to tranquilize the Shen.

b. Descending the Turbid Yin Qi. When the Middle and Lower Dantians are Opened and Closed, it allows the Turbid Yin Qi to descend and facilitates the storing of the Clear Yang Qi.

3. The hand postures for Opening and Closing the Dantian is described as follows. The action for the Opening of Qi is directed by the outward movement of the hands and arms, and the mind's intent on Opening (or expanding) the abdomen.

   The Closing of Qi is directed by the inward movement of the hands and arms towards the abdomen, and the mind's intent on Closing (or contracting) the abdomen.

   The function of the Dao Yin Posture training, while regulating the breath and reinforcing the body's Qi, relies primarily on the movement of the hands and arms to arouse and invigorate the Qi circulating in the Large Intestine and the Lungs' channels. The action of opening and closing the Upper, Middle and Lower Dantians causes the Turbid Yin Qi to descend and be expelled, while the Clear Yang Qi ascends to the upper orifices of the body.

4. The focus is placed on the breath. The Opening and Closing exercise uses Reversed Breathing, i.e., the abdominal muscles contract while inhaling and relax and expand during exhalation (Figures 15.45 and 15.46).

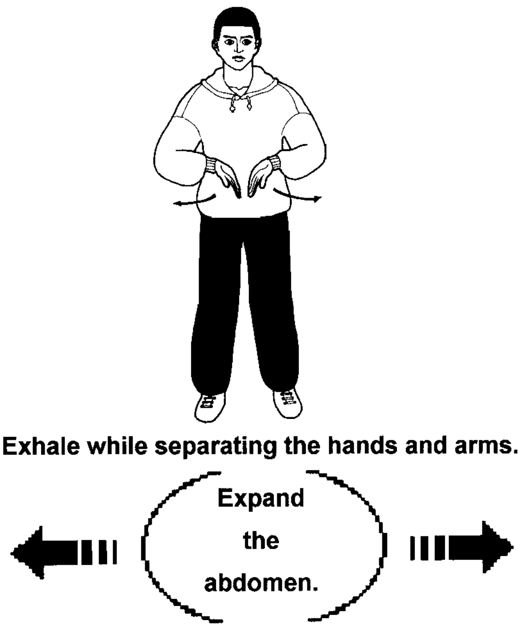

Exhale while separating the hands and arms.

Figure 15.45. The Opening of Qi is directed by the outward movement of the hands and arms and an opening (or expanding) intention in the abdomen.

This method of breathing is also referred to as Daoist breathing. The action of abdominal compression acts as a bellow to purge any Toxic (Yin) Qi from the body. These contractions and expansions of the abdomen are combined with the mental focus on the Lower Dantian. The nerve plexuses in the Lower Dantian area are considered, in modern Medical Qigong theory, to connect the Tai Yang Channels (Major Yang channels). If the mind is concentrated on the Lower Dantian for a long period of time, a conditioned reflex will occur that causes the nerve plexuses to produce bioelectricity, which spreads into the surrounding areas of the Tai Yang channels. This reflex causes the small blood vessels and capillaries to dilate. This reaction reduces the pressure in the blood vessels and improves Blood circulation to the Heart.

- Contraindications:

It is important to note that Reversed Breath-

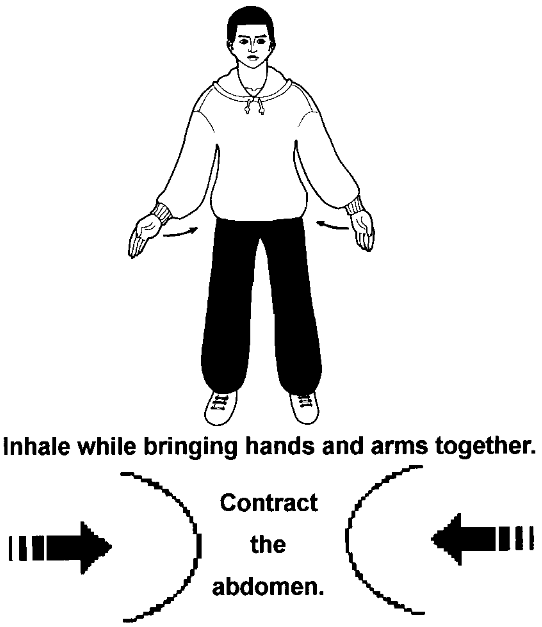
Inhale while bringing hands and arms together.
Contract the abdomen.

Figure 15.46. The Closing of Qi is directed by the inward movement of the hands and arms toward the abdomen for a closing (or contraction).

ing is contraindicated for young children, older patients and for people suffering from high blood pressure, stroke, or Heart diseases. In these cases, the Reversed Breathing techniques should be replaced with Natural Breathing. When practicing with Natural Breathing, the patients imagine Heavenly Qi descending into their body, filling and expanding the Lower Dantian, as they inhale. This expansion of the abdomen (upon inhalation) allows their Lower Dantian to Open, causing the Heavenly Qi to naturally purge the body of Toxic Qi. Upon exhalation, the patients contract their lower abdominal muscles to Close and seal their Lower Dantian.

THE OPEN AND CLOSE THE THREE DANTIANS METHOD

Normally an exercise begins with an equal number of inhalations and exhalations receptively, as the arms move toward the body, and away from the body. In this particular exercise, however, when approaching each separate Dantian, the first Open and Close action is to be performed by inhaling through the entire movement, without any exhalation. This action stimulates the opening of the Dantians. For the second through fourth Open and Close movements, the patients exhale when Opening and inhale when Closing. Practice the Opening and Closing of the exercise, step by step.

- Begin from Wuji posture. Raise your arms and hands to face the Upper Dantian. As you inhale, sperate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand, inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.
- Place your arms and hands in front of your Middle Dantian. As you inhale, separate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand; inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.
- Place your arms and hands in front of your Lower Dantian. As you inhale, sperate your arms outward away from the body as you contract the abdominal muscles, then while still inhaling, bring the arms back to the Dantian. Exhale as you separate the arms away from the Dantian, allowing the abdomen to expand; inhale and contract the abdomen as you bring the hands back together. Repeat this action two more times.

STATIONARY REGULATING EXERCISES (PREPARATION FOR WALKING)

This next preparation exercise is good for the patient's general tonification and for increasing respiration for Lung cancer patients.

- Start from a Wuji posture, inhale and exhale through the nose. Place your palms facing your lower abdominal area and begin extending Qi into the naval. Imagine the energy flowing into your navel, extending deep into your Mingmen and Lower Dantian area (Figure 15.47).

- Next, face your palms toward the ground and absorb the Earth Qi into your Lower Dantian (Figure 15.48). Absorb the Earth's energy for a few minutes.
- Turn your palms upwards, towards the Heavens, and begin absorbing the Universal Qi for several minutes (Figure 15.49). Finally, begin to close the exercise by allowing your palms to turn back towards your navel, absorbing and rooting the Qi back into the Lower Dantian

WALKING COMBINED WITH COLOR THERAPY

In China, to help visualize specific colors for therapeutic use, patients will carry brightly colored silk ribbons while walking. The colors and their affect on the internal organs are matched in accordance with the Five Elemental colors.

WALKING COMBINED WITH SOUND THERAPY

In China, in order to help the patients step with a structured rhythm, coordinate breathing patterns, and improve mental intention, Medical Qigong practitioners beat large wooden drums, click wooden sticks, and clang large metal cymbals. These rhythmic wooden and metal sounds resonate the walking area with an energetic field vibrating with Yin and Yang Qi. The patients, while walking in a circle, naturally absorb the Yin and Yang energy resonating from the instruments and seek to naturally regulate their internal organs.

THE THERAPEUTIC EFFECTS OF MEDICAL QIGONG WALKING THERAPY

The purpose of "raising your toes when stepping" is to invigorate the Yin and Yang channels and vessels of the legs. The energetic nature of the Spleen necessitates walking slowly during this exercise. The Spleen is the source of Postnatal Energy and is responsible for the activity of the muscles and extremities. The movements of the four limbs therefore strengthen the Spleen; this is why "toe-raised walking" and the "hand swaying" methods have both been adopted into the exercises and postures of most Medical Qigong therapies.

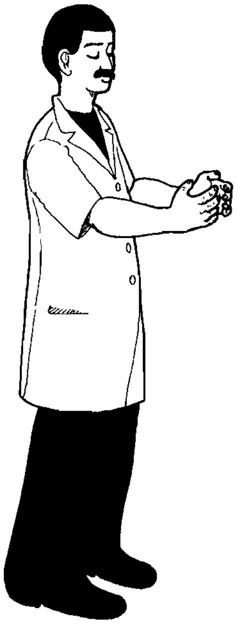

Figure 15.47. Begin by placing the palms toward the navel, in order to stimulate the Lower Dantian Qi.

Waist rotation plays a leading role in regulating the Liver Channels, as well as invigorating the circulation of Kidney Qi. This is achieved through all activities that initiate the actions of dropping, bending, relaxing, and turning the waist.

The opening and closing actions, combined with the rising and falling movements, and the swaying action of all four limbs, promote a pumping action within the muscles. This action is further intensified with the bending and leaning of the upper body, squatting and crouching of the lower body, and the alternate weight shifting when stepping. This pumping action greatly benefits the circulation of the lymphatic system.

Clinically, the Medical Qigong walking exercises result in increased appetite, strengthening

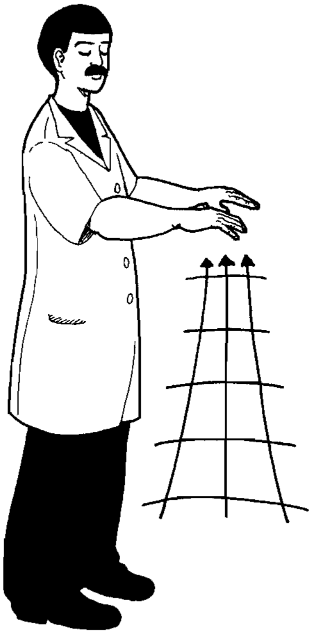

Figure 15.48. Next, begin to absorb the Earth energy into the Lower Dantian and Mingmen areas.

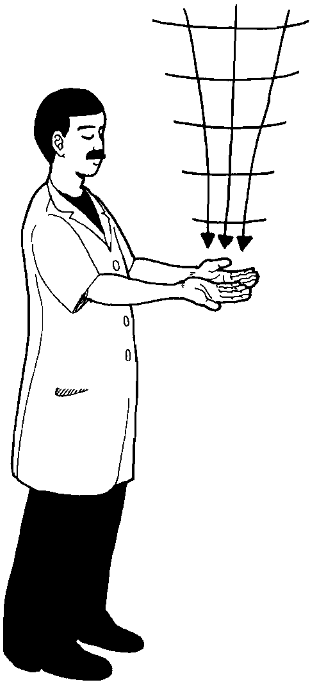

Figure 15.49. Turn the palms up to absorb Heaven Qi into the Lower Dantian Area.

of the patient's body, tonification of the Kidneys, Lungs, Liver, and Spleen, and the strengthening of Heart function.

Cancer patients need to regulate the function of their tissues and Yin and Yang organs to increase their healing potential. The initial goal in treating cancer with Medical Qigong, is to have the patient focus on the condition of the tumor, and its collection of energy. The patient imagines that the tumor area is beginning to soften, and the tissues are gradually transforming back into insubstantial energy. This technique is used to gradually dissolve tumors and certain types of less aggressive cancer.

The following are examples of Toe-Raised Slow Walking therapeutic exercises and prescriptions used in Medical Qigong therapy:

- Dynamic Regulation Walking Therapy
- Lung Tonification Walking Therapy
- Kidney Tonification Walking Therapy
- Liver and Spleen Tonification Walking Therapy
- Heart Tonification Walking Therapy
- Cancer-treatment Prescriptions

The explanation of these walking and breathing exercises and prescriptions are described as follows:

1. Dynamic Regulation Walking Therapy is a good exercise for regulating the Yin and Yang Qi of the patient's entire body. It can be prescribed for general tonification and nourishing of the patient's body, as well as strengthening the "anti-pathogenic" or True Qi. The Dynamic Regulation Walking Therapy is the primary walking exercise used to treat cancer patients.

   In every exercise the patient starts and ends with the Opening and Closing the Lower Dantian method. When stepping, the toes should be raised with the heels touching the ground (rocking back and forth), stimulating the Yin and Yang Heel Vessels of the legs, to strengthen the Kidney and Bladder Channels. As you step forward with the heel of the right foot, the Kidney and Bladder Channels open up. Swing your left hand up to the center of your Heart (CV-17 point) while your right hand swings to your right hip around and near the GB-30 point (Figure 15.50).

It is important to utilize the mind's concentration in conjunction with the breath. The breathing pattern is as follows:

- First step—inhale through the nose
- Second step—inhale through the nose
- Third step—exhale through the mouth
- Fourth step—hold the breath

The same details of stepping and hand movements apply to stepping forward with the left foot.

2. Lung Tonification Walking Therapy is a good exercise for tonifying the Qi of the patient's entire body. It can be prescribed for nourishing the patient's skin, as well as strengthening the True Qi. It is used for treating Lung Qi Deficiencies, bronchitis, tuberculosis, asthma, emphysema, and sinusitis.

The heel stepping pattern is similar to the Dynamic Regulation Walking exercise, however, when stepping, the hand postures will now switch from a normal relaxed position to the thumbs and first fingers touching at the tips (Figure 15.51).

Some Medical Qigong hospitals encourage the patients to rub their thumb and index fingers together when practising the Lung Tonification Walking Therapy. This rubbing action causes friction to develop between the Shaoshang Lu-11 and the Shangyang LI-1 points, which in turn stimulates and invigorate the flow of Qi along the Lung and Large Intestine Channels, strengthening the Lungs.

The breathing pattern is as follows:

- First step—inhale, inhale
- Second step—inhale, inhale
- Third step—exhale
- Fourth step—no breath

3. Kidney Tonification Walking Therapy. This exercise will purge the Liver Fire and tonify the Kidney Yin. It can be prescribed for patients who suffer from Kidney Qi Deficiencies, edema, blood diseases, gynecological diseases, and cancer of the Kidneys.

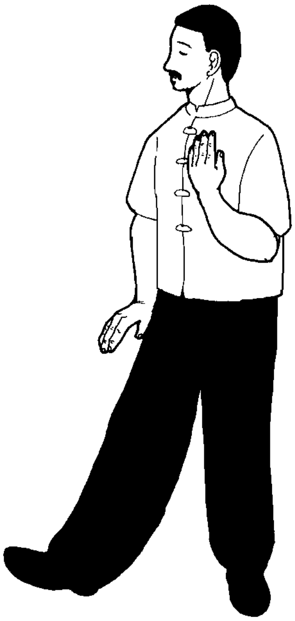

Figure 15.50. With each step, the opposite hand swings up to the Heart and Middle Dantian area.

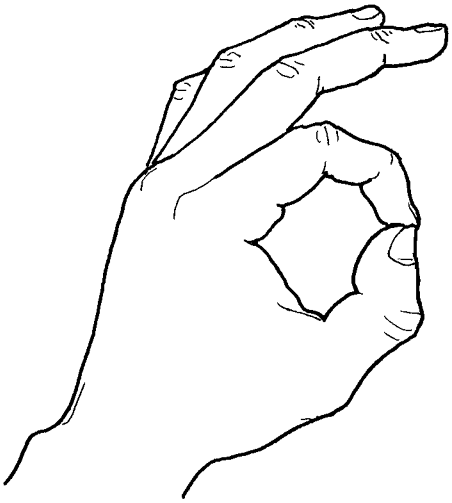

Figure 15.51. Tonification Hand Position for the Lungs

Begin this exercise with the palms embracing an imaginary ball at the naval and Lower Dantian area. This walking therapy causes the Qi to tonify your Lower Dantian.

Step forward so that the toes of the right foot touch the ground first, followed by the heel. This stepping action will open up the Yang Channels on the outside of your foot. Next, step with the left heel and point the toes up. This stepping pattern should be toe, heel, then step with heel (Figure 15.52).

On the first step, inhale through the nose while imagining the Qi filling your thorax. On the second step, sound the tone "Xu" (pronounced Shu) as you exhale. This particular sound is used to purge Toxic Liver Qi out the patient's body, allowing the Kidneys to fill with the rising Earth Qi.

- First step—inhale
- Second step—exhale "Shu"

The breathing pattern is as follows: inhale for the toe and heel, step and exhale on the heel step (all three actions equal one set). After walking in a circle for nine complete rotations, pause and begin again, starting with the left foot as the downward foot that makes the toe heel step as you exhale.

While stepping, keep both palms on your Lower Dantian. Regulate your breathing, while drawing the Earth Qi upward from the Yongquan Kd-1 point at the bottom of each foot. Press the Mingmen area backwards while breathing in the Earth Qi; this reinforces the Kidneys' Yuan Qi.

When ending the Kidney exercise, place both hands at the center of your chest at the CV-17 point. For men the right hand is on the outside, the left hand is on the inside, touching the chest. The hand position is opposite for woman. Make the sound "Shu" three times. End with the Opening and Closing of the Lower Dantian.

4. Liver and Spleen Tonification Walking Therapy activates the Liver's function of storing Blood and the Spleen's function of transforming Blood. It is used to treat digestive and blood diseases, hepatitis, Liver cancer, sclerosis of the Liver, and Gall Bladder stones.

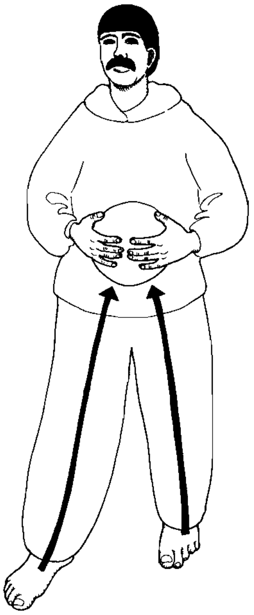

Figure 15.52. The Kidney Strengthening Step is as follows. While stepping, place both palms on the Lower Dantian, regulate your breath and draw the Qi upward from the bottom of each foot. Press your Mingmen area backwards while breathing in the essence of the Qi.

The starting position is similar to the exercise for tonifying the Kidneys with the toes pointing down, except that you step forward with the left toes. Step forward so that the toes of the left foot touch the ground first followed by the left heel. While this step is taken inhale twice through the nose. Then step out 45 degrees with the right heel while inhaling twice more through the nose. Next, step out with your left heel 45 degrees, exhale through your nose twice, then step out with your right foot with the toes pointed down as you exhale once.

- First step—inhale, inhale
- Second step—inhale, inhale
- Third step—exhale, exhale
- Fourth step—exhale

When walking with the toes raised, let the big toe touch the ground to stimulate the Yinbai Sp-1 and the Dadun Lv-1 points on the big toe. This promotes the ascension of the Qi flowing upward along the Liver and the Spleen Channels. The internal energy of these two channels is then reinforced, and the Liver and Spleen are strengthened.

After walking in a circle for nine complete rotations, pause and begin again starting with your right foot. Spleen and Liver Channels are activated when the toes pull up and the heel steps down.

The hand postures are as follows:

a. For Spleen problems the thumbs touch your ring fingers only when you step out to the fourth step (Figure 15.53).
b. For Liver problems the thumbs continuously touch your ring finger.

5. The Heart Tonification Walking Therapy calms the mind while tonifying the Qi and Blood. It is used for treatment of coronary Heart diseases, arthritis, palpitations, and angina pectoralis.

The starting position is the same as for the Liver and Spleen Walking therapy. The stepping is exactly like the last exercise, with the exception of the hand positions. In this exercise, the hands are opened and relaxed until they reach the chest (CV-17 point) at which point the middle fingers of each hand touches the Pc-8 point at the center of each palm (Figure 15.54).

When breathing, focus your Mind on inhaling the Qi upward from the Earth. When exhaling, imagine that the Qi within your chest is flowing downward.

- First step—inhale
- Second step—exhale

This walking exercise is practiced at a slightly slower speed then the previously described exercises. On the third step, the middle finger of each hand touches the center of your palms. This stimulates the Zhonchong Pc-9 point on each finger and is used to invigorate the flow of internal energy along the Pericardium Channels.

THREE STEPPING METHODS USED TO TREAT CANCER

These supplementary Medical Qigong Step-

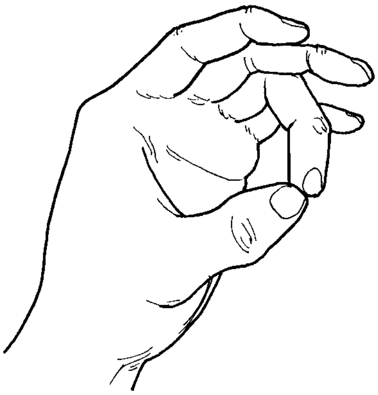

Figure 15.53. Hand Position for Tonification of the Liver and Spleen

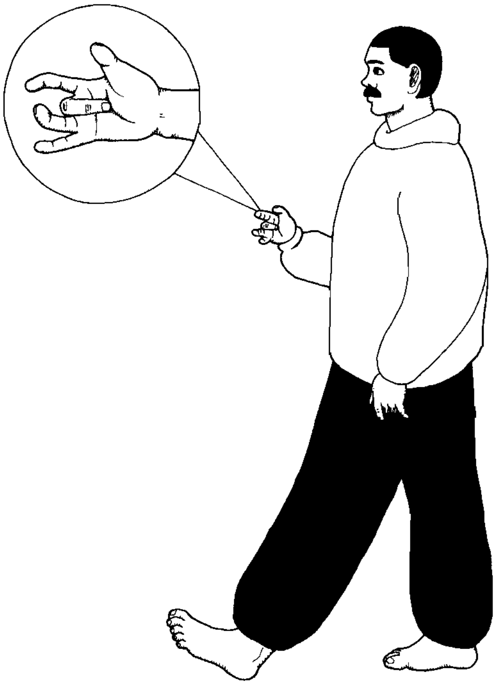

Figure 15.54. The Heart Strengthening Step is practiced at a slightly slower speed. On the third step, the middle finger of each hand touches the center of the palm.

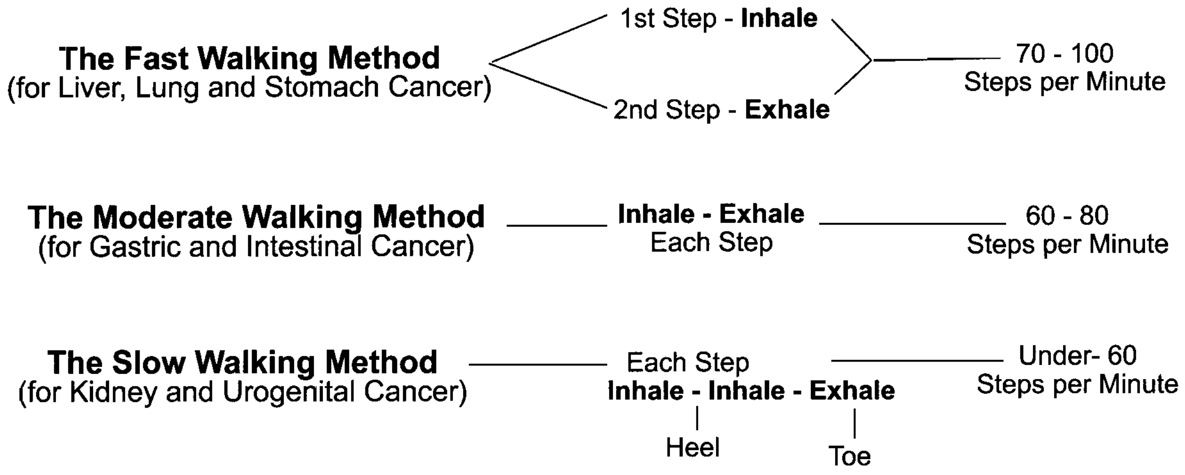
The Fast Walking Method
(for Liver, Lung and Stomach Cancer)
1st Step - Inhale
2nd Step - Exhale
70 - 100
Steps per Minute
The Moderate Walking Method
(for Gastric and Intestinal Cancer)
Inhale - Exhale
Each Step
60 - 80
Steps per Minute
The Slow Walking Method
(for Kidney and Urogenital Cancer)
Each Step
Inhale - Inhale - Exhale
Heel
Toe
Under- 60
Steps per Minute

Figure 15.55. Cancer Walking Therapy

ping exercises are given to cancer patients for breath regulating and Qi reinforcement. These breathing methods are applied with the Dynamic Regulation Walking therapy (the first walking exercise taught in this section). These walking methods use the "exhaling method" to strengthen the Kidneys, increase oxygen intake and enhance the patients' immune system.

The stepping methods are performed at three different speeds depending on the patients' symptoms (Figure 15.55). The patients are encouraged to walk at least once a day for 15-20 minutes. These walking exercises are classified into three kinds of prescriptions according to the following breathing rates.

THE FAST EXHALING METHOD

This walking method is mainly prescribed for patients in the early stages of cancer and is especially suitable for patients with Liver, Lung, and Stomach cancer. This exercise is practised using even breathing, inhaling for one step, followed by exhaling for another step. The patient steps out with the left foot, while inhaling through the nose. Then steps out with the right foot, and exhales through the nose. Both hands swing naturally up to the chest, while walking.

The main point to remember while stepping, is that the head leads the waist in turning and swinging the arms. The speed should be 70-100 steps per minute. When practicing the Fast Exhaling method it is important not to overexert yourself, as this would be counterproductive, and would lead to depleting the immune system.

THE MODERATE EXHALING METHOD

This walking method is mainly prescribed for patients in the early stages of cancer and is especially suitable for patients with gastric and intestinal cancer. The patient breathes in and out through the nose with each step taken. When taking each step, focus is on the respiratory rhythm. The speed should be between 60-80 steps per minute.

THE SLOW EXHALING METHOD

This walking method is mainly prescribed for patients with Kidney and urogenital cancer. The patient steps out with the left foot first and then with the right foot. Each step is accompanied by two inhalations and one exhalation. The patient steps out with the heel while taking two inhalations and then touches the ground with the ball of the foot, while exhaling. Because one step involves three breathing movements, the walking speed should be a bit slower, under 60 steps per minute.

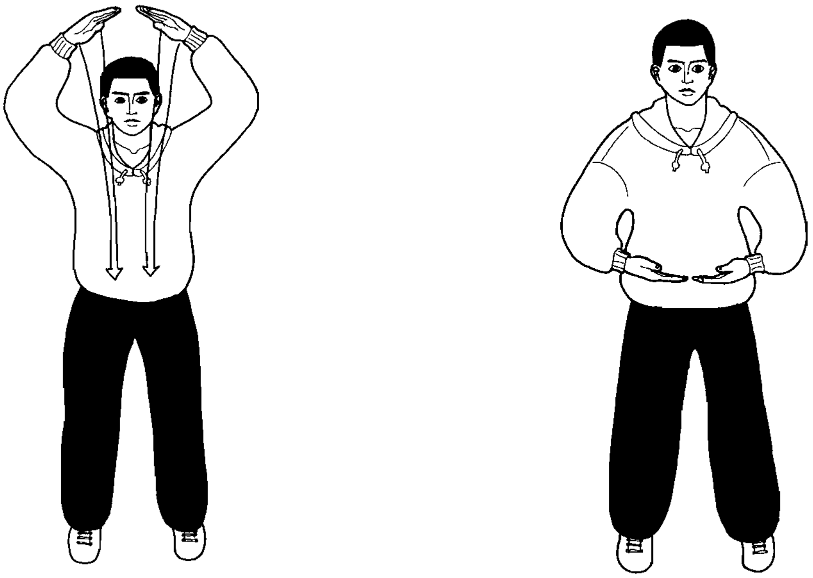

Figure 15.56. Secure the Upper and Middle Dantians in order to come to a tranquil state of relaxation.

ENDING AND CLOSING THE WALKING THERAPY

To end the Medical Qigong Walking therapy, the following exercises are taught to the patients in order to bring their Qi back into their Lower Dantian and to return to a state of quiescence. The Ending and Closing procedures are divided into three forms of practice: (1) Leading the Qi back into the Lower Dantian; (2) The official Closing of the prescription; and (3) Self-massage Regulation. These three ending procedures are described as follows.

LEADING THE QI BACK INTO THE LOWER DANTIAN.

This particular exercise is divided into four separate stages: (1) Securing the Three Dantians, (2) Kneading and Rooting the Lower Dantian, (3) Circle Massaging the Lower Dantian, and (4) Rooting the Qi.

1. The Securing the Three Dantians exercise has three purposes: to initiate a tranquil state of relaxation, to secure a harmonious balance of the body's Yin and Yang organs, as well as to bring the Qi back into the Lower Dantian.

   From a standing Wuji Posture, imagine embracing a ball of energy at the Lower Dantian (hold this posture for several breaths). After exhaling, separate the hands without inhaling, moving, and massaging the Qi along the hub of the body's Belt Vessel (at the waist).

   Inhale and begin raising the palms above the head. Imagine inhaling the Earth Qi into the body while the palms are facing the ground. At hip level, rotate the palms to face the sky, and begin inhaling Heaven Qi into the body. Continue raising the palms until they are positioned above the Baihui area on top of the head (left hand on top, opposite for women).

   Exhale and imagine the combined Earth and Heaven Qi flowing into the Taiji Pole, connecting all Three Dantians together. Descend the palms in front of the body, facing the Yintang, throat, Heart, upper abdomen (at the Yellow Court), and navel, then end by embracing the Lower Dantian. Begin again and repeat this exercise nine times (Figure 15.56).
2. Kneading and Rooting the Lower Dantian fills the Liver and Spleen with Blood and Qi, helping to further detoxify these organs.

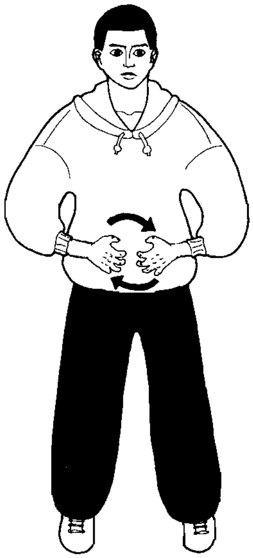

Figure 15.57. Knead the abdomen 36 clockwise then 24 counterclockwise.

After performing the Securing the Three Dantians exercise, secure the Qi in the Lower Dantian and knead the abdomen until it is sufficiently stimulated.

3. Circle Massaging the Lower Abdomen combines the body's organ energy together, thus forming the One True Qi. This allows the body to alleviate any Excess or Deficient conditions which may be caused from the excessive accumulation of Qi, due to the previous Qigong Walking exercises.

   Begin at the navel and start to circle massage the lower abdominal area thirty-six times in a counterclockwise direction (the movement should be directed against the flow of Qi in the Large Intestine). Begin with small and end with large circles, completing the movements at the base of the Lower Dantian. Next, reverse the direction of Qi flow by circle massage twenty-four times in a clockwise direction (with the flow of the Large Intestine), starting with big and finishing with small circles that end at the base of the Lower Dantian (Figure 15.57).

4. Rooting the Qi is sometimes referred to as "leading the Qi back to its origin." Rooting, extending, and stabilizing the energy into the Lower Dantian will facilitate the regulation of the Conception Vessel, Lower Burner, and Water pathways. Rooting the Qi also tonifies the Kidneys, Qi, Blood, and Yuan Qi, and allows the Toxic Qi to be easily expelled from the body.

   Leave the hands on the Lower Dantian and push the Mingmen backwards while inhaling the Qi and heat from the hands into the Lower Dantian (right hand on outside, opposite for women).

THE CLOSING

The closing procedure should be initiated after each practice period. After three deep breaths, click your teeth together thirty-six times (to stimulate the Kidney Qi), then circle the tongue eighteen times in each direction (to stimulate the Spleen Qi). Circle thirty-six times from the bottom of the teeth to the back of the throat. Tilt your head forward slightly, then swallow the saliva three times while imagining the Clear Qi of Heaven and Earth mixing with the saliva and flowing down to the Lower Dantian.

SELF-MASSAGE REGULATION

This method shifts the attention slowly off the area of concentration, while leading the Qi back down to the Lower Dantian. After the mind disengages from the specific point, relax the body, slowly open the eyes and perform self-massage.

Self-massage includes rubbing the hands, bathing the face (rubbing the face with the palms), combing the hair with the fingertips, and dredging the Twelve Primary Channels.

1. After rubbing the hands to heat the palms, place both hands on the face, cup the eyes, and draw the heat in through the eyes, into the Upper Dantian.
2. Next, comb the hair with your fingertips and drain the Excess Qi from off the head and neck area (Figure 15.58).
3. Then, rub from the feet to the abdomen to dredge the three Yin channels of the feet (Figure 15.59).

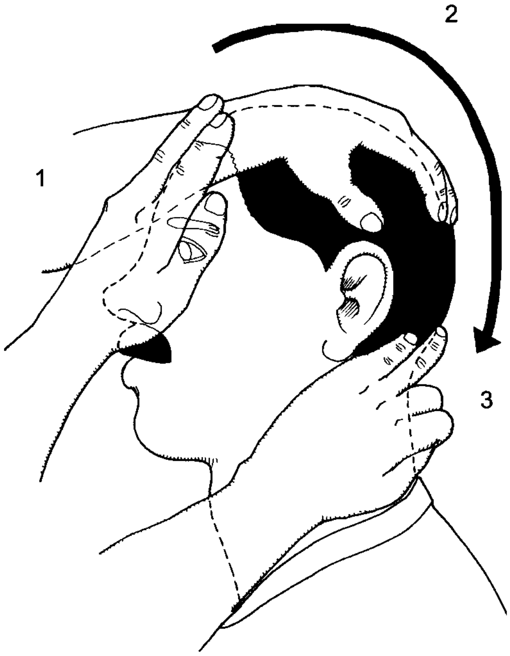

Figure 15.58. After rubbing the hands to heat the palms, place both hands on the face and draw the heat in through the eyes and into the Upper Dantian (1). Next, comb the hair with the fingertips (2), and drain the excess Qi from the head and neck area (3).

4. Rub from the chest to the hands to dredge the three Yin channels of the hands (Figure 15.60).
5. Rub from the hands, shoulders and lateral sides of the head, down the sides of the chest and abdomen to dredge the three Yang channels of the hands (Figure 15.61).
6. Rub from the waist and hips to the feet to dredge the three Yang channels of the feet (Figure 15.62).

These closing and self-massage procedures should be performed seven times each, followed by stretching to end the exercise.

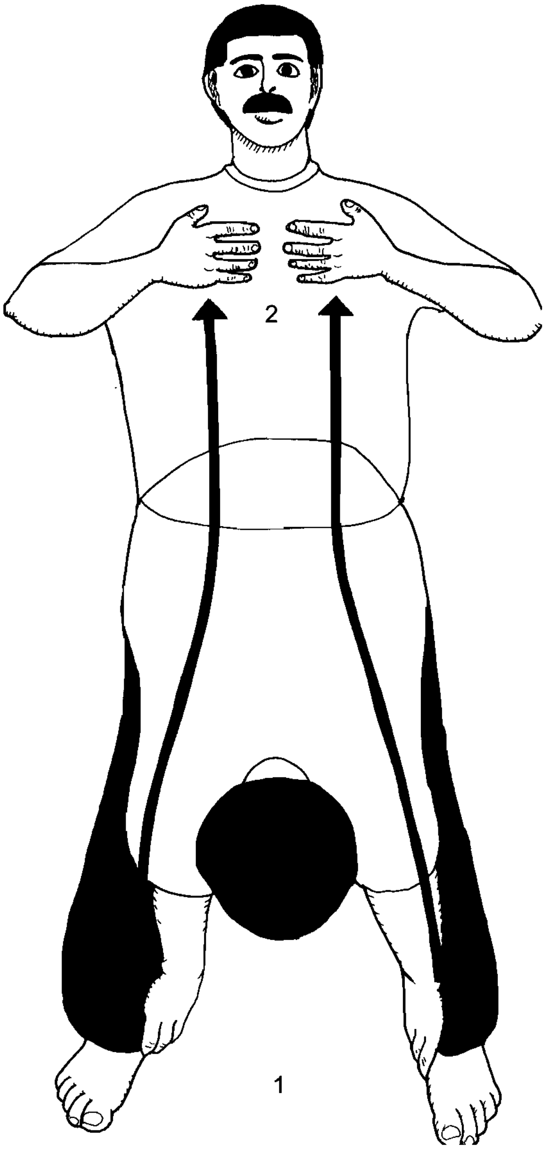

Figure 15.59. Rub from the feet to the abdomen (1–2) to dredge the three Yin channels of the foot.

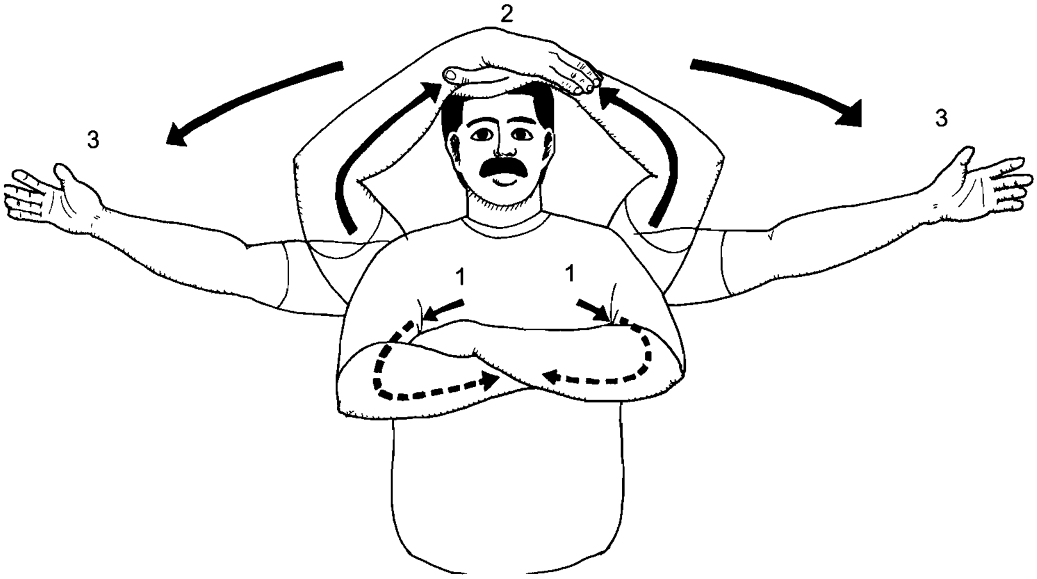

Figure 15.60. Rub from the chest (1) to the hands (2-3) to dredge the three Yin channels of the hand.

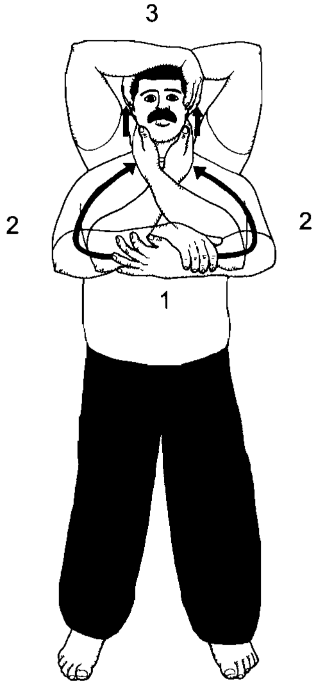

Figure 15.61. To dredge the three Yang channels of the hand, rub from the hands (1), shoulders (2), and lateral sides of the head (3).

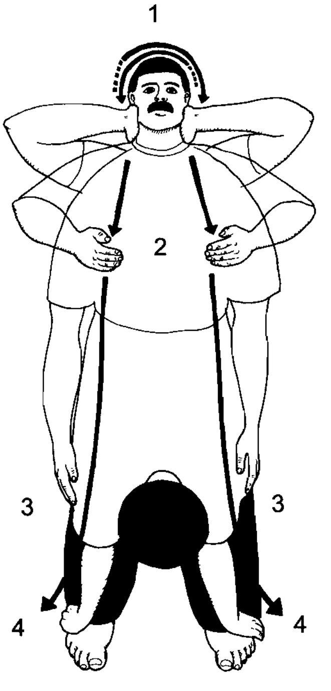

Figure 15.62. Continue down the sides of the chest (1) and abdomen (2). Next, rub from the waist and hips (3) to the feet (4) to dredge the three Yang channels of the foot.

CHAPTER 16
RESPIRATORY DAO YIN

THE PURPOSE OF RESPIRATORY DAO YIN

Respiratory Qigong therapy is often called Tu Gu Na Xin which means "expelling the old, drawing in the new." The purpose of Respiratory Dao Yin is to increase the intake and absorption of Clear Qi through regulating the breath. The extra absorption of Qi enhances the body's vitality, and harmonizes the Blood and Qi to promote health and eliminate disease.

Because the breath is also the link between the body and mind, by controlling the pace and quality of the breath, individuals can affect deep changes in their physiological functions. Breathing exercises have proven to be effective in reducing anxiety, depression, irritability, muscle tension, and fatigue. These techniques are also used in the treatment and prevention of agoraphobia, hypertension, hyperventilation, shallow breathing, and cold hands and feet.

THE ROLE OF THE LUNGS IN QI CULTIVATION

The Lungs control the body's Qi, although the Kidney Qi is responsible for aiding the Lungs by regulating the respiration (i.e., upon inhalation, as the Qi descends into the thorax, the Kidneys hold down and stabilize each breath). The Yuan Qi stored in the Kidneys must be continuously supplemented by the Gathering Qi derived from air, food, and water.

The internal absorption of air, food, and water rely on the smooth circulation of the Lungs' energy. Because of the significant role the Lungs play in absorbing universal and environmental Qi, the ancient Qigong masters placed a great deal of importance on breathing exercises. They believed that a patient's health and emotional condition were deeply affected by their breathing patterns, and by the amount of oxygen consumed in proportion to the amount of carbon dioxide released.

Through the study of these ancient Qigong observations, modern research confirms that relaxed and natural breathing patterns cause every tissue and cell to decrease its consumption of energy, while increasing its storage of Qi.

The ancient Qigong masters recommended that "vigorous breathing," in the form of Dynamic Breath Regulation, should be used when training Qi, and "gentle breathing," in the form of Quiescent Breath Regulation be used after training practice for nourishing and replenishing the Qi. Each breathing regulation exercise has long been considered a powerful prescription used to balance the patient's health.

DEEP INHALATION AND EXHALATION BREATHING METHODS

In breathing regulation exercises and prescriptions, the patient's inhalation is used to gather universal and environmental Qi into the body for tonification. Exhalation is used to eliminate Turbid or Toxic Qi from the body through purgation. Once the patients are in the prescribed posture, relaxed and free from distractions, they should, for example, exhale and imagine releasing Turbid Qi from out the mouth, nose, and pores of the body.

Clinical studies show that the Deep Inhalation of Medical Qigong prescriptions can have the following effects on the patient's body:

- increase stimulation of the sympathetic nervous system,
- cause the blood vessels to contract,
- raise the blood pressure, and
- increase the pulse rate.

On the other hand, the Deep Exhalation methods of Medical Qigong prescriptions can also have the opposite affects on the patient's body:

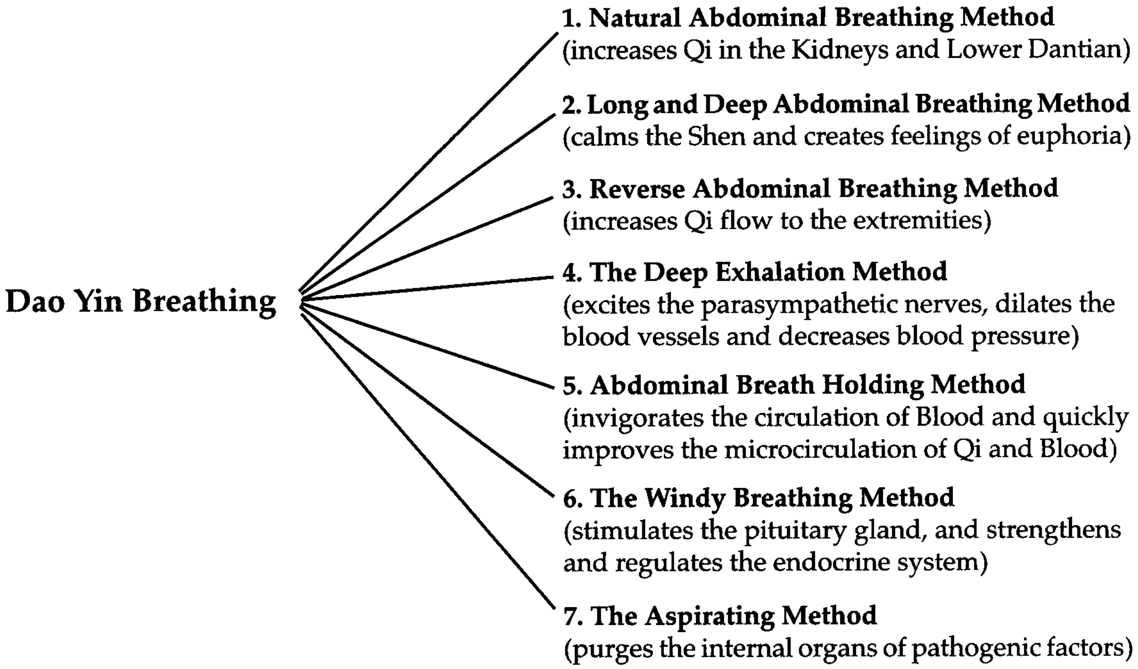
Dao Yin Breathing
1. Natural Abdominal Breathing Method
(increases Qi in the Kidneys and Lower Dantian)
2. Long and Deep Abdominal Breathing Method
(calms the Shen and creates feelings of euphoria)
3. Reverse Abdominal Breathing Method
(increases Qi flow to the extremities)
4. The Deep Exhalation Method
(excites the parasympathetic nerves, dilates the blood vessels and decreases blood pressure)
5. Abdominal Breath Holding Method
(invigorates the circulation of Blood and quickly improves the microcirculation of Qi and Blood)
6. The Windy Breathing Method
(stimulates the pituitary gland, and strengthens and regulates the endocrine system)
7. The Aspirating Method
(purges the internal organs of pathogenic factors)

Figure 16.1. The Seven Dao Yin Breathing Techniques

- increase the stimulation of the parasympathetic nervous system,
- dilate the blood vessels,
- lower the blood pressure, and
- decrease the pulse rate.

Deep exhalation helps to keep the body's Yin energy moving downward, and so has a downward regulating action on diseases caused by Yang Qi domination. Deep inhalation helps to keep the body's Yang energy moving upward, and has the effect of sustaining the Yang and tranquilizing the mind.

Students beginning Medical Qigong practice should focus their mind's intention primarily on the method of exhalation, to simultaneously stimulate the parasympathetic system and to relax. Once health is regained, patients will begin to alternate between deep inhalation and deep exhalation. This alternation produces a diastolic and systolic action of the blood vessels that also enhances the elasticity of the blood vessels' walls.

SEVEN DAO YIN BREATHING TECHNIQUES

The Seven Dao Yin Breathing Techniques are intended to regulate the symptoms caused by the Excess of Yang and Deficiency of Yin, which are characterized by too much energy in the upper part of the body and a weakness in the lower portion of the body.

Today, in China's Medical Qigong clinics, the breathing regulation exercises employ seven types of Dao Yin breathing techniques (Figure 16.1). Normally, all inhalation is done gently through the nose and exhalation is released slowly through the mouth. Inhalation and exhalation through the nose is also important, but should only be used when a higher state of "stillness" is obtained by the patient. This is to prevent the patient from falling back into his or her old shallow breathing patterns, until the Deep Abdominal Breathing method has been perfected.

1. The Natural Abdominal Breathing Method is also called Natural Breathing, and is used in

conjunction with mental concentration. In this method the abdomen expands upon inhalation and contracts upon exhalation. It naturally increases the body's peristaltic action, massages the internal organs, invigorates and increases the movement of Qi from the Kidneys into the Lower Dantian. Because of the differences in energetic physiology between men, women, and the breathing habits of each individual, Natural Breathing is further divided into Natural Thoracic Breathing, Natural Abdominal Breathing, and a combination of the two.

2. The Long and Deep Abdominal Breathing Method is also called Wen Huo Breathing. It is a gentle breathing method that consists of natural, deep, and full breathing. This results in a relaxed and comfortable feeling of euphoria. This method of breathing is performed as follows:
    a. At the beginning of the breath, inhale all the way down into the perineum, feeling the lower abdomen expanding in all six directions:

- from the bottom - at the base of the perineum,
- from the front - to the pubic bone and navel,
- from the back - to the coccyx and Mingmen,
- from the right - to the hip and lower ribs,
- from the left - to the hip and lower ribs, and
- from the top - at the base of the diaphragm.

    b. While continuing to inhale, feel the oxygen expanding and filling the upper thoracic cavity completely. The entire torso, from the perineum to the collar bones is expanded in six directions:

- from the bottom - at the top of the diaphragm
- from the front - to the solar plexus, Heart, and collar bones,
- from the back - to the base of the ribs, spine and top of the shoulders,
- from the right - to the rib area,
- from the left - to the rib area, and
- from the top - to the throat and neck area (Figure 16.2).

    c. Upon complete expansion, tuck the pelvic bowl under to increase the capacity of storing more energy. On the inhalation the pelvis performs a slight posterior tilt, on the exhalation the pelvis tilts slightly in the anterior direction. These actions will naturally ripple the spine. Exhale the breath from the upper thoracic cavity downward, to the lower abdomen.

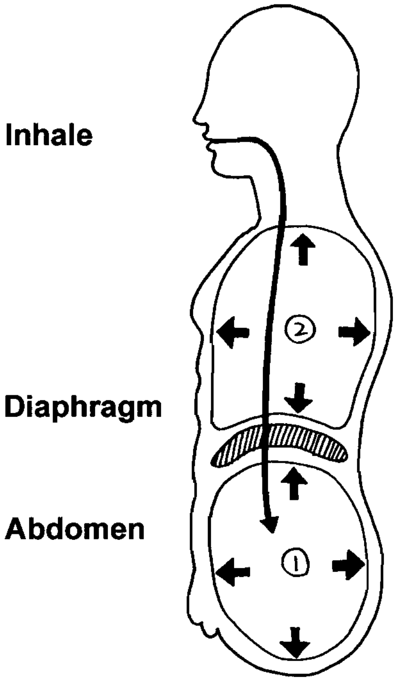

Figure 16.2. Inhale all the way down to the perineum. Make sure that the anal sphincter is closed in order to keep the Qi full and expanded in the Lower Dantian.

3. The Reverse Abdominal Breathing Method is an example, and subcategory of Long and Deep Abdominal Breathing. This method is performed by contracting the abdomen (and anus) when inhaling and expanding the abdomen upon exhaling. This is the primary breathing method used in Qigong training for guiding and emitting Qi, because the quality of air pressure and exertion of energetic pressure is greater. The Reverse Abdominal Breathing method results in superior efficiency in leading Qi to the extremities and has a much greater effect in raising the Qi from the legs into the brain.

However, when prescribing Medical Qigong therapy to treat diseases and alleviate physical stress, the Reverse Abdominal Breathing Method is prohibited except for very specific conditions. The reason it is not encouraged in Medical Qigong prescriptions is because patients with clinical hypertension or Heart disease often practice Reversed Breathing automatically. Scientific studies confirm that 90% of these patients practice reverse breathing chronically, as do most individuals with anxiety problems. Reverse Abdominal Breathing can also cause a constriction of the blood vessels and thus exacerbate problems such as cardiac diseases (i.e., angina) and migraine headaches.

4. The Deep Exhalation Method is often referred to as Wu Huo, which means "vigorous breathing." It is conducted with strong conscious intent. It can excite the parasympathetic nerves, dilate the blood vessels and decrease blood pressure.

   When practicing the Deep Exhalation method, the patient expels the Qi through the mouth, followed by a shallow inhalation.
5. The Abdominal Breath Holding Method invigorates the circulation of Blood and reduces swelling within the tissues, quickly improving the microcirculation. This method is also used to deepen the breath by stopping or holding the respiration for a short period of time.

   One Abdominal Breath Holding Method, known as Two Breathings, consists of inhaling through the nose, deep into the Lower Dantian, with one breath, pausing, then exhaling out through the mouth. The pause should be sustained for as long as possible without feeling tension (although a feeling of slight pressure building up is normal). After practicing deep abdominal breath holding, the breath will begin to develop its own stopping patterns. This stage involves the discipline and interdependence of the Heart and mind (Emotion and Spirit) with the breath.
6. The Windy Breathing Method consists of inhaling and exhaling through the nose. It is often referred to as "Windy" because it employs shallow breathing which resonates in the nasal passages. This kind of breathing purifies and warms the air as it passes through the nasal tract, resulting in air oscillation which stimulates the pituitary gland, thus strengthening and regulating the endocrine system. This method is designed to be used by patients with Lung disease because "the Lung energy opens at the nose."
7. The Aspirating Method refers to the method of drawing out by suction. This popular method can be used with healing sounds (or "tones") to purge pathogenic factors, or to regulate the body's internal organs. The pronunciation is based on resonating the sounds to stimulate specific organs such as the Heart, Spleen, Lungs, Kidneys, Liver and Triple Burners (Figure 16.3). There are two primary

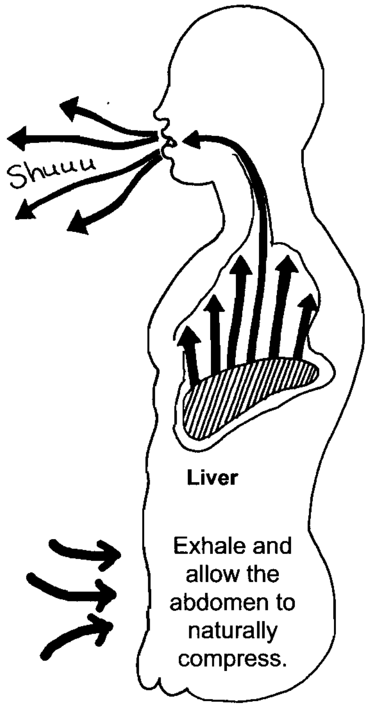

Figure 16.3. An example of tone resonation by aspirating the body's Excess Heat from the Liver is by using the sound "Shu."

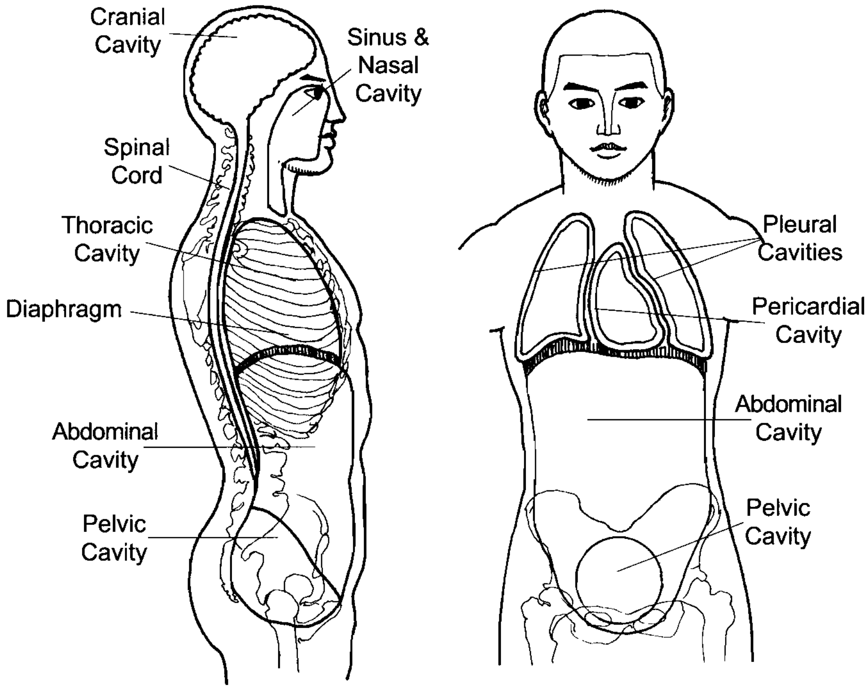

Figure 16.4. The body has seven main spatial cavities, as well as individual spatial cavities that surround the Yin and Yang organs.

sound systems of tone resonation used in Medical Qigong therapy, The Six Healing Sounds and the Tumor and Cancer Tone Resonations. These two healing sound methods are described as follows.

- The Six Healing Sounds are used to regulate, strengthen or detoxify the body's Yin and Yang Organs. This healing sound system is generally prescribed for patients with acute, or less serious conditions such as indigestion, insomnia, or Excess Heat in the organs, etc.
- Tumor and Cancer Tone Resonation Therapy is used to vibrate and destroy malignant tissue and cell growth in patients with chronic and serious organ diseases (see Chapter 40). This healing sound system is prescribed for diseases such as cysts, tumors, and cancer.

The choice of organ, and the number of times to pronounce each sound or character, is determined by the physical and energetic constitution of each individual. Each sound vibrates a specific organ or organ system, and surrounding set of involuntary muscles. Through repetition, the muscles acquire a kinesthetic memory of the vibrational sound. The doctor, or patient, can then trigger this memory by mental concentration alone to produce the desired results. As the patient performs the Sound Therapy, not only are the channels and organs cleansed, but also the spatial cavities (or Orbs) which surround the specific organs (Figure 16.4). For more information on Sound Therapy, see Chapter 40.

RESPIRATORY GUIDING AND REGULATING

There are four Respiratory Dao Yin methods the Qigong doctor can prescribe to regulate the

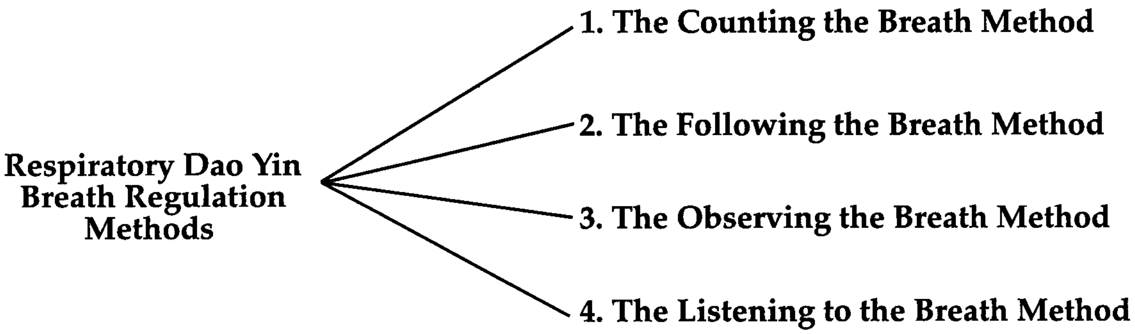
Respiratory Dao Yin Breath Regulation Methods
1. The Counting the Breath Method
2. The Following the Breath Method
3. The Observing the Breath Method
4. The Listening to the Breath Method

Figure 16.5. The Four Methods to Prescribe Respiratory Dao Yin For Breath Regulation

patient's breath (Figure 16.5): Counting the Breath, Following the Breath, Observing the Breath, and Listening to the Breath.

1. The Counting the Breath Method can be utilized to eliminate distractions of the mind during Walking Qigong Therapy by having patients silently count the number of their breaths. This method is also performed silently to regulate the body. In this particular method, one inhalation and one exhalation equals one breath.
2. When practicing the Following the Breath Method, the patients will contemplate each inhalation and exhalation of breath without counting. The patients mind, free from distractions or thoughts, follows each breath into and out of the body.
3. When practicing the Observing the Breath Method, the patients will observe their body's tissues conforming to the tranquil mental images from their subconscious imagination (e.g., relaxing, sinking, etc.).
4. When practicing the Listening to the Breath Method, patients are encouraged to listen to the "sounds" of their breathing (used in the Windy Breathing Method) to achieve a state of tranquillity.

RESPIRATORY TONIFICATION AND PURGING METHODS

Medical Qigong Respiration Therapy has its own unique methods and techniques for tonifying and purging the body's Qi. The primary technique used is focused concentration on the breathing. Through breath regulation, syndromes of Excess and Deficiency may be brought into balance either through tonification or through purgation.

1. Inhalation is used for tonifying the Qi in cases of Deficiency. The duration of the patient's inhalation should be longer than that of the exhalation. Tonification or Qi reinforcement patterns are created by applying the following methods:
   - more inhalation, and less exhalation
   - long inhalation, and short exhalation
   - focusing the mind on the inhalation
2. Exhalation is used for purging and reducing the Qi in cases of Excess. The duration of the exhalation should be longer than that of the inhalation. Purgation or Qi reducing patterns are created by applying the following methods:
   - more exhalation, and less inhalation
   - long exhalation, and short inhalation
   - focusing the Mind on the exhalation

BASIC ESSENTIALS OF RESPIRATORY TRAINING

Medical Qigong Respiratory training consists of the regulating, tonifying or purging exercises practiced in accordance with the doctor's treatment and specific homework prescriptions. There are four basic essential guidelines used to assist the patient in achieving maximum results in respiratory therapy, these guidelines consists of: Ensuring the Correct Posture Before Beginning, Purifying and Cleansing the Body's Qi, Tonifying and Purging the Qi, and Regulating the Body's Qi.

1. Before beginning Respiratory Dao Yin training, it is important that the doctor make sure the patient's posture is correct. When prescribing dynamic and static Medical Qigong therapy, it is best to start with a solid foundation and proper structure. Breath training should only begin when the patient becomes proficient in Postural Dao Yin, otherwise adverse effects may occur, (i.e., emotional trauma, headaches, chest stuffiness, or respiratory distress).
2. After their posture has been corrected, have the patients open their mouth and focus on their exhalation. As the patients exhale, they imagine all of the obstructed channels and collaterals being simultaneously purged of Turbid Qi. Each time the patients exhale they are to relax their body, more and more.

   Next, the patients close their mouth, and focus on inhaling clean, fresh Qi in to their body (through the nose). Have the patients practice this method using Natural Breathing. Each time the patients inhale and exhale they are to continue to relax the body.
3. After the body's Qi has been cleansed, the patients will focus their intention on any and all internal organs that specifically need to be purged. Breathing exercises used for purgation include focusing the patients' intention on the exhalation, as well as exhaling specific healing sounds. The patients continue to purge their body in this manner in accordance to the doctor's instructions.

   Having completed the purgation exercises, the patients are free to begin their tonification prescriptions. Breathing exercises used for tonification include the focus of the patients' intention on the inhalation, as well as visualizing specific colors being ingested into certain internal organs.
4. After practicing the tonifying or purgation prescriptions, the final goal of respiratory training is to achieve a long, natural, even and deep respiration. Urgent respiration causes the body's internal Fire to escape into the pores, resulting in uncontrolled Heat, feverish conditions and irritability. Normal breath regulation consists of even inhalation and exhalation patterns. The patient must avoid speeding up their breath, as this will also interrupt the breathing rhythm and lead to deviations of Qi.

BREATHING PATTERNS FOR COLLECTING AND REGULATING QI

The following are examples of exercises used for regulating and collecting Qi in the Lower Dantian. The doctor or patient will perform the breathing patterns as follows:

1. Collecting Qi in Lower Dantian is used to tonify and increase the collection of Qi gathered in the Lower Dantian.
   a. Inhale through the nose, filling the lower Dantian (expanding the abdomen), pause, then exhale through the mouth. Repeat and continue this breathing pattern for three times then proceed to b.
   b. Take three partial inhalations through the nose, filling the lower Dantian, pause, then complete one exhalation through the mouth. Next, inhale 50% of the breath into the Lower Dantian (expanding the abdomen), hold for three heartbeats, inhale remaining 50% into the Lower Dantian and hold for five heartbeats, relax and slowly exhale. Repeat and continue the pattern for thirty minutes a day (for three months), or until it becomes effortless.
2. Regulating Qi in Lower Dantian is used to stabilize the energetic balance of the Lower Dantian.
   a. Inhale through the nose, filling the lower Dantian (expanding the abdomen), exhale through the mouth, and

then pause. Repeat and continue this breathing pattern for three times then proceed to b.

   b. Inhale through the nose, then perform three partial exhalations through the mouth, pause, then one complete inhalation through the nose. Next, exhale 50% of the breath out of the Lower Dantian (contracting the abdomen), hold for three heartbeats, exhale remaining 50% out of the Lower Dantian and hold for five heartbeats, relax and slowly inhale. Repeat and continue the pattern for thirty minutes a day (for three months), or until it becomes effortless.

3. The Mysterious Pass is the pause between the inhalation and the exhalation. The ancient Qigong Masters believed that in the pause was the realm of infinite space (Wuji), existing between the Earthly Heaven (the congenital form of energy) and the Later Heaven (the acquired form of energy) where time and space stands still.

   For clinical practice, the Qigong doctors train in the method of accessing the Mysterious Pass in order to extend Qi through the entire body and to access the energy of the Wuji (Infinite Space). In the clinic, the training is divided into two stages (a and b).

   a. Begin this training by holding the breath from three to twelve counts (known as the "Small Count Breathing Regulation").
   b. Eventually increase to the "Big Count Breathing Regulation," which consists of 120 counts. This method enriches the Yuan Qi by breathing in more and exhaling less, and is sometimes used in Medical Qigong prescriptions to cure specific types of illness. When used with a strong intent, breath holding can constrict the blood vessels and raise the blood pressure; it is therefore sometimes prescribed for patients with hypotension. This Qigong method is, however, contraindicated for patients or doctors with hypertension.

THE FOUR PROGRESSIVE STAGES OF RESPIRATION

There are four progressive phases of difficulties that patients may experience during the beginning stages of their Medical Qigong Respiratory training: The Wind Phase, The Gasp Phase, The Unbalanced Air Phase, and The Quiescent Phase.

1. During the Wind Phase, the breathing is audible and may prove distracting to a beginner. The goal of focused concentration is to allow the energy to become gathered and collected. If the goal of the breathing prescription is to focus on a Deficient internal organ or specific tissue area, for example, and the patient's concentration becomes diverted (shifting to the noise of their own breath), it may disrupt the patient's mind. This can result in the patients collected Qi being dispersed instead of gathered.
2. During the Gasp Phase, the breathing is no longer audible, but any stagnant air, or obstructed air in the throat may disrupt the mind's concentration. Focused concentration on the stagnant air or obstruction may cause a sense of physical obstruction and mental anxiety, resulting in Qi stagnation in the throat (known as Plum Pit Qi).
3. During the Unbalanced Air Phase, the breath remains silent and there is no sensation of obstructed air in the throat. The patient's breathing, however, may be uneven, or unbalanced. Focused concentration at this phase may cause strain in the patient's mental awareness. If the mind becomes overstrained (trying to balance the inhalations and exhalations) the result will be fatigue.
4. During the Quiescent Phase, patients are able to achieve a long, natural, even and deep respiration without sound or stagnation, resulting in a state of extreme quietness. Concentration on this phase will result in a restful peace of mind.

DURATION TIME FOR RESPIRATION THERAPY

Patients in relatively good health should practice for about five minutes a day in the first week, when seeking to regulate their body's internal energy. In the second week patients should increase to ten minutes a day, and in the third week, fifteen minutes a day. Thereafter, five more minutes each day should be added until the patients reach a minimum of twenty minutes and a maximum of forty minutes per day for optimal Qi development.

When treating patients (for general tonification) using multiple standing meditation postures, have them begin by counting up to sixteen breaths per each posture. As the patients progress, the respiratory rate during meditation begins to slow down naturally. Normally, within a few months of training, respiratory rate drops from an average of sixteen breaths per minute, to three or five breaths per minute. This naturally increases the time spent in each posture, extending each posture from one to five minutes in duration.

Qigong doctors, on the other hand, should practice a minimum of forty minutes of standing posture each day to recharge and increase the efficacy of their treatments.

EMOTIONAL EFFECTS ON BREATHING PATTERNS

Breathing patterns are directly related to both the patients' thoughts and emotions. Any shift in one will affect the other. Each of the five primary emotions is associated with a specific breathing pattern as follows.

1. An angry patient's exhalation will be much stronger than the inhalation.
2. A grieving or sad patient's inhalation will be much stronger than the exhalation.
3. A fearful or scared patient's breathing will be fast, held high in the Lungs, and shallow because of the Kidneys' inability to hold the Lung Qi down.
4. A happy patient's inhalation and exhalation will be moderate, and irregular, with quick bursts.
5. When worried or deep in thought, the patient's inhalation will be shallow and weak, sometimes held for long periods of time, followed by a long gulping burst of inhalation and exhalation. The patient may also experience spontaneous sighing for prolonged periods of time.
6. When a patient experiences fright or shock, the breath immediately stops as the Qi attacks the Heart.

RESPIRATORY DAO YIN CANCER PRESCRIPTIONS

Respiratory prescriptions such as the Deep Relaxed Breathing Method described below, are used in treating cancer, and require the patient to combine the use of two methods of exercise and treatments:

1. The Opening and Closing Methods of Qi regulation (see Chapter 41), and
2. The Deep Relaxed Breathing methods for purgation, rather than tonification purposes.

The purpose of these treatment methods is to purge Excess Qi from the body, remove Blood stasis, and disperse Qi stagnations.

THE DEEP RELAXED BREATHING METHOD

The function of this exercise is to purge the channels and collaterals and to promote the circulation of Qi. It is also used to harmonize the ascension of Yang and the descending of Yin. Calming the Yin and suppressing the Yang promotes the flow of Qi and Blood through the organs, and has the effect of eliminating Evil pathogenic factors, and supporting healthy energy.

The Deep Relaxed Breathing method requires the patient to inhale through the nose and exhale through the mouth. The main points to remember in this method are as follows:

1. The patients inhalation and exhalation should be gentle, thin, even and long. They should focus their mind on the Hibernation Breathing Method. This method requires inhaling and exhaling through every pore on the body's surface while lying supine. The patients should not focus on their lower abdomen or the expansion of the upper chest and

Lung region. Instead, they should imagine inhaling through the pores into the Taiji Pole, and exhaling out from the Taiji Pole through the tissues and out the pores.

2. When breathing, the patients' whole body should stay relaxed; the key part to relaxing the whole body is to relax the waist.

Through training in these respiratory methods, the patients' immune system is enhanced and their vital Qi is nourished.

CHAPTER 17
MENTAL DAO YIN TRAINING

INTRODUCTION

There are two divisions of the Mind (Yin and Yang), each consisting of three different levels. The Yin part belongs to the energy of the Earth and is a more body-oriented type of mind. The Yang part of the Mind belongs to the energy of Heaven and is a more consciousness-oriented type of mind. The three levels of the Mind are as follows.

1. The first and deepest level exists within the Jing and the Shen, and represents the Original (Yuan) and primordial form of the mind.
2. The second or middle level exists within the Seven Corporeal Souls (Po) and Three Ethereal Souls (Hun) and represent the moving and active aspects of the mind (e.g., body movements, reflexes, instincts, drives, and astral projection - when accompanied by the Yuan Shen).
3. The third and superficial level exists within the Will (Zhi) and Intention (Yi) and represents the mind's everyday function (i.e., cognitive thinking). The Zhi is considered the "thinking body," while the Yi is considered the "thinking mind" (for more information refer to the Five Yin Organ components of Shen in Chapter 13).

These different levels of energies create the body, as well as activate the mental and emotional transitional states of the Mind. All three levels of the Mind are interactive and interdependent (Figure 17.1).

Energetic Division (Qi)	Yin	Yang
Energy of the Mind	Earth	Heaven
First Level of the Mind	Essence (Jing) Kidneys	Spirit (Shen) Heart
Middle Level of the MInd	Corporeal Soul (Po) Lungs	Ethereal Soul (Hun) Liver
Third Level of the Mind	Will (Zhi) Kidneys	Intention (Yi) Spleen

Figure 17.1. Mental and Emotional Transitional States of the Mind

Traditional Chinese Medicine believes that the Heart stores the Spirit (Shen). The Shen is further divided into Yuan Shen (the intuitive congenital spirit) and Zhi Shen (the analytical acquired spirit). The Shen is responsible for feeling and controlling the body's mental and emotional activities. The Zhi Shen is generally strong and can be stubborn and suspicious, therefore, it must be led by a master; the Yuan Shen must be that master, and must lead the Heart's Shen as one unit.

The brain is considered the house of the Yuan Shen, and the seat of mental function. It is believed that the senses and control of the body's physical movements are related to the brain. If the Heart has any functional disorder, the mental activities of the brain, and the emotional activities of the Heart also become disorganized. The Heart, which is the seat of all emotions, controls not only the Shen, but also the blood vessels, therefore Blood circulation is also influenced by the emotions.

The preservation of Jing depends on the cultivation of Qi, which in turn depends on the mental focus of the Shen. The mental focus of the Shen is to Qi as a mother is to her child. Mind concentration must combine with Qi; without the Shen

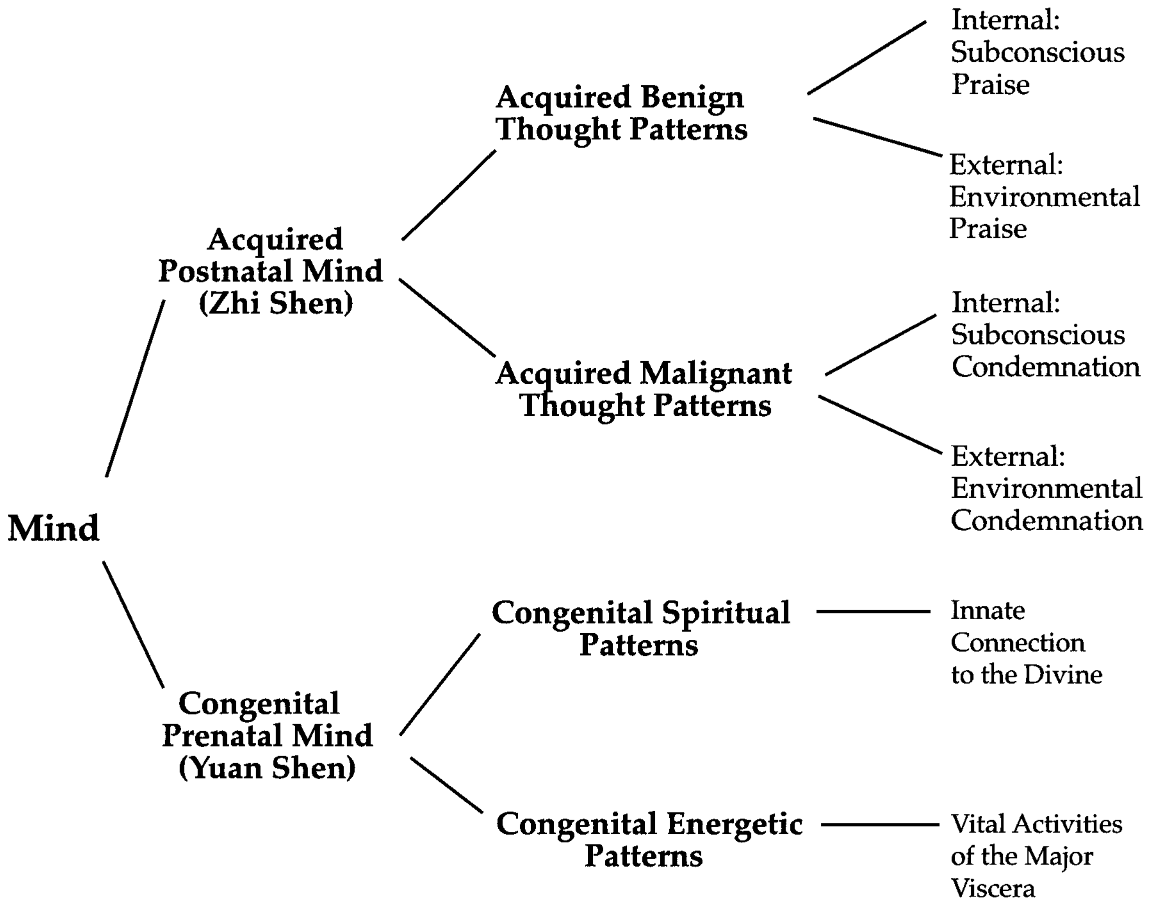
Mind
Acquired Postnatal Mind (Zhi Shen)
Congenital Prenatal Mind (Yuan Shen)
Acquired Benign Thought Patterns
Acquired Malignant Thought Patterns
Congenital Spiritual Patterns
Congenital Energetic Patterns
Internal: Subconscious Praise
External: Environmental Praise
Internal: Subconscious Condemnation
External: Environmental Condemnation
Innate Connection to the Divine
Vital Activities of the Major Viscera

Figure 17.2. The Differentiation Between the Prenatal and Postnatal Patterns of the Mind

guiding it, the body consumes the energy. The concentration of the Shen and Yi (intention) causes Qi to consolidate, while any distraction to the Shen and Yi causes Qi to disperse.

This concept is considered similar to that of regulating the emission of light, because Qi is emitted similar to the wave patterns of rippling water, and Shen is emitted similar to the wave patterns of light. When the Mind begins to focus concentration on the projection of Qi, both Qi and Shen unite, enabling the Qigong doctor to regulate the emitted Qi, changing its form from an expanded energy field to a finely honed beam.

THE PRENATAL AND POSTNATAL MIND

In Traditional Chinese Medicine, it is believed that the cerebral activity of the mind falls into two distinct categories: the Prenatal (Yuan Shen) Mind, and the Postnatal (Zhi Shen) Mind.

1. The Prenatal Mind (Yuan Shen) is inherited from the Jing of both parents and dominates the vital activities of the major viscera, as well as the function of the body's entire energetic organism and spiritual matrix.
2. The Postnatal Mind (Zhi Shen) is developed through interaction with people and the en-

vironment after birth. It dominates thought and interaction, and engages in unlimited mental activity.

The Yuan Shen and Zhi Shen are interactive and interdependent. If the Zhi Shen is active and chattering (sometimes called the Monkey Mind), it is difficult for the Yuan Shen to convey its perceived information. Regulating the Mind is therefore needed to harmonize the analytical and intuitive aspects of the Mind.

BENIGN AND MALIGNANT THOUGHT PATTERNS

Mental activity and thought processes can also be categorized into two clinical syndromes: benign and malignant thought activity (Figure 17.2).

BENIGN EMOTION AND THOUGHT ACTIVITY

This is a gentle or favorable belief structure of the subconscious mind, resulting in relaxed, pleasant thoughts. This thought pattern will both improve and regulate the excitation and inhibitory processes of the cerebral cortex. These positive thoughts and beliefs help to restore strength to the Yin and Yang organs, and thus promote health. Benign emotional thought activity can be further divided into internal and external influences:

1. Internal benign thought activities stem from the inner Hun's influences over the subconscious mind. These positive thought patterns tend to enhance self-esteem, by complementing and encouraging our thinking and actions.
2. External benign thought activities consist of praise and encouragement received from others (parents, friends, co-workers, etc.), which have been accepted and believed. These positive thought patterns promote confidence and self-esteem.

MALIGNANT EMOTION AND THOUGHT ACTIVITY

This is a harmful belief structure causing excited, nervous, stressful, and unpleasant thought patterns that interfere with the performance of the cerebral cortex, resulting in internal organ dysfunction and disease. Malignant emotional thought activity can be further divided into internal and external influences.

1. Internal malignant thought activities are the thoughts stemming from the subconscious mind, which tend to destroy self-esteem, cause obsessions, and discourage productive thoughts and actions. This thought pattern extends from the influences of the Po.
2. External malignant thought activities are internalized negative cynical beliefs and criticisms that originated from the surrounding environment (work, parents, teachers, friends, mate, etc.). External malignant thought activities cause thought distortions that destroy our self-esteem and discourage us.

THE IMPORTANCE OF MIND CONCENTRATION

An essential requirement for training the Qi is the training of mind concentration. If the concentration is forced, the Qi from the Lower Dantian rises up to the head causing the Stomach to become nauseated. If the Qi continues to ascend to the head, the result will be dizziness, headaches, and Qi deviations.

Thoughts and mental activities are reflexes of the brain and cerebral cortex. A Mind (whole body consciousness and awareness) agitated or depressed over a long period of time causes a long-term disorder to the whole nervous system, and leaves the body vulnerable to various kinds of chronic diseases. Long-term anger, for example, leads to the contraction of the blood vessels, sclerosis of the arterioles, gastric ulcers, and hypertension. Long-term emotional disturbances disrupt the endocrine system, and lead to a weakened immune system.

Psycho-neuro immunological research confirms that the Mind (whole body consciousness and awareness) can both cause and cure diseases. Healing visualizations sometimes act as a catalyst, increasing the rate of chemical reaction within the tissues. Improper concentration, however, can transform good Qi into Toxic Qi, and can lead to chemical imbalances that perpetuate the diseased condition. This is especially true when guiding and emitting energy.

According to medical statistics in China, with certain types of cancer, patients whose worry and anxiety were the main cause of their disease made up about 52% of the patients. These patients had been emotionally upset for half a year before becoming sick. The focus of the patients' mental concentration was directed on detrimental thoughts. It is a fact that chronic diseases are closely related to the patients' thoughts, emotions, and mood swings. To treat chronic diseases, it is necessary for the Qigong doctor to first relieve the patients' depression, irritability, and resentment and then place the patients, when possible, in a quiet comfortable environment.

It is important for the patients to relax their Mind (i.e., mind and body). Relaxation can be achieved through tissue manipulation (massage), or by inducing meditations, or trance states. To relax the Mind the Heart must be regulated. To regulate the Heart means to adjust the agitated or depressed Mind and place it into tranquility. One of my teachers in China, Qigong Master Zheng Zhan Ding would often say, "When the spirit, like a lake, is undisturbed, it mirrors the Heavens perfectly." Meaning that, the clarity of the Mind emerges gradually (or slowly) when entering into stillness, much like mud slowly settling in water. This allows ones true inner nature to be revealed.

According to research published by Dr. Karl Lashley, specific memory is not located in any one place in the brain. He found that destroying a portion of the brain does not destroy the memory assigned to that specific area, and also that memory could not be located in specific brain cells. His observation is that memory is distributed all over the brain as an energy field. Other researchers have made similar observations that support the conclusion that the human brain functions as a hologram, collecting and reading information from a holographic universe.

REGULATING THE MIND

Mental Dao Yin training involves regulating the Mind. This requires the Qigong doctor to diminish the mental activities (or judgements) of the Zhi Shen to prevent interference with the intuitive understanding of the Yuan Shen. The diminished mental activity allows for true relaxation, peace, and inner tranquility.

If the Zhi Shen is allowed to become unbridled through excessive internal chatter, it can become injured by the Excess accumulation of energy attributed to the Seven Internal Emotions: anger, worry, joy, fear, fright, sorrow, and grief; as well as the Four Desires that become pathogenic when in Excess: sex, money, fame, and power. These Seven Internal Emotions and Four Desires engage the mind, robbing the body of its life-force energy by depleting the Yin and Yang organs, disrupting their balance, and causing obstruction in Qi and Blood circulation that results in disease.

All thoughts carry within them emotional reactions, these in turn have a significant physiological as well as psychological effect on an individual's health. The degree of intensity as well as the frequency of returning thoughts determines the extent of the internal transitions experienced in the individual's body.

To illustrate this difficulty, when I began my initial Qigong training, my instructor informed me that the internal training would keep me warm in the winter and cool in the summer. Excited about the possibility of possessing my own internal air conditioner, I practiced diligently.

After three years of difficult training I had not yet developed this ability. I slowly began to lose trust in my instructor, and approached him to inquire about the promised skill. He informed me that it was the "imagination" that changes our thoughts and belief structures, enabling the body's tissues to feel what the mind directs. Stating that, "The Imagination leads the mind, the mind leads the Qi." My teacher further stated, "Whenever training in the summer when the heat is unbearable, focus your mind on the coolness of the sweat and imagine the air around your body as a cloud of cold air. In the winter time, when the cold is severe, focus your mind on the heat that your body is producing and imagine the air around yourself as a circle of hot fire." After applying these principles to my training, I experienced both the warmth and cool sensations that I was trying to achieve, and regained trust in my teacher.

Thoughts and emotions create electromagnetic waves in the brain. Your breath can also create or control these waves and cause your mind to take

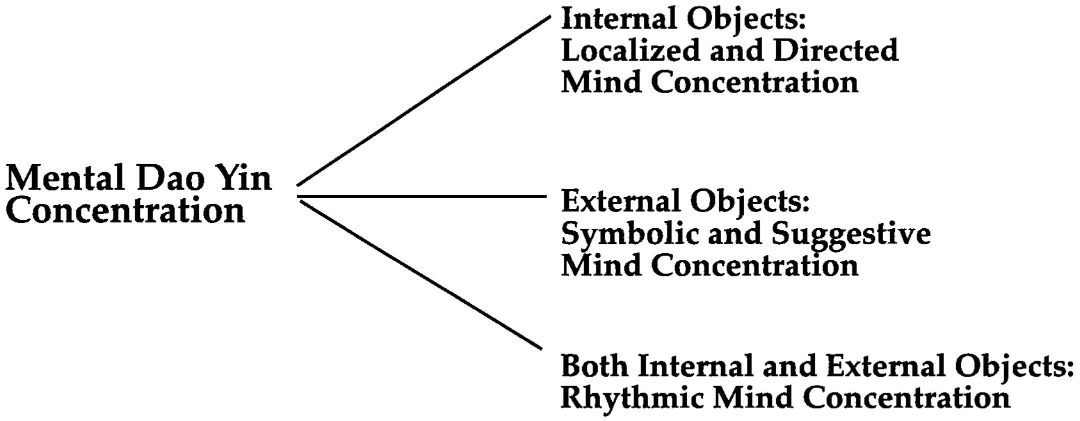
Mental Dao Yin Concentration
Internal Objects: Localized and Directed Mind Concentration
External Objects: Symbolic and Suggestive Mind Concentration
Both Internal and External Objects: Rhythmic Mind Concentration

Figure 17.3. The Three Categories of Mental Concentration

on the thought wave patterns of specific emotions. When the mind is tranquilized after certain Medical Qigong exercises, all disturbances stop, there is a quiet and comfortable feeling sensation (produced by the deep inhibitory state of the cerebral cortex) that facilitates the expansion of consciousness and awareness throughout one's entire being. This state of being is referred to as the Mind.

THE THREE CATEGORIES OF MENTAL CONCENTRATION

Mental Dao Yin training can be summarized by the following three foundational modalities: concentrating on internal objects, concentrating on external objects, and concentrating on both internal and external objects (Figure 17.3).

These three modalities of mental concentration encompass regulation of both the benign and malignant emotional thought activities which influence our everyday lives. People focus their mental concentration everyday on either benign or malignant internal objects (inner thoughts and feelings), external objects (outer thoughts and feelings), or engage in both internal and external thoughts and feelings through conversation.

CONCENTRATING ON INTERNAL OBJECTS

In China, the method of concentrating on internal objects is called "localized" and "directed" mind concentration. In this type of training, the mind is absorbed in concentrating on a specific area of the body. Patients, for example, are directed to focus their eyes downward along the nose in the direction of their navel, while simultaneously using inner-vision to concentrate on their Lower Dantian. Patients can also focus their concentration on other external or internal areas of the body such as:

1. The palms or feet to lead Excess Qi away from the head or torso.
2. A specific channel point (called "energetic point therapy") or pathway, to direct Qi into or away from specific areas of the patient's body, or
3. A Yin or Yang organ to direct Qi into or away from specific internal organs.

The ancient Qigong Master Li Shizhen once said, "When the mind is concentrated upon the inner channels, the practitioner is able to look within his own self." Qigong doctors and patients who are energy sensitive can feel the channels along which their own vital energy flows. This sensitivity is experienced in the more advanced stages of tranquility.

CONCENTRATING ON EXTERNAL OBJECTS

In China, the method of concentrating on external objects is called "symbolic" and "suggestive" mind concentration, and is focused through vivid imagery. In this type of training the patient's mind is absorbed in concentration on some form of object outside of the body. As the mind enters a state of tranquility, it focuses through vivid imagery, or energizes through sound vibrations. Some excellent examples of external objects are: music, the sound of rain or water flowing, a calm ocean, the sky, clouds, light, sound vibrations, mantras, a golden or white glowing ball, specific colors, trees, etc. In this type of concentration, the patient's internal vibration connects with and matches the external form of some object in nature or some specific symbol (such as a lit candle).

When focusing on external objects, it is important for patients with high blood pressure to choose low lying objects, such as a meadow, quiet stream, flower, etc. Patients with low blood pressure should choose objects that are eye level, such as a small pine tree, to balance the energy of the Heart.

When practicing this particular method of mental concentration it is important to focus on external objects that can be physically observed without straining. This way if distracting thoughts interfere with the concentration, the patients can return to the real image, or sound, to reestablish their concentration. The focus should never be forced or too intense. Dr. Xu, of the Xi Yuan Hospital in China, once stated, "In the state of emptiness you will see an image through your imagination. This image will arouse your emotions. These emotions help you to form a more colorful image. From these colorful images you will obtain a deeper state of emptiness. This is the state of Mind sought after in order to obtain emptiness and maintain tranquility." In this particular case the patient's mind is like still water, free from expectations or desires. In quietness you become the observer of images, allowing what will unveil itself to occur naturally.

CONCENTRATING ON BOTH INTERNAL AND EXTERNAL OBJECTS

These methods employ "rhythmic Mind concentration" and are the primary methods used in Medical Qigong therapy. They are simple, easy, safe, and effective. In this type of mental training, the patient's body responds to their thoughts. Patients focus their attention on the breath as it flows in and out of the body, listening to the sounds of their breathing and sensing the expanding and contracting movements of their abdomen. During this type of mental concentration the patients are sometimes given a mantra, such as thinking the word "quiet" during the inhalation and "relax" during the exhalation, or they are instructed to silently count numbers. Sometimes, the patients are also directed to imagine the feeling of quietness, like a mist, flow from their feet into their head (absorbing quiescent Earthly Qi), or from their head down into their feet (absorbing quiescent Heavenly Qi).

When first teaching patients the methods of mind concentration, keep the images simple. When the patients have recovered from their illness, they can begin to concentrate on either an internal or external object to further strengthen their health.

ESSENTIALS OF TRAINING MIND CONCENTRATION

These three mental Dao Yin techniques can be utilized to their full potential once the Qigong doctor understands the essentials of training the mind, the three stages of mental tranquility, and the skill of regulating the mind.

Mental activities should be coordinated naturally with respiration and posture. In Dynamic Qigong Dao Yin exercises, for example, mental activities must be adapted to the posture and the lifting, opening, and closing manipulations of the hands. When using an active (Yang) posture (if the goal is to regulate the patients' Qi), the inner concentration should be motionless (Yin). The ultimate goal is to combine both physical and mental activities that promote harmony in the body's energetic fields. The Qigong doctor should monitor their patients' mental concentration, making sure that they comprehend the following three concepts: Relaxing the Mind, Being Self-Confident, and Being Mentally and Emotionally Stable.

1. The purpose of relaxing the mind is to allow all mental activities to be carried out (without straining) in a composed state of mind. The mind should be kept clear and all distracting thoughts expelled gently. During concentration, the mental focus should never be forced. Mental concentration can be either strong (active) or mild (motionless).
2. Self-Confidence is a prerequisite, for without it, the training of mind concentration cannot happen. No matter what kind of mental activity the patients train in, they should be confident that they can reach their goal. This establishes a strong faith, which is the root of all mental Qigong projection abilities. Patients should also be realistic about not expecting quick results. Although miracles do happen, healing often takes time. The dynamic or active postures and exercises train the body's Qi, while quiescent or motionless exercises sustain and nourish the patients' Qi.
3. The purpose of being mentally and emotionally stable is to assist patients in not becoming overjoyed or frightened if something unexpected happens, or is perceived, during the Qigong exercise. It is important that patients take advantage of this time to keep their mind concentrated and the Qi consolidated. If patients feel weary, remind them that sleep and food can reinforce their mind and Qi.

CONCENTRATIVE MEDITATION AND INSIGHT MEDITATION

When Medical Qigong practitioners begin their meditative practices, they are generally encouraged to keep their mind focused and under control, this is known as "concentrative meditation." Concentrated power is manifested through silence; when the mind is noisy (distracted by physical sensations and chatty), the internal power becomes diffused. When Qigong doctors reach the place of silence in their Mind, they can connect with the divine power, a place of true power, and a place where "all is one." Through concentrated intention, Qigong doctors can bring all of their energies to bear on one point of focused power or force. At the same time, they contact and connect with the divine in silence (the hookup), and become one with the divine healing power.

This transition and conscious union with the divine occurs only when the Qigong doctors turn from the chaotic external world of sensory input (sight, hearing, smelling, etc.), to the quiescent state of silence. This divine energy is then manifested as power, substance and true intelligence, flowing through the Qigong doctors for healing.

Contact with the divine is only obtained in the deepest part of silence. Within oneself, and within this silence abides a state, or dimension, where the true "Gate of Heaven" can be opened and enlightenment obtained. Ideas conceived in this divine state of consciousness come into expression and can manifest as form. All Jing, Qi, and Shen can be transformed and transmuted into form, through this change in consciousness. When the doctor's Intention (Yi) and Will (Zhi) are aligned with the Dao (or divine will) all things are possible. The doctor is in direct contact with the universal Mind, and can manifest in physical form, that which is needed through faith. The divine energy resides within the body's Eternal Soul as power, substance, and intelligence and is brought into form and expression through consciousness. The degree to which the doctor can tap into and utilize the consciousness imbedded in the infinite Mind of the divine, is determined by the concept, or belief, that is held in the doctor's as well as the patient's conscious and subconscious mind.

In the silent state of quiescence, when the Qigong doctor sees a mental pattern or mold, a template is created into which will flow the substance (Qi) needed to bring it into being. The pattern will manifest into the form intended by the doctor's consciousness. Through the power or process of faith and thought, the doctor (or patient) can transmute and evolve the body or outer conditions and surroundings, by recognizing the divine consciousness within themselves.

As the practitioners advance in ability, they are encouraged to focus their attention indiscriminately on all sensual stimuli they receive, this is known as "insight meditation." In practicing insight meditation, practitioners are encouraged to reevaluate their experiences and world views according to the cause and effect of their disease(s). Patients are also taught to use this type of insight to observe their physical and mental processes. This self-observer is used as a vehicle through which to observe, scrutinize, and carefully examine the fundamental energy transformations within the body for spiritual growth.

THE THREE STAGES OF MENTAL TRANQUILITY

Before the mind surrenders to a state of tranquility, it must first transition through three stages. The Qigong doctor must experience all three of these stages in his or her effort to become effective when treating patients. In China, Qigong masters have a saying, "When you root the Mind, the Heart will open up to ten thousand voices." This means that when the Mind is removed from all distractions and excess chatter, the Yuan Shen is free to receive intuitive perceptions. This state of "tranquility" or "stillness" also dissolves time perception, allowing time to disperse into nothingness (Wuji). The three stages of mental tranquility are described as follows:

STAGE 1

In the first stage, the doctor's four extremities and Lower Dantian may experience heat and tingling as the mind slowly eliminates distracting thoughts and begins to withdraw from the external world. The doctor's mental focus and concentration, however, is not consistent. While emitting Qi, distracting thoughts still emerge, and the doctor must patiently wait for the chattering mind to subside.

In this stage of training, especially in male Qigong doctors, the cognitive, analytical thinking and speaking may sometimes become sporadically impeded. This is due to the thin energetic field within the corpus callosum that separates the right (intuitive) from the left (analytical) hemispheres of the brain. This makes it more difficult for men to access both hemispheres simultaneously. Men should therefore train to rely more on their intuitive perceptions instead of analyzing (they must feel the energy, and not think it).

This however, is not a problem in female Qigong doctors, as the energetic field which separates the corpus callosum is much wider and already active. This allows female Qigong doctors to easily access both sides of their brain simultaneously, enabling them to draw upon the perceptive, intuitive as well as cognitive, analytical thinking and speaking hemispheres of their brain.

STAGE 2

In the second stage, there is an increase in heat to all four of the doctor's extremities from the Lower Dantian. Through intention, the connection to the Lower Dantian is strengthened and the body's Qi feels like a rushing wind flowing throughout the doctor's torso.

While diagnosing and treating, the doctor experiences one or several of the patient's "eight manifestations of Qi" (also called the Eight Energetic Touches). These manifestations consists of the following physical and sensory phenomena of energy: shaking, rippling, heat, coldness, sinking, floating, itching, and heaviness. As the doctor focuses on extending energy, he or she will be able to remove internal distracting thoughts, but will still hear the distracting sounds from the external world.

STAGE 3

In the third stage, all of the doctor's sensory input (seeing, hearing, tasting, touching, smelling, and perceiving) is sealed from outside distractions. The doctor's Yuan Shen and Zhi Shen are now relaxed and in a state of mental tranquility. The Qigong doctor can now approach the patient free of chattering thoughts, judgments and all outside distractions. This enables the doctor to intuitively perceive and sense the patient's energetic traumas, stagnations, deviations, and energetic clusters.

THE THREE PROCESSES OF THE MIND

The mind can be divided into three separate but interdependent processes, which are responsible for our everyday actions, responses and health (Figure 17.4). Thoughts, ideas, and feelings are constantly being implanted, influencing and affecting the spiritual, emotional, mental, energetic, and physical process of healing. The heal-

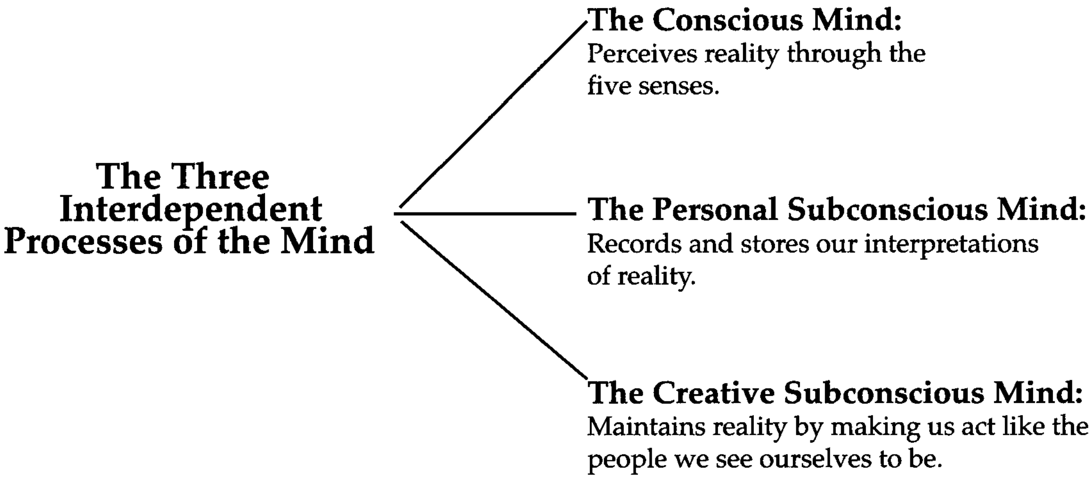
The Three Interdependent Processes of the Mind
The Conscious Mind:
Perceives reality through the five senses.
The Personal Subconscious Mind:
Records and stores our interpretations of reality.
The Creative Subconscious Mind:
Maintains reality by making us act like the people we see ourselves to be.

Figure 17.4. The three interdependent processes of the Mind, which are responsible for our everyday actions.

ing process includes reprogramming and regulating the conscious, personal subconscious and creative subconscious mind.

1. The Conscious Mind perceives reality through the five ordinary senses. It evaluates reality by investigating and interpreting the data through the filter of the past experiences. The conscious mind makes decisions based on personal goals, and manifests conscious thinking patterns. It is considered to be the Postnatal Mind (related to the Zhi Shen).
2. The Personal Subconscious Mind records and stores the conscious mind's interpretations of reality. It consists of the personal, perceptual data that were received through our five senses. The personal subconscious mind shares responsibility for how we think, what we say and imagine about ourselves and our emotional and behavioral reactions to those experiences. It is also responsible for the automatic function of the living mechanism (i.e., heartbeat, breathing, digestion, etc.) through its energetic connection to the autonomic nervous system. It is considered to be influenced by the spiritual aspects of the Wu Jing Shen (especially the Hun and Po).
3. The Creative Subconscious Mind maintains our reality by making us act like the people we see ourselves to be. It solves problems and provides the drive and energy to succeed or fail, heal or become sick. It is the center core energy of our innate spirit or soul consciousness. It is considered to be influenced by the spiritual aspects of the Yuan Shen.

CREATING THE BRIDGE OF LIGHT

The Bridge of Light is an energetically patterned wave frequency which consists of three main components of the body's energetic field. Its energetic connection to the body's tissues functionally envelops the conscious, personal subconscious and creative subconscious mind, and is considered to be the bridge between the three process of the mind. It consists of three separate but intertwining threads: the thread of life, the thread of consciousness, and the thread of creativity.

1. The thread of life comes directly from the Eternal Soul and is rooted in the Heart during conception. It is connected to all feelings that stem from the divine higher energy field and is considered the "seat of life."
2. The thread of consciousness also comes directly from the Eternal Soul and is rooted in the pineal gland within the brain. It embod-

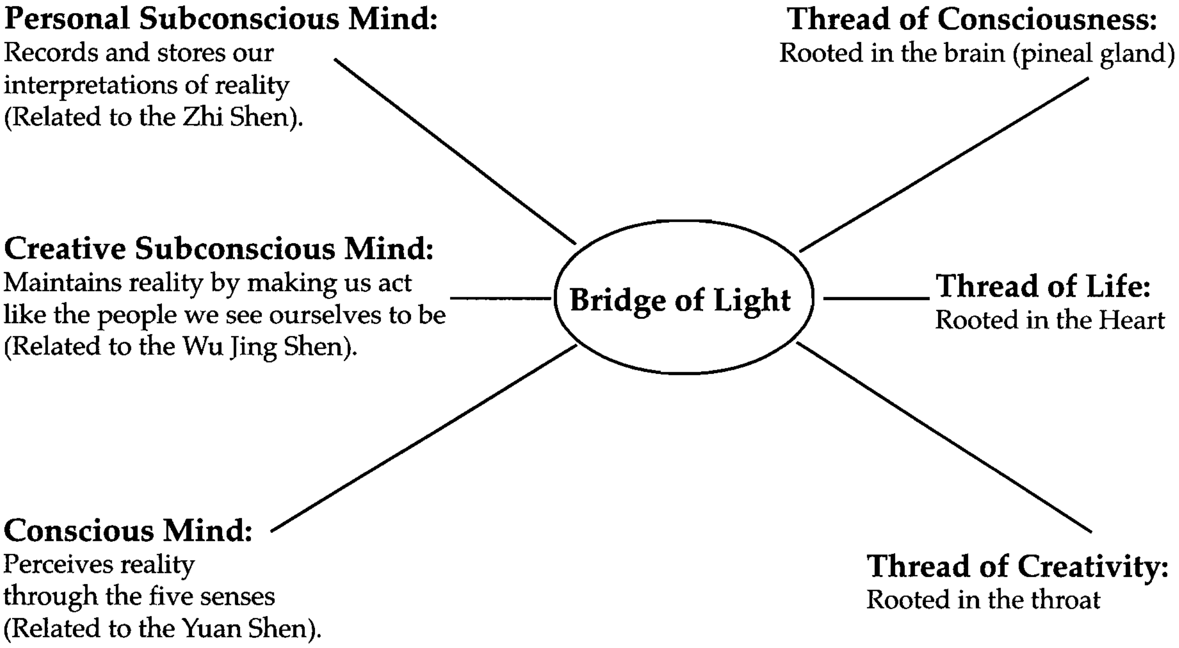
Personal Subconscious Mind:
Records and stores our interpretations of reality (Related to the Zhi Shen).
Thread of Consciousness:
Rooted in the brain (pineal gland)
Creative Subconscious Mind:
Maintains reality by making us act like the people we see ourselves to be (Related to the Wu Jing Shen).
Bridge of Light
Thread of Life:
Rooted in the Heart
Conscious Mind:
Perceives reality through the five senses (Related to the Yuan Shen).
Thread of Creativity:
Rooted in the throat

Figure 17.5. The Bridge of Light

ies portions of the energies of consciousness and is considered to be the "seat of consciousness."

3. The thread of creativity is rooted in the throat and is unique in that it is created and specifically constructed by each individual. It is also an extension, or synthesis, of the thread of life and thread of consciousness.

   The thread of creativity is in itself made up of three main components. These three components, or additional threads, intertwine as one unified thread of creativity and are explained as follows.

   a. The first portion of this thread extends from the physical Heart to the Spleen, and is connected from the body to the first external field of Wei Qi.
   b. The second portion of this thread extends from the solar plexus to the Heart and eventually extends and connects from the body to the second external field of Wei Qi.
   c. The third portion of this thread extends from the Yintang (Third Eye) through the brain, ascending through the Baihui region and beyond. This energy field is connected from the body to the third external field of Wei Qi.

After all three threads of creativity have been activated, energized, and developed, the next goal is to align and integrate the Qi of the thread of life and the thread of consciousness, with the frequencies of the thread of creativity.

Because the thread of life is connected to the Eternal Soul, the integration of the three threads allows an individual the direct connection to his or her core self (Figure 17.5). This enables the individual the ability to consciously access his or her true inner voice (the voice of his or her soul) and receive information and guidance from the divine. The goal in accessing the Bridge of Light is to unite these three energetic threads into one powerful and functional energy field.

When all three major threads (which compose the Bridge of Light) are connected as one harmonious cord of light (extending upward through the Taiji Pole), a spiritual gateway is opened that enables access to the Pure Heavenly Sound (also known as Hearing the Sounds of the Universe) (see the Six Transportations of Shen in Chapter 13).

REPROGRAMMING THE MIND TO HEAL

The mind holds specific energetic patterns which maintain a patient's physical, mental, emotional, and spiritual belief structures. These patterns are maintained by the dictates of the personal subconscious mind and can only be changed when the creative subconscious mind is engaged.

To change specific destructive patterns, the personal subconscious mind must be reprogrammed. There are three things that are needed in order to reprogram the personal subconscious mind: imagination, visualization, and positive affirmation. Imagination combines with vivid pictures, sounds, sensations, etc., and creates a new reality on a creative subconscious level. It is the awakening to this new reality that reconnects the individual to the creative subconscious mind.

If patients do not use imagery and affirmations to change their present personal subconscious representation of reality, the patient's personal subconscious mind will automatically seek to correct any newly made change. The personal subconscious mind views any new deviation from the normal energetic pattern as a mistake, and will automatically resist, or sabotage, any change of pattern, returning the patient back to his or her original state of mind and body sickness.

It is a simple fact that "we bring about what we think about," and act in accordance with the truth that we have come to believe. If we act out of a particular new state of mind and feeling over a sufficient time period, it will become a permanent reality of the self. In using the mind to make transitions, it is important to note that words, images, and sensations have tremendous power, and the spirit behind the intent is the key to any permanent transformation.

Because patients trust their doctor, it is the doctor's responsibility to direct the patients into a healing mental state for the purpose of restoring health. The doctor must never be guilty of "clinical hexing." Clinical hexing occurs when a patient (who completely trusts the doctor) is told by that doctor, that there is no hope. The patients, believing the doctor, may give up their desire to change and heal. Thus, through accepting the inevitability of death, they suppress their desire to live. Qigong doctors are therefore taught to always encourage their patients on their recovery, and to never destroy their hope.

Through positive affirmation and reinforcement, it is possible for the patients to turn around the progression of their disease. Scientific studies in China have proven, for example, that through imagination, visualization (which include using all of the five senses) and positive affirmation, the rate of phagocytosis (the ability of the immune cells to engulf and destroy invading germ cells) greatly increases and enhances the body's immune function.

USING THE IMAGINATION

The Yi (which consists of the imagination and intention) leads the Shen (composed of the thought, emotion and spirit) which in turn leads the Qi (the body's life-force energy). The power of the imagination has a strong influence on health, and can be used to help cure diseases, or to further complicate the condition. If, for example, patients imagine that their immune cells are effectively destroying the cancer cells, Western medical treatments such as chemotherapy or radiation can become more effective. Research from China reveals that patients with a positive belief structure produce certain physical transformations, which manifest in the increased production and release of hormones and immune cells. The imagination is the inner form of all things, the primal inspiration from which reality later manifests. In a sense, patients bring about what they think about. While treating disease, Qigong doctors encourage their patients to imagine healing energy from the divine being directed into the diseased area. Once the treatment is finished, the patients are instructed to imagine that their disease has become completely healed.

During the state of relaxed meditation, the foundational aspect of the patients' disease, which is normally hidden from consciousness, reveals itself in the form of images and impressions. After a Medical Qigong treatment, or meditation, the patients may describe to the doctor any unusual thoughts, images, or feelings that they may have experienced while in the deep state of relaxation. The doctor then helps the patients find the origin of the image (a past incident, unhealthy belief structure, or trauma, etc.) through the doctor's ability to trace the images and emotions associated with it to the patients' internal organ or organ systems involved in the creation of the illness. Once the connection is made to the patients' disease, and the Medical Qigong homework is prescribed.

Medical Qigong prescriptions initiate energy movement. This energetic movement brings about the release of trapped emotion and energetic feelings. As the emotions unravel, a new level of awareness surfaces that allows the patients to become cognizant of certain mental functions that were previously unconscious. This awareness in turn allows patients to access their creative subconscious Mind to change the conditioned programming of their Zhi Shen (Figure 17.6).

Through creative imagination, the patients' spirit is expressed and the patients' body is spiritualized. The imagination is an energetic world founded on the interrelation of various levels of spiritual, imaginative and physical realities. These energetic dimensions lie between what the Chinese call the "realm of the unfathomable hidden mystery" and the "world of animated physical forms." The imagination is an energetic and spiritual world that exists between the reality within oneself, and the reality outside of oneself. This energetic and spiritual world seems imaginary, but in fact it is very real, in that the patients live in it, and through it create a new belief system and self.

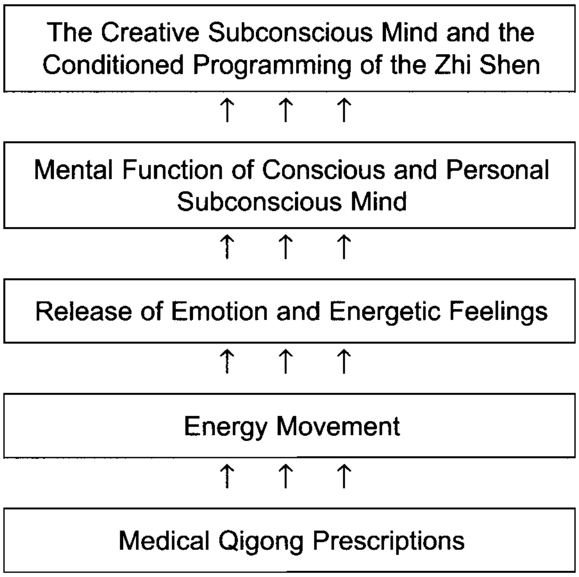
The Creative Subconscious Mind and the Conditioned Programming of the Zhi Shen
Mental Function of Conscious and Personal Subconscious Mind
Release of Emotion and Energetic Feelings
Energy Movement
Medical Qigong Prescriptions

Figure 17.6. Five stages of transition and energy awareness allow patients to access their creative subconscious energy in order to reprogram their initial belief structures.

TWELVE STAGES TO TRANSCEND AND TRANSFORM

During the Ming Dynasty there began to surface twelve specific poems accompanied with twelve pictures, describing the twelve stages to enlightenment through contemplation. In this series there are ten pictures which describe the first ten stages of mental, emotional, and spiritual transcendence, and two additional pictures which describe the last two stages of mental, emotional, and spiritual transformation. These pictures and poems use the image of the ox (water buffalo) to describe man's animal nature, which in the first stage of spiritual training is identical to man's spiritual nature. The image of a young man is used to depict human intention and will. The interaction of the ox and young man reflects the idea that by coming to terms with our own animal nature, we can move toward transformation and arrive at a new way of being.

The images of the ox changing from dark, Turbid Yin, to bright, white Yang, represent the field of consciousness transforming through the controlled deliberate intention of one's Zhi (Will). Such metaphors as capturing and controlling the ox represent an explanation of the difficulties and dangers of transforming one's self, as well as the inner changes of alchemical purification and the reconciliation with one's darker side. These pictures suggest the combination of both the sacred and the natural within Man.

These stages of controlling the mind are described as follows.

1. In the wild, the first stage of training is divided into three separate phases:
   a. The Searching for the Ox phase depicts the state in which the young man is unaware of his own true nature. Although he is not sure what he is seeking, he has begun searching for something he is desperately missing within his life. He is in a state of omnidirectional searching, without an instinctive direction.
   b. The Finding the Tracks phase depicts the young man undergoing the first traces of mental, emotional and spiritual transformation. Realizing that other people have walked along this path, the young man begins to rediscover his own spiritual instincts. He is in the beginning phase of achieving focus and direction.
   c. The First Glimpse of the Ox phase depicts the young man experiencing his first "awakening." He has the first glimpse of the true union of nature and spirit in the form of a spiritual vision. The animal nature, however, is still uncontrollable and in the wild state (due to its connection to the painful memories of the past). The young man must chase after and entice the ox, which displays its horns, bellows aloud, runs away, and is overshadowed by a dark cloud (anger, despair, grief and fear). It tramples the wheat seedlings (beginning moments of spiritual insight) wherever it goes. In this phase the ox is unruly, Yin, and pure black in color (Figure 17.7).
2. The Initial Training Begins with this next stage. The ox is controlled by a rope through its nose, and runs swiftly under the young man's whip. The young man struggles hard to control the ox's willful temper. He begins talking and listening to the ox, matching its consciousness with his own. At this stage the ox is still Yin, and pure black (Figure 17.8).
3. The Ox Comes Under Control through con-

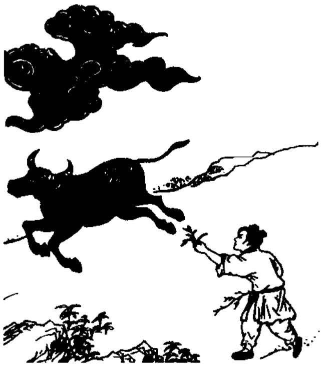
Figure 17.7. The Ox is in the Wild

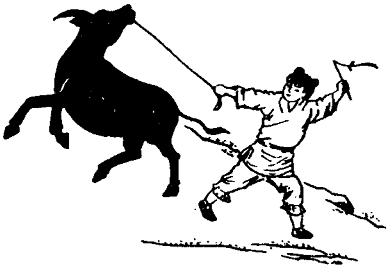
Figure 17.8. The Initial Training Begins

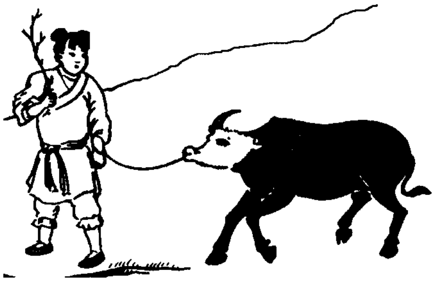
Figure 17.9. The Ox Comes Under Control

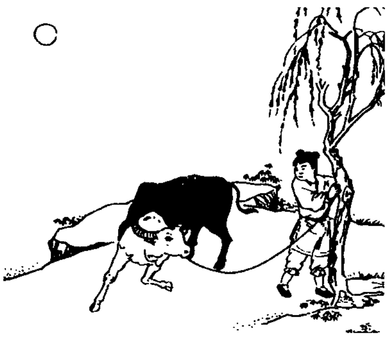
Figure 17.10. A Time of Transition

stant training. The ox stops its dashing about, and begins to follow the young man. In spite of his fatigue, the young man tends the ox all day long, not daring to loosen his control of the whip or his grip on the rope. Gaining control of the ox requires the young man to also become more vulnerable and honest, using less force to control the situation. In this stage only the ox's head, which is under control, is Yang and white, the rest of his body is Yin and black (Figure 17.9).

4. During the Time of Transition, the young man feeds and takes care of the ox with respect. He also does not abuse it. At this stage, the natural strength of the ox harmonizes with the will of the young man. A long time has passed and the restless temper of the ox gradually has turned gentle. However, still not trusting the ox, the young man maintains control of the rope. The ox is one third Yang and white and two thirds Yin and black (Figure 17.10).
5. The Ox Is Now Tamed and moves in harmony with nature. It follows the young man, who no longer uses a rope to lead the ox, but still maintains control of the whip. The ox is half Yang-white and half Yin-black (Figure 17.11).
6. The Ox Is Freed of Worldly Hindrance. The ox no longer needs the whip for control. The young man can begin to experience and enjoy the inherent rhythm of life, because he now has the ox's undivided attention. The ox is three quarters Yang-white and one quarter Yin-black (Figure 17.12).
7. The Ox Is Under Complete Control. After a long stage of struggling against its instincts, the ox is now tamed and free from worldly distractions. It drinks when it is thirsty and eats when it is hungry. The young man can now enter into deep sleep. If he so desires, he can sit atop the ox and play his flute, because the ox knows where its going and the young man does not have to direct it. The ox is now completely Yang-white (Figure 17.13).
8. The Young Man and the Ox Unite With Heaven. The white ox wanders among the heavenly clouds. The young man is free from

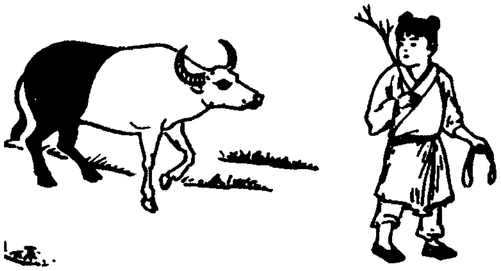

Figure 17.11. The Ox Is Now Tamed

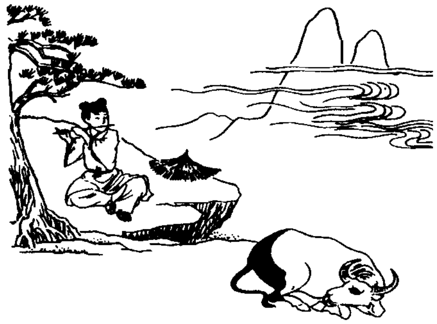

Figure 17.12. The Ox Is Freed of Worldly Hindrance

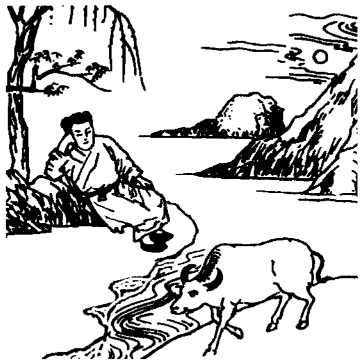

Figure 17.13. The Ox Is Under Complete Control

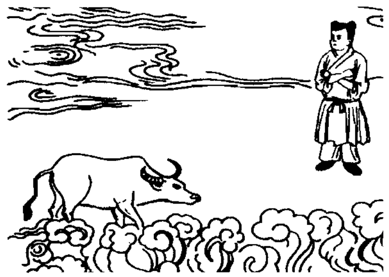

Figure 17.14. The Young Man and the Ox Unite With Heaven

worldly concerns, and so is the ox. Penetrated by moonlight, the Heavenly clouds grow whiter and drift away (Figure 17.14).

9. The Ox Transcends and A Single Light Remains. The clouds drift away and the ox vanishes. The young man finally enjoys his leisure time, singing under the moon (Figure 17.15).
10. Both the Ox and Young man Transcend and Return to the state of Wuji. As the moon continues to illuminate the vast void, both the young man and the ox vanish into Wuji and are nowhere to be found. Everything is formless, quiet, and pure. Existence is nonexistence and vice versa (Figure 17.16).
11. The Young Man and Ox Transform and Return to the Source. There is a difference between transcendence and transformation. Transcendence is a temporary or periodic experience of unity within the divine, the heavens and the earthly environment. It is a state, or condition of consciousness. Whereas transcendence implies a unifying experience that informs one of a new way of being, transformation is a process and a means to arrive at a new way of being. Mature transformation leads to an all-inclusive way of being that embraces the physical, mental, emotional, energetic, and spiritual existence of everything. In this stage man returns back to the original source of prenatal understanding, and is able to harmonize with Heaven and Earth (Figure 17.17).
12. The Young Man and Ox Transform and Live in the World. This stage depicts what happens after the transformation experience, and the reentry through nature into the everyday world. In this stage, man reenters the world as a transformed person with a new connection to his center core. He is able to become a guiding light to others. When needed, he can instantly return to a deeper state of enlightenment for additional instruction and personal guidance. Living in the world he can be in the seductions and temptations of the world, without being in any way affected by them (Figure 17.18).

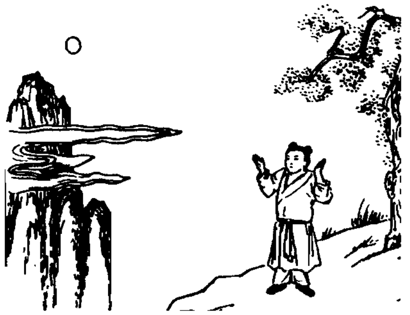

Figure 17.15. The Ox Transcends and A Single Light Remains

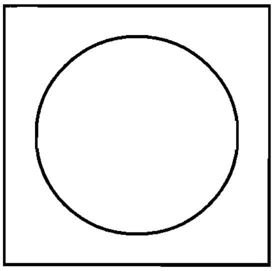

Figure 17.16. Both the Ox and Young Man Transcend and Return to the State of Wuji

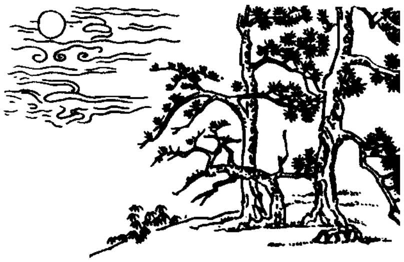

Figure 17.17. The Young Man and Ox Transform and Return to the Source

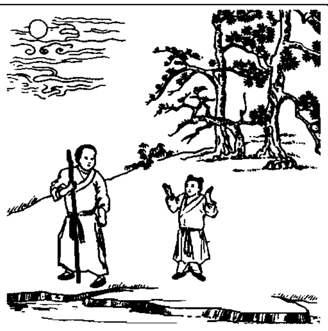

Figure 17.18. Transformation and Living in the World

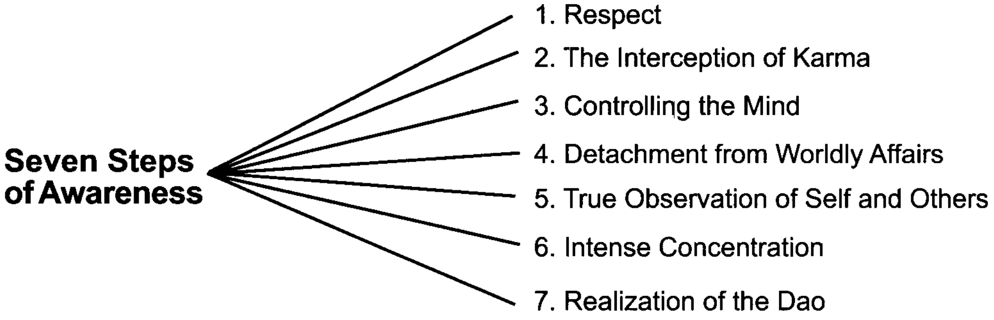
Seven Steps of Awareness
1. Respect
2. The Interception of Karma
3. Controlling the Mind
4. Detachment from Worldly Affairs
5. True Observation of Self and Others
6. Intense Concentration
7. Realization of the Dao

Figure 17.19. The Seven Steps Needed to Transform the Mind

SUMMARY

During mental, emotional and spiritual transformations, individuals undergo stages of physical and psychological purification, as they begin to experience the active, conscious introspection into their mind, body and soul. These individuals are able to eventually identify and connect their own personal will to that of the divine, and begin to lose their attachment to the ego, developing a new and more expansive identity. These individuals will also begin to see themselves as truly spiritual beings that are merely housed in fragile, physical frameworks, which are subject to all the transformations that their spirit must transcend. Generally, when using Medical Qigong as a vehicle for transformation, individuals will evolve through seven steps to awaken the spirit and reach a stage of enlightenment: Respect, The Interception of Karma, Controlling the Mind, Detachment from Worldly Affairs, True Observation of Self and Others, Intense Concentration and Realization of the Dao (Figure 17.19).

1. The first stage, Respect, encompasses the self, others and the divine. Through the cultivation of respect a stronger faith develops.
2. During the second stage, The Interception of Karma, individuals begin to accept accountability for all their actions, deeds, and thoughts. At this stage individuals receive instant karma for all their actions.
3. During the third stage, Controlling the Mind, individuals begin to discipline their conscious actions (Taming the Ox).
4. The fourth stage begins the process of Detachment from Worldly Affairs. At this stage individuals begin to develop and maintain a spiritual priority.
5. The fifth stage of True Observation of Self and Others, enables individuals to perceive themselves and others clearly. At this stage they begin perceiving the universe and environment from a deeper, more spiritually evolved understanding.
6. The sixth stage involves the development of Intense Concentration. At this stage individuals transform thought, through faith, into manifestation. They can now initiate into reality what they focus their intention on.
7. In the final stage to transformation, Realizing the Dao, individuals become aware of their relationship with the divine, not only believing in God, but personally knowing God or the Dao.

SECTION IV

QI DEVIATIONS IN QIGONG TRAINING

CHAPTER 18
INTRODUCTION TO QI DEVIATIONS

Energetic deviations occur naturally through the course of everyday life, however, the body's energetic constitution quickly corrects any disharmony in energy flow. The difficulty when cultivating Qi is not in the development of large amounts of energy, but in the creation of a constitutional system strong enough to use this energy without being damaged by it. This is one reason why Qigong doctors initially spend a lot of time developing the safety measures and internal reservoirs necessary to ensure that their body will not be damaged by creating too much energy too fast. These alterations of energetic patterns can result in feelings that can range from simple discomfort to an abnormal susceptibility to disease. Through proper Medical Qigong therapy, Qi deviations can be rectified.

Qi deviations can be precursors to disease. In recent years, medical researchers in China have reported that some patients admitted to the Medical Qigong clinics had developed mental as well as physical disorders as a result of Qi deviations. A deviation refers to abnormal phenomena occurring within the patient or Qigong practitioner. The clinical symptoms include abnormalities in perception, thinking, and behavior.

If the Qigong doctor has not been adequately trained in internal purging techniques, his or her body's increased energetic field can act as a magnet and begin to attract the toxic energy from the surrounding environment (including the patient's pathogenic Qi). Once toxic energy enters into the body, it can immediately create a Qi deviation. Qi deviations should be treated immediately if possible. If Qi deviations occur during Qigong exercises or meditations, the practitioner should stop the practice immediately and guide the Turbid Qi (and bad feelings) out from the body, discharging the toxic energy out through the four extremities and into the ground.

The process of Medical Qigong meditations and exercises is said to cause an alchemic transformation of Shi Shen (Turbid Spirit) into Yuan Shen (Original Spirit). This transformation occurs when first beginning Qigong practice, at the "laying a foundation" stage of the workout. If the Turbid Qi is not removed at the beginning of practice, the Qi concentrated in the Dantian areas (or other parts of the body) will also become Turbid. This may bring about internal disturbances when the practitioner encounters unfavorable external factors such as suddenly being startled or becoming angry. Too much concentration and hard focus on the exercises and meditations can also lead to Qi deviations. Turbid Qi can create Turbid Shen.

When Turbid Qi travels along the Governing Vessel to the head, there is a feeling of heavy pressure in the head. In severe cases, psychosis can occur. Turbid Qi that escapes into the body's channels can result in distending or numbing of the body. To avoid such side effects, it is extremely important to establish an equilibrium between the Five Yin Organs. No matter which Medical Qigong system is followed, the primary task is always to dredge or purge the channels of pathogenic Qi, to regulate the Yin and Yang organs and to cultivate the body's Jing, Qi, and Shen.

THE MAIN CAUSES OF QI DEVIATIONS

Mental disorders and personality disorders may exist in some individuals prior to their study of Medical Qigong. Sometimes, individuals with a family history of psychosis and certain other mental disorders (such as bipolar disorder and depressive disorder) may be more predisposed towards these diseases. The practice of Medical Qigong meditations and exercises may trigger the onset of symptoms in predisposed individuals, especially if Qi deviations are not properly addressed.

1. A Week Constitution	From severe disease and weak respiration
2. Improper Dao Yin Qigong Regulation	From forced practice, hastily and indiscriminately training
3. Disbelief and Suspiciousness	From a closed mind, superficial mind, or preoccupied mind
4. Too Much Mental Concentration	From concentrating too hard, resulting in Qi and Blood to stagnation

Figure 18.1. The Main Causes of Qi Deviations

The main causes believed to contribute to Qigong deviations, according to research gathered by Dr. Wei Lin Shen of Shanghai, China, include the person's constitution and pathological state, improper Dao Yin Qigong regulation, disbelief in and distrust of the Qigong exercises and meditations, or too much mental concentration (Figure 18.1).

1. A week constitution can lead to a pathological condition. Patients with a weak constitution should not be required to sit or stand for long periods of time. Patients with severe disease and weak respiration incur Qi deviations if they are allowed to stand too long.
2. Improper Dao Yin Qigong regulation can result in Qi deviations. Qigong deviations may also occur as a result of forced practice (when the state of illness does not allow relaxation), or when individuals hastily and indiscriminately train in the techniques that circulate internal Qi. Lack of proper instruction from a qualified master of Medical Qigong therapy often results in a failure to train in the correct way. The failure to correctly apply the three Dao Yin principles and methods (proper regulation of the body's Jing, breath's Qi, and mind's Shen) in training practice, for example, leads to Qi deviations. Creating new techniques, training recklessly before mastering the primary techniques, or changing the techniques at random can also result in Qi deviations. For this reason the practitioners should be patient when practicing Dao Yin regulations. The goal is to achieve a tranquil, quiet, and empty state.
3. Disbelief, and suspiciousness can induce Qi deviations. Disbelief or an inability to adopt a correct attitude toward Medical Qigong effects (such as extreme suspiciousness), results in mental tendencies that will become deviations over the course of time. Qi deviations can result from: A Closed Mind, A Superficial Mind, and A Preoccupied Mind.
    a. A closed mind, full of fear, pride or bitterness may prevent the patients from connecting with their True Self.
    b. A superficial mind, or lack of commitment and belief in energetic healing, may cause the patient to fail in taking the Medical Qigong prescriptions seriously.
    c. A preoccupied mind may cause the patient's spirit (Hun) to "wander off" while practicing Medical Qigong therapy.
4. Too much mental concentration is the most frequent cause of Qi deviations. The strong focus of mental attention on the flow of Qi can distort its path and lead to Qi deviations. It can cause too much concentrated Qi and Blood to flow into the brain. Concentration on internal objects with too intense a focus can actually cause a stagnation, or Excess of

Qi in the area (e.g., concentrating too hard on a tumor can actually cause it to grow instead of dispersing it). Mild concentration is always preferred when practicing Medical Qigong exercises and meditations. A relaxed mind and body allows the Qi to flow freely.

CORRECTING QI DEVIATIONS

To correct any Qi deviation, it is important to understand the root cause of the deviation. The Medical Qigong doctor should check for the following causes of Qi deviations.

1. Postural Dao Yin deviations result from incorrect postural changes.
2. Respiratory Dao Yin deviations result from improper breathing methods.
3. Mental Dao Yin deviations result from improper mental activities and emotional disturbances. These emotional disturbances (due to the accumulation of toxic emotions within the internal organs) may manifest as spontaneous emotional outbursts.
4. Stressful life activities resulting from an improper balance of work and play, lack of sleep, or an improper diet may likewise cause Qi deviations.
5. Environmental disturbances during Medical Qigong practice such as the phone or doorbell ringing, may startle the practitioner and cause Qi deviations.

To correct the Qi deviation, the Qigong doctor addresses each possible cause separately, beginning with the patient's posture.

POSTURAL DAO YIN DEVIATIONS

To diagnose Qi deviations due to incorrect postural alignment it is important to observe whether:

- relaxation is being achieved,
- the posture is structurally correct to facilitate relaxation, and
- the appropriate tonifying, purging, or regulating technique is being used.

When a deviation occurs, the patient should examine it in accordance with the basic laws of postural alignment. A beginner is more prone to some disorders due to incorrect postures that cause headaches, dizziness, and a stiff neck.

MANIFESTATIONS OF POSTURAL QI DEVIATIONS

If postural deviations occur, first review the Eighteen Rules of Proper Medical Qigong Structure in relationship to the sitting or standing positions, and then review the other Dao Yin Qigong patterns (respiratory and mental Dao Yin training). The following are examples of specific Qi deviations related to improper postural alignment.

1. Eye distention results from excessively staring at external objects during the meditations and exercises. To prevent this type of Qi deviation, encourage the patients to always use a soft focus.
2. A stiff or painful neck is usually caused by forcefully attempting to suspend the head. To prevent this type of Qi deviation, have the patient relax the neck muscles and sink the shoulders.
3. Shoulder and back pain can occur when patients fail to relax the muscles of the shoulders and back, stretch the back excessively, hollow their chest too much, or hold an unfamiliar posture too long. To prevent this type of Qi deviation, have the patients relax and sink their shoulders and imagine the back melting down into the hips and legs.
4. Waist and hip pain (Lumbago) come from several factors.
    a. Tension in the hip muscles causes pain in the hips that extends to the waist.
    b. Twisting the waist without first relaxing the hips can result in both hip and waist pain.
    c. Forcefully twisting the waist may also cause pain.

    To prevent this type of Qi deviation, have the patients relax their waist and hips and sink the energy into the Earth via the feet.
5. Abdomen and lower extremity distention or flatulence results from the forceful pulling-in of the Stomach, or too much concentration of the mind on the lower extremities. To prevent this type of Qi deviation, have the patients relax their abdomen and lower extremities, allowing the Qi to sink naturally.
6. A swollen and painful anus, or dry stool, can

result from raising the anus too forcefully. Also, hemorrhoids can develope from holding the breath too forcefully while focusing on the lower abdominal area during the Abdominal Breathing Holding method. To prevent this type of Qi deviation, have the patients close their anal sphincter softly and relax their buttock muscles.

7. Sore knees are a result of crouching too low, over-bending the knees, or extending them past the toes. To prevent this type of Qi deviation, have the patients keep their knees in proper alignment and guide the pressure of the body weight to flow behind the legs and not in front.
8. Heel pain can be caused by shifting the body's weight excessively onto the heels during walking or standing practice. Heel-ache can also occur if the patients focus too much attention on their heels instead of the center of their feet for balance. Knocking, stomping, or excessively dropping the heels against the ground, as well as sexual Excess, or a Kidney Deficiency can all cause pain in the heels. To prevent this type of Qi deviation, have the patients relax their legs and waist and imagine that their feet are kneading the ground. This squashing and kneading action is used to encourage the free flow of Yin Qi from the Earth into the body.
9. Improper hand movements occur when patients fail to distinguish between Yin and Yang energetic properties, or apply either too much or too little strength in the raising and lowering of the arm movements during Qigong practice. Both the alignment and energetic function of the patients' arms, hands, and fingers should be checked.
   a. If the arm movements are too dynamic, they will cause an abundant amount of energy to become either gathered or dispersed; this results in Qi deviations.
   b. While performing a tonification exercise, if the palms of the hands are turned outward, away from the body (instead of inside facing the body) this can easily cause diarrhea or a Qi Deficiency.
   c. If the fingers move up to point at the chest, the energy released from the hands can cause tightness in the patients' chest. If the patients' hands point obliquely at their neck and face, the emitted Qi can cause dizziness, nausea, and edema of the face in those patients who are sensitive to the energy movement within their channels and collaterals.

RESPIRATORY DAO YIN DEVIATION

Deviations in breathing often result from the incorrect practice of the deep inhalation and exhalation methods, when first starting Qigong meditations. Both inhaling and exhaling should be gentle, thin, even, and long. When practicing, the patient should breathe properly and naturally to regulate the breath and should avoid holding the breath.

1. Deviations can be due to improper exhalation. Determine whether the exhalation is long enough for relaxation and quiescence when sinking the breath. Excessive sinking of the breath, however, can cause pain in the legs by depleting Qi flow to the upper torso, and by increasing the weight and gravitational pressure within the tissues of the lower torso. Exhalations that are too long or too deep can cause shortness of breath, headaches, tightness in the chest, discomfort in the Heart, and abdominal distention.
2. Deviations can be due to improper inhalation. Determine whether the inhalation is long enough for relaxation and quiescence. For example, when practicing the Abdominal Breath Holding technique (the Two Inhaling and One Exhaling Method, which employs Reverse Breathing), the two inhalations should not be connected, for tightness in the chest will occur. Pause, therefore, between each breath and allow the exhalation to be fast and short rather than sustained and drawn out. Otherwise, the condition will cause general weakness and fatigue.

To rectify this condition, when you inhale twice through the nose, simultaneously pull in the navel and perineum area as if to lift your body. When you exhale through the nose, guide the Qi down into the Middle Dantian.

If this method is applied, the symptoms such as tightness in the chest, shortness of breath, and abdominal distention will disappear (The Rectifying Respiratory Qi Deviations exercise can also be used). When the breath is well regulated there will be no difficulty sinking the Qi down to the Lower Dantian.

3. Deviations can be due to improper tone resonation. Another Respiratory and Mental Dao Yin deviation can be created through the improper use of healing words or tone resonations. The improper use of tones (with too much breath or mental concentration placed on the rising and falling sound) may cause symptoms such as dizziness and slight fever, tightness in the chest, weakness in the legs, foul breath, and bloody stool.

These symptoms cannot be relieved by Western medicine. Some patients, however, have succeeded in rectifying such deviations by practicing the Opening and Closing the Three Dantians exercise (see Chapter 15). This exercise is used to regulate the Excess Yang, and to restore the body's energetic equilibrium.

Some patients are given the prescription of uttering one pitch sounds (known in China as the first or straight tone). This straight sound has a steady tone, and a rising volume, which can easily cause the energy to ascend. After the patients have completed their prescription and ended the exercise, if the energy continues to rise, the Qi deviation will cause dizziness and nausea. In this particular case, the patients should stop the exercise immediately. To rectify this condition, have the patients utter the sound in the third tone (the descending and ascending tone), and by deepening his or her voice, then raising the pitch. This increasing volume and deepening tone relaxes the body and causes the Qi to descend.

This example illustrates the fact that if the patient is allowed to select any prescription at random for mental concentration, without understanding the theory of Medical Qigong, he or she can easily get into trouble.

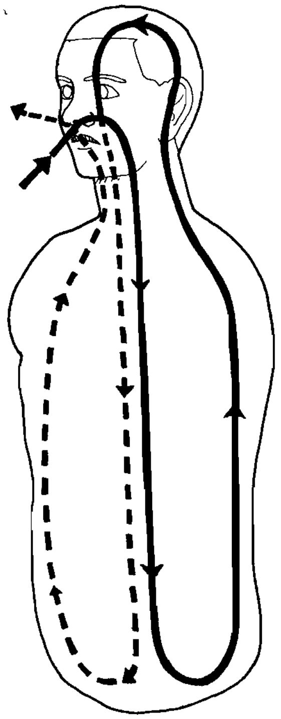

Figure 18.2. Rectifying Respiratory Qi Deviations

4. To rectify respiratory Qi deviations, the following exercise can be used. It will amend respiratory deviations and balance the Yin and Yang Qi, as well as clear any stagnant Qi.

When the practitioner feels uncomfortable after Medical Qigong exercises, practice this method for 20 minutes. This is an alternate nostril breathing exercise and requires both inhaling and exhaling through the nose. The abdomen expands upon inhalation and contracts upon exhalation.

From a seated posture, begin by imagining that there are three channels that start at your tailbone and travel up the body. The first channel intersects with the right nostril, the second channel joins the Baihui GV-20 point and the third channel intersects with the left nostril (Figure 18.2).

Place your left hand on your Lower Dantian, with the right hand resting on the nose. It is important to focus the mind on the energetic flow of the breath. Inhale through the left nostril and allow the Qi to flow down the Conception Vessel into the Lower Dantian and tailbone area. Hold the breath, pinch the nose and draw the Qi up the Governing Vessel into the Baihui area, and then circle the energy down to the Conception Vessel to the coccyx to complete the Microcosmic Orbit. Exhale through the right nostril. Repeat this sequence, alternating the focus of the mind's concentration to follow the beginning flow of energy from the left to right nostril and visa versa.

When training in Medical Qigong prescriptions, it is equally as important to relax the mind as the body. Therefore, this exercise should be practiced in a leisurely manner, with a carefree attitude, and a smile.

MENTAL DAO YIN DEVIATIONS

Deviations in mental activities are primarily caused from excessive thinking and too much focused concentration. These type of mental deviations should be rectified by controlling one's mental activities. It is important to fully understand the foundational principle of the Shen controlling all mental activities that determine the physical, emotional, and energetic activities. Any mental deviation should be regulated and controlled through intention.

It is sometimes difficult for patients who have just started Medical Qigong prescriptions to become absolutely tranquil; this may lead to mental Qi deviations. One way mental Dao Yin deviations can be rectified is by a mantra, or counting breaths (using one thought to replace many thoughts). After this technique is mastered the patients can advance to concentrating their mind on an external object in the surrounding area.

1. The cerebral cortex plays an important role in Qi deviations of the mind. Medical Qigong regulates the cerebral cortex to restore balance to the energetically unbalanced parts of the brain. If the mind is agitated or depressed for a long time, it can cause a long term disorder of the whole nervous system and, thus, cause various kinds of chronic diseases. A healthy nervous system takes the leading role in adjusting, regulating, and unifying all functional activities. It is responsible for maintaining a dynamic equilibrium.

   If deviations occur during Medical Qigong practice, the cerebral cortex loses control of the nervous system. Abnormal phenomena may occur and mental and emotional activities may become uncontrollable. If patients are not careful, they can actually develop meditation induced diseases, or meditation induced psychosis known as Tsou Huo Ru Mo, which translates to "the Fire becomes Excessive and the demons enter." In other words, the patient's Heart and Liver become overheated, causing the Hun to leave and Po to take control of the body. This in turn can lead to chronic psychoses or demon possession.

   Because Qi deviations can have serious consequences, the Qigong doctor must have an extensive knowledge of Medical Qigong principles to prevent and correct Qi deviations, stagnations, and adverse Qi flow. The mind should, for example, always be free from over-concentration or "distortions of consciousness" during practice. If a beginning patient feels a hot sensation in certain parts of the spine, and tries to force its energetic movement up through the Five Passes of the Governing Vessel (see Chapter 30), this can cause "distortions of consciousness" and mental confusion, which may result in hallucinations.

   A deviation caused by excessive focused mental concentration can also lead to a loss of self-control and produce such disorders as uncontrollable head-shaking, shoulder shrugging, hand and leg quivering, stumbling, staggering, uncontrolled crying or laughing, or even stiff and twisted extremities and opisthotonosis. To avoid this loss of self-control, do not allow the patients to use their mental intention and focused concentration at random, but rather have them focus their attention on quiescence to tranquilize the mind and recuperate their health.

Failure to properly lead and guide energy during Medical Qigong practice may also cause adverse deviations of Qi in the patient's Middle Dantian. In the Middle Dantian, the patient's Lung Qi rises and returns through the opening and closing actions of the Shan Zhong CV-17 point. If the Qi within the Middle Dantian begins to deviate, it can condense and cause stiffness, pain, and a feeling of suffocation in the chest region. To correct this condition the Qigong doctor extends energy into the patient's Middle Dantian to regulate and tonify the patient's Qi, and especially the Gathering (Zong) Qi. This treatment also regulates and tonifies the patient's Lungs and Upper Burner, as well as diffuses any Rebellious Lung Qi (see Chapter 22) as it unburdens the emotions of the Heart.

2. There are three rules for determining proper mental focus. The following three rules determine the focus of the patient's mental activities and are implemented to help avoid Qi deviations while practicing Mental Dao Yin training:
    a. Concentrate the mind on a motionless object. A patient suffering from hypertension can, for example, concentrate on still lake water, but not on a stormy sea.
    b. Concentrate the mind on a near object. A patient with Liver trouble can, for example, concentrate on a nearby pine tree. If the patient is at a park, absorbed in concentration on a pine tree that is far away, the patient must mentally trace and locate the pine tree with painstaking effort. This is achieved through expanding and extending the imagination. If the patient suddenly becomes mentally distracted or preoccupied, the purpose of the meditation is defeated. Once distracted, it will be very difficult for the patient to reconnect to the pine tree because of its distance. As a result, the patient's mind wanders and the patient may become discouraged.
    c. Concentrate the mind on a familiar object. Concentrate on something which creates a peaceful, emotional connection. The emotional connection will facilitate the proper attitude and diminish any likelihood of mental strain.

    When deciding on which familiar object to focus the attention on, patients must not choose a person, their own work or creations, or obscene things. Objects that have been used to do harm or evil should also be avoided. Patients are also forbidden to shift their thoughts from one thing to another or back and forth without stopping. Remember that it is the individual's Shen (Spirit) which guides all mental activity and determines the body's energetic actions, therefore the mental concentration should have purpose.

Adjust each technique according to the specific case. When prescribing Medical Qigong therapy, it is important for the doctor to adjust each meditation and exercise according to the patient's specific case. In all cases, the Yi (imagination and intention) must lead the Shen and the Shen must lead the Qi.

The following examples demonstrate how to adjust prescriptions for different patient's needs.

- Patients with hypertension should concentrate on a low place instead of a high place, or on an object below eye level.
- Patients with hypotension (including other Deficient conditions such as anemia) should concentrate on a high place instead of a low place, or on an object placed above eye level.
- Patients with Lung problems such as tuberculosis should concentrate on white and light objects (e.g., white clouds).
- Patients with Spleen and Stomach diseases should concentrate on yellow objects (e.g., yellow chrysanthemum, dahlia).
- Patients with Liver problems like hepatitis should concentrate on green objects (e.g., pine and cypress trees).
- Patients with a Heart Deficiency should concentrate on red objects. Patients with an Excess conditions of the Heart, however, should

never concentrate on the color red, and should only choose pink objects instead. A pure red color is too energetic for patients with Excess conditions and may overexcite their Heart, making it difficult for the patient to become relaxed.

- In Medical Qigong, the color black is sometimes assigned to the Kidneys. Focussing on this color can, however, easily cause the Kidney energy to sink downward, creating an emotional state of depression or anxiety. Therefore, when treating patients with Kidney diseases, it is preferable to choose things of purple, indigo and vibrant blue color.
- Mentally depressed patients (in a severe cases) should not practice any dynamic Qigong exercises until their depression has passed. Otherwise, they may develop headaches, tightness in the chest, or in severe cases, shock. To rectify this condition, the Qigong doctor must lead the toxic and stagnant Qi downward along the channels and out of the body.
- Patients with uncomfortable light or heavy sensations should concentrate on the opposite side on their body instead of the afflicted area. If they feel that their lower limbs are too heavy to lift during the exercises, for example, they can correct this symptom by focusing on the Baihui GV-20 point at the top of the head. If the patients feel that their body is too light and unsteady, they can focus on the Yongquan Kd-1 point located on the soles of the feet.

STRESSFUL LIFE ACTIVITIES

Stressful life activities create a flood of adrenaline within the body. This occurs not only from the core of the adrenal gland in the medulla (which responds to stress by releasing adrenaline in to the blood stream), but also from the endings of the vertebral lumbar division at the core segment of the major internal organs. Adrenaline spreads throughout the body and activates not only the body's sympathetic nervous system, but also the adrenal glands. The Heart rate, respiratory rate, blood pressure, and Blood flow to the muscles are all increased as they respond to the "fight or flight syndrome."

When this happens, the immune system is called upon to balance the whole stress mechanism and is, therefore, called away from its normal activities which include monitoring cellular activity and fighting the growth of disease. A life of constant stress, therefore increases susceptibility to illness. Meditation takes the patient out of the stress mode, thereby shutting down the adrenal glands, and allowing the immune system to return to its normal functioning.

Chronic stress wears down the mind and body, thereby producing anxiety or depression. Stress begins with an overexcitement of the nervous system and then slowly works its way deep into the body like wet cement, causing the shoulders and upper back to become tense and rigid. Happy occasions (such as weddings) and other normally pleasant events, also produce stress, which can be detrimental when the patient's life style is not properly balanced. An unstable life style is very stressful and can lead to depression, anxiety, and Qi deviations. The capacity to bear stress depends upon the individual's constitution, diet, support system (or lack thereof), the nature of the stress, and the congenital strength of the nerves. When stress levels surpass the nervous system's capacity to handle them, "the nerves break down", resulting in all sorts of physical, mental, and emotional disturbances. Certain neurotransmitters are depleted, leading to a chemical imbalance which can be aggravated by self medication (e.g., alcohol or drug abuse). These disturbances, if unheeded can eventually lead to organ malfunctions, cysts, tumors, cancer, and premature death. In order to avoid such states, a balance between work and rest must be maintained.

At a conference of the International Society for Neuro-immunomodulation (entitled *The Effects of Stress and Depression on Physical Disease*) Dr. Philip Gold, of the National Institute of Mental Health, stated that stress and depression cause the body to secrete high levels of hormones; (for example, cortisol) into the bloodstream. These hormones, if unregulated, can destroy the appetite, cripple the immune system, shut down the processes that repair tissue, inhibit sleeping patterns, breakdown bone density, aid the onset of infection, and even instigate cancer cell growth.

An individual's emotional past determines his or her present state of health, both physical and emotional. Past emotional belief structures are responsible for both creating and healing diseases. Just as thoughts drift through the mind but do not manifest until they reside within a belief structure, so emotional wounding resides within the tissues and cells waiting to be activated by emotional upheavals. Our thoughts materialize via choices and manifest within the body.

In American society, sharing traumatic wounds is often the first act of intimacy and bonding. It provides a common ground for relationships and supports the "pain structure" that deepens the bond. People can become addicted to empowering their wounds which, in turn, support and empower the disease. This exchange of painful personal traumas is reinforced by receiving sympathy and strong empathy, but does not fully release the pain from the tissues.

Although talking about past traumas can begin the process of healing, problems can emerge when these past traumas manifest through the following belief structures.

1. The individuals believe that they can only receive affection and attention when they suffer, and use their "neediness" to keep their significant other(s) from leaving or becoming more independent.
2. The individuals become stuck in one emotion -- anger, grief, fear, worry, or hurt -- and are unable to express a healthy range of emotions.
3. The individuals do not have an adequate support system.
4. The individuals do not have the incentive to heal themselves or the belief that they can be healed.
5. The individuals bond to anyone, out of fear of being alone, as long as they receive the attention they crave.
6. The individuals have low self-esteem and do not seek help or work out their emotional issues; they have difficulty trusting anyone to understand their pain.

The desire for sharing ones pain in order to feel loved can lead to a habitual way of relating to those closest and dearest. When such individuals feel taken for granted, or unappreciated, they may immediately resort to past traumas. If this fails to illicit the desired response, a new trauma or wounding may be unconsciously created to get back the attention and love and need. The individuals may, for example, physically injure themselves in an accident, or manifest a physical illness. This in turn can become an addiction, empowering old wounds and creating new ones, and may lead to chronic illness. The Shen becomes divided to support the new belief structure, and the Qi deviates from its assigned job of protecting the body, thus creating illness. This is not a conscious process, but an unconscious one, which needs to be brought into the patient's awareness to be healed. Then, and only then, can the Yuan Shen and Zhi Shen act as one unit directing the body's Qi towards its proper function of healing.

Because energetic healing occurs in the present and not in the past, victims of past traumas will never heal if their energy continuously regresses to the armored traumatic belief structure. Both the doctor's and the patient's Yi (Intent) must therefore, focus on forgiveness (of self, others and the specific tragic situations) to free the patient from toxic wounds. The patient's desire and determination to be well must reach the degree that they are willing to do anything to get well, including giving up on the "secondary gains" of illness (i.e., risking abandonment, or loss of attention received during the illness). Patients must be encouraged to find new ways to fulfill their needs, and must be encouraged and supported through this difficult transition. This point is illustrated effectively in the following traditional Chinese story.

> There once was a palace official named Cheng Guang Wei who owned a servant named Song Li. As the pressures of life wore on Director Cheng, he began to sink into a deep depression. Soon he began to lose all hope in life and desperately sought relief. Not content with the wine and festivities at hand, his attention soon turned to his servant Song Li, who always seemed to be happy. Song Li was plowing the fields, up to his ankles in

mud, singing and seemingly enjoying his life. Upon observing this, Director Cheng walked up to the edge of the field and called out, "Song Li, why are you so happy? What is the secret of your contentment?" To which Song Li responded, "Sir, come into the field with me, and I will tell you." The director, not wanting to get himself muddy said, "No, tell me from where you are." To which Song Li replied, "I cannot explain everything to you if you do not come into the field." Director Cheng, aggravated by this reply, answered, "Very well, since you refuse to tell me, tomorrow you will work in the stables." Song Li answered, "If that is your wish, sir," and went about his work contentedly.

Two weeks went by, and the pressures of the palace were starting to take a serious toll on Director Cheng. All of his problems, it seemed, were insurmountable. Locked in turmoil, he began to roam his property seeking peace until he heard singing coming from one of the stalls in the stables. Director Cheng went to investigate and sure enough, it was Song Li, now up to his knees in horse manure and still singing as beautifully and joyfully as a bird. At first Director Cheng was enraged at Song Li's contentment, but then became perplexed, because the duties of Song Li's new job were much more difficult than the previous work.

Director Cheng approached the edge of the stables and called, "Song Li, why are you so happy? What is the secret of this contentment you possess?" To which Song Li responded, "Sir come into the stall with me, and I will tell you." Repulsed by the stench of the horse manure and filth, not to mention feeling protective of his royal robe, Director Cheng answered, "Song Li, tell me the secret of this joy and contentment that you possess even in the midst of a barn full of stench and dung." Song Li answered, "I will tell you sir, but first you must come in to where I am." "If that is your answer, then tomorrow you will work in the hog pens!" shouted Director Cheng. To which Song Li answered, "If that is your wish, sir," and went about his work contentedly.

Well, things went from bad to worse for Director Cheng in the palace; seeing that he was about to lose everything, he contemplated taking his own life. Making one last round of his property, he again heard singing, this time from the hog pens. Realizing it was Song Li, Director Cheng slowly approached. Sure enough, there was Song Li, up to his thighs in pig dung, slop and filth, and still singing. Leaning against the fence, Director Cheng called out, "Song Li, why are you so happy? What is the secret of your contentment and joy?" To which Song Li responded, "Sir, come into the pen with me, and I will tell you." Director Cheng, now at the end of his rope, answered, "Very well, I will come in with you," and began to climb the fence to enter the hog pen. "Stop!" shouted Song Li, "I will come over to you, it is not right that you defile yourself on my account. But now you will truly listen to what I have to share and are ready to hear the words I speak. For now you are willing to sacrifice all, even your physical comfort, to receive spiritual knowledge and insight."

The moral of the story is clear: unless we are ready to abandon all masks (pretenses and defence mechanisms) of the emotional ego, it is extremely difficult to free ourselves from the toxic emotions that steal our life-force energy and direct our lives. If we are not intimate with our emotions, we cannot perceive the dynamics behind those emotions. After all, emotions are only energy currents which flow in, out, and through our bodies. Because our emotions reflect our intentions, any awareness of unprocessed emotions will lead to an awareness of deep-seated intentions. It is through this understanding that Qigong doctors train their emotions and mental thinking to regulate the Shen.

SHEN DISTURBANCES AND EMOTIONAL DYSFUNCTIONS

All Shen disturbances relate to emotional energetic dysfunctions. Generally, Shen disturbances can be divided into two categories: Yin disturbances and Yang disturbances (Figure 18.3). These

Yin and Yang disturbances affect the patient's thoughts and emotions.

A thought is energy that has been shaped by consciousness. Every experience, or change in experience, reflects intention, desire, and will. Any discrepancy between intention and emotions leads to a splintering reaction of the energetic self, which can cause a breakdown of the body's life-force energy and ultimately lead to disease.

Emotional energetic currents have different frequencies, resonating between low (fear, anger, worry, etc.) and high (love, forgiveness, joy, compassion, etc.). Lower frequencies deplete the physical body, draining it of its precious reserve of life-force energy. This deterioration begins to manifest in the spiritual fields of energy surrounding the body, and percolates down to the emotional, mental, and finally the physical level. Conversely, high frequency emotional energy raises the body's energetic frequency and begins to radiate outside the physical body. This does not mean that the patient must learn to suppress feelings of anger, hurt, etc. These are natural emotions that serve a positive function in protecting the individual from harm. It is only when emotions are suppressed, denied, and accumulated that they become highly toxic to the body. When this happens, new anger is fueled by old anger. The new hurts are suddenly blown out of proportion (related to the event), fueled by ancient emotional pain, until the patient's physical structure is overwhelmed.

Patients may become overactive and overemotional when under stress, when their energy becomes disrupted. Every emotional disturbance alters the breathing pattern, changing from slower to faster respirations (each breathing pattern varying according to the internal organs involved).

According to *The Yellow Emperor's Inner Cannon* (Spiritual Axis), the Blood, Ying, Jing, Qi, and Shen are stored in the Five Yin Organs (Liver, Heart, Spleen, Lungs, and Kidneys). An Excess condition will cause any of these components to leave their respected organs. This will:

- deplete the patient's Jing (Essence),
- cause the patient's Hun (Ethereal Soul) and Po (Corporeal Soul) to become unsettled,
- cause the patient's Zhi (Will) and Yi (Intention) to become muddled, and
- cause the patient's Shen (Spirit) to leave the patient's body.

Shen Disturbances	
Yin	**Yang**
Qi is compressed.	Qi is expanded.
Patient is depressed, withdrawn.	Patient is manic, expressive, impulsive, volatile.
Patient is too quiet.	Patient is too loud.
Patient avoids feelings.	Patient inappropriately expresses feelings.
Patient is too introverted.	Patient is too extroverted.

Figure 18.3. Shen Disturbances

When emotionally upset, energy literally rises upward, affecting the organs and releasing additional emotions stored within the tissues. To become free from these emotional reactions, patients must be taught to relax and release the energetic charge of emotions. This will allow stagnant Qi to become energetically discharged from the body. By cultivating a healthy attitude, and not holding onto, or suppressing, the emotions, the body will begin to seek its energetic balance naturally. This is initiated by having the patient sink his or her Mind and breath deep into the Lower Dantian to rebalance the life-force energy.

TREATMENT TECHNIQUES

When treating patients with Shen disturbances, the Qigong doctor categorizes the emotional and thought dysfunctions into Yin and Yang Shen disturbances.

1. In Yin Shen disturbances the patient's energy compresses inward. These patients have a tendency to avoid feelings, becoming quiet, withdrawn, and depressed. They also tend towards an introverted personality and posture.

   When treating patients with a Yin Shen disturbance it is important for the Qigong doctor to redirect the patient's Shen towards specific actions, and encourage them to use spiritual mantras and affirmations to gain inner strength and confidence.
2. In Yang Shen disturbances the patient's energy expands outward. These patients have a tendency to express their feelings, are loud, impulsive, volatile, and manic. They also tend to be extroverted in personality and posture.

   When treating patients with a Yang Shen disturbance it is important for the Qigong doctor to redirect the Shen (with a "reality check" towards the patient's actions) and set parameters for their boundaries. A sudden awakening to the responsibility of personal actions will generally send the patient into emotional shock.

   One treatment technique commonly used in the clinic to calm and restore harmony in the patient's Shen (due to a Deficient condition) is for the Qigong doctor to extend Qi into the patient's Middle Dantian and Yellow Court areas. Once this area has been sufficiently balanced, the doctor then ends the treatment by rooting the patient's Qi, by drawing it into the Lower Dantian.

PRESCRIPTIONS AND HOMEWORK

Several of the most recommended meditations used for the purpose of calming the patient's Shen are listed as follows.

1. The Energy Melting Meditation is used to relax the body and reduce stress. Relax and imagine melting from the top of the head to the bottom of the feet, like ice melting into water and water flowing downward into the Earth. This image centers, grounds and roots the Mind.

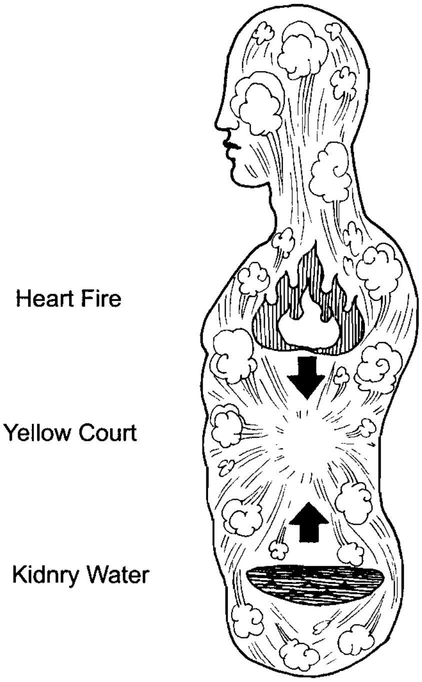

Figure 18.4. During the exchange of Fire and Water meditation, the fusion of Heart Fire and Kidney Water creates steam.

   If the patients are lying down, have them feel the upper layers of their body melting through the lower layers and into the Earth. Ice melts into water and pours into the Earth.
2. The Exchange of Fire and Water Meditation unifies the energies of the Kidneys (Water) and the Heart (Fire) and is often called the Fusion of Kan and Li. It is used to unify opposite principles within the body, to balance the energies of the mind, body, and Prenatal and Postnatal Shen for creating wholeness. It also strengthens the immune system and the bones, thus slowing or preventing osteoporosis.

   Sit and begin by breathing naturally, while inhaling and exhaling through the nose. Visualize the Lower Dantian as an ocean of water, and the Heart as a ball of fire. Imagine the waters of the Lower Dantian beginning to flow up the center of the body, reversing their natural flow. Now imagine the Fire of the Heart beginning to descend the center of the body, reversing its natural flow. As the Water and the Fire join in the Yellow Court (located in the lower part of the solar plexus), the Water vaporizes and becomes steam (Figure 18.4). This hot mist travels throughout the entire body. It should feel warm and pleasant.

Next allow the hot mist to enter into your bones and flow through the Marrow. Imagine this mist beginning to solidify and harden making your bones solid and radiant with white light. Upon completion of the meditation, relax and allow the images to disperse as you sink into the Wuji.

These following exercises are a series of meditations used in the clinic for detoxifying emotional stasis.

3. The Sun and Moon Meditation is used to detoxify emotional stagnation and establish physical, emotional, and spiritual clarity.

   Begin from a sitting or Wuji posture (Figure 18.5). Inhale and exhale through the nose. Imagine the sun over the left eye, and the moon over the right eye, appearing side by side over the head. Feel their presence over the Baihui point, at the top of the head. Imagine that the sun (on the left side of the body) pours golden light, and the moon (on the right side of the body) pours silver light. Both streams of light meet at the Baihui point on the top of the head, combining into white light energy. This white light energy pours into the body, filling the entire body from the feet to the top of the head (like a pitcher of water filling a glass). Once the body is completely full, the energy begins to spill out of the pores. At first a thick, black, sticky energy pours out of the pores, and flows down to the ground. This energy contains all the physical, mental, emotional, and spiritual toxins that are being released from the body. As more and more of these toxic substances leave the body, they start turning from black to gray, then from gray to white. Feel the body radiate this white energy in all six directions (front, back, right, left, up, and down) filling the entire room. After several minutes of purging and tonifying the body, relax the mind and focus your attention on to the Lower Dantian, before ending the meditation.

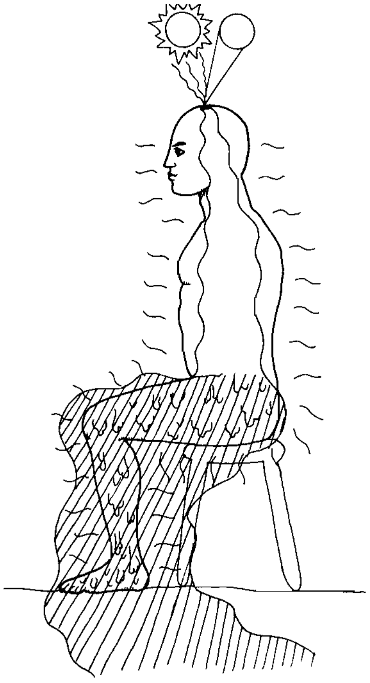

Figure 18.5. During the Sun and Moon Meditation imagine that the white light energy that is pouring down the body, is purging toxins from the internal organs, releasing the Evil Qi out from the pores.

4. The Releasing Emotional Blockages and Energetic Armoring Meditation is used to dissolve emotional blockages from the internal organs. When these blockages are released, the emotions are restored to balance. Begin from a Wuji posture, and inhale through the nose and exhale through the mouth.
   a. Direct your attention to your Liver. Every time you inhale, visualize and feel the emotions of kindness and compassion entering into your Liver. As you exhale, release anger or jealousy. Repeat for 18 breaths.
   b. Direct your attention to your Heart. Every time you inhale, visualize and feel the emotions of joy, contentment, and tranquillity entering into your Heart. As

you exhale, release nervousness and excitement. Repeat for 18 breaths.

c. Direct your attention to your Spleen. Every time you inhale, visualize and feel the emotions of trust, openness, and sincerity entering into your Spleen. As you exhale release worry, obsessiveness, and self-doubt. Repeat for 18 breaths.

d. Direct your attention to your Lungs. Every time you inhale, visualize and feel the emotions of dignity, integrity, and courage entering into your Lungs. As you exhale release grief and anxiety. Repeat for 18 breaths.

e. Direct your attention to your Kidneys. Every time you inhale, visualize and feel the emotions of self-confidence and inner strength entering into your Kidneys. As you exhale release fear and loneliness. Repeat for 18 breaths.

f. Finally, inhale, visualize and feel the emotion of a full-bodied state of peace and bliss. As you exhale, ripple and resonate this feeling into the room, filling the surrounding energetic space with this emotional enlightenment. Repeat for 18 breaths.

DISTURBANCES DURING QIGONG PRACTICE

The patient should carefully choose a safe and quiet environment for Medical Qigong exercises and meditations. Most often a patient is already in a state of relaxation when they become startled by disturbances generally caused by an unexpected sound or event. This sudden disturbance can disrupt the Heart Qi. When training, if you become startled, do not open your eyes. If you open your eyes, the disturbed energy will be thrown into disarray, causing a blockage of vital Qi (usually in the Heart). This reaction will cause deviations which can be difficult to rectify.

If instead you continue exercising with your eyes shut as if nothing had happened, you can restore your mind to a normal state of quiescence, by guiding the Qi up the Governing Vessel and down the Conception Vessel (i.e., the Microcosmic Orbit). Even if you were shocked by the un-

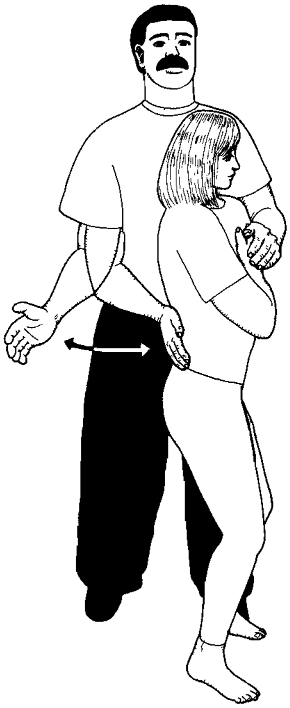

Figure 18.6. The doctor strikes the patient's Mingmen area and focuses his or her attention to lead the disruptive Qi downward into the Earth. This helps the patient stop the uncontrollable muscle deviation spasms.

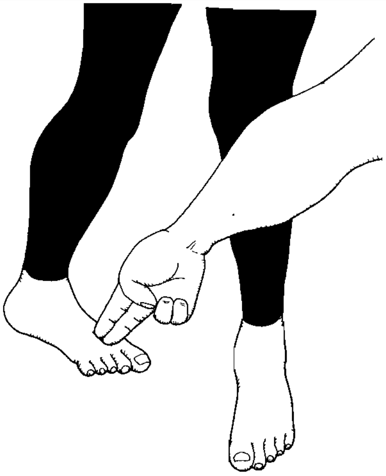

Figure 18.7. The doctor directs energy into the patient's foot, through the Lv-3 point, rooting the patient's Rebellious Qi into the Earth.

expected sound or event, restore the mind to a normal state of calmness by immediately performing the Opening-and-Closing the Two Dantians technique. The two Dantians refer to the Upper Dantian (Yintang point) and the Lower Dantian (Qihai point) areas (see Chapter 20).

To perform this technique, raise both hands to the lower jaw and breathe in with a hissing sound as if inhaling a mouthful of cold air. After the inhalation, the hands will move up to the Upper Dantian, and perform the opening exercise while exhaling slowly. Next, exhale while closing slowly. After four repetitions move the hands to the lower abdomen and perform the Opening and Closing of the Lower Dantian four times.

Usually the emotions will calm down after performing this exercise four times. If you still feel uncomfortable, you can perform the exercise eight times or more, until you become calm.

Note: It is good to ensure a quiet, safe environment in which to practice, preferably a place outside, with fresh air, among flowers, grass or thick, healthy green trees. Because the inside of the body matches the outside environment, when searching for specific directions to face while practicing Medical Qigong, it is important to avoid training in or near:

- extremely intense direct hot sunshine,
- a cold strong wind,
- a thunder and lightning storm,
- old withered trees,
- old graves or tombs,
- electric generators, power lines or power plants,
- dump sites,
- polluted sites, rivers, etc., and
- do not practice shortly after earthquakes, tornadoes, hurricanes, hailstorms, etc.

UNCONTROLLABLE MUSCLE DEVIATIONS

Sometimes a patient's body begins to automatically quiver, shake, tremble, or vibrate, revealing an energetic re-patterning causing muscle movement which cannot be stopped. When this happens, part of the body often keeps moving for several hours, causing exhaustion. Sometimes Qi in the legs and Heart rushes into the head, compounding the problem.

For older patients this is considered a dangerous symptom because their lower extremities are generally weaker. In mild cases the patients feel energy rushing upward into their head. In severe cases the whole body trembles because the Qi leaves the Conception and Governing Vessels and empties into the body itself. In these extreme cases energy muscle patterning should be controlled in the following manner.

1. Advise the patient by saying, "It is time to relax and calm down. You need to end your exercise and rest." This method is usually very effective when repeated over and over, again. If the patient is unable to stop, perform the next method.
2. Direct the patient to place their left hand on the Middle Dantian and the right hand on top of the left hand. Tell the patient that you will help them stop by striking their Mingmen area (Figure 18.6). Press their right hand with your left hand, to compress their Middle Dantian, while using your right palm to strike the patient's Mingmen three times, directing your attention down to their feet. This can help them stop moving.
3. If the above two methods do not help, direct the patients to turn their right foot inward with the big toe pointing at the heel of the left foot. If they cannot do it without help, take their left hand in your right hand and with your foot push the big toe of the patient's right foot to point at the left heel. That should stop the movement.

   Next, touch the patients' Taichong Lv-3 point, between the big toe and second toe, with the "sword fingers" Qi emitting technique. In this technique, the doctor uses the index and middle fingers to emit Qi into the patient and guide their Toxic Qi down into the ground (Figure 18.7).

RECTIFYING QI DEVIATIONS USING TAIJI RULER

If patients are experiencing too much Heat, they are concentrating too much and need to change the position of their tongue and perhaps

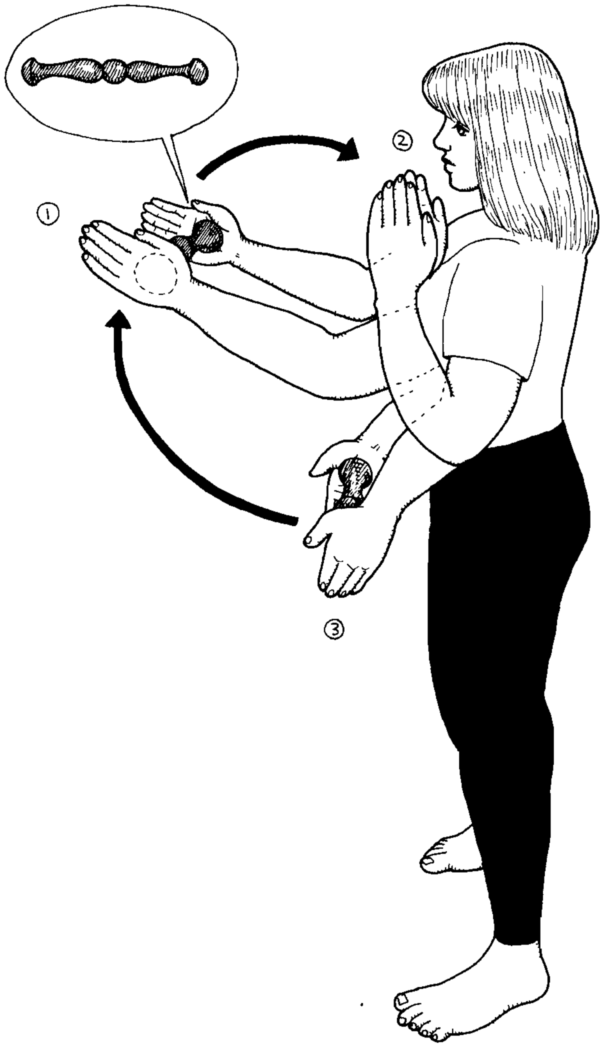

Figure 18.8. The Qi regulation exercises of the Taiji Ruler system are excellent for correcting Qi deviations.

switch meditations. The Taiji Ruler is very effective in rectifying deviations caused by inappropriate tonification or purgation, and can be used to dredge the channels and collaterals to promote the flow of Qi and Blood.

When performing this exercise, the rotation of the Ruler upward and outward (away from the body) is generally used for leading and increasing the energetic flow of Qi and Blood in the upper extremities. The downward and inward rotation of the Ruler is generally used for leading and increasing the energetic flow of the Qi and Blood in the lower extremities.

The purgation of energy blocks are usually cleared through downward movements (from Heaven to Earth). Tonification is implemented through upward movements (from Earth to Heaven). Using these methods, the patients can amend their deviations and promote the increase of Qi (Figure 18.8).

CHAPTER 19

DISORDERS AND DEVIATIONS OF THE SOUL AND SPIRIT

INTRODUCTION TO SPIRITUAL DIMENSIONS

An accepted reality in Medical Qigong therapy is the body's relationship to both the physical and spiritual world. The physical world of the living is considered a Yang state of existence, while the supernatural-spiritual world is in a Yin state of existence. Once the Qigong doctors' Shen has reached a high state of sensitivity, they can transcend their normal conscious perceptions and senses, to observe, and even communicate with the spiritual Yin world (see Six Transportation's of Shen, Chapter 13).

Within the subtle realm of the spiritual world, the existence of spiritual entities and ghosts are divided into two divisions of Yin and Yang (Figure 19.1). There is a significant difference between spiritual entities and ghosts.

1. A spiritual entity belongs to the Yang category of the Subtle Realm, and comes from the formation of nature. Spirits are considered the ethereal beings of the universe, who depend on the absorption of natural energy and are affected by the energetic cycles of the Earth. They are conscious beings, able to transform themselves into any size, shape, animal, or being (shape-shifting). A spiritual entity can affect changes in electrical current within the body's energy fields, or control certain areas on the body's nervous system. Spiritual entities can be separated into two divisions: Yang (good) and Yin (evil) influences.
    a. Yang spiritual entities are commonly referred to as angels, guides, spirits of the light, enlightened masters, and similar positive beings. These spiritual entities assist mankind (communicating through the body's Hun) in their quest for enlightenment. They assist individuals to make beneficial decisions that promote spiritual maturity.

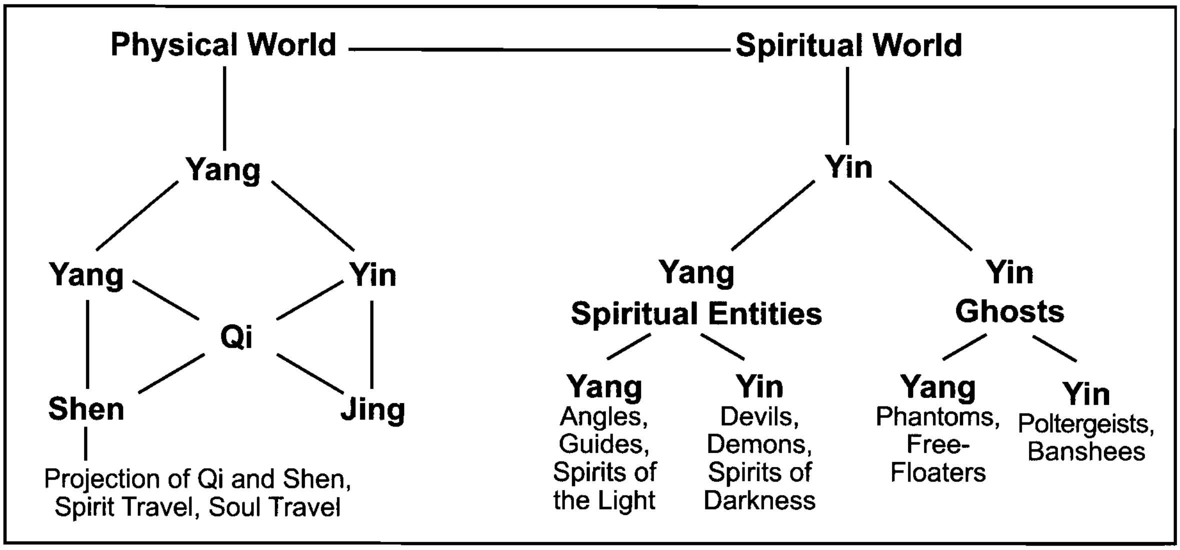
Physical World
Spiritual World
Yang
Yin
Yang
Yin
Qi
Shen
Jing
Projection of Qi and Shen, Spirit Travel, Soul Travel
Yang Spiritual Entities
Yin Ghosts
Yang
Angles, Guides, Spirits of the Light
Yin
Devils, Demons, Spirits of Darkness
Yang
Phantoms, Free-Floaters
Yin
Poltergeists, Banshees

Figure 19.1. The Spirit World

   b. Yin spiritual entities are commonly referred to as devils, demons, spirits of darkness, evil spirits, and other similar negative spiritual manifestations. Yin spiritual entities challenge mankind (communicating through the body's Po) by confusing the human spirit. If given the opportunity, these spiritual entities can possess an individual and become extremely territorial.

2. Ghosts or Gui, as they are commonly called in China, belong to the Yin category of the subtle realm. Ghosts are developed from the subtle form of the combined spiritual essence of an individual's Eternal Soul. When the body dies, the energies of the Corporeal Soul (Po) enter into a state of rest, returning back to the Earth, thus allowing the Ethereal Soul's (Hun) to return back to Heaven. The individual's Eternal Soul then fuses with the energies of the Three Dantians (within the Taiji Pole) and is projected outside the body, entering the tunnel of light to return back to the divine. Sometimes, however, because of unprocessed emotional issues, traumatic death, unfinished business, or extreme attachment to people, places, or possessions, an individual's Eternal Soul can begin to wander the Earth, becoming a Gui or ghost. Generally, it is believed that it takes three days for the Corporeal Soul to withdraw their energy completely from a deceased body. Ghosts are considered Eternal Souls that have departed from human life and are unable to detach from their life experiences; they are bound by their unprocessed issues. Ghosts can also be separated into two divisions of Yang (good) and Yin (evil).
   a. Yang Ghosts are commonly referred to as phantoms and free floaters. These disembodied souls can either be lost, confused, delusional, or purposely return in order to guide and protect the living.
   b. Yin Ghost are commonly referred to as poltergeists, banshees, or other names varying by location. These disembodied souls are angry, malicious, mischievous, and destructive.

It is difficult, if not impossible, to be immune to the influence, or encroachment, of the spiritual world. Spiritual manifestations are usually evident through audible sounds, unusual unexplainable smells, extreme cold, and the displacement of objects. Other phenomena include tactile sensations, visual images, voices, and the apparent psychokinetic movement of objects. Spiritual entities seem to move through solid matter and can appear or disappear abruptly. They can also cast shadows and be reflected in mirrors; some seem corporeal, while others are luminous, transparent, or ill defined. Many spiritual entities have jerky or limited movements, while others are lifelike in movement and speech. Encounters with spiritual entities can be divided into five main categories.

1. Spiritual manifestations of the dead generally occur within a short time after the death of the person. Visits from the departed are usually to bring comfort, to announce their role as a guardian, or to complete unfinished business. If an individual's life ends in an untimely manner, or so violently that the soul does not realize the body is dead, the disembodied soul often goes about its daily routines. Occasionally these souls will manifest as mischievous entities, attempting to get the attention of the living, but they lack the malevolence of an evil poltergeist manifestation.
2. Deathbed phenomena are generally of divine beings, religious figures, or luminous apparitions. Visions of previously deceased loved ones, who come to guide the dying soul into the next level of existence, are common occurrences.
3. Manifestations of impending tragedy are visual images of an entity which usually appears in a "waking vision" or in dreams. When they appear, these spiritual entities either communicate about the death or severe crisis of a loved one, or to give warning of an impending tragedy.
4. Reincarnating spiritual entities sometimes appear in a dream to a member of the family into which they will be born. Such dreams are referred to as "announcing dreams."

5. Manifestations of poltergeists are generally mischievous and often malevolent. They may be either disembodied souls, or demon-like entities (see Chapter 18). These disembodied souls were often extremely dissatisfied and bitter while living, and may be unable to reintegrate into the Earth after death, or to ascend into the Heavens. This may cause spiritual hauntings, either around the location of their death or around a person, or persons, that the disembodied souls blame for their life's misery. They make their presence known by assaulting the living, either human or animal. The most common phenomena include: battering with a rain of small stones or sand, throwing and moving of objects, loud noises and shrieks, strange lights and vile smells.
6. The ghostly images of the astral body can be projected and manifested in distant locations as a disembodied apparition. Sometimes the astral body can travel to other locations and actually appear in holographic-like form to others. This image is not actually considered a ghostly apparition (although it may appear so), but the extension and material form of an individual's Eternal Soul.

   One example of this type of phenomenon was demonstrated in France in 1908 by scholar Hector Durville, who was researching "traveling clairvoyance." In one series of tests, Mr. Durville and a colleague hypnotized a female psychic to facilitate her astral projection to another location. An observer was placed in another part of the house. The experimenter's instructed the female psychic's to touch, hit, or pull at the observer's body while in her astral body. The observer had no idea what to expect, but felt the touches, strikes and pulling on his body by invisible hands. The astral body was even visible as a whitish apparitional figure to some of the people Durville used as witnesses.

EVIL STATES AND MENTAL DELUSIONS

Sometimes the "spirit world" itself can affect the physical body through a subconscious emotional connection with the individual's Shen. These affected subconscious emotions can never be totally repressed because they will continually surface as dreams, visions, flashes of images, etc., and can only be redirected into either positive or negative actions.

Seeing a spiritual entity or ghost is a normal part of Medical Qigong training and is accepted as a natural energetic phenomena; however, disorders such as schizophrenia and psychosis (delusions and hallucinations, etc.) are considered Qi deviations and can be caused by improper Medical Qigong practice. These mental states are known as "evil states" and are difficult to correct. Evil states also include: infatuation, delusions, obsessions, encounters with seductive spirits, demon oppression or possession, and multiple personality disorder. This does not mean, however, that the individuals suffering from these disorders are themselves evil.

INFATUATION WITH THE TREATING DOCTOR

Infatuation is culturally regarded in China as an evil state. This generally refers to the occurrence of deceptions of the subconscious mind during, or after, Medical Qigong healing, which lead to mental derangement. While under the influence of these self-deceptions, patients display intensely amorous affection towards the doctor. As the doctor's energy becomes more intensified within the patients' nervous system, the patients experience a type of biological euphoria which gives way to feelings of unexpressed emotions, suppressed fantasies, and symptoms of eccentric disposition, such as sluggishness, apathy, or elation.

During treatment, patients become extremely open and vulnerable, as their boundary systems completely dissolve under the influence of the doctor's emitted Qi. The Qigong doctor's ability to channel divine light and compassion may easily be mistaken for displays of personal love toward the patients. Sometimes this divine love has never before been experienced by these patients, making them infatuated with the doctor, who is the apparent source of this love. This is an illusion which can usually be dispelled by the doctor's explanation. When the doctor is unable to dispel this illusion by conversing with the patients, this indicates an underlying, often long standing, mental disorder. Such patients should be referred out to a mental health practitioner for professional counseling.

ENCOUNTERS WITH SEDUCTIVE SPIRITS

On rare occasions, individuals who do not have a strong enough connection with the divine may become susceptible to encounters with seductive spirits. These spirits are known as Incubus (male) or Succubus (female), and tend to manifest in dreams. These dreams can range from the romantic and seductive, to rape and sodomy. Although these encounters occur during the individual's sleep, physical sensations are always a large part of the encounter. Often, once the victim has awakened, the physical sensations continue.

DELUSIONS

This is the occurrence of delusions of the subconscious mind during or after Medical Qigong treatment or training, which lead to mental derangement. A delusion is a false belief brought about without appropriate external stimulation (seen most often in psychoses), and is inconsistent with the individual's own knowledge and experience. Patients suffering from delusions display unsociable and eccentric dispositions, sluggishness, apathy, and trance. Some patients may become very disillusioned and depressed. They may lose confidence in life, and think about committing suicide; others may have symptoms similar to those of psychosis, such as persistent visual or auditory hallucinations. These symptoms occur because of preexisting emotional problems, which the patient has not worked through, or has consistently avoided facing.

In all truth, the practice of Medical Qigong does not create these problems; Medical Qigong can, however, occasionally allow patients to become extremely sensitive to the preexisting unresolved problems that continue to surface into their subconscious mind due to their overactive denial system.

SPIRIT DEMONS

An encounter with the phenomenon known as demon oppression or demon possession rarely occurs during clinical practice, but never the less, it must be addressed. Demon oppression or possession describes the state of mind attributed to the patient who has come under the influence of the "spirits of the objective world." These spirits of the objective world (also called "fallen angles," "spiritually hostile forces," or "demons") can either influence, or inhabit, a patient's thoughts and body. They can affect anyone who becomes open to their influence. Any practice or situation which opens the individual's subconscious mind may increase the person's vulnerability to such states. Spirit demon oppression or possession can occur from:

- misguided or improper use of meditations,
- misguided or improper use of hypnosis, or self-hypnosis,
- indulging in the practice of astral projection,
- meditations that cause adverse psychological side-effects,
- general anesthesia,
- drug and alcohol abuse,
- indulging in the practice of channeling, or Ouija board playing,
- deep emotional or physical traumas,
- indulging in the practice of deviant psychosexual encounters (Tantric sex with the wrong individual)
- exposure to places that are haunted by evil spirits when the individual is very stressed out, or deeply fatigued and depleted, and
- indulging in the practice of "black magic."

When the patient's connection to their Shen becomes extremely Deficient, to a point that they surrender their Hun, the patient's Hun may leave their body and wonder, making them vulnerable to the possibility of spirit oppression or possession.

One of the first things a demon seeks, once it enters a body, is to fulfill its sensual appetite (be it sexual pleasures, rage, or emotional addictions, etc.). A disembodied spirit lacks the ability to feel physical sensations due to the disconnection with its bodily senses (touching, feeling , smelling, tasting, etc.). Thus, it can only satisfy these cravings through someone else's living body. Consequently, if a patient is disassociated from the divine center core, or has little or no boundary system, a condition of oppression or spirit possession may occur. In China this condition is commonly described as, "the Spirit (Hun) leaves and the demon enters, uniting with the body's Po" and must be rectified with spiritual atonements and soul retrieval.

Although such encounters in the clinic are rare, it is still important to recognize and treat these conditions when they occur. While treating patients in China, I was fortunate enough to share my experiences with demon oppressed or possessed patients with certain doctors at the Xi Yuan Hospital. It was refreshing to talk with another doctor about this particular subject, as in America it is difficult, if not impossible, to find other individuals who can discuss such a topic in a clinically sound manner.

In understanding the phenomenon of spirit demons, it is important to realize that in the clinics in China, encounters with these types of spirits are divided into two categories: Projected Spirit Demons from the Subconscious Mind, and Actual Encounters with Spirit Demons.

PROJECTED SPIRIT DEMONS FROM THE SUBCONSCIOUS MIND

Sometimes through meditation, the subconscious mind will release visions into the conscious mind in the form of ghosts, phantoms, spirits, demons, etc. These energetic visions are subjective projections of part of the self, released through emotional arousal. Each color image represents certain feelings and repressed memories within the organ which the subconscious mind is trying to communicate to the conscious mind. This also applies to dreams and nightmares in which the patient is continually being haunted by certain spirits, demons, or monsters. For example:

- A green demon represents the deeply suppressed emotions of anger, rage, and irritation which are connected to the Liver.
- A red demon represents the deeply suppressed emotions of excitement, panic, and even anxiety which are connected to the Heart.
- A yellow demon represents the deeply suppressed emotions of worry and pity which are connected to the Spleen.
- A white demon represents the deeply suppressed emotions of sadness, grief, sorrow, guilt, anxiety, distress, and heartache which are connected to the Lungs.
- A black demon represents the deeply suppressed emotions of fear, paranoia, horror, panic, and terror which are connected to the Kidneys.

The patient's subconscious demonic images may tend to be culturally influenced. Each patient may find himself or herself observing something quite different, but with the same underlying emotion. In modern Western culture, individuals are more likely to dream of thugs and assassins that are threatening and pursuing them, as well as monsters.

ACTUAL ENCOUNTERS WITH SPIRIT DEMONS

These spiritual encounters have an objective reality, as they are true spiritual entities which exist outside an individuals thoughts or feelings. For obvious reasons it is extremely important for Medical Qigong doctors to have a strong spiritual foundation. Without a powerful connection to the divine, the doctor risks absorbing the patients' pathogenic Qi, and may become more vulnerable to "evil states" that may place him, or her, at the mercy of spirit demons who seek to dominate, control, and feed off of the negative emotions of the human body.

The spiritual state of the energetic field controls the emotional field, the emotional field in turn controls the mental, and the mental controls the physical field. This is the spiritual template used in the Medical Qigong clinics in China for initiating the energetic laws of dominance and control over the body's tissues. When there is a spiritual interaction involved, the emotions manifest the outcome. There are two ways a spirit demon can attach itself to the body; through spirit oppression, or through spirit possession.

1. Demon or spirit oppressed patients tend to first have had severe emotional trauma, resulting in extremely low self-esteem, depression, fear, rage, and so on (these are the emotional energies on which the evil spirit feeds). The spirit demon becomes attached to the patient's second field of Wei Qi, drawing energy and sustenance from the patient's chronic release of negative emotions.

   These spirit demons are external energy feeders. They are spiritual parasites which provoke and increase negative emotional reactions. Once the patient discharges these negative emotions into their second field of Wei Qi, the spirit demons then ingest and absorb this energy as food. If, for any reason, the spirit demons leave or abandon the patient's second field of Wei Qi, they will, like a predator, continue to look for their next feeding ground. This is why, in certain families, oppressive dynamics are passed from generation to generation, due to certain congenital spiritual patterns.

2. Demon or spirit possession occurs when the patient's Spirit/Mind (i.e., the focused power of their consciousness) becomes too attached to something; it may leave the patient (becoming a wondering spirit) and not return. The vacuum that is left by the original consciousness leaving its residence can become filled by the spirit of other beings. These alien energies tend to enter the patient's body, replacing their original consciousness (Yuan Shen), which is now lost. The resulting state is one of demon or spirit possession. Because the spiritual quality and stability of the patient's mind is lost, the patient will become increasingly dependent on the alien spiritual consciousness for orientation towards others, and the outside environment.

When either a demon or spirit entity possesses a patient's body, it tends to be extremely territorial and protective. This is because, as the demonic spirit surrounds the very core of the patient's energetic body, it initially has a strangle hold on the patient, feeding and distorting the patient's reality. This results in wild mood swings, and sometimes, in displays of extraordinary intuitions, perceptions, and incredible physical powers.

In most cases the spirits causing the pathology are not of a high order of intelligence; they are more on the level of spiritual bacteria, viruses, or parasites. In some cases, however, spirits can be of a high order of demonic intelligence, such as depicted in the movie "The Exorcist."

EXAMPLES FROM PERSONAL CLINICAL EXPERIENCE

Through my personal clinical practice, I have treated several patients who were actually demon possessed, or demon oppressed. One example occurred in 1992, when a patient came to me with a tumor located on her pituitary gland. The Western doctor who was treating her had suggested that she receive immediate surgery, so as a last resort she came to me.

Knowing that she had written several books that were currently being used in a major university as a standard text, I was quite impressed and even intimated by her reputation as an intellectual. She, being the mentor of several of my students who were psychotherapists, did not ease my apprehension. Although she was elderly, she was very strong willed, and quite opinionated. She requested that I lower the treatment table to the floor before starting the therapy. I thought that this was a little strange, but complied never the less, and so we began the initial treatment with the patient lying supine on the table, placed on the floor.

I always begin each treatment with a connection or "hookup" to the divine After connecting with the divine, the energy of the treatment room changes, as the room's energetic field transforms into a divine healing energetic field. Immediately upon "hookup" something in the patient shifted. Although her eyes were closed, the patient began to thrash about on the table like a wild animal, literally snarling, and gnashing her teeth. The power radiating from her body was quite incredible. I knew and believed, that if she wanted to, this elderly, frail woman, could now lift my body up, and toss it across the room like a rag-doll. I was quite aware of the supernatural power and spiritual demonic form that I was now facing, and knew that my 27 years of martial arts training would not serve me in this situation. To my advantage, I had previous exposure to such spiritual demonic phenomena, and had been successful in dealing with demon possessed individuals. Believing that the current healing work was part of my ministry and life purpose, and having participated in several demonic exorcisms, I was aware of the procedure which needed to transpire.

I began dredging and casting out the demons

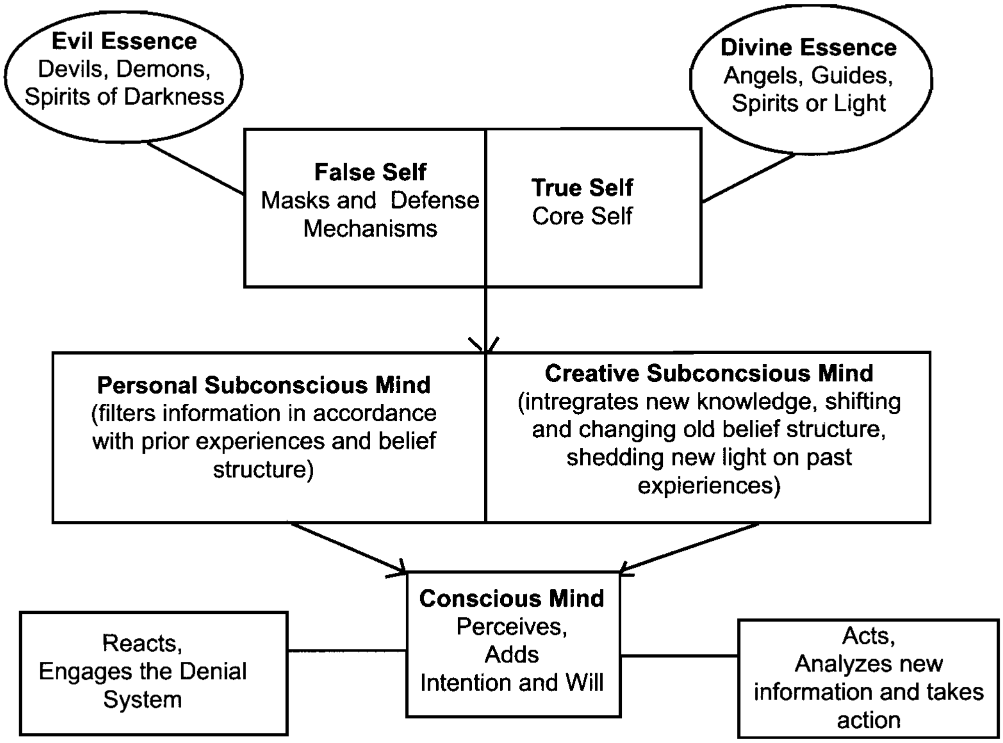
Evil Essence
Devils, Demons,
Spirits of Darkness
Divine Essence
Angels, Guides,
Spirits or Light
False Self
Masks and Defense Mechanisms
True Self
Core Self
Personal Subconscious Mind
(filters information in accordance with prior experiences and belief structure)
Creative Subconcsious Mind
(intregrates new knowledge, shifting and changing old belief structure, shedding new light on past expieriences)
Conscious Mind
Perceives,
Adds
Intention and Will
Reacts,
Engages the Denial System
Acts,
Analyzes new information and takes action

Figure 19.2. All human beings are subject to divine and evil spiritual influences. These influences can affect the subconscious and conscious mind, as well as the sensory perceptions and physical performance.

one by one, claiming each portion of her body for the glory of Christ (my divine authority). This procedure continued for six weeks, each treatment involved reclaiming certain areas of her body. Beginning at her head, I continued treatment down the patient's body. On the sixth week, when I reached her Lower Dantian, the last spirit demon left her body, and she immediately curled up into a fetal position and began to cry. Her Hun then returned to her body, and for the first time in many years she could feel loving emotions again.

It is important to note, that while the treatments were being administered, the patient was not consciously aware that anything out of the ordinary was happening. She had been so completely detached from her body, that she had no conscious recollection of her own emotional-spiritual field of existence, or its affect on her physical body. She had literally hid within the confines of her intellect, where she felt safe. She was also not aware of being demon possessed, or even that her body thrashed about during each treatment. In effect, she had disassociated from her Yuan Shen during childhood, and thus became dissociated from her emotions. After six weeks of treatment her brain tumor had completely dissolved. But, far more remarkable, was the change in her personality, for she became both friendly and courteous.

OVERVIEW

As mentioned before, once the Qigong doctors' Shen has reached a high state of sensitivity, they are able to transcend their conscious perceptions to sense, observe, and communicate with the spiritual world (Figure 19.2). This ability to see disembodied spirits or ghosts is a normal part of Medical Qigong training and a natural part of energy observation.

These spiritual presences can be seen or felt for either brief moments in time (usually as a hovering image of light, or a dark shadow), or for extended periods of time (usually an encounter). Some of these spiritual entities can also take the form of animals.

Some doctors view these energetic entities as either individuals who are astral-projecting (from the past, present, or future), or forms of energy that can either be ascended spirits of light (divine guides or angels), or fallen spirits of light (demons and hostile forces). In any case, encounters with such entities of the spiritual world should not startle or scare the Qigong doctor, but be expected as a part of the existence of the spiritual domain.

A final important note to make is that, although demon possession is real, the location, or type of the patient's disease, has nothing to do with demonic possession. Demonic possession is related to the disassociation from the Yuan Shen and the absence, or suppression, of the patient's Hun. It is not caused by obstructed tissue formations or physical disease.

After helping the patient uncover any severe emotional trauma, the Qigong doctor should assist the patient in "soul retrieval" meditations, in order to return the patient's Hun back into its residence (the patient's physical body) and to facilitate a strong reconnection to the Yuan Shen. During soul retrieval meditations, patients sometimes talk about viewing certain traumas from different visual perspectives, such as: watching the incident from above their body, or watching from a different corner of the room. The reason for this change in visual perspective is due to the displacement of the patient's Hun once it leaves the body. Although the Hun are outside the body, they still observe, listen, and record. This accounts also for the numerous cases of astral projection during surgery, as well as during near death experiences. Under anesthesia, many patients travel out-of-body and observe in fine detail what the doctors and nurses are saying and doing. They accurately recall the doctors' and nurses' conversations, and the experience of viewing the operation from above the body. These experiences are becoming more documented in recent years partly due to modern techniques in resuscitation, that have made such experiences more frequent. First hand accounts of out-of-body experiences abound in "near death experience" literature.

Similar experiences have been recounted by many physically and sexually abused patients, who, as children, learned to dissociate from the body, and observe the abuse from the ceiling, or corner of the room. In most cases, these adults, and children, do not suffer from multiple personality disorder. Only in the most severe cases does the personality fracture and become multiple.

MULTIPLE PERSONALITY DISORDER (DISSOCIATIVE IDENTITY DISORDER)

Occasionally the Qigong doctor may encounter a patient with a multiple personality disorder. The condition of a multiple personality disorder is generally due to extreme sexual, physical, and emotional abuse during childhood; in some cases, demon possession may also occur. This is a condition in which the patient has developed a split in their consciousness, resulting in the development of two or more distinct identities or personalities that recurrently take control of the patient's behavior and consciousness, one at a time.

It is believed that the original personality splinters during childhood (usually around the age of 6 or 7 years, or as early as 3) creating several different personalties, or personality states, that are disassociated from the patient's original whole personality. The different personalities created can number as few as two, or upwards of one hundred. They are often referred to as alternate identities or simply 'alters.' These alters, in fact, often themselves give birth to more fractured alters, each with its own gender, age (corresponding to the fracturing trauma) and function (or role). Some alters may be very aggressive and protective, others will be more passive and submissive.

Some alters will have different medical conditions from the others, which inhabit the same body. As the patient changes personalities (alters), the internal and external energetic fields likewise undergo changes. Each personality can differ in age and stage of development. In most cases, the personalities are usually unaware of the words, actions, and feelings of the other personalities, especially at the beginning of the treatment. Each personality can exhibit a different set of symptoms, as well as diseases. One personality may have chronic asthma and allergies, but when switching to another personality, the patient may suddenly display healthy lungs but display nearsightedness or high blood pressure.

Usually one personality becomes a helper to the doctor, while another personality may be psychotic, indifferent, and so on. Some of the alters may have access to highly advanced transpersonal states of consciousness, i.e., a finely honed and active psychic realm which may have provided a haven, or sanctuary during the abuse.

TYPES OF ALTERS

Over time, an alter's role in the patient's personality changes. There is a difference between an established personality which maintains both an internal and external role and just a fragment personality (e.g., washes the dishes only). Also, there will usually be a number of child personalities included, as well as adult personalities. The patient's past history of violence is the best indicator of whether there are dangerous alters. Some of the most prevalent personality types of alters are described as follows.

1. The first personality encountered by the doctor, or psychologist, is usually the patient's host personality, not the patient's original personality.
2. The demonic or spiritual personality type is generally found in very religiously oriented patients.
3. The persecutor personality type is generally responsible for the patient's attempts at self mutilation (a very common type syndrome with multiple personality disorder) or suicide. They tend to be children or teenagers. It is believed that these types of personalities are either interjections of the patient's original abuser, or have evolved from "helper" personalities.
4. The helper personality type is generally responsible for providing the patient with the emotional support to counteract some of the self-destructive behavior of the patient's persecutors.
5. The administrator or obsessive-compulsive personality type is often responsible for allowing the patient to earn a living. They generally appear cold and aloof.
6. The imitator, or impostor, personality type generally mimics other personalities and is responsible for handling certain situations that are difficult for the patient to bear.
7. The autistic or handicapped personality type generally emerges when the patient feels that he or she is under intense scrutiny, or feels that he or she is being controlled or confined. They may manifest autistic or catatonic states or become functionally deaf or blind.
8. The promiscuous personality type is generally responsible for the patient's overactive or uncontrollable sexual urges (e.g., nymphomania).

While treating patients with multiple personality disorder, the therapeutic goal is usually to integrate all the different personalities, merging them together through creative visualization (usually using hypnosis). This is achieved by having the patient imagine each personality and subset "alters" merge into one.

Some patients, however, opt to remain multiple, as merging incurs the "death" of others. In such cases, contracts are made with the different alters to keep each other informed of events occurring during their "coming out," and some of the more responsible personalities keep watch over the more self-destructive ones. Due to the complexity of treating multiple personality disorder, and due to the high potential of suicide in multiple personality cases, it is recommended that the Qigong doctor refer these type of patients out to specialists who are familiar with such conditions and their recovery.

SOUL RETRIEVAL AND MEDICAL QIGONG THERAPY

According to the Medical Qigong perspective, one of the major causes of illness is "soul loss" (the loss of parts or memories of the Eternal Soul). On the spiritual plain, the Eternal Soul is the manifestation and sum total of our energetic associations and the energy of the Wu Jing Shen (see

Chapter 2). The Eternal Soul serves as a middle point between the source and organization of all life-force energy. The Eternal Soul is considered the seat of all emotions, feelings, and sentiments. As a crystal grows around a central matrix, so the body grows around the Eternal Soul. In revue, the Eternal Soul radiates energy in all directions, like the sun, and is responsible for life, health, and the formation and growth of the body. In the scope of energetic manifestations, the Eternal Soul expresses its innate qualities as energetic movement, functioning through the physical form (led by the Shen). It is connected to all parts of the body, and its light or energy is reflected through the eyes via the Hun. It is through the influence of the Eternal Soul that all of the body's energetic processes (organs and organ systems) seek wholeness. It knows exactly what is needed in every situation for survival and health preservation, and takes action via the Po. Whole body consciousness is the main characteristic of the Eternal Soul. Without the process of the Eternal Soul, energy would have no specific direction and remain in meaningless activity.

"Soul loss" is not demon possession, but a spiritual illness that causes emotional, mental, and physical disease. When disturbed, the Hun leave the body and the Mind (Shen) wanders off on its own. When the Hun leave and the Shen wanders off for very short periods of time, it is sometimes called, "dissociating" or "spacing out." When the Hun leave for extended periods of time, it is called soul or spirit travel, which generally happens when the patient is asleep. It is believed that if the Hun completely vacate the body, the patient will die. A coma is one example of a chronic physical state where the Hun have left, but are still attached to the body, through the Hun's connection with the patient's energetic Qi field. When a patient is in a coma, the Hun and Shen abandon their residence, wandering about in the astral plane of existence.

The Hun and Po create the sustaining energy for the Wu Jing Shen, which provides the energy for the Yuan Shen. If the Qigong doctor can retrieve the lost parts of the patient's Eternal Soul (the suppressed memories attached to the Hun and Po), then the individual's Shen can be restored back to harmony and health.

REASONS THE ETHEREAL SOUL LEAVES THE BODY

In ancient times, disease was attributed to the Hun having strayed away, or having been stolen. It was believed that the Hun would leave the body through three main incidents: they could be stolen (through deep infatuation), frightened away, or simply wander off. In addition to soul theft or "rape of the soul," some Qigong doctors relate Hun loss to the interference of Gui (ghosts, demon spirits), as well as other human beings (i.e., a broken heart or severe trauma).

Soul retrieval is an essential part of Medical Qigong therapy; the body and mind without the Hun, or the spiritual connection of the Yuan Shen, remain in a suspended state of spiritual and emotional shock. When the spiritual needs of the Eternal Soul are neglected, the Shen begins to disassociate, and symptoms of addiction to drugs or alcohol, obsessions, violence, or deep depression take over (see Chapter 32).

Because we are born understanding hidden spiritual realities, patients will inevitably succumb to various mood swings; these are due to deep emotional pain and experiences that stay within their memory and touch their Heart. A spiritual life of some kind is absolutely essential for psychological health. Traditionally, the Hun are said to suspend themselves midway between the conscious and the unconscious, and that their perceptual modality is neither the mind's nor the body's, but the imagination's (i.e., the Creative Subconscious).

From a clinical perspective it has been noted that the Hun leave the body during severe emotional traumas (i.e., surgery, the death of a loved one, miscarriages, abortions, incest, molestations, physical and emotional abuse, the end of a deep relationship, and the stress of combat). Whenever we experience trauma, our Hun, which are part of our vital spiritual essence, separates from us to survive the experience and to escape the full impact of the pain (Figure 19.3). When patients be-

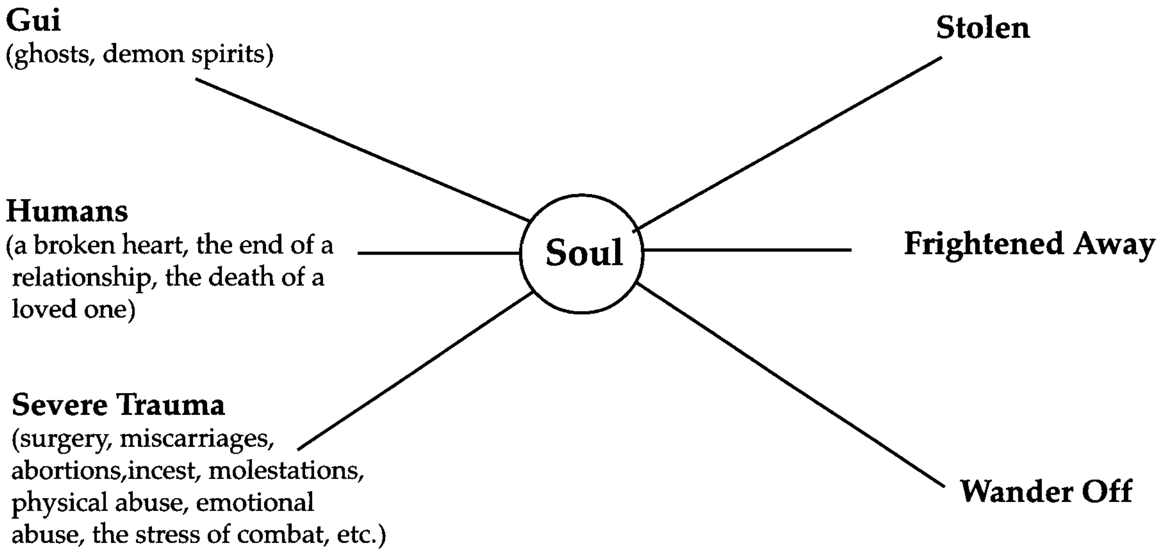
Gui
(ghosts, demon spirits)
Humans
(a broken heart, the end of a relationship, the death of a loved one)
Severe Trauma
(surgery, miscarriages, abortions,incest, molestations, physical abuse, emotional abuse, the stress of combat, etc.)
Soul
Stolen
Frightened Away
Wander Off

Figure 19.3. Reasons for the Ethereal Soul Leaving the Body

come dissociated, aspects of their personality separate from the mainstream of their consciousness, resulting in gaps of memory loss. An extreme example of personality separation occurs in multiple personality disorder.

When a patient has been violated or wounded in some way and remains in an emotionally present state of mind, within the body, the result usually encompasses feelings of anger from past hurts, grief from present pain, or fear of future harm. If an over-stimulation of the energetic organs occurs, it may wound the patient's Shen, causing the patient to close off from the Yuan Shen, to shut off all emotions. This action is similar to that of a sea-anemone, whose natural response (for the purpose of survival) is to immediately close itself after being touched.

THE PURPOSE OF SOUL RETRIEVAL

The purpose of Medical Qigong soul retrieval therapy (called age regression therapy in Western psychology) is to bring the awareness of emotions back into areas of the physical body, allowing the patient to express and release the feelings and reactions that are the cause and symptoms of his or her disease. This will help the patient transform the trapped, or frozen energy (stagnated Qi), back into its energetic potential to be reintegrated with the Shen.

A Qigong doctor will accompany and tend to the patient's Eternal Soul in times of emotional transformation, including times of crisis, illness, and emotional death and rebirth. This allows the patient to remain in the energetic present time mode. This also results in the relief from symptoms of illness and the return of the patient's personal power, creating the potential for fulfillment in work and rewarding relationships. Some patients may come to the realization that they must quit their job and find another that is more spiritually rewarding. They may also find themselves discarding old friendships and making new ones, as their energetic, emotional, and spiritual fields undergo growth and change.

WOUNDING AND CLOSING THE SPIRIT

The body's Yuan Shen is the main expression of the Eternal Soul as it moves from its inactive state to a state of dynamic activity. The Yuan Shen is the medium or active environment of the Eternal Soul. The Yuan Shen also expresses its energetic influence through a continual interchange between the personal-self and the higher-self through the act of breathing. The Eternal Soul expresses itself specifically through the Yuan Shen, and the Yuan Shen manifests itself through the physical body.

If a person's Yuan Shen becomes wounded (feels attacked, violated, taken for granted, etc.), it begins to close itself off from its environment. When the Yuan Shen "closes", the emotions begin to shut down, and the Hun wander. Depending upon the patient's set patterns, this entire process can occur within seconds. Once the Hun wander, the acquired (analytical) Zhi Shen and surviving Seven Corporeal Souls (Po) take over the body, and the patient responds like a machine (i.e., the patient can still think and function; reflexes remain intact, but he or she is void of interpersonal feelings). Patients whose Shen have closed down cannot make emotional attachments, and live in a state of perpetual hunger. These patients go through three stages of isolation before their spirit leaves.

1. The first stage involves anger and grief.

   Once a patient has been hurt (before his or her spirit begins to close), his or her natural survival energy immediately affects the Liver and Lungs, causing emotional reactions that are a natural response to feeling attacked. Hurt, sadness and anger are natural responses to feelings of rejection, betrayal and abandonment. These emotions reflect a spiritual, emotional and mental protest to feeling insignificant, excluded, and dis-empowered.

   a. The patient's congenital emotions of kindness and compassion, stored within their Liver, instantaneously shut down as the Liver Qi expands violently outward. This energetic reaction gives way to releasing the patient's acquired feelings of anger and rage, stored within the Liver. The energy of anger and rage is projected outward to protect the patient's emotional boundaries. If the patient has deep seated emotional problems, however, the energy can also implode inward creating self-hate.

   b. If this protective emotional reaction is not effective in obtaining the needed results (being heard, honored, valued, etc.), the patient's energetic field rebounds inward affecting the Lungs. Once the congenital emotions of integrity and dignity stored within the patient's Lungs become effected, the Lung Qi becomes obstructed, and the acquired emotions of grief and sorrow are released. The energy of grief and sorrow is encapsulated internally to encase the patient's wounded spirit.

2. During the second stage anger and grief transform into depression and despair.

   As the patient begins to lose hope, the Shen begins to close, causing two reactions.

   a. The patient's Liver Qi becomes obstructed and stagnant. This in turn transforms the emotion of anger into depression;

   b. At the same time the patient's Lung Qi transforms grief into despair.

   The patient slowly begins to feel helpless, hopeless, and empty inside. Because of the increased depression and despair, the patient becomes physically, mentally, emotionally, energetically and spiritually tired, and gives in to feelings of insignificance, unfulfillment and loss of control over his or her life.

3. During the third stage depression and despair transform into indifference and detachment.

   In the final stage, the patient has lost all hope of resolution and gives up entirely. The Shen closes down, and the Hun wander. The patient feels emotionally detached from people, the surrounding environment, and his or herself. The patient wanders around "dead to the world," isolated, internally unfilled, and alone. This creates what is known as a "depersonalization disorder."

   A depersonalization disorder involves the persistent, recurring episodes of depersonalization, characterized by a feeling of detachment, or estrangement from one's self. The individual may feel as if he, or she, is living in a dream, or movie, and responds to life like

Three Stages of Emotional Transitions before the Spirit Closes

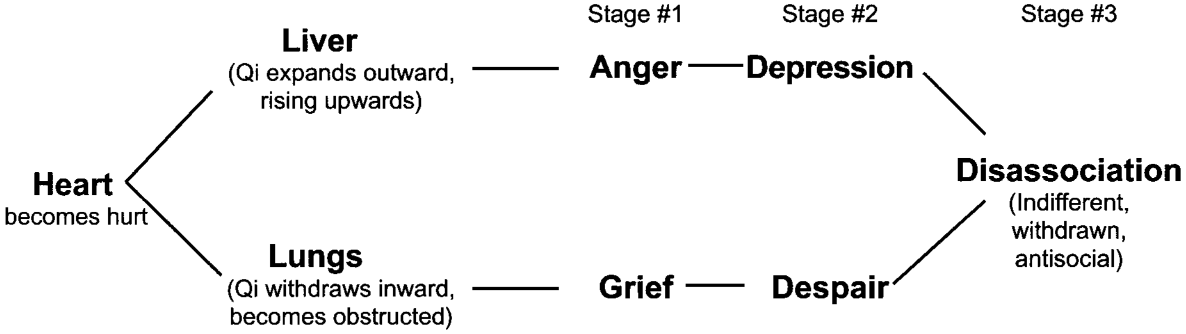
Stage #1
Stage #2
Stage #3
Liver
(Qi expands outward, rising upwards)
Anger
Depression
Heart
becomes hurt
Disassociation
(Indifferent, withdrawn, antisocial)
Lungs
(Qi withdraws inward, becomes obstructed)
Grief
Despair

Figure 19.4. There are three stages of emotional transition the patient must experience before their Shen closes and their denial system engages. After these stages are experienced the patient's defense mechanisms take over.

a robot. They may also have the sensation of being an outside observer (to their mental processes, body, or specific parts of the body). This feeling is often accompanied by sensory anesthesia, a lack of affective response, and a feeling of lack of control over their actions (including speech).

Voluntary induced expressions of depersonalization or derealization that are meditative, or cultural trance practices, should not be confused with this involuntary disorder. It is believed that approximately 50% of the population may experiences a single, brief episode of depersonalization, usually precipitated by severe stress. Approximately one-third of individuals exposed to life-threatening danger incur a brief episode of depersonalization. Depersonalization episodes may be accompanied by visual distortions, such as the shape or size of objects, and other people may appear unfamiliar or mechanical. Other common associated features include: anxiety or depressive symptoms, obsessions, somatic concerns, and time distortion. Depersonalization and derealization are some of the symptoms of panic disorders (not due to drugs or medication), which are usually caused by long-standing anxiety and stress. Panic disorder can also be brought on by emphysema and asthma due to the difficulty in breathing.

When the Eternal Soul becomes wounded, the first two stages in closing down the Yuan Shen are caused by the patient trying to repress feelings of anger (stored within the Liver), and sadness and grief (stored within the Lungs). This reaction is incurred because the Liver stores the body's Three Ethereal Souls (Hun), while the Lungs store the body's Seven Corporeal Souls (Po). These two emotions are expressions of the patient's natural protest against the lack of love that their Yuan Shen (which encompasses or contains both the Hun and Po) is experiencing (Figure 19.4).

HEALING THE SPIRIT

The first stage of healing in soul retrieval is to open the patient's Yuan Shen by giving them hope and encouragement. Hope helps the patient's perspective to be positive and restores meaning to life. Once the patient has closed his or her Yuan Shen, it cannot be opened through argument or logic. Instead the Qigong doctor must use a gentle, soft tone of voice expressing compassion and high regard for the patient. The patient must feel honored and valued by the doctor to feel safe enough to release the hurtful and negative emotions.

Praise must be used carefully and gradually, since some patients may be as yet, unable to accept a higher concept of themselves. Too much praise may cause the patient to feel "mocked" or "lied to", given his or her very low self-esteem. This type of healing only happens when the patient's spirit is engaged, stimulating the inner emotions and not just the intellect.

Once the patient's Yuan Shen begins to open, it is important for the doctor to increase the patient's hope for interpersonal relationships. This is initiated by using silent mental affirmation, or prescribing positive affirmations in conjunction with vivid visualizations. The patient's imagination can reengage his or her feelings and return them to the body's organs and tissues. Instilling hope and giving encouragement allows the doctor's influence on the patient to be increased.

As the patient's Shen begins to open, it is important for the Qigong doctor to take the time to energetically and emotionally travel with the patient back into the memories of the past, to encounter the precise moment in which the patient's Hun left the body. The choice to return the emotional spirit back into the body must come from the patient's desire to retrieve his or her original feelings, no matter how painful the process. The patient must be ready to reexperience and release these feelings. It is important that the patient's "adult observer" be present in retrieving these painful memories. The "adult observer" refers to the patient's ability to suspend his or her judgemental attitude towards his or herself, and to observe the events that are being released with compassion (i.e., as if they were happening to a friend to whom loving support can be given). This allows the patient to experience true healing with the help of a compassionate companion (his or herself).

Tears often come as the memory of the original trauma releases from the patient's body or flows back into the Mind. Many patients notice that after these experiences their dreams change significantly, becoming more vivid. Also, their life environment as well as their relationships tend to change drastically.

EMOTIONAL REBIRTHING PATTERNS

Often, when patients experience deep emotional and spiritual traumas, they also experience physical sensations which reflect the trauma of their biological birth. These psychophysical sensations are due to the physical, energetic, and spiritual connection each individual has with the maternal host while inside the womb. Although these experiences are natural, they can sometimes be disturbing and even frightening to patients unfamiliar with deep emotional and spiritual transitions. The following are but a few examples of physical rebirthing patterns, and the emotional reactions related to these types of spiritual transformations.

- In patients born with the umbilical cord wrapped around their neck, each time they experience deep emotional and spiritual changes it will feel as if they are being strangled.
- In patients born through C-section, each time they experience deep emotional and spiritual changes it will feel as if they are being torn asunder.
- In patients born through the natural birth cannel, each time they experience deep emotional and spiritual changes it will feel as if they are falling through space with nothing to support them.
- In patients born breech, each time they experience deep emotional and spiritual changes it will feel as if they are being crushed in by all sides and suffocated.
- In patients with mothers who were heavily sedated, each time they experience deep emotional and spiritual change they will feel mentally disoriented and unstable.

SOUL RETRIEVAL MEDITATION

The following comes from my clinical experiences in dealing with patients who are very eager to heal, but reluctant, or unable to tolerate rapid change within themselves. The patients' family dynamics, as well as the quality of traumatic and nourishing experiences they received while growing up, play an important role in shaping their individual personalities. It is an accepted fact within modern psychology that an individual's present emotional life is shaped by the events occurring in the "formative" years (especially the years before the individual can articulate their thoughts and feelings). These experiences sink deep into the individual's psyche, hidden from consciousness through repression and denial. This repression is generally due to the inability to understand and express in words the traumas they suffered.

This particular Shengong (which literally means working with spirit) technique involves creative visualization (in a very relaxed, quiescent state of mind), and is used to gain access to the patient's unconscious past emotional experiences. The goal of this meditation is to allow the patient to free him or herself from the repressed emotional material which systematically surfaces during daily life. These programed emotional responses systematically affect the patient's behavior, causing neurotic symptoms, slips of the tongue, associated dreams, and fantasies.

Many patients believe they are ready for this experience, but in reality they are not yet able to deal with such intense pain. These patients are likely to unconsciously choose the least painful memories in order to test how the doctor will handle their feelings, or fail in the attempt to retrieve the lost parts of the soul. Gradually, as trust grows, these patients will be able to access more painful and relevant memories (this is similar to peeling an onion layer by layer). Sometimes patients try very hard to access the memories and obtain nothing, or very little information, leaving them feeling frustrated. This frustration must then be processed and resolved.

The doctor must instill hope and dispel this frustration by explaining to the patients that they have made an important step forward by seeding the unconscious mind with the request, or concept, that past memories must be recalled. In time, this seed will sprout and give forth results. It just needs a little time to incubate. It is extremely important that the doctor never show disappointment or take away the patients' hope.

In such cases, the patient must be reassured that their higher-self (the observer self) senses that the timing is not right. There may be current issues or events, that must first be worked through for a successful soul retrieval to take place. To have done so now would have resulted in more trauma, or more precisely, re-traumatization. They may need to strengthened and develop more self-trust before trying again. Counseling may be recommended, and in time they will succeed.

Some patients may need several sessions before feeling comfortable enough with just entering and surfacing from trance states, before undergoing soul retrieval. It may be necessary to first retrieve pleasant memories that build self-trust and self-confidence before re-attempting soul retrieval (i.e., learning to fly a kite, etc.).

The Qigong doctor should have some training in hypnosis and hypnotic trances before giving this type of trance inducing meditation to their patients. The doctor should know what to do and be able to identify whether or not the patient is able to connect with their Yuan Shen or "higher self." As the doctor induces trance in the patient through this meditation, the following changes in the patient may be noted:

- a change of color in the complexion,
- a relaxation of the facial muscles,
- a change in breathing patterns, and
- sporadic eye movement (behind closed eyes).

If none of the above signs are noted, the patient may need more time to relax deeply before continuing with the meditation. Suggestions of imagining and feeling the body's tissues melting into the Earth may be given, along with suggestions of taking deep relaxing breaths to calm the patient's Shen.

While performing this meditation the patient may experience all the sensory, emotional, and physical sensations of these repressed memories, causing specific core memories (that have been repressed from infancy and childhood) to be brought to the surface and relived. As the patient experiences these memories, his or her facial expressions and demeanor, will often change, regressing to reveal the appropriate age and period of time during which the original trauma occurred.

This particular method of soul retrieval is divided into two stages. The first stage involves locating the body's three storage chambers of past, present, and future emotions. The second stage focuses on accessing the location of, and coming in contact with, the patient's unconscious repressed emotional traumas.

This meditation must first be understood in depth before it is given to the patient. It is very complex, and is not meant to be simply read through to the patient. Also, the doctor should set aside sufficient time to process what the patient experienced after ending the meditation.

STAGE 1. ACCESSING THE ENERGETIC CHAMBERS

The purpose of this meditation is to locate the patient's three storage chambers which access his or her past, present, and future emotions. This meditation also allows the patient to acquire a starting point, or outlet, for removing these painful experiences.

In this meditation, the patient first concentrates on feeling his or her entire body. Begin by having the patient sit comfortably in a chair, both feet on the floor, hands resting on the thighs, eyes closed, tongue on the upper pallet, and breathing naturally through the Lower Dantian. It is important that the patient relax, feel, and experience the body completely. The patient should also be encouraged to accept and feel everything as one unit, to connect with the Yuan Shen or "higher self."

The Storage Chamber of Past Emotional Memories

In order to find the chamber of past emotional feelings, the patient begins to focus attention on remembering and feeling a painful childhood experience. Once the patient has retrieved the memory and is experiencing the painful trauma, he or she points to the tissue area of the body where the feelings are located. This area is considered to be the current storage chamber, where the patient stores the energy of past emotional memories.

The storage areas change over a period of time, given the fact that the patient continues to process emotions throughout life. Each emotional healing causes the previous energetic holding pattern to die, and a new pattern is to be born within the patient's energetic structure. The chamber of the patient's past usually coincides with a tissue area where the patient is already experiencing disease or discomfort.

The Storage Chamber of Present Emotional States

In order to find the chamber of the present emotional states, the patient focuses attention on feeling the doctor's voice within his or her body. As the doctor continues to speak, the patient feels where the sounds are resonating. Once the patient experiences and locates this tissue resonation, he or she points to that area. This area is considered to be the current storage chamber, where the patient stores the energy of present emotional states.

The Storage Chamber of Future Emotions

To find the chamber of future emotional states, the patient focuses on a specific project that he, or she, is planning to do in the near future. Once the patient experiences slight physical sensations, he or she points to the area where the feelings are located. This area is considered to be the current storage chamber, where the patient stores the energy for future endeavors and future expectations. The patient's current ambitions may change as the patient begins to heal from past emotional wounds and makes life-style changes.

STAGE 2. RETRIEVING THE LOST MEMORIES

The final part of this meditation explores the patient's unconscious feelings and reprograms the destructive subconscious patterning. It is used to access the repressed memories stored within the patient's individual Personal Subconscious, which is hidden from through their denial or defense mechanisms. Sometimes after practicing this meditation, patients will not only gain access to the repressed memories of their infancy and childhood, but will also connect to their life within the womb, as well as their birth-process.

1. Begin by having the patient sit comfortably in a chair, both feet should be placed on the floor, hands resting on the thighs, eyes closed,

tongue on the upper pallet, anal sphincter lightly closed. The patient should breathe naturally, expanding the Lower Dantian with each inhalation, contracting the abdomen with each exhalation.

2. The patient focuses attention on the area of the body where the energy of the past emotions is stored. The doctor directs the patient to imagine opening the chamber of the past emotions, allowing the energy, like steam, to release itself from this area. This steam is gathered to the left side of the body. Once all of the energy has been released, it begins to form into a cocoon shaped cylinder of past emotional memories.
3. Next, the patient focuses attention on the area in the body that stores the energy of the present emotions. The patient imagines releasing this energy from out of the body like steam, creating a second cocoon shaped cylinder. The patient places this second cocoon in front of the body, next to the first cylinder.
4. The patient then focuses attention on the area in the body that stores the energy of future emotions. The patient imagines releasing this energy from out of the body like steam, forming a third cocoon shaped cylinder. The patient places this third cocoon to the right side of the body, next to the second cylinder.
5. Next, the patient imagines that all three separate energetic cocoons are joined together. The energetic cocoons are to be connected end to end, in a straight line, with the past on the left, the present in the middle, and the future on the right side of the patient's body.
6. The patient then focuses attention on his or her Taiji Pole, imagining the energy of the Upper Dantian descending and combining with the energy of the Middle Dantian (where the Eternal Soul is stored). Next, the patient imagines the collected energies of the Upper Dantian, Middle Dantian, and his or her Eternal Soul, descending and combining with the energy of the Lower Dantian. The patient uses the collected energies of the Three Dantians to create an "energy bubble" within the Lower Dantian to surround the Eternal Soul. While focusing on the center of the energy bubble, the patient imagines the Eternal Soul transforming into a small energetic form of his or herself.
7. Encapsulated within the energetic bubble, the small energetic form of the patient's self manifests within the physical body. The patient's Eternal Soul, now in small energetic form, begins slowly ascending the Taiji Pole, leaving the body through the Baihui GV-20 point. As the patient's Eternal Soul leaves, it is still connected to the physical body through a silver energetic cord.
8. The patient's Eternal Soul floats over the three energetic cocoons, hovering over the cocoon shaped cylinder on the left side of the body (to access the past emotional memories). Next, the patient observes the black, oval shaped cylinders, imbedded within the structures of the energetic cocoon of past emotional memories. These dark cylinders are energetic emotional clusters. They contain the memories of hurtful feelings and specific times in the life of the patient when he or she felt humiliation, degradation, shame, rejection, or emotional deprivation.
9. The patient descends into the energetic cocoon of past memories and allows the Eternal Soul to find the particular cylinder which currently needs addressing. In choosing a specific area to work on, the patient's Yuan Shen automatically selects the most relevant material from the patient's unconsciousness.
10. The patient then removes his or her Eternal Soul from the energetic cocoon, bringing with it a specific dark cylinder. The cylinder is removed from its webbing within the walls of the energetic cocoon and placed in front of the hovering soul. The cylinder is still connected to the energetic cocoon through an attached energetic cord. Through vivid concentration the patient focuses on the color of the cylinder emitting light from the Yin Tang (Third Eye) point of the little energetic self. This projected light pulsates and changes the dark cylinder, from black to white. As the cylinder pulsates and changes its color, it begins

to open and reveal the lost memories and unresolved feelings. These memories of emotional traumas are brought to the surface. Some of these experiences may be memories of events where the survival or integrity of the patient's physical body was threatened.

11. Next, the patient connects with the divine (God or "higher power") and begins the process of feeling, expressing, understanding, forgiving, and letting go of the emotional charge attached to the memories. This process of emotional purging is connected to the emotions stored within the patient's Five Yin Organs, and are expressed through the following stages:
    a. Anger, blame, and resentment are stored in the patient's Liver.
    b. Hurt, pain, sadness, and disappointments are stored in the patient's Lungs.
    c. Fear and insecurity are stored in the patient's Kidneys.
    d. Remorse, regret, and responsibility for keeping the painful memories and feelings alive are stored in the patients's Spleen.
    e. Love, understanding, and forgiveness are stored in the patient's Heart.
12. As the emotions begin to release themselves from the energetic cylinder, the energetic debris transforms into a vaporous cloud. After understanding and resolving the emotions contained within the energetic cloud, the patient divides the released energy into either a receiving or discharging cloud of energy.
13. At this point, the patient focuses on the energetic cloud and begins separating the pain and hurtful memories from the knowledge and wisdom gathered from the experience.
    a. First, the patient places those memories gathered from the knowledge and wisdom of the experience on the right side of the room, imagining them as sparkling gold, silver, and white light energy.
    b. Second, the patient places those memories gathered from the pain and hurt of the experience on the left side of the room, imagining them as thick, dark, black energy.
14. The patient focuses on the right side of the room, imagining this energy as bright, pure, and illuminating. The patient then inhales, absorbing the knowledge and wisdom gathered from the past experience into his or her small energetic body. Since this is actually the energy of the patient's soul, the patient's body immediately feels the emotion, energy and spirit becoming empowered, as it can now review these past experiences without the feelings of hurt, pain, and judgment.
15. Next the patient focuses on the left side of the room where the feelings of hurt and pain are stored. The patient imagines this hurt and painful energy as a dark, turbid cloud. Because the cylinder has a direct life line connected to the patient's past, the energetic cord attached from the cylinder to the energetic cocoon must be severed. Once the cord connecting the cylinder to the patient's past is severed, the patient is then truly released from the spiritual and energetic connection of the memory. This is initiated by imagining an angelic being, with the "sword of truth," severing the energetic life line connected to the patient's cylinder. The sword of truth represents the patient's decision to regain his or her power by deciding to no longer suppress and carry the painful memories. Once the energetic cord is severed, it causes the cylinder to vaporize, forming an energetic cloud. The patient then imagines severing any last connection or attachment to these toxic emotions, and releases this cloud of hurt and pain, visualizing it floating up out of the room, through the ceiling, out into space.
16. The patient then imagines that God (the patient's divine healing power) is seated in the Heavens, with ripples and pulsations of love and mercy emanating from the center of this divine presence. The patient visualizes one of these waves of love and mercy form the divine descending towards the Earth. As the wave of love and mercy touches the as-

cending dark cloud of hurt and pain, it immediately explodes and destroys the dark cloud. The moment that this dark cloud explodes, the patient sees the cloud's negative energy transforming into blue crystal drops of divine healing energy.

17. Next, the patient visualizes these bright, fluorescent blue, drops of pure, clean energy descending from the heavens and pouring down like rain on his or her small energetic self, saturating it completely. This energy immediately transfers into the patient's physical body. The patient inhales and absorbs this energy flowing into every tissue of the body, allowing the divine healing light to penetrate and radiate throughout his or her entire being.
18. Next, the patient observes the changes created within the three energetic cocoons. The energy in each cocoon will have shifted slightly, revealing an energetic alteration of the patient's present state of mind, and a change in the patient's future. By transforming the past, the patient has altered the present and initiated a change in the future course of events.
19. Finally, the patient returns his or her energetic bubble (containing the small energetic self) back down through the Baihui point into the Taiji Pole, drawing the energy back into the Lower Dantian. From the Lower Dantian the patient's soul energy will return back into the Middle and Upper Dantians, as well as the surrounding tissues. Finally, the patient reabsorbs the energetic cocoons back into the body, and returns to a quiescent meditation.

THE DOOR TO THE CHAMBER OF MEMORIES

A simple visual meditation can also be used to access the patient's suppressed past memories. This exercise is a form of guided meditation, in which the doctor leads the patient into the imagined hidden parts of his or her own soul.

Begin by having the patient lie supine. Ask the patient to close the eyes and imagine walking down a flight of stairs. Lead the patient through three separate doors, descending three separate staircases (this is performed in order to arrive at the patient's "chamber of memories").

After entering through the last door, the patient descends one final staircase until he or she reaches the chamber of memories. This is where all of the memories of the patient's past traumas have been kept, hidden deep within the subconscious mind.

Inside the Chamber of Memories, the patient discovers dark cylinders of past emotional pain. Each cylinder represents a specific trauma experienced and stored within the patient's energetic field. The patient then makes a decision to heal and free the body of the enormous energetic field used to suppress certain memories. To heal and dissolve the energetic cyst, the patient follows the same process of eliminating past memories (repeating section 8 through 15 from the previous meditation).

Once this process is completed, the Qigong doctor guides the patient back up the flight of stairs, through the three doors, and back to consciousness. Because these meditations have a profound affect on the patient's constitution (affecting their physical, mental, emotional, energetic, and spiritual nature) it is important for the patient to be referred to a psychotherapist while undergoing this type of Qigong therapy to further facilitate the patient's healing transformation.

FINDING THE LIFE PURPOSE

Another aspect of Soul Retrieval consists of an exercise that assists the patient in finding his or her purpose in life. Finding the purpose of the patient's life encourages interpersonal healing, and fills the empty space deep within the patient's heart. This in turn increases the healing potential by providing the patient with hope.

1. The first step is to ask patients to write down on paper what is was they wanted to be when they were very young. Have them list the people that they imagined or fantasized being while they were still children (e.g., a super hero, doctor, fireman, etc.). This exercise allows patients to access their emotional past, and to feel and recover pleasant, lost memories, and phantasies.

2. Next, have the patients link all the peak experiences of each imaginary character (e.g., a doctor and fireman possess power to save lives, and can be considered heroes, as they are both admired for their skills, etc.). This knowledge informs patients regarding the underlying nature of each imaginary character and may also reveal hidden talents, passions and ideas.
3. Then, have the patients write down a list of their personal priorities. This list should contain the patients' present state of active priorities, and their current value system.
4. Then, have the patients write down their own "code of conduct," i.e., their conscious value system. The patients' code of conduct represents the ethical standards by which they strive to act. This code of conduct influences feelings, thoughts, and beliefs. It affects the patients' personal priorities and reflects the quality and nature of their interpersonal relationships. It is from this personal code of conduct that they accept, or condemn, themselves and others. By physically writing down these moral conducts, patients can emotionally connect with new found feelings of moral integrity, as well as a establish a stronger connection with their True Self.
5. Finally, have the patients write down a "life purpose" statement, linking together the collected emotions gathered from the peak experiences of their childhood and their code of conduct.

It is important that patients know and feel that what they are writing is a true and accurate account of their thoughts and emotions. The patients' live purpose statement can always be modified and changed, as they undergo personal transformation and spiritual growth through experience and expanding knowledge.

SECTION V

DIFFERENTIAL DIAGNOSIS OF ENERGETIC PRINCIPLES

CHAPTER 20
INTRODUCTION TO DIFFERENTIAL DIAGNOSIS

DIAGNOSIS IN ENERGETIC MEDICINE

Chinese Medicine developed into a comprehensive and extremely effective diagnostic system, over many centuries of accumulated clinical experiences. The symptomatology to identify disease patterns and their underlying disharmony follows the way of nature in its relationship to cause and effect. To identify the patterns of disharmony, the Qigong doctor combines diagnosis of pathology and treatment principles into one continuum. The identification of the disease pattern is not only developed through categorizing a list of symptoms, but also through observing the pathogenesis of the disease. A symptom is a feeling, or reaction, that is a sign of disease, resulting from certain disorders (physical, mental, emotional, energetic, and spiritual).

Treatment must be aimed at finding and treating the cause of the disease. This is one of the most fundamental principles of Traditional Chinese Medicine in terms of differential diagnosis and treatment. The fundamental purpose of diagnosis is to find:

- the cause of the disease,
- the root of the disease, and
- the location of the disease.

Finding the primary cause (or root) of the disease requires understanding its nature, and allows the doctor to focus treatment on the most important etiological and pathological aspects of the disease. When diagnosing a disease, emphasis is placed on observing the energetic movement (or lack thereof), as well as energetic transformation, and not on the present set structures of the existing tissues. The tissues's structures are temporarily generated by the existing energetic changes and are therefore of secondary interest to the changes themselves.

ETIOLOGY IN MEDICAL QIGONG THERAPY

The etiology, or study of the causes of diseases in Medical Qigong therapy can be divided into what is commonly called in Traditional Chinese Medicine as the "three periods of life." These three periods encompass the developmental activities of the patient's Jing, Qi, and Shen formations during the transitional periods of their life, i.e., within the womb, childhood, and adulthood. Each patient's constitution results from the continuous interaction between his or her inherited form and the environment.

Within each period of life Three Stars develop and determine some of the major characteristics and traits of each person. Each group of Three Stars accounts for one third of what, and who, the person is. They combine together to form and complete a Nine Star System of life evaluation.

ETIOLOGY WITHIN THE WOMB

Problems arising during the Three Star Period within the womb (and during birthing) encompass the formative time span of 40 weeks. This time period can be influenced by the following factors.

1. The state of the patient's Jing prior to conception. If the Jing of one or both parents is deficient, the sperm and/or ovum will be weakened.
2. The mother's mental state during pregnancy has a deep affect on fetal formation, and directly affects the developmental activities of the fetus's Jing, Qi, and Shen. If the mother is malnourished, ill, emotionally disturbed, or taking alcohol or drugs during pregnancy the development of the fetus will be affected.
3. If the birth is abnormal, difficult and/or prolonged, it can cause an emotional shock within the newborn.

Diseases may originate from one or more of these factors causing the baby to be born with an illness or a predisposition to illness. This predisposition can be manifested as a particular type of emotional or mental imbalance, a diseases of particular Yin and Yang organs (or organ systems), or a general weakness as a whole. These conditions are commonly observed as "prenatal shock," and are indications of active fetal toxins. Prenatal shock can also be indicated when newborn babies are restless sleepers, continually opening and closing their eyes, and / or having fevers.

These problems are sometimes difficult to diagnose; simple observations however, can help the Qigong doctor to determine if the patient has experienced prenatal shock. A bluish color on the forehead of the patient, for example, is one indication of this condition.

ETIOLOGY DURING CHILDHOOD

Problems arising during the Three Star Period of childhood encompass the formative time span of 18 years. These internal problems can result from abuse, malnutrition, trauma and excess sex (during the teenage years). The subconscious mind of a child is extremely impressionable to both positive and negative influences.

If, as children, the patients experienced any physical, emotional, or sexual abuse, suffered a traumatic loss (of a relationship with either or both parents, siblings, peers, etc.), then they may become predisposed to illness later in life, or develop a disease during this time period. These traumas can affect the adolescent developmental activities of the patient's Jing, Qi, and Shen formations, as this is the child's greatest time of physical growth and development.

ETIOLOGY DURING ADULTHOOD

Problems arising during the Three Star Period of adulthood (from the age of 18 onwards), encompassing the time span of 60 years. These problems are caused by the patient's internal emotions and reactions to surrounding environmental factors. Environmental factors include: diet, relationships (be they social, work related, family or other), frequency of sex, and so on, and will affect the continual developmental activities of the patient's Jing, Qi and Shen formations. An adult, although less impressionable than a child, is less resilient than the child to trauma and illness.

PROGNOSIS

If the patient's constitution has been poor since birth, or if a specific disease has been present since birth, the prognosis is generally poor. If, however, the specific disease is acute, or has only developed during the patient's adult time period, the prognosis is more favorable.

There are always energetic interactions between these Three Periods of life. If, for example, a young female patient has a hereditary imbalance between her Conception and Governing Vessels, and also has severe emotional strain during puberty, she may suffer mental problems later on in her life.

THE GROWTH AND PEAK CYCLES OF JING, QI AND SHEN

By studying the physical, mental, emotional, spiritual, and energetic qualities inherent within the human body, the Qigong doctor is able to observe and predict the various changes according to the Nine Star System of life evaluation. It is believed that the body is given specific time periods in which to flourish during the three periods of life. The following chart depicts the body's natural time frame of each physical, mental, emotional, spiritual, and energetic peak (Figure 20.1).

- From birth to age ten represents a time of growth and development on each dynamic level. The child is in a formative stage, growing physically, mentally, emotionally, spiritually, and energetically.
- From age ten to age twenty represents a time of peak physical growth. Within this time frame the child physically enters into puberty, and is able to reproduce; the growing process is stabilizing.
- From age twenty to age thirty represents a time of peak mental growth. Within this time frame the adult actively acquires knowledge for survival sake. He or she is now able to work, produce results and begins planning for the future.

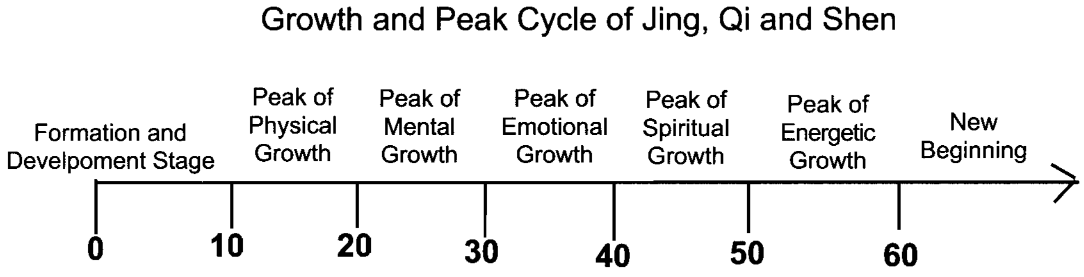
Growth and Peak Cycle of Jing, Qi and Shen

Figure 20.1. Chart of Physical, Mental, Emotional, Spiritual, and Energetic Peak

- From age thirty to age forty represents a time of peak emotional growth. Within this time frame the adult actively acquires knowledge for emotional comprehension of his or her surrounding environment. The adult is now able to creatively express his or her emotions in a civil manner and deal well with emotional disruptions.
- From age forty to age fifty represents a time of peak spiritual growth. Within this time frame the adult actively acquires knowledge for spiritual growth. Now comprehending his or her mortality, the adult begins to actively seek spiritual peace.
- From age fifty to age sixty represents a time of peak energetic growth. Within this time frame the adult's previous experiences come into fruition. If the adult has successfully acquired physical, mental, emotional, and spiritual harmony, the energetic peak will prepare the adult for the rebirth. If the physical, mental, emotional, and spiritual energy is not in harmony, the life-force is terminally exhausted, and the patient dies.

  It is commonly noted that when martial arts masters practice their training excessively or incorrectly, it is within this energetic time frame that they die. The cause of death is generally due to the over exertion of the master's Liver and Lungs, and the weakening of the Spleen, Heart, and Kidneys. This energetic disharmony stems from an improper balance of combat training with little or no spiritual training.
- From age sixty on is a time of new beginnings. This time frame represents new challenges of leadership within the family and the community.

DISORDERS OF THE HUMAN BODY

Disorders of the human body are generally due to three factors: constitutional failure of the body's Prenatal or Postnatal Qi, pathogenic factors and unrestrained Evil Qi, or one or more of the Eight Miscellaneous Factors.

CONSTITUTIONAL FAILURE OF PRENATAL AND/OR POSTNATAL QI

Prenatal and postnatal constitutional failure results in a constitutional Qi Deficiency, disease, or stagnation of the vital energy.

The Prenatal Constitution is determined by the strength of the parents when a child is being conceived, or during the pregnancy. If the parents are weak, sick, drunk, or on drugs during pregnancy or conception, the parents' Qi, Blood, and Body Fluids will be deranged and the child will be affected.

In uterus, if the mother is sick, the energy of the child will be affected. Also, negative emotions, poor nutrition, smoking, drugs, and alcohol used by the mother can affect the child within the womb and predispose the child to illness after birth (e.g., prenatal alcohol syndrome and crack babies are born addicted and with mental and physical defects). The prenatal constitution is inherited from both parents and is unchangeable.

The Postnatal Constitution is determined once the child is born. Proper nutrition and a nurturing environment enhances his or her constitution. While it is not possible to entirely erase the effects of the patient's prenatal constitutional patterning, it is possible to enhance it through a balanced life-style and development of their postnatal constitution.

PATHOGENIC FACTORS AND UNRESTRAINED EVIL QI

Pathogenic factors and Evil Qi are derived from Six External Factors and Seven Internal Factors.

The Six External Factors arise from the six climatic changes (Wind, Summer Heat, Damp, Dryness, Cold, and Fire). When the patient's energy system is not well balanced, or the Qi is Deficient, the body is unable to adapt to climatic factors. These climatic factors can then invade the patient's Wei Qi fields and penetrate deeply into the channels and internal organs of the body.

The Seven Internal Factors arise from excessive internal emotions (anger, joy, worry, grief, sadness, fear, and shock). These feelings overwhelm the patient's mental state as the individual responds to social, emotional, and environment interactions. Under normal circumstances, emotions in and of themselves do not cause problems, if they are allowed to be expressed and released. The excessive accumulation of one or more emotions can, however, cause Qi deviations, affecting the corresponding organ or organs of the body.

EIGHT MISCELLANEOUS CAUSES OF DISEASE

Sometimes a patient's disease is caused by what is known in Traditional Chinese Medicine as the Eight Miscellaneous Factors (Diet, Overexertion, Sex, Excessive Child Bearing, Exposures to Poisons, Parasites, and Iatrogenic Disorders). The Eight Miscellaneous Factors offset the patient's balance of Yin and Yang energy and are described as follows.

1. An improper diet can throw the patient's body and mind out of harmony. Nutritional problems can be divided into three main categories:
    a. The excess intake of certain particular foods, as well as overeating food in general.
    b. Unwise eating habits such as:
    - eating foods that are inconsistent with the seasons,
    - eating a diet of too many Cold, Damp, Dry, Hot, greasy, spicy or raw foods,
    - eating irregular meals,
    - oscillating from feasting to fasting, or
    - eating while hurrying or emotionally upset.

    c. Malnutrition, which is developed from a deficient intake of the basic requirements needed to maintain a balanced diet. This condition can be caused from ignorance, poverty, or an impaired digestion, absorption and metabolism.
2. Overexertion consumes the Qi and results in energetic deviations. This condition pertains to both mental/emotional and physical exhaustion, and can stem from loss of sleep, inadequate rest, etc.
    a. Mental and emotional overexertion can stem from a dissatisfaction with one's work (or the lack of work), or a dissatisfaction with one's life. It can lead to stress, frustration, boredom, apathy, or depression.
    b. Physical overexertion varies according to the individual's occupation, as each job has its own characteristics hazards to the patient's health. Problems arising in this area can range from structural damage to organs and tissues (caused from Qi and Blood stagnation), to sensory impairment.
    c. Lack of sleep and inadequate rest can also be either a contributing factor or direct cause in the development of Qi deviations.
3. An unbalanced sex life can cause disease. An individual's sex life is innately connected to the personal spiritual qualities of his or her Hun and Po. Therefore, the majority of sexual problems are rarely of purely physical origin, and reflect the energetic balance of each individual's emotional and spiritual passions.
    a. An Excessive sex life: what is excessive for one individual may not necessarily be excessive for another. The condition of Sexual Excessiveness should therefore be based on the individual's con-

stitution. When an individual exceeds his or her body's normal sexual function, the excessive sexual act can burn up the body's Yin, Jing, and Kidney Qi, causing Deficient syndromes (this condition is more prone in men then women). The act of sex itself is considered a Hot and Damp state.

   b. A Deficient sex life: when an individual abstains from sex, his or her Jing production begins to increase. If this increased energy buildup is not transformed into spiritual energy, it can overflow into the Sea of Marrow affecting the brain. The prolonged lack of sex sometimes creates an emotional need and dependency towards addictions to replace the need for intimacy.

4. Excessive childbearing injures the mother's Blood and weakens her Qi. Women suffer loss of Jing during childbirth, therefore excessive childbirth can cause a Deficient condition.
5. Traumatic accidents and injuries pertain to physical traumas which disperse or congeal the body's Qi and Blood, resulting in energetic deviations. Even after the physical trauma has healed, a site for potential weakness often remains. As the patient ages, the site becomes vulnerable to further injury due to old age, physical exhaustion, stress, and exposure to cold. All traumatic accidents and injuries also pertain to emotional trauma as they are energetically inseparable.
6. Exposure to poisons pertains to any substance taken into the body by ingestion, inhalation, injection, or absorption that interferes with the patient's normal physiological functions. These poisons include both chemical as well as environmental toxins (e.g., pesticides, carbon monoxide, electromagnetic fields, and radiation).
7. Parasites include infections and infestations, and pertain to the numerous organisms capable of living within the body.
8. Iatrogenic disorders pertain to any adverse mental, emotional, or physical condition induced through medical treatment.

THE DIFFERENTIATION OF SYNDROMES

The essence of Chinese medical diagnosis is holistic and based on the concept that all things and events are interrelated. Thus the Chinese physician takes into account not only the physical symptoms of the disease, but also spiritual, mental, emotional, and environmental factors or events, that both interrelate and interact to form "patterns of disharmony" within the patient. The doctor studies and seeks to understand these patterns of disharmony, called "syndromes," to provide the framework for treatment.

Syndromes can include External or Internal factors, congenital or acquired disorders, and symptoms of stagnation, Excess, and Deficiency, or traumatic injuries to the organs and tissues. The Qigong doctor must first determine whether the patient's disease is primarily due to an energetic disorder (e.g., an Excess or Deficient condition) or a physical disorder (i.e., structural damage). Once this determination has been made, the goal or goals of treatment can be set, and the means of treatment chosen. In the treatment of the lower back, for instance, if lumbar pain is due to a subluxation of the vertebra, caused by external injuries, the doctor should first reposition the vertebra through hand manipulations, by using Tui Na therapy, then remove the Qi stagnation. If the doctor is unfamiliar with such techniques, it is better that he or she first refer the patient out for spinal adjustment, and then remove the Qi stagnation. Only in this way can a satisfactory curative effect be obtained. If, on the other hand, the back pain is due to abdominal distention or a Kidney malfunction (i.e., primarily an energetic disorder) then treatment must proceed through Qi emission.

Several methods can be used singly, or in combination, to diagnose syndromes. They are described as follows.

1. Diagnosis according to the identification of the Eight Energetic Principles.
2. Diagnosis according to the identification of Qi, Blood and body fluid distinctions.
3. Diagnosis according to the identification of

the Five Elemental patterns.

4. Diagnosis according to the identification of the Six Stages.
5. Diagnosis according to the identification of the Four Levels.
6. Diagnosis according to the identification of the Triple Burners' System.

THE CATEGORIZATION OF SIGNS AND SYMPTOMS

In order to categorize symptoms into syndromes, the Qigong doctor must determine: The Onset, Location, Aggravating or Relieving Factors, Type of Pain or Sensation, and Course Since Onset.

1. The Onset includes the date the symptoms first appeared, whether they appeared suddenly, or had a gradual onset, and the order of their appearance.
2. The Location specifies whether the symptoms remain fixed, or migrate throughout the body.
3. The Aggravating or Relieving Factors indicate the specific conditions that affect the symptoms in a positive or negative way. Aggravating or Relieving Factors may include:
   - if the symptoms feel better or worse during daytime or nighttime,
   - if the condition is better with application of warmth or cold,
   - if the condition is better with pressure or worse with pressure,
   - if the condition is relieved by movement or aggravated by movement, and
   - if the condition is relieved by eating or made worse by it.
4. The Type of Pain or Sensation describes the nature of the symptoms, such as:
   - sharp or dull
   - pounding
   - stiffness
   - burning
   - itching
   - intermittent (coming and going)
   - electrifying
   - pulling or shooting
   - nauseated
   - distended
5. The Course Since Onset includes the incidence, progress, and the effectiveness of the treatments received.

The syndromes do not equal the patient's disease, but are the patterns of the illness. In determining the patient's syndrome, the Qigong doctor takes into account that these patterns can occur simultaneously, and vary in degree of severity. The purpose for assigning the patient's disease a syndrome is not merely to classify the disease, but to understand its process. Not all the symptoms listed for a particular illness need to be present in order to properly determine the patient's syndrome.

DIAGNOSIS AND TREATMENT OF THE PATIENT

The assessment and evaluation of the patient generally proceeds in eight stages. The first seven of the eight stages are performed prior to the initiation of the actual treatment. The final stage is undergoing the treatment plan and the periodical reevaluation of the course of the disease since the beginning of the treatment (Figure 20.2). During these eight stages, the Qigong doctor assess: The Whole Person, The Patient's Constitution, The Signs and Symptoms of the Disease, The Syndrome, The Patient's Organs, The Etiology of the Disease, The Treatment Goal, and the Treatment Plan.

1. The Whole Person must first be evaluated according to his or her whole being. This process involves assessing the patient's:
   a. physical appearance (e.g., is the patient disheveled or appropriately groomed),
   b. energetic demeanor (is the patient very nervous, calm, manic, etc.),
   c. structural build:
   - according to the physical Yao formations,
   - introverted or extroverted features, and
   - Yin or Yang appearance.
2. The Patient's Constitution is assessed according to the Five Elemental Constitutions and their subdivisions of balance, and Yin or Yang predominant element.
3. The Signs and Symptoms of the Disease are then classified as:
   a. Objective Signs that are apparent to the doctor,

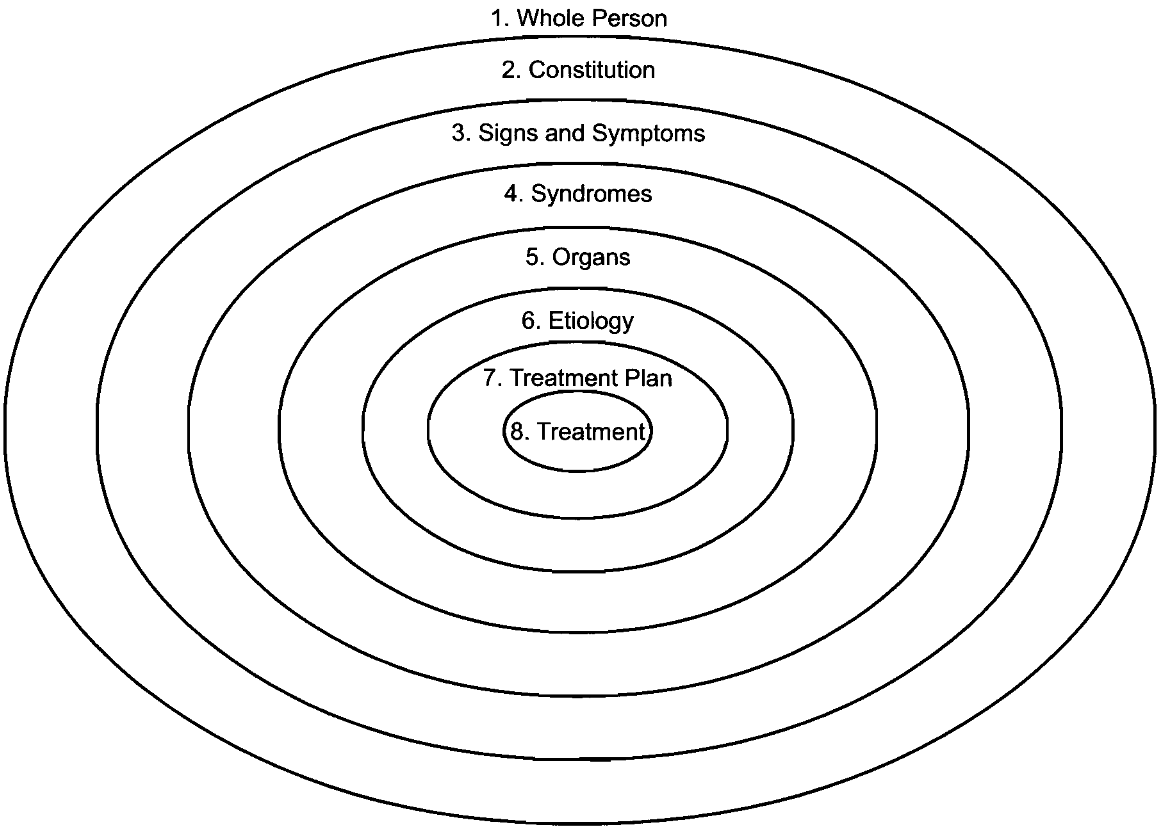

Figure 20.2. The Qigong doctor begins to diagnose the patient by first observing the patient as a whole. Next the doctor observes the patient's constitution, symptoms, syndrome, health and state of the internal organs, and the cause of the energetic dysfunction, before treating the disease. All the acquired knowledge is placed against the different diagnostic templates of the Eight Parameters, Five Element Theory, Qi, Blood and Body Fluid Dysfunctions, etc. This entire process is initiated in order to assist the Qigong doctor in prescribing the correct treatment. The treatment focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal being to relieve the patient's symptoms, strengthen the constitution, and return the patients to a state of wholeness.

---

   b. Subjective Symptoms that the patient reports to the doctor,
   c. Cardinal Symptoms and Signs that are used for differential diagnosis of disease, and
   d. Constitutional Symptoms and Signs that are indicative of a systemic disorder.
4. The Syndrome (a pattern of disease) is then determined according to the patient's symptoms and signs of disordered function, by assessing their relationship to one another.
5. The Patient's Organs are then evaluated according to any pathological changes occurring within the internal organs.
6. The Etiology (or root causes) of the syndrome is determined to complete the comprehensive diagnosis.
7. The Treatment Plan is then formulated, taking into account all the data gathered by the doctor. The Qigong doctor devises a treatment goal and determines a strategy of treatment

that addresses the patient's syndrome or disease, in accordance with the agreed upon goal and needs of the patient.

8. The Treatment is implemented and prescriptions are assigned to initiate the healing process. Over time, the treatment plan is reevaluated, as the patient improves or suffers setbacks (for various reasons). Prescriptions, likewise undergo changes in accordance with the patient's changing condition.

Once treatment begins, the process of the development of the disease is usually reversed, or the symptoms are alleviated. After the initial treatment begins, the causes of the disease are addressed, thus affecting the organs, which in turn changes the syndrome and the symptoms. This process strengthens the constitution and establishes a dynamic balance of health and well being.

UNDERSTANDING THE THEORY OF ENERGETIC COUNTERACTION

When treating patients it is important for the Qigong doctor to understand the theory of Energetic Counteraction. This theory describes the body's natural regulation, wherein no two specific symptoms (affecting the patient's general constitution) can occur at the same time, for any considerable period, within the same system.

Generally as a disease progresses, the old symptoms will give way to the newer ones. Consider, for example, a patient who has asthma, which is brought about by Wind Cold invading the Lungs (resulting in a dyspnea, wheezing, coughing, and severe mucous). If this patient unexpectedly suffers from an attack of gout (a painful inflammation and swelling of the joints), caused from the Wind Damp invading the joints and channels, the sudden attack of gout will cause the patient's asthmatic symptoms to disappear. When the gout subsides, the asthmatic symptoms will return. The new disease naturally prevails over the original condition.

CHAPTER 21
DIAGNOSIS ACCORDING TO THE EIGHT ENERGETIC PRINCIPLES

The method of diagnosis according to the Eight Energetic Principles was formulated by doctor Cheng Zhong Ling during the early Qing Dynasty. The Ba Gan or Eight Principles of Differential Diagnosis is a system using four pairs of opposite symptoms viewed in Traditional Chinese Medicine as being essential to the understanding and treatment of disease. In most cases, knowledge of the diagnostic systems and treatment principles of general Yin and Yang conditions is enough to give the Qigong doctor some good guidelines for steering the patient's training program. Similar principles can be found in most ancient healing traditions.

The four opposite pairs are Yin-Yang and the three subdivisions of Yin and Yang into: Internal-External, Hot-Cold, and Deficient-Excess.

1. Internal and External Symptoms inform the doctor of the origin of the pathogenic condition.
2. Hot and Cold Symptoms inform the doctor of the disease's onset (how it came on) and whether the patient's temperature has been altered due to a viral infection or a metabolic change.
3. Excess or Deficient Symptoms inform the doctor whether the condition is either to be tonified or purged, and the strength of the patient's constitution.
4. Yang and Yin Symptoms inform the doctor how to establish a physical (external), as well as energetic (internal) balance.

A detailed knowledge of the Eight Energetic Principles is required to pinpoint more subtle patterns of disharmony. This understanding allows the Qigong doctor to unravel complicated patterns and identify the basic contradictions within them. The main purpose for applying these principles is to understand the etiology of the disease and the nature of the dysfunctions; not simply to categorize the illness. Only then can a treatment pattern be initiated.

INTERNAL AND EXTERNAL PATHOGENIC FACTORS

Diagnosis according to internal and external pathogenic factors is not based on the etiology but on the basis of the disease's location. If a disease is caused, for example, by an "external" pathogenic factor but is currently affecting the internal organs it is classified as an Interior condition. An Exterior condition affects the patient's skin, muscles and channels, known as an "exterior pattern" of pathogenic factors. An Interior condition will affect the bones and internal organs, known as an "interior pattern" of pathogenic factors.

EXTERIOR CONDITIONS: SIX EXOGENOUS PATHOGENIC FACTORS

The external pathogenic factors—Wind, Cold, Dampness, Fire, Heat, and Dryness—are the six "evil" factors that come from the environment outside of the body. Under normal circumstances they are good for the health, but can be very harmful when they become abnormal.

Unseasonable weather (weather that is too cold or too hot for the body to tolerate) can make people ill by impeding the normal flow of internal energy thus making the body vulnerable to disease.

Each external factor, once it enters the body can manifest as an internal factor (i.e., External Wind can penetrate and become Internal Wind, etc.). It is also important to know that an external invasion once it penetrates the organ can transform into another internal factor (i.e., an External Cold invasion can give rise to Internal Heat).

These external factors are seen as not only the cause of disease, but also as manifestations of the disease. A disease may be due to any one of these Six External Pathogenic Factors, or simply manifest as an external "evil," even though that pathogenic factor was not present at the time the patient got sick (e.g., Damp-Heat). Treatment is therefore based on how a disease manifests itself, rather then the specific cause. However, it is wise to have a sense of the etiology of the disease to determine where to intervene and what might be expected next.

The Six External Pathogenic Factors are categorized in China according to the seasons, temperature, and climate. These transitions are explained as follows.

1. The Heat Season is characterized by temperatures that are usually above 90 degrees Fahrenheit. The Heart, Small Intestines, and blood vessels are the most vulnerable at this time. This is most evident in summer.

   Heat is a Yang pathogenic factor that rises and spreads excessively, consuming and diminishing the Prenatal (Yuan) Qi, Body Fluids and Lung Qi. As Heat rises, the body's Qi is drawn upward and dispersed, causing the Yin energy to descend. Symptoms can range from restlessness, anger, delirium, loss of consciousness to even coma. This condition is most evident during the summer. The Heart organ benefits from bitter tasting foods which directly stimulate and enhance the Heart's Qi.

   a. Mild Heat that has been created within the body, causes an ascending action of Qi. This ascending action causes the body's Yin energy to become disturbed creating an energetic imbalance.
   b. Fire (Extreme Heat) is a Yang pathogenic factor and spreads excessively, consuming and diminishing the Yuan Qi, Body Fluids, and the Lung's Qi. It produces an explosive reaction, causing the Qi to flow up into the head and disturb the Blood. This action causes symptoms ranging from nosebleeds, skin infections, swelling, and skin eruptions, to hemorrhaging and vomiting Blood.

2. The Damp Season most often affects the Spleen, Stomach and muscles that are at their most vulnerable time. The Spleen organ benefits from sweet tasting food. The temperature is usually between 75 and 90 degrees Fahrenheit. This is evident in the late summer, during the rainy season.

   Dampness is a Yin pathogenic factor that easily obstructs the body's production of Qi and impairs the Yang. It is considered a Yin pathogenic factor because, as the invasion of Dampness moves downward, it manifests as a heavy and turbid energy, resulting in a full chest, nausea, vomiting, and a sticky sweet taste in the mouth. During the damp season, an invasion of Dampness may affect the urinary organs, causing stagnation of Qi. Once pathogenic Dampness enters the body, it becomes foul in nature. If the body becomes obstructed, the results will become evident in skin diseases with abscesses, and oozing ulcers with turbid, cloudy urination. Disease also manifested through infections, in combination with Heat.

3. The Dry Season is characterized by temperatures that are usually between 60 and 70 degrees Fahrenheit. The Lungs and skin are the most vulnerable at this time. This is most evident in autumn, because of the lack of humidity.

   Dryness is a Yang pathogenic factor because as the Heat rises, dryness consumes the body's Yin Qi, especially in the Lungs. This results in dry, chapped skin, cracked skin, and dry cough with sputum, throat pain, or asthma. The Lungs themselves benefit from pungent tasting foods, which directly stimulate and enhance the Lung's Qi.

4. The Cold Season is characterized by temperatures that are usually below 60 degrees Fahrenheit. The Kidneys, Bladder, and bones are most vulnerable at this time. This is most evident in the winter.

   Cold is a Yin pathogenic factor characterized by astringency of the tissues. During the Cold season the Yin pathogenic factors consume the Yang Qi. Cold blocks the channels and collaterals, causing pain and damage to the Heart, as well as consuming the Kidney Yang. Pathogenic Cold also causes the channels and collaterals to contract, resulting in the slowing down of Qi and Blood circulation. As the Yin energy moves downward consuming the body's Kidney Yang, the body's Yang is unable to generate heat causing chills, pallor, shivering, diarrhea with undigested food, stiffness, pain, numbness, and clear urine. Internal Cold is characterized by conditions of overall stagnation and contraction. The Kidneys themselves benefit from salty tasting foods which directly stimulate and enhance the Kidney's Qi.

5. The Wind Season affects the Liver, Gall Bladder, and tendons that are at their most vulnerable time. The beneficial taste is sour. The climate in the Wind season, depending on the area, is usually between 60 and 70 degrees Fahrenheit. This is most evident in the spring. Wind is Yang in nature and tends to injure the patient's Blood and Yin. It is divided into both internal and external pathogenic Wind conditions. External Wind is called "the mother of one thousand pathogenic factors," and is responsible for the propagation and growth of the pathogenic environment, because it combines with other pathogenic factors to create Wind Heat, Wind Cold, Damp Wind, Dry Wind, etc.

   In the Wind season the Yang pathogenic factors are in the upper portion of the body, weakening the Wei Qi, and causing derangement of the opening and closing of the pores on the body's surface. The pathogenic Wind tends to move constantly causing abnormal motion and rigidity in the trunk and limbs. External Wind affects the upper part of the body, shifting and moving the pain like blowing a leaf. Once External Wind invades the body, the symptoms are stiff neck, itchy sore throat, watery eyes, headache, nasal obstruction, facial puffiness, aversion to Wind, irritability, abnormal sweating, migrating joint pain, tremoring, convulsions, rigidity, and facial paralysis.

   Internal Wind is manifested through the same symptoms. Because of its serious nature, it can also cause tremors, convulsions, and paralysis throughout the body (with the exception of the face, which is due to External Wind invasion).

INTERNAL PATHOGENIC FACTORS

An interior disharmony is diagnosed when the internal organs have become affected by pathogenic factors, creating an Interior syndrome. Interior syndromes are the pathological conditions which have resulted from one of three factors: (1) the invasion of the channels by external pathogens, which can be transmitted to the body's Yin and Yang organs; (2) the direct attack on the Yin and Yang organs by external pathogens; and (3) sudden emotional changes due to improper diet and stress which directly affect the body's Yin and Yang organs leading to functional disturbances.

1. A Persistent Exterior Invasion by one of the Six Exogenous pathogenic factors can penetrate deeper into the body creating an Interior syndrome, if it is not purged from the patient's body.
2. A Direct Invasion of the Yin and Yang organs can also occur from an Exterior Invasion of pathogens. In such cases, the affected organ will manifest as an Interior Excess or Deficient condition depending on the syndrome.
3. Emotional Factors can likewise cause damage to one of the Yin organs, which in turn can cause damage to the other Yin and Yang organs. People encounter emotional upheavals in everyday activities. Under normal circumstances, the Seven Emotions are good for health (see Chapter 21). Appropriate anger helps to disperse stagnant Liver Qi, preventing stasis of Liver energy and helps the patient to establish a healthy boundary system. Rage, on the other hand, leads to the abnormal increase and ascension of Liver Qi, resulting in a flushed face and dyspnea (shortness of breath). If the pattern continues (i.e., the person is a rage-aholic), the violent and uncontrolled anger may lead to serious disorders of the Liver, including eye injury, fainting due to inadequate Blood flow to the brain,

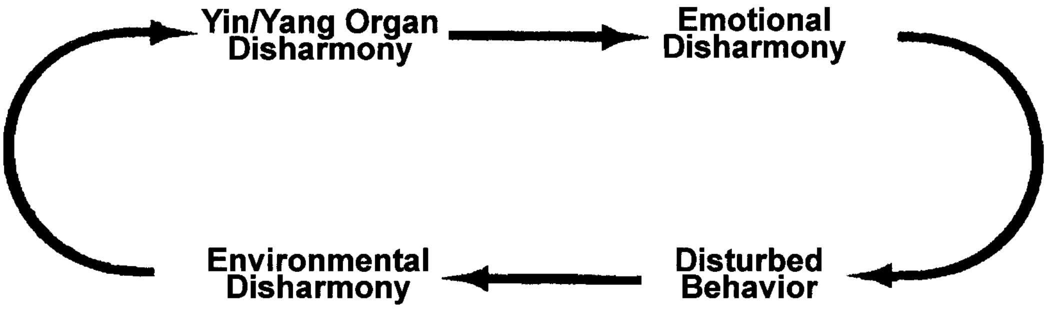

Figure 21.1. Cycle of Disharmony

cerebral hemorrhaging, or even death. It is believed therefore, that Internal conditions can arise from either an Excess emotional condition caused by emotional suppression, or a Deficient emotional condition caused by excessive emotional discharging.

It is also believed that a person's personal emotional history determines his or her biology (life processes). Each person creates this reality through his or her belief system. Hence the patient creates the disease by embracing traumatic past wounds and formulating belief structures that support these traumas. These energetic clusters are then fed via the major viscera by the excessive emotional energy. This pathological emotional process can result in the formation of cysts, tumors, cancer, etc.

Any imbalance of the Seven Emotions leads to the unnecessary consumption, or blockage of, the internal Qi flowing through the channels of the viscera. In either instance, this results in abnormal Qi activity and creates a functional disorder of the cerebral cortex.

Emotional disharmony induces a vicious cycle which affects the mind, body and spirit. Any factors (Yin and Yang disharmony, emotional disharmony, disturbed behavior, or environmental disharmony) can trigger the cycle (Figure 21.1). When caught in this cycle, emotional energy can deplete the physical body.

Anger, for example, generally causes the Qi to rise; however, when anger turns inward it can transform into depression. Mental depression may obstruct the functional activities of Qi and this stagnation may cause Fire. The Fire syndrome is caused by the disorders of the Seven Emotions and may burn up the fluid of the organ related to the predominant emotion.

In Western culture, people are generally not encouraged to acknowledge painful emotions; feelings are usually disregarded and conquered by a strong will. The storage of emotional energy in the muscles and organs of the body leads to tension, stress, and illness. The Qi cannot circu-

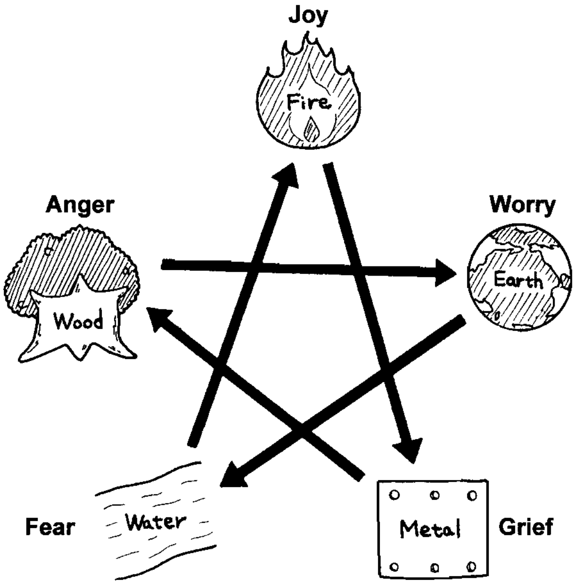

Figure 21.2. Five Elemental Controlling Cycle

late with emotional blocks. Many specialists in Western medicine today agree that a large percentage of diseases being treated are, indeed, caused or aggravated by stress.

The Five Elements have a creating (generating) and restrictive (controlling) relationship. The sequence of the creative relationship of the Five Elements is: Wood creates Fire; Fire creates Earth; Earth creates Metal; Metal creates Water; Water creates Wood. This circle is endless.

The sequence of the controlling relationship of the Five Elements is: Wood controls Earth; Earth controls Water; Water controls Fire; Fire controls Metal; Metal controls Wood. This circle is also endless.

The controlling relationship can be applied to problems with emotions. Anger, for example, belongs to the rising of Excess Liver energy and can be counteracted by descending the Lung's Qi (Figure 21.2). The Lungs control sadness, and sadness may lead to the consumption of abnormal rising Liver Qi (Metal controlling Wood).

Excessive sadness may lead to the extravagant consumption (and depletion) of Lung Qi, which damages the body's Yin and the Heart. The consumption of the Lung Qi can be stopped by the Heart's energetic release of the emotion of joy. Joy causes the rapid rising of the Heart's Qi. It can be regulated by fright and terror, which are controlled by the rapid descending energy of the Kidneys. Terror leads to the abnormal falling of Qi, it therefore counters the abnormal rapid rising of Heart's Qi. Excessive fright and terror may lead to the abnormal sinking of Gathering Qi, which damages the Kidney's Qi (Figure 21.3).

Because over-thinking can cause Qi to gather, it can be used as a tool to help bring back dispersed energy. Over-thinking and anxiety, however, may lead to the depression and stagnation of Qi, which weakens the transporting and converting ability of the Spleen's energy, and causes a lack of appetite. Mild cases of this condition can manifest as dyspepsia (indigestion), abdominal distention and diarrhea. In severe cases, the Qi

The Seven Emotions

Chinese Word	English Translation	Organ Association	Qi Manifestation
Xi	joy, happiness, excitement, pleasure	Heart	slows down and relaxes the Qi
Nu	anger, irritation	Liver	causes Qi to rise
Si	meditation, contemplation, pensiveness, worry	Spleen	stagnates the Qi
You	anxiety, sorrow, grief	Lungs	obstructs the Qi
Kong	fear, extreme anxiety	Kidneys	causes Qi to descend
Jing	shock (sudden intense fear), fright	Heart then Kidneys	scatters and deranges the Qi
Bei	grief, affliction, sadness	Lungs	disperses and consumes the Qi

Figure 21.3. The Seven Emotions Chart

and Blood stagnate in the chest and abdomen, causing a feeling of fullness in the chest or upper abdomen. This stagnated, ever-sinking Spleen energy can be smoothed and aroused by anger (the Liver's energy expanding upward and outward). In short, the hyperactivity of Yin is treated with Yang to check the unnecessary consumption of internal energy, and to promote a mild Yin energy. Then the Qi and Blood can flow without interruption and disease can be prevented.

One effective technique to avoid an emotional and energetic overload or depletion of a Yin organ's Qi, is to apply the Five Elemental counter-emotion to restore emotional balance (see Chapter 53).

OVERVIEW OF EXTERNAL AND INTERNAL CONDITIONS

The internal and external diagnostic parameters locate the depth of the condition of the disharmony and are fairly easy to distinguish, relative to the other pairs of opposites. Thus, determining the internal or external nature is usually the first step in diagnosis of a disorder.

Internal conditions are usually caused by internal disharmonies, such as excessive emotional energy in the form of anger, excitement, sorrow, anxiety, fear, and mourning, or other imbalances associated with the Seven Emotions. Excessive thinking and stress can also lead to an imbalance that is characterized as internal. Internal conditions are often chronic and tend to have a gradual onset and a longer duration.

Conditions of External disharmony are typically caused by the external influences on the body, known in Chinese medicine as "pernicious influences." They are Wind, Cold, Heat, Damp, Dryness, and Summer Heat. Pernicious influences usually invade the body when the body's Wei Qi is weak.

Typical External diseases are colds, influenza, skin eruptions, or diseases caused by injury or shock. Excessive amounts of eating, drinking, working, and sexual activity can also lead to both Internal and External disharmony. External conditions are often acute, with sudden onset, and of short duration.

Many times Internal conditions will develop when an External condition moves inward. This represents a further weakening of Wei Qi and the body's natural immune system. Before the condition has developed into a full Internal sickness, it will go through a stage of being half-external and half-internal.

Examples of conditions which may be considered as being half-internal and half-external are: alternating fever / chills, heaviness in the chest, restlessness, nausea, vomiting, no appetite, dry throat, irritation of the mouth, and dizziness. A bounding pulse is also an indication of this type of condition.

This is one reason why treatment of an External disease (colds and flu) requires sweating therapy to encourage the disease to remain external. The contractile mechanism of the skin is a specific external representation of the body's Yang immune system at work.

When treating extreme cases, such as internal injuries and shock, conserve the patient's Yang Qi by keeping them warm.

HOT AND COLD

Hot and Cold describes whether the condition is the result of a Hot or Cold state and is determined by hot or cold sensations. The nature of this clinical manifestation depends on whether it is combined with a Full or Empty condition.

HOT CONDITIONS

Hot conditions usually arise when there is excess activity of the body's Yang functions, or if there is an insufficiency of Yin Qi or Fluids in the body. A Hot condition can be further divided into either Full Heat or Empty Heat.

1. A Full Heat conditions manifest as the following physiological and mental / emotional symptoms.
    a. Physical symptoms include:
    - increased energy and metabolism, i.e., physiological hyperfunction
    - fever with thirst, and desire for cold liquids
    - dry stool and constipation
    - dark yellow and reduced urine
    - flushed face, red eyes, hot red skin
    - warm extremities
    - aversion to warmth
    - a rapid, full pulse and a red tongue with yellow fur

b. Mental and emotional symptoms include:

- irritability
- restlessness
- loud voice, talkativeness
- extroverted and aggressive behavior

Inflammations, ulcers, and infections are typical of Full Heat diseases. There are many other symptoms depending upon which organ is affected.

These Full Heat symptoms arise when there is an Excess of Yang energy in the body. Excess Yang energy can be caused from overconsumption of Hot foods (which cause Stomach or Liver Heat), or long-standing emotional problems (which cause Liver or Heart Heat) due to Qi stagnation. Full Heat can also develop from the invasion of external pathogenic factors which have transformed into Heat.

2. An Empty Heat conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- dry mouth and throat (at night)
- night sweats
- afternoon fever
- a feeling of heat in the chest, palms and bottoms of feet (Five Palms Hot)
- dry stool
- scanty, dark urine
- a floating, empty and rapid pulse, with a red, peeled tongue

b. Mental and emotional symptoms include:

- fidgeting and mental restlessness
- vague anxiety
- insomnia

There are many other symptoms depending upon which organ is affected.

These Empty Heat symptoms arise from a Deficiency of Yin. If the patient's Yin becomes Deficient (usually due to a Deficiency of Kidney Yin), this affects the Yin of the Liver, Heart, and Lungs and causes a relative Excess Yang condition.

COLD CONDITIONS

Cold conditions usually arise when there is excess activity of the body's Yin functions, or if there is an insufficiency of Yang Qi in the body. A Cold condition can be further divided into either Full Cold or Empty Cold.

1. A Full Cold conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- diminished physiological function and lowered immunity
- decreased energy
- aversion to cold and craving for warmth
- cold extremities
- lack of thirst
- clear urine, watery stool
- desire for hot drinks
- white face
- a slow pulse and a pale, white tongue with moist fur

b. Mental and emotional symptoms include:

- timid behavior and soft speech
- slow movement
- a lack of motivation

Generally, discomfort is increased by cold and reduced by warmth. Cold symptoms may also appear at later stages of fevers and chronic peptic ulcers, denoting a weakening of the body's ability to overcome the disease.

These Full Cold symptoms arise when there is an excess of Yin Qi in the body. Excess Yin Qi can be caused from the direct invasion of External Cold into the body's interior. In this particular condition, the External Cold can either invade the Stomach causing vomiting and epigastric pain, invade the Intestines causing diarrhea and abdominal pain, invade the uterus causing dysmenorrhea, or it can invade the Liver Channels causing pain and swelling in the scrotum. One of the main manifestations of Interior Full Cold is abdominal pain caused from the Cold constricting and obstructing the circulation of Yang Qi. Full Cold can also develop from the invasion of external pathogenic factors which have transformed into Cold.

2. An Empty Cold conditions manifest as the following physiological and mental/emotional symptoms.

a. Physical symptoms include:

- cold limbs and chills
- pale face
- fatigue
- no thirst
- loose stools
- clear, abundant urination
- sweating
- pale tongue with thin, white coating and a deep, slow or weak pulse

b. Mental and emotional symptoms include:

- tiredness and lack of motivation
- slow movement

An Empty Cold disharmony usually arises when the Yang Qi of the body is insufficient, and fails to warm the body. This is caused by a Deficiency of Spleen Yang, Kidney Yang, or Heart Yang (sometimes Lung Qi Deficiency). The most common cause is related to the Spleen Yang Deficiency, which fails to warm the muscles and thereby causes chills. If the Spleen does not receive heat for its function of transforming food, the result will also be loose stools.

COMBINATION OF HOT AND COLD SYMPTOMS

While distinguishing Hot diseases from Cold diseases may appear to be fairly straightforward, the doctor may find that there are many circumstances where both Hot and Cold symptoms appear simultaneously. The use of sensory input is just a general indicator, however, because it provides the doctor with only partial data; it can sometimes be misleading if other considerations are not taken into account. The doctor needs to consider the other symptom characteristics according to Yin-Yang or Deficient-Excess patterns.

1. An External Cold with Internal Heat syndrome manifests when a patient has a preexisting internal Heat condition and is subsequently invaded by External Wind Cold. This condition can also occur in attacks of latent Heat combined with the invasion of Wind Cold. Symptoms can include fever with an aversion to cold, body aches, irritability, and thirst.
2. An External Heat with Internal Cold syndrome manifests when a patient with a preexisting Cold condition is attacked by exterior Wind Heat. Symptoms can include fever with an aversion to cold, chills, headache, sore throat, and thirst.
3. A Heat Above-Cold Below syndrome manifests when a patient's pathogenic Heat rises resulting in Heat above and a Cold syndrome below. Symptoms can include thirst, bitter taste, sour regurgitation, irritability, Herpes Simplex I, borborygmus, loose stools, and pale and profuse urination.

Qigong patients must be aware of these signs of Hot and Cold disharmony. If a patient's training routine is too Yang, this will burn the body's Fluids, thus diminishing the Yin Qi's ability to contain the Excess Yang, which in turn produces a Hot condition. An overactive metabolism can also induce a Hot condition, while a sluggish metabolism can produce a Cold condition. Less extreme conditions are classified as warm, cool, or neutral.

EXCESS AND DEFICIENT

The capacity of an individual to maintain resistance to the invasion of pathogens is relative to the strength of the disease and the strength of the individual. An Excess condition is characterized by the presence of a pathogenic factor (interior or exterior) and by the fact that the body's Qi is still functioning normally. The battle against the pathogenic factors results in the symptoms and signs of an Excess condition. A Deficient condition is characterized by a weakness of the body's Qi and the absence of a pathogenic factor. If the body's Qi is weak, but the pathogenic factor continues to attack, the condition is considered Deficient and is characterized by an Excess.

EXCESS CONDITIONS

An Excess condition will usually occur when a bodily function becomes overactive or Qi accumulates unnaturally due to a blockage. It is characterized by strength, buoyancy of spirits and often acute symptomatic reactions. Disharmonious conditions which begin and end suddenly are due to Excess. Yang Excess can be quickly improved, in most cases, with a diet of predominately vegetables and fruits.

It is not unusual for patients whose diets are heavy in meat proteins, to suffer from Excess diseases. Many Naturopaths advocate the predominant use of raw fruits and vegetables and have experienced outstanding results due to the elimination of Excess Yang (from meat toxins) from the body.

Such a diet is not good therapy for individuals who eat a relatively balanced vegetarian diet, because it serves to make them more Yin and threatens to throw them off balance. This is a good reason to avoid "fad" diets. A diet or method that works well for one person may have the opposite affect for someone else; each case must be considered separately.

Some indications of an Excess condition are as follows: flushed face, loud voice, coarse and full breathing, acute pain, extreme menstrual pains, a tender abdomen, which worsen with the application of heat. A strong pulse can also be a sign of an Excess condition.

In cases of Excess, to expel pathogenic evils, the patient imagines that the Toxic Qi is being expelled from specific channel points upon exhalation. The results are always more effective when Qigong prescriptions are given in conjunction with the treatments.

Any interior pathogenic factor (Heat, Cold, Damp, Wind, Fire, and Phlegm) can give rise to an Excess condition, including the stagnation of Qi and stasis of Blood.

DEFICIENT CONDITIONS

Although Deficiency is characterized by insufficient Jing, Qi, Shen, Blood, and Fluids, it does not generally mean a lack of some nutrient as it does in Western medicine. It is more a statement about the body's inability to find, or produce, what is necessary for its immune function to provide a functional integrity in the body. Deficiency, therefore, is taken in a more general sense and denotes the overall condition of the patient, or of the disease.

Conditions which linger are usually due to Deficiencies. The general indications are: great weakness, inconsistent energy, shallow and rapid respiration, pain characterized by soreness, weak or inconsistent volume of voice, moderate menstrual pains, withered face, low spirits, and abdominal pain. A Deficient condition will generally respond well to heat and pressure. A weak pulse can also be an indication of a Deficient condition. There are four types of deficiencies: Deficient Qi, Deficient Blood, Deficient Yin, and Deficient Yang.

1. Deficient Qi is the first and least severe stage of Deficiency. Most symptoms arise from a weakness of the Lungs' Qi (which fails to control the breathing) and a weakness of the Spleen's Qi (which fails in its function of transforming and transporting).
2. Deficient Blood is caused from a dysfunction of various internal organs. Most symptoms arise from a weakness of the Liver, Heart, and Spleen.
3. Deficient Yin is caused by a dysfunction of various internal organs. Most symptoms arise from a deficiency of Yin within the patient's Liver, Heart, Stomach, Lungs, and Kidneys.
4. Deficient Yang is caused by a dysfunction of various internal organs. Most symptoms arise from a deficiency of Yang in the patient's Liver, Heart, Spleen, Lungs, and Kidneys.

In cases of Deficiency, when replenishing Qi is required, the best effect can be obtained when the patient cooperates with the doctor's extension of Qi through mental concentration and focused inhalation.

COMBINED EXCESS AND DEFICIENT CONDITIONS

Sometimes Excess and Deficiency conditions can occur simultaneously. Although Excess is one of the characteristics of a Yang disease, this condition can also give rise to certain symptoms of Deficiency. These are mostly a result of Qi being blocked in the utilization of its own vital nutrients.

YIN AND YANG THEORY

Yin and Yang energy is usually represented through the image of the Taiji symbol. Yang Qi is represented through the color white and Yin Qi is represented through the color black. Through the image of the Taiji symbol the interrelationship of Yin and Yang represent the stages of waxing and waning of energy (Figure 21.4). In Traditional Chi-

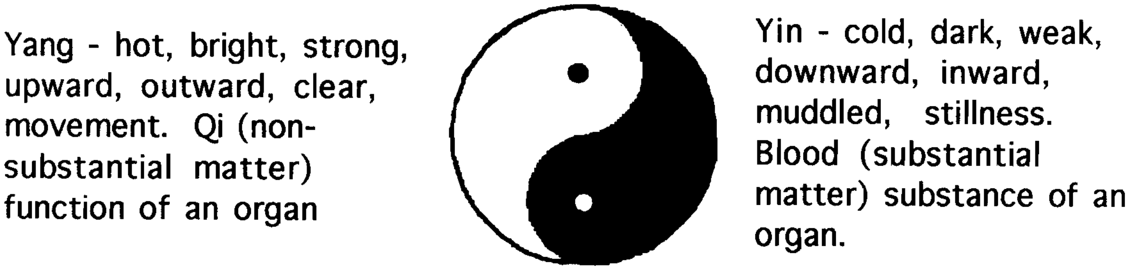
Yang - hot, bright, strong, upward, outward, clear, movement. Qi (non-substantial matter) function of an organ
Yin - cold, dark, weak, downward, inward, muddled, stillness. Blood (substantial matter) substance of an organ.

Figure 21.4. In Traditional Chinese Medicine the theory of Yin and Yang represent the dynamic duality of balance and harmony within the body, as well as within the universe itself, and as represented through the image of the Taiji Symbol. There is always Yin within Yang and Yang within Yin. These two energies are always changing and transforming the body's life force energy.

nese Medicine, the theory of Yin and Yang is also used to explain both the psychological and pathological phenomena of the body. It is considered a major principle for diagnosing and treating diseases.

Generally speaking, an Excess of Yang may lead to a Deficiency of Yin, and vice versa. A Hot syndrome can be the result of either Excess Yang or Deficient Yin. A Cold Syndrome can be the result of either Excess Yin or Deficient Yang (Figure 21.5).

The Qi activities of these Yin and Yang aspects unite and regulate the body's internal organs and tissues. The direction and energetic nature of Qi (Excess or Deficient, strong or weak, evil or vital) can be detected at corresponding body surfaces so that information for diagnosis and treatment can be obtained. For example:

A strong, dense, dry, and hot sensation of Qi in the head area may be a symptom of Excess Yang in the upper portion of the body and a Yin Deficiency in the lower portion. This would require the Qigong doctor to guide the Yang Qi to descend and the Yin Qi to ascend to restore the balance of Yin and Yang energy.

To treat patients with an overabundance of Yang Qi and hyperactivity of Fire, the Qigong prescription meditations should be practiced during the Yin periods (11 a.m. to 11 p.m.) while facing north, with emphasis placed on exhaling to purge the Heat. The doctor should adhere to the principle of replenishing the patient's Yin to regulate the Yang, or leading the Yang to descend to supplement the Yin.

Once the nature of Yin and Yang disequilibrium is determined, the proper method of treatment can be selected to regulate the Excess or Deficiency. The doctor always treats Yin for Yang diseases and Yang for Yin diseases. A Yang syndrome, for example, should be treated with Yin herbs, and a Yin syndrome should be treated with Yang herbs. The treatment should be aimed at purging the Excess, replenishing the Deficiency, and regulating the Deficient Yin or Yang to restore a normal balanced state of energy with in the body.

In Medical Qigong therapy, the principle of "using Yin for treatment of a Yang disease and using Yang for treatment of a Yin disease" must be strictly adhered to. It is important to dredge or purge the Excess before replenishing and tonifying any Deficiencies. The intake of cold foods (fruit, salad, ice cream, etc.) or drinks, for example, may lead to an Excess of Yin, which gives rise to a Cold syndrome of the Spleen and Stomach. This manifests as pain, diarrhea, aversion to cold, and cold extremities.

To treat this Excess, the doctor guides the Cold Qi out of the body by way of St-36 and Sp-9 points, then projects Hot Yang Qi towards specific channel points on the patient's body, such as CV-12 and St-25.

DIAGNOSING YIN AND YANG IN THE BODY

Every person has both Yin and Yang elements, but will tend to be predominantly one or the other in terms of personality, physique, life-style pref-

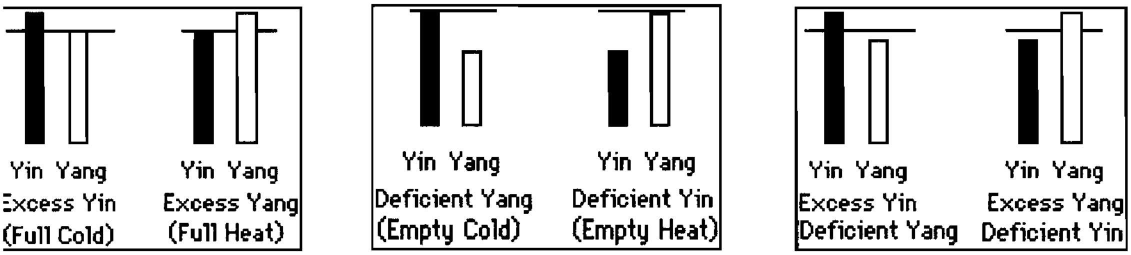

In cases of Excess, the doctor should Purge.

In cases of Deficiency, the doctor should Tonify.

In cases of Excess and Deficieny combined, the doctor should Regulate.

Figure 21.5. Yin and Yang Energetic Diagnosis

erences, speech patterns, mannerisms, etc. Within a general constitution there will be more subtle fluctuations within the Yin/Yang continuum as the body reacts to external and internal energetic movements. In designing the patient's exercise program, diet, or herbal remedies, it is necessary to first determine whether the patient is predominantly Yin or Yang in constitution.

YIN/YANG CONSTITUTION CHART

There are general indications that will help the doctor diagnose a patient's condition as predominantly Yin or predominantly Yang. The patient's basic Yin/Yang constitution, and an analysis of the current Yin/Yang balance of the body will play an important role in determining the type of exercise and the time of day that it will be performed. In most cases, a Yang condition should be balanced by a Yin exercise and vice versa.

There are varying levels of disharmony in the patient's body. Yin and Yang are only relative terms, and they are often not enough to adequately describe the character of the body's condition. Other factors such as Hot - Cold, Deficient - Excess, Internal - External must be considered in more detail. Yin and Yang are never in a permanent state; there is always dynamic movement. All relationships based on Yin/Yang are relative, and their mutual interaction must be considered.

Yin and Yang theory is valuable because it provides a tool for understanding the patient's constitution at a given point in time. This information enables the doctor to determine the best course of treatment, including exercise, food, or herbal formulas that the patient's body needs. Consideration of the patient's inherited and acquired attributes (developed through diet and life-style, environmental influences, natural energy cycles, and other miscellaneous influences such as sleep, work conditions, and relationships) should be based on observation and contemplation of the patient's physical, mental, emotional, and spiritual state of mind.

Designing a program to establish balance and harmony requires a keen awareness of energetic and mental states. In evaluating the patient's condition the doctor takes into account the location and nature of the disharmony, the severity of the condition and the overall physical, mental, and emotional condition of the patient (Figure 21.6). Furthermore, the doctor may notice a combination of both Yin and Yang symptoms or any of the other contrasting pairs be they Hot/Cold, Internal/External, Excess/Deficient. A combination of such symptoms creates a challenge when diagnosing a patient, since symptoms and indicators tend to be more extreme in a person who is ill (Figure 21.7).

YIN CONDITIONS

Primary Yang generates primary Yin. Everything in the universe needs support from Yin. On a basic level, Yin conditions in the body tend to be Cold, internal, Deficient, and degenerative in nature. If a condition becomes overly Yin, it may express some Hot or Yang symptoms. Extreme cold in the form of snow, for example, can cause frostbite. This is called a "false Yang" condition because the condition created is manifested as Yang instead of Yin.

Type of Sign	Yin	Yang
General	• Cool skin temperature • Pale, dark, dull, or yellow complexion • Cold hands and feet • Aversion to cold • Discomfort relieved by warmth and pressure • Senses dull • Curled lying posture	• Warm skin temperature • Red, oily, shiny complexion • Warm hands and feet • Aversion to heat and pressure • Five senses strong • Stretching posture in sleep
Diet	• Prefers warm food and drinks • Poor appetite • Poor sense of taste • Indigestion	• Prefers cold drinks and cold food • Strong appetite • Thirsty, dry mouth
Energy	• Tired • Weak	• Forceful, restless • Strong, stout
Behavior	• Underactive • Quiet, withdrawn	• Overactive • Noisy, outgoing
Breathing & Voice	• Weak, shallow breathing • Weak voice, quiet	• Deep, heavy breathing • Loud, coarse voice, talkative
Excretion	• Copious clear urine • Watery stool	• Scanty dark urine • Hard solid stool
Reproduction	• Sexually underactive • Scanty, pale menses • White leukorrhea	• Sexually overactive • Profuse, red menses • Yellow leukorrhea
Pulse	• Deep, slow • Empty, weak • Fine	• Superficial, rapid • Full, strong • Flooding
Tongue	• Pale, fat body • Moist, thin white coat	• Red, solid, sharp body • Dry, thick yellow coat

Figure 21.6 Signs of Yin and Yang Conditions

Yin conditions are typically characterized by the following: watery stool, clear and copious urine; depression; pale, dull, and yellow complexion; weak, tired, and thin constitution; cool skin, cold hands and feet; dulled senses; curled lying posture when sleeping; soft voice; silent, shallow breathing; preference for warm food; and a tendency towards indigestion. A Yin condition will also be reflected in a deep, weak, fine, slow, or sluggish pulse and a fat, pale, moist, and smooth tongue with possibly a white coating.

YANG CONDITIONS

Yang conditions tend to be Hot, External and Excessive. A Yang disease is generative. Under cer-

Attribute	Progression from Extreme Yang to Extreme Yin
Spirit	Aggressive, outgoing, joyful, sad, fearful, anxious, depressed (The degree of outward expression determines the yangness of the spirit, not so much the emotion itself.)
Body Type	Large and muscular, normal build, thin, emaciated
Posture	Stiff, erect, relaxed, hunched over, limp
Activity	Very animated, normally active, little movement, still
Respiration	Heavy and loud, loud sighing and stretching, breathing heavily through mouth, shallow light breathing, soft intermittent sighs
Voice	Loud, rough, high pitched, regulated and moderate, soft, murmuring, almost inaudible
Skin Cast	Red, yellow, green, pink, white
White of Eye	Red, yellow, blue-green, pink, white
Lip Color	Red, yellow, blue-green, pink, white
Moistness of Lips	Cracked, dry, moist, overly wet
Sputum	Yellow and thick, white and thick, thin, clear
Smell of Breath	Strong, mild, faint
Body Odor	Strong, mild, faint
Urine	Thick, yellow, smelly, thin, whitish, no odor
Feces	Dark, strong smelling, soft, loose, light color

Figure 21.7. The Continuum of Yin and Yang Constitutional Signs and Symptoms

tain conditions, an overabundance of Heat (Yang) may transform into Cold (Yin) to bring about a condition known as "false Yin." Extreme heat and sun, for example, can cause sunstroke. This is called a "false Yin" condition because the condition created manifests as Yin instead of Yang. If an acute Yang condition persists, it will become Yin.

Yang conditions are typically characterized by the following: hard, solid, dry stool; thick, yellow urine; outwardly excited appearance; strong, stout, muscular constitution; restless, active, irritable in character; warm hands and feet and overall warm skin and temperature; strong senses; tendency to stretch frequently; loud voice, talkative; heavy breathing; thirsty, dry mouth; and a preference for cold drinks. A Yang condition will also reflect a floating, flooding, fast, or strong pulse, and a red or yellow, solid, dry tongue with cracks and little or no coating.

SIGNS OF YIN AND YANG

Determining true from false Yang may seem confusing. Yin Deficient people who are very thin and emaciated are often quite hypertensive, this can be seen in their excessive loud talking, rest-

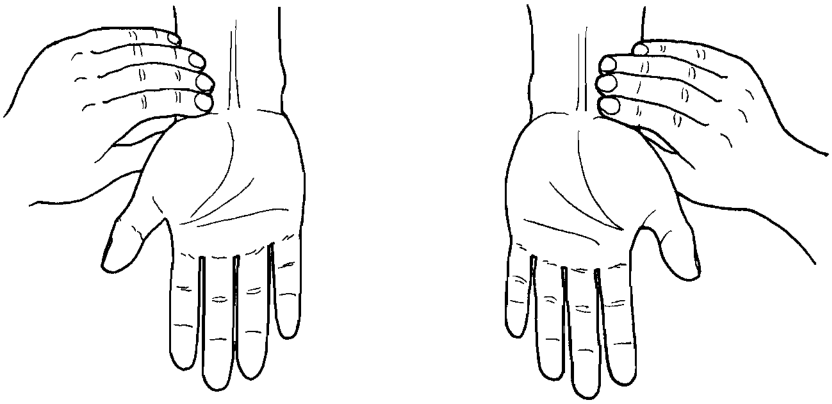

Figure 21.8. The Nine Regions of Pulse Diagnosis

lessness and insomnia. Such individuals can also develop erratically feverish symptoms which usually change from Hot to Cold.

Generally, a true Yin or Yang condition is determined by three or more symptoms that simultaneously manifest either one of those conditions.

False Yin symptoms occur when the Yang becomes overly strong, causing a weakening of the Yin. (Again, the overall condition should be assessed before making a differential diagnosis.) False Yin symptoms include weakness, coldness, wetness, and other Yin symptoms in an otherwise Yang condition. One should bear in mind that there are seldom purely Yin or Yang symptoms because Yin and Yang are so interdependent on each other.

Furthermore, stages of Yin and Yang may change during the course of the day, or over the course of an illness. If a Yang stage is allowed to progress, it can degenerate into weakness described as a Yin stage. Therefore, if a Yang condition changes to Yin it is probably not a good sign, but if a Yin condition changes to an acute Yang condition it may be a positive sign of movement towards overall balance. The focus is to understand that it takes active energy (Yang) to manifest an acute condition, and the stronger the manifestation, the stronger the energy.

YIN AND YANG CATEGORIZATION OF TRAINING

Chinese herbs and foods are typically classified in terms of Yin and Yang, as are Qigong exercises and meditations. With exercises and meditations the state of mind and level of relaxation during the exercise or meditation is critical in determining the Yin or Yang categorization. It is important to be aware of this when developing a patient's program to provide the proper energy adjustments to maintain a healthy balance. If a disharmonious condition develops and the wrong prescription is given, the patient can be thrown further off balance, making it more difficult to restore good health.

PULSE DIAGNOSIS OF THE EIGHT ENERGETIC PRINCIPLES

The body's pulse in considered another form of energy manifestation. The pulse is imagined as waves of various Fluids flowing within the body, originating from within the Heart, where the Shen resides. Pulse diagnosis can give the Qigong doctor detailed information on the state of the patient's internal organs, as it reflects the patient's flow of Qi and Blood, as well as Yin and Yang energy. The pulse is felt at the radial artery, which is divided into three areas (front, middle, and lower), and at three different levels (superficial, middle, and deep), creating "nine regions" of pulse diagnosis (Figure 21.8).

The three areas of the wrist relate to the energetic manifestations of the Triple Burners. The front position corresponds to Heaven and reflects the diseases from the head to the chest; the middle position corresponds to Man and reflects the diseases

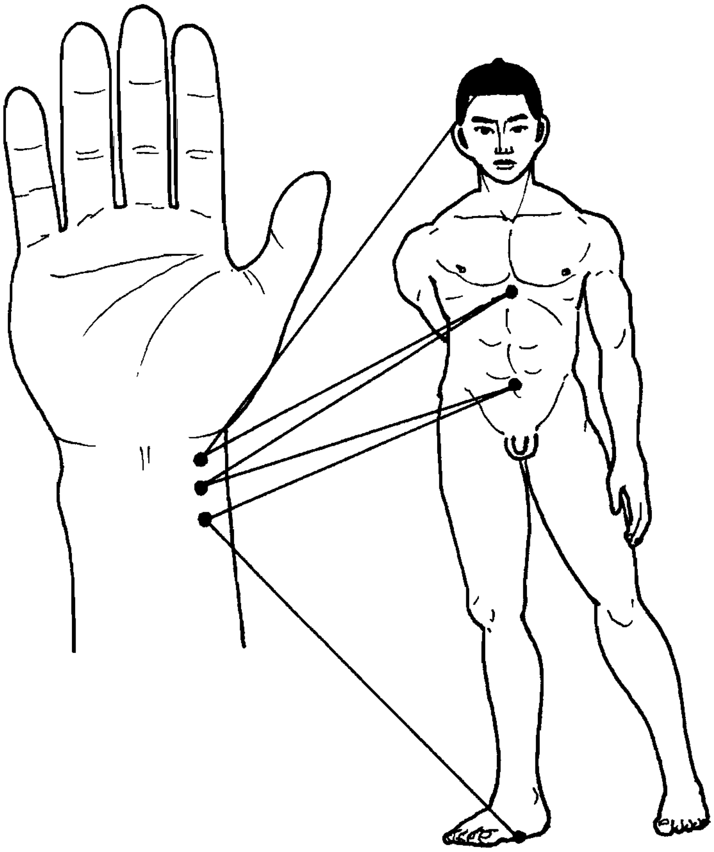

Figure 21.9. Pulse Correspondence with the Triple Burners

from the diaphragm to the umbilicus; the lower position corresponds to Earth and reflects the diseases from the umbilicus to the feet (Figure 21.9).

The three different levels relate to the energetic manifestations of the body's diseases. The superficial level corresponds to exterior diseases, the body's Qi and Yang organ energy; the middle level corresponds to the condition of the Blood, as well as the energy of the Stomach and Spleen; the lower level corresponds to interior diseases, the body's Yin energy and Yin organ energy.

The important thing is to diagnose how the body's energy is flowing, taking into consideration the relationship of the Yin and Yang on the pulse, i.e., whether it is superficial or deep, fast, or slow, Excess or Deficient. The quality of the pulse may also vary according to the patient's:

1. General Constitution and Body Build, with bigger frame patients generally having a stronger pulse then smaller frame individuals,
2. Level of Activity, with patients who engage in heavy physical work having a stronger pulse then those who engage in mental work, and
3. Exposure to Seasonal Weather, as a deeper pulse is generally felt in the winter and a more superficial pulse is felt in the summer.

PULSE QUALITIES AND CLASSIFICATIONS

In China, attention is placed primarily on the overall quality of the pulse. While reading the superficial pulses, it is important to note that they indicate the condition of the patient's Qi and Yang Organs. The deep pulses indicate the condition of the patient's Jing, and Yin Organs. Pulses can be divided into three main classifications, which are related to either Yin and Yang qualities: The Depth, The Speed, and The Strength (Figure 21.10).

1. The Depth at which the pulse can be felt is considered first.
    a. Superficial/Floating Pulses are felt with a light touch, barely touching the artery. They indicate an Exterior syndrome (e.g., Wind-Cold or Wind-Heat), chronic illness, or a general weakness. This informs the doctor that there is an Excess at the patient's Qi level (or a Deficiency at the Yin level).

- Superficial and Weak (Empty) pulses are felt in Yang Excess diseases (superficial diseases), when the patient's Wei Qi is fighting the pathogenic factors at the superficial level.
- Superficial and Forceful (Full) pulses are felt in interior diseases due to Yin or Jing Deficiency.

    b. Deep Pulses are felt on heavy pressure, near the bone. They indicate an interior syndrome (problems within the Yin Organs), and relate to the patient's Jing.

- Deep and Weak (Empty) pulses are felt in diseases due to a Deficiency of Qi and Yang.
- Deep and Forceful (Full) pulses are felt in deep diseases (when the disease has penetrated into the patient's internal organs). They confirm an Excess at the Yin level, or Deficiency at the Qi level, and point to a stasis of Qi or Blood, and to a Cold or Heat condition.

2. The Speed of the pulse is considered next.
    a. Slow Pulses are less then 4 beats per the patient's breath; they indicate a Cold syndrome.

- Slow and Weak (Empty) pulses indicate Empty Cold and Yang Deficiency.
- Slow and Forceful (Full) pulses indicate Full Cold and Yin Excess.

    b. Rapid Pulses are more then 5 beats per patient's breath and indicate a Heat syndrome.

- Rapid and Weak (Empty) pulses indicate Empty Heat and Yin Deficiency.
- Rapid and Forceful (Full) pulses indicate Full Heat and Yang Excess.

	YANG	YIN
DEPTH	Superficial — Exterior	Deep — Interior
SPEED	Rapid — Hot	Slow — Cold
STRENGTH	Forceful — Excess	Weak — Deficient

Figure 21.10. The Eight Energetic Principles are manifested through the body's pulses.

	LEFT HAND **Superficial - Deep**	**RIGHT HAND** **Superficial - Deep**
FRONT POSITION	Small Intestine - Heart	Large Intestine - Lungs
MIDDLE POSITION	Gall Bladder - Liver	Stomach - Spleen
LOWER POSITION	Bladder - Kidneys' Yin	TB & Pc - Kidneys' Yang

Figure 21.11. Left and Right Hand Correspondence

3. The Strength of the pulse is considered last.
    a. Weak (Empty) Pulses are weak and disappear on heavy pressure; they indicate a Deficient syndrome.
    b. Forceful (Full) Pulses are forceful and are felt on deep pressure; they indicate an Excess syndrome.

- Forceful (Full) and rapid pulses indicate Full Heat.
- Forceful (Full) and slow pulses indicate Full Cold.

THE CLASSICAL PULSE DIAGNOSTIC POSITIONS AND DEPTH

Traditionally the best time to take the patient's pulse is when the patient is calm and relaxed. There are several schools of pulse diagnosis, each one with its own specific methods of approach to understanding the patient's symptoms. The most important aspect in listening to the patient's pulse is to feel the strength, quality, and spirit of the pulse. Only then can the doctor receive accurate information about his patient's condition (Figure 21.11).

Pattern	Signs	Pulse	Tongue
Internal	Vomitting, high fever, gradual onset, chronic, long duration	Deep	Changes in texture and coating
External	Cold, flu, skin eruptions, sudden onset, acute, short duration	Superficial	Changes in coating
Cold	Diminished physiological function, decreased energy, lowered resistance, lowered immunity, white face, aversion to cold, cold extremities, lack of thirst, craving for warmth, clear urine, watery stool, desire for hot drinks, diarrhea, coolness of limbs, timid behavior, soft voice, slow movement, lack of motivation	Slow, tight	White coat, pale body, moist fur
Heat	Physiological hyperfunction, increased energy, increased metabolism, acute pains, fever, thirst, desire for cold liquids, flushed face, red eyes, restlessness, irritabiliity, dry stool, dark yellow urine, reduced urine secretions, aversion to warmth, hot red skin, constipation, warm extremities, loud voice, talkative, extroverted manner, aggressive behavior	Rapid	Red tongue with yellow fur
Deficiency	Great weakness, low syptomatic reactions, inconsistent energy, shallow and rapid respiration, less severe pain, pain characterised by soreness, weak or inconsistent voice, dull menstrual pain, withered face, low spirits, abdominal pain	Empty, weak	Little or no coat
Excess	Flushed face, loud voice, coarse and full breathing, sharp and acute pain, extreme menstrual pain, tender abdomen, condition worsens with apllication of heat	Full	Thick coat

Figure 21.12. Signs of Internal, External Cold, Hot, Deficient and Excess Conditions

Usually, in men, the pulse on the left (Yang) side of the body should be slightly stronger then that of the pulse on the right (Yin) side, and vice versa for women. Also in men, the front (Yang) positions should be slightly stronger, while in women the lower (Yin) positions should be stronger.

OVERVIEW

In the clinic we have a saying, "Where the Qi goes, the Blood flows." Although a patient's pulse is produced by the movement of Blood in the arteries, the activity of that movement is initiated by the force of the patient's Qi. By feeling the movement of Blood in the vessels, the activity of the patient's Qi, Blood, and Body Fluids are diagnosed. This internal communication is summarized in the above chart (Figure 21.12).

CHAPTER 22
DIAGNOSIS ACCORDING TO QI, BLOOD, AND BODY FLUID DYSFUNCTIONS

INTRODUCTION

This method of diagnosis identifies patterns on the basis of disharmonies according to Qi, Blood, and Body Fluids. It is used in clinical practice (especially for internal diseases) for diagnosing according to the pathological changes occurring in the internal organs.

The essential cause of disease is due to an imbalance of Qi, Blood, or Body Fluids. When one or more of these substances are unbalanced, the body's normal physiological function is disrupted and sickness results. This can be caused by the effect of the Six Exogenous Pathogenic Factors, Seven Emotional Pathogenic Factors, and is influenced by the patient's general constitution. Each of the patient's organs and channels have sets of symptoms or syndromes of these types of imbalances, which can then be identified for treatment. The Qigong doctor's ability to identify problems based on the pathological manifestations of Qi, Blood, and Body Fluids is founded on clinical observations of Excess, Deficient, Stagnant, and Rebellious patterns (Figure 22.1). These patterns can be further defined as predominantly Yin or Yang. These patterns also described as the various disorders of Qi, Blood, and Body Fluids, and sometimes overlap, for example:

- Excess conditions can give rise to Rebellious Qi and also create stagnation;
- Deficient conditions can give rise to Rebellious Qi and also create stagnation.

DISORDERS OF THE QI

Whenever energy movement is impeded, disorders in Qi functions occur. Generally, these dis-

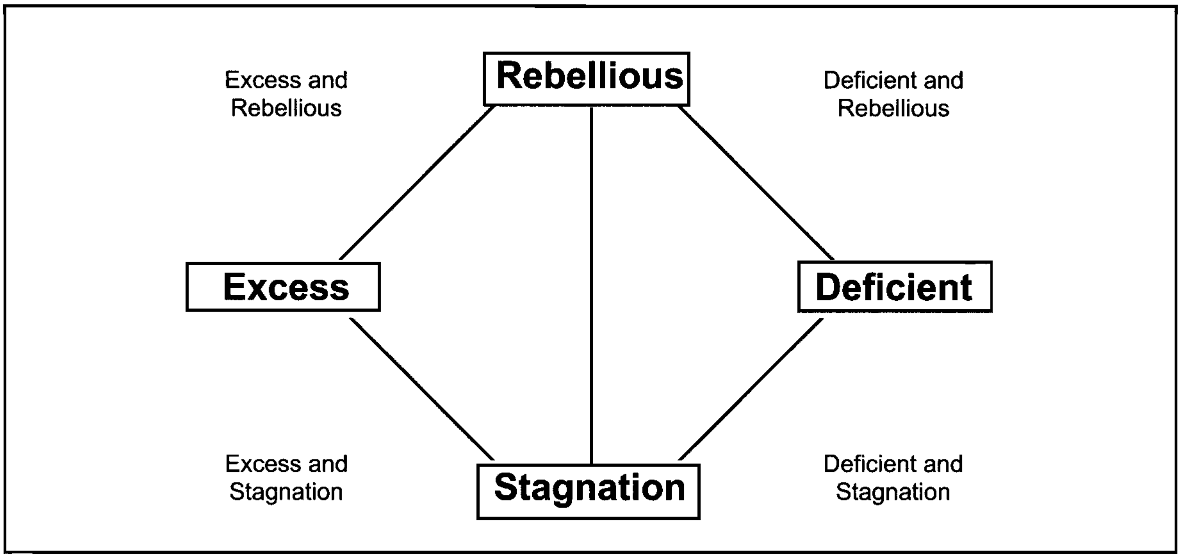

Figure 22.1. Pathological Manifestations of Qi, Blood, and Body Fluids

Fire		**Water**
Constraint of Qi	**First Stage**	Stasis of Qi
Arrogant/Agitated	**Second Stage**	Fluid Builds Up
Full on Fire is Created	**Third Stage**	Phlegm is Created
Heart Fire is Created	**Final Stage** ——> Mania <——	Phlegm Fire is Created

Figure 22.2. Disorders caused from the obstruction of Qi flow result in either Fire or Water energetic dysfunctions.

turbances cause either disorders of Fire, or disorders of Water (Figure 22.2).

The cause and progression to these disorders can be further broken down and classified according to their energetic dysfunction (Figure 22.3). There are several types of Qi disorders observed in the clinic and are generally described as follows.

LEAKAGE AND LOSS OF QI

During or after the training practice, the patient may feel that Qi is continuously and uncontrollably leaking out of the body from the genitalia and anus or anywhere else on the body. This loss of Qi can lead to Qi Deficiencies, and manifest through the following symptoms: emaciation, weakness of the limbs, grayish and dull complexion, nervousness, mental disturbances, distractibility, failing memory, spontaneous sweating, night sweats, seminal emission, insomnia, lassitude, and sluggishness.

DEFICIENT QI DISORDERS

Deficient Qi exists when the entire body, or particular organ, is not being sufficiently nourished by the Qi. This results in there being too little energy to maintain proper organ function. This condition is usually caused by a weakness of the Lungs, Spleen, or Kidneys, by chronic illness, or a weak constitution, poor diet, or fatigue. Deficient Qi is unable to:

- transform, resulting in symptoms such as diarrhea,
- transport, resulting in symptoms such as stagnation of Qi,
- hold, resulting in symptoms such as bruising,
- warm, resulting in symptoms such as cold extremities,
- protect, resulting in symptoms such as sickness, and
- raise, resulting in symptoms such as prolapse.

Deficient Qi is divided into three subcategories from which the patient can suffer. These subcategories are Empty Qi, Sinking Qi, and Qi Collapse. Each of these subdivisions can further be divided into Yin and Yang types.

1. Empty Qi is considered the first type of Deficiency. If not treated, Deficient Qi may progress to a condition known as Empty Qi. Diseases that are classified as Empty Qi are characterized by a serious weakness or Deficiency of the body's Qi. Empty Qi, especially Empty Yang Qi manifests as a weakness

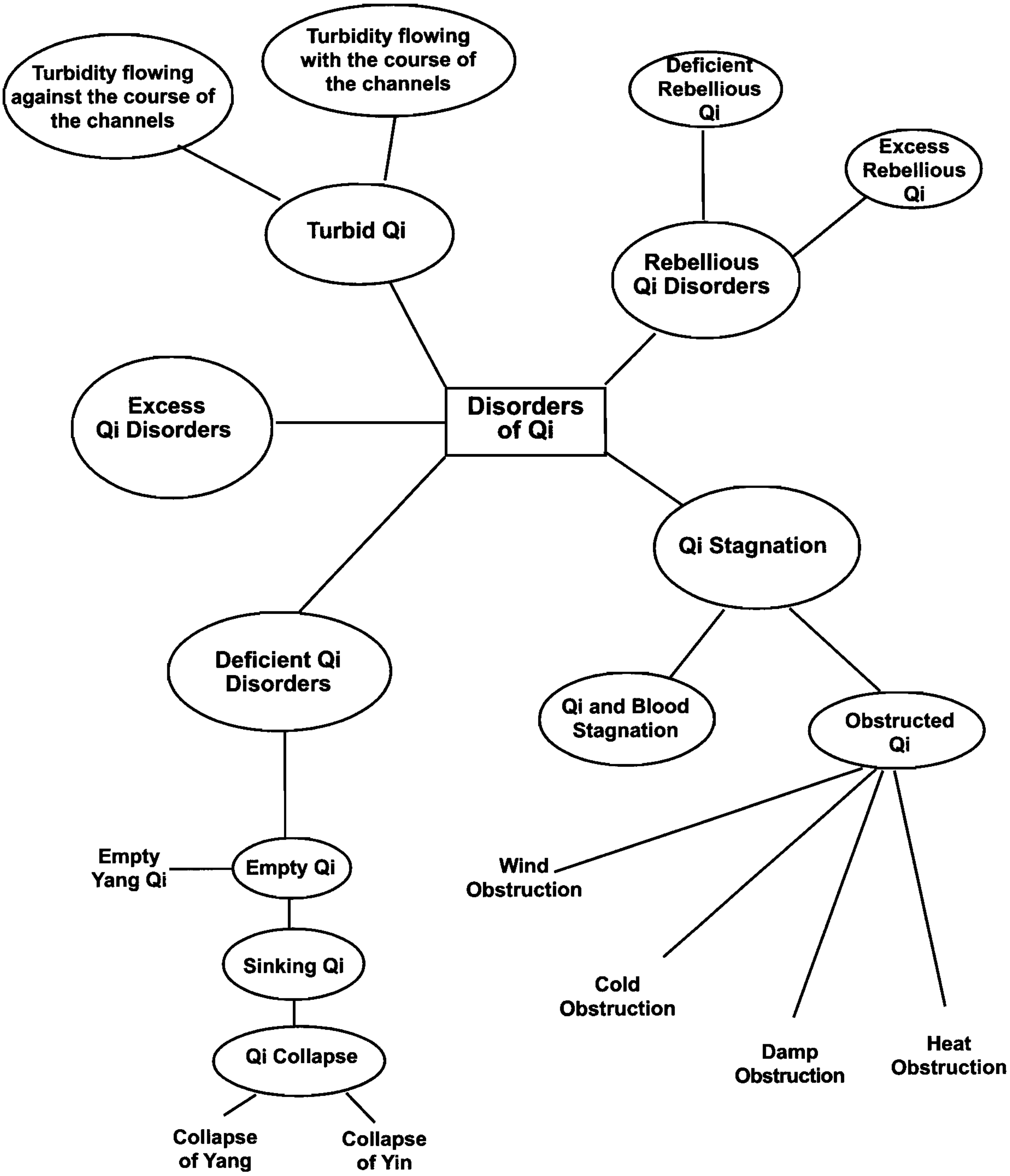
Turbidity flowing against the course of the channels
Turbidity flowing with the course of the channels
Deficient Rebellious Qi
Excess Rebellious Qi
Turbid Qi
Rebellious Qi Disorders
Excess Qi Disorders
Disorders of Qi
Qi Stagnation
Deficient Qi Disorders
Qi and Blood Stagnation
Obstructed Qi
Empty Yang Qi
Empty Qi
Wind Obstruction
Sinking Qi
Cold Obstruction
Qi Collapse
Damp Obstruction
Heat Obstruction
Collapse of Yang
Collapse of Yin

Figure 22.3. Whenever energy movement is impeded or in excess, disorders in Qi function occur, resulting in Qi Deviations.

in the function of holding, moving, and transforming Qi.

   a. Empty Qi originates primarily from the patient's constitution and manifests as malnutrition, fatigue, premature signs of old age, and an excess of extreme emotions.
   b. Empty Yang Qi belongs to the category of Yang. Because Yang Qi has the function of warming, Empty Yang manifests as an inability to adequately warm the energy, resulting in Cold symptoms. Empty Yang is generally caused by a weak constitution, poor diet (excessive raw, cold food and drink), overwork, old age, and excess sex.

2. Qi Sinking is also a subcategory of Deficient Qi, and is considered the second type of Deficiency. If not treated, Empty Qi may progress to a more serious Deficient condition known as Qi Sinking. This condition results in symptoms such as tiredness, listlessness, and mental depression, and may eventually lead to the prolapse of one or more of the organs.
3. Qi Collapse is a subcategory of Deficient Qi and is considered the third and most severe type of Deficiency. A Qi collapse exists when the Qi is so Deficient that it can no longer hold the internal organs in place. This Qi depletion not only causes a weakness of the body's Righteous Qi, which leads to the prolapse of the organs (Stomach, uterus, intestines, rectal area, vagina, or Bladder), but also induces prolonged diarrhea, upper abdominal distention, and dizziness.
   a. The Collapse of Yin results in pure Fire. The main manifestations are abundant perspiration and external tissue, which is hot to the touch.
   b. The Collapse of Yang results in pure Cold. The main manifestations are chills and external tissues which are cold to the touch.

EXCESS QI DISORDERS

An Excess Qi condition is caused from too much Qi, which overacts on an organ, affecting its function, and usually produces too much transformation and Heat in a given organ. This can be due to over-energizing the organs and channels, from excessive Qigong exercises and meditations, or too much energy stimulation during a Qigong doctor's treatment.

QI STAGNATION DISORDERS

Qi stagnation exists when the Qi is not flowing smoothly, and becomes sluggish when moving within the body's channels, organs, or tissues. Qi stagnation can be caused by many different factors such as: suppressed emotions, trauma, contraction of microorganisms, poor diet, and weak digestion. Qi stagnation, over time, results in localized pain, distention, cysts, tumors, etc. Qi stagnation disorders have two subcategories: Qi and Blood Stasis, and Obstructed Qi.

1. Qi and Blood Stasis refers to the sensation of local pain, heaviness, soreness, distention, and compression, which are caused by the stagnant Qi and Blood in certain areas of the body. These sensations do not disappear by themselves. This stasis may occur during or after Medical Qigong training practice from Qi disorders and visceral malfunctions.
2. Obstructed Qi is also considered a subcategory of Qi stagnation, and exists when the Qi becomes so stagnant that it no longer flows. At this stage the Qi can not move the Blood. This can be caused by trauma, surgery, and acute Qi stagnation.

When Qi becomes obstructed inside the body, certain clinical manifestations result, such as:

- Cold Obstruction causes Qi to become fixed with local numbness that improves with heat;
- Damp Obstruction causes Qi to become fixed with local swelling;
- Heat Obstruction causes Qi to become inflamed, red, and swollen;
- Damp Heat Obstruction is a combination of Damp and Heat. It is called "summer Damp Heat," and results in poor appetite, dizziness, diarrhea, heaviness of the head, and sensations of chest suffocation;
- Wind Obstruction causes migrating pain that moves from joint to joint.

Organ	Normal Flow	Rebellious Flow	Symptoms
Liver	Upwards	Excessively Upwards Horizontally to Stomach Horizontally to Spleen Horizontally to Intestines Downwards	Dizziness, Headache, Irritation Nausea, Belching, Vomiting Diarrhea Dry Stool Burning Urination
Heart	Downwards	Upwards	Insomnia, Mental Restlessness
Spleen	Upwards	Downwards	Diarrhea, Prolapse of Organs
Stomach	Downwards	Upwards	Hiccup, Nausea, Belching, Vomiting
Lungs	Downwards	Upwards	Coughing, Asthma
Kidneys	Downwards	Upwards	Asthma

Figure 22.4. Chart of Rebellious Qi Flow

REBELLIOUS QI DISORDERS

Rebellious Qi disorders manifest as Qi flowing in the wrong direction (a direction different from its normal physiological progression); this causes a disturbance within the harmonious flow of energy and affects the organs' functions. The symptoms have to be analyzed according to the internal organ which is involved. There are two types of Rebellious Qi, the Excess and Deficient type. Rebellious Qi is categorized as follows:

DISORDERS OF EXCESS REBELLIOUS QI

Disorders of Excess Rebellious Qi occur when the vital Qi increases and forces the pathogenic Qi to move; leading to local sensations of pain, soreness, distention, heaviness, cold or heat, which can develop during a patient's Qigong practice (Figure 22.4).

DISORDERS OF DEFICIENT REBELLIOUS QI

Disorders of Deficient Rebellious Qi refers to the disorderly flow of Qi and Blood during or after the training practice, which causes dizziness, fright, a heavy feeling in the chest, shortness of breath, shaking limbs, trembling of the hands and feet, or fainting. Generally, energy sensitive patients know the location and direction of the disorderly flow of Qi in their tissues.

TURBID QI

While transforming and transporting energy, the body divides ingested and absorbed Qi into Clear and Turbid energy for distribution (keeping and redistributing the clean, while disposing of the turbid). Turbid Qi refers to any dark, murky energy which can originate within, or from outside of the body's organ system, and sometimes travels throughout the body's channel system. Turbid Qi attaches to the body's wounds (physical, emotional, or spiritual), manifesting as dark, discolored clouds of energy. This energy can adversely affect the body's physical, emotional, and energetic balance, especially when combined with any pathogenic internal or external factor.

PATHOGENIC SENSING OF TURBID QI

Feeling the sensations of Turbid Qi is referred to as "pathogenic sensing." This skill is acquired by the doctor while projecting energy to his patients. While emitting energy, the Qigong doctor can sense the patient's Yuan Qi, as well as the Turbid Qi inside the patient's body. This enables the doctor to diagnose and treat their patients' diseases according to the various Qi sensations, and to categorize the patients' Turbid Qi. The Turbid Qi is also called Evil or "pathogenic Qi." The flow of Turbid Qi within the patient's channels is characterized as follows:

- Turbidity flowing with the course of the channel indicates transference of Turbid Qi towards the body's interior.
- Turbidity flowing against the course of the channel indicates the beginning production of Turbid Qi.

Turbid Qi differs from the pathogenic factors of infectious diseases referred to in Western medicine. According to clinical experience, the pathogenic sense of Turbid Qi can be classified as: Dry or Hot Sensations, Cold or Chilly Sensations, Soreness or Numbing Sensations, Impure or Foul Sensations, The Sensations of The Seven Emotions, and The Sensations of The Six Climate Conditions.

1. Dry or Hot Sensations are considered Yang signs or symptoms of Heat. This type of Turbid Qi feels feverish and flaccid; it can cause perspiration, irritation, etc. The sensations of dry-heat on the body or hands of the doctor often cause the doctor to fidget, as if being scorched by a fire. It may also serve as an indicator of syndromes of Excess Heat and hyperactivity of Fire due to a Yin Deficiency.
2. Cold or Chilly Sensations are considered Yin signs or symptoms of Cold. This type of Turbid Qi feels cold, rigid, and stagnant. It may be so cold that the doctor's fingertips turn cold immediately. This coldness transmits from the doctor's fingertips upward, causing shivering and contractions of the sweat glands ("goose-bumps"). This gives the doctor a particular feeling of cold and discomfort. Cold feelings often serve as a message of Deficiency and Cold syndromes such as rheumatism, rheumatoid disease, Deficiency of the Kidney Yang, and certain types of cancer.
3. Soreness or Numbing Sensation often indicate syndromes of hyperactivity of the Liver Yang and Excess of Wind, Cold, Dampness or Phlegm, as well as arthralgia syndrome and cancer. The doctor may experience an uncomfortable feeling of local soreness or pain, heaviness, numbness, tingling and itching.
4. Impure or Foul Sensations can be felt when the doctor is standing opposite the patient, or when extending energy towards the patient. It is an indescribable unpleasant feeling that is often felt when the open palm and fingers are used to detect a disease.
5. The Seven Emotions (joy, anger, worry, anxiety, fear, shock, and grief) in excess will cause discomfort to the doctor. During treatment, the doctor may detect and feel the different natures of these emotional pathogenic Evils flowing through the doctor's body. The doctor may, for example, have a specific feeling of anger if the disorder of the patient is due to excessive anger. Once these pathogenic emotions are discovered, the doctor must purge or disperse them, as well as determine their origin, and cause.
6. The Six Climate Conditions (Wind, Cold, Summer Heat, Dampness, Dryness, and Fire) will also cause discomfort to the doctor. During treatment, the doctor may detect and feel the different natures of these pathogenic Evils. The doctor may, for example, have a specific feeling of slippery wet ooze if the disorder of the patient is due to Dampness.

RECTIFYING QI DISORDERS

The primary treatment principles for rectifying Qi Disorders are as follows:

- for Qi leakage - tonify
- for Deficient Qi - tonify
- for Empty Qi - tonify
- for Qi sinking - raise and tonify
- for Collapsed Qi - tonify
- for Excess Qi - reduce
- for Stagnant Qi - move
- for Obstructed Qi - move
- for Rebellious Qi - calm and subdue
- for Turbid Qi - remove, dredge, and purge

These are often found in combination with other patterns. Here are some examples of treating combined patterns:

- for Excess and Rebellious Qi, reduce Excess and calm Rebellious Qi;
- for Deficient and Rebellious Qi, tonify Deficient and calm Rebellious Qi;

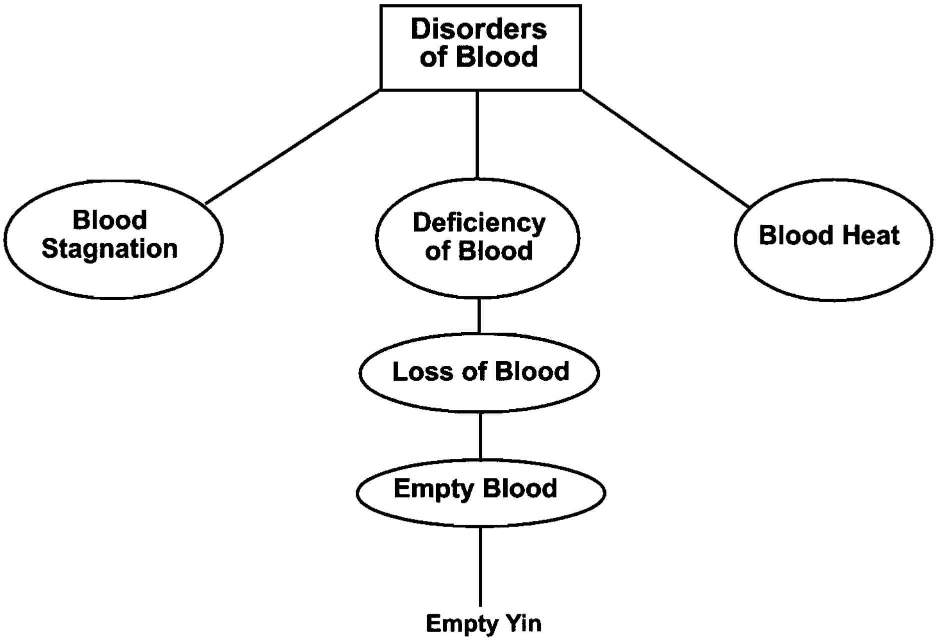

Figure 22.5. Whenever Blood movement is impeded or in excess, disorders in Blood function occur.

- for Excess and Stagnant Qi, reduce Excess and move the Stagnant Qi;
- for Deficient and Stagnant Qi, tonify Deficient and move the Stagnant Qi;
- for Excess and Deficient Qi, drain Excess and tonify Deficient Qi; and
- for Rebellious and Stagnant Qi, calm Rebellious and move the Stagnant Qi.

DISORDERS OF THE BLOOD

To understand Blood disorders it is important to understand the concept of Blood from the perspective of Traditional Chinese Medicine. In Chinese Medicine, Blood is different in concept, characteristics and function from that of Western medicine. Blood originates from the transformation of food and drink by the Spleen, which then transfers the refined food energy (Gu Qi) to be further enhanced by the Heart, Lungs, and air energy. Blood is therefore considered a transformation of constructive energies.

Disorders of the Blood can relate to disorders of the Qi. Qi and Blood flow together, Qi being the active force which makes the Blood circulate and keeps it within the blood vessels. Both Qi and Blood flow within the channels and blood vessels, continuously circulating throughout the body nourishing, maintaining, and moistening the tissues. Qi is an energetic form and is considered a Yang substance, while Blood is a liquid form of energy and considered a Yin substance. The cause and progression to Blood disorders can be further broken down and classified according to their energetic dysfunction (Figure 22.5). There are several types of Blood disorders observed in the clinic and are generally described as follows.

DEFICIENCY OF BLOOD

A Deficiency of Blood exists when the entire body, or a particular organ, is insufficiently nourished by the Blood, or not enough Blood is present to nourish the organs and channels, to support the Qi. Blood Deficiency is usually caused from Blood loss, chronic illness, poor diet, or Spleen Qi Deficiency, which weakens the Qi. This in turn affects the Liver (causing symptoms of dizziness, blurred vision, and numbness) and Heart (causing symptoms effecting the mind, memory, and Shen).

LOSS OF BLOOD

Loss of Blood is considered a subcategory of Deficient Blood. Loss of Blood is caused by Spleen Qi failing to control the Blood, resulting in the Blood spilling out of the vessels and organs, etc. Loss of Blood is divided into two main types:

- Deficient loss of Blood is caused by an insufficiency of Qi to hold the Blood.
- Excess loss of Blood, is caused by either Blood-Heat pushing the Blood out of the vessels, or by Blood stagnation and Yin Deficiency which causes the Blood to pool.

EMPTY BLOOD

Although Empty Blood is a subcategory of Deficient Blood, it is considered a serious Deficient condition, mostly due to the Blood of the whole body becoming Deficient. Empty Blood corresponds to a Deficiency of Qi, caused by:

- Inadequate nutrition,
- Loss of Blood through accident or disease,
- Loss of Body Fluids through excessive sweating, urination, vomiting, or diarrhea,
- Extreme emotions (e.g., rage, terror and shock),
- Blood stagnation, which leads to poor Blood circulation.

Since the Empty Blood fails to nourish the tendons and muscles, it results in shaking in the extremities and head, dizziness, and blurred vision.

Empty Yin

Blood belongs to the category of Yin. Because Yin Qi has the function of cooling and moistening, when an Empty Yin condition is present, the symptoms manifest as Excess Heat and Dryness. Empty Yin is caused by a weak constitution and a poor diet (overcooked foods and hot liquids). The organs most likely to be effected by the Empty Yin condition are the Kidneys, Lungs, Heart, Liver, and Stomach.

BLOOD STAGNATION

Blood Stasis exists when the Blood is unable to flow smoothly, and becomes obstructed or congeals. Blood stasis is usually caused by an impact trauma, Deficient Blood, stagnant Blood, Blood Heat, Blood Cold, or stagnant Qi. It results in localized pain which worsens with pressure, swelling of the internal organs, cysts, and tumors. The organ most often affected by Blood stasis is the Liver.

BLOOD HEAT

Blood Heat is mostly due to Liver Heat and External Wind. When toxic Heat enters the Blood, the Blood may break out of the vessels and cause: nose bleeding, hemorrhaging, and broken blood vessels. Heat consumes the Blood and Yin, injuring the tendons and muscles, causing spasms and rigidity. The Blood Heat also injures the Governing Vessel, Pericardium, and the Shen.

The Four Main Treatment Principles

- for a Deficiency of Blood, tonify
- for Empty Blood, tonify
- for Blood Stasis, move Stasis
- for Blood Heat, cool down

THREE DISORDERS OF BODY FLUIDS

To understand disorders of the Body Fluids it is important to understand the concept of Body Fluids, from the perspective of Traditional Chinese Medicine. Body Fluids are a denser form of Qi, and depend on Qi for transformation and transportation. Qi occupies and directs the movement of Blood and Body Fluids. The function of the Fluids is to moisten and nourish the body's tissues, as well as thin the Blood to prevent stasis. Body Fluids are formed from ingested food and drink, and are then transported by the energy of the Spleen to the Lungs and Small Intestines for distribution.

Disorders of the Body Fluids can relate to disorders of the Blood. There is a constant interchange between Fluids and Blood, such as, for example, a Deficiency of Fluids can be developed from a heavy loss of Blood. Body Fluids are divided into two categories:

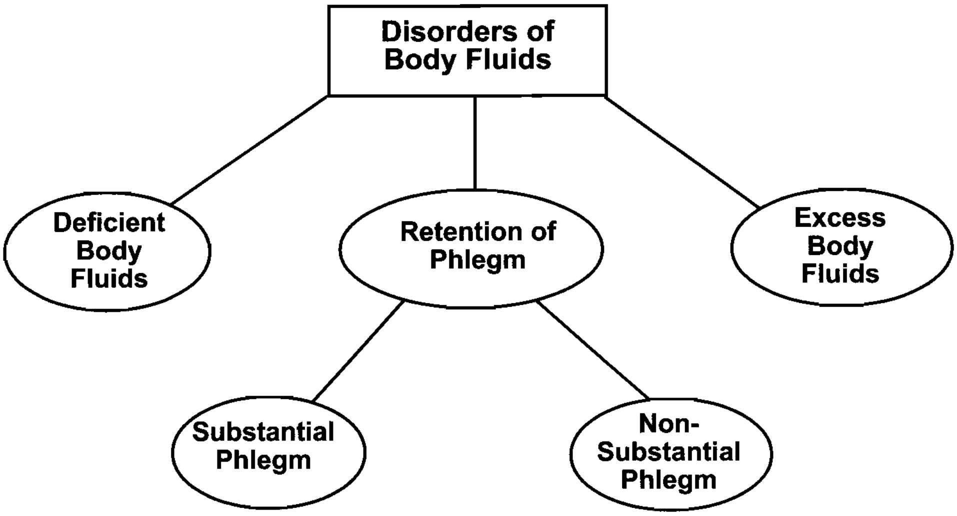

Figure 22.6. Whenever Body Fluid movement is impeded or in excess, disorders in Body Fluid function occur.

1. Jin refers to the light, thin, watery, clear, quick-moving Fluids and is considered Yang. These Fluids are under the control of the Lungs which circulate them with the Wei Qi. These Fluids also permeate and moisten the skin and have the function of warming and nourishing the muscles. The Upper Burner is responsible for controlling the Jin transformation and moving it towards the skin. The Jin is manifested in the body in the form of sweat, tears, saliva, mucus, and parotid serum.
2. Ye refers to the heavier, thicker, turbid, dense, slow moving Fluids and is considered Yin. These Fluids are under the control of the Middle and Lower Burners, which circulate the Ye with the Ying Qi. The Spleen and Kidneys are responsible for controlling the Ye transformation. This Fluid lubricates the joint cavities (synovial fluid), nourishes and lubricates the brain and spinal cord (cerebrospinal fluid), bone marrow and the orifices of the sensory organs (eyes, ears, nose, and mouth).

There are three primary disorders of Body Fluids observed in the clinic, they are described as follows (Figure 22.6).

1. Deficient Body Fluids exists when the entire body, or a particular organ, is insufficiently nourished by Body Fluids. This is generally due to insufficient Yin to create or nourish the body's organs or joints. Deficient Body Fluids are usually caused by prolonged loss of Fluids (sweating, diarrhea, and vomiting), prolonged illness, a Heat condition, or a Yin Deficiency which leads to Dryness. A Deficiency of Body Fluids usually affects the Lungs, Kidneys, Stomach, and Large Intestines.
2. Excess Body Fluid accumulation is usually caused by an Excess Yin and one of three factors, or a combination of all of them. They are:
   - A Lung Qi Deficiency, which affects the top of the body, face and hands,
   - A Spleen Qi Deficiency, which affects the middle part of the body, and abdomen,
   - A Kidney Qi Deficiency, which affects the lower part of the body, legs and ankles.

   If there is a Deficiency within one or all three of these organs, and the Body Fluid is not able to transform properly, it will overflow the channels and settle under the skin.
3. Retention of Phlegm is due primarily to a Deficiency of the Spleen failing to transport

and transform the Body Fluids. If Phlegm is retained for long periods of time, disease of the organs (Substantial Phlegm), as well as channels, skin, and joints (Non-Substantial Phlegm), can form.

The Three Main Treatment Principles

- for Deficient Body Fluids, moisten;
- for Excess Body Fluids, dry or drain;
- for Retention of Phlegm, resolve.

TYPES OF PHLEGM

The concept of Phlegm is an important pathological and etiological factor in Traditional Chinese Medicine. If Phlegm is allowed to accumulate and becomes retained within the body, it causes disease.

The main cause of Phlegm formation is Spleen Deficiency. If the Spleen becomes weak and fails to transform and transport the Body Fluids, these will accumulate and transform into Phlegm. However, Phlegm formation can also occur if the Lungs fail to disperse and lower Fluids, or if the Kidneys fail to transform and excrete Fluids.

In Traditional Chinese Medicine, there are two types of Phlegm: Substantial Phlegm and Non-Substantial Phlegm.

1. Substantial Phlegm is described in ancient texts as "having a form." It can be observed as the thick mucus that collects within the Lungs and spat out during bronchial infections and other Lung diseases. As the Lung disease progresses, the patient's mucus changes color, progressing from clear watery to yellow and thick, then to green and pussy, and finally brown and red.
2. Non-Substantial Phlegm is described in ancient texts as "having no form." This type of Phlegm can accumulate within:
   - the body's organs forming stones,
   - the joints forming arthritic bone deformations,
   - the skin forming lumps or cysts, and
   - the channels causing stagnation of Qi and Blood, resulting in numbness.

Both Substantial and Non-Substantial Phlegm can assume different forms of disease, developing in accordance to its associations with Heat, Cold, Damp, Wind, Qi, and Fluids. For example:

1. Phlegm Heat is a type of Substantial Phlegm that affects the Lungs, Stomach, or Heart. It manifests as yellow sticky phlegm, a red face and a red tongue with sticky yellow coating.
2. Cold Phlegm is a type of Substantial Phlegm that affects the Lungs and Stomach. It manifests as white watery phlegm expectoration, cold feelings in the back and extremities, and a pale tongue with a white and wet coating.
3. Damp Phlegm is a type of Substantial Phlegm that affects the Lungs. It manifests as a profuse white sticky phlegm, feelings of stuffiness in the chest and epigastrium region, and a sticky tongue coating.
4. Wind Phlegm is a type of Non-Substantial Phlegm and is seen in Wind Stroke. It causes dizziness, nausea, vomiting, coughing up of Phlegm, numbness of the limbs, rattling throat, and aphasia.
5. Qi Phlegm is a type of Non-Substantial Phlegm which affects the throat. It manifests as a "plum pit" syndrome within the throat. Its formation is related to emotional problems, created when the Liver Qi rises and stagnates within the throat. Its condition manifests as a feeling of swelling in the throat, difficulty in swallowing, and stiffness in the chest and diaphragm.
6. Phlegm Fluids is a type of Substantial Phlegm which has a fluid or watery nature, and is called Yin in Chinese medicine. There are four types according to their location:
   - Phlegm Fluids located in the Stomach and intestines manifest as abdominal fullness, distension, and vomiting;
   - Phlegm Fluids located above the diaphragm manifest as coughing, asthma, edema, and dizziness;
   - Phlegm Fluids located within the hypochondrium manifest as distension in the hypochondrium area, with pain and coughing;
   - Phlegm Fluids located in the four limbs manifest as heaviness and pain in the muscles.

CHAPTER 23
DIAGNOSIS ACCORDING TO THE FIVE-ELEMENT THEORIES

INTRODUCTION TO THE FIVE ELEMENTS

This method of diagnosis is based on the interpretation of clinical manifestations according to the creative, controlling, invading, and insulting sequences of the Five Elements. The Five Elements are five phases of transition used in Chinese Medicine and are sometimes referred to as the Wu Xing. Wu translates to mean the number five, Xing translates to mean a process of movement or walking. The concept of the Five Elemental theory began to first appear in documents in China during the Zhou Dynasty (from 1000 to 770 B.C.). This theory classified tangible and intangible substances into five categories for observation and study, as well as diagnosis and treatments.

1. The Wood Element is associated with the active functions of birth in nature's growth stage of development.
2. The Fire Element is associated with maximum functional activity in nature's growth stage of development.
3. The Earth Element is associated with the function of stabilization in nature's growth stage of development.
4. The Metal Element is associated with nature's declining functions of the growth stage of development.
5. The Water Element is associated with the maximum state of rest in nature's growth stage of development.

The Five Elemental theory is used to explain the classifications and characteristics along with the laws of universal cycles (creating, controlling, invading and insulting). These interactions are studied today in Traditional Chinese Medicine to assist the doctor's diagnosis in comprehending the growth and development of the body's energetic anatomy, physiology, disease processes, and symptom development (Figure 23.1).

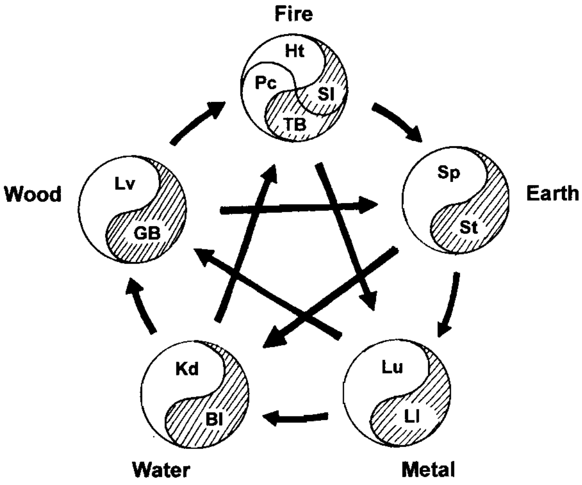

Figure 23.1 The Five Elements

THE FIVE ELEMENTS AND YIN AND YANG THEORY

It is important to understand the Five Elemental theory's energetic origin and its interaction with Yin and Yang Qi to appreciate its energetic categorizations.

The ancient Chinese Qigong masters believed that the creation, development, and declining transitions of all things throughout the universe were the result of the interaction of Yin and Yang Qi. The Yin and Yang theory was therefore used to describe the opposing, interdependent, as well as waxing and waning transitions of all things, especially the interactions of the Five Elements (Wood, Fire, Earth, Metal, and Water). Fire and Wood are considered active Yang elements, while Water and Metal are considered quiescent Yin elements. The Earth is considered a balance point between Yang and Yin.

Qigong doctors assign the Five Elements of Traditional Chinese Medicine to the body's organs and tissues. The Five Elements are considered the Jing of the body's Yin and Yang energy pools. Imagine each element organ as a pool of liquid energy. The energetic movement, much like water, flows from element to element. This flowing movement of liquid energy is considered to be the body's energetic pulse, moving Qi to and from these pools.

The rivers of Qi cannot exist without the pools of Jing to nourish their flow and function; and the pools of Jing cannot act without the rivers of Qi stimulating their action. Therefore the principle action of the Five Elements is the circulation of the Yin and Yang Qi.

The human body is like a living Taiji symbol (see Chapter 21). Its back channels are considered Yang; its front channels are considered Yin; and its center channel is contained through the energetic actions of the Taiji Pole. The Taiji Pole produces Yang Qi by the inhalation of life-force energy through the breath. When Yang Qi reaches its peak, it rests, while the Yin Qi begins to increase. When the Yin Qi reaches its peak and comes to rest, then the Yang Qi begins to increase once more. This continuous interaction of Yang energy with Yin energy produces and enhances within the body, the Five Elemental organ energies of the Liver (Wood), Heart (Fire), Spleen (Earth), Lungs (Metal), and Kidneys (Water).

The first stage of Qigong exercises and meditations create an energetic transformation which regulate, refine, and reduce these Five Elemental organ energies into two primary Yin and Yang energies known as Fire and Water Qi. The second stage of transformation consists of combining the Water and Fire energy into one union of Three Human Treasures (Jing, Qi, and Shen). This energetic transformation is responsible for "bringing 10,000 things into being" (Figure 23.2). In China, these meditation stages are used for exploring the stimulation of the energetic functions of the central nervous system, the peripheral nervous system, cardiovascular system, digestive system, reproductive system, etc.

The Five Elemental theory assists the Qigong doctor in examining the energetic correspondences for clinical evaluation. This is only one phase of clinical observation however, and should always be applied with the Yin and Yang theory.

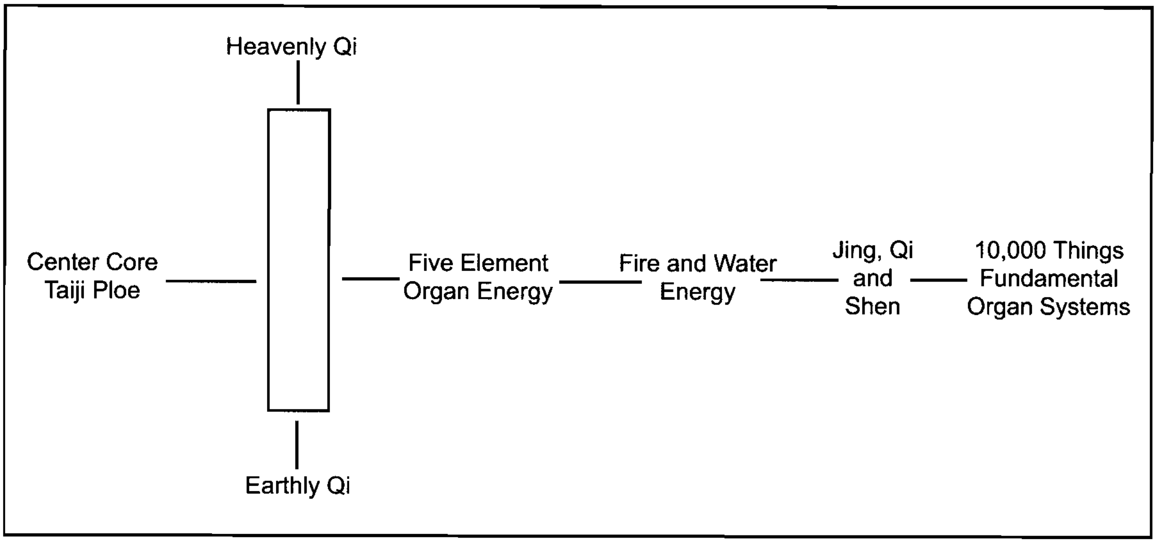

Figure 23.2 Stages of Energetic Transformation

THE CREATIVE OR GENERATIVE CYCLE

Within every Qi cycle each element has two main aspects: it is either (1) creating or being created, or (2) restricting and/or being restricted. In the process of growth and development, both creative and restrictive patterns are needed. Without creation, growth would have no origin, and growth without restriction can create harm. The creative element is called the Mother, and the element created is called the Child. This is the "Mother and Child" relationship referred to in Traditional Chinese Medicine. The Creative Cycle is generally used for tonification, as the stimulation of one organ enhances the function of the next (Figure 23.3). The Five Elemental Creative Cycle and its organ correspondences are as follows.

1. The Liver's element is Wood, which is used to create the Heart's element Fire. The Liver is the Mother of the Heart, which is the Child. The Liver stores the Blood and Blood houses the Shen. If the Mother (Liver) becomes Stagnant, the Child's Heart (Shen) will suffer. Clinical manifestations include indecision, timidity, lack of courage, palpitations, and early morning insomnia.
2. The Heart's element is Fire, which is used to create the Spleen's element Earth. The Heart is the Mother of the Spleen. The Heart's Qi pushes the Blood, thereby helping the Spleen in its function of transporting. If the Mother (Heart) becomes stagnant, the Child (Spleen) will suffer, creating a Spleen Yang Deficiency. This is due to failure of Heart Fire in providing Heat to the Spleen. Clinical manifestations include chills, weakness in the limbs, and loose stools.
3. The Spleen's element is Earth, which is used to create the Lungs' element Metal. The Spleen is the Mother of the Lungs. The Spleen Qi provides Gu Qi (food energy from the Earth) to the Lungs where it combines with air (energy from Heaven) to form Gathering (Zong) Qi. If the Mother (Spleen) becomes Deficient, the Child (Lungs) will suffer, creating Phlegm which obstructs the Lungs. Clinical manifestations include Phlegm in the chest, coughing, and general weakness.

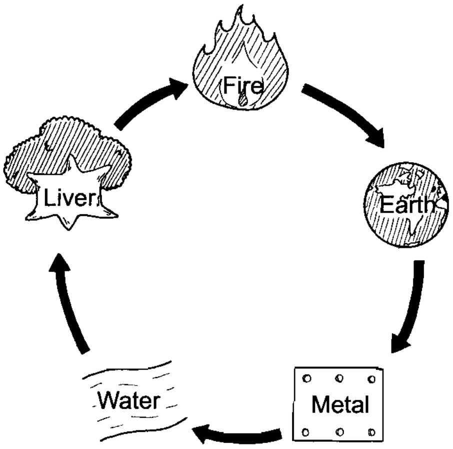

Figure 23.3 The Five Elemental Creative Cycle

4. The Lungs' element is Metal, which is used to create the Kidneys' element Water. The Lungs are the Mother of the Kidneys. The Lungs send Qi down to meet the Kidney Qi, as well as Fluids to nourish the Kidneys. If the Mother (Lungs) becomes stagnant, the Child (Kidneys) will suffer, due to the Kidneys not receiving enough Qi. Clinical manifestations include breathlessness, loss of voice, coughing, and asthma.
5. The Kidneys' element is Water, which is used to create the Liver's element Wood. The Kidneys are the Mother of the Liver. The Kidneys' Yin nourishes the Liver Blood. If the Mother (Kidneys) become Deficient, the Child (Liver) will suffer (i.e., Liver Yin Deficiency). Clinical manifestations include headaches, blurred vision, dizziness, and vertigo.

THE CONTROLLING OR RESTRICTIVE CYCLE

This is a condition in Traditional Chinese Medicine where one element brings under control, or restrains, a corresponding element, thus ensuring that a dynamic balance is maintained among the Five Elements. The element that restricts is considered successful, while the element being controlled is in check. It is commonly referred to as "the strong controlling the weak," and is generally used for sedation, to ensure that the Creative Cycle does not over-create and cause imbalances. The element prior to the Mother is commonly referred to as the Grandmother. In the Controlling Cycle, it is the Grandmother who controls the Child (Figure 23.4). The Five Elemental Controlling Cycle and its organ correspondences are explained as follows.

1. The Wood absorbs nutrients from the Earth, i.e., the Liver (Grandmother) controls the Spleen and its paired organ the Stomach (Child). Two of the Liver's energetic functions are to help the Stomach "rot and ripen" the food and the Spleen to transform and transport the Gu Qi.
2. The Earth restricts Water, i.e., the Spleen (Grandmother) controls the Kidneys (Child). Both the Spleen and Kidneys control the transformation of Body Fluids.
3. The Water extinguishes Fire, i.e., the Kidneys (Grandmother) control the Heart (Child). The energetic aspect of the Kidneys' Water controls the energetic balance of the Heart's Fire.
4. The Fire melts Metal, i.e., the Heart (Grandmother) controls the Lungs (Child). The Heart governs the Blood, while the Lungs govern Qi (both Qi and Blood mutually assist and nourish each other).
5. The Metal chops Wood, i.e., the Lungs (Grandmother) control the Liver (Child). If the Lungs' Qi is weak it will not descend thus allowing the Liver Qi to rise.

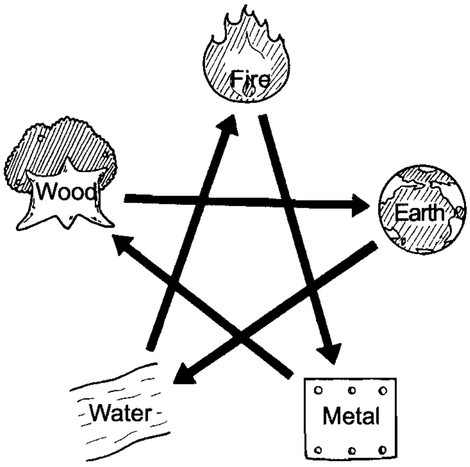

Figure 23.4. The Five Elemental Controlling Cycle

THE INVADING OR OVERACTING CYCLE

This is a condition in Traditional Chinese Medicine where an element overacts, overcontrols and invades, the element it is supposed to control (Figure 23.5). Instead of the role of checking and restraining, one element severely attacks another element, stealing Qi from its energetic pool. This is generally observed in cases of Excess or Deficiency conditions of the organs. The Five Elemental Overacting Cycle and its organ correspondences are as follows.

1. The Wood invades the Earth, i.e., the Liver (Grandmother) overcontrols the Spleen (Child) and its paired organ the Stomach. In the clinic, this is commonly called "Liver invading the Spleen." Clinical manifestations include hypochondriac and epigastric pain and distension, irritability, poor appetite, loose stool, and a green face color.
2. The Earth invades the Water, i.e., the Spleen (Grandmother) overcontrols the Kidneys (Child). This occurs when a Deficient Spleen fails to transform and transport Fluids. These Fluids accumulate and obstruct the Kidneys' function of transforming and excreting Fluids. Clinical manifestations include edema, difficulty urinating, and a yellow face color.
3. The Water invades the Fire, i.e., the Kidneys (Grandmother) overcontrol the Heart (Child). This pattern is almost never seen in clinical practice since the Kidneys are rarely in Excess.
4. The Fire invades the Metal, i.e., the Heart (Grandmother) overcontrols the Lungs (Child). This results in Full Heat in the Lungs.

Pulmonary disease can be caused by cardiac disorders, this is known as "Fire Overacting and Invading Metal." Clinical manifestations include cough with yellow sputum, a feeling of Hot sensations, and a red face color.

5. The Metal invades the Wood, i.e., the Lungs (Grandmother) overcontrols the Liver (Child). Clinical manifestations include irritability, a feeling of distension, fatigue, and a white face color.

THE INSULTING OR COUNTERACTING CYCLE

This is a condition in Traditional Chinese Medicine where the elements reverse their restrictive cycle and the Child insults the controlling element, the Grandmother (Figure 23.6). This is generally observed in cases of Excess or Deficiency within the Five Elements. The Five Elemental Insulting Cycle and its organ correspondences are as follows.

1. The Wood dulls the Metal ax, i.e., the Liver (Child) insults the Lungs (Grandmother). If the Child (Liver Qi or Liver Fire) becomes stagnant, it can obstruct and prevent the Lung's Qi (Grandmother) from descending. Clinical manifestations include coughing, a feeling of distension in the chest and hypochondrium area, and asthma. If pulmonary disease is due to the impairment of the Liver, it is explained as "Wood Insulting or Counteracting Metal."
2. The Metal controls the spread of Fire, i.e., the Lungs (Child) insult the Heart (Grandmother). If the Child becomes Deficient, it can cause the Heart Qi (Grandmother) to become Deficient. Clinical manifestations include palpitations, breathlessness and insomnia.
3. The Fire dries up the Water, i.e., the Heart (Child) insults the Kidneys (Grandmother). If the Grandmother (the Kidneys' Yin) becomes Deficient, then the Heart (Child) can become excessive, giving rise to Heart Empty Heat. Clinical manifestations include dizziness, lower back pain, night sweating, insomnia, malar flushes, and dry mouth (at night).
4. The Water erodes the Earth, i.e., the Kidneys

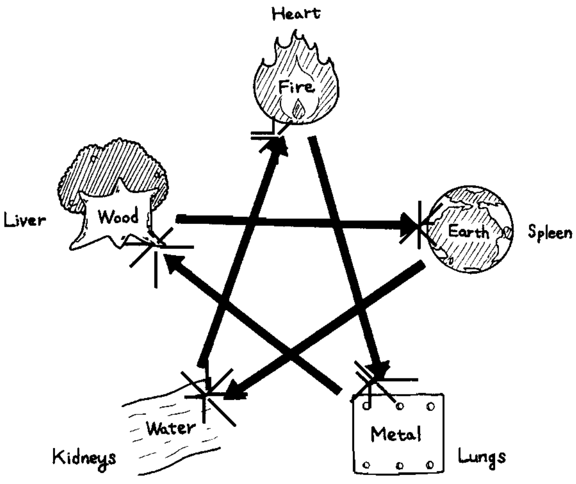

Figure 23.5 The Five Elemental Invading Cycle

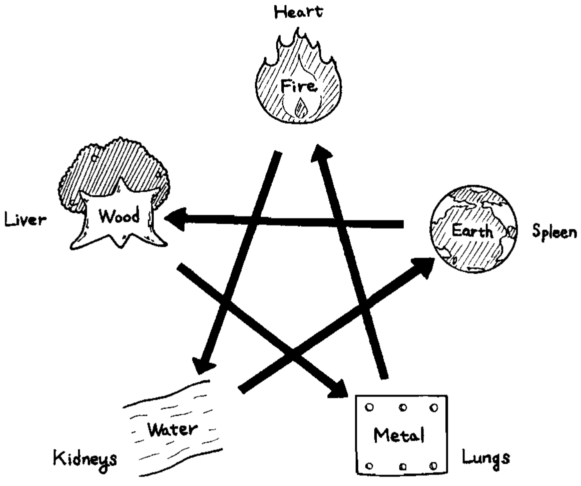

Figure 23.6 The Five Elemental Insulting Cycle

(Child) insult the Spleen (Grandmother). In this case, the Child (the Kidneys' Yang) becomes Deficient causing the Grandmother (Spleen) to become Deficient. Clinical manifestations include weak limbs, fatigue, edema, and loose stools.

5. The Earth resists the Wooden plow, i.e., the Spleen (Child) insults the Liver (Grandmother). If the Spleen (Child) fails to transform Fluids, a Damp condition can result. This Dampness can accumulate and begin to obstruct the flow of the Liver's Qi (Grandmother). Clinical manifestations include hypochondriac pain and distension, bile obstruction, and jaundice.

THE FIVE ELEMENTAL TONIFICATION CYCLE

The Five Elemental tonification cycle can be used for balancing the life-force energy of the body in conjunction with nature. Wood corresponds to spring, Fire corresponds to summer, Metal corresponds to autumn, Water corresponds to winter, and Earth corresponds to the late stage of each season (late spring, late summer, late autumn and late winter). The element Earth is centrally placed, associated with the 18 days of transformation at the end of each of the four seasons. The object of using the Five Elemental Tonification Cycle is to support and strengthen the next organ in the cycle (Figure 23.7).

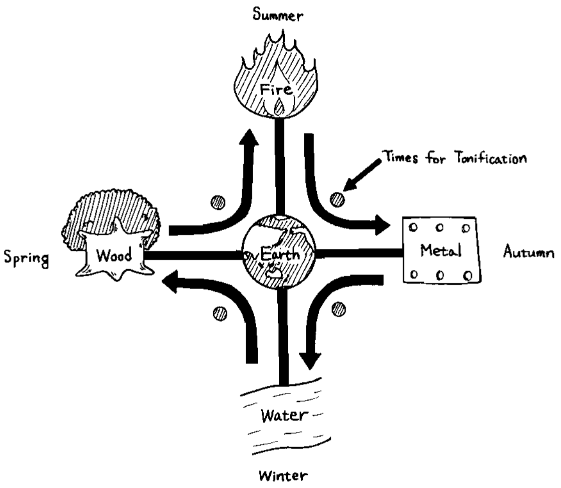

Figure 23.7 The Five Elemental Tonification Cycle

FIVE YIN ENERGETIC ORGAN MANIFESTATIONS

The Five Elemental energies are manifested through five different directions of movement. The Qigong doctor studies the movements and actions of a patient while they are in a meditative state (usually a sitting or standing posture), to diagnose their energetic disposition. After being in a meditative state for awhile, the patient's organs will overflow with Qi. When the organs overflow with Qi, certain physical reactions are evident, indicating which particular element has a problem.

1. Wood or Liver Qi is responsible for the dispersing effect of Qi, and is manifested by expansive outward movements in all directions, causing the body to sway with gentle and soft motions. When the Liver's energy is overabundant the torso makes a gentle, soft movement like a pine tree swaying in the wind. This is caused by the sprouting and branching manifestation of the Liver Qi's rising warm current.
2. When Fire or Heart Qi is overabundant, it is evident through the physical manifestation of the upward ascending actions of the torso caused by the flaring up of Yang Heat. This affects the energetic body and is manifested by upward movements and sometimes jerky actions (like a flame crackling in a fire).
3. Earth or Spleen Qi is generally manifested by stable movements, however, when overabundant, it is evident through movements that are both smooth and swift. This allows the body to progress in a limber, flexible, and free moving action (think of the body movement similar to a spinning ball).
4. When Metal or Lung Qi is overabundant, it is evident through the twitching actions of the torso and is manifested by contracted, inward movements, causing the body to violently jerk from side to side (think of the body as a metal bell being violently struck).
5. When Water or Kidney Qi is overabundant, it

is evident through the clumsy, awkward, heavy downward moving action of the torso (think of a waterfall descending onto rocks).

When the patient's body, unintentionally or unconsciously, rocks or sways gently as a result of a meditative state, the patient does not feel tired upon completing the meditation. However, when a patient consciously tries to force relaxation or augment the movement of his or her body while meditating, he or she will be tired by the end of the meditative session.

Note: While interning at the Medical Qigong Ward at the Xi Yuan Hospital in China, I found it fascinating to observe meditating patients displaying the results of both energetic organ manifestations or consciously induced augmentation of movement. All the above listed energetic manifestations were evident.

CHAPTER 24
OTHER DIAGNOSTIC SYSTEMS AND TECHNIQUES

DIAGNOSIS ACCORDING TO THE SIX STAGES

This system of diagnosis was formulated by Dr. Zhang Zhong Jing in the early Han Dynasty. The application of the six stages of diagnosis provides the key link in treating febrile diseases which are due to an attack of Cold. The six stages are described as Tai Yang, Shao Yang, Yang Ming, Tai Yin, Shao Yin, and Jue Yin.

THE THREE YANG SYNDROMES

The signs and symptoms of these pathological changes result from attacks by exogenous factors, especially cold weather. The first three syndromes describe the pathological changes occurring in the six Yang organs and channels (Small Intestine, Bladder, Large Intestine, Stomach, Triple Burners and Gall Bladder), and are called the Three Yang Stages. In the Three Yang Stages, the pathogenic factors predominate, but the body's energy is still relatively strong. The major symptoms are signs of Heat, and the treatment is directed at eliminating the pathogenic factors. These are the external conditions which can progress from the superficial Tai Yang Stages to the deeper Shao Yang Stages.

1. The Tai Yang or Great Yang Stage affects Qi on the exterior surface of the body and all three Wei Qi levels. This energetic disturbance affects the Small Intestine and Bladder Channels and organs that govern the exterior of the body. Although they are the most superficial channels, they are very rich in Wei Qi.
   - Clinical Manifestations of an Attack of Cold include: Aversion to cold, stiff neck, headache, no sweating, fever, body aches, floating-tight pulse, and shortness of breath.

   The focus of the Qigong treatment is to first open the pores to eliminate pathogenic factors by purging and dispersing, followed by tonification of the Lungs.
   - Clinical Manifestations of an Attack of Wind include: Aversion to wind, stiff neck, sweating, body aches, floating pulse, and fever.

   The focus of the Qigong treatment is to first open the pores to eliminate pathogenic factors by purging and dispersing, then to relax the muscles for diaphoresis (profuse sweating) in order to regulate the Ying and Wei Qi, and finally to prescribe herbs (or refer out for herbal treatment).
2. The Shao Yang or Lesser Yang Stage affects the interior and exterior levels of the body equally. This energetic disturbance affects the Triple Burners and Gall Bladder organs and channels. The patient experiences an alteration of fever and chills with a fullness of the costal and hypochondriac regions, blurred vision, loss of appetite, irritability, and wiry pulse with a white-slippery tongue coating.

   The focus of the Qigong treatment is on dispersing Heat with purging techniques and promoting the production of Body Fluids using herbs.
3. The Yang Ming or Bright Yang Stage affects Qi at the interior level of the body. This energetic disturbance affects the Large Intestine and Stomach organs and channels. The pathogenic factors transform into Heat.
   - Clinical Manifestations of the Bright Yang Channel Stage include: Pure Interior Heat known as the Four Bigs (fever, thirst, sweating, and pulse). In addition the patient also experiences an aversion to heat and has a red tongue with a yellow coating.

   The focus of the Qigong treatment is on dispersing Heat with purging techniques and on promoting the production of Body Fluids through herbs.
   - Clinical Manifestations of the Bright Yang Or-

gan Stage include: Interior Heat and the Four Bigs with constipation, profuse sweating, high fever which worsens in the afternoon, fullness and pain in the abdomen (which becomes worse if you press it), and a red tongue with a dry, yellow or black coating.

The focus of the Qigong treatment is on eliminating pathogenic factors with purging and dispersing techniques, and relieving constipation.

THE THREE YIN SYNDROMES

The second three syndromes are those of the Three Yin Channels. They are based on the pathological changes of the Six Yin Organs and channels (Lungs and Spleen, Heart and Kidneys, Pericardium and Liver). These syndromes are the reflection of pathological changes within the Yin and Yang organs, channels, and collaterals. In the three Yin stages, the pathogenic factors are still present but diminishing. The body's Qi has become weakened, and there are Deficient signs of Cold. The treatment is focused on strengthening the body's Qi. These are internal conditions that progress from the Tai Yin stages deeper into the body until they reach the Jue Yin stage.

1. The Tai Yin or Great Yin Stage affects the Yin organs, as the penetration of pathogenic factors flow deep into the energetic layers. This energetic disturbance affects the Lung and Spleen organs and channels, resulting in a Deficiency and Cold syndrome. The clinical manifestations are: Vomiting and diarrhea, abdominal fullness, absence of thirst, and a slow, deep pulse with a pale tongue.

   The focus of the Qigong treatment is on heating and energizing the Middle Burner and dispersing the Cold using herbs. It is also important to tonify the patient's organ Qi.
2. The Shao Yin or Lesser Yin Stage affects the Heart and Kidney organs and channels. Clinical manifestations are as follows:
   - The transformation of Cold Deficiency to Kidney Yang Deficiency: Aversion to cold, chills, cold limbs, diarrhea, listlessness, lethargy, abundant-pale urine, deep-fine pulse, and pale tongue.

     The focus of the Qigong treatment is on tonifying the patient's organ Qi, preventing the Yang from collapsing and treating with herbs.
   - The transformation of Heat Deficiency to Kidney Yin Deficiency: insomnia, fever, irritability, dry throat and mouth, scanty-dark urine, fine-rapid pulse, and a red tongue without coating.

     The Qigong treatment should focus on tonifying the patient's organ Qi, nourishing the Yin, and dispersing Fire using herbs.
3. The Jue Yin or Terminal Yin Stage affects the Pericardium and Liver organs and channels. This is the final stage of pathogenic invasion before the life-force energy vanishes altogether.

   Clinical manifestations are as follows: cold limbs, diarrhea, vomiting, pain, feeling of heat and energy expanding in the chest, and feeling of hunger with no desire to eat.

   The focus of the Qigong treatment is on purging, tonifying and regulating the patient's organ Qi. Applying Cold and Hot herbal medicines simultaneously can be used to drive out the patient's pathogenic factors and aid in purgation. Herbal tonics can also be used to increase the patient's resistance to pathogenic diseases and aid in tonification (Figure 24.1).

DIAGNOSIS ACCORDING TO THE FOUR LEVELS

This system of diagnosis was formulated by the famous Chinese physician, Dr. Ye Tian Shi in the late 1600's. These syndromes are used to explain the etiology and pathogenesis of febrile diseases. The differentiation of syndromes pertaining to the Four Levels of Diagnosis is concerned with febrile diseases caused by externally contracted Heat. Because exogenous diseases are always developing and changing, by studying the transmission of pathological changes the Qigong doctor is able to deduce the prognosis. These four levels are identified by four types of energy, and are classified as follows: The Wei Qi Level, Qi Level, Ying Qi Level, and Blood (Xue) Level.

1. At the Wei Qi Level (Protective Energy), the

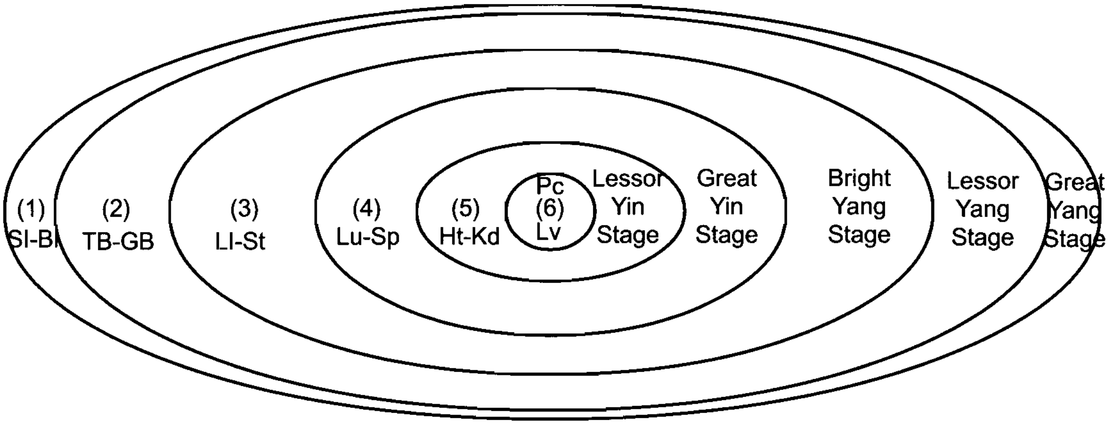

Figure 24.1. The Six Stages of Pathogenic Invasion are illustrated here. The invasion of external pathogenic factors follows a systematic progression through the body's Six Levels of Qi.

1. The Great Yang Stage (Tai Yang)

Affects the Small Intestine and Bladder Organs and channels. At this stage the pathogens are at the superficial level of the body's Wei Qi, caused by an attack of Wind and/or Cold Invasion. This stage is considered an External Cold Invasion at the Wei Qi level.

2. The Lessor Yang Stage (Shao Yang)

Affects the Triple Burners and Gall Bladder Organs and channels. At this stage the pathogens are half Exterior and half Interior, flowing in and out of the body, causing alternating chills and fever. This stage is also considered an Internal condition at the Qi level.

3. The Bright Yang Stage (Yang Ming)

Affects the Large Intestine and Stomach Organs and channels. At this stage the pathogens absorb deeper into the body's tissues, transforming into Heat affecting the internal organs and/or channels. This stage is considered an Internal condition at the Qi level.

- Bright Yang Organ Stage: Results in Interior Heat with constipation.
- Bright Yang Channel Stage: Results in pure Internal Heat with "Four Bigs" (fever, thirst, sweating, and pulse).

4. The Great Yin Stage (Tai Yin)

Affects the Lungs and Spleen Organs and channels. At this stage the penetration of pathogenic factors flow deep into the energetic layers affecting the Yin organs, invading the Middle Burner (affecting the Spleen) causing a Deficient condition. This stage is considered an Internal condition at the Ying Qi level.

5. The Lesser Yin Stage (Shao Yin)

Affects the Heart and Kidney Organs and channels. At this stage the pathogenic factors transforms either into Cold or Heat, causing further Deficient conditions. This stage is also considered an Internal condition at the Ying Qi level.

- Pathogens will transform into Cold due to a Deficiency of Kidney Yang.
- Pathogens will transform into Heat due to a Deficiency of Kidney Yin.

6. The Terminal Yin Stage (Jue Yin)

Affects the Pericardium and Liver Organs and channels. At this stage the pathogenic factors continue to deplete the body's life-force energy. This is the final stage of pathogenic invasion before death occurs. This stage is also considered an Internal condition at the Blood level.

pathogenic Heat invades the outer level of Wei Qi, causing the human body to manifest certain disorders which are clinically observed as fever, aversion to wind and cold, excessive or absence of sweating, sore throat, coughing, headache, red tongue with thin coating, and a floating and rapid pulse.

The Qigong treatment should focus on dispersing pathogenic factors and tonifying the exterior by utilizing pungent tastes and cooling herbs. Points are generally selected from the Lungs, Large Intestine, and Bladder Channels, as well as from the Governing Vessel.

2. At the Qi Level, the pathogenic Heat continues to progress deeper into the body, attacking and affecting the Qi system. At this stage, the body system is still strong, but the normal functions of the Yin and Yang organs are impaired. Because of the different functions of the Yin and Yang organs the manifestations differ. The three conditions associated with this stage of pathogenic invasion are as follows:

- Heat in the chest and diaphragm is the first stage and is not considered a severe syndrome. The symptoms manifest as fever, burning sensation in the epigastrium, thirst, and a yellow coated tongue.

  The Qigong treatment should focus on dispersing pathogenic Heat accumulated in the chest and diaphragm.
- Heat in the Stomach Channel is the second stage, and is considered a more severe Heat syndrome. These symptoms manifest as Heat, thirst, a dry tongue with yellow coating, and a slippery, rapid pulse.

  The Qigong treatment should focus on purging pathogenic Heat, and tonifying the body's Fluids.
- Heat in the Lesser Yang Channel is considered the final stage of Qi invasion. This syndrome is similar to the Lesser Yang Stage of the Three Yang Syndromes (according to the Six Stages), except that there is Dampness.

  The Qigong treatment should focus on purging pathogenic Heat from the Lesser Yang Channel, dispelling Dampness and dispersing Phlegm.

3. At the Ying Qi Level (Nutritive Energy), the pathogenic factors have penetrated deeper to the level of the Ying Qi and the organs, the Blood, blood vessels, and the Heart are therefore all affected. The symptoms in this stage are generally caused by the burning up the Yin. Clinical manifestations include: irritability, insomnia, mental restlessness, fever that worsens at night, dry mouth, absence of thirst, faint skin eruptions, a deep red tongue with little or no coating, and a fine-rapid pulse.

   Focus the Qigong treatment on purging pathogenic Heat from the Ying Qi level, using dispersing techniques. As an auxiliary method, bleeding may be performed by pricking the Heart and Pericardium Channels, as well as the Governing Vessel.
4. The Blood (Xue) Level, is the deepest and final level of pathogenic invasion. This stage contains the most serious manifestations that the disease can cause. Clinical manifestations include: high fever, skin that is hot to touch, Bloody stool, vomiting of Blood, Blood in urine, delirium, skin eruptions of purple or black color, a deep red tongue, a fine-rapid pulse, and in severe cases convulsions.

   The Qigong treatment should focus on purging pathogenic Heat from the Blood using bloodletting techniques (points are mainly selected from the Heart, Pericardium, Liver, Stomach, and Large Intestine Channels, along with the Governing Vessel), Cool the Blood, relieve convulsions, calm the mind, and reduce pathogenic factors with herbs.

DIAGNOSIS ACCORDING TO THE TRIPLE BURNERS

This system of diagnosis was developed by the famous Chinese physician, Dr. Wu Ju-Tong in the late 1700's. This system is concerned with febrile diseases caused by externally-contracted Heat, invading the body's Triple Burners. This Heat syndrome generally proceeds from the Upper Burner to the Middle and ends up at the Lower Burner. However, in cases of diseases due to Damp Heat, the pathogenic invasion of Heat begins at the Middle Burner in the Spleen.

1. The Upper Burner encompasses the body from the top of the head to the diaphragm. This area includes the Heart, Lungs, Pericardium, throat and head, and is responsible for respiratory and cardiac functions. The Upper Burner moves the body's Clean Qi, circulating and distributing nutrients and Qi throughout the body like a mist.

   Clinical manifestations of diseases in the Upper Burner correspond to those of a pathogenic invasion of the Lungs and Pericardium.

   a. When Wind Heat invades the Lungs, it is either manifested at the Wei Qi level or deeper in the Lungs themselves. The symptoms include: fever, sweating, coughing, stiffness and pain in the chest, wheezing, thirst, a red tongue with a yellow coat, and a rapid pulse.
   b. When pathogenic Heat invades the Pericardium, it is manifested at the Wei Qi level. These symptoms include: fever, coldness of the limbs, delirium, aphasia, burning sensation in the epigastrium, and a deep red tongue with spots.

   The Qigong treatment should focus on purging pathogenic Heat and dispersing Phlegm.

2. The Middle Burner encompasses the area from the diaphragm to the umbilicus. This area includes the Stomach, Spleen, pancreas, and Gall Bladder and is responsible for digestion, fermentation, and transformation of food and drink into nutrients for distribution. It moves the body's energy, circulating it like a swamp.

   Manifestations of diseases in the Middle Burner from pathogenic Heat are divided into two syndromes: Heat in the Bright Yang and Damp Heat in the Spleen.

   a. Heat in the Bright Yang include: symptoms of interior Heat and the Four Bigs (severe excess fever, sweating, thirst, and pulse) with constipation, profuse sweating, high fever which worsens in the afternoon, fullness and pain in the abdomen (which becomes worse when pressure is applied), and a red tongue with a dry, yellow or black coating.
   b. Damp Heat in the Spleen include: symptoms that are identical to the early stages of diseases caused by externally-contracted Heat.

   The clinical manifestations include: aversion to cold, fever that worsens in the afternoon, a heavy sensation in the head, body and limbs, fullness of the chest and epigastrium, nausea and vomiting, a white, sticky tongue, and floating or slippery pulse.

   The Qigong treatment should focus on Heating the Middle Burner, tonifying any Deficiency, purging Dampness, and prescribing herbs.

3. The Lower Burner encompasses the area from the umbilicus to the feet. This area includes the Liver, Kidneys, Bladder, intestines, external genitalia for men, and the uterus for women. The Lower Burner is responsible for filtering and eliminating waste products and for reproductive functions. It moves the body's Turbid Qi, circulating it like a drainage ditch.

   Manifestations of diseases in the Lower Burner are caused from pathogenic Heat attacking the Kidneys or Liver and are divided into two syndromes:

   a. Invasion of the Kidneys by pathogenic Heat dries up the Kidneys' Yin causing Heat symptoms due to a Yin Deficiency. Clinical manifestations include: lingering fever, hot hands and feet, dryness of the mouth, deafness, lassitude, a red-peeled tongue, and a rapid, floating or empty pulse.

   Focus the Qigong treatment on tonifying Yin, and purging the Heat.

   b. Invasion of the Liver by Heat occurs after the Kidney Yin has been exhausted. A condition known as Liver Wind is created. Clinical manifestations include: low-grade fever, cold limbs, dry and cracked lips, convulsions and trembling, a dry, reddish-purple tongue, and a rapid-deep pulse.

   The Qigong treatment should focus on tonifying Yin, and purging the Heat and Wind.

CHAPTER 25
CLINICAL EXAMINATION AND DIAGNOSIS

SENSORY, INTUITIVE AND PERCEPTUAL DIAGNOSIS

Clinical diagnosis can be categorized into two primary methods of energetic examination and treatment: Sensory Diagnosis and Intuitive and Perceptual Diagnosis. The clinical methods of Sensory Diagnosis are quite prevalent throughout China, especially within the acupuncture, massage, and herbal clinics. The methods of Intuitive or Perceptual Diagnosis, however, have been kept secret for many years, and are primarily found only within the Medical Qigong clinics. The latter methods (Intuitive or Perceptual Diagnosis) will be explained later in this chapter.

The root of any disease can be drawn from its complex symptoms and signs, by using the two main methods of Sensory and Intuitive and Perceptual Diagnostic principles. These two main methods of diagnosis serve as guidelines for the Medical Qigong doctor in clinical evaluation. Once the cause and extent of the disease is known, a treatment plan can be developed. There are four main types of Qigong treatment, which are chosen according to the treatment goals:

- a routine treatment (basic or nonspecialized),
- a specialized treatment (designed for the specific patient and his or her condition),
- a treatment aimed at etiology, or
- a treatment focused on alleviating the symptoms and signs.

It is therefore important to determine and assess the cause (root) and the manifestations (branches) of the patient's disease.

SENSORY DIAGNOSIS

Before treatment begins, the Qigong doctor performs a clinical intake, examining and evaluating the patient's condition, using various methods of sensory diagnosis. This clinical format involves taking a history of the patient's complaints. Next the doctor inquires about the patient's medical history. The doctor then conducts an examination for diagnosis, and devises a treatment principle on which the treatment plan is based, before implementing the treatment.

CLINICAL INTAKE, EXAMINATION AND EVALUATION

It is important that the doctor clearly identify the patient's problem before any treatment may be given. Never treat a patient without first making a diagnosis. The diagnosis is developed from the doctor's sensory examination and is evaluated according to symptom observation.

Both the root (the source and cause of the disease) and its branches (the symptoms or manifestations of the disease) should be examined carefully before treating the patient. In the case of an emergency, the doctor may choose to treat the urgent symptom first before treating its source.

The diagnosis and treatment of each patient traditionally begins by observing the patient as a whole, followed by an evaluation of the patient's individual constitution. Next, the doctor observes the symptoms and syndromes of the individual's organs. From these observations the cause of the disease is determined and treatment is begun. The treatment usually focuses on addressing the cause of the disease, to return the organ or organ system back into harmony with the rest of the patient's body. This affects changes within the patient's constitution and returns him or her back to a normal state of health.

Sensory examination for diagnosis is divided into five primary methods: diagnosis by visual observation, diagnosis by listening, diagnosis by smelling, diagnosis by palpation, and diagnosis by questioning the patient. By looking, listening, smelling, touching, and asking the patient specific questions, a Qigong doctor is able to diagnose the patient's syndrome (this includes the etiology of the disease).

DIAGNOSIS BY VISUAL OBSERVATION

1. Observe the patient's Shen, overall appearance, spirit, emotional nature, energy field, and composure. Look at the Shen of the patient's face; a healthy patient should have a brightness or vibrant shininess to his or her countenance. The Shen of the face should be checked against the state of the patient's eyes.
2. Observe the patient's eyes, especially the spirit in their eyes -- the glitter or dullness, and the expression of a controlled or uncontrolled emotional state. The eyes manifest the Essence of all the organs, and are the messengers of the Heart. There are two things to look at in the eyes -- glitter and focus.
   a. Are the patient's eyes bright and do they glitter? The less the glitter, the more long-standing the emotional and mental problems. It is said that there are three emotions that make the eyes lack glitter: sadness, grief, and shock. If the patient's eyes are dull, it looks as if he or she is seeing through a mist or film.
   b. Are the eyes focused, is the attention sustained and penetrating or uncontrolled? Uncontrolled attention is actually worse then having no glitter. If the patient has unfocused eyes it means that there is an obstruction of the patient's Shen and mind, generally due to guilt, preoccupation, or anxiety. It is said that Excess joy makes the "eyes uncontrolled."
3. Observe the color of the patient's face, complexion, eyes, ears, nose, mouth, tongue, skin zones, and ask about the discharges (mucus, vaginal, urine, and stool). Anger manifests as a greenish tinge on the cheeks or forehead (if it is affecting the Stomach); worry manifests as a grayish color, pensiveness as a sallow complexion, fear as a white complexion (unless combined with Kidney Yin which turns the cheeks red). Shock creates a bluish tinge on the forehead. A change in the complexion often indicates a deeper and longer-standing psycho-emotional problem. Also, if the eyes show no Shen but the face does, it indicates a recent problem.
4. Observe the patient's body and any outstanding structural features. Are they thin, emaciated, obese, or paralyzed? Are there deviations or obstructions in or on the body, face, head, mouth, teeth-gums, eyes, nose, and ears? Are there blemishes on the skin? Observe the appearance of the nails, hair and so on.
5. Observe the way the patient moves. Do they shake and tremor? Are they rigid or loose? Notice if the movements are fast or slow, and so on. When you shake hands upon meeting for the first time, are his or her hands clammy or dry; is his or her grip strong or weak?
6. Observe the patient's tongue. Is it cracked, serrated, fat, thin, or thick? If you observe a central crack, the deeper it is, the more long-standing the emotional problem. A red tipped tongue can also indicate a Shen disturbance.
7. Observe the patient's channels and points. Are there darkened red blotches, blemishes, swelling, sweat, etc.?

DIAGNOSIS BY LISTENING

1. Listen to the patient's speech for slurring or clarity.
2. Listen to the patient's tone. Is it high, low, loud, or quiet?
3. Listen to the patient's tempo. Is it fast or slow?
4. Listen to the patient's words. Do they make sense? Are his or her thoughts well ordered, or does the patient change subjects abruptly and frequently (if so, the patient may be suffering from a severe mental disorder).
5. Listen to the patient's respiration. Is it loud, strong, weak, or wheezing?
6. Listen to the patient's cough. Is it loud, dry, or wet?
7. Listen to the patient's Stomach and Intestines. Are they gurgling or quiet?

DIAGNOSIS BY SMELLING

1. Does the patient smell strong, weak, or odorless? Does the patient smell of medications, or alcohol?
- For Liver problems, the patient's body or breath will smell like a goat.
- For Heart problems, the patient's body or breath will smell like something has been burnt.
- For Spleen problems, the patient's body or breath will smell fragrant or sweet.
- For Lung problems, the patient's body or breath will smell like fish or a tide pool.
- For Kidney problems, the patient's body or breath will smell rotten.

DIAGNOSIS BY PALPATION

1. How does the patient's pulse feel? Is it strong or weak? The pulse shows more about the patient's Qi than the state of his or her Shen. Because the pulse reflects the short-term state of the patient's Qi, it should always be integrated with tongue and complexion diagnosis.
2. How do the patient's channels and points feel? Are there tender areas, movable nodules (Qi stagnation), or fixed nodules (Blood stagnation) along the patient's channels?
3. How does the patient's abdomen feel?

DIAGNOSIS BY QUESTIONING

1. Ask about the history of the patient's main complaint (illness, disorder, or injury).
- When did it begin?
- What other signs or symptoms accompany it?
- Ask about the course of the illness, is it improving or getting worse?
- What makes it better or worse (i.e., pressure, warmth, cold, and so on)?
- What other treatments has he or she tried?

2. Ask about the patient's general medical history.
- Has the patient had any surgeries?
- Has the patient experienced any past or recent physical or emotional traumas?
- Ask if the patient is on any medication(s).
- Ask about self-medication with alcohol or drugs.
- Ask about his or her current coffee, soft drink and sugar intake.
- Ask about the patient's extremities, do they feel hot or cold?
- Ask about his or her sweat.
- Ask about his or her thirst.
- Ask about his or her appetite and diet (how often does the patient eat, and how regular are the meals?).
- Ask about his or her digestion (gas, belching, etc.).
- Ask about his or her stool (loose, well formed, hot, smelly, cold, fishy, color, etc.).
- Ask about his or her urine (frequency, urgency, and color).
- Ask about his or her sleeping patterns (difficulty falling asleep, remaining asleep, or recurring nightmares).
- Ask for an energetic self-evaluation on a scale of 1 to 10.
- Ask about current emotional and mental problems or concerns.
- Ask about his or her current physical exercise program.

EXAMINATION FOR DIAGNOSIS

There are certain general principles used by all doctors of Traditional Chinese Medicine to become proficient in diagnostic skills. In Medical Qigong therapy, becoming skilled and accurate in clinical diagnosis requires the Qigong doctor to become proficient in combining perceptual and mental skills with spiritual understandings and insights. The following is a thorough and systematic approach for arriving at an accurate diagnosis.

1. Record all information observed and gathered as meticulously and impartially as possible. When first encountering a patient, be open minded, but note all first impressions.
2. Organize all the information according to the Eight Energetic Principles or Five Elemental Theories.
- When diagnosing according to the Eight Energetic Principles, the doctor should first determine whether the condition is Yin or Yang (chronic or acute), then classify the patient's

symptoms according to the condition's general location (External or Internal), followed by the disease's symptomatic appearance (Hot or Cold), and finally, the physio-energetic nature of the disease (Excess or Deficient condition).

Next, the Qigong doctor focuses on the Yin or Yang aspects of the patient's body, in terms of the physiological properties of the patient's disease, noting which is more affected, the Qi and Blood, Yin and Yang organs, etc.

The Qigong doctor then determines the precise depth and severity of the patient's disease, observing whether it is the patient's surface channels which are being affected, or the patient's internal organs. If it is determined that the condition is an internal problem, the Qigong doctor then differentiates the energetic level of penetration into the patient's body. This level of pathogenic penetration is diagnosed according to either the Six Stages, Four Levels, or the Triple Burners method of diagnosis.

- If the doctor is diagnosing according to the Five Elemental Theory, he or she should first determine the condition of the patient's pulse to assess which organ or organs are Excess or Deficient. The doctor must then decide whether the imbalance is attributed to either an individual internal organ or a combination of internal organs. If the patient's pulse does not match with the information gathered through the examination, the doctor must reconsider whether the patient's disorder will in fact conform to a Five Elemental pattern.

3. Evaluate all information: The doctor should evaluate the gathered information according to the patient's signs and symptoms, determining what is of primary and secondary importance. It is in this stage of evaluation that the Qigong doctor determines the "root and branch" of the patient's disease. If the disease is External in origin, the doctor determines what pathogenic factors are involved. If the disease is Internal in origin, the doctor determines whether the cause is emotional, hereditary, dietary, or an energetic charge remaining from a previous external disruption.

   Finally, the doctor should determine whether the cause of the disease is still active within the patient's life. If it still is, the doctor must take this factor into consideration before treating the patient and prescribing homework.

DEVISING A TREATMENT PRINCIPLE

In devising a treatment principle, the Qigong doctor should always include the Condition of the disease, Location of the disease, and Possible Cause of the disease.

1. The Condition of the Disease refers to the categorization of the disease, and may be diagnosed according to a combination of the Eight Energetic Principle's criteria. The patient's condition could, for example, be considered Excess or Deficient, Hot or Cold, Internal or External, or a combination thereof, such as: Excess Heat, Internal Heat, Deficient Heat, etc.
2. The Location of the Disease refers to the physio-energetic level of the patient's body that the disease is affecting (whether it is in the Upper, Middle or Lower Burner, or affecting certain Yin or Yang organs, channels, or muscles, etc.).
3. The Possible Cause of the Disease is either a result from a congenital or acquired illness. If congenital, consider the patient's DNA patterns, and all hereditary and ancestral traits. If acquired, consider environmental exposure to toxic poisons, diet, trauma (physical or emotional), electromagnetic fields, etc.

TREATMENT STRATEGY

Once the working diagnosis and treatment principle has been established, it is time to determine a treatment strategy. A treatment strategy consist in applying the treatment principle. It is the foundation from which the Qigong doctor chooses the approach in treating the patient's disease (Figure 25.1). The doctor will choose to treat the internal organs, channels, collaterals, points, and tissue areas which are the primary cause of the patient's disease. These areas are specifically chosen in order to facilitate a rapid affect on the patient's disease, to purge, tonify, or regulate the tissue area.

TREATMENT

The treatment usually focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal is to relieve the patient's symptoms, strengthen the constitution and return the patient to a state of wholeness. This process may involve restructuring the patient's belief structure, to emotionally release old traumas.

AFTER THE TREATMENT

After treating the patient's condition, the Qigong doctor explains to the patient in clear, comprehensive terminology, what the doctor has discovered, and his or her approach in rectifying the condition. The more the patient understands the

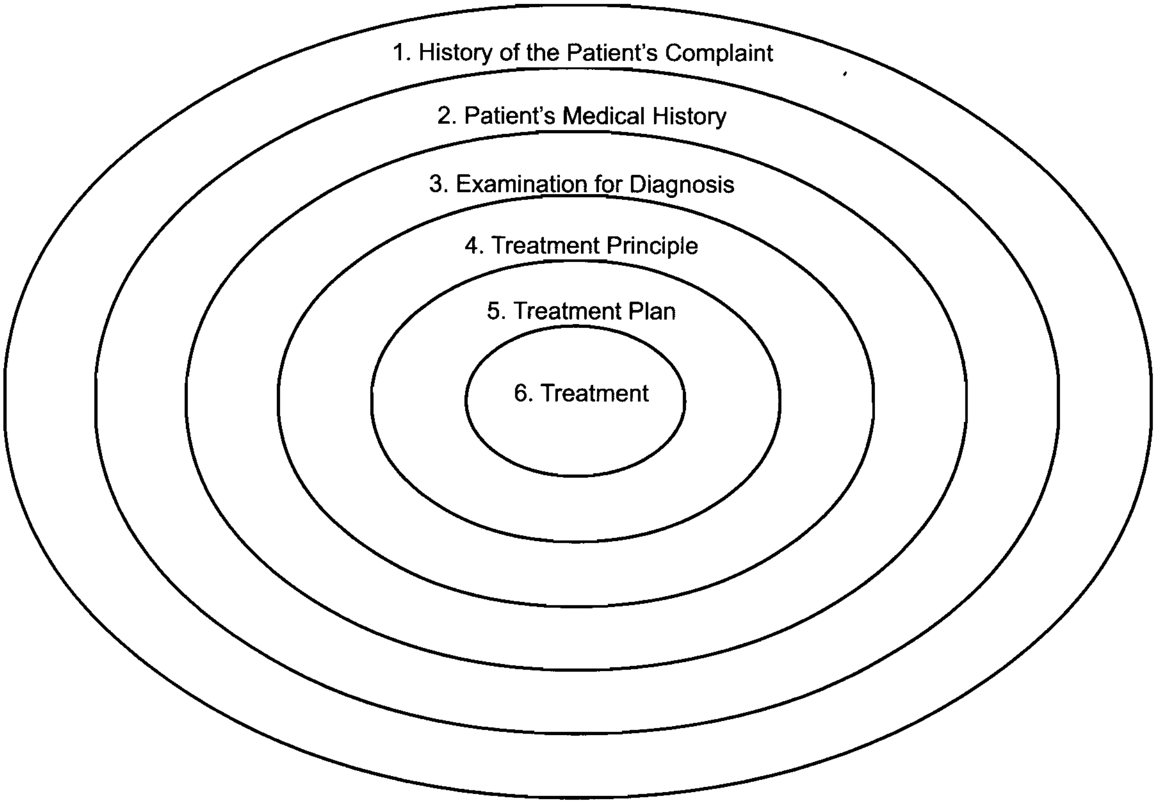
1. History of the Patient's Complaint
2. Patient's Medical History
3. Examination for Diagnosis
4. Treatment Principle
5. Treatment Plan
6. Treatment

Figure 25.1. Examination, Diagnosis, and Treatment: The Qigong doctor begins diagnosing the patient by taking a "history of the patient's complaint." Next the doctor inquires about the "patient's medical history." Then, the doctor begins the "examination for diagnosis," observing the patient's constitution, symptoms and signs and determines the syndrome(s), health and state of the patient's internal organs, and the cause of the energetic dysfunction. Next, all the acquired knowledge is assessed in accordance with the different diagnostic templates of the Eight Parameters, Five Elemental Theory, Qi, Blood and Body Fluid Dysfunctions, etc. and the doctor "devises a treatment principle." Before treating the disease the doctor "develops a treatment plan." The treatment focuses on the specific cause of the disease and how it affects the patient's internal organs. The primary goal is to relieve the patient's symptoms, strengthen the constitution and return the patient to a state of wholeness. This entire process is initiated to assist the Qigong doctor in prescribing the correct treatment.

doctor's diagnosis and approach in addressing the treatment, homework prescriptions, and suggestions for life-style modifications, the greater the likelihood of a successful healing.

After the initial treatment, it is important for the Qigong doctor to establish a treatment schedule. This establishes a realistic goal and time frame for the therapy. With each treatment, the doctor monitors the patient's progress, and may choose to modify the treatment approach in accord with the patient's changing condition.

As the patient's symptoms improve, the doctor lengthens the time between treatments. This continues until the symptoms are completely relieved, and the doctor is satisfied that the disease will not return.

ALTERNATIVE TREATMENT STRATEGIES

As the human body is in a constant state of change, the doctor's diagnosis and treatment approach will sometimes have to be altered. The doctor may also choose to modify each treatment according to the patient's changing symptoms or disease development. When choosing alternative treatment strategies, it is important that the doctor choose adjunctive therapies that enhance the primary goal of purging, tonifying, and regulating the patient's condition.

INTUITIVE AND PERCEPTUAL DIAGNOSIS

THE METAPHYSICAL DIMENSIONS OF MEDICAL QIGONG HEALING

The dictionary describes "metaphysical" as "the study of psychic phenomena beyond the limits of ordinary or orthodox psychology." There are other dimensions of reality that are very tangible, and access to them is only limited by fear and ignorance. Since theories on the nature of the mind and spirit cannot be easily verified or disproved by intellectual analysis or scientific experiment, sometimes certain kinds of knowledge are disbelieved, ridiculed, and dismissed. The advanced energy theories that are presented in this textbook are based upon the five thousand years of cultivation and experience of Chinese medical healing. The energetic form of Qi is neither good nor evil, it is simply another aspect of matter. The ancient Chinese Qigong masters refined the ability to analyze and diagnose these different forms of energy into an effective healing modality.

The existence of energy and its multidimensional patterns is embraced and accepted by many other cultures and societies, with the exception of Western conventional thinking, which is unfamiliar with this unorthodox approach to healing.

These energetic realms, which form the reality of the Medical Qigong doctor's clinical healings, are actually normal perceptional skills that the doctors have developed and trained within themselves. These doctors have simply rediscovered and trained their sensitivity to energies that they were open to during childhood. These energies encompass more then just the physical world. These perceptive skills and abilities of Shen are actually our natural, energetic "birthright," and have been with us since birth. Shen Gong skills, and all psychic powers, lie within the Eternal Soul of an individual, although in most people these skills are undeveloped and largely unused. Often we fear the most the things we understand the least. It has also been my observation that individuals are taught to ignore these skills because of one or more of the following reasons:

1. They have become frightened of them, because as children they were told that these abilities do not exist;
2. They were told that such abilities were evil, or of the 'devil';
3. Unable to understand them, they do not know how to integrate them into their personal lives;
4. They may fear the emotional and spiritual responsibility of knowing the truth about themselves and others; or
5. The known is familiar and therefore more comfortable then the unknown and new.

Sometimes even facing our own human potential alarms us so greatly (fear bordering on terror) that we are willing to deny empirical evidence that these paranormal skills and abilities are real.

We live, therefore, with a paradox. These paranormal abilities seem to be part of our nature, yet they can disturb us so greatly that we often violently reject them. Medical Qigong doctors are healers who have learned to accept these paranormal abilities and direct their energy, spirit and psychic skills towards the focus of alleviating pain and suffering. They are "wounded healers" who have learned to face their fears, and have the courage to walk through and overcome them.

THE AWAKENING

For the Qigong doctor to understand spiritual insights, three phases of energetic transformation must first be experienced. These three phases are collectively known as "the awakening" and are described as follows.

1. The awakening describes the beginning phase or "opening" of the Qigong doctors' Mind and Spirit to their energetic potential. This acceptance becomes rooted within the doctors' intention, emotion, and thought, and sets into motion the correct vibrational field for accessing their higher consciousness.
2. Entering the inner sanctuary describes the phase wherein Qigong doctors begin to receive greater clarity of mind and an elevated consciousness. The doctors receive information on a multidimensional level and are now able to access, as well as better understand their inner core being. This sets into motion the Qigong doctors' understanding of their interconnectedness to everything in the universe. At this point, the doctors must accept responsibility for their actions and non-actions.
3. Accessing the Wuji through the Creative Subconscious Mind describes the Qigong doctors' ability to access and reprogram their Mind and core vibration. This sets into motion the doctors' personal contact with their higher-self, or Eternal Soul.

LEARNING INTUITIVE AND PERCEPTUAL DIAGNOSIS

Before beginning the primary methods of Intuitive and Perceptual Energy Diagnosis, it is important to understand the initial training needed to access and train these skills. Without an understanding of these subtle energies, the risk of misdiagnosis of the microprojections of life-force energy is high. The subconscious mind communicates within ourselves and the outside world through the Three Dantians, by way of three distinct mechanisms:

- Physical or kinetic communication is established through the Lower Dantian.
- Emotional or empathic communication is established through the Middle Dantian.
- Spiritual or intuitive communication is established through the Upper Dantian.

These three modalities of energetic communication must be finely honed into effective, receptive listening tools.

Memories, thoughts, emotions, and knowledge are stored in the patient's tissues throughout the body at the pre-atomic and pre-cellular levels. They are arranged and structured similar to the way sounds and images are recorded on the magnetic molecules of a video cassette tape. Not only are emotions and sounds stored, but tastes, smells, and touch can be invoked simultaneously in vivid detail as the cells are stimulated. The patient feels that they are actually reliving the event from the past exactly as if it were occurring now.

The Qigong doctor must be ready to recognize, uncover, and be able to feel his or her patient's pain when interceding as an empath. Sometimes the doctor will not only stumble upon pockets of unresolved pain, but also suppressed memories hidden deep within the patient's tissues. When this occurs, the doctor diagnoses the patient's toxic emotional symptoms through his or her own body and spirit. To perform this type of diagnostic technique successfully, doctors must be secure in their established emotional boundaries. Otherwise they might project their own feelings onto the patient, or absorb the patient's pathogenic energy into their own body, where it may become trapped.

The goal is to absorb and receive the patient's symptoms briefly for diagnosis, and then release them by exhaling out the mouth (away from the patient). This expels the Toxic Qi into the ground, and allows the patient's feelings to pass through the doctor's body without harm.

The state of consciousness in which the doctor listens and feels the patient is more important than the technique or modality used. This is why there are so many diagnostic tools available to assist doctors in their evaluation. The patient's body maintains the energetic pattern of trauma until it is released and dispersed. The amount of energy required to neutralize this traumatic charge and return the tissues back to their normal state is in direct proportion to the intensity of the original trauma.

RECEIVING AND INTERPRETING THE PATIENT'S FIELDS OF BIO-INFORMATION

The human body emits several "bio-fields" of energy, varying in density and frequency and resonating within a holographic field. Like radio waves and other modern technological wave transmissions, these fields are encoded with information. The key to what we absorb and radiate lies in the energetic qualities of the rate or frequency of vibration, amplitude, and wave length of these bio-fields.

The physical world as we know it is made up of energetic fields. These energetic fields hold our universe together (Figure 25.2). Matter is simply the temporary expression of these energetic fields as they interact. The human body consists of Shen, Qi, and Jing actively resonating within the cells and tissues. Qi, as energy, is considered both substantial and insubstantial. Qi creates matter and acts as a medium between matter and spirit; it carries their mutual interactions in the form of waves. Because matter manifests as a distinctive type of consciousness, it operates at discreet levels of awareness, differing only in its complexity of structure and degree of freedom. Matter also manifests itself at every level of organization, from fundamental particles in physics, to biological organisms within the human body. The cells are governed by these bio-fields which determine their growth.

It is the conscious aspect of energy that dictates how, where and in what form energy is manifested within the body. This energetic consciousness releases bio-information which transmits encoded information about the patient both internally and externally. Form, force, and medium are three aspects of energy held in dynamic patterns by consciousness. The bio-information released from the patient's consciousness is read and diagnosed by the Qigong doctor before treatment begins.

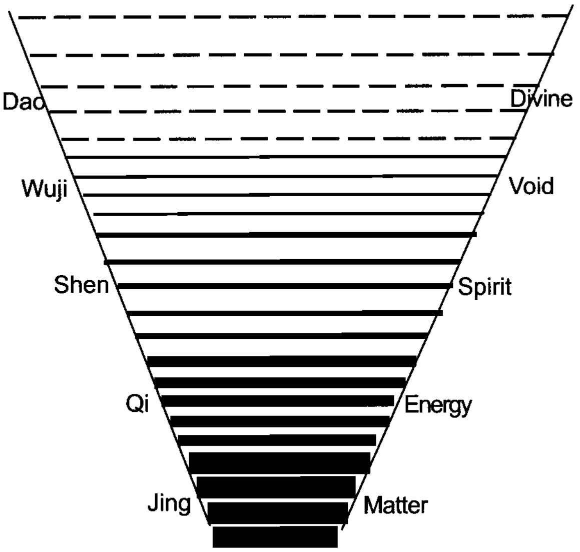

Figure 25.2. The Energetic Fields

The body's bio-information resonation Xin Xi or the "message" as it is called in China, is not limited by space or time. This Message refers to the energy fields that emanate to and from the Wuji or Void (infinite space embodied in between matter and energy).

The Wuji is the state of no boundaries, of pure openness, of complete oneness. Its essence is emptiness (as depicted by the empty circle), which relates to the awareness aspect of our Yuan Shen. When it begins to manifest within the relative universe of Qi, Yin, and Yang (referred to as Taiji), its nature appears as clear light which dispels darkness. When it manifests as a pure and tangible physical form, it appears as an enlightened being (saint or immortal) or as the "divinity" inherent within each one of us.

All of these levels exist within each individual, on both a pure level and pathological level. These levels can be detected and interpreted by the trained Qigong doctor as bio-information resona-

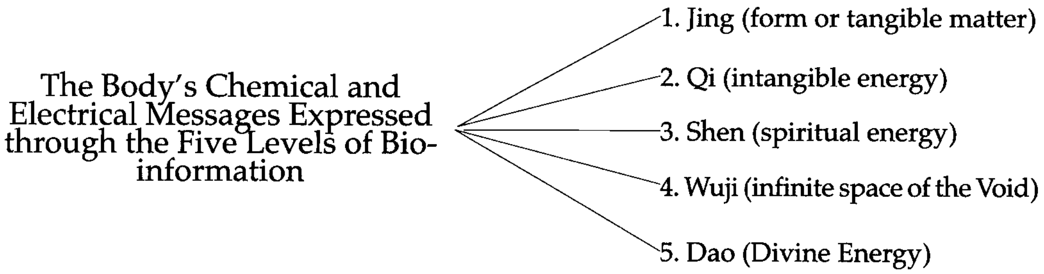
The Body's Chemical and Electrical Messages Expressed through the Five Levels of Bio-information
1. Jing (form or tangible matter)
2. Qi (intangible energy)
3. Shen (spiritual energy)
4. Wuji (infinite space of the Void)
5. Dao (Divine Energy)

Figure 25.3. The Five Levels of Bio-information

nce. Bio-information resonance is subdivided into five levels of expression:

1. Jing (form or tangible matter) contains the least amount of encoded information which is emitted at the lowest frequency.
2. Qi (intangible energy) contains more encoded information which is emitted at a higher frequency than Jing.
3. Shen (spiritual energy) emits an even higher frequency containing still more encoded information than Qi.
4. Wuji (the infinite space of the Void) contains even more encoded information than Shen which is emitted at a still higher frequency.
5. Dao (Divine Energy) emits the highest, most complete encoded information or Message and is the most powerful of all the emissions.

Medical Qigong views congenital disease as a form of distorted bio-informational vibration. The body's cellular systems internally regulate themselves via communication through these coded Messages. Messages are sent and received both on the energetic level (Shen, Qi, and Jing) and on the biochemical-electrical level. These two different levels interact continuously. Distortions on the energetic level interfere with the necessary transformations of Jing into Qi, Qi into Shen, and vice versa. Failures in energetic transformations also affect and distort chemical and electrical Messages between the different types of cells in the body (blood cells, tissue cells, nerve cells, and so on).

The maturity of the physical body depends on the transference of these Messages as the source of its evolution. Each time the egg cell divides (morula, embryonic, and fetal stages) the Message is transcribed into the newly-generated cells. Some encoded Messages are transformed but, due to the adverse evolutionary conditions or pathogenic factors, some become latent. Germ cell Messages, however, are not restricted by these evolutionary conditions since they multiply, evolve, and transform at an accelerated pace.

Information of the body's disharmonies is received as a distorted, biological resonation, which indicates the location and severity of the patient's condition. All diagnostic scanning techniques utilize this type of perception. The body's cellular systems internally regulate themselves via communication through coded bio-informational signals. This biological information exists in energetic, electrical, and chemical forms which are constantly being transformed from one form into another (Figure 25.3). The function of Medical Qigong therapy is to provide the correct bio-informational instructions (or Message) to the patient's body in the form of Qi emission (or coded energetic bio-informational signals) to initiate healing. By receiving Messages from the patient, the Qigong doctor can project energy while receiving and diagnosing the patient's energetic patterns.

ACCESSING THE KNOWLEDGE STORED WITHIN THE WUJI

When the Qigong doctors emit healing energy to the patient, they are sending healing Messages encoded within the Qi. When these Messages are received by the patient's cells the healing process is initiated. Qigong doctors access the knowledge needed to treat patients, by remaining receptive to the Messages stored within the patient's Jing, Qi, Shen, and surrounding energetic space (Wuji or Void).

Within the energy of the Wuji is the knowledge stored throughout time. In Medical Qigong practices, accessing the knowledge of the Wuji is the Chinese equivalent of the Sanskrit concept of accessing the Akashic Records. "Akasha" is a Sanskrit term used to describe the all-pervasive space of the universe. There are two kinds of Void: the first type of Void is internal, limited by the material plane (or matter) and can be associated with the enveloping of the internal human form. Matter can be described as consisting of more space then actual physical matter and, therefore, is considered infinite in its internal division of energetic properties. Within this Void forms the foundation for the matter or the body's personality by way of physical forms, sensations, perceptions, mental formations, and consciousness.

The second type of Void is external, unlimited, beyond all description, unbound by the material, yet contains all things material. It is the vehicle for all life, resonating through sound and light, permeating everything in the universe. In the practice of Medical Qigong the Wuji is one of three universal principles, along with Qi and Shen, which form a trinity of sources of intuitive and psychic power. These three universal principles also manifest the Eternal Soul, allowing divine thought to infuse matter.

The knowledge of the Wuji records the vibrational resonance of every action (all thought and emotion), as well as light and sound. These energetic impressions are stored within the astral plane. These records exist as impressions in the astral / spiritual dimension, providing a sort of accessible filing system for those who wish to receive information about past history or past lives, or even for the examination of their own spiritual progress. Admittance to this sea of knowledge and wisdom can be triggered in the hypothalamic limbic system of the brain, and is accessible through the energetic stimulation of the third ventricle of the Qigong doctor's brain.

The body is composed of literally billions of energetic molecules. Each molecule is a hologram of ancestral particles, knowledge, and experiences existing throughout time, spanning our entire history as we know it. These molecules gather together to form and create matter. They function for specific purposes of life transitions and energetic interaction, and then dissolve, and transform back into Qi and Shen. Each molecule stores its energetic experiences, to be later accessed through spiritual intention.

As the molecules gather to form a fetus, energy as well as ancestral history is stored within the tissues and cells of the child via the environmental, universal, mother and father's energetic fields.

Once the Qigong doctor has fused with his or her patient, he or she will be able to access specific information about the patient through the internal connection to the energetic impressions of the patient's Void. This allows the doctor to study and learn about the patient's past history of disease formations, as well as the information contained within the molecular structures of the patient's tissues.

UNDERSTANDING PAST-LIFE REGRESSION

Although seldom talked about, and much less understood, the clinical significance about past-life regression is quite important. Psychiatrists Sigmund Freud and Carl G. Jung both believed that the patients' worst fears, pain, and traumas are buried deep within their unconscious mind. Sometimes the patients' emotional traumas are so deeply entrenched that they can only be expressed through the energetic presence of a different time-space reality. One metaphor used by patients to describe this other form of reality is through a "past-life trauma."

During Medical Qigong treatments, patients will sometimes describe traumas or situations that do not relate to present life experiences. These experiences might be explained through past life theories. There are six major theories used to explain past-life recall (Figure 25.4). These theories are described as follows.

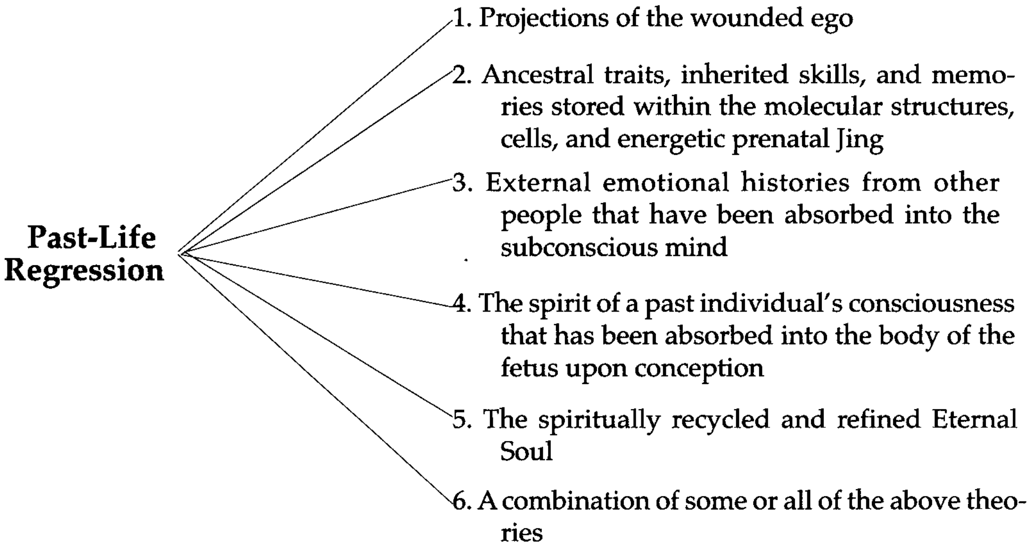

Past-Life Regression

1. Projections of the wounded ego
2. Ancestral traits, inherited skills, and memories stored within the molecular structures, cells, and energetic prenatal Jing
3. External emotional histories from other people that have been absorbed into the subconscious mind
4. The spirit of a past individual's consciousness that has been absorbed into the body of the fetus upon conception
5. The spiritually recycled and refined Eternal Soul
6. A combination of some or all of the above theories

Figure 25.4. The Six Major Theories of Past Life Regression

1. Some Qigong doctors contend that these past-life memories are the projections of the patient's wounded ego, used by the subconscious mind to assist the patient in dealing indirectly and symbolically with life's present painful experiences. These are painful memories that are consciously avoided by the patient, but are necessary for healing to take place.
2. Other doctors believe that these past-life phenomena are the ancestral traits, inherited skills, and memories stored within the molecular structures, cells, and energetic Prenatal Jing of the patient, that periodically surface from time to time and generation to generation.
3. Some doctors believe that apparent memories of past-life regressions, as well as of past-life traumas, are external emotional histories from other people that have been absorbed into the patient's subconscious mind. The patient responds like an energetic empath to sensations of telepathy, registered by the patient's hypothalamic limbic system. These external memories are triggered when the patient accesses the knowledge of the Wuji.
4. Other doctors believe that the spirit of a past individual's consciousness can be absorbed into the body of the fetus upon conception. This means that wherever a child is conceived, the environmental energy, as well as any native hovering spiritual entity, can be implanted into the forming energetic field of the fetus.
5. Other doctors believe that one's Eternal Soul is constantly being recycled and refined through generation after generation, constantly being reborn in order to evolve into a higher plane of consciousness.
6. Still other doctors believe that past-life memories can be a combination of some or all of the above theories.

Some patients feel that they experience spontaneous past-life recall through the treatment of Medical Qigong therapy. While receiving treatments or performing Medical Qigong prescriptions these patients may experience intuitive flashes or visions. Whether the patients' memories actually are of their own historical past lives, or are reconstructions from subconscious material, is a matter of controversy. The primary objective is for the doctor to accept the patients' reality, supporting them in this belief system to facilitate their healing process.

It is the doctor's responsibility to respect the patient's reality and not impose his or her own interpretations on everything the patient says. By accessing the knowledge of the Wuji, the doctor can assist the patient to connect with and benefit from these past life recollections or "reconstructions" to facilitate the healing process. Ultimately it matters not if the memories are reconstructions of subconscious material, true past life events, or even telepathic experiences. What is important, however, is that these memories are processed, learned from, and integrated to promote a better balance within the patient.

PAST-LIFE HEALING THERAPY

Clinically, past-life regression therapy has been effective in treating such problems as chronic fears and phobias, detrimental habits and cravings, suppressed hostility and guilt, depression, the fear of death, sexual dysfunctions, chronic pains and headaches, as well as many physical disorders. It is also used for understanding relationship problems, since the theory maintains that we often meet the same souls in different lives (and bodies) until the problems are worked out.

The Qigong doctor is not qualified to perform past-life regression therapy. This requires a specialist or at least supervision by an expert. The Qigong doctor can, however, facilitate healing by accepting the patient's experience and validating it to prepare the way for the healing process. In Medical Qigong therapy, the doctor accepts the patient's belief that the causes of his or her present physical or psychological problems can be traced to past-life traumas or death experiences. The improvement attained through past-life memories is said to be achieved through the release of the energetic charge that is trapped in the patient's body from a former life. This energetic charge keeps the patient's emotions prisoner, chained to the old feelings that are associated with the painful past-life experience. It is important for the Qigong doctor to realize that the same therapeutic benefits as the retrieval of current-life real memories can be achieved by allowing the patient to relate any images, impressions and so on, that may or may not stem from his or her present life experiences. The foundational base of the emotional trauma will often be rooted in feelings associated with rejection, betrayal, and abandonment. The psychological and physiological reactions to the patient's past-life memories, whether they be real or contrived, will have a profound affect on the patient's healing.

LEARNING MEDICAL QIGONG TECHNIQUES

According to Qigong Master Zheng Zhan Ding from Beijing, China, when learning Medical Qigong techniques, the doctors experience what is known as the Three-Part Wisdoms.

1. The Upper Dantian Wisdom is achieved through divine calling. When the full potential of the Upper Dantian is tapped, the doctor suddenly finds him or herself able to utilize skills and techniques that were previously completely beyond his or her ability. This is a major energetic breakthrough for the doctor, enabling him or her to perform supernatural feats that he or she immediately recognizes and accepts as part of a personal spiritual calling.
2. The Middle Dantian Wisdom is achieved through inspirational observation. Upon seeing energetic techniques performed, the doctor immediately establishes an emotional belief that he or she possesses that same energetic ability. This is a kind of "if they can do it, I can do it" observation and belief structure.
3. The Lower Dantian Wisdom is achieved through practice. This is a type of "practice makes perfect" method of transition which allows the doctor time to accept and integrate new beliefs and abilities through long hours of constant practice, observation, and tutelage.

All Qigong techniques are based on the imagination and spiritual belief. It is necessary to create an image (a spiritual belief) within the mind, to create a positive environment for healing transformations (the spirit transforms energy, and energy transforms matter). Sometimes the Qigong doctor will want to utilize a certain image of his or her belief, but because of certain doubts, is unable to bring it into existence. In order to successfully cultivate this image, the doctor must use creative thinking to focus his or her life-force energy on this specific objective, accept it as true and know it will happen. To create this image the doctor must use his or her imagination, emotional feelings, unwavering belief, visualization, affirmation, and acceptance.

Energy (Qi) is the bridge between the material and the spiritual, and is the medium through which consciousness interacts with the world. Through intention, Qi is gathered, focused, and used to transform light, sound, and other resonations. Through intention, the Qigong doctor is able to change structural formations, and to transform tangible and intangible forms and vice versa, to dissolve tumors, cysts, and so on. Of the many methods of healing available in Medical Qigong, it is important for the Qigong doctor to explore the healing skills suited to his or her body and personality, to become most effective. Once the doctor has learned and mastered these healing techniques, he or she will no longer need to think about the methods, or when they need to be employed, they will become second nature.

PERCEPTUAL COMMUNICATION TRAINING

Subconscious communications are divided into physical, emotional, and intuitive forms of interaction and perception. The internal-vision techniques required to utilize energetic perception in a clinical setting are expressed and enhanced through these types of communication. These types of communications are realms of perception that can be explored, wherein, the Qigong doctor is so present in "the moment" that his or her attention begins to dissolve the perceptual realities of all five senses, and begins to perceive and diagnosis patients on a spiritual level. At this level, all feelings, sounds, colors, or any combination of the doctor's senses take on new dimensions of reality.

DIAGNOSIS THROUGH DANTIAN COMMUNICATION

Any treatment, exercise, or meditation the Qigong doctor prescribes must connect with the energetic resonation of the patient's disease. This connection is established through the doctor's ability to communicate with the patient's body via the Three Dantians.

1. The Upper Dantian's perceptional awareness is received through a type of spiritual/intuitive communication. Through this intuitive subconscious communication the doctor is able to interact and perceive the patient's spiritual and energetic disturbances. This allows the doctor to perceive the patient's spiritual condition without bias.
2. The Middle Dantian's perceptional awareness is received through a type of emotional/empathic communication. Through this empathic subconscious communication the doctor is able to interact and perceive the patient's feelings, actively empathizing, listening, and hearing what patients communicate about their pain and problems. This allows the doctor to perceive the patient's true emotional condition by bypassing the person's ego.
3. The Lower Dantian's perceptional awareness is received through a type of kinesthetic communication. Through this physical subconscious communication the doctor is able to interact and perceive the active condition of the patient's physical tissues. This allows the doctor to perceive (through his or her hands and physical body) the patient's physical symptoms, such as pain, Heat, and so on.

UTILIZING INTERNAL-VISION AS A CLINICAL MODALITY

The brain has the ability of receiving and recording the vibrations of any object that the senses see, hear, smell, taste or feel. The vibrations of light, sound, smell, color, and object's shape are all recorded. This also holds true for thoughts, words, and actions. The brain also has the ability of reproducing these vibrations and projecting them out again through the use of inner vision and the imagination.

Before beginning internal Qi diagnosis it is important that the Qigong doctor have a thorough understanding of the internal energy's function and flow within his or her own body. Dr. Li Shi Zhen pointed out during the Ming Dynasty that, "The internal organs and channels can be perceived as unobstructed and free from disease only by those who can see internally."

This is achieved by performing the Nei Guan or Internal Viewing Meditation. The purpose for internal-vision is to carefully examine, through conscious introspection, the energetic flow and structure of the body and mind. Most doctors in the Ming Dynasty studied the Nei Guan Meditation as part of their clinical practice to develop the light for internal viewing. It is believed that light enters the body through the Baihui Point (at the top of the head), the Yin Tang Point (the Third Eye area) and through both eyes. Since internal-vision uses light that has a higher vibration than visual light, it is able to penetrate through the skin deep into the tissues.

Many years ago I was truly skeptical of this phenomenon, although I met many Qigong masters who possessed such skills. After acquiring this skill, however, I was able to demonstrate its effectiveness quite thoroughly while interning in the Qigong hospitals of Beijing, China. It is actually a simple matter of consistent practice and focused concentration which allows this sense of perception to become tangible. It is important to remember that to know something, one must first think it, then feel it in order to integrate the knowledge fully. Always watch as an observer, void of emotional attachment, noticing what is actually there.

Nei Guan—Internal Viewing Meditation

The best way to practice the Nei Guan Meditation is through relaxed, quiescent breathing. Begin from a Wuji posture. After quieting the mind, relax and focus the imagination. Imagine that the left eye is a bright radiant sun, while the right eye is a bright and luminous moon. Both the sun and moon join together at the Yin Tang (Third Eye point). As their energies come together, they form a bright white ball. This light actually gathers around the pituitary gland, illuminating the optic chiasm, filling the occipital lobes (for internal perceptual vision) and the thalamus (for oculomotor control) (Figure 25.5).

As the light expands it stimulates the corpus callosum and penetrates the pineal gland (which acts as a projector for internal vision) (Figure 25.6).

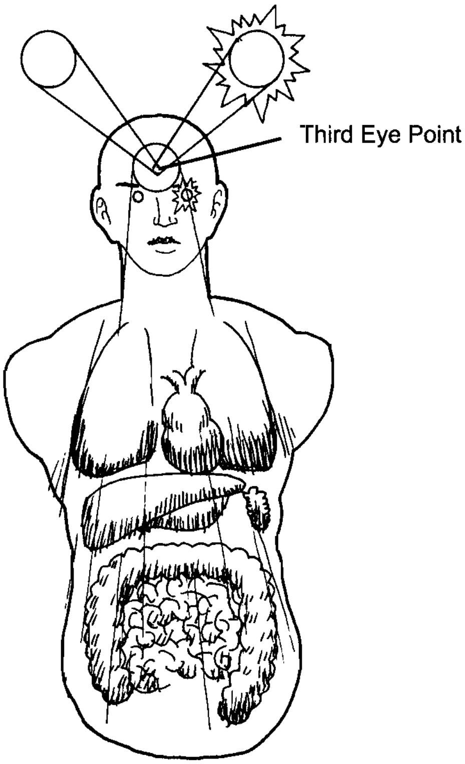

Figure 25.5. The Internal Viewing Meditation

Next, focus your intention on directing this ball of white light, allowing it to shine down into the body, illuminating all the internal organs. This enables you to see directly inside your own body and view each organ. Observe and feel the energetic strengths and individual weaknesses of the bones, organs, and tissues. Allow both energy and information to gather and expose itself through the Third Eye region of the Upper Dantian. After a while, relax and return to Wu Wei (i.e., No Mind), allowing all of these images to settle.

After a relatively short time of constant practice Qigong doctors learn to discern the state of their own organs, and are able to regulate their own physical and energetic body. This internal diagnostic ability sets the foundation for doctors to extend their Internal-viewing capability outside their own body to use on patients.

Once this occurs, doctors find it possible to effectively diagnose patients while scanning and using the Flat Palm Detection technique. The patient's points and areas for diagnosis are revealed to the doctors through this modality of energetic extension.

When using internal-vision, extend your intention as deep into the patient as possible, to slowly scan the patient's tissues, layer after layer, observing any cellular changes and distortions. Use the internal light resonating from inside the patient's body to identify the pathogenic factors. Then, look deeper into the energetic currents to find the root and origin of the patient's disease.

When you use your hands to read the patient's internal organs through Flat Palm Detection or any other diagnostic modality, it is extremely important to use internal-vision to determine the severity of the disease. You may observe a specific color, texture, or feel pools of emotional energy. According to Dr. Zheng Zhang Ding, doctors will be able to determine the condition of the diseased organ by its shade of color. The internal-vision colors most commonly observed in internal-vision diagnosis are as follows:

- A red or yellow color represents a healthy or normal organ;
- A white color signals the beginning of energetic dysfunction, but is not serious enough for treatment;
- A gray color signifies sickness; and
- A black color denotes, tumors, cancer, etc.

APPLICATIONS OF INTUITIVE AND PERCEPTUAL QI DIAGNOSIS

There are a variety of methods concerning the application of intuitive and perceptual diagnosis of diseases. The specific physiological methods such as remote sensing, hand detection and physical body observation are very popular in the Medical Qigong clinics in China today. Five of the most favorite modalities of intuitive and perceptual Qi diagnosis are described as follows.

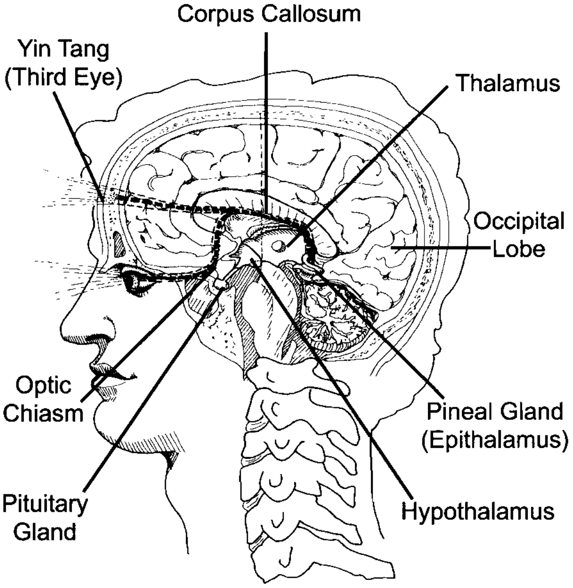

Figure 25.6. The inner light gathers at the pituitary gland, illuminating the optic chiasm, filling the occipital lobes and thalamus, stimulating the corpus callosum and penetrating the pineal gland.

DIAGNOSIS BASED ON FLAT-PALM DETECTION (HAND SENSING)

Because the body's energetic field reflects the state of health or dysfunctions of the physical body, any changes occurring in the tissues and organs are detectable through sensing this energetic field. Detection of these changes with the palm is termed Flat-Palm Detection.

To perform this technique, the doctor begins by asking the patient to take a standing, sitting or lying posture. The patient's whole body should be relaxed, the mind focused on breathing naturally. The doctor positions his or her palm several inches from the patient's energy field, with the Laogong Pc-8 point or fingertips facing the direction of the patient's body.

The doctor then emits a little energy, slowly, followed by Pushing, Pulling and Leading manipulations used to detect the Qi of the patient's energetic body. The patient's surface tissue is divided into upper and lower, left and right aspects. The doctor focuses on the surface tissue area first, specifically the area corresponding to the region of the patient's disease. The doctor then determines the energetic nature of the corresponding regions by making comparisons between the sensations gained in different areas to find where energy feels abnormal. Gradually, the doctor's emitted energy is extended deeper into the body to purge, disperse, tonify or regulate the patient's organs and organ systems (Figure 25.7).

The Qigong doctor then analyzes these energetic sensations, comparing how they feel from the different areas of the body (both on the superficial and deep layers). The doctor categorizes the pathological findings (the location, nature, and severity of the disease) in accordance with the Eight Principles of Diagnosis and the energetic patterns of The Five Elemental Theory.

Feeling a great difference in energy intensity between the two sides of the Stomach Channels, for example, indicates a Yin and Yang imbalance of energy circulation between the left and right sides of the body. The Deficiency and Cold syndrome of the Spleen and Stomach may be manifested by a cold and contracted sensation in the epigastric region.

When an acupuncture needle is inserted into an obstructed point the energy grabs the needle and vibrates it. Once this energy block is removed, the needle ceases to vibrate because the channel is now open. This same phenomenon occurs when the palm is placed on an area of stagnant Qi. The palm vibrates until the obstructed energy is dispersed.

Energy blocks are commonly cleared by dredging in a downward direction from the head (Heaven) to the feet (Earth). When treated this way, patients will experience a temporary feeling of fatigue as the stored toxins are released from the body. Tonification is facilitated by bringing the energy current up the patient's body from the feet to the head (Earth to Heaven). This action will sometimes cause patients to have psychic experiences and see visions, as the energy enters the patient's Upper Dantian region.

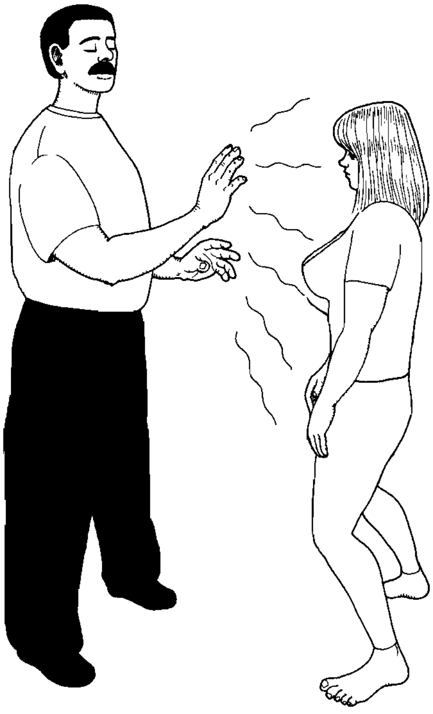

Figure 25.7. In Flat Palm Detection, the Qigong doctor uses the entire palm to sense the condition of his or her patient's internal organs and tissues.

DIAGNOSIS BASED ON KINESTHETIC AND EMPATHIC QI ABSORPTION

There are some individuals who tend to easily absorb the feeling and emotions of other people, even without trying. Sometimes when treating a patient, a Qigong doctor will inadvertently ingest a small amount of their patient's energy. As this energy absorbs into the doctor's body, it infuses throughout his or her entire energetic field. This energetic fusion sometimes causes the doctor to experience abnormal feelings in the locations that correspond to the pathogenic areas of the patient's body. From these feelings, the doctor can diagnose the location and the nature of the patient's disease.

To actively perform this method of Kinesthetic and Empathic Qi Absorption, the Qigong doctor stands facing the patient (Figure 25.8) and projects white healing light energy toward the top of the patient's head (Baihui GV-20 point). The patient is thus enveloped by the white healing light energy, from the top of the head, to the bottom of the feet. Next, the doctor directs the Qi to flow from the bottom of the patient's feet into the bottom of his or her own feet, absorbing the patient's energy. This energetic action forms a complete cycle of energy between the doctor and the patient's body. The Qigong doctor then begins to carefully feel the sensations of the patient's energy in his or her own tissues and organs. The locality of abnormal feelings, their shape, and/or prominent characteristics reflect the locality and nature of the patient's illness.

Note: It is generally encouraged to avoid this type of body detection on cancer patients, because the risk of pulling in the patient's excessively Hot pathogens into the doctor's body is too high, especially if the doctor's energy has become depleted.

HAND DIAGNOSIS BASED ON BODY ASSOCIATION

A small percentage of energy absorbed from the patient can suffice to diagnose the whole energetic system. Because the law of bio-holograms holds that one part of the body contains information about the whole (including various parts of the body, such as the viscera, Qi, and Blood, etc.), information obtained from certain isolated areas can reflect the condition of the whole body. Such parts may include the hand, foot, eye, ear, nose, face, etc. The location on these parts that represent different areas and organs are termed as "points" or "zones:" they include hand points, foot points, ear points, and so forth.

Hand diagnosis (called Shou Zhen) is used to evaluate the physical predispositions and ailments of patients. This diagnostic technique first appeared in Chinese medical literature during the Sui Dynasty (618-581 B.C.) from the clinical writings of Dr. Chao Yuan Fang, who authored the "General Treatise on the Origin and Symptoms of Various Diseases." To begin diagnosis through hand-body association, the Qigong doctor first neutralizes his or her left palm (by releasing the energy into the Earth) to balance and stabilize its sensitivity. The left hand is generally used for reviewing energy and diagnosing diseases.

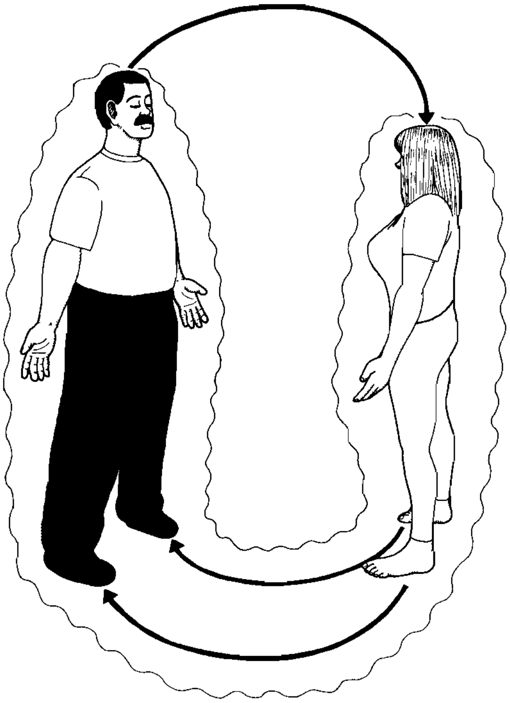

Figure 25.8. In Diagnosis based on Kinesthetic or Empathic Absorption of the patient's Qi, the Qigong doctor temporarily absorbs the patient's energy field into his or her own body.

Once the left palm has been neutralized, the doctor uses his or her intention to gather the patient's Qi into the left palm, distributing the absorbed energy evenly throughout the hand. Focused concentration on any one specific area of the hand should be avoided. The doctor imagines the patient's energy enveloping the left hand and fingers, like an energetic glove. The doctor carefully begins to sense the sensations of the patient's energy on his or her palm. The center and front of the doctor's palm is considered the center and front of the patient's body, while the back of the hand is considered the back of the patient's body. The doctor feels various kinds of sensations on specific areas of the left palm, which indicate the location of the patient's disorders. These different sensations,

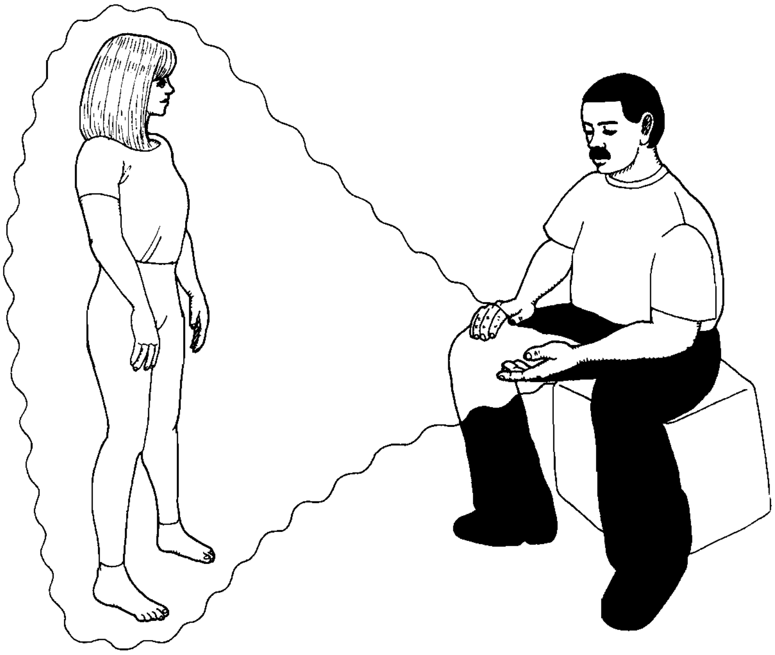

Figure 25.9. Diagnosis through Hand/Body Association

such as distention, heaviness, numbness, heat, cold, itching, and throbbing correspond with the features of the disorders (Figure 25.9).

While in Beijing, China, I was fortunate to study under Qigong Master Zheng Zhan Ding. Master Zheng had a remarkable gift for diagnosis and was compassionate enough to share his particular skills and training with me and two of my senior Medical Qigong students.

There are many schools of teaching for hand diagnosis. I have used several methods with great success in diagnosing and treating patients in both China and North America. The primary requirement for correct diagnosis is the use of the mind's intention. The imagination sets the foundation for the "energetic blueprint" of the patient's body (Figure 25.10). The method for employing this technique is described as follows.

1. Begin to scan the left palm and feel for Hot (Excess) or Cold (Deficient) areas. These areas will relate to specific locations of illness. The doctor begins to treat these regions of disease by using one of the following three methods:
    a. The doctor uses intention to tonify or purge the troubled areas inside his or her own left palm;
    b. The doctor uses his or her right hand (usually with the index and middle finger) to tonify or purge the energy from his or her left palm; or
    c. The doctor extends and treats the patient's body with his or her right hand while reading and diagnosing with the left palm.
2. Another variation of hand diagnosis is practiced by first having the patient breathe evenly into the Lower Dantian. Next, the patient focuses on his or her entire body, while placing the right palm (facing outward) towards the direction of the doctor. The patient is encouraged to relax while keeping his or her palm steady.

The doctor slowly moves the index or middle finger of his or her right hand towards the patient's right palm. While continuing to focus on the patient's energetic field (emitted from the patient's extended right palm), the

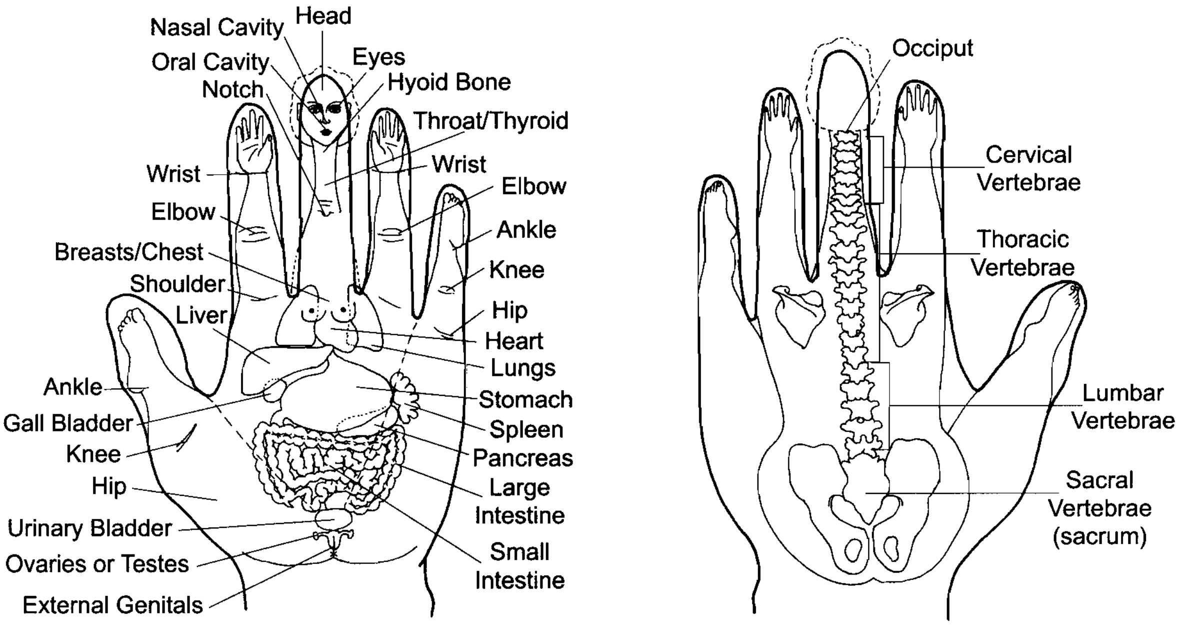
Nasal Cavity
Head
Oral Cavity
Eyes
Notch
Hyoid Bone
Throat/Thyroid
Wrist
Wrist
Elbow
Elbow
Ankle
Breasts/Chest
Knee
Shoulder
Hip
Liver
Heart
Lungs
Ankle
Stomach
Gall Bladder
Spleen
Knee
Pancreas
Hip
Large Intestine
Urinary Bladder
Ovaries or Testes
Small Intestine
External Genitals
Occiput
Cervical Vertebrae
Thoracic Vertebrae
Lumbar Vertebrae
Sacral Vertebrae (sacrum)

Figure 25.10. Front and Back View of the Hand/Body Association

Qigong doctor gently points to where he or she feels a particular energetic sensation (Figure 25.11). This special feeling or sensation that the doctor is receiving, represents the location of an energetic problem or disease. While scanning the patient's right palm, each different feeling or sensation represents a diversity of energetic manifestations related to the patient's disorder.

DIAGNOSIS BY OBSERVING AURA FIELDS

Some Medical Qigong doctors use aura readings to determine the physical, mental, emotional, energetic, and spiritual health of the patient. Aura colors will sometimes appear in the form of waves or other energy patterns (for example: evenly layered, blotchy, mixed together, etc.). Aura colors are in a state of constant flux, shifting their colors and patterns with each shift of the patient's moods and emotional thoughts. These colors merge and blend as they fluctuate, weaving in and around the patient's field of Wei Qi.

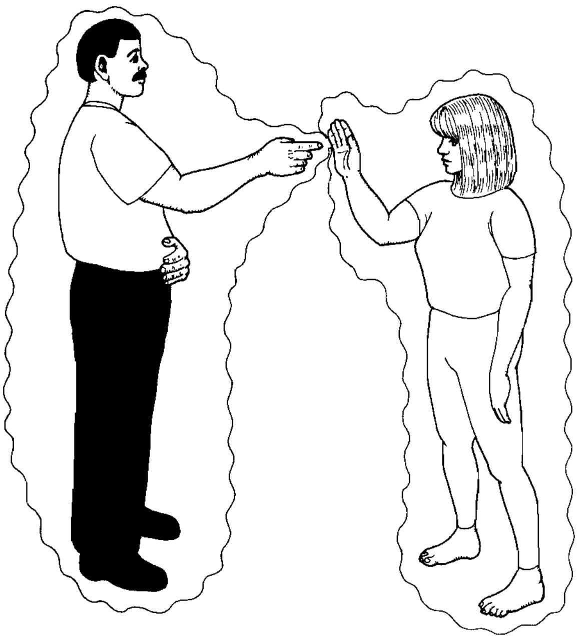

Figure 25.11. Variation of Finding the Patient's Disease Area by Diagnosing the Patient's Hand

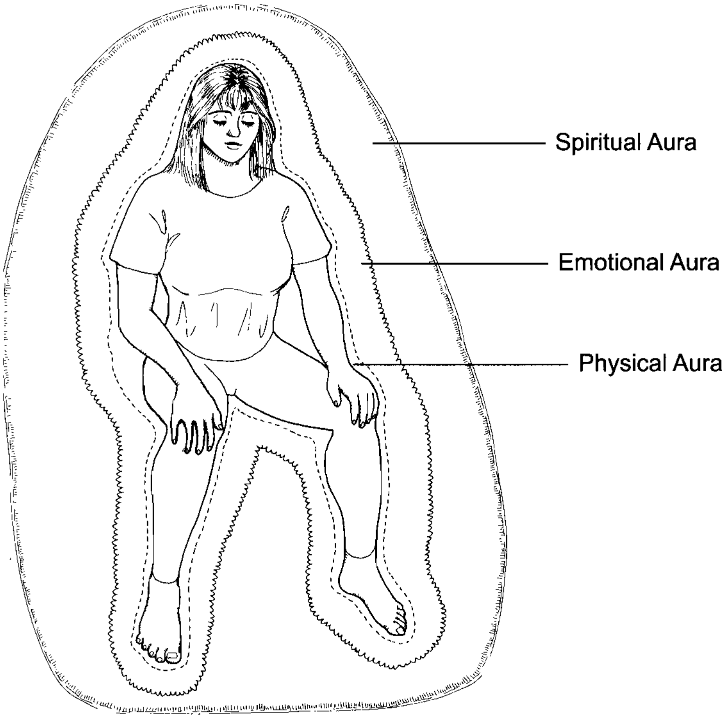

Figure 25.12. Diagnosis by Observing the Patient's Aura Fields.

To begin observing auras, ask the patient to relax his or her whole body, while sitting or standing in front of you. The patient should breathe evenly, to release any distractive thoughts, and then concentrate on the location of the specific disorder. The Qigong doctor focuses his or her attention on the space surrounding the patient's body, while simultaneously concentrating on his or her own Upper Dantian.

Focusing on the Upper Dantian to view the patient's body facilitates seeing the patient's aura field. The patient's aura field may be observed in the form of transparent or opaque Qi flow, or masses of different densities and colors. This may be used as a basis for determining the location and features of diseases (Figure 25.12). Each Yin and Yang organ will have its own color vibration.

When the energy of an organ is healthy, the Qi color is clean, clear and lucid. When an organ is diseased its Qi becomes dark gray and turbid. These colors reveal whether an organ is in an abnormal state and to what extent it has been affected.

COLOR VIBRATIONAL CLASSIFICATIONS

The variety of shifting colors can be interpreted according to their density and tone. The following chart displays the color frequency correlation in nanometers. All living organisms vibrate at a frequency between 300-2,000 nanometers. When analyzed, specific colors emanate consistent frequencies and wave forms. The slower frequencies register in the infrared light spectrum. The highest vibrational frequencies register in the ultraviolet light spectrum (for more information see Chapter 4).

Red	700 nm	Infrared
Orange	600 nm	
Yellow	550 nm	
Green	500 nm	
Blue	450 nm	
Violet	400 nm	Ultraviolet

SEEING AURAS AND EMOTIONAL ENERGY

Seeing auras is a visual phenomenon which occurs after many hours of regular Qigong training. This ability to see the external energy field is not limited to Medical Qigong training; many people are born with this special gift. The word "aura" literally means "breeze," manifesting as shimmering layers of luminous, colored energy that circulate around and penetrate into the physical body. These colors are part of the body's biophoton emissions (radiant light energy that emanates from all living systems). These colors are constantly in motion, reacting to thoughts, feelings, emotional patterns, and environmental influences.

Through the stimulation and physical dilation of the occipital lobe of the brain, the doctor begins to develop this unique perceptual ability. In the beginning stages the doctor may see energy coming off the patient's body like steam.

Later, brilliant, luminous colors become visible. The doctor is actually observing the infrared and ultraviolet radiation color patterns that the patient's body emits. The infrared spectrum is just below the average body's visual spectrum, while the ultraviolet is just above the average visual spectrum. By stimulating and dilation the occipital lobe, the visual spectrum expands to include the infrared frequency, allowing the doctor to observe from a much greater range of colors.

It has been my personal observation that students who possess a large bump on the back of the cranium, specifically between the channel points GV-16 and GV- 17, begin to see auras at a very early stage. These points are below and just above the external occipital protuberance and surround the internal branches off the occipital arteries, veins and nerves. When energy fills this area, the stimulation and dilation of these nerves and arteries causes the visual receptors of the brain to observe certain phenomena at a faster rate, which causes time to seem to slow down (Figure 25.13).

It is very important to note that these observations are all possible due to the occipital membranes being dilated. To maintain this dilation the doctor must remain calm and relaxed. Any tension brought about by stress only diminishes this ability. Auras are generally seen through the peripheral field of vision.

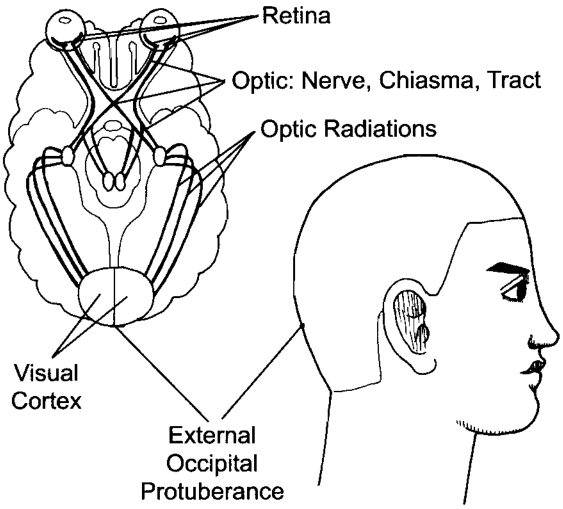

Figure 25.13. As the occipital lobe fills with Qi, the visual spectrum changes.

If the doctor begins to stare at any particular object, focused concentration will contract the occipital membranes, causing tension and pressure that inhibit the dilation process. The secret in maintaining this altered state of observation is to anchor the mind deep into the ground. By focusing the mind on a specific area, focus can be shifted to receiving and observing auras in addition to extending energy without distraction.

The following meditation can be used to stimulate the occipital membrane and enhance the perception of auras.

Seeing Aura Meditation

This meditation can be practiced either sitting or standing. The mind must be free of all distractions and tensions in order for the energy to permeate the occipital lobe. Draw an imaginary line from the top of the head (Baihui) down the spine, down the back of the legs, out the heels and deep into the Earth. This is your mental anchor line which is used to drain your mind of all thoughts (Figure 25.14).

By continuing to drain your thoughts deep into the Earth, you free your Yuan Shen to rely on intuitive perceptions that allow for this altered enhanced peripheral field of vision. This meditation should be practiced for a period of 15 to 30 minutes each day to establish a pattern of relaxation.

As you scan a patient, look to the sides of the patient's body, observing the silhouette of the outer shell of the patient's Wei Qi. It is important not to stare, but just observe the patient's external Wei Qi field (e.g., observing where the field is large, full, thin, or broken). Sometimes different lighting will affect the dilation of the occipital lobe, enhancing visual receptivity.

Next, allow the silhouette to become different colors. Usually, in the beginning, you can see only two or three different colors. With time and practice you will be able to differentiate more colors, with finer detail.

DIAGNOSIS BASED ON THE BODY'S EXTERNAL CHANNELS AND AURA COLORS

Each of the five emotions are related to both the Yin and Yang organs and their Yin and Yang natures. These five emotions are also related to five major colors, both inside and outside of the body's physical structure. By comprehending the colors and understanding their connection to the emotions of the body, the Qigong doctor can successfully analyze the abundance or depletion of energy in his or her patient's major organs. Not only will this reveal the patient's present psychological state of being, but will also expose which internal organs are weakened.

The color chart (Figure 25.15) is used to correlate the patient's energetic fields. These colors are similar, yet have their own unique variations from the physical body's tissue colors.

Think of these colors as extensions from the original organ pools. The mist of the organ pools flows into the body's aura fields, and can be observed in the patient's external energy fields. They are generally very vibrant colors and veil themselves around the first two layers of the Wei Qi. Observe their location around the body, whether they are static or pulsing. If a particular body organ malfunctions, or if any disease is present, then certain parts of the aura may become dim, murky, or produce a dark void.

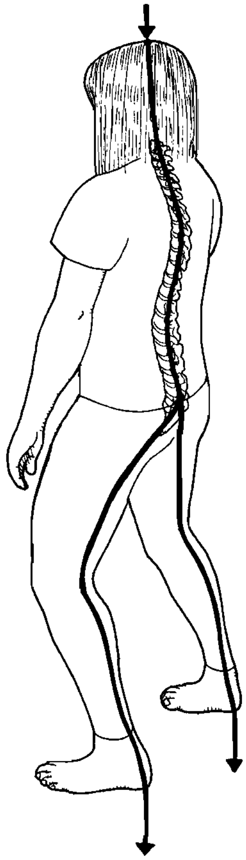

Figure 25.14. Root the Mind deep into the Earth.

It is important to remember that each patient has his or her own color schemes. When observing the color spectrums of the patient's energy field the most important characteristic is the clarity of the colors. Observe if the colors are sharp, murky, or too concentrated.

The Yang patient's colors will vary in tone and brilliance from the color spectrums of Yin patients. Each aura adjusts to match the patient's mood and temperament. People also often choose for their environment colors that reflect the characteristics of their aura and emotional spectrum (e.g., clothing, furniture, house, and room color, etc.).

When using color energy for healing, the Qigong doctor chooses which specific color to emit, e.g., using blue to cool the Qi or red to heat the Qi (see Chapter 4). Generally, when treating patients, the Qigong doctor will either wear all white (which is the combination of all colors, and is used to naturally project energy), or all black (which is the absence of all colors, and is used to naturally absorb energy), but never the color red (the color red is too intense or "Hot" for Qi extension).

DIAGNOSIS BASED ON INTENTION

The doctor's own developed inner vision can provide a diagnosis through observation of a mental image of the patient. The doctor begins by observing an image of the patient as a complete and whole energy system. Next, the doctor imagines the energetic form of the patient's internal organs (Figure 25.16).

If any part or organ is dark or gives a feeling of turbidness, this indicates the location of the disease. When using this type of energetic diagnosis, the sense of smell, the sense of taste, direct perspective, and remote sensing are also helpful diagnostic skills to enhance detection.

LONG DISTANCE MEDICAL QIGONG THERAPY

Long Distance Medical Qigong Therapy has been with us for centuries. One of the ministers of Huang Di, the Yellow Emperor and patriarch of Chinese medicine, was a shaman named Zhu You. Zhu You was a famous doctor who advocated exorcistic prayer over the use of acupuncture needles and herbs to treat illness. Some scholars believed that Zhu You practiced Qi emission healing at the same time that he prayed for his patients. So effective was this healing approach that The Yellow Emperor's Classics states that, in ancient times, most illnesses were treated according to the methods of Zhu You. In fact, at one time, professional "prayer healers" were once widespread throughout China.

There are several schools of thought on the modalities of long distance healing. Each school

Internal Organ	Aura Color
Lungs	Pure White
Large Intestines	Off-white
Stomach	Deep Yellow
Spleen	Orange Yellow
Heart	Red with Slight Blue Tint
Small Intestine	Pink
Bladder	Deep Blue
Kidneys	Light Flame Blue
Pericardium	Purple Red
Triple Burners	Orange Red
Gall Bladder	Yellow Green
Liver	Deep Green

Figure 25.15. The internal organs are each associated with a specific aura color.

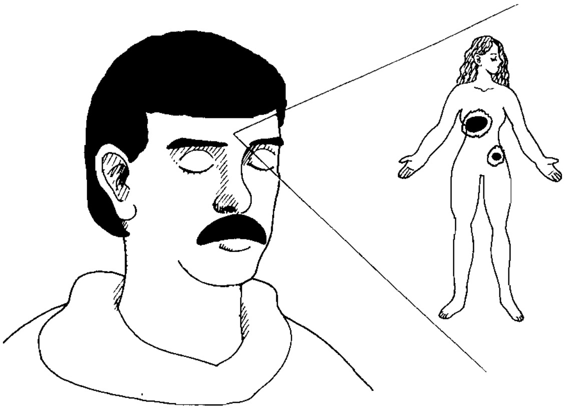

Figure 25.16. The Qigong doctor can diagnose the patient's body through energetic observation and intention.

has its own unique, yet extremely effective application of this esoteric skill.

Although I believe in the effectiveness of prayer, at one time I sincerely doubted the existence of long distance healing. I was thoroughly convinced that it was a matter of projected influential visual and auditory input, similar to hypnotherapy. I expected the doctor to talk to the patient and have him or her close the eyes to facilitate an emotional and mental healing over the telephone. The difficulty in this kind of healing is that if the patient does not trust you, his or her consciousness will not cooperate and will sabotage any healing energy extended in his or her direction.

As fate would have it, one of my Medical Qigong instructors called me one day. As we were conversing, my Kidneys began to heat up. This heat was quite intense, and I was extremely startled because no prior warning of this particular kind of treatment had been given. Yet, I could not dismiss the overwhelming searing heat that I felt in my lower back, Kidneys, and Mingmen area. This healing was done without my conscious consent or foreknowledge! My instructor laughed and explained that as she was talking to me, she was holding a pillow (which represented my body) and visualizing treating my Kidneys with her right hand.

At that point, my beliefs surrounding long distance Medical Qigong healing began to change radically. I began to inquire as to the qualifications needed to treat with that modality of therapy. I learned from my instructor that long distance energy transformation is similar to close range energy transformation. The only difference is that in long distance healing, the doctor's faith and psychic faculty has to be extremely developed and accurate.

To become proficient at long distance therapy, the Qigong doctor must extend his or her thought vibration into the atmosphere, which connects with the thought vibration of the distant patient. Both vibrations collect and blend, drawing together at a selected point on, or within, the patient, creating and reproducing the patient's energetic fields, organs, and tissues.

Zhu You's long distance healing through prayer may be compared to the effectiveness of prayer groups today. Larry Dossey, M.D., for example, documents one double-blind experiment involving nearly 400 patients admitted to the coronary care unit of San Francisco General Hospital for heart attacks or suspected heart attacks. These patients were evenly subdivided into two groups. Both groups received "state of the art" medical care; the only difference was that one group was prayed for by Protestant and Catholic prayer groups throughout the country. These prayer groups had only the patients' first names and sketches. The patients in the experimental group had no idea they were being prayed for; yet when the study was concluded, they experienced significantly fewer complications and deaths than did the non-prayed for control group.

SPACE/TIME AND ENERGY

To study long distance energetic healing, we need to look at a three-dimensional energetic pattern which includes Positive Space / Time, Negative Space / Time, and Energy.

1. Positive Space / Time Energy includes commonly accepted forms of energy (such as electromagnetism, nuclear energy and gravity), and all vibrations ranging from extremely slow to the speed of light. It includes lower vibrations (from inanimate objects) and the higher vibrations of thoughts.
2. Negative Space / Time Energy includes all vibrations which travel faster then the speed of light (super-luminal), and accounts for such phenomenon as telepathy, psychokinesis, precognition, and levitation. In negative spacetime, particles have a negative mass and are known as antimatter. When activated to superluminal speeds, these particles appear to dematerialize, and produce a levitation force which balances the gravitational effects of positive space / time. This also accounts for the Qigong doctor's ability to transmit energy through all known force shields (excluding mirrors, due to their ability to reflect light-energy).
3. Energy and matter form a continuum as does space / time. Energy is the medium between

the positive and negative forms of space. It is the medium between which matter and spirit interact with the space/time continuum, through which heat, light, sound, and resonant vibrations travel.

Matter is composed mostly of space with a minute amount of mass, bound in a particular pattern. This pattern is unique for every form of matter, including the human body. Particles of matter in the human body are constantly vibrating in intermediate stages of energy interaction. Consciousness (the qualitative and directional aspect of energy), acts as a medium for the Qigong doctor's intention to project into the time/space energy field. Consciousness functions as the energy source for resonant vibrations which create the physical world.

To perform long distance healing, the Qi in the Lower and Middle Dantians (the body's Fire and Water energy) must overflow and transform into Shen. The Shen is then united with the divine Qi and extended outside the body for long distance healing of the patient. This allows the energy to cross both spatial and temporal boundaries, thus enabling the Qigong doctor to heal patients in the absence of physical contact.

The first priority in long distance healing, is to establish a connection between all three Dantians through the absorption of divine Qi into the Taiji Pole; this process is called the "empowerment with divine Qi," or simply, the "hookup." This action surrounds the doctor's Wei Qi with Tian Qi (Heavenly Energy), which then is transformed into Tian Shen (Heavenly Spirit).

The doctor then visualizes a long stream of white light extending from the Yin Tang (Third Eye) point, which connects to the patient's body (Figure 25.17).

Once the doctor feels this connection, he or she may now begin the treatment. The physical manifestations of long range healing are the result of the Shen affecting the Qi, and the Qi affecting the body's Jing.

LONG DISTANCE SCANNING EXERCISE

Long distance scanning can be develop by practicing the following mediation exercise.

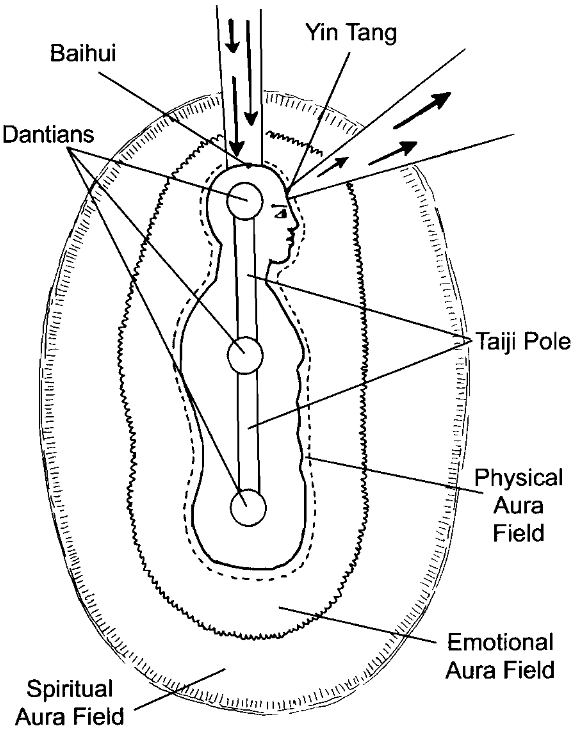

Figure 25.17. The Qigong doctor absorbs divine Qi through the Baihui area, filling the center Taiji Pole, then projects the divine energy out the body through the Yin Tang or Third Eye area.

- The patient sits in front of the doctor. After connecting with the divine, close your eyes and start to scan the patient's external and internal energetic body. Try to feel the patient's different shades of light and different color spectrums, as well as the patient's temperature, projections, resonant vibrational essence, and varying qualities of texture in the patient's electromagnetic fields.
- Next, open your eyes and scan the patient, comparing and contrasting any differences from your initial readings. Continue to practice this modality of distant scanning until your reading and diagnosing abilities become very accurate. Perfecting your long distance scanning abilities may require several months of training.

LONG DISTANCE THERAPY

To dredge the channels, tonify and regulate the patient's energetic fields using long distance energy transformation, the Qigong doctors sometimes practice the two following meditation exercises.

Long Distance Dispersing Meditation

The doctor begins by placing his or her patient in the next room. After the doctor connects with the divine, he or she imagines his or her body surrounded by a ball of brilliant white light. The doctor then extends a stream of this white light energy out from the top of his or her head and onto the top of the patient's head, enveloping the patient's entire body.

As this white light energy descends down the patient's body, the doctor visualizes the patient's diseased energy dispersing and melting away into the ground. The patient's pathogenic energy is extended into the center core of the Earth, where it is transformed by the Earth's center core fire and then recycled back into the patient as pure, refined Earth Qi.

Next, the doctor focuses on specific areas of the patient's body. The doctor visualizes any area which needs specific attention such as dark areas in the patient's body. The doctor imagines these areas becoming brighter until all the dark areas are dispersed. After energizing the patient's diseased area, it is important for the doctor to disconnect from the patient and withdraw his or her projected Shen (Spirit) back into his or her body.

Long Distance Purging Meditation

The doctor begins by sitting on the edge of a chair with both feet firmly on the ground. Next, the doctor brings both palms up in front of his or her body, at throat level (middle fingers pointing straight upwards), and imagines a ball of white light suspended in-between the palms.

The doctor now pulls the Heavenly Qi towards the center of the ball via the fingers. As the ball fills completely, the doctor imagines the Excess Qi beginning to overflow into the center of his or her palms (Laogong points), descending down the arms and spine, and collecting into his or her Kidneys and Lower Dantian area. The goal is to first collect and store Heavenly Qi in the doctor's Lower Dantian and Taiji Pole (Figure 25.18).

To facilitate long distance healing, the doctor imagines the patient in the center of the ball of light. The doctor focuses on the Excess energy no longer overflowing into the palms, but flowing into his or her patients body, cleansing, filling, and healing the disease (Figure 25.19).

FAITH PROJECTION AND MEDICAL QIGONG THERAPY

This technique is a combination of both psychosomatic and visceral healing. It entails speaking to the patient and establishing a curative relationship through belief modalities, similar to those of "faith healing" used in the West.

It is a form of mind projection; the stronger the Qigong doctor's faith and intention, the better the healing will be. It is important for the patient to have an extremely strong faith in the doctor and to feel secure in the healing environment. The stronger the patient's and doctor's faith, the better the healing.

One popular example of this faith projection ability comes from the famous Qigong Master Shen Chang. Master Shen, lecturing at a conference sponsored by the China Ministry of Broadcast, Film and Television (with hundreds of witnesses present), extended Qi into a woman patient (in her 50's) with a three inch tumor on her leg. As Master Shen began his transmission he began to shout "Gone!" Instantly the tumor started to shrink. After shouting "Gone!" two more times, the tumor had completely disappeared. To the astonishment of everyone, Master Shen then shouted "Grow!" and caused the tumor to reappear. After this amazing demonstration, he then removed the tumor, and completely healed the patient.

Perhaps the body manifests what the mind believes and the Heart feels. It then adjusts itself according to the belief structure and forms a new holding pattern (both physically and energetically). Usually, the doctor will say an instructional phrase four times to the patient, intensifying his

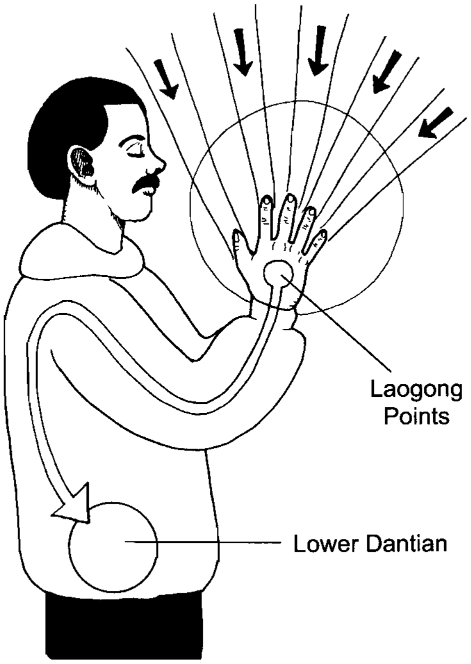

Figure 25.18. The Qigong doctor imagines creating a ball of divine healing white light, suspended between his or her palms.

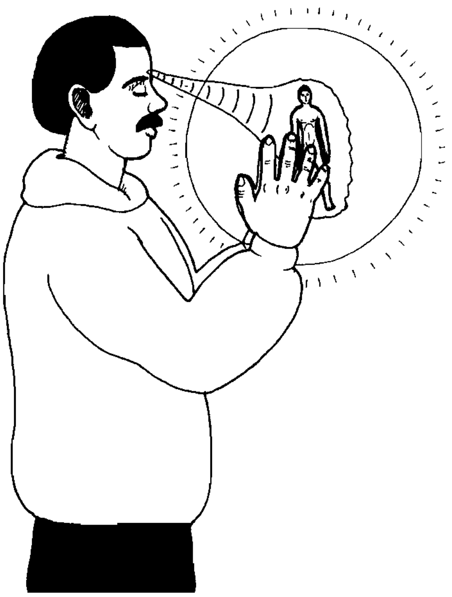

Figure 25.19. The Qigong doctor imagines that his or her patient is positioned in the center of the ball of divine healing light.

or her intention and instruction each time. The doctor's intention begins to reprogram the holding patterns of the tissues by shocking the system into a new and healthier pattern; a kind of "Be Healed Now!" approach. It is important to know that the energetic field surrounding the patient is strongly affected by either positive emotions and support, or negative emotions, suspicion and disbelief.

To illustrate this fact, I would like to recall an incident that took place in a Santa Cruz bible college in the late 1980's. One morning a colleague was lecturing in an auditorium full of 250 young seminary students. Suddenly, the back door of the auditorium swung open, and a young woman carrying a small baby came walking in. She was crying, and asked for help. She told my colleague that "God" had spoken to her in a dream and told her that if she came to this particular bible college and asked for help, that he would heal her child of a serious disease.

My colleague was both startled and shocked. Not knowing what to do, or even how to handle the situation, he quickly turned the meeting over to a local spiritual elder who was monitoring the lecture. This elder, seasoned in these types of spiritual phenomena, again asked the woman what the problem was. After hearing her story a second time, he informed her that while "God" had not spoken to him directly about the situation, he would be obedient to whatever "God" wished.

Next he looked up to the 250 young seminary students sitting in the auditorium and stated, "If anyone here does not believe that God can and will heal this child instantly, please leave now." All but three seminary students left the auditorium. The elder called to the three remaining students to join him with my colleague in praying for the baby. The child was instantly healed (a follow-up visit confirmed this fact), and the mother left, smiling. My colleague, perplexed by the whole event, turned to the elder and inquired as to how he knew what to do. The elder responded, "Even Jesus, who was the greatest healer of all time, couldn't heal in his own home town because of everyone's disbelief. Here in this auditorium it would have been five of us who had the faith in healing, versus 247 who were skeptical and full of doubt. It is better to have a few with great unwavering faith than to have many with little faith at all."

AUXILIARY MEDICAL QIGONG HEALING MODALITIES

The following are auxiliary Qigong modalities used by various Qigong doctors in hospitals and clinics, as well as their personal practices throughout China.

One of the methods used for these Medical Qigong therapies is the projection of Qigong Messages into specific material objects (i.e., a liquid, food, a scroll of calligraphy, a talisman, a cloth, a necklace, etc.). Once the object contains the Qigong doctor's Message it will have a powerful effect on the patient's body. This is a form of "transmission of intention" which is a form a Qi and Shen projection. Qi and Shen projection transcend time and space, lasting an indefinite amount of time depending upon the belief of the Qigong doctor and patient.

Certain Qigong doctors can transfer their energy or "thought message" through works of art. I experienced this technique quite accidently while lecturing at the Third World Conference on Medical Qigong in Beijing, China in 1996. I was in the back of the lecture hall sketching while one speaker was sharing his dissertation. Another Qigong Master seated in front of me noticed my artwork and asked to see it. Curious as to what he would do, I gave him my notebook. He placed his hand above my drawing and began projecting Qi into the picture. Immediately energy flooded my being from head to toe. I was literally vibrating all over, being filled with light and sound resonation (that felt similar to a strong electrical current). Since my original piece of art contained a direct connection to my physical energy, the Qigong doctor was able to extend energy into my body, by simply projecting his Qi into my artwork.

The doctor can transmit healing energy through any object which has had contact with or was created by the patient. This technique is possible because creative artwork, or any object touched or created by hand, is imprinted with the person's energy. This energy contains a unique blend of the individual's encoded energy patterns. This information regarding the individual can be retrieved through psychometry (the act of sensing the thoughts, images and so on, with which the object has been imprinted). A connection to the person who created or held the object can thus be established through the Wuji. This object can then serve as a focusing point to send healing to that person, through the doctor's intent.

Energy, in the form of waves, is expressed in the abstract form of thoughts, accessible to the vibrational medium of space and time. This allows for the energetic transformation of the individual's emotional patterning.

USING MEDICAL QIGONG TO HEAL BY PROJECTING CHINESE CHARACTERS

There are other Qigong doctors who draw specific Chinese characters in the air and send them encapsulated in a bright white light ball of energy into their patient's body to initiate healing (Figure 25.20). This projected energy is created through the doctor's intention and is absorbed into the patient's body, through the patient's receptive attitude and desire to heal.

USING MEDICAL QIGONG WITH SUPPLEMENTARY APPARATUSES

There are certain Qigong doctors who will use specific objects ranging from such supplementary equipment as rubber acupuncture dolls to pillows, in order to perform long distance healing. To utilize these objects for long distance healing, these Qigong doctors place the object, a doll for example, on the table in front of them and begin to draw their patient's Essence (from many miles or even across other countries) into the doll. The doctors then treats the patient by draining off the pathogenic excess from the doll (Figure 25.21). Once the doctors feel that the patient has been sufficiently cleansed and the disease dispersed, they begin to tonify the patient's weak organs. While

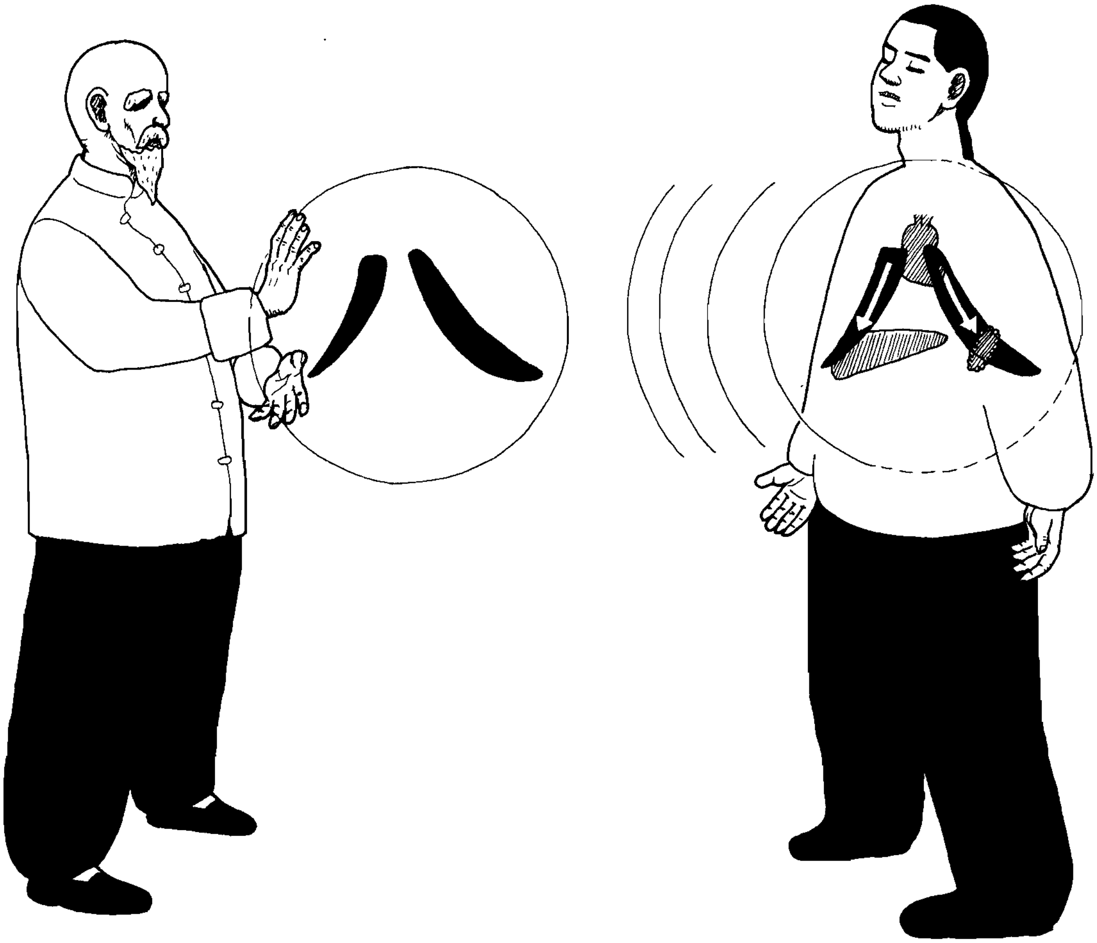

Figure 25.20. Another skill utilized by Qigong doctors, is to create different colored balls of energy to encapsulate specific Chinese characters (or Messages) within them, which are then projected into the patient's body as a form of treatment.

the Qigong doctors work on a patient, you can actually feel hot and cold spots emitting from the rubber acupuncture doll. Several of my students have been able to successfully implement this modality of training and find it fascinating.

USING MEDICAL QIGONG TO ENERGIZE HEALING CLOTHS

Some Qigong doctors extend Qi into pieces of paper or cloth, to be placed on the patient's body to invoke a healing. This is similar to the "prayer

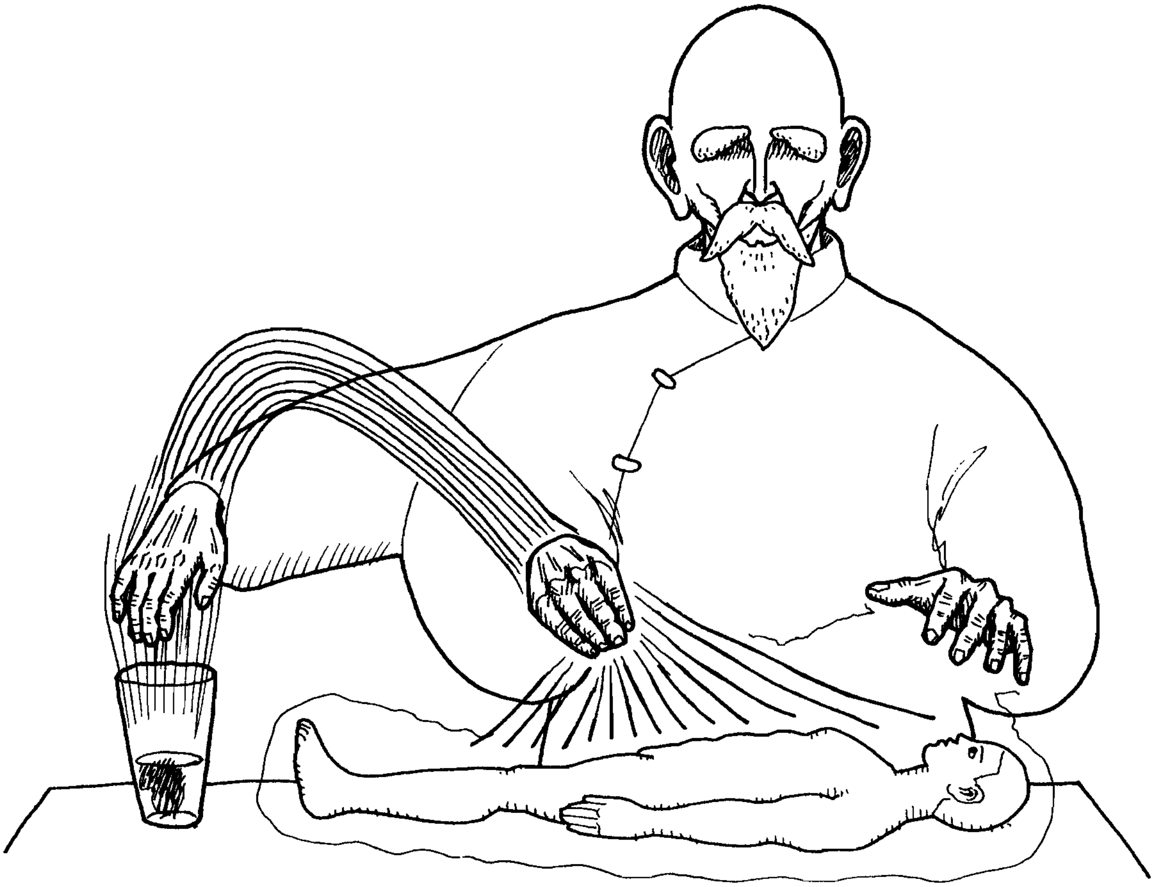

Figure 25.21. Some Qigong doctors use a rubber acupuncture doll (man or woman) as an energetic medium, while treating patients at a distance. The Qigong doctor is removing pathogenic Qi and throwing it into a glass of salt water (used to neutralize the Turbid Qi).

cloths" and other techniques used today in the Pentecostal Churches. The energy stored within the cloth or paper, will last indefinitely, however, the most important component of this type of healing is the patient's belief structure (Figure 25.22-23).

USING MEDICAL QIGONG WITH CRYSTALS

The energy in the human body produces a polarity similar to that present in crystals. Because crystals emit a very strong white light energy, they can be used for energetic surgery, allowing the

Qigong doctor to cut into the patient's energetic field like a laser beam. Crystals can additionally be used for scooping energetic toxins, cleansing and purging the patient's body, and for cutting loose pathogenic Qi within the patient's tissues and external fields of Qi. Crystals can also be used for tonification, by enhancing the doctor's energetic extension and projection abilities.

USING MEDICAL QIGONG TO ENERGIZE LIQUID COMPONENTS

Other Qigong doctors extend their energy into herbal teas, wines, and I.V. fluids, as well as simple components such as water, to energize and utilize these liquids for healing. These doctors send the energized liquid substance to the patient, who ingests the medicine to tonify deficiencies. Water is an excellent repository for electromagnetic energy, and absorbs vibration better then any other substance (Figure 25.24).

USING MEDICAL QIGONG TO ENERGIZE FOOD

There is an ancient Chinese proverb which poses the question, "Is food medicine or medicine food?" Because a basic principle of Traditional Chinese Medicine states that anything ingested is transformed into energy (Gu Qi) by the body, some Qigong doctors extend their energy into the patient's food before it is ingested. This belief is not isolated to the Orient, but is also part of the Western culture, hence the practice of praying and blessing over food before eating it.

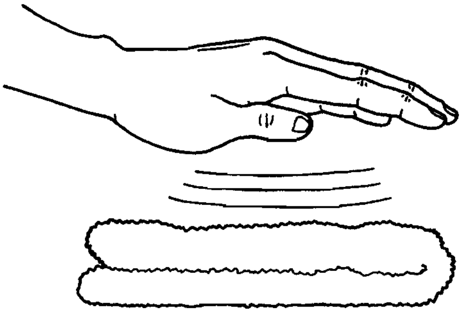

Figure 25.22. Emitting Qi into Cloth

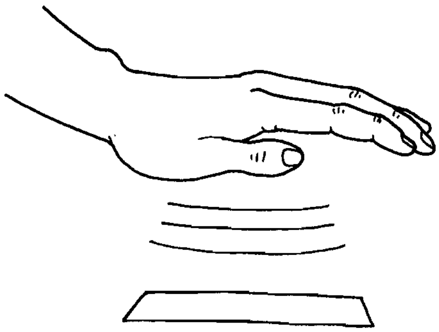

Figure 25.23. Emitting Qi into Paper

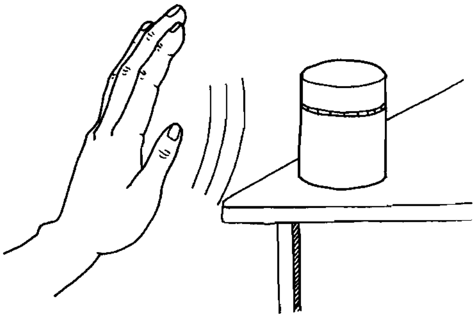

Figure 25.24. Emitting Qi into Liquid

SECTION VI

ESTABLISHING THE MEDICAL QIGONG CLINIC

CHAPTER 26
TREATING THE PATIENT

SETTING A SPACE/TIME FOUNDATION

Before treating patients, and to establish a sacred healing space, the Qigong doctor must first bring his or her true self into what is called the center of space and the center of time before performing the divine hook-up.

This establishes a foundation for treating patients, and allows the doctor to fuse with the patient, while accessing the energies of the divine, as well as the universal and environmental energies. Once this is achieved the doctor can allow him or herself to act as an observer, viewing the patient from a non-judgemental position.

FINDING THE "TRUE SELF"

The Qigong doctor's "true self" is defined as his or her pure innate spiritual existence, void of all ego masks and personalities. The "true self" is spiritually connected to, and rooted in, the energy of the divine. It is the part of the soul that unconditionally accepts self and others, and is more in the role of a compassionate observer than a judge. The "true self" allows the doctor to observe the patient's condition (spiritual, emotional, mental, energetic, and physical), void of any personal emotional biases. The "true self" is a state where unconditional love and acceptance is fostered and supported by the divine.

To discover the "true self", the Qigong doctor is trained through special Shengong (Spirit skill) meditations. One such meditation asks the question "Who are You?" This is repeated over and over until the doctor realizes and experiences his or her "true self". Who you are is not determined through expressing a correct answer, but by actually experiencing the spiritual quest. It is through the realization of this innate truth, which is inherent within one's own enlightenment, that the doctor is freed from the ego (known as the "Death of the Ego Fear"). Once the doctor realizes who he or she truly is, the answer is reflected from the light within the doctor's eyes and not through any specific mental concepts. There is a saying from Lao Zi, written in the *Dao De Jing* (The Way and Its Power) that explains this phenomenon: "The Dao that can be named is not the true Dao."

FINDING THE CENTER OF SPACE

Before treating a patient, the Qigong doctor must also place his or her body, mind, emotions, energy, and spiritual essence into a place of infinite existence called the center of space. The center of space is actually a spiritual and energetic concept, wherein the doctor imagines him or herself as being the actual center of the universe. This concept and image maintains that all energy, light, and energetic fields extend from the middle of the doctor's center core, expanding outward into infinite space (Wuji). This energy continues to extend and envelop the entire universe and all the energetic dimensions beyond.

To access the center of space the doctor must place his or her "true self" into the Taiji Pole, connecting the core self with the divine. It is from the orientation of the core self that the Qigong doctor accesses the deeper, hidden, spiritual, and energetic fields of the patient's body.

FINDING THE CENTER OF TIME

The final image that the Qigong doctor visualizes before the divine hook-up is to place himself or herself into the space known as the center of time. The center of time is the place of "Now." It is a dimension where infinite future is forever flowing into the doctor's center core, while the infinite past is forever rushing away.

To grasp the center of time concept, think of time as a linear progression of light and energy. The concept of future time is rushing directly into the front of your body. Within a split second, the Yang changes into Yin and the future transforms into the past, and continues rushing out the back of your body. You must find the still point, inside of your body, where Yang transforms into Yin, and the future transforms into the past - this is the center of time. From the center of time the Qigong doctor brings the patient into the present moment. This process enables the doctor to access the hidden past memories trapped deep within the patient's tissues.

ESTABLISHING THE MEDICAL QIGONG TREATMENT ROOM

Before receiving patients it is important to establish an environment that is conducive to healing. The density and quality of energy in the healing room is a very critical factor. Filling both the treatment room and clinic with the energy that flows from the Heavens to the Earth sets up an energetic vortex. Through the creation of this energetic vortex the patient's pathogenic factors will automatically flow into the ground once they are dispersed from the patient's body. The more energy that the healing room contains, the more successful the treatment will be. It is important to cleanse the healing room before and after each patient. This cleaning is achieved through the doctor's divine hook-up.

When establishing a treatment room it is also important to have adequate lighting and clean air ventilation. Avoid fluorescent lighting; if this is impossible, then use a full spectrum light.

It is important to also keep the patient's living quarters and bed energetically cleansed, as pathogens have a habit of collecting in areas of sickness and disease. The patient is encouraged to burn salt with a little alcohol to cleanse the treatment room (at least once a day) and fill the room with fresh air and light. Often neglected is the living space of the patient. Energetically cleaning the entire house or apartment often, during and after there has been an illness, allows the toxic energy to be purified from the patient's home.

One example of how energy stagnates is observed in hospital wards. Each ward has an energetic "feeling," associated with the different types of diseases or conditions treated there. The cancer ward, for example, feels entirely different from the maternity ward, emergency room, etc.

The treatment room must also be kept uncluttered, as energy stagnates in cluttered environments. Also, after each treatment, always change the bedding and energetically clean the clinic or treatment room.

WHEN TO AERATE, ILLUMINATE, OR SMUDGE

It is important to daily cleanse the energetic field of the treatment room. This facilities a fresh uncontaminated environment, allowing the next patient to feel safe and protected from any previous contaminants which may have been left from a prior patient. The three most common ways to cleanse the clinic before and after each patient (besides changing the table coverings), are:

1. Aerate by ventilating the clinic, making sure that the air circulation is fresh and clean (free from pathogens, Toxic Qi, dust and pollens, foul orders, smoke, etc.).
2. Illuminate by exposing the clinic to bright sunlight, or divine healing light (through prayer).
3. Smudging is performed by burning fragrant herbs in the form of certain incense (such as, sandalwood, myrrh, frankincense, or sagebrush, etc.). Remember to aerate after smudging.

The Qigong doctor may wish to incorporate either one or several modalities of cleansing. Each particular technique offers its own advantages and disadvantages, depending on the size and location of the clinic and its surrounding environment.

ESTABLISHING A RAPPORT WITH THE PATIENT

It is important to establish a foundation of trust between the doctor and patient before the treatment begins. This trust is enhanced by: Talking with the Patient to Establish Treatment Goals, Explaining and Reaching Agreements on the Treatment Modalities to be Used, Honoring the Patient's Spiritual Belief System, Receiving the Patient's Consent Before Treating, Matching the

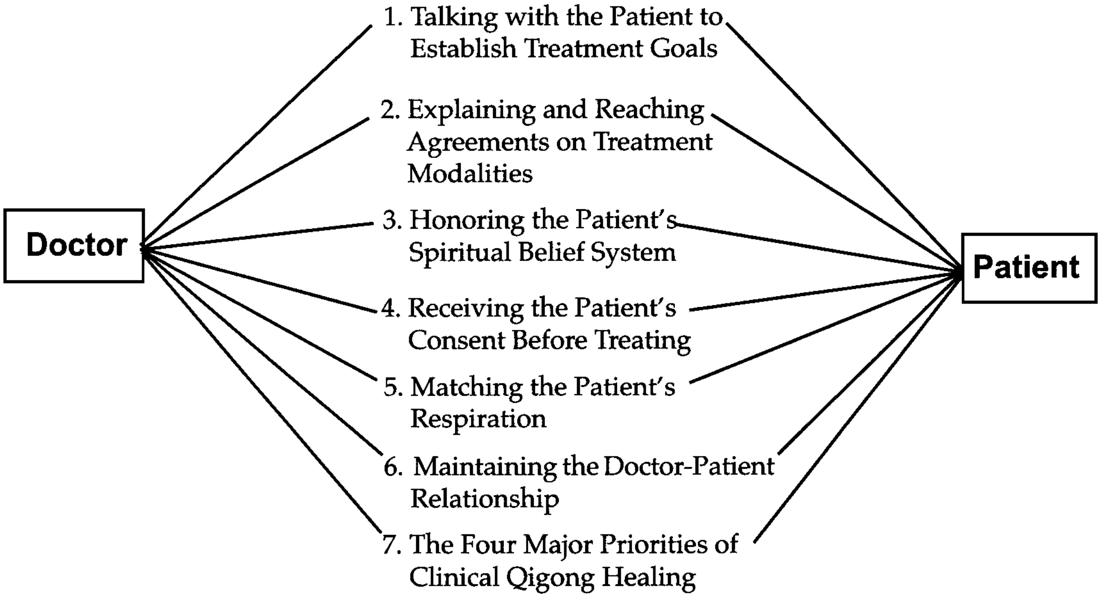
Doctor
1. Talking with the Patient to Establish Treatment Goals
2. Explaining and Reaching Agreements on Treatment Modalities
3. Honoring the Patient's Spiritual Belief System
4. Receiving the Patient's Consent Before Treating
5. Matching the Patient's Respiration
6. Maintaining the Doctor-Patient Relationship
7. The Four Major Priorities of Clinical Qigong Healing
Patient

Figure 26.1. Establishing a rapport with the patient begins with trust.

Patient's Respiration, Maintaining the Doctor-Patient Relationship, and The Four Major Priorities of Clinical Qigong Healing (Figure 26.1).

Talking with the Patient to Establish Treatment Goals

After completing the diagnosis, establishing the treatment principle, and determining the treatment strategy, the Qigong doctor discusses with the patient the treatment goal (i.e., which internal organs, channels, and tissue areas are affecting the primary cause of the patient's disease, and how to address it). The treatment areas are specifically chosen to facilitate a rapid affect on the functioning process of the patient's disease. The doctor explains the intended goal of using the purging, tonifying, or regulating techniques, and how they will affect the patient's body.

Explaining and Reaching Agreements on the Treatment Modalities to be Used

The Qigong doctor explains, in clear, comprehensive terminology, what he of she has determined as to the best way of approaching and rectifying the patient's condition (e.g., having the patient in a lying, sitting or standing posture), and what treatment modalities the doctor will be using (i.e., through Qi Emission, Qigong Massage, Sound Therapy, Light Therapy, etc.).

The more the patient understands the doctor's approach in addressing the condition, the greater the likelihood of the patient relaxing during the treatment. This explanation also assists the doctor and patient in developing trust, which will encourage the patient when practicing the homework prescriptions, or in receiving suggestions on life-style modifications.

Honoring the Patient's Spiritual Belief System

When treating patients, it is important that the Qigong doctor not become distracted by the illusion of solid tissue matter, remembering that we are spirit, and that healing begins in this fifth dimension. The primary belief in Medical Qigong therapy is that tangible and intangible energy, through faith, will merge at the highest level of cultivation. Prayer is an invocation that extends the mind, emotion, and spirit to the divine, while meditation is a state of mental, emotional, and spiritual receptivity for divine inspiration, guidance, and manifestation. Both actions (prayer and meditation) connect the doctor and patient to the divine to bring patients back into awareness of their body and to initiate divine healing.

In China, Qigong doctors have a saying, "train your virtue and morality to increase your energy, allow your Spirit to lead your Qi and the Qi will flow pure." Meaning, the stronger the individual's virtue has been developed, the more powerful his or her spiritual intention will flow, and Qi move.

The healing modality choosen by the doctor must be congruent with the patient's spiritual belief structure. The patient must be connected to his or her own unique interpretation of universal consciousness (God) in order for divine healing to take place.

When the patient disbelieves in a Higher Power -- be it God or a universal consciousness -- the doctor can still help the patient heal, providing the patient has established sufficient trust and hope in the doctor. If the patient feels understood and listened to by the doctor, this can empower the patient to heal, regardless of his or her own conscious belief system.

If the doctor, on the other hand, does not believe that a true healing can take place unless the patient believes in a Higher Power or God - then no healing will take place, for the doctor will lack confidence, hope, and trust. The therapeutic bond between the doctor and patient will be broken and replaced by a battle of wills.

For patients who lack faith, applying a hands on Qigong technique may be more effective. Gradually, over the course of time, the doctor may incorporate other healing modalities as greater trust is built.

The Qigong doctor should always begin treatment in harmony with the patient's spiritual belief system (that may change in time), through touch, verbal conversation and unconditional love (empathy). The doctor should also communicate mind to mind, for this will facilitate trust at the subconscious level.

RECEIVING THE PATIENT'S CONSENT BEFORE TREATING

The treatment modalities should be decided together by the doctor and the patient. It is important for the Qigong doctor to provide the patient with enough information about the type of therapy he or she will receive, so that informed consent can be given for the treatment. This empowers the patient and allows relaxation and trust to occur.

MATCHING THE PATIENT'S RESPIRATION

Another technique to relax the patient is called "matching and pacing." The Qigong doctor matches the patient's breath and breathing patterns, and gradually shifts his or her own breathing to a slower and deeper pace. It is important to note that the patient's physiology affects his or her energy, which in turn affects his or her emotions and respiration. This technique of matching the patient's and doctor's breath and leading it into a state of quiescence is very often used by Medical Qigong doctors to relax their patients. It not only connects the doctor with the patient on an energetic and emotional level, but also provides a strong spiritual connection.

Note: Do not match an asthmatic patient's breathing, as this is disruptive to the flow of Qi within your own body.

MAINTAINING THE DOCTOR-PATIENT RELATIONSHIP

Whenever the doctor begins to judge or formulate negative opinions of a patient, he or she is no longer operating in the divine "present" state of spiritual intuition, but has shifted into the "past" (as all judgements are based on past experiences and emotional encounters). The doctor is then in danger of absorbing the discharged energetic pathogens from the patient, or projecting subconscious biased opinions onto the patient. Whenever the doctor relates, relives, or orients from his or her own personal past (the feelings can be either positive or negative), the doctor is at high risk of absorbing the patient's energetic pathogens.

When the patients come to the doctor, they are giving their consent for the doctor to learn about personal things which the patient has often withheld from their mate, family, or even themselves. They are further giving the doctor permission to utilize their combined energies to affect changes within their physical, mental, emotional, energetic, and spiritual fields and body.

Each treatment will be different for the patient, as each experience will vary according to the patient's present state of mind and emotional spirit. The patient's willingness to allow the Qigong doctor's emitted energy to penetrate deeper into his or her body with each session will depend upon the level of trust established, as well as any outside influences that may have a positive or negative effect on the patient's ability to be open and vulnerable. Patients who experience ridicule or criticism from family and friends for their choice of a non-Western form of treatment, may temporarily shut down their receptivity in self-defence against others. At times, patients may also close-off as a result of an event that has a strong impact on their emotional and spiritual well-being, such as a death of a friend, a physical trauma, or other upsetting news.

The patient always has the final control over the doctor's emitted energy, and can choose to either absorb and utilize the energy, or disperse and sabotage the doctor's treatment. It is for this very reason that the doctor-patient trust factor, as well as the goal orientation, must be securely established before the treatment begins.

Be sure the patient's arms and legs are not crossed, as this can impede, or "short-circuit," any energetic flow, causing Qi obstructions. The patient should close his or her eyes, but should not be allowed to internally guide or direct the Qi unless specifically instructed to do so by the doctor. A patient's unguided imagery can cause the body's energy to increase, alter, or obstruct the flow of the doctor's energy, which may cause Qi deviations or aggravate an existing condition. Usually the patient is given the image of dissolving his or her tissues into space, which encourages relaxation and freedom from distractions. If the patient is lying down, he or she is given certain visualizations, such as floating on a sea of warm water (or basking in the sun on the beach or shore of a lake). These images relax the patient and enable the Qigong doctor to disperse the patient's armoring. Once the patient's "street armor" is down, the doctor will have easier access into the patient's tissues.

After the treatment, the doctor and patient evaluate each session, discussing any and all feelings, observations, transitions, and sensations that both the doctor and patient experienced. Homework, in the form of Medical Qigong prescriptions are then generally given, in addition to follow-up treatments (depending on the patient's case and condition).

THE FOUR MAJOR PRIORITIES OF CLINICAL QIGONG HEALING

According to China's prominent Qigong Grand Master Pang Ming, clinical Qigong healing is based on four basic priorities. In establishing a clinical environment it is important to inform the patient about the structures of these four priorities: A Strong Faith, Group Healing, Medical Qigong Therapy, and Medical Qigong Prescriptions and Exercises.

1. A strong faith that Qi or life-force energy can heal the patient's specific ailment is built up by the patient listening to testimonials. These testimonials are given by other patients who have recovered from the same type of disease, as well as those patients who are experiencing improvement through Medical Qigong therapy.

   Qigong deals directly with the psychological aspects of diseases. It is well known that two side effects of cancer are depression and fear. A very popular theory maintains that depression (being the suppression of anger) can cause cancer. Even before the cancer patients begin their treatment program, their confidence is bolstered by the positive experiences of more advanced students of Qigong, who at one time also suffered from these particular kinds of cancer. In addition, as part of the Qigong exercises, the new patient's mind is focused on pleasant images or memories. All of this creates a positive mental attitude in the patient.

Sometimes patients come to the doctor for healing, yet at a deeper level they feel that they need or deserve the illness. Sometimes the secondary gain of illness (for example, to keep the wife or husband underfoot) outweighs the benefits of healing completely. If a patient is subconsciously unwilling to be healed, his or her attitude or disbelief will sabotage all clinical modalities of healing (both Western and Eastern). You cannot make someone better against his or her own will. Illness, like health, is often a matter of choice.

2. Before a group of patients begin Qigong, the doctor verbally synchronizes the thinking of the group by absorbing Heavenly Qi (the divine hook-up), and bringing it down into the healing energy field which surrounds everyone including the doctor. The healing effect is enhanced because the group is acting as one unit, with one faith.
3. The doctors initiate the actual Qigong healing by bringing divine healing energy to each individual patient.
4. The patients are taught how to easily follow Medical Qigong exercises and meditations, practicing them over and over again.

When a patient enters a Qigong hospital in China, they are diagnosed by a doctor and then assigned a specific Qigong class for a twenty-four day treatment period. The patient spends eight hours a day practicing Qigong without television, newspaper, or telephone. This is required to help the patient avoid regression into a previously established belief structure or subconscious pattern. Those who are able, practice Qigong from a standing posture. Other patients with limited mobility practice Qigong sitting or while lying in bed. After the twenty-four day treatment program is over, the patient is again diagnosed by a Qigong doctor. The diagnoses is classified into one of four categories for statistical purposes.

1. The patient has been cured: symptoms have disappeared and appropriate instruments (e.g., EKG, ultrasound, etc.) register normal tissue.
2. The treatment has been very effective: symptoms have almost disappeared and allopathic tests indicate great improvement.
3. The treatment has been effective, in that the patient is now able to eat, sleep, and feel good.
4. The treatment has been noneffective: there has been no change, or the condition has worsened.

PREPARING THE SELF TO TREAT THE PATIENT

This section prepares the Medical Qigong doctor to treat patients. The "preparation" sets the energetic ambiance for the clinical treatment and secures the foundation for dispersing pathogens. The preparation includes: The One through Ten Meditation, Creating a Sacred Healing Space, The Divine Hook-Up, Rooting to Prevent Energetic Shock, and The Energetic Circle Drawing Pattern (Figure 26.2).

THE ONE THROUGH TEN MEDITATION

The following is called the One through Ten Meditation and is used to prepare the Qigong doctor for clinical work. It allows the Qigong doctor to relax, sink, and root the mind. This meditation allows the doctor to extend his or her energetic field before connecting with the divine. This meditation is used to create a sacred healing space and is taught in a rhyming one through ten counting format, described as follows.

1. One Is Fun, therefore take pleasure in your work. Imagine a smile melting down the front of your body like warm oil, melting all stress and tension into the Earth. A second smile melts down the back of your body, dissolving all stress and tension down into the Earth. A third and final smile melts down the center of your head, relaxing all thoughts, judgments and feelings, and flows down your shoulders and arms and out your hands. The melting continues to flow down the center of your torso, like warm oil, melting all emotional blocks and tension, then flows down your legs and into the ground. At this point your mind and body should feel content, relaxed, and peaceful.
2. Two Is Shoe, thus imagine your feet melting into the Earth, like ice melting on a hot side-

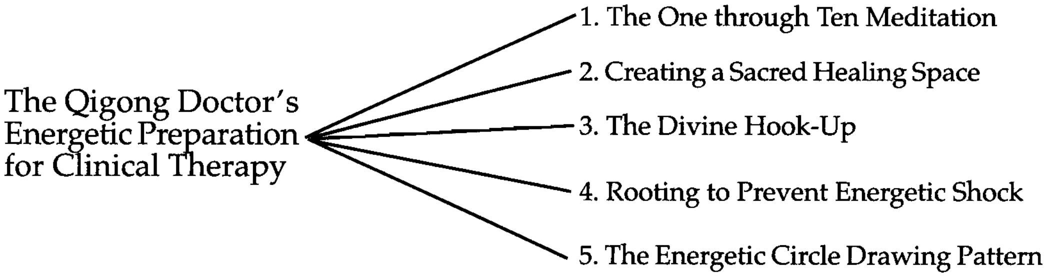
The Qigong Doctor's Energetic Preparation for Clinical Therapy
1. The One through Ten Meditation
2. Creating a Sacred Healing Space
3. The Divine Hook-Up
4. Rooting to Prevent Energetic Shock
5. The Energetic Circle Drawing Pattern

Figure 26.2. The Qigong Doctor's Preparation for Treating the Patient

walk. The energy of the feet fuses with the Earth, melting in five directions (forward, backward, right, left, and down). This connects your mind and body with Earth energy.

3. Three Is Tree, therefore feel yourself extending tree-roots twice your body's height deep into the ground. The roots expand in five directions (forward, backward, right, left, and down), securing your rooting and connection with the Earth energy.
4. Four Is Core, thus imagine your roots extending deep into the core of the planet. The Earth energy flows into these roots, ascending up the legs, spine, over your head, down the chest, and enters your Lower Dantian. This action allows you to draw Earth energy into your body and circulate it through the Microcosmic Orbit. A solid connection to the Earth roots your physical, mental, emotional, and spiritual energies, and fills up your Lower Dantian with Earth energy.
5. Five Is Alive, therefore feel the Earth energy building up within your Lower Dantian. The energy increases to the point that your Dantian can no longer contain its power. It rushes up the center of your chest like a mighty river, dividing into two streams which flow out each palm into the earth. This action releases and emits the energy stored within your Lower Dantian.
6. Six Is Thick, thus imagine the room or environment where you are standing being filled with energy. This energy becomes so thick, it is as if you are submerged under water. This fuses the environmental energy with your energy, and the Earth's energy into one dynamic field of energy.
7. Seven Is Heaven, therefore allow your Baihui point to open up and draw in the divine healing white light into your Upper Dantian. Let the divine light fill your head and the entire Upper Dantian area till it floods and shines down from the Upper Dantian into your chest, filling up the entire Middle Dantian. Finally, the light shines down from the Middle Dantian filling up the lower abdominal area and the entire Lower Dantian. This light connects you with the Heavenly healing white light energy from the divine, filling up all three Dantians and your Taiji Pole. Thus Heavenly and Earthly energies fuse within your body.
8. Eight Is Open the Gates, therefore imagine opening every pore on your surface tissues, and begin to draw the environmental energy into your Taiji Pole upon inhalation. Upon exhalation imagine your center core vibrating and glowing like a neon light. This visualization energizes your center core, harmonizes your energy and breath, and prepares you for Qi and Shen (energy and light) projection.
9. Nine Is Shine, therefore imagine that your center core becomes completely full of energy and overflows by shining (from the Taiji Pole outward) through your pores until it fills up the entire room, like a bright phosphorus flare. This visualization fuses Heaven Qi, Earth Qi, and Man's Qi with the external clinical environment.

10. Ten Is Begin, with all things now in order, you are physically, mentally, emotionally, energetically, and spiritually in harmony, with your body, the Heavenly energy, and Earthly energy. You can now begin treating patients.

CREATING A SACRED HEALING SPACE

After completing the preparation, the Qigong doctor creates a sacred healing space in which to treat patients. Creating the sacred healing space allows the doctor and patients to interact in a safe environment without outside distractions.

The Qigong clinic should be a place of complete sanctuary, a consecrated place of refuge, where the patient feels safe, warm, and cared for (spiritually, emotionally, mentally, energetically, and physically). The patient should experience the peaceful ambiance of the treatment room as being enveloped in unconditional love and security.

To create the sacred healing space, the Qigong doctor performs the divine hook-up. Only after completing the divine hook-up should the Qigong doctor treat his or her patients.

THE DIVINE HOOK-UP

Once the doctor has established energetic grounding, the next step in establishing a clinical environment is to connect with his or her spiritual center through the divine hook-up. From this spiritual center the Qigong doctor observes the patient, void of personal thoughts, feelings, and judgments. This spiritual state of open receptivity allows patients to feel secure and to release their guarded emotions. Lacking a safe environment, patients may feel the need to repress emotions or to be distracted from experiencing and letting go of the original trauma.

The Qigong doctor, once healing has begun, enters into a state of "oneness" with the patient and the divine, in which the doctor's will and intention envelops and fuses together with the divine will and wisdom. Both doctor and patient begin to communicate through deep subconscious interactions. The doctor's success or effectiveness as a Qigong healer depends on a willingness and ability to connect with the patient and the divine healing light. Drawing from this interpersonal communication with the patient, and the deep spiritual connection with the divine, the doctor begins the treatment.

According to Dr. Meng Xian Tong of the Beijing Chengjian Integrated Traditional Chinese Medicine and Western Medicine Experts Clinic, when Qigong doctors perform the divine hook-up, they increase their own body's Righteous Qi, which then extends the energetic Message encoded in their Qi into the patient's body. The doctor's Righteous Qi is then absorbed by the patient's body and fused together with his or her Qi. This fusion empowers the patient to heal his or her disease.

Before contact with any patient the doctor first connects with the divine to purify and stabilize, not only him or herself, but also the treatment room (with a constant flow of Heavenly Qi pouring into the room). This divine hook-up enables the Qigong doctor to not only establish a sacred healing space, but also to access both spiritual insights and divine healing power.

ROOTING THE DIVINE HEALING LIGHT

Rooting the divine healing light is accomplished by visualizing a beam of healing light (the doctor can also choose specific colors) streaming into his or her body via the Baihui and Sishencong points (at the top of the head), and entering into the Taiji Pole. The doctor holds this frequency until the divine healing light becomes rooted or "anchored" within the Lower Dantian.

Next, the doctor brings the healing light vibration up the Taiji Pole to energize each Dantian until the healing light infuses and harmonizes all of the body's tissues. As the doctor feels the healing light reach the Middle Dantian, he or she visualizes the light intertwining with the "bridge of light," which connects the doctor's Heart (Middle Dantian), throat and Yintang (Upper Dantian) areas together. This energetic connection fuses at the Upper Dantian, becoming one unified cord. The healing light is then stored within the doctor's body and released during Qi emission.

USING THREE DIVINE INVOCATIONS

To assist the doctor in rooting the divine healing light, I encourage my students to begin each treatment with a small prayer and three invocations, for example:

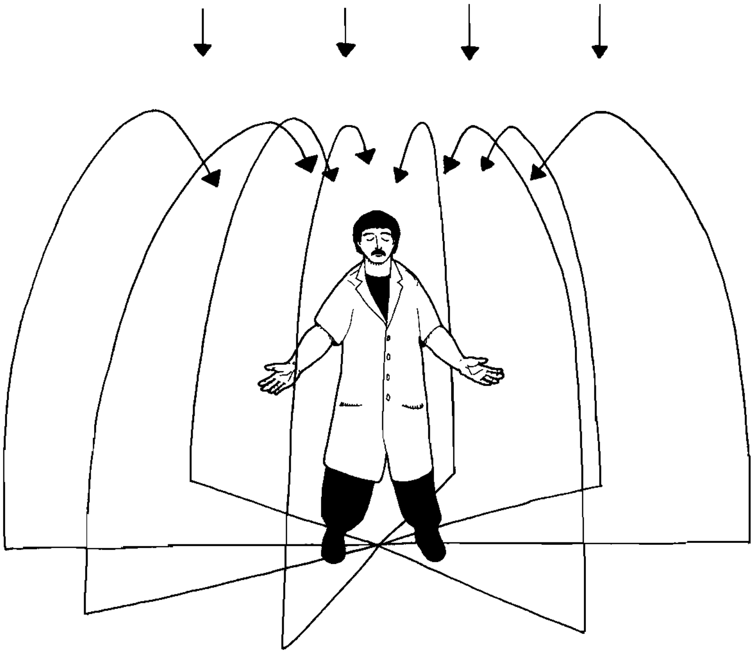

Figure 26.3. As the divine healing Qi absorbs into his or her body, the Qigong doctor expands this field to envelop the entire treatment room.

1. "Lord God, fill me with your Holy Spirit and divine healing power. Increase my skill, power, knowledge, and wisdom and use me for Your Glory." The first invocation connects the doctor to the healing energy of the divine and roots the doctor deep into the Earth. This prevents the patient's pathogenic Qi from entering into and contaminating the doctor's body.
2. "Fill this room with your Holy presence and healing virtue, and let your Glory shine through me into this room." This second invocation connects the doctor and the treatment room to the divine and secures a sacred healing space (or sanctuary) for the patient. It also establishes an expansive divine energetic field by causing the doctor's Wei Qi field to overflow with divine healing light, filling the treatment room. This divine energetic field produces a dynamic field of healing energy (Figure 26.3).
3. "I lift (the doctor will mention the patient's name) before Your Holy Presence, and ask that you may heal these wounds and free him or her from these chains of despair, agony and disease. Thank you, God. Amen." After this final invocation, the doctor envelops the patient with an energetic bubble of light and love.

Next, the doctor roots the patient's energetic bubble deep into the Earth, and creates an energetic vortex underneath the patient. The energetic vortex will spiral clockwise, deep into the center of the Earth. This energetic rooting technique is practiced to purge and drain the patient's Toxic Qi into the Earth, where it is cleansed and then recycled back into the patient's body.

ROOTING TO PREVENT ENERGETIC SHOCK

Rooting (sometimes called "energetic grounding") is an important technique used by the Qigong doctors to establish and maintain a firm physical, mental, emotional, energetic, and spiritual connection to the Earth. Connecting to the Earth is essential to prevent energetic shock (a jolt of energy which flows from the patient into the doctor's body), which can result in the doctor ab-

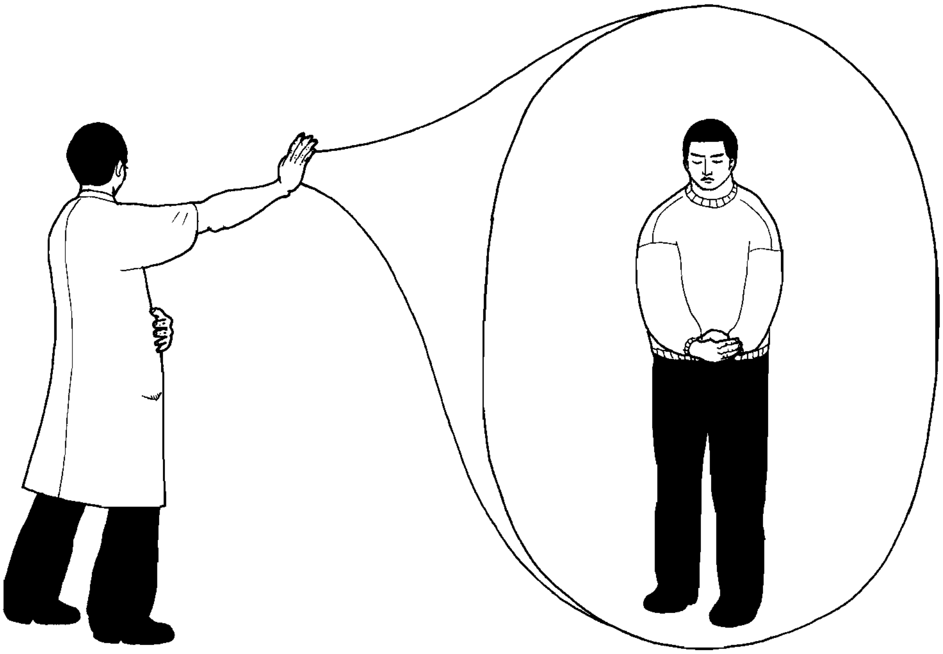

Figure 26.4. The Qigong doctor envelops the patient in an energetic bubble (or healing circle) before beginning the treatment.

sorbing the patient's energetic pathogenic disease. To prevent energetic shock from happening, Qigong doctors go through several steps to prepare themselves and the clinical environment before treating patients. Energetic rooting is utilized in all phases of meditation and allows doctors to maintain a strong physical connection to the Earth. It is especially used when the doctor is absorbing and emitting environmental and universal energy.

- From a Wuji posture, imagine an energetic cord extending from your Lower Dantian, down the Taiji Pole and out the Huiyin point (at the base of the spine) deep into the center of the Earth. This energetic cord will root your energy into the Earth. It is also affected and maintained by your will and intention. The energetic cord follows the doctor's movements smoothly and easily.
- If the patient is lying, sitting, or standing in front of you, extend your energy to envelop the patient within an energetic cocoon called "energetic circle drawing."
- Imagine extending an energetic cord from the patient's Lower Dantian through this energetic cocoon, to root the patient's body deep into the Earth. This process energetically grounds the patient's body, mind, emotion, and spirit.

THE ENERGETIC CIRCLE DRAWING PATTERN

Once the divine hook-up is completed, the doctor grounds the patient by drawing an energetic circle around the patient's body. This is used to charge the patient's energy field. This energetic projection also roots the doctor's intention around the patient and strengthens the doctor's concentration. This circular field changes the patient's static field to a divine dynamic field and enables the Qigong doctor's energy to enter the patient's body with ease.

The enveloping of the patient's body allows the patient's body and energetic fields to become even more accessible to the Qigong doctor, while keeping the patient's pathogens isolated from the doctor's body (Figure 26.4). As the doctor steps into the energetic bubble, he or she is able to treat the patient's tissues, feeling absorbed and enveloped within the Messages expressed from the patient's energetic fields.

The energetic bubble is created by emitting Qi towards the patient, and enveloping and circle wrapping the patient in a clockwise energetic rotation. This energetic bubble forms a shield, maintaining protection for the patient and the doctor. Next, the doctor creates an anchoring line to root the patient's energetic field deep into the ground. Finally, the doctor creates an energetic vortex (spiraling in a clockwise direction) underneath the patient, to pull any dispersed pathogens deep into the Earth (this helps when the doctor is dredging and purging the patient's energetic fields).

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE PATHOGENS

The energetic circle drawing pattern helps to facilitate energetic cleansing. Energetic purging facilitates the absorption of Clear Qi and also reduces the risk of contamination within the body's channels and collaterals by expelling energetic toxins and other pathogenic factors.

After creating the energetic circle, the doctor begins using dredging and purging techniques to expel pathogens from the patient's body. Next, the doctor treats the patient with tonification and regulating techniques, ending the treatment by closing the diseased area and sealing the patient's three energetic fields (three Wei Qi fields). This ending technique will securely close the energetic holes in the patient's outer field of Wei Qi (see Chapter 13). Without this sealing, the healing process is hindered due to the patient's Qi having a tendency to leak or disperse itself (even after tonification). Once the sealing is completed, any Excess Qi will be evenly distributed throughout the patient's body, establishing a Yin and Yang balance within the visceral organs.

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE TOXIC HEAT

Before treating patients who have received radiation or chemotherapy, the Qigong doctor should wrap the patient with an energetic bubble and then purge the Toxic Heat from the patient's body. The Toxic Heat is dispersed into the energetic vortex created under the patient's body, and then released deep into the Earth. The Qigong doctor can then step into the energetic bubble and begin combing, dredging and purging the patient's tissues.

USING THE ENERGETIC CIRCLE DRAWING PATTERN TO PURGE TOXIC COLD

When using the circle drawing method to purge pathogenic Cold, the doctor stands outside of the energetic bubble, and extends energy deep into the patient's Taiji Pole. Once the doctor feels the patient's center core Qi, he or she draws the pathogenic Cold or Hot energy out from the patient's tissues (in all directions), filling the energetic bubble. The doctor, standing outside the energetic bubble, observing the patient's energetic field, waits for an energetic shift which is immediately followed by Toxic Cold (or Hot) Qi filling up the energetic bubble. Once this occurs, the Qigong doctor must drain the pathogenic Cold (or Hot) Qi down into the Earth to rid the patient of the Toxic Qi.

TREATING THE PATIENT

After the Qigong doctor has "circle wrapped" the patient, the treatment begins. First, the doctor dredges the patient's external energetic fields, channels and collaterals (from head to feet). The doctor senses the patient's Qi with his or her hands, changing the depth of penetration of the touch by extending intention deeper into the patient's tissues. Once the patient's pathogenic Qi has been reached, the doctor feels various energetic sensations (vibration, warmth, thickness, etc.) and removes any pathogenic energy from around the patient's body.

The doctor continues the treatment by extending his or her intention deep into the patient's body, purging and dispersing the toxic pathogens from out of the patient's internal organs and channels. It is usually during this deep internal purging that the patient experiences profound emotional discharges, observes flashes of light, patterns of colors, see visions, or relives deep seated past emotional traumas. This purging releases toxic emotions which are generally the cause or contributing factor to the patient's disease.

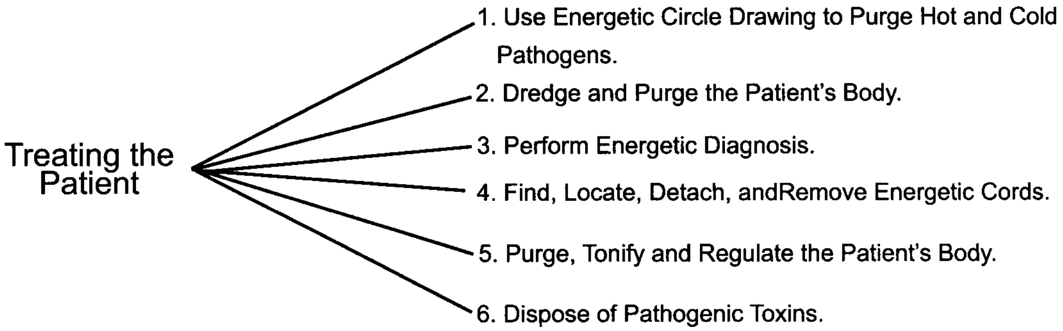
Treating the Patient
1. Use Energetic Circle Drawing to Purge Hot and Cold Pathogens.
2. Dredge and Purge the Patient's Body.
3. Perform Energetic Diagnosis.
4. Find, Locate, Detach, andRemove Energetic Cords.
5. Purge, Tonify and Regulate the Patient's Body.
6. Dispose of Pathogenic Toxins.

Figure 26.5. The Different Steps Involved in Treating a Patient

After the patient has been sufficiently purged and cleansed of pathogenic energy, the Qigong doctor begins to tonify the patient's internal organs, channels, collaterals, and Righteous Qi. This is done to help the patient build a stronger immune system, and fight or avoid any future complications caused by an Excess or Deficient internal organ condition (Figure 26.5).

DREDGING AND PURGING TECHNIQUES

When treating a patient, the doctor purges an internal organ and dredges a channel. The technique of dredging is a particular type of purging method. The energetic circle drawing techniques are always performed with dredging techniques (using the pulling and stroking methods of energy manipulation), to drain the energy into the Earth. These techniques can be initiated with the patients either sitting, standing, or lying.

It should also be noted that the diseased area of a patient's body must be thoroughly dredged and prepared before any kind of tonification or energizing is performed. This technique is similar to a Western doctor cleaning a wound before operating.

If the Qigong doctor does not dredge and purge the pathogens from the patient's body before treating, the patient can suffer adverse reactions when the doctor's clean energy fuses with the patient's Turbid Qi.

DREDGING AND PURGING THE PATIENT

There are two stages of clinical purgation which are described as follows.

1. The doctor dredges and purges the surface pathogens from the patient's body. The doctor moves to the right lateral side of the patient's body and energetically combs the patient's body from head to toes using the Tiger Kneading and/or Extended Fan Palm techniques (in this particular example the patient is lying supine on a treatment table). The combing technique is performed to energetically evaluate the patient's reaction to the doctor's emitted Qi, while dredging the patient's three energetic fields of Wei Qi.
2. The doctor purges from the bottom of the patient's feet. The Qigong doctor stands by the patient's feet and continues to dredge the patient's body by energetically combing the patient from head to toes. This action further purges the patient's body of pathogenic energy.

ENERGETIC DIAGNOSIS

After performing the clinical purgation, the Qigong doctor begins the energetic diagnosis. Some examples of energetic diagnosis are described as follows.

- The doctor can begin energetic diagnosis by first connecting with patient's heels (either by touching or not touching). This particular method and approach to diagnosing allows the Qigong doctor to energetically connect with the patient's inner-fascia. It is through this connection that the Qigong doctor feels the energetic flow of Qi moving through the patient's internal organs. By extending energy from the patient's heels up to the head (like a rippling ocean wave or sonar), the doctor can "listen" for obstructions or deviations of Qi flow.

- Next, the doctor can use the Flat Palm diagnosing technique to scan the patient's entire body for any Excess or Deficient areas within the patient's Wei Qi fields. This second type of diagnosis serves to verify the obstructions or deviations of Qi flow that the doctor discovered while first connecting with the patient's heels.

PURGING, TONIFYING AND REGULATING THE PATIENT'S BODY

After diagnosing the patient's body, the Qigong Doctor purges, tonifies, or regulates the energetic fields, channels, tissues, organs, and organ systems, according to the etiology and prognosis of the disease.

While dredging and purging the patient's body, the Qigong doctor may discover several energetic cords attached to the patient's tissues and internal organs. These energetic cords may sometimes contribute to the cause of the disease.

UNDERSTANDING THE BODY'S ENERGETIC CORDS

The human body attaches, extends, absorbs, and collects energetic cords. The size and shape of these cords can be big, small, thick, or thin, depending on the amount of emotion invested in their attachment. These cords are located within the Three Dantians and Chakra Gates, and form the spiritual, emotional, mental, and physical connection of the patient to the outside world. Generally, what makes the patient sick is not his or her genetic predisposition, but the negative, co-dependent bonds the patient creates in his or her relationship with others. It is the spiritual, emotional, mental, and physical reactions to these relationships that either promote health or disease. Energetic cords are also formed towards objects, pets and places that have special sentimental or emotional value (be they positive or negative) to the patient. Energetic cords are also sometimes used to feed the formation of the preexisting energetic clusters and cysts, which later transform into the patients' diseases.

Energetic cords are the psychophysical connection between what the patient feels, believes and knows to be true; they represent the connection to the patient's personal reality. This does not mean that the patient's reality is valid, only that the patient has come to believe it to be true. By removing these cords an energetic shift is initiated, allowing the patient to reprogram his or her feelings, thinking processes, and behavior, by becoming proactive instead of reactive. Energetic cords are full of light and energy and can be categorized into two distinct divisions: prenatal (congenital) and postnatal (acquired) energetic cords.

- Prenatal Energetic Cords originate in the fetus, when the baby is still in the womb. They form the strong congenital attachment and permanent bond the child has with the mother and father. Because of the mother's energetic/spiritual connection with her egg, and the father's energetic/spiritual connection with his sperm, when the egg and sperm unite, the energetic cords of each parent fuse together with the universal and environmental energy creating the energetic structure of the child. It is through the parents' energetic cords that inherited talents and defects are passed along to the infant.
- Postnatal Energetic Cords are acquired after the baby is born through interaction with siblings, relatives, and other individuals (in some cases animals or places are involved). These cords will either form the emotional attachments which can help the child through life by being a means of support, or set the foundation for disease by depleting the child of his or her life-force energy. They represent the emotional bonding (positive or destructive) inherent within each one of us.

In a clinical environment, the attachment of these cords (prenatal and postnatal) will affect the healing relationship of both the patient and the doctor.

- The Patient's Energetic Cords are still connected to the patient after the pathogenic Qi has been removed from the patient's body. These small cords are the energetic attachments to part of the patient's old belief structure (which created the disease or emotional attachment). Although there is still an energetic connection to the departed pathogenic Qi, these energetic cords will soon disperse once the patient's Creative Subconscious is able to reprogram the individual's belief structures. Only then can the physical body, as well as the mental, emotional and spiritual energetic Wei Qi fields maintain the new energetic pattern. This new belief structure creates an energetic transformation, allowing the patient to release his or her hold on to the painful past and begin to heal. The patient is encouraged to emotionally disconnect from old emotional belief structures through creative visualization, and by affirming the new belief structure. Through affirmation and positive emotional, spiritual visualizations, the patient is able to "sever" any energetic cords that connect to the pain and disease.

- Energetic Cords between the doctor and patient are formed during healing sessions. These new energetic cords are passed back and forth between the Qigong doctor and the patient. Sometimes this occurs without either party being aware of this energetic connection. Usually, as the doctor steps away from the patient and reconnects with the divine, an energetic purging is automatically initiated, freeing the doctor from the attachment to the patient's energetic cords. If, however, there has been an emotional connection between the doctor and the patient, these cords can remain attached even when the patient leaves the office.

  The doctor should make every effort to disconnect from the emotional attachment of the patient's cords, otherwise the increased energetic pull on the doctor's center will affect the doctor on every level. Staying connected to the patient through an energetic cord can cause not only emotional, but physical and spiritual influences on the doctor. Any remaining cord can be used by the patient to unconsciously project thoughts and feelings onto the doctor. This can result in the doctor being depleted of energy either through direct draining of his or her Qi, or through obsessing about the patient's well-being, long after the treatment has ended. The drain of the energetic cord can lead to either clinical "burnout" from absorbing excessive negative emotions, or deep emotional bonding, which can lead to unethical sexual indiscretions.

FINDING, DETACHING, AND REMOVING ENERGETIC CORDS

Energetic cords manifest through their effects on behavior and emotional responses of the patient. They are relationship based patterns, located within the patient's tissues. Energetic based patterns can evoke certain reoccurring emotional responses, such as attractions, addictions, etc.

1. To assist the patients in removing the toxic energetic cords, the Qigong doctor must first locate the cords.
   a. Through the emission of Qi, the doctor may elicit (from the patient) either an attraction or repulsion response. This response occurs when certain memories are stimulated within the patient's body, indicating that there is an energetic cord attached within the tissues.
   - The attraction response causes the patients to experience deep emotional release through laughing, crying, shouting, etc.
   - The repulsion response causes a physical jerking of the patient's muscles, anywhere within the patient's body (observed by the wincing and grimacing of the patient's face). Often the patient will "Shen-out" (emotionally and spiritually leave their body) to avoid feeling the painful surfacing emotions.

   b. The doctor locates the origin and entry point of the energetic cord through Flat Palm detection. The root of an energetic cord acts as an open receptor to a specific psychophysical problem that the patient is experiencing. When stimulated, these cords may manifest physical symptoms such as nausea, hot flashes, vertigo, dizziness, headache, and muscle contractions. The patient may also experience emotional reactions such as momentary panic, fear and so on.

2. Before detaching the energetic cords, the doctor requests the patient's permission to do so either verbally, or silently through mind to mind communication.
3. To remove the energetic cord, the doctor dredges and purges the patient's body, while silently communicating through the Creative Subconscious mind. The doctor encourages the patient to let go of, and fully release, the toxic cords along with the memories that sustain them. To ensure against their recreation, the doctor discusses with the patient any images, emotions or sensations that surfaced during the removal of these cords; through this process the doctor brings the patient's conscious awareness of the fears, pain and major unresolved issues that resulted in the disease. The patient thus gains a clear identification of his or her traumas to facilitate the reprogramming of his or her dysfunctional relationships through affirmations and creative visualizations.

If the patient is not ready for such a transition (feeling scared, unprepared, undeserving, etc.), his or her subconscious mind will sabotage any effort made in that direction, and the energetic cord, emotional cluster and all energetic attachments to the disease will be recreated and reintegrated. It is important in cases such as these, that the Qigong doctor refer the patient out to a psychotherapist for additional support.

In Western society, the Twelve Step Programs (such as Alcoholics Anonymous) are but one example of a pattern based reprogramming of energetic cord attachments. Alcohol and Narcotics Anonymous groups, for example, help detach energetic cords toward illegal substances, whereas, CODA meetings help reprogram the energetic cord attachments to co-dependant relationships. In a codependent relationship, both parties rely too heavily on one another, instead of being interdependent. One person usually dominates the other, as both parties base their self-esteem on the partner's approval, or disapproval, thus avoiding the personal responsibility of their own actions.

Energetic reprogramming encourages the participants to take responsibility for filling and replacing the energetic gap that the toxic cord has left with a connection and attachment to the divine, or "Higher Power."

DISPOSING OF PATHOGENIC TOXINS

One problem a doctor may confront when dredging and purging is where to dispose of the patient's toxic and noxious Qi. To maintain a clean treatment room it is essential for this pathogenic energy to be disposed of properly. This will keep both the doctor and other patients from contamination. If the doctor's Qi becomes toxic or contaminated during treatment, this turbid energy may be transmitted to other patients.

Because the patient is in a state of energetic shock (after the therapy), attention must be placed on the patient avoiding contact with the contaminated Qi. Otherwise, there is a possibility of the diseased Qi returning into the patient's body via the energetic cords of attachment connected to the Toxic Qi.

Imagination and visualization are two of the most powerful and widely used techniques employed in healing today for disposing of Toxic Qi. Since we experience reality through creative imagination, pictures, words, and feelings are extremely powerful techniques which can be used to restructure our physical, mental, emotional, energetic, and spiritual foundations. This also holds true for the disposing of the patients' pathogenic Qi.

There are several methods for disposing of the patients' pathogenic energy in the treatment room. These methods include purification by either sound, light, or the Five Elements, in conjunction with disposing of the toxic energy through the energetic vortex.

PURIFICATION BY SOUND AND LIGHT

- Sound penetrates any substance and causes profound shifts in energy, quickly breaking up clusters of stagnant energy. Some examples of sounds used for purifying the treatment room include: music, toning, chanting, singing, wind chimes, gongs, singing bowls, bells, drumming, and clapping the hands.

- Light can be used to cleanse and purify the treatment room of toxic energy, due to the radiant heat emanating from the light source. Some examples of light being used to purify the clinic include: exposure to sunlight, candles, and imagining the divine healing light incinerating any energetic toxic debris.

PURIFICATION BY THE FIVE ELEMENTS

Some Qigong doctors utilize the energy stored within the Five Elements to cleanse the treatment room of toxic debris. This works best in conjunction with the light and sound purification methods. When using one or more of the elements, it is best to consider how much debris has accumulated from the previous treatment and choose accordingly.

1. Dispose of the Toxic Qi through Fire. By burning fire-water (alcohol with salt), lighting a candle, or releasing the Toxic Qi into the Earth's molten core, Toxic Qi can be transformed and purified.
   - To create fire-water, the Qigong doctor places a small amount of alcohol in a round dish and sprinkles a cone of salt (or camphor) in the center of the dish. Next, the doctor lights the salt and allows the smoke to purge the treatment room, making sure that the doors and windows are open to secure clean ventilation.
   - Other Qigong doctors may imagine the center of the Earth as an incinerator. This image is used to purge the patient's Toxic Qi by directing it to the Earth's molten core. Once the patient's Toxic Qi is incinerated (at the speed of thought), the transformed / purified energy is reabsorbed back into the patient's body.
2. Dispose of the Toxic Qi through Water. Water can also be used to absorb toxic energy and for healing and purifying. Some examples of water usages include: Holy Water and salt water. Some doctors use a container of salt water as their disposal unit for the patient's Toxic Qi (Figure 26.6). Once the patient's toxic energy reaches the container of salt water (usually placed underneath the treatment table), it is immediately absorbed due to the molecular structure of the salt crystals. Sometimes it is advisable for the Qigong doctor (after a full day at the clinic, or at the end of a busy week) to submerge him or herself in a baking soda or take an Epsom salt bath in order to cleanse and purify the body.
3. Dispose of the Toxic Qi through Wind / Air. Wind and air can be used to purify the room of toxic energy. Some examples of this include open air ventilation along with the placement of indoor plants.
4. Dispose of the Toxic Qi through Divine Healing Light. Some doctors dispose of the patient's pathogens into the air, where it is immediately incinerated by a divine healing light. Because of the preestablished divine healing field of energy, once the patient's toxic energy leaves the doctor's hand, it is immediately absorbed by the divine healing light and is instantly transformed into purified energy and reabsorbed back into the patient's body.
5. Dispose of the Toxic Qi through Wood. Wood can also be used to purge toxic energy from the treatment room. Some examples of the Wood Element used to cleanse the room are: the burning of incense or herbs such as myrrh, frankincense, cedar, sage (an old American Indian method), sandalwood or camphor.
6. Dispose of the Toxic Qi through the Earth. Certain stones can be used in the clinic because of their ability to absorb Toxic Qi. Strategically placed in the clinic, these Earth Elements can be used as a natural filtering device. Some examples of this include: amethyst, salt, clear and colored quartz, as well as many other minerals or gemstones. These crystals require cleaning at the end of each treatment day to dispose of the toxic Qi they have absorbed, and should be cleansed by placing them in salt water till they are clean, or burying them in the ground for three days.

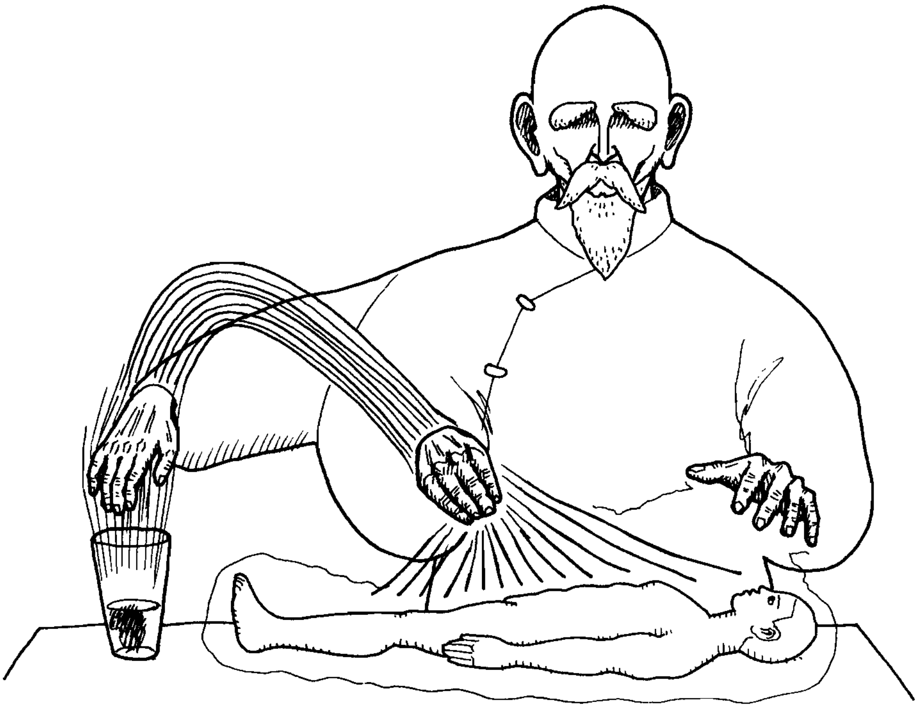

Figure 26.6. One example of a Qigong doctor using a rubber acupuncture doll for treating and disposing of the patient's toxic energy into salt water.

MEDICAL QIGONG TREATMENTS

The following is a clinical procedure that a Medical Qigong doctor will use when treating a patient. The clinical procedure known as the "preparation" remains the same for all Medical Qigong treatments (i.e., the hook-up, dredging and purging the patient, as well as energetic diagnosing). However, the initial treatment will vary depending on the patient's constitution, disease, and personal goals.

After scanning the patient the doctor has a solid understanding of which areas or organs in the patient's body are currently in a state of Excess or Deficiency. The doctor then chooses which area to begin dredging and purging in accordance with the treatment plan. Only after purging all areas of Excess should the doctor then move to tonifying those areas that are Deficient.

TONIFYING A PATIENT'S BODY

After completing the diagnosis, if you have determined that the patient's body is not in an Excess condition, but is instead constitutionally weak or depleted due to Kidney exhaustion, the treatment may be continued as follows.

1. Begin by extending energy through the patient's Yongquan Kd-1 point at the bottom of each foot. Imagine that the Qi is ascending up the patient's legs, via the Yin channels, filling the Lower Dantian, Kidneys, thorax and spinal column into the brain. Hold this position for several breaths, then proceed to the next area.
2. Physically embrace the area of the patient's ankle with one palm, while the other palm embraces the Kd-1 point on the same foot. Start with the right ankle first and then the

left. Project energy into the foot and ankle, by holding the mind's intention on that specific area for several breaths, then begin circling the energy up the Yin channels and down the Yang channels. Hold this position for several breaths, then proceed to the next area.

3. Energize each of the patient's knees (back and sides) by holding the mind's intention on those specific areas for several breaths, then connect each knee to the Yongquan Kd-1 point at the bottom of each foot (the right foot first, then the left). Connect and emit energy into the bottom of each foot to the knee, circling the Qi up the Yin channels and down the Yang channels. Hold this position for several breaths, then proceed to the next area.
4. Energize each of the patient's hips, by holding the mind's intention on that specific area for several breaths, then connect each hip to the Yongquan Kd-1 point at the bottom of each foot (right first, then the left). Connect and project energy into the bottom of each foot to the hips, circling the Qi up the Yin Channels and down the Yang Channels. Hold for several breaths, then proceed to the next area.
5. Energize the patient's Lower Dantian by holding the mind's intention on that specific area for several breaths, allowing the Qi to completely fill up the energetic chamber. Place your left hand on the patient's Mingmen GV-4 point. Your right hand is placed on the patient's navel (CV-8) with your fingers extending down to the patient's Qihai CV-6 point. Mentally connect the Lower Dantian with both Yongquan Kd-1 points on the bottom of the patient's feet and imagine the energy simultaneously traveling down the outside of both legs into the feet. Imagine the energy from the feet continuing to travel along the inside of both legs back into the patient's Lower Dantian area. Hold this intention for several breaths, then proceed to the next area.
6. Energize the patient's Yellow Court by holding the mind's intention there for several breaths. Your left hand is placed on the Jizhong GV-6 point, your right hand should be placed on the Juiwei CV-15 point. Mentally connect the Yellow Court with the patient's Lower Dantian, allowing this energy to connect and fuse both areas. Hold this intention for several breaths, then proceed to the next area.
7. Energize the patient's Middle Dantian by holding the mind's intention there for several breaths. Place your left hand on the Shendao GV-11 point, while your right hand is placed on the Shanzhong CV-17 point. Mentally connect the Middle Dantian area with the patient's Lower Dantian. Hold this position for several breaths, then proceed to the next area.
8. Energize the patient's throat area by holding the mind's intention on that specific area for several breaths. Your left hand is placed on the Dazhui GV-14 point, while your right hand is placed on the Tiantu CV-22 point. Mentally connect the throat area to the Middle Dantian area, then to the patient's Lower Dantian. Hold this intention for several breaths, then proceed to the next area.
9. Energize the patient's Upper Dantian area by holding the mind's intention there for several breaths. Your left hand should be placed to stimulate the occipital region at the Naohu GV-17 point. Your right hand is simultaneously stimulating the Yintang (Third Eye) and Baihui (GV-20) areas. Mentally connect the Upper Dantian area to the Middle Dantian first, and then to the patient's Lower Dantian. Hold this intention for several breaths, then proceed to the next area.
10. Regulate the patient's Microcosmic Orbit, to balance the Heart Fire and Kidney Water energy. Remove your hands from the patient's body, yet still remain energetically connected through your intention. Having stepped away from the treatment table, continue treating the patient with external Qi emission, regulating the patient's Microcosmic Orbit (extending energy up the Governing Vessel and down the Conception Vessel). Hold this intention for several breaths, then proceed to the final step in the treatment.

11. Envelop and seal the patient's three external fields of Wei Qi by connecting the patient's body with the divine energetic field. Hold this intention for several breaths, then proceed to the "closure."
12. Disconnect yourself from the patient and close the healing session. Be sure to disconnect physically, mentally, emotionally, energetically, and spiritually from the patient by willfully detaching all energetic connections to the patient.
13. After disconnecting from the patient, center yourself, then regulate your own energetic fields to balance, recharge, and replenish any energetic depletion.
14. Finally, after you have completed centering yourself, finish the cleansing by washing your hands to purge any of the patient's lingering pathogens which may still be attached to your body.

GENERAL THERAPEUTIC TREATMENT

The following is an example of a clinical routine used in the Medical Qigong Clinic for "emotional purging." The following purging routine is separate from the previous clinical modalities for tonifying a patient. It is important to note that when treating a patient, the Qigong doctor usually purges the patient's Liver and Lungs before tonifying the Spleen and Kidneys.

After completing the diagnosis, if you have determined that the patient has both Excess and Deficient conditions, you may decide to continue the treatment as follow.

1. Begin by dredging, purging, and activating (energizing) the patient's Upper Dantian area to reconnect the patient to the divine, to activate the patient's spiritual/emotional presence.
    a. Start by purging the patient's Upper Dantian.
    b. Then regulate and energize the patient's Upper Dantian by performing a small Microcosmic Orbit within the patient's head.
    c. Finally energize the patient's Taiji Pole through the Baihui area.
2. Purge and dredge the patient's Lungs, to disperse the emotions of grief and despair caused by "rejection, betrayal, and abandonment issues" which are stored in the patient's Lungs.
    a. Start by dredging the patient's neck (trapezius, clavicle, and manubrium) of grief and despair. Begin by standing on the right side of the patient's body, and use the Kneading Tiger Palm and Dragon's Mouth Palm hand techniques (see Chapter 33) to pull the Toxic Qi from the patient's body. Switch from the right to the left side and continue purging.
    b. Then roll the patient's shoulders back to open the Lungs and to facilitate their purging of any deep seated grief. Start on the left Lung first. This area is responsible for storing the emotions related to the patient's mother.
    c. Next dredge the inside Yin Channels of the patient's left arm by using Qi Massage techniques (see Chapter 37).
    d. Regulate and treat the patient's body through "hand diagnosis and treatment" (based on energetic hand/body association, see Chapter 25) by focusing on the patient's left hand.
    e. Finally, begin the entire Lung purging procedure again, this time focusing the attention on the right side of the patient's body. This side reflects the patient's issues with his or her father.
3. Regulate and tonify the patient's Heart to calm the Spirit.
    a. Your left hand connects and extends energy deep into the patient's Shendao GV-11 point, the area in-between the patient's shoulder blades. Your right hand connects to the Shanzhong CV-17 point above the patient's Heart, purging emotional toxins.
    b. Finally, regulate the Heart by circulating Qi through the patient's upper Microcosmic Orbit (between patient's Upper and Middle Dantian). This reconnects the divine light emanating

from the Upper Dantian back with the spiritual emotions (unconditional love, devotion, etc.) stored within the upper chamber of the Heart (see Chapter 5).

4. Purge and dredge the patient's Yellow Court to rid the body of toxic emotions stored within the tissues of the Yellow Court. Your left hand remains secure on the patient's back, at the Shendao point to maintain contact with the Heart Qi. The patient's true feelings are stored within the Yellow Court and expressed through the patient's Heart. Place your right hand on patient's Juiwei CV-15 point, located below the patient's xiphoid process. Sometimes you may see images of the patient's past as traumas are released from the tissues.
5. Your left hand remains secure on the patient's back, at the Shendao point while your right hand purges the patient's Liver organ of anger and depression caused through the storing of "rejection, betrayal and abandonment issues." As you dredge the Liver's Hot toxic emotions from out of the patient's body, purge the pathogens out the patient's right Gall Bladder Channel.
6. Keep your left hand on the patient's back, at the Shendao point while you tonify and regulate the patient's Spleen with your right hand.
7. With your left hand on the patient's back, at the Shendao point, tonify and regulate the patient's Kidneys with your right hand.
8. Maintain contact with the patient's back, at the Shendao point, as your right hand connects with the patient's Mingmen (GV-4) area to regulate all three Dantians. This is initiated through the Microcosmic Orbit (extending energy up the Governing Vessel and down the Conception Vessel), to balance all of the patient's energetic reservoirs.
9. Then slowly remove your hands from the patient's body, but remain energetically connected through intention. Step away from the treatment table, as you continue to regulate the patient's Microcosmic Orbit. Hold this intention for several breaths, then proceed to end the treatment.
10. Envelop and seal the patient's three external fields of Wei Qi, by connecting the patient's body with the divine energetic field. Hold this intention for several breaths, then proceed to the "closure."
11. Disconnect physically, mentally, emotionally, and spiritually from the patient to close the healing session. It is important to shake or whip your palms to completely disconnect from the patient's energy. The shaking or whipping action frees the doctor from any energetic attachment which may have occurred while treating the patient.
12. After disconnecting from the patient, center yourself, then regulate your own energetic field. This balances, recharges, and replenishes any energetic depletion you may have suffered during the treatment session. If you are affected by any pathogenic Qi, immediately expel it from your body.
    a. Next, focus your attention on your Taiji Pole, imagining both the universal and environmental energy collecting, absorbing, and resonating deep within your Lower Dantian and center core. This image draws Qi into your body like a magnet, replenishing any energetic depletion you may have experienced. By placing the fingertips together in front of the Lower Dantian, you will further facilitate the replenishment of the energy.
    b. During this time the patient is allowed to continue relaxing on the treatment table and to slowly come out of the trance state induced by the treatment. The patient should then allow the Yuan Qi to settle back to its origin (the Lower Dantian), by relaxing and calming the mind, breathing, and posture. Instruct the patient not to wash the area that was treated for several hours because water absorbs some of the Qi that was projected into the diseased area.
13. Finally, wash your hands, using tepid water (hot water allows turbid Qi to enter the

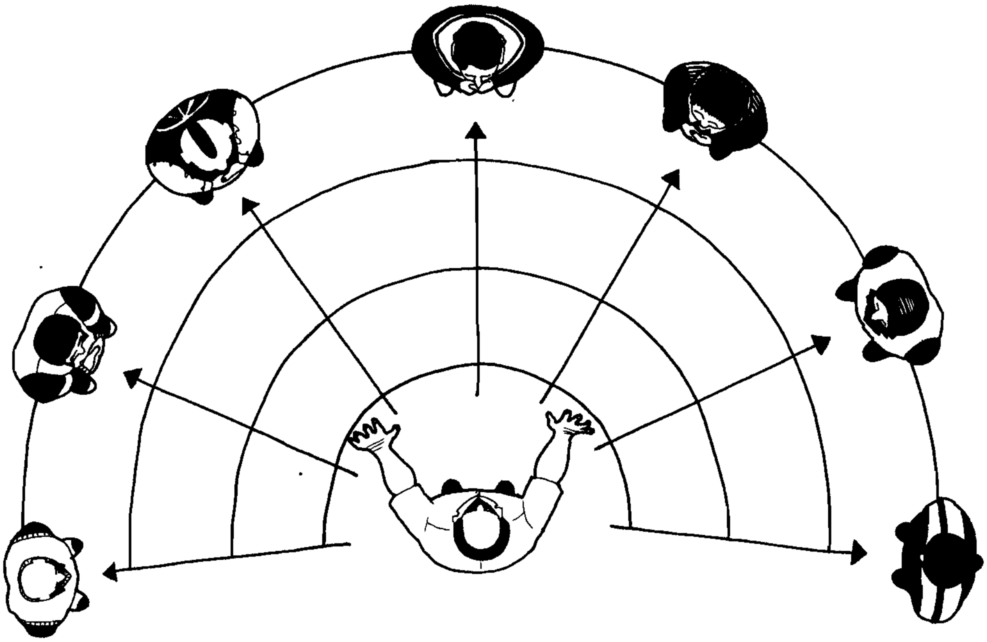

Figure 26.7. As the patients practice their own individual Medical Qigong prescriptions, the Qigong doctor extends energy into the patients, helping to facilitate a stronger therapeutic reaction.

doctor's channels, cold water closes off the channels) to purge any of the patient's lingering pathogens.

ADJUSTING AND FORTIFYING THE PATIENT'S ENERGETIC GRIDS

At the end of the treatment the doctor regulates the patient's energetic fields by adjusting and fortifying the patient's energetic grids (the webbing within the patient's energetic fields). This keeps the patient's energy in harmony, allowing both his or her internal and external fields to operate at their maximum potential for health and healing.

The Qigong doctor adjusts and fortifies the patient's energetic grids, to initiate a permanent energetic transformation for healing within the patient's energetic fields, tissues, and cells. The body's energetic grids are responsible for tissue formation and for maintaining the production, growth, and development of energy. These grids are influenced by thoughts and emotions.

The creative subconscious mind maintains the body's energetic grid formations and causes the physical body to act (grow stronger or weaker) in accordance with the energetic "blueprint" encoded within the structure or webbing. After the Qigong doctor has changed, or corrected, the energetic grid, the patient is given Medical Qigong prescriptions (homework) in the form of visualizations and affirmation meditations, as well as Qigong exercises.

TREATING MULTIPLE PATIENTS

Once an energetic field in the treatment room is established, the patients can begin their individual Qigong regulations. The doctors position themselves according to the number of patients needing treatments; for example, one doctor may be positioned in front of several patients and extend Qi into the patients' energetic fields (Figure 26.7). The doctors' postures and energy extension varies according to their individual skill and energy projection level. It is important for Qigong

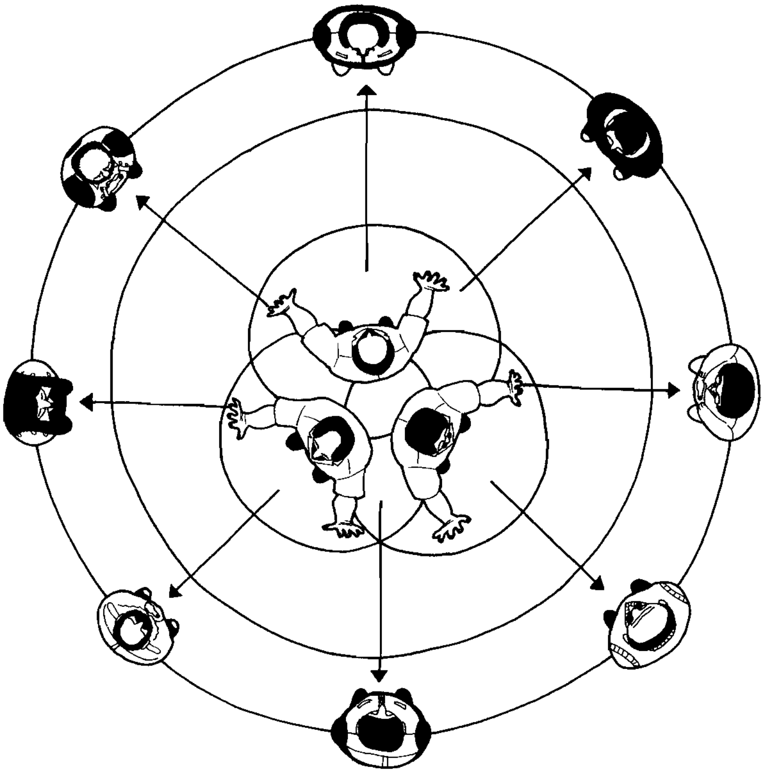

Figure 26.8. The Qigong doctors will position themselves back to back, in the center of a circle, in order to induce energetic movement within the multiple patients' tissues.

doctors to set the pace for treatment. The patients' rhythm is based on the pulse and vibration of the Qigong doctors.

In China, doctors usually rotate every fifteen minutes when treating patients. This prevents the doctors from fatiguing while allowing the patients to absorb new and fresh Qi every fifteen minutes.

If only one or two Qigong doctors are available when treating several patients, then the doctors will position themselves back to back in the middle of the room, with their patients surrounding them in a circle (Figure 26.8). After several minutes the doctors will slowly begin to rotate the center circle extending their energy, and treating each patient individually. The doctor's emitted Qi combines and synergizes with the patient's Qi to produce a much stronger energetic healing field.

By establishing a treatment room that facilitates the regulation of the patients' Qi, the treatments tend to become more effective with faster results. In China, patients are treated in a group treatment room environment three times every day (sunrise, sunset, plus one time before bed).

ENDING THE TREATMENT

When the treatment is over, encourage each patient to relax and then return the patients' Qi back to its origin (the Lower Dantian). The duration of the treatment time, and proper recuperation after Medical Qigong treatment will be addressed next.

DURATION OF TREATMENT TIME

The duration of treatment time varies according to each patient's disease, condition, constitution, and age. A long treatment is usually between twenty and thirty minutes and is generally performed on cancer patients, senior citizens, or for severe traumatic or chronic illnesses. A short treatment is generally between three to fifteen minutes and is usually administered for athletic sprains and ligament dislocations.

PROPER RECUPERATION AFTER MEDICAL QIGONG TREATMENT

Proper recuperation is very important for the replenishment of strength and energy after the treatment session (even after the patient's own prescription practice). This may include having the patient change his or her life activities. The patient must incorporate a balance of work and rest, and acquire regular healthy eating habits (with reasonable nutritional goals). Establishing adequate sleeping patterns is also necessary. Of equal importance is an open-minded positive attitude by the patient, for this reason, a supportive network can be invaluable.

After the Qigong treatment the patient is also encouraged to avoid cold showers, the consumption of alcohol, excessive sex, and recreational drugs. Also prohibited is anything that will emotionally shock the patient, as remaining in a quiet state of mind and spirit is essential.

CHAPTER 27
COMBINING MEDICAL QIGONG THERAPY WITH OTHER HEALING MODALITIES

COMBINING MEDICAL QIGONG THERAPY WITH TCM MODALITIES

Traditional Chinese Medical treatments in conjunction with clinical Qigong modalities are generally broken down into exclusive Medical Qigong Therapy, Medical Qigong and Chinese Massage Therapy, Medical Qigong and Acupuncture/Moxa Therapy, and Medical Qigong and Herbal Medicine Therapy.

EXCLUSIVE MEDICAL QIGONG THERAPY

Medical Qigong therapy can be administered by the Qigong doctor (Figure 27.1) to tonify and promote energy circulation or to sedate the body's energetic channels and collaterals. Qigong prescriptions can also be given for the patient to practice alone. Medical Qigong therapy by itself is excellent for the treatment of ailments such as:

- Insomnia
- Neurasthenia (Chronic Fatigue)
- Hypertension
- Stroke
- Acute abdominal pain
- Injuries to the nerves
- Hemopleura (Blood in the walls of the thorax and diaphragm, along the pleural space)
- Myopia
- Migraines
- Muscle atrophy
- Sprains
- Cancer
- Tumors
- Cysts

MEDICAL QIGONG AND CHINESE MASSAGE THERAPY

Medical Qigong therapy in conjunction with Chinese Massage therapy is a powerful form of combined clinical modalities (Figure 27.2). It enhances tissue response as well as the activation of psychophysical reactions within the body. These therapies can be combined to clear the joints before setting the bones (as in Jie Gu therapy), to purge and dredge the channels to adjust the Qi circulation before regulating the body's muscular system (as in Tui Na and Gua Sha therapies), or to disperse toxins before regulating the body's internal viscera

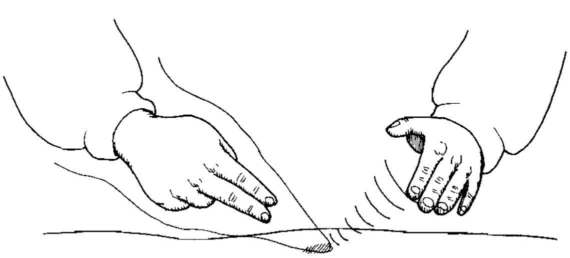
Figure 27.1. Medical Qigong Therapy

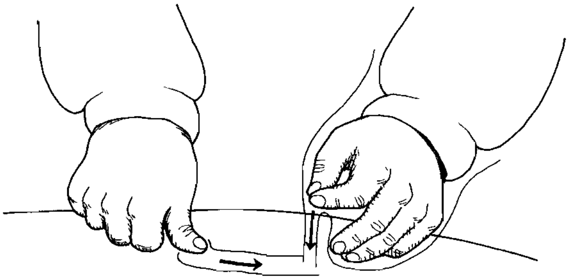
Figure 27.2. Medical Qigong Therapy with Chinese Tissue Regulation Therapy

(as in An Mo and Jing Point therapies). Qigong and Chinese Massage therapy are most commonly used in the treatments of patients with:

- Deep tissue obstruction
- Stiff muscles
- Frozen shoulder
- Low sensitivity to emitted energy

MEDICAL QIGONG AND ACUPUNCTURE/MOXA THERAPY

Acupuncture needles are inserted into the patient's body before the doctor extends Qi into the patient to facilitate healing (Figure 27.3). This also applies in the treatment of Moxa sticks in which the doctor extends energy into the patient through the burning herb. The combined energies (the Moxa heat and doctor's own Qi) initiates increased Qi flow within the patient's channels resulting in a stronger tonification. Moxa is especially beneficial in the treatment of arthritis. Qigong and Acupuncture therapy are combined in tonification or purgation treatment.

MEDICAL QIGONG AND HERBAL MEDICINE

Medical Qigong Therapy in combination with herbal therapies (or with pharmaceutical prescriptions) are becoming more popular. There is ample evidence that the combination of self-applied Medical Qigong therapy and drug therapy is far superior to that of drug therapy alone. In China, this data is reported in several studies of patients with hypertension as well as in cancer patients.

Medical Qigong therapy relaxes the body, promotes the flow of Qi, Blood, oxygen, and nutrients to all cells of the body, as well as promotes the removal of waste products from the cells. Medical Qigong also promotes drug uptake by the cells and tissues by means of increased microcirculation within the body's microcirculatory system. This increased Qi flow through the body's microcirculation nourishes diseased or stressed tissues.

One diagnostic herbal technique I observed at the Xi Yuan Hospital was performed by a Qigong doctor who absorbed into his body the effects of a specific medicine by touching the bottle that contained the herbs. He then extended energy to his patient and began absorbing the patient's Qi to see if the patient's body was affected by the herbal prescription (Figure 27.4).

The Qigong doctor is encouraged to prescribe herbs along with certain of the Medical Qigong exercises. If the Qigong doctor is not qualified to prescribe Chinese medical herbs, then a referral can be made to an acupuncturist/herbalist. The herbal prescriptions will vary according to the patient's condition, constitution and illness.

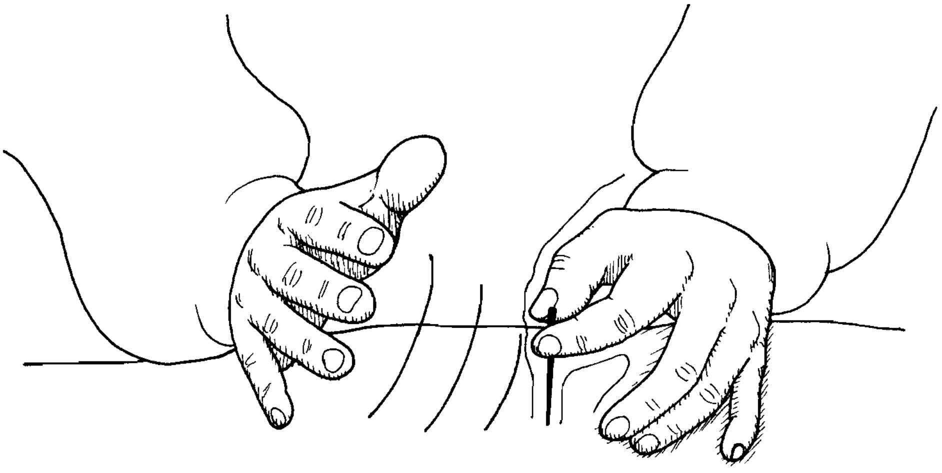

Figure 27.3. Medical Qigong Therapy with Acupuncture

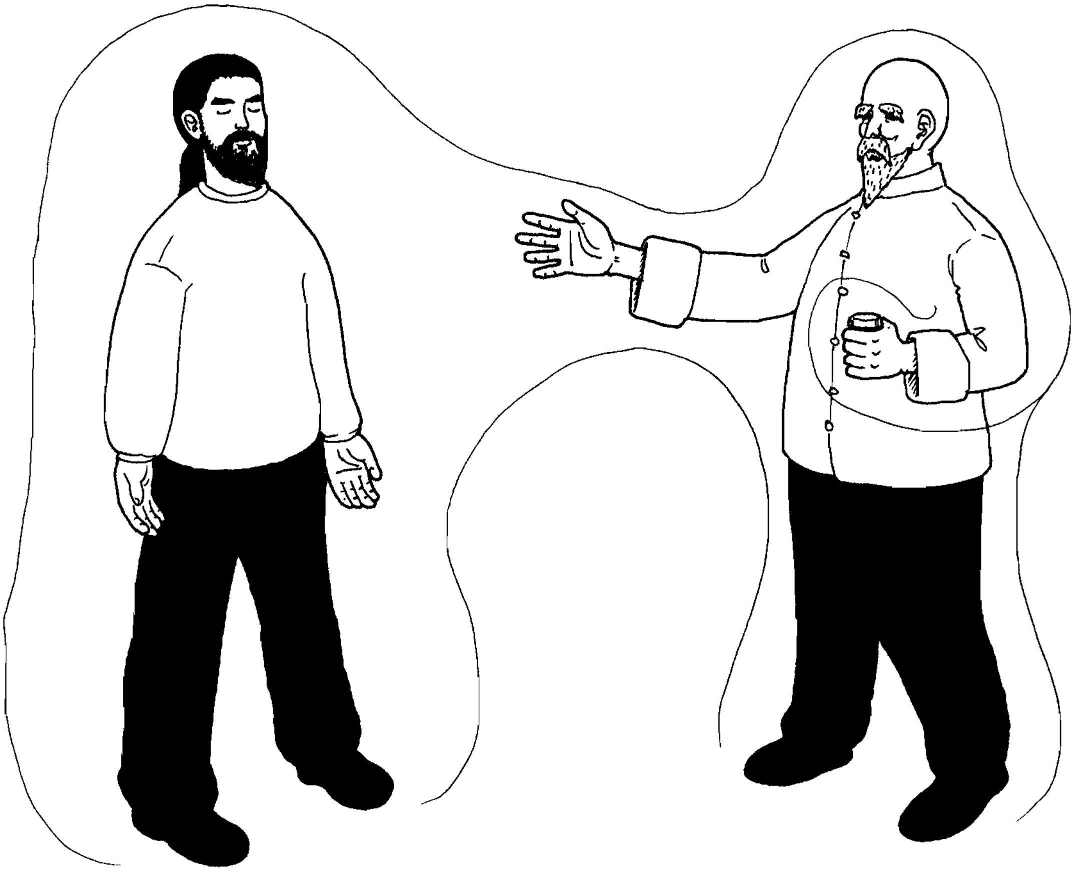

Figure 27.4. A Qigong Doctor Testing the Efficacy of a Specific Herbal Medicine for his Patient

Qigong and herbal therapies are most commonly used in the treatments of:

- Tumors
- Stomach cancer
- Lung cancer
- Hepatitis
- AIDS
- Stroke

COMBINING MEDICAL QIGONG THERAPY WITH WESTERN MEDICAL MODALITIES

Traditional Chinese Medical Qigong serves as a powerful tool in aiding Western medical approaches in alleviating patients' suffering. Because of its emphasis on mental, emotional, energetic, and spiritual approaches to healing, Medical Qigong therapy has been successfully combined as a complementary source of medical treatment with the following Western professional approaches:

- Pediatrics
- Geriatrics
- Gynecology
- Neurology
- Psychology
- Oncology
- Surgery

CHAPTER 28

PRECAUTIONS AND ETHICAL CONSIDERATIONS FOR THE QIGONG DOCTOR

AVOIDING ENERGETIC DEPLETION

Because Medical Qigong therapy consumes energy, as long as the Qigong doctor's quantity of energy is higher than that of the patient's, the doctor can transmit energy into the patient with no ill effect. If, however, the doctor is very tired or depleted and tries to treat a patient, the energy that the doctor is able to produce may become weaker than that of the patient.

It is important to note that energy currents flow from high to low (much like water). Therefore, the doctor may, if in a weakened state, absorb the Turbid Qi or disease energies from the patient. Any imbalance in the doctor's organ system can increase the doctor's vulnerability to foreign pathogens. These pathogenic sensations may also be transmitted into the body of healthy people, or other patients as well. Individuals who have not practiced Qigong, however, and whose energy circulation and channel points are not open, are not as prone to interference by these sensations. These individuals have a natural barrier, or to put it plainly, they are not sensitive to energetic fields of transmission and because of their denial system, cannot perceive energy sensations.

If the Qigong doctor already has a tendency towards a particular illness, exposure to pathogens may instigate or intensify the disease. If the doctor is suffering from any energetic Deficiencies, for example, the end result may be that the doctor performing the treatment may experience Qi deviations due to absorbing the patient's Toxic Qi. This means that more Qi and energy have been extended outside of the doctor's body than can be replenished, weakening the doctor's energetic constitution.

When treating patients the doctors receive and diagnose information from their Six Openings (also called the Six Windows of the Body). These Six Openings are described as follows:

- The eyes -- open for seeing energy patterns and colors,
- The ears -- open for hearing energy patterns and tones,
- The nose -- opens for smelling different odors,
- The tongue -- opens for tasting or flavor sensing,
- The body -- opens for energetic feeling sensations, and
- The Spirit (Heart/Mind) -- opens for energetic perceptions of emotions, thoughts, and spiritual insight.

While using the Six Openings for diagnosis, the doctor naturally depletes some life-force energy. Since all five organs (eyes, ears, nose, tongue, and body) are directed by the doctor's Spirit (Shen), the expenditure of energy is considerable. Any imbalance due to organ depletion will cause the doctor to absorb Turbid Qi from the patient. This Turbid Qi can move within the doctor's body causing Qi deviations, especially if it mixes with the doctor's Clear Qi.

1. If this happens within the doctor's Upper Dantian, the symptoms may include mental fatigue, dizziness, headaches, etc.
2. If this happens within the doctor's Middle Dantian, the symptoms may include emotional fatigue, discomfort in the Liver (hepatic region), etc.
3. If this happens within the doctor's Lower Dantian, the symptoms may include physical fatigue, cold in the extremities, discomfort in the lower abdomen, etc.

This is why Qigong doctors must constantly be refilling and rebalancing themselves energetically. It is important to note that if the Qigong doctor has indeed ingested the patient's pathogens, and Qi deviations have resulted, no Western doctor, acupuncturist, or herbalist can help him or her. Only the doctor with the Qi deviation (or with the help of another Qigong doctor) can change the Turbid Qi back to Clear Qi, by using divine energy and purging the body's Shen, Qi, and finally the Jing. This supports the doctor's Righteous Qi and fights the invasions of the patient's pathogens.

Emission of Qi should also be avoided if the doctor becomes sick, tired, hungry, overfed, deeply grieved, indignant, or drunk. Also, if the doctor is unable to stay in the present mind-frame he or she will inevitably fall asleep or "Shen out." To avoid this state of depletion, the Qigong doctor should:

- Cut back on the number of patients,
- Decrease treatment time—from 25 to 20 minutes maximum,
- Maintain a consistently healthy diet with supplements,
- Receive frequent exposure to fresh air and sunlight, and
- Perform the Microcosmic or other regulating Qi meditations in between patients.

Note: When treating patients with Medical Qigong therapy the doctor should never wear the color red. The color red is too hot and has a dispersing affect on Qi, which tends to scatter the body's energetic fields, distorting the Qigong doctor's Qi emission.

PREVENTING THE INVASION OF TURBID QI

When Turbid Qi enters the doctor's body, it interferes with normal Qi circulation, causing disorders in part of, or in all of the energy circulatory system. In mild cases, Qi may stagnate in a certain areas of the body (i.e., shoulder, arm, chest, or back), causing physical and mental symptoms.

- Physical symptoms may include: tingling, pain, cold, contraction, heaviness, soreness, and distention and stuffiness in the chest; and
- Mental symptoms may include: interference with mental activity, causing dizziness, headache, heaviness in the head, vexation, and restlessness.

In severe cases, the doctor may experience the same symptoms as the patient, but upon physical examination there is no verification of any existing disease. The doctor should therefore be on the alert to avoid any danger of developing symptoms from the patient after Turbid Qi is perceived.

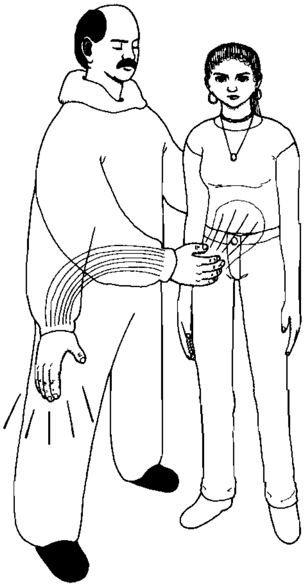

Figure 28.1. In order to disperse the patient's Evil Qi, the Qigong doctor exhales while slinging the toxic energy into the Earth.

It is very important for doctors who treat patients with Medical Qigong therapy to possess the ability to prevent and expel Turbid Qi so that it will not disturb their own energetic activities. Qi deviations are most often seen in those individuals who treat patients after they have gained some knowledge of Medical Qigong but have no experience, though they can occur even to some veterans of Medical Qigong practice. The interference by Turbid Qi is often an important factor in the the doctor's health.

When the interference by Turbid Qi is perceived, the doctor should use proper hand manipulations and readjust energy activities to expel the Turbid Qi immediately. If the Turbid Qi invades the doctor's fingers, or penetrates into certain channels or points, the doctor should guide the Qi to the specific points, channels, and infected areas and then relax and shake the hands to discharge the Turbid Qi while exhaling.

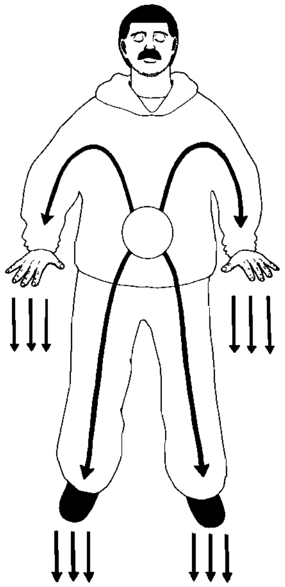

Figure 28.2. The Qigong doctor gathers energy into his or her Lower Dantian and expels any and all pathogenic invasions out through the hands and feet.

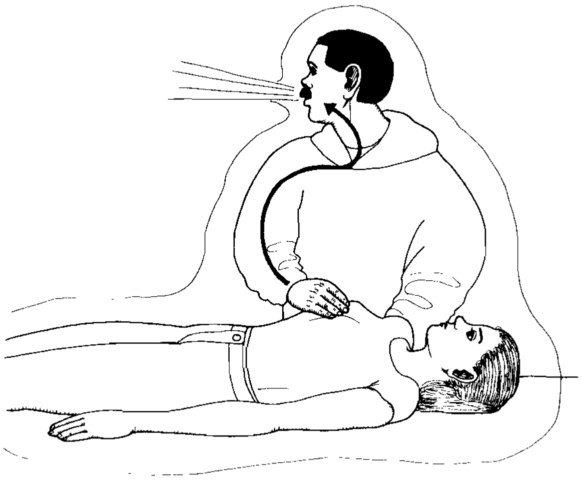

Figure 28.3. To avoid absorbing the emotions from an energetic discharge (released from dissolving an energetic cluster), the Qigong doctor turns his or her head and expels the patient's toxic emotions.

The doctor should be able to stop the Turbid Qi before it reaches the Dazhui GV-14 point on the upper back, Tiantu CV-22 point at the base of the throat, Quepen St-12 points on the front of the shoulders, and the Fengfu GV-16 and Fengchi GB-20 points on the back of the head. It is important to expel the Turbid Qi when it has reached the fingers, wrists, elbows, and at most the shoulders.

When this occurs the doctor should drive the Evil Qi out and regulate his or her Qi circulation as soon as possible. There are several ways to do this:

- Exhale while swinging the arms toward the ground, driving the Evil Qi deep into the center of the Earth. When treating tumors, be careful to disperse the pathogens out your fingers and do not allow the toxins to stick to your palms (Figure 28.1).
- Exhale while "blowing Qi" (extending Qi) out of the palms (Pc-8) and feet (Kd-1) (Figure 28.2).
- Exhale out of the mouth (but never over the patient, or towards anyone else). A shift in breathing is a common reaction when the emotions are discharged, or when an energetic cluster is reached (Figure 28.3).

GUIDELINES FOR THE MEDICAL QIGONG DOCTOR

1. The Qigong doctor must avoid practicing Medical Qigong in wet clothes. If the doctor perspires a little when training, he or she should avoid standing in the wind; if the doctor perspires profusely, practicing should be discontinued as it is liable to damage the doctor's Yin Qi.
2. The doctor should not go to the toilet within half an hour after practicing in order to avoid developing a conditioned reflex. This will help prevent the leakage of vital energy.
3. Patients suffering from cancer may affect the doctor's Qi circulation more seriously. Care should be taken when treating these patients, or avoid treating them completely, until sufficient energy has been cultivated, and sufficient experience has been gained.
4. The doctor must always wash his or her hands

with soap after each patient's treatment to prevent the energetic transference of diseases. It is important to note that washing the hands in cold water immediately after Qi emission is harmful to the doctor's energetic field because of the sudden shock to the system. Washing with hot water immediately after treatment, opens the channels on the doctor's hands and arms, sending the Toxic Qi deep into the doctor's body. To avoid these complications, the doctor should wait several minutes after completing the treatment, before washing his or her hands.

While waiting, the doctor should touch his or her finger tips together and focus on the Lower Dantian, allowing the Qi to settle back into its origin (minimum of three minutes). The doctor then washes up to the elbows with water that is room temperature to reduce the possibility of absorbing the patient's Turbid Qi.

If the doctor is not able, during treatment, to expel the Turbid Qi right away because of some distraction, or if the doctor's internal energy is not substantial, the Turbid Qi may enter his or her body through the fingers traveling up the wrists, elbows, or entering the Baihui GV-20, Tanzhong CV-17, Fengchi GB-20, Yintang (Extra Point) and Yongqauan Kd-1 points. If this begins to happen, the doctor should not continue to extend energy, but should stop the treatment immediately and perform a "closing," expelling the Turbid Qi by flicking the fingers, wrists, and shaking the arms until all discomfort has been eliminated. If any Turbid Qi still lingers, the doctor should practice Medical Qigong exercises to regulate the energetic activities.

5. It is important that doctors, as well as patients, avoid strenuous activities (such as running immediately after Medical Qigong exercises) to prevent the wrong flow of vital energy or pain in the legs. It is always better for the doctors and patients to be active first (running, swimming, forms training, etc.) and then cool down with the Medical Qigong Quiescent exercises (unless otherwise prescribed).

WU ZANG (FIVE ORGAN) MEDITATION FOR ENERGETIC PROTECTION

Even after the treatment of diseases, the doctor must perform certain Medical Qigong exercises to prevent and expel any undetected pathogenic Qi. Otherwise the doctor may develop the symptoms of the illness of the patient in corresponding locations. For instance, if the patient has a headache or pain in the hepatic area, the doctor may also acquire a feeling of discomfort, numbness, itching, and pain in the head and hepatic region. Therefore, in clinical practice, a deep understanding and correct judgment of the pathogenic Evils is compulsory.

Today in China, the Wu Zang meditation is still taught to Qigong doctors for clinical use to prevent invasion from the patient's pathogenic Qi. This meditation requires the doctor's Shen to guide the Qi of the Five Yin Organs (also known as the Wu Jing Shen, see Figure 28.4), to make the body's Qi substantial. The ending of the meditation, which causes the doctor's Qi to be reabsorbed into the Taiji Pole, allows the transformed Qi to flow into the appropriate organs, strengthening the doctor's energetic fields.

When a Qigong doctor enters the space of a diseased patient, the potential for absorbing the patient's pathogenic Qi is high. Noting this potential for harm, the ancient Chinese Qigong masters developed certain meditations to protect themselves from the patient's diseased or Turbid energy.

The Yellow Emperor's Classics on Internal Medicine states that before treating patients of any kind, the doctor should first perform the Wu Zang Energy meditation. It encouraged doctors to practice this meditation before treating patients in a clinical environment and before entering the treatment room, because it surrounds them with a strong field of protective energy (Wei Qi).

The Wu Zang meditation focuses on six directions (north, south, east, west, Heaven and Earth). It is very powerful and can be used to gather environmental Qi at the beginning of each season. The focus of the meditation is to extend each of the five organs' energy far into the horizon to gather the Qi. When practicing the Wu Zang

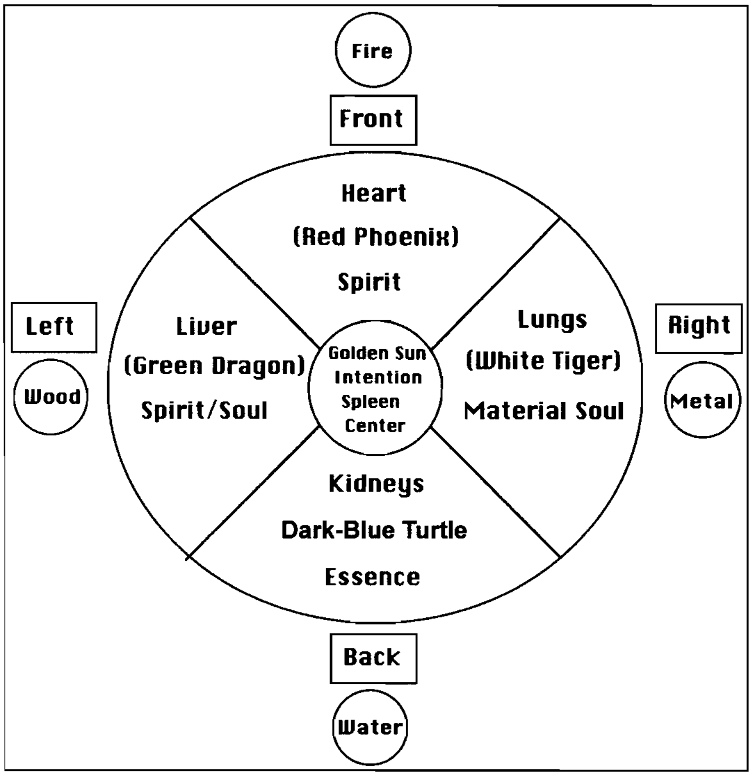
Fire
Front
Heart
(Red Phoenix)
Spirit
Left
Wood
Liver
(Green Dragon)
Spirit/Soul
Golden Sun
Intention
Spleen
Center
Lungs
(White Tiger)
Material Soul
Right
Metal
Kidneys
Dark-Blue Turtle
Essence
Back
Water

Figure 28.4. The Wu Jing Shen Relationship of the Wu Zang Meditation

meditation the Qigong doctor should:

- face east during the spring equinox;
- face south on the day of the summer solstice;
- concentration on the center of the Earth during late summer, before the fall equinox;
- face west on the day of the fall equinox;
- face north on the winter solstice.

Start by focusing on the center of the body. Imagine opening up the Baihui point (at the top of the head) and begin to pull in the Qi from the Heavens. Imagine the divine light as a bright shining white light, illuminating and filling your entire body through the Baihui. Feel the body radiating this divine white light energy. Next, imagine this white light energy coalescing into the body's center core, forming an energetic tube of Heavenly energy that extends from the Baihui point to the Huiyin point. The center core vibrates and resonates with the divine white light energy.

1. Now, imagine a golden yellow mist of Qi arising from under the Earth and filling your body and connecting with the Spleen organ. Next, feel this golden light Earth energy envelop the center core of divine white light energy, merging together and synergizing. This represents the energy of your Yi (Intention), to root and stabilize your power.
2. Begin to focus on your Heart and imagine a portal opening and the Qi flowing out of your Heart like a red swirling wind in front of you, full of power, protecting you with your Shen and the fire of a red phoenix. This represents your innate spirit, alive, graceful yet powerful.
3. Focus your attention on your back and the Kidneys and Mingmen area. Imagine a portal opening and the Qi flowing out of your Mingmen behind you like water. Out of this water grows an enormous dark blue turtle,

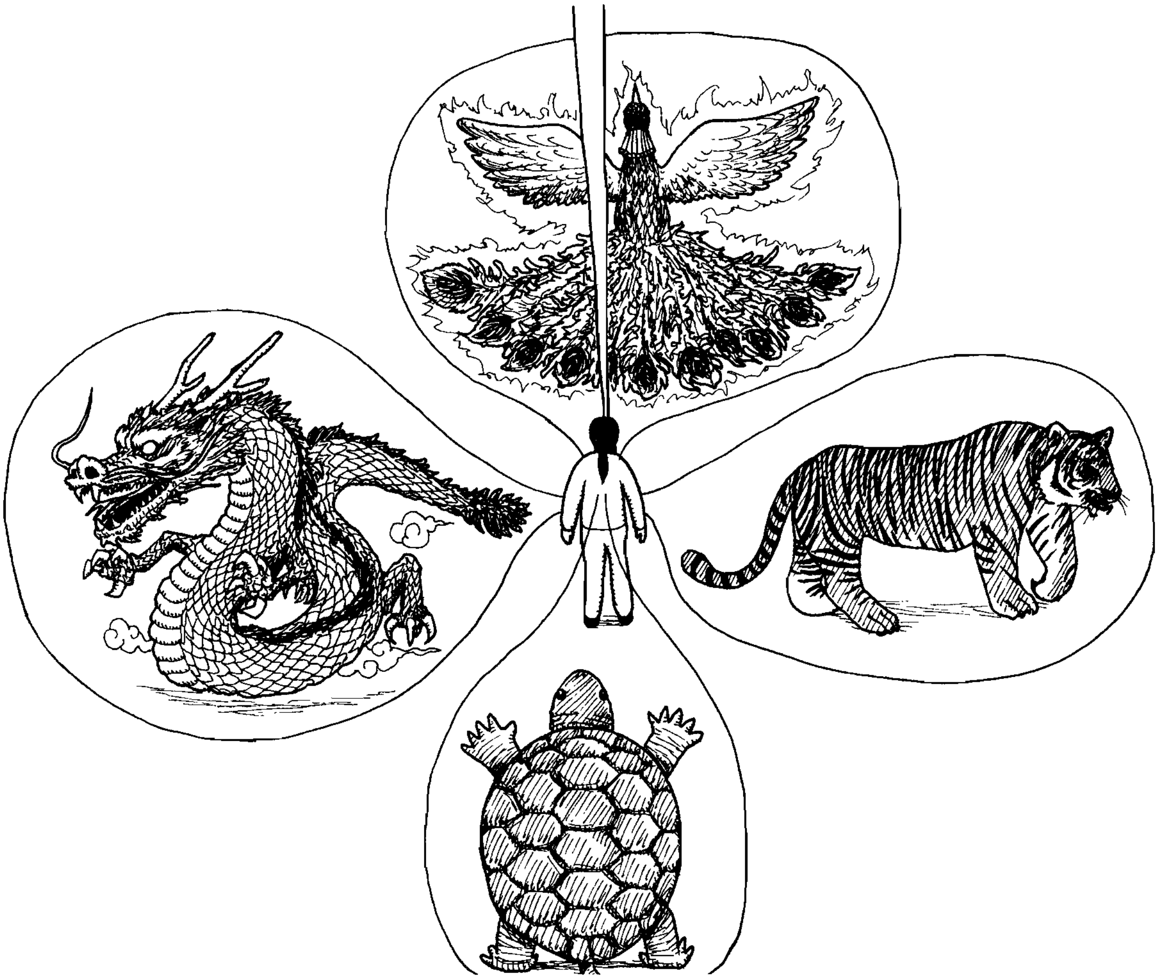

Figure 28.5. The Wu Zang Meditation

whose shell protects you like a mighty shield. This represents the energy of your Zhi (Willpower), and the Jing and Zhi of all your ancestors, backing you and supporting you.

4. Place your attention on your Lungs and visualize a portal opening on your right side of your body, under the right ribs. Imagine the Lung Qi flowing out of to the right side of your body like steam, forming a white tiger, as strong as steel. This represents the Po and the body's animal nature, that guards and protects you with an animal passion for survival.
5. Next, place your attention on your Liver and visualize a portal opening on your left side of your body, under the left ribs. Imagine the Liver Qi flowing out of to the left side of your body like steam, forming a green dragon, sinewy and resilient as bamboo. This represents the Hun and the body's divine nature, guarding and protecting you with a spiritual passion for victory.
6. Each animal begins to rotate to the left, protecting, stalking, and defending the previous animal's position (Figure 28.5). Slowly begin

to circle these energies, then increase their speed (like a mighty wind) counterclockwise around your body, blending these colors together, to form an energy bubble.

After forming a protective rainbow hue around your body, draw the energies back in through the Baihui point and return the energy of each organ color back to its origin (the color red returns back to the Heart, dark blue to the Kidneys, white to the Lungs, and green back to the Liver). As the colors return back to their organ of origin, imagine steam (white light) flowing out of the pores and filling up the energy bubble created by the animal rotation. This forms a solid connection between the body's internal organs and the body's external field of energy.

7. When the protection of the body with the Five Elements is done, imagine that there are many sparkling lights like the Big Dipper above your head. Then you can enter the room of the patient, safe and protected.

Keep in mind that you can send the five organ energy outside of your body instantly, any time you need protection. Some schools teach the doctor to extend his or her energy like mist out through the eyes, enveloping the body with the five organ colors (green, red, yellow, white, and dark blue/indigo).

CREATING A PROTECTIVE ENERGETIC FORCE FIELD

This second meditation also establishes a powerful energetic force field, needed to protect the doctor from any attack or attachment of external pathogenic Qi. It consists of wrapping the doctor's external fields of Wei Qi to increase the energetic field's power.

This is similar to wrapping a magnet with coils of wire to increase its magnetic field (Figure 28.6).

1. Begin from a standing Wuji posture. Focus on all the body's energy collecting into the Lower Dantian. Concentrate the Qi in the Lower Dantian, condensing it into a dense ball of white light energy. Imagine this ball of white light energy exiting the body through the navel and circling the body, spinning in a counterclockwise direction nine times, wrapping the Belt Vessel, Mingmen GV-4 and Shenque CV-8 points.
2. Move the energy ball down to the perineum and continue circling the lower body in a counterclockwise direction nine times, spinning and wrapping the Huiyin CV-1 and Changqiang GV-1 points.
3. Move the energy ball down to the knees and continue circling in a counterclockwise direction nine times, spinning and wrapping the ascending Yin channels and descending Yang channels of the legs.
4. Move the energy ball down to the ankles and continue circling in a counterclockwise direc-

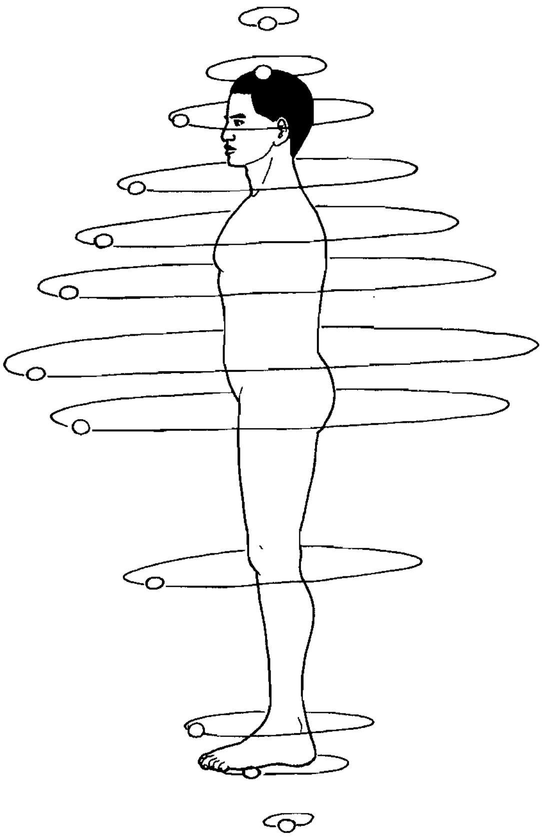

Figure 28.6. Creating a Protective Energetic Forcefield

tion nine times, spinning and wrapping the ascending Yin channels and descending Yang channels of the legs.

5. Move the energy ball down to the bottom of the feet and continue circling in a counterclockwise direction nine times, spinning and wrapping the Yongquan Kd-1 points.
6. Move the energy ball several feet into the ground and continue circling in a counterclockwise direction nine times, spinning and wrapping the body's Earth Transpersonal point. This energetic point (under the ground) is responsible for establishing the body's rooted connection to the Earth and for absorbing the Earth energy into the body's internal organs, channels, and collaterals.
7. Next, imagine the energy ball slowing down and stopping its rotation, pulsing for several heartbeats, then reversing its energetic rotation, and beginning to spin in a clockwise direction nine times. The direction and flow of the energetic ball now reverses its orbit, ascending and circling the same areas and points nine times, until it returns back to the navel.
8. From the navel move the energy ball up to the solar plexus area and continue circling the body in a clockwise direction nine times, spinning and wrapping the Jiuwei CV-15 and Zhiyang GV-9 points.
9. Move the energy ball up to the Middle Dantian at the center of the chest, and continue circling in a clockwise direction nine times, spinning and wrapping the Heart, Shanzhong CV-17 and Shendao GV-11 points.
10. Move the energy ball up to the throat and continue circling in a clockwise direction nine times, spinning and wrapping the Tiantu CV-22 and Dazhui GV-14 points.
11. Move the energy ball up to the Upper Dantian at the Third Eye area, and continue circling in a clockwise direction nine times, spinning and wrapping the Yintang and Naohu GV-17 points.
12. Move the energy ball up to the top of the head and continue circling in a clockwise direction nine times, spinning and wrapping the Baihui GV-20 point.
13. Move the energy ball several feet above the head and continue circling in a clockwise direction nine times, spinning and wrapping the body's Heavenly Transpersonal point. This energetic point (above the head) is responsible for establishing the body's extended connection to the Heavens and for absorbing the divine energy into the body's internal organs, channels, and collaterals.
14. Next, imagine the energy ball slowing down and stopping its rotation, pulsing for several heartbeats, then reversing its energetic rotation, and beginning to spin in a counterclockwise direction nine times. The direction and flow of the energetic ball now reverses its orbit, descending and circling the same areas and points nine times, until it returns back down to the navel. The body should now be encircled, above and below, with several haloes of white light energy.
15. Finally, imagine the white light energy ball absorbing back into the body through the navel. As the energy ball descends back into the Lower Dantian, imagine energetic steam filling the entire body.
16. Once the body is completely full, the steam begins to exit the body through the pores and begins connecting and filling the space outside the body with white light energy, forming an energetic cocoon. This cocoon creates an energetic force field which protects the Qigong doctor from pathogenic invasion.

CLINICAL ETHICS

The study of clinical ethics refers to the morality of the Qigong doctor and the ability to set boundaries. The study of morals in the therapeutic relationship involves reflecting upon the ideals the doctor assumes as a professional care giver and the enforcement of these professional standards in his or her actual behavior. Both the morality and the self reflection are engaged, both consciously and unconsciously, to align the doctor's values with his or her actions.

Ethical behavior stems from the doctor's internal values and actions (emotionally, cognitively, and spiritually) and his or her sense of external connection (physically, existentially, and socially). The Qigong doctor's behavior should be always congruent with high ethical standards and in harmony with his or her values, intuition, knowledge, and feelings.

The importance of establishing ethical standards for the doctor-patient relationship is to avoid those behaviors which prove to be counter therapeutic, and encourage those that are most effective therapeutically for the patient. The patient, as well as the doctor, rely upon each others' trust and boundaries. It is the Qigong doctor's responsibility to create a context for the patient, in which the patient can make a change in his or her life in safety.

When patients are treated energetically, they relax into an altered state of consciousness. This causes the patient's boundaries to become relaxed or nonexistent. Because Medical Qigong therapy often reveals intimate knowledge of the patient's life and life-style, there is a danger of enmeshment, i.e., of co-dependent relationships forming (see Chapter 19).

When working with patients, doctors sometimes encounter intense physical, mental, emotional, and spiritual reactions from the patients. These experiences tend to disperse the patients' boundaries and can confuse the roles between the doctor and patients, intensifying transference and triggering countertransference.

When both the doctor and patient have a personal relationship with a divine higher power, it naturally brings them into a state of divine healing and spiritual wisdom. The acknowledgment and honor of this divine relationship should establish ethical boundaries. This is especially true as the doctor begins to see the bigger picture of how his or her intentions and actions (in relationship with patients) affect others in a rippling outward motion. It is important that doctors also see the effect on themselves when they take certain actions towards others, including their patients.

CONSCIOUS AND SUBCONSCIOUS CLINICAL INTERACTIONS

Clinical experiences are divided into two states of interactions: the body / mind and emotional / spiritual.

- The body / mind interaction deals with ordinary states of consciousness, such as: talking, thinking, analyzing, working, interacting with the material world and other human beings. These different qualities of interactions help the doctor to define and categorize, educate, and establish his or her solid belief structures.
- The emotion / spirit interaction deals with subconscious states of communication, such as: perceiving, sensing, intuiting, interacting with the spiritual / metaphysical world, as well as with energetic fields. Every day people go into states of light trance (daydreaming, spacing-out, etc.) allowing their focus of attention, thoughts, feelings, sensations, and intuitions to interconnect with their environment and that of the divine. This allows access to information and facilitates the healing and understanding of themselves and their life experiences.

Because most wounding involves an emotional / spiritual trauma at the time of the incident, the patients must relive this trauma during their healing process. Through the emotional / spiritual interaction between the Qigong doctor and his or her patients, ingrained habits of thought, feelings, and perceptual understandings begin to recede, diffuse and break down. This causes the patients to find a new understanding of certain traumas and allows them to reclaim disconnected parts of their past (soul retrieval).

Each time patients transcend to a new level of understanding of their fears and belief structures, they change, mature, and expand to adjust to the new mental, emotional, and spiritual growth. The moment that patients break out of their old boundary structure, feelings of emotion rush through their body / mind, followed by a sense of either expansiveness or disorientation as they find themselves in unfamiliar territory.

It is not useful for patients to relive their emotional traumas without the awareness, understanding, and insight of the process they are undergoing. To relive the trauma without this awareness leads to re-traumatization. That is why the Qigong doctor coaxes the patients through the "five stages of emotional healing" (see Chapter 19). The patients can then experience their past emotions in the present time, and become free of them through present awareness, rather then repeating the past through a continuous loop of repression and denial. In the clinic, this is known as "emotional time traveling" and refers to the practice of allowing the patients the experience of reliving past traumas through their presently increased state of awareness and knowledge. At this point of transition, the patients will not only experience the emotions and sensations from their past, but may also experience the motivations and emotions of the perpetrators as well.

The empathic knowledge of the perpetrator's emotions is due to the victim's identification process with the abuser. This phenomenon is well recognized in helpless hostage situations, and is a major contributor to childhood victims of physical, sexual or emotional abuse "acting out" their repressed fears, angers, and pain, in similar ways to the manner in which they were abused.

ESTABLISHING DOCTOR AND PATIENT AGREEMENTS

For the protection and integrity of both the doctor as well as the patients, it is important that a description of the clinical procedures that are to be used, their physical and energetic boundaries, as well as their limitations, be firmly established. The following seven principles are used to establish the foundation of clinical ethics.

1. The Qigong doctor will cause no harm (physical, mental, emotional, or spiritual). The Qigong doctor will maintain a high standard of skill, knowledge, and professional conduct, offering his or her services without favoritism, prejudice, or discrimination with regards to race, religion, nationality, gender, or sexual orientation.

   The Qigong doctor is committed to the patients' education and personal development of practicing Medical Qigong methods that relieve pain, suffering, and other symptoms of illness. The doctor will exercise his or her best judgement to determine if the therapeutic techniques prescribed for patients are contraindicated, dangerous, or not in the best interest of the patients.

   The doctor will never make negative suggestions, criticize, or tell the patients that they will get worse or are in bad shape, as that destroys the patients' hopes and is counterproductive to healing (this is known as "clinical hexing"). Clinical hexing will cause patients, who trust the doctor, to sabotage their own healing potential by succumbing to the doctor's particular beliefs about their conditions.

   One example of trusting a doctor's diagnosis to ruin, happened last year in 1999, when two patients' lab tests were accidently switched. The healthy patient, whose lab test had been switched, was informed by the Western doctor that she had terminal cancer; and the patient with terminal cancer, on the other hand, was given a clean bill of health. The healthy patient died several months later as a result of believing the doctor's diagnosis of having a terminal condition; whereas the patient with terminal cancer (believing that there was nothing wrong) resumed a normal, active life-style, and the patient's cancer went into remission.
2. The Qigong doctor will keep all sessions in confidence. The Qigong doctor will keep all patient information strictly confidential, in keeping with ethical and legal standards held by other health care professionals. This allows patients to experience the treatment in safety and give honest feedback.

   Exceptions to the rule of confidentiality are as follows: when the patient is a danger to self (intending suicide), or others (intending homicide), when the patient has a life threatening emergency, in cases of suspected child abuse (sexual, physical, or neglect), and elderly abuse (sexual, physical, neglect or fiduciary).

3. The Qigong doctor will first receive informed consent from the patient before starting the treatment, or informed consent of the patient's legal guardian. Before treatment, the Qigong doctor explains to the patient the therapeutic goal of the treatment, and what the patient may energetically and emotionally expect to experience. The doctor will respect the patient's boundaries at all times, before, during, and after each treatment. At times the doctor may have to impose boundaries for patients who cannot do so for themselves, in order to avoid total enmeshment, which would endanger the professional relationship.
4. The Qigong doctor and the patient will tell the truth to each other. No part of the Medical Qigong diagnosis should ever be hidden from the patient, yet the truth should always be delivered with hope, since no one can be certain of the future.

   The Qigong doctor should never make unrealistic, inaccurate, or false claims about Medical Qigong therapy, or give the patient false hope. A realistic assessment of the patient's condition should always be conveyed, and if need be, the patient should be referred for additional medical or psychological modalities of treatment.
5. The Qigong doctor will honor all agreements established with the patient. It is important for the patient to be able to rely on the doctor's trust. Because this is a critical factor, the doctor is never allowed to borrow money, cars, etc. from patients.
6. The Qigong doctor and the patient will not act sexually or romantically with each other. The Qigong doctor will not instigate or tolerate sexual advances while interacting with his or her patients. Although there are times that the patient, or doctor, may experience certain attractions towards each other, acting out on these feeling will cause an emotional/sexual entanglement which will be detrimental to the patient's healing.
7. The Qigong doctor and the patient will agree on the time, place, duration of treatment, and fee. The Qigong doctor must be consistent in maintaining an established clinical treatment time and location. This instills a stable security for the patient and doctor. If this boundary is not maintained, patients without strong boundaries will take advantage and hurt both doctor and themselves. The Qigong doctor is responsible for giving adequate advance notice (two weeks) to the patient of any and all treatment or fee changes. If, for example, the original agreement was for six or fewer sessions and the doctor changes fees after the first session, the original agreement must be upheld (new patients however, will be charged the new fee).

POWER DYNAMICS AND ETHICAL BEHAVIOR

In any clinical setting there is an active power dynamic at work. Generally it is the doctor who maintains the power over the patient, as the purpose of the patient's visit is for help or assistance in specific healing. As the patient has come to the doctor with expectations, it is important for the doctor to be congruent with these expectations. These expectations set both the standard of the patient's healing and the power dynamics of the relationship. Through the power dynamics of the doctor-patient relationship there can develop three classical problems: Transference, Counter-transference, and Emotional-Sexual Entanglement.

1. Transference is the process whereby a patient unconsciously transfers feelings, thoughts, beliefs, and patterns of behavior that had been previously experienced towards others onto the doctor.
2. Counter-transference is the process whereby a doctor loses his or her objectivity and unconsciously transfers feelings, thoughts, beliefs, and patterns of behavior to the patient.
3. Emotional-Sexual Entanglement is the process whereby the doctor and patient both lose their objectivity and transfer emotional-sexual feelings, thoughts, beliefs, and patterns of behavior onto each other.

In order to prevent these types of enmeshment and co-dependent relationships from happening, it is important for the doctor to constantly maintain his or her boundaries. The doctor and patient both must remember that it is the patient who actually does the healing in energy work, the doctor merely creates the situation in which this energetic transformation can transpire.

PITFALLS

There are certain pitfalls the Qigong doctor should be aware of and may experience, if a boundary has already been, or is beginning to be violated. These experiences are described as follows.

- The doctor begins to identify with the patient's problem and feels the need to discuss his or her own similar problems with the patient.
- The doctor begins to become emotionally attached or attracted to a patient sexually and feels the need to act on it (i.e., asks for a date, inappropriate touching or sexual contact).
- The doctor begins to feel the need to go out of the way above and beyond the normal course of responsible treatment to meet a patient's needs.
- The doctor begins to treat the patient for free.
- The doctor begins to find him or herself becoming overprotective, or worrying about a patient at inappropriate times.
- The doctor begins to need validation from his or her patient (this approval can be either as a healer, emotionally, or sexually).
- Arranging an exchange of services with a patient can often lead to the patient feeling abused or hurt.
- Allowing the patient to build up a large bill to be paid back later is unethical and may lead to legal action.
- Disclosing too much personal information regarding self (the doctor) in order to help the client can make the patient feel responsible for the doctor's well-being. Self-disclosure can be used very sparingly when the patient can benefit from solutions that worked for the doctor. It is safer, however, to pretend that these situations came from someone else.
- The doctor does not listen carefully to what the patient is saying, thinking that he or she knows what the problem is in advance.

It is important for the doctor to remember that the patient's emotional / spiritual safety and health must always come first. To avoid sinking into one of these pitfalls, the Qigong doctor is encouraged to constantly monitor his or her own actions and motivations. This is a broad and important topic that is often neglected in professional training and education. Because it is well dealt with in Kylea Taylor's book entitled, *The Ethics of Caring*, I encourage the reader to refer to this work for in-depth treatment of this subject.

CLINICAL BURNOUT AND MEDICAL QIGONG THERAPY

One classic problem all healers encounter is clinical burnout. The condition known as clinical burnout results from chronic job stress, and may sometimes occur to the Qigong doctor. This condition is characterized by extreme frustration, disappointment, physical and emotional exhaustion, and sometimes physical illness. The result of this condition is the doctor's loss of concern for the patients, or the loss of fulfillment with his or her performance as a healer. The Qigong doctor is susceptible to experiencing clinical burnout, especially if he or she feels that:

- There are too many or too few patients;
- Too little time is spent for recreation;
- There is too much stress in his or her life;
- Not enough attention is placed on self care and personal psychological growth; and
- Not enough attention is placed on a personal spiritual path and practice.

If the Qigong doctor's life revolves primarily around the clinic (which is usually the case), he or she will disproportionately come into contact with the patients' pain and emotional traumas, as well as be chronically bombarded with toxic energetic discharges released from the patients.

Sometimes patients, in order to heal their intense emotional traumas, require an enormous amount of spiritual, emotional, and mental support from the doctor. If the doctor does not carefully monitor his or her own energetic reservoir, this may result in Qi depletion.

If Qigong doctors expend all their spiritual/emotional energy from intense personal interaction and enmeshment with their patients during treatment, the toxic interaction will deplete the doctors, resulting in complete physical and mental exhaustion. If Qigong doctors become depleted in the middle of a treatment, they may abandon their Upper Dantian (psychic) connection with patients (spiritually, emotionally, and/or energetically), and begin to suddenly express feelings of criticism, anger, disappointment, or grief. This reaction occurs when the Qigong doctors' own internal organ energy, depleted and worn, becomes susceptible to absorbing the patients' powerful emotional discharges. The problem of clinical burnout usually stems from energetic boundary invasion. Remember, Medical Qigong healing requires that the doctor heal him or herself first.

Often, a Qigong doctor's clinical burnout is primarily related to the unresolved personal issues and not the number of patients on the case load, or the amount of environmental stress the doctor experiences.

ENERGETIC BOUNDARY INVASION

In understanding clinical burnout due to energetic boundary invasion and its relationship between the doctor and patient interaction, it is important to review seven important factors which control and transform the body's energetic fields.

1. Energy exists on different planes. The body is like a complex hologram, composed of physical, mental, emotional, and spiritual planes of energy. The spiritual field of energy controls the emotional and mental energetic field, which in turn controls the physical energetic field. These energetic planes affect and control the energetic formation of the body's cells, tissues, organs, and organ systems.
2. Because the body produces heat, light, electricity, magnetic energy, and resonate vibrations, these natural energetic releases create a basic Yin (negative) and Yang (positive) polarity, which is found in all manifestations of energy.
3. Everything in the body is in vibratory motion. Various forms of tissue develop from different energetic vibrations. When the body's energy begins to slow down, it begins to take on a denser form of resonant vibration. The impediment of this resonant energetic vibration can stem from trauma or shock to the system (physical or emotional), resulting in energetic stagnations and disease.
4. The body's internal organs collect and store emotional energy, both positive and negative. As any suppressed, or "stuffed," emotional energy begins to culminate, the body will naturally (and quite frequently) disperse these internal emotional charges. They are generally released externally through: the Respiration, the Posture, and the Mental Dialogue.
   - The respiration, (which includes sighing and laughing), is released from the Heart, shouting released from the Liver, groaning from the Kidneys, crying from the Lungs, and singing from the Spleen.
   - The posture includes the body's physical structure adapting or conforming to an Expanded-Yang or Contracted-Yin energetic influence.
   - The mental dialogue includes both benign and malignant thought patterns.
5. The body has three primary energy centers known as Dantians. Each Dantian resonates at its own distinct vibration and is responsible for specific modes of interacting with others in interpersonal communication.
   - The Lower Dantian is considered the most physical energetic base. It is the most dense and has the slowest vibratory rate of the Three Dantians and communicates kinetically with the outside environment.
   - The Middle Dantian is considered the emotional and mental base. It has moderate resonances and communicates empathetically with the outside environment.
   - The Upper Dantian is considered the spiritual base and has the fastest and most delicate of the vibrational interactions. The Upper Dantian functions at an intuitive level.

6. When communicating, the body will interact from all five levels: physical, mental, emotional, energetic, and spiritual. This interactive communication will sometimes energetically fuse with the external environment, including people.
7. Our external and internal thoughts and feelings, as well as tissues and cells are all byproducts of energetic interactions, stemming from communication from within ourselves and our environment.

After reviewing the following facts, the Qigong doctor can understand why every time he or she connects with a patient, the patient's energetic fields will automatically be energetically ingested by the doctor. The doctor does this consciously in order to analyze the patient's emotional past, as well as present emotional outbursts. Although the doctor and patient's energies fuse together, the fields of energy around the doctor's internal organs generally protect the doctor from absorbing negative emotions. These energetic boundaries also extend out into the doctor's Wei Qi field.

Both the patient's and doctor's Wei Qi fields are fused during interaction and treatment. It is therefore imperative that the doctor's energetic boundary system maintain its structure (both internally and externally), yet be permeable while treating the patient. The doctor must not lose sight of his or her purpose. Any criticism aimed at the doctor should not be taken seriously, nor should any inappropriate behavior by the patient. The doctor's Wei Qi field should be at maximum strength. If not, the cords of energetic attachment flowing from both the doctor and the patient allow energetic emotional transference to happen. If the doctor, for example, has a preexisting Liver Heat condition, it will be further compounded by the inflowing of the patient's released of anger and rage. If and once this happens, the patient may feel relieved, but the doctor will become quite agitated.

STRENGTHENING THE DOCTOR'S ENERGETIC BOUNDARIES

It is important for the Qigong doctors to allow the patients' emotions to constantly flow through their own bodies, feeling, understanding, then releasing these feelings. If at any one point in this transition, a doctor (because of personal issues) suddenly stops the flow of emotions and begins to focus, or dwell on, a particular type of energetic charge, the patients' discharging emotions will invade the doctor's body.

To avoid energetic boundary invasion the doctor should maintain a regular Qigong meditation and exercise program. This program should include the following exercises and meditations.

1. Practice Qi dredging and emotional dispersing exercises and meditations to release your own emotional traumas.
2. Perform Qi and organ strengthening exercises and meditations to strengthen both your internal organ energies, as well as the three fields of Wei Qi. Since the emotions are stored within the body's internal organs, strengthening these organs will facilitate a stronger control of the body's emotional energy.
3. Practice Qi regulation and balancing exercises and meditations to balance the body's Yin and Yang, as well as internal and external energetic channels, organs, and organ systems. This helps to control the energetic surges released from internal emotional discharges.
4. Maintain emphasis on the energetic grounding and the divine hook-up meditations before treating patients. These meditations are important for establishing an emotional as well as spiritual boundary, and will assist you in avoiding clinical burnout and energetic boundary invasion.
5. Maintain emphasis on personal time, including any and all spiritual, social, and recreational times away from the clinic.

CAUSES OF SPIRITUAL DRYNESS

Sometimes clinical burnout can result from "spiritual dryness." This type of burnout can result from experiencing a "dry period" of spiritual insight and perception. It can occur during clinical observations and treatments, resulting in frustration for the doctor. Spiritual dryness is commonly due to one or more of the following five conditions.

1. If the doctor is tired, hungry, or sick, the fo-

cus will tend to be on the doctor's own particular needs instead of the patient's.

2. A disconnection from the divine is caused by guilt, lack of forgiveness (self-blame), shame, etc., resulting in a loss of spiritual vision. It is important for the Qigong doctor to have an open and honest connection with the divine to receive the more subtle spiritual insights. Without this type of transpersonal relationship, the Qigong doctor's ability to dissolve into the Wuji surrounding the patient's tissues can be hindered.
3. Rushing personal meditation time is caused by failing to relax during the preparation meditation practice (the divine hook-up). This condition is normally caused by worries or preoccupations due to hidden agendas, that disturb the doctor's Shen prior to meditation. It is important for the doctor to relax and return to a state of quiescent peace, in order to truly listen to the divine.
4. Getting into a rut is caused by the doctor performing repetitive Shengong meditations in a mechanical way (i.e., the same way an individual will routinely brush his or her teeth), thus losing the emotional and spiritual connection with the core self. The meditation then becomes meaningless and powerless (just words and no spirit). To avoid getting into a rut, the doctor should try to experience each meditation as a time of new beginnings and enlightenment, and not as a mandatory routine.
5. It is important for the doctor to be able to share emotional and spiritual insights with other colleagues in the same field of work, and when appropriate, with the patient. This sharing of spiritual insights, observations, fears, and triumphs renews the doctor's confidence in his or her spiritual gifts, and maintains a strong faith in the healing potential of Medical Qigong therapy.

SECTION VII

THE TREATMENT PRINCIPLES OF MEDICAL QIGONG THERAPY

CHAPTER 29

PRINCIPLES AND PARAMETERS OF MEDICAL QIGONG THERAPY

The foundational factors of the patient's state of health originate from their actions, thoughts, and feelings. A Qigong doctor, studying a patient's physical problems, understands that the origin and solution to these problems are to be found within a patient's recurring thought patterns, both conscious and unconscious.

The brain processes data consistent with what it is familiar with; patients will generally experience and accept information that coincides with their own belief system. Their consciousness then, creates its own reality, resulting in either health or disease.

Good health is a result of bringing the patient's own unique spiritual and emotional essence back into consciousness and increasing the awareness of the physical body. Both health and disease are created through this process which is the foundational base for clinical Qigong treatments.

THE FIVE METHODS FOR TREATING PATIENTS

Qigong doctors generally use five methods for treating their patients' diseases (Figure 29.1). These methods initiate an energetic restructuring within the patient's body. The five methods are described as follows.

1. The doctor uses his or her own Qi to initiate the healing.
2. Qi is drawn, by the doctor, from the surrounding natural environment (from the energetic fields of the Earth) to be emitted to the patients.
3. Qi is drawn, by the doctor, from the surrounding universe (from the energetic fields of the Heavens) to be emitted to the patients.
4. Divine energy is channeled through the doctor to treat the patients.
5. Homework is prescribed in the form of self-regulation exercises and meditations to support the healing process.

THE NINE STAGES OF TREATMENT

Using these five methods of treatment, the Qigong doctor initiates nine stages of healing within the patient's body. The healing process will always vary according to the belief structure of the patients, their acceptance and performance of the Medical Qigong prescriptions, and their willingness to change. The Qigong doctor's nine specific stages in treating a patient are described as follows.

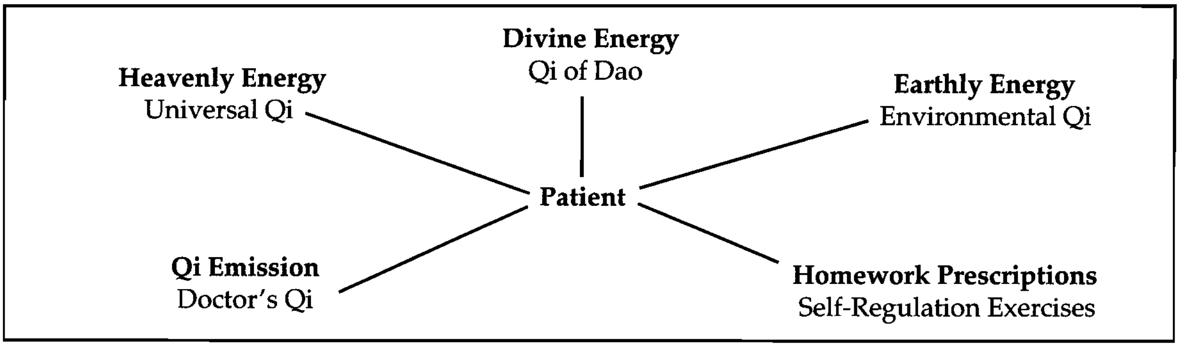

Figure 29.1. Five Methods Used For Treating Patients

The Subtle Wonders	Expensive or Contracted, Heavy or Weightlessness, Cold or Hot, and Tingling, or Vibrating Sensations
The Mysterious Wonders	Flashing Light Sensations, Sound Sensations, Smell Sensations, and Fainting (Syncope) Sensations
The Incredible Wonders	The patient will experience various colored lights, sounds, smells and other sensations according to the doctor's intention.

Figure 29.2. The Three Wonders

1. Activate, connect, and energize the patient's energetic field;
2. Dredge the patient's channels and collaterals;
3. Purge and eliminate any energetic Excess;
4. Tonify any energetic Deficiencies;
5. Regulate the Qi and Blood content of the patient's Yin and Yang organs (harmonize the organ Jing);
6. Regulate the patient's Prenatal and Postnatal Qi (harmonize the Qi of the Three Dantians and Taiji Pole);
7. Balance the patient's Yin and Yang Shen (harmonize the Heart/Mind and Emotions);
8. Seal the patient's external field of energy;
9. Give prescriptions.

THE THREE WONDERS OF MEDICAL QIGONG THERAPY

The cultivation of Qi and Shen through Medical Qigong therapy generates what is described as the Three Wonders. These Three Wonders consist of the three separate transformations of energetic abilities that the Medical Qigong doctor and, sometimes, the patients experience (Figure 29.2). The Three Wonders are explained as follows:

THE SUBTLE WONDERS

The Subtle Wonders consist of subtle changes inside and outside the body, known as the Eight Energetic Touches. These interactions of tissue and energy form the physical sensations and manifestations of Yuan Qi. When the doctor emits Qi to treat diseases, involuntary phenomena can occur instantly, or gradually, in local areas of the limbs, or the patient's whole body. In some cases the muscles stir slightly, while in other cases the limbs, or the body, will perform extreme, harsh movements. This results from the patient's external energy (Wei Qi) inducing spontaneous movements.

When the Qigong doctor emits energy to treat diseases, some patients produce, in local areas, electromagnetic, or stimulating, sensations of Qi similar to those felt in Qigong circulation/cultivation exercises and meditations. This includes: cold or hot sensations, feelings of compression or tugging, impressions of creeping or tingling, experiences of heaviness or lightness, and of floating or sinking, as well as other sensations. The Subtle Wonders result from energy in the channels circulating and responding to the doctor's intention.

Generally, the effects of Qi on the body are manifested in physical sensations and sensory phenomena during Qigong exercises. These phenomena are all manifestations of the free movement of Qi in the channels, collaterals, and inner body circulation.

The doctor may also experience these same sensations while extending energy into the patient. The doctor should be able to sense the patient's Qi circulation, density, direction, as well as his or her own energy. The sensations known as the Eight Energetic Touches of the Subtle Wonders are: Expansive, Contracted, Heavy, Light or Weightlessness, Cold, Hot, Tingling or Itching, and either Vibrating, Shaking, Rippling or Moving.

1. When the Qi is circulating freely within the body, the capillaries expand as the body's energetic field expands, resulting in the sensation of being very large, tall, and expanded.
2. When the Qi enters from the outside of the body to gather in the Lower Dantian, the patient experiences the sensation of feeling very small or contracted.
3. When the Qi sinks down (upon exhalation), the body feels heavy, condensed, and solid as a rock.
4. When the Qi rises upward (overflowing the channels), the body feels light and weightless.
5. When the Qi circulates through the Microcosmic Orbit (balancing the Qi of the Heart and Kidneys), the Kidney-Yang becomes sufficient to allow the Kidney Yin to well up; this results in a pleasant cool sensation.
6. When the Qi is vigorous and thermal energy is gathered, the parts of the body where the Qi passes will feel hot.
7. Once Qi circulates vigorously and passes through the body's obstructed channels, collaterals and minute collaterals, the result is itching or tingling of the skin and scalp. This is an energetic manifestation of the body's obstructed channels, collaterals, and minute collaterals being cleansed.
8. When the Qi passes through the channels, collaterals, and parts of the body it has opened, the patient feels electrical and creeping sensations in the extremities as well as the body, causing the tissues to vibrate, shake, ripple, or move, etc.

These are all normal phenomena of Qi cultivation and circulation. As long as the patient remains calm and present, these sensations will appear and disappear as spontaneously as they were formed.

THE MYSTERIOUS WONDERS

Patients may also experience elusive as well as dynamic transitions inside and outside of the body, called the Mysterious Wonders. These energetic changes are sometimes present, but are rare, and may be confusing to the patient. The sensations known as the Mysterious Wonders are: Flashing Light Sensations, Sound Sensations, Smell Sensations, and Fainting (Syncope).

1. Flashing Light Sensations sometimes occur in some patients during Qigong treatment. The photoelectric phenomena is due to the interaction of light with matter. In some cases, patients feel electric sensations in the limbs and body, and in other cases light patterns may be seen in different forms, most of which present as round, sheet light, or lightning patterns.

   One of the manifestations of the Qigong doctor's cultivation through Shen Gong meditations is a flashing light sensation which is seen and felt within the doctor's head and body. This glow may sway back and forth, pulse like a white neon light, or pierce into the doctor's brain, and illuminate the spinal column. It may also appear as a ball of white light rotating at high speed through the Microcosmic Orbit.

   This ball or beam of light may exist continuously at the Qigong doctor's Baihui GV-20 point throughout the duration of the meditation and several hours afterwards. The degrees of light sensations and their colors will vary in accordance with the doctor's skill level. Also, it is this beam of light that the Qigong doctor will later utilize for inner vision diagnosis.
2. Sound Sensations are sometimes experienced during treatment, such as a rustling, rumbling, or high-pitched sound.
3. Specific Smell Sensations may occur, such as the sweet scent of sandalwood, or fragrances of various flowers, or incenses.
4. Fainting (Syncope) can occur during Qigong treatment. A few patients may suffer from perspiration and accelerated heart rate, fol-

lowed by syncope (fainting). When this occurs, the doctor lightly presses the patient's channel points: Baihui (GV-20), Mingmen (GV-4), Jiangjing (GB-21), and Yintang (Extra Point). The doctor then grasps the shoulders and presses down on the GB-21 points with force. Finally, the doctor emits Qi along the Governing and Conceptual Vessels and leads the patient's Qi to the Lower Dantian. This allows the patient to recover consciousness.

THE INCREDIBLE WONDERS

The Qigong doctor's ability to have complete conscious and automatic control over the patient's reaction to energetic projections constitutes what is known as the Incredible Wonders. When the Qigong doctor takes control over the patient's energetic projections, he or she can willfully initiate the patient to experience various colored lights, sounds, smells and other sensations according to the Qigong doctor's will. This procedure is commonly known as Cultivating the Spirit to Generate Wonders.

When the Qigong doctor emits energy towards patients to treat their diseases, most patients will either instantly or gradually feel some form of energetic sensation from the increased life-force energy. The degree of energetic sensation felt by patients is in accordance with their emotional/spiritual sensitivity and degree of connection to their own physical body.

The sensations of Yuan Qi are the most common reactions felt by both the Qigong doctors and their patients. In rare instances some patients display no physical effects and experience no Qi sensations during treatment. These patients can still achieve good therapeutic results by improving the degree of their emotional/spiritual connection to their body.

In some cases, patients may even feel miserable or uncomfortable. When this occurs, they should continue training so that their Righteous Qi can overcome their Evil Qi. This ongoing battle will also affect their spiritual and emotional being. Fear usually arises at this point causing them to question their ability to continue in their practice. This is a normal transition and can be resolved through encouragement, patience and continual practice.

CONTRAINDICATIONS FOR MEDICAL QIGONG THERAPY

There are several contraindications that the Qigong doctor should be aware of before treating patients, or even prescribing Medical Qigong exercises. The following is a list of different personality types that may be encountered in the clinical settings that require special treatment modifications.

1. The very "nervous" type of patient should not practice Medical Qigong, or must be under strict supervision due to the patient's emotional instability.
2. The "extreme temperament" type of patient who has a fanatical type of personality should also avoid Medical Qigong training or remain under strict supervision. Patients of extreme temperament generally suffer from Excess disorders and compulsive overtraining. They are therefore prone to Qi deviations.
3. The "extremely introverted" type should be careful when practicing Medical Qigong meditations because strong emotions may be set into motion, and even certain mental disorders may be violently released that the patient may have been unaware of.
4. The "narrow minded" type will generally condense Qi into smaller areas of the body causing stagnations. Such patients should be instructed to relax and let go of their stress to allow the condensed energy to disperse.
5. Patients with the following conditions should be restricted to very specific Medical Qigong therapy: serious mental disorders, congestive Heart failure and severe organ failure, severe hemorrhaging, menstruation, nerve suppression, bone fracture, pregnancy, geriatric patients, young children, and cancer patients (Figure 29.3).

- Patients with mental disorders such as schizophrenia and bipolar (previously called manic depression) disorders (or individuals with a family history of such mental diseases), should refrain from training. The Qigong exercises and meditations could induce symptoms of these said disorders, or could enhance

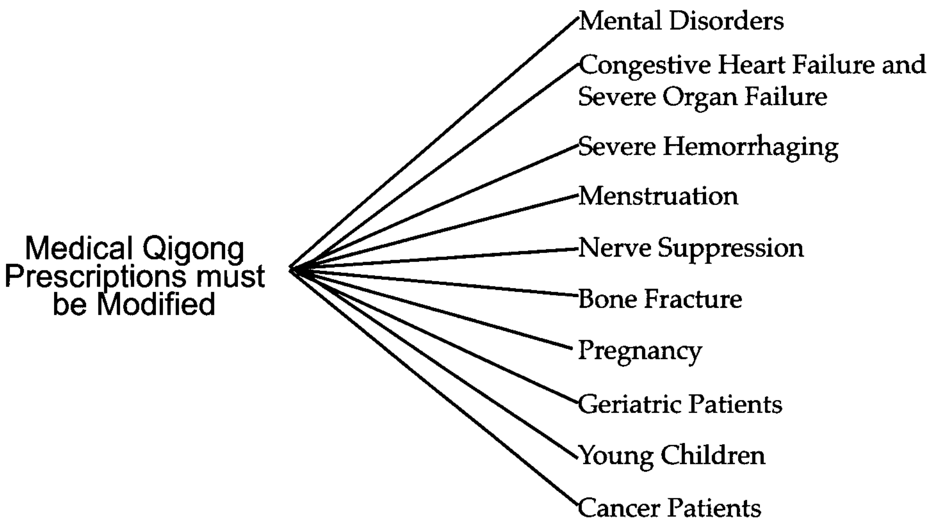
Medical Qigong Prescriptions must be Modified
Mental Disorders
Congestive Heart Failure and Severe Organ Failure
Severe Hemorrhaging
Menstruation
Nerve Suppression
Bone Fracture
Pregnancy
Geriatric Patients
Young Children
Cancer Patients

Figure 29.3. Restricted Conditions for Prescribing Medical Qigong Exercises and Meditations

the symptoms due to the patient's mental and emotional instability.

- Patients with congestive Heart failure and severe organ failure (any internal organ that has experienced an energetic collapse), require that the treatment be specifically oriented towards this condition.
- Patients with severe hemorrhaging require prescriptions that prevent the leakage of Blood and Qi, since Blood and Qi flow together.
- Women during their menstruation cycle should not store Qi in their Lower Dantian. Any Blood and Qi stored in the uterus will purge from the woman's body naturally, drawing with it any Qi stored within the Lower Dantian.
- Patients suffering from nerve suppression (cervical spondylosis) may experience a worsening of their condition. In severe cases, the additional pressure on the nerve roots may complicate the patient's condition, especially when the body's energy field expands causing the Marrow within the spinal column to ascend.
- Patients with bone fractures must first have the bones set correctly and immobilized to prevent further injury. The subtle involuntary movements of the limbs during Qigong treatment or practice can disturb the bone's healing process.
- Pregnant women must be treated by a Medical Qigong doctor who specializes in obstetrics, otherwise the risk of involuntary abortion (miscarriage) or other complications can arise.
- The treatment of geriatric patients prohibits Dynamic Qigong because of the delicate nature of their internal tissues. All exercises and meditations should be kept slow, quiescent, and tranquil.
- The treatment of young children requires that all exercises and meditations be age appropriate, and be specifically oriented toward healing his or her disease. Some Qigong exercises, meditations and treatment methods may interfere with the natural development of the child's internal organs and tissues.
- The treatment of cancer patients requires modified prescriptions in accordance to the specific disease. With certain cancers, Dynamic Qigong is contraindicated to prevent the cancer from further metastasizing.

CHAPTER 30
BALANCING THE BODY'S ENERGETIC SYSTEM

PURGING, TONIFYING, AND REGULATING

In Medical Qigong therapy, successful results can be achieved only when the doctor adheres to the principle of "treating Deficient syndromes with tonification and Excess syndromes with purgation." Tonification supplements the insufficiency and strengthens the body's resistance, while purgation reduces the Excess and expels pathogenic factors. The final aim is to balance Yin and Yang to regulate the physiological functions of the body.

Medical Qigong Therapy adheres to the traditional Chinese medical explanations of illnesses and general principles of healing. Only the techniques used during treatment differ. Qigong doctors Purge the patient's body, then emit Hot or Cold Qi in conjunction with the Five Elemental energies, using various hand manipulations, including: pushing, pulling, rotating, leading, and shaking.

Through energy extension, the doctor can reinforce the patients body's resistance to pathogenic factors, replenish the Yin and Yang organ energy and enhance mental clarity.

THE THREE TREATMENT MODALITIES

Medical Qigong therapy is divided into three treatment modalities (Figure 30.1). These treatment modalities enable the Qigong doctor to purge, tonify, and regulate the patients' body. These techniques are described as follows.

1. Medical Qigong Purging techniques are used by the doctor to treat the patient's Excess syndromes, and to detoxify the body of pathogenic Qi. The "lower aspect" of classical Chinese medicine governs the treatment of the patient's illness and corresponds to Earth. It involves purging pathogenic influences and dispersing stagnations in order to cure the patient's illnesses. This allows the patient the ability to remove Excess Qi from organs and tissues.
2. Medical Qigong Tonification techniques are used by the doctor to treat the patient's Deficient syndromes, as well as strengthen and stabilize the organs and organ systems. The "middle aspect" of classical Chinese medicine governs the nourishment of the patient's nature and corresponds to Man. It involves helping the patient prevent illness by tonifying

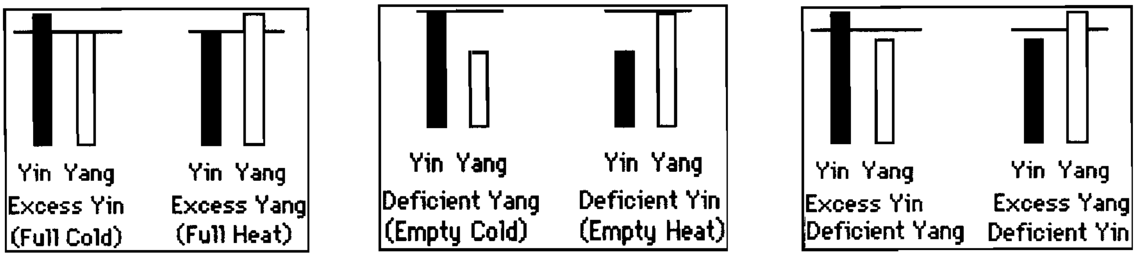

In cases of Excess, the doctor should Purge.

In cases of Deficiency, the doctor should Tonify.

In cases of Excess and Deficieny combined, the doctor should Regulate.

Figure 30.1. Purgation, Tonification, and Regulation

Heaven —— Regulation ——— Balance Qi and Blood —— Nourish the Patient's Destiny

Man ——— Tonification —— Treat Deficient Conditions — Strengthen the Body to Prevent Illness

Earth —— Purgation ——— Treat Excess Conditions —— Detoxify the Body to Treat Illness

Figure 30.2. Ancient Chinese Perspective on Purgation, Tonification, and Regulation

organ energy Deficiencies. This allows the patient the ability to replenish organ energy depletions and tissue emaciation.

3. Medical Qigong Regulation techniques are used by the doctor to balance the patient's Yin and Yang energy, as well as organ Qi. The "highest aspect" of classical Chinese medicine governs the nourishment of the patient's destiny and corresponds to Heaven. It involves regulating the patient's Three Dantians, Taiji Pole, Yin and Yang organs and energy fields. Regulation techniques promote a balance of Jing, Qi and Shen for the optimal development of all the aspects of man, i.e., body, mind, emotion, spirit and energy. Thus, in China, Medical Qigong regulation is said to "nourish the patient's destiny" (Figure 30.2).

PURGATION TECHNIQUES

The patient's body can be dredged and purged by using two different expelling methods. The first method is the direct removal of pathogens from the patient's tissues and organs. The second method is the indirect purging of the patient's tissues and organs through the channels and points of the patient's body. These two methods serve two different purposes: The Direct Removal of Pathogenic Factors, and The Purgation of Channels and Points.

1. The Direct Removal of Pathogenic Factors is achieved through extending Qi into the tissues and organs where they are lodged, and dredging the Toxic Qi directly out of these locations. This method is used to expel pain due to Excess Heat or Cold. The various techniques used in this type of purging employs stationary Qi-guiding manipulations. The various hand manipulations used for the direct removal of pathogenic factors consists of such techniques as: The Five Thunder Fingers Hand Posture, Shaking and Trembling Hand Posture, Bellows Palm, Vibrating Palm, Dragon's Mouth Palm, Tiger Kneading Palm, and Spiraling Palm (see Chapter 33).
2. The Purgation of Channels and Points techniques are used to expel External pathogenic Evils (such as the invasion of Wind, Fire, Heat, Dry, Damp, Cold, and environmental toxins, etc.), or Internal pathogenic Evils (such as fear, anger, and grief, etc.) from the organs and tissues of the body. It is also used to redirect Rebellious Qi (e.g., Stomach Qi). The various techniques used in this type of purging employs dynamic-linear Qi-guiding manipulations such as the Pulling, Leading, and Guiding methods (see Chapter 34).

TONIFICATION AND PURGATION METHODS

The doctor should always purge before tonifying the patient's body. In the clinical setting, these two methods are often used together in combination with other methods in treatment. These combined methods are as follows.

TONIFICATION AND PURGATION WITH RESPIRATION

Tonification occurs as the doctor exhales while guiding and directing Qi into the channels, points, and the internal organs and tissues of the patient's body. Purgation happens when the doctor inhales while guiding, emitting, and leading the patient's Qi, or pulling the pathogenic Evils out of the patient's body.

During treatment, it is important that the patient cooperate with the doctor by matching (if possible), per the doctor's instruction, his or her breathing rhythm to the doctor's treatment. The patient should assume a proper posture (lying, sitting or standing), relax, and concentrate his or her mind on the location being worked on. When the doctor exhales and emits Qi toward the patient, the patient should inhale to absorb the doctor's Qi into the designated location. When the doctor performs purgation, the patient should exhale to release Toxic Qi out of his or her body.

TONIFICATION AND PURGATION ALONG AND AGAINST THE CHANNEL FLOW

With regard to acupuncture therapy, the classics say that, "puncturing along the direction of the channels yields a tonifying effect, while going against it induces purgation." In Medical Qigong, however, due to the Qigong doctor's nature of emitting, leading and guiding the energetic flow of Qi, certain modifications are used. The following are a few examples of tonification and purgation techniques used along and against the energetic flow of the channels:

- Causing Qi to flow up the Yin channels of the legs and into their associated organs in the torso is considered tonification (think of following the river to the sea). One example of this type of treatment therapy is emitting energy into the base of the patient's feet (at the Kd-1 points); this causes the energy to flow into the torso to strengthen the Kidneys.
- Moving Qi from the torso down the legs against the flow of the leg Yin channels is considered purgation (following the river from the sea). One example of this treatment therapy would be pulling Excess Qi from the Liver organ, down the right Liver channel and out the feet to sedate the Liver Fire (the Gall Bladder Channels can be used to purge Excess energy from the Liver organ down the leg Yang channels).
- Moving Qi against the flow of the arm Yin channels into their associated organs is considered tonification (think of following the river to the sea). One example of this treatment would be emitting energy from the base of the thumbs, at the Lu-11 points, into the torso to strengthen the Lungs.
- Moving Qi against the flow of the arm Yang channels into their associated organs is considered purgation (following the river from the sea). One example of this treatment would be pulling Qi against the flow of the Large Intestine Channels down and out the hands to purge Excess Qi in the head.

TONIFICATION AND PURGATION POINT THERAPY WITH THE HORARY CYCLE

When prescribing Jing Point therapy as a means of self-regulated homework, the Qigong doctor encourages the patients to follow the formula of tonification and purgation according to the energetic flow of the Horary cycle. According to ancient Chinese medical theory, the ebb and flow of Qi and Blood along the different channels is related to designated days and hours (see Chapter 11). This theory applies to two specific tissue manipulations: Holding the Point for Tonification, and Closing the Point for Purgation.

- Holding the point for tonification requires the patient to use the Massage Tapping method (see Chapter 37) on specific points two hours prior to its opening or peak flourishing time (high-tide).
- Closing the point for purgation of the internal organs or channels requires that the Qigong doctor drain the diseased area at high-tide (when it is most full of Qi and Blood). When the patient is applying Jing Point therapy for homework prescriptions, however, the application is quite different. This technique requires the patient to use the Massage Tapping method on specific points two hours after its opening time.

The patients may also be instructed in self-massage using the Tapping method for self-healing. If the patient has a Kidney disease, for example, the specific points should not be tapped during its peak time (5–7 p.m.), but rather two hours prior to its peak time (3–5 p.m.). This will improve the Kidneys' function.

TONIFICATION AND PURGATION GUIDELINES

When the Qigong doctor purges and tonifies a patient, he or she must remember to adhere to the following guidelines carefully.

GUIDELINES FOR THE DOCTOR

When treating the patient, the doctor must always consciously direct and be aware of the movements of each hand.

- Generally, the right hand is used for tonification, while the left hand is for purgation. Diseases that require purgation are treated by draining Toxic Qi with the left hand. Diseases that require tonification are treated by extending Qi through the right hand.
- Turning hands towards the body is used for tonification, while turning the hands away from the body is used for purgation.
- Moving the hands away from the body and upward can regulate the Yang energy, causing it to rise, and is inclined to tonify. Moving the hands downward pulls the Yang energy down, and is inclined to purge.
- Clockwise drilling (the palm or extended fingers move in a clockwise direction over an area) is used for tonification and increasing Qi; counterclockwise drilling is used for purgation and decreasing Qi.

GUIDELINES FOR THE PATIENT

The patient, (like the Qigong doctor), must be aware of hand movements when practicing Qigong prescriptions. In addition, the patient must be aware of his or her posture, respiration, and mental attitude.

- The eyes are open for purgation; closed eyes are used for tonification. The patient should apply these principles according to the symptoms of the disease.
- Inhalation is used for tonification, exhalation is used for purgation. During purgation treatments, the patients assume the proper postures, relax, and concentrate their mind on the disease area while exhaling the pathogenic Qi. This exhalation technique is used not only to promote the removal of the surfacing toxins, but is also used to enhance the doctor's efficacy in purging the diseased area.
- The audible and silent tone resonations (see Chapter 40) also have the function of either tonification and purgation. Speaking the silent tone when inhaling can be used for tonification; while exhaling, the tone is used for purgation.

UNDERSTANDING MEDICAL QIGONG REGULATION

The Qigong doctor must master three main techniques of Qi regulation to facilitate clinical healing through Qi emission. These techniques include: Training the Qi, Guiding and Inducing, Tonifying and Purging the Qi, and Emitting the Qi.

1. Training the Qi exercises and meditations are used to gather Qi and to promote its circulation within the body. The Qigong doctor must have an extensive comprehension of several Medical Qigong techniques used to gather and circulate energy throughout the body. This includes a complete understanding of Postural, Respiratory, and Mental Dao Yin modalities, as well as the static and dynamic clinical modalities. These training techniques provide the Qigong doctor with the ability to regulate and balance the patients' Qi, and set the foundation for the patients' Medical Qigong prescriptions.
2. Guiding and Inducing, Tonifying and Purging the Qi are used to circulate Qi within the external field of energy (Wei Qi). These techniques require a complete understanding of guiding and inducing (leading the Qi), and tonifying and purgation techniques, to being able to rectify Qi deviations and remove energy blocks.

These training techniques provide the Qigong doctor with the foundation for the patients' Medical Qigong therapy.

3. Emitting Qi techniques must be learned and perfected by the doctor. The Qigong doctor must have an extensive understanding of the multiple techniques used to extend Qi into the patient's body. These include electromagnetic energy, light, sound, and hot and cold Qi projections. These training techniques and abilities also set the foundation for the patients' Medical Qigong therapy (see Chapter 34).

QI REGULATION AND GUIDING PRINCIPLES

Qi regulation is considered a combination technique for reinforcing and reducing. It is used to readjust and balance the Yin and Yang energy. If the patient's Qi rises adversely, the doctor should guide and secure it downward. The same principle is applicable for any Deficiency or Excess of Qi in the left or right sides of the patient's body.

Qi regulation also includes purgation after tonification, and tonification after purgation. This method of guidance is used when Qi stagnation, stasis of Blood, or disorders of the channel Qi develop. Under pathogenic conditions, Qi of the viscera and channels may become weak and stagnant. This is manifested by symptoms such as listlessness, general weakness, and the hypofunction of certain organs or tissues. Stagnation of Qi and stasis of Blood can also obstruct the channels causing pain. Pain may also be caused by outside pathogenic factors such as trauma, inflammation, muscular spasms, and adhesive degeneration. In these circumstances, the channels are dredged to promote Blood circulation, and to remove the stasis in order to relieve spasms, and dispel pain.

Every organ in the body has its own Qi that circulates and interacts with the rest of the body. When the normal function of the organ is affected, the circulation of Qi may become disturbed. This disturbance of Qi in one organ may influence the functional balance of the whole Qi system.

The doctor regulates the patient's channel Qi by dredging the channels and collaterals. This facilitates Qi circulation and leads Qi back to its origin. In the case of a weak and hypofunctional Lung Qi, for example, emitting Metal-Qi plus Pushing, Shaking and Leading manipulations (see Chapter 34) can be adopted to activate the Lungs' Qi. The combination of these techniques causes the channels' Qi to overflow freely, thus improving the function of the Lungs.

ENERGY BALANCING PROCEDURES

When extending energy into the patient the main emphasis is placed on balancing the patient's energetic fields. The doctor must have both physical and emotional releases from the patient's body, to detoxify and liberate the trapped Qi.

It is important to not let the patients avoid facing their problems; keep them centered around the troubled area, focused on the awareness of what they are feeling. Basic energy balancing procedures are as follows.

1. If the treatment area is Hot, cool it down. The doctor must continue treating the patient until the Hot area has cooled. A strong and quick moving hand gesture is best used to treat a Heat syndrome.
2. If the treatment area is Cold, heat it up. For a Cold syndrome, use a longer treatment to warm up the channel in the area treated, to eliminate the Cold.
3. If the tissue's energy is dark or cloudy, change it to light and clear.
4. If you feel tingling on any area or organ, remain there until the tingling subsides and the organ or area completes its energetic unwinding. Tingling in an area indicates that the patient's body is releasing an energetic cluster. The patient may or may not be aware of any feelings or sensations when this occurs.
5. If the area is emitting electric shocks, dampen or sedate it by connecting it with its associated organ. Connecting Yang (positive) into Yin (negative) will ground the energy.
6. If you sense pressure or Excess, dredge the area, and circulate the Qi.
7. If you feel intense pulsation, decrease the intensity and make it rhythmical.
8. When treating an energetic field that is De-

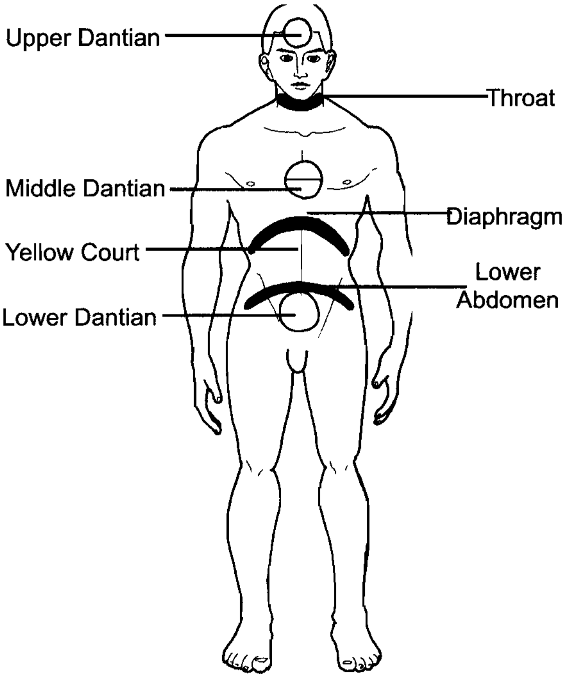

Figure 30.3. Locations of Energetic Blockages

pressed (sinking or contracting inward), use an "ascending method" of Qi manipulation for tonification (i.e., move and emit Qi from the lower areas of the body upwards). Some examples of this type of ailment are:

- Lower Burner Depression
- Middle Burner Depression
- Immune System Depression and Breakdown
- Autoimmune Diseases (Lupus and Fibromyalgia)
- Chronic Fatigue Syndrome
- Heart Failure
- Kidney Failure
- Prolapse of the Stomach

9. Assign Medical Qigong prescriptions for patients, including meditations and exercises, as needed.
10. Treat patients with severe diseases by:
    - Stopping and controlling the spread and development of the disease,
    - Stopping the patient's Excess consumption of Qi,
    - Treating the cause and root of the disease, after the patient's condition is under control, and
    - Assigning homework.
11. For chronic diseases, use the method of purging and dredging stagnation to move any impacted Qi.
12. For sprains, use Qi compression (using the Qi to press the tissues) around the local area. Close the channels surrounding the sprain first, isolating the trauma. Then pull out the stagnant Qi. If the doctor immediately treats the local area, he or she will disturb and further complicate the injury. It is important to close off the surrounding channels around the wound first, leaving only the affected channel open, before pulling out the stagnant Qi. This controls the flow of energy in to the stagnated area.

Check the patient after seven treatments to determine if·the therapy has been effective. Reevaluate each patient according to each individual's constitution.

CLINICAL OBSERVATIONS OF PATHOLOGICAL ENERGETIC PATTERNS

When the Qigong doctor begins to treat patients, certain energetic observations and phenomenon will be experienced while diagnosing the energy inside the patient's tissues. These energetic experiences are quite different from the Eight Energetic Touches which focus on the feelings of the body after meditation practice (see Chapter 29). The Internal conditions of the patients' disease, its energetic nature, as well as its clinical manifestation, are expressed through the following: Hot, Cold, Deficient, Excess, moving, stagnant, Damp, Dry, Clear (Clean), Turbid (Impure), open (accessible), closed (armored), and leaking.

LOCATIONS OF ENERGETIC BLOCKAGES

In clinical practice we generally observe six main areas in the patient's body where energy becomes trapped and blocked. Each energetic blockage is an expression of the patient's ability to armor and protect him or herself from emotional wounding. These six main areas in the patient's body can be the foundational base and the root cause for the patient's disease (Figure 30.3). These six main areas are described as follows.

THE AREA OF THE HEAD

The area in the cranium that is most commonly obstructed is within the base of the occiput, at the back of the head. This obstruction is usually noted within patients who suffer from migraine headaches and brain tumors. This barrier is often created through chronic stress patterns which cause the Liver's Heat to rise up the patients' back and become trapped inside the head. This condition is usually caused by intense concentration. Too much hard focusing can also cause Qi to condense and stagnate. The location of Qi obstructions due to physical trauma vary according to each type of trauma, and the and the severity of the injury.

THE AREA OF THE THROAT

This area separates the Upper Dantian (the spiritual/perceptive center) from the Middle Dantian (the emotional/empathic center). Messages of divine inspiration and awareness descend downward from the Heavens through the Baihui area seeking to stir the emotions of the Heart. If the patient is not receptive (due, for example, to a disconnection or even a resistance to spiritual growth) a barrier is sometimes formed within the patient's throat, dividing the Upper Dantian from the Middle and Lower Dantians.

This barrier can also be created to stop emotions from flowing through the Heart and rising upward to be released through the windows and caverns of the soul (the eyes, ears, nose, tongue, body, and spirit). If, for example, the teeth become clenched and the throat tightens from feelings of grief or anger, these emotions can become trapped in the throat, forming a "plum seed" or "knot." When the Lungs' Qi congeals in the areas of the upper chest, shoulder blades, neck and the back of the occiput, this indicates chronic suppression of these emotions. It is also evidence of the mind and body struggling to reconnect and express the emotional and spiritual self.

THE AREA OF THE HEART

The Heart is energetically divided into two main chambers. The upper chamber houses and responds to spiritual emotions, while the lower chamber houses and responds to sensual emotions.

Patients will tend to suffer not only from repressing their basic biological drives (the sensual emotions based in the lower chamber of the Heart), but also by repressing their emotional connections with the divine (by failing to accept their higher spiritual calling or spiritual enlightenment associated with the upper chamber of the Heart). The higher self is the focal point of the superconscious realm; it is all-inclusive and at the very center of our being where individuality and universality blend (center core).

Patients who have chronically armored their Heart tend to exist in a state of guarded suspicion. This pattern of emotional suppression tends to lead to Qi stagnation and the creation of disease. When patients armor their upper chamber, they tend to hide in the emotions of their lower chamber and visa versa. Patients who armor the upper chambers of the Heart, for example, may become involved in relationships of extreme sensual passions related to the Heart's lower chamber, to avoid painful unresolved issues hidden within the past. The deeper and more energetically charged the issues, the stronger the sensual passions. Whereas, patients who armor the lower chamber of the Heart, might become involved in relationships of extreme spiritual, or religious passions related to the Heart's upper chamber to hide from their painful unresolved issues. The deeper and more energetically charged the issues, the stronger the spiritual or religious passion.

THE AREA OF THE DIAPHRAGM

The diaphragm area separates the Middle Dantian (the emotional center) from the Lower Dantian (the physical/sexual center). When a barrier in this particular area is created, it separates the emotional Heart from feeling the sexual passion which rises up from the Lower Dantian and the Kidneys' area.

It is interesting to note that patients with a

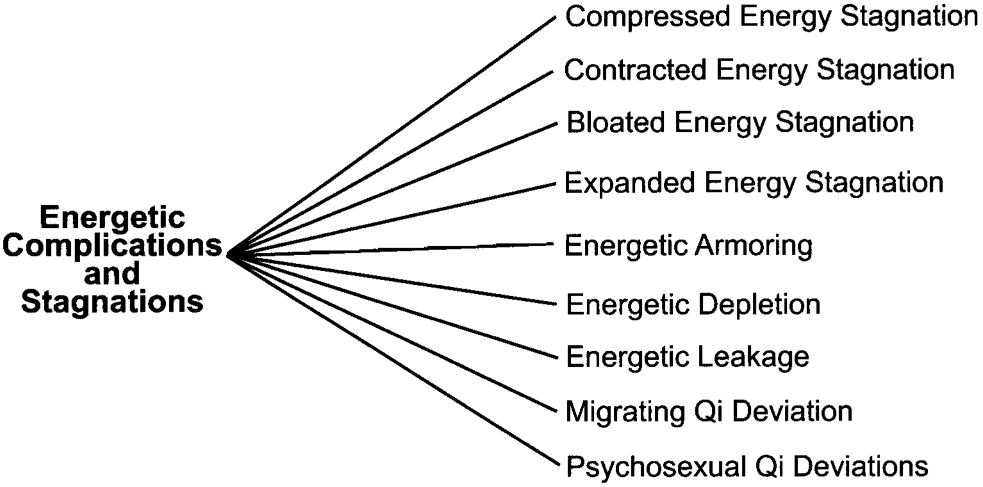
Energetic Complications and Stagnations
Compressed Energy Stagnation
Contracted Energy Stagnation
Bloated Energy Stagnation
Expanded Energy Stagnation
Energetic Armoring
Energetic Depletion
Energetic Leakage
Migrating Qi Deviation
Psychosexual Qi Deviations

Figure 30.4. The Body's Energetic Complications and Stagnations

barrier in the diaphragm sometimes enjoy the physical sensations of sexual intercourse, but are separate from their spiritual/emotional connection to true bonding or "oneness" with their partner. Again, the connection between the emotional spirit and body/mind has been severed.

THE AREA OF THE YELLOW COURT

The Yellow Court (CV-12) is an area located in the center of the chest, just below the diaphragm and xiphoid process of the sternum. Its function is that of being the access area to release the body's internal organs' emotional memories. Its location is also attributed to the Third (solar plexus) Chakra.

To access the patient's stored emotional memories and release them from the Yellow Court, the Qigong doctor supports the patient's Shendao GV-11 (Spirit Gate) with the left hand while energetically manipulating the patient's Yellow Court area with the right hand. If the patient is not energy sensitive, the doctor uses a slight Compression/Release technique (see Chapter 37) to disperse the patient's armoring system.

The patient should be in a supine position, with knees pointed upward, feet flat, arms by his or her sides, mouth open, inhaling deeply into the Lower Dantian. As the patient exhales, the doctor lightly compresses the patient's Yellow Court until a deep groaning sound begins to resonate from the patient's thorax. The doctor circulates the energy from the patient's Yellow Court into the Shendao, filling up the patient's Heart with energy. The doctor then moves this energy back into the patient's Yellow Court. This process continues until the patient's suppressed emotions surface and are released.

Note: It is important to first purge and release the Lungs of grief before attempting to access the patient's Yellow Court.

THE AREA OF THE LOWER ABDOMEN

This area also separates the Lower Dantian from the Middle Dantian, insulating the sexual feelings from the middle and lower torso. This energetic freezing of the pelvis is usually found in victims of rape, sexual molestation, or in women who have had one or more abortions. The Qi flowing into the lower pelvis and legs becomes blocked due to stored fear, shame, guilt, and sorrow. Sometimes the patient's physiology in that particular area may become obese; the abdomen, thighs or hips may be heavily padded and armored due to Qi stagnation. Often times, victims of childhood sexual abuse become overall obese to avoid attracting the opposite sex, or to block and pacify these hidden traumatic memories.

TYPES OF ENERGETIC COMPLICATIONS

Each organ and channel has the potential to carry or create several types of energetic complications and stagnations. The doctor should observe and feel the shifts in the patient's energetic balance to determine the specific conditions of the disorder (Figure 30.4).

COMPRESSED AND CONTRACTED ENERGY STAGNATION

Compressed or Contracted energy stagnations are sometimes considered a Yin condition due to the type of energetic compression involved in its formation.

- Compressed Stagnation is caused by the patient's energy pressing downward (externally moving inward) and will feel armored and hollow (Figure 30.5). When this type of energetic stagnation disperses it does not energetically unwind, but disperses and dissolves into the surrounding tissue areas. Because the patient's surrounded consciousness has armored the tissue area, the energetic dispersion can sometimes cause spontaneously released emotions or behavior such as crying, laughing, shouting, screaming, sighing, coughing, etc.
- Contracted Stagnation is caused by the patient's energy pulling inward, and will generally feel armored and solid (Figure 30.6). This type of energetic stagnation can be formed by an emotional core, or focused concentration, that draws additional Qi into its center, compacting the belief structure, layer after layer with energy. When this type of energetic stagnation disperses it can sometimes begin energetically unwinding throughout the entire body. This energetic unwinding sometimes creates involuntary muscle twitches and physical undulations.

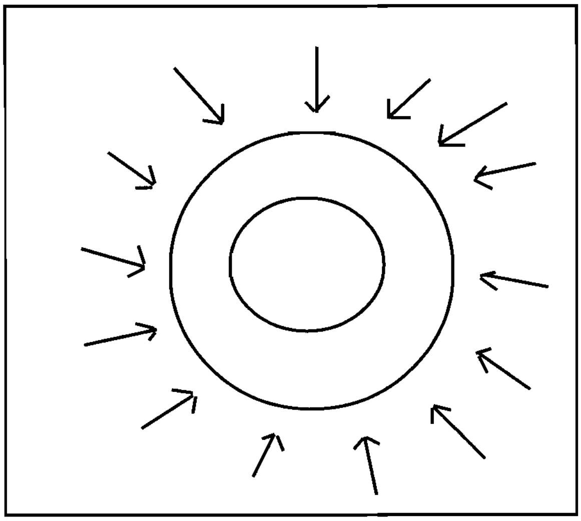

Figure 30.5. Compressed Stagnation

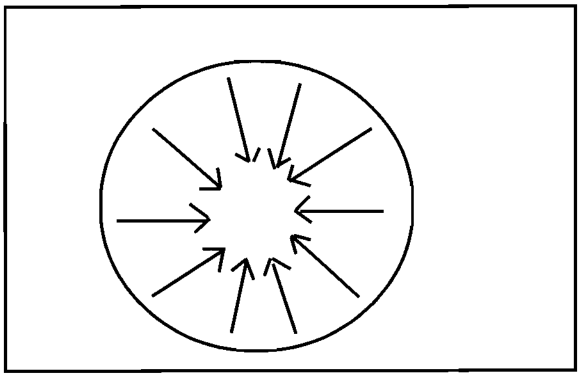

Figure 30.6. Contracted Stagnation

Body tissues and muscles accumulate emotional energy, becoming hard and tense in the adjacent areas. Denseness indicates the accumulation of pathogenic stagnant Qi. If this energy block remains, disease can manifest as ovarian cysts and tumors, or even in extreme cases as cervical cancer. Breast cysts, tumors, and cancer can result from stagnant Qi in the Lungs, or from a combination of Lung and Liver Qi stagnation. This type of stagnant energy feels very Hot and volatile because of the accumulated anger and rage.

Treatment Principals for Compressed and Contracted Energy Stagnation

The Qigong doctor disperses and purges the patient's stagnant Qi using the Thunder Palm technique, Tiger Kneading techniques, and counterclockwise Circular Spiraling (see Chapter 34).

BLOATED AND EXPANDED ENERGY STAGNATION

Bloated and Expanded energy stagnations are sometimes considered a Yang conditions due to the type of energetic expansion involved in their formation. Qi stagnation, with a characteristically expansive or bloated appearance, can be caused from an accumulation of Phlegm and Body Fluids (in addition to Qi and Blood) in the adjacent tissue areas of the body (Figure 30.7 and 30.8). This type of stagnation usually occurs in the Lower Dantian and abdominal areas. These are examples of Excess syndromes of stagnant energy and manifest as: distending, turbid, leaking, tingling, and dense Qi.

If this type of energy stagnation remains obstructed, the resulting disease can manifest as edema within the tissues, colitis, or angina pectoralis due to unprocessed despair and grief (in the Lungs) with anxiety (in the Spleen). The energy of grief and despair feels thick, heavy, and slimy like phlegm or mucus. Often this energy is felt in the Lungs, as well as in the Spleen.

Treatment Principals for Bloated and Expanded Energy Stagnation

The Qigong doctor disperses or dissolves the patient's stagnant Qi by using the Tiger Kneading, Vibrating Palm, and Shaking Palm techniques (see Chapter 34).

ENERGETIC ARMORING

This condition can result from the patient protecting specific tissues, organs, or areas of the body. Energetic armoring is initiated when the patient freezes certain emotional feelings to maintain the denial system. Patients may be unaware of this armoring condition because they dissociate from the original trauma, or minimizing its severity. Emotions attributed to energetic armoring are generally manifested as anger, fear, guilt, and shame. The angry patient's denial system is maintained through physically or emotionally attacking others; the anxious patient, full of fears, guilt, and shame, copes by "Shening out." Both of these reactions shield the patients in emotional situations from their own feelings, while reinforcing the energetic armoring.

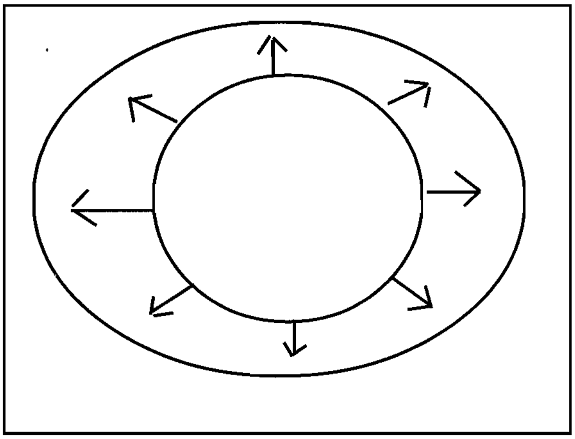

Figure 30.7. Bloated Stagnation

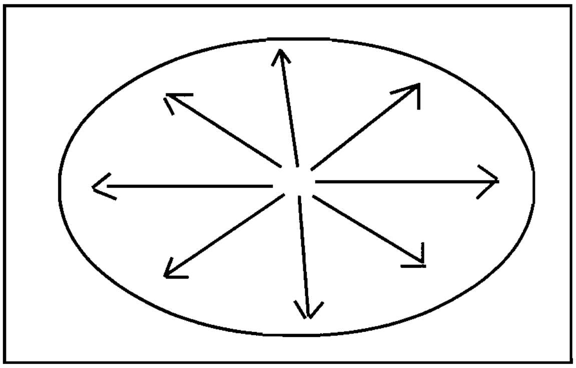

Figure 30.8. Expanded Stagnation

Armoring results in tense organs and tissues. When patients cannot feel their body, other complications can arise, such as compressed, contracted, bloated or expanded energetic stagnations.

Treatment Principals for Energetic Armoring

The Qigong doctor can assist the patient through "soul retrieval" to unlock the feelings logged deep within the body's tissues (see Chapter 19). The primary goal is to first help the patients acquire a sense of security within their own body, then address the root of the disease.

ENERGETIC DEPLETION

Whole body depletion results from the following conditions:

- Not receiving enough Postnatal Qi for restoration, resulting in improper internal organ functioning,
- From suppressing emotions (e.g., worry or fear),
- Experiencing a damaged energetic field due to trauma, or
- Being born with a Deficient prenatal constitution.

Generally, energetic depletion can result in physically undeveloped areas in the body. In some cases, emotional traumas can cause specific organs or tissues to become energetically depleted to the degree that their natural development cannot take place.

Treatment Principals for Energetic Depletion

The Qigong doctor, if qualified, can assist the patient by prescribing herbal tonics. The patient should also be encouraged to eat regular, nutritional meals and to balance work with relaxation time. If the patient's energetic field has been damaged, the Qigong doctor can repair it by using Flat Palm and Circle Spiraling techniques (see Chapter 34).

ENERGETIC LEAKAGE

Energetic leakage is a result of the patients releasing Qi from out of their joints, due to injury or unconscious sabotage. When energy drains from the patients' joints, the specific areas along the channels become depleted, resulting in a Deficient Qi which can cause an atrophy syndrome. Any type of energetic thinness along the Wei Qi field indicates an insufficiency of Qi. A Deficient energetic field can result in an empty, numb, or weak feeling of Qi within the tissue areas.

Treatment Principals for Energetic Leakage

The doctor envelops the patient's depleted tissue area (which is leaking) with Spiral Wrapping techniques to create an energetic cast that allows the patient's energetic field to completely heal (see Chapter 34).

MIGRATING QI DEVIATION

Migrating Qi Deviations can result from either the patient's suppressing and avoiding feelings, or from Wind invasion (either Internal or External). In Traditional Chinese Medicine this condition is known as Wind Bi (Wind Pain) or Wandering Bi (Migrating Pain). Once patients feel cornered by a specific memory, they subconsciously shift the toxic energy to another part of their body; this energetic shifting requires the treatment to be restarted. When treating these patients, it is important to first connect the original trauma to the Heart to prevent these energetic shifts, and to stabilize the condition. Once stabilized, disperse the Wind Invasion and tonify any Deficient organ.

Treatment Principals for Migrating Qi Deviation

The Qigong doctor first envelops the primary area of complaint, isolating the pain, and then begins to internally Spiral Circulate the patient's internal Qi. Once the energy is in motion, the doctor leads the Toxic Qi out of the patient's body.

PSYCHOSEXUAL QI DEVIATION

Psychosexual Qi Deviations can result from an immediate energetic tissue over-stimulation. Patients with this condition experience intense sexual undulations and orgasms when being treated in a safe clinical environment. This reaction is due to their avoidance issues, brought about by an energetic trigger mechanism. This psychosexual reaction usually takes place when touching any part of the torso or extremities. Because the energy seeks balance (much like water), when it flows into the genital area, the sexual neurons become overstimulated, activating and releasing the energetic turbulence already stored therein. This condition is mostly found within patients who have either been raped or molested as children.

Treatment Principles for Psychosexual Qi Deviation

The doctor is encouraged to refer the patient to a psychotherapist who specializes in such trauma. Meanwhile, the doctor continues to treat the patient by rooting the patient's Heart (emotional chamber) to the patient's Lower Dantian (sexual chamber). Once this connection is established, the doctor reconnects the patient's Lower Dantian to the patient's Upper Dantian (spiritual chamber), forming a complete body connection within the patient's Taiji Pole. By reconnecting the sexual chamber with the spiritual chamber, the doctor empowers the patient, allowing the patient to reclaim his or her true sexual/ spiritual identity.

PATIENT'S ENERGETIC FACIAL DISTORTION

Sometimes, when treating a patient, the Qigong doctor will notice the patient's facial features energetically distort, changing their facial pattern. These facial distortions can appear and disappear rapidly, or maintain their illusion for several seconds. The cause for the patient's energetic facial distortion can be brought on by several factors: The Patient Reliving A Past Trauma, The Ancestral Patterning of the Patient's Jing, or The Energetic Manifestations of the Patient's Active Spiritual Guardians.

1. The patient reliving a past trauma can release an energetic distortion that expresses itself on the energy field of the patient's face. This projected image reflects the age and time frame wherein the patient received the initial trauma. Often, as the doctor connects with the energetic root of the patient's disease, the patient's face will regress to reveal the delicate features of the patient as a young child. This visual facial pattern of energetic age regression reveals the exact age the patient suffered the initial trauma.
2. The ancestral patterning of the patient's Jing can also be released as energetic distortions that express themselves on the energy field of the patient's face. Occasionally, these projected images can reflect the patient's ancestry, or reveal an ancestral guide. Multiple projections of racially identifying features; images of men, women, and sometimes children will appear to shadow the patient's features, much as a projected image over an original form.
3. The energetic manifestations of the patient's spiritual guardians are believed to project their image onto the face of the patient during treatment. Some cultures accept that these guides manifest in the form of animals. The animal form will appear to flicker across the patient's face and then dissipate. This phenomenon is the most rarely seen of all the energetic facial distortions.

PATIENT RESISTANCE TO ENERGETIC THERAPY

Sometimes, in order to avoid the painful memories, the patient will subconsciously sabotage the treatments. Resistance can manifest in one of four ways.

1. Falling asleep is a form of physical escape.
2. Excessive talking is a mental form of avoidance behavior.
3. Irritability or sexual response is a form of emotional escape.
4. "Shening out," also called "spacing out," is a form of spiritual escape.

When observing these patterns, the doctor encourages the patient back into the body to feel the emotions and memories trapped in the particular areas being treated. It is important for the doctor to let the patient know the cause and effect of the patient's chosen defense mechanism. These personal conversations are necessary for the patient's growth and emotional healing. In cases of sexual response, tact and appropriate professional boundaries are especially important.

When emitting Qi, the doctor can encourage the patient's Shen to stay in its residence by asking the patient to focus his or her breathing on the tissue areas being treated. This focused attention causes the patient's Po to settle and calm themselves within the patient's body, allowing the patient's Hun to communicate the feelings surfacing within the patient's Heart.

UNBLOCKING THE FIVE PASSES

The Wu Guan or Five Passes refer to the five important gates on the Governing Vessel. These gates, or passes, are sometimes called Tie Bi, which translates as the "iron wall," and are the areas of the body where it is most difficult for the energy to pass through when circulating the Microcosmic Orbit. Free passage through the Governing Vessel is important, especially when tonifying the patient's body or regulating the patient's Qi. To circulate the Qi up the Governing and down the Conception Vessels, these five gates must remain open.

Opening these Five Passes are essential for the treatment of diseases in the neck and lumbar-sacral regions. In Medical Qigong therapy, disease is sometimes an indication that both the Governing and Conception Vessels have been, or are, in a state of disconnection. If disorders of Qi, along with Qi obstruction, develop along the Governing and Conception Vessels, the doctor should emit Qi into one of the Five Passes (depending upon the affected area) to facilitate the regulation of Qi (Figure 30.9).

THE LOWER SPINAL PASS

The Lower Spinal Pass or "Coccyx Pass" (Wei Lu Guan) is located on the lowest segment of the spine just posterior to the anus, near the Chang Qiang GV-1 point. If Qi becomes blocked in this area the patient suffers a dull persistent pain by the coccyx, along with feelings of heaviness.

To treat, have the patient direct energy up the Lower Spinal Pass, using mild or gentle intention; have the patient position the tongue against the upper palate and inhale deeply while gently contracting the anus and abdominal muscles. These gentle contractions of the muscles along with the position of the tongue on the upper palate create an energetic sacral pump, which facilities the release of the energetic block.

THE MIDDLE SPINAL PASS

The Middle Spinal Pass consists of the Mingmen (GV-4) area and two points (Jia Ji Guan) located on the lateral sides of the Mingmen. If Qi becomes blocked in these areas the patient suffers a dull persistent pain in the middle of the spine, or on one or both sides of the Mingmen. This obstruction causes a physical sensation that feels like a spinal rupture, break, or herniated disc near the waist area when Qi circulates through these points.

To treat this disorder, the doctor extends Qi into the patient's Middle Spinal Pass, while leading the patient's energy upward along the Governing Vessel.

THE UPPER SPINAL PASS

The Upper Spinal Pass is located between the shoulders on the Shendao point (GV-11) Spirit Gate. If Qi becomes blocked in this area, the patient suffers a dull persistent pain on the midline of the upper back below the fifth thoracic vertebra. This obstruction causes a physical sensation that feels like a spinal rupture, or causes intense Heart pain when Qi circulates through this point.

To treat this condition, the doctor extends Qi into the patient's Upper Spinal Pass, while leading the patient's energy upward along the Governing Vessel.

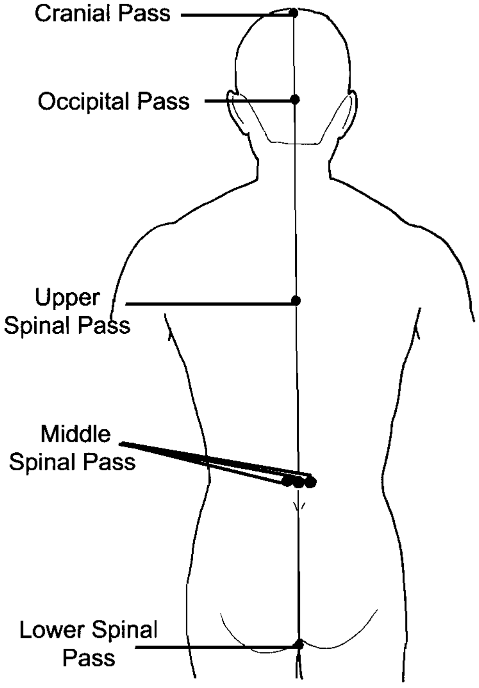

Figure 30.9. The Five Spinal Passes

THE OCCIPITAL PASS

The Occipital Pass (Yu Zhen Guan) is located just inferior to the occipital bone where the brain originates. If the Qi becomes blocked at this area, the patient suffers a dull persistent pain around the occiput, stiff neck, feelings of heaviness, as if there is something were stuck inside the back of the head.

To treat this problem, the doctor should instruct the patient to close the eyes, raise the head slightly and look upward, while gently concentrating on the Baihui GV-20 point. Have the patient imagine the energy block as a cube of ice. The ice melts into water, the water dissolves into vapor, and travels upward along the Governing Vessel into (and through) the Baihui area, then proceeds down the Conception Vessel toward the Lower Dantian.

THE CRANIAL PASS

The Cranial Pass is located at the vertex on the midline of the skull. If the Qi becomes blocked at this area, the patient suffers a dull persistent pain around the eyes and the top of the head, as well as feelings of heaviness, as if there is something were stuck inside the top of the head.

To treat this condition, the doctor instructs the patient to close the eyes, raise the head slightly, while gently concentrating on the Lower Dantian. Have the patient imagine the energy block as a cube of ice. The ice melts into water, the water dissolves into vapor, and travels downward along the Conception Vessel toward the Lower Dantian.

FURTHER COMPLICATIONS

In balancing the patient's energetic body, the doctor first locates the Excess or Deficient condition, determines if the tissues should be tonified, purged, or regulated, then ascertains which energetic complications have initiated the patient's disease.

The doctor must also determine if the patient is sensitive to treatment or resistant to energetic therapy. Below is a summary listing several energetic complications and suggestions on how to rectify them.

1. For Compressed Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
2. For Contracted Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
3. For Bloated Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
4. For Expanded Stagnation, break up the stagnation, purge the pathogens, then tonify and regulate the patient's Qi.
5. For Energetic Armoring, access the patient's Heart (through the Shendao GV-11 area), and induce "soul retrieval" (see Chapter 19).
6. For Energetic Depletion, tonify and regulate the patient's Qi.
7. For Energetic Leakage, envelop the patient's tissues creating an energetic cast to stabilize the energetic fields.
8. For Migrating (Wandering) Qi Deviations, purge the pathogens and then isolate and stabilize before tonifying and regulating the patient's Qi.
9. For Psychosexual Deviations, root the patient's Heart, then regulate the patient's Taiji Pole with divine healing light.

CHAPTER 31
THE MAGIC SQUARE OF THE YELLOW RIVER

The configuration known as the Lo Shu (Lo Writings) or Magic Square of the Yellow River, is shown below. It is said to have been derived from markings on the shell of a "spiritual turtle" that crawled out of the River Lo when the Emperor Yu was draining off the floods. It is considered a mathematical model of the universe. The relationship between the numbers presented in this chart may be used to diagnose the energetic relationship between universal and environmental phenomena, as well as the various pathologies within the occurring internal organ correspondence.

It corresponds to the Later Heaven sequence of the trigrams of the Yi-Jing (I-Ching) and is arranged in accordance with the directions of the compass, with the number "5" at the center of the compass, in an arrangement known as the "Magic Square."

The numerals 1–9 are assigned to each position in accordance with the Lo Scroll markings and represent the Acquired Essences (Postnatal Jing).

Understanding the concept of balance is paramount in Qigong medicine. When Yin and Yang energies in the body are balanced, health is established. The Chinese have given numerical relationships to Yin and Yang properties representing this dynamic connection.

Yang energy is represented by all the odd numbers (1, 3, 5, 7, and 9). Yin energy is represented by all the even numbers (2, 4, 6, and 8).

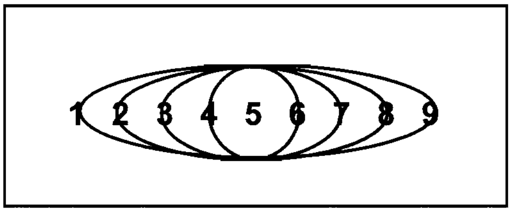

Figure 31.1. Yin and Yang Number Balance

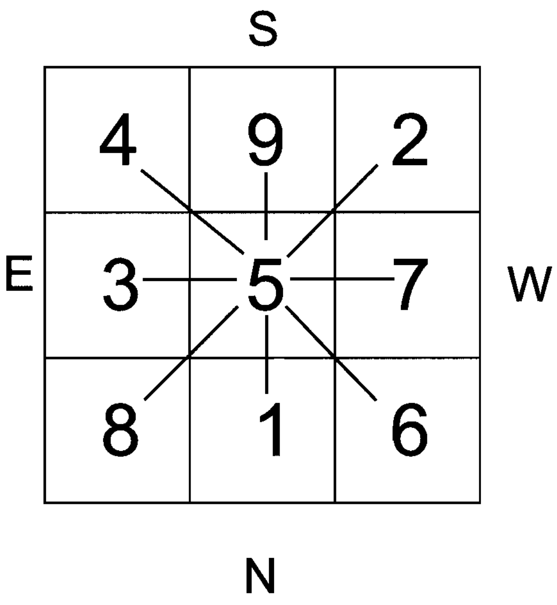

Figure 31.2. The Magic Square

The cycling of Yang (odd numbers) and Yin (even numbers) indicates the law of motion of Yin and Yang in all things. In Chinese medicine the numerical configuration is based on the sum totaling ten (Figure 31.1).

The numbers arranged together create what is known as the Magic Square. Investigation will uncover several numerological correspondences (Figure 31.2). The sum of any two numbers opposite each other in a row, column, or diagonal (with the number "5" between them) equals 10; thus the sum in any of these rows, columns, or diagonals equals 15. Fifteen is the number that represents "Man," who lives between Heaven above (the number being 20, which equals the sum total of the four primary directions) and Earth below (the number also being 20, which equals the sum total

SW = 2 } SE = 4}
NE = 8 } NW = 6}
10 + 10 = 20 Yin Earth Qi

S = 9 } W = 7}
N = 1 } E = 3}
10 + 10 = 20 Yang Heaven Qi

Figure 31.3. Heaven and Earth Numbers

of the four secondary directions). The numbers in the four primary directions (north, south, east and west) are odd and represent Yang Heaven Energy, while the numbers on the corners are even and represent Yin Earthly Energy (Figure 31.3).

EARTHLY YIN ENERGETIC FLOW

The numbers on the corners of the Magic Square are "terrestrial" in design. All the terrestrial numbers are even numbers. They represent the Earth's energetic field and flow counterclockwise in their circular progression. These terrestrial cycle phases are expressions of Earthly Yin energy and are designed in progressions of "times two." For example, (2) Fire: 2 × 2 = (4) Metal: 4 × 2 = (8) Wood: 8 × 2 = (16) Water (which is reduced to 6, as the ten position is not counted): 16 × 2 = 32 (which is reduced to 2). This completes one cycle in the terrestrial progression, which then returns to the element Fire and begins all over again in the same fashion, continuing in an endless progression of "times two." (Figure 31.4).

Through understanding the Yin energetic progression, the Qigong doctor rotates the patient's energy in a counterclockwise direction to purge Excess conditions and expel pathogenic Qi (called "Yin destroying Evils").

HEAVENLY YANG ENERGETIC FLOW

The numbers on the sides of the Magic Square are "celestial." All the celestial numbers are odd numbers. They represent the Heaven's energetic field and flow in a clockwise circular progression. These celestial cycle phases are expressions of Heavenly Yang energy and are designed in progression of "times three." For example, (1) North: 1 × 3 = (3) East: 3 × 3 = (9) South: 9 × 3 = (27) West (which is reduced to 7, as the ten position is not counted): 7 × 3 = 21 (which is reduced to 1). This completes one cycle in the celestial progression, which then returns to the North and begins all over again in the same fashion, continuing in an endless progression of "times three."

In understanding the Yang energetic progression, the Qigong doctor will rotate the patient's energy in a clockwise direction to tonify. This technique facilitates the continuous circulation of Yang Qi and is used in order to help the tonification of Qi and Blood and replenish Deficiencies (called "Yang promoting growth").

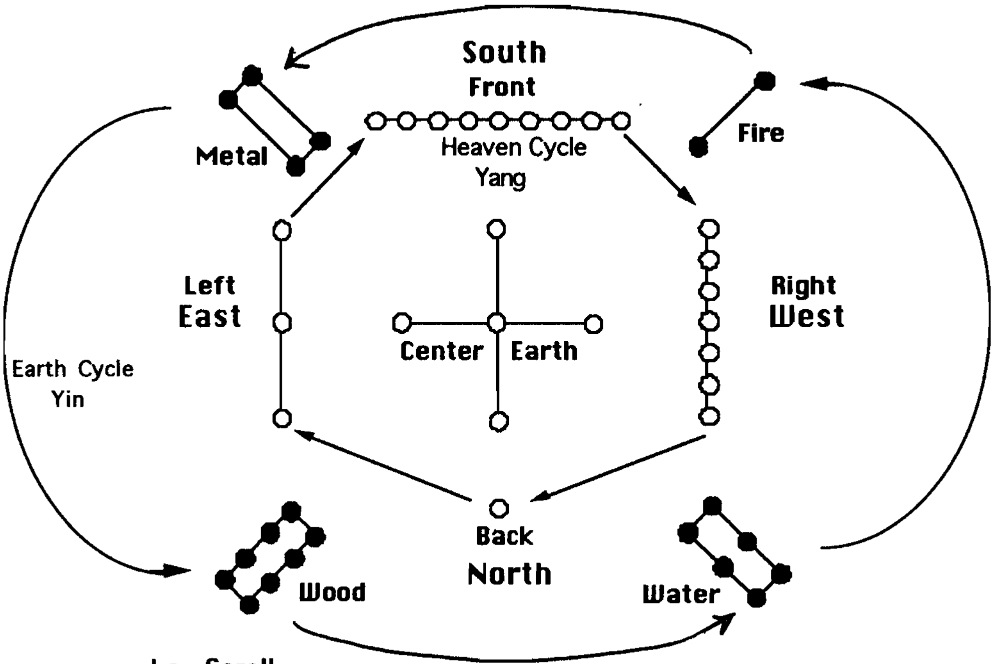
South
Front
Heaven Cycle
Yang
Metal
Fire
Left
East
Center
Earth
Right
West
Earth Cycle
Yin
Back
North
Wood
Water

Lo Scroll

	South East	South	South West	
	☴ (GB-41) 4 Belt Vessel	☲ (Lu-7) Heart 9 Fire Conception	(Kd-6) ☷ 2 Yin Heel	
East	Liver (TB-5) ☳ 3 Wood Yang Linking	5 Spleen Earth	(SI-3) Metal Lungs 7 ☱ Governing	West
	Yin Linking 8 ☶ (Pc-6)	Yang Heel (Bl-62) Kidneys 1 Water ☵	Thrusting 6 (Sp-4) ☰	
	North East	North	North West	

The Magic Square

Figure 31.4. The reason the South is positioned on the top of the page is because, in China, maps were specifically designed for the convenience of the Emperor, whose throne was always positioned facing South, the direction of Yang energy. As the magistrates of his court unrolled the maps before him, the East would be positioned on his left, the West to his right, and the South furthest away at the top of the map (whereas, in the Western culture North is at the top).

THE MAGIC SQUARE AND EIGHT TRIGRAMS

The following Eight Trigram Hand chart represents the terrestrial aspect of the body's acquired structures, evolution, and exchange of Yin / Yang energy.

It should be noted that if the trigrams associated with the numbers shown in the Magic Square are placed in their correct positions, the postnatal or "Later Heaven" trigram arrangement will form. The middle number "5" does not correspond to a trigram because it is the unifying number placed in the center.

In order to use the Eight Trigrams in conjunction with the Magic Square, the patient's palm is divided into nine sections. The Eight Trigrams form the exterior of the palm, while the heart of the palm (known as the Bright Hall or Ming Tang) aligns the Magic Square to the patient's internal organs. The Qigong doctor observes the color, luster, sinking or bulging of skin, as well as any prominent blue veins appearing on the surface of the patient's skin. These observations are compared to the organs and functions associated with each trigram that rules that particular area of the patient's palm (Figure 31.5).

1. The Trigram Kan, located at the base of the palm's heel, is associated with the patient's Kidney and Bladder organs, as well as the function of their urogenital and reproductive systems.
2. The Trigram Kun, located on the palm just below the little finger, is associated with the patient's left Lung organ and the left side of the chest (the left breast in women).
3. The Trigram Zhen, located on the upper part of the patient's thenar eminence, is associated with the function of the patient's Liver and Gall Bladder organs.
4. The Trigram Sun, located on the palm, just under the index finger, is associated with the patient's right Lung, and the right side of the chest (the right breast in women).
5. The Bright Hall (Ming Tang), located on the center of the palm is associated with the patient's Heart, blood vessels, and mental health.

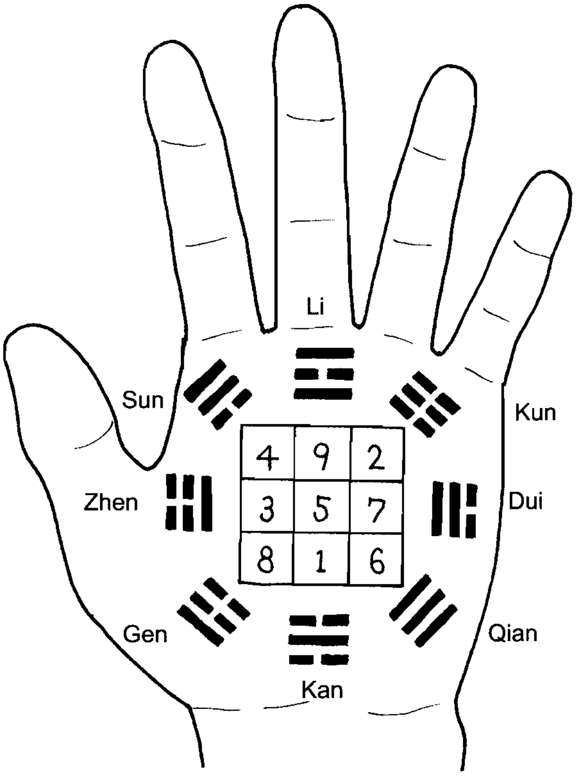

Figure 31.5. The "Magic Square" can be superimposed on the human body for observation, diagnosis, and treatment. For example:

2 and 4 represent the location of the Lungs

3 and 7 represent the location of the Liver and Spleen

8 and 6 represent the location of the Kidneys

9 represents the location of the Upper Burner

5 represents the location of the Middle Burner

1 represents the location of the Lower Burner

6. The Trigram Qian, located on the lateral edge of the palm's heel, just above the wrist crease, is associated with the patient's descending and sigmoid colon.
7. The Trigram Dui, located on the lateral edge of the palm's heel, midway between the little finger and the wrist crease, is associated with the patient's Spleen, pancreas, and Stomach organs.
8. The Trigram Gen, is located on the lower half of the thenar eminence, is associated with

Trigram Correspondences

Trigram	Name	Meaning	Channel	Shu Point	Direction	Number	Season
☲	Li	Fire (Clinging)	Conception	Lu-7	South	9	Mid Summer
☷	Kun	Earth (Receptive)	Yin Heel	Kd-6	South West	2	Start of Fall
☱	Dui	Lake (Joyful)	Governing	SI-3	West	7	Dividing of Fall
☰	Qian	Heaven (Creative)	Thrusting	Sp-4	North West	6	Beginning of Winter
☵	Kan	Water (Abysmal)	Yang Heel	Bl-62	North	1	Absolute Winter
☶	Gen	Mountain (Keeping Still)	Yin Linking	Pc-6	North East	8	Start Of Spring
☳	Zhen	Thunder (Arousing)	Yang Linking	TB-5	East	3	Dividing of Spring
☴	Sun	Wind (Gentle)	Belt	GB-41	South East	4	Start of Summer

Figure 31.6. The Eight Trigram Correspondence

patient's appendix and ascending colon.

9. The Trigram Li, located on the palm between the middle and ring finger, is associated with the patient's blood circulation, Heart and eyesight.

THE MAGIC SQUARE AND THE EIGHT EXTRAORDINARY VESSELS

In Chinese medicine, the Lo Writings and their correspondences form the model which is the basis for the Ling Gui Pa Fa (The Eight Techniques of the Mysterious Turtle) system of acupuncture. In this system there are Eight Confluential Points that are associated with the Eight Trigrams (Figure 31.6). These points are where the Eight Extraordinary Vessels and Twelve Primary Channels intersect each other.

This method describes the theory that all the points on the body are dominated by the Shu Points, which are in turn dominated by the Eight Confluential Points. The confluential points and their trigram correspondences are Kd-6 (Kun Gua), Lu-7 (Li Gua), GB-41 (Sun Gua), TB-5 (Zhen Gua), Pc-6 (Gan Gua), Sp-4 (Qian Gua), SI-3 (Dui Gua), and Bl-62 (Kan Gua).

UNDERSTANDING THE METAPHYSICAL ASPECT OF NUMBERS

The Chinese medical text of the Nei Jing explains numerology and its mystical significance, as well as the practical utilization of the correspondences between numbers and natural phenomena. Man's physical, mental, emotional, energetic, and spiritual form manifest as part of the microcosm of these nine correspondences, and are explained as follows.

1. Number One corresponds to Heaven, the Center Taiji Pole, and represents the skin.
2. Number Two corresponds to Earth and represents the muscles. The number two also cor-

responds to the two primary forms of Yin and Yang.

3. Number Three corresponds to Man and represents the blood vessels. The number three also corresponds to the Triple Burners, Three Powers, Three Treasures, Three Dantians, and Three External Fields of Wei Qi.
4. Number Four corresponds to the four seasons and represents the muscles corresponding to the body's four extremities, the Four Earthly Quadrants and the Four Divisions of Heaven (N.S.E.W.), and the Four Prenatal Trigram Essences (Great Yang, Small Yang, Great Yin, and Small Yin).
5. Number Five corresponds to the Five Notes of the Chinese pentatonic scale and represents the voice. The number five also corresponds to the Five Elements, Five Colors, Five Tastes, Five Planets, and the Five Elemental Constitutions.
6. Number Six corresponds to the Six Divisions (Tai Yang, Shao Yang, Yang Ming, Tai Yin, Shao Yin, Jue Yin) and represents the union of Yin and Yang. The number six also corresponds to the Six Yao lines and their correspondence to the body, and the Six Accomplishments (intelligence, humanity, holiness, sincerity, moderation, and benevolence).
7. Number Seven corresponds to the Seven Luminaries (sun, moon, and the five planets), which represent the Seven Openings of Man (two eyes, two ears, two nostrils, and one mouth). The number seven also corresponds to the Seven Emotions, and the Seven Stars of the Big Dipper.
8. Number Eight corresponds to the Eight Extraordinary Channels from which the body's prenatal energy flows and represents the breath. The number eight also relates to the Eight Trigrams and their various correspondences.
9. Number Nine corresponds to the formless Void or Wuji, as well as the Nine Heavens (one is centered, and the others are in the eight separate divisions of the Lo Pan compass) and the Nine Lo River Chart (Magic Square). The number nine also represents the nine openings of man (two eyes, two nostrils, two ears, one mouth, one anus, and one urethra).

THE FIVE ELEMENT ENERGETIC FLOW OF THE MAGIC SQUARE

The number formations on the Magic Square are not only considered the energetic ebb and flow, evolution and exchange of Yin and Yang Qi, but also correspond to the energetic harmony of the Five Elements. The energetic harmony of the Five Elements are arranged as follows.

- The Kidneys are positioned in the North; their element is Water and number is 1.
- The Liver is positioned in the East; its element is Wood and number is 3.
- The Heart is positioned in the South; its element is Fire and number is 9.
- The Lungs are positioned in the West; their element is Metal and number is 7.
- The Spleen is positioned in the Center; its element is Earth and number is 5.

These five numbers are arranged in accordance to their energetic progression and are used as a templet for observation and diagnosis of the patient's body, according to their Five-Elemental postnatal structure.

CROSSOVER ENERGY PATTERNS

According to the theory of energetic harmony, when one area of the body lacks Qi, there is generally an Excess energy accumulation in another area. Because the body is symmetrical, one can naturally look on the opposite side for the misplaced Qi. The Qigong doctor may dredge the full side in order to drain the Excess and support the Deficient area.

If the imbalance is not on the opposite side, the doctor can scan for Excess Qi using a figure eight pattern (Figure 31.7). Classically, the Crossover Energy Pattern is one of the most powerful pathways by which Qi flow in the body can be stimulated, and is excellent for balancing internal body energy. The figure "8" pattern can also be used in order to balance all sides of the body (top and bottom, as well as left and right).

When the Qigong doctor must balance the energy in the left and right sides of the brain, the patient may be assigned homework that requires them moving Qi in a figure "8" pattern sideways (i.e., drawing the infinity symbol). This exercise prescription stimulates, strengthens, and balances the right and left hemispheres of the patient's brain. It is used successfully in the treatment of patients with dyslexia.

When patients have internal problems that affect the function of the whole body, it may be necessary to treat two areas simultaneously. The Spleen and Stomach areas, for example, may be tonified along with the Lung areas to strengthen and increase the patient's resistance to External pathogenic factors.

Because organ imbalances are deeper, and tend to be the result of years of energetic deviations, they may take longer to reverse.

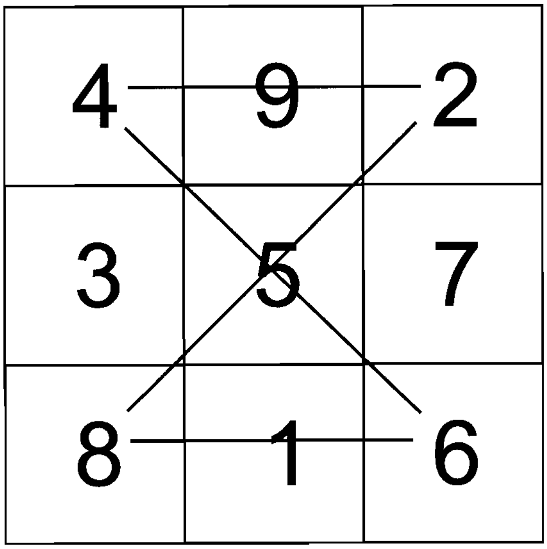

Figure 31.7. In this Crossover Energy Pattern, area 4 is treated using area 6 and so on.

CHAPTER 32
HEALING THE PATIENT'S EMOTIONAL TRAUMAS

INTRODUCTION

In my personal practice I feel the main problem for patients is in dealing with their own "hurt," which may have been inflicted upon them several years prior to the manifestation of their disease. It is a recognized truth that when people isolate themselves from each other, their environment, and their spiritual life, it is because they feel the need to guard themselves.

They guard and protect themselves and their belief structures from invasion, criticism, exploitation, or attacks, whether real or imagined, allowing interchange only if it supports their beliefs.

People build energetic walls, and when the walls are penetrated they often experience anger, self-hatred, pain, fear, and shame. Consequently, masks and defence mechanisms are created to hide and support these hidden belief structures.

How the patients develop and maintain their energetic patterning will determine which clinical modality the Qigong doctor will utilize to initiate healing. To free themselves from personal hurt the patients must reconnect with their anger, pain, fear, or worry. Eventually, with the unconditional loving support of the doctor, patients are able to reintegrate the hurt and anger, and accept parts of themselves they had rejected for so long. This acceptance of painful feelings and memories affects the Heart and frees the patients from the emotional traumas of the past.

To heal and release these pathogenic emotions, the patients go through a five-stage detoxifying process during which they feel, experience, and release the emotional trauma that initiated their disease. This is a natural progression in the process of Medical Qigong therapy and healing.

THE PHYSIOLOGICAL EFFECTS OF BENIGN AND CHRONIC EMOTIONS

Within the clinic it is observed that benign emotions (love, joy, peace, etc.) bring about a calm and restful condition within the patient's physiology. Chronic emotions however, create a detrimental stressful physiology, adversely affecting the quality of the patient's life-force energy. As discussed previously, emotional internal and external environmental factors can and do affect the quality of the body's Qi.

1. Excessive joy from the Heart slows down the Qi.
2. Excessive worry and thinking from the Spleen stagnates the Qi.
3. Excessive sadness from the Lungs obstructs the Qi.
4. Fear from the Kidneys descends the Qi.
5. Excessive anger from the Liver ascends the Qi.

FIVE LEVELS OF FEELING AND THEIR GOVERNING ORGANS

To complete their initial healing, the patients go through five stages of emotional resolution. These stages act as a template for the patients' emotional healing, allowing them to progress through, and detoxify each of their Five Yin Organs. The Five Yin Organs store the patients' emotional history in their tissues. Each of the Five Yin Organs also affects different functions and parts of the body (Figure 32.1).

1. The Liver governs the emotions dealing with the past. The Liver stores the negative emotions of irritability, jealousy, anger, rage, blame, and resentment. These emotions and mental states affect the flow of Qi in the channels, influencing the patient's tendons, ligaments, nails, small muscles, peripheral nerves,

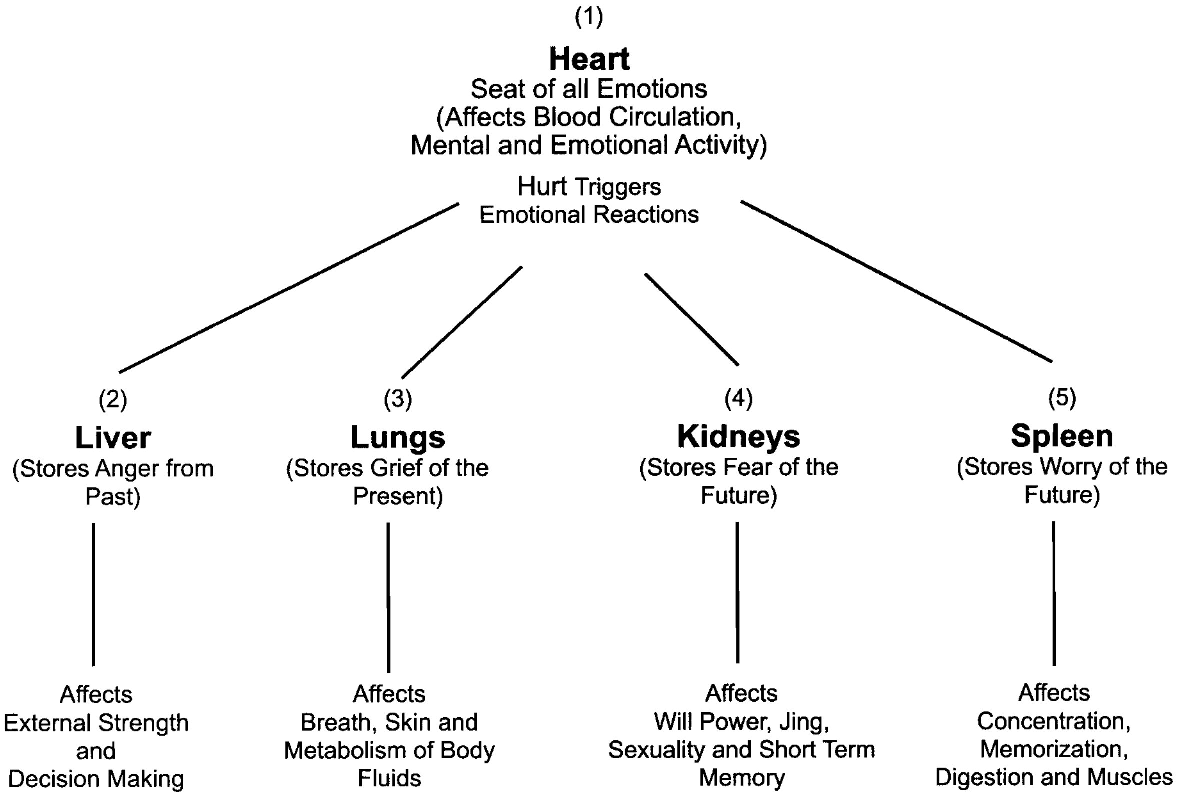
(1)
Heart
Seat of all Emotions
(Affects Blood Circulation,
Mental and Emotional Activity)
Hurt Triggers
Emotional Reactions
(2)
Liver
(Stores Anger from Past)
Affects
External Strength and Decision Making
(3)
Lungs
(Stores Grief of the Present)
Affects
Breath, Skin and Metabolism of Body Fluids
(4)
Kidneys
(Stores Fear of the Future)
Affects
Will Power, Jing, Sexuality and Short Term Memory
(5)
Spleen
(Stores Worry of the Future)
Affects
Concentration, Memorization, Digestion and Muscles

Figure 32.1. Diagram of Five Elemental Emotional Transitions

external strength, the iris of the eyes, vision, tears, the Gall Bladder, bile, decision making, the external genitalia, and the Hun (Three Ethereal Souls).

2. The Lungs govern the emotions dealing with the present. The Lungs store the negative emotions of grief, sorrow, sadness, anxiety, shame, guilt and disappointment. These emotions and mental states affect the flow of Qi in the channels, as well as the patient's breath, skin, Large Intestine, mucous membranes, the metabolism of the Body Fluids, body hair, smell, and the Po (Seven Corporeal Souls).
3. The Kidneys govern the emotions dealing with the future. The Kidneys store the negative emotions of fear, loneliness, and insecurity. These emotions and mental states affect the patient's energetic channel flow, as well as the Jing, bones, brain, bone marrow, inner ear, hearing, head hair, pubic hair, spinal cord, cerebrospinal fluid, ovaries, testes, anus and urethra, sexual fluid, reproduction, short term memory, and the Zhi (Will Power).
4. The Spleen governs the emotions dealing with resolution and responsibility. The Spleen also stores the negative emotions of worry, remorse and regret, which can create obsessions. These emotions and mental states affect the flow of Qi within the channels, as well as the patient's large muscles, lymph and saliva secretions, the mouth, lips, taste, the Stomach, digestion, concentration, the ability to memorize, and the Yi (Intention).
5. The Heart governs the emotions dealing with

shock, nervousness and excitement, as well as the final stages of healing (i.e., peace, love and forgiveness). These emotions and mental states affect the patient's energetic channel flow, as well as the patient's blood vessels and circulation, complexion, perspiration, mental and emotional activity, long term memory, Small Intestine, and the Shen (Spirit).

In the process of forgiveness, the anger and resentment are traversed to reach through to the levels of original hurt, sadness and grief. Only after reaching this level, and working through the pain, can the patient begin the process of letting go of the anger, blame, and resentments that protect them from feeling the full impact of the original trauma. Forgiveness begins only after the anger and crying ends. There are three stages of forgiveness: forgiving oneself, accepting the situation, and forgiving the abuser.

1. The forgiving oneself stage deals with the energy involved in healing oneself, and is divided into two specific stages:
   - Forgiving oneself for allowing the hurt to occur in the first place (even if the patient had no choice in being victimized), or for not fighting back strongly enough, if at all, and
   - Forgiving oneself for hanging on to all that anger and pain for so many years.
2. The accepting the situation stage deals with releasing the energetic emotional charge within the environment, or the unwanted consequences connected to the original trauma (e.g., forgiving the accident, injury, handicap, illness, trauma) within the patient.
3. The forgiving the abuser stage is the most difficult final step. Many factors can interfere with this process, such as the patient's fear of re-abuse or continued abuse if forgiveness is genuinely given. Another major factor is the negative, toxic, external energy projected by the abuser upon the patient, sometimes over many years. This energy is absorbed by the patient over time and often leads to the patient identifying with the abuser. The fear of being, or becoming, just like the abuser can be overwhelming. Anger at the abuser may seem to be the only thing standing between them; and allowing any feelings of love, compassion, or understanding may appear to be an acceptance of "evil," not only within the abuser, but within the patient's own self. An acceptance of a dark side within each of us is necessary to forgive the abuser. Often victims of abuse cannot tolerate the idea that they are anything like their abuser. Only the conscious mind, working in concert with the patient's spirit, can reintegrate this essential part of the self that has been utterly rejected and denied.

Children who have experienced abuse may be especially hard pressed to accept anger as a natural, integral part of themselves without which they cannot feel whole. They frequently view all and any anger as not only dangerous but an "evil" force. Thus, forgiveness of the abuser requires restructuring of belief systems, and the acceptance of one's own dark side, as well as the abuser's.

Through conscious intent and creative subconscious intuition, it is possible to come to an understanding that the "evil" experienced stems not from the devil, but from ignorance and pain. In most cases the abuser's intent was not to harm and destroy, but ironically, to teach and discipline for the good of the patient. The abuser did the best that he or she could at the time, given all the unresolved pain and anger within the self.

When patients experience extreme abuse, or prolonged abuse, they may lose the ability to distinguish the difference between wrong actions and the abuser. The abuser is not evil incarnate though evil deeds were done. This often leads victims to see the world only in terms of black and white with no shades or colors. The true healing requires acquiring the ability to see anew (as a little child), to reawaken to the amazing diversity of life.

These skills sometimes require teaching the patient and may necessitate referring the person out to a professional, especially if the patient suffers from a character disorder (e.g., borderline personality disorder, schizo-affective disorder, and so on).

Generally, after performing specific Qigong exercises and meditations which release the patient's pent-up, toxic emotions, the patient begins to come out of denial and awaken to genuine spontaneous emotions and feelings. This transition can be very painful at first. I usually tell the patients that this transition is similar to having their arm frozen, and then thawed out.

Imagine if the arm is placed into ice cold water, and left there. The first thing the body would feel is shock, followed by pain. As the feelings slowly begin to leave, the pain begins to diminish, and the arm becomes numb. As the mind shifts its attention away from the feelings in the arm, it begins to establish a denial system, ignoring, minimizing, or suppressing the condition in which the arm has been placed. When first removing the arm from the ice water, the tissues are still numb. After awhile, however, as the numbness begins to wear off, feelings begin to return to that particular area of the body, and pain is felt. The body is no longer in shock and is now feeling the pain and the memory of the past experience.

Our lives are the same way, in regard to physical, mental, emotional, and spiritual pain. While growing up, we are subject to constant trauma on all four levels of existence. Through an instinctive ability, we adapt and learn to become numb to our true feelings in order to survive. When we begin to awaken, we are suddenly thrust into shock, then pain, as we begin to face the source and cause of our trauma.

From a Traditional Chinese Medical perspective, pain is a means by which the body communicates its needs to the mind. In cases of chronic pain, the patient generally learns to divert awareness away from the area of pain through distracting thoughts and emotions. This helps the patient to cope with the pain, but also interferes with the healing process. If the patient reduces conscious awareness of the painful area(s) in the body, the mind/body connection becomes blocked and slows or prevents healing. The initial clarification of the "Message" regarding the disturbance can cause both a sensation of increased pain and an activation of the healing process.

By comprehending the five levels of feeling and their progressive transitions, the patient gains understanding and healing from these past traumas.

STAGES OF EMOTIONAL HISTORY AND RESOLUTION

After the patient understands the five levels of feeling and how they affect their respective organs, the Qigong doctor can assist patients in the cognitive restructuring of their disease-causing belief structures. This emotional re-patterning is initiated through the unveiling of several stages of painful emotional history, which affect the energetic healing abilities within the patient's body.

It is important for the Qigong doctor to remember that all physical matter is conscious energy activated and focused by consciousness. The patient's perceptions create the physical presence of the disease. Since all matter and energy is unified, all matter down to the smallest subatomic particle has awareness, makes choices, and communicates simultaneously with all other matter in the universe. Choices are initiated in the patient's body through the active fields of environmental and universal energy.

Understanding these subatomic communications allows the Qigong doctor to assist the patient in discharging toxic energy trapped and stored within the tissues. The experience of the emotional trauma, and its destructive emotional energetic patterning, is clinically explained and described in the following stages.

1. The patient is hurt (physically, mentally, emotionally, and/or spiritually). The patient receives the original trauma through either an aggressive, direct attack (i.e., being struck, screamed at, belittled, or shamed, etc.), or by a passive, indirect attack (i.e., being abandoned, rejected, ignored, betrayed, etc.). In this stage of development, the patient's trusting connection to life is violated.
2. The patient registers the hurt, feels the pain, and naturally expresses this trauma in an outburst of emotion (i.e., screaming, crying, laughing, groaning, etc.). In this stage the patient recognizes the wounding.
3. The patient seeks attention and looks for a

human emotional response, seeking validation and self-affirmation. In this stage, if the patient does not seek and receive attention for healing the trauma or, if the individual is not free to express his or her real feelings, healing will not take place.

4. The patient's hurt remains unresolved and becomes history. At this point in the transition it is important that the patient feel the hurt, understand the circumstances surrounding the hurt, then discharge the pain. If the patient's hurt does not become completely discharged from the body, it becomes stored within the patient's tissues. Thus, the patient's hurt becomes internalized; the emotions of unexpressed anger, fear, pain, and shame begin to accumulate in the body. The trauma remains as an active charge of unexpressed emotional energy.
5. The patient's whole body stores the distressful emotional memories. This distress gets restimulated in the present (through various trigger mechanisms), and alters or blocks the patient's flow of energy and thinking.
6. The patient learns defense mechanisms to survive. The patient develops strong defense mechanisms instead of healing from the emotional wounds (known as a pathological recovery). At this point the patient's denial system is reinforced, and kept active to protect the patient from any further hurt, in order to ensure survival and maintain at least the semblance of a normal, functional life.
7. A dynamic polarity arises that seems to take on a life of its own. The patient swings back and forth between overcontrolling emotions and loosing control over them completely. The patient becomes obsessed with the painful feelings, either hiding from them or becoming overwhelmed by them. The patient then chooses either one of two options:
   - To become isolated, emotionally removed and distant, or
   - To become needy, co-dependent and attention hungry.
8. The patient develops addictions to continually hide from the pain. The patient's actions become based on the unhealed feelings resulting in four overlapping ways to hide:
   - Physically through sports, excessive exercise, food, sleeping, etc.,
   - Mentally through compulsive behavior (e.g., studying, being a work-a-holic, compulsive book reading, etc.).
   - Emotionally through either chronic rage, grief, obsessive worrying, and chronic fear, or through a compulsive search for love, joy, and excitement (sex / love addiction), etc., and
   - Spiritually through fanaticism of various types, or extreme spirituality.
9. The patient begins to recover from the trauma. To heal and recover, the patient must cease concentrating on the pain and redirect his or her full attention towards setting healthy boundaries and positive affirmations. Some patients' boundaries are too solid and immovable (i.e., they won't allow for any meaningful interaction with others). These patients do not really need to set boundaries, they need to allow their boundaries to become more permeable. Patients' boundaries need to be flexible in order for them to be healthy. Patients must learn when to open and when to close their boundary system, and not just constantly shut everything out (people, relationships, responsibilities, etc.) or let everyone and everything in.

It is important for patients to understand that perpetrators of abuse are in a state of denial of their own shame and imperfection; thus they inflect pain and suffering onto anyone present at the time. Some perpetrators single out one family member as a recipient for their pain. Sometimes the chosen victim reminds the perpetrator of his or her younger innocent self; thus all the accumulated self-hate is projected onto this one child through no fault of the victim. The resemblance may be real (slight physical similarities for instance) or imaginary (the child's vulnerability, for example, or a singled out personality trait that triggers the abuser's anger). The one common factor among abusers is their inability to see their victims for who they really are. Abusers make no effort to know the victim, to really see the whole person; instead they see a distorting mirror which reflects their own image back to them, magnifying their own flaws and fears. Through setting boundaries and affirming their own self-worth, the patients will be able to release years of toxic emotions stored within their tissues. This can initiate health and healing on every level of their being (physical, mental, emotional, and spiritual).

ASSISTING THE PATIENT IN ESTABLISHING HEALTHY BOUNDARIES

The patient is generally informed that in order to make it through the spiritual, emotional, and mental transitions brought about by Medical Qigong therapy, it is important to establish an active but flexible boundary system. The patients are then taught to divide their relationships into three categories: True Friends, Acquaintances, and Toxic Relationships (Figure 32.2).

1. True Friends consist of people with whom the patient can share and expose his or her most intimate hidden secrets and emotional feelings, without worrying about being judged. In the presence of true friends, the patient feels emotionally accepted and loved unconditionally. The boundary system should be down, and the patient should feel completely open. It is safe to be completely vulnerable.
2. Acquaintances consist of people with whom the patients should share only limited amounts of information regarding their personal/emotional life experiences. The most hidden secrets could be judged, misconstrued, and gossiped about. All conversation with these individuals should, and must be, kept on a fairly superficial level. The patient's boundary system should be up, and the patient should monitor the conversation, as it is not safe to be totally vulnerable.
3. Toxic Relationships consist of people with whom the patient cannot and should never share any knowledge of their personal/emotional life. These individuals chronically abuse, mistreat, belittle, and attack others (whether they are members of the patient's own family, work, or neighborhood). All conversation with these individuals should be kept to a minimum. The patient's boundary system should be on full armor, and whenever possible, these individuals should be avoided.

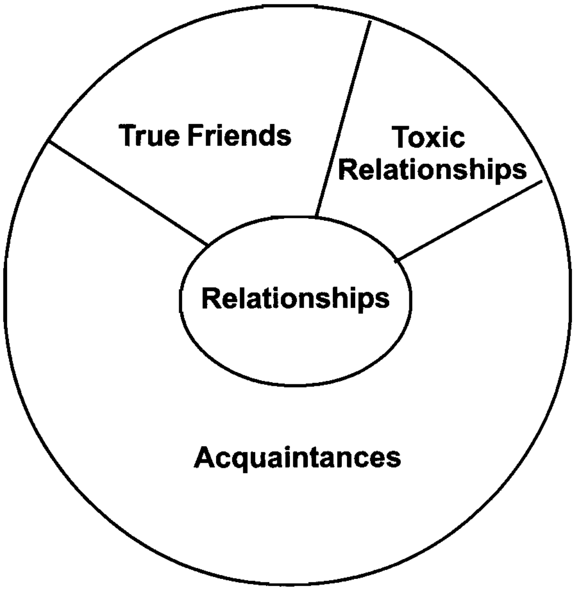

Figure 32.2. Types of Relationships for Setting Flexible Boundaries

REJECTION, BETRAYAL, AND ABANDONMENT

From a young age, people are introduced to the multidimensional experience of being hurt (physical, mental, emotional, and spiritual). This affects the individual's center core and initiates a coping response due to feelings of rejection, betrayal, and abandonment. The normal reaction to such emotions results in experiencing emotions ranging from fear to frustration. These individuals begin to develop masks, that warp and scar their developing personality. As these individuals continue to grow, the illusions of these masks demand more Qi to sustain their form, which in turn drains vital energy from the growing body, and weakens the immune system. This denial system, fortified by the individual's emotional patterning,

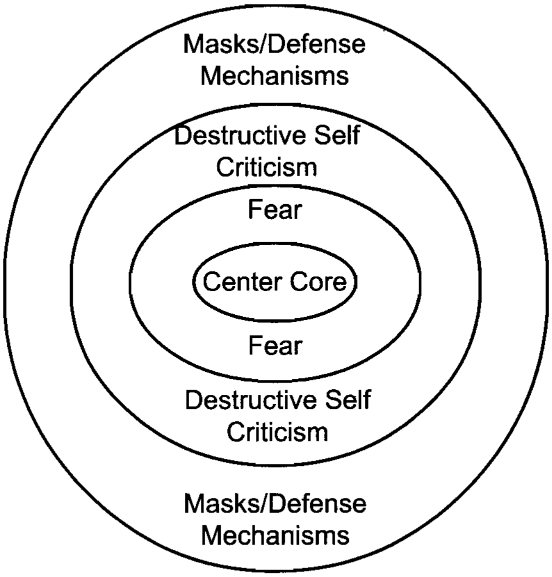

Figure 32.3. The True Self is contained within the Core Self.

sustains him or her until they can heal and re-pattern the no longer beneficial belief system.

To re-pattern their belief systems, patients and Qigong practitioners must first be made aware of their initial programming patterns and the reasons for which these were created in the first place. The following "spiritual template" is applied to all patients, regardless of culture, race, or religion.

To access the patient's center core, the Qigong doctor must help the patient identify the masks which have taken on a life of their own and created the false "ego self." Next, the patient must address personal self-hate issues, which have driven the patient to develop the masks and defense mechanisms. Finally, the patient must face the foundational fear that has severed his or her connection with their "true self" contained within the center core (Figure 32.3).

In the Medical Qigong clinics in China there is a saying, "Patients must first stop running from their fears, and face the "Paper Tiger" (i.e., the illusions and fantasies that have been chasing them); in the West, it is considered "stopping and facing your internal demons."

MASKS AND DEFENSE MECHANISMS

To armor the ego, individuals resort to what Sigmund Freud labeled as "ego defense mechanisms." There are nine strategies; and with the exception of the last one listed, i.e., "sublimation," they serve to mask the core self. The nine defense mechanisms and their strategies are: Repression, Displacement, Projection, Intellectualization, Regression, Fixation, Denial, Reaction-Formation, and Sublimation.

1. Repression is the pushing down of unwanted ideas and emotions into the unconscious;
2. Displacement is the shifting of impulses aroused by one person, or situation on to a safe target;
3. Projection is the attribution of unacceptable impulses within oneself to other people;
4. Intellectualization is an elaborate rationalization of a naked impulse, to justify it;
5. Regression is the return to an earlier childhood stage of behavior to reduce the demands on the ego;
6. Fixation has the same result as Regression, but the person becomes fixated at a particular stage of mental and emotional development;
7. Denial is a conscious refusal of an impulse-evoking fact, feeling or memory;
8. Reaction-Formation is the conversion of one feeling into its opposite, typically seen in love turning into hate;
9. Sublimation is the channeling of unacceptable impulses into acceptable, refined social forms and is the only defense mechanism considered to be a healthy reaction.

The ego mediates between the "id" (i.e., the primordial survival drives and impulses) and the "super ego." The super ego consists of the ideals and conscience of the individual.

EMOTIONAL SECURITY

Each individual has basic emotional needs that must be met to feel secure. Without the comfort of these needs being met an individual will sometimes lash out and react in an antisocial behavior. This reaction is generally due to the individual fostering feelings of rejection, betrayal, and abandonment. The three personal needs required

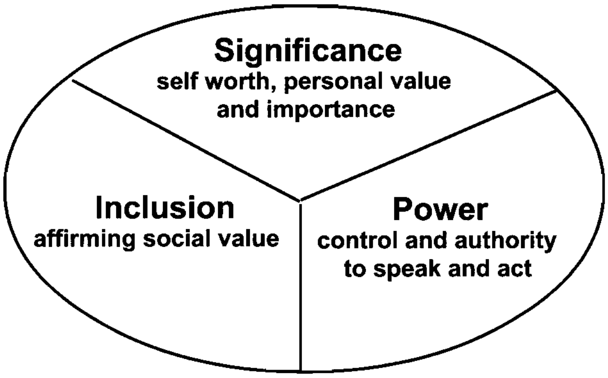
Significance
self worth, personal value and importance
Inclusion
affirming social value
Power
control and authority to speak and act

Figure 32.4. The Three Personal Needs That Foster Emotional Security

to foster emotional security are as follows: Significance, Inclusion and Power (Figure 32.4).

1. For an individual to feel significant, he or she must be given self worth, personal value and importance. Without personal value the individual will feel insignificant, unimportant, and worthless.
2. For an individual to feel included, he or she must be given affirming social value (knowing that they play an important role in the family, relationship, or society as a whole). Without personal acknowledgment the individual feels isolated, segregated, and separated.
3. For an individual to feel powerful, he or she must be given the ability to control and the authority to speak and act. Without personal power the individual feels powerless, weak, and incapable.

In addressing the patient's disease, it is important to sometimes address his or her living and work environment, as both can contribute and influence the patient's healing potential.

HEALING EMOTIONAL TRAUMAS

There are several techniques and methods used to assist patients in healing through emotional traumas (Figure 32.5). Such meditations as Pulling out the Pain are often prescribed as homework along with other treatment methods of Medical Qigong therapy. Some of these methods used to heal emotional trauma are described as follows.

1. Medical Qigong Distance Therapy affects the body, mind and emotions. The Qigong doctor alters the patient's energetic tissues and fields to re-pattern the patient's emotional responses.
2. Medical Qigong Massage Therapy and Bodywork involves lightly touching and stimulating the patient's tissues, to release the trapped emotions so that they may be appropriately reprocessed.
3. Medical Qigong Respiration Therapy unites the body, mind, and emotions by moving the energy through specific breathing techniques. By changing the breathing pattern, the Qigong doctor changes the patient's energetic emotional patterns, releasing trapped memories deep within the patient's tissues.
4. Medical Qigong Dynamic Postural Therapy expands and stretches the patient's tissues. This active adjustment of the patient's body alignment increases the flow of Qi and Blood through his or her internal organs, facilitating an emotional release (e.g., the Dry Crying exercise).
5. Medical Qigong Imagery and Meditation Therapy uses awareness meditations and images to sedate the patient's reactive mind in order to connect with the higher self-identity. The doctor interrupts the patient's emotional

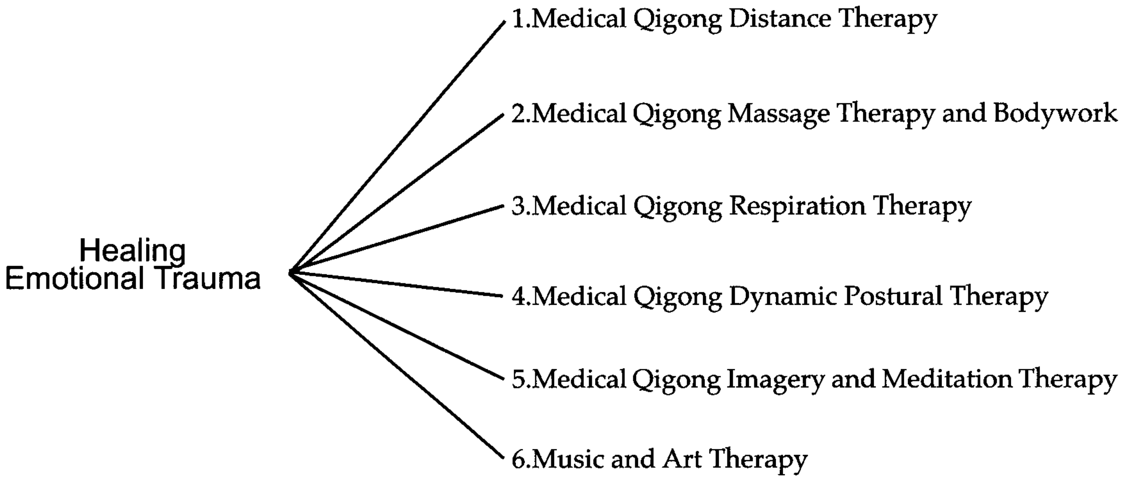
Healing Emotional Trauma
1.Medical Qigong Distance Therapy
2.Medical Qigong Massage Therapy and Bodywork
3.Medical Qigong Respiration Therapy
4.Medical Qigong Dynamic Postural Therapy
5.Medical Qigong Imagery and Meditation Therapy
6.Music and Art Therapy

Figure 32.5. Six Methods Used For Healing Emotional Trauma

thought patterns by changing the patient's sequence of sensations, word pictures and feelings. This emotional reforming changes the meaning of the patient's interpretation and challenges his or her self-limiting beliefs.

6. Music and Art Therapy are sometimes prescribed to change the patient's mood and awareness. Through playing a musical instrument, dancing, singing, or creative free-association drawings, sculptures, pottery, and so on, patients can access their Creative Subconscious Mind to repattern old belief structures and enhance the body's energy. Some patients will resist this form of therapy as they do not consider themselves to be good artists. It is vital to explain that these creative endeavors are not for the purpose of creating art, but for expressing feelings and emotions. These art forms need not be pretty, perfect, or shared with anyone if the patient does not desire to do so. They need not even be kept so long as the patient has released some feelings and emotions. The process itself is the healing. This type of therapy is used to help the patient recreate, identify, and release trapped emotions deep within the subconscious mind.

PULLING OUT THE PAIN MEDITATION

This meditation uses part of the guided imagery used in the Soul Retrieval Meditation (found in Chapter 19). It releases and transforms the patient's conscious feelings and helps to change or remove the Personal Subconscious toxic holding patterns.

After the meditation, patients should feel as if they have been cleansed and purified of toxic emotional debris. This is, however, only the first level. Patients can only release and transform the surfacing emotions one level at a time. Like peeling an onion, the patients must continue to work through each level until reaching their core issues. These core issues surround the suppressed feelings of hurt and pain.

The doctor instructs and guides patients through this meditation matching the rhythm of his or her words to the patient's exhalations. Patients draw their painful memories and traumas out of their body in the form of dark smoke; they then separate the distilled wisdom and knowledge from this dark cloud, reabsorbing the healing light into every pore and cell of their body. The dark cloud must then be released with the help of an angelic being who severs the energetic cord(s) still attaching the patients to their pain and suffering. When the patients reach this sixth step, the doctor assists them by clapping his or her hands as the patients visualize the "sword of truth", held by this holy being, severing the dark energetic cord(s). The sword of truth represents the patients' decision to regain their power by letting go of their attachment to their illness. This attachment creates very real energetic cords that bind the patients to their painful memories and toxic emotions. Severing these cords, or cord, symbolizes the end of the patients' investment in their disease, and the beginning of true healing. This releases all the pain and despair, i.e., the black cloud ascending into the Heaven into the hands of God or a "higher power," in accordance with the belief structure of the patients. As the cloud reaches the first wave of love and compassion emanating from God or a "higher power," a mighty explosion transmutes the black despair and pain into divine healing light. The doctor then assists the patients in accepting this healing, loving light back into their body to experience, perhaps for the first time in their lives, true forgiveness and a state of grace.

1. Begin by sitting comfortably in a chair, with both feet on the floor, hands resting on your thighs, eyes closed, tongue on the upper palette. Breathe naturally through the Lower Dantian.
2. Focus your attention on the Yellow Court area below the diaphragm where you store painful memories and trauma. Imagine opening up this area to release the toxic energy out of your body as a stream of dark steam releasing from a pot.
3. As the dark steam flows out of your body you are releasing feelings and memories of guilt, anger, rage, humiliation, abandonment, degradation, rejection, insecurity, and sorrow. These toxic emotions have, until now, prevented your healing through creating a deep distrust of self and others.
4. Focus your attention on this energetic cloud, and begin separating the pain and hurtful memories from the knowledge and wisdom gathered from these experiences. Visualize this knowledge and wisdom in the form of golden, white, and silver light energy gathering on the right side of the room. Continue to drain the dark cloud of misery and pain, now occupying the left side of the room, until you have extracted all the new insights from it. As you do so, you will notice the dark cloud becoming heavier and darker.
5. Focus on the right side of the room. Through your intention, begin to inhale and imagine this bright, illuminating energy flowing back into your body. Absorb this knowledge and wisdom gathered from past experiences, void of any feeling of hurt, pain, and judgement into every cell of your body. This distilled knowledge and wisdom empowers you to heal from your wounds on a physical, mental, emotional, energetic, and spiritual level.
6. Next focus your attention on the dark black cloud containing all your pain and suffering. Imagine an angelic being, righteous and holy, holding a "sword of Truth," standing by your side, ready to sever the dark energetic cord(s) still connecting your physical body to this dark cloud of despair. The sword descends and severs the cord(s) (*the doctor claps his or her hands as the sword severs the cord(s)*.
7. The cloud of darkness begins to float up, through the ceiling, through the sky into space. Far in the distance you begin to perceive God or a "higher power" emanating waves of compassion, love, and mercy descending towards the Earth. One of the waves touches the dark cloud of hurt and pain, exploding it into a brilliant light. The cloud is immediately transmuted into fluorescent crystal blue drops of light. This pure, clean energy descends from the Heavens like gentle rain. Breathe in this crystal blue healing energy, let it penetrate deeply into every pore, every tissue and cell, saturating your body completely. Feel the light cleansing and radiating throughout your being.

UNDERSTANDING WHY A PATIENT WON'T HEAL

Normally, in a clinical setting, 80% of the patients will respond favorably to Medical Qigong therapy. In some cases, complete remission of the disease is achieved. An additional 10% will improve, but stabilize in their condition, while a final 10% will not respond at all to the treatments and will progressively get worse. When the patient does not respond to the treatments, the reason can usually be related to the following conditions: Failing to Carry Out the Prescriptions, The Pathogens are Just Too Strong, and Divine Will Guides the Outcome.

- Failing to carry out the prescriptions can be caused by the patient subconsciously sabotaging the healing. Refusing to practice is usually due to deep unresolved emotional issues, e.g., feeling unworthy to heal, fear of abandonment by spouse or others, and so on.
- The pathogens are just too strong to be healed by Medical Qigong therapy alone; when an illness is long-standing, additional complementary treatments, i.e., herbs, Western medicine, or surgery may become necessary to achieve healing.
- Divine will guides the outcome in some cases. The patient's time in the physical realm may be complete, and the soul is preparing to return back to the divine.

Sometimes the healing process takes more time then expected. I explain to patients that permanent transitions can sometimes be instantaneous, but are more likely to be gradual. I explain that they are like a sea captain, standing at the helm of a great ocean liner. If the captain decides that the present course is unsafe, and destructive to the ship and its crew, he quickly orders an immediate change in the course of the ship's direction. With a flick of a switch and a hard spin of the steering wheel, the direction of the rudder is immediately changed. However, because of the perpetual inertia of the massive ship, it will take some time to observe any alteration in the course of the ship's current direction. Although the eyes of the captain can be focused on the new direction, the ship still requires enough time to effectively move in the new direction. It is the patient's spiritual intentions and focused goals that eventually direct the patient's emotional belief system, which in turn, redirects their thoughts and body's physical reactions.

Sometimes, toxic emotions and memories are released through diligent Qigong practice and treatments. With some patients, however, although major changes have already occurred in their conscious attitude, belief structure, and spirit, it may take longer for the unconscious reprogramming to take hold and stop reinforcing the old fears, and traumas. This impediment of unconscious reprogramming is due to the patient's conscious mind not being able to fully integrate the new reality and belief structure. A cancer patient may, for example, have been diligently practicing the prescriptions and the disease may now be in remission, but, the patient still visualizes the cancer as all-invasive. Rather then starting with an image of less cancer in the body, the patient continues to start the meditations with the cancer condition visualized at its worst peak. This sends a double message to the subconscious mind (heal and don't heal).

The same thing applies to observing the patient's transitions of chronically programmed patterns. New behaviors and attitudes cannot become habits overnight. It simply takes time, endurance, and constant encouragement in order to make progress. It is important for the patients to keep their eyes focused on their intended goal and not give up hope.

It is also important not to force mental imagery when feeling angry, depressed or upset. To be effective, creative visualization requires a quiescent mind and soft focus; it should not be hard work. Time allotted to creative visualization can be brief (only a couple of minutes), several times a day, for it to be effective.

FACING THE DEATH OF A PATIENT

Because human beings are created from the energies of Heaven and Earth, they are subject to the transitions of nature, death being part of that transition. Sometimes, no matter who the doctor is, or what his or her specific training has been, there is nothing that can be done to physically heal the patient. Because patients have made attachments during their stay on Earth (physical, men-

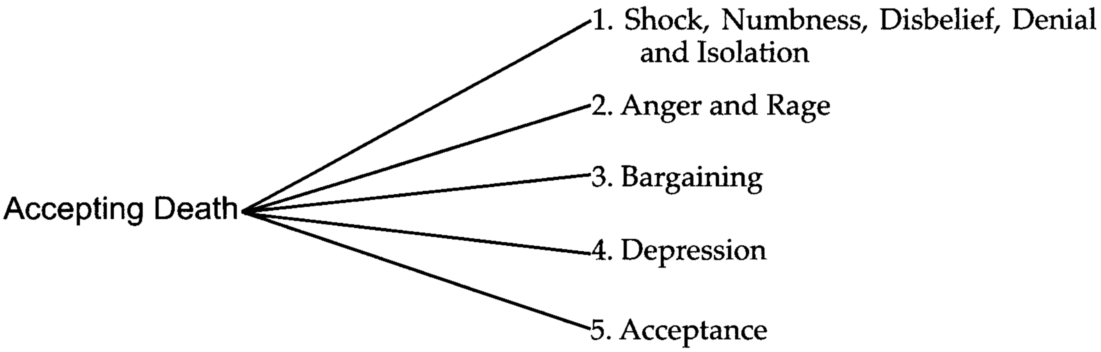

Figure 32.6. The Five Stages and Transitions of Accepting Death

tal, emotional, and spiritual), it is important for the doctor to assist them in their final transitions of releasing and letting go. The fear of death often causes patients to feel loss of control of their life, facing and conquering this fear allows them to regain a sense of control.

According to Dr. Elisabeth Kubler Ross, there are five stages of dying that facilitate closure for patients, especially in the life and death transition (denial, anger, bargaining, depression, and acceptance). These transitions also occur when facing the death and closure of any personal relationship or job, as well as the termination of one's life (Figure 32.6). Although these stages were originally believed to follow each other in an orderly fashion, further research has indicated that these stages need not, and most often do not occur one at a time. The different "stages" can occur at any time going from "denial" to "acceptance," back to "bargaining," etc., (sometimes in just a few minutes). Every individual reacts differently. Some never accept death and struggle to the very end. Medical Qigong therapy describes these transitions as follows: Shock, Numbness and Disbelief, Denial and Isolation, Anger and Rage, Bargaining, Depression, and Acceptance.

1. Shock, numbness, and disbelief occur when an individual is faced with the fact that his or her own death is near. Because of shock, the mind will generalize, rationalize, delete, or distort information to make the information acceptable to the patient's belief system. When these initial feelings of shock, numbness, or stoicism begin to disappear, most patients then experience disbelief. They can not believe or will not accept what is happening to them.

   Some patients bypass this stage all together. They feel great relief at learning of their impending death. They look forward to the cessation of pain and rejoining those close to them who have passed on.
2. Denial and isolation is the first semi-conscious choice the patient makes after hearing the news. The patients' disbelief allows them to establish their own denial system. This denial system functions as a shield to protect them, as they emerge from the shock. Denial and isolation further allows these patients time to gather additional knowledge, hoping that the original diagnosis was incorrect.
3. Anger and rage arise as the denial system breaks down. Patients experience anger, rage, envy, and resentment, as they view their terminal progress as an injustice. The patients' discomfort and anger are usually displaced and projected towards anything and everyone (usually people and God). Patterned, fear-based behaviors get stored in the body and mind. These emotions affect the patients' future perceptions, diminishing their belief system in personal survival.

4. Bargaining is the next stage in the process of adjusting to the verdict. Patients offer counterproposals (usually with God), hoping to change or alter the course of events and postpone the inevitable from happening.
5. Depression replaces hope and the patient experiences a deep sense of loss. Still resisting the outcome, patients sink into depression, sometimes lingering for quite some time before death. Depression is divided into two different categories, each one is different in nature.
   - Regretful depression is caused by guilt and shame for not having lived life to the fullest. The patients regret having given in to fears and having held on to past hurts (both real and imaginary) for so long, that many dreams were abandoned. Opportunities for healing and enjoyment were missed, and old sorrows and conflicts never resolved. Patients with this type of depression often seek to voice their regrets. Many ask to let go of the bitterness and open up to a spiritual healing, rather than a physical healing.
   - Preparatory depression is not caused by regrets, but by mourning the loss of projected future experiences. These patients are usually quiet and reflective, as they face the process of losing everything and everyone that they love.
6. Acceptance is the final stage before death. Having nowhere else to go, patients evaluate what was learned in life. They are now able to express their feelings of envy for the living and the healthy, as well as anger at those who do not have to face their fate so soon. They finally accept the inevitability of death. Depression gives way to a time of rest before the divine journey home. The patients, having faced their life's work, fears, and feelings, now have a new outlook on life and are prepared for the spiritual transition of death.

Facing death is an important transition in life and must be approached with a reverent respect for spiritual transformation. When a patient is dying it is extremely important for the doctor to accept the "higher order" or "divine will" for the patient, and not harbor a win or lose attitude. No matter what stage of transition the patients are in, they will almost always maintain some glimmer of hope, even though they may claim otherwise, to the very last moment of their life. It is the doctor's responsibility to accept and support patients to their inevitable end and not cease to treat them for reasons of personal fears, sorrows, or regrets. Listen to your patients, and allow them to release their fears. This continued support, acceptance, and encouragement for the patient allows them to die in peace and with dignity.

SUMMARY

Death is considered merely a change of residence for the patient's spirit. The ailing body is sometimes viewed as "a house with rotting walls." Once the dying body becomes unusable, it is best to abandon it, and look for another place to stay. The spirit-body is immortal and will exists forever in accordance with the will of the divine. This is why one is taught to be unattached to the physical transformations of this world and accept and embrace death as part of the natural process of existence.

Think of the transition of death as similar to that of a baby chick which is about to be hatched. Although it has lived and existed within its shell all its formative life, as the baby chick continues to develop, it has slowly begun to outgrow its place of existence. The baby chick is now becoming harshly compressed and crushed within the security of its own shell. As the little chick exhausts itself, trying to get free from what has now become its prison, it seemingly faces death. With its apparent last breaths, the little chick shatters the surrounding shell. Immediately, an intense light appears, followed by a whole new world that reveals itself in new wonderment. The baby chick is now reborn into a fresh new realm of existence, with new room to grow and the ability to experience true freedom.

Once, while in China, I was talking to Dr. Wong of the Xi Yuan Hospital, in Beijing. He made the profound statement that everyone he had ever treated had gotten healed. I have never heard such a bold statement being made by a Qigong master, or by any doctor for that matter, so I immediately began to inquire as to what his special techniques were. Dr. Wong responded, "Everyone I treat gets healed; some get healed spiritually, others emotionally, some mentally, and some physically. Not all get cured, but all do get healed. Some patients that come to me for therapy are supposed to die. It is their time of passage, and I help them accept death with dignity and honor. This is achieved by helping the patients heal their spiritual, mental and emotional wounds before they pass on. Other patients that come to me are in an active stage of energetic transition, and I help these patients to see where they've become stagnant, Excessive, or Deficient in their bodies and release their pathogenic Qi." Life is a pathway of transition and change, and so, in the field of energetic medicine, all patients are moving toward or away from healing their relationship with themselves and others.

SECTION VIII

QI EMITTING METHODS

CHAPTER 33

ENERGY EXTENSION AND QI MANIPULATION TECHNIQUES

ENERGY PROJECTION

The body's energetic constitution is a combined matrix of sound, light, heat, and electromagnetic energy. Studies on the effects of Qi projection conducted by Chinese scientists concluded that energy released by a Qigong doctor into a patient carries the properties of infrasound, electromagnetics, static electricity, infrared radiation, gamma rays, particle and wave flows, organic ion flows, and light. These properties produce the physiological changes in the patient's tissues during the Qigong treatment.

A Qigong doctor can emit energy during a treatment with or without touching the patient. The treatment with the doctor's hand moving at a distance from the patient's body is considered Qi Emission. If the doctor's hand touches the patient's body, it is considered Qi Massage, which combines both tissue manipulation and energy flow. Functionally, it is also helpful if the Qigong doctor also has a thorough grasp of the tissue manipulation skills of Chinese massage therapy (Jie Gu for bone setting, Tui Na for traumatology, and An Mo for Internal diseases and visceral manipulation), as well as acupressure and TCM theory, to ensure a highly proficient quality of treatment and to avoid misdiagnosis and erroneous treatments.

Qi is projected outside the body and travels in ripples and vibrational pulses similar to the ripples of an ocean wave. The frequency of this energy wave can be modulated into dense or dispersed resonating patterns depending on the Qigong doctor's needs.

Shen is projected outside the body as light and travels in light beam particles. This particle light extension can change in color and intensity depending on the doctor's intention. When Qi and Shen are combined, the light and wave frequencies facilitate maximum energy projection.

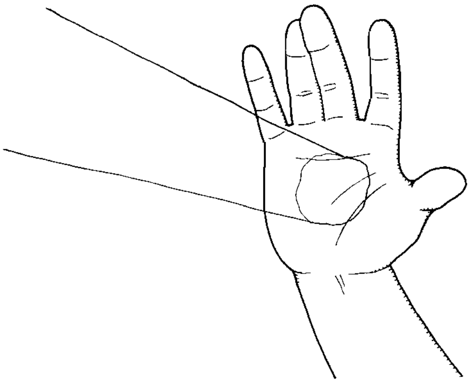

Figure 33.1. Palm Qi Extension

TYPES OF ENERGY EXTENSION

There are nine types of energy extension and projection skills that are divided into three categories based on the source of the projected energy. The first three are projected from the extremities: palms, fingers, and soles of feet. The next category of Qi projection originates from the three Dantians: Upper, Middle, and Lower Dantian. The third category includes Qi emission from the eyes, throat (sound resonation), and thought projection.

ENERGETIC PROJECTION FROM THE EXTREMITIES

The first three energetic projections are released from the extremities (palms, fingers, and the soles of feet) in accordance with the doctor's imagination and intention.

- Palm extension is the ability to release emitted energy through the center of the doctor's palm (Pc-8). It is considered the primary hand posture used by all Medical Qigong doctors for emitting Qi (Figure 33.1).
- Finger extension is the ability to release emitted energy through the extended fingers (usually through the index and middle fingers).

The finger extension techniques allow the Qigong doctors the ability to finely hone their energetic projections into a smaller, condensed energetic beam (Figure 33.2).

- Soles of feet extension is the ability to release energy through the bottom of the feet (Kd-1); this method is usually used to surround and envelop the treatment room as well as the patient (Figure 33.3).

ENERGETIC PROJECTION FROM THE DANTIANS

The next three energetic projections are released from the Upper, Middle, and Lower Dantians. These energetic projections are based upon the doctor's kinesthetic, empathic and intuitive abilities.

- Upper Dantian (Yin Tang) extension from the Qigong doctor's Third Eye area can be used in conjunction with the eyes for intensifying the doctor's Shen projection towards the patient.
- Middle Dantian (CV-17) extension from the Qigong doctor's Heart area can be used in conjunction with projection from the doctor's hands or throat, for vibrational resonation. This technique intensifies the emotional connection between the doctor and the patient.
- Lower Dantian (CV-6) extension from the Qigong doctor's abdominal area can be used in conjunction with Qi emission from the feet for expansion or extension of Wei Qi. This method intensifies the projected External field of Qi around the patient (Figure 33.4).

ENERGETIC THOUGHT PROJECTION OF QI AND SHEN

The last three energetic projections are released from the eyes, throat and mouth (sound resonation) and thought projection. These techniques are a form of long-distance thought projections, based upon the doctor's ability to project Qi and Shen.

- Projection from the eyes releases energy through the eyes and is generally used to increase the doctor's intention and focus of Qi (Figure 33.5).
- Sound resonation projects the voice to tonify or disperse the patient's energy through visceral resonations (Figure 33.6).

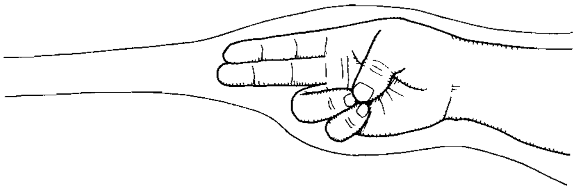

Figure 33.2. Finger Qi Extension

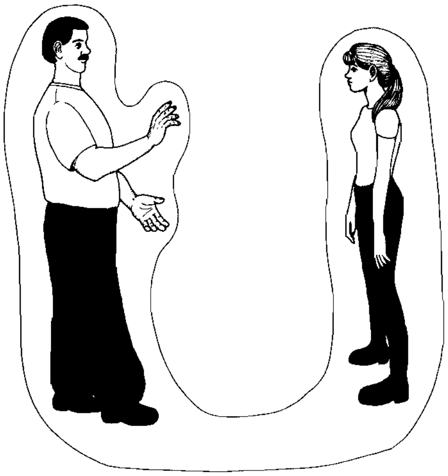

Figure 33.3. Extending Energy from the Soles of the Feet

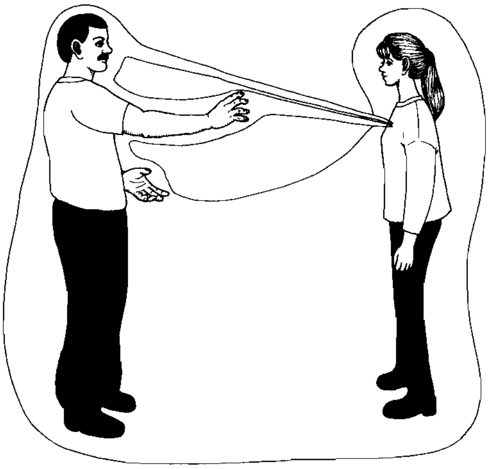

Figure 33.4. Qi Extension. When treating a patient, all Three Dantians extend their energy into the patient.

- Thought projection is another way to release Qi. It is used for long distance energy therapy and is a form of transmuting Qi into Shen, which is transmuted through the Wuji into the patient's Yuan Shen (Figure 33.7).

CONTROLLING ENERGETIC PROJECTIONS

The problem with energy extension lies not in releasing life-force energy but in controlling the power and flow of its projection. If the doctor is distracted or unstable in his or her projection of energy, the doctor may induce more problems in the patient's body. Qi projection can be stabilized in two specific ways:

- First, by visualizing and projecting specific colors (blue, red, violet, etc.) with the energy extension. This causes the Qi to take on a dense and fuller projection, which facilitates the rooting and stabilizing of the energy into the patient's tissues.
- Second, by using intention to mentally focus the size and shape of the projected energy (like a laser), to intensify the Qi projection.

When treating a patient, the diseased tissue must be sufficiently energized. Inadequate energizing results in only slight improvement or a slower rate of healing; over-energizing may cause Qi congestion.

The degree of curative effect depends on the competence and versatility of the doctor, as well as the responsiveness of the patient. If the patient is not responsive, repeated stimulation and regulation of Qi is needed to raise their sensitivity.

Sometimes the patient's emotion and spirit are outside and disconnected from their body and mind. Although this is a perfect adaptation for a military environment, it is not conducive for an energetic healing environment. The patient must undergo a series of Shen Gong training sessions to retrieve their emotional spirit. This training is commonly referred to as "soul retrieval" because the patient's soul is in a state of shock (or suspended animation) due to severe emotional trauma, and must be returned to its Yuan Jing within the patient's tissues. (See Chapter 19).

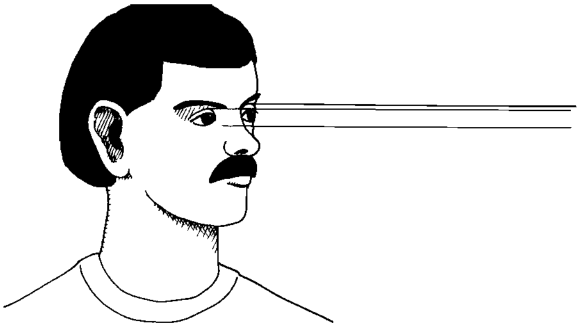

Figure 33.5. Energy Projection from the Eyes

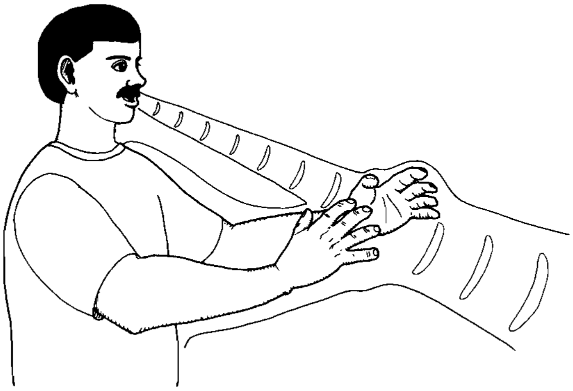

Figure 33.6. Sound Resonation Projection

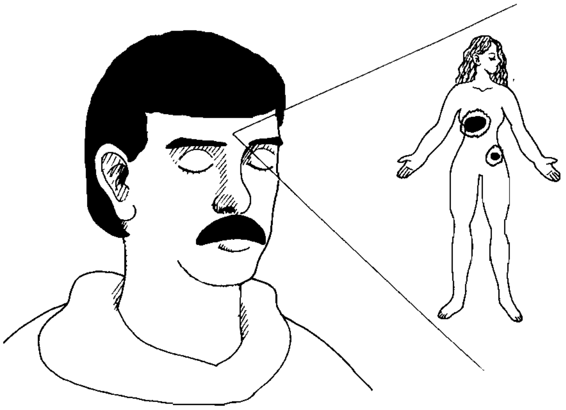

Figure 33.7. Thought Projection

In addition to Shen Gong training, the Qigong doctor must make the patient's Qigong meditations and prescriptions more active so that a satisfactory effect can be achieved. However, a desirable effect can only be attained when the patient is cooperative with the doctor and is consistent with the assigned homework.

HAND POSTURES FOR QI EMISSION

There are many hand postures used in emitting and drawing in energy. Each variation depends on the skill and particular background of the doctor. Generally the doctor uses the entire hand to extend energy, constantly changing the intensity of the Qi being extended. By expanding and contracting the hand and fingers, the "beam of Qi" being emitted into the patient's body will expand and contract. Presented here are six hand postures most preferred by myself and many other Medical Qigong doctors in clinics throughout China.

THE EXTENDED FAN PALM

This first hand posture is the most popular in terms of Qi emission. It releases the widest field of Qi emission. In the Extended Fan Palm hand posture, all five fingers stretch naturally, separate, and open like a Chinese fan (Figure 33.8). There should be no tension, and the focus of the energy should be in the center of the palm (Pc-8). This Extended Fan Palm is the primary hand posture for extending energy because it can be easily used for emitting either Hot or Cold energy. It can also be used in conjunction with other techniques as a reinforcing method to strengthen the flow of energy extension.

Note: Before beginning these exercises, it is important not to eat a heavy meal (at least one and a half hours before practicing).

PALM BREATHING ENERGY EXERCISE—STAGE 1

1. From a Wuji posture, bring both hands up in front of the body, with the palms facing the Lower Dantian. Breathe naturally while concentrating on the Lower Dantian. Next, place

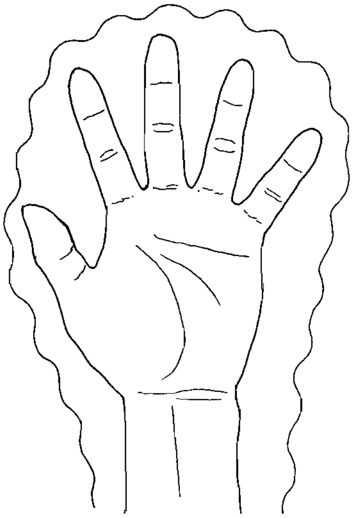

Figure 33.8. The Extended Fan. This palm is the primary hand posture used for emitting Qi.

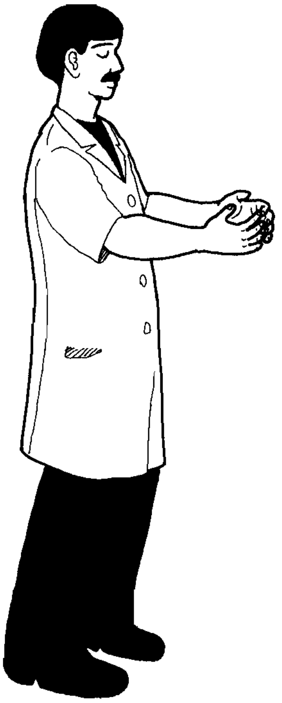

Figure 33.9. The Image of Holding a Ball of Energy

the left palm facing the right palm, as if holding an imaginary ball. Allow the energy between both hands to build and increase its energetic field (Figure 33.9).

2. Press the Mingmen backwards and feel the energy absorbing through the arms into the Lower Dantian. Next, release the pressure from the Mingmen and feel the Qi return back into the center of the palms.
3. Shift the attention to the center of the right palm, and begin to extend energy into the center of the left palm. Next, shift the attention from the center of the right palm, and begin extending energy into the center of the left palm.
4. As a feeling of pressure begins to build between the hands, perform the "Pushing" (energy extension) and "Pulling" (energy retraction) manipulation technique. This technique is practiced by pushing the Qi of the right palm into the center of the left palm. Hold it there, feeling the pressure mount, then withdraw the Qi back into the center of the right palm. Repeat for several minutes, then move to step five.
5. After several minutes switch the mind's focus to the opposite hand and perform the exercise again. Do this exercise twice a day for 15-30 minutes each practice. Be careful not to let the hands touch when running the energy back and forth between the palms. After 100 days go to Stage #2.

When training to increase their own energetic field, the Qigong doctors must strive to overcome the Pituitary Pain Syndrome. This is a state of transition during which the body is overcome by pain to the extent that the mind gives up and the doctor quits practicing (e.g., feeling fatigue when practicing a Holding the Ball posture). It is important that as the doctor become stronger, the training posture be maintained longer, so that the pituitary gland is challenged to activate the hypothalamus. The hypothalamus causes the body to produce endorphins such as ACTH, which is a natural form of cortisone that helps to escalate the doctor's healing potential. This physical reaction is similar to a runner who must overcome physical cramping when racing to obtain a second wind.

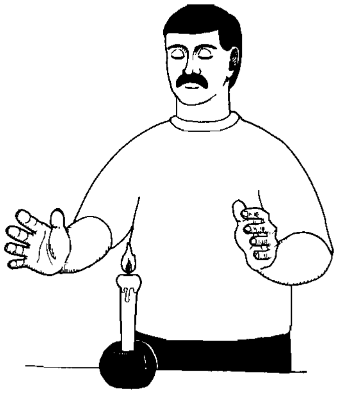

Figure 33.10. Posture for Palm Breathing Exercise

PALM BREATHING ENERGY EXERCISE—STAGE 2

1. Light a candle and place it on a table. From a Wuji posture place the center of the right palm several inches to the right of the tip of the flame (Figure 33.10). The left palm should be placed on the opposite side of the candle, facing the right palm. The Laogong points and the tip of the candle should be in a straight line with each other.
2. Breathe naturally while concentrating on the Lower Dantian. Shift the attention onto the center of the right palm, and begin extending energy into the center of the left palm.
3. Once a feeling of pressure begins to build between the palms, begin the Pushing and Pulling energy manipulation technique. Push the Qi of the right palm into the left palm. Hold it there for several minutes, feeling the pressure mount, then draw the Qi back into the center of the right palm. Practice this exercise for several minutes.
4. Next place the left palm on the Lower Dantian while continuing to extend and retract energy from the right palm. After several minutes, switch hands. Notice that as the Qi is being extended out of the palm that the flame bends away, and as the Qi is retracted back

into the palm the flame bends towards the center of the palm. Practice this exercise twice a day for 15-30 minutes each time.

Note that the Yin Tang area (the Upper Dantian's projection point) can grasp the flame and hold it stationary. Since this technique (grasping and rooting) is used for rooting and securing the patient's Qi, its application can be counter productive when trying to move the flame tip. To avoid this condition, the doctor should relax the mind and use the imagination when projecting Qi.

THE SWORD FINGERS

In this second hand posture, the expansive field of Qi being released is reduced and condensed into a more solid beam of light. In the Sword Fingers hand posture, the index and middle fingers of the right hand are extended and joined together, the ring and little fingers curl into the center of the palm (Figure 33.11). The thumb should rest on the nails of the ring and little fingers forming a circle where Qi is gathered.

The index finger indicates Wood or Liver's Qi, and the middle finger indicates Fire or the Heart's Qi. Together, both fingers activate the Qi of Fire and Wood, which is a catalyst for the Yang energy needed for guiding Yang Heat with the intention (Yi). When extending energy, the Qi is gathered first in the palm and the circle between the thumb and last two fingers, then it is released through the tips of the index and middle fingers.

SWORD FINGER QI EXTENSION—STAGE 1

1. From a Wuji posture, extend both hands to form an imaginary circle of energy in front of the Lower Dantian. As soon as the Dantian is activated and full, bring the ball of energy in front of the body, level with the Middle Dantian (Heart / solar plexus level). Direct the Qi from the Lower Dantian into both hands (Figure 33.12).
2. Point the middle and index fingers of the right hand towards the left palm (Pc-8) and feel the energy press into the tissues (Figure 33.13). After a while, neutralize the energy by turning the palms towards each other, then switch and feel the energy of the left sword fingers

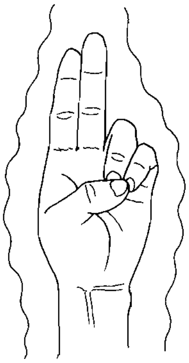

Figure 33.11. The Sword Fingers Technique

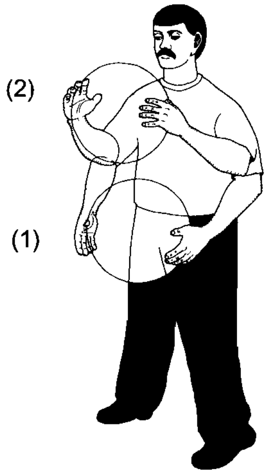

Figure 33.12. Imagine a ball of energy in front of the Lower Dantian, then raise it up to the level of the Middle Dantian.

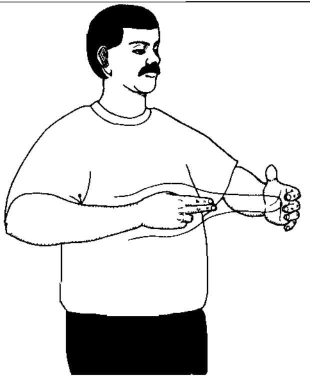

Figure 33.13. Point the Sword Fingers of the right hand towards the center of the left hand.

(index and middle finger) press into the center of the right palm. Practice this exercise twice a day for 15-30 minutes each practice. After 100 days go to the next stage.

SWORD FINGER QI EXTENSION—STAGE 2

1. Light a stick of incense and place it in a standing upright position on a table. From a Wuji posture, place the center of the left palm (Pc-8) several inches to the left of the incense. The right hand sword fingers point through the tip of the incense, extending Qi into the center of the left palm. All three points - left center palm (Pc-8), tip of the incense, and the right sword fingers - should be in a straight line (Figure 33.14).
2. Breathe naturally while concentrating on the Lower Dantian. Shift your attention onto the tip of the incense and focus your concentration there. Continue to extend Qi until you have a strong feeling of Qi in the center of your left palm. After a while switch to the right palm. Practice this exercise twice a day for 15-30 minutes each time.

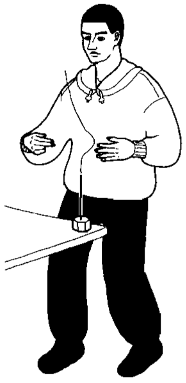
Figure 33.14. The Sword Fingers point at the tip of the lit incense.

THE ONE FINGER SKILL

In this third hand posture, the field of Qi being released is further condensed into an more solid beam of light. In the One Finger Skill hand posture, the hand forms a soft fist with one appendage (finger or thumb) extended (Figure 33.15). The energy is projected in a straight line (like a laser beam) and can be used to Heat the patient's tissues, channels, or points, as well as to lead, extend, and tonify the energetic field (Figure 33.16).

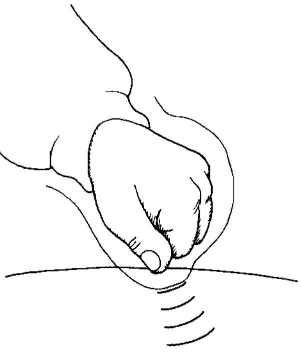
Figure 33.15. The use of the finger or thumb to emit Qi is an example of the One Finger Skill technique.

One Finger Skill can be used to extend long range energy into the patient's channels, or for tissue and channel point regulation. The former technique is regarded as an internal application, and the later technique is regarded as an external application.

- The external application requires that the doctor make physical contact with the patient's tissues, extending the mind deep into the specific areas, channels, or points to be treated. The Qigong doctor physically rotates and shakes a specific finger (or thumb) from side

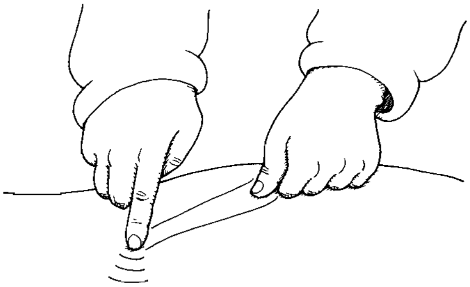
Figure 33.16. Energy is extended through the One Finger Skill technique.

to side while emitting Qi into the patient's tissues or energetic fields (Figure 33.17).

- The internal application requires that the Qigong doctor extend his or her arm, pointing a specific finger at a particular area on the patient's body. The doctor then focuses on that specific channel point, leading and directing the emitted Qi into or out of the patient's body.

ONE FINGER SKILL ENERGY EXTENSION EXERCISE

1. From a Wuji posture, focus the concentration on the Lower Dantian area and bring both arms up to chest level and form an energy ball. Once the Qi in the Lower Dantian becomes activated, direct its energy flow to the tip of the right index finger. Once a feeling of heat and expansion of energy is felt within the right index finger, point the energy beam into the heart of the left palm (Figure 33.18).
2. Once the energy is felt pressing into the left palm, begin to circle-rotate the energy beam in a clockwise direction for several minutes, then in a counterclockwise direction for several minutes.
3. Next, extend and focus the Qi from the Dantian into the left index finger, and repeat the exercise.

SOARING DRAGON TECHNIQUE

The Soaring Dragon technique is a variation of the One Finger Skill energy extension technique. If the pathogens are strong and too much Heat is being emitted from the patient's tissues (as in certain types of cancer), the Qigong doctor may decide to use this particular variation to avoid absorbing the dangerous toxins being released from the patient's body. The Qi is emitted out the doctor's middle finger into the patient's body while the other four fingers disperse the pathogenic Heat into the ground (Figure 33.19).

Before initiating the Soaring Dragon technique, it is important for the Qigong doctor to use his or her mind's intention to seal the wrist. This enables the doctor to pull the patient's Qi into the palm and to disperse it into the Earth via the other four fingers, while avoiding the absorption of the patient's pathogenic Qi.

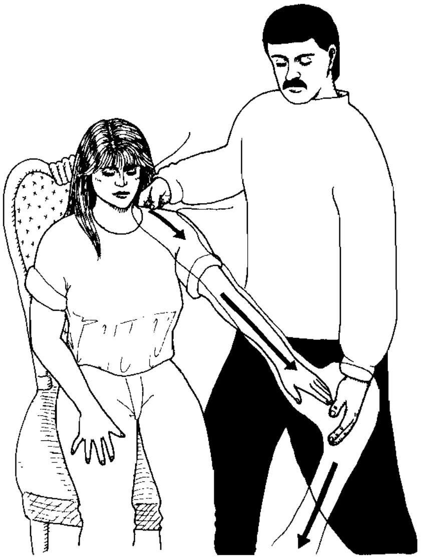

Figure 33.17. The doctor makes contact with the patient's tissues, extending his or her intention along the patient's channels, leading the Toxic Qi out her body.

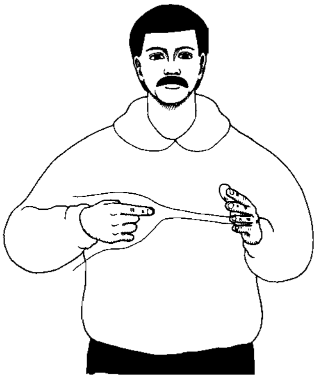

Figure 33.18. One Finger Skill Energy Extension Exercise

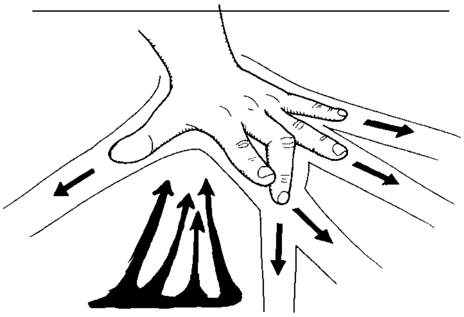

Figure 33.19. Soaring Dragon Hand Posture Technique

I have found this technique very effective in my personal practice. As the pathogenic Heat from the patient's tissues releases upward, the dispersing energy that flows from the four fingers cause the patient's toxic Heat to dissipate and be carried away from the doctor.

BELLOWS PALM TECHNIQUE

Another variation of the One Finger Skill is the Bellows Palms technique, which uses the Lao Gong Pc-8 point at the center of the palm as a pump, to expel Qi out the middle finger. Originally, while treating patients, the palm is Hot (in a Yang state). However, when pulling Qi from the center of the palm and directing it out the middle finger, the polarity of the palm changes from Yang to Yin, causing the center of the palm and the middle finger to turn white (Figure 33.20).

Because this technique is used to pull out Heat from the patient's body, before initiating the Bellows Palm it is important for the Qigong doctor to use his or her mind's intention to seal the wrist. The doctor then pulls the patient's Qi into the palm and disperses it into the Earth via his or her middle finger to avoid absorbing any of the patient's pathogenic Qi (Figure 33.21).

THE BACK BRIDGE BAR TECHNIQUE

The Bellows Palm technique is also used with the "back bridge bar" (where the trapezius muscles join both arms across the doctor's shoulders) to aspirate toxic energy from the patient's body. In this particular technique, the doctor's left palm is used as a vacuum to aspirate a specific area inside the patient's body. The toxic energy travels through the doctor's left arm, across the shoulders and back (never the front), and out through the right palm into the Earth. The right palm applies the "bellows" technique used to create the energetic suction applied by the left palm, for removing Toxic Qi. This technique is contraindicated when treating cancer patients.

THE INVISIBLE NEEDLE PALM

In this fourth hand posture, the field of Qi being released is further condensed into an intense beam of light, similar to that of a laser beam (Figure 33.22). In the Invisible Needle Palm hand posture,

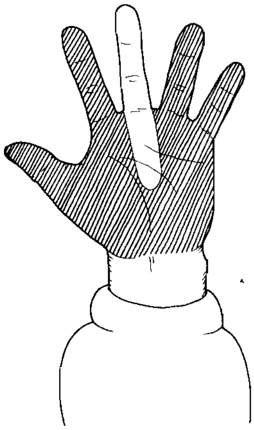

Figure 33.20. When using the Bellows Palm technique, the center of the palm and middle finger remain white in color, while the rest of the palm remains red.

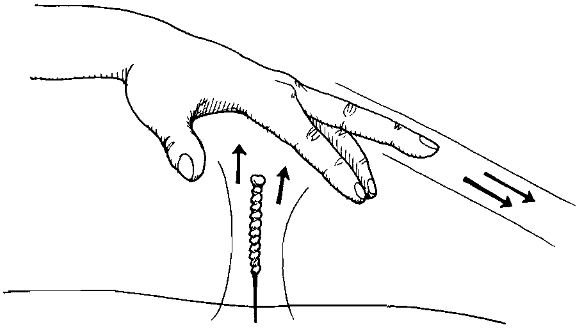

Figure 33.21. Absorb the patient's Toxic Qi from the middle of the palm and disperse the pathogens out the middle finger into the Earth.

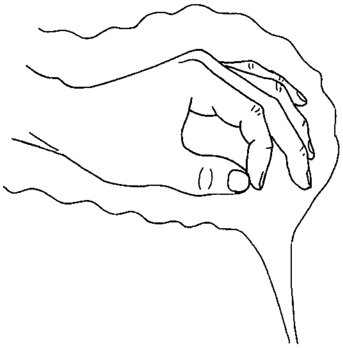

Figure 33.22. The Invisible Needle Palm

three fingers on the right hand are curled into the center of the palm with the thumb resting against the tip of the index finger (see Chapter 35 for Invisible Needle Theory and Application).

THE DRAGON'S MOUTH PALM

This hand posture is sometimes also called the Duck's Bill Palm, and is used to pull energy (like pulling a silk thread). The thumb and all four fingers perform the action of "pulling the golden thread" in order to lead or pull energy along and through the body's channels and points. The intensity of the Qi and movement between the thumb and fingers is determined by the relaxed movements of the wrist and finger joints. This hand technique is often used to purge energy away from specific points when treating areas of Excess (Figure 33.23).

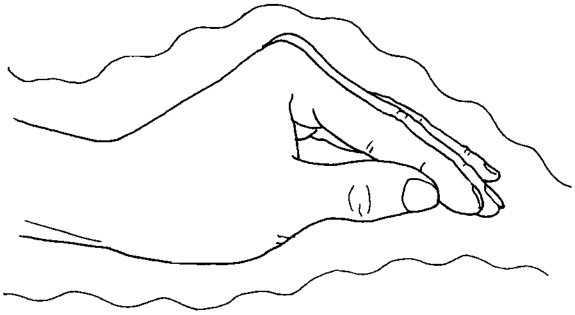

Figure 33.23. The Dragon's Mouth Palm

THE KNEADING TIGER PALM

In this hand posture, all five fingers naturally curl as if embracing a ball. The wrist action is very fluid and supple, allowing the fingers to expand open like the Extended Fan Palm, and then gently close as if kneading a cotton ball. The entire action should be similar to a jellyfish moving in the ocean. This is an excellent technique for pulling, shaking, and rotating energy (Figure 33.24).

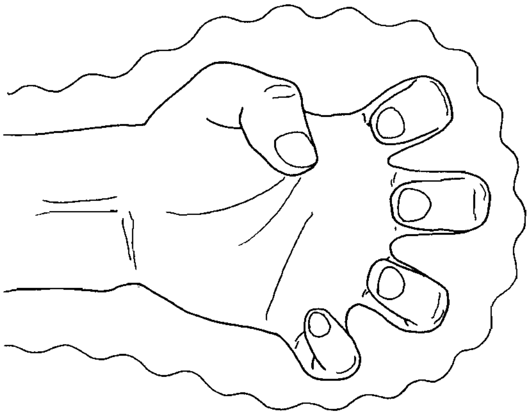

Figure 33.24. The Kneading Tiger Palm

KNEADING TIGER TECHNIQUE

This hand posture is used for dispersing stagnations and dissolving energy blocks. Once the doctor has selected the proper hand posture and the area to be treated, the doctor places his or her hands and fingers above the patient's body, performing a circular kneading action (slightly opening and closing the palm) while simultaneously extending energy into the patient's body. The spiraling energetic frequency should be synchronized with the doctor's hand kneading and the patient's pulse and respiration. Both the pressure and the strength of energy extended into the patient are determined according to the severity of the illness. The Tiger Kneading technique is used for extending energy into energetic points of the chest, abdomen, and extremities to purge Qi stagnation and improve energetic circulation.

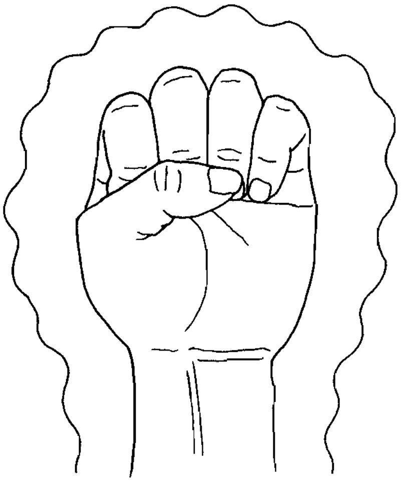

Figure 33.25. The Five Thunder Finger Palm

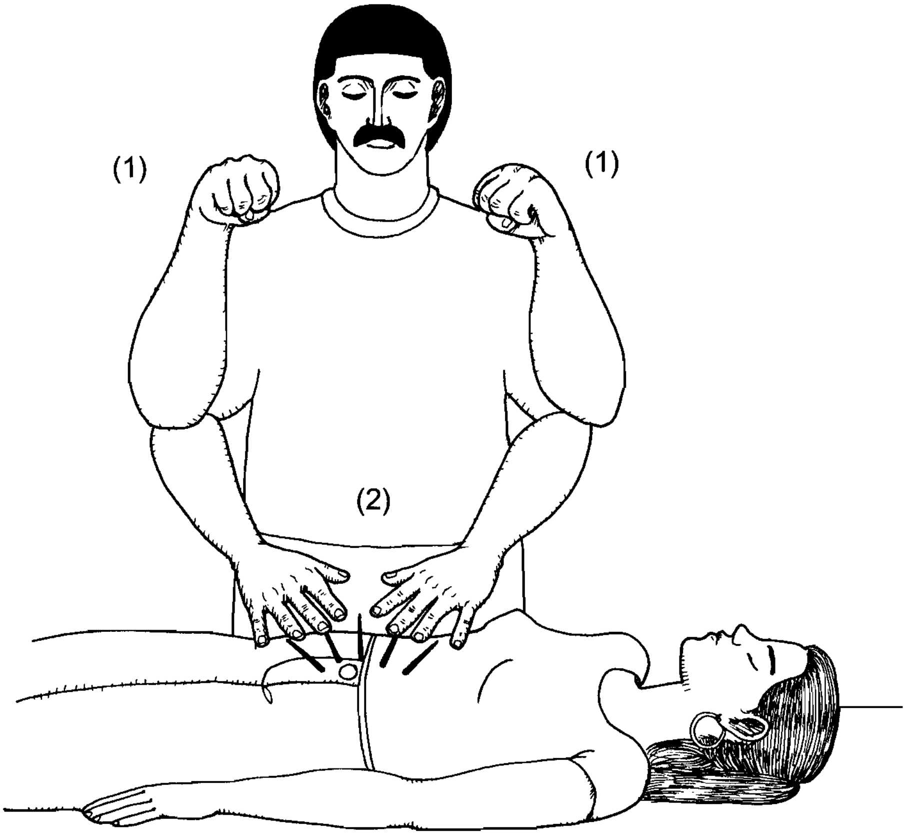

Figure 33.26. The Five Thunder Fingers. This technique is used primarly to destroy cysts, tumors and pathogens.

THE FIVE THUNDER FINGERS

This hand posture is one of the most popular techniques used for dispersing stagnations, dissolving energy blocks and treating tumors and cysts. In the Five Thunder Fingers hand posture, all four fingers curl into the center of the palm with the thumb resting against the nail of each finger. Keep the center of the palm hollow. Qi is gathered and collected there upon inhalation (Figure 33.25).

When emitting or extending Qi, the doctor exhales while suddenly stretching the palm open like "a tiger exposing its claws." After extending the energy, the doctor inhales and resumes the Five Thunder Fingers posture while gathering Qi back into the center of the palm (Figure 33.26).

Note: Because it builds up the doctor's guiding and gathering skill, this is an important technique and should be practiced consistently.

QI BLAST EXERCISE

From a Wuji posture, raise both palms upward to shoulder level. Both hands form the Five Thunder Fingers posture. Concentrate on the Lower Dantian when inhaling. Imagine drawing the Qi up through the chest, gathering the energy into both palms as the arms raise upward. Upon exhalation, sink the body's weight, and suddenly stretch out the fingers (like a claw) at chest level, striking the air. This action should look like each hand is heaving a rubber ball. Perform this exercise once or twice a day, using 24 or 48 breaths each practice time.

CHAPTER 34
QI EXTENSION, GUIDANCE, AND REGULATION TECHNIQUES

Qi extension is most effective in activating and balancing the patient's flow of Qi, since both the doctor and patient are enveloped in a strong field of Wei Qi. In Qigong massage the patient's channels are more easily dredged and their Yin and Yang energy are often adjusted much more quickly because of the stimulation produced by the doctor's hand on the patient's channels and points.

Therefore, both modalities - Qigong massage and Qi extension - are usually combined to provide the most effective and efficient means of energetic therapy.

QI EXTENSION TECHNIQUES

The clinical guiding and directing of Qi is performed by extending energy outside of the doctor's body. The energy follows the dictates of the doctor's mind as to the pattern, amount, and direction in which to flow. Qi is emitted in three basic formats: linear, circular/spiral, and stationary flows of energy. The combination of these three forms insures that the patient is receiving the maximum benefit from the treatments, and may be combined with Hot and Cold energy extension, the Five Elemental energy extension, as well as with Qigong massage. These following exercises lay a foundation for guiding and directing the patient's Qi within the channels, points, and Wei Qi fields.

GUIDING THE LINEAR FLOW OF ENERGY

This method refers to a straight line of energy being emitted either between both hands or between one of the doctor's hands (usually the right) and a certain point or area on the patient's body. This is the primary Qi guiding method used in Medical Qigong therapy to sweep over the patient's body in a straight line in the direction of the outgoing flow of energy.

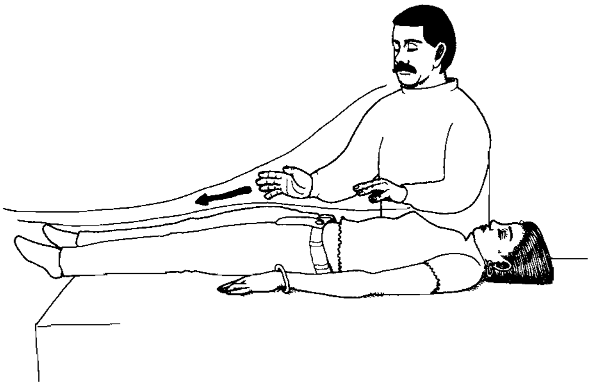

Figure 34.1. An example of a Qigong doctor using the Pushing Energy technique in a Straight Linear energy flow

Because patients vary in their sensitivity to transverse energy wave flow, the doctor should be aware of the influence of the different wave patterns affecting the patient during treatment. Linear energy is relatively mild and often gives the patient a sensation of constriction, of heat or cold, or of heaviness, pressing, tugging, tingling, or pulling. This is the basic means of inducing the channel Qi to purge Excess and to supplement any Deficiencies. Energy is directed within the body through either pushing, pulling, or leading the patient's Qi.

PUSHING ENERGY

Pushing energy is initiated by first extending energy over the surface area of the patient's body (Figure 34.1). Once the doctor has selected the appropriate hand posture above the patient (between 4 inches to 3.5 feet from the channel point or area), he or she will decide whether to push the energy in a linear or circular fashion.

Using intention, the Qigong doctor gently extends energy to the desired area and begins treatment. Pushing energy is an essential energetic

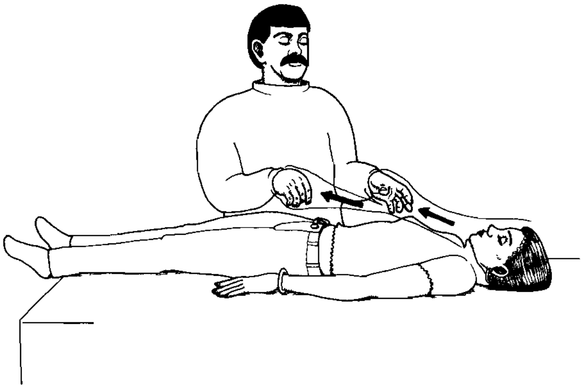

Figure 34.2. An example of a Qigong doctor using the Pulling Energy technique in a Straight Linear energy flow

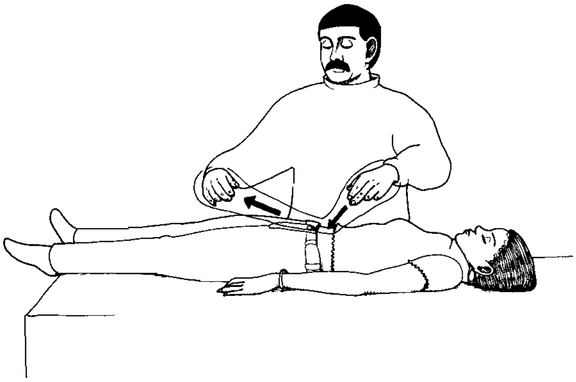

Figure 34.3. An example of a Qigong doctor using the Leading Energy technique (with the left hand) while Pulling Toxic Energy (with the right hand) from the patient's body

manipulation used to open up the points to activate or replenish the body's Qi. The patient feels pressure from the doctor's Qi, and may experience heat, tingling, condensation, expansion, swelling, or heaviness.

Sometimes when the doctor uses the pushing energy technique, the patient may spontaneously lift the stimulated body area towards the hand of the doctor. This spontaneous reaction enables the doctor to implement a combination of manipulation therapies such as push-pull, push-pull-rotate (spiral), push-pull-shake, etc.

PULLING ENERGY

Pulling energy is initiated first by extending energy over the surface area of the patient's body. After selecting the appropriate hand position, the doctor chooses to either pull the energy in a straight line or spiral it along the channels.

Using intention, the Qigong doctor usually begins to pull the patient's toxic energy along the channels while extending clean energy to the affected area (Figure 34.2). Pulling energy is an essential energetic manipulation used for dredging and expelling pathogenic factors from the patient's body, and can also be used to open up channel points to activate or replenish the patient's Qi.

During treatment the doctor actually feels the pathogenic factors being pulled out of the patient's body. This causes the patient to feel heavy, tingling, dryness, heat, cold, or chills. When combining multiple manipulation methods, the pushing and pulling actions cause the energy inside and outside of the patient's body to flow into each other, opening up the patient's channel points.

The combination of pulling and shaking techniques promotes the flow of the patient's Qi, and induces spontaneous reactions of the tissue's energetic fields.

LEADING ENERGY

Leading energy is initiated first by extending energy over the surface area of the patient's body. After selecting the appropriate hand posture, the doctor extends energy towards the affected area and begins to lead the patient's channel Qi up, down, left, right, with, or against the flow of the body's channels (Figure 34.3).

The Qigong doctor bases the decision of whether or not to lead the patient's energy (and which technique should be used) upon the patient's condition and severity of the illness. This manipulation technique is used for guiding the circulation of channel Qi, regulating any Excess or Deficiency of Yin and Yang energy and to transfer Qi back to its origin. Once the channel Qi and point Qi have been stimulated, it is important that the doctor immediately use the "leading energy" technique to guide the patient's energy in as smooth a transition as possible.

CIRCULATING THE QI IN A CIRCLING PATTERN

Emitting a flow of energy in a circling pattern is used to move and circulate stagnant Qi. The flow of energy can be moved in a continuous pattern with or against the channel flow, and is used to regulate the patient's channel and organ Qi (Figure 34.4).

This method refers to techniques of rotating Qi in a circle pattern inside the patient's body. When using circling energy, the doctor extends Qi into the patient's body and implements the circle technique with the right (Yang) hand, while rooting and discharging the patient's toxic Qi with the left (Yin) hand, which points to the Earth. There are many variations of this technique and the doctor must become familiar with each modality of healing and its potential for treating the patient.

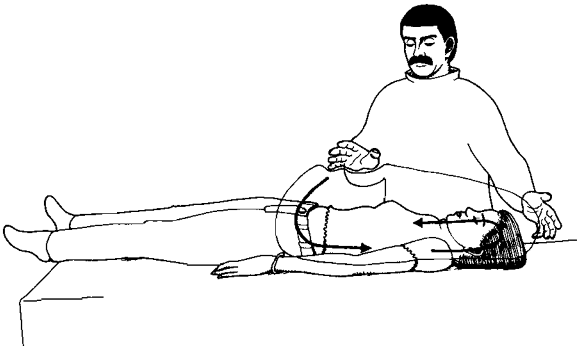

Figure 34.4. An example of a Qigong doctor using the Circling Energy technique to regulate the Heart Fire and Kidney Water, through the Microcosmic Orbit

ENERGETIC CIRCULAR PATTERNS

The following patterns are specific templates used for circulating the body's internal and external Qi. Each pattern has an advantage over the others when used for the regulation of Yin and Yang disharmony. The depth of energetic penetration will vary in accordance with the doctor's intention. Each pattern's movement will utilize the ring of energy that surrounds the center of each Dantian (Figure 34.5).

1. **The Eternity Pattern** is used for balancing the energy between the Lower Dantian and the Middle Dantian. The doctor moves the energy in a figure "8" pattern, Guiding and Leading the Qi from the Lower Dantian

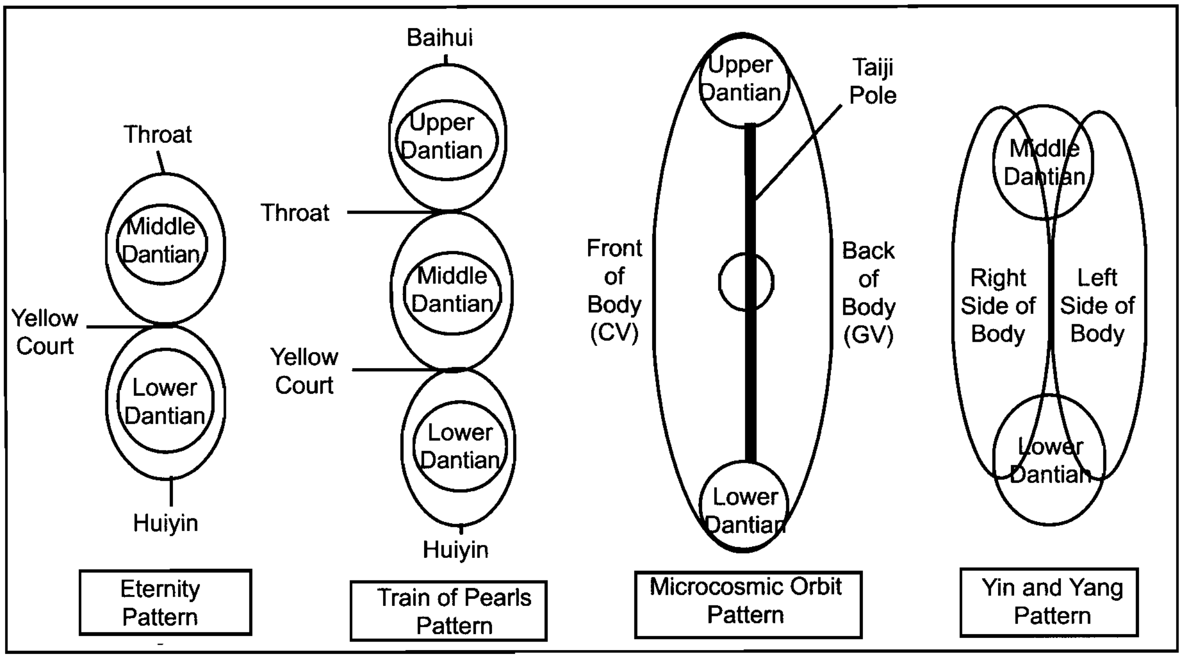
Throat
Middle Dantian
Yellow Court
Lower Dantian
Huiyin
Eternity Pattern
Baihui
Upper Dantian
Throat
Middle Dantian
Yellow Court
Lower Dantian
Huiyin
Train of Pearls Pattern
Upper Dantian
Taiji Pole
Front of Body (CV)
Back of Body (GV)
Lower Dantian
Microcosmic Orbit Pattern
Middle Dantian
Right Side of Body
Left Side of Body
Lower Dantian
Yin and Yang Pattern

Figure 34.5. A Qigong Doctor Using Circling Energy

(through the Yellow Court) to the Middle Dantian, and back again. The doctor continues this pattern until the patient's energy is regulated. If the patients are suffering from an Excess Yang condition in the upper part of the body, and a Deficient Yin condition in the lower part of the body, the doctor must regulate this condition by using the Eternity Pattern of Qi regulation.

2. **The Train of Pearls Pattern** is used for balancing the energy between all three Dantians. The doctor moves the energy from the Lower Dantian (through the Yellow Court) to the Middle Dantian (through the throat), and into the Upper Dantian, then back down again. The Qi is swept up and down the entire torso, combining and regulating the energies of the patient's Wei Qi, Channel Qi and Organ Qi. The doctor continues this pattern until the patient's energy feels regulated. If the patients are energetically disconnected from the upper and lower parts of their body due to emotional congestion and energetic stagnation in the throat and Yellow Court areas, the doctor reconnects these areas by using the Train of Pearls Pattern of Qi regulation.

3. **The Microcosmic Orbit Pattern** is used for balancing the energy between the Three Dantians, Taiji Pole, the Heart's Fire and the Kidneys' Water. The doctor moves the energy from the Lower Dantian (Mingmen area) to the Middle Dantian (Shendao Area), then back down again, completing the Fire Cycle of the Microcosmic Orbit. The doctor continues this pattern until the patient's energy is regulated. The Microcosmic Orbit Pattern is used for the complete regulation of the body's Yin and Yang energy, naturally purging, tonifying, and regulating any Excess or Deficient condition along the Governing and Conception Vessels.

4. **The Yin and Yang Pattern** is used for balancing the energy between the right and the left sides of the body. The doctor moves the energy in a horizontal figure "8" pattern, Guiding and Leading the Qi through the left and right quadrants of the patient's body. In cases of paralysis after a stroke, for example, the doctor regulates the Qi of the patient's left and right quadrants and energetically balance these areas by using the Yin and Yang Pattern of Qi regulation.

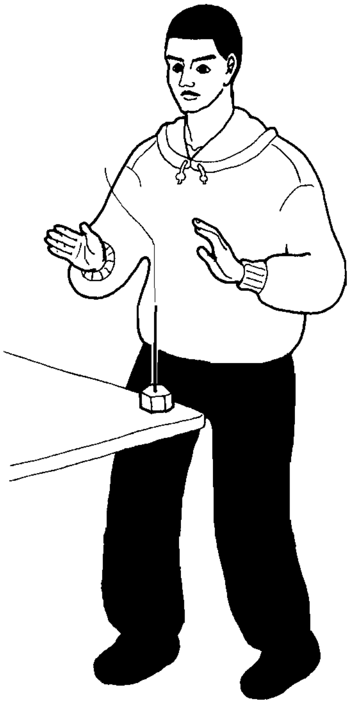

Figure 34.6. An example of a Qigong doctor moving the incense smoke, using the image of an energy ball

The Qigong doctor selects the appropriate energetic pattern according to the patient's condition.

ENERGY BALL ROTATION FOR DEVELOPING CIRCLE ENERGY FLOW

- Light a stick of incense and stand it on a table. From a Wuji posture, place the center of both palms on opposite sides of the tip of the incense. All three points, the tip of the incense, and the heart of each palm (Pc-8), should form a triangle (a tree or flower can be substituted for the incense).
- Breathe naturally while focusing on the Lower Dantian. Slowly shift the attention to lead the energy into the center of each palm to form a ball of energy (basketball size). Concentrate and

imagine the ball condensing to form a powerfully compacted ball of energy (about the size of a marble). Next, imagine a circle or rounded triangle between both palms and the tip of incense. Exhale lightly while extending the energy ball toward the tip of the incense. Imagine the ball of energy moving counterclockwise from the right palm to the incense and then to the left palm, moving in a circle. The right hand pushes while the left hand pulls. The incense smoke should move in accordance with the Qi movement. After a while reverse the flow of the circle. Practice this exercise twice a day for 15-30 minutes each time (Figure 34.6).

SPIRALING ENERGY FLOW

A spiraling flow of energy moves the energy in a spiral pattern (clockwise or counterclockwise) to drill the energy deep inside the patient's body (Figure 34.7). This "drilling" action, caused by the spiraling energy, allows for deeper penetration and is excellent for breaking up stagnations when combined with the Kneading Tiger Palm or the Five Thunder Fingers techniques.

Spiraling energy can also be used in the clinic to regulate the patient's Qi activities. Patients may feel the energy penetrating deep into their body and experience light, sound, and mild electric shock. When patients are lying supine, the doctor may choose to spiral the energy superior to the internal organs to pull toxic Qi out of the tissues of the organ; or the doctor may spiral the energy inferior to the internal organs to push the Qi back into the tissues of the organs (Figure 34.8).

Spiraling energy is an essential energetic manipulation used for activating the channel Qi, guiding the energy to spiral in, up, or down, much like a cyclone. The Qigong doctor chooses to use either the circle or spiral energy flow, based upon the patient's condition. Most often these two energetic flows are combined to achieve a more functional treatment. The Qigong doctor will, for example, use "spiraling" energy to drill deep into the body to access the toxic energy stored deep inside the patient's organs. Once the patient's toxic Qi has been reached, the doctor utilizes purging techniques to remove the energetic debris. After purging the doctor can circulate the patient's energy to move the stagnant Qi out of the organs. This may be followed by the Microcosmic Orbit to regulate the patient's energy.

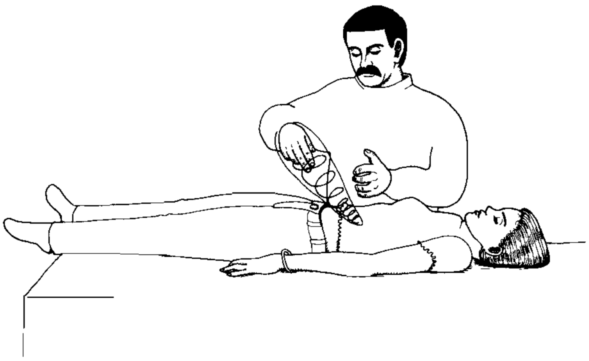

Figure 34.7. A Qigong doctor using the Spiraling Energy technique, to purge Toxic Qi from the patient's body

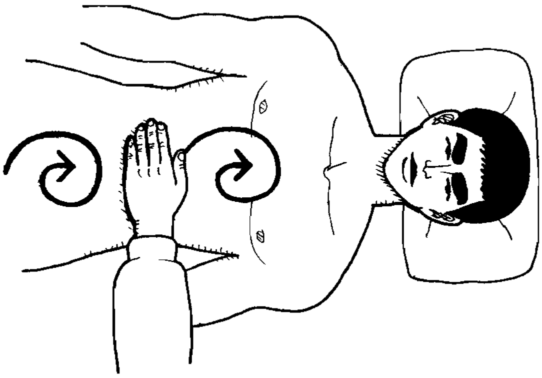

Figure 34.8. A Qigong doctor Spiraling the Qi in order to penetrate the patient's tissues

CLOCKWISE AND COUNTERCLOCKWISE ENERGY FLOW

The doctor focuses on sensing and creating an internal energy vortex within the Lower Dantian that moves up the body in a clockwise or counterclockwise spiral pattern. Only then is the spiraling Qi projected out through the doctor's right hand into the patient's body.

Using the appropriate hand position, the doctor slowly spirals the energy in either a clockwise direction to guide the patient's energy into the affected area, or in a counterclockwise direction

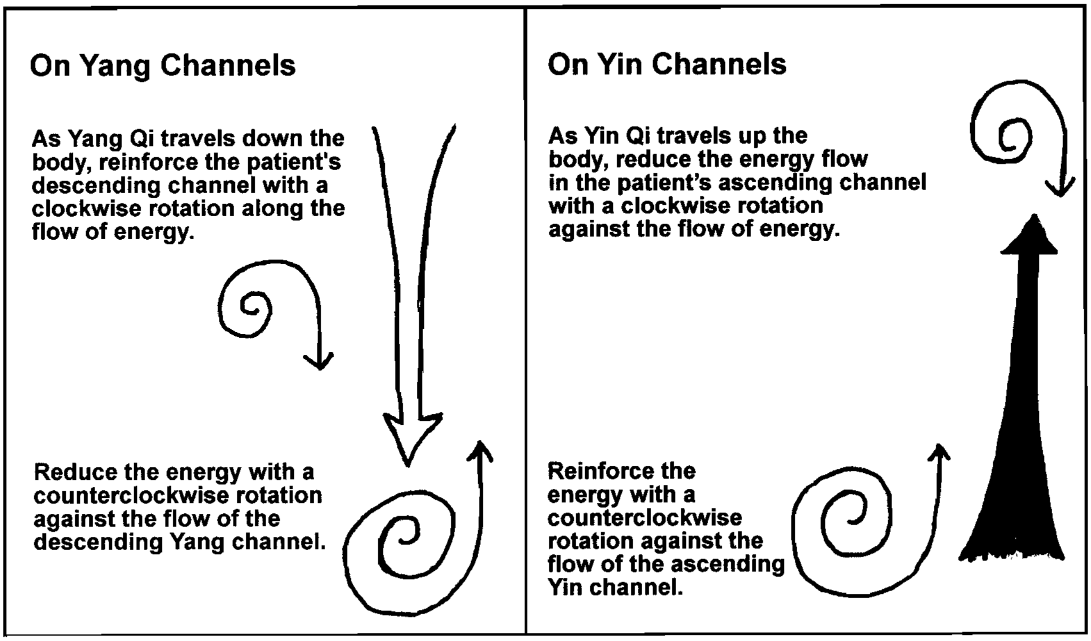

On Yang Channels

As Yang Qi travels down the body, reinforce the patient's descending channel with a clockwise rotation along the flow of energy.

Reduce the energy with a counterclockwise rotation against the flow of the descending Yang channel.

On Yin Channels

As Yin Qi travels up the body, reduce the energy flow in the patient's ascending channel with a clockwise rotation against the flow of energy.

Reinforce the energy with a counterclockwise rotation against the flow of the ascending Yin channel.

Figure 34.9. The reinforcing and reducing techniques of moving energy along the patient's channels is determined by the direction of the Spiraling Energy flow.

to remove the patient's Toxic Qi out of the diseased area.

- The counterclockwise rotation is considered Yin and is used to reduce Heat and to cool the patient. It is generally used to purge conditions of Excess (think of unscrewing the pathogens from the diseased area when purging).
- The clockwise rotation is considered Yang and is used to increase energy and warm the patient. Generally, it is used to tonify conditions of Deficiency (think of drilling and filling the Deficient area when tonifying).

The clockwise spiraling energy originating in the Lower Dantian is emitted and synchronized with the doctor's circling hand movements to reinforce the patient's Qi. To reduce the flow of the channel's Qi, the counterclockwise spiraling technique is used (Figure 34.9).

The Qigong doctor must cultivate these movements into a conditioned reflex before applying these treatment methods in a clinical environment.

SPIRALING ENERGY EXERCISE

- From a Wuji posture, place the left palm on the Lower Dantian. Slowly extend the right palm (keep elbow bent) in front of the chest, aligned with the center line at the Middle Dantian (CV-17), palm facing outward.
- Breathe naturally, while focusing on the Lower Dantian. With the mind's intention, begin to spiral the Qi in the Lower Dantian in a clockwise circle. Lead the energy up through the chest and out the extended right palm. Make sure that the spiraling of the Lower Dantian is synchronized with that of the right palm, making the inner Lao Gong Pc-8 points the center of concentration (Figure 34.10). Start off slowly, and gradually increase speed. The energy circle spirals outward with increasing or decreasing radius, depending upon the doctor's intent.

Repeat the exercise, only this time use counterclockwise spiraling.

ENERGETIC CUPPING

One popular clinical technique called "energetic cupping" is initiated by creating an energetic vortex which remains on the patient, pulling or drawing in energy, similar to that of an acupuncturist's cupping method. These energetic vortices can be used to drain or purge pathogens from a specific point or area on the patient's body. These energetic vortices can also be used to fill or tonify the patient's body. For example, when the patient is prone, the Qigong doctor can stimulate points LI-15, GV-4, and GV-14 (Figure 34.11). Then, using his or her intention, the doctor can pull with one hand while the other hand pushes, increasing the spiraling action already set in motion. This initiated energetic torque can sometimes cause the patient's body to turn, twist, bend, and sway. This modality of treatment simultaneously activates the patient's channel Qi, regulates the Yin and Yang organs, and balances the patient's Yin and Yang energies.

The doctor is advised to rely on intuition or divine guidance in choosing the right method, after careful consideration of the patient's condition (age, sex, personality, illness, etc.).

Note: If combined with the "shaking" technique, the circle / spiraling energy becomes especially effective for relieving pain.

GUIDING THE STATIONARY FLOW OF ENERGY

Stationary energy is emitted from a non-mobile position and can be used to stimulate the channel Qi, points, and Dantians, while restoring and replenishing the patient's Qi. After selecting the appropriate hand posture, the doctor positions his or her hand above the area being treated and emits energy from a fixed or stationary position. This is the most common energy projection technique, in which Qi flows like dense waves or is spaced like a chain of moving pearls flowing outward one after another (Figure 34.12). The patient often feels the sensation of vibration, tingling, or heat. It is not uncommon for the patient to develop spontaneous involuntary reactions such as full body spasms or shaking, leg jerking, etc.

Stationary energy is emitted from a station-

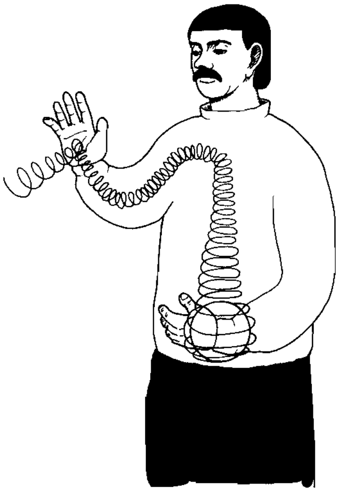

Figure 34.10. An example of a Qigong doctor practicing the Spiraling Energy exercise

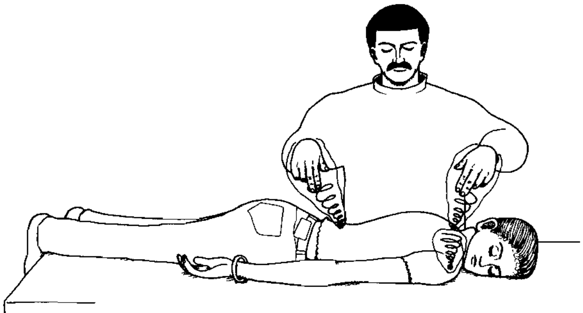

Figure 34.11. An example of a Qigong doctor performing multiple point stimulation with the Energetic Cupping Technique

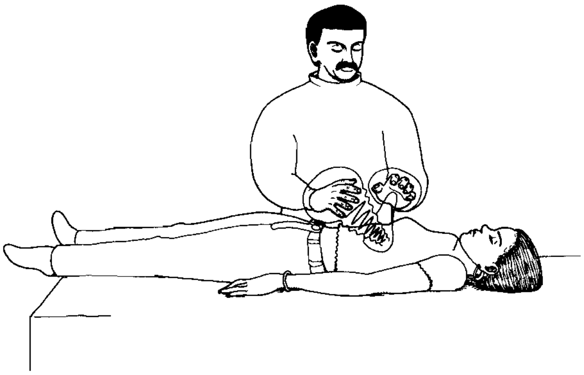

Figure 34.12. When a Qigong doctor extends energy from a stationary position, the Qi flows either like a dense wave, or is spaced like a chain of moving pearls.

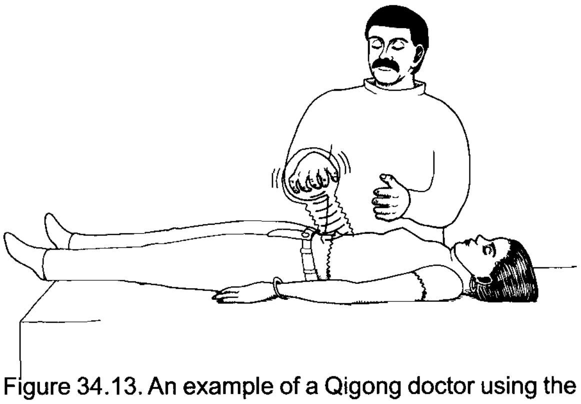

Figure 34.13. An example of a Qigong doctor using the Shaking Energy technique along with the Kneading Tiger Palm

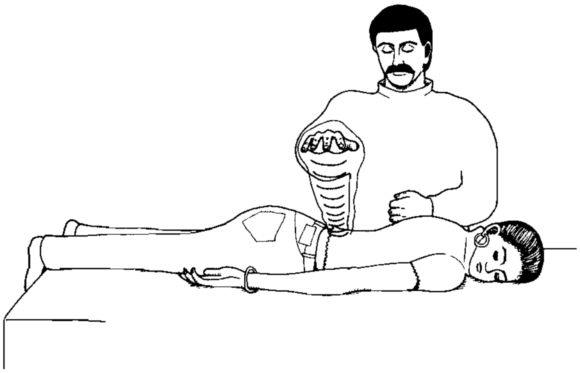

Figure 34.14. An example of a Qigong doctor using Vibrating Energy through the Extended Fan Palm technique

ary palm position. There are two main energetic pulsations utilized in the Stationary Palm techniques: Shaking Energy technique and Vibrating Energy technique.

SHAKING ENERGY TECHNIQUE

The Shaking Energy technique is initiated by the doctor first extending his or her energy over the surface area of the patient's body. After using the Tiger Kneading Palm technique to grasp the Turbid Qi, the doctor slowly begins to guide the patient's Qi out of the tissue area with a stationary manipulation. The doctor then shakes his or her hand from side to side, while continuing to pull the Qi from the area being treated (Figure 34.13).

To practice this particular skill, the Qigong doctor begins from a Wuji posture, using natural and slow breathing. The doctor imagines the waist as the axis and the abdomen as a pump and vibrates the Qi inside his or her Lower Dantian. The doctor then allows the energy to flow upward and out the palm, into the patient. It is important for the doctor to use the mind's intent to follow, guide, and direct this vibration of energy flow, instead of forcing the muscles and palm to shake.

VIBRATING ENERGY TECHNIQUE

Vibrational trembling is one way that memory is stored or processed in the physical body. The Qigong doctor's vibration penetrates the patient's field of "vibrational activity" (the body's natural vibrational rhythm), awakening this memory, allowing the patient to release trapped thoughts and emotions (Figure 34.14).

Qi naturally oscillates and vibrates in the doctor's Three Dantians, each Dantian vibrating at its own unique frequency pattern. The body's bone structure is built with natural formations similar to that of "tuning forks," positioned in both the arms and the legs. These tuning forks are connected together through the body's spinal column and are used to vibrate and increase the body's energetic resonance. Through a balanced interplay of stretching and releasing the body's tendon, ligament, and muscle fibers (in a rhythmic pattern), a natural vibration is produced. After some time of training (through naturally twisting and stretching the tendons), the Qigong doctor, using proper hand postures and intention, is able to direct and lead this vibration throughout his or her entire torso and extend it outwards through the hands.

Vibration is released through three forms of resonation:

1. Jing vibrational resonance is felt within the patient's external tissues and bones. It is related to the energetic release of Heat and the energy of the doctor's Lower Dantian.
2. Qi vibrational resonance is felt deep within the patient's internal organs. It is related to the energetic release of the patient's emotions and the energy of the doctor's Middle Dantian.
3. Shen vibrational resonance is felt within the patient's center core. It is related to the ener-

getic release of the patient's spirit and the energy of the doctor's Upper Dantian.

The release of energetic vibration is primarily used to activate the patient's channel Qi. Clinically, it is most effective when the frequency of the doctor's hand vibrations matches and synchronizes with the vibrational frequency of the patient.

While treating the patient, the doctor's palm usually gets hot and vibrates when coming in contact with Qi stagnation. If a patient's channel point becomes blocked, energy builds up causes Qi stagnation. Once the Qi stagnation is removed, the doctor's hands will cease to vibrate as the channels open and flow freely. When cleared, the patient feels fatigue as the toxins are released from the body and the energy seeks to balance itself.

To apply the Vibrating Technique, the doctor selects the proper hand posture and the area to be treated, then barely touches the patient's body. Any firm pressure should be avoided so as not to hinder the extending and directing of Qi. The doctor initiates the vibrational extension of Qi from the Lower Dantian, slowly turning clockwise or counterclockwise, while directing the vibration out through the hands. The doctor focuses on his or her own breath to adjust the frequency and amplitude of the vibration, the shape and nature of the emitted Qi (Yin or Yang, etc.), and the strength of the energy.

Some examples of hand postures and tissue regulation techniques for emitting vibrational energy include:

a. The Extended Fan Palm is used for activating and regulating Qi activities by:
   - Vibrating the Extended Fan Palm on the Baihui GV-20 point, to descend the patient's Yang Qi, tranquilize the patient's mind, regulate Qi in the Upper Burner, and induce sleep;
   - Vibrating the Zhong Wan CV-12 point to replenished the patient's Qi, to regulate the Stomach's energy and to induce sleep;
   - Vibrating the patient's Lower Dantian, to Lead the Qi back to its origin through supplementing its Qi;
   - Vibrating the Mingmen GV-4 and Da Zhui GV-14 points, to activate Qi in the Governing Vessel.

b. The Sword Finger Palm is used for directing and emitting concentrated vibrational energy into channel points, to stimulate the patient's energetic field.

c. The Invisible Needle Palm is used for directing and emitting the most condensed form of concentrated vibrational energy into the channel points, to stimulate the patient's energetic field.

VIBRATING PALM EXERCISE

The training for this kind of vibrational extension usually begins from a static Wuji posture, then gradually progresses into a dynamic state of energy projection.

- From a Wuji posture, press both palms together at the center of the chest forming a "Buddhist greeting" posture (see Chapter 15). Press the Baihui point upward and the Mingmen backward, while slightly drawing in the chest. Keep hips relaxed, knees bent, tongue on palate, and eyes closed.
- Breathe naturally while focusing the mind on the Lower Dantian. When the Heat and Qi begin to circulate, exhale while mentally guiding the Qi up the chest into both of the palms, by sending the energy through the three Yin channels of the hands.

When inhaling, mentally guide the Qi along the three Yang channels of the hands, back down the back into the Lower Dantian. Maintain natural breathing, and continue to focus attention on both the palms and the fingertips. The palms should heat up and the fingertips tingle, as they expand with Qi (Figure 34.15).

Lead the vibrating energy in the Lower Dantian up into the palms and fingertips, feeling the tissues dissolve inch by inch, cell by cell. The Qi is always centered on the inner Lao Gong Pc-8 point, gathering and collecting but not dispersing. Practice this exercise once or twice a day for 3-10 minutes.

Note: It is imperative that the doctor, when using the technique of "stationary energy pro-

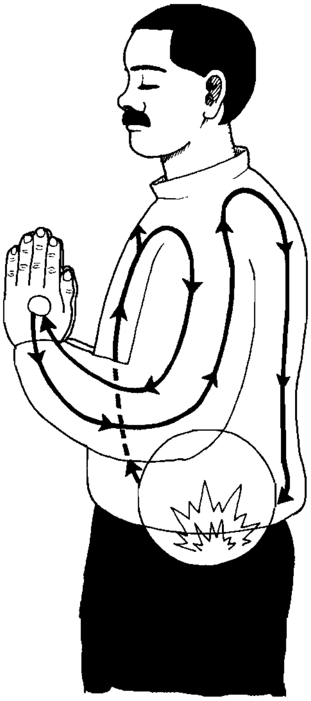

Figure 34.15. The Vibrating Palm exercise helps to increase the Qigong doctor's projected vibration.

jection," never hold his or her breath, or purposely make the hand vibrate by forcefully shaking the muscles. To do so causes Qi stagnation which results in pain and stiffness in the chest and sharp pain in the arms that feels similar to having a fractured bone or a laceration of the muscles. To ensure that there are no difficulties, the doctor should first master the Vibrating Qi method, allowing the vibrating energy to extend and project naturally.

GUIDING THE HOT AND COLD FLOW OF ENERGY

This method refers to the doctor's release of energy being emitted into the patient's body to either regulate Hot syndromes with Cold energy, or Cold syndromes with Hot energy. These two Yin and Yang energetic properties are used to develop the emission of both Hot and Cold energetic flows, through breathing, mental concentration and finger placement.

When Yin Qi is applied, the patient will feel "cold as iron," when Yang Qi is applied, the patient will feel "hot as fire." The mind plays an important role in regulating the doctor's and patient's Yin and Yang energy. When the Qigong doctor thinks of fire, the doctor's body will naturally become hot; conversely, when the doctor thinks of water, his or her body will naturally become cold.

HAND MANIPULATION OF YIN AND YANG ENERGY EXTENSION

The Yin (Cold) and Yang (Hot) energy extension utilizes specific palm and finger arrangements. The index and middle fingers together are assigned Yang properties, whereas the ring and little fingers together are assigned Yin properties. The doctor's left thumb connects with the selected two fingers, to determine the release of Yin or Yang energies. This energy is first gathered from the doctor's left hand, through the arm into the Lower Dantian, then emitted out through the right hand. The Qi that is emitted when the doctor's thumb, index and middle fingers connect, for example, releases the Heart's Yang Qi through the right hand. This technique is used to replenish the patient's Yang and regulate the patient's Heart energy.

GATHERING AND ISSUING YANG (HOT) ENERGY

Before emitting Hot energy into the patient, the doctor's Qi is first gathered into his or her left palm by imagining a small "fiery red sun" radiating heat, light, and fire. This small sun is imagined within the circle created between the thumb and first two fingers on the left hand. The hot energy of the sun is then absorbed through the doctor's left arm into the Lower Dantian. The Heat is then blended with the Lower Dantian's Qi and released out the doctor's right arm through either the tips of the index and middle fingers (the Yang Sword Fingers hand technique) for a deeper, concentrated type of penetration, or through the Extended Fan Palm hand technique for a dispersing type of penetration.

To extend Hot energy deeper into the patient, the right hand Yang Sword Finger posture is used. The doctor's index and middle fingers are extended and joined together, the ring and little fingers curl into the center of the palm. The thumb should rest on the nails of the ring and little finger, forming a circle.

The first or "index" finger indicates the Wood Element, i.e., the Liver's energy, and the middle finger indicates the Fire Element or the Heart's Qi. Together, both fingers activate the Qi of Fire and Wood, which catalyze the Yang energy needed for guiding Heat during point concentration.

Guiding Hot Yang Qi Exercise

This particular meditation is used to increase Heat and light in specific areas of the patient's body (usually for chronic, Deficient, or Cold syndromes). The following meditation, for example, can be used to Heat and tonify the Kidneys in chronic fatigue patients.

- Sit on the edge of a chair in front of a table, feet flat, eyes closed, with the left hand resting comfortably on the left thigh. The right palm is extended face down suspended over the table.
- Breathe naturally while concentrating the mind on the Lower Dantian. Imagine the energy in the Lower Dantian is a hot, burning sun expanding out and enveloping the entire body. Gather the energy back into the Lower Dantian, transforming it into heat and light. Draw the heat and light up into the chest, down the arms, and out the right palm. Concentrate and imagine the heat and light burning and extending into the table, penetrating deep into it. After awhile, switch palms. In China, this exercise is practiced while standing or sitting in the sunshine, facing south or east. Practice for 20-40 minutes (Figure 34.16).

GATHERING AND ISSUING YIN (COLD) ENERGY

Before extending Cold energy into the patient, the doctor's Qi is first gathered into his or her left palm by imagining a small, watery blue moon, radiating cool light. This small moon is imagined within the circle created between the thumb and last two fingers. The cool blue energy of the moon is then absorbed through the doctor's left arm, into the Lower Dantian. The cool blue energy is then blended with the Qi of the doctor's Lower Dantian and released out the doctor's right arm through either the tips of the ring and little fingers (the Yin Sword Fingers hand technique) for a deeper, concentrated type of penetration, or through the Extended Fan Palm hand technique for a dispersing type of penetration.

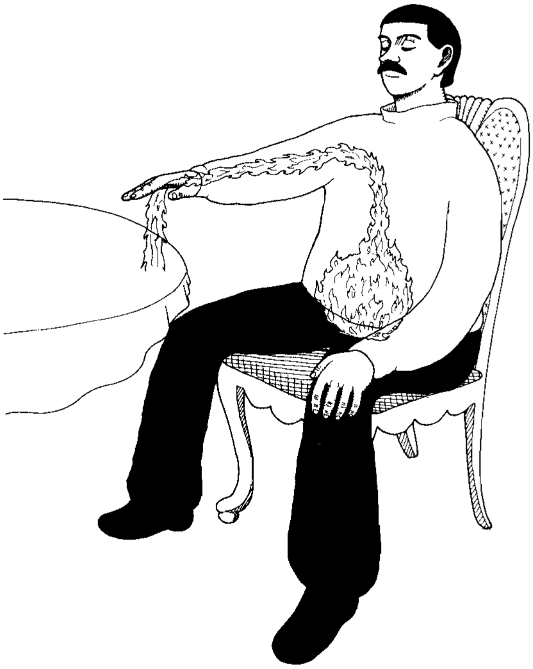

Figure 34.16. Guiding Hot Yang Qi Exercise

In extending cool blue energy into the patient, the right hand Yin Sword Finger posture is generally used. The doctor's ring and little fingers are extended and joined together, and the index and middle fingers curl into the center of the palm. The thumb should rest on the nails of the index and middle finger forming a circle.

The ring finger indicates the Metal Element or the Lungs' energy, and the little finger indicates the Water Element or the Kidneys' energy. Together, both fingers activate the Qi of Metal and Water, which is a catalyst for the Yin energy needed for guiding cool Yin Qi during point concentration.

Guiding Cold Yin Qi Exercise

This meditation is used to cool specific areas in the patient's body in acute, inflamed Excess or Heat syndromes. It can also be used, to cool down and dredge the Liver in cases of chronic depression.

- Sit on the edge of a chair in front of a table, feet flat, eyes closed, with the left hand resting comfortably on the left thigh. The right palm is extended face down suspended over a table.
- Breathe naturally while concentrating the mind on the Lower Dantian. Begin to gather energy from the heels and Bubbling Well (Kd-1) points, directing the energy through the chest (CV-17), out the arms, and into the right palm. Imagine the right palm turning as cold as ice. Focus the mind on the coldness of the right palm and imagine the cold energy penetrating deep into the table. In China, this exercise is practiced at night, facing north or west for 20-40 minutes (Figure 34.17).

Note: While practicing the Cold Yin Guiding exercise, never imagine that the whole body is cold, also do not direct and keep the cold energy into your body (only through it). Otherwise the Cold energy will affect the balance and energetic harmony of the Qi's flow. You can, however, use the image of cool blue water or vapor flowing through your body into the right hand (which is freezing cold), without causing harm to your body's energetic balance.

CLINICAL APPLICATION OF YIN AND YANG ENERGY EXTENSION

Traditionally, the Qigong doctor would emit Hot Yang Qi during the Yang period of the day (from midnight to high noon) to warm the patient's Cold areas. During the Yin periods of the day (from high noon to midnight) the doctor would apply the Cold Yin Qi to cool the Hot areas of the patient's body.

In China today, the clinical application of issuing Hot or Cold energy is no longer limited to the position of the sun and moon, but is applied according to the doctor's ability to regulate the Excess or Deficient condition of the patient's disease.

GUIDING THE ENERGETIC FLOW OF THE FIVE ELEMENTS

This particular Medical Qigong healing system is divided into two popular methods of energy extension: The Five Elemental Organ Energies Method and The Five Elemental Channel Energies Method.

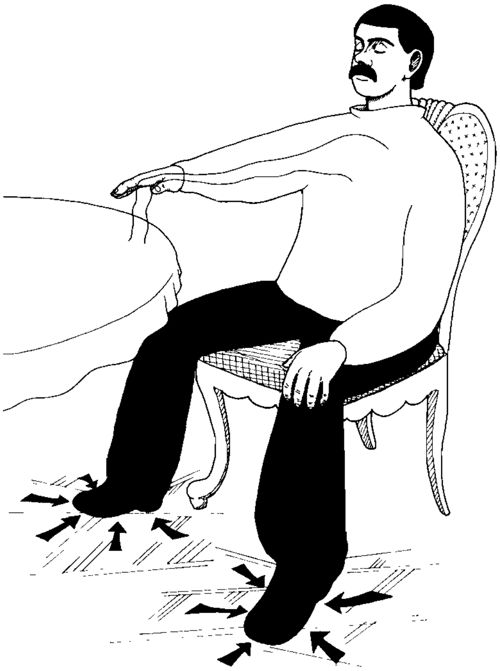

Figure 34.17. Guiding Cold Yin Qi Exercise

Both systems can be used for diagnosing the patient (using palm detection) or for treating Excess or Deficient conditions within the patient's body (using energy extension techniques). Each of these two powerful systems of treatment has its own unique advantage in application. Both systems combine healing sound resonances with energy light extension of various colors, and are used in accordance with the theory of the Five Elements.

THE SIX CHARACTER FORMULA

When utilizing either the Five Elemental Organ, or Five Elemental Channel system of energy projection, the Qigong doctor will guide the energy of the Five Elements in conjunction with the Six Character Formula. The Six Character Formula includes: The Five Elemental Energies, Time of Treatment, Direction of Energy, Position of the Thumb, Guiding Energy, and Emitting Energy and Sound.

1. The Five Elemental Energies of Wood, Fire, Earth, Metal, and Water relate to either the Five Elemental Channel system of energy projection and the Blood / Heat Cycles; or the relate to Five

Elemental Organ system and the energy of the Five Yin Organs. The decision regarding which of the two methods to choose is determined by the patient's condition.

2. The time of treatment is traditionally determined by the doctor in accordance with the patient's Blood/Heat Cycle, the time of day (sunrise, noon, sunset, and midnight) and the season (summer, winter, autumn, or spring).
3. The direction of energy emitted into the patient's body is determined by the doctor in accordance with his or her intention to purge, tonify or regulate the patient's tissues.
4. The position of the thumb is determined in accordance with the type and amount of energy the doctor wishes to absorb into his or her Lower Dantian.
5. Guiding (Pushing, Pulling and Leading) the energy is determined by the doctor in accordance with the Excess or Deficiency of the patient's Five Yin Organs: Liver (Wood), Heart (Fire), Spleen (Earth), Lungs (Metal), and Kidneys (Water).
6. The type of energy (Qi or Shen) and the choice of sound to be emitted are determined by the doctor's intention and the condition of the patient's organs and channels. Qi and sound are emitted like water, Shen is emitted like light.

EMITTING THE ENERGY OF THE FIVE ELEMENTAL ORGANS

The method of projecting the Five Elemental Organ Energies is utilized by blending a specific organ energy together with the Qi stored in the doctor's Lower Dantian. Both of these energies are combined and then issued out the doctor's right palm into the patient. The fusion of the doctor's Lower Dantian Qi with the selected organ Qi is very powerful. This synergized Qi is further combined with sound resonation and colored light to treat patients with extreme conditions of Deficiency.

The left thumb connects with either the tip of the index (Wood), middle (Fire), ring (Metal), little finger (Water) or upper pad of the palm (Neutral) to connect with the selected organ energy (Figure 34.18).

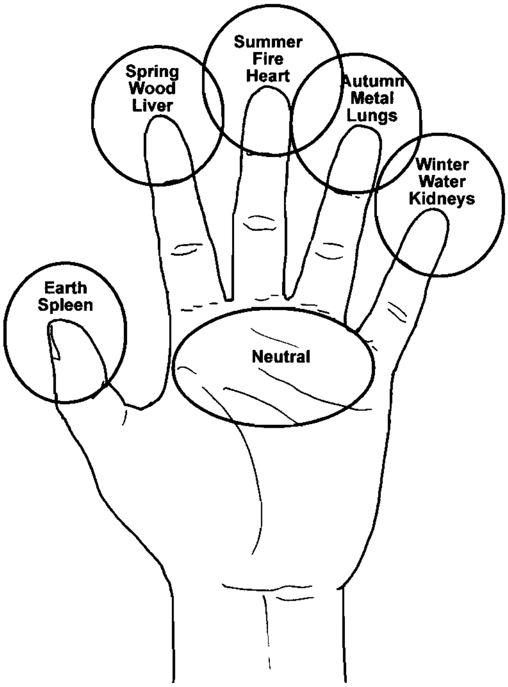

Figure 34.18. Left Palm - The Five Element Organ Energy Pattern: the left hand is used in order to absorb the organ Qi into the doctor's Lower Dantian. This arrangement corresponds to the Five Seasonal Pattern Cycle and the Five Elements.

Five Fingers	**Five Elements**	**Five Viscera**
Thumb	Earth	Spleen
Index	Wood	Liver
Middle	Fire	Heart
Ring	Metal	Lungs
Little	Water	Kidneys

Figure 34.19. Hand Chart of the Five Elements

After choosing the appropriate thumb and finger position on the left hand to tap into the selected organ's energy, the doctor descends this energy into the Lower Dantian. This energy is then blended together with the energy of the Lower Dantian and led out of the doctor's body, to be emitted into the patient using the right hand, along with the sound and specific organ color (Figure 34.19).

EMITTING THE ENERGY OF THE FIVE ELEMENTAL CHANNELS

The method of projecting the Five Elemental Channel Energies is utilized by blending a specific channel energy together with the Qi stored in the doctor's Lower Dantian. To tap into the selected channel's energy, the thumb is positioned on the corresponding finger, or palm segment (Figure 34.20). The selected channel's energy is then blended with the Qi stored in the doctor's Lower Dantian. These combined energies are then issued out the doctor's right palm into the patient. The blending of channel and Lower Dantian Qi is strong but not as powerful as the doctor's organ energy.

The resonant sound and light emitted from the combined channel and Lower Dantian energy is considered appropriate for patients with moderate Deficiency conditions. Healing sounds in combination with specific colors can be emitted by the Qigong doctor to stimulate the internal tissues. When healing sounds are directed into the patients tissues they have a tonifying action, as opposed to the purging affect which happens when the patient's practice healing tones by themselves (see Chapter 40).

The treatment is conducted according to the Five Elemental Creative Cycle. First the energy of the "mother" element is gathered into the doctor's Lower Dantian to be combined with the doctor's Lower Dantian energy. After the doctor collects and mixes the mother's elemental channel energy with the Lower Dantian's energy, the doctor switches the left thumb's position to the mother's corresponding "child" element position. This enables the mother's channel energy to support the strength of the child's channel energy, which is emitted into the patient's body (Figure 34.21).

Note: The doctor first connects to the mother element, then to the child element. Thus the mother's elemental channel energy, Lower Dantian energy and the child's elemental channel Qi are synergized and projected into the patient's corresponding "child" organ. The patient's organ, into which the Qi is emitted, has the same elemental attribute as the doctor's "child" transmitted energy.

1. To Emit Liver Energy and Guide Wood Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Water bands (9, 10). Inhale this Water energy into the Lower Dantian, mixing it together with your Lower Dantian energy.
   - Second, extend the right hand facing outward from the chest, facing the patient. Then exhale and press the thumb onto the Wood channel bands (1, 2) while uttering the sound "Shu." Guide the Wood energy out of the right palm into the patient's Liver organ or channel. In this particular technique, the doctor is using Water energy to support the extension of Wood energy.
2. To Emit Heart Energy and Guide Fire Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Wood bands (1, 2), while inhaling the Wood energy into the Lower Dantian, mixing it together with your Lower Dantian energy.
   - Second, extend the right hand facing outward in front of the chest, towards the patient. When exhaling, press the thumb onto the Fire channel bands (7, 8) with the thumb while uttering the Heart sound "Haa." Guide the Fire energy out the right palm into the treatment area. In this particular technique, the doctor is using Wood energy to support the extension of Fire energy.
3. To Emit Spleen Energy and Guide Earth Qi proceed as follows:
   - First, from a Wuji posture, press the left thumb into the channel Fire bands (7, 8) gathering this Fire energy into your Lower Dantian.
   - Second, extend the right hand in front of the

Colors		Element	Channel
1.	Green	Wood	Gall Bladder
2.	Green	Wood	Liver
3.	White	Metal	Lungs
4.	White	Metal	Large Intestine
5.	Yellow	Earth	Stomach
6.	Yellow	Earth	Spleen
7.	Red	Fire	Heart
8.	Red	Fire	Small Intestine
9.	Blue	Water	Bladder
10.	Blue	Water	Kidney
11.	Red	Fire	Pericardium
12.	Red	Fire	Triple Burner

Figure 34.20. The Twelve Channels are superimposed onto the Qigong doctor's left palm, arranged according to the progression of the Two Hour Qi and Blood/Heat Cycle.

This Blood/Heat Cycle also corresponds to the Four Principle Time Periods and can be divided into quarters and arranged according to the Midnight, Sunrise, Noon, and Sunset patterns of the Sun.

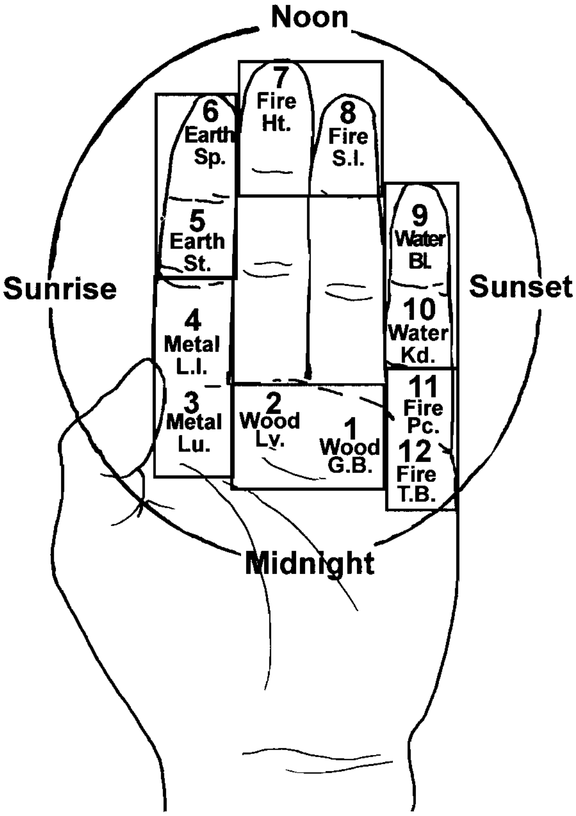

Figure 34.21. Left Palm - The Five Elemental Channel Pattern:

The left palm is used in gathering and absorbing the channel Qi of the doctor's Five Yin Organs. As the Qigong doctor's thumb touches the specific areas on the fingers, the energy is gathered into the doctor's Lower Dantian and released out from the right palm.

chest, facing outward towards the patient. When exhaling, press the thumb into the Earth's channel bands (5, 6) while uttering the Spleen sound "Who." Guide the Earth energy out of the right palm into the treatment area. In this particular technique the doctor is using Fire energy to support the extension of Earth energy.

4. To Emit Lung Energy and Guide Metal Qi proceed as follows:

- First, from a Wuji posture, connect the left thumb onto the tip of the left index finger (i.e., the Earth channel bands #5, 6), thus forming a circle. Imagine this circle full of golden light. When inhaling, draw this golden Earth energy into the Lower Dantian, while also inhaling Heavenly Qi into the Yellow Court area (CV-12) below the solar plexus. Mix all three of these energies together.
- Second, extend the right hand facing outward in front of the chest, towards the patient. When exhaling, press the thumb the Metal channel bands (3, 4) while uttering the Lungs' sound "Sss." Guide the Metal energy out of the body by way of the right palm into the treatment area. In this particular technique, the doctor is using

imagination to draw the Earth's golden light energy into his or her Lower Dantian to support the extension of Metal energy.

5. To Emit Kidney Energy and Guide Water Qi proceed as follows:
    - First, from a Wuji posture, press the left thumb into the Metal channel bands (3, 4). Inhale the Metal energy into the Kidneys, mixing it together in the Lower Dantian.
    - Second, extend the right hand facing outward in front of the chest. As you exhale, press the Water channel bands (9, 10) while uttering the "Chree" sound (in a very low pitch). Guide the Water energy out of the body by way of the right palm. In this particular technique, the doctor is using Metal energy to support the extension of Water Qi.

Once skilled in these five exercises, the Qigong doctor should continue to practice them in combination with the Linear, Circular, or Spiral Energy Guiding Methods, creating an individualized, structured technique.

COMBINING THE ENERGY OF THE FIVE ELEMENTAL TONGUE POSITIONS

The doctor can also use the Five Elemental Tongue Positions to increase the power and energetic potential of the doctor's internal organ. Each tongue position connects with the organ's Orb (the complete organ system), thereby increasing the power of the doctor's Qi projection. These tongue positions are described as follows.

1. To connect with Earth (Spleen) energy—the tongue is placed on the bottom of the jaw at the base of the lower palate, beneath the teeth and gum line.
2. To connect with Metal (Lungs) energy—the tongue is placed between the maxilla and mandible bones, suspended, between the teeth.
3. To connect with Fire (Heart) energy—the tongue is placed on the hard palate, at the front of the upper palate, behind the teeth and gum line.
4. To connect with Wood (Liver) energy—the tongue is placed on the hard palate, at the middle of the upper palate at the center of the roof of the mouth.
5. To connect with Water (Kidneys) energy—the tongue is placed on the soft palate, at the back of the upper palate.

COMBINING THE ENERGY OF THE FIVE ELEMENTAL COLORS

The doctor can also include the Five Elemental Colors to increase the power of his or her internal Qi projections (see Chapter 4). The Five Elemental Colors are described as follows.

- Earth (Spleen)—The organ color bright yellow, the channel color is soft yellow.
- Metal (Lungs)—The organ color bright white, the channel color is soft white.
- Fire (Heart)—The organ color bright red, the channel color is soft red.
- Wood (Liver)—The organ color bright green, the channel color is soft green.
- Water (Kidneys)—The organ color bright blue, the channel color is soft blue.

SUMMARY OF QI EMITTING METHODS

When the Qigong doctor emits Qi into the patient, it effects the quality of light, sound, heat, vibration, and electromagnetic energy stored within the patient's tissues and cells. There are many variations of Qi emission hand postures and techniques used by Qigong doctors in China. Each differs according to the various schools, colleges, and styles of Medical Qigong training. These changeable and diverse hand postures, as well as the various Qi extension, guidance, and regulation techniques can all be combined and grouped together to assist the Qigong doctor in understanding the clinical treatment potentials. The following is an outline of the aforementioned Hand Postures and Extension, Guidance, and Regulation techniques (described in Chapter 33 and 34), along with their applications.

HAND POSTURES USED FOR EMITTING QI

1. The Extended Fan Palm Hand Posture energy emission has the widest beam of Qi extension.
2. The Sword Finger Hand Posture energy emission has the width and beam reduced into a more condensed form of Qi extension.
3. The One Finger Skill Hand Posture energy emission is further reduced into an even more condensed form of Qi extension.

4. The Invisible Needle Hand Posture energy emission is condensed even further into a finely honed needle of light, similar to that of a laser beam.

HAND TECHNIQUES USED FOR BREAKING UP ENERGETIC STAGNATIONS

1. The Thunder Fingers Hand Technique is the strongest, most aggressive hand technique, and is used to break into the energetic boundary of the Qi stagnation. It is used as a preparation technique before purging the stagnation.
2. The Shaking and Trembling Hand Technique is the second most aggressive hand technique, and is used to purge stagnation and toxic Qi from the patient's body.
3. The Vibrating Hand Technique is a powerful, yet subtle hand technique, which can be used to expand, contract, vibrate, and loosen the patient's Stagnant Qi.
4. The Kneading Tiger Hand Technique is used to separate the Toxic Qi, and to dredge and purge it out of the patient's body.

ENERGY EXTENSION AND GUIDANCE TECHNIQUES FOR MOVING QI

1. The Pushing Technique is used to push and move the energy within the body's tissues, channels, and energetic fields.
2. The Pulling Technique is used to pull and move the energy within the body's tissues, channels, and energetic fields.
3. The Leading Technique is used to lead, guide, and move the energy within the body's tissues, channels, and energetic fields.
4. The Circling Technique is used to lead, guide, and move the energy in a circling pattern within the body's tissues, channels, and energetic fields.
5. The Spiraling Technique is used to spiral or drill energy into or out of the body's tissues, channels, and energetic fields.
6. The Energetic Cupping Technique is used to lead, guide, and move the energy within the body's tissues, channels, and energetic fields.

HAND POSTURES USED FOR PURGING, DREDGING, AND REMOVING TOXIC QI

1. The Bellows Palm Hand Posture is used like a vacuum to remove toxic Qi
2. The Dragons Mouth Palm Hand Posture is used to grasp, pull, and lead toxic Qi from the body.
3. The Sword Fingers Hand Posture is used to circle-wrap toxic Qi (like a fork wrapping spaghetti) and remove it from the body.
4. The Kneading Tiger Hand Posture is used to grasp, seize, and pull toxic Qi from the body.
5. The Shaking and Trembling Hand Posture is used in order to grasp, seize, and pull toxic Qi from the body.
6. The Thunder Fingers Hand Posture is used to shake and release toxic Qi from the doctor's body.
7. The Extended Fan Palm Hand Posture is used to dredge, pull, and purge toxic Qi from the body.

TYPES OF ENERGY RELEASED THROUGH QI EMISSION

1. Hot Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
2. Cold Qi Emission is used to cool down or sedate a specific area, organ, channel, or energetic field.
3. Color Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
4. Sound Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
5. Five Elemental Organ Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.
6. Five Elemental Channel Qi Emission is used to tonify or sedate a specific area, organ, channel, or energetic field.

CHAPTER 35
THE INVISIBLE NEEDLE THEORY AND APPLICATION

INTRODUCTION

In China today, the use of Invisible Needles in Medical Qigong therapy is usually reserved for patients who are hypersensitive to energy flow. As the patient's sensitivity increases, the treatments are modified in order to complement their perceptive abilities.

To perform the Invisible Needle technique, the Qigong doctor must first connect with the divine energy. This is performed by drawing energy from the Heavens through the Baihui GV-20 point at the top of the head and extending it out through the hands, enveloping the patient. Some Qigong doctors imagine that needles of light, from the Heavens, descend through their Baihui point. As this light fills their body, they become empowered with a never-ending reserve of Invisible Needles to treat the patient. In order to use these needles, the doctors rotate their wrist clockwise and the Invisible Needle flows out the Pc-8 Point at the center of their palm (Figure 35.1).

Other Qigong doctors use a single needle image. The doctor imagines needles of light coming out of the Heavens, spiraling around the head like a golden halo. This golden halo is positioned above the head, above the Heavenly Transpersonal Point (located about a foot above the head). To use these needles, the doctors reach above their head and pluck the needles from the Heavens. These needles are then inserted into the patient's body (Figure 35.2). When using either visualization, the important factor is to focus on how deep the needles should penetrate. When treating an area deep in the major organs, a clockwise rotation reinforces and tonifies, while a counterclockwise rotation sedates the organ area.

PREPARATION

Before inserting an Invisible Needle into the

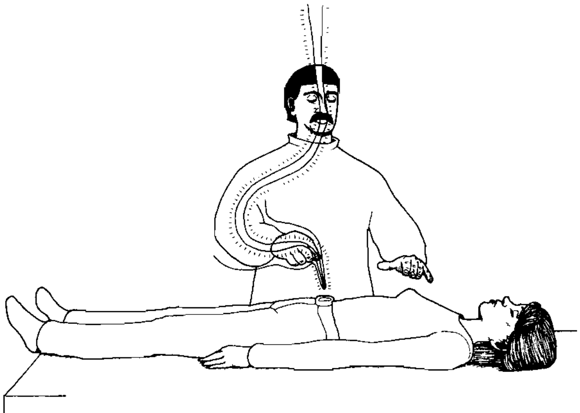
Figure 35.1. The Qigong doctor imagines divine healing light descending from the Heavens, filling the doctor's body with the energy to create the Invisible Needle.

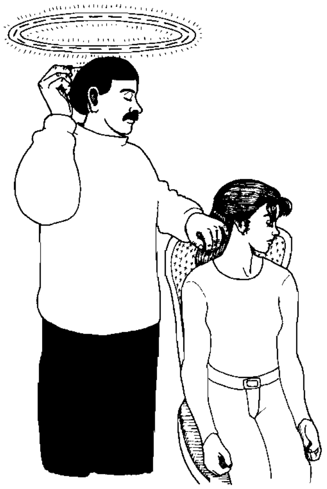
Figure 35.2. Another variation is for the Qigong doctor to imagine divine healing light descending from the Heavens and encircling the doctor's head like a golden halo. Contained within this halo are the Invisible Needles.

patient, it is important for the doctor to dredge the diseased area first. This is similar to swabbing an area with alcohol (or disinfectant) before inserting an acupuncture needle. As one hand extends the Invisible Needle, the other is used to reinforce its energy. The doctor works the Invisible Needle into the patient's tissues until he or she feels it inside the patient's body. Sometimes the patient may feel a cold sensation while being treated by the doctor. This is a normal reaction to the Invisible Needle insertion. The longer that the doctor keeps the Invisible Needle inserted inside the patient's body, the better the healing effect.

ANGLE OF INSERTION FOR TONIFICATION OR SEDATION

The skill of using Invisible Needles encompasses a complete system of energetics that utilizes both reinforcing (tonifying) and reducing (sedating) techniques. The polarity of the patient's channel (Yin or Yang) determines the angle at which the Invisible Needle is inserted.

- When performing reducing or sedating techniques, it is important to insert the Invisible Needles perpendicularly into the patient's points or channels, forming an energetic dam to slow the flow of Qi (Figure 35.3). The doctor removes the needle slowly, and leaves the point open (this allows the point to continue to release Qi naturally).
- When performing any reinforcing or tonifying techniques, insert the Invisible Needle at an angle, pointed in the direction of the channel's flow (provided that the channel is flowing towards the direction of the intended organ). Then, when retracting the Invisible Needle, remove it quickly and seal the point. This quick retraction method is done to prevent the Qi from escaping the area once it is stimulated, and is followed by pressing the needle hole to seal the point (Figure 35.4).

One favorite tonification technique used by certain Qigong doctors in Beijing, China is to leave the Invisible Needles inside the patient, while regulating the Conception and Governing Vessels. Once they have strengthened the patient's Microcosmic Orbit (Fire Cycle), they retract the Invisible Needles and disperse them one by one into the Earth. The insertion and manipulation of the Invisible Needles is performed with the Invisible Needle Palm technique.

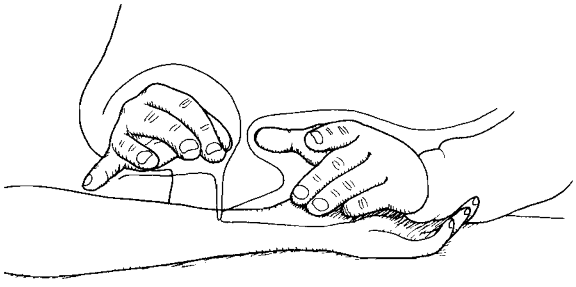

Figure 35.3. The doctor's right hand extends the Invisible Needle perpendicularly into the patient's point or channel, to reduce or sedate the energy flow. The doctor's left hand is used to reinforce the energetic purging of the patient's toxins.

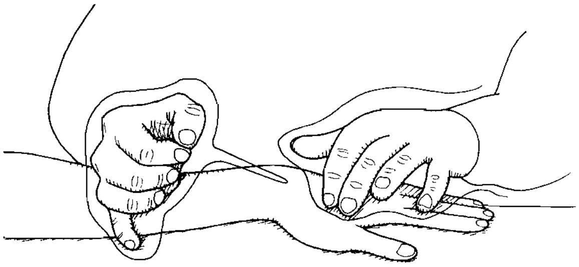

Figure 35.4. Insert the Invisible Needle at a 45 degree angle with the flow of the patient's channel to reinforce or tonify, provided the channel's flow is towards the organ.

In order to prevent the Qi from escaping, when performing any reinforcing or tonifying techniques, the Qigong doctor presses the patient's channel point after retracting the Invisible Needle.

Another favorite technique used for tonification is for the Qigong doctor to first insert the Invisible Needle deep into the patient's Kd-1 point at the bottom of the feet. Once the needle is inserted, the doctor extends his or her intention, causing the needle to grow and extend up the patient's legs and into the Kidneys. Next, the doctor emits energy up the patient's legs to further tonify the Kidneys and Mingmen area.

The Invisible Needles can also be transformed into the specific colors of the Five Elements to enhance tonification. A blue needle, for example, can be used to intensify Kidney tonification. Each needle can additionally be spiraled in a clockwise or counterclockwise direction to generate an energetic vortex used for either tonification or purgation (Figure 35.5).

THE HAND POSTURES OF THE INVISIBLE NEEDLE

The following is a description of the various hand postures used when treating patients with the Invisible Needle technique.

THE DOCTOR'S RIGHT HAND

The doctor's right hand is generally used for controlling and projecting the energy of the Invisible Needle into the patient. It is not only responsible for the formation but also the insertion of the Invisible Needle. The action and movement of the right palm projecting and stimulating Qi is sometimes referred to as the Chicken Pecking technique. In this hand posture, the thumb connects with the index finger or index and middle finger to form the shape of a bird's beak (Figure 35.6). While these fingers are joined together, the other fingers on the hand relax in an extended, straight position. The doctor's intention is focused on the image of an Invisible Needle which is projected into the patient's body.

The pricking, insertion, and manipulation of the Invisible Needle is only effective when the Qigong doctor connects and roots his or her intention deep within the patient's Jing, Qi, and Shen. This intention connects the doctor with the patient's Qi and Blood, and their animated Spirit.

THE DOCTOR'S LEFT HAND

The doctor's left hand is used to either manipulate the patient's tissues to enhance the Qigong treatment, or to collect Qi into the doctor's Lower Dantian to be projected into the patient. This technique is similar in method to collecting the Yin (Cold) and Yang (Hot) energy before emitting it into the patient. Once the thumb connects with any of the fingers, the released energy bears the nature of the associated element and its Yin and Yang properties (Figure 35.7). The Qi that is collected when the thumb and middle finger connect, for example, is the Yang Qi within the Heart.

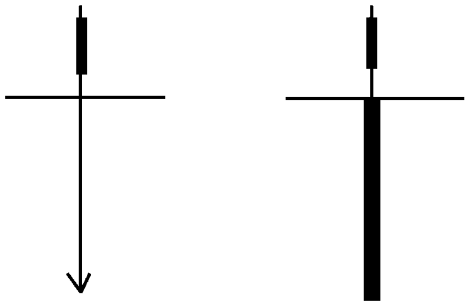

The needle can be made to grow and extend deep inside the patient's tissues or channels.

The needle can be made to grow fat inside the patient's tissues or channels to increase Qi projection.

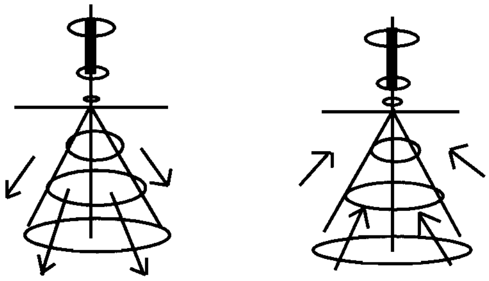

Clockwise spiraling the needle is used to create an energetic vortex for tonification.

Counterclockwise spiraling the needle is used to create an energetic vortex for purgation.

Figure 35.5. The Invisible Needle Techniques

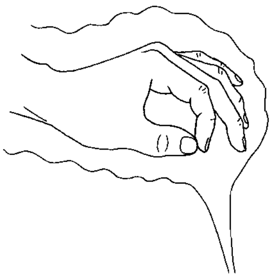

Figure 35.6. The Invisible Needle Palm

This Qi replenishes the Yang Qi and regulates the energy of the Heart (Figure 35.8). If the doctor's hand is positioned over the Pericardium Channel and energy is emitted into the patient, the patient's Shen will become peaceful. When treating with the Five Elemental organ energy, the index and middle finger stand for Fire and Wood; this combination is considered Yang. The ring and little finger stand for Water and Metal; this combination gathers and creates Yin energy.

TIME AND DURATION OF TREATMENT

In general, Medical Qigong Invisible Needle Technique and Point Therapy is applied once a day. A routine therapeutic course includes six to eighteen treatments. Patients with mild diseases may continue the therapy from six to twenty-four treatments. For chronic patients, however, treatment may last as long as one to three months depending on the patient's condition (i.e., for paraplegic patients the treatment is generally from three to six months).

As far as the time sequence of when to treat the patient, or when the patient is to treat themselves, the chronometric rate is based on the high-tide sequence of the patient's channels (i.e., 11 am to 1 pm is the Heart time, which is the best time period to treat patients with Heart disease).

REMOVING THE INVISIBLE NEEDLES AND ENDING THE TREATMENT

During tonifying treatment, the Qigong doctor removes the Invisible Needle quickly from its point of origin while the patient is inhaling. This allows the patient to keep the Qi and Shen within the body's channels and tissues. After extracting the needle, the doctor presses down on the acupuncture point with emitted Qi and lightly rubs it in a clockwise direction until the open point has closed.

During a sedating treatment, the Qigong doctor has the patient exhale to expel the Evil Qi and Heat from the patient's tissues. After extracting the needle slowly, the doctor allows the acupuncture point to naturally drain and eventually close by itself. This aids in draining the Evil Qi and Excess Heat from the patient's body.

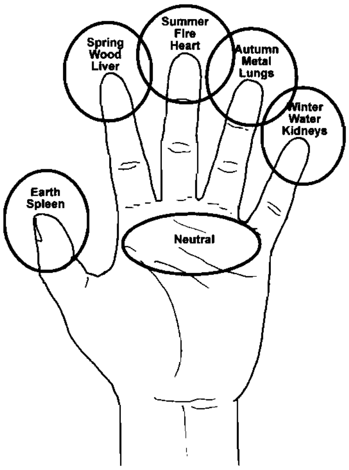

Figure 35.7. When treating patient's with the Invisible Needle technique for tonification, the Qigong doctor emits the required amount of energy into the patient, according to the specific Deficient organ and channel's condition.

The thumb, being the Earth Element and neutral, allows the Qigong doctor the ability to connect and combine the energy of the Five Yin Organs into his or her Lower Dantian, before extending the Qi into the patient.

Five Fingers	Five Elements	Five Viscera
Thumb	Earth	Spleen
Index	Wood	Liver
Middle	Fire	Heart
Ring	Metal	Lungs
Little	Water	Kidneys

Figure 35.8. Hand Chart of the Five Elements

CHAPTER 36
ENERGETIC POINT THERAPY

The importance of Energetic Point therapy as a Medical Qigong modality can best be understood when exploring the amount of information stored within the body's tissues. According to theoretical physicist Michio Kaku, the body's DNA contains over one hundred trillion times the information stored in current computers. There is an even greater amount of information stored within the three subtle energy fields outside of the human body. Most of this information is not readily translatable or decipherable, and is hidden from the conscious mind.

By stimulating certain energy points, for a specific length of time, an energetic trigger mechanism, or response reflex is created. This energetic reaction releases memories and stored information unique to each patient. To facilitate the unlocking of this stored information, the Qigong doctor must maintain a deep state of Alpha consciousness when stimulating the patient's energy points.

Energetic Point therapy involves two distinct modalities of treatment: Channel Point treatments and Channel Point meditations. Either can be used in conjunction with the Invisible Needle therapy. In these particular modalities of treatment, the Qigong doctor can either emit Qi into specific channel points on the patient's body to enhance the Medical Qigong treatment, or have the patient meditate on a specific point while the doctor administers the therapy. Both approaches can be used by the Qigong doctor to gather energy, move Qi, or release trapped energy from the patient's tissues.

CHANNEL POINT TREATMENTS AND PRESCRIPTIONS

The following treatments are generally used in the clinic by the Qigong doctor for Invisible Needle therapy, but are not necessarily limited to that modality of treatment. These Energetic Point treatments are arranged and categorized according to specific diseases. The Energetic Points are located on the patient's front and back as depicted in the anatomical graphs (Figures 36.1 and 36.2). These points are clinically used for the following treatment goals:

POINT THERAPY TO TONIFY YIN ENERGY

1. To Tonify Yin energy and to reduce a Deficient type of Fire, extend Qi into the patient's Kd-3 point.
2. To nourish the Yin energy and to reduce Fire, extend Qi into the patient's Kd-3 and Lu-10 points.
3. To treat a Deficiency of Yin of the Liver and Kidneys and to reduce the causes of Deficient Fire, extend Qi into the patient's Sp-6 and Lv-3 points.
4. To treat a Deficiency of Yin causing a dryness of the Lungs, extend Qi into the patient's Lu-1 and Bl-13 points on the front and back of the body. Regulate the respiratory tract, as well as the patient's Lu-7 and Kd-6 points, thereby tonifying the body's Yin energy and activating the descending function of the Lungs.

POINT THERAPY TO TONIFYING OR DISPERSE YANG ENERGY

1. To strengthen the Yang of the patient's Kidneys, extend energy into the patient's GV-4, Bl-23, and Kd-3 points.
2. To warm, and tonify the Yang of the patient's Spleen and Kidneys, extend energy into the patient's Bl-20 and -23 points. This treatment is ideal for eliminating Damp fluid caused from Deficient types of edema.
3. To fortify the Yang and avert a state of Yang collapse, tonify the patient's Yuan Qi by extending energy into the patient's Lower Dantian, focusing specifically on the patient's CV-4, -6, and -8 points.

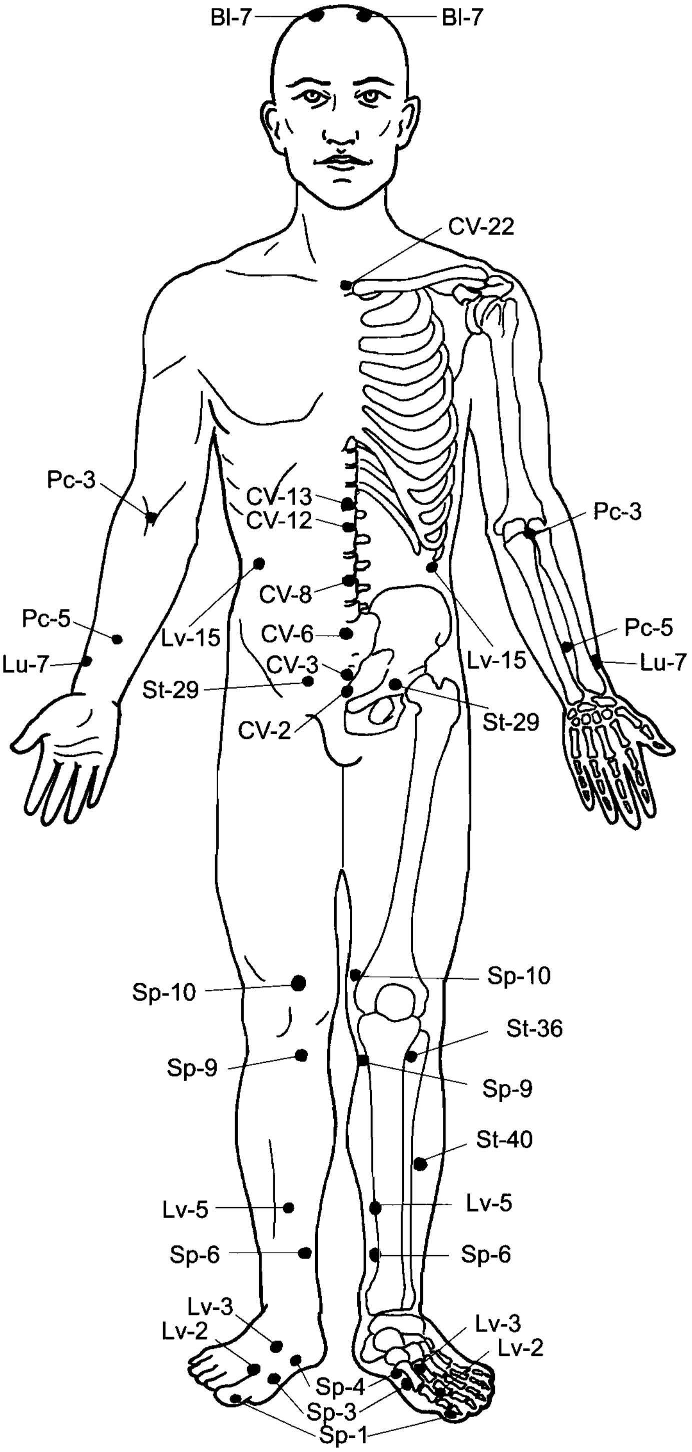
Bl-7
Bl-7
CV-22
Pc-3
CV-13
CV-12
Pc-3
CV-8
Pc-5
Lv-15
CV-6
Pc-5
Lu-7
CV-3
Lv-15
Lu-7
St-29
CV-2
St-29
Sp-10
Sp-10
St-36
Sp-9
Sp-9
St-40
Lv-5
Lv-5
Sp-6
Sp-6
Lv-3
Lv-3
Lv-2
Lv-2
Sp-4
Sp-3
Sp-1

Figure 36.1. Chart of Specific Channel Points

4. To treat a condition of Deficiency of Yang of the Spleen, strengthen the patient's Stomach and Spleen to eliminate Damp and Phlegm. This is done by extending energy and tonifying the patient's St-36, CV-12, and Bl-20 points. Then activate and warm the Qi of the patient's Lungs by extending Qi into Bl-13, and -43 points.
5. To strengthen the Yang of the Spleen, promote the Spleen's transporting function by extending energy into the patient's CV-12, Bl-20, Lv-13, and Sp-3 points.
6. To raise the patient's Yang Qi or the Yang Qi flowing within the patient's Governing Vessel, extend energy into the GV-20 point.
7. To reduce hyperactive Yang of the patient's Liver, first dredge and disperse the patient's GB-20, Bl-18, and Lv-2 points to pacify the Yang of the Liver; then tonify by extending energy into the patient's Bl-23 and Kd-3 points to strengthen the Kidneys.
8. To reduce Excess Yang within the patient's Governing Vessel which has caused a hyperactivity of Heat, dredge and sedate the patient's GB-2O point. Then purge the Heat from the patient's GV-14 and LI-4 points.

POINT THERAPY TO TONIFY OR DISPERSE QI

1. To strengthen the patient's Yuan Qi, extend Qi into the patient's Lower Dantian, focusing on CV-4 and -6 points.
2. To fortify the patient's Qi and reestablish their Yang, extend energy into the patient's GV-20, CV-6, and St-36 points.
3. To subdue the patient's ascending Qi, extend Qi into the patient's CV-22 and Bl-17 points.
4. To disperse the patient's stagnant Qi and Phlegm, extend Qi into the patient's Bl-13 point.
5. To tonify the Kidneys' Qi and Jing and to reduce any Kidney Deficiency, extend Qi into the patient's GV-4, Bl-52, and Kd-3 points.
6. To stimulate the circulation of Qi and Blood, extend Qi into the patient's LI-4 and Lv-3 points.

POINT THERAPY TO TONIFY AND NOURISH THE BLOOD

1. To activate the patient's Blood circulation, extend energy into the patient's Sp-10 point.
2. To stimulate the function of conducting Blood, extend energy into the patient's Sp-1 point.
3. To control the production of Blood, extend energy into the patient's Bl-20 point to control the Blood, as well as Bl-15 to increase the production of Blood.
4. To activate Blood circulation and remove stasis, extend energy into the patient's Sp-6 and Bl-17.
5. To strengthen the source of Blood formation, extend energy into the patient's St-36 and Sp-6 points.
6. To remove Blood stasis of the uterus, extend energy into the patient's Bl-32 and St-29 points.

POINT THERAPY TO ELIMINATE COLD AND DAMP

1. To warm the Spleen and Stomach to eliminate Cold, extend energy into the patient's Lower Dantian and CV-12 and -13 area.
2. To disperse Cold, relieve pain, pacify the Stomach and strengthen the Spleen, extend energy into the patient's Sp-4 and Bl-20 points.
3. To strengthen the Spleen to disperse Dampness, extend energy into the patient's Sp-6 and St-36 points.
4. To eliminate edema above the waist, extend energy into the patient's Bl-20 and Lv-13 points.
5. To eliminate edema below the waist, extend energy into the patient's Bl-28, LI-6, and Sp-9 points.

POINT THERAPY TO ELIMINATE HEAT

1. To eliminate Heat from the patient's body, extend energy into the patient's LI-11 points, or dredge and disperse energy from the patient's Shi Xuan points (at the tips of the fingers).
2. To move Heat downwards from the patient's body, dredge and disperse energy from the patient's Kd-1 points.
3. To disperse External Heat, dredge and disperse energy from the patient's SI-3 points.
4. To eliminate Internal Heat, dredge and disperse energy from the patient's Pc-5 points.
5. To eliminate Damp Heat, extend energy into the patient's Bl-20 and Sp-9 points.

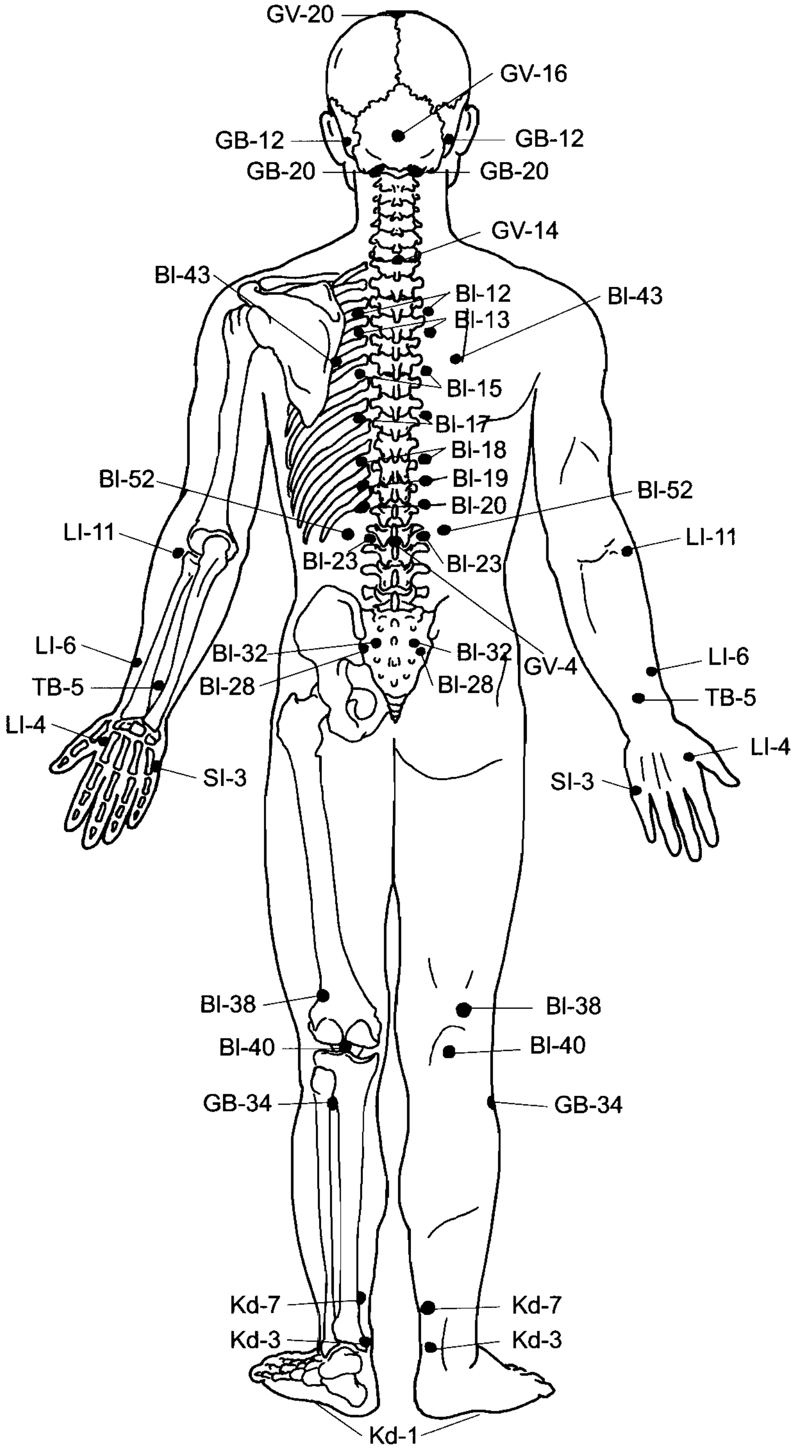
GV-20
GV-16
GB-12
GB-12
GB-20
GB-20
GV-14
Bl-43
Bl-12
Bl-13
Bl-43
Bl-15
Bl-17
Bl-18
Bl-19
Bl-52
Bl-20
Bl-52
LI-11
LI-11
Bl-23
Bl-23
GV-4
LI-6
Bl-32
Bl-32
LI-6
TB-5
Bl-28
Bl-28
TB-5
LI-4
LI-4
SI-3
SI-3
Bl-38
Bl-38
Bl-40
Bl-40
GB-34
GB-34
Kd-7
Kd-7
Kd-3
Kd-3
Kd-1

Figure 36.2. Chart of Specific Channel Points

6. To eliminate Damp Heat from the Blood, dredge and disperse energy from the patient's Bl-40 points, then extend energy into the Pc-3 points.
7. To reduce Fire of the Liver, dredge and disperse energy from the patient's CV-3 and Lv-5 points.
8. To reduce Liver and Gall Bladder Upward Fire disturbance, dredge and disperse energy from the patient's GB-12, as well as Bl-19 and -18 areas.

POINT THERAPY TO ELIMINATE WIND

1. To eliminate Wind Cold, dredge and disperse energy from the patient's GV-16, Bl-12, GB-20, Lu-7, LI-4, and Kd-7 points.
2. To eliminate Wind Heat, dredge and disperse energy from the patient's GV-14, LI-4, TB-5, and GB-20 points.
3. To eliminate Wind from the upper portion of the patient's body, dredge and disperse energy from the patients Bl-7, GV-20, and -16 points.
4. To eliminate pathogenic Wind, extend energy into the patient's Lu-7 points to activate the natural dispersing function of the Lungs.
5. To eliminate Wind and reduce Fire, dredge and disperse energy from the patient's GB-20 points.
6. To calm the Wind of the Liver, dredge and disperse energy from the patient's Lv-3 points.

POINT THERAPY TO ELIMINATE PHLEGM

1. To eliminate Damp Phlegm in the Interior, first strengthen the function of the patient's Spleen and Stomach to eliminate the Dampness, by extending energy into the patient's Bl-20 and CV-12 points, then eliminate the Phlegm by dredging both of the St-40 points.
2. To eliminate Phlegm Fire in the patient's Stomach, dredge the patient's GB-34 and St-40 points.

CONTRAINDICATIONS

Never apply Energy Point therapy to patients who are in an acute stage of inflammatory diseases (especially of the abdomen). Also avoid using Energy Point therapy on patients with hypertension, heart diseases, severe cases of pulmonary tuberculosis, hemophilia, purpura hemorrhagic, purpura allergic, and severe skin diseases.

CHANNEL POINT MEDITATION

After applying the Invisible Needle technique or Energetic Point therapy, the Qigong doctor may decide to initiate Channel Point meditation to intensify the treatment. This is achieved by having the patient concentrate on specific points along the channels. This concentration will produce two distinct results. First, it will get rid of any of the patient's distracting thoughts, and second, it will cause different fluctuations of the Qi and Blood circulation through the patient's Yin and Yang organs. This focused concentration will allow the internal organs to supply the patient's body with either: a more concentrated flow of Qi and Blood into the tissue area; or the focused attention will have a stronger dispersing effect on the patient's tissues (depending on the patient's focus).

Each time a patient switches the focus of concentration to a different point in a different area of the body, the circulation and regulation of Qi and Blood will alter to accommodate the mental transitions of intention. In giving the patient a specific point meditation as a prescription, the Qigong doctor should be careful to select the areas and points in accordance with the particular Qigong exercises that will arrest the patient's Yin and Yang organ disease.

Generally speaking, when treating an Excess condition with Channel Point meditation, the Qigong doctor has the patient imagine draining the Excess Qi out through the extremities. In the case of a Deficiency, the doctor has the patient focus on tonifying the Deficient organ(s) by absorbing energy and light into the Deficient area. Below is a list of channel point regulations and their applications.

QIHAI (CV-6) SEA OF QI

The CV-6 point, located in the Lower Dantian, is the most frequently selected point (Figure 36.3). It is chosen for normal conditions to lead the Qi back to its origin. It is this area from which Qi emanates and returns, hence its name, Sea of Qi. Focusing on this point will allow the patient to tonify the Kidneys (Yuan) Qi, and can be used to regulate the Lower Burner, Conception Vessel, and Water pathways. The Qigong doctor may also extend energy into this area to:

- restore collapsed Yin or Yang Qi,
- raise the patient's Middle Burner Qi,
- warm a patient's Yang Qi,
- warm a patient's Cold condition,
- treat mental disorders,
- treat male sexual problems,
- treat urinary problems,
- treat local pain,
- treat fatigue, and
- treat Qi and Blood Deficiency.

MINGMEN (GV-4) GATE OF LIFE

This point is also called the Palace of Essence (Figure 36.4). Located between the two Kidneys, this point is used for patients with a Deficiency of the Kidney Yang and declining Fire of the Mingmen (this condition is manifested as lumbago pain, seminal emissions, and an aversion to cold). Focus the patient's concentration on the navel first, then gradually have the patient focus deeper into the Mingmen area. This area is known as the Sea of Blood and Essence. Focusing on this point will allow the patient to tonify and stabilize the Kidney (Yuan) Qi and Jing. A Qigong doctor extending energy into this area can regulate the patient's water pathways, heat the body's Yang Qi, and lower blood pressure. This point is also used to treat lower back pain, sciatica, as well as sexual/genital problems.

SHAO SHANG (LU-11) LESSER SHANG

This point is also called Ghost Sincerity, and is considered the second of eleven ghost points (Figure 36.5). The sound "shang" is a musical note and corresponds to the Element Metal. This particular point is used for patients with a Deficiency of Lung Qi (this is manifested by asthma and coughing). It reinforces the Qi and regulates the Lungs. A Qigong doctor can pull energy from this area to dredge and clear Lung Fire, Heat, and Summer Heat, as well as dispel Wind Heat from the patient's body.

ZHONG CHONG (PC-9) MIDDLE RUSHING

This point is used for patients with a Deficiency of Heart Qi (which manifests as palpi-

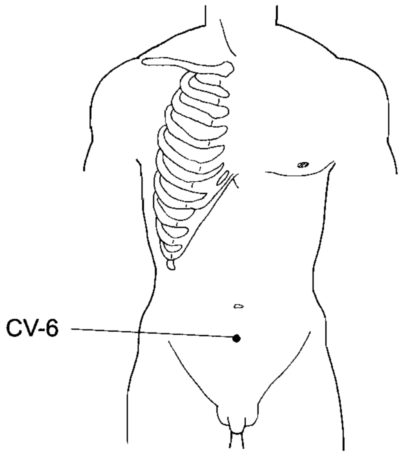

Figure 36.3. Lower Dantian (CV-6)

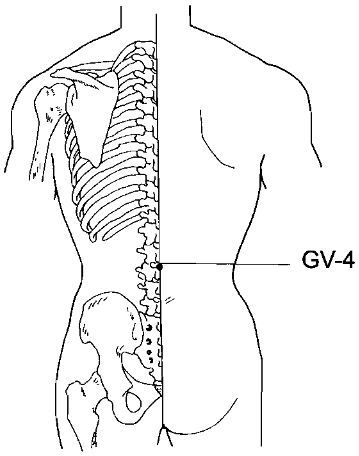

Figure 36.4. Mingmen (GV-4)

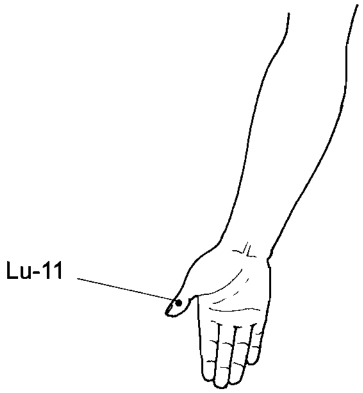

Figure 36.5. Shao Shang (Lu-11)

tations and restlessness). It reinforces the Qi and calms the Heart and mind. This area is good for regulating the Heart's Qi, reviving consciousness, and restoring collapsed Yang. It can be used to dredge and clear the patient's body of Heart Fire, Heat, and Summer Heat conditions (Figure 36.6).

ZU SAN LI (ST-36) FOOT THREE MILES

This point is used for patients with a disorder of transporting and transforming Spleen and Stomach Qi (which manifests as abdominal distention and pain). This point has a strong tonifying effect (especially for the Spleen) and regulates the Stomach. It also tonifies the Ying (Nutritive) Qi, and regulates the Lower and Middle Burners. It can be used by a Qigong doctor to dredge and reduce digestive stagnation, to redirect Rebellious Qi, to drain pathogenic influences from the Stomach, and to soften hard abdominal masses or tumors (Figure 36.7).

DA DUN (LV-1) GREAT PILE

This point is used for patients with hyperactive Liver Yang or overactive Liver and Heart Fire. It calms the Liver to treat diseases of the upper portion of the body (primarily the head). By focusing on this area, the patient can regulate and tonify the Liver Qi and Blood. The Qigong doctor may dredge this area to disperse Liver Qi and transform Damp Heat in the Lower Burner (Figure 36.8).

YONG QUAN (KD-1) GUSHING SPRING

This point is used for patients with Yin Deficiencies of the Liver and Kidney, hyperactivity of Fire due to Yin Deficiencies, or Excess in the upper torso and Deficiency in the lower torso. Concentration on this area will also calm the spirit, clear Fire and Heat from the head area, restore collapsed Yang, and transform Heart Phlegm (Figure 36.9).

BAIHUI (GV-20) HUNDRED MEETINGS

This point is used for patients with a Deficiency of Qi in the Middle Burner (manifested by a shortness of breath, dizziness, intractable diarrhea, and a prolapse of the internal organs). This point is known as the Sea of Marrow. Focusing on this area will clear the brain and calm the spirit. If the Qigong doctor treats this area, the patient's

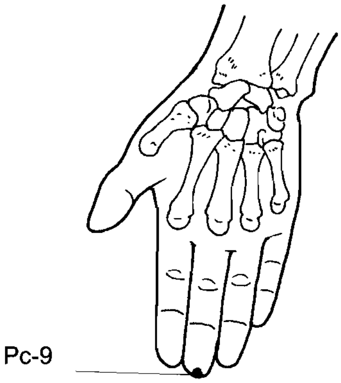

Figure 36.6. Zhou Chong (Pc-9)

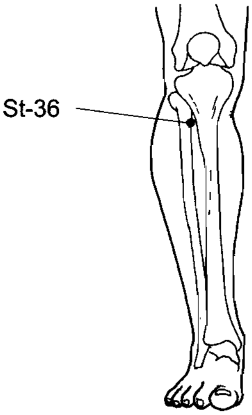

Figure 36.7. Zu San Li (St-36)

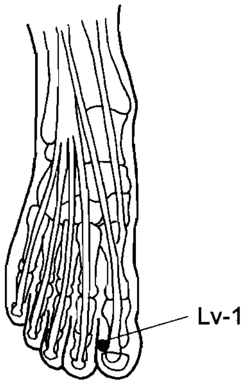

Figure 36.8. Da Dun (Lv-1)

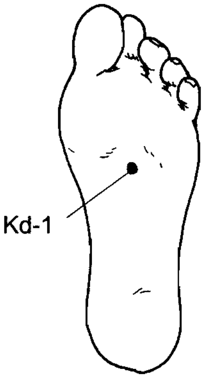

Figure 36.9. Yong Quan (Kd-1)

lower orifices can be stabilized, the Qi can be warmed and tonified, and Yang Qi can be warmed and a collapse can be prevented. This point is also used in order to treat hypertension, insomnia, seizures, dizziness, headaches, and the prolapse of internal organs (Figure 36.10).

HUIYIN (CV-1) MEETING OF YIN

This point is used for patients with a Deficiency of Lung Qi and Kidney Yin (manifested by asthma and coughing). Focusing the patient's concentration on this area stabilizes the patient's Jing and lower orifices, as well as calms the spirit and clears the brain. It can be used by the Qigong doctor to tonify and regulate the patient's Qi, to clear Heat from the patient's body, to treat irregular menses, urethritis, prostatitis, and to raise the blood pressure (Figure 36.11).

YIN TANG (EX. HN. 3) SEAL HALL

This point is used for patients to calm the spirit, and can be used by the Qigong doctor to dispel Wind and clear Heat from the patient's body. This point is also used to treat epilepsy, nausea and vomiting, insomnia, sinus headaches, dizziness, and vertigo (Figure 36.12).

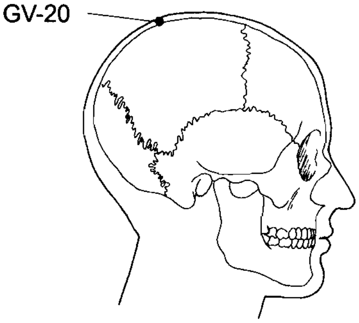

Figure 36.10. Baihui (GV-20)

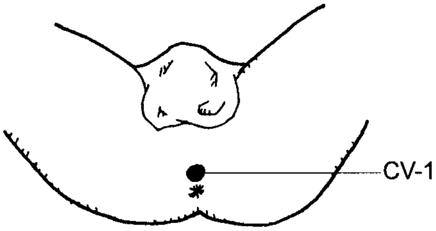

Figure 36.11. Hui Yin (CV-1)

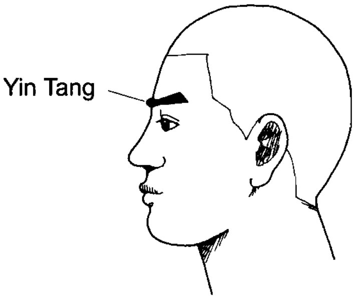

Figure 36.12. Yin Tang (Ex.Pt.)

CHAPTER 37
MEDICAL QIGONG MASSAGE

INTRODUCTION TO TISSUE FUNCTION

To understand the energetic function of surface and deep tissue stimulation through Medical Qigong Massage, it is important to first understand the reflexes and spontaneous energetic reactions within the body's tissues. The tissues of the human body are in a constant state of perpetual change. No matter where illness or damage has occurred, everything in the body is linked to the organs' and bowels' functions of producing and governing energy, and their interaction with the body's surface tissues (Figure 37.1–37.2).

In Traditional Chinese Medicine, the human body is viewed as a conglomeration of Qi, Blood, channels, tendons, fascia, bones, Marrow, bowels (Yang organs), and viscera (Yin organs). Each part is interdependent, each having inseparable Yin and Yang relationships and all of them interacting with each other. The following describes the interaction of: Qi and Blood, Tendons and Fascia, Bones and Marrow and the Bowels and Viscera.

1. Qi is considered the Blood's leader: as the Qi leads, the Blood follows. If the Qi ceases to flow, the Blood stops.
2. The tendons and fascia are used for channel energy movement. As the tendons and fascia stretch, they cause the channel Qi to increase its energetic movement and flow. When the tendons and fascia are calm, the channel Qi slows down.
3. The bones are the Marrow's residence. When the bones are hard, the Marrow is solid; when the bones are soft, the Marrow is hollow.
4. The bowels (Yang organs) are the viscera's (Yin organs') manifestations. If the bowels are strong, the organs become robust; when the bowels become weak, the viscera begin to fail.

Medical Qigong massage takes into consideration the function and flow of the body's tis-

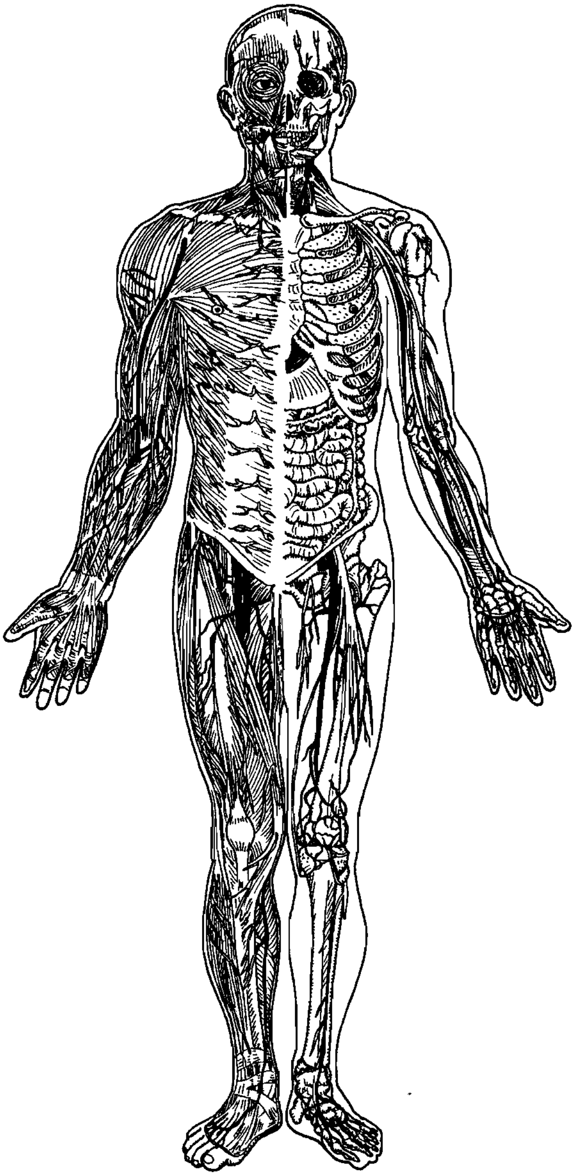

Figure 37.1. The Human Body (Anterior)

sues, in conjunction with the interacting systems of the body's energetic fields, striving to regulate both systems at the same time.

TWO APPROACHES IN TREATMENT

Medical Qigong massage can be divided into two schools of treatment modalities: Surface Tissue Massage Therapy and Visceral Tissue Massage Therapy.

1. Surface Tissue Massage Therapy utilizes the Five Elemental Qi Massage Therapy techniques in conjunction with external tissue stimulation. All external tissue manipulation techniques used in Qigong massage should be light as a feather and should not exceed the pressure one would place on an eyeball. When treating with Surface Tissue Massage, the doctor's focus must be centered on the dispersion of pathogenic Qi and the expansion of the patient's Wei Qi, while softly touching the patient's tissues.
2. Visceral Tissue Massage Therapy utilizes the Five Elemental Qi Massage Therapy techniques in conjunction with deep tissue stimulation. By matching the patient's pulse frequency and respiration with that of the doctor's, a stable, energetic resonance is secured and treatment begins. When treating with Visceral Tissue Massage, the doctor's focus must be centered deep into the patient's internal organs without touching the patient's external tissues.

After external tissue massage is used to treat the patient's tissues, it is important for the doctor to maintain focused intention on the patient's channels while slowly removing his or her hands. If a break in contact with the mind's intention is made, the patient's tissues will return to their pretreatment state. Once the doctor has physically disconnected from the patient in this manner, Distance Qi Emission and Qi manipulation techniques can be applied.

It is believed that the treatment of diseases should be carried out primarily by Qi emission as the main approach, and secondarily through Qigong massage. By combining both external and internal tissue manipulation skills with energy

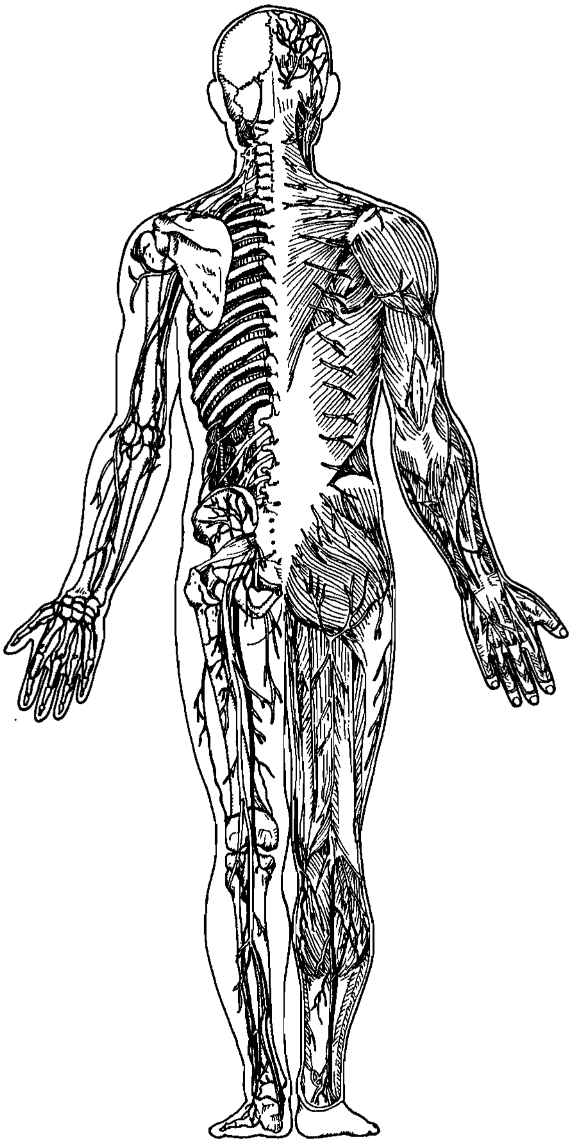

Figure 37.2. The Human Anatomy (Posterior)

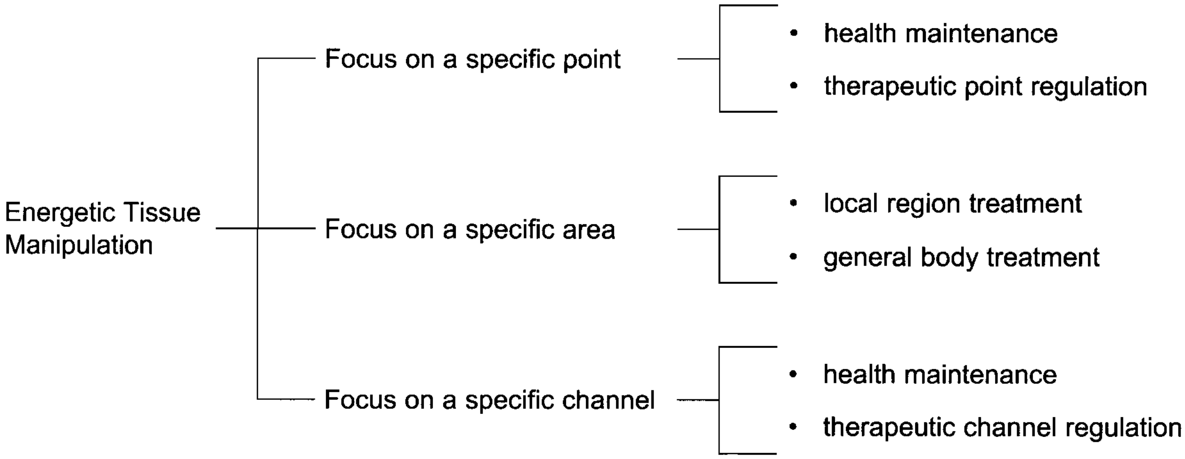
Energetic Tissue Manipulation
Focus on a specific point
health maintenance
therapeutic point regulation
Focus on a specific area
local region treatment
general body treatment
Focus on a specific channel
health maintenance
therapeutic channel regulation

Figure 37.3. Energetic Tissue Manipulation Chart

extension techniques, the Qigong doctor greatly improves the quality of treatments. In the clinical setting, this type of energetic tissue manipulation is utilized before or after energy extension. This facilitates the relaxation of the patient's muscles and joints and the opening or closing of their points. It also aids in dredging the channels, guiding the flow of energy, and in regulating Yin and Yang energy.

TISSUE REGULATION THROUGH MEDICAL QIGONG MASSAGE

The external treatment of the patient's tissues encompasses a wide variety of body manipulation and alignment methods to provide a complete treatment program (Figure 37.3). Used for toning the organs, glands, nervous system, and joints, Qigong massage is applied to treat soft tissue injury. The following are the six treatment goals of Qigong Massage:

- To relax the tendons, activate the channels, promote circulation of Qi and Blood, and relieve pain,
- To disperse Blood stagnation,
- To relieve muscle spasms,
- To expand the tendon sheaths,
- To treat fascia in order to move stagnations and separate adhesions, and
- To correct dislocations.

The following is a list of the five predominant external tissue manipulation techniques used by Qigong doctors in China today.

THE FIVE ELEMENTAL QI MASSAGE THERAPY

Traditionally there are five different techniques for Medical Qigong massage therapy. These five techniques are very important for reestablishing the energetic vitality of the patient's body. Each of these techniques relates to a specific action and organ associated with one of the Five Elements. The objective of these techniques is to influence the muscles, nerve fibers, and deep-lying tissues of the body in order to aid metabolism and stimulate energy flow (Figure 37.4). The applications of these five methods of energy manipulation and their affect on the body's tissues are described as follows.

THE CIRCLE-TWISTING METHOD

This technique relates to the Spleen, which rules and affects the muscles, and is implemented as a quick drilling motion into the body's surface tissues, using the fingers, palms, knuckles, or heel of the palm.

When treating a patient's tissue area, a circle-twisting movement will stimulate the Qi and Blood within the channel points. This will generate a numbing and tingling sensation that will

Circle-Twisting Method	Spleen-Earth Element: affects the muscles, stimulates and improves Blood circulation.
Compression Release Method	Kidneys-Water Element: affects the bones, and regulates the Wei and Ying Qi.
Thrusting Method	Heart-Fire Element: affects the Blood, pulse, and regulates and activates the Qi.
Grasping and Shaking Method	Liver-Wood Element: affects the tendons, ligaments and muscles, promotes, clears, and activates Qi flow, and balances the Yin and Yang energy.
Tapping Method	Lungs-Metal Element: affects the Qi and skin.

Figure 37.4. The Five Elemental Qi Massage Methods

cause the channel points to induce the energy to counterattack the foreign invasion of Turbid Qi and disease. The circle-twisting action will also stimulate and improve Blood circulation and is divided into two modalities of treatment:

- To stimulate, massage the area in a clockwise, spiraling, circular motion from the outside to the center of the point. The purpose of this action is to gather energy from the surrounding areas of the patient's body and collect it into the center point the doctor is treating.
- To sedate, massage the area in a counterclockwise, spiraling circle from the center of the point outward. The purpose of this action is to disperse the energy from the area the doctor is treating. This Excess energy will later be either moved into other organs, or dredged outside the patient's body through an adjacent channel (Figure 37.5).

THE COMPRESSION RELEASE METHOD

This technique relates to the Kidneys, which rule and affect the bones, and is performed by the doctor extending his or her intention into the patient, softly pressing the body's surface tissues, using the whole palm, parts of the fingers, or elbows. The Compression Release method is commonly used to purge or drain the tissues of pathogenic Qi.

The Compression Release techniques are generally used to regulate the Ying and Wei Qi, which travel along the channels and collaterals. Deep pressure reaches the Ying Qi, while shallow pressure stimulates the Wei Qi. Once damage has occurred in the body, the damaged channel points and organ regions cause certain changes in the Wei Qi, Ying, Qi, and Blood areas of the body, resulting in numbness, aches, pain, swelling, etc. Both Circle-Twisting and Compression Release techniques can eliminate these symptoms and create a more stable energy flow.

The Compression Release Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy. The doctor then initiates a rhythmic compression over the patient's tissues stimulating the patient's energetic field. This action has the function of relieving stagnation or congestion and can be used to dredge the

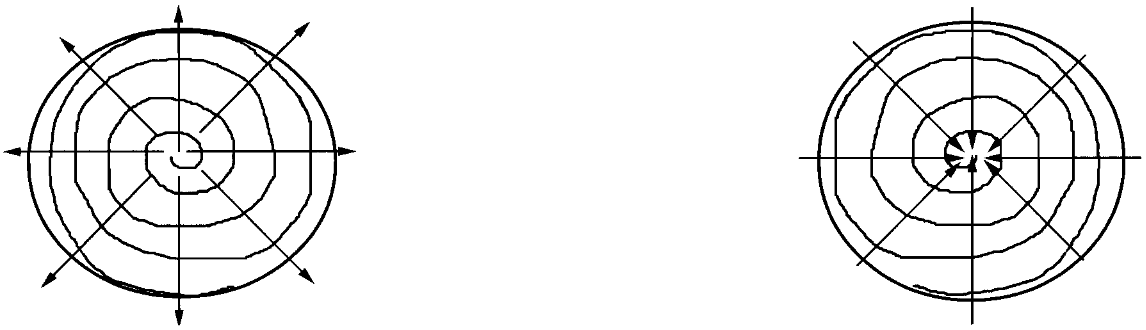

To sedate, purge, or disperse, move the energy in a counterclockwise direction.

To stimulate, tonify, or gather, move the energy in a clockwise direction.

Figure 37.5. Circle Twisting Method of Energy Therapy

channels and energetic points, as well as to relieve muscle spasms and physical pain.

The doctor must apply the correct amount of pressure for the patient's physical condition:

1. A slow Compression Release method is applied to any kind of hyperactive illness, nervous system disorder, mental disorder, or loss of Blood.
2. A medium Compression Release method is applied to illness for Spleen and Stomach disorders.
3. A fast Compression Release method is applied to hypoactive or Cold illnesses.

It is important to note that when applying the Compression Release method to tonify a specific area, the doctor should energetically press on the channel at a forty-five degree angle, slightly above where the channel flows into the injured area. This augments the flow of energy along the channel.

To disperse or drain a particular area, the doctor should apply slight pressure at an angle past where the channel flows into the injured area. This downward pressure will cause the energy to flow away from the injury. This application is used when treating Excess conditions.

THE THRUSTING METHOD

This technique relates to the Heart, which rules and affects the Blood and Blood pulses. It is utilized to gently push the tissues, extending the Qi and Blood along the patient's surface channels, using the thumb, ball of the thumb, or heel of the palm. The thrusting method can draw energy into a Deficient area or move Qi out of an Excess area.

The Thrusting Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy. The doctor then rubs lightly or lightly touches above the patient's tissues while simultaneously emitting Qi.

The technique should be synchronized with the frequency of the patient's pulse and respiration. The Thrusting technique has the function of regulating and activating the patient's Qi, alleviating pain, and dredging the channels.

THE GRASPING AND SHAKING METHOD

This technique relates to the Liver, which rules and affects the tendons. It is administered by lightly vibrating the skin (gently at first, then slowly increasing the intensity). The Grasping and Shaking method sends a wave of vibration throughout the patient's body, and is used to increase Qi and Blood flow to the tissue area.

The Grasping and Shaking Technique

Once the proper hand posture and area to be treated has been selected, the doctor guides Qi into his or her palm and fingers using Hot, Cold, vibrating, or spiraling energy.

The doctor then lightly grasps and shakes the energetic field within the patient's tissues, while extending energy. The Grasping and Shaking technique is used for directing Qi to flow with or against the natural currents in the body's channels. It promotes, clears, and activates the channels and flow of Qi, relaxes the muscles and tendons, and balances the Yin and Yang energy.

THE TAPPING METHOD

This technique relates to the Lungs, which rule and affect the Qi and skin. This technique is administered by lightly and rhythmically tapping the patient's skin (with varying degrees of force), using the fingers, palms or fists (Figure 37.6). The Tapping method sends a pulsating ripple throughout the patient's tissues and is used to disperse stagnation from a point, channel, or organ.

YIN AND YANG ENERGETIC POINT MANIPULATION

Before treating patients with Five Elemental Regulation therapy, the Qigong doctor should be aware of the principals of Yin and Yang energetic polarity involved within the Energetic Point Manipulation. This method of treatment is used to regulate the patient's Wei Qi, Ying Qi, and Qi and Blood systems (as a whole), in order to form a governed systemic balance. In the Energetic Point Manipulation methods, the treating techniques are directly based on Yin and Yang theory.

THE YIN THEORY OF QI MANIPULATION

This technique is expressed as a passive, quiet, or motionless action (more energy and less physical movement). The energetically passive Yin movements are used to push, fill, or tonify the patient's Righteous Qi, located within the patient's tissues. It is used for treating external tissue conditions and for internal organ conditions.

- If the external tissue area is in an inactive or Yin state, apply a Yin technique. Treat a Cold area with soft, passive manipulations, while extending the mind superficially onto the patient's surface tissues to fill or tonify anti-pathogenic factors, and to revitalize the Deficient area.
- If the internal organ area is either in an active or Yang (Hot or hard) state, the doctor must apply a Yin, passive, or slow action technique, while extending his mind deep into the patient's tissues and organs.

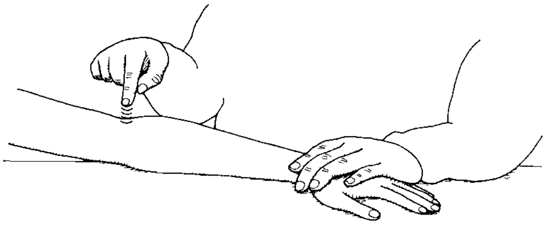

Figure 37.6. In the Massage Tapping Method, the Qigong doctor will extend his or her intention deep into the patient's tissues in order to stimulate the channel point.

THE YANG THEORY OF QI MANIPULATION

This technique is expressed as an active technique, in motion. The active energy of the Yang movements are generally used to pull or purge pathogenic factors. It is used for treating external tissue conditions and for internal organ conditions.

- If the external tissue area is in an active or Yang state, apply a Yang technique (treating an External Hot or hard area with fast dredging manipulations) while extending the mind superficially onto the patient's surface tissues. This is done to pull or purge pathogenic factors, or break up the stagnation.
- If the internal organ area is either in an inactive or Yin (Cold or flaccid) state, the doctor must apply a Yang, quick, or active technique, while extending his or her mind deep into the patient's tissues and organs.

This combination of movement (Yang) and non-movement (Yin) is embodied in both the Circle-Twisting and Compression Release Methods of the Five Elemental Regulation Techniques.

APPLICATION FOR YIN AND YANG TECHNIQUES

The following text gives examples of specific Yin or Yang hand techniques, and when to apply them:

1. The Qigong doctor uses Yin hand techniques for tonifying, filling and pushing Qi and

Blood through the patient's organs, tissues, and channels. The Yin hand actions are demonstrated through slow or inactive hand movements.

2. The Qigong doctor uses Yang hand techniques for sedating, purging, and pulling Qi and Blood. The Yang hand actions are demonstrated through fast or active hand movements.

YIN AND YANG TREATING METHODS

The following text gives examples of specific Yin or Yang hand treatment methods, and when to apply them:

1. When a patient's tissues are in an Excess Yang state (Hard or Hot), the Qigong doctor sedates the area with counterclockwise rotations, leading the Excess energy from inside the tissue area out of the body.
2. When a patient's tissues are in a Deficient Yin state (Flaccid or Cold), the Qigong doctor tonifies the area with a clockwise rotation, to bring energy into the Deficient tissue area.

TREATING THE TISSUE AREA OR ENERGETIC POINT

In treating the patient's surface tissue area or specific energetic points, some Qigong schools teach that with a male patient, one should begin with points on the left side, then on the right side; whereas for female patients, one begins working on the points on the right side of the body, then on the left.

DEEP AND SHALLOW PRESSURE

Energy compression (deep or shallow) is controlled by the doctor's release. A deep compression relates to the Ying Qi, while a shallow compression pertains to the Wei Qi. This Compression Release technique affects the body's Blood and vital energy harmony. It seeks to balance the body's energy by preventing the Qi from flowing in the wrong direction. This technique also leads the pathogenic Qi outside the patient's body and is commonly called "The Dispersing Evil Qi Technique."

If stagnant energy remains in the center of a channel point, then the doctor must employ a Compression Release method. The compression causes the channel point to contract, leading it toward a calm, stable state. When the doctor releases the compression on the channel point, it expands and becomes stimulated into an active state.

THE GRADUAL ENERGETIC PENETRATION OF TISSUES

When treating a patient with External Tissue Regulation, there should always be a gradual penetration of energy permeating the patient's tissues. The Qigong doctor first sinks his or her focused intention onto the patient's surface tissues. The intention is then slowly sunk deeper, first into the patient's muscles and inner fascia, then into the bones or internal viscera.

At no time should pressure damage the tissues or compound pain. Treatment or pressure time at each area or energetic point should not be excessive. To establish the correct amount of pressure, determine the normal energetic pulsation in accordance with the Five Elemental Qi Massage methods, then decide if the technique is to be light, medium, or heavy.

ADDITIONAL GUIDELINES

The following are specific guidelines of Qigong massage treatment methods, and when to apply them:

1. Before beginning Qigong massage, the patient must be comfortable and relaxed. Otherwise the patient's body will not be receptive to the tissue stimulation or Qi emission.
2. When first touching the patient, the touch should be slow, as if the doctor is dipping his or her hand to test the water. The movements should have an established rhythm.
3. The doctor should never hold his or her arm, or the patient's arm, rigid or fully extended; the elbows should always be bent to some degree.
4. A symptom or damaged area occurring on one area or point at the extremity of a channel may be effectively treated by stimulating the point at the opposite end of the channel (see Magic Square, Chapter 31).
5. Points which are painful with light pressure

are in a state of Yang and need to be purged and sedated.

6. Points which are painful to deep pressure are in a state of Yin and need stimulation.
7. When an organ is functioning normally, it should receive weak, short stimuli that will increase the function, action, and energetic potential of the organ.
8. Cool skin over a point or channel indicates a Deficiency in that channel and should be tonified.
9. Hot skin indicates Excess Qi accumulation. It also indicates a current or pending illness in an organ, or area associated with that channel, and should be purged immediately.

SECTION IX

MEDICAL QIGONG EXERCISES AND PRESCRIPTIONS

CHAPTER 38

INTRODUCTION TO MEDICAL QIGONG PRESCRIPTIONS AND REGULATION EXERCISES

HISTORY OF MEDICAL QIGONG PRESCRIPTIONS

Many of the earliest known Medical Qigong prescriptions were derived from the movements of animals. Dating back to the second century B.C., the Qigong Classic (Dao Yin Tu) illustrates, in manuscripts written on silk, over 45 Qigong postures with descriptions of the movements as well as the names of the diseases which they treat. Over half of these illustrated postures are animal movements.

In the ancient Chinese text called the *Spring and Autumn Annals*, it is written, "Flowing water never stagnates, and the hinges of an active door never rusts. This is due to movement. The same principle applies to the body's Jing and Qi. If the body does not move, Jing does not flow. When Jing does not flow, Qi stagnates." The soft flowing movements in the Medical Qigong prescription exercises enhance and increase the flow of the body's Qi and Blood.

THE FUNCTION OF MEDICAL QIGONG PRESCRIPTIONS

When a disease develops, certain active points of the disease correspond to different organs, muscles and limbs of the body, and a pathological reflex is formed. Trigger and reflex points which are painful to touch are perfect examples of this phenomenon. The cerebral cortex can be inhibited in specific ways by regulating the Mind and entering a state of tranquility through Medical Qigong prescriptions, thus overriding the pathological reflex pattern, allowing the body time to heal.

Although acupuncture and herbs can, in addition to relieving pain, move Qi, Blood, and Phlegm (the roots of the disease), they fail to address the patient's life-style which is often a major contributing factor. Problems such as stagnation, Excess or Deficient conditions are often exacerbated by a patient's sedentary life style. A responsible doctor of Medical Qigong therapy addresses this particular element by prescribing physical therapy, exercise, and meditations as part of the overall treatment strategy.

Tonify	Deficient Conditions	Color Visualizations
Purge	Excess Conditions	Sound Resonations
Regulate	Yin and Yang	Qigong Massage

Figure 38.1. Medical Qigong Treatments

The standing rule for treatment in Medical Qigong prescriptions is to:

- Tonify the Deficient organs and organ systems with color visualization, by using the Mind's intention to focus on moving Qi into a Deficient organ, strengthening, thus illuminating its tissues;
- Purge the Excess organs and organ systems with sound resonation, by using the Mind's intention to focus on moving the Qi out from an Excess organ, resonating, and draining its tissues;
- Regulate the body's Yin and Yang organs and organ systems with Qigong massage, by using the Mind's intention to balance the energetic fields by moving the Qi up and down, right and left, and inside and outside of the tissues (Figure 38.1).

The therapeutic effects of Medical Qigong meditations and prescriptions are designed to move Qi and Blood stagnation, to tonify the internal organs and organ systems, and to enhance the autonomic nervous system. These prescriptions stimulate the patient's body, thereby gradually eliminating fatigue, as the normal body functions are restored. Another important aspect of the Medical Qigong meditations and prescriptions is the enhancement of the body's immune system functions.

MEDICAL QIGONG AND THE IMMUNE SYSTEM

The many health benefits of Medical Qigong include increased white blood cell production, stabilized blood pressure, and an improved immune function. The immune system, unlike the vascular circulation system, defies a clear anatomical description. It can only be understood through its functions and interrelationships with the mind and various other parts and systems of the body, such as the lymph nodes, Spleen, tonsils, appendix, Peyer's patches (in the ilium), thymus gland and the bone marrow.

The lymph nodes contain specialized compartments - some contain B cells (B stands for Blood), some T cells (T stands for Thymus), and some contain macrophages. The tonsils likewise contain B and T cells. The appendix and Peyer's patches are components of the lymphatic drainage system, and serve as locations where the B cells mature and where antibodies are produced by the B cells. The body's T cells mature in the thymus gland.

Both B and T cells begin their life in the Liver, when the fetus is only nine weeks old. They then migrate to the bone marrow where they form precursor cells. Afterwards, the T cells migrate to the thymus and the B cells remain in the bone marrow.

The immune system utilizes two major strategies for protecting the body from an infectious disease: The Humoral Immunity and the Cell Mediated Immunity.

- In humoral immunity the B cells play the primary role, and the T cells as well as other related cells play supporting roles.
- In cell mediated immunity the T cells play the main role, while B cells and other related cells play auxiliary roles.

Both T cells and B cells are called "Lymphocytes." They are two types of white blood cells that originate in the body's bone marrow and travel in the blood and lymph circulatory system (Figure 38.2). Both T cells and B cells are concentrated in the lymph nodes, which are found in clusters near groups of veins, and are heavily concentrated under the arm pits, in the adenoids, the Spleen and in the groin area.

When B cells encounter an antigen or foreign substance, they differentiate into plasma cells. These plasma cells secrete antibodies that fight and neutralize bacteria, or coat the antigen to mark it for destruction by other components of the immune system (such as the T cells).

T cells specialize in destroying cells infected by viruses, in attacking tumors, and in rejecting transplanted organs and tissues. The T cell, after leaving the bone marrow, travels to the thymus gland where it undergoes further development before returning to the lymph and blood circulatory systems. The T cells respond to an antigen by differentiating into several kinds of T cells, three of which are of primary importance: Killer T cells (also called cytotoxic T cells), Helper T cells and Suppressor T cells. Killer T cells bind to viral antigens displayed on the surface of infected cells and destroy them. Helper T cells bind to the antigens on the surface of a B cell that has already bound itself to the antigen. Each Helper T cell then releases Lymphokines (one well know class of lymphokines is interferon), that enhance the maturation of the B cells, or signal them to differentiate and multiply into plasma cells that secrete antibodies. The Suppressor T cells function to dampen the immune response and to signal the "all clear."

A third group of important immune cells are the phagocytes. These cells are the body's natural scavengers. They engulf antigens and destroy them. Those that specialize in cleaning up the debris are called macrophages. The macrophages can also signal B or T cells to the presence of foreign material.

Another important group of cells are called Null cells, among which the N.K. or "natural killer" cells are found. These cells are able to identify tumors and virally infected cells, and kill them without having any prior encounter with such an antigen.

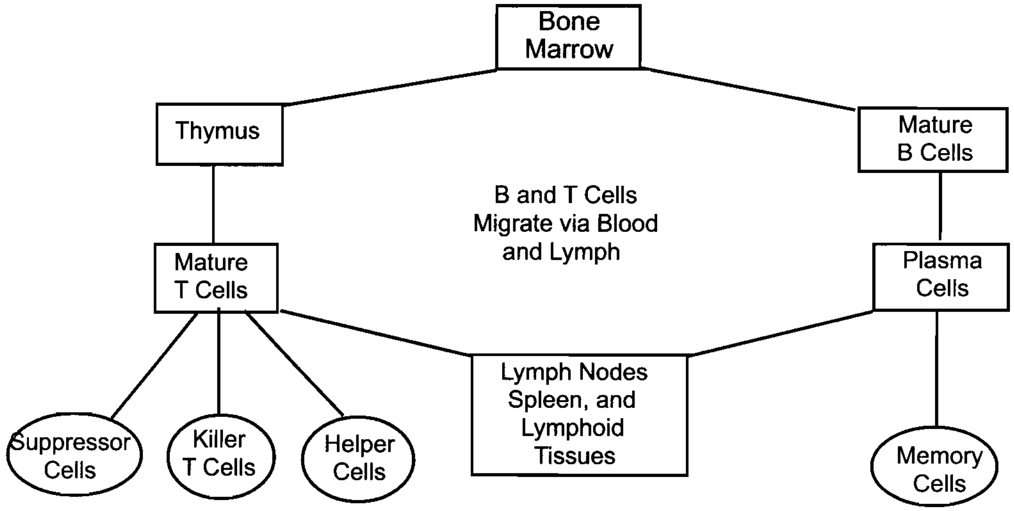

Figure 38.2. The Body's Immune System

A final important group consists of Mast cells. They contain powerful chemicals: heparin - an anticoagulant, serotonin - a neurotransmitter, and histamine - a biochemical released in overabundant quantities in allergic reactions.

These white blood cells live in the blood and lymph fluid, and are commonly known as the leukocytes. There are two main varieties of white corpuscles: The Polymorphonuclear Leucocyte (Polymorphs) and Lymphocyte.

- Polymorphs are found in the bone marrow stem cells; they destroy invading bacteria and remove dead or damaged tissue.
- Lymphocytes are found in the lymph glands, Spleen and thymus gland and circulate in the blood along with the polymorphs. They are the smallest of the leucocytes, and designated the lymphocyte because it is identical with the lymphoid cell derived from the lymphatic glands.

Both the white blood cells called polymorphs and lymphocytes move into the body's surrounding tissues as a part of the immune function. The purpose of these cells is to work together to defend the body from invasion of pathogenic factors. Their function is to identify, ingest, destroy and eliminate disease-causing organisms, as well as to remove any and all dead, damaged, or irregular cells.

Immune cells are produced in the lymphoid tissue and red bone marrow, and are stored in the lymphatic organs and blood. In order to perform their function of protecting and cleaning, these cells must circulate throughout the entire body. The capillaries and lymphatic vessels carry the lymph fluid through the lymph nodes and the lymphatic organs of the thymus gland, spleen, and tonsils, where the lymph fluid is filtered, cleansed of impurities, and infused with new lymph cells.

In order for an external pathogen to affect the body, it must first pass through the protective layers of cells. Generally, the macrophages (specialized white blood cells) present within the tissues ingest the pathogens and dead tissues. This action of the macrophages is reinforced by the action and movement of the Wei Qi. When a pathogenic intruder invades the body, the macrophages are the first to attack the intruder. If the macrophages are not strong enough, more blood and fluid, which contain more immune cells, are brought to the area. The increased fluid engorges the tissue, and brings more cellular activity. The increased

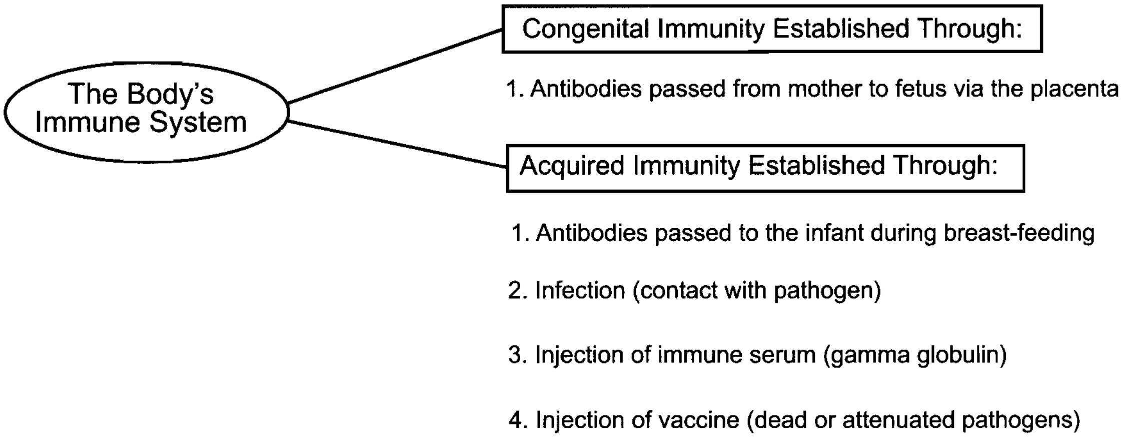
The Body's Immune System
Congenital Immunity Established Through:
1. Antibodies passed from mother to fetus via the placenta
Acquired Immunity Established Through:
1. Antibodies passed to the infant during breast-feeding
2. Infection (contact with pathogen)
3. Injection of immune serum (gamma globulin)
4. Injection of vaccine (dead or attenuated pathogens)

Figure 38.3. Congenital and Acquired Immunity

blood and cellular metabolism increases the local temperature. The immune cells try to break down and ingest the pathogen, creating a process which continues until the disease is eliminated.

Once the battle is won, some of the Killer T cells and Helper T cells will remain circulating in the blood and lymph circulatory system as Memory cells that will be able to deal more efficiently with any future antigen of the same kind. Likewise, some of the mature B cells will also become Memory cells that hover around lymph nodes which they use as a base of operation.

Because Qigong movements generate heat and soften the body's connective tissues, this encourages the flushing of the intercellular fluid throughout the body's tissues and organs. When combined with Breathing, Mental, and Postural Qigong exercises, the entire body is affected. These energetic Qigong actions increase oxygenation of the blood and improve circulation, thereby enhancing lymphocyte production and strengthening the body's immune system.

CONGENITAL AND ACQUIRED IMMUNITY

The body has two ways of establishing or maintaining its immune functions: Congenital Immunity and Acquired Immunity (Figure 38.3).

- Congenital Immunity (Prenatal) is established during gestation, through antibodies that are passed from the mother to the fetus via the placenta.
- Acquired Immunity (Postnatal) is established in the following ways: immediately after birth during breast-feeding (through the colostrum-mother's first milk), through contracting an infection (contact with external pathogens), the injection of an immune serum (gamma globulin), and the injection of vaccines (dead or attenuated pathogens).

Due to the importance of establishing a strong immune system within the body's Prenatal Jing, Traditional Chinese Medicine strongly endorses fetal education, instruction on diet and nutrition, maintaining emotional well-being during pregnancy, and Qigong exercises specifically designed to enhance the child's Prenatal Essence, Energy, and Spirit.

OVERVIEW

One specific goal of Medical Qigong treatments, exercises and meditations is to activate and strengthen the immune function. Qigong therapy can be used to not only strengthen the metabolism of the organs and cells, but also increase cellular energy, thus making the immune system stronger.

In China, even though the lymphatic function itself was unknown and unnamed, its effects were

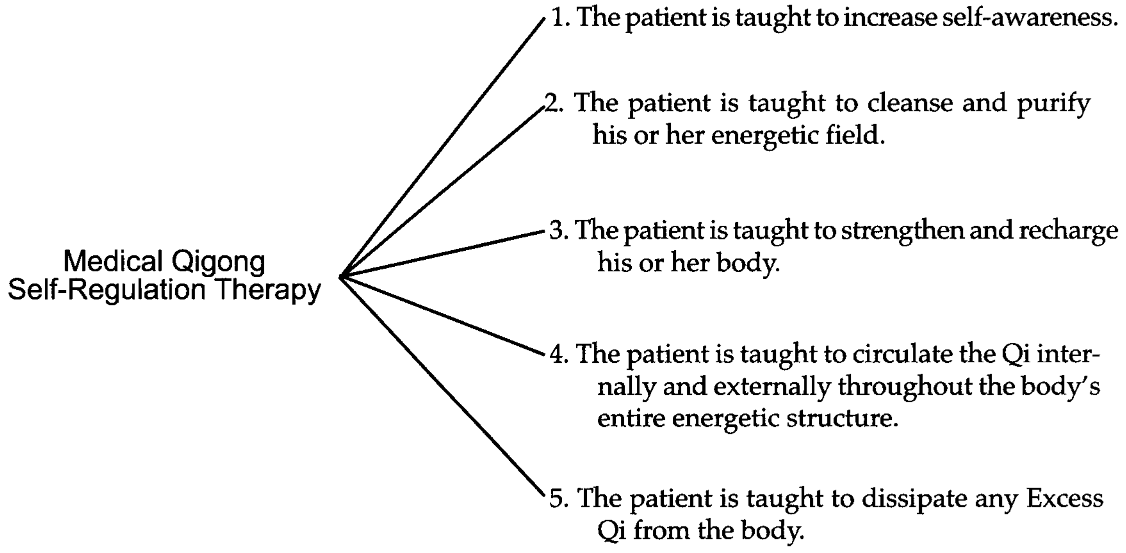
Medical Qigong
Self-Regulation Therapy
1. The patient is taught to increase self-awareness.
2. The patient is taught to cleanse and purify his or her energetic field.
3. The patient is taught to strengthen and recharge his or her body.
4. The patient is taught to circulate the Qi internally and externally throughout the body's entire energetic structure.
5. The patient is taught to dissipate any Excess Qi from the body.

Figure 38.4. The Five Stages of Healing Transitions That Occur Through Self-Regulation Therapy

generally ascribed to the proper action of the Qi and Body Fluids. Therefore, elaborate methods for generating and circulating lymph were developed, and faithfully practiced through Qigong. Breath, movement and postural modification as well as meditations have specific effects on the circulation of the lymph and therefore enhance the overall function of the lymphatic system.

THE FIVE STAGES OF HEALING

When practicing Medical Qigong exercises and meditations, there are five basic stages of healing transitions which occur through self-regulation therapy (Figure 38.4).

1. The patient is taught to increase his or her awareness of the body (physically, mentally, and emotionally) and its current condition.
2. The patient is taught to cleanse and purify the body's energetic fields to rid it of Qi stagnations and toxic pathogenic factors.
3. The patient is taught to strengthen and recharge his or her body to replenish the Qi.
4. The patient is taught to circulate the Qi internally and externally throughout the body's entire energetic structure, thus moving any stagnant Qi and strengthening the body.
5. The patient is taught to dissipate any Excess Qi from the body by way of self-massage (lightly brushing the skin to release stagnant Qi).

THE THREE CATEGORIES OF MEDICAL QIGONG PRESCRIPTIONS

The Medical Qigong prescriptions are divided into the same three categories (or goals) used in Medical Qigong therapy: tonification, purgation and regulation. These three goals are achieved through both Dynamic and Quiescent Qigong exercises, meditations, prayers, and mantras. These prescriptions enable the patient to reinforce the treatments performed by the Qigong doctor, and prevent the patient's energy from reverting back to old destructive patterns.

The Qigong doctor selects the proper Qigong prescription to fit the patient's illness (to Tonify, Purge, or to Regulate) and formulates a rational program that suits the patient's case. The patient should never be allowed to practice at random, otherwise pathological changes or Qi deviations may occur.

The following are but a few examples of Qigong prescriptions, which are interchangeable, as there are certain variations of purging exercises that can be modified and used for tonification and visa versa.

QIGONG PURGING EXERCISES

Qigong Purging exercises are practiced by the patient to detoxify pathogenic Qi. Some examples are already presented in this textbook are:

- Channel Purging exercise (Chapter 15)
- Cleansing the Body meditation (Chapter 12)
- Dry Crying meditation (Chapter 53)
- Energy Melting meditation (Chapter 18)
- Internal Organ Sound Therapy exercises (Chapter 4 and Chapter 40)
- Old Man and Tide Pool meditation (Chapter 40)
- Pulling Out The Pain meditation (Chapter 28)
- Releasing Emotional Blockages and Energetic Armoring meditation (Chapter 18)
- Sun and Moon meditation (Chapter 18)
- Tumor and Cancer Tone Resonating exercises (see Chapter 40)

QIGONG TONIFICATION EXERCISES

Qigong Tonification exercises are practiced by the patient to strengthen and stabilize the internal organs and organ systems. Some examples are already presented in this textbook are:

- Channel Point meditations (Chapter 36)
- Color Visualization meditations (Chapter 4)
- Cultivating Yuan Qi meditation (Chapter 50)
- Daoist Five Yin and Yang Organ exercises (Chapter 42)
- External Qi Cultivation meditations (Chapter 41)
- Gaining Wisdom meditation (Chapter 49)
- Gathering Energy from the Five Planets meditations (Chapter 11)
- Gathering Moon Cream meditation (Chapter 11)
- Gathering Sun's Essence meditation (Chapter 11)
- Gathering Qi from Mountains, Valleys, and Deserts (Chapter 12)
- Gathering Qi from The Four Directions (Chapter 12)
- Gathering Qi from Trees, Bushes, and Flowers (Chapter 12)
- Improving the Intellectual Facilities meditation (Chapter 49)
- Internal Qi Cultivation meditations (Chapter 41)
- Nei Guan - Inner Viewing Meditation (Chapter 25)
- One Through Ten meditation (Chapter 26)
- Opening and Closing the Three Dantians exercise (Chapter 41)
- Organ Massage exercises (Chapter 39)
- Qigong Walking exercises (Chapter 15)
- Seeing Auras meditation (Chapter 25)
- Soul Retrieval meditations (Chapter 19)
- Tian Wu Zang meditation (Chapter 11)

QIGONG REGULATION EXERCISES

Qigong Regulation exercises are practiced by the patient to balance their Yin and Yang, as well as their internal organ energy. Regulation exercises usually follow Tonification or Purgation exercises, because the goal is to reestablish balance. Some examples already presented in this textbook are:

- Exchange of Fire and Water meditation (Chapter 18)
- Gathering Qi in the Upper and Lower Dantians (Chapter 41)
- Microcosmic Orbit meditations Chapter 41)
- Opening and Closing the Three Burners (Chapter 41)
- Opening and Closing the Three Dantians exercise (Chapter 15)
- Pulling Down the Heavens (Chapter 41)
- Qigong Self-Healing massage (Chapter 41)
- Taiji Ruler exercises (Chapter 18)
- Wu Zang meditation (Chapter 28)

CHOOSING PRESCRIPTIONS ACCORDING TO THE PATIENT'S QI CIRCULATION

When prescribing Medical Qigong prescriptions, the priority is to get the patient to move and circulate the Qi and Blood superficially and deeply within the body. According to Chinese physician, Professor Yan De-xin, the motion and quality of Blood flow in young and middle-aged patients is different from that in the elderly, which should be kept in mind when prescribing exercises.

- When prescribing Medical Qigong therapy for young patients, use mostly Dynamic and some Quiescent Qigong regulations.
- When prescribing Medical Qigong exercises for both adolescent and middle aged patients, it is important to increase Yin Qi cultivation while maintaining and storing the patient's Yang Qi.
- When prescribing Medical Qigong for senior citizens, focus is placed on Quiescent Qigong, with a minor emphasis on Dynamic breathing exercises, depending on their age and constitution.
- When prescribing Medical Qigong therapy for patients with very weak health or constitution, the doctor should prescribe mostly Quiescent Qigong regulations to start with.
- When prescribing Medical Qigong therapy for senior citizens with a strong constitution and health, the doctor can prescribe Qigong regulations that increase Yin Qi cultivation while maintaining and storing the patient's Yang Qi.

MEDICAL QIGONG PRESCRIPTIONS FOR GENERAL SYNDROMES

The following are some examples of the proper application of Dynamic and Quiescent Qigong prescriptions as presented by Dr. He Si Hai of the Zhe Jiang Qigong Hospital in China:

1. For Excess and Heat Syndromes it is recommended to prescribe static (quiescent) Medical Qigong exercises. After the gradual decline of the patient's fever or disease, the Qigong exercises are gradually increased in numbers and duration, and the purging training method is added (provided they have not excessively consumed their constitution).
2. For Deficiency and Cold Syndromes it is recommended to prescribe Dynamic Medical Qigong exercises. Gradually increase the exercises to include methods of lifting the limbs. If however, the pathological state is more serious, then static exercises are prescribed first.
3. For Cerebral Hemorrhage, have the patient practice static exercises such as the Hibernation Breathing exercise (quiescent lying posture, see Chapter 15) while combining it with the Yongquan (Kd-1) purging exercise (once the patient has regained consciousness). Dynamic Medical Qigong exercises that use methods of passing, circulating, or distributing Qi through the head are forbidden.
4. For Cerebral Embolism combine the training methods to include exercises that accelerate Blood flow and strengthen the function of the limbs as well as quiescent tonifying methods.
5. For Malignant Tumors (in a state of non-diffusional transference, in situ), combine an effective Dynamic exercise with a Quiescent exercise. Some examples of non-aggressive tumors/cancers are considered energetically warm and in a non-aggressive state of transition are:

- Prostate cancer
- Cervical dysplasia
- Endometrial adenocarcinoma (uterine)
- Basal cell skin cancer
- Squamous cell skin cancer

6. For Malignant Tumors (in a state of diffusion, i.e., metastatic) prescribe Quiescent Medical Qigong exercises which are similar to the Hibernation Breathing exercise, in combination with strengthening exercises for patients with chronic diseases. Dynamic Medical Qigong exercises which accelerate the Blood flow are forbidden. Some examples of tumors that are considered energetically Hot and in an aggressive state of transition are:

- Pancreatic cancer
- Leukemia
- Cervical cancer
- Clear cell carcinoma (uterine)
- Melanoma (skin cancer)

NUMBER OF BREATHS REQUIRED IN MEDICAL QIGONG PRESCRIPTIONS

Each Medical Qigong treatment will generally last up to three days. The Medical Qigong prescriptions that the doctors gives their patients will extend the treatment's effects indefinitely, providing the patients do all their homework.

The number of breaths required to Regulate the patient's Yin and Yang energy in a Medical Qigong prescription is often calculated in respiratory cycles, and on the basis of the number of the Yin and Yang aspects of diseases. This is often determined by the Yin and Yang energy numbers. Yang energy is represented by all the odd numbers (1, 3, 5, 7, 9, etc.). Yin energy is represented by all the even numbers (2, 4, 6, 8, etc.). These numbers are related to the Five Elements (Figure 38.5). For example: 6 is the number of the Kidneys, 7 the Heart, 8 the Liver, 9 the Lungs, and 10 the Spleen (Figure 38.6). The cycles of Yin and Yang numbers can be implemented through the utilization of the Magic Square (see Chapter 31).

The Yang number is used to calculate the respiratory cycles for Tonification and Purgation of the Yang channels; likewise, the Yin number is used for the Yin channels. This calculation should only be used as a general reference guide in treatment. Clinically, respiratory cycles should be determined based on the location and nature of the disease (if it is Yin or Yang, Exterior or Interior, Deficient or Excessive). The primary goal for using respiratory calculations is to ensure that sufficient Qi is administered when applying Medical Qigong treatments and therapeutic prescriptions. A disease due to a Deficiency of Yin, for example, should be treated with a tonification method for 18 or 64 respiratory cycles. To treat a Deficiency of Yang, a tonification method is used by the patient for 24 to 36 breaths.

	General Treatment	Moderate Treatment	Strong Treatment
Yin	6	18 or 36	64
Yang	9	27 or 49	81

Figure 38.5. Number of Yin and Yang Respirations

5 Yin Organs	Kidneys	Heart	Liver	Lungs	Spleen
Number of Respirations	6	7	8	9	10
Element	Water	Fire	Wood	Metal	Earth

Figure 38.6. Yin and Yang Five Element Respiration

CHAPTER 39
SPECIFIC ORGAN SELF-MASSAGE PRESCRIPTIONS

INTRODUCTION

The next few chapters of this book include some of China's most popular clinical therapies. It is the author's sincere desire to educate the reader as to the diversity, yet effectiveness of these prescriptions. It is important to remember that it is not necessarily the exercise or prescription's history that matters, what does matter is the initial goal of the prescription (whether to purge, tonify, or regulate) and the correct posture, breath, and mental focus.

This section begins with energy massage techniques, which are prescriptions that require the patient to regenerate his or her own body energy for self-healing, cultivation, and circulation. These exercises are designed to purge, tonify and regulate the Five Yin Organ System. They are not arranged in any particular order, and can be singled out by the doctor and given to patients for specific ailments. Each massage prescription is named after the organ it stimulates. For clinical therapy, perform these exercises three to five times a day.

HEART MASSAGE

The doctor can chose to give the patient either the One Through Four meditation (1-fun, 2-shoe, 3-tree, 4-core; from Chapter 26), or the Stacking the Bones meditation (described below) to prepare for the following meditations.

1. Begin in a Wuji posture to perform the Stacking of the Bones meditation. Imagine the energy in your body sinking and melting into the ground; the bones vertically stack, one by one, on top of each other. Each of the body's articulations stacks on top of the next section of bone, from the bottom of the feet, ankles, knees, etc., to the top of the head.

   The spine is kept straight but relaxed, the anal sphincter is closed, the tongue is touching the top of the upper palate, the Mingmen is slightly pressing towards the back, and the body is completely relaxed.
2. Perform the Heart Regulation. When prescribing these exercises for male patients, have them place the left hand over the Heart with the right hand on top of the left hand (for women, the hand positions are opposite, and the beginning placement of the hands is below the left breast). Massage in twelve circular rotations clockwise to the left, then twelve rotations counterclockwise. The hand can either lightly touch the skin (using Qi massage), or be slightly off the skin, up to several inches away from the patient's body (using Qi Distance therapy), depending on the degree of energetic sensitivity and projection skill (Figure 39.1). Focus the mind's intention on the Heart organ, allowing the energy within the Heart area to flow and circulate with the movement of the hands.

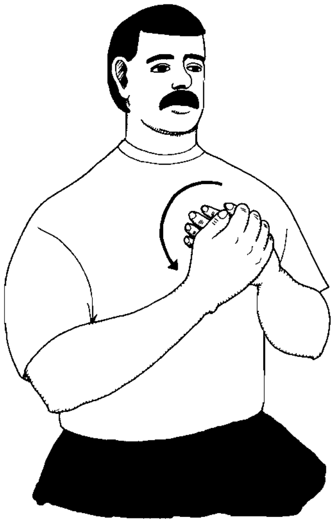
Figure 39.1. The patient uses the mind's intent to regulate the Qi of the Heart.

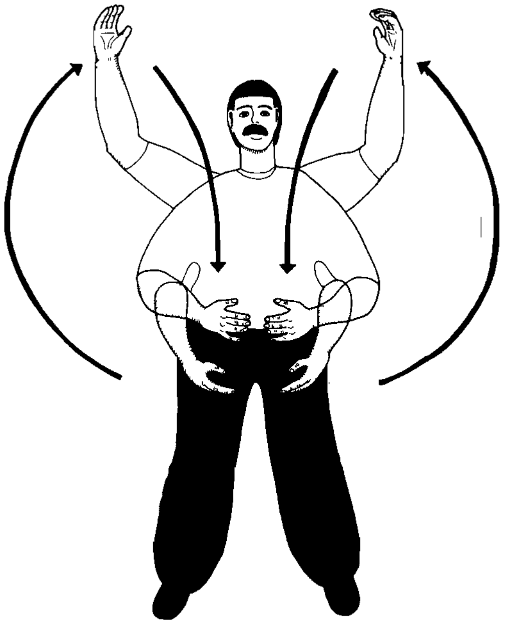

Figure 39.2. Pulling Down the Heavens allows the patient to sink the Qi, returning the excess energy back into the Lower Dantian.

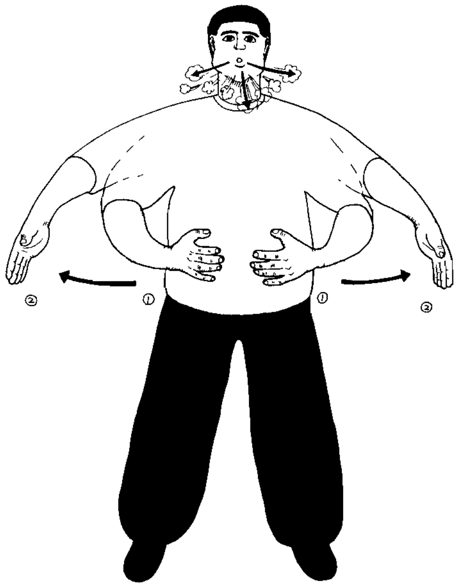

Figure 39.3. Exhaling and Opening the Lower Dantian allows the body to expel the Evil Qi from the internal organs. Expand the abdomen.

The purpose of the rotating action of the Heart's Qi is for promoting Blood circulation and dispersing Blood Stasis, while dredging the channels and collaterals.

3. Perform the Point Respiration. This technique focuses breathing and imagination into the channel points for the specific treatment area. For this particular exercise, the patient should leave the hands over the Heart, while concentrating the focused intention and breath deep into the tissue of the organ. The patient should then exhale while lightly squeezing and pressing the area of the Heart, and imagining divine healing light radiating outward from the tissues. While inhaling, the patient should lift the hands away from the body and imagine divine healing light flowing into the Heart organ area. This action is similar to that of a bellows. This exercise should be performed twelve times, then lead the Qi down into the Lower Dantian.
4. Perform the Pulling Down the Heavens method. Inhale, pause, then exhale while slowly bringing the arms down the center line of the torso, while quietly drawing the energy back down into the Lower Dantian. When lowering the arms, imagine the energy of the Heart like water, melting and flowing down the torso.

   Pause for a breath at the Lower Dantian, rooting the energy. Then slowly raise the hands above the head, while inhaling. Exhale slowly and imagine the energy in the arms and head melting down the body, beginning at the top of the head and tips of the fingers, flowing down and into the Lower Dantian.

   This action is important, because if the Qi is not brought back into the Lower Dantian after each exercise, the energy will be easily dispersed (Figure 39.2).
5. Perform the Opening and Closing of the Lower Dantian. Finally exhale and imagine the Evil Qi being expelled through the mouth. As you exhale, separate both of the hands away from the abdomen (as you contract the abdomen). This is called the Opening (Figure 39.3).

   After completing the Opening of the Lower

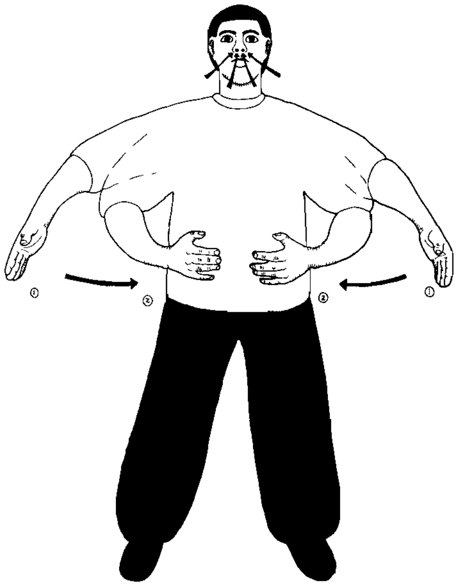

Figure 39.4. Inhaling and Closing the Lower Dantian allows the body to seal itself from outside pathogenic factors. Contract the abdomen.

Dantian, inhale through the nose while bringing both hands back to face the Lower Dantian (as you expand the abdomen). This action is called the Closing. When Closing, imagine sealing the body in a protective energetic field, so that the pathogenic factors cannot get inside (Figure 39.4). Perform the Opening and Closing method three times.

Note: The techniques 1 through 5 are one complete set. Perform a total of three sets. Three sets are necessary because the body will naturally progress through three stages. The first time the body is introduced to a new pattern it will shock the system. The second time the body begins to understand the pattern. The third and final time the body recognizes and accepts the changes the patient is attempting to reprogram.

LIVER MASSAGE

1. Begin from a Wuji posture and perform either the One Through Four or Stacking the Bones method as described in Heart Massage (1).

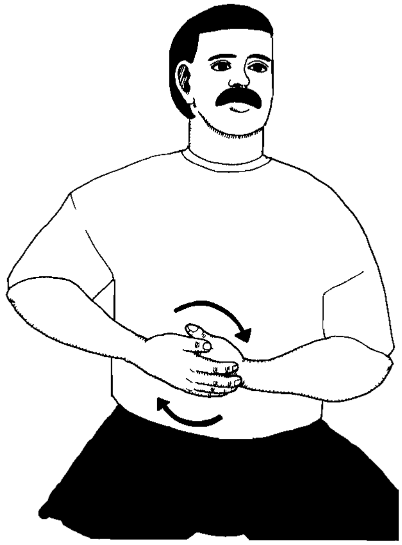

Figure 39.5. The patient will focus their mind's intent on the Liver area while massaging, allowing their energy to circulate into the organ's tissues.

2. Perform the Liver Regulation. Place the left hand on the Liver with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, focus the mind on the Liver. Inhale and imagine divine healing light coming down from the Heavens and filling the Liver. Exhale and purge the Liver of Turbid Qi through the mouth.

   The purpose of this exercise is to smooth and regulate the activities of the vital energy flowing inside the Liver, encouraging the hyperactive and Rebellious Liver Qi to flow downward (Figure 39.5).
3. Perform Point Respiration with both hands still over the Liver, for twelve breaths, as described in the Heart Massage (3).
4. Perform Pulling Down the Heavens by leading Qi down into the Lower Dantian, as described in the Heart Massage (4).
5. End with Opening and Closing method as described in the Heart Massage (5).

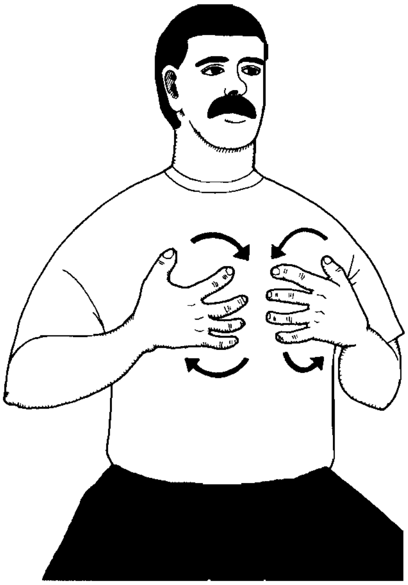

Figure 39.6. The patient focuses the mind's intent on the Lung area while massaging, allowing the energy to circulate into the organ's tissues.

Note: The techniques 1 through 5 form one complete set. Perform two more sets for a total of three to complete the prescription.

LUNG MASSAGE

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Lung Regulation. Place both palms on the Lungs (specifically on an injured or diseased region, if known). Massage in twelve circular rotations from the bottom, up the center to the top and then to the outside of the chest. Then massage twelve rotations in the opposite direction, from the top, down the center and to the outside. When massaging, concentrate on the Lung region. Inhale and imagine divine healing light coming down from the Heavens and filling the Lungs. Exhale and purge the Lungs of Turbid Qi.

   The purpose of this exercise is to bring Blood and Qi into the Lung area, clean and purify the inhaled breath and support the Lungs' Qi flowing downward (Figure 39.6). TB patients with pulmonary holes should concentrate on abundantly filling and expanding the vital energy in the Lungs while visualizing the Qi activating the body's cells to close the holes. If the patient is unable to visualize this, or the reaction is too uncomfortable, have the patient mentally pull the Qi downward.
3. Perform the Point Respiration with both hands still over the Lungs, for twelve breaths, as described in the Heart Massage (3).
4. Perform Pulling Down the Heavens method, leading the Qi back down into the Lower Dantian as described in the Heart Massage (4).
5. End with Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 5 are one complete set. Perform two more sets for a total of three to complete the prescription.

SPLEEN & STOMACH MASSAGE

SPLEEN MASSAGE

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Spleen Regulation. Place the left hand on the Spleen with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, concentrate the mind on the energy within the Spleen area, imagining the Spleen Qi movement flowing with the movements of the hands (Figure 39.7). Inhale and imagine divine healing light coming down from the Heavens and filling the Spleen. Exhale and imagine the Spleen absorbing the divine Qi.

   The purpose of this exercise is to strengthen and increase the Spleen's function of transporting and converting Qi. This is an excellent massage for patients with splenomegaly (enlargement of the spleen) and dyspepsia (painful digestion).
3. Perform the Point Respiration with both hands still over the Spleen, for twelve breaths,

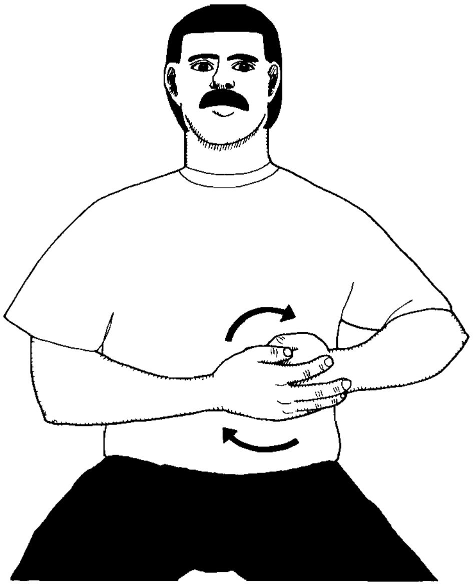

Figure 39.7. The patient focuses the mind's intent on the Spleen area while massaging, allowing the energy to circulate into the organ's tissues.

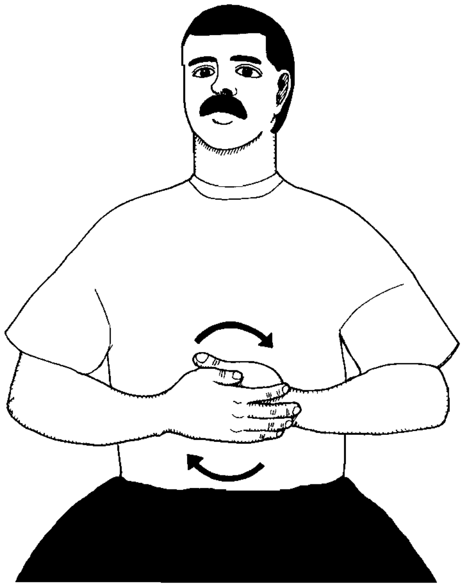

Figure 39.8. The patient will focus their mind's intent on the Stomach area while massaging, allowing their energy to circulate into the organ's tissues.

as described in the Heart Massage (3). Then move both hands to the Stomach area.

STOMACH MASSAGE

4. Perform the Stomach Regulation. Place the left hand on the Zhongwan CV-12 point at the middle of the Stomach with the right hand on top (opposite for women). Massage in twelve circular rotations to the left in a clockwise direction, then twelve rotations in the opposite direction. When massaging, concentrate on warming the Stomach area (Figure 39.8). Inhale and imagine divine healing light coming down from the Heavens and filling the Stomach. Exhale and purge the Stomach of Turbid Qi.

   This is an excellent massage for patients with thoracic depression, abdominal distention, gastritis, and gastric ulcers.
5. Perform the Point Respiration with both hands still over the Stomach, for twelve breaths, as described in the Heart Massage (3).
6. Perform the Pulling Down the Heavens method to draw the Qi down into the Lower Dantian, as described in the Heart Massage (4).
7. End with the Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 7 are one complete set. Perform two more sets for a total of three to complete the prescription.

KIDNEY MASSAGE

This exercise is excellent for tonifying the Kidneys, as well as healing impotence, premature ejaculation, lumbago, lower back problems, and pelvic inflammatory disease (for PID use reverse breathing with static Qigong training).

This is an extremely important exercise for patients and for Qigong doctors to practice; it enables them to increase their Qi projection when treating patients.

1. Begin from a Wuji Posture and perform either the One Through Four or Stacking the Bones method as described in the Heart Massage (1).
2. Perform the Kidney Regulation. Close the eyes and place both hands on the back (waist

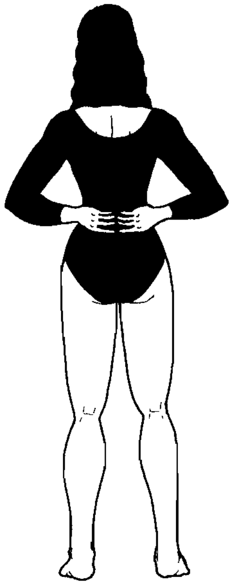

Figure 39.9. The patient will focus the mind's intent on the Kidney area while massaging, allowing the energy to circulate into and warm the organ's tissues.

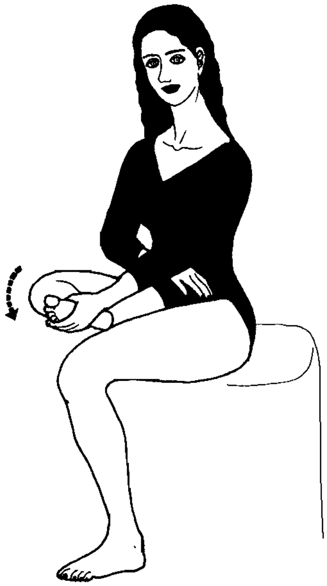

Figure 39.10. The patient will focus the mind's intent on the bottom of the foot while massaging, allowing the energy to flow through the Kidney Channel into the Lower Dantian.

level), middle fingers touching the Shen Shu Bl-23 points. Focus on the Kidneys. Begin massaging twenty-four times inward, then twenty-four times outward, until the area becomes warm (Figure 39.9).

3. Perform the Point Respiration with both hands still over the Shen Shu Bl-23 points for thirty-six breaths. Exhale as you press the palms into the Kidneys, inhale as you raise the palms from the back. While inhaling, imagine divine healing light coming down from the Heavens and filling the Kidneys completely. While exhaling, imagine the Kidneys absorbing the divine Qi.
4. Perform the Pulling Down the Heavens method, drawing Qi into the Lower Dantian as described in the Heart Massage (4).
5. Perform the Opening and Closing method as described in the Heart Massage (5).

Note: The techniques 1 through 5 are one complete set. Perform two more sets for a total of three to complete the prescription, then proceed to the Closure.

6. Perform the Closure. After completing three sets of the Kidney exercise, place the right hand on the navel and the left hand on the Mingmen. Focus on gathering Heat and Qi in the Lower Dantian. Next, place the right foot on the left knee (opposite for women). With the right hand still on the Lower Dantian, place the left hand at the bottom of the right foot and gently massage the Yongquan Kd-1 point. Massage this area 100 times in a clockwise direction, then 100 times counterclockwise. When massaging the Kd-1 point, work slowly, focusing the mind's intention through the center of the palm into the center of the foot, imagining the Qi flowing up the leg into the Lower Dantian (Figure 39.10).

Note: Some Medical Qigong Clinics in China perform this same exercise while slapping the bottom of the feet (instead of using circular massage). They begin this treatment with 100 strikes on each foot for the first day, increasing an additional 100 strikes each consecutive day (not exceeding 300 strikes) until the desired result is achieved.

7. Perform the Point Respiration for twelve breaths, with the right hand still on the Lower Dantian and the left hand still on the bottom of the right foot, as described in the Heart Massage (3).
8. Switch feet, placing the left hand on the Lower Dantian while the right hand gently massages the Kd-1 at the bottom of the left foot, 100 times clockwise then 100 times counterclockwise.
9. Perform the Point Respiration for twelve breaths, with the left hand still on the Lower Dantian and the right hand still on the bottom of the left foot.
10. End with the Opening and Closing method as described in the Heart Massage (5).

Note: Do not use this ending method (5 and 10) for patients who suffer from insomnia, as it will have an awakening effect as it invigorates the vital energy. When performed at bedtime it can be used for treating insomnia by following the simple prescription below.

TREATMENT FOR INSOMNIA

This is a very famous prescription for insomnia. I have used it in my clinical practise for many years and have found it to be extremely effective. If the insomnia is due to a Deficiency of Kidneys' Yin, or the patient is restless during the night due to the Liver's Heat rising and causing the patient's mind to wander, this prescription will help. After clinically treating patients for insomnia, the Qigong doctor can prescribe the following exercise to be performed by the patient.

This exercise is to be practiced just before the patient is going to sleep. Begin by sitting at the edge of the bed, repeating the previous Kidney exercises 2, 3, and 6.

1. Perform the Kidney Regulation. Close the eyes and place both hands on the back (waist level), middle fingers touching the Shen Shu Bl-23 points. Focus on the Kidneys. Begin massaging twenty-four times inward, then twenty-four times outward, until the area becomes warm (Figure 39.9).
2 Perform the Point Respiration with both hands still over the Shen Shu Bl-23 points for thirty-six breaths. Exhale as you press the palms into the Kidneys, inhale as you raise the palms from the back. While inhaling, imagine divine healing light coming down from the Heavens and filling the Kidneys completely. While exhaling, imagine the Kidneys absorbing the divine Qi.
3. Perform the Closure. After completing three sets of the Kidney exercise, place the right hand on the navel and the left hand on the Mingmen. Focus on gathering Heat and Qi in the Lower Dantian. Next, place the right foot on the left knee (opposite for women). With the right hand still on the Lower Dantian, place the left hand at the bottom of the right foot and gently massage the Yongquan Kd-1 point. Massage this area 100 times in a clockwise direction, then 100 times counterclockwise. When massaging the Kd-1 point, work slowly, focusing the mind's intention from the Lower Dantian through the leg into the center of the foot, imagining the Qi flowing down the leg into the bottom of the foot (Figure 39.10). As the mind begins to relax, close the eyes and fall sleep. If you become fatigued, do not complete the exercise.

An explanation for the reason for this exercises success is as follows:

1. The Kidney exercise 1 and 2 will root the body's Blood, Heat, and Heart Qi back into the Kidneys and Lower Dantian, where they will become stabilized.
2. The Kidney exercise 3 will draw the body's Heat, Blood, and the Heart Qi away from the torso and Lower Dantian into the extremities of the feet, quieting the mind and causing sleep.

CHAPTER 40
HEALING SOUND THERAPY AND PRESCRIPTIONS

INTRODUCTION AND HISTORY

As early as the Qin Dynasty (221-207 BC) healing sounds have been recorded in China. For thousands of years, Qigong masters have studied occurrences in nature and in the physical body to develop a holistic approach to health and longevity. One of the many discoveries made is that all humans produce similar sound patterns in certain situations. For instance, after a tense situation many people utter a sigh of relief. This sigh is the body's way of releasing emotional stress. Specific situations will release different emotional sounds, because emotions are stored as tone resonances in the different chambers of the Five Yin Organs.

During the Sui Dynasty (590-618 AD), ancient Qigong Master Zhi Zhuan pointed out that there are six ways of expelling toxic energy from the internal organs, using the six different sounds formula. The silently spoken six sounds are coordinated with the breathing and the movements of the body's extremities to purge and cleanse the organ cavities and particular channels of stagnant Qi. These exercises are excellent for treating muscular injuries and for the Yin and Yang organs. The impure Qi produced by external pathogenic factors (from food, air, and drink) and internal pathogenic factors (toxic emotions and negative mental attitudes) will be expelled directly from each organ.

The healing sound techniques can either be used by the patient as a form of static Medical Qigong prescription (for purgation), or projected into the patient by the doctor as a clinical form of energetic stimulation (for tonification or purgation). The six words are: Xu (Shu), Ke (Haa), Hu (Who), Si (Sss), Chui (Chree), and Xi (Shee). These sounds are also used in chanting (in spiritual Qigong). Ancient documents state that these six sounds denote the Five Yin Organs as well as the five seasons.

It is important to note that there are several healing sound systems used by different Medical Qigong schools and other cultures. These sound systems all affect the body's sympathetic and parasympathetic nerves to some degree. The main emphasis, when performing any healing sound, should be placed on the connection of the mind, breath, and imagination to the area or internal organ focused on by the patient. This will reduce the potential for causing Qi deviations. The patient must have an internal image and actually hear and feel the resonating sound inside the body. It is also important that the patient feel enveloped in the sound, vibration, energy, and light of the therapeutic tone.

SOUND VOLUME RESONATION

The pronunciations of the sounds must be correct. The resonant effects of the sounds must be brought into full play to be successfully transmitted to the diseased areas of the internal organs. Therefore, it is extremely important that the patient focus the mind and intention on the area being treated. There are three volume resonations commonly used in Medical Qigong sound therapy: soft, moderate, and loud.

1. Soft volume is spiritual in its energetic nature and is related to the projection of Shen. A soft volume is generally used when beginning sound therapy treatments. With soft volume pronunciation, the patient can easily feel where the sound waves reach and acquire resonant effect, but the amplitude of the sound wave is not wide enough to produce the desired healing effects, and its vibrational force is very weak.
2. A moderate volume of sound is emotional and mental in its energetic nature and is related to the projection of Qi. A moderate volume is

considered the standard for all healing sound treatments. Moderate volume is decided by the individual's own voice. It should be limited to the range in which the patient's voice does not strain and in which the voice can be raised or lowered effortlessly.

3. Loud Volume is related to the projection of Jing. A loud volume is used to shock the patient into feeling the diseased tissue area. This is due to the fact that loud volume is more tangible (or physical) in its energetic nature. With loud volume pronunciation, the patient can easily feel the body's internal organs and tissues vibrate; however, this sound is only used to reintroduce a desensitized patient to his or her body. The problem with prolonged use of loud volume is that the increased volume is liable to create tension and cause Qi deviations.

The volume of sound uttered also depends on the patient's current state of health. People who are physically weak can usually practice the breath regulating and Qi reinforcing exercises, but should not practice the sound pronouncing method until they feel appropriately energetic.

To practice the word pronouncing method, the patient should start from the soft volume and increase it gradually. Diseases of different parts of the body require different volumes. For example:

- Patients with Heart or Lung trouble can comfortably take the soft to moderate volume pronunciation;
- Patients with Liver, Spleen, Stomach, Kidney and intestinal diseases can take the moderate to louder volume pronunciation.

HIGH-, MIDDLE-, AND LOW-PITCHED SOUND

The healing sounds have high, middle, and low pitches with different penetrating potentials. When correctly vocalized, they cause resonant vibrations in the body, producing certain states of consciousness that have both a psychological and a physiological effect on the nervous system.

As the tones change from a high to low pitch, they regulate the depth of the tone penetration. Similar to a dimmer switch on a light bulb, the tone adjusts the depth of the vibration resonance in the body.

The three pitch sounds also affect the energy movement of the body's Triple Burners, as well as the internal organs positioned within the thorax.

1. The high-pitched sound is pronounced in a straight tone with the head slightly drooped. The high pitch focuses the energy resonation into the specific organs of the patient's Upper Burner from the sternum to the manubrium. After completing the sound resonation exercise, return back to the Wuji position and relax; regulate the body using natural breathing.
2. The middle-pitched sound is pronounced in either a rising or descending tone, depending on the patient's specific conditions. The rising tone will cause the patient's Qi to ascend, while the descending tone will cause the Qi to drop. The middle pitch focuses the energy resonation into the specific organs of the patient's Middle Burner from the navel to the sternum within patient's body. Middle-pitched sounds are only prescribed for specific conditions (organ prolapse or organ energetic overload).
3. The low-pitched sound is pronounced in a dropping, then rising tone, within one breath. This causes the energetic resonation to saturate the diseased organ from top to bottom, then from bottom to top. This focuses the energy resonation into the specific organs of the patient's Lower Burner from the navel to the perineum. The method of pronouncing the low-pitched sound is usually practiced immediately after sounding the high-pitched tone.

After pronouncing either the high-, middle- or low-pitched sounds, perform the Open and Close the Three Dantians method once, then close with the Pulling Down the Heavens exercise. This completes one set.

BREATH AND MIND CONTROL

Healing sounds produce an aspirating effect analogous to vacuuming a room. This analogy describes to the patient the desired intention needed to successfully utilize this modality of healing. Toning without proper focused intention, breath, and mind control placed on the specific internal organ is like trying to vacuum a room while waving the hose in the air. Although the hose is sucking the air, it is not performing the desired result of purging the room of dust and dirt. To become proficient and establish the desired result, the individual must focus the attention and place the vacuum hose onto the specific area that needs cleaning.

Healing sound and toning therapy work the same way. It is not just sounding the tone which allows the technique to work, but the combination of the breath and mind focused on the intended internal organ. When practicing sound therapy, the patient generally uses the Natural Breathing method, along with focused intention and color visualization.

While inhaling, the patient imagines breathing in divine, white light healing energy through the nose and absorbing this healing Qi into the diseased area, vibrating it.

When exhaling, the patient imagines breathing out the dark, diseased, Turbid Qi through the mouth. It is important for patients to maintain the healing color (i.e.: green, red, golden yellow, white, or blue) vibration inside the diseased organ, even when they exhale.

SEPARATING THE BREATH INTO YIN AND YANG

The breath can be divided into either Yin or Yang qualities. When breathing shallow and focusing the mind on the front of the throat, the breath feels cool and is therefore considered Yin; however, when shifting the mind's attention towards the back of the throat, the breath, when released, feels hot, and is considered Yang. This change in temperature is due to the front of the body being Yin in nature and the back of the body being Yang.

AUDIBLE AND INAUDIBLE SOUND RESONATION

Before projecting sound therapy in a clinical setting, the Qigong doctor must vibrate the three Dantians while focusing on his or her Taiji Pole. Once the sound vibration has filled the doctor's body, the resonate healing sounds can be projected into the patient's tissues and organs.

Another way to produce sound is to imagine that the sound is appearing as a specific color and causing a vibration within the body. Clinical sound therapy is further divided into three distinct levels of emitted energies: Obvious Sound Resonation, Hidden Sound Resonation and Mysterious Sound Resonation.

1. The Obvious (Ming) Sound Resonation is audible sound. This energy sound resonation is considered Yang and is produced through the vocal organs. When toning audibly, the healing sounds fills the Qigong doctor's energetic fields. These fields can then be transferred onto the patient to produce a very strong, physical affect on the patient's tissues and organs. The Qigong doctor connects (through intention) with the patient's internal organ, and imagines his or her voice penetrating the diseased organ's tissues. The doctor then imagines the patient's tissues vibrating like rippling water with the sound exhalation, to purge the pathogenic Qi.
2. The Hidden (An) Sound Resonation is slightly audible sound. This energy sound is considered Yin transitioning into Yang, and is produced by softly exhaling sound energy resonation into the patient's body. This slightly audible energy sound resonation affects the patient's tissues and organs more on an emotional plane, enveloping and penetrating the patient's tissues like a mist.
3. The Mysterious (Hua) Sound Resonation is inaudible sound. This energy sound resonation is considered Yin and is mentally produced inside of the doctor's body. The Qigong doctor first focuses and fills his or her body and energetic field with the vibrant sound. The doctor then focuses the mind on the healing sound like a mantra. This inner sound fills the Qigong doctor's field with resonant vibration which can be transferred into the patient's body. This energy sound resonation affects the patient's tissues and organs on a more spiritual dimension, enveloping and purging like a gentle light.

The Qigong doctor can also coordinate the projection of the healing sounds with the patient's toning to synchronize their energetic fields. All sound techniques have specific postures, breathing, and mind regulations to facilitate whole body integration.

Patients and students are usually encouraged to practice the audible sound vibrations until they can vibrate the tissues of the organs at will, before they proceed to the inaudible sound resonation.

THE SIX HEALING SOUND PRESCRIPTIONS

Before starting any organ sound exercise, there are certain grounding and cleansing techniques that the patient should practice. These preparation exercises are described as follows.

1. Always begin sound resonation from a Wuji standing posture.
2. Breathe naturally from the abdomen, inhaling through the nose and exhaling through the mouth. The body should be relaxed, and respiration should be gentle, thin, even, and extended. The mind should not concentrate on the expansion of the Lungs, but on the Lower Dantian first, and then on the treatment area. Also, breath through every pore of the body into the Lower Dantian.

   Upon exhaling, sound the healing word with a low volume pronunciation, feeling where the sound waves reach. Once the tissue area is stimulated, increase the volume. At the same time draw in the abdomen, draw up the anus, slightly knead the ground with the feet, sink the shoulders, and then shift the body weight to both heels. This will access the Kidneys' Water Qi. When inhaling, close the mouth, raise the tongue against the hard palate, and shift the body weight to the anterior soles of the feet. This will access the body's Fire energy. Exhale the specific sounds through the mouth while imagining the impurities of that specific organ being expelled.
3. Pull Down the Heavens (the ending breath regulation). Inhale and exhale through the nose. When inhaling, slowly raise both arms forward and upward from the sides of the body, with both palms facing downward until reaching shoulder level. Then turn the palms up, and continue to circle the arms over the head. When exhaling, bend the elbows to make a curve in front of the chest, and press the abdomen. Then separate the arms to the sides of the legs, and return to the Wuji posture (Figure 40.1).

   Regulate the breath three times in the Wuji posture after completing the exercise. Breathing Regulation (Pulling Down the Heavens) is required after the exercise of each sound is completed and before the next exercise begins.

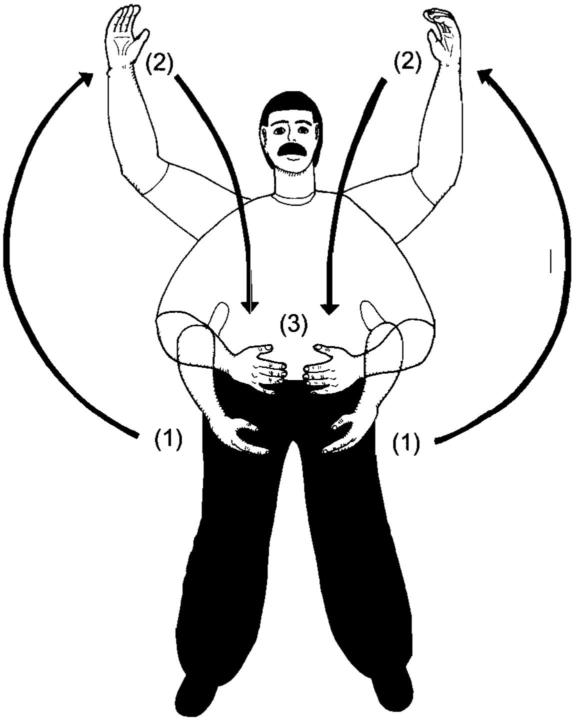

Figure 40.1. Pulling Down the Heavens allows the patient to sink the Qi, returning the excess energy back into the Lower Dantian.

THE SOUND "XU" (SHU) NOURISHES THE LIVER

This sound relieves Liver stagnation, aids in correcting certain types of visual distortions, and purges the accumulation of Phlegm in the body (e.g., the condition of Lung Hot Phlegm, due to an already existing cold).

Both hands overlap the inner Laogong Pc-8 points, with the left hand on the outside for men (opposite hand position for women). The Shaoshang Lu-11 point of both thumbs presses against the navel, with the Laogong points pointing at the umbilicus.

When inhaling, expand the abdomen and close the eyes, feeling the breath envelop and penetrate the Liver organ. At the same time direct the Qi of the Liver's Channels to flow upward from the Dadun Lv-1 point on the lateral portions of the big toes. Visualize the Qi rising through the lower abdomen, linking the Gall Bladder to the Lungs, traveling to the throat, and eyes, to the forehead, and then to the Baihui GV-20 point.

When exhaling, contract the abdomen and release the pathogenic factors. Focus the attention and imagination on the Qi flowing down the face, to connect with the Lungs' Channels, flowing downward to the Shaoshang Lu-11 points on the medial portions of the thumbs. The Qi is then directed and rooted into the Lower Dantian. Open the eyes wide, and sound the word "Shu" (the lips are slightly brought together, the tongue tip stretches forward with the lateral sides slightly curling toward the middle). Exhale completely.

In this particular exercise, the breath should feel lukewarm as you exhale (Figure 40.2). Perform six times.

THE SOUND "KE" (HA) REPLENISHES THE HEART

This sound relieves Fire from the Heart and expels Heat from the body; it also improves poor circulation and reduces fevers. The respiration patterns are the same as those in the previous (Liver) sound prescription.

When inhaling, raise both hands to shoulder level as if embracing a post. Both palms should face each other with the arms making a curve in front of the chest. Expand the abdomen and close the eyes; feel the breath envelop and penetrate the Heart organ. At the same time direct the Qi of the Spleen's Channels to flow from the Yinbai Sp-1 point on the medial portion of the big toes, along the inner portions of the legs, upward to the abdomen to meet with the Thrusting Channels, and enter into the Heart.

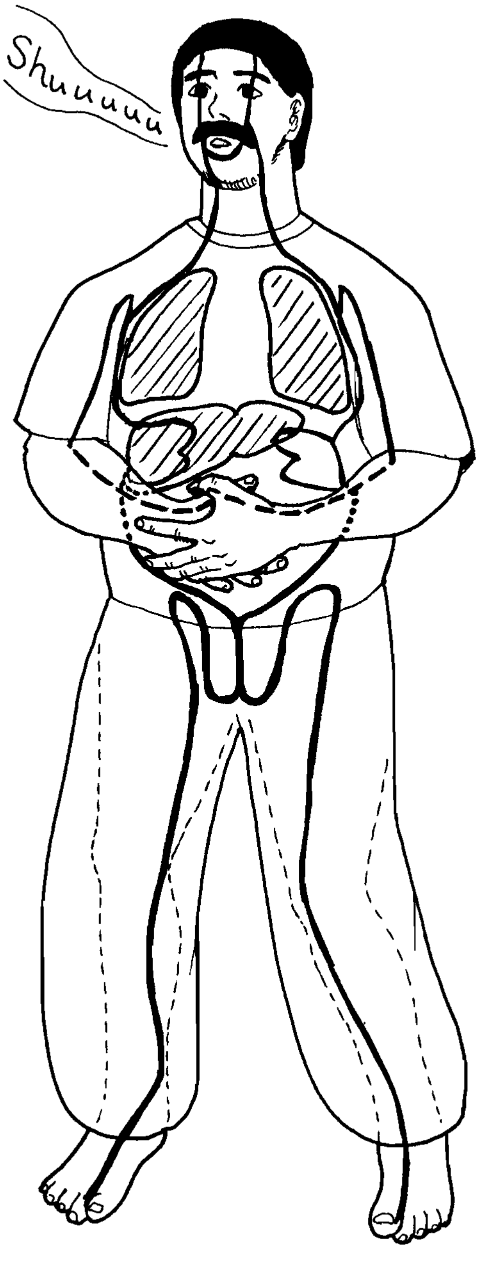

Figure 40.2. The sound "Xu" (Shu) is used to purge Liver stagnation as well as to disperse the accumulation of Phlegm in the body (Lung Hot Phlegm).

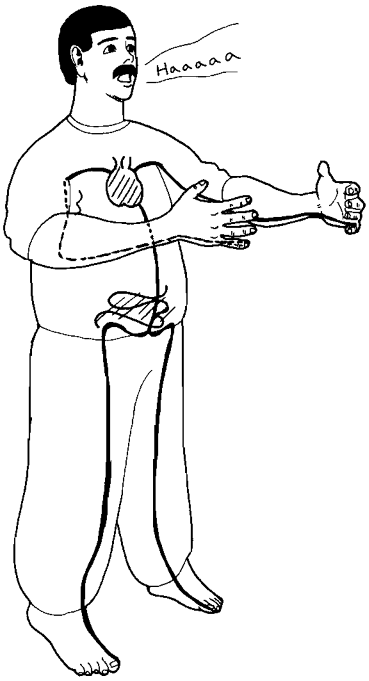

Figure 40.3. The sound "Ke" (Ha) is used to purge Heart Fire as well as expel Heat from the body.

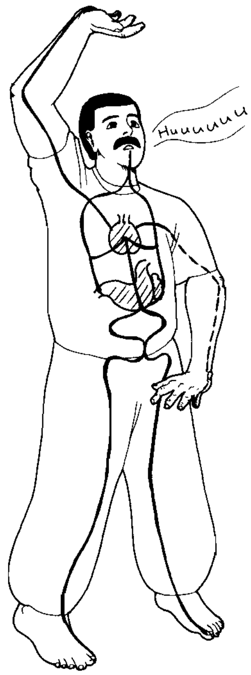

Figure 40.4. The sound "Hu" (Who) is used to treat digestive problems, by purging Stomach and Intestinal stagnation.

When exhaling, softly open the eyes and draw the Qi out from the Heart, past the Lungs, through the armpits, along the inner portions of the arms, to the Shaochong Ht-9 point on the medial portion of the little fingers. Point the fingers forward to release the pathogenic Qi, and sound the word "Ha." The mouth is half closed with the tongue placed against the lower mandible. Exhale completely. This breath should feel hot (Figure 40.3).

Return to the beginning position and perform six times.

THE SOUND "HU" (WHO) STRENGTHENS THE SPLEEN

This sound is used to correct digestive problems, especially when the Stomach or intestines feel full or sluggish.

When inhaling, raise both hands upward in front of the abdomen, as if embracing a ball in the Lower Dantian area. Both palms should face each other with the arms making a curve. Expand the abdomen and close the eyes; feel the breath envelop and penetrate the Spleen organ. Direct the Qi of the Spleen's Channels to flow from the Yinbai Sp-1 points on the medial portion of the big toes, up to the abdomen to enter the Spleen, and to link up with the Stomach.

When exhaling, the right palm turns outward, rolling over and pushing above the head, thus forming the posture of Holding the Heavens. At the same time, imagine the Qi ascending up to the throat, to the root of the tongue, and spreading under it. Imagine a secondary river of Qi flowing from the Stomach to the Heart, and ascending along the Heart's Channels to the Shaochong Ht-9 point on the medial portion of the little fingers, and exhale the sound "Who." The lips are brought together as if playing a flute, the tongue is level and stretched forward as much as possible to allow the ascending energy of the Thrusting Channels to rush out. Exhale completely while the left palm turns downward and presses down to the outside of the left hip (Figure 40.4).

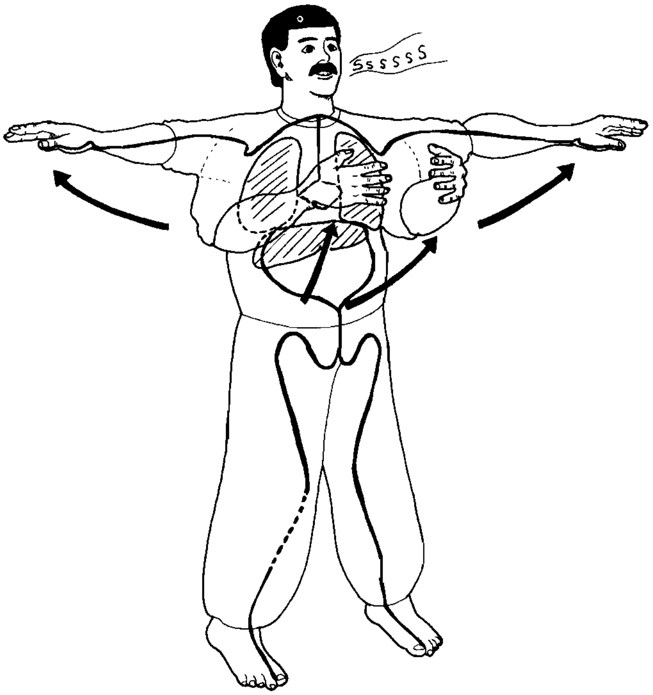

Figure 40.5. The sound "Si" (Sss) is used in order to Purge the Lungs.

Next, inhale and turn the right palm over to face the body; slowly descend the right arm in front of the body, while the left arm turns towards the body and ascends. The two hands cross in front of each other (with the ascending arm on the inside, and the descending arm outside). Now the left hand pushes up and the right hand pushes down. At the same time, exhale and repeat the sound "Who." Pressing both left and right hands form one set. Perform six times. At the end, cross both arms in front of the chest and return to the beginning position.

THE SOUND "SI" (SSSS) MOISTENS THE LUNGS

This sound is used to nourish the Lungs, treat problems such as tuberculosis, abscesses, and dermatitis.

From the beginning Wuji posture, inhale and raise both hands as if holding a ball, moving it upwards, until the hands face the Shanzhong CV-17 point at the front of the chest. Direct the Qi of the Liver's Channels to ascend from the Dadun Lv-1 point on the lateral portions of the big toes, along the inner portion of the legs, through the lower abdomen into the Lungs.

When exhaling, turn over both palms to face downward, then separate the palms horizontally to the sides. Guide the Qi along the Lungs' Channels to the Shaoshang Lu-11 points on the inner portion of the thumbs, while exhaling the sound "Sss" The lips are slightly drawn back, the upper and lower teeth are slightly brought together forming a small slit; the tip of the tongue is placed against the slit in-between the teeth to articulate the sound. Exhale completely, then allow both arms to descend to the sides of the body (Figure 40.5). Repeat six times.

THE SOUND "CHUI" (CHREE) STRENGTHENS THE KIDNEYS

This sound is used to regulate the Kidneys, expel chills, and to maintain the Internal Heat of the Lower Dantian and Mingmen Fire.

Inhale and raise both arms to embrace an imaginary ball in front of the chest (Middle Dantian area).

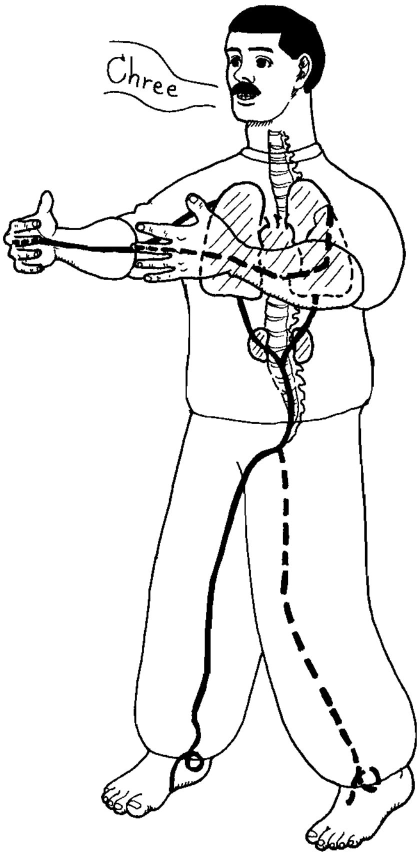

Figure 40.6. The sound "Chui" (Chree) is used to regulate the Kidneys and the Mingmen Fire, as well as to expel chills from the body.

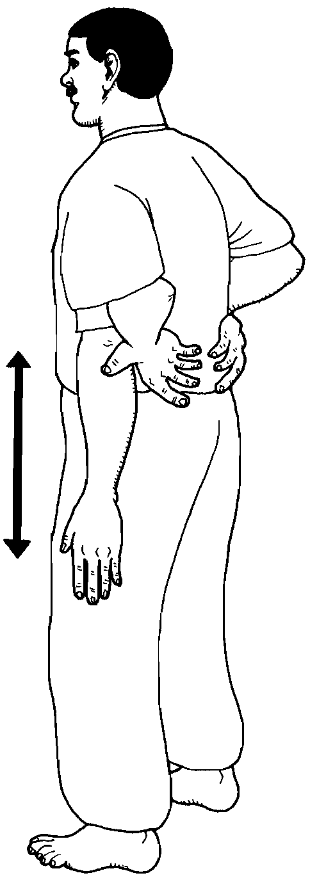

Figure 40.7. Massage the sides of the body.

The joints in between the thumb and index finger of both hands are stretched wide open. Lead the Qi to flow upward from the Yongquan Kd-1 points of the Kidney Channels, through the inner portions of the thighs, along the spinal column, and into the Kidneys.

When exhaling, hold the posture and stretch the fingers, while leading the Qi from the Kidneys into the chest and then to the Lungs. The channel Qi then transfers into the Pericardium Channels and flows down the arms to the Zhongchong Pc-9 points at the tips of the middle fingers. While exhaling, sound the word "Chree." The mouth is nearly closed with the corners slightly drawn back; the tongue moves forward. The breath should feel cold (Figure 40.6).

Slowly straighten the body as the hands circle downward, from the lower back to the thighs. Rub the lower lumbar region (Shenshu Bl-23 points) and the sides of the thighs (Huantiao GB-30 points), moving the hands upward while massaging the buttocks. Return to the holding the ball posture and repeat six times (Figure 40.7).

THE SOUND "XI" (SHE) REGULATES THE TRIPLE BURNERS

This sound is used to eliminate Excess Heat, for treating rheumatism, poor circulation, and problems of the Triple Burners. This exercise helps to recharge the body's Qi metabolism, regulate Internal Qi, and balance the eight systems of the body (respiratory, digestive, reproductive, circulatory, lymphatic, nervous, muscular/skeletal, and excretory systems).

The "She" sound of the Triple Burners is always prescribed for patients undergoing radiation or chemotherapy. This allows the patient to disperse the toxic Heat left within the tissues.

From a beginning Wuji posture, inhale and raise both hands as if holding an imaginary ball up in front of the chest at the Middle Dantian. Direct the channel Qi of the Triple Burner Channels in the arms to flow downward to the Lower Dantian, then down the legs via the Gallbladder Channels to the tips of the fourth toes of the feet at the Qiaoyin points (GB-44).

Exhale the sound "She," at the same time rotating both arms upward to form the Embracing the Heavens posture. The lips are slightly open and drawn back; the tongue suspends behind the teeth. When exhaling, direct the Qi to flow from the Qiaoyin GB-44 points on the outer portion of the fourth toes upward, along the Gall Bladder Channels, through the outer portions of the legs to the head, and transfer the energy flow into the Triple Burner Channels, directing it along the outer portions of the arms to the Guanchong TB-1 points at the ends of the ring fingers. Exhale completely.

Inhale, while turning the palms over to face downward, and begin lowering both arms along the front of the chest. Imagine the Qi flowing down from both arms into the Lower Dantian. Exhale without sounding "She," and press both palms down along the front of the abdomen to the sides of the hips, returning to the Wuji posture (Figure 40.8). Repeat the entire sequence six times.

This exercise is different from the others in that the energy circulates back and forth through the channels but does not disperse outside the body. Meaning you inhale Qi from fingers to toes, exhale Qi from toes to fingers and sound, inhale Qi from fingers to Lower Dantian, exhale Qi from Lower Dantian to toes, without sound, then begin entire sequence again.

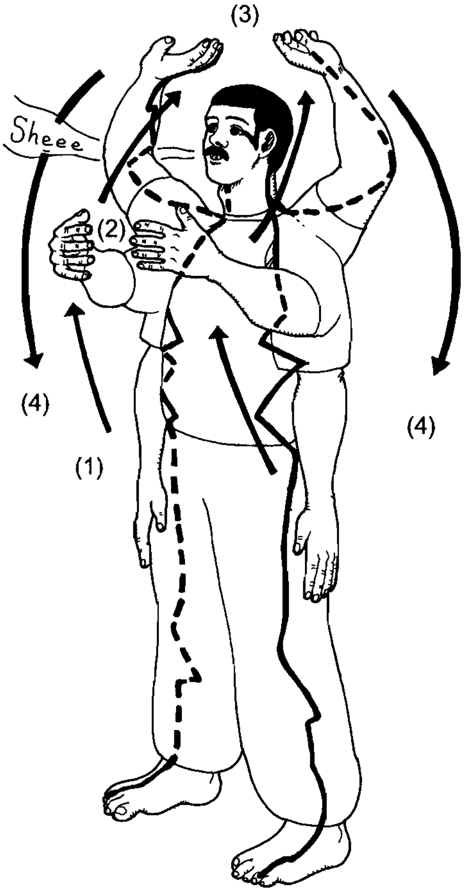

Figure 40.8. The sound "Xi" (She) is used to purge Excess Heat from the body.

THE PROPER NUMBER OF BREATHS FOR HEALING SOUND PRACTICE

When practicing all six healing sounds, each sound should be practiced a minimum of three to six times before proceeding on to the next sound. If the patient wishes to practice the six healing sounds daily, the patient should perform each sound no more than twenty-four times. If the patient is practicing one or more healing sounds for a specific therapy, then the number of times for tone resonances should be thirty-six.

Practice the six healing sounds in the sequential order of the Fire Cycle (i.e., Liver, Heart, Spleen, Lungs, Kidneys, and Triple Burners) to enhance the

Yang energetic field. Practice the six healing sounds in the sequential order of the Water Cycle (i.e., Liver, Kidneys, Lungs, Spleen, Heart, and Triple Burners) to enhance the Yin energetic field.

It is advisable to practice the healing sounds at the end of a meditation practice. This is because the free circulation of Qi is already in motion and freely accessible after meditating.

COMBINING MULTIPLE ORGAN SOUNDS

The six healing organ sounds may be performed as in the prescribed series, or as individual therapies ranging from single to multiple organ sounds. Some examples are:

1. Single organ sound regulation, like the "Haa" sound for the Heart, is used followed by swallowing the collected saliva (after Qi purgation) to relieve a fast, irregular heartbeat.
2. Multiple organ sound regulations, like the sounds of "Shu" (Liver), "Haa" (Heart), and "Sss" (Lungs), are used in sequence to quench the pathogenic Fire inside the patient's body (for mild symptoms).
3. All six healing sounds are used for regulation and balance to treat "False Cold" (a feverish sensation followed by a feeling of cold all over the body). This symptom is caused by Excess Fire. The Internal Fire drives the Yin to the body's surface, thus producing the cold sensation. To treat, the Qigong doctor guides the patient's Yin Qi back to its original organs and uses all six sounds to regulate the body.
4. Sounding according to the Five Seasons will also be beneficial to the patient's Qi regulation. The five sounds are divided according to each season, and the vulnerability of each organ. They are categorized as follows:
    a. "Shu" sounding in the springtime will benefit the Liver.
    b. "Ha" sounding in the summer will benefit the Heart.
    c. "Ssss" sounding in the autumn will benefit the Lungs.
    d. "Chree" sounding in the winter, and during all seasons, will vitalize the Kidneys and harmonize Qi and Blood.
    e. "Who" sounding in all seasons helps vitalize the Spleen and regulates the function of Postnatal Qi.
5. Combining healing sounds according to the specific prescriptions is similar to taking combinations of Chinese herbs. The Six Healing Sounds may be combined in order to target specific areas on the body's energetic system. Generally, the doctor gives the patient specific healing sound combinations as prescriptions to purge stagnation, and to increase the energetic movement within the diseased tissues or organs.

One such combination prescription used to treat overall Shen disturbances due to emotional depression, from suppressed anger, guilt or grief is called, "Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool." In Medical Qigong Soul Retrieval, this healing sound prescription is used to bring the patient back in touch with his or her painful emotions.

OLD MAN SEARCHING FOR THE REFLECTION OF THE MOON AT THE BOTTOM OF TIDE POOL

This particular exercise was introduced into the United Stated in the early 70's by Dr. Her Yue Wong. Doctor Wong believed that after the Qigong student has spent a sufficient amount of time practicing the Five Elemental Daoist Qigong tonification exercises, his or her deep seated emotions will begin to surface and must be purged from the body. In order to purge the toxic emotions, this healing sound prescription was given:

1. Begin this exercise prescription from a standing posture. Both feet are facing forward, shoulders width apart. Inhale, and imagine divine healing light filling the Lungs. While inhaling, separate and stretch both arms to the sides of the body, forming the shape of a "T" (Figure 40.9).
2. Lean over and begin to exhale, making the "Sh-h-h-h" sound. While exhaling imagine Toxic Qi from both Lungs flowing down each arm and pouring into the ground. Both arms should swing from one side to the other, crisscrossing

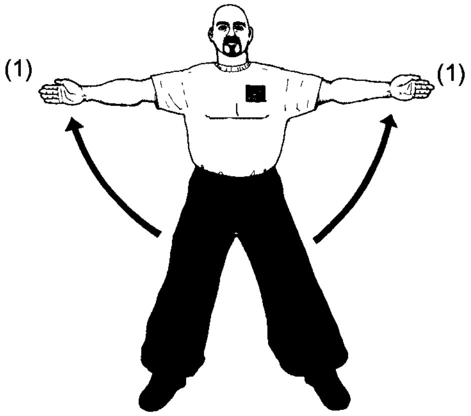

Figure 40. 9. Old Man and Tide Pool (a)

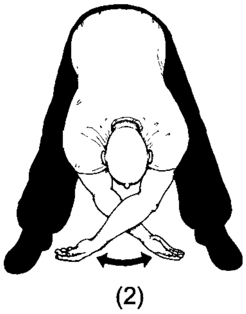

Figure 40. 10. Old Man and Tide Pool (b)

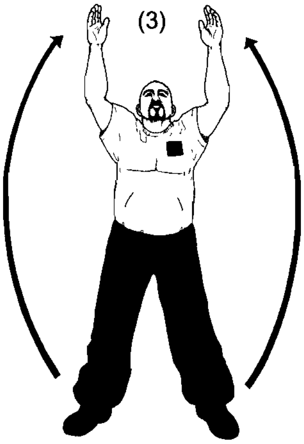

Figure 40. 11. Old Man and Tide Pool (c)

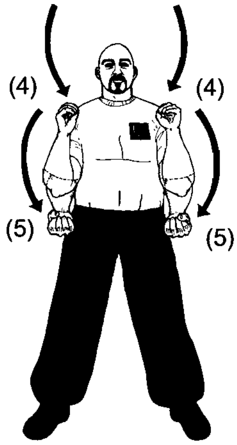

Figure 40. 12. Old Man and Tide Pool (d)

while making the purging sound (Figure 40.10).

3. Return to an upright position, while inhaling and imagining divine healing light filling the chest (especially the Heart, Liver and Spleen areas). When inhaling, look upward and raise the hands over the head (Figure 40.11).
4. When exhaling, both hands should descend to shoulder level, making soft fists (as if embracing two sparrows). Focus the mind's intention on the center of the chest and imagine the Toxic Qi releasing from the Heart while making the "Haa-a-a-a" sound. Exhale only half of your breath. Next, exhale the second half of your breath while imagining Toxic Qi releasing from the Liver and Spleen and making the "Who-o-o-o" sound. While exhaling, move both hands down to the hips while keeping the fists soft (Figure 40.12). Repeat the last two steps (c. and d.) three times. This equals one set. Practice 36 sets to complete the exercise. It should take about 25 minutes.

There is an story associated with this ancient healing sound prescription; it is described as follows.

One night, on a full moon, an old man on the southern coast of China, went looking for sea crustaceans to feed his family. Holding his lamp upward, he gazed down and saw a large beautiful pearl at the bottom of a tide pool. "Oh My!" He exclaimed, as he quickly placed his lamp on a rock. "Look at the size of that pearl! With a pearl that large, I'll be able to retire! Even my family won't have to work, and we'll all live in the lap of luxury for the rest of our lives!"

So the old man stretched his hands wide and began splashing his arms in the tide pool, trying to grasp the large pearl. Sh-h-h-h-h-h-h-h! went the water as the old man frantically grasped at nothing. Perplexed, the old man looked up. Reaching into the night sky, the old man's eyes widened as he saw the full moon. "Ha-a-a-a-a-a!" exclaimed the old man, "It's the moon!" "Whooo-o-o-o-o-o" cried the old man, "There's no money!"

CONTRAINDICATIONS FOR THE SIX HEALING SOUND THERAPY

There are several counter-indications for the use of the Six Healing Sound Therapy:

- The resonating sounds have a strong affect on the tissues, therefore patients who have broken or fractured bones are forbidden to practice sound therapy. Only after the bones tissues have mended are the patients allowed to practice the healing sounds.
- Due to the stimulating effect of the Six Healing Sound Therapy, it is advisable for the Qigong doctor not to prescribe the practice of the six healing tones in a very acute stage of illness. Vibrating the patient's internal organs in the beginning stages of a disease (i.e., a high fever) may stimulate the patient's virus or bacterial growth.
- Women are prohibited from practicing Tone Resonation therapy when pregnant, as the tone resonations have a pattern of dispersing clots and may lead to spontaneous miscarriages.
- Women are prohibited from practicing Tone Resonation therapy when menstruating, as it is important for the woman's body to be allowed to naturally Purge Toxic Blood from the body.

SOUND RESONATION THERAPY FOR TUMOR AND CANCER

Cysts, tumors, and cancer are characteristic of Excess stagnation of Qi and Blood, as well as stagnations of Heat and Cold. Chronic diseases, marked by obstruction of channel Qi, disorders of Qi and Blood, and the breakdown of Yin and Yang balance, are also included as Excess stagnation syndromes. The following sound methods involve a special series of tone resonation exercises (similar to, but stronger then the six healing sounds) which are utilized to disperse stagnations.

In order to find the correct pitch, the patient starts with a soft volume sound to experience the physical characteristics of the internal organ vibrating. While pronouncing the tone with the vocal cords relaxed, the patient gradually increases the sound to a louder volume.

While exhaling the sound, the patient focuses on an internal image of the diseased tumor dissolving. The patient must actually hear and feel the sound resonating the body's tissues, saturating the organ with sound vibration, energy, and light.

The following sounds are prescribed for the treatment of tumor and cancer therapy: the "Guo" sound is used by patients with Liver diseases, the "Zheng" (pronounced as jang) sound is pronounced by patients with Heart trouble, the "Gong" sound by patients with Spleen diseases, the "Shang" sound by patients with Lung trouble, and the "Yu" sound by patients with Kidney diseases. The relationship between the five pronunciations and the five solid organs as well as the number of pronunciations are shown in Figure 40.13.

PREPARATION FOR TUMOR AND CANCER SOUND THERAPY

Have the patient begin this exercise in a Wuji posture, and encourage the patient through the One Through Four meditation (One-Fun, Two-Shoe, Three-Tree, and Four-Core) described in Chapter 26.

Each tumor and cancer healing sound is pronounced in a straight tone, facing the left and exhaling the tone while rotating the body to the right. This allows the patient to purge the specific diseased organ. After completing the required number of straight tone exhalations, the patient faces forward and pauses.

The patient then begins the second half of the

		Elements				
		Wood	Fire	Earth	Metal	Water
Yin organs	solid organs	Liver	Heart	Spleen	Lung	Kidney
Yang organs	hollow organs	Gall Bladder	Small Intestine	Stomach	Large Intestine	Bladder
body turns left to right body turns right to left	high-pitched straight tone low-pitched descending-ascending tone	guo	zheng	gong	shang	yu
number of repetitions		8	7	10	9	6

Figure 40.13. Tumor and Cancer Sound Therapy Chart

prescription by facing the right direction and exhaling a descending tone while rotating the body to the center of the room, then rising tone while rotating the body towards the left side of the room. This application causes the patient to purge the diseased organ's channels of Toxic Qi.

Upon completion the patient performs the Pulling Down the Heavens technique to cleanse the body of any leftover energetic residue.

THE PROPER AMOUNT OF TONE REGULATIONS

The beginning patient should not be too anxious to perform the required number of pronunciations. By gradually reaching the required number, and slowly adapting themselves to the Tone Resonation method, the patients avoid getting tired. The sound "Shang," for example, is pronounced nine times for each exercise in the case of Lung disease. However, a beginner may start by pronouncing the tone three times, and then gradually progress to six times in twenty days. Then, in another thirty days, the patient can reach the required number of times. The number of repetitions is always dependent upon the patient's age, type of constitution and the severity of the disease.

In other words, the amount of times in a day that the patient performs the healing sounds depends upon the patient's general condition. The sounds used in tumor and cancer therapy, for example, will be prescribed more frequently with less volume then that of the regular six healing sound therapies. For Lung cancer, for example, the patient should perform the "Shang" sound eighteen times a day, nine breaths in each direction, right and left; for Lung congestion, perform the "Sss" sound three times a day, 36 breaths for each set.

SPECIALIZED PRESCRIPTIONS FOR TUMORS AND CANCER

The sound therapies used for treating tumors and cancer are quite different and much more powerful then the six healing sounds previously prescribed to purge and regulate the body's internal organs. Because of the dramatic difference in power and energetic potential between these two healing sound systems, the cancer patient will be given a lower number in tone respirations, but an increased number of times of practice per day.

The healing sounds used for treating tumors and cancers are as follows: for Liver/Gall Blad-

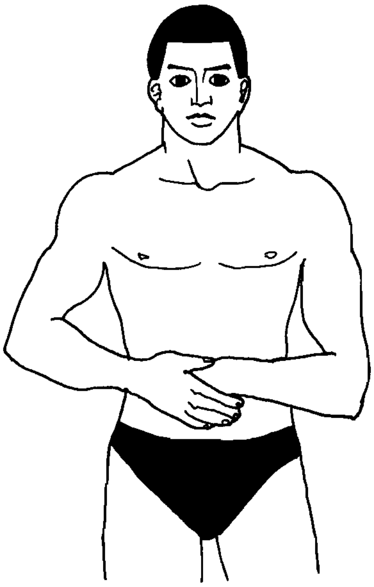

Figure 40.14. For Liver diseases the sound is "Guo."

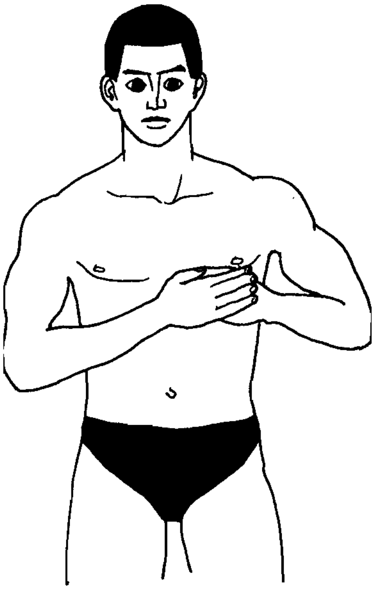

Figure 40.15. For Heart trouble, the sound is "Zheng" (Jang).

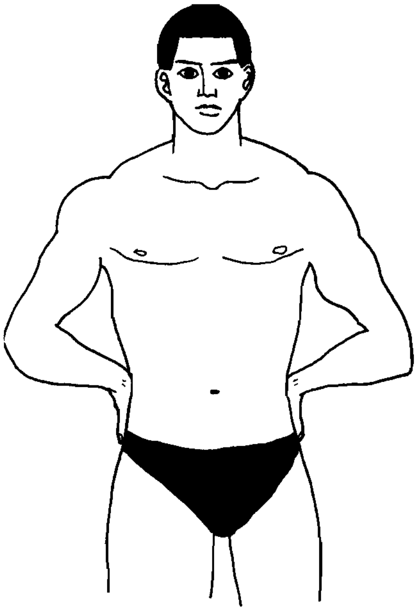

Figure 40.16. For Spleen diseases, the sound is "Gong."

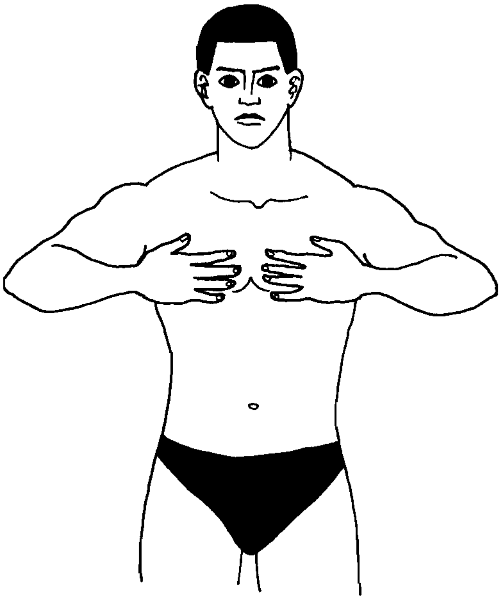

Figure 40.17. For Lung trouble, the sound is "Shang."

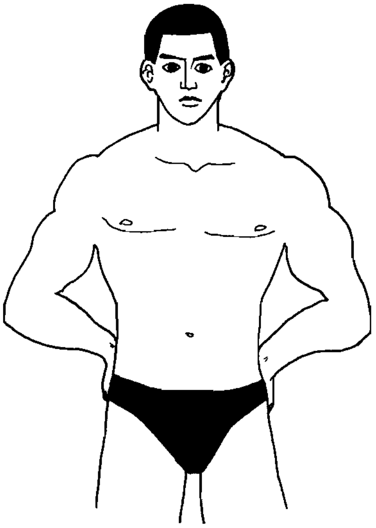

Figure 40.18. For Kidney diseases, the sound is "Yu."

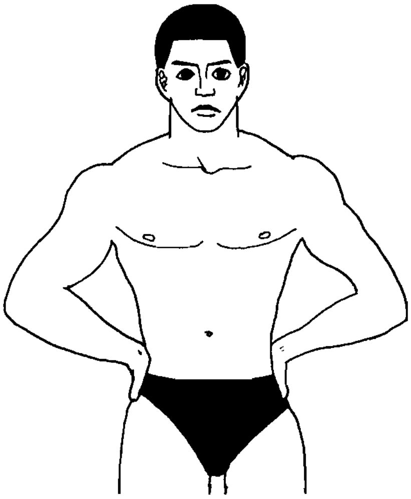

Figure 40.19. For chemotherapy, the sound is "Xi" (She).

der tumors, use the sound "Guo" (Figure 40.14); for Heart/Small Intestine tumors, use the sound "Zheng" (Jang - Figure 40.15); for Spleen/Stomach tumors, use the sound "Gong" (Figure 40.16); for Lung/Large Intestine tumors, use the sound "Shang" (Figure 40.17); and for Kidney/Bladder tumors, use the sound "Yu" (Figure 40.18). The Triple Burners' sound "Xi" (She) is used to purge Heat from the body after radiation and chemotherapy (Figure 40.19).

The choice of sound resonation must suit the patient's specific disease. Generally, in the beginning, patients with cancer should mainly pronounce the purgation sound "Haa" nine times to purge the Heart of stagnant emotions (Figure 40.20). After pronouncing the sound "Haa" for a certain time period, the patient begins to become emotionally vulnerable. The doctor then adds the specific sounds for the treatment of the diseased internal organs. The "Shang" sound, for example, is prescribed for patients with Lung cancer.

1. The tones for Radiation and Chemotherapy patients should include the Triple Burners' sound "She." This toning should be done when their blood count is low after chemotherapy or radiation therapy and is used to purge Excess Heat from the body. The sound "She" is pronounced six times. Its high-pitched sound is in the straight tone (perform six times), and its low-pitched sound is in the dropping and rising tone (perform six times).

   After the patient has stabilized and the red and white blood count has reached a normal level, the "She" sound is dropped and replace by the sound "Haa." This is done to stabilize the patient's emotions. Instead of "Haa," the doctor can also replaced the healing sound with the specific sound needed to treat the diseased internal organ.
2. The tones for Stomach cancer patients should include the sound "Dong." The Dong sound has a deeper penetrating affect than the "Gong" sound (which is usually prescribed for Stomach diseases). The "Dong" sound should be pronounced ten times (Figure 40.21). Its high-pitched sound is "Dong" in the straight tone (perform ten times), and its low-pitched sound is "Dong" in the dropping and rising tone (perform ten times).
3. The tones for Brain tumor patients (with cerebral carcinoma) should include the sound "Duo." The "Duo" sound should be pronounced ten times (Figure 40.22). Its high-pitched sound is "Duo" in the straight tone (perform ten times), and its low-pitched sound is "Duo" in the dropping and rising

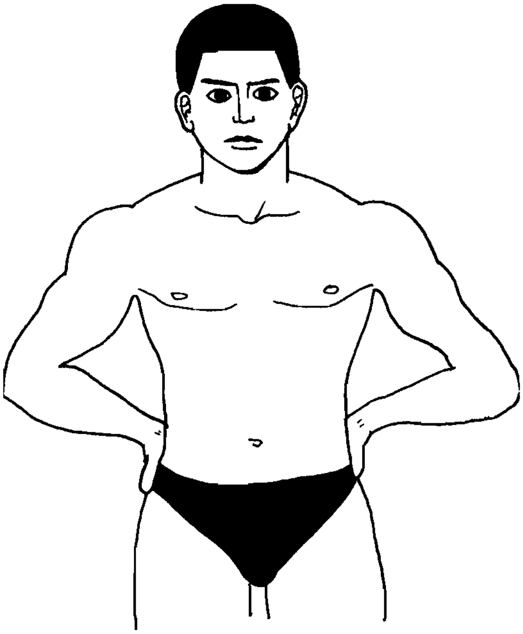

Figure 40.20. For cancer in general, the sound is "Ha."

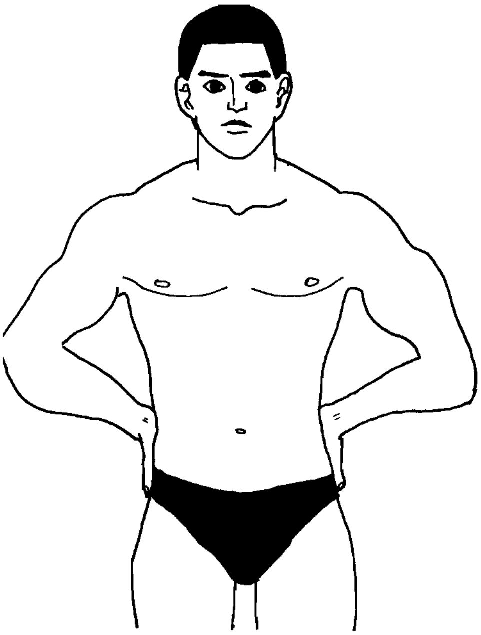

Figure 40.21. For Stomach cancer, the sound is "Dong."

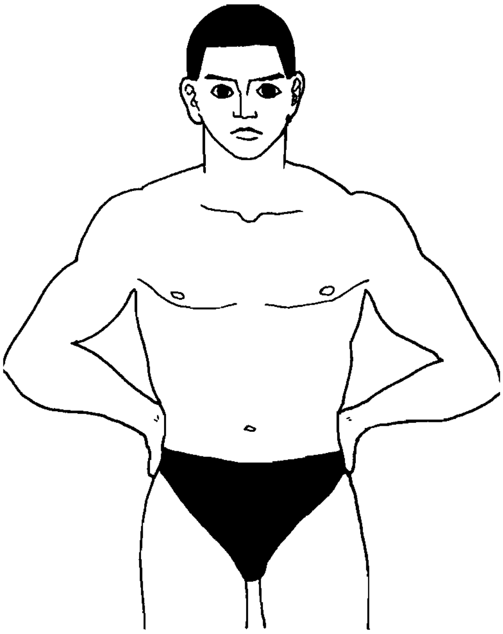

Figure 40.22. For brain cancer, the sound is "Duo."

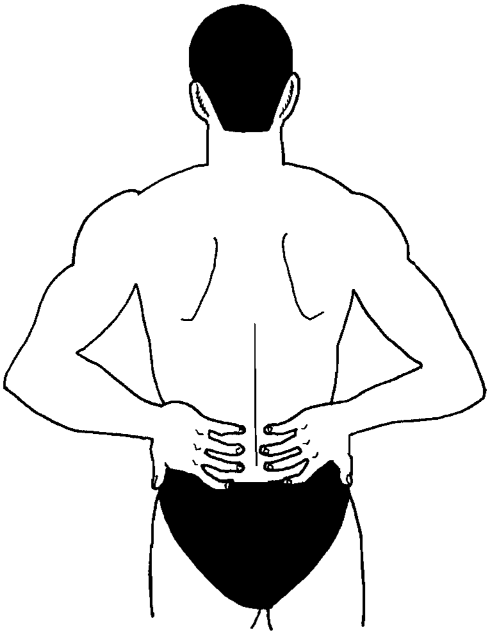

Figure 40.23. Back View for the Cancer Prescriptions

tone (perform ten times). In addition the patient should also practice techniques to strengthen the Kidneys and Heart.

4. The tones for Lung tumor patients (with carcinomas of the Lungs) should include the "Shang" sound. The "Shang" sound should be pronounced nine times. The patient should also use the Fast Breathing Method technique (2 inhalations, followed by 1 exhalation of the sound "Shang"), in addition to practicing Lung and Kidney strengthening exercises.
5. The tones for Large Intestine tumor patients (with carcinomas of the Large Intestine) should include the "Shang" sound. The "Shang" sound should be practiced nine times using the Fast Breathing Method technique, in addition to practicing Lung strengthening exercises.
6. The tones for Nasopharyngeal tumor patients should include the "Shang" sound. The "Shang" sound should be practiced nine times using the Fast Breathing Method Walking technique (see Chapter 15), in addition to practicing Lung strengthening exercises.
7. The tones for Small Intestine tumor patients (with carcinomas of the Small Intestine) should include the "Zheng" sound. The "Zheng" sound should be practiced seven times, using the Fast Breathing Method technique, in addition to practicing Heart strengthening exercises.
8. The tones for Liver Cancer and Gall Bladder tumor patients (with carcinomas of the Liver and Gall Bladder) should include the "Guo" sound. The "Gou" sound should be practiced eight times, using the Fast Breathing Method technique, in addition to practicing Lung, Liver, and Spleen strengthening exercises.
9. The tones for the Bladder and uterus tumor patients (with carcinomas of the Urinary Bladder and uterus) should include the "Yu" sound. The "Yu" sound should be practiced six times, using the Fast Breathing technique, in addition to practicing Kidney strengthening exercises.

   While in a Wuji posture, the patient places both hands on the Kidneys and presses the middle fingers at the Shenshu Bl-23 point beside the Mingmen before pronouncing the sounds (Figure 40.23).
10. The tones for Prostate cancer patients should include the "Yu" sound. The "Yu" sound should be practiced six times, using the Fast Breathing Method technique, in addition to practicing Kidney strengthening exercises.
11. The tones for Skin Cancer patients (with dermal carcinomas) should include the "Shang" sound. The "Shang" sound should be practiced nine times, using the Fast Breathing Method technique, in addition to practicing Lung, Liver, and Spleen strengthening exercises.

CONTRAINDICATIONS FOR TUMOR AND CANCER SOUND THERAPY

As previously mentioned, the tumor and cancer healing sounds are only prescribed for purging Excess conditions within the patient's diseased organs, and not for treating Deficient organ conditions. If a specific internal organ is already in a state of Deficiency and the patient is allowed to focus on that organ with one of the tumor and cancer sounds, the patient's condition will worsen. This increases the potential for rapid growth of the disease.

CHAPTER 41
MEDICAL QIGONG REGULATION EXERCISES

INTRODUCTION

The following chapters explore several systems of Medical Qigong exercises used in regulation, purgation, and tonification prescriptions. Conceived in ancient China, these exercises originate from the various medical, martial, and Daoist schools of energy cultivation. It is from these ancient energetic systems that today's modern clinical Qigong prescriptions originate.

Each Qigong system has its own unique approach to circulating, balancing, and storing the body's life-force energy. No one particular style of energetic healing is superior to the next, only different in its approach. The primary goal of each of these systems is removing stagnant pathogenic energy and replenishing and circulating clean life-force energy.

The Xi Yuan Hospital in Beijing, China is famous for its advanced Medical Qigong clinics and Medical Qigong prescriptions. These exercises utilize both Static and Dynamic Postures and are generally taught in both the inpatient and outpatient clinics. The static exercises are used to store the Qi, while Dynamic Qigong exercises are used to open the patient's channels and circulate the Qi.

Each prescription should be practiced for five to ten minutes for each organ or organ system on a daily basis; or if the patient is practicing only one set (e.g., Regulation of Spleen and Stomach Qi), they should practice for thirty minutes. Unless otherwise indicated, all breathing should be "natural."

The goal is for patients to empty their body and mind, so that their emotions can be felt and activated. In the Medical Qigong clinic there is a saying, "The patient can only heal what they feel."

STARTING AND ENDING THE PRESCRIPTIONS

Before starting any Medical Qigong prescription, it is important for the patient to understand that every Qigong exercise must start and end with specific preparations.

- Before starting the Medical Qigong prescription, the patient must always purge his or her body. It does not matter if the exercise they are about to begin is used to purge, tonify, or regulate the internal energy; they must always purge first. This is similar to washing a dirty glass before filling it with clean water.
- After ending the Medical Qigong prescription, the patient must always gather the generated Internal Qi and the return to its origin. The term "Return to its Origin" refers to a three fold process:
  - cultivating, collecting, and accumulating the prenatal energy,
  - developing and transforming it into postnatal Jing, Qi, and Shen, and
  - returning it back to prenatal Jing, Qi, and Shen.

This allows the Post-Heaven Kan and Li (Water and Fire energies) to return to their origin and restores them to Pre-Heaven Qi (Heaven and Earth energies). This fusion into one energy, often called the Hun-Yuan or Blended Originals, requires the patient's original substances to condense into one energetic substance and be merged with the patient's Lower Dantian energy.

In order to collect the energy, the patient is taught to apply the Three Natures Become One technique of focused concentration (all thinking, seeing, and listening is focused into the Lower Dantian). The Three Natures are as follows: the eyes represent the perceptive nature, the ears represent the spiritual nature, and the Heart represents the courageous nature. When all three natures unite into one, the energetic refinement is called "collecting the body's true nature." The patient can then collect the body's power.

MEDICAL QIGONG CULTIVATION EXERCISES

The body is composed of different densities of energy and matter in the form of internal organs, organ systems, channels, and collaterals. When practicing Qi cultivation methods both students and patients are instructed to train in a progressive order of Qi cultivation.

First, the students and patients are taught how to purge, tonify, and regulate their internal and external channels. Once the Yin and Yang energy of the channels are balanced, the practitioners learn how to balance the Yin and Yang energy of the internal organs. This training likewise involves purging, tonifying and cultivating methods. The directed focus of intention on specific points, channels, or areas of the body causes the Qi to gather in those locations and to be absorbed into the body's tissues through the Mind's intention (whole body awareness and consciousness). The final step of training the Yuan (Original) Qi is undertaken only after the first two steps have been accomplished.

All cultivation practices of Medical Qigong therapy can be divided into two main categories of gathering and training Qi: the External Elixir Cultivation (Wai Dan Shu) and the Internal Elixir Cultivation (Nei Dan Shu). These two Qi cultivation training methods are described as follows.

EXTERNAL QI CULTIVATION

The External Elixir Cultivation (Wai Dan Shu) involves several methods of gathering energy from the external environment (Heaven and Earth) to strengthen the body's organs, organ systems, channels, and collaterals.

HERBS AND FOOD

The first method is developed through food and herbal soups, teas, and pills to increase the body's energy. When herbs are prescribed as a tonic, they stimulate the organ's energy by increasing the flow of Qi in the patient's channels. Qigong and herbology go hand-in-hand with respect to the energetic regulation and balance of the body's life-force energy. Herbs can be extremely important for maintaining the physiological balance of the metabolism and restoring any depletion of the body's energetic mechanism. Herbs act as a form of energetic nutrient, allowing the Earth energy to be ingested into the body through a natural substance.

MEDITATION

The second method is developed through meditation. Qi is cultivated through the body's extremities from energy originating outside of the body's torso (i.e., both palms, both soles of the feet, and the top of the head). The External Elixir Cultivation meditation focuses the concentration on either the Heaven or Earth energy fields to draw their energy into the torso for increased Yin or Yang Qi regulation. The Postural Dao Yins in External Elixir Cultivation may be either static or dynamic.

Two examples of External Elixir Cultivations are: the Earth Yin Qi Regulation and the Heaven Yang Qi Regulation. The Earth Yin and Heaven Yang Qi Regulations are used to balance the energy along the body's ascending and descending energetic channels. To balance along the horizontal energetic channel, perform the Turning and Winding the Belt Vessel exercise described later on in this chapter.

- **Earth Yin Qi Regulation**

  The Earth Qi is considered Yin Qi or electronegative energy. Blood also pertains to Yin Qi. Through Earth Regulating Qigong meditations, the body's Blood cells can be replenished with Earth Yin energy.

  Begin the Earth Qi Regulation meditation from a Wuji posture, while using Natural Breathing. In this particular exercise, the practitioner takes advantage of the Kd-1 points at the bottom of the feet, and the Pc-8 points at the center of the palms which are sensitive to the magnetic pull of the Earth.

  After standing in the Wuji posture for several minutes, inhale and begin to absorb the

Earth Qi up from the ground through the centers of the palms and the bottom of the feet (Figure 41.1). After several more minutes of drawing the Earth energy into the Lower Dantian, begin circulating the energy through the Microcosmic Orbit's Fire Cycle. The palms should begin to expand and feel almost swollen; the red and white blotches on the palms and feet indicate that the exercise has been practiced correctly.

- **Heaven Yang Qi Regulation**

Heaven energy pertains to the sun's Yang Qi or electropositive energy, which is used to tonify the body's Yang Qi.

Begin from a Wuji posture, placing both palms above the head, facing upwards, like two antennas collecting the sun or Heaven's Qi into the Middle Dantian. Use Natural Breathing. Imagine the sun's golden light showering the outside of your body. Open your mouth and imagine swallowing the sun's Qi, allowing it to mix with your saliva. As you swallow, feel the energy flow down into your Lower Dantian, and then out of your pores. Once the Dantian is full, the body should feel expansive. Always end the meditation by returning the energy to the Lower Dantian to avoid any Qi deviations (Figure 41.2).

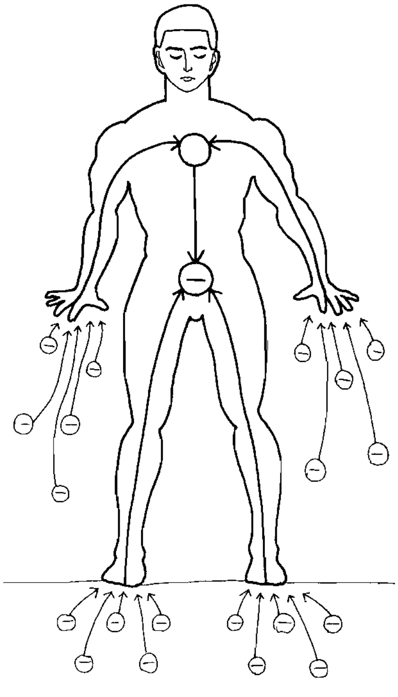

Figure 41.1. Absorb the Earth Qi into the body through both hands and feet, and store the Qi in the Lower Dantian.

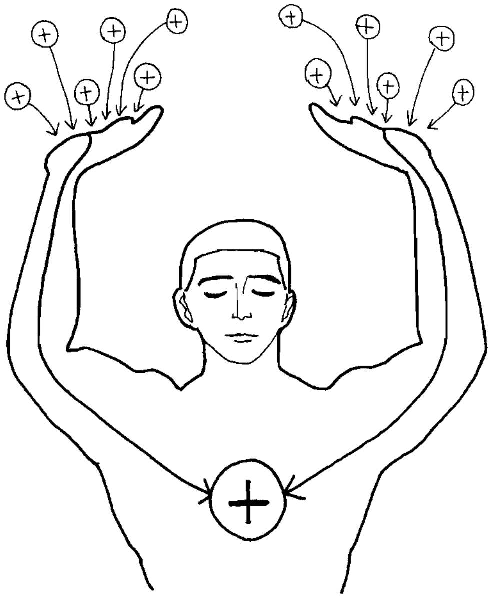

Figure 41.2. Absorb the sun's energy into the body through both palms, moving it into the Middle Dantian, and store the Qi in the Lower Dantian.

INTERNAL QI CULTIVATION

The Internal Elixir Cultivation (Nei Dan Shu) method focuses on first training the energetic substances already gathered and existing inside the torso, and then extending this accumulated energy out into the extremities. Since it first appeared in the Han Dynasty, the training and theory of Internal Elixir Cultivation has become the primary focus of many Qigong masters, working with postnatal energy transformations. Internal Elixir Cultivation focuses on transforming the body's Jing, Qi, and Shen. When practicing this exercise, ancient Qigong masters imagined the Lower Dantian as a stove, the Shen as a Fire, and the body's Jing, Qi, and Shen as the internal medicine.

At birth, the energy flows up the front of the body (Conception Vessel) and down the back (Governing Vessel). This is considered the circulation of the Yuan Qi (generated from the prenatal Jing stored in the Kidneys). It is considered the Perceptual Yin or Water Cycle of the Microcosmic Orbit. This Perceptual Yin Cycle allows the newborn child to possess powerful psychic, empathic, and kinetic abilities; it enables the child to energetically connect with the environment in order to ensure survival.

As the child advances in age, he or she begins to encounter and experience some hostile messages in the form of physical or emotional frustration and pain from both parents, siblings, and the environment. For protection, the child armors and creates a denial system, which insulates him or her from the energetic and emotional wounding. The child begins to instinctively stop the energetic flow of the perceptual Yin Water Cycle and begins to activate the acquired rational mind, trying to cope with his or her feelings and the stress of his or her surroundings. Whenever the child begins to experience painful emotions, he or she learns to avoid experiencing these hurtful feelings through certain mental and physical actions by contracting the throat and thorax. These acquired actions further create an energetic restriction and impede the flow of Qi along the Conception and Governing Vessels. This physical activity naturally causes the Yin Water Cycle to stop its perceptual flow of energy, and allows the child's body to deal with the trapped emotional experiences by storing them within the tissues. Although this is not a healthy response, it allows the child to survive and develop in a sometimes difficult environment. In cases of deep traumatic wounding (physical, emotional or sexual), the child becomes void of deep emotional connections, while still allowing for certain perceptual survival instincts.

Also, our culture emphasizes left brain thinking for educational purposes, and later to enable young adults to rival others in the competitive job market. The child thus must forgo most of his or her innate intuitive perceptions to succeed in school and to appropriately integrate society's values and belief structures. Very little encouragement is given to right brain activity; it is in fact discouraged by society at large.

REGULATING THE BODY'S QI

The first stage in circulating and cultivating the Qi is to regulate the Yang or Fire Cycle of the Microcosmic Orbit. This allows the trapped emotions within the tissues to be released. This is initiated through mental concentration, focusing the flow of the Qi up the back and down the front of the body. It is extremely important to allow the trapped energy to release itself naturally from the body through the Fire Cycle of the Microcosmic Orbit. Otherwise, trapped emotions deep within the tissues can cause Qi deviations within the body, which may lead to energetically induced psychosis.

When regulating the Microcosmic Orbit, there are three primary methods of energetic cultivation that harmonize the body's Jing, Qi, and Shen. They relate to the sun (Fire and Heaven), moon (Water and Earth) and stars (Wind and Man).

THE FIRE PATH OF THE MICROCOSMIC ORBIT

The Yang or Fire Path is the most popular method of cultivating energy along the Small Heavenly Cycle or Microcosmic Orbit (Figure 41.3). Because it joins together the body's Yin and Yang rivers of Qi, which regulate the Twelve Primary Channels, it is considered the foundational meditation for the Internal Elixir Cultivation methods. The circulation of Postnatal Qi is generated from the Acquired Essence (food, air, and water). In this cycle, the Qi moves up the Governing Vessel and down the Conception Vessel, and is considered the ascension of Yang or the opening of the Fire Channel. The initial goal is to circulate the Fire energy from Yang to Yin to convert Jing to Qi and then to Shen. The Yang (Fire) Qi supports awareness of the emotional side of consciousness.

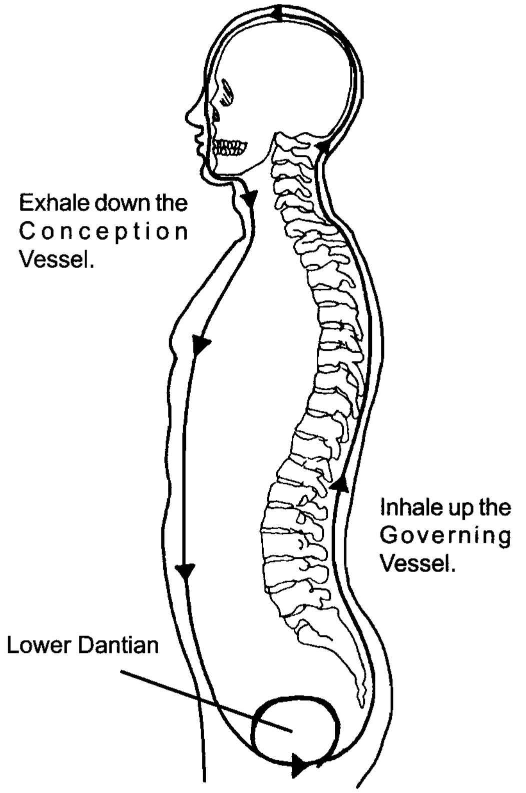
Exhale down the Conception Vessel.
Inhale up the Governing Vessel.
Lower Dantian

Figure 41.3. The Fire Path of the Microcosmic Orbit flows up the back and down the front of the body.

The Heart and Middle Dantian are considered the House of Shen; Fire and passion (Qi that manifests as emotions) and are related to the flow of the Conception Vessel. The Kidneys and Lower Dantian are considered the House of Jing, Water and pleasure and are related to the flow of the Governing Vessel. It is important to use Reverse Breathing while uniting the Heart's Fire and Kidneys' Water, to cause the alchemical transformation of Jing into Qi in the lower abdomen.

When moving the Qi along the Governing Vessel, practitioners must move the energy through the Five Gates (see Chapter 30). The energy first gathers at the Lower Dantian, then overflows into the Governing Vessel, collecting, building, and then passing through to the next gate, one gate at a time. As the energy gathers at these gates, the Qi and Blood begins to Heat and vibrate the tissues. This energetic reaction affects the patient's emotions that correspond to or are embedded in each specific gate, causing a powerful emotional release. As the Qi passes through to the next gate, a new set of emotions is experienced, then released.

THE JIAJI POINTS' EFFECT ON THE GOVERNING VESSEL'S ENERGETIC FLOW

Jia translates to mean "besides." Ji translates as "the spinal column." These points are located lateral to the lower border of each spinous process of the cervical, thoracic, and lumbar vertebrae. Each point has its own Luo pathway (energetic channel), which intersects with the spine and connects to the Jiaji point located on the opposite side of the Governing Vessel. It is believed that one pair of the ascending Governing Vessel's Luo Channel intersects and connects to all of the Jiaji points along the spine.

This being the case, when the two energetic flows from the Jiaji points (via the Luo pathways) are superimposed, the reaction causes a crisscrossing and intertwining of Qi flow, which resembles the intertwining energies of the caduceus or double snakes.

THE TWELVE EARTHLY BRANCHES' RELATIONSHIP WITH THE MICROCOSMIC ORBIT

Within the Fire Cycle of the Microcosmic Orbit, there are twelve channel points located along the Governing and Conception Vessels which relate to the Twelve Earthly Branches and the waxing and waning of the Yin and Yang cycles (see Chapter 11). The Fire phase of the Microcosmic Orbit starts at the Zi branch located at the Sheng Si Qiao area at the root of the penis in men and at the vagina in women. Moving up the back of the body, there are four phases of Qi ascent in the Governing Vessel. The four phases are: from Zi to Chou, from Zi to Yin, from Zi to Chen, and from Zi to Si. The Mao point is one of the four cardinal points (along with Zi, Wu, and You) and is a point of purification. The path up the Governing Vessel cleanses the energy through this point of purification.

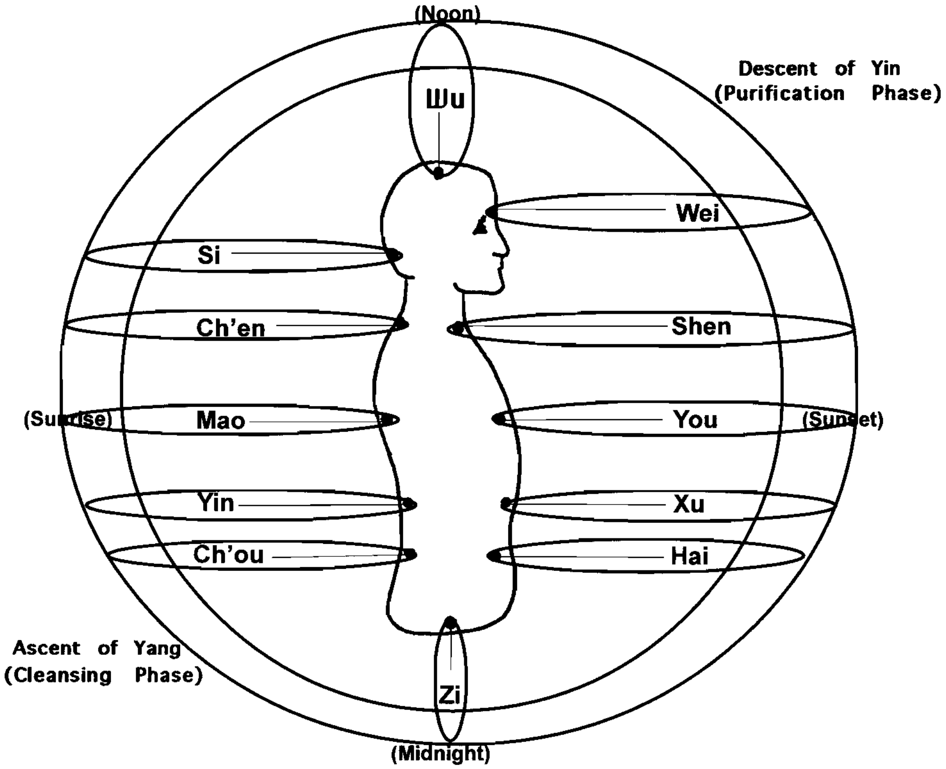

Figure 41.4. Twelve Earthly Branch Relation with the Microcosmic Orbit: the main channels of the back and front of the body correspond to the ecliptic path of the sun. These twelve special energy points are associated with the Twelve Chakra Gates.

When the energy reaches the Wu point at the top of the head (also known as the Ascending point), it stays there briefly before beginning the descent of Yin in the Conception Vessel. The descent down the Conception Vessel purifies the energy (Figure 41.4).

While at the perineum and Zi point in the orbit, the Yang Qi begins to develop as the Yin Qi reaches its zenith. Once the energy reaches the Mao point at the middle of the back, the Yang energy has developed fully and begins to prosper. At the Wu point at the top of the head, the Yin Qi begins to develop while the Yang Qi reaches its zenith. This point represents the turning point from Yang to Yin.

If the energy is to flow properly down the Conception Vessel, there must be an opening of the energy centers via a connection with the Shen. This purification of energy commences at the Wu point when the Yang Qi turns to Yin Qi. While the Earth's daily orbit around the sun has the same Yin-Yang transitions, the body absorbs the Yang energy during the hours of Zi and Wu, and expels Yin energy during the hours of Mao and You.

At the Zi hour (11 p.m.–1 a.m.), the Jing naturally collects at the Sheng Si Qiao point at the root of the penis (prostate area) in men causes a natural erection while the man is sleeping. This hour also marks the commencement of the Yang half of the day. This period of the day is the best for collecting Jing and, thus, is the best time to perform Wuji standing practice.

CONTRAINDICATIONS

Whenever the practitioners are weak, it is important that they practice Natural Dantian Breathing and avoid practicing the Microcosmic Orbit. If patients attempt to practice Circular Breathing while in a depleted state, the Qi may become trapped in their head causing Qi Deviations, and initiating more problems. Circular Breathing is achieved by guiding the Qi in a circular pattern through even inhalation (e.g., up the spine) and exhalation (e.g., down the front of the body).

THE WATER PATH OF THE MICROCOSMIC ORBIT

The Water Path is trained only after the completion of the Fire Path circulation exercise. After the patient's Prenatal Qi has been cultivated and accumulated to a substantial level, the patient may lead the Qi up the Conception Vessel and down the Governing Vessel (Figure 41.5).

This is opposite the direction of the flow of the Fire Path, and is considered the ascension of Yin or the opening of the Water Channel. The purpose for the Water Path circulation is to change, regulate, and charge the body's Yuan Shen. The Water cycle supports the perceptional side of consciousness, cools down the body's overheated Yang Qi, and regulate the Qi circulation of the Fire Path to rebalance the body.

THE WIND PATH OF THE MICROCOSMIC ORBIT

The Wind Path of the Microcosmic Orbit is the third method of circulation and requires practitioners to lead energy up the center of their body through the Taiji Pole (Figure 41.6). This energy pathway flows up through the Marrow of the spine, through the brain, and washes over the falx cerebri into the pineal and pituitary gland. It then descends down along the Conception Vessel and back into the Lower Dantian.

The purpose of the Wind Path is to increase divine perception (by accessing the energy of the Wu Jing Shen), and to regulate the body's glandular and endocrine systems, to enable the practitioner to slow down the aging process.

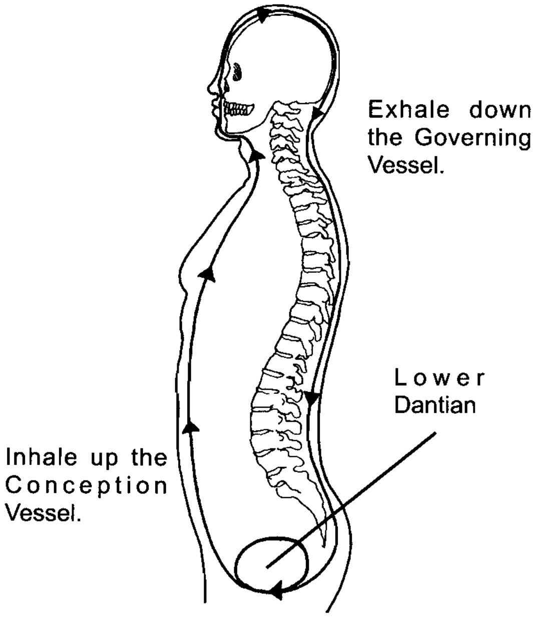
Exhale down the Governing Vessel.
Lower Dantian
Inhale up the Conception Vessel.

Figure 41.5. The Water Path of the Microcosmic Orbit travels up the front and down the back of the body, ending at the Lower Dantian.

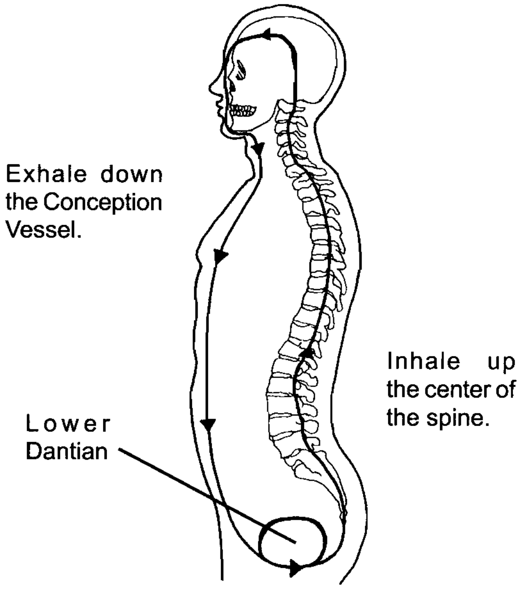
Exhale down the Conception Vessel.
Inhale up the center of the spine.
Lower Dantian

Figure 41.6. The Wind Path of the Microcosmic Orbit travels up the center of the spine into the brain and third-eye point, down the front of the body, ending at the Lower Dantian.

MEDICAL QIGONG PRESCRIPTIONS

The following exercises are taken from various hospitals, universities, and institutes of Medical Qigong. These regulation prescriptions are used throughout China.

DESCEND THE QI AND CLEANSE THE ORGANS

This exercise was developed to utilize the body's Hun-Yuan Qi (the body's internal combined organ energy) to flush and cleanse the Triple Burners and the Five Yin Organs. This is a purging exercise, used to:

- eliminate the body's Turbid Qi,
- unclog the channels and collaterals,
- wash the bones,
- transform the Marrow, and
- combine the remaining pure energy with the body's Yuan Qi.

It can be prescribed for patients with Excess conditions to descend the Liver's or Heart's Fire.

PULLING DOWN THE HEAVENS

This exercise requires the focus of energy to flow down the body in three directions (front, center, and back) to complete one set.

1. Begin in a Wuji posture, with both feet facing forward, a little wider than shoulders width apart. Both hands are at the sides of the body. The spine is kept straight, with the body relaxed and centered (see Chapter 15 for proper alignment). Stand with the eyes focused on the horizon. Pull the Qi from the horizon into the eyes and down to the Lower Dantian, filling the entire body. Hold this posture for one to three minutes (Figure 41.7).
2. Inhale through the nose. With the palms facing downward, begin to scoop the energy from the Earth into the body, gathering and absorbing the environmental energy. Focus on slowly filling the body with Earth Qi while moving the hands upwards to about waist level (Figure 41.8).
3. Continue inhaling, turning the palms to face the Heavens and drawing both hands upwards at a forty-five degrees angle. As both palms continue to move upwards, begin pulling in and absorbing the Heavenly Qi while guiding the hands above the head. Complete the inhalation with the arms above the head, palms facing the top of the head. Pause when the hands are above the Baihui area (Figure 41.9).
4. Exhale through the nose, and begin descending the palms in front of the body. Allow the palms to face the body and direct the Qi flow through the head and throat area (Figure 41.10).
5. Imagine the energy continuing to pour down through the chest and abdominal areas (Figure 41.11).
6. While exhaling, slowly lower the hands in front of the body in one continuous movement, until they reach just above the knees (Figure 41.12).
7. Inhale and begin the sequence again. Each time the hands pass in front of the body, exhale and imagine the Qi being emitted through the hands into the tissues. Three passes over the front of the body equal one set, practice 36 sets. The energetic pathway of each set is explained as follows:
    a. The first pass begins by leading the Qi down the front of the body from the Baihui area. Each palm facing the body divides the energy into two rivers, which flow from the top of the head, over the ears, and down the throat. Continue guiding the two rivers down the front of the chest, joining again at the navel. From there lead the Qi down into the Huiyin CV-1 point at the perineum, then separate them again into two rivers which continue to flow down the Yin channels on the inside of each leg, ending at the Yongquan Kd-1 point at the bottom of each foot.
    b. Start the next pass at the middle of the head above the Baihui area. Lead the Qi down the center of the body, and absorb the Heavenly Qi as it passes through all the body's internal organs, till it reaches the Lower Dantian area. From the Lower Dantian divide the energy into two rivers and descend the Qi down the center of the legs ending

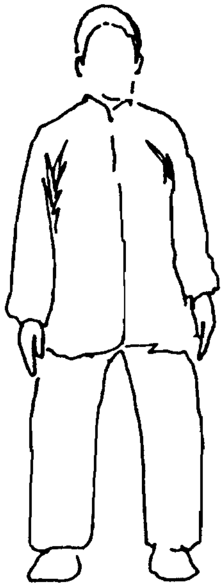

Figure 41.7. Wuji Posture

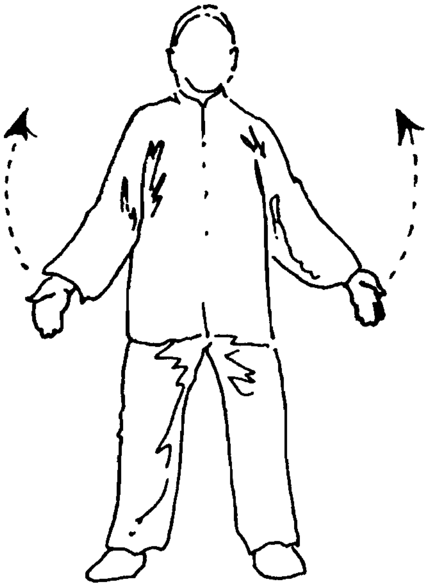

Figure 41.8. Pulling Down the Heavens (A)

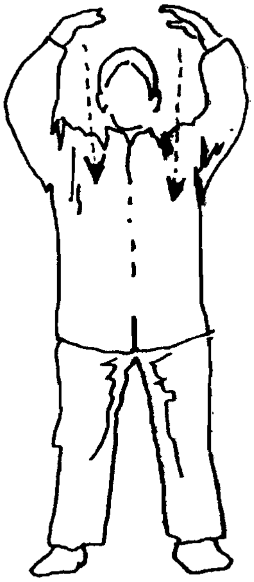

Figure 41.9. Pulling Down the Heavens (B)

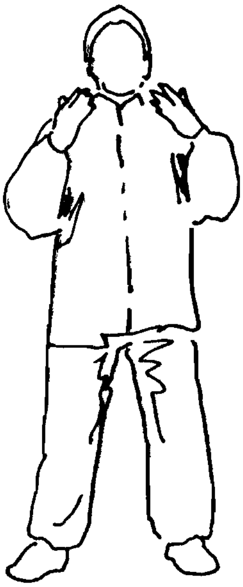

Figure 41.10. Pulling Down the Heavens (C)

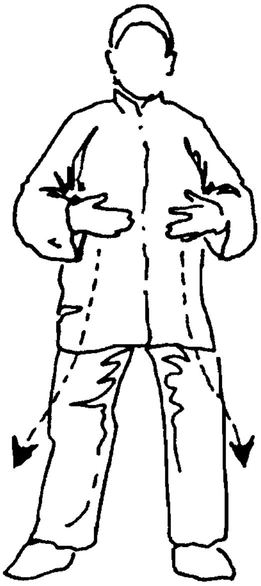

Figure 41.11. Pulling Down the Heavens (D)

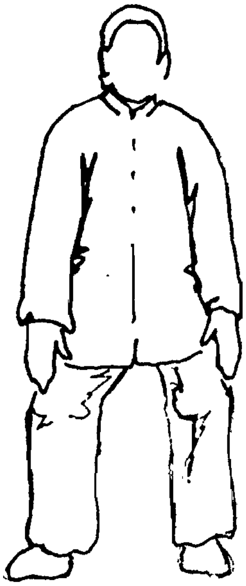

Figure 41.12. Pulling Down the Heavens (E)

at the Yongquan Kd-1 point at the bottom of each foot.

c. The final pass likewise begins at the top of the head, at the Baihui area. Imagine the energy flowing like a river, down the midline of the back of the head and back, until it reaches the Mingmen GV-4 point. From there it divides into two rivers, descending into the Huantiao GB-30 point at the side of each buttock, and continues to flow down the outside of each leg and ends at the Yongquan Kd-1 point at the bottom of each foot. As the Qi descends (on each pass), it washes the organs from the top of the head (Baihui point) to the bottom of the feet.
d. To end the exercise, relax the whole body and return to the Wuji posture, allowing the Qi to return to its origin. As the Qi descends through the body, imagine the environmental energy fusing with the body's Yuan Qi, forming a mist and dew. Imagine this mist slowly flowing down (from top to bottom) and bathing, cleansing, and purifying the entire body. As the mist descends, if a certain organ is diseased, pause for a while to bathe the area with the vapor before continuing the flow of energy. Consciously disperse the pathogens out the extremities.

GATHERING QI IN THE UPPER AND LOWER DANTIANS

This exercise is designed to strengthen, tonify, and regulate the body's internal organs and energetic fields. It is excellent for promoting and maintaining health, and can be used for tonifying the Kidneys, as well as to treat impotence, premature ejaculation, lumbago, lower back problems, and pelvic inflammatory disease (PID). When treating PID, use reverse breathing with Static Qigong training.

This is a tonification exercise, used to increase the production of Jing (sperm or ovary energy), transforming it into Qi, and to lead the Jing up the spinal cord to support the brain, as well as nurture the bone marrow. This exercise adds Fire to the Mingmen, simmering the condensation of energy within the Kidneys. When the Kidney (Water) and Heart (Fire) energies are sufficient, their energies merge to form the True Qi which cultivates one's nature, spirit, and life. This exercise also assists the Microcosmic Orbit and establishes a base for refining and gathering energy in the Upper and Lower Dantians and Huiyin.

GATHERING QI IN THE UPPER DANTIAN

Begin from a Wuji posture. Inhale, as both hands gather energy while moving up the sides of the body and embrace the Qi into the Upper Dantian. Pause and exhale while separating the hands (Figure 41.13).

Inhale and imagine the middle fingers of each hand gathering the Heaven and Earth Qi into the Upper Dantian. Imagine the body absorbing these energies through the Yin Tang point (Third Eye), into the pineal gland at the center of the brain. Imagine the light of the Upper Dantian getting brighter as the energy begins to increase (Figure 41.14). Repeat this exercise nine times.

Upon completion, focus on the Upper Dantian and begin to gather and fuse the body's Yuan Shen along with the collected Heaven and Earth energies into one energy. As this energy collects, imagine it forming into an energetic ball of white light and allow the hands to guide the light downward into the Lower Dantian.

During the process of descending the Qi from the Upper to Lower Dantian, the focus should be on feeling, guiding, listening, and seeing the energy traveling downward. As the energy of the Shen descends, the Jing energy naturally ascends. This descending and ascending action of Jing and Shen energies cultivates an abundant amount of Qi in the Lower Dantian. After the Qi has been gathered and stored there, pause a moment, allowing the energy to settle. Then proceed to regulate the Lower Dantian.

GATHERING QI IN THE LOWER DANTIAN

Both hands sink down to waist level and form an embracing posture in front of the navel. Inhale while slowly drawing the palms inward towards the navel, gathering the Qi into the center of the Lower Dantian. Pause a moment and relax, allow-

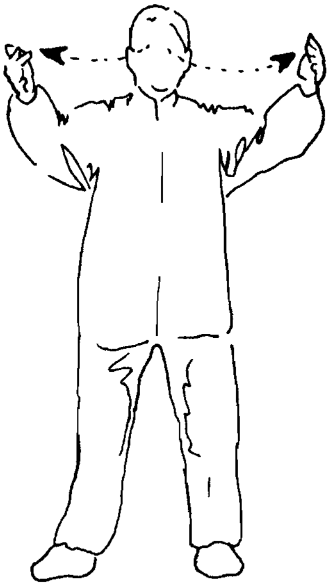

Figure 41.13. Gathering Qi in the Upper Dantian (A)

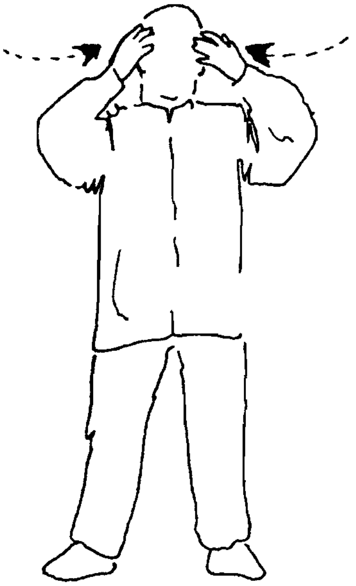

Figure 41.14. Gathering Qi in the Upper Dantian (B)

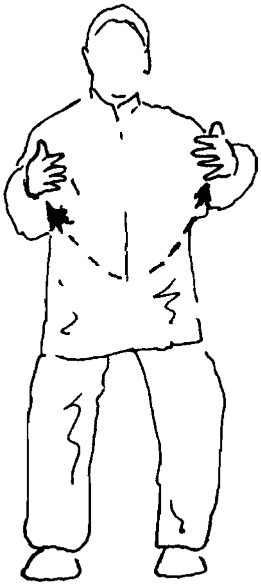

Figure 41.15. Gathering Qi in the Lower Dantian (A)

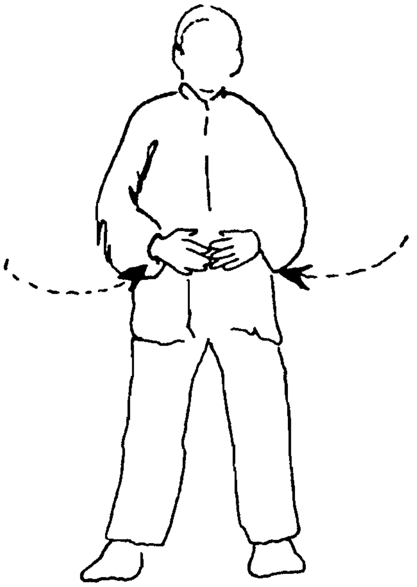

Figure 41.16. Gathering Qi in the Lower Dantian (B)

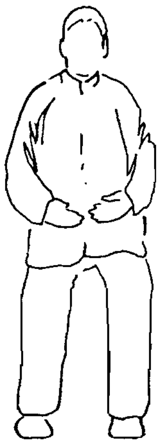

Figure 41.17. Gathering Qi in the Huiyin Area (A)

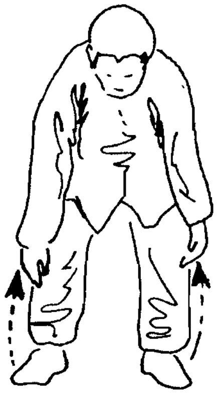

Figure 41.18. Gathering Qi in the Huiyin Area (B)

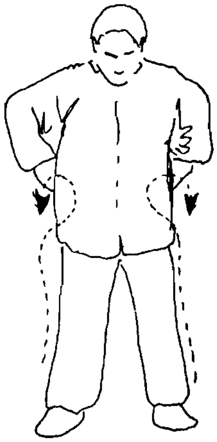

Figure 41.19. Gathering Qi in the Huiyin Area (C)

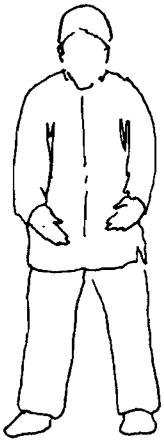

Figure 41.20. Gathering Qi in the Huiyin Area (D)

ing the energy to settle, then exhale and separate the hands (Figure 41.15).

Inhale and imagine the body absorbing the universal and environmental energies in through the navel and Lower Dantian areas (Figure 41.16). Repeat this exercise nine times.

Imagine the Lower Dantian as a reservoir, storing but never overflowing. The expansion and contraction of the abdominal muscles creates a "beating and drumming" movement which improves the abdominal muscles' elasticity while massaging the digestive system. The "beating and drumming" movement strengthens and increases peristalsis.

GATHERING QI INTO THE HUIYIN AREA

Continuing from the last posture, both hands begin to move downward from the Lower Dantian (Figure 41.17).

As you inhale, bend over and imagine scooping the Earth energy along the inside of the legs, directing the Qi through the three Yin channels into the body's Huiyin, coccyx, and Mingmen areas. While inhaling, it is important to pull up on the anal sphincter and imagine gathering the Earth Qi directly through the lower body's orifices (testicles and penis, or vagina) into the Kidneys and Mingmen area (Figure 41.18).

Move both hands from the insides of the thigh, groin, and lower abdomen backward, along the Belt Vessel, expanding the energy into the Kidneys and Mingmen area. Hold this posture and exhale, imagining the Qi flowing into the Kidneys and Mingmen areas (Figure 41.19).

Next, circle the Qi down to the Lower Dantian and Huiyin area. Pause for a moment to allow the Qi to settle and root (Figure 41,20). Repeat this exercise nine times and end this exercise.

ROLLING THE BALL WITH BOTH PALMS

The purpose of this exercise is to regulate the Yin and Yang balance of energy in the body. This exercise also causes the capillaries along the Microcosmic Orbit to open up, improving the Qi circulation and nerve stimulation in the extremities.

It is excellent for purging and regulating emotions from the Heart and Yellow Court, as well as for descending Rebellious Lung Qi downward and purging Yang Qi Rising.

SHAKING THE SUN AND MOON

Begin from a Wuji posture. Both arms raise up in front of the shoulders, palms and fingertips pointing upward. Imagine the left palm embracing the sun and the right palm embracing the moon. Vigorously shake the palms, focusing the mind's intention on the center of the palms and fingers (Figure 41.21).

Place both palms in front of the chest and Middle Dantian area, and imagine the energy of the sun and moon melting into each other, creating a huge ball of energy. Make sure the eyes focus on the center of the ball (Figure 41.22).

ROLLING THE BALL

Wait until you can feel the sensation of the energy ball before beginning the rotations. Imagine the energy ball in the hands and the energy ball in the Lower Dantian blending naturally into one energy, simultaneously rotating and rolling together both internally and externally.

Begin rolling the ball, keeping the mind's intention focused on the center of each palm. Imagine the ball is slowly growing between the hands. Roll the ball, with the hands opposite each other, (one hand is up while the other is down), eighteen times in a clockwise direction, then reverse the direction and repeat another eighteen times (Figure 41.23). After the last rotation, relax the arms and return to Wuji posture.

OPENING AND CLOSING THE UPPER, MIDDLE, AND LOWER BURNERS

This exercise is used to regulate the body's Yin and Yang energy, tonify the Yin organs, calm and harmonize the Yang organs, Regulate the Triple Burners, and help open the Conception and Governing Vessels. This is an excellent exercise for purging and regulating Excess conditions within the Triple Burners, as well as for purging the patient's Liver Fire.

There is an ancient Chinese saying, "Open, close, and come and go, one hundred illnesses will be healed." When opening, the Qi circulates into the muscles. When closing, the Qi sinks and gathers into the bones. The purpose of this exercise is to train the mind's intention, making it easier to increase the body's external field of Wei Qi.

OPENING AND CLOSING THE UPPER BURNER

Begin from a Wuji posture. Both arms raise up to shoulder level. Vigorously shake the hands, palms facing down. Before opening and closing the Upper, Middle and Lower Burner areas, the awareness must settle in the Laogong area (center of the palms) for a few minutes. Wait until heat

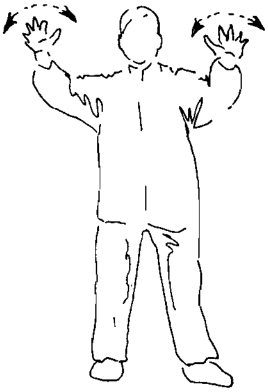

Figure 41.21. Shaking the Sun and Moon (A)

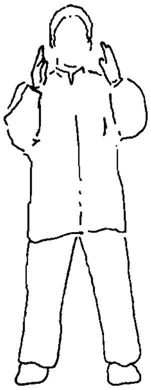

Figure 41.22. Shaking the Sun and Moon (B)

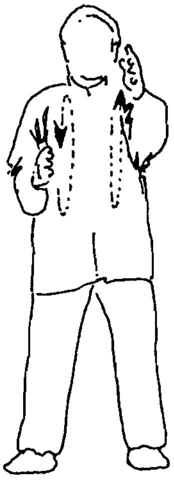

Figure 41.23. Rolling the Ball with Both Palms (A)

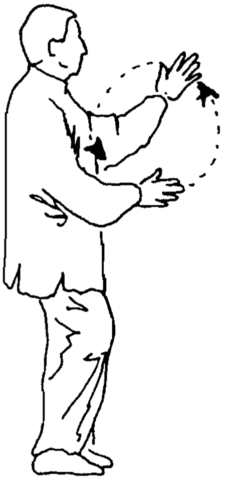

Rolling the Ball with Both Palms (Side View)

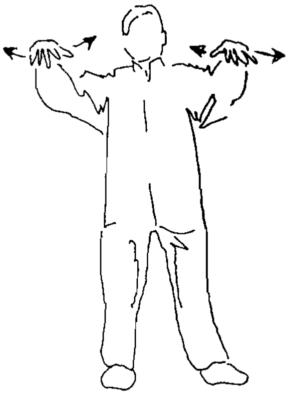

Figure 41.24. Vigorously Shake the Hands

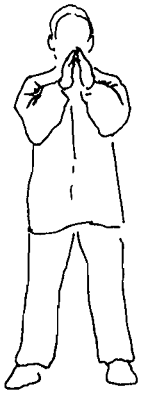

Figure 41.25. Open and Close the Upper Burner (A)

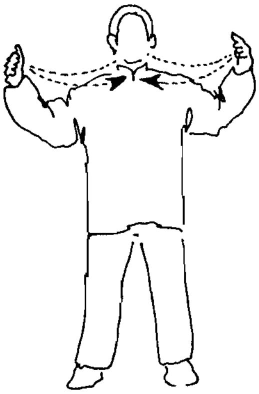

Figure 41.26. Open and Close the Upper Burner (B)

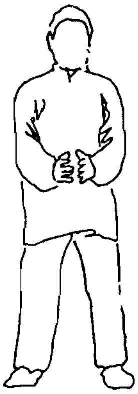

Figure 41.27. Open and Close the Middle Burner (A)

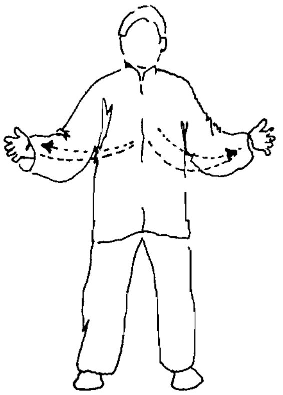

Figure 41.28. Open and Close the Middle Burner (B)

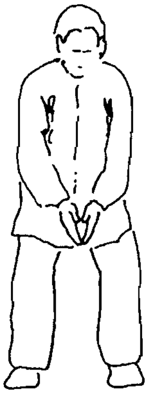

Figure 41.29. Open and Close the Lower Burner (A)

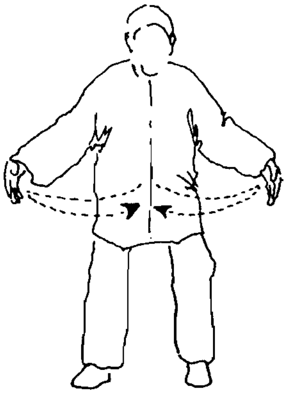

Figure 41.30. Open and Close the Lower Burner (B)

is felt in the center of the palms before beginning (Figure 41.24).

After a while, bring both hands close to each other in front of the throat and manubrium sterni area (the upper segment of the sternum). Both palms face each other, with the fingers facing upwards. Pause a moment, focusing the mind's attention on the center of the palms, forming an energetic ball (Figure 41.25).

Slowly begin to open by pulling the arms apart, while imagining the energy between the center of the palms stretching like warm taffy. When the hands separate and open, the mind and body (chest and abdomen) also open. Slowly begin to close by bringing both palms together, pushing and condensing the stretched energy to form an energetic ball. When the hands close, the mind and body close by removing the attention away from the outside world (Figure 41.26).

When inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. When exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body. Repeat the exercise eighteen times.

OPEN AND CLOSE THE MIDDLE BURNER

Both hands sink down in front of the diaphragm and xiphoid process (just above the navel). Pause a moment, keeping the eyes on the energy ball and an imaginary line which connects the inner Laogong Pc-8 points at the center of the palms (Figure 41.27).

Slowly begin to open by pulling the arms apart, while imagining the energy between the center of the palms stretching like warm taffy. While exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body (Figure 41.28).

Slowly close by bringing both palms together, forming an energetic ball as you push and condense the energy. While inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. Repeat the exercise eighteen times.

OPENING AND CLOSING THE LOWER BURNER

Both hands sink down in front of the reproductive organs and perineal area. Pause a moment, keeping the eyes on an imaginary line which connects the inner Laogong Pc-8 points at the center of the palms (Figure 41.29).

Slowly begin to open by pulling the arms apart, imagining the energy between the center of the palms stretching like warm taffy. When exhaling, separate the hands while imagining the Excess Heat and toxic energy being purged from the body (Figure 41.30).

Slowly close by bringing both palms together, forming an energetic ball by pushing and condensing the Qi. While inhaling, imagine extending the fingers deep into the body and the Qi pouring deep into the tissues. Repeat eighteen times.

To end the exercise, draw the Qi back into the Lower Dantian, and return to a Wuji posture.

SUN AND MOON ROTATING TECHNIQUE

This exercise will massage the internal organs of both the Heart and the Lungs. In Medical Qigong, the sun is sometimes referred to as the energy of the Heart, which is the considered the master of the three Yang energies. The moon pertains to the Lungs, which are the masters of the three Yin energies. Through the harmony of these dual powers the body's Yuan (Original) Qi is born. This exercise increases the elasticity of the Heart muscles and arteries and improves Blood circulation. It also flushes the Eight Extraordinary Vessels, increases vitality, and causes the upper half of the body to become lighter while making the lower half more solid.

This exercise can also be used in conjunction with the healing sound "Shang" to treat Lung diseases (i.e., asthma and chronic bronchitis), as well as breast cysts and tumors by purging Toxic Qi from the tissues of the Lung.

ROTATING THE PALMS

Begin from a Wuji posture, start by vigorously shaking the hands and then rubbing the palms together (Figure 41.31).

Place both hands on the chest, and draw the heat from the center of the palms into the chest, inhaling three times. Once the chest has absorbed a certain amount of heat, both palms begin to ro-

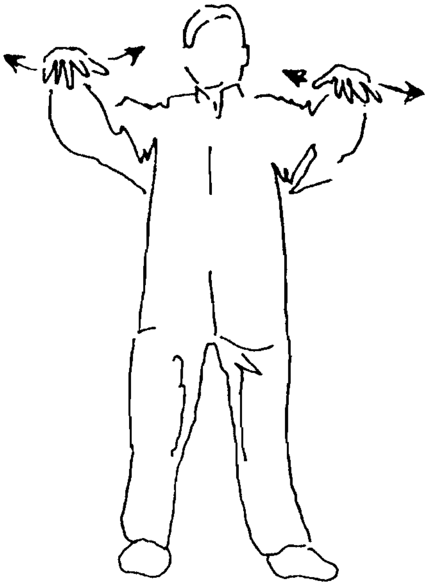

Figure 41.31. Vigorously Shake the Hands

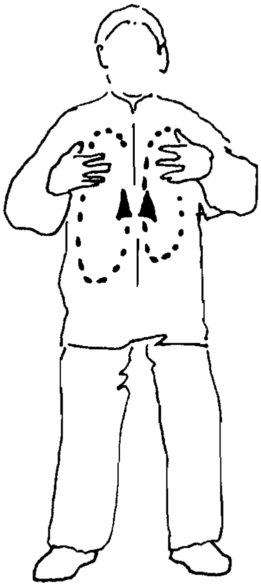

Figure 41.32. Sun and Moon Rotating Technique (A)

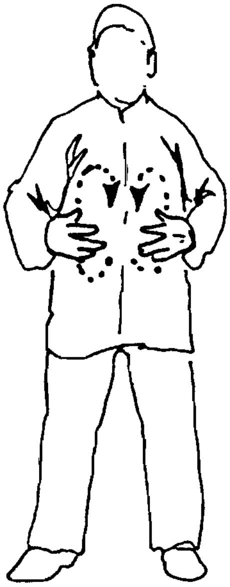

Figure 41.33. Sun and Moon Rotating Technique (B)

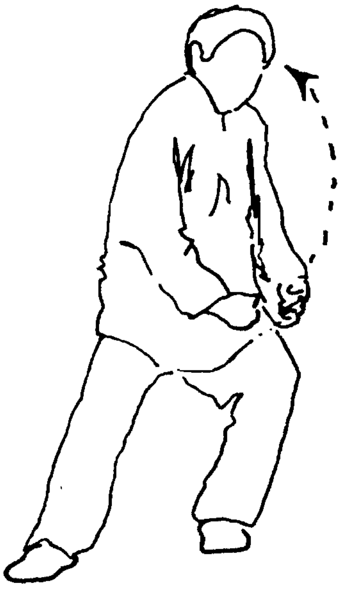

Figure 41.34. Expanding and Contracting the Rings (Forward-A)

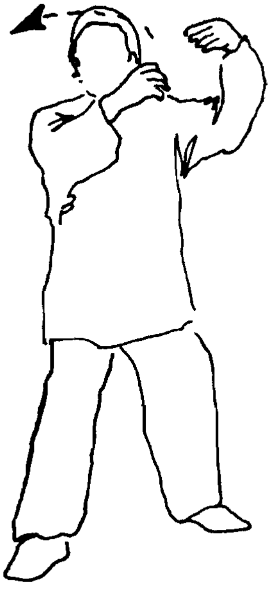

Figure 41.35. Expanding and Contracting the Rings (Forward-B)

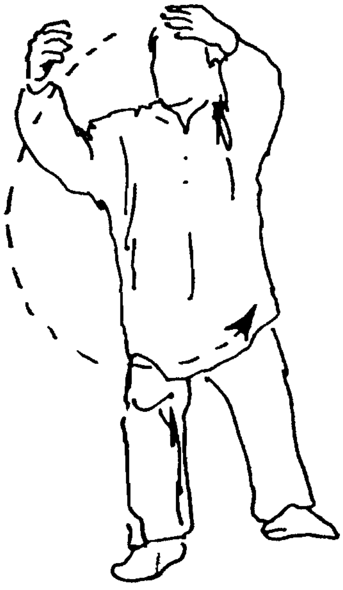

Figure 41.36. Expanding and Contracting the Rings (Forward-C)

tate up the center line of the chest. The rotation of the arms forms two circles, that circulate the energy from the inside to the outside of the breast. The focus of the intention should be on the "chest opening and the back closing" of the thorax. Repeat this exercise 18 to 36 rotations, depending on the severity of the condition (Figure 41.32).

Next, rotate the circles in the opposite direction, moving from the outside towards the center line of the chest and then down, circulating the energy from the outside to the inside of the breast. Repeat 18 to 36 rotations, depending on the severity of the condition (Figure 41.33).

When the rotations are completed, the middle fingers of both hands slightly press the Conception Vessel downward to descend the Qi into the Lower Dantian. Rub the abdomen three times and resume a Wuji posture.

EXPANDING AND CONTRACTING THE RINGS

Expanding and contracting the rings refers to the expansion and contraction of the spinal vertebra. This exercise reinforces the flow of the Microcosmic Orbit, invigorates and strengthens the Yuan Qi, aids in transforming Jing to Qi, and strengthens the entire spine. The patient's coordination and balance are also positively affected.

While performing this exercise, the body should move like a "dragon playing in the ocean waves" by appearing and disappearing, rising and sinking, and twisting and coiling. The goal is to stretch and loosen the spinal column, rib articulations, and the back. This will cause the body's Fire to descend and Water energy to ascend naturally through the body's center channels.

This exercise can be modified into a stationary position for convalescing patients or patient's confined to a wheelchair. The patient bends forward, allowing the spine to slowly stretch, shifting the body to a forty-five degree angle.

Begin this exercise from a right Bow Stance (weight is shifting 70% to the back, 30% forward) with both hands positioned in front of the Lower Dantian (Figure 41.34).

Both hands raise upward and stretch forward, creating an arc and stretching the spinal column. When extending and stretching the spine forward, imagine the Earth Qi traveling up the back leg, through the spine and Governing Vessel to the top of the head (Baihui point), and extending down to the Yin Tang and both Laogong points at the center of the palms (Figure 41.35).

While extending the energy outward, use intention to connect with the Heavenly Qi, the energy of the horizon, or a tree (Figure 41.36).

As both hands withdraw, shift the weight to the back leg, allowing the spinal column to contract. As you shift your weight back to the starting position, imagine drawing Qi up from the Earth into your body. Repeat nine times. End in the Wuji posture. Then switch directions by switching the position of the feet and repeat the entire sequence while reversing the rotation.

To perform the second half of the exercise, shift the body's weight forward and begin moving the body in the reverse circular direction (Figure 41.37).

Imagine the Heaven Qi traveling from the Baihui point down the Conception Vessel and out the front foot, extending deep into the Earth (Figure 41.38).

As the weight shifts back, imagine drawing Qi from the Heavens back into the body through the head and hands (Figure 41.39).

Sink the Qi deep into the Earth allowing the body to be cleansed and purified (Figure 41.40). Repeat the exercise nine times. End in the Wuji posture. Then switch directions and repeat the entire sequence.

DESCENDING THE YANG AND ASCENDING THE YIN TECHNIQUE

This exercise cultivates Qi in the channels and collaterals of the lower body, aids the free flow of Qi to the Lower Dantian, strengthens the Kidneys, Spleen, Stomach, and the Liver Qi, while regulating the Gall Bladder Qi.

It is an excellent exercise for pulling the Earth Qi (Spleen Qi) up while sinking the Stomach Qi, and is good for treating problems such as irritable bowel syndrome and problems caused from Excess Heat in the Middle and Lower Burners. In order to treat such conditions, the exercise should be performed for twenty minutes, twice a day.

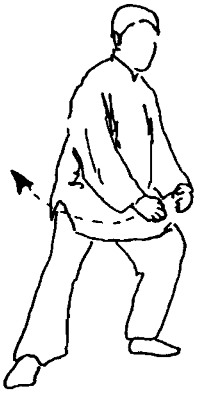

Figure 41.37. Expanding and Contracting the Rings (Reverse-A)

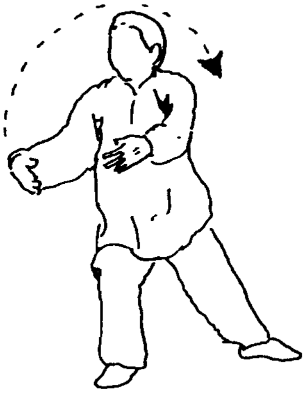

Figure 41.38. Expanding and Contracting the Rings (Reverse-B)

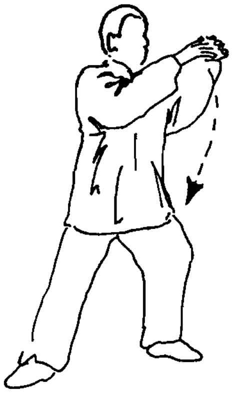

Figure 41.39. Expanding and Contracting the Rings (Reverse-C)

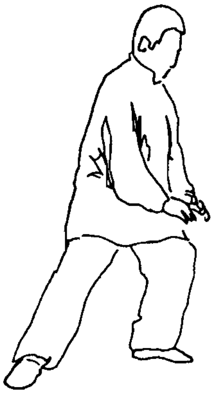

Figure 41.40. Expanding and Contracting the Rings (Reverse-D)

This exercise can furthermore be used in conjunction with the healing sound "Yu" to treat reproductive organ diseases (i.e., uterine and ovarian cysts, tumors, and cancer, as well as prostate cancer). The patient should practice a total of one to four hours each day, depending on the severity of the condition.

Begin from a Wuji posture and embrace the abdomen (Figure 41.41).

Gather Qi into the Lower Dantian and turn towards the left, both hands move into the Huantiao GB-30 point at the side of the hip (Figure 41.42).

Both hands slowly descend the three Yang channels down the outside of the left leg with a "pushing" action, ending on the outside of the left foot at the Bubbling Well point (Kd-1) at the bottom of the foot. As both hands reach downward to the foot, stretch the body and allow the waist to follow the descending Qi with a downward crouching movement (Figure 41.43).

Next, shift the hands around the foot and begin ascending the three Yin channels with the hands inside the left leg. Continue the movement all the way up into the Huiyin point, pubic bone, and Lower Dantian area with a "pulling" and "drawing" action. Pull the anal sphincter upward while drawing the Earth Qi into the body. When both hands raise and gather Qi upward, the body and waist must follow the ascending and rising movements (Figure 41.44).

Pause for a few seconds and let the Qi settle into the Lower Dantian (Figure 41.45).

Shift the body's weight, and repeat the entire sequence on the opposite leg. The turning of the waist should match the movements of the thighs. The toes should naturally grasp the ground and the soles of the feet should feel full of Qi. Repeat nine times on each leg, and end in a Wuji posture.

End the exercise by ascending and descending the Earth Qi through both legs.

From a Wuji posture, both hands embrace the abdomen to gather the Qi of the Lower Dantian and to separate the energy into two balls of energy (Figure 41.46).

Both hands then slowly descend the outside Yang channels of each thigh, all the way to the outside of the feet, ending at the Bubbling Well points (Figure 41.47).

Move the hands along the front of the feet, toward the inside of the foot, slowly ascending the inside Yin channels of each thigh all the way up the body into the Huiyin and Lower Dantian areas (Figure 41.48)

Pause and allow the Qi to settle in the Lower Dantian, then begin again (Figure 41.49). Repeat

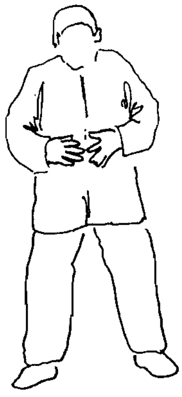

Figure 41.41. Descend the Yang and Ascend the Yin Technique (Single-A)

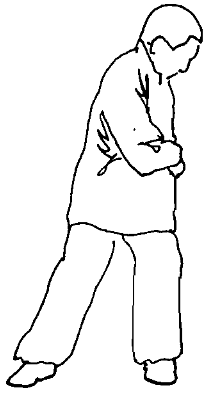

Figure 41.42. Descend the Yang and Ascend the Yin Technique (B)

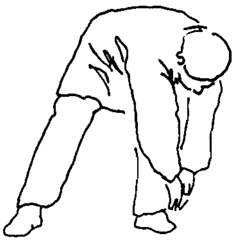

Figure 41.43. Descend the Yang and Ascend the Yin Technique (C)

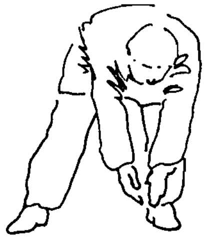

Figure 41.44. Descend the Yang and Ascend the Yin Technique (D)

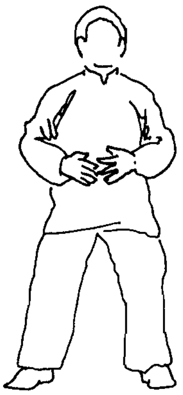

Figure 41.45. Descend the Yang and Ascend the Yin Technique (E)

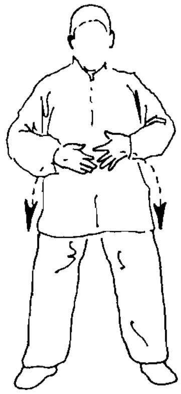

Figure 41.46. Descend the Yang and Ascend the Yin Technique (Double-A)

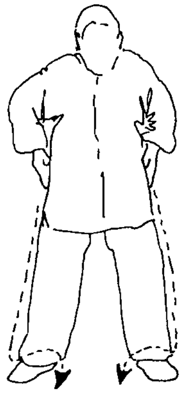

Figure 41.47. Descend the Yang and Ascend the Yin Technique (B)

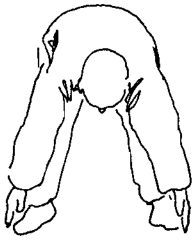

Figure 41.48. Descend the Yang and Ascend the Yin Technique (C)

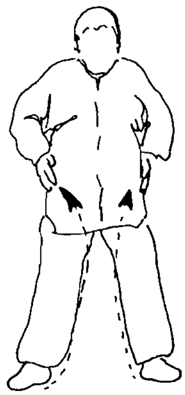

Figure 41.49. Descend the Yang and Ascend the Yin Technique (D)

this exercise 18 times, then return to the Wuji posture and end the exercise.

TURNING AND WINDING THE BELT VESSEL TECHNIQUE

This tonifying exercise is divided into two methods: the Small Turning and Winding, and the Large Turning and Winding techniques. Each method of exercise causes a different action between the rotation of the energy's circle, the form's circle, and the internal turning action of the Lower Dantian.

Practice the Small Turning and Winding technique first, then proceed to the Large Turning and Winding technique. The small method uses a 180 degree semicircular rotation in addition to the Belt Vessel and Lower Dantian circle rotation.

When practicing the Large Turning and Winding method, the patient's mind must imagine a 360 degree circle rotation of the Belt Vessel. The body moves externally in one direction, while the energy in the Lower Dantian rotates in the reverse direction.

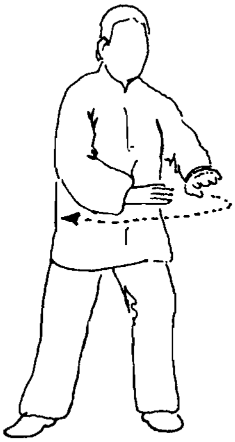

Figure 41.50. Small Turning and Winding (A)

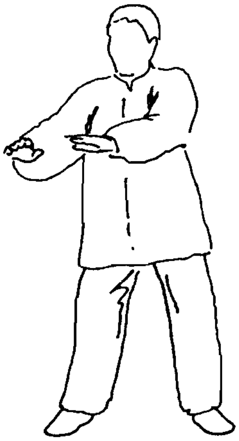

Figure 41.51. Small Turning and Winding (B)

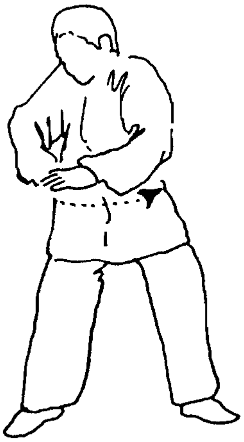

Figure 41.52. Small Turning and Winding (C)

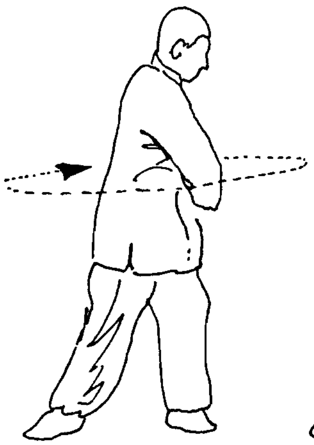

Figure 41.53. Large Turning and Winding (A)

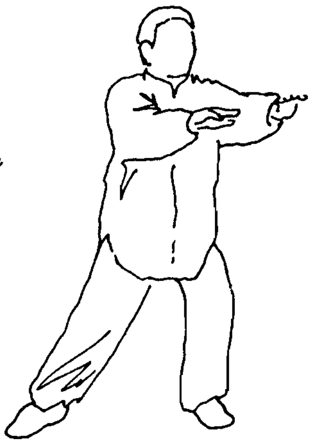

Figure 41.54. Large Turning and Winding (B)

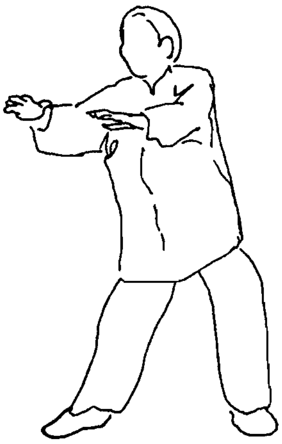

Figure 41.55. Large Turning and Winding (C)

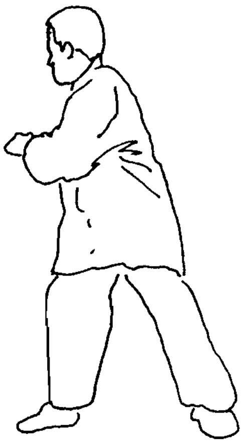

Figure 41.56. Large Turning and Winding (D)

The back and forth, side to side movements of the Turning and Winding action move the body's Qi, and stimulate the energy moving within the Belt Vessel's Qi circle. This energy circle fuses with the external rotation of the hands, allowing the energy from both external and internal Qi circles to be absorbed into the Lower Dantian. Once absorbed into the Lower Dantian, the Qi expands into the body's center core and blends with the Yuan Qi. This strengthens the waist, spine, and enables the Middle Burner's Qi to pervade through the upper and lower parts of the body. This exercise also develops strong and vigorous Kidneys and thus promotes the transportation of the Kidney's Jing and Qi.

SMALL TURNING AND WINDING TECHNIQUE

From a Wuji posture, step forward with the left foot to form a Leaning Horse stance. Raise both palms to waist level, in front of the Lower Dantian, with the palms facing the ground (Figure 41.50).

Inhale and shift the body forward towards the right direction and begin to rotate, turning and twisting the body in a clockwise direction. Imagine gathering the Earth's Qi into the Lower Dantian (Figure 41.51)

Exhale, but allow the mind to focus on absorbing the Earth Qi into the Lower Dantian. Pause for a minute, allowing the Qi to settle after filling the Lower Dantian (Figure 41.52). After nine circle rotations switch directions, moving in the opposite direction nine times.

LARGE TURNING AND WINDING TECHNIQUE

Bring the hands to the right side of the Mingmen and right Kidney area, while placing the mind's intention onto the Lower Dantian and navel area (Figure 41.53).

Inhale, as you extend both hands forward, and begin gathering the Earth's Qi (Figure 41.54).

Begin to circle rotate a 270 degree turn towards the left, while internally the energy of the Lower Dantian circles in the opposite direction to the right (Figure 41.55). Imagine the navel to be the center of the external circle and the Belt Vessel to be the internal circle. The mind's intention leads the energy which guides the hands and moves the body and the waist. The hand movements and center of weight shifting must be synchronized.

Exhale and gather the Earth's Qi into the left Kidney and Mingmen area; repeat nine times. Allow the Qi to fill the Mingmen and Kidneys, then switch directions for nine times (Figure 41.56). End in the Wuji position.

OPENING AND CLOSING THE HEAVEN AND EARTH TECHNIQUE

The purpose of this exercise is to cultivate and strengthen the Lower Dantian energy by blending it together with the Heaven and Earth Qi using the Beating and Drumming the Qi method of Qi cultivation. This cultivation method roots and stabilizes the body, causing the Five Hearts to return to their origin. The Five Hearts (sometimes called the Five Palms) are as follows:

- The Heart of the Baihui area (the Crown Chakra)
- The Hearts of each Laogong area (the palm Chakra)
- The Hearts of each Yongquan area (the bottom of each foot Chakra)

When the energy of the Lower Dantian disperses through the body, as the Mingmen pushes back, the Yuan Qi returns and gathers into the navel. This gathering and extending, back and forth action of energy of the Five Hearts transforms the body's Yuan Qi, strengthens the Yin and Yang energy, opens the body's channels, and encourages the merging of the Heart's Fire and the Kidneys' Water Qi. This method is divided into two techniques: Open the Heaven and Close the Earth, and Open the Earth and Close the Heaven techniques.

OPEN THE HEAVEN AND CLOSE THE EARTH

Begin from a Wuji posture with the left hand on the Lower Dantian and the right hand resting on top of it, both Laogong points face the navel (Figure 41.57).

Keep the mind's intention focused on the navel as the center base. Inhale as both arms raise up the center line of the body (Figure 41.58).

Extend the intention into the Heavens and begin to gather the universal Qi (Figure 41.59). Imagine Heavenly Qi descending and completely

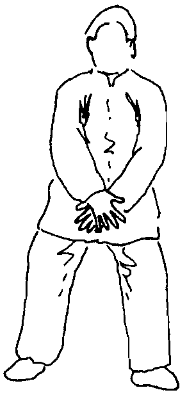

Figure 41.57. Open the Heavens and Close the Earth (A)

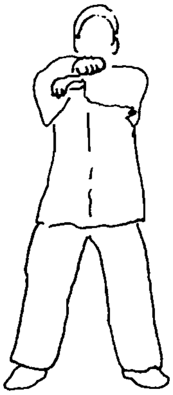

Figure 41.58. Open the Heavens and Close the Earth (B)

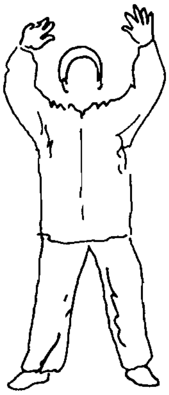

Figure 41.59. Open the Heavens and Close the Earth (C)

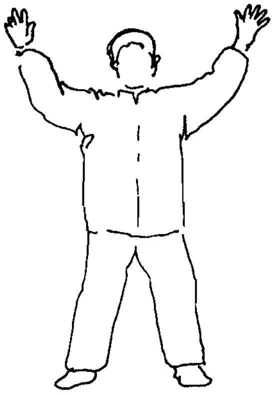

Figure 41.60. Open the Heavens and Close the Earth (D)

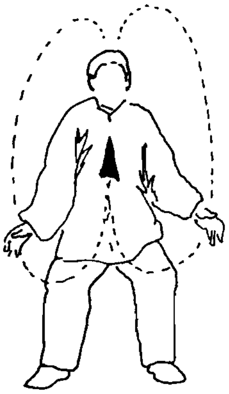

Figure 41.61. Open the Heavens and Close the Earth (E)

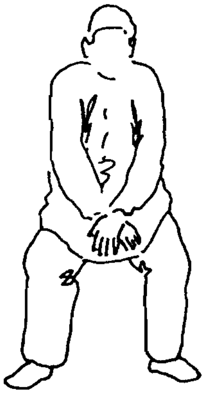

Figure 41.62. Open the Heavens and Close the Earth (F)

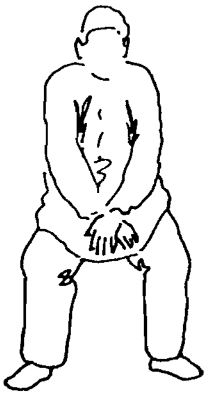

Figure 41.63. Close the Heavens and Open the Earth (A)

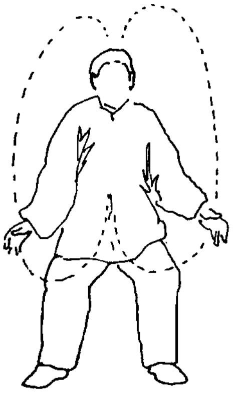

Figure 41.64. Close the Heavens and Open the Earth (B)

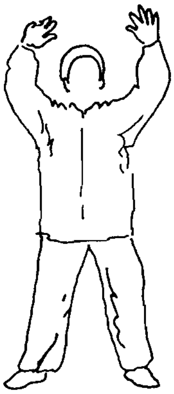

Figure 41.65. Close the Heavens and Open the Earth (C)

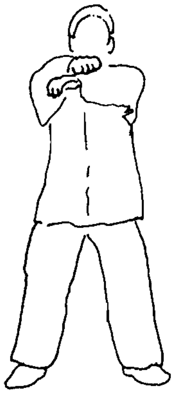

Figure 41.66. Close the Heavens and Open the Earth (D)

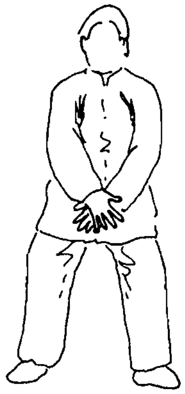

Figure 41.67. Close the Heavens and Open the Earth (E)

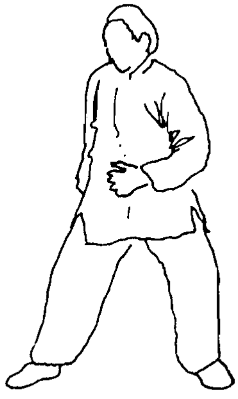

Figure 41.68. Collect the Qi and Return it to the Lower Dantian (A)

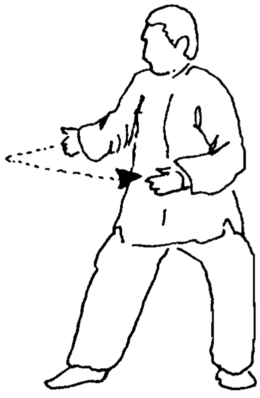

Figure 41.69. Collect the Qi and Return it to the Lower Dantian (B)

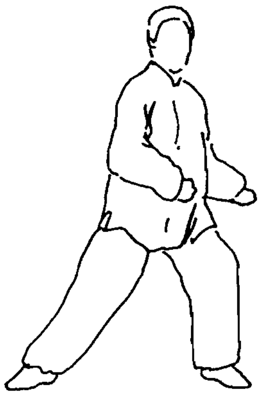

Figure 41.70. Collect the Qi and Return it to the Lower Dantian (C)

filling the body (Figure 41.60). Exhale and gather the Heavenly Qi deep into the body's center core (Figure 41.61).

Lead the energy from inside of the body's center core to the outside, imagining Heavenly Qi, like steam, releasing through the pores and filling up the body's external energy bubble like a mist (Figure 41.62). Repeat 18 times.

OPEN THE EARTH AND CLOSE THE HEAVENS

This is the exact reverse action of the Open the Heaven and Close the Earth technique. Begin from a Wuji posture with the left hand on the Lower Dantian and the right hand resting on top of it, both Laogong points face the navel (Figure 41.63).

Inhale and begin to separate the hands, moving them to the outside of the body (Figure 41.64). Imagine gathering the Earth's Qi into the body as the hands raise upwards towards the head (Figure 41.65). Exhale and gather the Earth's Qi deep into the body's center core (Figure 41.66).

Lead the energy from inside of the body's center core to the outside, imagining the Earth's Qi, like steam, releasing through the pores and filling up the body's external energy bubble like a mist (Figure 41.67). Repeat 18 times, then pause and relax, returning to the Wuji posture.

COLLECT THE QI AND RETURN TO THE LOWER DANTIAN TECHNIQUE

The purpose of this exercise is to gather both universal and environmental energy and to store it in the navel. This exercise is used for treating Deficient conditions.

COLLECT AND RETURN

Begin from a Wuji posture and step out to the right, forming a Horse Stance. The left hand swings to embrace the navel and Lower Dantian as the body shifts its weight and twists toward the right (Figure 41.68).

The body slowly begins to shift from side to side while the opposite hand gathers the universal and environmental energy, gathering and packing it into the navel and Lower Dantian area (Figure 41.69).

During the exchange of left and right hands collecting Qi, imagine that the body is swimming in water. Imagine the energetic consistency of the air, like water, following each arm as it moves, flowing into the center of a whirlpool located in the navel and Lower Dantian area (Figure 41.70). Repeat 18 times. End in Wuji posture.

QIGONG SELF-HEALING MASSAGE AND POINT THERAPY

Qigong Massage and Point therapy is usually practiced after the Qigong meditations and exercise prescriptions. It involves having the patient lightly massage or tap the skin, while extending Qi into the various channels, points, and tissues. These self-healing massage techniques are specifically designed to disperse stagnations and Excess conditions, as well as to activate and regulate the energetic flow within the body's external and internal channels, thereby increasing the energy flow.

These massage techniques can be practiced from either a lying, sitting, or standing posture.

WASH THE FACE

This exercise stimulates the skin and facial muscles. Begin from a Wuji posture. Shake the hands vigorously to remove any External pathogenic Qi. Then slap the palms and rub the hands together vigorously until they get hot. Gently place both hands over the eyes and inhale, absorbing the heat into the face and Upper Dantian. While inhaling, allow the Qi to enter the Taiji Pole.

Exhale and imagine the energy gathered in the Upper Dantian descending down the center of the body, flowing into the Lower Dantian. Perform this gathering and descending action three times.

Next, both hands begin to wash the face, in an outward circular motion. Allow the hands to massage the face, up from the nose, eyes, and forehead, down along the temples, cheeks, and across the chin and mouth, returning back to the nose. Perform this face massage technique for 9 breaths.

Shake the hands vigorously to remove any External pathogenic Qi, then proceed to the next exercise.

MASSAGE THE HEAD

Guide all ten fingers over the head and down over the back of the neck (this is called Ten Dragons Run Through the Forest). Repeat nine times. This can be performed either by touching or not touching the head (Figure 41.71).

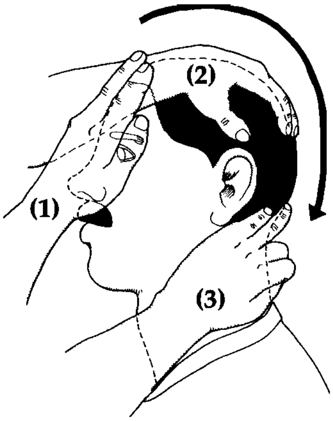

Figure 41.71. Wash the Face and Massage the Head: After rubbing the hands to heat the palms, place both hands on the face and draw the heat in through the eyes and into the Upper Dantian (1). Next, comb the hair with the fingertips (2), and drain the excess Qi from the head and neck area (3).

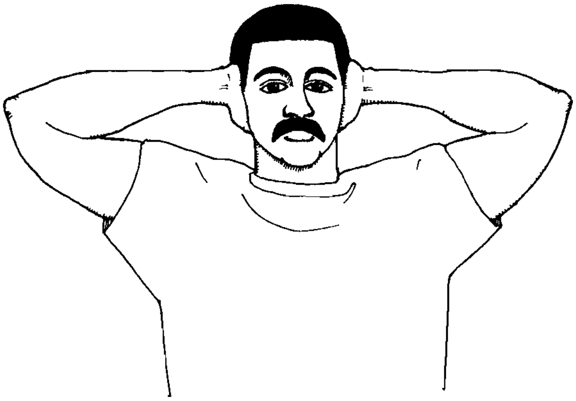

Figure 41.72. Beating the Heavenly Drum

BEATING THE HEAVENLY DRUM

This exercise balances the air pressure in the eustachian tubes and ear canals, relieves tinnitus, helps prevent vertigo, relaxes the mind, helps to improve hearing, expels stagnant Qi collected at the back of the head (within the Jade Pillow), and also stimulates the pineal gland and Kidneys' Qi.

Begin by interlacing the fingers and place both hands on the back of the head. The palm of each hand should cover the ears. While compressing the ears, close your eyes and practice the Beating and Drumming thumping method at the back of the head in the rhythmic sequence of a heart beat. This causes relaxation and creates an altered, in-utero state of consciousness. The index fingers of each hand should thump the Jade Pillow (Bl-9) at the back of the occiput pass 36 times (Figure 41.72).

After thumping the back of the head, clench the teeth, close the mouth, and bend the body over. Exhale while bending over and bring the head between the legs. The legs are straight, waist and buttocks are relaxed, and both eyes look towards the back from in-between the legs, while holding the breath (Figure 41.73).

Next, raise the body up, inhale and again perform the Beating the Heavenly Drum sequence. Practice this exercise 14 times.

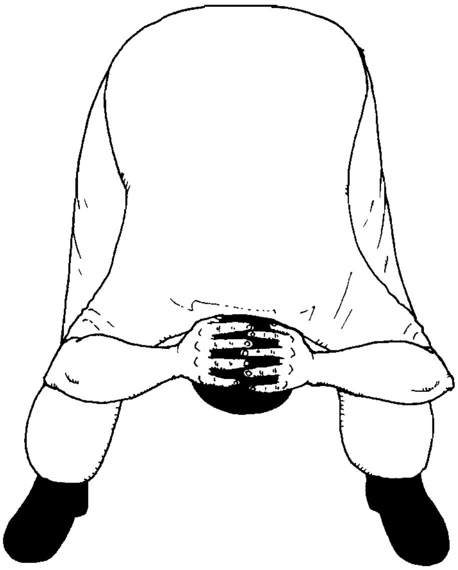

Figure 41.73. Bend the Body Over While Holding the Breath

PRESSING THE EARS

This exercise is used to enhance hearing, stimulate mental energy and also to stimulate the Kidneys' Qi. Use the Laogong areas at the center of each palm to compress the ears (like a plunger). Repeat this exercise nine times (Figure 41.74).

Next, move the right hand behind the head and pull the left ear (at a 45 degree angle) towards the back, with the left hand is placed on the Lower Dantian. Hold this posture for 9 breaths, then reverse for 9 breaths.

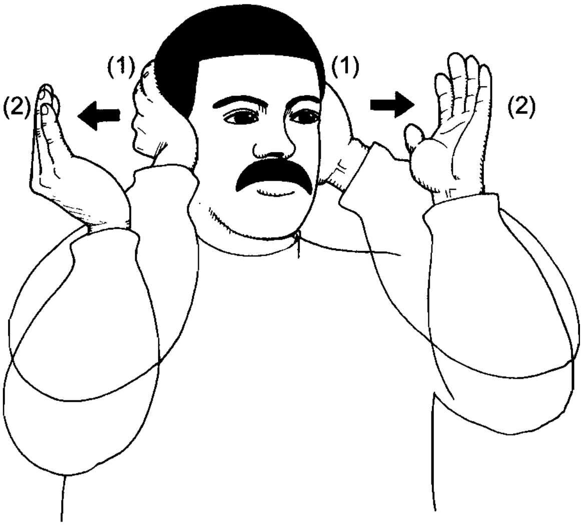

Figure 41.74. Pressing the Ears

RUBBING THE EYES

RUBBING THE EYES (SEQUENCE #1)

This exercise is good for enhancing Qi and Blood circulation to the eyes for improving vision, treating eye diseases (especially glaucoma), relieving fatigue (caused by eye strain), and also for stimulating the Liver Qi.

With both thumbs resting on the cheekbones, form two soft fists and, using the edge of the index knuckles, begin to massage in a circle the top of the eyelids (from the bridge of the nose out along the eyebrows). Next massage just bellow the bottom of the eyelids from the bridge of the nose. Alternate from top to bottom 18 times (Figure 41.75).

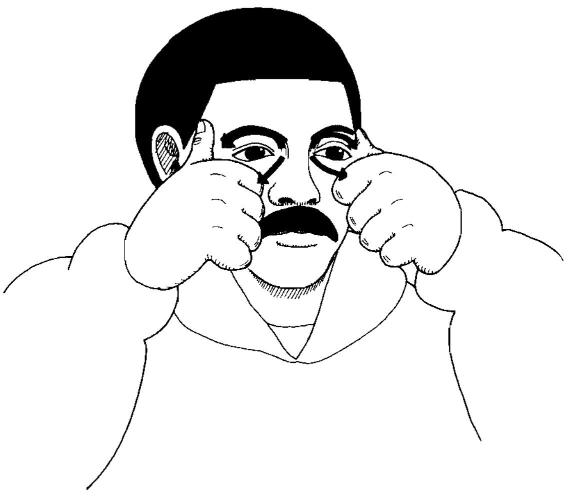

Figure 41.75. Rubbing the Eyes (Sequence 1)

RUBBING THE EYES (SEQUENCE #2)

With the thumbs still resting on the cheekbones, use the tip of the index fingers to massage the points around the eyes, from the inner corners of the skin, in a circular motion. Start rotating in a small circle, at the Bl-1 point. Circle nine times towards the inside, then nine times towards the outside, for a total of three breaths.

Next massage the middle of the eyebrows at the Extra Point Yu Yao. Then proceed to the outside of the eyes at the GB-1 point. (3) End at the base of the eyes at the St-1 point. Each area should be massaged 18 times (9 times towards the inside, then 9 times towards the outside) (Figure 41.76).

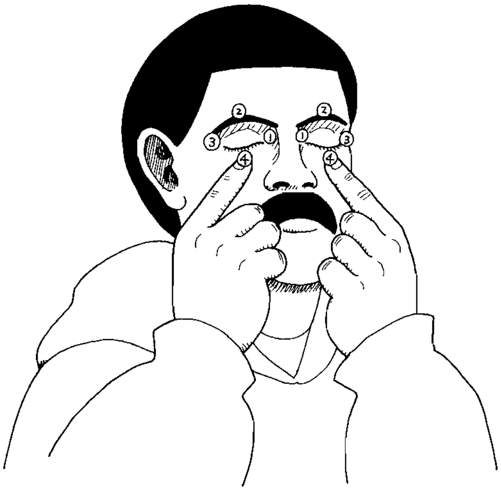

Figure 41.76. Rubbing the Eyes (Sequence 2)

MASSAGING THE NOSE

This exercise is performed to treat colds, reduce swelling of the sinus tissues, eliminate mucus from the airways, stimulate the sinus nerves, open the nasal passages, and also stimulate the Lungs' Qi.

Begin by making soft fists and use the sides of the thumbs to massage the sides of the nose at the LI-20 points in an up and down motion 9 times pressing inward, and then 9 times pressing outward (Figure 41.77).

Next, use the tips of the index fingers and press deeply at the base of the nose at the LI-20 points, for a period of ten seconds, then circle-massage for 18 rotations. Press deeply midway up the nose, on both sides at the Bitong points for a period of ten seconds, then circle-massage 18 times. Finally, both fingers press deeply at the upper point at the bridge of the nose at the Shangen points for ten seconds, then circle-massage 18 times.

CLICKING THE TEETH

This exercise is good for strengthening the teeth and gums, tonifying the jaw muscles, improving the Qi and Blood circulation to the gum tissues, and also to enhance the Kidneys' Qi. Clamp the jaws shut so that the teeth are pressing together and the jaw muscles flex. Click the teeth 36 times.

WAGGING THE TONGUE

This exercise is good for stimulating the salivary ducts beneath the tongue, and also for stimulating the Heart's Qi. Wag the tongue on the outside gums 18 times in one direction, 18 times in the opposite direction, then 18 times from the lower to upper palate, and finally, 18 times pressing the tip of the tongue against the back of the teeth (Figure 41.78).

SWALLOWING THE SALIVA

Saliva contains many kinds of enzymes which aid in digestion. In ancient China, this practice was sometimes called the Immortal's Water or Juice of Jade, because the saliva is energized during the Qigong meditation and believed to contain special healing properties.

To begin, inhale and swallow the saliva with a big gulp (three times), using the breath as a seal or bubble to enclose the air, then send it down the Conception Vessel to bathe the Five Yin Organs, ending at the Lower Dantian. It is important to tilt the head slightly forward (after inhaling) before swallowing the Immortal's Water.

WHIRLING THE NAPE AND TURNING THE NECK

The purpose of this exercise is to massage the thyroid and larynx as well as to stretch the muscles and tendons throughout the neck and cervical vertebrae.

To begin, place the back of the hands on the Kidneys. While relaxing, allow the head to relax and naturally hang, slowly rotating its motion to the left and right nine times (Figure 41.79).

Next, suspend the head and slowly twist the nape of the neck to the right and left direction nine times. While performing this exercise, focus the mind's intention on the Lower Dantian. When finishing, focus the mind's attention on the Bubbling Well points at the center of the feet.

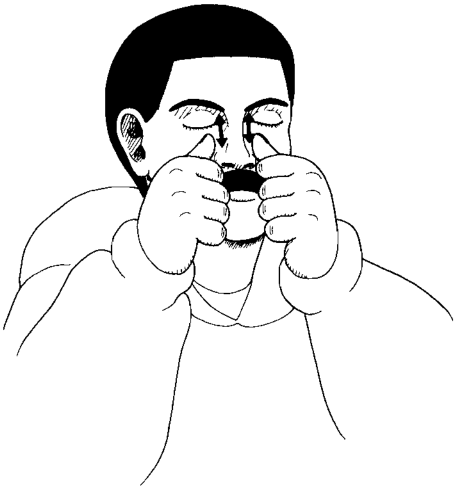

Figure 41.77. Massaging the Nose

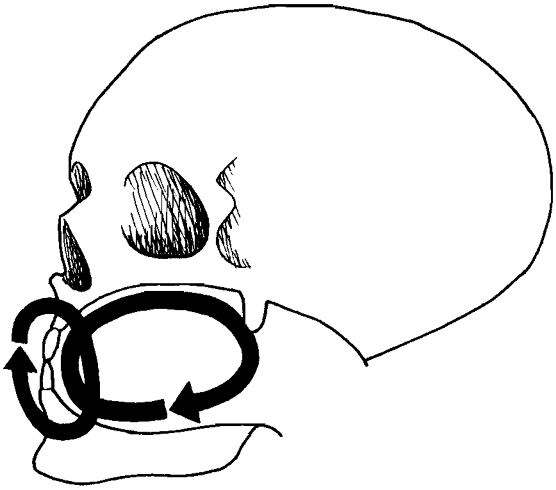

Figure 41.78. Wagging the Tongue

MASSAGING THE CHEST AND ABDOMEN

The purpose of this exercise is to disperse Excess Qi from the chest and abdominal area. This allows the body to balance the energy of the upper torso with that of the lower torso.

Beginning with the mind's attention focused on the Laogong areas at the center of each palm (Pc-8), brush the chest and abdomen downward nine times, starting from the collar bone and ending by the pubic bone (Figure 41.80).

MASSAGE TAPPING THE HANDS (LI-4)

Stimulating the Hegu or Joining of the Valleys points. The purpose of this exercise is to clear and activate the channels and collaterals, regulate, tonify and promote the circulation of Qi, to dispel pathogenic Wind, clear Fire and Heat, and to tranquilize the mind.

Begin by extending both arms in front of the body and allow both hands to collide into each other, thus tapping the Hegu points located between the thumb and first finger. Repeat nine times (Figure 41.81).

MASSAGE TAPPING THE FOREARMS (PC-6)

Stimulating the Neiquan or Inner Border Gate points. The purpose of this exercise is to regulate and tonify the Qi and Blood of the Heart and Stomach, to clear the Heart Fire, transform the Heart Phlegm, and to calm the Heart.

Begin by forming a hollow fist with the right hand and then lightly tap the Neiquan point located three fingers up from the wrist fold, on the lower left forearm area, nine times. Switch arms and repeat (Figure 41.82).

MASSAGE TAPPING THE ELBOWS (LI-11)

This exercise stimulates the Quchi or Crooked Pond points. The purpose of this exercise is to help dispel pathogenic Wind from the body, to regulate the Stomach and intestines, eliminate stasis, relax the muscles and tendons, lubricate the joints, and to cool the Heat in the Blood.

To begin, the right hand forms a hollow fist

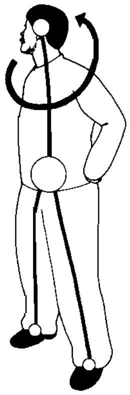

Figure 41.79. Whirling the Nape and Turning the Head

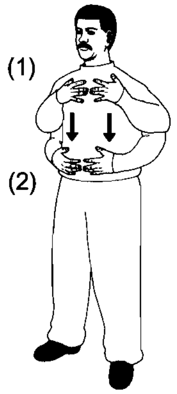

Figure 41.80. Massaging the Chest and Abdomen

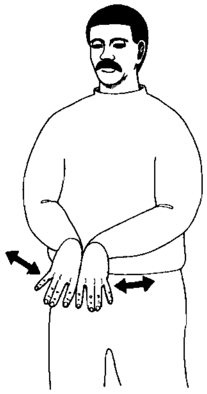

Figure 41.81. Massage Tapping LI-4

and lightly taps the Quchi point located at the end of the left elbow-fold, nine times. Switch arms and repeat (Figure 41.83).

MASSAGING UNDER THE ARMS (HT-1)

This exercise stimulates the Jiquan or Supreme Spring points. The purpose of this exercise is to regulate the Qi of the Lungs and Heart, promoting the circulation of Qi and Blood, as well as to regulate the blood vessels.

To begin, the right hand reaches under the left arm pit and circle massages nine times. Switch sides and repeat nine times (Figure 41.84).

MASSAGE TAPPING THE UPPER SHOULDERS (GB-21)

This exercise stimulates the Jianjing or Shoulder Well points. The purpose of this exercise is to regulate the body's Qi activities, disperse Liver Qi stagnation, extinguish Liver Wind, dispel Wind and Cold, redirect Rebellious Qi downward, and to dredge Excess Qi from all the Yang channels.

To begin, the right hand forms a hollow fist or soft palm and taps the left Jianjing point located on the upper ridge of the neck, while the left hand alternately taps the right Jianjing point. Both left and right tapping count as one set. Perform nine times (Figure 41.85).

MASSAGING THE ARMS

This exercise stimulates the six channels of the arms (the three Yang channels located on the outside, three Yin channels located on the inside).

Begin by placing the right palm on the inside of the left shoulder. Exhale and rub the entire arm from inside the shoulder and arm to the fingers. Next inhale and rub the entire outside of the arm from the fingers to the outside of the shoulder. Continue this movement for 18 breaths, then repeat on the other side (Figure 41.86).

MASSAGE TAPPING THE LOWER DANTIAN (CV-6)

This exercise stimulates the Qihai or Sea of Qi points. The purpose of this exercise is to energize and tonify the lower abdominal organs, as well as to stimulate the Qi of the Lower Dantian.

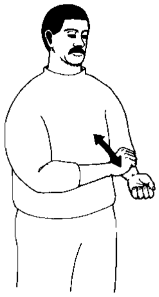

Figure 41.82. Massage Tapping Pc-6

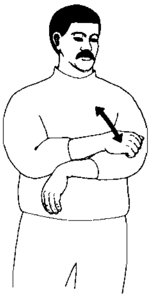

Figure 41.83. Massage Tapping LI-11

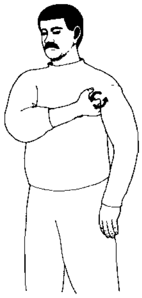

Figure 41.84. Massaging Under the Arm Ht-1

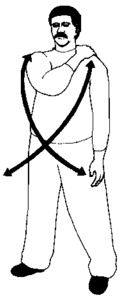

Figure 41.85. Massage Tapping GB-21

Begin by stroking the chest, gathering the body's Qi down into the Lower Dantian. Using either soft fists, palms, or fingers, both hands lightly tap the Lower Dantian and navel area for three breaths (Figure 41.87).

MASSAGE TAPPING THE BACK OF THE WAIST (GV-4)

This exercise stimulates the Mingmen or Gate of Life point. This exercise stimulates and tonifies the Kidneys, adrenal glands, and the Mingmen. It is used to regulate the Governing Vessel, excite and strengthen Kidneys' Yang Qi, to replenish the Kidneys, as well as to strengthen the lower back and bones.

Begin by slightly leaning the body forward. Using the right and left soft palms or hollow fists, lightly tap the back of the waist, Mingmen and the Kidneys' area for three breaths (Figure 41.88).

MASSAGE TAPPING THE SIDES OF THE HIPS (GB-30)

This exercise stimulates the Huantiao or Jumping Circle points. The purpose of this exercise is to dispel Wind and Cold, reduce pain in the thighs and legs, as well as to strengthen the lower back, and tonify the lower extremities.

Using the right and left hollow fists or soft palms, lightly tap the hip bone Huantiao points nine times (Figure 41.89).

MASSAGE TAPPING THE THIGHS (GB-31)

This exercise stimulates the Fengshi or Wind's Market points. The purpose of this exercise is to dispel Wind and Cold, clear Heat, tonify the legs, reduce flaccidity and numbness of the lower extremities, and to transform Dampness.

Using the right and left hollow fists or soft palms, lightly tap the upper thighs at the Fengshi Points nine times (Figure 41.90).

MASSAGE TAPPING THE OUTSIDE OF THE KNEES (GB-34)

This exercise stimulates the Yanglingquan or Yang Mound Spring points. The purpose of this exercise is to relax the muscles and tendons, acti-

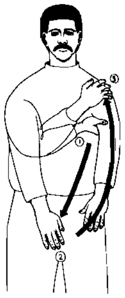

Figure 41.86. Massaging the Arm Channels

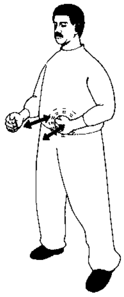

Figure 41.87. Massage Tapping the Lower Dantian

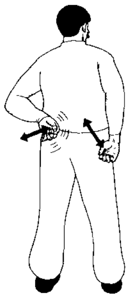

Figure 41.88. Massage Tapping the Mingmen GV-4

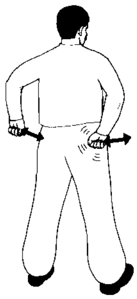

Figure 41.89. Massage Tapping GB-30

vate the flow of Qi and Blood in the channels and collaterals, relieve spasms and pain, clear away Damp-Heat, and disperse stagnant Liver and Gall Bladder Qi.

Using the right and left hollow fists or soft palms, lightly tap the outside of the knees at the Yanglingquan points nine times (Figure 41.91).

MASSAGE TAPPING THE INSIDE OF THE KNEES (SP-9)

This exercise stimulates the Yinlingquan or Yin Mound Spring points. The purpose of this exercise is to regulate and tonify the Spleen (Yang), Stomach (Yin), and Lower Burner, to resolve Dampness and Damp Heat, as well as to treat knee pain and problems with difficulty in urination.

Using the right and left hollow fists or soft palms, lightly tap the inside of the knees at the Yinlingquan points nine times (Figure 41.92).

MASSAGE TAPPING THE BACK OF THE KNEES (BL-40)

This exercise stimulates the Weizong or Entrusting Middle points. The purpose of this exercise is to relax the muscles and tendons, activate the flow of Qi and Blood in the channels and collaterals, cool the Blood to stop bleeding, transform Damp Heat in the Bladder and intestines and to strengthen the waist and knees.

Using the right and left hollow fists or soft palms, lightly tap the back of the knees at the Weizhong points nine times (Figure 41.93).

MASSAGE TAPPING BELOW THE KNEES (ST-36)

This exercise stimulates the Zusanli or Foot Three Miles points. The purpose of this exercise is to regulate the Stomach and intestines, facilitate the flow of Qi to relieve bloating and distention, to dissipate stagnation and obstructions, to promote the circulation of Ying Qi, to alleviate water retention and subdue swelling, and to activate the flow of Qi and Blood in the channels and collaterals.

Using the right and left hollow fists or soft palms, lightly tap the lower leg Zusanli points nine times (Figure 41.94).

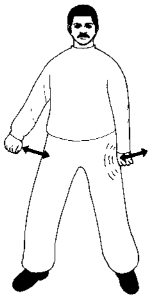

Figure 41.90. Massage Tapping GB-31

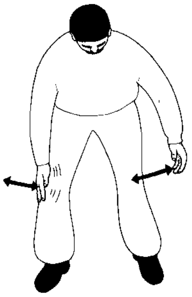

Figure 41.91. Massage Tapping GB-34

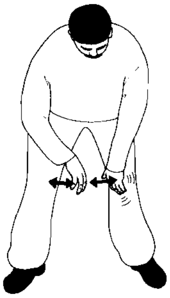

Figure 41.92. Massage Tapping Sp-9

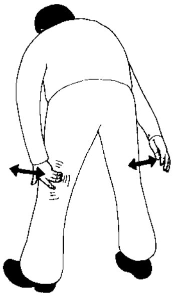

Figure 41.93. Massage Tapping Bl-40

MASSAGE TAPPING ABOVE THE MEDIAL MALLEOLUS (SP-6)

This exercise stimulates the Sanyinjiao or Three Yin Crossing points. The purpose of this exercise is to tonify and regulate the Qi and Yang of the Spleen, as well as regulate the Qi of the Stomach, Middle and Lower Burners, reduce digestive stagnation, and to facilitate Blood flow.

Using the right and left hollow fists or soft palms, lightly tap the inside of the lower leg at the Sanyinjiao points nine times. (Figure 41.95).

MASSAGE TAPPING THE HEELS (KD-1)

This exercise stimulates the Yongquan or Bubbling Spring points. The purpose of this exercise is to tonify the Kidneys' Yin and Jing, tranquilize the mind, clear away Heat, purge pathogenic Fire, calm the Liver and to stop pathogenic Wind.

To begin this exercise make sure that the knees are slightly bent. Quickly rise and fall on both heels to stimulate the Yongquan points and shake the back nine times (Figure 41.96).

TREMBLING THE KNEES

The purpose of this exercise is to disperse stagnant Qi, as well as to relax the perineum.

To begin this exercise, relax, then bend and straighten the knees quickly for nine times. Next shake them from side to side quickly in order to tremble and shake the entire body. Repeat nine times (Figure 41.97).

CIRCLING THE ABDOMEN

The purpose of this exercise is to balance the Lower Dantian Qi, stimulate the Stomach and bowels, and enhance digestive peristalsis.

To begin, place the center of both palms on the Lower Dantian, embracing the navel (for men, the left palm is on the top, opposite for women). Focus the mind's intention deep into the Lower Dantian. While using the navel as the center, connect the energy of the Lower Dantian with the center of the palms as they rotate.

Begin to circle 36 times counterclockwise for men (from the patient's point of view), whereas the rotation is clockwise for women. Allow the circle to gradually move from a smaller to a larger

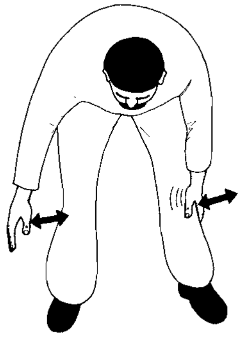

Figure 41.94. Massage Tapping St-36

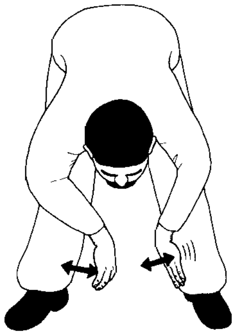

Figure 41.95. Massage Tapping Sp-6

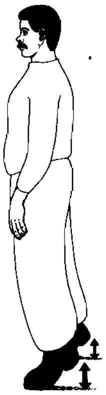

Figure 41.96. Dropping the Heels to Stimulate Kd-1

rotation (as the hands circle, imagine that they are dispersing the Qi). Pause and reverse the circle, rotating 24 times in the opposite direction, gradually reducing the circle's size. This will draw the Qi from the outer portion of the body and collect it into the Lower Dantian. Pause for a moment then return to the Wuji position (Figure 41.98).

COMPLETING THE EXERCISE

Upon completion of the exercise, close the Four Doors (the center of each palm and foot) and cultivate the Qi with stillness, allowing the energy to return to its origin. Hold a quiet Wuji postue for about 15 minutes.

The purpose for the first circle rotation is to release the Yuan Qi which has been collected and gathered in the Lower Dantian. This is to avoid any stagnation which may cause the energy flow to slow down. This also disperses and releases the Yuan Qi, allowing it to blend with the Essential Qi of the Five Yin Organs.

When performing the reverse or second circle rotation, allow all the dispersed energy to gather together and return back to the Lower Dantian. This accumulates and stores the organ energy, blending it together with the Original Qi.

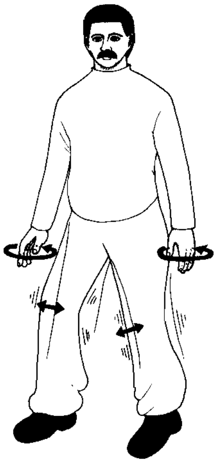

Figure 41.97. Trembling the Knees

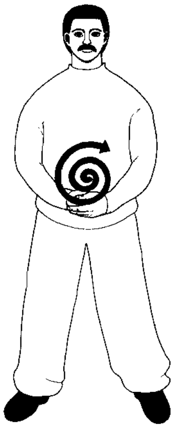

Figure 41.98. Circling the Abdomen to Finish

CHAPTER 42
DAOIST QIGONG REGULATION EXERCISES

This particular school of Daoist Medical Qigong Regulation was first introduced in the West by one of my teachers, Dr. Her Yue Wong. These techniques have become increasingly popular, as they are simple, yet extremely effective in tonifying and regulating the body's organ energy.

The interesting thing about Western healing practices is that they focus on the neuromuscular skeletal system to the exclusion of the visceral system, not recognizing that the visceral system is necessary to supply the fuel for the neuromuscular system. These following exercises focus on the visceral system and enhance the quality of life of the viscera.

The system as a whole is divided into two sections. The first section stimulates, energizes, and regulates the body's Yang organs and channels. The second section stimulates, energizes, and regulates the body's Yin organs and channels.

The exercises can be practiced as a complete system, or singled out individually and practiced as tonification prescriptions. As with any Medical Qigong prescription, the patient must purge his or her body of Toxic Qi before starting the tonification exercise and return the Qi back to the Lower Dantian upon completion.

DAOIST FIVE YANG ORGAN REGULATION EXERCISE

This exercise massages the five Yang digestive organs, strengthens the peristaltic action of the body's digestive system, as well as increases the capillary circulation through the stimulation of the autonomic nervous system. The Yang digestive organs (Stomach, Small Intestine, Large Intestine, Urinary Bladder, and Gall Bladder) are used by the body to release waste and Turbid Qi; therefore, it is important to keep the Yang organs strong and functioning smoothly.

The Beating and Drumming the Qi Breath Regulation (which is employed within this exercise) is beneficial for those patients who suffer from digestive problems and irregularity (i.e., spastic colon, chronic diarrhea, irritable bowel syndrome, etc.).

When the patient's arms swing rhythmically throughout the exercise, the nerves and Qi channels of the shoulder joints become stimulated. The swaying action of the arms increases the circulation in the energetic channels which connect the different internal organs and terminate at the hands.

Although this exercise is simple, the result in strengthening the body's peristaltic action and curing digestive illnesses is very effective. The patient is therefore cautioned not to eat prior to practicing this exercise.

SWAYING THE ARMS WHILE BEATING AND DRUMMING THE QI

Assume a wide stance with the arms suspended by the sides of the body. While inhaling, swing the arms up straight out in front of the body and then bring the hands in by the chest (Figure 42.1). Expand the abdomen fully focusing the mind's inten-

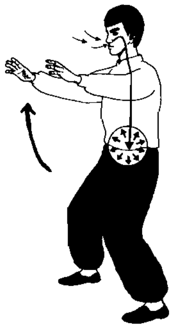

Figure 42.1. The Daoist Five Yang Organ exercise. Inhale and draw the Qi into the Lower Dantian while swinging the hands forward.

tion on filling the lower abdomen with Qi.

When exhaling, swing the arms back behind the body while compressing the abdomen down and inward directing the Qi to flow from the lower perineum, circulate back into the sacral area, then to the navel and back down to the perineum (Figure 42.2).

Continue to swing the arms back and forth at a pace that is comfortable using natural breathing. The abdomen should expand and contract without forcing the respiration. The patient must have the anal sphincter closed throughout the entire exercise. Continue this repetition for at least 50 breaths. Try to work up to 250 breaths or more for chronic conditions of abdominal obstruction and Qi stagnation within the digestive system.

The pressing and releasing action of the abdomen stimulates the energetic functions of the autonomic nervous system. The autonomic nervous system has two divisions:

1. The fight or flight mechanism of the sympathetic division -- which shuts down the digestive system due to the production of adrenaline or norepinephrine and redirects Qi and Blood flow to the brain and skeletal muscles, and
2. The vegetative, renewal mechanism of the parasympathetic division -- which draws Qi and Blood from the brain into the digestive system.

The centers for neuronal control of the autonomic nervous system and also control the digestion and tubal contraction are located in the cranium and in the sacrum. The accessory nerve and vagus nerve carries from the brain stem, its nucleus (or center), all of the information to supply the digestive system with impulses to stimulate peristaltic contraction. At the lower end of the digestive system, the sigmoid colon, rectum, Bladder and uterus are controlled by nerve fibers coming from the first, second, and third sacral foramina. The action of the cranial-sacral division of the parasympathetic nervous system is to stimulate peristaltic action (hence the movement of nutrients through the digestive system).

The movement of the abdominal cavity, diaphragm, and thoracic cavity changes the position of each organ and enhances its function. By tightening the anal sphincter and compressing and releasing the urogenital diaphragm, the reproductive organs are also stimulated, including the uterus, prostate, urethra, and testicles. Finally, the Daoist Five Yang Organ Regulation exercise also enhances the peristaltic action of the body's lymphatic system, circulatory system, and tubal systems.

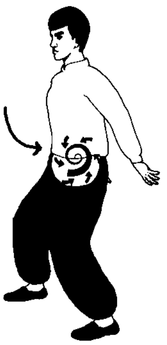

Figure 42.2. Exhale and compress the abdomen, allowing the hands to swing behind the body.

DAOIST FIVE YIN ORGAN TONIFICATION AND REGULATION EXERCISES

Each of the following exercises are designed to stimulate the specific internal Yin organ, as well as the paths of energy that flow from the channels of each organ. Since the Yin organs have the responsibility of storing energy, each exercise is named for the organ it tonifies.

The Yin exercises are performed in the sequence shown below without pausing in-between transitions. Unless specifically prescribed as a therapeutic modality, each exercise should flow directly into the next. The transitions are very important for continuity of energy flow. This particular Five Yin Organ series begins with an opening exercise and ends with a closing exercise.

OPENING, COLLECTING, AND MOVING THE QI

The opening exercise is designed to stretch and loosen the spinal vertebrae and to facilitate the energy flow through the patient's Governing Vessel. The spinal stretch also stimulates the flow of fluids and electrical impulses in the central nervous system (brain and spinal cord), which is nurtured by the cerebral spinal fluid. When the fluid starts to move, it awakens the nervous system causing more nutrients to enter the cellular structure of the nerves. The stretching, compressing, and pumping action of the spinal cord stimulates the movement receptors located in all of the ligaments and fascia energizing the body's entire proprioceptor mechanism.

The opening exercise is also designed to stretch the connective tissue, warm the body's intercellular matrix and tissues, stimulate the somatic system, free adhesions, as well as enhance the elasticity and compressibility of the ligamentous structure.

After finishing the Daoist Five Yang Organ Regulation exercise, return both hands to the Lower Dantian. Next, raise both hands above the head as you inhale (Figure 42.3).

Imagine energy (like a rushing river) flowing up from the feet, filling the legs, hips, waist, chest, arms, and head. Every square inch of the body is absorbing and being saturated with this Earth energy. When the entire body is completely full, the hands should still be positioned above the head (Figure 42.4).

Begin to exhale and feel the hands getting very heavy. The heaviness pulls the hands forward and then slowly starts to pull the body over. Slowly bend the head forward and feel the cervical vertebrae stretch. Let the hands pull the arms, which in turn will pull the shoulders and torso, followed by bending at the waist. The purpose of this movement is to feel each vertebra of the spine stretch sequentially so that a rippling effect descends down the spine. This will facilitate maximum stretch of the spine. While bending over, exhale and imagine the energy melting away from the entire body (like ice melting into water) slowly dissolving and rushing down through the feet and out into the ground (Figure 42.5).

Once completely bent over, imagine picking up a ball. Bend the knees and slowly stand up. While standing up, reverse the rippling of the spine from the coccyx, sacrum, and lower lumbar vertebrae to the base of the skull (Figure 42.6). Repeat these movements five times. Remember to inhale while raising the body, bring both arms over your head, and exhale as the body descends.

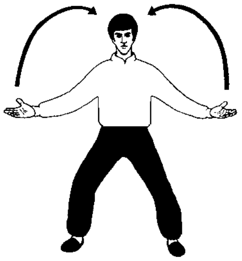

Figure 42.3. The Daoist Five Yin Organ Opening exercise. Begin by stretching the arms above the head.

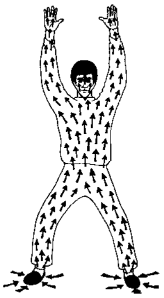

Figure 42.4. Inhale and imagine the Earth Qi like water filling the entire body.

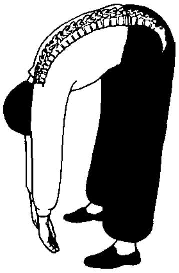

Figure 42.5. Exhale, bend the body and imagine any tension melting out the body into the Earth.

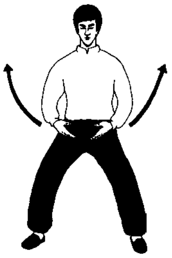

Figure 42.6. For the transition movement to the Lung Tonification exercise: inhale, bend the knees and slowly stand up.

LUNG TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Lungs' tissues and nerves, as well as tonify the entire Lung organs and channels. In the Lung exercise, both the sympathetic and parasympathetic divisions of the autonomic nervous system are stimulated. The compression and release action of the scapulae and thoracic vertebrae stimulate the nerves, arteries, veins and capillary system in the thoracic region.

Compressing and releasing of the muscles and blood vessels in the back and chest areas stimulates the transforming of the Qi and Blood and enhances the health of the individual tissue cells. The compression and release of the thorax also increases the stimulation of the nervous system in the Lung area, as well as stimulates the "release and return" action of the lymphatic system. All the major lymphatic ducts are located in the posterior portion of the thoracic wall.

The compression and release of the second, third, and fourth thoracic vertebrae stimulates the sympathetic preganglionic fibers from the spinal cord, which in turn stimulates the postganglionic fibers and the sympathetic outflow to the Lungs and bronchi. Patients practice this exercise to strengthen their respiratory system as well as increase the production of their Wei Qi to enhance their body's immune system.

Upon completion of the Opening exercise, the patient brings the hands above the head and places them in front of the torso at shoulder level while exhaling. The patient's palms should be facing downward parallel with the floor (Figure 42.7).

Inhale and bring the arms straight out to the sides (both hands should not go any higher than the shoulders) (Figure 42.8).

As the arms reach the sides of the body, rotate the palms until they face upward. Keep the shoulders stable and relaxed (Figure 42.9).

While exhaling, bring the arms straight out in front of the body returning to the beginning posture (Figure 42.10).

Allow the scapulas to gently push the arms forward. The sternum at this point is pushed inward so as to hollow the chest. Once the hands are aligned in the front of the shoulders, rotate the arms and

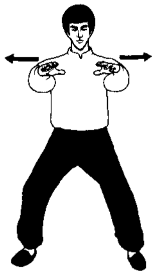

Figure 42.7. The Lungs Tonification exercise: Inhale as the hands separate.

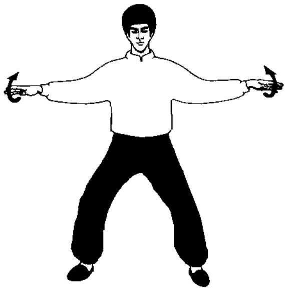

Figure 42.8. Separate the Hands and turn the palms upward.

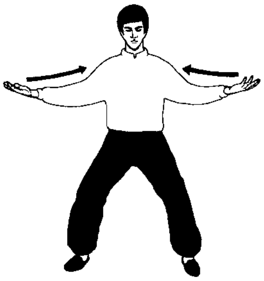

Figure 42.9. Exhale as you bring the arms toward each other.

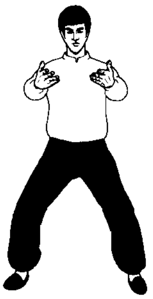

Figure 42.10. Draw the hands toward each other and turn the palms facing the Earth.

inhale to begin again. Repeat 20 times.

Note: Do not let the hands touch when bringing the arms back together towards the center line of the body. The body will short circuit itself, resulting in a mild shock or a sick feeling inside the thorax by the Heart.

KIDNEY TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Kidneys' tissues and nerves, as well as tonify the entire Kidney organs and channels. According to Western physiology, the Kidneys maintain the chemical, pH (potential of hydrogen), and temperature balance of the Blood. The Kidneys function like a rheostat changing the body's ability to absorb, retain, or release water according to the environment (e.g., retaining water in hot climates).

The twisting back and forth, pumping action of the exercise stimulates an opening and closing action within the Kidney organ. The Kidneys receive the major Blood supply from the abdominal aorta. In the Kidney exercise, the action of raising the torso upwards will literally suck Qi and Blood into the Kidney area.

Patients practice this exercise to strengthen their reproductive system, as well as increase the production of their Jing. It also helps balance the body's general energy condition, and can be used to treat sexual disorders and malfunctions.

Upon completing the Lung exercise, both arms are extending straight out in front of the body (Figure 42.11).

Inhale and drop the elbows and bring the hands down in front of the Lower Dantian (Figure 42.12).

Exhale and shift the weight to the left. As the left outer palm wraps around the back, resting itself on the right Kidney, the right palm extends towards the left direction (Figure 42.13).

The right palm, facing outward towards the left direction, begins to circle upwards in front of the face at eye level. As the palm begins to move, the eyes watch the back of the right palm (Figure 42.14).

Shift the weight towards the right side of the body and begin leaning towards the right direction. The eyes continue to follow the right palm as it begins to circle downwards in front of the body. Ex-

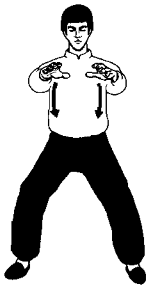

Figure 42.11. After completing the Lung exercise, center the body to prepare for the Kidney Tonification exercise.

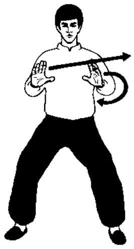

Figure 42.12. Preparation for the Kidney Tonification Exercise

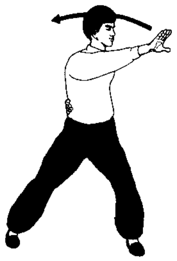

Figure 42.13. Kidney Tonification Exercise: exhale as the left outer palm wraps around the right kidney, and the right palm extends outward in a left direction.

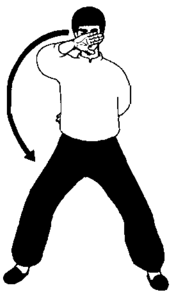

Figure 42.14. Begin to inhale as the right palm circles upwards and turns outward at eye level and begins to descend.

hale while shifting the weight (Figure 42.15).

Shift the weight, bend the knees and begin to lean over. While continuing to exhale, imagine scooping water with the right palm. Bend at the waist and scoop with the right arm. Keep the back relaxed and straight (Figure 42.16).

Allowing the right arm to follow the motion of the body, begin rising up slightly while turning forming an arc in front of the torso. As the palm raises, imagine the energy (which has been scooped into the arm) draining down the right arm across the shoulders and down the left arm into the right Kidney. Repeat this side for ten breaths (Figure 42.17).

After completing 10 breaths switch directions by shifting the weight onto the right leg (Figure 42.18 and Figure 42.19).

The right hand wraps the torso ending at the left Kidney and Mingmen area while the left palm is positioned in front of the face. Begin to inhale and repeat 10 times on the opposite side.

LIVER TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Liver's tissues and nerves, as well as tonify the entire Liver organ and channels.

According to Western physiology, the Liver is the body's master filter, collecting all of the nutrients that are being absorbed from the digestive system via the hepatic portal vein. The Liver functions like a series of channels. The cells of the channels are considered the processing factory for the entire body. If the body has a need for a particular nutrient, the Liver will create the enzymes for digestion, breaking down into components that which the body needs (or storing the components until needed). The Liver is also responsible for detoxifying, producing antibodies, and creating new cells to handle foreign substances recently absorbed by the body (storing substances of which it has no need).

The compression and release action of the torso causes a physical compression within the Liver organ itself, flushing Qi and Blood from the digestive system through the Liver and into the Gall Bladder where bile is stored.

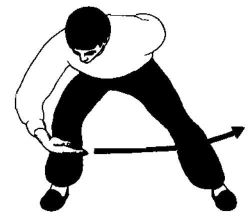

Figure 42.15. Continue inhaling as you bend over at the waist. Once your head sinks below your hips begin to exhale. Allow your right palm to circle downward to the left.

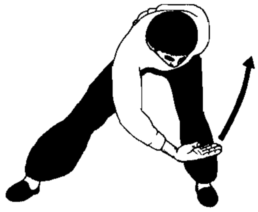

Figure 42.16. Continue exhaling as your weight shifts to the other foot and imagine the palm scooping up water as the body begins to circle upward.

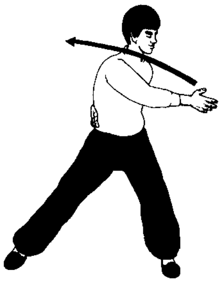

Figure 42.17. As the torso begins to arch upwards, allow the right arm to follow the body's movement. As the arm reaches the chest level, turn the outer palm towards the face and begin inhaling.

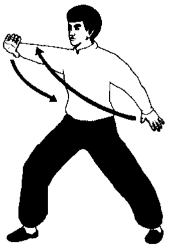

Figure 42.18. Preparation to switch hand positions: Shift the body's weight onto the right leg as the right arm begins to move to the back left Kidney and the left arm leaves the right Kidney to move forward.

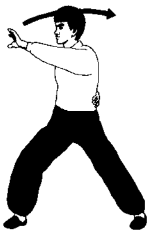

Figure 42.19. Kidney Regulation exercise final movement: begin to inhale as the left palm circles upwards and turns outward at eye level and begins to descend to the left.

Patients practice this exercise to strengthen their tendons and ligaments. This exercise helps to regulate the patient's emotional state.

Upon completing the Kidney exercise, begin to unwind the arm from behind the back and rotate the hips so they face forward (Figure 42.20).

Place the right arm straight out in front of the body, shoulder level, palm facing down. Place the left hand by the left hip, palm facing up (Figure 42.21).

Simultaneously draw the right palm back and extend the left palm forward while inhaling. The left hand moves forward passing the right hand which is moving backwards (Figure 42.22).

The hands continue in this motion extending outward from the center line of the body. Use long, slow inhalation and exhalation, inhaling and exhaling every three moves. Relax and continue to repeat this sequence for 20 breaths (Figure 42.23).

HEART TONIFICATION AND REGULATION

This exercise is designed to massage the Heart's tissues and nerves, as well as tonify the entire Heart organ and channels.

The compression and release of the skeletal muscles squeezing and releasing the cardiac muscle energetically stimulates the pericardium surrounding the heart. The twisting movement and compression of the torso also enhances the fluidity of all the connective tissues that are involved with the Heart. Every time the thoracic wall moves, expands, contracts, or twists, the tissues of the Heart are pulled, stretched, and released.

Patients practice this exercise to strengthen their circulatory system. This exercise helps to

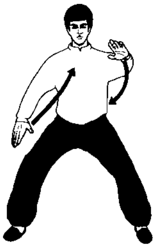

Figure 42.20. Preparation for the Liver Tonification Exercise: upon completing the Kidney exercise, begin to unwind the arm from behind the back and rotate the hips so they face forward.

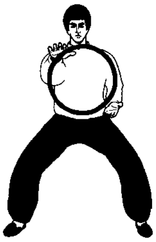

Figure 42.21. Liver Tonification Exercise: place the right arm straight out in front of the body, shoulder level, palm facing down. Place the left hand by the left hip, palm facing up.

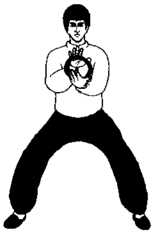

Figure 42.22. Simultaneously draw the right palm back and extend the left palm forward while inhaling. The left hand moves forward, passing the right hand which is moving backwards.

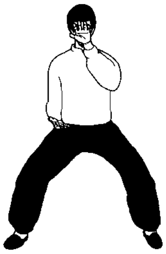

Figure 42.23. Use long, slow inhalation and exhalation, inhaling and exhaling every three moves.

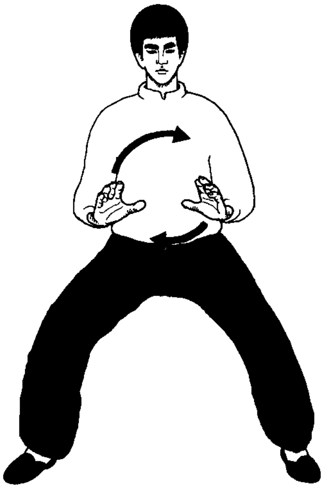

Figure 42.24. Preparation for the Heart Regulation Exercise: upon completing the Liver exercise allow both hands to sink down in front of the lower abdomen.

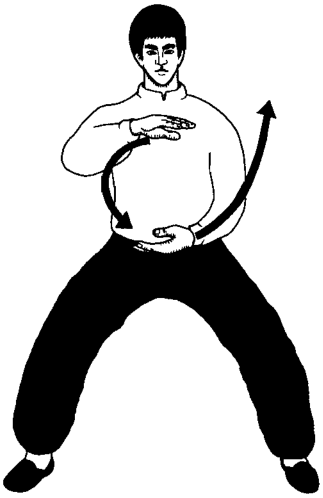

Figure 42.25. Heart Tonification Exercise: imagine the hands embracing a ball, placing the right hand on top, and the left hand on bottom: inhale.

stimulate the patient's intellectual activity and mental acuity or clarity.

Upon completing the Liver exercise allow both hands to sink down in front of the lower abdomen (Figure 42.24).

Imagine the hands embracing a ball, placing the right hand on top, and the left hand on bottom (Figure 42.25).

Exhale and twist the hips to the left, simultaneously raising the left hand up over the head while extending the right hand towards the left (Figure 42.26).

Continue pressing until both hands extend to the furthest point (Figure 42.27).

Inhale and turn the right palm upwards facing the sky as the body twists back to the center allowing the right hand to end up in front of the navel (Figure 42.28).

The left hand simultaneously turns so that the palm faces the center of the right hand following it to the center line of the body (Figure 42.29).

As the body reaches the point where it faces forward, twist to the right side and push in the same manner while exhaling. Repeat 10 times on each side (Figure 42.30).

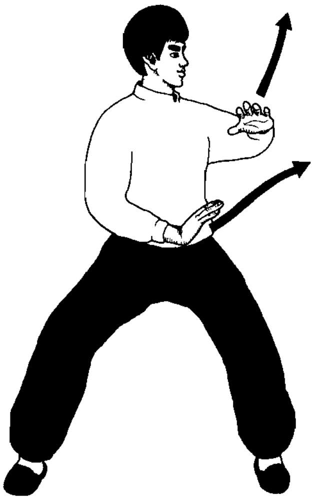

Figure 42.26. Exhale and twist the hips to the left simultaneously raising the left hand up over the head while extending the right hand towards the left.

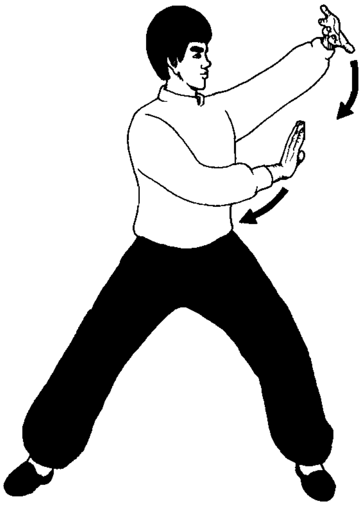

Figure 42.27. Continue pressing until both hands extend to the furthest point.

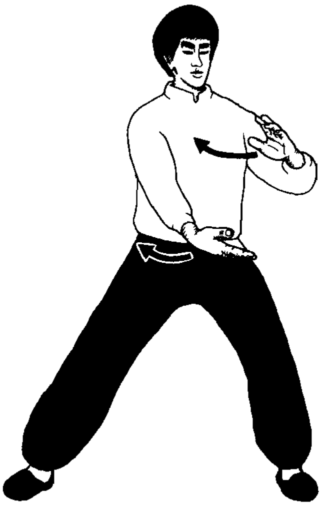

Figure 42.28. Inhale and turn the right palm upwards facing the sky as the body twists back to the center allowing the right hand to end up in front of the navel.

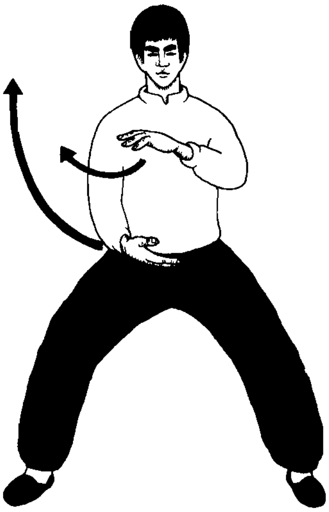

Figure 42.29. The left hand simultaneously turns so that the palm faces the center of the right hand following it to the center line of the body.

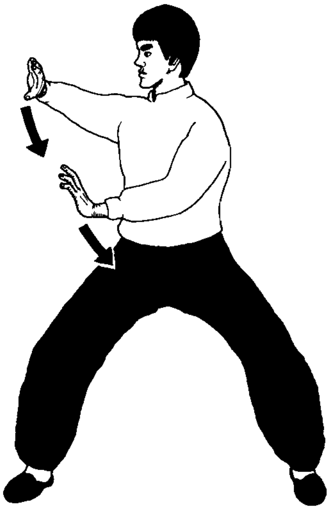

Figure 42.30. As the body reaches the point where it faces forward, twist to the right side and push in the same manner while exhaling.

SPLEEN TONIFICATION AND REGULATION EXERCISE

This exercise is designed to massage the Spleen's tissues and nerves, as well as tonify the entire Spleen organ and channels.

Raising the arms above the head stretches all of the suspensory ligaments of the Spleen making the Spleen energetically accessible by the twisting movements of the torso. The compression and release action on the connective tissue and fascia that support the placement of the Spleen allows the internal channels to become dilated facilitating a greater flow of Qi and Blood into the organ.

While looking up and turning to the right, the left brain is stimulated. Consequently, by looking up and turning to the left, the right brain is stimulated. According to Western physiology, light waves come into the eyeballs through the retina stimulating the optic nerve, chiasma, tract, lateral geniculate body (stimulating the thalamus), superior colliculi (in the brain stem where visual reflexes are initiated), optic radiations, cerebral cortex, and occipital (visual) cortex, which integrates both visual and memory impulses responsible for the perception of the image. Half of the information received from the right eye goes to the right side of the brain, while the other half of the information crosses over and stimulates the left brain.

Patients practice this exercise to strengthen their digestive system, as well as to enhance their visual equilibrium.

Upon completion of the Heart exercise, bring both hands down to the waist level, palms facing downward. Circle them above the head and perform the Pulling Down the Heavens exercise. Next, move the palms towards the outside of the knees while twisting the thumbs to point towards the back of the body (this stimulates the Luo points of the arms and squeezes the Wei Qi of the hands and arms deep into the bones) (Figure 42.31).

The thumb and index fingers touch and form a triangle as the hands slowly raise above the head (Figure 42.32).

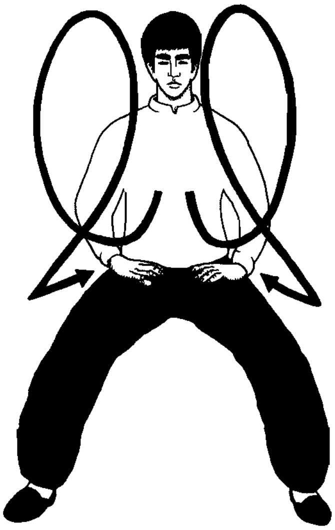

Figure 42.31. Preparation for the Spleen Regulation Exercise

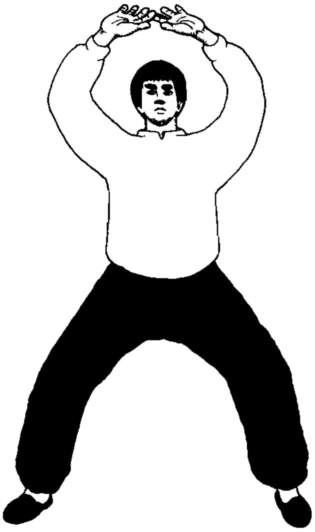

Figure 42.32. Spleen Regulation Exercise: the thumb and index fingers touch and form a triangle as the hands slowly raise above the head.

The eyes look through the center of the triangle as the upper torso is turned from side to side. It is important that the hips remain stationary and that the upper torso rotate in order to massage the digestive organs. Inhale in the center. Exhale on each side. Repeat 10 times on each side (Figure 42.33).

ENDING, ROOTING, AND STABILIZING THE QI

The purpose of the ending is to gather the Qi back into the patient's Lower Dantian rooting and stabilizing the patient's Qi. This ending exercise is very important as its initial goal is to gather any excess Qi developed from the over stimulation of the Five Yin Organs and to collect the runoff energy into the Lower Dantian (where it can be absorbed and used by the Eight Extraordinary Vessels). Otherwise, if the individual unknowingly has a preexisting Excessive condition in a particular Yin organ, the additional tonification can exasperate the organ's condition. To avoid this Excess condition, certain Medical Qigong schools will use external massage and lightly brush the energetic channels after practicing certain Qigong exercises and meditations to disperse any excess Qi.

The soft compression used during the ending exercise increases the partial pressure of oxygen in the Blood as more oxygen passes into the cells. The action of moving oxygen into the cells creates a movement of carbon dioxide out of the cells increasing the exchange rate and establishing a greater metabolic breakdown.

The breathing initiated for the ending exercise is known as "Turtle Breathing." The Turtle Breathing method requires slow respiration and stimulates the center hub of the Belt Vessel.

Upon completion of the Spleen exercise, exhale and start to separate both hands (as if embracing a ball) (Figure 42.34 and 42.35).

Bring both hands down to waist level and separate the arms, moving the hands away from the center of the body (Figure 42.36). Inhale and raise the arms up over the head. Each hand is placed above the head as if embracing a ball (Figure 42.37).

Exhale and form loose fists with the hands (Figure 42.38).

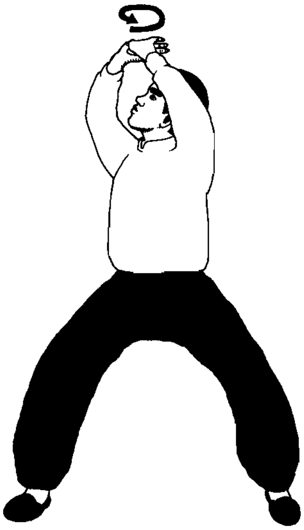

Figure 42.33. The eyes look through the center of the triangle as the upper torso is turned from side to side. Inhale in the center, exhale on each side.

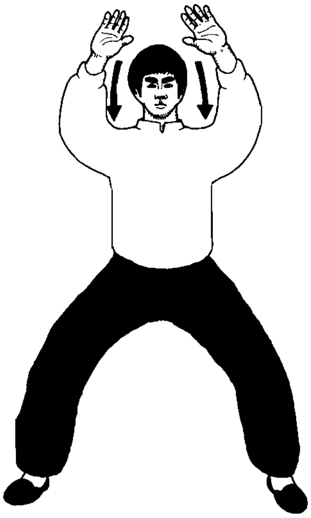

Figure 42.34. Preparation for Ending the Daoist Five Yin Organ Exercise (a)

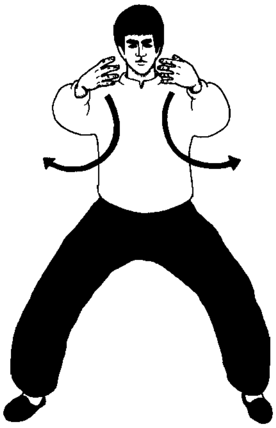

Figure 42.35. Preparation for Ending the Daoist Five Yin Organ Exercise (b)

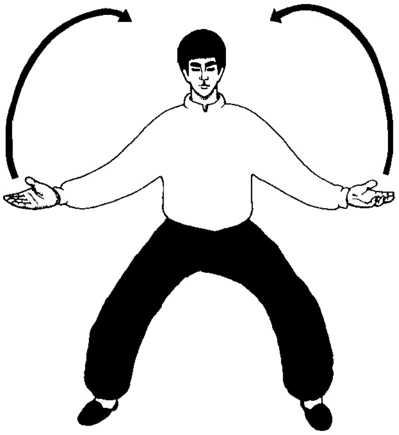

Figure 42.36. Preparation for Ending the Daoist Five Yin Organ Exercise (c)

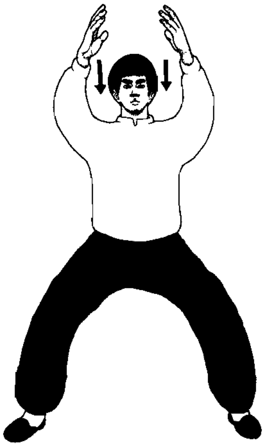

Figure 42.37. Ending the Daoist Five Yin Organ Exercise: each hand is placed above the head as if embracing a ball.

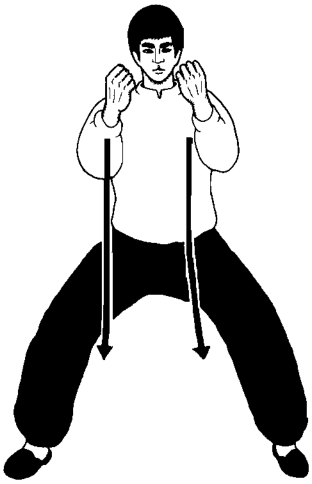

Figure 42.38. Exhale and form loose fists with the hands.

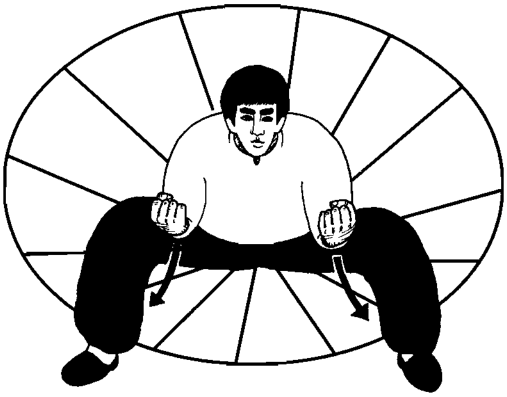

Figure 42.39. As you bend over, slowly turn the soft fists so that the curled fingers face the body and begin squatting, forming a Turtle Posture.

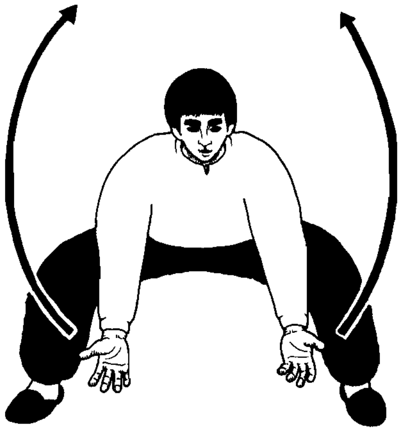

Figure 42.40. Inhale and slowly begin to raise the body up while opening the fists, and allow the hands to return back to their original position at the sides of the body.

As you bend over, slowly turn the soft fists so that the curled fingers face the body and begin squatting forming a Turtle Posture (Figure 42.39).

Inhale and slowly begin to raise the body up while opening the fists and allow the hands to return back to their original position at the sides of the body (Figure 42.40). Repeat five times and end in the Wuji posture.

When practicing the Turtle Breathing method, the mind's intention and respiration should be focused onto the Lower Dantian, navel, and Mingmen areas of the body.

The Lower Dantian is divided into eight sections (see Figure 42.41). Each section is numbered, beginning with the front of the Lower Dantian at the navel area (1), increasing in number as one continues over the left and right sides of the waist (4), ending at the Mingmen area (8).

Think of the anal sphincter as a great funnel, attached in eight different segments (like canals) which progress up the torso. As you focus your mind on each section (begin with the navel at section 1), allow any excess energy to flow downwards from that particular section to be gathered into the Lower Dantian. With your imagination, direct the breath and Qi from your torso to flow into each of the eight areas.

The breath and Qi should be slowly inhaled and guided into the Lower Dantian via the anal sphincter area for an even three counts. Then, proceed to gather Qi from the next consecutive area of the body. By pulling upwards on the anal sphincter and perineal areas, an energetic vacuum is created within the pelvic diaphragm and urogenital diaphragm. This energetic vacuum is used to gather the internal organ's excessive Qi.

Begin with collecting Qi from the front of the body, to the navel via the anal sphincter. Hold the Qi in the Lower Dantian area for three counts, then proceed to the next areas at the sides of the body absorbing the Qi into the Lower Dantian via the anal sphincter for three counts.

Next, proceed to the sides of the body continuing in progression until you reach the Mingmen area (8). Then, reverse the progression beginning with the Mingmen and working around

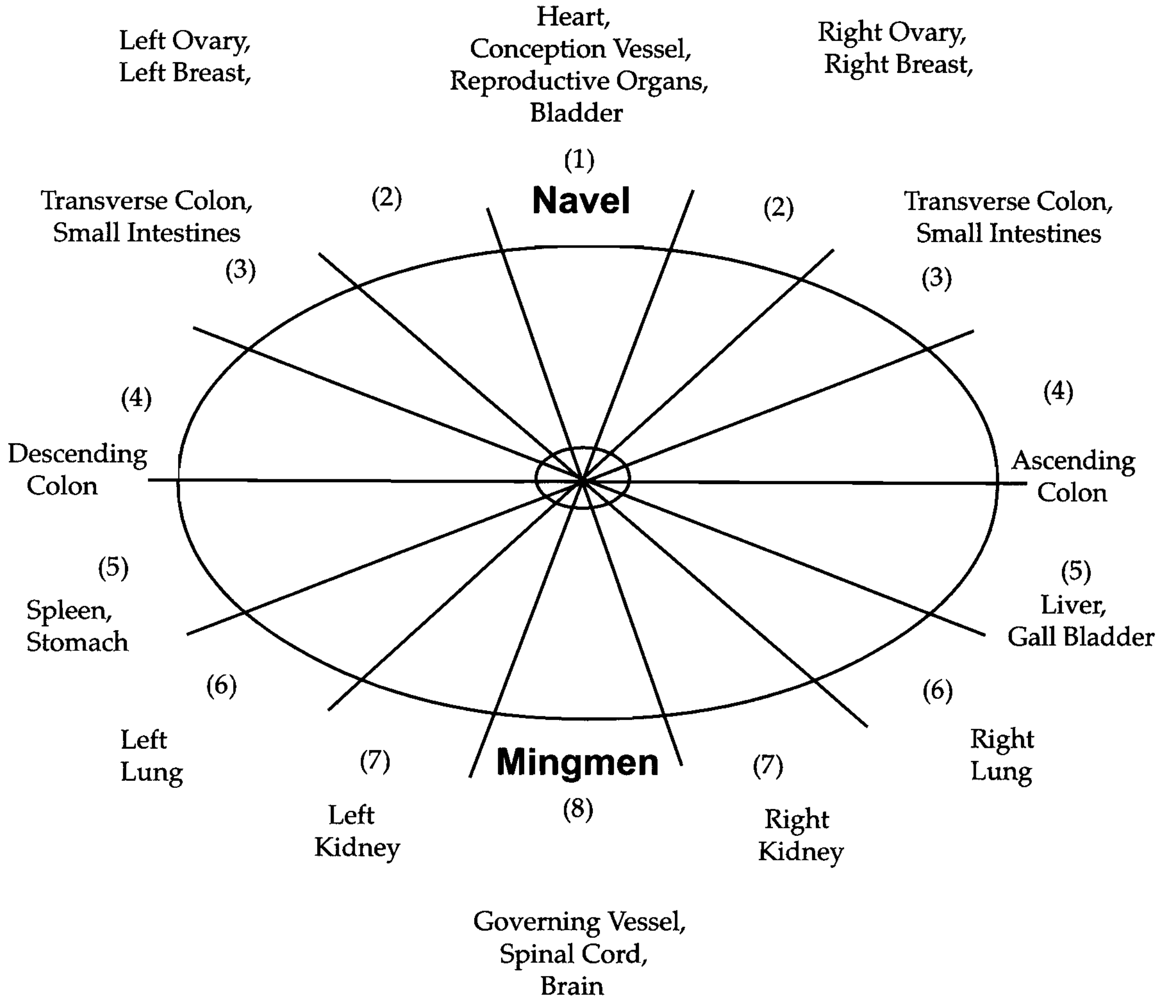
Heart,
Conception Vessel,
Reproductive Organs,
Bladder
(1)
Navel
Left Ovary,
Left Breast,
Right Ovary,
Right Breast,
(2)
(2)
Transverse Colon,
Small Intestines
(3)
Transverse Colon,
Small Intestines
(3)
(4)
Descending
Colon
(4)
Ascending
Colon
(5)
Spleen,
Stomach
(5)
Liver,
Gall Bladder
(6)
Left
Lung
(6)
Right
Lung
(7)
Left
Kidney
Mingmen
(8)
(7)
Right
Kidney
Governing Vessel,
Spinal Cord,
Brain

Figure 42.41. When practicing Turtle Breathing, inhale and slowly begin drawing the Qi from the front of the body into the navel area (1), absorbing the energy from the Heart, Conception Vessel, reproductive organs, and Bladder into the Lower Dantian. This absorbing action is created by pulling upwards on the anal sphincter in the direction of the intended energy. Next, proceed to the sides of the body. In women, the next area (2) will include the excess energy gathered from both the left and right ovaries, as well as the left and right breasts. In men, however, this area (2) is considered general and not specific to certain internal organs or tissues. Continue in the progression of circling the waist until you have reached the Mingmen area (8). Then, reverse the order of the tissue areas (8, 7, 6, etc.).

the waist until you reach the navel (8, 7, 6, etc.). Make sure that you push your anal sphincter towards the direction of the area of the body from which you are absorbing Qi.

After perfecting the Turtle Breathing, a practitioner can create the energetic vacuum and initiate the absorption of excess Qi into the Lower Dantian by moving his or her intention from the anal sphincter through the Lower Dantian and torso at an extremely fast speed.

SECTION X

TREATMENT OF INTERNAL DISEASES WITH QI EMISSION THERAPY

INTRODUCTION

Clinical records from China verify that Medical Qigong treatments are most effective on nervous system diseases, migraines, bronchial asthma, nocturnal enuresis, psychosomatic diseases, gastric ulcers, arthritic pains, sterility, and in the elimination of ovarian cysts and benign tumors.

The following chapters are designed to assist the Medical Qigong doctor in his or her professional clinical practice, by explaining the various techniques and prescriptions used in several Medical Qigong Hospitals, Institutes and Colleges throughout China.

These Medical Qigong techniques are presented as basic guidelines for the Qigong doctor to use, and need not be followed exactly to the letter. They are presented as an example to advise Qigong doctors in treatment and prescription protocols.

I have personally found these various approaches of Chinese energetic medicine to be extremely effective in my own clinical practice, and have therefore encouraged the directors of the International Institute of Medical Qigong to instruct their students in the following techniques.

This particular section of the book (Section X) deals with the treatment of Zang (Yin) and Fu (Yang) disharmonies in clinical practice, and discusses general and specific diagnosis, treatments, and prescriptions. I will begin by exploring several examples of the Five Elements' disease patterns and the various techniques used to treat them (Chapter 43-48). Starting with the treatment of Liver diseases, I will progress to the treatment of Heart, Spleen / Stomach, Lung and Kidney Diseases.

In the ending chapter in this section emphasis will be placed on the practical application and theoretical principles of "Miscellaneous" diseases and their treatments.

CHAPTER 43
QI EMISSION THERAPY FOR LIVER DISEASES

INTRODUCTION

The Five Yin Organs' interrelationship with each other, and their complex interrelationship with the energetic ebb and flow of Qi must constantly be regulated to maintain health. The constant flux of energetic flow within the body sometimes causes an Excess or Deficiency of Qi requiring specific organ therapy for restoration and balance. This therapy is designed to bring the specific organ or organ systems back into balance by using various meditations, visualizations and breathing exercises.

This section focuses on specific Liver diseases and how to rectify them. The Liver is responsible for the circulation and smooth movement of the body's internal Qi.

The Liver is a solid (Zang) organ which stores the Blood and is attributed to the element Wood. The Liver is sometimes referred to as the Green Emperor. The Liver is in charge of strategy without hesitation. It is energetically paired with the Gall Bladder (a hollow organ) which carries out the decisions made by the Liver (Figure 43.1).

The Liver houses the Hun. The Liver's association with the environment is through the eyes. Through observation, the Qigong doctor's eyes can receive images that reveal the status or intention of the patient. These images are absorbed into the doctor's Shen via the Hun. There is an old Chinese saying, which goes, "When the Hun swim to the eyes, a person can see."

In the springtime, Liver Qi becomes more active in individuals with strong Liver Qi; whereas, it may become Deficient in those with weak Liver Qi. When Liver Qi is Deficient, the patient becomes fearful; when the Liver Qi is in Excess, the patient becomes consumed by anger. The excessive eating or drinking of sour foods, and overexposure to wind depletes the Liver.

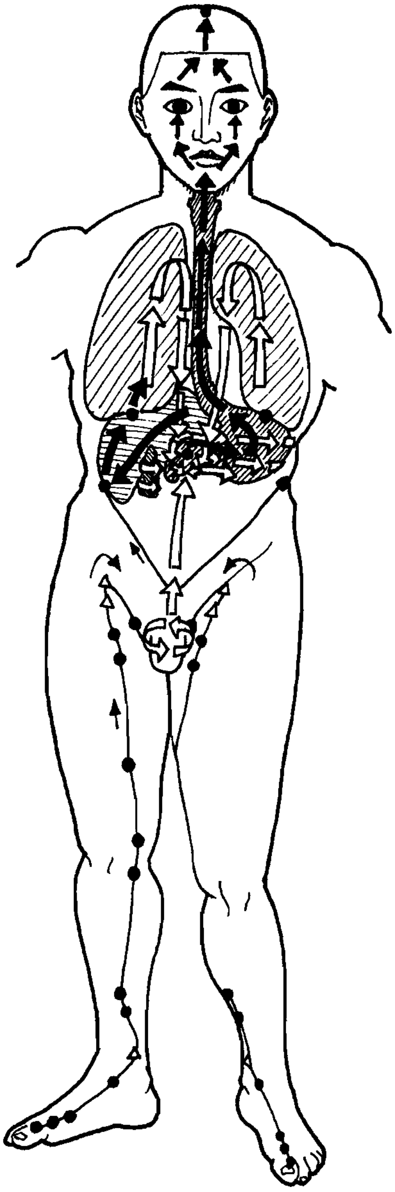

Figure 43.1. The Liver (Lv) Channel

LIVER COMPLICATIONS AND SYMPTOMS

Liver diseases often manifest with the following symptoms which are common to the organ's energetic dysfunction: headaches, dizziness, tiredness, and mental and emotional problems. When Liver disorders occur, the patient may experience Qi movement below the left side of the navel (which feels hard upon pressure), pain and distention in the hypochondria, cramps, drowsiness, blurred vision, etc. A list of Liver syndromes and their symptoms are described as follows.

1. Stagnation of Liver Qi can cause such symptoms as: headaches, mental and emotional problems, hypochondriac pain, abdominal pain and masses, diarrhea, constipation, painful urination, dysmenorrhea, and premenstrual tension.
2. Stagnant Liver Qi Invading the Stomach can result in epigastric pain.
3. Stagnant Liver Qi Invading the Lungs can cause wheezing.
4. Liver Yang Rising can cause such complications as: headaches, dizziness, and tiredness.
5. Liver Fire Blazing can cause such symptoms as: severe headaches, dizziness, breathlessness, wheezing, coughing, mental and emotional problems, insomnia, tiredness, painful or scanty urination, and constipation.
6. Liver Wind can cause headaches, dizziness, seizures, tremors, muscle spasms and tiredness.
7. Liver Blood Deficiency can cause such symptoms as headaches, mental and emotional problems, tiredness, hypochondriac pain, constipation, and premenstrual tension.
8. Liver Yin Deficiency can cause such symptoms as wheezing, mental and emotional problems, insomnia, tiredness, and hypochondria pain.
9. Stagnation of Liver Blood can cause such symptoms as headaches, mental and emotional problems, hypochondriac pain, abdominal pain, and dysmenorrhea.
10. Damp Heat in the Liver and Gall Bladder can cause such symptoms as jaundice, headaches, mental and emotional problems, hypochondriac pain, and abdominal pain.

GENERAL TREATMENT FOR LIVER DISEASES

The following is a description of a general protocol used for treating common Liver diseases. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted or Turbid Qi is removed. Once the Qigong doctor becomes proficient at locating these specific points, the treatment flows more smoothly.

The reader will notice that after comparing the various Liver and Gall Bladder therapies, certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Liver through either direct energetic insertion into the Liver's anatomical location, or through the patient's channels or external energetic fields.
2. After purging, tonifying, and regulating the patient, the Qigong doctor stimulates the back gates to the patient's Liver and the Yellow Court.
3. Next, the doctor roots the patient's Middle Burner Qi into the Lower Dantian.
4. Finally, the Qigong doctor leads any remaining Turbid Qi down the right Gall Bladder Channel and out of the patient's body (Figure 43.2).

Keep in mind that these areas are general, and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR LIVER QI DYSFUNCTION

The primary therapies used to treat Liver diseases include the Windy Breathing Method (see Chapter 16) to strengthen the Liver, Spleen, Lungs, and Kidneys, and Quiescent Qi Regulating meditations (see Chapter 14).

1. Patients with Deficiency of Qi and Cold syndromes of the Liver may feel cold and are li-

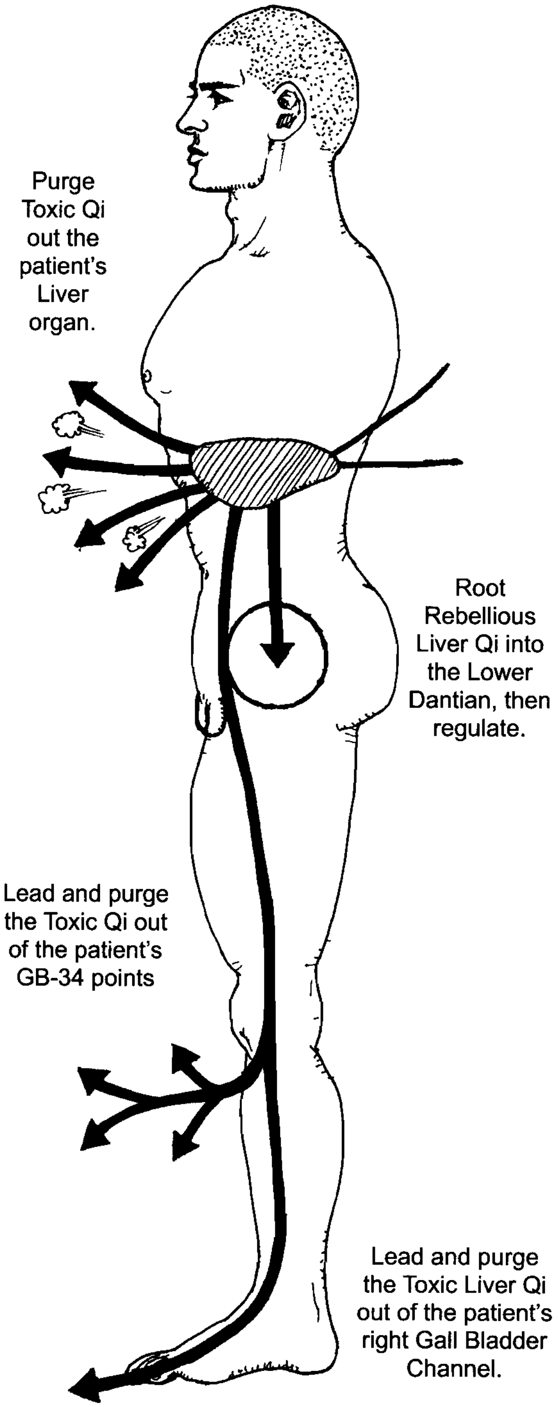
Purge Toxic Qi out the patient's Liver organ.
Root Rebellious Liver Qi into the Lower Dantian, then regulate.
Lead and purge the Toxic Qi out of the patient's GB-34 points
Lead and purge the Toxic Liver Qi out of the patient's right Gall Bladder Channel.

Figure 43.2. Basic Treatment for Liver Stagnation

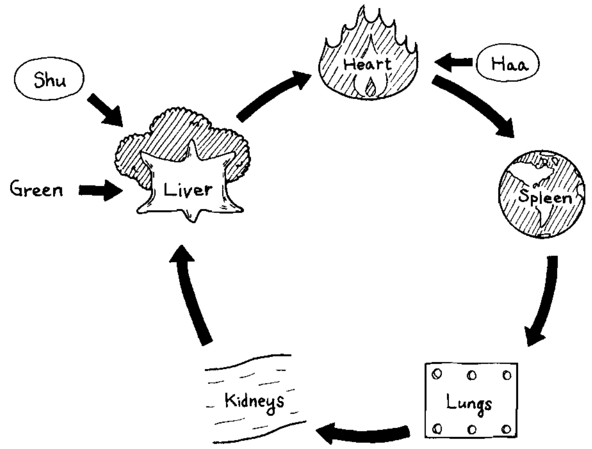
Shu
Green
Liver
Heart
Haa
Spleen
Lungs
Kidneys

Figure 43.3. Purging the Child to Treat the Mother

able to feel frightened. To treat this condition, the patient can practice the meditation of Taking in the Green Qi to nourish the Liver and also Taking in the Blue Qi to nourish the Kidneys (Formula: nourishing the mother to replenish the child).

2. Patients with an Excess of Qi and Heat syndromes of the Liver may have dry eyes, pain in the hypochondria that may radiate down to the lower abdomen, and changing moods accompanied by flushed cheeks. To treat this condition, the patient can practice uttering "Xu" (pronounced Shu) for the Liver, and "Ha" for the Heart (Figure 43.3. Formula: purging the child to treat the mother).

   Note: In treating Liver diseases, exhaling the sound "Shu" purges and sedates, while inhaling "Shu" replenishes and tonifies (see Chapter 40).

3. Stagnation of Liver Qi and Liver Fire Rising (anger or rage impairing the Liver) may cause an abnormal dispersing and impair the Liver's function. The patient may experience headaches, deafness, and swelling cheeks. To stabilize the hyperactive Liver Yang (and guide the flaring up Liver Fire to descend), the Qigong doctor guides the patient's Qi back to its origin with external energy projection. If treatment of Liver stagnation is not undertaken, there may be an organic change in the

Liver, resulting in hardening of the Liver or Liver cancer.

4. Patients with Rebellious Liver Qi resulting in hypochondriac pain and stiffness with spasms, should be treated by guiding their Qi back to the origin of the Liver with external Wei Qi therapy.

LIVER QIGONG PRESCRIPTIONS

The following system of exercises can be used to regulate the Qi and Blood of the Liver Channels, soothe and regulate Qi circulation, and suppress hyperactive Yang. These exercises can also be used to prevent and treat diseases and syndromes such as hypertension, psychoneurosis, chronic hepatitis, and hepatosplenomegaly (the enlargement of both Liver and Spleen), as well as bitter taste in the mouth, dry throat, dizziness, vertigo, and fullness in the chest and hypochondrium.

COLOR VISUALIZATION AND ORGAN CORRESPONDENCE

In China, the clinical use of color therapy is prescribed according to the Five Color Correspondence Theory. Each of the five major colors are assigned to specific organs and their channels (see Chapter 4). If an organ is diseased, it will generally appear in various shades from gray to black. To heal the disease, the patient imagines taking in the organ's clean color, while dispersing the turbid pathogenic color. This visual meditation helps to return the organ's Qi back to its origin, in order to achieve balance and health.

PRESCRIPTIONS FOR TONIFYING THE LIVER

Taking in the Green Qi can be used to tonify the Liver. The properties inherent within the color green have a persistent vibratory rate also used for tonifying the Liver and Gall Bladder.

1. Assume a standing, sitting, or lying posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts. Place the tongue against the soft palate on the roof of the mouth.
2. Visualize green energy in front of you. Draw from a visual/emotional association with green trees, grass, bushes, etc. from your imagination, or if possible, actually look at a green object. Inhale the green light energy in through the nose and down into the Liver organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Liver; however, the bright, clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.

   Next breathe the color green in through the nose, filling the entire mouth. When exhaling, send the Qi slowly to both costal portions at the sides of the Lungs, and then down to the Lower Dantian. Practice eight to sixteen times.

PRESCRIPTIONS FOR PURGING THE LIVER

Purge the Liver by having the patient rub his or her chest while sounding "Shu." Have the patient assume a sitting or standing posture, and place both palms flatly on both sides of the chest and inhale slowly. When exhaling, the patient should sound the word "Shu," while gently rubbing both costal regions in a clockwise circular motion with both palms. Practice for ten to twenty breaths.

PRESCRIPTIONS FOR SOOTHING THE LIVER AND CONDUCTING QI

The following exercise is used to regulate and soothe the Liver Qi; it is used after the purging or tonifying exercises.

1. Stand relaxed with both arms naturally hanging down, palms facing downward, with the five fingers of each hand slightly held up. Press down with slight force and imagine Qi reaching the palms and extending to the fingertips. Press the palms down three times (Figure 43.4).
2. Lift both hands up in front of the chest, with each palm facing forward. Focus the mind on both palms. Push the palms forward three times, first drawing them closer to the body and then pressing them further away from the chest (Figure 43.5).
3. Stretch both hands horizontally out to the sides of the body (as if imitating a bird that is stretching out its wings) with all the fingers pointing upward, and the palms pushing to

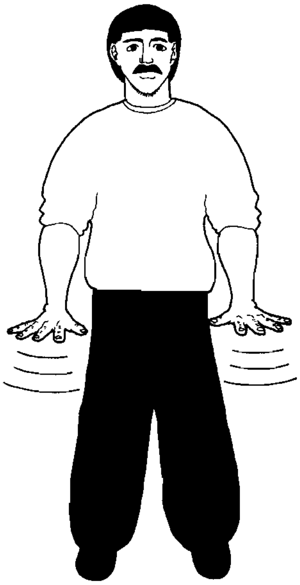

Figure 43.4. Press the palms downward three times.

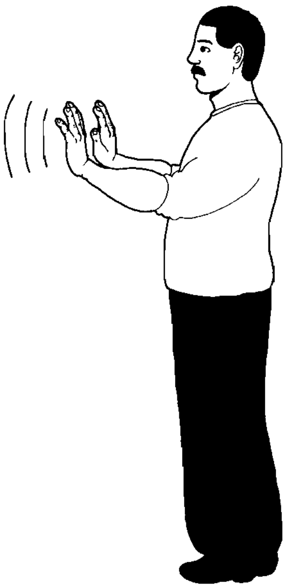

Figure 43.5. Press the palms forward three times.

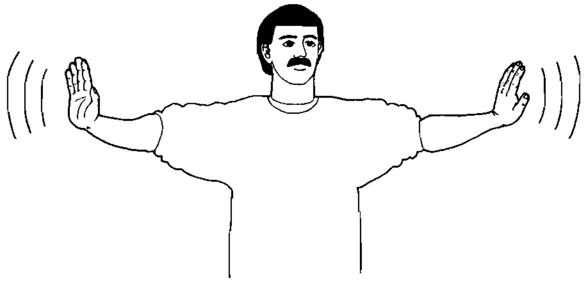

Figure 43.6. Press the palms to the sides of the body three times.

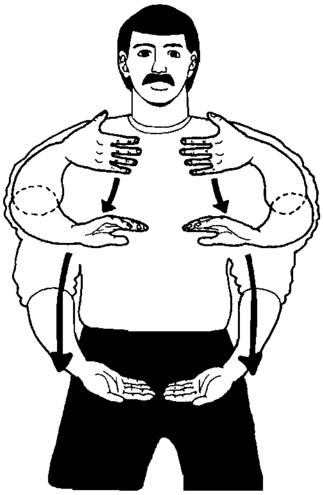

Figure 43.7. Direct the Qi to flow downward from the Middle to the Lower Dantian three times.

the sides. Focus your attention on extending the Qi out the palms to the fingertips. Press the palms out to the sides three times (Figure 43.6).

4. Draw both palms back to the front of the chest, palms facing upwards, elbows down. Rotate the palms to face downwards, and circle them down to the pubic arch, while focusing the mind on both palms (Figure 43.7). When Qi flows to the Lower Dantian turn the palms to face upward as if embracing a ball at the Lower Dantian.

Perform this exercise three times, then place both hands against the sides of the body. This completes one set. Do three sets.

MEDICAL QIGONG THERAPY FOR SPECIFIC LIVER DISEASES

The following are several Medical Qigong treatments used in China to treat specific Liver diseases.

DISORDERS OF THE BILIARY TRACT

The biliary tract includes the organs and ducts that participate in the secretion, storage, and delivery of bile into the duodenum.

ETIOLOGY

Disorders of the biliary track mainly include cholecystitis (inflammation of the Gall Bladder), cholelithiasis (formation of calcium or bile stones in the Gall Bladder), and ascariasis of the biliary tract (infestation by ascaris lumbricoides parasites). Traditional Chinese Medicine teaches that the first two diseases belong to the category of hypochondriac pain or jaundice, while the last condition is called biliary ascariasis.

Disorders of the biliary track can be caused by mental depression, stagnation of Gall Bladder Qi, the excessive intake of fatty foods, exopathogenic invasion, failure of biliary drainage system resulting from stagnation of Dampness and Heat, or obstruction by ascarid parasites. Although the causes and symptoms of these disorders are different, the Medical Qigong treatments and prescriptions are the same.

SYMPTOMS

The onset of biliary tract disorders are usually acute, and the pain is felt on the right upper abdomen and the right hypochondrium area of the thorax. Other symptoms include nausea, vomiting, rigor, high fever, yellow-colored skin pigmentation (jaundice), yellow sclera of the eyes, and whitish-grey stool.

Patients with biliary ascariasis may experience severe colic or a tearing pain below the xiphoid process. The pain is sometimes so severe that the patient experiences extreme perspiration accompanied by nausea and vomiting. If the ascarid parasites withdraw from the biliary tract, the pain is immediately relieved, but can return intermittently. If the ascarid parasites move completely into the Gall Bladder, the pain will be continuous, causing Gall Bladder distention. Symptoms such as jaundice, rigor, and high fever may appear, and a tenderness on the upper right quadrant of the xiphoid process can be felt with deep pressure.

TREATMENT

Cholelithiasis (gall stones) and cholecystitis (inflammation) are often treated by guiding the Toxic Qi downward and out of the patient's lower extremities along the Liver and Gall Bladder Channels, and by balancing the Qi activities between the upper and lower quadrants of the patient's body, as well as the right and left sides.

1. To begin treating disorders of the biliary tract, place the patient in a sitting position on the table. Stimulate the Bl-18, -19, -20, -21 points beside the patient's lower thoracic vertebrae, paying particular attention to the right side of the patient's body.
2. Next, extend Qi into the Gall Bladder area for 24 breaths (on the front of the patient's body), stimulating this area with the Tiger Kneading Palm technique, using Pushing, Pulling, and Shaking Qi manipulations.
3. To regulate the patient's Qi, and relieve epigastric distension, the doctor purges the patient's Toxic Qi, guiding it to flow out of the body downwards along the patient's Gall Bladder and Stomach Channels and out of the feet.
4. Using the Vibrating Palm technique, stimulate the patient's Bl-18, -19, -20, -21 points on the right side of the body once more, this time for 28 breaths. Then use the same Vibrating Palm technique on the front side of the body while directing the energy towards the Gall Bladder for 28 breaths.
5. Finally extend energy into the patient's Bl-19 point on the right side of the body for 12 breaths, and using the Sword Fingers technique, lead the Qi down the right Bladder Channel to the foot and out the small toe three times.

TREATMENT MODIFICATIONS

When treating the condition of ascariasis of the biliary tract, also include rubbing the abdomen 18 times in a counterclockwise direction.

PRESCRIPTIONS AND HOMEWORK

In prescribing homework, place the patient in a standing Wuji posture, with both palms placed against the hypochondria. Have the patient slowly massage the Gall Bladder area while purging the organ with the sound "Shu" for 24 exhalations.

Next, have the patient sit or lie down in a comfortable setting. As the patient inhales, he or she should meditate on the word peace; as the patient exhales, he or she should focus on relaxation. The patient should focus on releasing all tension in the head, chest, Liver, Gall Bladder, back, abdomen, and waist, for 36 breaths.

GALLSTONES

A gallstone is a Gall Bladder stone formed as a result of an excess of cholesterol in relation to the bile acids. The formation of calculi which results from this skewed relationship of cholesterol and bile is believed in Traditional Chinese Medicine to be dependent on the Liver Qi.

ETIOLOGY

Stagnant Liver Qi is believed to be a prerequisite for the formation of gallstones. If the Liver Qi stagnates, the bile will not secrete properly, leading to an accumulation of Damp Heat in the Gall Bladder. The steaming action of Heat on Dampness over prolonged periods of time, leads to the formation of gallstones.

It is generally believed that gallstones are related to cholestasis, an infection of the biliary tract, and metabolic disturbances of cholesterol. Gallstones form when the bile contains more cholesterol than it can keep in solution. The cholesterol precipitates out to form gallstones. The incidence of gallstones in women is approximately twice that in men. Any factor which increases the patient's cholesterol or decreases bile acids may lead to the formation of gallstones (i.e., obesity, high fat diet, Liver disease, and bile stasis increase the risk of developing gallstones). There are three types of gallstones: Cholesterol Stones, Mixed Stones, and Pigment Stones.

1. Cholesterol stones are usually single stones, which are composed almost entirely of cholesterol, and correspond to the pattern of Damp Heat in the Liver and Gall Bladder.
2. Mixed stones are the most common type of stones, consisting of lamellated layers of cholesterol, bilirubin, and calcium.
3. Pigment stones are less common and are always numerous. They are composed of bile pigment.

Gallstones are clinically manifested as pain in the right upper abdominal quadrant and right costal region, and are classified in Traditional Chinese Medicine as being placed in the categories of either hypochrondriac pain or jaundice.

SYMPTOMS

- Liver Qi stagnation is manifested by pain and distention in the right upper abdominal quadrant and right hypochondriac region. The pain can range from mild to severe. The patient may experience a distention and fullness in the Stomach with no obvious fever or jaundice.
- Damp Heat in the Liver and Gall Bladder is manifested by persistent pain in the right upper abdominal quadrant and right hypochondriac region. The pain manifests proximally or radiates up to the right shoulder. The patient experiences abdominal distention, fever, and jaundice.

TREATMENT

Have the patient sit on the edge of the table. Begin stimulating the patient's middle-back area, focusing on the Bl-21, Bl-20, BL-19, and Bl-18 points, on the right side of the back and drain the Toxic Qi down the torso and out the GB-34 point below the knee.

Use the Vibrating Palm hand technique and emit Qi into the patient's mid-back area for 14 breaths at each point. Then emit Qi into the painful area of the front side of the patient's body for 28 breaths.

Next, applying the Extended Fan Palm hand technique, emit Qi into the patient's Gall Bladder for 24 breaths using the Pulling and Leading manipulation skills to draw the Qi down through the Gall Bladder and Stomach Channels and out the patient's feet.

TREATMENT MODIFICATIONS

When treating a patient with a Liver Qi stagnation, use the Pulling and Leading manipulations to draw Qi down the lower limbs of the patient's body.

When treating patients with Damp Heat in the Liver and Gall Bladder: use the Extended Fan Palm hand technique and emit Qi into the patient's CV-12 and St-21 points, leading the Qi down the right Stomach Channel to the feet.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice Liver Massage described in Chapter 39, while exhaling the sound "Shu" for 36 breaths.

PRESCRIPTION MODIFICATIONS

- For patients with Liver Qi stagnation, have the patient practice rubbing the hypochondrium while imagining the Toxic Qi descending down and out the body. Instruct the patient to practice the Descending the Yang and Ascending the Yin Technique (see Chapter 41).
- For patients with Damp Heat in the Liver and Gall Bladder, have the patient practice rubbing the hypochondrium while imagining the Toxic Qi descending down and out the body as he or she sounds the "Hu" sound to purge and regulate the Spleen.

CIRRHOSIS

Cirrhosis is a generalized disease marked by hepatic lesions. It is a chronic disease of the Liver characterized by the formation of dense lobular connective tissue, degenerative changes in the parenchymal cells, structural alterations in the Liver lobules, and sometimes fatty and cellular infiltration within the Liver.

ETIOLOGY

Cirrhosis refers to a series of pathological changes of degeneration, necrosis and regeneration of Liver cells, as well as the proliferation of fibrous tissue due to prolonged or recurrent damage of the Liver caused by various pathogenic factors. Cirrhosis is clinically manifested as a series of symptoms caused by a hypofunction of the Liver and portal hypertension.

SYNDROMES

1. Liver Qi stagnation and Spleen Deficiency results in lassitude, weakness, fullness in the chest and abdomen, hypochrondriac distension and pain, occasional nausea, belching, and vomiting.
2. Qi and Blood stagnation results in poor appetite, hypochrondriac distention and pain, and enlargement of the Spleen and Liver marked by spider nevus and Liver spots.
3. Water retention results in hardness and fullness of the abdomen with a severe fullness in the Stomach, distention of the abdomen or intestines (by gas), nausea, vomiting, and anorexia. Patients with an insufficiency of Spleen and Kidney Yang suffer from symptoms of cold limbs, pale complexion, and an aversion to cold. Patients with a Deficiency of Liver and Kidney Yin experience sporadic fever, irritability, and a dark complexion.

TREATMENT

Place the patient on the edge of the table and stimulate the Bl-18 and BL-20 points on the mid-back, as well as the Lv-13 and Lv-14 points (at the sides of the body), and the CV-12 point (at the center of the body). Use a Vibrating Palm hand technique to emit Qi into these points for a period of 14 to 18 breaths.

Use the Extended Fan Palm technique and emit Qi into the patient's mid-back at the Bl-18 points, as well as the sides of the torso at the Lv-13 and Lv-14 points. Purge and drain the Toxic Qi down and out the patient's body along the Gall Bladder and Liver Channels.

TREATMENT MODIFICATIONS:

When treating patients with Liver Qi stagnation and Spleen Deficiency, use the Vibrating Palm to emit Qi into the Bl-20 and CV-12 points.

When treating patients with Qi and Blood stagnation, use the Extended Fan Palm technique to emit Qi into the Lv-13 and Lv-14 points, conducting the Qi along the Gall Bladder Channel down the leg to GB-34, and out the body.

When treating patients with water retention, use the Extended Fan Palm technique to emit Qi into the Sp-21 points. Lead the Qi down the Spleen Channels into the lower limbs. For patients with a Spleen and Kidney Yang Defeciency or a Liver and Kidney Yin Deficiency, use a Vibrating Palm technique to emit Qi into the Lower Dantian and Mingmen areas.

PRESCRIPTIONS AND HOMEWORK

Have the patients practice self-regulation exercises to tonify the entire body's energetic system.

PRESCRIPTION MODIFICATIONS

1. For patients with Liver Qi stagnation and Spleen Deficiency, have the patients practice the Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool (see Chapter 40).
2. For patients with Qi and Blood stagnation, have the patient perform the Liver Massage while exhaling the sound "Xu."
3. For patients with water retention, have them practice rubbing the Yellow Court area (CV-12) and sounding the word "Hu" to dredge the Spleen and Stomach. For patients with a Spleen and Kidney Yang Deficient, have them practice Gathering the Sun's Essence (see Chapter 11), and for patients with a Kidney and Liver Yin Deficient, have them practice Gathering the Moon Cream. (see Chapter 11).

CHOLECYSTITIS

In Traditional Chinese Medicine, both sides of the hypochondrium are related to the Liver Channels. In Western medicine, however, only the right side reflects a possible Liver pathology. Thus, it is important to have a distinction between left and right hypochondriac pain. Cholecystitis consists of the inflammation of the Gall Bladder with or without gallstones. It is related to hypochondriac pain on the right side of the body. It is created when cholesterol is maintained in the bile solution within the bile acids. Hypochondriac pain on the left side of the body extending to the epigastrium and left scapula is often related to acute pancreatitis.

ETIOLOGY

Cholecystitis is mostly caused by a bacterial infection or an obstruction of the bile duct. Clinically it is characterized by pain in the upper right quadrant of the abdomen and dyspepsia, and may be accompanied by nausea, vomiting, and sweating.

SYNDROMES

- Damp Heat manifestations are alternating attacks of chills and fever. Sometimes there is fever without chills, nausea, vomiting, loss of appetite, and pain in the right hypochondrium region.
- Qi stagnation manifestations are distending or moving pain in the right hypochondrium, belching, acid regurgitation, nausea, and anorexia.

TREATMENTS

1. Have the patient sit at the edge of the table. Stimulate the patient's mid-back, emitting Qi into the Bl-18, Bl-19, and Bl-20 points as well as the GB-34 point below the right knee.
2. Using the Vibrating Palm, first emit Qi into the patient's mid-back for 14 breaths, then directly into the painful area on the front side of the patient's body for 28 breaths.
3. Using the Extended Fan Palm hand technique, emit Qi into the patient's Gall Bladder region for 24 breaths, then purge the Toxic Qi down their right Gall Bladder Channel to the base of the foot and out of the body.

TREATMENT MODIFICATIONS

- For patients with Damp Heat, use the Extended Fan Palm hand technique to emit Qi into the patient's CV-12 area, then purge the Toxic Qi down the Stomach Channels to the base of the feet and out the body.
- For patients with Qi stagnation, use mostly purging and dredging Qi manipulations, focusing on the patient's lower torso.

PRESCRIPTIONS AND HOMEWORK

It is advisable to have the patient perform the Six Healing Sound Regulation exercise (see Chapter 40).

PRESCRIPTION MODIFICATIONS

- For patients with Damp Heat, it is advisable to have them perform the Spleen Massage (see Chapter 39) while exhaling the sound "Hu" to purge the Spleen and Stomach, allowing the Toxic Qi to flow down the Spleen and Stomach Channels into the ground.
- For patients with Qi stagnation, it is advisable to have them perform the Liver Massage (see Chapter 39) while exhaling the sound "Shu."

HYPOCHONDRIAC PAIN

Hypochondriac pain is a syndrome characterized by pain in one or both sides of the hypochondrium. The hypochondrium is located on the part of the abdomen beneath the lower ribs on each side of the epigastrium.

Chronic hepatitis, hepatolithiasis, cholelithiasis, cholecystitis and other diseases related to the Liver and Gall Bladder can be differentiated and treated in accordance with the exercises described in the prescription section.

ETIOLOGY

Hypochondriac pain may be caused by different factors such as trauma, the accumulation or blockage of Qi, stagnation of Phlegm and the Deficiency or Excess of Liver Qi (commonly seen in patients with Excess Liver Fire or Liver Qi stagnation). The symptoms of hypochondriac pain may occur in one or both sides.

SYNDROMES

- An Excess syndrome occurs when a patient suffers from Excess Liver Fire. This condition manifests with symptoms such as severe pain in the hypochondrium area with difficulty in breathing. The pain may radiate from both sides of the body, or from one side only.
- A Deficient syndrome occurs when a patient suffers from a Liver Yin and Kidney Yin Deficiency (often caused by mental depression or hemorrhage). This condition is observed clinically with symptoms such as a dull or sharp pain in the hypochondrium areas.

TREATMENT

1. Begin by stimulating the patient's Middle Dantian and Lv-14 and Lv-13 points at the front of the chest in the hypochondrium area (on both sides or the right side only). Next, stimulate the patient's mid-back around the Bl-17 and Bl-18 points and the base of the outside of the knees around the GB-34 areas.
2. Using the Extended Fan Palm hand technique, emit Qi into the patient's Lv-13 and Lv-14 points and the painful region of the chest and hypochondrium and begin purging the Toxic Qi, dredging it down the body and out of the extremities through the patient's Liver and Gall Bladder Channels.

TREATMENT MODIFICATIONS

- For patients with an Excess syndrome, extend energy into the patient's mid-back at the Bl-18 points. Then grasp and softly rock this tissue area to free the trapped Qi.
- For patients with a Liver and Kidney Yin Deficiency, emit Qi into the patient's Lower Dantian and Mingmen area using the Extended Fan Palm technique for 12 to 24 breaths.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient perform the Liver Massage (see Chapter 39) and sound the healing tone "Shu" for 36 breaths.
2. If the pain is on the left side of the patient's chest, have the patient place his or her palms on the painful area and begin to purge the Toxic Qi from the midline of the thorax, down the left side of the body, following the Gall Bladder Channel to the ground, for nine breaths.
3. Finally, have the patient stand in a Wuji posture and focus the mind's intention on his or her palms. The patient then raises the arms sideways, like a bird stretching its wings (creating a "T" formation). Next, both arms embrace (as if holding an imaginary ball) at the Middle Dantian level. The patient imagines the Qi from the middle chest flowing downward and rooting into the Lower Dantian.

CHAPTER 44
QI EMISSION THERAPY FOR HEART DISEASES

INTRODUCTION

The Heart is a solid (Zang) organ and controls Blood circulation. The Heart corresponds to the element Fire and is sometimes referred to as the Red Emperor. The Heart houses the Shen and is in charge of mental activities and long term memory. It is energetically paired with the Small Intestine (a hollow organ), which is in charge of separating the pure, clean energy from the impure Qi (Figure 44.1).

The Heart's upper external connection with the outside is through the tongue, hence sensations of the tongue (from food, kissing and so on), as well as talking can travel directly into the spirit, mind, and emotion, affecting the Heart's desires and senses.

In the early summer months, the Heart energy becomes more active in individuals who already have strong Heart Qi, but can become Deficient in those who already have weak Heart Qi. Excessive eating and drinking of bitter foods, and overexposure to heat depletes the Heart.

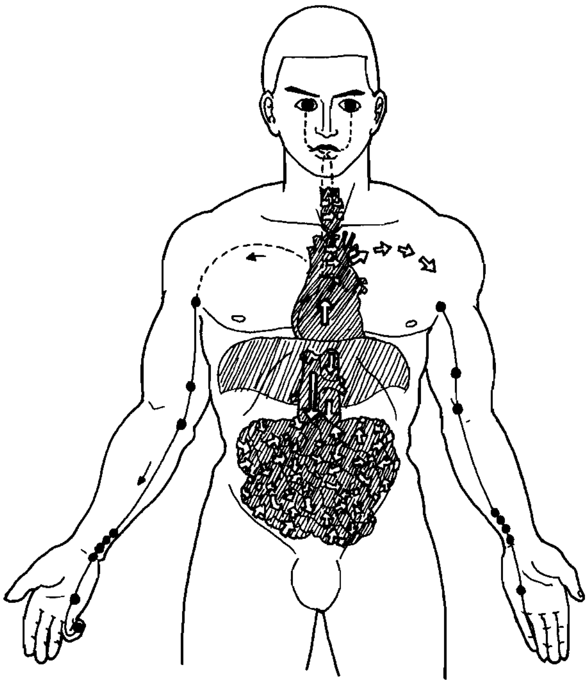

Figure 44.1. The Heart (Ht) Channels

HEART COMPLICATIONS AND SYMPTOMS

The Heart is responsible for governing the flow of Blood through the body's arteries and veins. Coronary artery disease (CAD) is a disease in which the patient's coronary arteries begin to harden, or impede adequate vascular flow to the myocardium, resulting in an insufficient supply of Qi and Blood throughout the body. This causes such diseases as hypertension, vasculitis, myocarditis, congenital Heart disease, rheumatic Heart disease, nervous malfunctioning of the Heart, organic pathological changes of the Heart, and arteriosclerosis of the brain.

In general, patients with Heart disease may have an arterial pulse above the navel which often disappears upon pressing, as well as a hot sensation in the soles of the feet and palms, dry mouth, stiffness of the tongue, and absentmindedness. A list of Heart syndromes and their symptoms are described as follows.

1. Heart Qi Deficiency can cause such symptoms as: physical tiredness, palpitations, sweating, pallor, and shortness of breath when exerting oneself.
2. Heart Yang Deficiency can cause such symptoms as: tiredness, tightening in the chest, and edema.
3. Heart Blood Deficiency can cause such symptoms as: headaches, mental and emotional problems, tinnitus, and tiredness.

4. Heart Blood stagnation can cause such symptoms as: mental and emotional problems, as well as tightness or pain in the chest.
5. Heart Yin Deficiency can cause such symptoms as: tiredness, insomnia, and mental and emotional problems.
6. Heart Fire can cause such symptoms as: insomnia, painful urination, and mental and emotional problems.

GENERAL TREATMENT FOR HEART DISEASES

The following is a description of a general protocol used for treating a patient's Heart. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi is removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Heart therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor can access the patient's Heart through four methods:
   - direct energetic insertion into the Heart's anatomical location,
   - the patient's channels,
   - the external energetic Wei Qi fields, or
   - the patient's Taiji Pole (Figure 44.2).

   To avoid energetic armoring (while accessing the Qi in the front of the Heart), it is important for the doctor to already be connected to the patient's Shendao point (through the back of the Heart).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Heart and Shendao area, and roots the Upper Dantian Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Pericardium and Heart Channels and out from the patient's hands (Figure 44.3).

Keep in mind that these areas are general and

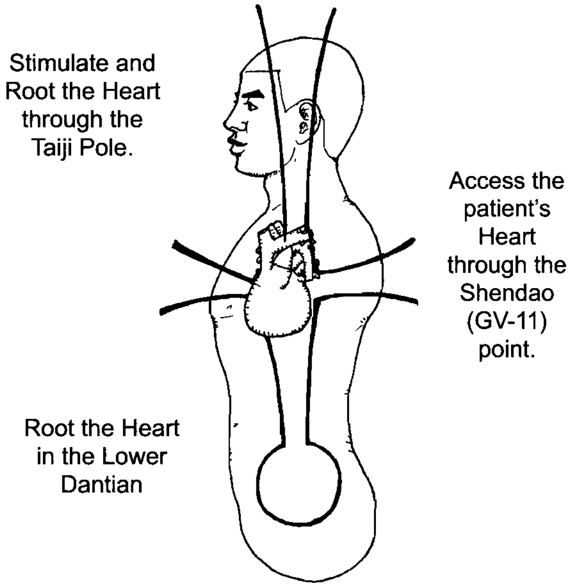
Stimulate and Root the Heart through the Taiji Pole.
Access the patient's Heart through the Shendao (GV-11) point.
Root the Heart in the Lower Dantian

Figure 44.2. Treating Heart Conditions

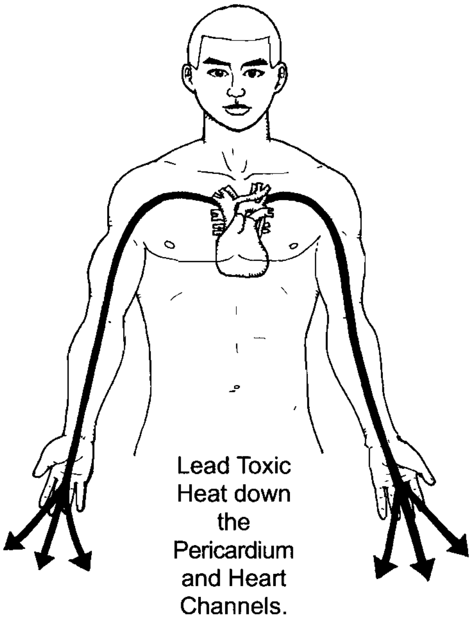
Lead Toxic Heat down the Pericardium and Heart Channels.

Figure 44.3. Lead the Toxic Qi out of the body through the patient's Pericardium Channels.

that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR HEART QI DYSFUNCTION

The primary therapies used to treat cardiovascular diseases include: the Windy Breathing method to strengthen the Heart and Kidneys, Toe-Raised Walking, Heart Massage, and Quiescent meditation (see Chapter 15).

- Patients with an Excessive Qi syndrome of the Heart with extreme Heat may have a flushed complexion, full pulse, and excess sputum; they may feel pain in the chest, hypochondrium, back, shoulders, and arms. This syndrome can be treated by having the patient utter the "Ha" sound while guiding the Kidney Water Qi upwards to suppress the Heart Fire (Figure 44.4). Alternatively, the Qigong doctor may guide the Evil Qi out of the patient's body along the Heart and Small Intestine Channels with Pulling and Leading Qi manipulations.
- Patients with Deficient Heart Qi may dream of flames and red colored objects and have a radiating pain in the abdomen. This condition can be treated by having the patient perform the Taking in Yellow Qi meditation (see Chapter 4) which replenishes the child (Spleen), and Taking in Pink Qi meditation to nourish the mother (Heart); or, by using the Pushing and Guiding Qi manipulations to reinforce the Heart Qi and the Qi in the Lower Dantian.

In this particular case (Deficient Qi syndrome of the Heart), the reason the patient should not take in the energetic color red is because the Heart has already established a draining pattern, and any energy pouring into the Heart will only continue to drain away.

HEART QIGONG PRESCRIPTIONS

To tonify the Heart, tranquilize the mind, promote Blood circulation, and remove obstructions in the channels, prescription exercises can be used. The following exercises can be used to prevent and treat diseases and syndromes such as coronary Heart disease, hypertension, angina, arrhythmia, rheumatic Heart disease, cardiac neurosis as well as palpitations (ranging from mild to severe), pericardial pain, and insomnia.

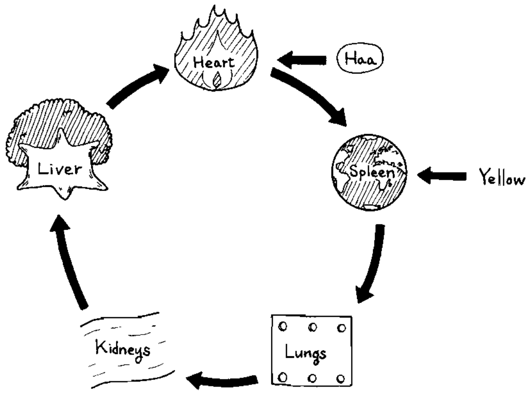

Figure 44.4. Treating Patients with an Excess Qi Syndrome of the Heart

PRESCRIPTIONS FOR TONIFYING THE HEART

Taking in the Red Qi can be used to tonify the Heart. The properties inherent within the color red have a persistent vibratory rate used for tonifying the Heart and Small Intestine.

1. Assume a standing, sitting, or lying posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts.
2. As you inhale, visualize red energy in front of you. Inhale the red light energy in through the nose and down into the Heart organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Heart; however, the bright, clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the red energy in through the nose and fill the entire mouth with it. When exhaling through the nose, send the Qi slowly to the Heart, then to the Lower Dantian, to connect the energy of the Heart and Kidneys and to remove obstructions in the whole body. Perform this exercise seven or fourteen times. Return to the first step and close the training.

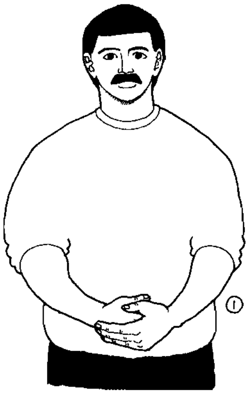

Figure 44.5. Hands rest in front of the Lower Dantian.

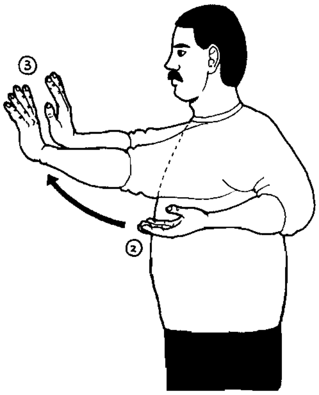

Figure 44.6. Both palms simultaneously press forward.

PRESCRIPTIONS FOR PURGING THE HEART

Purge the Heart by having the patient rub his or her chest while sounding "Ha."

1. The preparatory postures are the same as above.
2. Begin by clicking the teeth, stirring the tongue, and swallowing the saliva, then place the palms on the pectoral region of the left side of the chest (left hand on the outside for men, opposite for women). Inhale slowly, then gradually exhale the sound "Ha," focusing the mind on the Heart area. Repeat for six to twelve breaths altogether while rubbing with the palm lightly in a clockwise circle.

PRESCRIPTIONS FOR REGULATING THE HEART AND CONDUCTING QI

The following exercise is used to regulate the Heart after purging or tonifying.

1. Assume a standing or sitting posture. Relax the whole body, breathe naturally, and raise the tongue against the hard palate, just behind the teeth. Bring both palms loosely together (left over right for men, opposite for women) at the Lower Dantian; stand still for several minutes and concentrate the mind on the Lower Dantian (Figure 44.5).
2. Turn both palms outward, and push the arms along the sides of the body towards the back. Hold this posture for several minutes.
3. Turn both palms upward and lift them to the sides of the chest.
4. With both palms facing forward, stretch both hands level and forward. Focus your attention on the ends of the middle finger of each hand, while the thumb and little fingers stretch and separate (Figure 44.6).
5. Clench both hands into fists and pull them along the sides of the body towards the back, as if dragging heavy weights. Perform this movement three times (Figure 44.7).
6. Raise the right palm, as if holding a heavy weight in front of the chest, and push the palm out vertically to the right. Then, draw the right hand back to the side of the body near the hip. Repeat on the left side, and finally return to the starting posture. Perform these movements three times (Figure 44.8). This completes one set. Perform three sets.

SPECIFIC MEDICAL QIGONG THERAPY FOR HEART DISEASES

The following are several Medical Qigong Treatments used in China to treat specific Heart diseases.

HYPERTENSION

Hypertension is considered a condition in which the patient has higher blood pressure then what is determined to be normal for his or her

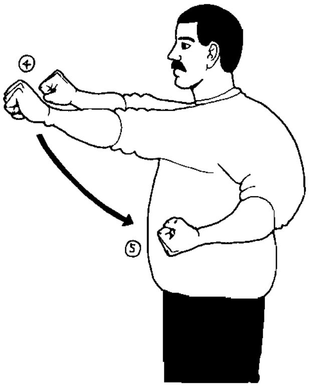

Figure 44.7. Both hands simultaneously clench into fists and pull to the hips.

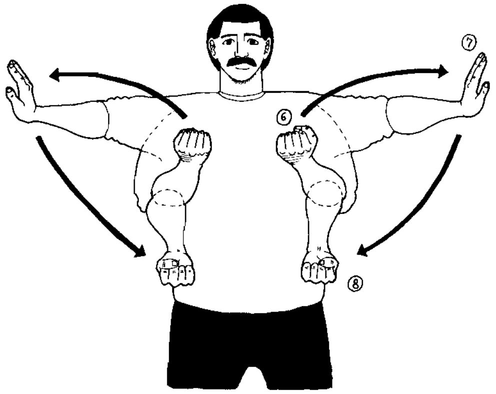

Figure 44.8. Alternate each palm as it presses towards the side of the body.

chronological age and physique. Generally, Traditional Chinese Medicine places it in the categories of vertigo and headache.

ETIOLOGY

Hypertension is one of the most common problems impacting the Heart. It is the result of an imbalance of the Yin and Yang functional aspects of Deficient Kidney Yin and Excess Liver Yang, and/or an overabundance of Phlegm and Dampness within the body.

Hypertension can be caused by the patient eating too much food or salt, by too much stress, or not enough exercise. A renin imbalance within the patient's Kidneys, increased cholesterol, a preexisting genetic predisposition, a secondary illness (diabetes, nephritis, or hyperthyroidism), or even pregnancy can also lead to hypertension. These causes of hypertension can result in one of three internal factors: high cardiac output due to stress, reduced elasticity of the vessels, and resistance to Blood flow.

1. When the body experiences emotional tension it releases adrenaline, which speeds up the Heart rate and causes the blood vessels to narrow, increasing the body's blood pressure.
2. A buildup of plaque and normal aging reduces the vessel's elasticity and can lead to hypertension.
3. Old age, certain hormonal imbalances or disorders, or a genetic predisposition can likewise cause high blood pressure.

The overall goals in treating hypertension patients are to:

1. Lower blood pressure,
2. Stabilize blood pressure,
3. Establish a quiescent state of mind (generally, meditating 20 minutes, twice a day provides the patient with a prolonged state of deep relaxation needed to combat stress), and
4. Encourage life-style changes (i.e., monitor potassium intake, decrease sodium intake, weight control, and stress reduction management). Treatment programs are established as described below.

SYNDROMES

As previously stated, there are three underlying causes or syndromes relating to hypertension: Excess Liver Yang, Deficient Yin, Excess Liver Fire, and Kidney Yin and Yang Deficiency.

1. If the body's Kidney Yin is Deficient, it fails to nourish the Liver Yin, which leads to a condition of hyperactivity of Liver Yang. The clinical manifestations are feelings of fullness in the head, daydreaming, headaches, dizziness, tinnitus, insomnia, and dysphoria with feverish sensations in the chest.
2. Excess Liver Fire can cause headaches, dizziness, restlessness, irritability, flushed face, and constipation.
3. If the body's Kidney Yang is Deficient the Yin (Water) accumulates, if the Yang does not move, the blood vessels cannot relax. This results in

obstructed Blood flow. The clinical manifestations are dizziness and tinnitus, daydreaming, palpitations, amnesia, lassitude and weakness in the legs, as well as insomnia.

TREATMENT FOR PATIENTS SENSITIVE TO ENERGY

1. Use Medical Qigong therapy, and have the patient sit in a chair with his or her feet flat on the ground. Begin the treatment by cleaning the patient's Conception and Governing Vessels, using purgation methods. Focus specific attention on clearing the patient's head area (Figure 44.9). Literally sweep with the entire hand, and comb the patient's Wei Qi fields with the fingers.

   By scanning the occipital region of the patient, you can verify that the Excess Qi in the patient's upper extremities has abated. Once these vessels are clean, you can focus on treating the hypertension.
2. Stand behind the patient, using the Extended Fan Palm technique (with both hands) extend energy from the left hand into the top of the patient's head at the Baihui GV-20 area, while your right hand extends energy into the back of the patient's head, beginning at Yuzhen Bl-9. The two streams of Qi flow downward to the Dazhui GV-14 area, where they join into one river, which continues to flow downward through the patient's Taiji Pole. Continue to visualize these two streams joining and filling the Lower Dantian, for a period of 6 to 12 breaths (Figure 44.10).
3. Next, move your left hand and extend energy into the patient's Yellow Court, while your right hand is placed above the patient's Mingmen area. Lead the Qi from the Yellow Court into the patient's Lower Dantian, continuing to root the Qi, for a period of 6 to 12 breaths.

TREATMENT MODIFICATIONS

These modifications are additional formulas which are added onto the previous treatment according to the specific cause and condition of the patient's Heart.

1. When treating patients with Excess Liver Yang and Deficient Yin, use the Extended Fan Palm

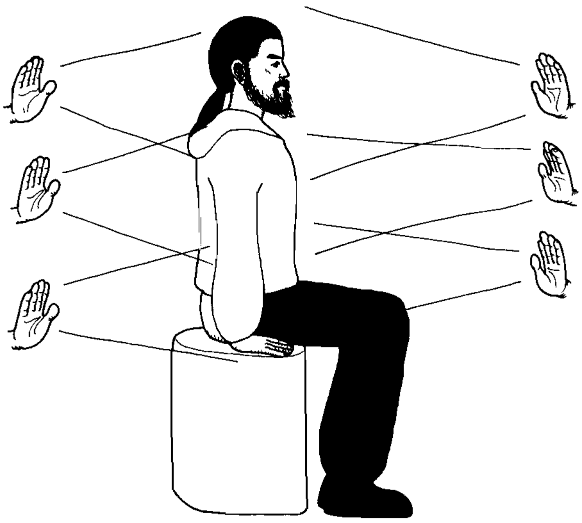

Figure 44.9. The Qigong doctor begins by dredging, purging and cleaning the patient's Conception and Governing Vessels.

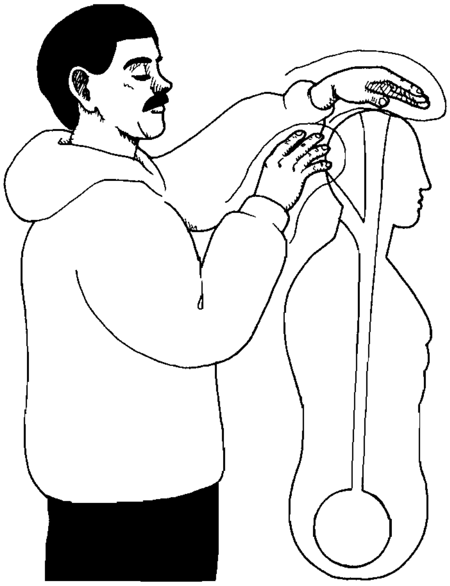

Figure 44.10. The Qigong doctor emits Qi into the patient's Baihui, Yuzhen and Dazhui points, in order to access the Taiji Pole and fill the patient's Lower Dantian.

hand technique and emit Qi into the patient's mid-back, drawing the energy down and rooting it in the patient's Lower Dantian and Mingmen.

2. When treating patients with Excess Liver Fire, it is important to treat the Lv-13 and -14 points, by dredging and purging the Excess Qi, and leading it down the patient's right Gall Bladder Channel, through the foot and into the Earth.
3. When treating patients with Kidney Yin and Yang Deficiency, use the Extended Fan Palm technique to emit Qi into the patient's mid-back, drawing the energy down and rooting it in the patient's Lower Dantian and Mingmen.

TREATMENT FOR PATIENTS NOT SENSITIVE TO ENERGY

Use Jing Point therapy in conjunction with Medical Qigong therapy (see Chapter 36). Have the patient lie prone. Using the Sword Finger technique, touch the patient's head with the left hand and extend energy into the Baihui GV-20 point. Guide this Qi down to the Lower Dantian through the Taiji Pole, while your right hand embraces and energizes the patients' Mingmen GV-4 point, thereby filling the Lower Dantian and rooting the patient's Heart Qi (Figure 44.11). Next, treat the patient's LI-11, St-36, Lv-2, and Sp-6 points on both sides of the body to lower the patient's blood pressure.

PRESCRIPTIONS AND HOMEWORK

1. A static posture can be given to convalescing patients as part of a meditative homework prescription for treatment of high blood pressure. The prescription is as follows.
   a. Have the patients lie supine (the patients may use a pillow under the head for comfort) and have him or her imagine that they are floating in warm water.
   b. The patients should perform the toning sound "Zheng" (prenounced Jang), and imagine dispersing Qi out of his or her body. The patient should use natural breathing while toning and dispersing the Toxic Qi. There are ten major areas, from the top of the head to the bottom of the feet, from which the patients must discharge the toxins. These areas include: the top of the head, Upper Dantian, throat, Middle Dantian, Yellow Court, Lower Dantian, the center of the thighs, knees, ankles, and bottom of the feet (Figure 44.12).
   c. After several minutes the patients will relax and focus their attention on the Lower Dantian or Bubbling Well points at the bottom of the feet.

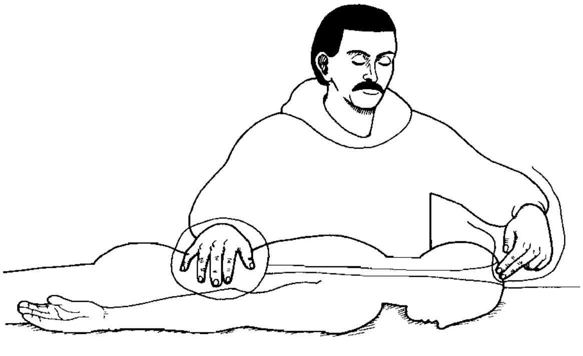

Figure 44.11. If the patient is not sensitive to energy, have them lie prone, and apply Jing Point Therapy in conjunction with Qi emission.

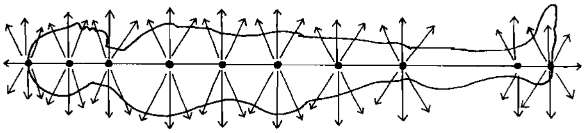

Figure 44.12. In this Static posture used to treat high blood pressure the patient lies supine and performs the toning sound "Zheng" (Jang), for dispersing Qi.

2. Another meditation prescription given to patients with hypertension is to have them sit and imagine the feeling of warm water pouring down over the head, torso, extremities, and feet. This meditation should be practiced several times a day, using 36 breaths each time.
3. The patient can also be given the prescription of rubbing the hypochondrium and toning the sound "Shu" to purge Excess Heat from the Liver.

PALPITATION

A palpitation is an abnormally rapid throbbing or fluttering of the Heart. The symptom is characterized by nervous feelings or restlessness due to strong anxitey or fright. Clinically, it mostly attacks paroxysmally and is caused by emotional stress or overexertion. Symptoms such as insomnia, vertigo, tinnitus, etc. often accompany palpitations.

Diseases with palpitation as the main manifestation include arrhythmia, iron-deficiency anemia, aplastic anemia, hyperthyroidism and anxiety disorders. These diseases may all be differentiated and treated in reference to the prescriptions described in this section.

ETIOLOGY

There are several factors which contribute to palpitations. These factors and their symptoms are described as follows.

1. In patients suffering from Deficient Qi and Blood, and Deficient Heart Qi, palpitations may occur following sudden mental irritation. The patient may note unduly rapid heartbeat accompanied, in cases of Deficient Heart Qi, by pallor, restless sleep at night, dizziness and vertigo. This condition may also be caused by Deficient Heart Blood that fails to nourish the Heart.
2. Phlegm Fire disturbing the Heart may also give rise to palpitations. In cases with internal disturbances of Phlegm Fire, symptoms also include fidgeting and disorientation.
3. Another factor causing palpitations is a Deficiency of Heart Yang that leads to fluid retention. In cases of fluid retention with epigastric stiffness symptoms will also include: dizziness, excessive salivation and mental fatigue.

TREATMENT

1. Begin by extending Qi into the patient's Bl-14, Bl-18, CV-15, and CV-17 points. To open up these points, emit Qi into the tissue areas surrounding these points using the Extended Fan Palm technique.
2. Next, use the Sword Fingers hand technique and vibrate Qi into each of these points for the duration of six to twelve breaths at each point. Upon completion, the doctor will then guide the Qi back down the patient's torso and root it into the patient's Lower Dantian.
3. Finally, the doctor emits Qi into the patient's Yellow Court area using the Extended Fan Palm technique and then guide the Heart and Pericardium Channel Qi down the patient's arms using Pushing, Pulling and Leading techniques. This ending is performed to balance the Qi in the upper and lower aspects of the patient's body.
4. As with every treatment, end the procedure by regulating the patient's Microcosmic Orbit (Fire Cycle).

TREATMENT MODIFICATIONS

These modifications are additional formulas which are added onto the previous treatment according to the specific cause and condition of the patient's Heart.

1. In cases of insufficiency of Heart Blood, add Bl-17 and Bl-20 to the treatment points.
2. In cases of internal disturbance of Phlegm Fire, purge from the CV-12 to St-40 points.
3. In cases of Fluid retention, add Bl-21 and Bl-22 to the treatment points, for 14 to 24 breaths. Then, guide the Qi to flow down the patient's Gall Bladder Channels into their Bl-40 points.

PRESCRIPTIONS AND HOMEWORK

1. For treatment of palpitations due to Deficient Heart Blood, have the patient practice the method of Taking in the Red to nourish the Heart (see Chapter 4). Have the patient send the red Qi slowly down into the Heart and then to the Lower Dantian during the exhalation. Practice should be for several minutes, according to the patient's strength and constitution.
2. For treatment of palpitations due to Phlegm Fire, have the patient place both hands on the Heart and slowly inhale. As the patient exhales, he or she should pronounce the sound "Ha." This exercise should continue for several minutes.
3. For Heart regulation, have the patient practice the Heart exercise from the Daoist Five Yin Organ prescriptions (see Chapter 42).

RHEUMATIC HEART DISEASE

Rheumatic valvular Heart disease is a condition where there is chronic damage of the cardiac valve caused by rheumatosis. It clinically manifests with palpitations, dyspnea and edema.

SYNDROMES

1. Heart Blood stagnation is manifested through symptoms such as severe palpitations, bad cough with hemoptysis, lassitude and weakness, purplish-red color of the lips, nails and zygomatic facial regions, pain in the Heart, chest, or both.
2. Deficient Qi and Blood manifests through such symptoms as shortness of breath, palpitations which are aggravated on exertion, perspiration and pale complexion.
3. Heart and Kidney Yang Deficiency manifest through such symptoms as palpitations, pale or darkish complexion, edema, cough, cold hands and feet and dyspnea.

TREATMENT

1. Have the patient sit at the edge of the table, and begin to stimulate the Shendao (GV-11) and Middle Dantian areas.
2. Use the Vibrating Palm hand technique and emit Qi into the patient's Yellow court and Shendao (GV-11) area for 6 to 12 breaths.
3. Use the Extended Fan Palm hand technique and emit Qi into the patient's Heart region, leading Qi down the Heart and Pericardium Channels to the tips of the fingers (on both sides of the body).

TREATMENT MODIFICATIONS

1. When treating patients with Heart Blood stagnation, use the Extended Fan Palm hand technique and emit Qi into the patient's Middle Dantian, leading the stagnant Qi down the Heart and Pericardium Channels and out the hands (on both sides of the body).
2. When treating patients with Deficient Qi and Blood, use the Extended Fan Palm technique and emit Qi into the patient's Middle Dantian (both Shendao GV-11 and CV-17 areas), leading the Qi down the Conception Vessel to the Lower Dantian to root and stabilize the patient's Qi.
3. When treating patients with Heart and Kidney Yang Deficiency, use the Extended Fan Palm technique and emit Qi into the patient's Shendao (GV-11) and Mingmen (GV-4) for 24 breaths.

PRESCRIPTIONS AND HOMEWORK

When prescribing homework for patients with rheumatic Heart disease, have them begin with a quiescent sitting meditation, while focusing on their Lower Dantian.

PRESCRIPTION MODIFICATIONS

1. Have patients with Heart Blood stagnation perform the Heart Massage exercise (see Chapter 39), while toning the sound "Ha."
2. It is advisable for patients with Qi and Blood Deficiency to practice the method of Taking in the Yellow Qi into the Spleen and Red Qi into the Heart from the natural environment.
3. Have patients with Heart and Kidney Yin Deficiency practice Gathering the Moon Cream (see Chapter 11), as well as the method of Taking Yellow Qi into the Spleen and Red Qi into the Heart from the natural environment (see Chapter 4).

CONTRAINDICATIONS

The Heart may either have functional or structural problems. In either condition, Medical Qigong is contraindicated when working directly over a pacemaker.

Also, when treating Heart conditions, make sure the patient is under a Western doctor's care, and inquire as to any medications the patient may be taking.

CHAPTER 45
QI EMISSION THERAPY FOR SPLEEN/STOMACH DISEASES

INTRODUCTION

The Spleen is a solid (Zang) organ and it stores and nourishes Ying (Nourishing) Qi. The Spleen corresponds to the element Earth and is sometimes referred to as the Yellow Emperor. It houses the Yi, which governs the intention, thought, study, concentration, and memorization. It is energetically paired with the Stomach (a hollow organ). The Spleen is in charge of transporting, distributing and transforming nutrients and Qi into Gu Qi, as well as controlling Blood circulation (Figure 45.1).

The Spleen's external connection with the outside world is through the mouth, hence any connection with the mouth can affect the body's acquired constitution, developing a tolerance or intolerance to specific substances (allergies), as well as affect concentration and thought.

In the late summer months, Spleen Qi becomes more active in individuals who already possess strong Spleen Qi, but can become Deficient in those individuals who already have weak Spleen Qi. Excessive eating and drinking of sweet foods, and overexposure to dampness weakens the Spleen.

SPLEEN COMPLICATIONS AND SYMPTOMS

Because of the importance of both the Spleen and Stomach, this chapter is divided into two separate sections. The first section focuses on specific Spleen diseases and how to rectify them. The second section explores the energetic dysfunctions of the Stomach.

The following discussion of treatment begins with an explanation of Spleen complications and symptoms. Spleen syndromes have symptoms which are common to the organ's energetic dysfunction such as: diarrhea, tiredness, and edema. A list of Spleen syndromes and their symptoms are described as follows.

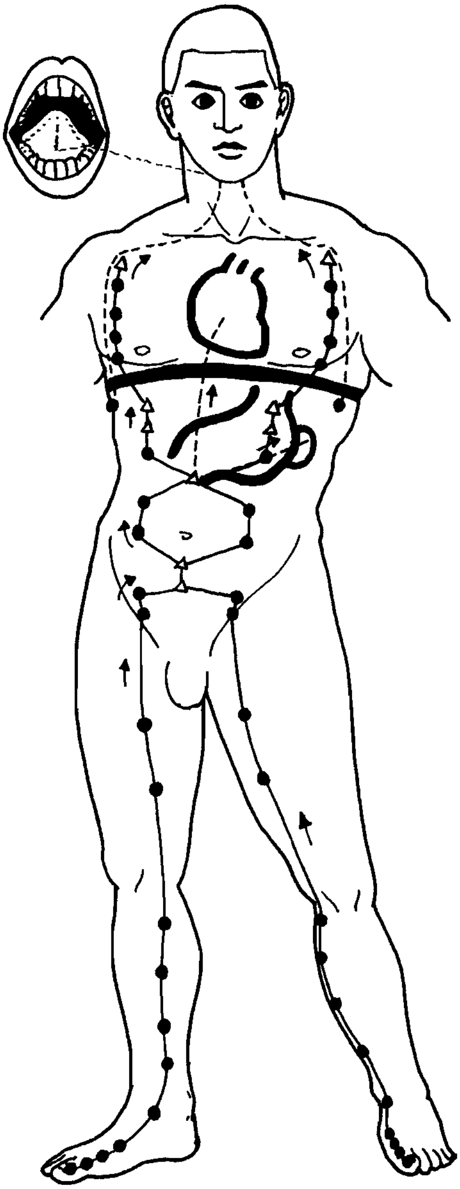

Figure 45.1. The Spleen (Sp) Channels

1. Spleen Qi Deficiency can cause such symptoms as: wheezing, tiredness, abdominal pain, diarrhea, constipation, painful urination, enuresis, bleeding and menorrhagia.
   a. Patients with a failure of normal transporting and transforming Qi activity in the Spleen may have a poor appetite, abdominal distention, loose stools, edema and retention of Phlegm.
   b. Deficient Spleen Qi may cause the Qi in the Middle Burner to sink. This results in shortness of breath, unwillingness to talk, persistent diarrhea, prolapse of the anus or uterus, or gastroptosis (downward displacement of the Stomach).
   c. Spleen Qi Deficiency can result in failure of the Spleen to control the Blood circulation and may be manifested by the following symptoms: bloody stool, bleeding from the uterus, and subcutaneous hemorrhaging.
2. Spleen Yang Deficiency can cause such symptoms as: tiredness and edema.
3. Spleen Blood Deficiency can result in fatigue.
4. Spleen Yin Deficiency can result in tiredness.
5. Damp Heat in the Spleen can cause such symptoms as: sinusitis, edema, common cold, and influenza.
6. Dampness in the Spleen can cause such symptoms as: diarrhea, tiredness, edema, and Wind Stroke.

GENERAL TREATMENT FOR SPLEEN DISEASES

The following is a description of a general protocol used for treating the Spleen. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Spleen therapies, the reader will notice that certain patterns repeat

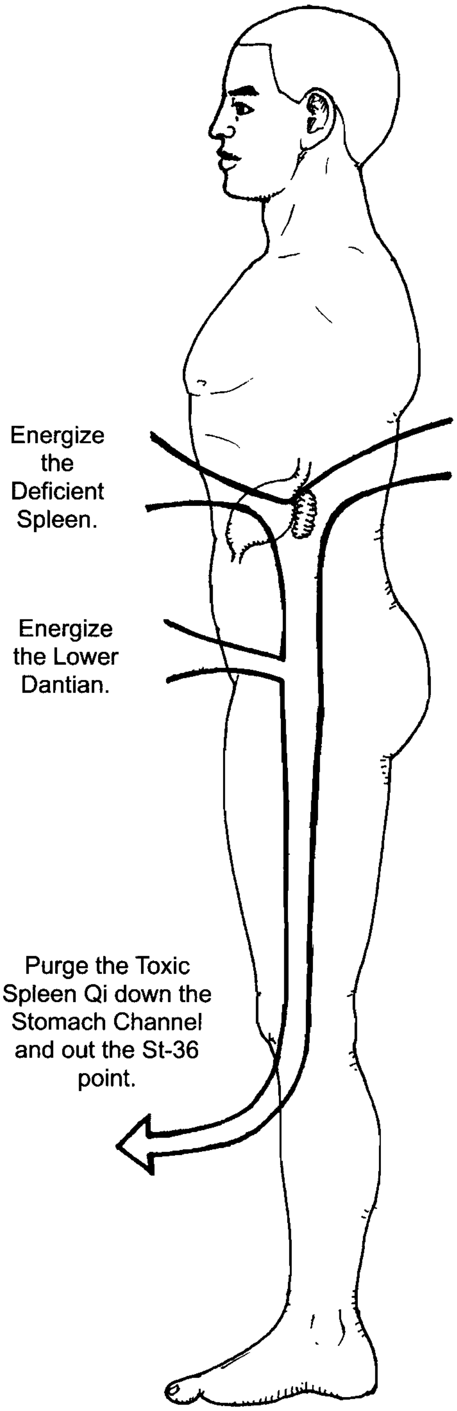
Energize the Deficient Spleen.
Energize the Lower Dantian.
Purge the Toxic Spleen Qi down the Stomach Channel and out the St-36 point.

Figure 45.2. Treating Spleen Conditions

themselves.

1. Generally, the Qigong doctor accesses the patient's Spleen through either direct energetic insertion into the Spleen's anatomical location, or through the patient's channels and external energetic fields.
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Spleen and Yellow Court areas, and roots the patient's Middle Burner Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Stomach Channels and out of the patient's St-36 points (Figure 45.2).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR SPLEEN QI DYSFUNCTION

The primary therapy used to treat the Spleen and gastric diseases include the Windy Breathing method (see Chapter 16) to strengthen the Liver, Spleen, Lungs, and the Kidneys.

1. Spleen diseases should be treated by vocalizing the "Hu" sound, which is effective for removing Turbid Qi accumulated in the Spleen, as well as for aiding digestion (Figure 45.3).
2. To treat insufficiencies of the Middle Burner, the doctor emits Qi into the patient's Zhong Wan CV-12 point and Lower Dantian points to reinforce the patient's Qi in the Middle Burner area.
3. Patients with an Excess Heat syndrome of the Spleen should be treated by guiding the Qi out along the Spleen and Stomach Channels with Pulling and Leading Qi manipulations to expel the Turbid Qi from both the Spleen and Stomach.

SPLEEN QIGONG PRESCRIPTIONS

The following exercise regulates the Qi and Blood of the Spleen Channels, strengthens the Spleen and replenishes Qi. It also regulates the Stomach and promotes digestion. It can be used to prevent and treat diseases and syndromes such as: gastritis, gastric and duodenal ulcers, colitis and gastrointestinal neurosis, abdominal distention, diarrhea and constipation.

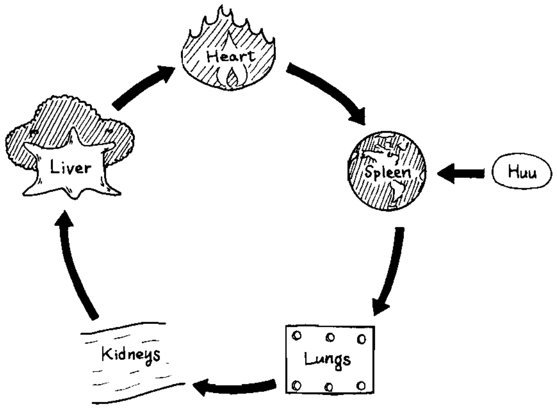

Figure 45.3. The "Hu" sound will purge the Spleen.

PRESCRIPTIONS FOR TONIFYING THE SPLEEN

The method of Taking in the Yellow Qi can be used to tonify the Spleen. The properties inherent within the yellow color have a persistent vibratory rate used for tonifying the Spleen and Stomach.

1. Assume a standing or sitting posture. Relax the whole body, breathe naturally, and get rid of any stray thoughts.
2. When inhaling, imagine yellow energy in front of you. Inhale the yellow light energy in through the nose and down into the Spleen organ. As you exhale through the mouth, the dark, pathogenic Qi leaves the Spleen; however, the bright clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the color yellow in through the nose, filling the entire mouth. When exhaling through the nose, send the Qi slowly to the Zhongwan CV-12 point, then out to the four limbs, skin, and hair. Repeat five to ten times. Return to the beginning posture and close the training.

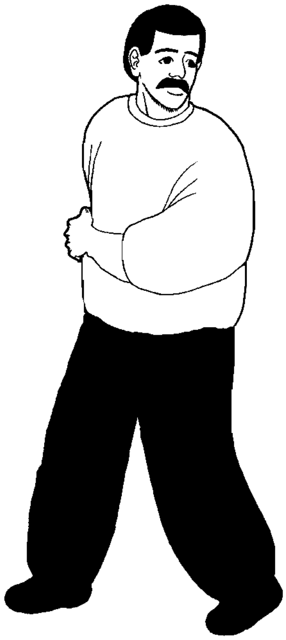

Figure 45.4. Twist at the waist, swing the arms, and look towards the opposite direction, focusing the mind's intention on the heel.

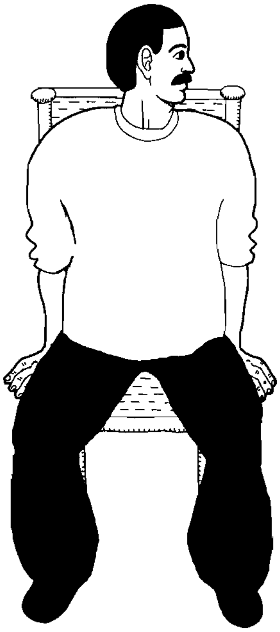

Figure 45.5. Turn and look over the shoulder with both eyes wide open.

PRESCRIPTIONS FOR PURGING THE SPLEEN

Purge the Spleen by having the patient rub his or her epigastrium while toning the sound "Hu."

Assume a sitting or standing posture. Gently place both palms (left on the outside for men, opposite for women) flatly on the Zhongwan CV-12 point on the upper abdomen. Exhale, rubbing the palms in a clockwise direction while sounding "Hu". Perform ten to twenty breaths.

Dredging the Spleen and Stomach

The following exercise is used to purge Toxic Qi from the Spleen and Stomach.

1. Assume a standing posture and breathe naturally. While relaxing the whole body, twist at the waist as you swing both arms in front of the body to the left, then to the right. Look in the opposite direction when turning the torso. As you turn your torso, your mind should be focused on your heels. Continue until you feel all tension has left your body (Figure 45.4).
2. Assume a sitting posture. While sitting on a bed, press the palms level on the mattress and meditate until you feel peaceful and relaxed. Turn the head towards the left and look over your left shoulder into the distance. Then turn the head to the right and look back over your right shoulder into the distance. Pause in the center, facing forward (Figure 45.5). When looking over the shoulders, both eyes should be wide open. Repeat five times in each direction.

PRESCRIPTIONS FOR REGULATING THE SPLEEN AND CONDUCTING QI

The following exercise is used to regulate the Spleen after purging or tonifying.

1. In a Wuji posture, begin by placing the hands on the Yellow Court and begin circle rotation in a clockwise direction for 36 rotations, ending at the Lower Dantian. Pause for 18 breaths to root the Spleen Qi in the Lower Dantian.
2. Next, rotate the abdomen in a counterclockwise direction for 36 rotations, ending at the Lower Dantian. Pause for 18 breaths and return to Wuji posture.

SPECIFIC MEDICAL QIGONG THERAPY FOR SPLEEN DISEASES

The following Medical Qigong treatment is used in China to treat specific types of Spleen disease.

DIABETES

Diabetes refers to a syndrome characterized by polydipsia (excessive thirst), polyphagia (excessive eating), and emaciation. It is a general term used to describe diseases characterized by insufficient secretion of insulin and excessive urination.

The prescriptions described in this section can also be used for the treatment of such disorders as: psychogenic polyuria, polyphagia due to hyperthyroidism, hyperhidrosis (excessive sweating), and emaciation.

ETIOLOGY

According to Traditional Chinese Medicine, diabetes is caused by disturbances in the Spleen and Stomach functions of transforming and transporting energy and the accumulation of toxic Heat in the body. These malfunctions can be due to eating too much sweet and greasy food, or drinking too much alcohol.

From the Western medical perspective, there are two types of diabetes, one type can be controlled through diet, the other cannot (usually hereditary). Diabetes is included in metabolic diseases (the Beta cells in the pancreas do not produce enough insulin) and caused by pathogenic factors relating to genetic predispositions, environmental stress, diet, obesity, and alcohol consumption.

In the Medical Qigong clinic there are three main factors that can contribute to the origination of diabetes: improper diet, emotional factors, and congenital predisposition.

1. The first factor is due to improper diet, including the excessive intake of alcohol, acrid, sour, pungent and greasy foods. The excessive intake of these types of foods can cause stagnation at the Middle Burner (which houses the Stomach and Spleen), and failure of the Stomach and Spleen to transport the accumulated Heat inside the body. This in turn can give rise to failure of the body's Yin fluids to nourish the Lungs and Kidneys, resulting in diabetes.
2. The second factor is due to a state of constant emotional upheaval. Emotional upsets can lead to stagnation of Qi which in turn impairs the body's Fluids, causing a Yin Deficiency with a hyperactivity of Evil Fire, resulting in diabetes.
3. The third factor is due to a genetic predisposition that can eventually surface if the patient is negligent with his or her diet and emotional health.

The Qigong doctor focuses on strengthening the patient's Spleen and Stomach, regulating the Liver Qi, and tonifying the Lung, Stomach, and Kidney Yin (which has been consumed by the Toxic Heat). It also is helpful to monitor the patient's eating habits to correct the disturbed metabolic rate and promote the regeneration and secretion by the patient's pancreatic islet cells.

SYMPTOMS

1. Diabetes involving the Upper Burner is characterized by excessive thirst, dry throat, dry tongue, and frequent urination.
2. Diabetes involving the Middle Burner is characterized by consuming abnormally large quantities of food during meals, emaciation, or constipation.
3. Diabetes involving the Lower Burner is characterized by frequent sweet smelling urination, or chyluria (passing of fat globules in the urine), and weakness and softness in the waist and knees.

TREATMENT

1. Begin by stimulating the Bl-23, Bl-20, Bl-13, St-36, and Sp-6 points on both sides of the patient's body.
2. Then purge and dredge the patient's CV-12, CV-4, and St-21 points.
3. Next, use the Extended Fan Palm hand manipulation with the Vibrating Palm technique to emit Qi into the patient's Bl-23, Bl-20, Bl-13, CV-12, and CV-4 for six to twelve breaths.
4. Continue emitting Qi into those points and into the Middle Dantian to regulate the Qi

activities of the Upper, Middle, and Lower Burners.

5. End the treatment by softly holding on to the patient's LI-11, LI-4, and Ht-1 points on both sides of the body, and with a gentle rocking or shaking motion allow the Qi to flow into the patient's body, settling into the patient's Lower Dantian.

PRESCRIPTIONS AND HOMEWORK

The "Inner Health" cultivation meditation is given to patients with diabetes, and is prescribed as follows.

1. Have the patient begin in a sitting posture, eyes closed, body relaxed using natural abdominal breathing.
2. While inhaling, the patient places the tongue against the hard palate, and begins to mentally repeat the mantra "I am quietly sitting."
3. While holding the breath, the patient mentally continues by saying the words "I am healthy."
4. The patient then exhales, releasing the tongue from the palate, while thinking the last phrase.

MEDICAL QIGONG THERAPY AND STOMACH DISEASES

This second section focuses on specific Stomach diseases and how to rectify them (Figure 45.6). The following discussion of treatment begins with an explanation of Stomach syndromes and symptoms. Several Stomach diseases have symptoms which are quite common to the organ's energetic dysfunction, such as epigastric pain.

STOMACH COMPLICATIONS AND SYMPTOMS

A list of Stomach syndromes and their symptoms are described as follows.

1. Stomach Heat can cause headache, epigastric pain, and bleeding.
2. Stomach Damp Heat can cause sinusitis and epigastric pain.
3. Stomach Phlegm Fire can cause epigastric pain and mental and emotional problems.
4. Stomach Deficiency can cause epigastric pain, diarrhea, atrophy syndrome, and bleeding.

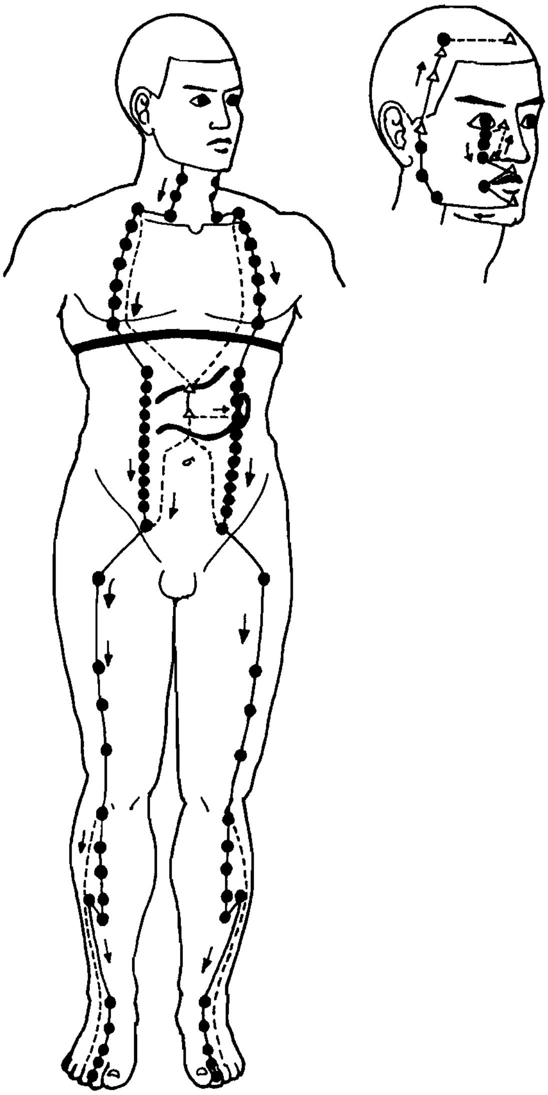

Figure 45.6. The Stomach (St) Channels

5. Stomach Yin Deficiency can cause tiredness, epigastric pain, and bleeding.
6. Cold invading the Stomach can result in epigastric pain.
7. Stagnant Blood in the Stomach can result in epigastric pain.

GENERAL TREATMENT FOR STOMACH DISEASES

The following is a description of a general protocol used for treating the patient's Stomach. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Stomach therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Stomach through either direct energetic insertion into the Stomach's anatomical location, or through the patient's channels and external energetic fields.
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Stomach, Spleen, and Yellow Court areas, and then roots the patient's Middle Burner Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Stomach Channels and out of the patient's St-36 points (Figure 45.7).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

SPECIFIC MEDICAL QIGONG THERAPY FOR STOMACH DISEASES

The following are several Medical Qigong treatments used in China to treat specific Stomach diseases.

ABDOMINAL PAIN

Abdominal pain is an abnormal condition of the abdomen in which there is a sudden, abrupt onset of severe pain in any part of the abdomen below or just around the umbilicus.

Diseases such as acute pancreatitis, gas-

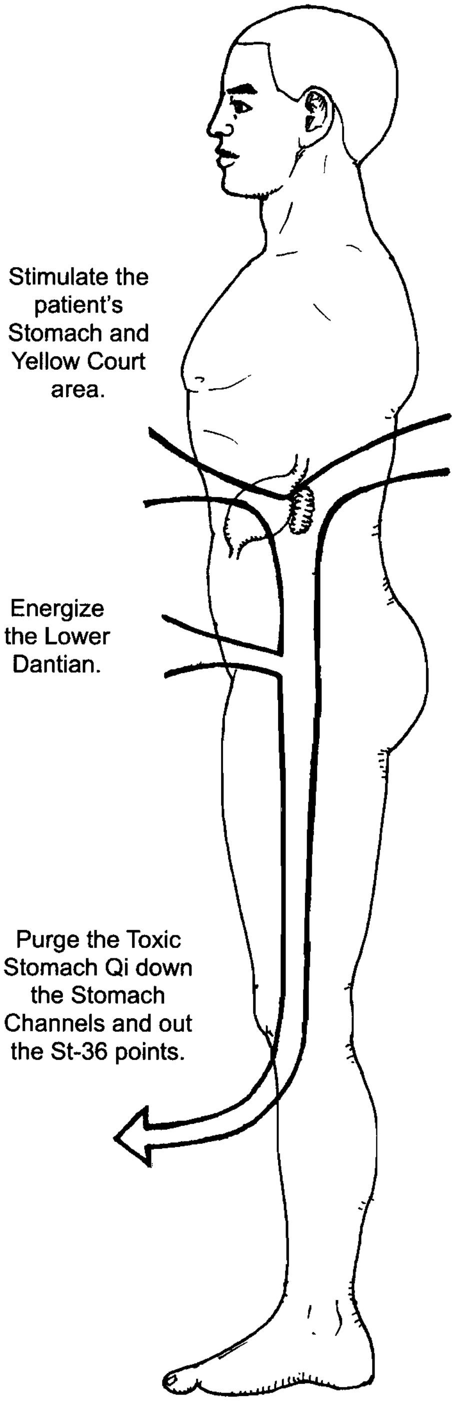
Stimulate the patient's Stomach and Yellow Court area.
Energize the Lower Dantian.
Purge the Toxic Stomach Qi down the Stomach Channels and out the St-36 points.

Figure 45.7. Treating Stomach Conditions

trointestinal spasm, stress induced abdominal pain, dyspeptic abdominal pain, and peritonitis, etc., in which abdominal pain symptoms appear, should be differentiated. Each differential diagnosis can be treated according to the exercises taught in the prescription section.

ETIOLOGY

The treatment for abdominal pain that is primarily intestinal in origin is different then that of gynecological abdominal pain, which is related to the menstrual cycle (see Chapter 51). There are four main factors which can contribute to the intestinal type of abdominal pain: abdominal pain due to the internal accumulation of Cold in the Lower Burner, abdominal pain due to dysfunction of Spleen Yang, abdominal pain due to food stagnation, and abdominal pain due to emotional stress.

1. Internal Cold in the Lower Burner may accumulate from eating too many cold foods and drinks, or can be due to the invasion of pathogenic Cold caused by exposure to the elements. Women in particular are more prone to Cold invading the abdomen during or immediately after menstruation. When Cold contacts the tissues, it slows down the movement of Qi and Blood, thus causing pain. External Cold may invade the patient's intestines directly without causing exterior symptoms, resulting in a sudden onset of pain and diarrhea.
2. Patients who are usually Deficient in Yang Qi often have a weak digestive system and a dysfunction of the Spleen Yang Qi. In this condition, pain often follows a pathogenic invasion, no matter how mild it may be. The pathogenic invasion may be due to overeating, overexertion, or brought on by eating after fasting and dieting, or simply by an Exterior Cold invasion.
3. Food stagnation is related to food intake or constipation, and is induced by Qi stagnation. The Stomach fails in digesting the food and the Spleen fails in separating and transporting the Clear and Turbid Qi.
4. Emotional stress is related to the circulation of Qi and Blood in the abdomen, and thus affects both the Liver and Spleen. Anger, frustration, and resentment can cause Liver Qi stagnation in the intestines, resulting in abdominal pain. Pensiveness and worry affect the Spleen and Lungs, impairing the transforming ability of the Spleen Qi and the downward descending ability of the Lung Qi, and likewise result in abdominal pain.

SYMPTOMS

1. Abdominal pain due to the internal accumulation of Cold, occurs abruptly, and may be relieved by defecation.
2. Abdominal pain due to Spleen Yang dysfunction, is characterized by continuous recurring pain, which may be relieved by pressure. Symptoms include loose stool, mental fatigue, and aversion to Cold.
3. Abdominal pain due to food stagnation, is characterized by gastric and abdominal distention, severe tenderness in the abdominal area, aversion to food and belching with acid regurgitation.
4. Abdominal pain due to emotional stress, is characterized by crying and vomiting.

TREATMENT

1. Have the patient lay on his or her side and begin stimulating the Bl-21, Bl-20, CV-12, St-36, and St-25 points on both sides of the body.
2. Purge the patient's Bladder Channels from the mid-back and waist area down to the feet three to five times.
3. Using the Vibrating Palm, emit Qi into the patient's CV-12, St-25, Bl-21, and Bl-20 points on both sides of the body. Purge and guide the Toxic Qi through the abdomen and down the lower extremities along the Stomach and Bladder Channels and out the patient's feet.
4. Using the Extended Fan Palm technique, extend Qi into the patient's CV-12 and both St-25 points for 14 breaths. Using Pulling and Leading manipulations, guide the patient's Qi down along the Stomach Channels and out both of the patient's feet.
5. Finally, Circle Rotate the patient's abdomen for 81 breaths in a clockwise direction.

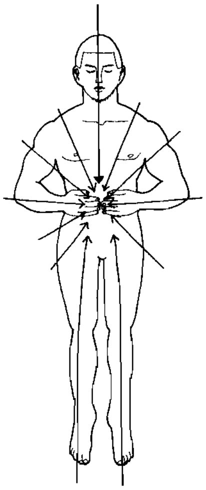

Figure 45.8. To tonify Deficient Stomach Qi, the patient must imagine drawing Qi into the body through the Baihui point at the top of the head, and the Yongquan points at the bottom of the feet. All three rivers of energy must meet to form a pool of energy within the Stomach area, which the body will absorb and store.

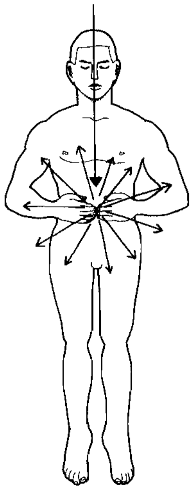

Figure 45.9. To treat Excess Stomach Qi, begin from a supine posture. While inhaling, imagine the Qi flowing down the Stomach Channels from the Baihui point ending just below the sternum at the Stomach area. Exhale and imagine the Turbid Qi dissolving and melting into the Lower Dantian, then disperse it from the body, out the pores.

TREATMENT MODIFICATIONS

1. In cases of abdominal pain due to the internal accumulation of Cold, add the method of extending Hot Qi.
2. In cases of pain due to Spleen Yang dysfunction, extend Hot Qi into the patient's Bl-21, Bl-20, and Mingmen area.
3. In cases of pain due to food stagnation, emit Qi using the Extended Fan Palm with clockwise then counterclockwise Spiraling methods into the patient's CV-12, CV-4, and St-25 points.
4. In cases of pain due to emotional stress, emit Qi using the Extended Fan Palm hand technique into the patient's Ht-7 and Pc-6 points on both sides of the body. Purge and Circle Massage the Shendao (GV-11) area in a counterclockwise direction.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient place his or her right palm on the CV-12 area and begin to Circle Massage 36 times in a clockwise and then 36 times in a counterclockwise direction.
2. After completing the circle rotations, the patient pauses for a moment, and then leads the Qi down to the navel and continues the 36 rotations, in both clockwise and counterclockwise directions.
3. Next, the patient inhales and presses the fingers of both hands gently below the xiphoid process. The patient exhales and directs the Qi down along the midline of the body to the pubic symphysis area for 36 times.
4. Finally the patient overlaps his or her hands (right over the left in men and opposite in women) and places them on the midpoint of the lower abdomen to knead the tissue area clockwise and then counterclockwise 36 times.

PRESCRIPTIONS MODIFICATIONS

- In cases of Stomach pain due to Deficient Stomach Qi, have the patient lie supine. The patient must imagine drawing Qi into his or her body through the Baihui point at the top of the head, and the Yongquan points at the bottom of the feet. All three rivers of energy must meet to form a pool of energy within the Stomach area, which the body will absorb and store into the Stomach's spacial cavity (Figure 45.8).

- In cases of Stomach pain due to Excess Stomach Qi and to treat Excess Stomach Qi symptoms, have the patient lie supine. As the patient inhales, he or she is to imagine the Qi flowing down the Stomach Channels (from the Baihui point) ending just below the sternum at the stomach area. As the patient exhales, he or she imagines the Turbid Qi trapped within the Stomach dissolving and melting into the Lower Dantian. Next, the patient disperses the Toxic Qi from the body out through the pores (Figure 45.9).

CHRONIC GASTRITIS

In Traditional Chinese Medicine, this disease belongs to the category of epigastralgia. Gastritis is described as an inflammation of the Stomach. Inflammation of the gastric mucosa is the main clinical manifestation of chronic gastritis, and results in chronic epigastralgia and dyspepsia (air hunger, resulting in difficult breathing, shortness of breath, sometimes accompanied by pain).

ETIOLOGY

Generally, emotional factors such as worry, anxiety, or anger may cause stagnation of Qi. Stagnant Qi, by impeding the function and flow of the body's energy, causes Qi deviations and obstructions that result in a stomach ache.

Gastritis can be superficial (affecting only the superficial layers of the Stomach) or atrophic (causing thinning of the Stomach mucosa). Its main manifestations are epigastric pain, which occurs about one hour after eating, followed by nausea, belching, and vomiting with sour regurgitation.

SYNDROMES

1. Liver and Stomach Qi stagnation, manifests as a distending pain, fullness, and discomfort in the Stomach, which is aggravated after a meal. The pain is not fixed to any specific Stomach location and worsens under emotional stress, causing belching with acid regurgitation and nausea.
2. Stomach Yin Deficiency due to Stomach Heat manifests as a stomach ache accompanied by a scorching sensation. The pain is not fixed to any specific Stomach location and is aggravated when the Stomach is empty, and is alleviated after a meal. The symptoms are expressed as irritability, dysphoria, and excess appetite.
3. Weakness of the Spleen and Stomach manifests as a dull stomachache which is relieved by Heat and pressure. Symptoms are expressed as abdominal distention, dim complexion, lassitude, cold limbs, and anorexia. If the patient has congenital insufficiency of Yang in the Middle Burner (the functional area of the Spleen and Stomach), Cold may originate from the interior of the body.

TREATMENT

1. Have the patient lie supine and relax his or her whole body, focusing especially on the Stomach area. The patient imagines the Toxic Qi dispersing upon exhalation.
2. Begin to purge and dredge the patient's CV-15 point (just below the xiphoid process), leading the Qi down to the Lower Dantian.
3. Next, use a Vibrating Palm hand technique to emit Qi into the patient's CV-12 point and both St-21 points for 14 to 28 breaths.
4. Applying the Extended Fan Palm hand technique to the patient's CV-12 and Middle Dantian, begin drawing Qi downward along the Conception Vessel and Stomach Channels to the bottom of the patient's feet.
5. Finally, have the patient lie prone and stimulate the Bl-20 and Bl-21 points on the lumbar-dorsal region, directing the energy from the upper back down to the Mingmen area.

TREATMENT MODIFICATIONS

1. For Stagnant Liver and Stomach Qi, use the Extended Fan Palm hand technique to emit Qi into the patient's Bl-18 points, pulling and conducting Qi along the Gall Bladder and Liver Channels down the legs and out the feet on both sides.
2. For Stomach Yin Deficiency due to Stomach Heat, use the Vibrating Palm hand technique

to emit Qi into the patient's Bl-20 and Bl-23 points, leading the Qi down the patient's Bladder Channels to the feet.

3. For Spleen and Stomach Deficiency, use the Extended Fan Palm hand technique to emit Qi into the patient's Lower Dantian area.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Descending the Yang and Ascending the Yin technique (see Chapter 41).

PRESCRIPTION MODIFICATIONS

1. For patients with Liver and Stomach Qi stagnation, have the patient practice the Spleen and Stomach Massage (see Chapter 39). Then perform the healing sounds "Hu" to regulate the Spleen and "Shu" to regulate the Liver.
2. For treating patients with Stomach Yin Deficiency due to Stomach Heat, have the patient first dredge the Spleen and Stomach and then perform the "Chui" (pronounced Chree) healing sound to regulate the Kidneys. Next, have the patient practice the tonification method of Taking in the Yellow Qi (see Chapter 4).
3. For treating patients Deficient Spleen and Stomach Qi, have the patient first practice the Daoist Five Organ Regulation exercise (see Chapter 42). Next, have the patient practice the tonification method of Taking in the Yellow Qi to strengthen the Spleen and Taking in the Red Qi to strengthen and regulate the Heart.

GASTRIC AND DUODENAL ULCERS

An ulcer is an open sore or lesion of the skin or a mucus membrane, accompanied by sloughing of inflamed, dead, or dying tissue, that causes the Stomach's lining to become more vulnerable to acids. If the sore becomes infected puss is discharged.

ETIOLOGY

A gastric or duodenal ulcer refers to a chronic ulcer seen only in the regions where the gastrointestinal tract contacts the patient's gastric juice. Its formation and development is closely related to the digestive functions of the acidic gastric juices and pepsin. Clinically this type of ulcer is characterized by chronic recurring epigastralgia, lasting for a few weeks, with pain occurring from 30 minutes to two hours after eating.

Currently, Stomach ulcers are also being treated through either herbs or antibiotics (from Western medicine). They are caused by bacterial infections resulting from either Qi and Blood stagnation or a Deficient Cold syndrome of the Spleen and Stomach.

SYNDROMES

- Qi and Blood stagnation results in a distended pain and fullness in the Stomach, tenderness, belching, and acid regurgitation, which becomes aggravated by emotional stress.
- A Deficient Cold syndrome of the Spleen and Stomach results in a dull pain within the Stomach which is relieved by heat and compression and aggravated by cold. There can also be vomiting with watery regurgitation, shallow complexion, cold extremities, lassitude, and weakness.

TREATMENT

1. Have the patient sit at the end of the treatment table, and begin extending energy into the patient's mid-back along their Bl-18, Bl-20, and Bl-21 points using the Extended Fan Palm technique.
2. Next, use the Vibrating Palm to extend energy into the patient's Yellow Court area. When the area has been sufficiently stimulated, purge the Toxic Qi out of the patient's body by leading the Qi down to the Lower Dantian for 14 to 28 breaths.
3. Finally, use the Extended Fan Palm hand technique to emit Qi into the patient's CV-12 and St-34 points for 14 breaths. Then lead the Qi down the Conception Vessel into the Lower Dantian to root and stabilize the patient's energy.

TREATMENT MODIFICATIONS

- When treating patients with Qi and Blood stagnation, use the Extended Fan Palm hand method to emit Qi into the CV-12, CV-13, and St-21 points. Next, direct the Qi along the patient's Conception Vessel and Stomach Channels, leading the Toxic Qi down the patient's legs and out of the body via the St-36 points.

- When treating patients with a Deficient Cold syndrome of the Spleen and Stomach, use the Extended Fan Palm hand method to emit Qi into the patient's Mingmen and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

In treating patients with ulcers, have them practice the Old Man Searching for the Reflection of the Moon at the Bottom of the Tide Pool exercise (see Chapter 53).

PRESCRIPTION MODIFICATIONS

- For patients with Qi and Blood stagnation, have them perform the Spleen and Stomach Massage (see Chapter 39) while sounding "Hu." Next, have the patient dredge the Spleen and Stomach using the sound "Shu" to purge the Liver and regulate Liver Qi.
- For patients with a Deficient Cold syndrome of the Spleen and Stomach, have them practice the method of Taking in the Yellow Qi from the natural environment.

CHRONIC CONSTIPATION

Chronic constipation refers to the prolonged retention of feces in the intestinal tract and difficulty in discharging the fecal matter. This condition is not only painful to the patient, but also increases toxicity in the body. It is clinically characterized by decreased times of fecal discharge or small amounts of feces which are considered either too dry and hard, or too difficult to be discharged.

Generally, the transitional time for food to pass through the digestive tract for humans takes 24 to 40 hours. The postponement of defecation over two days is considered constipation, although some normal patients may have regular defecation every two to three days.

ETIOLOGY

The common causes of constipation are: Diet, Febrile Diseases, Emotional Stress, and A Sedentary Life-style.

1. Diet plays a major role in causing constipation. Constipation can result from the excessive consumption of hot, spicy, or greasy foods. This excess consumption of hot and greasy foods can dry up the fluids of the Stomach and intestines, thereby drying up the stool so that it cannot properly move through the intestinal walls.

   An excess consumption of Cold foods can block the Spleen's function of transporting and thus prevent the stool from moving downwards.

   Finally, a lack of fiber in the food, lack of sufficient fluid intake or excessive alcohol consumption can cause constipation.
2. Febrile diseases can be caused by External Wind Heat invasion, which if not rapidly expelled from the body, can turn into an Interior Heat condition. The Interior Heat usually attacks the Lungs, Stomach, and intestines causing the Body Fluids to dry up. A Deficiency of Body Fluids from the retention of Heat within the intestines (even after the febrile disease has passed) can cause dry stool and constipation.
3. Liver Qi stagnation (due to mental depression, anger, resentment, or frustration) can cause a Qi obstruction in the Lower Burners that affects bowel movements and leads to constipation, abdominal distention, and pain. Excessive studying, worrying, and thinking affects the Spleen, slowing down its transportation of food in the intestines, this in turn can lead to constipation.
4. A sedentary life-style, whether from choice or due to a weak constitution from a chronic or severe illness (or surgery), can cause constipation. Patients in convalescent homes often suffer from constipation due to a lack of exercise and poor diet, which can lead to Spleen Qi and Blood Deficiency. Lack of exercise weakens the Spleen Qi and can cause Liver Qi stagnation. When Spleen Qi becomes Deficient for long periods of time, the Spleen energy fails to provide the Qi to move the stool, resulting in constipation.

SYNDROMES

- Constipation due to an Excess condition is caused by decreased times of fecal discharge (once every three to seven days), and is usually accompanied with fever, thirst, poor appetite, and emotional anxiety.

- Constipation due to a Deficient condition is mostly seen in patients who have just recovered from an illness, or have just given birth and have become Deficient in Qi and Blood. It may also be observed in patients who suffer from a congenital insufficiency of Yang Qi in their Lower Burner and an accumulation of Yin Cold, which makes it difficult for the intestines to perform their normal transporting function. Patients with a Deficient condition may have bowel movements every three to seven days. The stool may be very dry, hard, and difficult to pass. This condition is clinically manifested by a pale and dull complexion, listlessness, and cold pain in the abdomen which is relieved by heat. The patient may also experience emotional fright due to a Kidney Qi Deficiency.

TREATMENT

1. Have the patient sit at the end of the treatment table, and begin to stimulate the patient's mid-back between the Bl-23 and Bl-25 points. Then stimulate the patient's lower lumbar and coccygeal areas, dredging and purging the Toxic Qi down the patient's thighs and the feet, while stimulating the Bl-31 through Bl-34 points.
2. The doctor will then have the patient lie supine and begin to stimulate the patient's Yellow Court and Lower Dantian areas, using the Vibrating Palm hand technique for 14 breaths.
3. Next, apply the Kneading Tiger Palm with the Vibrating and Quivering hand manipulations to emit Qi into the back of the patient's head at the Bl-10 points for 12 breaths, and then onto the patient's Lower Dantian for 14 breaths.
4. After energizing the patient's Lower Dantian, emit Qi, using the Extended Fan Palm technique, into the patient's middle epigastrium and navel area, as well as into the Lower Dantian. Finally, use the Spiraling and Quivering hand manipulation in a clockwise direction to gather the stagnant Qi and to lead it down the patient's Stomach Channels and out the legs.

TREATMENT MODIFICATIONS

- When treating patients with an Excess type of constipation, use the Extended Fan Hand Palm technique to emit Cold Qi into the patient's lower abdomen. Rotate the hand in a counterclockwise direction, while purging the Toxic Qi from the patient's body. Repeat for 14 breaths.
- When treating patients with a Deficient type of constipation, use the Vibrating Palm technique to emit Qi into the patient's Lower Dantian and Mingmen area for 8 to 16 breaths.

PRESCRIPTIONS AND HOMEWORK

To treat chronic constipation, have the patient practice the Daoist Five Yang Organ Regulating exercise (see Chapter 42) for 250 breaths twice a day.

PRESCRIPTION MODIFICATIONS

- For patients with an Excess type of constipation, place them in the Wuji posture with their hands resting on their lower abdomen. Begin with a clockwise rotation, imagining the Toxic Qi draining from the patient's Spleen and Stomach areas, down into their Stomach Channels and out their feet.
- For patients with a Deficient type of constipation, have them practice rubbing their Lower Dantian to strengthen the abdominal Qi, and have them practice the Gathering the Sun's Essence meditation (see Chapter 11).

CHAPTER 46

QI EMISSION THERAPY AND LUNG DISEASES

INTRODUCTION

The Lungs are solid organs that control the cycles of Qi circulation in the body, and are sometimes referred to as the White Emperor. They correspond to the element Metal. The Lungs house the Po (Corporeal Souls) and are in charge of respiration and regulating the metabolism of the Body's Fluids (Figure 46.1). The Lungs are also responsible for keeping Qi pure and extending it downward. The Lungs are energetically paired with the Large Intestine (a hollow organ).

The Lungs' upper external connection with the outside world is through the nose, hence any fragrance or odor inhaled through the nose affects the senses and evokes emotional responses from the Po.

In the fall months, the Lung Qi becomes more active in individuals that already possess strong Lung Qi, but becomes Deficient in those who already have weak Lung Qi. The excessive eating and drinking of pungent foods and the overexposure to dryness drains the Lungs.

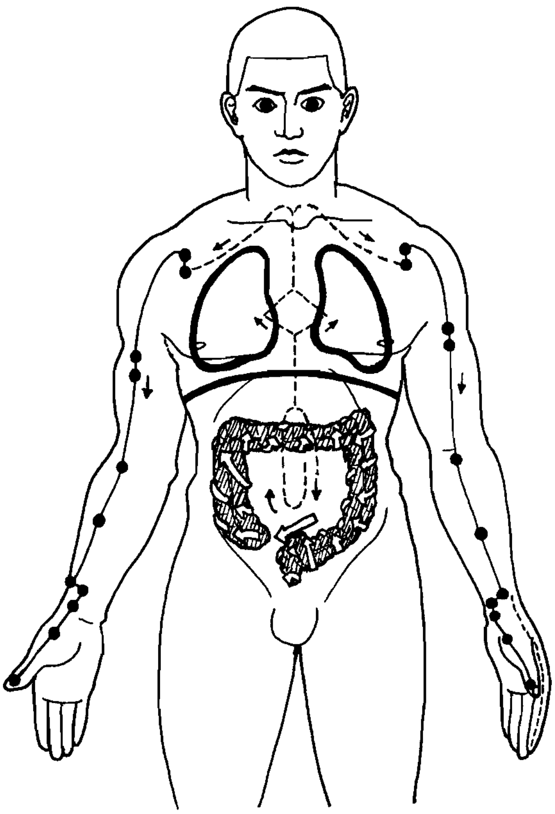

Figure 46.1. The Lung (Lu) Channels

LUNG COMPLICATIONS AND SYMPTOMS

This section focuses on specific Lung diseases and how to rectify them. The following discussion of treatment begins with an explanation of Lung syndromes and symptoms. Several Lung diseases have symptoms which are quite common to the organ's energetic dysfunction, such as headaches, cough, the common cold and influenza. A list of Lung syndromes and symptoms are described as follows.

1. Wind Cold invading the Lungs usually enters the patient through the pores of the skin and body hair (which are ruled by the Lungs). If the Lungs' energetic field is invaded by Wind Cold, it may cause the Lung Qi to become impure, ascend, and cause obstructions in the pores, resulting in a Wind Cold syndrome. This can cause the following symptoms: severe aversion to cold, headache, cough, the common cold or influenza, breathlessness, asthma, and allergic rhinitis.
2. Wind Heat Invading the Lungs usually enters the body by way of the mouth and nose; the nose being the orifice to the Lungs. If the Lungs' energetic field is invaded by Wind Heat, it may cause the Lung Qi to become impure, ascend, and impede the ability of the

pores to expel pathogenic Evils, resulting in a Wind Heat syndrome. This can cause the following symptoms: slight aversion to cold, headache, cough, the common cold or influenza, asthma, allergic rhinitis and sinusitis.

3. Wind Dampness is a type of Wind Cold syndrome. It consists of both Wind and Dampness together, invading either the skin (causing itching and rashes) or the channels and joints (causing painful obstruction syndrome). If the Lungs' energetic field is invaded by Wind Dampness, it causes the following symptoms: skin rashes, swollen joints, body aches, headaches, the common cold and influenza.
4. Wind Dryness invading the Lungs causes the following symptoms: dry, itchy sore throat, dry lips, dry mouth, the common cold and influenza, cough and bleeding.
5. Lung Heat causes the following symptoms: restlessness and a feeling of heat, common cold and influenza, cough, sinusitis, breathlessness and atrophy syndrome as the body's functions begin to shut down.
6. Lung Phlegm Heat causes the following symptoms: common cold and influenza, cough with expectoration, breathlessness and a sensation of oppression in the chest.
7. Cold-Damp Phlegm invading the Lungs causes the following symptoms: a feeling of heaviness, cough with profuse expectoration, wheezing, nausea, poor appetite and breathlessness.
8. Lung Qi Deficiency causes the following symptoms: breathlessness, wheezing, slight cough with no phlegm, asthma, allergic rhinitis, tiredness, spontaneous sweating, and enuresis or incontinence.
9. Lung Yin Deficiency causes the following symptoms: night sweating, extreme tiredness, breathlessness, asthma, dry cough in short bursts, and Five Palms Hot (a condition also known as "Five Center Heat," in which the patient feels a hot sensation in the palms and soles of the feet, accompanied by heat and agitation of the chest and/or head area).

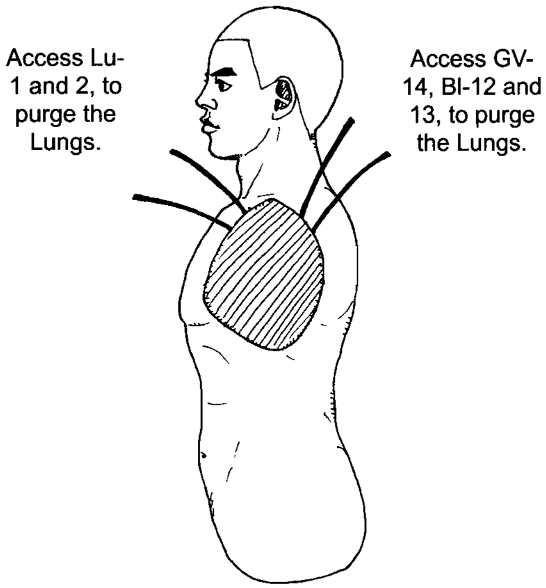
Access Lu-1 and 2, to purge the Lungs.
Access GV-14, Bl-12 and 13, to purge the Lungs.

Figure 46.2. Treating Lung Diseases

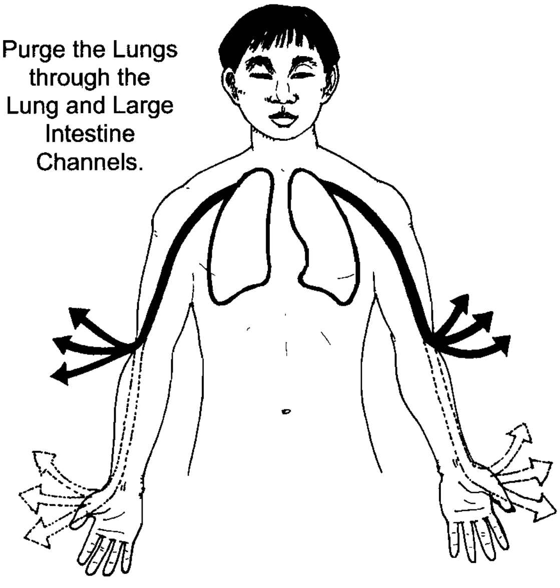
Purge the Lungs through the Lung and Large Intestine Channels.

Figure 46.3. Drain the Turbid Qi out of the Lungs, down the Lung Channels and out of the hands.

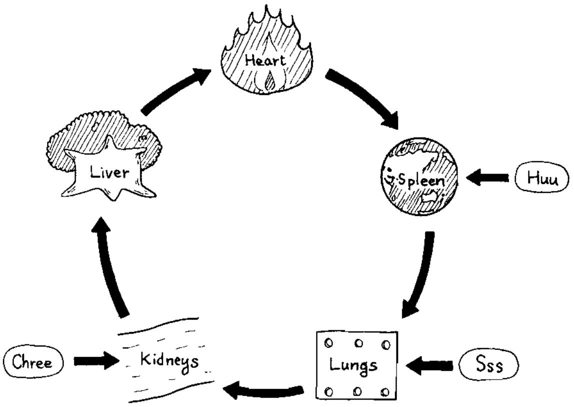

Figure 46.4. Replenish the mother (Spleen) to nourish the child (Lungs).

GENERAL TREATMENT FOR LUNG DISEASES

The following is a description of a general protocol used for treating a patient's Lungs. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body towards which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

After comparing the various Lung therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Lungs through either direct energized insertion into the Lungs' anatomical location, or through the patient's channels and external energetic fields (Figure 46.2).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor stimulates the patient's Lungs and Upper Burner areas and roots the patient's Qi into the Lower Dantian.
3. Next, the Qigong doctor leads any remaining Turbid Qi down the Lung Channels and out the patient's hands (Figure 46.3).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

MEDICAL QIGONG THERAPY FOR LUNG QI DYSFUNCTION

Patients with Lung disease may have a pale complexion, dry hair, dyspnea (difficult breathing), hyposmia (a deficient sense of smell), stuffy nose, dryness and itching of the skin, or discomfort and pain in the chest and back. Patients may dream of the sun, moon, clouds, or of having sex. The primary therapies used to treat Lung diseases include the Breath regulation and Qi reinforcing meditations (see Chapter 16).

- Patients with an Excess Heat syndrome of the Lungs may have a flushed right cheek. To treat this type of syndrome, have the patient utter the sounds "Si" (Sss) for the Lungs and "Chui" (Chree) for the Kidneys. Guide the Evil Qi out of the patient's body along the Lung and Large Intestine Channels using the Pushing, Pulling and Guiding Qi manipulations.
- Patients with a Deficient syndrome of the Lungs may have shortness of breath and be unable to regulate their breathing. To treat this type of syndrome, have the patients inhale the sound "Hu," as they visualize absorbing the color yellow to replenish the Spleen (the mother); this will nourish the Lungs (the child). Next, the patient guides the Qi back to the Lungs (Figure 46.4).

LUNG QIGONG PRESCRIPTIONS

These exercises regulate the Qi in the Lung Channels, tonify the Lungs, replenish Qi, ventilate the Lungs, and relieve asthma, coughing, and sputum. These exercises can be used to prevent and treat the following diseases and syndromes: bronchitis, pulmonary emphysema, and bronchial asthma, as well as dyspnea and abundant expectoration.

PRESCRIPTIONS FOR TONIFYING THE LUNGS

The Taking in the White Qi meditation can be used to tonify the Lungs. The properties inherent within the color white have a persistent vibratory rate used for tonifying the Lungs and Large Intestine.

1. Assume a standing, sitting, or lying posture.

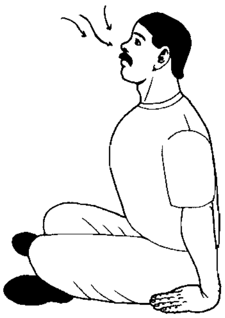

Figure 46.5. Inhale

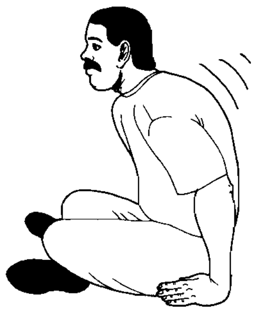

Figure 46.6. Exhale

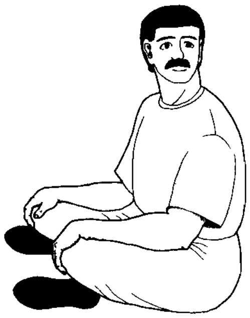

Figure 46.7. Inhale

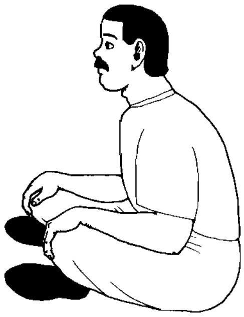

Figure 46.8. Exhale

Relax the whole body, breathe naturally, and get rid of any stray thoughts.

2. When inhaling, imagine a white energy in front of you. Inhale the white light energy in through the nose and down into the Lungs. As you exhale through the mouth, the dark, pathogenic Qi leaves the Lungs; however, the bright clean color remains in the Lungs, stimulating and vitalizing them. With each breath the Lungs retain more clean energy and begin to glow. Repeat for five breaths.
3. Next breathe the color white in through the nose, filling the entire mouth. When exhaling through the nose, send the Qi slowly into the Lungs, then down to the Lower Dantian, and then further out into the skin and hair of the whole body. Repeat nine to eighteen times. Return to the beginning posture and close.

PRESCRIPTIONS FOR PURGING THE LUNGS

Purge the Lungs by having the patient rub his or her chest while sounding "Sss."

The patient should assume a sitting or standing posture. Both palms are placed flat on the sides of the chest as the patient inhales slowly. When exhaling, the patient sounds the word "Sss" while rubbing the chest with both palms (circling up and down, from the center outward). This exercise is performed for six to twelve breaths.

PRESCRIPTIONS FOR REGULATING THE LUNGS AND CONDUCTING QI

The following exercise is used to regulate the Lungs after purging or tonifying.

1. Sit with the legs crossed, breathe in with the spine arched backwards and the hands by the sides pressing against the ground; extend the chest, and inhale until the chest is full (Figure 46.5).

   Pause, then push the upper back in a posterior direction and draw in the chest while exhaling through the nose (Figure 46.6). Repeat four to nine times.
2. Next, still sitting with the legs crossed, place the palms on the knees. Inhale with the spine erect and turn the head allowing the torso to turn slightly to the left (Figure 46.7). Then bend forward and exhale while facing the center (Figure 46.8). Next, inhale and turn to the right, then again bend forward and exhale while facing the center. Always inhale when turning to the sides, and exhale while facing the center.

SPECIFIC MEDICAL QIGONG THERAPY FOR LUNG DISEASES

The following are several Medical Qigong treatments used in China to treat specific Lung diseases.

TREATING THE COMMON COLD AND INFLUENZA

Upper respiratory tract infections belong to the category of common colds, while influenza belongs to the category of seasonal epidemic colds. Both conditions may be treated by referring to the prescriptions described within this section.

ETIOLOGY

The common cold is a generic term used to describe an Exterior condition caused from the invasion of pathogenic Wind (Wind Cold, Wind Heat, Wind Damp Heat or Wind Dry Heat). It is generally associated with coryza or inflammation of the respiratory mucous membranes.

Upper respiratory tract infection is caused by viruses or bacteria. Influenza is known as an acute, contagious respiratory infection, characterized by sudden onset. The chief clinical manifestations are stuffy nose, nasal discharge, coughing, sneezing, sore throat, fever, headache, chills, and an aversion to wind.

If the pathogenic Qi is on the Exterior level of the body, the internal organs are not affected, and only the patient's Wei Qi is involved. If the pathogenic factor becomes Interior, the patient's internal organs are affected, especially the Lungs and Spleen. If the pathogen is not cleared and purged from the patient's body, it may either penetrate more deeply into the patient's body and cause serious problems, or give rise to residual Heat which causes chronic post-viral fatigue syndromes.

SYNDROMES

- A Wind Cold syndrome has the following symptoms: chills, headache, aches and painful joints, coughs and sore itchy throat, stuffy nose, low fever, and no sweating with profuse, watery nasal discharge.
- A Wind Heat syndrome has the following symptoms: fever, cough, headache, slight aversion to cold, swollen and sore throat, thick yellow sputum, slight thirst, and perspiration. Once the pathogenic factor penetrates into the patient's interior, the body's Qi begins to fight against it, causing a fever and feeling of heat inside the body, with an aversion to cold, and shivering.

TREATMENT

1. Have the patient sit at the end of the treatment table.
2. Begin by stimulating the energy within the patient's head, purging and leading the Toxic Qi down the patient's arm and out the LI-11 and LI-4 points.
3. Then stimulate the patient's Yintang (Third Eye) and Taiyang (temple) points. Using Pulling and Leading Qi manipulation techniques, purge and guide any remaining Toxic Qi downward, along the Conception Vessel and the Stomach Channels, out the patient's feet, three to seven times.
4. Next, to balance the functional activities of the body's Qi, stimulate the tissue area on the patient's upper back and head (GV-14 and GV-16, Bl-12 and Bl-13) using the Extended Fan Palm technique, and guide the Qi downward along the Bladder Channels to the patient's feet.
5. Finally, softly grasp the patient's LI-4 and LI-11 points and gently shake the arm, stimulating the patient's Lung Qi. This technique is performed on each of the patient's arms.

TREATMENT MODIFICATIONS

1. When treating patients with Wind Cold syndrome, extend Hot Qi into the GV-14 and Bl-12 points, using the Extended Fan Palm technique, and guide the Qi downward along the Bladder Channels to the patient's feet and out the body.
2. When treating patients with Wind Heat syndrome, extend Cold Qi into the GV-14, Bl-13, and GB-20 points, using the Extended Fan Palm technique, and guide the Qi downward along the Stomach Channels to the patients' feet and out the body.
3. When treating patients with nasal obstruction, extend Vibrating Qi, using the Sword Fingers technique, into the Ll-20 points for 9 breaths.
4. When treating patients with cough, chest stuffiness, and Phlegm, extend Vibrating Qi, using the Sword Fingers technique, into the CV-17, Lu-1, and Lu-2 points for 8 breaths.

PRESCRIPTIONS AND HOMEWORK

The patients should be given the Ten Dragons Run Through the Forest exercise (see Chapter 41) to sweep and purge the Gall bBladder

Channels. Most of the attention should be placed on feeling the Qi beneath the hands, with emphasis placed on exhalation to expel the pathogenic Wind from the body. Perform this prescription for 24 to 50 times.

CHRONIC BRONCHITIS

Chronic bronchitis is a condition caused by the chronic inflammation of the mucous membranes of the bronchial tubes.

ETIOLOGY

Chronic inflammation of the bronchial mucosa and peribronchial tissues is mostly due to lingering and recurrent acute bronchitis, or may occur after an infection from other diseases. The chief clinical manifestations are chronic recurring cough, expectoration and dyspnea. Bronchitis can be divided into acute and chronic conditions.

1. Acute bronchitis corresponds to the Qi level of infection within the four level identification patterns. It is the easiest to treat, and normally manifests with Hot Phlegm symptoms.
2. Chronic bronchitis takes a little longer to heal and the doctor should focus treatment on attending to the root and manifestation of the disease simultaneously (purging Phlegm and dispersing the Heat or Cold).

SYNDROMES

- For Phlegm Dampness attacking the Lungs, the main symptoms are abdominal distention, cough with copious white sputum, stuffiness in the chest and loss of appetite.
- For Lung and Spleen Deficiency, the main symptoms are cough with sputum, shortness of breath, spontaneous sweating, loss of appetite and loose stool.

TREATMENT

1. Have the patient sit at the edge of the table. Begin to purge the patient's Lungs, dredging the Toxic Qi down the Lung Channels and out of the hands.
2. Use the Vibrating Palm technique to stimulate the patient's Middle Dantian, Lu-1, -2, Bl-13, GV-16, and St-40 points. Use 8 to 12 breaths for each point.

TREATMENT MODIFICATIONS

- For Phlegm Dampness attacking the Lungs, use the Extended Fan Palm technique to emit Qi into the patient's Lungs; purge the Spleen and Stomach Channels, leading the Toxic Qi down the channels and out of the St-40 and Sp-6 points on each leg.
- For Spleen and Lung Deficiency, emit Qi using the Vibrating Palm technique into the patient's Bl-20, Bl-23, Bl-13 and CV-12 points for 6 to 12 breaths.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Lung exercise from Daoist Five Yin Organ Tonification and Regulation exercises (see Chapter 42) to strengthen and balance the Lungs.

PRESCRIPTION MODIFICATIONS

- For patients suffering from Phlegm Dampness attacking the Lungs, have the patient perform the Sun and Moon Rotating technique (see Chapter 41), sounding "Sss" to regulate the Lungs and the sound "Hu" to regulate the Spleen.
- For patients suffering from Lung and Spleen Deficiency, have them perform the Taking in the White Qi to tonify the Lungs, and the Taking in the Yellow Qi to tonify the Spleen.

ASTHMA

Asthma is a chronic disorder of the respiration organs characterized by difficulty in breathing, wheezing, and tightness in the chest. Bronchial asthma, chronic bronchitis and pulmonary emphysema are all included in the category of this disease.

ETIOLOGY

Asthma is an Excess condition which is caused mainly by a retention of Wind, Cold, Heat or Damp Phlegm in the Lungs. When a patient is affected by invasion of pathogenic Dampness, disorders in Qi mobility develop. Phlegm mixes with Qi and obstructs the air passages, causing dyspnea (shortness of breath) and wheezing.

Asthma may also result from an obstruction of the Lungs due to Phlegm retention caused by a failure of the Spleen to transport, or from a Deficient condition of the Kidneys. The diagnosis of this disease can be divided into two types of asthma: early onset asthma and late onset asthma.

1. Early onset asthma (also called extrinsic or atopic asthma), usually starts during childhood and is divided into non-allergenic and allergenic asthma. Allergic asthma is associated with eczema and it tends to be difficult to treat because it stems from a congenital Lung and Kidney Qi Deficiency and is often hereditary. Early onset asthma can be affected by diet (e.g., lactose intolerance), emotional problems, fatigue, chronic illness and invasion by external pathogenic factors. It is related to an allergic hypersensitivity of the immune system.
2. Late onset asthma (also called intrinsic asthma) normally starts later in life and is due to bronchial hyperreactivity. In late onset asthma there is no hereditary basis. It is generally characterized by a Lung, Spleen, or Kidney Deficiency and a retention of Phlegm. In some cases, asthma can be caused by Liver Qi stagnation or Liver Fire obstructing the descending action of Lung Qi. Late onset asthma can be affected by overwork, excessive sexual activity, emotional stress and the excess consumption of sour, greasy or cold foods.

SYNDROMES

1. For Wind Cold asthma due to External Evils, the main symptoms are dyspnea, coughing, abundant, thin expectoration and clear Phlegm accompanied with external Wind Cold syndrome.
2. For asthma due to the accumulation of Phlegm Dampness in the Lungs, the main symptoms are dyspnea, cough, abundant, thin and sticky expectoration, stuffiness in the chest and in the epigastric region.
3. For asthma due to Deficient Lung and Spleen Qi, the main symptoms include dyspnea, spontaneous perspiration, aversion to Wind, an expectoration of clear, thin Phlegm, weak voice, loss of appetite and general lassitude.
4. For asthma due to Deficient Lung and Kidney Yin, the main symptoms are dyspnea, dry throat and mouth, dry cough, inconsistent fever, night sweating, and with some patients there may be hemoptysis (the expectoration of Blood).

TREATMENT

The therapeutic effect of Medical Qigong is better for patients who have had this disease for a short time period. In chronic asthma, improvement usually takes more time.

1. Begin by stimulating the patient's CV-22, CV-17 and CV-4 points along with the Bl-13 points.
2. Using the Vibrating Palm technique and emit Qi with the left palm into the patient's lower neck and upper back area, while simultaneously extending energy into the patient's Lower Dantian with the right hand for 14 to 28 breaths.
3. Purge the Toxic Qi from the patient's CV-22 and Bl-13 areas, then lead and guide the patient's Lung Qi down the Governing Vessel and Bladder Channels into the Mingmen GV-4 and Bl-23 points. Repeat for 3 to 7 respirations.
4. Using the Sword Fingers hand posture, vibrate energy into the patient's CV-22, -17, Lu-1 and Lu-2 points. Then guide the patient's Qi to flow downward from the chest to the Lower Dantian along the Stomach Channels and Conception Vessel to stabilize the patient's Qi.

TREATMENT MODIFICATIONS

1. To treat a case of Wind Cold syndrome due to exogenous Evils, purge and sweep the patient's Gall Bladder Channels, concentrating on the head area. Then extend Qi into the patient's Lu-1 and Lu-2 points. Use Pulling and Leading manipulations to draw the Toxic Qi along the patient's Lung Channels, expelling the pathogenic Qi out of the fingertips on both hands.
2. In treating cases of Deficient Lung and Spleen Qi, use the Extended Fan Palm hand posture and Vibrating Palm technique to emit Qi towards the patient's CV-12, CV-6, Bl-20 and Bl-13 points for 9 to 18 breaths.

3. In treating cases of Deficient Lung and Kidney Yin, use Extended Fan Palm hand posture and Vibrating Palm hand technique to emit Qi towards the patient's Lower Dantian, Mingmen, and Bl-13 points for 6 to 12 breaths.

PRESCRIPTIONS AND HOMEWORK

Patients are encouraged to practice physical exercises to strengthen the body and enhance their resistance to pathogenic invasion. These exercise prescriptions however, must not be too strenuous. Two prescriptions used for relieving asthma and respiratory pain are described as follows.

Exercise to Relieve Asthmatic Pain. For chronic conditions of asthma, patients are given the following prescription:

1. From a sitting or standing posture, slightly close the eyes and relax. Stimulate the energy of the chest by kneading the sternum from the sternal notch to the xiphoid process 36 times.
2. Inhale and place the right palm on the left side of the chest. Exhale, and begin dredging the left Lung Channel using a sweeping motion to purge the Toxic Qi down the arm and out through the hand. Perform this Purging exercise 10 times, then repeat 10 times on the opposite side.
3. Finally, perform the Kidney Rub exercise (see Chapter 39) for 18 breaths.

Exercise to Relieve Asthma Attacks

For acute asthmatic attacks, patients are given the following prescription.

From a sitting or standing posture, slightly close the eyes and relax the muscles of the chest and back. Take a deep breath and say "relax" silently, while simultaneously imagining the Lungs relaxing and descending the body's Qi down to the toes along both lateral sides of the body. As the Qi descends from the Lungs, imagine this energy flowing like water, down your body into the Earth. Repeat this procedure until the asthma attack has receded.

PRESCRIPTION MODIFICATIONS

- Patients with Deficient Lung Qi should practice the previous exercises in combination with Taking in the White Qi.
- Patients with a domination of pathogenic factors should practice the previous exercises in combination with rubbing their chest and uttering the sound "Sss."

BRONCHIECTASIS

Bronchiectasis is the chronic dilation of the bronchi, with a secondary infection that usually involves the lower portion of the Lungs. Dilation may involve an isolated segment or can spread throughout the entire bronchi.

ETIOLOGY

This disease is characterized by hypertrophy of the mucus-secreting glands in the bronchial tree. The dilation and deformation of the bronchi are caused by the damage of the vessel walls due to chronic inflammation of the bronchi and the peribronchial Lung tissue. In advanced cases the bronchi themselves can become inflamed with infection. The clinical manifestation is a chronic cough with an excessive expectoration of sputum. If followed by infection, there may be night sweating, fever and lassitude.

SYNDROMES

1. Liver Fire attacking the Lungs can manifest with pain and distention in the chest and hypochrondium, flushed face, cough, expectoration of yellow, thick sputum and repeated hemoptysis (Blood discharging from the larynx, trachea, bronchi, or Lungs).
2. Dry Heat scorching the Lungs manifests as a dry throat and dry cough with little sputum, which is bright red and bloody.
3. Qi and Yin Deficiency manifests through shortness of breath, lassitude, emaciation, repeated hemoptysis, cough with little sputum, erratic fever and flushed face (especially around the cheeks).

TREATMENT

1. Have the patient sit at the edge of the table, and begin to stimulate the upper chest at the Lu-1 and Lu-2 points and the upper back around the patient's Bl-13 points.
2. Using the Extended Fan Palm hand technique, emit Qi into the patient's Middle Dantian directing the energy to flow into the Lu-1 and Bl-13 points for 6 to 12 breaths.

3. Using the Leading and Guiding hand techniques, purge and dredge the Toxic Qi out the Lung Channels (i.e., down the arms and out of the hands).

TREATMENT MODIFICATIONS

1. When treating patients with Liver Fire attacking the Lungs, include the Vibrating Palm technique, and emit Qi into the patient's Bl-13 and Bl-18 points for 12 breaths. Then purge the Liver Channels down the torso and out of the legs.
2. When treating patients with Dry Heat scorching the Lungs, apply the Vibrating Palm technique to emit Qi into the patient's Bl-23 and Bl-13 points for 24 breaths.
3. When treating patients with Qi and Yin Deficiency, apply the Vibrating Palm technique to emit Qi into the Bl-13, Bl-20, and Bl-23 points for 12 breaths.

PRESCRIPTIONS AND HOMEWORK

It is advisable for patients to practice regulating Lung exercises with Natural Breathing to strengthen the Lungs.

PRESCRIPTION MODIFICATIONS

1. For Liver Fire attacking the Lungs, have the patient perform the Sun and Moon Rotating technique (see Chapter 41), while sounding the word "Sss" for 36 breaths, to purge the Lungs in addition to performing the Lung tonification and regulation exercises (see Chapter 42). Next, have the patient rub the hypochrondrium area and sound the word "Shu" for 36 breaths.
2. For Dry Heat scorching the Lungs, (from a Wuji posture) have the patient bring both hands up to the sternum and begin kneading the sternum down to the xiphoid process 36 times. Next, the patient places both palms on the chest with the fingers touching the sternum to allow the Qi to flow into the Lungs for 10 breaths. Finally, the patient rubs the hypochrondrium and purges the Toxic Qi from the body while exhaling and rubbing the sides of the chest, starting at the armpits, to direct the Qi to flow downward into the Lower Dantian.
3. For Qi and Yin Deficiency, it is advisable for the patient to practice Lung tonifying exercises and the Gathering the Moon Cream exercises (see Chapter 11), as well as the methods of Taking in the White Qi.

SINUSITIS

Sinusitis is described as the inflammation of the sinuses. Infections stemming from the common cold or influenza frequently cause secondary infections in the sinuses, these infections tend to become chronic. The body's frontal and maxillary sinuses are very prone to infection and inflammation.

ETIOLOGY

An invasion of pathogenic Wind Cold into the Lungs through the nose (the orifice of the Lungs) can bring about the accumulation of Heat and cause an obstruction of the Lungs' Qi. This condition further impairs the Lungs' dispersing and descending function, giving rise to sinusitis.

The excessive consumption of greasy, hot foods leads to a buildup of Phlegm and Heat toxins within the body and may predispose the patient to sinusitis. Greasy, hot foods may also lead to the formation of Damp Heat within the Stomach and Spleen, which can be transported upwards into the sinuses by way of the Stomach Channels.

Excess conditions, such as Wind Heat, Lung Heat, Stomach and Spleen Damp Heat, and Liver and Gall Bladder Fire, should all be purged from the patient's body, whereas chronic conditions caused by Spleen Deficiency should be tonified.

SYNDROMES

Sinusitis is characterized by consistent yellowish and foul smelling nasal discharge accompanied by cough, impairment of smell, and dull pain in the upper part of the cranium. The headache caused from sinusitis is located on the face and corresponds with either the frontal ethmoidal or the maxillary sinuses.

TREATMENT

1. Begin by stimulating the patient's frontal sinuses, focusing specific attention on the LI-20, GB-20, and Yintang points.

2. Then emit Qi into both of the patient's hands and wrist areas, focusing specific attention on the LI-4 and Lu-7 points.
3. Next, using either the Sword Fingers or Extended Fan Palm hand posture, emit Qi into the patient's LI-20 points. Begin using the Pushing, Pulling, and Leading manipulations to guide the patient's Toxic Qi down the Stomach Channels, purging it out through the St-36 point on each leg.

PRESCRIPTIONS AND HOMEWORK

The following prescription is used for relieving sinusitis.

The patient, from a seated posture, rubs his or her palms until the dorsal sides of each thumb are very hot. Then the patient places the thumbs on the sides of the nose and begins to gently rub up and down for 10 breaths. This exercise is called Bathing the Nose.

Next, the patient places the tips of the middle fingers on each side of the nose (LI-20), and begins kneading these points. This is followed by using circular rotations from the cheeks to the nose five times during inhalation, and then from the nose to the cheeks five times during exhalation. This part of the exercise is called Kneading the Apex of the Nose. This exercise is performed for 10 breaths.

CHAPTER 47
QI EMISSION THERAPY FOR KIDNEY DISEASES

INTRODUCTION

The Kidneys are solid (Zhong) organs that store the body's Jing. They correspond to the element Water and are sometimes referred to as the Black Emperor. The Kidneys house the Zhi, which is the will power, and they are in charge of life and reproduction. The Kidneys are also responsible for bone, brain and Marrow formation. They govern the Mingmen Fire and are the root of the Yin and Yang energies within all the organs (Figure 47.1). The Kidneys are energetically paired with the Bladder (a hollow organ).

The Kidneys are responsible for inhalation. Upon inhalation the Kidneys grasp the Lungs' Qi and hold it. This "holding" action secures the body's ability to maintain deep, slow inhalation. Therefore, when the Kidneys become weak (due to shock, trauma or disease), one of the clinical manifestations of Deficient Kidneys includes dyspnea (shortness of breath and gasping for air).

The Mingmen, or Gate of Life, is the root of the Yuan Qi, the residence of Water and Fire energy. The Yin Qi of the Five Yin Organs cannot be nourished without it, and the Yang Qi of these organs cannot be lifted without it. Kidney Qi flows to the brain. The ears (Yang orifice), anus and urethra (two Yin orifices) are the orifices of the Kidneys.

The Kidneys' upper external connection with the outside world is through the ears, hence any sound or tone resonation affects the core essence of the individual's spiritual and emotional will power.

In the winter months the Kidneys become more active in individuals who already possess strong Kidney Qi, but becomes Deficient in those with weak Kidney Qi. The excessive eating and drinking of salty foods and overexposure to cold depletes the Kidneys.

Kidney Yin energy flows to the Liver, Heart, and Lungs. It is responsible for the body's Jing

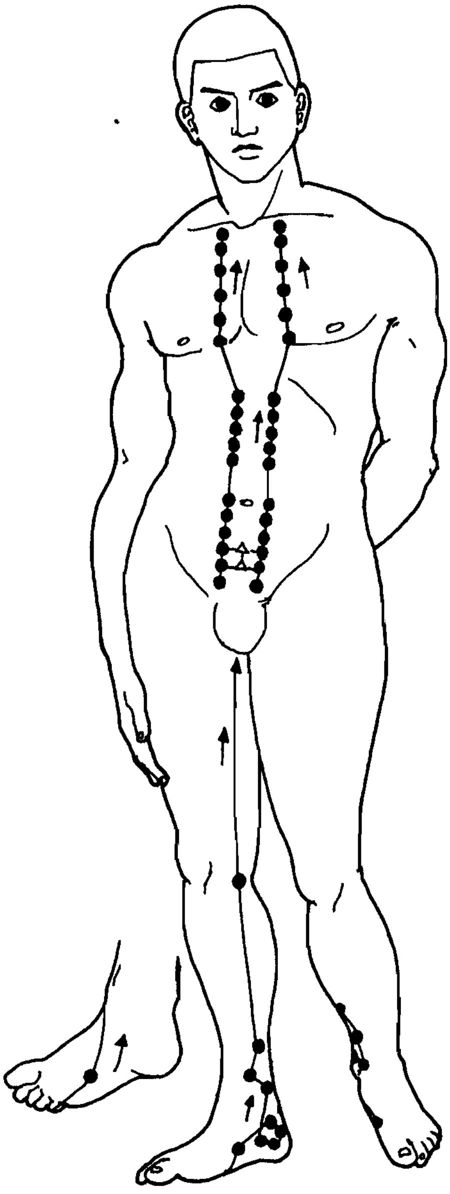

Figure 47.1. The Kidney (Kd) Channels

and produces the birth, growth, maturation, and reproduction cycle. Kidney Yang energy flows to the Spleen, Liver, Heart, and Lungs. It supports the Yang of all the body's organs via the Mingmen (Figure 47.2).

KIDNEY COMPLICATIONS AND SYMPTOMS

This section focuses on specific Kidney diseases and how to rectify them. The following discussion of treatment principles begins with an explanation of Kidney syndromes and symptoms. Several Kidney diseases have symptoms which are common to the organ's energetic dysfunction, such as mental and emotional problems, lower back pain, tiredness, and breathlessness. A list of Kidney syndromes and their symptoms are described as follows.

1. A Kidney Yang Deficiency syndrome manifests through such symptoms as headaches, breathlessness, wheezing, asthma, tiredness, mental and emotional problems, lower back pain, diarrhea, constipation, painful urination, enuresis, edema, menorrhagia and premenstrual tension.
2. A Kidney Yin Deficiency syndrome manifests through such symptoms as tiredness, chest tightness, asthma, breathlessness, mental and emotional problems, constipation, enuresis, lower back pain, bleeding, Parkinson's disease, menorrhagia and premenstrual tension.
3. A Kidney Jing Deficiency syndrome manifests through such symptoms as tinnitus, dizziness, and mental and emotional problems.

GENERAL TREATMENT FOR KIDNEY DISEASES

The following is a description of a general protocol used for treating a patient's Kidneys. The goal is to introduce the Qigong doctor to the basic hand positions and areas of the patient's body through which Qi is emitted (or Turbid Qi removed). Once the Qigong doctor becomes proficient at locating these specific points, the treatment will flow more smoothly.

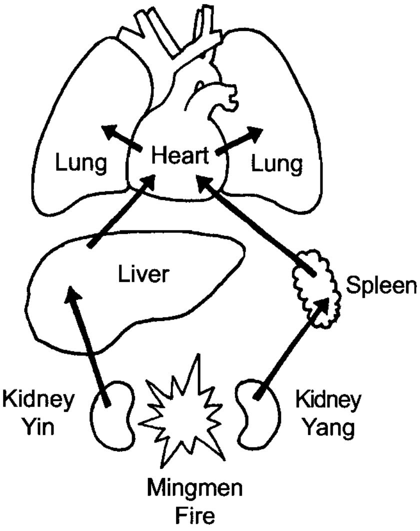

Figure 47.2. Kidney Yin flows from the Kidneys to the Liver, then to the Heart and Lungs. Kidney Yang flows from the Kidneys to the Spleen, Liver, Heart and the Lungs.

After comparing the various Kidney therapies, the reader will notice that certain patterns repeat themselves.

1. Generally, the Qigong doctor accesses the patient's Kidneys through either direct energetic insertion into the Kidneys' anatomical location on the back, through the naval and Mingmen access gates, or through the patient's channels and external energetic fields (Figure 47.3).
2. After treating the patient (purging, tonifying, and regulating), the Qigong doctor will stimulate the patient's Kidneys, Mingmen, and Lower Burner areas, and roots the patient's Qi into the Lower Dantian.
3. Next, the Qigong doctor stimulates the patient's Kidneys and Lower Dantian through the Taiji Pole and Bubbling Spring (Kd-1) points, regulating the body's Yin and Yang Qi via the Microcosmic Orbit (Fire Cycle).

Keep in mind that these areas are general and that specific treatment patterns must change as they apply to the patient's exact condition (Excess, Deficient, etc.).

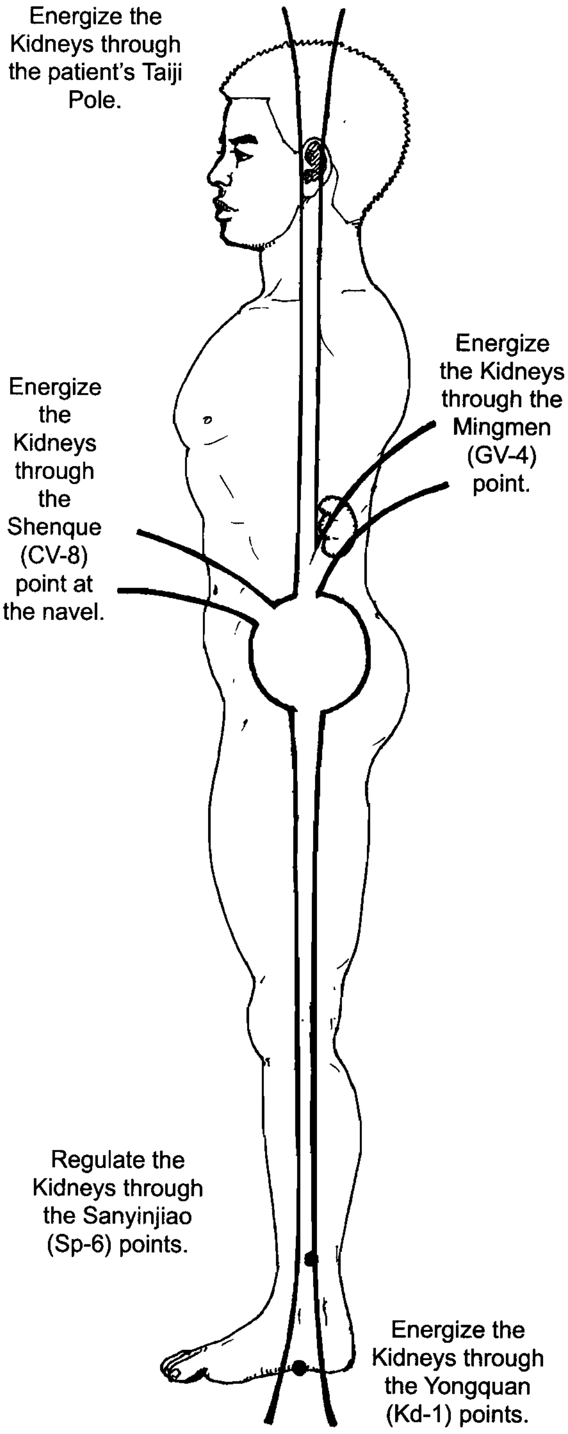
Energize the Kidneys through the patient's Taiji Pole.
Energize the Kidneys through the Mingmen (GV-4) point.
Energize the Kidneys through the Shenque (CV-8) point at the navel.
Regulate the Kidneys through the Sanyinjiao (Sp-6) points.
Energize the Kidneys through the Yongquan (Kd-1) points.

Figure 47.3. Treating Kidney Diseases

MEDICAL QIGONG THERAPY FOR KIDNEY QI DYSFUNCTION

Some symptoms of Kidney disease can manifest as dark circles under the eyes, decaying teeth, distended abdomen, dyspnea, and perspiration. The primary therapy used to treat Kidney diseases include: the Windy Breathing method to strengthen the Kidneys, Toe-Raised Walking, and Kidney massage (see Chapter 15).

1. An Excess Heat syndrome of the Kidneys should be treated by purging through uttering the "Chui" (Chree) sound, and reinforcing its effectiveness by focusing intent on the exhalation (Figure 47.4). This gradually relieves sluggishness and stagnation of the Kidney's Qi.
2. A Kidney Deficiency syndrome manifests mid-lower back pain at the waist level, and should be treated by Taking in the Blue Qi (see Chapter 4) into the Kidneys, Mingmen, and Lower Dantian.
3. A Deficiency of Mingmen Fire should be treated by emitting Qi into the patient's Lower Dantian and Mingmen area with the Pushing technique along with clockwise Qi rotation. This manipulation invigorates the Kidneys' Qi and replenishes the Fire of the Mingmen.

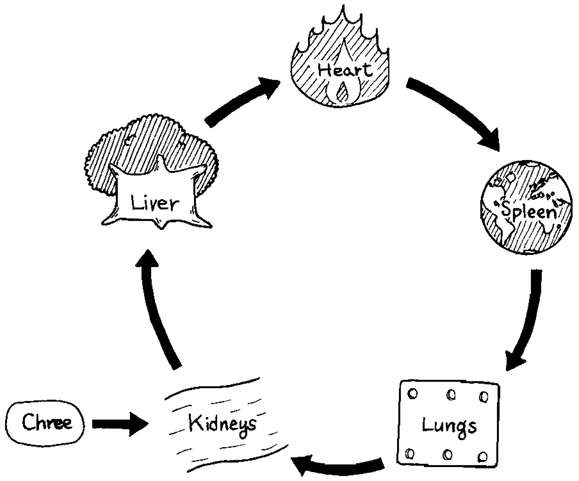
Heart
Spleen
Lungs
Kidneys
Liver
Chree

Figure 47.4. The Kidney sound is "Chui" (Chree).

KIDNEY QIGONG PRESCRIPTIONS

These exercises promote the flow of Qi and Blood of the Kidney Channels, nourish the Kidneys, strengthen Yang, and increase the Yuan Qi. These exercises can also be used to prevent and treat the following diseases and syndromes: nephritis, psychoneurosis, lumbago, rachialgia (spinal inflammation), tinnitus, deafness, frequent micturition and aversion to cold due to having a Cold condition.

PRESCRIPTIONS FOR TONIFYING THE KIDNEYS

The Taking in the Blue Qi can be used to tonify the Kidneys. The properties inherent within the color blue have a persistent vibratory rate used for tonifying the Kidneys and Bladder.

1. Assume a standing, sitting, or lying posture. Relax the whole body, place the tongue against the hard palate behind the teeth, and get rid of any stray thoughts.
2. When inhaling, visualize blue energy in front of you. Inhale the blue light energy in through the nose and down into the Kidney organs. As you exhale through the mouth, the dark, pathogenic Qi leaves the Kidneys; however, the bright clean color remains in the organ, stimulating and vitalizing it. With each breath the organ retains more clean energy and begins to glow. Repeat for five breaths.
3. Next breathe the blue color in through the nose, filling the entire mouth. When exhaling through the nose, send the blue Qi slowly down to the Kidneys, then into the Lower Dantian. Perform six to twelve times. Return to the beginning posture and close.

PRESCRIPTIONS FOR REGULATING THE KIDNEYS

Remove stagnation from the Kidneys by having the patient rub the abdomen while sounding "Chree."

Assume a sitting or standing posture. Place both palms flatly on the lower abdomen (left hand on the outside for men, opposite for women), inhale slowly, and exhale the sound "Chree" while rubbing the lower abdomen clockwise, gently with the palms. Perform ten to twenty breaths.

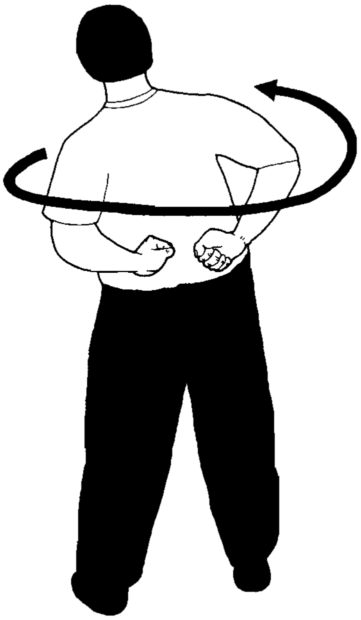

Figure 47.5. Rotate the waist in a clockwise and counterclockwise direction.

PRESCRIPTIONS FOR STRENGTHENING THE KIDNEYS AND CONDUCTING QI

The following exercise is used to tonify the Kidneys.

1. Assume a standing posture. Clench both hands into fists, and place them against the soft tissue on both sides of the Mingmen. Rotate the waist, turning towards the left in a clockwise direction for six rotations. After the six rotations pause in the center, then rotate in the opposite direction for six times (Figure 47.5).
2. Assume a standing or sitting posture. Rub the sides of the Mingmen with both hands, up and down thirty-six times, focusing the mind on the lumbar region (Figure 47.6).
3. Tuck the sacrum under (to stimulate the renal capsule), placing both palms on the groin, under the junction of the pubic bone (left hand on the outside for men, right for women).

Support and gently pull up the area around the perineum with both hands eighty-one times (Figure 47.7). Switch hands and repeat. Note: The Kidneys are injured by irregular

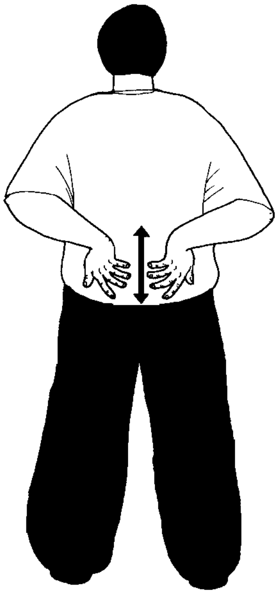

Figure 47.6. Rub the Kidney and Mingmen area.

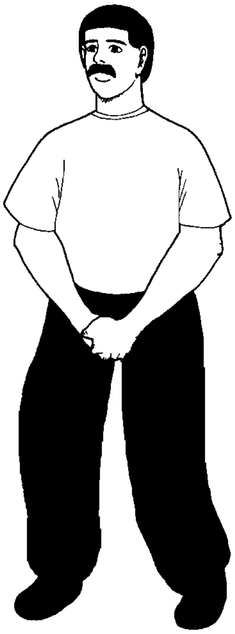

Figure 47.7. Support the groin and perineal area.

sleep patterns, fear, excessive caffeine, sex, drugs, alcohol, and smoking. Excess saltiness and Cold (climate and food) have a draining effect on the Kidneys.

SPECIFIC MEDICAL QIGONG THERAPY FOR KIDNEY DISEASES

The following are several Medical Qigong treatments used in China to treat specific Kidney diseases.

TINNITUS AND DEAFNESS

Tinnitus is a condition of a high-pitch or rushing-wind sound within one or both ears. The sound may be a consistent pitch or vary in intensity and modulation.

ETIOLOGY

Although the ears are the opening of the Kidneys, there are other influences which must be addressed when dealing with tinnitus. Tinnitus is most commonly caused by stagnation of the Triple Burner and Gall Bladder Channels. These channel stagnations are due to three adverse conditions: the rising of Liver Wind Fire which may be induced by violent rage, the obstruction of the orifices in the head due to External Wind invasion, or by failure of the Kidney Jing to reach the ears due to a Kidney Deficiency.

This disease can be divided into Full and Empty Types of tinnitus:

1. The Full Type of tinnitus is caused by the flaring up of pathogenic Wind, Yang, Fire, Phlegm, or Phlegm Fire. It is most noted in the Excess rising of Liver and Gall Bladder Fire, or Phlegm Fire flaring upwards. This type of tinnitus is experienced as a sudden onset with a loud sound and is aggravated by cupping one's hands over one's ears.
2. The Empty Type of tinnitus is caused by a Deficiency of Kidney Qi, Kidney Jing, Lung Qi, or Heart Blood that is unable to extend to the ears. It is most noted in Kidney Jing Defi-

ciency, Deficient Upper Burner Qi and Heart Blood Deficiency. This type of tinnitus manifests clinically through a gradual onset with a low sound and is improved by cupping one's hands over one's ears.

SYMPTOMS

1. The Excess Type of tinnitus manifests as a sudden deafness or distending pain in the ears accompanied with a loud-pitched ringing noise. The adverse rising of Liver Wind Fire or Gall Bladder Wind Fire causes vexation and irritability, and often manifests as a flushed face and dry mouth.

   The Excess condition caused by an invasion of External Wind also includes symptoms such as chills, fever, and headache.
2. The Deficient Type of tinnitus can either occur constantly or intermittently, and may be aggravated by overstraining and relieved by hand pressure. This condition is often accompanied by dizziness, soreness of the waist, seminal emission, or leukorrhea.

TREATMENT

1. Begin by stimulating the patient's TB-3, TB-17, SI-19, and LI-4 points on both sides of the body.
2. Emit Qi towards the patient's Lower Dantian, using the right Sword Finger hand posture and left Extended Fan Palm hand posture.
3. Next, extend energy into the patient's affected ear and focus the Qi emission around the patient's TB-17 and TB-21 point areas. Next, guide the Qi to flow down the head, neck, shoulders, and arms along the Triple Burner Channels and out of the hand 3 to 7 times.
4. Finally, extend Qi into the patient's affected ear and guide the Qi down and out of the body along the corresponding Gall Bladder Channel ending at the corresponding foot.

MODIFICATIONS

- In cases of adverse rising of Liver Wind Fire, use the Vibrating Palm to emit Qi into the patient's Bl-18 and -19 points for 14 breaths on both sides of the body.
- In cases of Kidney Deficiency, emit Qi into the patient's Bl-23 points for 14 breaths.

PRESCRIPTIONS AND HOMEWORK

In the following three prescriptions make sure the patient heats his or her hands before beginning.

1. Have the patient perform the Beating the Heavenly Drum exercise (see Chapter 16) for 24 breaths.
2. Have the patient press the ears with the palms in a compression-release fashion 10 times each. Be sure to warn the patient to avoid forceful or violent compression. The patient's intention should focus deep within the orifice of the ears.
3. Have the patient pinch the top of the auricles of the ears gently, and massage them from the top downwards 24 times.

LUMBAGO

Lumbago is a general non-specific term for dull aching pain in the lumbar region of the back.

ETIOLOGY

Lumbago is a common symptom observed in many diseases, such as acute and chronic muscular strain of the lower back, subluxation of the discs, prolapse of the lumbar vertebra, etc. Patients may suffer from lumbago with pain radiating down to their lower limbs. Lumbago pain can also be aggravated by weather changes. Traditionally, there are three general causes of lumbago: lumbago due to Wind Cold and Dampness, lumbago due to Kidney Deficiency, and lumbago due to trauma.

1. Lumbago due to Wind Cold and Dampness, is caused by retention of Cold and Damp pathogens in the channels and collaterals. This in turn leads to the obstruction or stagnation of Qi and impairs Blood circulation.
2. Lumbago due to Kidney Deficiency is caused by an over consumption of the Kidney Jing due to an overactive sex life. This results in the failure of the Kidneys to nourish the Triple Burner Channels.
3. Lumbago due to trauma occurs from an injury to the tissues and muscles (e.g., dislocated vertebra, fractures, ruptured spinal disks, etc.), and leads to stagnation of Qi and Blood. Qi and Blood stagnation leads to obstructions in the channels and collaterals.

SYMPTOMS

1. Lumbago, due to Wind Cold and Dampness, is manifested by heaviness and pain at the waist, which may radiate to the hips and down the legs, or may be accompanied by muscular stiffness. In chronic cases, the pain is intermittent with various degrees of severity. On wet and cold days the patient often feels cold at the waist and lumbar area.
2. Lumbago, due to Kidney Deficiency, is manifested by continuous, dull pain, accompanied by soreness and weakness of the waist and knees, cold limbs and nocturnal emission.
3. Lumbago due to trauma is manifested by localized pain along with tenderness and limited mobility.

TREATMENT

1. Place the patient in a sitting position on the table. Begin stimulating the patient's Mingmen area with your left hand while the right hand guides the energy down to the patient's Bl-40 points behind the knees.
2. Using the Extended Fan Palm hand manipulation emit Qi into the patient's Mingmen and Bl-18 areas, guiding the Qi to flow along the Bladder Channels down the legs.
3. Next, place the patient in a supine posture, knees pointing upwards and feet flat on the table. Position yourself over the patient allowing the weight of your chest to support and compress the physical structure of the patient's knees, lower legs and feet. Embrace the patient's hips and imagine separating the inner fascia and tissues. The purpose for this compression is to initiate a strong internal stretching action within the tendons and smaller muscles of the patient's lower back and sacral area. This technique is used to disperse stagnant Qi and Blood within the patient's lumbar vertebra, sacrum and femur articulations. At the same time, direct the Qi upwards along the patient's spine and Governing Vessel. Hold this compression for several minutes.
4. To avoid respiratory congestion (while your weight is stretching the patient's sacrum and lower lumbar), the patient is instructed to raise the arms above the head.
5. As the patient experiences the lower back compression, he or she imagines the stagnant energy (trapped within the lumbar area) melting like ice into water and flowing through the Microcosmic Orbit (Fire Cycle). With each exhalation the patient relaxes to a deeper level, imagining him or herself sinking deeper into the table.

TREATMENT MODIFICATIONS

1. For lumbago due to Wind Cold and Dampness, purge the pathogenic Qi out of the patient's body first, and then emit Heat into the channel points associated with the specific location of the patient's pain.
2. For lumbago due to Kidney Deficiency, use a Vibrating Palm technique to emit Qi towards the patient's Mingmen, Bl-23 and Bl-18 points for 14 breaths each point on both sides of the body.
3. For lumbago due to trauma, extend Qi for 28 breaths into any Ah Shi point (painful tissue area).

INSOMNIA

Insomnia is described as a sleeping disorder that disrupts the patient's normal sleeping pattern. In mild cases, the patient may have difficulty falling asleep or remaining asleep (waking numerous times during the night). In severe cases the patient may be unable to fall asleep at all. Diseases with insomnia as part of their manifestation include anxiety disorders and menopausal syndrome.

ETIOLOGY

The amount and quality of sleep depends on the state of the patient's Shen, which is rooted in the Heart (specifically, the Heart Blood and Heart Yin). If the Heart is agitated or Deficient, the Shen cannot root, and sleep will be affected. Any disharmony of the body's internal organs due to an Excess or Deficient condition affects the patient's Jing and Blood. Since the Jing and Blood root the Shen, the Shen will have no residence and insomnia may result.

Insomnia can be caused by insufficient Blood due to impairment of the Spleen Qi, or from malnourishment of the Heart Qi due to a Deficiency of Heart Blood. This condition can also result from overfatigue, anxiety or irritability. This may be due to a disharmony between the Heart and Kidneys (resulting from Deficient Kidneys), irritability (due to mental depression) or from disharmony between the energetic function of the Stomach and Spleen (resulting from an improper diet).

Since the Hun (rooted in the Liver Blood or Liver Yin) play an important role in the length and quality of sleep, Liver Blood Deficiency can cause the Hun to be deprived of their residence and to wander at night causing restless sleep.

SYMPTOMS

The main symptom is an inability to sleep, which based on the different conditions of this illness can be further classified as:

- An inability to get to sleep,
- Sleepwalking, and an
- inability to fall asleep after waking.

The symptoms vary with the causes of the disease.

1. Insomnia due to a Deficiency of the Heart and Spleen (resulting from severe worry) may cause shallow sleep, or dream disturbed sleep. It is manifested by palpitations, poor memory, vertigo, dizziness and spontaneous sweating.
2. Insomnia due to Kidney Deficiency (resulting from excessive sexual activity) is manifested by dizziness, tinnitus, soreness in the loins and back area, as well as involuntary seminal emission.
3. Insomnia due to upward stirring of Liver Yang is manifested by pain and distention of the chest and epigastrium, belching, and acid regurgitation, as well as mental hyperirritability, dizziness and headaches.

TREATMENT

1. Begin by having the patient sit at the edge of the table. Your right hand emits Qi into the patient's CV-12 point, directing the energy to flow down into the patient's Lower Dantian. The left hand extends energy into the patient's mid-back around the area of the patient's Bl-18, Bl-19, Bl-20 and Bl-21 points, and leads the Qi down the Bladder Channels to the feet on both sides.
2. Next, use the Vibrating Palm technique to extend energy into the patient's Baihui, GV-14, Mingmen, CV-12 and Lower Dantian areas for 9 to 18 breaths each.
3. Emit Qi into the patient's Baihui area and leads the Qi from the top of the patient's head down the Conception Vessel into the Lower Dantian.
4. Finally, emit Qi into the patient's GB-8 point by the top of the ears and guide the Qi down the patient's Stomach Channels to the St-36 points.

TREATMENT MODIFICATIONS

1. For treatment of insomnia due to Deficiency of Heart and Spleen, emit Qi into the patient's Bl-20 and Bl-15 points, guiding the Qi down along the Bladder Channels to the feet.
2. For insomnia due to Kidney Deficiency emit Qi into the patient's Mingmen and Bl-23 points for 18 respirations.
3. For insomnia due to emotional depression, emit Qi towards the patient's Lv-3 and Lv-4 points (on both sides of the body), guiding the Qi down to the patient's lower extremities along the Liver and Gall Bladder Channels.
4. For insomnia due to Spleen and Stomach disharmony, use the Vibrating Palm hand technique to emit Qi into the patient's CV-12, St-36, and St-21 points on both sides of the body.

PRESCRIPTIONS AND HOMEWORK

For the best results, have the patient practice Medical Qigong prescriptions one to two hours before going to bed. For chronically ill patients, the causes of the insomnia should be removed first, followed by Medical Qigong prescriptions and regular physical exercise for lasting therapeutic results.

Have the patient perform the Kidney Massage for treating insomnia prescribed in Chapter 39.

PYELONEPHRITIS

Pyelonephritis is an infectious disease caused by bacteria invading the renal parenchyma and renal pelvis unilaterally or bilaterally. It clinically manifests through the following symptoms: fever, lumbago and abnormal discharge of urine.

SYNDROMES

1. Damp Heat in the Bladder is clinically manifested through symptoms such as frequent and painful urination, fever, distention and pain in the lower abdomen, lumbago and aversion to cold.
2. Stagnation of Liver and Gall Bladder Heat is clinically manifested through symptoms such as fever with or without chills, nausea, loss of appetite, lumbago or pain in the lower abdomen, frequent urination with cloudy urine and feeling agitated.
3. Spleen and Kidney Deficiency with Heat stasis is clinically manifested through such symptoms as edema of the face, swelling of the feet, abdominal distention, thin stools and frequent urination, lassitude and weakness in the entire body (especially the legs), dizziness, tinnitus and anorexia (lack of appetite).

TREATMENT

1. Have the patient sit at the edge of the table and begin stimulating the Mingmen and Lower Dantian areas using a Vibrating Palm technique for 11 to 22 breaths.
2. Next, have the patient lay supine on the table and stimulate the patient's Sp-6 and Kd-3 areas on both sides of the body.
3. Use the Extended Fan Palm hand technique to emit Qi into the patient's Lower Dantian, guiding it down the Kidney Channels to the bottom of the feet.

TREATMENT MODIFICATIONS

1. When treating patients with Damp Heat in the Bladder, use a Sword Fingers hand posture and apply the Vibrating technique, emitting Qi into the patient's Mingmen and Lower Dantian for 14 to 28 breaths.
2. When treating patients with stagnant Heat in the Liver and Gall Bladder, use the Extended Fan Palm hand posture to emit Qi into the patient's mid-back and Liver, purging Toxic Qi down the Liver and Gall Bladder Channels and out of the body via the legs.
3. When treating patients with Deficient Spleen and Kidneys, use the Extended Fan Palm hand posture to emit Qi into the patient's mid-back Bl-20, and Bl-23 points, CV-12 and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

Have the patients practice the Standing Post Posture with arms suspended at the Middle Dantian, and the mind's intention focused on the Lower Dantian. Use natural abdominal breathing (see Chapter 16).

PRESCRIPTION MODIFICATIONS

1. For patients with Damp Heat in the Bladder, have them practice rubbing the costal area and sounding "Who."
2. For patients with stagnant Liver and Gall Bladder Heat, have them practice massaging the costal area and sounding the word "Shu."
3. For patients with Spleen and Kidney Deficiency, have them practice the method of Taking in Yellow Qi, as well as the Gathering the Moon Cream exercise (see Chapter 11).

IMPOTENCE

Impotence is a weakness and inability of the male to achieve or maintain an erection. Symptoms vary from premature ejaculation to partial penile erection (the penis is erect but not hard, resulting in the incapability of sexual intercourse), to a complete failure to achieve any erection at all. Impotence indicates a weakness in all Five Yin Organ energies. The combined energies must be sufficient to culminate an erection, as described below.

1. The Heart energy quality is strongly affected by the nature of the sexual encounter. Casual sex may not involve the highest spiritual energy of the Heart's upper chamber (see Chapter 5), only the basic instincts and drives of the lower chamber. The Heart may also be affected by the individual's moral values and beliefs, as well as the quantity and quality of emotional passion (i.e., with or without guilty feelings) at the time of arousal. Heart energy also pertains to the energy of the Blood.
2. The Spleen energy pertains to what the man

is thinking and his intention at the time of arousal. This also affects the man's muscles.

3. The Liver energy pertains to visual stimulation that affects a man's sexual drive. This affects the man's strength. Many aspects of man's sexuality are relate to the Liver's functions, i.e., visual images evoking emotions, Blood filling the erectile tissue, and the power of desire.
4. The Kidney's Qi pertains to the Jing energy stored within the Marrow, and relates to the man's will power and endurance. It is associated with audible stimulation and affects the man's endurance.
5. The Lung Qi is man's somatic expression of sexual passion and is stimulated through touch and smell. This energy also affects the man's endurance.

ETIOLOGY

There are two main factors ascribed to impotence: psychogenic and organic. In the first case, the patient's physical, mental, emotional, and spiritual energies are out of harmony. This can be caused by fear, excitement, performance anxiety, etc. In the second case, impotence can be caused by medications, drugs, alcohol, etc. In most organic cases, however, impotence is caused by prostate failure, cardiovascular disease or diabetes. If this condition persists long enough the man looses sexual interest and the condition becomes psychosomatic.

SYNDROMES

- Insufficiency of Kidney Yang is clinically manifested through symptoms such as impotence, lumbago, weakness in the knees, dizziness, tinnitus, aversion to cold, cold limbs and a cold sensation in the external genitalia, listlessness and shortness of breath.
- The downward flow of Damp Heat is clinically manifested through such symptoms as impotence, itching and pain around the external genitalia and scanty, dark urine.

TREATMENT

1. Stimulate the patient's Mingmen and Lower Dantian areas using the Vibrating Palm technique for 12 breaths.
2. Next, stimulate the patient's Sp-6 areas on each leg, leading the Qi up into the Lower Dantian for 12 breaths.
3. Finally, return back to the Mingmen and Lower Dantian areas using the Extended Fan Palm hand technique. Emit Qi for 24 breaths, rotating in a clockwise direction to gather the Qi in the Lower Dantian.

TREATMENT MODIFICATIONS

- When treating a patient with an insufficiency of Kidney Yang, vibrate the energy using the Sword Fingers Hand technique, while emitting Qi into the Lower Dantian for 16 to 18 breaths.
- When treating a patient with a condition of Damp Heat flowing downward, purge the Qi along the Kidney Channels down into the Lower Dantian. Use the Extended Fan Palm hand technique and emit Qi into the patient's Lower Dantian, Kd-13 and Kd-14 points on both sides of the body, as well as the navel area, for 24 breaths.

PRESCRIPTIONS AND HOMEWORK

It is advisable to have the patient perform the Deer exercise, prescribed for increasing the body's Jing.

1. Have the patient sit on the edge of the chair, rubbing the palms to create heat. This right hand will cup his testicles so that the palm completely covers them with slight pressure. The left hand is placed on the area of the Lower Dantian just below the navel. The left hand moves in a circular clockwise direction 81 times. Next, the hands are rubbed together again, reversing the hand positions so that the left hand cups the testicles and the right hand is on the Lower Dantian. The circle rubbing is repeated in the opposite direction 81 times. Concentrate on the heat and Qi filling up the Lower Dantian and genital area.
2. The patient should tighten and draw up the anal muscle while inhaling, feeling the Qi being drawn up the rectum and prostate area, filling up the Lower Dantian. He should hold the breath as long as possible, then exhale, release and relax. The anal contraction exer-

cise should be repeated beginning with 25 breaths, and eventually working up to 250 breaths.

PRESCRIPTION MODIFICATIONS

1. When treating patients with Kidney Yang Deficiency, it is advisable to have them practice Beating and Drumming the Qi with lower abdominal breathing (see Chapter 42).
2. When treating patients with a downward flow of Damp Heat, it is advisable to have the patient rub the Lower Dantian and sound the healing tones "Chree" for 36 breaths and "Shu" for 36 breaths.

CHAPTER 48

QI EMISSION THERAPY FOR MISCELLANEOUS DISEASES

INTRODUCTION

This chapter deals with miscellaneous diseases that are classified as either complicated multiple diseases, or energetic conditions which have not yet been classified. This section describes the conditions, symptoms and treatment for these syndromes. The discussion of these various treatments begins with energetic dysfunctions stemming from the Qi of the Three Dantians, and expands further to include other tissues, internal organ and energetic organ system complications.

DEFICIENT DANTIAN QI

One of the most common problems found in the body's energetic system is Deficient Dantian Qi. The symptoms vary according to the location of the Deficiency, as well as the patient's constitution.

1. If the energy is Deficient within the Lower Dantian, the symptoms and syndromes cover a large range that includes: chronic diarrhea, asthma due to Kidney Deficiency, hypertension due to Yin Deficiency, Blood Deficiency, insomnia, spermatorrhea, collapsing syndrome due to Deficiency, abandoned-type Wind Stroke, abandoned-type coma and all types of organ prolapse.
2. If the energy is Deficient within the Middle Dantian, the problems may include shortness of breath, bronchitis, asthma, chest pain, palpitations, intercostal neuralgia, all breast disorders including acute mastitis and breast abscesses.
3. If the energy is Deficient within the Upper Dantian, the problems may include frontal headaches, dizziness, vomiting, fainting due to Blood loss, insomnia, eye soreness, febrile convulsions and insanity.

PRESCRIPTIONS FOR REGULATING THE THREE DANTIANS

The following meditations are prescriptions used to regulate the heat, vibration, and light in the Lower, Middle, and Upper Dantians.

1. The Lower Dantian meditation solidifies the Qi in the Lower Dantian that allows for stable energy transference to take place. The purpose for Heat Regulation in the Lower Dantian is to cause the internal organs to steam; this transfers energy from organ to organ to nurture and regulate the body's Qi.

   Have the patient focus on breathing into the center of the Lower Dantian. Upon inhaling, the patient imagines drawing Qi into the body from the nose, urethra and anus, and directs it into the Lower Dantian. After inhaling, the patient closes the lower orifices and circulates the Qi up through the coccyx and Mingmen area, then down into the navel and into the Lower Dantian. It is important that the patient visualize a ball of energy rotating in the lower abdomen like a spinning wheel. As this Qi begins to collect in the Lower Dantian, heat and vibration begin to resonate in the lower abdominal area.
2. The Middle Dantian meditation expands the patient's Wei Qi, which in turn enhances the body's protection against an attack of Hot or Cold pathogenic factors. The patient should draw the Toxic Qi out of the Yang Channels of the arms, through the center point of the palms. This is an excellent meditation for treating diseases caused from Excess Heat, such as insomnia, headaches, hypertension, and arthritis of the joints:

   To regulate the heat of the Heart Fire and Middle Dantian area, the patient focuses the breath and mind's intention on the center of the chest (CV-17) while placing the right thumb inside the heart of the left palm (Pc-8), while resting both hands on the lap.

3. The Upper Dantian meditation not only expands the patient's Wei Qi, but also fills the body with sound and light vibration. Sound and vibration are used in the Qigong clinic to purge Toxic Qi out of the patient's body. This is an excellent meditation for balancing and energizing all Three Dantians.

   The patient begins in a standing or sitting Wuji posture, with the spine straight, Mingmen pressed back, anal sphincter closed, tongue on upper palate, and breathing naturally. This particular exercise follows the opening and closing pattern of Qi regulation. The focus is placed on each Dantian area for a period of 18 to 24 breaths. Each Dantian area needs a specific sound and light visualization to open, purge, close, and regulate its energy. The sounds and physical actions are as follows.

- To open, energize and regulate the Upper Dantian, use the sound of "Om." The arms separate and come together above the Baihui area (above the head) in order to open and close the Upper Dantian. The focus of the mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling, imagine the energy rushing into the Upper Dantian while toning the first vowel sound "O." Then imagine light shinning outside of the body while sounding the consonant sound "m." On the beginning of the vowel sound, the tongue is placed on the lower palate. As the tone and breath reach its midpoint, the tongue rises to connect with the upper palate to finish the sound "Om."
- To open, energize and regulate the Middle Dantian use the sound of "Ha." The arms separate and come together above the chest at the CV-17 point while opening and closing the Middle Dantian. The focus of the mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling the tone "Ha", separate the arms and imagine that the Qi rushes into the Middle Dantian and then expands to shine outside of the body. On the beginning of the tone "Ha," the tongue should be placed on the lower palate. As the tone and breath reach its midpoint, the tongue maintains its connection to the lower palate.
- To open, energize and regulate the Lower Dantian use the sound of "Hai-riem." The arms separate and come together above the lower abdomen at the CV-8 point while opening and closing the Lower Dantian. The focus of mind's intention is placed on drawing in divine healing light while inhaling, to fill the Taiji Pole. When exhaling, imagine the Qi rushing into the Lower Dantian with the first syllable "Hai," then shines outside of the body when the syllable "reim" is pronounced. At the beginning, the tongue is placed on the lower palate to pronounce the sound "Hai." As the tone and breath reach its midpoint, the tongue changes its position to behind the teeth to tone the sound "reim."

All three sounds are used to activate and energize the body's Three Dantians by practicing these sounds in the following two step progression:

1. Upon inhalation, focus the mind's intention on drawing divine healing light into the body to fill the Taiji Pole.
2. While exhaling, sound the specific tone of each Dantian, while both hands perform the opening (the expanding movements of the arms and hands) and closing (the contracting movements of the arms and hands) technique at each Dantian. The mind visualizes the light resonating and shining out from the body at each Dantian during the toning.

IMPROPER YIN AND YANG ENERGY BALANCE

Another common problem observed in the clinic is an imbalance of Yin and Yang Qi within the patient's body. Converting and balancing the Yin and Yang energy diverts the Qi before it has a chance to overload the body's system. The Qigong doctor first removes any obstruction in the patient's Governing Vessel and then harmonizes the patient's Yin and Yang energy by extending Qi into the patient's back, on the GV-13 and -14 points.

The following meditations are used as prescriptions to develop a strong electrical charge in the body, that balances the Yin and Yang currents:

Have the patient place the thumb in the heart of his or her palm (Pc-8), fingers touching the back (dorsal) side of the hand, opposite the (Pc-8) point (Figure 45.1). To energize the Lower Dantian the patient concentrates on his or her lower abdominal area. To energize the Middle Dantian the patient should concentrate on the area at center of the chest. The only difference between the Lower Dantian exercise and the meditation used to heat the Middle Dantian is the focal point of the patient's concentration. In these particular exercises, the hand postures are the same, however, the mind is focused on either the hands near the Lower Dantian, or on the middle of the chest.

- For men to develop a stronger electropositive Yang energy field, the right hand should cover the top and center of the left hand. To develop a stronger electronegative Yin energy field, the left hand should cover the top and center of their right hand.
- For women to develop a stronger electropositive Yang energy field, the left hand should cover the top and center of the right hand. To develop a stronger electronegative Yin energy field, the right hand should cover the top and center of the left hand.

Note: In generating energy within each Dantian area it is important to understand basic Yin and Yang polarities and their relationship to male and female. The female chest is considered Yang while the female genital area is considered Yin. The female Middle Dantian (or Heart area) is the focal point of female Yang (positive) energy; the genital area is the focal point of female Yin (negative) energy. Conversely, the male chest is considered Yin while the male genital area is considered Yang. The Middle Dantian (or Heart area) is the focal point of male Yin (negative) energy; the genital area is the focal point of male Yang (positive) energy (see Chapter 5).

DISORDERS OF THE QI

Whenever energy movement is impeded, disorders in Qi functions occur. Generally, these disturbances cause either disorders of Fire, or disorders of Water (see Chapter 22).

The cause and progression to these disorders can be further broken down and classified according to their energetic dysfunction. There are several types of Qi disorders observed in the clinic and they are generally described as follows.

LEAKAGE AND LOSS OF QI

During or after training practice, the patient may feel that Qi is continuously and uncontrollably leaking out of the body from the genitalia and anus or somewhere else on the body. This loss of Qi can lead to Qi Deficiencies and manifest through the following symptoms: emaciation, weakness of the limbs, grayish and dull complexion, nervousness, mental disturbances, distractibility, failing memory, spontaneous sweating, night sweats, seminal emission, insomnia, lassitude and sluggishness.

To treat this condition apply the Extended Fan Palm technique, use the Pushing, Pulling, and Shaking manipulations to emit Qi into the Mingmen; then use Vibrating and Shaking manipulations to emit Qi into the Lower Dantian.

QI STAGNATION DISORDERS

Qi stagnation exists when the Qi becomes sluggish when moving within the body's channels, organs, or tissues. Qi stagnation can be caused by many different factors such as suppressed emotions, trauma, contraction of microorganisms, poor diet, and weak digestion. Qi stagnation, over time, results in localized pain, distention, cysts, tumors, etc. Qi stagnation disorders have two subcategories: Qi and Blood Stasis, and Obstructed Qi (see Chapter 22).

1. Qi and Blood Stasis refers to the sensation of local pain, heaviness, soreness, distention, and compression, which are caused by the stagnant Qi and Blood in certain areas of the body. These sensations do not resolve by themselves. This stasis may occur during or after

Medical Qigong training practice from Qi disorders and visceral malfunctions.

2. Obstructed Qi is also considered a subcategory of Qi stagnation, and exists when the Qi becomes so stagnant that it no longer flows. At this stage the Qi can't move the Blood. This can be caused by trauma, surgery, and acute Qi stagnation.

Reconstructive Qi therapy, for conditions of Qi stagnation use the following two methods:

1. Select corresponding channel points in the region of the Qi and Blood stasis. Knead the selected channel points, then push and rub along the obstructed channel.
2. Finish the treatment by applying the Extended Fan Palm technique; use the Pushing, Pulling and Shaking manipulations to emit and conduct Qi along the channels to open them.

REBELLIOUS QI DISORDERS

Rebellious Qi disorders manifest as Qi flowing in the wrong direction (a direction different from its normal physiological progression); this causes a disturbance within the harmonious flow of energy and affects the organ functions. The symptoms have to be analyzed according to the internal organ which is involved. There are two types of Rebellious Qi, the Excess type and the Deficient type. Rebellious Qi is categorized as follows.

DISORDERS OF EXCESS REBELLIOUS QI

Disorders of Excess Rebellious Qi occur when the vital Qi increases and forces the pathogenic Qi to move. This causes local sensations of pain, soreness, distention, heaviness, coldness or hotness which can develop during Qigong practice.

To treat for conditions of Rebellious Qi, use the following two methods:

1. From the head and torso, follow the routes of the channels and collaterals out to the extremities. Pinch and lightly pull on the extremities (fingers and toes). Then press, knead, push and rub the location of the discomfort, guiding and leading the Qi along the channels away from the area of discomfort.
2. Apply the Extended Fan Palm method; use the Pulling and Leading manipulation to pull and guide the Qi, or open up the channel points to disperse the Qi.

DISORDERS OF DEFICIENT REBELLIOUS QI

Disorders of Deficient Rebellious Qi refer to the disorderly flow of Qi and Blood during or after training practice, which can cause dizziness, fright, a heavy feeling in the chest, shortness of breath, shaking limbs, trembling of the hands and feet or fainting. Generally, energy sensitive practitioners know the location and direction of the disorderly flow of Qi in their tissues.

To treat conditions of Deficient Rebellious Qi use the following two methods:

1. Open the appropriate channel points to regulate the functional activities of the Eight Extraordinary Vessels.
2. Select appropriate regions and channel points and apply the Extended Fan Palm or Sword Fingers technique. Use Pushing, Pulling, and Shaking manipulations to activate the channel Qi of that particular channel or region. Then, use the Leading manipulation to direct the Qi into specific channels or return it to the Lower Dantian.

ARTHRITIS

Arthritis is defined as inflammation of the joints, usually accompanied by pain, swelling and frequently changes in structure. This is a chronic, multiple arthropathy and can be divided into the peripheral type (pertaining to the four limbs), and the central type (pertaining to the spinal column, i.e., rheumatoid arthritis or ankylosing spondylitis).

ETIOLOGY

Arthritis is caused by pathogenic Wind, Cold and Dampness attacking the body's channels and hindering the Qi and Blood circulation. This invasion can give rise to pain, soreness, heaviness, and numbness of the tendons and muscles, as well as swelling and difficulty in moving the joints. This condition can be caused from an unstable lifestyle weakening the body's Wei Qi, overexposure to a draft after sweating, wading in water, exposure to cold weather, or laying on a damp ground for long periods of time.

SYMPTOMS

Arthritis is clinically manifested primarily as arthralgia and dysfunction of the joints and can be accompanied by muscle spasm, loss of muscle tone, loss of muscle mass and deformity. Symptoms can be divided into early and late stages.

- In the early stages of arthralgia, the local swelling usually starts from the small joints of the fingers and toes, wrists and ankles, or from the iliosacral joint and spreads upward into the body. Several weeks or months later, the patient may feel local tenderness and inflexibility of the joints with some clicking or cracking sounds. In this stage the pain is worse at night and early in the morning, and may be alleviated with gentle exercise. In the early stages of arthritis, the patient's condition can be controlled and even restored back to normal through Medical Qigong prescriptions.
- In the later stages of arthritis, the articulation of the joint movement is markedly limited, and patients may experience deformity in the joints, as well as stiffness, muscular atrophy and paralysis. The phalangeal and carpal joints of the patient's hands may become rigid and deviate to the ulnar side of the arm. The fingers may also become deformed with restricted mobility, and the elbow and shoulder joints may also be affected. In the late stages of arthritis, the treatment can only control the progression of the disease and alleviate local symptoms. The patient's locomotive function can only be gradually improved.

If the patient's lower limbs are affected, the ankle, knee and hip joints may become rigid, and he or she may experience motor impairment. If arthritis occurs in the patient's spinal column, it may cause kyphotic deformity and dyspnea, as well as restriction in neck movement. At this stage a sudden backward flexing of the neck may cause sudden death in patients with rigid forward bending neck deformity.

When treating an External invasion of pathogenic factors, the Qigong doctor observes three main symptoms of arthritis:

1. Wind or migratory arhtralgia is manifested by pain that has no particular fixed location, but moves throughout the body's extremities and joints, and is sometimes accompanied by chills and fever.
2. Cold arthralgia is manifested by localized pain of the joints, which may be relieved by heat and aggravated by cold.
3. Damp arthralgia is manifested by fixed pain and swelling in the joints and extremities, or by numbness of the skin which may be aggravated in cloudy and rainy weather.

TREATMENTS

1. Place the patient in the standing posture and begin dredging the Governing Vessel, drawing the energy into the ground.
2. Use the Extended Fan Palm hand technique and begin to emit Qi into the patient's GV-14, Mingmen, and Lower Dantian areas.
3. While emitting Qi into the painful joint areas, purge the pathogenic energy out of the patient's body, leading the Toxic Qi along the channels and out of the extremities.

PRESCRIPTIONS AND HOMEWORK

In treating arthritis it is important to prescribe exercises which lubricate the joint articulations to increase the production of Qi and Blood to the extremities. The Daoist Five Yin and Yang Organ exercises will accomplish this goal quite easily.

The patients are required to perform Qigong prescriptions which improve the body's resistance, but do not cause fatigue. Patients should also improve their diet and should be encouraged to consult a nutritionist for natural supplements that can help rebuild healthy cartilage and reduce pain. They should also make every effort to keep their body warm.

For patients with osteoporosis, stretching techniques, as well as sudden forward flexing of the neck are absolutely prohibited.

PRESCRIPTION MODIFICATIONS

1. When treating patients who suffer from arthritic pain in the upper limbs, advise them to practice shoulder and arm Dynamic Medical Qigong, moving at a slow, fluid pace.
2. When treating patients who suffer from ar-

thritic pain in the lower extremities, advise them to practice Dynamic Walking Medical Qigong (see Chapter 15), moving at a slow, fluid pace.
3. When treating patients who suffer from arthritic pain in the waist and lower back, advise them to practice the Expanding and Contracting the Rings exercise (see Chapter 41).
4. When treating patients who suffer from arthritic pain in the neck, advise them to practice slow-moving non-forceful neck rotation exercises in conjunction with stationary neck massage.

VERTIGO

Vertigo is a syndrome manifested mainly by dizziness and dim eyesight. It is defined as a sensation of moving around in space or having objects move about the person as a result of a disturbance of equilibrium. Vertigo is sometimes used as a synonym for dizziness, feeling light headed, or giddy.

Patients with mild vertigo may experience relief from this condition after a short bed rest, with their eyes closed. In severe cases, patients may suffer from a serious rotary vertigo and will not be able to stand or walk straight (e.g., sea and air sickness), sometimes followed by nausea, vomiting, perspiration, or even syncope (fainting).

Diseases that have vertigo as their main manifestation include cerebral arteriosclerosis, auditory vertigo, hypertension, anemia, and psychosomatic illnesses such as hyperventilation (which often accompanies anxiety disorders, especially panic disorders). These diseases may be diagnosed, categorized, and treated with reference to the prescriptions described in this section.

ETIOLOGY

Several factors contribute to the vertigo syndrome. In ancient times, Chinese physicians believed that vertigo was caused by Yin Deficiency and Excess Liver Wind, or by a head trauma. In Excess conditions, vertigo is caused by Phlegm Fire or Wind Phlegm. Currently, the clinical protocol for treating vertigo is divided into three conditions:

1. Hyperactive Liver Yang causing an insufficiency of Kidney Yin,
2. A Deficiency of Qi and Blood in the Heart and Spleen causing a Deficiency in the Marrow reservoir, and
3. Stagnation of Phlegm Dampness caused by a Deficiency of Spleen and Stomach Qi (which fail to transport the Phlegm).

Vertigo may be caused by a variety of factors affecting the ear, the eighth cranial nerve, and the brain stem; including middle ear disease, infectious diseases, trauma, tumors, toxemia due to food poisoning, hypertension, sun stroke, and toxic conditions such as those caused by alcohol and drugs. Over medication and malnutrition (especially in the elderly who suffer from dehydration) also cause vertigo.

SYMPTOMS

1. In cases where the patient has hyperactive Liver Yang, the syndrome of vertigo is characterized by severe dizziness, splitting headache, blurred vision, nausea, soreness and weakness of the waist and legs, mental fatigue (caused by emotional upset and anger), flushed face, constipation and tinnitus.
2. In cases of Qi and Blood Deficiency, the symptoms include rotary vertigo, blurred vision or night blindness, listlessness, disinclination to talk, palpitations, insomnia and anorexia.
3. In cases of stagnation of Phlegm Dampness, the symptoms include chest pain and distress, nausea and vomiting, poor appetite, listlessness of the body and heaviness in the head.

TREATMENT

When treating patients with vertigo, begin with the upper body working downward and from the medial aspect of the body towards the lateral part of the body.

1. Begin with the patient sitting on the edge of the table. Stimulate the patient's Baihui, the sides of the head, (GB-4, GB-8, and Taiyang areas), GV-14, Shendao, Mingmen, and the Lower Dantian.
2. Next, use the Vibrating Palm hand manipulation to emit Qi into the patient's Baihui, GV-14, Yellow Court and Lower Dantian for 12 to 24 breaths.

3. Use the Extended Fan Palm hand technique to emit energy into the patient's St-8, St-21 and St-36 points on both sides of the body.
4. With the left palm, emit Qi into the patient's Baihui point while using the right hand Sword Fingers to guide the patient's Qi from the Yellow Court down the Conception Vessel into the Lower Dantian.
5. Next, direct the Qi from the patient's St-8 points down to the St-36 points via the Stomach Channels.
6. Purge the energy of the upper torso out through the hands via the Large Intestine Channels.
7. Finally, lightly pat and knead the patient's Baihui and GV-14, softly rocking their upper limbs to end the treatment. Softly pressing and kneading the patient's tissues along their channels can push Qi and Blood downward to free the patient's mind from Toxic Qi and can improve the eyesight. The soft grasping can relieve stasis of Qi and Blood in the patient's blood vessels and promote circulation, thus reducing and relieving vertigo.

TREATMENT MODIFICATIONS

1. In cases of hyperactive Liver Yang, use the Extended Fan Palm hand method to emit Qi into the patient's Bl-18, Mingmen, and Lower Dantian. Guide the Qi to flow along the channels (or in a clockwise direction) to nourish the patient's Yin and to suppress the hyperactive Yang.
2. In cases of Deficient Qi and Blood, use the Extended Fan Palm hand method to emit Qi into the patient's mid-back (Bl-17, Bl-18, and Bl-20 points) and Lower Dantian to nourish the patient's Qi and to replenish the Blood.
3. In cases of stagnation of Phlegm Dampness, use the Extended Fan Palm hand method to emit Qi into the patient's Yellow Court. Guide the Toxic Qi down the Stomach Channels and expel the pathogens out from the patient's St-36 points.

PRESCRIPTIONS AND HOMEWORK

To treat vertigo, have the patient practice the Descend the Qi and Cleanse the Organs exercise (see Chapter 41).

PRESCRIPTION MODIFICATIONS

When treating patient's who suffer from a Deficiency of Kidney Yin and a hyperactivity of Liver Yang, prescribe the following meditation. Have the patient take a standing or sitting posture, breathing naturally, and relaxing. While inhaling through the nose, the patient imagines blue energy filling the mouth. When exhaling, the patient imagines this dark blue energy slowly pouring down the throat, like warm water, filling the Kidneys and Lower Dantian area. The patient should practice this meditation for 7 breaths.

HEADACHES

A headache is defined as a diffused pain in different portions of the head which is not confined to any one nerve distribution area. It may be acute, chronic, frontal, temporal, occipital, confined to one side of the head, or affect the region immediately over one or both eyes. The character of the pain may vary from dull and aching, to acute stabbing pain that is almost unbearable. There may be intermittent, intense pain, a throbbing pain, a pressure in the head which feels as if it will burst the head, or a penetrating pain driving through the head.

ETIOLOGY

Transient, acute headaches may vary in their etiology. Disease of the perinasal sinuses, teeth, eyes, ears, nose or throat, acute infections, invasion of Wind Cold pathogens into the channels of the head, or trauma to the head are among the most frequent causes of headaches.

Chronic headaches may also be caused by a variety of conditions including physical, emotional, psychosomatic, or psychogenic factors, including fevers, metabolic conditions (hypertension, etc.), or even exposure to toxic chemicals.

SYNDROMES

Headache syndromes can be divided into five categories which are described as follows.

1. A Shaoyang headache (migraine) is clinically manifested through symptoms such as a severe headache (splitting pain) in the temple region accompanied by Excess Heat in the

head, conjunctive congestion, and sometimes intercostal fullness or pain. Nausea often accompanies a severe migraine attack.

2. A Taiyang headache (occipitocervical headache) is clinically manifested through such symptoms as a headache and pain in the back of the head which can be accompanied by fever, aversion to cold, and rigidity at the back of the neck.
3. A Jueyin headache (vertical ache) is clinically manifested through such symptoms as a vertex headache, vertigo, insomnia, dizziness, and irritability with a flushed face.
4. A Yangming headache (frontal ache) is clinically manifested through such symptoms as a headache in the upper part of the cranium accompanied by dysphoria (depression and anxiety) with a smothering sensation, thirst, bad breath and constipation.
5. A Deficient Qi and Blood headache is clinically manifested through such symptoms as a lingering headache with a heavy sensation in the head, blurred vision, lassitude, and weakness.

TREATMENT

Begin by purging and dredging the patient's Upper Dantian, focusing on the Yintang, Taiyang, Baihui, GV-16, GV-20, and GB-4 points. Purge the pathogenic Qi down the patient's arms and out the LI-11 and LI-4 points, as well as down the Gall Bladder Channel out of the GB-41 points. The point and channel purging is specific to the location (right or left side) of the headache.

Next, using the Extended Fan Palm hand technique, emit Qi into the patient's headache region, purging it down the Conception Vessel to the Lower Dantian.

TREATMENT MODIFICATIONS

1. When treating a patient with a Shaoyang headache, apply the Extended Fan Palm hand technique to emit Qi into the GB-4, GB-8, and TB-20 points, leading the pathogenic Qi out of the Triple Burners and Gall Bladder Channels.
2. When treating a patient with a Taiyang headache, apply the Extended Fan Palm hand technique to emit Qi into the LI-11 and Bl-10 points, Purging pathogenic Qi down the Large Intestine and Bladder Channels and out of the hands and feet.
3. When treating a patient with a Jueyin headache, start at the head using the Circle Spiraling hand technique in a counterclockwise direction, purging the Toxic Qi out the patient's Gall Bladder Channels and out of the legs.
4. When treating a patient with a Yangming headache, use the Extended Fan Palm hand technique to emit Qi into the patient's Taiyang, Yintang, and St-8 points, purging the Toxic Qi down the Stomach Channels and out of the legs.
5. When treating patients with Deficient Qi and Blood, use the Extended Fan Palm hand technique to emit Qi into the patient's mid-back, at the Bl-17 and Bl-20 points, and the chest at the CV-17 point. Then emit Qi into the patient's CV-12 and Lower Dantian areas.

PRESCRIPTIONS AND HOMEWORK

1. It is advisable to have the patient practice the Ten Dragons Run Through the Forest exercise found in the Wash the Face and Massage section of Chapter 39.
2. It is beneficial to prescribe the following meditation to prevent the advancement of an acute attack. The main focus of this meditation is to direct Blood, Heat, and Qi to the body's upper extremities (the arms and hands), away from the original source of pain (i.e., the head).
   a. From a sitting, semi-reclining, or lying posture, the patient begins to relax and quiet the mind, with the eyes closed.
   b. The patient imagines that it is summertime and he or she is at the beach. Next, the patient focuses on both arms and hands, imagining them growing warmer and warmer in the hot sun. The patient imagines feeling massive amounts of heat radiating from the arms and hands, redirecting the Blood flow away from the head.
   c. Then, the patient imagines the face becoming cooler, as the pressure and pain

melts down the neck, shoulders, arms, forearms and out of the hands.

Prescription Modifications

1. For patients with a Shaoyang headache, focus the prescription meditations and exercises on treatments that purge the bilateral sides of the patient's head along the Triple Burners and Gall Bladder Channels, dredging the pathogenic Qi down and out of the patient's arms and legs.
2. For patients with a Taiyang headache, focus the prescription meditations and exercises on treatments that purge the patient's occiput and cranium, the Small Intestine and Bladder Channels, and that dredge the pathogenic Qi down and out of the patient's arms and legs.
3. For patients with a Jueyin headache, focus the meditations and exercises on treatments that purge the patient's vertex cranium, dredging the Toxic Qi down the Liver Channels and out of the legs.
4. For patients with a Yangming headache, focus the prescription meditations and exercises on treatments that purge and dredge the energy from the patient's forehead down the Large Intestine and Stomach Channels and out of the legs.
5. For patients with a Deficient Qi and Blood headache, focus the prescription on Lower Dantian Cultivation exercises and meditations.

MIGRAINE HEADACHES

This type of headache initially consists of a constriction in the head arteries, followed by vasodilation of the extracerebral cranial arteries.

ETIOLOGY

Migraine headaches are usually brought about by stress, and can also be initiated by an allergic reaction to what the patient's body considers to be a toxic invasion (e.g., cheese, chocolate, red wine, contraceptive pills, etc.). These types of stresses on the body's system result in internal organ dysfunction, which is generally caused by one or more of the following three factors.

1. The energy of the Heart and Spleen are not regulated.
2. A rising up of Liver Heat and Fire occurs.
3. A Deficiency of Kidney Yin and rising of Heart Fire occurs simultaneously.

The result of stress-related internal organ dysfunctions causes the initial constriction of the arteries in the head, followed by vasodilatation and distention of the vessels.

SYMPTOMS

The clinical manifestations of migraine headaches are hypersensitivity to light and sound, followed by severe, throbbing pain (most commonly felt on only one side of the head), nausea, dizziness, and sometimes vomiting. Some migraine patients experience seeing an aura or visual disturbances (colors, shapes, lines, flashing spots, or temporary reduction in the field of vision) before the head pain begins.

TREATMENT

For treatment, the Qigong doctor must pull the pathogenic energy out of the patient's head, from the location of the pain; then lead the Excess Toxic Qi out of the patient's extremities, and open up the channels so that Qi can rise and fall to the patient's head unobstructed.

1. In a severe case, combine Medical Qigong therapy with Jing Point therapy.
   a. Begin the treatment by pulling out (with strong intention) the stagnant Qi trapped in the patient's head. First, purge the patient's Wei Qi field, then enter the patient's Jade Pillow at the base of the head. Next, emit Qi along the falx cerebri and circulate it using your intention to create a small Microcosmic Orbit (circling from the back of the head to the nose and back). This encourages the patient's stagnant Qi to move.
   b. Next, dredge around the area of the patient's eye (from the inner corner of the eye to the side of the eye), starting at the Bl-1 point to the Yin Tang point (Third Eye), following the sphenoid arch across to the temple at the Taiyang point, then to the back of the ear at the

TB-17 point. Next, lead the pathogenic Qi down the neck and pull out the stagnant Qi through the extremities (Pc-6 and LI-4 on the hand). If the patient is feeling nausea, purge the Toxic Qi from the Yellow Court and lead the Toxic Qi down and out the body through the St-36 point on the leg and Lv-3 point on the foot.

c. To remove sluggish Qi stagnation, squeeze the patient's trapezius muscles, draining the pathogenic Qi out of the axillary fold of the patient's arm. The patient should breathe into the painful areas, and imagine any remaining Toxic Qi descending down the arm into the fingertips when exhaling. The doctor should squeeze the patient's arm as the patient exhales, to support the descending action of the Toxic Qi.

d. The patient should be given homework, depending on his or her constitution.

2. In treating chronic migraine headaches, combine Medical Qigong therapy with An Mo therapy, working from the branch to the root.

   a. Begin the treatment by pulling out the stagnant Qi trapped in the patient's head. This is done by first purging the patient's Wei Qi field, then entering the patient's Jade Pillow (at the base of the head) to purge and remove Toxic Qi and stagnation. Then emit and guide the Qi with your intention, to flow along the patient's falx cerebri. This forms a small Microcosmic Orbit (circling from the back of the head to the nose and back) and encourages the patient's stagnant Qi to move.

   b. Extend Qi into the patient's Bl-1 point (on the inside of the eye).

   c. Next, focus energy into the patient's Yin Tang point and begin circulating the Qi along the orbicularis oculi muscle of whichever eye is in pain (along the eyebrows to the ear). Work from the inside of the eye, out along the eyebrows (Figure 48.1).

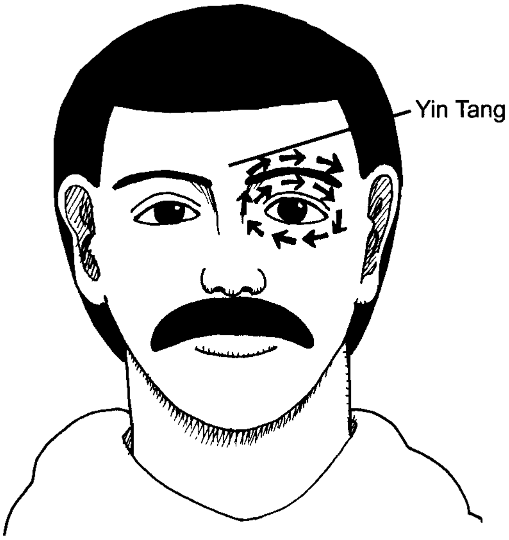

Figure 48.1. For treating chronic migraine conditions (which dominate behind the left eye), begin by extending Qi into the patient's Bl-1 point and start to dredge the Toxic Qi from the patient's Yin Tang area, circulating the orbicularis oculi muscle of the eye in a clockwise direction.

   d. After several rotations, follow the sphenoid arch to the back of the ear at the TB-17 point (if the pain is bilateral, for men treat the left side first; for women treat the right side first).

   e. Next, beginning at the TB-21 point, outline the outside of the ear, then massage the lateral aspect of the ear's ridge (upper auricle) (Figure 48.2).

   f. Hold both of the patient's ears firmly and begin to apply slight pressure at a 45 degree angle towards the sides of the patient's head. This allows the Qi trapped within the patient's head to escape. Using intention, move the Qi from the Jade Pillow (along the tentorium cerebelli) into the patient's pituitary gland. Then divide the energy into two rivers of Qi that circulate through the middle of the cerebrum and end back at the Jade Pillow.

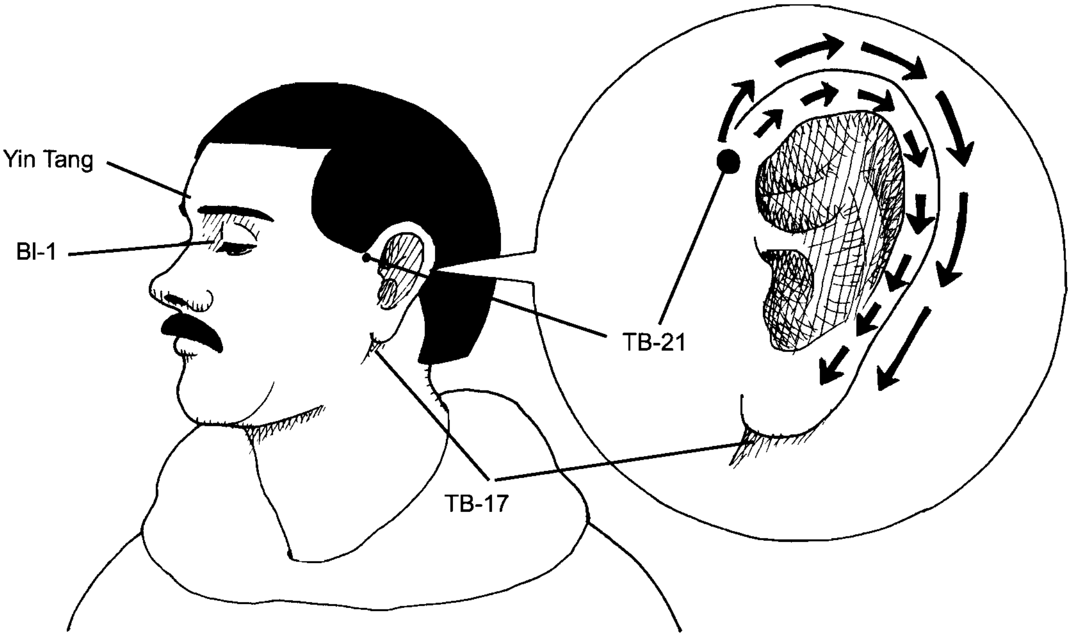

Figure 48.2. Outline the outside of the patient's ear, then massage the lateral aspect of the ear's ridge.

g. From the front of the head, at the edge of the hair, gather the Qi and lightly scratch the patient's head, as you use the Kneading Tiger technique. Visualize pulling the Toxic Qi down the front and sides of the patient's head (down the Gall Bladder Channels) to the base of his or her neck (Figure 48.3).

h. From the base of the neck, pull the Toxic Qi down and out of the arms.

i. The patient should be given seated Quiescent Qigong meditation homework, the length of meditation time will depend on his or her constitution.

Note: When treating migraine headaches, the head should always be treated first before treating the extremities.

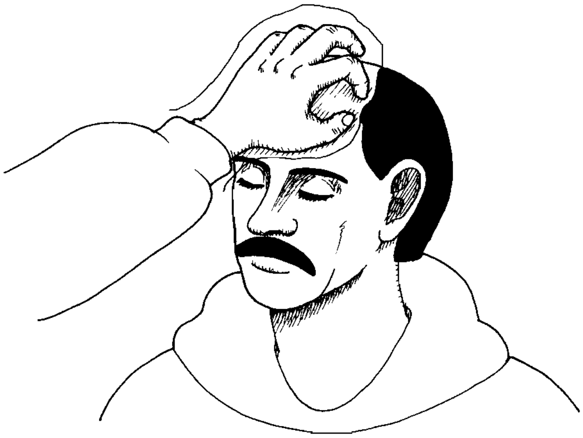

Figure 48.3. Using the Kneading Tiger technique, visualize pulling the pathogenic Qi out of the patient's head, down the side of the neck and down along the sides of his or her arm.

SECTION XI

SPECIALIZED QI EMISSION THERAPY FOR PEDIATRICS, GERIATRICS, GYNECOLOGY, NEUROLOGY, PSYCHOLOGY, ONCOLOGY, AND SURGERY

INTRODUCTION

In this section we will explore disease treatment techniques and prescriptions unique to Medical Qigong doctors who specialize in Pediatrics, Geriatrics, Gynecology, Neurology and Oncology.

The last two chapters will address Medical Qigong therapy and the treatment of radiation, chemotherapy, and surgery.

CHAPTER 49
QI EMISSION THERAPY FOR PEDIATRICS

INTRODUCTION

Pediatrics is the branch of medicine which deals with the care of children and the treatment of childhood diseases.

From a traditional Chinese medical perspective, children are different than adults in several important ways. First, the Yin and Yang energy in children is fresh and pure, and their vitality usually flourishes. Second, although the internal organs of children have begun to take shape, the organs are still in the process of developing Qi; the organs are still maturing and are not yet fully developed. Because children are still growing organs and tissues, stimulating the channels and channel points has a different effect on their body than on an adult.

Specifically, the Lungs, Spleen, and Kidneys are especially delicate in children. The Lungs and Spleen are weak, and the Kidneys are particularly vulnerable. The Lungs control the Qi of the whole body. The Spleen is responsible for the transformation of food and the transportation of Gu Qi (food Qi). The Lungs and Spleen are the postnatal foundation of health. The Kidneys are in charge of the bones and Marrow and are the prenatal foundation of health. Children are still full of Prenatal Qi; they are just beginning to use their Lungs and Spleen for the function of acquiring Postnatal Qi. Consequently, their energetic system is vulnerable and unstable.

If children are born without congenital defects or weakness, are nursed properly, and avoid contracting major diseases, the internal organs should develop normally. However, because of inherent weakness or immaturity of the digestive system, almost all pediatric diseases of children under the age of six usually begin with some element of indigestion. A child's tiny digestive system makes him or her prone to the creation of Phlegm. Since the Spleen is the root of Phlegm production, and the Lungs are the storehouse of Phlegm, then a congenitally weak Spleen can lead to the creation of Turbid Dampness within the child's body. This Turbid Dampness can congeal into Phlegm and lodge within the child's Lungs. These symptoms can be observed in young children as runny noses, ear infections, phlegmy coughs, and other respiratory difficulties.

EARLY STAGES OF CHILD GROWTH AND DEVELOPMENT

In childhood, growth and development occur rapidly. In the ancient Chinese medical classic entitled *"Prescriptions Worth a Thousand Gold Ducats"* it states, "A newborn is feeble and tender; two months later, its pupils begin to become well developed, it can smile and distinguish parents from strangers; 150 days later, its sacral bones have been well developed, it can sit independently; 210 days later, its metacarpal bones have been well developed, it can crawl; 300 days later, its patellae have been well developed, it can stand alone; at one full year of life, its shank and occipital bones have been well developed, it can walk and speak."

In the natural course of childhood growth and development, there may be fever, irregular pulse signs, or perspiration due to the imbalances of Yin and Yang. The child usually recovers spontaneously within a day or two. Such episodes are called Changing and Steaming. Changing implies the transformation of the Five Yin Organs, with the accompanying variations of emotions, and the development of the child's intelligence. Steaming implies the vaporous exchanges within the six Yang organs induced by accumulated Heat. The ancient Chinese discovered that there is one Changing every 32 days, and one Steaming every 64 days. Thus, within 320 days, there are 10

Changings, and within 576 days, there are 9 Steamings. Every Changing or Steaming is an ordinary event in the course of growth and development, and should not be mistaken for a condition of childhood disease. However, if children are not cared for properly during Changing and Steaming, they are more apt to fall ill.

PATHOLOGY

In "*Detailed Analysis of Epidemic Febrile Diseases*," Dr. Wu Tang stated that children's diseases are apt to change because the internal organs are frail and their defense mechanisms are incomplete; children contract infections easily because the skin and muscles are still tender, and the vitality is timid and weak. Propensity to illness, inclination to Excess and Deficiency, and being prone to suffer from Heat and Cold syndromes are common features of pediatric pathology. Since children are full of pure Yang, the course of disease is changeable and unpredictable. Nevertheless, recovery of health is generally easy.

DEFICIENT LUNGS

Children are apt to suffer from internal organ disorders, especially of the Lungs, manifested by symptoms of cough, dyspnea, retention of Phlegm and Dampness in the Lungs, seasonal febrile diseases, allergies, asthma, and other Lung disorders.

The Lung is a delicate organ in charge of the skin and hair. Therefore, particularly in children, junctures between the skin and muscles are not compact enough to resist the invasion of Exogenous Cold, Heat, Dampness, Wind, and other pathogens. Seasonal pathogens may attack the Lungs through these slackened junctures and through the nose, thereby interfering with the descending and dispersing functions of the Lungs. This results in the accumulation of Heat, fullness in the chest, productive cough, or dyspnea.

DEFICIENT SPLEEN

Since the Spleen is not yet matured and thus usually Deficient, it may be harmed or damaged due to improper diet, irregularity of meals, or imbalances caused by excessively Hot or Cold diets. Consequently, Clear Yang fails to rise, Turbid Yin fails to descend, and Stomach Qi becomes disharmonious. Therefore, children frequently suffer from abdominal masses, abdominal pain and distension (stomachache), belching, hiccups, vomiting and diarrhea, infantile malnutrition, and other digestive disorders. These conditions often lead to children becoming hypersensitive to the foods they consume.

According to Five Elements theory, the Spleen (Earth) generates the Lungs (Metal) and the Lungs (Metal) generate the Kidneys (Water). The elements work in coordination and influence each other. In children suffering from Spleen Deficiency, there will also be a Deficiency of Lung Qi; in children suffering from Lung Deficiency, there will also be a Kidney Deficiency.

DEFICIENT KIDNEYS

Kidney problems can manifest in two different ways: Kidney Jing Deficiency, which can cause retardation, or Congenital Qi Deficiency (i.e., Deficiency in the Sea of Marrow), which can lead to Down's Syndrome, Attention Deficit Disorder (ADD) and learning disabilities. Since the Kidneys control the lower orifices, Deficient Kidney Qi can also lead to chronic bed wetting.

If the child was improperly nursed when they were young, the Deficient Kidney Yin not only fails to nourish the bones and Marrow, but also fails to nourish tendons. During depletion of Liver and Kidney Yin, there can be atrophy, weakness of tendons, five types of developmental retardation, as well as five kinds of flaccidity.

LIVER YANG EXCESS

By contrast, Liver Yang is usually Excessive, and Liver Wind may be brought on by Heat or Fire, resulting in infantile convulsions, twisting, and even opisthotonos.

HEART (SHEN) IMBALANCE

The Heart is responsible for the child's mental stability. The Heart Fire and the Kidney Water must constantly be balanced to stabilize the child's Wu Jing Shen. An imbalance of the Heart can lead to Shen disturbances which can result in mental instability or disability.

According to the active energetic balance of the child's Wu Jing Shen, the Hun tend to naturally be overactive, and the Shen of the Heart is generally unstable. The reason for this instability is because the Hun are very active within a child. The Hun are responsible for creativity, enthusiasm, dreaming, going in and out of the body, and seeing into the Spirit world. The Shen, which is responsible for the child's discriminating mind, respect and conveying the appropriate behavior, is not fully developed. Consequently, young children tend to often be wild, hard to control, and get easily bored.

In treating a child's overactive Hun and unstable Shen, the Qigong doctor harmonizes the child's Heart Fire and Kidney Water energy, and opens the Heart's orifices (using Qi massage on the child's eyes, ears, nose, mouth and all upper sensory orifices) to stabilize the Wu Jing Shen.

HOT AND COLD EXOGENOUS FACTORS

Susceptibility to Exogenous Cold and Heat reflects the characteristics of puerile Yin and Yang. When children are invaded by Exogenous Evils, for example, there will be Excess Heat syndromes; when the child's healthy energies are depleted, there will be Deficiency Cold syndromes.

Although various defects and Deficiencies do exist in children, their Yang, Liver, and Heart are in a hyperfunctional state. Therefore, children are full of vitality and their anti-pathogenic responses are sensitive and immediate. The causes of childhood diseases are usually comparatively simple, relative to adults, and their convalescence is easier and quicker provided they receive appropriate treatment and care.

THERAPY

Because the internal organs of children are still forming and are, therefore, still frail and tender, children are much more sensitive than adults, and cannot stand strong energetic stimulation. This is why the primary modality for treating young children is gentle pediatric Tui Na, which is usually on the child's hands, feet or abdomen, or Qigong massage (lightly touching the child's body). Acupuncture is used only infrequently, and when it is, the needles are usually not retained for long - the points are quickly and lightly stimulated and then the needles are removed. Similarly, when using Medical Qigong emission therapy, a milder stimulation is applied than would be used for adults when purging, tonifying, and regulating.

Medical Qigong is useful for treating pediatric neurological disorders, and particularly for such conditions as epilepsy, hyperactivity, attention deficit disorder (ADD), and mental retardation. Effective points for tonification include Tianzhu (Bl-10), Sishencong (four extra points, located around the Baihui area at the top of the head), Baihui (GV-20), Yamen (GV-15) and Fengfu (GV-16).

When children under the age of six practice self-regulation Medical Qigong exercises (such as Qigong to increase intelligence), they frequently develop special abilities. These abilities include ESP (e.g., telepathy and telekinesis), and special Qigong healing abilities.

ENERGETIC MOVEMENT AND FLOW

The child's Microcosmic Orbit naturally flows in the direction of the Water cycle, moving up the front (CV) and down the back (GV) in order to facilitate the child's spiritual, intuitive, and psychic perceptions. This energetic movement continues in this direction until the child reaches puberty, at which time the energy reverses its direction and flows up the back (GV) and down the front (CV), in the direction of the Fire cycle. This energetic switch in direction facilitates the child's cognitive development and the ability to control emotions and impulses. The time of the Microcosmic Orbit's energetic reversal varies depending on the child's physical constitution, state of health, and his or her environment.

MEDICAL QIGONG EXERCISE PRESCRIPTIONS FOR CHILDREN

The following Medical Qigong exercise is considered a basic training method for children. This exercise is specifically developed for children of all ages and physical conditions, and is used to improve a child's physical and mental health, as well as aid in curing some common childhood diseases. This exercise can also help to cure food allergies, loss of appetite, malnourishment, a weak immune system, neurasthenia, insomnia, dreaminess, and disturbances of the Spleen and Stomach. It is believed that this exercise also helps moderate a child's stubbornness and impatience.

IMPROVING INTELLECTUAL FACILITIES QIGONG (DA-ZHI-GONG)

This particular Medical Qigong exercise is excellent for children who range from 3 to 12 years old. The time and frequency of the following Qigong exercise is not fixed. It is important, however, that the child be consistent with the practice, and that the body and mind remain relaxed while training.

It is important that the child practice the entire Qigong sequence in its proper order. The child is not allowed to practice only a part of the exercise.

Prohibited from this particular exercise are children who have weak constitutions (i.e., have difficulty in standing up) or suffer from serious diseases.

1. Begin by having the child assume a standing Wuji posture. Have him or her relax the body, with the eyes either open or closed. The child then raises both hands in front of the Upper Dantian. Boys will place the left hand facing the forehead, about three inches from the Yintang point. Girls will place the right hand facing the forehead, about three inches from the Yintang point. At the same time the opposite hand should be level with the other palm and should be facing outwards (Figure 49.1).
2. Both hands begin to make a clockwise circle rotation (from the child's point of view). The child keeps the Yin Tang point as the center of the circle, for 18 breaths. Next, the child moves both hands to the front of the right eye and continues the same clockwise movement while keeping the right eye at the center of the circle. Rotate for 18 breaths. Then, the child moves both hands to the front of the left eye and continue the clockwise movement for 18 breaths. Finally, the child returns to the front of the forehead and continues the rotation, keeping the Yintang point as the center for 18 breaths, then

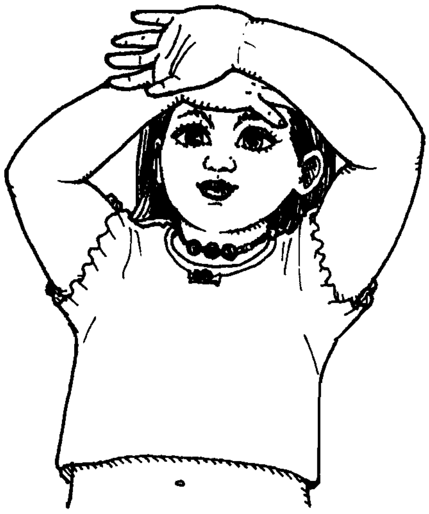

Figure 49.1. Improving Intellectual Facilities (1)

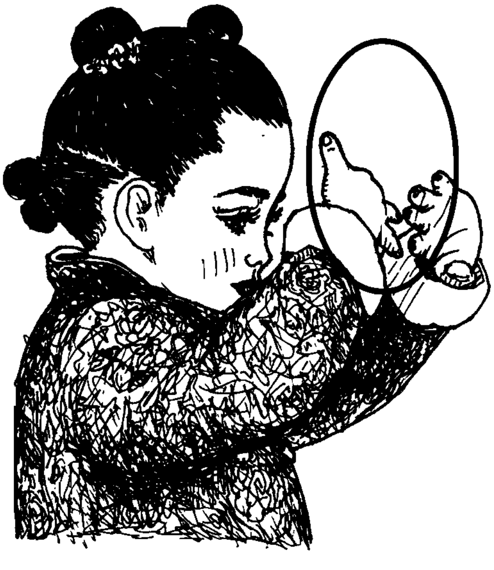

Figure 49.2. Improving Intellectual Facilities (2)

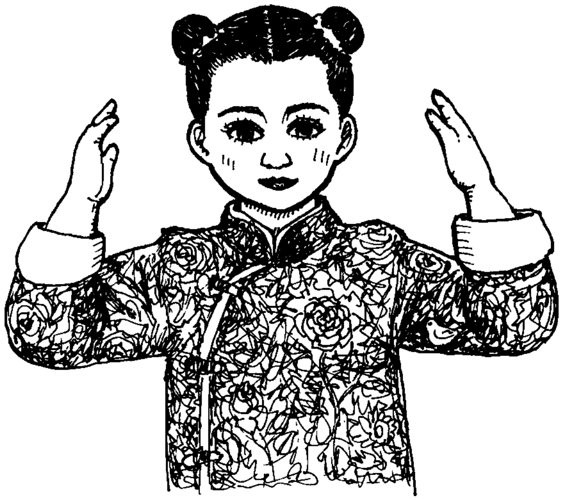

Figure 49.3. Improving Intellectual Facilities (3)

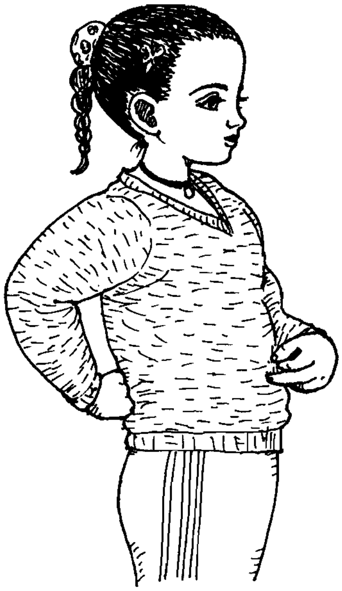

Figure 49.4. Improving Intellectual Facilities (4)

ends the meditation (Figure 49.2).

3. The child places both hands in front of the ears, with the center of the palms facing the external opening of the ear canals; breathing naturally, the child moves both hands forward and backwards while emitting Qi into the ears. Make sure the child does not concentrate on any specific thing, but simply relaxes (Figure 49.3).
4. Next, the middle fingers softly touch the navel with the left hand, and the right hand softly touches the Mingmen. Have the child imagine that there is a big red ball of fire in his or her abdomen (where the child's two middle fingers are pointing). The child's fingers should remain is this position until warmth is felt in the abdomen. If the heat becomes too hot, have the child exhale the sound "Ha" three times to release the Excess Heat (Figure 49.4).
5. To end the exercise, have the child cover his or her tailbone with both hands and jump up and down gently for several minutes. Do not allow the child to become tired from jumping. Again, make sure that the child does not concentrate on any specific thing (Figure 49.5).

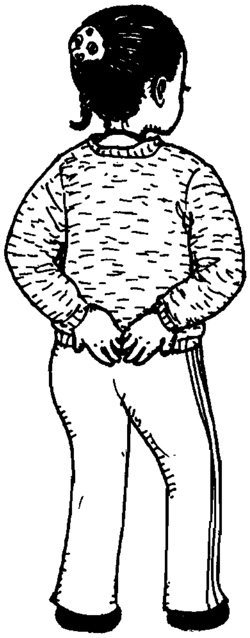

Figure 49.5. Improving Intellectual Facilities (5)

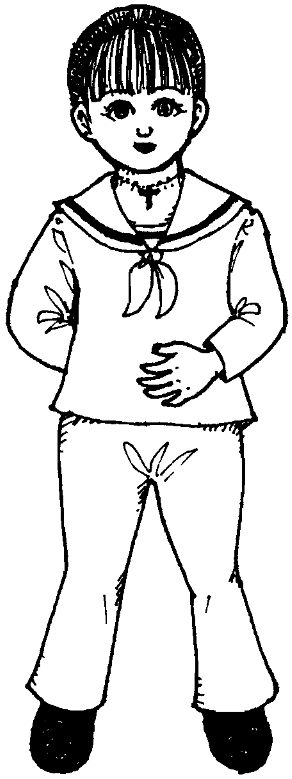

Figure 49.6. Gathering Wisdom Qigong (1)

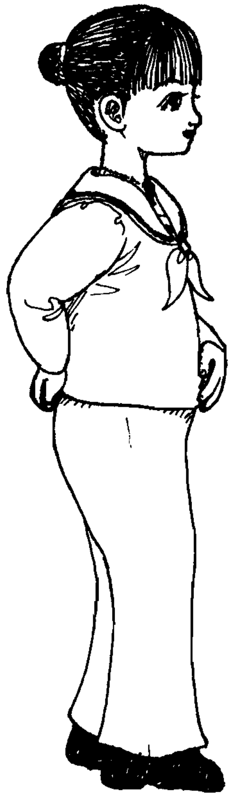

Figure 49.7. Gathering Wisdom Qigong (2)

MEDICAL QIGONG FOR GAINING WISDOM QIGONG (TONG-LING-GONG)

This particular Medical Qigong exercise is excellent for children who range from 3 to 16 years of age. The time and frequency of this particular Qigong exercise is not fixed, and may be decided by the child. Also, there are no contraindications for this particular Qigong exercise.

1. The child begins by standing with his or her feet flat, shoulders width apart. The child should place one hand on the Lower Dantian (below the navel) and the other hand at the back, below the Mingmen (opposite the Lower Dantian). Have the child close his or her eyes and relax the body and mind for about three minutes. The child is to imagine that the upper part of the body (above the waist) is extended into the blue sky above, and that below it (around waist level) are clouds (Figure 49.6).
2. Next, the child imagines that he or she has changed into an Immortal (a Super-Hero). Encourage the child's imagination to flow freely with this concept, allowing the child to move about, gently shaking the knees (Figure 49.7). After several minutes end the meditation.

CHAPTER 50
QI EMISSION THERAPY FOR GERIATRICS

Geriatrics is a branch of medicine which deals with the problems of the aging. Geriatric specialists addresses the physiological, pathological, psychological, economic, and sociological problems of the elderly. The importance of geriatrics is emphasized by the fact that the expected life-span of the average person is increasing.

Diseases of the internal organs can spread throughout the body via the channels and collaterals, and vice versa, as channel diseases can be transmitted to the internal organs. If Qi and Blood lose their harmony due to old age, and the patient's internal organs, channels, and collaterals become obstructed, then Phlegm and Stasis will develop internally. The accumulation of Phlegm, Qi and Blood stasis may lead to a chain reaction of pathological transformations, such as Hot, Cold, Excess or Deficiency Syndromes in the patient's internal organs. The following are examples of some of the pathological transformations.

1. If the patient's Qi and Blood in the Liver lose their harmony, the patient may experience fatigue and depression with dry eyes or diminished vision. The patient may also experience numbness in the extremities and muscle spasms due to a weakening of the sinews and blood vessels.
2. If the patient's Qi and Blood in the Heart lose their harmony, the patient may experience poor memory, insomnia, heart palpitations, and shortness of breath (which may be aggravated by movement).
3. If the patient's Qi and Blood in the Spleen lose their harmony, the patient may experience fatigue and general weakness in all four limbs. The patient may also experience loose or bloody stools, and there will be abdominal distention after food consumption.
4. If the patient's Qi and Blood in the Lungs lose their harmony, the patient may experience coughing with Phlegm and panting, as well as spontaneous sweating.
5. If the patient's Qi and Blood in the Kidneys lose their harmony, the patient may experience poor memory, anxiety, tinnitus or deafness, as well as a sore or weak lower back, knees, and teeth.

The Chinese have a saying, "When the Heaven (Heart Qi) connects with the Earth (Kidney Qi), all seven emotions are kept in moderation." When Qi becomes scattered the Heart Fire and the Kidney Water cannot converge. This results in an unbalanced and unstable energy flow.

The Kidneys are extremely important for health maintenance and healing in all patients, and especially in senior citizens. As one gets older, the Kidneys, lower back, and legs are the first things to be affected.

It is important to stabilize the Heart because the Heart is responsible for mental and emotional orientation. Traditional Chinese Medicine holds that, "when the Heart is moved, all the other organs will be shaken."

In teaching Medical Qigong to senior citizens, first quiet the Heart, so that the other organs can begin to relax. When the mind becomes disturbed, the body becomes disturbed. To affect one is to affect the other.

The eyes are considered the seedlings of the Heart as they receive information, gathering it and feeding it to the brain. It is important for the Heart's Qi to descend into the Lower Dantian to calm the mind. When the patient is in a hurry, the Heart is in a hurry, thus inducing the Qi to rise to the head. This results in nervous tension and anxiety. It is only through rest and quiet training that the mind and the Heart can begin to settle.

One method of quieting the mind is through prayer and meditation. This is considered "active rest." Sleep is considered "passive rest." When focusing inward and not using the eyes to see or the ears to hear, the Qi begins to converge in the middle of the body and the energy is drawn back into the organs, settling down into the Lower Dantian. This convergence of Qi nourishes and cultivates the natural energetic structure of the internal organs. Senior citizens are encouraged to sit for longer periods of energy cultivation. If they practice Taijiquan they are to avoid low postures and big motions. The arm movements should be kept small, soft, and slow.

TREATING THE PATIENT'S EMOTIONAL STABILITY

When treating the elderly, the Qigong doctor bases the regulation therapy prescriptions on two major priorities. The first is to achieve emotional stability and the second is to improve the Qi and Blood circulation.

The first priority in working with healing prescriptions is to focus on the patient's psychological profile. When teaching Medical Qigong to senior citizens, first teach them how to quiet the Heart. Once the Heart is calm the other internal organs begin to relax. Always keep in mind that when the mind becomes disturbed, the body becomes disturbed. To affect one is to affect the other.

Depression and sadness can have a profound effect on the patient's healing ability. Medical Qigong regulation can be a strong aid in releasing an elderly patient from a chronic, detrimental attitude that complicates the healing cycle. Take for example the three stages of emotional transition the patient endures before his or her Spirit closes (see Chapter 18).

If anger and grief are not allowed to be expressed by the patient (which is the first stage of energetic expression and release of emotional pain), then anger transforms into depression, and grief transforms into despair. This is the second stage of the energetic/emotional transformation due to emotional pain. If the patient's depression and despair are not alleviated, then indifference, the final stage of energetic/emotional transformation, is developed. The patient becomes withdrawn, apathetic and either antisocial, or obsessive/compulsive in his or her thinking and behavior.

When the patient's Spirit closes, they tend to be extremely difficult to work with, and can easily sabotage their own healing potential. Though the use of Soul Retrieval meditations (see Chapter 19), Medical Qigong prescriptions can be used to reconnect the patient with their disassociated emotions. This allows the patient the ability to feel and regain contact with his or her true feelings.

Pleasant emotions bring about a calm and relaxing physiology, while strong, chronic negative emotions cause Qi and Blood stagnation, as well as Excess conditions within the patient's body, effecting the quality of the body's life force energy.

ENERGIZING THE PATIENT'S QI CIRCULATION

The second priority is to get the patient to move and circulate the Qi and Blood deep and superficially within his or her body. Static Blood is considered the primary mechanism of senility. All growth, development, decline, and debility of the human body, is related to the condition of the patient's Qi and Blood.

Pathophysiologically, both Qi and Blood affect each other; if one becomes diseased, the other is affected. Due to the Kidney's influence, as seniors get older, their hearing, hair, memory, lower back, and legs become adversely affected. Traditionally, all senility-combating formulas of Medical Qigong stress treating the patient's Spleen and Kidney Deficiency by supplementing the patient's Yin and/or Yang. Then, by applying the balancing methods of activating the patient's Blood and rectifying the Qi, the doctor can also combat the patient's senility.

Senility is also frequently caused by dehydration and malnutrition. Both conditions are frequently overlooked by Western medicine. The dehydration is usually due to a breakdown of the body's thirst signal. The patient does not feel thirsty, and thus forgets to drink. Malnutrition may be caused by a lack of enzymes that aid in absorbing nutrition from foods. Overmedication likewise can cause senility. These issues should always be addressed when treating the elderly, along with Medical Qigong exercises and meditations.

One example of a Medical Qigong exercise for aiding Qi and Blood circulation for geriatric patients is the Microcosmic Orbit (Fire Cycle) which can be practiced in a sitting position.

The elderly are more subject to musculoskeletal disorders and visceral diseases, especially in the Kidneys and Liver. As the Kidney Yang rises, it begins to restrict the heat in the upper portion of the body, thereby affecting the Heart. Also, as the patient gets older the Lungs cannot sustain liquid retention; this affects the Wei Qi, and results in a thinner layer of skin covering the body.

The Qigong movements prescribed for the elderly should be kept slow and smooth. Qi regulation can be cultivated safely by focusing the patient's intention on the center of the palms while keeping the mind's intention moving downward. The patient must not exert strain when using the muscles. When the patient turns his or her body, make sure that the center of each palm (Pc-8) faces the other. Slow and even breathing will calm the patient's emotions and sedate the mind.

It is important to begin Medical Qigong regulation training for senior citizens with natural breathing. Natural breathing allows their respiration to become soft, natural, gentle and quiet. The focus of the mind should be on relaxation. In dealing with senior citizens, if the focus of their mind's intention is allowed to become too concentrated, it can cause the Qi in their body to constrict and develop into Qi or Blood stagnation. Redirect their focus of intention on their breathing by practicing the Small Heavenly Cycle meditation. The Small Heavenly Cycle meditation is used to regulate the body's Yin and Yang energy and to strengthen the patient's mind and spirit. The Large Heavenly Cycle meditation can be prescribed to increase the patient's Qi and Blood circulation, to rejuvenate the body and the sense organs, to clear the complexion, and to improve mobility.

TAPPING THE KNEES TO PREVENT SENILITY

One exercise prescription that the doctor can use to assist the patient in combating senility is to have the patient lightly tap above, below, and around his or her knees. This exercise stimulates the Kidneys' energy and can also be used as an important adjunct to maintaining health for senior citizens, especially when combined with the foot tapping (Kd-1) Kidney Tonification exercise (see Chapter 39). Combined, both of these exercises can help to strengthen the patient's Kidneys, lower back, and legs. When combining both exercise prescriptions, the Knee Tapping exercise should follow after the Kidney Tonification exercises.

To begin the Knee Tapping exercise, have the patient sit in a chair, with both knees lower then the thighs. Encourage the patient to direct the focus of his or her attention in and around the knee area. As the patient taps or softly slaps the tissue surrounding the knees, he or she should imagine the energy and vibration flowing into the Lower Dantian. After several minutes (about 36 breaths) the patient can stop.

CULTIVATING THE PATIENT'S YUAN QI BY ABSORBING ENERGY FROM NATURE

Another priority of focus for older patients is the gathering, storing and cultivation of his or her Yuan Qi. In China, in order to tonify the Yuan Qi, elderly patients practice their Qi Tonification and Regulation Exercises facing a cypress tree daily (a cypress tree is considered very powerful and full of energy). If a cypress tree is not available, then the patient can use a pine or oak tree. This exercise prescription is considered a slow and gentle Dynamic cultivation. Any Dynamic exercise for senior citizens should always be practiced slowly, and gently, in order to replenish the Yuan Qi during the Static Qigong posture (in this case, standing).

The flexion and extension of the muscles relates to the interaction of Yin and Yang within the body's energetic tissues and muscles. This flexion and extension of the muscles becomes the foundation of energy balance. Relaxation of the body results in soft, gentle movement, peacefulness and tranquillity. Muscular rigidity, on the other hand, results in hardness, restlessness and disease.

In order to practice Medical Qigong exercises, senior citizens are encouraged to keep their body's energy connected with the energy of the Earth. This connection is obtained via their energetic attachment through the bottom of their feet and the center of their palms. If the patients are confined to wheelchairs, or are bedridden, they can still focus the mind's intention on the Lower Dantian, and then extend their energy deep into the Earth.

As with all Medical Qigong prescriptions, the patients must first purge and cleanse their body before gathering, collecting, tonifying and regulating the body's energy. To purge the body of Toxic Qi, the patients imagine divine healing energy pouring down from the Heavens, filling the entire body, as they inhale. This vibrant healing energy is absorbed into their tissues and the Toxic Qi stored within the patient's body is dispersed into the Earth.

As the patients exhale, they imagine the Toxic Qi melting out their tissues and flowing downward, deep into the ground. The patients perform this purging and dispersing sequence for several breaths until they feel cleansed. Next, the patients begin to fill and regulate their body.

To fill and regulate the body, the patients imagine circulating the energy downward, deep into the Earth, and into the root system of a tree.

Next, the patients imagine the Earth's energy ascending up through the top of the tree, blending with the energy of Heaven, and then descending through their head, filling their entire body from the feet upward like water being poured into a glass and filling it up (Figure 50.1).

The inhalation and exhalation should follow the Natural Abdominal Breathing method, and the mind's intention should be focused on filling, gathering and circulating the universal and environmental Qi.

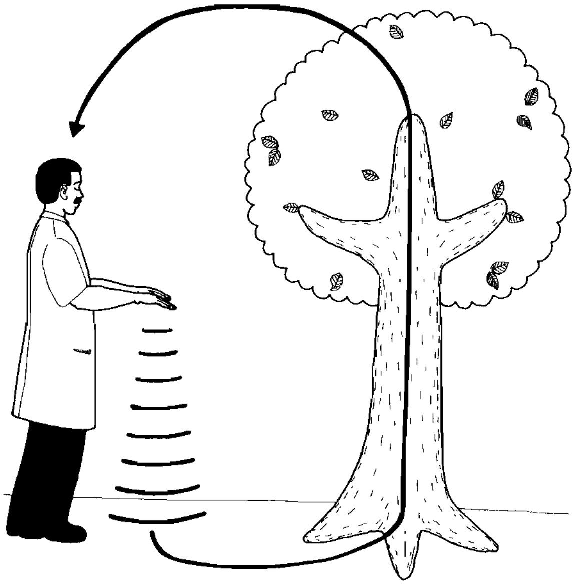

Figure 50.1. Extend Qi and Intention deep into the root system of the tree and absorb its natural energy through the Baihui point at the top of the head.

FURTHER OBSERVATIONS AND INSIGHTS ON GERIATRIC PATIENTS

According to Mark Johnson (an expert who has been treating "Seniors" with Medical Qigong Therapy and Taijiquan since 1987), any encouragement to undertake new experiences is very cathartic for older patients. Getting involved in anything creative seems to prolong the quality of their life.

Understanding the physical and emotional transitions that geriatric patients undergo in their every day life helps the Qigong doctor establish a better awareness of the patient's energetic dynamics. These physical and emotional transitions can best be understood when realizing the internal and external conflicts senior patients must constantly address.

COMPANIONSHIP

Many people tend to lose meaning in their life when they retire, lose a spouse, witness the death of countless friends, or when their children move away. Patients who lose their life purpose (or their meaning of life) usually do not live long. Elderly patients, who have lost their friends, usually have an extremely strong emotional attachment to any pet they may own. If the pet dies, this is usually as devastating as the loss of a close family member.

According to clinical observation, most older patients who become involved in spirituality, social service or voluntary work tend to be the most fulfilled and therefore most healthy. Humor and laughter as a means of internal emotional medicine for older patients cannot be overemphasized. Socializing with an understanding yet active support group also tends to bring healing to their heart.

In addressing the issue of being lonely, it has been found that the healing quality of having a pet to love can be very helpful for seniors who have become solitary or isolated.

OVERMEDICATING

Too many older patients are overly medicated. Taking up to ten pills a day is not unusual for many seniors. In a book initiated by Ralph Nader, titled "Worst Pills, Best Pills" the author emphasizes the fact that if an older person is taking over three different drugs, they are taking too many and should consult their doctor with the expressed purpose of cutting down to a maximum of only three types of medication.

The drugs given to seniors are usually tested on 30 year old individuals who are generally much stronger than most 70 year olds. Therefore half dosages are usually considered safer. In addition, there is a minimum of thirty-three drugs on the market today that can give the symptoms of Parkinson's disease. Several healers have gotten their patients off Parkinson's medication after demonstrating the fact that one of their other medications initiated the patient's symptoms.

SEDENTARY LIFE-STYLE

Too many older patients are sedentary, especially men. It is generally agreed that people living today move only 35% as much as people living 100 years ago. This has got to have a major impact on one's health, especially on older people whose circulation is not what it once was. Our bodies are designed to move; to become sedentary violates one of our basic needs. The studies conducted at Emory College confirm that the safest and most beneficial exercises for seniors is Medical Qigong, Taijiquan, walking, mild aerobics and light weight training. Seniors are encouraged to stay away from running.

IMPROPER DIET

Because of the patient's age and changing metabolic needs, diet is as equally important for seniors as exercise. Proper nutrition can be used to augment the benefits from the patient's prescriptions, as well as to increase the patient's healing potential. The diet in senior housing communities is usually atrocious. Sugar, canned vegetables and too many heavy meats are the norm. When prescribing Medical Qigong exercises for seniors, the Qigong doctor must also take into consideration the patient's diet and fluid intake, making sure that it is not sabotaging the healing effect of the Qigong prescriptions.

SLEEP DEPRIVATION

As patients gets older, their Kidney Jing becomes weak, their hair begins to turn grey, the hearing and sight becomes affected, the bone structure, mental comprehension and sleeping patterns also undergo changes. These physical transitions can and do affect the patient's emotional outlook on life. This however, is all part of the natural course of ageing.

Due to the weakening of the patient's Kidneys, sleep deprivation is more common in seniors than most people realized. More training in dealing with insomnia and changing sleeping patterns is necessary for Western doctors, as the serious impact on health by sleep deprivation is poorly understood in Western society.

Sleep is needed to invigorate and stabilize the patient's Wu Jing Shen (Five Essence Spirits). Without sleep, the patient's Hun (Spirit Soul), Shen (Spirit), Yi (Intention), Zhi (Willpower) become depleted and they lose control of their Po (Corporeal Soul). This affects the patient's intuitive perceptions, mental comprehension, drive and willpower, and can result in dementia.

UNDERSTANDING DEATH AND DYING

Classes and workshops on death and dying should be more available, and both patients as well as doctors should be encouraged to attend them. The doctor can benefit greatly by learning specific meditations being taught to facilitate this major transition, and also how to relate better to their elderly and young dying patients. The patients can benefit greatly by increasing their awareness and understanding of this process, as well as learning the numerous meditations that help resolve unprocessed emotions. Furthermore, patients can meet others with whom they can share their fears and hopes openly.

Too often the elderly do not wish to share these feelings, even with those closest to them for fear of burdening their family and friends. Patients may also be very reluctant to disclose their personal concerns for fear of being alienated, even with Qigong doctors, especially if the doctor's fear of death becomes evident through too much encouragement to get well and not lose hope.

CHAPTER 51
QI EMISSION THERAPY FOR GYNECOLOGY

INTRODUCTION

Gynecology is the study of diseases of the female reproductive organs. The earliest records of Chinese gynecological medical writings date as far back as the Shang Dynasty (1500–1000 BC). From a Traditional Chinese Medical perspective, a woman's physiology is characterized and diagnosed in accordance with her menstrual flow and vaginal discharge (also included are the energetic formations and transitions during pregnancy and childbirth). A woman's physiology is dominated by, and rooted in Blood. Her Lower Burner houses the uterus, which stores Blood, and corresponds to her Lower Dantian.

MENSTRUAL FLOW

Normally, a woman will follow the cycles of the moon and will menstruate every 28 days. Her cycles can be categorized into either new moon or full moon periods. The length of days between menstruation and the time of duration vary; four days of flow is considered average.

A woman's menstrual cycle can be categorized into four distinct time periods: Irregular Menstruation, Early Menstruation, Late Menstruation, and Prolonged Menstruation.

1. An irregular menstrual cycle is never consistent: sometimes early, sometimes late. The cause of an irregular menstrual cycle can be due to either a stagnation of Liver Qi (from suppressed anger or emotional shock), Liver Qi invading the Spleen (emotions affecting the Liver invades the Spleen), or a constitutional weakness of the Kidneys.
2. The cause of an early menstrual cycle can be due to either Heat in the patient's Blood or Qi Deficiency.
3. The cause of a late menstrual cycle can be due to either Cold stagnation of Blood, Blood Deficiency, stagnation of Qi (caused from emotions), or Deficiency of Qi.
4. The cause of a prolonged menstrual cycle can be due to either Qi Deficiency or Heat in the Blood (this is because Blood that is Hot wanders out of the vessels).

QI AND BLOOD REGULATION

Due to periodical menstruation and the complex changes that occur in the uterus and the organs (ovaries, cervix, and vagina) associated with reproduction, women tend to become especially vulnerable to Blood damage. The formation, circulation, and control of Blood depends upon the transformation, regulation, and production of the woman's Qi. This interdependence of Qi and Blood is regulated through the harmony of the body's Five Yin Organs. For example:

1. The Liver stores the Blood.
2. The Heart controls the Blood.
3. The Spleen contains or keeps the Blood in its vessels. In addition, the body's Spleen and Stomach are also known as the source of reproduction and transformation of Qi and Blood.
4. The Lungs store the Qi, which commands or moves the Blood.
5. The Kidneys store the Jing, which transforms or creates the Blood.

The woman's Sea of Blood can only flow freely when the Qi and Blood of the Five Yin Organs are in harmony. When a woman's Sea of Blood is full and exuberant, then her menstruation will be healthy and normal.

Regarding the Eight Extraordinary Vessels, a woman's Governing, Conception and Thrusting Vessels all originate within her uterus. The Governing Vessel governs the body's Sea of Yang and represents the Yang aspect of a woman's reproductive functions. The Conception Vessel governs the body's Sea of Yin and provides the Yin substances for all physiological and hormonal transitions, including puberty, conception, pregnancy, childbirth, and menopause. The Thrusting Vessels function as the Sea of the Twelve Primary Channels, and influence the supply and proper movement of Blood in the uterus, and control all aspects of menstruation.

The Belt Vessel encircles the Governing, Conception, and Thrusting Vessels, and governs, guides, and supports the energetic flow of Qi and Jing to the uterus.

Regarding the Twelve Primary Channels, their internal network is distributed throughout the uterus and is connected to both the Heart and Kidneys. In addition, the energetic connection with the uterus is further strengthened by the convergence of the Liver, Spleen, and Kidney Channels, along with the Conception Vessel, through the connecting vessels in the genitalia.

ETIOLOGY AND PATHOLOGY OF GYNECOLOGICAL DISEASES

In order for the Qigong doctor to make a correct diagnosis pertaining to a woman's menstrual flow and reproductive organ function, attention must be placed on identifying the primary etiology of the imbalance in the patient's Qi and Blood.

The Governing, Conception, Thrusting, and Belt Vessels are the most important vessels in terms of diagnosing gynecological pathophysiology. Consequently, any damage or disease of the these vessels are believed to be the main causes of gynecological disorders.

Either a direct or indirect attack to the Conception and Thrusting Vessels can cause disharmony of Qi and Blood, resulting in a loss of internal organ regulation.

DIRECT CAUSE

A direct cause of damage to the Conception and Thrusting Vessels can result from either Evil Toxin infections or from hygienically unclean sexual encounters. If the Conception and Thrusting Vessels become damaged, the result can manifest in symptoms such as profuse abnormal vaginal discharge or uterine bleeding.

INDIRECT CAUSE

Indirect causes of damage to the Conception and Thrusting Vessels can result from internal damage due to: The Seven Emotional Factors, Invasion of the Six Exogenous Factors, and An Inconsistent and Undisciplined Diet.

1. The Seven Emotional Factors tend to be the major factors in determining the root causes in most gynecological diseases.

   The suppression of anger, rage, resentment, and jealousy often results in Liver Qi depression, which eventually leads to Blood stagnation. This can result in such symptoms as delayed, painful, or blocked menstruation. If a hyperactive Liver Yang condition occurs, the patient can develop symptoms as serious as pre-eclampsia or eclampsia (coma and convulsive seizures that can occur between the 20th week of pregnancy and first week postpartum; symptoms include edema of the legs and feet, puffiness of the face, hypertension, severe headaches, dizziness, epigastric pain, nausea, sudden convulsive seizures and coma). The Liver's Blood Heat can lead to profuse uterine bleeding, leaking of Blood, hemoptysis (throwing up Blood) or epistaxis (bleeding of the nose).

   The suppression of worry and regret, or obsessiveness, can harm the Heart and Spleen. This can lead to amenorrhea (the absence or suppression of menstruation) or a fetal leakage (after conception, if a small amount of bloody fluid discharges from a woman's vagina, it is known as Tai Lou or fetal leakage).

   The suppression of fright, fear, loneliness, and insecurity can damage the Kidneys and lead to miscarriages.

2. Invasion of the Six Exogenous Factors, in particular, Cold, Heat, and Dampness, are the primary causes for gynecological diseases.

   If the Blood is invaded by Cold, it congeals, causing delayed menstruation, painful menstruation, or severe contractions and thick conglomerations of Blood.

   If the Blood is invaded by Heat, it moves, causing the Blood to flow recklessly. This reckless movement of Blood leads to early menstruation, excessive menstruation, and uterine bleeding.

If the Blood is invaded by Dampness (Damp Evils invading the Spleen), it can lead to abnormal vaginal discharge, uterine bleeding, and irregular menstruation.

3. An Inconsistent and Undisciplined Diet can lead to a Spleen and Stomach disharmony. If the Spleen and Stomach's ability to transform and transport food becomes hindered or weakened, the Sea of Blood becomes Empty. This can lead to delayed menstruation or scanty menstruation.

   If the Spleen loses its ability to restrain and contain the Blood, both Blood and Qi leak out of the vessels. This leaking of Qi and Blood can lead to excessive menstruation and uterine bleeding.

   If the Spleen's Yang does not perform its transporting function, Dampness and Turbidity will descend, resulting in abnormal vaginal discharge.

VAGINAL DISCHARGE

In Traditional Chinese Medicine, abnormal vaginal discharge (Dai Xia) is considered a disease of the Belt Vessel. If the Belt Vessel loses its restraint on the other vessels, then abnormal vaginal discharge is produced. This is usually caused by a Spleen Deficiency, Dampness, and descending Turbid Qi.

CONSISTENCY

The consistency of the patient's vaginal discharge can be divided into three categories: Turbid, Watery, and Clotting (thick and pasty).

1. A Turbid discharge refers to Blood Heat and an infection in the uterus. It also refers to signs of Cold stagnation.
2. A Watery discharge is considered a Cold Damp and / or Deficient condition.
3. Clots (thick and pasty) are considered a Damp Heat and Excess condition.

SMELL

The smell of the patient's vaginal discharge can be divided into conditions of either Damp Heat, indicated by a leathery smell, or Cold Dampness, indicated by a fishy smell. Also, if the patient's menstrual Blood has a strong, foul smell, it usually indicates a Heat condition.

COLOR

Vaginal discharge color is divided into five colors (Wu Se Dai): white, yellow, red, green-blue, and dark brown or black. The pathology of the discharge color is mostly caused from a Deficient Conception Vessel, loss of restraint from the Belt Vessel, Spleen Dampness, or Evil Wind invading the uterine vessels. These conditions create Heat, which descends into the uterine area causing disease.

Only profuse vaginal discharge, which bears an unnatural color and gives off a malignant odor, is considered pathological. The five colors of vaginal discharge are as follows.

1. White vaginal discharge (Bai Dai) is the most commonly seen type of vaginal discharge. It is a Cold condition caused by a Spleen or Kidney Yang Deficiency, or from Exterior Cold Dampness. It manifests as a white, sticky, slimy fluid discharge from inside the patient's vagina with a fishy smell. Symptoms include sore lower back, weakness, and irregular menstruation.
2. Red vaginal discharge (Chi Dai) is caused from Damp Heat in the Lower Burner or Fire Blazing through the Heart and Liver channels. The latter eventually causes a Yin Blood Deficiency. Red vaginal discharge manifests as thick and sticky, white with red mixed together, pale red, or red vaginal discharge with a foul odor. Symptoms include thirst, bitter taste in the mouth, and reddish urine. Continued red vaginal discharge may be a sign of cancer, and the patient should immediately be referred out to an oncologist.
3. Yellow vaginal discharge (Huang Dai) is caused from Heat, usually Damp Heat in the Lower Burner produced from a Deficient Spleen not transporting the Dampness. This condition causes Water Dampness to brew and bind, which then transforms into Damp Heat. The Damp Heat invades the Conception Vessel which causes the Belt Vessel to lose its ability to restrain. It manifests as a yellow, sticky, and

slimy vaginal discharge with a foul odor. A vaginal discharge that has yellow and red with white pus (after menopause) is considered a manifestation of Toxic Damp Heat in the uterus and can indicate endometrium cancer.

4. Green-blue vaginal discharge (Qing Dai) is caused by Damp Heat in the Liver or Gall Bladder Channels. A yellow-green color with pus accompanied by itching and soreness is also a sign of Damp Toxic invasion.
5. Dark brown or black vaginal discharge (Hei Xia) is seen as a form of Blood stagnation.

All gynecological disorders are closely related to the patient's whole constitution, even though the manifestations appear in the reproductive organs. The patient's general constitution, age, dietary habits, home and work environment, should all be taken into account when diagnosing clinical pathophysiologies.

MENSTRUAL COMPLICATIONS

DYSMENORRHEA

Dysmenorrhea is one of the most frequent gynecologic disorders observed in the clinic today. It is described as pain associated with menstruation and refers to cyclic pain (before, during or after menstruation) felt in the lower abdomen or lumbosacral area of the woman's body before, during, or after menstruation. Serious cases of dysmenorrhea can be accompanied by headache, nausea, vomiting, and even syncope.

ETIOLOGY

From a Traditional Chinese Medical perspective, a woman's body is more susceptible to an attack of external pathogens and injury from internal emotional disturbances during her menstrual period. Mental depression may injure the Liver and cause Qi and Blood stagnation.

The invasion of Cold into the Thrusting and Conception Vessels may also cause stagnation and coagulation of Blood, which in turn causes Qi and Blood Deficiency (leading to poor nourishment of the uterus). Excessive exposure to Cold and Dampness can cause Cold to invade a woman's uterus, causing stasis of Blood. A women is prone to Cold invasion when her uterus and Blood are in a weakened state, during and soon after her period.

There are four different phases identified during a woman's menstrual cycle: Before Menstruation, During Menstruation, After Menstruation, and During Mid-cycle.

1. Before menstruation, the woman's Yang rises in preparation to move Blood during her menses; Liver Qi also moves into the lower abdomen in preparation to move the Blood. If a woman's Liver Qi stagnates, it can cause the Blood stagnation in the uterus, leading to pain, especially before her period. This type of stagnation is the most important pathological condition causing dysmenorrhea. Liver Qi stagnation can be caused from emotional strain resulting from suppressed emotions, such as anger, resentment, hatred, and frustration.
2. During menstruation, the woman's Blood is moving. The movement of Blood relies on Liver Qi and Liver Blood. If a woman's Liver Blood stagnates, it will cause pain during her period.
3. After menstruation, the woman's Blood and Yin are Empty.
4. During mid-cycle, the woman's Blood and Yin gradually fill up the Thrusting and Conception Vessels.

According to Western medicine, dysmenorrhea is related to mental stress, emotional instability, mental depression, and fright. Dysmenorrhea can also be due to diseases of the ovaries, uterine, or endocrine glands.

The menstrual pain may appear one to two days before the onset of the menstrual period, and remain until the start of the menstrual flow. It usually lasts from just a few hours to one to two days.

SYMPTOMS

In Traditional Chinese Medicine, dysmenorrhea is classified into two types of syndromes: the Excess type and the Deficient type.

- The Excess type of dysmenorrhea can be caused by either Qi and Blood stagnation, Cold stagnation, or Damp Heat. An Excess condition manifests as distention and pain in a woman's lower abdomen and / or lumbago, occurring before or during menstruation. Symptoms also include obstructed menstruation with pale-purplish color and clots.

- The Deficient type dysmenorrhoea can be caused by either Qi and Blood Deficiency or Kidney and Liver Yin Deficiency. Symptoms of this condition manifest as pain in the lower abdomen, occurring before or during menstruation, which can be relieved by heat and compression. The patient will experience soreness and distention of the waist and legs with pale and scanty menses.

TREATMENT

1. To begin, have the patient sit at the edge of the table and begin stimulating the Yellow Court, Mingmen, and Lower Dantian areas, using the Vibrating Palm technique, for 6 to 8 breaths.
2. Emit Qi, using the Extended Fan Palm technique, into the patient's Lower Dantian, purging the Qi down the Stomach Channels to her feet.

TREATMENT MODIFICATIONS

- When treating patients with the Excess type of dysmenorrhea, emit Qi into the patient's Lower Dantian, St-21, and Lv-14 points, then purge the Stomach and Liver Channels down to and out of the feet.
- When treating patients with the Deficient type of dysmenorrhea, emit Qi into the patient's Lower Dantian, then guide Qi along the Stomach and Spleen Channels to regulate the body.

PRESCRIPTIONS AND HOMEWORK

Have the patient practice the Descending the Yang and Ascending the Yin techniques (see Chapter 41).

PRESCRIPTION MODIFICATIONS

1. For patients with the Excess type of dysmenorrhea:
    a. Have the patient sit at the edge of a chair, relax and breath naturally, with the tongue placed up against the hard palate. The patient's mind should concentrate on her Lower Dantian and navel area.
    b. The patient should imagine the navel as the center of a wheel. When inhaling, she should move the Qi upwards, rotating the energy clockwise, from the lower portion of the right side of the abdomen. The patient silently chants the phrase: "The white tiger hides in the east." Exhaling, she rotates the Qi from above the navel to the left side of the abdomen back to the beginning point, while silently chanting the phrase: "The blue dragon shelters in the west." The patient continues circulating the Qi around the navel, rotating it in a clockwise direction for 36 breaths. In this particular meditation, it is important the patient start with small circles and gradually increase the circumference of the energy circle, in order to disperse the accumulated Heat in the uterus.
2. For patient's with a Deficient type of dysmenorrhea:
    a. Have the patient sit at the edge of a chair, relax and breath naturally, with her tongue placed up against the hard palate. The patient's mind should concentrate on the Lower Dantian and navel area.
    b. Imagine the navel as the center of a wheel. When inhaling, she should move the Qi upwards, rotating the energy counterclockwise, from the lower portion of the left side of the abdomen. The patient silently chants the phrase: "The blue dragon shelters in the west." Exhaling, she rotates the Qi from above the navel to the right side of the abdomen back to the beginning point, silently chanting the phrase: "The white tiger hides in the east." She continues circulating Qi around the navel, rotating from large to small circles in a counterclockwise direction for 36 breaths. In this particular meditation, it is important that the patient start with large circles and gradually decrease the en-

ergy circle's circumference in order to gather energy into the uterus.

MENOXENIA

Menoxenia refers to the pathological changes of menstruation occurring in the cycles, such as the color, quantity, and quality. This includes extended or delayed menstrual cycles, as well as menorrhagia (excess bleeding) and scanty menstruation.

SYMPTOMS

1. Blood Heat symptoms include: prolonged menstrual cycle, profuse menstruation of bright red or purplish color, viscous in quality, and a heavy feeling in the chest.
2. Deficiency of Qi symptoms include: prolonged menstrual cycle, profuse, pale and watery menstruation, listlessness, palpitations, shortness of breath, and an empty or dropping sensation in the lower abdomen.
3. Blood Stasis symptoms include: delayed menstrual cycle, purplish and massed menstruation with pain and distention in the lower abdomen.

TREATMENT

1. Have the patient sit at the edge of the table and begin to stimulate her Yellow Court, Lower Dantian, and Mingmen areas using the Vibrating Palm, for 8 to 16 breaths.
2. Use the Extended Fan Palm hand posture to emit Qi into the patient's Mingmen and Lower Dantian using the Spiral Rotating method of Qi manipulation. Rotate the energy in a counterclockwise direction for conditions of Excess and a clockwise direction for Deficient conditions.

TREATMENT MODIFICATIONS

- When treating a patient with Blood Heat, emit Qi into her Yellow Court and Lv-14 points, drawing the energy down her Liver Channels and out of her feet.
- When treating patients with a Deficiency of Qi, use a Extended Fan Palm hand method while tonifying and filling her Mingmen and Lower Dantian areas.
- When treating a patient with Blood Stasis, use the Extended Fan Palm hand method to emit Qi into her Lower Dantian and rotate the Qi in a clockwise direction.

PRESCRIPTIONS AND HOMEWORK

In order to move the energy and allow Qi and Blood to flow smoothly in the uterus, have the patient practice the Microcosmic Orbit Fire Cycle (see Chapter 41).

PRESCRIPTION MODIFICATIONS

- For patients with Blood Heat, it is advisable to have them practice the Liver Massage method (see Chapter 39) and sound the word "Xu" (Shu).
- For patients with Qi Deficiency, have the patients perform Natural Breathing meditations, as well as the Daoist Five Yin and Yang Organ Exercise (see Chapter 42).
- For patients with Blood Stasis, have the patient practice Circling the Abdomen, for 36 breaths in a clockwise direction (see Chapter 41).

CYSTS AND TUMORS

BREAST CYSTS AND BENIGN TUMORS

A breast cyst or benign tumor is a lump or nodule found within the breast tissue.

ETIOLOGY

The fibrocystic changes that can occur in a woman's breast tissue are commonly divided into three levels of benign breast disorders. Beginning with Level-1 breast cysts (which include the development of benign lumps, nodules or fibroadenoma), progressing onto the development of Level-2 breast cysts (which include the development of mastitis or carbuncles), and ending with Level-3 breast tumors.

The Qigong doctor's treatment will vary according to the type and condition of the lump enclosed within the patient's breast tissue.

DIAGNOSIS ACCORDING TO DIFFERENT TYPES OF BREAST LUMPS

The Qigong doctor differentiates among the types of breast lumps and initiates treatment accordingly. Breast lumps can be divided into three major categories: Phlegm, Blood Stasis, and Toxic Heat.

1. Phlegm lumps are soft with distinct edges, mobile and slippery. They are not swollen, and the skin surface around the area is white.
2. Blood Stasis lumps are hard, immobile nodules. The skin surface around the surrounding area is purple.
3. Toxic Heat lumps are hard, immobile, with indistinct edges. The skin surface of the surrounding area is indented and red.

LEVEL-1 BREAST CYSTS (BENIGN LUMPS, NODULES OR FIBROADENOMA)

These are lumps that feel like hard nodules; they are usually painless, round, flat, and mobile. The lumps do not feel hot or cold to touch, the skin color is unchanged, and the size of the lump may vary depending on the patient's mood changes.

ETIOLOGY

Level-1 cysts are usually caused by the patient's Liver and Spleen being affected by anger, worry, and depression. The Spleen Qi is weakened, thus disrupting the free-flow of the Liver Qi. This leads to conditions of nodules due to both Qi and Phlegm stagnation. Long term Qi stagnation eventually leads to the accumulation of Phlegm and Stasis of Blood.

SYMPTOMS

These type of cysts are characterized by their process of development; they generally start off small and increase in size after several months.

TREATMENT

Eliminate the stagnation and remove the Phlegm using Qi purging and dispersing techniques.

LEVEL-2 BREAST CYSTS (MASTITIS OR CARBUNCLE)

This condition is a painful inflammation of the breast skin and underlying tissue.

ETIOLOGY

Level-2 cysts are usually caused by Qi and Blood stagnation, resulting from the stagnation of the Liver or Gall Bladder Qi, which is accompanied by excessive stagnation of Toxic Heat in the Stomach Channels. Breast cysts can be internally induced by suppressing the emotions, especially grief. Generally mother issues relate to the left breast and Lung area, while father issues relate to the right breast and right Lung area.

TREATMENT

1. After clearing the diseased area of Toxic Qi, the Qigong doctor uses the Sword Fingers technique to stimulate CV-17, St-18, Lu-1, Bl-18, and -21 points, to open the points along the Liver, Lung, and Stomach Channels.
2. Next, the Qigong doctor emits Qi, using the Vibrating and Shaking techniques, to disperse the cyst.
3. The doctor then applies the Dragon Mouth manipulation to dredge the patient's body, by pushing, pulling, and leading the Stagnant and Toxic Qi down the patient's Lung Channels (out the arms), and Stomach Channels (out her legs), via the St-36 points.
4. The patient is given homework, consisting of Jing Point Therapy, and of kneading the surface and surrounding tissue of the cyst at least once a day. This prescription should also include the use of the "Shang" sound tone resonation to further dissolve the cyst, and to dispel the Toxic Qi out of the breasts.

LEVEL-3 BREAST TUMORS

This type of breast cyst corresponds to a severe condition of dysplasia.

ETIOLOGY

Level-3 cysts are usually caused by Liver Fire, accompanied by Heart and Lung Qi stagnation due to the accumulation of grief, sorrow, exasperation, and being upset, or due to the derangement of the patient's Thrusting and Conception Vessels.

TREATMENT

1. After clearing the diseased area of pathogenic Qi, the Qigong doctor uses the Thunder Palm technique to disperse the tumor.
2. After several minutes of blasting the area, the Qigong doctor uses the Spiraling and Circling Energy techniques to circle the tissue areas

surrounding the tumor, drawing the pathogenic Qi down and out of the patient's nearest arm via the Lung Channel (Figure 51.1).

3. The Qigong doctor then uses energetic Pulling and Shaking techniques to release the pathogenic Qi from the patient's body and to draw it out of the arm. The doctor must then pull and snap each of the patient's fingers and thumbs, to release and disperse the Excess Heat and pathogenic Qi.
4. Next, circle rotate the energy within the patient's Lower Dantian, pulling the Qi down the lower legs and out the St-36 points.
5. Finally, fill her Lower Dantian and circle regulate her Governing and Conception Vessels using the Microcosmic Orbit to balance the Yin and Yang energy.
6. Give the patient homework consisting of Jing point therapy that stimulates the GB-21, St-36, CV-17, and Sp-6 points, once a day. If the patient is using magnetic therapy, she should place the magnets on the points for 1 hour. This prescription should also include herbs, and the use of the "Shang" sound tone resonation (straight tone and descending/ascending tone prescriptions) to further dissolve the tumor.

OVERVIEW OF TREATMENT

The general treatment for breast diseases, depending on the severity of the condition, is divided into five main components: Medical Qigong therapy, Medical Qigong massage, prescriptions, Jing point therapy, and herbal teas.

MEDICAL QIGONG THERAPY

The doctor's emitted energy can assist the patient in dispersing stagnations, as well as dissolving breast tumors and cysts. Generally, the doctor's treatment lasts for a period of three days, and can be indefinitely prolonged and increased in effectiveness if the patient is given additional Medical Qigong prescriptions as homework.

MEDICAL QIGONG MASSAGE PRESCRIPTION

This therapy is most important as it allows the patient to reconnect with her tissues and further allows her to feel and release emotions trapped

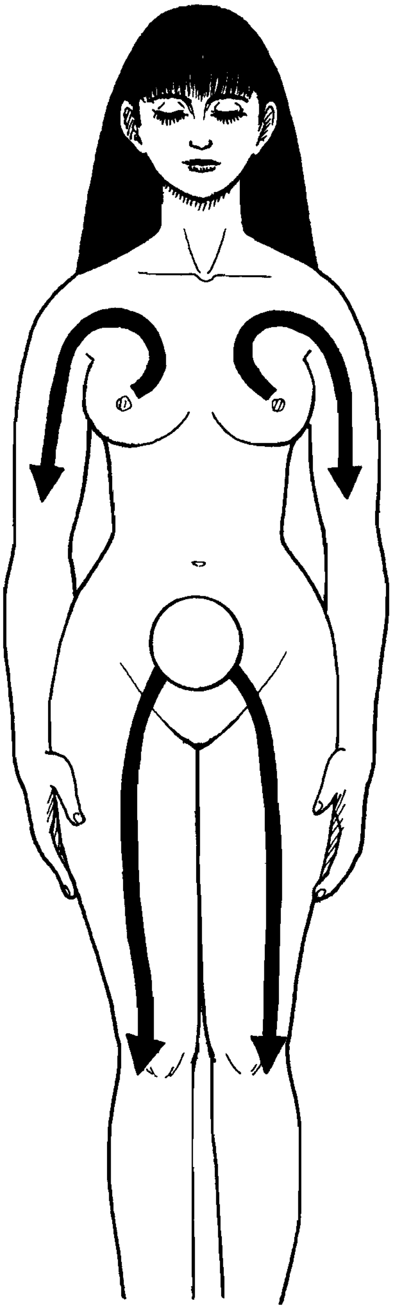

Figure 51.1. After purging the Toxic Qi from the patient's upper body, the Qigong doctor continues to disperse the patient's Toxic Qi from the lower abdomen, leading it down the patient's leg and out of the St-36 points.

within her body (remember, you can only heal what you feel). This healing process is as follows.

1. The patient begins by rubbing her hands together to get them hot.
2. Placing her hands on her breast, the patient inhales and draws the heat from her hands into the breast tissue, while imagining divine healing light flowing from the Heavens, and pouring into her body to fill up her chest.
3. The patient circle massages her breasts 36 times towards the outside of her body, concentrating on the breast tissue filling up with the divine healing energy that destroys any and all pathogens stored within her breast tissues.
4. After completing 36 times in one direction, the patient reverses directions but maintains the same visualization.
5. Next, the patient exhales the straight healing sound "Shang" for 9 breaths, toning, while turning the body from the left to the right side. The patient focuses on purging the pathogens out of her breast area through her mouth. This purges her Lungs and the surrounding breast tissues.
6. The patient then exhales the descending/ascending healing sound "Shang" for 9 breaths as she turns at the hips, toning from the right to the left side of her body, and focuses on purging the pathogens out of her breast area. This purges her Lung Channels and the surrounding energetic connections to the breast area.
7. The patient then rubs 36 times in a clockwise direction, from her abdomen to her Lower Dantian, leading the Qi downwards. This is performed to root the energy back into her Lower Dantian.
8. The patient ends the prescription in a quiescent state of mind.

MEDICAL QIGONG PRESCRIPTIONS

This therapy allows the patient to purge the organs that are found to be the root cause of the disease. In Medical Qigong, the doctor views the patient's cysts or tumors as a flower. If, through surgery, the top of the flower is plucked from its stem (and the branch, stem, and root cause of the disease is left still intact), the tumor or cyst will return again. In order to heal the patient's condition, the doctor must also address the original cause of the disease and destroy the entire support system. The purpose for these prescriptions is to destroy the disease's entire root system to kill the flower.

1. The Dry Crying prescription is given to purge the emotions of grief and sorrow from the patient's Lungs (see Chapter 53).
2. The "Guo" healing sound is given to purge the patient's Liver Fire.
3. The "Shang" healing sound is given to dissolve the cyst or tumor.
4. Prescribe Jing Point Therapy. Specific points are prescribed for the patient to enhance the doctor's treatments, and increase her own healing potential. The patient is encouraged to stimulate specific areas of her body, by emitting Qi while pressing and focusing her intention into the following points:
    a. The Small Intestine 3 point affects the patient's breast disease caused by Rebellious Qi and Qi depression.
    b. The Stomach 18 point frees Qi flow locally in the patient's breast.
    c. The Stomach Channel transverses the breast, and the Stomach 36 point regulates Qi of the patient's entire body.
    d. The Spleen 6 point is specifically chosen for treating breast cysts.
    e. The Gall Bladder 21 point (Shoulder Tapping) is prescribed to assist the patient in sinking Liver Qi Heat.

HERBAL TEAS

All herbal teas are administered to move the Qi and assist the doctor's treatment. Dandelion tea, for example, is mixed with green tea to clear Heat from the patient's body; or green orange peel is added to help move the Qi.

OVARIAN CYSTS AND UTERINE TUMORS

In an ovarian cyst, there is a closed sack or pouch containing fluid, semifluid, or solid material, usually characterized by Phlegm, Damp Heat, and Stasis of Blood.

ETIOLOGY OF OVARIAN CYSTS

Ovarian cysts can be caused by psychological upheavals which lead to enlarged follicles that fail to rupture, as well as by Liver Fire stagnation. Ovarian cysts are usually related to the anger from the Liver turning inward and attacking the body.

ETIOLOGY OF UTERINE TUMOR

A uterine tumor, or myoma, is a solid benign growth in the myometrium, and is often called a fibroid. It contains muscle tissue. Excessive uterine bleeding is a common effect of myomas. Ovarian cysts and uterine tumors are generally divided into three types of myomas:

1. The subserous type is located on the outside wall of the uterus;
2. The interstitial type is located within the wall of the uterus;
3. The submucous type is located on the inside of the uterus.

TREATMENT

The same treatment is used for both ovarian cysts and uterine tumors. To treat these conditions, the doctor uses Medical Qigong therapy in conjunction with Jing point therapy and herbs.

1. Begin with the patient in a supine position and purge and clean the diseased area of pathogenic Qi, using the Thunder Palm technique in conjunction with the Vibrating Palm, to disperse the stagnation in the ovaries or uterus.
2. Disperse pathogenic Qi out the patient's body via the Gall Bladder Channels.
3. Tonify the patient's Lower Dantian with Qi and circulate the energy through the Microcosmic Orbit.
4. Press and stimulate with Jing point therapy the Sp-6, Sp-9, St-36, GB-34, and GB-38 points (Figure 51.2).

PRESCRIPTIONS AND HOMEWORK

The patient should be given homework, in accordance with her constitution. The prescriptions should include:

1. The "Guo" sound to disperse Liver Fire.
2. The "Yu" sound to disperse ovarian cysts.
3. Prescribe Lower Dantian regulation exercises to access and regulate the ovarian energy.
4. Prescribe herbs for chronic conditions.

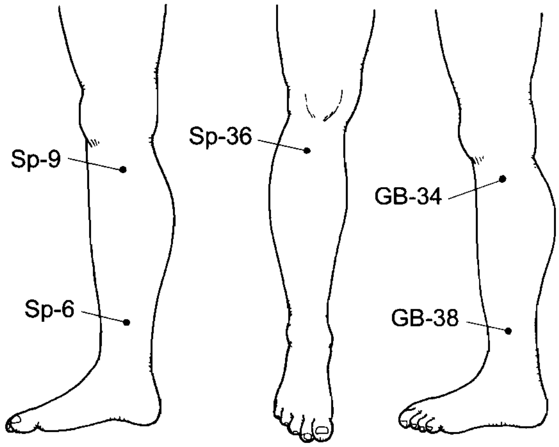

Figure 51.2. When treating ovarian cysts, use Jing Point therapy in conjunction with Qi emission, to stimulate Sp-6,-9, St-36, G.B.-34 and 38 points.

UTERINE COMPLICATIONS

MISCARRIAGE

A miscarriage is the termination of a pregnancy, at any time before the fetus has attained extrauterine viability. Miscarriage occur usually between conception and the fourth month of pregnancy.

In Traditional Chinese Medicine, a miscarriage can have several stages and levels of severity, ranging from such conditions as Fetal Leakage, Restless Stirring of the Fetus, Falling Fetus, and Small Birth. Within each of these conditions there are several patterns of disharmony.

ETIOLOGY

There is an ancient Chinese saying which states "A miscarriage is more serious than child birth." According to Traditional Chinese Medicine, there five basic mechanisms associated with the cause of a miscarriage: Empty Conception and Thrusting Vessels, Blood Deficiency, Depressed Liver Fire, Blood Stasis, and External Injuries.

1. An Empty condition of the Conception and Thrusting Vessels causes miscarriages. The mother's Kidney Qi consolidates the fetus and is responsible for constricting the anal and vaginal orifices. If the Conception and Thrust-

ing Vessels become empty, they cannot consolidate and secure the fetus within the womb. If any of the following three conditions occur singly or in concert, both the Conception and Thrusting Vessels may indeed become Empty: (1) the mother's Yin Essence is depleted or consumed; (2) the Mingmen Fire is insufficient, and (3) both the Yin and Yang Qi are Deficient. Any of these three conditions may be due to: congenital weakness and insufficiency, chronic disease, extreme or prolonged stress, unrestrained sex after conception, or the mother's age (over 40).

2. Blood Deficiency from an irregular diet, excessive worry and anxiety may deplete the Spleen and Stomach organs resulting in a miscarriage. Since the Spleen is the postnatal root for Blood generation and transformation, an Empty or Deficient condition of Spleen Qi may cause Blood Deficiency. Furthermore, as the Spleen Qi restrains and supports the mother's abdomen, a Spleen Qi Deficiency, or Empty Spleen Qi, may lead to the prolapse of the Central Qi and produce insufficient Blood to nourish the fetus. Both the Central Qi prolapse and the Blood Deficiency can present the Conception and Thrusting Vessels from consolidating the fetus and securing it in the womb.
3. Depressed Liver Fire due to suppressed emotional frustration (with mental agitation, anger, jealousy, and envy) may cause a miscarriage. The excessive Heat generated by the depressed Liver Fire may in turn attack the mother's Sea of Blood, scorching and injuring the Thrusting (Chong) Vessel. This injury, in turn, may cause the mother's Yin Blood to flow recklessly downward, leaving no place for the fetus to rest.
4. Blood Stasis may be due to three different factors: past internal injuries, iatrogenic disorders, and long term Qi stagnation. The excessive accumulation of Qi and Blood within the mother's womb during conception can sometimes aggravate a preexisting condition, and thereby cause hemorrhaging, as well as initiate energetic patterns which fails to stop her from bleeding.
5. External Injuries from accidents (sustaining a sprain or strain in the abdominal region), or battery (being hit or struck) can lead the body into a state of shock. This shock to the mother's body may damage either the Conception or Thrusting Vessels, preventing them from securing the fetus within the womb.

TREATMENT

In the Medical Qigong clinics in China, doctors are cautioned to be extremely careful when treating pregnant patients. Qi emission can cause a spontaneous miscarriage if done incorrectly. It is therefore recommended that only Medical Qigong doctors, who have specialized in the specific treatment of pregnant women, be allowed to diagnose and treat them. The interdependencies of the mother's and the fetus' energy circulation is too complex and delicate for a generalized practitioner to handle.

Treatments for miscarriage vary, based on pattern discriminations of Kidney Deficiency, Qi and Blood Deficiency, Blood Heat or physical trauma. An example of an ancient Chinese prescription requires the doctor to treat Yingu Kd-10 for miscarriages with vaginal bleeding, dark urine, abdominal distention, and fullness with an inability to lie down.

If the patient has already experienced a miscarriage, and is in the state of recovery, specific treatments vary according to the doctor's diagnosis concerning the etiology and pathology of the patient's condition.

HOMEWORK AND PRESCRIPTIONS

Each consecutive miscarriage causes further injury to the patient's Kidneys, creating or aggravating the stagnant Blood condition in the uterus. Doctors of Traditional Chinese Medicine therefore encourage their patients to take herbal medicines, and to practice Medical Qigong for six months before attempting to conceive again.

When they do become pregnant again, they are encouraged to continue to practice Medical Qigong meditations and exercises for strengthening and balancing their body's energetic fields and internal organs.

A patient recovering from a miscarriage is given specific prescriptions according to her emotional condition, organ dysfunction, and the etiology of the condition.

UTERINE COMPLICATIONS DUE TO ABORTIONS

Another important aspect of Medical Qigong and gynecology is helping a patient face the loss, or termination of a pregnancy. This can occur either intentionally (through an abortion), or unintentionally through a miscarriage, spontaneous abortion, or the death of a fetus by other causes.

When a woman has undergone an elective abortion, the Qigong doctor must not only consider the potential for scar tissue formation and physical complications within her uterus, but also the energetic and emotional complications resulting from the abortion.

While treating the patient, the doctor may discover within the uterus what appears to be the energetic remains of the aborted embryo. These phantom embryos may in fact be thought forms created by the mother's feelings of guilt and shame at having undergone an abortion. These difficult emotions are further compounded by deeper feelings of self-hate, that go unnoticed by others.

Most people feel little compassion for women who suffer a miscarriage, and discourage them from grieving. They are frequently advised by well meaning friends to believe that the miscarriage was all to the good, preventing a possible deformed or sickly baby from being born.

If women who miscarry are discouraged from grieving, those who make the painful decision to abort often face scorn and disbelief when they grieve their loss. Silently they mourn as best they can, haunted by feelings of shame imposed upon them by others and their own conscience.

Most women struggle long and hard before resorting to abortion. Afterwards, they often obsess about the living embryo, and what it felt like to be pregnant, centering all their attention on the uterus. Such intense, emotional concentration creates phantom embryos that, in fact, take on the form of ghostly embryos. These thought forms or energetic imprints are clearly visible to the doctor's inner vision. The energetic imprint contains the location, exact age and size at the time of the embryo's termination. The unprocessed energetic charge of the embryonic ghost's location creates an energetic vortex. This vortex, in turn, draws emotions such as rage, anger, guilt, and shame into itself. If this area is not purged, the resulting accumulation of negative emotions will develop into cysts, tumors, and possibly cancer.

This energetic reaction to surgical abortions has been confirmed by several Western doctors (all wishing to remain anonymous). These doctors observed that after many years of performing abortions, within a very short time period several patients (whose ages ranged from 16 to the mid 30's) developed reoccurring cysts and tumors. The location of the cysts or tumors in the majority of the patients uteri was at the exact tissue area of the embryo.

HEALING THE ENERGETIC TRAUMA

When the patient has decided to heal from her loss, she must not only address her physical condition (scar tissue formation, etc.), but also heal from the spiritual and emotional woundings as well. To assist the patient to heal emotionally and disperse the energetic charges, the doctor must help the patient transform the energy stored within her Wu Jing Shen, which is responsible for maintaining the toxic energetic pattern. Only after experiencing and releasing the emotions stored within the Wu Jing Shen (see Chapter 13) is the patient ready to heal. The healing process is divided into three distinct steps:

1. The patient is encouraged to completely forgive herself for losing the child and to reconnect with her higher power.
2. The patient is encouraged to completely forgive herself for being in the situation, and assume the responsibility of the pregnancy and the loss.
3. The patient is encouraged to completely forgive the person responsible for making her pregnant.

It is important for the patient to completely feel and experience all three stages of forgiveness in order to begin healing these energetic wounds.

Withholding the smallest bit of guilt or anger will seek to sabotage the healing process, and will also allow the toxic energetic charge to remain within the patient's tissues. Even the smallest bit of anger or guilt that remains will continue to recreate a destructive energetic cycle, storing and retaining the patient's toxic emotions.

RELEASING THE EMBRYO'S ETERNAL SOUL

The purpose and goal of releasing the embryonic thought form is to free the patient from the energetic remnants of the phantom embryo. The patient is encouraged to make peace with herself and with the energetic form of her unborn child, as well as the divine (God or higher spiritual force incorporated in the patient's own belief system). The patient does this through deep, meditative prayer, positive affirmation, and an attitude of honor and respect for the relationship between herself and the embryo's energetic remains. Even though the Eternal Soul does not remain present during the abortion, the mother may still need to atone for her decision not to bring the baby to term. This atonement may be necessary to fully release all feelings of guilt in accordance with the patient's religious belief system.

The patient must also make peace with the embryo's Eternal Soul, by enveloping the phantom embryo in a circle of divine healing light and love, to return it back to the divine. In order to achieve this, it is helpful for the patient to visualize creating a small energetic bubble, full of love and healing light, to house the departed embryo's energetic remains before returning it back to the divine. This serves to complete the emotional closure. Once closure is established, the energy of the embryo's energetic remains will leave her body, allowing the patient to feel a fulfilling sense of release, completion, calmness, and inner-peace.

UNDERSTANDING ENERGETIC ABORTIONS

The gift of life is an incredible miracle. If a difficult decision is made to not keep a pregnancy, however, then the patient should consider contacting the embryo's Eternal Soul. Through prayer, meditation, imagination, and visualization, the mother and father can connect to the developing embryo and request that the embryo's Eternal Soul return back to the light of the divine. Sometimes called an energetic abortion, this energetic transformation can take place within the first developing stages of the mother's pregnancy.

Since the Eternal Soul is ageless, it is possible to carry on a mind to mind conversation. When a baby is born, its brain must have time to develop before it can grasp complex information; the mind itself must have time to direct the brain's growth and the body's maturation. Once the Eternal Soul is freed from the body, however, it is not confined by the physical limitations of the infantile brain and should be spoken to as an equal. This method is used by psychotherapists on the leading edge today, using hypnosis to facilitate the patient's connection to the Eternal Soul through visualization.

Patients who have undergone age regression psychotherapy to the state of conception and embryonic state often describe visiting the mother's womb during the time of conception, then leaving and returning at various times during the embryonic stage of development. The Eternal Soul's residence does not appear to be permanent until the end of the first trimester (see Chapter 2). Until then, the Eternal Soul may come and go as it chooses, staying, at times, just long enough to insure the continued growth and development of its physical body.

It may be for this reason, that talking mind to mind with the embryo's Eternal Soul to request its departure, is often successful in precipitating a miscarriage. As the Eternal Soul leaves the embryo's body (sometimes observed as a small ball of white light energy), all that remains in the mother's womb is an empty, lifeless mass of tissue. Usually the mother's body will naturally abort the lifeless tissue mass, either immediately, or by the next morning.

CHAPTER 52

QI EMISSION THERAPY FOR NEUROLOGY (STROKE, COMA, PARALYSIS, AND MULTIPLE SCLEROSIS)

INTRODUCTION

Neurology is a branch of medicine that deals with the nervous system and its diseases, as well as neuropathology. Neuropathology is the study of diseases of the nervous system and the structural and functional changes occurring within them. Neurological diseases are divided into either congenital predispositions that develop over time into full-blown disorders, or those in which acquired destructive pathogenic factors act upon the brain. The latter are usually caused by toxic, traumatic, mechanical, inflammatory factors or neoplastic types of diseases (abnormal tissue formation).

The following is a description of several treatments and prescriptions used in various Qigong clinics and hospitals throughout China. Please be aware that specific treatments will vary according to each patient's constitution and the nature of his or her diseases.

STROKE (WIND-STROKE)

A stroke is a sudden loss of consciousness followed by paralysis, and may be caused by one of several different factors, including a brain hemorrhage, the formation of an embolus or thrombus in an artery, or the rupture of an extracerebral artery causing subarachnoid hemorrhage.

ETIOLOGY

Although a stroke occurs suddenly, in actuality, the causal factors build up for several years. A stroke can be caused from high blood pressure, diabetes, high cholesterol, chronic dehydration, overwork, emotional stress, irregular diet, excessive sexual activity and severe migraines. Strokes usually occur at night, or early in the morning (the same is true regarding heart attacks).

There are four pathogenic factors involved in causing stroke. These factors are: Wind, Phlegm, Fire and Stasis. The Liver is dependent on the Kidney Water (Yin Qi) to moisten and soften it. If, however, the Kidney Yin becomes Deficient, this may cause the Liver Yang energy to grow, and to generate Fire. Excess Liver Fire may stir up Internal Wind which causes the Qi and Blood to rebel upwards. Phlegm is then formed, which obstructs the cavities and vessels, causing the patient to lose consciousness. This can further develop into Penetrating Wind (i.e., stroke). Internal Wind may also develop from Heat and Phlegm caused by the patient eating too many sweet and fatty foods (irregular diet), overwork, emotional stress, excessive sexual activity, and inadequate rest.

In the clinic, conditions referred to as stroke (Wind-Stroke) include cerebral hemorrhage, cerebral thrombosis, cerebral embolism, and cerebrovascular spasm. These four conditions are termed in Western Medicine as "Cerebro-Vascular Accident" (CVA), and include: Cerebral Hemorrhage, Cerebral Thrombosis, Cerebral Embolism, and Cerebrovascular Spasm.

1. Cerebral Hemorrhage consists of bleeding, which is caused from a rupture of a sclerosed, or diseased vessel in the brain;
2. Cerebral Thrombosis is an obstruction of a cerebral artery by a thrombus (blood clot);
3. Cerebral Embolism is a condition which occurs when an embolus (bubble of air, or piece of a thrombus) detaches from a thrombus and obstructs a cerebral artery; and
4. Cerebrovascular Spasm (Ischemia) is a condition which occurs when a cerebral vessel temporarily contracts, or constricts, impeding Blood flow. Hence, the oxygen supply is obstructed sufficiently to cause symptoms. When reversible, or when there is no permanent damage, they are known as Transient Ischemic Attacks, or TIA's. The TIA's are pre-stroke symptoms, commonly known as a mini-stroke.

SYMPTOMS

In dealing with stroke patients, it is important to differentiate between the severe type (which attacks the internal organs and channels), the mild type (which attacks only the channels), and the chronic type.

Stroke manifestations can either arise independently from an attack on the channels, or as a result of an attack on the internal organs. After a severe type of stroke, the patient enters into the next stage, clinically manifesting symptoms similar to that of the mild type. The shorter the time lapse before beginning treatment on a stroke patient, the better the chances are for recovery.

THE SEVERE TYPE OF STROKE

In a severe type of Wind-Stroke, the energy attacks the patient's internal organs and channels. It is characterized by a loss of consciousness, aphasia, numbness, paralysis and/or coma. If the right side of the brain is damaged, there will be paralysis on the left side of the patient's body with accompanying memory defects and language difficulties. If the left side of the brain is damaged, there will be paralysis on the right side of the patient's body, speech and language difficulties, slow, cautious movements, and memory problems.

Clinically, severe stroke patients are further classified into two categories: Tense or Flaccid. The Tense type (also referred to as the Closed type) corresponds to the Collapse of Yin. The Flaccid type (also referred to as the Open type) corresponds to the Collapse of Yang. These conditions and symptoms are further described as follows.

1. The Tense Stroke Symptoms include: clenched hands, open eyes, clenched jaw, red face, coarse breathing, and a forceful pulse.
   - **Treatment for Tense Stroke conditions.** Rescue the patient's Yin by giving herbs, as well as extending Qi into the patient's GV-20, -26, Kd-1, Lv-3, St-40, and the 12 Jing-Well points on both hands (Lu-11, LI-1, Ht-9, SI-1, Pc-9, TB-1). The patient should be given homework upon recovering, depending on his or her constitution.
2. The Flaccid Stroke Symptoms include: limp hands, closed eyes, open mouth, weak breathing, and a weak pulse.
   - **Treatment for Flaccid Stroke conditions.** Rescue the patient's Yang by giving herbs, as well as extending Qi into CV-8, CV-6, and CV-4 points, and GV-20 points (with or without the application of Moxa), to increase the patient's Yang. Patient should be given homework upon recovering, depending on his or her constitution.

THE MILD TYPE OF STROKE

In the Mild Type of Wind-Stroke, the energy attacks only the patient's channels. This is characterized by numbness, slurred speech, and unilateral paralysis. Clinically, mild stroke patients are further classified into two categories that are determined by the type of channels that are attacked.

1. An attack of the main channels causes no loss of consciousness or coma; however, there will be unilateral paralysis (hemiplegia), numbness of the limbs and deviation of the eye and mouth. Clinical manifestations vary according to the location of the channels and collaterals that have been affected.
   a. Symptoms of an attack on the channels include: facial paralysis, hemiplegia, motor impairment and numbness of the limbs.
   b. Symptoms of an attack on the collaterals consists of a unilateral numbness of the face and extremities.
2. An attack of the connecting channels is characterized by numbness.
   - **Treatment for mild stroke conditions.** In order to treat a Mild stroke condition combine Medical Qigong therapy with acupuncture therapy and herbs, to:
     a. Remove obstructions from the patient's channels,
     b. Purge Wind and transform Phlegm,
     c. Invigorate the patient's connecting channels,
     d. Regulate the circulation of Qi and Blood in the patient's channels, and
     e. Prescribe homework in accordance with the patient's constitution and condition.

In the acute stage, stroke patients should immediately be treated in combination with Western medicine, utilizing Qigong, acupuncture and herbal therapy. Treatments should be given everyday (usually every six hours) for the first month, and every other day for the second month. The general treatment is aimed at relieving the patient's spasms, and lowering his or her blood pressure (if the condition is due to hypertension). The primary points treated are determined by the type of stroke, which internal organs were affected, and the extent of damage to the patient.

THE CHRONIC TYPE OF STROKE

This condition is characterized by slurred speech or an inability to talk or eat, and involves either a unilateral (hemiplegia) or bilateral (paraplegia) paralysis.

- **Treatment for Chronic Stroke conditions.** In order to treat a chronic stroke condition the Qigong doctor must invigorate the patient's Blood, tonify the Spleen and open the patient's channels. The specific points will vary according to the area and location where the channels are blocked. The most commonly used points include: St-4, St-6, St-7, TB-17, GB-14, LI-4, SI-18, Tai Yang (Temple), and Yin Tang (Third Eye Point).

  If the body is in a state of unilateral paralysis, the doctor also treats the opposite side of the patient's head using Jing Point Scalp Therapy, in order to move stagnation, purge any Excess, and stimulate and fill any Deficiencies.

COMA

A coma is an abnormal, deep stupor occurring during illness, as a result of the illness, or due to an injury.

ETIOLOGY

From a Traditional Chinese Medical perspective, a coma results from Qi obstruction to the brain. More then 50% of coma cases result from trauma to the head, or circulatory accidents in the brain. A coma can be initiated from several different types of conditions.

1. A trauma induced coma is caused by an injury to the head as a result of accidents and internal hemorrhages. Electrocution can also induce a coma.
2. Insufficient Blood flow to the brain can cause a coma. This can result from circulatory accidents within the brain caused from hypertension, arteriosclerosis, thrombosis, and tumors.
3. An overdose of drugs, alcohol or medications, poisoning, gases or fumes such as carbon dioxide or carbon monoxide can all induce a coma.
4. Temperature abnormalities, such as hypothermia or sunstroke can cause vomiting, diarrhea, shock and then coma.
5. Insulin shock, epilepsy, electrolyte abnormalities, endocrine problems, meningitis, uremia or renal problems, and hyperglycemia can all induce a coma.
6. An infection induced coma is caused by an acute infection and bacterial intoxications as in fevers, botulism, and other infectious diseases; or as a result from abscess formations.
7. Psychogenic comas are also known as a hysterical coma.
8. Windstroke comas are caused by tension, hyperactive Liver Yang and Phlegm Fire overheating the brain. This results in urine and stool incontinence, cerebral hemorrhaging, and thrombosis. This is often located at the union of the anterior and posterior cerebral arteries, called the Circle of Willis.

SYMPTOMS

When a patient is in a coma, the Shen is completely out of its residence and the patient cannot be aroused by external stimuli. The patient's Shen is inactive, yet the Hun and Po are still there. For death to occur, it is not enough for the patient's Shen to leave, but the energy of the Hun and Po must also leave.

TREATMENT

In treating coma patients, it is mandatory to use methods that will activate the patient's Qi, remove the Blood stasis, and return the Shen back into the body. In the acute stages of treatment, use Medical Qigong therapy in addition to the Invisible Needle technique (see Chapter 35). It is most important to treat the coma patient as soon as possible. I have trained with three doctors of Traditional Chinese Medicine who were proficient in coma revival. Below are two accounts of coma revival cases from Dr. Lu Guo Hong's clinic at the Hai Dan Medical Qigong College in Beijing, China.

Case Study #1. Dr. Lu's patient had been in a coma for four hours prior to treatment. The patient regained consciousness within fifteen minutes during the treatment. Dr. Lu treated his patient using the Vibrating Palm technique (on points GV-20, Yin Tang, GV-15, CV-17, Bl-15, Pc- 6, TB- 5, and the extra points Shi Xuan at the tips of the fingers). By squeezing the Shi Xuan points, the Excess Heat was allowed to disperse (Figure 52.1).

Case Study #2. A female patient apparently incurred a series of Transient Ischemic Attacks (TIA) over a period of two years prior to coming to Dr. Lu. As the condition progressed, she developed a thrombosis on the left side of the brain, which paralyzed the right side of her body. The patient then apparently incurred thrombosis of the right side as well.

Because there were thrombosis on both sides of the brain, the patient was expected to die within a month. Dr. Lu, assisted by two other doctors, treated the patient for three days using the Extended Fan Palm technique, Invisible Needles, and Vibrating Palm technique. In this particular case, the doctors used the same points as in the previously mentioned case to cultivate the patient's Yuan Qi, adding points Bl-23, GV-4, CV-4, and CV-8. The patient was treated twice a day for 15 minutes, both morning and evening.

On the fourth day, the patient regained consciousness, and the treatment was continued once a day for a month using the Conception and Governing Vessel circulation points (to move the Qi along the Microcosmic Orbit), in addition to points on the arms including LI-4, LI-11, LI-15, TB-5, and TB-14 points.

When treating the lower limbs, he used GB-30, GB-31, GB-34, Sp-6, Sp-10, and St-41 points. The patient was given prescriptions and homework, specific to her constitution, state of consciousness, and motor control.

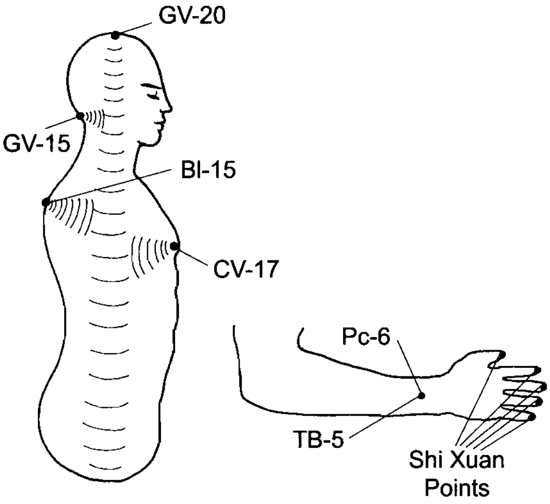

Figure 52.1. Common Points used for Coma Revival

The patient gradually regained movement in her limbs. Once the patient could move her limbs, she was given Dynamic Qigong to increase the flow of energy to her extremities. This included meditations and exercises that used guided imagery to regain muscle cellular memory. After two months, the treatment time was changed to every other month until the patient was fully restored.

FACIAL PARALYSIS (BELLS PALSY)

This condition is known as the sudden onset of unilateral facial paralysis.

ETIOLOGY

The patient experiences the swelling of the seventh facial nerve due to the External Wind and Cold invasion.

SYMPTOMS

The disease is characterized by hemiparalysis of the face which manifests as numbness of the affected area, deviation of the eye and mouth, distorted mouth with the corners tilted on the healthy side, and a half closed eye with dacryorrhea (excess tear flow).

TREATMENT

For this type of paralysis, it is important that the Qigong doctor treat the patient as soon as possible. The shorter the duration time from the initial onset and the treatment, the more successful the therapy tends to be.

1. Begin by purging the Toxic Qi out from the patient's face and head area, dredging the Toxic Qi down the arm and out the LI-4 point.
2. Using the Extended Fan Palm hand technique, emit Qi into the affected area, guiding the Qi to circulate in the Microcosmic Orbit from the anterior to the posterior of the patient's head.
3. Finally, initiate the Fire Cycle Microcosmic Orbit to regulate the patient's Qi.

TREATMENT MODIFICATIONS

- For treatment of the first stage of paralysis, the doctor is advised to apply gentle Pressing and Kneading techniques to the patient's face before using the Extended Fan Palm hand technique.
- To treat patients whose paralysis has remained in a chronic state, the doctor is advised to use deep pressing and kneading techniques before using the Vibrating Palm technique.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient massage the affected side of the face from the corner of the mouth to the ear and forehead, down over the side of the ear and down the side of the neck 81 times.
2. The patient then focuses on specific areas on the face, circle massaging with the tips of the thumbs for 24 times.
3. Finally, the patient is given the Wash the Face and Massage the Head exercise to perform for 18 times (see Chapter 39).

MULTIPLE-SCLEROSIS (MS)

In Chinese medicine, multiple sclerosis is considered a type of Atrophy Syndrome. It is an inflammatory disease of the central nervous system in which infiltrating lymphocytes (predominantly T cells and macrophages) deteriorate the myelin sheath of nerves. Its pathology consists in the partial destruction of the myelin sheath around the spinal cord, brain and optic nerves.

ETIOLOGY

Multiple sclerosis is suspected to be an autoimmune disease, somehow linked to a viral infection originating from an invasion of External Dampness (especially in the beginning stages). As the External Dampness invades the body it begins to obstruct the channels, causing heaviness, numbness, and tingling in the limbs.

Diet is another contributing factor. The excessive consumption of greasy fried foods, dairy foods, or Cold foods can impair the Spleen, leading to the formation of Dampness.

In the later stages of MS, excessive sexual activity can also be a contributing factor for weakening the body and escalating the disease's progression. Excessive sexual activity can weaken the patient's Kidneys, contributing to such symptoms as dizziness, blurred vision, urgency or hesitation of urination, and extreme weakness of the legs. If excessive Liver Yang develops, symptoms will also include stiffness of the legs, severe vertigo, and vomiting. In the later stages of development, Liver Wind can cause tremors and severe spasms of the patient's legs.

Clinically MS is categorized as either a passive aggressive type of disease, in which episodes of neurologic dysfunctions are followed by times of remission, or a steady progression of neurologic dysfunction.

SYMPTOMS

The symptoms of MS were once thought to develop gradually; however, this is not necessarily true. In about 40% of clinical cases, the onset may occur in less than a few hours. In about one half of the patients, at onset the symptoms include weakness, or numbness in one or more limbs; later, the tendon reflexes become hyperactive.

In young patients the first obvious symptom is often retrobulbar neuritis (inflammation of the nerves behind the eyeball) causing blurred vision. In older patients the most common symptom is weakness of the legs.

Patients may experience vertigo, retrobulbar or optic neuritis, double vision, unstable walking, vomiting, and difficulty urinating. Also included are numbness or tingling of the extremities, as well as periodical electric shocks radiating through the limbs. These symptoms may be due to optic nerve, brain stem, cerebellar and spinal cord involvement.

Traditional Chinese Medicine divides these symptoms into two categories of clinical manifestations: Damp Phlegm with Spleen Deficiency, and Liver and Kidney Deficiency.

Clinical manifestations in patients with Damp Phlegm and Spleen Deficiency include feeling weak and tired, dizzy, as well as numbness, tingling and feeling of heaviness in the legs.

Clinical manifestations of patients with Liver and Kidney Deficiency include progressive weakness of the back, legs and knees, blurred vision, dizziness, poor memory, and a hesitancy, or an urgency to urinate.

TREATMENT

There are two basic treatment patterns in Medical Qigong for multiple sclerosis. If the treatment is started within its early stages, the symptoms can be completely eliminated, and the progression of the disease indefinitely stopped. This, of course, must be done in conjunction with an alteration in the patient's eating habits (no red meat, processed meat, sugar additives, or high fat diary products). The patient should reduce the amount of sexual activity, get plenty of rest and exercise, and avoid allergy causing agents. Treatments are described as follows.

1. To treat a patient with Damp Phlegm and Spleen Deficiency, first resolve the Dampness by purging the patient's torso and extremities. Next Dredge the Toxic Qi down the legs and out the Sp-9, and Sp-6 points. After purging, emit Qi into the Yellow Court (CV-12) to tonify the Spleen. Finally tonify the patient's Lower Dantian and regulate his or her Microcosmic Orbit (Fire Cycle).
2. To treat a patient with Liver and Kidney Deficiency, first tonify the patient's Kidneys and Liver by emitting Qi into the patient's Mingmen and Lower Dantian. Next emit energy into the patient's Bubbling Spring (Kd-1) points and direct the Qi to flow up the patient's Yin Channels into the patient's Lower Dantian and Kidney area. Finally, regulate the patient's Microcosmic Orbit (Fire Cycle).

PRESCRIPTIONS AND HOMEWORK

Patients often benefit from Medical Qigong exercises such as the Descending the Yang and Ascending the Yin Technique, Walking Therapy, Daoist Five Organ Exercises, as well as Taijiquan (Tai Chi Chuan) practice. One prescription I have found to be very effective in treating MS patients is as follows.

1. Have the patient sit in a chair, feet flat on the ground, hands resting on the knees, anal sphincter closed, tongue touching the upper palate while inhaling and exhaling through the nose.
2. While using natural breathing, the patient focuses on the Lower Dantian to create an energy ball. With each inhalation, divine healing light is absorbed into the body and stored in the Lower Dantian. With each exhalation, this healing light radiates within the Lower Dantian. After several minutes, the patient notices the Lower Dantian becoming full of heat and light.
3. Next, the patient imagines the energy ball in the Lower Dantian leaving the body through the coccyx into the Earth upon exhalation. Upon inhalation, the patient imagines the energy flowing up the center of both legs and filling the Lower Dantian. This meditation should continue for at least 20 to 30 minutes per sitting.

CHAPTER 53

QI EMISSION THERAPY FOR PSYCHOLOGY

INTRODUCTION TO ENERGETIC PSYCHOLOGY

The object of Chinese psychology is to study and understand the interplay of internal and external events on physical, mental, emotional, energetic and spiritual dimensions of human existence. These interactions include the human qualities and characteristics, as well as the fundamental attributes that exist and evolve within the changing environment and evolutionary development of nature. These interactions at the most fundamental level can be viewed as the transformational interactions of Yin into Yang, and Yang into Yin vibrational energy.

All matter vibrates regardless of its form (gaseous, liquid, solid, or in a transitional state such as ice melting into water). The human body receives these vibrations through the skin, eyes, ears, nose, and mouth, and records them in its cells. The mind interprets these incoming vibrations as images, sounds, smells, and sensations, and organizes these into memories. The extent to which the mind interprets the incoming data, depends upon the level of consciousness and awareness of the individual at the time of the recording. In times of danger, alertness normally increases and consciousness either expands or focuses more sharply until the danger is past; or if an injury occurs consciousness and awareness may be lost (e.g., a car crash).

When a Qigong doctor, attunes his or her vibrations to the patient's energetic field and emits Qi, the body's cells responsible for storing the vibrations are stimulated. The doctor's intention to heal the patient by releasing traumatic memories, combined with his or her projected Qi, often causes the patient to relive the original traumatic experience as the recorded vibrations are released. The Qigong doctor (still in tune with the patient's energy) is receptive to the patient's reaction and is able to interpret the released energy in the form of images, sensations, smells, sounds, and emotions while they are being reexperienced by the patient. If the doctor is sufficiently attuned, he or she may even pick up the conscious thoughts of the patient during this process.

As doctors tune in to their patients' cell emissions they are also able to project their own thoughts and interpretations of these memories back to the patients. This creates a positive feedback system, wherein the doctors' thoughts can influence the patients' thoughts and emotional reactions. Through thought transmission, Qigong doctors are often able to break the vicious cycle of negative thinking that so often leads to disease in the body.

In Western psychology, it is believed that a person's tendency to interpret events - even negative imaginary events (e.g., assigning malicious intentions to peoples' actions), can result in a self-fulfilling prophecy. By assuming and expecting the worst behavior in people, such a person always feels on the defensive and provokes others around him or herself to anger or irritation, thus reinforcing the negative viewpoint.

Chinese energetic medicine adds to this simple behavioral understanding of self-fulfilling prophesies, a new level of comprehension based on energetic interchange between people. When someone observes, or imagines, a specific negative thing happening, certain brain cells record this information in the form of vibrations. These vibrations can be received by another person's corresponding brain cells and projected back to the original sender, reinforcing the negative interpretation of the event(s). If this feedback loop continues - an imaginary wrong may become manifest, i.e., the negative event imagined actually occurs.

It is believed that different cells store different types of information, received in the form of vibrations. The silent thought communications between doctors and patients is believed to involve different sets of cells than those used to receive and record emanations of divine energy. These cells interrelate with the patient's emotional changes that are stored and released from the Yin and Yang organs and tissues, and are expressed through the patient's Wu Jing Shen.

The patient's ontology also includes the patient's growth and self-realization in relationship to the Wu Jing Shen (Five Essence Spirits) and the emotions stored within the body's Yin and Yang organs. Each of the Five Yin Organs has a Yin and a Yang aspect within its energetic substance, as well as emotional factors, and dynamic interconnections. The Yin and Yang aspects give rise to the Creative and Controlling Cycles and the Five Yin organs with their associated elements and emotional energies (Figure 53.1 - 53.3).

Summary of the Five Elements		
Wood	Liver	Anger, irritation and rage
Fire	Heart	Joy, happiness, pleasure and excitement
Earth	Spleen	Worry, sympathy, anxiety, contemplation and meditation
Metal	Lungs	Grief, sorrow, anguish, distress and heartache
Water	Kidneys	Fear, terror, panic, horror and fright

Figure 53.1. The Five Elements, Organs and Acquired Emotions

THE EMOTIONAL ENERGY STORED WITHIN THE FIVE YIN ORGANS

Traditional Chinese Medicine classifies the etiology of diseases into Internal and External physical, mental, emotional, energetic, and spiritual influences. In Chinese psychology, the primary focus is placed on the "root" of the energetic/emotional dysfunction. The root of the energetic/emotional dysfunction is often the unprocessed emotions which are stored in their associated organs. Once the patients' mental and emotional symptoms have been alleviated, the energy that was being used to suppress the excessive emotions, is then freed to be focused on their physical healing.

Traditional Chinese Medicine also recognizes the power of thoughts, images (including psychotic delusions and hallucinations), and fantasies to generate and release emotions. Qigong doctors study these images and emotional reactions of attraction or aversion to evaluate their patient's condition. When traumatic memories and emotions are suppressed over a long period of time, or are still fresh in the mind, any minor incident can precipitate a tidal wave of emotions. A rapid influx of thoughts and images sweeps through the mind, too fast for the conscious brain to register. These thoughts and images are of past painful experiences, and are loaded with emotional energy. Getting cut off by someone on the freeway, for example, may cause a momentary reaction of surprise, or an explosive enduring rage, depending upon the condition of the Five Yin Organs. In the latter case, the driver is overwhelmed by the Liver's sudden release of rage, triggered by the traumatic images and emotions from the past.

Clinical pathology is considered an Excess or Deficiency of the emotions at the level of the Five-Elements and their corresponding organs. The Excess or Deficiency of an internal organ can cause a release of thoughts or images that sweep through the patient's mind, bringing a sudden emotional response. In some instances these emotions first affect a specific organ, and then the entire balance of energy, as the first organ's energetic field deteriorates or adversely expands. When emotionally charged Qi is trapped within the tissues of the patient's internal organs, it travels through the channels, affecting the metabolic order of the viscera, glands and nervous system.

Creative Cycle			
Wood	Anger	Creates Joy	Fire
Fire	Joy	Creates Worry	Earth
Earth	Worry	Creates Grief	Metal
Metal	Grief	Creates Fear	Water
Water	Fear	Creates Anger	Wood

Figure 53.2. The Creative Cycle

Controlling Cycle			
Wood	Anger	Controls Worry	Earth
Earth	Worry	Controls Fear	Water
Water	Fear	Controls Joy	Fire
Fire	Joy	Controls Grief	Metal
Metal	Grief	Controls Anger	Wood

Figure 53.3. The Controlling Cycle

Emotions consist of energetic light that is either tempered and softened by the patient's belief structure and energetic fields. When the emotions and life purpose are in harmony, the Wu Jing Shen become focused; the Hun and Po do not disperse, grief and anger do not arise, and the Five Yin Organs effectively combat pathogenic influences. If this energetic light becomes unstable, i.e., too Excessive or Deficient, the patient's inner strength and individual nature begins to undergo dramatic changes.

The body's internal organs and their relationship to the Five Elements are as follows.

EXCESSIVE ANGER INJURES THE LIVER (WOOD)

The Chinese character for anger (Nu) is depicted by the Heart ideograph underneath two characters, a woman (placed on the left) and someone who has the upper hand (placed on the right). It expresses the feeling and frustration of an enslaved woman being dominated by someone.

PATHOLOGICAL MANIFESTATIONS OF ANGER

Pathological anger is considered the perverse movement of the element Wood. Anger is a generated force that has the potential to unleash explosive amounts of Qi, pushing the energetic manifestations to the extreme.

THE YIN AND YANG ASPECTS OF ANGER

The emotion of anger can be divided into Yin and Yang tendencies. The Yang type of anger can burst out due to accumulated pressure, resulting in the upward flow of massive amounts of Blood and Qi. The Yin type of anger implodes, remaining buried or held inside, thereby causing internal agitation and dissatisfaction which generates blockages.

The Liver's negative nature is one of being stubborn and rude; its positive nature is one of being compassionate, unselfish, kind, and merciful. The rising of Heat from the Liver's Qi corresponds to the level of anger; since the Liver stores the Blood and the Blood stores the Hun (Ethereal Soul). When the Hun leaves the body due to anger and rage, the animal nature of the Po take over. As the individual's Liver Fire rises, it causes the eyes and the face to become red. Liver Fire can also injure the Liver as well as the Blood, bile and Body Fluids associated with this organ.

ANGER AFFECTING THE LIVER

When Liver Qi becomes depleted it retracts, causing the emotion of fear to arise. When Liver Qi fills to Excess, the result is progressive restlessness, irritability, anger, and rage. In some instances, these emotions affect the entire balance of energy so that the organ begins to deteriorate.

Anger can cause trembling of the entire body due to the Liver's control of the tendons and ligaments.

When an individual experiences any type of boundary invasion, the Liver Qi triggers the Heart (Shen) to recall from its long term memories past traumas (containing both anger and fear). At the same time, the Liver also drains the Kidney Qi to meet the apparent life-threatening situation; this prepares the body for the flight or fight response. The power of the emotional response causes the Heart to go into shock and prevents the Shen from directing the mind. Thus another trauma is experienced by the patient's interpretation of the event (e.g., getting cut off on the road) and will automatically trigger an exaggerated emotional response in the future.

Given that the Liver is the mother of the Heart (responsible for long term memory) and the child of the Kidneys (responsible for short term memory), the Liver can drain the Kidneys and flood the Heart with anger, causing one's mind to go blank and the face to turn red.

ANGER AFFECTING THE SPLEEN AND STOMACH

Anger can damage the Spleen and Stomach, causing digestive problems, loss of appetite, and a sour taste in the mouth. These symptoms are directly related to the level of anger.

ENERGETIC PATTERN OF ANGER

The Qi pattern of anger has a very hot energy associated with it. When discharged, this hot energy swells up like a tidal wave expanding from behind the person's back over the head (Figure 53.4). It crests over the head and rushes forward like a wild animal attacking. The emotion also travels from the groin up the back of the body and out through the eyes.

Grief controls anger as metal cuts wood. At the root of anger is pain and sorrow. The anger protects the patient from feeling the initial hurt that festers like an infected wound. When the doctor lances the wound, allowing the toxic anger to drain, the uncovered original hurt can heal cleanly. Anger is transformed to sorrow and grief. The compassionate, non-judgmental attitude of the doctor allows the patient to grieve openly. The release of toxic anger and pain enables the patient to listen to the Hun. By focusing on the patient's emotional hurt from past experiences, which are the root of his or her anger, a shift from anger to hurt occurs. This exchange of emotions allows the patient to confront the hurtful past memories and discharge the toxic anger.

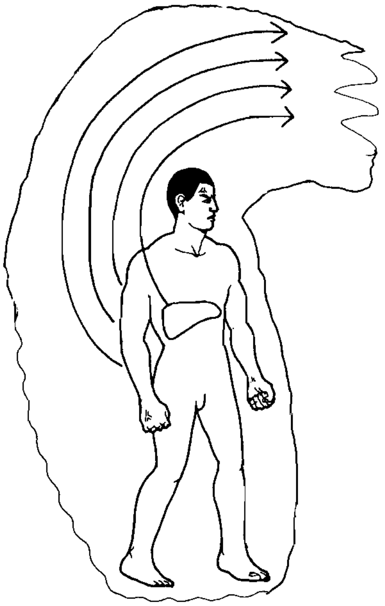

Figure 53.4. Anger rushes from the Liver over the back like a tidal wave, causing the Qi to ascend.

EXCESSIVE JOY INJURES THE HEART (FIRE)

The Chinese character for joy (Le) is depicted by the ideograph of a large drum, framed by bells mounted on a wooden stand. It expresses the feeling of rhythmical ceremonies and excitation. The character gives the image of a hand beating a drum during a festival (this same character is used for music).

THE YIN AND YANG ASPECTS OF JOY

When the Heart's emotions are balanced, they can be experienced as love, joy, happiness, excitement, and pleasure. This warmth radiates through the body affecting the expansion of free and easy Qi circulation. Joy is communicated and felt throughout the body, mind and soul.

PATHOLOGICAL MANIFESTATIONS OF JOY

Excessive joy and laughter can injure the Heart and scatter the spirit; the Qi becomes congested and the Heart flutters erratically. When the Heart Qi becomes depleted, it can no longer control the Lungs: the emotion of grief arises, and when the Qi fills to Excess, hysteria results. Water extinguishes Fire, therefore the emotion of fear controls the emotion of joy. By supporting a hysterical patient to become aware of his or her underlying fears, the doctor can startle the patient into a sense of sobriety.

Pathological joy and laughter are considered the perverse movement of the element Fire. If the patient's Fire becomes uncontrollable, the Qi - which expands outward towards the periphery - will cause the patient to become hysterical. After Fire has exhausted the patient's life-force, it leaves the patient exhausted, dazed, and disenchanted. The Heart's negative nature is confused, doubtful, and greedy; the positive nature is open-minded, loving, non-judgemental, and trustful.

ENERGETIC PATTERN OF JOY

Joyful feelings make the body's Qi become soft and relaxed. This energetic pattern is expansive, and causes a slower movement of the Qi. As the Heart controls the blood vessels and stores the Shen, a joyful spirit produces a clear, calm mind and a healthy body. This energy bubbles out from the center core of the body and ripples out in all directions (Figure 53.5). A light energy gently flows from the recipient back to the person expressing the emotion. While in this state, the patient's head is usually held erect, the face is full of color, and the limbs are relaxed.

EXCESSIVE WORRY INJURES THE SPLEEN (EARTH)

When the Spleen Qi becomes depleted, all four limbs lose their strength and the Five Yin Organs become unstable. When the Spleen Qi becomes Excessive, the body becomes swollen and the flow of urine and menses loses its smoothness.

THE YIN AND YANG ASPECTS OF WORRY

The emotions associated with the Spleen are relative to the energy expended in contemplation, meditation, anticipation, and worry. Too much worry and thought produces Qi stagnation. A stagnant condition causes Stomach disorders, such as poor appetite, indigestion, ulcers, constipation, or diarrhea. Stagnation can also block the Shen, which is responsible for emotional and mental balance, and thus result in a lack of common sense. The Spleen's negative nature is suspicious and self-centered; the positive nature is truthful, sincere, and compromising.

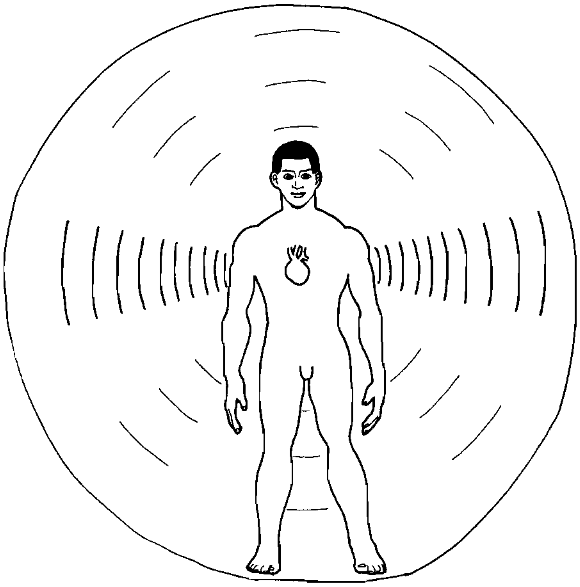

Figure 53.5. Joy bubbles out from the center core of the body, rippling out in all directions, causing the Qi to slow its flow.

WORRY AFFECTING THE MEMORY

The Spleen stores the Yi, and the Yi stores ideas. The Spleen transfers short term memory to the Kidneys; the Kidneys later transfers the acquired memory to the Heart, which is responsible for long term memory. In trauma victims, the Kidneys' role in transferring memories, can also account for the loss of memories that are too terrible to accept, since these memories are transferred along with great fear, or fail to get transferred at all.

THE ENERGETIC PATTERN OF WORRY

When one is worried, the energy comes up from the Earth and congests in the Middle Dantian area, thus weakening the Wei Qi (Figure 53.6).

CONTROLLING THE EMOTION OF WORRY

Because the element Wood controls the element Earth, the emotion of anger controls worry. It is very important that any anger uncovered be directed at the root cause, and not at the doctor who is trying to help a patient end a negative pattern. The doctor will gently challenge toxic belief structures held by the patient, until the patient begins to confront whatever issue is at the root of the anger.

When the patient becomes angry with the doctor, it is generally a condition known as "transference." If and when transference arises, the Qigong doctor can ask the patient, "Who do I remind you of right now?" or "When was the last time you felt this way?" Trace the patient's feelings as far back as possible. If the doctor gets mad at the patient, it is called "countertransference." Both transference and countertransference can involve any emotion (e.g., falling in love, feeling afraid, etc.).

EXCESSIVE SADNESS AND GRIEF INJURES THE LUNGS (METAL)

The Chinese character for sadness (Bei) is depicted by the Heart ideograph underneath the character for the back (placed on the top). It portrays a picture of a person who has turned his or her back on the true feelings of the Heart. The Chinese character for grief (Ai) is depicted by the mouth ideograph (wailing, groaning, and lamenting) placed within the center of the character of someone dressed in a special garment for mourning. Pathological grief is considered the perverse movement of the element Metal.

When grief is experienced, its natural movement of condensation begins to compress the Heart, interfering with Blood circulation. This obstruction destroys both the respiratory flow of the Lungs, as well as the body's liquids. The Lungs store the Qi, the Qi stores the Po (Corporeal Soul). When the Lung Qi becomes depleted, shortness of breath develops. When the Lung Qi becomes excessive, the result is a feeling of discomfort in the chest, panting, and difficult breathing.

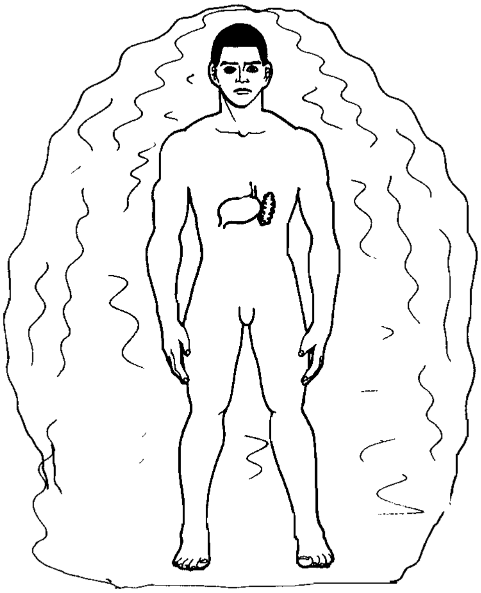

Figure 53.6. Worry binds the Qi, causing the Qi to stagnate.

THE YIN AND YANG ASPECTS OF GRIEF

The feelings associated with the Lungs are sadness, melancholy, grief, sorrow, and anxiety. The negative nature is cunning, jealousy, and pessimism; the positive nature is generosity and righteousness. In the Five Elemental Controlling Cycle, the Lungs (Metal) affect the Liver (Wood). Excessive sadness from the Lungs can cause the Liver Qi to stagnate the Blood (Metal Overcontrolling Wood), leading to possible intercostal and hypochondriac pain, bloating, and depression.

GRIEF'S AFFECT ON THE INTERNAL ORGANS

Grief disperses the Qi, and thus injures the Lungs. Crying can lead to the release of toxic anger (that suppresses the feelings of sorrow) and grief. Depression and despair may also be manifestations of abnormal Liver Qi stagnation. When grief from the Lungs is excessive, the Spleen may be affected (this almost always happens after a major loss, such as a death in the family).

THE ENERGETIC PATTERN OF GRIEF

The external energetic patterns of grief and sorrow have a heavy, downward moving energy (Figure 53.7). The internal energetic patterns and manifestations of grief have a very thick, sticky energy, and a denseness like gelatin. Long periods of extreme grief also injure the Heart, Pericardium, and the Triple Burner Channels.

CONTROLLING THE EMOTION OF GRIEF

Because Fire melts Metal, the emotion of joy controls grief. The grieving process is a very important part of healing; however, chronic grief can be extremely detrimental, and when it is extreme and prolonged it can lead to spiritual oppression or possession. If the survivor cannot accept the death, he or she may prevent the departure of the loved one's spirit for many years on end. The patient may beg the Eternal Soul not to leave and cling with unrelenting grief to the lost one's energy. This type of dependency deepens, rather than lessens over time, as the patient opens more to receive the loved one's spiritual essence. In so doing, other Gui (spirits) are given power to enter as well.

After a death of someone loved, the patient may consciously suppress any joy or pleasure. The patient may believe that to allow any relief to suffering is tantamount to a betrayal of the deceased. This belief structure must be addressed by the doctor and patient. Establishing a sense of empathy with the patient will provide a safe environment for the patient to release the emotion of grief. The doctor can work towards empowering the patient to begin to recognize times of joy, and learn, or remember, how to count his or her blessings, to let go of sadness and grief, and reclaim lost joy.

Grief may give way to despair and anxiety. Because anxiety suppresses respiration, anxiety lowers the body's resistance to pathogenic factors by weakening the Wei Qi; anxiety also damages the Large Intestine, resulting in symptoms that range from constipation to ulcerative colitis.

Just as too much grief can cause inappropriate laughter, or gallows humor; too much laughter masks the accumulated and suppressed grief. Gallows humor is a syndrome whereby the patient laughs instead of grieving about hurtful past, or current experiences. Gallows humor is a defence mechanism of the patient's denial. Grief and humor are examples of opposites (Yin and Yang) transforming into each other.

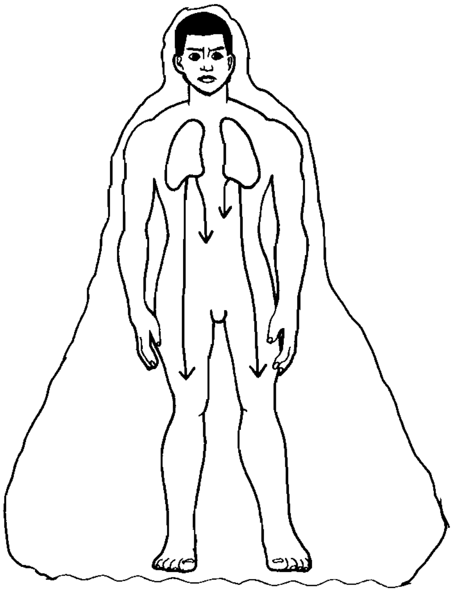

Figure 53.7. Grief causes the Qi to become obstructed.

The concern in dealing with many grieving people (pertaining to the death of a patient's mate or close friend, etc.) is that the patient may be using grief to suppress other unresolved emotions, such as: guilt, fear and loneliness.

1. Fear of forgetting the deceased voice, appearance, etc. may be due to feelings of guilt over not having expressed enough love towards the person when it was possible to do so.
2. Guilt over any expressed feelings of joy or happiness, may seem morally wrong because they cannot be shared by the departed.
3. Guilt may arise from too many hurtful things said, or too many things left unsaid to the departed.
4. Guilt over never having had the chance to really say good-bye (as with victims of accidents and sudden death) may become all consuming.
5. Loneliness and isolation may result from the loss of friendships.

6. Worry that loss of income and security may cause financial difficulties and health problems may cause the patient to cling to the past.

To assist the patient to heal from these toxic beliefs, the doctor prescribes creative visualizations, such as the following.

"Imagine, see and feel yourself and your loved one clothed in your astral bodies. Embrace and express your feelings and thoughts, knowing that these will be heard and received. Listen to your loved one's answer.

Ask forgiveness for any cruel remarks or failures to be as loving as you wish you could have been. Visualize your loved one accepting the request for forgiveness; embrace with love, and release him/her to be with God with your blessings."

In cases where the patients could not attend the funeral, have the patient imagine the funeral and perform a ritual closure (e.g., lighting a candle at home and saying a prayer).

Finally, encourage the patient to give him or herself permission to experience some joy in life and to go on living.

EXCESSIVE FEAR INJURES THE KIDNEYS (WATER)

The Chinese character for fear (Kong) is depicted by the Heart ideograph underneath two characters, a structure (placed on the left) and repeated strokes by the small blows of a hammer (placed on the right).

THE AFFECT OF FEAR ON THE INTERNAL ORGANS

The Kidneys store the Jing, the Jing stores the Zhi (actions that the mind expresses). Pathological fear is considered the perverse movement of the element Water. Fear is the unrestrained sinking and uncontrolled descent of Water Qi. When the Shen of the Heart is no longer supported by the Kidneys' Jing, the Shen leaves its residence. This results in the emotional release of panic. If the Heart Fire becomes weak, the Kidney Water rapidly draws downward and freezes. Insecurity, hesitation, immobility, and paralysis ensue. If the Heart Fire become Deficient, it cannot radiate warmth, causing the patient to crouch and hide.

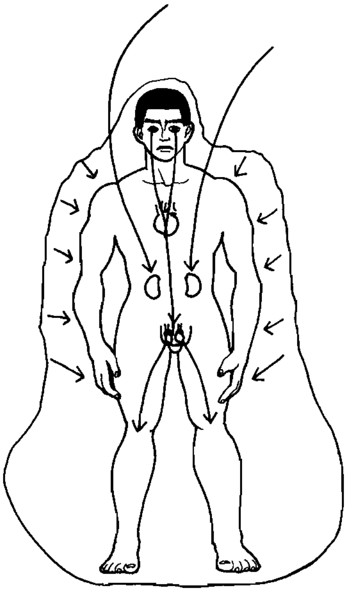

Figure 53.8. Fear causes the Qi to move inward and downward.

If the Kidney Yin is Deficient, the vision deteriorates, there is a ringing in the ears, dizziness, night sweats, and nocturnal emissions. If the Kidney Yang is Deficient, the Mingmen Fire becomes Deficient, and causes a counter-flow of energy affecting the hearing. Kidney Yang Deficiency also results in shortness of breath, cold limbs, and impotency. When the Kidney Qi becomes Excessive, swelling occurs and the Five Yin Organs become unstable. The feelings associated with the Kidneys are fear, paranoia, and panic.

THE YIN AND YANG ASPECTS OF THE KIDNEYS

The Kidneys' negative nature is arrogant, troublesome, and ignorant; the positive nature is peaceful, soft, and tender. Fear can cause loss of Bladder control, weakening of the knees, renal failure, and permanent damage to the Kidneys. It can injure the Heart, affecting the speech, eyes, and the Shen.

When fear injures the Kidneys, this can affect the Liver, causing muscle spasms, trembling, and disruption of the menses cycle. When the Gall Bladder is affected, decision-making becomes more difficult.

THE ENERGETIC PATTERN OF FEAR

The energetic pattern of fear is very cold and implodes into the center of the body, drawing all the energies inward and downward. Fear travels from the eyes down the front of the body into the groin (Figure 53.8).

Because Earth absorbs Water, contemplation controls fear. Therefore, the doctor can help the patient control the fear through visualization and meditation. The purpose of the visualization is to resolve and reprogram the patient's thinking patterns.

THE EMOTION OF FRIGHT

The emotion of fright is distinguishable from fear by the sudden, unexpected shock to the system. The Chinese character for fright (Ju) is depicted by the ideograph of a small bird opening its frightened eyes, in fear, anxiety, and dread. Fright startles the Shen, which then scatters the Qi and injures the Heart. Persistent fright causes damage to the Kidney Qi and results in a negative cycle of chronic fear.

FIVE YIN ORGAN EMOTIONAL RESPONSES

The common ingredient in any emotional dysfunction is the power element maintaining it. Holding on to original traumas, whether consciously or unconsciously, requires a great deal of energy. It is possible to stop this negative pattern by releasing the toxic emotional energy, thereby creating the opportunity for the mind, body, and spirit to heal. A new pattern of harmonious balance can be maintained, however, only after the old dysfunctional patterns are consciously recognized during their occurrence. Old habits die hard; new patterns require time and continuous effort before becoming as automatic as the previous dysfunctional habits. Every emotion is meant to serve a positive function, even anger, fear, and grief. Patients often require reminding of this fact. Every emotion also has its darker side when repressed for too long and when avoided at all costs.

1. Learning how and when to express anger is as important as not holding on to it for extended periods of time. Anger serves as a warning sign that personal boundaries are being violated, threatened, wounded, or that injustices are being committed. Patients must be taught better communication skills so they need not resort to screaming or insulting others to protect themselves. Appropriate assertiveness skills must be learned, and may require either psychotherapy, or classes in assertiveness training. Patients must also learn how to ask for what they need and desire, without expecting those close to them to mind read. Without learning how to set appropriate boundaries, without the ability to say "no" to those who take advantage of them, they will fall into the old habit of accumulating resentments; or they will alienate their employer, friends, and family through unguarded candor when angry.
   - On its negative side, anger can cause rage and destructive patterns (both internally and externally).
   - On its positive side, anger can cause constructive boundary setting and the righting of wrongs.
2. Fear, like anger, serves as a warning sign. Fear informs patients when they are in danger, so that appropriate measures can be taken to insure safety and survival. Some patients do not recognize the danger signals until it is too late to avoid falling victim to unscrupulous people. Some may require counseling and instruction in how to establish healthy relationships, and in how to identify individuals who are potential abusers before it becomes too late to avoid the ensuing violence.
   - On its negative side, fear can cause disabling panic.
   - On its positive side, fear can cause an awareness of the need for self-protection.
3. Joy in excess can lead to disharmony. Nervous laughter, or always playing the clown to gain acceptance, can be extremely difficult to unlearn. Often professional help is needed to increase socialization skills and self-esteem.

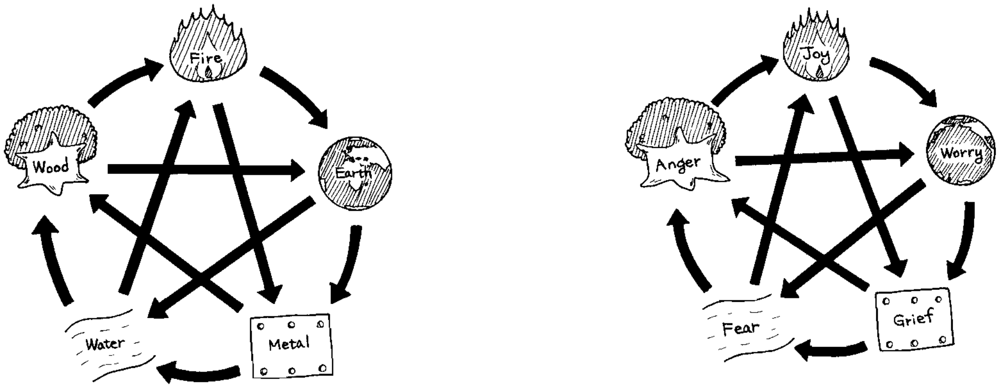

Figure 53.9. The Five Element Cycles

- On its negative side, joy can cause hysteria.
- On its positive side, joy can be a source of happiness to the spirit.

4. Sorrow is necessary at times to motivate patients to change. The old self must periodically die to make room for new experiences and learning. Risking sorrow is part of the process of bonding and maturing. Sadness is often a signal of imbalance in the patient's life, indicating insufficient rest or recreation, and far too much work.
   - On its negative side, sorrow can cause self-pity, rather than compassion for oneself.
   - On its positive side, sorrow can release distress.
5. Worry and anxiety are normal reactions to change. When these emotions become pathological in intensity of duration, patients may need professional help in the form of medication (e.g., for obsessive compulsive disorders), and/or psychotherapy (e.g., in the case of generalized anxiety disorder and posttraumatic stress disorder).
   - On its negative side, worry can cause great anxiety.
   - On its positive side, worry can be a motivating source for change.

The healing stages used in Medical Qigong Therapy employ these transitions in order to assist the patient's emotional and spiritual growth.

The picture in Figure 53.9 summarizes the Five Elements, their associated organs, and their associated emotions.

The cycles serve a functional importance for emotional transitions. The elements can be arranged in a number of ways. For clinical purposes, the main emphasis is on the Creative and Controlling Cycles.

EMOTIONAL DISHARMONY

It is becoming more common in the Western belief system that if a patient concentrates his or her mind and emotions on healing the illness, they can accelerate the recovery. Some Western research suggests that when the patient stops thinking and trains the mind to relax (reaching an alpha state), the healing is accelerated.

Emotional disharmony can resonate and penetrate one's entire being, thereby creating a vicious

Major Organ Disharmony	→	Emotional Disharmony
↑		↓
Environmental Disharmony	←	Disturbed Behavior

Figure 53.10. The Cycle of Emotional Disharmony

circle affecting the body, mind, emotion, energy, and spirit. As evidenced by the diagram (Figure 53.10), any disharmony in one of these four factors can trigger the cycle.

This vicious circle can be set in motion by any major disharmony, be it environmental, physical, emotional, or behavioral. Once in motion the disease can feed on itself. The power of emotional energy generated can at times overwhelm the conscious mind. The mind then has three options: to give in to the emotional tidal wave, to become eventually paralyzed through the effort of controlling it, or simply to deny that anything is wrong.

When emotional energy is suppressed it accumulates within the body, causing the Qi to flow incorrectly or to condense. The unresolved emotions cause chronic blocks in the body, particularly in the back, neck, Stomach, and intestinal areas. This toxic energy stored in the muscles and organs leads to tension, stress, and illness. The root cause of chronically tense shoulder muscles, for example, can be traced to: tensing the muscles to suppress the emotion(s), a strong negative belief system, and an accumulation of unresolved past traumatic memories, which continue to resurface.

Unless the initial cause of tension is released, the tissue memory associated with the initial trauma will repeat its programmed pattern, resulting in chronic tension and pain. Painful and negative emotions are not the only emotions suppressed. Many people also suppress feelings of joy and love. Most do so out of fear, guilt, and sense of unworthiness. Some do so out of religious beliefs, believing that suffering is good for the soul, or that the are being punished by God for sins of omission and / or commission to test their faith.

All emotions are triggered by sensory associations with pleasant and unpleasant past memories. In many cases patients are completely unaware of the triggering smell, image, sensation, or sound. To them it seems that the trauma, or memory, resurfaces from nowhere. This inability to pinpoint the trigger mechanism may indicate the progression of the patient's illness. Patients suffering from panic attacks, for example, have no idea what stimulus may have caused the overwhelming terror that they feel. Quite often panic attacks increase in frequency, partly out of the anticipatory fears of having another one, and partly due to their fear spreading and generalizing.

The first panic attack may have taken place in a restaurant, triggered by one of the five senses registering similar data (a smell or image for instance) which occurred during one of their traumatic memories. Henceforth they will avoid this restaurant, or perhaps all restaurants because new data has been added to the original trigger mechanism. New smells and images occurring during the attack may now stimulate the old memory of the trauma. Eventually, such patients may become home bound as they try to limit themselves to only safe places; their panic attacks have generalized to agoraphobia (fear of places or situations from which escape might be difficult or embarrassing). Each panic attack requires more energy to suppress both the original trauma and all of the new traumas added to it from surfacing.

Outgoing, successful people can thus transform themselves into anxiety ridden hermits. They can also develop physical symptoms, or turn to alcohol and drugs in attempts to cope. In less extreme mental / emotional diseases, changes in skin colors, tone of voice, mannerisms, and nervous twitches may manifest. As the mind tries to suppress uncomfortable emotions, these unresolved feelings often manifest as physical symptoms.

At times, however, the symptoms manifest more on the mental level, as in obsessive compulsive disorder. The logical mind, failing to completely suppress the emotion, may develop frantic obsessive thinking patterns, that seem to take on a life of their own. Some patients may obsess about dying, even if they seem by Western medical standards to be physically completely healthy; or they may obsess about killing their own child, or forgetting to lock all the doors when leaving the house and so on. They have no control over their repeating thoughts, and are aware of the unrealistic and unreasonable nature of these obsessions. Nevertheless, they are helpless to stop themselves.

Obsessions often manifest with compulsive addictions such as washing the hands so often during the day that they are raw and bleeding, or pulling out the hair on their head, or folding and unfolding their clothing. Any ordinary daily activity may turn into a compulsion, requiring the patients to continue performing a meaningless action again and again.

Another frequent disease found in traumatized patients goes by the name of posttraumatic stress disorder. Patients suffering from this condition have flash backs, i.e., they can not stop having moments of releasing the original traumas again and again with full intensity.

In posttraumatic stress disorder patients' symptoms are often disabling. Anxiety is chronic, they are in a state of hyper-alertness. They have become good at disassociating from their body, since it has become a very uncomfortable place to be. The Shen wanders and the patients may, or may not, be aware of where their spirit goes or what it experiences at these times. In extreme cases the ego fractures, usually during early childhood, and a full blown identity disorder (previously called multiple personality disorder) ensues.

Patients suffering from obsessive compulsive disorders, posttraumatic stress disorder, agoraphobia and panic attacks, or identity disorders should be referred out for extra help to a mental health professional specializing in these disorders.

Ideally, the doctor can teach and encourage the patient to get in touch with their body, and learn how to achieve and maintain a balance between the physical, mental and emotional planes. The doctor can also teach the patients how to transition detrimental energy into a higher level of consciousness, when the patients become aware of an unpleasant emotion. When patients learn how to channel their own energy in a more healthy way, the cycle of reliving the past can be broken. A negative emotion can be transmuted and expressed as a positive emotion; emotional energy can be converted to positive life-force energy.

EMOTIONAL FREQUENCY RESONATION

When the Qigong doctors are appropriately sensitive to the patients' emotional transitions, they are able to detect, and recognize, recurring energetic patterns. These energetic patterns are in constant resonation throughout the body's tissues. As patients shift their emotional thinking, their energetic frequency resonation (the body's Hz rhythm) shifts. These different emotions establish specific patterns of energy in the body. Each individual has a brain wave pattern that is as unique as their fingerprints. According to Richard Lee, director of the China Healthway's Institute, individuals resonate at different frequencies depending on their state of consciousness. Mr. Lee's research reveals the following.

1. Those motivated by fear and instinct will resonate at the delta wavelength, between 0 and 4 Hz (hertz, i.e., cycles per second). This level of energetic resonation, as measured by an electroencephalograph (EEG), is normal in an awake infant and a sleeping adult. This delta wavelength, however, indicates brain pathology in adults that are awake.
2. Those motivated by attachment and desire will resonate at the theta wavelength, between 4 - 7 Hz. This level of energetic resonation is generally common in children and in adults in the early stages of sleep. This frequency range also occurs when an individual creates mental imagery, and often manifests when a person is involved in deep altered states of consciousness. This level of energetic resonation also occurs in many brain disorders.
3. Those motivated by service and compassion will resonate at the alpha wavelength, between 8 and 13 Hz. These low amplitude alpha waves usually occur when the brain is in a calm, and the mind is relaxed (in a state of wakefulness). It was discovered that Qigong masters emit energy within this frequency range.
4. Those motivated by organization and planning will resonate at the beta wavelength, between 14 and 25 Hz. This level of energetic

resonation usually occurs when the nervous system is active and the body is awake. It manifests during sensory input and mental activity. This frequency range is usually common for those suffering from nervous overload and excessive thinking.

When the patients' vibratory rate shifts towards the alpha rhythm, they begin to experience deep relaxation and clarity of mind. Relief from chronic pain and symptoms associated with low energy is achieved.

According to Richard Lee, the infrasonic sound produced by Qigong masters is effective because it causes decreased resonation in the theta and beta ranges. This simultaneously decreases attachments and desires, calms the emotions, and reduces excess thinking, thereby freeing the patient's vital sources to be redirected towards healing.

DIFFERENTIATION OF PSYCHO-EMOTIONAL DISORDERS

Traditional Chinese Medicine has, for centuries, recognized the fact that psycho-emotional states, energy, and physical health are intertwined. In the Huang Di Nei Jing, the emperor's medical adviser, Qi Po, outlined the Internal Factors (emotions) as primary causes of disease.

Within the context of Traditional Chinese Medicine, Medical Qigong therapy is useful for treating many types of psycho-emotional disorders. Qi emission therapy can be used to calm the patients' Shen, as well as accelerate the patients' recovery time. In China, psycho-emotional problems are commonly called Jing Shen disorders and are differentiated as follows: Obstructed Shen, Unsettled Shen, and Weakened Shen.

1. Obstructed Shen is a condition due to the Shen becoming obstructed with Phlegm, or in mild cases, becoming obstructed by Qi or Blood. Symptoms are characterized by confused thinking, clouding of the mind, and in severe cases, complete loss of touch with reality. Degrees of severity range from very mild disorders to severe schizophrenia or mania.
2. Unsettled Shen is a condition due to the Shen becoming unsettled from Blood or Yin Deficiency (in mild cases), Qi stagnation, Blood Stasis, Fire, Empty Heat, Phlegm Fire, or Internal Wind. Symptoms are characterized by agitation, restlessness, and anxiety.
3. Weakened Shen is a condition due to Deficiency of Qi, Yang, Blood, or Yin. Symptoms are characterized by depression, mental exhaustion, and melancholy (dysthymia).

ETIOLOGY

The cause for psycho-emotional disorders can range from a variety of factors, including prenatal or postnatal constitutional imbalance, diet, unbalanced sexual activity, overwork, alcohol or drug abuse, trauma, stress, emotional immaturity, and instability, as well as a lack of love or spiritual perspective.

TREATMENT PRINCIPLES

There is a considerable degree of overlap among Jing Shen disorder symptoms. Therefore, the treatment principle must be based on a clear distinction between Excess and Deficiency, as well as the cause (root) and manifestation (branch) of the symptoms. The main treatment principles are as follows:

1. Nourish the Heart and calm the Shen to treat Shen weakness due to Qi, Yin, Yang, Blood, or Body Fluids Deficiency.
2. Purge the Excess, move stagnation, and calm the Shen to treat Shen obstruction or Shen agitation (unsettled Shen) due to Excess conditions of Qi, Blood, Phlegm, and/or Fire stagnation.
3. Purge the Excess, nourish the Heart and calm the Shen to treat Shen obstruction due to Yin Deficiency with Empty Heat.
4. Resolve the Phlegm, open the Heart's orifices, and calm the Shen to treat Shen obstruction due to Phlegm or Phlegm Fire.
5. Sink and calm the Shen to treat unsettled Shen due to rising Qi.

The term "calm the Shen" is used to describe the treatment principles in all psycho-emotional disorders. It should be understood to mean not only calming the mind, as in anxiety, but also lifting the patient's mood (as in depression), or clearing the mind (as in schizophrenia or mania).

SELF-TREATMENT PRESCRIPTIONS FOR PSYCHO-EMOTIONAL DISORDERS

Most meditations are difficult for patients with psycho-emotional disorders due to several factors. One factor is the problem of maintaining focused attention, as well as retaining the memory of specific medical Qigong prescriptions. The success, or failure of the treatment will also depend on the severity of the disorder.

POINT MEDITATION FOR EMOTIONAL TRAUMA

In China, Channel Point meditation is used in conjunction with the Massage Tapping method for treating patients with conditions such as anxieties, phobias, and addictions. In acupuncture, 135 of the original 365 acupuncture points are used to treat mental, emotional, energetic, and spiritual disharmonies. In fact, many of the "new" points can be applied for the treatment of Jing Shen disorders.

The cause of fear and anxiety is the disruption to the body's energy system from emotional trauma. When a patient is focusing on a particular anxiety or fear, the Shen becomes disrupted, and the energetic harmony of the body becomes unbalanced. This energetic imbalance causes certain physical reactions within the patient (abnormal breathing, pulse, etc.), and results in pathological behaviors.

TREATING EMOTIONAL ANXIETIES, PHOBIAS, AND ADDICTIONS

To treat anxieties, phobias, and addictions, the doctor encourages the patient to focus attention on the specific fears, or cravings for the particular addiction, to initiate a maximum emotional charge of distress. When the patient reaches maximum emotional distress, the doctor calls out specific channel points for the patient to tap.

The patient is given a specific mantra to repeat while tapping, to discharge the accumulated emotionally charged Qi, such as; "I accept this fear (or other emotion or craving)" or "I am being healed." The treatment continues for five to ten minutes, as the patient focuses on feeling the distress as much as possible, until all the trapped emotion has been discharged. Some energetic points commonly chosen for treating emotional anxieties, phobias, and addictions are GV-26, Ht-7, and Pc-6. These specific points are used to draw the energetic charge away from the Heart and into the extremities; this action facilitates calming the patient's Shen.

The initial tapping method, in conjunction with vivid imagination and positive affirmation, disperses the energy away from the Heart and other organs and calms the Shen. The spontaneous dispersing of the emotional charge, stored within the patient's tissues, releases all feelings of stress and anxiety. This type of guided Tapping Meditation therapy not only reduces and disperses the patient's emotional trauma, but also causes an energetic reprogramming within the patient's psychophysical dynamics.

MEDICAL QIGONG TREATMENT OF THE PHOBIA AND ANXIETY OF DEMON POSSESSION

These specific treatments are used in China to treat specific mental, and emotional disorders, that lead the patient to imagine that they are being haunted by spirit demons. In rectifying this type of psycho-emotional disorder, it is recommended that the Qigong doctor proceed as follows.

1. Perform Channel Point therapy to open the energetic points of the Eight Extraordinary Vessels (Figure 53.11).
2. Press and knead the Baihui GV-20, Dazhui GV-14, Lingtai GV-10, and Feishu Bl-13 points. Then apply the Extended Fan Palm or Sword Fingers method (see Chapter 33), using the Pulling and Shaking manipulations (see Chapter 34) to emit and conduct Qi along the Governing Vessel.
3. Press and knead the following channel points Baihui GV-20, Yintang (Ex.Pt.), Renzhong GV-26, Tinggong SI-19, Jiachè St-6, Quchi LI-11, Hegu LI-4, Weizhong Bl-40, and Chengshan Bl-57 points.
4. Apply the Sword Finger technique, using the Vibrating and Shaking manipulations to emit Qi into the Jiuwei CV-15 and Zhongwan CV-12 points and conduct Qi along the Conception Vessel to return it to the Lower Dantian.

MEDICAL QIGONG TREATMENT FOR SPECIFIC SPIRIT/GHOST HALLUCINATIONS

These specific treatments are also used in

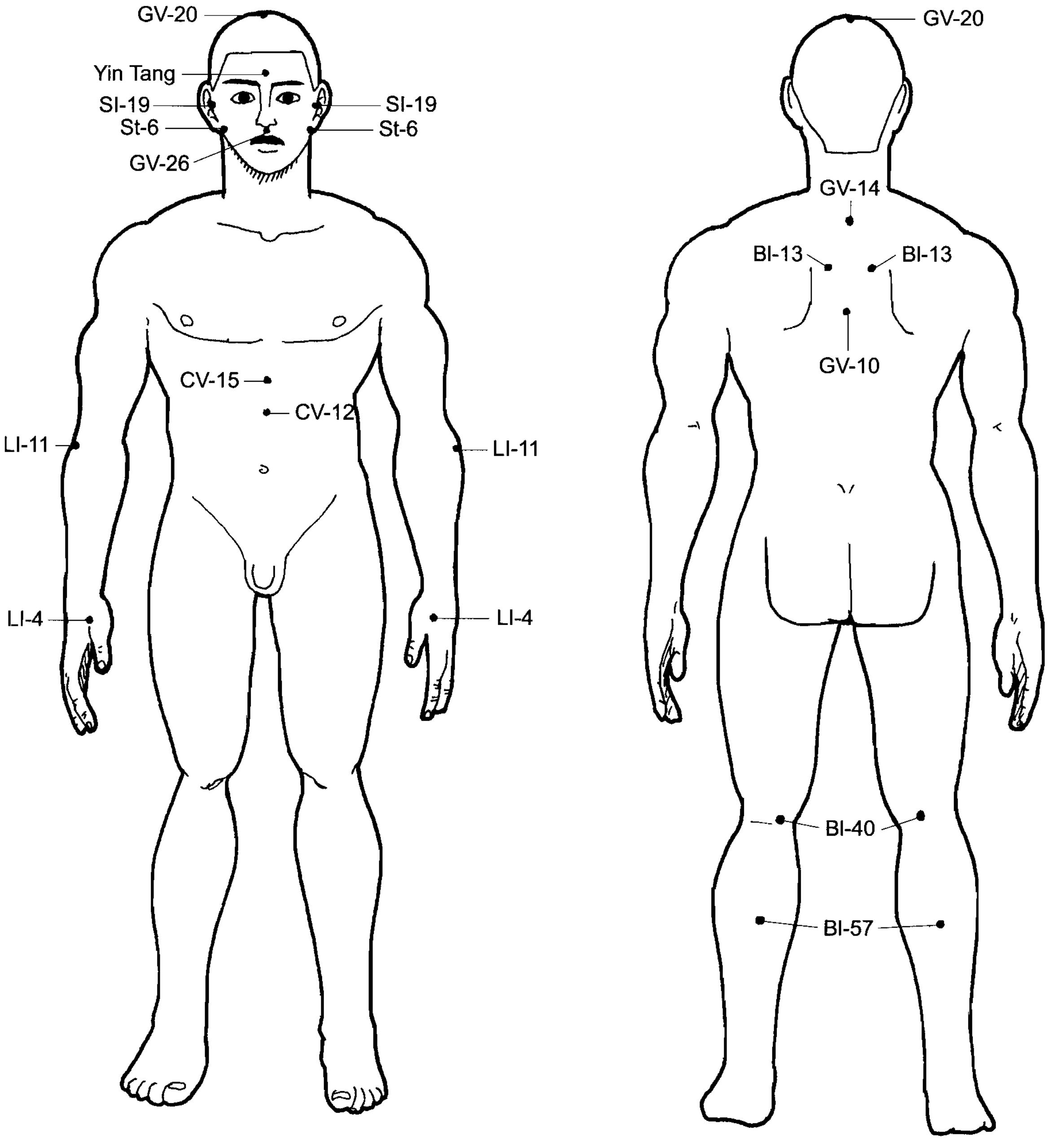
GV-20
Yin Tang
SI-19
St-6
GV-26
SI-19
St-6
CV-15
CV-12
LI-11
LI-11
LI-4
LI-4
GV-20
GV-14
Bl-13
Bl-13
GV-10
Bl-40
Bl-57

Figure 53.11. Medical Qigong Channel Point Therapy for the Treatment of Energetic Delusions

China to treat specific mental and emotional disorders that lead the patient to see spirit/ghost hallucinations. The Qigong doctor may additionally treat the patient according to the following therapeutic point stimulation:

- If the patient is disoriented, sees and hears floating corpse ghosts, or hears crying ghosts, treat the Tianfu Lu-3 points.
- If the patient is manic-depressed, sees and hears ghosts talking, treat the Tianchuang SI-16 points.
- If the patient is having seizures, sees and hears ghosts talking, treat the Tianzhu Bl-10 points.
- If the patient is manic, suicidal with incessant wild talking and actions, treat the Fengfu GV-16 point.
- If the patient experiences chronic nightmares, confused dreaming, and restless sleep, treat the Tianyou TB-16 points.

MEDICAL QIGONG TREATMENT FOR DEMON OR SPIRIT POSSESSION

Since the creation of Traditional Chinese Medicine in the People's Republic of China, the notion of spirits has been down-played. However, within the confines of the Medical Qigong clinics, I have been fortunate enough to be able to discuss with several Qigong doctors the cause, effect, and treatment of demon and spirit possession.

These following Thirteen Ghost Points are listed in the *Thousand Ducat Formulas*, prescribed by the famous physician Dr. Sun Simiao in the 7th century China, and were once used for the treatment of spirit possession (Figure 53.12). Today in China, these points are now used in the TCM clinics for the treatment of manic and depressive mental disorders, as well as for treating epilepsy. In extreme manic cases (with Heat) the doctor bleeds all of these points as well as all twelve Jing-Well points on the patient's hands. Classically, the patient is treated by bleeding each of these points, or by inserting an acupuncture needle into each point and then immediately withdrawing it. These Ghost Points are stimulated in a sequence of progressive point therapy, wrapping and enfolding the patient's body in five consecutive circles.

1. The Guigong (Ghost Palace), also known as the Renzhong GV-26 (Middle of Man) point, is the first point to be stimulated and is treated to calm the Shen and clear the brain.
2. The Guixin (Ghost Faith), also known as the Shaoshang Lu-11 (Lesser Metal's Note) points on both thumbs, are treated to calm the Shen and restore Collapsed Yang.
3. The Gullei (Ghost Fortress), also known as the Yinbai Sp-1 (Hidden Clarity) points on both big toes, are treated to calm the Shen, clear the brain, in order to instill clarity of thought and mind.

These first three points are used to complete the energetic pass of the first circle, covering the head, tips of the thumbs and tips of the big toes. The GV-26 point is considered the meeting point of Yin and Yang in the body, while the Lu-11 points on both hands and the Sp-1 points on both feet are located on the body's Hand and Foot Tai Yin Channels.

Next, the following two points are used to complete the second circle:

4. The Guixin (Ghost Heart), also known as the Daling Pc-7 (Big Mound) points on both wrists, are treated to calm the Shen and clear the brain.
5. The Guilu (Ghost Path), also known as the Shenmai Bl-62 (Extending Vessel) points under both outside ankles, are treated to calm the Shen.

The Ghost Heart and Ghost Path points complete the energetic pass of the second circle, covering and wrapping the patient's wrists and ankles. The Pc-7 points on the wrists are located on the body's Hand Jue Yin, and the Bl-62 points on the outside of the ankles are located on the body's Foot Tai Yang Channel points.

The following set of four points are stimulated to complete the energetic pass of the third circle:

6. The Guizhen (Ghost Pillow), also known as the Fengfu GV-16 (Wind's Palace) point just below the occiput, is treated to clear the brain.
7. The Guichuang (Ghost's Bed), also known as the Jiache St-6 (Jaw Vehicle) points on both sides of the jaw, are treated to dispel Wind and Cold and to clear Heat.
8. The Guishi (Ghost Market), also known as the Chengqiang CV-24 (Receiving Liquid) point under the lower lip, is stimulated to treat mental disorders, dispel Wind and Cold, clear

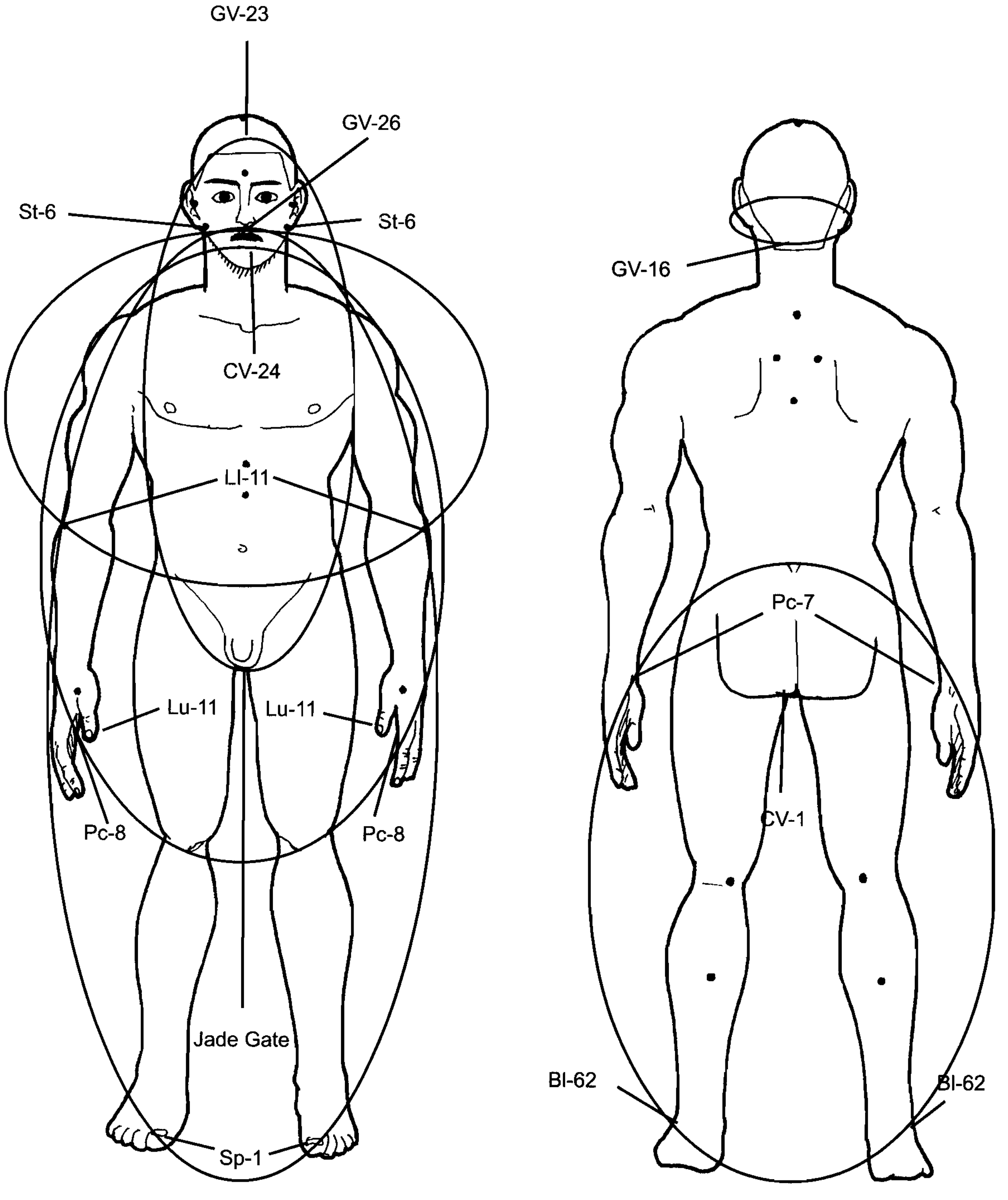
GV-23
GV-26
St-6
St-6
CV-24
LI-11
Lu-11
Lu-11
Pc-8
Pc-8
Jade Gate
Sp-1
GV-16
Pc-7
CV-1
Bl-62
Bl-62

Figure 53.12. Medical Qigong Treatment for Spirit Possession

Heat, and to transform Dampness and Phlegm.

9. The Guicu (Ghost Cave), also known as the Laogong Pc-8 (Palace of Labor) points at the center of both palms, are stimulated to treat mental disorders and to clear the brain.

These four points complete the energetic pass of the third circle, by covering and wrapping the patient's head, then drawing the energy into the center of the patient's palms. The GV-16 point on the back of the patient's head is located on the back gate of the Upper Dantian, the St-6 points are located on the sides of the mandible, and the CV-24 point is located in the depression in the center of the patient's mentolabial groove (jaw).

The next two points complete the energetic pass of the fourth circle:

10. The Guitang (Ghost Hall), also known as the Shangxing GV-23 (Upper Star) point, is treated to brighten the Shen and clear Heat.
11. The Guicang (Ghost Store), also known as the Yumentou (Jade Gate) in women, and Yinxiafeng (Below the Hidden Seam) in men, is used to treat mental disorders. The Jade Gate is located at the opening of the vagina, the Below the Hidden Seam is located in the Huiyin area in men.

These two points complete the energetic pass of the fourth circle, by covering and wrapping the patient's body from the head to the lower perineum. The GV-23 point is located on the anterior hairline, at the top of the patient's forehead, and the Ghost Store point is located in the lower perineum.

The following two points are stimulated to complete the fifth energetic pass:

12. The Guitui (Ghost Leg), also known as the Quchi LI-11 (Crooked Pool) points, are treated in order to drain the pathogenic influences from the Lungs and to calm the Po (Corporeal Soul).
13. The Guifeng (Ghost Seal), also known as the Haiquan (Sea Spring) and She Xia Zhong Feng (Under Tongue Middle Frenulum) are two Extra points located below the tongue. They are stimulated in order to treat mental disorders.

These final two points complete the energetic pass of the fifth circle, by covering and wrapping the patient's body from the arms to the head. The LI-11 points are located in the depression on the lateral end of the elbow crease, and the Haiquan Sea Spring points are located under the tongue.

TREATING PHOBIAS AND DEPRESSION USING "THE WINDOWS OF THE SKY" POINTS

If we observe the body as a figure 8, with the head being the upper portion and the torso being the lower portion, the neck would represent the area where the two circles meet. The points and orifices of the head, face and neck are like great windows in which Qi moves.

The trunk of the body is considered Earth and the realm of the physical; the head pertains to Heaven and the realm of the spirit. The neck is where all of the classic "window" points are located. The neck is the junction between Heaven and Earth, and is considered the Earth's window to Heaven. With the exception of Lu-3 (upper arm) and Pc-1 (chest), eight of the ten Windows of the Sky points are located on the neck itself, and seven of the ten points have the word Tian (Heaven) as part of their name.

Over the last quarter century, the Chinese have used the Windows of the Sky points in treating such psychological disorders as depression and phobias. In the Qigong clinic, we look for three indications before prescribing the Windows to the Sky points.

1. According to the Nei Jing, the Windows of the Sky points are used when the Yang Qi is unable to ascend to the patient's head. The result is that the patient will have physical symptoms in the head or Heart (i.e., loss of hearing, or pain described as being behind the Heart.
2. The second indication is that there is something missing in the Heavenly aspect of the patient's spiritual life. When the patient becomes stuck in their own negative patterns, these particular points can be used to open a window of hope by bringing into the patient's body the illuminating light from Heaven.

   In these cases, the patient's disconnection from emotions and spiritual insight is due to the chronic negative patterns that block the descent of Heavenly light.

3. The third indication is a constant attitude of despondency by the patient. This behavior is not in accordance with his or her own values and beliefs, but the patient cannot control this pattern. Some patients know that something is amiss but cannot discern the source of the problem.

The Window points induce a hypnagogic state (a transitional state between sleeping and awaking) that allows some patients to experience out-of-body projection and past life regressions.

TREATMENT

When treating patients with the Windows of the Sky points, the Qigong doctor should use slow, gentle pressure on the skin to dip into the tissues, and then proceed with pure intention. The ten Windows of the Sky points are described as follows.

1. The Tian Fu Lu-3 (Heaven's Residence) point is sometimes called the Celestial Palace. The word Heaven represents divinity, clarity of vision, and the sense of the spirit (Figure 53.13). This point has a powerful psychological effect on all emotional problems deriving from Lung Qi disharmonies (e.g., depression, claustrophobia, agoraphobia, mental confusion, and forgetfulness).
2. The Fu Tu LI-18 (Support the Prominence) point is also known as Support and Rush Out. It is used on patients who feel out of touch with their own inner strength (Figure 53.14). This point relieves coughing, resolves Phlegm, and disperses masses (cysts and tumors) in the neck area. It is also used to treat a patient's inability to speak.
3. The Tian You TB-16 (Heaven's Opening) point can be used to calm the patient of anxieties and to soothe Heart palpitations (Figure 53.15). This point is also used for dizziness, sudden deafness, and diminished visual acuity.
4. Tian Zhu Bl-10 (Heaven's Pillar) is a Sea of Qi point. It is used to clear the brain, open the orifices, brighten the patient's eyes, and re-

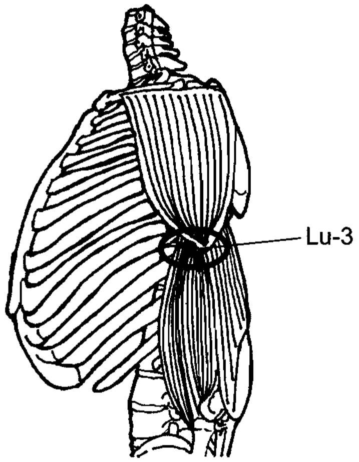

Figure 53.13. Heaven's Residence

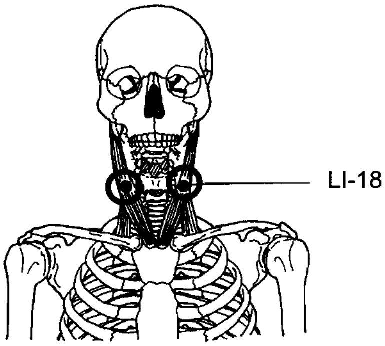

Figure 53.14. Support the Prominence

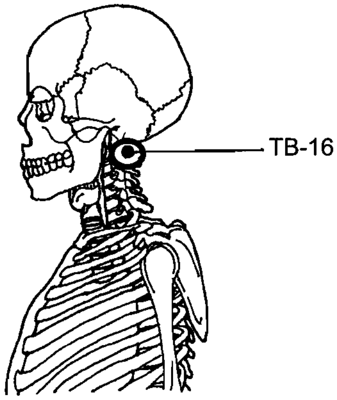

Figure 53.15. Heaven's Opening

move obstructions from the channels. The energy that moves and is located at these points has to do with the patient's will, ambition, and inner fortitude (Figure 53.16). Energetic dysfunction located at the Heaven's Pillar may manifest as deep sadness or anxiety, and be related to a lack of inner strength, or will. This point is also used when the patient's feet are unable to support the body.

5. Ren Ying St-9 (Man's Welcome) is a Sea of Qi point. The ancient Chinese philosophers associated this area of the body with Man in the triad of Heaven, Man, and Earth. This point is known as a place where energy is being welcomed into the body (Figure 53.17). It is a very sedating point of Blood and Qi, and is used to reduce blood pressure. Opening up this point allows the patient to become receptive to nurturing energy (Ying Qi). The doctor should be gentle and only treat one side at a time. It is also used to disperse Lung Qi and Regulate Qi.
6. The Tian Tu CV-22 (Heaven's Rushing Out) point is also known as the Celestial Chimney, and is an area where the body's internal energy and spirit leaves the body through sound (Figure 53.18). Heaven Rushing Out is related to emotional and spiritual expression, and is relevant when the patient's spiritual self (i.e., the individual's life purpose) is not being supported by physical actions (his or her code of conduct). In other words, the patient's life purpose is not in alignment with the patient's behavior. This point can also be used to regulate the Lungs, stimulate the descending action of the Lung Qi, clear Heat, and resolve Phlegm.
7. Feng Fu GV-16 (Wind's Palace) is a Sea of Marrow point, and is also a very good point for strengthening the body's Wei Qi (Figure 53.19). This point connects directly to the brain, and is used to eliminate Wind, clear the mind, open the sensory orifices, and benefit the energetic function of the brain (used for treating convulsions, epilepsy, or palsy).
8. The Tian Chuang SI-16 (Heaven's Window)

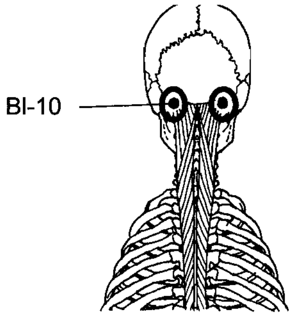

Figure 53.16. Heaven's Pillar

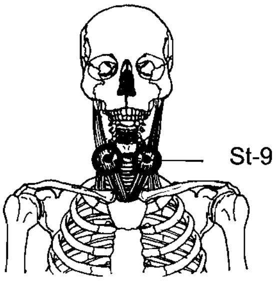

Figure 53.17. Man's Welcome

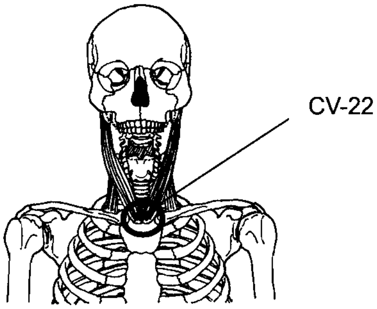

Figure 53.18. Heaven's Rushing Out

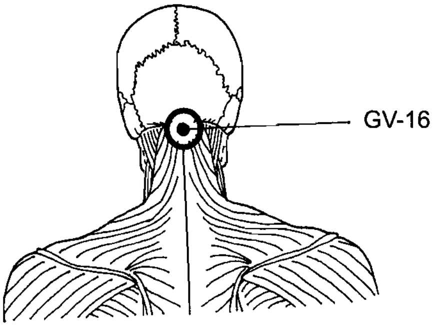

Figure 53.19. Wind's Palace

point is used to open a window to Heaven, especially in patients who need help in perceiving how to change a negative behavior pattern into a productive and harmonious one (Figure 53.20). This point is used for dispersing the Liver's Qi and for eliminating buzzing, pain, or deafness of the ears.

9. The Tian Rong SI-17 (Heaven's Appearance) point is also known as the Celestial Countenance. This point is where the Small Intestine's Channel Qi is received by, and enters into, the Heaven's Body Chamber (the head or skull). It can be used to produce a strong movement of energy from the head downward into the body.

   This point is used in spiritual emergencies, when patients have lost their connection with Earth energy, and remain blissed out, and oblivious of reality. It is often used in cases of addictions, once the patient is conscious of the addiction (Figure 53.21). This point is also used in order to expel Excess Fire and remove obstructions from the patient's channels.

10. The Tian Chi Pc-1 (Heaven's Pond) point is also known as the Celestial Pool. The term pond represents a source, resource, or reservoir of energy for the spirit. This point influences the area where the Shen gathers or pools in the chest region (Figure 53.22). It is capable of enhancing the transition from normal consciousness to feelings of being submerged in a sea of energy.

    It affects the Heart and Pericardium which govern the Shen, and whose symbolic home is Heaven. The doctor uses this point to remove the patient's street armor, and to begin the process of opening the patient's Heart protector at its source. This point is also used to disperse the Lungs' Qi, expand and relax the chest, regulate Qi, and clear Heat.

    Note: Because of the proximity of the breast tissue, the Heavenly Spring points (Pc-2) are sometimes used in women instead of the Heaven's Pond points (Figure 53.23). The Heavenly Spring point is located halfway down the anterior face of the biceps muscle.

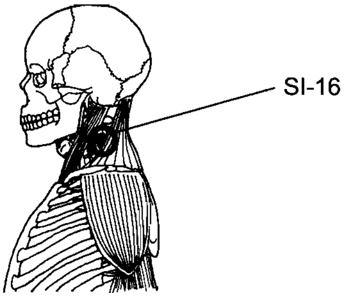

Figure 53.20. Heaven's Window

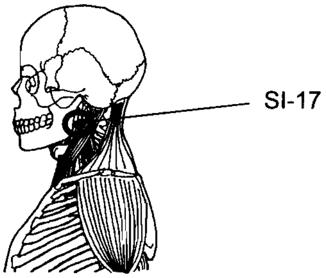

Figure 53.21. Heaven's Appearance

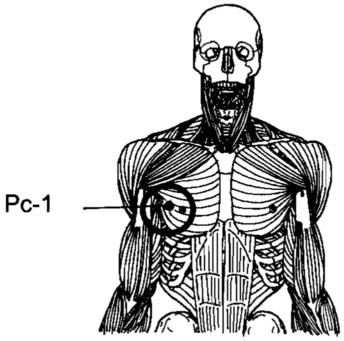

Figure 53.22. Heaven's Pond

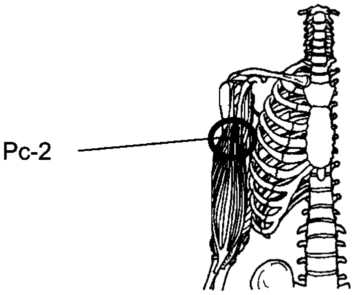

Figure 53.23. Heavenly Spring

TREATING PSYCHO-EMOTIONAL DISORDERS RELATED TO STRESS

The following prescriptions are used in order to assist the patient in dealing with suppressed anger and grief, for additional prescriptions used in order to treat emotional problems (see Chapter 19).

TREATING EXCESS LIVER FIRE AND THE REPRESSION OF ANGER

One of the most common problems observed in the clinic is Excess Liver Fire. One of the many causes of Excess Liver Fire is the repression of anger and a suppression of the spiritual components of the Hun. When anger is suppressed, its volatile Heat will sometimes attack and drain the Kidney Yin (Water). Therefore, Excess Liver Fire is generally related to a Deficiency of the Kidney Yin energy. When Qigong doctors reduce the Excess Liver Heat, they must also increase the patient's supply of Kidney Yin (Water) Qi (see note #2).

If the Liver's Yin becomes Deficient, the body's Qi will not be able to keep the patient's Internal Fire from rising. This causes the Yang to rise to the patient's head, and the Blood to become Deficient. The patient becomes irritable as the Liver's Yang begins to rise. This in turn, causes high blood pressure, which is associated with such problems as headaches, dizziness, insomnia, tinnitus, hypertension, and migraine headaches. Deficient Liver Yin, if unchecked, turns into Liver Fire, which causes severe high blood pressure, and heats the patient's Blood. As the Blood heats, it becomes Deficient, and causes Liver Wind. Unchecked Deficient Liver Yin results in shaking and trembling. If the Liver Wind mixes with the patient's Phlegm, it blocks the patient's channels resulting in paralysis, aphasia, memory problems, and eventually stroke.

The method of treatment is chosen according to the patient's syndromes:

1. For patients sensitive to energy, use Medical Qigong therapy while the patient is lying supine.
    - Purge and disperse Qi from the patient's Liver;
    - Pull pathogenic Qi out the patient's Liver Channels;
    - Tonify the patient's Kidneys through the Lower Dantian;
    - Give Liver purging and Kidney tonification prescriptions and homework.
2. For patients not sensitive to energy, use Tui Na and An Mo in conjunction with Medical Qigong therapy while the patient is lying prone.
    - Purge and reduce the Fire in the patient's Governing Vessel and Bladder Channels;
    - Start at the Mingmen and dredge the patient's Liver Fire, ending at the sacrum, using a counterclockwise circle rotation to reduce the Liver's Qi;
    - Give Liver purging and Kidney tonification prescriptions and homework.

Note #1. If the Liver Fire attacks the Stomach, resulting in vomiting and migraine headaches, have the patient lie supine. Because the Liver's Fire can travel transversely along the hypochondrium area, the Qigong doctor rocks and gently shakes the patient from the center of the thorax down to the Kidneys using intention, lead and disperse the Heat out from the patient's torso. Next, pull the Stomach Qi down and out of the patient's feet.

Note #2. Tonify the patient's Kidneys by using Jing Point therapy on the following points: St-36, GB-34, Sp-6, Lv-3, and Kd-1 on both sides of the body.

BEATING THE BAG AND SHOUTING TO DETOXIFY LIVER HEAT AND RELEASE TRAPPED ANGER AND RAGE

Beating the Bag and Shouting is the most common Medical Qigong homework given to patients with Excessive Liver Fire. The concept of Beating the Bag and Shouting is best comprehended when the function and flow of Liver Qi Heat is understood. When striking the bag and shouting, the shaking movements of the body will:

- facilitate an energetic discharge from the Liver organ,
- help alleviate pressure stored within the Liver organ itself (detoxifying suppressed feelings), and
- purge the body of Excess Liver Heat.

It is important for the patients to understand, that when shouting and striking on the surface of the bag (while focusing their intention deep into the bag's center), they are releasing deep seated pathogenic rage. This is why it is prohibited for patients to substitute such things as pillows, bed, etc., for the bag. If a patient takes a pillow and pounds it, for example, screaming and shouting to release the aggression, he or she is releasing and discharging pathogenic Qi into the pillow. If after releasing this Toxic Qi into the pillow, it is returned to the bed for a person to sleep on, the patient will reabsorb the discharged energy. To avoid this occurrence, patients are encouraged to set aside a space where they can safely discharge internal pathogenic Qi (preferably outdoors). This is similar to isolating and storing garbage away from living quarters.

While striking the bag, in addition to feelings of anger and rage, sometimes the patient may start crying or experience overwhelming feelings of grief and despair. This is also a normal reaction to dispersing toxic emotion. It sometimes occurs while discharging Stagnant Liver Qi, because the Lungs, which store the emotion of grief, have the function of keeping the Liver organ in check, in accordance with the Five Elemental Controlling Cycle.

What the patient needs to perform this exercise is two wooden dowels (1–1.5 inches in diameter and three feet long), a large bag of rice from the grocery store, and some duct tape. Completely encase the bag of rice with the duct tape to prevent spilling. Place the bag on a table, or a hard surface, level with the Lower Dantian. Hold a dowel in each hand, strike the bag, and focus on the Liver's Qi traveling up the back, down the arms, and through the wooden dowels.

With each impact, imagine the body releasing Toxic Liver Qi into the bag. It is important for the patient to exhale completely and deeply from the lower abdomen when shouting and striking the bag. Sometimes the patient is encouraged to shout "no" or "stop." These words will help the patient to facilitate complete emotional discharging, while also reestablishing a patient's broken energetic/emotional boundary system.

Bag striking should be performed for 10–15 minutes to achieve maximum benefit. Afterwards patients may feel euphoric and exhausted due to the heightened endorphin activity now occurring within the tissues. If the patient still feels angry, he or she should try writing out the rest of the rage on paper, which is then discarded. In order to end the exercise, patients should sit for five minutes in quiet solitude. It is important for the patient to reflect on any emotional memories which have been released and brought back into consciousness. The patient may then begin sharing these feelings and memories with the doctor, and commence the process work for healing the unexpressed emotions.

Some patients may become absolutely horrified at themselves for getting so angry during bag beating that they feel terrified afterwards; especially if so much rage was built up that is seems there is no end to it. Often patients are forced to stop from overwhelming fatigue, aches, and pain caused by this purging exercise. If the patient allows the rage to build and never completely processes the rage, it may cause the patient to feel that without the anger he or she would literally cease to exist, or that he or she is innately evil. The patient may feel that he or she "is" the anger, and that the exercise is making it worse, and may strive to control the anger even more than before.

Some patients will get so mad, they will walk out on the Qigong doctor without processing anything. This unprocessed rage can be extremely dangerous, especially if the patient is prone to beating his or her mate, kids, or pets. Many people are completely unaware of how much anger they have suppressed. Some patients are like bottomless wells, especially patients with borderline personality disorder. Patients with this disorder should never be given this exercise, as it could lead to self-mutilation. Refer such patients out for psychotherapy.

In our culture, women are not allowed to express anger, much less rage. If the patient has ever been beaten, or witnessed a loved one being abused, he or she may believe that he or she is just like the abuser, and that he or she has been lying to him or herself about not being bad and may become suicidal.

The Qigong doctor should never prescribe this exercise unless he or she fully understands the complexity of the patient's circumstances. Some patients require a period of time, in a safe place, with supervision to help them deal with such intense emotions.

Other possible reactions might be that the patient goes into stronger denial of the rage, blames the doctor for causing him or her to beat their mate or children, or blames the doctor for causing the rage. The Qigong doctor must make sure there is sufficient time for the patient to process all the reactions aroused. The patient must not feel hurried.

ANGER MANAGEMENT

With some patients it is best to start them off journalizing the anger, not in a dairy, but on sheets of paper. They address the letters to the person, or persons, with whom they are angry. They continue writing until all the anger has been discharged. They are not to mail the letters, or give them to the addressee, or show them to anyone including the Qigong doctor. In writing the letter, they allow themselves to fully express the anger without judging themselves. The patient should not be concerned about grammar, foul language, etc. No matter how trivial the upsetting incident was, it should be written down. The flow of consciousness is initiated and continued until the end of the letter. No event (no matter how silly or insignificant) that comes to mind should be left unexpressed.

When finished, the letter is to be torn up, or safely burned in a ritual of healing and letting go (e.g., light some incense or candles, then blow them out after the paper has been incinerated). Symbolically the patient must let go of the anger for good, and that is why the letter is destroyed in a ritualistic act.

Only later, as the patient continues to practice this purging technique, slowly gaining control of the emotions, can the angry letters be kept for a few days and reread. In rereading the letters the patient engages his or her discriminating, judging faculties (i.e., "Was I right to get upset over this incident? Did I over-react? What did this incident remind me of from my childhood?"). This is a safe and very powerful method of discharging anger from the patient's body. The patient is encouraged and allowed to say everything he or she thinks of when writing, holding nothing back. It is nonthreatening, and allows the patient to come to understand him or herself better in time.

TREATING STAGNANT LUNG QI AND THE REPRESSION OF GRIEF

The Lungs dominate the Qi and respiration. A common problem observed in the clinic is stagnant, or obstructed Qi within the Lungs, from the toxic buildup of sorrow, grief, and worry. Sadness and sorrow can cause a Deficiency of Lung Qi, while worry can cause the Qi in the chest to stagnate. Stagnant Qi within the Lungs is usually directly or indirectly related to the suppression of emotional issues.

The right Lung is considered more Yang than the left Lung; it therefore tends to reflect, and sometimes manifests physically, any unresolved, or suppressed emotional conflicts experienced within the father-child relationship (or other main male figures in the patient's childhood). These unresolved emotional issues tend to carry over and help shape the nature of subsequent conflicts experienced in other male relationships. These ongoing emotional reactions provide a further opportunity in life to express, work through, and resolve these emotional issues.

The left Lung is considered more Yin and thus reflects emotional issues stemming from the mother-child relationship (or main female figure in childhood). The left Lung may likewise manifest these issues in subsequent female relationships.

Generally, cysts, tumors, or cancer originating in these corresponding regions reflect the emotional history. Because of these emotional patternings, the method of treatment for Lung diseases (that are Internally induced) is chosen according to the patient's syndromes in relationship to the root of the original trauma. After determining the specific Lung which dominates and manifests the patient's pathogenic condition, the Qigong doctor proceeds as follows.

1. For patients sensitive to energy, use Medical Qigong therapy while the patient is supine.
   - Disperse Qi over the patient's throat and Lungs;
   - Purge pathogenic Qi out the afflicted Lung Channel;
   - Tonify the patient's Spleen and Kidneys;
   - Prescribe Homework.
2. For patient's not sensitive to energy, use Medical Qigong therapy in conjunction with An Mo Qi therapy while the patient is supine.
   - Disperse Qi over the patient's throat and Lung area;
   - Purge pathogenic Qi out the afflicted Lung Channel;
   - Dredge Fire from the patient's Liver;
   - Tonify the patient's Spleen and Kidneys;
   - Prescribe Homework.

DRY CRYING TO DETOXIFY STAGNANT LUNG QI AND RELEASE TRAPPED SORROW AND GRIEF

The most common homework prescription given to patients with stagnant Lung Qi is the Dry Crying exercise. The concept of "dry crying" is best understood if we look back to when we were little children. When little children cry hard, they generally release full body resonances. They shake and undulate from the middle of the body, allowing the pain to resonate outwards. When children reach puberty, they learn (through parental or peer pressure) to suppress their feelings, and to restrain full body resonance, by tensing the muscles in the throat. The constant pressures from society, parents and peers, result in self-conscious feelings and shame regarding the loss of control over the emotions. The following exercise allows the patient to experience full body resonation and to release deeply suppressed emotions.

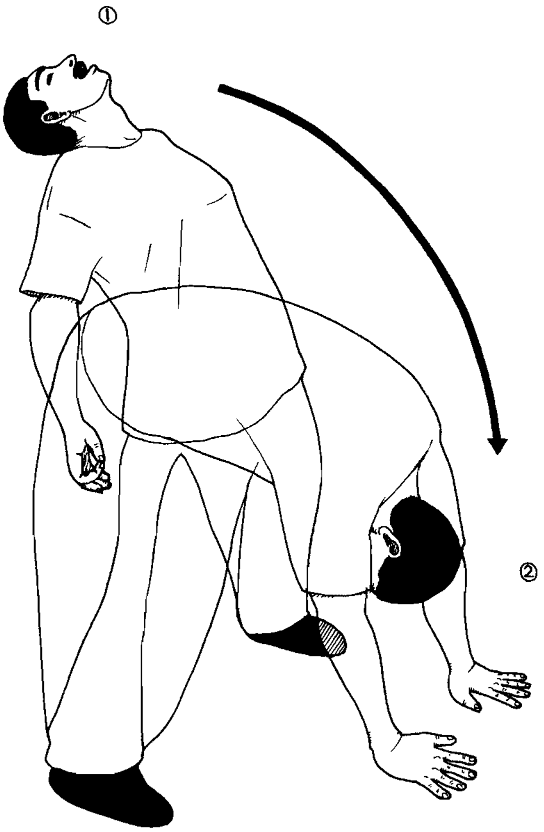

Figure 53.24. Dry Crying

Begin the exercise from a standing position. Stand with the feet more than shoulders width apart, toes pointing straight and the knees comfortably bent. With each hand, touch the thumb to the index finger to form a circle. Relax the arms letting them hang naturally. Close the anal sphincter, and breath naturally from the abdomen. Tilt the head backwards, opening the mouth wide, and relax the jaw. Allow the upper body to lean and bow backwards (without straining the spine), but stretching the psoas muscles. Hold this posture from 5 to 15 minutes, concentrating on relaxing the body, and allowing it to resonate (Figure 53.24).

The psoas muscles are located in the lower back and Mingmen area. They articulate from the transverse process of the lower lumbar vertebra. These muscles travel near the center of the abdomen down to the inner thigh. These particular muscles are associated with the storing and releasing of powerful emotions. As the psoas muscles stretch, the body begins to resonate and shake from the pelvic cavity upward. This action returns patients back to the preadolescent state of energy release, allowing them to detoxify deep seated grief within the tissues.

In order to complete the exercise, it is extremely important to ground. This prevents the uprooted toxic emotions from causing Qi stagnations in the throat and chest area. After holding the posture for the prescribed amount of time, slowly lean forward and tilt the chin down. Allow the arms to continue to hang naturally. Slowly lean forward until the hands touch the floor. Imagine Qi, flowing like water, from the torso out your arms, cleansing your body, sending the pathogenic Qi deep into the center of the Earth.

CONTRAINDICATIONS

Patients should avoid this particular exercise if they have previously injured their back. For patients who have difficulty with the stretching of the cervical or lumbar vertebrae, it is encouraged that they wear a cervical collar or lumbar support to prevent injury to the spine. Also, the exercise can be modified to allow the patient to sit comfortably on the edge of a chair while stretching the psoas muscles if the standing stretch is too difficult.

CHAPTER 54
QI EMISSION THERAPY FOR ONCOLOGY

HISTORY OF CANCER TREATMENTS

The ancient doctors of Traditional Chinese Medicine understood about the formation of tumors. The earliest records found were inscriptions recorded on "Oracles of Bone," unearthed from the ruins of the Yin Dynasty (16th–11th century B.C.). The word "ai" (cancer) first appeared in the Chinese Medical classic *Wei Ji Bao Shu* during the Song Dynasty, in the year 1171.

In Traditional Chinese Medicine, cysts, tumors, and cancers are considered multifaceted diseases, originating from imbalances in one or more of the five following sources: environmental, chemical, biological, physical, and psychological. According to the *General Treatise on the Etiology and Symptomatology of Diseases*, a tumor is one of the many diseases that responds to Dao Yin (Medical Qigong) exercises and meditations. To prevent the formation of cysts, tumors, or cancer, and to increase the strength of the immune system, Qigong doctors encourage patients to monitor their diet, maintain emotional stability, increase physical exercises through Medical Qigong prescriptions, and avoid exposure to pollutants.

From a Western Medical perspective, tumors and cancer are likewise considered multifaceted diseases. They are generally viewed as being created and formed through a constant bombardment, attack, and weakening of the tissues, caused by the following three factors:

- A congenital tendency towards specific types of tumors and cancers (inherited weakness);
- Exposure to radiation, environmental pollutants, and chemical poisoning, and
- An inability to repair weakened DNA, leading to uncontrolled replication.

In Traditional Chinese Medicine, cancer cells are considered "infant bodies" which absorb nourishment from the host (mother) body for their own growth; thus they are called "pathogenic or evil embryos." Cancer cells are actually components of the "original" or "mother" body.

CATEGORIZATION OF TUMORS

In Chinese Medicine, tumors and cancers are commonly referred to as "a retention of mass." Retention has the meaning of stagnation. Tumors and cancer originate from flowing substances within the body such as Qi, Blood, Phlegm, and food. These substances stagnate when they are attacked by pathogenic factors, and they gradually take on the form of a lump as a result of excessive accumulation. If the lump subsides spontaneously, it is commonly referred to as a cyst, or benign tumor. Many benign cysts and tumors, however, persist indefinitely.

BENIGN TUMORS

A benign tumor grows slowly, and can spontaneously stop growing, or disappear by itself. A benign tumor also grows in expansive ways. Its margin is clearly defined with firm boundaries, and is easily manipulated through palpation. There is no temperature differentiation between the skin surrounding a benign tumor and normal skin tissue.

Often, no symptoms surround the internal location of a benign tumor. As the tumor grows in certain areas of the patient's body, however, it can disrupt the functional activity of an affected organ. Benign tumors have been known to cause bleeding and inflammation. If the tumor is located within the endocrine glands, it can cause an abnormal function within the gland itself. A tumor located within an endocrine gland does not usually endanger the patient's life. Benign tumors can, however, become life threatening if the location compresses specific areas of the patient's brain or vital organs.

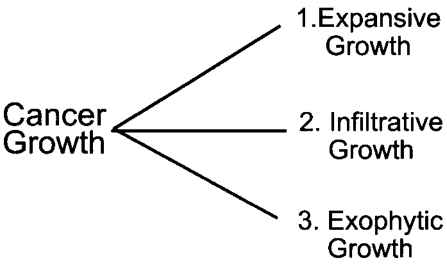

Figure 54.1. The Three Formations of Cancer Growth

MALIGNANT TUMORS

A malignant tumors grow very fast (potentially doubling every two months). Degeneration, festering, and hemorrhaging often accompany malignant tumors. A malignant tumor is infiltrative in its growth, with no clearly defined boundary. Its appearance is irregular and fastened (not easy to move through palpation). The skin temperature surrounding the malignant tumor is considered energetically Hot, as it is warmer than the normal surrounding tissue.

Most malignant tumors either metastasize to nearby, or distant areas of the body. The body's symptoms include fever, pain, anemia, fatigue, and general weakness. In many cases, if left untreated, they can cause death. Malignant tumors, and cancers can be divided into two main categories: a Carcinoma, and a Sarcoma.

1. A carcinoma is a malignant growth or tumor that occurs in the epithelial tissue (the outer surface or first layer of tissue that lines the body's cavities, as well as the principal tubes and passageways leading to the exterior of the body). Skin cancer, esophageal squamous cancer, adenocarcinoma of the Stomach, and breast cancer all belong to this type of carcinoma. A carcinoma may affect any organ, or part of the body, and may spread by direct extension, or indirectly through the lymphatic system and bloodstream.
2. A sarcoma is a malignant growth, or tumor, that occurs within the connective or mesenchymal tissue. It may affect the muscles, bones, fat, blood vessels, lymph system, Kidneys, Bladder, Liver, Lungs, Spleen, or parotid glands. Rhabdomyosarcoma (sarcoma of the muscles), osteosarcoma (sarcoma of the bones), and lymphosarcoma (sarcoma of the lymphatic system) all belong to this type of sarcoma.

THREE FORMATIONS OF CANCER GROWTH

According to Dr. Pan Mingji, Director of the Fuzhou Cancer Institute of Integration of Traditional Chinese and Western Medicine, there are three forms of cancer growth: Expansive Growth, Infiltrative Growth and Exophytic Growth (Figure 54.1).

1. Cancers of the Expansive Growth type consist of tumor cells crowded together on certain parts of the body. These cells proliferate, grow constantly, and expand towards the outside. The cells of the surrounding tissue form a wall to resist the tumor's expansion. The wall becomes the external membrane of the tumor, which has an evident boundary separating it from the normal tissue. Tumor growth begins in this way in the early stages of both benign tumors and malignant sarcomas.
2. Cancers of the Infiltrative Growth type consist of cells that not only group together, but, also separately invade the intercellular spaces, and tubular tissue of normal cells. They establish their foundation within the tissues, and proliferate daily, expanding outward with no boundary system between the cancer tumor and the normal tissue. Since there is no surrounding wall of membrane, the growth spreads in mass. Malignant tumors, especially carcinoma growth, infiltrate in this way.
3. Cancers of the Exophytic Growth type consist of cancer cells that occur on the skin, body cavity, or gastrointestinal mucosa; they often form on the surface of the body's tissues through papillary masses, projecting outwards on the skin and mucosa. Polyp, papilla-

shaped tumors, papilloma, gingival cancer, esophagal cancer, Stomach cancer, and intestinal cancer all grow in this way.

Benign tumors do not spread; they only grow in an expansive or exopathic way. The malignant tumor is different, however, as the cancer cells do not remain in their original location but expand from the mass outward. This expansion extends into other organs and tissues by way of the body's lymphatic plexus and blood circulation. This phenomenon is known as metastasizing.

ETIOLOGY, THE SIX PATHOGENIC EVILS THAT CAUSE CANCER

The following is an analysis of the causes of tumor formation. Traditional Chinese Medicine maintains the concept that a tumor is a systemic disease (affecting the whole body). It can be summarized as a disorder of the internal organs caused by emotional strains, and an imbalance of the viscera, that has allowed the pathogenic Evils to enter the body and obstruct the channels and collaterals. These pathogenic Evils cause problems, that may eventually lead to the stagnation of Blood and Qi and develop into tumors. The pathology of cancer is as follows: Qi stagnation, Blood stagnation, Retention of Toxic Dampness, Retention of Phlegm, Toxin Attack, and Weakness or Deficiency of Energy (Figure 54.2).

1. Emotional strains, depression of Liver Qi, or External pathogenic Evil factors can lead to a stagnation of Qi. Long term Qi stagnation can cause stagnation of Blood, which eventually causes clots. It is known that pathogenic Cold and Heat, and the prevailing pathogenic Evil factors, inevitably cause Blood clots.
2. Since Blood and Qi flow together, the obstruction of Qi can lead to an interference in Blood circulation. This interference gradually leads to Blood stasis; Blood Stasis is always further complicated by stasis of Qi. Blood stasis is sometimes referred to as "Dead Blood," due to the fact that the Blood is not moving. When Dead Blood accumulates in a certain area of the body, it eventually develops into an immovable mass or lump.
3. A weakness of the Spleen and Stomach, as well as indigestion, can cause retention of fluid in the body. Long term fluid retention will produce toxic materials that can accumulate and form lumps.

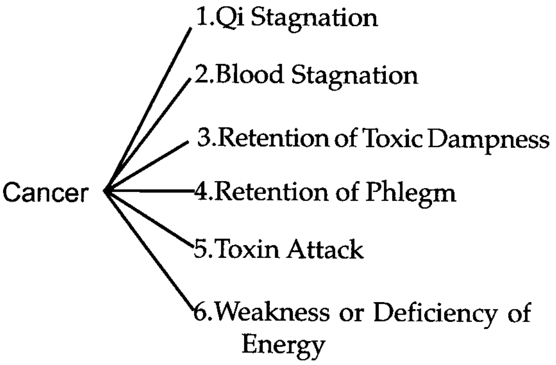

Figure 54.2. The Six Pathogenic Evils That Cause Cancer

4. Functional disorders of the Lungs and Spleen lead to poor digestion of watery food, and poor distribution of Body Fluids. Fluid retention causes Phlegm, especially when it is complicated by Excess Heat. If Phlegm is unable to move upward to be dispersed out of the body, it collects and stagnates. The stagnation of Phlegm in the Lung(s) leads to asthma. The stagnation of Phlegm in the Stomach causes gastric disorder and nausea. If Phlegm goes beneath the skin, it will take the form of a movable mass or lump.
5. Toxic Evils may come from poisonous external environments (exposure to chemicals, asbestos, ultraviolet rays, radiation, etc.) or may be generated by the stagnation of Excess Heat within the body. This Excessive Heat causes abnormal cell fusion and growth within the body, leading to cancer formation, and a systemic breakdown of normal cell growth.
6. Deficiency of both Qi and Blood, as well as a weakness of Yuan Qi, make the body more vulnerable and susceptible to the formation of new growth. When pathogenic Evils enter the body, they damage the Blood and Qi. If the Yuan Qi cannot be restored, the disease will lead to a vicious, destructive cycle, mak-

ing the individual more and more debilitated.

These six pathological changes may occur either alone or in combination: i.e., Qi and Blood stasis, Phlegm stagnation and Toxic Dampness, Phlegm stagnation and Blood stasis, a weakness of the body's Yuan Qi and strong pathogenic Evils.

The choice of which Medical Qigong therapy to use should be based on differentiation of symptoms and signs. Since cancers are primarily caused by emotional strains, strong pathogenic Evils, and a weakness of the patient's Yuan Qi, clinical therapy should be focused on tonifying the patient's Yuan Qi, and purging the Evil pathogenic factors.

EMOTIONAL COMPONENTS OF CANCER FORMATION

From the clinical perspective, if the cancer is internally induced through emotional suppression, there will be a series of emotional transitions the Qigong doctor will observe during its development. When there is an accumulation of negative emotions, for example, the patient's body begins to wall off this type of energy, and creates an energetic cyst, instead of dissipating and disposing of the Toxic Qi. The patient's suppressed internal emotions intensify the production and formation of the cyst, and feed the cancer cells, which result in an uncontrolled, rapid growth rate of these cells.

The body creates potential tumors and cancer cells every day. The macrophages or monocytes (large phagocytes) in the immune system aggressively attack and dissolve the cancer cells once the body becomes aware of them. Areas in the body, however, which are armored in denial, tend to attract and protect these distorted cells, allowing them to cluster and mass together.

Cancer cells can grow and accumulate wherever emotions are being repressed. The patient's denial maintains the constant unconscious flow of energy to these tissue areas. The phagocytes and other components of the immune system do not attack these toxic pockets of emotional energy, because of the patient's continuous denial of their existence. Consciousness and awareness exists even at the cellular level of the body. The continuous emotional energy bombarding the cells and tissues causes these tissues to go into an emotional state of shock. Emotional shock causes a release of endogenous opiates and corticosteroids, that deplete the white blood cells, and prevent the immune system from protecting the body. Because of the immune system's weakened state of resistance, infection and even the formation of cancer cells are now free to enter and coagulate within the tissue areas.

Scientific studies based in China, and the West, have shown that stress can stimulate the neuroendocrine system to the extent that it leads to depression of the immune system. Furthermore, the creation and growth of cancer can be influenced by stress. Removal of psychological stress alone can alter the neuro-immunologic functioning of the body to the extent that it can prevent and even change the course of cancer development.

In a patient with emotional blocks, chronic tension begins to interrupt the flow of Qi. If the emotional block is deep and long standing, this tension becomes constant. The capacity to express or discharge the trapped emotion becomes very difficult.

As the awareness dims, the natural communication between the conscious and subconscious mind breaks down. The original conscious decision to suppress certain emotions becomes a habit; it thus becomes an unconscious aberrant automatic function. When emotions are expressed, the conscious and subconscious mind are in harmony and synergized. When emotions are suppressed, the subconscious mind can no longer perform its job of healing the body; it diverts all its energy to shield the patient's denial system. Thus, the Shen itself, becomes misaligned with the patient's life purpose. The patient's posture, movements and lack of spiritual growth, reflect this misalignment. Changes also occur in skin temperature, body chemistry, and skin tone. The overall level of vital life force visibly drops.

Dr. Xu, at the Xi Yuan Hospital in Beijing, China, often pointed out that emotional upheaval, and the restriction of the breath, could cause the patient's energy to stagnate or deviate from its path. The deviated Qi draws more energy to its pathway, compounding the disease. The unstable emotional and mental focus of the patient further causes adverse effects to the body. Medical Qigong therapy and exercises expose the patient's deepest secrets, and thus reveal the underlying causes of the disease, which can then be treated. The patient's being is energized through his or her personal perceptions and belief structures. Therefore, it is necessary to change the patient's belief structure to increase the level of awareness and perceptual accuracy. The patient must be enabled to let go of suppressed emotions to fully experience deep emotional and spiritual changes. The Qigong doctor initiates the release of Toxic Qi to begin this healing process that can thereafter be maintained through appropriate prescriptions.

When cysts, tumors, or cancer are internally induced, they are formed from long-standing suppressed emotions which have become trapped within the tissue itself. These emotions usually consist of anger, guilt, grief, shame, and disappointment. The energetically formed cyst, or tumors, will remain walled off until the emotional energy is dissipated. Cysts, or tumors, can lodge themselves anywhere, penetrating deep into the major viscera causing visceral dysfunction; or they can superficially lodge themselves beneath the skin.

If the nature of the emotional energetic matrix (i.e., belief structure) is negative, the patient retains the energy of the injury, making it difficult to heal the trauma. If, however, the emotional matrix is positive, then the energy of the injury can be easily dispersed.

Once the Qigong doctor's energy reaches the cyst or tumor, the patient's energetic resonance supporting the tumor formation is interrupted. As the doctor continues to break up the cyst or tumor's physical mass, the frozen energetic pattern begins to unwind. The doctor continues to unravel the energetic mass until all the Heat has been released from the tissues, and the energetic pattern's unwinding is completed.

The underlying pathophysiology of healing can be broken down into four components:

1. The mind affects tissue matter through the emotions.
2. The emotions cause neuropeptides to be released.
3. The emotions (via the neuropeptides) influence the direction and movement of the immune cells (e.g., monocytes), sending them to various parts of the body.
4. The immune system (e.g., monocytes and phagocytes) heal the body.

The neuropeptide receptor sites are located on the immune cells, linking the body's immune system with the functions of the mind, emotions, and spirit. Neuropeptides are produced in response to the individual's state of mind and emotions; they influence the state of health of all the internal organs.

Through Dynamic Medical Qigong training, neuropeptides, in the form of endorphins, are released within the body's system. This release of energy and endorphins helps to facilitate the patient's healing.

Through the use of Medical Qigong therapy, exercises, and meditations, areas of denial and subconscious energetic patterning can be altered. The energetic matrix which supports the tumor(s) or cyst(s) thus dissolves.

When a healing session occurs, the Qigong doctor's external energy field envelops the patient's tissues, thus allowing the patient's cells to receive new information to reprogram the toxic energetic patterns. This reaction occurs providing the patient's tissues are open to receive the informational structures needed to either rebuild them, or disperse the disease's cell formation.

CANCER FORMATION AND THE PATIENT'S MIND, EMOTIONS, AND SPIRIT

Emotions strongly influence the increase or decrease of tumor size; therefore, it is important to regulate the patient's mind, emotions, and spirit. When the patient's mind, emotions and spirit are free from distractions and are in harmony, the patient is able to manage negative emotions. Regulating the Mind (whole body awareness and consciousness) is considered the primary Medical Qigong technique used for teaching patients how to avoid creating the dark abyss of negative emotions.

The patient must learn to let optimism prevail over pessimism. The practice of Medical Qigong not only cultivates vital energy and builds up the patient's health, but promotes the development of optimism through creating a quiescent mind. Tranquility and optimism improve circulation of Qi and Blood (which strengthens the immune system) thereby inhibiting cancer cells until they gradually perish.

While visiting in Maui, Hawaii, I came in contact with four people who had been diagnosed with terminal cancer, and had a life expectancy of two weeks according to their physicians. Each person had their own unique story as to how they had come to the Hawaiian Islands to die. What struck me as fascinating was the fact that each one of these individuals had either left successful but extremely stressful businesses, or long term abusive relationships. Each made amends the best they could with relationships from their past (those who had anger against them or those whom they had hurt). They had come to paradise to live their last weeks in peace. That was several years ago! It seems that upon resolving all personal conflict, and leaving their stressful environment, their cancers went into spontaneous remission. Each individual was now concentrated on enjoying the pleasures of a simpler life, and on focusing on their own spiritual growth.

STRESSFUL LIFE ACTIVITIES AND THE IMMUNE SYSTEM

Stressful life activities create a flood of adrenaline within the body. This occurs not only from the endings of the vertebral lumbar division at the core segment of the major internal organs, but also from the core of the adrenal gland in the medulla (which specifically is a sympathetic ganglion, and responds to stress by releasing adrenaline into the blood stream). Adrenaline spreads throughout the body and activates not only the body's sympathetic nervous system, but also the adrenal glands; the heart rate, respiratory rate, blood pressure, and Blood flow to the muscles are increased in response to the Fight or Flight Syndrome.

When the release of adrenaline causes the mind and body to react, the immune system must then seek to balance the chemical imbalances created within the body, before it can return to its normal job of monitoring cellular activity, and defending against pathogens. If a patient is already vulnerable to a specific disease, added stress causes the patient to get sick quicker. Meditation takes the patient out of the stress mode, deactivates the adrenal glands, and normalizes the function of the immune system.

An unstable life style is very stressful and can lead to depression and Qi deviations. Conflict produces stress which causes anxiety; chronic stress wears down the mind and body, which leads to anxiety, or depression. Stress begins with an overexcitement of the nervous system, slowly working its way deep into the body like wet cement. It causes the shoulders and upper back to become tense and rigid. Stress can result from happy occasions (such as weddings) and other normally pleasant events. When the patient's life style is not properly balanced, over time Qi deviations, and illness can result. The capacity to bear stress is related to the strength of the nerves. When stress levels surpass the nervous system's capacity to handle them, the nerves "break down"; this results in all sorts of physical, mental, and emotional disturbances. These disturbances, if unheeded, can eventually lead to organ malfunctions, cysts, tumors, cancer, and premature death. To avoid such states, a balance between work and rest must be maintained.

At a conference of the International Society for Neuro-immunomodulation (*The Effects of Stress and Depression on Physical Disease*), Dr. Philip Gold, of the National Institute of Mental Health stated that stress and depression send hormones (e.g., cortisol) flowing into the bloodstream. An imbalance of hormones, if unregulated, can:

- destroy the appetite,
- cripple the immune system,
- shut down the processes that repair tissue,
- inhibit sleeping patterns,
- break down bone density,
- aid the onset of infection, and
- instigate the growth of cancer cells.

An individual's emotional past determines his, or her, present state of health, both physical and emotional. Past emotional belief structures are responsible for both creating and healing diseases. Just as thoughts drift through the mind (but do not become active until they reside within a belief structure), so emotional wounding resides within the tissues and cells waiting to be activated by emotional upheavals. Our thoughts become form via choices, and manifest within the body.

In most societies, sharing traumatic wounds is often the first act of intimacy and bonding. Traumas provide a common ground for relationships, and support the pain structure that cements the relationship. People become addicted to empowering wounds, which in turn supports and empowers the disease. The exchange of painful personal traumas is reinforced by receiving sympathy and strong empathy, which binds rather than releases the pain from the tissues. Desire for sharing one's pain in order to receive love can lead to a habitual way of relating to those closest and dearest. When such individuals feel taken for granted, or unappreciated, they may immediately resort to sharing past traumas to illicit the desired response. When such techniques fail to elicit the desired response, a new wounding is experienced. Many of these individuals will continue to unconsciously create new traumas, in the hope that, this time, the desired outcome will be achieved. This in turn can become an addiction, empowering old wounds as it creates new ones, including illness. Thus, the spirit becomes divided to support the new belief structure (i.e., that pain and misery grin love and attention), and the life-force energy deviates from its assigned job of protecting the body, which creates illness. Setting the self up for new wounds is not a conscious process, but an unconscious one, which needs to be brought into the patient's awareness in order to be healed. Then and only then can the mind and spirit act as one unit, directing the body's Qi towards its proper function of healing.

Because energetic healing occurs in the present, not in the past, victims of past traumas will never heal if the energy continuously regresses to the armored traumatic belief structure.

USING MEDICAL QIGONG TO STRENGTHEN THE IMMUNE SYSTEM

Medical Qigong increases leukocyte production, stabilizes the body's blood pressure, and improves the immune function. The immune system is not so much an independent system, but is rather an interactive functional system throughout the body. Immune cells which live in the Blood and lymph fluid are commonly known as leukocytes (white blood cells); lymphocytes are a type of leukocytes that is created in the Spleen. The purpose of these cells is to work together to defend against pathogenic factors. The function of leukocytes and lymphocytes is to identify, ingest, destroy, and eliminate disease-causing organisms, as well as remove all dead, damaged, or irregular cells.

Immune cells are produced in the lymph tissue and red bone marrow, and are stored in the lymphatic organs and the Blood. To perform the function of protecting and cleaning, immune cells must circulate throughout the entire body through the capillaries and vessels that carry Blood and lymph fluid. Immune cells travel from the lymph nodes and the lymphatic organs of the thymus gland, Spleen, and tonsils, which are responsible for filtering the lymph fluid to remove impurities, and produce more lymph cells.

For an External pathogen to affect the body, it must first pass through protective layers of cells. Generally, the macrophages present within the tissues ingest the pathogens and dead tissues. Macrophages are reinforced by the action and movement of the body's Wei Qi. When a pathogenic intruder invades the body, the macrophages are often the first to attack the intruder. If the macrophages are not strong enough, more Blood and fluid are brought to the area, which contain more diverse immune cells. The increased fluid engorges the tissue, which engenders more cellular activity. The increased Blood and cellular metabolism increases the local temperature. The immune cells break down and ingest the pathogens, creating a process which continues until the disease is eliminated.

The dynamic movements of Medical Qigong

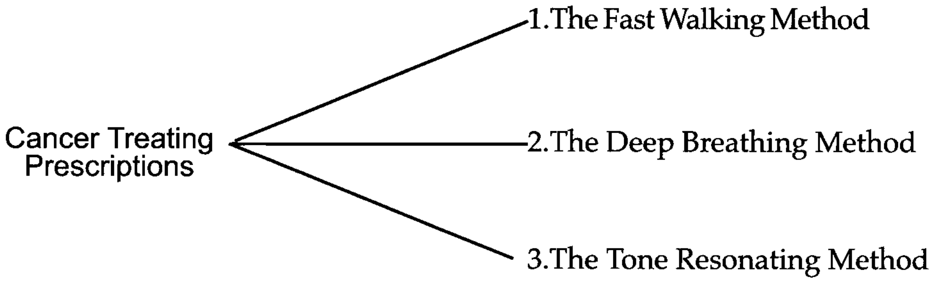
Cancer Treating Prescriptions
1.The Fast Walking Method
2.The Deep Breathing Method
3.The Tone Resonating Method

Figure 54.3. Three Prescriptions Used For Treating Cancer

exercises generate Heat, soften the body's connective tissues, and encourage the flushing of the intercellular fluid throughout the body's tissues and organs. Energetic Qigong practices increase oxygenation in the Blood and improve circulation, thereby enhancing leukocyte production, and strengthening the body's immune system.

CANCER TREATMENT METHODS

Tumor and cancer cells make their own room within the body's tissues by crowding the normal tissues. The tumor and cancer cells grab nourishing energy at the expense of the host, thus damaging the patient's body.

Medical Qigong cancer treatment methods are techniques designed to purge the energy that supports the tumor. These techniques are based on the fundamental Qigong patterns mentioned previously. The goal of these patterns focuses on treating the patient's Excess syndrome first, by purgation. Purgation is then followed by tonification, then regulation. The rapid growth of healthy energy, and the fast expulsion of Evil factors, should lead to normal body function restoration, and create more vigorous metabolic activity. More nourishment can be absorbed, delivered and distributed throughout the body, leaving no room to feed the tumor cells.

In Medical Qigong clinics, tumors are viewed as being similar in their expansion and growth to that of the energetic pattern of a plant's Qi flow. When tracing the origin of a tumor, the Qigong doctor begins at the tumor's location (considered the flower or fruit of the disease) and traces its energetic branches to the organ of origin (the actual root of the disease). If the disease is internally induced, the original source, i.e., the dysfunctional internal organ, feeds the tumor through suppressed emotional trauma.

To address this problem, the doctor prescribes the following three cancer treatment methods as homework, in addition to Medical Qigong treatments and herbal therapy: the Fast Walking Method, Deep Breathing Method, and the Tone Resonation Method (Figure 54.3).

1. The Fast Walking Method stimulates the electromagnetic fields within the tissues and reinforces the patient's Yuan Qi. By walking fast and changing the normal breathing patterns, the cancer cells can be destroyed. This is because the dual action of shocking the system with more Qi and Blood flowing through the body tends to stimulate the electromagnetic fields within the tissues, causing the tissue cells to become stimulated and the immune system to be aroused and thus enhanced.
2. The Deep Breathing Method tranquilizes the patient's mind, and enhances their Yuan Qi. The Deep Breathing method is used to treat various cancers due to mental and emotional factors, by changing the patient's normal breathing patterns.

By changing the normal breathing into rapid exhaling, slow deep exhaling, or relaxed toning, Blood clots and tumors can be dissolved. This breathing technique has a special effect on reinforcing the patient's Yuan Qi. When prescribing Medical Qigong therapy, it is important to match the patient's Yin disease with a Yang method of treatment, and a Yang disease with a Yin method of therapy. For example: rapid exhalation (a Yang method) would be used when treating an abdominal mass caused by a Cold (Yin) condition.

3. The Tone Resonating Method is utilized to purge any residual Evil pathogenic factors from the patient's body.

When first starting the Medical Qigong Cancer Tone therapy, a patient's practice time should be short. The performing time can be extended gradually with the increase of the patient's physical strength. To bring the Medical Qigong exercises to peak efficiency, a patient should practice a minimum of four hours (including 15 minute breaks), and up to six hours each day. The patient can take a break after half an hour to avoid fatigue.

Through various kinds of breathing and fast walking therapies, Qi can be energized to flow so vigorously through the Conception Vessel that the patients' Yin and Yang is balanced and their Lung Channels are strengthened. In the course of treating the tumor or cancer, True Qi (Zhen Qi) is fostered and the immunity of the body strengthened. Research from China proves that persistent training, vigorous enough to increase a person's heartbeat up to 70-80% of its maximum rate, helps dissolve Blood clots. Other reports state that long distance running can be effectively used to cure cancers. These reports may help explain why fast walking, and breathing methods, are effective for preventing and treating cancer.

SOCIAL ONCOLOGY—ESTABLISHING A CANCER SUPPORT GROUP

In China, cancer patients join support groups to practice Medical Qigong. This form of treatment is called social oncology; it was created by the Cancer Recovery Society for the purpose of supporting human interaction and the recovery of health by cancer patients.

Cancer recovery groups meet daily in parks throughout China, practicing self-healing (in particular, Guo Lin's Cancer Recovery Qigong and Walking Qigong Therapy). After practice, the groups gather together for about an hour to socialize in a nearby tea house. The patients sing songs, tell jokes, laugh, introduce new members, read poems, and tell stories that highlight the possibilities of recovery from cancer. As part of the social meeting, the group regularly sings "Happy Birthday" to each patient who has an anniversary after his or her cancer diagnosis. Laughter becomes a top priority in social oncology, and is considered the strongest medicine for healing.

According to Dr. Roger Jahnke, O.M.D., the advantages of social oncology are three fold: stress reduction, positive reinforcement of hope, and the establishment of deep interpersonal relationships.

1. Stress is reduced through socializing rather than isolating and fretting. The social healing aspect of the patient's recovery program serves as a constant reminder that the physical healing is enhanced when he or she becomes lighthearted and free of worry.
2. Increased hope and faith is achieved through the power of personal testimony. Each day during the social healing session, patients hear stories of close friends whose recovery process proves the benefits of healing from various forms of Medical Qigong.
3. The value of love and encouragement from friends mitigates depression, fear, and worry. Isolation is like a food to cancer; social oncology provides fun, encouragement and a deep, caring interpersonal connection within the recovery group.

SOCIAL ONCOLOGY WITHIN THE HOSPITALS

In several Medical Qigong hospitals in China, when the patients first check-in, they are assigned a support group. The support group consists of previous and current patients, who have survived the same type of cancer as the entering patients.

To assist patients in healing and to give them every emotional advantage during recovery, visitors, phone calls, letters, TV, and newspapers are forbidden for six weeks; this ensures a controlled environment in which neither family nor friends can adversely affect the healing process (through critical comments or attitudes). Patients are immersed in a support group full of caring, motivated individuals all working towards the same goal. Furthermore, patients who have been there the longest share their success stories, boosting everyone's moral.

The concept of clinical social oncology is based on the belief that the patients' high spirits affect the emotional perspective, and thus mental thinking. These emotional and mental changes alter the physical body. It was recently discovered that the group of breast cancer patients, who participated in the Stanford Hospital Cancer Support Group, lived twice as long as the women who refused to join the group. This difference in longevity was attributed to the love, support, and social aspect of the cancer support group. This recently discovery has given rise to the increase of several support groups in the United States now being established for patients recovering from AIDS, heart attacks, and many other illnesses including several types of cancer.

CANCER RESEARCH

Noted research scientist Feng Lida pioneered cancer research in China by showing that emitted Qi from Qigong masters produced marked changes in cancer cell cultures from mice. Several studies reported the effects of emitted Qi on tumors in animals. The emission of Qi, for example, was reported to inhibit the growth of implanted malignant tumors in mice, but did not destroy the tumors. Encouraged by the results with animals, researchers carried out clinical research on the effects of Medical Qigong on human subjects with cancer.

In one study, 127 patients with medically diagnosed cancer were divided into a Qigong group of 97 patients, and a control group of 30 patients. All patients received drugs, and the Qigong group practiced Medical Qigong for more than two hours a day over a period of 3 to 6 months. Both groups improved, but the Qigong group showed improvements four to nine times greater than the control group for strength, appetite, and weight gain. The Qigong group also did not experience bouts of diarrhea. The phagocytic rate, which is a measure of the immune function, increased in the Medical Qigong group, but decreased in the control group. These cancer treatment results were gathered and compiled from clinical research by Dr. Kenneth M. Sancier, available in the first Appendix of this text.

MEDITATIONS FOR CANCER PATIENTS

The following meditations are frequently used within the Medical Qigong clinic to aid healing. As previously discussed, it takes focused concentration to initiate an energetic transformation. Focused concentration is achieved through proper imagination, visualization and positive affirmation; it is practiced to reprogram the body's energetic cell patterns, and initiate healing.

DIVINE HEALING LIGHT MEDITATION

A very effective Medical Qigong meditation called Divine Healing Light is practiced by cancer and tumor patients throughout China to dissolve and disperse energetic pathogens. This meditation begins as follows:

Begin in a seated posture. Relax, and feel a quiet comfort. Imagine yourself seated on a quiet mountain overlooking the ocean, resting under a clear, dark blue evening sky. At the rim of the ocean, the full moon begins to rise. It continues to ascend until it faces you, suspended in the sky above your head. Slowly the moon begins to shrink, becoming smaller and smaller, yet brighter and brighter until it becomes the size of a small luminous pearl. This brilliant white pearl begins to pulse and expand, as it grows it opens to form a large orb, vibrating with divine power. From the center of the divine power flows unconditional love, healing, and compassion.

Breathe in this vibrant, healing light; allow it to penetrate deep into your body, permeating every cell, illuminating and dispersing all illness. Open fully to the divine healing. Feel the love and compassion enveloping you, and penetrating all the diseased areas of your body. Continue to breathe the healing energy in through your nose and exhale out through your mouth (imagining all the unclean energy descending into the ground). Continue this breathing pattern until all the negative energy has been expelled and replaced with divine love and healing light.

PULLING OUT THE PAIN MEDITATION

This meditation is an effective treatment prescription for treating many illnesses, including cancer and tumors. The goal is to eliminate all toxic belief structures that stand in the way, and sabotage the patient's healing. Gentle acceptance of the illness and emotional pain opens the pathway to understanding and unveiling hidden traumas. Force does not work in trying to change the tumor or cancer energetic patterns. Only honesty, respect, tenderness, and understanding will begin the healing process. As pain and suffering are allowed to surface into consciousness, the immune system is able to recognize diseased cells, and can mobilize an army of immune cells. These are then able to effectively identify, dissolve, and remove any tumors and disease.

The doctor instructs and guides patients through this meditation matching the rhythm of the words, to the patients' exhalations. Patients draw their painful memories and traumas out of their body in the form of dark smoke; they then separate the distilled wisdom and knowledge from this dark cloud, reabsorbing the healing light into every pore and cell in their body. The dark cloud then is released with the help of an angelic being who severs the energetic cord(s) that still attach the patients to their pain and suffering. The doctor assists the complete severance by clapping his or her hands as the patients visualize the "Sword of Truth" (held by this holy being), severing the dark energetic cord(s). The sword of truth represents the patients' intent to regain their power by letting go of their attachment to their illness. This attachment created very real energetic cords that bound the patients to their painful memories and toxic emotions. Severing these cords, or cord, symbolizes the end of the patients' investment in their disease, and the beginning of true healing. All the patients' pain and despair is released, as the black cloud ascends into the Heavens, and into the hands of God (or a higher power, in accordance with the belief structure of the patients). As the cloud reaches the first wave of love and compassion emanating from God (or a higher power), a mighty explosion transmutes the black despair and pain, into divine healing light. The doctor then assists the patients in accepting this healing, loving light back into their body, to experience, perhaps for the first time in their lives, true forgiveness and a state of grace.

1. Begin by sitting comfortably in a chair, with both feet on the floor, hands resting on the thighs, eyes closed, tongue on the upper palette. Breathe naturally through the Lower Dantian.
2. Focus your attention on the Yellow Court area below the diaphragm where you store painful memories and traumas. Imagine opening up this area to release the toxic energy out of your body as a stream of dark steam releasing from a pot.
3. As the dark steam flows out of your body you release feelings and memories of guilt, anger, rage, humiliation, abandonment, degradation, rejection, insecurity, and sorrow. These toxic emotions have, until now, prevented you from healing through creating a deep distrust of yourself and others.
4. Focus your attention on this energetic cloud, and begin separating the pain and hurtful memories from the knowledge and wisdom gathered from these experiences. Visualize this knowledge and wisdom in the form of golden, white, and silver light energy gathering on the right side of the room. Continue to drain the dark cloud of misery and pain, now occupying the left side of the room, until you have extracted all the new insights from it. As you do so, you notice the dark cloud becoming heavier and darker.
5. Focus on the right side of the room. Through you intention, begin to inhale and imagine this bright, illuminating energy flowing back into your body. Absorb this knowledge and wisdom gathered from past experiences, void of any feelings of hurt, pain, or judgement, into every cell of your body. This distilled knowledge and wisdom empowers you to heal from

your wounds on a physical, mental, emotional, and spiritual level.

6. Next focus your attention on the dark black cloud containing all your pain and suffering. Imagine an angelic being, righteous and holy holding a "Sword of Truth" standing by your side, ready to sever the dark energetic cords still connecting your physical body to this dark cloud of despair. The sword descends and severs the cords (the doctor claps his or her hands at this moment).
7. The cloud of darkness begins to float up, through the ceiling, through the sky into space. Far in the distance you begin to perceive God, or a higher power, emanating waves of compassion, love, and mercy descending towards the Earth. One of the waves touches the ascending dark cloud of hurt and pain, exploding it into a brilliant light. The cloud is immediately transmuted into fluorescent crystal blue drops of light. This pure, clean energy descends from the Heavens like a gentle rain. Breathe in this crystal blue healing energy, let it penetrate deeply into every pore, every tissue and cell, saturating your body completely. Feel the light cleansing, healing, and radiating throughout your being.

SUMMARY

Cancer is defined by Traditional Chinese Medicine as, "an obstruction of Qi and Blood circulation resulting in stagnation." The root cause for cancer is the breakdown and stagnation of the patient's Yuan Qi. Once cancer becomes established within the patient's body, or metastasizes, it is difficult to heal.

In order for cancer to form, it must elude the body's immune system, undergo many divisions, and produce countless generations of cells without resistance. By the time the cancer cells have formed into a cluster mass, a fixed energetic pattern has been established to support its existence and survival.

Only when the patient's Yuan Qi is returned to its original state, can the body begin to heal itself. Medical Qigong Cancer treatment methods strengthen the True Qi which in turn strengthens and cultivates the Yuan Qi. In this manner, the body is returned to its natural state of Yin and Yang balance.

It is the author's hope that these effective methods of treating cancer will soon be available to the American public. To this end, I have established a three-year master's and plan to offer a five-year doctorate program in Medical Qigong Science at the Five Branches Institute, College of Traditional Chinese Medicine, in Santa Cruz, California.

TREATING TUMORS AND CANCER WITH MEDICAL QIGONG THERAPY

The following are several Medical Qigong techniques and treatment modalities, currently used in China, to treat different types of energetically induced cysts, tumors, and cancers. It is advisable to accompany the prescriptions with various herbal prescriptions.

BRAIN TUMORS

A brain tumor is a generic term used to describe any intracranial mass, which includes: neoplastic, cystic, inflammatory (abscesses), or syphilitic.

ETIOLOGY

There are many causes for intracranial tumors. The most common cause is Kidney Yin (Water) failing to wash over the brain to disperse the pathogens. This condition can be caused by Excess Liver Fire dissolving the body's Kidney Water.

SYMPTOMS

Intracranial tumors produce symptoms primarily by two mechanisms: the infiltration and destruction of the normal brain tissue; or the tumor mass affecting the surrounding tissue, which causes edema, thus increasing the intracranial pressure.

General symptoms include headache (due to intracranial pressure), changes in the retina, and vomiting (without nausea). Additional symptoms include mental and emotional changes such as dullness, giddiness, or epileptiform convulsions.

TREATMENT

1. The doctors first step is to analyze the syndrome and the patient's emotional outlook.

When treating brain tumors, the growth of the disease must first be stopped. Check the patient's external living conditions, and do a psychological profile to determine any detrimental patterns.

2. Dredge and disperse toxic Heat out of the patient's Yang channels. Remove any Excess Liver Fire by drawing it out the Liver organ; then dredge the patient's Liver Channels, dispersing toxins out through the Lv-3 points.
3. The doctor must reverse the process of tumor development to dissolve the tumor. Instruct the patient to visualize energy from the Lower Dantian rising up to the tumor and transforming the mass back into Qi. The tumor growth must be reversed by first converting the solid tumor into soft tissue, then into fluid Qi that is then dispersed.

   In cases of cranial thrombosis, tumors must be dissolved slowly; otherwise other tumors will be created. The tumors must be softened, then slowly dispersed to avoid any recreation of the energetic channels' dysfunctions.
4. Extend Qi into the patient to reinforce the Kidneys' Qi by extending Qi into the patient's feet via the brain tumor point, Shihmein, at the bottom of the heels (Figure 54.4). Next, lead the patient's Kidney Water up into the brain to rinse over and slowly dissolve the tumor.
5. Finally, regulate and balance the patient's Yin and Yang energy.

PRESCRIPTIONS AND HOMEWORK

1. Have the patient practice the healing sound "Duo" (10 times in a straight and 10 times descending / ascending tone) five times a day.
2. Assign tonifying exercises for the Kidneys and Heart.
3. Prescribe herbal teas to enhance the patient's Kidney Yin.

BREAST CANCER

Mammary cancer is one of the most common types of malignant tumors in women; it is considered a malignant neoplasm of the breast.

ETIOLOGY

There are multiple factors associated with breast cancer. These factors can be categorized under genetic family factors, hormonal factors, dietary factors, environmental factors, and benign breast disease. Heredity can play an important role, especially if the patient's mother (or close relative) had bilateral breast cancer diagnosed prior to menopause. Also, a history of chronic breast disease (especially epithelial hyperplasia), or obesity can predispose the patient to breast cancer.

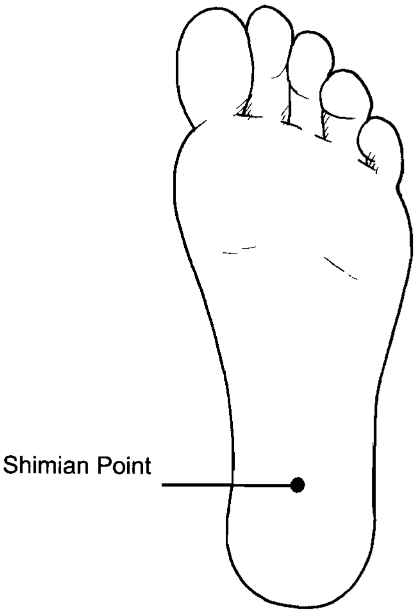

Figure 54.4. The Qigong doctor will emit Qi into the patient's "Brain Tumor" point (also called the Shimian, meaning insomnia point) located at the bottom of the patient's Heel.

SYMPTOMS

Symptoms consist of hard lumps which lack mobility, and progressively increase in size. The hard lumps may be concave, convex, or protrude; or the nipple may be inverted. Breast cancer is usually related to stagnation of the Liver's Qi, stasis of Blood, Phlegm and Toxic Heat, as well as weak Kidneys. The emotional components which lead to breast cancer are suppressed grief, worry, sorrow, and depression.

TREATMENT

1. Purge and eliminate the stagnation from the breast tissue (Figure 54.5).
2. Subdue the Rebellious Qi.
3. Clear the Liver and tonify the Kidneys. When treating breast cancer, the doctor should use the Vibrating Palm in conjunction with projected Tone Resonation (projecting both vibration and the "Ni" tone resonation into the cancerous area).
4. Tonify the patient's Conception and Thrusting Vessels, and end with the Microcosmic Orbit regulation (Fire Cycle).

UTERINE CANCER

Uterine cancer is considered a malignant neoplasm of the uterus.

ETIOLOGY

The etiology of uterine cancer is unknown. Possible causes of the disease may be:

- chronic stress
- toxins
- suppression of emotions
- pathological changes of the endometrial hyperplasia, or
- cervical carcinoma spreading onto the uterus.

SYMPTOMS

A uterine carcinoma is generally more common in women after menopause, roughly 75% of uterine cancers occur in postmenopausal women, 15% peri-menopausal women and 10% of menstruating women. The main clinical manifestations include pain in the lower abdomen, waist and thighs, abnormal uterine bleeding with a bloody, purulent discharge.

TREATMENT

One Medical Qigong therapeutic technique used in China to treat uterine cancer is as follows.

1. Purge the diseased area of stagnant Qi using the Thunder Palm technique in conjunction with the Vibrating Palm technique.
2. Project the sound "Yu" into the uterus to disperse Heat from the cancer cells.
3. Disperse Excess Heat from the patient's Liver and uterus, leading the Toxic Qi out the patient's body via the right Gall Bladder

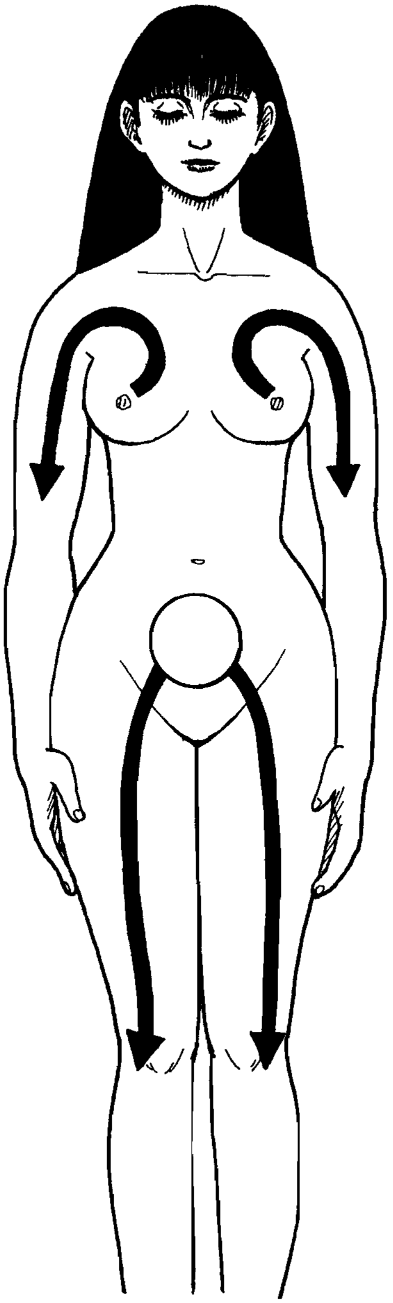

Figure 54.5. After Purging the pathogenic Qi from the patient's upper body, the Qigong doctor continues to disperse the patient's Toxic Qi from her lower abdomen, leading it down her legs and out her St-36 points.

Channel.

4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi. Circulate the energy through the Microcosmic Orbit to improve Qi and Blood circulation through the Governing and Conception Vessels.

PRESCRIPTIONS AND HOMEWORK

The Conception Vessel moves Qi in the Lower Burner and uterus; it is the primary Vessel used in treating uterine and cervical cancer. Sexual activities are therefore prohibited from the start of the treatments, until treatment is no longer needed.

1. Have the patient practice the Descending the Yang and Ascending the Yin Technique (see Chapter 41), ending with the healing sound "Yu" for 24 breaths, 9 times a day.
2. Prescribe the Fast Exhaling Breathing Method along with the Cancer Walking methods for treating Kidney disease (see Chapter 15).
3. Have the patient practice the method of Taking in the Blue Qi, three times a day for 15 minutes each session, to strengthen the Kidneys.
4. Have the patient practice the Filth Dispelling meditation, while focusing on her uterus. This meditation is practiced as follows:
    a. From a sitting posture, with the eyes closed, and the body relaxed, place the tongue up against the upper hard palate, behind the teeth.
    b. Breathe naturally and evenly.
    c. Imagine Qi whirling in through the upper portion of the uterus, circulating clockwise, building momentum as the energy flows downward to the base of the uterus.
    d. Once the energy has reached just above the cervix area, imagine the whirling Qi absorbing the noxious Heat and Toxic Qi from the uterus, transforming it into wind, and dispelling it out the vagina. Repeat this sequence 10 times.
    e. Next, gently close the vagina and anal sphincter, then inhale, and imagine the Qi flowing into the body through the vagina. As this energy flows into the uterus and Lower Dantian area, it transforms into wind. This wind circulates in a counterclockwise direction, spiraling up the body, through the Thrusting Vessel and exits the body through the mouth with each exhalation. Repeat this visualization 10 times.
    f. Finally, focus attention on the Lower Dantian, imagine the Qi returning to its origin, and rub the abdomen 36 times in a clockwise direction to end the prescription.

CERVICAL CANCER

Cervical cancer is considered one of the most common malignant tumors in China. It is considered a malignant neoplasm of the cervix of the uterus, and can occur at any age.

ETIOLOGY

Cervical cancer is suspected to be an oncogenic papilloma virus that is transmitted sexually. Risk factors include:

- sexual intercourse at an early age,
- multiple sexual partners and promiscuous male sexual partners (especially those whose previous partner had cervical cancer),
- serious cervical lacerations caused from the penis being too long, or the insertion of foreign objects,
- unsanitary sexual relationships,
- closely spaced births, and
- a family history of cervical cancer.

Excessive sexual activity predisposes young teenage girls to cervical cancer, because, with the onset of ovulation and the changes in vaginal PH, active squamous metaplasia is taking place in the cervix. During this time of cellular immaturity and vulnerability, a carcinogen is most likely to have an influence on the squamous epithelium; this can predispose the teenager girl to cervical cancer later on in life.

THE DEVELOPMENT OF CERVICAL CANCER

There are two main types of cancer of the cervix. The squamous cell carcinomas make up 85-90% of cervical cancer, the rest (a small 10-15%)

are developed from a malignant adenoma arising from a glandular organ known as a adenocarcinoma.

Cervical cancer begins with the growth of abnormal precancerous cells and tissues called dysplasia. Dysplasia is first detectable in the squamous epithelium of the cervix (the muscle lining of the cervix), and spreads to create a tumor called a cervical intraepithelial neoplasm (CIN). The cervical intraepithelial neoplasm progresses through three levels of severity, from mild, to moderate, to severe. At the severe level of CIN, the precancerous growth turns malignant; it becomes a carcinoma in situ (localized). If left untreated, the carcinoma progresses through four more stages of development, each of which is subdivided into two grades of severity. These grades indicate the location and spread of the carcinoma. The four stages and their subdivisions are described as follows:

In the beginning stages of the disease, cervical dysplasia is characterized by three stages of abnormality and severity (CIN-1,-2,-3), which later progress onto four stages of invasive carcinoma (Stages I through IV).

- CIN-1: Corresponds to a mild condition of dysplasia.
- CIN-2: Corresponds to a moderate condition of dysplasia.
- CIN-3: Corresponds to a severe condition of dysplasia, and carcinoma (in situ). The cancer, however, is found only in the layers of the cells lining the cervix, and has not yet extended deeper into the tissue lining.
- Stage I-A: Corresponds to a microinvasive carcinoma. The cancer is beginning to spread from the lining of the cervix into the deeper connective tissue.
- Stage I-B: Corresponds to an invasive carcinoma confined to the cervix. The cancer has now spread from the lining of the cervix into the deeper connective tissues.
- Stage II-A: Corresponds to a tumor extending beyond the cervix, into the upper third of the patient's vagina, but is still inside the pelvic area.
- Stage II-B: Corresponds to a tumor extending beyond the cervix, into the loose connective tissues around the patient's uterus (parametrium), but is still inside the pelvic area.
- Stage III-A: Corresponds to a tumor extending into the lower third of the patient's vagina.
- Stage III-B: Corresponds to a tumor extending into the patient's pelvic wall. At this stage, the cancer may be blocking the ureters (responsible for carrying urine from the Kidneys to the Bladder).
- Stage IV-A: Corresponds to a tumor extending into the patient's Bladder or rectum.
- Stage IV-B: Corresponds to a tumor extending into the patient's Liver or Lungs.

SYMPTOMS

Clinical symptoms include bleeding after sexual intercourse and between periods, abnormal bowel movements, or fatigue. Bleeding may also be noticed after severe exertion or straining during defecation. As the disease develops, the leukorrhea (vaginal discharge) may become filthy and mixed with blood, or foul and mixed with sloughed pieces of Body Fluid or tissue. Additional symptoms include lumbar, sacral and abdominal pain, Bladder irritation, frequent and urgent urination, blood-tinged vaginal discharge, unilateral lymphoedema, and unilateral ureteral obstruction (this indicates a very advanced stage).

TREATMENT

When Righteous Qi is increased through Medical Qigong therapy and prescription homework, the energy feeding the patient's cancer begins to decrease.

One Medical Qigong therapeutic technique used in China to treat cervical cancer is as follows:

1. Purge the diseased uterine area of pathogenic Qi using the Thunder Palm technique in conjunction with the Vibrating Palm to disperse the stagnation.
2. Use the "Yu" sound Qi projection into the uterine and cervix area to disperse Heat from the cancer cells.

3. Disperse the Excess Heat from the patient's Liver and cervix areas, leading the Toxic Qi out the patient's body via the right Gall Bladder Channel.
4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi. Circulate the energy through the Microcosmic Orbit to improve Qi and Blood circulation through the Governing and Conception Vessels.

PRESCRIPTIONS AND HOMEWORK

The Conception Vessel moves Qi in the Lower Burner and uterus, and is the primary vessel used to treat uterine and cervical cancer; therefore, sexual activities are prohibited from the start of the treatments, until treatment is no longer needed.

The prescriptions and homework required for cervical cancer treatment are exactly the same exercises that are used for the treatment of uterine cancer.

PROSTATE CANCER

Prostate cancer is considered a malignant neoplasm, and is usually an adenocarcinoma of the prostate gland. It is the major type of cancer in males. Prostate cancer occurs in less than 1% in men under 50. Men with a family history of prostate cancer are three times more likely to develop it then the general public. Many prostate cancers remain dormant, but once awakened, can become life-threatening.

ETIOLOGY

The etiology of prostate cancer is unknown. Speculations suggest that prostate cancer can be related to an imbalance between male and female hormones. From a Western Medical perspective, the carcinogenesis is a multistep accumulation of genetic lesions. These lesions may result in uncontrolled cellular proliferation, a decrease in cellular death or an apoptosis, invasion, and the metastatic spread of the disease.

From a Traditional Chinese Medical perspective, prostate cancer is due to Excess Liver Fire and Kidney Deficiency. The suppression of anger and fear, as well as a diet consisting of a high alcohol intake, or consumption of too many hot spicy foods, may also cause, or contribute to prostate cancer.

SYMPTOMS

In the early stages there are no symptoms; however, as the disease progresses the inflamed prostate gland impedes urination and results in outflow obstruction, preventing the Bladder from emptying completely. Dull pain is experienced around the external genitals and Bladder. There is frequent burning, unfinished and dripping after urination with white discharge. Also, in some cases, there will be sexual function disorders (an inability to attain erection) as well as sciatica-like symptoms with pain radiating downward to the perineum and anus.

If the prostate is hard and firm, it indicates cancer; if the prostate is large and swollen, it indicates an infection. Inflammation of the prostate can be divided into acute and chronic conditions.

In acute prostatitis caused from Damp Heat in the Lower Burner, symptoms include pain and discomfort in the genital and Bladder areas, fever, and chills.

In chronic prostatitis, symptoms include dull pain in the urogenital area, discharge from the penis, discomfort during sexual intercourse, and frequent urination (especially during the night).

TREATMENT

One of the Medical Qigong therapeutic techniques used in China to treat prostate cancer is as follows.

1. With the patient supine, begin to purge the diseased prostate area of pathogenic Qi using the Thunder Palm technique in conjunction with the Vibrating Palm to disperse the stagnation.
2. Next, use "Yu" sound projection into the prostate area to disperse the Heat from the cancer cells.
3. Disperse Excess Heat from Liver and pathogenic Qi from the prostate, out of the patient's body via the right Gall Bladder Channel.
4. Fill and tonify the patient's Lower Dantian and Kidneys with Qi, and circulate the energy through the patient's Microcosmic Orbit to improve Qi and Blood circulation within the Governing and Conception Vessels.

PRESCRIPTIONS

Depending on the patient's condition, one or more of the following prescriptions may be given:

1. The "Guo" sound resonation is used to disperse the Liver Fire.
2. The Kidneys' "Yu" sound resonation is used to disperse Toxic Qi from the prostate.
3. Tonification and regulation exercises are prescribed for the Lower Dantian, Spleen, and Kidneys.
4. The following three exercises are used to treat prostate cancer, and were made popular in the West by Dr. Hong Liu.
    a. Sit at the edge of the chair with your feet flat on the floor. The right leg crosses over your left leg. With your left hand grasp the bottom of the right foot.

       Then extend your right hand out in front of the body, palm facing upward, while you imagine releasing toxic energy from your body through the mouth as you exhale. As you inhale, twist your hand as far as possible towards the right side keeping your eyes on the palm, while simultaneously pulling your foot towards the body and contracting the anus and perineum. This opens the Liver Channel and directs Qi into the Kidneys and prostate area.

       After completing the movement, hold your breath for a moment. Imagine Blood and Qi flowing into the prostate area and exhale while returning to the starting position. The exercise is repeated for 18 breaths per each side.
    b. The next exercise begins from a Wuji posture. Bend over as you flex your toes upwards. Hold this position as long as you can before standing erect, focusing on breathing naturally from the abdomen. This movement increases the flow of Qi and Blood into the prostate area. It is important to make sure this exercise is performed slowly. The exercise is repeated 9 times.
    c. This last exercise is done sitting on the floor while facing a wall. Place the balls of your feet against the wall while resting your heels on the floor. Wrap your arms around your bent knees, while keeping the back straight and breathing naturally through the Lower Dantian.

       While inhaling, focus on breathing from the lower abdomen instead of the chest. This causes the Qi of the Thrusting Vessel to stimulate the prostate gland, and increases the flow of Qi and Blood into the prostate area. The exercise is repeated for 18 breaths.
5. The "Filth Dispelling Meditation," is also prescribed, with the patient focusing on his prostate area. This meditation is practiced as follows:
    a. From a sitting posture, with the eyes closed, and the body relaxed, place the tongue up against the upper hard palate, behind the teeth.
    b. Breathe naturally and evenly.
    c. Imagine Qi whirling in through the upper area of the Bladder, circulating clockwise, building momentum as the energy flows down to the prostate area.
    d. Once the energy has reached the prostate, the patient will imagine the whirling Qi absorbing the noxious Heat and filthy Qi from the Bladder and prostate areas, transforming it into wind, which is dispelled out through the penis. The exercise is repeated 10 times.
    e. Next, gently close the anal sphincter while inhaling and imagine the Qi flowing into the body through the penis. This energy flows into the Lower Dantian, becoming transformed into wind. This wind begins circulating in a counterclockwise direction, spiraling up the body, through the Thrusting Vessel and exits through the mouth with each exhalation. This is also practiced for 10 repetitions.

f. Finally, focus the attention on the Lower Dantian; imagine the vital Qi returning back to its origin, while rubbing the abdomen in a clockwise direction to end the prescription.

LUNG CANCER (PULMONARY CARCINOMA)

Lung cancer, also called a "Bronchial Carcinoma," is a malignant Lung tumor that originates in the bronchi. It can also spread to the trachea appearing as an ulcer in the windpipe, a small flattened bump or nodule. It can extend into the body's lymphatic system, as well as into the blood vessels.

The most common type of bronchogenic tumor is the squamous cell or epidermoid (on the airway lining). The oat cell and adenocarcinoma are the next most common.

ETIOLOGY

Squamous cell cancer is usually associated with a history of cigarette smoking and second-hand smoke. Statistically it is the most common of all carcinomas and accounts for 40% of all male deaths. Other chemical carcinogens may cause Lung cancer, especially among workers in industrial and mining areas.

SYMPTOMS

The main clinical manifestations include: chest pain, coughing with scanty sputum (sometimes with blood), pectoralgia, fever, loss of appetite, loss of weight, weakness and breathlessness in the later stages due to a Lung Yin Deficiency.

TREATMENT

Purge the Excess Heat from the patient's Lungs and Liver. Tonify and regulate the patient's Kidneys and Mingmen area, then strengthen the patient's Spleen and Lungs.

PRESCRIPTION AND HOMEWORK

1. Have the patient practice the Dry Crying exercise for a period of 15 minutes twice a day (see Chapter 53).
2. Have the patient practice the Sun and Moon Rotation Technique (see Chapter 41), ending with the healing sound "Shang" for 24 breaths, 9 times a day.
3. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Stepping Method for treating Lung diseases (see Chapter 15).
4. Have the patient practice the method of Taking in the White Qi, three times a day for 15 minutes each time to strengthen the Lungs.
5. Have the patient practice the method of Taking in the Blue Qi, three times a day for 15 minutes each time in order to strengthen the Kidneys.

LIVER CANCER

The Liver is the most usual site for the metastatic spread of tumors that disseminate through the Blood system. Carcinoma of the Liver is quick in development, high in mortality, and is regarded as one of the common malignant tumors.

ETIOLOGY

The etiology of Liver cancer is unknown. There are speculations suggesting that the following list of items may cause, or contribute, to the creation of Liver cancer:

- exposure to chronic stress and the suppression of anger,
- toxic poisoning (Liver cancer is usually associated with a long history of alcohol, drug abuse and chronic hepatitis), and
- a congenital weakness, with heredity being the predisposing factor.

SYMPTOMS

The Liver may have a single nodule, or multiple nodules. The main clinical manifestations include: distension and severe hypochondriac pain, hemorrhage of the digestive tract, and progressive hepatomegaly (enlargement). Its surface feels irregular on palpation (central depression or umbilications can often be detected).

TREATMENT

Purge the Excess Heat from the patient's Liver and Gall Bladder Channels, then tonify and regulate the patient's Kidney and Mingmen areas. Strengthen the patient's Spleen and Lungs, then regulate the patient's Liver.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see

Chapter 41), ending with the healing sound "Guo" for 24 times, 9 times a day.
2. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Walking Method for treating Liver disease (see Chapter 15).
3. Have the patient face, or embrace, a tree while practicing the method of "Taking in the Wood Element Qi," but only upon inhalation. With each exhalation the patient should purge the toxic Qi, sending it into the ground. This prescription is practiced three times a day for 15 minutes each time.
4. Have the patient practice the method of "Taking in the White Qi," three times a day for 15 minutes for each session to strengthen the Lungs.
5. Have the patient practice the method of "Taking in the Green Qi," three times a day for 15 minutes for each session to strengthen the Liver.
6. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes for each session to strengthen the Spleen.

ESOPHAGEAL CANCER

Carcinoma of the esophagus is one of the common malignant tumors, caused from a carcinoma of the squamous epithelium of the esophagus. This type of carcinoma is called "Ge Shi" or dysphagia in Traditional Chinese Medicine.

ETIOLOGY

The cause of esophageal cancer is unknown. There is speculation that the following list of items may cause or contribute to the creation of esophageal cancer:

- consumption of foods with a high content of nitrates,
- large quantities of spicy, hot foods, or mouldy food,
- chronically eating too fast, thus irritating the throat,
- strong alcoholic drinks,
- smoking,
- poor oral hygiene, resulting in tooth decay, and
- chronic stress and the suppression of emotions.

There are three types of esophageal cancer that invade the wall of the esophagus.

1. An ulcerous carcinoma of the esophagus that grows towards the outside of the esophagus. Dysphagia (difficulty in swallowing) is not evident in this type of cancer.
2. A tumor which grows in a circular formation. In this type of cancer, the lumen is narrow and easily obstructed.
3. The tumor attaches itself to one side of the esophagus, while the other side is still healthy and can still dilate. In this type of cancer, symptoms of obstruction appear slowly.

SYMPTOMS

More common in men over 40, carcinoma of the esophagus causes pain in the epigastric and sternal area. Swallowing aggravates the symptoms. The main clinical manifestations include the following symptoms:

- enlarged lymph nodes in the neck,
- difficulty in swallowing (dysphagia),
- excessive salivation,
- poor digestion,
- vomiting,
- loss of weight,
- pain in the chest and back, and
- emaciation and weakness.

TREATMENT

Purge the Excess Heat from the patient's throat, Lungs and Liver area. Tonify and regulate the patient's Kidney and Mingmen area, then regulate the Qi of the patient's esophagus, rooting it within the Lower Dantian.

PRESCRIPTIONS

1. Have the patient practice the "Dry Crying" exercise for a period of 15 minutes twice a day (see Chapter 53).
2. Have the patient practice the healing sound "Shang" for 24 breaths, 9 times a day (see Chapter 40).
3. Prescribe the Fast to Moderate Exhaling Breathing Method along with the Cancer Walking method for treating Lung disease (see Chapter 15).

4. Have the patient practice the method of "Taking in the White Qi," three times a day for 15 minutes for each session to strengthen the Lungs.

STOMACH CANCER

Carcinoma of the Stomach is regarded as one of the common malignant tumors. It is sometimes categorized as a carcinoma, lymphoma or sarcoma.

ETIOLOGY

The etiology for Stomach cancer is unknown. Speculations suggest exposure to chronic stress, and a suppression of emotions, may cause, or contribute to this disease. There is also an association between the development of gastric cancer and previous partial gastrectomies for benign diseases.

SYMPTOMS

The symptoms of Stomach carcinoma often do not appear until they have developed to the middle and late stages. Most common in men over 40, carcinoma of the Stomach causes a continual pain and burning sensation in the epigastric and sternal area, which is usually aggravated by swallowing. The main clinical manifestations include: shallow complexion, lack of appetite, loss of weight, extreme anemia, blood in the stool, debility, vomiting of partially digested food, and sometimes Blood that looks like coffee-grounds. The Liver may also be enlarged with nodules.

TREATMENT

Purge the Excess Heat from the patient's Stomach, Liver and Gall Bladder Channels, then tonify and regulate the patient's Heart and Spleen. End by regulating the patient's Kidney and Mingmen areas.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see Chapter 41) ending with the healing sound "Dong" for 24 times, 9 times a day (see Chapter 40).
2. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes each time in order to strengthen the Spleen.
3. Have the patient practice the "Taking in the Red Qi," three times a day for 15 minutes each session to strengthen the Heart.

PANCREATIC CANCER

Carcinoma of the pancreas is regarded as one of the common malignant tumors. Statistics state that 22% of deaths from gastrointestinal cancer are due to cancer of the pancreas.

ETIOLOGY

The etiology for pancreatic cancer is unknown. Speculations suggest cigarette smoking and diet (in particular, a high intake of fat and / or meat) may cause, or contribute to the cancer. Recent investigations have identified a number of factors that may also contribute to the cause of pancreatic cancer, including: environmental factors, surgical or associated medical factors, genetic factors, and occupational exposures to toxins.

SYMPTOMS

This type of cancer is more common in men between 50 and 70. A carcinoma of the pancreas causes pain in the epigastric area, which starts as a dull ache and usually becomes aggravated by lying down. The main clinical manifestations include dyspepsia, diarrhea and constipation, loss of weight, nausea, vomiting and fatigue.

The symptoms of pancreatic cancer can sometimes vary in accordance to its exact location. The head of the pancreas is close to the duodenum, while its tail reaches to, and energetically connects with, the Spleen. The middle of the pancreas lies horizontally across the upper aspect of the abdomen.

1. Carcinoma of the Pancreas's head has as for its main symptom jaundice, due to the common bile duct of the Gall Bladder becoming obstructed. This obstruction causes the bile to back flow into the Liver and thus contaminates the Blood. This causes both the Gall Bladder and Liver to swell. As the amount of bile and Blood increases, it causes the skin, mucous membrane and sclera to become yellow, and the urine to turn brown. During the later stages, the skin color changes from a yellowish color to orange or dark yellow.
2. Carcinoma of the Pancrea's body has for its

main symptoms sharp abdominal pain and swelling. The abdominal pain reacts to the compression of the abdominal cavity nerve. This type of pain can occur sporadically, radiating to the patient's shoulder and back. The patient will notice this type of pain becoming aggravated when sitting up or lying down. Bending forward, however, can ease the pressure on the celiac nerve plexus and reduce the pain.

3. Carcinoma of the Pancrea's tail usually does not have detectable, early symptoms. At times, it can only be discovered when the tumor has metastasized to the peritoneum, Lungs, bones or other organs. Once the tumor has developed to sufficient size, the manifestation of symptoms include anorexia, fatigue, and loss of body weight.

TREATMENT

Purge the Excess Heat from the patient's pancreas, then dredge the Stomach, Liver and Gall Bladder Channels. Next, tonify and regulate the patient's Heart and Spleen, then regulate the patient's Kidney and Mingmen areas.

PRESCRIPTIONS

1. Have the patient practice the Descend the Yang and Ascend the Yin Technique (see Chapter 41), ending with the healing sound "Dong" for 24 times, 9 times a day (see Chapter 40).
2. Have the patient practice the method of "Taking in the Yellow Qi," three times a day for 15 minutes each session to strengthen the Spleen.
3. Have the patient practice the "Taking in the Red Qi," three times a day for 15 minutes each time in order to strengthen the Heart.

COLON AND RECTUM CANCER

Carcinoma of the Large Intestine or rectum is regarded as one of the common malignant tumors of the digestive tract.

ETIOLOGY

It has been long accepted that colorectal cancer is caused, or promoted, by environmental factors, especially diet. It is suspected that carcinogens are present in the patient's feces.

Also, chronic ulcerative colitis is considered a predisposing factor for colorectal cancer. The following chronic diseases are believed to be related: benign tumors of the colon and rectum (especially frequently occurring Large Intestine polyps), and chronic colitis (especially chronic amoebic dysentery and chronic ulcerative colitis).

SYMPTOMS

Colorectal cancer is more common in patients over 50 years of age. A carcinoma of the colon causes pain in the lateral abdominal area, which also may be umbilical. The main clinical manifestations include a sudden change of bowel habit (constipation or diarrhea), loss of weight, fatigue and Blood in the stool.

Large Intestine cancer can be divided into two quadrants of right and left intestinal dysfunction.

- Cancer on the right half of the Large Intestine is measured from the cecum to the right half of the transverse colon. The early symptom of this disease is dull pain and discomfort within the right epigastrium, or lower abdominal area. As the disease progresses, the pathological changes of the tumor initiate poisoning symptoms which appear as: weakness of the body, poor appetite, loss of weight, and anemia. In the midterm, or late stages, the patient can sometimes feel the tumor on the right side of their abdomen (located where the right Large Intestine stores food residue from the Small Intestine).
- Cancer on the left half of the Large Intestine is measured from the center of the transverse colon to the junction of the sigmoid colon and rectum. Symptoms of cancer in the left half of the Large Intestine appear earlier because the frequency of defecation increases with mucus and Blood in the stool. Moreover, because the narrow opening of the Large Intestine is now occupied by the tumor, it is difficult for the fecal matter to pass through, causing inflammation, as well as alternating diarrhea and constipation. If the tumor continues to grow, the intestinal cavity can be clogged, causing abdominal distention and toxicosis of the whole body.

TREATMENT

Purge the Excess Heat from the patient's Large Intestine and rectum area, purging and dredging the toxic Qi out the body and down the Stomach, Liver and Gall Bladder Channels. Next, tonify and regulate the patient's Kidney and Mingmen areas, emitting Qi through the soles of the patient's feet. End with regulating the patient's Fire and Water Qi through the Fire Cycle of the Microcosmic Orbit.

PRESCRIPTION AND HOMEWORK

Have the patient practice the "Filth Dispelling Meditation," focusing on the Large Intestine and rectum area. The meditation should be practiced as follows.

1. Begin from a sitting posture, eyes closed, body relaxed, with the tongue placed up against the upper hard palate, behind the teeth.
2. The breathing should be natural and even.
3. Imagine Qi whirling in through the upper orifice of the Stomach, circulating clockwise through the Small Intestine, building momentum as it flows to the Large Intestine.
4. Imagine the whirling Qi driving through the colon. As it moves through the colon, the whirling Qi begins absorbing the noxious Heat and filthy Qi from the Large Intestine, transforming it into wind, and dispelling it out the anus. This is practiced for 10 repetitions.
5. Next, gently close the anal sphincter while inhaling; imagine the Qi in the lower abdomen again becoming transformed into wind. This wind begins circulating in a counterclockwise direction, spiraling up the body, through the Stomach, and is expelled out the mouth. This is also practiced for 10 repetitions.
6. Finally, focus the attention on the Lower Dantian; imagine the vital Qi returning back to its origin, and begin rubbing the abdomen clockwise to end the prescription.

CHAPTER 55

QI EMISSION THERAPY FOR RADIATION AND CHEMOTHERAPY

The purpose of Cytotoxic treatments (radiation and chemotherapy) is to kill the dividing cells responsible for the formation of the tumor mass. The well-known side effects of these types of therapy are loss of hair, loss of appetite, nausea, vomiting and damage to the patient's skin and gastrointestinal tract. When facing radiation and chemotherapy, the damage to the patient's immune system, although less obvious, is the major concern for the Qigong doctor. The cure of the cancer inevitably lies in activating and increasing the function of the immune response, so that it can recognize and eliminate the malignant tissue.

In China, Medical Qigong Therapy is commonly prescribed as an adjunct to both radiation and chemotherapy. The value of Medical Qigong Therapy is that it is very useful in reducing the harmful side effects caused by both radiation and chemotherapies, as well as enhancing the patient's immune system. Both radiation and chemotherapy assist the body in weakening a pathogenic cluster, but do not completely kill all of the tumor or cancer cells. That function is still left to the patient's immune system (which has consequently been weakened by the radiation or chemotherapy). Because Medical Qigong treatments, prescriptions and meditations enhance the body's immune system, its utilization as a adjunctive therapy is paramount.

MEDICAL QIGONG THERAPY AND RADIATION THERAPY

In most cases, radiation is much safer than chemotherapy because it can be directed into one specific area of the patient's body. However, it may also cause scarring that can interfere with the patient's internal organ function. Although radiation therapy is strong, its effect on the patient's body can be corrected by the Qigong doctor with dredging, purging, dispersing and tonifying techniques. The tissue damage left after radiation therapy should be rectified by dredging, purging and dispersing the afflicted area first. Then, after the toxic Heat has been purged, begin to tonify the patient's supporting organs and tissues surrounding the treatment area.

Sometimes the harmful side-effects of radiation therapy can be avoided by giving the patient Resonant Sound Therapy. In order to alleviate the patient's toxic Heat and pain, it is important for them to practice the "She" sound to cool the tissues after radiation treatments. After the organs have been sufficiently cleansed, the doctor may then tonify the patient's tissues, as well as the Wei Qi Fields, which have been damaged by the radiation therapy.

MEDICAL QIGONG THERAPY AND CHEMOTHERAPY

All forms of chemotherapy are cell-killing agents that damage the DNA and injure actively dividing cells, including those of the body's immune system. The tissue damage after chemotherapy is extremely extensive; the white blood cells decrease, causing secondary infections. Unlike radiation therapy, which is generally isolated to specific areas of the patient's body, chemotherapy affects all of the body's internal organs.

When treating patients who have undergone chemotherapy, it is important for the Qigong doctor to dredge, purge and disperse all the patient's Yin and Yang organs. As with radiation therapy, Sound Resonation Therapy is also mandatory in alleviating the patient's pain.

It is important for the patients to practice the "She" sound to cool the tissues after chemotherapy treatments. Specifically, cancer patients should pronounce the "She" Healing Sound therapy several times a day when their blood count is low, after both chemotherapy or radiation therapy. The sound "She" is pronounced six times. Its high pitched sound is in the straight tone, and its low pitched sound is in the dropping and rising tone. This sound is dropped and replaced by the sound "Ha," or the sound for the specific internal organ, as soon as the red and white blood count reaches a normal level.

After pronouncing the sound "Ha" for a certain period of time, add the sounds for the specific internal organs that have been affected (for example, add "Shang" for a patient with Lung cancer, and "Guo" for a patient with Liver cancer, etc.).

When treating the patient, after the diseased organs have been sufficiently cleansed, the doctor may then tonify the patient's Blood, Jing and tissues, as well as their energetic fields, which have been damaged by the chemotherapy.

RELIEVING SYMPTOMS ASSOCIATED WITH RADIATION AND CHEMOTHERAPY

After receiving radiation or chemotherapy, patients may suffer from several internal complications due to the severe organ and tissue trauma caused by these therapies. The following are six common problems resulting from radiation and chemotherapy.

1. A gastrointestinal infection can result from both radiation and chemotherapy, which have a tendency to burn up the body's Yin. This frequently results in causing Rebellious Qi, which is accompanied by such symptoms as nausea, vomiting, and poor appetite. Treat by tonifying the patient's Stomach Yin by prescribing herbs, and use Jing Point Therapy to quiet the digestion by tonifying Pc-6, St-36 points, as well as the Ear Shenmen points.
2. An infection of the Large Intestine can occur from the accumulation of toxins in the Lower Burner creating Damp Heat Symptoms (caused from radiation and chemotherapy). These toxins can cause abdominal pain, constipation or diarrhea with bloody mucous. Treat by tonifying the patient's Yin to prevent loss of Body Fluids (also give astringent herbs), and use Jing Point Therapy to stop any internal bleeding, by tonifying St-25, St-36, St-37, St-38, LI-4 and LI-11 points.
3. A Urinary Bladder infection can be caused by both radiation and chemotherapy that have a tendency to create Damp Heat in the Lower Burner. Damp Heat in the Lower Burner can result in cloudy, bloody and scanty urination, with a feeling of an urgency to urinate but an inability to do so. To treat, use Jing Point Therapy and tonify Sp-6, Sp-9, CV-2, CV-3, Bl-23 and Bl-29 points.
4. Edema in the limbs can occur from both radiation and chemotherapy that have a tendency to create Wind Damp and Deficiency, which leads to an accumulation of Body Fluids in the extremities. Treat by tonifying the Spleen and draining the body of Dampness. Use Jing Point Therapy to tonify Sp-6, Sp-9, Lu-7, LI-4, and Bl-20; also prescribe herbs.
5. A diminished white blood cell count can be caused by radiation and chemotherapy and can lead to secondary infections. Treat by tonifying the patient's Blood and Jing, and prescribe herbs.
6. Pneumonia can be caused by radiation and chemotherapy that have a tendency to cause damage to the Lungs. When both the Lung Qi and Yin Qi are damaged, pneumonia can result. Treat by tonifying the patient's Yin and Lung Qi, and prescribe herbs to enhance the immune system.

CHAPTER 56

QI EMISSION THERAPY FOR SURGERY

Medical Qigong Therapy is useful for treating patients before, during and after surgery. Qi emission can be used to reduce the patient's bleeding, enhance the immune system, minimize the risk of infection, strengthen the body, and to accelerate the recovery rate. Medical Qigong modalities are involved in preoperative therapy, surgery, postoperative therapy, follow-up therapies and remedial prescriptions.

MEDICAL QIGONG AND PREOPERATIVE THERAPY

Although Medical Qigong Therapy has had incredible results in healing and in stopping the progression of certain diseases, it is not a "cure-all." As with other clinical modalities, it has its own limitations. Currently, conventional medicine only utilizes three modalities for treating tumors and cancer; these include surgery, radiation and chemotherapy. In the healing process, undergoing surgery is sometimes unavoidable, and should never be viewed by the patient as a defeat, but only as a necessary step in the healing transition.

Having to face surgery should not be viewed as a negative reflection of a Qigong doctor's healing potential, nor as a patient's inability to heal. Surgery should be viewed as a necessary life saving intervention, that can firmly establish the patient on the road towards health and recovery. It is a positive affirmation of the patient's commitment to getting well and taking responsibility for the healing. Sometimes, surgery is a necessary step towards health and healing.

Before surgery, it is important for the doctor to consider the emotional relationship that the patient has developed with his or her diseased tissues. The mental, emotional, energetic and spiritual aspects of a disease must not be ignored.

The doctor's involvement with the patient and the tissues which are about to be surgically removed is extremely important, since this emotional interaction affects the patient's healing potential. For healing to take place, a trusting relationship with the patient must first be solidly established; this includes a compassionate attitude of the doctor towards the tissues that will be removed. The doctor models the correct attitude to facilitate an accelerated healing for the patient. It is believed that scar tissue formation is in direct proportion to the patient's fears and negative expectations surrounding the surgery. These fears and expectations form energetic and spiritual attachments to the physical body. Often the patient attempts to disconnect these spiritual attachments out of fear and survival instinct.

A patient is spiritually, emotionally and energetically interconnected with his or her body. This energetic attachment includes any and all organs, organ systems, or areas of the body which can be involved in the operation.

A patient may sometimes feel disillusioned about his or her own interpersonal relationship with the body, and try to disconnect all energetic attachments to the body's tissues out of fear and survival instinct. It is important for the Qigong doctor to help the patient get in touch with the suppressed feelings of loss surrounding the tissues which will soon be operated on. This is because surgery represents the death of a relationship with part of the self. Although the relationship that the patient has with the diseased organ is not considered a healthy one, it exists nevertheless. Allowing the patient to deny his or her true feelings related to the surgery interferes with the healing process. Denial may lead to increased scar tissue formation, and can promote the return of certain disease formations.

The Qigong doctor therefore encourages the patient to talk to, and "brief," each organ and tissue area regarding the surgical procedure to prevent further denial, and to promote the release of feelings. Otherwise, the patient may experience feelings of loss, grief and depression; this reaction can of course also be due to the aftereffects of the anesthesia. It is important that these feelings be experienced and expressed before the surgery. After the surgery, the patient may then experience feelings of relief and gratitude about the new changes in his or her body.

To prepare for surgery, the patient is usually given Medical Qigong meditations and prescriptions that strengthen the Lower Dantian, Kidney and Mingmen areas. One, or several sessions may be specifically devoted to allowing the patient to connect with the diseased organ(s), as well as the regions that are to be operated on. A request for divine intervention can facilitate the patient's relinquishing control, and for the removal of any and all energetic armor surrounding the diseased area (and the subsequent release of emotions associated with this armoring). The patient is then encouraged to disconnect any energetic attachments that the diseased tissues have formed with the surrounding cells, facilitating a final closure. This gradual process allows for an easier transition, as the patient's body must completely release any attachment to the diseased tissues during surgery.

Just before surgery, the patient is given a meditation, and encouraged to imagine a divine white light energy submerging and enveloping the regions of the body which are to undergo the operation. The Lower and Middle Dantians, as well as the areas which will soon be operated on, should also be energized. In addition, the Liver should also be energized to make sure that the Blood coagulates properly during the surgery.

USING MEDICAL QIGONG FOR ANESTHESIA

According to research on Western Surgeries (presented at the Second International Symposium on Memory and Awareness in 1992), patients who undergo anesthesia may be aware of both the pain and conversations within the operating room. Many kinds of drugs used to supplement anesthesia, such as neuromuscular blocking agents, paralyze all the muscles in the body but do not affect the central nervous system and consciousness. During surgery, patients have been observed flinching, twitching and making other facial expressions. Many patients later can give an accurate detailed account of the surgical procedure and conversations between the doctors and nurses. Even when unconscious, patients (whether energetically inside or outside of their bodies), can record, see and remember everything.

Qigong energy emitted during surgery, can however, reduce pain and wound shock, as well as calm the patient's Shen. Qi emission also reduces postoperative complications such as respiratory tract infection, functional disorders of the gastrointestinal tract, retention of urine, etc. The energy received by the patient also speeds up the healing process.

Medical Qigong therapy is found to be very useful in relieving pain by affecting the body's neuro-chemicals. The energy stimulates the peripheral and cutaneous nerves that carry sensory information, via the spinal cord, to the brain. This stimulation of the cutaneous nerves activates neurotransmitters which carry pain messages to the brain, and facilitates the closure of the body's pain-relay gates. The brain produces endorphins, or endogenously generated morphine-like chemicals in response to this closure. The body's opiate receptors are then able to dull the pain. This process is the basis for Qigong Anesthesia. Because Medical Qigong therapy safely produces an analgesic affect on the body's cutaneous tissues, its use in hospitals for surgery as well as for pre and postoperative procedures is gradually increasing.

The first operation in which Qigong anesthesia was applied (recorded in modern times), was performed successfully for the removal of a thyroid tumor in the Shanghai #8 People's Hospital, on May 9, 1980. Qigong Master Lin Hou-sheng demonstrated the ability to induce anesthesia during surgery by pointing his fingers at specific acupuncture points, using external Qi projection without the use of any Western anesthetic. The patient's blood pressure, pulse and respiration rate remained stable during the operation, and the patient's physiological function remained normal after the operation. The success of the Qigong anesthesia boosted the resident doctor's confidence and laid a solid foundation for future applications of Qigong anesthesia in surgical operations.

Qigong anesthesia has begun to follow in the wake of the widely accepted acupuncture anesthesia, as a holistic alternative to conventional pharmaceutical methods. The number of cases treated in China with Qigong anesthesia is steadily increasing, primarily for the following six benefits.

1. Arresting the patient's pain,
2. Resisting the pulling reflex action or shock reaction of the body's internal tissues and organs when the scalpel is inserted,
3. Reducing infection,
4. Reducing the body's opposition or contrary action/reaction to operative wounds,
5. Preventing wound shock (response of the body to the tissue trauma), and
6. Promoting the healing of wounded tissues.

Since Qigong analgesia is effective through the stimulation of certain points on the body's surface, both the location of the points, and the type of energetic application are crucial. There are three methods commonly used in selecting points: According to the Channels, According to the Segmental Innervation, and According to the Auricular Points.

1. The analgesic approach used according to the patient's channels includes two main applications:
   - The selection of points according to the course of the channel which traverses the site of the operation, and
   - The selection of points according to the differentiation of symptoms and signs of a disease, as well as the responses that may be elicited from the patient's body during the operative procedure.
2. The analgesic approach used according to the segmental innervation includes three main applications:
   - The selection of points according to the adjacent segment, or an area that is supplied by the same spinal nerve, or an adjacent spinal nerve of the operative site,
   - The selection of points according to the remote segment, in an area not supplied by the same spinal nerve, or an adjacent spinal nerve (on the operative site), and
   - The selection of points according to the stimulation of the nerve trunk within the same segment, stimulating directly the peripheral nerve of the operative site.
3. The analgesic approach used according to the Auricular Points employs the insertion and stimulation of acupuncture needles with Medical Qigong Therapy, and includes three main applications:
   - The selection of points according to the corresponding Auricular areas pertaining to the operative site and its involved internal organs,
   - The selection of points according to the theory of the Yin and Yang Organs pertaining to the operative site and its involved internal organs, and
   - The selection of points according to such things as tenderness, reduction of electro-resistance, the appearance of deformation and/or discoloration when a certain internal organ or area of the body is affected.

It is important preoperatively that the patient have confidence in the doctor's method of Qigong application. Any misgivings should be dispelled by explaining the procedure to gain the patient's full corporation and trust. In order to know the level of response that the patient will experience, it is necessary to test the patient's response to Qigong anesthesia prior to any surgery.

To use external Qi as an anesthetic, the Qigong doctor must first focus Qi into his or her Lower Dantian, increase its charge and density and then extend it out through the palms, or sword fingers, into the patient's body through specific channel points. The goal is to numb a specific area on the patient's body without touching the patient.

When Qigong anesthesia is being administered successfully, 96% of the nerve fibers responsible for pain sensation become anesthetized and the patient loses consciousness within ten minutes. Ten minutes after cessation of the Qigong anesthesia emission, 24% of the anesthesia has disappeared; while 72% of the tissue remains numb. The remaining tissue quickly recovers from the anesthetic effect and the patient quickly returns to consciousness.

Patients suffering from pain caused by such conditions as cancer or dysmenorrhea also respond positively to Qigong anesthesia.

MEDICAL QIGONG THERAPY AND SURGERY

Surgery has been performed in China for thousands of years. In ancient times, the legendary physician Hua Tuo was noted, not only for his expertise in Medical Qigong, acupuncture, massage and herbs, but also for his surgical skills as well. In Western culture, surgical therapies are applied for the removal of tumors, cysts, and abnormal tissue growth, as well as the transplantation of organs (heart, lungs, and kidneys) and the replacement of dysfunctional systems (joint replacement). Surgery, however, having a potential to both hinder and obstruct the body's flow of life force energy, is generally considered a last resort modality.

During surgery, the Qigong doctor assists the surgeon by applying External Qi Emission to the patient's body, energizing and strengthening specific areas. This is done to reduce pain and alleviate or eliminate the pulling reflex action (shock reaction) of the body's internal tissues and organs when the scalpel is inserted, prevent infection, and reduce bleeding.

When assisting surgeons during an operation, I personally have found that it is important to maintain a physical connection with the patient's body before beginning the Qi emission, until the anesthesia has fully taken effect. As the anesthesia relaxes the patient's tissues, the Qigong doctor can feel and see the patient's spirit leave the body.

Just before the scalpel begins its first incision, it is important for the Qigong doctor to physically disconnect from the patient's body. Even though the Qigong doctor has physically disconnected from the patient's tissues, it is extremely important for the Qigong doctor to increase his or her energy extension deeper into the center core of the patient's body. This protects the Qigong doctor from experiencing the shock, and initial trauma from the scalpel's first incision, which resonates throughout the patient's entire body.

It is also important for the Qigong doctor to focus the attention on extending and circulating his or her energy deep into the patient's Taiji Pole while the surgery is being performed. As the patient begins to lose Qi and Blood due to the operation, the Qigong doctor replenishes the patient's lost Qi, as well as energizes the incoming Blood from the transfusion. The Qigong doctor should also slow down the bleeding by slowing the patient's pulse.

During surgery, the surgeon's scalpel will sever several of the patient's channels as the doctor cuts through the surface of the skin and continues into the deeper regions of the patient's body. Each channel that is being disconnected will weaken the patient's energetic field corresponding to that specific organ or organ systems. It is the Qigong doctor's responsibility to maintain connection with the patient's energy field, and also to monitor and maintain the patient's energy level. The Qigong doctor maintains this connection as the surgeons open up the patient's body cavity, move internal organs, remove the diseased tissues (and any associated structures), and then sutures and closes the patient's body cavity.

When the patient is being sutured, the Qigong doctor must facilitate the patient's circulation of Qi through the Microcosmic Orbit; this stimulates the re-connection of any of the patient's internal and external channels which have been severed. The Qigong doctor must also change and rebuild the Energetic Grid formations in the body's internal and external energetic fields. The patient's Energetic Grid is reconstructed to a pre-diseased pattern to prevent the regrowth of the disease.

MEDICAL QIGONG AND POSTOPERATIVE THERAPY

The separation and restructuring of the inner fasciae that occurs during surgery can cause serious postoperative problems by either stopping the Qi from flowing (causing stagnations), or altering the energy's natural course (causing deviations). This is why after the surgery has been completed, it is extremely important for the Qigong doctor to immediately dredge and disperse the Turbid Qi from the patient's body. This is performed by energetically combing over the fresh incisions, to reconnect the patient's energetic pre-operative patterns. Energetic combing also promotes the rapid healing of the patient's wounded tissues, reduces the formation of scar tissue, and continues to alleviate the effects of the surgery.

To facilitate an escalated healing, the Qigong doctor must also energize the patient's Kidneys and any energetic field which may have been depleted. Also, to bring the patient's temperature back to normal, and to disperse any feelings of nausea after the anesthesia and shock of surgery wear off, certain Heart, Pericardium and Stomach Channel points require stimulation.

It is important to note that after an operation, the patient's Dantians and all the body's tissues have been depleted, and are in need of revitalization. While the patient is healing, the color of the aura (located in the Wei Qi field surrounding the area of the surgical trauma) is usually dark to light gray. To increase the healing potential, the Qigong doctor gives the patient specific prescriptions in accordance with the internal organ(s) affected, such as white light meditation images. This is prescribed to energize and replenish the patient's damaged tissues and energy fields. The Qigong doctor also continues to work on rebuilding, reconnecting and reinforcing the patient's channels and Energetic Grids which the surgeon has severed during the operation.

MEDICAL QIGONG THERAPY AND RECREATING ORGAN ENERGY

Once an organ has been removed, there remains an energetic "void" within the spatial cavity of the patient's tissues. The energy of the organ, however, still exists within the patient's body as a "Phantom Organ." This energetic phenomena is due to the creative "blue-printing" of the prenatal and postnatal formation of the body's tissues. Because mass and energy are interconnected and interchangeable, it is therefore possible to energetically stabilize, as well as energize, the spatial cavity wherein the surgically removed organ once resided. To assist the patient in recreating active organ energy, the Qigong doctor prescribes the following meditation:

Lungs	Brilliant Pure White
Large Intestines	Abalone-Shell White
Stomach	Deep Yellow
Spleen	Orange Yellow
Heart	Red with Blue Tint
Small Intestine	Pink
Bladder	Deep Blue
Kidneys	Light Pastel Blue
Pericardium	Purple Red
Triple Burners	Orange Red
Gall Bladder	Pale Chartreuse
Liver	Dark Green

Figure 56.1. Organ Color Chart

STAGE 1 - THE HOOKUP

Generally several days after the surgery, while the patient is still recovering, the Qigong doctor has the patient imagine him or herself being connected to the Heavens and rooted to the Earth. Next, the patient uses color visualization.

STAGE 2 - CHOOSING THE APPROPRIATE COLOR

To prescribe the correct color, it is important for the Qigong doctor to know which organ or organ systems have either been removed, or seriously affected by the surgery. The patient is then assigned specific color meditations to tonify and strengthen the energetic fields that support the traumatized tissues. These meditations are also used to reestablish a physical, mental, emotional, energetic and spiritual connection to the area of trauma.

Figure 56.1 categorizes the general colors associated with the body's organs and channels. These colors can be viewed as extensions of energy originating from the original organs.

STAGE 3 - CLEANING AND ENERGIZING THE SPATIAL CAVITY

After the doctor has selected the proper organ color, the patient is given specific images used to cleanse, purify, and energize the tissues. These meditations create a solid energetic transformation that stabilizes the patient's energetic field.

1. The Cleansing and Purifying meditation is as follows. The patient imagines that his or her fingers and toes are straws. As the patient inhales, he or she will imagine divine white light energy flowing from the Heavens and Earth into the fingers and toes, filling the specific area which has been traumatized by the surgery (Figure 56.2). While exhaling, the patient imagines any pathogenic Qi leaving the tissues and body via the fingers and toes. The patient continues to perform this cleansing meditation for several minutes until he or she feels that the specific area has been purified.
2. The Creating and Energizing meditation requires that the patient imagine a clean, vibrant color (the Five Element Color associated with the particular organ), flowing into the spatial cavity, and areas, surrounding the removed organ. As the patient inhales, the organ color fills the vacant cavity, energizing the remaining energetic structure and surrounding tissues. The patient continues with this imagery until the specific organ area is full of heat, light and vibration (Figure 56.3). Because the surgery has permanently removed the organ from the patient's body, it is important for the patient to continue with this meditation every day for the rest of his or her life.

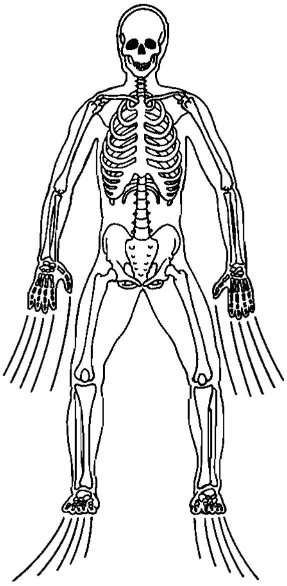

Figure 56.2. As the patient inhales, he or she will imagine the divine white healing light flowing into his or her body through the fingers and toes. This image is used to purge the patient's body and cleanse the tissues.

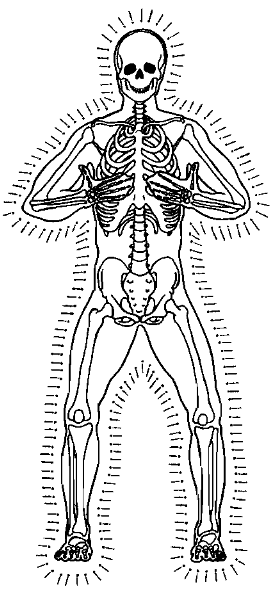

Figure 56.3. Once the patient is cleansed, in order to energize the tissues, he or she will imagine a clean vibrant energy color flowing into the specific spatial cavity associated with the removed organ.

MEDICAL QIGONG THERAPY AND ORGAN TRANSPLANTATION

Each particular organ in the human body has its own unique history and experience. This emotional memory and experience is active in the donated organ, as well as in the patient's diseased organ. Before the body can, or will receive a foreign organ, there must first be a time of reconciliation between the memories stored within the negative (as well as positive) physical cells of the organ soon to be replaced, and that of the donor's organ. To assist the surgeon in transplanting an organ, the Qigong doctor must perform visceral cleansing, preparation and purification of the transplanted organ. This method is explained in the following three stages:

STAGE 1

The patient must release all attachment to the departing organ. This includes all negative (as well as positive) physical, mental, emotional, energetic and spiritual memories and experiences.

STAGE 2

The history and emotional memory of the transplanted organ must be cleansed. This includes all negative (as well as positive) physical, mental, emotional, energetic and spiritual memories stored within the donor's organs.

STAGE 3

The relationship between the patient's tissues and the spatial cavity wherein the initial diseased organ was, or is located, must be established and harmonized.

The patient's body will either receive or reject the transplanted organ due to unresolved emotional trauma stored within the patient's tissues, or within the transplanted organ itself. Current research reveals that emotions trapped within transplanted organs can surface within the patient's subconscious mind, affecting his or her behavioral patterns. This is manifested through certain unnatural behavioral traits, which are transferred via the donated organ to the patient.

A famous example of a lingering emotional connection between an organ donor, the organ, and the recipient is described in the following true story. A few years ago a young woman (vegetarian) received a heart transplant from a young man who died in a motorcycle accident. Upon being discharged from the hospital, the woman began craving Chicken McNuggets, which was the young man's favorite food. This was in direct opposition to her previous diet and eating habits.

It is important to note here that organ rejection is not always due to the donor or patient's unresolved emotional issues. Organ rejection can also be caused from biological reactions complicating the acceptance of the foreign tissue.

FOLLOW-UP AND REMEDIAL PRESCRIPTIONS

The follow-up and remedial prescriptions will be determined according to the specific type of surgery the patient has undergone. The prescriptions must be designed to purge, strengthen and regulate any organ(s) that have been weakened by the surgery. Follow-up and remedial prescriptions must also circulate the patient's energy to reduce any energetic stagnations which may have been caused by internal organ movement and fascia damage. This will assist the patient in an escalated recovery, as well as reduce postoperative depression, which often follows after major surgery.

Sometimes, when assisting a patient who is healing from a severe disease (i.e., malignant tumors), additional Western therapies such as radiation or chemotherapy are required. The following is a testimonial by a patient of mine, Thea Sagen (female - age 50), illustrating the use of Qigong therapy in combination with radiation therapy.

Here is a summary of events surrounding my pancreatic surgery, June 16, 1997 at Community Hospital of the Monterey Peninsula.

"On June 16, 1997, emergency surgery to remove a fist-sized tumor from the tail of my pancreas, resulted in my former Sifu (Martial Arts Master), Jerry Alan Johnson, changing roles to become an "Energy Therapist." Thanks to my admitting physician, Dr. Stephanie Taylor, who facilitated the arrangement, Dr. Johnson was the first representative of Chinese Medical Qigong therapy allowed to enter surgery at the Community Hospital of the Monterey Peninsula.

As my Energy Therapist, Dr. Johnson conducted a preoperative session during which he facilitated the energetic and emotional release of my tumor. He then accompanied me into the four hour surgery, during which he maintained my energy levels, while reminding my tissues of their preoperative energetic patterns in order to facilitate rapid healing. He continued to work with me through the postoperative healing, and followed up several days after surgery. It was his vital support and insights that helped me to undergo the recommended course of radiation therapy.

The following is a synopsis of the Medical Qigong therapy that I experienced along with the traditional Western modalities used to treat my pancreatic islet-cell tumor.

Phase 1. The Preoperative Stage. Consisted of a Medical Qigong session directed toward putting me in touch with the tumor, my relationship to it, and the emotions associated with it. This was followed with an atonement with the Divine and relinquishment of control, removal of the armoring around my tumor, and subsequent release of emotions with it. I actually "saw" the disconnected energetic attachments of the tumor as they detached from my organs and were blowing like ostrich feather segments in a wind. I also saw a bright neon sign on the tumor that identified its emotional source. This entire process took only an hour, yet it facilitated closure with the tumor in order for me to easily release it during the surgery, which immediately followed the session. It also provided me with insights that helped my emotional healing over the months following surgery.

Phase 2. Surgery. I entered surgery without any sedatives and was filled with a tremendous sense of peace and empowerment in knowing that the hospital was honoring my energetic, as well as my physical needs. The surgery lasted for four hours during which Dr. Johnson connected with my energy fields and monitored, as well as maintained, my energy levels. He continued treating me, and working with my energetic fields, as the anesthesia was administered, and as the two surgeons opened my cavity, moved structures, removed the tumor (along with the tail of my pancreas and the spleen), and then sutured and closed my abdominal cavity.

Throughout the surgery Dr. Johnson stood at my head, where energy enters through the crown, and facilitated the running of energy into my center channel in the core of my body. He described his work as following the actions of the surgeons, and my body's response to these actions: the incision, the movement of internal organs, the resection of the tumor, and the final suturing. He used this opportunity to change the energetic grid formations in the tumor bed to prevent regrowth of the tumor, which was malignant.

Phase 3. The Postoperative Stage. This stage involved a continuation of the mitigation of the effects of surgery. He worked to tonify my Kidneys, to bring my temperature up, and to offset the shock of surgery, as well as used acupressure points to calm my heart and stomach during the nausea. I was reminded upon awakening, to do my Microcosmic Orbit exercise. Dr. Johnson then proceeded to "dredge" or comb over the incision to reconnect my pre-energetic channel and tissue patterns in order to facilitate healing, and to reduce scar tissue. The healing of the incision was fast. My doctors later remarked at how quickly I healed, and how little scar tissue there was compared to similar surgeries.

Phase 4. Follow-Up and Remedial Prescriptions. During the days following surgery, Dr. Johnson taught me several meditations: The "Divine Healing Light Meditation" for healing the tumor bed and incision, and the "Heaven and Earth Spleen Meditation" that fuses a ball of energy where the spleen once was – allowing me to continue to produce the energetic pattern of the spleen within my body. I began doing these meditations the day following surgery for the week of hospitalization, and continued to practice them along with the Microcosmic Orbit exercises throughout my recovery.

Phase 5. Radiation. After several weeks of recuperation, I was scheduled to received radiation treatments daily (Monday through Friday) for six weeks. Each day I would perform the Medical Qigong exercises (vocal and physical prescriptions) outdoors in a field overlooking the ocean.

Then I would drive to the hospital for the 15 minutes of radiation, followed by a session with Dr. Johnson every other day. I understood that the Medical Qigong exercises and sessions would offset the heat in the tissues while protecting the surrounding organs from unwarranted harm. There was concern about the possible loss of function in one kidney, but blood tests during the treatments showed no noticeable effects on them, or on my liver. In addition, unexpected by-products of the Medical Qigong sessions involved the release of emotional memories and an experience of the Divine in my healing process.

On alternate days after radiation, I had acupuncture and pranic healing with Jason Gill L.Ac. who teaches classes in Chinese Herbs and Traditional Chinese Medicine at Dr. Johnson's International Institute of Medical Qigong. Through these treatments, and the use of Chinese herbs, my immune system was strengthened. These complimentary modalities, combined with my body's own healing abilities and visualization practice, enabled me to avoid the typical side-effects of nausea and diarrhea. Aided by the positive energy from these two therapists, I was able to maintain my energy levels in sufficient quantity to continue driving to the radiation treatments on my own.

It is hard for me to imagine what this entire experience might have been like without the support of a doctor of Medical Qigong. I am truly grateful for Dr. Johnson's dedication, discipline and expertise that enabled him to be present for me throughout all aspects of this illness. It was a gift that I would encourage other patients to utilize for themselves in combination with any traditional allopathic program for cancer. I had the best of both worlds working for me in a remarkable manner thanks to all of my doctors."

PHANTOM PAINS OF AMPUTATED ORGANS

Amputation may be necessary as a result of severe injury or massive infection, such as gangrene. During surgical amputation, the soft tissues are cut in such a way that they cover the end of the remaining bone to heal completely. This allows the body's tissues to seal and protect themselves from the external environment. "Ghost" or "phantom" pains are caused by the patient's energetic connection with the amputated tissues, through the internal "energetic spacial cavity," or external "energetic boundary system." The phantom limb (usually one of the body's extremities, e.g., finger, hand, arm, etc.) exists in an energetic state and continues to interact with the patient's external Wei Qi field, as well as with the patient's Shen. Any interaction with the space, or energetic field of the amputated organ, will produce a direct response in the patient, similar to that of contacting the actual physical limb.

Rehabilitation programs generally do not prepare the patient for such reactions as phantom pains, which tend to leave the patient perplexed and frustrated. To alleviate the symptoms of phantom pains, the Qigong doctor sedates the patient's Wei Qi field that surrounds the amputated limb, and assists the patient in re-patterning the old energetic field to fit the postoperative tissue formation. The Qigong doctor can also prescribe Juniper Berries, which will relieve the phantom pain symptoms.

MEDICAL QIGONG THERAPY AND THE TREATMENT OF SCAR TISSUE

Medical Qigong therapy may be effective for treating pathologies caused by scarring due to surgical procedures. The formation of scar tissue often becomes the source of other physical problems by either obstructing, or weakening, the flow of Qi within the body's channels. In most cases of pathological scar tissue, the scar itself transverses several channels within the superficial and deep tissues.

The topological relationships of each scar will vary according to each patient. Therefore, different therapeutic approaches may be required for treatment, depending on the patient's strength, weakness, Elemental Constitution, and the patient's emotional state surrounding the surgery. Manifestations of scar tissue pathology include some of the following symptoms:

- a feeling of discomfort, or unpleasant sensations that originates around the scar area,

- radiating sensations that occur when the scar is being pressed, and
- a sharp, pulling type of pain that originates from the scar.

These conditions can be present, even if it has been several years since the surgery was performed.

TREATMENT

Medical Qigong treatment for scar tissue formation follows the principle of emitting Qi along the pathway of the channels. Before treatment, the doctor should take into consideration the energetic flow and function of the channels that have been affected by the surgery. If, for example, a patient's Yang channel (e.g., Stomach) becomes obstructed due to surgery, it can create an Excess condition within the Stomach Channel, which in turn creates a Deficiency of the Spleen Channel. Therefore, the treatment of both the Stomach and Spleen channels should be addressed.

After surgery, some scar tissue formation is inevitable, since both the body's inner-facia and channels (which are responsible for the lubrication of the tissues) have been severed. To minimize the degree of scar tissue formation, the Qigong doctor uses Qi extension techniques on the sutured area, extending his or her intention deep into the patient's body. The doctor's focuses on dispersing stagnations and regulating the flow of the undercurrent of Qi through the sutured tissues. Emphasis should be placed on reconnecting the patient's channels and collaterals. Next, the doctor focuses on regulating the external and internal flow of energy from within and without the patient's tissue area.

In treating scar tissue formation which has laid dormant for quite some time, tissue regulation through Tui Na therapy is required. After using the proper tissue regulation method (see Chapter 37), the Qigong doctor ends the treatment by using the One Finger Skill technique along the line of the patient's external scar. To perform this method, one thumb is placed on the end of the scar, while the doctor's other thumb presses into tissue, slowly outlining the entire scar (this heats and energizes the patient's tissues). The doctor ends the treatment with both thumbs placed at the beginning and ending of the scar, extending his or her energy through the patient's tissue in a circular pattern. This technique naturally induces an energetic emotional release of the memories trapped within the scar tissue.

SECTION XII

MEDICAL QIGONG APPENDICES

INTRODUCTION TO THE APPENDICES

There are many documents discussing the experimental research of Medical Qigong Therapy written within the past ten years. Among these documents, are listed a collection of abstracts from five national congresses on Medical Qigong researches as well as four World Conferences. The research has covered a variety of topics ranging from Medical Qigong and its relationship with: Physics, Chemistry, Biology, Western and Traditional Chinese Medicine, Optics, Magnetism, Electromagnetism, Acoustics and the Projection of Sound Resonation, Fluid Mechanics, Nuclear Physics, Electro-Chemistry, Organic Chemistry, Inorganic Chemistry, Analytical Chemistry, Physiology, Bio-Chemistry, Molecular Biology, Micro-Biology, Cytology, Genetics, Pathology, Pharmacology, Psychology, and Sports Medicine.

The experimental research of Medical Qigong is actually a study with multiple sciences. Of all the research reported, Biology and Medicine are in the forefront of the material collected.

The following research was conducted by distinguished scholars at various Universities, Scientific Academies, and Institutes through out the world (i.e.: Beijing Medical University, Qinghua Medical University, Nanjing Medical University, Tokyo Medical University, Tokyo Electro-engineering University, Higher Energy Institute, Physics Institute, International Institute of Medical Qigong, China Healthways Institute, etc.).

This research was focused on the Extension of Emitted Qi by various Medical Qigong Masters / Doctors, the result being positive in it's conclusion on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality.

This section of the text is devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality.

For this purpose, I have elicited the help of my good friend, the renowned scientist and research investigator, Dr. Kenneth M. Sancier. Dr. Sancier (a Material Scientist) is the president of the Qigong Institute, of San Francisco, California, and an active Council Board Member of the World Academic Society of Medical Qigong (Beijing, China). His many years of extensive research in the field of Energetic Medicine has led him to help organize the first International Congress of Qigong in conjunction with the Fifth International Congress of Chinese Medicine (at the University of California, Berkeley) including maintaining an active annual Qigong Science Forum for scientists. He is also responsible for developing the first computerized Qigong Data Base, which is currently being sold worldwide.

Dr. Sancier has actively collected scientific information on Qigong by attending international conferences, and meeting scientists in China, Japan, Canada and the United States. He has published several reviews of clinical and experimental studies of the medical aspects of Qigong therapy in many scientific journals and magazines. Although the following information in this appendix represents but a fraction of Dr. Sancier's findings, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

APPENDIX 1A
MEDICAL APPLICATIONS OF QIGONG

Kenneth M. Sancier, Ph.D.
Qigong Institute
Menlo Park, California

ABSTRACT

This article reviews selected scientific studies of medical applications of Chinese Qigong. The intention of the review is to outline research on Qigong and its potential for improving health care in western countries. The review centers on clinical and experimental studies to show that Qigong exercise can beneficially affect many functions of the body and improve health. The studies were selected to illustrate the following points: medical applications of Qigong are diverse, some studies were conducted in depth, and many applications hold promise to improve western health care.

Several specific clinical and experimental studies are outlined. The study using electroacupuncture according to Voll shows that almost all organs and functions of the body can be balanced by Qigong. The research on hypertension serves as a model to illustrate the many ways that the effects of Qigong on the body can be measured. Studies on brain waves are included to show that the mind is involved in Qigong and probably in the healing process. Research on Qigong's role in high altitude adaption suggests some practical applications for treating cardiovascular disease. A combination of Qigong and drug therapies is superior to drug therapy alone, as indicated in the case of two diseases, hypertension and cancer. In conclusion, there is a plea to medical practitioners and scientists to collect existing information on medical Qigong, promote research if needed, and identify applications that hold promise for improving western health care.

INTRODUCTION TO MEDICAL QIGONG

Traditional Chinese Medicine (TCM) is a holistic system for promoting health and for healing that includes several therapies such as acupuncture and moxibustion, herbal medicine, acupressure massage, nutrition, and Medical Qigong. These therapies are often used in combinations. The underlying theory of TCM is based on balancing Qi, the vital energy in the body, according to the theory of "yin-yang" and the Five Elements, that has been used for more than three thousand years. In practice, the flow of Qi is regulated, and "blocks" to the flow of Qi are removed. Energy blocks or Qi in excess or deficiency may result from disease, injury or stress. Qigong (pronounced chee gong) has a history of thousands of years, and it is unique among TCM therapies because almost anyone can learn and practice Qigong. However, it best to study Qigong with a qualified teacher to avoid adverse side effects. It is claimed that 60 million people in China practice Qigong daily, primarily to maintain health and achieve long life.

The word Qigong is a combination of two ideas. Qi is the vital energy of the body, and gong is the skill of working of the Qi. Medical Qigong for health and healing consists primarily of meditation, physical movements, and breathing exercises. Qigong practitioners develop an awareness of Qi sensations in their bodies and use their mind, i.e., intention, to guide the Qi in the body. It is said that the benefits of Qigong extend beyond health and healing to enhance spiritual life and even special abilities, such as psychic powers. Qigong is also used in martial arts to help develop physical and mental powers for self defense and healing.

Medical Qigong is divided into two parts: internal and external. Internal Qi is developed by individual practice of Qigong exercises. When Qigong practitioners become sufficiently skilled, they can use external Qi to "emit" Qi (Wai Qi in Chinese) for the purpose of healing another person. There are many reports of the medical efficacy of emitted Qi. This therapy has limited application on a large scale because there are a limited number of skilled Qigong masters. The present article will focus mainly on internal Qi because almost everyone can learn Qigong exercises for maintaining health and for self-healing.

In the early 1980s, scientists in China began investigating the many medical benefits claimed for Qigong. Since then, research on hundreds of medical applications of Qigong has been reported in the literature, which, unfortunately for the West, is written primarily in Chinese. A wealth of material is published in English in the proceedings of international conferences of Qigong. The proceedings contain abstracts of talks given at the conferences, and since 1986 there are 837 abstracts, more than half in English.

EXAMPLES OF MEDICAL APPLICATIONS OF QIGONG

Some examples of the applications of medical Qigong and emitted Qi on humans, animals, cell cultures, and plants were reviewed. Clinical and experimental evidence showing that Qigong exercise and external Qi affects various functions and organs of the body have been discussed. A short list of some of the functions and organs affected by Qigong, and the measurement techniques employed, include: the brain (EEG and magnetometer); blood flow (thermography, sphygmography, and rheoencephalography); heart functions (blood pressure, EKG, and UCG); kidney (urinary albumin assay); biophysical (enzyme activity, immune function, sex hormone levels); eyesight; and tumor size in mice.

In the following discussion, some clinical and experimental research studies will be discussed to illustrate the scope of research on medical applications of Qigong. The studies have been selected because they contain scientific information on important chronic medical conditions, such as hypertension and aging. Most of these studies on Qigong used controls and statistical analysis. This review focuses on the main accomplishments of selected research studies. Details may be omitted because of space limitations or because relevant information is not available. The critical evaluation of research studies will be left to medical specialists.

THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY

Electroacupuncture According to Voll (EAV) can monitor the effects of Qigong practice on therapeutic balancing of the Channel system, the classical TCM organs, and body functions. In EAV the electrical conductance of the skin above individual acupuncture points is measured using a low voltage and low current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an "indicator drop" that may occur during the measurement when the conductance reaches an apparent maximum value but then decreases before leveling off.

Measurements were made at 24 acupuncture points at the ends of the Channels of the fingers and toes of subjects and were made by the same operator and equipment. The subjects were asked to perform a Qigong exercise of their choosing, for example, sitting or standing meditation or moving Qigong. Two series of EAV measurements were made before and after healthy subjects practiced Qigong. In the first series, four subjects were examined by EAV before and after they practiced Qigong for 10 to 15 minutes. Qigong exercise decreased the average of the EAV measured values in the range of -19 to -31% ($p<0.004$) for the four subjects. Qigong essentially eliminated the indicator drops. In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced Qigong before the second or third examination. In the second series, Qigong exercise changed the average EAV measured values in the range of -17% to -35% for four subjects and in the range of 4% to 15% for three subjects. Indicator drops again were decreased.

These preliminary results show that Qigong can make significant changes in the therapeutic balancing of the Channel and organ systems, the goal of TCM.

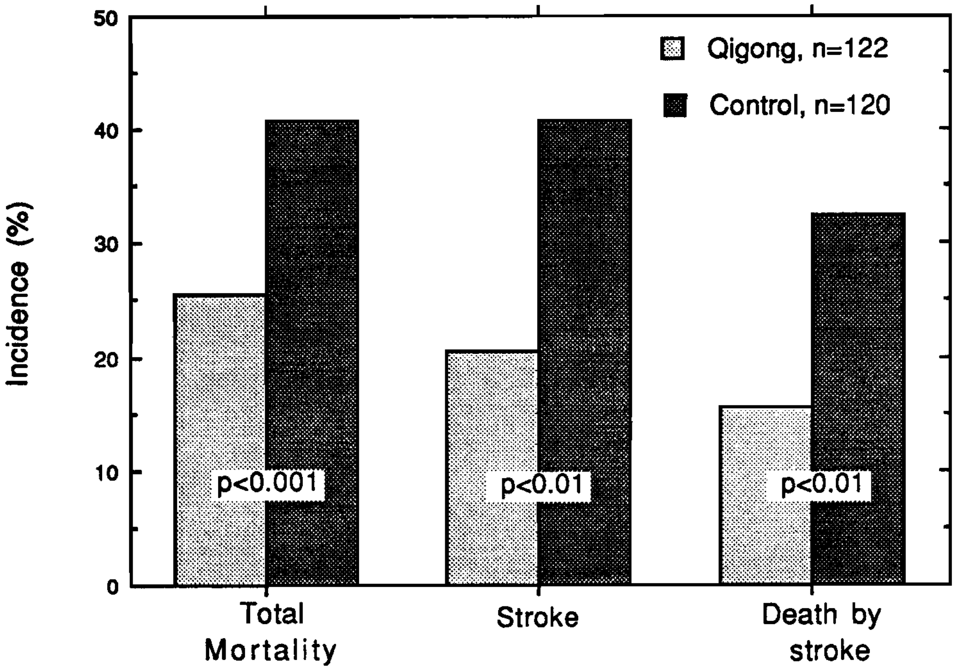

Figure 1. A 30-year folow-up of the effect of Qigong on mortality and stroke of 242 hypertensive patients. Both groups received drug therapy. (Qang. Xu, et al. ref. 7)

HYPERTENSION

Several groups in China have investigated the effects of Qigong on hypertension (i.e., high blood pressure). Research on the short- and long-term effects of Qigong practice on hypertensive patients has been carried out at the Shanghai Institute of Hypertension by Wang Chongxing, Xu Dinghai and their colleagues. Their research is reviewed extensively because it serves as a model of the effects of Qigong on many functions of the body. For these studies, the patients practiced "Yan Jing Yi Shen Gong" for 30 minutes twice a day. This Qigong is claimed to be especially valuable for therapeutic purposes and delaying senility. The Qigong exercise consists of a combination of sitting meditation and gentle physical movements that emphasizes a calm mind, relaxed body, and regular respiration.

PREVENTION OF STROKE AND MORTALITY

In 1991, the researchers reported a 20-year controlled study of the anti-aging effects of Qigong on 204 hypertensive patients. Recently, the researchers performed a 30-year follow-up on 242 hypertensive patients who were divided randomly into a Qigong group (n = 122) and a control group (n = 120). All patients were given drug therapy to control blood pressure, but only the experimental group practiced Qigong 30 minutes a day, twice a day. The results show that the accumulated mortality was 25.41% in the Qigong group and 40.79% in the control group ($p<0.001$). The incidence of stroke was 20.49% and 40.73% ($p<0.01$), and death rate due to stroke was 15.57% and 32.50% ($p<0.01$), respectively. These results, which are shown graphically in Figure 1, indicate that Qigong has potential for preventing strokes and extending life.

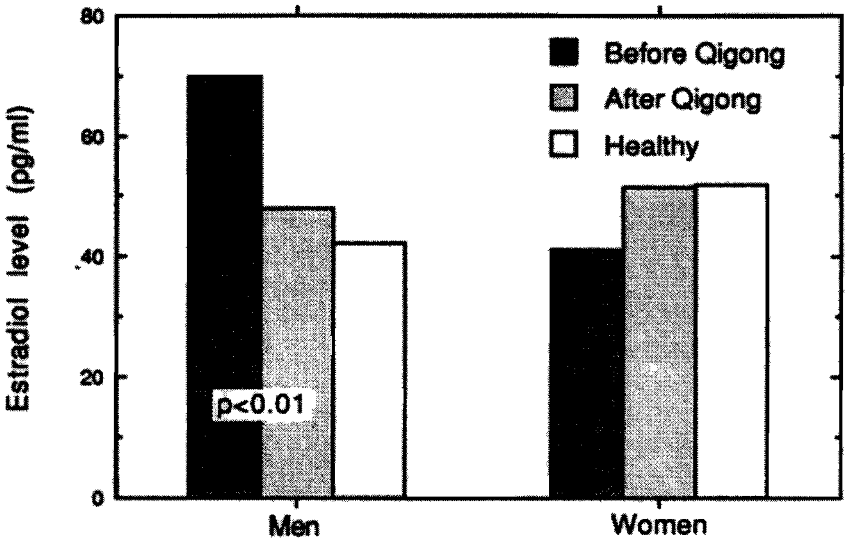

Figure 2. Serum estrdiol levels of hypertensive men (ages 40–69) and women (ages 51–57) who practice Qigong one year compared to healthy controls. (Ankun, Eang, et al. ref. 6)

The researchers also reported that over the 20-year period, blood pressure of the Qigong group stabilized while that of the control group increased. Remarkably, during this period the drug dosage for the Qigong group could be decreased and for 30% of the patients could be eliminated. However, the drug dosage for the control group had to be increased. These results suggest that Qigong activated the natural healing process of the body.

IMPROVEMENTS IN HEART FUNCTION AND MICROCIRCULATION

Aged hypertensive patients usually are found to have a deficiency of Heart-energy, which often leads to a weakened function of the left ventricle and a disturbance of microcirculation. The researchers evaluate the effects of Qigong by using ultrasonic cardiography (UCG) and indices of microcirculation. The 120 male subjects, 55 to 75 years of age, were divided into three groups: 46 cases of hypertensive subjects with Heart-energy deficiency, 34 cases without Heart-energy deficiency, and 40 cases with normal blood pressure. Patients whose blood pressure measured more than 160/95 mm mercury were accepted as subjects after regulation with anti-hypertensive drugs for 4 weeks.

The results showed that subjects with Heart-energy deficiency experienced several improvements: increases in cardiac output, ejection fraction mitral valve diastolic closing velocity, and mean velocity of circumferential fiber shortening, while the total peripheral resistance decreased ($p<0.05$-0.01). Significant changes did not occur in the group without Heart-energy deficiency.

Multiple quantitative evaluation of nailfold disturbance in microcirculation was made on the above three groups by observing 10 indices of abnormal conditions: configuration of micrangium, micrangium tension, condition of blood flow, slowdown of blood flow, thinner afferent limb, efferent limb and afferent limb ratio, color of blood, hemorrhage, and petechia. At the beginning of the study, the incidence of microcirculation obstruction for the above three groups was 73.91%, 26.47% and 17.50%, respectively. After practicing qigong for one year, the group with

Heart-energy deficiency showed a decrease in nailfold microcirculation obstruction from 73.91% to 39.31% ($p<0.05$). Significant changes did not occur in the group without Heart-energy deficiency.

These studies show that proper TCM diagnosis is essential for qigong therapy. The investigators emphasized that the kind of qigong must be selected according to the patient's condition.

IMPROVEMENT IN SEX HORMONE LEVELS

One consequence of aging is that the levels of sex hormones change in unfavorable directions. For example, female hormone (estrogen) levels tend to increase in men and decrease in women. Two studies indicate that this trend can be reversed by qigong exercise. The effect of qigong exercise on plasma sex hormone levels was determine for hypertensive men and women. The sex hormones levels were measured before and after qigong practice for one year.

The 70 male patients with essential hypertension (ages 40 to 69; disease stage II) were divided into two groups. For the qigong group (n = 42), the estradiol level (E2) decreased from 70.12 to 47.71 pg/ml, a decrease of 32% ($p<0.01$), while no significant changes occurred in the control group (n = 20). The testosterone levels (T) of both groups decreased about 7%. The value of E2 for the qigong group (47.71 pg/ml) approached that of healthy men (42.23±5.82 pg/ml) of the same age but without hypertension or cardiovascular, pulmonary, hepatic, renal, or endocrine diseases ($p<0.05$). For women (ages 51 to 67, the number in group not available), the aging process was associated with failure of ovarian function manifested by decreased E2 and increased T levels. Qigong resulted in an increase of E2 from 40.9±.3.5 to 51.6±3.5 pg/ml, a value about equal to that of normal menopausal controls without hypertension or cardiovascular, pulmonary, hepatic, renal, or endocrine diseases. The value of T was also increased by qigong from 25.5±2.2 to 37.2±2.2 ng/dl. These favorable changes in estradiol levels, E2, brought about by qigong are summarized in Figure 2 for both men and women.

In an auxiliary study, the 24-hour urinary estradiol levels were determined in 30 men ages 50 to 69. Qigong for one year resulted in a decrease of 31% in E2 and a decrease of 54% in the estradiol/testosterone ratio (E2/T). These changes were accompanied by improvements in symptoms associated with Kidney deficiency hypertension, such as soreness, dizziness, insomnia, hair loss, impotence, and incontinence. The average score for these symptoms was changed by qigong from 5.53±2.31 to 2.83±1.32 ($p<0.001$).

Ye Ming and co-workers reported similar favorable changes in plasma sex hormone levels E2 in 77 male and female qigong exercisers after 2 months qigong compared with 27 controls. They did not observe significant changes in testosterone.

The above studies show that qigong exercise can help restore the sex hormone levels that had deteriorated because of aging.

INCREASED BONE DENSITY

Aging may result in a decrease in bone density, especially in women. As a consequence, bones become more brittle and subject to fracture.

Bone density of male subjects was found to increase by practicing qigong for one year. For subjects 50 to 59 years of age (18 cases), the bone density increased from 0.627±0.040 to 0.696±0.069 g/cm3 ($p<0.01$). For subjects 60 to 69 years of age (12 cases), the bone density increase was somewhat less, namely, from 0.621±0.039 to 0.672±0.083 g/cm3 ($p<0.05$). For both age groups, the bone density increased to values exceeding those of normal men of the same age, 0.695±0.096 and 0.657±0.102 g/cm3, respectively. The results are shown graphically in Figure 3.

It is likely that qigong therapy also would help restore the bone density of women, especially menopausal women. If so, hormone replacement therapy and its side effects could be reduced.

CHANGES IN BLOOD CHEMISTRY IN HYPERTENSIVE PATIENTS

Auxiliary studies by Wang, Xu and co-workers on the effects of qigong exercise on blood chemistry of hypertensive subjects include improvements in plasma coagulation firbrinolysis indices, blood viscosity, erythrocyte deformation index, levels of plasma tissue-type plasminogen activator (tPA), plasminogen activator inhibitor (PAI), VIII factor related antigen (VIII R:AG), and anti-thrombin (AT-III). In another study, they reported that qigong exercise significantly and beneficially changed the activities of two messenger cyclic nucleotides (cAMP and cGMP).

These results provide further evidence that qigong affects many complex functions of the body.

CANCER

Feng Lida pioneered in research showing that emitted qi from qigong masters produced marked changes in cell cultures of cancer cells from mice. Several studies reported the effects of emitted qi on tumors in animals. For example, emitted qi was reported to inhibit the growth of implanted malignant tumors in mice but did not destroy the tumors. Encouraged by the results with animals, researchers carried out clinical research on the effects of qigong on human subjects with cancer. Detailed results are not available in English for all these clinical studies.

Some results are available for a clinical study of qigong as a therapeutic aid for patients with advanced cancer. In this study, 127 patients with medically diagnosed malignant cancer were divided into a qigong group of 97 patients and a control group of 30 patients. All patients received drugs, and the qigong group practiced qigong for more than 2 hours a day over a period from 3 to 6 months. Both groups improved, but the qigong group showed improvements four to nine times greater than the control group for strength, appetite, diarrhea free, and weight gain of 3 kg. The phagocytic rate, which a measure of the immune

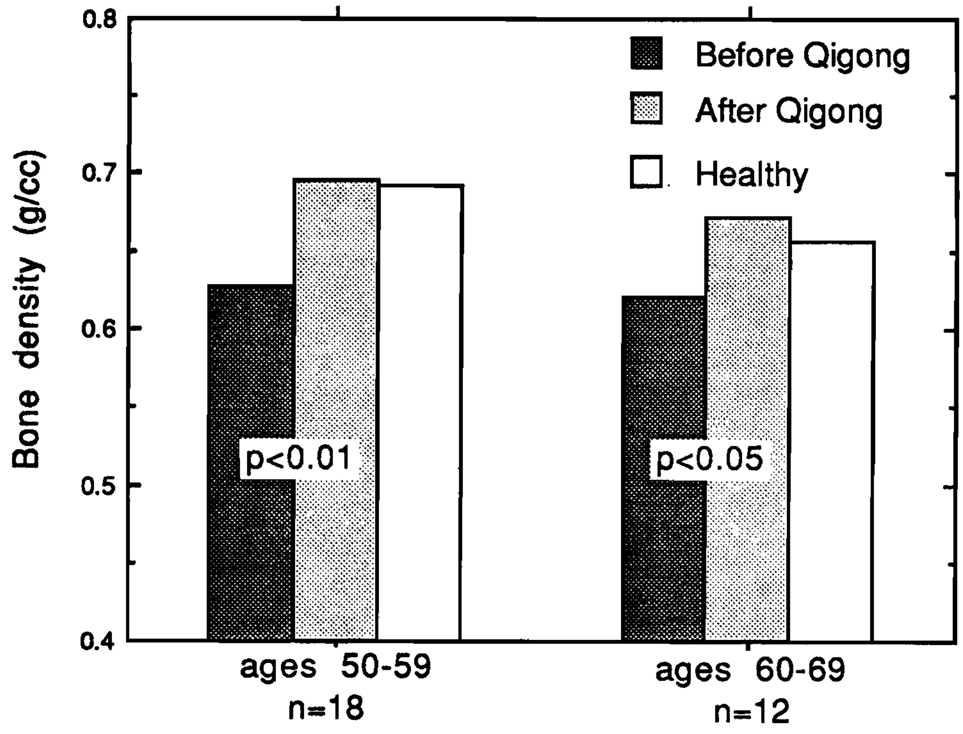

Figure 3. Changes in bone density of hypertensive men who practice Qigong one year compared to healthy males of the same age. (Xu Wang, et al. ref. 9)

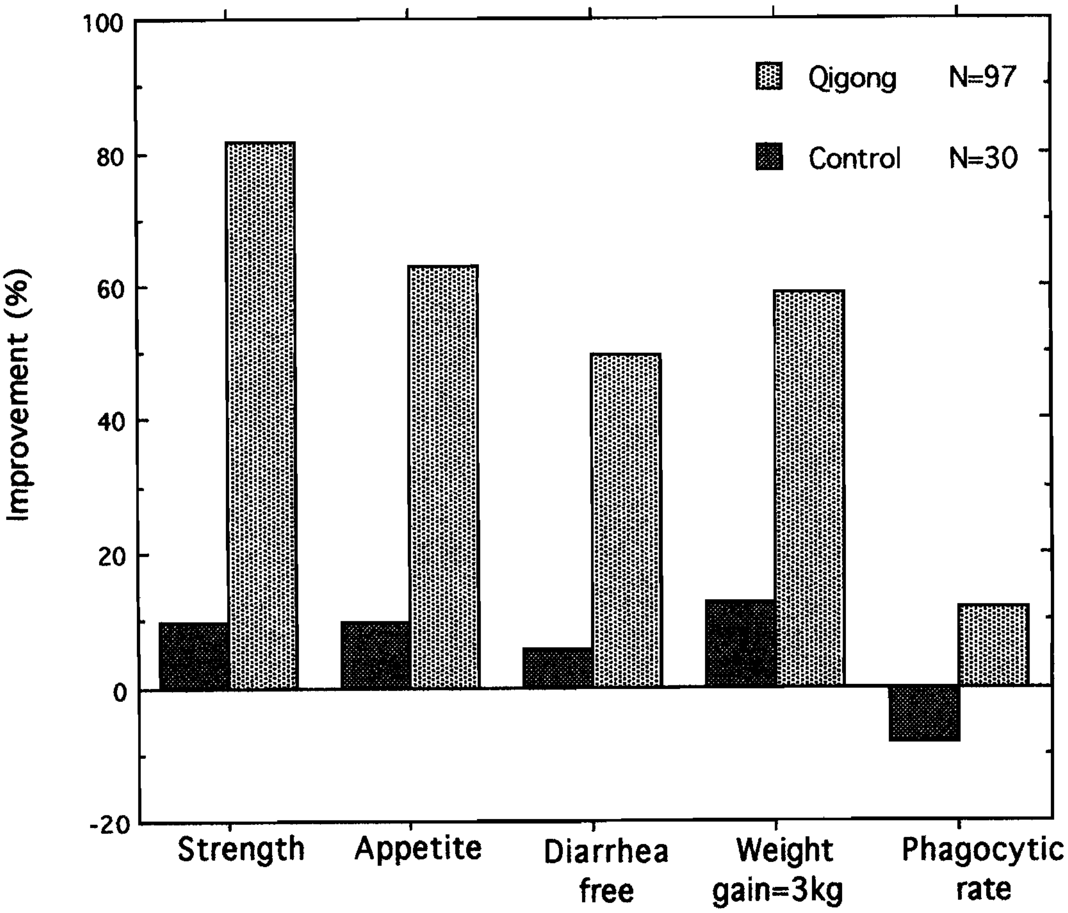

Figure 4. Comparison of Qigong, drug and drug-alone terapies for patients with advanced cancer. (ref. 25)

function, increased in the qigong group but decreased in the control group. These results are shown in Figure 4.

REVERSING SENILITY

To study the mechanism of keeping fit by qigong, a controlled study was made of 100 subjects classified either as presenile or with senile impaired cerebral function. The subjects were divided into two groups of 50 people each with a mean age of 62.7 years and with a similar distribution of age and sex. The qigong group practiced a combination of static and moving qigong. The control group exercised by walking, walking fast, or running slow. According to TCM method of classifying the vital energy, more than 80% of the patients in each group were classified as deficient in vital function and vital essence of the Kidney. Criteria for judging outcome were based on measuring clinical signs and symptoms including cerebral function, sexual function, serum lipid levels, and function of endocrine glands.

After six months, 8 of the 14 main clinical signs and symptoms in the qigong group were improved above 80%, whereas none of the symptoms in the control group was improved above 45%. These results suggest that qigong can reverse some symptoms of aging and senility.

MIND-BODY REGULATION

The main function of qigong is to regulate the mind, and according to TCM theory the mind leads the qi. The mind-regulating function of qigong mainly refers to the regulation of the functions of the brain and related body reactions. One tenant of qigong is that "the mind leads the qi, and the qi leads the blood." This somewhat mysterious statement can be interpreted to mean that intention (the mind) can direct the qi within the body. This mechanism is perhaps similar to the role of volition in biofeedback self-regulation.

BRAIN WAVES

The main focus of research has been on the effects of qigong on brain waves as measured by electroencephalography (EEG). During static, e.g., sitting meditation, alpha brain waves dominate over beta waves and spread to the frontal areas of the brain. Kawano and Wang have found differences in the EEG of Zen Buddhist priests and qigong masters. During almost all types of qigong training, the frequency of the alpha waves increased in the range from +0.6 to 1.0 Hz. During deep Zen meditation, the frequency decreased in the range from -1.0 to -1.5 Hz, and sometimes theta waves appeared. Also, frontal and occipital alpha waves tended to synchronize with a phase difference that depended on the type of meditation. This phase difference became smaller with qigong meditation (i.e., better synchronization) and larger for Zen meditation. According to Kawano and Wang these differences in brain function suggest that internal qigong involves a "semi-conscious" process, which involves some awareness and activity, while Zen meditation is a neutral process that releases the meditator from all concerns. This difference may relate to qigong being a healing art, whereas Zen is generally not considered to be one.

As mentioned earlier, a qigong master can emit qi to heal a patient, and the literature contains many references to this subject. The interaction between the qigong master and his subject has been followed in double blind tests. Both qigong master and subject were simultaneously measured by EEG, polygraphs, biochemical blood tests, and psychological tests. The EEG studies showed that brain wave patterns of the subject and qigong master were synchronized in type of brain waves and their location in the brain. Such synchronism may be required for healing by emitted qi.

Machi has studied qigong masters by combining EEG measurements with simultaneous measurements of physiological changes. He found that while the qigong master was emitting qi, the alpha-1 waves showed extremely high potential on the right frontal lobe, and there were increases in blood pressure, heart rate, and skin surface temperature. He also detected a far infrared emission with 1 Hz modulation signal coming from the Laogong point (a major acupuncture point in the palm of the hand.)

BLOOD FLOW TO THE BRAIN

Qigong exercise has been shown by rheoencephalography to increase blood flow to the brain. For 158 subjects with cerebral arteriosclerosis who practiced qigong for 1 to 6 months, improvements were noted in symptoms such as memory, dizziness, insomnia, tinnitus, numbness of limbs, and vertigo headache. During these studies, a decrease in plasma cholesterol was also noted. These results may offer hope to people with cerebral arteriosclerosis.

EFFECT OF RAPID AND LARGE CHANGES IN ALTITUDE

Studies have been made to determine whether qigong exercise would protect pilots from altitude stress when they increased altitude by going rapidly from low altitudes to the Tibetan highlands.

CARDIAC FUNCTION

Before entering the Tibetan highland, 66 persons were divided into two groups: a qigong group of 32 young men who did Qiyuan Qigong exercise for 4 weeks, and a control group of 34 young men who exercised to radio music. The two groups of men suddenly entered the highlands from a lower altitude. Before and after entering the highland, measurements were made of symptoms of altitude sickness and physiological changes. The qigong group suffered less altitude stress than the control group as measured by blood pressure, heart rate, oxygen consumption, microcirculation on epex of tongue and nail fold, and temperature at Laogong point (P8) of the left hand ($p<0.01$). The researchers suggest that qigong can prevent stresses from altitude changes.

These results with healthy subjects suggest that qigong also should be effective in improving the health of people with cardiovascular conditions. In fact several research studies have been reported the beneficial effects of qigong on cardiovascular diseases.

MICROCIRCULATION DISORDERS

Air force pilots were randomly divided into two groups: a qigong group of 22 persons who had practiced Qiyuan Qigong exercise for eight weeks, and a control group of 18 persons who did physical exercise for eight weeks before entering the Tibetan highlands. Microcirculation was measured at tongue apex and the nail fold, and also from the temperature at the Laogong point in palm of the left hand. When the men entered the high altitude, abnormal blood pressure and microcirculation of tongue apex and nail fold occurred in both groups. However, the abnormalities were statistically less in the qigong group than in the control group ($p<0.01$). The temperature at Laogong kept steady in the qigong group, but was reduced in the control group ($p<0.05$).

STUDY OF LUNG FUNCTION

The men were divided into three groups. Forty young males were in the qigong group and practiced Qi Yuan qigong for 4 weeks prior to entering the highlands; 40 young men were in the control group and exercised to radio music for 4 weeks prior to entering the highlands; and 40 young males were residents living at high altitudes. The results show that the integral value of symptoms of acute mountain sickness was lower in the qigong than in control group ($p<0.05$—0.01). Pulmonary ventilation of the qigong group was significantly improved compared with the control group ($p<0.05$—0.01), and nearly equal to the resident group.

The favorable outcome of these studies suggests that qigong may be a valuable therapy for people suffering from cardiac, microcirculation and lung disorders.

SUMMARY AND CONCLUSIONS

This review deals with only a small fraction of a large collection of research already performed on medical applications of qigong, mainly in China. Hopefully, the information presented is sufficient to illustrate the potential of qigong exercise for restoring normal body functions. The main conclusion from many studies is that qigong enables the body to heal itself.

Medical practitioners and scientists are encouraged to evaluate the potential value of medical applications of qigong. They should collect as much information as possible from original sources, develop bibliographies, and then determine whether additional research is required to validate given medical applications. They should then recommend those applications most likely to be suitable for western health care. It is recommended that TCM and qigong be considered to complement western medicine.

There are several issues that should be addressed for the successful introduction of qigong into western health care. It is necessary to win the support of scientists, medical practitioners and the public who know very little about TCM and qigong. A crucial step is to collect scientific evidence from research already performed. Regrettably, there are few research reports in English, and original scientific reports in Chinese are often difficult to obtain and translate. Further, some of the best technical research papers are incomplete scientifically. Recently this has become less of a problem in China. Although some scientific reports have flaws, we must not reject all research on qigong. We must give credit to China for developing the science of qigong and recognize that it is a science in development.

On the positive side, qigong is emerging rapidly among a public interested in improving their health. Qigong is a subject being addressed by the public press and in books and magazines. Increasing numbers of medical practitioners are becoming active in alternative medicine, although few physicians have studied TCM. At least one insurance company has paid for qigong therapy. HMOs and state and federally funded medical programs would benefit from paying for qigong therapy because it is effective and inexpensive. Pharmaceutical companies are already exploring many herbs, and these companies also could benefit from qigong research because of the synergic effect between qigong and drugs mentioned in this review. The Office of Alternative Medicine (OAM) of the National Institutes of Health is working to develop a basis for evaluating the merits of different alternative therapies, and in 1994 it funded one grant involving qigong. Incidentally, the Japanese Ministry of International Trade and Industry (MITI) is reported to be taking a pragmatic approach to alternative therapies by encouraging the development of industrial applications of therapies such as acupuncture and qigong.

There are many medical applications of qigong. Some of the most promising for introducing into western health care are chronic problems such as hypertension, cardiovascular disease, aging, asthma, allergies, neuromuscular problems, and cancer. These areas of public health deserve consideration by the western medical establishments.

*The p-value is the probability that two quantities are not different: the smaller the p-value greater the difference.

REFERENCES

1. Qigong Database™ is available from the Qigong Institute,450 Sutter Street, Suite 2104, San Francisco, CA 94108.
2. Sancier K M, Hu B. Medical Applications of Qigong and Emitted Qi on Humans, Animals, Cell Cultures, and Plants: Review of Selected Scientific Studies. Am J Acupuncture.1991:19 (4) 367-377.
3. Sancier K M. The effect of qigong on human body functions. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:179.
4. Sancier K M. The effect of qigong on therapeutic balancing measured by electroacupuncture according to Voll (EAV): a preliminary study. Acupuncture & Electro-Thera Res Int J. in press.
5. Wang Chongxing, Xu Dinghai, Qian Yuesheng, Medical and health care qigong, J Traditional Chinese Medicine. 1991:11(4) 296-301.
6. Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yuesheng. Research on the anti-aging effect of qigong. J. Traditional Chinese Medicine. 1991:11 (2) 153-158.
7. Hong Shunhua, et al. Microcirculation of nail fold and immunogenicity after qigong practice for short periods. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 56-57.
8. Li Ziran, Li Liziang, Zhang Boli. Group observation and experimental research on the prevention and treatment of hypertension by qigong. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 113-114.
9. Jing Guinian. Observations on the curative effects of qigong self adjustment therapy in hypertension Proceedings, Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 115-117.
10. Bian Huangxian. Clinical observation of 204 patients with hypertension treated with qigong. Proceedings, First International Congress of Qigong. Berkeley, Calif., 1990: 101
11. Wang Chongxing, et al. The beneficial effect of qigong on the hypertension incorporated with coronary heart disease. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990: 40.
12. Wu Renzhao, Liu Zhewei. Study of qigong on hypertension and reduction of hypotensor. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.
13. Bornoroni Corrado, et al. Treatment of 30 cases of primary hypertension by qigong techniques. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 126.
14. Zhang Guifang. Development and application of a series of qigong feedback tapes Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.
15. Wang Chongxing, et al. Effects of qigong on preventing stroke and alleviating the multiple cerebro-cardiovascular risk factors—a follow-up report on 242 hypertensive cases over 30 years. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 123-124.
16. Wang Chongxing, Xu Dinghai, et al. Beneficial effects of qigong on the ventricular function and microcirculation of deficiency in heart-energy hypertensive patients. Pri-

vate communication.

17. Xu Dinghai, Wang Chongxing, et al. Clinical study of delaying effect on senility by practicing "Yang Jing Yi Shen Gong" in hypertensive patients. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, 1994:109.
18. Ye Ming, et al. Relationship among erythrocyte superoxide dismustase activity, plasma sexual hormones (T, E2), aging and qigong exercise. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990:28-32 (in English and Chinese).
19. Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yueshang. Research on "anti-aging" effect of qigong. J Traditional Chinese Medicine. 1991:11 (3) 224-227.
20. Feng Lida, Effect of emitted qi on human carcinoma cells. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 1-4.
21. Feng Lida, Effect of emitted qi on the L 1210 cells of leukemia in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 4-5.
22. Zhao Shan, et al. Preliminary observation of the inhibitory effect of emitted qi on transplanted tumors in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 46-48.
23. Liu Tehfu,Wan Minsheng,Lu Oulun. Experiment of the emitted qi on animals. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 60-61.
24. Feng Lida,Peng Liaomin. Effect of emitted qi on prevention and treatment of tumors in mice. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 106-107.
25. Sun Quizhi, Zhao Li. Clinical observation of qigong as a therapeutic aid for advanced cancer patients. Proceeding, First World Conf Academic Exchange of Medical Qigong, Beijing, China. 1988: 97-98.
26. Chen Guoguang. The curative effect observed for 24 (cancer) cases under my emitted qigong treatment. Proceedings, Second International Conference on Qigong, Xi'an, China. 1989: 141-142.
27. Lo Jifeng, et al. Changes of peripheral blood cell population and immune functions in 31 nasopharyngeal carcinoma (NPC) patients treated with radiotherapy and qigong. Proceedings ,Third National Academic Conference on Qigong Science, Guangzhou, China. 1990: 94-95.
28. Yu Yi, et al. Effect of self-controlling qigong therapy on the immune function of cancer patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 128.
29. Zhao Hongmei, Bian Jingnan. Curative effect of intelligence qigong on 122 tumor patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993:130.
30. Wang Ying. Clinical observation on 30 cases of cancer treated by qigong therapy Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131.
31. Xong Jing, Lu Zhong. Curative effect on 120 cancer cases treated by Chinese-Western medicine and qigong therapy. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131
32. Fu Jingzhi. Treatment of advanced gastric cancer in the aged by the combination of qigong and medicinal herbs. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 132-133.
33. Zhang Shengbing. Effects of mind-regulation by qigong on the human body. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994:68.
34. Kawano Kimiko, Wang Fengfong. Difference between the EEG of Zen priests and internal qigong masters. Society of Mind-Body Science. 1994: 3 (1) 99-104. (in Japanese).
35. Machi Yoshio. Various measurements to qigong masters for analyzing qigong mechanism. Society for Mind-Body Science. 1994:3 (1) 65-87. (in Japanese).
36. Kashiwasake Masaki. Double-blind tests of qi transmission from qigong masters to untrained volunteers. Society for Mind-Body Science. 1993:2 (1) 81-87 (in Japanese).
37. Machi Yoshio. Various measurements of qigong masters for analyzing qigong mechanism Society of Mind-Body Science. 1994:3 (1) 65-87. (In Japanese).
38. Liu Yuanliang, He Shihai, Xie Shanling. Clinical observation on the treatment of 158 cases of cerebral arteriosclerosis by qigong. Proceedings, Second World Conference on Academic Exchange of Medical Qigong, Beijing, China. 1993:125.
39. Liu Yuanliang. Private communication, 1993.
40. Mo Feifan, Xu Yongchun, Lu Yongpin, Xu Guang. Study of prevention of cardiac function disorder due to immediate entry into highlands by qigong exercise. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 78.
41. Chu Weizong, et al. Changes of blood viscosity and RCG in 44 cases with cardiovascular diseases after qigong ex-

ercises. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 57-58.

42. Qin Chao, et al. Bidirectional adjustment of blood pressure and heart rate by daoyin tuina on the arterial blood and heart rate. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 107.
43. Wang Chongxing, et al. Beneficial effect of qigong on improving the heart function and relieving multiple cardiovascular risk factors. Proceedings ,Third International Symposium on Qigong. Shanghai, China. 1990: 42 (in Chinese).
44. Mo Feifan, Wan Lurong, Jia ZiZi, Xu Guang. Study of prevention of microcirculation disorders of pilots in highlands by qigong. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 78.
45. Mo Feifan, Lu Yongpin, Zhao Guoliang. Effect of exercise with qigong on lung function of persons entering highland. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:186.
46. Wang Chongxing, Xu Dinghai, et al., private communication.
47. Wang Chongxing, Xu Dinghai, et al. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994::22 (Chinese),109.
48. Wen-hsien Wu, Intractable reflex sympathetic dystrophy
49. McGee C T, Chow E P Y. Miracle Healing from China — Qigong. 1st ed. Coeur d'Alene, ID: Medipress; 1994: 8-19.

APPENDIX 1B
ANTI-AGING BENEFITS OF QIGONG

Kenneth M. Sancier, Ph.D.

INTRODUCTION

In the early 1980s, scientists in China began to study the medical benefits claimed for qigong. Since then, research on hundreds of medical applications of qigong have been reported in the Chinese literature. Of special interest for the present article are clinical reports of the medical benefits of qigong that claim to retard or reverse some diseases associated with aging.

Most of the original research was reported in Chinese, but access in English to most of this material is possible by reference to the proceedings of international conferences of qigong. Since 1986, ten such proceedings contain about 840 abstracts of talks given at the conferences, more than half of which are in English. These abstracts, along with about 160 abstracts of articles in the scientific literature, have been organized as a computerized database. The database enables searches and development of bibliographies across this entire body of information by using any key word. The clinical outcomes reported in this article are partly based on material in the database and partly on the author's person contacts with researchers.

The word qigong is a combination of two ideas: qi the vital energy of the body, and gong the skill of working of the qi. Medical qigong for health and healing consists primarily of meditation, physical movements, and breathing exercises. Qigong practitioners develop an awareness of qi sensations in their bodies and use their mind, i.e., intention, to guide the qi in the body. The benefits of qigong are said to extend beyond health and healing to enhance spiritual life and even special abilities, such as psychic powers.

Medical qigong is divided into two parts: internal and external. Internal qi is developed by individual practice of qigong exercises. When qigong practitioners have sufficiently mastered the skill, they can "emit" qi (external qi or waiqi in Chinese) for the purpose of healing another person. There are many scientific reports of the medical existence and efficacy of emitted qi. The present article focuses mainly on internal qi because almost everyone can learn qigong exercises for maintaining health and for self-healing, whereas, there are a limited number of skilled qigong masters available for healing.

There are numerous reports of the effects of emitted qi on living systems and the functions and organs of the human body. The present author reviewed some examples of medical applications of qigong and emitted qi on humans, animals, cell cultures, and plants, and he also published some of his experimental research on physiological effects of qigong. He discussed some clinical and experimental evidence showing that qigong exercise and external qi affects various functions and organs of the body. A short list of some of the functions and organs affected by qigong, and the measurement techniques employed (in parentheses), include: the brain (EEG and magnetometer); blood flow (thermography, sphygmography, and rheoencephalography); heart functions (blood pressure, EKG, and UCG); kidney (urinary albumin assay); biophysical (enzyme activity, immune function, sex hormone levels (laboratory analysis); eyesight (clinical); and tumor size in mice.

CLINICAL STUDIES INDICATING THE ANTI-AGING BENEFITS OF QIGONG

Several clinical studies will be described to illustrate the scope of research on medical applications of qigong to treat chronic medical conditions that may affect the aging process. Some details may be omitted because of space limitations. The critical evaluation of the research studies will be left to medical specialists.

THERAPEUTIC BALANCING OF THE CHANNELS AND FUNCTIONS OF THE BODY BY QIGONG

The profound effect that internal qigong practice may have on balancing the energies of the organs and functions of the body is illustrated by measurements using Electroacupuncture According to Voll (EAV). In EAV the electrical conductance of the skin above individual acupuncture points is measured using low voltage and low current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an "indicator drop" that may occur during the measurement when the conductance reaches an apparent maximum value but then decreases before leveling off.

Measurements were made at 24 acupuncture points at the ends of the 12 Channels in the fingers and toes of subjects and were made by the same operator and equipment. The subjects were asked to perform a qigong exercise of their choosing for 10 to 15 minutes, for example, sitting or standing meditation or moving qigong. Two series of EAV measurements were made before and after healthy subjects practiced qigong.

In the first series, four subjects were examined by EAV before and after they practiced qigong. Qigong exercise decreased the average of the EAV measured values of the four subjects in the range of -19 to -31% ($p<0.004$). Qigong eliminated almost all the indicator drops.

In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced qigong before the second or third examination. The results showed that qigong exercise changed the average EAV measured values in the range of -17 to -35% for four subjects and in the range of 4 to 15% for three subjects. Indicator drops again were markedly decreased.

These preliminary results show that internal qigong practice can make significant changes in the therapeutic balancing of the Channel and organ systems.

In a similar type of study, the electric current at acupuncture points on 14 Channels was measured using a single square wave voltage pulse technique. Both a qigong master, who emitted qi, and a qi-receiver were measured simultaneously and continuously. The results show that internal and external qigong produce different values in some measurement parameters, and also some synchronous behavior was observed between the sender and receiver.

BLOOD PRESSURE

The effect of qigong exercise on blood pressure is shown graphically in Fig. 1. The blood pressure (systolic and diastolic) in millimeters mercury is plotted as a function of time over 20 years for the group consisting of 242 patients, 122 in the qigong group and 120 in the control group. During the first two months, the blood pressure of all patients dropped in response to the hypotensive drug. Subsequently, and over the period of 20 years, the blood pressures of the qigong group stabilized while that of the control group increased. Remarkably, during this period the drug dosage for the qigong group could be decreased, while the dosage for the control group had to be increased.

Fig. 1. Effect of qigong on blood pressure of hypertensive patients over 20 years. Qigong group (n = 104) practiced 30 min/day twice/day, control group (n = 100).

MORTALITY AND STROKE

The incidences of mortality and stroke for the 30-year study are shown in Fig. 2. These results show that qigong exercise decreased by about 50 percent the incidence of total mortality, mortality due to stroke, and morbidity due to stroke. At the end of 30 years, 86 patients survived in the qigong group and 68 in the control group. These results clearly show that qigong has significant potential for preventing strokes and extending life.

Fig. 2. Effect of qigong on mortality and stroke of hypertensive patients over 30 years. Both groups received drug therapy; qigong group (n = 122) practiced qigong 30 min/day twice/day; the control group (n = 120).

IMPROVEMENTS IN HEART FUNCTION AND MIRCROCIRCULATION

Aged hypertensive patients usually are found to have a deficiency of Heart-energy, which often leads to a weakened function of the left ventricle and a disturbance of microcirculation. The researchers evaluated the effects of qigong for120 aged patients by using ultrasonic cardiography (UCG) and indices of microcirculation.

Experiments showed that the left ventricular function (LVF) in the hypertensive aged group (80 cases) was lower than that in the aged normal blood pressure group (40 cases), while the LVF in the deficiency of Heart-energy hypertensive patients (46 cases) was lower than in the non-deficiency Heart-energy hypertensive patients (34 cases).

After practicing qigong for one year, cardiac output was increased, the total peripheral resistance decreased, and the ejection fraction mitral valve diastolic closing velocity and the mean velocity of circumferential fiber shortening tended to be increased. Significant changes did not occur in the group without Heart-energy deficiency.

Quantitative evaluation of nailfold disturbances in microcirculation was made on the above groups by observing 10 indices of abnormal conditions: configuration of micrangium, micrangium tension, condition of blood flow, slowdown of blood flow, thinner afferent limb, efferent limb and afferent limb ratio, color of blood, hemorrhage, and petechia. The results showed that hypertension had an accelerating effect on the disturbance of microcirculation. The incidence of disturbance of microcirculation disturbance was 73.9% in the deficiency of Heart-energy hypertensive patients. After a year of qigong practice, the incident of disturbance was 39.1% ($p<0.01$).

The results suggest that qigong exercise has beneficial effects on Heart-energy and regulation of the blood channel, and qigong seems to have improved abnormal conditions of blood circulation.

ENHANCED ACTIVITY OF ANTI-AGING ENZYME SOD

Superoxide dismutase (SOD) is produced naturally by the body but its activity declines with age. SOD is often called an anti-aging enzyme because it is believed to destroy free radicals that may cause aging. The effects of qigong exercise to treat disorders of retired workers were studied by Xu Hefen and coworkers and included determinations of plasma SOD.

For their study, 200 retired workers,100 males and 100 females, ranging in age from 52 to 76 were divided into 2 groups: the qigong exercise group and the control group, and each group consisted of 50 males and 50 females. The main qigong exercise was Emei Nei Gong (one kind of qigong exercises of the Emei School), and was practiced at least 30 minutes a day.

The result showed that the mean level of SOD was increased by qigong exercise. For example, the SOD level was larger in the qigong group (about 2700 μ / g Hb) and than in the control group (1700 μ / g Hb), and this difference was significant ($p<0.001$). This study shows that qigong exercise can stimulate physical metabolism, promote the circulation of Channels and regulate the flowing of qi and blood, thus preventing and treating disorders of aging and promoting longevity.

CARDIOVASCULAR FUNCTION

Several studies reveal the potential benefits that qigong may have for improving the cardiovascular function of those with heart disease as well as old people. This conclusion is based on three studies reporting that qigong exercise can protect healthy pilots from altitude stress when they flew rapidly from a low altitude to the high altitude of the Tibetan highlands.

Before entering the Tibetan highland, 66 healthy young men were divided into two groups: a qigong group of 32 men who did Qiyuan Qigong exercise for 4 weeks, and a control group of 34 men who exercised to radio music. The two groups of men rapidly entered the highlands from a lower altitude. Before and after entering the highland, measurements were made of symptoms of altitude sickness and physiological changes.

The qigong group suffered less altitude stress than the control group as measured by blood pressure, heart rate, oxygen consumption, microcirculation on the epex of tongue and the nail fold, and the temperature at the Laogong point of the left hand ($p<0.01$). The researchers suggest that qigong can prevent stress from altitude changes.

In another study of changes in altitude, healthy young men were divided into three groups. Forty males were in the qigong group and practiced Qiyuan qigong for 4 weeks prior to entering the highlands; 40 men were in the control group and exercised to radio music for 4 weeks prior to entering the highlands; and 40 males were residents living at high altitudes. The results show that the integral value of symptoms of acute mountain sickness was lower in the qigong than in control group ($p<0.01$ to 0.05). Pulmonary ventilation of the qigong group was significantly improved compared with the control group ($p<0.01$ to 0.05), and nearly equal to the resident group.

In another study, air force pilots were randomly divided into two groups: a qigong group of 22 men who had practiced Qiyuan Qigong exercise for eight weeks, and a control group of 18 men who did physical exercise for eight weeks before entering the Tibetan highlands. Microcirculation was measured at tongue apex and the nail fold, and also from the temperature at the Laogong point in palm of the left hand. When the men entered the high altitude, abnormal blood pressure and microcirculation of tongue apex and nail fold occurred in both groups. The abnormalities were statistically less in the qigong group than in the control group ($p<0.01$). The temperature at Laogong kept steady in the qigong group, but was reduced in the control group ($p<0.05$).

The results of these three studies with healthy subjects lead to the conclusion that qigong also should be effective in improving the health of people with cardiovascular conditions including the aged. In fact several research studies have reported such beneficial effects of qigong on cardiovascular diseases. The three studies also provide evidence that qigong exercise is superior to physical exercise such as calisthenics.

COMBINATION OF MEDICAL QIGONG AND DRUGS THERAPY

There is ample evidence in the literature that therapy by a combination of qigong exercise and drugs is superior to that of drugs alone. The advantages of a combination therapy of qigong and drugs over drugs alone were discussed earlier in this paper for hypertension and cancer.

The mechanism of this apparent synergism is not entirely understood, but undoubtedly relates to the fundamental mechanism of qigong. Qigong is believed to remove blocks to the ready flow of the qi (energy), blood, oxygen and nutrients to all cells of the body as well as to promote removal of waste products from cells of the body. Blocks to energy (qi) flow may result from injury, disease or stress.

Increases in qi flow and blood circulation help nourish diseased or stressed tissue, providing a means for the body to heal itself. This mechanism suggests that qigong also could promote drug uptake to tissue and cells via increased blood circulation. Omura's research shows that drug uptake was increased by using qigongized paper (i.e., paper to which emitted qi was sent) applied to afflicted area of the body.

CONCLUSIONS

This review deals with a small fraction of the large collection of clinical research on medical applications of qigong. The information presented is intended to illustrate the potential of qigong exercise for restoring normal body functions in people with chronic conditions, many of which accelerate the aging process. The main conclusion from many studies is that qigong exercise helps the body to heal itself. In this sense, qigong is a natural anti-aging medicine. Two studies indicate that qigong exercise is superior to some physical exercises.

Qigong can complement Western medicine in many ways to provide better healthcare. For example, qigong has special value for treating chronic conditions and as a preventive medicine, whereas Western medicine has special value for treating acute conditions. There are many medical applications of qigong that can complement Western medicine to improve health care. Some examples include chronic problems such as hypertension, cardiovascular disease, aging, asthma, allergies, neuromuscular problems, and cancer. These areas of public health deserve consideration by the Western medical establishment.

REFERENCES

1. Kenneth M. Sancier, Ph.D., Copresident & Director of Research, 561 Berkeley Avenue, Menlo Park, CA 94025, USA. Phone/Fax +1-415-323-1221.
2. Some of the material in this article was adapted from the article, "Medical Applications of Qigong," by K.M. Sancier, Ph.D., and published in Alternative Therapies in Health and Medicine. 1996;2 (1): 40-46.
3. Qigong Database™ is available from the Qigong Institute, East West Academy of Healing Arts, 450 Sutter Street, Suite 2104, San Francisco, CA 94108, USA.
4. The p-value is the probability that two quantities are not the same: the smaller the p-value smaller the probability and the more significant the result.
5. Wang C, Xu D. Private communication, May 1995.
6. Duration of the study is not available.

ENDNOTES

[1] Sancier K M, Hu B. Medical Applications of Qigong and Emitted Qi on Humans, Animals, Cell Cultures, and Plants: Review of Selected Scientific Studies. Am J Acupuncture.1991:19 (4) 367-377.

[2] Sancier, KM, Medical applications of qigong. Alternative Therapies in Health & Medicine. 1995; 2(1) in press.

[3] Sancier KM, Chow EPY. Healing with qigong and quantitative effects of qigong. J. American College of Trad. Chinese Medicine. 1989: 7(3):13-19.

[4] Sancier KM. The effect of qigong on therapeutic balancing measured by electroacupuncture according to Voll (EAV): a preliminary study. Acupuncture & Electro-Therapy Res Int J.:1995;19:119-127.

[5] Sancier K M. The effect of qigong on human body functions. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:179.

[6] Kido M. Channel measurements of qi-gong operation and synchronous phenomena. J. Mind-Body Science: 1993: 2(1): 19-26.

[7] Wang Chongxing, Xu Dinghai, Qian Yuesheng, Medical and health care qigong, J Traditional Chinese Medicine. 1991:11(4) 296-301.

[8] Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yuesheng. Research on the anti-aging effect of qigong. J. Traditional Chinese Medicine. 1991:11 (2) 153-158.

[9] Hong Shunhua, et al. Microcirculation of nail fold and immunogenicity after qigong practice for short periods. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 56-57.

[10] Li Ziran, Li Liziang, Zhang Boli. Group observation and experimental research on the prevention and treatment of hypertension by qigong. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 113-114.

[11] Jing Guinian. Observations on the curative effects of qigong self adjustment therapy in hypertension Proceedings, Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 115-117.

[12] Bian Huangxian. Clinical observation of 204 patients with hypertension treated with qigong. Proceedings, First International Congress of Qigong. Berkeley, Calif., 1990: 101

[13] Wang Chongxing, et al. The beneficial effect of qigong on the hypertension incorporated with coronary heart disease. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990: 40.

[14] Wu Renzhao, Liu Zhewei. Study of qigong on hypertension and reduction of hypotension. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.

[15] Bornoroni Corrado, et al. Treatment of 30 cases of primary hypertension by qigong techniques. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 126.

[16] Zhang Guifang. Development and application of a series of qigong feedback tapes Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 125.

[17] Wang Chongxing, et al. Effects of qigong on preventing stroke and alleviating the multiple cerebro-cardiovascular risk factors—a follow-up report on 242 hypertensive cases over 30 years. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 123-124.

[18] Wang Chongxing, Xu Dinghai, et al. Beneficial effects of qigong on the ventricular function and microcirculation

of deficiency in heart-energy hypertensive patients. Private communication, January 1994.

[19] Xu Dinghai, Wang Chongxing, et al. Clinical study of delaying effect on senility by practicing "Yang Jing Yi Shen Gong" in hypertensive patients. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, 1994:109.

[20] Ye Ming, et al. Relationship among erythrocyte superoxide dismustase activity, plasma sexual hormones (T, E2), aging and qigong exercise. Proceedings, Third International Symposium on Qigong, Shanghai, China. 1990:28-32 (in English and Chinese).

[21] Kuang Ankun, Wang Chongxing, Xu Dinghai, Qian Yueshang. Research on "anti-aging" effect of qigong. J Traditional Chinese Medicine. 1991:11 (3) 224-227.

[22] Xu, Hefen; Xue, Huining; Bian, Meiguang; Zhang, Chengming; Zhou, Shuying. Clinical study of the anti-aging effect of qigong. Proceedings, Second World Conference for Academic Exchange of Medical Qigong. Beijing, China. 1993: 137.

[23] Mo Feifan, Xu Yongchun, Lu Yongpin, Xu Guang. Study of prevention of cardiac function disorder due to immediate entry into highlands by qigong exercise. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 78.

[24] Mo Feifan, Lu Yongpin, Zhao Guoliang. Effect of exercise with qigong on lung function of persons entering highland. Proceedings, Fifth International Symposium on Qigong, Shanghai, China.1994:186.

[25] Mo Feifan, Wan Lurong, Jia ZiZi, Xu Guang. Study of prevention of microcirculation disorders of pilots in highlands by qigong. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 78.

[26] Chu Weizong, et al. Changes of blood viscosity and RCG in 44 cases with cardiovascular diseases after qigong exercises. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 57-58.

[27] Qin Chao, et al. Bidirectional adjustment of blood pressure and heart rate by daoyin tuina on the arterial blood and heart rate. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 107.

[28] Wang Chongxing, et al. Beneficial effect of qigong on improving the heart function and relieving multiple cardiovascular risk factors. Proceedings ,Third International Symposium on Qigong. Shanghai, China. 1990: 42 (in Chinese).

[29] Zhang Shengbing. Effects of mind-regulation by qigong on the human body. Proceedings, Fifth International Symposium on Qigong, Shanghai, China, September. 1994:68.

[30] Liu Yuanliang, He Shihai, Xie Shanling. Clinical observation on the treatment of 158 cases of cerebral arteriosclerosis by qigong. Proceedings, Second World Conference on Academic Exchange of Medical Qigong, Beijing, China. 1993:125.

[31] Liu Yuanliang. Private communication, 1993.

[32] Feng Lida, Effect of emitted qi on human carcinoma cells. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 1-4.

[33] Feng Lida, Effect of emitted qi on the L 1210 cells of leukemia in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 4-5.

[34] Zhao Shan, et al. Preliminary observation of the inhibitory effect of emitted qi on transplanted tumors in mice. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 46-48.

[35] Liu Tehfu,Wan Minsheng,Lu Oulun. Experiment of the emitted qi on animals. Proceedings, First World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1988: 60-61.

[36] Feng Lida,Peng Liaomin. Effect of emitted qi on prevention and treatment of tumors in mice. Proceedings, Second World Conference for Academic Exchange of Medical Qigong, Beijing, China. 1993: 106-107.

[37] Sun Quizhi, Zhao Li. Clinical observation of qigong as a therapeutic aid for advanced cancer patients. Proceeding, First World Conf Academic Exchange of Medical Qigong, Beijing, China. 1988: 97-98.

[38] Chen Guoguang. The curative effect observed for 24 (cancer) cases under my emitted qigong treatment. Proceedings, Second International Conference on Qigong, Xi'an, China. 1989: 141-142.

[39] Lo Jifeng, et al. Changes of peripheral blood cell population and immune functions in 31 nasopharyngeal carcinoma (NPC) patients treated with radiotherapy and qigong. Proceedings ,Third National Academic Conference on Qigong Science, Guangzhou, China. 1990: 94-95.

[40] Yu Yi, et al. Effect of self-controlling qigong therapy on the immune function of cancer patients. Proceedings,

Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993: 128.

[41] Zhao Hongmei, Bian Jingnan. Curative effect of intelligence qigong on 122 tumor patients. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China.1993:130.

[42] Wang Ying. Clinical observation on 30 cases of cancer treated by qigong therapy Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131.

[43] Xong Jing, Lu Zhong. Curative effect on 120 cancer cases treated by Chinese-Western medicine and qigong therapy. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 131

[44] Fu Jingzhi. Treatment of advanced gastric cancer in the aged by the combination of qigong and medicinal herbs. Proceedings, Second World Conf on Academic Exchange of Medical Qigong, Beijing, China. 1993: 132-133.

[45] McGee C T, Chow E P Y. Miracle Healing from China—Qigong. 1st ed. Coeur d'Alene, ID: Medipress; 1994: 8-19.

APPENDIX 1C
THE EFFECT OF QIGONG ON THERAPEUTIC BALANCING MEASURED BY ELEXTROACUPUNCTURE ACCORDING TO VOLL (EAV): A PRELIMINARY STUDY

Kenneth M. Sancier, Ph.D.

ABSTRACT

Electroacupuncture According to Voll (EAV) was used to monitor the effects of qigong practice on therapeutic balancing of subjects. In EAV the electrical conductance of the skin above individual acupuncture points is measured using low voltage and current. Diagnosis depends on measuring the relative electrical conductance and its time dependence. An important diagnostic criterion of degeneration of an organ is an indicator drop which occurs during the measurement when the conductance decreases from an apparent maximum value and then levels off. Two series of EAV measurements were made before and after healthy subjects practiced qigong. Measurements were made at 24 acupuncture points at the ends of the Channels of the fingers and toes of a subject and were made by the same operator and equipment. The subjects were asked to perform a qigong exercise of their choosing. In the first series, four subjects were examined by EAV before and after qigong exercise. Qigong exercise decreased the average EAV measured values of the four subjects in the range of -19 to -31%. Qigong eliminated indicator drops for three subjects and reduced the indicator drop by 80% for the fourth subject. In the second series, each of seven subjects was examined by EAV three times in a blind protocol so that the operator did not know whether a subject had practiced qigong before the second or third examination. Qigong exercise changed the average EAV measured values in the range of -17 to -35% for four subjects and 4 to 15% for three subjects. Indicator drops appeared for three subjects, and they were eliminated for two subjects and reduced by 50% for the third subject. These preliminary studies indicate that EAV can monitor the effects of qigong on changes in the therapeutic balancing of the Channel-organ system. Suggestions are made of how EAV can provide basic information about qigong and its applications.

INTRODUCTION

Electroacupuncture According to Voll (EAV) is one of several electroacupuncture techniques for measuring local skin electrical conductance for diagnostic and therapeutic purpose(1). In general, EAV emphasizes preventive medicine by providing the medical practitioner with a diagnosis of a subject's health and its change at every stage of therapy. In the present study, EAV was used to monitor the effects of qigong practice on therapeutic balancing of subjects (2,3). The equipment is basically a D.C. resistance meter that measures relative changes in electrical conductance of the skin above an individual acupuncture point. Diagnosis depends on measuring the relative electrical conductance and its time dependence. The relative conductance is measured by a meter that is calibrated from 0 to 100. The "normal" value of 50 is expected when the acupuncture point and associated organ are in a physiologically balanced condition. Higher values (higher conductance) are associated with inflammation and lower values (lower conductance) with degenerative disturbances. The indicator drop is an important diagnostic criterion in EAV of a functionally disturbed organ. An indicator drop occurs during a measurement when the conductance of a given acupuncture point increases to an apparent maximum value and then decreases to a lower equilibrium value. In the present study, EAV was used to monitor the effects of qigong practice on therapeutic balancing of healthy subjects.

Two series of EAV measurements were made by the same operator and equipment. In both series, the subjects were asked to perform a qigong exercise of their choosing, usually sitting or standing meditation or moving qigong. The first series was performed during an international conference (4) at which qi and the human Channel system were among the main topics. About six months later, the second series was conducted in San Francisco with different subjects (5).

MATERIALS AND METHODS

The EAV instrument, a Computronix Acupro II system (Synergy Health Systems, 1223 Wilshire Boulevard #321, Santa Monica, CA 90403) was designed and operated by Douglas C. Leber. The instrumental test parameters were: 1.25 volts D.C. output voltage, 12.7 microamperes current output at full scale, and 95,000 ohms resistance at midscale of 50. A schematic of the electrical circuit is shown in Fig. 1. The instrument was calibrated from 0 to 100, and the value of 50 indicates that an organ associated with the acupuncture point was free of pathological problems. The tester pressed the brass probe electrode (0.13 cm diameter and connected to the positive side of the circuit) gently onto the acupuncture point as shown in Fig. 2. To minimized variations of skin conductance due to perspiration, tap water was applied to the skin in the region of the acupuncture point before the measurement. The cylindrical brass hand electrode (1-inch diameter and connected to the negative side of the circuit) was held in the subject's free hand, which was also moistened with water. Twenty four acupuncture points at the ends of the Channels of the fingers and toes of a subject were tested in about 5 minutes. The computer was programmed to provide information on the right and left side of the body for twenty organs and physiological functions of the body.

The subjects were 30 to 65 years of age. They were examined by EAV before and after practicing a qigong exercise for 10 to 15 minutes. The subjects went to a secluded spot to practice their own style of qigong, which was sitting or standing meditation or a form of moving qigong. The subjects were asked to balance their energies, and they were examined within 15 minutes after practicing qigong. One subject was examined by EAV before and after receiving qigong therapy from a qigong master and was also examined by EAV three more times within 22 hours.

RESULTS AND DISCUSSION

Series 1: In Series 1, four subjects were examined by EAV before and after qigong. Subjects A, B, and C, who had practiced qigong for more than 10 years, practiced their own style of qigong for 10 to 15 minutes. Subject D, who had studied qigong for about one year, was given a qigong balancing treatment for about 10 minutes by subject A, a qigong master who used a combination of acupressure massage and emitted qi to balance the Channels and organs of subject D. Subject A was not aware of that variations in the polarity of qi may exist in different parts of the his hands and of the body of the subject (6).

Examples of the EAV measurements before and after qigong are shown for subject D in Fig. 3 and for subject C in Fig. 4. The 20 pairs of horizontal bars represent the measured values at the acupuncture points associated with specific organs and functions of the body as described below the pair of bars. For a given pair of bars, the upper bar represents the value on the left side of the body and the lower bar that on the right side of the body. For example, LUCL and LUCR represents the lung Channel on the left and right side of the body, respectively. The scale for the measured values is shown at the bottom of each graph in the figure. A vertical line at the "normal" reading of 50 is a convenient reference for comparing changes that resulted from qigong practice. The two numerical values to the left of each bar are the actual measured values. The maximum value is on the left and the equilibrium value, which may be lower if there is an indicator drop, is on the right. The magnitude of the indicator drop is represented by a "white" bar and its numerical value is the difference between the maximum and the minimum measured values.

Visual inspection of Fig. 3 and Fig. 4 shows that qigong had decreased all measured values and had either eliminated indicator drops or reduced their numbers. However, qigong did not restore balance to some of the Channels and the corresponding organs. For example after the qigong balancing treatment for subject D, the measured values for the spleen, liver and gall bladder Channels on the right side of the body remained elevated with respect to the others at values of 66, 69 and 67, respectively. In the case of subject C, the measured values remained elevated for the stomach at values of 77 and 66 and for the urinary bladder on the right side of the body at a value of 68, and two of the nine indicator drops remained. These observations suggest that EAV can be used to monitor the effects of qigong therapy on individual organs of the body.

The results of the EAV measurements obtained before and after qigong for each of the four subjects are summarized in Table 1. For each subject, average measured values were calculated from the 40 values, 20 on the left and 20 on right side of the body (see Fig. 3). For each subject, a percentage change (change %) was calculated from the average measured values obtained before and after qigong. The results show that qigong had changed the average measured values of each practitioner in the range of -19 to -31%. The sum of all indicator drops (in meter divisions) for a given EAV examination were either eliminated or reduced markedly by qigong. The style of qigong of each subject and the duration of the practice are noted in a separate column in Table 1.

In Table 1, the plus and minus value that follows an average EAV value is an apparent "standard deviation" that indicates that individual Channels or body functions have measured values greater or less than the average value. This standard deviation would be zero if all the Channels and body functions were balanced and had the same measured AEV values. Individual Channels and body functions may not be balanced for many reasons, including chronic, acute, stress, and emotional factors. The magnitude of the apparent standard deviation may be regarded as a semiquantitative indicator of imbalance in the Channel system.

In a preliminary study to determine possible lasting effects of qigong, subject D was examined by EAV a total of four times over a period of 22 hours. The average measured value was 75.1±5.2 before qigong, 51.5±9.9 twenty minutes after qigong balancing, 55.7±10.4 four hours later after lunch, and 59.0±8.1 twenty-two hours later after breakfast. The increases in the standard deviations after the first 20 minutes were mainly due to large elevations in the measured EAV values of the Channels of the liver at a value of 85, spleen at 85, large intestines at 68, and kidney at 65. These changes are probably caused by food eaten at lunch.

Series 2. A blind protocol was used in Series 2 so that the EAV operator did not know when the seven subjects practiced qigong. This was accomplished by measuring the EAV of each subject three times, and in between measurements the subject left the examination room. After the initial EAV examination, the subjects were given a random selected card instructing them to practice qigong after the first or after the second EAV examination. Later the subjects informed the examiners when they practiced qigong. Of the seven subjects, E, F, G, H, I and K were experienced qigong practitioners, while subject J had studied qigong one year. Computer printouts similar to those in Fig. 3 and Fig. 4 were analyzed for each subject. The averages of the 40 measured values and the sum of indicator drops are tabulated in Table 2. The average measured values obtained just after qigong practice are indicated in boldface.

The results show that qigong practice changed the average measured readings of most subjects. The average measured values decreased for four subjects (E, F, G, H) in the range of -17% to -35%, increased for two subjects (J and K) in the range of 12% to 15%, and remained essentially unchanged for subject I. For subject I, a decrease in the average measured values was observed after qigong, that is, from the second to the third examination (82.7±2.7 to 61.2±11.8, respectively). This decrease may be due to a delay in the effects of qigong exercise. Indicator drops were present in the EAV of only three subjects before qigong, and these drops were either eliminated or decreased by qigong.

In series 2, the average measured values for four subjects (E, F, G, H) showed similar decreases in magnitude as those in series 1. However, the average measured values for the other three subjects (I,J, K) increased. This difference in response of individual qigong practitioners can be accounted for in many ways, such as by differences in qigong style, intention of the practitioner, and physical and emotional states. For example, we had asked the subjects to balance their energies, but subject J later told us that she had tried to increase her qi, thus perhaps explaining the increase of 12%.

Because of scheduling problems, we were unable to carry out control experiments. However, the reproducibilities of successive AEV measurements before or after qigong were satisfactory for five of the subjects. This can be seen in Table 2 by comparing the second and third measurements (after qigong) for subjects E, F, H, and J, and by comparing the first and second measurements (before qigong) for subject G. These successive readings changed less than 4%.

CONCLUSIONS

Our preliminary studies suggest that the EAV technique can monitor the effects of qigong on the therapeutic balancing of the Channel-organ system. EAV also should be valuable in other applications of qigong: 1) evaluating the effectiveness of different types of personal qigong exercises for balancing organs of the body, 2) measuring the receptivity of subjects to emitted qi, 3) evaluating the effects of different kinds of emitted qi, for example, positive and negative qi (8), 4) determining the duration of the balancing effects of qigong, 5) determining the relative healing power of qigong masters for certain medical problems, 6) comparing the relative effectiveness of energy balancing by qigong with other therapies, such as acumassage, acupuncture, healing with hands, Reiki, as well as conventional physical exercise, 7) investigating the therapeutic effectiveness of qi energy stored on materials such as paper (7), 8) investigating the therapeutic effectiveness of electrical, mechanical and laser stimulation(8), and 9) investigating the interaction between a qigong master and a subject (9).

ACKNOWLEDGEMENTS

The author acknowledges the cooperation of Douglas C. Leber of the Computronix Electro-Medical Systems, who designed and operated the EAV instrument. He also acknowledges support in arranging the experiments by Jim Jose of Synergy Health Systems, qigong master Gong Xin Xiong, Harry Kung, Linda Kung, and the qigong practitioners.

REFERENCES

1. Fred M. K. Lamb, Jr., M.D., Julia J. Tsuei, M.D., Zixian Zhao, M.D., M.S.: Bioenergetic Regulatory Measurement Instruments and Devices. Am. J. Acupuncture, Vol.16, No. 4: pp. 345-349, Oct.- Dec., 1988.
2. R. Voll: Twenty years of electro-acupuncture diagnosis in Germany: A progress report. Am. J. Acupuncture, Vol. 3, No. 1: pp. 7-17, 1975.
3. See collection of nine articles on Electroacupuncture According to Voll in the Am. J. Acupuncture, Special EAV issue, pp. 5-99, Second printing, 1989.
4. International Conference on Bioenergetic Medicine, Past, Present and Future, University of Hawaii, September 30, 1989.
5. Presented at the Second World Conference on Academic Exchange of Medical Qigong, Beijing, China, September 1993.
6. Omura,Y., Common factors contributing to intractable pain and medical problems with insufficient drug uptake in areas to be treated, and their pathogenesis and treatment: Part I. Combined use of medication with acupuncture, (+) Qigong energy-stored material, soft laser or electrical stimulation, Acupuncture & Electro-Thera. Res. Int. J., Vol. 17: pp. 107-148, 1992.
7. Omura, Y., Storing of Qigong energy in various materials and drugs (Qigongnization): its clinical application for treatment of pain, circulatory disturbance, bacterial or viral infections, heavy metal deposits, and related intractable medical problems by selectively enhancing circulation and drug uptake, Acupuncture & Electro-Thera. Res. Int. J., Vol. 15: pp. 137-157, 1990.

8. Omura, Y., Connections found between each Channel (heart, stomach, triple burner, etc.) & organ representation area of corresponding internal organs in each side of the cerebral cortex; release of common neurotransmitters and hormones unique to each Channel and corresponding acupuncture point & internal organ after acupuncture, electrical stimulation, mechanical stimulation (including Shiatsu), soft laser stimulation or Qigong, Acupuncture & Electro-Thera. Res. Int. J., Vol. 14: pp. 155-186, 1989.
9. Omura, Y., Unique changes found on the Qigong (chi gong) master's and patients's body during Qigong treatment; their relationship to certain Channels and acupuncture points and the re-creation of therapeutic Qigong states by children and adults, Acupuncture & Electro-Thera. Res. Int. J., Vol. 14: pp. 61-89, 1989.

Table 1. The effect of qigong exercise on the average measured EAV values and sum of indicator drops for each of 4 subjects before and after qigong (Series 1).

Each subject was tested before and after practicing qigong. Subject D was given qigong balancing therapy by subject A. The change (%) refers to the difference between the measured values before and after qigong. The indicator drop (ID) divisions are the sum of all such drops for a given subject. The style and duration of qigong practice is shown in the right-hand columns. Average measured values for before and after qigong were calculated from the average of 40 readings, 20 on the left side of the body and 20 on the right side (see Fig. 3).

Average measured values Sum ID Divisions Qigong
Before After Change Before After Time
SubjectQigongQigong(%)QigongQigongStyle(min.)
A69.0±3.351.4±4.4-26200Moving15
B72.0±3.953.3±7.1-26220Standing10
C67.0±6.654.6±5.7-1912928Sitting10
D75.1±5.251.5±9.9-31530Therapy10

Table 2. A blind study of the effect of qigong exercise on the average measured EAV values and sum of indicator drops for each of 7 subjects before and after qigong exercise (Series 2).

Each subject was tested 3 times. The bold faced numbers indicate measured values obtained just after qigong practice. Secret instructions informed the subjects to practice qigong either after the second or third EAV test. The change (%) refers to the difference in measured values before and after qigong practice. Indicator drop (ID) divisions are summed for a given subject. The style and duration of qigong practice are shown in the right-hand columns. The average measured values before or after qigong were calculated from the average of 40 measured values, 20 on the left side of the body and 20 on the right side (see Fig. 4).

Average Measured Values Sum ID Divisions Qigong
Change Before After Time
SubjectFirst testSecond testThird test (%)QigongQigongStyle(min.)
E61.3±15.439.6±11.238.6±9.7-35140Moving 15
F74.8±3.052.0±5.251.7±4.0-3000Sitting10
G76.1±2.6 79.1±4.258.1±8.6-2700Sitting10
H74.4±3.262.1±12.758.9±10.5-1700Sitting10
I79.2±4.182.7±2.761.2±11.8400Sitting10
J67.5±8.974.5±7.071.1±4.2122814Moving15
K67.2±11.5 53.9±13.562.1±14.71580Sitting10

APPENDIX 2

INTRODUCTION

This second Appendix includes research information on Medical Qigong therapy, collected by my good friend Richard H. Lee, Director of China Healthways Institute, San Clemente, California. Mr. Lee, a council board member of the World Academic Society of Medical Qigong, in Beijing, China, is a specialist in Western medical technology. He has worked extensively with the research department of the Beijing College of Traditional Chinese Medicine, studying the relationship of infrasonic waves emitted as Qi.

This research, focused on the Extension of Emitted Qi by various Medical Qigong Masters / Doctors, was motivated by the pioneering research of Professor Lu Yan Fang at the National Institute of Electro-Acoustics in Beijing, China. Professor Lu developed prototype devices which simulated the frequency output of Medical Qigong doctors.

Mr. Lee has spent many years in China and abroad researching and collecting scientific data which led to the development of several Infratonic Medical Qigong Therapeutic Machines (in Chaotic resonances), which are currently being sold worldwide.

This second Appendix is a continuation of the research devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality. The result on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality being positive in conclusion.

The following information in this second appendix represents but a fraction of Mr. Lee's findings, however, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

APPENDIX 2

SCIENTIFIC INVESTIGATIONS INTO CHINESE MEDICAL QIGONG THERAPY

Richard H. Lee
Director China Healthways Institute
San Clemente, California

INTRODUCTION

In the years since the first edition of "Scientific Investigations" was published, the debate over the existence of Qi, the legitimacy of healers, and the reality of PSI phenomena has heated up. While most readers embrace energetic concepts like Qi as a useful framework for understanding otherwise unexplainable phenomena, some continue to reject Qi. Even in China, while an estimated 100 million people practice Qigong, many professionals remain skeptical. Scientists who actually investigate Qigong generally conclude that the phenomena really do happen, but still try to maintain an arms-length distance from it. Many doctors are also reluctant to discuss Qi openly even though most of the general public is quick to identify Qi as feelings of abundant vitality, and the absence of Qi as fatigue.

There appear to be three principal barriers to serious consideration of Qi.

1. Inertia: "There is not enough data." "These studies have not been replicated." This seems to stem from an unwillingness to explore the studies that have been done. This book presents enough information to satisfy the open-minded investigator and the Qi explorer, but nothing was added in this edition to satisfy those who demand overwhelming evidence.
2. Hidden Bias: This is probably the biggest resistance that Qigong research faces. Many people simply refuse to consider the possibility of action at a distance, as exhibited in Qigong and PSI research. On the basis of this single assumption, all empirical data which supports such phenomena must be rejected as mistakes, experimenter bias, or falsification. Nothing was added to this edition to try to overcome this hidden bias.
3. Need for a framework: While the information presented in the first edition of this book answered the question, "Is the phenomenon of Qigong scientifically measurable?," It left readers without a framework other than faith by which they can understand these phenomena. Without a framework of understanding, many are reluctant to accept these strange experimental results.

Others are less concerned about physical substance and seek to understand Qi as an aspect of intention or consciousness. This more closely parallels the approach of most Qigong masters who, through years of training, gradually gain mastery over their own mind (through focus of attention), emotions (through calmness and releasing desires), and physical body (through discipline). Interestingly, this second viewpoint is similar to the view of many skeptics who insist that phenomena of Qi are simply "a product of the mind." Combining the idea that Qi is strongly influenced by the mind, with the research in this book which shows that Qi is measurable in the physical world, we get a picture of Qi as a mind / body connection which is appealing to most readers.

The Chinese do not view Qi as a scientific substance, but rather, as the essence of life, the bridge of consciousness between mind and body, and the "eternal now" in which all activity occurs. The physical body supports physical fields (Qi) which support vibrational frequency patterns (Qi) which allow communication (Qi) between some kind of abstract consciousness (mind) and a physical body which is our vehicle.

Richard H. Lee, January, 1999

QIGONG MASTERS AND ENERGETIC MEDICINE

For thousands of years, China has been a major source of the world's technology. We are surrounded by Chinese inventions such as paper, movable type, gun powder, and surgical anesthesia. And while the West has shot past China in many technological areas, there are still areas where China holds an edge. The Chinese investigations into the phenomenon of Qigong over the last ten years have, in some ways, placed Chinese researchers at the forefront of energetic medicine. Many findings of this research are presented on the following pages.

Qigong has been practiced for thousands of years in China. Yet it has never before had the broad acceptance that it has today. This is because the scientific investigations into the nature of Qigong show clearly that something is going on beyond what physics might predict. Measurable energy is emitted from the hands of those who have mastered these Qigong techniques. These masters can influence events from a distance with energy (or signal) emitted from their hands. Further, these energy emissions and Qigong training exercises have scientifically measurable, beneficial effects on health and healing.

Qigong masters have demonstrated many remarkable feats. Some can soften and bend steel. Some can light fluorescent tubes with their bare hands. Others can kill cancer cells in laboratory experiments, and many can, from another room, alter the brain wave patterns of test subjects including test animals.

The energy they emit from their hands is measurable with scientific equipment in the forms of infrared, magnetic, and acoustical energy. In one test, a strong pulse of static electricity was measured at a point 50 mm above the head of a Qigong master. In another, a strong magnetic field was measured just at the instant that a heavy iron bar hit a Qigong master's head. The bar bent as it would against Superman's head. He was unharmed with no sign of a lump or a bruise.

In 1979 and 1980, the leaders of China decided to put Qigong to the test. The Western scientific test. High technology government organizations across the country, many as prestigious as the National Atomic Energy Lab in Shanghai, and the Space Science and National Electro-Acoustics Institutes in Beijing, started testing Qigong masters and promptly discovered that the phenomenon of emitted Qi was measurable. Some of these laboratories went further and actually developed therapy equipment such as the Infratonic QGM which reproduces the infrasonic component of the emitted Qi. This research has led me into Qigong and taught me many things I didn't think I ever wanted to know.

Today's Western scientists know little about Qigong, but it is an important healing technology in the world today. It is the basis of acupuncture and traditional Chinese herbal medicine, yet it presents effective medical intervention when used alone. Americans have been radiating cancer with energy emitted from machines for the last few decades. The Chinese have been using energy emitted from the hands of Qigong doctors to treat cancer for thousands of years, finding that, in many cases, tumors can be shrunk without side effects.

These scientific experiments are repeatable. Scientific laboratories in America will reach the same conclusions as those in China. The evidence is persuasive that systematic Qigong training produces real benefits. And advanced practitioners can emit energy from their hands that is measurable with scientific equipment. Further, therapeutic equipment has successfully reproduced many of the beneficial effects observed. The Chinese scientific investigations into medical Qigong are indeed at the heart of a medical revolution that is improving the science of health and healing around the world.

THE SCIENCE OF CHINA

For thousands of years China has been a center for the advancement of civilization. The science of China has produced a wide variety of technologies. Gunpowder is the central ingredient to modem warfare. Paper and printing with movable type have for centuries been at the center of education, communication, and information storage. Despite the growth of computers, this system will continue to be the world's dominant information storage method.

Many centuries ago, a physician named Hua Tou invented surgical anesthesia and performed many kinds of surgeries like appendectomies and other of today's popular surgeries. While the Chinese contributed to what has today become modem medicine, they also created "Chinese medicine," a system of medicine based on the Chinese science of "Qi" or vital energy. It is this system from which acupuncture and Chinese herbal medicine have evolved. And at the heart of this system lie the science and practice of Qigong.

QIGONG'S LONG HISTORY

Through the millennia Qigong has been an influential force in the molding of China. Until recently it has been hidden from public view. Like many other disciplines, the art of Qigong has been taught from father to son or from master to disciple only after a strict oath of secrecy has been given. Knowing a science that others don't know gives a practitioner a decided advantage in earning a living. Thus, until recently, very few books have been written about Qigong, particularly about the more advanced techniques of study. The exception is the secret books that have been maintained in religious monasteries, being copied and recopied by studious monks through the centuries.

Even this activity in monasteries came to an abrupt halt with the Communist government in China when Mao put an end to most religious activities. Through the Cultural Revolution, monasteries and temples were shut down. Many of the ancient texts and artifacts were destroyed. Others were stored in government libraries. Many of the Qigong masters were killed. A few went into hiding or escaped from the country.

At the end of the Cultural Revolution, Deng Xiao Peng set about rebuilding much of the cultural heritage of China, allowing limited freedom of religious expression. He also promoted the rediscovery and preservation of traditional Chinese science and medicine, which included Qigong.

It is interesting to note that much of the willingness to share the secret teachings of Qigong came from outside countries. Masters who had emigrated to the West had started teaching the techniques openly without the strict code of silence. And some of the secret Qigong texts that had been taken from the temples and monasteries were published for people around the world to read. Qigong was no longer a secret so there was no longer any value to secrecy in China and Qigong masters became much more willing to teach what they knew.

The most influential factor in the Qigong revival was, perhaps, the new policy on the part of the China government to allow freedom of religion, particularly religion that had a basis in China's cultural heritage. Thus, Taoists and Buddhists who had preserved its teachings in secret for centuries were free to share their knowledge. This allowed Qigong practitioners to demonstrate their skills freely in public without fear of being condemned to death as witches or evil beings. Paranormal skills were again publicly accepted as a product of China's cultural heritage.

The big boost to Qigong came in the early 1980's when modern science turned to study the tradition of Qigong. The leaders of China reasoned that, if Qigong is to be accepted, it must pass the rigorous screening of scientific scrutiny. Studies were funded in the prestigious scientific institutes of China to determine whether the special skills of the Qigong masters could be scientifically verified. The results were most impressive. Many forms of energy were measured coming from the hands or bodies of Qigong practitioners in reproducible scientific experiments.

Scientific investigations into the effectiveness of medical Qigong quickly showed that, for many illnesses, medical Qigong was unsurpassed by any other medical system in the world. The discovery that energy emitted from the hands of Qigong doctors under controlled laboratory conditions was highly effective in killing cancer cells, gained extensive press coverage around the world.

Qigong science is a very new field with valuable technologies just beginning to surface. But given the fundamental nature of these discoveries, Qigong science could easily dominate the world of the 21st century.

SIMULATION OF EMITTED QI

Simulation of emitted Qi has been the goal of much of the Chinese Qigong research that has been conducted since about 1980. Many devices have been developed and tested, some of which have considerable potential. The following is a description of the most common forms:

Infrared devices have been developed some of which, through rare-earth emission surfaces, emit characteristic frequency bands, while others modulate the infrared output according to recorded patterns emitted by Qigong masters. These devices tend to have the most effect where coldness in the body is effectively treated with heat.

Electrical devices are often adapted to deliver current through the skin at frequency patterns that follow patterns produced by Qigong masters. These units have been used principally by acupuncturists.

Acoustical devices, which reproduce the subsonic output of the emitted Qi in the form of deep acting massage action, have been shown to be effective in pain management, circulation of Qi, blood, and endocrine fluids, and relaxation. The Infratonic QGM was used in several of the laboratory tests described in this book to simulate the emitted Qi in experiments. Because this device offers many insights into the nature and function of emitted Qi, and because many of the readers of this book use the Infratonic QGM at home or in their medical practices, much of the research around this device is included in this book.

SCIENTIFIC PLAUSIBILITY OF QI, THE HUMAN ENERGY FIELD

Some readers might tend to reject the findings of this study based on their assumption that the existence of human energy fields and the altering of a person's energy field are outside the laws of physics, and therefore impossible. It is therefore appropriate here to briefly review some of the findings of researchers mostly within the last decade to show that these energy fields are quite measurable with existing equipment and that there is strong evidence that known biophysical mechanisms can account for the observed phenomena.

First, the human body clearly has a vibrating electrical field, measurable for decades through such techniques as EMG, EEG, and EKG. How is this possible? The classical view of the human cell as a bag of water with a nucleus has been, for the most part, rejected in favor of models in which a highly structured collection of filaments, tubes and fibers connect the cell's nucleus not only to the surface of the cell but beyond, in a matrix which connects and interrelates the cells of the entire body (Oschman 1994) and (Pinta and Coffee 1991). Second, helical molecules within the body, such as DNA, keratin, collagen, actin, and myosin have piezoelectric qualities. As they elongate and contract they produce voltages. Also, when voltages are applied, they expand or contract producing mechanical activity and acoustical energy.

While it is well established that electrical signals are produced and received by the body, readers are more likely to question whether a magnetic field interpenetrates and surrounds the body. In 1963 researchers at Syracuse University reported the first measurements of the magnetic field of the human heart, just a millionth the strength of the earth's magnetic field, yet highly coherent and measurable throughout the body and beyond. Then in 1971, the SQUID, a highly sensitive superconducting magnetometer, was developed which was used to measure the magnetic field of the brain, 100 times weaker than that of the heart (Cohen 1972).

Many researchers now believe that the magnetic field of the heart, which extends throughout the body plays a role in regulating biological functions. Then, in 1991, Seto and colleagues measured the magnetic field between the hands of healers to be as strong as 1 milligauss. This is less than 1% the strength of Earth's magnetic field, but 1000 times as strong as the magnetic field produced by the heart. If the heart regulates biological functions with it's weak field, it is certainly feasible that a magnetic field from the hands of healers, measured to be many times stronger than the heart field might influence biological functions.

But, how can a magnetic field be produced from the hands? The Josephson junction, a superconducting semiconductor which is the basis of the SQUID's extremely high sensitivity to weak magnetic fields, has been predicted in quantum physics calculations to be present in living tissue. This was verified just ten years ago. (Del Giudice et al 1989). DNA is one of the biochemical structures shown to act as a superconducting magnetic field detector, able, through the Hall effect, not only to detect very weak magnetic signals, but also to produce magnetic signals (McCraty 1993). Thus DNA may act as a signal amplifier, releasing its energy coherently into passing magnetic waves, much as the ruby rod of a laser releases its light energy coherently with a passing photon to produce a laser beam. This provides a mechanism by which healers might produce the observed relatively strong biomagnetic signals, and by which a test subject or patient might be influenced by such signals (Oschman 1993).

Thus, modern advances in science now provide mechanisms by which it is possible for healers to produce magnetic and electric signals of significant magnitude to influence patients. For those who seek a more indepth overview of this rapidly advancing field of knowledge, "The Scientific Basis of Energy Medicine" by James L. Oschman, 1998, is highly recommended.

REFERENCES

Cohen D 1972 Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer. Science 175: 664-666

Del Guidice ES, Doglia S, Milani M, Smith JM, Vitiello G 1989 Magnetic flux quantisation and Josephson behaviour in living systems, Physica Scripta 40: 786-791

Dumitrescu I, Kenyon JN Electrographic imaging in medicine and biology, Nevil Spearman Limited, Suffolk England, 1983

Konikiewicz L, Griff L Bioelectrography a new method for detecting cancer and monitoring body physiology, Leonard Associates Press, Harrisburg, 1984

Lee R Bioelectric vitality- exploring the science of human energy, China Healthways Institute, 1997

Liu G, Wan P, Peng X, Zhong X The influence of the emitted Qi on the auditory brainstem evoked responses (ABER) and auditory middle latency evoked responses (MLR) in cats, Proceedings of the First World Conference for Academic Exchange of Medical Qigong, Beijing College of Traditional Chinese Medicine, 1988.

McCraty R, Atkinson M, Tiller WA 1993. New electrophysiological correlates associated with intentional heart focus, Subtle Energies 4: 251-268

Niu Xin, Liu G, Yu Z Measurement and Analysis of the Infrasonic Waves from the Emitted Qi, Proceedings of the First World Conference for Academic Exchange of Medical Qigong, Beijing College of Traditional Chinese Medicine, 1988.

Oschman JL 1993 A biophsysical basis for acupuncture. Proceedings of the First Symposium of the Committee for Acupuncture Research, CAR, Boston.

Oschman JL, Oschman NH 1994, Somatic recall, parts I-II Soft tissue memory / Soft tissue holography, Massage Therapy Journal, American Massage Therapy Association, Lake Worth, FL 34: 36-45, 66-67, 101-116

Pinta KJ, Coffee DS 1991 Cellular harmonic information transfer through a tissue tensegrity-matrix system. Medical Hypotheses 34:88-95

Seto A, Kusaka C, Nakacato S et al 1992 Detection of extraordinarily large bio-magnetic field strength from human hand. Acupuncture and Electro-Therapeutics Research International Journal 17:75-94

THE FUTURE OF QIGONG

Chinese research has found that Qigong training increases balance, quickness and endurance in athletes, and that fighter pilots trained in Qigong have quicker reactions and greater tolerance for the G-forces of sudden turns. Qigong has provided revolutionary treatments for hypertension and cancer, and has the potential for much more.

If, as the Chinese claim, Qi is the fundamental substance of the universe making up both matter and energy, and if we can through our bodies and our awareness, control and direct this Qi, not only within our own bodies, but as has been seen, beyond the limits of our bodies, the fantasy of "the Force" in Star Wars comes within the grasp of reality. Psychic phenomena like telekinesis and mental telepathy become minor extensions of the human manipulation of Qi. The Western ideas of positive thinking take on new meaning:

"The mind is the builder" and "Your thoughts create your reality."

The Chinese science of Qigong gives us proven tools to strengthen our ability to structure the world around us by training our minds and bodies to feel and understand Qi.

Remember that Western science and technology gave us electric motors which, when coupled with man-made equipment, can do the work of a thousand men. This came with our understanding of the principles of electricity. Emitted Qi simulation devices such as the Infratonic QGM provide this work-saving feature of machines. As we learn more about Qi, how it functions, and how it is directed by the human body, we will learn how to construct more and more powerful machines. We will see Qigong devices not only in hospitals but also in agriculture, in building air conditioning systems, in chemical processing plants and in space flight. The scientific Qigong research that started just a decade ago will grow and grow, and perhaps dominate our world of the 21st century.

However, the real emphasis of the Qigong movement and the intent of most Qigong teachers is not just to heal patients with illness, but to expand human potential. The research presented in this book shows that Qigong exercises help people to develop, not only the ability to heal themselves and others, but to do things which scientists considered impossible just a few years ago. Through a study of Qigong, there is strong evidence that we can become more than we ever thought possible.

INFRASONIC QIGONG RESEARCH

The discovery that specific bands of very low frequency sound influence healing and vitality was an accident. A "friend of the family" walked in to the National Electroacoustics Laboratory in Beijing and asked Prof. Lu Yan Fang, an acoustics scientist, if she would test his hand to see if any sound was coming out of it when he emitted Qi. She thought he was crazy but went along because she didn't want to dishonor her family by turning out a friend.

Prof. Lu's discovery that he did indeed emit a very strong signal in the infrasonic range was quite a surprise. This signal was 100 times stronger than that of a healthy, normal person. Elderly and sick people had even lower readings.

To verify that this infrasonic sound had a physiological effect, she built a device to simulate this output and tested it in hospitals. It worked to relieve pain and assist recovery in a wide variety of patients.

Prof. Lu's patent for the Infratonic QGM was the first ever filed in China and her device was awarded three awards for outstanding contribution to science and medicine from Chinese national science organizations.

It is clear from the Chinese infrasonic research presented in this chapter that infrasonic sound is an important component of emitted Qi.

INFRASONIC SIMULATION OF EMITTED QI FROM QIGONG MASTERS

by Professor Liu Guo-long, MD, PhD
Beijing College of Traditional Chinese Medicine
with Richard Lee

SUMMAY OF RESEARCH

Emitted Qi from Qigong masters clearly has a strong effect on the central nervous system, not only in humans but also in animal subjects. Qigong meditation and the Infratonic QGM have similar effects. These findings are summarized below:

Repetition of these tests with animals confirmed that these changes are physiological and not psychological phenomena by eliminating placebo effects.

DOUBTS ABOUT EMITTED QI

At the beginning of my research with Qigong, I was confident that a study of neurophysiology would prove that the "Wai-Qi" emitted by Qigong masters was nothing more than a psychological factor induced by the waving of hands and hypnotic suggestion.

Qigong is a system of physical and mental exercises which has been practiced in China for thousands of years. Up until the 1980's, most Chinese considered Qigong masters (those who have mastered this Qigong training) to be mythical story book characters with super-human powers. However, as the Chinese government began to support scientific research into Qigong, the few remaining true Qigong masters began to surface, amazing researchers with feats like killing bacteria in test tubes and causing previously paralyzed people to get up and walk.

I thought that these were all tricks of one sort or another, perhaps magicians' sleight-of-hand, hypnosis, or optical illusion. When I got the chance to work with people claiming to possess these abilities, I was fascinated. This was my chance to test with scientific equipment, whether there was any truth behind the excitement about Qigong masters. I was confident that I could disprove the myths.

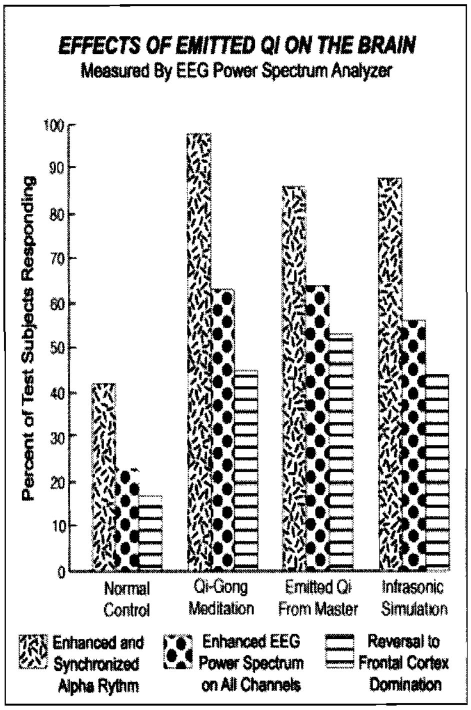

Emitted Qi has a pronounced and repeatable effect on EEG. It enhances frontal and occipital EEG power spectra, and often enhances the frontal so much that frontal becomes the dominant EEG activity whereas occipital dominance is more common. Emitted Qi also enhances and synchronizes the Alpha.

BACKGROUND

I was a specialist in Western medical technology assigned to the research department of the Beijing College of Traditional Chinese Medicine. The project was sponsored by the China Government Department of Education and the Department of Natural Sciences. My research group was assigned to study the relationship of infrasonic waves to emitted Qi. These research grants were motivated by the pioneering research of Prof. Lu Yan Fang at the National Institute of Electro-Acoustics in Beijing, who developed prototype devices to simulate the output of Qigong doctors.

Research has confirmed that humans have a very high degree of acoustic activity in the subsonic range below 20 Hertz (infrasonic), similar to the alpha rhythm of EEG. Also, people with chronic illnesses were found to have a much lower level of infrasonic activity, while Qigong masters had a much higher level of infrasonic output when they were emitting Qi.

The implication was that infrasonic sound might be related to human functioning, and further, that it might be in some way involved in the mechanism of brain functioning. We searched the available scientific literature for research papers for relationships between sound waves and brain waves but found none that had been done.

Extensive clinical research, based on infrasonic Qigong simulation, showed it to be effective for a wide variety of hospital problems. These results are what motivated the National Departments of Education and Natural Sciences to look further into the importance of infrasonic sound and what led to the research I am about to describe:

THE BRAIN AS A DETECTOR OF EMITTED QI

I had learned in my many years of research with the Electro-Encephalograph (EEG) that the human brain responds to even the most subtle of stimuli to the body, so I reasoned that, if there were really any scientific basis to emitted Qi, it would show up in the brain waves of test subjects who were placed in the path of these emissions. I expected to see no difference between the resting states and the Qi emission states.

What we saw was extraordinary. Within a few seconds after the Qigong master began to emit Qi, the subject's EEG would begin to shift. The EEG power spectrum was enhanced on all channels while the most pronounced increase was in the frontal lobe. Also, there was an enhancement and synchronization of the Alpha Rhythm in all channels. When the Qigong master stopped emitting Qi the EEG would gradually shift back toward the baseline readings.

To determine whether infrasonic energy was a significant part of the emitted Qi, we used the infrasonic Qigong prototypes in the same experiment. It was located 18 inches away, directly behind the back of the head of the test subject. The EEG electrodes were attached as before. The simulator was activated for short periods of time and the results recorded. We found that the effects on the receiver's EEG were quite similar to those of the emitted Qi.

Our further research involved monitoring the various sensory-cortical evoked potentials during Qigong meditation, emitted Qi, and infrasonic Qigong simulation. We again found very similar results from all three stimuli. We found that a large portion of the cerebral cortex was inhibited while other somatosensory cortex were excited. One of the significant findings of this study is that the inhibition of the cerebral cortex during Qigong meditation is clearly different from the excitation of the cerebral cortex that is measurable during sleep.

Through Acoustical Brainstem Evoked Response (ABER) it was found that the brainstem structures from the medulla to the hypothalamus were significantly facilitated. The brainstem plays an important role in regulating the functions of the inner organs, motor function, and emotion.

The implications of these studies were startling. Qigong masters can, without touch, voice, eye contact or any other traditional communication means, induce a clear, strong, and highly measurable change in a subject's brain functioning. A synchronization of alpha rhythm indicates deep relaxation, and is closely associated with accelerated healing. Enhanced power spectrum in the frontal lobe is especially significant because the association cortex of the frontal lobe is concerned with higher motor action, higher sensory function, emotional and motivational aspects of behavior, and integration of autonomic function. Facilitation of the brain stem, with its regulation of internal organs, may be a mechanism by which physical healing is induced or accelerated.

Despite these highly significant changes in EEG and evoked potentials, the subject had felt nothing and had no idea of the profound changes taking place within them.

The findings of these studies are solid evidence that a Qigong master can induce real physiological changes in a subject from several feet away, and further, may help to explain the high rate of recovery from chronic degenerative diseases in groups of hospital patients under the care of Qigong masters. These studies also show that the infrasonic Qigong simulator can induce similar changes in brain function and that, through Qigong meditation, a Qigong master can induce these same changes in his own brain.

SCIENTIFIC CONTROLS

There is much disagreement on how emitted Qi affects the brain. Many doctors insist that brain changes are psychologically induced, and that verbal suggestion, impressive hand motion, and expectation of the subject account for the observed phenomena.

To test this, we had several Qigong masters and people pretending to be Qigong masters treat the test subjects. The subjects were told that all were Qigong doctors, and all moved their hands in similar ways. We saw no significant changes in brain wave patterns with the fake Qigong masters, but when the real doctors emitted their Qi, or when we used the infrasonic Qigong simulator, we repeatedly got the highly significant changes.

Even this did not satisfy many of the doctors who reviewed our work, so we repeated the study with animals. We monitored EEG in awake rabbits and ABER in anesthetized cats as Qigong masters emitted Qi toward them. Even though there was no voice or eye contact between the Qigong masters and the animals, and the masters emitted Qi from several feet away, we saw shifts in EEG and ABER (Acoustical Brainstem Evoked Response) similar to those observed in the human subjects. This is a highly convincing result because all kinds of placebo effects are eliminated, yet modification of brain function at a distance remains.

INFRASONIC VS. ELECTROMAGNETIC INTERFERENCE

Extremely low frequency (ELF) electromagnetic signals can affect EEG. There is growing concern about the low frequency radiation produced by 60 Hz electrical power lines located close to people's homes and schools because of apparent disruption of brain and cellular function. ELF signals will cause the principal EEG power spectrum to show a spike at the frequency of the ELF signal. This is because the EEG is easily entrained by ELF signals. When the signal is discontinued, the EEG abruptly returns to normal. Electrical power lines operate at 60 Hz, which corresponds to EEG in the high Beta range, associated with mental anxiety and confusion. The entrainment of EEG to 60 Hz around power transmission lines may be why researchers are finding that people who live close to these power lines show a higher incidence of brain tumors.

The results using the infrasonic Qigong simulator were quite different from ELF signal entrainment. The shifts in EEG were gradual rather than abrupt, and while dominant EEG frequency did drift toward the dominant peak infrasonic frequency, it was a broad spectrum of EEG activity rather than a spike. The enhanced power spectrum continued after simulation was stopped, gradually decreasing and returning to the pretest state. From these observations it is clear that the effects of infrasonic are quite different from the entrainment of ELF electromagnetic signals.

SUMMARY

When I started this project in 1976, I had serious doubts about Qigong masters and emitted Qi. Now I am convinced that emitted Qi is very real and that Qigong is a very valuable art. With the support of the China government, an estimated 50 million Chinese are practicing Qigong daily, and many Western hospitals have added Qigong departments for patients with chronic and degenerative diseases. In addition, the China government has funded extensive scientific research into the nature of emitted Qi with the goal of advancing science and medical technology.

The study of emitted Qi and infrasonic sound, as they relate to human health and functioning, is a broad and exciting field. The use of Qigong in treating chronic degenerative diseases such as cancer and hypertension in China has proven very effective. It has been employed to accelerate healing for thousands of patients with a wide variety of diseases in Chinese hospitals. I am confident that emitted Qi, Qigong meditation, and the infrasonic Qigong simulator will play an increasing role in health care around the world.

SECONDARY ACOUSTIC BIOLOGICAL RESPONSE

(The following is from a technical paper presented in Beijing, China by Professor Lu Yan Fang, the inventor of the Infratonic Qigong Machine.)

"People often talk about sound. Common sense suggests that sound is what the ear can hear, the frequencies between 60 Hz. and 20,000 Hz.

Secondary sound is that sound that is outside of the range of the human ear, specifically low frequency sound below 60 Hz. Secondary sounds travel long distance with little attenuation.

In the 20th century, in the 1930's, this secondary sound was discovered, but it wasn't until the 1960's that it started drawing much attention. Since then the study, measurement, and analysis of the physical features of secondary sound have included:

1. Physical Features
2. Secondary Sound Measurement
3. Technology of Analyzing, and
4. Protecting the Body Against Strong Secondary Signals

Initial secondary sound biological response studies emphasized protecting against high level secondary sound which can damage the human body. Laboratory studies showed that strong secondary sounds beyond 150 dB are harmful to biological bodies. The American Environmental Protection Agency sets standards for secondary sound, concluding that secondary sounds below 130 dB are not harmful.

To this date there had been almost no studies investigating how secondary sounds affect biological processes. And never before has there been any thought of testing of the human body as a secondary sound source. Since 1983, we have worked with more than 70 Qigong masters, and have discovered for the first time that the human body actually emits a secondary sound signal. The frequency distribution peak of the signal is located in the secondary sound region of the sound spectrum, at about 10 Hz.

Through more than 100 repeated lab experiments we found almost identical results of frequency distribution. According to the analysis of test results in 1984, the first infrasonic Qigong prototypes were constructed. These machines used electroacoustic technology to simulate the frequency distribution of the Qigong master's low frequency signal, but stronger. This simulation of the human body life signal emission is then fed back to the human body. In the two years that followed, many clinical and animal tests proved that weak secondary sounds benefit the human body. The discovery that infrasonic sound is a basic substance of Qigong emission opens Qigong science's big door, opening an exciting new field for medical science.

Through hospital and clinical use, 1,134 patients were treated by the Infratonic QGM. Animal studies were also performed. Clinical use and animal use prove that this kind of secondary sound signal can strengthen many functions of the human body. The Qigong machine can have significant treatment value to the human body and works by reproducing the acoustic waves produced by the human body. It can promote circulation, regulate Qi and blood, open channels, and relieve pain.

The Qigong master is not alone in this ability to emit the secondary acoustical signal. Most everybody generates a similar signal. Young people with good health can emit the signal to a degree similar to that of a Qigong master. People with a weak body and elderly people have similar signals but much lower in power. People having Qigong training have a signal that is strong all the time, but especially strong during Qi emission. However, in normal people, as age increases or health condition worsens, the signal decreases. This explains why the human body is only sometimes a source of secondary sound. In the meantime, the body is also a receptor of secondary acoustical signals, and can be benefited by receiving the weak secondary acoustic signals.

Very early in the development of traditional Chinese medicine, this effect had been shown to exist. Anmo, and tuina, Chinese therapeutic massage techniques, give the human body this kind of low frequency stimulation. This stimulation through acupuncture points follows the cell as the basic pathway, and then goes to the meridian, and reaches the problem area. Then the problem can improve. Just like an old Chinese saying: "If you can get enough energy into the problem area, you can use that energy to treat the problem."

The human body's secondary acoustical biological effect is a scientific reality. The human body produces, delivers, emits, receives and controls this low frequency energy. This weak secondary acoustic effect has beneficial biological effect on the human body. It and its relationship with traditional Chinese medicine are of great value to the medical world. It is hoped that all scientists do research in this area.

MEASUREMENT AND ANALYSIS OF INFRASONIC WAVES FROM EMITTED QI

Niu Xin, Liu Guolong and Yu Zhiming
Beijing College of Traditional Chinese Medicine,
Beijing, China

The theory of traditional Chinese medicine suggests that Qi is one of the fundamental substances in human bodies. Modem scientific research of the essence of the emitted Qi has yielded some positive results. In order to find out the relation between the emitted Qi and infrasonic waves, to explore the mechanism of the emitted Qi, to find out how a person generates and receives the emitted Qi, and to provide a quantitative physical scale for indicating the strength of the emitted Qi for experimental studies of the effect of it, we measured and analyzed the infrasonic waves from the Qi emitted by Qigong masters.

The test, which was done by an infrasonic testing system made in Denmark, was conducted in a noise-proof room in the Institute of Sound and Electronics under the Ministry of Electronic Industry. The background noise in the room was lower than 30 dB (decibel). The microphone was hung in the air over Laogong (P-8, acupuncture designation for a point in the center of the palm), the distance being 2 cm with no contact.

The experiment tested the release of emitted Qi at Laogong (P-8) in the palm, Mingmen (Du-4), Baihui (Du-20), Dantian, and Jianzhi. Special attention was paid to the test of Laogong (P-8).

1. The peak frequency of the infrasonic waves from the tested Qigong and non-Qigong masters ranged from 8 to 12.5 Hz. In one case the frequency reached 16 Hz. In another two, it reached 6 Hz.
2. The infrasonic waves from the Qigong masters ranged from 45 to 76 dB and those from the non-masters, 45-50 dB. Comparison of the intensity of infrasonic waves during the Qigong state and the non-Qigong state before and after the emission of Qi showed a statistically significant increase ($P<0.01$).* The increase of wave intensity of the Qigong masters compared with that of the non-Qigong masters was also obviously significant ($P<0.01$).* **The energy of the Qigong masters was over 100 times higher than that of ordinary persons.**
3. The masters who had practiced Qigong for many years and often emitted Qi to treat patients had a higher intensity of infrasonic waves, reaching over 70 dB. Those who started to practice Qigong a short time before and mainly practiced *Nei Yang Gong* had a lower intensity of the infrasonic waves (lower than 60 dB.)

Infrasonic and ultrasonic waves are all sound waves which cannot be detected by human ears. The frequency of infrasonic waves is below 20 Hz. Many natural phenomena and artificial actions may generate infrasonic waves. Human bodies may act as a source or a receptor of infrasonic waves giving rise to a biological effect. The infrasonic information we acquire a from the measurement of the Qi makes it possible to study the effect of the emitted Qi.

The Valsalva state in which a Qigong master emits his Qi, is the breath regulating state of Qigong and also the state of emitted Qi in the breath holding exercise. Non-Qigong masters who simulate the Valsalva state also send out more intense infrasonic waves. It shows that every person has infrasonic characteristics. A long period of practicing Qigong helps increase the radiative intensity of infrasonic waves. Entering the Valsalva state helps in the emission of Qi. The Qi emitted by masters who adopted Song Jing Gong (relaxed Quiescent pattern) had more intense infrasonic waves (reaching 72 dB). Thus, the mechanism of the emitted Qi released by different exercise patterns is different.

We have found by a series of tests that very able Qigong masters can keep the energy of infrasonic waves at a relatively high level (over 70 dB). So tests on infrasonic waves can be used to screen Qigong masters.

The human body can generate and emit infrasonic waves. As far as acoustics is concerned, the most suitable resonant frequency of human tissues is within the range of infrasonic waves. It shows that the human body easily receives infrasonic waves. Infrasonic waves are a strong, effective part of the emitted Qi because of their quick, long distance transmission, strong penetration, and non-decreasing vibration. It is possible that infrasonic waves themselves transmit the messages between the Qigong masters and the subjects, or serve as a carrier.

[(P<0.01) means that the probability that these test results would occur by chance is less than 1%. This measure of "statistical significance" tells us that the sample size was large enough and causative factors strong enough that the results of this study are reliable.]*

NEURO SCIENCE—NERVE IMPINGEMENT AND EMITTED QI

China Healthways Newsletter's Interview with Liu Guo Long, MD, PhD
Beijing College of Traditional Chinese Medicine
by Richard H. Lee in Palos Verdes, CA, August, 1991

(This interview came about as a result of requests by readers of China Healthways Newsletter for an explanation of neuroscience as it relates to vital energy and healing.)

Dr. Liu provides a fresh view into the role of vital energy in the nervous system and how vertebral nerve impingement affects Qi.]

China Healthways Newsletter: How has your research changed your views on Qi?

Dr. Liu: I started my research with the intention of proving that there was no physiological basis for Qigong, that it was nothing more than a circus trick. As my research progressed, I became more and more convinced that emitted Qi is both measurable and effective for healing. This was not before I eliminated all possible psychological effects by replacing human subjects in my experiments with laboratory animals and got similar results. In my view, emitted Qi is quite real because it is easily measurable using the human brain as a detector.

CHN: How does Qigong affect the function of the nervous system?

Dr. Liu: This is a very complex question because emitted Qi affects different organs quite differently. First, it clearly increases Electro-Encephalograph (EEG) activity throughout the brain and nervous system, principally in the frontal cortex and in the range of alpha, 8 to 14 Hz. This greater mental activity in the alpha range is associated with clear relaxed state of deep thought, which many agree is effective to relieve tensions and clear the mind. The frontal cortex of the cerebrum is at the top of the hierarchy. It is the center of awareness within the body, directing the cerebellum and brain stem. Increased EEG activity in the frontal cortex would indicate expanded awareness.

Secondly, emitted Qi inhibits the sensitivity of the cerebrum to outside stimulation while facilitating the sensitivity of the hypothalamus and medulla of the brainstem. The cerebrum performs high level conscious activities. When inhibited it receives less visual, auditory, and sensory input, and presumably sends fewer and clearer signals to the brain stem, which, in turn, is responsible for automatic functions such as respiration, vasodilation, and the complex functioning of internal organs. Because the brainstem is facilitated, it is more able to "tune-in" to the needs and illnesses of the body and to activate the needed repairs to restore health. I believe that this enhanced awareness of internal stimuli in the brain stem, combined with reduced sensitivity to outside stimuli, is the primary product of emitted Qi that brings about accelerated healing throughout the body.

CHN: Then, would you say that emitted Qi directed only to the central nervous system can enhance healing in all parts of the body?

Dr. Liu: That's right.

CHN: How can Qi applied to the local injury accelerate healing?

Dr. Liu: Neurons in all parts of the body are in constant oscillation, producing measurable output in the range of 8 to 14 Hz, the frequency of the alpha pattern. Neurons can both emit and receive energy in this frequency band. Therefore nerves throughout the body can be facilitated by the emitted Qi. The application of emitted Qi to the site of an injury facilitates the signals to the brain stem. Stronger and more accurate signals to the brain means a more effective healing response directed to the area of injury, including greater vasodilation, increased lymph flow, and an increased supply of nutrients for cell regeneration.

CHN: Is there any direct effect of emitted Qi on the injured tissue.

Dr. Liu: Yes, of course. Qi is the vital energy that drives the regeneration of tissue. Supplementing the available Qi can directly increase the metabolic activity of the injured tissue just as it increases the EEG activity of the brain.

CHN: How can emitted Qi strengthen the immune system against bacteria and viruses?

Dr. Liu: The mechanisms are the same. A facilitated brainstem and nervous system are able to identify the invader more quickly and accurately and to direct the immune system response more precisely to the area of infection. And, as before, the added supply of Qi fuels a more furious response from the immune system. Chinese herbs such as Huang Qi (Astragalus) are also highly effective at enhancing Qi to accelerate the immune response.

CHN: Can mis-position of spinal vertebrae disrupt this healing process?

Dr. Liu: Most definitely. There are basically three types of nerves, the sensory, the motor, and autonomic. Vertebral pressure on any type of nerve cell can inhibit sensitivity or create false signals. If sensory nerves are affected, important information may never reach the brain, or may arrive confused, misdirecting the healing effort. Affected motor nerves can result in spasm or dropsy, and most importantly, impingement of autonomic nerves affects the function of internal organs and if it continues over a long period, can result in serious degeneration or dysfunction.

Often, impingement of motor nerves can perpetuate the problem. As muscles spasm from false motor nerve signals, they may tend to pull the vertebrae further out of alignment, increasing the impingement. In this case, the problem tends to get worse with time as greater impingement further increases the muscle spasm.

The best way to deal with this sort of nerve impingement is through adjusting the position of spinal vertebrae. Often just the vertebral adjustment is sufficient to release pressure on the motor nerve and release the muscle spasm, allowing complete recovery. However, in many cases the impinged nerve is slow to recover, showing some immediate release of spasm but not complete release. The remaining muscle spasm is sufficient to pull the vertebra out of adjustment again and re-establish the motor nerve impingement.

In these cases of muscle spasm aggravating the nerve impingement, better results are achieved if a nerve facilitation technique is used in conjunction with the adjustment. Emitted Qi or the Infratonic QGM applied a few minutes before or after the adjustment facilitates the nerve to provide greater muscle relaxation and quicker regeneration of the nerve. In upper back disturbances the Chinese herbal formula, Hsiao Yao Pien is also excellent to accelerate progress by promoting improved Qi circulation through the motor nerves and through the muscle itself.

INFRASONIC TREATMENT OF ASTHMA

by Su Cheng Wu,
Department of Pediatrics,
The First Affiliated Hospital Guangxi Medical University,
Nanning, Guangxi, China.

(This clinical study finds that infrasonic sound is effective in treating symptoms of childhood bronchial asthma, and in strengthening immune function as measured by T-lymphocyte population.)

Fifty cases of child bronchial asthma were treated with the Infratonic QGM and routine drugs. T lymphocyte subpopulation (CD2 , CD4, CD8, and CD4/CD8) of peripheral blood was measured to evaluate this treatment's clinical effectiveness on raising immunity. The control group of 15 cases was treated with routine drugs only.

TREATMENT METHOD

Based on the features of child bronchial asthma and a dialectical therapeutics of traditional Chinese medicine, the following four points were treated for 5-10 minutes each morning for 5 days:

Dingchuan (located on the back, 0.5 cun lateral to the lower border of the spinous process of the 7th cervical vertebra). Indications: Facilitating the flow of the lung-qi to relieve asthma. Feishu (Located on 1.5 cun lateral to the lower border of the spinous process of the 3rd thoracic vertebra). Indications: Facilitating the flow of the lung-qi to resolve cough. Tiantu (located on the neck, on the anterior midline, in the center of the suprasternal fossa). Indications: Cough, asthma, chest pain and so on. Danzhong (located on the chest, on the anterior midline, on the level of the 4th intercostal space, on the midpoint of the line connecting the two nipples). Indications: Checking upward adverse flow of the lung or the stomach-qi).

OBSERVATION CRITERIA—FOUR GRADES

1. Clinically controlled: Remission from symptoms. Occasional mild asthma attacks are remitted without taking any medicine.
2. Obviously curative effect: Asthma attack is relieved within a short time.
3. Improvement: Some abatement.
4. Inefficacy: Symptoms, signs and reproduction rate of T lymphocyte subpopulation are unimproved or exacerbated.

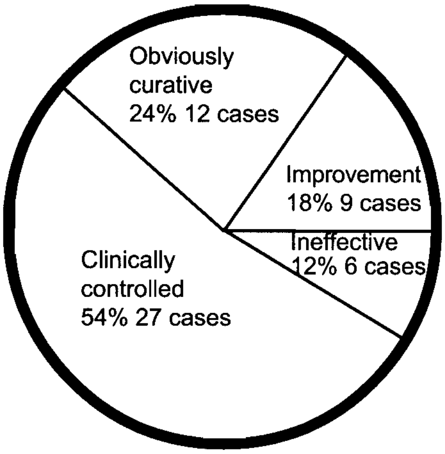

T- lymphocytes in Control vs Experimental Groups

THERAPEUTIC RESULTS

94% of the experimental group showed improvement (as shown to the right). In the control group 80% showed improvement as follows: Clinically controlled - 6 cases; obviously curative - 3 cases, improvement - 3 cases, inefficacy - 3 cases. X^2 determination indicates that there is a more obvious therapeutic effect in the treatment group than that in the control group (P<0.01).

LABORATORY RESULTS

T Lymphocyte subpopulation showed a significantly greater increase in the experimental group than in the control group treated by routine drugs alone as shown below:

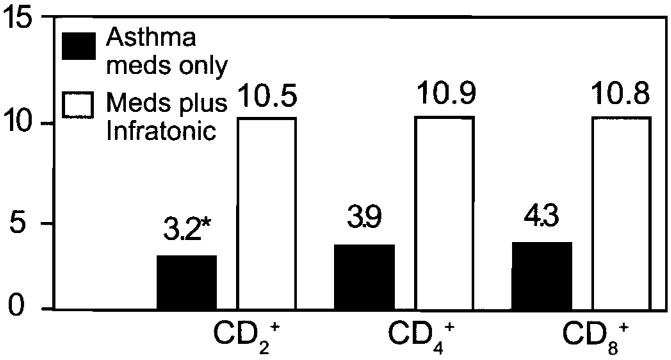

*Notes: Comparison before/after treatment $P<0.01$ unless noted: * = $P>0.05$*

In this study, clinical symptoms of wheezing, coughing and asthma caused by infection of the upper respiratory tract are obviously relieved using the Infratonic device. Testing of peripheral blood showed that the reproduction rate of T lymphocytes subpopulation in treatment group is greater than those receiving routine treatment, indicating increased level of immunologic function of human cells. The Infratonic QGM reduces the course of treatment without discomfort or side-effect. Its biological mechanism of action should be explored in the future.

THE INFRASONIC SIGNAL FROM THE QIGONG PERSPECTIVE

Sun Hua Ling, MD, Beijing Military Qigong Research Institute

(Dr. Sun, a physician and gifted Qigong doctor presents her analysis of the mechanism of emitted Qi and infrasonic signal.)

With the development of human society, health care has become an important part of daily life. Qigong training and techniques and devices like the Infratonic QGM can provide valuable influence. This instrument simulates the low frequency sound of Qigong doctors, which penetrates tissues and organs of the human body, influencing the function of cells, providing clinical results. This study reflects clinical treatment with the QGM of more than 1000 patients.

Cells are units of the human body. Biological electric waves transfer information between cells, forming a biological field of consciousness. The aim of Qigong training is the unification of this field, making the human body more in tune with nature and able to affect nature.

Some "hard Qigong" masters can break stones but can't treat disease because their emitted Qi's amplitude and frequency can not be absorbed by the human body. To be effective, emitted Qi or simulated emitted Qi must be close to the biological waveform of the patient. Then it is easily absorbed and treatment is more likely to be successful. Effective treatment protocols for 30 disease conditions were developed.

MECHANISMS OF THE QGM

1. Relieves pain and edema.
2. Strengthens immune function and reduces inflammation.
3. Normalizes autonomic nervous system.
4. Relieves some symptoms of heart disease.

RELIEVES PAIN AND EDEMA

When applied to sprain of joints, fracture, and arthritis, the QGM has the obvious result of reducing pain distention and edema. It also relieves pain due to sciatic neuralgia, cancer, lumbago, periarthritis of shoulder and stiffness of neck and shoulder muscles. Thus the QGM can effectively treat disease caused by disturbance of blood circulation and can improve local blood circulation.

Patients may feel a sensation of heat, relaxation of muscles, and gradual reduction of pain. This phenomenon may be due to excitation of cells, nerves, and blood vessels around the diseased area. This excitation may contribute to the reduction in pain.

STRENGTHENS IMMUNE FUNCTION AND REDUCES INFLAMMATION

The QGM has a function similar to short wave and ultrashort wave diathermy which strengthens immune function and reduces inflammation. It strengthens metabolism of organs and cells, increasing cellular energy, thus making the immune system stronger.

Diseases such as nephritis, inflammation of intestines and gall bladder, pancreatitis, gas, duodenal ulcer, and infantile dyspepsia, all respond favorably to treatment with the QGM, probably because of this factor of increased cellular energy.

NORMALIZES AUTONOMIC NERVOUS SYSTEM

Where the autonomic nervous system is disturbed, the QGM is often useful. This includes neurasthenia, tenseness of the cerebral cortex, and fatigue. Research shows that the low frequency signals emitted by the QGM penetrate the tissues and organs, altering the internal environment of cells, improving cellular metabolism, and relaxing nerves and muscles, thus regulating the balance between internal and external environment of nerve cells.

Some scholars suggest that emitted Qi may increase cerebral bioelectricity, balancing cell environment, resulting in disappearance of disease. In traditional Chinese medicine, this corresponds to "regulating the meridians and collaterals", "improving blood circulation", and "nourishing internal organs".

RELIEVES SOME SYMPTOMS OF HEART DISEASE

The QGM can relieve symptoms such as lack of blood in the heart muscle, arrhythmia, pressure on the chest, palpitation, and shortness of breath for pulmonary heart disease. This result may be due to improved metabolism of cells and improved local blood circulation.

EMITTED QI TRAINING INCREASES INFRASONIC SOUND EMISSION OF HEALERS

Prof. Lu Yan Fang, National Institute of TV and Electroacoustics, Beijing
Richard H. Lee, Director, China Healthways Institute, USA

(This study supports the premise, proposed by researchers earlier in this chapter, that infrasonic sound can be used as a measure of Qigong ability and indicates that Qigong skills, in this case, the ability to emit infrasonic sound from the palms, can be increased through intensive training.)

Can students in an intensive Qigong training increase the output of infrasonic sound from their hands after one week of training?

To answer this question (and others), 29 Americans traveled to China. The strength of their infrasonic emission was measured before and after the training at the Institute of TV and Electroacoustics, a national institute of technology and standards. Prof. Lu Yan Fang, who conducted the seminal research on infrasonic sound emission two decades ago, conducted the testing. In her research she had found that Qigong masters, powerful healers in China, were able to emit high intensity low frequency sound, 100 to 1000 times stronger than average individuals. Before the experiment she expressed strong doubts as to the possibility of increasing energy emission within one week because Qigong masters require many years to cultivate powerful Qi emission.

Prof. Lu had seen that strong young adolescent males could produce sound as high as 60 dB while average individuals emitted in the 40's and 50's. She was not surprised when in the initial testing 21 of 25 produced scores in the 40's. She felt that an improvement to 60 dB (only one participant had scored in the 60's in the pretest) would suggest improvement to supernormal ability, and that scores above 70 dB would be in the range of the powerful and well known Qigong healers.

A powerful Qigong healer who had scored well in this test before was invited to demonstrate. After several tries, the best he could do was 67 dB, which was 100 times what an average adult could produce, and still in the range of powerful healers, but not nearly as good as he had done previously. He was rather embarrassed and the translator explained that he had worked late the night before and had exhausted his Qi. An hour later he returned covered with perspiration. Apparently, in an effort to redeem himself, he had been building up power doing Qigong exercises for the past hour. This time his score was 78 dB, 10 times the power he had generated just an hour earlier. Participants applauded as he came out of the testing chamber. He had redeemed himself. Interestingly, he was massaging his shoulder as if he had strained it during the Qi emission.

This was a most interesting scenario. First the failure of an experienced Qigong doctor to produce high power, then his performing Qigong exercises to build up his Qi, and finally, his powerful sound emission, offered a persuasive picture that some sort of energy can be stored in the body during Qigong exercises and discharged rapidly through the shoulders to the hands to create intense sound, and possibly straining muscles.

Supporting this finding, one participant was experimenting with the equipment afterwards, emitting Qi into the microphone. At first he could produce high levels of infrasonic sound in front of the microphone, but after five minutes of emitting he could produce only 10% of the power he had produced earlier. He did not feel tired, but as he tried to emit sound, he found that his hand would no longer tremble as it had before. He had apparently exhausted his supply of this form of Qi!

Test Scores: The average score for the group before the training was 47 dB with only four participants scoring above 50. Under the guidance of Master Wan Sujian, a well known Qigong teacher and physician, the group began a 10-day intensive which included climbing mountains and the Great Wall, practicing emitted Qi, and receiving extensive emitted Qi. The group also received physical therapy treatments to open meridians and remove potential blockages in the arms and shoulders which would otherwise restrict the flow of Qi to the hands. The results of both tests are shown below.

After the emitted Qi intensive, the number scoring above 50 dB tripled from four to 13. Five scored above 60 and two above 70. An average increase of 7 dB (to 54 dB) means that students were producing five times the infrasonic energy they had produced before the training. (Statistical significance t < 0.0005) Professor Lu was quite impressed. She felt that five of the participants had both the amplitude and spectral distribution associated with powerful Qigong healers.

An interesting technique: While four of the five participants who scored above 60 were men with strong musculature, one petite woman scored 64 dB, which is 10 times the power that anyone else of similar build had scored. When asked her secret, she answered, "Well, to be honest, as I was waiting to emit Qi I silently asked our teacher for help. As soon as I did, I felt a surge of energy flowing through my arm and my hand started trembling powerfully. When I was done, my arm was sore clear up to the shoulder." She had massaged her shoulder after emitting her Qi much as the Qigong master who first demonstrated had done.

SCIENTIFIC QIGONG RESEARCH

Measuring the emitted Qi from Qigong masters has been the purpose of scientific research in many major scientific research institutes in China,

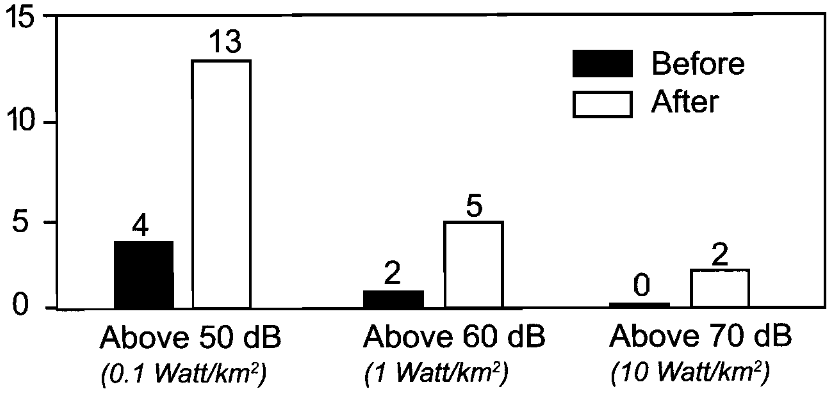

Participants were able to emit substantially higher levels of infrasonic sound after a week-long Qigong training. On the average, scores increased from 47 dB to 54 dB, a five-fold increase in infrasonic energy.

including the following:

- China Immunology Research Center, Beijing
- Institute of Space Medico-Engineering, Beijing
- National Research Institute of Sports Science, Beijing
- Institute of Psychology, Academia Sinica, Beijing
- China Academy of Traditional Chinese Medicine, Beijing
- Institute of Traditional Chinese Medicine, Beijing
- Beijing College of Traditional Chinese Medicine, Beijing
- Tangshan Health Institute for Women and Children, Hebei
- Zhejiang Research Institute of Traditional Chinese Medicine, Hangzhou
- Guiyang College of Traditional Chinese Medicine, Guizhou
- Shanghai Academy of Traditional Chinese Medicine, Shanghai
- The Second Military Medical College, Shanghai
- Shanghai Qigong Institute, Shanghai
- Xing Cheng Air Force Sanitarium
- Zhao Baofeng Beidaihe Qigong Rehabilitation Hospital, Hebei
- Shanghai Institute of Hypertension, Shanghai
- Shenyang Institute of Traditional Chinese Medicine, Shenyang
- Institute for Industrial Health, Xian
- Shandong Medical University, Shandong
- Anti Epidemic Centers of Henan Province, Xin Xiang
- Institute of High Energy Physics, Beijing
- Jiangxi Chinese Medicine and Pharmacy Institute, Jiangxi
- Nanjing Aeronautical Institute, Nanjing
- Institute of Space Medical Engineering, Sichuan
- Chinese Academy of Sciences, Chengdu

While thousands of investigations have been done, just a few have been described here, principally those that are reproducible either with Qigong masters or with Qigong simulation equipment. These studies show that emitted Qi is effective in treating digestive problems, infectious diseases, and bone fractures; that measurable changes in blood chemistry occurs a result of Qigong exercises; and that emitted Qi affects electricity and light in physics experiments.

EFFECTS OF THE EMITTED QI ON HEALING OF EXPERIMENTAL FRACTURE

Jia Lin and Jia Jinding

National Research Institute of Sports Science, Beijing, China.

(This study offers laboratory evidence that Qigong masters can accelerate recovery from fractures. From 1988 FWCAEMQG Proceedings)

The emitted Qi has been found to have a good curative effect on soft tissue injuries such as muscle soreness, scleroma in muscles, acute muscle sprain, muscle contusion and pains. Fracture is also a common injury in sports medicine. We have cured some cases of fracture with the emitted Qi. The purpose of this experiment was to investigate the biological effect of the emitted Qi on healing of fracture.

Sixteen healthy male rabbits, weighing between 1.9 and 2.5 kg, were divided into two groups: the control group and the emitted Qi group. A fracture with a gap of 3 mm was made in the left radius. The rabbits in the emitted Qi group were given the emitted Qi treatment for three minutes per day after fracture. X-ray films were taken every week. Each week specimens for light microscopy from two rabbits were cut into slices of 8*u*m thickness along vertical axis of radius, stained with H.E., and studied by a light microscope. The morphological observations are as follows:

1. Based on some radiographic indexes, such as reaction of fracture section, periosteal reaction and amount of callus formation and callus density, we found that the amount and density of callus formation were better in the emitted Qi group than in the control group. The difference was significant in the 2nd week

(n=16. P<0.01) and 3rd week (n=12. P<0.05).

2. Ultrathin sections were produced with LKB-III ultrotome, stained with uranyl acetate and lead citrate, and studied with a DXB electron microscope (EM). The changes in myofibrils were observed and photographed (x13000). Injured myofibril volume density in the pictures were calculated by means of point-count method.

The ultrastructural examination revealed that overstrain caused pathological changes such as muscle fiber edema, shortening or lengthening of sarcomeres, disorganization, breaking and disappearance of myofibrils as well as Z lines, accompanied by edema and damage of mitochondria. These changes could be seen less frequently in the emitted Qi group than in the control group. The result of quantitative analysis showed that the volume density of myofibrils of the injured muscle in the emitted Qi group was 2.47%, and that in the control group, 20.41 %.

MEASURING QI IN THE BLOOD

A Preliminary Study of the Relationship Between Qigong and Energy Metabolism—The Changes of the Blood ATP Content in Qigong Masters in the Qigong State

by Wang Zhenchang, Huang Jian, and Wu Zijuan
Shanghai Qigong Institute, Shanghai, China

[This revealing study quantifies a physiological change in the blood during Qigong exercise and Qi emission. When Qigong masters do Qigong exercises, blood ATP level increases, indicating a storing of energy. When they emit Qi, blood ATP level drops, indicating a depletion of energy. This is measurable evidence that physiological changes are occurring when a healer says he is collecting energy or emitting energy. It is also evidence that healers may be exhausted after treating patients and need rest and energy building activities to remain in good health themselves. From the 1988 FWCAEMQG Proceedings.]

Recently it has been reported that the emitted Qi by Qigong masters may have material foundation and the internal Qi cultivation can promote health. In this paper, the relationship between Qigong and energy metabolism was studied. The ATP content was measured with 20*ul* of blood drawn from the ring finger of the Qigong masters before and after the Qigong exercise, and, at rest, by the method of bioluminescence assay.

The instrument (FG-30 Type Luminescence Meter) and the chief reagent (luciferase-luciferin) were produced by Shanghai Plant Physiology Institute. The experiments were conducted in three groups:

1. The experimental group: The blood was drawn from each Qigong master's ring finger after he had taken a short rest, and before he exercised Qigong and entered the Qigong state. If he could emit Qi, he would do it for 5-10 minutes. If he exercised internal Qi cultivation, he would do it for half an hour. After the Qigong exercise the blood was drawn again. After rest for half an hour the blood was drawn once more.
2. The control group: The above procedure was repeated on the second day but the Qigong masters did not exercise Qigong.
3. The healthy group: Healthy persons did as the Qigong masters did at the same time. The results are as follows:

In 10 healthy persons the blood ATP contents were constant during the experiment. For nine Qigong masters in the control group the blood ATP contents did not change markedly. The blood ATP contents in 11 Qigong masters after emitting

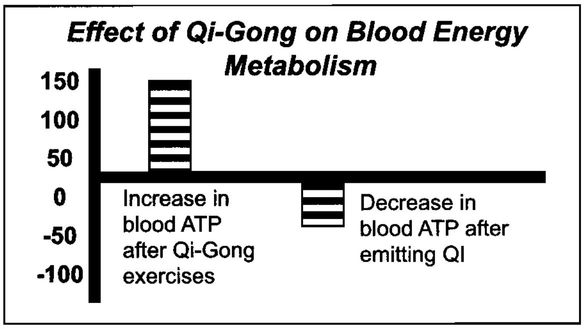

their Qi decreased markedly (the mean decrease was .000131 M compared with that of the healthy group ($P<0.01$). The blood ATP contents in 16 Qigong masters after cultivating the internal Qi increased markedly (the mean increase was .000054 M) compared with that of the healthy and the control group ($P<0.05$). The duration of Qigong practice influenced the changes of the blood ATP content.

The blood ATP contents in both groups of Qigong masters recovered after rest. Qigong masters were observed continually. First they exercised Qigong, cultivating the internal Qi. Then they emitted their Qi. The blood ATP contents changed as above. Conclusions were discussed. The emitted Qi may have a material foundation because the blood ATP contents decreased after the Qi was emitted. The internal Qi cultivation process promotes anabolism and increases CAMP level so that health may be improved.

EFFECTS OF QIGONG ON NEUROTRANSMITTERS

Liu Binghuai, Jiao Jie and Chen Quanzhu
Anhui College of Traditional Chinese Medicine, Anhui Province, China
Li Yefu, Shang Lili (Affiliated Hospital)

[This study shows that you don't need to be a Qigong master to gain healing benefits from Qigong exercises. New Qigong practitioners with chronic illnesses showed significant normalization of neurotransmitter concentration after Qigong practice. From the 1988 First World Conference for Academic Exchange of Medical Qigong (FWCAEMQG)]

It is well accepted that Qigong contributes to preventing and curing of diseases and prolonging life. To study its mechanisms, the authors observed by means of fluorescence spectrophotometry in 68 subjects the variations in blood content of the following monoamine neurotransmitters before and after the Qigong exercises. (5-hydroxytamine: 5-HT, norepinephrine: NE, and dopamine: DA) Forty-four of the subjects were male, 24 females. Their age ranging from 25 to 68 years old, with an average of 48. They all had practiced the Qigong exercise for one month. All were patients of common chronic diseases such as hypertension, coronary heart disease, gastric ulcer, gastritis, pulmonary emphysema, chronic bronchitis, joint system diseases, neurasthenia and neurosis. Results are as follows:

1. A comparison of monoamine neurotransmitter contents in the blood of the subjects pre- and post-exercise showed a general reduction in 5-HT, averaging from 0.42 to 0.21 *ug*/ml, close to the normal value of 0.15 *ug*/ml. The difference was notable ($P<0.001$) Variations in NE and DA tended to go up. The average NE was from 0.27 to 0.35 *ug* /ml, and DA increased from an average of 0.86 to 1.19 *ug* / ml. The above description showed that the effect of the Qigong exercise is closely related to the monoamine neurotransmitters in the body fluid.
2. Effects of the Qigong exercise on the blood monoamine neurotransmitters of patients of different diseases: The subjects were divided into five groups according to the nature of their disease. Subjects in each group showed reduction in blood 5-HT content after they had practiced the Qigong exercise. In all the groups, the content of vasoconstrictor substance 5-HT dropped to the normal concentration level, resulting in sufficient blood supply for the local tissues and organs. It also improved blood flow, promoted metabolism and strengthened the physiological function of the cells in the tissues thus helping to restore health, adding to the efficacy of Qigong.
3. The Qigong exercise helps regulate the neuro-body fluid inner environment. Through regulating the 5-HT blood concentration in the neurotransmitters of parasympathetic nerve, it causes the extra high blood 5-HT of the cardiovascular patients to drop to the normal level. It also dilates the coronary artery and improves the physiological functions, resulting in the relief of symptoms.
4. The Qigong exercise helps build or strengthen

the physiological function of the digestive system. The exercise exerts an influence on the control of nerves that govern the activities of the internal organs. It is shown by variation in the blood content of neurotransmitters NE, DA and 5-HT, with 5-HT getting lower and DA getting higher, facilitating local blood flow and metabolism in the tissues.

5. Through regulating the concentration of pain producer 5-HT in the peripheral blood, the Qigong exercise achieves an analgesic effect. The drop of 5-HT content is relative to the result of analgesia. This fact shows that the drop of 5-HT is good for invigorating blood circulation and reducing extravasation, which is a proof of the TCM principle: "pain is removed when block is eliminated."

The Qigong exercise results in preventing and curing of diseases. It helps regulate the neuro-body fluid inner environment, promotes activities of organs and maintains dynamic equilibrium of the organism through self stabilization.

PHYSICAL CHARACTERISTICS OF EMITTED QI

Gui Yongfan, Chen Qi, Li Yinfa, and Jiang Shen
Nanjing Aeronautical Institute, Nanjing, China

[Emitted Qi can alter the results of common physics experiments shown in this study from 1988 FWCAEMQG Proceedings.]

In this paper, three functions of the emitted Qi which were discovered with the aid of some physical instruments, such as the Van de Graaff generator and light sensitive plate are discussed.

1. The emitted Qi can restrain the discharge process. The discharge between the two balls in the Van de Graaff generator will be stopped or the period of discharge will become longer when the emitted Qi is exerted on the instrument. The Van de Graaff generator was used for testing. The two discharge balls in the generator were pulled apart until the critical state was reached. Then the Qigong master aimed his emitted Qi at the discharge space. The distance between the fingers or palm of the Qigong master and the discharge space was 10 cm to 3 M. We have discovered the following phenomena:
    a. The emitted Qi stopped the discharge or made the period of discharge longer.
    b. The Qigong master had no sensation of getting an electric shock when he emitted his Qi near the discharge space. Nevertheless, when he stopped emitting his Qi, he had the sensation of an electric shock if his hand was still near the discharge space.
    c. In the restraining process, a series of restraint and discharge, such as restraint—intermittent bursts of discharge—restraint, sometimes appeared.
2. The emitted Qi can cause the exposure of light sensitive plates wrapped in a piece of black paper. By false color image processing with a picture processing instrument, we obtained a color photograph, in which the image of fingers and the Laogong (P-8, center of palm) can be seen clearly. The Qigong master emitted his Qi at the emulsion layer of a light sen-

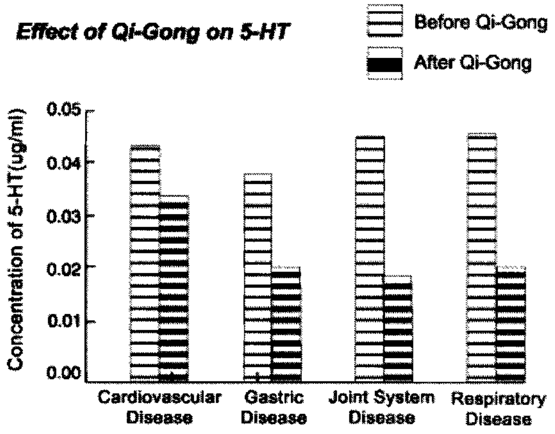

The concentration of the monoamine neurotransmitter 5-hydroxytamine decreased toward the normal value of 0.15 ug/ml in the above groups of patients after one month of Qigong exercises.

sitive plate from about 10 cm away for about 10 minutes.

RESULTS

a) The color of the controls was darker than that of the affected plates. b) After development, white speckles, various patterns or orderly stripes appeared on the affected plates.

SUPERCONDUCTING MAGNETOMETER MEASURES CHANGES DURING QIGONG

Wu Benjie, Dept. of Biophysics, Beijing Medical University
Wang Xiubi, Institute of Somatic Sciences, Beijing,
Wang Zijun, Nanjing University
Liu Jianben, Institute of Geophysics, Beijing

[Here we see that Qigong masters can produce magnetic fields from their palms, and that these fields have characteristic frequency bands. From 1988 FWCAEMQG Proceedings.]

A dynamic study of Qigong (QG) magnetic signals (MS) by a flux-gate magnetometer was conducted in a magnetically shielded room. Under the QG state, MS was <2Hz in frequency and several *n*T (T = Tesla, the unit of magnetic flux density, 1*n*T = 10^{-9} T) in amplitude. For deeper understanding of the MS dynamic properties, human weak MS were investigated by a superconducting biomagnetometer which is wider in frequency response (DC to 10 kHz instead of DC to 10 Hz) and higher in sensitivity. Through a frequency-power spectrum analysis by a computer, the MS changes under QG state were analyzed quantitatively.

Twenty persons (38 tests) were selected as subjects, and divided into two groups. (1) The control group consisting of three persons did not practise Qigong, 46.3 years old in average; (2) The Qigong group (practising for 1-20 years) consisted of 17 persons (31 tests), 48.3 years old in average.

The subjects took off all magnetic things; then came into a magnetically shielded room, with the acupoint to be measured aiming at the probe. The distance between the probe and Laogong (P 8) was about 6 cm, and that between the probe and head acupoints was about 4.5 cm. The MS dynamic changes were recorded continuously, before, during, and after QG state (3 minutes each).

MS was measured by a superconducting biomagnetometer, Model M-601. The frequency power spectra were analyzed by a signal processor, Model 7T17-S. In this paper the power changes were all of relative values.

The experimental results indicate that under the QG state three kinds of MS (1Hz, 3 Hz and <2 kHz) are measured from different acupoints Laogong (P 8), Baihui (Du 20), Yintang (Extra); there are three kinds of MS changes in intensity: increasing, decreasing and no change; and there is a good repetition of the change patterns and signal frequency of MS. Two novel phenomena were found, i.e. decrease of power and the change of frequency spectrum in MS during QG.

QIGONG AND CANCER

Cancer is one area where western science can use some help. Our government has pumped billions into cancer research, yet the problem continues to grow. Chemotherapy and radiation have such negative side effects that the time has come to look beyond therapies of killing "bad" cells, toward therapies with which the individual patient brings his own body into order.

In China, the obvious answer is Qigong. Cancer occurs when the cells of the body follow the will of an outside influence, perhaps a bacteria or a virus or a genetic predisposition. If Qi is strong and flowing smoothly within the individual, many Chinese believe that the cells will do the will of the organism, not of an outside influence.

A very interesting experiment that illustrates the will of an individual influencing cell growth is presented in a study by Dr. Feng Lida, MD, Ph.D., a famous Beijing researcher. She had Qigong masters treat vials of E-coli bacteria. The first vial was held but not treated. The Qigong master emitted Qi to the second one to try to "kill" the bacteria. He directed "health-promoting" Qi to the third one. Dr. Feng reported that, with over 40 repetitions, the health promoting dose resulted in a seven- to tenfold increase in the number of E. coli. The health destroying dose lowered the number of bacteria by one half or more. This experiment illustrates the power of the Qigong master in influencing cell growth. Every one of us has the latent ability to do the same, directing Qi to control disease within our own bodies. This can be seen from the success of programs emphasizing visualization and meditation.

If Qi is weak in an area of the body, abnormal growth will appear. Qi is weakened by poor nutrition and lack of exercise. It is also weakened by mental and environmental stress. Chemotherapy and radiation put heavy additional loads on the body, draining it of Qi. This is why methods to strengthen Qi are so valuable.

This chapter is devoted to Qigong as it is applied to cancer. It is clear from these studies that Qigong is an option that every person trying to beat cancer should consider, regardless of whether western techniques are also pursued. Chemotherapy and radiation in many cases have been shown to be of little value improving quality or length of life. They are not the only option. The following evidence shows that Qigong is an important therapeutic modality for treatment of cancer.

THE EFFECT OF THE EMITTED QI ON CARCINOMA CELLS

Feng Lida, Qian Ju Qing, Chen SuQing, et.al.
China Immunology Research Center, Beijing, China

[This influential paper is the first presented in the 1998 FWCAEMQG Proceedings, and shows that emitted Qi can kill cancer cells and alter chromosomes in vitro, providing clinical evidence that cancer patients would do well to learn Qigong.]

We used the techniques of tissue culture, cytogenics and electron microscopy to study the Hale cells and SCG-7901 human gastric adenocarcinoma cells, the change of the cellular ultrastructure and abnormal structure of the chromosomes of the cells. The results showed:

1. The effect of the emitted Qigong on the Hale Cells:

We repeated the experiment 20 times under identical conditions. After the Hale cells received the emitted Qi for 20 minutes, the survival rate of the cells in the experimental group was 69.28% of that in the control group, i.e. 30.72% of the cells in the experimental group were killed. At the same time we observed the Hale cells with a translation electron microscope, and found that degeneration and swelling took place in some of the cells which received the emitted Qi. The endoplasmic reticula were dilated obviously, and the nuclei of the cells were destroyed.

THE EFFECT OF EMITTED QI ON THE GASTRIC ADENOCARCINOMA CELLS:

After the gastric adenocarcinoma cells received the emitted Qi for one hour, the survival rate of the cells was 74.98% of that in the control group. i.e. the average destruction rate of the cells was 25.02%. We repeated the experiment 41 times under identical conditions, and through statistical analysis, found a difference between the two groups ($P<0.01$). In the meantime, a scanning electron microscope was used to study the cells. The results showed that the microvilli of the cells which received the emitted Qi had remarkably dwindled or disappeared and tiny holes could be seen on the surface of the cells.

3. The effect of the emitted Qi on the chromosomes of the gastric adenocarcinoma cells: We repeated the experiment 10 times under identical conditions. 1420 tumor cells in the control group and 1,428 tumor cells which received the emitted Qi for 60 minutes were analyzed. The subdiploid in the control group was 2.11%, the superdiploid was 0.14%, whereas the subdiploid in the experimental group was 3.29% and supertriploid was 0.56%. The statistical significance in these findings ($P<0.05$).

The structure of the chromosomes of the gastric adenocarcinoma cells also varied. In the experimental group, the total abnormality rate was

5.39%. The total abnormality rate in the control group was 1.40%. ($P<0.01$).

SUMMARY:

Tumor is a common disease which endangers man's life and health. The study on the effects of the emitted Qi on carcinoma cells, especially on the study of the structure and the chromosomes of tumor cells, has not been reported in the world before. The human chromosome is a material base of human heredity and a carrier of the genetic material. The results of our studies showed that the abnormality of the chromosomes of the gastric adenocarcinoma cells was significant after they received emitted Qi. The rate of exchange, breaking, and the dicentromere in the structure of the chromosomes of the gastric adenocarcinoma cells is increased. The results suggested that the emitted Qi could kill the chromosomes of the gastric adenocarcinoma cells. The study may start a new way to cure carcinoma.

QIGONG AS A THERAPEUTIC AID FOR ADVANCED CANCER PATIENTS

by Sun Quizhi and Zhao Li, Kuangan Men Hospital China Academy of Traditional Chinese Medicine, Beijing, China

[This article shows that, in clinical trials, cancer patients were far better off when they combined Qigong exercises and chemotherapy as compared to chemotherapy alone.]

One hundred and twenty three cases of advanced cancer treated simultaneously with drugs and Qigong has been reported in this paper. These were inpatients of our department during the past five years. Among them there were 60 males and 63 females. The youngest was 20 years old. The average age was 47 years old. These patients were diagnosed pathologically as malignant cancer, 70 cases were in the III stage, 53 were in the IV stage, of which, cancer of the stomach accounted for 42.23%, cancer of the large intestine 31.70%, mastocarcinoma 11.29%, and cancer of the esophagus and cardinal carcinoma, 8.1% each.

The 123 patients were divided into two groups. The experimental group included 93 cases treated with drugs and Qigong exercises. The control group, which consisted of 30 cases, was treated by drugs alone. Similar drugs were given to the two groups. Each patient who received the therapy of drugs and Qigong did the Qigong exercises for over two hours a day, three months as a course. Symptoms, signs, body weight, and immunological indices, etc. were recorded before and after treatment. The results were as follows:

1. Amelioration of general symptoms: In the experimental group, 81.7% of the patients regained strength, 63% improved appetite, 33.3% were free from diarrhea or irregular defecation in the experimental group, while in the control group the rates of patients were

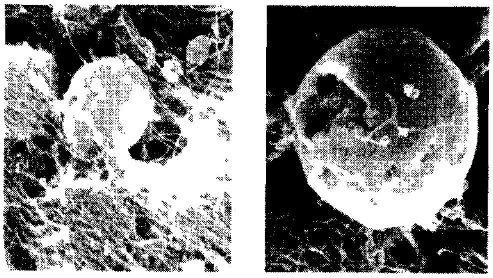

Cells of lung cancer before irradiation with emitted Qi

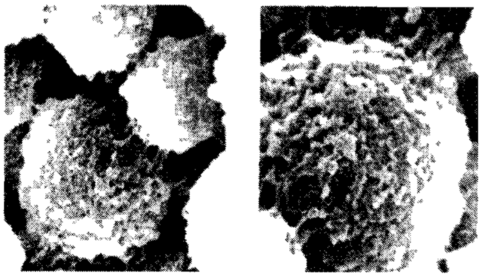

Villus exfoliation of lung cancer cells after irradiation with emitted Qi.

10%, 10% and 6%, respectively. The difference between the two groups is significant. (P<0.01)

2. Body weight: In the experimental group, 50.54% of the patients' body weight increased more than 3 kg, and 5.4% of the patients' body weight decreased more than 3 kg. In the control group, the rates were 13.3% and 30.0%, respectively. The difference between the two groups is, again, quite significant.
3. Phagocytosis of macrophages: In the experimental group, the phagocytic rate of macrophages increased by 12.31 % and the index of macrophages increased from 0.45 before the combined therapy to 0.63 after. In the control group, the phagocytic rate did not elevate, but decreased by 7.87% while the phagocytic index also dropped from 0.63 to 0.50.
4. In addition, it has been observed that, in the experimental group, the erythrocyte sedimentation of 23, and hepatic function of 20 of the 93 patients returned to normal, while in the control group, there were only three patients with normalized erythrocytic sedimentation, and two patients with normalized hepatic function, out of the 30 control cases.

To sum up, these studies suggest that the Qigong therapy is helpful to some extent to ameliorating the symptoms, improving appetite, strengthening constitution, and increasing the ability of self-cure. Qigong can regulate the mind, heart, Qi, and blood so as to get rid of pessimism and eliminate evil factors. Therefore, it is really an effective, simple supplementary treatment with no side effects.

THE EFFECT OF EMITTED QI ON LEUKEMIA IN MICE

Feng Lida and Zhao Xiuzhen
China Immunology Research Center, Beijing, China

[This study by Feng Lida, MD, PhD shows that emitted Qi can decrease the reproduction rate of leukemia cells in mice and that infrasonic Qi simulators can play a similar role. From the 1988 FWCAEMQG proceedings.]

With the rapid development of modern science and technology and the enhancement of people's living standard, the constitutive proportion of diseases has changed and most infectious and parasitic diseases have been brought under control or eliminated, whereas, tumors are becoming a common problem and are endangering man's life and health. In our country, the number of patients suffering from tumor diseases is over 1,000,000 a year, and the number of patients who died from tumors has exceeded 800,000. Therefore, it is important for us to explore a new route to cure tumors.

Leukemia is a malignancy, which has a high incidence and is difficult to cure. As there are some similarities between leukemia in mice and human leukemia, we chose DBA mice as a research model to study the effect of the emitted Qi on L_{1210} cells of leukemia in mice.

The experiment chose DBA mice with a weight of 20±2 gm, and both male and female mice were used. In the experiment we killed the mice, which had been injected with the L_{1210} cells 7 to 9 days before we collected the celiac liquid and adjusted the concentration of cells to 4.70-31.00 million/ml (avg. concentration was 1.20 million/ml). 0.2 ml L_{1210} cells was injected i.p. into the abdominal cavity of mice. And after one day we randomly divided the mice into groups.

One Group Treated With Emitted Qi

Among these groups, the mice of the experimental group received the emitted Qi once per day, for between 10 and 40 minutes for 10 days, while in the control group, no treatment was given. After 10 days the mice were killed and the number of L_{1210} cells were counted with a light microscope. The average value in the control group was 200.435x0.5 million per mouse, while in the experimental group 66.458x0.5 million per mouse and there is significant meaning between the two groups. (P<0.01).

Second Group Treated With Infrasonic Qi Simulator

At the same time, we used an infrasonic qi

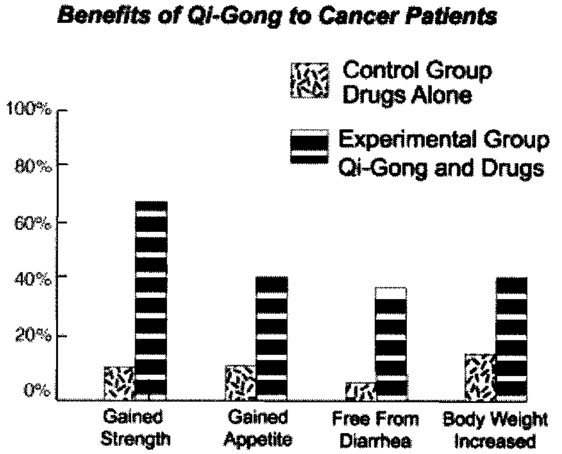

123 advanced cancer patients were treated for 6 months. The control group with drugs, the experiemental with drugs and Qigong exercises. The results show the value of Qigong exercises in cancer treatment.

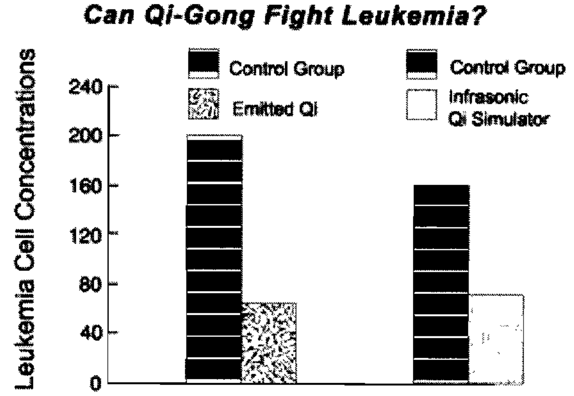

Concentration of leukemia cells in mice after 10 days' treatment with emitted Qi and the infratonic Qi stimulators

simulator to stimulate the mice injected with the L_{1210} cells, once per day, two hours at a time for 10 days, after which the number of L_{1210} cells was observed. The number in the control group was 160.826x0.5 million per mouse, while in the experimental group 70.870x0.5 million per mouse. This difference is of statistical significance ($P<0.1$). The results showed that the number of L_{1210} cells which were injected into mice could be remarkably reduced after the mice received emitted Qi, and the fact suggested that the emitted Qi could kill or inhibit the L_{1210} cells in mice.

Qigong is a part of traditional Chinese medicine, which is a therapy with Chinese characteristics. In recent years, Qigong has been used to cure a lot of diseases, especially difficult diseases and tumor diseases. It is an important issue whether Qigong can kill or inhibit malignant tumor cells or not. The result of this experiment shows that the emitted Qi could kill or inhibit the L_{1210} cells of leukemia in mice. After a body received it many times, the emitted Qi could significantly reduce the number of L_{1210} cells in mice. The experimental study laid a theoretical basis for the treatment of tumors with Qigong. Qigong therapy will be a new way to cure carcinoma. However, the mechanism and way that the emitted Qi kills or inhibits L_{1210} cells of malignant tumor cells in mice needs to be further investigated.

EFFECTS OF QIGONG ON MALIGNANT TUMOR

Luo Sen, Tong Tianmin, et al.
Zhejing Institute of Traditional Chinese Medicine
Hangzhou Qigong Hospital, Hangzhou, China.

[This study, also from the 1988 FWCAEMQG proceedings shows the benefits of combining traditional and modern cancer therapies in maintaining healthy blood cells.]

Under observation in this experiment were 80 patients with 17 kinds of malignant tumors including nasopharyngeal carcinoma, breast cancer, lung cancer, and gastric carcinoma. Their diagnosis was confirmed by cytological and radiological examinations. Within this group, 48 were male and 32 female. Most of them had been operated on before hospitalization. They were at their I or II stage of the disease and had received radiation or chemotherapy before. These patients were randomly divided into three groups. 30 of them, as the first group, received Qigong treatment.

The second group, composed of 25 cases, were treated with routine chemotherapy. And the third group, another 25, received chemotherapy combined with Qigong exercises. The parameters adopted included WBC and RBC count, serum hemoglobin and platelet count in blood before and after treatment, T-lymphocyte conversion test (H3TdRLT) and urine excretion of 17-hydroxy steroids within 24 hours before and after the last week of Qigong therapy in some patients. The total period of observation lasted 60 days. Results are reported as follows:

1. Group I, the Qigong group, developed a significant rise in their WBC, RBC count and serum hemoglobin after treatment ($P<0.01$). Group II, treated with chemotherapy, presented a significant lowering in these measures ($P<0.01$). Group III, which both practiced Qigong and received chemotherapy, showed an obvious elevation of serum hemoglobin, RBC and platelet count ($P<0.005$-0.01). Their WBC remained at the same level as before the treatment ($P>0.05$)
2. T-lymphocyte conversion test in Group I also showed a significant rise ($P<0.05$). It was seen to be maintained at the original level in Group III ($P>0.05$).

This experiment proves that this kind of Qigong exercise, by exciting the circulation of Qi along meridians and modulating the functions of blood circulation, immune response and endocrine tissues, can minimize the toxic side-effects of chemotherapy and enhance energy supply to the body. This practice is effective in treatment of cancer and supported by confirmed scientific experiment.

EFFECTS OF PATIENT TUMORS ON KIRLIAN IMAGES OF HEALTHCARE PROVIDERS

Su Cheng Wu, Guangxi Tumor Hospital, Guangxi, China
Richard H. Lee, China Healthways Institute, USA

The process of emitting Qi during a Qigong treatment, while valuable at jump-starting a patient's recovery, is exhausting and requires extensive Qigong exercises to replenish the depleted Qi. However, as Qigong practitioners become more skilled, they can treat more patients with less depletion, probably through a combination of becoming more proficient at collecting Qi and more efficient at utilizing the Qi that they do have. This is an important field of study because of the prevalence of "burnout" which affects many practitioners in all the medical professions. Perhaps doctors and nurses can be taught the principles of efficient use of their Qi. The following is offered to illustrate this point:

The top two Kirlian photographs on this page show images of fingertips of healthy individuals and healers. The light recorded by the photos is created as electrons moving between the fingers and the film ionize air. In the first photo, the balls are created as electrons leaving the finger create pools of light. And the streamers are created as electrons stream back from the film to the finger. A balance between balls and streamers indicates that electrons move both to and from the fingers. The second photo shows that, around the fingers of healers, the electrons travel more as waves than particles, creating a smooth glow instead of balls and streamers.

In cancer patients, research has shown that balls appear around the cancer (not illustrated, reference Dumitrescu and Kenyon in "Electrographic Imaging in Medicine and Biology", 1983) and streamers appear around all other parts of the body, as shown in the fourth picture. (reference "Bioelectrography" by Konikewicz and Griff 1984 and Dumitrescu, 1983). This strongly suggests that the rapid growth of cancer cells is fueled by high electrical vitality at the cancer site, and that the rest of the body is devitalized as the high energy electrons are somehow attracted by the cancer.

The lower three Kirlian photos were taken in a tumor hospital. A typical cancer patient (the third photograph) shows virtually no Kirlian image. This indicates a very low flow of electrons, either in or out of the finger, and thus, low vitality. After increasing the patient's electrical conductivity using an Infratonic QGM on the foot (K-1), the patient's image brightened, but showed only streamers but no balls, as electrons shot to the fingers but did not return from the fingers to the film. Had high energy electrons returned to the film, we would have seen balls in addition to streamers in the photographs. This suggests that the entire patient, not just the photographed finger, is depleted in electrical vitality, and may be drawing electrical vitality from the Kirlian camera as high voltage electrons stream into the patient's body, but do not return to the film. This supports the common belief that cancer draws vitality from everything in the surrounding area while fueling the very rapid growth of the cancer cells.

Several doctors in the tumor hospital were photographed and those with patient contact showed a similar image dominated by streamers with almost no balls. This suggests that the doctors were in some way devitalized by the patients' low energy condition, and were, themselves, drawing electrical vitality from the camera. It appears that cancer drains the vitality of everyone, the patient, health care providers, family, and even those who visit to show their support.

Why does vitality flow to the cancer site? Qigong practitioners believe that intentionality moves Qi, which suggest that intentionality might be acting. We might suspect the intentionality of the patient, the family or the doctor. All seem to be convinced that the cancer will win. It seems that we should also look at the possibility that the cancer, itself, has intentionality, and is causing the patient to give up hope and the doctor to say things like, "You have no more than 3 to 6 months

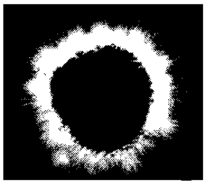

Kirlian image of typical healthy person shows balance between balls and streamers.

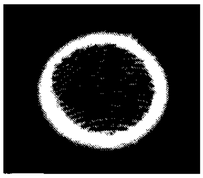

Kirlian image of typical healer

Cancer patient shows low conductivity.

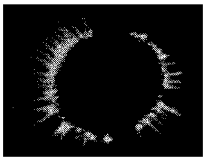

Cancer patient shows increased streamers after QGM treatment.

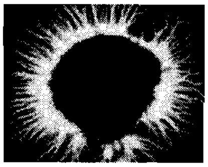

Doctor with patient contact also shows streamer dominance.

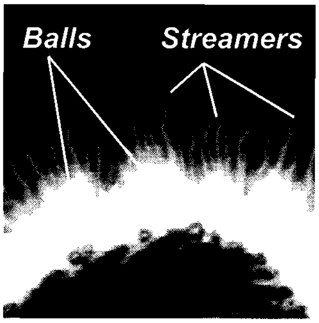

to live." Perhaps the fight against cancer is a battle of wills over the flow of vitality more than simply the biologically understood multiplication of ignorant cancer cells.

Can we reverse this flow of vitality, and will this reverse the growth of cancer? The preceding research in this chapter shows that Qigong practice and emitted Qi can kill and decrease the growth rate of cancer. Perhaps intentionality is a new weapon against cancer. And maybe the Kirlian camera will be the biofeedback technique which shows patients and researchers when their intentionality is properly applied and when they are winning the battle.

As is concluded in most research studies, "More research is necessary."

CHAOS AND CONSCIOUSNESS

The study of Qi as the relationship between the physical world and consciousness is often avoided by scientists because this violates a basic tenet of science, the experimental method, that the experimenter is not to interfere with the experiment. However, this appears to be exactly what Qigong masters are doing. Whereas religious groups have long held that our thoughts, wishes, desires, and prayers can influence our own lives and the lives of those around us, the scientific world has, for the most part, denied this possibility.

If, in our study of Qigong, we discount the conscious intent of the practitioner, we are ignoring what Qigong practitioners claim is the vital ingredient, intentionality. Thus, we are led to study the physics of the phenomenon, the interaction of electric, magnetic, and acoustical fields. However, if we allow intentionality to enter the equation, a very different view of Qigong emerges with enormous implications.

To accept that consciousness can influence the physical world, science needs some sort of field or substance which can be shaped by consciousness such as the rhythmic movement of electrical charges in the body as measured by EEG, which is shown to be influenced by Qigong masters.

From a broader perspective, this field is chaos, which might be described as uncertainty regarding the future, or instability such that subtle influences can guide the field's reorganization. In the case of EEG, the chaos is the available instability of free electrons, molecular ions, and neurochemicals. There are probably many other forms of chaos within the body including magnetic, acoustical, and electromagnetic. It is chaos, or uncertainty in these fields, which allows consciousness to enter and shape reality.

Qigong therapy is often broken into three phases. First, the practitioner "adds Qi" increasing the available energy, or chaos of the patient. Then he "sweeps the bad Qi out the feet" to remove crystallized structures which are interfering with the patient's health. Finally, he "smooths the Qi" to help bring order out of the chaotic state which results from the adding of chaos and removing of "bad Qi."

Thus, a Qigong treatment might be viewed as adding chaos to liquefy a crystallized structure, removing undesired, loosened crystals, then smoothing and harmonizing the field to "bring a higher order out of the chaos." Softening and bending metal, altering EEG in anesthetized animals, and accelerating recovery from cancer might all be explainable by "consciousness creating and guiding chaos."

LOW DIMENSIONAL CHAOS IN PRACTICING QIGONG

Yagu Lin, Jijun Gao, and Huo Lu
Shanghai Qigong Research

[In our research at CHI, we have found that chaos is a key ingredient in the effectiveness of a therapeutic signal, allowing greater penetration into the body. This Chinese chaos research shows that Qigong exercises result in reduced complexity of the EEG in the brain, providing calmness and mental clarity (bringing order out of chaos). From the FWCAEMQ.]

Recently, researchers such as Kaczmark, I.K. (1977), Nicolis, G. (1985), Skarda, C.A. and Freeman, W.J. (1987), and Xu Jinghua and Xu Nan (1987) have reported the systematic complexity and chaos of the brain electroencephalographic techniques (EEG) by the analysis of dimensionality in studies of chaotic dynamics of biological systems. However, there are few reports analyzing the Qigong functional state using this method.

In this paper, we analyze change in attractor dimensionality d values of the occipital EEG before, during, and after Qigong exercises. The results show that before exercises, d=1.8, during exercises d increases to 2.1, and after Qigong exercise, it remains at the increased value of 2.1. These results show that Qigong exercises can lower the complexity under conscious conditions, increase the orderliness, and decrease the randomness. This may explain Qigong's effect of "warding off distraction, thoughts, or evils," and may be a principal mechanism of Qigong's effectiveness.

QIGONG TRAINING AND COHERENCE OF EEG

Yang Sihuan, Yang Qinfei, Shi Jiming, Cao Yi
Institute of Qigong Science,
Beijing College of Traditional Chinese Medicine,
100029, Beijing, China

[This study shows that Qigong training increases coherence between the right and left sides of the brain. From the perspective of chaos theory, this means that Qigong helps to bring "order out of chaos" in the brains of practitioners. From Second World Conference on Academic Exchange of Medical Qigong.]

In this study, we analysed the coherence of EEG to observe the trainees' EEG regularly. The young students, who were 17 to 20 years old, had been practicing "Zhanzhuang Gong" for one year. We try to find out the effects of the Qigong training period on coherence of EEG. Thirty two persons in the Qigong group and 35 persons in the control group were involved in this experiment. During one year period of observation the subjects of the Qigong group practiced Qigong for 40 minutes every day. The EEGs of the Qigong group were analysed every half year in meditation, and the EEGs were also recorded before learning Qigong. The students in the control group did not take part in the Qigong training and their EEGs were investigated at rest twice with an interval of one year. In the test, eight channels of EEGs were simultaneously processed by a computer on line for 20 minutes. The program, "computer evaluation system for the Qigong state" was provided by the "Laboratory of Bio-Control, Department of Electrical Engineering, Zhejiang University".

After one year of Qigong training, total coherence between the left and right frontal regions increased from 0.84±0.07 to 0.87±0.06 ($p<0.05$). Before Qigong training, the total coherence between the left and right occipital areas was 0.68±0.14. After half year's training, it increased to 0.79±0.10, and after a year's training, it was 0.76±0.10. Self comparison showed the probability $p<0.001$. The total coherence between the left and right temporal areas before Qigong training was 0.48±0.17. Half year after Qigong training it was 0.55±0.13, compared with that before Qigong training ($p<0.05$). One year after Qigong training it was 0.64±0.12. Comparing before and a half year after Qigong training $p<0.001$.

Total coherence did not change significantly in the control group

VIBRATIONAL TREMBLING, CHAOS, AND EEG

Dale M. Patterson, MS, CACB, BCIA Certified
Richard H. Lee, China Healthways Institute, USA

[This article, published in China Healthways Newsletter, provides data collected from biofeedback research conducted by Dale Patterson which shows that chaos increases alpha EEG induction.]

Vibrational trembling appears to be a way memory is stored or processed in the physical body. Chaotic sound penetrates this field of vibrational activity, softening this memory and allowing us to more easily let go of fixated thoughts, addictive desires, and compulsive actions.

When someone is full of anger at being blocked from acquiring the object of desire, he may tremble extensively in the theta range of EEG. And when someone is overloaded with the stress of excess thinking (worry) and approaching a nervous breakdown, he will also tremble, but this time in the beta range. This excess vibrational activity consumes the body's Qi and causes the person to become emotionally, mentally, and physically rigid.

Applying chaotic sound breaks up these repeating patterns, often freeing the patient from frustration and worry. As the patient relaxes, capacity to process thoughts and emotions increases, new answers are found, and the increased freedom and flexibility are experienced as increased vitality. Thus, where physical memory inhibits healing, adding a chaotic alpha signal can accelerate recovery.

LABORATORY TESTING

Dale Patterson, a board certified EEG technician trained in biofeedback became intrigued with the QGM when one of his patients brought one in. He recorded EEG brain maps first with an early model of the QGM which produced a clinically significant 18% increase in Alpha. He then used the Q4 with chaos which provided eight times the increase in alpha activity, showing that chaos increases induction of alpha rhythm in the EEG.

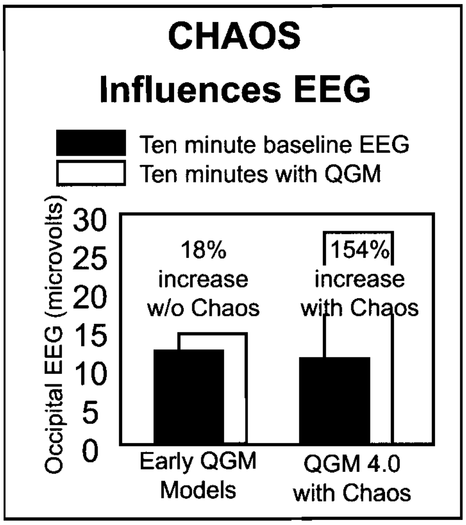

CHAOS AND THE COLLEGE ENTRANCE EXAM

Su Cheng Wu, Guangxi Medical Hospital

[From China Healthways Newsletter: A medical researcher applied the Infratonic QGM to high school students preparing for the college entrance examination.]

There was little chance that Li Wen Chao could ever attend college. As a high school senior, he had achieved only moderate grades in his courses and had done poorly on previous national examinations. Without very high scores in the China College Entrance Examination, Li would almost certainly spend his life in a menial factory or farming job. The College Entrance Examination is terrifying, because everyone who faces it knows that only half of high school graduates will enter the nation's college system. *Failure in this examination is a common cause of suicide in China.*

Enter Su Cheng Wu, a medical researcher and Li's uncle. He had just completed a successful study involving infrasonic treatment of bronchial asthma (see page 28) and was searching for other subjects to study. He felt that stress, nervousness, and mental overload were the principal reasons that high school seniors failed the examination, and believed that the Infratonic QGM would induce the deep calmness and mental clarity of Alpha into the students' minds and bodies, helping them to relax and overcome the mental overload that causes brains to lock up during tests.

For three days before the examination, Dr. Su went to visit his nephew's class and treated each patient by holding the machine on the upper back for five minutes and encouraging the student to talk about fears and tensions regarding the upcoming test. For students suffering from symptoms such as headaches, dizziness, poor appetite, insomnia, and menstrual pain, he instead selected and treated an appropriate point, usually along the front midline of the body.

The results of this small test drew much attention. Li's class was average among the seven senior classes, yet performed the best by far. Whereas the average pass rate for the other six senior classes was 50%, 86% of the students in this class passed the examination and were placed in five year colleges! The other 14% got high enough scores that they were accepted to three year trade schools. Thus, 100% made it to some sort of college. Li Wen Chao enrolled the next year at Guangxi Medical University to become a physician.

While it sounds unreasonable that a little chaotic infrasonic sound can cause such a large increase in scores, the opposing view makes more sense. *Stress and worry can shut down the thinking process.*

SEARCHING FOR JING, QI, AND SHEN IN WESTERN SCIENCE

Traditional Chinese medicine (TCM) holds that three "vital treasures", Jing, Qi, and Shen, are real substances which envelop the human body and are essential to life and healing. These are the fields through which Qigong masters work. Jing provides the genetic structure or the physical strength of the body. Qi is the life or vitality of the body, our energy level. And Shen provides the mind, or conscious aspect. TCM goes further, introducing three seeds of consciousness, the lower, middle and upper Dan Tians located in the lower abdomen, chest, and above the head respectively, which are born and mature through cultivation, providing awareness of the physical world through Jing, Qi, and Shen.

This parallels the writings of Alice Bailey which present a three fold personality: the etheric body (Jing) is shaped by the emotional body (Qi) which, in turn, is shaped by the mental body (Shen). The substance of each body is produced by the physical body whereas the function of each body is to support a mode of consciousness.

In "A New Science of Life" by Rupert Sheldrake a "morphogenetic field" is described as a vibrational field which surrounds and shapes living things as they organize. He argues that our genetic structure doesn't contain, but rather tunes into the blueprint of the physical body through this morphogenetic field. He supports this with laboratory findings that as more and more animals and humans are taught a new behavior, others of similar genetic makeup but physically separated from the first, learn the new behavior more and more quickly. Jing appears to be the TCM equivalent to the morphogenetic field of the physical body. Qi and Shen might also support morphogenetic fields of their own. According to Sheldrake, these morphogenetic fields are programmable, obtaining their initial programming from genetic structure, and being reprogrammed through experience and learning.

Jing, from the standpoint of physics, might be a field of electrons in constant flux suspended in salt water throughout our bodies. The rhythmic movement of this electric field is measurable by EEG, EMG, and EKG. Electrostatic forces are very strong and it is conceivable that such a field can shape a physical body by catalyzing certain chemical reactions and moving certain molecules to the right place at the right time, causing, for example, differentiation between liver cells and brain cells in the embryo.

If Jing is to vibrate freely it requires sufficient energy to have free flowing electrical charges both to resonate with the morphogenetic field and to fuel biochemical reactions. Ilya Prigogine, in "Order Out of Chaos", proposes that life requires chaos, that higher forms of organization emerge out of chaos, and that raising the energy of a system allows it to reconfigure into a higher order of organization. From the standpoint of Jing, when biochemical activity exhausts the available electrochemical potential, chaos in the field of electrons has been exhausted. Then the physical body may become unresponsive, and illness may result. Research during long airline flights shows that, where crowding and recirculated cabin air reduce the availability of electrical energy in the air, people become depleted and develop exhaustion and succeptibility to illness and infection. When passengers wear portable air ionizers, which add electrical instability (chaos) to the air by ionizing it, they do not suffer from exhaustion and avoid illnesses from airborne bacteria. From the standpoint of physics, Jing might be the available electrical instability or chaos in the sea of electrons in the body.

Qi might be a field of magnetic substance which can be measured because it makes the body more electrically conductive. Such diagnostic equipment as Ryodoraku and EAV are claimed to measure the Qi by measuring the electrical conductivity of meridians. High and balanced conductivity are associated with good health and vitality, and low or imbalanced conductivity is associated with illness and debility. While the Qi may have little direct influence on the physical body, its ability to increase local electrical conductivity gives it the ability to provide selective pathways of decreased resistance which allows it to influence the flow of electrons, and thus, shape the Jing. Electrically conductive needles may give acupuncturists a similar ability.

Analysis with Kirlian photography shows that washing the hands with tap water causes the electrical conductivity of the hand to decrease considerably. However, when the hands are washed with the same water which is first passed through a strong magnetic field, the electrical conductivity of the hands increases. (ref: Bioelectric Vitality — The Science of Human Energy by Lee) This tells us, first, that water can store and transport some sort of magnetic substance, and second, that this substance can be removed from or added to the human body. To illustrate this, Kirlian research on long jet flights reveals that reduced magnetic fields on airliners often cause those who have a TCM condition called Yin deficiency to have severe exacerbations of symptoms such as anxiety, nausea, headaches, dizziness, and rising heat. Placing a magnet on the body during the flight provides a source for magnetic substance and reduces the symptoms considerably. Qi might be describable by physics as available magnetic activity within the body, though there appears to be much about magnetism that physics doesn't yet know.

Shen is described in traditional literature as mind or awareness, and often as light. If it is light, why can't we see it? Modern physics tells us that an electron is really an electromagnetic wave, an x-ray, high energy light which has been somehow captured by the nucleus of an atom. Thus, it is conceivable that a field of light or electromagnetic energy could resonate around our physical bodies and remain unmeasurable, just as electrons orbit around a nucleus undetected.

Shen may become visible when excited by electricity. Most people have Kirlian images which show small balls of light surrounding the finger during the negative pulse and streamers, lines of light perpendicular to the finger surface, during the positive pulse. However, healers and artists who use their hands with enhanced awareness, often have a smooth glow around their fingers. (Kirlian photos of the balls and streamers and smooth glow are shown on page 58.) Some even have certain fingers with a smooth glow while others show the balls and streamers. It appears that only those fingers used intelligently have the glow. For instance, a person who principally uses three fingers for massage will find that those three will have the glow while the others have the balls and streamers. The fact that electrons act differently around a healer's fingers is a big opportunity for physicists to investigate the nature of Shen, and to give us a better understanding of the special function abilities of Qigong masters and other healers.

CONCLUSION

Jing, Qi, and Shen are measurable with physical equipment, and thus are subject to scientific study. These measurements correlate with states of health and rates of healing. Thus monitoring the strength of these fields in hospitalized patients and supplementing them when low may prove of value in reducing recovery time and saving lives. In addition to studying, monitoring and supplementing these fields, it might also be productive to study how to increase a patient's resonance with his own morphogenetic field while decreasing his resonance with that of a disease process. This may be a principal area where Qigong masters and other healers excel, while western medicine has little knowledge. Jing, Qi, and Shen, long considered nothing more than useful metaphors by western scientists, may be very real substances which unlock the door for western scientists to understand human vitality.

THE ROLE OF CHAOS IN DISEASE AND HEALING

Richard H. Lee, China Healthways Institute

In the book, "Order Out of Chaos," Ilya Prigogine offers that, when a structured system is energized, its chaos increases, allowing it to reconfigure spontaneously into a higher order of organization. This may explain the evolutionary progression on Earth from the simple structures to the highly organized life forms, from the helpless child to the capable adult who chooses to serve a yet higher organization, humanity.

How can chaos be healing? Chaos is often thought of as unpredictable and destabilizing. It is the enemy of structure and overturns the law of the land. How is it possible that chaos can be harnessed to serve us? As an example: When someone's life is not working, when the basic order upon which they depend is failing them, and when they just want to run away, it may be time to "let go of life" in a way that nurtures the creation of a new order within which the world again makes sense.

Chaos is opportunity. Where there is rigid organization, there is little opportunity for restructuring. The best solutions often go ignored because their implementation would require destruction of some of the structure which holds an organization together. An organization which is flexible, receptive to new ideas, and free to adjust to changing circumstances, is willing to break down old structures for the sake of growth and efficiency. Structure provides for certainty. Chaos injects uncertainty. With a high degree of chaos, an organization can quickly adapt to strange and unfamiliar circumstances without being held back by the inertia of structure.

Within the human body, there is a great deal of structure. Physically we can have unnecessary chronic muscle tension, purposeless lumps, and other structures which don't serve us. Emotionally, we can hold onto desires which repeatedly carry us down pathways of misery. Mentally, we can so fill ourselves with conflicting thoughts that we are unable to sleep and on the edge of nervous breakdown. Whenever a structure within us becomes inflexible, we experience pain. Blocked desire provides emotional pain, and conflicting world views cause mental pain. Flexibility allows us to face the unexpected with capability and enthusiasm. We can run from the path of an approaching car, or stop to pick up a dime. We can eat a bowl of bean sprouts and tofu as comfortably as a hamburger, attend a religious service of any denomination without concern for the clothing of our creator, and be comfortable considering philosophies of strict child rearing alongside newsletters promoting chaos.

Random, unexpected events provide the chaos which exposes our rigidity, our unyielding habits, our crystallized beliefs. Sometimes these crystallized structures within us yield to the pressure but often they resist, sustaining the old ways despite conflicting forces and pain. And sometimes the strain even becomes so great that a person's entire world view shatters. This can be a hugely transformative event:

The initial stress is experienced as pain, uncertainty, or fear of the unknown. We become desperate to avoid further stressing of the crystallized structure so we hang on even tighter. Then the order governing our emotions and thoughts snaps, pops, or shatters. We burst into tears and watch as our life falls to shambles before our eyes, with seemingly nothing left for us to hang onto. We may feel an abrupt shift. Perhaps we experience an immediate release or an exhilarating feeling of freedom and relaxation. At other times, we find ourselves drowning in the unknown, trying to hang onto the structure which just crumbled, unwilling to relax and trust that a new better structure will form in its place. Sometimes we spend years trying to hang onto a crumbled, useless, and painful structure.

Chaos is an opportunity to restructure our beliefs, habits, and physical bodies into a more flexible arrangement, more suitable to the unknown but anticipated future. Whenever we face a new situation, we have this opportunity. Every time we make a new choice or break an old habit, we increase our flexibility and expand our opportunities.

How do we nurture chaos? We start by looking for patterns which restrict our choices, consume our resources, and limit our beliefs. Watching TV, smoking, recounting past experiences or gossip, consuming countless grams of animal fats and alcohol, spending endless hours or repetition in a job or relationship. Changing these patterns may be quite painful because of deep and extensive crystallization, but the result will almost certainly be free time, vitality, new experiences, and an opportunity to make less restricted new choices. While choices made in anger are easiest, and often create more chaos, choices made with love usually provide a greater opportunity for growth. If shattering and reorganization occur within the emotion of anger, the new order will embody anger. If we immediately forgive and "love our enemy", the new order within us will embody love and harmony.

THERAPEUTIC CHAOS

Richard H. Lee, China Healthways Institute

[This article describes how chaotic infrasonic sound might work to penetrate the body, break up stagnation and accelerate healing.]

Over the past ten years CHI has been studying the Infratonic QGM, trying to determine why it is effective at accelerating recovery and decreasing pain. Until recently, our focus has been on the broad spectrum frequency band of infrasonic sound produced by the QGM which infuses the local area, facilitating nerves and inducing Alpha brainwave activity, facilitating nerve transmission between the brain and the local site of pain. This is a plausible mechanism, but there appears to be more to it.

As we experimented chaos surfaced as an "active ingredient." We built prototypes with advanced microprocessor technology which produced an almost purely chaotic signal. In testing we found that these new units penetrated more deeply and stimulated resistant areas of the body which other therapies could not penetrate.

It appears that our thoughts and emotions are strongly affected by sound, and that the body defends itself against outside vibrational signals by anticipating and setting up signals which cancel them out. This may be why someone speaks loudly when he wants to persuade us of an idea, and why we may shout back to defend our own position, and why it is difficult to read or think clearly while sitting in a noisy park in a big city, or in a windstorm, or with loud music playing. It seems that it is difficult to sustain one's own complex thoughts when noise is jumbling them.

Many of the emotions and thoughts we hold within us can be the cause of chronic illness, pain, and slow healing. We might even be carrying some of these patterns from childhood, or pick them up from chance encounters or spooky movies. Like a computer virus, they are lurking among thoughts and beliefs which make up our personality. Those emotions and thoughts which are causing health problems are part of what Qigong therapists call "bad Qi."

Our ability to filter out recognizable sounds, while allowing us to maintain our own thoughts and emotions, is a major barrier to the effectiveness of therapy devices like TENS, ultrasound, and massagers. If we receive a therapeutic treatment from such a device, it may provide some relief but will only affect those surface vibrational patterns which are ready to let go. Thus many therapy devices relax muscles to some degree. But as soon as the therapy begins to intrude on dominant beliefs or emotions we automatically identify the invading signal and filter it out. Thus, predictable therapeutic signals can only do so much.

While we can defend against rhythmic signals, our old attitudes and emotions are defenseless against the random sound of a chaos therapy device because it is unpredictable. It simply goes past our vibrational defenses and softens all thoughts and emotions in the target frequency range. This may be why, after a treatment, people feel calm, clear thinking, and liberated from the problems and worries of life. A big part of what they are feeling may be the abundance of energy which is released when old unnecessary thought and emotion patterns dissolve, releasing the Qi, the substance that was vibrating to maintain them.

It is not just "bad Qi" that is softened by chaotic infrasonic sound. All structures, thoughts and emotions, may be softened as acoustical chaos increases in the body. Since accelerated healing of broken bones and sprained tendons is often reported, chaotic sound may soften these solid physical structures as well. This parallels the first step in Qigong therapy, emitting Qi to soften and liquefy the stagnant Qi.

The second step is to sweep "bad Qi" downward and out the feet. This may be a process of removing old vibrational structures which have been freed to move but have not entirely dissolved. Therapists have found that using chaotic infrasonic sound to sweep from the area of pain down the body and out the feet after treating the pain is often more effective than treating the point of pain alone.

The third step in Qigong therapy is to create a nurturing Qi field to allow the remaining Qi to harmonize, nurturing the emergence of a new order out of the chaos. The Infratonic QGM helps to do this by infusing the body with alpha EEG activity. Whereas the theta rhythm is associated with our individual identity and our relationship with the physical world (and our sense of ego and self importance), and the beta rhythm is associated with planning and thinking (stress and worry), alpha is all about relationships, putting the group before the individual. This may be why, after a QGM treatment, patients become more "compliant" in a doctor's office and people find that interpersonal conflicts are easier to resolve. This process may also occur on the cellular level toward healing and disease resolution as "bad Qi" is removed and the body's cells and organs can reorganize along the lines of teamwork and harmony.

Qigong therapy can be understood in terms of chaos theory. As we let go of old structure, we have an increase of chaotic or undefined energy, which then can transform into a higher order of organization which handles all previous requirements in a new way and has left over capacity for new endeavors.

IMPLICATIONS OF SCIENTIFIC QIGONG RESEARCH

The findings of the scientific research into human vitality have implications in broad areas of science and medicine.

1. Emitted Qi has strong measurable effects in health, healing, physiology and mental functioning. The presented research will encourage scientists around the world to seek deeper understanding of this phenomenon.
2. As human vitality becomes better understood, both in terms of measurable fields which surround the body, and chaos and consciousness, it will be integrated into the American health care system. Monitoring and treating patients accordingly will significantly improve recovery rates and shorten hospital stays.
3. Qigong exercises are effective tools for the individual to cultivate rapid healing and vitality. In addition, many who overcome chronic illness through Qigong practice will develop the ability to help treat others through the emitted Qi.
4. For those who are caught in the frenzied thinking of busy city life, practice of Qigong and other forms of concentration and meditation exercises is an answer. Qigong simulators help to get people started in their relax-

ation and Qigong exercises by relieving pain, promoting deep relaxation and mental clarity, and providing evidence that the phenomenon is real.

5. As the principles of human vitality are accepted in families, children will learn that their feelings of energy flowing through their bodies are normal, and that their awareness of these feelings is valuable. Just as children quickly learn the language that their parents speak, these children will learn to understand and use Qi if their parents accept and discuss it.

The Chinese scientific investigations into Qigong establish the phenomenon of emitted Qi as a real and fertile area of scientific study. Qigong teaching centers and Qigong simulation devices are already making a difference. Further studies into the nature of human vitality and inventions coming out of this ongoing research will do much to shape the 21st century.

APPENDIX 3

INTRODUCTION

This third Appendix includes research information on the physiological and energetic mechanisms triggered within the body, when stimulated through Medical Qigong therapy, Taijiquan or Yoga. This information was collected by my good friend Roger Jahnke, OMD, Director of the Qigong Department at the Santa Barbara College of Oriental Medicine, Santa Barbara, California. Mr. Jahnke is a doctor of Acupuncture and Oriental Medicine, and has studied extensively in China and abroad, learning Traditional Chinese Medicine and refining his clinical ability since 1967.

Doctor Jahnke is also the Director of the Health Action Clinic in Santa Barbara, California, and was one of the founding Board members of the National Qigong Association. He has written a Medical Qigong book entitled, "The Healer Within: Using Traditional Chinese Techniques to Release Your Body's Own Medicine."

This third Appendix is a continuation of the research devoted to the scientific approach of analyzing and scrutinizing the claims made as to the actual effectiveness of Medical Qigong Therapy as a clinical modality. The result on the effectiveness of Medical Qigong Therapy as a solid and effective clinical modality being positive in conclusion.

The following information in this third appendix represents but a fraction of Dr. Jahnke's findings, however, I am confident that the reader will find the data fascinating.

Dr. Jerry Alan Johnson May 2000

THE PSYCHOLOGICAL AND ENERGETIC MECHANISMS TRIGGERED IN THE HUMAN SYSTEM BY THE SELF-APPLIED HEALTH ENHANCEMENT METHODS QIGONG, TAIJI, AND YOGA

Roger Jahnke, OMD
Director of the Qigong Department
Santa Barbara College of Oriental Medicine

Western science has very clearly delineated numerous physiological systems: the digestive system, the nervous system, the cardiac system, etc. However, what might be called the "healing system" has not been discussed or clearly defined until recently. Medicine has uncovered many of the pathological processes that occur within the recognized systems and defined clinical strategies for their resolution. Strategies for enhancing the "healing system," however, have not been addressed by conventional science or medicine.

Numerous cultures, typically considered to be scientifically unsophisticated, actually noted and described the "healing system" long ago. The Chinese in traditional Chinese medicine, and the Asian Indians in Ayurvedic medicine, simultaneously developed elaborate theories of the function of the "healing system," and over thousands of years refined strategies for enhancing its function to produce a remarkable endogenous resource for self-repair. This internally generated healing resource is called an "inner elixir," literally a medicine produced within the human system.

In addition to specific clinical methods such as acupuncture, herbal therapies and tissue manipulation (massage and articular adjustment), these ancient cultures also created comprehensive health enhancement and self-healing systems to assist people in improving their own capacity to heal. Collectively such methods have been called the self-applied health enhancement methods (SAHEM). 1 Typically SAHEM includes breath practice, relaxation and regulation of the mind and nervous system through meditation, purposeful adjustment of the body posture with gentle movements and self-applied massage.

Qigong (also Chi Kung) from China and Yoga from India are ancient methodologies developed and refined over thousands of years specifically for sustaining and enhancing the "healing system." Qigong and Yoga, while they are historically based on the cultivation of factors unknown to the West (China—Qi and India—Prana), actually have specific effects on well known physiological mechanisms that are fully understood by Western science.

While there are many subtle aspects of human physiology that are triggered or accelerated by the practice of the SAHEM (including an emerging new understanding of "energetics," "bio-energetics" or "the electro-physiological system) the following three areas of conventional physiology make up a major portion of the physiological function of the "healing system" and are enhanced by the practice of the SAHEM.

1. Oxygen Delivery
2. Lymph Generation and Propulsion
3. Brain and Neurological Activity

The following sections will address each of these aspects. Other important mechanisms effected by the SAHEM including co-enzymes, antioxidants, connective tissue, precursor hierarchies, stem cell morphology , etc. do not fall into the scope of this paper and their discussion must be held for the future. The domain of subtle energy based mechanisms for the SAHEM effect will be explored briefly in conclusion.

OXYGEN

The Chinese knew about the circulation of the blood approximately 2000 years before William

Harvey described it in 1616 (Temple and Needham, 1986; Veith, 1972)). They knew about the energy generating relationship of food and air 2300 years before the elaboration of the oxidation feature of the Krebs cycle. The simplicity of the ancient Chinese formula for metabolic activity, understandable even by children, encourages wise use breath and food as health enhancing factors This is how Chinese grandmothers have passed health wisdom to children. Note that such simple wisdom is uncommon in the grandmothers of scientific cultures today.

It is often asked what difference there is between Qigong, Yoga (or other mild SAHEM) and the more vigorous conventional fitness methods like running, lifting weights, power biking and aerobic exercise. One way to discuss this is to trace the oxidation and energy cycle. In both aggressive, vigorous exercise and mild exercise the body naturally produces a powerful mix of metabolic resources. In vigorous exercise this is typically called a fuel. In the mild fitness systems of ancient cultures this inner resource has historically been called a medicine, or an elixir.

In vigorous exercise this oxygen induced resource is spent as fuel for hungry muscles. In the milder forms of SAHEM this resource is not completely spent, instead it is circulated internally and utilized as an internal reserve of self-repair factors to sustain and heal the tissues, organs and glands. At the advanced levels of Qigong this endogenous medicine is believed to be transformed into an elixir of longevity and wisdom.

Increased oxygen availability from practice of SAHEM has three potential effects, it:

1. supports energy (ATP, AMP, ADP) generation.
2. generates water as a by product of energy metabolism.
3. enhances immune function.

ENERGY GENERATION

The energy necessary for cellular processes and body heat regulation is supplied through the reaction of oxygen and glucose, in the presence of adenosine triphosphate (ATP). The combination of oxygen from air and sugar from food is catalyzed by ATP releasing chemical energy (ergs) which accomplishes all cellular processes. ATP is the basic energy storage and energy transfer molecule in the body.

The oxygen carrying capacity of the blood and the energy catalyzing ability of ATP are critical factors in the ability of the human organism to sustain a high level of vitality. This suggests that oxygen and ATP may be related to what the Chinese call "Qi" and what the Asian Indians call "Prana," the vital force or life energy. Chinese medicine distinguishes several aspects of Qi. Zhen Qi (Genuine, True or Normal Energy) is the basic dynamic force of all vital function. Ying Qi (Nutrient Energy) is the dynamic force which fuels cellular work (Kaptchuk, 1983). The interaction of oxygen, nutrients and ATP is apparently a physiological process that is a direct correlate of one particular aspect (there are many) of the activity of the Qi.

In Chinese research it was observed that blood levels of ATP increase with practices which cultivate Qi (Wang, 1988). When the mysterious practice of Qi emission (projecting Qi) was performed, the practitioner's ATP levels decreased significantly. Emission of Qi results in a decrease in stored ATP (or potential energy) as recognized by Western physiological science.

The SAHEM activate the body's energy metabolism cycle. The Chinese call this the "cycle of the transformation of energy" (Jahnke, 1989). In the West it is called the Krebs cycle, and has been nicknamed the "bioenergetic cycle."

In the West, ancient medicine is generally characterized as unscientific, superstitious or even primitive. Therefore, the Chinese "formula" for the transformation of energy (Qi) seems overly simplified:

Gu Qi + Kong (Qing) Qi = Zhen Qi

or energy of food + energy of air = essential body energy

Gu Qi (grain Qi, or food Qi) is the essence or life force of food. It mixes with Kong Qi (Kaptchuck, 1983) or Qing Qi (Kendall, 1989) (natural air Qi), the essence or life force of air to form Zhen Qi (Genuine, True or Normal Qi)) which is the life force of the body (Veith, 1972;

Kaptchuck, 1983; Kendall, 1989).

However, it is this same basic formula, hidden in the complex vocabulary of Western science, that is used in modern physiology:

$6O_2 + C_6H_{12}O_6 = ATP = ergs + 6CO_2 + 6H_2O$
air + glucose = catalyst = energy + carbon dioxide + water

It seems that the Chinese knew, without a particularly refined scientific method, that only a portion of the air and food, the Qi or essence, was employed in the process. In fact only 21% of air is oxygen and glucose is approximately 60% of food. (Guyton, 1992)

Science has been exploring energies, beyond metabolic energy (ergs), within the human system including the migration of ions, the ionic gates in cell membranes, the ionic discharge along neurological pathways and the electrical discharge that causes the beating of the heart. The overall understanding of body energetics will increase significantly as science explores the extent to which Qi is linked to these known physio-energetic mechanisms. The interaction of oxygen, glucose and ATP is likely only one energetic aspect of the beneficial effects of Qigong and other SAHEM.

WATER PRODUCTION

A second critical benefit of increased oxygen metabolism is linked to the cellular production of water as a by-product of the energy production cycle noted above. (Shields, 1989) This water is incorporated in interstitial fluid, lymph, blood serum, cerebrospinal fluid, synovial fluid, tears, sweat, etc.

Coincidentally, the energy which fuels the work of the cells as well as much of the water which contributes to the lymphatic fluid and other body fluids have the same source: metabolism. Metabolism is gently increased by all of the SAHEM. In addition, while the metabolic production of ergs also produces water, it is also interstitial water that has been found to conduct the physiologic flow of ions internally.

The benefits of the practice of health enhancement and self-healing methods upon the water component of the endogenous self-repair resource will be further explored in the section below on the lymphatic system. The bionic conductance will be explored briefly under energetic mechanisms.

IMMUNE FUNCTION

ATP drives the activity of every cell. Immune function as well as the production of immune resources (white blood cells, lymphocytes, T-cells, killer cells, etc.) are dependent on oxygen availability. It has been shown that exercise can mobilize the effect of natural killer (NK) cells. (Pedersen, 1988) However, in individuals who exercise so vigorously that they exceed the aerobic level and cross the anaerobic threshold, immune function is actually decreased (Brahmi, 1985; Fehr, 1989).

The scientific evidence is strong on two points: 1) oxygen deficiency leads to decreased immune function and 2) moderate amounts of mild exercise increase immune function.

In his research, Nobel Prize recipient Otto Warburg found that oxygen deficiency was typically supportive of cancer cell proliferation (Warburg, 1966). There are numerous studies that associate reduced lung volume and oxygen exchange capacity with increased mortality and reduced resistance to disease (Gordon, 1970; Cullen, 1983). In studies with elders, immunodeficiency was found to be one consequence of reduced oxygen metabolism. (Saltzman, 1987).

Oxygen's effect on immune function was demonstrated in a German study. It was found that in elderly injured, stressed and hospitalized individuals the arterial oxygen content is often reduced from normal levels. Administration of oxygen was found to elevate the arterial oxygen content and increase recovery rates. Interestingly, the experimental addition of germanium to the treatment protocol in the study further accelerated the healing process (Liem, 1985). Germanium is known to accelerate oxygen metabolism.

These mechanisms associated with oxygen's role in the endogenous "healing system," are naturally occurring and optimal in healthy individuals. One or several of these mechanisms may be deficient in individuals who are ill or at risk for disease. These mechanisms are all enhanced by the gentle movement, focused breath practice and relaxation that occur in the practice of Qigong, Yoga and other SAHEM.

LYMPH SYSTEM FUNCTION

The lymphatic system has been largely neglected in the conventional Western scientific tradition. Contrasted with cardiac function, for example, the lymph function is relatively unexplored. Perhaps, because lymph and lymph vessels are generally translucent, they drew little attention in early anatomical study compared to organs, blood vessels, muscles and bones.

Hippocrates and Aristotle referred to "white blood" and "colorless fluid." In the Middle Ages medical knowledge declined and the lymph was temporarily forgotten. Not until 1627 did Asellius, in Milan, recover the knowledge of the lymph (Rusznyak, 1976). The structure and action of the lymph system was still undefined by 1900 (Yoffey, 1970) and both the immunological function of the lymph and the actual lymphogenic process are still not clearly understood even as we enter the 21st century.

However, in Asia, medical practice is based on meticulous observation of whole organisms—the whole person and the totality of integrated functions. Even though the discrete lymphatic function itself was unknown and unnamed, its effects were generally ascribed to the proper action of the Qi and fluids in China or the Prana and fluids in India. Even though the lymph, as such, was unnamed by Asian physicians, elaborate methods for generating and circulating lymph were developed, and faithfully practiced through Qigong, Yoga and other self-care practices.

When these practices are viewed in relation to what is now known about lymphatic function and its healing role, it appears that much of Qigong, Tai Chi, Yoga and Pranayama were developed specifically with the enhancement of lymphatic function in mind. Breath, movement and postural modification as well as meditation have specific effects on the circulation of the lymph and therefore enhance the overall function of the lymphatic system.

In the West science has divided the body fluids into specific categories: blood, lymph, cerebrospinal fluid, synovial fluid, extracellular fluid, intracellular fluid. From the paradigm of the West it seems that the unsophisticated Chinese may have overlooked important information with their broad, non-specific view of "Qi, blood and body fluids." However, the lymph fluid is largely interstitial biological water. This becomes an aspect of blood plasma (Rusznyak, 1976). Plasma and lymph both contribute water to the cerebrospinal fluid, part of which migrates back into the lymphatic system (Bradbury, 1985). This cycling of fluid (essentially water) from one system into another validates the Chinese approach of thinking of the all various fluids in a single continuum: simply as "body fluids."

Recent discoveries have demonstrated that electrical energy in the form of ion streams migrate throughout the body (Nordenstrom, 1983). The conductance medium for this energy flow system (which will be discussed in the energetic section of this article) is biological water in its numerous contexts including lymph, blood serum, interstitial fluid, etc.

The practice of SAHEM activates a number of remarkable self-healing mechanisms associated with the lymphatic system:

1. Lymph generation
2. Lymph propulsion
3. Immune function
4. Cerebrospinal fluid circulation
5. Nutritive function

LYMPH GENERATION

The actual generation of the lymph has long been attributed to the filtration of blood plasma from the capillaries (Adair and Guyton, 1985). As recently as 1985 this was recognized as the primary source of the lymph. This idea overlooks a significant detail that is at the crux of the tremendous health benefits of Qigong and Yoga: a significant portion of the water in the body's fluids is produced by the identical physiological process that generates the body's biological energy as discussed in the previous section on oxygen (Shields, 1989). Just as an individual respires to sustain life, each cell is also respiring. For each gram of glucose metabolized, in excess of a gram of water is produced.

In a moderately active 70 Kg human, between 2100cc and 2800cc of lymph enters the blood stream daily at the sub-clavian vein through the thoracic duct. In a resting individual the calculations of the Krebs cycle show that the cells produce approximately 950cc (Shields, 1989) of this daily water flow. In a person engaging in moderate exercises, such as Qigong or Yoga (or walking), up to 1400cc of aerobically generated water can be produced daily. When an individual is practicing some sort of moderate body activity, as much as one half of the water that is produced and propelled through the lymph system is a by-product of cellular metabolism. Increased lymph flow improves the elimination of the by-products of metabolism and pathogenic factors (both considered toxins) and increases the circulation of immune factors.

LYMPH PROPULSION

The blood's circulatory system has the powerful heart muscle to propel its fluid. The lymph, however has no distinct heart in humans. It was understood that the movement of lymph against gravity is accomplished by the contractions of the skeletal musculature through vessels with one way valves in a uni-directional system. Nothing new emerged until the mid 1900's when studies revealed that birds and reptiles have specific lymph hearts (Shields, 1980). Even as late as 1941 important aspects of lymph propulsion mechanism remained obscure (Drinker and Yoffey, 1941).

By 1949 "spontaneous, intrinsic pulsatory contraction of the peripheral lymphatic vessels" was demonstrated in humans with a rhythm unassociated with either the heart or the breath (Smith, 1949; Olszewski, 1979).

The current literature is crowded with a wide range of questions raised by research specifically on lymph propulsion: What factors might stimulate the intrinsic contractile mechanism? Is it consistent throughout or differentiated, by what regional factors? By what autonomic and neuroendocrine factors is intrinsic contractility modified? These are just a few of the questions. The study of SAHEM will accelerate our ability to answer these and other more general questions on the lymph system such as: What regulates lymph protein concentration? What effect does passage through the lymph nodes have on the proliferation of immune cells from within the nodes (Olszewski, 1985)? And what is the nutritive role of the lymph (Shields, 1972)?

The breath, through two mechanisms, has a significant effect on the propulsion of the lymph: 1). aerobic production of water and 2). mechanical pumping of the breath apparatus, lungs and diaphragm. Several additional propulsive mechanisms are initiated by body movement and body posture.

Tentative agreement now exists on at least five mechanisms which comprise the "lymph heart" and are activated by the SAHEM, including:

1. aerobic production
2. intrinsic smooth muscle contraction
3. contraction of striated skeletal muscles
4. gravity
5. breath apparatus: lungs, rib cage and diaphragm

Aerobic Propulsion. The aerobic production of water contributes to lymph propulsion mechanically. The cellular production of water, as a by-product of oxygen metabolism, is increased by dynamic Qigong and Yoga practice. The liquid holding capacity of the tissue spaces is naturally limited. As the limit is reached, the presence of additional lymph drives the excess into the terminal lymphatic vessels, the initial vascular openings of the lymphatic system (Yoffey, 1970: Adair, 1985; Shields, 1980, Olszewski, 1985). Like a cup that is running over, the interstitial space fills, building volume and pressure, and then flows over into the lymph vessels.

Intrinsic Smooth Muscle Contraction. The automatic response of the smooth muscle tissue of the lymphatic vessels and the peripheral lymphatic capillaries is to contract when filled and stretched to a certain tolerance (Olszewski, 1985). This contraction moves the lymph forward, with the assistance of the one way valves, in much the same way as the contraction of the heart moves the blood (Bradbury, 1985). In Qigong and Yoga this mechanism is accelerated by the increase in metabolism from gentle movement which contributes to lymph volume.

A number of studies have demonstrated that the intrinsic mechanism is also triggered when the overall sympathetic aspect of the autonomic nervous system is modified by the parasympathetic. Therefore, he intrinsic contractility mechanism is accelerated or enhanced by the shifting into the deeply relaxed state that is a primary feature of SAHEM practice.

Striated Skeletal Muscle Contraction. Even slight contraction of skeletal muscles in mild health enhancement practices propels the lymph in the one way vessels. In the more dynamic practices the pumping of the skeletal muscles and compression of the body parts increases interstitial. The effect of muscle contraction on lymph is one of the classic explanations for lymph motion (Adair, 1985, Bradbury, 1985). In Qigong and Yoga this mechanism is triggered by the mechanical action of the musculature and the action of the skeleton through the relaxation and contraction of the striated muscles in the movement forms of the practice.

Gravitational Propulsion. Because gravity exerts such a substantial force and because lymph has so far to climb to get to the thoracic duct's entry into the sub-clavian vein (at the base of the neck), any inversion of the limbs or even the prone body position allows for a more free flow of lymph. In keeping with this, elevation of the limbs is often prescribed for health problems characterized by a pooling of interstitial fluids, edema.

In SAHEM, many postures and movements create this mechanical dynamic where the lymph is actually propelled by gravity. In certain walking forms, Tai Chi and Guo Lin Qigong are the most renowned, the practitioner is constantly but slowly moving all of the limbs in beautiful circular motions that recurrently activate this mechanism. In Yoga many of the asanas (postures) invert the limbs. In the head and shoulder stands the whole body is inverted.

Breath Apparatus: Mechanical Propulsion. The most powerful of the array of mechanisms that work together to form the "lymph heart" is the mechanical action of the breathing apparatus itself (Sheilds, 1980). The concentration of lymphoid tissue above and below the diaphragm is many times more dense, and contains greater fluid volume, than any of the lymphoid tissue at the periphery, or even in the moderately prolific lymphoid areas of the axilla or groin (Adair, 1985). Lymph that has been carried from all over the body, accumulates centrally and is then propelled by the breath / diaphragm in a final rush through the thoracic lymph duct into the blood at the subclavian vein where it leaves behind its identity as lymph and is transformed into blood serum (Shields, 1981).

Above the diaphragm the thoracic duct of the lymphatic system is a central collecting vessel. Its size is many times that of the peripheral lymph vessels. Below the diaphragm a substantial dilation of the thoracic duct forms a collecting capsule for lymph, called the cisterna chyli. Chyle is a milky fluid of nutrients absorbed from the small intestine, which is passed into the circulating blood through the thoracic duct.

When full inspiration of the breath occurs, the diaphragm drops downward. A tremendous pressure is exerted on the cisterna chyli and a great negative pressure is generated in the thoracic cavity. The compressed cisterna chyli shoots lymph upward through the thoracic cavity. As air rushes in to balance the negative thoracic pressure, the lungs are fully expanded. This creates a powerful positive pressure which compresses the thoracic duct. Due to the one way lymphatic valvular system, lymph is forced upward where it enters into the sub-clavian vein with a rush.

Simultaneously, when the diaphragm drops downward on full inspiration it also compresses the abdominal and pelvic organs. The contents of the lymphoid reservoirs and vessels are forced by the same arrangement of one way valves forward in the system. In research with moving X-ray film, study subjects explored various actions and breath patterns. It was demonstrated that deep inspiration pumps the lymph at a rate that is dramatically greater than resting inspiration and other activities, including massage and walking (Shields, 1981).

It is fascinating to note that in Traditional Chinese Medical theory it has been taught for several millennia that the "lungs regulate the water passages" (Xie Zhu Fan, 1988). To Westerners this seems quite strange. The kidneys are considered to be responsible for water regulation. How could the Chinese have known that the activity of the lungs and diaphragm were linked to moving the body's internal water supply? Even now as we enter the 21st century Western science, and especially average Western citizens, have little idea that the simple act of regulating the breath is such a profound healing tool. However, here we can see that the Chinese somehow arrived at a kind of remarkable wisdom about how the lungs, the breath and the oxygen literally cooperate to produce and circulate the healing resource of water in the body.

IMMUNE FUNCTION

The immunoactive aspect of the lymphatic system is well represented in the conventional literature of Western physiology (Drinker, 1941; Bradbury, 1985; Olszewski, 1985; Van Rooiijen, 1987). The bone marrow, thymus, spleen and lymph nodes participate in the interaction of the lymph and immunity. The composition of the lymph fluid itself includes a number of immune active agents such as lymphocytes and macrophages (Olszewski, 1985).

Lymphocytes that exit with the lymph fluid from the nodes come from three sources: 1) inflowing with lymph from the tissues in the peripheral vessels, 2) exchanged from the blood that enters the node's own vascular system and 3) formed by local proliferation in the node itself (Bradbury, 1985). Lymphocytes naturally collect within the node, especially when lymph flow is sluggish. Greater numbers of cells proliferate when lymph flow is greater and the numbers circulated out of the node increase with flow volume as well (Adair, 1985). Both the numbers immunoactive cells and their circulation rate are increased by the deepening of the breath, relaxation and the alternating contraction and releasing of the musculature in the gentle practices of Qigong and Yoga.

Excellent recent research has clearly delineated, localized and quantified the development of specific antibody forming cells in lymph nodes (Van Rooijen, 1987). In addition, it has been found that there are neurotransmitter receptor sites on lymphocytes where they actually interface with the neuropeptides that drive immune function (Smith, 1985). This is one of the important links between neurochemistry and immunity that is mediated through the lymph system and optimized by the practice of the SAHEM which also increases the numbers of circulating immune cells and accelerates their rate of delivery.

CEREBROSPINAL FLUID (CSF)

The cerebrospinal fluid (CSF) system was classically perceived as a closed system, though another view, unconventional for many years, held that CSF was actually circulating in an open system that allowed the fluid to flow through the aracnoid villi and into the venous blood. By the 1970's it was generally acknowledged that the CSF travels along the cranial and spinal nerves and into the perineural lymphatics (Kimber, 1977).

Research using the microinjection of tracers has suggested several possible pathways for the passage of both the CSF and the cerebral interstitial fluid (CIF) to exit the aracnoid space (Bradbury, 1985). By 1985 the flow of CSF and CIF into the lymphatics was well documented (Bradbury, 1985). Consideration has even been given to the effects of pressure and posture on this flow (Bradbury, 1985), both of which are primary effects that are enhanced in the practice of Qigong, Taiji and Yoga.

The presence of CSF in the lymphatic system and the presence of neurotransmitter receptors on immune cells (Covelli, 1988) suggests a powerful benefit of the SAHEM on the interaction between neurotransmitters and immune function in the reticulo-endothelial system.

NUTRITIVE (TROPHIC) FUNCTION

The importance of a broad availability of nutritional factors to the tissues is fully accepted. However, the role of the lymphatic system in this activity was barely understood before 1972. (Shields, 1972) The original findings of Assellius in dogs revealed the route of absorption of nutrients from the small intestine via the lacteals which are specialized lymphatic ducts. (Adair, 1985) The effect of the breath apparatus through the action of the diaphragm, during deep inspiration, on the cysterna chyli and the small intestine may significantly enhance the rate or effectiveness of nutrient absorption.

The breath gentle body movements are considered especially important in traditional Asian systems of medicine and self-care as techniques for accelerating the circulation and absorption of healing nutritional factors in the diet and in medicinal and tonic herbal formulas. The clarification and understanding of the trophic function of the lymph suggests a simple but profound effect of the SAHEM on both the volume and flow rate of the lymphatic fluid and its nutrifying potential.

These mechanisms associated with lymphatic function play a significant role in the endogenous "healing system." They are naturally occurring and optimally functional in healthy individuals. One or several of these mechanisms may be deficient in individuals who are ill or at risk for disease. Each of these mechanisms are enhanced by the gentle movement, focused breath practice and relaxation that occur in the practice of Qigong, Yoga and other SAHEM.

BRAIN AND NERVOUS SYSTEM FUNCTION

Much of what is described in traditional medical systems as the "balance" of forces or energies, such as yin and yang in the Chinese system, can be associated with the dualistic components of the nervous system. In the central nervous system yin is rest and yang is action. Balance is the state between rest and action called dynamic equilibrium. This is the state that training in Qigong and Yoga seeks to refine. In the autonomic nervous system yin may be associated with the parasympathetic and yang may be associated with the sympathetic. The balance of yin and yang is associated with homeostasis—the state of balanced function and vitality.

Because the Western world view has generally had a difficult time understanding and accepting the concepts of Qi (chi), prana or vital force from the Asian systems, there has been a strong trend toward explaining the effects of Yoga, Qigong, acupuncture, etc. through the mechanisms of the nervous system (Kendall, 1989; Mann, 1973; Melzak and Wall, 1973; NIH, 1997). While these practices do have a definite effect upon neurological function, the neurological mechanism may actually be only an intermediary feature of a more refined and less easily quantified system of dynamic subtle energy interactions. However, much research has been done that reveals the neurological mechanisms activated in Qigong and Yoga, and it is appropriate to explore them here.

Qigong and Yoga and many of the SAHEM enhance several nervous system functions. These include:

1. initiation of the "relaxation response" (RR)
2. shift of the neurotransmitter profile
3. dilation of blood capillaries
4. support of the brain/neurological aspects of immune function
5. balance of right/left brain hemispheres
6. induction of alpha, and sometimes theta waves in EEG.

INITIATION OF THE RELAXATION RESPONSE

When the predominance of autonomic nervous system activity is sympathetic (adrenergic) the human system is working, expending energy and breaking down tissue. This is associated with the action phase of metabolism, and referred to as catabolic. In the extreme, this is the "fight or flight response" with increased heart rate, breath rate and blood pressure. It is also called the stress state and can lead to adrenal exhaustion and collapse (Selye, 1978). This state, when overactive and out of balance causes internal acidosis. Biological stress is conducive to the proliferation of a number of diseases or syndromes including hypertension, pain, depression, immune deficiency and inflammation (Benson, 1975).

The opposite aspect of autonomic activity, parasympathetic (cholinergic), is a phase of rest and tissue regeneration. It is associated with the conservative anabolic phase of metabolism. In its extreme this state is associated with the "relaxation response" (Benson, 1975), characterized by decreased heart and breath rate and a lowering of blood pressure. This is also associated with the resting phase of the basic resting activity cycle (BRAC), (Kleitman, 1967). Conscious deactivation of the sympathetic function with the activation of certain parasympathetic features of autonomic activity can neutralize the negative effects of "fight or flight" overactivity. The primary steps to initiate this state are deep, slow breathing coupled with the intention to relax. (Benson, 1975) These are the identical initiating steps for the practice of Qigong and Yoga. The literature alludes liberally to traditional Asian health maintenance practices as the historic source of techniques for generating the relaxation response (RR) and the typical biofeedback response (Green and Green, 1977).

NEUROTRANSMITTER PROFILE

Much of the new science of psychoneuroimmunology is based upon findings in the area of neurohormones, neuropeptides or neurotransmitters. It has already been mentioned that neurotransmitter receptor sites have been found on lymphocytes. A characteristic profile of neurotransmitters is present in a person who is experiencing pain, anxiety or depression. (Shealy, 1989) Joy, comfort or celebration produce their own neurotransmitter profiles as well. (Ornstein and Sobel, 1987) In patients suffering from pain, increased norepinephrine, reduced cholinesterase and depressed beta endorphin were found to be typical. (Shealy, 1989)

When SAHEM methods are employed that regulate the sympathetic function through the hypothalamus (see below) a neurotransmitter profile emerges which is characterized by decreased norepinephrine, elevated cholinesterase and elevated beta endorphin. (Shealy, 1989) The neurotransmitter profile present in the predominantly parasympathetic and usually more anabolic (alkaline) environment is recognized as able to reduce pain and depression (Shealy, 1989), reduce cravings for addictive substances (Kovacs, 1988) and promote healing.

Chinese research has quantified neurotransmitter activity specific to Qigong exercise. It was found that the Qigong effect, the biological state during Qigong practice, is associated with specific shifts in the monoamine neurotransmitter content of the blood. (Liu, 1988) Serotonin and 5HE generally tend to be decreased by Qigong practice. Noradrenaline and dopamine tend to increase. This demonstrates that the aspects of Qigong and Yoga that quiet the mind and relax the body tend to induce a neurotransmitter profile that is conducive to healing.

INCREASED MICROCIRCULATION

A classic body response in Qigong and Yoga is the elevation of skin temperature. In the fight or flight state, hyper-sympathetic, the arterioles in the skin, muscles and certain organs constrict. During the systematic deactivation of sympathetic function, typical in Qigong and Yoga, vasodilation occurs with the accompanying warmth of the surface of the skin. This is one of the primary goals in biofeedback training and was found as a typical response when the skin temperature of meditators was evaluated in research. (Green, 1977)

A number of studies from China explore the microcirculatory mechanism very thoroughly and conclude that this microcirculatory dilation, increasing blood volume in a particular region, is a major reason for the continued successful application of Qigong and other ancient health maintenance methods (Zhou, 1989; Xiu, 1988; Hong, 1988; Wei, 1988; Zhao, 1988; Chu, 1988).

An ancient saying from traditional Chinese medicine is, "The Qi is the commander (leader) of the blood, and the blood is the mother of the Qi" (Kaptchuk, 1983). It is interesting that the Chinese were aware of a direct relationship between healing energy and blood circulation long ago. They were able, without science as we know it, to develop specific self-healing methods that purposefully increase the circulation of Qi and blood more effectively to the tissues, organs and glands.

BRAIN/NEUROLOGICAL ASPECTS OF IMMUNE FUNCTION

In the classic tradition of Western science, it has been thought that the immune system was an autonomous self-regulating system. A tremendous amount of research now demonstrates that mental and emotional states alter resistance to disease and infection (Ornstein, 1987). Immune organs including the thymus gland, spleen, lymph nodes and bone marrow are invested with nerve tissue (Ornstein, 1987). Lymphocytes and macrophages have receptors for neurochemicals, including catacholamines, prostaglandins, serotonin and endorphin (Roszman, 1985; Ornstein, 1987). Clearly, there is a definite relationship between brain, nervous system function and immune capability—a major interaction of the "healing system."

In the practice of Qigong and Yoga the hypothalamus regulates the autonomic nervous system function to decrease sympathetic activity. (Green, 1977; Benson, 1975) A number of studies have demonstrated that the hypothalamus has an influence over immune function (Ornstein, 1987). The hypothalamus has 40 times more receptor sites than any other part of the brain or nervous system (Pert, 1985 and 1986).

Meditation, progressive relaxation, deep breathing and slow relaxed movement all tend to move the practitioner out of the sympathetic state and induce the relaxation response. Research on the effect of relaxation and visualization sheds some light on the effect that the Qigong and Yoga states may have on immune function. Groups of elders who received relaxation training had significant increases in the activity of "natural killer cells" while control groups did not. Chinese research has corroborated the positive effect of Qigong and other SAHEM on the status of immune function as well (Hong, 1988; Yao, 1989; Xu, 1988; Wang, 1988; Shannahoff-Khalsa, 1988).

BRAIN HEMISPHERE DOMINANCE

Thousands of years ago the Oriental practitioners of self-care disciplines intuitively developed an awareness of an alternating cycle of the predominance of body activity from the right side of the body to the left side. One particular Qigong method, Taiji (Tai Chi), is founded on a constant, flowing of the limbs in circular motions, alternating from right to left. The side of the body that bears the weight is planted, stable, heavy and associated with the Yin. The side that is free to move and kick is active, light, flexible and associated with the Yang. Constant alternation of right and left side activity are thought to balance the forces of yin and yang in the body. Focusing on the right and left sides, alternatively, activates and balances the right and left motor centers in the brain.

In Yogic medicine (Ayurveda), the channels or circuits that conduct the resonating human energy field, are called nadis. Ida nadi and Pingala nadi associate with right and left brain activities (Shannahoff-Khalsa, 1988; Werntz, 1983). This association effects right and left nasal passage activity. These channels alternate dominance in a 2-3 hour cycle, causing the dominant nostril to be clear while the other nostril to swells and becomes congested. (Werntz, 1983) This phenomenon was not noted in the Western world until 1889 when the German physician R. Kayser recorded his observation of the "nasal cycle" (Kayser, 1889). Much of the research on this phenomenon up through the 1980's was motivated by the quest to develop pharmaceuticals for nasal congestion (Shannahoff-Khalsa, 1988).

It has been demonstrated that the nasal cycle is coupled with the alternating lateralization of cerebral hemispheric activity (Werntz, 1983). It was found with research subjects that when a shift occurred in either nasal dominance or brain hemisphere dominance, there was an associated shift, within moments, in the other as well. The right nasal cavity, tends to be more open and the left more congested when the left hemisphere of the brain is more active. This is associated with the active phase of the BRAC and increased general sympathetic tone (Shannahoff-Khalsa, 1988). In contrast the right brain hemisphere is more active when the left nostril is open and dominant and the individual is in the resting phase of the BRAC or the para-sympathetic mode.

A number of different physiological states have been found to be associated with the dominance of one or the other nostril.

- deep sleep is initiated more quickly with left nostril dominance.
- appetite and digestive ability are enhanced during right nostril dominance.
- left nostril, right brain dominance is more conducive to receiving new ideas, while right nostril, left brain dominance is an advantage during discourse.
- it is possible to alter the pattern of "thought waves" by consciously alternating nasal dominance by exercising the congested nostril through forced nostril breathing (Shannahoff-Khalsa, 1989).

A specific Qigong and Yoga breath technique which has been practiced for centuries is the right and left singular nostril breathing. The Salk Institute has done extensive research with this technique, originally prompted by the researcher's work with the Kundalini Yoga tradition. The studies have shown that forcing the breath through the constricted nostril can increase the EEG amplitude of the contralateral hemisphere of the brain (Werntz, 1987). It has been demonstrated that certain diseases and psychopathologies are brain hemisphere specific (Flor-Henry, 1979; McIntyre, 1976). It may be possible, therefore, that the use of alternating single nostril breathing may be applicable as therapy in cases where lateralized dysfunction has been found.

In addition, it was discovered that there is a direct correlation between nasal dominance, brain hemisphere dominance and the lateralized biochemical activity in the peripheral body parts. Recent studies of the nasal cycle comparing plasma catecholamine levels in the venous circulation of the right and left arms found that levels of norepinephrine alternated in direct correlation with the rhythm of sympathetic dominance of the nasal passages (Kennedy, 1986).

INDUCTION OF ALPHA/THETA BRAIN WAVE ACTIVITY

The intention to relax and deepening of the breath are the classic initiating actions that trigger the "relaxation response." Research with practitioners of Yoga (Green, 1977) and Qigong (Collected Proceedings, 1988) has shown that during practice brain wave frequency tends toward the alpha range and in certain cases theta frequency brain activity is achieved.

Alpha level brain function is a result of relaxation and is conducive to healing. The slowing of heart rate, reduction of blood pressure and elevation of skin temperature are common physiological features of the alpha state. Theta is a deeper trance like state that has been found in research with individuals with extraordinary capabilities such as sitting on beds of nails or immediate wound healing without bleeding. (Green, 1977)

In Qigong and Yoga the goal is to bring the lowest achievable frequency of brain wave activity to the practice. In standing Qigong, where there is no body movement, deep states of consciousness with low frequency brain waves are more likely than in moving Qigong. Similarly, in Yoga, there are methods involving movement and methods that primarily involve stillness. The pure meditation state lends more easily to the theta range of brain activity.

Several EEG studies from China have concentrated on the quiescent state—meditation with no movement (Pan, 1988; He, 1988; Wang, 1988). However, the dynamic or moving methods are effective, as well, when the alpha or theta state is achieved with the movement. In both Qigong and Yoga it is a primary focus to "allow the body and energy to sink and relax" and to "relax into the posture."

These neurological mechanisms are associated with the production and circulation of an endogenous healing resource. In healthy individuals they are optimally functional. In those individuals who are ill or at risk for disease, one or several of these functions may be deficient. The gentle movement, focused breath practice and relaxation that occur in the practice of Qigong, Yoga and other SAHEM enhance all of the neurological activities.

ENERGETIC MECHANISMS

As has been shown, the benefits of practicing health enhancement and self-healing methods are significant even though we have little or no understanding of the traditional Asian mechanisms based in Qi and Prana. There is, however, a robust exploration of energetic biophysiology in the Western world that is not new. The biological energy for the work of the cell, called ergs, is fully embraced by conventional science. The electrical discharge in the sinus node of the heart and the ability to track it electrically through EKG has an unquestionable place in science and medicine. It is fully recognized that the movement of the neurological impulse along the nerve pathway includes a discharge of ions. The permeability of the cell membrane is directly linked to the electrical charge associated with the channels of entry and exit In diagnostic medicine the MRI's ability to manipulate the electrons within the cells is clearly based on energetic subtlety that is beyond ionic exchanges.

There is really no question that there has been a profound discussion of energetic physiology that is quite evolved in Western science. As early as 1935 Harold Saxton Burr described a system of electro-dynamic fields (Burr, 1935) from studies using DC current with both salamanders and humans (Burr, 1973). Previous to that numerous European and American scientists and natural philosophers, including Hippocrates, Paracelcus, Van Helmut, Fludd, Stahl, Mesmer, Galvani, Hahnamann discussed a non-chemical basis to life and medicine: pneuma, physis, munia, magnale magnum, magnetic fluid, anima, animal magnetism, life force, odic force, vital force, elan vitale, life field, etheronic force, body electric. This is a much richer history of energetic physiology than most people in the Western world are aware.

Robert Becker confirmed Burr's work and used low frequency DC currents to regenerate salamander tails, and foster healing in wounds and non-healing bone fractures. In his work with the National Institutes of Health (NIH) Becker demonstrated that the perineural (nerve sheath) network is highly conductive. The points of highest conductivity on this network are in the same spot, in many instances, as the major acupuncture points (Becker, 1985). B.E.W. Nordenstrom has delineated the vascular-interstitial closed circuit (VICC) as a system of preferential ion conductance pathways (PICP), comprising a network of biological circuitry (Nordenstrom, 1983) which is nearly identical to the energy channels of traditional Chinese medicine.

It is unlikely that much historic time will pass before the Western scientific community turns full strength on the issues of energetic physiology. Already the notion of inherent electrodynamic fields has gained significant attention. A concept is arising in both China and the west of "bioinformation," which is an intrinsic "informing" capacity across the life continuum, all the way from molecules through mind. (Rubik, 1995) It has been proposed that the energetic mechanism of acupuncture, energy medicine and the inner capacity for self-repair is mediated through an electromagnetic vector field, intrinsic to each living entity. (Liboff, 1997)

Qigong and Yoga with their roots in energy based systems and their obvious benefits in health and healing will only accelerate the movement toward a scientific clarification of the interface between the biochemical and bioenergetic aspects of physiology.

CONCLUSION

Qigong, Taiji, Yoga and various other self-applied health enhancement methods (SAHEM) including Feldenkrais and Alexander methods have demonstrated significant potential for disease resolution, health improvement, health maintenance and performance enhancement. The efficacy of these ancient techniques and the newer innovations can be explained clearly through our conventional scientific understanding of the activation of the body's basic self-regulatory mechanisms. Though the traditional Asian explanations for such efficacy has been based on mysterious and invisible forces, Qi and Prana, the methods are, nonetheless, fully congruent with our prevailing model of physiology.

One or several of the following mechanisms (see chart) are activated during the practice of various forms of SAHEM: enhancement of oxygen availability, cellular water production, lymph propulsion, cerebrospinal fluid motion, acceleration of immune activity, modification of the neu-

Physiological Mechanisms	Structures, Processes, and Substrates	Qigong or Yoga Activity
I. Oxygen		
1. Energy generation	Oxygen uptake	Movement/breath
2. Aerobic water	Oxygen uptake	Movement/breath
3. Immune enhancement	Oxygen uptake	Movement/breath
4. Free radical neutralization	Oxygen uptake	Movement/breath
II. Lymph		
1. Aerobic generation	Oxygen uptake	Movement/breath
2. Propulsion		
a) Aerobic	Oxygen uptake	Movement/breath
b) Intrinsic contraction	Interstitial fluid volume	Breath and relaxation
c) Muscle pump	Muscle contraction	Movement
d) Gravitational	Body position	Postures and movement
e) Breath apparatus	Lungs, diaphragm, cisterna chyli	Breath activity
3. Immune function	Propulsion (2a–e)	Breath, movement, and posture
4. Cerebrospinal fluid flow	Propulsion (2a–e)	Breath, movement, and posture
5. Nutritive function	Propulsion (2a–e)	Breath, movement, and posture
III. Nervous system		
1. Autonomic	Brain, neurochemistry, nervous system	Relaxation and breath
2. Neurotransmitter profile	Hypothalamus	Relax/visualize
3. Microcirculation	Hypothalamus	Relaxation
4. Immunity	Macrophages, lymphocytes, neurotransmitters	Relaxation, meditation
5. Brain hemisphere control	Brain, nervous system	Nostril breath and right/left side movement
6. Brain wave frequency	Brain, nervous system	Meditation
7. Neuroreflex stimulation	Neuroreflex system	Rubbing points

Appendix Figure 3.1. Physiological mechanisms and the activities in Qigong and other self-applied health enhancement methods (SAHEM) which are their triggers:

rotransmitter profile, initiation of the relaxation response, alpha range brain wave frequency, etc. Each of these mechanisms is a component of a powerful internal "healing system." Together they produce and circulate internal healing resources which are historically referred to in Asia as an "inner medicine" or "healer within." In Western terms we might call this an endogenous medicine or a naturally occurring self-repair resource.

It may be that simple SAHEM practices are among the most powerful healing tools ever developed by the human race. It is clear that triggering such an array of self-regulating internal mechanisms has a profound effect on the capability of an individual to maintain health or resolve disease. Given 70% of disease is preventable (DHHS, 1991) and 8 of the 9 major causes of disease are preventable (Fries and Koop, 1993), the SAHEM may be more powerful than post symptomatic medical intervention in reducing disease overall.

Using science to study Qigong, Yoga and acupuncture will very likely help to rewrite the textbooks of medicine and physiology. Direct study of the activity of the physiological states triggered in subjects trained in Qigong or Yoga is an obvious research priority. In light of the crisis in health care costs, it is the mandate of the medical and social sciences to purposefully remove barriers to the implementation of these practices and the study of their benefits.

It was agreed in 1997 that acupuncture had health enhancing potential (NIH, 1997) even though neither its traditional Asian explanation nor its Western scientific explanation were conclusive. Given the conventional scientific understanding of the SAHEM is very much more conclusive than the understanding of acupuncture, it is likely that Qigong and Yoga will be even more easily embraced in the emerging health care delivery paradigm that is based in health improvement. This promises to eventually create a radical transformation in health care and medicine, a major redesign of public health and a significant decrease in the cost of medicine.

For now we have ample evidence to confirm the value of these practices physiologically. Eventually, perhaps, our capacity to explore more subtle realms of human energy will confirm that there is an energetic feature of anatomy and physiology that we do not yet understand. Either way, it is clear that Qigong, Yoga and other SAHEM are practical, timely and inexpensive tools for increasing health and healing disease.

REFERENCES

Adair, T.H., Guyton, A.C., Introduction to the Lymphatic System. Experimental Biology of the Lymphatic Circulation. Johnston (ed), Amsterdam, Elsevier Science Publishers, 1985.

Becker, R. O., The Body Electric: Electromagnetism and the Foundation of Life. New York, William Morrow and Co., 1985.

Benson, H.R., The Relaxation Response, New York, William Morrow and Company, 1975.

Bradbury, M.W.B., Cserr, H.F., Drainage of Cerebral Interstitial Fluid and of Cerebral Spinal Fluid into Lymphatics. Johnston (ed), Biology of the Lymphatic Circulation, Elsevier Science Publishers, 1985.

Brahmi, Z., et al., Effect of Acute Exercise on Natural Killer Cell Activity in Trained and Sedentary Human Subjects, Journal of Clinical Immunology, 5(5):321, 1985.

Burr, H.S., Northrop, F.S.C., The Electro-dynamic Theory of Life. Quarterly Review of Biology, 8: 322-33, 1935.

Burr, H.S., Fields of Life. New York, Ballentine, 1973.

Chu Weizong, et al., Changes of cerebral blood flow and microcirculation under the Qigon state. Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Collected Proceedings, The First World Conference for the Academic Exchange of Medical Qigong. Beijing, China, 1988.

Covelli, V., Jirillo, E., Neuropeptides with Immunoregulatory Functions: Current Status of Investigation. Functional Neurology. 3(3):253-261, 1988.

Cullen, K., et al., Multiple Regression Analysis of Risk Factors for Cardiovascular Disease and Cancer Mortlality in Busselton, West Austrailia: A 13 Year Study, Journal of Chronic Diseases, 36:371-377, 1983.

DHHS, Healthy People 2000. Department of Health and Human Services, 1991.

Drinker, C.K., Yoffey, J.M., Lymphatics, Lymph and Lymphoid Tissue. Cambridge MA, Harvard University Press, 1941.

Fehr, H.G. Human macrophage function and physical exercise: phagocytic and histochemical studies. European J Applied Physiology 1989;58(6):613-7.

Flor-Henry, P., Shizophrenic-like reactions and affective psychoses associated with temporal lobe epilepsy. American Journal of Psychiatry, 126:400-403, 1979.

Fries, J., and Koop, C.E., Refducing Health Care Costs by Reducing the Nedd and Demand for Medical Services. New England Journal of Medicine,

329: 321- 325, 1993.

Gordon, T., Kannel, W.B., The Framingham Study. An Epidemiological Investigation of Cardiovascular Disease. Sections 1-26, Bethesda MD, National Heart and Lung Institute, 1970.

Green, E., Green, A., Beyond Biofeedback. New York, Delacorte Press, 1977.

Guyton, A.C., Medical Physiology, Phila, Saunders and Co., 1992

Hong Shunhua, et al., A primary observation on microcirculation of nail fold and immunogenicity after Qigong practice for a short period. Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Jahnke, R., Qi Gong: Awakening and Mastering the Profound Medicine that Lies Within. American Journal of Acupuncture, 17(2):139-151, 1989.

Kaptchuk, T.J., The Web That Has No Weaver. New York, Congdon and Weed, 1983.

Kayser, R., Die Exakte Messung der Lufterchgangigkeit der Nase, Z Ohrenheilk, 20:96-106, 1889.

Kendall, D.E., Understanding Traditional Energetic Concepts. Energy Fields in Medicine, Proceedings of the Fetzer Foundation1989 Conference coordinated by Michael Morton, 10-43, 1989.

Kennedy, B, Ziegler, M.G., Shannahoff-Khalsa, D.S., Alternating Lateralization of Plasma Catecholamines and Nasal Patency in Humans, Life Sciences, 38:1203-1214, 1986.

Kimber, D.C., Kimber, Gray, Stackpole's Anatomy and Physiology, 17th Edition, New York, Macmillan Publishing Co., 1977.

Kleitman, N., Basic Rest Activity Cycle in Relationship to Sleep and Wakefulness. In Kales, A., Sleep: Physiology and Pathology, Philadelphia, Lippincott, 1967.

Kovacs, G.L., Telegdy, G., Hypothalamo-neurohypophyseal Neuropeptides and Experimental Drug Addiction, Brain Research Bulletin, 20(6):893-895, 1988.

Liboff, A., Bioelectromagnetic Fields and Acupuncture. The Journal of Alternative and Complementary Medicine, 3(Supp 1), S-77—S-87, 1997.

Liem, S. H., The Effect of Sanumgerman (Germanium) on the Oxygen Partial Pressure of the Oxygen Multistep Process. paper presented at the International Conference on Organic Germanium. 1985.

Liu Binghuai, et al., Effects of the qigong exercise on the content of the monoamine neurotransmitters in the blood. Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Mann, F, Acupuncture: The Ancient Chinese Art of Healing and How It Works Scientifically. New York, Random House, 1973.

McIntyre, M., et al., Left and right temporal lobe epileptics: a controlled investigation of some psychological differences. Epilepsia 17:377-86, 1976.

Melzak, R and Wall, The Puzzle of Pain. 1973 Harmondsworth : Penguin Books, 1973.

NIH, National Institutes of Health Consensus Panal on Acupuncture, 1997.

Nordenstrom, B.E.W., Biologically Closed Electrical Circuits: Clinical, experimental and theoretical evidence for an additional circulatory system. Stockholm, Sweden, Nordic Medical Publications, 1983.

Olszewski,W.L., Lymphatic Contractility. New England Journal of Medicine, 300, 316, 1979.

Olszewski, W.L., Peripheral Lymph: Formation and Immune Function. Boca Raton, FL, CRC Press, 1985.

Ornstein, R., Sobel, D., The Healing Brain. New York, Simon and Schuster, 1987.

Pedersen, B.K., et al., Modulation of Natural Killer Cell Activity in the Peripheral Blood by Physical Exercise. Scandanavian Journal of Immunology, 27(6):673-678, 1988.

Pert, C.B., et al., Neuropeptides and their receptors: a psycho somatic network. The Journal of Immunology, 135(2):820-826, 1985.

Pert, C.B., The wisdom of the receptors: neuropeptides, the emotions

and the bodymind. Advances, Institute for the Advancement of Health, 3(3):8-16, 1986.

Rubik, B., The unifying Concept of Information in Acupuncture and Other Energy Medicine Modalities. The Journal of Alternative and Complementary Medicine, 3(Suppl), S-67—76, 1997.

Roszman, T.L., et al., Neuroanatomic and neurotransmitter influences on immune function. The Journal of Immunology, 135(2):769-772, 1985.

Rusznyak, I., Foldi, M., Szabo, G., Lymphatics and Lymph Circulation. London, Pergamon Press, 1976.

Saltzman, R.L., et al., Immunodeficiency of the Elderly. Reviews of Infectious Diseases, 9(6):1127-1139, 1987.

Selye, H., The Stress of Life, New York, McGraw-Hill Book Co., 1978.

Shannahoff-Khalsa, D., A Contemporary View of Life Force Biology: The Merging of Kundalini Yoga and the Neurosciences, in Energy Mediddcine Around the World, ed T.M. Srinivasan, Pheonix, AZ, 1988.

Shannahoff-Khalsa, D., Rhythems of the Mind and Breath, Advances, Journal of Mind-Body Health, 6, #2, 51-55, 1989.

Shealy, C.N., et al., Depression: A Diagnostic Neurochemical Profile and Therapy with Cranial Electrical Stimulation, Journal of Neurological and Orthopedic Medicine and Surgery, 10(4):319-321, 1989.

Shields, J.W., The Trophic Function of the Lymphoid Elements. Springfield, Il., International Foundation of Lymphology, 1972.

Shields, J.W., "Central Lymph Propulsion," Journal of the International Society of Lymphology, 13(1):9-17, 1980.

Shields, J.W., Human Central Lymph Propulsion, Journal of the American Medical Association, 246 (18), 1981.

Shields, J.W., The Aerobic Production and Central Propulsion of Lymph. Proceedings of the 12th International Congress of Lymphology, Tokyo, Japan, September, 1989.

Smith, E.M., et al., Lymphocyte production of endorphins and endorphin-mediated immunoregulatory activity. The Journal of Immunology, 135(2):779-782, 1985.

Smith, R.O., Lymphatic Contractility: A Possible Intrinsic Mechanism of the Lymph Vessels for Transport of Lymph. Journal Of Experimental Medicine, 90, 497, 1949.

Temple, R, Needham J, The Genius of China: 3,000 years of science, discovery and invention, New York, Simon and Schuster, 1986.

Van Rooijen, n., The "In Situ" Immune Response in the Lymph Nodes: A Review. The Anatomical Record, 218(4):359-364, 1987.

Veith, I., The Yellow Emperor's Classic of Internal Medicine (Nei Ching). Berkeley, University of California Press, 1972.

Wang, Zhengchang, et al., A preliminary study of the relationship between Qigong and energy metabolism: the changes of blood ATP content in Qigong masters in the Qigong state. Proceedings of the First World Conference for the Academic Exchange of Medical Qigong, Beijing, China, Pg 58, 1988.

Warburg, O., The Prime Cause and Prevention of Cancer, 1966 Lindau Lecture, Wurzburg, Germany, K. Triltsch.

Wei Shouzhang, et al., An experimental research of the changes of blood volume in the brain and the changes of the heart rate in the Qigong and PSI state. Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Werntz, D., Bickford, R.G., Bloom, F.E., Shannahoff-Khalsa, D.S., Alternating Cerebral Hemispheric Activity and the Lateralization of Autonomic Nervous Function. Human Neurobiology, 2:39-43, 1983.

Werntz, D., Bickford, R.G., Shannahoff-Khalsa, D.S., Selective Hemispheric Stimulation by Unilateral Forced Nostril Breathing. Human Neurobiology, 6:165-171, 1987.

Yoffey, J.M., Courtice, F.C., Lymphatics, Lymph and the Lymphomeyloid Complex. London, Academic Press, 1970.

Xie Zhu Fan, et al., Reference Guide to Acupuncture: Zang Fu Principles and Diagnosis. Wonder Lake, IL, Northern Star Press, 1988.

Xiu Ruijuan, et al., Studies of Qigong effect on the human body by the computerized synchronous system for macro and microcirculatory parameters measurement. Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Zhao Baofeng, et al., Effects of Qigong on cerebral blood flow (CBF) and extremitic blood flow (EBF). Proceedings of the First International Conference for the Academic Exchange of Qigong, Beijing, China, 1988.

Zhou, M.R., Observation of Qigong treatment in 60 cases of pregnancy induced hypertension. Chung Hsi i Chieh Ho Tsa Chih, 9(1):16-18, 1989.

1Note: Taiji (Tai Chi) is a kind of Qigong. Pranayama is an aspect of Yoga. Feldenkrais Method, Alexander Technique and numerous systems of self-initiated health enhancement including puroseful walking and very low impact aerobics are more recently developed self-applied health enhancement methods (SAHEM).

GLOSSARY OF TERMS

A

Abdominal Breathing - to breathe from the abdomen (expand with inhalation, contract with exhalation).

Acupuncture - one of the Four branches of Traditional Chinese Medicine, which involves treating patients through the use of Needling, Cupping, Bloodletting, Moxa and Magnets in order to stimulate energy flow.

Acupuncturist - a doctor of acupuncture therapy.

Acquired Essence - also called Postnatal Jing, is acquired energetic tissue substance, developed after birth.

Acquired Force - energy pertaining to the Earth and surrounding environment.

Acquired Qi - also called Postnatal Qi, is the body's energy derived from food, air and drink, acquired after birth.

Acquired Rational Mind - also called postnatal mind, are thoughts and feelings derived from the acquired experiences of one's environment.

Acute - a rapid onset and short duration of a particular condition.

Adenoma - a neoplasm (abnormal formation of tissue) of glandular epithelium.

Adenocarcinoma - a malignant tumor arising from a glandular organ.

Affirmation - a word, phrase or sentence that is repeated frequently to influence, or change, a belief held deeply within the unconscious mind.

Agoraphobia - fear of places or situations from which escape might be difficult or embarrassing. Symptoms include panic like symptoms and a precondition towards panic attacks.

Akashic Records - a Sanskrit term used to describe the detailed knowledge of all the historical events of the world recorded within the "all-pervasive space of the universe" or "Wuji," also called the "knowledge of the infinite Void."

Alchemy - internal transformation of body and energy brought about by: 1. Nei Dan (Inner Alchemy) through internal Qigong training, and 2. Wai Dan (External Alchemy) using herbal formulas.

Amenorrhea - the absence or suppression of menstruation.

Ancestral Channel - a term used to describe one of the Eight Extraordinary Vessels.

Ancestral Qi - energy inherited from both parents at the time of conception, responsible for innate talents and skills.

Ancestral Traits - pertaining to the ancestral spiritual influence which affects the patient's body. Ancestral traits are developed according to the geographic location at the time of conception.

Anemia - a reduction in the number of circulating red blood cells.

Angina Pectoris - severe pain and a sensation of constriction about the heart.

An Jing - see Hidden Power.

Ankylosing Spondylitis - inflammation of the vertebrae, giving rise to stiffness of the back.

An Mo Therapy - a tissue manipulation therapy that focuses primarily on the treatment of internal organ disorders.

Anorexia - loss of appetite.

An Sound Resonation - Hidden or inaudible sound resonation.

Antibody - any of the numerous proteins produced by the immune system that defend against antigens.

Anxiety - emotional distress, resulting in Heart palpitations, inability to concentrate, muscle tension causing muscle aches.

Aphasia - the absence or impairment of the ability to communicate through speech, writing, or signs, due to a dysfunction within the brain center.

Aplastic Anemia - anemia caused by deficient red cell production, due to bone marrow disorders.

Apoptosis - the disintegration of cells into membrane-bound particles that are then phagocytosed by other cells.

Archetypes - a term coined by the psychiatrist Carl G. Jung to describe the collective unconscious images and motifs (e.g., warrior, healer, priest, whore, etc.). An inherited idea or mode of thought derived from the experiences of the race which is present in the unconscious of the individual.

Arrhythmia - an irregularity or loss of rhythm pertaining to the Heart.

Arteriosclerosis - term pertaining to a number of pathological conditions in which there is a thickening, hardening, and loss of elasticity of the artery walls.

Arthralgia - pain in the joints.

Arthritis - pain and inflammation of the joints, followed by progressive stiffness.

Arthropathy - pertaining to any joint disease.

Ascariasis - infestation of ascaris lumbicoides parasite (pinworm).

Ascending Qi - the action and flow of energy moving upward.

Ashi Points (also called Trigger Points) - places on the body which are tender spots or painful areas near diseased or injured tissue.

Asthma - a disease caused by increased responsiveness of the tracheobronchial tree within the Lungs, due to various stimuli, causing severe difficulty in breathing.

Astral Body - also called the Soul Body, describes the energetic vehicle in which the Eternal Soul can journey outside of the physical body. The Astral Body is connected to the Middle Dantian by a silver "cord of life."

Astral Matter - the energetic substance located within the second field of the body's external Wei Qi and attributed to the emotional energy body.

Astral Plane - an energetic and spiritual plane of existence parallel to the physical plane.

Astral Travel - also called Soul Travel (or Astral Projection) describes the condition of the Eternal Soul journeying outside of the physical body, connected to the Middle Dantian by a silver "cord of life."

Astringent - any substance or agent that causes tissues to contract or that inhibits secretion of Body Fluids such as mucus or Blood.

Antigens - any substance able to provoke an immune response in the human body.

Atrophic - pertaining to atrophy.

Atrophy Syndrome - a disorder characterized by flaccidity and weakness of the limbs and a progressive loss of strength and muscle tone.

Attention Deficit Disorder - a learning disorder manifesting through the following symptoms: habitual failure to pay attention, very distractible, inability to organize, extreme impulsiveness, difficulty in studying, often accompanied by hyperactivity.

Aura - the energetic field which radiates light and circulates around the second field of the body's external Wei Qi

Aura Colors - the body's energetic luminous colors ranging from red, orange, yellow, green, blue, violet to white.

Auspicious Powers - the energy potential contained within the Five Yin Organs.

B

Ba Gan - eight diagnostic principles used in Traditional Chinese Medicine in order to differentiate symptoms.

Bai Dai - white vaginal discharge.

Baihui Point (One Hundred Meetings) - the Governing Vessel point at the top of the head (GV-20).

Bellows Palm Technique - a palm technique in which the thumb and little finger compress and release like a bellows.

Bells Palsy - unilateral facial paralysis of sudden onset.

Benign - gentle or kindly, not aggressive, the opposite of malignant.

Bile - a secretion stored in the Gall Bladder released into the duodenum as a digestive juice.

Bio-Rhythm - three distinct cycles and energy flow that pertain to the body's physical, emotional and intellectual rhythms.

Bipolar Personality Disorder - formerly known as Manic Depressive Personality Disorder, a state of extreme euphoria or pervasive irritability, with racing thoughts, inability to sleep, and impulsive behavior (that may last for days or months), that alternate with morbid depression with suicidal ideation or attempts at suicide (see Depression). During the manic phase there may be hallucinations.

Bird's Bridge - pertaining to the energetic connection between the tongue and the upper "hard" palate, behind the teeth.

Birth Energetic Patterns - pertaining to the energetic patterns developed according to the influence of the time and geographic location of the patient's birth.

Bladder Fire - also called Common Peoples' Fire, or Perineal Fire, is located in the lower abdomen by the perineum, and is responsible for evaporating water.

Blended Originals (Hun Yuan) - the body's "internal combined energy" fused into the Lower Dantian area.

Bloated and Expanded Stagnation - stagnation with a characteristically expansive or bloated appearance, can be caused from an accumulation of Phlegm and Body Fluids (in addition to Qi and Blood) in the adjacent tissue areas of the body.

Blood (Xue) - the dense fluid which nourishes the body, transmits Qi, and provides the material for the mind and emotions.

Blood Heat - a condition categorized by Heat and Blood signs (retching of Blood, expectoration of Blood, Bloody stool or urine, nosebleeds, and menstrual irregularities).

Bloodletting - a technique used in acupuncture therapy which entails pricking the skin to release and remove Blood Heat and Blood stagnation.

Blood Stagnation - the impairment or cessation of normal Blood flow.

Blood Stasis - the same as Blood stagnation.

Blood Vessels - the body's transportation system for Qi and Blood nutrition and regeneration.

Body Fluids (Ye) - these are clear, light, and watery. They originate from food and drink and are transformed and separated by the Spleen (aided by the Kidneys) and dispersed by the Lungs and Triple Burners. (i.e., sweat, tears, saliva, and mucus.)

Body Liquids (Jin)- these are a heavier denser form of Body Fluids. Their function is to moisten the joints, spine, brain, and bone marrow. They lubricate the orifices of the sensory organs.

Bone - the body's skeletal material related in essence to the Kidneys.

Bone Marrow - sustains and nurtures the bones, composed of Kidney Jing (Essence) and Marrow.

Book of Commentaries - consisting of ten commentaries from Confucius and his disciples, pertaining to the study of the eight trigrams, sixty-four hexagrams, and the Yi-Jing.

Book of Oracles - written by King Wen and the Duke of Zhou, pertaining to the study of the eight trigrams, sixty-four hexagrams, and the Yi-Jing.

Borborygmus - a gurgling, rumbling sound heard over the Large Intestine, caused by the passing of gas through the liquid contents of the intestines.

Borderline Personality Disorder - a psychological disorder characterized by a pervasive pattern of intense, unstable relationships, and an unstable self-image. Such patients suffer from chronic feelings of emptiness stemming from abandonment issues. They exhibit self-destructive behavior and transient paranoia, or dissociative symptoms.

Brain - pertaining to the Sea of Marrow issuing from the Kidneys that collects within the cranium.

Brain Tumor Point - Shihmien Point located on the heel of each foot.

Bronchial Asthma - asthma caused by a hypersensitivity to an allergen.

Bronchiectasis - chronic dilatation of a bronchus or bronchi, with a secondary infection (usually involving the lower portion of the Lungs).

Bronchitis - inflammation of mucous membrane of the bronchial tubes.

C

Caduceus - medical insignia picturing double snakes wrapping a winged staff.

Calculi - the plural of calculus or stones, usually composed of mineral salts.

Cancer - an obstruction of Qi and Blood circulation resulting in stagnation and the formation of a malignant tumor that tends to spread.

Carbuncle - a circumscribed inflammation of the skin and deeper tissues.

Carcinogens - any substance or agent that produces or increases the risk of developing cancer.

Carcinoma - a malignant growth or tumor that occurs in the epithelial tissue (the outer surface or first layer of tissue that lines the body's cavities, as well as the principal tubes and passageways leading to the exterior of the body)

Catatonic - totally withdrawn, almost unconscious, frozen and unable to move.

Celestial Stems - see Ten Celestial Stems.

Center Core - pertaining to either 1. The core of light within the Taiji Pole which joins the body's three Dantians through the center of the body, or 2. The True Self (the essence of a person's spirit).

Cerebral Embolism - a condition which occurs when an embolus (bubble of air, or piece of a thrombus) detaches from a thrombus and obstructs a cerebral artery.

Cerebral Hemorrhage - bleeding caused from a rupture of a sclerosed or diseased vessel in the brain.

Cerebral Thrombosis - an obstruction of a cerebral artery by a thrombus (blood clot).

Cerebro-Vascular Accident (CVA) - in the clinic, conditions referred to as stroke (or Wind-stroke) include cerebral hemorrhage, cerebral thrombosis, cerebral embolism, and cerebrovascular spasm. These four conditions are termed in Western Medicine as "Cerebro-Vascular Accident."

Cervical Spondylosis - a degenerative arthritis (osteoarthritis) of the cervical vertebrae and related tissues.

Chakra - an energetic vortex, spiraling out from the body's Taiji Pole, manifesting in the body and extending through the external field of Wei Qi.

Chakra Gates - the twelve major energy gates located on the center line of the anterior and posterior aspects of the body, as well as at the lower perineum and the top of the head.

Chakra System- the seven major Chakra or energy centers that connect to the Taiji Pole. Five Chakras extend to the front and back of the torso, with an additional one located at the top of the head and on the perineum.

Channels - the body's energetic rivers responsible for transporting Qi, also called "meridians."

Channel Points - areas or points (similar to small pools of Qi) found along the streams of a Channel, through which energy of the Yin and Yang (Zang / Fu) organs and channels are transported internally and externally.

Channel Qi - pertaining to the Qi found within the energetic flow of a channel.

Charts of the Hall of Light - ancient diagram depicting the body's internal organs and channels.

Chemotherapy - the application of chemical agents that have a specific and toxic effect upon the disease-causing microorganism, as well as the patient's tissues, energetic fields and immune system.

Chicken Pecking Palm Technique - a palm technique wherein the doctor's hand resembles the head of a chicken while manipulating the energy flow of the Invisible Needle.

Chi Dai - red vaginal discharge.

Child Element - pertaining to the Five Element Creative Cycle of Traditional Chinese Medical, the primary organ is considered the Mother and its sequential organ is considered the Child.

Chinese Massage and Tissue Regulation - one of the four branches of Traditional Chinese Medicine, which involves treating patients through the use of tissue manipulation, including Jie Gu Therapy for bone disorders, Tui Na Therapy for muscle disorders, Gua Sha Therapy for febrile diseases, An Mo therapy for internal organ disorders, and Jing Point Therapy for channel and internal organ regulation.

Cholecystitis - inflammation of the Gall Bladder.

Cholelithiasis - formation of calcium, i.e., bile stones in the Gall Bladder.

Cholestasis - an infection of the biliary tract.

Chronic - the long duration of a specific disease or condition, showing slow or little improvement.

Chronic Fatigue Syndrome - debilitating fatigue that is not the result of physical or mental exertion, and does not diminish by resting.

Chrono-biology - the science that deals with the study of the body's biological clocks and fluctuations in accordance with the cycles of the sun, moon and nature's rhythms.

Chyluria - the passing of fat globules in the urine.

Circle of Willis - the union of the anterior and posterior cerebral arteries, forming an anastomosis at the base of the brain.

Circulating Energy Technique - this method refers to rotating Qi in a circular pattern (clockwise or counterclockwise) to move Qi and Blood stagnation, or to gather energy for tonification.

Cirrhosis - Cirrhosis is a generalized Liver disease marked by hepatic lesions, characterized by the formation of dense lobular connective tissue, degenerative changes in the parenchymal cells, structural alterations in the Liver lobules, and sometimes fatty and cellular infiltration within the Liver.

Clairaudience - the ability to hear sounds, music, and voices not audible to normal hearing (for example, receiving Messages and / or inspirations from the divine).

Clairasentience - the ability to perceive smells, taste, touch, emotions and physical sensations that contribute to an overall psychic and intuitive impression.

Clairvoyance - the ability to perceive current objects, events and / or people that may not be discerned through the body's normal senses. Both time and space are perceived on a clairvoyant spiritual dimension.

Clean Qi - Energy which has been purified.

Clinical Ethics - the moral principles and standards governing the doctor's conduct with patients in or away from the clinic.

Coccyx Pass (Wei Lu Guan) - is located on the lowest segment of the spine just posterior to the anus, near the Chang Qiang (GV-1) point.

Colitis - the inflammation of the colon.

Cold Constitution - pertaining to a physical body innately prone towards coldness.

Collapsed Qi - this is considered a subcategory of Deficient Qi, and is regarded as the third and most severe type of Deficiency.

Collaterals (Luo) - the body's energetic streams that branch off the Twelve Primary Channels and the Conceptional and Governing Vessels.

Collective Unconsciousness - concept of psychiatrist Carl G. Jung pertaining to the memories of mental patterns that are experienced and shared by all mankind.

Coma - an abnormal deep state of unconsciousness with some possible awareness of surroundings, but a total inability to communicate with the outside environment. Comas result from a Qi obstruction to the brain caused by illness or injury.

Common People's Fire - pertaining to the Bladder Fire, located in the perineal area of the body.

Compressed Stagnation - stagnation caused by the patient's energy pressing downward (externally moving inward), feels armored and hollow.

Concentrative Meditation - keeping the mind focused and under control.

Congealed - when Qi and/or Blood become thick and solid within the body's tissues, energetic fields, or both.

Congenital Qi (Prenatal Qi) - energy existing before the fetus is born, acquired from the mother's, father's energies, as well as from the environmental and universal energies. It is sustained through prayer, meditation and sleep.

Conscious Mind - acquired mental reasoning, created through learning via the five senses and interactions with other people.

Constipation - difficulty or infrequent defecation, with the passage of unduly hard and dry fecal material.

Constitutions - see Five Elemental Constitutions.

Contracted Stagnation - stagnation caused by the patient's energy pulling inward, feels armored and solid.

Contraindications - any symptom or circumstance indicating the inappropriateness of a form of treatment that would be otherwise advisable.

Controlling Cycle - pertaining to the Five Element Cycle, where one organ controls the second elemental organ in the Five Elements' Circle.

Corporeal Soul - see Po.

Coryza - the inflammation of the respiratory mucous membranes.

Cosmology - the study of the universe.

Countertransference - this is the process whereby a doctor loses objectivity and unconsciously projects feelings, thoughts, beliefs and patterns of behavior onto the patient.

Couple Point - the Master Point's secondary point of connection affecting the Eight Extraordinary Vessels.

Cranio-Sacral Rhythm - Western term given to the fluctuating rhythm of the Sea of Marrow flowing from the Kidneys.

Creative Cycle - pertaining to the Five Element Cycle, where one organ creates the energy for the next.

Creative Subconscious Mind - the part of the mind that maintains the patient's reality by making him or her act like the person they believe themselves to be.

Creative Visualization - the process of using visualization as a tool for transforming energy and spirit.

Crown Center - the name given to the Baihui Point (GV-20) and Upper Chakra Gate, at the top of the head.

Crystal Palace - the energetic field of Qi located within the third ventricle of the brain.

Cupping - a technique used in acupuncture therapy in which wooden, clay, or glass cups are adhered to the patient's skin by suction, to drain or remove pathogenic Qi from the body's pores.

Curious Organs - the body's Six Extraordinary Yang Organs which function like Yin Organs as they store Yin Essence (i.e., Blood, Marrow, or Kidney Essence), but look like Yang Organs (because they are hollow). These organs consist of the uterus, brain, Marrow, bones, blood vessels and Gall Bladder.

Cycle of Disharmony - an emotional state which induces a vicious cycle of physical, mental, emotional, energetic and spiritual disharmony.

Cyst - a closed sac that forms in tissue or a body cavity.

Cystitis - inflammation of the Bladder, usually occurring secondary to ascending urinary tract infection.

Cytotoxic Treatments - treatment containing toxins which attack the cells of the body (for example: radiation and chemotherapy).

D

Dacryorrhea - excess tear flow.

Damp - a internal pathogenic condition relating to the storage of Damp or Wet toxins.

Dantians - the body's three main energetic pools, or reservoirs located in the head, chest and lower abdomen.

Dantian Regulation - the principle of balancing the energy governing the body's three main energetic pools (Dantians).

Dao - pertaining to God or divine consciousness.

Daoist - a student of the "Dao" or way of life, pertaining to living in harmony with the universe and environment.

Dao Yin - energy regulation exercises consisting of training the body, mind, and breath.

De - pertaining to an individual's personal Virtue.

Defence Mechanisms - see Ego Defence Mechanisms.

Deficiency - a condition relating to the inadequate degree of a particular substance, e.g., Qi, Blood, Yin, Yang, Heat, etc.

Delusions - refers to the occurrence of a mental derangement in the patient resulting in a false belief based on incorrect inference about external reality. Regardless of the evidence to the contrary, the belief is strongly maintained.

Demon or Spirit Oppression - the condition of having a foreign spirit attach to a patient's external Wei Qi field, resulting in emotional disharmony.

Demon or Spirit Possession - the condition of having a foreign spirit invade and inhabit the patient's body, mind, emotion, and spirit.

Denial - a conscious refusal of an impulse-evoking fact, feeling or memory.

Depersonalization Disorder - persistent, recurring episodes of depersonalization, characterized by a feeling of detachment, or estrangement from one's self.

Depression (Major) - a psychological disorder resulting in major sadness and pessimism, feelings of worthlessness, helplessness and hopelessness. Symptoms include overeating or under-eating, insomnia or hyper-insomnia, difficulty concentrating, and fatigue.

Descending - the action of energy moving downward.

Diabetes - a general term used to describe diseases characterized by excessive urination and a sugar imbalance in the Blood.

Diaphoresis - profuse sweating.

Diathermy - the therapeutic use of a high-frequency current to generate heat within a certain area of the body.

Di Qi - Earth (Environmental) Energy.

Disharmony - pertaining to a lack of adequate balance of energy.

Disorder - an abnormal state of physical, mental, emotional, energetic or spiritual disharmony.

Dispersing- the spreading of Qi to other parts of the body, or purging of pathogenic energy from the body.

Displacement - the shifting of impulses aroused by one person, or situation on to a safe target.

Dissociated Identity Disorder - new terminology used for multiple personality disorder.

Distance Therapy - also called External Qi Therapy, Qi Emission and Outgoing Qi Therapy, is defined as extending or projecting energy into a patient from a distance.

Divergent Channels - twelve secondary channels that parallel the Twelve Primary Channels.

Divine - pertaining to God.

Divine Center - referring to the North Star's stable position in the sky.

Divine Hook-Up - the Qigong doctor's initial preparation for therapy, wherein he or she connects with the divine for guidance.

Divine Therapy - long distance Qigong healing.

D.M.Q. - a licensed Doctor of Medical Qigong Therapy, presently only obtainable in China.

Dong - the Yang method of dynamic Postural Dao Yin training.

Dragon's Mouth Palm Technique - hand technique wherein the thumb touches the other four fingers of the hand, forming an image of the head of a dragon. This hand manipulation is used for leading and pulling the Qi.

Draining Qi - drawing off pathogenic energy from a specific organ area or channel of the patient's body.

Dredging - a type of energetic purging, used to clean the patient's energetic fields and channels.

Drilling Energy Technique - this method refers to rotating Qi in a spiraling pattern (clockwise or counterclockwise) to access the energy deep inside the patient's body.

Dryness - a internal pathogenic condition relating to the storage of Dry toxins.

Dynamic Qigong - energy gathering which utilizes active movements of the body.

Dysmenorrhea - painful menstruation.

Dyspepsia - painful digestion.

Dyspeptic - one afflicted with dyspepsia.

Dysphagia - difficulty in swallowing.

Dysphoria - exaggerated feeling of depression, anxiety and unrest.

Dysplasia - the abnormal development of tissue.

Dyspnea - air hunger, resulting in difficult breathing, shortness of breath, sometimes accompanied by pain.

Dysthymia - a chronic, form of depression (lasting at least two years), for children and adolescents the mood can be irritable rather than sad.

E

Earth Element - one of the Five Elements, pertaining to Spleen and Stomach.

Earth Energy (Qi) - Energy pertaining to the Earth and surrounding environment.

Earth Jing - energy that supervises the maturation phase of the fetus's ability for emotional and spiritual bonding during the seventh month of pregnancy.

Earthly Branches - twelve energies of the Earth represented in the human body as the Twelve Primary Channels.

Earth Transpersonal Point - pertaining to the body's energetic connection to the Earth, located several feet beneath the feet.

ECG (or EKG) Eletrocardiogram - a graphic record made by an instrument that measures the Heart's electrical activity; usually used to confirm a diagnosis of a Heart condition.

Eclampsia - coma and convulsive seizures (between the 20th week of pregnancy and first week postpartum). Symptoms result in edema of the legs and feet, puffiness of the face, hypertension, severe headaches, dizziness, epigastric pain, nausea, sudden convulsive seizures and coma.

Eczema - an acute or chronic inflammation of the skin.

Edema - an acute or chronic cutaneous inflammatory condition.

EEG Electroencephalogram - a graphic record made by an instrument that measures the brain's electrical activity and records it as patterns of fluctuating waves.

Ego - the ego is the mediator between the id and the superego. According to Dr. Sigmund Freud's psychoanalytical theory, the superego combines the critical inner parent aspect with the idealistic aspect of the individual's conscience; the id consists of unconscious drives and instincts. The ego as meditator is responsible for ensuring rational behavior.

Ego Defense Mechanisms - according to Dr. Sigmund Freud's psychoanalytical theory, the ego defence mechanisms include: Repression, Displacement, Projection, Intellectualization, Regression, Fixation, Denial, Reaction-Formation, and Sublimation.

Eight Confluential Points - the areas where the Eight Extraordinary Vessels and Twelve Primary Channels intersect with each other.

Eight Energetic Principles - a system of differential diagnosis using four pairs of opposites (Yin and Yang, Cold and Hot, Deficient and Excess, Internal and External).

Eight Energetic Touches - pertaining to the somatic tissue response to energy stimulation (tingling, sensations of heat, coldness, expansiveness, contracted, heaviness, lightness, and vibration).

Eight Extraordinary Vessels - the primary channels responsible for the formation of the fetus, which after birth, are considered the body's reservoirs for collecting the overflow energy from the Twelve Primary Channels.

Eight Miscellaneous Factors - eight factors that can off set the patient's Yin and Yang balance (diet, overexertion, excessive sex, child bearing, traumatic injuries, exposure to poisons, parasites, and iatrogenic disorders).

Eight Trigrams (Bagua) - eight cosmological patterns of three lines (solid and / or broken), called Yaos, used to diagnose as well as predict future transitions.

Emaciation - the state of being malnourished and extremely lean.

Embolus - the obstruction of a blood vessel by foreign substances or a blood clot.

EMG Electromyogram - a graphic record made by an instrument that measures the muscle's electrical activity and records its function.

Emitting Qi - the Qigong doctor extending energy outside the body for the purpose of treating a patient.

Emotional Energy Body - is the external energy existing in the body's second field of Wei Qi, which is attached to the internal organs.

Emotional Spirit - pertaining to the Emotional Energy Body.

Empathic Communication - the doctor's ability to experience the feelings and thoughts of his / her patients.

Empty Qi - a serious weakness or Deficiency of the body's Qi.

Encephalomyelitis - acute inflammation of the brain and spinal cord.

Endrometrial Hyperplasia - excessive proliferation of the cells within the lining of the uterus.

Energetic Armoring - a condition resulting from the patient protecting specific tissues, organs, or areas of the body. Energetic armoring is initiated when the patient freezes certain emotional feelings to maintain the denial system.

Energetic Barriers (Energetic Boundaries) - the protective barriers existing within and outside of the body's tissues.

Energetic Complications - pertaining to energetic imbalances within the patient's tissues (e.g., compressed energy stagnation, energetic armoring, migrating Qi deviations, etc.)

Energetic Cords - energetic bands of light and vibration which form an emotional attachment, connecting the patient to certain people, places, or things.

Energetic Grids - an energetic net covering the surface of Heaven , Earth or the human body.

Energetic Leakage - a condition resulting from the patient releasing Qi out of the joints, due to an injury or unconscious sabotage.

Energetic Medicine - any and all medicine having to do with the stimulation, cultivation, tonification, purgation, balance and maintenance of the body's Qi.

Energetic Point Therapy - extending energy from a distance into specific channel points, or vessels on the patient's body.

Enteric Nervous System - pertaining to the nervous system of the Small Intestine.

Enuresis - the involuntary discharge of urine.

Environmental Energy (Qi) - Energy pertaining to the Earth and surrounding environment.

Environmental Force - energy pertaining to the Earth and surrounding environment.

Epigastric Pain - pain in the region over the pit of the stomach.

Epileptiform - having the form of epilepsy.

Epistaxis - bleeding of the nose.

Epithelial Hyperplasia - excessive proliferation of the cells within the outer surface of the body, including the secreting portions of the glands and ducts.

Essence (Jing) - referring to either Prenatal and Postnatal energetic tissue mass.

Eternal Soul - the individual's True Self, which is always connected to the divine. It is absorbed into the mother's egg at the time of conception, and is rooted within the body's Taiji Pole.

Ethereal Matter - the body's foundational energetic substance.

Ethereal Souls - see Hun.

Etiology - the causes of a disease.

Eustachian Tubes - the auditory tube, extending from the middle ear to the pharynx.

Evil Embryo - pertaining to a toxic formation in the form of a tumor or cancer mass.

Evil Influences - pathogenic factors that can be either physical, mental, emotional, energetic or spiritual.

Evil Qi (Xie Qi) - also called Pathogenic Qi, Perverse Qi, Toxic Qi, and Heteropathic Qi, is energy that causes disease or harmful effects to the body.

Evil States - a condition wherein the patient experiences mental delusions, obsessions, infatuations with the doctor, spiritual oppression or possession.

Evil Wind - toxic wind that invades the body, tissues or organs.

Excess - a condition relating to the over abundance of a particular substance. (e.g., Heat, Wind, Damp, Cold, etc.).

Exopathogenic - a disease or pathogen originating outside of the body.

Extended Fan Palm Technique - hand technique where the fingers separate like a Chinese fan, used for extending energy through the doctor's palm.

External Pathogenic Factors - an external invasion of Heat, Damp, Cold, Dryness, or Wind, or a combination thereof.

External Qi Therapy - a technique used in Medical Qigong therapy which pertains to Qi being emitted onto a patient.

Extraordinary Organs - also called Curious Organs, these six organs are shaped like Yang (Hollow) organs but store Yin Essence. The brain stores Marrow, the Marrow stores Kidney Jing, the bones store Marrow, the blood vessels store the Blood, the Gall Bladder stores the bile, and the uterus stores Kidney Jing, Blood, and Qi.

Extraordinary Vessels - secondary channels that flow in conjunction with the body's Twelve Primary Channels.

Extra Point - a point with a definite location, but not originating on the fourteen main channels.

F

False Cold - a feverish sensation followed by a feeling of cold all over the body.

False Self - pertaining to the dark emotional side of the self, i.e., the masks and defence mechanisms that serve to protect the individual from dealing with his or her issues.

False Yang - a clinical condition where an overabundance of Cold (Yin) transforms into Heat (Yang).

False Yin - a clinical condition where an overabundance of Heat (Yang) transforms into Cold (Yin).

Fascia - a fibrous membrane covering, supporting, and separating the muscles, as well as uniting the skin with the body's underlying tissues.

Febrile Diseases - any and all diseases which cause the body to produce a fever.

Feng Shui - the study of harmonizing the energetic flow of Wind and Water, and the healing art of adjusting the person's environment to create improvements in the person's health and life.

Fetal Education - regulating the mother's behavior to improve her child's physical, emotional, and mental health is called "fetal education" in Traditional Chinese Medicine, and is important in the development of the child's Prenatal Essence, Energy, and Spirit.

Fetal Leakage - after conception, if a small amount of bloody fluid discharges from a woman's vagina, it is known as Tai Lou or fetal leakage.

Fibroadenoma - a tumor with fibrous tissue, forming a dense covering.

Fire Element - one of the Five Elements, pertaining to the Heart, Small Intestine, Pericardium, and Triple Burners.

Fire Jing - energy that controls the development phase of the fetus's emotional and spiritual well-being during the fifth month of pregnancy.

Five Agents - the five energies that are linked to the moral qualities of a person's inner characteristics (the five virtues stored within the body's Wood, Fire, Earth, Metal and Water elements).

Five Elements - Wood, Fire, Earth, Metal and Water.

Five Element Animals - Green Dragon, Red Phoenix, White Tiger, Dark Blue Turtle, and Yellow Phoenix.

Five Elemental Constitutions - physical constitutions based upon the observation of the Five Elemental formations within the human body.

Five Element Organs - the organs related to the Five Elements, including: Wood - Liver and Gall Bladder, Fire - Heart and Small Intestine (also included are the Pericardium and Triple Burners), Earth - Spleen and Stomach, Metal - Lungs and Large Intestine, Water - Kidneys and Bladder.

Five Energetic Fields - pertaining to the body's five levels of energy, including: 1. External Wei Qi Fields, 2. Internal Ying Qi Field, 3. Sea of

Blood, 4. Sea of Marrow, 5. Center Core of Light (Taiji Pole).

Five Flavors - also called the Five Tastes, are sour, bitter, sweet, pungent, and salty.

Five Major Yang (Fu) Organs - also called the Five Bowels, they are the body's five hollow organs: Gall Bladder, Small Intestine, Stomach, Large Intestine, and Bladder.

Five Major Yin (Zang) Organs - are the body's five solid organs: Liver, Heart, Spleen, Lungs and Kidneys.

Five Orbs - pertaining to the 5 Yin internal organs (Liver, Heart, Spleen, Lungs and Kidneys), their complete organ system, and the surrounding areas that they influence.

Five Palms Hot - a condition in which the patient feels a hot sensation in the palms and soles of the feet, accompanied by heat and agitation of the chest and / or head area.

Five Passes - five important gates on the Governing Vessel located at the coccyx, Mingmen, Shendao, occiput and Baihui, where energy tends to stagnate.

Five Portals of the Earthly Yin Gate - pertaining to the five points at the bottom of the perineum through which the Qi of Earth enters into the body.

Five Portals of the Heavenly Yang Gate - pertaining to the five points at the top of the head through which the Qi of Heaven enters into the body.

Five Sense Organs - eyes, tongue, mouth, nose and ears.

Five Thunder Fingers Technique - hand manipulation technique wherein the fingers and thumb are rapidly extended from a closed soft fist, to strike with Qi for dispersing stagnations.

Five Thrusting Channels - see Thrusting Channels.

Five Tissues - tendons, blood vessels, muscles, skin and bones.

Five Zhi - in connection with the five mental aspects of the Mind, the Hun (Ethereal Soul), Po (Corporeal Soul), the Shen (Spirit), the Yi (Intellect), and the Zhi (willpower) are sometimes referred to as the Five Zhi.

Fixation - has the same result as regression, but the person becomes fixated at a particular stage of mental and emotional development.

Flat Palm Detection - an extended palm technique used for sensing and diagnosing.

Four Bigs - pertaining to severe excess of fever, sweating, thirst, and pulse.

Four Doors - the center of each palm and foot.

Four Winds - pertaining to the Energy of the four compass points. (North - back, South - front, West - right, and East - left.)

Fu Organs - Yang or hollow organs (Gall Bladder, Small Intestine, Stomach, Large Intestine, and Bladder). The Fu Organs operate primarily to relieve the Zang (Yin) Organs of toxic energies and wastes.

G

Gallow's Syndrome - laughing at a very painful experience instead of grieving or crying.

Gastritis - the inflammation of the stomach.

Gastroptosis - the downward displacement of the stomach.

Gate of Access - the passageway between life and death, believed to be related to the stars of the Big Dipper.

Gate of the Moon - the western energetic region, which the sun must pass through in order to create Autumn.

Gathering the Immortal's Water - pertaining to the energetic production of the saliva or Juice of Jade to create the Immortal Pill.

Gathering Qi - also called Respiratory Qi, Collection Qi, Chest Qi, Pectoral Qi, and Big Qi of the Chest. It is derived from the conversion of the purest and most potent forms of the body's Jing (particularly sexual fluids, hormones, and neurochemicals).

Ghosts (Gui) - disembodied spirits.

Ghost Points - points used for the treatment of emotional and spiritual disorders.

Gland - an organ that produces a hormone or other secretion.

Golden Gate in the East - the eastern energetic region, which the sun must pass through in order to create spring.

Gong - meaning skill or study.

Gout - sudden intense pain in a joint, usually the big toe or ankle, followed by swelling, inflammation and heat in the joint (in extreme casts alternating chills and fever are experienced).

Grandmother Element - pertaining to the Five Element Creative Cycle of Traditional Chinese Medical, the primary organ is considered the Mother and its previous organ is considered the Grandmother.

Grain Qi (Gu Qi) - Energy derived from food and drink.

Grounding - see Rooting.

Gua Sha Therapy - an external "surface" tissue scraping technique commonly used to clear Excess Heat from the body.

Guiding Qi - the technique of leading Qi.

Gu Qi - Energy derived from food and drink.

Gui - ghosts or spirits.

H

Hai - means sea.

Healing Tones - resonant sounds used to purge the body of pathogens.

Heart Fire - pertaining to the energy of the Heart, which is responsible for transforming the body's Energy into Spirit. This occurs in the chamber of the Heart's courtyard (the Yellow Court).

Heaven Qi - also known as Heavenly Qi, this energy pertains to the Heavens, the divine and the celestial influences.

Heavenly Stems - the ten energies of Heaven represented in the human body as the Yin and Yang aspects of the Five Elements or ten major internal organs.

Heavenly Transpersonal Point - pertaining to the body's energetic connection to the Heavens, located two to five feet above the head.

Hei Xia - dark brown or black vaginal discharge.

Hemiparalysis - paralysis on one side of the body

Hemiplegia - paralysis on only one side of the body.

Hemoptysis - the expectoration of Blood.

Hepatitis - inflammation of the Liver.

Hepatolithiasis - calculi or concretions in the Liver.

Hepatomegaly - enlargement of the Liver.

Hepatosplenomegaly - the enlargement of both the Spleen and Liver.

Herbal Therapy - one of the four branches of Traditional Chinese Medicine, which involves treating patients through the use of formulas created through teas, soups, tinctures, wines, oils, balms, liniments and pills to stimulate energy flow.

Herbalist - a doctor of Herbal Therapy.

Herpes Simplex I - an infectious disease caused by the herpes simplex virus. This disease is characterized by thin-walled vesicles that occur in the skin, usually at a site where the mucus membranes joins the skin.

Hexagram - a six line symbol representing the function and flow of Yin and Yang energy, formed by stacking two trigrams on top of each other.

Hibernation Breathing - a breathing method which includes inhaling and exhaling through every pore on the body's surface, from the body's Center Core (Taiji Pole).

Hidden Power (An Jing) - techniques that emphasize stretching and twisting the tendons and ligaments (known as Reeling and Pulling the Silk) to cultivate resonant vibration within the body for striking and issuing power.

Hollow Organs - the body's Yang organs, which consist of the Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder. Also included in this list are the Triple Burners.

Hologram - pertaining to the body's energetic three dimensional image.

Hook-Up - see divine Hook-Up.

Hostile Forces - dark spiritual forces which seek to influence the physical, mental, emotional, energetic and spiritual life of an individual.

Hot Constitution - pertaining to a physical body innately prone towards heat.

Hot Evil - also known as Evil Heat, a pathogenic condition causing Excess patterns that are Hot and Yang in nature.

Hou Tian Zhi Qi (Postnatal Qi) - translates to mean after the baby sees the Heavens.

Hua Jing - see Mysterious Power

Huang - any membranous tissue.

Hua Sound Resonation - Mysterious or Spiritual Sound Resonation.

Huang Dai - yellow vaginal discharge.

Hui Yin Point (CV-1) - Conception Vessel point located between the scrotum and the anal sphincter.

Human Force - the energy or force manifesting from inside of the human body, as well as the within the human energetic field.

Hun (Ethereal Soul) - the Three Ethereal Souls which are the spiritual part of man that ascends to Heaven upon the death of the body.

Hunter Killer Cells - the body's neutrophils and macrophages, as well as the interferons and antibodies.

Hun Yuan - the body's internal energies that has been combined and fused into the Lower Dantian area.

Hyperbilirubinemia - excessive amounts of bilirubin (the orange-colored or yellowish pigment in bile) in the Blood.

Hyperhidrosis - excessive sweating due to an over-activity of the sweat glands.

Hypertension - high blood pressure.

Hyperthyroidism - a condition caused by excessive secretions of the thyroid glands, resulting in an increased metabolic rate and the consumption of food to support this increased metabolic activity.

Hypertrophy - the increase in the size of an organ or structure that does not involve tumor formation.

Hypochondriac Pain - pain in the upper lateral region on each side of the body below the thorax and beneath the ribs.

Hyposmia - a deficient sense to smell.

Hypotension - low blood pressure.

I

I-Ching - see Yi-Jing

Id - Dr. Sigmund Freud's terminology for one of the three divisions of the psyche in Psychoanalytic Theory that is the unconscious source of psychic energy responsible for the body's drives and instincts.

Immortal's Pill - pertaining to the energetic production of Heaven and Earth Qi, in conjunction with the saliva.

Immortal's Water - when meditating, another word for energized saliva.

Incontinence - an inability to control urination, involuntary urination when coughing, laughing, sneezing, running, or performing some other physical activity. This condition can also refer to involuntary defecation.

Infatuations - refers to the patient feeling intensely amorous towards the doctor.

Influenza - an acute contagious respiratory infection, characterized by a sudden onset, with chills, fever and headache.

Inner-vision - the skill of observing images of the internal organs, and energetic fields relating to the body, mind, emotion and spirit, and their transition stages.

Insight Meditation - focusing on sensual stimuli (sounds, smells, colors, etc.) while meditating.

Insomnia - a sleeping disorder.

Insulting Cycle - pertaining to the Five Elemental Cycle, wherein the Child Element counter attacks the Grandmother Element.

Intellectualization - an elaborate rationalization of a naked impulse, to justify it.

Interferons - a group of proteins released by the white blood cells and fibroblasts, responsible for fighting infection.

Interjection - the insertion or interpose the energetic pattern.

Internal Dialogues - internal conversations, which are part of the patient's personal belief structure.

Internal Pathogenic Factors - pertaining to diseases originating from the body's internal organs and emotions (e.g., anger, fear, grief, worry, etc.).

Internal Viewing - technique used by the Qigong doctor to view the patient's internal organs.

Interpersonal Relationships - close personal relationships, relating to self and others.

Invading Cycle - pertaining to the Five Elemental Cycle, wherein the Grandmother Element overcontrols the Child Element.

Invisible Needle Therapy - the insertion of invisible energetic needles into the patient's body to stimulate energy flow.

Invisible Needle Palm Technique - Qi emission in which energy is emitted in a very fine line, to stimulate specific channel points.

J

Jaundice - a condition due to deposition of the bile, characterized by the yellowing of the skin, eyes, mucous membranes and Body Fluids.

Jia Ji Guan (Spinal Pass) - two points located on the lateral sides of the Mingmen (GV-4) where Qi can become stagnant.

Jie Gu Therapy - a tissue manipulation therapy that focuses on the adjustment of bones and ligaments.

Jin - thin Body Fluids.

Jing - the human body's Essence, divided into Prenatal and Postnatal Essence.

Jing Luo - the body's channels and collaterals.

Jing Point Therapy - the original term for Channel Point Therapy or Acupressure.

Jing Shen Bing - pertains to all types of mental illness.

Jiu Wei (Yellow Court) - located in the center of the diaphragm, below the xiphoid process of the sternum. Its function is that of being the access point to release emotional memories from the body's internal organs. Its location is also attributed to the 3rd Chakra.

Jue Yin (Reverting Yin) - associated with the most severe diseases, indicates Yin Qi developing its final stage and then reverting into Yang.

Juice of Jade - Energized saliva produced in meditation practices.

K

Karma - the manifestation of consequences to our actions and beliefs: "As you sew, so shall you reap."

Karmic Related Illness - pertaining to spiritual illnesses, which can be either congenital or acquired.

Ketheric Matter or Substance - pertaining to the spiritual energy located within the third external field of Wei Qi.

Kidney Fire (Mingmen Fire) - the energy that heats the body's Essence (Jing), and dominates all Twelve Primary Channels. It is the motivating force of the body.

Kinetic Communication - the intuition of the physical body, felt by the Qigong doctor as a movement in or of his or her own body.

Kneading Tiger Palm Technique - hand manipulation technique resembling the movement of a tiger kneading the ground, used for dispersing Qi stagnations.

Kyphotic - the exaggeration or angulation of the normal posterior curve of the spine (humpback).

L

Laogong - channel point located at the center of each palm.

Large Heaven Cycle (Macrocosmic Orbit) - Qigong Meditation which connects the Qi of the extremities to the Qi within the Governing and Conception Vessels.

Leading Qi - technique of manipulating the patient's Qi by using a guiding gesture with the hands.

Leukocyte - the body's white blood corpuscles, which included lymphocytes and other immune system cells.

Leukorrhea - an acute or chronic disease caused by the unregulated clonal proliferation of stem cells within the Blood forming tissues.

Ley Lines - the energetic pathways that connect energy spots on the planet.

Life Force Energy - Qi.

Light Energy Therapy - color, light projection and visualization used for healing.

Light of the Dao - divine healing light energy.

Lipid Bilayers - the outer membrane of most cells, includes two layers of lipid molecules.

Lithotripsy - crushing of a calculus in the bladder or urethra.

Liver Wind - terminology used to describe excess Heat generated from a toxic Liver condition which can cause pathogenic symptoms.

Lobular - composed of small lobes.

Lords of the Three Dantians - three spiritual energies used to describe the energetic aspects of the human soul (Tai Yi, Si Ming and Xia Tao Kang).

Lo Scroll (Magic Square) - a tool used for diagnosis and treatment of disorders through number configurations, which correspond to the Late-Heaven sequence of the Trigrams of the Yi-Jing (I-Ching).

Lower Burner - area of the body in the Lower Dantian, responsible for the separation of Clean and Dirty Fluids. It also facilitates the production of urine.

Lower Dantian - area in the center of the lower abdomen, attributed to the body's chamber of heat and physical power. The Lower Dantian is also known as the Sea of Energy, Pill of Immortality, Root of life, Source of Generating Qi, Five Qi Collection Seat, Progenitor of Life, Stove of Spirit, Root of Heaven, and Cinnabar Field.

Lower Unconsciousness - pertaining to the acquired mind which is connected to the primal senses.

Lumbago - dull, aching pain in the lumbar region of the lower back.

Luo - translated to mean a net or web, and in Traditional Chinese Medicine it refers to the Connecting Vessels, i.e., the Fifteen Collaterals. These vessels are the major "passage ways" for the circulation of the body's channel energy, emerging out of the Luo (pathway) points on the Twelve Primary Channels (plus the Governing and Conception Vessels).

Luo Points - are the major intersecting points of the Fifteen Collaterals. The Luo points are located below the elbows and knees and provide an additional energetic barrier to keep Evil Winds from affecting the Twelve Primary Channels, being somewhat deeper then the Muscle/Tendon Channels.

Lymphocytes - immune cells present in the Blood and lymphatic tissue.

Lymphoma - a group of malignant solid tumors of the lymphoid tissue.

Lymphosarcoma - a sarcoma of the lymphatic system.

M

Macrocosmic Orbit - Qigong Meditation which connects the Qi of the Governing and Conception Vessels with the extremities of the body.

Macrophages - the major phagocytic cells of the immune system (also known as Hunter Killer Cells).

Magic Square - also known as the Lo Scroll, a tool used for diagnosis and treatment of disorders through number configurations corresponding to the Late-Heaven sequence of the trigrams of the Yi-Jing (I-Ching).

Magnetic Energy Therapy - magnetic energy affecting the body's channels and points via electromagnetic field stimulation.

Malar Flushes - pertaining to flushed skin along the cheeks.

Malignant - detrimental; growing worse; threatening to produce death.

Man Qi - general term used to describe the energy relating to the body, mind, emotion, energy and spirit of both men and women. In the body, the area between the navel and the lower sternum correspond to "Man."

Manic Depressive Personality Disorder - see Bipolar Personality Disorder

Mantra - a Sanskrit word, meaning a spiritual phase or sound repeated internally or externally, used as a tool in meditation to induce an altered state of consciousness.

Marrow - derived from the Kidneys, nourishes the brain, spinal cord and forms the bone marrow.

Master Point - the main point of energy interaction on a specific channel, used to affect another organ system or channel energy flow.

Mastitis - inflammation of the breast.

Medical Qigong - one of the four branches of Traditional Chinese Medicine.

Menorrhagia - excessive bleeding during the time of menstruation.

Menoxenia - the pathological changes of menstruation occurring in a woman's cycle, affecting the color, quantity and quality of Blood flow.

Mental Delusions - the occurrence of mental derangement in the patient resulting from a false belief based on an incorrect inference regarding external reality. This belief is firmly sustained despite incontrovertible evidence to the contrary.

Meridians - the body's channels or rivers of Energy.

Message (Xin Xi) - knowledge stored within the Wuji or the Void.

Metal Element - one of the Five Elements, relating to the Lungs and Large Intestine.

Metal Jing - Energy that supervises the development phase of the fetus's ability for emotional attachment and bonding during the sixth month of pregnancy.

Metrorrhagia - bleeding from the uterus.

Microcosmic Orbit - energetic orbits that circulate the Qi within the body's energetic channels; divided into Fire, Water and Wind pathways.

Micturition - discharging urine.

Middle Burner - area of the body pertaining to the body's digestive system, responsible for transporting Gu Qi (derived from food and drink).

Middle Dantian - area in the center of the chest, attributed to the body's chamber of emotional and vibrational power. The Qi of the Middle Dantian is called Zong Qi. Zong Qi is translated as Gathering Qi, Ancestral Qi, Genetic Qi, or Essential Qi. The Middle Dantian is also known as Middle Field of Elixir, Scarlet Palace, Central Altar, Middle Sea of Energy, Courtyard of the Heart, Opening of Suspended Gold, and the Seat of Emotion.

Middle Emotional/Mental Barrier - the second and middle energetic barrier of the Wei Qi fields.

Mind Regulation - the principles governing the conduct, action or functions of the mind.

Ming Jing - see Obvious Power.

Mingmen (Gate Of Life) - area responsible for heating the body, in particular the Kidneys and Lower Dantian.

Mingmen Fire (also known as Kidney Fire, Advisor Fire, or Ministerial Fire,) - the Energy that heats the body's Essence (Jing), and dominates all Twelve Primary Channels. It is the motivating force of the body.

Ming Sound Resonation - Obvious or Audible Sound Resonation.

Ministerial Fire - pertaining to the Energy responsible for heating the Middle Burner.

Moon Cream - energy gathered during meditation from the moon's Essence in the form of cool light.

Morphogenic Field - pertaining to the form of the energetic field of Jing.

Mother and Child Therapy - the Traditional Chinese Medical description of the primary organ (Mother) and its sequential organ (Child) in the Five Elemental Creative Cycle.

Mother Element - pertaining to the Five Element Creative Cycle, the primary organ is considered the "mother."

Moxa Therapy - herb heated and applied in a clinical setting for tonification.

Multiple Personality Disorder - also called Identity Disorder, is a mental state in which the patient develops "alter" personalities as a coping mechanism in dealing with severe emotional traumas.

Multiple-Sclerosis (MS) - an inflammatory disease of the central nervous system in which infiltrating lymphocytes degrade the myelin sheath of nerves.

Muscle/Tendon Channels - channel connections to the body's muscles, tendons, ligaments and other connective tissues.

Myalgic Encephalomyelitis - acute inflammation of the brain and spinal cord.

Myocarditis - the inflammation of the cardiac muscle (located in the middle layer of the walls of the heart).

Myoma - a uterine tumor that is a solid benign growth in the myometrium, often called a fibroid, containing muscle tissue.

Myophagism - a condition where the macrophages destroy (eats) muscular tissue.

Mysterious Pass - the space between Yin and Yang where infinite space and time (Wuji) exists.

Mysterious Power (Hua Jing) - techniques which emphasize training and conditioning the mind's imagination and intention, to project and utilize the power of the individual's Shen (Spirit).

N

Necrosis - part of an area of tissues or bone that is dying or dead and may spread to healthy tissues or bones.

Nei Dan Shu - internal elixir cultivation, that focuses on cultivating Qi from within the individual's body.

Nei Gong (Internal Skill) - the training of the body's tendons, bone, breath, mind, emotion, and spirit to facilitate internal power.

Nei Guan (Internal Viewing) - see inner vision.

Nei Jing - the Yellow Emperor's classics on Chinese internal medicine.

Neoplasm - a new or abnormal formation of tissue, as in a tumor growth.

Nephritis - inflammation of the Kidneys.

Neurasthenia - unexplained chronic fatigue and lassitude.

Neutrophils - the most common type of granulocytic white blood cell, responsible for fighting infection.

Nine Dantian Chambers - the nine internal cavities established within the energetic matrix of each Dantian.

Nine Palaces - the Later-Heaven sequence of the Trigrams of the Yi-Jing (I-Ching), represented in the human body as the Eight Extraordinary Vessels and the Taiji Pole.

Nine Star System - pertaining to the total development of the three periods of life and the three star developmental sequence.

Noxious Qi (Turbid Qi) - coarse, Toxic, Evil, unrefined, polluted, or dirty Energy.

O

Objective World - pertaining to the spiritual world existing outside an individual's thoughts or feelings.

Obstructed Qi - Energy that is immobile.

Obstruction - the inhibition of the flow of Qi or Blood, caused by Cold, Damp, Heat and Wind, etc.

Obvious Power (Ming Jing) - techniques that emphasize the training and conditioning of the muscles, strengthening the bone structure, and increasing the individual's overall stamina. This school also includes such techniques as pounding the body (arms, hands, legs, and torso) to strengthen and toughen the tissues.

Occipital Pass (Yu Zhen Guan) - the area located just inferior to the occipital bone where the brain originates (according to energetic embryology), known as a specific point where Qi often stagnates.

Oliguria - diminished amount of urine formation.

Omniscient Sight - the ability to see 360 degrees simultaneously.

One Finger Skill Technique - clinical modality involving Energy extension employed through a single finger.

Ontology - the study of the historical development of an individual.

Opening and Closing - the method of leading Qi into and out of specific internal organs, the Triple Burners areas, or one of the Three Dantians, via the hands.

Opisthotonosis - a form of spasm in which the patient's head and heels are bent backwards, and the body is bowed forward.

Organ Dysfunction - the impaired or abnormal function of an internal organ.

Organ Regulation - technique for balancing the action or functional principles of the internal organs.

Organ Qi - Energy of the body's Yin and Yang organs.

Original Spirit (Yuan Shen)- see Prenatal Spirit

Original Qi (Yuan Qi) - pertaining to the body's Prenatal Qi acquired from both parents at conception, and from the mother during gestation.

Original Force - pertaining to the Heavenly energy, manifesting as the energy of the entire cosmos.

Original Yang - pertaining to the body's Prenatal Kidney Yang.

Original Yin - pertaining to the body's Prenatal Kidney Yin.

Osteoarthritis - a chronic disease involving the joints and the deterioration of the articular cartilage.

Osteoporosis - a general term used for describing any disease process that results in the reduction of bone mass.

Osteosarcoma - a sarcoma of the bones.

Outer Spiritual Barrier - the third and furthermost energetic barrier of the Wei Qi fields.

Overcontrolling Cycle - pertaining to one of the Five Elemental Cycles, where one organ overcontrols the second elemental organ in the Five Elements' Circle.

P

Palace of Eternal Frost - the northern energetic region, which the sun must pass through in order to create winter.

Palace of Universal Yang - the southern energetic region, which the sun must pass through in order to create summer.

Palpitations - an abnormal rapid, throbbing, or fluttering of the Heart.

Pancreatitis - inflammation of the pancreas.

Panic Attack - overwhelming panic and sense of impending doom, resulting in hyperventilation (breathlessness), heart palpitations and visual distortions.

Papillary Masses - small, nipple-like protuberances or elevated tissue masses.

Paraplegia - paralysis on both sides of the body.

Parenchymal Cells - the essential parts of an organ's cells that are concerned with the organ's function.

Parkinson's Disease - a chronic nervous disease characterized by muscular weakness, rigidity and a fine, slow tremor.

Past Life Regression - pertaining to the patient's exploration and experiences of previous lives.

Pathogenic - disease-causing; see Internal Pathogenic Factors and External Pathogenic Factors.

Penetrating Wind - pertaining to the external pathogen of Wind invading the tissues.

Peribronchial - surrounding the windpipe (bronchus).

Perineural Cells - the sheath of cells around a bundle of nerve fibers within the perineurium.

Peristalsis - a progressive wave like movement that occurs involuntarily in the hollow tubes of the body.

Peritonitis - inflammation of the abdominal cavity.

Pernicious Influences - pertaining to the Six External Factors that cause disease.

Personal Subconscious Mind - part of the mind associated with the recording and storing of personal interpretations of reality.

Peyer's Patch - an aggregation of lympth nodes found chiefly in the ilium.

Phagocytes - cells that have the ability to destroy and ingest bacteria, protozoa, unhealthy cells and cell debris.

Phantom Embryo - an energetic thought form in the shape of an embryo created through the woman's feelings of grief, guilt or remorse after a surgical abortion.

Phantom Organ - the energy of a particular organ which still exists, even after surgical removal.

Phantom Pain - the feeling of pain relating to a particular organ which still energetically exists, even after surgical removal.

Phlegm - pathogenic factor responsible for the formation of diseases including tumors.

Physical Barrier - the first level and closest to the body of the three Wei Qi energetic barriers.

Piezoelectric - pertaining to the electricity created from pressure, especially pressure on or within the bones.

Po (Seven Corporeal Souls) - sometimes called the Seven Turbid Demon Natures, this spiritual energy manifests as the physical or material soul of the human body that returns to the Earth at death.

Points - specific areas on the body where energy can intersect to travel externally to internally, or visa versa.

Point Respiration - exercise which requires breathing into a specific channel point, organ, or area of the body.

Polarity - opposite negative and positive qualities of power.

Polergeists - malevolent spiritual entities. Parapsychology research indicates that poltergeist activity is often the manifestation of a psychokinetic ability.

Polydipsia - excessive thirst.

Polyphagia - eating abnormally large amounts of food at a meal.

Polyuria - the excessive secretion and discharge of urine.

Portal Hypertension - the increased pressure in the portal vein resulting from an obstruction of the Blood flow through the Liver.

Postnatal Energy (Postheaven Qi) - sometimes called the Acquired Qi, it is the Energy acquired after birth from food, drink, and air.

Postnatal Essence (Postheaven Jing) - sometimes called the Acquired Essence, it is the Essence acquired after birth from food, air and drink.

Postnatal Spirit (Zhi Shen) - also called the body's Mental Spirit, Acquired Spirit, and Conscious Spirit. This spiritual essence is acquired after birth through the refinement of one's Qi.

Post Traumatic Stress Disorder - characterized by the reexperiencing of an extremely traumatic event or events, accompanied by symptoms of increased arousal, and by avoidance of stimuli associated with the traumas. This includes the general numbing of the patient's emotional responsiveness.

Prenatal Energy (Yuan Qi) - sometimes referred to as Congenital Qi, Pre-Heaven Qi, Inherited Qi, Source Qi, Ancestral Qi, Primordial Qi, Genuine Qi, and Kidney Qi. It is Energy existing before the fetus is born, acquired from the mother's, father's, environmental and universal energies, and sustained through prayer, meditation and sleep.

Prenatal Essence (Yuan Jing)- also called Preheavenly Essence, Original Essence, Inherited Essence, Congenital Essence, Primordial Essence, and Ancestral Essence. It is the Original Essence existing before the fetus is born, acquired from the mother and father.

Prenatal Spirit (Yuan Shen) - also called the Intuitive Spirit, Perceptual Spirit, Primordial Spirit, Congenital Spirit, and the Original Subconscious. It is the Spirit essence existing before the fetus is born, acquired from fusing the mother's, father's, environmental and universal energies. The Prenatal Spirit also relates to the individual's ability to perceive and intuit information.

Prescriptions - directions given to the patient with regard to the manner of Medical Qigong exercises and meditations that must be practise after the initial Medical Qigong treatment.

Primal Senses - pertaining to the gross physical, animalistic survival senses (seeing, hearing, feeling, smelling, etc.).

Primary Channels - the body's twelve main channels, containing six Yin and six Yang rivers of Energy.

Primary Posture - the main posture, in a series of Medical Qigong prescriptions, that the patient focuses on.

Primitive Unconsciousness (Lower Unconsciousness) - pertaining to the acquired mind and related to the primal senses.

Projection - the attribution of unacceptable impulses within oneself to other people.

Proliferative Arthritis - the rapid reproduction and growth of arthritis.

Prostatitis - the inflammation of the prostate.

Protective Qi (Wei Qi) - the body's external field of defensive, protective energy (divided into three external fields of Qi).

Pruritus - severe itching.

Psychogenic - a condition developed from the beliefs originating within the mind.

Psychogenic Polyuria - pertaining to the belief that one must frequently secrete and discharge urine.

Psychometry - the act of sensing the thoughts, images and so on, with which the object has been imprinted.

Psychoneurosis - emotional disfunction caused from unresolved unconscious conflicts.

Psychosexual Qi Deviation - a condition resulting from an immediate energetic tissue overstimulation of the sexual organs. Patients with this condition experience intense sexual undulations and orgasms when being treated in a safe clinical environment.

Psychosomatic - pertaining to the relationship between the physical tissues and the emotions.

Pulling Down the Heavens - an opening and closing meditation used to energize and clear the body from the top of the head to the bottom of the feet, with breath, mind and hand movements.

Pulmonary Emphysema - a chronic disease of the lungs characterized by a destructive increase in the normal size of air spaces distal to the terminal bronchiole.

Purpura - a condition characterized by hemorrhages of the internal organs, skin, mucous membranes and other tissues, with various manifestations and diverse causes.

Purgation (Purging) - technique used in order to reduce Excess and expel pathogenic Evils located within the energetic fields and tissues of the body.

Pyelonephritis - the inflammation of the Kidneys and pelvis.

Q

Qi - is the energetic medium existing between matter and spirit (also known as Life Force Energy, when pertaining to the physical body).

Qi Collapse - pertaining to the complete absence (void) of either Yin or Yang Qi.

Qi Compression - using the Qi to press the tissues.

Qi Deviations - an alteration of energetic patterns and flow of energy that affects the body, mind, emotion and spirit, resulting in disease.

Qi Dysfunction - the impaired or abnormal function of the body's energy.

Qi Extension - the emission of energy from the body.

Qigong - Energy Skill, pertaining to exercises that cultivate Life Force Energy. There are three primary schools of Qigong training - Martial, Medical and Spiritual.

Qigong Clinic - a facility for diagnosis and treatment of outpatients with Medical Qigong therapy.

Qigong Doctor - in China, a person who medically treats patients for mental or physical disorders using Qi.

Qigong Massage - soft tissue regulation wherein the Doctor's hand lightly skims the patient's body. This gentle surface tissue stimulation is used to energize, stimulate or dredge the patient's Wei Qi fields. It is used with purging and tonifying techniques.

Qigong Therapy - one of the four branches of Traditional Chinese Medicine, which involves treating patients through the use of Energetic Point Therapy, Qigong Massage, Distance Therapy, Self-Regulation Therapy, and Invisible Needle Therapy, to stimulate energy flow.

Qigong Therapist - in America, a person who medically treats patients for mental or physical disorders using Qi.

Qi Hai - Sea of Qi point (CV-6).

Qi Manipulations - techniques used to treat or influence the flow of energy in the body.

Qing Dai - green-blue vaginal discharge.

Qi Projection (Qi Extension) - the emission of energy from the body.

Qi Regulation - energetically balancing the action or functions of the body's Yin and Yang energies.

Qi Stasis - the total stagnation of energy.

Quiescent - a meditative state wherein the individual's mind and body becomes quiet and peaceful.

R

Rachialgia - spinal inflammation.

Reaction-Formation - the conversion of one feeling into its opposite, typically seen in love turning into hate.

Rebellious Qi - energy that does not follow the correct flow or current, acting recklessly.

Reconstructive Qi Therapy - pertaining to the reconstruction and energizing of the body's energetic fields and organ systems (especially after surgery).

Reducing Qi - to lessen or decrease an organ or channel's energy.

Reflexology - the skill of pressing specific areas of the body's hands and feet to initiate internal energetic movement.

Regression - the return to an earlier childhood stage of behavior to reduce the demands on the ego.

Regulating - pertaining to the balancing of the body's Yin and Yang Energies.

Reinforcing Qi - to strengthen and support the body's organ or channel Energy.

Repression - the pushing down of unwanted ideas and emotions into the unconscious.

Respiratory Qi - Energy of the chest.

Restrictive Cycle - pertaining to the Five Elemental Controlling Cycle, where one organ restricts the energy of another organ (as depicted in the pentagram drawing).

Retrobulbar Neuritis - inflammation of the nerves behind the eyeball.

Returning To The Origin - see Rooting the Lower Dantian.

Reverse Breathing - opposite of abdominal breathing, wherein the patient will contract the abdomen with the inhalation, and expand the abdomen with the exhalation.

Rhabdomyosarcoma - a sarcoma of the muscles.

Rheumatic - pertaining to an rheumatism (a general term used to describe an acute or chronic condition characterized by inflammation, soreness and stiffness of the muscles, and pain in the joints and associated structures.

Rheumatoid Arthritis - a form of arthritis, characterized by inflammation of the joints, swelling, stiffness, cartilaginous hypertrophy, and pain.

Rheumatoid Spondylitis - a chronic, progressive disease, characterized by inflammation of the joints between the articular processes, costovertebral joints, and sacroiliac joints.

Rheumatosis - an acute or chronic condition characterized by inflammation, soreness and stiffness of the muscles, and pain in the joints.

Rhinitis - the inflammation of the nasal mucosa.

Righteous Qi (Zheng Qi) - is also called Upright Qi and Correct Qi. It is energy that heals the body and fights diseases.

Rigor - a sudden, chill with high temperature, followed by heat and profuse perspiration. Can also be referred to a state of hardness and stiffness, as in the muscles.

Root - the original cause of a disease; or to energetically secure into the Earth by extending the body's Energy deep into the ground, as if growing tree roots.

Rooting - the process of extending the body's Qi into the Earth to either establish a solid energetic foundation, or if need be, disperse Toxic Qi.

Rooting the Lower Dantian (Returning to the Origin) - returning the body's collected Qi back into the Lower Dantian.

S

San Bao (Three Treasures of Man) - pertaining to Jing (Essence), Qi (Energy) and Shen (Spirit).

San Jiao (Triple Burners) - corresponding to three main body cavities, responsible for heating the body and transporting the Body Fluids.

Sarcoma - a malignant growth, or tumor, that occurs within the connective or mesenchymal tissue. It may affect the muscles, bones, fat, blood vessels, lymph system, Kidneys, Bladder, Liver, Lungs, Spleen, and/or parotid glands.

Schizophrenia - a mental disorder, that induces hallucinations - usually auditory - through can also be visual, accompanied by very disordered thinking, delusions, disorganized speech, irrational or catatonic behavior, such as stupor, rigidity, or flaccid movement of the limbs. The ability to interact with others is greatly impaired.

Sclera - a tough white fibrous tissue that covers the white of the eyes.

Sea of Blood (Sea of the Twelve Channels) - pertaining to the Energy located in the Thrusting Vessel.

Sea of Energy - Energy located in the Lower Dantian, or Qi Hai area.

Sea of Grain and Water (Sea of Nourishment) - pertaining to the Energy located in the Stomach.

Sea of Marrow - pertaining to the Energy flowing in the spinal column and brain, originating from the Kidneys.

Sea of Qi - the chest center. Some Medical Qigong schools maintain that there are two reservoirs of Qi: the Middle Dantian, being the Sea of Postnatal Qi, and the Lower Dantian, being the Sea of Prenatal Qi (which is regulated by the Qihai CV-6 point).

Sea of Yang Channels - pertaining to the Governing Vessel.

Sea of Yin Channels - pertaining to the Conception Vessel.

Self Regulation Therapy - pertaining to the patient's Qigong prescriptions (meditations and/or exercises).

Seven Emotions - see Seven Internal Factors.

Seven Essential Stars - the Sun, Moon, Mars, Venus, Mercury, Saturn and Jupiter, associated with the body's seven orifices.

Seven Internal Factors - pertaining to the seven emotional pathogenic factors that cause disease, when in an Excess condition (Joy, Sorrow, Worry, Grief, Fear, Fright, and Anger).

Seven Material Souls - pertaining to the seven Earthly spirits that reside in the body as the Po.

Seven Orifices - ears, eyes, nostrils, mouth, anus and urethra, which are considered the gates and windows of Essence, Energy and Spirit.

Seven Turbid Demon Natures - see Po.

Shaman - an ancient Tungus term meaning "between the worlds." A Shaman is a tribal priest or priestess who heals the physical, mental, emotional, energetic and spiritual aspects of the patient.

Shao Yang - Small Yang, also called Lesser Yang, Minor Yang, or Young Yang, is affiliated with the sunrise and the waxing-moon phase. Modern physicists associate the Lesser Yang with a light force and electromagnetism.

Shao Yin - Small Yin, also called Lesser Yin, Minor Yin, or Young Yin, is affiliated with the sunset and the waning-moon phase. Modern physicists associate the Lesser Yin with a heavy force, and gravity.

Shen - meaning Spirit; when speaking about physical development, it is derived from Qi, and can be divided into both Prenatal and Postnatal Shen.

Shen Deviations - mental and emotional disorders which have caused the Three Ethereal Souls (Hun) to leave the patient's body.

Shengong - training of the spirit through meditation and visualization.

Shening Out - terminology used to describe the Ethereal Soul (Hun) wandering away from the body.

Shi Qi (Turbid Qi) - also known as Evil Qi, Toxic Qi, and Pathogenic Qi, it is coarse, unrefined, polluted or dirty energy.

Sishencong (Four Alert Spirit) Points - four points at the top of the head (that surround the Baihui Point) used to absorb Heavenly Qi into the body.

Shou Zhen (Hand Diagnosis) - a form of diagnosis, wherein, the doctor assess the "energetic blueprint" of the patient's body transformed onto the doctor's left hand.

Shu Points - five specific points below the elbows and knees identified as the Well, Spring, Stream, River and Sea points. Each point has an affect on the quantity of the energy of an organ.

Sishencong (Four Spirit Hearings) Points - a group of four points located at the top of the head (surrounding the Baihui point), used for absorbing Heavenly Qi into the body's Taiji Pole.

Six External Factors - also known as the Six Pernicious Influences, these factors pertain to the six climatic changes (Wind, Summer Heat, Heat, Damp, Dryness, Cold, and Fire).

Six Storage Areas- body's Yang organs constantly fill and empty, and include the Bladder, Gall Bladder, Stomach, Large Intestine, Small Intestine, and Triple Burners.

Skin Zones - twelve dermal-zones, based upon the surface location of the body's Twelve Primary Channels.

Soaring Dragon Technique - hand technique for Qi emission, where the energy is emitted through the middle finger bent and pointing downward, while the other fingers are extended straight outwards.

Solid Organs - the body's Yin organs, which include the Liver, Heart, Spleen, Lungs, and Kidneys (also included in this list is the Pericardium).

Soul - immaterial Spiritual Essence of an individual's life, stored within the Heart and Middle Dantian.

Soul Body- see Astral Body.

Soul Extensions - the Shen develops and contains Twelve Soul Extensions. These Twelve Soul Extensions contain the body's different personality characteristics.

Soul Loss - the loss of parts of the Eternal Soul.

Soul Retrieval - to spiritually search for and bring back one's forgotten memories (soul), which have been isolated from consciousness due to trauma and shock.

Soul Travel - see Astral Travel.

Sound Energy Therapy - sound projected as audible and inaudible tone resonation, used for healing.

Sound Resonation - healing tones used for tonifying or dispersing the patient's Energy.

Spider Nevus - a branched growth of dilated capillaries on the skin, that resemble a spider.

Spinal Pass (Jia Ji Guan) - two points located on the lateral sides of the Mingmen (GV-4), where energy has a potential to stagnate.

Spinous Process - the single midline posterior projection arising at the junction of each vertebra.

Spiraling Energy Technique - hand manipulation, that extends and spirals the doctor's projected energy.

Spirit - the energetic manifestation of the Eternal Soul.

Spirit Body - the energetic vehicle inwhich the body's Shen can travel throughout the Astral Plane. The Spirit Body can manifest through many forms (Body of Light, animal forms, etc.).

Spirit Demons - see Demon Possession and Oppression.

Spirit Soul - the Three Ethereal Souls, accompanied by the individual's consciousness, acting as one unit for spirit travel.

Spirit Travel - the spirit (Hun) journeying outside of the physical body.

Splenomegaly - the enlargement of the Spleen.

Squamous Metaplasia - the conversion of tissue into a form of scalelike cells, that is abnormal for that tissue.

Stacking the Bones - allowing the bones to stack upon each other from the bottom of the feet to the top of the head.

Stagnation - not moving, inactive; pertaining to Qi, Blood, or thought patterns.

Static Qigong - the process of stationary, quiescent Energy gathering.

Stroke (Wind Stroke) - caused by the buildup of Excess Liver Fire creating Internal Wind.

This Internal Wind causes Qi and Blood to rebel upwards causing Phlegm to form and obstruct the cavities and vessels, creating Penetrating Wind or Stroke.

Subarachnoid Hemorrhage - bleeding internally, within the spaces at the base of the brain, between the pia proper and arachnoid contain the cerebrospinal fluid.

Subconscious Mind - part of the mind associated with the recording and storing of personal interpretations of reality (not readily accessible to the conscious mind).

Sublimation - the channeling of unacceptable impulses into acceptable, refined social forms and is the only defence mechanism considered to be a healthy reaction.

Substances - pertaining to the body's essential parts of physical and energetic material.

Sui - Marrow.

Super Ego - Dr. Sigmund Freud's terminology for the division of the psyche in psychoanalytic theory, responsible for the psychic reward and punishment system.

Sun's Essence - energy gathered from the sun, ingested as warm light.

Sword Fingers Technique - hand manipulation that emits Qi through the extended index and middle fingers.

Symptoms - a subjective manifestation of a pathological condition, reported by the patient.

Syndromes - a grouping of signs and symptoms, based on their frequent reoccurrence, that may suggest a common underlying pathogenesis.

Systemic - affecting the entire body.

T

Taiji Pole - the Center Core of light which joins the body's three Dantians and the Eternal Soul together.

Tai Yang - Great Yang, also called Strong Yang, Major Yang, or Old Yang, is affiliated with high noon and the full-moon phase. Modern physicists associate the Strong Yang with a strong nuclear force.

Tai Yi - meaning Great Divinity or God.

Tai Yin - Great Yin, also called Strong Yin, Major Yin, or Old Yin, is affiliated with midnight and the new-moon phase. Modern physicists associate the Great Yin with a weak nuclear force.

Ten Heavenly Stems - the ten energies of Heaven that rule the changes of the Five Elemental seasonal transitions, and are represented in the human body as the Yin and Yang aspect of the Five Elements (represented in the human body as the ten major internal organs).

Ten Thousand Voices - pertaining to the state of open receptivity of the Qigong doctor's Heart, after rooting the mind.

Ten Wings - consisting of ten commentaries from Confucius and his disciples, pertaining to the study of the eight trigrams, sixty-four hexagrams, and the Yi-Jing.

Third Eye Point (Yin Tang) - located in the center of the forehead, responsible for spiritual intuition and communication.

Thought-forms - images of concentrated thought patterns that manifest on the vibrational resonance of the Astral Plane.

Three Ethereal Souls (Hun) - pertaining to the three heavenly spirits that reside in the body.

Three Fires - the heat in the body, generated from the energy radiating from the Heart Fire, Kidney Fire, and Bladder Fire.

Three Outer Forces - pertaining to the three natural powers of Heaven, Earth and Man.

Three Parts Wisdom - knowledge obtained through the doctor's connection and communication with his or her Upper, Middle, and Lower Dantians.

Three Periods of Life - the developmental stages of the patient's Jing, Qi and Shen divided into the womb, childhood and adulthood.

Three Stars - pertaining to the three periods of life, each period is divided into three stages of development, known as the three stars.

Three Treasures of Earth - pertaining to the energy of Soil, Water and Wind, and the study of Feng Shui (Wind and Water).

Three Treasures of Heaven - pertaining to the energy of the sun, moon and stars, and the study of Chinese astrology.

Three Treasures of Man - pertaining to the energetic interaction of the body's Essence, Energy and Spirit, and the study of the Yi-Jing (I-Ching or Book of Changes).

Three Wonders - Clinical manifestations of Qi, categorized as Subtle, Mysterious and Incredible Wonders.

Thrombosis - the formation and development or existence of a Blood clot (thrombus) within the walls of the vascular system.

Thrusting Channels - the Five Energy Channels which surround and penetrate the body's center core via the Taiji Pole.

Thrusting Vessels - also called the Chong Vessels, they are the Five Energy Vessels which originate from the center of the body and internally transverse the legs and torso.

Ti - referred to as the Divine Center.

Tian Qi (Heavenly Energy) - the transformed energy of the Yuan Qi and the divine.

Tian Shen (Heavenly Spirit) - the transformed energy of the Yuan Shen and the divine.

Tie Bi (Iron Wall) - the areas of the body where it is most difficult for the energy to pass through when circulating the Microcosmic Orbit.

Tinnitus - a ringing, tinkling, or buzzing sound in the ear.

Toe Raised Stepping - pertaining to energetic walking therapy, wherein the toes are stretched when stepping in order to facilitate the increase of Qi flowing into the body via the leg channels.

Tonification (Tonify) - to supplement the insufficiency and strengthen the body's resistance.

Traditional Chinese Medicine - Chinese Energetic Medicine, divided into four branches of healing modalities (Acupuncture, Herbal Therapy, Medical Qigong Therapy, and Tissue Regulation Therapy (Chinese Massage).

Transference - the process whereby a patient unconsciously transfers feelings, thoughts, beliefs and patterns of behavior that had been previously experienced with others onto the doctor.

Transient Ischemic Attacks (TIA) - temporary interference with the Blood supply to the brain.

Treatment - the medical care given to a specific condition.

Trigger Points (Ashi Points) - places on the body which are tender spots, or painful areas near diseased or injured tissue.

Trigram - pertains to three Yao lines stacked upon one another forming a specific symbol, which represents certain characteristics.

Triple Burners (San Jiao) - also known as the Triple Heaters and Triple Warmers, they correspond to three main body cavities (perineum to navel, navel to base of solar plexus, solar plexus to throat), and are responsible for heating the body and transporting Body Fluids.

True Fire - the original Heat or Fire Energy that regulates the body's Yin and Yang Qi, created from the radiating energy of the Heart's Fire, Kidneys' Fire and Bladder's Fire.

True Nature - one's innate nature in harmony with life.

True Qi - the energy that circulates in the body's channels and collaterals which nourishes the Yin and Yang organs and fights disease.

True Self - one's true nature, connected to the subconscious mind.

True Spirit - pertains to the spiritual nature of the True Self. The Hun and Po are expressions of the body's True Spirit.

Tsou Hou Ru Mo ("the Spirit leaves and the Demon enters") - describes self induced psychosis, pertaining to improper Qigong training, wherein the patient's Hun leave the body and the Po take over.

Tui Na Therapy - a tissue manipulation therapy that focuses on the adjustment and / or stimulation of the muscles and tendons.

Tumor - an abnormal growth, either benign or malignant, caused by a retention of mass due to the cultivation and stasis of Qi, Blood, Phlegm, etc.

Turbid Qi - also called Evil Qi, is coarse, unrefined, polluted, and dirty energy.

Twelve Pi Hexagrams - the twelve symbols pertaining to the twelve time periods of the day and year.

Twelve Primary Channels - the body's twelve main energetic rivers (Liver, Lungs, Large Intestine, Stomach, Spleen, Heart, Small Intestine, Bladder, Kidneys, Pericardium, Triple Burners, and Gall Bladder).

Twelve Earthly Branches - twelve energies of the Earth that determine the six Qi factors of the seasonal transitions (represented in the human body as the Twelve Primary Channels).

Two Breathings - pertaining to the abdominal breathing method of holding the breath.

U

Umbilications - a depression resembling a navel.

Universal Qi - energy pertaining to the Heavens, the divine and the celestial influences.

Upper Burner - pertaining to the body's complex system of Fluid distribution via the Lungs and located within the upper chest cavity.

Upper Dantian - area within the center of the head, attributed as the body's chamber of light and door to psychic and intuitive powers. The Upper Dantian is also known as Seal Palace, Ancestral Opening, Calm Fountain, Heaven's Valley, Inner Source, and Clay Pill Palace.

V

Vasculitis - the inflammation of a blood or lymph vessel.

Vertigo - the sensation of moving in space, resulting in such symptoms as dizziness and light-headedness.

Virtue (De) - pertaining to the function of the divine in man.

Virtue of Dao - pertaining to the commendable quality of the divine.

Viscera - the body's internal organs.

Void - also called Wuji, it pertains to the infinite space between matter and energy.

W

Wai Dan Shu - external elixir cultivation, that focuses on cultivating Qi from outside the individual's body.

Wai Qi - external, extended energy.

Walking Therapy - Postoral Dao Yin walking exercises and dynamic "moving" meditations used for the treatment of organ Deficiencies.

Wandering Bi - migrating pain within the body's cavities.

Waning - to grow smaller.

Water Element - one of the Five Elements, pertaining to Kidneys and Bladder.

Water Jing - energy that controls the genetic development phase of the fourth fetal month.

Waxing - to grow larger.

Wei Lu Guan (Coccyx Pass) - located on the lowest segment of the spine just posterior to the anus, near the Chang Qiang (GV-1) point.

Wei Qi - the body's external field of Defensive and Protective energy (divided into three fields of Qi).

Wen Huo - pertaining to the gentle breathing method of Respiratory Dao Yin training.

White Blood Cell - any of a group of Blood cells that have no hemoglobin and migrate into tissues to fight infection and digest cell debris.

Wind Bi - pain in the body created by toxic Wind invasion.

Wind Stroke - Stroke caused by the buildup of Excess Liver Fire creating Internal Wind. This Internal Wind causes Qi and Blood to rebel upwards causing Phlegm to form and obstruct the cavities and vessels, thus creating Penetrating Wind or Stroke.

Windy Breathing Method - pertaining to the method of breathing through the nose.

Wood Element - one of the Five Elements, pertaining to the Liver and Gall Bladder.

Wood Jing - energy that controls the development phase of the direction of the fetus's emotional and spiritual aspects during the seventh month of pregnancy.

Wu Guan (Five Passes) - five important gates on the Governing Vessel located at the coccyx, Mingmen, Shendao, occiput, and Baihui areas where energy tends to stagnate.

Wu Huo - pertaining to the vigorous breathing method of Respiratory Dao Yin training.

Wuji - pertaining to infinite space or the formless Void.

Wuji Posture - a quiet standing posture used in meditation to allow the practitioner to return to a state of tranquility.

Wu Jing Shen (Five Essence Spirits) - the spiritual energy radiating from the core of the Five Yin Organs. Combined, these energies create the foundation of the body's Shen (Spirit).

Wu Se Dai - pertaining to the five colors of vaginal discharge - white, yellow, red, green-blue, and dark brown or black.

Wu Wei - a state of "no mind," i.e., no thoughts.

Wu Zang - the Five Yin Organs. Wu translates to mean "five," Zang translates to mean "to store or hold."

X

Xie Qi (Evil Qi) - energy that causes disease or harmful effects to the body.

Xin Xi (The Message) - knowledge stored within the Wuji or the Void.

Xiphoid Process - the lowest part of the sternum bone (sometimes referred to as the Doves Tail).

Xue - Blood.

Y

Yang - the positive charged energetic polarity, opposite of its companion Yin, pertaining to man, hard, light, hot, etc.

Yang Channels - the body's Yang energetic rivers, consisting of the Governing Vessel, Belt Vessel, Yang Linking Vessels, Yang Heel Vessels, Large Intestine Channels, Triple Burner Channels, Small Intestine Channels, Stomach Channels, Gall Bladder Channels, and Bladder Channels.

Yang Fire - also called Emperor's Fire, energy of the Heart Fire.

Yang (Fu) Organs - also known as Hollow Organs, that consist of the Gall Bladder, Small Intestine, Stomach, Large Intestine, Bladder. Also included is this category are the Triple Burners.

Yang Shen Disturbances - an emotional Yang state of energetic dysfunction.

Yang Ming (Yang Brightness) - indicates Yang Qi developing its final stage and then reverting into Yin.

Yao - a solid or broken line which is representative of either Yang or Yin energy, used in combination of three as Trigrams or six as Hexagrams.

Yao Cycles - the progression of twelve hexagrams (six Yang and six Yin) flowing in a waxing and waning cycle.

Ye (humor) - thick, turbid Body Fluids.

Yellow Court - located in the center of the diaphragm, just below the xiphoid process of the sternum. Its function is that of being the access point to releasing the body's internal organ emotional memories. Its location is also attributed to the 3rd Chakra.

Yi - the intention or thought (the cognitive mind).

Yi Jing - Chinese "Book of Changes," pertaining to the natural transitions of life.

Yin - the negative charged energetic polarity, opposite of its companion Yang, pertaining to woman, soft, dark, cold, etc.

Yin Channels - Yin energetic rivers, consisting of the Conception Vessel, Thrusting Vessel, Yin Linking Vessels, Yin Heel Vessels, Lung Channels, Pericardium Channels, Heart Channels, Spleen Channels, Liver Channels, and Kidney Channels.

Yin (Zang) Organs - also known as the Solid Organs, that consist of the Liver, Heart, Spleen, Lungs and Kidneys. Also included in this category is the Pericardium.

Yin Shen Disturbances - an emotional Yin state of energetic dysfunction.

Yin Tang (Third Eye Point) - located in the center of the forehead, responsible for projecting the Spirit for psychic intuition and communication.

Ying Qi (Nutritive Qi) - the body's nourishing energy.

Yu (Surplus) Vessels - secondary vessels that branch away from the energetic flow of the major Linking Vessels (at the chest and back), connecting the Linking Vessels energetic flow to the hands.

Yuan Jing (Original Essence) - the Original Kidney or Prenatal Essence.

Yuan Shen (Original Spirit) - the Original Prenatal Spirit.

Yuan Qi (Original Energy) - the Original Kidney or Prenatal Qi.

Yun - the Yin method of dynamic postural Dao Yin training.

Yu Zhen Guan (Occipital Pass) - the area located just inferior to the occipital bone where the brain originates, known as a specific point where Qi often stagnates.

Z

Zang Organs - Yin or solid organs (Liver, Heart, Spleen, Lungs, Kidneys and Pericardium).

Zang/Fu Organs - the body's Yin and Yang organs.

Zhang Xiang Xue Shou - in Chinese medical science, the study of energetic physiology.

Zhen Qi - see True Qi

Zheng Qi - Righteous Qi, pathogenic fighting Energy.

Zhi - the Will power, mental drive and determination.

Zhong Qi - Center Qi, Energy of the chest.

Zhou Qi - Turbid Qi, Evil Qi, Impure Qi

Zong Qi - Gathering Qi, and / or Respiratory Qi.

Zygomatic Facial Regions - pertaining to the sides of the cheeks below the eyes.

BIBLIOGRAPHY

1. Anodea Judith
   Wheels of Life
   Liewellyn Publications, 1990
   St. Paul, Minnesota 55164-0383
2. *A Barefoot Doctor's Manual*
   A Guide To Traditional Chinese and Modern Medicine
   A Cloudburst Press Book
   Madrona Publishers
   Seattle, Washington 1977
3. Robert O. Becker, M.D.
   Cross Currents
   Jeremy P. Tarcher, Inc., 1990
   Los Angeles, Ca. 90036
4. Harriet Beinfield, L.Ac. and
   Efrem Korngold, L.Ac., O.M.D.
   Between Heaven And Earth
   A Guide To Chinese Medicine
   Ballantine Books, NY 1992
5. Barbara Ann Brennan
   Hands of Light
   A Guide to Healing Through the Human Energy Field
   Bantam Books Publishing 1988
   New York, New York
6. Paul Carus
   Chinese Astrology
   Early Chinese Occultism
   Open Court Publishing 1989
   La Salle, Illinois
7. Dr. Chen Yan-Feng
   Pre-Natal Energy, Mobilizing Qigong
   China Taoist Ancient Qigong
   Chinese Kung Fu Series
   Guandong Xin hua Printing House 1992
8. *Chinese Qigong*
   Publishing House of Shanghai
   College of Traditional Chinese Medicine
   China 1990
9. *Chinese Qigong*
   Outgoing Qi Therapy
   Shandong Science and Technology Press
   China 1992
10. *Chinese Qigong Therapy*
    Shandong Science and Technology Press
    China 1995
11. Dr. Choa Kok Sui
    Pranic Healing
    Samuel Weiser, Inc. 1990
12. Kenneth S. Cohen
    The Way of Qigong
    Ballantine Books 1997
13. Peter Deadman
    Selected Articles from the
    Journal of Chinese Medicine
    England
14. Peter Deadman & Maazin Al-Khafaji
    with Kevin Baker
    A Manual of Acupuncture
    Journal of Chinese Medicine Publications
    England 1998
15. Luc DeSchepper MD, PhD,CA
    Acupuncture for the practitioner
    Asnta Monica, California 91403
16. Vincent T. DeVita, Jr., Samuel Hellman,
    Steven A. Rosenberg
    Cancer Principles and Practice of Oncology
    Fifth Edition - Vol 1
    Lippincott - Raven Publishers
    Philadelphia - New York 1997

17. Vincent T. DeVita, Jr., Samuel Hellman, Steven A. Rosenberg
Cancer Principles and Practice of Oncology
Fifth Edition - Vol 2
Lippincott - Raven Publishers
Philadelphia - New York 1997

18. Paul Dong and Thomas Raffill
Empty Force
Element Books, Inc., 1996
Rockport, Mass. 01966

19. *Essentials of Chinese Acupuncture*
Beijing College of Traditional Chinese Medicine
Foreign Languages Press, 1980
Beijing, China

20. Richard Feit and Paul Zmiewski
Acumoxa Therapy
A Reference and Study Guide
Volume II The Treatment of Diseases
Paradigm Publications
Brookline, Massachusetts 1990

21. Richard Gerber, M.D.
Vibrational Medicine
Bear and Company, 1988
Santa Fe, New Mexico 87504-2860

22. Stanislav Grof,M.D. with Hal Zina Bennett
The Holotropic Mind
Harper Collins Publishers, Inc. 1993
New York City, New York

23. Rosemary Ellen Guiley
Harper's Encyclopedia of
Mystical and Paranormal Experiences
Harper Collins Publishers Inc. 1991
New York City, New York

24. Dr. Guo Liang Cao
Essentials Of Tuinaology
Chinese Medical Massage and Manipulation
Cao's Fire Dragon
Hilo, Hawaii 1983

25. Leon Hammer, M.D.
Dragon Rises and Red Bird Flies
Psychology and Chinese Medicine
Staton Hills Press, Inc.
Barrytown, New York 1990

26. *Illustration of Channels and Points For Acupuncture and Moxibustion and Qigong*
Hunan Science and Technology Press
China 1992

27. Roger Jahnke D.O.M.
The Healer Within
Harper Publishing
San Francisco 1997

28. Dr. Jiao Guo Rui
Qigong Essentials for Health Promotion
China Reconstruction Press 1990

29. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Chi Kung Correspondence Course, Vol. 1-6
Pacific Grove, CA 1988

30. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
The Essence of Internal Martial Arts, Vol. 1
Esoteric Fighting Techniques and
Healing Methods
Pacific Grove, CA 1994

31. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
The Essence of Internal Martial Arts, Vol. 2
Energy Theory and Cultivation
Pacific Grove, CA 1994

32. Ted J. Kaptchuk, O.M.D.
The Web That Has No Weaver
Congdon and Weed, Inc., 1983
New York, New York, 10001

33. Dr. Arnie Lade, DOM
Acupuncture Points
Images & Functions
Eastland Press 1989

34. Claude Larre and Elisabeth Rochat de la Vallee
The Seven Emotions
Psychology and Health in Ancient China
Monkey Press
Cambridge, CB4 3PU 1996

35. Dr. Mariam Lee, D.O.M.
Insights of a Senior Acupuncturist
Blue Poppy Press
Boulder, Colorado 1992

36. Richard H. Lee
Qi and Kirlian Photography
China Healthways Institute, 1996
San Clemente, Ca 92672

37. Master Shou-Yu Liang and Wen-Ching Wu
Qigong Empowerment
A Guide to Medical, Taoist, Buddhist
and Wushu energy Cultivation
The Way of Dragon Publishing, 1997
East Providence, Rhode Island 02914-0561

38. Dr. Lin Huo Sheng & Dr. Luo Pei Yu
Three Hundred Questions on
Qigong Exercises
Guandong Science and Technology Press
China 1994

39. Dr. Hong Liu with Paul Perry
Mastering Miracles
Warner Books Inc., 1997
New York, New York, 10020

40. Liu Zheng-Cai and Hua Ka
A Study of Daoist Acupuncture
Blue Poppy Press
Boulder, Colorado 1999

41. Livia Kohn & Yoshinobu Sakade
Taoist Meditation and
Longevity Techniques
Center for Chinese Studies
University of Michigan 1989

42. Ke Yun Lu
The Essence of Qigong
A Handbook of Qigong Therapy
and Practice
Abode of the Eternal Tao
Eugene, Or. 1998

43. Giovanni Maciocia
Tongue Diagnosis in Chinese Medicine
Eastland Press Inc., 1987
Seattle, Washington 98111

44. Giovanni Maciocia
Obstetrics & Gynecology in
Chinese Medicine
Churchill Livingstone Inc., 1998
New York City, New York, 10011

45. Giovanni Maciocia
The Foundations of Chinese Medicine
Churchill Livingstone Inc., 1989
New York City, New York, 10011

46. Kiiko Matsumoto & Stephen Birch
Extraordinary Vessels
Paradigm Publications
Brookline, Massachusetts 1986

47. Kiiko Matsumoto and Stephen Birch
Hara Diagnosis: Reflections on the Sea
Paradigm Publications, 1988
Brookline, Massachusetts 02146

48. Hugh Milne
The Heart of Listening
A Visionary Approach to Craniosacral Work
North Atlantic Books
Berkeley, California 1995

49. Joseph Needham,F.R.S.,F.B.A.
Science and Civilization in China
Volume 5 Chemistry and
Chemical Technology
Part V: Spagyrical Discovery and Invention:
Physiological Alchemy
Cambridge University Press 1986
Cambridge, London

50. Master Ni Hua Ching
The Book of Changes and
the Unchanging Truth
Malibu, CA 1983

51. John O'Connor and Dan Bensky
Acupuncture, A Comprehensive Text
Shanghai College of Traditional Medicine
Eastland Press
Chicago, Illinois 1981

52. Yoshiaki Omura, Sc.D.,M.D.
Acupuncture Medicine
Japan Publications, Inc. 1982
Tokyo, Japan

53. Pan Mingji
How to Discover Cancer Through Self-Examination
Fujian Science and Technology Publishing House
Beijing, China 1992

54. *Point Therapy*
A Chinese Traditional Therapeutic Skill
Shandong Science and Technology Press 1990

55. Daniel Reid
Harnessing the Power of the Universe
A Complete Guide to the Principles and Practice of Chi-Gung
Shambhala Publications, Inc.
Boston, Massachusetts 1998

56. D. Scott Rago
Leaving the Body
A Complete Guide to Astral Projection
Prentice-Hall, Inc., 1983
Englewood Cliffs, New Jersey

57. Elisabeth Kubler-Ross,M.D.
On Death and Dying
Collier Books, 1969
Macmillan Pub. Co., New York

58. Jeremy Ross, C.Ac. (Nanjing), B.Ac. (M.B.Ac.A.), C.Ed.
Zang Fu,The Organ Systems of Traditional Chinese Medicine
Functions, Interrelationships and Patterns of Disharmony in Theory and Practice
Churchill Livingstone
Edinburgh, London, Melbourne and New York 1985

59. Stephano Sabetti
The Wholeness Principle
Exploring Life Energy Process
Life Energy Media 1986
Sherman Oaks, California

60. Mikio Sankey Ph.D., L.Ac.
Esoteric Acupuncture
Gateway to Expanded Healing Vol.1
Mountain Castle Publishing
Los Angeles, Ca. 1999

61. Jack Schwark
Human Energy Systems
The Aletheia Foundation, 1980

62. Kylea Taylor
The Ethics of Caring
Hanford Mead Publishers 1995
Santa Cruz, California

63. *Textbook of Medical Qigong Therapy*
Haidian Qigong University
Medical Qigomg College
White Cloud Monestry, 1988
Beijing, China

64. Clayton L. Thomas, M.D., M.P.H.
Taber's Cyclopedic Medical Dictionary
F.A. Davis Company, 1993
Philadelphia, Pa., 19103

65. Michael Tierra
East West Master Course In Herbology
Santa Cruz, California 1981

66. Dr. Tzu Kuo Shih
Qigong Therapy
The Chinese Art of Healing With Energy
Station Hill Press 1994

67. Ilza Veith
The Yellow Emperor's Classic of Internal Medicine
University of California Press, 1966
Berkeley and Los Angeles, Ca.

68. Andrew Weil,M.D.
Spontaneous Healing
Fawcett Columbine
Published by Ballantine Books 1995
New York City, New York

69. Nigel Wiseman and Feng Ye
A Practical Dictionary of Chinese Medicine
Second Edition
Paradigm Publications 1998
Brookline, Massachusetts

70. Henry S. Wu, D.C.
Differentiation of Syndromes
California Acupuncture College 1982
Los Angeles, Ca.

71. Yan De-Xin
Aging & Blood Stasis
A New TCM Approach to Geriatrics
translated by Tang Guo-shun & Bob Flaws
Blue Poppy Press
Boulder, Co. 1995

72. Dr. Yang Jwing Ming, Ph.D.
The Root of Chinese Qigong
The Secrets of Qigong Training
Yang Martial Arts Association 1989

73. Yin Hui He
Fundamentals of
Traditional Chinese Medicine
Foreign Languages Press, 1995
Beijing, China

74. Yu Gong Bao
Chinese Qigong Illustrated
New World Press 1995

CLINICAL REFERENCES

1. Master John Fey
   Clinical Notes from Lectures
   Medical Qigong Therapy
   Reference: Differential Healing Modalities
   Los Angeles , CA 1984 to 1987
2. Michael Gershon, M.D.
   Professor of Anatomy and Cell Biology
   Columbia Presbyterian Medical Center
   Research presented by Sandra Blakeslee
   New York Times, New York 1996
3. Dr. Jason Gill, L.Ac.
   Lectures on Traditional Chinese Medicine
   Clinical Notes on Differential Diagnosis
   Pacific Grove, CA 1994 to 1997
4. Dr. Jason Gill
   Clinical notes and Lectures
   Reference: Dr. Jeremy Ross Lectures
   on Traditional Chinese Medicine
   Pacific Grove, CA. 1994 to 1997
5. Dr. Jiang Gui Zhong and
   Dr. Cai Chang Rong
   Clinical Notes from Hu Guo Hospital
   Medical Qigong Therapy, An Mo Therapy
   and Differential Healing Modalities
   Beijing, China 1993
6. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
   Clinical notes from the First World
   Conference on Medical Qigong
   Medical Qigong Therapy and Differential Healing Modalities
   Reference Conference No. 1, Sept. 1988
7. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
   Clinical notes from the International
   Symposium on Chinese Qigong Health
   Preservation and Qigong Techniques
   Medical Qigong Therapy and Differential Healing Modalities
   Reference Symposium, Nov. 1990
8. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
   Clinical notes from the Second World
   Conference on Medical Qigong
   Medical Qigong Therapy and Differential Healing Modalities
   Reference Conference No. 2, Sept. 1993
9. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
   Clinical notes from Xi Yuan Hospital
   on Medical Qigong Therapy and
   Differential Healing Modalities
   Internship July 1993 through Sept. 1995
10. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
    Clinical notes from Hai Dian Medical
    University on Medical Qigong Therapy and
    Differential Healing Modalities
    July 1993 through Sept. 1995
11. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
    Clinical notes from the Third World
    Conference on Medical Qigong
    Medical Qigong Therapy and Differential Healing Modalities
    Reference Conference No. 3, Sept. 1996

12. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Clinical notes from the Second World Congress on Medical Qigong
Medical Qigong Therapy and Differential Healing Modalities
San Francisco, California Nov. 1997

13. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Clinical notes from the Fourth World Conference on Medical Qigong
Medical Qigong Therapy and Differential Healing Modalities
Reference Conference No. 4, Sept. 1998

14. Jerry Alan Johnson, Ph.D., D.T.C.M., D.M.Q. (China)
Clinical Notes from International Institute of Medical Qigong
Medical Qigong Therapy and Differential Healing Modalities
Pacific Grove, CA 1988 - 1999

15. Sifu Mark Johnson
Clinical Notes in Geriatric Medical Qigong Therapy and Esoteric Daoist Studies of the Yi-Jing and Feng Shui
Mill Valley, CA 1998 - 1999

16. Dr. Lu Guo Hong, Dr. Niu Yu Hua, and Dr. Li Fu Dong
Clinical notes from Hai Dian Medical University on Medical Qigong Therapy and Differential Healing Modalities
Beijing, China 1993 - 1996

17. Dr. Lu and Dr. Shu Zong Wei
Clinical Notes from Xi Yuan Hospital
Medical Qigong Therapy and Differential Healing Modalities
Beijing, China 1993

18. Dr. Hugh Milne
Clinical Notes from Cranial-Sacral Therapy
Level 1 through 6, 1990-1995
Monterey, Calif.

19. Dr. Ren Shun Tu & Dr. Xu Hong Tao
Clinical Notes from Xi Yuan Hospital
Medical Qigong Therapy and Differential Healing Modalities
Beijing, China 1995

20. Dr. Teng Ying Bo
Clinical Notes from Beijing Medical Qigong Science and Research Institute
Medical Qigong Therapy and Differential Healing Modalities
Beijing, China 1995

21. Dr. Zheng Zhan Ding
Clinical Notes on Medical Qigong Diagnosis, Therapy and Differential Healing Modalities
Beijing, China 1993 - 1996

INDEX

A

Abdominal Breathing, 392-394, 422
Abdominal Pain, 64, 206, 208, 211, 471, 473, 565, 768, 788,793-795, 838, 910, 916, 917, 920
Abortions, 444, 445, 600, 860, 861
Absorbing Qi From Nature, 125, 126, 293, 304, 306, 307
Acquired Cellular Patterns, 115
Acquired Constitution, 787
Acquired Disorders, 461
Acquired Energy, 317
Acquired Essence, 319, 607, 720
Acquired Force, 270
Acquired Mind, 91, 201, 333, 335, 336, 138, 344, 345
Acupressure, 631, 928
Acupuncture, 49, 50, 52, 54, 56, 59, 66, 126, 128, 129, 133, 159, 164, 199, 215, 245-249, 309, 507, 522, 529, 536, 537, 557, 565, 566, 595, 611, 662, 664, 839,683, 864, 865, 882, 884, 922, 923, 924, 929
Adenocarcinoma, 689, 896, 910, 911, 913
Adrenaline, 426, 781, 900
Affirmation, 87, 89, 90, 104, 115, 411, 430, 448, 519, 554, 555, 561, 619, 861, 882, 904, 921
Akashic Records, 516
Alarm Points, 252
Alchemy, 131, 132, 249, 267, 293
Alternating Current, 128
An Jing (Hidden Power), 48
An Mo, 57, 61-65, 566, 631, 832, 890, 893
Anal Sphincter, 149, 154, 200, 342, 363, 365, 366, 393, 422, 451, 691, 728, 734, 752, 824, 868, 893, 909, 912, 917
Ancestral Channel, 197
Ancestral Qi, 137
Ancestral Traits, 274, 510, 517
Anemia, 425, 784, 828, 896, 915, 916
Anesthesia, 129, 438, 442, 447, 922, 923, 924, 925, 928
Anger, 48, 64, 75, 88, 99, 108, 138, 149, 154, 155, 168, 348, 403, 404, 413, 427, 429, 431, 439, 445-447, 452, 460, 466-470, 488, 508, 545, 560, 578, 581, 582, 594, 599, 601, 602, 615, 617, 619, 624, 626, 627, 708, 767, 769, 794, 796, 798, 828, 844, 849, 850, 852, 855, 858, 859, 860, 861, 869, 871, 872, 874, 877, 890, 891, 892, 899, 900, 905, 911, 913
Angina, 384, 394, 602, 779
Anorexia, 774, 775, 796, 819, 828, 916
Antique Shu Points, 249, 250, 251 370
Anxiety, 75, 88, 108, 123, 129, 135, 170, 174, 176, 204, 208, 214, 250, 253, 391, 394, 398, 404, 426, 432, 439, 447, 469-471, 488, 508, 602, 616, 784, 796, 880, 817, 818, 820, 828, 830, 843, 859, 874, 875, 877-882, 888, 900
Aphasia, 202, 212, 219, 220, 492, 505, 864, 890
Aplastic Anemia, 784
Archetypes, 48, 87, 334, 336
Arrhythmia, 779, 784
Arteriosclerosis, 777, 828, 865
Arthritis, 64, 384, 566, 823, 826, 827
Ascariasis, 772, 773
Ascending Qi, 201, 375, 376, 667
Asthma, 172, 174, 198, 204, 251, 360, 382, 443, 447, 464, 466, 492, 495, 497, 544, 670, 672, 731, 766, 801-803, 806-808, 812, 823, 838, 897
Astral Body, 84, 85, 348, 349, 351, 437
Astral Matter, 328
Astral Plane, 344, 348, 350, 351, 444, 516
Astral Travel, 85, 91, 328, 337, 345, 346, 435
Astrology (Chinese), 99, 268, 293, 297
Atman, Adam, 50, 54
Atrophy Syndrome (Disorder), 603, 792, 802, 867
Aura, 137, 143, 147, 155, 328, 525 529, 531, 688, 831, 925
Aura Colors, 525, 528
Auspicious Powers, 280
Autumn Equinox, 280
Awakening, 74, 150, 411, 413, 430, 513, 650

B

Ba Gan, 465
Back Pain, 154, 194, 202, 219, 421, 461, 497, 670, 812, 813
Bai Yuan, 84, 86
Baihui Point, 84, 147, 201, 229, 284, 370, 371, 431, 453, 520, 547, 573, 575, 846
Base Chakra, 151
Becker, Robert O., 128, 129, 141, 245, 246
Beijing Medical University, 274

Bellows Palm Technique, 594, 639, 659
Bells Palsy (Facial Paralysis), 866
Belt Vessel, 70, 71, 185, 197, 198, 200, 206-209, 210, 253, 260, 386, 575, 718, 728, 735, 737, 850, 851
Benign Tumors, 766, 854, 895-897, 916
Big Dipper, 135, 273, 290-294, 575, 612
Biochemical, 128, 138, 515, 685
Biorhythm, 272-276
Bitter Foods, 304, 310, 311, 466, 472, 770, 777, 851
Black, 46, 105, 247, 308, 310, 311, 327, 413, 414, 426, 431, 438, 439, 451, 452, 473, 502, 504, 505, 521, 529, 617, 624, 770, 811, 851, 852, 905, 906
Bladder, 57, 78, 82, 123, 136, 146, 154, 161, 182-185, 193, 312, 324
Bladder Channel, 71, 82, 174, 182 184, 194, 201, 210, 212, 219, 221, 236, 252, 381, 501, 504, 707, 772, 794, 797, 805, 807, 817, 818, 830, 831, 890
Bladder Divergent Channel, 222, 225
Bladder Fire, 127
Bloated Energy Stagnation, 600
Blood Deficiency, 170, 214, 242, 489, 670, 768, 777, 785, 788, 798, 816, 818, 823, 828, 849, 851-853, 858, 859
Blood Heat, 490, 851, 854, 859, 890
Blood Pressure, 64, 274, 362, 377, 379, 391, 394, 398, 406, 426, 670, 672, 684, 780-783, 863, 865, 888, 890, 900, 901, 922
Blood Stasis, 180, 399, 486, 490, 667, 692, 825, 843, 854, 855, 858, 859, 881, 897, 898
Blood Vessels, 73, 76, 128, 138, 163, 170, 176, 178, 180, 189, 198, 217, 241, 243, 257, 262, 263, 322-324, 326, 329, 378, 391, 392, 394, 398, 401, 403, 466, 489, 490, 504, 610, 612, 617, 781, 843, 873
Blue, 74, 91, 101, 121-124, 147, 155, 217, 241, 247, 293, 294, 307, 326, 426, 453, 524, 526, 529, 561, 573, 575, 610, 624, 633, 653, 654, 658, 701, 769, 813, 814, 829, 842, 851, 852, 853, 904, 906, 909, 913, 925
Body Fluids, 73, 74, 92, 119, 159, 163-166, 170, 182, 184, 191, 193, 318, 324, 333, 459, 466, 481, 483, 490-492, 496, 501, 616, 687, 798, 897, 920
Bone, 79, 81, 95, 129, 198, 260-263, 319, 650, 837, 895 896
Bone Marrow, 76, 141, 166, 260, 319, 324, 330, 331, 491, 684, 685, 726, 901
Bones, 257, 260, 261, 263
Book of Commentaries, 271
Borborygmus, 472
Brain, 77, 138, 140, 141, 146, 184, 198, 201, 202, 257-260, 324, 328, 330, 331, 349, 401, 403, 404, 408, 410, 519, 527, 587, 599, 605, 671, 843, 861, 863-870, 880, 888, 895, 906
Brain Tumor, 441, 713, 906, 907
Breast Cancer, 896, 904, 907, 908
Breast Cysts, 108, 731, 601, 854, 855, 857
Breath Regulation, 355, 356, 391, 396,, 397 803
Bridge of Light, 409, 410, 548
Bronchial Asthma, 766, 803, 806
Bronchiectasis, 808
Bronchitis, 360, 382, 731, 803, 806, 823

C

Caduceus, 721
Calculi, 773
Cancer Formation, 898, 897, 899
Cancer Research, 904
Cancer Treatments, 895
Carbuncle, 854, 855
Carcinogens, 913, 916
Carcinoma, 204, 689, 713, 715, 896, 908-916
Categorization of Signs and Symptoms, 462
Categorization of Tumors, 895
Celestial Stems, 296, 300
Center Core, 131, 148, 149, 150, 151, 161, 307, 330-332, 409, 415, 438, 494, 541, 547, 573, 620, 621, 650, 737, 873
Center Qi, 321, 322, 324
Center Thrusting Channels, 120, 293, 295
Centrifugal and Centripetal, 249
Cerebral Embolism, 689, 863
Cerebral Hemorrhage, 689, 863
Cerebral Thrombosis, 863
Cerebro-Vascular Accident (CVA), 863
Cervical Cancer, 601, 689, 909-911
Chakra, 135, 139, 142-144, 151-157, 553, 600, 722, 737
Chakra System, 151, 153, 157
Channel Pathology, 193, 194, 252
Channel Point Meditation, 665, 669, 688, 882
Channel Point Therapy, 883
Channel Points, 59, 66, 126, 128, 159, 228, 246, 249, 250, 644, 651, 665-668, 675, 676, 692, 721
Chao, Yuan Fang, 71, 523
Chemotherapy, 411, 551, 707, 713, 919-921, 927
Cheng, Zhong Ling, 465
Chicken Pecking Technique, 663
Child Element, 656
China Healthways, 47, 119, 125, 129
Chinese Massage, 49, 50, 52, 56, 57, 61, 62, 129, 164, 309, 631, 565, 566
Cholecystitis, 772, 775, 776

Cholelithiasis, 772, 776
Cholestasis, 773
Chronic Bronchitis, 360, 731, 806
Chronic Constipation, 799,798
Chronic Fatigue Syndrome, 360, 598
Chronic Gastritis, 796
Chronobiology, 274
Chyluria, 791
Circle of Willis, 865
Circle Twisting Method, 677
Circular Breathing , 723
Cirrhosis, 774
Clairaudience, 146, 155, 343, 346
Clairsentience, 155
Clairvoyance, 146, 343, 346, 351, 437
Classification of Energetic Points, 249
Clean Qi, 191, 505
Climate, 215, 356, 466, 467, 488, 815
Clinical Ethics, 576, 578
Coccyx Pass (Wei Lu Guan), 605
Cold, 465-467, 470-475, 477, 480, 481, 551, 552, 597, 598
Cold Deficiency, 502
Colitis, 602, 789, 875, 916
Collapsed Qi, 488
Colon and Rectum Cancer, 916
Color Imagery, 123
Color Vibration, 119, 120, 122, 526
Color Visualization, 683, 688, 701, 770, 925
Coma, 64, 146, 444, 466, 823, 850, 864-866
Common Cold and Influenza, 801, 802, 804
Common People's Fire, 127
Compass School (Feng Shui), 270
Compassion, 75, 104, 155, 168, 429, 431, 437, 446-448, 524, 541, 617, 624, 860, 871, 872, 878, 880, 904, 905, 906, 921
Compressed Energy Stagnation, 600
Compression Release Method, 676-679
Concentrative Meditation, 407
Conception; 67-70, 72, 74, 84, 86, 119, 132, 200, 274, 275, 318, 320, 321, 333, 339, 409
Conception Vessel, 151, 157, 198, 200-208, 213, 217, 221, 242, 252, 257, 284, 294, 342, 372, 387, 432, 605, 606, 646, 670, 720-723, 733, 743, 785, 850-852, 855, 856, 882, 903, 909, 911
Congealed Blood, 241
Congenital Cellular Patterns, 115
Congenital Constitution, 99
Congenital Jing, 99, 115
Congenital Qi, 838
Congenital Spirit, 401
Connecting Channels, 208, 239, 864
Conscious Mind, 343, 409, 410, 439, 441, 617, 625, 665, 879
Constipation, 154, 174, 470, 502, 503, 505, 768, 781, 788, 789, 791, 794, 798, 799, 812, 828, 830, 873, 875, 915, 916, 920
Constitution, 61, 68, 99, 100-105, 176, 319, 356, 360, 362, 395, 419, 420, 426, 453, 457-459, 460, 462-465, 475-477, 479, 483, 484, 486, 490, 507, 511, 557, 563, 566, 569, 598, 603, 631, 689, 711, 784, 787, 798, 823, 832, 833, 839, 852, 858, 863, 864, 866, 929
Constitutional Failure, 459
Contracted Energy Stagnation, 600, 601
Contraindications for Medical Qigong, 590
Controlling Cycle, 100, 123, 313, 468, 496, 871, 874, 891
Corporeal Soul, 138, 250, 273, 429, 436, 847, 874, 886
Coryza, 805
Cosmology, 268, 348
Counter-Transference, 579
Creative Cycle, 100, 123, 249, 251, 281, 316, 495, 496, 656, 871
Creative Subconscious, 409, 412, 444, 513, 554, 555, 561, 617, 623
Crown Center, 156
Crown Chakra, 151, 153, 154, 737
Cupping, 57-59, 649, 659, 815, 816
Curious Organs, 257
Cycle of Disharmony, 468
Cycle of the Moon, 276, 284
Cycle of the Sun, 276, 277
Cycles of the Stars, 290
Cysts and Tumors, 64, 601, 731, 854, 860, 887, 895
Cytotoxic Treatments, 919

D

D.M.Q., 50, 54
Damp Heat, 174, 184, 193, 210, 486, 504, 505, 667, 669, 671, 747, 768, 773-776, 788, 792, 805, 809, 819-821, 851, 852, 857, 911, 920
Dantian Breathing, 723
Dao, 47, 84, 102, 132, 141, 142, 144, 148, 149, 199, 267, 268, 271, 317, 318, 323, 326, 333, 407, 416, 514, 515, 541, 587
Daoist, 99, 118, 131-133, 143, 249, 333, 348, 350, 378, 688, 708, 717, 751, 752, 762, 763, 784, 797, 799, 806, 827, 854, 868
Death, 86-89, 92, 93, 95, 132, 134, 135, 142, 144, 268, 270, 276, 292, 334, 339, 344, 411, 426, 436, 437, 442-445, 459, 468, 503, 518, 530, 541, 545,

625, 626, 627, 628, 827, 846, 848, 860, 865, 874, 875, 896, 900, 911, 913, 915, 921
Defense Mechanism, 149, 447, 450, 604, 619, 621, 838
Deficient Dantian Qi, 823
Delusions, 437, 438, 870, 883
Demon Oppression, 437, 438
Demon Possession, 424, 438, 442,
Denial, 142, 326, 438,441, 444, 447, 449, 450, 569, 578, 602, 618-621, 626, 720, 875, 882, 892, 898, 899, 922
Depersonalization Disorder, 446
Depression, 89, 119, 129, 156, 168, 172, 191, 202, 214, 219, 245, 335, 337, 341, 391, 404, 426, 427, 439, 444, 446, 460, 468, 469, 476, 486, 518, 545, 560, 590, 598, 626, 627, 653, 695, 708, 772, 776, 798, 818, 830, 843, 844, 850, 852, 855, 857, 874, 881, 886, 887, 897, 898, 900, 903, 907, 913, 922, 927
Diabetes, 64, 155, 311, 781, 791, 792, 820, 863
Diagnosis According to the Triple Burners, 504
Diagnosis and Treatment of the Patient, 462
Diaphoresis, 501
Diet, 46, 61, 69, 95, 99, 115, 176, 260, 308-310, 313, 421, 426, 458, 460, 467, 472, 473, 475, 484, 486, 489, 490, 509, 510, 570, 686, 773, 791, 798, 807, 818, 838, 847, 850, 851, 859, 863, 867, 881, 895, 911, 915, 916, 927
Different Schools of Qigong, 47
Differentiation of Psycho-Emotional Disorders, 881
Differentiation of Syndromes, 461, 502
Digestion, 111, 168, 174, 176, 178, 182, 189, 191, 260, 309, 310, 312, 333, 360, 361, 372, 409, 460, 486, 505, 509, 616, 694, 743, 789, 897, 914, 920
Direct Current, 128, 246
Disorders of Body Fluids, 491,490
Disorders of Qi, 483
Disorders of the Blood, 489, 490
Disorders of the Body Fluids, 490
Disorders of the Human Body, 459
Disorders of the Qi, 483, 489, 825
Divergent Channels, 160, 161, 222-227
Divine Center, 290, 438
Divine Energy, 46, 48, 91, 132, 147, 161, 260, 317, 323, 407, 515, 531, 570, 576, 587, 661, 870
Divine Guidance, 649
Divine Healing, 407, 440, 452, 453, 533, 542, 544, 546-549, 556, 577, 606, 624, 661 697, 708, 709, 846, 857, 861, 868, 904, 905, 928
Divine Hook-Up, 541, 542, 546-548, 550, 582, 583
Divine Light, 86, 115, 326, 437, 547, 559, 573
Divine Qi, 323, 531, 694, 696, 697
Divine Will, 142, 155, 340, 407, 548, 625, 627
Dossey, Larry, 530
Dragons Mouth Technique, 659
Draining Toxic Qi, 596
Dreams, 335-337, 340, 344, 347 436, 437-439, 448, 449, 627
Dry Crying Prescription, 622, 688, 857, 893, 913, 914
Durville, Hector, 437
Dynamic Qigong, 278, 355, 356, 372, 373, 406, 426, 591, 717, 866
Dysmenorrhea, 206, 471, 768, 852, 853, 924
Dyspepsia, 469, 694, 775, 796, 915
Dyspeptic, 794
Dysphoria, 781, 796
Dyspnea, 172, 198, 214, 251, 253, 360, 464, 467, 785, 803, 806, 807, 811, 813, 838

E

Earth Element, 98, 100, 293, 493, 663, 676
Earth Jing, 80, 81
Earth Qi, 67, 73, 82, 83, 132-134, 301-303, 305, 307, 308, 324, 362, 372, 380, 383, 386, 532, 656, 718, 719, 724, 726, 728, 733, 734, 737, 752
Earth Transpersonal Point, 576
Earthly Branches,156, 157, 161, 268, 269, 270, 281, 290, 291, 296, 297, 299, 721
Eczema, 241, 807
Edema, 96, 172, 242, 382, 422, 492, 496, 498, 602, 665, 667, 777, 785, 787, 788, 812, 819, 850, 906, 920
Ego, 142, 155, 331, 340, 344, 416, 428, 517, 519, 541, 621, 880
Eight Confluent Points, 253, 611
Eight Energetic Principles, 308, 461, 465, 478, 480, 509
Eight Energetic Touches, 359, 408, 588, 589, 598
Eight Extraordinary Vessels, 70, 71, 83, 84, 107, 133, 148, 159, 160, 197, 198, 199, 200, 215, 216, 253, 257, 284, 296, 332, 611, 731, 826, 849, 882
Eight Influential Points, 253
Eight Miscellaneous Causes of Disease, 460
Eight Miscellaneous Factors, 459, 460
Eight Principles of Differential Diagnosis, 465
Eight Trigrams, 270, 271, 286, 610-612
Eighteen Classifications of Qi, 321
Eighth Lunar Month, 80
Electricity, 117, 124, 126, 128, 129, 141, 581, 631
Elemental Constitutions, 83, 95, 99, 100, 106, 107, 114, 115, 320, 462, 612
Elements, 57, 75, 98, 99, 107, 118, 119, 249, 250, 251, 268, 269, 270, 273, 276, 281, 290, 299, 308,

310, 311, 313, 469, 493, 494, 496, 497, 555, 556, 575, 612, 654, 655, 663, 675, 690, 838, 870, 871, 878
Emaciation, 206, 484, 594, 791, 808, 825, 914
Embryological Development, 67, 71, 200, 206
Emotional Components of Cancer, 898
Emotional Disharmony, 468, 878
Emotional Disorder, 882
Empathic Communication, 84, 138, 139, 513, 519
Emperor, 90, 91, 138, 144, 268, 293, 296, 333, 529, 607, 609, 767, 777, 787, 801, 811
Empty Qi, 484, 486, 488
Endocrine System, 392, 394, 403
Endorphins, 129, 635, 899, 922
Energetic Armoring, 431, 600, 602, 606, 688, 778
Energetic Barriers, 327
Energetic Blockages, 598
Energetic Boundary Invasion, 581, 582
Energetic Circle Drawing Pattern, 546, 547, 550, 551
Energetic Complications, 600, 601, 606
Energetic Cords, 552-555, 624, 905, 906
Energetic Cupping, 649, 659
Energetic Depletion, 559, 560, 569, 600-603, 606
Energetic Disharmony, 459
Energetic Disorder, 461
Energetic Grids, 88, 91, 272, 326, 327, 561, 925
Energetic Leakage, 600, 603, 606
Energetic Medicine, 84, 157, 259, 628, 766, 869
Energetic Point Therapy, 57, 64-66, 247, 405, 665, 669
Energy Balancing Procedures, 597
Energy Projection, 118, 120, 155, 247, 326, 343, 348, 349, 351, 561, 631, 633 649, 651, 654, 769
Enuresis, 184, 219, 220, 766, 788, 802, 812
Environmental Energy, 48, 67-69, 88, 125, 149, 161, 199, 268, 274, 302, 307, 327, 517, 547, 553, 560, 724, 726, 739
Environmental Force, 270
Environmental Qi, 86, 144, 152, 286, 302, 306, 307, 324, 351, 587, 391, 572, 846
Epigastric Pain, 471, 496, 768, 792, 796, 850
Esophageal Cancer, 914
Esophagus, 176, 180, 914
Eternal Soul, 48, 74-76, 80, 86, 87, 90, 92, 93, 138, 251, 334-338, 340, 341, 401, 436, 447, 349, 444, 616
Ethics, Clinical, 576, 580, 578
Etiology During Adulthood, 458
Etiology During Childhood, 458
Etiology Within the Womb, 457
Eustachian Tubes, 741
Evil Qi, 290, 322-324, 357, 377, 431, 459, 460, 570, 571, 590, 664, 679, 692, 779, 803
Evil States, 437, 439
Evil Wind, 851
Excess Body Fluids, 492
Excess Heat, 51, 126, 134, 260, 394, 395, 488, 490, 510, 594, 664, 706, 707, 713, 731, 733, 783, 789, 803, 813, 823, 829, 839, 841, 856, 866, 897, 908, 911, 913, 914, 915, 916, 917
Excess Liver Fire, 364, 776, 781, 783, 863, 890, 906, 907, 911
Excess Qi, 58, 59, 61, 107, 108, 129, 140, 197, 243, 310-312, 377, 387, 388, 399, 405, 486, 488, 532, 551, 593, 595, 612, 669, 680, 687, 740, 744, 745, 779, 782, 783
Expanded Energy Stagnation, 600, 602
Extended Fan Palm Technique, 552, 634, 640, 650-653, 658, 659, 773-776, 782-785, 791, 794-799, 805-808, 810, 816, 817, 819, 820, 825 827, 829, 830, 853, 854, 866, 867, 882
External Pathogenic Factors, 48, 64, 239, 251, 312, 322, 323, 327, 328, 465, 466, 471, 503, 613, 699, 807
External Wei Qi Fields, 64, 151, 263
Extra Point, 572, 590, 742
Extraordinary Organs, 141, 197, 198, 257
Extroverted Structure, 114

F

Facial Paralysis (Bells Palsy), 194, 212, 467, 864, 866
Facing the Death of a Patient, 625
Faith Projection, 532
Fall Equinox, 573
False Self, 441
False Yang, 475, 477
False Yin, 478, 477
Fascia, 61, 67, 95, 150, 159, 163, 164, 228, 229, 241, 552, 673, 675, 679, 817, 927
Fear, 64, 75, 76, 89, 99, 104, 106-108, 129, 142, 149, 151, 154, 172, 182, 187, 327, 330, 335, 348, 365, 368, 404, 413, 427, 429, 432, 439, 445, 452, 460, 470, 488, 508, 512, 518
Febrile Diseases, 59, 501, 502, 504, 798, 838
Feng, Lida, 904
Feng Shui, 119, 270
Fetal Toxins, 68, 77, 458
Fetus, 68-77, 80-83, 86, 88, 99, 148, 197, 204, 206, 208, 257, 269, 275, 317, 320, 322, 457, 516, 517, 553, 684, 686, 858, 859, 860
Fibroadenoma, 854, 855
Fifteen Collaterals, 217
Fifth Lunar Month, 77

Fire Cycle, 157, 201, 260, 293, 294, 646, 662, 707, 719-721, 784, 812, 817, 839, 845, 854, 867, 868, 908, 917
Fire Element, 79, 98, 100, 137, 251, 293, 493, 653, 676
Fire Jing, 77
First Lunar Month, 72
Five Agents, 74-78, 86, 87, 99, 118, 137, 292, 333-335, 347
Five Dominions of Energy, 45
Five Element Cycles, 878
Five Element Organs, 296
Five Element Theory, 313, 333, 463
Five Elemental Animals, 270
Five Elemental Constitutions, 83, 95, 99, 100, 106, 107, 114, 115, 320, 462, 612
Five Energetic Fields, 47, 326
Five Essence Spirits, 75, 90, 139, 141, 334, 346, 847, 870
Five Flavors, 310, 311, 313
Five Orbs, 77, 78
Five Palms Hot, 471, 802
Five Passes, 424, 604, 605
Five Portals of the Earthly Gate, 148, 342
Five Portals of the Heavenly Gate, 342
Five Sense Organs, 217, 242
Five Stages of Healing, 687
Five Thunder Fingers, 594, 641, 647
Five Tissues, 242
Five Yang Organs, 82
Five Yin Organs, 75, 77, 78, 80, 82, 84, 90, 137, 141, 204, 292, 295, 329, 333, 339, 341-344, 372, 376, 419, 429, 452, 572, 615, 655, 657, 663, 699, 724, 743, 749, 811, 837, 849, 870, 871, 873, 876
Fixation, 621
Flat Palm Detection, 521, 522, 554
Fluids, 60, 73, 74, 76, 92, 119, 137, 159, 163-166, 170, 174, 182, 184, 189, 191-193, 241, 242, 310, 311, 312, 315, 318, 322-324, 333, 459, 466, 470, 472, 473, 478, 481, 483, 490-492, 495, 496, 498, 501, 504, 537, 602, 616, 687, 752, 791, 798, 801, 871, 881, 897, 920
Food and Color, 123
Food Qi, 137, 321, 324, 837
Form School (Feng Shui), 270
Four Bigs, 501-503, 505
Four Categories of Energetic Points, 246
Four Doors, 749
Four Levels, 57, 462, 502, 510
Four Principle Time Periods, 278, 657
Four Seasonal Peak Transitions, 279
Four Winds, 306
Fourth Lunar Month , 76
Freud, Sigmund, 516, 621

G

Gall Bladder, 160, 166-168, 182, 191, 192, 194, 257, 260
Gall Bladder Channel, 73, 166, 168, 194, 215, 219, 222, 229, 253, 375, 560, 595, 707, 768, 772, 774, 815
Gall Bladder Divergent Channel, 222, 223
Gallstones, 119, 773, 775
Gastric and Duodenal Ulcers, 789, 797
Gastritis, 208, 695, 789, 796
Gastroptosis, 360, 788
Gate of Access, 292
Gate of Life (Mingmen), 135, 154, 670, 746, 811
Gate of the Moon, 279, 280
Gathering Qi, 137, 303, 305, 306, 322, 324, 391, 469, 641, 688, 726-728
Gathering Sun, 688
Generated Heat, 126
Geriatrics, 567, 836, 843
Gershon, Michael, 134, 135
Ghost Points, 254, 670, 884
Ghosts (Gui), 351, 435, 436, 439, 441, 444, 445, 884
Gill, Jason, 929
Gold, 120, 131, 132, 148, 156, 283, 293, 294, 307, 331, 406, 431, 452, 573, 624, 640, 657, 658, 661, 701, 719, 905
Gold, Philip, 426
Governing Vessel, 198, 200-203, 215, 220, 221, 253, 259, 260, 849, 866
Grasping and Shaking Method, 676, 677
Green, 53, 101, 121-123, 155, 293, 296, 307, 308, 310, 425, 433, 439, 492, 496, 508, 574, 575, 658, 701, 770, 851, 852, 857
Grief, 48, 64, 75, 88, 99, 104, 105, 108, 129, 155, 170, 172, 339, 404, 413, 427, 432, 439, 445-447, 460, 488, 508, 559, 581, 594, 599, 600, 602, 616, 617, 619, 708, 844, 855, 857, 871-875, 877, 890-892, 894, 899, 907, 922
Gu Qi (Food Qi), 83, 137, 163, 176, 180, 191, 309, 319, 321, 322, 324, 489, 495, 496, 537, 787, 837
Gua Sha, 57, 61, 62, 63, 565
Gui (Ghosts), 86, 93, 334, 338, 436, 444, 445, 611, 875
Guiding Qi, 779, 803, 368
Gynecology, 567, 836, 849, 860

H

Hai Dian University, Medical Qigong College, 51, 69
Hand Postures for Qi Emission, 634
He, Si Hai, 689
Headaches, 54, 64, 111, 115, 155, 198, 202, 211, 214, 219, 394, 397, 403, 421, 422, 426, 495, 518, 569, 599, 672, 768, 769, 777, 781, 801, 802, 812, 818, 823, 829, 831, 832, 833, 850, 890
Healing Emotional Trauma, 623
Healing Sound Therapy, 118, 699, 710, 919
Heart, 80, 84, 99, 140, 152, 179, 214, 220, 281, 293, 337, 338, 340, 341, 343, 430, 490, 495, 496, 497, 560, 603, 606, 616, 664, 726, 769, 779, 817, 838, 872, 886
Heart Channel, 178, 180, 194, 220, 235, 279, 778
Heart Diseases, 379, 384, 669, 777, 778, 780
Heart Divergent Channel, 222, 226
Heart Fire, 123, 127, 137, 151, 180, 310, 311, 430, 495, 558, 671, 704, 744, 778, 779, 823, 831, 838, 839, 843, 876
Heart Massage, 691, 693, 694, 695, 696, 697, 779, 785
Heat Deficiency, 502
Heaven Qi, 67, 83, 87, 125, 132, 139, 140, 141, 273, 308, 324, 372, 381, 386
Heavenly Stems, 161, 268, 270, 281, 286, 296, 297, 299
Heavenly Transpersonal Point, 576, 661
Hemiplegia, 864, 865
Hemoptysis, 785, 807, 808, 850
Hepatitis, 383, 425, 567, 770, 776, 913
Hepatolithiasis, 776
Hepatomegaly, 913
Hepatosplenomegaly, 770
Herbal Medicine, 49, 50, 60, 164, 336, 565, 566
Herbalist, 60, 61, 566, 569
Herpes Simplex, 472
Hexagram, 109, 111-115, 277, 278, 279, 363
Hibernation Breathing, 399, 689
Hidden Power (An Jing), 48, 260
High Blood Pressure (Hypertension), 64, 379, 406, 443, 781, 783, 863, 890
History of Cancer Treatments, 895
Hollow Organ, 767, 777, 787, 801, 811
Hologram, 325, 326, 404, 516, 581
Hong, Liu , 912
Hostile Forces, 89, 438, 442
Hou Tian Zhi Qi, 321
Housing the Shen, 48, 178, 337, 342, 343
Howell, Madeleine, 53, 56
Hua Jing (Mysterious Power), 48
Hua, Tuo, 924
Huiyin Point, 145, 201, 284, 342, 550, 573
Hun (Ethereal Soul), 48, 74-78, 80, 84, 86, 87, 88, 90-92, 132, 138, 139, 141, 146, 168, 250, 251, 273, 333-341, 344, 347, 371, 401, 409, 420, 424, 429, 435, 436, 438, 441, 442, 444, 446-448, 460, 574, 604, 616, 717, 724, 767, 818, 839, 847, 865, 871, 872, 890
Hyperhidrosis, 791
Hypertension, 51, 54, 99, 180, 211, 214, 360, 375, 391, 394, 403, 425, 565, 566, 669, 672, 770, 774, 777, 779, 780, 781, 782, 783, 823, 828, 829, 850, 865, 890
Hyperthyroidism, 781, 784, 791
Hypochondriac Pain, 215, 498, 768, 770, 772, 775, 776, 874, 913
Hyposmia, 803
Hypotension, 210, 214, 376, 425

I

I Ching, 109
Id, 621
Immortal, 842
Immune System, 47, 115, 137, 154, 198, 241, 260, 274, 322, 341, 385, 400, 403, 426, 430, 470, 552, 598, 620, 684-686, 753, 807, 840, 895, 898-902, 905, 906, 919-921, 929
Impotence, 154, 202, 230, 695, 726, 819, 820
Improper Yin and Yang Energy Balance, 824
Incredible Wonders, 588, 590
Indago, 575
Infatuation, 437, 444
Influenza, 470, 788, 801-805, 809
inner-vision, 85, 259, 325, 346, 371, 405
Insight Meditation, 407
Insomnia, 64, 138, 180, 204, 211-214, 250, 322, 335, 395, 471, 478, 484, 495, 497, 502, 504, 565, 672, 697, 768, 778-782, 784, 817, 818, 823, 825, 828, 830,840, 843, 847, 890
Insulting Cycle, 497
Intent (Yi), 49, 80, 134, 191, 278, 332, 361, 378, 394, 398, 411, 427, 534, 617, 648, 650, 691-696, 813
Internal Dialogues, 86, 87
Internal Organ Disorders, 838
Internal Pathogenic Factors, 64, 323, 467, 699
Internal Viewing (Inner-Vision), 520
Internal Wind, 202, 465, 467, 863, 881

Introverted Structure, 114
Intuition, 101, 102, 134, 140, 143, 144, 146, 154-156, 197, 201, 286, 328, 336, 343, 544, 577, 617, 649
Intuitive and Perceptual Diagnosis, 507, 512, 513, 521
Invading Cycle, 497
Invisible Needle Therapy, 65, 66, 247, 639, 640, 651, 659, 661-665, 669, 865
Iron Wall, 604

J

Jahnke, Roger, 903
Jaundice, 498, 768, 772, 773, 915
Jia Ji Guan, (Spinal Pass) 605
Jie Gu Therapy, 61, 63, 565
Jing Point Therapy, 57, 61-63, 228, 247, 595, 783, 831, 855-858, 890, 920
Jing-Luo, 159, 323
Jing-Shen Bing (Mental Illness), 141
Johnson, Mark, 846
Joy, 75, 99, 129, 178, 189, 193, 404, 428, 429, 431, 460, 469, 488, 508, 615, 619, 872-879
Jue Yin, 57, 160, 222, 501-503, 612, 884
Juice of Jade, 372, 743
Jung, Carl G., 516

K

Kaku, Michio, 665
Karma, 70, 86, 268, 416
Ketheric Matter, 328
Kidney Channel, 81, 107, 154, 184, 187, 194, 206, 208, 210, 219, 237, 706, 814, 819, 820, 850
Kidney Diseases, 426, 710, 713, 766, 811-815
Kidney Divergent Channel, 208, 222, 225
Kidney Fire, 127, 133
Kidney Jing, 133, 137, 170, 184, 260, 319, 320, 812, 815, 816, 838, 847
Kidney Massage, 695, 813, 818
Kidney Qi, 133, 184, 185, 312, 377, 380, 382, 387, 391, 461, 491, 495, 498, 799, 807, 811, 813, 815, 838, 843, 858, 872, 876, 877
Kinetic Communication, 84, 134, 347, 513
Kneading Tiger Palm, 559, 640, 647, 650, 659, 799, 833
Knees, 109, 111, 113, 215, 217, 249, 251, 360-365, 368, 422, 428, 558, 575, 600, 651, 691, 724, 746, 747, 748, 749, 753, 757, 761, 776, 783, 791, 804, 817, 820, 842, 843, 845, 868, 876, 893, 912

L

Lao, Zi, 541
Large Intestine, 71, 78-82, 98, 123, 136, 160, 161, 166, 170, 172-174, 180, 191, 194, 215, 219, 220, 222, 224, 232, 240, 252-254, 257, 276, 281, 182, 312, 339, 366, 378, 382, 387, 501-504, 525, 595, 616, 713, 715, 751, 801-803, 829, 830, 831, 875, 916, 917, 920
Large Intestine Channel, 80, 81, 172, 174, 194, 219, 232, 382, 595, 802, 803, 829
Large Intestine Divergent Channel, 222, 224
Lashley, Karl, 404
Lee, Richard H., 119, 125, 129, 880, 881
Leukocyte, 901, 902
Leukorrhea, 206, 816, 910
Ley lines, 271
Li, Fu Dong, 51
Li, Shi Zhen, 200, 519
Light, 166, 326, 331, 342, 347, 410, 442, 530, 542, 548
Light Meditation, 342, 904, 925, 928
Liu, Bing, 274
Liver, 168, 170, 174, 182, 260, 310, 929
Liver Cancer, 383, 715, 770, 913, 920
Liver Channel, 72, 168-170, 187, 194, 219, 222, 230, 380, 384, 471, 770, 775
Liver Diseases, 710, 712, 766-769, 772
Liver Divergent Channel, 222, 223
Liver Fire, 117, 155, 215, 364, 382, 497, 595, 729, 768, 769, 776, 781, 783, 807-809, 855, 857-859, 863, 871, 890, 906, 907, 911, 912
Liver Massage, 693, 774, 775, 776, 854
Liver Wind, 117, 505, 745, 768, 815, 816, 828, 838, 867, 890
Lo Scroll, 607
Locations of Energetic Blockages, 598
Long Distance Therapy, 530, 532
Lords of the Three Dantians, 84, 85, 86
Lower Burner, 57, 182, 191-193, 204, 211, 322, 387, 504, 505, 510, 598, 610, 670, 671, 700, 729-731, 747, 791, 794, 799, 812, 849, 851, 909, 911, 920
Lower Dantian, 363, 397
Lu, Guo Hong, 51
Lumbago, 362, 421, 670, 726, 814, 816-820, 853
Lung Cancer, 379, 567, 711, 713, 913, 920
Lung Channel, 79, 166, 170, 172, 194, 220, 231, 282, 802, 803, 806-809, 857, 893, 903
Lung Diseases, 155, 172, 360, 492, 731, 801-804, 892, 913

Lung Divergent Channel, 222, 224
Lung Massage, 694
Lungs, 123, 124, 137-139, 155, 161, 166, 168, 170, 172, 174, 178, 184, 191-193, 198
Luo points, 217, 218, 220, 761
Lymphocytes, 684, 685
Lymphocytes, 178, 685, 867, 901
Lymphoma, 915
Lymphosarcoma, 896

M

Maciocia, Giovanni, 194
Macrocosmic Orbit, 189
Macrophages, 684, 685, 867, 898, 901
Magic Square, 200, 270, 607, 608, 610-612, 679, 690
Magnetic Energy, 117, 124, 125, 581
Magnetic Therapy, 57, 58, 126, 856
Malar Flushes, 497
Malignant Tumors, 689, 896, 904, 907, 909, 913-916, 927
Mantra, 89, 406, 424, 701, 792, 882
Marrow, 57, 64, 141, 143, 184, 198, 202, 253, 254, 257, 259, 260, 263, 288, 315, 318, 319, 324, 326, 330, 331, 359, 431, 461, 591, 616, 671, 673, 723, 724, 837, 838, 888
Master and Couple Points, 252
Master Points, 213, 215, 252
Mastitis, 823, 854, 855
Meditation, 282, 288, 294, 307, 332, 376, 407, 426, 430, 431, 448, 520, 527, 532, 546, 547, 572, 573, 574, 591, 622, 623, 669, 688, 718, 882, 900, 904, 905, 912, 917, 928
Meditations for Cancer Patients, 904
Memory, 113, 114, 119, 140, 141, 149, 176, 178, 187, 202, 259, 319, 329, 337, 338, 340, 341, 351, 395, 404, 444, 445, 448, 450, 452, 484, 490, 603, 616-618, 621, 650, 685, 686, 777, 818, 825, 843, 844, 864, 866, 868, 872, 873, 879, 882, 890, 922, 927
Meng, Xian Tong, 548
Menopause, 204, 850, 852, 907, 908
Menoxenia, 854
Menstrual Complications, 852
Menstruation, 204, 206, 257, 274, 363, 590, 591, 794, 849-854
Mental Dao Yin Training, 357, 401, 421, 404, 405, 425
Mental Delusions, 437
Mental Disorder, 438, 508, 677
Mental Restlessness, 214, 250, 471, 504
Meridians, 159
Message (Xin Xi), 80, 134, 139, 142-245, 327, 340, 343, 344, 346, 349, 352, 488, 514-516, 534, 535, 551, 599, 618, 625, 720, 922
Metal Element, 98, 293, 493, 653, 676
Metal Jing, 78
Microcosmic Fire Cycle, 201
Microcosmic Orbit, 120, 157, 189, 199, 242, 260, 293, 294, 304, 365, 366, 424, 432, 547, 558-560, 589, 604, 645-647, 662, 688, 720-723, 726, 728, 733, 784, 812, 817, 831, 832, 839, 845, 854, 856, 858, 866-868, 908, 909, 911, 917, 924, 928
Microcosmic Water Cycle, 201
Middle Burner 57, 87, 139, 170, 191-193, 242, 321, 322, 502-505, 598, 610, 670, 671, 700, 730, 731, 768, 788, 789, 791, 793, 796
Middle Dantian, 131, 132, 137-140, 145-147, 154, 155, 161, 189, 198, 776, 783, 785, 874
Migraine, 54, 64, 394, 599, 829, 830-833, 890
Migrating Qi, 600, 603
Ming Jing (Obvious Power), 48
Ming, Qing Zhu, 50
Mingmen, 560, 590, 605, 670, 785, 812, 813, 842
Mingmen Fire, 127, 133, 368, 705, 706, 811-813, 859, 876
Moon Cream, 288, 688, 775, 785, 809, 819
Morphogenic Field, 315
Mother Element, 309, 401, 495, 496
Moxa Therapy, 57-59, 128, 212, 239, 565, 566, 864
Multiple Personality Disorder, 437, 442, 443, 445, 880
Multiple Sclerosis (MS), 863, 867, 868
Muscle Tendon Channel, 60, 161, 163, 217, 228, 239
Myocarditis, 777
Myss, Caroline, 141
Mysterious Pass, 398
Mysterious Power (Hua Jing), 48
Mysterious Wonders, 588, 589

N

Nader, Ralph, 847
Natural Breathing, 361, 379, 391-393, 397, 651, 700, 701, 718, 719, 752, 783, 809, 845, 854, 868
Necrosis, 774
Nei Dan Shu, 718, 719
Nei Guan, 520, 688
Nei Jing, 355, 611, 881, 886
Neoplasm, 907, 908, 909, 910, 911

Nephritis, 781, 814
Neurology, 50, 567, 836, 863
Neuropeptides, 135, 899
Nine Chambers, 82, 135, 136, 139, 140, 143, 144, 292
Nine Palaces, 296
Nine Stars, 135
Nordenstrom, Bjorn, 128
Ninth Lunar Month, 81
Niu, Yu Hua, 51
Nutritive Qi, 137

O

Objective World, 438
Obstructed Qi, 486, 488, 825, 826, 892
Obvious Power (Ming Jing), 48
Occipital Pass, 605
Old Age, 461, 486, 781, 843
Oncology, 50, 567, 836, 895, 903, 904
One Finger Skill Technique, 637, 638, 639, 658, 930
Opening and Closing the Three Dantians, 152, 170, 239, 322, 359, 377-383, 399, 423, 425, 433, 458, 467, 640,688, 692-697, 729, 731, 737, 824
Opisthotonosis, 424
Orange, 121, 122, 154, 311,526, 857, 915
Organ Dysfunction, 403, 831, 860
Organ Regulation, 62, 63, 751, 752, 797,850
Organ Transplantation, 927
Orifices, 79-81, 92, 159, 187, 292, 378, 491, 672, 728, 811, 815, 823, 838, 839, 858, 881, 886-888
Original Energy (Yuan Qi), 316, 317
Original Essence (Yuan Jing), 316, 317, 318, 331
Original Force, 268, 320
Original Spirit (Yuan Shen), 84, 85, 90, 316-318, 419
Original Yang (Kidney), 321
Original Yin (Kidney), 321
Osteoporosis, 430, 827
Outer Spiritual Barrier, 327
Ovarian Cysts, 64, 601, 734, 766, 857, 858
Ovarian Cysts and Uterine Tumors, 857

P

Palace of Eternal Frost, 279
Palace of Universal Yang, 279
Palpitation, 180, 384, 495, 497, 670, 777, 779, 782, 784, 785, 818, 823, 828, 843, 854, 887
Pancreas, 111, 122, 153, 154, 176, 191, 311, 505, 525 610, 791, 915, 916, 927, 928
Pancreatic Cancer, 689, 915
Pancreatitis, 775, 793
Pang, Dong Hui, 51
Paraplegia, 865
Parenchymal Cells, 774
Parkinson's Disease, 812, 847
Past Life, 344, 516-518, 887
Pediatric, 62, 567, 837-839
Penetrating Wind (Stroke), 863
Perceptual Communication Training, 519
Peribronchial, 806, 808
Pericardium, 71, 74, 79, 122, 140, 160, 161, 166, 178, 187-194, 213, 214, 219,220, 222, 227, 238, 240, 252, 253, 257, 276, 282, 375, 384, 490, 502-505, 664, 706, 778, 784, 785, 875, 889, 925
Pericardium Channel, 74, 187, 189, 194, 213, 220, 238, 375, 384, 504, 706, 778, 784, 785
Pericardium Divergent Channel, 222, 227
Perineum, 83, 109, 110, 111, 125, 127, 132, 133, 135, 145, 147, 148, 149, 151, 154, 191, 200, 201, 205, 206, 221, 284, 294, 295, 331, 332, 342, 363, 365, 393, 422, 575, 700, 722, 724, 748, 752, 814, 886, 911, 912
Peristalsis, 728, 748
Peritonitis, 794
Pernicious Influences, 470
Personal Subconscious Mind, 409, 410, 411,441
Peyer's Patch, 684
Phagocytes, 684, 898, 899
Phantom Embryo, 860, 861
Phantom Organ, 925
Phantom Pain, 929
Phlegm, 473, 492, 495, 504, 505, 602, 667, 669, 671, 683, 702, 703, 744, 781, 788, 802, 805-809, 815, 828, 829
Phlegm Fire, 176, 669, 784, 792, 815, 828, 865, 881
Phobias, 518, 882, 886
Piezoelectric, 129, 263
Pineal Gland, 141, 144, 156, 366, 409, 410, 520, 521, 726, 741
Pink, 426, 779
Pituitary Gland, 155, 366, 392, 394, 440, 520, 521, 635, 723, 832
Po (Corporeal Soul), 74-76, 80, 83-92, 134, 138, 139, 141, 146, 170, 250, 251, 273, 333-335, 338-341, 344, 347, 365, 371, 401, 403, 409, 424, 429, 436, 438, 444, 446, 447, 460, 574, 604, 616, 801, 847, 865, 871, 874, 881, 886
Point Names, 249
Point Respiration, 692, 693, 694, 695, 696, 697
Polarity, 70, 129, 152, 199, 200, 250, 268, 536, 581, 619, 639, 662, 678

Poltergeists, 435,436, 437
Polydipsia, 791
Polyphagia, 791
Polyuria, 791
Portal Hypertension, 774
Postnatal Jing, 137, 316-320, 324, 333, 607, 717
Postnatal Mind, 346, 402, 409
Postnatal Qi, 133, 137, 174, 198, 199, 309, 316, 320, 321, 333, 338, 459, 588, 603, 708, 720, 837
Postnatal Shen, 316, 333-337, 430
Postoperative Therapy, 921, 924
Posttraumatic Stress Disorder, 878, 880
Postural Dao Yin Training, 355, 359, 363
Prayer, 46, 87, 90, 91, 99, 118, 127, 146, 148, 149, 260, 305, 317, 318, 331, 333, 337, 529, 530, 535, 542, 544, 548, 687, 844, 861, 876
Pregnancy, 68, 69, 72-74, 81, 99, 204, 206, 257, 274, 275, 321, 457, 459, 590, 591, 686, 781, 849, 850, 858, 860, 861
Prenatal Jing, 141, 316, 318, 319, 320, 324, 333, 517, 686
Prenatal Jing, 517, 717, 720
Prenatal Qi, 68, 84, 120, 138,198, 199, 260, 279, 316, 317, 320, 321, 333, 723, 837
Prenatal Shen, 316, 333, 335, 337
Preoperative Therapy, 921
Prescriptions, 561, 784, 857, 902
Primary Channels, 71, 83, 84, 127, 133, 148, 159, 160-166, 189, 194-199, 206, 217, 218, 220, 222, 228, 239, 240, 241, 246, 249, 250, 253, 257, 281, 282, 284, 296, 387, 611, 720, 850
Prostate Cancer, 689, 715, 734, 911, 912
Protective Qi (Wei Qi), 132, 191, 215, 322, 323, 324
Psoas Muscles, 893, 894
Psychic Powers, 512
Psychogenic Polyuria, 791
Psychology, 49, 55, 56, 98, 117, 344, 445, 449, 512, 567, 869, 870
Psychometry, 534
Psychoneurosis, 770, 814
Psychosexual Qi Deviations, 438, 600, 603, 606
Psychosomatic, 532, 766, 820, 828, 829
Psychotherapy, 336, 861, 877, 878, 891
Pulling Down The Heavens, 283, 288, 688, 692-696, 700, 702, 711, 724, 725, 761
Pulmonary Emphysema, 803, 806
Pulse, 122, 148, 151, 152, 247, 248, 274, 312, 316, 326, 330, 331, 341, 349, 359, 391, 392, 470-473, 476-481, 494, 501-505, 509, 510, 562, 589, 640, 674, 676, 677, 777, 779, 837, 864, 882, 904, 922, 924
Pungent Foods, 311, 312, 801
Purgation Techniques, 594, 595, 596
Purging, 84, 284, 396, 397, 419, 452, 537, 551-559, 839
Purpura, 669
Pyelonephritis, 818

Q

Qi, Blood and Body Fluid, 463, 511
Qi Compression, 598
Qi Deficiency, 247, 422, 459, 472, 489, 491, 777, 788, 799, 802, 807, 838, 849, 854, 859
Qi Deviation, 419-426, 433, 434, 437, 460, 545, 569, 570, 590, 596, 600, 603, 606, 687, 699, 700, 719, 720, 723, 796, 900
Qi Dysfunction, 768, 779, 789, 803, 813
Qi Extension Techniques, 643, 930
Qi Manipulation, 598, 631, 674, 678, 772, 775, 779, 789, 803, 805, 854
Qi Regulation, 320, 374, 375, 399, 434, 582, 596, 597, 646, 708, 718, 719 824, 845
Qi Reservoirs , 197
Qi Stagnation, 471, 486, 509
Qigong Clinic, 50- 56, 64, 303, 304, 392, 419, 439, 507, 521, 548, 559, 621, 696, 717, 791, 824, 859, 863, 884, 886, 902, 904
Qigong Massage, 65, 543, 622, 623, 643, 673-675, 679, 683, 740, 839, 856
Qihai, 135, 198, 377, 433, 558, 669, 745
Quiescent, 278, 306, 318, 344, 355, 356, 391, 398, 406, 407, 449, 453, 493, 520, 572, 583, 591, 625, 687,689, 768, 781, 785, 779, 833, 857, 900

R

Rachialgia, 814
Radiation, 121, 128, 273, 302, 411, 461, 527, 551, 631, 707, 713, 836, 895, 897, 919-921, 927-929
Rebellious Qi, 208, 325, 432, 483, 487, 488, 594, 671, 745, 826, 857, 908, 920
Reconstructive Qi Therapy, 826
Red 124, 147, 154, 194, 198, 290, 294, 307, 308, 426, 470, 471, 477, 521, 529, 570, 573, 575, 633, 639, 658, 701, 751, 752, 777, 779, 784, 785, 797, 841, 851, 852, 854, 855, 857, 864, 868, 872, 915, 916
Reflexology, 61, 62, 112
Regression, 329, 445, 516-518, 546, 604, 621, 861, 887
Repression, 449, 578, 621, 890
Respiratory Dao Yin Training, 357, 397

Restrictive Cycle, 495, 497
Reverse Breathing, 394, 422, 695, 721, 726
Rhabdomyosarcoma, 896
Rheumatic Heart Disease, 777, 779, 785
Rheumatosis, 785
Rhinitis, 801, 802
Righteous Qi (Zheng Qi), 323, 324, 357, 486, 548, 552, 570, 590, 678, 910
Rigor, 772
Rooting, 380, 430, 432, 546-550, 603, 633, 636, 761
Ross, Elisabeth Kubler, 626

S

Sacred Healing Space, 541, 546, 547, 548, 549
Sadness, 75, 88, 99, 138, 170, 172, 339, 439, 446, 447, 452, 460, 469, 508, 615-617, 844, 874, 875, 878, 888, 892
Sagen, Thea, 927
Saliva, 75, 166, 324, 312, 372, 387, 491, 616, 708, 719, 743, 780
Salty Food, 313
San Bao (Three Treasures), 315
Sancier, Kenneth M., 904
Sarcoma, 896, 915
Sclera, 772, 915
Sea of Blood, 64, 71, 198, 199, 206, 253, 326, 329, 330, 670, 849, 851, 859
Sea of Energy, 326, 889
Sea of Grain and Water, 198
Sea of Marrow, 57, 64, 141, 143, 198, 253, 254, 259, 326, 330, 331, 461, 671, 838, 888
Sea of Qi, 133, 135, 137, 198, 252, 253, 669, 745, 887, 888
Sea of Yang Channels, 201
Sea of Yin Channels, 204
Seat of Emotions, 328
Second Lunar Month, 73
Self-Massage, 377, 386, 387, 388, 596, 687, 691
Self-Regulation Therapy, 57, 64, 65, 687
Seven Corporeal Souls, 75, 80, 83, 86-90, 134, 170, 251, 334, 338, 340, 341, 365, 401, 446, 447, 616
Seven Emotions, 98, 99, 467, 468, 469, 470, 488, 612, 843
Seven Essential Stars, 79, 80
Seven Internal Factors, 460
Seven Orifices, 292
Seventh Lunar Month, 79
Shaman, 529
Shang Ling (Pleasent Soul), 86, 87, 347
Shao Yang, 57, 109, 160, 222, 501-503, 612
Shao Yin, 57, 109, 160, 222, 501-503, 612
Shealy, C. Norman, 141
Shen, Chang, 532
Shengong (Spirit Skill Training), 56, 318, 326, 332, 342, 343, 348, 449, 541, 583
Shening Out, 342, 602, 604
Shimian Point, 907
Shu Point, 193, 249, 250, 251, 370, 611
Si Ming (Administrator of Destiny), 84-86, 90, 92, 347
Silver 84, 85, 151, 288, 307, 337, 348, 349, 431, 451, 452, 624, 905
Sinusitis, 382, 788, 792, 802, 809, 810
Sishencong Points, 149, 342, 548, 839
Sitting Meditation, 242, 337, 360, 785
Six External Factors, 460
Six Functions of Qi, 323
Six Stages, 57, 308, 462, 501, 503, 504, 510
Six Storage Areas, 78, 79
Six Transportations of Shen, 144, 344, 345, 346, 411
Sixth Lunar Month, 78, 79
Skin Zones, 160, 161, 163, 239, 240, 241, 276, 508
Sleep, 89, 91, 99, 166, 178, 210, 211, 212, 254, 274, 317, 323, 333, 335-337, 341, 360, 361, 407, 414, 421, 426, 438, 444, 458, 460, 475, 476, 509, 546, 563, 570, 604, 619, 651, 697, 722, 784, 815, 817, 818, 844, 847, 880, 884, 887, 891, 900
Small Intestine, 79, 98, 122, 136, 146, 160, 168, 172, 174, 178, 180-182, 191, 192, 194, 217, 219, 220, 222, 226, 235, 322
Small Intestine Channel, 180, 182, 194, 215, 219, 235, 779
Small Intestine Divergent Channel, 222, 226
Soaring Dragon Technique, 638
Social Oncology, 903, 904
Solid Organs, 165, 710, 801
Soul Extensions, 85, 90, 347
Soul Loss, 443, 444
Soul Retrieval, 91, 439, 442-450, 453, 577, 602, 606, 623, 633 688, 708, 844
Sound Vibration, 65, 117, 118, 155, 406, 701, 702, 710
Spider Nevus, 774
Spinal Pass, 605
Spiraling Energy, 647-649
Spirit Demon, 438-441, 882
Spirit Oppression, 438, 439
Spirit Possession, 143, 254, 438-440, 884, 885
Spirit Soul, 87, 273, 847
Spiritual Dryness, 582
Spleen, 75, 77-83, 99, 111, 122, 123, 160, 163, 166, 174-178, 184, 191-194, 281, 286, 290, 294, 311,

312, 380, 381, 383, 425, 432, 439, 452, 466, 469, 470, 472-474, 479-498, 503, 505, 509, 522, 525, 615, 616, 656-658, 665, 667, 669, 671, 675, 676, 677, 704, 707, 708, 709, 710, 713, 715, 717, 733, 747, 748, 761
Spleen and Stomach Massage, 797, 798
Spleen Channel, 77, 176-178, 194, 206, 209, 217, 219, 221, 234, 775, 789, 853, 930
Spleen Diseases, 710, 712, 787, 788, 789, 791
Spleen Divergent Channel, 222, 225
Splenomegaly, 694
Spring Equinox, 279, 280, 573
Stacking the Bones, 691, 693-695
Standing Meditation, 360, 399
Static Qigong Training, 355, 356, 695,726
Sternum, 139, 204, 260, 261, 600, 700, 731, 755, 795, 796, 808, 809
Stomach, 78, 79, 82, 83, 174-177, 180, 182, 191-194, 198, 276, 281, 293, 309, 311, 319, 321, 322, 363, 368, 385, 403
Stomach Cancer, 385, 567, 713, 714, 897, 915
Stomach Channel, 78, 108, 174, 176, 194, 217, 219, 233, 325, 504, 522, 776, 788, 793, 795
Stomach Diseases, 425, 713, 787, 792, 793
Stomach Divergent Channel, 222, 225
Stream Points, 251, 253, 370
Stroke, 64, 202, 250, 379, 492, 565, 567, 646, 788, 823, 828,863-865, 890
Subconscious Mind, 118, 134, 336, 344, 345, 403, 407, 409-412, 437-439, 441, 453, 458, 513, 517, 555, 561, 623, 625, 898, 927
Subtle Wonders, 359, 588, 589
Summer Solstice, 279, 280, 573
Sun, Simiao, 143, 309, 884
Super Ego, 621
Surgery, 95, 108, 124, 440, 442, 444, 445, 486, 536, 567, 625, 798, 826, 836, 857, 92-929, 930
Sweet Food, 311, 787
Sword Fingers, 433,636, 637, 652, 653, 659, 772, 784, 805, 807, 810, 819, 820, 826, 829, 855, 882, 923
Symptoms, 457, 461-464, 546, 570, 572, 578, 591, 596, 708
Syndrome, 461-464, 467, 468, 472, 474, 480

T

Tai Guang (Eminent Light), 86, 87
Tai Yang, 57, 109, 160, 222, 378, 501, 503, 612, 865, 884
Tai Yi (Great Divinity), 84-86, 92, 347
Tai Yin 57, 109, 160, 222, 501-503, 612, 884
Taiji Pole, 64, 69-71, 83, 84, 86, 88, 89, 91, 92, 118-120, 122, 125, 131, 132, 135, 144, 145, 147-154, 157, 161, 187, 197, 243, 280, 284, 288-290, 293, 294, 296, 306, 307, 326, 328, 331, 332, 334, 341, 342, 344, 347, 363, 364, 366, 386, 400, 410, 436, 451, 453, 494, 531, 532, 541, 547, 548, 550, 551, 559, 560, 572, 588, 594, 603, 606, 611, 645, 646, 701, 723, 740, 778, 782, 783, 812, 813, 824, 924
Taiji Ruler, 433, 434, 688
Tang, Rongchuan, 111
Tapping Method, 595, 596, 676, 678, 882
Taylor, Kylea, 580
Taylor, Stephanie, 53
Ten Thousand Voices, 408
Ten Wings, 271
Tenth Lunar Month, 82
The Six Stages, 308, 462, 501-504, 510
Third Eye, 92, 120, 142, 144, 332, 350, 410, 451, 520, 521, 531, 558, 576, 632, 726, 805, 831, 865
Third Lunar Month, 74, 86
Three Fires, 127
Three Formations of Cancer Growth, 896
Three Functions of Energetic Points, 247
Three Outer Forces, 132, 267-271
Three Periods of Life 457, 458
Three Treasures of Earth, 301, 313
Three Treasures of Heaven, 273, 297
Three Treasures of Man, 69, 267, 315, 316
Three Wonders, 588
Thrombosis, 863, 865, 866, 907
Thrusting Channels, 120, 148, 293-295, 342, 703, 704
Thrusting Method, 676, 677
Thrusting Vessel, 70, 168, 198, 200, 204, 206,-208, 219, 366, 909, 912
Tinnitus, 202, 215, 235, 254, 260, 741, 777, 781, 782, 784, 812, 814-820, 828, 843, 890
Tongue Positions, 372, 658
Tonification and Purgation Guidelines, 596
Tonification and Purgation Methods, 594
Tonification Techniques, 593
Traditional Chinese Medicine, 45, 47-50, 54, 56-58, 61, 67, 69, 86, 93, 95, 97, 100, 119, 138, 144, 159, 163, 164, 165, 176, 193, 217, 228, 249, 268, 272, 281, 312, 321, 330, 332, 401, 402, 457, 460, 465, 473, 474, 489, 490, 492-497, 509, 537, 548, 603, 673, 686, 772, 773, 775, 781, 791, 796, 843, 851, 852, 858, 859, 866, 867, 870, 881, 884, 895, 897, 906, 914, 929

Transference, 65, 68, 160, 315, 338, 488, 515, 572, 577, 579, 582, 689, 823, 874
Transformation, 48, 56, 91, 92, 95, 98, 99, 117, 119, 128, 132, 133, 137, 145, 172, 176, 182, 191-193, 199, 206, 259, 263, 276, 297, 299, 318, 319, 320-322, 338, 411-413, 415, 416, 419, 445, 453, 454, 457, 486, 489, 490, 491, 494, 496, 498, 502, 505, 513, 530, 532, 534, 554, 561, 580, 627, 721, 837, 844, 849, 859, 861, 904, 926
Transient Ischemic Attacks (TIA), 863
Transpersonal Point, 576, 661
Transportation, 111, 163, 172, 176, 192, 338, 490, 737, 798, 837
Trappier, Arthur, 126
Treating Phobias and Depression, 886
Treating Psycho-Emotional Disorders, 890
Treating Tumors and Cancer, 711, 906, 921
Treatment of Scar Tissue, 929
Trigger Points, 246
Trigrams of the Yi-Jing, 607
Triple Burner Channel, 76, 189, 190, 193, 194, 215, 220, 238, 707
Triple Burners, 57, 76-79, 127, 133, 135, 189-193, 222, 227, 282, 321, 374, 478, 479, 503, 504, 510, 612
Triple Burners Divergent Channel, 222, 227
True Fire 127, 135
True Qi (Zhen Qi), 81, 286, 321-324, 381, 382, 387, 726, 903, 906
True Self, 106, 139, 142, 541, 150, 331, 347, 420, 441, 454, 541, 621
True Spirit, 334
Tui Na, 57, 61-63, 65, 461, 565, 631, 839, 890, 930
Tumor, 53, 55, 327, 381, 395, 421, 440, 532, 688, 710, 711, 715, 855-857, 895-899, 902-907, 910, 913, 914, 916, 919, 922, 927, 928
Turbid Qi (Shi Qi), 191, 282, 288, 303, 391, 397, 419, 487, 488, 505, 536, 552, 569-572, 650, 676, 693-695, 701, 724, 751, 768, 778, 788, 789, 793-796, 802, 803, 812, 851, 925
Twelve Chakra Gates, 151, 152, 156, 157, 722
Twelve Divergent Channels, 222
Twelve Earthly Branches, 156, 157, 268-270, 281, 290, 291, 296, 297, 299, 721
Twelve Entry Points, 252
Twelve Exit Points, 252
Twelve Muscle/Tendon Channel, 163, 228
Twelve Pi Hexagrams, 166, 276, 277
Twelve Primary Channels, 71, 83, 84, 127, 133, 159, 160, 163-166, 189, 194-199, 206, 217, 222, 228, 239-241, 253, 281, 282, 284, 296, 611, 720, 850
Twelve Skin Zones, 163, 239, 240, 241, 276
Twelve Yuan-Source Points, 253
Two Gates of Energetic Points, 246

U

Ulcers, 155, 403, 466, 471, 695, 766, 789, 797, 798, 873
Universal Qi, 321, 380, 587, 737
Upper Burner, 57, 166, 191-193, 321, 322, 425, 491, 504, 505, 610, 651, 700, 729, 730, 791, 803, 816
Upper Dantian, 83, 84, 86, 87, 89, 120, 125, 131, 132, 137, 139-147, 155, 156, 161, 199, 283, 288, 289, 317, 328, 332, 342, 347, 350, 366, 371, 372, 377-379, 387, 388, 433, 451, 453, 513, 518-520, 522, 526, 547, 548, 558-560, 569, 576, 581, 598, 599, 603, 632, 636, 645, 646, 651, 726, 727, 740, 778, 783, 823, 824, 830, 840, 886
Uterine Cancer, 908, 911
Uterine Tumor, 858
Uterus, 69, 72-74, 76-82, 108, 133, 136,146, 366, 459, 471, 486, 505, 591, 667, 715, 908-911

V

Vaginal Discharge, 194, 210, 849-852, 910
Vertigo, 254, 495, 554, 672, 741, 770, 781, 784, 818, 828-830, 867
Violet, 121-123, 128, 155, 156, 526, 527, 633, 897
Virtue, 75, 76, 88, 89, 91, 138, 146, 147, 268, 332-334, 339, 340, 544, 549
Visualization, 49, 65, 87, 89, 243, 336, 343, 348, 356, 371, 403, 411, 443, 448, 449, 519, 545, 547, 554, 555, 561, 625, 683, 688, 701, 767, 770, 824, 861, 876, 877, 904, 929
Vlamynck, Jean Ruth, 50, 53, 54
Void (Wuji), 108, 316, 317, 331, 332, 356, 514-516, 612

W

Walking Therapy, 376, 377, 380-386, 868
Wandering Bi, 603
Water Cycle, 201, 269, 708, 720, 723, 839
Water Element, 98, 133, 251, 268, 293, 493, 653, 676
Weather, 124, 215, 270, 279, 280, 281, 290, 293, 301, 307, 322, 465, 479, 501, 816, 826, 827
Wei Lu Guan (Coccyx Pass), 605
Wei Qi (Protective Qi), 48, 57, 64, 126, 132, 137, 140, 145, 149, 151, 164, 322, 324-329, 549, 551, 552, 554, 559, 572, 575, 612

White, 46, 104, 120, 122-124, 137, 138, 147, 148, 206, 210, 241, 260, 288, 293, 294, 306, 307, 311, 327, 331, 338, 406, 412, 414, 415, 425, 431, 439, 451, 452, 471-473, 476, 492, 497, 501, 505, 508, 520-523, 529, 531-534, 536, 547, 573-576, 589, 617, 624, 639, 658, 684, 685, 701, 713, 719, 726, 803, 804, 806, 851-853, 855, 861, 898, 901, 904, 905, 911, 919, 920, 922, 925, 926
Willpower (Zhi), 151, 338, 340, 341, 348, 574, 847
Wind Bi, 603
Wind Cold, 217, 464, 467, 472, 669, 801, 802, 805, 807, 809, 816, 817, 829
Wind Damp, 464, 802, 805, 920
Wind Heat, 217, 467, 472, 505, 669, 670, 798, 801, 802, 805, 809
Wind Phlegm, 492, 828
Wind Stroke, 202, 492, 788, 823
Window of the Sky Points, 254, 886, 887
Windy Breathing Method, 392, 394, 396, 768, 779, 789, 813
Wong, Her Yue, 708
Wood Element, 98, 100, 293, 493, 556, 653, 676, 914
Wood Jing, 79
Worry, 64, 75, 88, 99, 101, 104, 108, 129, 138, 172, 178, 348, 403, 404, 427, 429, 432, 439, 460, 488, 508, 603, 615, 616, 619, 794, 796, 798, 818, 850, 855, 859, 873, 874, 876, 878, 892, 903, 907
Wu Guan (Five Passes), 604
Wu Huo Breathing Method, 394
Wu Jing Shen (Five Essence Spirits), 80, 84, 90, 141, 409, 443, 57, 847, 860, 870, 871
Wu, Ju-Tong, 504
Wu Ying (Without Essence), 84, 86
Wu Zang (Five Yin Organs), 294, 572, 573, 574, 688
Wuji, 45, 85, 132, 144, 148, 149, 267, 316, 317, 332, 344, 345, 356, 371, 398, 415, 514-516, 520, 633

X

Xi Yuan Hospital of Beijing China, 51, 52, 67, 303, 406, 439, 499, 566, 627, 717, 898
Xia Tao Kang (Below Healthy Peach/Life), 84-86, 92, 347
Xiphoid Process, 109, 111, 154, 600, 772, 808, 809
Xu, Hongtao, 52

Y

Yan, De-xin, 688
Yang Channels, 71, 72, 160, 161,165, 166, 172, 174, 180, 182, 189, 558, 595, 651, 690, 734, 745
Yang Deficiency, 472, 480, 495, 502, 777, 781, 783, 785, 788, 812, 821, 823, 851, 876
Yang Fire, 103, 137, 200, 242
Yang Heel Vessels, 70, 71, 107, 197, 198-200, 210-212, 215, 251, 259, 377, 381
Yang Linking Vessels, 70, 71, 108, 197, 200, 213-215, 251
Yang Ming, 57, 160, 222, 501, 503, 612
Yang Organs, 197, 198, 222, 242, 247, 253, 254, 257, 837, 870
Yang Polarity, 152, 268
Yang Shen Disturbances, 430
Yang Yao, 166, 276-278
Yao Cycles, 166, 276, 278
Yao Lines, 109, 612
Ye, Tian Shi, 502
Yellow, 49, 68, 99, 106, 122-124, 136, 139, 140, 147, 154, 193, 198, 241, 319, 355, 386, 429, 430, 431, 526, 529, 558, 560, 572, 598, 600, 607, 624, 645, 646, 657, 728, 768, 775, 779, 782-785, 787, 789, 790, 793, 797-799, 806, 819, 828, 829, 832, 851, 853, 854, 868, 905, 914-916
Yellow Emperor, 49, 68, 99, 106, 193, 198, 319, 355, 429, 529, 572, 787
Yi (Intent), 75, 80, 84-86, 92, 109, 139, 141, 176, 333-335, 337, 338, 340, 341, 347, 401, 402, 407, 411, 425, 427, 429, 573, 607, 616, 636, 787, 847, 873
Yi-Jing (Book of Changes: I-Ching), 268, 271
Yin Channels, 71, 160, 161, 165, 168, 170, 176, 178, 184, 187, 213, 214, 223, 251, 690, 868, 884
Yin Deficiency, 204, 242, 474, 480, 488, 490, 491, 495, 502, 505, 768, 775, 776, 778, 785, 788, 791, 792, 796, 797, 802, 808, 809, 812, 828, 853, 881, 913
Yin Essence, 200, 257, 278, 285, 321,859
Yin Heel Vessels, 71, 107, 108, 199, 210-212
Yin Linking Vessels, 70, 71, 108, 200, 213, 214
Yin Organs, 250, 253, 257
Yin Shen Disturbances, 430
Yin Tang (Third Eye Point), 120, 131, 143, 144, 332, 350, 451, 520, 521, 531, 632, 636, 672, 726, 733, 831, 832, 833, 840, 865, 866, 883
Yin Yao, 277-279
Ying Qi (Nutritive Qi), 57, 64, 137, 163, 166, 191, 193, 322, 324, 326, 329, 330, 491, 502-504, 676, 678, 679, 747, 888
Yu Jing (Hidden Essence), 86, 87, 347
Yu Vessels, 213, 215

Yu Zhen Guan, 605
Yuan Jing, 133, 316-318, 324, 331, 333, 633
Yuan Qi, 70, 133, 135, 137-139, 191, 197, 199, 201, 204, 252, 253, 279, 316-324, 333, 368, 371, 383, 387, 391, 398, 466, 487, 560, 588, 590, 665, 667, 688, 720, 724, 726, 733, 737, 749, 811, 814, 845, 866, 897, 898, 902, 903, 906
Yuan Shen, 84-86, 90-92, 118, 138, 142, 187, 201, 259, 268, 316-318, 333, 335, 337, 338, 343, 344, 345-347, 371, 401-404, 408-410, 419, 427, 440-451, 514, 528, 633, 723, 726

Z

Zhang, Zhong Jing, 501
Zhao, Joanna, 50, 54
Zhao, Xianke, 127
Zhen Qi (True Qi), 81, 286, 322, 324, 903
Zheng Qi (Righteous Qi), 282, 323, 324
Zheng, Zhan Ding, 404, 518, 524
Zhi (Will), 75, 76, 79, 80, 85, 90, 102, 118, 139, 141, 151, 187, 201, 333-335, 338, 340, 341-347, 811, 847, 876
Zhi Shen (Postnatal Spirit), 85, 90, 201, 333, 335, 343-347, 401-404, 408-412, 427, 446
Zhi, Zhuan, 699
Zhong Qi (Center Qi), 321, 322, 324
Zhou Qi (Turbid Qi), 321, 324
Zhou, Qianchuan, 49, 99
Zong Qi (Gathering Qi), 80, 137, 138, 191, 322, 324

ABOUT THE AUTHOR

- Dean of Medical Qigong Science - Five Branches Institute, College and Clinic of Traditional Chinese Medicine (U.S.A.)
- Director of Medical Qigong Clinic - Five Branches Institute, College and Clinic of Traditional Chinese Medicine (U.S.A.)
- Director and Founder of the International Institute of Medical Qigong - (United States, Canada, South Africa, and Europe).
- Professor of Medical Qigong Science and Philosophy - Licensed by the Beijing Western District Qigong Science and Research Institute (China)
- Physician in Charge of Traditional Chinese Medicine - Licensed by the Beijing Bureau of Scientific Technology (China)
- Doctor of Medical Qigong and Anmo Therapy - Licensed by the Beijing Western District Qigong Science and Research Institute (China)
- Graduate Degree in Medical Qigong Therapy - Medical Qigong College - Hai Dian University of Beijing, (China)
- Council Board Member (United States Representative)- World Academic Society of Medical Qigong, (China)
- Council Board Member (Medical Qigong Committee)- National Qigong Association (U.S.A.)
- Ambassador of Chinese Medical Qigong Therapy to North America - Commissioned by The World Health Organization Collaborating Center for Traditional Chinese Medicine, Beijing, China (sponsored by the United Nations)

MEDICAL BACKGROUND

Dr. Jerry Alan Johnson is one of the few internationally recognized non-Chinese Grand Masters, practicing doctors and Director/Professors of Medical Qigong Therapy. Having studied for more than twenty-eight years, he is recognized both in China and the West as America's leading authority on Medical Qigong therapy. Dr. Johnson is licensed as a Doctor of Traditional Chinese Medicine (D.T.C.M.) in Beijing, China, and has served with national and international committees to promote and encourage the practice of Medical Qigong.

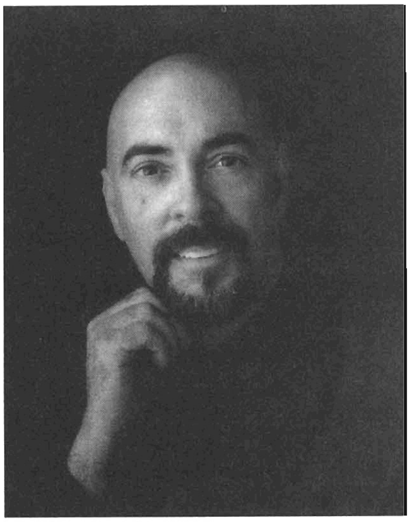
Dr. Jerry Alan Johnson

Dr. Johnson began his initial training in energetic medicine in 1972. He started his formal clinical studies of Traditional Chinese Medicine in Monterey, California in 1974, where he focused his studies on acupuncture, herbology, Medical Qigong therapy, traumatology and Chinese massage. After receiving certification as an Acupuncturist and Herbalist from an 1978-81 internship, Dr. Johnson operated clinics in Colorado Springs, Colorado (1981-84), working as a consultant and associate in Acupuncture and Medical Qigong Therapy to Western Medical Doctors and Chiropractors.

In 1993 Dr. Johnson furthered his clinical studies at the China Beijing International Acupuncture Training Center, and at the Acupuncture Institute of China, Academy of T.C.M. He was the first foreigner from the Acupuncture Institute invited to treat patients in the Medical Qigong Clinic at the Xi Yuan Hospital of T.C.M. (featured in Bill Moyer's special "Healing and the Mind"). There he interned as a Doctor of Medical Qigong Therapy, specializing in the treatment of cysts, tumors, and kidney dysfunctions.

Dr. Johnson also interned as a Doctor of Anmo Therapy at the Orthopedic Traumatology Clinic of the Hu Guo Si Hospital of T.C.M. (in Beijing, China) specializing in Chinese bone-setting, traumatology and tissue manipulation. His training includes Neuro-muscular Therapy, Psychophysical Integrational Therapy, Advanced Visceral Manipulation and Advanced Cranio-Sacral Therapy from the Milne Institute of Healing Arts.

In 1995 Dr. Johnson received his Graduate Degree in Medical Qigong Therapy from the Medical Qigong College - Hai Dian University of Beijing, China, and was licensed as a Doctor of Medicine Qigong. The Beijing Western District Qigong Science and Research Institute awarded Dr. Johnson a license as "Professor of Medical Qigong Science and Philosophy" after his completion of both academic doctoral thesis, and numerous clinical field studies (1995).

In 1997 Dr. Johnson became the first Qigong doctor allowed to assist in surgery at the Community Hospital of the Monterey Peninsula, in Carmel, California, where he continues to work closely with Western Doctors.

In 1997 Dr. Johnson was elected to the Board of Directors of the National Qigong Association-USA. In 1998 he was appointed the Chairman of the Medical Qigong Committee, responsible for establishing the guidelines and national standards for Medical Qigong practitioners working in clinical settings.

In 1998 Dr. Johnson was elected to the Board of Directors of the World Academic Society of Medical Qigong, centered in Beijing, China, and serves as one of four Council Board Members that represent the United States at the World Conferences. Also, in 1998, his clinical status was elevated to the position of "Physician in Charge of Traditional Chinese Medicine" from the Beijing Bureau of Scientific Technology, in China.

In the Spring of 1999 Dr. Johnson accepted the position of Dean of Medical Qigong Science from the Five Branches Institute, College and Clinic of Traditional Chinese Medicine, in Santa Cruz, California. Dr. Johnson developed a Masters Program which includes instruction and training from his textbook entitled: *Chinese Medical Qigong Therapy, A Comprehensive Clinical Text*. The course also includes the specialization of specific Medical Qigong Therapies such as Pediatrics, Geriatrics, Gynecology, Neurology, Psychology, Oncology and Surgery.

In the Spring of 2000 Dr. Johnson established within the Five Branches Institute the first Medical Qigong Clinic at a T.C.M. College in North America, facilitating the combined use of acupuncture, herbs and Medical Qigong therapies to the general public.

Dr. Johnson is also the founder and director of the International Institute of Medical Qigong, in Monterey, California, and has designed a three-year Medical Qigong program which has been implemented by several colleges and Medical Qigong Institutes throughout the United States, Canada, Europe and South Africa.

Dr. Johnson serves as the official representative for North America for The World Health Organization Collaborating Center for Traditional Medicine, sponsored by the United Nations for the promotion and development of Traditional Chinese Medical Qigong Therapy.

An active member of the World Academic Society of Medical Qigong, Dr. Johnson was a key speaker at the 1996 Third World Conference on Medical Qigong, held in Beijing, China, the 1997 Second World Congress on Medical Qigong held in San Francisco, California, the First American Qigong Association Conference, held in San Francisco, California, the 1998 Fourth World Conference on Medical Qigong, held in Beijing, China, and the 1999 National Qigong Gathering held in Baltimore, Maryland. He has lectured and taught many seminars at The American College of Traditional Chinese Medicine in San Francisco, California, The University of California San Francisco, The Qigong Institute of Menlo Park, California, The International Training Center at San Diego State University, The Monterey Institute of International Studies, The Asilomar Conference Center of Pacific Grove, California, as well as several other colleges, institutes and hospitals throughout North America.

MARTIAL ARTS BACKGROUND

Dr. Johnson is also internationally renowned as a Shifu (Master Instructor) of several Chinese Martial Arts Systems, having studied the martial arts for over 34 years.

In Shaolin, he has studied Praying Mantis, Hung Gar, Wing Chung, and Mi Zongquan Shaolin styles of fighting, weapons and Martial Qigong, and is a certified Master Instructor in Mi Zongquan Shaolin, teaching through the authority of the Mi Tsung-I Northern Shaolin Federation of Taiwan.

In Baguazhang he has studied several schools of fighting, including the Yin Fu, Cheng Ting-Hua, Liu De Kuan, Fu Zhen-Song, and the Chang Zhao-Dong Styles of combat, weapons and Martial Qigong. He is a certified Master Instructor in Chang Zhao-Dung Baguazhang, teaching through the authority of the Ching Yi Baguazhang Association, the Hsiao Mien Hu Martial Arts Association, and The Canadian Martial Arts Society.

In Taijiquan, Dr. Johnson has studied several schools of fighting, including the Yang Pan-Hou, Yang Chien-Hou, and Chen Fa Ke styles of combat, weapons and Martial Qigong. He is a certified Master Instructor as well as a "Twentieth Generation Inner-Door Disciple" from Feng Zhi Qiang's Chen Style Taijiquan, teaching through the authority of the Beijing Martial Arts Association and the Beijing Chen Style Taijiquan Association of China. He is also a certified Master Instructor in the Yang Style Taijiquan, teaching through the authority of the Tai Chi Chuan Federation of Taiwan.

Excelling in "Empty Hand," as well as "Weapons Fighting" from both the Shaolin and Wudang Martial Arts Systems, Dr. Johnson competed regularly in the tournament circuit and was titled: 1980 Southbay National Champion, 1982 Tournament of Fighters Champion, 1982 Universal Tae Kwon Do/Tang Soo Do Invitational Grand Champion, 1983 Colorado National Kung Fu Champion, 1983 Tournament of Fighters Grand Champion, 1984 Colorado National Kung Fu Champion, and the Tournament of Fighters Kung Fu Champion in 1984, when he retired from competition.

In addition to his own championships, many of Dr. Johnson's students also went on to compete at the national level, bringing several championship titles to the school. Like their teacher, Dr. Johnson's students excelled in free-form fighting, Taiji Push-hands and Chinese Gongfu forms, as well as weapon competitions.

In 1986, the state of California issued Dr. Johnson teaching credentials, allowing him to teach adult education classes in Taijiquan and Baguazhang. Since then, he has been a council member of the International Baguazhang Research and Teachers Exchange, and served as a judging official and member of the Executive Committee of the U.S. Chinese Kuo Shu Federation, as well as a coaching official, national certifier and council member of the Amateur Athletic Union (affiliated with the U.S. Olympics). Dr. Johnson also participated on the Board of Advisors for the United States of America Wushu-Kung Fu Federation, and is a member of the International Congress of Oriental Medicine and Martial Arts. He was also a standing member of the North American Chinese Martial Arts Federation.

Author and Publisher:

Dr. Johnson was a contributing author, selected to write the Medical Chi Kung section, for the book *The Complete Illustrated Guide to Chi Kung*, a cooperative project between Element Books (UK) and author James MacRitchie. Dr. Johnson was also one of the selected contributing authors for the book *Qigong-Essence of the Healing Dance*, a cooperative project between P.B.S. and Documentary Producer/Director Garri Garripoli.

Dr. Johnson has been interviewed many times on local radio programs, and has been featured in numerous local and international newspapers and magazines, including: *Newsweek Magazine* "The International Addition," *The World Medical Qigong Quarterly*, *Qi-The Journal of Traditional Eastern Health & Fitness*, *The Empty Vessel Magazine*, *Pent-*

house Magazine, Self Magazine, Billboard Magazine, The Mystical World of Chinese Martial Arts, Inside Kung Fu Magazine, Combat Karate Magazine, Internal Arts Magazine, Black Belt Magazine, The Pa Kua Chang Journal, and Inside Karate's Master Series Magazine.

Televised specials include: *NBC News* "Dateline," *CBS/Channel 46* "Eye on America," featured in a documentary on Medical Qigong filmed for the *"Discovery Channel,"* featured in a documentary on Medical Qigong filmed for SBS "Qi-into the world of the unknown," and *K.P.B.S.* "Managing Health and Productivity; New Approaches In Technology," as well as *The David Letterman Show.*

In 1988, Dr. Johnson created and published a six level "Pa Kua Chang Chi Kung Correspondence Course," and in 1989, founded the "Pa Kua Chang News Letter," which later became known as the "Pa Kua Chang Journal," and was distributed worldwide. Dr. Johnson has written and published five books on martial arts entitled:

- The Secrets of the Eight Animals,
- The Masters Manual of Pa Kua Chang,
- Classical Pa Kua Chang Fighting Systems and Weapons,
- The Essence of Internal Martial Arts Volume 1: Esoteric Fighting Techniques and Healing Methods, and
- The Essence of Internal Martial Arts Volume 2: Energy Theory and Cultivation.

From 1984 to 1986, Dr. Johnson produced and directed six Baguazhang instructional videos entitled:

- The Eight Animal School of Pa Kua Chang
- The Original Form of Pa Kua Chang
- The Eight Circular Pa Kua Staff
- The Eight Circular Pa Kua Broadsword
- The Dragon School of Pa Kua Chang, and
- The Fighting Techniques of Pa Kua Chang.

In 1994 he was featured in two videos on Taijiquan, as well as one Qigong video narrated by actor John Saxon, entitled:

- Tai Chi-The Empowering Workout,
- Power Tai Chi- Total Body Workout, and
- Chi Kung-The Healing Workout,

He additionally created two Meditation CD / cassette tapes with composer John Serri entitled:

- Tai Chi Meditation - [1] Life Force Breathing, and
- Tai Chi Meditation - [2] Eight Direction Perception.

Dr. Johnson's latest project includes 10 Medical Qigong videos which contain instructional information and training from his current textbook *Chinese Medical Qigong Therapy, A Comprehensive Clinical Text*, entitled:

- Gathering Energy from Heaven and Earth,
- Stationary and Dynamic Medical Qigong Posture Training,
- Treating Patients with Medical Qigong Therapy (Vol.1),
- Treating Patients with Medical Qigong Therapy (Vol.2),
- Medical Qigong Energy Techniques and Qi Emitting Methods,
- Medical Qigong Invisible Needle Technique, Five Element Qigong Massage, and Energetic Point Therapy,
- Medical Qigong Healing Sound Therapy and Prescriptions,
- Treatment of Internal Organ Diseases with Medical Qigong,
- Treatment of Cysts, Tumors, and Cancer with Medical Qigong Therapy, and
- Soul Retrieval.

Dr Johnson's books, video tapes, meditation C.D.'s and cassette tapes have been translated into other languages and are currently being sold around the world. They have been featured in *Publishers Clearing House, Columbia House, Borders Book and Music Co., The Book of the Month Club, T'ai Chi Magazine, Century Martial Arts, Qi-The Journal of Traditional Eastern Health & Fitness, The Complete Guide to Exercise Videos, the Double Day Book and Music Co, and Direct TV (both in the United States and Japan).*

For more information about the author, the reader can turn to his web page at:

www.qigongmedicine.com

ADDITIONAL BRANCHES OF THE INTERNATIONAL INSTITUTE OF MEDICAL QIGONG

The following is a list of graduating students and instructors, as well as national and international directors of the International Institute of Medical Qigong.

BRANCHES WITHIN THE UNITED STATES

Jerry Alan Johnson, Ph.D.,D.T.C.M.,
D.M.Q. (China)
Director and founder of the International
Institute of Medical Qigong
(Main Branch)
P.O.Box 52144
Pacific Grove, California 93950 USA
Phone: (831) 646-9399
Fax: (831) 646-0535
E-Mail: drjerryalanjohnson@eatrhlink.net
Web Page: www.qigongmedicine.com

Seth Lefkowitz, D.C., D.M.Q. (China)
Director of the International Institute of
Medical Qigong
(South San Francisco Branch)
351 South Baywood Avenue,
San Jose, California 95128 USA
Phone (408) 243-1565
Fax: (408) 243-1568
E-Mail: naturespath4me@yahoo.com

Arnold Tayam, D.M.Q. (China)
Director of the International Institute of
Medical Qigong
(San Jose/Silicon Valley Branch)
c/o The Longevity Center
P.O. Box 26712
San Jose, California 95159-6712 USA
Phone (408) 295-5911
Fax: (408) 971-0136
E-Mail: chi@longevity-center.com
Web Page: www.longevity-center.com

Jean Ruth Vlamynck, L.Ac., Dipl.Ac.,
M.T.C.M., M.Q.T.
Director of the International Institute of
Medical Qigong
(Santa Cruz Branch)
257 B Center Ave.
Aptos, California 95003 USA
Phone: (831) 689-9093
Fax: (831) 689-9094
E-Mail: jeanhere@bigplanet.com
Web Page: www.relaxationresources.com/
assoc/jeanv.html

Dale M. Hopkins, M.Q.P.
Director of the International Institute of
Medical Qigong
(Northern Santa Cruz Branch)
24345 Loma Prieta Ave.
Los Gatos, California 95033
Phone: (408) 353-2566

Wendy Lang, M.Q.P.
Director of the International Institute of
Medical Qigong
(Marin County Branch)
12 Skylark Dr. Apt 33
Larkspur, California 94939
Phone: (415) 924-7883
E-Mail: wendylang@altavista.com

Andy Balestracci, M.Q.P., and
Linda MacElwee, M.Q.P.
Director of the International Institute of
Medical Qigong
(Mendicino Branch)
P.O. Box 43
Philo, California 95466
Phone: (707) 895-9115

Geoffery Greenspahn, M.Q.T.
Director of the International Institute of
Medical Qigong
(Los Angeles Branch)
Los Angeles, California
Phone: 310-775-5022
E-Mail: rumblej@hotmail.com

Todd Gedryn, M.Q.T.
Director of the International Institute of
Medical Qigong
(San Francisco Branch)
c/o Qigong Institute of North America
9 White Tail Lane
Monterey, Ca 93940 USA
Phone (408) 221-0628
Fax: (831) 373-8149
E-Mail: todd@qigonginstitute.net
Web Page: www.qigonginstitute.net

Matthew B. Weston, M.Q.T.
Director of the International Institute of
Medical Qigong
(Big Sur Branch)
c/o South Coast Qigong Center
P.O. Box 223373
Carmel, California 93922 USA
Phone (831) 626-3347
E-Mail: qigongr@mbay.net

Ted Cibik, M.Q.P.
Director of the International Institute of
Medical Qigong
(Pennsylvania Branch)
c/o Inner Strength Inc.
R.D. #4 Box 64
Leechburg, Pa 15656 USA
Phone (724) 274-0586
Fax: (724) 845-7739
E-Mail: ted@inner-strength.com
Web Page: www.inner-strength.com

Judith "Lacy" Storey, L.Ac.
Director of the International Institute of
Medical Qigong
(Colorado Branch)
Breckenridge, Colorado USA
Phone (303) 274-0586
E-Mail: dakini@colorado.net

Ellie Christianson, L.L.C., M.Q.P.
Director of the International Institute of
Medical Qigong
(Arizona Branch)
481 West Greys Road
Tuson Arizona 85718
E-Mail: ellie@MIOBI.com

Donald Mead, M.Q.P.
Director of the International Institute of
Medical Qigong
(Georgia Branch)
1008 Courtenay Drive NE
Atlanta, Georgia 30306-3330
Phone: (404) 873-6557
E-Mail: dmm10@mindspring.com

Connie Comer, M.Q.P.
Director of the International Institute of
Medical Qigong
(Oregon Branch)
1427 N.W. 23rd Ave.
Portland, Oregon 97210
Phone: (503) 226-6822
E-Mail: conniecomer@hotmail.com

INTERNATIONAL BRANCHES

Cynthia McMullen, C.M.T., M.Q.P.
Director of the International Institute of
Medical Qigong
(Alaska Branch)
3713 E. 19th Ave.
Anchorage, Alaska 99508
Phone: (907) 279-0135
E-Mail: touchoftaoCM@netscape.net

Robert C. Bawol, M.Q.P.
Director of the International Institute of
Medical Qigong
(New Mexico Branch)
1145 N. Alameda
Las Cruces, New Mexico 88005
Phone: (505) 647-8827
E-Mail: arrow@HTG.net

Shirley Norway
Director of the International Institute of
Medical Qigong
(Illinois Branch)
P. O. Box 1349
Fairfield, Illinois 52556
Phone: (515) 472-8404

Andy James, M.Q.T.
Director of the International Institute of
Medical Qigong
(Canadian Branch)
c/o Canadian Practitioners' Association
of Asian Medicine
Toronto, Canada M4K IK4
Phone: (416) 463-9360
Fax: (416) 463-0045
E-Mail: emerge@powerofbalance.com

Richard Garnier
Director of the International Institute of
Medical Qigong
(Europe Branch)
c/o Chariot D'or
Villa Samicky 171 che du Suveran
06140 Vence, France
Phone/Fax: 33 493 58 51 20
E-Mail: chardor@club-internet.fr

Susan Vial, M.Q.P.
Director of the International Institute of
Medical Qigong
(South Africa Branch)
P.O. Box 235
Magaliesburg, 2805 Gauteng
South Africa
E-Mail: saspect@iafrica.com

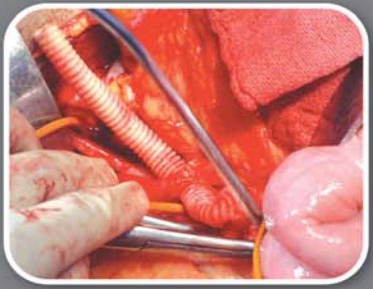

SRB's
Manual of Surgery

MANUAL OF SURGERY

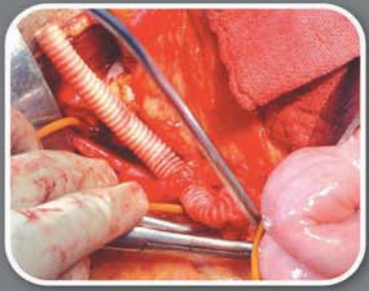

Fourth Edition

SRB's Manual of Surgery

Sriram Bhat M MS (General Surgery)
Professor
Department of Surgery
Kasturba Medical College
Mangalore, Karnataka, India

Honorary Surgeon
Government Wenlock Hospital
Mangalore, Dakshina Kannada, Karnataka, India
e-mail: meera_sriram2003@yahoo.com

Foreword
Prakash Rao

JAYPEE BROTHERS MEDICAL PUBLISHERS (P) LTD
New Delhi • Panama City • London • Philadelphia (USA)

Jaypee Brothers Medical Publishers (P) Ltd

Headquarters
Jaypee Brothers Medical Publishers (P) Ltd
4838/24, Ansari Road, Daryaganj
New Delhi 110002, India
Phone: +91-11-43574357
Fax: +91-11-43574314
Email: jaypee@jaypeebrothers.com

Overseas Offices

J.P. Medical Ltd
83, Victoria Street, London
SW1H 0HW (UK)
Phone: +44-2031708910
Fax: +02-03-0086180
Email: info@jpmedpub.com

Jaypee Brothers Medical Publishers (P) Ltd
17/1-B Babar Road, Block-B, Shaymali
Mohammadpur, Dhaka-1207
Bangladesh
Mobile: +08801912003485
Email: jaypeedhaka@gmail.com

Jaypee-Highlights Medical Publishers Inc
City of Knowledge, Bld. 237, Clayton
Panama City, Panama
Phone: +507-301-0496
Fax: +507-301-0499
Email: cservice@jphmedical.com

Jaypee Brothers Medical Publishers (P) Ltd
Shorakhute, Kathmandu
Nepal
Phone: +00977-9841528578
Email: jaypee.nepal@gmail.com

Website: www.jaypeebrothers.com
Website: www.jaypeedigital.com

© 2013, Jaypee Brothers Medical Publishers

All rights reserved. No part of this book may be reproduced in any form or by any means without the prior permission of the publisher.

Inquiries for bulk sales may be solicited at: jaypee@jaypeebrothers.com

This book has been published in good faith that the contents provided by the author contained herein are original, and is intended for educational purposes only. While every effort is made to ensure accuracy of information, the publisher and the author specifically disclaim any damage, liability, or loss incurred, directly or indirectly, from the use or application of any of the contents of this work. If not specifically stated, all figures and tables are courtesy of the author. Where appropriate, the readers should consult with a specialist or contact the manufacturer of the drug or device.

SRB's Manual of Surgery

First Edition : 2004
Second Edition : 2007
Third Edition : 2009
Reprint : 2010
Fourth Edition : **2013**

ISBN 978-93-5025-944-3

Printed at

God could not be everywhere all the time,
so, He created Mothers

This book is in the memory of my mother
Late Mrs Devaki Krishna Bhat
who continues to inspire and guide me

Foreword

> *A true surgeon is never fearless; he fears for his patients, he fears for his shortcomings, his own mistakes, but never fears for himself or his professional reputation.*
>
> **—Samule J Mixter**

The manual is written for the students in this firing line for their MBBS degree. Common conditions seen on a day-to-day basis have been dealt with a little more detail. The uncommon and rare conditions have been dealt briefly. The mode of clinical presentation and the physical signs are mentioned vividly so as to register in the mind of the reader. There is a brief mention of necessary investigations in each of the surgical conditions. The accepted mode of recent advances in investigation and treatment has been included. This manual is meant for Final MBBS degree examination and not a consolidated textbook. Certain chapters are included at the end of the book for the more enthusiastic students and for those interested in participating in quiz programmes.

The clinical photographs are of very good quality, which are self-explanatory of the conditions. The clinching physical signs have been highlighted in box forms with different shades of colour. This mode of presentation emphasises the point to be conveyed to the students which they can recollect when required either while arriving at a diagnosis in their examination or while in practice after obtaining the degree. This book gives a good foundation in surgery for those who wish to pursue surgery as a career.

What does the surgeon ought to be?

The conditions necessary for the surgeon are four—*first*, he should be learned; *second*, he should be an expert; *third*, he must be ingenious; and the *fourth*, he should be able to adapt himself.

Foremost, it is required that the surgeon should know not only the principles of surgery, but also those of medicine in theory and practice; second, he should have seen others operate; third, he should be ingenious, of good judgement and memory to recognise conditions; and the fourth, he should be adaptable and able to accommodate himself to circumstances.

> *Let the surgeon be bold in all sure things, and fearful in dangerous things, let him avoid all faulty treatments and practices. He ought to be gracious to the sick, considerate to his associates, cautious in his prognostications. Let him be modest, dignified, gentle, pitiful, merciful; neither covetous nor an extortionist of money; but rather let his reward be according to his work, to the means of the patient, to the quality of the issue, and to his own dignity.*
>
> **—Ars Chirurgica**

I thank Sriram Bhat M for giving me this privilege of writing a foreword to his excellent manual.

Prakash Rao MS
Retd. Professor of Surgery
Kasturba Medical College
Mangalore, Karnataka, India

Preface to the Fourth Edition

It is my pleasure to release the *fourth edition* in four years after the release of third edition of *SRB's Manual of Surgery*. Earlier editions were well-accepted by undergraduates and postgraduates. I felt that I should bring out the fourth edition in the intention of making it much more updated with texts and illustrations. I have done extensive corrections in this edition. I added many new topics, text details, and additional topics in different chapters. Morbid obesity, gastrointestinal fistulas, and surgical audit are the newer topics added. Malignancies have been discussed in much more detail with new concept of staging and with present oncological trends. Mistakes are corrected and insufficient topics are upgraded. I have referred many books, journals and taken the help of my colleagues in surgery and other departments. I have retained all chapters with some rearrangements of few chapters and topics. Self-assessment questions at the end of each chapter are removed which allowed me to add much more text material in the same space.

I sincerely express my thanks to everybody who have helped me to bring out this edition, and also to the publishers who are the backbone of this upgraded edition. I hope this new book will be well-accepted by the teachers, undergraduates and postgraduates of surgery department. I sincerely welcome all criticisms.

Sriram Bhat M

Preface to the First Edition

This book is born out of a desire to provide a complete, authoritative, current textbook on Surgery. For the past fifteen years of my teaching profession, watching students hurriedly jotting down notes during classes, an urge was born within me to put my thoughts in print. I am fortunate to have been guided by my teachers and supported by my students towards this goal.

An attempt has been made to make this book simple and reader-friendly while not compromising on the aspect of providing students with necessary information and a better clinical understanding of surgery. Recent advances in surgery till date have been included to the best of my knowledge. Text has been presented in a simple and lucid language so as to help students understand and recapitulate the subject better. Since a picture is worth than thousand words, over 500 photographs, X-rays and illustrations have been incorporated in this work to make surgery more interesting and understandable. Inspirational quotes have been interspersed to motivate students to go the extra mile.

Discussion on Instruments and Operative Procedures has been dealt with in necessary detail separately. A chapter containing a list of interesting new developments in surgery and uniquely new topics like Fascinating Signs, Misnomers, Triads and other interesting chapters have been included.

While this book has been compiled with undergraduates in mind, I am confident that it will serve as a useful reference for postgraduates and practitioners.

In compiling this book I have consulted many authoritative books and publications and I sincerely express my appreciation and gratitude to all of them.

Suggestions and constructive criticisms towards improving this book in subsequent editions are always welcome.

Sriram Bhat M

Acknowledgements

I am happy to bring out one more edition (fourth) of the book *SRB's Manual of Surgery*. This is due to constant help and support of many.

I thank our Chancellor, Dr Ramdas M Pai; Pro Chancellor, Dr H S Ballal; Vice Chancellor of Manipal Academy of Higher Education (MAHE), Professor K Ramnarayan; beloved Dean, Professor Venkatraya Prabhu; our associated Deans, Professor Anand Kini and Professor Vivian D'Souza for their academic support.

I thank Dr Alfred Augustine, Professor and Head, Department of Surgery, Kasturba Medical College (KMC), Mangalore, Karnataka, India, for his constant encouragement in academic work and progress.

I always remember my senior teachers, Professor CR Ballal and Professor Suresh Kamath, Professor K Prakash Rao, Professor Subramanya Bhat, Professor Bhaskar Shetty and Professor Ramachandra Pai for their constant help.

Professor Prakash Rao is a guide and encourager to everybody who will be remembered as a special person by me always.

My special thanks to Dr Jayaram Shenoy and Dr Jayaprakash Rao for their affection and support.

Surgical unit heads in our college Dr BM Nayak, Dr Thangam Verghese, Dr Harish Rao, Dr Shivaprasad Rai, Dr Yogishkumar, and Dr Shivananda Prabhu are always supportive for my work and are worth to be remembered always.

I am grateful to all my teachers and colleagues in surgery department who directly or indirectly helped me to bring out this fourth edition.

I appreciate District Medical Officer and Resident Medical Officer of Government Wenlock Hospital, Mangalore, Dakshina Kannada, Karnataka, India, for their kind help.

I will never forget my close associates, Dr Ganapathy, MD, Mangala Hospital, Kadri, Mangalore, Karnataka, India, and Dr Ashok Pandith, MCh, Urologist, for their affectionate help and encouragement in all my endeavours. They always stood with me in my difficulties.

I thank very much the Faculty, Department of Surgery and Paediatric Surgery, JJM Medical College (JJMMC), Davangere, Karnataka, India for providing the needed photographs.

I sincerely thank Professor Navinchandra Shetty, Senior Radiologist, Mangalore, Karnataka, India, and also other radiologists for their help in providing and guiding me in X-rays, CT scans and imaging methods.

I acknowledge Professor Kishore Chandra Prasad, Senior ENT Surgeon, for his help, guidance and encouragement in bringing out this book.

I thank Dr Manohar Pai, Dr Ashfaque Mohammed, Dr Kalpana Sridhar, Dr Ramesh, Dr K Akbar, Dr Keshava Prasad, Dr Kishore Reddy, Dr Achaleshwar Dayal, Dr Raghav Pandey, Dr Rupen Shah, Dr Ashwini Mallya, Dr Praveen, Dr Shibu Mohan, Dr Poornachandra, Dr Rahul Bhat, Dr Ranjith Rao, Dr Irshad, Dr Amar, Dr Rajesh, Dr Suyog, Dr Chandini, Dr Adithya, Dr Ashok Hegde, Dr Rajesh Ballal, Dr Devidas Shetty, Dr Venkatesh Sanjeeva, Dr Sunilkumar Shetty, Dr Shanbogh, Dr Harish Nayak, Dr Subraya Kamath, Dr Venkatesh Shanbogh, Dr Tantri, Dr Sandeep Gopal, Dr Ganesh and Dr Sampath, for their help in various aspects.

My sincere thanks to well-known in our state and country, Senior Professor Narayanaswamy Srinivasan, MS, FRCS who earlier worked as—Head, Department of General Surgery, Bangalore Medical College; Director of Sanjay Gandhi Accident Hospital and Research Institute, Byrasandra, Bengaluru; Dean and Director of Bangalore Medical College and Research Institute, Bengaluru; Joint Director of Medical Education, Government of Karnataka, India, for his affectionate encouragement.

I sincerely appreciate Dr Raghavendra Bhat and Dr Ravichandra, Consultants, Department of Radiology, Yenepoya Medical College, Mangalore, Karnataka, India, for their contribution and affectionate help.

I thank Dr Ganesh Pai, MCh (Paediatric Surgeon), Dr Suresh Pai, MCh; Dr Ashok Shetty (Cardiothoracic Surgeon); Dr Muralidhar Pai, MCh (Neurosurgeon); Dr Vivekanand Prabhu (Paediatric Surgeon); Dr Sadashiva Rao (Paediatric Surgeon); Dr Narayana Bhat (Paediatric Surgeon); and Dr Prashanth Mallya (Anaesthetist), for their help and contributions.

I thank my friend Dr Jagadish, MDS, for his contributions to X-ray and opinion on dental and faciomaxillary topics.

My special thanks to Dr Harikiran Chekuri, Hyderabad, for his help in correcting the mistakes.

My wife, Dr Meera Karanth, helped me day and night in revising this edition and without her help this could not have been possible. My beloved daughter, Ananya, helped me in drawing new diagrams artistically. I enjoy her love and affection towards me.

I remember my students, Dr Ravi CR, Dr Ashwini Polnaya, Dr Ishwara Keerthi, Dr Sudesh, Dr Narasimha Murthy, Dr Nithin, Dr Ashish Singh and Dr Aromal, for their special contributions.

I thank all my students especially postgraduates of surgery department who were helping regularly in bringing out this edition.

Words are not sufficient to remember all my patients who are the main material for the book. I pray for their good health always.

I appreciate Shri Jitendar P Vij Sir (Group Chairman), Mr Ankit Vij (Managing Director), Mr Tarun Duneja (Director-Publishing) and all staff of M/s Jaypee Brothers Medical Publishers (P) Ltd, New Delhi, India, for doing appreciable work in their respective field of printing and publishing. Also, my appreciation to working team in the Branches of Jaypee Brothers Medical Publishers at Bengaluru and Mangalore, for their timely help.

My sincere thanks to editing team members—Ms Sunita Katla (Publishing Manager); Mr KK Raman (Production Manager); Ms Seema Dogra (Cover Designer); Ms Geeta Srivastava and Mr Laxmidhar Padhiary (Proofreaders); Mr Kapil Dev Sharma and Mr Kulwant Singh (DTP Operators), and Mr Sumit Kumar and Mr Ram Murti (Graphic Designers) of M/s Jaypee Brothers Medical Publishers (P) Ltd, New Delhi, India, for their dedicated and affectionate work in bringing out the fourth edition of the title.

Contents

1. General Surgery 1

A. Wounds and Wound Healing 1

Wounds 1; Classification of Wounds 1; Wound Healing 5; Compartment Syndrome 8; Crush Injury 9; Crush Syndrome 10; Degloving Injuries 10; Keloid 11; Hypertrophic Scar 12; Problems with Wound Healing 13

B. Ulcer 14

Ulcer 14; Granulation Tissue 17; Investigations for an Ulcer 19; Traumatic Ulcer 20; Trophic Ulcer (Pressure Sore/ Decubitus Ulcer) 21; Ulcer due to Chilblains 23; Ulcer due to Frostbite 23; Martorell's Ulcer 23; Arterial/Ischaemic Ulcer 23; Bairnsdale Ulcer 23; Carcinomatous Ulcer (Epithelioma, Squamous Cell Carcinoma) 24; Marjolin's Ulcer 24; Rodent Ulcer 24; Melanotic Ulcer 24; Diabetic Ulcer 25; Meleney's Ulcer (Postoperative Synergistic Gangrene) 26; Lupus Vulgaris ('Lupus'—Wolf) 26; Tuberculous Ulcer 27; Bazin's Disease (Erythrocyanosis Frigida/ Erythema Induratum) 27; Tropical Ulcer 27; Venous Ulcer (Gravitational Ulcer) 28; Syphilitic Ulcer 28; Soft Chancre/ Soft Sore/ Ducrey's Ulcer/Chancroid/Bubo 28; Climatic Bubo/Tropical Bubo 29; Clinical Examination of an Ulcer 29

C. Sinus and Fistula 30

Sinus 30; Fistula 30; Median Mental Sinus 33; Sequestrum 34; Preauricular Sinus 35

D. Infectious Diseases 36

Surgical Infection 36; Cellulitis 37; Erysipelas 38; Erysipeloid Disease 39; Lymphangitis 39; Abscess 40; Bacteraemia 45; Septicaemia 45; Pyaemia 45; Metastatic and Pyaemic Abscess 46; Boil (Furuncle) 46; Hidradenitis Suppurativa 46; Carbuncle 47; Pott's Puffy Tumour 48; Pyogenic Granuloma (Granuloma Pyogenicum) 49; Impetigo 49; Scrum Pox 49; Tetanus 49; Gas Gangrene 52; Tuberculosis 54; Leprosy 55; Syphilis (Great Pox) (French Disease) 56; Actinomycosis 56; Madura Foot (Mycetoma Pedis) 57; Rabies (Hydrophobia) 58; Rabies in Dogs 60; Anthrax 61; Nosocomial and Opportunistic Infections 61; HIV Infection and AIDS (Acquired Immunodeficiency Syndrome) 62; Necrotising Fasciitis 65; Acute Pyomyositis 67; Surgical Site Infection (SSI) 67

E. Swellings 71

Lipoma 71; Cysts 74; Dermoids 75; Sebaceous Cyst (Wen, Epidermoid Cyst) 79; Glomus Tumour 82; Papilloma 83; Warts 84; Fibroma 84; Bursae 85; Semimembranosus Bursa 86; Morrant Baker's Cyst 87; Lymph Cyst (Lymphatic Cyst) 88; Lymphangioma 88; Calcinosis Cutis 89; Neuroma 89; Neurofibroma 90; Neurilemmoma (Schwannoma) 92; Ganglion 93; Chordoma 93; Epignathus 93

F. Electrolyte and Nutrition 95

Normal Physiology 95; Water Loss (Volume Loss) 95; Water Excess (ECF Volume Excess) 95; ECF Loss 96; ECF Excess 96; Hyponatraemia 96; Hypernatraemia 97; Hypokalaemia 97; Hyperkalaemia 98; Hypermagnesaemia 98; Hypomagnesaemia 98; Acid-Base Balance 98; Metabolic Alkalosis 99; Respiratory Alkalosis 99; Metabolic Acidosis 99; Respiratory Acidosis 100; Anion Gap 100; Fluid Therapy 100; Nutrition 101; Gastrostomy 103; Jejunostomy 104; Total Parenteral Nutrition (TPN) 104; Refeeding Syndrome 106; Obesity and Morbid Obesity 106; Different Surgeries 107

G. Shock 111

Shock 111; Stages of Shock 112; Effects of Shock 112; Clinical Features of Shock 115; Investigations in Shock 115; Central Venous Pressure (CVP) 116; Pulmonary Capillary Wedge Pressure (PCWP) 116; Systemic Inflammatory Response Syndrome (SIRS) 117; Multiple Organ Dysfunction Syndrome (MODS) 117; Oxygen Therapy 118; Topical O_2-Therapy 118; Cardiac Arrest 118

H. Haemorrhage and Blood Transfusion 121

Haemorrhage 121; Blood Transfusion 124; Blood Substitutes 126; Massive Blood Transfusion 126; Autologous Blood Transfusion 127; Artificial Blood 127; Erythropoietin 127; Tourniquets 127; Disseminated Intravascular Coagulation (DIC) 128; Mechanism of Blood Coagulation (Haemostasis) 129

I. Burns 130

Burns 130; Management of Burns 134; Eschar 136; Contracture in Burn Wound 137; Electrical Burns 139; Inhalation Injury 141; Chemical Burns 141

J. Trauma **143**

Triage **143**; Spinal Injury **145**; Neck Injuries **146**; Bullet Injuries **146**; Blast Injuries **147**; Penetrating Injuries **148**; Abdominal Trauma **148**; Blunt Trauma of Abdomen **150**; Duodenal Injury **152**; Pancreatic Injury **153**; Small Bowel Injury **153**; Colonic Injury **154**; Liver Injury **154**; Splenic Injury **154**; Renal Injury **154**; Urinary Bladder Injury **154**; Abdominal Compartment Syndrome **154**; Seatbelt Injuries **155**

K. Hand and Foot **156**

Hand **156**; Hand Infections **157**; Acute Paronychia **160**; Chronic Paronychia **161**; Apical Subungual Infection **161**; Terminal Pulp Space Infection (Felon) **161**; Infection of Web Spaces **162**; Deep Palmar Space Infection **162**; Space of Parona Infection **164**; Acute Suppurative Tenosynovitis **164**; Compound Palmar Ganglion **165**; Orf **166**; Milker's Nodes **166**; Hand Injuries **166**; Dupuytren's Contracture **168**; Volkmann's Ischaemic Contracture **168**; Syndactyly **169**; Mallet Finger (Base Ball Finger) **170**; Heberden's Nodes **170**; Spina Ventosa **170**; Foot **170**; Callosity **170**; Corn **171**; Plantar Fasciitis (Policeman's Heel) **172**; Ingrowing Toe Nail **172**; Onychogryphosis **172**; Onychomycosis **172**; Athlete's Foot **172**; Hallux Valgus **173**

L. Arterial Diseases **174**

Surgical Anatomy of Thoracic Outlet **174**; Arteries of Upper Limb **175**; Arteries of Lower Limb **175**; Arterial Diseases **175**; Intermittent Claudication **175**; Rest Pain **176**; Limb Ischaemia **176**; Pregangrene **177**; Gangrene **177**; Different Levels of Arterial Obstruction **178**; Other Features of Poor Circulation **178**; Investigations for Arterial Diseases **179**; Diseases of the Arteries **182**; Atherosclerosis **182**; Thromboangiitis Obliterans (TAO) Syn. Buerger's Disease **184**; Takayasu's Pulseless Arteritis **188**; Raynaud's Phenomenon **188**; Temporal Arteritis **190**; Treatment of Arterial Diseases **190**; Subclavian Steal Syndrome **196**; Acute Arterial Occlusion **196**; Traumatic Acute Arterial Occlusion **197**; Embolism **197**; Reperfusion Injury **199**; Saddle Embolus **200**; Embolectomy **200**; Fat Embolism **201**; Air Embolism **201**; Therapeutic Embolisation **201**; Caisson's Disease or Decompression Disease **202**; Aneurysm **202**; Mycotic Aneurysm **203**; Abdominal Aneurysm **203**; Abdominal Aortic Aneurysm **204**; Peripheral Aneurysm **208**; Carotid Artery Aneurysm (Extracranial) **209**; Dissecting Aneurysm **209**; Erythromelalgia/Erythralgia **210**; Livedo Reticularis **210**; Polyarteritis Nodosa **210**; Scleroderma/Systemic Sclerosis **211**; Acrocyanosis (Crurum Puellarum Frigidum) **211**; Gangrene **211**; Diabetic Foot and Diabetic Gangrene **213**; Trophic Ulcer **214**; Bedsores (Decubitus Ulcer) (Pressure Sores) **214**; Frostbite **215**; Ainhum **215**; Endovascular Surgeries **215**; Upper Limb Ischaemia **216**; Arterial Substitutes **220**

M. Vascular Lesions and Hamartoma **221**

Vascular Anomalies **221**; Haemangioma **221**; Vascular Malformations **225**; Cirsoid Aneurysm **225**; Arteriovenous Fistula (AVF) **225**; Campbell de Morgan Spots **229**; Parry-Romberg Disease **229**; Hamartomata **229**

N. Venous Diseases **230**

Anatomy of Veins of Lower Limb **230**; Physiology of Venous Blood Flow in Lower Limb **231**; Deep Vein Thrombosis (DVT) **231**; Varicose Veins **235**; Venous Ulcer (Gravitational Ulcer) **249**; Compression Therapy for Varicose Veins **251**; Thrombophlebitis **251**; Klippel-Trenauny Syndrome **252**; Anticoagulants **252**; Heparin **252**; Low Molecular Weight Heparin (LMWH) **252**; Oral Anticoagulants **253**; Warfarin **253**; Direct Thrombin Inhibitors **253**; Antiplatelet Drugs **253**; Pulmonary Embolism **254**

O. Lymphatics **255**

Surgical Anatomy **255**; Lymphangiography **257**; Isotope Lymphoscintigraphy **257**; Lymphoedema **257**; Lymphomas **264**; Hodgkin's Lymphoma (HL) **265**; Non-Hodgkin's Lymphoma (NHL) **269**; Malt Lymphoma (Maltoma) **270**; Burkitt's Lymphoma (Malignant Lymphoma of Africa) **270**; Cutaneous T Cell Lymphoma **270**; Chylous Ascites **271**; Chylothorax **271**; Chyluria **271**; Sarcoidosis **272**

P. Peripheral Nerves **273**

Peripheral Nerve Injuries **273**; Tinel's Sign **275**; Brachial Plexus Injuries **275**; Causalgia **275**; Median Nerve Injury **276**; Carpal Tunnel Syndrome **276**; Ulnar Nerve Injury **278**; Claw Hand **278**; Radial Nerve Injury **279**; Common Peroneal Nerve Injury **280**; Foot Drop **280**; Medial Popliteal Nerve Injury **281**; Axillary Nerve Injury **281**; Long Thoracic Nerve Injury (Nerve of Bell) **281**; Meralgia Paraesthetica **281**

Q. Neoplasm **282**

Definition **282**; Dysplasia **283**; Carcinoma In Situ **284**; Aetiologic Factors **284**; Spread of Malignant Tumours **284**; Grading of Tumour **285**; Staging of the Tumour **285**; Paraneoplastic Syndromes **286**; Investigations for Neoplasm **286**; Management Strategy for Cancers **291**

R. Skin Tumours **293**

Anatomy **293**; Classification of Skin Tumours **293**; Skin Adnexal Tumours **294**; Dermatofibroma (Sclerosing Angioma or Subepithelial Benign Nodular Fibrosis) **296**; Dermatofibrosarcoma Protuberans **296**; Keratoacanthoma (Molluscum Sebaceum) **296**; Rhinophyma (Potato Nose) (Bottle Nose) **297**; Seborrhoeic Keratosis (Seborrhoeic Wart, Basal Cell Papilloma) **298**; Squamous Cell Carcinoma (Epithelioma) (Khangri Cancer) (Chimney Cancer) **299**; Marjolin's Ulcer **302**; Basal Cell Carcinoma (Rodent Ulcer) **303**; Turban Tumour **306**; Naevi **306**; Melanoma **308**

S. Sarcomas 318

Sarcoma 318; Liposarcoma 325; Fibrosarcoma 326; Malignant Fibrous Histiocytoma (MFH) 326; Leiomyosarcoma 327; Rhabdomyosarcoma 327; Chondrosarcoma 327; Haemangiosarcoma 328; Lymphangiosarcoma 328; Synovial Sarcoma 328; Malignant Peripheral Nerve Sheath Tumour (MPNST) 328; Kaposi's Sarcoma 328

T. Amputations 330

Amputation 330; Complications of Amputations 337; Prosthesis 339

U. Reconstruction 340

Graft 340; Skin Grafts 340; Flaps 344; Tendon 351; Tendon Repair 351; Tendon Transfer 352; Tendon Graft 352

V. Transplantation 353

Preoperative Evaluation 353; Organ Procurement 354; Renal Transplantation 354; Immunosuppressive Agents 355; Liver Transplantation 357; Bone Marrow Transplantation 359; Pancreatic Transplantation 359; Small Bowel Transplantation 359; Dialysis 359; Cimino Fistula (Cimino-Brescia) 360

W. Stings and Bites 362

Snake Bite 362; Spider Bite 362; Bee Bite 363; Mammalian Bite 363

X. Pain 364

Gate Control Theory 364

2. Faciomaxillary Diseases 367

Diseases of the Palate 367; Orthopantomogram (OPG) 367; Cleft Lip and Cleft Palate 368; Maxillofacial Injuries 372; Primary Care (Early Care) in Maxillofacial Injuries 374; Fracture Middle Third Area 375; Zygomatic Complex Fracture 376; Fracture of the Mandible 378; Dislocation of the Mandible 380; Preauricular Sinus 380; Jaw Tumours 381; Epulis (Greek—Means Upon Gum) 382; Ameloblastoma (Adamantinoma, Eve's Disease, Multilocular Cystic Disease of the Jaw) 383; Dentigerous Cyst (Follicular Odontome) 384; Dental Cyst (Radicular Cyst, Periapical Cyst) 385; Osteomyelitis of Jaw 385; Alveolar Abscess (Dental Abscess) 386; Fibrous Dysplasia of Bone/Jaw 386; Cherubism (Cherub—Angelic Being) 388

3. Oral Cavity 389

Ranula 389; Sublingual Dermoids 390; Stomatitis 390; Cancrum Oris (Noma) 391; Syphilitic Lesions of Oral Cavity 392; Leukoplakia 392; Erythroplakia 393; Oral Submucosal Fibrosis 393; Premalignant Conditions of Oral Cavity 394; Oral and Upper Aerodigestive Cancers 394; Cheek 396; Carcinoma Cheek/Buccal Mucosa 396; Lip 408; Neoplasm of Lip 410; Carcinoma Lip 410; Tongue 414; Tongue Ulcers 415; Benign Tumours of Tongue 416; Tongue Fissure 416; Glossitis 416; Tongue Tie 417; Carcinoma Tongue 417; Carcinoma of Posterior One-Third/Base of the Tongue 422; Nasopharyngeal Carcinoma 423; Maxillary Tumours 424; Malignant Tumours of Tonsil 426; Carcinoma Hard Palate 427; Laryngeal Tumours 427; Malignant Tumours of Larynx 427; Trismus 430

4. Salivary Glands 431

Anatomy 431; Saliva 434; Sialography 434; Salivary Calculus and Sialadenitis 435; Sialosis 437; Sialectasis 437; Recurrent Childhood Parotitis 438; Parotid Abscess (Suppurative Parotitis) 438; Parotid Fistula 438; Sjögren's Syndrome 439; Mikulicz Disease 440; Salivary Neoplasms 440; Pleomorphic Adenoma (Mixed Salivary Tumour) 441; Adenolymphoma (Warthin's Tumour, Papillary Cystadenolymphomatosum) 444; Oncocytoma (Oxyphil Adenoma) 445; Basal Cell Adenoma 445; Mucoepidermoid Tumour 445; Adenoid Cystic Carcinoma (10% of Salivary Tumours) 446; Acinic Cell Tumour 446; Malignant Mixed Tumour (MMT) 446; Adenocarcinoma of Salivary Glands 446; Squamous Cell Carcinoma of Salivary Glands 447; Submandibular Salivary Gland Tumours 447; Management of Malignant Salivary Tumours 447; Minor Salivary Gland Tumours 450; Parotid Lymphoma 451; Parotidectomy 451; Frey's Syndrome (Auriculotemporal Syndrome, Gustatory Sweating); (Lucie Frey—Polish Surgeon—1932) 452; Facial Nerve Injury (Lower Motor Nerve Lesion, Surgically Related) 453

5. Neck 455

Anatomy of Lymphatics of Head and Neck 455; Thoracic Outlet Syndrome (TOS) 456; Cervical Rib 457; Branchial Cyst 459; Branchial Fistula 460; Pharyngeal Pouch 462; Laryngocele 463; Cystic Hygroma (Cavernous Lymphangioma) 464; Ludwig's Angina 465; Parapharyngeal Abscess 466; Retropharyngeal Abscess 466; Subhyoid Bursitis 467; Carotid Body Tumour (Potato Tumour, Chemodectoma, Nonchromaffin Paraganglioma) 468; Torticollis (Wry Neck) 469; Sternomastoid Tumour 469; Tuberculous Lymphadenitis 470; Cold Abscess 474; Secondaries in Neck Lymph Nodes 475; Chemotherapy for Head and Neck Cancers 482

6. Thyroid 484

Development 484; Anatomy 484; Physiology 486; Congenital Anomalies 486; Thyroid Function Tests 489; FNAC of Thyroid 489; Classification of Goitre 490; Diffuse Hyperplastic Goitre 491; Multinodular Goitre (MNG) 491; Discrete Thyroid Nodule 494; Solitary Thyroid Nodule 494; Retrosternal Goitre 497; Thyrotoxicosis and Hyperthyroidism 499; Radioactive

Iodine 508; Thyroid Neoplasms 509; Differentiated Thyroid Carcinoma (DTC) 509; Papillary Carcinoma of Thyroid (PCT) 510; Follicular Carcinoma of Thyroid (FCT) 512; Anaplastic Carcinoma of Thyroid 515; Medullary Carcinoma of Thyroid (MCT) 518; Malignant Lymphoma 519; Hashimoto's Thyroiditis (Struma Lymphomatosa) 519; De-Quervain's Subacute Granulomatous Thyroiditis 519; Riedel's Thyroiditis (0.5% Common) 520; Thyroid Incidentaloma 520; Thyroidectomy 521; Emil Theodor Kocher 526; Kocher's Test 526; Hypothyroidism 526; Recurrent Laryngeal Nerve Palsy 527

7. Parathyroids and Adrenals 531

Anatomy 531; Calcium 531; Hyperparathyroidism (HPT) 532; Parathyroidectomy 534; MEN Syndrome (MEA Syndrome) 536; Apudomas 536; Hypoparathyroidism 537; Tetany 537; Adrenals 538; Adrenal Cortical Tumours 538; Adrenocortical Carcinoma 539; Cushing's Syndrome 539; Conn's Syndrome 540; Virilising Syndrome or Adrenogenital Syndrome 540; Neuroblastoma 540; Pheochromocytoma 542

8. Breast 544

Anatomy 544; Mammography 547; Aberration of Normal Development and Involution (ANDI) of the Breast 548; Fibroadenoma 548; Fibrocystadenosis (Fibrocystic Disease of the Breast/Mammary Dysplasia/Cyclical Mastalgia with Nodularity) 551; Sclerosing Adenosis 552; Phyllodes Tumour (Cystosarcoma Phyllodes/Serocystic Disease of Brodie) 552; Mastalgia ("Pain in the Breast") 553; Traumatic Fat Necrosis 554; Galactocele 554; Mastitis 555; Antibioma 558; Duct Ectasia 558; Mondor's Disease 558; Tuberculosis of the Breast 559; Breast Cysts 559; Galactorrhoea 560; Gynaecomastia (Greek—Women Breast) 560; Duct Papilloma 562; Zuska-Atkins Disease 563; Mammary Fistula of Atkins 563; Carcinoma Breast 563; Management of Early Carcinoma Breast 586; Advanced Carcinoma Breast 590; Prognostic Factors in Carcinoma Breast 592; Prophylactic Mastectomy 593; Carcinoma of Male Breast 593; Breast Reconstruction 594; Breast Implants 596; Nipple Retraction 597

9. Peritoneum 599

Anatomy 599; Physiology 599; Acute Peritonitis 600; Primary Peritonitis 600; Secondary Peritonitis 600; Tertiary Peritonitis 600; Spontaneous Bacterial Peritonitis (SBP) 605; Sclerosing Peritonitis 606; Biliary Peritonitis 606; Post-operative Peritonitis 606; Other Forms of Peritonitis 607; Pelvic Abscess 607; Subphrenic Spaces and Subphrenic Abscess 608; Mesenteric Cysts 611; Mesenteric Panniculitis 612; Acute Mesenteric Lymphadenitis 612; Mesenteric Malignancy 612; Mesenteric Trauma 613; Peritoneal Malignancy 613; Omental Cyst 614; Omental Torsion 614; Omental Tumour 614

10. Abdominal Tuberculosis 615

Abdominal Tuberculosis 615; Ileocaecal Tuberculosis 616; Ileal Tuberculosis 621; Peritoneal Tuberculosis 621; Tuberculous Mesenteric Lymphadenitis 624; Ano-Recto-Sigmoidal Tuberculosis 626; Tuberculosis of the Omentum 626

11. Liver 627

Surgical Anatomy of Liver 627; Liver Function Tests (LFT) 628; Alfa-Feto Protein (AFP) 628; Liver Biopsy 629; Liver Injury 629; Infections of Liver 631; Liver Tumours 642; Liver Cysts 649; Portal Hypertension 651; Oesophageal Varices 654; Portal Hypertensive Gastropathy 661; Ascites 661; Ascites in Portal Hypertension 663; Budd-Chiari's Syndrome 664; Hepatic Failure 664; Hepatic Encephalopathy 665; Hepatorenal Syndrome 665; Hepatic Resection 665; Portal Biliopathy 668

12. Gallbladder 669

Surgical Anatomy 669; Oral Cholecystogram (OCG; Graham-Cole Test) 670; IV Cholangiogram 671; Endoscopic Retrograde Cholangio-pancreatography (ERCP) 671; Percutaneous Transhepatic Cholangiography (PTC) 672; Magnetic Resonance Cholangio-pancreatography (MRCP) 673; Radioisotope Scan Study 673; Peroperative Cholangiogram 673; Postoperative T-Tube Cholangiogram 673; Congenital Anomalies of Gallbladder 673; Choledochal Cysts 675; Caroli's Disease 677; Biliary Atresia 677; Gallstones 679; Acute Cholecystitis 684; Acute Acalculous Cholecystitis 686; Empyema Gallbladder 687; Mucocele of the Gallbladder 688; Chronic Cholecystitis 688; Cholecystoses 689; Dissolution Therapy for Gallstones 690; Choledocholithiasis 690; Sump Syndrome 694; Courvoisier's Law 694; Surgical Jaundice (Obstructive Jaundice) 695; CBD Strictures (Biliary Strictures) 697; Sclerosing Cholangitis 699; Gallbladder Polyp 699; Benign Biliary Papilloma 699; Carcinoma Gallbladder 699; Cholangiocarcinoma (Bile Duct Carcinoma) 702; Klatskin Tumour 702; Biliary Fistulas 703; Gallstone Ileus 703; Hemobilia 704; White Bile 705; Cholecystectomy 705; Laparoscopic Cholecystectomy 706; Single Incision Laparoscopic Surgery (SILS) in Cholecystectomy 708; Bile Duct Injuries 708; Post-cholecystectomy Syndrome 709; Biliary Dyskinesia 710

13. Spleen 711

Surgical Anatomy 711; Functions of the Spleen 711; Splenunculi 712; Splenic Injury (Rupture Spleen) 712; Splenomegaly 716; Hereditary Spherocytosis 716; Immune Haemolytic Anaemia 717; Thalassaemia (Mediterranean Anaemia/Cooley's Anaemia/Erythroblastic Target Cell Anaemia) 718; Sickle Cell Disease 718; Idiopathic (Immune) Thrombocytopenic Purpura (ITP) **718**; Thrombotic Thrombocytopenic Purpura (TTP) **720**; Splenectomy **720**; Overwhelming Postsplenectomy Infection (OPSI) **722**; Splenic Artery Aneurysm **722**; Splenic Abscess **723**; Hypersplenism **723**; Splenic Cyst **724**

14. Pancreas **725**

Anatomy **725**; Serum Amylase **726**; Magnetic Resonance Cholangio-pancreatography (MRCP) **727**; Pancreatitis **727**; Acute Pancreatitis **727**; Complications of Acute Pancreatitis **732**; Pseudocyst of Pancreas **734**; Chronic Pancreatitis **738**; Pancreatic Tumours **746**; Exocrine Pancreatic Tumours **746**; Carcinoma Pancreas **747**; Endocrine Pancreatic Tumours **755**; Insulinomas **756**; Gastrinomas **757**; Glucagonomas **758**; Zollinger-Ellison Syndrome **758**; Cystic Fibrosis **758**; Annular Pancreas **759**; Ectopic (Accessory) Pancreatic Tissue **759**; Pancreatic Divisum **759**; Pancreatic Calculus **760**; Pancreatic Ascites **760**; Pancreatic Fistulae **760**; Pancreatic Abscess **761**; Pancreatic Necrosis **761**; Pancreatic Trauma **762**; Cystic Lesions of Pancreas **763**; Exocrine Pancreatic Disease **763**

15. Retroperitoneal Space **764**

Anatomy of Retroperitoneum **764**; Retroperitoneal Fibrosis **764**; Retroperitoneal Swellings **765**; Retroperitoneal Tumours **766**; Psoas Abscess **769**

16. Differential Diagnosis of Mass Abdomen **771**

Mass in the Right Hypochondrium **773**; Mass in the Epigastrium **775**; Mass in the Left Hypochondrium **778**; Mass in the Lumbar Region **778**; Mass in the Umbilical Region **780**; Mass in the Right Iliac Fossa **780**; Mass in the Left Iliac Fossa **781**; Mass in the Hypogastrium **781**

17. Abdominal Wall and Umbilicus **784**

Diseases of the Umbilicus **784**; Omphalitis **784**; Umbilical Granuloma **786**; Anomalies of Vitellointestinal Duct **786**; Umbilical Sinus **787**; Umbilical Adenoma (Raspberry Tumour) **787**; Umbilical Fistula **788**; Patent Urachus **788**; Burst Abdomen (Abdominal Dehiscence) (Acute Wound Failure) **788**; Abdominal Wall Tumours **791**; Desmoid Tumour **791**; Exomphalos (Omphalocele) **792**; Gastroschisis (Belly Cleft) **793**; Rectus Sheath Haematoma **794**; Abdominal Wall Abscess **794**; Meleney's Progressive Synergistic Bacterial Gangrene of Abdominal Wall **795**; Diverication of Recti (Diastasis Recti) **796**

18. Hernia **797**

Hernia **797**; Aetiology **797**; Parts of Hernia **798**; Classification of Hernia **799**; Inguinal Hernia **800**; Surgical Anatomy of Inguinal Canal **800**; Classification of Inguinal Hernia (Earlier) **802**; Newer Classification of Hernia **803**; Indirect (Oblique) Inguinal Hernia **803**; Direct Hernia **811**; Malgaigne Bulging **817**; Incarcerated Hernia **817**; Strangulated Hernia **817**; Sliding Hernia (Hernia-En-Glissade) **820**; Pantaloon Hernia (Double Hernia, Saddle Hernia, Romberg Hernia) **821**; Infantile Hernia **821**; Femoral Hernia **822**; Ventral Hernia **824**; Incisional Hernia **825**; Umbilical Hernia **829**; Paraumbilical Hernia (Supra- and Infraumbilical Hernia) **831**; Epigastric Hernia (Fatty Hernia of Linea Alba) **831**; Spigelian Hernia **833**; Obturator Hernia **834**; Richter's Hernia **834**; Lumbar Hernia **835**; Phantom Hernia **835**; Sciatic Hernia **836**; Complications of Hernia Surgery **836**; Parastomal Hernia **836**; Different Types of Hernia **837**

19. Oesophagus **838**

Anatomy **838**; Lower Oesophageal Sphincter (LOS) **839**; Dysphagia **840**; Contrast Study of Oesophagus **841**; Oesophagoscopy **842**; Oesophageal Endosonography **842**; Gastro-oesophageal Reflux Disease (GORD/GERD) **842**; Hiatus Hernia **847**; Rolling Hernia (Paraoesophageal Hernia) **848**; Reflux Oesophagitis **848**; Barrett's Oesophagus (Norman Barrett, British, 1950) **849**; Barrett's Ulcer **849**; Oesophageal Motility Disorders **849**; Achalasia Cardia (Cardiospasm) **850**; Plummer-Vinson Syndrome (Paterson-Kelly Syndrome) **853**; Corrosive Stricture of Oesophagus **854**; Schatzki's Rings **856**; Boerhaave's Syndrome **856**; Mallory-Weiss Syndrome **856**; Tracheo-oesophageal Fistula **857**; Oesophageal Diverticulum **857**; Carcinoma Oesophagus **858**; Benign Tumours of the Oesophagus **866**; Oesophageal Perforation **867**; Foreign Body Oesophagus **867**

20. Stomach **869**

Anatomy **869**; Blood Supply of Stomach **870**; Nerve Supply of Stomach **870**; Histology **870**; Lymphatic Drainage of Stomach **871**; Duodenum **871**; Gastric Physiology **871**; Gastric Function Tests **872**; Gastrin **873**; Barium Meal Study **873**; Gastroscopy **874**; Helicobacter Pylori **875**; Congenital Hypertrophic Pyloric Stenosis **876**; Gastritis **877**; Acute Peptic Ulcer (Duodenal or Gastric Ulcer) **878**; Gastric Ulcer **879**; Duodenal Ulcer **882**; Pyloric Stenosis due to Chronic Duodenal Ulcer **885**; Perforated Peptic Ulcer **887**; Bleeding Peptic Ulcer **891**; Haematemesis **893**; Complications of Gastric Surgery **894**; Trichobezoar (Rapunzel Syndrome) **898**; Chronic Duodenal Ileus (Wilkie's Syndrome) **898**; Durban's Syndrome **899**; Acute Gastric Dilatation **900**; Gastric Volvulus **900**; Gastric Polyp **901**; Menetrier's Disease **901**; Duodenal Diverticula **902**; Carcinoma Stomach **902**; Gastric Lymphoma **915**; Gastric Sarcomas **916**; Gastrointestinal Stromal Tumours (Gist) **917**; Pyloroplasty **918**; Gastrostomy **918**; Gastrectomy **919**; Gastrojejunostomy (GJ) **920**; Retrograde Jejunogastric Intussusception **920**; Vagotomy **920**

21. Small Intestine 922

Anatomy **922**; Meckel's Diverticulum **923**; Regional Enteritis (Crohn's Disease) **925**; Surgical Complications of Typhoid **928**; Surgical Complications of Roundworm (Ascaris Lumbricoides) **929**; Pneumatosis Cystoides Intestinalis **931**; Mesenteric Vessel Ischaemia **931**; Necrotising Enterocolitis **934**; Small Bowel Tumours **935**; Benign Tumours of Small Bowel **936**; Malignant Tumours of Small Bowel **937**; Carcinoid Tumour **939**; Short Bowel Syndrome (Short Gut Syndrome) **941**; Massive Bowel Resection **941**; Small Bowel Enema (Enteroclysis) **942**; Capsule Endoscopy **942**; Small Bowel Enteroscopy **943**; Enteric/Gastrointestinal Fistula **943**

22. Large Intestine 948

Anatomy **948**; Hirschsprung's Disease (Congenital Megacolon) **949**; Diverticular Disease of the Colon **951**; Ulcerative Colitis **954**; Ischaemic Colitis **959**; Pseudomembranous Colitis **960**; Surgical Complications of Intestinal Amoebiasis **960**; Tumours of Colon **961**; Benign Tumours/Polyp of the Colon **961**; Juvenile Polyps **961**; Metaplastic/Hyperplastic Polyp **961**; Peutz-Jegher's Polyp **962**; Adenoma of Colon **962**; Familial Adenomatous Polyp (FAP) **963**; Gardner's Syndrome **964**; Carcinoma Colon **964**; Angiodysplasia of Colon **974**; Ogilvie's Syndrome **974**; Colostomy **975**; Stoma Care **977**; Stoma Appliances **978**; Faecal Fistula **979**; Preparation of Large Bowel for Surgery **980**; Surgical Pouches **980**; Barium Enema **981**

23. Intestinal Obstruction 983

Intestinal Obstruction: Types **983**; Dynamic Obstruction **984**; Duodenal Atresia **989**; Small Intestine Atresia (Intestinal Atresia) **990**; Malrotation **991**; Meconium Ileus **992**; Intussusception (ISS) **994**; Volvulus **997**; Sigmoid Volvulus (Volvulus of Pelvic Colon) **997**; Paralytic Ileus (Adynamic Intestinal Obstruction) **1000**; Adhesions and Bands **1000**; Internal Hernias **1003**

24. Appendix 1005

Surgical Anatomy **1005**; Duplication of Appendix (Wallbridge Classification) **1005**; Acute Appendicitis **1006**; Incidental Appendicectomy **1013**; Appendicular Mass (Periappendicular Phlegmon) **1014**; Appendicular Abscess **1015**; Faecal Fistula After Appendicectomy **1016**; Mucocele of Appendix **1016**; Neoplasms of the Appendix **1017**; Laparoscopic Appendicectomy **1017**

25. Rectum and Anal Canal 1020

Anatomy **1020**; Per-rectal Examination (Digital Examination of the Rectum) **1021**; Proctoscopy (Kelly's) **1022**; Sigmoidoscopy **1022**; Colonoscopy **1022**; Carcinoma Rectum **1023**; Solitary Ulcer Syndrome **1029**; Rectal Prolapse **1030**; Anorectal Malformations (ARM) **1034**; Pilonidal Sinus/Disease (Jeep Bottom; Driver's Bottom) **1036**; Piles/Haemorrhoids **1038**; Anal Fissure (Fissure-In-Ano) **1046**; Anorectal Abscess **1049**; Fistula-In-Ano **1051**; Anorectal Strictures **1056**; Condyloma Acuminata **1057**; Anal Intraepithelial Neoplasia (AIN) **1057**; Malignant Tumours of Anal Area **1057**; Anal Margin Tumours **1058**; Sacrococcygeal Teratoma **1058**; Anal Incontinence **1059**; Descending Perineal Syndrome **1059**; Proctitis **1060**; Proctalgia Fugax **1060**; Hidradenitis Suppurativa of Anal Region **1060**; Pruritus Ani **1060**; Gastrointestinal Haemorrhage (GI Bleed) **1061**

26. Urology 1065

A. Kidney 1065

Anatomy of Kidney and Ureter **1065**; Kidney—Anatomy **1065**; Ureter—Anatomy **1066**; Plain X-ray—Kidney, Ureter and Bladder (KUB) **1066**; Intravenous Urogram (IVU) **1067**; Retrograde Pyelography (RGP) **1068**; Renal Angiogram **1069**; Micturating Cystourethrography (MCU) **1069**; Ascending Urethrogram **1069**; Isotope Renography **1070**; Cystoscopy **1071**; Catheters **1071**; Foley's Catheter (Fredrick Eugene Basil Foley—American Urologist) **1072**; Malecot's Catheter **1073**; Red Rubber Catheter **1073**; Nephrostomy **1074**; Suprapubic Cystostomy (SPC) **1074**; Haematuria **1075**; Horseshoe Kidney **1076**; Cystic Diseases of the Kidney **1076**; Polycystic Kidney Disease (PCKD) **1077**; Duplication of Renal Pelvis and Ureter **1078**; Retrocaval Ureter **1079**; Ureterocele **1080**; Injuries to Kidney **1080**; Renal Tuberculosis **1081**; Hydronephrosis (HN) **1084**; Pyonephrosis **1088**; Carbuncle of Kidney (Renal Carbuncle) **1088**; Perinephric Abscess **1089**; Renal Calculus **1089**; Ureteric Calculi **1094**; Staghorn Calculus **1097**; Benign Tumours of Kidney **1098**; Wilm's Tumour (Nephroblastoma) **1099**; Renal Cell Carcinoma (RCC) **1100**; Approaches to Kidney (Surgical) **1104**

B. Urinary Bladder 1105

Anatomy **1105**; Ectopia Vesicae (Extrophy of the Bladder) **1106**; Urachal Anomalies **1106**; Vesical Calculus **1106**; Cystitis **1110**; Recurrent Cystitis **1110**; Interstitial Cystitis (Hunner's Ulcer, Elusive Ulcer) **1110**; Schistosoma Haematobium (Endemic Haematuria, Urinary Bilharziasis) (Swimmer's Itch) **1111**; Thimble or Systolic Bladder **1111**; Bladder Tumours **1112**; Transitional Cell Carcinoma (TCC) **1112**; Ureterosigmoidostomy **1115**; Rupture Bladder (Bladder Injury) **1115**; Residual Urine **1116**; Malakoplakia **1116**; Neurogenic Bladder **1117**; Vesicoureteric Reflux **1117**; Bladder Diverticula **1118**; Urinary Diversion **1119**; Urinary Fistulas **1120**

C. Prostate 1121

Anatomy **1121**; Acid Phosphatase **1121**; Prostate Specific Antigen (PSA) **1121**; Benign Prostatic Hyperplasia (BPH) **1122**; Prostatitis **1125**; Bladder Outlet Obstruction (BOO) **1125**; Carcinoma Prostate **1126**

D. Urethra 1129

Anatomy **1129**; Urethral Injury **1129**; Rupture of Membranous Urethra and/or Prostatic Urethra (Posterior Urethra) **1129**; Rupture of Bulbous Urethra (Anterior Urethra) **1130**; Stricture Urethra **1131**; Hypospadias **1132**; Epispadias **1133**; Posterior Urethral Valve **1133**; Urethral Calculi **1134**; Urethritis **1135**; Extravasation of Urine **1135**; Retention of Urine **1136**

E. Penis 1138

Phimosis **1138**; Paraphimosis **1139**; Circumcision **1140**; Balanoposthitis **1141**; Chordee (Cordee) **1141**; Priapism **1141**; Peyronie's Disease (Induratio-Penis Plastica) **1141**; Ram's Horn Penis **1142**; Carcinoma Penis **1142**; Buschke-Lowenstein's Tumour **1147**

F. Scrotum 1148

Anatomy **1148**; Fournier's Gangrene **1148**; Hydrocele **1149**; Primary Vaginal Hydrocele **1150**; Secondary Hydrocele **1151**; Haematocele **1154**; Pyocele **1155**; Cyst of Epididymis **1155**; Spermatocele **1156**; Varicocele **1157**

G. Testis 1160

Anatomy **1160**; Undescended Testis **1161**; Ectopic Testis **1163**; Retractile Testis **1164**; Torsion of the Testis **1164**; Testicular Tumours **1165**; Paratesticular Tumours **1170**; Orchitis **1170**; Epididymitis **1171**

27. Neurosurgery 1172

Head Injuries **1172**; Extradural Haematoma **1177**; Subdural Haematoma **1178**; Subarachnoid Haemorrhage (SAH) **1179**; Depressed Skull Fracture **1180**; Hydrocephalus **1181**; Intracranial Abscess **1182**; Intracranial Aneurysms **1182**; Intracranial Tumours **1183**; Pituitary Tumours **1186**; Craniopharyngiomas **1187**; Spinal Dysraphism **1187**; Meningocele **1188**; Spina Bifida **1188**; Intervertebral Disc Prolapse (IVDP) **1189**; Tuberculosis of Spine (Caries Spine) **1190**; Spinal Tumours **1193**

28. Thorax 1194

Chest Injuries **1194**; Fracture Ribs **1195**; Flail Chest and Stove in Chest **1196**; Pneumothorax **1197**; Tension Pneumothorax **1197**; Haemothorax **1198**; Pleural Tap **1198**; Bronchoscopy **1199**; Empyema Thoracis **1199**; Empyema Necessitans **1200**; Lung Abscess **1201**; Intercostal Tube Drainage **1202**; Shock Lung (Stiff Lung) **1203**; Pulmonary Embolism (PE) **1203**; Surgical Emphysema **1204**; Lung Cysts **1204**; Mediastinal Tumours **1205**; Thymomas **1207**; Pancoast Tumours (Superior Sulcus Tumour) **1208**; Chest Wall Tumours **1208**; Pericarditis **1209**; Acute Pericarditis **1209**; Chronic Constrictive Pericarditis (Pick's Disease) **1210**; Pericardial Tap **1210**; Cardiac Tamponade **1210**; Diaphragmatic Hernia **1211**; Pulmonary Complications during Postoperative Period **1213**; Surgical Management of Pulmonary Tuberculosis **1214**; Video Assisted Thoracoscopic Surgery (VATS) **1214**

Cardiac Surgery 1215

Anatomy **1215**; Preoperative Assessment and Preparation of the Cardiac Patient **1216**; Cardiopulmonary Bypass **1217**; Congenital Heart Diseases **1217**; Patent Ductus Arteriosus (PDA)—10% **1217**; Coarctation of Aorta—5% **1218**; Atrial Septal Defect (ASD)—7% **1218**; Ventricular Septal Defect (VSD)—15% **1219**; Pulmonary Stenosis **1219**; Transposition of Great Vessels **1219**; Tetralogy of Fallot **1220**; Acquired Heart Disease **1220**; Mitral Regurgitation **1221**; Aortic Stenosis **1221**; Aortic Regurgitation **1221**; Valve Replacement Surgery **1222**; Ischaemic Heart Disease (IHD) **1222**; Cardiac Pacemakers **1223**; Postoperative Care **1223**

29. Adjuvant Therapy 1224

Radiotherapy **1224**; Chemotherapy **1227**; Cell Cycle **1227**; Antimalignancy Drugs **1227**; Hormone Therapy in Cancer **1227**; Immunosuppression **1229**; Immunotherapy **1229**; Hybridoma **1229**; Gene Therapy **1230**

30. Anaesthesia 1231

Preoperative Assessment **1231**; General Anaesthesia **1232**; Anaesthetic Agents **1232**; Oxygen **1232**; Muscle Relaxants **1232**; Reversal Agents **1232**; Instruments in Anaesthesia **1232**; Complications of General Anaesthesia **1234**; Postoperative Care **1234**; Monitoring the Postoperative Patient **1234**; Regional Anaesthesia **1235**; Topical Anaesthesia **1235**; Infiltration Block **1235**; Field Block **1235**; Nerve Block **1235**; Intravenous Regional Anaesthesia (Bier's Block) **1235**; Spinal Anaesthesia **1235**; Saddle Block **1236**; Epidural Anaesthesia **1236**; Caudal Anaesthesia **1237**

31. Advanced Imaging Methods 1238

Ultrasound **1238**; Doppler **1240**; CT Scan **1240**; Magnetic Resonance Imaging (MRI) **1242**; Radionuclide Imaging **1243**; Positron-Emission Tomography (PET Scan) **1243**

32. Operative Surgery 1244

A. Asepsis and Sterilisation **1244**

Sterilisation **1244**; Disinfection **1244**; Antisepsis **1244**; Asepsis **1244**; Different Methods of Disinfection/Sterilisation **1244**

B. Instruments **1247**

Cheatle's Forceps **1247**; Sponge Holding Forceps (Rampley's) **1247**; Mayo's Towel Clip **1247**; Artery Forceps (Haemostat) **1247**; Right Angle Forceps **1248**; Kocher's Forceps **1248**; Allis' Tissue Holding Forceps **1248**; Babcock's Forceps **1248**; Lane's Tissue Holding Forceps **1249**; Morant-Baker's Appendix Holding Forceps **1249**; Volkmann's Retractor **1249**; Langenbeck's Retractor **1249**; Czerny's Retractor (Hernia Retractor) **1249**; Morris' Retractor **1249**; Deaver's Retractor **1249**; Doyen's Retractor **1250**; Self-retaining Retractor **1250**; Single Hook Retractor **1250**; Plain Non-toothed Dissecting Forceps **1250**; Toothed Dissecting Forceps **1250**; Surgical Needles **1250**; Needle Holder **1251**; Joll's Thyroid Retractor **1251**; Moynihan's Occlusion Clamp **1251**; Payr's Crushing Clamp (Gastric) **1251**; Desjardin's Choledocholithotomy Forceps **1251**; Bake's Dilator **1251**; Sinus Forceps (Lister's) **1252**; Scissors **1252**; Volkmann's Scoop **1252**; Tracheostomy Tube **1252**; Drains **1252**; Foley's Catheter **1254**; Malecot's Catheter **1254**; Simple Red Rubber Catheter **1254**; Lister's Urethral Dilator **1254**; Ryle's Tube **1254**; Infant Feeding Tube **1255**; Kehr's 'T' Tube **1255**; Proctoscope **1255**; Flatus Tube **1255**

C. Suture Materials **1256**

Classification I **1256**; Classification II **1257**; Classification III **1257**; Classification IV **1257**; Classification V **1257**

D. Diathermy (Electrocautery) **1258**

Types **1258**

E. Operative Procedure **1259**

Abdominal Incisions **1259**; Vasectomy **1260**; Circumcision **1260**; Hydrocele **1261**; Inguinal Hernia **1261**; Appendicectomy **1262**; Thyroidectomy **1262**; Tracheostomy **1263**; Cryosurgery **1264**; Lasers in Surgery (Light Amplification Stimulated Emission of Radiation) **1265**; Staplers in Surgery **1265**; Nasojejunal Tube Feeding **1266**; Gossypiboma (Gossypium—Cotton Based in Latin) **1266**

F. Laparoscopic Surgery **1267**

Advantages of Laparoscopic Surgery **1267**; Laparoscopic Cholecystectomy **1268**; Laparoscopic Appendicectomy **1269**; Advanced Laparoscopic Surgeries **1269**; Diagnostic Laparoscopy **1270**; Retroperitoneoscopy **1270**; Natural Orifice Transluminal Endoscopic Surgery (NOTES) **1270**

G. Dressings and Bandages **1272**

Dressings **1272**; Bandages **1272**

H. Day Care Surgery **1274**

Day Care Surgery **1274**; Surgical Audit **1275**; Surgeon and Law **1276**

33. Miscellaneous 1277

A. Fascinating Signs in Surgery **1277**

B. Triads in Surgery **1287**

C. Misnomers in Surgery **1288**

D. Triangles in Surgery **1289**

E. Drugs at a Glance **1291**

Antibacterials **1291**; Sulfonamides **1291**; Quinolones **1291**; Penicillin **1292**; Ampicillin/Amoxycillin/Talampicillin/Pivampicillin **1292**; Methicillin **1292**; Cloxacillin/Dicloxacillin/Flucloxacillin **1292**; Carbenicillin/Ticarcillin **1292**; Piperacillin/Azocillin/Mezlocillin **1292**; Clavulanic Acid **1292**; Sulbactum **1292**; Tazobactum **1292**; Carbapenem/Imipenem/Meropenem **1292**; Lincomycin/Clindamycin **1292**; Vancomycin/Teicoplanin **1293**; Cephalosporins **1293**; Macrolides **1293**; Aminoglycosides **1293**; Tetracyclines **1294**; Chloramphenicol / Thiamphenicol **1294**; Drugs For Tuberculosis **1294**; Antiamoebic Drugs **1295**; Important Antihelminthics **1295**; Anticoagulants **1296**; Other Drugs **1296**

Appendix *1299*

Further Reading *1303*

Index *1305*

Chapter 1 General Surgery

A. Wounds and Wound Healing

> *One having a wound in his eyebrow. An ailment which I will treat. Treatment [of a wound in the eyebrow]: Now after thou hast stitched it, [thou shouldst bind] fresh meat upon [it] the first day. If thou findest that the stitching of this wound is loose, thou shouldst draw it together for him with two strips (of plaster), and thou shouldst treat it with grease and honey every day until he recovers.*
>
> **—[Anonymous],** circa 2500 BC

CHAPTER OUTLINE

- Wounds
- Classification of Wounds
- Wound Healing
- Compartment Syndrome
- Crush Injury
- Crush Syndrome
- Degloving Injuries
- Keloid
- Hypertrophic Scar
- Problems with Wound Healing

WOUNDS

Wound Definition

A wound is a break in the integrity of the skin or tissues often, which may be associated with disruption of the *structure and function.*

Wound is simply a disruption of any tissues—soft tissue or bone or internal organs. Ulcer is disruption or break in the continuity of any lining—may be skin, mucous membrane or others. Ulcer is one of the types of wounds.

CLASSIFICATION OF WOUNDS

I. Rank and Wakefield Classification

a. **Tidy wounds**
- They are wounds like surgical incisions and wounds caused by sharp objects.
- It is incised, clean, healthy wound without any tissue loss.
- Usually primary suturing is done. Healing is by primary intention.

b. **Untidy wounds**
- They are due to:
  - Crushing.
  - Tearing.
  - Avulsion.
  - Devitalised injury.
  - Vascular injury.
  - Multiple irregular wounds.
  - Burns.
- Fracture of the underlying bone may be present.

Wound dehiscence, infection, delayed healing are common.
- Liberal excision of devitalised tissue and allowing to heal by secondary intention is the management.
- Secondary suturing, skin graft or flap may be needed.

II. Classification based on Type of Wound

a. ***Clean incised wound*** is a clean cut wound with linear edge.

b. ***Lacerated wounds*** have ragged edges with devitalisation of some part of tissues. Wound excision and primary suturing is done.

c. ***Bruising and contusion***: Minor soft tissue injury with discoloration and haematoma formation without skin break.

I dressed him and God healed him.— **Ambroise Pares**

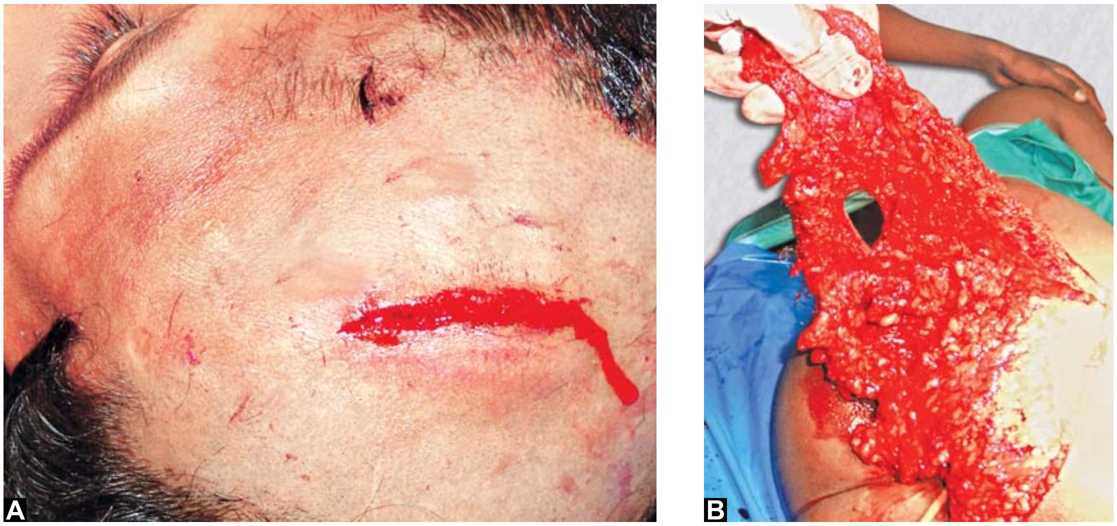

Figs 1.1A and B: A. Tidy wound; B. Untidy wound.

d. ***Haematoma***: It may be subcutaneous/intramuscular/subfascial/intra-articular. Small haematoma will get absorbed. Large haematoma once get infected forms an abscess and so it should be drained under general/regional anaesthesia adequately. Often haematoma contains only reddish plasmatic fluid which can be aspirated with wide bore needle.

e. Closed blunt injury.

f. Puncture wounds and bites.

g. ***Abrasion***: It is superficial and is due to shearing of skin where the surface is rubbed off. It heals by epithelialisation. It is only epidermal injury exposing dermis and dermal nerves.

h. Traction and avulsion injury.

i. ***Crush injury:*** It is caused by war wounds, road traffic accidents, tourniquet. It leads to:
   - Compartment syndrome.
   - Muscle ischaemia.
   - Gangrene, loss of tissue.

j. War wounds and gunshot injuries.

k. Injuries to ***bones and joints, may be open or closed.***

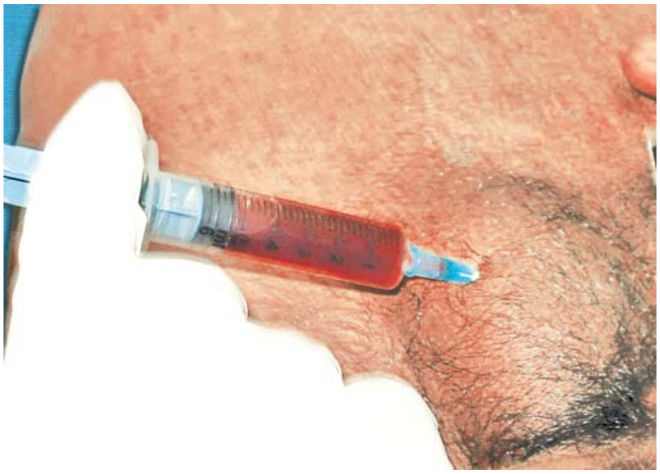

Fig. 1.3: Haematoma in the groin being aspirated. Haematoma may get infected or can compress adjacent structures.

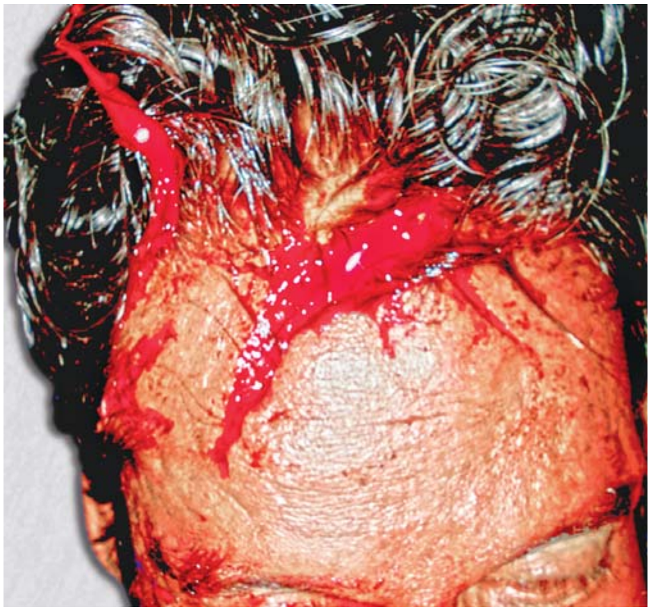

Fig. 1.2: Lacerated wound in the scalp.

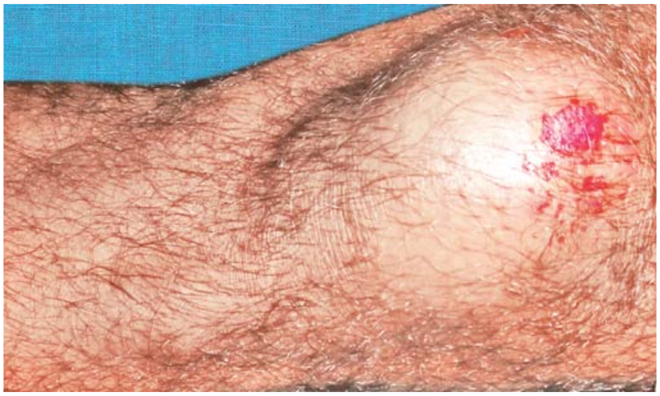

Fig. 1.4: Traumatic haematoma in the leg with abrasion over its surface.

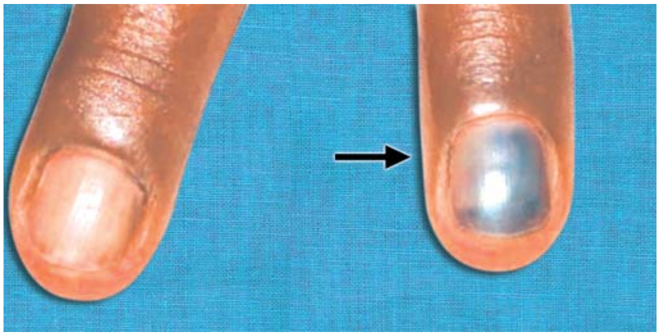

Fig. 1.5: Subungual haematoma. Often nail may need to be removed to evacuate the blood clot.

l. Injuries to ***nerves, either clean cut or crush.***
m. Injuries to ***arteries and veins*** (major vessels).
n. Injury to internal organs, may be of penetrating or nonpenetrating (blunt) types.
o. ***Penetrating wounds***—commonly due to stab injuries. Common example is stab injury abdomen. Stab wound may be small on body surface but damage to the deeper organs like liver, spleen, major vessels like inferior vena cava/mesenteric vessels or intestines may be extensive and life-threatening. Patients with stab injury over abdomen should be admitted and evaluated by ultrasound or CT scan abdomen and observed. Often wound exploration is needed under general anaesthesia in operation theatre.

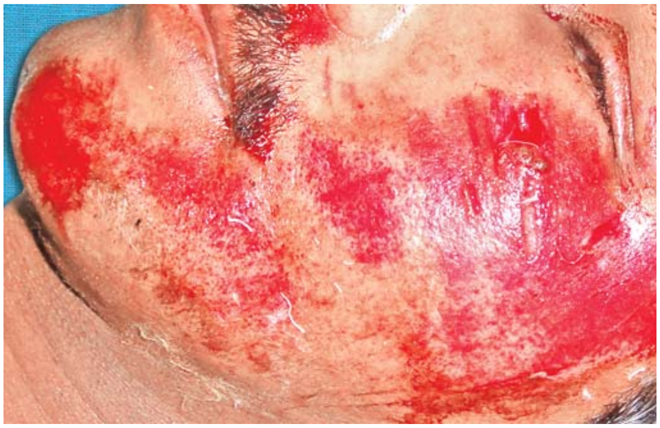

Fig. 1.6: Typical look of abrasion in face.

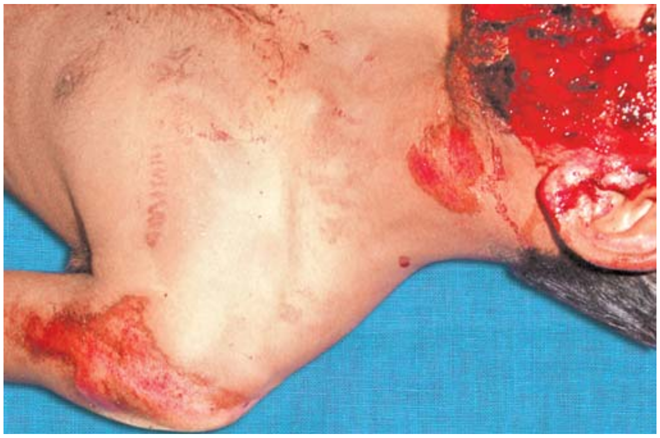

Fig. 1.7: Multiple abrasions

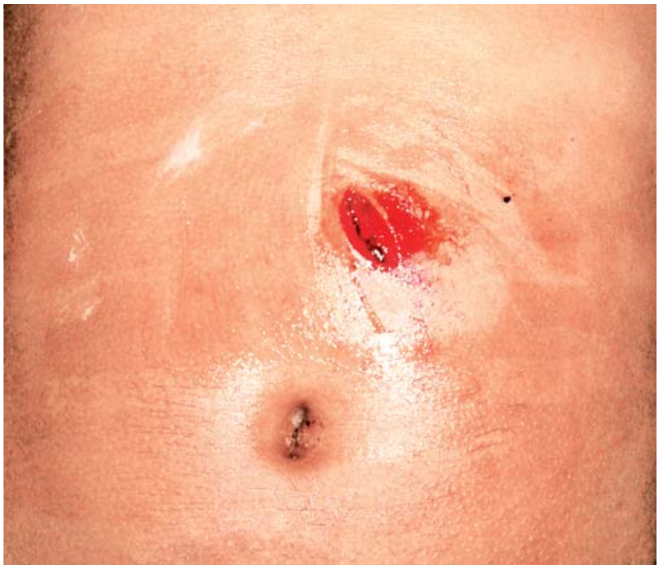

Fig. 1.8: Penetrating wound of the abdomen.

III. Classification based on Thickness of the Wound

- ***Superficial wound*** involving only epidermis and dermal papillae.
- ***Partial thickness*** wound with skin loss up to deep dermis with only deepest part of the dermis, hair follicle shafts and sweat glands are left behind.
- ***Full thickness*** wound with loss of entire skin and subcutaneous tissue causing spacing out of the skin edges.
- ***Deep wounds*** are the one extending deeper, across deep fascia into muscles or deeper structures.
- ***Complicated wounds*** are one associated with injury to vessels or nerves.
- ***Penetrating wound*** is one which penetrates into either natural cavities or organs.

IV. Classification based on Involvement of Structures

- ***Simple wounds*** are one involving only one organ or tissue.
- ***Combined wounds*** are one involving mixed tissues.

V. Classification based on the Time Elapsed

- ***Acute wound*** is up to 8 hours of trauma.
- ***Chronic wound*** is after 8 hours of trauma.

VI. Classification of Surgical Wounds

a. Clean wound
  - Herniorrhaphy.
  - Excisions.
  - Surgeries of the brain, joints, heart, transplant.
  - Infective rate is less than *2%.*

b. Clean contaminated wound
  - Appendicectomy.
  - Bowel surgeries.
  - Gallbladder, biliary and pancreatic surgeries.
  - Infective rate is 10%.

c. Contaminated wound
  - Acute abdominal conditions.
  - Open fresh accidental wounds.
  - Infective rate is 15-30%.

A great part, I believe, of the art of medicine is the ability to observe.—**Hippocrates, Father of Medicine**

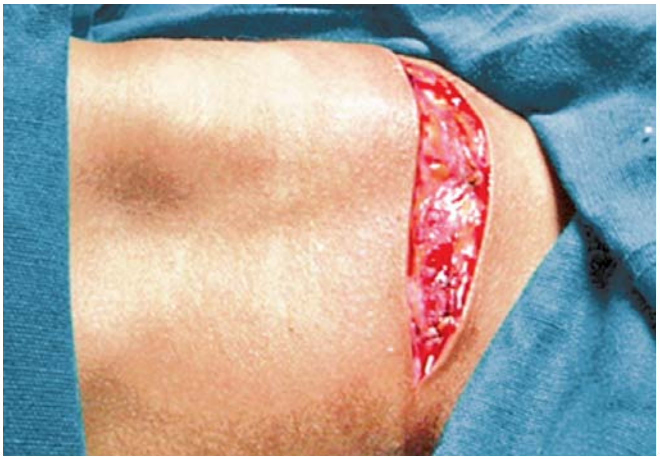

Fig. 1.9: Clean wound of thyroidectomy surgery.

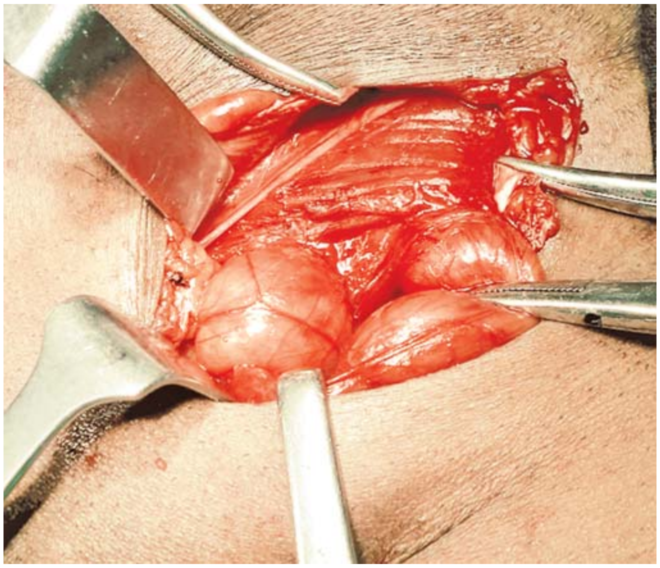

Fig. 1.10: Hernia surgery with hernia sac. It is a clean wound.

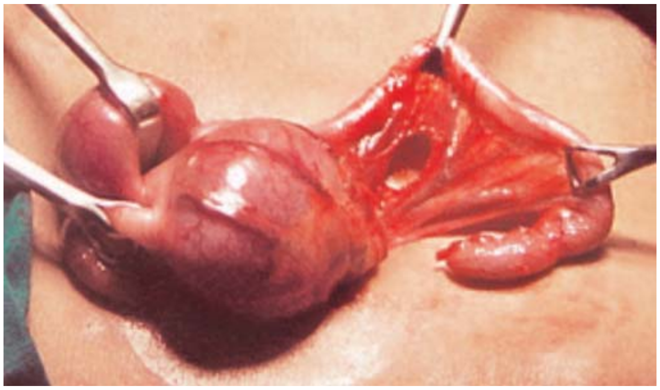

Fig. 1.11: Appendicectomy wound, which is a clean contaminated wound.

d. Dirty infected wound

- Abscess drainage.
- Pyocele.
- Empyema gallbladder.
- Faecal peritonitis.
- Infective rate is 40-70%.

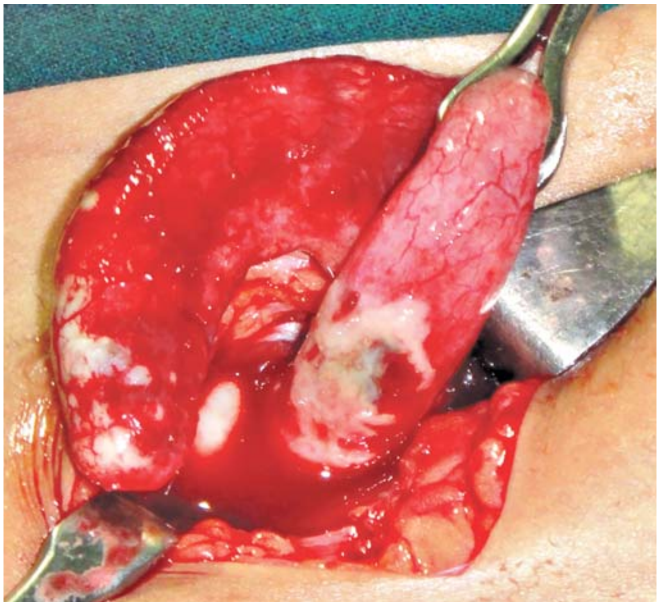

Fig. 1.12: Contaminated wound—burst appendicitis.

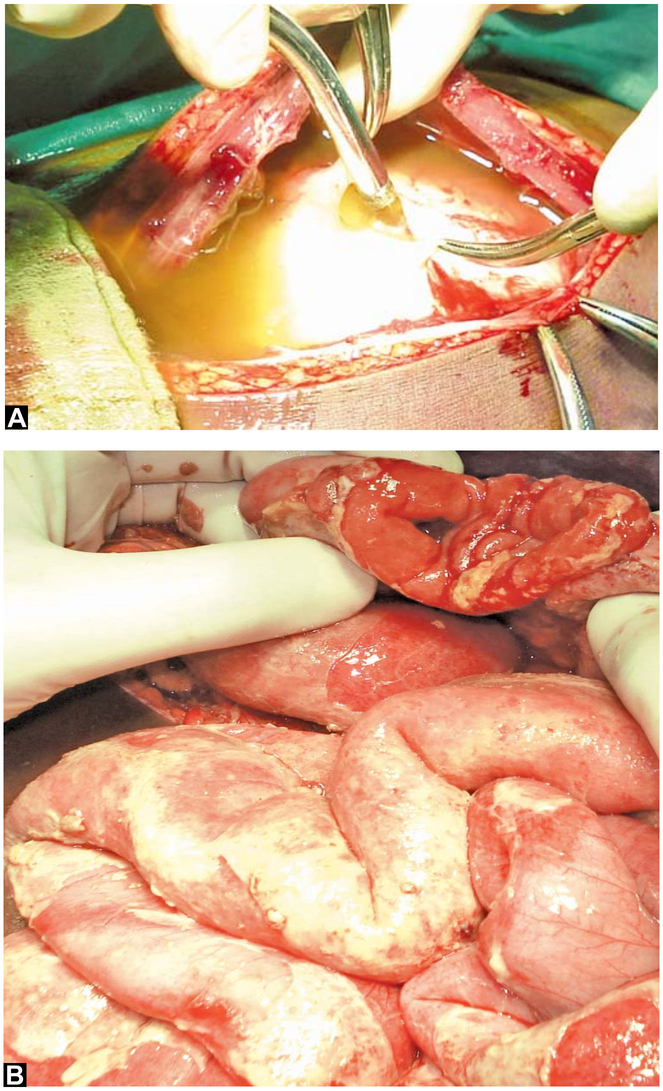

Figs 1.13A and B: Acute peritonitis with frank pus in peritoneal cavity due to bowel perforation is a dirty wound.

Wound classifications	
- *Simple wounds:* Only skin is involved.	- Tidy wounds
- *Complex wounds:* Vessels, nerves, tendons or bones are involved.	- Untidy wounds
- *Closed wounds:*	- Clean wound
– Contusion	- Clean contaminated wound
– Abrasion	- Contaminated wound
– Haematoma	- Dirty wound
- *Open wounds:*	
– Incised wounds	
– Lacerated wounds	
– Crush injuries	
– Penetrating wounds	

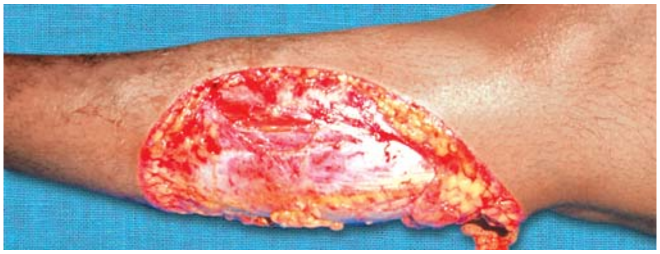

Fig. 1.14: Degloving injury in the leg.

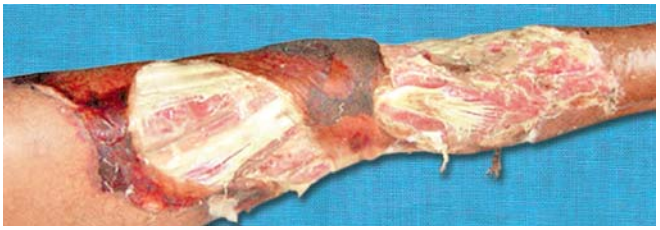

Fig. 1.15: Wound causing extensive skin loss and necrosis.

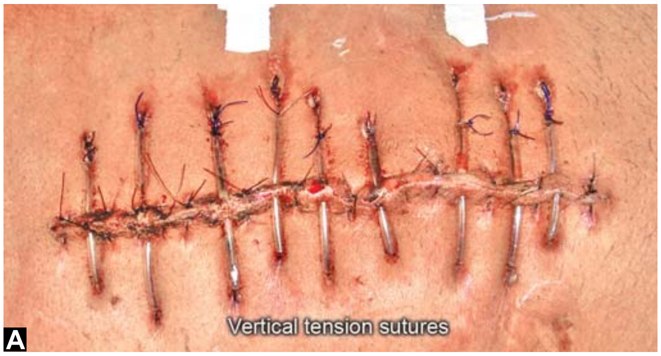

Fig. 1.16A

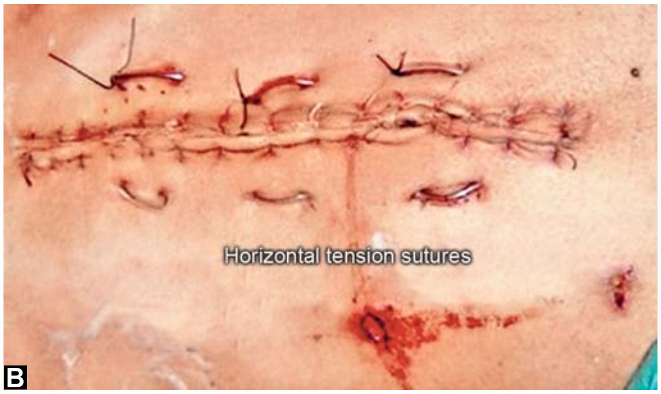

Fig. 1.16B

Figs 1.16A and B: Wound with tension sutures after surgeries for acute abdomen/peritonitis. It is a contaminated wound. Both vertical and horizontal tension sutures are shown.

WOUND HEALING

> *A scab is a beautiful thing—a coin the body has minted, with an invisible motto: In God We Trust. Our body loves us, and, even while the spirit drifts dreaming, works at mending the damage that we do.*
>
> **—John Updike,** 1984

Wound healing is complex method to achieve anatomical and functional integrity of disrupted tissue by various components like neutrophils, macrophages, lymphocytes, fibroblasts, collagen; in an organised staged pathways—haemostasis → inflammation → proliferation → matrix synthesis (collagen and proteoglycan ground substance) → maturation → remodelling → epithelialisation → wound contraction (by myofibroblasts).

Types of Wound Healing

Primary Healing (First Intention)

- It occurs in a clean incised wound or surgical wound. Wound edges are approximated with sutures. There is more epithelial regeneration than fibrosis. Wound heals rapidly with complete closure. Scar will be linear, smooth, and supple.

Secondary Healing (Second Intention)

- It occurs in a wound with extensive soft tissue loss like in major trauma, burns and wound with sepsis. It heals slowly with fibrosis. It leads into a wide scar, often hypertrophied and contracted. It may lead into disability.
- Re-epithelialisation occurs from remaining dermal elements or wound margins.

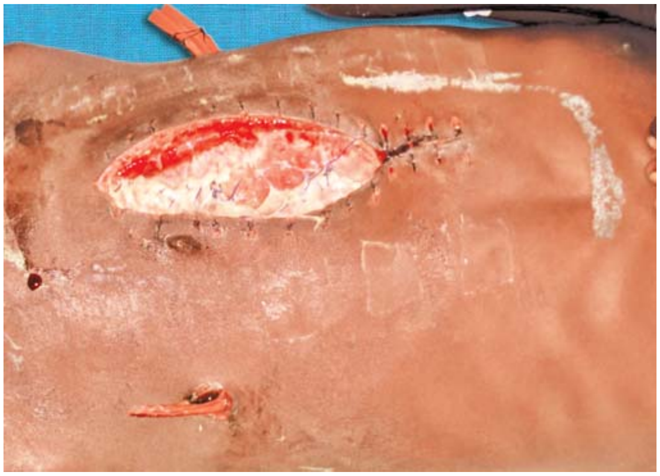

Fig. 1.17: Wound in the abdomen healing with second intention which requires secondary suturing once it granulates well. Secondary suturing is done after 10-14 days, once wound granulates well with proper control of infection. Scar in such type is prone to form incisional hernia.

Healing by Third Intention (Tertiary Wound Healing or Delayed Primary Closure)

After wound debridement and control of local infection, wound is closed with sutures or covered using skin graft. Primary contaminated or mixed tissue wounds heal by tertiary intention.

Stages of Wound Healing

- Stage of inflammation.
- Stage of granulation tissue formation and organisation. Here due to fibroblastic activity synthesisation of collagen and ground substance occurs.
- Stage of epithelialisation.
- Stage of scar formation and resorption.
- Stage of maturation.

Phases of Wound Healing

Inflammatory Phase (Lag or Substrate or Exudative Phase)

- It begins immediately after wound healing. It lasts for 4-6 days.
- Features of inflammation are rubor, calor, tumour, dolor and loss of function.
- Macrophages secrete fibroblastic growth factor which enhances angiogenesis.
- Polymorphonuclear leukocytes (PMN leukocytes) appear after 48 hours which secrete inflammatory mediators and bactericidal oxygen derived free radicals.
- These cells also remove clots, foreign bodies and bacteria.
- Chemical factors involved in wound healing are:
    - Growth factor—platelet derived, epidermal, transforming.
    - Interleukin.
    - Tumour necrosis factor.
    - Prostaglandins.
    - Collagenase.
    - Elastase.

Here *haemostasis, coagulation and* chemotaxis occurs. Coagulation begins in wound haematoma → formation of platelet fibrin thrombus → release of cytokines, PDFG (platelet derived growth factor), transforming growth factor β (TGF-β), platelet activating factor, fibrin, serotonin. Chemotaxis causes neutrophil migration first, and then activation of macrophages, lymphocytes leading into phagocytosis, wound debridement, matrix activation, angiogenesis. Chemotaxis factors are complement factors, interleukin-1, TNF-α (tumour necrosis factor-α) TGF-β and platelet factor. Activated macrophages produce free radicals and nitric oxide; release cytokine to activate lymphocytes which release interferon and interleukin (called as lymphokines). These actions are reduced in diabetes mellitus, Cushing's syndrome and immunosuppression increasing the rate of sepsis.

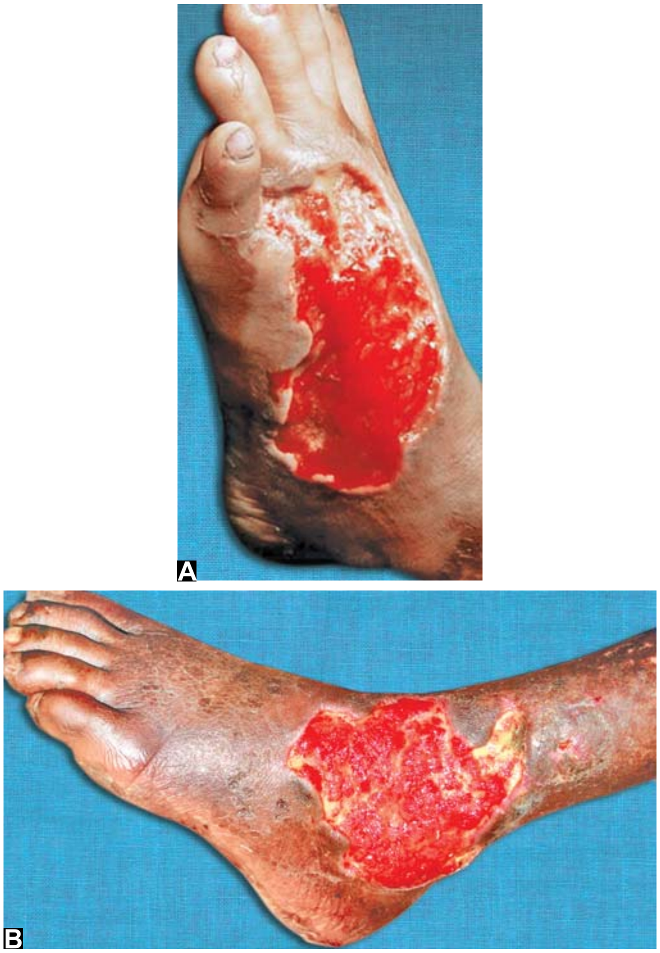

Figs 1.18A and B: Healing ulcer with healthy granulation tissue which is ready for skin grafting.

Proliferative Phase (Collagen/Fibroblastic Phase)

- Collagen and glycosamines are produced by fibroblasts.
- It begins in 7 days and lasts for 6 weeks.
- Hydroxyproline and hydroxylysine are synthesised by specific enzymes using iron, alpha ketoglutarate and vitamin C.
- Tropocollagen is produced which aggregates to form collagen fibrils.
- 80–90% of their final strength (in postoperative wounds) is achieved in 30 days.

Here proliferation of venular endothelial cell with fibroblast at wound margin and bed occurs by the action of cytokines and released growth factors. This *angiogenesis and fibroplasia* causes formation of granulation tissue which contains fibroblasts, neocapillaries, collagen, fibronectin and hyaluronic acid. Later neutrophils lead into apoptosis and die which are phagocytosed by macrophages; later macrophages go for apoptosis which are cleared by lymphocytes into draining lymph nodes.

Remodelling Phase (Maturation Phase)

- It begins at 6 weeks and lasts for 2 years.
- There is maturation of collagen by cross-linking which is responsible for tensile strength of the scar.
- Collagen production is not present after 42 days of wound healing.

Initially fibrin, fibronectin, proteoglycan deposition occurs; later collagen protein develops to form scar. Normal dermal skin contains 80% type I (20% type III) collagen; granulation tissue contains mainly type III collagen; scar contains both type I and III collagen equally. Basic essential components of collagen are proline and lysine. Hydroxylation of lysine and later glycosylation of this hydroxylysine decides the collagen molecule type. Hydroxylation of both proline and lysine as essential step needs adequate concentration of vitamin C, iron and α-ketogluteric acid. Collagen deposition in the wound is assessed by quantity of hydroxyproline excreted in urine. There is a balanced activity of collagen production and degradation of collagen (collagenolysis). Collagen is broken down by collagenase and *MMPs (matrix metallo prpteins)*. Procollagen through procollagenase → collagen fibril → cross-linking → collagen fiber → deposition. Deposited collagen → through collagenase → degradation and collagenolysis.

Scar strength is 3% in 1 week; 20% in 3 weeks; 80% in 12 weeks.

Phases of wound healing

Vascular response → blood coagulation/thrombosis → inflammation → new tissue formation → epithelialisation → wound contraction → remodelling.

Factors Affecting Wound Healing

Local factors

- Infection
- Presence of necrotic tissue and foreign body
- Poor blood supply
- Venous or lymph stasis
- Tissue tension
- Haematoma
- Large defect or poor apposition
- Recurrent trauma
- X-ray irradiated area
- Site of wound, e.g. wound over the joints and back has poor healing
- Underlying diseases like osteomyelitis and malignancy
- Mechanism and type of wound—incised/lacerated/crush/avulsion
- Tissue hypoxia locally reduces macrophage and fibroblast activity

General Factors

- Age, obesity, smoking
- Malnutrition, zinc, copper, manganese
- Vitamin deficiency (Vit C, Vit A)
- Anaemia
- Malignancy
- Uraemia
- Jaundice
- Diabetes, metabolic diseases
- HIV and immunosuppressive diseases
- Steroids and cytotoxic drugs
- Neuropathies of different causes

Management of Wounds

a. Wound is inspected and classified as per the type of wounds.
b. If it is in the vital area, then:
   - The airway should be maintained.
   - The bleeding, if present, should be controlled.
   - Intravenous fluids are started.
   - Oxygen, if required, may be given.
   - Deeper communicating injuries and fractures, etc. should be looked for.
c. If it is an incised wound then *primary suturing* is done after thorough cleaning.
d. If it is a lacerated wound then the *wound is excised* and primary suturing is done.
e. If it is a crushed or devitalised wound there will be oedema and tension in the wound. So after wound debridement or wound excision by excising all devitalised tissue, the oedema is allowed to subside for 2-6 days. Then *delayed primary suturing* is done.
f. If it is a deep devitalised wound, after wound debridement it is allowed to granulate completely. Later, if the wound is small *secondary suturing* is done. If the wound is large a split skin graft (Thiersch graft) is used to cover the defect.
g. In a wound with tension, *fasciotomy* is done so as to prevent the development of compartment syndrome.
h. Vascular or nerve injuries are dealt with accordingly. Vessels are sutured with 6-zero polypropylene nonabsorbable suture material. If the nerves are having clean cut wounds it can be sutured primarily with polypropylene 6-zero or 7-zero suture material. If there is difficulty in identifying the nerve ends or if there are crushed cut ends of nerves then marker stitches are placed using silk at the site and later secondary repair of the nerve is done.
i. Internal injuries (intracranial by craniotomy, intrathoracic by intercostal tube drainage, intra-abdominal by laparotomy) has to be dealt with accordingly. Fractured bone is also identified and properly dealt with.
j. Antibiotics, fluid and electrolyte balance, blood transfusion, tetanus toxoid (0.5 ml intramuscular to deltoid muscle), or antitetanus globulin (ATG) injection.

Wound debridement (wound toilet, or wound excision) is liberal excision of all devitalised tissue at regular intervals

> *Clinical diagnosis is an art, and the mastery of an art has no end: you can always be a better diagnostician*
> **—Logan Clendening**

(of 48-72 hours) until healthy, bleeding, vascular tidy wound is created.

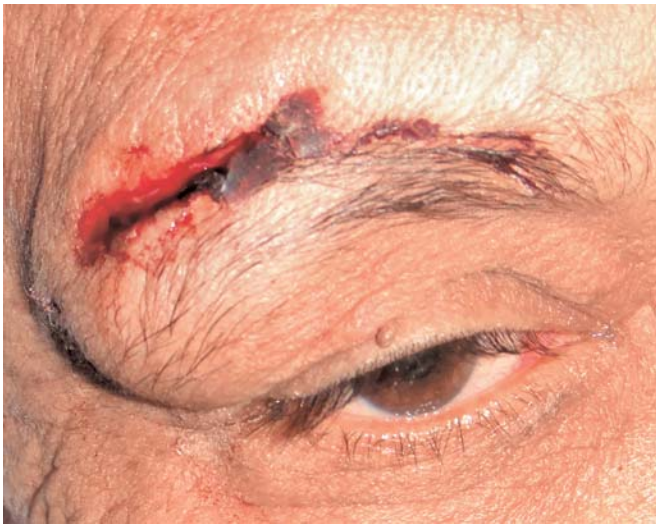

Fig. 1.19: Lacerated wound over the eyebrow. Eyebrow should not be shaved during suturing.

Primary suturing means suturing the wound immediately within 6 hours. It is done in clean incised wounds.

Delayed primary suturing means suturing the wound in 48 hours to 10 days. It is done in lacerated wounds. This time is allowed for the oedema to subside.

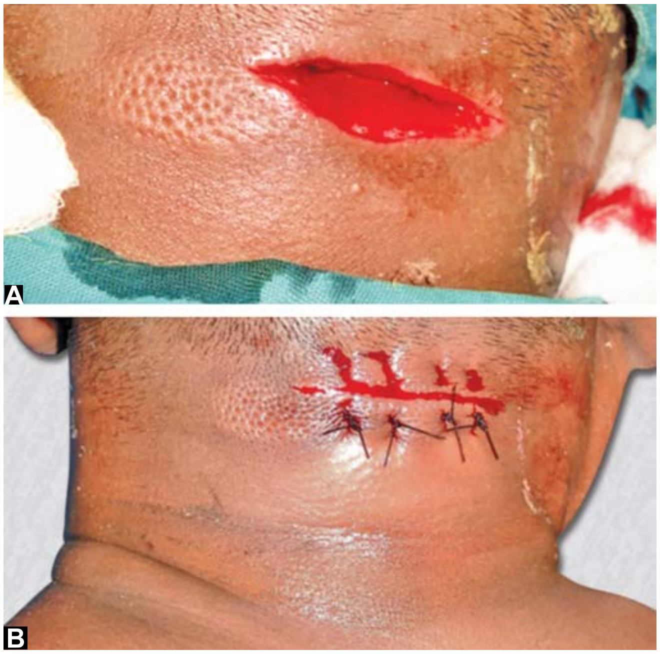

Figs 1.20A and B: Delayed primary suturing is done once oedema over the wound subsides. It is done as single layer interrupted deep sutures using monofilament polypropylene or polyethylene.

Secondary suturing means suturing the wound in 10-14 days or later. It is done in infected wounds. After the control of infection, once healthy granulation tissue appears, secondary suturing is done.

Principles of wound suturing

- Primary suturing should not be done if there is oedema/infection/devitalised tissues/haematoma
- Always associated injuries to deeper structures like vessels/nerves or tendons should be looked for before closure of the wound
- Wound should be widened by extending the incision whenever needed to have proper evaluation of the deeper structures—*proper exploration*
- Proper cleaning, asepsis, wound excision/debridement
- Any foreign body in the wound should be removed
- Skin closure if it is possible without tension
- Skin cover by graft/flap—immediate or delayed
- Untidy wound should be made tidy and clean before suturing
- Proper aseptic precautions should be undertaken
- Antibiotics/analgesics are needed
- Sutured wound should be inspected in 48 hours
- Sutures are removed after 7 days

Remember

- *Wound toilet* is washing the wound thoroughly using normal saline—ideal
- *Wound debridement* (french-letting loose) is allowing content to come out by release incisions or faciotomies. But commonly debridement is used for wound excision
- *Wound excision* is actually correct terminology for excision of devitalised tissues once or serially
- *Radical wound excision* is (pseudotumour approach) is excising entire devitalised tissues leaving tissues with visible bleeding from all layers

COMPARTMENT SYNDROME

Compartment syndrome is a special entity; common in leg, forearm, thigh and arm; is a syndrome due to increased intracompartmental pressure within a limited space area.

Causes are—narrowed space due to tight dressings/plaster cast, lying on one limb in comatous patient; increased content within the compartment due to trauma like fractures, oedema, ischaemic injury, haematoma, positioning after trauma, burn injury, etc.; high pressure injection injuries like gun injury, oil based material injury, extravasation of chemotherapeutic drugs; snake bite.

It compromises circulation and function mainly of muscles and nerves. It often maintains the normal colour and temperature of the fingers and distal pulses may not be obliterated in spite of severe muscle ischaemia. Muscle ischaemia *more than 4 hours* causes muscle death and myoglobinuria. Irreversible nerve damage develops if ischaemia persists for 8 hours. Progressive, persistent severe pain which is aggravated by passive muscle stretching is the diagnostic sign. Tense tender regional lymph node is typical. Pulse will be usually normally felt in compartment syndrome; but may become absent if there is associated arterial injury. Compartment pressure more than 30 mmHg is an indication for *fasciotomy*.

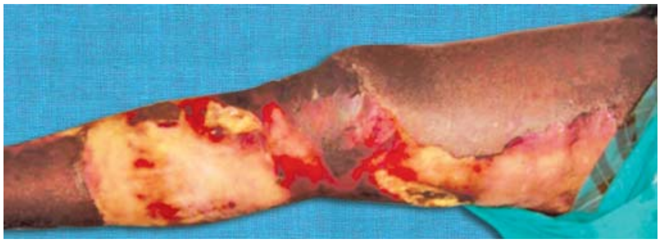

Fig. 1.21: Necrotizing fasciitis with extensive skin involvement which requires adequate wound excision and eventual skin coverage.

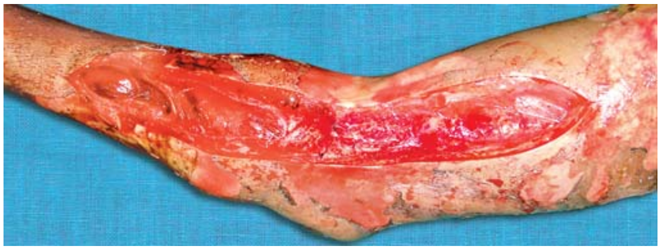

Fig. 1.22: Fasciotomy for compartment syndrome should be longitudinal, deep and lengthy and should decompress the compartment to expose the underlying muscle. It should be done early.

It is common in calf and forearm. Closed injuries cause haematoma leading to increased pressure. It is often associated with fracture of the underlying bone which in turn compresses the major vessel further aggravating the ischaemia causing ***pallor, pulselessness, pain, paraesthesia, diffuse swelling*** and ***cold limb***.

If allowed to progress it may eventually lead to ***gangrene*** or ***chronic ischaemic contracture*** with deformed, disabled limb.

Problems with the compartment syndrome

- Infection, septicaemia and abscess formation
- Renal failure
- Gangrene of the limb
- Chronic ischaemic contracture
- Disabled limb, Volkmann's ischaemic contracture

Muscle necrosis releases myoglobulin which is excreted in the urine, damages the kidneys leading into renal failure.

Note:
Affected muscle when passively stretched worsens the pain—the most reliable clinical sign.

Treatment

- Compartment pressure will be persistently more than 30 mm Hg. It can be measured by placing a fine catheter in the compartment and using a pressure monitor. This is an indication for *fasciotomy*. Adequate lengthy incision involving skin, fat and deep fascia should be done until underneath muscle bulges out properly. Multiple incisions should be made if needed. Separate incision in each compartment should be done.

Fasciotomy done in forearm anterior compartment is a specific method. Carpal tunnel *should be released by cutting flexor retinaculum.* Incision begins at the junction of the thenar and hypothenar area; extends proximally initially transverse across flexion crease of the wrist at the ulnar border; then across forearm towards radial side of forearm; then in proximal forearm towards medial side creating convex flap towards lateral side. In the elbow it crosses along the medial border to reach the arm where it runs in arm along the medial part of the anterior arm. Injury to major nerves, palmar cutaneous branch of median nerve should be avoided while placing the incision. Incision should be deepened by cutting the deep fascia along the entire length of the incision. *Dorsal fasciotomy* should be added by placing longitudinal lengthy incision in the midline. Two longitudinal incisions on the dorsum of the hand also should be made.

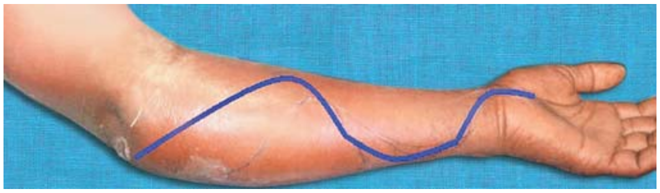

Fig. 1.23: Incision for fasciotomy in upper limb begins at flexor retinaculum extending into the forearm with a convex flap towards radial side eventually leading towards medial epicondyle of the elbow joint.

- Antibiotics.
- Catheterisation.
- Mannitol or diuretics to cause diuresis, so as to flush the kidney.
- Fresh blood transfusion.
- Hyperbaric oxygen.

CRUSH INJURY

Crush injury is one where a part of the body is being squeezed/compressed between two high force or pressure systems. It causes extensive lacerations, bruising, compartment syndrome, crush syndrome, fractures, haemorrhage, etc. with extensive tissue destruction and devitalisation. Renal failure, hypovolaemic shock and sepsis are the most dreaded problems in crush injuries.

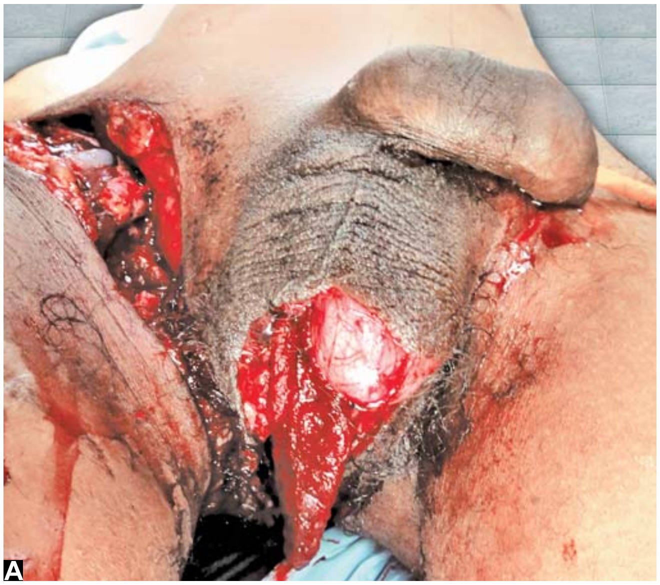

Fig. 1.24A

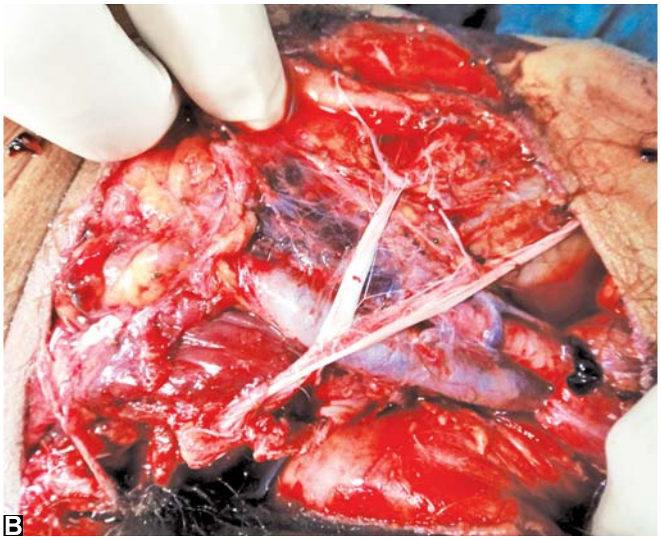

Fig. 1.24B

Figs 1.24A and B: Traumatic crush injury pelvis exposing testis and iliac vessels. Patient underwent hemipelvectomy.

CRUSH SYNDROME

It is due to crushing of muscles causing extravasation of blood and release of myohaemoglobin into the circulation leading to *acute tubular necrosis* and *acute renal failure*.

Causes

- Earthquakes.
- Mining and industrial accidents.
- Air crash.
- Tourniquet.

Initially tension increases in the ***muscle compartment*** commonly in the limb, which itself impedes the circulation and increases the ***ischaemic damage***. In 3 days, urine becomes discoloured and scanty, patient becomes restless, apathy and delirious with onset of ***uraemia***. Crush syndrome is often life-threatening. *Injury* is much worser than initial look.

Effects of crush syndrome

- Renal failure
- Toxaemia
- Septicaemia
- Disability with extensive tissue loss
- Gas gangrene

Treatment

- Tension in the muscle compartment is relieved by placing *multiple parallel deep incisions* in the limb so as to prevent further damage.
- Rheomacrodex, or mannitol is given to improve the urine output by improving the renal function.
- Alkalisation of urine is done by giving sodium citrate or sodium bicarbonate. It increases the solubility of acid haematin in the urine and so promotes its excretion. Urinary pH should be above 6.5 until urine does not show any myoglobin. Mannitol-alkaline diuresis should be 8 litre/day.
- Initial aggressive volume load using saline about 1-1.5 litres/hour is ideal in these patients.
- Haemodialysis is done sometimes as a life-saving procedure.
- Other measures:
  - Catheterization.
  - Oxygen therapy.
  - Antibiotics.
  - Blood transfusion.
  - Correction of severe hyperkalaemia.

Note:

Doing fasciotomy several days after crush injury may not be safe as it may lead to sudden release of myoglobulin causing myoglobulinuria and renal failure.

DEGLOVING INJURIES

- It occurs due to shearing force between tissue planes as traction—avulsion injury. It usually occurs between subcutaneous tissue and deep fascia or between muscle and bone. It can be localised or circumferential.

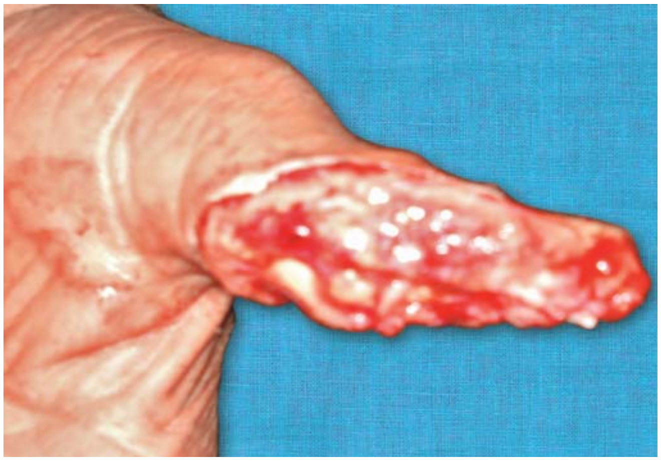

Fig. 1.25: Degloved injury.

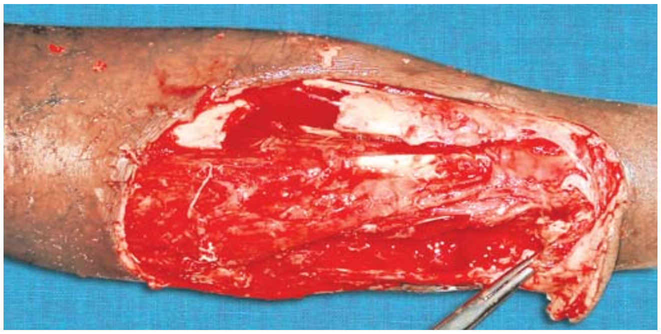

Fig. 1.26: Degloving injury wherein skin often with soft tissues gets avulsed off the deeper plane to create extensive raw area which may bleed/get infected.

- It can be in one plane or multiple planes.
- It is commonly observed in machinery injuries or major road traffic accidents. It is much more extensive than of on initial presentation.
- Under anaesthesia fluoroscein is injected intravenously and viable skin is visible as fluorescent yellowish—green colour under ultraviolet light. As injection of fluoroscein is not fully safer, serial excision is better to look for dermal punctate bleeding.
- It needs examination under general anaesthesia, wound excision/radical excision, flap coverage, microflap surgeries,

skin grafting, with proper asepsis, and blood transfusion as there is significant blood loss in these injuries.

KELOID: *'Like a claw'*

- Keloid is common in blacks. Common in females.
- Genetically predisposed. Often familial. Very rare in Caucasians.
- There is defect in maturation and stabilization of collagen fibrils. Normal collagen bundles are absent.
- Keloid continues to grow even after 6 months, may be for many years. It extends into adjacent normal skin. It is brownish black/pinkish black (due to vascularity) in colour, painful, tender and sometimes hyperaesthetic; spreads and causes itching.
- Keloid may be associated with Ehlers-Danlos syndrome or scleroderma.
- When keloid occurs following an unnoticed trauma without scar formation is called as *spontaneous keloid*, commonly seen in Negroes.
- Some keloids occasionally become nonprogressive after initial growth.
- Pathologically keloid contains proliferating immature fibroblasts, proliferating immature blood vessels and type III thick collagen stroma.

Site: Common over the sternum. Other sites are upper arm, chest wall, lower neck in front.

Differential diagnosis: Hypertrophic scar.

Treatment: Controversial.

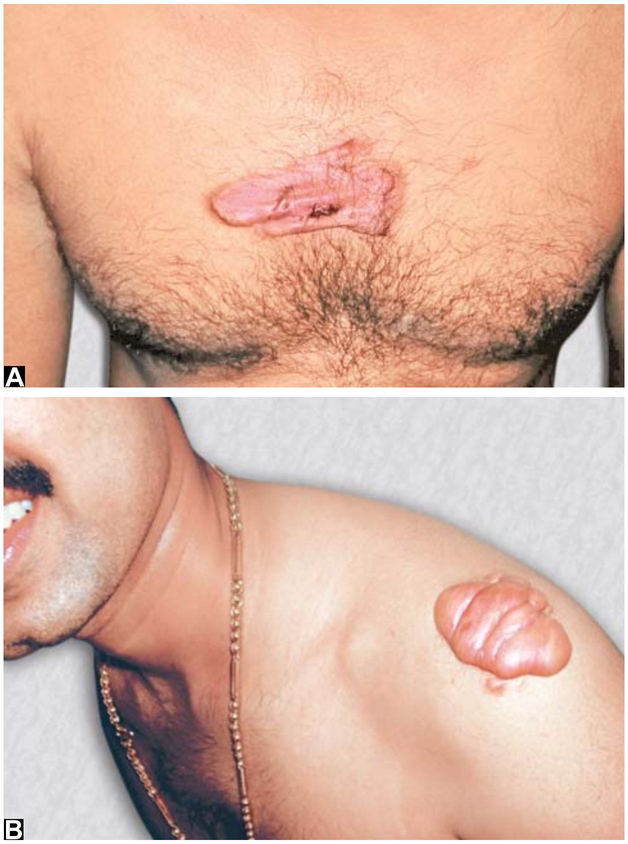

Figs 1.27A and B: Keloid in chest below the sternum and upper part of arm near shoulder—common sites of occurrence.

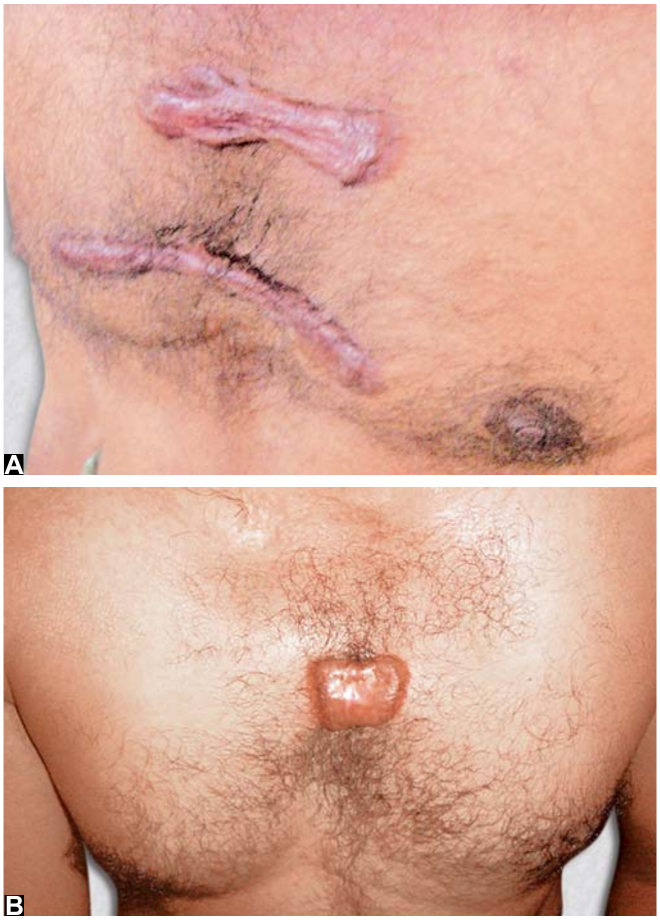

Figs 1.28A and B: Keloid over sternum (Butterfly shaped). Note the typical common site.

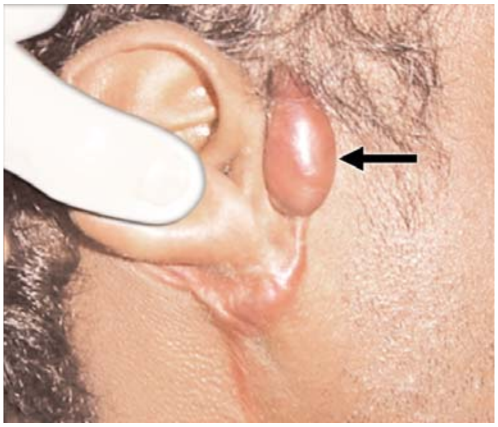

Fig. 1.29: Keloid in the upper part of the scar. It is the previous parotidectomy scar.

a. Steroid injection—***intrakeloidal triamcinolone***, is injected at regular intervals, may be once in 7-10 days, of 6-8 injections.
b. Steroid injection—excision—steroid injection.
c. Methotrexate and vitamin A therapy into the keloid.
d. Silicone gel sheeting; topical retinoids.
e. Laser therapy.
f. Vitamin E/palm oil massage.
g. *Intralesional excision* retaining the scar margin may prevent recurrence. *It is ideal and better than just excision.*

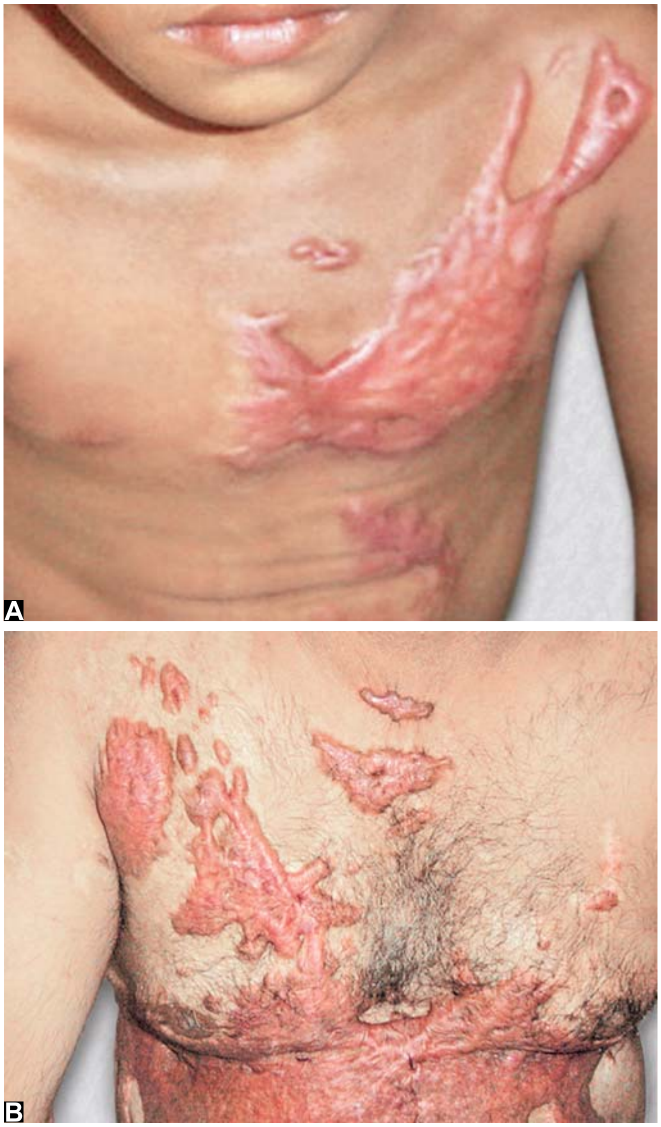

Figs 1.30A and B: Keloid in the chest wall due to burn wound.

h. Excision and irradiation or irradiation alone.
i. Excision and skin grafting may be done.

***Note*:**

Excision and primary suturing has got high recurrence rate; hence it is not usually practiced.

Recurrence rate is very high—more than 50%.

HYPERTROPHIC SCAR

- Occurs anywhere in the body.
- Not genetically predisposed. Not familial.
- Growth usually limits up to 6 months.
- It is limited to the scar tissue only. It will not extend to normal skin.

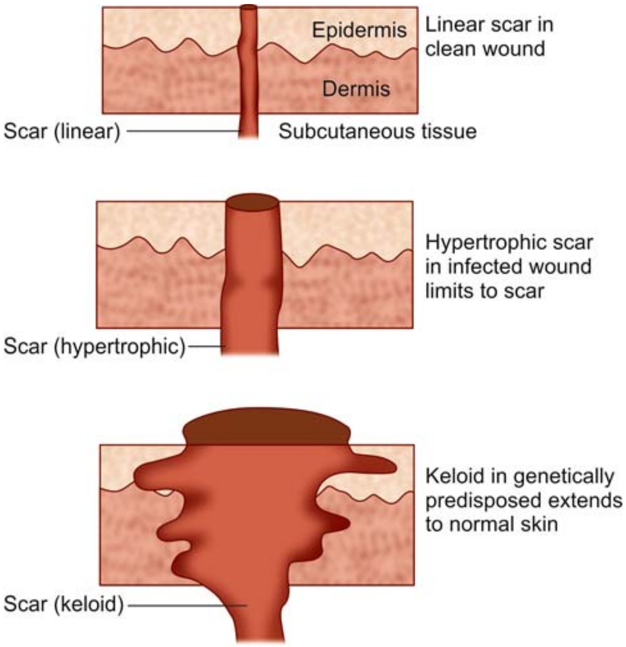
Epidermis
Linear scar in clean wound
Dermis
Scar (linear)
Subcutaneous tissue
Hypertrophic scar in infected wound limits to scar
Scar (hypertrophic)
Keloid in genetically predisposed extends to normal skin
Scar (keloid)

Fig. 1.31: Diagrammatic representation of linear, hypertrophic and keloid scar.

Differences between keloid and hypertrophic scar

		Keloid	***Hypertrophic scar***
a.	Genetic predisposition	Yes	No
b.	Site of occurrence	Chest wall, upper arm, lower neck, ear	Anywhere in the body, common in flexor surfaces
c.	Growth	Continues to grow without time limit	Growth limits for 6 months
		Extends to normal skin	Limited to scar tissue only
d.	Treatment	Poor response	Good response to steroids
e.	Recurrence	Very high	Is uncommon
f.	Collagen synthesis	20 times more than normal skin (Type III thick)	6 times more than normal skin (Type III fine collagen)
g.	Relation of size of injury and lesion	No relation. Small healed scar can form large keloid	Related to size of injury and duration of healing
h.	Age	Adolescents, middle age	Children
i.	Sex	Common in females	Equal in both
j.	Race	More in blacks (15 times)	No racial relation
k.	Structure	Thick collagen with increased epidermal hyaluronic acid	Fine collagen with increased alpha actin
l.	Features	Vascular, tender, itching	Not vascular, nontender, no itching
m.	Natural history	Progressive	Shows regression
n.	Problems	Hyperaesthesia, ulceration	Not much

- It is pale brown in colour, not painful, nontender.
- Often self-limiting also. It responds very well for **steroid** injection.
- Recurrence is uncommon.
- It is common in wounds crossing tension lines, deep dermal burns, wounds healed by secondary intention.

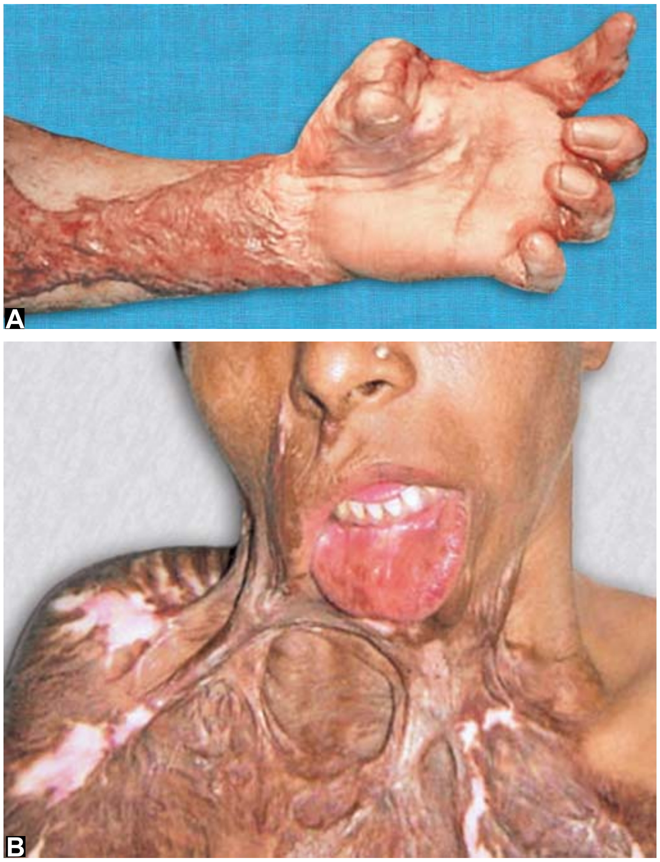

Figs 1.32A and B: Different contractures in hypertrophic scars over forearm, finger and neck.

Complications

- Often this scar breaks repeatedly and causes infection, pain.
- After repeated breakdown it may turn into ***Marjolin's ulcer***.

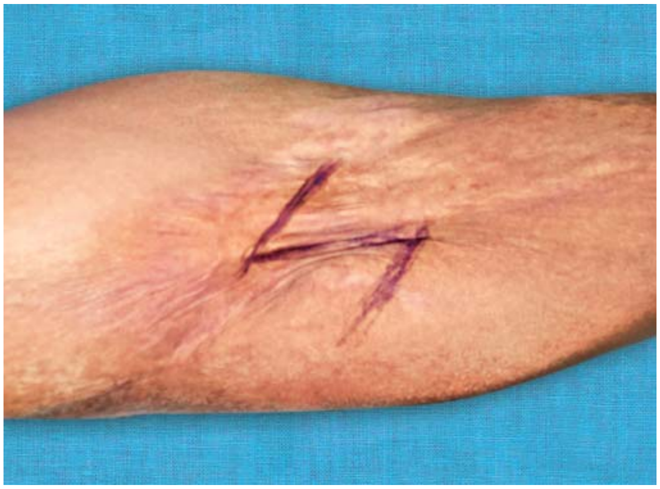

Fig. 1.33: 'Z' plasty is very useful method to release small contractures.

Treatment

It is controlled by ***pressure garments*** or often revision excision of scar and closure, if required with skin graft.

PROBLEMS WITH WOUND HEALING

- Wound infection is common in devitalized deep difficult wounds. Diabetes, immunosuppression, cytotoxic drugs, anaemia, malnutrition, malignancy increases the chances of wound infection.
- Wound dehiscence is common in all above said adverse factors. Wound suddenly gives away with pain causing copious serosanguineous discharge. After laparotomy when done specially as an emergency in trauma, acute abdomen and also in malignancy, abdominal closed wound may burst in 5-7 days. Usually all layers of abdomen give away causing discharge, occasionally bowel will extrude out. It needs emergency closure of the abdominal wound using specialized sutures or retention sutures.
- Hypertrophic scar or keloid formation due to altered collagen synthesis in the wound healing process. Collagen synthesis is increased *3 times* in hypertrophic scar and *20 times* in keloid.
- Deeper wound will cause specified problems like paraesthesia, ischaemia, paralysis, etc.

B. Ulcer

CHAPTER OUTLINE

- Ulcer
- Granulation Tissue
- Investigations for an Ulcer
- Traumatic Ulcer
- Trophic Ulcer
- Ulcer due to Chilblains
- Ulcer due to Frostbite
- Martorell's Ulcer
- Arterial/Ischaemic Ulcer
- Bairnsdale Ulcer
- Carcinomatous Ulcer
- Marjolin's Ulcer
- Rodent Ulcer
- Melanotic Ulcer
- Diabetic Ulcer
- Meleney's Ulcer
- Lupus Vulgaris
- Tuberculous Ulcer
- Bazin's Disease
- Tropical Ulcer
- Venous Ulcer
- Syphilitic Ulcer
- Soft Chancre/Soft Sore/Ducrey's Ulcer/Chancroid/Bubo
- Climatic Bubo/Tropical Bubo
- Clinical Examination of an Ulcer

ULCER

Definition

An ulcer is a break in the continuity of the covering epithelium, either skin or mucous membrane due to molecular death.

Parts of an Ulcer

a. *Margin*: It may be regular or irregular. It may be rounded or oval.

b. *Edge:* Edge is the one which connects floor of the ulcer to the margin. *Different edges are:*

Sloping edge. It is seen in a healing ulcer.

Its inner part is *red* because of red, healthy granulation tissue.

Its outer part is *white* due to scar/fibrous tissue.

Its middle part is *blue* due to epithelial proliferation.

Undermined edge is seen in a tuberculous ulcer. Disease process advances in deeper plane (in subcutaneous tissue) whereas (skin) epidermis proliferates inwards.

Punched out edge is seen in a gummatous (syphilitic) ulcer and trophic ulcer. It is due to endarteritis.

Raised and beaded edge (pearly white) is seen in a rodent ulcer (BCC). Beads are due to proliferating active cells.

Everted edge (rolled out edge): It is seen in a carcinomatous ulcer due to spill of the proliferating malignant tissues over the normal skin.

c. *Floor:* It is the one which is seen. Floor may contain discharge, granulation tissue or slough.

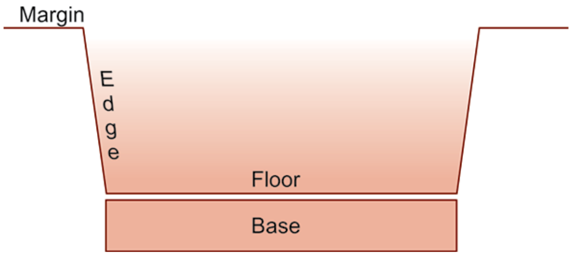

Fig. 1.34: Parts of an ulcer.

d. *Base:* Base is the one on which ulcer rests. It may be bone or soft tissue.

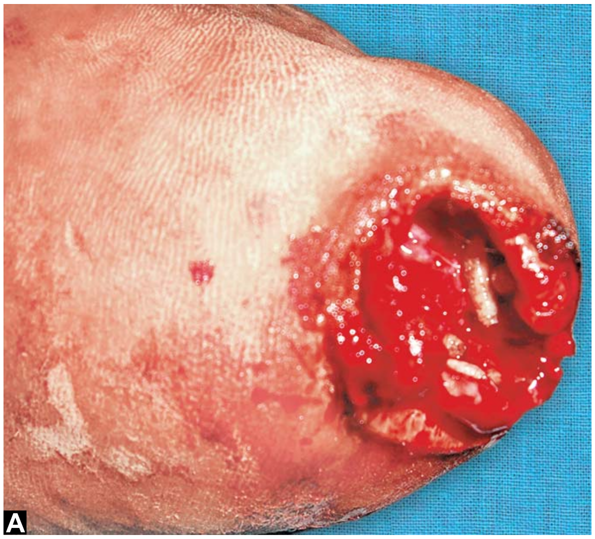

Fig. 1.35A

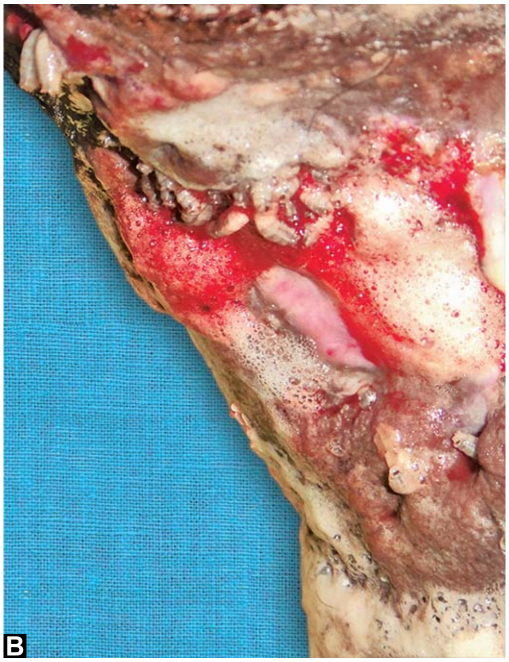

Fig. 1.35B

Figs 1.35A and B: Maggots on the ulcer bed in two different ulcers. Maggots eat only dead tissue/slough.

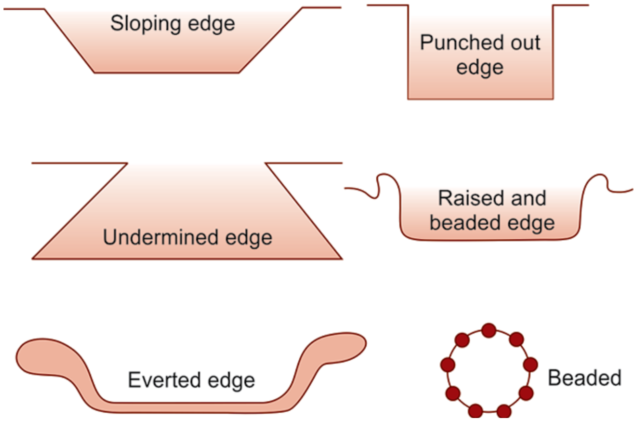

Fig. 1.37: Ulcer edges.

Induration of an Ulcer

Induration is a clinical palpatory sign which means a specific type of hardness in the diseased tissue. It is obvious in well-differentiated carcinomas. It is better felt in squamous cell carcinoma. It is also observed in long standing ulcer with underlying fibrosis. It is absent or less in poorly differentiated carcinomas and malignant melanoma. Less indurated carcinoma is more aggressive. Specific types of indurations are observed in venous diseases and chronic deep venous thrombosis. Brawny induration is a feature of an abscess. Induration is felt at edge, base and surrounding area of an ulcer. Induration at surrounding area signifies the extent of disease (tumour). Outermost part of the indurated area is taken as the point from where clearance of wide excision is planned.

Classification I (Clinical)

a. *Spreading ulcer:* Here edge is inflamed and oedematous.

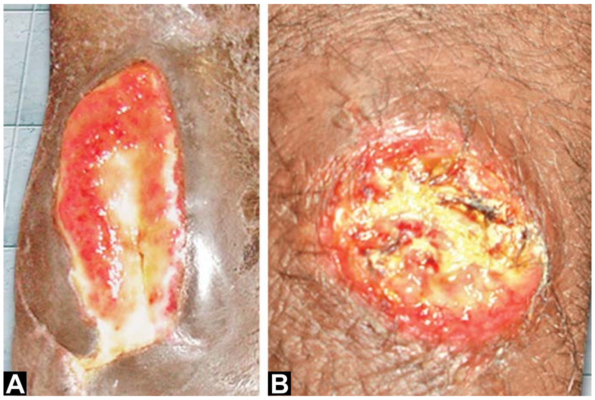

Figs 1.36A and B: Callous ulcer without any sign of healing, without any granulation tissue. *It is due to callous attitude of the patient.*

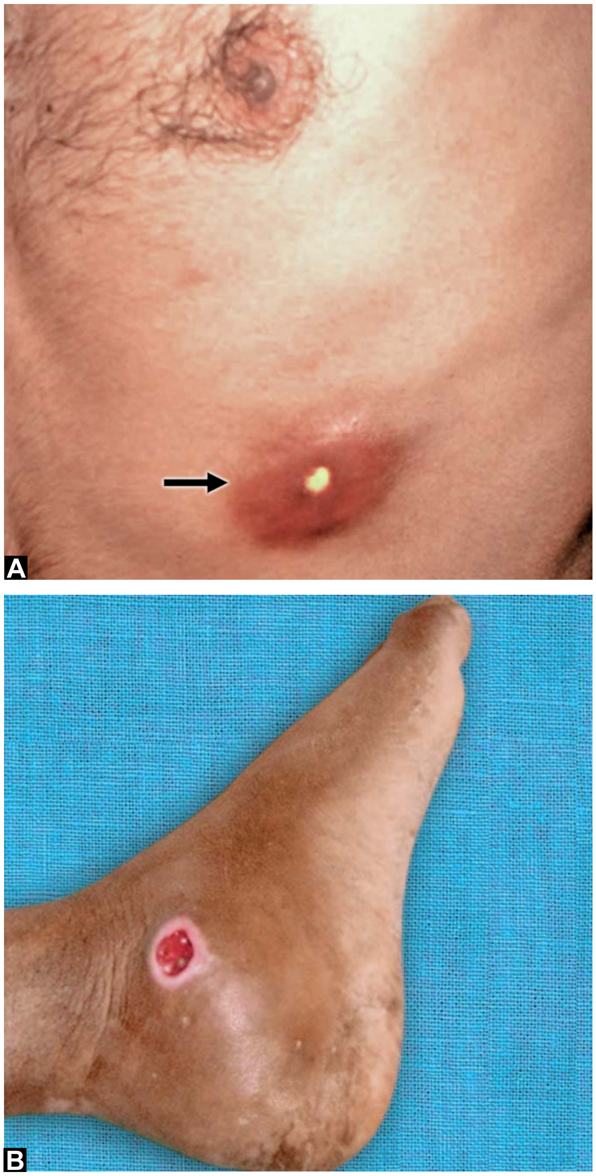

Figs 1.38A and B: Tuberculous ulcer in chest wall and ankle in two separate patients. Note the undermined edge. Discharge study, biopsy and later antituberculous drugs are the treatment. They are *usually painful*.

b. *Healing ulcer:* Edge is sloping with healthy pink/red granulation tissue with serous discharge.

c. *Callous ulcer*: Floor contains pale unhealthy granulation tissue with indurated edge/base. It lasts for many months to years. *Ulcer does not show any tendency to heal.* It is due to callous attitude of the patient.

Classification II (Pathological)

a. *Specific ulcers:*
  - Tuberculous ulcer.
  - Syphilitic ulcer: It is punched out, deep, with "wash-leather" slough in the floor and with indurated base.
  - Actinomycosis.
  - Meleney's ulcer.

b. *Malignant ulcers:*
  - Carcinomatous ulcer (Figs 1.40 and 1.41)
  - Rodent ulcer (Fig. 1.39).
  - Melanotic ulcer.

c. *Nonspecific ulcers:*
  - *Traumatic ulcer*: It may be mechanical, physical, chemical—common.
  - *Arterial ulcer:* Atherosclerosis, TAO (Fig. 1.42)
  - *Venous ulcer:* Gravitational ulcer, post-phlebitic ulcer (Fig. 1.43).
  - *Trophic ulcer/Pressure sore.*
  - *Infective ulcers*: Pyogenic ulcer (See Fig. 1.44).
  - *Tropical ulcers*: It occurs in tropical countries. It is callous type of ulcer, e.g. Vincent's ulcer.
  - *Ulcers* due to chilblains and frostbite (*cryopathic ulcer*).
  - *Martorell's hypertensive ulcer.*
  - *Bazin's ulcer.*
  - *Diabetic ulcer.*
  - *Ulcers* due to leukaemia, polycythemia, jaundice, collagen diseases, lymphoedema.
  - *Cortisol ulcers* are due to long time application of cortisol (steroid) creams to certain skin diseases. These ulcers are callous ulcers last for long time and require excision and skin grafting.

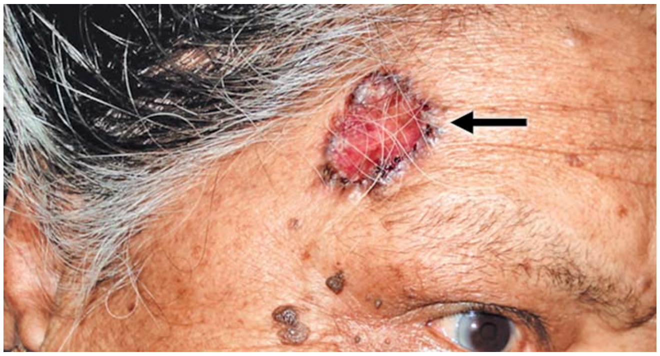

Fig. 1.39: Basal cell carcinoma of face (rodent ulcer). Ulcer edge is raised and beaded in appearance.

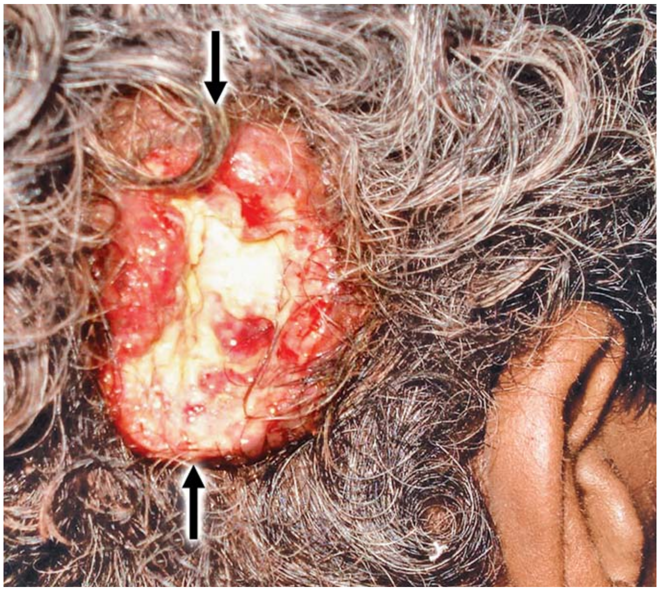

Fig. 1.40: Squamous cell carcinoma scalp. Note the ulceroproliferative lesion with everted edge.

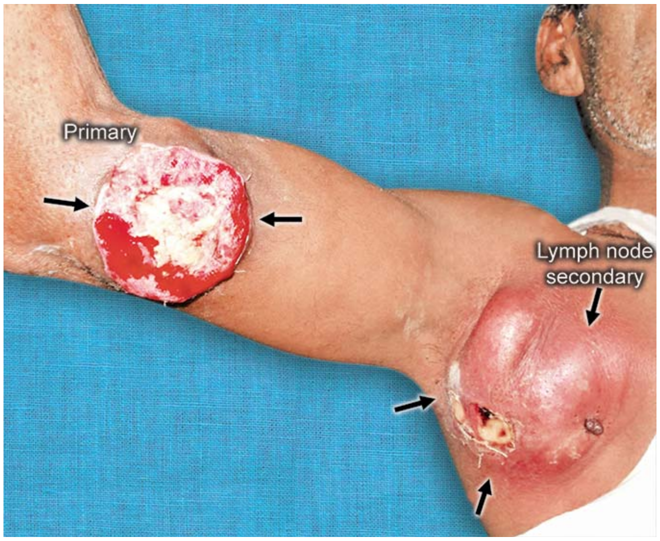

Fig. 1.41: Squamous cell carcinoma in the arm with secondaries in the axillary lymph node. Friable tumour tissues in the floor cause bleeding after trauma. Secondaries are fixed with ulceration. It is advanced disease.

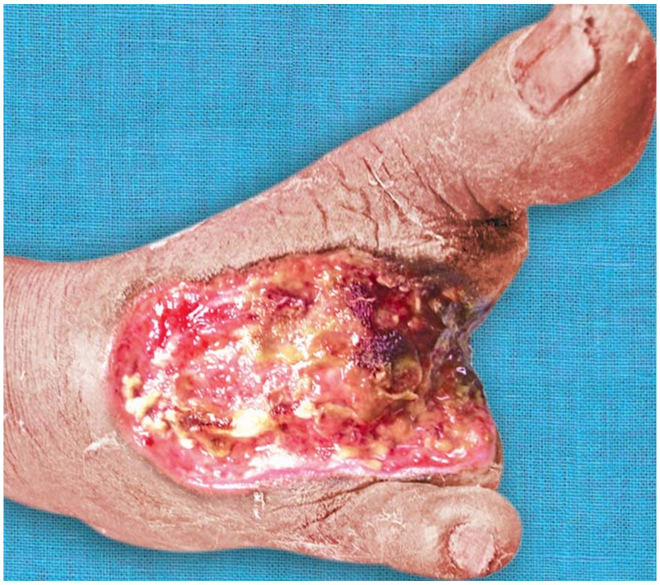

Fig. 1.42: Ischaemic ulcer foot. Middle three toes are already amputated because of gangrene.

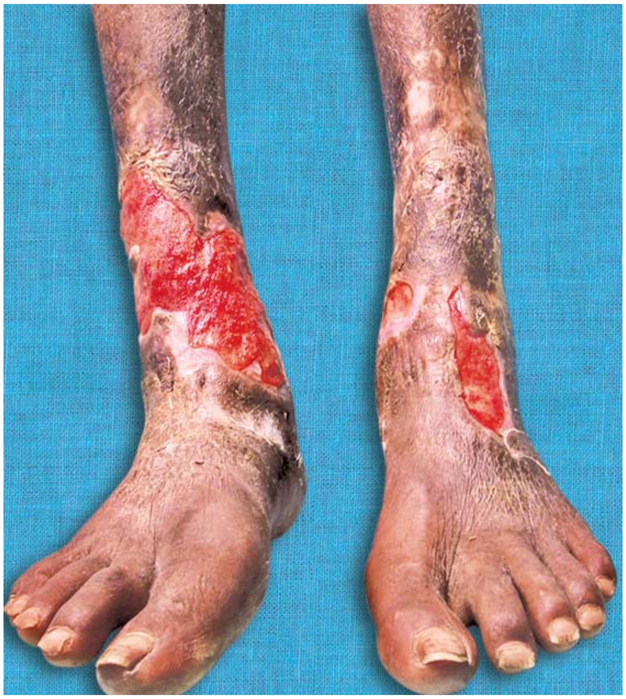

Fig. 1.43: Venous ulcers in both feet. Site is around ankle (Gaiter's zone). There are healthy granulation tissues. It needs skin grafting and definitive procedure for varicose veins after evaluation.

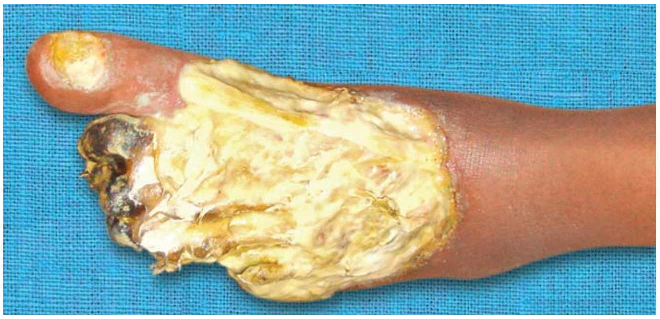

Fig. 1.44: Infective ulcer in the foot. Note the quantity of slough, exposed tendon and gangrenous toes. Patient requires below or above knee amputation.

Wagner's Grading/Classification of Ulcer

Grade 0 – Preulcerative lesion/healed ulcer
Grade 1 – Superficial ulcer
Grade 2 – Ulcer deeper to subcutaneous tissue exposing soft tissues or bone
Grade 3 – Abscess formation underneath/osteomyelitis
Grade 4 – Gangrene of part of the tissues/limb/foot
Grade 5 – Gangrene of entire one area/foot

GRANULATION TISSUE

It is proliferation of new capillaries and fibroblasts intermingled with red blood cells and white blood cells with thin fibrin cover over it (See Figs 1.51 and 1.52)

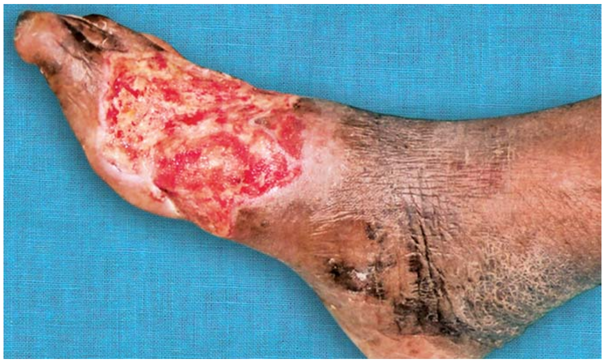

Fig. 1.45: Nonhealing ulcer with pale unhealthy granulation tissue.

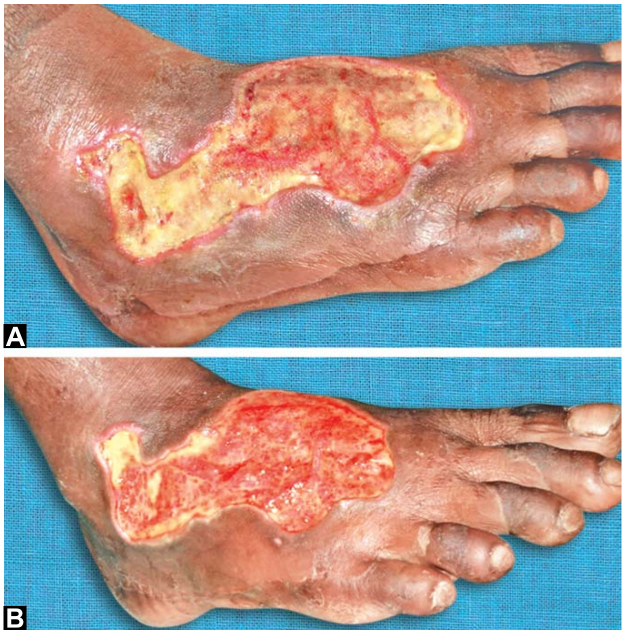

Figs 1.46A and B: Ulcer in the foot, initially with slough; later after slough excision and regular dressings. Area requires skin grafting.

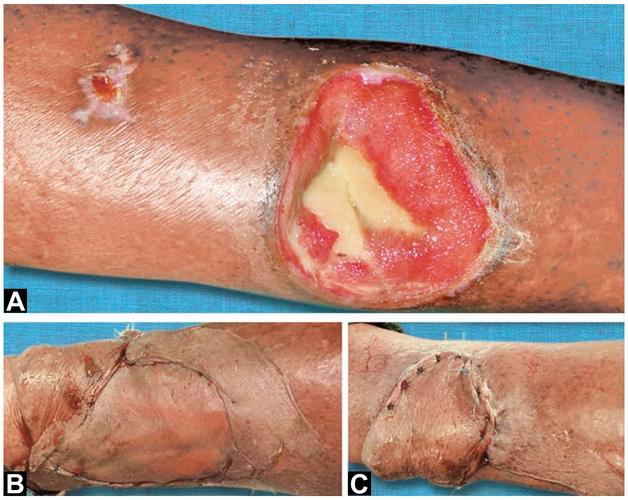

Figs 1.47A to C: Ulcer leg with exposed bone. Patient underwent local rotation flap to cover. Area from where flap is rotated is covered with split skin graft. When the bone is exposed, skin grafting is not possible.

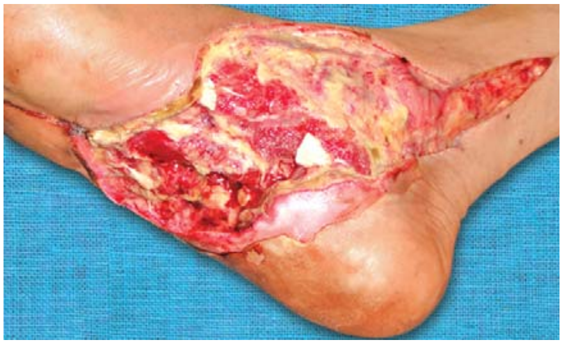

Fig. 1.48: Large ulcer in the foot and leg with exposed tendon.

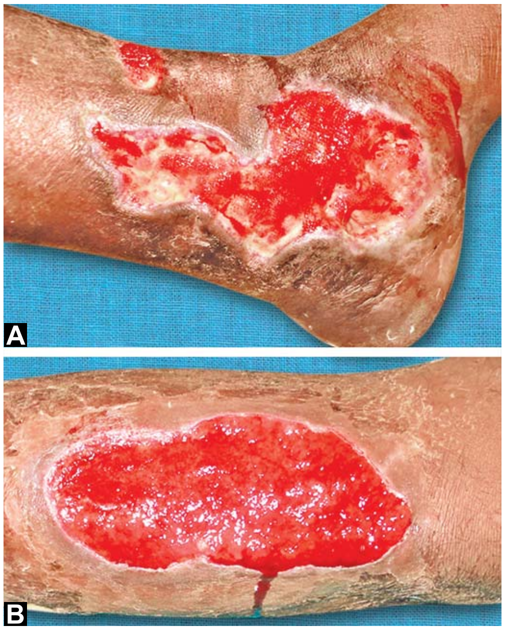

Figs 1.49A and B: Ulcers with healthy granulation tissue. Ulcer is ready for split skin grafting.

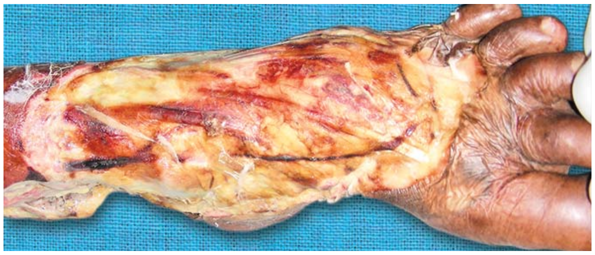

Fig. 1.50: Nonhealing ulcer foot in a diabetic patient with *Pseudomonas* infection. Note the greenish discharge in the wound. *Pseudomonas* infection is commonly hospital acquired.

Different discharges in an ulcer (as well as from a sinus)

a. *Serous:* In healing ulcer
b. *Purulent:* In infected ulcer
    Staphylococci: Yellowish and creamy
    Streptococci: Bloody and opalescent
    Pseudomonas: Greenish colour due to pseudocyanin
c. *Bloody:* Malignant ulcer, healing ulcer from healthy granulation tissue
d. *Seropurulent*
e. *Serosanguinous:* Serous and blood
f. *Serous with sulphur granules:* Actinomycosis
g. *Yellowish:* Tuberculous ulcer

Types

- ***Healthy granulation tissue***: It occurs in a healing ulcer. It has got sloping edge. It bleeds on touch. It has got serous discharge. 5 Ps of granulation tissue—*Pink, Punctate haemorrhages, Pulseful, Painless, Pin head granulation.* Skin grafting takes up well with healthy granulation tissue. Streptococci growth in culture should be less than 10^5/gram of tissue before skin grafting.

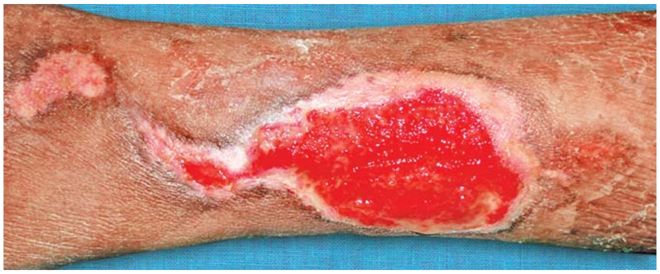

Fig. 1.51: Healing ulcer with healthy granulation tissue. Note the sloping edge.

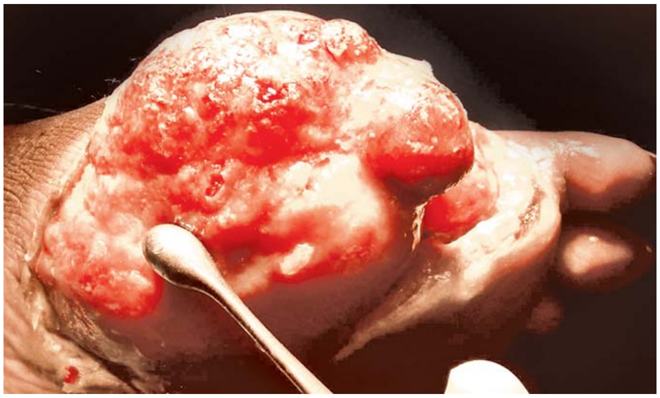

Fig. 1.52: Exuberant granulation tissue (proud flesh) in an ulcer. It should be scooped out using Volkmann's scoop prior to skin grafting.

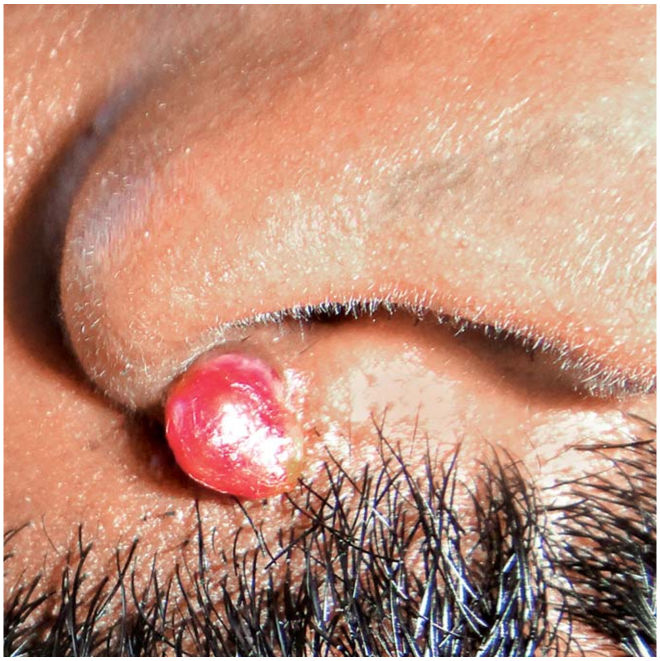

Fig. 1.53: Pyogenic granuloma.

- ***Unhealthy granulation tissue:*** It is pale with purulent discharge. Its floor is covered with slough. Its edge is inflamed and oedematous. It is a spreading ulcer.
- ***Unhealthy, pale, flat granulation tissue***: It is seen in chronic nonhealing ulcer (callous ulcer).
- ***Exuberant granulation tissue (Proud flesh)***: It occurs in a sinus or ulcer wherein granulation tissue protrudes out of the sinus opening or ulcer bed like a proliferating mass. It is commonly associated with a retained foreign body in the sinus cavity.
- Sprouting granulation tissue of sinus.
- ***Pyogenic granuloma***: It is a type of exuberant granulation tissue. Here granulation tissue from an infected wound or ulcer bed protrudes out, presenting as a well-localised, red swelling which bleeds on touching.
  Differential diagnosis: PapillomXa, skin adnexal tumours.
  Treatment: Antibiotics, excision and sent for biopsy.

INVESTIGATIONS FOR AN ULCER

- *Study of discharge*: Culture and sensitivity, AFB study, cytology.

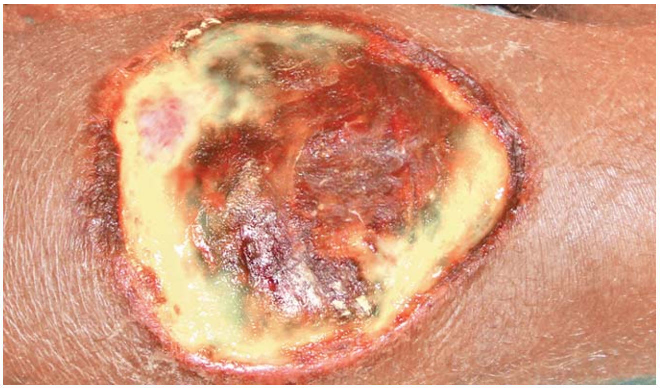

Fig. 1.54: Typical greenish coloured ulcer due to *Pseudomonas* infection.

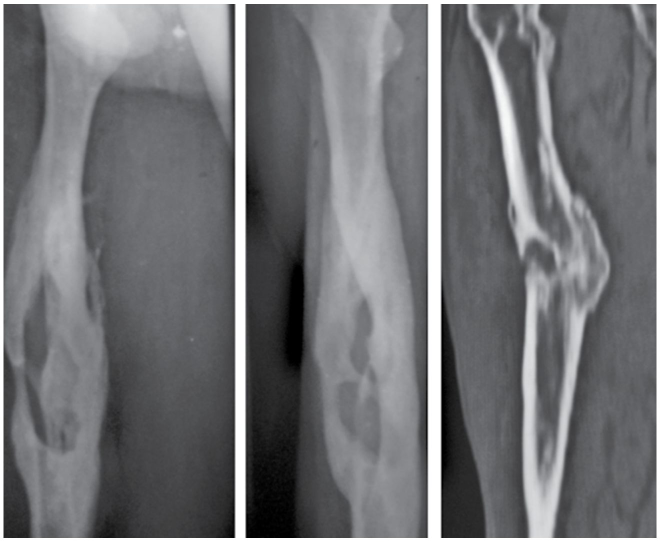

Fig. 1.55: X-ray showing osteomyelitis with sequestrum inside. Osteomyelitis prevents ulcer healing. Bone thickening on clinical examination is typical.

- *Edge biopsy*: Biopsy is taken from the edge because edge contains multiplying cells. Usually two biopsies are taken. Biopsy taken from the centre may be inadequate because of central necrosis.
- X-ray of the part to look for periostitis/osteomyelitis.
- FNAC of the lymph node.
- Chest X-ray, Mantoux test in suspected case of tuberculous ulcer.

Note:
Ulcer will not granulate if haemoglobin is less than 10 gm% and serum albumin is < 3 gm%.

Assessment of an Ulcer

- Cause of an ulcer should be found—diabetes/venous/arterial/infective.
- Clinical type should be assessed.
- Assessment of wound is important—anatomical site; size and depth of the wound; edge of the wound; mobility; fixity; induration; surrounding area; local blood supply. *Wound perimeter* may be useful in assessing this.
- *Wound imaging* is done by tracing it on a transparent acetate sheet at regular intervals.
- Presence of systemic features; regional nodal status; function of the limb/part; joint movements; distal pulses; sensations should be assessed.
- Severity of infection should be assessed—culture of discharge.
- Specific investigations like edge biopsy; X-ray of part; blood sugar; arterial/venous Doppler; angiogram.

Management of an ulcer

- Cause should be found and treated
- Correct the deficiencies like anaemia, protein and vitamins deficiencies
- Transfuse blood if required
- Control the pain
- Investigate properly
- Control the infection and give rest to the part
- Care of the ulcer by ***debridement, ulcer cleaning and dressing is done***
- Remove the exuberant granulation tissue
- Topical antibiotics for infected ulcers only like framycetin, silver sulphadiazine, mupirocin
- Antibiotics are not required once healthy granulation tissues are formed
- Once granulates, defect is closed with secondary suturing, skin graft, flaps

Ulcer cleaning is done using dilute povidone iodine and normal saline (*better and ideal*). It should be done daily or two times a day depending on the severity.

Conflict is the primary engine of creativity and innovation.—**Ranold Heifetz**

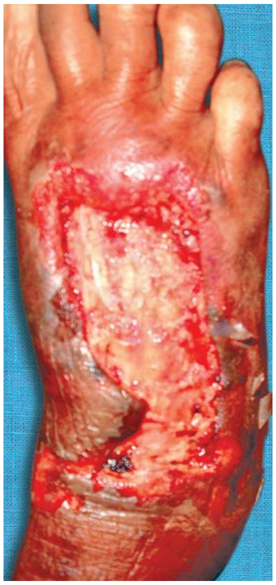

Fig. 1.56: Ulcer after wound debridement. Adequate wound excision is essential whenever there is slough and necrosis to allow granulation tissue to develop.

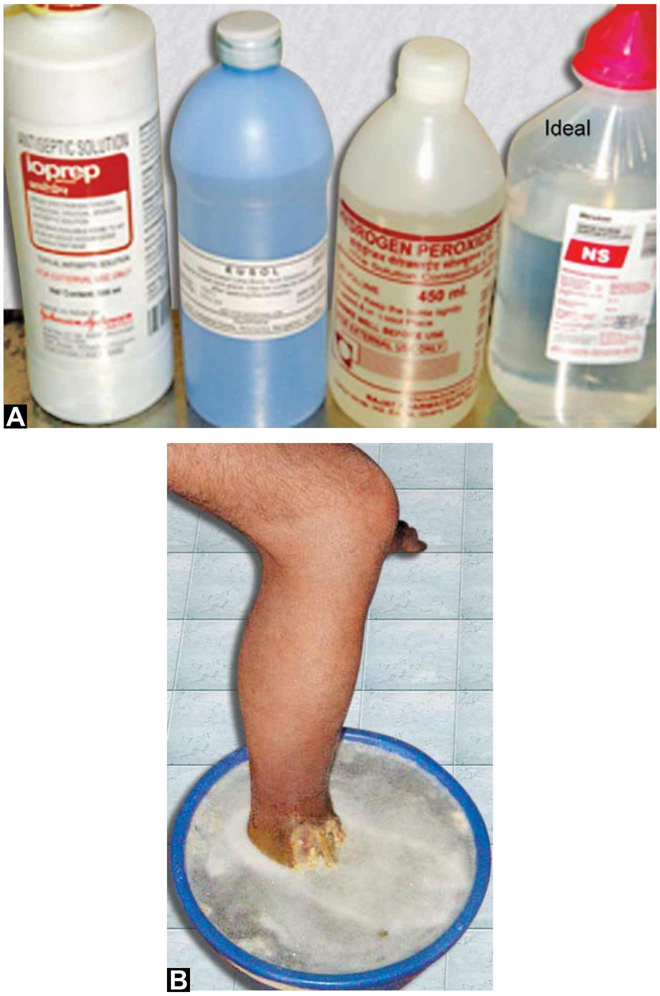

Figs 1.57A and B: Solutions commonly used for ulcer dressing—povidone iodine (brownish); EUSOL (colourless); hydrogen peroxide and normal saline. Note the EUSOL bath. Dilute EUSOL solution in a basin is used wherein ulcer foot is dipped and kept in place for 20-30 minutes. EUSOL removes the slough and cleans the ulcer bed. Hydrogen peroxide releases nascent oxygen and helps in removing necrotic material.

Debridement of an ulcer

- It is removal of devitalised tissue
- Small ulcers are debrided in ward
- Large ulcers are debrided in operation theatre under general anaesthesia
- All dead, devitalised, necrotic tissues are removed
- Slough should be separated adequately before debridement
- Often devitalised tissue separates on its own by autolysis
- Enzymes like collagenase are used for debridement
- Hydrotherapy and dressings are mechanical nonselective method of debridement

Dressing of an ulcer is done

- To keep ulcer moist
- To keep surrounding skin dry
- To reduce pain
- To soothen the tissue
- To protect the wound
- As an absorbent for the discharge

Note:

Debridement can be surgical, mechanical, autolytic or enzymatic.

Ulcer dressings

- *Cotton dressing*—cheap but traumatic
- *Paraffin dressing*
- *Polyurethane dressings* used in clean wounds
- *Alginates (seaweed) dressing* used when there are heavy exudates
- *Type 1 collagen dressings* cause haemostasis, proliferation of fibroblasts and improve the blood supply
- *Foam dressings* are highly absorbent, decrease the wound maceration, and reduce the frequency of dressing—hydrophilic polyurethane foam
- *Hydrocolloid dressings* help in separation of slough and autolysis of dead tissues
- *Transparent film dressings* are waterproof, permit oxygen and water vapour across and prevent contamination
- *Hydrogel dressings* used for clean wounds

Causes of formation of chronic ulcers in the skin

- Recurrent infection
- Trauma
- Absence of rest
- Poor blood supply
- Hypoxia
- Oedema of area
- Loss of sensation
- Malignancy
- Specific cause like tuberculosis
- Fibrosis
- Periostitis or osteomyelitis of the underlying bone

TRAUMATIC ULCER

- Such ulcer occurs after trauma. It may be *mechanical*—dental ulcer along the margin of the tongue due to tooth injury; *physical* like by electrical burn; *chemical* like by alkali injury.

- Such ulcer is acute, superficial, painful and tender. Secondary infection or poor blood supply of the area make it chronic and deep.
- *Footballer's ulcer* is a traumatic ulcer occurring over the shin of males due to direct knocks on the shin. It is staphylococcal infection with a chronic and deep ulcer.
- Traumatic ulcers can occur anywhere in the body due to trauma.

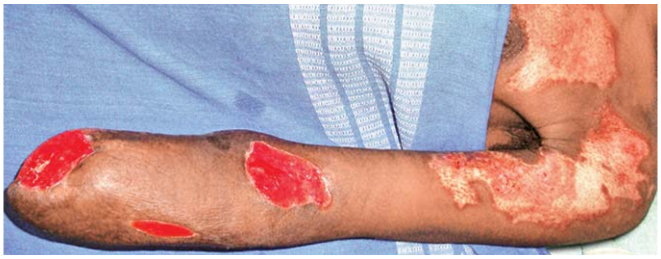

Fig. 1.58: Traumatic ulcers in the upper limb.

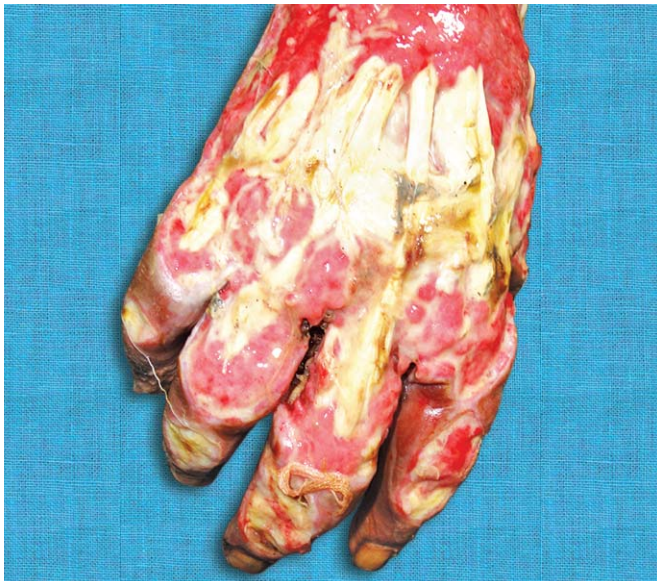

Fig. 1.59: Degloving ulcer with devitalised exposed tendons on the dorsum of the hand.

- Trauma causes infection, necrosis, fasciitis, crush injury, endarteritis of the skin leading into formation of large/deep nonhealing ulcer.
- Treatment depends on size and extent of ulcer. Regular dressing, later skin grafting is done.

TROPHIC ULCER (PRESSURE SORE/ DECUBITUS ULCER)

Pressure sore is tissue necrosis and ulceration due to prolonged pressure. Blood flow to the skin stops once external pressure becomes more than 30 mmHg (more than capillary occlusive pressure) and this causes tissue hypoxia, necrosis and ulceration. It is more prominent between bony prominence and an external surface.

It is due to:
- Impaired nutrition.
- Defective blood supply.
- Neurological deficit.

Sites

- Over the ischial tuberosity.
- Sacrum.
- In the heel.
- In relation to heads of metatarsals.
- Buttocks.
- Over the shoulder.
- Occiput.

Due to the presence of neurological deficit, trophic ulcer is also called as *neurogenic ulcer/neuropathic ulcer.* Initially it begins as callosity due to repeated trauma and pressure, under which suppuration occurs and gives way through a central hole which extends down into the deeper plane up to the underlying bone as *perforating ulcer (penetrating ulcer*).

Bedsores are trophic ulcers.

Factors causing pressure sore

- Normal stimulus to relieve the pressure is absent in anaesthetised patient
- Nutritional deficiencies worsens the necrosis
- Inadequate padding over the bony prominences in malnourished patients
- Urinary incontinence in paraplegia patient causes skin soiling-maceration-infection-necrosis

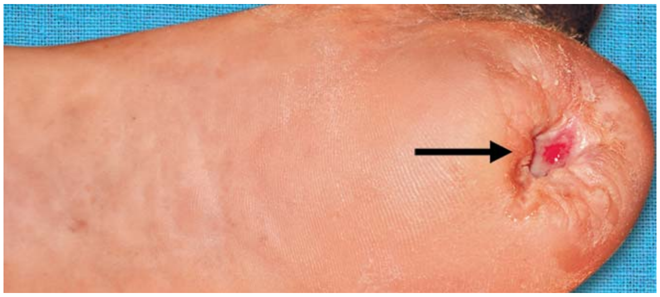

Fig. 1.60: Trophic ulcer—heel. It is deep punched out ulcer. It often requires rotation flap/transposition flap.

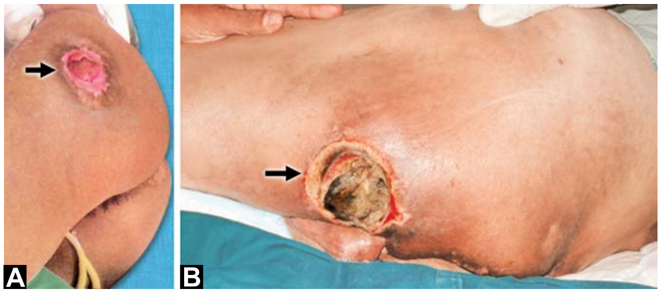

Figs 1.61A and B

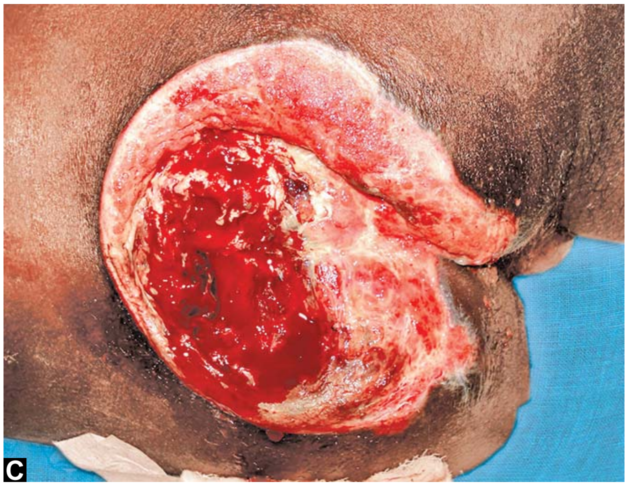

Fig. 1.61C

Figs 1.61A to C: Bedsore (decubitus ulcer) over the greater trochanter, ischium and sacrum. Bedsore is a trophic ulcer. It is usually with punched out edge.

Clinical Features

- Occurs in 5% of all hospitalised patients.
- Painless ulcer which is punched out.
- Ulcer is nonmobile with base formed by bone.

Neurological causes

- Diabetic neuropathy
- Peripheral neuritis
- Tabes dorsalis
- Spina bifida
- Leprosy
- Spinal injury
- Paraplegia
- Peripheral nerve injury
- Syringomyelia

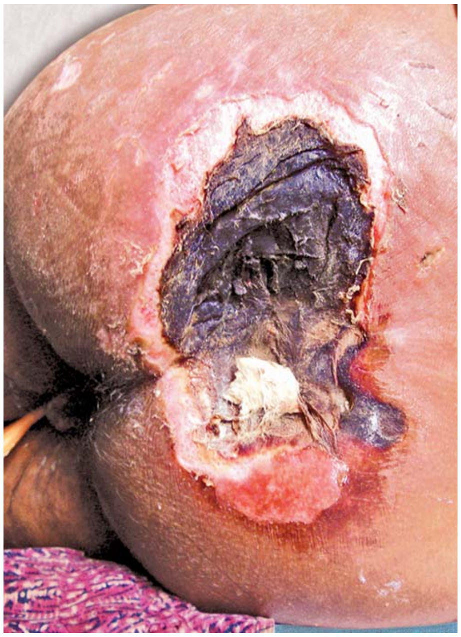

Fig. 1.62: Bedsore over the sacrum with thick black slough on the surface.

Staging of pressure sore

- Nonblanching erythema—early superficial ulcer
- Partial thickness skin loss—late superficial ulcer
- Full thickness skin loss extending into subcutaneous tissue but not through fascia—early deep ulcer
- Full thickness skin loss with fascia and underlying structures like muscle/tendon/bone, etc—late deep ulcer

Investigations

Study of discharge, blood sugar, biopsy from the edge, X-ray of the part, X-ray spine.

Treatment

- Cause should be treated.
- Nutritional supplementation.
- Rest, antibiotics, slough excision, regular dressings.
- *Vacuum assisted closure (VAC):* It is the creation of intermittent negative pressure of minus 125 mmHg to promote formation of healthy granulation tissue. Negative pressure reduces tissue oedema, clears the interstitial fluid and improves the perfusion, increases the cell proliferation and so promotes the healing. A perforated drain is kept over the foam dressing covered over the pressure sore. It is sealed with a transparent adhesive sheet. Drain is connected to required vacuum apparatus.

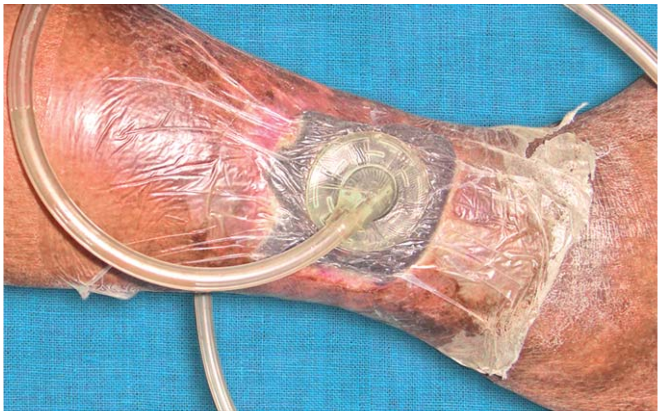

Fig. 1.63: Vacuum assisted closure of an ulcer.

- Once ulcer granulates well, flap cover or skin grafting is done.
- Excision of the ulcer and skin grafting.
- Flaps—local rotation or other flaps (transposition flaps).
- Cultured muscle interposition.
- *Proper care:* Change in position once in 2 hours; lifting the limb upwards for 10 seconds once in 10 minutes; nutrition; use of water bed/air bed/air-fluid floatation bed and pressure dispersion cushions to the affected area; urinary and faecal care; hygiene; psychological counselling. Regular skin observation; keeping skin clean and dry (using regular use of talcum powder); oil massaging of the skin and soft tissues using clean, absorbent porous clothing; control and prevention of sepsis helps in the management.

ULCER DUE TO CHILBLAINS

- It is due to exposure to intense cold causing blisters and ulcerations in the feet.
- These ulcers are ***superficial.***
- It is due to excessive cutaneous arteriolar constriction.
- The condition is also called as ***perniosis.***

ULCER DUE TO FROSTBITE

- It is due to exposure of a part to wet cold below the freezing point (cold wind).
- There is arteriolar spasm, denaturation of proteins and cell destruction.
- It leads to gangrene of the part.
- Ulcers here are always ***deep.***

MARTORELL'S ULCER (1945)

- It is seen in *hypertensive* patients often with atherosclerosis.
- It is seen in *calf*. Often it is *bilateral* and *painful*.
- Necrosis of calf skin occurs with sloughing away and formation of deep, punched out ulcers extending into the deep fascia.
- There is sudden obliteration of the arterioles of the calf skin.
- All *peripheral pulses* are present.
- *It takes months to heal.*

  Treatment: Once ulcer granulates well, skin grafting with lumbar sympathectomy is done.

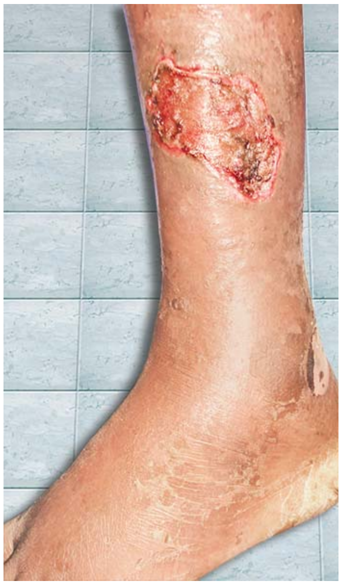

Fig. 1.64: Typical Martorell's ulcer.

ARTERIAL/ISCHAEMIC ULCER

- It is common in toes, feet or legs; often can occur in upper limb digits. It is due to poor blood supply following blockage of the digital or medium sized arteries.
- Atherosclerosis and TAO (Thromboangiitis obliterans) are common causes in lower limb.
- Cervical rib, Raynaud's phenomenon and vasculitis are common causes in upper limb.
- Ulcer initially occurs after trauma, soon becomes nonhealing, spreading with scanty granulation tissue.
- Ulcer is very painful, tender and often hyperaesthetic. Digits may often be gangrenous. Intermittent claudication, rest pain are common. Other features of ischaemia are obvious in the adjacent areas. They are—pallor, dry skin, brittle nail, patchy ulcerations, and loss of hair.
- Ulcer is usually deep, destructs the deep fascia, exposing tendons, muscles and underlying bone. Dead tendons look pale/greenish with pus over it.
- **Management:** Specific investigations like arterial Doppler, angiogram, lipid profile, and blood sugar are done. *Treatment* is done accordingly—drugs like vasodilators; arterial surgeries may be needed.

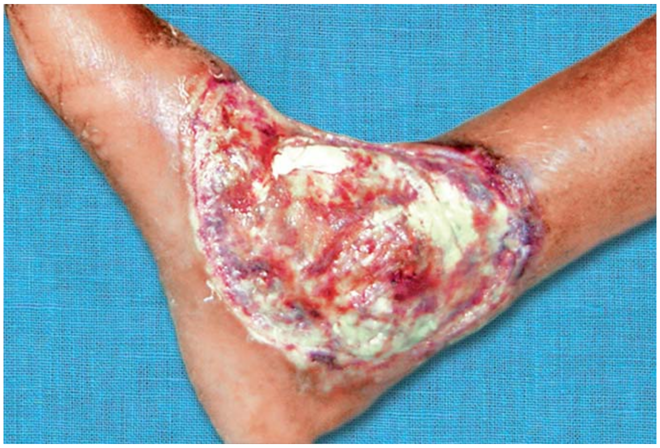

Fig. 1.65: Ischaemic ulcer foot. This ulcer is likely to lead into gangrene and eventual amputation.

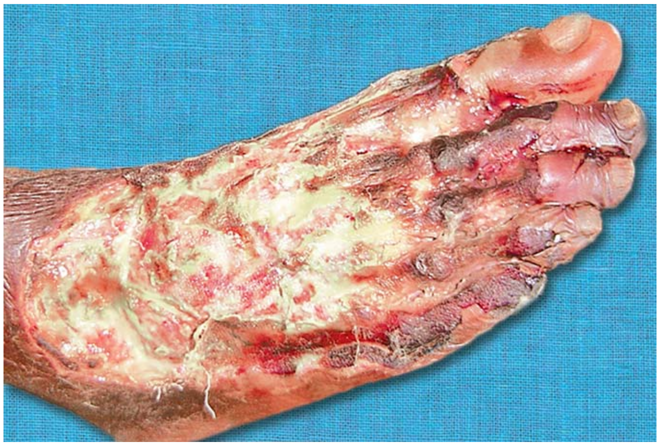

Fig. 1.66: Ischaemic ulcer foot with gangrene of toes without evidence of any bleeding with exposed tendons

BAIRNSDALE ULCER

- It is a chronic, irregular, undermined ulcer due to *Mycobacterium ulcerans* infection.
- Deep severe form, with extensive dermal necrosis is called as ***'Buruli ulcer.'***
- Discharge study will show acid-fast bacilli.
- Antituberculous drugs resolve the ulcer usually. Skin grafting may be required later.

CARCINOMATOUS ULCER (EPITHELIOMA, SQUAMOUS CELL CARCINOMA)

- It arises from prickle cell layer of skin. It may initially begin as a nodule or ulcer; but later forms an ulcerative lesion with rolled out/everted edge. Floor contains necrotic content, unhealthy (tumour) granulation tissue and blood.
- Ulcer bleeds on touch and is vascular and friable. Induration is felt at the base and edge. It is usually circular or irregular in shape. Initially ulcer is mobile but becomes nonmobile once it infiltrates into deeper tissues. The typical foul smell is due to necrotic material, infection and release of polyamides from the tumour cells.
- Hard, discrete regional lymph nodes are often palpable, initially mobile but later become fixed. Lymph nodes can fungate eventually. Ulcer and lymph nodes are initially painless; but becomes painful and tender once there is deeper infiltration or secondary infection. Systemic spread is rare. It is a *locoregional malignant* disease.
- *Verrucous carcinoma* is exophytic, locally malignant well differentiated squamous cell carcinoma without lymphatic spread.
- **Management:** Edge biopsy; FNAC of regional lymph nodes are the investigations. Treated with wide local excision with skin grafting and regional lymph node block dissection.

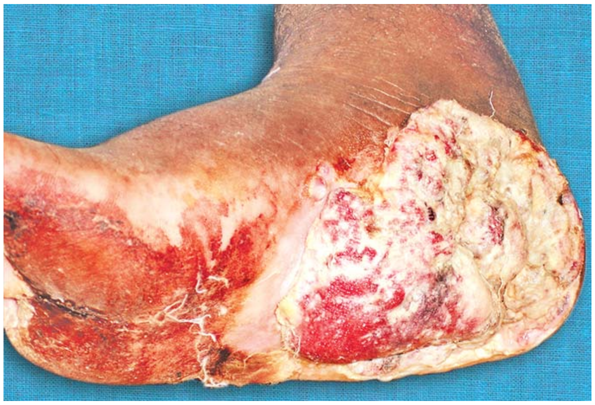

Fig. 1.67: Squamous cell carcinoma heel. It is proliferative ulcer. Note the raised and everted edge.

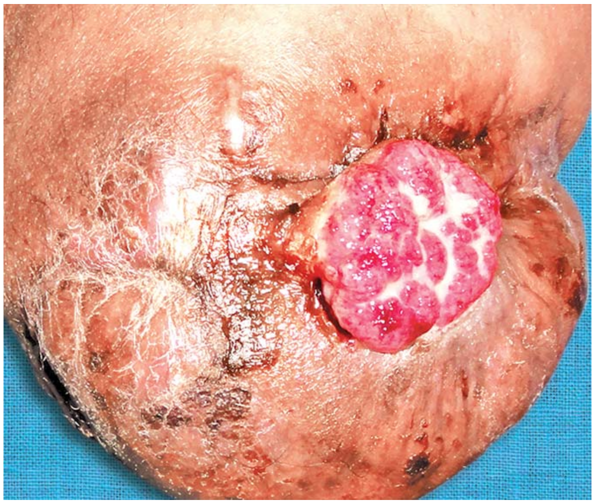

Fig. 1.68: Proliferative squamous cell carcinoma in the amputation stump. Patient has undergone amputation earlier for distal carcinomatous ulcer. This is recurrence at the stump site. SCC can be proliferative, ulceroproliferative or proliferative ulcer on clinical look.

MARJOLIN'S ULCER (Rene Marjolin,1828, Paris)

- It is *slow* growing *locally* malignant lesion—a very well differentiated squamous cell carcinoma occurring in an *unstable scar of long duration*.
- It is commonly seen in chronic venous ulcer scar. Often it is observed in burns scar and scar of previous snake bite. Lesion is ulcerative/proliferative.
- Edge may be everted or may not be. It is painless as scar does not contain nerve fibrils. It does not spread into lymphatics as scar is devoid of lymphatics. Induration is felt at the edge and base. There is marked fibrosis also.
- Once lesion spreads into adjacent normal skin, it can spread into regional lymph nodes behaving like squamous cell carcinoma.
- *Managed by* edge biopsy and wide local excision and grafting.

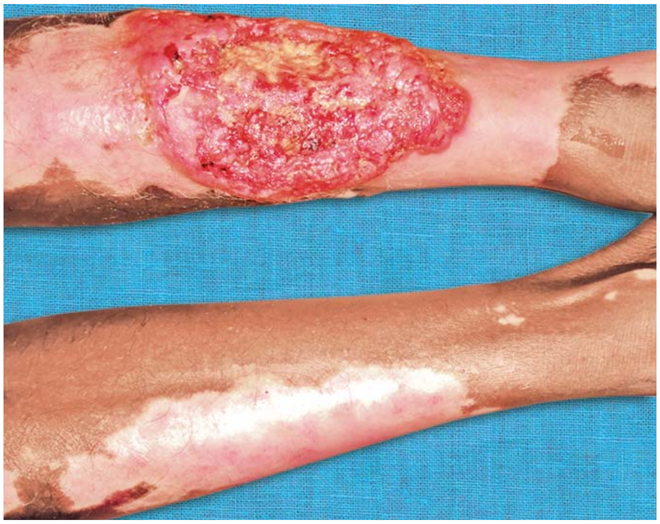

Fig. 1.69: Marjolin's ulcer in a chronic unstable scar (of long duration) in the leg. It does not spread through lymphatics.

RODENT ULCER

- It is ulcerative form of basal cell carcinoma which is common in face.
- Ulcer shows central area of dry scab with peripheral raised active and *beaded (pearly white)* edge. Often floor is pigmented. It erodes into deeper plane like soft tissues, cartilages and bones hence the name—*rodent ulcer*.
- As lymphatics are blocked early in the disease by large tumour cells, it does not spread to regional lymph nodes. Blood spread is absent. It is only *locally malignant.*
- It is common in face; rarely can it occur over tibia, external genitalia, mucocutaneous junction. It does not occur in mucosa.
- **Management:** Edge biopsy, CT scan of the part to see the depth, wide excision

MELANOTIC ULCER

- It is ulcerative form of melanoma. It can occur in skin as *de novo* or in a pre-existing mole. Ulcer is pigmented often with a halo around.
- Ulcer is rapidly growing, often with satellite nodules and '*in-transit*' lesions. It is very aggressive skin tumour arising from melanocytes.

- It spreads rapidly to regional lymph nodes which are pigmented. Blood spread to liver, lungs, brain, and bones is common. It can occur in mucosa, genitalia, and eye. It is a systemic malignant disease.
- **Management:** Excision biopsy (usually *incision biopsy is not done*), FNAC lymph node, US abdomen. *Treatment* is wide local excision, regional node block dissection, chemotherapy.

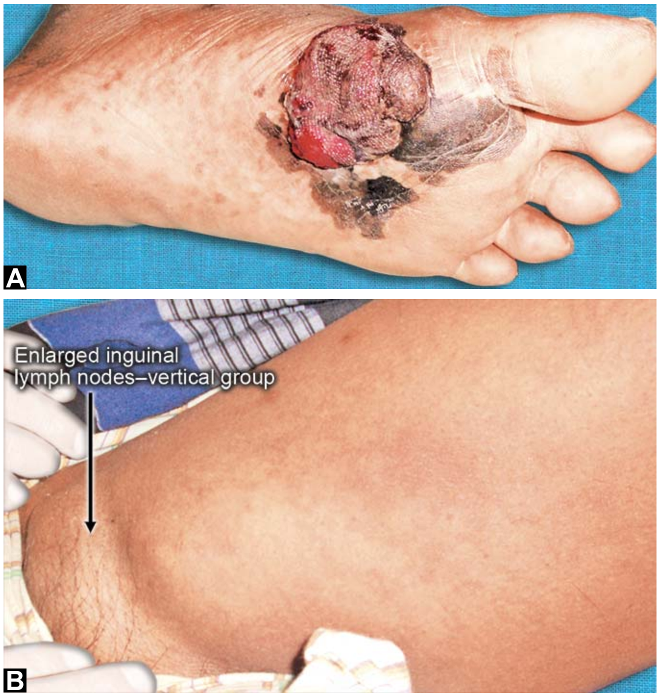

Figs 1.70A and B: Melanotic ulcer in the foot with enlarged inguinal lymph nodes.

DIABETIC ULCER

Causes

- Increased glucose in the tissue precipitates infection.
- Diabetic microangiopathy which affects microcirculation.
- Increased glycosylated haemoglobin decreases the oxygen dissociation.
- Increased glycosylated tissue protein decreases the oxygen utilization.
- Diabetic neuropathy involving all sensory, motor and autonomous components.
- Associated atherosclerosis.

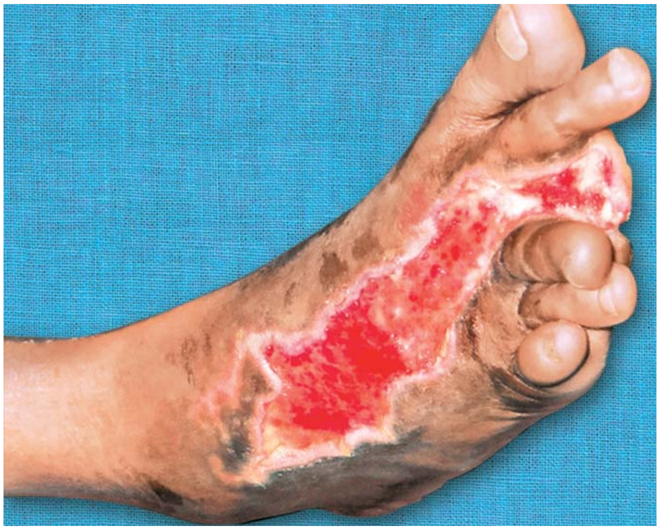

Fig. 1.71: Diabetic ulcer foot.

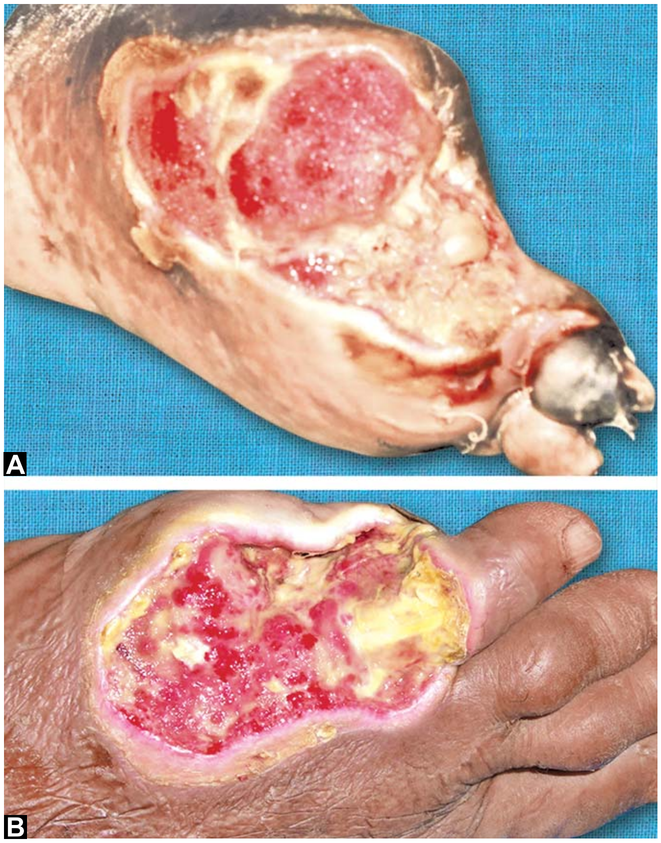

Figs 1.72A and B: Foot is the most common area for diabetic infective problems. It can cause abscess, ulcer, osteomyelitis, gangrene, septicaemia. Initially patient undergoes toe amputation but later eventually land with below knee or above knee amputation.

Sites

- Foot-plantar aspect—is the most common site.
- Leg.
- Upper limb, back, scrotum, perineum.
- Diabetic ulcer may be associated with ischaemia.
- Ulcer is usually spreading and deep.

Investigations

- Blood sugar both random and fasting.
- Urine ketone bodies.
- Discharge for culture and sensitivity.
- X-ray of the part to see osteomyelitis.
- Arterial Doppler of the limb; glycosylated haemoglobin estimation.

When you cease to dream, you cease to live.

Problems with diabetic ulcer

- Neuropathy, in foot—clawing of toes, hammer toe (due to intrinsic muscle paralysis)
- Multiple deeper abscesses; osteomyelitis of deeper bones are common
- Reduced leukocyte function; resistant infection; spreading cellulitis
- Arterial insufficiency
- Septicaemia; diabetic ketoacidosis
- Associated cardiac diseases like ischaemic heart disease

Treatment

- Control of diabetes using insulin.
- Antibiotics.
- Nutritional supplements.
- Regular cleaning, debridement, dressing.
- Once granulates, the ulcer is covered with skin graft or flap.
- Toe/foot/leg amputation.
- Microcellular rubber (MCR) shoes to prevent injuries; care of foot.

MELENEY'S ULCER (POSTOPERATIVE SYNERGISTIC GANGRENE)

- It is commonly seen in *postoperative wounds* in abdomen and chest wall like empyema drainage or after surgery for peritonitis.
- It is an acute rapidly spreading ulcer with destruction and deep burrowing of subcutaneous tissues.

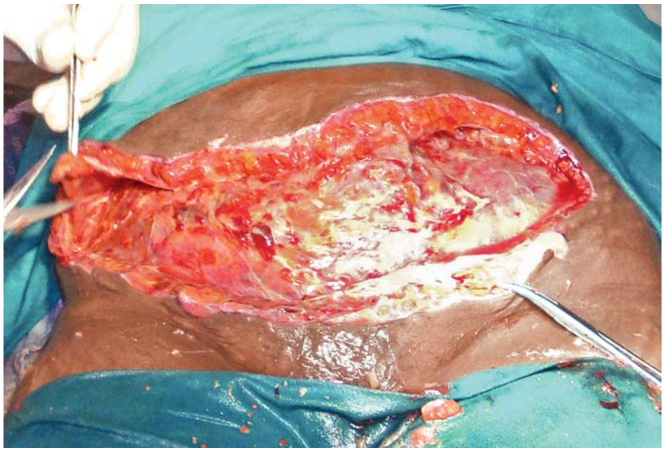

Fig. 1.73: Meleney's ulcer.

- Occasional symbiotic infection may develop in leg or hand. This can be as *de novo* association of ulcerative colitis or on a pre-existing venous ulcer.

Aetiology

- It is common in old age and immunosuppressed individuals and after surgery for infected cases.
- It is caused by microaerophillic streptococci, *Staphylococcus aureus* and anaerobes.

Sites

- It is common in abdomen and thorax. It begins in wound margin and spreads rapidly. It can also occur in other areas of skin.
- Infection is severe, often with endarteritis of the skin leading to ulcer and destruction.

Clinical Features

- Features of toxaemia.
- Spreading painful ulcer with discharge.
- Abundant granulation tissue with purple and red zones.

Management

- Random blood sugar is checked, if diabetic it has to be controlled.
- Antibiotics.
- Blood transfusion, critical care.
- Adequate excision of dead tissues until it bleeds.
- Once healthy granulation tissue is formed skin grafting is done.

LUPUS VULGARIS ('Lupus'—Wolf)

- It is *cutaneous tuberculosis* which occurs in young age group.
- Commonly seen on face, hand and forearm; starts as *typical* ***apple-jelly nodule*** with congestion of skin around. Eventually a superficial ulcer with *undermined edge* is formed.
- Glass slide pressed firmly on the diseased area to eliminate the surrounding hyperaemia causes clinically obvious *apple-jelly* appearence.

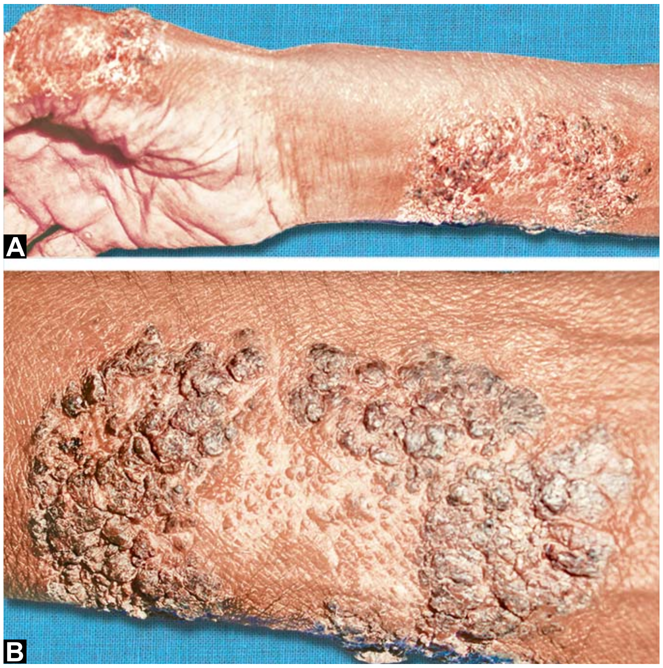

Figs 1.74A and B: Lupus vulgaris in forearm and thumb. Biopsy confirmed tuberculosis. Patient needs antituberculous drugs (*Courtesy:* Dr Ashfaque, DNB, Surgeon, KMC, Mangalore).

- The ulcer is active and destruction occurs at the periphery with healing takes place at the centre.
- Often lesion extends into nose and oral cavity involving the mucosa.
- Due to lymphatic obstruction facial oedema can occur.
- Long-standing lupus vulgaris can turn into *squamous cell carcinoma.*

Investigation

ESR, discharge study, biopsy, chest X-ray.

Treatment

a. Antituberculous drugs.
b. If complete healing does not occur, then ***excision and skin grafting*** is required.

TUBERCULOUS ULCER

It is due to *Mycobacterium tuberculosis*. It is usually due to cold abscess later forming ulcer in the neck, chest wall, axilla and groin. It can also be primary tuberculosis of the skin (commonly in face). Ulcer can be single or multiple; oval or rounded; with undermined edge (due to progression of disease outwards underneath and healing inwards by skin), painful and tender with caseating material on the floor. Ulcer is usually not deep. Regional lymph nodes may be enlarged matted, firm, and nontender.

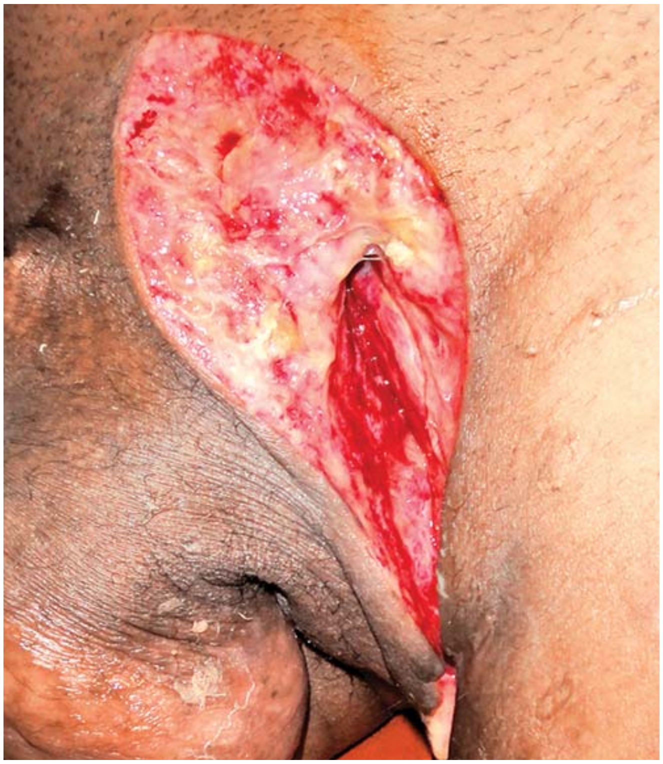

Fig. 1.75: Tuberculous ulcer is typically undermined.

Management: Discharge study for epithelioid cells (modified histiocytes), AFB; edge biopsy, antituberculous drugs.

BAZIN'S DISEASE (ERYTHROCYANOSIS FRIGIDA/ERYTHEMA INDURATUM)

(Pierre Bazin, 1850, Paris)

- It is localised area of fat necrosis with chronic ischaemia of ankle skin affecting exclusively adolescent girls. It may be due to tuberculosis. It is observed in girls with more / thick subcutaneous fat around ankle.
- Bluish pink leg which becomes bluish mottling in extreme cold season. On warming, skin turns bright red and painful which is typical due to hyperaemia. In these patients perforating arteries perfusing the skin around the ankle are small/ poor/not existing causing ischaemia of skin around ankle which becomes hyperaesthetic and sensitive for temperature alteration.
- Symmetrical, purple nodules develop in ankles and lower leg which later break down forming multiple, small, painful, superficial ulcers often with ankle oedema and pigment scars.
- Treatment is antitubercular drugs and lumbar sympathectomy.

TROPICAL ULCER

- It is endemic in monsoon hit humid tropics with repeated epidemics but sporadic in subtropics. Trauma or insect bite leads into infection exclusively in the lower part of the leg and foot.
- It is an acute ulcerative lesion of the skin observed in tropical regions like Africa, India and South America. It is associated with lower socioeconomic group, anaemia, and malnutrition and vitamin deficiency.
- It is commonly caused by *Fusobacterium fusiformis* (Vincent's organisms) and *Borrelia vincentii*.
- There are abrasions, redness, papule and pustule formation, acute regional lymphadenitis and severe pain.
- Pustule bursts in 3 days along with necrobiosis and phagedena causing a spreading painful ulcer with an undermined edge, brownish floor and serosanguineous discharge. Spreading stops in few weeks with ulcer persisting for many months to years. Eventually a chronic, large nonhealing/callous ulcer forms with persistent pain, profuse serosanguineous discharge, extremely unpleasant odour, long existing firmly adherent slough in the floor without any obvious constitutional symptoms. During healing it causes a slight pigmented, parchment like round scar.
- Often destruction is progressive without cessation (*phagedena*) to extend into entire soft tissues of foot and leg inviting amputation. ***Phagedena*** (Greek—to eat) is also seen in chancroid and cancrum oris.
- Occasionally squamous cell carcinoma can develop in it.
- **Treatment:** Improvement in nutrition, penicillin, metronidazole, Eusol dressing, skin grafting at a later date.

VENOUS ULCER (GRAVITATIONAL ULCER)

- It is common around ankle (*gaiter's zone*) due to chronic venous hypertension. It is due to *varicose veins* (long saphenous vein/short saphenous vein/perforators) or *post-phlebitic* limb.
- Post-phlebitic limb consists of veins that is been partially recanalised following deep venous thrombosis which causes increased venous pressure around ankle through perforators.
- Varicose veins are common in *females*. 50% of venous ulcer is due to varicose veins; 50% are due to post-phlebitic limb (previous DVT). Pain, discomfort, pigmentation, dermatitis, lipodermatosclerosis, ulceration, periostitis, ankle joint ankylosis, talipes equinovarus deformity and Marjolin's ulcer are the problems of varicose veins and later of venous ulcer.

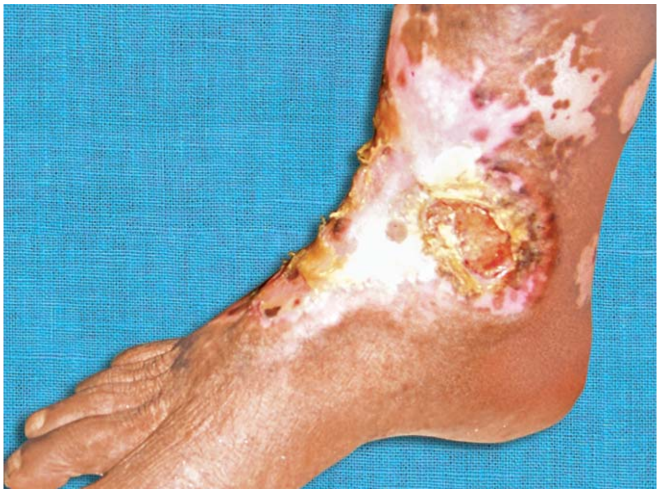

Fig. 1.76: Venous ulcer around ankle with skin changes over surrounding area. It is the most common site of venous ulcer.

- Ulcer is initially painful; but once chronicity develops it becomes painless. Ulcer is often vertically *oval*; commonly located on the *medial side*; occasionally on lateral side; often on both sides of the ankle; but *never* above the middle third of the leg. Floor is covered with pale or often without any granulation tissue. When well-granulated, edge is sloping. Induration and tenderness is seen often at the base of an ulcer.
- Inguinal lymph nodes (*vertical group*) are often enlarged. Ulcer often attains very large size which is nonhealing, indolent and callous.
- Ulcer heals on rest and treatment; but reforms again. Scarring is common due to repeated healing and recurrent ulcer formation. This *unstable scar* of long duration may lead into squamous cell carcinoma (*Marjolin's ulcer*).
- **Management**: Venous Doppler, regular dressing, skin grafting, specific treatment for varicose veins.

SYPHILITIC ULCER

- Nowadays it is a rare entity. It is caused by *Treponema pallidum* bacterium. It is a sexually transmitted disease. It is named as '*Syphilis*' after a shepherd named *Syphilus* who acquired the disease as was written in a poem by Francastorius of Verona. Many clinical lesions are observed in different stages of syphilis.
- *John Hunter* inoculated himself with syphilis organism to study the clinical features and effects. After 24 years of inoculation, he died from rupture of syphilitic aortic aneurysm at the age of 65.
- Genital chancre (*Hard chancre, Hunterian chanre*) is painless, hard, button like, indurated, nonbleeding ulcer; usually seen in corona or frenum of penis, often on lips, breasts and anal region; appears 4 weeks after initial infection in ***first stage*** of the disease (***primary syphilis***). Shotty, painless, firm, discrete groin lymph nodes may get enlarged along with genital chancre. Suppuration in these nodes will not occur. Extragenital chancres in lips and breasts show enlarged neck/axillary nodes which are inflamed, painful and also often may be matted.
- During ***second stage*** (***secondary syphilis***) white, thickened *mucous patches* appear commonly in the mouth like small, circular, superficial *snail track* ulcers. Also there appears raised, flat, hypertrophied, and warty like epithelium at mucocutaneous junctions (mouth, genitalia) called as *condyloma lata*. Generalised, shotty, hard, discrete, painless lymph nodes are palpable, epitrochlear and suboccipital nodes in particular are enlarged. *Epitrochlear nodes* are felt 1-2 cm above the medial epicondyle (It is also enlarged in non-Hodgkin's lymphoma/NHL). Iritis, arthritis, hepatitis (massive liver in syphilis is called as *hepar lobatum*), meningitis, syphilitic osteitis with *'ivory' sequestrum*, *coppery red skin rash*, *moth-eaten alopecia* are other features of second syphilis.
- ***In tertiary/late stage syphilis*** *gummatous ulcer* develops. It is deep, punched out, painless, nontender ulcer with *wash leather slough* in the floor, with *'silvery tissue paper'* like scar around and occurs over the subcutaneous bones like tibia, sternum, skull, palate or other area. It also can occur over the tongue, anterior aspect of the scrotum. It is due to delayed hypersensitivity reaction with endarteritis obliterans and vasculitis. Perforation of nasal septum/palate can occur. *Clutton's joint* and *Sabre tibia* are often seen. Lymph nodes are not affected in tertiary syphilis. Neurosyphilis (*tabes dorsalis*), *aneurysm of arch of aorta* are other features of tertiary syphilis.
- Tabes dorsalis presenting as generalized paralysis of insane is often called as ***late tertiary or quaternary*** syphilis.
- Long quiescent asymptomatic period from secondary to tertiary is called as ***latent syphilis***.
- Secondary syphilitic stage shows plenty of circulating *Treponema spirochaetes* in blood where as in tertiary stage spirochaetes are less or absent.

SOFT CHANCRE/SOFT SORE/DUCREY'S ULCER/CHANCROID/BUBO

- These multiple irregular genital ulcers appear 3 days after infection with *Haemophillus ducreyi* as a venereal disease.
- They are acute painful, tender, nonindurated ulcers. Floor shows yellowish slough with purulent discharge. Edge is oedematous and inflamed. Acute regional lymphadenitis with suppuration presenting as tender, soft or firm swelling is common. Such soft fluctuant inguinal swelling is termed as ***bubo***.
- *Treatment* is by drugs like co-trimoxazole, erythromycin, ciprofloxacillin, ceftriaxone; aspiration of bubo.

CLIMATIC BUBO/TROPICAL BUBO

- It is due to *lymphogranuloma inguinale*, a *venereal* spreading organism (LGV, *Chlamydia* type L1, 2, 3).
- In LGV, lesion of *primary genital stage* is small, painless and commonly unnoticed.
- Lesion *of secondary stage* develops in 2 weeks. In males inguinal lymph nodes; in females intrapelvic and pararectal nodes are involved. Suppuration of inguinal nodes eventually occurs leading into discharging sinuses. *Frei intradermal test* becomes positive in 6 weeks and remains positive for life time.
- *In tertiary stage,* eye, joint, meninges may get involved after many years. Repeated chronic inflammation, lymphatic blockage, scarring can cause rectal stricture and vulval elephantiasis (*esthiomene*) in females.
- *Treatment* is tetracycline for 3 weeks.

OTHER ULCERS

- Ulcers can occur, in various parts like over shin, legs, feet, face, chest wall, in various diseases like anaemia, polycythaemia, sickle cell disease, hereditary spherocytosis, leukemia, vasculitis, autoimmune diseases like rheumatoid arthritis, Paget's disease of bone (deep, nonmobile, fixed to bone; common in tibia), ulcerative colitis, etc.
- *Treponema pertenue* causing *Yaws* (Frambesia) can have multiple painless ulcers in leg and feet due to walking with bare foot (organism enters through abrasion) which heals spontaneously leaving a *tissue paper like scar*.
- Poor hygiene and dressings can cause multiple, small, red often scabbed *Staphylococcus aureus ulcers* on the skin over the leg and feet which is often recurrent and disturbing.

CLINICAL EXAMINATION OF AN ULCER

History

- Mode of onset
- Duration
- Pain—its time of onset, progress, severity
- Discharge from ulcer
- History suggestive of associated disease/treatment history

Local examination of an ulcer

Inspection

- Site of ulcer—arterial ulcer over the digits; venous ulcer over the malleoli; trophic ulcer over heel/pressure points
- Size of ulcer
- Shape of ulcer
- Number
- Margin whether regular/irregular/well-defined/ill-defined
- Edge of ulcer
- Floor of the ulcer—floor is the one what is seen. It rests on the base (Base is not seen; it is only felt). Red color in floor—healing ulcer; slough with pale/purulent discharge—nonhealing ulcer or tubercular; wash leather slough—syphilitic ulcer/proliferative and nodular floor—squamous cell carcinoma; pigmented—melanoma, pigmented basal cell carcinoma
- Discharge from ulcer bed—serous, serosanguinous, bloody, purulent; color of discharge—greenish in *Pseudomonas* infection
- Surrounding area to be examined for inflammation, oedema, eczema, scarring, pigmentation
- Inspection of the entire part/limb

Palpation

- Tenderness over edge, base and surrounding area
- Warmness over surrounding area
- Edge palpation for induration
- Palpation of base for induration/fixity
- Depth of ulcer—trophic ulcer is deep with bone as its base—often it is measured gently in mm
- Bleeding on palpation and touching
- Palpation for deeper structures and its relation to ulcer
- Surrounding skin and tibia/calcaneum/other related bones for thickening
- Examination of adjacent joint for mobility
- Examination of regional lymph nodes is essential—tenderness (acute infection), mobility, consistency may be hard (carcinoma metastasis)/ firm/soft and nontender (inflammatory); fixity (malignancy); ulceration or fungation (malignancy); sinus (nonspecific, tuberculosis or carcinoma)

Specific systems

- Examination of arterial pulse, peripherally in relation to ulcer and cardiovascular system for murmur
- Examination for varicose veins in standing position
- Examination of the abdomen for splenomegaly (sickle cell disease); hepatomegaly
- Examination of spine (gibbus, paraspinal spasm, movements) and neurological system like sensation and muscle power

Excellence is never granted to man but given as the reward of labour.

C. Sinus and Fistula

CHAPTER OUTLINE

- Sinus
- Fistula
- Median Mental Sinus
- Sequestrum
- Preauricular Sinus

SINUS

It is a blind track lined by granulation tissue leading from an epithelial surface into the surrounding tissues.

Sinus means ***"hollow"*** or ***"a bay" (Latin)***.

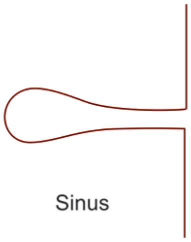

Fig. 1.77: Sinus.

Causes

Congenital: Preauricular sinus.
Acquired: Actinomycosis, tuberculosis, pilonidal sinus, chronic osteomyelitis, median mental sinus.

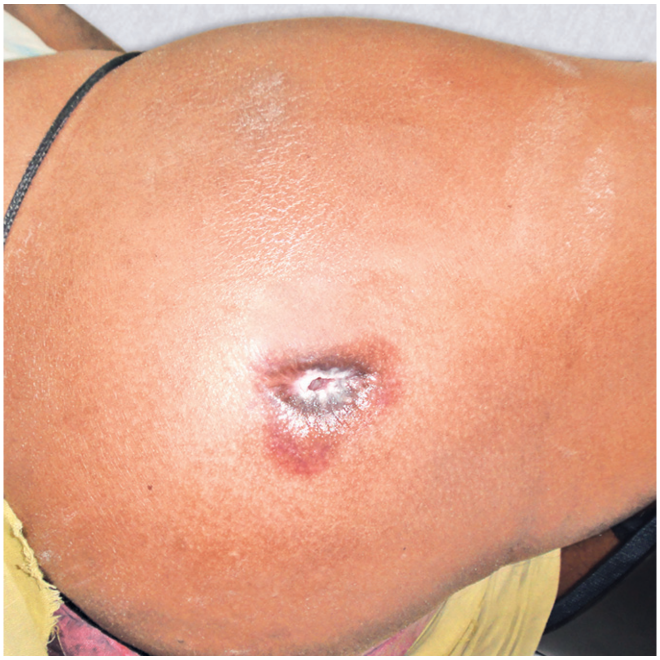

Fig. 1.78: Typical sinus in the thigh due to osteomyelitis of the greater trochanter.

FISTULA

It is an abnormal communication between the lumen of one viscus to another or the body surface or between the vessels. Fistula means "***flute"*** or ***"a pipe*** *or* ***tube***."

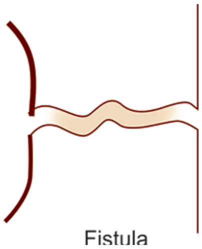

Fig. 1.79: Fistula.

Causes

Congenital:

- Branchial fistula.
- Tracheo-oesophageal fistula.
- Congenital AV fistula.
- Umbilical fistula (patent vitello-intestinal duct).

Acquired:

a. Traumatic:
   - Following surgery—intestinal fistulas (biliary, pancreatic, faecal).
   - Following instrumental delivery or difficult labour (vesicovaginal fistula, rectovaginal fistula, ureterovaginal fistula).

b. Inflammatory—intestinal actinomycosis, tuberculosis.

c. Malignancy—when the growth of one organ penetrates into the nearby organ (Rectovesical fistulas as in carcinoma rectum, vesicouterine fistulas as in uterine cancer).

External fistula	***Internal fistula***
❖ Orocutaneous	❖ Tracheo-oesophageal fistula
❖ Branchial fistula	❖ Cholecystoduodenal fistula
❖ Thyroglossal fistula	❖ Colovesical fistula
❖ Enterocutaneous fistula	❖ Rectovesical fistula
❖ Appendicular fistula	

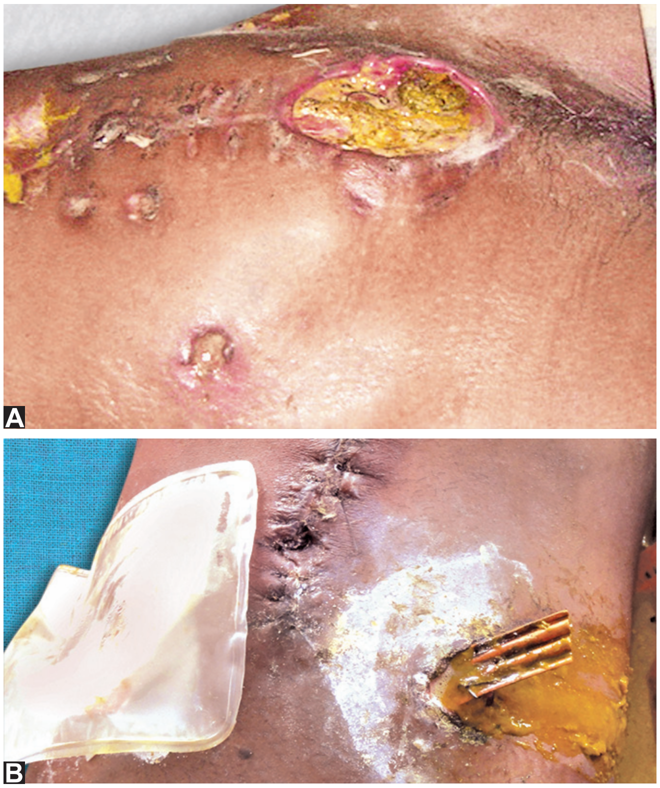

Figs 1.80A and B: Faecal fistula through laparotomy wounds. One is through main wound; other is through drain site. It often occurs after surgery for severe peritonitis, tuberculosis, malignancy.

Clinical Features

- Discharge from the opening of sinus.
- No floor.
- Raised indurated edge, indurated base, nonmobile.
- Often *sprouting granulation tissue over the sinus opening.*
- Bone thickening in osteomyelitis.
- Surrounding skin may be *erythematous* in inflammatory; *bluish* in tuberculosis; *excoriated* in faecal fistula; *pigmented* in chronic sinuses/fistulas.
- Discharge typical of the cause will be evident which will be obvious after *applying pressure* over surrounding area.
- Induration is a feature of all chronic fistulas *except* tuberculosis.
- *Thickening of the bone underneath* on palpation if sinus is adherent to bone or if there is osteomyelitis.
- Enlargement of *regional lymph nodes* will be evident.
- Sinus may be single or multiple.

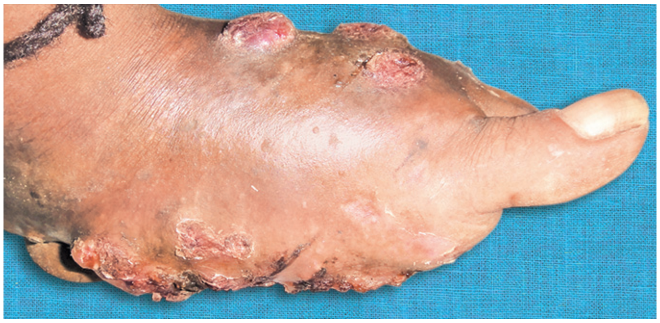

Fig. 1.81: Multiple discharging sinus foot. It is commonly due to mycetoma (Madura foot). It could also be due to tuberculosis, chronic pyogenic osteomyelitis or malignancy.

Causes of persistence of a sinus or fistula

- A foreign body or necrotic tissue underneath, e.g. suture, sequestrum
- Insufficient or nondependent drainage
- Persistent obstruction in the lumen, e.g. in faecal fistula, biliary fistula (distal obstruction)
- Lack of rest, persistent infection
- Wall become lined with epithelium or endothelium
- Dense fibrosis prevents contraction and healing
- Specific infections: Tuberculosis, actinomycosis
- Presence of malignant disease, post-irradiation

***Note*:**

The most common cause of sinus in neck is tuberculosis. Commonly it is tuberculous lymphadenitis. It shows *yellowish cheesy* discharge with bluish margin. Usually tuberculous sinus/ulcer do not show any induration.

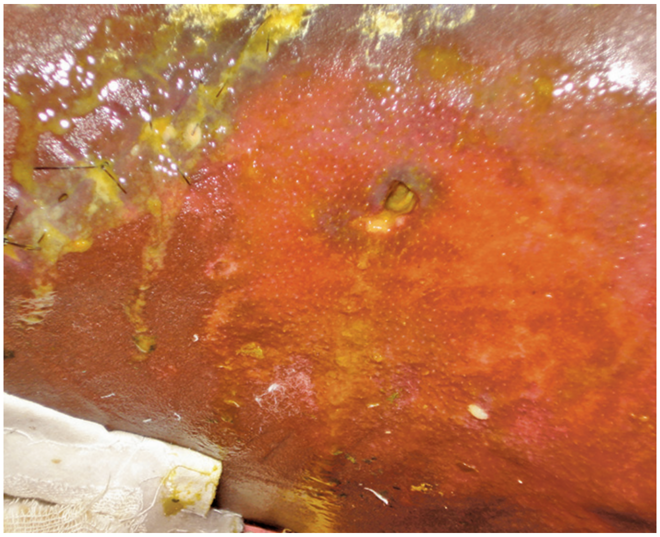

Fig. 1.82: Discharging abdominal faecal fistula showing extensive skin excoriation.

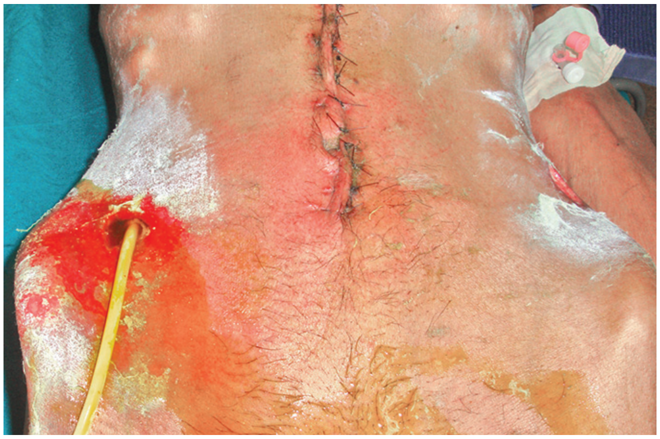

Fig. 1.83: Abdominal faecal fistula in a patient who underwent surgery for acute intestinal obstruction due to tuberculosis.

When a man looses his health he begins to take care of it.— **John Billings**

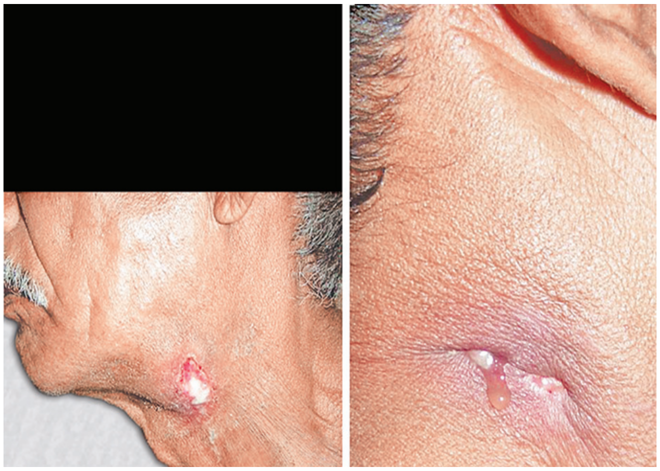

Figs 1.84A and B: Secondaries in neck causing discharging sinus. (A) and (B) in two different patients.

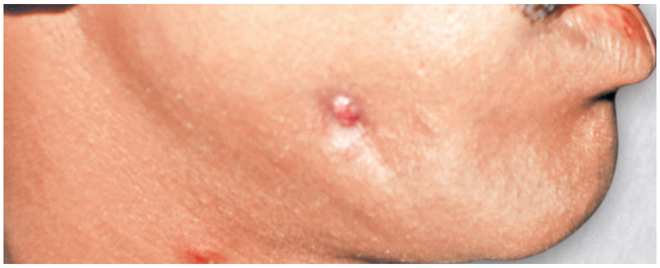

Fig. 1.85: Mandibular sinus. It is usually due to infected tooth causing osteomyelitis of mandible. It also could be due to tumour, trauma, actinomycosis and radiation. X-ray (orthopantomogram), study of discharge and biopsy are relevant investigations. Such fistula should be excised with extraction of the causative tooth.

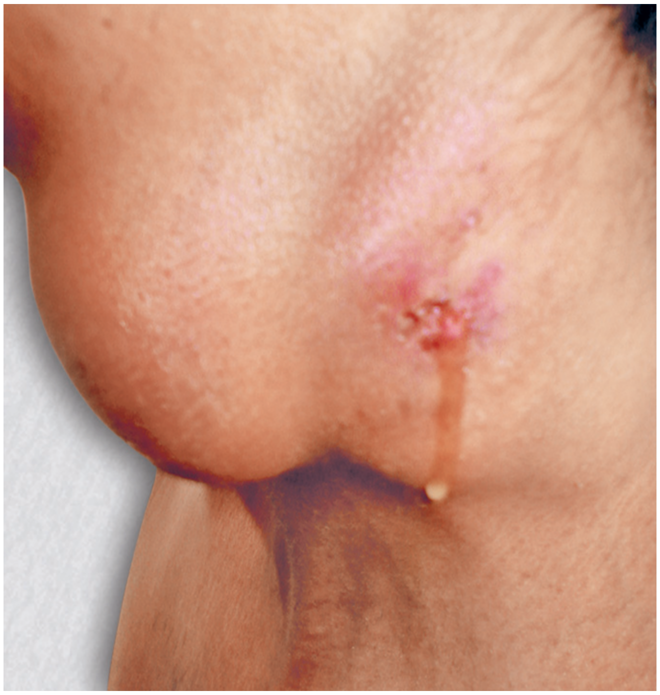

Fig. 1.86: Discharging sinus in the neck due to tuberculosis of lymph nodes with a cold abscess underneath.

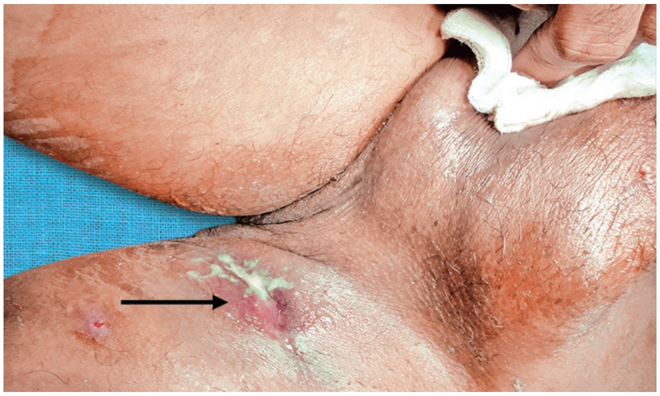

Fig. 1.87: Sinus opening in the thigh. Note there is no floor.

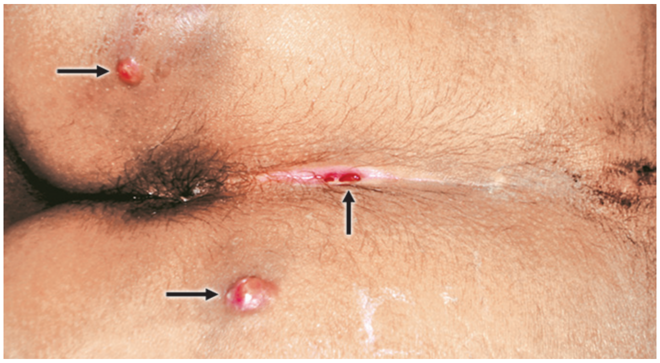

Fig. 1.88: Pilonidal sinus showing primary and secondary sinus.

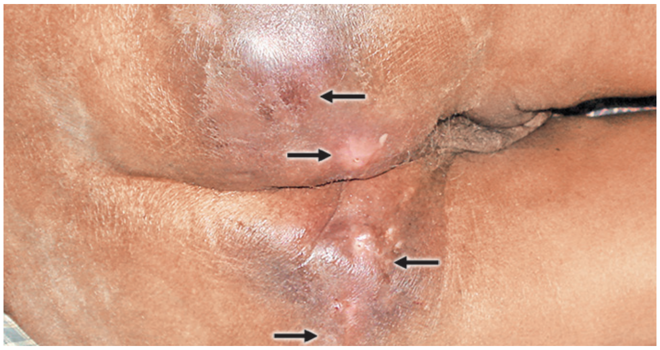

Fig. 1.89: Fistula in ano both sides.

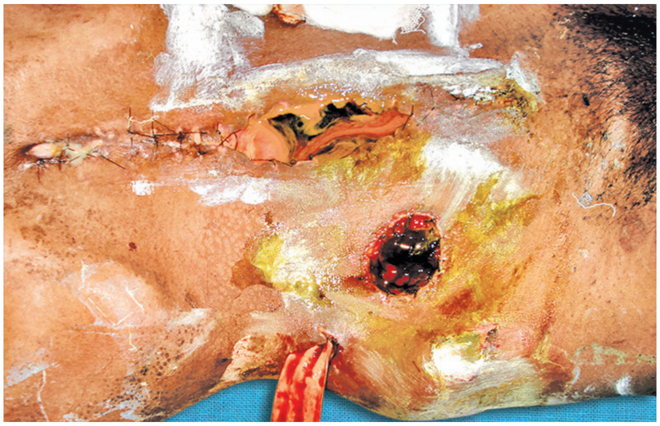

Fig. 1.90: Postoperative gastrointestinal fistula. Note the skin excoriation. It can be controlled by using zinc oxide cream local application.

Different discharges in a sinus/fistula

- *Purulent*—bacterial infection
- *Caseous*—tuberculous
- *Sulphur granules*—actinomycosis
- *Mucus*—branchial fistula
- *Saliva*—parotid fistula
- *Faeces*—faecal fistula
- *Bile*—biliary, duodenal fistula
- *Bone*—osteomyelitis sinus
- *Urine*—urinary fistula

Note:

- Streptococcal pus is *watery* with blood stain.
- Staphylococcal pus is *yellow* and creamy.
- Green or *greenish blue* pus is due to *Pseudomonas aeruginosa* infection.
- *Anchovy sauce pus* is seen amoebic liver abscess.
- Gas gangrene produces *sickly sweet odour—decayed apple* like.
- *E. coli* pus is usually *odourless.*
- Anaerobic bacteria and proteus vulgaris cause typical odour due to proteolysis. Bacteroids cause typical *over ripe Camembert cheese* like odour.
- Faecal fistula causes *foul smelling discharge with gas bubble* in it.
- Tuberculous sinus discharges *caseating cheesy* material.

Investigations

a. Fistulogram/sinusogram using ultrafluid lipiodol or water soluble iodine dye (Lipiodol is poppy seed oil containing 40% iodine).
b. Discharge for C/S, AFB, cytology, staining.
c. Biopsy from the edge for tuberculosis and malignancy.
d. Chest X-ray; X-ray of the part; MRI (most reliable) of the part.
e. ESR.
f. CT sinusogram.
g. Probing gently with care.
h. Digital examination of the rectum and proctoscopy in fistula in ano.

Treatment

- Treat the cause.
- ***Excision*** of sinus or fistulas. Always specimen should be sent for histology.
- Antibiotics, antitubercular drugs, rest, adequate drainage.

MEDIAN MENTAL SINUS

It is a chronic infective acquired condition wherein there is infection of roots of one or both lower incisor teeth forming root abscess which eventually tracks down between two halves of lower jaw in the midline presenting as discharging sinus on the point of chin at midline.

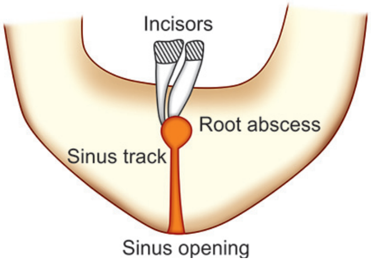

Fig. 1.91: Median mental sinus. Note the origin of the sinus from the root/roots of the lower incisor/incisors.

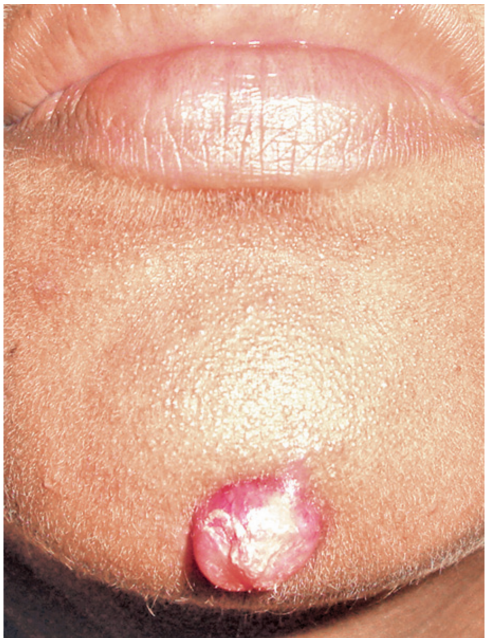

Fig. 1.92: Median mental sinus.

Clinical Features

- Usually painless discharging sinus in the midline on the point of chin.
- Often incisor infection may be revealed (in many patients clinically tooth looks normal even though root is infected invariably).
- It is often mistaken for infected sebaceous cyst.
- Osteomyelitis of the mandible is the possible complication.

Differential Diagnosis

- Infected sebaceous cyst.
- Tuberculous sinus.
- Osteomyelitis.

Investigations

- Dental X-ray is diagnostic (Plain X-ray mandible may not reveal the disease).
- Discharge study—C/S, cytology, AFB.

Be strong enough to face the world each day.
*Be weak enough to know you cannot do everything.***—Nabil Mourad**

Treatment

- Antibiotics, after doing discharge study (C/S).
- Lay opening and excision of the sinus track with *extraction of incisor tooth/teeth*.

SEQUESTRUM

- Sequestrum is dead bone *in situ*.
- It can be pyogenic, tubercular (feathery), *Salmonella* (granular), syphilitic (ivory), tubular and ring (in amputation stump).
- It can be *unformed*—means separation between sequestrum and adjacent normal bone has not occurred or *formed*—means there is proper adequate separation between normal bone and sequestrum by forming granulation tissue. Radiologically formed sequestrum shows clear lucent area/zone of demarcation.
- Sequestrum is denser because of the absence of decalcification in the dead bone as there is no blood supply (*dead bone is dense bone*).
- Sequestrum should be *formed* prior to surgical intervention—sequestrectomy and saucerisation.

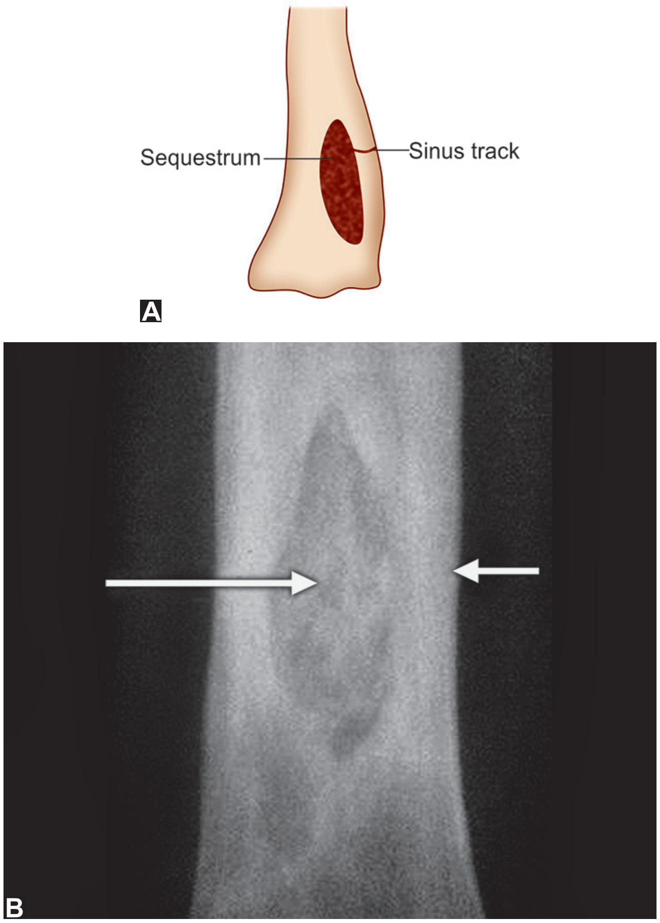

Figs 1.93 A and B: Diagram and X-ray showing osteomyelitis with sequestrum and sinus. Sequestrum is dead bone *in situ*.

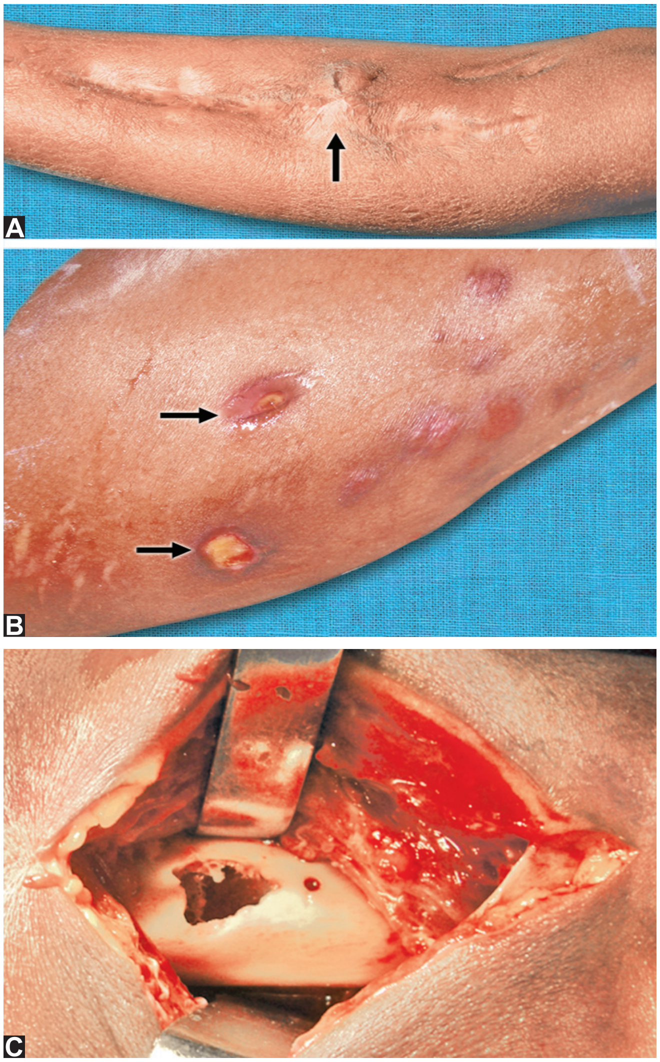

Figs 1.94A to C: Osteomyelitis patient with scar. Also showing on table photo of sequestrectomy and saucerisation.

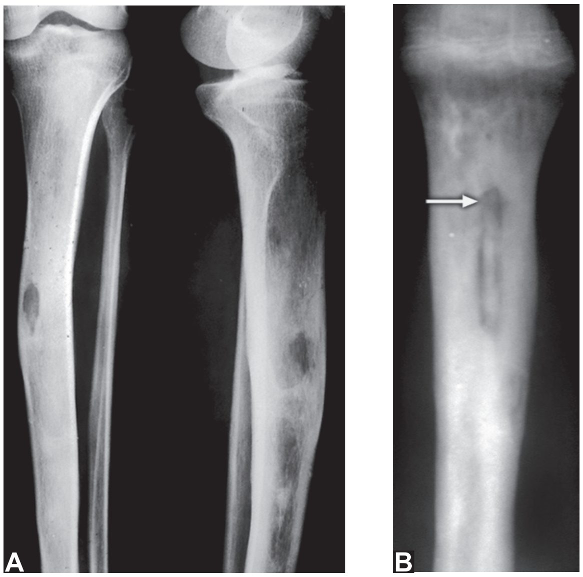

Figs 1.95A and B: X-ray pictures showing features of osteomyelitis with sequestrum, sinus, cavity.

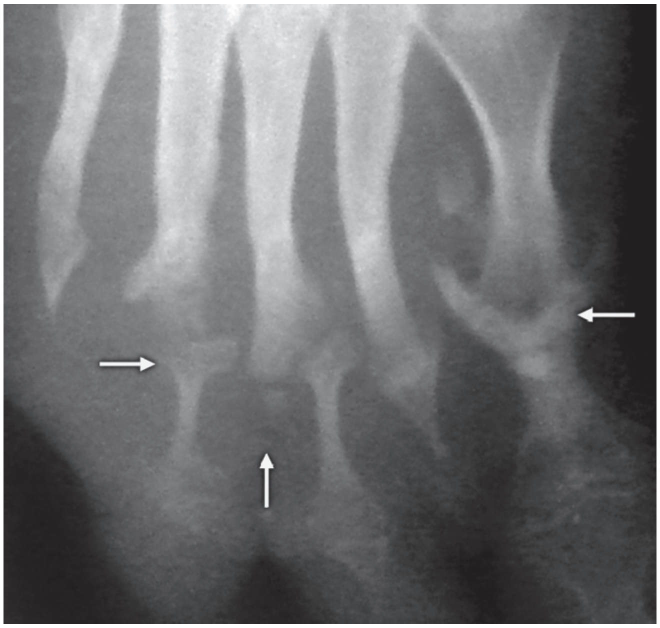

Fig. 1.96: X-ray of foot showing osteomyelitis of metatarsal bones. Metatarsal osteomyelitis is commonly seen in diabetes mellitus and trauma.

PREAURICULAR SINUS

It is a *congenital entity* occurring due to imperfect fusion of the six tubercles which form ear cartilage. Sinus opening may be seen at the root of the helix or on the tragus. Track is quiet deep running backwards, slightly upwards towards the helix. It usually ends *blindly*. Outer opening of the sinus often closed causing formation of a cystic swelling which contains fluid which is often infected. Preauricular sinus in *no instance* will communicate with the external auditory meatus. Bursting of this swelling leads into formation of ulcer like lesion. It can be unilateral or bilateral. *Occasionally* multiple sinuses are seen. Opening of the sinus occurs in a *small triangular area* in front of the ear at the level of the tragus. *Scarring* is common around the opening due to repeated infection.

Features

- It is seen since childhood.
- Often swelling appears and apparently disappears repeatedly.
- Pain and discharge is common.
- It causes a cosmetic problem in young individual.
- Differential diagnosis is tuberculosis.
- Discharge study, ESR, sinusogram to assess the track is needed. MR sinusogram is beneficial.

Treatment

Excision under general anaesthesia with removal of entire track is essential. If track is not removed properly recurrence will occur.

D. Infectious Diseases

> *But however secure and well-regulated civilized life may become, bacteria, protozoa, viruses, infected fleas, lice, ticks, mosquitoes, and bedbugs will always lurk in the shadows ready to pounce when neglect, poverty, famine, or war lets down the defenses.*
>
> **—Hans Zinsser,** 1934

CHAPTER OUTLINE

- Surgical Infection
- Cellulitis
- Erysipelas
- Lymphangitis
- Abscess
- Bacteraemia
- Septicaemia
- Pyaemia
- Metastatic and Pyaemic Abscess
- Boil
- Hidradenitis Suppurativa
- Carbuncle
- Pott's Puffy Tumour
- Pyogenic Granuloma
- Impetigo
- Scrum Pox
- Tetanus
- Gas Gangrene
- Tuberculosis
- Leprosy
- Syphilis
- Actinomycosis
- Madura Foot
- Rabies
- Anthrax
- Nosocomial and Opportunistic Infections
- HIV Infection and AIDS
- Necrotising Fasciitis
- Acute Pyomyositis
- Surgical Site Infection

SURGICAL INFECTION

- Surgical infection is major problem in surgical practice. Asepsis (prevention of entry of organisms) and antisepsis (killing of the bacteria in the skin or tissues) has made a difference in surgical practice. Epithelial surfaces act as mechanical barrier and phagocytes, antibodies; complements, macrophages, leukocytes, opsonins, etc. act as protective mechanisms.
- Malnutrition, diabetes mellitus, obesity, uraemia, jaundice, malignancy, immunosuppression, radiotherapy, chemotherapy, HIV, ischaemia, foreign body, haematoma are the risk factors for surgical infections. Virulence of organisms, blood supply, body immunity and support of antibiotics are the decisive factors in proper response to control infection.
- Surgical infection can be *superficial surgical site infection* in the wound or *deep surgical site infection* in deeper fascio-muscular layers or *organ space infection* like abdomen/thoracic cavity, etc.
- Health care associated infection occurs after hospital admission in intensive unit/postoperative ward, etc.
- *Southampton wound grading system for healing and infection:*
    - Grade 0 is normal healing;
    - Grade 1 is with bruising/mild erythema;
    - Grade 2 is severe erythema with other features of inflammation at or around wound;
    - Grade 3 is serous or bloody discharge;
    - Grade 4 is presence of pus or deep infection or tissue breakdown or significant haematoma.

ASEPSIS wound score system is used to assess the wound infection.

ASEPSIS wound scoring

Parameters	*Score*
• **A**dditional treatment	
1. Antibiotics	10
2. Pus drainage	05
3. Wound debridement	10
• **S**erous discharge—for 5 days of first 7 days of wound infection	0-5 daily
• **E**rythema—for 5 days of first 7 days of wound infection	0-5 daily
• **P**urulent fluid—for 5 days of first 7 days of wound infection	0-10 daily
• **S**eparation of deep tissues—for 5 days of first 7 days of wound infection	0-10 daily
• **I**solation of bacteria	10
• **S**tay in the hospital (in-patient) more than 14 days due to infection	05

CELLULITIS

- It is ***spreading inflammation*** *of subcutaneous and fascial planes.*
- Infection may follow a small scratch or wound or incision or insect/snake/scorpion bite.

Causative Agents

- Commonly due to *Streptococcus pyogenes* and other Gram +ve organisms. Release of streptokinase and hyaluronidase cause spread of infection.
- Often Gram –ve organisms like *Klebsiella, Pseudomonas, E. coli* are also involved (usually Gram –ve organisms cause secondary infection).
- It can be superficial or deep. More common superficial type is easier to diagnose.
- It is common in diabetics, immunosuppressed people and old age.
- It is common in face, lower limb, upper limb and scrotum.
- Cellulitis occurring in children is never primary but secondary to an underlying bone infection—*Morison's aphorism* (James Morison, 1939—Surgeon, Durham, Newcastle).

Sequelae

- Infection can get localised to form ***pyogenic abscess***.
- Infection can spread to cause ***bacteraemia, septicaemia, pyaemia***.
- Often infection can lead to ***local gangrene.***
- Extensive necrosis of skin and subcutaneous tissue—*necrotizing fasciitis*.

Clinical Features

- Fever, toxicity (tachycardia, hypotension).
- Swelling is diffuse and spreading in nature.
- Pain and tenderness, red, shiny area with stretched warm skin.
- Cellulitis will progress rapidly in diabetic and immunosuppressed individuals.
- Tender regional lymph nodes may be palpable which signify severity of the infection.
- No edge; no pus; no fluctuation; no limit.

Management

- Elevation of limb or part to reduce oedema so as to increase the circulation
- Antibiotics—penicillins, cephalosporins.
- Dressing (often glycerine dressing is used as it reduces the oedema because of its hygroscopic action glycerine magnesium sulphate dressing).
- Bandaging.

Cellulitis in Special Areas

Orbital Cellulitis

Cellulitis in orbit causes *proptosis,* leading to impairment of ocular movements and blindness. It can spread through ophthalmic veins into cavernous sinus causing *cavernous sinus thrombosis.* It requires admission and immediate aggressive treatment with higher generation antibiotics (Penicillins, cephalosporins).

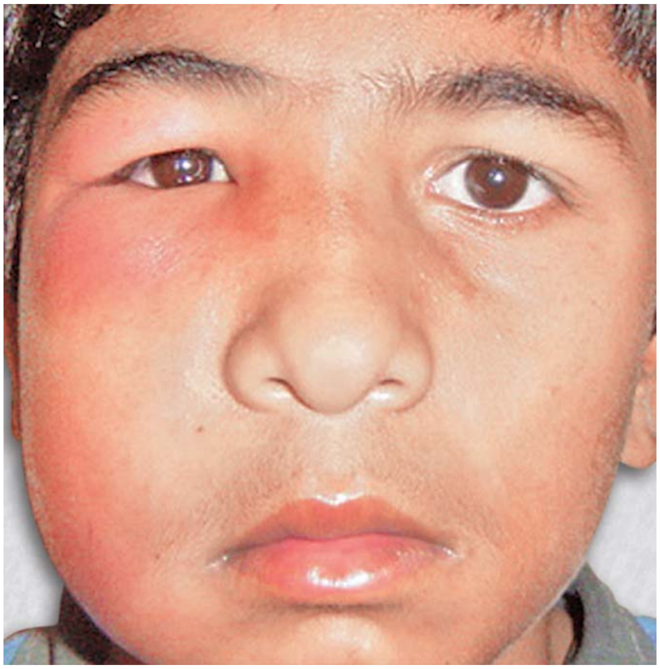

Fig. 1.97: Cellulitis face. Note the oedema of the face and eyelids.

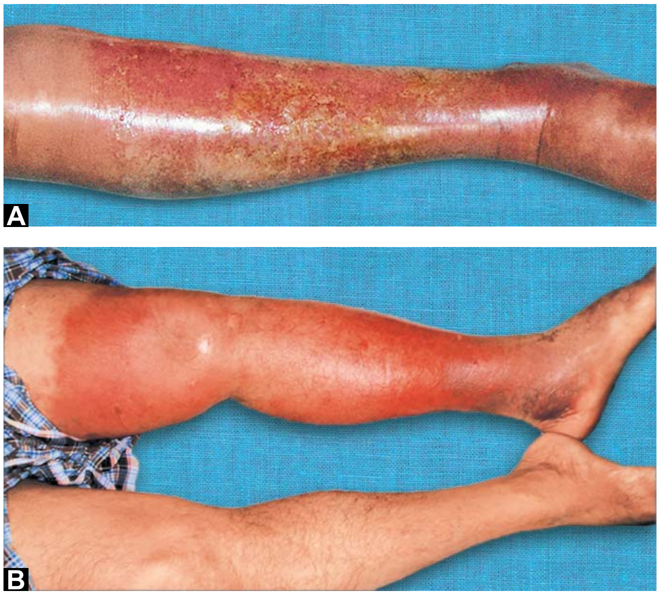

Figs 1.98A and B

*Oedema gives rise to soft pitting, while if pus present, induration can always be felt.***—Allan B Kanavel**

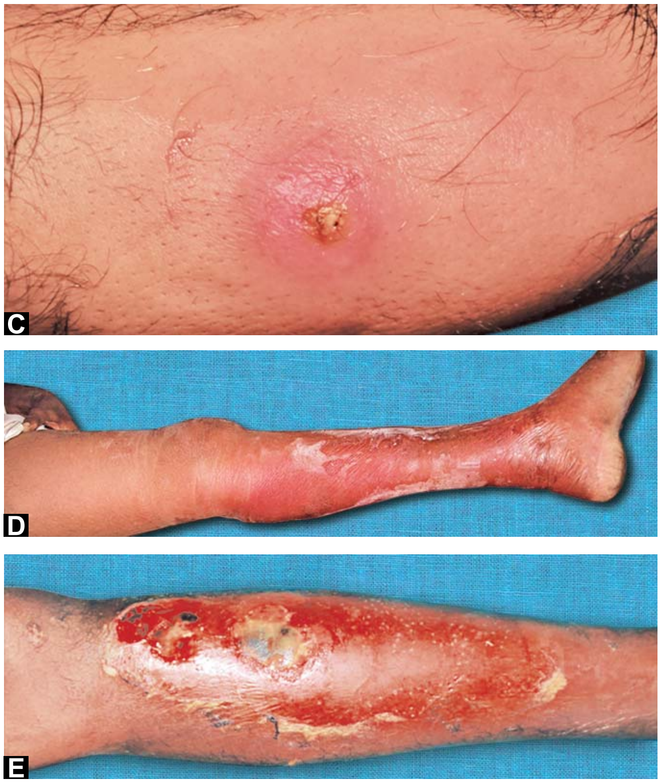

Figs 1.98C to E

Figs 1.98A to E: Note the cellulitis in different patients. It is common in lower limbs. There is no edge without any formed pus. It should never be incised with certain exceptions like Ludwig's angina. It is treated by antibiotics. It can cause bacteraemia/septicaemia.

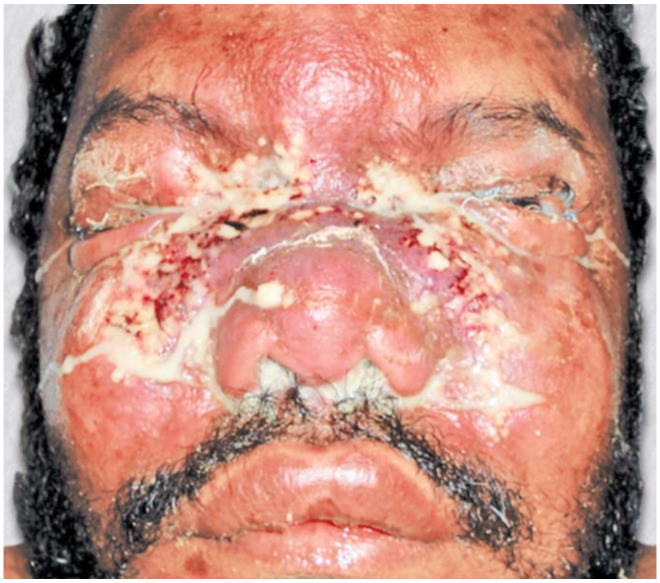

Fig. 1.99: Severe cellulitis face involving nose, eyelids and facial skin. It can cause septicaemia, cavernous sinus thrombosis or intracranial spread of infection. This area is called as *dangerous zone* in face. It needs hospitalisation for treatment.

Ludwig's Angina

It is cellulitis of upper part of the neck involving submandibular region and floor of the mouth along the fascial planes It may be precipitated by tooth extraction, oral cancer, submandibular gland infection, diabetes mellitus, chemotherapy, etc.

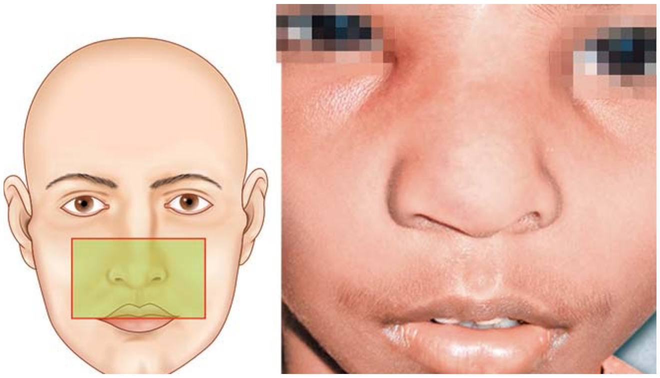

Fig. 1.100: Dangerous area of face—area of upper lip and lower part of nose. Infection from this area spreads through deep facial vein → pterygoid plexus → communicating vein → cavernous sinus causing its life-threatening thrombosis.

Clinical Features

- Diffuse swelling, redness, tenderness and induration in the floor of the mouth and submandibular region.
- Difficulty in opening of the mouth (***Trismus***).
- Toxic features like fever, tachycardia and tachypnoea.
- ***Severe laryngeal oedema*** (presents with respiratory distress, stridor and cyanosis).
- Dysphagia and *putrid halitosis*.

Complications

- ***Septicaemia.***
- Spread of infection into the parapharyngeal space leads to ***thrombosis*** of internal jugular vein which may extend above into the sigmoid sinus which may be fatal.

Treatment

- Antibiotics—intravenous.
- Early surgical intervention (decompression) is required. Under GA, horizontal incision is placed in submandibular region extending on both sides, deepened to include deep fascia. ***Mylohyoid*** muscles on both sides are cut. It releases the tension, preventing laryngeal oedema and so further spread of infection is prevented. Antibiotics should be continued.
- Incised wound is closed by ***secondary suturing*** once infection is controlled completely (not earlier).

ERYSIPELAS

It is a **spreading inflammation** of the skin and subcutaneous tissues due to infection caused by *Streptococcus pyogenes*. There will be always cutaneous lymphangitis with development of ***rose pink rash*** with cutaneous lymphatic oedema. Vesicles which form eventually will rupture to cause serous discharge.

Sites

- *Orbit, face and ear lobule*—most common.
- Hands and scrotum.
- Umbilicus in infants.
- Decubitus ulcer of lower limb occasionally.

Clinical Features

- ***Toxaemia*** is always a feature.
- ***Rash*** is fast spreading and blanches on pressure.
- ***Rash*** *is raised with sharp margin.*
- Redness becomes brown and later yellow with vesicles.
- ***Discharge*** *is serous (In cellulitis discharge is purulent).*
- In the face and orbit it causes severe oedema.
- ***Milian's ear sign*** is a clinical sign used to differentiate erysipelas from cellulitis wherein ear lobule is spared. Skin of ear lobule is adherent to the subcutaneus tissue and so cellulitis cannot occur. Erysipelas being a cutaneous condition can spread into the ear lobule (Gaston Milian—dermatologist Paris, 1945).
- Disease is common in poorly hygienic debilitated individuals.
- Septicaemia, localised cutaneous and subcutaneous gangrene are the dangerous problems.
- Lymphoedema of face and eyelids can occur later due to lymphatic fibrosis.

Treatment is with ***penicillins.***

ERYSIPELOID DISEASE

- Also called as *'Fish handler's disease.*
- Occurs following any cuts or scratches.
- It has features of both erysipelas and cellulitis.
- It is self-limiting with relatively mild symptoms.

LYMPHANGITIS

- It is an acute *nonsuppurative* infection and *spreading* inflammation of lymphatics of skin and subcutaneous tissues due to *beta haemolytic streptococci*, staphylococci, clostridial organisms. It is commonly associated with cellulitis. Erysipelas is a type of lymphangitis.
- In endemic areas, filariasis is the most common cause (coastal India). It is caused by *Wuchereria bancrofti*. It is transmitted through bites of Culex mosquito. Microfilaria reaches the lymph node forming adult worm which blocks the lymph node causing obstruction, fibrosis and lymphangitis.
- Usually infection occurs following a small trauma in bacterial cause. Rapidly fever and redness at the area develops.

Features

- Streaky redness which is spreading is typical. On pressure area *blanches*; on release redness *reappears*.
- Oedema of the part, palpable tender *regional lymph nodes* are obvious.
- Fever, tachycardia, features of toxaemia.
- Groin lymph nodes are enlarged and tender in *lower limb* lymphangitis. In *upper limb,* as lymphatics are mainly located on the dorsum of hand, oedema and redness develops on the dorsum. Infection in thumb and index finger causes palpable tender axillary nodes; in little and ring finger causes first tender palpable epitrochlear nodes to appear; infection in middle finger causes first deltopectoral nodes to enlarge.
- Regional lymph nodes (only) may eventually suppurate to form an *abscess*.
- *Toxaemia, septicaemia* may occur. Rapidity may be more in diabetics and immunosuppressed.
- *Chronic lymphangitis* occurs due to repeated attacks of acute recurrent lymphangitis leading into acquired *lymphoedema*.

Management

- Antibiotics like penicillin, cloxacillin.
- Elevation, rest, glycerine magnesium sulpha dressing.
- Management of toxaemia or septicaemia with critical care.

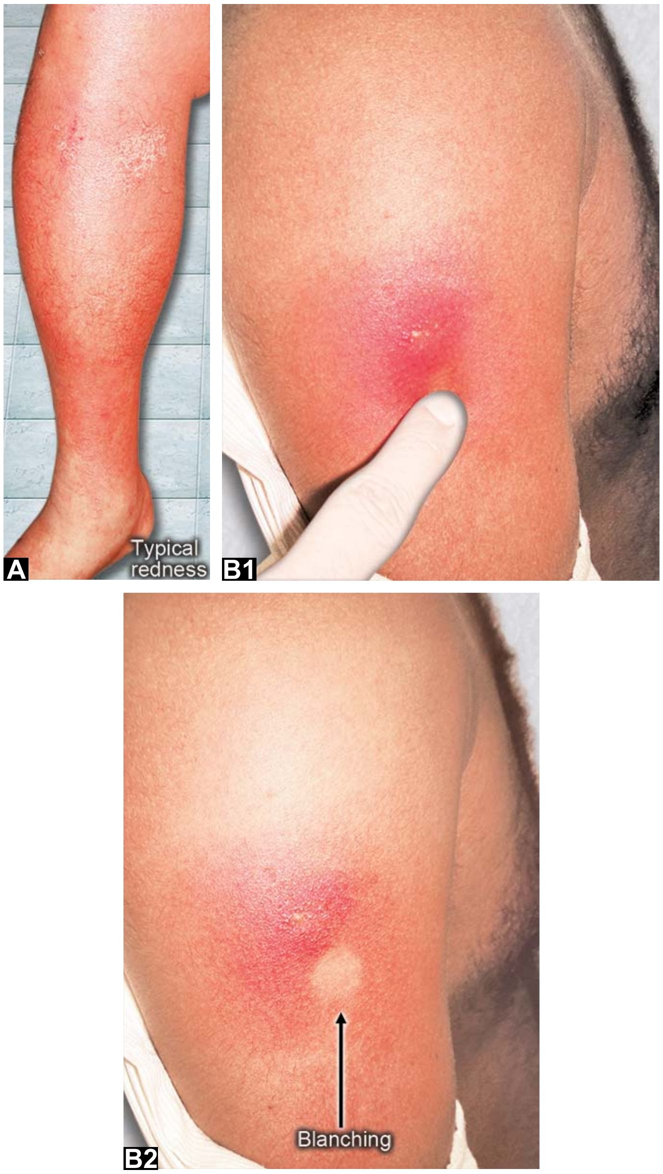

Figs 1.101A and B

Hope puts our feet on the path when our eyes cannot see it.

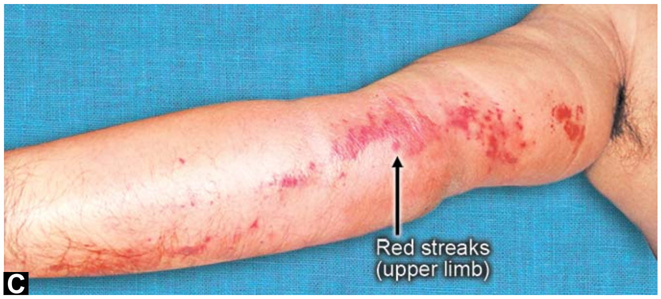

Figs 1.101C

Figs 1.101 A to C: (A) Lymphangitis of lower limb showing redness. (B1) Lymphangitis of upper limb; (B2) Blanching is seen immediately after pressure; (C) Typical red streaks in an upper limb lymphangitis.

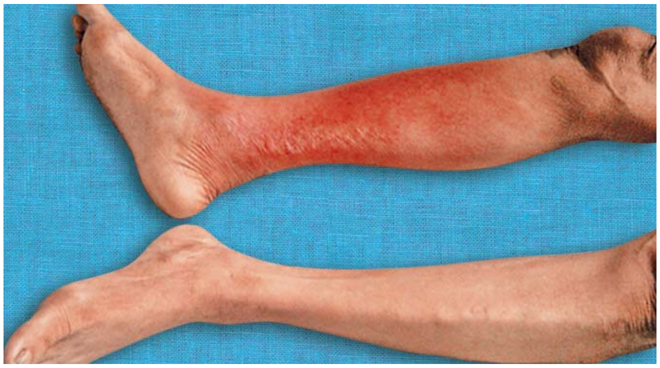

Fig. 1.102: Lymphangitis of one leg. Note the typical streaky redness compare to normal side.

ABSCESS

> *If you examine a swelling of pus in any limb of a man and you find it (with) its head raised and it is enclosed and it is rounded, you shall say concerning it: "a swelling of pus, an illness which will be treated by me with the knife treatment. There is something in it like mucus. Something comes forth like wax. It makes a pocket. If anything (remains) in its pocket, it recurs."*
>
> —**(Anonymous),** circa 1500 BC

Types

- Pyogenic abscess
- Pyaemic abscess
- Metastatic abscess
- Cold abscess due to chronic infection like tuberculosis

Pyogenic Abscess

It is a ***localised collection*** of pus in a cavity lined by granulation tissue, covered by pyogenic membrane. It contains pus in loculi. Pus contains dead WBC's, multiplying bacteria, toxins and necrotic material.

- Protein exudation causes fibrin deposition and formation of pyogenic membrane.
- Macrophages and polymorphs release lysosomal enzymes which cause liquefaction of tissues leading into pus formation.
- Toxins and enzymes released causes tissue destruction and pus formation.

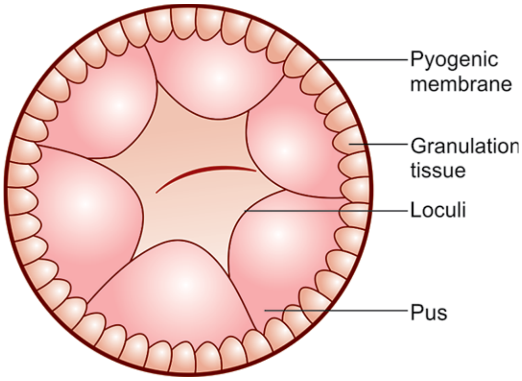

Fig. 1.103: Pyogenic abscess—parts.

Mode of Infection

- Direct
- Haematogenous
- Lymphatics
- Extension from adjacent tissues

Bacteria Causing Abscess

- *Staphylococcus aureus.*
- *Streptococcus pyogenes.*
- Gram-negative bacteria (*E. coli*, *Pseudomonas, Klebsiella*).
- Anaerobes.

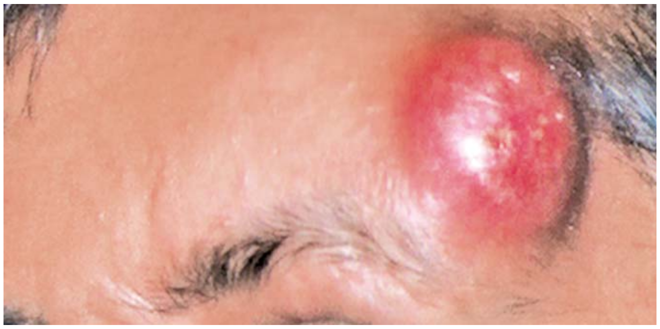

Fig. 1.104: Abscess forehead. Patient is diabetic and has had trauma to the site.

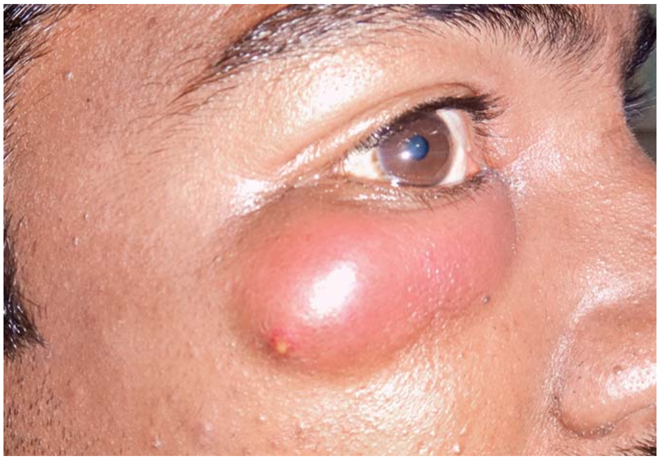

Fig. 1.105: Abscess in the face. Note the localisation and redness.

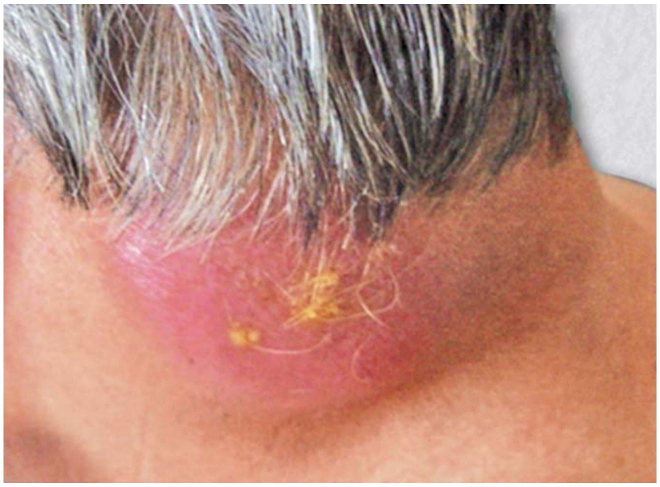

Fig. 1.106: Abscess in the nape of the neck—suboccipital region. Note the redness and visible pus.

Factors precipitating abscess formation

- General condition of the patient: Nutrition, anaemia, age of the patient
- Associated diseases: Diabetes, HIV, immunosuppression
- Type and virulence of the organisms
- Trauma, haematoma, road traffic accidents

Clinical Features

- Fever often with chills and rigors.
- Localised swelling which is smooth, soft and fluctuant.
- Visible (pointing) pus.
- Throbbing pain and pointing tenderness.
- Brawny induration around.
- Redness and warmth with restricted movement around a joint.
- *Rubor* (redness); *dolor* (pain); *calor* (warmness); *tumour* (swelling) and *functiolaesa* (loss of localised and adjacent tissue/joint function) are quiet obvious.

(Commonly cellulitis occurs first which eventually gets localised to form an abscess.)

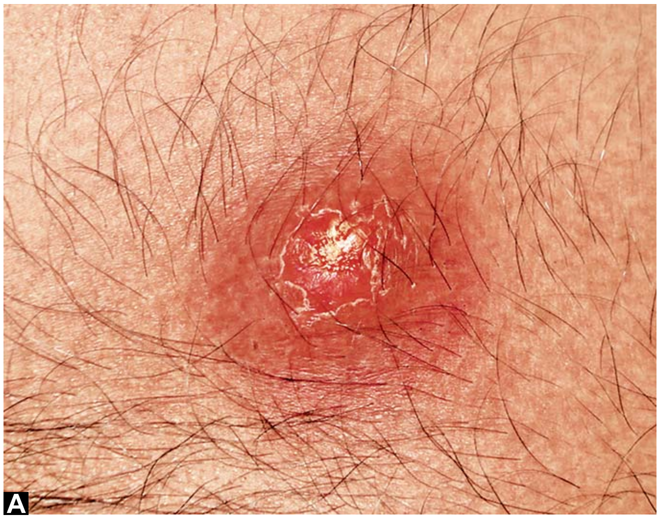

Fig. 1.107A

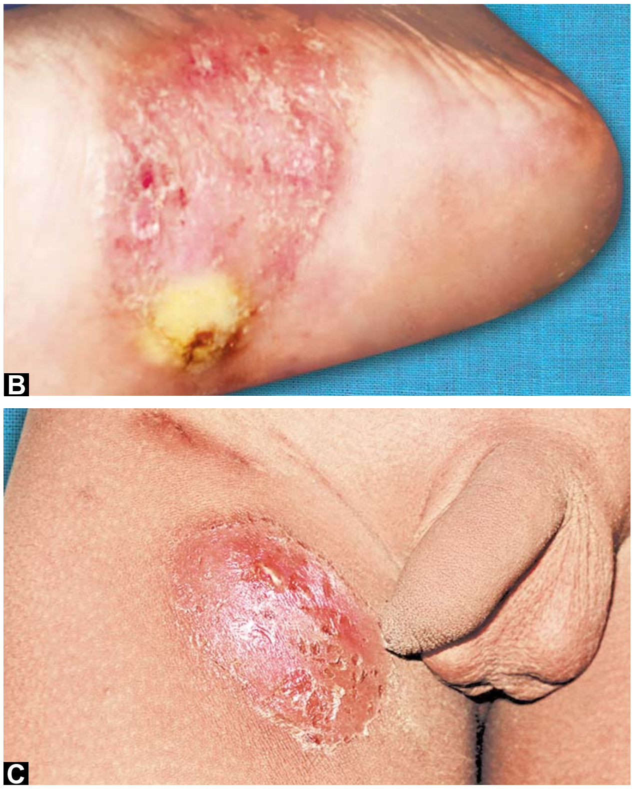

Figs 1.107B and C

Figs 1.107A to C: Pyogenic abscess in different locations in chest wall; in plantar aspect of the foot and in the thigh. Note the redness and visible pus.

Visible (pointing) pus, tenderness, fluctuation are the features of formed abscess.

Cellulitis	*Pyogenic abscess*
• Diffused—no edge	• Well-localised with clear edge
• Pus not formed—non-suppurative initially	• Formed pus
• Not fluctuant	• Fluctuant
• Spreading—SIRS can occur	
• Blood culture may be negative	• Culture of pus is usually positive
• Never incise; if done danger of bacteraemia, antibiotics, elevation, dressing—are treatment	• Drainage is essential

Sites of Abscess

a. *External Sites*

- Fingers and hand.
- Neck.
- Axilla.
- Breast.
- Foot, thigh—here it is deeply situated with brawny induration.
- Ischiorectal and perianal region.
- Abdominal wall.
- Dental abscess, tonsillar abscess and other abscesses in the oral cavity.

Happy is he who has no serious consequences of his erroneous diagnosis to regret.—**Fredrick H Marsh**

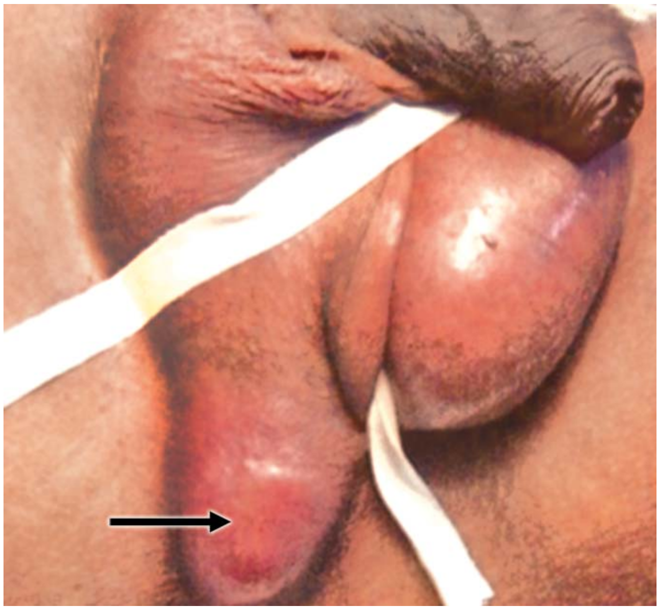

Fig. 1.108: Perineal abscess.

b. *Internal Abscess*

- Abdominal: Subphrenic, pelvic, paracolic, amoebic liver abscess, pyogenic abscess of liver, splenic abscess, pancreatic abscess.
- Perinephric abscess.
- Retroperitoneal abscess.
- Lung abscess.
- Brain abscess.
- Retropharyngeal abscess.

Investigations

- Total count is increased.
- Urine sugar and blood sugar is done to rule out diabetes.
- USG of the part or abdomen or other region is done when required.
- Chest X-ray in case of lung abscess.
- Gallium isotope scan is *very useful*.
- CT scan or MRI is done in cases of brain and thoracic abscess.
- Investigations, relevant to specific types: Liver function tests, PO_2 and PCO_2 estimation, blood culture.

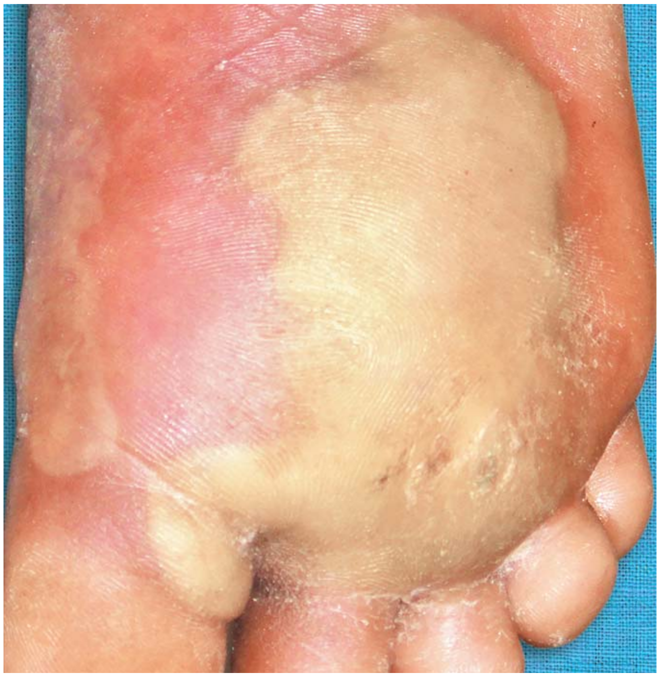

Fig. 1.109: Abscess in plantar aspect of foot. Note the visible pus.

Complications of an Abscess

- Bacteraemia, septicaemia, and pyaemia.
- Multiple abscess formation.
- Metastatic abscess.
- Destruction of tissues.
- *Antibioma formation* (common in breast abscess). Once abscess forms, thick fibrous tissue develops around abscess cavity because of antibiotics. Cavity contains sterile pus as thick flaques. It is nontender, localised, smooth, hard swelling which may mimic carcinoma (in breast—carcinoma breast). It should be differentiated clinically from carcinoma (it is not progressive whereas carcinoma is progressively increasing). FNAC is essential to differentiate. Antibioma should be excised.
- Sinus and fistula formation.
- Large abscess may erode into adjacent vessels and can cause life-threatening torrential haemorrhage, e.g. as in pancreatic abscess.
- Abscess in head and neck region can cause laryngeal oedema, stridor and dysphagia.
- Specific complications of internal abscess:
  - Brain abscess can cause intracranial hypertension, epilepsy, neurological deficit.
  - Liver abscess can cause hepatic failure, rupture, jaundice.
  - Lung abscess can lead to bronchopleural fistula or septicaemia or respiratory failure or ARDS.

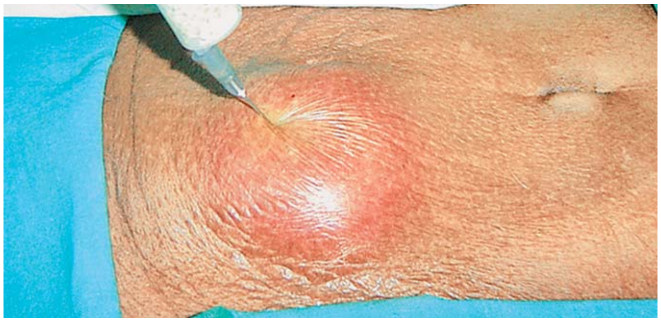

Fig. 1.110: Abdominal wall abscess being aspirated. Note the aspirated pus in the syringe. Abscess should be aspirated prior to drainage.

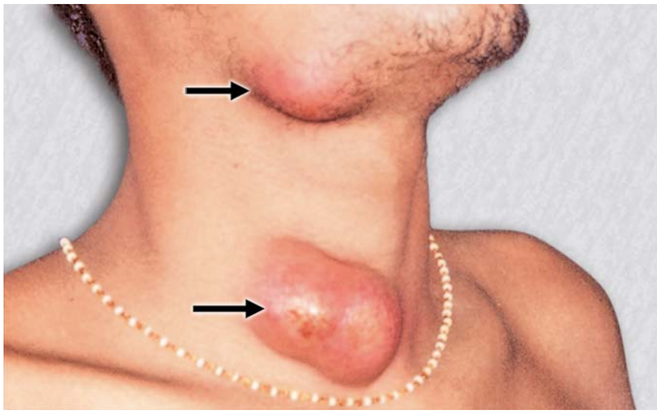

Fig. 1.111: Multiple abscesses in neck, chin, face. It is common in HIV, diabetes and immunosuppressed.

Treatment of an abscess

Abscess should be formed before draining. Exceptions for this rule are:

- Parotid abscess
- Breast abscess
- Axillary abscess
- Thigh abscess
- Ischiorectal abscess

Features of a formed abscess

- Visible pus
- Pointing tenderness
- Fluctuation
- Excruciating pain

Note:
Pus anywhere will come to surface; pus anywhere should be drained.

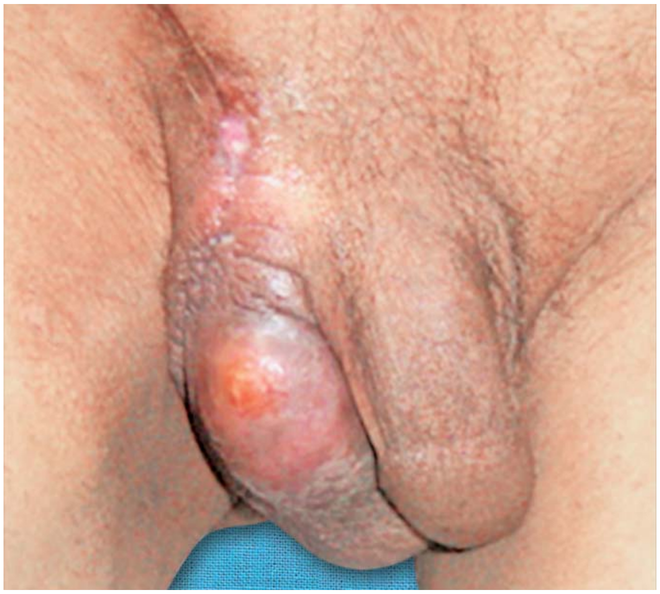

Fig. 1.112: Scrotal abscess, which is well-localised and ready for drainage. Patient has undergone surgery for hernia earlier.

Procedure

***Hilton's** method* of draining an abscess.

- Initially broad spectrum antibiotics are started (depending on severity, extent and site of the abscess).
- Under general anaesthesia or regional block anaesthesia,* after cleaning and draping, abscess is aspirated and presence of pus is confirmed.
- Skin is incised *adequately*, in the line parallel to the neurovascular bundle in the most dependent position.
- Next, pyogenic membrane is opened using ***Sinus forceps***** and all loculi are broken up. Abscess cavity is cleared of pus and washed with saline.
- A **drain** (either gauze drain or corrugated rubber drain) is placed.
- Wound is ***not closed***. Wound is allowed to granulate and heal. Sometimes *secondary suturing* or *skin grafting* is required.
- Pus is sent for ***culture and sensitivity***.
- Biopsy should be done in suspected tuberculosis or malignancy.
- Antibiotics are continued.
- Treating the ***cause*** is important.

Counter-incision is placed in breast abscess which is placed in upper quadrant.

Incision should be deeper while draining pus in radial and ulnar bursae, palmar spaces and tenosynovitis.

Problems in drainage

- Improper drainage
- Bleeding
- Residual abscess or sinus formation

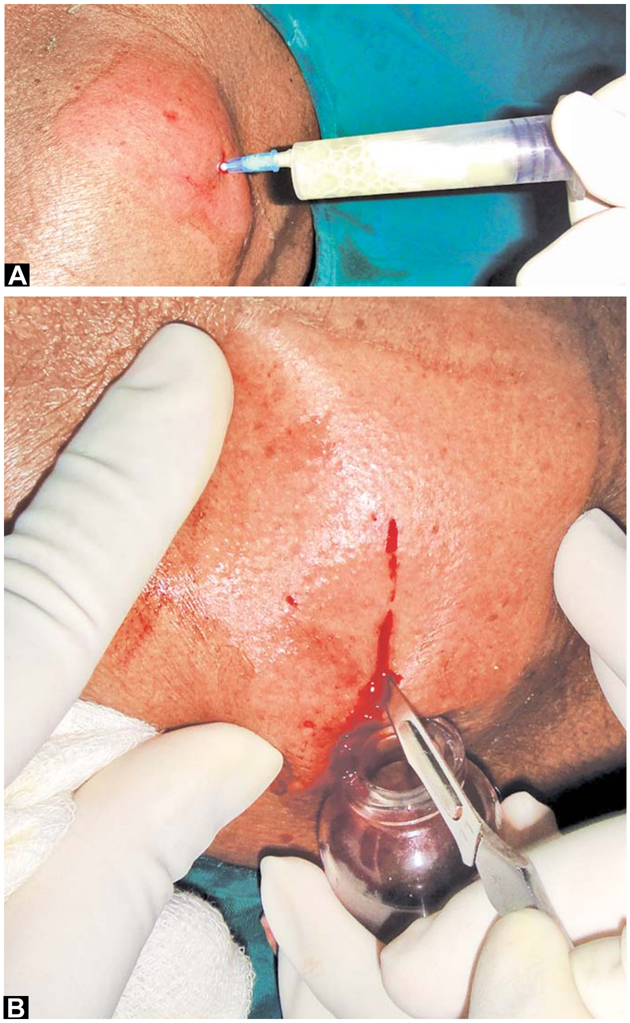

Figs 1.113A and B

---

* As the pus is acidic local anaesthetic agent will not act and hence it is not used.

** Sinus forceps do not have lock and has got serrations in the tip. It is called as sinus forceps because it was initially designed and used to pack sinuses.

Joseph Lister is father of modern surgery and started antiseptic surgery in 1912.

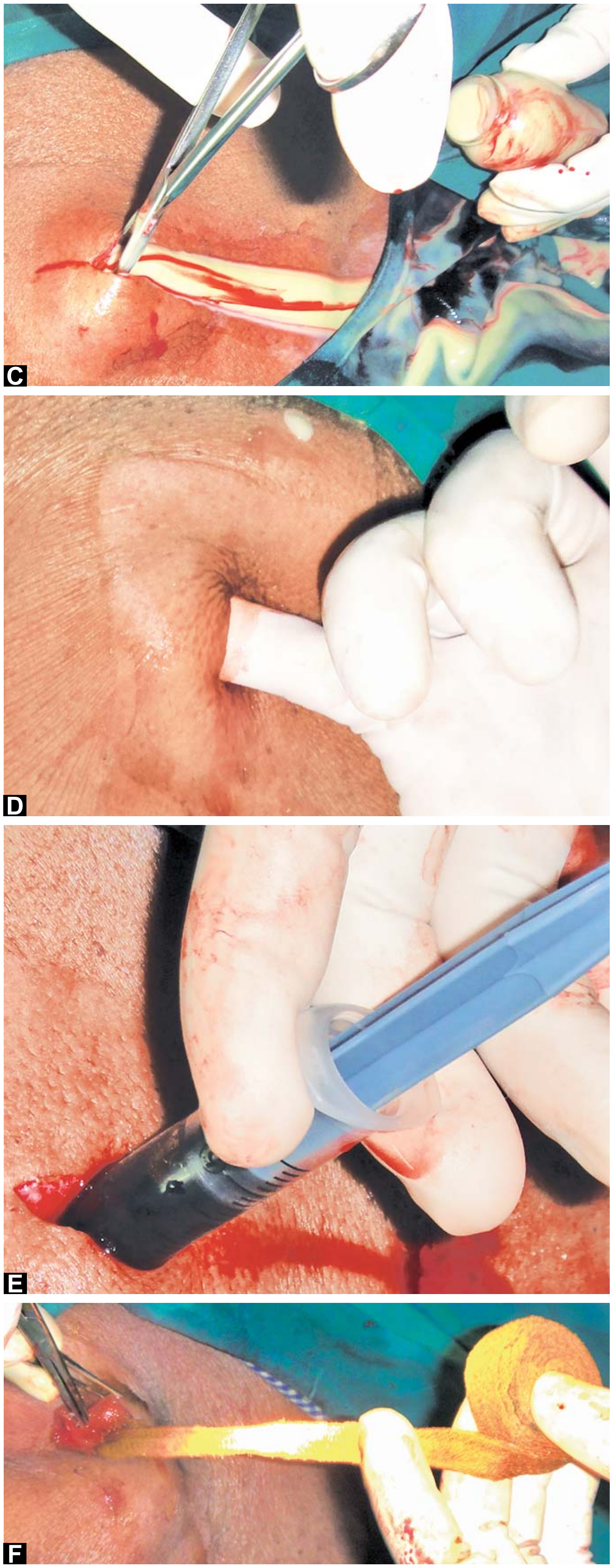

Figs 1.113C to F

Figs 1.113 A to F: Technique of incision and drainage of a pyogenic abscess. Abscess should be aspirated first to get pus; No. 11 blade is used to incise; using sinus forceps pyogenic membrane is opened; pus is collected for culture; loculi are broken using sinus forceps and little finger; cavity is irrigated with *normal saline (ideal)* and povidone iodine; cavity is packed with roller gauze; wound is not sutured.

Differential Diagnosis

- ***Aneurysm:*** Especially in popliteal, femoral and axillary regions. So before draining the abscess, presence of pus is confirmed by aspirating with a needle. It should be remembered that thrombosed aneurysm may not be pulsatile but can be warm, soft and tender.
- ***Soft tissue tumours:*** Sarcomas may be smooth, soft and warmer.
- Haematoma.

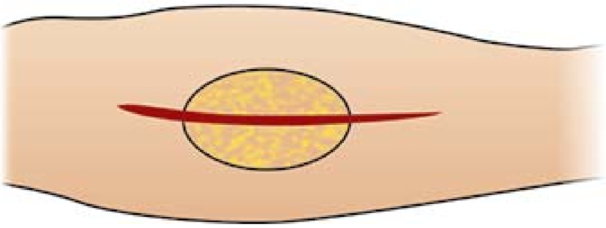

Fig. 1.114: Incision for draining an abscess by Hilton's method. Note the longitudinal incision.

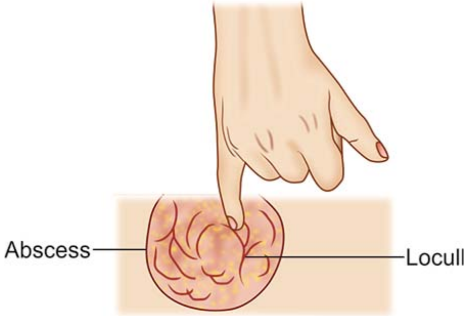

Fig. 1.115: Abscess with loculi and breaking of loculi using finger for drainage.

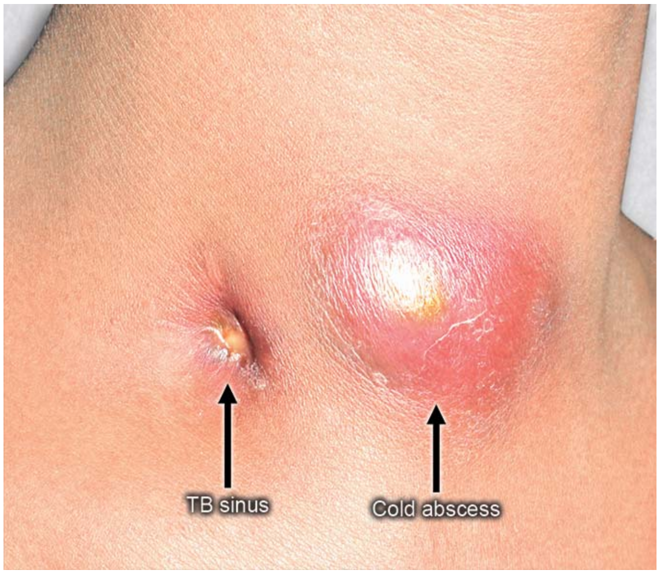

Fig. 1.116: Typical cold abscess and tuberculous sinus in the neck. It does not show signs of acute inflammation. It should be drained (under cover of antituberculous drugs) through *nondependent incision* and incision should be closed without placing a drain.

Differences between pyogenic abscess and cold abscess

Pyogenic abscess	*Cold abscess*
a. Red, warm, tender, with signs of acute inflammation	a. No signs of acute inflammation
b. Pyogenic bacteria are nonspecific organisms (*Streptococcus, Staphylococcus*)	b. Tuberculous bacteria
c. For drainage, dependent incision is used	c. Nondependent incision is used
d. Suturing of the wound is not done	d. Wound is sutured
e. Drain is placed	e. Drain is not placed (otherwise sinus will form which is difficult to treat)

BACTERAEMIA

Presence of bacteria in blood.

SEPTICAEMIA

Presence of overwhelming and multiplying bacteria in blood with toxins causing SIRS (Systemic inflammatory response syndrome) or MODS (Multiorgan dysfunction syndrome).

Types

a. ***Gram +ve septicaemia*** is due to staphylococci, streptococci, pneumococci, etc. It is common in children, old age, diabetics and after splenectomy. After splenectomy, overwhelming post-splenectomy sepsis (OPSI) is not uncommon. Common origin is skin, respiratory infection.

b. ***Gram –ve septicaemia*** is common in acute abdomen like peritonitis, abscess, urinary infections, biliary infections, postoperative sepsis. It is commonly seen in malnutrition, old age, diabetics, immunosuppressed people. Common bacteria are *E. coli, Klebsiella, Pseudomonas, Proteus.* Condition is also called endotoxic shock. Common focus of Gram-negative sepsis are urinary infection, abscess or infected wounds, biliary sepsis, postoperative wounds.

Stages of Gram-negative Septicaemia

a. ***Warm stage*** is *reversible stage*. Here existing fever is due to pyrogenic response. Patient is toxic with fever, chills and rigors.

b. ***Cold stage*** is *irreversible stage*. Here fever is not present due to absence of pyrogenic response. Patient is having renal failure, ARDS, liver failure and multiorgan failure.

Investigations

- Urine/pus/discharge culture.
- Blood culture.
- Haematocrit.
- Electrolyte assessment.
- PO_2 and PCO_2 analysis.
- Blood urea, serum creatinine, liver function tests.

Treatment

- Antibiotics like cefoperazone, ceftazidime, cefotaxime, amikacin, tobramycin, metronidazole.
- Fresh blood transfusion.
- Adequate hydration.
- Oxygen supplementation.
- Ventilatory support.
- Electrolyte management.
- Parenteral nutrition (TPN).
- CVP line for monitoring and perfusion.
- FFP or platelets in case of DIC.

Complications of septicaemia

- Disseminated intravascular coagulation (DIC)
- ARDS
- Liver dysfunction
- Renal failure
- Bone marrow suppression—thrombocytopaenia
- Multiorgan failure

PYAEMIA

Presence of multiplying bacteria in blood as emboli which spread and lodge in different organs in the body like liver, lungs, kidneys, spleen, brain causing ***pyaemic abscess.*** This may lead to ***multiorgan dysfunction syndrome (MODS***). It may endanger life if not treated properly.

Clinical features

- Fever with chills and rigors
- Jaundice, oliguria, drowsiness
- Hypotension, peripheral circulatory collapse and later coma with MODS

Common Causes

- Urinary infection *(most common)*.
- Biliary tract infection.
- Lower respiratory tract infection.
- Abdominal sepsis of any cause.
- Sepsis in diabetics and immunosuppressed individuals like HIV, steroid therapy.

> *In tuberculus arthritis, starting pain at night is very important, as it attracts more attention than occasional discomfort during the day.*
>
> **— Royal Whitmann**

Investigations

- Total leucocyte count.
- Pus culture.
- Blood culture.
- Urine culture.
- Blood urea and serum creatinine.
- LFT.

Treatment

- Monitoring of vital parameters.
- Antibiotics (ceftazidime, cefoperazone, ceftriaxone sodium).
- IV fluids, maintenance of urine output.
- Hydrocortisone.
- Blood and plasma transfusion.
- Nasal oxygen, ventilator support, monitoring of pulmonary function.

METASTATIC AND PYAEMIC ABSCESS

Metastatic abscess

- It is an abscess which occurs as a spread from other abscess. For example, lung abscess causing metastatic abscess in the brain (common example).
- Presentation here is of features of focus abscess and of metastatic abscess (localised features).

Pyaemic abscess

- It is from any infective focus which need not be always from an abscess (from cellulitis or skin infections, etc.) causing pyaemic emboli leading into multiple abscess in different places like brain, kidneys, liver, etc.
- Presentation here, is mainly of systemic features involving multiple organs with toxicity.
- These emboli contain bulk of multiplying organisms often derived from infective thrombus or vegetations. Focus may be an abscess, cellulitis, skin infection, acute osteomyelitis, and acute bacterial endocarditis. Acute appendicitis with severe sepsis can cause infective pyaemia in liver called as *pyelphlebitis or portal pyaemia.*
- Pyaemic abscess are multiple, deeper, beneath the fascia or in the internal organs. When it is on the surface, it is less tender without any clear signs of inflammation.
- **Management:** Evaluation for focus of infection, pus for culture, blood culture (three samples), antibiotics, critical care, systemic therapy, drainage of surface abscesses.
- Pyaemic abscess carries high mortality with SIRS and MODS.

BOIL (Furuncle)

It is an acute staphylococcal infection of a hair follicle with perifolliculitis which usually proceeds to suppuration and central necrosis.

- Often boil opens on its own and subsides (*S. aureus* infection).
- Furuncle in external auditary canal is very painful because of rich cutaneous nerves. Here skin is adherent to perichondrium.
- Boil often heals spontaneously; suppuration will not occur in such boil; it is often called as *blind healed dull boil.*
- Boil is *common* over back, neck, thigh, and forearm even though it can occur anywhere. Boil in eyelash follicle is called as *stye*. Boil can occur in perianal region which can lead into abscess and fistula. Boil can lead into *hidradenitis, common in axilla and pubic region*. Boil can cause *cellulitis* of local area. Overlying skin undergoes necrosis. But during healing re-epithelialisation occurs.
- Boil commonly *subsides* spontaneously often with the support of suitable antibiotics; occasionally it requires incision and drainage. Regional enlarged tender lymph nodes may be palpable due to secondary infection.
- Systemic features are *not common* unless it is multiple/recurrent/severe or in diabetics and immunosuppressed.
- Multiple/recurrent boils are *common* in diabetics.

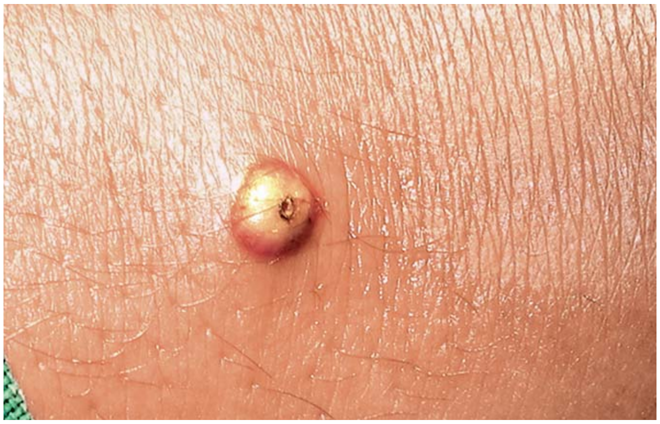

Fig. 1.117: Furuncle/boil are infection of hair follicle with perifolliculitis due to *Staphylococcus aureus.*

Treatment

- Antibiotics given if boil is not resolving spontaneously—*cloxacillin*/amoxycillin.
- Rarely drainage of boil is needed in severe persistent form.

Complications

- Cellulitis.
- Lymphadenitis.
- Hidradenitis (*Infection of group of hair follicles*).
- Boil in dangerous zone can cause cavernous sinus thrombosis.

HIDRADENITIS SUPPURATIVA

It is a chronic infective and fibrous disease of the skin bearing *apocrine* sweat glands. Apocrine sweat glands are coiled glands which open into the hair follicles.

Sites of apocrine sweat glands

- ❖ Axilla
- ❖ Groin
- ❖ Areola
- ❖ Umbilicus
- ❖ Scalp, chest
- ❖ Perineum

Aetiology

- Obesity, smoking.
- Poor hygiene.

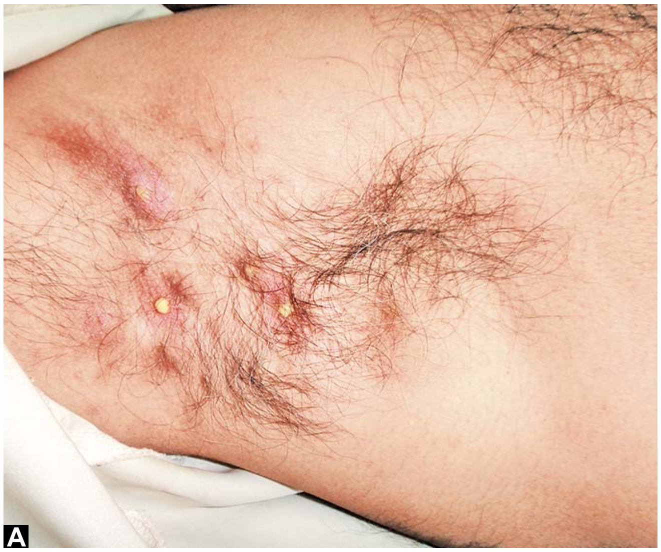

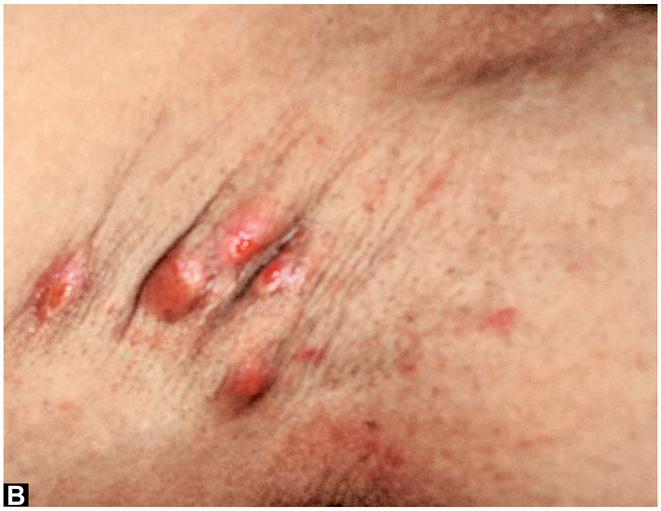

Figs 1.118A and B: Hidradenitis suppurativa. It is chronic infection of apocrine sweat glands of the skin. It is common in axilla (*Courtesy*: Dr Achaleshwar Dayal, MS, Itarsi, MP).

- Diabetes mellitus.
- Steroids.

Common bacteria involved are staphylococci, streptococci, *Staphylococcus aureus*, *Propionibacterium acnes*.

Pathogenesis

Obstruction of duct of apocrine sweat gland by keratin
↓
Dilatation of the duct of gland
↓
Infection and abscess formation
↓
Involvement of subcutaneous tissue and adjacent apocrine glands
↓
Fibrosis, scarring, sinus formation
↓
Spread to surrounding tissues

Clinical Features

- Common in females 4 : 1.
- The most common site is axilla. Often it is bilateral.
- Multiple discharging sinuses, with nodules in the skin which is tender.
- Induration due to fibrosis.

Investigation

- Discharge study—culture/sensitivity and AFB.
- Biopsy to rule out tuberculosis or malignancy.

Differential Diagnosis

- Tuberculous sinus.
- Malignancy (squamous cell carcinoma of skin).
- Lymph node mass in the region which are in *deeper plane*.

Treatment

- Antibiotics—cephalosporins, long acting penicillin.
- Excision of the involved area widely followed by skin grafting or flaps.
- Excision of the involved area widely followed by skin grafting or flaps (radical wound excision).
- Antiandrogen drugs.

CARBUNCLE

Word meaning of carbuncle is ***charcoal.***

- It is *an infective gangrene of skin and subcutaneous tissue*.
- ***Staphylococcus aureus*** is the main culprit.
- Common site of occurrence is ***nape of the neck and back.*** *Skin* in this area is thick. Condition also can occur in shoulder, cheek, hand, forearm.
- It is common in ***diabetics*** and after forty years of age**.**
- It is common in males.

Infection
↓
Development of small vesicles
↓
Sieve like pattern
↓
Red indurated skin with discharging pus
↓
Many fuse together to form a central necrotic ulcer (*ash-grey slough*) with peripheral fresh vesicle looking like a "***rosette***" (cribriform)
↓
Skin becomes black due to blockage of cutaneous vessels
↓
Disease spreads to adjacent skin rapidly

Patient is ***toxic*** and in diabetics they are ***ketotic***

Investigations

- Urine sugar and urine ketone bodies.
- Blood sugar.
- Discharge for C/S.

> *An external sore may be obvious on external genitalia—hot and tender in gonorrhea and chancroid, painless and indolent in syphilis.*
> **— Angus E.W. McLachlam**

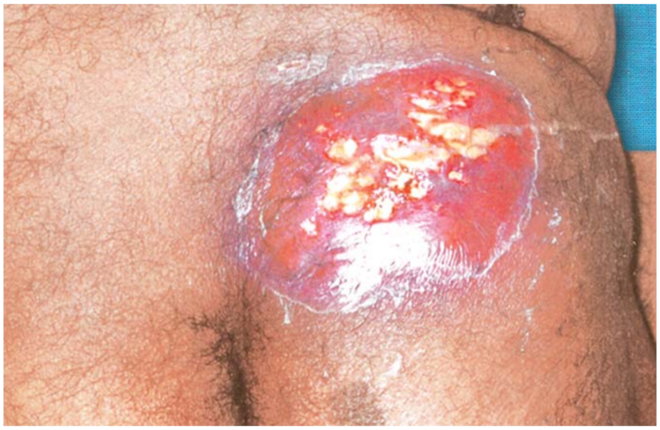

Fig. 1.119: Carbuncle in the buttock. Note the vesicles and typical colour.

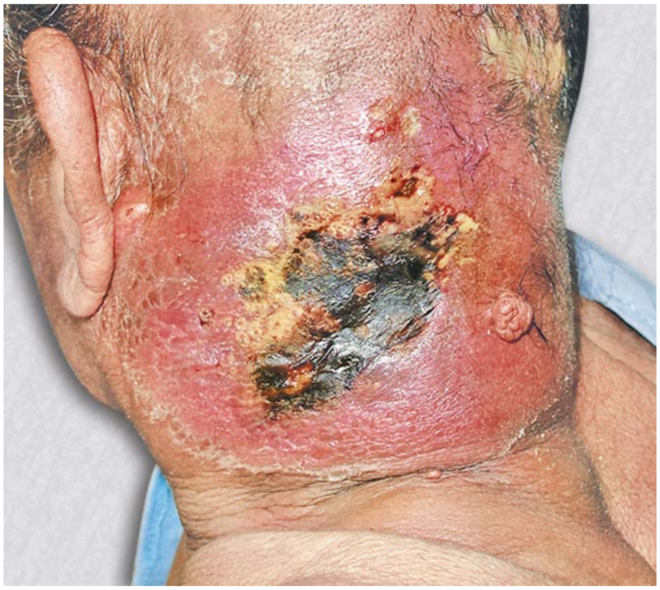

Fig. 1.120: Carbuncle in the nape of the neck—typical site. Note the wide area of involvement and dark area—charcoal like. *Ash-grey slough* is specific.

Treatment

- Control of diabetes is essential using insulin.
- Antibiotics like penicillins, cephalosporins or depending on C/S is given.
- Drainage is done by a cruciate incision and debridement of all dead tissues is done. Excision is done later.
- Once wound granulates well, ***skin grafting*** may be required.

Renal Carbuncle is an entity which occurs in kidney due to infection, forming localized infective mass lesion.

POTT'S PUFFY TUMOUR

- It is a misnomer. It is not a tumour.
- It is formation of diffuse external swelling in the scalp due to **subperiosteal pus** formation and **scalp oedema**.
- It originates commonly in frontal region and may extend into other regions.
- Often there is acute osteomyelitis of frontal bone.

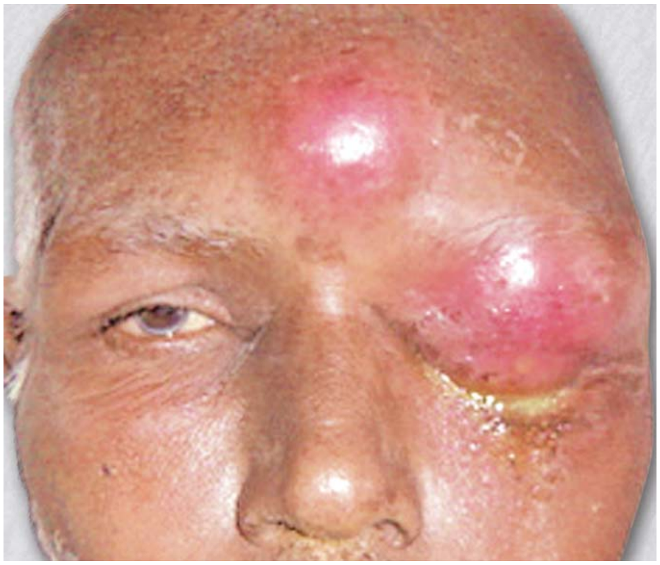

Fig. 1.121: Pott's Puffy tumour. Note the swelling over the frontal region and eyelid.

Causes

- Chronic frontal sinusitis which eventually suppurates and extends into subperiosteal region.
- Trauma—subperiosteal haematoma.
- Chronic suppurative otitis media.

Clinical Features

- Pain and boggy swelling in frontal region which is warm, tender.
- Toxicity and drowsiness.
- Pitting scalp oedema is typical.

Complications

- *Osteomyelitis* of frontal bone.
- Spread of infection into intracranial cavity leads to *intracranial abscess* (Extradural or subdural abscess) (*dumb-bell* abscess). So may present with features of raised intracranial tension like headache, coning and convulsions.

Investigations

- Total leucocyte count—increased. ESR—raised.
- X-ray skull. CT scan.

Differential Diagnosis

Secondaries in the skull or brain.

Treatment

- Antibiotics and drainage under general anaesthesia before it spreads into cranial cavity.
- Once it extends into cranial cavity, it is treated accordingly by formal neurosurgical decompression (often using ***Dandy's brain cannula***).
- Osteomyelitis of skull bones requires radical removal with reconstruction of skull defect.

PYOGENIC GRANULOMA (GRANULOMA PYOGENICUM)

- It is a common condition which occurs on the face, scalp, fingers and toes.
- It may be due to minor trauma or minor infection.
- Infection leads to formation of unhealthy granulation tissue which protrudes through the wound.
  It is also called as acquired lobular capillary haemangioma.

Clinical Features

- Usually single, well localised, red, firm, nodule, which bleeds on touch.
- May or may not be tender.

Sites

- Face
- Scalp
- Fingers
- Toes

Differential Diagnosis

- Haemangioma.
- Papilloma.
- Skin adnexal tumours.
- In recurrent cases, haemangioendothelioma and melanoma.

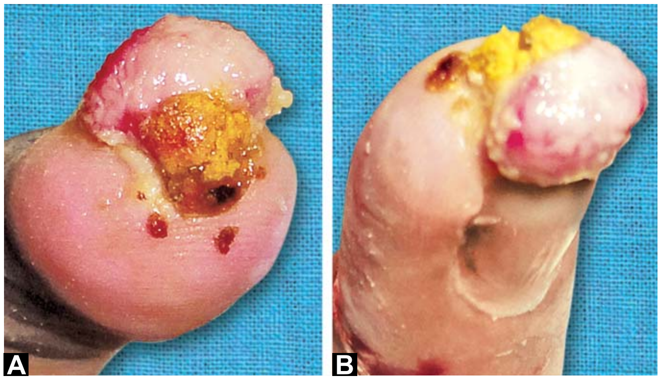

Figs 1.122A and B: Pyogenic granuloma finger.

Treatment

- Excision, laser surgery.
- Tissue to be sent for histopathological study.

IMPETIGO

- It is highly infectious superficial skin infection caused by staphylococci/streptococci organisms.
- Usually seen in children, with formation of multiple blisters that rupture and coalesce, to cover as *honey-coloured* crust.
- Treatment is oral antibiotics and topical antiseptics.

SCRUM POX

It is a type of impetigo seen in Rugby players (Staphylococcal infection).

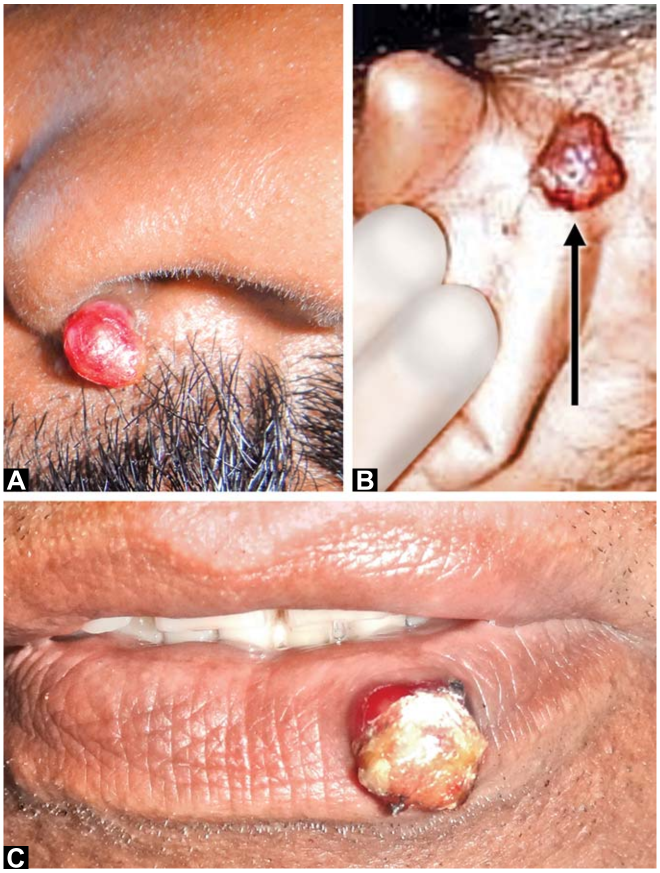

Figs 1.123A to C: Pyogenic granuloma (A) adjacent to nose; (B) in the—scalp post-auricular region and (C) lower lip.

TETANUS

> *It is painful for him to open his mouth. His heart beats too slowly (or weakly) for speech. You observe his saliva falling from his lips, but not falling completely.…He suffers stiffness in his neck. He does not find he can look at his two shoulders and his breast.*
>
> —**(Anonymous),** Circa 1500 BC

It is an infective condition caused by *Clostridium tetani* organisms leading to reflex muscle spasm, often associated with tonic clonic convulsions.

Predisposing Factors for Tetanus

- Absence of prior tetanus toxoid immunisation.
- Trauma with lacerations, deep wounds, crush devitalised wounds, presence of foreign body, wounds with anaerobic environment in the tissues.
- Chronic suppurative otitis media with perforation, caries teeth.
- Improper sterilisation in the ward, labour (septic abortion) and operation theatre.
- Tattooing, rusted nails (there is a myth that only rusted instrument will cause tetanus; this is not true), ear lobe prick, colloquial perianal therapies,

Note:
Tetanus is *not communicable* from person to person.

Even if you're on the right track, you'll get runover if you just sit there. — **Arthur Godfrey**

Organism

- ***Clostridium tetani*** is a *Gram-positive, anaerobic, motile, noncapsulated,* organism with *peritrichous flagella*, and ***terminal spores*** *(Drum stick appearance*).
- Spore is the infective agent. They are found in soil, manure, dust.
- Spore can gain entry through any wound, pricks, injuries resulting from road traffic accidents, penetrating injuries.

Toxins

- Exotoxins released by *Clostridium tetani* are *tetanospasmin* and *tetanolysin* (Exotoxin is the one which is released from bacteria without their destruction/death. *Endotoxin* is the one which is released with death of the bacteria). Tetanus bacteria release only exotoxins.
- Tetanospasmin, the *common exotoxin* released by the multiplying bacteria once spore germinates travels along the *perineural* sheath, *lymphatics* along the nerve, and through blood to cause various effects.

Pathogenesis

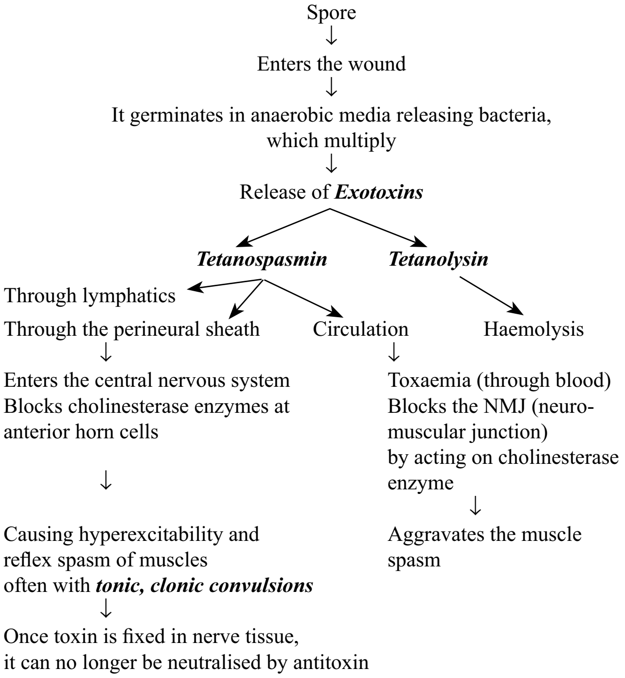
Spore
Enters the wound
It germinates in anaerobic media releasing bacteria, which multiply
Release of ***Exotoxins***
Tetanospasmin
Tetanolysin
Through lymphatics
Through the perineural sheath
Circulation
Haemolysis
Enters the central nervous system
Blocks cholinesterase enzymes at anterior horn cells
Toxaemia (through blood)
Blocks the NMJ (neuromuscular junction) by acting on cholinesterase enzyme
Causing hyperexcitability and reflex spasm of muscles often with ***tonic, clonic convulsions***
Aggravates the muscle spasm
Once toxin is fixed in nerve tissue, it can no longer be neutralised by antitoxin

Clinical Features

Symptoms

- *Trismus* is the most common symptom
- Jaw stiffness, pain and stiffness in the neck and back muscles
- Anxiousness, sweating
- Headache, delirium, sleeplessness
- Dysphagia
- Dyspnoea

Signs

- ***Trismus,*** due to spasm of masseter and pterygoids.
- ***Risus sardonicus*** (smiling facies), due to spasm of the facial muscle—zygomaticus major. Looks as if patient is smiling.
- Neck rigidity.
- Spasm and rigidity of all muscles.
- Hyperreflexia.
- Respiratory changes—due to laryngeal muscle spasm, infection, aspiration.
- ***Tonic clonic convulsions.***
- Abdominal wall rigidity often with haematoma formation.
- Severe convulsion may often lead to fractures, joint dislocations and tendon ruptures.
- Fever and tachycardia.
- Retention of urine (due to spasm of urinary sphincter), constipation (due to rectal spasm).
- Rarely features of ***carditis*** are seen due to involvement of the cardiac muscle, which is dangerous, as it often leads to cardiac arrest and death. Here ***steroids*** are very useful. It presents with refractory *bradycardia.*
- Symptoms will be aggravated by stimuli like light, noise.

Incubation Period

- Time between the ***entry of spore and appearance of first symptom***.
- Usually 7-10 days.
- Shorter the incubation period worser the prognosis and more severe the course of disease. If it is less than 3 days it is commonly fatal.

Period of Onset

- *Time between appearence of first symptom and appearance of first sign.*
- Shorter the period of onset worser the prognosis and vice versa. If it is less than 2 days it is commonly fatal.

Effects on Respiratory System

Diaphragm and other muscles of respiration undergo spasm causing tachypnoea, respiratory distress, respiratory infections, aspiration, cyanosis, respiratory failure with altered PO_2 and PCO_2 levels.

Types of Tetanus

- ***Early tetanus*:** It is a severe form with a short incubation period and poor prognosis.
- ***Latent tetanus*:** Wound is healed and forgotten. After a long incubation period, may be years later, under favourable environment, spores release bacteria and cause tetanus. It carries better prognosis. Latent tetanus is often called as *delayed tetanus*.
- ***Late tetanus*:** Disease develops many months after injury.
- ***Ascending tetanus:*** Symptoms and signs progress from below upwards.
- ***Descending tetanus*:** Symptoms and signs progress from above downwards.
- ***Cephalic tetanus:*** Facial muscles are involved first (3rd, 4th, 6th and 7th cranial nerves can get involved). Facial nerve is commonly involved in this type. Oculomotor nerve—3rd nerve (*ophthalmoplegia*), hypoglossal nerve—12th nerve (*spasm of tongue*) are other cranial nerves involved.
- ***Localised tetanus.*** Here muscles adjacent to the wound or muscles of one segment or one area develop spasm. It is due

to less virulent toxin or released toxins are of less concentration or only one or few segments of anterior horn cells of the spinal cord are affected.

- ***Bulbar tetanus*:** When muscles of deglutition and respiration are involved. Highly fatal.
- ***Tetanus neonatorum*** (7th day tetanus): Tetanus occurring in neonates. Spread is through umbilical cord. It carries very high, nearly 100% mortality.
- ***Urban tetanus***. Due to repeated injections in IV drug abusers.
- ***Postoperative tetanus*** due to improper sterilisation.
- Even though it is classified as *acute tetanus* (tetanus develops within 10 days) or *chronic tetanus* (tetanus develops from 10 days to 3 months) it is only of outcome value as therapy will be same in both. *Post-abortion or puerperal tetanus* develops due to practice of improper sterilisation during abortion or delivery. If patient is not given tetanus toxoid during the first attack, he can develop second attack of tetanus at a later period called as *recurrent tetanus* as patient who had once tetanus is not immune for development of second attack of tetanus. In children and adolescents chronic suppurative otitis media (CSOM) with perforation of tympanic membrane can cause tetanus—*otitis tetanus.*

Different Postures in Tetanus

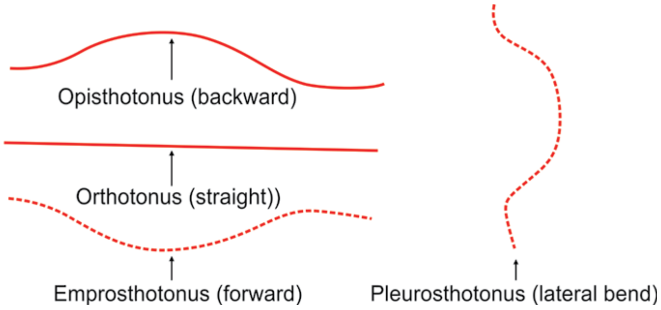

Fig. 1.124: Different types of postures seen in tetanus.

- ***Opisthotonus**: Posterior muscles are acting more, so backward bending.*
- ***Orthotonus***: *Straight posture*. Both front and back muscles are acting equally.
- ***Emprosthotonus:*** *Forward bending* as front muscles are acting more.
- ***Pleurosthotonus***: *Lateral bending* as lateral muscles act more.

Complications of tetanus

- Fracture bones
- Haematoma
- Aspiration pneumonia, respiratory failure, ARDS
- Carditis, arrhythmias—life-threatening
- DVT, pulmonary embolism
- Toxaemia
- Secondary infection—septicaemia
- Bedsore. Malnutrition, stress gastric ulcers
- Mortality used to be very high in earlier days up to or more than 50%. It is reduced to 15-20% at present; but it is still higher in children and elderly
- Repeated uncontrollable convulsions can lead into coma and death.

Staging of Tetanus

- ***Mildly ill**:* Rigidity, spasm, trismus and different postures.
- ***Seriously ill***: Spasm, rigidity, severe respiratory infections, dysphagia.
- ***Dangerously ill**:* Cyanosis with ***respiratory failure*** and tonic-clonic convulsions, cyanosis.

Differential Diagnosis

- Strychnine poisoning (here patient is normal in between).
- Trismus due to other causes like—dental, oral, tonsillar sepsis, oral malignancy, temporomandibular joint dysfunction.
- Meningitis.
- *Hydrophobia.*
- Convulsive disorders.

Culture media for *Clostridium tetani* are RCM (Robertson's cooked meat) media and nutrient agar.

Cause of death in tetanus

- Respiratory failure with aspiration pneumonia and ARDS
- Severe carditis—an ominous sign
- Mortality is 45-50%—becoming less now in will equipped centers (15%)

Treatment

- Patient is admitted and isolated in a dark, quiet room.
- Antitetanus globulin (ATG), 3,000 units IM single dose is given. Test dose is not required (It is a human Ig). It is 100 times more effective than ATS. Anaphylaxis is not observed. Once toxin is *fixed to nerve* tissue it *cannot* be neutralised; but *circulating toxins* are *very well* neutralised.
- Antitetanus serum (ATS): When ATG is not available or when patient cannot afford, after ***IV test dose*** (1,000 units of ATS), full dose is given, i.e. 1,00,000 units, half of it is given IM and half of it is given IV. It is a horse serum, so possibility of anaphylactic reactions should be borne in mind.
- Wound debridement, drainage of pus, injection of ATG 250-500 units locally to reduce the toxin effect.
- Ryle's tube has to be passed, initially to decompress, so as to prevent aspiration, but later for feeding purpose.
- Catheterisation.
- IV fluids and electrolyte balance has to be maintained. TPN is also required.

Ancients discovered much, and yet left much more to be discovered.

- ***Tetanus toxoid*** should be given as disease will not give immunity against further infections. To start—first dose, second dose after one month, third dose after six months. *Aluminium phosphate adsorbed tetanus toxoid* 0.5 ml is injected into the deltoid muscle (IM). Booster dose should be given every 4 years or after any significant trauma. In patients who have not been immunised earlier, it needs 30 days to develop antibody after tetanus toxoid injection.
- IV *diazepam* 20 mg 4th or 6th hourly. Dose is adjusted depending on severity and convulsions.
- IV *phenobarbitone* 30 mg 6th hourly.
- IV *chlorpromazine* 25 mg 6th hourly.
- Injection crystalline penicillin 20 lacs 6th hourly and injection. gentamicin and metronidazole to prevent secondary infection.
- Regular suction and clearance of respiratory tract.
- Nasal oxygen.
- In ***severe cases***, patient is ***curarised*** and placed in ***ventilator (IPPR).***
- ***Endotracheal intubation*** or ***tracheostomy*** are often ***life-saving procedures.***
- ***Good nursing care***—Change of position, prevention of bedsores, prevention of DVT (which is common in tetanus and often requires heparin injection).
- Chest (respiratory) physiotherapy during recovery period.
- ***Steroids*** are given when carditis is suspected.
- *Cardiac pacemaker* may be useful in refractory bradycardia and arrhythmias.

Following treatment patient often gets spasm of different muscles (**tics**) for a long period which can be prevented by giving methocarbamol for 6 months to one year.

General measures

- Isolation
- Avoid noise and light
- ATG 3,000 units IM
- ATS-50,000 IM and 50,000 IV—after test dose
- Antibiotics like inj penicillin 20 lacs 6th hourly
- Inj tetanus toxoid 0.5 ml IM—to deltoid muscle
- IV fluids with TPN
- Urinary catheterisation
- Nasogastric tube is passed to prevent aspiration initially, later for feeding
- Regular suction of throat
- Nasal oxygen when required
- Prevention of bedsore formation
- Prevention of DVT by low molecular weight heparin
- Good nursing care.

Specific measures

- IV diazepam 20 mg 6th hourly
- IV phenobarbitone 30 mg 6th hourly
- IV chorpromazine 25 mg 6th hourly
- Endotracheal intubation and ventilator support
- Tracheostomy if there is severe respiratory secretions
- Steroids
- Bronchodilators like deriphylline
- Wound care—debridement, drainage, and local injection of ATG

Prophylaxis against tetanus

- In adults, fresh immunisation to start—second in one month, next in 6 months period. Tetanus toxoid 0.5 ml IM. Booster dose should be given once in every 4 years or after any significant trauma
- In infants—triple antigen (DPT)—6 weeks, 10 weeks, 14 weeks, 18 months, and 5 years
- To pregnant mother tetanus toxoid injections are given in 4th and 6th months of pregnancy
- Additional booster dose given in major injuries or high-risk injuries
- ATG 500-1000 units IM given as a prophylaxis in road accidents, severe burns, crush injuries, war wounds, penetrating wounds and wounds of head and face. Here tetanus toxoid also should be given separately at separate site IM

Monitoring during Therapy

PCO_2, PO_2, haemotocrit, blood urea, serum electrolytes, chest X-ray, ECG are done at regular intervals.

GAS GANGRENE

It is an *infective gangrene* caused by clostridial organisms involving mainly skeletal muscle as *oedematous myonecrosis*. Earlier it was called as ***malignant oedema***.

Source and Predisposing Factors

- Contaminated, manured or cultivated soil, intestines are the sources. Faecal flora commonly contains clostridial organisms enters the wound; in presence of calcium from blood clot or silica (silicic acid) of soil, it causes infection.
- It is common in crush wounds, following road traffic accidents, after amputations, ischaemic limb, gunshot wounds, war wounds. Injury or ischaemia or necrosis of the muscle due to trauma predisposes infection.
- Anaerobic environments in the wound—initial infection with aerobic organism utilises existing oxygen in tissues creating anaerobic environment to cause clostridial sepsis.

Organisms

- *Clostridium welchii* (perfringens): Gram-positive, central spore bearing, nonmotile, capsulated organisms, most common—60%.
- *Clostridium oedematiens.*
- *Clostridium septicum.*
- *Clostridium histolyticum.*

***Clostridium welchii* produce toxins**

- Alpha (most common)
- Beta
- Epsilon
- Iota
- *Phi toxin*—myocardial depressant
- *Kappa toxin*—destruction of connective tissue and blood vessels
- **Bursting factor** and circulating factor

Various *strains* include—A, B, C, D, E.
'A' strain is most common.

***Note*:**
Nonclostridial gas producing organisms like coliforms can also cause gas gangrene.

Exotoxins

- ***Lecithinase*** is important toxin which is haemolytic, membranolytic and necrotic causing ***extensive myositis.*** It splits lecithin into phosphocholine.
- ***Haemolysin*** causes extensive haemolysis.
- ***Hyaluronidase*** helps in rapid spread of gas gangrene.
- ***Proteinase*** causes breaking down of proteins in an infected tissue.

Spores enter through the devitalised tissues commonly
in road traffic accidents, crush injury
↓
Spores germinate
↓
Released bacteria will multiply
↓
Exotoxins are released
cause their effects

Effects

- Extensive necrosis of muscle with production of gas (hydrogen sulphide; nitrogen; carbon dioxide) which stains the muscle brown or black *anaerobic myositis/myonecrosis*.
- Usually muscle is involved from origin to insertion.
- Often may extend into thoracic and abdominal muscles.
- When it affects the liver it causes necrosis with frothy blood—***foaming liver***, is characteristic.
- Rapidly spreading infection which is also often fatal.
- Limbs are commonly involved; but organs like liver can also be affected.
- Muscle glycogen is broken down into lactic acid, CO_2 and hydrogen. Proteinase released by organism forms amino acids which further releases ammonia and hydrogen sulphide. Acid released earlier is neutralised by ammonia and calcium to progress further multiplication of organisms.

Clinical Features

Incubation period is 1-2 days.

- Features of toxaemia, fever, tachycardia (*out of proportion* to fever) pallor.
- Wound is under tension with foul smelling discharge (sickly sweety/*decaying apple* odour).
- Khaki brown coloured skin due to haemolysis.
- ***Crepitus*** can be felt.
- Jaundice may be *ominous sign* and also oliguria *signifies renal failure.*
- *Frequent sites* are adductor region of the lower limb and buttocks and subscapular region in upper limb.

Clostridium welchii can infect limbs, abdominal wall, appendix, gallbladder, common bile duct, intestine, uterus (during septic abortion).

Clinical Types

- ***Fulminant type*** causes rapid progress and often death due to toxaemia, renal failure or liver failure or MODS or ARDS.
- ***Massive type*** involving whole of one limb containing fully dark coloured gas filled areas.
- ***Group type:*** Infection of one group of muscles, extensors of thigh, flexors of leg.
- ***Single muscle type*** affecting one single muscle.
- ***Subcutaneous type*** of gas gangrene involves only subcutaneous tissue (i.e. superficial involvement). It is mainly of *anaerobic cellulitis type* without muscle involvement usually caused by less virulent clostridial organisms other than clostridial welchii. It is usually superficial but may spread and involve fascial planes. It causes necrosis with foul smelling seropurulent discharge.

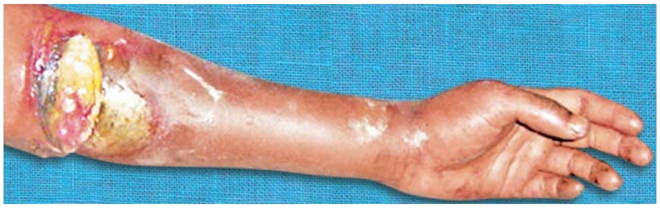

Fig. 1.125: Gas gangrene in upper limb. It is after effect of an assault using an axe by a drunken husband.

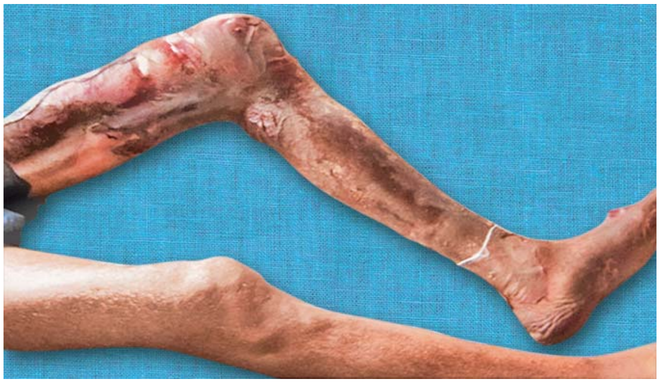

Fig. 1.126: Gas gangrene involving entire leg extensively. Note the black/khaki coloured area. Patient died of toxaemia.

Complications of Gas Gangrene

- Septicaemia, toxaemia.
- Renal failure, liver failure.
- Circulatory failure, DIC, secondary infection.
- Death occurs in critically ill patients.

Investigations

- X-ray shows gas in muscle plane or under the skin.
- Liver function tests, blood urea, serum creatinine, total count, PO_2, PCO_2.
- CT scan of the part may be useful especially in chest or abdominal wounds.
- *Gram's stain shows* Gram-positive bacilli.
- *Robertson's cooked meat media* is used which causes meat to turn pink with sour smell and acid reaction.

Right is right even if everyone is against it, and wrong is wrong even if everyone is for it.

- *Clostridium welchii* is grown in culture media containing 20% human serum in a plate. Antitoxin is placed in one-half of the bacteria grown plate sparing the other half. Zone of opacity will be seen in that half of the plate where there is no antitoxin. In the other half part of the plate where there is antitoxin there is no opacity—*Nagler reaction.*

Prevention of gas gangrene

- Proper debridement of devitalised crushed wounds
- Devitalised wounds should not be sutured
- Adequate cleaning of the wounds with H_2O_2 and normal saline
- Penicillin as prophylactic antibiotic.

Treatment

- Injection benzyl penicillin 20 lacs 4th hourly + Injection metronidazole 500 mg 8th hourly + Injection aminoglycosides (if blood urea is normal) or third generation cephalosporins or metronidazole.
- ***Fresh blood transfusion***.
- ***Polyvalent antiserum*** 25,000 units given intravenously after a test dose and repeated after 6 hours (Welchii 10,000 IU, oedematiens 10,000 IU, and septicum 5,000 IU).
- ***Hyperbaric oxygen*** is very useful.
- Liberal incisions are given. All dead tissues are ***excised*** and **debridement** is done until healthy tissue bleeds.
- Rehydration and maintaining optimum urine output (30 ml/hour) (0.5 ml/kg/hour).
- Electrolyte management.
- In severe cases ***amputation*** has to be done as a ***life-saving procedure***—stump should never be closed *(Guillotine amputation)*.
- Often ventilator support is required.
- *Once a ward or operation theatre is used for a patient with gas gangrene, it should be fumigated for 24-48 hours properly to prevent the risk of spread of infection to other patients especially with open wounds.*
- Hypotension in gas gangrene is treated with whole blood transfusion.
- Therapy should be concentrated in managing dehydration, hypotension, infection, toxaemia by hydration, fresh whole blood transfusion, passive immunisation, antibiotics, and hyperbaric oxygen, doing radical wound excision with removal of all dead tissues with foreign body or amputation with critical care.

TUBERCULOSIS

Whilst meagre Pthisis gives a silent bow, her strokes are sure, but her advances slow. No loud alarms, nor fierce assaults are shown: She starves the fortress first, then takes the town.

—**Samuel Garth**, 1699

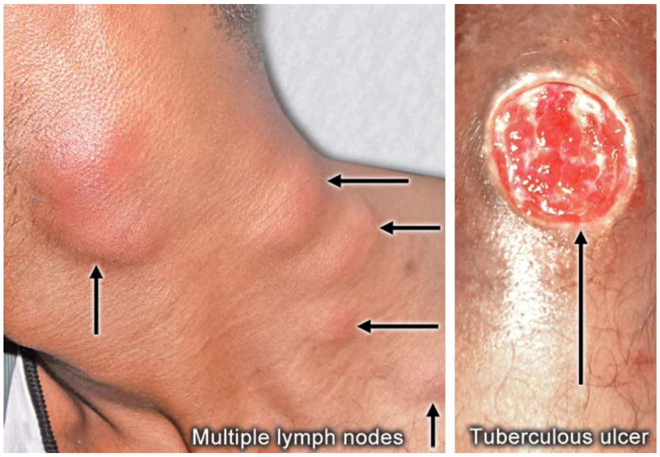

Fig. 1.127: Tuberculous lymphadenitis and tuberculous ulcer.

- It is commonly caused by *Mycobacterium tuberculosis;* occasionally by *Mycobacterium bovis, M. kansasii, M. fortuitum, M. marinum, M. ulcerans.*
- *M. tuberculosis* is ***gram neutral, acid fast, alcohol fast*** straight or slightly curved rods.
- It is prevalent in most of the developing countries and has made its resurgence in the developed countries with the advent of AIDS.

Pathogenesis

- The characteristic lesion here is *'tubercle',* which is an avascular granuloma composed of a central zone containing giant cells, with or without caseation necrosis, surrounded by a rim of epithelioid cells, lymphocytes and fibroblasts.
- It can occur in almost all organs in the body. Presentation may vary depending on the individual sites.

General Features

- Low grade fever with evening rise of temperature.
- Loss of appetite.
- Decreased weight.

Investigations

- ESR is often raised; peripheral smear—lymphocytosis.
- AFB (Acid fast bacillus) staining using Ziehl-Neelsen stain.
- Chest X-ray is done to rule out pulmonary tuberculosis.
- Culture of the organism in Lowenstein Jenson media.
- Mantoux skin test—read after 48 hours of inoculation.
- Guinea-pig inoculation.
- Relevant investigations depending on the site of the tuberculosis.

Treatment

- Antituberculous drugs are given for 6 months to one year.
- Specific treatment is given depending on the site of the tuberculosis.

(Please refer individual chapters for details about tuberculosis at different locations and also chapter drugs for Antituberculous Therapy.)

LEPROSY

- It is caused by *Mycobacterium leprae.* It is a Gram-positive, acid fast bacillus.
- It mainly involves skin, nasal mucosa and peripheral neural tissues.

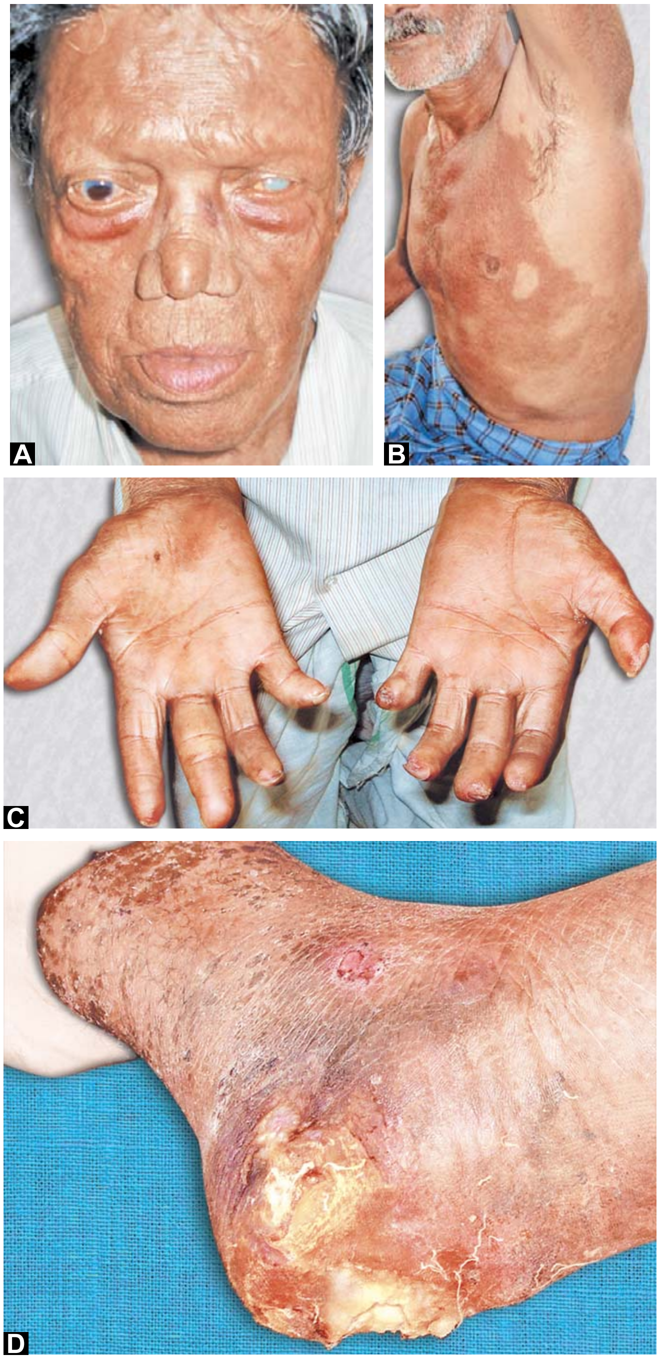

Figs 1.128A to D: A leprosy patient with the typical face (leonine); skin patches; hand deformities and trophic changes; trophic ulcer in heel.

- It involves only the cooler parts of the body (So axilla, gluteal region are not involved). Testicular involvement is seen but not the ovary. It does not involve the vital organs. Though not acutely fatal, the disease leaves the victim severely deformed and crippled for life.

Types

a. *Multibacillary Types*

- ***Lepromatous leprosy***: It denotes little or no host resistance. Bacilli are seen in large numbers in the superficial nodular lesions and the patient is highly infective.
- ***Borderline lepromatous.***
- ***Borderline.***

b. *Paucibacillary Types*

- ***Borderline tuberculoid.***
- ***Tuberculoid leprosy***: Here strong host resistance is observed. The disease is more localised, but it causes more deformities due to early involvement of nerves. Bacilli are scanty in the lesion and so infectivity is minimal.

Investigations

- Regular checking of sensation of the suspected area.
- Split skin smear, nerve biopsy.

Treatment

- Dapsone 100 mg daily.
- Rifampicin 600 mg once a month.
- Clofazimine 50 mg daily + 300 mg once a month.

For paucibacillary types, treatment is for 6 months.
For multibacillary types, treatment is for 2 years or more.

Surgical Complications in Leprosy

a. *Primary Deformities*

- Leonine facies.
- Collapsed nasal bridge.
- Upper branch facial nerve palsy (causes lagophthalmos).
- Keratitis and blindness.
- Claw hand either ulnar or combined ulnar and median nerves.
- Radial nerve palsy—wrist drop (1%).
- Clawing of toes due to involvement of posterior tibial nerve.
- Foot drop due to involvement of lateral popliteal nerve. (Medial popliteal nerve which supplies the tibialis posterior is never involved.)

b. *Secondary Deformities*

- Anaesthesia of the part makes it prone to trauma, infection, infective gangrene, destruction, autoamputation and functionless parts.
- Trophic ulcers in the foot are common.

Treatment

- Reconstructive surgeries.
- Release of contractures.
- Tendon transfers.
- Arthrodesis.
- Ulcer management.
- Physiotherapy and rehabilitation.

SYPHILIS (GREAT POX) (FRENCH DISEASE)

> *It is concluded…that this disease which amongst the Italians is called Gallicus, that is to say, the French disease, should now be named Patursa, …a disease filthie and Saturnall. It is a filthie disease, because it maketh women to be esteemed unchast and irreligious…. There is a twofold kinde of causes…. The first is the only influence or corruption of the aire, from whence we must charitably thinke, that it infected those which were religious. The second is conversation, as by kissing and sucking, as appeareth in children, or by carnall copulation.*
>
> —**Juan Almenor,** 1502

It is a venereal infection caused by *Treponema pallidum.*

Early syphilis: It lasts for 2 years and the patient is infective during this period.

Primary syphilis: It presents as a *Hunterian chancre* (Hard chancre). It is a shallow, painless, indurated, nonbleeding ulcer usually seen in the genitalia and often on the lips, breasts and anal region. It occurs in 3-4 weeks after the infection. It is confirmed by dark field microscopic study of the discharge for the organism.

Secondary syphilis: It occurs in 6-12 weeks.

It presents as:

- Cutaneous coppery rashes.
- Snail-track oral ulcers.
- Condyloma lata: Raised, flat, white lesions seen at the mucocutaneous junctions (mouth, anus vulva).
- Painless, shotty lymphadenopathy—epitrochlear suboccipital.
- Moth-eaten alopecia.
- Hepatitis, arthritis, iritis, meningitis. Massive, enlarged liver in syphilis is called as *hepar lobatum.*
- Syphilitic osteitis with "ivory" sequestrum.

Latent syphilis: It lasts between 2 years to lifetime. Serum tests are positive.

Late (tertiary) syphilis:

- Painless, punched out, *gummatous ulcers* are seen with *"wash leather base" and silvery tissue paper scar.*
- It is seen over the tibia, sternum, ulna, skull, scrotum.
- It is a hypersensitivity reaction occurs as a result vasculitis and obliterative endarteritis.
- *Neurosyphilis.*
- *Cardiovascular syphilis*—aneurysm of arch of aorta.

Investigations

- VDRL test; Kahn test.
- *Treponema pallidum* haemagglutination test (TPHA).
- *Treponema pallidum* immobilisation test (TPI).

Treatment

Penicillin for 15 days is the drug of choice.

Doxycycline 100 mg can be given thrice daily for 15 days. Others: Erythromycin, tetracycline, cephalosporins.

Jarisch-Herxheimer reaction is commonly seen after penicillin, which often requires steroid therapy.

Congenital Syphilis

Here the infection is transmitted from the mother to foetus through placenta.

Early congenital syphilis: It is seen in newborn.

- *Features are:*
  - Rash, syphilitic snuffles
  - Nasal discharge, weight loss
  - Periostitis, meningitis, hepatosplenomegaly
  - Pneumonia alba

Late congenital syphilis:

- *Hutchinson's triad:*
  - *Interstitial keratitis*
  - *8th nerve deafness*
  - *Hutchinson's teeth: Peg shaped upper incisors*
    Moon's molar, molars with cusps
- Congenital neurosyphilis
- Cutaneous, skeletal or visceral gummas
- Saddle nose
- Sabre tibia, Clutton's joint
- Perforated palate

Congenital syphilis is treated with penicillins.

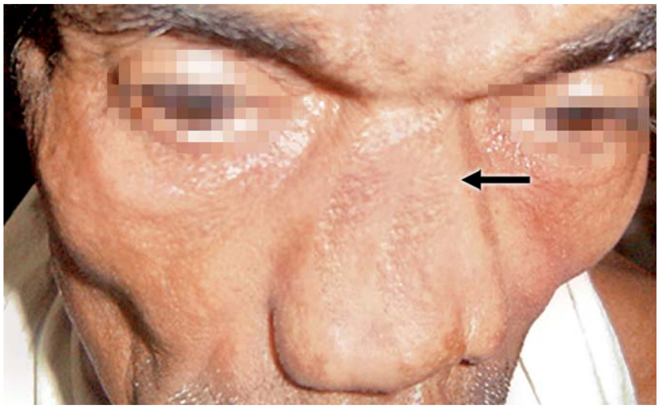

Fig. 1.129: Saddle nose.

ACTINOMYCOSIS

- It is caused by ***Actinomyces israelii***.
- It is an anaerobic Gram-positive ***fungal like bacterium***, which is a branching filamentous organism. It is called as ***"Ray fungus"*** because of ***sun-ray*** appearance.

Clinical Types

- **Facio-cervical**: It is the most common type. Infection spreads either from tonsil or from adjacent infected tooth. Initially an induration develops. Nodules form with involvement of skin of face and neck. It softens and bursts through the skin as ***sinuses*** which discharge *pus which contains sulphur granules* (60%).

- **Thorax**: Lungs and pleura get infected by direct spread from pharynx or by aspiration. Empyema develops. Later chest wall nodules appear leading to sinuses with discharge (20%).
- **In right iliac fossa:** It presents as a ***mass abdomen*** with discharging sinus.
- **Liver** is infected through portal vein (*Honey comb liver*).
- **Pelvic:** Pelvic actinomycosis can occur due to intra uterine devices.

Pathogenesis

Organism enters through deeper plane of the tissue, causes subacute inflammation with induration and nodule formation. Eventually discharging sinus forms at the surface. Pus collected in a swab or sterile tube will show *sulphur granules*.

Clinical Features

- ***Discharging sinus*** with induration and nodules.
- ***No lymph nodal involvement***.
- Through blood spread it may cause pyaemia and endanger life.

Investigations

- Pus under microscopy shows branching filaments.
- ***Gram's staining*** shows ***Gram-positive mycelia*** in centre with ***Gram-negative radiating peripheral filaments***. These clubs are due to host reaction which are lipoid material (antigen-antibody complex).
- Cultured in brain heart infusion agar and thioglycolate media.

Differential Diagnosis

- Chronic pyogenic osteomyelitis.
- Carcinomas at the site.
- Tuberculous disease.

Treatment

- Penicillins for longer period (6-12 weeks).
- Tetracyclines, lincomycin, streptomycin.
- Dapsone and iodides may be useful.
- Antifungals are often given because it is fungal like bacterium.
- Surgical debridement is occasionally required.

MADURA FOOT (MYCETOMA PEDIS)

- It is a chronic granulomatous condition of the foot involving subcutaneous and often deeper tissues causing multiple discharging sinuses.
- It was first identified in ***Madurai, Tamil Nadu (India)*** *by* ***Gill*** (1842, Madura mycosis).
- It is common in India and Africa.
- It is common in Tamil Nadu.
- It can be *fungal* (*more common*) or *bacterial origin. Bacterial* can be *Actinomyces* or *Nocardia*. Among bacterial *Nocardia madurae* is most common.
- *Fungal eumycotic mycetoma* is caused by *Madurella mycetomi* or *Madurella grisa*. Fungal mycetoma causes black granules, crushed smear of which shows thick stout filaments (5 µm). *Bacterial mycetomas* are by *Actinomyces* (*A. israelii* or *A. bovis*) or by *Nocardia* or *Actinomadura* (red granules). Occasionally it can be due to pyogenic bacteria like *Staphylococcus aureus* (*Botryomycosis*). Bacterial mycetoma shows white/yellow/red granules which on smear delineates thin filaments (1 µm). *Nocardia asteroids* can primarily involve lungs later spreading to brain, kidney and other organs as metastatic infection.

Organisms

- ***Nocardia madurae*** (most common)
- *Nocardia brasiliensis*
- *Nocardia asteroides*
- *Actinomyces israelii*

Pathogenesis

Organism enters through a prick in the foot usually who walks barefoot
↓
Reaches deeper plane in the foot
↓
Evokes chronic granulomatous inflammation
↓
Causes pale, painless, firm nodule
↓
Formation of vesicles
↓
Burst to form a discharging sinuses.

- Discharging granules may be *Black, Red, Yellow.*
- In black type of Madura foot, infection is mainly subcutaneous.
- In red and yellow types, it burrows into the deeper plane including bone causing bone necrosis (osteomyelitis). Eventually gross swelling of the limb with multiple discharging sinuses with disability will occur.
- Muscles and bones are involved.
- Tendons and nerves are affected.
- **Regional lymph nodes are not involved.**
- Condition will deteriorate by secondary bacterial infection.

Clinical Features

- Painless diffuse swelling in the foot of long duration.
- Later multiple discharging sinuses develop on the skin.
- Lymph node involvement will not occur unless secondary bacterial infection is present.
- Significant limb disability is common.

Differential diagnosis

- Chronic osteomyelitis
- Tuberculous osteomyelitis
- Carcinoma

Investigations

- Discharge study will show branching filamentous appearance of the organism.
- Culture in Sabouroud's dextrose agar medium.
- Gram stain for actinomycosis will show sun-ray appearance with Gram +ve centre and Gram –ve clubs.
- X-ray of the foot is done to look for osteomyelitis.

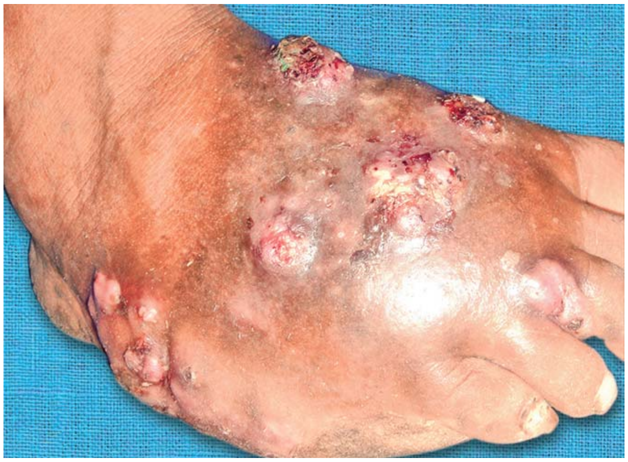

Fig. 1.130: Madura foot. Note the multiple sinuses.

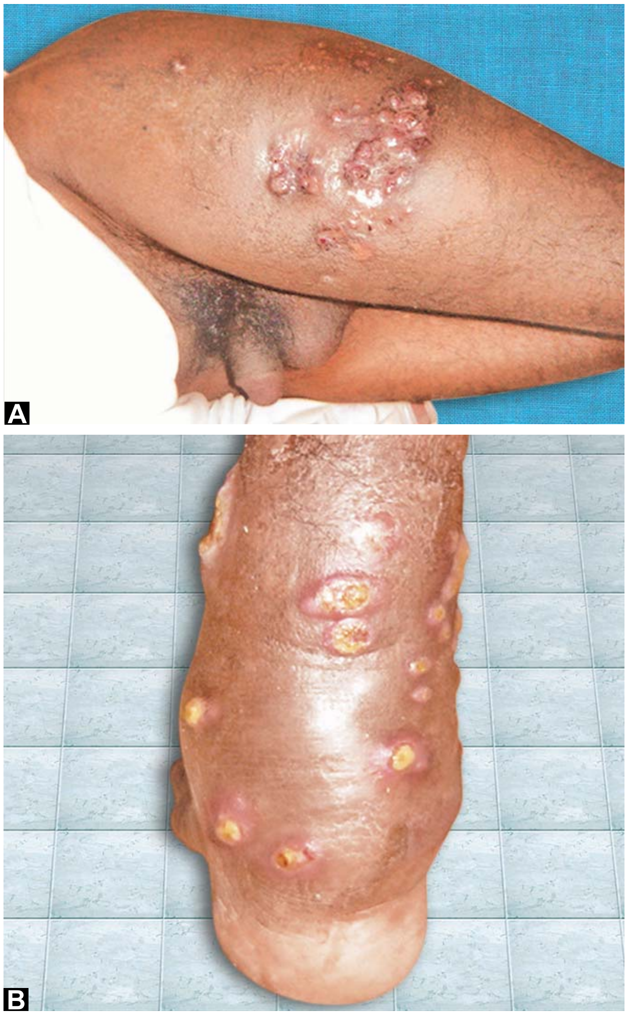

Figs 1.131A and B: Mycetoma thigh and foot in two different patients. Note the discharging sinuses.

Treatment

- Antifungal drugs—amphotericin.
- Long-term penicillins.
- Dapsone, iodides.
- In severe cases amputation may be required.

If infection occurs in the hand it is called as ***Madura hand***.

RABIES (HYDROPHOBIA)

It is an *acute fatal encephalomyelitis* caused by a single stranded RNA virus *Lyssa virus type 1*.

- It is a zoonotic disease transmitted to humans by bite/lick/scratch of infected animals (*commonly dogs*).
- It is an ancient disease mentioned even in Vedas. Rabies word is derived from Sanskrit word *'Rabhas'* means 'to do violence'.
- *Celsus* found relation of saliva of infected dog to human disease in 1st century AD. *Louis Pasteur* developed 1st vaccine against Rabies in 1885.
- Rabies is uncommon in developed countries. It is mainly seen in Indian subcontinent (80%) and Africa. In India it is not seen in Lakshadweep, Andaman and Nicobar islands. If disease is not seen for 2 years in humans and animals then that area is termed as *rabies free*. Maldives country does not have human or animal rabies.

Pathogenesis

- There is no predilection for age or sex even though it is observed more in children and adult males.
- 95% of rabies develops due to *bite of rabid dog* occasionally cat (**Urban rabies**). Other animals that can transmit rabies are monkey, horse, fox, cows and buffaloes, donkey, pig, sheep, camel, elephant, mongoose, jackal, bear (**Wild life/sylvatic rabies**). In India transmission is not observed through bats, rodents and birds. **Bat rabies** (vampire bat) is seen in parts of USA and Latin American countries.
- Asymptomatic carrier stage occurs only in animals but they are unlikely to be infective. Only symptomatic animals are considered to be infective.
- Rabies virus is bullet shaped envelop virus (75 nm × 180 nm) with numerous glycoprotein spikes to help in attachment of virus and also to induce antibodies. Natural occurring rabies virus is called as **street virus** which shows long incubation period of 20-60 days. Serial passage of this virus to brain of rabbits creates **fixed virus** which has got short incubation period of 4-6 days which does not show Negri bodies. This fixed virus which cannot multiply in extraneural tissues is inactivated to prepare vaccine.
- *Infection* commonly occurs by animal ***bite***, often by ***licks***, ***scratches***. Licks on *abraded skin* and licks on *abraded or unabraded* mucosa can cause infection. Licks are often *ignored* dangerously. Severity of infection depends on *viral load in the animal saliva* and class of wound. ***Aerosol*** transmission is found in bats or in lab workers. ***Person to person*** transmission can occur even though it is rare. Rabies may get transmitted through organ/corneal transplantation.
- Virus multiplies at the site of infection and passes (*ascends*) through the peripheral nerves into the CNS to develop **Negri bodies** in the brain leading into *fatal encephalomyelitis*.

From the brain virus *descends* into different tissues like *salivary glands*, muscles, heart, adrenals and skin. It also involves salivary glands to get secreted in the saliva to cause infection.

Clinical Features

- Incubation period is 3-6 weeks; but rarely can be up to many years.
- Prodromal symptoms like fever, headache.
- Pain, tingling sensation at the site of bite.
- Hyperexcitability and irritability; increased muscle reflexes and spasms.
- Increased salivation, sweating, lacrimation.
- Hydrophobia (fear of water) and **aerophobia** (fear of air) is **pathognomonic.**
- Mental instability, dilatation of pupils.
- Symptoms are aggravated by swallowing water or blowing air on them.
- Once disease starts, patient die in 72 hours.

Fear of water is seen only in affected human beings, not in animals.

Classification of Wounds

Class I: Touching or feeding the diseased animal, lick over intact skin or scratches without oozing of blood.

Class II: Licks on broken skin, scratches with blood ooze, and all bites except over head, face, palms and fingers. Minor wounds less than five in number.

Class III: All bites over head, face, palms and fingers, lacerated wounds, wounds more than five in number, wild animal bites, *and contamination of mucous membrane* with saliva.

Indications for Antirabies Vaccination

- All rabid animal bites.
- If animal is killed or dies during 10 days of observation period.
- Bite by an unidentified animal.
- If lab tests in animal show positive for rabies.
- All wild animal bites.

Vaccines for Rabies

1. Nervous Tissue Vaccine

(a) BPL inactivated vaccine: It is nervous tissue vaccine. It is 5% emulsion of the infected brain of the sheep containing the inactivated fixed virus. It is *Semple vaccine*.

Dosage of semple vaccine (as recommended by Pasteur Institute, Coonoor)

	Adult	*Children*	*Duration*
Class I	2 ml	1 ml	7 days
Class II	3 ml	3 ml	10 days
Class III	5 ml	3 ml	10 days

Mode of administration: Subcutaneously into the abdominal wall using long needle. Joseph Meister received first anti-rabies vaccine (ARV) in 1885. In Olden days, it was given for 15-21 days.

Antibody develops in 7-30 days. Protection lasts only for 6 months. Booster doses are given if needed.

Side effects

- Headache, palpitation, allergic reactions.
- Redness, tenderness and swelling at the site of the vaccination.
- *Post-vaccinal neuroparalysis*—a dangerous life-threatening complication.
- During therapy patient *should avoid alcohol and steroids*.

(b) Nervous tissue vaccine derived from suckling mouse (less than 9 days old) brain *(Fuezalida vaccine)* is used to reduce neuroparalytic complications as suckling mouse has low myelin neuron. It is given for 10 daily doses then on 20th and 30th day.

2. Avian Vaccines—Duck Embryo Vaccine (DEV)

It has got less neuroparalytic side effects. It can cause egg protein allergy. Purified duck embryo vaccine (PDEV-1 ml) is available in India.

3. Cell Culture Vaccines

They are more potent and safer.

- **Human Diploid Cell Vaccine (HDCV-1 ml):** Safest vaccine. It is prepared using fixed virus in human diploid fibroblast cells. But it is costly. It is available in India.
- *Second generation tissue culture vaccines*: They are potent and cost-effective. They are derived from nonhuman base sources. Examples are chick embryo fibroblast **(Purified Chick Embryo Cell Culture Vaccines—PCECV-1 ml)**, foetal bovine kidney, hamster kidney cells, vero cells **(Purified Vero cell Rabies Vaccine—PVRV-0.5 ml).**

Dosage: 2.5 IU in one ml. One ml is given IM into the deltoid on 0, 3, 7, 14, 28 and 90 (optional) days.

Side effects: Headache, redness at the site, fever. No other serious side effects.

Second generation tissue culture vaccine can be given intradermally also. Intradermal dose is one fifth of the intramuscular dose (0.1 ml).

Wound Treatment

- *Proper local wound care* reduces the chances of rabies infection by 80%. Immediate cleaning and washing of the wound with running water for 15 minutes is essential to reduce the viral load at the wound site. If soap is available soap water is also used. It is better to wash with warm water if available.
- Wound should be *cleaned with virucidal agents* like alcohol, tincture, povidone iodine. Savlon or carbolic acid or nitric acid should not be used.

- Wounds *should not be closed*. ARS should be injected locally. In deep wounds it may be closed only after 48 hours with loose sutures after thorough washing.
- ARS (horse or human) *should be injected* to all wounds locally.
- One *should not* scrub the wound.
- One *should not touch* the wound with *bare hands*. One should wear gloves to touch the wound.

Passive Immunity

It is used in all severe exposures, class II and class III and in all wild animal exposure. Present recommendation is injection of ARS with vaccine in all exposed patient irrespective of the class.

Types of Antirabies Serum (ARS)

Horse antirabies serum (horse/equine ARS): It is given on first day with a dose of 40 IU/ kg body weight (maximum up to 3000/units). Half is given into the wound and another half given into the gluteal muscle (IM)—single dose. It prevents the multiplication of the virus at the wound site. It also prolongs the incubation period. Passive immunity should always be combined with vaccine therapy. ARS needs test dose prior to injection of full dose. *Side effects*: Serum sickness, anaphylaxis. **Human rabies immunoglobulin (**HRIG). Dose is 20 units/kg body weight. Part is injected into the wound remaining part into the gluteal muscle (IM)—single dose. Patient should be immunised actively along with serum with additional booster doses. Side effects are rare here.

Post-exposure Prophylaxis

Cell culture and purified duck embryo vaccines are used as they are safe and efficacious.

All vaccines should be given to deltoid region (never to gluteal region as due to high fat content vaccine won't get absorbed into circulation rapidly and so immune response may not be optimum). Vaccines should be stored at 4-8°C after reconstitution and should be used immediately.

Mode of Injection

a. Intramuscular into deltoid region—**Essen regimen.** It is commonly used and technically easier but higher dose is required compared to intradermal. It is injected at a schedule of 0, 3, 7, 14, and 28 days and booster at 90 days. *First dose should be combined with ARS preferably HRIG.* Multisite IM regime is often used as follows—first dose on day 0 two doses of IM vaccine is injected one on each side deltoid. Later single doses on 7 and 21 days (as 0 {2}, 7 {1}, 21 {1}).
b. *Intradermal route*: (1) *Two site intradermal method* is used. 1/5th of the IM dose of selected vaccine is used. Two sites on the day 0, 3, 7 and one site on the days 28 and 90 . PVRV 0.1 ml; PCECV 0.2 ml; PDEV 0.2 ml. (2) *Eight site intradermal method* is used. On day 0, eight sites intradermal injections at both deltoids, both suprascapular, both thighs, both lower quadrants of abdomen are given. On day 7, on 4 sites—both deltoids, both thighs intradermal injections are given. On days 28 and 90 one dose on each day intradermal vaccine is injected at one site. HDCV 0.2 ml is used. It is like 0 (8); 7 (4); 28 (1); 90 (1). In whatever type, on first day (0) rabies immunoglobulin should be injected locally as well as IM.

Postexposure vaccination if individual *has been vaccinated earlier*: Doses on days 0, 3, and 7 are given. But ideally assessed by serum antibody level (should be more than 0.5 IU/ml. Passive immunity is not given in individuals who had vaccination earlier.

Pre-exposure Prophylaxis

It is given to veterinarians, animal handlers. Dose: I ml of cell culture vaccine IM or 0.1 ml intradermally on days 0, 7, 28. Serum titre for antibodies should be assessed after 1 month. If it is less than 0.5 IU/ml then one booster dose is injected. Booster doses are given once in every 2 years.

RABIES IN DOGS

Incubation period: 10 days to 8 weeks.

Types

a. ***Furious type***: Here dogs are aggressive like a mad dog. Dog changes its behaviour with loss of fear of human beings; bites objects, eat, mud, etc. Running amok, voice change, inability to bark properly, excessive salivation and foaming, paralysis and death.
b. ***Dumb type***: Dog is paralytic and sleepy. There is no aggressiveness at all. Dog sleeps for 3 days and dies.

Once symptoms of rabies develop in a dog *it rarely survives more than a week.*

Dog brain is sent for study in 50% glycerol-saline solution.

Lab Tests to Confirm Rabies in Dogs/Animals

- *Fluorescent antibody test (FRA test)*: It is reliable test. If FRA test is negative in brain of animal then even if dog is rabid its saliva does not contain virus. FRA test is positive at any stage of the disease.
- *Microscopic* examination of the brain of the infected dead animal to look for NEGRI bodies. It is seen in 90% of dead rabid dogs.
- *Mouse inoculation test* is very sensitive test. 10% brain tissue emulsion in saline is centrifuged at 2000 rpm for 10 minutes; 0.03 ml top fluid is injected intracerebrally into the suckling mouse to demonstrate rabies in 8 days in mouse.
- Corneal test is simpler but negative result does not rule out the infection possibility.

Immunisation in Animals

a. *BPL inactivated nervous tissue vaccine* (20% infected sheep brain suspension): Single dose 5 ml to dogs; 3 ml to cats. Second dose after 6 months. Then once a year regularly.
b. *Modified live virus vaccine* (33% infected chick embryo suspension): Dose—3 ml single dose which is repeated once in 3 years.
c. *Oral vaccines* are used successfully to control wild foxes in Canada by placing vaccine in food through baits.

Remember about rabies

- Control of stray dogs and immunisation of all dogs will reduce the incidence of rabies
- Prevention is the only way in rabies. Established rabies cannot be treated—100% mortality
- HDCV and PCECV dose is 1 ml; PVRV is 0.5 ml—IM into deltoid or anterolateral aspect of thigh in children (NEVER GLUTEAL region)
- Day 0 is the day of first date of vaccination not day of bite
- Interchanging of vaccines is even though acceptable but not well recommended and ideally should be avoided
- Vaccine dose is same in all age groups
- Reconstituted vaccine should be used immediately
- Vaccine dose may be doubled in first dose if ARS is not used in bites of face, head, hands and genitals, in malnourished patients, in patients who are on steroids, antimalignancy drugs, antimalarials; in pregnancy, lactation, infants, elderly, HIV and immunosuppressed individuals
- Consumption of unboiled milk of a rabid cattle amounts for class III bite and needs prophylaxis vaccine
- Pet dog should be examined periodically by veterinary doctor. Usual vaccination method used in dog is—at 3rd month of age, 1 month later, and later yearly booster doses. 0, 6 months and then yearly is also used. Pet dog should be prevented from coming into contact with stray dogs
- Dietary or alcohol restriction is not needed during vaccination as it will not alter immune response
- Concomitant other vaccine injection along with rabies vaccine can be done but at different injection site as there is no interference with immune response
- There is no contraindication for rabies vaccine as it is life-saving method
- All cases of dog bite should receive initial vaccine and ARS immediately. If the dog remains healthy even after 10 days of observation period vaccine is discontinued. Virus can be present in saliva 3 days before onset; once symptom begins dog cannot live for more than 4 days. Safe period of 3 days is added and so total 10 days is observed
- Virus can present in semen of a rabid man. So if he had contact with his wife within 5 days prior to symptoms then wife should be vaccinated as class III with HRIG
- If by mistake vaccine is given to gluteal region, fresh full course of vaccine to deltoid region should be started
- There are no single shot vaccines available. It is a myth
- Antiviral drugs are of no use
- Modern CCVs are very safe and efficacious
- Intradermal route of 0.1/0.2 ml is the best route
- Equinus ARS is cheaper and equally effective; but test dose should be given to prevent reaction
- Rabid dog *will never* have hydrophobia. Rabid dogs can swim through water or can even drink water. Aerophobia and photophobia are present in rabid dog. Hydrophobia is observed *only in* human rabies
- Bite by cats and cows also should be vaccinated
- Steroids and antimalarials should be avoided during vaccination as it may alter the immune response

ANTHRAX

- It is caused by *Bacillus anthracis*, which is a *Gram-positive*, aerobic, spore forming, capsulated, nonmotile, nonacid-fast bacillus and is *resistant to heat and antiseptics*.
- Disease is common in cattle and seen in people who handle carcasses, wool, hairs.
- It is often used in *biological war*.

Types

a. **Cutaneous type** (*Hide porter's disease*)
  - It is the most common type and occurs within 3-4 days after infection.
  - Indurated papule with black slough surrounded by vesicles—***malignant pustule.*** Itching is common in papule—black colour eschar (Anthrax means *charcoal*).
  - Regional lymph nodes are involved.
  - ***Toxaemia*** is common.

b. **Respiratory type** (*Wool Sorter's disease*) is due to inhalation of spores, causing ***haemorrhagic pneumonia***. It is more dangerous and life-threatening.

c. **Alimentary type** due to ingestion of spores.
  - Fatal *septicaemia and* meningitis can occur in any type.

Diagnosis

- Culture of fluid will show ***Medusa head appearance.***
- It shows positive M'Fadyean's reaction and positive Ascoli's thermoprecipitation test.

Treatment

- *Ciprofloxacillin, doxycycline*, penicillins.
- Alum precipitated Anthrax toxoid is used in humans.
- Scalvo's serum prepared by active immunisation of asses are used.

NOSOCOMIAL AND OPPORTUNISTIC INFECTIONS

Nosocomial Infection (Hospital Acquired Infection)

It is an infection acquired because of *hospital stay.*

Sources

- Contaminated infected wounds.
- Urinary tract infections.
- Respiratory tract infections.
- Opportunistic infections.
- Abdominal wounds with severe sepsis.

Spread can occur from one patient to another, through nurses or hospital staff who fail to practice strict asepsis.

It is more common in:

- Diabetics
- Immunosuppressed individuals
- Patients on steroid therapy and life-supporting machines
- Instrumentations (indwelling catheter, IV cannula, tracheostomy tube)
- Patients with artificial prosthesis

Organisms

- *Staphylococcus aureus* is the most common organism causing hospital acquired wound infection. Others are *Pseudomonas, Klebsiella, E. coli, Proteus.*
- *Streptococcus pneumoniae, Haemophilus*, Herpes, *Varicella, Aspergillus, Pneumocystis carinii* are the most common pathogens involved in hospital acquired respiratory tract infection which spreads through droplets.
- *Klebsiella* is the most common pathogen involved in hospital acquired UTI which is highly resistant to drugs.

Management

Most of the time, organisms involved are *multidrug resistant, virulent* and hence, cause *severe sepsis*.

- Antibiotics.
- Isolation.
- Blood, urine, pus for culture and sensitivity to isolate the organisms.
- Blood transfusion, plasma or albumin therapy.
- Ventilator support.
- Maintaining optimum urine output.
- Nutritional support.

Prevention

- Isolation of patients with badly infected open wounds, severe RTI/UTI.
- Following strict aseptic measures in OT and in ward by hospital attendants.
- Proper cleaning and use of disinfectant lotions and sprays for bedpans, toilets and floor.
- The precipitating causes has to be treated, along with caring for proper nutrition and improving the anaemic status by blood transfusion.

Opportunistic Infections

They are normally of *low pathogenicity,* occur through therapeutic invasive procedures and are common in immune deficiency status.

Immune deficiency may be due to:

- Diabetes
- HIV
- Steroid therapy
- Radiotherapy
- Immunosuppressive therapy in transplantation
- Cytotoxic chemotherapy
- Starvation and old age

Therapeutic invasive procedures may be in the form of IV cannula, bladder catheter, tracheostomy and other minor surgical procedures which permit the skin organisms like *Staphylococcus epidermidis* to penetrate the skin and invade the deeper tissues.

Organisms

Bacteria: *Gram-negative: E. coli, Pseudomonas, Klebsiella, Proteus, Serratia.*
Gram-positive: *Staphylococcus epidermidis, Streptococcus pneumoniae.*
Viruses: *Herpes, CMV, Varicella zoster*, may cause fatal pneumonia.
Fungal: *Candida, Aspergillus,* yeast.
Protozoal: *Cryptosporidia* (causes diarrhoea), *Pneumocystis carinii.*

Because of the poor defence mechanism, infection is severe and often life-threatening.

Investigations

- Swab culture, blood culture, pus culture.

Treatment

- These infections are difficult to treat as they are often *multidrug resistant.*
- Combination of broad spectrum antibiotics—cephalosporins, aminoglycosides, metronidazole are given.
- Depending on culture and sensitivity appropriate antibiotics are given.
- Often ventilatory support and critical care are necessary.

HIV INFECTION AND AIDS (Acquired Immunodeficiency Syndrome)

> *Even if I'd made no money, even if it was the pariah of specialties by virtue of its lack of procedures, an unexpected fringe benefit had become evident with the appearance of AIDS: In those early days, dealing with AIDS made us an elite group, an unexpectedly glamourous group. Even the cardiac surgeons could not approach our kind of heroism. Yes, they dealt with death every day. But it was somebody else's death they had to worry about. Never their own.*
>
> —**Abraham Verghese,** 1994

History

- 1983: Discovery of the virus
- First case of AIDS detected in UK
- 1984: Development of an antibody test

Definition of AIDS

Confirmed HIV infection with CD_4 T lymphocyte count $< 0.2 \times 10^6$/L with symptoms.

Human Immunodeficiency Virus

It was discovered by *Barre-Sinoussi and Montagnier in 1983.*

Types

It is classified under HTLV Type III. HIV is subdivided into Type 1 and Type 2.

They are *retroviruses*.

Clinical classification of HIV Infection

- Acute infection
- Asymptomatic but positive HIV
- Persistent generalised lymphadenopathy
- AIDS (HIV related diseases)
  - Constitutional diseases like weight loss, fever, diarrhoea
  - Neurological diseases, dementia, neuropathy, myelopathy
  - Opportunistic infections
  - Malignancies. Kaposi's sarcoma, non-Hodgkin's lymphomas, primary cerebral lymphomas
  - Other diseases attributable to HIV infection

Mode of Transmission

- Sexual intercourse—vaginal or anal.
- Needle pricks—using unsterilised needles for injections, in IV drug abusers, careless handling.
- Mother to child—during birth through vaginal secretion, transplacental, through breast milk.
- Blood transfusions, organ transplantations.

Disease is common in Africa and Asian countries.

HIV mainly harbours in semen, genital secretions, blood, pus, sputum, saliva and other body fluids.

Anti-HIV antibody

- It appears about 3 weeks to 3 months after exposure
- It always indicates infection
- It has got weak neutralising capacity
- It persists throughout the HIV infection

Tests for HIV

1. ELISA test (**screening test**).
2. Western blot test (**diagnostic test**).
3. Polymerase Chain Reaction (PCR).
4. Anti-HIV antibody detection.
5. *Viraemia quantification*—to start treatment and to see the response of antiviral drugs (useful if it is within 0.5 log 10).
6. *CD_{4+} count*
   - Normal value > 500/mm^3.
   - Values between 200-500/mm^3 is seen in Kaposi sarcoma, *Candida* infection, *Mycobacterium tuberculosis.*
   - Values between 50-200/mm^3 is seen in *Pneumocystis carinii* and *Toxoplasma* infections.
   - Values < 50/mm^3 is seen in atypical mycobacteria, cytomegalovirus, lymphomas.

After HIV infection, there is a time gap for the patient to become reactive to tests. This time gap is called as *"Window period"*. This period is variable. But during this period, the individual is infective.

Pathogenesis

Envelope glycoprotein of HIV binds with the surface molecule CD_4 of 'T' lymphocytes, monocytes, macrophages, cutaneous Langerhan's cells, dendritic cells of all tissues.

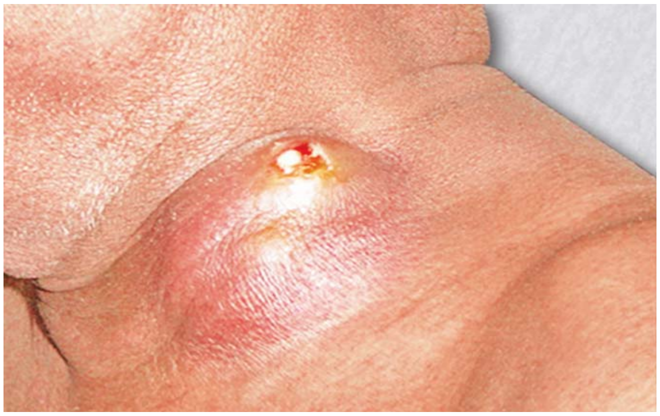

Fig. 1.132: HIV patient presented with tuberculous cold abscess with discharging sinus.

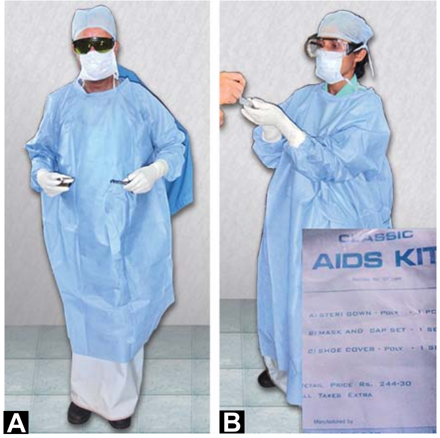
A
B
CLASSIC
AIDS KIT
A: STERI GOWN - POLY - 1 PC
B: MASK AND CAP SET - 1 SET
C: SHOE COVER - POLY - 1 SET
RETAIL PRICE Rs. 244-30
ALL TAXES EXTRA

Figs 1.133A and B: Picture showing HIV kit and wearing of the HIV kit as a precaution while doing surgery for HIV infected individuals.

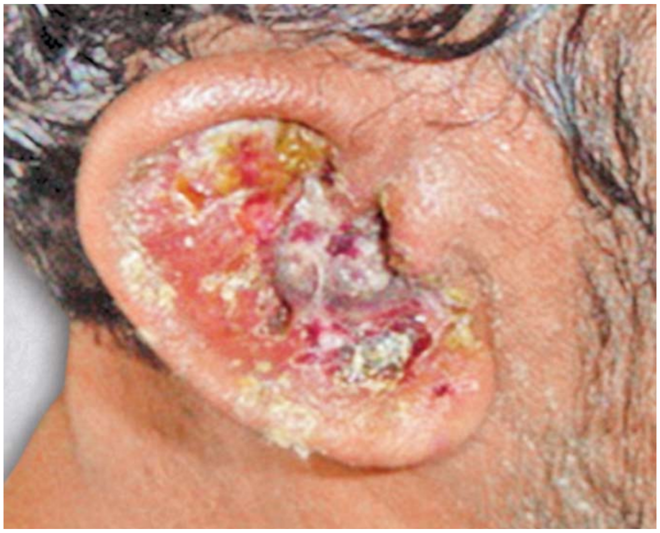

Fig. 1.134: Severe infection of ear cartilage in a HIV infected patient.

CD_4 of lymphocytes—T helper cells induce and control the normal immune response.

HIV suppresses immune response completely directly by suppressing 'T' cell, indirectly by suppressing 'B' cell. Finally it dismantles and destroys the immune system making the individual prone to opportunistic infections.

The fragrance always remains in the hand that gives the rose.

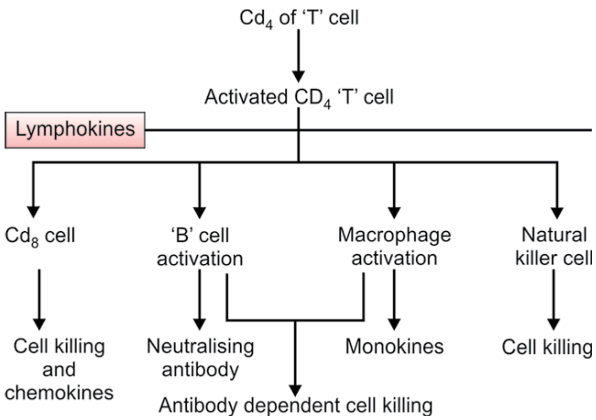

After HIV infection, antibodies develop to virus envelope and core proteins which persists throughout life.

General Features in HIV

- Weight loss more than 10%.
- Fever more than 1 month.
- Diarrhoea more than 1 month.
- Neuralgia, arthralgia, headache.
- Lymphadenopathy.
- Cutaneous rashes, dermatitis, fungal (*Candida*), bacterial, viral (herpes simplex 1 and 2) infection.
- Dental infection, gingivitis, candidiasis of oral cavity and oesophagus.
- Varicella zoster infection.
- Opportunistic infections.
- Poor healing after surgery, trauma, infection with more complications.

Tumours in HIV Infection

- Kaposi's sarcoma—40% common.
- Lymphomas (NHL common) (3-4%).
- Cervical cancer.
- CNS lymphomas.
- Ano-genital squamous cell carcinoma.
- Testicular tumours (Germ cell types).
- Lung cancer.
- GIT lymphomas and adenocarcinomas.
- Squamous cell carcinoma of anal canal and cervix.

Pulmonary Problems in HIV Infection

- Pneumonia.
- Tuberculosis.
- Fungal infections.
- *Pneumocystis carinii pneumonia.*
- Cytomegalovirus pneumonia.

GIT Problems in HIV Infection

- GIT infections—bacterial, protozoal, viral.
- Kaposi's sarcoma, lymphomas, adenocarcinomas.
- Hepatitis ('C' virus), cholestasis.
- Anorectal diseases.
- Abdominal tuberculosis.

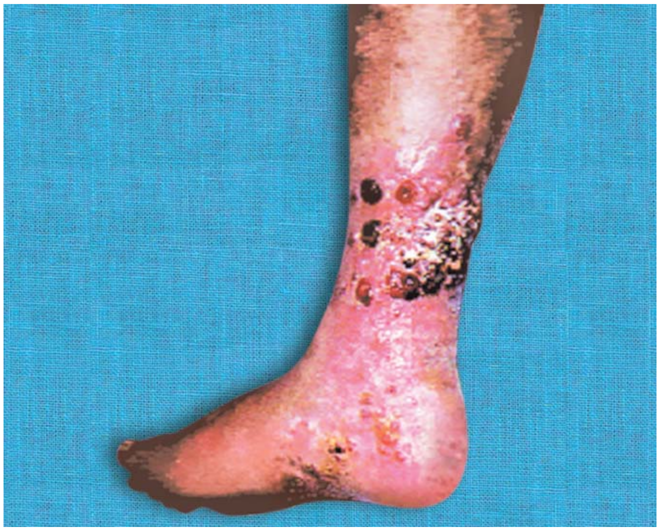

Fig. 1.135: Kaposi's sarcoma.

Neurological Problems in HIV Infection

- Encephalitis, aseptic meningitis, myelitis.
- Neuropathies with demyelination.
- Opportunistic infections like *Toxoplasma, Cryptococcus* causing severe meningitis.
- Primary CNS lymphomas.
- CNS tuberculosis (Tuberculomas).
- Visual problems.

Management

Investigations

- Tests for HIV.
- Tests for specific and opportunistic infections.
- Tests relevant for associated tumours.

Treatment

1. Antiviral therapy:
    - *Nucleoside reverse transcriptase inhibitor (NRTI):* Zidovudine, didanosine, abacavir, lamivudine, stavudine.
    - *Non-nucleoside reverse transcriptase inhibitor (NNRTI)*: Nevirapine, delavirdine.
    - *Protease inhibitors*: Ritonavir, indinavir, amprenavir.
2. Treatment of **opportunistic infections**.
3. Treatment of **tumours.**
4. Immunotherapy:
    - Alpha and gamma interferons.
    - Interleukins.
5. Bone marrow transplantation.
6. Anti-CD_3 or IL-2 after **HAART** (**H**ighly **A**ctive **A**nti-**R**etroviral **T**herapy).
7. ***Psychotherapy***.
8. Counselling of HIV patients and their families.
9. Life-expectancy after initial HIV infection is 8-10 years.

Prevention

Continues to be our best weapon in combating the menace of HIV infection.

- Safe sex. Condom usage reduces the risk of transmission.
- Health education.
- Use of disposable needles to prevent infections.

Universal precautions against HIV

- Care in handling sharp objects like needles, blades
- All cuts and abrasions in an HIV patient should be covered with a waterproof dressing
- Minimal parenteral injections
- Equipments and areas which are contaminated with secretions should be wiped with sodium hypochlorite solution or 2% glutaraldehyde
- Contaminated gloves, cottons should be incinerated
- Equipments should be disinfected with glutaraldehyde
- Disposable equipments (drapes, scalpels, etc.) should be used whenever possible
- Walls and floor should be cleaned properly with soap water
- Separate operation theatre and staff to do surgeries to HIV patients is justifiable
- Avoid shaving whenever possible before surgery in HIV patients
- All people inside the theatre should wear disposable gowns, plastic aprons, goggles, overshoes and gloves
- Surgeons, assistants and scrub nurse should wear in addition double gloves
- Suction bottle should be half-filled with freshly prepared glutaraldehyde solution
- Spilled body fluids should be diluted with glutaraldehyde
- Accidental puncture area in surgeon or scrub nurse should be immediately washed with soap and water thoroughly
- Theatre should be fumigated after surgery to HIV patient

HIV, hospital and surgeon

- Isolation per se of HIV patient is not required.
- Proper care should be taken to prevent transmission of the virus.
- Open wounds, disposal of excreta, fluids, discharge, pus and other infective materials should be taken care of properly.
- Risk of HIV infection through needle prick is very less (0.03%).

Following measures should be taken while managing HIV patients:

- Wearing double gloves.
- Wearing proper spectacles (as HIV can get transmitted through eyes directly).
- Wearing proper head mask, theatre shoes, apron.
- Measures to prevent spread of infection from patient to patient in the hospital.
- Disposal of needles through a sharp disposing container.

Disinfection

- Autoclave is ideal.
- Boiling.
- Sodium hypochlorite solution.
- 2% glutaraldehyde solution.

NECROTISING FASCIITIS

- It is spreading inflammation of the skin, deep fascia and soft tissues with extensive destruction, toxaemia commonly due to *Streptococcus pyogenes* infection, but often due to mixed infections like anaerobes, coliforms, Gram-negative organisms.
- It is common in old age, smoking, diabetics, immunosuppressed, malnourished, obesity, steroid therapy and HIV patients. Trauma is a common precipitating factor/cause—80%.
- It can occur in limbs, lower abdomen (Meleney's infection), groin, perineum. There is acute inflammatory response, oedema, extensive necrosis and cutaneous microvasculature thrombosis.
- Muscle is usually *not involved* in necrotising fasciitis.

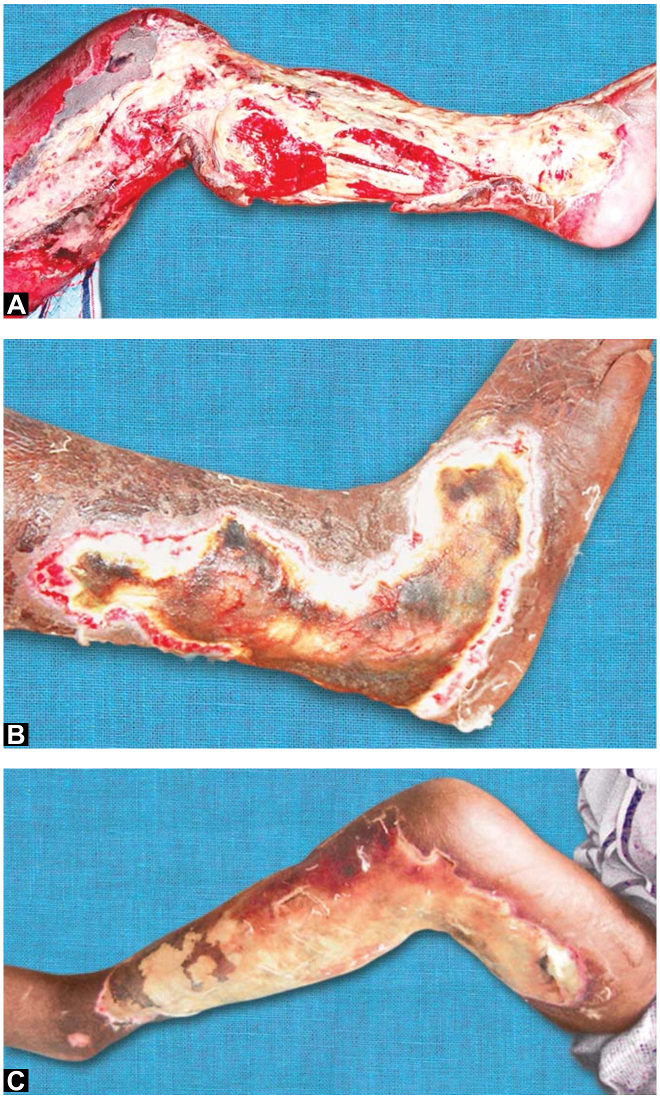

Figs 1.136A to C: Necrotising fasciitis of skin, deep fascia and soft tissues without involvement of muscle.

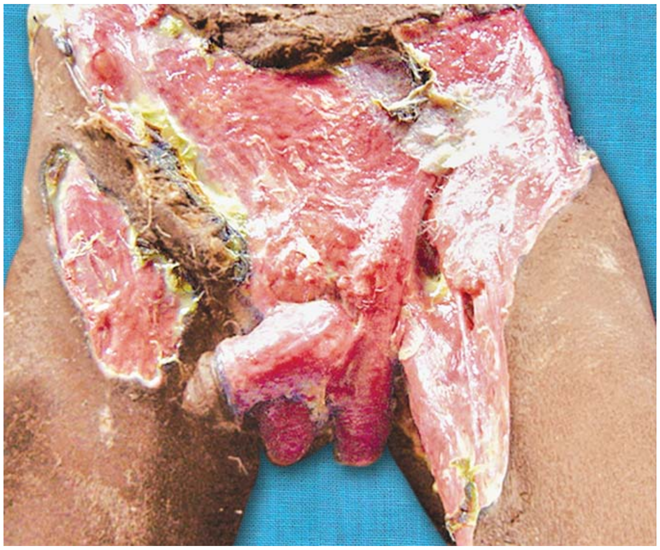

Fig. 1.137: Fournier's gangrene. Note the destruction of skin over penis, scrotum and lower abdominal wall.

Types

Type I—It is due to mixed infection.

Type II—It is due to *Streptococcus pyogenes*, usually due to minor trauma like abrasions.

Clinical Features

- Sudden swelling and pain in the part with oedema, discoloration, necrotic areas, ulceration.
- Foul smelled discharge.
- Features of toxaemia with high-grade fever and chills, hypotension.
- Oliguria often with acute renal failure due to acute tubular necrosis.
- Jaundice.
- Rapid spread in short period (in few hours).
- Features of SIRS, MODS with drowsy, ill-patient.
- Condition if not treated properly may be life-threatening.

Management

- IV fluids, fresh blood transfusion.
- Antibiotics depend on C/S or broad-spectrum antibiotics. High dose penicillins are very effective. Clindamycin, third generation cephalosporins, aminoglycosides are also often needed.
- Catheterisation and monitoring of hourly urine output.
- Haematocrit, serum creatinine assessment.
- Pus culture, blood culture.
- Electrolyte management and monitoring.
- Control of diabetes, if patient is diabetic.
- Oxygen, ventilator support, dopamine, dobutamine supplements whenever required.
- Radical wound excision of gangrenous skin and necrosed tissues at repeated intervals.

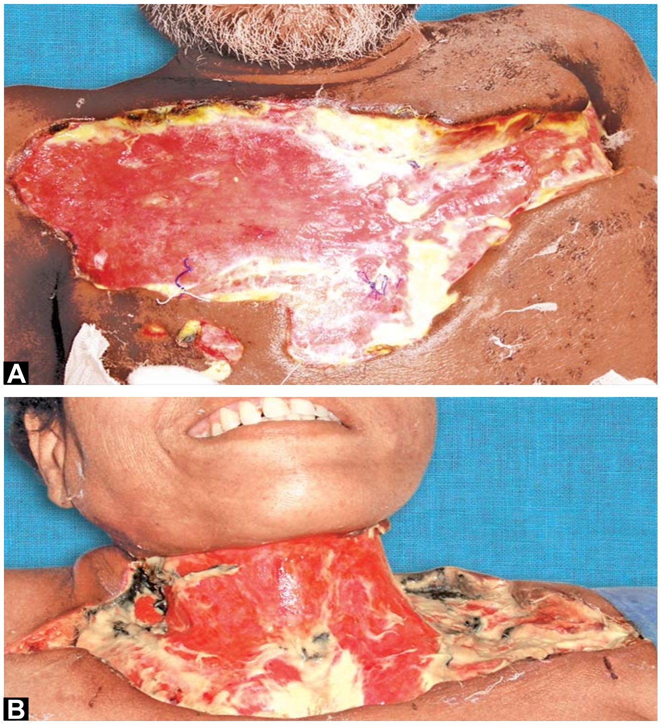

Figs 1.138A and B: Necrotising fasciitis over chest wall and neck in two different patients.

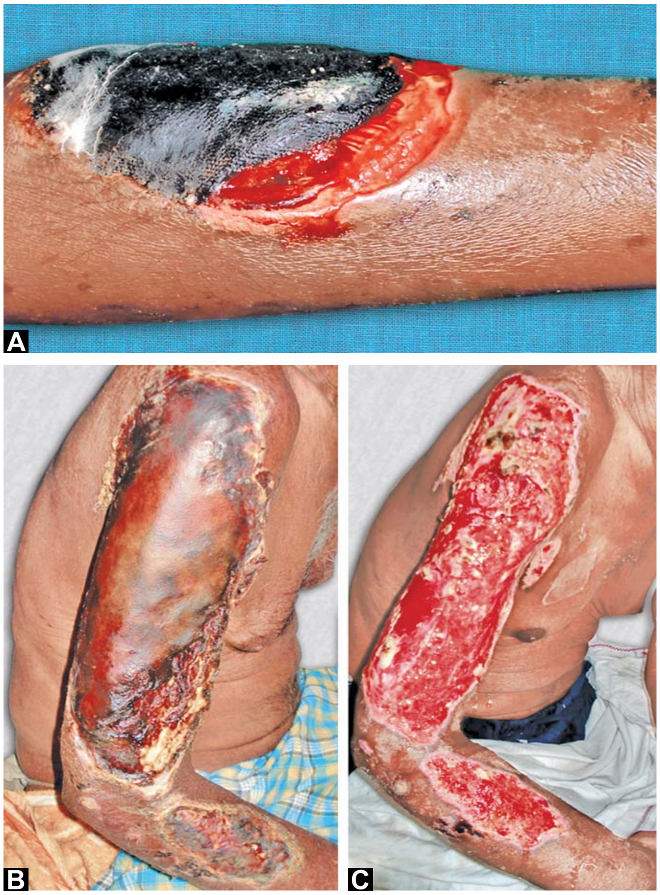

Figs 1.139A to C: The typical necrotising fasciitis showing gangrenous skin. In second photo, arm is extensively involved. Also note the granulating area after extensive wound excision of the necrotic skin (debridement).

Necrotising fasciitis

- 80% are polymicrobial—*streptococci, staphylococci, E. coli, Pseudomonas, Proteus, Clostridium*
- It is a surgical emergency condition as it is very rapidly progressive
- Lower limb is the most common site—60%
- Oedema beyond erythema, woody hard texture on palpation
- Crepitus with subcutaneous emphysema, skin vesicles, dish water pus with grayish discharge are common
- Lymphangitis is usually absent
- Pink/orange skin stain and later focal skin gangrene
- Shock, multiorgan failure
- Resuscitation, wound excision, antibiotics, critical care (oxygen, intubation, ventilator) is needed
- Hyperbaric oxygen given in high pressure chamber with 100% oxygen in 2-3 atmospheric pressure reduces the mortality to 10-20%. It is bactericidal and promotes the neutrophil function
- In spite of adequate therapy mortality is 30-50% or more

- Vacuum assisted dressing is better.
- Once patient recovers and healthy granulation tissue appears, spilt skin grafting is done. As it commonly involves large area, mesh graft (meshing of SSG) is needed.

ACUTE PYOMYOSITIS

- It is infection and suppuration with destruction of the skeletal muscle, commonly due to *Staphylococcus aureus* (90%) and *Streptococcus pyogenes*, occasionally due to Gram-negative organisms.
- It is common in muscles of thigh, gluteal region, shoulder and arm.
- Precipitating factors are similar to necrotising fasciitis—trauma, malnutrition, anaemia, and immunosuppression.
- Pain, oedema, tenderness over the site with apparently normal overlying skin.
- Induration and muscle spasm is typical.
- Fever, jaundice, uraemia (*acute renal failure*) are common.

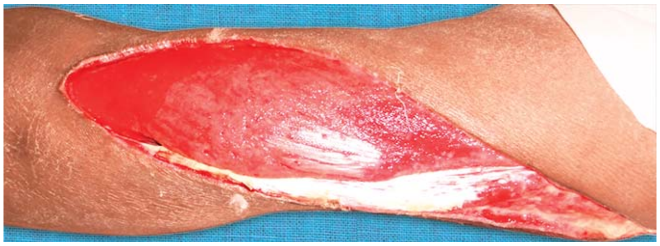

Fig. 1.140: Pyomyositis of thigh muscle.

Management

- *Creatine phosphokinase* will be very high and signifies acute phase (more than 50,000 units); due to *rhabdomyolysis*.
- MRI is useful. US guided pus aspiration is also done.
- *Treatment* is antibiotics, radical wound excision with removal of pus and all necrosed muscle and compartment release. *Haemodialysis* is needed until recovery if there is renal failure. Later secondary suturing or skin grafting is needed.

SURGICAL SITE INFECTION (SSI)

Surgical site infection is the second most common complication following surgical procedures (*first being postoperative pneumonia*) due to virulent bacterial entry, altered wound microenvironment, and changed host defense. Prevention of SSI can be achieved by better preoperative preparation; proper infection control during surgery; adherence to principles of preventive antibiotic therapy; better surgical techniques to reduce hematoma, tissue injury and foreign bodies within the surgical site; prevention of tissue hypoxia with enhanced oxygen support.

Common Sources of Infection

- Surgical wards, wounds, ulcers, catheters, drains, sputum, urine, faeces, open wounds.
- Operation room without proper ventilation, nurses, surgeons. Operation methods, sterilisation of instruments.

Organisms Causing SSI

- Commonly *Staphylococcus aureus*. Any organisms like clostridia, Gram-negative bacteria can cause SSI.
- Bacteria present in a wound with no signs or symptoms of systemic inflammation is called as *colonization,* usually less than 10^5cfu/ml. Transient exposure of a wound to bacteria (usually less than 6 hours) is called as *contamination* with varying concentration.

Sequence of Events (in Surgical Wounds)

- *Activation* of inflammation occurs by cuts, incisions, abrasions, burns. This initiates *inflammation* by protein coagulation, platelet aggregation, mast cell activity, release of complements and bradykinin. *Phase I of inflammation* begins with vasodilatation, increased bulk flow, increased vascularity. Later *Phase II of inflammation* proceeds with phagocytic infiltration and bacterial phagocytosis, removal of dead tissue with release of proinflammatory cytokines. Here tissue injury from incision mobilizes phagocytes before bacterial contamination leading into prior preparation against infection. If contamination is controlled monocytes activate to *regulate wound healing* using myofibrocytes and collagen.
- If bacterial *contamination is not controlled*, proinflammatory cells release TNF-α to stimulate neutrophils for phagocytosis. It also causes release of reactive oxygen and acid hydrolases from lysosomal vacuoles to result in lipid peroxidation, release of interleukins, evoking acute inflammatory response with creation of space containing pus which contains necrotic tissue, neutrophils, bacteria and proteinaceous fluid with *all signs of inflammation*—rubor, dolor, calor, tumour. It is *typical surgical site infection* (SSI).

Factors Related to SSI

- *Bacterial entry* (inoculum) into the wound occurs through air in operation room, through instruments, through surgeons and theatre staffs, patient's endogenous bacteria like perineum, urine, etc.
- *Bacterial virulence* plays major role in causing SSI.
- *Microenvironment in the wound* like haemoglobin level at surgical site; presence of necrosis which interferes with phagocytosis; presence of dead space and or foreign body in the wound.
- *Host defenses* both natural (Innate) and acquired, when altered SSI occurs. Acquired causes are—shock, hypoxia, chronic illness, hypoalbuminemia, malnutrition, hypothermia, hyperglycemia, corticosteroids, HIV infection, malignancy and certain drugs.

Classification of Surgical Wounds

- *Clean wounds*—operative procedure does not enter into normally colonised viscus.
- *Clean-contaminated*—operation enters into a colonised viscus but under elective controlled circumstances.
- *Contaminated wounds*—gross contamination is present at the surgical site in the absence of obvious infection.
- *Dirty wounds*—surgical procedures performed when active infection is present.

Risk Classification and Identification System

It is based on three categories of variables—(1) Those that estimate intrinsic degree of microbial contamination at the surgical site. (2) Those that measure the duration of operation. (3) Host susceptibility markers.

Variables that Influence SSI

Variables that influence SSI	Point
1. An abdominal operation.	1
2. Operation lasting more than 2 hours.	1
3. Surgical site classified as contaminated or dirty/ infected.	1
4. Operation performed on a patient with more than three discharge diagnosis.	1
Total index	**4**
All variables have equal significance. This index twice better at predicting SSI than wound classification. Disadvantage is that it is not operation specific and variables collected at discharge.	

The National Nosocomial Infections Surveillance (NNIS) System as Basic SSI Risk Index

NNIS system	Point
Operation classified as contaminated or dirty.	1
The patient has an ASA (American Society of Anaesthesiology) preoperative assessment score of 3, 4, or 5.	1
Duration exceeds 75th percentile of 'T' point. 'T' point defined as length of time in hours that represents 75th percentile of procedures in NNIS survey.	1

Note:
T point for common surgical procedures are—Coronary artery bypass graft—5; Bile duct, liver or pancreatic surgery, craniotomy, head and neck surgery—4; Colonic surgery, joint prosthesis surgery, vascular surgery—3; Abdominal or vaginal hysterectomy, ventricular shunt, herniorrhaphy—2; Appendicectomy, limb amputation, caesarean section—1.

Physical Status classification

Class I: A patient in normal health.

Class II: A patient with mild systemic disease resulting in no functional limitations.

Class III: A patient with severe systemic disease that limits activity, but is not incapacitating.

Class IV: A patient with severe systemic disease that is a constant threat to life.

Class V: A moribund patient not likely to survive 24 hours.

Classification of Surgical Site Infection (SSI)

A. According to the Depth of the Wound Infection

1. Superficial incisional SSI

It occurs within 30 days of operation; involves only skin and subcutaneous tissue; and one of following:

Purulent drainage (culture documentation not required), organisms isolated from fluid/tissue of superficial incision, at least 1 sign of inflammation, wound is deliberately opened by the surgeon, surgeon or attending physician declares that the wound is infected.

A wound not considered a superficial site infection—if stitch abscess is present; if infection is at episiotomy site; burn wound, SSI extends into the fascia or muscle.

2. Deep incisional SSI

It occurs within 30 days of operation or 1 year if an implant is present; involves deep soft tissues of the incision; and at least one of the following—purulent drainage from the deep incision site without organ/space involvement, fascial dehiscence or deliberate separation by surgeon, deep abscess, identified by—reoperation/histopathology/radiology, surgeon or attending physician declares deep infection present.

3. Organ space infection

It occurs within 30 days or 1 year if an implant is present; involves anatomic structures not opened or manipulated during surgery; and one of the following—pus from a drain placed into organ/space, organism isolated by culture, identification of abscess by direct examination, reoperation, histopathology, radiology, diagnosis by surgeon or attending physician.

B. Classification of Wound Infection According to the Aetiology

a. *Primary infection* where the wound is the primary site of infection.

b. *Secondary infection* arises following a complication that is not directly related to the wound.

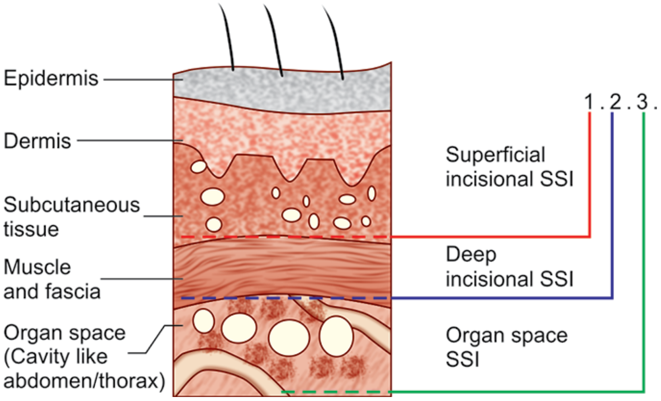

Fig. 1.141: Surgical site infection classification as—superficial incisional, deep incisional, organ space.

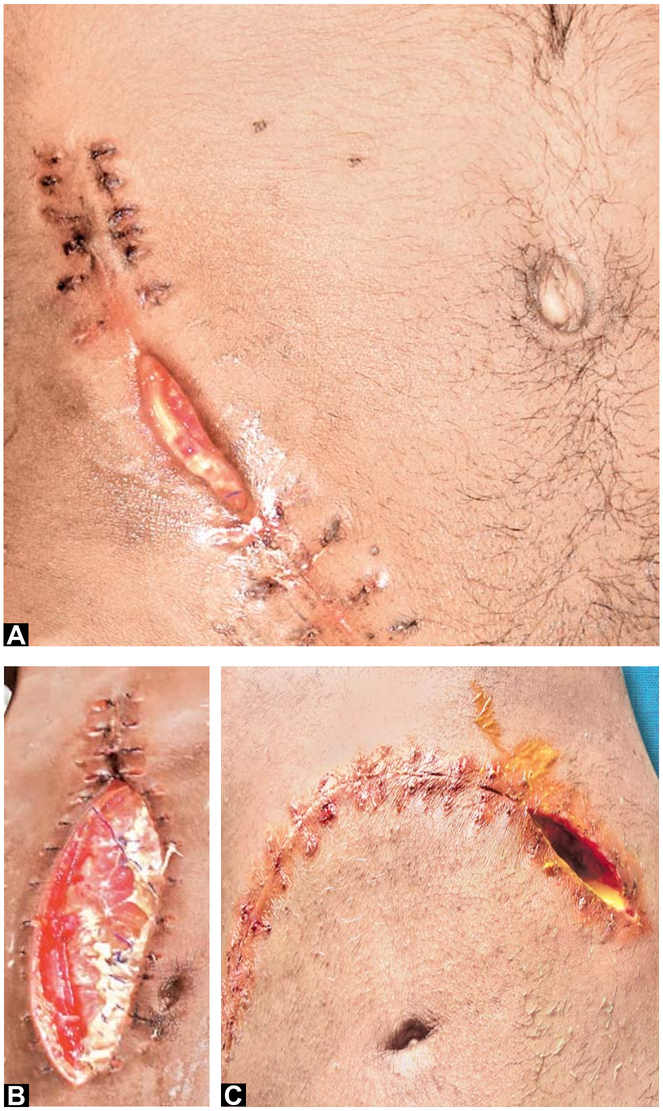

Figs 1.142A and C: Surgical site infection (SSI) types: (A) Superficial incisional SSI. (B) Deep incisional SSI. (C) Organ space SSI.

C. Classification of Wound Infections According to the Time

a. An *early infection* presents within 30 days of a surgical procedure.
b. An *intermediate infection* occurs between 1-3 months afterwards.
c. *Late infection* occurs in more than three months after surgery.

D. Classification of Wound Infections According to the Severity

a. *Minor* wound infection if there is discharge without cellulitis or deep tissue destruction.
b. *Major* if the discharge of pus is associated with tissue breakdown, partial or total dehiscence of the deep fascial layers of the wound, or if systemic illness is present.

Note:
Please refer first page of this Chapter for Southampton wound grading system and Asepsis wound score system.

Prevention of SSI

1. Preoperative

- Preoperative cleaning and antiseptic scrub of surgical site. Skin is colonised by various bacteria mainly *Staphylococcus aureus* (50%). Preoperative skin wash using chlorhexidine decreases bacterial colonisation by 80% and so wound contamination.
- Surgical site to be shaved or clipped in the operation theatre. Shaving should be done in the theatre itself or within 2 hours of beginning of the surgery otherwise infection rate may raise. Clean wound infection after shaving is 2.3%; after clipping it is 1.7%; without shaving or clipping it is 0.9%. However, selective shaving is definitely needed in area like scalp, axilla, groin, and perineum.
- Surgery should be avoided or postponed if fingers or hand of surgeon has open wounds or infection.
- Obvious infection in patient if exists should be treated.
- Prolonged preoperative admission should be avoided for an elective surgery.

2. Care in the Operation Theatre

- One should ensure that sterile caps, masks, gowns and sterile gloves are used.
- Proper skin cleaning is needed on table after anaesthesia using antiseptics like povidone iodine. One should ensure that all drapes are dry throughout the procedure and all instruments are thoroughly sterilised.
- Unimpregnated plastic drapes are avoided as it is found that it does not have any advantage.

- Gentle tissue handling, absolute haemostasis, holding tissues using instruments as much as possible, using appropriate suture materials, avoiding dead space during closure are certain essential on table tips to reduce SSI.
- One should consider leaving wounds open if it is severely contaminated.

3. *Preventive Antibiotic Therapy*

- It is used whenever high-risk of infection is associated with the procedure and consequences of infection if possibly severe and if patient has a high NNIS risk index.
- Antibiotics should be administered as close to the incision time as possible, before induction of anaesthesia.
- Selected antibiotic should have activity against likely pathogens.
- Postoperative systemic antibiotics for 24 hours (beyond 24 hours not shown to reduce SSI).
- Benefit of preoperative antibiotics in NNIS risk 0 index is difficult to assess and quantify.
- Proper techniques and wound microenvironment are more important than antibiotics.
- Preventive systemic antibiotics not to be used to prevent nosocomial infections.
- Oral antibiotic bowel preparation with appropriate mechanical bowel preparation.
- If systemic antibiotics are to be used antibiotics of longer half-life are to be chosen.
- Very long procedures should have a redosing strategy during the procedure.

4. *Enhancement of Host Defences*

- Increased oxygen delivery facilitates phagocytic eradication of microbes.
- Optimising core body temperature is important as warmer patients resist bacteria better.
- Blood glucose control is essential even to nondiabetics as well.

Management of SSI

- SSI is managed depending on the type of SSI—superficial, deep or organ space.
- All infected material and pus should be removed from the wound site—*debridement*.
- Sutures are removed to allow *free drainage* of infected material.
- Infected fluid is sent for culture and sensitivity and suitable antibiotics are started.
- Once wound shows signs of healing by healthy granulation tissue, secondary suturing is done. Often it is allowed to heal by scarring.

E. Swellings

CHAPTER OUTLINE

- Lipoma
- Cysts
- Dermoids
- Sebaceous cyst
- Glomus Tumour
- Papilloma
- Warts
- Fibroma
- Bursae
- Semimembranosus Bursa
- Morrant Baker's Cyst
- Lymph Cyst
- Lymphangioma
- Calcinosis Cutis
- Neuroma
- Neurofibroma
- Neurilemmoma
- Ganglion
- Chordoma
- Epignathus

LIPOMA

- It is a benign tumour arising from yellow fat.
- Tumour arising from brown fat is called as ***hibernoma (reddish brown).***
- It is called as ***universal tumour*** (ubiquitous tumour) as it can occur anywhere in the body (except in brain).
- It is the ***most common benign tumour (Karyotype 12q change)***.
- It can be ***diffused*** or ***localised***.
- *Diffuse lipomas* are not encapsulated, not well localised. Common in palm, sole, head and neck region, difficult to be removed. It is seen in subcutaneous and intermuscular tissues (*pseudolipoma*).
- It can be *single or multiple (5%)*. Multiple lipomas often associated with many syndromes like ***MEN syndrome*** *(Multiple endocrine neoplasia syndrome).*

Types

- Painful lipomas are called as ***neurolipomas.***
- ***Dercum's disease*** is tender deposition of fat especially on the trunk, is also called as ***adiposis dolorosa***. *It is basically multiple neurolipomatosis.*
- *Fibrolipoma:* Lipoma with fibrous component.
- *Naevolipoma:* Lipoma with telangiectasis.
- *Lipoma arborigens* is pedunculated lipoma.
- Lipomas attain large size in thigh, shoulder, retroperitoneum, back and often may turn into sarcoma.

Sites

- Subcutaneous.
- Subfascial.
- Intramuscular.

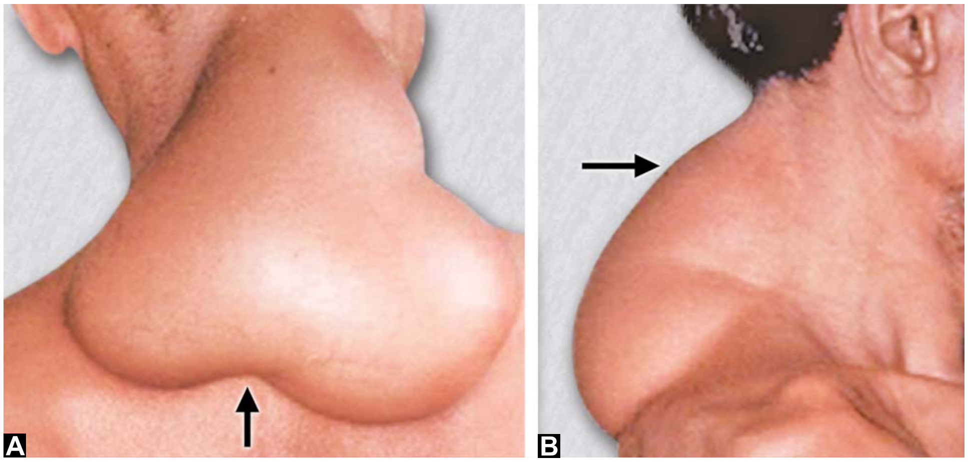

Figs 1.143A and B: Large lipoma in the nape of neck.

- Intermuscular in anterior abdominal wall.
- Parosteal.
- Subserosal.
- Submucosal (GI tract).
- Extradural (not intradural).
- Intra-articular.
- Subsynovial.
- Subperiosteal.
- Intraglandular—breast, pancreas, kidney

Clinical Features

- Localised swelling, which is lobular (surface), nontender.
- Semifluctuant (because fat in body temperature remains in semiliquid condition). It is usually *nontransilluminant*.
- Mobile, with edges slipping between the palpating fingers (slip sign).
- Skin is free.
- Lipomas may be pedunculated at times.
- It is *rare* in children.
- *Pain* in lipoma may be due to neural element or compression to nerves or adjacent structures. *Angiolipomas* being highly vascular is commonly *tender*.
- *Trunk* is the most common site; nape of neck and limbs are next common.
- Clinically lipoma can be single, multiple or diffuse.

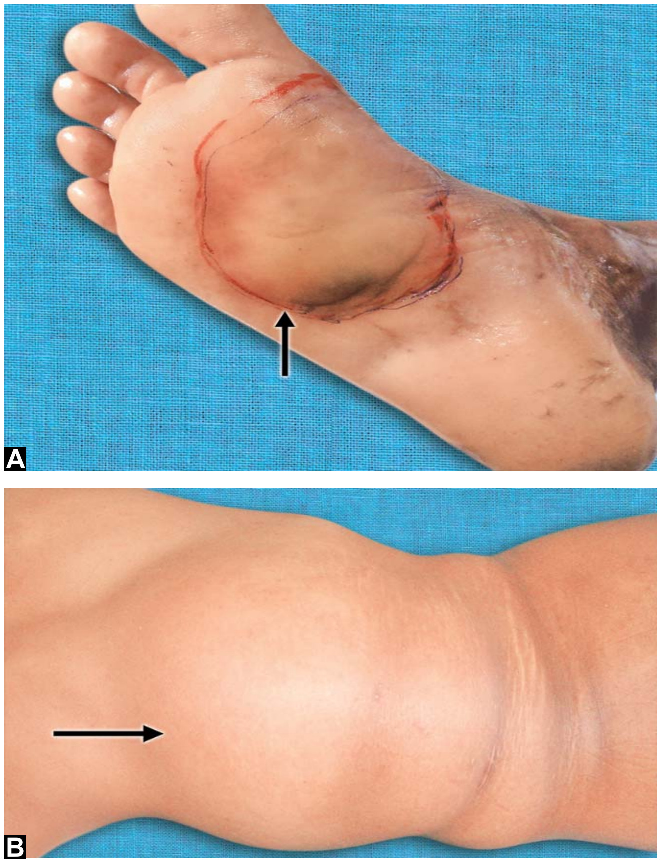

Figs 1.144A and B: Diffuse lipoma on the plantar aspect of foot and back.

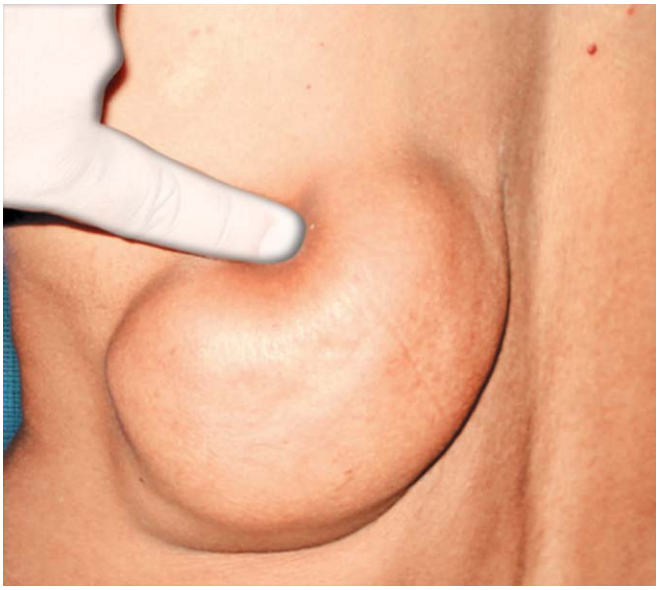

Fig. 1.145: Slip sign of lipoma.

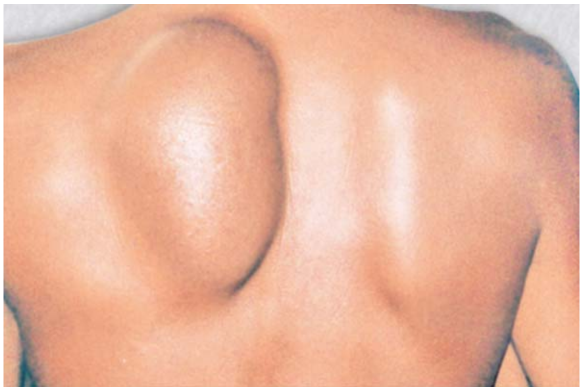

Fig. 1.146: Large lipoma in the back over scapula.

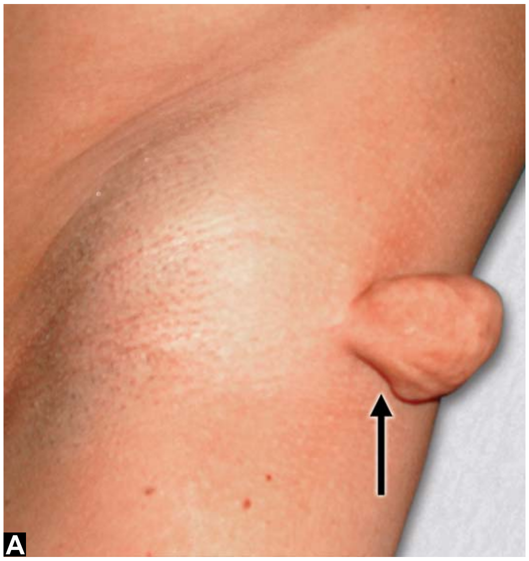

Fig. 1.147A

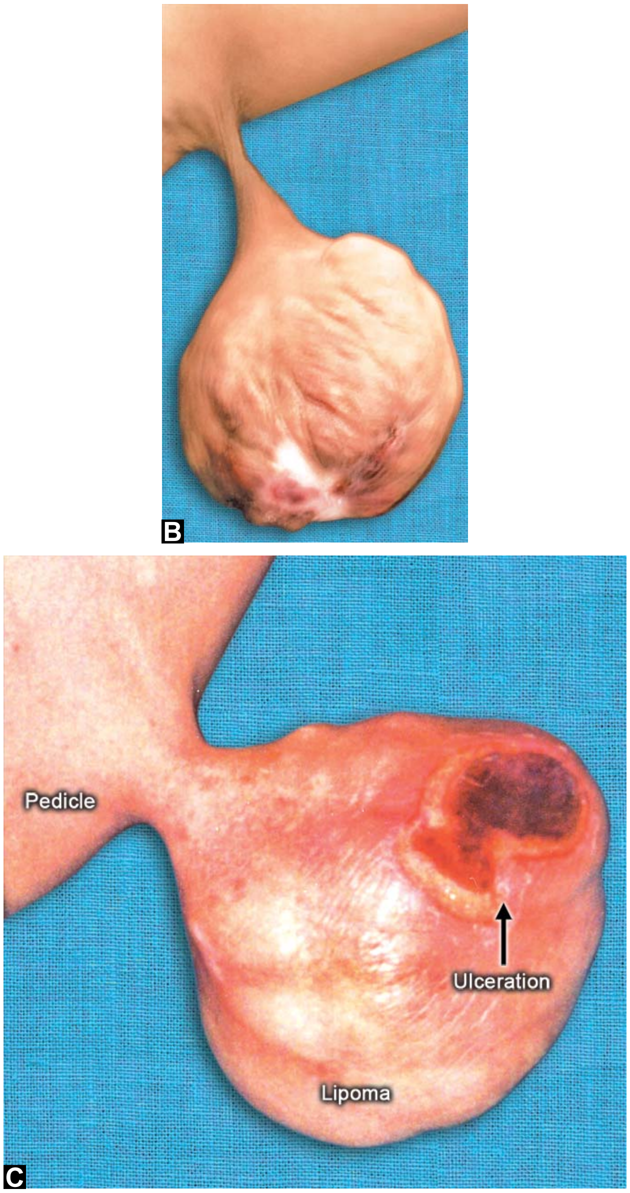

Figs 1.147B and C

Figs 1.147A to C: Pedunculated lipoma. It is common in axillary region. Lobular surface with narrow pedicle is typical. Often ulceration can occur in the surface due to repeated friction. It often mimics papilloma.

Differential Diagnosis

- Neurofibroma,
- Cystic swellings.

Note:

FNAC/incision biopsy in suspected cases and CT /MRI scan in cases of deeply located/intracavitary lipomas/large lipomas is done.

- Application of ice over the swelling hardens lipoma but not neurofibroma.

Complications

- Myxomatous changes—occurs in *retroperitoneal* lipoma.
- Saponification.

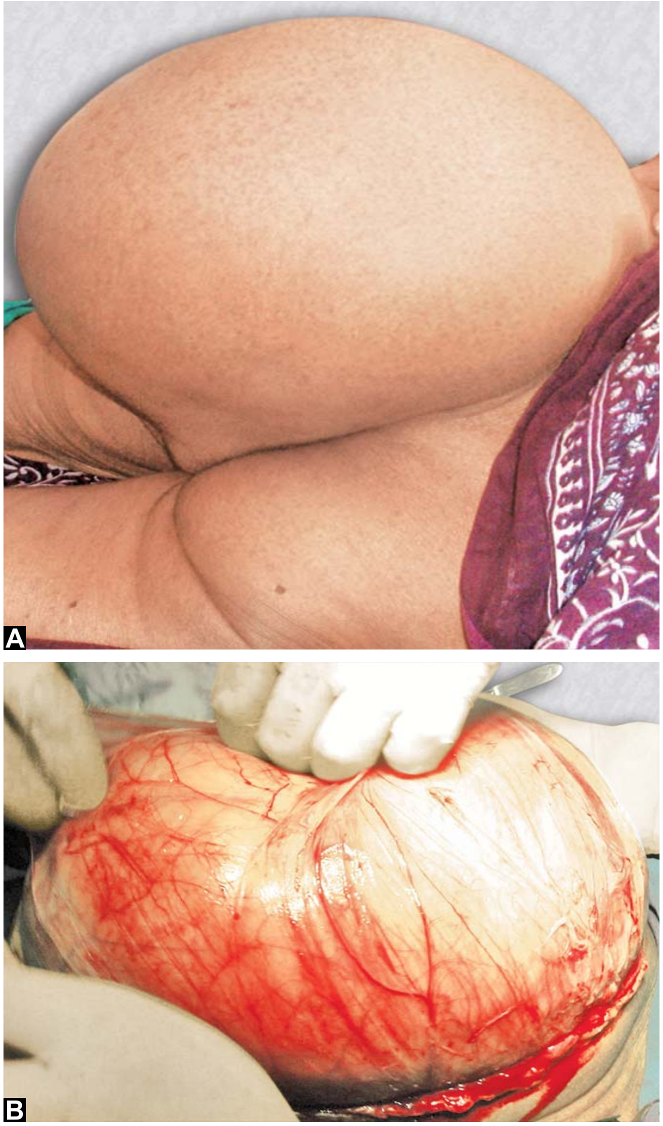

Figs 1.148A and B: Large lipoma in buttock region.

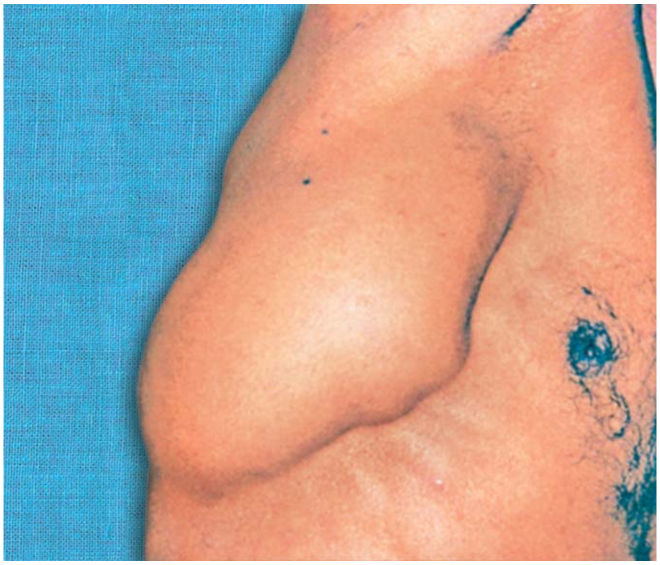

Fig. 1.149: Lipoma in lateral chest wall near posterior axillary fold.

There can be no substitute for detailed appraisal of the history of clinical signs.— **Harold Ellis**

- Calcification.
- Submucosal lipoma can cause intussusception and so intestinal obstruction.
- Sarcomatous changes—liposarcoma (most common type of sarcoma).

***Note**:*
It is now considered that all sarcomas are of *de novo* in origin to begin with at mitochondrial level; so benign soft tissue tumour turning into sarcoma is under debate or rare or not existing; but in *clinical practice* such an entity is still considered *by few*.

Liposarcoma

- Common in retroperitoneum, thigh and back
- Rapid growth
- Warm and vascular
- Dilated veins over the surface
- Infiltration in to deeper plane with restriction of the mobility
- Skin fixation and fungation
- Blood spread to lungs

Treatment

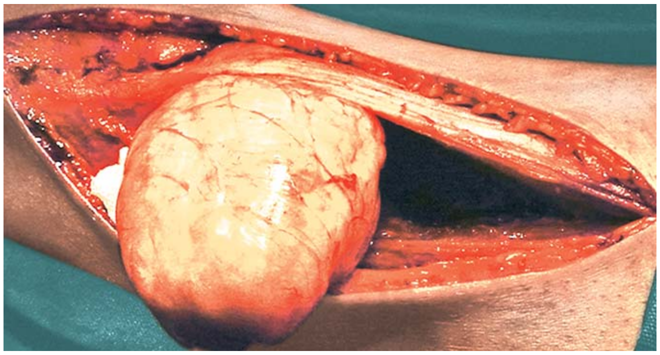

Fig. 1.150: Submuscular lipoma under gastrocnemius muscle.

- Excision.
- Small lipoma is excised under local anaesthesia and larger one under general anaesthesia.
- If it is liposarcoma, CT chest should be done to see secondaries in lungs. Later wide excision is done alongwith adjuvant chemotherapy and radiotherapy.

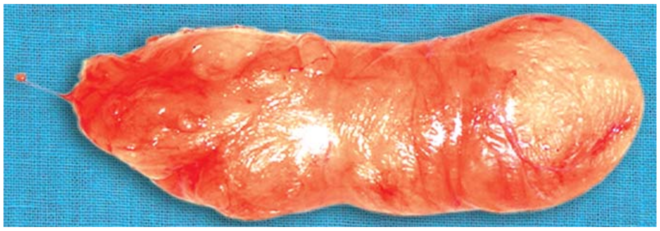

Fig. 1.151: Gross feature of specimen of lipoma.

***Note**:*
Lipoblastoma, a benign tumour is seen in infant boys in extremity subcutaneous tissue.

CYSTS

Cyst is a collection of fluid in a sac lined by epithelium or endothelium. **Word meaning of cyst is** "*bladder*" **(Greek)** **(**Greek word '*KUSTIS*' means bladder**).**

True Cyst

- Cyst wall is lined by epithelium or endothelium.
- If infection occurs, cyst wall also will be lined by granulation tissue.
- Fluid is usually serous or mucoid derived from the secretion of the lining.

False Cyst

- It does not have epithelial lining.
- Fluid collection occurs as a result of exudation or degeneration.

Example:

a. Pseudocyst of pancreas.
b. Wall of cystic swelling in tuberculous peritonitis.
c. Cystic degeneration of tumour.
d. After haemorrhage, in a haematoma, RBC's are lysed, gets absorbed and fluid remains as a false cyst.

"Apoplectic cyst" is formed in brain as a result of ischaemia, causing collection of fluid.

Classification

a. *Congenital cyst*
  - *Dermoids:* Sequestration dermoid.
  - *Tubulodermoids:* Thyroglossal cyst, postanal dermoid, ependymal cyst, urachal cyst.
  - *Cysts of embryonic remnants:* Cysts from paramesonephric duct and mesonephric duct, cysts of urachus and vitellointestinal duct.

b. *Acquired cysts*
  - *Retention cysts*: They are accumulation of secretions of a gland due to obstruction of the duct, e.g. sebaceous cyst, Bartholin's cyst, cyst of parotid, breast, epididymis.
  - *Distention cyst:* Lymph cyst, ovarian cyst, colloid goitre.
  - *Exudation cyst:* Bursa, hydrocoele, pancreatic pseudocyst.

c. *Cystic tumours*: Dermoid cyst of ovary, cystadenomas.
d. *Traumatic cyst*: Due to trauma, haematoma occurs usually in thigh, loin, shin. It eventually gets lined by endothelium containing brown coloured fluid with cholesterol crystals.
e. *Degenerative cyst*: Due to cystic degeneration of a solid tumour (due to necrosis of tumour).
f. *Parasitic cyst*: Hydatid cyst, trichiniasis, cysticercosis.

Clinical Features of a Cyst

- Hemispherical swelling which is smooth, fluctuant, nontender, well-localised.
- Some cysts are transilluminant.
- Presentation varies depending on its anatomical location and pathology.
- Cyst can be single or multiple. Sebaceous cysts are often multiple.

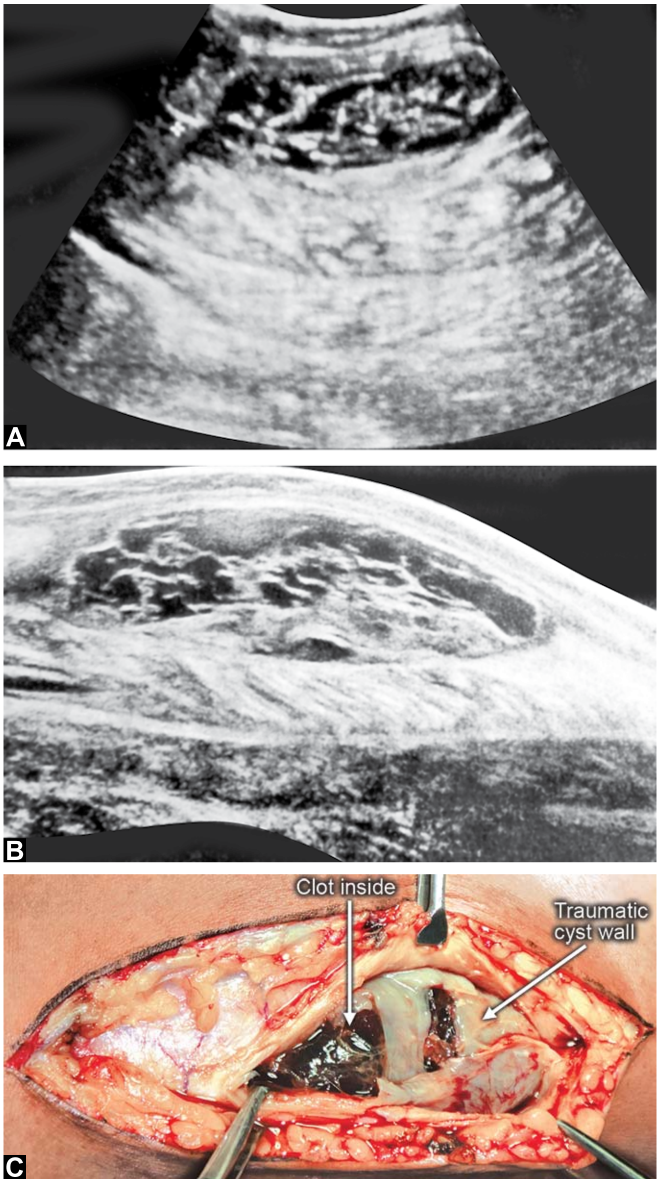

Figs 1.152A to C: Traumatic cyst thigh with false cyst wall with clot inside.

Effects of a Cyst

- Compression of adjacent structures (CBD): Choledochal cyst compressing over the CBD.
- Infection.
- Sinus formation.
- Haemorrhage.
- Torsion, e.g. ovarian cyst.
- Calcification, e.g. hydatid cyst, cysticercosis.
- Cachexia: In malignant ovarian cyst patient goes for severe cachexia.

Swellings which are brilliantly transilluminant

- Ranula
- Cystic hygroma and lymph cyst
- Hydrocoele
- Epididymal cyst (*Chinese-lantern* pattern)
- Meningocoele

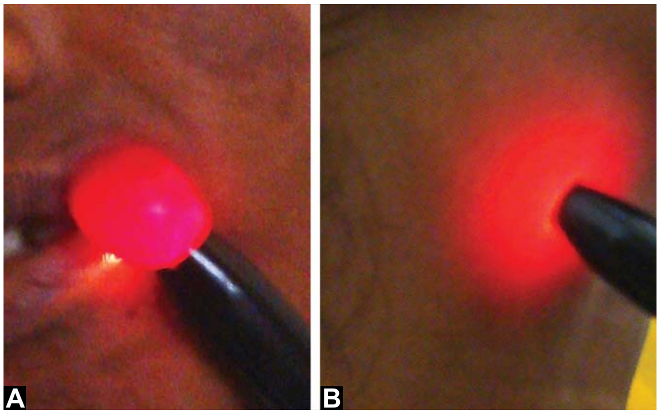

Figs 1.153A and B: Brilliantly transilluminant swellings—possibly lymph cyst.

DERMOIDS

Types

1. Sequestration Dermoids

- It occurs at the line of embryonic fusion, due to inclusion of epithelium beneath the surface which later gets sequestered forming a cystic swelling in the deeper plane.
- It is congenital type.

Common sites are:

a. Forehead, neck, postauricular dermoid.
b. External angular dermoid.
c. Root of nose.
d. Sublingual dermoid.
e. Anywhere in midline or in the line of fusion.
    - Dermoids occurring in the skull may extend into the cranial cavity.
    - When it occurs as an external angular dermoid, it extends into the orbital cavity, or it can extend into any cavity in relation to its anatomical location (e.g. thorax, abdomen).

Types of angular dermoid

a. ***External angular dermoid:*** It is a sequestration dermoid situated over the external angular process of the frontal bone. Outer extremity of the eyebrow extends over some part of the swelling. This typical feature differentiates it from the swelling arising from the lacrimal gland. It may extend into the orbital cavity also (Frontozygomatic suture).

b. ***Internal angular dermoid:*** It is a sequestration dermoid cyst in central position near the root of the nose. It occurs in frontonasal suture line. It is rare. It mimics swelling from *lacrimal sac* or mucocele of frontal sinus. *Mucocele of frontal sinus* is due to blockage of frontonasal duct.

- Dermoid cyst contains putty like desquamated material, hair follicle, sebaceous and sweat glands. It is lined by both dermal and epidermal components.
- Dermoid cyst in skull region has different anatomical types as it often may extend into cranial cavity.

Count your blessing not your problems.

- Cyst which is located entirely outside the skull bone over suture line but without bone indentation.
- Cyst located outside the skull bone but with a bone defect underneath. Bone defect may be either on outer table of skull or through both tables of skull with attachment to dura.
- Cyst lyeing partly outside and partly inside the skull with a connecting stalk between the two like a *dumb bell*.
- Cyst is entirely within the skull bone between skull and dura. It is very rare but known.

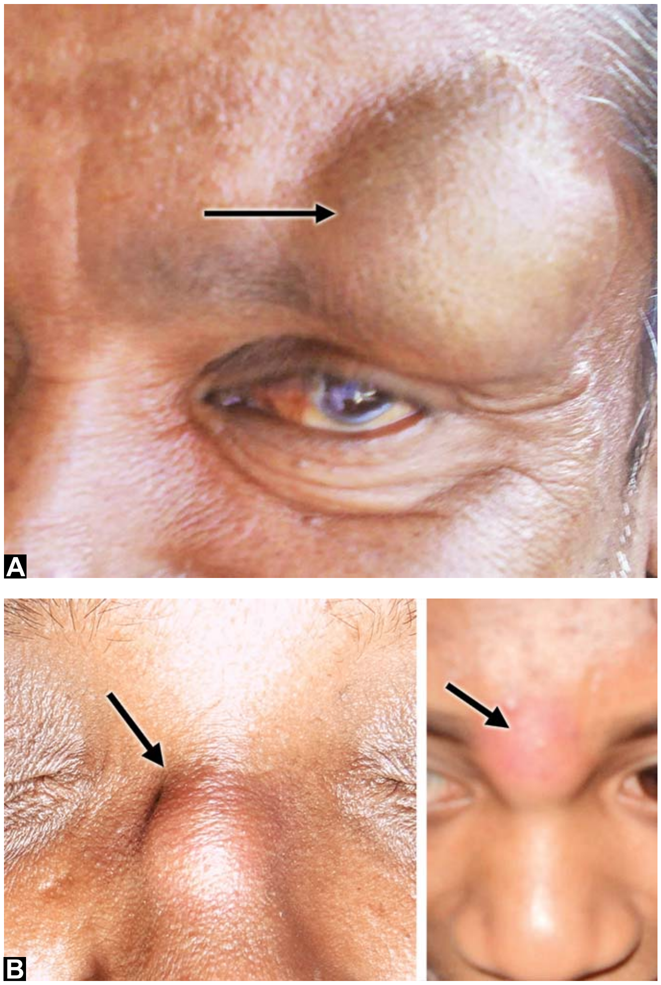

Figs 1.154A and B: Types of angular dermoid. (A) External angular dermoid. (B) Internal angular dermoid (midline).

Clinical features

- Painless swelling in the line of embryonic fusion.
- Presents in the second or third decade onwards.
- Smooth, soft, nontender, fluctuant (*Paget's test positive* i.e. swelling is fixed with two fingers and summit is indented to get yielding sensation due to fluid).
- Nontransilluminating.
- Free skin, often adherent into the deeper plane.
- There will be ***resorption and indentation*** of the bone beneath.
- Impulse on coughing may be present only if there is intracranial extension.

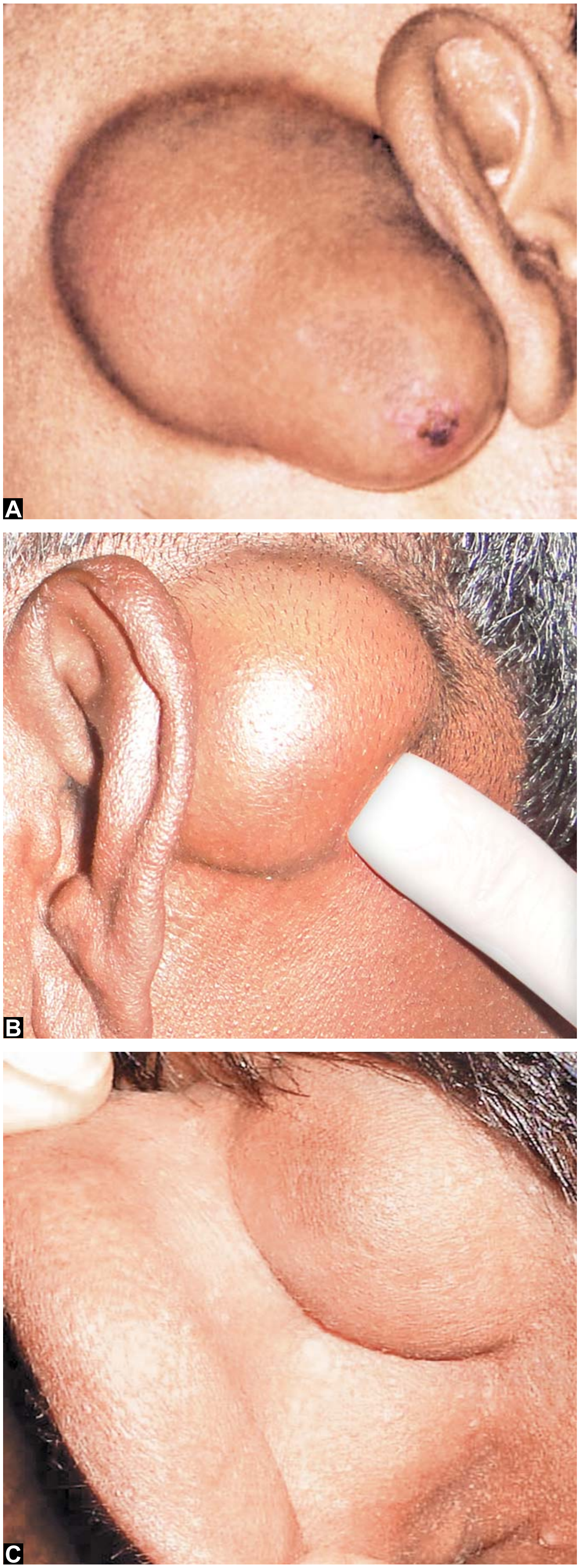

Figs 1.155A to C: Postauricular dermoid in different patients.

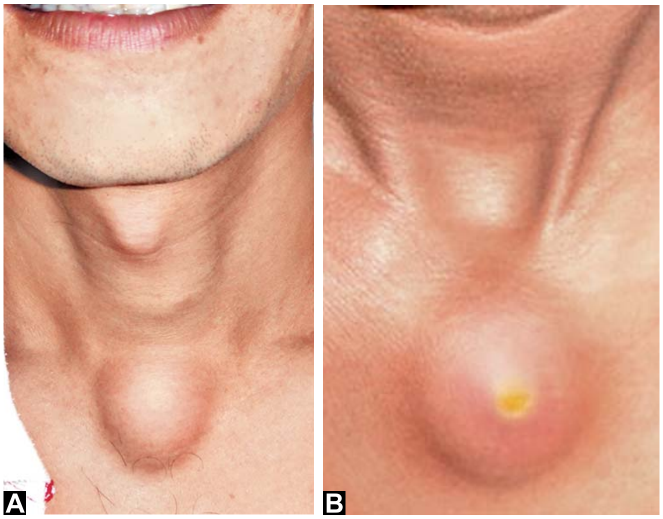

Figs 1.156A and B: Dermoid in the midline—suprasternal space of burns/sternum. It may extend into deeper plane and so it needs CT chest and evaluation.

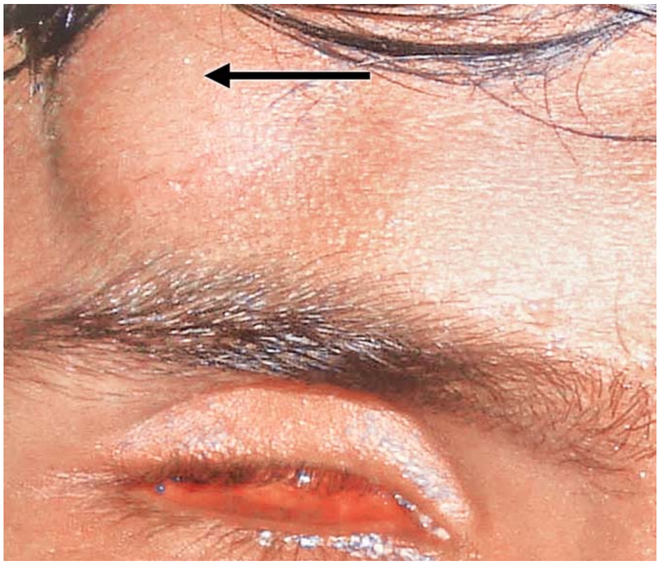

Fig. 1.157: Lipoma forehead. It looks like a dermoid.

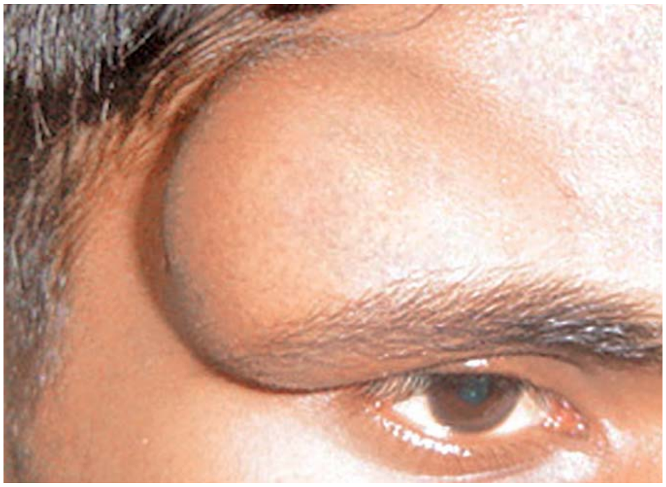

Fig. 1.158: External angular dermoid, right sided.

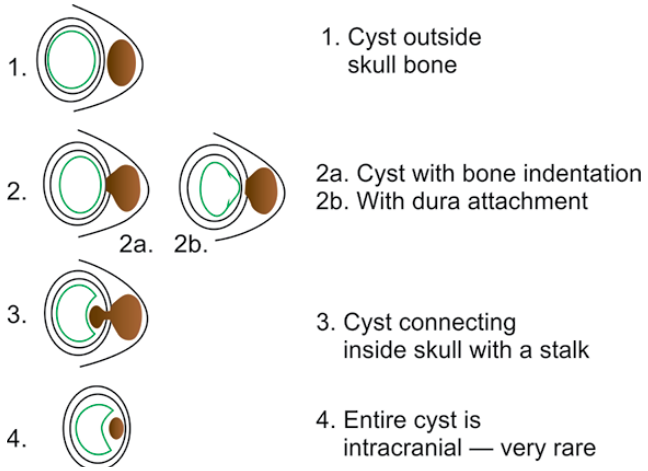
1.
1. Cyst outside skull bone
2.
2a.
2b.
2a. Cyst with bone indentation
2b. With dura attachment
3.
3. Cyst connecting inside skull with a stalk
4.
4. Entire cyst is intracranial — very rare

Fig. 1.159: Sequestration dermoid in skull—anatomical types.

Differential diagnosis

- Sebaceous cyst.
- Lipoma.

Investigations

- X-ray—skull or part.
- CT scan head or part.

Complication of sequestration dermoid

- Infection
- Haemorrhage, rupture
- Surface ulceration
- Pressure effects if there is intracavitary extension like into cranial cavity or thoracic cavity, etc.
- Calcification

Treatment

- Excision is done under *general anaesthesia.* Often formal neurosurgical approach is required by raising cranial osteocutaneous flaps.

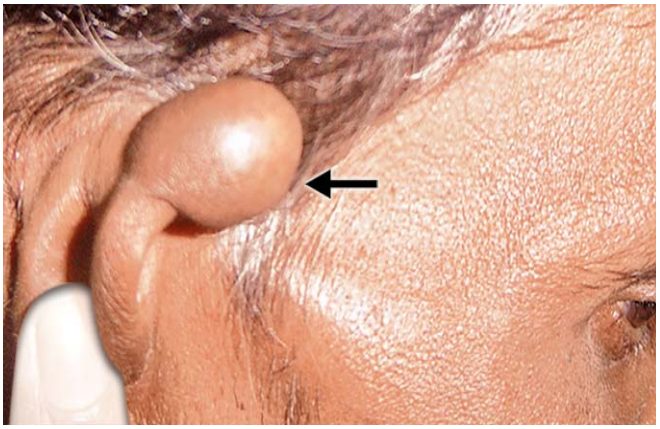

Fig. 1.160: Dermoid in the ear. It arises due to sequestration at the fusion line of the one of the six developmental ear tubercles (Each ear develops from SIX ear tubercles).

The most valuable diagnostic instrument is passage of time. — **Henry George Miller**

Submental dermoid

- ❖ It is a congenital sequestration dermoid occurs during fusion of 1st and 2nd branchial arches. It is deep to deep fascia of neck.
- ❖ It presents as soft, cystic, fluctuant, nontransilluminating, swelling in midline in submental region which does not move with deglutition nor moves while protruding the tongue out.
- ❖ It should be differentiated from thyroglossal cyst, cold abscess from submental lymph nodes or sebaceous cyst.
- ❖ It is excised under general anaesthesia with a curvilinear submental incision.

2. *Tubulodermoids*

It arises from the embryonic tubular structures.
Examples include:

- ♦ Thyroglossal cyst.
- ♦ Ependymal cyst.
- ♦ Postanal dermoid.
- ♦ Urachal cyst.

3. *Implantation Dermoid (Figs 1.161 and 1.162)*

- ♦ Due to minor pricks or trauma, epidermis gets buried into the deeper subcutaneous tissue which causes reaction and cyst formation (trauma is *forgotten often*). It is an *acquired* cyst.
- ♦ It is common in fingers (common in tailors, gardeners), toes and feet.

Clinical features

- ♦ Swelling is painless, observed after minor trauma, slowly progressing in fingers or toes.
- ♦ It is smooth, soft, mobile, tensely cystic, nontransilluminating and is often adherent to skin.
- ♦ It contains only squamous epithelium, without hair follicle/ sweat or sebaceous glands.
- ♦ Complications are infection, rupture and pressure effect over digital nerves.

Differential diagnosis

- ♦ Lipoma, Bursa.

Treatment

- ♦ Excision—under local anaesthesia.

4. *Teratomatous Dermoid*

- ♦ It arises from all germinal layers ecto, meso and endoderms.
- ♦ It occurs in ovary, testis, retroperitoneum, mediastinum.
- ♦ It contains hair, teeth, cartilage, sebum and muscle.
- ♦ It can be benign or malignant.

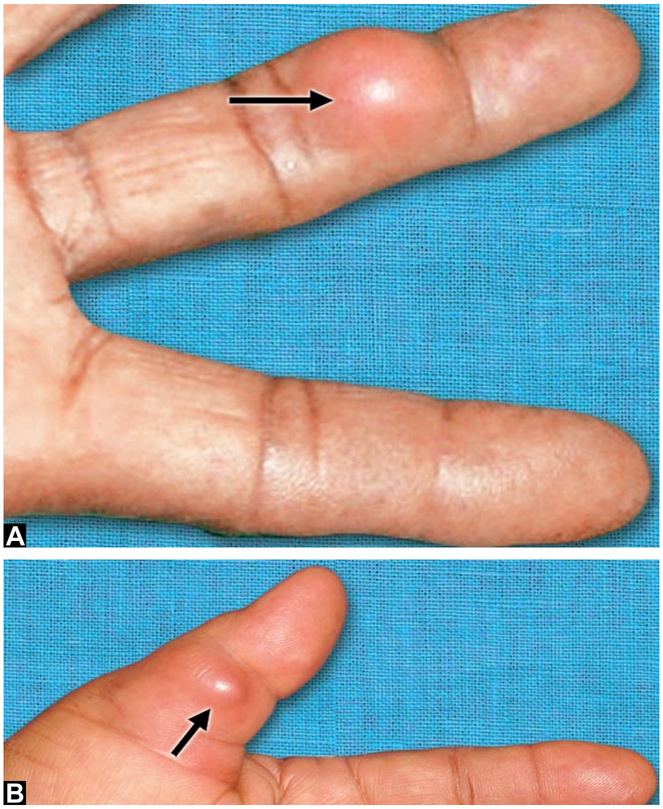

Figs 1.161A and B: Implantation dermoid finger and thumb—classical site.

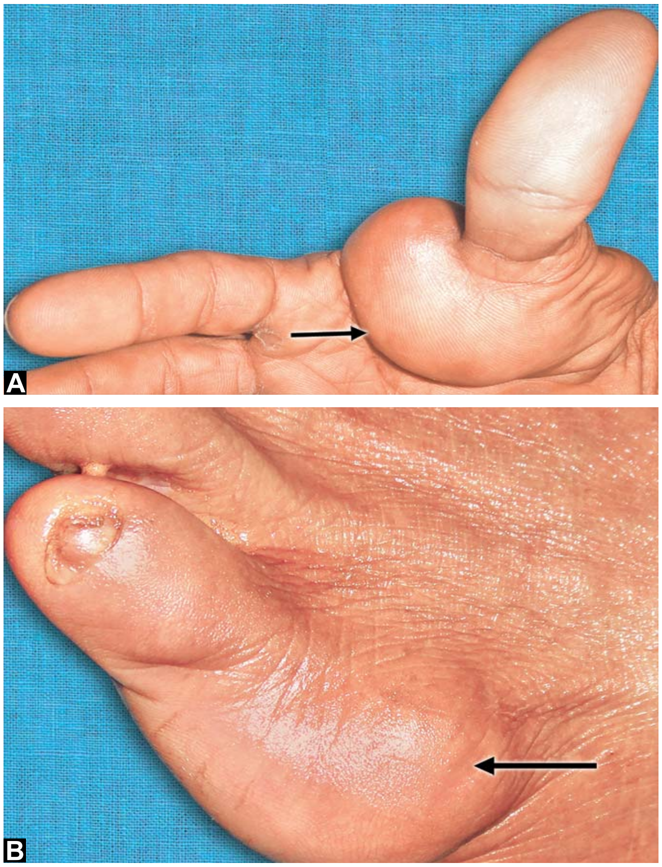

Figs 1.162A and B: Implantation dermoids in the hand and foot.

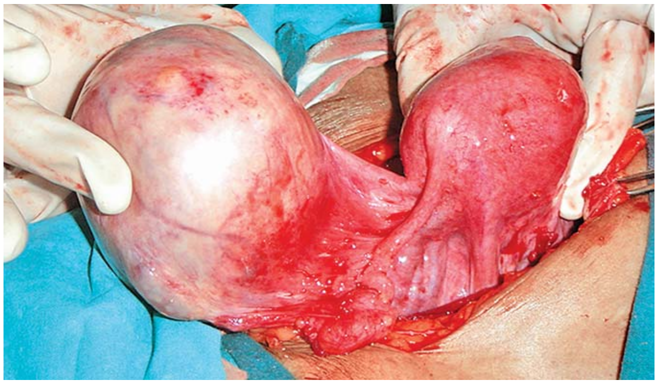

Fig. 1.163: Dermoid cyst ovary right sided.

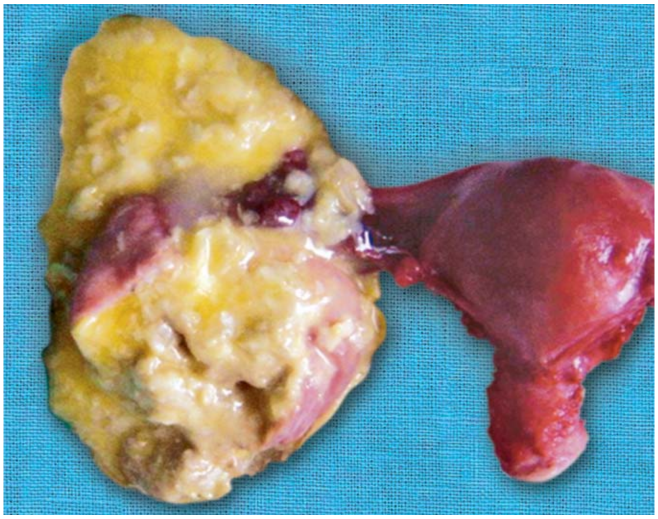

Fig. 1.164: Cut section of dermoid cyst ovary showing sebum.

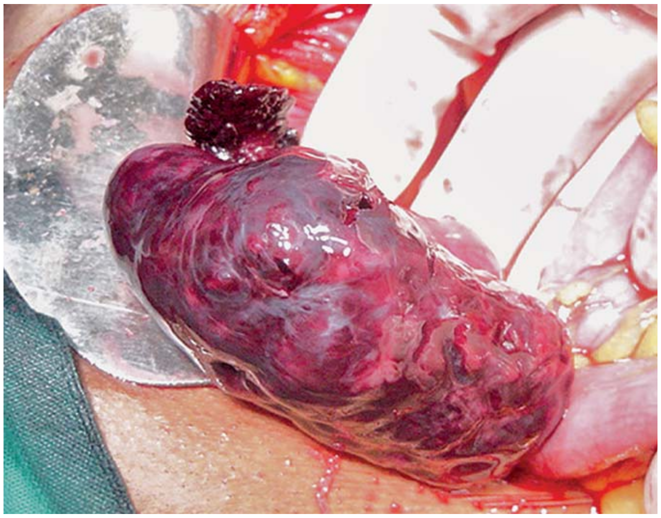

Fig. 1.165: Twisted gangrenous cyst ovary.

SEBACEOUS CYST (WEN, EPIDERMOID CYST)

- It is a *retention cyst.* It is due to blockage of the duct of sebaceous gland, causing a cystic swelling. Sebaceous glands are situated in dermis which secretes sebum through sebaceous duct which opens either directly to skin surface or in to a hair follicle.
- It is common in face, scalp, scrotum.
- *It is not seen in palms and soles* as there are no sebaceous glands.
- Sebaceous cyst contains yellowish white cheesy material with fat and epithelium. It has putty like consistency, with a parasite in the wall of the sebaceous cyst—***Demodex folliculorum***. It is lined by only epidermal layer of squamous epithelium.

Clinical Features

- Painless swelling which is *smooth, soft, nontender, freely mobile, adherent to skin especially over the summit, fluctuant (positive Paget's test), nontransilluminating with punctum over the summit.*
- It moulds on finger indentation.

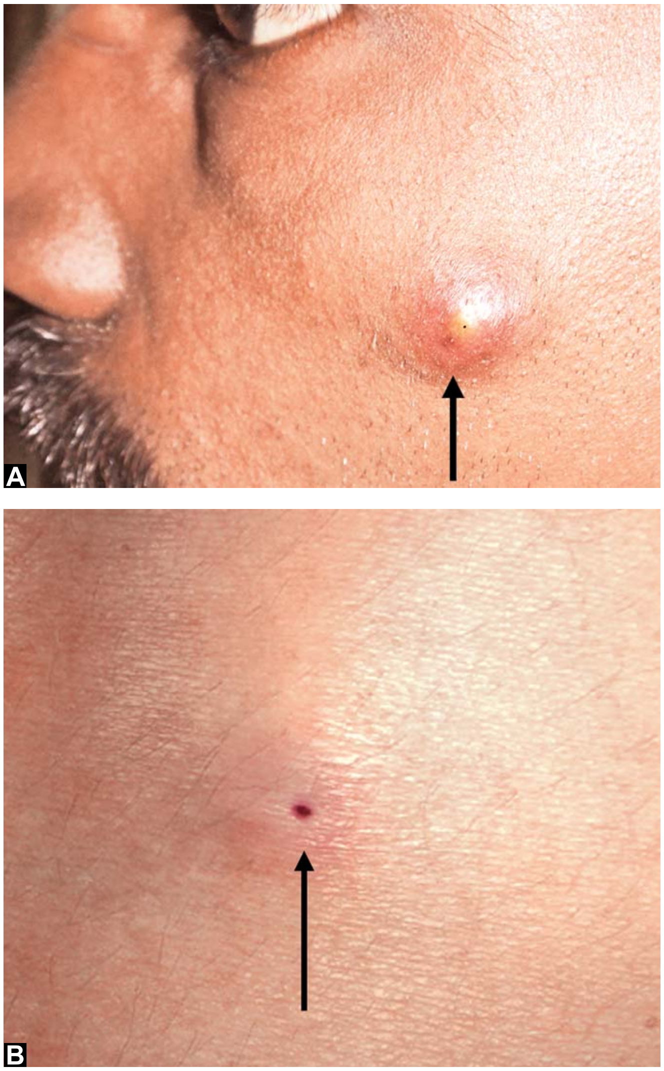

Figs 1.166A and B: Small sebaceous cysts in face and back with typical punctum over the summit.

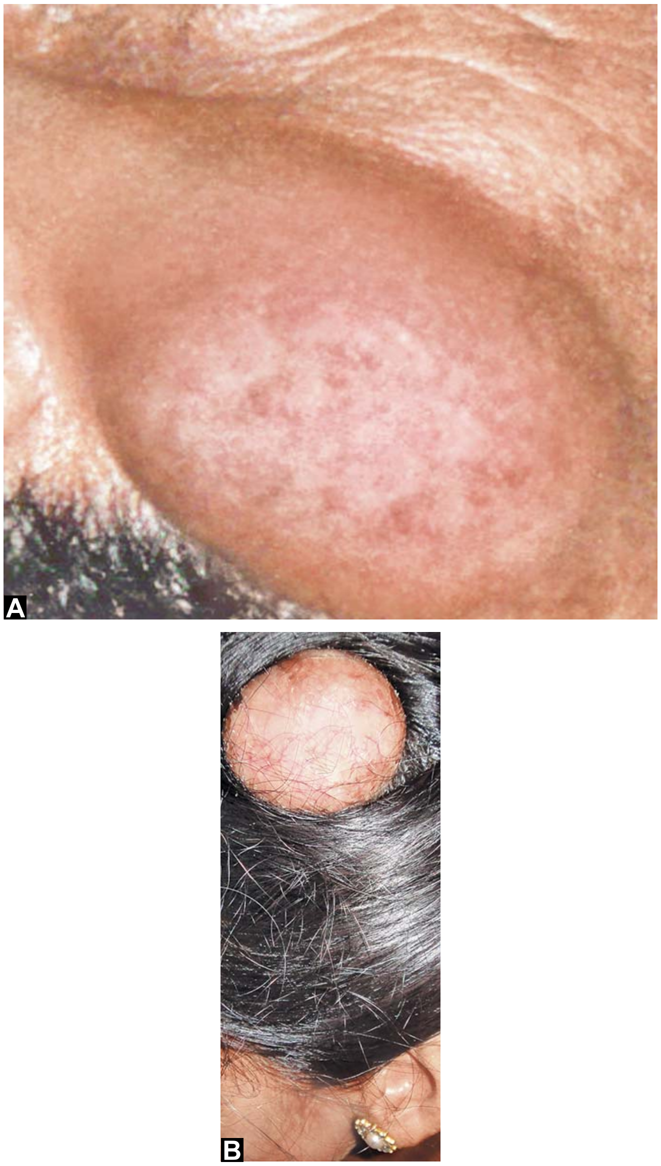

Figs 1.167A and B: Large sebaceous cyst in the face and scalp. Note the hair loss on the surface.

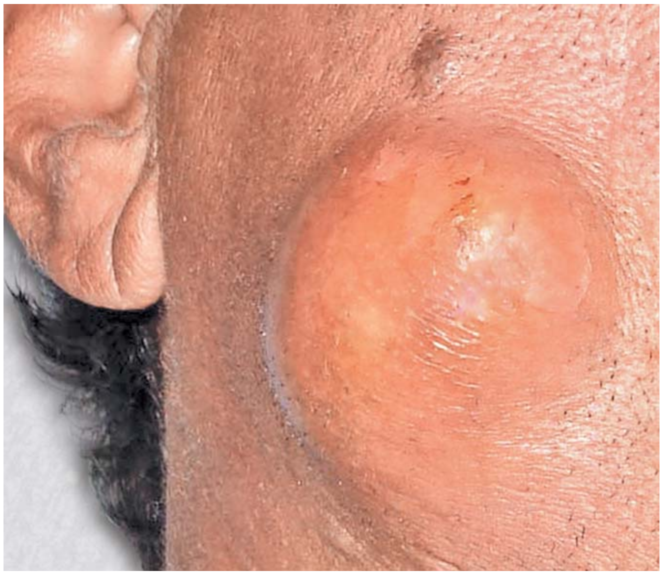

Fig. 1.168: Infected sebaceous cyst face.

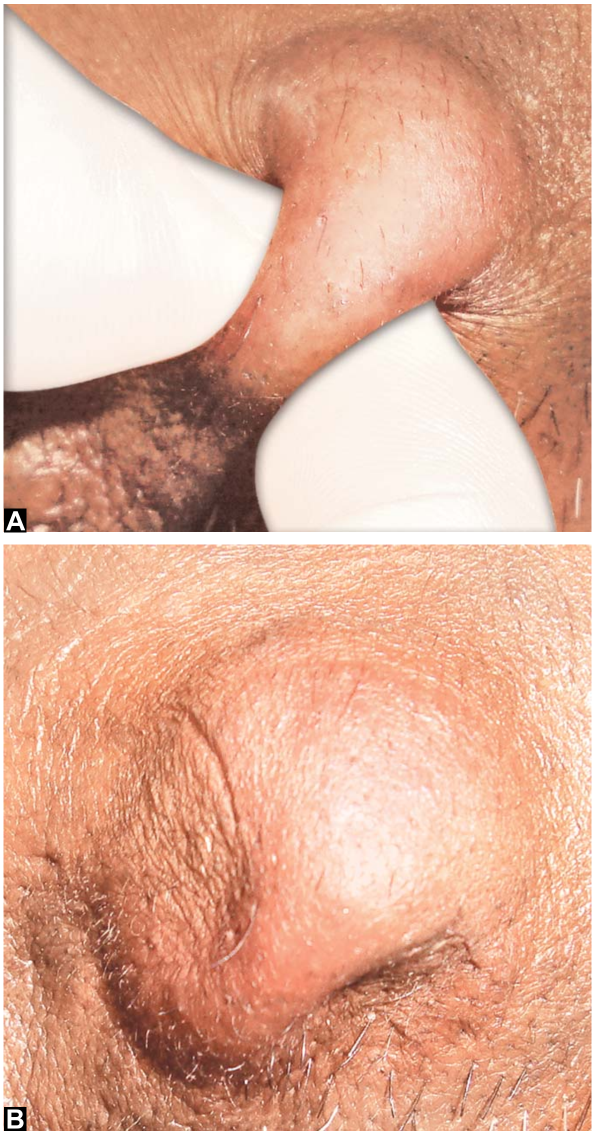

Figs 1.169A and B: Sebaceous cyst showing moulding sign.

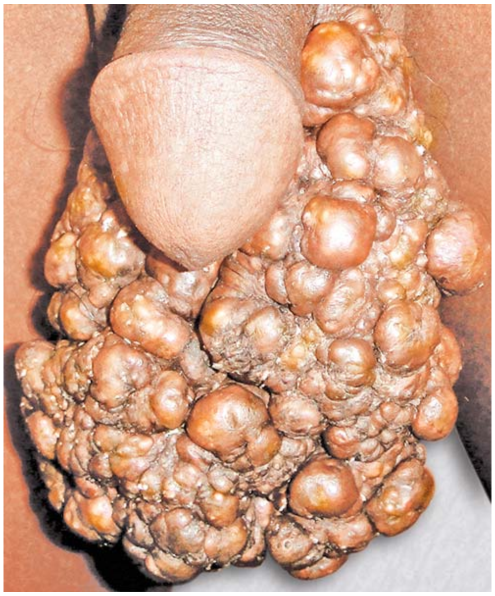

Fig. 1.170: *Strawberry scrotum*—multiple sebaceous cysts on the scrotum.

- ***Punctum*** is present over the summit in 70% of cases because here sebaceous duct opens directly into the skin which gets blocked. Punctum is depressed black coloured spot over the summit of the sebaceous cyst. Because of the denuded squamous epithelium (keratin) it is black in colour. In 30% cases sebaceous duct opens into the hair follicle and so punctum is not seen.
  (**Fordyce's disease** is heterotopic sebaceous glands in mucosa of the lip and oral cavity).
- Hair loss over the surface is common due to constant pressure over the roots of the hair follicles.
- Unpleasant odour of sebum content is typical.

Complications

- Infection and abscess formation.
- Surface gets ulcerated leading to formation of a painful, fungating mass with discharge called as—***Cock's peculiar tumour***—often resembles epithelioma. It is a misnomer as it is not a tumour. It is a chronic granuloma on an ulcerated surface of a sebaceous cyst.
- ***Sebaceous horn*** results from hardening of slowly discharged sebum through the punctum. Horn is one which has greater length than its base diameter. Cutaneous horn is keratin deposition.

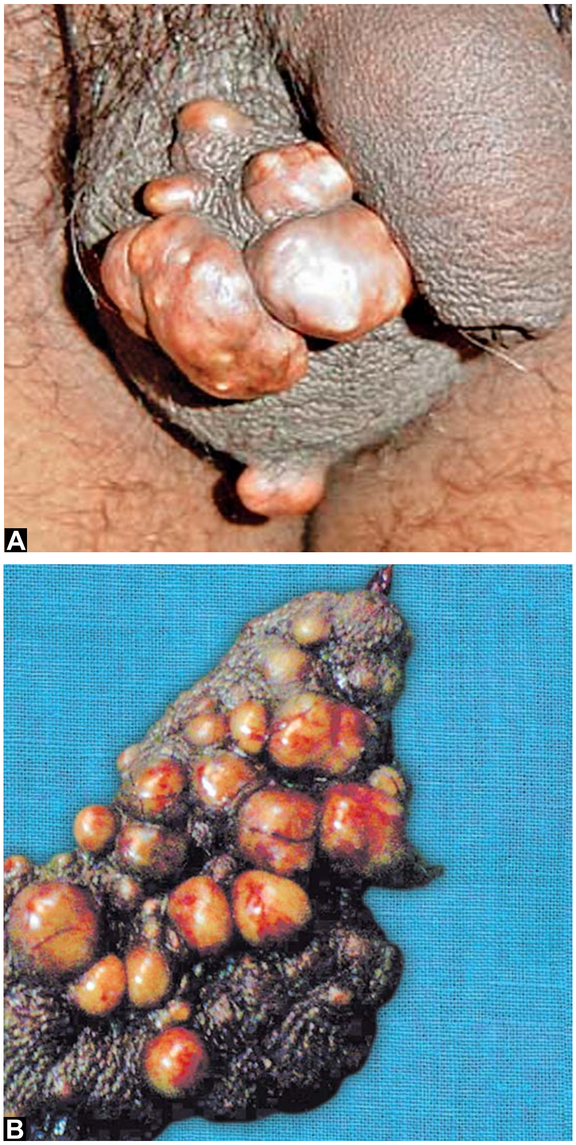

Figs 1.171A and B: Multiple sebaceous cysts in the scrotum. Occasionally partial scrotectomy is done in these patients.

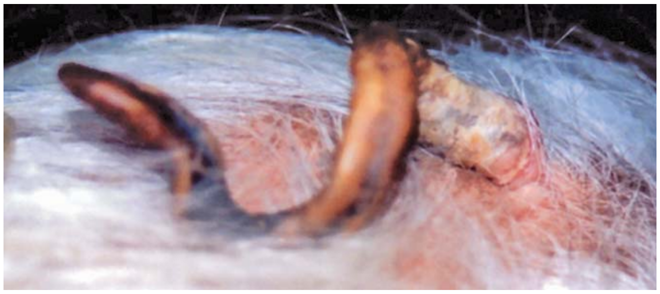

Fig. 1.172: Sebaceous horn scalp.

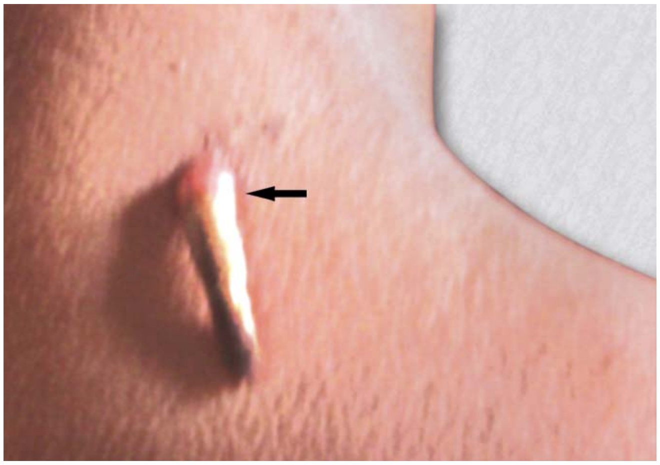

Fig. 1.173: Sebaceous horn—nape of the neck.

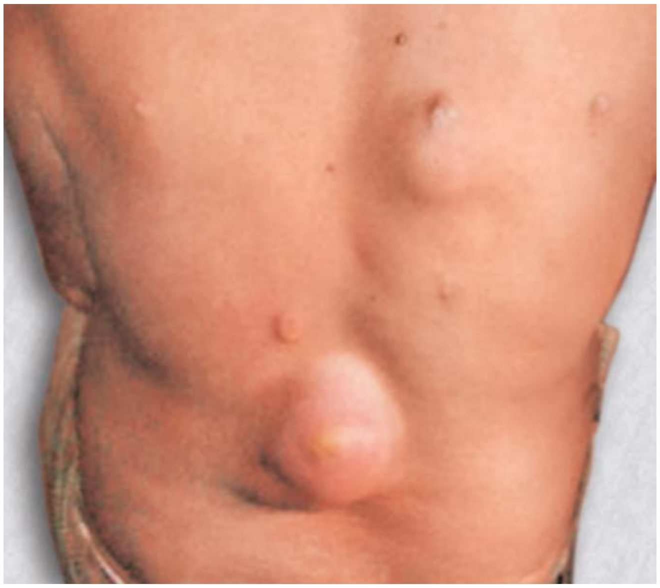

Fig. 1.174: Multiple sebaceous cysts back.

Note:

- Multiple sebaceous cysts may be associated with syndromes like Gardner's syndrome.
- Sebaceous cyst in the scrotum is usually multiple, firm and often calcified without any punctum. It is often treated by partial or total scrotectomy. Scrotum with multiple sebaceous cysts is *strawberry* scrotum.

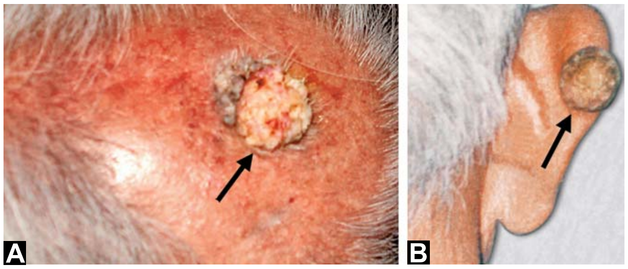

Figs 1.175A and B: Cock's peculiar tumour over scalp and ear.

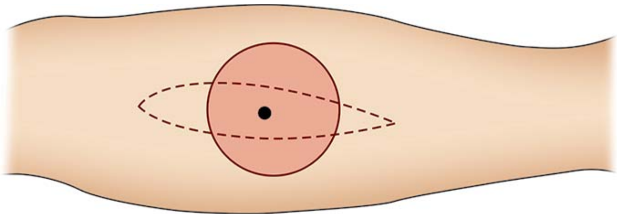

Fig. 1.176: Incision for sebaceous cyst is elliptical which includes punctum.

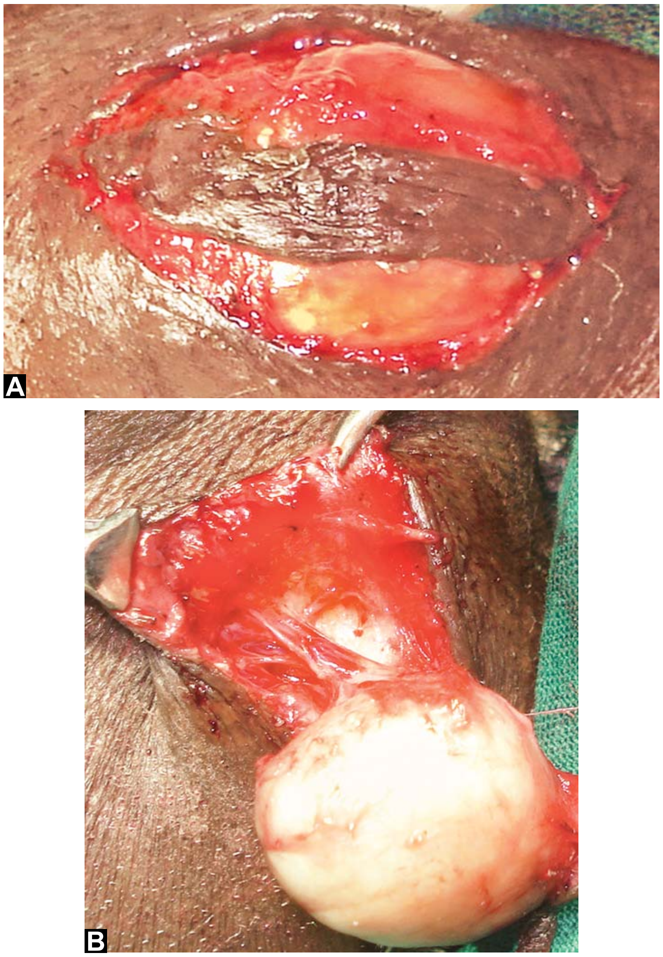

Figs 1.177A and B: Elliptical incision and removal of sebaceous cyst.

Treatment

- Excision including skin adjacent to punctum using elliptical incision—*dissection method*.
- Incision and avulsion of cyst wall.
- If abscess is formed, then drainage initially and later excision is done.
- If capsule is not removed properly the cyst will recur.

GLOMUS TUMOUR

- It is also called as ***glomangioma***.

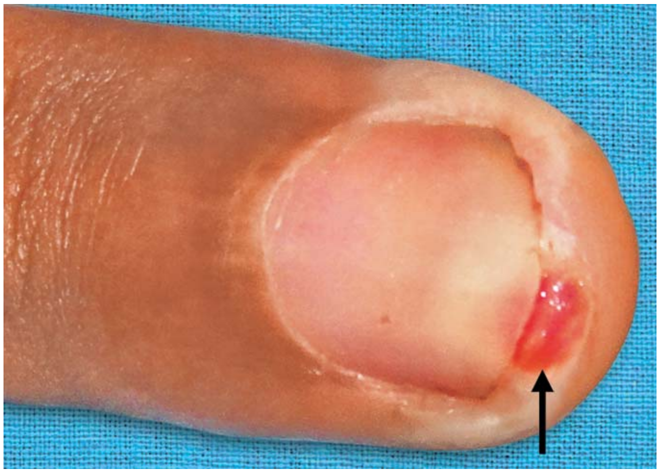

Fig. 1.178: Glomus tumour in finger tip—subungual region.

- It arises from the cutaneous glomus composed of a tortuous arteriole which communicates directly into the venule (*Sucquet-Hoyer* canal) and these vessels being surrounded by network of small nerves making this lesion very painful. This communication is called as *Sucquet-Hoyer canal*.
- Tumour consists of a mixture of blood spaces, nonmedullated nerve tissue, muscle fibres derived from the wall of the arteriole, with large cuboidal glomus cells—***angiomyoneuroma***.
- They are often seen in limbs and common in nail-beds near the finger tips, as purple red swelling.
- It is 2-3 mm in size.
- It does not turn into malignancy.

Clinical Features

- *Severe burning sensation and pain, out of proportionate to the size. The most common site is nail-bed.*
- Even the slightest pressure will give rise to severe pain. Dilated vessels compress over nerves.
- **It is compressible and pain is more when the limb is exposed to sudden changes in temperature (cold stimulus).**
- On increasing the pressure in the arm above systolic, pain disappears.
- It looks like a *reddish blue spot* which does not blanch on percussion.
- *Subungual* type may not be visible but only to cause episodic digital severe pain.
- It is usually single, but *rarely* multiplicity is observed as familial.

Sequestration dermoid	*Sebaceous cyst*
• Occurs in the line of fusion	• Occurs anywhere except palm and sole
• Skin is not adherent (free)	• Skin is adherent over summit
• Extends often into deeper plane or cavities through suture line	• Subcutaneous plane—do not extend to deeper plane
• Punctum is absent	• Punctum is present—70% cases
• Bone resorption and indentation is common	• Freely mobile without bone resorption
• With restricted mobility	• Superficial swelling, mobile
• Needs proper evaluation with X-ray/CT scan	
• Excision is done under general anaesthesia	• Excision is done under local anaesthesia

Differential Diagnosis

- Pyogenic granuloma—bleeds on touch.
- Subungual melanoma—painless pigmented lesion.

Treatment

- Excision cures the condition.

PAPILLOMA

- It is warty swelling from the skin or often from the mucous membrane.
- It has got a central axis of connective tissue, blood vessels and lymphatics.

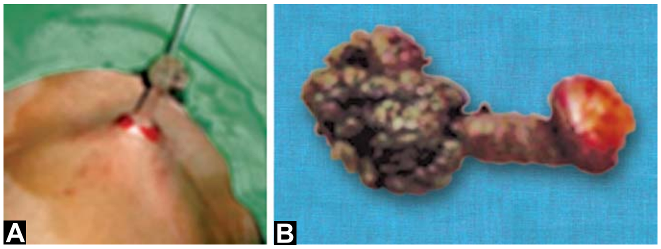

Figs 1.179A and B: Pedunculated papilloma. Note the incision for the same.

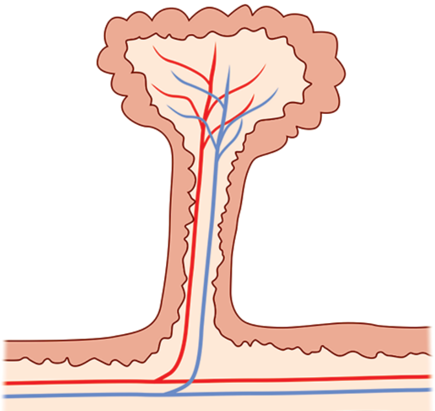

Fig. 1.180: Pedunculated papilloma with its pedicle.

True Papilloma

- It is a benign tumour with localised overgrowth of *all* layers of the skin. It is commonly pedunculated but rarely can be sessile.
- It contains sweat glands, sebaceous glands and hair follicles.
- Pedunculated papilloma is villous with a central axis of connective tissues, blood vessels and lymphatics.
- Papilloma can be *cutaneous or mucosal*. Cutaneous can be *squamous or basal* cell type. Squamous cell papilloma can be soft which is seen in eyelids in elderly; or can be *congenital* which can be sessile or pedunculated; single or multiple. Squamous papilloma also occurs in oral cavity. *Basal cell papilloma* is oily semitransparent brownish raised seborrhoeic keratotic lesion in skin around trunk of elderly.

Infective Papilloma

Infective papilloma is a warty lesion due to infection, e.g. condyloma acuminata.

Papilloma may be	single.
	multiple.
Papilloma may be	pigmented.
	nonpigmented.

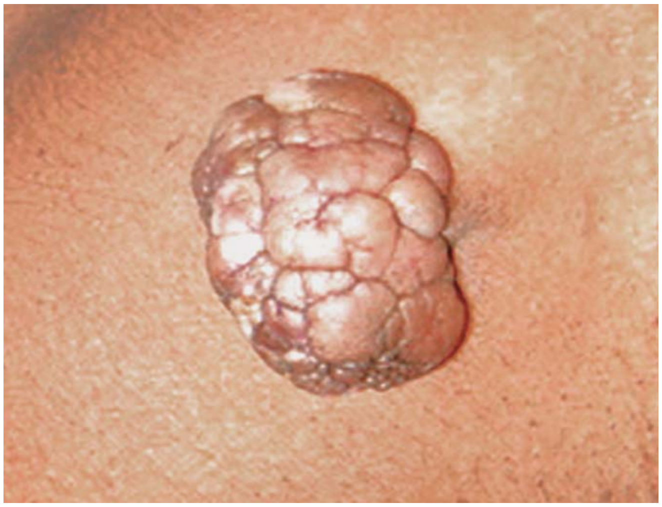

Fig. 1.181: Papilloma right waist. Note the warty surface.

> *Look well to this day, for yesterday is but a dream and tomorrow is but a vision. But today, well lived, makes every yesterday a dream of happiness and every tomorrow a vision of hope.* **—Kalidasa**

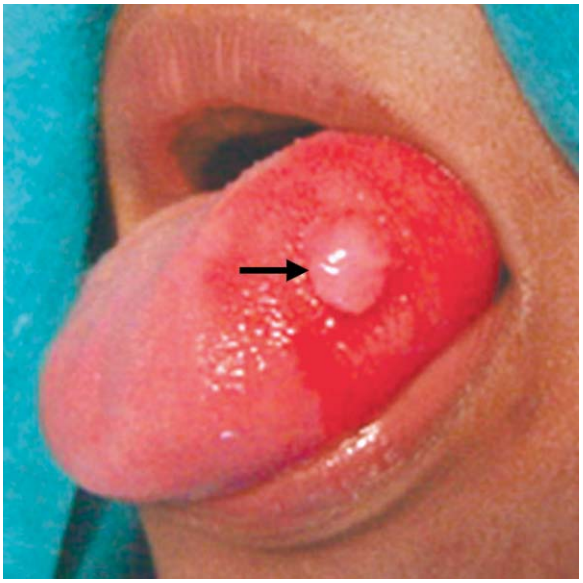

Fig. 1.182: Papilloma tongue. Note: Papilloma can occur in mucous membrane also like in oral cavity, urinary bladder, gallbladder and rectum.

- True papilloma may turn into squamous cell carcinoma occasionally. There will be sudden increase in size, bleeding or ulceration.
- Papilloma can occur in the breast called as *duct papilloma* which is the most common cause of bloody discharge from the nipple.
- Papilloma can occur in mucous membrane like in *oral cavity, urinary bladder* (transitional papilloma), in the *rectum* (columnar), in the *larynx,* in the *gallbladder* (cuboidal).

Differential Diagnosis

Amelanotic melanoma, pedunculated lipoma, carcinoma.

Treatment

- *True papilloma* is excised with its base along with surrounding 1 cm skin margin.
- *Infective warts* can be treated by excision or CO_2 snow or diathermy coagulation.

Complications of papilloma

- Bleeding
- Malignant transformation
- Ulceration
- Mechanical disability like voice change when it occurs in vocal cord

Fibroepithelial papilloma also called as *skin tags/achrochordon* is very common benign skin lesion of usually 5 mm diameter in size; common in adults; common in neck, axilla, thigh and groin. It is a vascularised keratinised squamous epithelium, can cause local irritation and insignificant bleeding. If it causes cosmetic problem it is removed by laser/cautery or excision.

WARTS

- They are usually multiple hyperkeratotic skin patches with finger like projections, common in children and adolescents.
- They are common in finger tips, face, axilla and sole of the feet. It may be familial but often stimulated by virus.
- They are dry, overgrown projections from the skin of finger often painful, tender and disfiguring. Repeated rubbing may cause infection. It can spread to other fingers and other parts of the body. Kiss lesions can occur.
- Warts attain their size in few weeks and then persist for many months to years. They may regress spontaneously also.
- **Plantar wart (*Verruca plantaris*)** is wart in the sole. Speciality of this is it gets pushed into the sole of the foot due to walking. It is common in ball and heel of the foot. It is pearly white in color with brownish haemorrhagic flecks. It is often covered by apparently normal skin because wart is buried into the skin. It looks like a circular pit. It is grey white finger/filiform like strands in the center of the lesion and is soft. Plantar warts can be multiple. It is painful and very tender on pressing (more than callosity or corns). A defined clear lump may not be felt.
- *Butcher's wart/pathologist's wart/verrucas necrogenica* is due to entry of *Mycobacterium tuberculosis* through broken skin which are common in *milkmaids*. It presents as bluish red warty lesion from which fluid oozes out on pressure between projections. It is common over *dorsum of hand*. Lesion is surrounded by pustules.
- *Senile* warts and *venereal* warts are other types. Venereal warts are same as infective papilloma.
- *Differential diagnoses* are true papilloma, callosities, neurofibromas.
- *Treatment* is excision. Salicylic acid, podophyllin cream applicatins are also used. Laser ablation of warts is commonly advocated.

FIBROMA

It is a benign tumour arising from fibrous tissue. It is capsulated.

Classification of True Fibroma

1. *Soft fibroma*—contains immature fibrous tissue. Common in face, presents as soft brown swelling.
2. *Hard fibroma*—contains well-formed fibrous tissue.

Note:

True fibroma is rare and cannot be diagnosed clinically. It is mostly combined with mesodermal tissues like nerve sheath (neurofibroma), fat (fibrolipoma), muscle (fibromyoma).

Treatment

Treatment of true fibroma is excision.

An entity called ***aggressive fibromatosis*** is known to occur as unencapsulated proliferation of fibrous tissue, common in abdominal and chest wall. It is considered presently *as locally malignant*. It does not spread through lymphatics or through blood. *But recurrence is common.*

Desmoid tumour is a variant of aggressive fibromatosis, seen in females, often associated with Gardner's syndrome (Desmos = tendon, eidos = appearance). *Refer chapter Abdominal Wall and Umbilicus*.

Recurrent fibroid of Paget's is a rare type of fibrosarcoma occurring in a scar tissue after many years.

BURSAE

Bursa is a sac like cavity containing fluid within, which in normal location prevents friction between tendon and bone.

- Minor injuries and pressure leads into bursitis, which will present as a swelling at the site.
- Inflammation of this bursa due to friction causes ***bursitis***, which commonly presents as swelling, pain, and restricted movements.
- Bursa secretes *synovia like clear fluid* in a cavity lined by flat endothelium. It reduces the friction at the site between tendon and bone. Normally fluid content is little to cause a swelling. Minor trauma or infection causes sudden increase in fluid secretion of the bursa making it to enlarge and clinically palpable as pathological bursitis. Bursa is common around knee, elbow, heel and hip.
- Long-standing bursitis leads into thickening of its wall often with calcification making it feel hard with indurated surface. Lining of bursa may become rough or fluid may contain loose fibrinous particles to create *grating sensation* (*crepitus*) on the surface.
- Often overlying skin becomes thick, cracked and horny due to repeated friction and inflammation.
- Bursa may get adherent to deeper tissue as well as overlying skin to make it immobile. Bursa is usually well-localised, smooth, fluctuant, nontender swelling. Often it can be bilateral—in knee or elbow.
- Joint related should be examined. Bursa *may be* communicating with the adjacent joint.
- Gout or rheumatoid arthritis can cause bursa. For example, olecranon bursa can develop in gout patient.
- Bursa should be differentiated from cold abscess, soft tissue tumour, aneurysm, synovial tumour (sarcoma) at different locations.

Complications of Bursa

- Infection of bursa can occur due to trauma to overlying skin or through blood.
- Mechanical disturbances and discomfort.

Management

- US of the anatomical site, X-ray of the part or MRI are very useful.
- Avoiding friction and other aggravating factors may control many bursae.
- Aspiration and steroid injection may be useful.
- Bursa which is felt indurated with thick wall or calcified or infected or attained large size or which interferes with joint movement or daily activities needs surgical excision. Subcutaneous bursa can be excised under local anaesthesia; large or deeper bursa requires general anaesthesia for excision.

Different Types

It can be ***anatomical or adventitious***.

Anatomical

- *Anatomical bursae* are located normally in a particular anatomical site with a purpose of reducing friction. They are commonly deep and adjacent to a bone or joint.
- They become *pathological* and clinically significant when it presents with *bursitis*.
- They are soft, cystic, well localised, nontransilluminating swelling at known anatomical site.
- *Subhyoid bursa*: An horizontally oval swelling situated below the hyoid bone and in front of the thyrohyoid membrane.
- *Subacromial bursa*: In front and lateral to humeral head in relation to supraspinatus tendon between acromion and greater tuberosity of humerus.
- *Bicipito radial bursa.*
- *Olecranon bursa* (*Student's elbow*, *Miner's elbow*): It is subcutaneous bursa in relation to olecranon which becomes distended due to prolonged periods of leaning over elbow. Gout may involve this bursa.
- *Psoas bursa*: A tensely cystic swelling situated beneath and below the inguinal ligament, in the lateral aspect of the femoral triangle. But it will not extend above the inguinal ligament in to the iliac region (unlike in psoas abscess which extends above and is *cross fluctuant*). Psoas bursa lies between the psoas tendon and lesser trochanter. When it is enlarged, it causes diffuse swelling over outer part of femoral triangle lateral to femoral vessels. When hip is moved swelling becomes painful. It also should be differentiated from femoral hernia.
- *Prepatellar bursa* (*Housemaid's knee/miner's beat knee*)*:* It lies subcutaneously in front of lower half of patella and upper half of patellar tendon (upper part) undergoes inflammation in people who do much kneeling. Joint is normal here.
- *Subcutaneous infrapatellar bursa* occurs between skin and lower part of the tibial tuberosity and ligamentum patellae. It is called as *Clergyman's knee*. 'Clergyman' is Christian priest who kneels down during prayer.
- Suprapatellar bursitis is deep to patella and vastus intermedius, in front of lower end of femur. It communicates with knee joint.

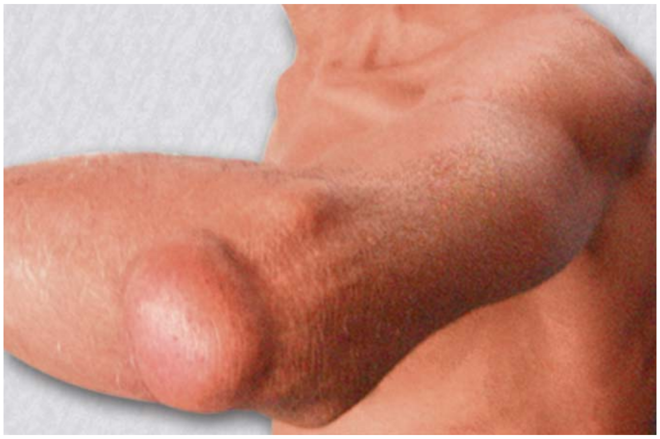

Fig. 1.183: Bursa near elbow joint.

There is only one pretty child in the world, and every mother has it.— **Chinese proverb.**

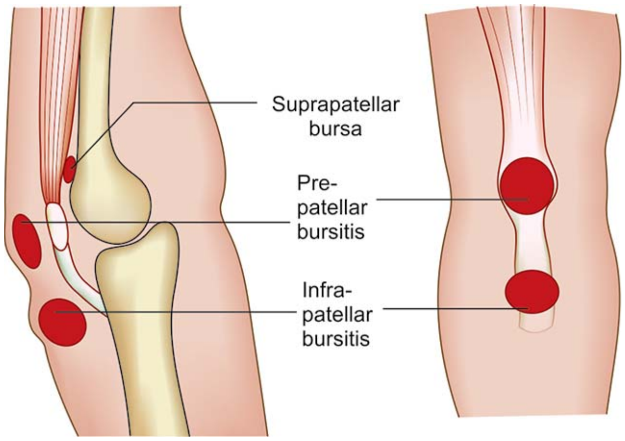

Fig. 1.184: Prepatellar (*Housemaid's knee*) and infrapatellar bursae (*Clergyman's knee*).

- ◆ *Brodie's bursa* lies deep to medial head of gastrocnemius.
- ◆ *Semimembranosus bursa*
- ◆ *Bursa anserina*—under the tendons of Guy ropes (sartorius, gracilis and semitendinosus tendons—*Goose's foot*).
- ◆ *Retrocalcaneum bursa*—between calcaneum and tendo-Achilis.

Adventitious Bursa

- ◆ Adventitious bursa occurs in an *unusual place/site* due to friction/pressure between two layers of tissue. Once it becomes chronic it may get adherent to overlying skin or tissue underneath.
- ◆ It may get infected, wall gets calcified, and fluid may become thick. It is well-localised, cystic usually nontender swelling. It becomes painful and tender if infected.
- ◆ Lining of bursa may become rough or fluid may contain loose fibrinous particles to create grating sensation/crepitus on the surface.

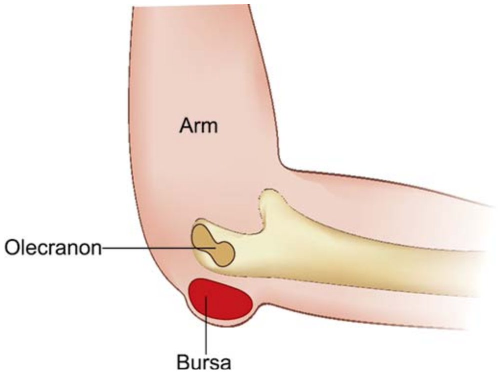

Fig. 1.185: Location of olecranon bursa (*Student's elbow*).

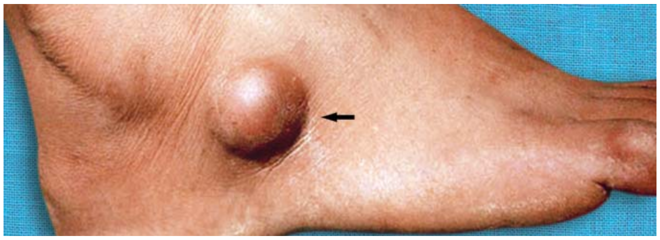

Fig. 1.186: Adventitious bursa over lateral aspect of foot—a common site.

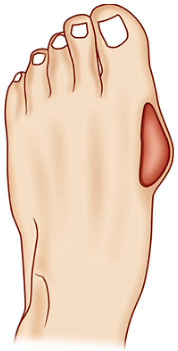

Fig. 1.187: Common site of adventitious bursa (*Bunion*).

- ◆ **Examples**
  - *Bunion* is adventitious bursa in patient with hallux valgus occurring between head of first metatarsal and skin.
  - *Tailor's bursa* occurs between lateral malleolus and skin.
  - *Porter's bursa* occurs between skin over shoulder and clavicle.
  - *Weaver's bursa* occurs between gluteus maximus, ischial tuberosity and skin.
  - *RetroAchillis bursitis* occurs between skin and Achilles tendon.
  - *Subcalcaneal bursitis* occurs between calcaneum and heel in long distance runners.
  - *Billing gate hump* appears over 7th cervical spine deep to overlying skin in people carrying weight over it. Billing gate is a large *fish market* in London.
- ◆ Condition should be *differentiated* from soft tissue tumour, sebaceous cyst, ganglion (depending on the location of bursa).
- ◆ **Management**: X-ray of the part and FNAC of swelling should be done. Later it is excised usually under local anaesthesia.

SEMIMEMBRANOSUS BURSA

It is a cystic swelling in the *upper medial* aspect of the popilteal fossa under the semimembranosus tendon. It is said to be due to friction under the tendon causing bursitis.

Features

- ◆ It is located between semimembranosus tendon and femoral condyle *above* the knee joint line.
- ◆ It is *common* in young adult; in *both* sexes. It is *most common* swelling of the popliteal fossa.
- ◆ It is nontender, cystic/tensely cystic (firm) swelling located above and on medial aspect of the popliteal fossa, fluctuant, noncompressible, often transilluminating, often with a fluid thrill. When it enlarges it comes out of semimembranosus tendon to become subcutaneous.
- ◆ Content of bursa *does not* communicate with knee joint. So fluid cannot be reduced into the joint cavity; but often appears *flaccid* on flexion of knee or by firm pressure probably due to displacement of fluid into deeper recesses of the bursa. The swelling becomes *tense* when knee is extended.
- ◆ Knee joint is *normal.*

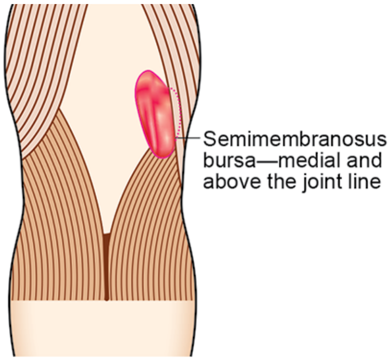

Fig. 1.188: Semimembranosus bursa typical location.

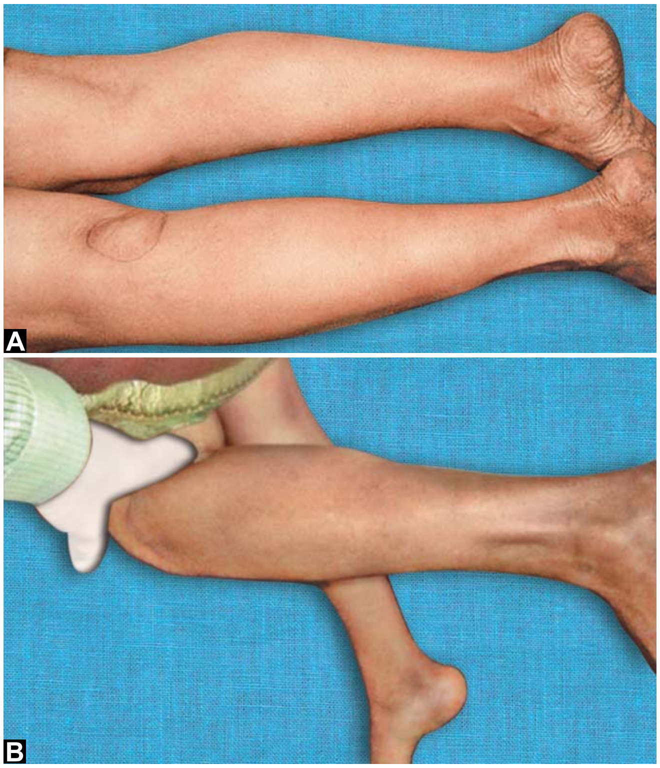

Figs 1.189A and B: Semimembranosus bursa—typical location. On flexion it only becomes flaccid but *does not disappear.*

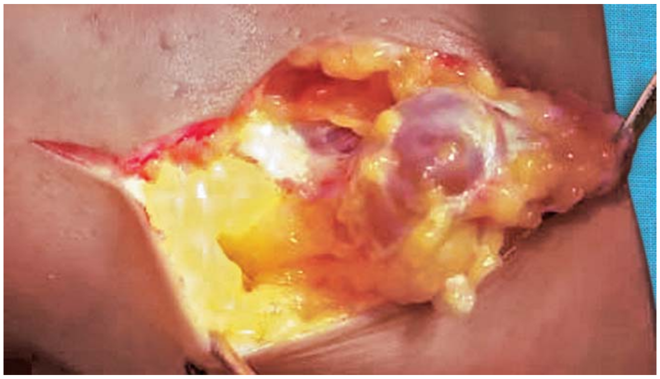

Fig. 1.190: Semimembranosus bursa on table look.

Differential Diagnosis

- Baker's cyst, popliteal aneurysm.

Treatment

- Ultrasound of popilteal fossa shows the cystic swelling under semimembranosus tendon.
- X-ray knee joint is normal.
- Excision is done under general anaesthesia using tourniquet. Complete excision of the sac is needed to prevent recurrence.

MORRANT BAKER'S CYST

It is a cystic swelling containing gel like fluid in the *lower midline* of the popliteal fossa. It occurs due to herniation of the synovial membrane of the knee joint as a result of chronic arthritis.

- It is *pulsion/pressure diverticulum* of the synovial membrane towards surface under the gastrocnemius through an opening in the joint capsule. It is *below* the joint line.

Clinical Features

- It is common in middle-aged individuals.
- It is smooth, soft and cystic, nontransilluminant, often tender swelling located below (the joint line) and in midline of the popilteal fossa.

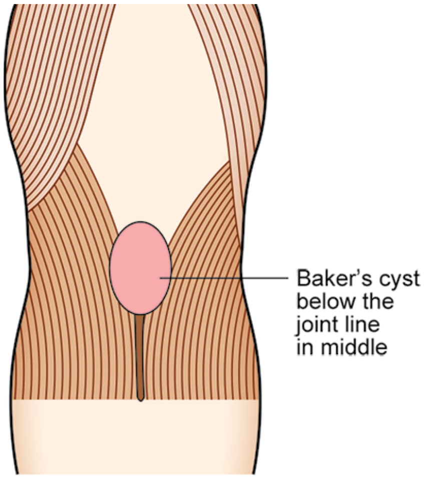

Fig. 1.191: Morrant Baker's cyst—typical location.

- On flexion swelling *disappears* and on extension swelling *increases* in size.
- Pain and tenderness are present in knee joint with effusion showing positive patellar tap.
- The knee joint movements are painful and restricted.
- Baker's cyst may *rupture* sometimes causing severe sudden pain and swelling in the calf mimicking *deep vein thrombosis.*

The mother's heart is child's schoolroom.—**Henry Ward Beecher.**

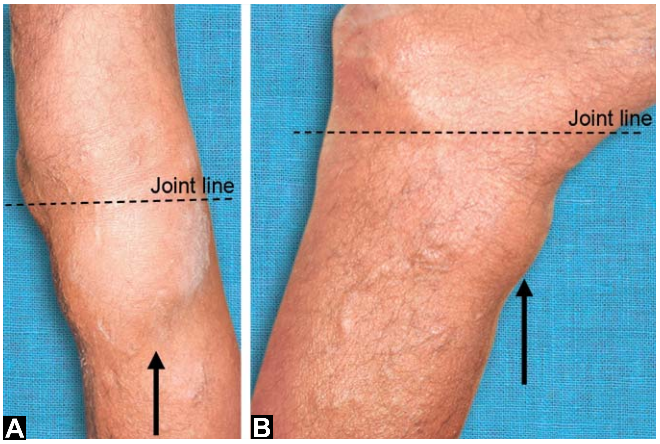

Figs 1.192A and B: Baker's cyst. It is below the joint line.

Differential Diagnosis for Baker's Cyst

- Semimembranosus bursa.
- Thrombosed popliteal aneurysm (often bilateral).

Management

- X-ray of joint shows arthritic changes.
- MRI may be needed occasionally.
- Arthritis is treated and Baker's cyst is excised under *general anaesthesia in prone* position.

LYMPH CYST (LYMPHATIC CYST)

- It is an *acquired type* of *distension cyst* wherein lymphatics form a localised swelling with a capsule around it.
- It usually occurs in subcutaneous plane, which is smooth, soft, nontender, mobile, and brilliantly transilluminant. It is usually not adherent to the skin.
- **Common sites** are neck and limbs.
- It can get infected and form an abscess.

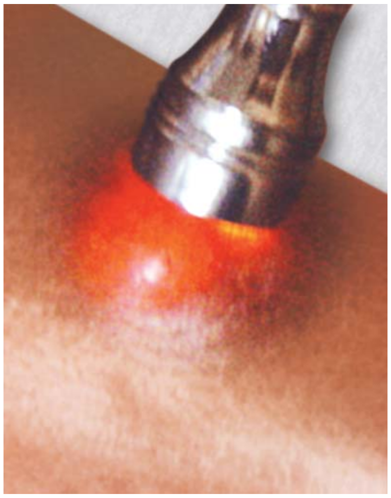

Fig. 1.193: Lymph cyst, which is transilluminant. It is an acquired condition.

Differential Diagnosis

- Cold abscess, dermoid cyst.

Treatment

- Excision.

LYMPHANGIOMA

It is congenital localised clusters of dilated lymph sacs in the skin and subcutaneous tissue that has failed to join the normal lymph system during development period.

Types

Capillary Lymphangioma

- It is simple type which can be present at birth but noticeable skin vesicles often develop in few years.
- It is common at the junction of body to limbs-like near shoulder, axilla, groin or buttock.
- Skin vesicles contain clear watery or yellow fluid. Bleeding within the vesicle may turn it into brown or black. Its features includes multiple, indistinct white/brown/black coloured vesicles of 0.5 to 4 mm size at typical locations in children involving around 5-20 cm area of skin in the particular location.
- If it is less than 5 cm in size it is called *as lymphangioma circumscriptum*.
- If it is more than 5 cm in size it is called as *lymphangioma diffusum*.
- If it is with reticulated ridges, it is called as *lymphoedema ab igne*.
- Area is soft, spongy, often fluctuant with fluid thrill and translucency. It is not compressible. Vesicles will not fade on pressure.
- Often lesion may get infected to make it painful and tender.
- Condition will not block the lymph drainage in normal lymphatics and skin oedema is absent.
- Regional lymph nodes are not enlarged.

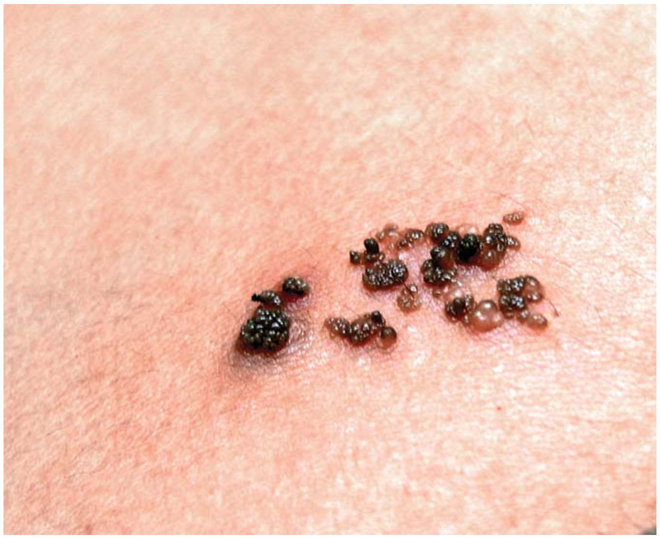

Fig. 1.194: Lymphangioma circumscripta (*Courtesy*: Balasaraswathy, MD, Consultant Dermatologist, Mangalore)

Cavernous Lymphangioma

- It is soft, lobulated, fluctuant, *brilliantly transilluminant* larger lymphatic swelling with often multiple communicating lymphatic cysts.
- It often extends into deeper plane like muscle. It is common in face, mouth, lips *(macrocheilia)*, tongue *(macroglossia)*.

Cystic Hygroma

It is collection of clustered sequestered lymph sacs (occurring during developmental period *in utero*) presenting in newborn as large swelling which is soft, smooth, fluctuant, brilliantly transilluminant, and *compressible* (For detail refer chapter 5, Neck).

CALCINOSIS CUTIS

- It is a type of calcification (*dystrophic*) in or under the skin. Usually presents as a circumscribed lesion in the skin.
- Commonly seen in females. Common site is in the waist (Fig. 1.195)
- Usually bilateral.
- It is said to be due to friction causing degeneration of skin and immediate deeper structure with increased local alkalinity of the tissue causes precipitation of the calcium leading to solid, hard, swelling in the skin. Cut section shows hard, yellowish material.
- It may mimic calcified lipoma or neurofibroma.
- **Treatment** is excision and closure of defect often with local flaps.

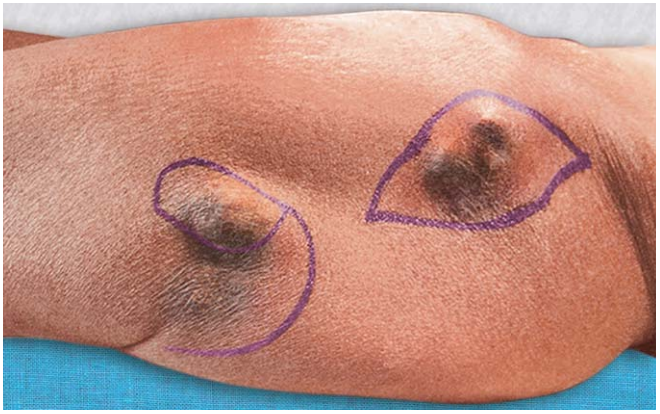

Fig. 1.195: Calcinosis cutis near waist is a common site. It is common in females.

NEUROMA

Two types of neuromas are found.

- **False neuroma.**
- **True neuroma.**

False Neuroma

- It occurs due to injury to the nerve either after trauma or amputation which presents as a tender swelling.
- It arises from the connective tissue of nerve sheath.
- It contains fibrous tissue with coiled nerve fibres.
- It can be ***end neuroma,*** commonly seen in amputation stump as a tender localised firm swelling often adherent to the stump scar. It can cause neuralgia in the stump. It prevents the proper usage of prosthesis. So during amputation, nerve should be pulled down and cut so as to make cut end to retract more proximally, which prevents the formation of neuroma.
- It can be ***lateral (side) neuroma*** which is observed after partial nerve injury, usually after trauma. It presents as a tender, firm swelling along the line of the peripheral nerve.

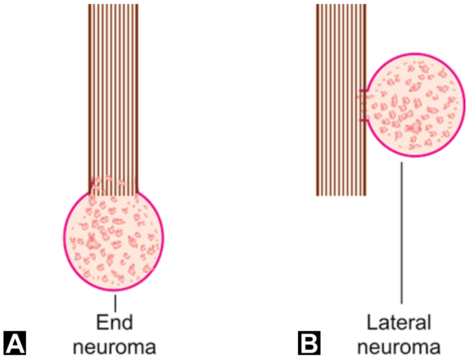

Figs 1.196A and B: Types of false neuromas: (A) End neuromas: (B) Lateral neuromas.

True Neuroma

- It is a rare tumour.
- It occurs in connection with sympathetic system.

Types

Ganglioneuroma

- It contains ganglion cells and nerve fibres.
- It occurs in connection with sympathetic chain. So it is observed in neck (parapharyngeal mass), thorax, retroperitoneum, adrenal medulla.
- It is relatively benign and symptomless and often attains a large size.
- Early complete excision can give a cure.

Neuroblastoma

- It is poorly differentiated, aggressive, embryonic type of tumour.
- It is seen in infants and children.
- It spreads through blood but can go for spontaneous remission occasionally.

Myelinic Neuroma

- It contains only nerve fibres.
- Ganglion cells are absent.
- It occurs in spinal cord or pia mater.

Move to the rhythm of soul and you'll never miss a beat.—**Vicki Virk**

NEUROFIBROMA

- **It is a benign tumour arising from connective tissue of the nerve containing ectodermal neural and mesodermal connective tissue components.**
- It can be single or multiple. Neurofibromas may be associated with pheochromocytomas, hypertension and few syndromes.

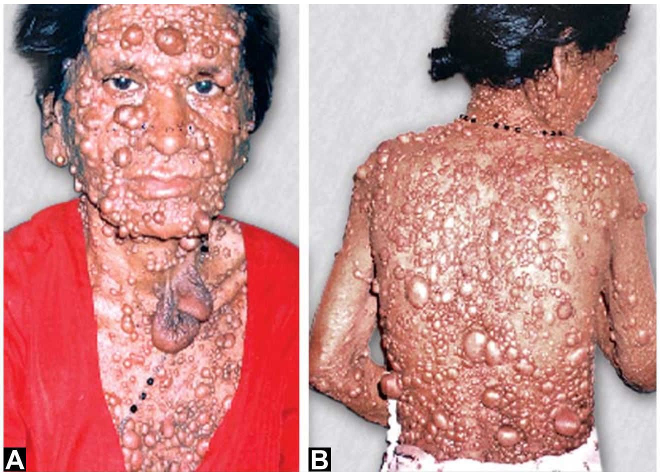

Figs 1.197A and B: Multiple neurofibroma.

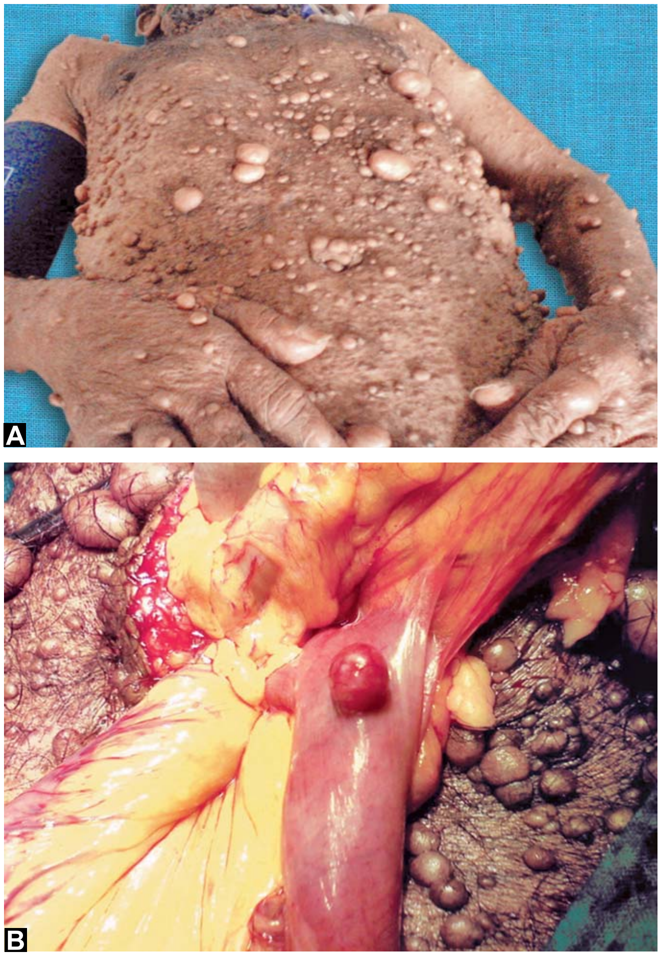

Figs 1.198A and B: Patient with multiple neurofibromatosis underwent laparotomy showing serosal neurofibroma.

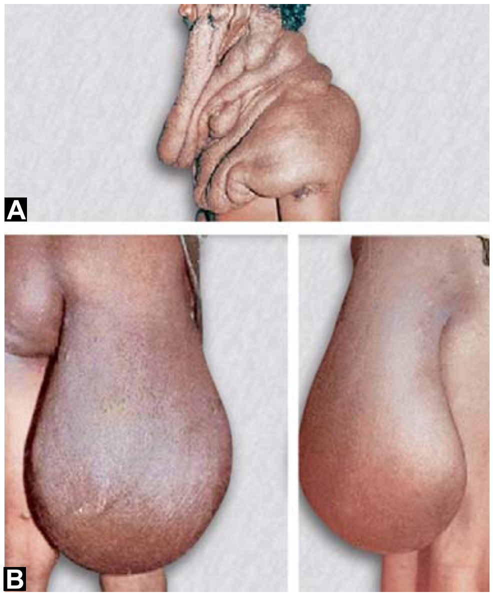

Figs 1.199A and B: Plexiform neurofibroma in the (A) Face and neck; and (B) Buttock.

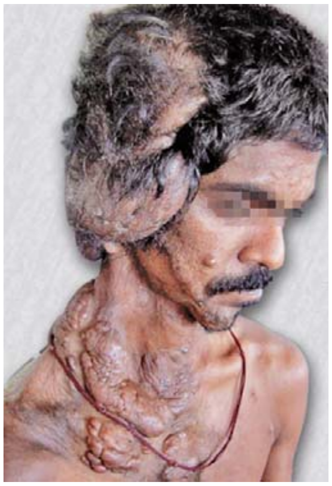

Fig. 1.200: Plexiform neurofibromatosis involving trigeminal (5th cranial) nerve.

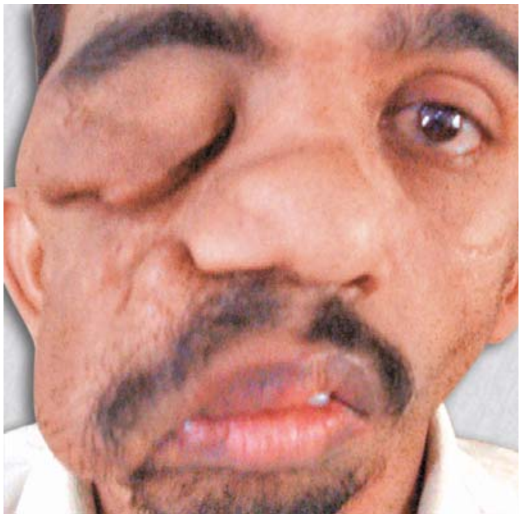

Fig. 1.201: Plexiform neurofibroma involving ophthalmic division of trigeminal nerve.

Sites

- Cranial.
- Spinal. Neurofibroma is the most common intradural extramedullary spinal tumour.
- Peripheral.

Types

- ***Nodular neurofibroma:***
    - It presents as single smooth, firm, tender (often) swelling which moves horizontally or perpendicular to the direction of the nerve, not in the direction of the nerve.
    - There is pain and hyperaesthesia in the distribution of the nerve.
- ***Plexiform neurofibroma:***
    - It commonly occurs along the distribution of 5th cranial nerve in the skin of face. It often occurs in the cutaneous distribution of the peripheral nerve.
    - It attains enormous size with thickening of the skin which hangs downwards. It causes erosion into the bone, orbit and deeper structure.
    - It may also undergo myxomatous degeneration. It causes cosmetic problem.
    - Paraesthesia along the distribution of the trigeminal nerve (commonly ophthalmic division) is common.
    - **Pachy dermatocele**—a variant of plexiform neurofibroma where neck is involved. There is thickening and oedema of skin, pigmentation, thromobosed veins with enormous proliferation of subcutaneous nerve fibers causing *folded pendulous hanging thickened skin* with tissues. It is *common* along trigeminal (5th cranial) nerve; rarely is it seen in limbs and scalp skin.

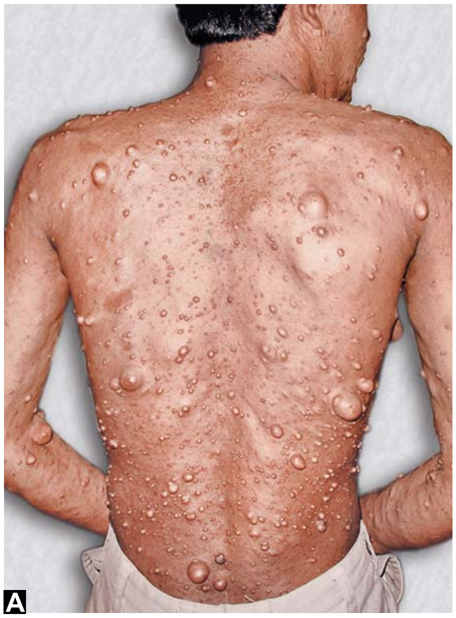

Fig. 1.202A

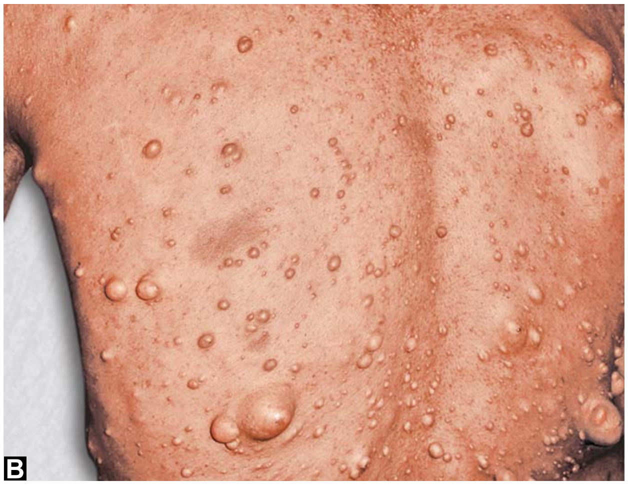

Fig. 1.202B

Figs 1.202A and B: Multiple neurofibromatosis with café-au-lait spots in the skin.

- ***Generalised neurofibromatosis (von Recklinghausen's disease***): (1:4000 births)
    - It is an inherited autosomal dominant disease wherein there will be multiple neurofibromas in the body—chromosome 17. It is called as type 1 in which more than 50% will be familial. *Type 1*, more commonly, called as *von Recklinghausen's* disease. Type 2 is *acoustic neurofibromatosis* (mutation of chromosome 22). All layers of nerve with Schwann cells and fibroblasts show tumour proliferation.
    - It may be cranial, spinal or peripheral.
    - It is associated with pigmented spots (coffee coloured) in the skin, commonly seen on the back, abdomen, thigh ***(café au lait spots)***. More than 5 in number, with each 1.5 cm or more in size is significant. *Café-au-lait* signifies common neuroectodermal origin of nerve sheath cells and melanocytes.
    - *Axillary/groin freckles and Lisch nodules,* pigmented iris, hamartoma may be present.
    - Neurological disturbance is uncommon.
    - It may be associated with MEN type IIb (Multiple neurofibromas of eyelids, lips, and face; medullary carcinoma of thyroid; pheochromocytoma, hyperparathyroidism), primary brain tumours and bone cysts.
- ***Elephantiatic neurofibromatosis***:
    - It is of congenital origin. It involves limbs.
    - Skin of the limb is greatly thickened, dry and coarse.
- ***Cutaneous neurofibromatosis*:**
    - They are small, multiple, firm/hard nodules arising from terminal ends of dermal nerves. Overlying skin is normal without any changes. It commonly occurs all over the body. It can be pedunculated or sessile. It can occur in scalp to result in a *turban tumour.*

Clinical Features

- Mild pain or painless swelling usually in subcutaneous or cutaneous plane with tingling, numbness and paraesthesia along the distribution of nerve.

- Round/oval/fusiform swelling along the peripheral/cutaneous nerves which moves perpendicular to the direction of the nerve but not along the longitudinal direction of nerve fibre with smooth surface and firm consistency. It is nontender or mild tenderness may be present. Skin is free unless it is of cutaneous type. Cutaneous type is often soft.

Complications

- Cystic degeneration.
- Haemorrhage into the tissues.
- Spinal and cranial neurofibromas can cause neurological deficits.
- Erosion into deeper planes, bone, orbit.
- Muscle atrophy.
- Spinal dumbbell tumour can cause compression of spinal cord and paralysis of the limb.
- *Sarcomatous changes*: Common in generalised type (5%). When it occurs it shows rapid enlargement, warmness, more vascularity with dilated veins. Secondaries in lungs can occur through blood spread.

Note:

- *The most common* spinal tumour is neurofibroma.
- von Recklinghausen disease of *primary hyperparathyroidism* is different as it is due to raised PTH level causing *osteitis fibrosa cystica*.
- Familial type of neurofibroma may be associated with *scoliosis*.
- Café-au-lait spots is also seen in *McCune Albright* syndrome.

Treatment

- Excision.

Indications

- Symptomatic neurofibroma—pain and pressure symptoms
- Cosmetically problematic lesion
- Recent increase in size
- Malignant transformation 5%

Remember about neurofibroma

- Neurofibroma is fusiform swelling which has got horizontal mobility but not longitudinal mobility (neurilemmoma has got free mobility in all directions)
- Café au lait spots signify common neuroectodermal origin of nerve sheath and melanocytes
- Kyphoscoliosis may be an association in familial variety
- Phaeochromocytoma with hypertension may be an association (20%)
- *Plexiform neurofibromatosis* commonly involves trigeminal cranial nerve mainly ophthalmic division causing paraesthesia, can obstruct the vision when it is large, have grossly thickened pendulous skin hanging down to various levels
- Calcification, saponification, myxomatous changes, sarcomatous transformation (fungation, rapid recent increase in size, warmness, dilated surface veins, redness with increased vascularity, fixity, persistent severe pain, immobility, lung spread) and cosmetic problem are the complications
- Intestinal neurofibroma may precipitate intussusception

NEURILEMMOMA (SCHWANNOMA)

It is arising from Schwann (neurilemmal cells) cells. They are lobulated, encapsulated, soft, whitish grey in appearance. They displace the nerve from which they arise and can be removed. They are common in acoustic nerve (cerebellopontine angle) but also can occur in a peripheral nerve. Occasionally they are multiple. Calcification is common. It is *ectodermal* in origin.

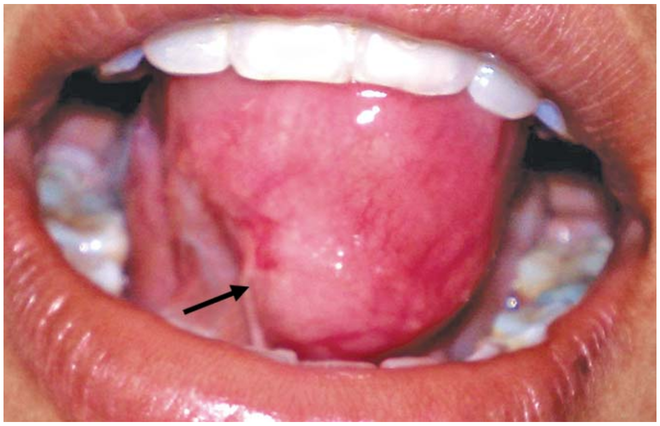

Fig. 1.203: Lingual schwannoma (*Courtesy:* Dr Harish Rao, Professor in Surgery, KMC, Mangalore).

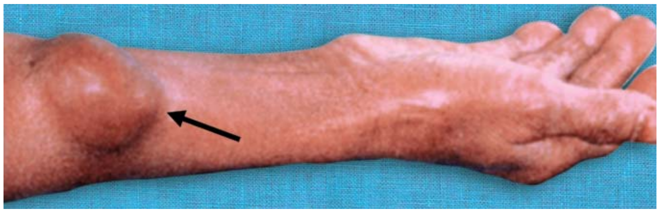

Fig. 1.204: Schwannoma in the forearm.

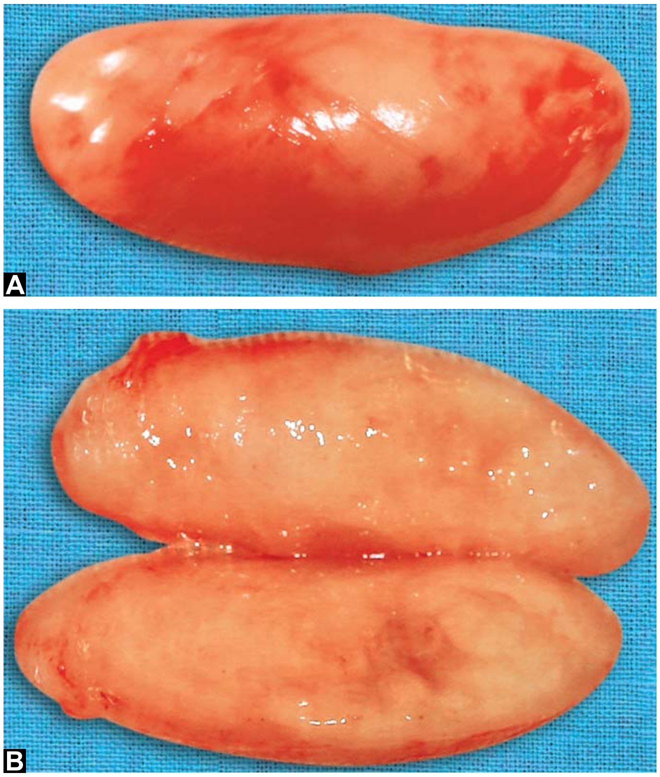

Figs 1.205A and B: Neurilemmoma—gross look and cut section.

Two types:

Anthony A—Two rows of spindle cells with central acellular area (Verocay bodies).

Anthony B—Acellular myloid areas.

Presentation is pain along the distribution of the nerve, hyperaesthesia and tenderness and as, soft or firm, lobulated, well localised mobile swelling.

Treatment

- Excision—easier without causing neurological damage.

Recurrent schwannoma could be malignant.

GANGLION

It is a cystic swelling occurring in relation to tendon sheath or synovial sheath or joint capsule. It contains clear gel like fluid.

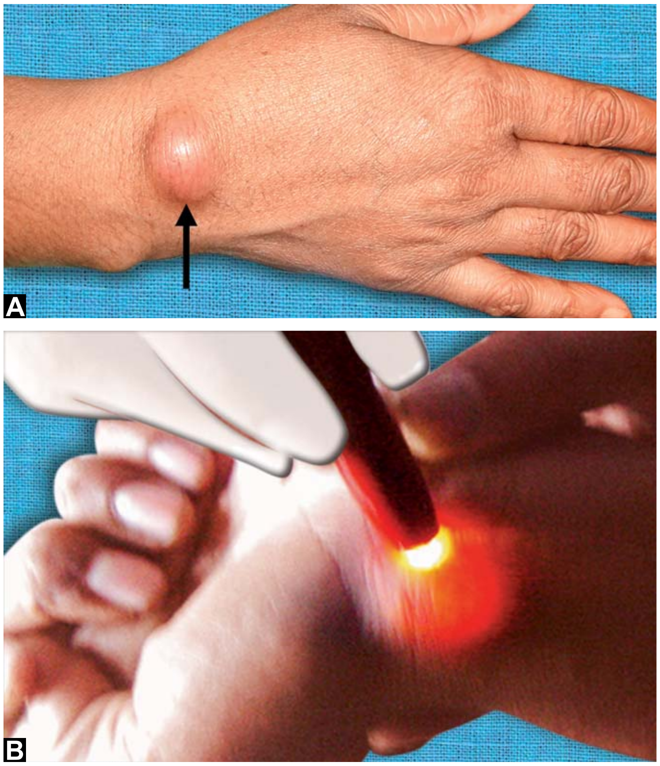

Figs 1.206A and B: Ganglion over the dorsum of the wrist—typical location. It can occur over the flexor aspect also. It can be transilluminant also.

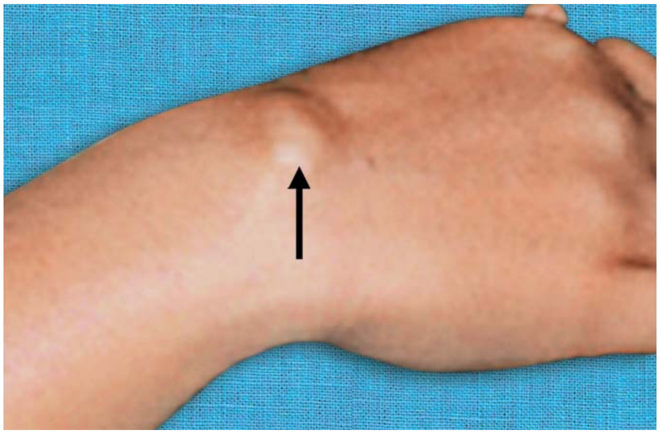

Fig. 1.207: Ganglion over the wrist.

Common sites

- Dorsum of wrist (Near scaphoid-lunate articulation)
- Flexor aspect of wrist
- Around ankle joint—occasionally

Pathogenesis

- Cystic degeneration of the tendon sheath.
- Leakage of synovial fluid through joint capsule.
- There are small islets of microspaces in synovial sheath which often fuse together or one of them gets enlarged to form ganglion.

Clinical Features

- Well-localised swelling which is smooth, soft, cystic, or tensely cystic, (Paget's test is +ve), nontender, transilluminant. It is mobile but mobility is restricted when tendon is contracted against resistance.
- Occasionally it is communicating with joint capsule.
- Often pain, tenderness and restricted joint movement may be the presentation (but rare).

Differential Diagnosis

- Lipoma.
- Lymph cyst.
- Sebaceous cyst.
- Small ganglion often mistaken for sesamoid bone or exostoses.
- Bursa.

Treatment

- Asymptomatic ganglion is better left alone because of high recurrence rate.
- **Excision.** Usually done under local anaesthesia (lignocaine plain 2%). Patient should be explained of **high recurrence** rate (30%). After excision always it should be sent for histopathology. Firm crepe bandage application for 4 weeks is better in these patients.
- Aspiration and sclerosant injection may be useful.

(In olden days people used to rupture the ganglion using bible book).

CHORDOMA

Tumour arising from the remnants of notochord.
It is seen in:

- Sacrococcygeal region.
- Sphenoid sinus region.
- Around the foramen magnum.

- It invades into the surrounding structure like nerves.
- Resection is difficult.
- Radioresistant.

EPIGNATHUS

This is a type of growth anomaly seen in neonates wherein growth from the base of skull protrudes through the mouth.

We seldom think of what we have, but always of what we lack.

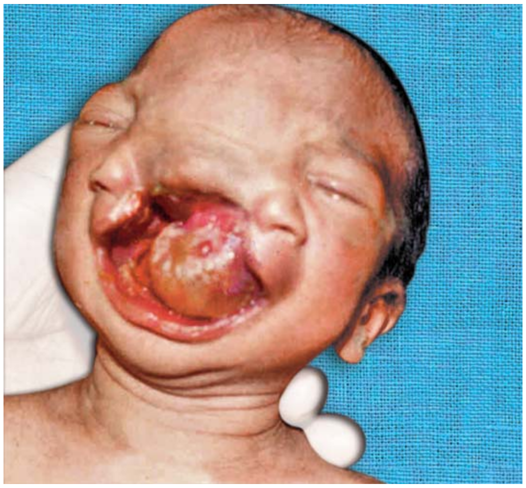

Fig. 1.208: Epignathus.

Note:

- *Enucleation* is removal of the swelling within the tissue of origin with normal part of tissue of origin is being retained, e.g. enucleation of prostate in benign prostatic hyperplasia (BPH).
- *Excision* is removal of tissue/tumour entirely with its capsule.
- *Wide excision* is removal of tumour with surrounding tissue margin adequately for clearance.
- *Compartment excision* is removal of tumour/diseased tissue with all adjacent soft tissues in one compartment except neurovascular bundle. It is done in limbs for soft tissue sarcoma as a curative but limb saving procedure.
- *Radical excision/radical block dissection* is removal of tumour widely with adjacent soft tissues with lymph node dissection.

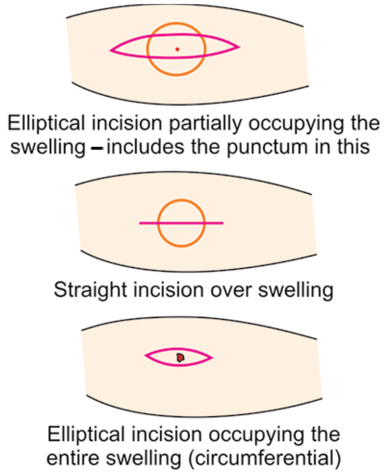
Elliptical incision partially occupying the swelling – includes the punctum in this
Straight incision over swelling
Elliptical incision occupying the entire swelling (circumferential)

Fig. 1.209: Different incisions used in surgical approaches to remove swelling.

Swellings which are cross fluctuant

- Psoas abscess
- Bilocular hydrocoele
- Ranula (plunging)
- Compound palmar ganglion

F. Electrolyte and Nutrition

CHAPTER OUTLINE

- Normal Physiology
- Water Loss
- Water Excess
- ECF Loss
- ECF Excess
- Hyponatraemia
- Hypernatraemia
- Hypokalaemia
- Hyperkalaemia
- Hypermagnesaemia
- Hypomagnesaemia
- Acid-Base Balance
- Metabolic Alkalosis
- Respiratory Alkalosis
- Metabolic Acidosis
- Respiratory Acidosis
- Anion Gap
- Fluid Therapy
- Nutrition
- Gastrostomy
- Jejunostomy
- Total Parenteral Nutrition
- Refeeding Syndrome
- Obesity and Morbid Obesity

NORMAL PHYSIOLOGY

Total body water is 60% of body weight in males, 50% of body weight in females, i.e. 30 litres.

- Intracellular water—20 litres (2/3).
- Extracellular water—10 litres (1/3).
  - Plasma (1/4) (2.5 litres).
  - Interstitial fluid (7.5 litres).

Ion	*ICF*	*ECF and plasma*
Sodium	10 mmol/L	140 mmol/L
Potassium	150 mmol/L	4.5 mmol/L
Chloride	Trace only	105 mmol/L

ECF volume and osmolality regulation is controlled by three hormones. Aldosterone, ADH, atrial natriuretic hormone.

WATER LOSS (VOLUME LOSS)

It is decrease in the whole body fluid volume which includes both ECF and ICF. It is usually ECF loss which is more important and assessed. It can be *isotonic volume depletion* with both salt and water loss leading into *hypovolaemia*, or *only water loss* with only minimal loss of electrolytes leading into *dehydration*.

Causes and Features

- *Isotonic volume depletion* occurs due to diarrhoea, vomiting, and excess diuresis. Here normal or decreased sodium is observed. Fluid loss is only of ECF and so early intravascular volume reduction occurs. This causes hypotension and decreased tissue perfusion. Features are—dry tongue, rapid pulse, cold clammy extremities, sunken eyes, hypotension, oliguria, raised blood urea, decreased urinary sodium. Hypovolaemia can be mild (< 2l fluid loss); moderate (2-3 L fluid loss); severe (>3 L fluid loss).
- *Only pure water* loss occurs due to poor fluid intake and diabetes insipidus. It causes dehydration with proportionate decrease in total body water (2/3rd ICF, 1/3rd ECF). As ECF including intravascular fluid loss is less, hypotension is less. Features here are—severe thirst, confusion and convulsions due to hypernatraemia; blood pressure is relatively normal. Dehydration can be mild (weight loss 5%); moderate (10%); severe (15%).

Management

- Evaluation is done by doing serum sodium, urinary sodium, and blood urea.
- Isotonic volume depletion is corrected by 0.9% normal saline.
- Pure water depletion is corrected by more water intake/intravenous 5% dextrose.
- Monitoring fluid therapy by skin and tongue examination, weight gain, pulse, blood pressure, CVP, PCWP.

WATER EXCESS (ECF VOLUME EXCESS)

It can be divided into *water and salt excess* or *predominantly water excess* called as *water intoxication*.

Water and salt excess occurs in CCF, cirrhosis, nephrotic syndrome, hypoproteinaemia, renal failure, excessive saline infusion.

Water intoxication occurs in TURP, excess infusion of 5% dextrose only, SIADH secretion, psychogenic polydypsia. It is managed by stopping fluid infusion or procedure (TURP); fluid restriction, and treating the cause.

Causes

- Excessive amount of intravenous dextrose (5%).
- During colorectal bowel wash for preparation of large bowel for surgery, if water is used instead of saline, especially in children.
- In TURP (Transurethral resection of prostate) when excess irrigating fluid water or glycine is used (commonly used).
- In syndrome of inappropriate antidiuretic hormone (SIADH) which is commonly associated with lobar pneumonia, empyema, oat cell carcinoma and head injury.

Clinical features

- Drowsiness, weakness
- Convulsions and coma
- Nausea, vomiting
- Passage of dilute urine
- Distended neck veins
- Pedal oedema
- *Gain in body weight*—most sensitive and consistent sign
- Circulatory overload—tachycardia, pulmonary oedema, hypertension
- Bilateral basal crepitations, ascites
- Raised CVP, PCWP

Investigations

- Haematocrit and sodium level (will show fall in level).
- Low potassium. Low blood urea.

Treatment

- Water and salt restriction and observation.
- Monitoring in ICU.
- Management of fluid and electrolyte balance.
- Infusion of hypotonic sodium chloride.

Administration of diuretics and hypertonic saline should be avoided, as it may cause rapid changes in serum sodium and water level which will lead to neuronal demyelination and *fatal outcome.*

ECF LOSS

- Here only ECF loss is present with normal ICF.
- It is seen in vomiting, diarrhoea, intestinal obstruction.
- Treatment is infusion of normal saline.

ECF EXCESS

- Only ECF excess without an ICF excess.
- Excessive infusion of saline with impaired excretion.
- Raised JVP (earliest and best clinical sign), cardiac failure and peripheral oedema.
- Treatment is fluid restriction and diuretics like frusemide.

HYPONATRAEMIA

Sodium level less than 130 mEq/litre. Hyponatraemia is said to be *severe* if serum sodium becomes lesser than 100 mEq/L in acute type; and lesser than 115 mEq/L in chronic type.

It can be due to water overload (dilutional) or sodium loss.

Types of Hyponatraemia

- Acute—presents as neurological manifestations
- Chronic—causes pontine myelinolysis. It presents as behavioural changes, progressive weakness, and cranial nerve palsies.

Types also may be:

- *Hypervolaemic hyponatraemia* wherein rapid absorption of fluid occurs into intravascular compartment leading into pulmonary and cerebral oedema. It is due to decreased osmolality causing movement of ECF into the cells. Serum sodium level lesser than 100 mmol/L is called as severe hyponatraemia, causes convulsions. Here urinary sodium will be less than 15 mmol/L. Acute hyponatraemia is corrected by fluid restriction, hypertonic saline, loop diuretics like frusemide. Monitoring the serum sodium level of the patient is essential. Sodium should be corrected up to above the level of 125 mmol/L. Correction should be slow and gradual at a rate of 2 mEq/L/hour with up to 20 mEq/L correction in 24 hours with 4th hourly assessment of serum sodium. Overcorrection of sodium should not be done. Rapid correction can lead into *irreversible myelin lysis of pontine*.
- *Hypovolaemic hyponatraemia:* It is due to hypovolaemia by diarrhoea, vomiting, wherein urine sodium level is less than 20 mmol/L; due to diuresis or renal causes wherein urine sodium level is more than 20 mmol/L or it may be due to correction of hypovolaemia using hypotonic fluid like 5% dextrose. Condition can be treated well using isotonic normal saline.
- *Normovolaemic hyponatraemia:* It may be due to renal failure or syndrome of inappropriate ADH secretion (SIADH). In mild asymptomatic patients it is corrected by fluid restriction (1 litre/day will raise the serum Na). Vasopressin antagonist *demeclocycline* which increases the diluting ability of kidney is used in severe cases.
- *Pseudohyponatraemia:* Plasma osmolality is mainly achieved by serum sodium; but small proportion, i.e. 25% of osmolality is due to other solutes like glucose, lipids, plasma proteins, urea which will not move easily between intracellular and extracellular spaces. When concentration of these molecules raise due to some pathology, proportionately relative concentration of sodium will drop causing pseudohyponatraemia. Here condition causing related to specific solutes mentioned above is treated, than hyponatraemia.

Causes

- Intestinal obstruction.
- Intestinal fistulas—biliary, duodenal, gastric, pancreatic.
- Gastric outlet obstruction with severe vomiting.
- Ryle's tube aspiration.
- Severe diarrhoea due to viral cause, in colitis, colorectal polyps.
- Syndrome of inappropriate antidiuretic hormone (SIADH).
- Immediately after surgery and trauma, sodium depletion occurs.
- Stroke.

Clinical features

- *Dry coated tongue*
- *Sunken eyes*
- *Dry wrinkled skin*
- *Hypotension*
- *Dark scanty urine*
- *Irritability, disorientation and neurological manifestations*
- *Convulsions*
- *In chronic hyponatraemia—hypothermia, reduced tendon reflexes, pseudobulbar pasly*

Investigations

- Serum electrolytes. Urinary sodium is low.
- Sodium *deficit* is calculated by: (125 – present serum sodium) × body weight in kg × 0.6.

Treatment

- Intravenous infusion of normal saline as a slow and gradual correction at a rate of 2 mEq/L/hour in acute cases and < 1 mEq/L/hour in chronic cases. Correction should not exceed more than 20 mEq/L/day in acute cases and more than 10 mEq/L/day in chronic cases. Hypertonic saline of 1.6% or 3% also can be used in severe cases. 0.9% normal saline contains 154 mEq of NaCl; 3% saline contains 500 mEq of NaCl.
- The cause is treated.

HYPERNATRAEMIA

Serum sodium level > 150 mEq/L. Excess infusion of normal saline causes overload in circulating salt and water. It is usually due to water deficit.

Causes

- Renal dysfunction.
- Cardiac failure.
- Drug induced like NSAID, corticosteroids.

It may be either primary sodium excess or primary potassium excess or primary water deficit.

Types of Hypernatraemia

- *Euvolemic* (pure water loss): It is due to failure of water intake like in comatous patients, bedridden people, postoperative patients and in patients with high fever leading into extrarenal loss of water. It can occur in diabetes insipidus or chronic renal failure as renal loss of water.
- *Hypovolaemic* (among loss of water and sodium, more water is lost than sodium): It is due to vomiting, diarrhoea, more undue sweating (extrarenal); osmotic diuresis by glucose/mannitol (renal).
- *Hypervolaemic* (both sodium and water gain but sodium gain is more than water gain) as seen in more salt intake, excess steroids, sodium bicarbonate/hypertonic saline infusion (salt gain).

Clinical Features

- Pitting oedema.
- Puffiness of face.
- Increased urination.
- Often dilated jugular veins.
- Features of pulmonary oedema.

Investigation

- Serum electrolytes, plasma and urine osmolality, renal function tests, haematocrit.

Management

- Restriction of saline and sodium. Treatment of pulmonary oedema.
- *Hypernatraemia should be corrected slowly as follows:*
  - Initial infusion of normal saline, then infusion of half strength saline (0.45%) and later with 5% dextrose, i.e. gradual controlled correction is done. Otherwise cerebral oedema and hyperglycaemia can develop.
  - Oral and nasogastric administration of water/fluids.

HYPOKALAEMIA

Sudden Hypokalaemia

Serum potassium level less than 3.5 mEq/L.

It occurs in patients in diabetic coma treated by insulin and saline infusion.

Gradual Hypokalaemia

Causes

- Diarrhoea of any causes, villous tumour of the rectum, ulcerative colitis.
- After trauma or surgery.
- Pyloric stenosis with gastric outlet obstruction.
- Duodenal fistula, ileostomy.
- After ureterosigmoidostomy.
- Insulin therapy.
- Poisoning.
- Drugs like beta agonists.
- Familial periodic paralysis.

Clinical Features

- Slurred speech.
- Muscular hypotonia—physical sign.
- Depressed reflexes.
- Paralytic ileus.
- Weakness of respiratory muscles.
- Cardiac arrhythmias.
- Inability to produce concentrated urine and so causes nocturia and polyuria.

Sir, it is plain that if he is too fat he has eaten more than he should have done.— **Samuel Johnson**

ECG shows prolonged QT interval, depression of the ST segment and inversion of T wave, prominent U wave.

Often hypokalaemia is associated with *alkalosis.*

Serum potassium will be decreased.

Treatment

- Oral potassium 2 g 6th hourly, 15 ml potassium chloride syrup (20 mmol of K).
- IV KCl 40 mmol/litre given in 5% dextrose or normal saline *slowly*, often under ECG monitoring [Total dose is 40 mmol (0.2 mmol /kg/hour). Maximum dose per hour is 20 mmol].
- *Hypokalaemic alkalosis which occurs in pyloric stenosis* should be treated carefully by IV potassium as there will be severe potassium loss.

HYPERKALAEMIA

Normal range of potassium is 4.0 to 4.5 mEq/litre.

Hyperkalaemia manifests when potassium ***exceeds 6 mEq/litre.***

Causes

- Renal failure.
- Rapid infusion of potassium.
- Transfusion of stored blood.
- Diabetic ketoacidosis.
- Adrenal insufficiency.
- Potassium sparing diuretics, cyclosporine, beta blockers.
- Metabolic acidosis.
- Insulin deficiency.
- Tissue destruction, burns, trauma, tumour necrosis, crush injury.
- *In vitro* haemolysis, thrombocytosis, tourniquet application, exercise—pseudohyperkalaemia.
- Familial hyperkalaemic periodic paralysis.

Potassium excess is a dangerous condition which can cause sudden cardiac arrest.

Investigations

High serum potassium level. Peak 'T' wave in an ECG.

Treatment

- IV administration of 50 ml of 50% glucose with 10 units of soluble insulin, slowly.
- Infusion of 10% calcium gluconate (as cardioprotection) intravenously.
- Calcium chloride is given in severe cases as calcium in this form is released immediately without hepatic metabolism.
- Diuresis using frusemide injection.
- Haemodialysis when required—very useful.
- Continuous ECG monitoring is a must.
- Polyesterene sulphonate ion exchange resin 30 g/hour in 50 ml of 70% sorbitol as an enema.
- Salbutamol nebulisation or intravenously 0.5 mg in 4 ml of saline/Albuterol nebulisation.
- IV sodium bicarbonate—shifts potassium in to cells. 7.5%, with 50-100 ml intravenously in 10 minutes.

HYPERMAGNESAEMIA

It is rare. Serum magnesium > 2.5 mEq/litre.

Normal *serum magnesium* is 1.5-2.5 mEq/L and intracellular magnesium which is more (2nd higher) is 26 mEq/L. Magnesium is mainly deposited in bone (60%). It is a cofactor for many enzymes necessary in phosphorylation of glucose in the cell and ATP utilisation in muscle fiber. Daily required dietary intake of magnesium is 0.4 gram. It is reabsorbed well in proximal renal tubule.

Causes

- Advanced renal failure treated with magnesium containing antacids, diabetic ketoacidosis.
- Intentionally produced hypermagnesaemia while treating pre-eclampsia.

Clinical Features

- Loss of tendon reflexes (most common).
- Neuromuscular depression.
- Flaccid quadriplegia.
- Respiratory paralysis.
- Somnolence.
- Hypotension.

HYPOMAGNESAEMIA

- Serum magnesium < 1.5 mEq/litre.

Causes

- Malnutrition, alcohol.
- Large GI fluid loss.
- Patients on total parenteral nutrition.

Clinical Features

- Hyperreflexia.
- Muscle spasm.
- Paraesthesia.
- Tetany.
- It mimics hypocalcaemia. It is often associated with hypokalaemia and hypocalcaemia.

Treatment

Two gram (16 mEq) of magnesium sulphate slow intravenously, in 10 minutes. Later maintenance dose of 1 mEq/kg/day as slow continuous infusion is given/oral magnesium is needed.

ACID-BASE BALANCE

Normal pH (– log 10 of H^+) is 7.36-7.44.

Factors which control the pH

- *Buffer system*
  - Bicarbonate buffer
  - Protein buffer
  - Phosphate buffer
- *Renal control of pH*
- *Respiratory control of pH*

Note:
When H^+ increases pH decreases.

An *acid* is a substance that dissociates water to release hydrogen ion. A *base* is a substance that takes hydrogen ion. A *buffer* is a combination of weak acid and conjugate base. These buffers maintain the H^+ concentration in blood within fine limits. Natural buffers are *extracellular or intracellular*. Bicarbonate/carbonic acid buffer, phosphate buffer and plasma proteins are *extracellular* natural buffers. Haemoglobin and other proteins are *intracellular* buffers. Bicarbonate/carbonic acid buffer is most important as carbonic acid levels are regulated by lungs which eliminates excess of it as CO_2. Bicarbonate part is separately controlled by kidney.

Acidosis is pH of blood less than 7.35. Alkalosis is pH more than 7.45.

Henderson Equation (Used to Assess Hydrogen Ion Concentration)

$$H^+ \text{ (nmol/L)} = \mathbf{K} \times \frac{H_2CO_3 \text{ mmol/L}}{HCO_3^- \text{ mmol/L}}$$

$$\text{OR} \quad \mathbf{K} \times \frac{\alpha \, PCO_2}{HCO_3^- \text{ mmol/L}}$$

Here constant **K** is 800 (for H_2CO_3 / HCO_3 buffer).

Carbonic acid (H_2CO_3) is solubility coefficient of CO_2 in blood (α) multiplied by partial pressure of CO_2 (PCO_2). α is 0.03 ml/mmHg/100 ml of blood; PCO_2 is 40 mmHg. $H_2CO_3 = \alpha PCO_2 = 0.03 \times 40 = 1.2$ ml. Normal blood bicarbonate/HCO_3 level is 24 mmol/L. so $\mathbf{H^+}$ is 800×1.2 divided by 24 = 40 mmol/L.

Henderson-Hasselbalch Equation (Used to Assess pH)

It is used to find out pH of the blood using logarithm. Negative logarithm of constant K (800 for carbonic buffer) is called as pKa. It is 6.1 for H_2CO_3 / HCO_3 buffer system.

$\mathbf{pH} = \text{pKa} + \log \frac{HCO_3^-}{H_2CO_3}$ means 6.1 + log 24 divided by 1.2 = 6.1 + log 20 = 6.1 + 1.3 = 7.4

METABOLIC ALKALOSIS

Primary base excess, i.e. HCO_3. *A standard bicarbonate above 27 mmol/litre.*

Causes

- Repeated vomiting due to any cause. Commonly seen in cases of pyloric stenosis. Here ***hypokalaemic alkalosis*** occurs which is an important aspect for managing the patient.
- Excess alkali ingestion, e.g. antacids.
- Cortisol excess either due to over administration or Cushing's syndrome.

Clinical Features

- Cheyne stokes breathing with period of apnoea of 5-30 seconds.
- ***Tetany*** due to alkalosis. More often *latent tetany* which is revealed by *Trousseau's sign.*

Investigations

Serum electrolytes, arterial blood gas analysis.

Treatment

- Normal saline or double strength normal saline IV infusion, with slow IV potassium chloride 40 mmol/litre in saline, slowly under ECG monitoring.
- pH more than 7.7 causes life-threatening alkalosis which requires rapid correction by infusing dilute hydrochloric acid or ammonium chloride, however, with care and monitoring.

RESPIRATORY ALKALOSIS

Arterial PCO_2 is below normal.

Causes

- Hyperventilation during anaesthesia, due to head injury/ severe pain.
- High altitude.
- Hyperpyrexia.
- Encephalitis, hypothalamic tumours, drugs like salicylates, due to cirrhosis of liver.
- Hysteria.

Features and Management

- Headache, tingling, circumoral anaesthesia, tightness in chest, tetany, arrhythmias are the features.
- Low $PaCO_2$, low HCO_3, high alkaline pH are typical. Serum HCO_3 will not fall below 15 mEq/L.
- It can be acute or chronic.
- It is managed by oxygen therapy, treating the cause, *acetazolamide* in high altitude.
- Respiratory suppression due to alkalosis is treated by CO_2.

METABOLIC ACIDOSIS

It is an excess acid or base deficit. **A standard bicarbonate below 21 mmol/litre.**

Causes

Increase in fixed acid:

- Diabetic ketoacidosis.
- Starvation.
- Hypoxia.
- Renal insufficiency.
- Cardiac arrest.
- Excessive exercise.
- Intestinal strangulation. Here anion gap is increased.

Less you eat, you are malnourished. More you eat, more you are diseased.

Loss of base:

- Diarrhoea.
- Ulcerative colitis.
- Gastrocolic fistula.
- Intestinal fistula.
- Ureterosigmoidostomy causes hyperchloraemic hypokalaemic acidosis. Here anion gap is normal.

Features

- Rapid, deep, noisy breathing (air-hunger)—*Kussmaul's* breathing.
- Cold clammy skin, tachycardia, right heart strain, altered level of consciousness.
- Cardiac arrhythmias, hypotension.
- Anorexia, muscle weakness, vomiting.
- pH below 7.2 is dangerous and life-threatening.
- Capillary stasis.
- Urine is strongly acidic.
- Low standard HCO_3 level.
- Base deficit.
- It is evaluated by doing arterial blood gas analysis (ABG) which shows low HCO_3, low pH; anion gap; urinary anion gap (UAG). Normal UAG is zero or positive. In metabolic acidosis due to GI cause, UAG becomes negative due to increased NH_4Cl excretion; if it is due to renal cause, UAG will be positive.

Treatment

- Correction of hypoxia.
- **50 mmol of 8.4% sodium bicarbonate infusion IV.**
  Sodium bicarbonate required in mEq/L = Body weight in kg × Base deficit × 0.3.
- Correction of electrolytes.
- Specific treatment for lactic acidosis (*type A* [shock / respiratory/CO/cyanide/anaemia]; *type B* [diabetes/hepatic/toxins/drugs])—only careful use of $NaHCO_3$ in severe cases, *dichloracetate* which stimulates pyruate dehydrogenase to reduce lactate.
- Specific therapies for diabetic ketoacidosis, alcoholic acidosis, aspirin poisoning, renal causes.

Astrup formula

Total base excess or deficit = Base excess/base deficit × body weight in kg × 0.3

RESPIRATORY ACIDOSIS

It is a feature of respiratory failure with high arterial PCO_2 causing fall in pH.

Causes

- During and after anaesthesia.
- Chronic bronchitis.
- Emphysema.
- Thoracic diseases.
- Upper abdominal surgeries and diseases.
- Respiratory airway obstruction.
- Myasthenia gravis.
- Poliomyelitis.
- Stroke, infection, obesity, hypoventilation.

Features and Treatment

- Features of hypercapnia like dyspnoea, confusion, psychosis, hallucinations, sleep disturbances, tremor, jerks, and personality changes. CNS manifestations are more severe in respiratory acidosis than in metabolic acidosis as lipid soluble CO_2 crosses blood-brain barrier easily than HCO_3.
- Acute respiratory acidosis is managed by oxygen therapy, ventilator support. Oxygen therapy should not be used in chronic hypercapnoea unless it is really indicated as hypoxia stimulated respiration may be suppressed. Alkali therapy also is not usually used unless acidosis is very severe (below 7.15) or there is severe bronchospasm.

ANION GAP

- It is calculated estimation of the undetermined or unmeasured anions in the blood.
- It is $(Na^+ + K^+) — (HCO_3^- + Cl^-)$.
- Normal anion gap is 10-16 mmol/litre.
- Anion gap is charge difference between unmeasured anion and cation. Important unmeasured anions are anionic protein, phosphate, sulphate, organic acids. Unmeasured cations are calcium and magnesium. Albumin is the main component of anion gap. When albumin decreases by one gram/dl then anion gap decreases by 2 mEq/L.

Increased anion gap is seen in	*Normal anion gap is seen in*
• Metabolic acidosis due to ketoacidosis	• Diarrhoea
• Lactic acidosis	• GIT fistulae
• Poisoning	• Hyperchloraemic acidosis
• Renal failure	

FLUID THERAPY

Osmolality of a solution is assessed by the amount of solute dissolved in a solvent like water measured in weight (kg).

Osmolarity of a solution is assessed by the amount of solute dissolved in a solvent like water measured in volume (litre).

Normal plasma Osmolality is 285 mOsm/kg (275-295).

Osmolality is calculated by two methods

a. *Osmolality of plasma* = $\frac{0.54}{1.86} \times 10^3$ mOsm/kg

It is based on the fact that solution of 1 mOsmol/kg freezes at – 1.86°C; whereas normal plasma freezes at – 0.54°C.

b. *Osmolality of plasma* = $2 \times (Na) + \frac{(\text{Glucose mg\%})}{18} + \frac{(\text{Blood urea mg\%})}{6}$

It is based on the concentrations of major solutes in plasma. So sodium concentration contributes mainly to the osmolality.

Colloidal osmotic pressure is difference in plasma osmotic pressure and interstitial fluid pressure which is normally 25 mm Hg, which is mainly by plasma albumin concentration. Plasma proteins do not go out of capillary wall into the interstitium.

Principles of Fluid Therapy

Indications

- For rapid restoration of fluid and electrolytes in dehydration due to vomiting, diarrhoea, shock due to haemorrhage or sepsis or burns.
- Total parenteral nutrition.
- Anaphylaxis, cardiac arrest, hypoxia.
- Post-gastrointestinal surgeries.
- For maintenance, replacement of loss or as a special fluid.

Advantage

Controlled, accurate and adjustable, rapid and predictable.

Problems in Fluid Therapy

- Needs hospitalisation; costly; needs asepsis.
- Fluid overload; pulmonary oedema and cardiac failure; infection.
- Thrombophlebitis; haematoma; cellulitis in local area.
- Pyrogenic reaction; air embolism; bacteraemia.
- Discomfort; poor patient acceptance.

Daily requirement of sodium is 100 mEq; potassium is 60 mEq; calcium is 5 mEq; magnesium 1 mEq.

One litre of normal isotonic saline contains 154 mEq of sodium.

Ringer's lactate is the most physiological fluid (*crystalloid*) containing sodium—130 mEq/lt; potassium—4 mEq/lt; chloride—109 mEq/lt; lactate (bicarbonate)—28 mEq/lt; and calcium—3 mEq/lt. It should be avoided in liver failure patients. As it does not contain glucose it can be used in diabetics.

Other *crystalloid* fluids—normal saline, dextrose saline, 5% dextrose, isolyte P, isolyte G, isolyte M.

Colloids are of large molecules which shift the fluid from interstitial compartment to intravascular compartment and are used as plasma expanders. Haemaccel, hetastarch, pentastarch, dextran 40/70 are colloids.

Special purpose fluids are sodium bicarbonate 7.5% and 8.4% used in metabolic acidosis, forced diuresis, hyperkalaemia; mannitol 10/20% used as an osmotic diuretic agent; hypertonic saline 1.6%, 3%, 5% and 7.5% used in hyponatraemia of different severity; albumin 4.5% as plasma expander; albumin 20% in severe hypoalbuminaemia.

Calculation of Drop Rate of IV Fluids

1 ml =16 drops in usual drip set. For microdrip set one ml = 60 drops.

a. Quantity of fluid required in liters per day × 10 = Drop rate/minute. 2.5 litres is usually used quantity of fluid/day. So 2.5 × 10 = 25 drops/minute.
b. Fluid volume in ml to be infused in one hour divided by four = Number of drops/minute. Example: 100 ml/hour means 25 drops/minute.
c. Number of microdrop/minute = Volume in ml/hour (50 microdrop/minute = 50 ml/hour)

Note:

- Weight loss more than 10% of individual's weight in 6 months is called as significant weight loss.
- Body Mass Index (BMI) is body weight in kilograms divided by height in meters squared. BMI less than 18.5 signifies nutritional impairment and below 15 signifies severe malnutrition.
- Daily fluid loss from kidneys is 1500 ml; from lungs is 400 ml; from skin is 800 ml; from stool is 60-150 ml.
- Energy requirement per day is 20-30 kcal/kg, i.e. around 2000 kcal/day total.
- Glucose requirement is 200 g per day; fat requirement is 200 g per week; nitrogen (protein) requirement is 0.15 g/kg per day. Nitrogen need increases to 0.25 g/kg/day in hyper catabolic status.
- Transit time is rapid in jejunum; three times slower in ileum; still slower in colon.
- Fluid absorption capacity is 40% in jejunum; 70% in ileum; 90% in colon.
- Electrolyte and vitamin B_{12} absorption and enterohepatic circulation occurs in ileum and so ileum is more important than jejunum.

NUTRITION

> *We (now) devote more attention to the patient's diet and habits, and more often send him away with good advice than with hastily-written prescriptions.*
>
> —**Robert Hall Babcock,** 1901

Principles of Nutrition

- Avoiding of malnutrition is the basic goal in nutrition therapy as malnutrition increases the morbidity and mortality of the disease process and prevents or delays the recovery. Malnutrition increases the chance of sepsis, prevents wound healing, increases the respiratory complications, and decreases the efficacy and tolerance to radiotherapy or chemotherapy.
- Whenever possible enteral route of nutrition should be used ideally. If that is not possible then parenteral nutrition is used.
- Overfeeding should be avoided as it leads into hyperglycaemia, hepatic steatosis, raised BUN, and excess CO_2 production.
- Timing and type of nutrition is also important.
- Nutrition therapy reduces protein wasting.
- Immunomodulators like glutamine, arginine and omega 3 fatty acids are also very useful. *Glutamine* is a nonessential amino acid synthesized in skeletal muscle. It is essential for cell proliferation during tissue repair. Glutamine helps GI mucosal cell proliferation, maintains mucosal integrity, improves immune function and prevents translocation of bacteria. It is useful in inflammatory bowel disease, short gut syndrome, burns, major trauma, and sepsis. Glutamine

is used commonly by enteral route even though IV preparations are now available (but it is very unstable in solutions).

Caloric requirement:

- Neonatal 100 kcal/kg/day.
- Adult 40 kcal/kg/day.
- Adult with catabolism 60 kcal/kg/day.

It is given as:

- Carbohydrates 50%.
- Fat 30-40%.
- Protein 10-15%.

Caloric values:

- Carbohydrate 4 kcal/g.
- Protein 4 kcal/g.
- Fat 9 kcal/g.

Indications for Nutritional Support

a. Preoperative nutritional depletion.
b. Postoperative complications: Sepsis, ileus, fistula.
c. Intestinal fistula: High type wherein output is more than 500 ml/day. It may be duodenal, biliary, pancreatic, intestinal.
d. Pancreatitis, malabsorption, ulcerative colitis, pyloric stenosis.
e. Anorexia nervosa and intractable vomiting.
f. Trauma—multiple fractures, fasciomaxillary injuries, head and neck injuries.
g. Burns.
h. Malignant disease.
i. Renal and liver failure.
j. Massive bowel resection causing short bowel syndrome.

Assessment

- Body weight
- Mid-arm circumference
- Triceps skin fold thickness
- Serum albumin
- Lymphocyte count

Nutritional requirements: Carbohydrates, fat, proteins, vitamins (includes fat soluble vitamins also), minerals, trace elements.

Methods of Feeding

Enteral:

a. *Gastrointestinal tract is the best route to provide nutrition.*
b. *Enteral feeding can be delivered by bolus, by gravity or using mechanical pump.*
   - *By mouth:* Requires:
      - Common sense,
      - Cleanliness,
      - Compassion.
   - *By nasogastric tube:* Confirmation of the tube in the stomach is made by injecting 5 ml of air down the tube and listening through a stethoscope for its bubbling entry into the stomach. Feeding rate is 30-50 ml/hours. 5 hours night time gap is given to allow gastric pH to return to normal.

Problems with tube feeding are:

- Blockage
- Nausea and vomiting, aspiration
- Hyperosmolarity
- Diarrhoea
- Tube discomfort
- Cholestasis

c. *By enterostomy:*
   - Gastrostomy.
   - Jejunostomy.

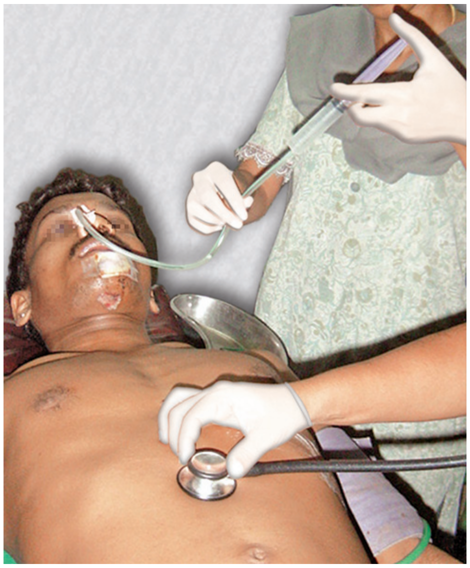

Fig. 1.210: Nasogastric tube passed should be confirmed in place using stethoscope. Tube is used for feeding purpose.

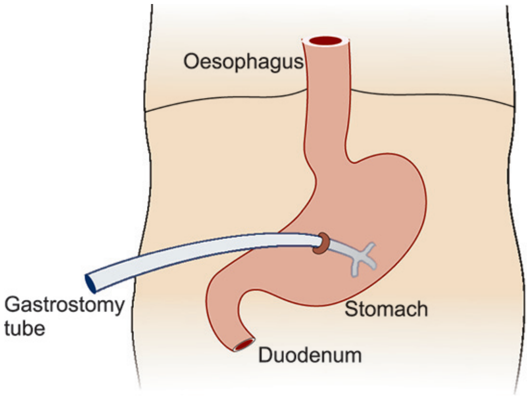

Fig. 1.211: Gastrostomy tube in place for enteral feeding.

Different preparations and formulas are available for enteral feeding. Soluble fibre containing diets along with nutrients are better to prevent diarrhoea.

Complications of enteral feeding

- Aspiration
- Wound infection and leak
- Diarrhoea due to rapid feeding or hyperosmolarity
- Hyperglycaemia
- Hypokalaemia
- Refeeding syndrome due to severe hypokalaemia and hypophosphataemia

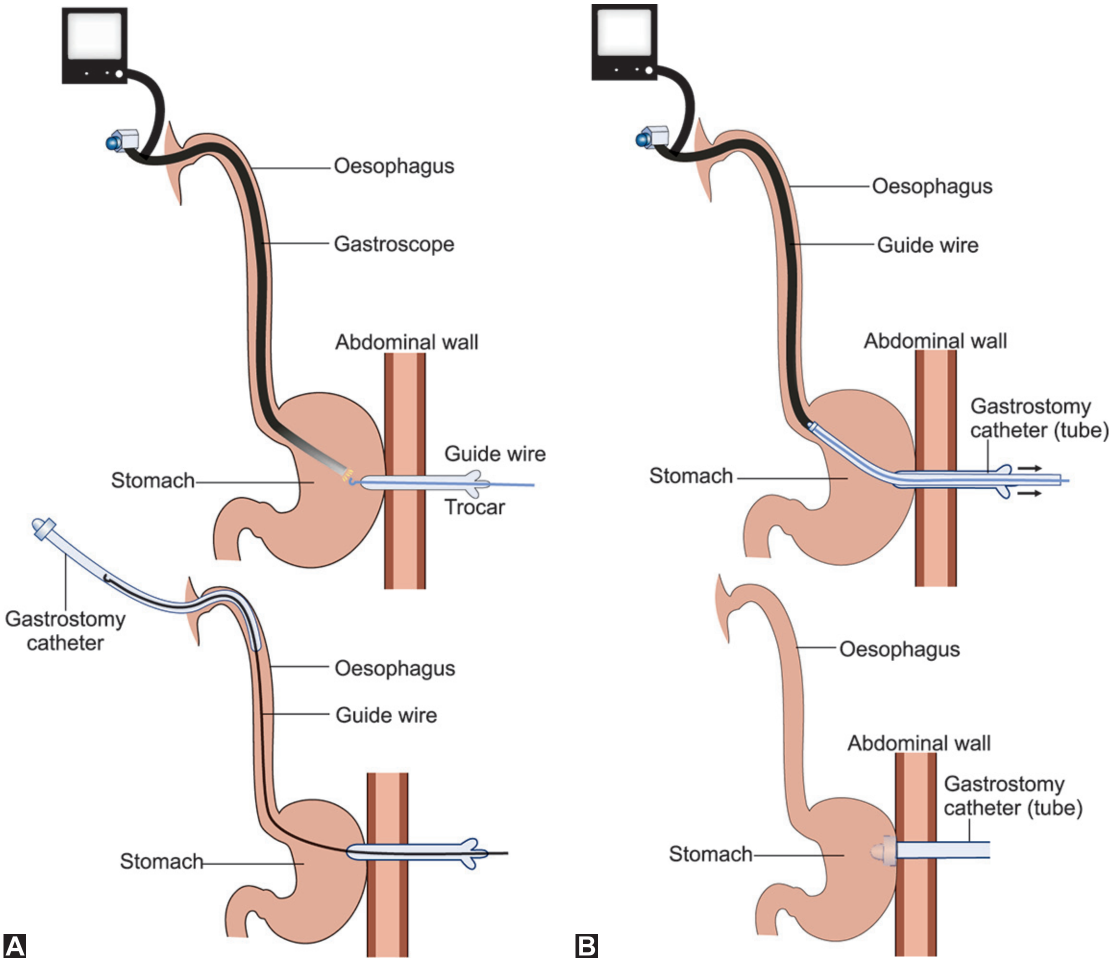
Oesophagus
Gastroscope
Abdominal wall
Guide wire
Stomach
Trocar
Gastrostomy catheter
Oesophagus
Guide wire
Stomach
A
Oesophagus
Guide wire
Abdominal wall
Gastrostomy catheter (tube)
Stomach
Oesophagus
Abdominal wall
Gastrostomy catheter (tube)
Stomach
B

Figs 1.212A and B: Percutaneous endoscopic gastrostomy (PEG). Guide wire passed through a trocar in the abdominal wall into the stomach is pulled out across the oesophagus and mouth under visualisation using gastroscope. Gastrostomy catheter (tube) is passed along the guide wire and pushed downwards across mouth and oesophagus to reach the stomach. It is further pushed along the abdominal wall through the trocar. Gastrostomy tube is pulled out through abdominal wall and fixed. Trocar is removed. Placement of tip of gastrostomy tube can be confirmed by doing gastroscope.

Advantages of Enteral Nutrition

- Enteral nutrition preserves mucosal protein, digestive enzymes, IgA secretion; prevents mucosal atrophy and bacterial translocation.
- It is more physiological as nutrients pass through liver, the first filter to process and store. Gallstone formation is prevented (unlike long-term TPN) by stimulating gallbladder motility.
- It has got less serious complications. It is cost-effective.
- It supplies glutamine and short chained fatty acids to gut.

Contraindications of Enteral Nutrition

- Intestinal obstruction, GI bleed, paralytic ileus, severe diarrhoea, high output fistula.
- Low cardiac output, haemodynamically unstable patient.
- If safe access to enteral feeding is not present.
- Anticipated complications if thought to be present should be avoided.

GASTROSTOMY

It is done if feeding is required for more than *one month.*

Indications

- Severe malnutrition
- Major surgeries
- Severe sepsis
- Trauma
- Head and neck surgeries

Types

Based on duration of use:

- Temporary.
- Permanent.

Based on lining:

- Mucus lined (permanent).
- Serosal lined (temporary).

Based on technique:

a. ***Stamm temporary gastrostomy:*** After opening the abdomen, anterior wall of the stomach is opened. Feeding tube (Malecot's catheter) is placed in position. Two layers of purse string sutures are put around the tube. Wound is closed.
b. ***Kader-Senn temporary gastrostomy***.
c. ***Percutaneous endoscopic gastrostomy*** (popular)—now becoming common method.
d. ***Janeway's mucus lined permanent gastrostomy*** by creating tunnel in stomach wall.

Problems in gastrostomy

- Leak—gastric fistula
- Infection
- Aspiration and pneumonia
- Diarrhoea is common—30%
- Bloating, abdominal cramps
- Displacement, blockage of the tube

Contraindications

- Previous gastric surgeries.
- Intestinal obstruction.
- Gastric outlet obstruction.

JEJUNOSTOMY

Jejunostomy for enteral nutrition is becoming more popular because of:

- Its comfort,
- Easy to do,
- Can be kept for long time,
- Lesser complication than gastrostomy.

Indications are same as gastrostomy.

Types

a. *Witzel jejunostomy:* Site of placing jejunostomy is 30 cm from duodenojejunal junction.
b. *Needle jejunostomy* using catheter of small gauge.

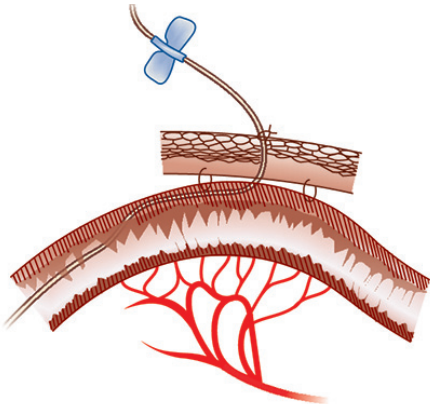

Fig. 1.213: Needle jejunostomy.

TOTAL PARENTERAL NUTRITION (TPN)

All nutritional requirements are given only through intravenous route, not through gastrointestinal tract. It can be through a central catheter through the subclavian/internal jugular vein where the tip of venous catheter is at distal part of superior vena cava.

It can also be a peripheral (Peripheral Parenteral Nutrition/PPN) through a peripherally inserted central venous catheter (PICC) or through a formal peripheral venous line.

Indications

- Failure or contraindication for any enteral nutrition for 7-10 days
- High output abdominal fistulas, duodenal, biliary, pancreatic fistulas
- Major abdominal surgeries of liver, pancreas, biliary, colonic
- Septicaemia
- Multiple trauma
- Short bowel syndrome
- Severe pancreatitis, bowel ischaemia, peritonitis, ileus
- Massive GI bleeding, unstable haemodynamically
- High-risk of aspiration
- Hyperemesis gravidarum
- Multiorgan failure, head injury, severe burns

About 5% of hospital admissions require TPN.

Technique

- Using a needle and guide wire a *Subclavian vein catheter* is passed just below the clavicle and fixed securely to the skin.
- TPN is given through *central vein* and not through a peripheral vein.
- Peripherally inserted central catheter (PICC) is also commonly used (PPN).

Goals, Factors and Assessment in TPN

- To decrease adverse effects of catabolism; to increase protein synthesis, to reduce protein breakdown, to prevent weight loss.
- To support ongoing metabolism.
- To improve immune function, cardiac and respiratory function.
- To maintain glycogen reserve in cardiac and respiratory muscles.
- To maintain acid, base and electrolyte metabolism.
- Age, premorbid state, muscle mass, weight, serum albumin should be assessed.
- Underlying disease, its severity, therapies for the disease, GI function should be assessed.
- *Fluid requirement* is assessed by—1500 ml for 20 kg weight + 20 ml/kg for additional weight.
- *Energy needed* is calculated by calculating resting energy expenditure (REE). (1) By *simple calculation*: REE in kcal/day = 25 × weight in kg. (2) *Harris Benedict equation*: REE in men = 66 + (13.7×weight in kg) + (5 × ht in cm) – (6.7 × age in years). In women = 655 + (9.6 × weight) + (1.8 × ht) – (4.7 × age). Activity/disease/thermal factors are also

added. (3) *Indirect calorimetry*: It is more accurate method done using special instrument. REE: = (3.9 × VO_2) + (1.1 × VCO_2) – 61.

Components used in TPN

- Carbohydrates
- Fat and amino acids
- Vitamins and trace elements, electrolytes, minerals

Components Used in TPN/PPN

Carbohydrates: Dextrose is less costly (1 gram dextrose 3.4 kcal); can be used in 50-70% concentration during PN. It supplies calories, stimulates insulin release and glucose oxidation, prevents muscle protein breakdown, has got nitrogen sparing ability. *Problems* of carbohydrate/dextrose are—low calorie value compared to fat, requires large fluid volume to infuse, hyperglycaemia, causes more CO_2 production, because of high osmolality it causes thrombophlebitis in 10% or above concentration. Rate of administration of dextrose is 5 mg/kg/min.

Fat: Fat gives high calorie (1 gram—9 kcal), essential fatty acids. It is given as emulsion containing long chain triglycerides. It contains soyabean/sunflower oil with egg yolk phospholipids (emulsifying factor), glycerin (isotonic). Fat has got low osmolality (260 mosm/L); it is available as 10%, 20%, 30% emulsions. *Advantages of fat* in PN are—high calorie, prevents hyperglycaemia, glucose and nitrogen sparing, less CO_2 production, less insulin production; it prevents essential fatty acid deficiency (for this purpose 3 days a week dose is given), reduces thrombophlebitis. *Problems of lipids in* PN are—hypertriglyceridaemia, sepsis, fat embolism, fat overload, hepatic dysfunction, pancreatitis, delayed gastric emptying. Lipid emulsions are avoided in hyperlipidaemia, anaemia, acidosis, obesity. Lipid emulsion is good culture media for bacteria and fungi; so care should be taken to prevent sepsis. Triglyceride level should be monitored weekly; if more than 400 mg%, infusion is discontinued. Mixture of long and medium chain fatty acids is better tolerated and efficient.

Amino acids: They are source of proteins. Calorie value of amino acid is 4 kcal/gram. 6.25 gram protein has 1 gram nitrogen. In PN 20% of energy comes from amino acids; rest from dextrose and fat. Daily protein need is 0.8-1.5 gram/kg. Protein supplement should be less in patients with CRF and hepatic encephalopathy. Its need is more in burns, trauma, enteropathy, sepsis. Protein supplement should not exceed 1.7 gram/kg/day; if so will cause raised urea production. *Uses of* amino acids in PN—in protein anabolism; prevents scatabolism. Proper monitoring by doing BUN or ammonia level is essential during amino acid therapy.

***Vitamins, electrolytes, trace elements and minerals*:** Electrolytes like sodium, potassium, magnesium, phosphate, calcium; fat soluble vitamins like A, D, E, K; water soluble vitamins; trace elements like chromium, copper, iodine, iron, manganese, selenium, zinc are all used in PN.

Monitoring the patient

- By body weight, fluid balance, blood glucose, electrolytes, blood urea, LFT, serum calcium, magnesium, phosphate should be done at regular intervals.
- A weight gain more than one kg/day signifies fluid overload.

Complications

Technical

a. Air embolism.
b. Pneumothorax.
c. Bleeding.
d. Catheter displacement, sepsis, blockage.
e. Infection, thrombosis.

Biochemical

a. Electrolyte imbalance: Hyponatraemia, hypokalaemia, hypophosphataemia.
b. Hyperosmolarity.
c. Hyperglycaemia—common.
d. Dehydration.
e. Altered immunological and reticuloendothelial function.
f. Azotaemia.

Others

a. Dermatitis.
b. Anaemia and increased capillary permeability.
c. Cholestatic jaundice: It is common.
d. Severe hepatic steatosis.
e. Metabolic acidosis.
f. Candida infection (candidiasis), staphylococcal infection (10-15%).

Contraindications

- Cardiac failure
- Blood dyscrasias
- Altered fat metabolism

Anabolic steroid durabolin 25 mg IM weekly is given to improve nitrogen balance.

Home parenteral nutrition

- It is becoming popular
- It is commonly used in western countries
- It is indicated in short bowel syndrome or any other conditions wherein enteral feeding is not possible but patient can be sent home with provision for home parenteral nutrition
- Patient himself uses the TPN fluids as advised at home. He will be with TPN catheter
- Patient should attend TPN clinic weekly for follow-up or immediately whenever complications arise
- Patient will be comfortable psychologically and often can attend his job also

The secret of happiness is to admire without desiring.

REFEEDING SYNDROME

- Refeeding syndrome is occurrence of severe fluid and electrolyte imbalance in severely malnourished individual while starting the proper feeding enteral or parenteral nutrition. It is more common in TPN.
- It causes hypomagnesaemia, hypocalcaemia and hypophosphataemia leading into myocardial dysfunction, respiratory changes, altered liver functions, altered level of consciousness, convulsions and often death.
- Gradual feeding and correction of magnesium, phosphate and calcium and other electrolytes is important.
- Condition is common in chronic starvation, severe anorexia and alcoholic patients.

OBESITY AND MORBID OBESITY

Obesity is weight more than 20% above the normal. Body mass index (BMI) is weight in kilogram divided by height in meters squared [wt in kg / (Ht in meters)^2].

Morbid obesity is a condition wherein BMI is more than 40 kg/m^2. It is often also defined as weight 100 lbs or more; or 100% over ideal body weight.

Weight more than double the expected weight to that age and height of the individual is also often called as morbid obesity. Body weight exceeding BMI 50 kg/m^2 is called as *superobesity*.

Causes of Obesity

- Familial, hyperinsulinism, hyperadrenocorticism, hyogonadism.
- Abnormal eating behaviour: Hormones which control eating are—ghrelin from stomach; insulin from pancreas; leptin from fat; PYY 3-36 from colon. Hypothalamus is the center in CNS which controls eating.

Nutritional status	***BMI (kg/m^2)***
• Underweight	• < 18.5
• Normal	• 18.5-24.9
• Overweight (Preobesity)	• 25.0-29.9
• Obesity	• > 30
– Class I	– 30.0-34.9
– Class II (Moderate)	– 35.0-39.9
– Class III (Severe / Morbid)	– 40.0
• Superobesity	• > 50
• Super superobesity	• > 60

Obesity is more common in young women. Child of a normal weight parents has 10% chance to become obese. If both parents are obese then child has got 80% chances to develop obesity.

Problems with Morbid Obesity

Obstructive sleep apnoea, degenerative joint disease, back pain, hypertension, GERD, gallstones, type II diabetes, hyperlipidaemia, arrhythmias, venous diseases, DVT, skin diseases, urine incontinence, infertility, ventral hernias, obesity hypoventilation syndrome, polycystic ovarian disease, hirsutism, gynaecomastia, steatohepatosis, malignancies—are common problems. All these conditions are called as ***comorbidities.***

Infertility is common in married morbid obese women due to improper ovulation, polycystic ovary disease. Urinary incontinence is common in morbid obese women (stress incontinence and detrusor instability). Pregnancy in morbid obese patient is considered as high-risk pregnancy. Gestational diabetes, hypertension, are common. It is better these women to undergo bariatric surgery prior to pregnancy which definitely reduces the complications during pregnancy. Such individual after bariatric surgery needs more iron and vitamin supplements.

Complications of obesity	***Obesity and surgery***	***Treatment for obesity***
• *General:* Difficulty in work, fatigue, depression, back pain, arthritis and gout • *Cardiovascular:* Hypertension, stroke, thrombophlebitis, pulmonary embolism • *Pulmonary:* Hypoventilation, poor respiratory effort • *GIT:* Hiatus hernia with reflux, changes in liver, pancreatitis, gallstones • *Endocrine:* Diabetes mellitus	• Hernia and gallstones are more common in obese individual • Burst abdomen, incisional hernia are more common in obesity • Delay in recovering from anaesthesia • Infertility is more common	• *General:* Dieting, exercise • Drugs: Orlistat, Sibutramine • Surgeries: – *Restrictive* ◆ Vertical banded gastroplasty ◆ Laparoscopic adjustable gastric banding (LAGB) ◆ Jaw wiring – *Malabsorptive* ◆ Biliopancreatic diversion (BPD) ◆ Biliopancreatic diversion with duodenal switch (BPD-DS) ◆ Jejunoileal bypass – *Combined* – Roux-en-Y gastric bypass (RYGB) open or laparoscopic

Evaluation of the patient—cardiac and respiratory assessment; lipid profile; blood glucose; renal and liver status assessment; anaesthesia risk assessment.

Medical Management

- General: Diet, life style changes, exercise.
- *Drugs*: (1) Orlistat is a selective inhibitor of gastric and pancreatic lipases that reduces the absorption of lipids from intestine. (2) Sibutramine is a noradrenaline and 5 HT reuptake inhibitor which act as appetite suppressor.

Surgical Treatment (Bariatric Surgery)

Bariatric surgery causes long-term durable weight loss. Patient's eating behaviour is reduced to slow ingestion of small boluses of food and or also reducing the absorptive surfaces.

Types of Bariatric Surgery

a. *Restrictive* wherein caloric intake is reduced. Purely gastric restrictive procedures are vertical banded gastroplasty (VBG) and laparoscopic adjustable gastric banding (LAGB).

b. *Malabsorptive* wherein absorption of calories and nutrients from food is reduced. Biliopancreatic diversion (BPD) and biliopancreatic diversion with duodenal switch (BPD-DS) are malabsorptive procedures.

c. *Combined* wherein both methods are used. Roux-en-Y gastric bypass (RYGB) is both restrictive and malabsorptive types. Gastric bypass reduces GI polypeptide ghrelin level secreted from fundus of stomach and duodenum. Ghrelin acts on specific receptor in CNS—hypothalamus to initiate appetite. This stimulation is reduced to decrease appetite. In restrictive only procedure ghrelin level raises and so appetite stimulation is not controlled.

Indications for Bariatric Surgery

BMI more than 40 kg/m^2 or BMI more than 35 kg/m^2 with comorbidity is indication for bariatric surgery. American Association of Bariatric Surgery (ASBS) published indications and approved surgeries.

Contraindications

Patients who are unfit for general anaesthesia (cardiac/renal/respiratory/hepatic causes) or who are unable to adjust postoperative life styles or psychiatric patients are contraindications for bariatric surgery.

Preoperative Preparations and Evaluation

- Complete cardiac, respiratory/renal/hepatic evaluation.
- Lipid profile and blood glucose assessment.
- Obstructive sleep apnoea in obese patient should be assessed using polysomnography and be treated.
- Risk assessment for DVT should be done.
- If GERD symptoms are present gastroscopy should be done.
- USG abdomen to identify gallstones should be done; if gallstones present it is of usual practice to do cholecystectomy along with bariatric procedure.
- Nutritional evaluation and dietician advice for preoperative and postoperative diet management.
- Psychological screening is needed to all patients to counsel their postoperative care and diet.
- Separate theatre table is needed for morbid obese patient. Equipments should be long and flexible. In laparoscopic surgery, special ports and instruments are needed.

DIFFERENT SURGERIES

Vertical Banded Gastroplasty (VBG); Mason (1982)

It is a purely restrictive type with creation of a calibrated stoma in the lesser curvature which is reinforced by an encircled mesh with a proximal gastric pouch. After laparotomy Ewald's stomach tube is passed per orally to place against lesser curve. A 2.5 cm circular window is created in the body of stomach near lesser curvature 8 cm below the *angle of His*. After this, four lines of linear vertical stapling are done from circular opening towards angle of His. This staple line ideally should be divided using another cutting linear stapler to reduce chances of dehiscence. This creates a 50 ml proximal gastric pouch. A 1.5 × 7 cm polypropylene mesh is placed around the lesser curve through circular opening and sutured to create a 5 cm collar stoma.

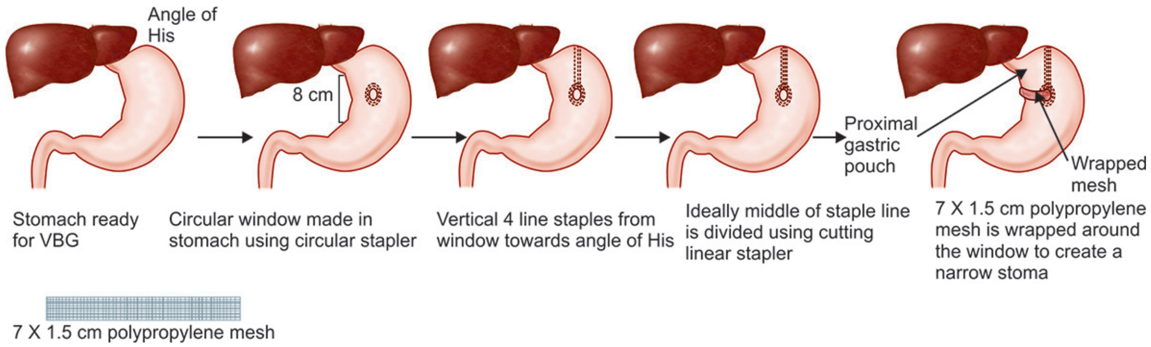

Stomach ready for VBG

Circular window made in stomach using circular stapler

Vertical 4 line staples from window towards angle of His

Ideally middle of staple line is divided using cutting linear stapler

7 X 1.5 cm polypropylene mesh is wrapped around the window to create a narrow stoma

Fig. 1.214: Vertical banded gastroplasty (VBG).

VBG causes only medium term weight loss; its efficacy is less compared to other procedures. *Complications of VBG are*—stricture at stoma (20%), vomiting (30%), reflux (20%), staple line dehiscence (40%), conversion into other procedures. Mortality is 0.3%.

VBG is technically easier to do; it has got very less chances of long-term metabolic and nutritional deficiencies.

Laparoscopic Adjustable Gastric Banding (LAGB)

It is also a restrictive type (1992, Guy Bernard) to create a narrow stoma just below the OG junction. It is used in adolescents and elderly. It is contraindicated in hiatal and paraoesophageal hernias. It is done using *laparoscopy.*

Under general anaesthesia, with patient in reverse Trendelenburg position six laparoscopic ports are placed. Using pars flaccida method, retrogastric tunnel is created; a silicone band is passed through the tunnel to encircle the cardia just below the OG junction; tail of the band is buckled and locked. Stoma diameter is determined by inserting a calibration tube. Stomach over the band is imbricated using interrupted sutures except the buckle area. Silicone tube end is brought out through the abdominal wall to connect access port. It is used for band volume adjustment by injecting or withdrawing the saline.

Gastrograffin study is done on 1st postoperative day to assess band position and lumen patency. Patient is advised to have liquid diet for one month. Band adjustment is done under fluoroscopic guidance in 2 months. Adjustment is done to achieve weight loss at a rate of 2 kg/week. Efficacy of LAGB is 55%.

Complications of LAGB are—spleen/stomach injury; bleeding; band slippage (10%); band erosion (7.5%); tube related complications; vomiting; pouch dilatation; reflux. Conversion rate is 3%; mortality is less than 0.5%.

Jejunoileal Bypass

It was the first malabsorptive procedure done for obesity. Now this technique is not done due to high incidences of complications.

Proximal jejunum is divided and proximal cut end is anastomosed to distal ileum just proximal to ileocaecal valve to reduce the absorptive surface area of small bowel.

Complication rates were higher and so procedure is not practiced. Colon in these patients absorbs high level of oxalate causing nephrocalcinosis. Bypassed bowel promotes bacterial growth causing endotoxic induced liver injury, cirrhosis, liver failure. Complications like protein, vtamin K, vtamin B_{12} deficiencies; gallstone formation; enteritis and diarrhoea; arthritis, osteoporosis are common.

Roux-en-Y Gastric Bypass

It is commonly done combined procedure. It can be done by open or laparoscopic method. Proximal stomach is dissected between 1st and 2nd branches of left gastric branches. Vagii nerves and *nerves of Latarjet* are retained carefully. Stomach is transected at this proximal site to create a proximal gastric pouch (15 ml if BMI is > 50; 30 ml if BMI is 40-50). It is usually carried out through linear stapler. Jejunum is transected 45 cm from *ligament of Treitz*. A side to side jejunojejunal anastomosis is done using stapler 75 cm distal to the distal cut end. Proximal Roux part of the distal jejunal cut part (75-150 cm, based on patient's preoperative weight) is brought out through the transverse mesocolon towards the created proximal gastric pouch and gastrojejunostomy is done to this proximal gastric pouch. Mesenteric defect is closed. Stomal integrity is checked on table by air distension and methylene blue infusion. Gastrograffin study is done in 24 hours to assess pouch size, stomal patency and distal obstruction. Oral food is started in 24 hours and patient is discharged in 4 days.

RYGB is more useful in weight loss compared to purely restrictive types. 5 years weight loss is 60-75%. It also prevents progression of noninsulin dependent diabetes mellitus, controls hypertension, sleep apnoea, hyperlipidaemia, asthma, arthritis, GERD.

Complications are—Roux obstruction, anastomotic leak, acute distal gastric dilatation, stomal stenosis, marginal ulcer, dumping syndrome, internal hernias, vitamin B_{12} deficiency, iron

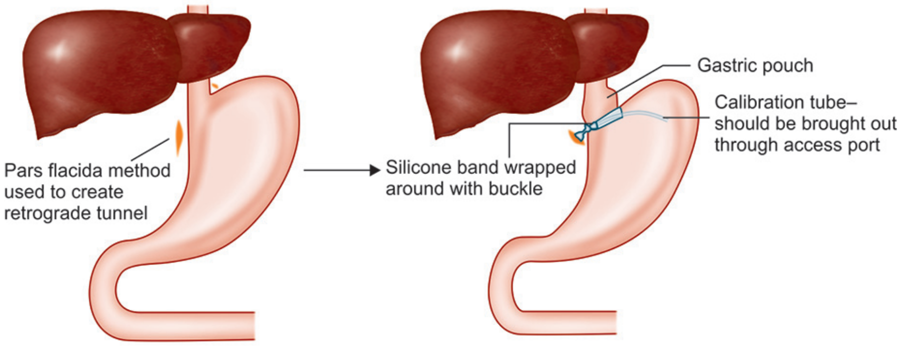

Fig. 1.215: Laparoscopic adjustable gastric banding (LAGB).

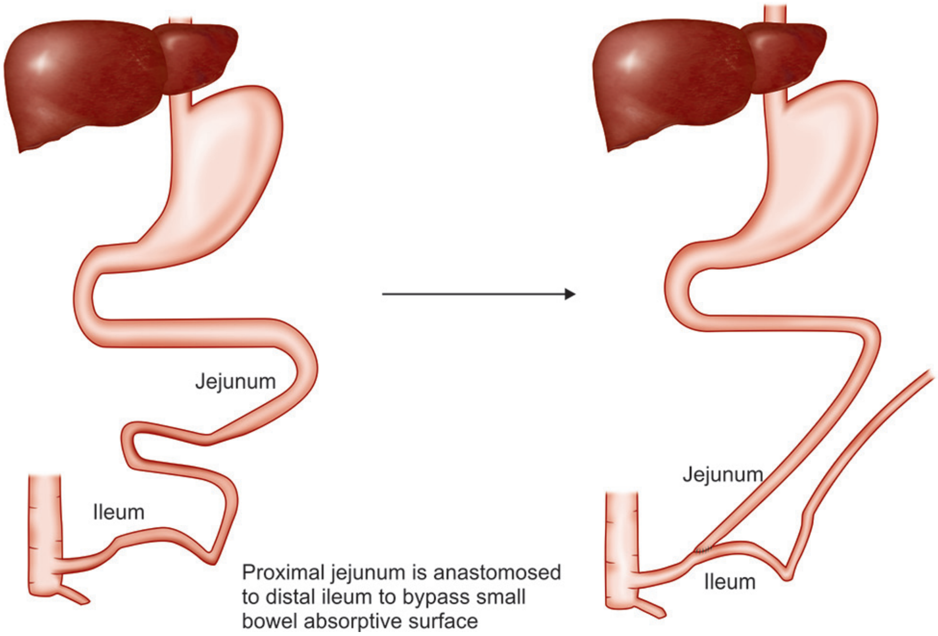
Jejunum
Ileum
Proximal jejunum is anastomosed to distal ileum to bypass small bowel absorptive surface
Jejunum
Ileum

Fig. 1.216: Jejunoileal bypass procedure (JIB).

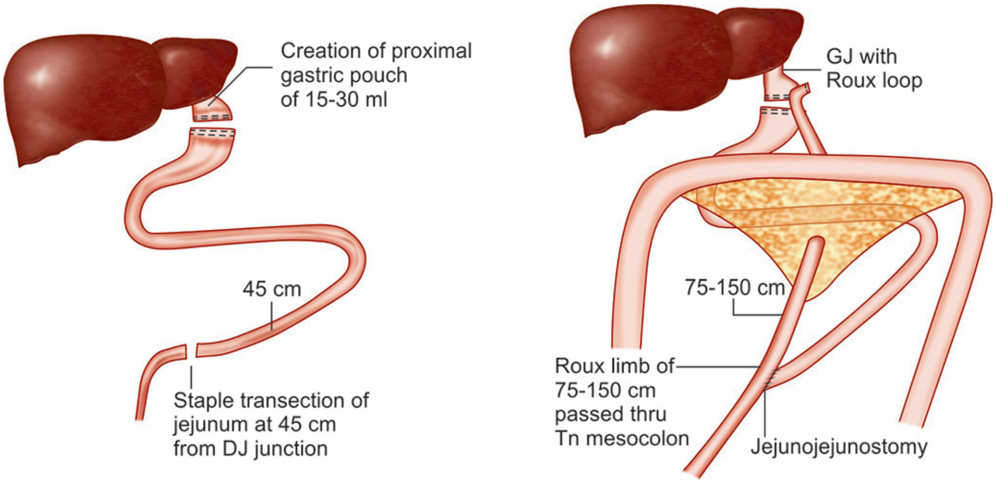
Creation of proximal gastric pouch of 15-30 ml
45 cm
Staple transection of jejunum at 45 cm from DJ junction
GJ with Roux loop
75-150 cm
Roux limb of 75-150 cm passed thru Tn mesocolon
Jejunojejunostomy

Fig. 1.217: Roux-en-Y gastric bypass procedure (RYBG).

deficiency anaemia. Distal gastric dilatation needs emergency intervention which is usually due to jejunojejunal obstruction.

Laparoscopic RYGB (1994, Wittgrove, Clark, Trembly)

Technique is similar to open RYGB. Anastomoses are done using endoscopic stapler. GJ between gastric pouch and Roux jejunum is done either using linear stapler through laparoscopic port after making a gastrotomy in the pouch which is later sutured after staple firing; or using circular stapler anastomosis is done wherein anvil is initially passed transorally often under endoscopic guidance across the pouch into the Roux jejunum; or using hand sewing with absorbable sutures. Omentum is released from the colon and is covered over the GJ. Mesenteric defect and *Patterson Brown defect* are closed. A *Bronlin antiobstructive stitch* is placed between Roux and biliopancreatic limbs. Integrity of anastomosis is checked using insufflation of air, methylene blue. *Complications ar*e similar to open RYBG. Conversion rate is 9%. *Advantages* of LRYBG to open RYBG are—faster recovery, less postoperative pain, less wound related complications, less morbid one. *Disadvantage* of LRYBG is availability of facility, technical expertise, and steep learning curve.

Now technique is modified to antegastric, antecolic one which has become popular (Gagner).

The lazy man is always occupied with his laziness.

Biliopancreatic Diversion (BPD) (Nicola Scopinaro, Italy)

It is done in patients who had failed restrictive procedure or who are superobese.

Distal subtotal gastrectomy is done with formation of proximal gastric pouch (of 400 ml in BMI 40-50; 200 ml in BMI > 50). Ileum is transected 250 cm proximal to ileocaecal valve; distal ileal segment is brought up to anastomose into the proximal gastric pouch. Proximal biliopancreatic jejunoileal limb is anastomosed into distal ileal segment 50 cm proximal to ileocaecal valve as end to side stoma. Additionally cholecystectomy should be done.

Modification of BPD with duodenal switch (BPD-DS) has become more popular. Here sleeve gastrectomy along the greater curvature is done to create gastric reservoir (200 ml) along lesser curve. Duodenum just distal to first part is divided using stapler; proximal cut end is sutured to proximal upward pulled end of the distal ileal segment of earlier transected ileum, 250 cm from ileocaecal valve. Biliopancreaticoduodenal with proximal jejunoileal segment is later stapled to distal ileum 50 cm proximal to ileocaecal valve. Duodenal switch reduces the rate of marginal ulcer and dumping syndrome.

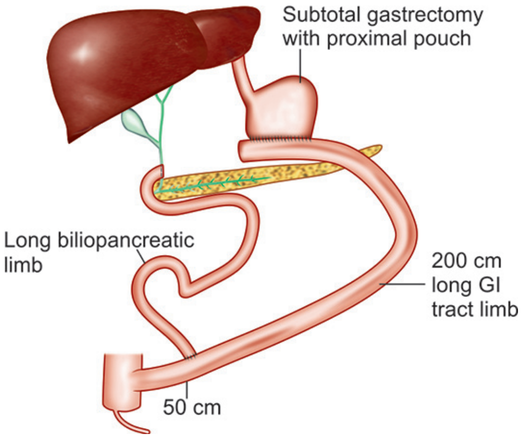
Subtotal gastrectomy with proximal pouch
Long biliopancreatic limb
200 cm long GI tract limb
50 cm

Fig. 1.218: Biliopancreatic diversion procedure (BPD).

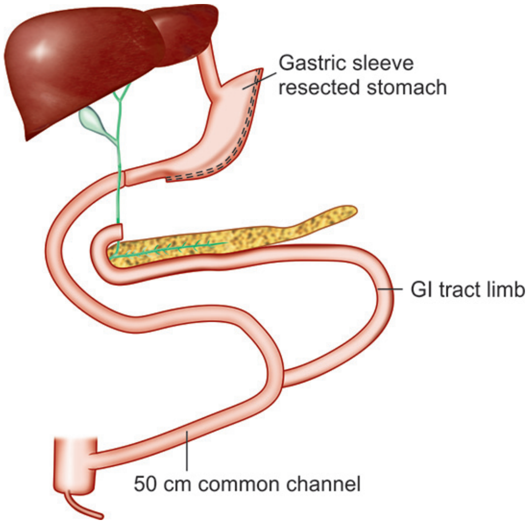

Fig. 1.219: Biliopancreatic diversion with duodenal switch procedure (BPD-DS).

Results of BPD/BPD-DS are—excellent for weight reduction compared to restrictive procedures. But they need lifelong supplement of vitamins, fat soluble vitamins, calcium, and iron. Technically BPD is easier to do when compared to BPD-DS.

Complications of BPD/BPD-DS are—anemia (30%); protein deficiency (20%); dumping syndrome; marginal ulcer (in BPD 10%; in BPD-DS it is 1%); osteoporosis; night blindness; biliopancreatic limb obstruction; staple line leak; staple line bleed; DVT; subphrenic abscess. Vitamin B_{12} deficiency is specific.

Ileal Interposition with Sleeve Gastrectomy

It is done *mainly in type II diabetes*. Often they are associated with obesity, dyslipidaemia, hypertension, nephropathy and neuropathy. Two types of ileal interposition with sleeve gastrectomy are done.

Type 1: Sleeve gastrectomy is done. 170 cm of ileum with mesentery is isolated 30 cm from ileocaecal junction. Jejunum is transected 50 cm from duodenojejunal flexure. Isolated 170 cm ileal segment is interposed 50 cm distal to DJ junction with end to end anastomosis on both ends.

Type 2: After doing sleeve gastrectomy, gastroduodenal junction is transected; cut proximal end of duodenum is closed. 170 cm ileal segment with mesentry is isolated and interposed between cut end of stomach and side of jejunum 50 cm distal to DJ flexure.

Cholecystolithiasis After Bariatric Surgery

Gallstone formation is common after bariatric surgery (50%). It is due to rapid weight loss. If gallstones are present at the time of bariatric surgery it is essential to do laparoscopic cholecystectomy during bariatric procedure. It is controversial about doing prophylactic cholecystectomy during bariatric surgery even though it is practiced in many places. Advantages are—it prevents future gallstone formation at the time of bariatric surgery (loss of access); it prevents difficulties in approaching CBD in case needed endoscopically due to surgery. Disadvantage is cholecystectomy adds additional 1 hour time for bariatric surgery increasing the risk of immediate complications. Patients who are not having gallstones at the time of bariatric surgery, should receive ursodeoxycholic acid 300 mg twice daily for 6-12 months of post-bariatric surgery period. Often stomach is anchored to abdominal wall as access part to biliary system for future need.

Note:

Many of these patients after bariatric surgery require plastic surgery for abdominal contour (panniculectomy, abdominoplasty) after weight reduction due to bariatric surgery.

G. Shock

> *The first noteworthy characteristic of the blood in shock is a high capillary red count…. When haemorrhage as a complicating factor tending to reduce the blood count is considered, these high counts are striking. They indicate that in shock a concentration of the blood occurs, at least in the superficial capillaries.*
>
> **—Walter Bradford Cannon, J Fraser, AN Hooper,** 1918

CHAPTER OUTLINE

- Shock
- Stages of Shock
- Effects of Shock
- Clinical Features of Shock
- Investigations in Shock
- Central Venous Pressure
- Pulmonary Capillary Wedge Pressure
- Systemic Inflammatory Response Syndrome
- Multiple Organ Dysfunction Syndrome
- Oxygen Therapy
- Topical O_2-Therapy
- Cardiac Arrest

SHOCK

Shock is a state of poor perfusion with impaired cellular metabolism manifesting with severe pathophysiological abnormalities. It is due to circulatory collapse and tissue hypoxia. Normal *aerobic metabolism* is not maintained due to hypoperfusion. Shock is meant by '*inadequate perfusion*' to maintain normal organ function.

At *cellular level* hypoxia causes change of normal aerobic to anaerobic metabolism causing lactic acidosis. Intracellular potassium is released into circulation. Lysosomes from cells get released into blood causing cell lysis. *Hypoxia and acidosis* through complements release free oxygen radicals and cytokines which damage capillary endothelium. Eventually cardiovascular, respiratory, renal, endocrine and GIT will be affected presenting as *systemic* features.

Causes of shock

1. **Hypovolaemic shock**—due to reduction in total blood volume.
   It may be due to:
   a. *Haemorrhage*
      – External from wounds, open fractures
      – Internal from injury to spleen, liver, mesentery or pelvis
   b. *Severe burns*, which results in loss of plasma
   c. Peritonitis, intestinal obstruction
   d. Vomiting and diarrhoea of any cause
2. **Cardiac causes**
   a. Acute myocardial infarction, acute carditis
   b. Acute pulmonary embolism wherein embolus blocks the pulmonary artery at bifurcation or one of the major branches
   c. Drug induced
   d. Toxaemia of any causes
   e. Cardiac surgical conditions like valvular diseases, congenital heart diseases
   f. Cardiac compression causes
      i. Cardiac tamponade due to collection of blood, pus, fluid in the pericardial space which prevents the heart to expand leading to shock.
      ii. Trauma to heart
3. **Septic shock**—is due to bacterial infections which release toxins leading to shock
4. **Neurogenic shock**—due to sudden anxious or painful stimuli causing severe splanchnic vessel vasodilatation. Here, patient either goes for cardiac arrest and dies or recovers fully spontaneously—spinal cord injury/anaesthesia can cause neurogenic shock
5. **Anaphylactic shock**—is due to Type 1 hypersensitivity reaction
6. **Respiratory causes**
   a. Atelectasis (collapse) of lung
   b. Thoracic injuries
   c. Tension pneumothorax
   d. Anaesthetic complications
7. **Other causes**
   a. Acute adrenal insufficiency (Addison's disease)
   b. Myxoedema

Shock may be *hypovolaemic, cardiogenic, obstructive, distributive or of endocrine* variety.

Pathophysiology of Shock

Any cause of shock

↓

Low cardiac output

↓

Vasoconstriction occurs as a compensation to perfuse vital organs like brain, heart, muscle, kidneys, liver

↓

Because of vasoconstriction and tachycardia

↓

Dynamic circulation increases

↓

Tachypnoea occurs to increase the oxygen saturation

↓

Peripheral veins (capacitance vessels) constrict diverting blood from splanchnic system towards essential vital organs

↓

Decreased renal blood flow reduces the GFR and thereby the urine output

↓

Renin angiotensin mechanism gets activated causing further vasoconstriction and aldosterone release

↓

Causes salt and water retention

↓

ADH is released

↓

Further concentration of urine occurs

When shock persists cardiac output falls further

↓

Hypotension and tachycardia occurs leading to poor perfusion of coronaries

↓

Hypoxia—***metabolic acidosis***

↓

Release of cardiac depressants

↓

Cardiac (pump) failure

Hypoxia

↓

Anaerobic metabolism

↓

Lactic acidosis

↓

Cell wall damage

↓

Sodium and calcium enter the cell

↓

Potassium leaks out of the cell

↓

Causes hyperkalaemia, *hyponatraemia and hypocalcaemia*

↓

Intracellular lysosomes break down releasing powerful enzymes which destroy own cell

↓

SICK CELL SYNDROME

Platelets are activated forming small clots in many places

↓

Disseminated intravascular coagulation (DIC) (Consumption coagulopathy)

↓

Further bleeding.

STAGES OF SHOCK

Factors like infection, trauma, burns, haemorrhage, hypovolaemia

↓

Hypoxia and its effects.

↓

SIRS (*Systemic inflammatory response syndrome*) is due to vasodilatation, increased endothelial permeability, thrombosis, leukocyte migration and activation.

↓

All these lead to altered cytokines level, abnormal **NO (nitric oxide)** synthesis, abnormal arachidonic acid metabolism, neutrophil activation, free radical production, altered complement activation, failure to have a localisation of inflammation. It is severe type of ***reversible shock.***

↓

Which will lead to established microvascular occlusion, cellular dysfunction, sick cell syndrome, DIC and PUMP failure.

↓

MODS (*Multiorgan dysfunction syndrome)* ***(Irreversible shock*)**—of lungs, kidneys, liver, clotting system and brain.

Note:

- *Distributive shock* is one in which there is vasodilatation, decreased vascular resistance, hypotension, altered microvascular perfusion with arteriovenous shunting, altered cellular oxygen metabolism. It is seen in septic shock, spinal trauma, adrenal crisis and anaphylaxis.
- *Obstructive shock* occurs due to mechanical impediment of circulation due to pulmonary embolism, tension pneumothorax or cardiac *tamponade*.

EFFECTS OF SHOCK

Heart: Low perfusion → low venous return → decreased cardiac output → hypotension → tachycardia. Persistent shock causes hypoxia and release of myocardial depressants leading to further cardiac damage.

Lung: Interstitial oedema → decreased gaseous exchange → pulmonary arteriovenous shunting → tachypnoea → *Adult/Acute respiratory distress syndrome (ARDS)* and pulmonary oedema.

Metabolic: Shock leads to hypoxia, which activates anaerobic metabolism leading to lactic acidosis. Antidiuretic hormone (ADH) is released which increases the reabsorption of water from renal tubules. Other hormones released are ACTH, prostaglandins, histamine, bradykinin, and serotonin to compensate the effects of shock to increase the perfusion of vital organs like heart, brain and lungs.

Cellular changes occur in persistent shock due to release of lysosomal enzymes, which alters the cell membrane permeability causing cell death—*sick cell syndrome*.

Sympathetic overactivity alters the microcirculation leading to capillary dysfunction.

Brain perfusion, when decreases the patient becomes drowsy. *Brain is the last organ to get underperfused in shock.*

Kidneys: GFR decreases and tubular reabsorption of salt and water increases for compensatory response. But in severe cases tubular necrosis sets in leading into irreversible damage.

Blood*:* Alteration in cellular components including platelets leads to *Disseminated intravascular coagulation (DIC)*. It causes bleeding from all organs.

Gastrointestinal tract: Mucosal ischaemia develops causing bleeding from GIT with haematemesis and malaena. It is aggravated by DIC. Hepatic ischaemia leads into increased enzyme levels.

Types of Hypovolaemia

a. *Covert compensated hypovolaemia:* When blood volume is reduced by 10-15% , there will not be significant change in heart rate, cardiac output and splanchnic blood compensates for the same.

b. *Overt compensated hypovolaemia*: Here patient has cold periphery, tachycardia, a wide arterial pressure, tachypnoea, confusion, hyponatraemia, metabolic acidosis, but systolic pressure is well-maintained.

c. *Decompensated hypovolaemia:* Here all features of hypovolaemia are present like hypotension, tachycardia, sweating, tachypnoea, oliguria, drowsiness, eventually features of SIRS is seen and often if not treated on time leads to MODS, i.e. irreversible shock.

Types of Shock

1. *Vasovagal Shock*

It is sudden dilatation of peripheral and splanchnic vessels causing reduced cardiac output and shock. Often it may be life-threatening due to hypoxia.

2. *Neurogenic Shock*

- It is usually due to spinal cord injury, which causes dilatation of splanchnic vessels.
  - This type can safely be treated with vasoconstrictor drugs to bring up the blood pressure. There will be bradycardia, hypotension, arrhythmias, and decreased cardiac output. Blood pressure control, oxygen delivery, maintenance of haemodynamics, airway, fluid therapy, intravenous methylprednisolone therapy should be done. Dopamine and or phenylephrine (α agonist) can be used.

3. *Hypovolaemic Shock—Most Common Type*

- Haemorrhage, may be due to injury to the liver, spleen, bone fractures, haemothorax, vascular injury, severe bleeding on table during surgeries of thyroid, liver, portal vein or major vessels.
- Vomiting, diarrhoea due to any cause.
- Burns.

4. *Cardiogenic Shock*

Cardiogenic shock is defined as circulatory failure causing diminished forward flow leading into tissue hypoxia in the setting of adequate intravascular volume with systolic blood pressure < 90 mmHg for 30 minutes; cardiac index < 2.2 L/minute / sq meter; raised PCWP (pulmonary capillary wedge pressure) > 15 mmHg. It is commonly seen in acute MI with a mortality > 50%. Cardiogenic shock develops within 24 hours of MI. it occurs when 50% of left ventricular wall is damaged by infarction. It leads to pulmonary oedema and severe hypoxia. Ischemic necrosis of left ventricular wall causes failure of pump thereby decreasing stroke volume.

Diagnosis is established by *ECG*, echocardiography, arterial blood gas analysis, *cardiac enzymes, PCWP* and electrolyte estimation (hypokalaemia and hypomagnesaemia are common) are the essential investigations.

Management

- Proper oxygenation with intubation, ventilator support, cardioversion, pacing, antiarrhythmic drugs, correction of electrolytes, avoiding fluid overload, prevention of pulmonary oedema as immediate measures.
- Dobutamine (β1 receptor agonist) is used to raise cardiac output provided there is adequate preload and intravascular volume (it is peripheral vasodilator and reduces BP). Dopamine is preferred in patients with hypotension. But it may increase peripheral resistance and heart rate worsening cardiac ischaemia. Often both dopamine and dobutamine combination may be required.
- Careful judicial use of epinephrine, norepinephrine, phosphodiesterase inhibitors (amrinone, milrinone) are often needed. Anticoagulants and aspirin are given. Thrombolytics can be used. β blockers, nitrates (nitroglycerine causes coronary arterial dilatation), ACE inhibitors are also used.
- *Intra-aortic balloon pump (IABP)* may need to be introduced transfemorally as a mechanical circulatory support to raise cardiac output and coronary blood flow.
- Relief of pain, preserving of remaining myocardium and its function, maintaining adequate preload, oxygenation, minimizing sympathetic stimulation, correction of electrolytes should be the priorities.

Tears shed for self are tears of weakness, but tears shed for others is a sign of strength.

- ♦ Percutaneous transluminal coronary angioplasty *(PTCA)* and coronary artery bypass graft *(CABG)* are the final choices.

5. *Cardiac Compression Shock*

- ♦ It is probably due to pericardial tamponade of any cause or kinking of great vessels, massive pulmonary embolism, tension pneumothorax, air embolism causes obstructive shock with reduced preload to heart.
- ♦ Acute massive pulmonary embolism from a thrombus or an air embolism (***50 ml of air***), obstructing more than 50% of pulmonary vasculature leads to severe shock and sudden death.
- ♦ Tachycardia, hypotension, pulmonary oedema, raised JVP, gallop rhythm are the features.

6. *Septic Shock*

Septic shock may be due to gram-positive organisms, gram negative organisms, fungi, viruses or protozoal origin.

Gram-negative septicaemia/gram-negative septic shock is called as *endotoxic shock*. It occurs due to gram-negative bacterial infections, commonly seen in strangulated intestines, peritonitis, gastrointestinal fistulas, biliary and urinary infections, pancreatitis, major surgical wounds, diabetic wounds and crush injuries.

Gram-positive septic shock	*Gram-negative septic shock*
• Due to exotoxin by gram +ve bacteraemia like *Clostridium tetani/welchii*, staphylococci, streptococci pneumococci • Fluid loss, hypotension is common; with normal cardiac output	Gram negative bacteria cause endotoxaemia and its effects. Urinary/gastrointestinal/ biliary and respiratory foci are common

Pathophysiology of septic shock

Toxins/endotoxins from organisms like *E. Coli*, *Klebsiella*, *Pseudomonas*, and *Proteus*

↓

Inflammation, cellular activation of macrophages, neutrophils, monocytes

↓

Release of cytokines, free radicals

↓

Chemotaxis of cells, endothelial injury, altered coagulation cascade—SIRS

↓

Reversible hyperdynamic warm stage of septic shock with fever, tachycardia, tachypnoea

↓

Severe circulatory failure with MODS (failure of lungs, kidneys, liver, heart) with DIC

↓

Hypodynamic, irreversible cold stage of septic shock.

Septic shock is typically a *vasodilatory shock* wherein there is peripheral vasodilatation causing hypotension which is resistant to vasopressors. This is due to toxin induced release of isoform of nitric oxide synthetase from the vessel wall which causes sustained prolonged release of high levels of nitric oxide. Magnitude of infection is quantified as (1) *Sepsis* which shows fever, tachycardia, leukocytosis. (2) *Severe sepsis* which shows low tissue perfusion with organ dysfunction (lactic acidosis, dysfunction of liver, kidney, lungs). (3) *Septic shock* with systemic hypotension (BP < 90 mm Hg in spite adequate fluid therapy), severe organ dysfunction (acute lung, kidney, liver injury), maldistribution of blood flow, shunting in microcirculation.

Stages of septic shock

a. *Hyperdynamic (warm) shock*: This stage is *reversible stage*. Patient is still having inflammatory response and so presents with fever, tachycardia, and tachypnoea. Pyrogenic response is still intact. Patient should be treated properly at this stage. Based on blood culture, urine culture (depending on the focus of infection), higher antibiotics like third generation cephalosporins, aminoglycosides, metronidazole are started. The underlying cause is treated like draining the pus, laparotomy for peritonitis, etc. Ventilatory support with ICU monitoring may prevent the patient going for the next *cold stage of sepsis*.

b. *Hypodynamic hypovolaemic septic shock (cold septic shock)*: Here pyrogenic response is lost. Patient is in decompensated shock. It is an irreversible stage along with MODS (Multi-organ dysfunction syndrome) with anuria, respiratory failure (cyanosis), jaundice (liver failure), cardiac depression, pulmonary oedema, hypoxia, drowsiness, eventually coma and death occurs (*Irreversible stage)*.

Treatment of septic shock

- ♦ Correction of fluid and electrolyte by crystalloids, blood transfusion. *Perfusion* is very/most important.
- ♦ Appropriate antibiotics—third generation cephalosporins/ aminoglycosides.
- ♦ Treat the cause or focus—drainage of an abscess; laparotomy for peritonitis; resection of gangrenous bowel; wound excision.
- ♦ Pus/urine/discharge/bile/blood culture and sensitivity for antibiotics.
- ♦ Critical care, oxygen, ventilator support, dobutamine/ dopamine/noradrenaline to maintain blood pressure and urine output.
- ♦ Activated C protein prevents the release of inflammatory mediators and blocks the effects of these mediators on cellular function.
- ♦ Monitoring the patient by pulse oximetry, cardiac status, urine output, arterial blood gas analysis.
- ♦ *Short-term (one or two doses) high dose steroid therapy* to control and protect cells from effects of endotoxaemia. It improves cardiac, renal and lung functions. Single dose of methylprednisolone or dexamethasone which often may be repeated again after 4 hours is said to be effective in endotoxic shock.

Septic shock

- ❖ Common causes are biliary, urinary, GIT sepsis (peritonitis, strangulation), respiratory (pneumonia)
- ❖ Common bacteria are *E.coli, Klebsiella, Pseudomonas*
- ❖ Common pathophysiologies are release of toxins, neutrophil activation, cytokine release, and sick cell syndrome, SIRS, MODS
- ❖ Clinical stages are hyperdynamic amd hypodynamic
- ❖ Find out the source of the infection by U/S, CT scan
- ❖ Do pus/blood/urine culture
- ❖ Start antibiotics of high generations like ceftazidime, amikacin, cefoperazone
- ❖ Dopamine/dobutamine infusion (slow)
- ❖ Monitoring by pulse, BP, respiration, urine output, level of consciousness
- ❖ Ventilator support, ICU management
- ❖ Treat the causes like peritonitis, abscess

7. Anaphylactic Shock

Injections—penicillins, anaesthetics, stings, venom, shellfish may be having antigens which will combine with IgE of mast cells and basophils, releasing histamine and large amount of SRS-A (Slow releasing substance of anaphylaxis). They cause bronchospasm, laryngeal oedema, respiratory distress, hypotension and shock. Mortality is 10%.

Rashes all over the body are commonly observed.

Anaphylactic shock

- ❖ Sudden onset
- ❖ Distributive shock
- ❖ Bronchospasm, laryngeal oedema
- ❖ Generalised rashes and oedema
- ❖ Hypotension, feeble pulse
- ❖ Mortality 10%
- ❖ To start adrenaline 100 ug IV, steroids, IV fluids, oxygen with foot end elevation
- ❖ Ventilator in severe cases
- ❖ Cardiac massage, defibrillation

CLINICAL FEATURES OF SHOCK

- ♦ In early stage—tachycardia, sweating, cold periphery, hypotension, restlessness, air hunger, tachypnoea, oliguria, collapsed veins.
- ♦ In late stage—cyanosis, anuria, jaundice, drowsiness.

Clinically shock may be:

- ♦ *Compensated* with mild tachycardia, normal blood pressure, urine output, normal respiration and mild lactic acidosis.
- ♦ *Mild* shock with mild lactic acidosis, tachycardia, tachypnoea and anxiousness.
- ♦ *Moderate* shock with significant lactic acidosis, decreased urine, tachycardia, tachypnoea, drowsiness, and mild hypotension.
- ♦ *Severe* shock with severe lactic acidosis, anuria, tachypnoea with gasping, severe tachycardia, profound hypotension and unconsciousness.

Shock index is ratio of pulse rate to blood pressure. Normal is < 1. In shock, it reverses.

INVESTIGATIONS IN SHOCK

- ♦ Regular monitoring of BP, pulse.
- ♦ Heart rate. But in young patients and patients on beta blockers, tachycardia may be masked.
- ♦ Respiratory rate.
- ♦ CVP line (Central venous pressure).
- ♦ PCWP (Pulmonary capillary wedge pressure).
- ♦ Pus, urine, blood culture in case of septic shock.
- ♦ U/S, CT, X-ray depending on location of pathology or septic focus.
- ♦ Measurement of urine output.
- ♦ Arterial PO_2 and PCO_2 analysis.
- ♦ Electrolyte estimation.
- ♦ Blood CBC, pH assessment, pulse oximetry. *Serum lactate estimation* is an important prognostic indicator. Level on admission, high level reach, and level becoming normal should be assessed.

Treatment

Guidelines

- ❖ *To treat the cause*
- ❖ *To improve cardiac function*
- ❖ *To improve tissue perfusion*

Treatment of shock

- ❖ First stabilize the patient with initial resuscitation
- ❖ Next evaluate the patient for cause and severity
- ❖ Lastly treat the specific cause to achieve cure

- ♦ Treat the cause, e.g. arrest haemorrhage, drain pus.
- ♦ **Fluid replacement:** Plasma, normal saline, dextrose, Ringer's lactate, plasma expander (haemaccel) (maximum 1 litre can be given in 24 hours).

  Initially *crystalloids* then colloids are given. Blood transfusion is done whenever required.

  Fluid therapy is ideally done with crystalloids like normal saline, Ringer's lactate, Hartmann's solution. Blood loss should be corrected by blood transfusion only. Crystalloids and colloids do not have O_2 carrying capacity. Hypotonic solutions like dextrose are poor volume expanders and so should not be used in shock. *Dynamic fluid response* is studied by infusing 500 ml of fluid rapidly in 10 minutes.

  Responders show improvement; *transient responders* show improvement temporarily but revert back to original status probably due to still existing fluid/blood loss or still existing fluid shift from intravascular space; *nonresponders* will not respond as fluid loss is severe and persistently ongoing.

- *Inotropic agents*: Dopamine, dobutamine, adrenaline infusions—mainly in distributive shock like septic shock.

Note:

- Dopamine improves renal and splanchnic blood
- Dobutamine improves the cardiac output
- Adrenaline/Levarterenol IV in anaphylaxis
- Amrinone and milrinone are newer inotropic drugs

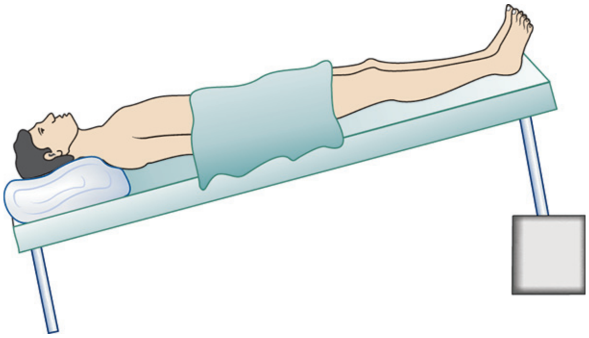

Fig. 1.220: Trendelenburg position—head down position, used in patient in shock.

- Correction of acid-base balance. Acidosis is corrected by using 8.4% sodium bicarbonate intravenously.
- *Steroid* is often life-saving. 500-1000 mg of hydrocortisone can be given. It improves the perfusion, reduces the capillary leakage and systemic inflammatory effects.
- Antibiotics in patients with sepsis; proper control of blood sugar and ketosis in diabetic patients.
- Catheterisation to measure urine output (30-50 ml/hour or > 0.5 ml/kg/hour should be maintained).
- Nasal oxygen to improve oxygenation or ventilator support with intensive care unit monitoring has to be done.
- CVP line to perfuse adequately and to monitor fluid balance. TPN is given when required.
- PCWP to monitor very critical patient.
- Haemodialysis may be necessary when kidneys are not functioning.
- Control pain-using morphine (4 mg IV).
- Ventilator and ICU/critical care management.
- Injection ranitidine IV or omeprazole IV or pantoprazole IV.
- *Activated C protein* even though costly is beneficial as it prevents the release and action of inflammatory response.
- *MAST (military antishock trouser): Provides circumferential external pressure of 40 mmHg. It is wrapped around lower limbs and abdomen, and inflated with required pressure. It redistributes the existing blood and fluid towards centre. It should be deflated carefully and gradually.*

Remember

- Response to dynamic fluid therapy can be checked in patient in shock by perfusing 500 ml of fluid rapidly in 10 minutes and patient is checked as nonresponder/transient only initial responder or proper responder.
- Vasopressor like dobutamine is used only in distributive shock like due to sepsis (not in hypovolaemic, haemorrhagic shock where there is low preload).
- Intubation and ventilator may be needed in shock.
- The patient is monitored with ECG, pulse oximetry, blood pressure/invasive blood pressure, CVP/ PCWP, urine output, pupillary reaction (dilated or not), serum electrolytes, arterial PO_2 and PCO_2 analysis.

CENTRAL VENOUS PRESSURE (CVP)

It is a method to measure the right atrial pressure by placing a venous catheter (20 cm) into the SVC (superior vena cava). Commonly for CVP monitoring, a venous catheter is passed through internal jugular vein or infraclavicular subclavian vein to the SVC (used for TPN purpose). Occasionally a long catheter (60 cm) can be passed through basilic vein (not commonly done). Under radiological guidance, initially a needle is passed 3 cm above the medial end of the clavicle, in the hollow between the two heads of sternomastoid muscles, directing towards the suprasternal notch into the right internal jugular vein. Then through a guide wire, a venous catheter is passed into the SVC through right internal jugular vein, which can also be confirmed by changes in flow during inspiration and expiration.

Catheter is connected to saline manometer, taking manubriosternal angle (angle of Louis) as zero point.

Normal value is 2-10 cm of saline.

If less than 2 cm, more fluid is infused.

If more than 10 cm, fluid infusion should be restricted.

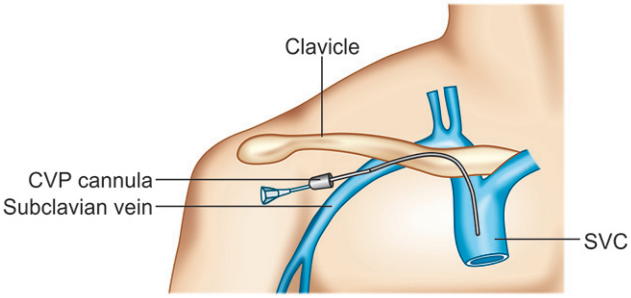

Fig. 1.221: CVP cannula. Note the location of tip of CVP line in the SVC (or can be in the right atrium).

Complications of CVP

- Pneumothorax
- Haemothorax
- Injury to brachial plexus and vessels
- Bleeding
- Sepsis
- Catheter displacement

PULMONARY CAPILLARY WEDGE PRESSURE (PCWP)

It is a better indicator of circulating blood volume and left ventricular function.

Catheter used is *Swan Ganz* triple channel pulmonary artery balloon catheter.

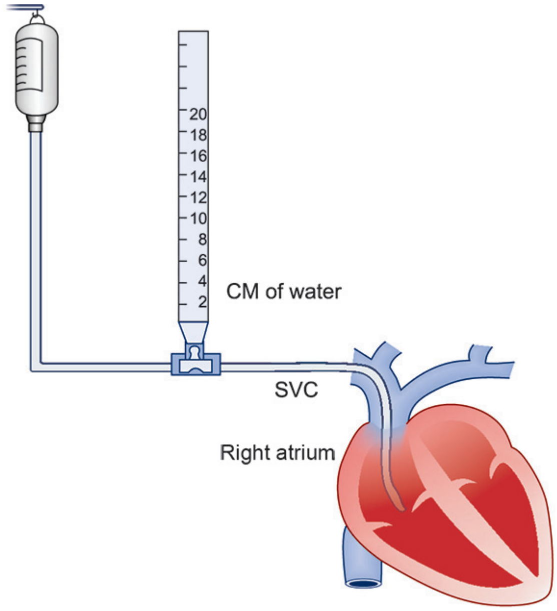

Fig. 1.222: CVP line for monitoring and perfusing the patient in shock.

It is used to:

- Differentiate right and left ventricular failure, pulmonary embolus, septic shock
- To measure and monitor cardiac output during the use of inotropic agents, vasodilators and fluid therapy

Procedure

Under strict aseptic precaution, using cannula and guide wire, catheter is passed through internal jugular vein, into the right atrium. Balloon is inflated by 1.5 ml of air and then negotiated into pulmonary artery, until it reaches a small branch and wedges it. Pressure at this point is called as *pulmonary capillary wedge pressure.*

PCWP normally is 8-12 mmHg, considering mid axillary point as zero reference point.

After that, balloon is deflated to get pulmonary artery pressure which is normally 25 mmHg systolic and 10 mmHg diastolic.

PCWP catheter can be kept *in situ* only for 72 hours.

Complications

- Arrhythmias
- Pulmonary artery rupture
- Balloon rupture
- Pulmonary infarction
- Pneumothorax
- Haemothorax
- Bleeding, sepsis, thrombosis

SYSTEMIC INFLAMMATORY RESPONSE SYNDROME (SIRS)

- *SIRS* is systemic manifestations of inflammation due to variety of causes like infection, pancreatitis, polytrauma, burns, transfusion reaction, and malignancy. So it is often categorised as *infectious cause SIRS or noninfectious cause SIRS*. It causes either hyperthermia (>38°C) or hypothermia (< 36°C); tachycardia (pulse > 90/minute); tachypnoea (> 20/minute); total white cell count > 12,000/cu mm, or count < 4000/cumm.

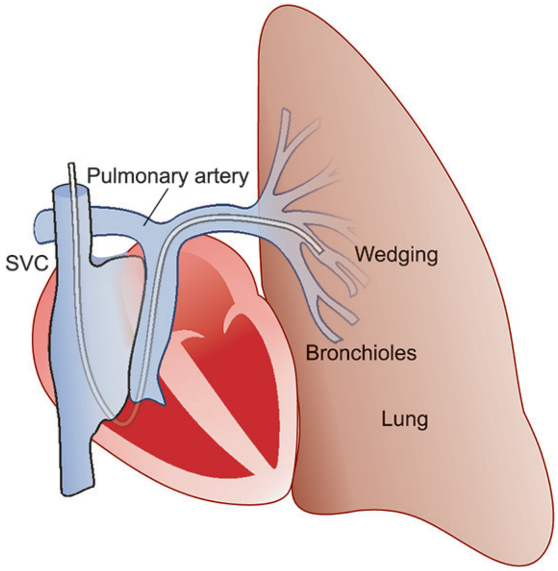

Fig. 1.223: Pulmonary capillary wedge pressure. Note the wedged balloon in the tip of the venous catheter in the pulmonary arteriole.

- It is final common pathway in shock due to any cause (trauma, sepsis, endotoxaemia, burns) where there is failure of inflammatory localisation with vasodilatation, increased endothelial permeability with damage, thrombosis, leucocyte migration and activation.
- It is associated with release of free radicals, abnormal arachidonic acid release, cytokine release, neutrophil sequestration, abnormal NO synthesis, complement activation, DIC.
- It is a part of severely decompensated *reversible shock* which eventually leads to MODS (Multiorgan dysfunction syndrome), a state of *irreversible shock* wherein patient is anuric, drowsy, cold and terminally ill.
- SIRS carries poor prognosis.

Distributive shock is sudden change in vascular capacity leading to shock like in anaphylaxis and adrenal crisis.

MULTIPLE ORGAN DYSFUNCTION SYNDROME (MODS)

- It is progressively becoming irreversible injury of all tissues like *kidney, lungs, liver, GIT*. Lungs and liver are commonly involved (70%). Next organs to be involved are kidney and GIT. Order of involvement of organs in MODS is—lungs → right ventricular failure → liver → kidney. But mortality is higher if kidney and lungs are involved.
- It occurs in critically ill patient after severe trauma, burns, acute pancreatitis, bleeding and sepsis.
- It is more common in elderly, diabetic, smokers, alcoholics, cirrhosis, malnutrition, patients under steroids and cytotoxic drugs, and uraemia.

Hope puts a smile on our face when the heart cannot manage.

- Multiple features related to multiple organ dysfunctions are typical. Oliguria, jaundice, hypotension, drowsiness, respiratory distress are common.
- Platelet microaggregation, acute pulmonary hypertension, ARDS, DIC, circulatory failure with reduced total oxygen utilization in spite of adequate oxygen supply, impaired defense mechanism are the pathogenetic features.
- Respiratory, renal, hepatic, circulatory, coagulative and cardiac failure occurs as an end stage MODS.
- *Primary MODS* is due to a well defined cause like pulmonary contusion, rhabdomyolysis, multiple transfusions.
- *Secondary MODS* occurs as result of host response in SIRS.
- *Management* of MODS is critical care in ICU with ventilator support, haemodialysis, transfusions, antibiotics, proper nutrition in the form of TPN or enteral. MODS stage has got high mortality.

OXYGEN THERAPY

Indications

1. Chest injuries, any severe haemorrhage.
2. Gas gangrene with toxic haemolysis.
3. Coal gas poisoning.
4. Over morphinisation.
5. Pulmonary embolism and fat embolism.
6. Spontaneous pneumothorax, pulmonary oedema, cardiac infarction, pneumonia, cor pulmonale.
7. Cardiogenic shock and acute bronchitis.

27% oxygen is delivered through ventimask (disposable polythene mask) at a rate of 4-6 litres per minute. Oxygen is also given along with positive pressure ventilation.

HYPERBARIC OXYGEN

It is administration of oxygen 1 or 2 atmospheres above the atmospheric pressure in a compression chamber. It increases the arterial oxygen saturation so that oxygen perfusion of tissues will be increased.

Indications

1. Carbon monoxide poisoning.
2. Tetanus, gas gangrene infections.
3. Bedsores, frostbites, necrotising fasciitis.
4. Drenching in paralytic ileus to reduce the nitrogen gas in distended bowel.
5. As a radiosensitizer in the treatment of cancer.

Complications

- Cerebral gas embolism.
- Rupture of tympanic membrane.
- Visual defects.
- O_2 toxicity.
- CO_2 narcosis.
- Respiratory depression.

Contraindication

- Asthma/emphysema.
- High fever.
- Chronic sinusitis.
- Viral infection.
- Pregnancy.

TOPICAL O_2-THERAPY

H_2O_2 to release nascent oxygen in ulcers and abscess.

CARDIAC ARREST

*It is the cessation of the heart. Heart stops contracting***.**

Causes: All causes for shock.

Differences between CVP and PCWP

	CVP		*PCWP*
1.	Technically easier	1.	Requires skilled experts
2.	Normal pressure is 2-10 cm of saline	2.	8-12 mmHg
3.	Gives gross idea about fluid balance	3.	Better and specific
4.	Left ventricular function is not assessed	4.	Left ventricular function is very well-assessed
5.	Not used to differentiate between right and left ventricular function	5.	Very well-differentiated
6.	Can be kept *in situ* as long as desired	6.	Cannot be kept *in situ* for more than 72 hours
7.	Catheter tip is in SVC	7.	Catheter tip is in pulmonary capillary with wedging
8.	Plain tip catheter	8.	1.5 ml air filled balloon tip
9.	Can be used for TPN, fluid infusion, etc.	9.	Can not be used for TPN, or fluid infusion
10.	Complications are easy to tackle	10.	Often difficult to tackle
11.	Not as sensitive and specific as PCWP	11.	Sensitive and specific

Features of cardiac arrest

- No palpable pulse
- Heart sounds not heard
- Cessation of respiration—cyanosis occurs
- Development of unconsciousness
- Pupils start dilating

Critical Period

Once heart and lungs stop, *brain death occurs in 3 minutes*.

Immediate measures

- **A**irway
- **B**reathing
- **C**ardiac compression
- **D**rugs and **D**efibrillator
- **E**CG, **E**ndotracheal tube and **M**onitor

a. **External cardiac compression (massage)**: Patient is laid flat on a *hard surface (never on soft surface)*. Manual compression is exerted over the lower sternum using both hands one over other without bending the elbow at a rate of 60 to 70 per minute. Rib cage damage during procedure can be very well ignored (Heel of right hand is placed over the sternum 8 cm above xiphoid process and left hand is placed over it).
b. Another person at the same time should give *mouth to mouth breathing* at a rate of 20 to 30 per minute after clearing the airway by removing froth and dentures. A bag with mask can be used to ventilate using air or oxygen.
c. Endotracheal intubation and ventilator support.
d. Injection of 1:10,000 adrenaline and 10% calcium chloride intravenously.
e. Sodium bicarbonate 8.4% injection, hydrocortisone injection.
f. Defibrillator, if there is ventricular fibrillation.
g. Analysis of blood gas (PCO_2 and PO_2), and serum electrolytes assessment at repeated intervals.
h. Urinary catheterization, Ryle's tube insertion.
i. Monitoring the patient with BP, pulse, respiration, and temperature chart.

Observations to be made are:

- Groin pulse
- Respiration and breath sounds
- Blood pressure
- Pupillary reaction
- ECG activity

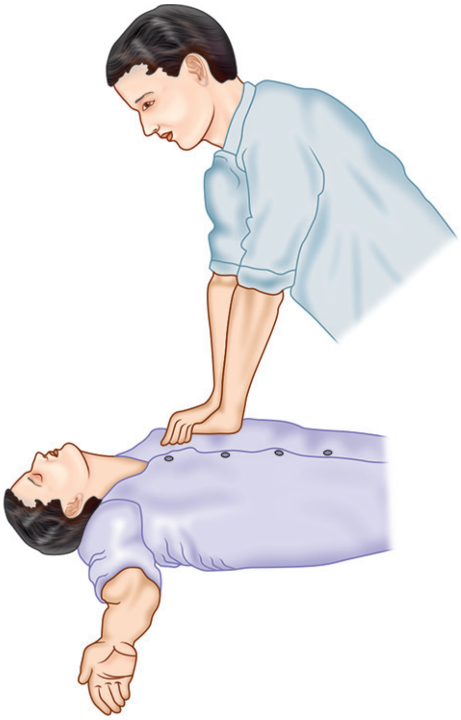

Fig. 1.224: Technique of external cardiac massage.

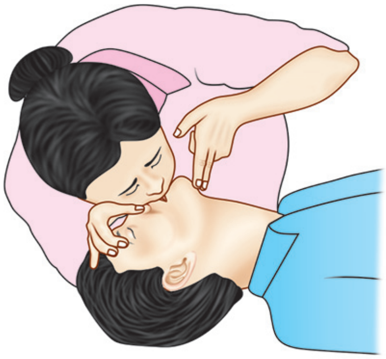

Fig. 1.225: Technique of mouth-to-mouth respiration. Note patient's nose should be held closed with right hand of the doctor/assistant and with left hand lower jaw should be pushed forward.

Once patient recovers, the cause and sequelae has to be managed properly.

Sequelae are due to hypoxia and circulatory collapse

- Cerebral oedema and permanent brain damage
- ARDS (Adult respiratory distress syndrome)
- Renal failure

Internal Open Cardiac Massage

This method is used when cardiac arrest occurs in the operation theatre during surgery, acute tamponade, and acute bilateral pneumothorax.

Left side thorax is opened through a lengthy incision along 4th or 5th intercostal space. Initially heart with intact pericardium is rhythmically compressed and relaxed using left hand against sternum. Mean while costal cartilages above and below are cut with a knife to have a better exposure. Pericardium is opened in front of the phrenic nerve. Direct cardiac massage is undertaken until heart regains its function and later shifted to ventilatory support and critical care.

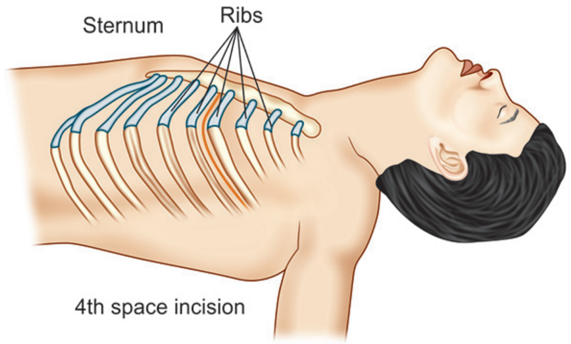

Fig. 1.226: Left thoracotomy for open internal cardiac massage.

Defibrillation Technique (Cardioversion)

Apply gelly to the site of electrodes. One electrode at the base of heart to the right of the sternum other over the estimated area of the apex of the heart. ***Ensure that nobody is in contact with the patient***. Activate the defibrillator. Resume ventilation and ECG monitor immediately. After that, monitor continuously, correct the acidosis, catheterise and observe urine output. Arrange for ICU care.

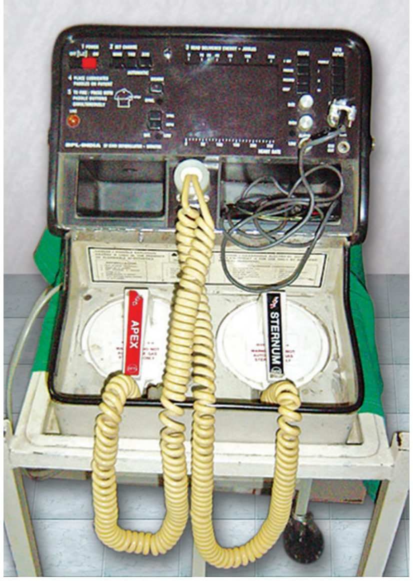

Fig. 1.227: Defibrillator used in case of cardiac arrest.

H. Haemorrhage and Blood Transfusion

> *There are canals (or vessels) in it (the heart) to (every) member. Now if the priests of Sekhmet or any physician put his hands (or) his fingers (upon the head, upon the back of the) head, upon the two hands, upon the pulse, upon the two feet, (he) measures the heart, because its vessels are in the back of the head and in the pulse; and because its (pulsation is in) every vessel of every member.*
>
> —**(Anonymous),** Circa 2500 BC

CHAPTER OUTLINE

- Haemorrhage
- Blood Transfusion
- Blood Substitutes
- Massive Blood Transfusion
- Autologous Blood Transfusion
- Artificial Blood
- Erythropoietin
- Tourniquets
- Disseminated Intravascular Coagulation
- Mechanism of Blood Coagulation

HAEMORRHAGE

Classification

I. Based on the source of bleeding:
   a. ***Arterial*** is bright red in colour, spurting like jet along with pulse of the patient.
   b. ***Venous*** is dark red, steady and continuous flow. Blood loss may be severe and rapid when bleeding is from femoral vein, jugular vein, other major veins, varicose veins, portal vein, oesophageal varices.
      Pulmonary arterial blood is dark red in colour and pulmonary venous blood is bright red in colour.
   c. ***Capillary:*** Here bleeding is rapid and bright red. It is often torrential due to continuous ooze.

II. Based on the time of onset of bleeding in relation to any operative procedure:
   1. ***Primary:*** Occurs at the time of injury or operation.
   2. ***Reactionary:*** It occurs within 24 hours after surgery or after injury (commonly in 4-6 hours).

Precipitating factors	*Common causes*
Coughing	Thyroid surgery
Vomiting	Cholecystectomy
Straining	Major abdominal surgeries
Rise of blood pressure	Circumcision
Restlessness	Hydrocele surgery
Venous refilling during recovery from anaesthesia	Tonsillectomy
Slipping of ligature	
Clot dislodgement	

   3. ***Secondary:*** It usually occurs in 14 days after surgery.

Factors	*Common causes*
Infection	Erosion of carotid artery by cancer (secondaries in the neck)
Pressure by drain or bone	Haemorrhoidectomy
Malignancy	Inguinal block dissection

III. Based on the type of haemorrhage:

Revealed haemorrhage	*Concealed haemorrhage*	*Initially concealed but later revealed*
It is visible	It constitutes	Haematuria
external	internal haemorrhage	Haematemesis
haemorrhage	• Liver injury	Melaena
	• Spleen injury	
	• Fracture femur	
	• Ruptured ectopic gestation	
	• Cerebral haemorrhage	
	• Haemothorax	

IV. Based on the duration of haemorrhage:
   1. ***Acute haemorrhage:*** It is sudden, severe haemorrhage after trauma, surgery.
   2. ***Chronic haemorrhage:*** It is chronic repeated bleeding for a long period like in haemorrhoids, bleeding peptic ulcer, carcinoma caecum, etc. They present with chronic

No problem is too large for God's intervention and no person is too small for God's attention anaemia with hyperdynamic cardiac failure. They are in a state of chronic hypoxia. It is corrected by *packed cell transfusion* not by whole blood itself. Cause has to be treated accordingly.

3. ***Acute on chronic haemorrage:*** It is more dangerous as the bleeding occurs in individuals who are already hypoxic, which may get worsened faster.

V. Based on the possible intervention:

1. *Surgical haemorrhage*—can be corrected by surgical intervention.
2. *Nonsurgical haemorrhage*—is diffuse ooze due to coagulation abnormalities and DIC.

Pathophysiology of Haemorrhage

Bleeding
↓
Hypovolaemia
↓
Low cardiac output
↓
Tachycardia and shunting of blood from splanchnic vessels by venoconstriction so as to maintain perfusion of vital organs like brain, heart, lungs, kidneys
↓
Hypoxia
↓
Activation of cardiac depressants
↓
Anaerobic metabolism and altered cell membrane function causing influx of more sodium and calcium inside the cell and potassium comes out of the cell
↓
Hyponatraemic, hyperkalaemic, hypocalcaemic metabolic acidosis
↓
Lysosomes of cell get lysed releasing powerful enzymes which is lethal to cell itself
↓
SICK CELL SYNDROME

Platelets and coagulants are activated leading to formation of small clots DIC and further bleeding.
↓
Progressive haemodilution leading to ***total circulatory failure.***

Initially there is compensatory hypovolaemic shock and later there is decompensatory hypovolaemic shock which will lead to MODS and death.

DIC, acidosis and hypothermia are the major factors in worsening the situation in haemorrhage.

Clinical Features of Haemorrhage

- Pallor, thirsty, cyanosis.
- Tachycardia, tachypnoea.
- Air hunger.
- Cold clammy skin due to vasoconstriction.
- Dry face, dry mouth and goose skin appearance (due to contraction of arrector pilorum).
- Rapid thready pulse, hypotension.
- Oliguria.
- Features related to specific causes.

Signs of significant blood loss

- Pulse > 100/minute
- Systolic BP< 100 mmHg
- Diastolic BP drop on sitting or standing > 10 mmHg
- Pallor/ sweating
- Shock index (ratio of pulse rate to blood pressure) > 1 (cardiac index)

Measurement of Blood Loss

- Clot size of a clenched fist is 500 ml.
- Blood loss in a closed tibial fracture is 500-1500 ml; in a fracture femur is 500-2000 ml.
- Weighing the swab before and after use is an important method of on-table assessment of blood loss.

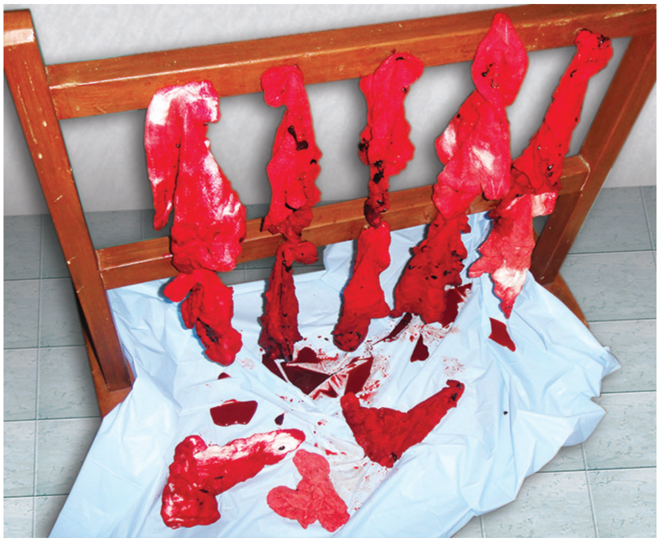

Fig. 1.228: Mops used during major surgery should be kept in a rack so that quantity of bleeding during surgery can be assessed approximately.

Classification of haemorrhagic shock (circulatory failure)

Class	*Blood loss*	*Features*
I	Up to 15% (< 750 ml)	Normal
II	Blood loss 15-30% (750-1500 ml)	Palor, thirsty, tachycardia
III	Blood loss 30-40% (1500-2000 ml)	Hypotension, tachycardia, oliguria, confusion
IV	Blood loss > 40% (> 2000 ml)	Rapid pulse, low BP, anuria, unconsciousness, MODS

Rains Factor

Total amount of blood loss = Total difference in swab weight × 1.5
or
Total difference in swab weight × 2
(For larger wounds and larger operations)

- Hb% and PCV estimation.
- Blood volume estimation using radioiodine technique or micro-haematocrit method.
- Measurement of CVP or PCWP.
- Investigations specific for cause: U/S abdomen, Doppler and often angiogram in vascular injury, chest X-ray in haemothorax, CT scan in major injuries, CT scan head in head injuries.

Effects of haemorrhage

- Acute renal shut down
- Liver cell dysfunction
- Cardiac depression
- Hypoxic effect
- Metabolic acidosis
- GIT mucosal ischaemia
- Sepsis
- Interstitial oedema, AV shunting in lung—ARDS
- Hypovolaemic shock—MODS

Treatment

- Restoration of blood loss: By blood transfusion, albumin 4.5%, SAG-M blood, saline, Haemaccel (Gelatin), dextran, plasma infusions.
  Note: One unit of blood should raise 1 gm% of haemoglobin.
- Catheterisation, foot end elevation, monitoring.
- Oxygen support/intubation/ventilator and critical care.
- Pressure, packing and *head down (Trendelenburg)* position to restore BP and blood supply of brain.
- Wound exploration and proceeding, i.e. ligation of the small vessel, suturing the wound part, vessel suturing (anastomosis), excision of the tissues.
- Absolute rest, analgesics, morphine 10-20 mg IM/IV to relieve pain, sedation.
- ICT placing for haemothorax.
- Laparotomy for liver or spleen or mesentery or bowel injuries, suturing, splenectomy.
- Topical applications for local ooze—Oxycel, gauze soaked with adrenaline, bone wax for oozing from bone and other local haemostatic agents (collagen, thrombin).
- In venous haemorrhage, elevation, ligation of vein or in case of large vein suturing of venous wall, pressure bandaging, packing will be helpful.
- Tourniquet are often used in operation theatre for control of haemorrhage in limbs. But it is not advisable as a first aid measure.
- TPN, CVP monitoring, electrolyte management are all equally important.
- Steroid injection, antibiotics, ventilator support are often required.

Management concepts

- Confirm shock, hypovolaemia and haemorrhage by clinical assessment
- Immediate resuscitation by blood, oxygen, fluid
- Identify site of haemorrhage—ultrasound, endoscopy, CT scan, diagnostic peritoneal lavage (DPL), blood tools
- Control of haemorrhage—surgery, endoscopic control, therapeutic embolisation
- Definitive treatment if any
- Sepsis control
- Prevention of coagulopathy by FFP, platelet concentrate, fresh blood
- Critical care management
- End point resuscitation, fluid and electrolyte management, prevention of organ failure

Local haemostatic agents

- Gelatin sponge (Gel foam)
- Oxidised cellulose (Surgicel)
- Collagen sponge (Helistat)
- Microfibrillar collagen powder (Avitene)
- Topical thrombin
- Bone wax (derived from bees wax + almond oil)
- Gelatin matrices (Floseal)
- Topical EACA, topical cryoprecipitate

A
Ab Gel®
ABSORBABLE GELATINE SPONGE USP
Size :80MM X 50mm X 10mm Approx.
Sterile, Haemostatic, Absorbable Gelatine Base Foam Specially Prepared For SURGICAL USE
Mfg. Lic. No. 130
WARNING : The Sterility of the product cannot be guaranteed if the package bears evidence of damage, or if the package has been previously opened. or wet
Manufactured by
SRI GOPAL KRISHNA LABS. PVT. LTD.
CUT HERE
B

Figs 229A and B

Every obstacle introduces a person to himself.

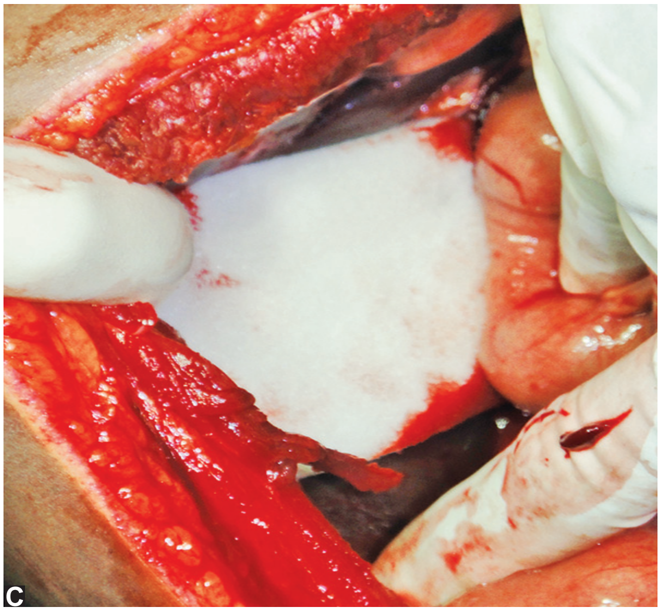

Fig. 229C

Figs 1.229A to C: Gel foam (gelatin sponge) is a good haemostatic agent on oozing surface like gallbladder bed after cholecystectomy.

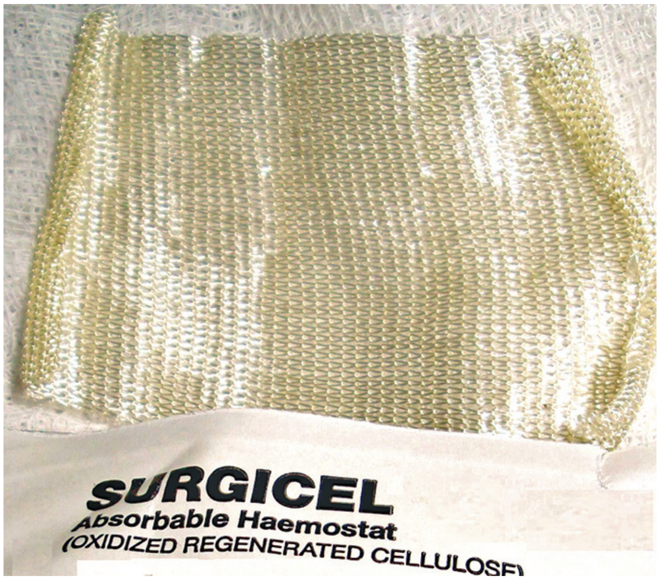

Fig. 1.230: Oxidised regenerated cellulose is very good local haemostatic agent; but it is much costlier than gelatin sopnge.

BLOOD TRANSFUSION

> *This noon I met with Mr (Robert) Hooke, and he tells me the dog which was filled with another dog's blood at the (Gresham) College the other day, is very well, and like to be so as ever. And doubts not its (i.e. blood transfusion) found being of great use to men.*
>
> —**Samuel Pepys,** 1666

Indications

- Acute blood loss following trauma, ≥15% of total body volume in otherwise healthy individuals (liver, spleen, kidney, GIT injuries, fractures, haemothorax, perineal injuries).
- During major surgeries—abdominoperineal surgery, thoracic surgery, hepatobiliary surgery.
- Following burns.
- In septicaemia.
- As a prophylactic measure prior to surgery.
- Whole blood is given in acute blood loss.
- Packed cells are given in chronic anaemia.
- Blood fractions are given in ITP, haemophilias.

Donor Criteria

- Donor should be fit without any serious diseases like HIV1 and HIV2 and hepatitis infections and malaria.
- Weight of donor should be more than 45 kg.

Collection of Blood

Blood is collected in a sac containing 75 ml of CPD (Citrate phosphate dextrose) solution and stored in special refrigerators at 4 degree celcius. CPD blood lasts for 3 weeks.

In stored blood

- RBC's last for 3 weeks
- WBC's are destroyed rapidly
- Platelets also get reduced in 24 hours
- Clotting factors are labile and so their levels fall quickly

Blood Fractions

1. ***Packed cells***
   - It is obtained by centrifuging whole blood at 2000-2300 g for 15-20 minutes.
   - It is used in chronic anaemias, in old age, in children.
   - It minimises the cardiac overload due to transfusion.
   - It can be stored for 35 days at 1°-6°C.
2. ***Plasma:***
   - This is obtained in the same way as packed cells by centrifugation.
   - It is indicated in *burns, hypoalbuminaemia, severe protein loss.*
   - It can be fractionalised into different fragments:
     a. ***Human albumin 4.5%*** is obtained after repeated fractionations and can be stored for several months in liquid form at 4°C**.**
     b. ***Fresh frozen plasma (FFP)*****:** Fresh plasma obtained, is rapidly frozen and stored at –40°C. It contains all coagulant factors. 1 unit of FFP increases the clotting factors levels by 3%. It can be stored for 2 year. Rhesus D positive FFP can be transfused to Rhesus D negative female.

     Uses:
     - Severe liver disease with abnormal coagulation function.
     - Congenital clotting factor deficiency.
     - Deficiency following warfarin therapy, DIC, massive transfusion.
     - To maintain prothrombin time at normal level. Dose of FFP is 15 ml/kg.

c. ***Cryoprecipitate*:** When fresh frozen plasma is allowed to thaw at 4°C, visible white supernatant layer develops and is called as cryoprecipitate which is rich in Factor VIII and fibrinogen. It is stored at minus 40°C and can be kept for 2 year. Cryoprecipitate is used to raise fibrinogen level at a dose to make plasma fibrinogen level 150 mg/dl. It is also used in inherited deficiency of Factor VIII, fibrinogen, Factor XIII, von Willebrand's disease.

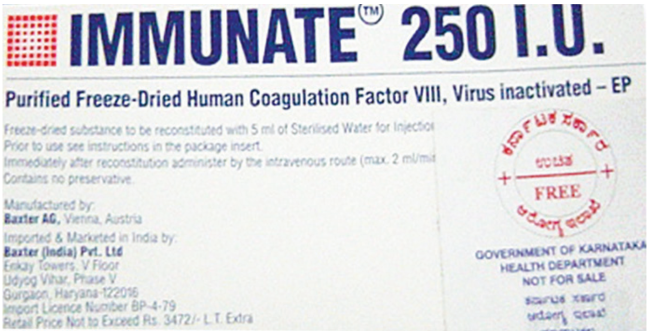
IMMUNATE(TM) 250 I.U.
Purified Freeze-Dried Human Coagulation Factor VIII, Virus inactivated – EP
Freeze-dried substance to be reconstituted with 5 ml of Sterilised Water for Injectio
Prior to use see instructions in the package insert.
Immediately after reconstitution administer by the intravenous route (max. 2 ml/mi
Contains no preservative.
Manufactured by:
Baxter AG, Vienna, Austria
Imported & Marketed in India by:
Baxter (India) Pvt. Ltd
Enkay Towers, V Floor
Udyog Vihar, Phase V
Gurgaon, Haryana-122016
Import Licence Number BP-4-79
Retail Price Not to Exceed Rs. 3472/- L.T. Extra
FREE
GOVERNMENT OF KARNATAKA
HEALTH DEPARTMENT
NOT FOR SALE

Fig. 1.231: Purified freeze dried human coagulation factor VIII is available for use in haemophilia patients

d. ***Fibrinogen*** is obtained by organic liquid fractionation of plasma and is stored in dried form. It is very useful in DIC and afibrinogenaemia. It has risk of transmitting hepatitis.

e ***Factor VIII and IX concentrate**:* They are freeze dried part from a large pooled plasma used in haemoplhilia and von Willebrand's disease.

3. ***Platelet rich plasma***: It is obtained by centrifugation of freshly donated blood at 150-200 g for 15-20 minutes.Platelet rich plasma contains 5.5 × 10^9/L platelets in 50 ml plasma. It can be random donor platelet or single donor platelet. Single donor platelet is prepared by plateletapheresis containing 3 × 10^9/L platelets in 200 ml of plasma. One single donor platelet is equal to 8 units of random donor platelet.
4. ***Platelet concentrate:*** It is prepared by centrifugation of platelet rich plasma at 1200-1500 g for 15-20 minutes. Used in thrombocytopenia and drug (aspirin, clopidogrel) induced haemorrhage. Platelet is transfused at a dose of 0.1 unit/kg, when platelet drops below 50,000 with episodes of bleeding.
5. ***Prothrombin complex concentrate (PCC)*** are derived from pooled plasma which contains factors II, IX and X; used in emergency reversal of warfarin therapy in uncontrolled haemorrhage.

SAG-M Blood

A proportion of donations will have plasma removed and will be replaced by crystalloid solution of SAG-M.

S — Sodium chloride.
A — Adenine.
G — Glucose anhydrate.
M — Mannitol.

Advantages

- This allows good viability of cells
- But it is devoid of any protein
- It is very useful in anaemias

Precautions

- For every four units of SAG-M blood, one whole blood has to be given.
- Later for every two units of SAG-M blood, one unit (400 ml) of 4.5% human albumin has to be given.
- Coagulation status and platelet count should be checked regularly.

After grouping and cross-matching, 540 ml of blood is transfused in 4 hours (40 drops per minute), using a filtered drip set.

One litre of blood contains 350 mg of iron. Normal excretion of iron is 1 mg/day. Iron overload can occur after many transfusions. Iron excretion can be increased by desferrioxamine infusion.

Complications of Blood Transfusions (Please also See Table for Entire List)

Febrile reactions: It is the most common complication due to impurities like pyrogens in the blood or in infusion set. Headaches, fever, chills and rigors, tachycardia, nausea are the features. Transfusion is temporarily stopped or the flow is slowed down with administration of antipyretic drug to reduce fever. Often transfusion of that unit needs to be discontinued.

Allergic reaction (3%): Utricaria and allergy to specific proteins in the donor's plasma can occur. Usually it is mild and is treated with steroid and antihistaminics. In severe utricaria that unit of blood is discarded; new washed RBC's and platelets are used.

Acute haemolytic reactions: It is the most dangerous complication. It is due to ABO incompatibility. Usually it is *nonfatal* but occasionally can be *fatal*. It is commonly due to technical error at different levels. It amounts for criminal negligence in court of law. Intravascular destruction causes haemoglobinemia, haemoglobinuria, acute renal failure and DIC. Dyspnoea, chest pain, sweating, fever with chills, tachycardia, hypotension, and cardiac arrest occurs in fatal type. *Jaundice is a common* feature in nonfatal type. Free haemoglobin level in blood will be above 5 mg/dl. Condition is treated as an emergency in critical care unit/ICU. Transfusion is stopped immediately; blood sample of recipient and transfusing blood is sent immediately for two laboratories for rechecking. *Smoky urine* of the patient is typical. Injection hydrocortisone/dexamethasone IV is given immediately. *Fluid therapy*, *alkalization of blood* is done using sodium lactate and sodium bicarbonate. Mannitol 20 gram in 100 ml is infused in 5 minutes; furosemide 120 mg is injected intravenously. Haemodialysis is needed if there is renal failure. Often ventilator support, defibrillator if cardiac arrest occurrs is needed. Correction of acidosis, electrolytes is needed.

The very act of believing creates strength of its own.

Transfusion related acute lung injury: It is due to donor plasma antibody against HLA and leukocyte specific antigens of recipient. Occasionally it is due to recipient's antibody against donor's leukocytes. Features are—breathlessness, saturation drop, fever, hypotension which is observed 4 hours after transfusion. Chest X-ray shows bilateral diffuse infiltrate. They need ventilator support for short period with eventual rapid and complete recovery.

Transfusion related graft versus host disease (TGVH): This *very serious*, very rare complication occurs due to recognition and reaction against host tissues by infused donor lymphocytes. It is common in immunosuppressed, lymphoma, leukaemic patients. Any type of blood products including leukocyte reduced blood can cause the condition. Features are—pancytopaenia, toxic epidermal necrosis, liver dysfunction with *more than 90%* mortality. It is difficult to treat.

Congestive cardiac failure (CCF): It occurs if especially large quantities of whole blood are transfused in chronic severe anaemia, pregnancy, elderly patients, in patients who have cardiac problems.

Complications of blood transfusion

- Congestive cardiac failure
- Transfusion reactions HBV, HCV
  - Incompatibility. Major and minor reactions with fever, rigors, pain, hypotension
  - Pyrexial reactions due to pyrogenic ingredients in the blood
  - Allergic reactions
  - Sensitisation to leukocytes and platelets
  - Immunological sensitisation
- Infections
  - Serum hepatitis
  - HIV infection
  - Bacterial infection
  - Malaria transmission
  - Epstein-Barr virus infection
  - Cytomegalovirus infection
  - Syphilis, Yersinia
  - *Babesia microti* infection
  - *Trypanosoma cruzi* infection
- Air embolism
- Thrombophlebitis
- Coagulation failure
  - Dilution of clotting factors
  - DIC
  - Dilutional thrombocytopenia occurs in patients with massive blood transfusion
- Circulatory overload causing heart failure
- Haemochromatosis in patients with CRF receiving repeated blood transfusions
- Citrate intoxication causes bradycardia and hypocalcaemia. For every four units of blood 10 ml of 10% calcium chloride or gluconate should be infused intravenously
- Iron overload

BLOOD SUBSTITUTES

- ***Human albumin* 4.5%**
  There is no risk of transmitting hepatitis.
  Plasma *fractionation* is done using organic liquids and heat to extract *albumin* which is stored at 4°C for many months. It can be used in patients with cirrhosis, burns, nephrotic syndrome, ovarian hyperstimulation syndrome (occurring after ovarian stimulation with gonadotrophin injections during *in vitro* fertilization (IVF) therapy). One gram of albumin binds with 14 ml of water so it increases the blood volume also. Albumin is infused daily as needed until good response is observed. Albumin is expensive.
- ***Dextrans*** are useful to improve plasma volume. They are polysaccharides of varying molecular weights.
  This is derived from *leuconostoc mesenteroides* bacteria after adding yeast. One gram of dextran binds with 20 ml of water to raise the plasma volume.
  a. *Low molecular weight dextran* (40,000 mol wt) (*Dextran 40, Rheomacrodex*).
  Dextran 40 is very effective in restoring blood volume immediately. But small molecules are readily excreted in kidney and so effect is transitory. It may be useful in prevention of sludging in kidney and hence renal shut down.
  b. *High molecular weight dextran* (*Dextran 110 and Dextran 70*).
  Less effective but long acting and so useful to have prolonged effect.

Precautions

1. Blood samples for blood group and cross-matching should be taken before giving dextrans as it interferes with Rouleaux formation of red cells.
2. Dextrans also interfere with platelet function and so may precipitate abnormal bleeding.
3. Total volume of dextrans should not exceed 1000 ml.

- ***Gelatin,*** in a degraded form of mol. wt. 30,000S, is used as a plasma expander. Up to 1000 ml of 3.4-4% solution containing anions and cations is given intravenously—***Haemaccel***. But it is less effective than dextran and after 4 hours of its infusion, only 30% remains intravascular.
- ***Hydroxyethystarch:*** It contains starch, sodium hydroxide, ethylene oxide. It is a good plasma volume expander but lasts only for 6 hours.

MASSIVE BLOOD TRANSFUSION

- It is defined as *replacement or transfusion of blood equivalent to patient's blood volume in < 24 hours corresponding to that particular age (In adult it is 5-6 litres, in infants it is 85 ml/kg body weight.)* Or single transfusion of blood more than 2,500 ml continuously
- Massive transfusion is used in severe trauma associated with liver, vessel, cardiac, pulmonary, pelvic injuries. Often it is required during surgical bleeding (primary haemorrhage on table) of major surgeries

- Adverse effects of massive transfusion
  - a. Severe electrolyte imbalance (hypocalcaemia, hyperkalaemia, acidosis)
  - b. Coagulopathy—altered platelet and coagulation factors.
    - – Dilutional thrombocytopenia
  - c. Citrate toxicity
  - d. Hypothermia
  - e. Poor oxygen delivery—due to reduced 2,3 DPG
  - f. Infections
  - g. Incompatibility and transfusion reactions
  - h. ARDS, DIC

AUTOLOGOUS BLOOD TRANSFUSION

An healthy individual with no infection and haematocrit of ≥ 30% can predonate blood few weeks prior to any elective surgeries which in turn can be used at the time of surgery.

Autologous blood is used in orthopaedic, gynaecologic and urologic surgeries. Patient donates one unit of blood weekly; last one if at all being 72 hours before the date of surgery.

Recycled Blood

In major surgeries if there is significant blood loss, then patient's bled blood is carefully sucked out through a sterile system and is filtered and reused again to the patient. This will reduce the number of transfusions.

ARTIFICIAL BLOOD

1. *Perfluorocarbon (Fluosoleda)*—abiotic substitute as synthetic oxygen carrier. Its half life is 7 days. It is RBC substitute.
   - It has got high affinity for O_2.
   - It is inert, colourless, odourless, dense, poorly soluble liquid.
   - It is biocompatible.
   - It is emulsified with albumin or lipids before infusion. Its emulsion alone injection can cause pulmonary embolism.
   - It can bind and release oxygen. But as it reduces the PPO_2 quickly, it is a disadvantage. Patient ideally to be kept in hyperbaric place.
2. *Stroma free haemoglobin*—biomimetic haemoglobin based substitute.
3. *Chelates which reverse bound O_2.*
   - Intraoperative—salvage of blood: On table blood is collected, washed, filtered and transfused. Used in trauma.

ERYTHROPOIETIN

- Injection 1000-3500 units preoperatively also used to increase the RBC count.
- It is used in CRF patients who are on haemodialysis. It is given twice weekly but it is costly.

TOURNIQUETS

Tourner means to turn (Greek). A tourniquet is used to cut off the blood supply to a limb temporarily so that a bloodless field is created while performing the surgery. Limb should be exsanguinated before applying/inflating the cuff of the tourniquet. It is done using a bandage or pressurized Rhys-Davis exsanguinator. A tourniquet is applied in mid-thigh above the knee joint in lower limb and in mid-biceps level above the elbow in upper limb. It should not be closer to joints. It is not applied over the forearm or leg. It is applied over layers of gauze or cotton, not over a bare skin. Pressure used in upper limb is 250 mmHg; lower limb is 300 mmHg (In children, it is 150 and 250 mmHg for upper and lower limbs respectively).

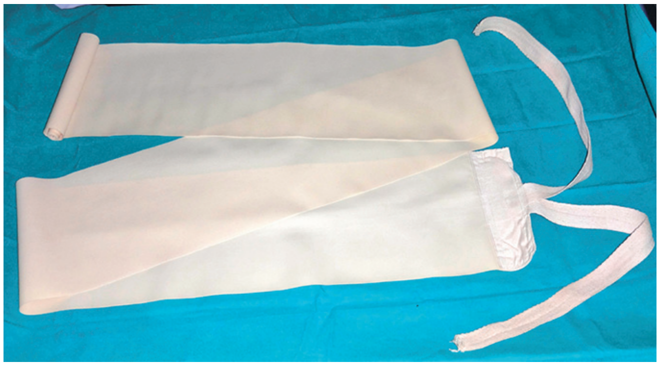

Fig. 1.232: Esmarch rubber tourniquet.

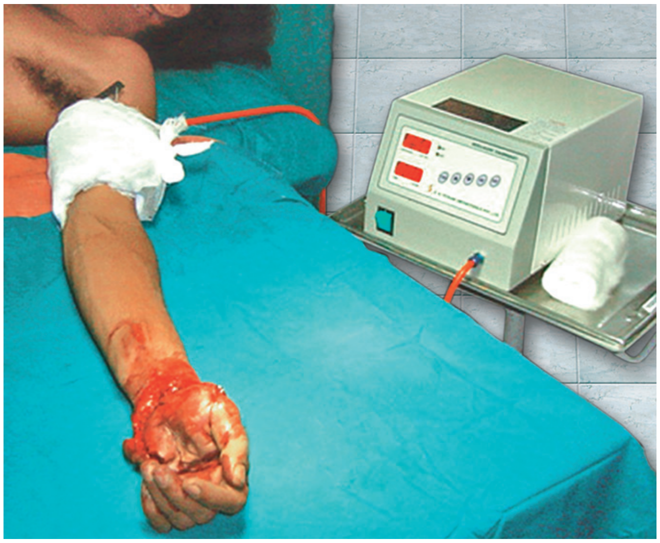

Fig. 1.233: Automatic gauged pressure tourniquet.

Uses

- To attain bloodless field in limb surgeries—upper and lower limbs, orthopaedic surgeries, soft tissue tumours, amputations.
- It is used (rubber tourniquet) to access veins for IV injections and IV sampling.
- Tourniquet is used in diagnostic tests for varicose veins, purpura (ITP), carpal tunnel syndrome, tetany.
- It is used as a first aid in bleeding conditions of limbs, snake bite (it is controversial).

- Tourniquets are often used for small procedures in fingers and toes.

Types

- **Rubber tourniquet:** Simple red rubber catheter is used for drawing blood, to have access to veins.
- Martin's tourniquet made up of India rubber.
- **Pneumatic tourniquet:** Used in limbs, will give the arterial pressure and also acts as a tourniquet (Sphygmomanometer cuff is simpler and easily available type).
- Esmarch rubber **elastic bandage tourniquet.**
- Conn pneumatic tourniquet is manually operated tourniquet where air is pumped up to the required pressure.
- Specialised sophisticated tourniquets are available which gauge pressure and time accurately—automatic tourniquet.

Tourniquet time for upper limb is one hour and for lower limb is two hours.

Contraindications

- In all peripheral vascular diseases and atherosclerosis.
- Infection.
- Deep venous thrombosis.
- Crush injuries.
- Sickle cell disease.

Complications

- Crushing effect on muscles in thigh occurs leading to crush syndrome.
- Tourniquet palsy in upper limb (radial nerve involvement)—neuropraxia.
- Infection.
- Improper application of tourniquet leads to more bleeding.
- Forgetting the removal of tourniquet or taking more time to release may compromise the blood supply of the limb leading to severe ischaemia and gangrene.
- Skin blistering and necrosis.

DISSEMINATED INTRAVASCULAR COAGULATION (DIC)

DIC is a manifestation due to wide spread intravascular coagulation resulting in microthrombi formation, consumption of platelets and clotting factors and production of breakdown products eventually leading into severe bleeding and tissue ischaemia.

Causes

- Major trauma causes DIC due to release of tissue thromboplastin. Burns, major surgery can also cause DIC.
- Sepsis is the most common cause of DIC. Common sepsis causing DIC are gram-negative, meningococcal, malarial, histoplasmosis, aspergillosis, etc.
- Acute pancreatitis can cause DIC by releasing proteolytic enzymes which activate prothrombin and factor X.
- Septic abortion, abruption, retained dead foetus, amniotic fluid embolism are obstetric causes of DIC.
- Carcinoma of pancreas, prostate, acute promyelocytic leukaemia often cause DIC.
- Haemolysis, snake bite, liver dysfunction are other causes.

Types of DIC

- *Acute DIC* presents with bleeding in gums, GIT, venepuncture site, haematuria, petechiae, oozing from surgical or traumatic wounds. Massive bleeding also can occur.
- *Chronic DIC* is a low grade type with thrombotic features.

Investigations

- In DIC—bleeding time, platelet counts are reduced. Thrombin time (TT), prothrombin time (PT) and activated partial thromboplastin time (APTT) are prolonged. Fibrinogen degradation product (FDP), D-dimer test are raised.
- Complete haematocrit, investigations relevant to cause, renal function tests, LFT, electrolyte estimation, blood/discharge/pus/urine culture.

Coagulation cascade system:

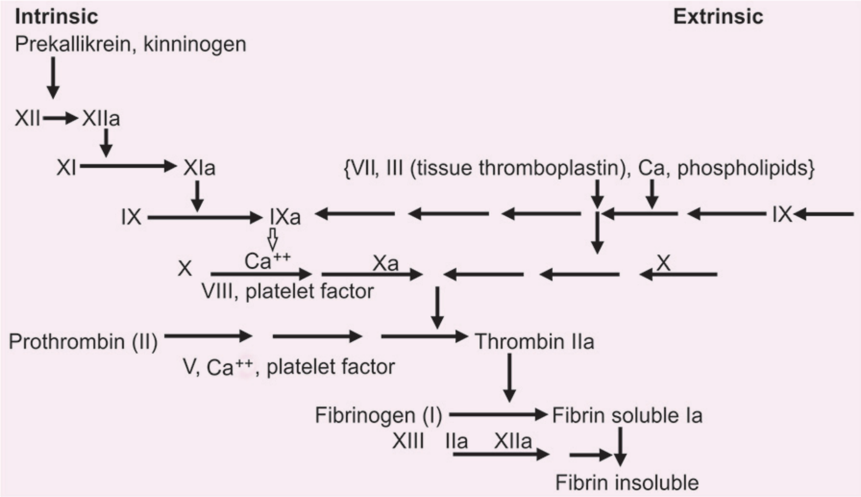

Treatment of DIC

- *Treatment of specific cause* as per protocol. Correction of haemodynamic instability by fluid therapy, transfusion of packed cells or whole blood. Dopamine/dobutamine therapy.
- *Factor replacement—specific therapy* for DIC—FFP, cryoprecipitate, platelet concentrate transfusions are essential. FFP is given at a dose of 15 ml/kg. Cryoprecipitate is used to raise fibrinogen level at a dose to make plasma fibrinogen level 150 mg/dl. Platelet is transfused at a dose of 0.1 unit/kg, when platelet drops below 50,000 with episodes of bleeding.
- Heparin use is often controversial. It is used *mainly* in chronic DIC, DIC with purpura, DIC of obstetric cause, cancer induced DIC, DIC due to acute antiphospholipid antibody syndrome.
- EACA, tranexamic acid can be used but with questionable benefits.

MECHANISM OF BLOOD COAGULATION (HAEMOSTASIS)

Haemostasis is the spontaneous arrest of bleeding. When an injury occurs ***platelet adhesion*** occurs to injured vessel/capillary wall which activate the release of ADP (Adenosine diphosphate) which makes more platelet to aggregate (***platelet aggregation***). These activated platelets release thromboxane A_2 which further increases the adhesion and aggregation of platelets. Circulating fibrinogen binds to an activated platelet receptors glycoprotein IIb and IIIa and fibrinogen gets converted into fibrin.

Clotting factors are proteins synthesized by the liver which with a series of cascade reaction activates clotting factors and achieves blood coagulation by a complex mechanism. Factor II, VII, IX and X are vitamin K dependent for their synthesis in liver (carboxylation of glutamic acid). In the process of coagulation each factor gets activated to an enzyme by partial proteolysis which in turn activates other needed coagulation factors. Eventually fibrinogen gets converted into soluble fibrin and later into insoluble ***fibrin.***

Two types coagulation system are there:

- *Intrinsic pathway*
- *Extrinsic pathway*

In vitro coagulation occurs by intrinsic coagulation system. Cascade gets activated by vessel wall injury, shear stress of vessel or other factors. It activates the cascade to get final result.

Coagulation Cascade System

Blood clotting factors

Factor No.	*Common name*
I	Fibrinogen
II	Prothrombin
III	Thromboplastin
IV	Ionic calcium
V	Hereditary labile factor, activator (AC) globulin, proaccelerin
VI	Accelerin,supposed to be active form of Factor V
VII	Proconvertin; serum prothrombin—conversion accelerator (SPCA)
VIII	Antihemophilic factor (AHF)
IX	Plasma thromboplastin component (PTC; Christmas factor)
X	Stuart-Prower factor
XI	Plasma thromboplastin antecedent (PTA)
XII	Hageman factor
XIII	Fibrin stabilizing factor, fibrinase
XIV	Prekallikrein
XV	Kallikrein
XVI	Platelet factor

Hemophilia and *von Willebrand's disease* are the two most common inherited bleeding disorders due to deficiency of factor VIII.

Factor VIII has two components; smaller one—factor VIII C is needed for activation of factor X in intrinsic coagulation pathway; its deficiency leads to *classic hemophilia*. It is inherited as an X-linked recessive trait, thus it occurs in males and homozygous females. The larger component of factor VIII called *von Willebrand's factor,* facilitates the adhesion of platelets to subendothelial collagen, hence crucial for hemostasis, its absence leads to *von Willebrand's disease.*

Classic haemophilia is called haemophilia A is caused by deficiency of factor *VIII C* with X linked *recessive* trait. It occurs in males or homozygous females. Recurrent haemarthroses is a common presentation. Petechiae and echymoses are absent. Bleeding time is normal but coagulation time is prolonged. Treatment is replacement of factor VIII haemophilic factor. If not known patient may go for a life-threatening bleeding even after dental extraction.

Von Willebrand's disease is deficiency of larger component (99%) of the factor VIII-vWF. It is an autosomal *dominant* disease with normal bleeding time and normal platelet count. Common presentations are spontaneous bleeding from mucous membrane, excessive bleeding from wounds and severe menorrhagia. Haemarthroses is not common in von Willebrand's disease. Treatment is replacement of specific factors.

Haemophilia B also called as *Christmas disease* is due to factor IX deficiency is inherited as X linked autosomal recessive trait.

Action, to be effective, must be directed to clearly conceived ends.

I. Burns

> *The infusion of large quantities of fluid, mainly a salt solution, as many as 3-4 litres daily, has proved satisfactory. Sometimes a life is saved if not more than one-quarter of the body surface has received a third-degree burn…. I have given blood transfusions to two patients with severe burns, with recovery in one case and postponement of death in the other.*
>
> **—Gustav Riehl,** 1925

CHAPTER OUTLINE

- Burns
- Management of Burns
- Eschar
- Contracture in Burn Wound
- Electrical Burns
- Inhalation Injury
- Chemical Burns

BURNS

Types of burns

- Thermal injury
  - Scald—spillage of hot liquids
  - Flame burns
  - Flash burns due to exposure of natural gas, alcohol, combustible liquids
  - Contact burns—contact with hot metals/objects/materials
- Electrical injury
- Chemical burns—acid/alkali
- Cold injury—frost bite
- Ionising radiation
- Sun burns

Classification of Burns

Depending on the Percentage of Burns

Mild (Minor):

- Partial thickness burns < 15% in adult or <10% in children.
- Full thickness burns less than 2%.
- Can be treated on outpatient basis.

Moderate:

- Second degree of 15-25% burns (10-20% in children).
- Third degree between 2-10% burns.
- Burns which are not involving eyes, ears, face, hand, feet, perineum.

Major (severe):

- Second degree burns more than 25% in adults, in children more than 20%.
- All third degree burns of 10% or more.
- Burns involving eyes, ears, feet, hands, perineum.
- All inhalation and electrical burns.
- Burns with fractures or major mechanical trauma.

I. Depending on thickness of skin involved

a. ***First degree***: Here the epidermis looks red and painful, no blisters, heals rapidly in 5-7 days by epithelialization without scarring.

Rule of Nine (Wallace's rule of "9")

	Adults	***Children***	***Infants***
Head and neck	9%	18%	20%
Front of chest and abdominal wall	9 × 2 = 18%	18%	10 × 2 = 20%
Back of chest and abdominal wall	9 × 2 = 18%	18%	10 × 2 = 20%
Lower limb	18 × 2 = 36%	13.5 × 2 = 27%	10 × 2 = 20%
Upper limb	9 × 2 = 18%	18%	10 × 2 = 20%
Perineum	01%	01%	
	100%	**100%**	**100%**

Note: It is head and lower limb percentage which differ in adults and children.

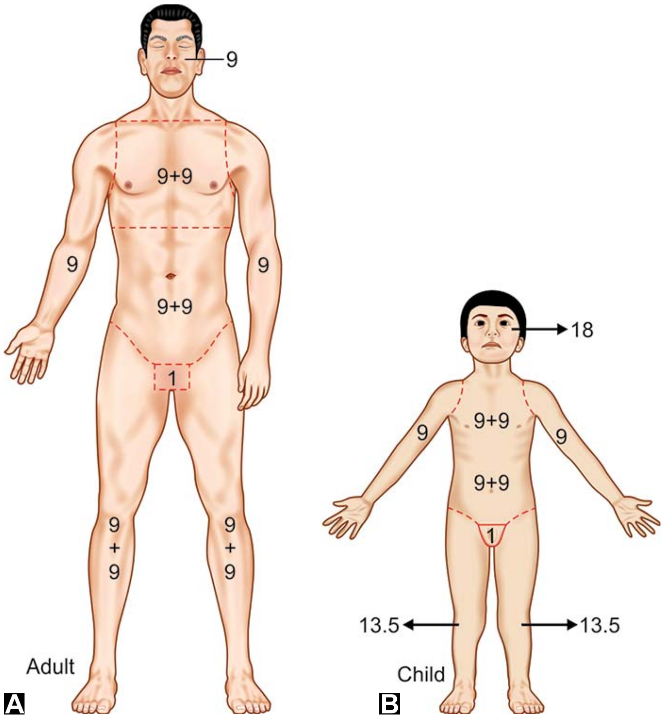
9
9+9
9
9
9+9
1
9
+
9
9
+
9
Adult
A
18
9+9
9
9
9+9
1
13.5
13.5
Child
B

Figs 1.234A and B: Percentage of burns in (A) Adults; (B) Children

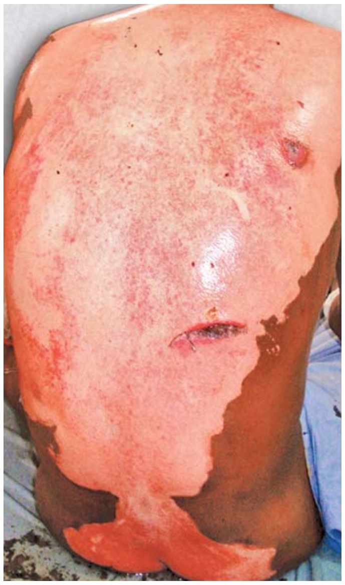

Fig. 1.235: First degree burn.

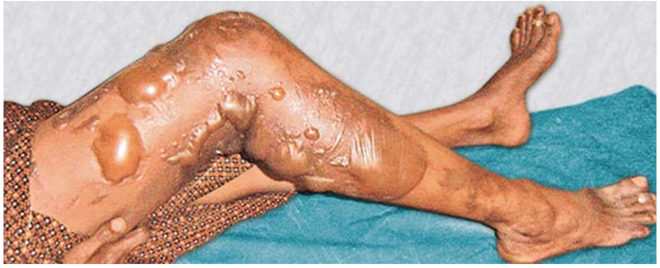

Fig. 1.236: Second degree burns with blisters.

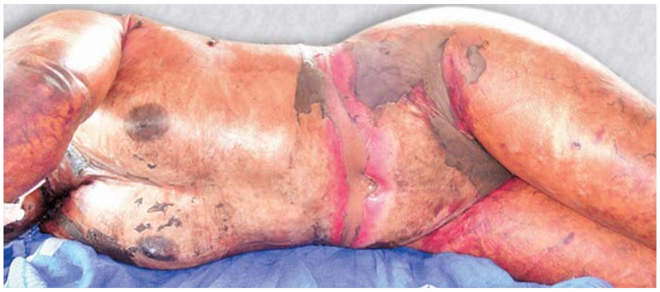

Fig. 1.237: Extensive third degree burns with eschar.

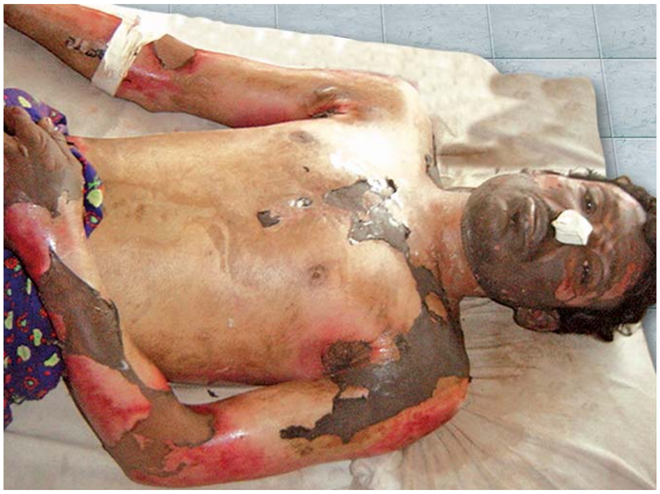

Fig. 1.238: Extensive burns more than 50%.

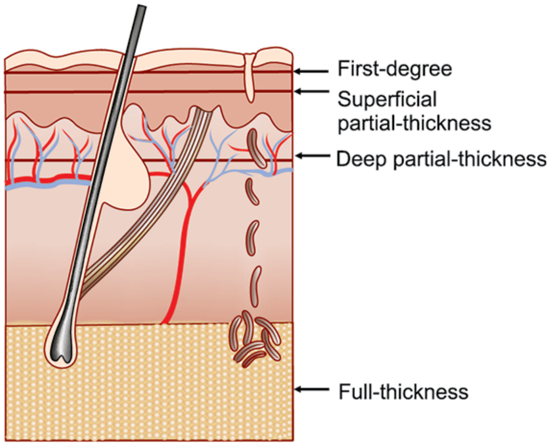
First-degree
Superficial partial-thickness
Deep partial-thickness
Full-thickness

Fig. 1.239: Degrees of burns.

b. ***Second degree***: The affected area is mottled, red, painful, with blisters, heals by epithelialisation in 14-21 days.
  - Superficial second degree burn heals, causing pigmentation.
  - Deep second degree burn heals, causing scarring, and pigmentation.

c. ***Third degree***: The affected area is charred, parchment like, painless and insensitive, with thrombosis of superficial vessels. It requires grafting. Charred, denatured, insensitive, contracted full thickness burn is called as **eschar.** These wound must heal by re-epithelialisation from wound edge.

d. ***Fourth degree***: Involves the underlying tissues—muscles, bones.

Success lies not in achieving what you aim at, but in aiming at what you ought to achieve.

II. Depending on thickness of skin involved

a. ***Partial thickness burns:*** It is either first or second degree burn which is red and painful, often with blisters.

b. ***Full thickness burns:*** It is third degree burns which is charred, insensitive, deep involving all layers of the skin.

Assessment of Burns

- Wallace's rule of nine is used for early assessment—refer figure.
- *Using the Lund and Browder chart* is better method for assessing the burns wound. Here each part of the body is individually assessed for involvement of burns.
- Patient's *entire hand area* is 1%. Clean piece of paper is cut to the size of hand and through that percentage of burns is assessed.

44°C temperature takes 6 hours to cause deep burns. 65°C takes 45-60 seconds to cause deep full thickness burn.

Clinical Features

- History of burn.
- Pain, burning, anxious status, tachycardia, tachypnoea, fluid loss.
- In severe degrees features of shock.

Tolerable temperature to human skin is 40°C for brief period.

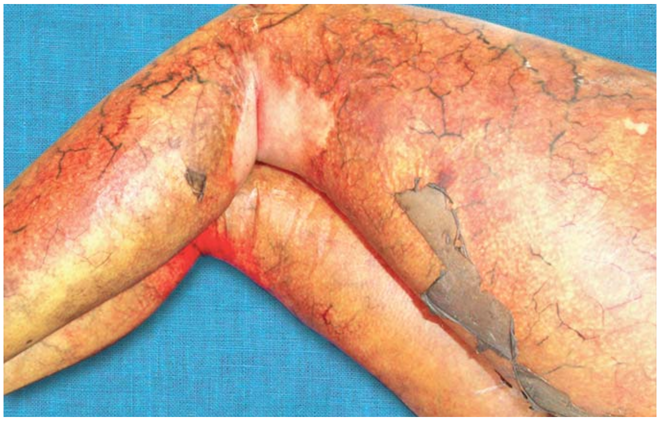

Fig. 1.240: 3rd degree burns with charred skin, with thrombosed veins and eschar.

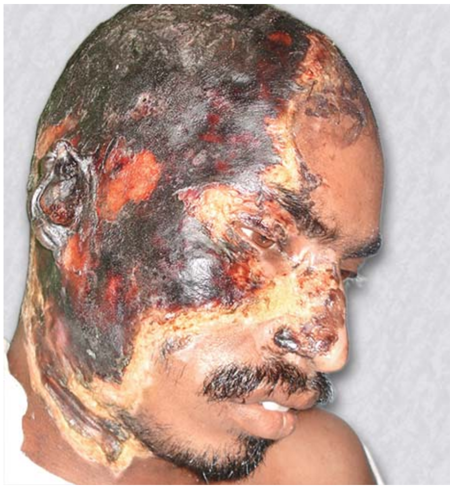

Fig. 1.241: Face burns in an adult—severe 3rd degree.

Pathophysiology

Heat causes coagulation necrosis of skin and subcutaneous tissue

↓

Release of vasoactive peptides

↓

Altered capillary permeability

↓

Loss of fluid → **Severe hypovolaemia**

↓

Decreased cardiac output → Decreased myocardial function

↓

Decreased renal blood flow (Renal failure) → Oliguria

↓

Altered pulmonary resistance causing pulmonary oedema

↓

Infection

↓

Systemic inflammatory response syndrome (SIRS)

↓

Multiorgan dysfunction syndrome (MODS).

There is increased capillary permeability, decreased plasma oncotic pressure causing loss of protein and fluid from intravascular space. Vasoconstriction occurs due to raised capillary hydrostatic pressure leading into cellular aggregation. Blockage of lymphatics causes poor clearance of fluid and proteins from interstitial spaces. Cell membrane function is impaired causing intracellular fluid accumulation. Activation and release of various complement factors, histamine, and prostaglandins results in myocardial dysfunction, oedema of tissues, reduced immunoglobulin synthesis. Catecholamine levels are raised drastically in patient with burn. There will be lipolysis, proteolysis, increased release of glutamine and alanine from skeletal muscles. Urea production is increased due to more proteolysis.

Massive oedema in the body is due to altered pressure gradient because of the injury to basement membrane.

Cardiac dysfunction is due to:

- Hypovolaemia.
- Release of cardiac depressants.
- Hormonal causes like catecholamines, vasopressin, angiotensins.

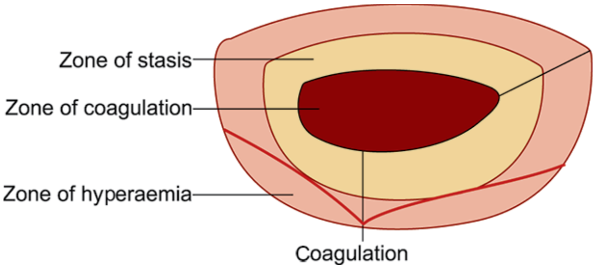

Fig. 1.242: Burn zones (Jackson's).

Renal changes are due to:

- Release of ADH from posterior pituitary to cause maximum water reabsorption.
- Release of aldosterone from adrenals to cause maximum sodium reabsorption.
- Toxins released from the wound along with sepsis causes acute tubular necrosis.
- Myoglobin released from muscles (in case of electric injury or often from eschar) is most injurious to kidneys.

Pulmonary changes are due to:

- Altered ventilation-perfusion ratio.
- Pulmonary oedema due to burn injury, fluid overload, inhalation injury.
- ARDS.
- Aspiration.
- Septicaemia.

Assessment of airway injury is important in burns

- Occurs in burns around face and neck, or trapped in burning room
- Presents with hoarseness or stridor
- Inhaled burning gases can cause upper airway burns and laryngeal oedema
- Smoke inhalation can cause chemical alveolitis, pulmonary oedema, ARDS and respiratory failure
- Steam inhalation can cause damage to respiratory epithelium and subglottic oedema
- Carbon monoxide inhalation more than 10% is dangerous as it forms carboxyhaemoglobin (CO has got 240 times more affinity to haemoglobin than oxygen) which blocks oxygen transport completely causing respiratory arrest, hypoxia and metabolic acidosis
- Chest wall burn causes mechanical block of ventilation—needs escharotomy in chest wall and procedure is painless
- Airway burn may require early elective intubation or tracheostomy or emergency cricothyroidotomy as a life-saving method

GIT changes are due to:

- *Acute gastric dilatation* which occurs in 2-4 days.
- Paralytic ileus.
- Curling's ulcer.
- Cholestasis and hepatic damage.
- Acute acalculous cholecystitis, acute pancreatitis can occur.

Metabolic Changes

- Hypermetabolic rate (BMR).
- Negative nitrogen balance.
- Electrolyte imbalance.
- Deficiencies of vitamins and essential elements.
- Metabolic acidosis due to hypoxia and lactic acid.

Sepsis in Burn Patient

- Focus may be at the burn site, catheter site, cannula/CVP line site, or respiratory infection.
- Low immunity, loss of proteins and immunoglobulins, loss of barrier causes sepsis. Opportunistic infection is also common.
- Associated conditions like diabetes, HIV infection, old age, respiratory diseases worsen the sepsis in burn injury.
- It may be *local infection commonly* by *Staphylococcus aureus* in early period, *Pseudomonas*, *Candida*, *Aspergillus*, herpes simplex virus in partial thickness nasolabial burns. It may be suppurative thrombophlebitis also.
- *Systemic infection* like pneumonia, bacteraemia, septicaemia can occur.
- Burns itself *creates* immunosuppression.
- Sepsis is identified by fever, lethargy, *leukocytosis, thrombocytopenia*.

Infections are commonly due to:

- Streptococci (Beta haemolytic—most common)
- *Pseudomonas*
- *Staphylococci*
- Other gram-negative organisms
- *Candida albicans*

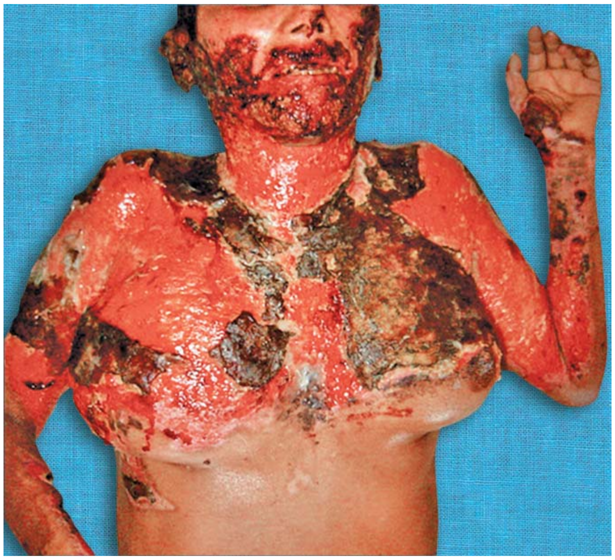

Fig. 1.243: Burns over chest wall with eschar. Note the extent of involvement.

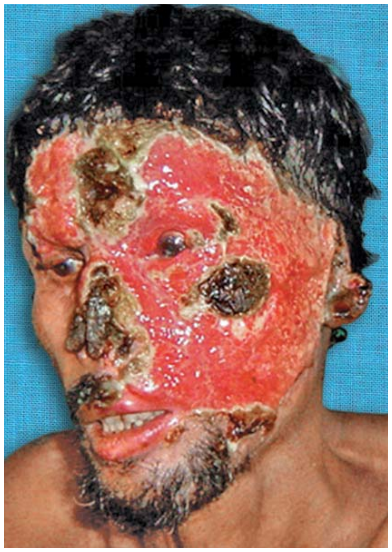

Fig. 1.244: Deep burns face with eyelid involvement.

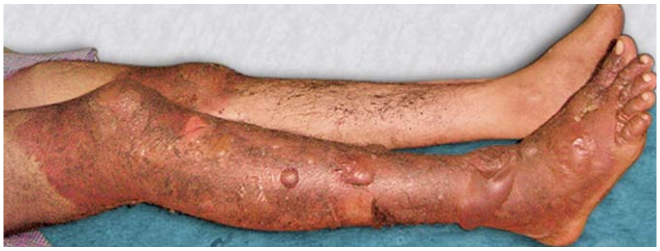

Fig. 1.245: Second degree burns in the leg with blebs.

Effects of Burn Injury

- Shock due to hypovolaemia.
- Renal failure.
- Pulmonary oedema, respiratory infection, adult respiratory distress syndrome (ARDS), respiratory failure.
- Infection by *Staphylococcus aureus*, beta haemolytic *Streptococcus, Pseudomonas, Klebsiella* leads to bacteraemia, septicaemia. Fungal and viral infections of dangerous types can also occur.
- GIT: Hypovolaemia, ischaemia of mucosa, erosive gastritis—**Curling's ulcer (seen in burns > 35%).**
- Fluid and electrolyte imbalance.
- Postburn immunosuppression predisposes to severe opportunistic infection.
- Eschar formation and its problems like defective circulation, ischaemia when it is circumferential.
- Electrical injuries often cause fractures, major internal organ injury, convulsions.
- Development of contracture is a late problem. It leads to ectropion, microstomia, disability of different joints, defective hand functions, growth retardation causing shortening.
- Inhalation burn causes pulmonary oedema, respiratory arrest, ARDS.
- Chemical injury causes severe GIT disturbances like erosions, perforation, stricture oesophagus (**alkali**), pyloric stenosis (**acid**), mediastinal injury.
- Other problems commonly seen are DVT, pulmonary embolism, urinary infection, bed-sores, severe malnutrition with catabolic status, respiratory infection.
- Complications of burns contracture itself like hypertrophic scar, keloid formation.
- **Toxic shock syndrome:** It is a life-threatening exotoxin mediated disease caused by *Staphylococcus aureus*. It is common in children, presents with rashes, myalgia, diarrhoea, vomiting, and multiorgan failure with high mortality.

Causes of death in burns

- Hypovolaemia (refractory and uncontrolled) and shock
- Renal failure
- Pulmonary oedema and ARDS
- Septicaemia
- Multiorgan failure
- Acute airway block in head and neck burns

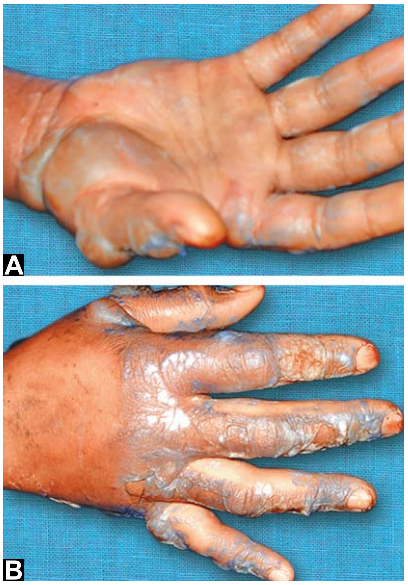

Figs 1.246A and B: Burns hand both palmar and dorsal aspect.

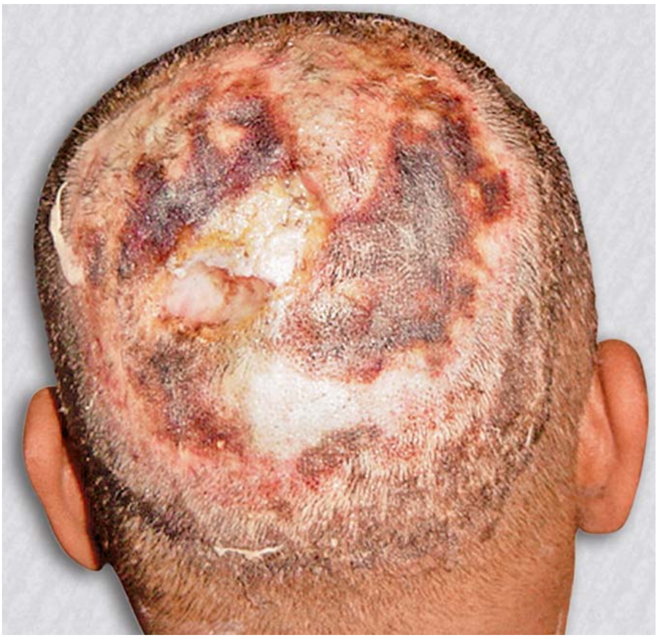

Fig. 1.247: Deep burn in the scalp involving extensively.

MANAGEMENT OF BURNS

First Aid

- Stop the burning process and keep the patient away from the burning area.
- Cool the area with tap water by continuous irrigation for 20 minutes (not cold water as it can cause hypothermia).

Indications for admission in burns

- Any moderate and severe burns
- Airway burns of any type
- Burns in extremes of age
- All electrical/deep chemical burns

Initial management

- Clothing should be removed
- Cooling of the part by running water for 20 minutes
- Cleaning the part to remove dust, mud, etc
- Chemoprophylaxis—tetanus toxoid; antibiotics; local antiseptics
- Covering with dressings by different methods
- Comforting with sedation and pain killer

Definitive Treatment

- Admit the patient.
- Maintain airway, breathing, circulation.
- Assess the percentage, degree, and type of burn.
- Keep the patient in a clean environment.
- Sedation and proper analgesia.
- Patient should be in ***burns unit*** (ideally air-conditioned) with barrier nursing, sterile clothes, bed sheets with all aseptic methods.

Fluid Resuscitation

Formulas to calculate the fluid replacement:

a. **Parkland regime: Commonly used:**
4 ml/% burn/kg body weight/24 hours.
Maximum percentage considered is 50%.
Half the volume is given in first 8 hours, rest given in 16 hours.

b. **Muir and Burclay regime:**

$$\frac{\%\text{ Burns} \times \text{Body weight in kg}}{2} = 1 \text{ Ration}$$

3 Rations given in first 12 hours.
2 Rations in second 12 hours.
1 Ration in third 12 hours.

c. **Galveston regime (pediatric):**
5000 ml/m^2 burned + 1500 ml/m^2 total

d. **Modified Brooke formula:**
First 24 hours:
RL: 4 ml/kg/% burns in 24 hours
(first half in first 8 hours)
Colloid—none.
Second 24 hours:
Crystalloids—to maintain urine output
Colloids—0.3 ml to 0.5 ml/kg/burns in 24 hours.
(Albumin in RL solution) (Albumin alone should be given with care if really indicated only).

e. **Evan's formula:**
In first 24 hours:
Normal saline 1 ml/kg/% burns
Colloids 1 ml/kg/% burns
5 % dextrose in water, 2000 ml in adult.
In second 24 hours:
Half of the volume used in first 24 hours.

Fluids used are normal saline, ringer lactate, Hartmann fluid, plasma. Ringer lactate is the fluid of choice. Blood is transfused in later period (after 48 hours).

First 24 hours only *crystalloids* should be given (Crystalloids are one which can pass through capillary wall like saline either hypo, iso or hypertonic, dextrose saline, Ringer lactate).

Sodium is assessed by formula: 0.52 mmol × kg body weight × % body burns, given at a rate of 4.0 to 4.4 ml/kg/hour.

After 24 hours up to 30-48 hours, *colloids* should be given to compensate plasma loss (colloids are one which are retained in intravascular compartment). Plasma, haemaccel (gelatin), dextrans, hetastarch are used. Usually at a rate of 0.35-0.5 ml/kg/% burns is used in 24 hours.

- Urinary catheterization to monitor output; 30-50 ml/hour should be the urine output.
- Tetanus toxoid.
- Monitoring the patient: Hourly pulse, BP, PO_2, PCO_2, electrolyte analysis, blood urea, nasal oxygen, often intubation is required.
- IV ranitidine 50 mg 8th hourly.
- **Ryle's tube insertion initially for aspiration purpose later for feeding** *(Enteral feeding).*
- Antibiotics: Penicillins, aminoglycosides, cephalosporins, metronidazole.
- Culture of the discharge; total white cell count and platelet count at regular intervals are essential to identify the sepsis along with fever, tachycardia and tachypnoea.
- In burns of oral cavity tracheostomy may be required to maintain the airway.
- **Total parenteral nutrition (TPN)** is required for faster recovery, using carbohydrates, lipids, vitamins ***(through a CVP line).***
- Tracheostomy/intubation tube may be required in impending respiratory failure or upper airway block.
- Intensive nursing care.

Local Management

- Dressing at regular intervals under general anaesthesia using paraffin gauze, hydrocolloids, plastic films, vaseline impregnated gauze or fenestrated silicone sheet or biological dressings like amniotic membrane or synthetic biobrane.
- **Open method** with application of silver sulfadiazine without any dressings, used commonly in burns of face, head and neck.
- **Closed method** is with dressings done to soothen and to protect the wound, to reduce the pain, as an absorbent.
- **Tangential excision** of burn wound with skin grafting can be done within 48 hours in patients with less than 25% burns. It is usually done in deep dermal burn wherein dead dermis is removed layer by layer until fresh bleeding occurs. Later skin grafting is done.
  Advantages of tangential excision: *It reduces*—the chance of secondary infection, the hospital stay, and formation of hypertrophic scar or contracture, the cost.
- In burns of head and neck region, exposure treatment is advised.
- Slough excision is done regularly.
- **After cleaning with povidone iodine solution silver sulfadiazine** ointment is used. It is an antiseptic and soothening agent. It causes neutropenia.

If anything is sacred, the human body is sacred.— **Walt Whitman**

Topical agent	*Advantages*	*Problems*
Silver sulfadiazine (1%)	• Antiseptic (G +ve and G –ve)	• Neutropenia, pseudoeschar • It diffuses poorly through the eschar
	• Soothening, good penetration, • Hydration and softening of eschar occurs	• Causes hypertrophic granulation and so not used once wound is clean and eschar is separated
	• It does not cause pain	• Causes wound maceration
		• It should be avoided in pregnancy, nursing mother and infants below 2 years of age
Sulfamylon (5%/11%)	• Antipseudomonal, anticlostridial	• Very irritant, painful
(Mafenide acetate)	• Penetrates very well in to tissues • It is water soluble • It is good antipseudomonal agent	• Causes acidosis, by inhibiting the carbonic anhydrase
Silver nitrate (0.5%)	• Antiseptic	• Stains burnt area
Povidone iodine (5%)	• Antiseptic 0 (G +ve, G –ve)	• Irritant and painful
	• Used on granulation tissue after eschar separation	• Not used in partial burns
	• Used in perineum and buttocks to prevent maceration	
Silver sulphadiazine and cerium nitrate	• Boosts cell mediated immunity and forms sterile eschar	

Other agents used are *Sulfamylon* (***Mafenide acetate***) and *Silver nitrate*.

- ***Sulfamylon*** is antipseudomonal and anticlostridial agent. It penetrates well into the tissues but it is very irritant. It causes acidosis.
- ***Silver nitrate*** causes staining of burnt area.
- 0.025% sodium hypochlorite (Dakin's solution) is effective against Gram +ve organisms; 0.25% acetic acid is effective against Gram – ve organisms, but both mildly inhibit epithelialisation.

Regular culture and sensitivity for bacteria is required, to see for streptococcal growth which should be less than *1,00,000 (10^5) per gram of tissues.*

Wound Coverage

- Once the area granulates well, in 3 weeks usually, split skin grafting is done (SSG, Thiersch graft).
- For wider area MESH split skin graft is used.
- If there is eschar, *escharotomy* is required to prevent compression of vessels.
- In certain areas like face and ear, full thickness graft (Wolfe graft) or flap is required.
- **Cultured skin:** Full thickness skin biopsy of patient's skin is done immediately after admission. By specialized culture technology sheets of skin can be manufactured in 3 weeks as cultured epithelial grafts. It can cover skin of almost entire body. It is usually useful in burns of > 80%. Take up of cultured graft is 60-75%. *Limitations are*—time taken to develop cultured graft; more vulnerability for mechanical trauma; costly; time taken to manufacture; scarring.

Synthetic dressings in burn wound

- Vaseline impregnated gauze dressing prevents stiffness of eschar.
- Hydrocolloid dressing (*duoderm*) helps moist environment, proper epithelialisation. It is useful in mixed deep burns. It is changed once in 3 days.
- *Opsite* is less expensive, with less pain, creates moist barrier. But it does not have antimicrobial effect and it causes accumulation of exudates.
- *Biobrane* is collagen coated silicone sheet which gets adherent to wound acting as barrier without any pain. But it does not have antimicrobial effect and it causes accumulation of exudates. It is used for 2nd degree burns.
- *Transcyte* has similar features of biobrane. It contains growth factor derived from cultured fibroblasts which promotes wound healing.
- *Integra* contains deeper collagen matrix as dermal substitute; outer silicone sheet as epidermal substitute. Inner collagen matrix acts as dermis whereas outer silicone sheet is removed 2 weeks after dressing and additional autograft should be placed. It provides complete wound cover. Scarring after healing is reduced significantly.

Biologic dressings for burn wound

It is used to cover the wound temporarily as a barrier and also to have some immunologic function. Eventually graft will slough. Later wound is covered with auto-skin graft. It is used for massive burn injuries more than 50%. Possible problem is transmission of viral diseases.

Xenograft is of pig skin. *Allograft* is of cadaver skin (homograft)—it gives all existing normal skin function for temporary period. It may leave a dermal equivalent in the wound later.

ESCHAR

- It is charred, denatured, full thickness, deep burns with contracted dermis.

- It is insensitive, with thrombosed superficial veins.
- **Circumferential eschar** in the upper limb, lower limb, neck, thorax can cause more oedema which initially causes venous compression and later arterial compression causing ischaemia, gangrene of the distal part. So distal area should be monitored for circulation.
- If required deep longitudinal full thickness incisions are made in different areas so as to prevent collection of oedema fluid and also to prevent compression over the vessels. This is called as **escharotomy**. Escharotomy causes large quantity of blood loss and so blood transfusion is needed while doing escharotomy. Incision should be of adequate length and depth during escharotomy. It should be placed in such a way so as to avoid injury to major neurovascular system. Release of muscle compartment is needed often in these patients.
- Multiple incisions or incisions over the joints may be needed.
- Early rapid separation of eschar indicates severe sepsis underneath.
- Eventually eschar should be excised and the area is allowed to granulate and skin grafting should be done.
- **Pseudoeschar** is thickened burnt skin due to repeated silver sulphadiazine application.

CONTRACTURE IN BURN WOUND

Contracture in burns can occur anywhere. It is more common wherein flexibility and mobility is present like along the joint, eyelids, cheeks, lips, neck, elbow, knee, etc. Contracture can be *intrinsic* by loss of tissue or *extrinsic* by pull during healing phase contraction. Contracture proceeds towards position of comfort until it meets or closely reaches opposite surface. There is clearly wound shortening. Disorganised over formation of compact collagen (3 times normal) causes hypertrophic scar leading further contracture.

Deficit of neck extension is graded, normal > 110°; E_1 95-110°; E_2 is 85-95°; E_3 is < 85° with mentosternal synechia.

Classification of Burns Contracture in the Neck (BM Achauer)

- Mild (less than 1/3rd)—inability to see ceiling.
- Moderate (1/3rd to 2/3rd)—flexion is possible but not extension.
- Severe (more than 2/3rd)—fully contracted in flexed position with pull on lower lip.
- Extensive—contraction is extensive with mentosternal adhesions.

Ifeanyichukwu has classified neck contracture into: Type 1—mild anterior with narrow contracting band less than fingerbreadth (1a) or broad band (1b); *type 2*—moderate anterior with narrow band (2a) or broad band (2b); *type 3*—severe anterior mentosternal adhesion with supple neck skin (3a) or without supple skin (3b); *type 4*—posterior with narrow band (4a) or multiple or broad band (4b).

Reconstruction territories in neck in burn contracture based on functional benefits are—central above; central below; central above and below; lateral.

Complications of Burns Contracture

- Ectropion of eyelid causing keratitis and corneal ulcer.
- Disfigurement in face.
- Narrowing of mouth **microstomia.**
- Contracture in the neck causing restricted neck movements.
- Disability and nonfunctioning of joints due to contracture.
- Hypertrophic scar and keloid formation.
- Repeated breaking of scar and infection, ulcer, cellulitis.
- Pain and tenderness in the scar contracture.
- ***Marjolin's ulcer:*** It is a very well-differentiated squamous cell carcinoma occurring in a scar ulcer due to repeated breakdown *(unstable scar of long duration).*
  - It is locally malignant.
  - As there are no lymphatics in the scar, so there is no spread to lymph nodes.
  - As there are no nerves in the scar it is painless.
  - It has raised and everted edge with induration.
  - Biopsy confirms the diagnosis.

Treatment: *Radiotherapy is not given* for Marjolin's ulcer. Treatment is either *wide excision* or *amputation. It is curable.*

Once it spreads out of the scar tissue it behaves like any other squamous cell carcinoma and so can spread to regional lymph nodes.

Treatment for Contracture

- Release of contracture surgically and use of skin graft or "Z" plasty or different flaps. *Different flaps* used are—transposition flaps, vertical or transverse; laterally based flap; bilobed flap; bipedicled flap; advancement flap; regional flap; random cutaneous flap (*Epaulette flap, Charretera flap*); fasciocutaneous/myocutaneous flap; tube flap; expanded skin flap; combined skin graft and flap; microvascular free flap.
- Proper physiotherapy and rehabilitation is essential.
- Pressure garments to prevent hypertrophic scars.
- Management of itching in the scar using aloe vera, antihistamines and moisturizing creams.

Problems in Managing Burn Contracture

- Giving proper anaesthesia is challenging.

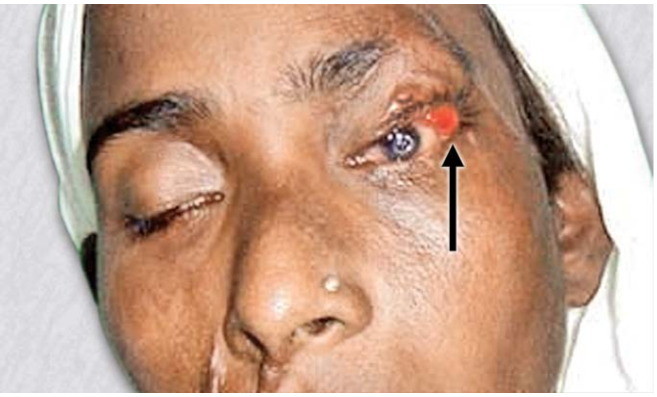

Fig. 1.248: Ectropion of upper eyelid due to burns.

Complete and lasting freedom from disease is but a dream remembered from imaginings of garden of Eden.
— Reni Jules Dubos

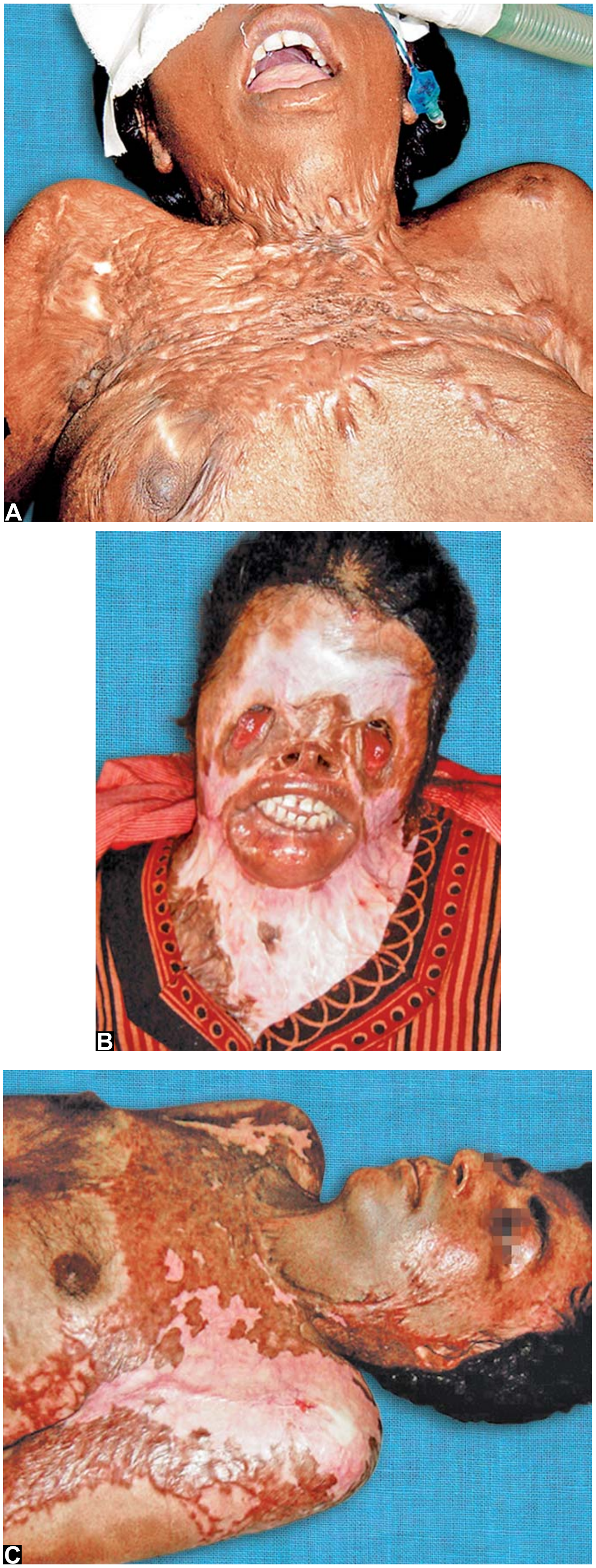

Figs 1.249A to C: Contracture at different parts of the body—chest, face and neck.

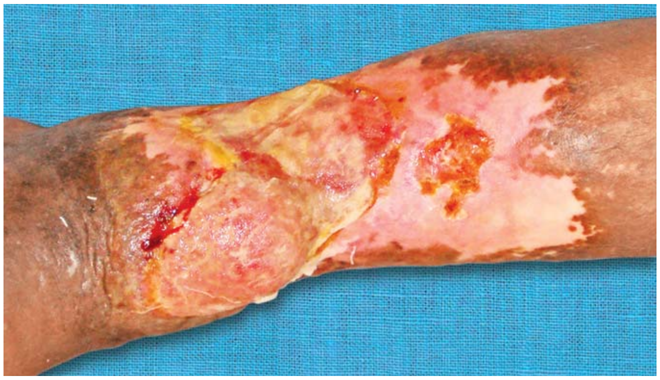

Fig. 1.250: Marjolin's ulcer developed over burns scar.

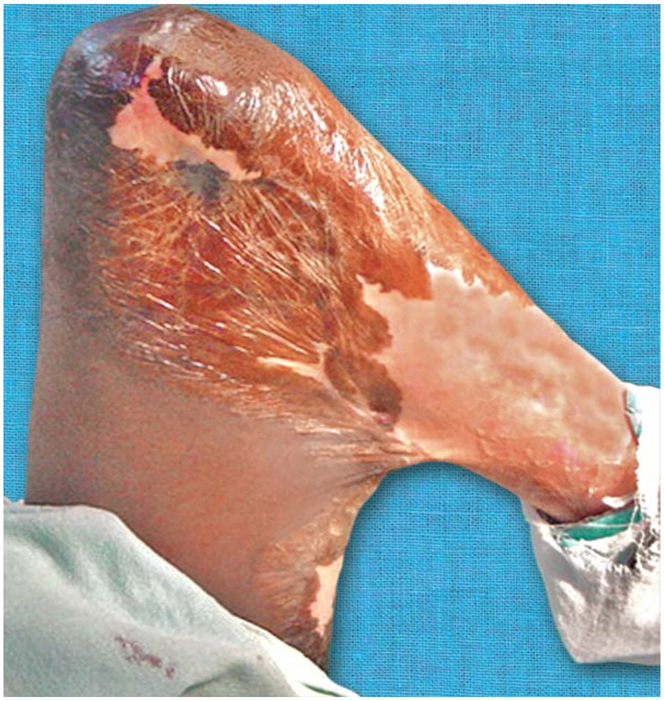

Fig. 1.251: Severe contracture at knee joint causing deformity.

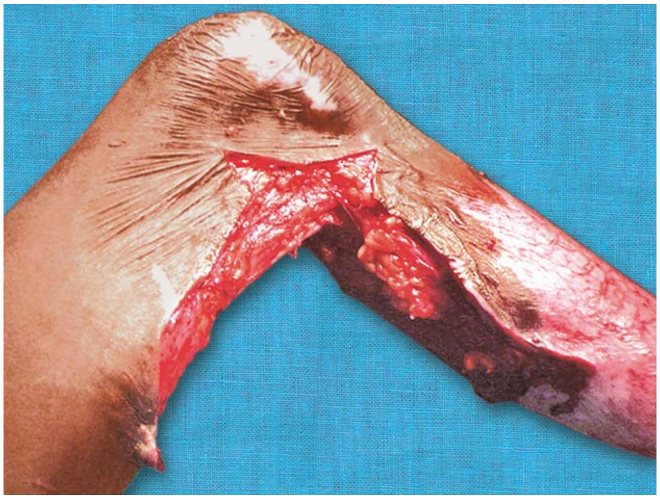

Fig. 1.252: Same patient with knee contracture undergoing release surgery.

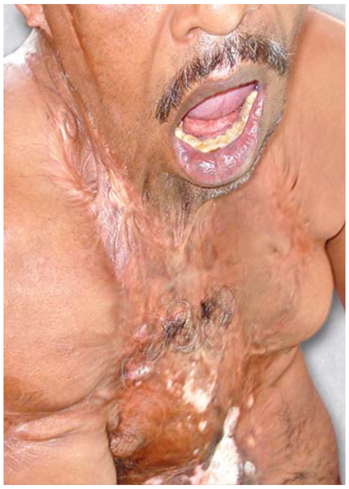

Fig. 1.253: Burn contractures over neck and chest wall causing cosmetic problem.

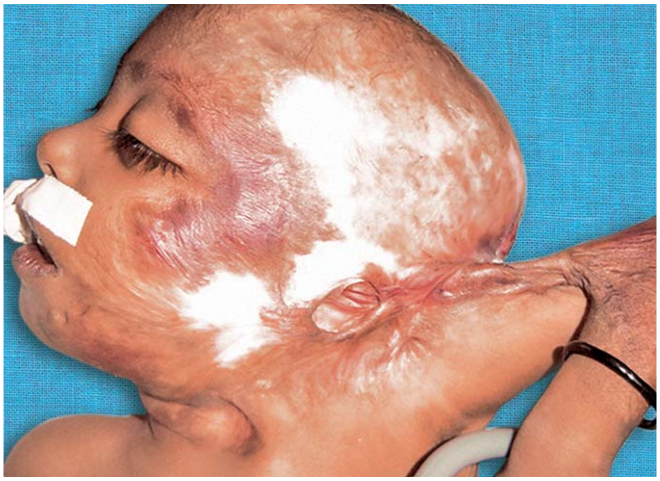

Fig. 1.254: Infant had burns in head and neck region causing severe contracture.

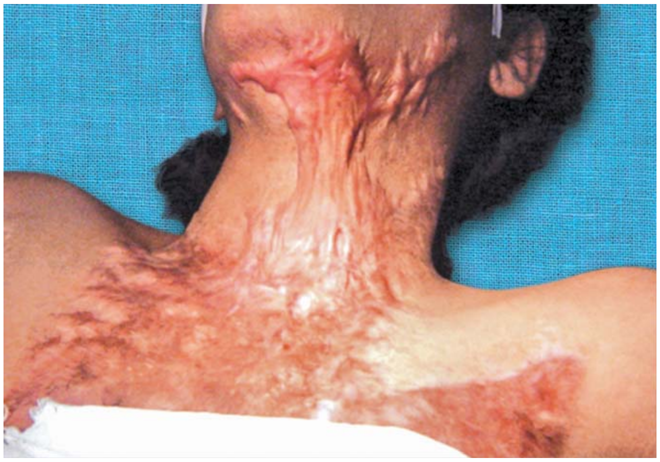

Fig. 1.255: Hypertrophic burn scar with contracture neck in a female.

- Scar excision can cause significant bleeding.
- Identifying major structures in the area and safeguarding vascular and others is often worrisome.

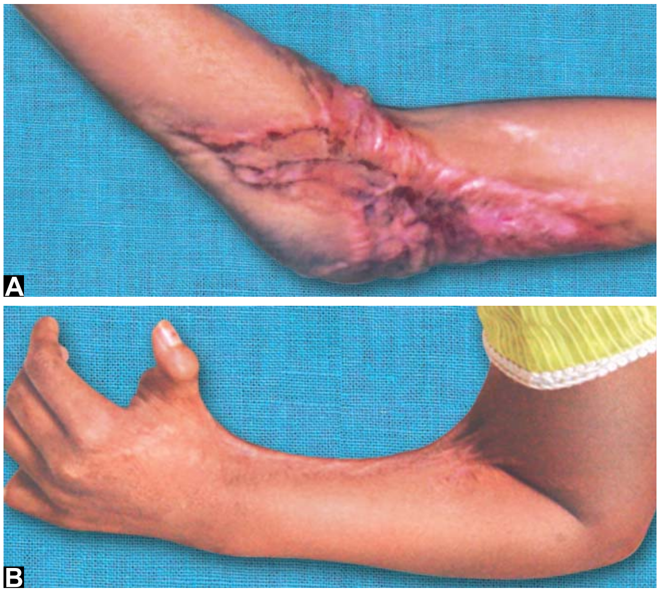

Figs 1.256A and B: Elbow contracture due to burns.

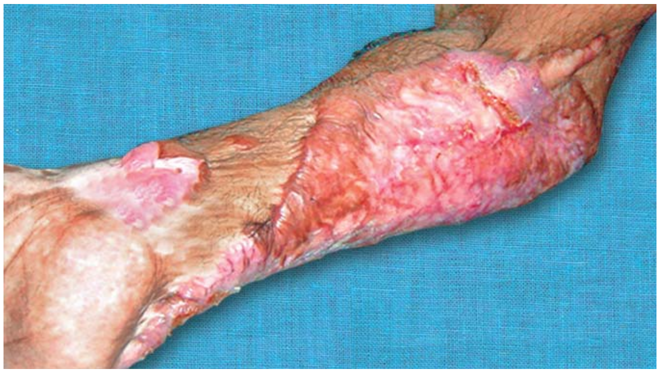

Fig. 1.257: Keloid in the hand after burn injury.

- Need for repeated surgeries as staged one.
- Maintaining the position with skeletal traction, fixation, collar, POP cast, etc.
- Psychological problems and needs counseling.
- Prolonged hospital stay, cost factors.

Prevention of development of contracture

- Joint exercise in full range during recovery period of burns
- Pressure garments for a long period
- Topical silicon sheeting
- Saline expanders for scars

ELECTRICAL BURNS

Types of Electrical Injury

- *Low tension injury:* Less than 1000 volts.
- *High tension injury*: More than 1000 volts—may be due to current itself causing intense damage on the tissues up to 2000°C; flash injury due to electrical arc up to 4000°C; flame injury by catching of fire to the clothing and body; traumatic injury like fractures and internal organ injuries.
- It is always a deep burn (***always a major burn***).
- There is a *wound of entry and wound of exit*.

There is no right way to do something wrong.

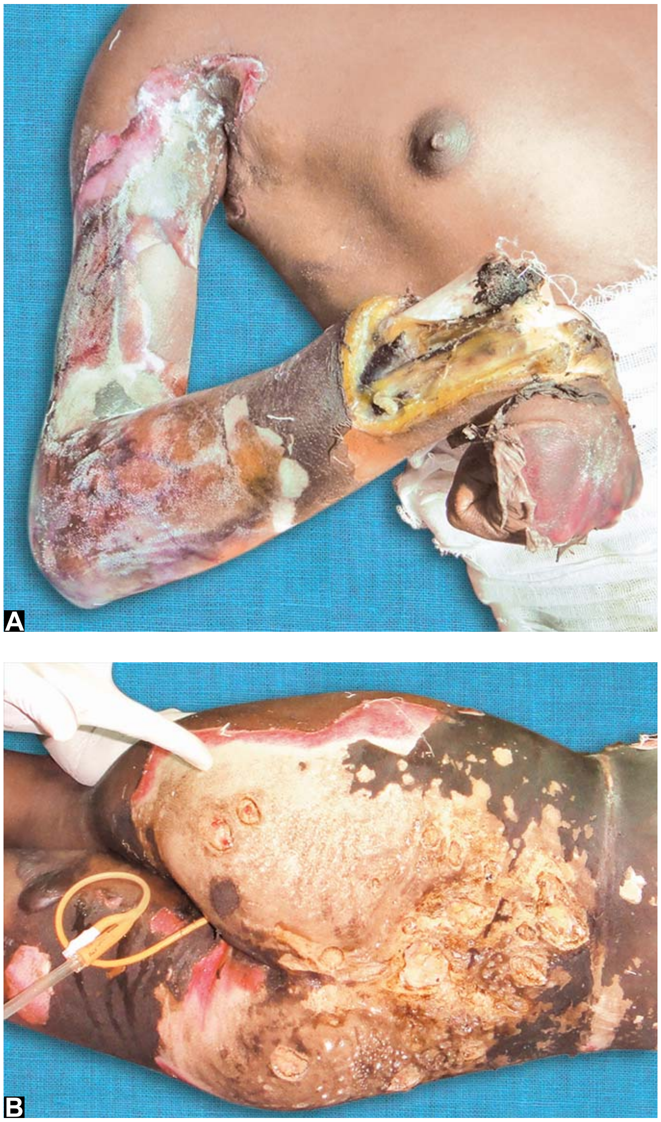

Figs 1.258A and B: High voltage electric burn wound of entry in the hand with extensive destruction. Some patient showing wound of exit in the buttock which is extensive and deep.

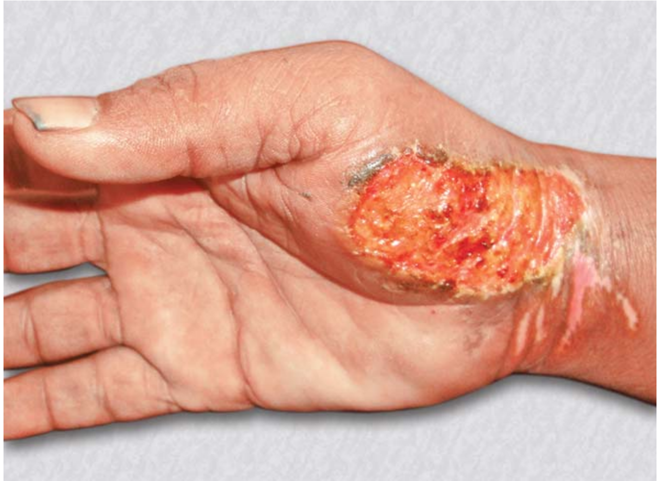

Fig. 1.259: Wound of entry in an electric burn.

- Patient may also have major internal organ injuries. GIT, thoracic injuries.
- Often convulsions can develop.
- Death may occur due to cardiac arrhythmias (***instant death due to ventricular fibrillation***).
- ***Gas gangrene*** is common after electric injury.
- Release of myoglobin can cause renal tubular damage and ***renal failure***.
- Acidosis is common and so often bicarbonate infusion is needed.
- Patient should *always be admitted* and should be assessed by ECG, cardiac monitor, U/S abdomen, chest X-ray, sometimes even CT scan head, cardiac enzyme analysis.
- Depending on the injury it is managed accordingly.
- Fractures and dislocations are common in electrical injuries which is treated accordingly.
- *Mafenide acetate* is better agent as it penetrates well and it is useful against clostridial infection.
- *Mannitol* is used to prevent myoglobin induced renal failure.

Electrical burns

Types	***Treatment***	***Complications***	***Monitoring and prevention***
High voltage—can cause ventricular fibrillation, cardiac arrest, death. Extensive organ and surface injuries with fractures. *Low voltage*—deep burn at the site of wound of entry.	Emergency resuscitation Assessment of burn Prevention of renal failure by hydration, dialysis, alkalization of urine to clear myoglobin, mannitol therapy, IV sodium bicarbonate. Extensive fasciotomy, debridement. Infection control. Reconstruction.	• Neurological like epilepsy, haemiplegia, aphasia, memory loss, headache. • Cardiac—arrhythmias. • Vascular injuries—major vessels, bleeding. • Compartment syndromes. • Ischaemia, gangrene of limbs. • Contracture development. • Bronchopneumonia, pleural effusion. • Abdominal—ileus, erosive gastritis, Curling ulcer, injury to liver, pancreas, spleen, GIT. • Bone and joint injuries. • Cataract can occur in high voltage burn. • Severe potassium deficiency is common.	**Monitoring:** ECG, echocardiography Relevant investigations like X-ray, ultrasound, CT, electrolytes, urine analysis, LFT, renal function tests. **Prevention:** Care during electrical work and with electrical system. Unused outlets should be sealed with plastic. Electrical system should be away from water source.

- Wound excision, amputation, surgery for internal organ injury, cardiac monitoring are essential part of the surgical management.

INHALATION INJURY

It occurs after major fire burns.
It is due to:

- Inhalation of heat.
- Noxious gases and incomplete products of combustion.
- At the site of fire, oxygen concentration is less than 2% which can cause death in 45 seconds due to hypoxia.
- Inhaled carbon monoxide binds with Hb immediately to form carboxyhaemoglobin causing severe anoxia and death.
- CO has got 240 times more affinity for haemoglobin than oxygen. Carboxyhaemoglobin in blood more than 10% is dangerous; more than 60% is life-threatening.
- Symptoms of carbon monoxide intoxication—headache, disorientation, visual changes, fatigue, vomiting, hallucinations, shock and cardiac arrest.
- Smoke contains hydrocyanide which causes tissue hypoxia and profound acidosis.
- Laryngeal oedema and laryngospasm.
- Bronchial oedema and bronchospasm.
- Formation of *bronchial cast* is typical which is due to oedema, lymph exudation, separation of ciliated epithelial cells from basement membrane. Inhaled gas causes supraglottic airway burn, laryngeal oedema, loss of respiratory epithelium, ARDS, CO poisoning, mechanical restriction of chest wall movement.

Later Problems

- ARDS, pneumonia.
- Atelectasis, pulmonary embolism.
- Pulmonary oedema, pneumothorax.

Clinical Features

- They have low oxygen saturation.
- Charring of mouth, oropharynx with facial burns.
- Carbon sputum.
- Change in the voice, singed facial and nasal hair.
- Decreased level of consciousness with stridor or dyspnoea.
- Acute pulmonary insufficiency with asphyxia, CO poisoning, upper airway obstruction. After 3 to 5 days, ARDS and hypoxia develops. Bronchopneumonia with septicaemia occurs after 5 days.

Management

- Replacing the patient from the site earliest.
- Ventilator support for several weeks.
- Antibiotics.
- Bronchoscopy, at regular intervals to remove bronchial cast.
- Tracheostomy whenever required.
- Hyperbaric oxygen.
- IV heparin to reduce bronchial cast. Heparin nebulisation (10,000 units in 3 ml saline 4th hourly) is also useful. N-acetylcysteine nebulisation—20% in 3 ml saline 4th hourly, bronchodilators like albuterol 2nd hourly is very useful. Hypertonic saline inhalation induces the effective coughing to remove casts. Racemic epinephrine is used to reduce mucosal oedema.

Note:
Steroids are not beneficial in inhalation burns.

- Monitoring the patient with arterial blood gas analysis regularly.

CHEMICAL BURNS

- In chemical burns, tissue destruction is more and progressive. It is always a deep burn.
- **Acid** burn occurs in skin, soft tissues and GIT. In GIT, it is common in **stomach** either due to nitric acid or sulphuric acid which may lead to severe gastritis or pyloric stenosis. Other acids are formic acid, hydrofluoric acid. They cause metabolic acidosis, renal failure, ARDS, haemolysis. Acidaemia should be corrected by IV sodium bicarbonate.
- *Hydrofluoric acid* is commonly used in industrial areas. It is strongest inorganic acid that can produce corrosion and dehydration. It chelates blood calcium causing hypocalcaemia and arrhythmias. It is managed with water irrigation, application of 2.5% calcium gluconate gel at 15 minutes interval, local intradermal and intra-arterial injection of 10% calcium gluconate. Continuous cardiac monitoring, IV calcium gluconate or calcium chloride administration is needed.
- **Alkali** burns occur in oral cavity and **oesophagus** which leads to multiple oesophageal strictures. Sodium hydroxide, lime, potassium hydroxide and bleach are common alkalis involved. They cause saponification of fat, fluid loss, release of alkali proteinates and hydroxide ions which are toxic.
- ***External chemical burns*** are always deep and cause extensive disfigurement with cosmetic problems.
- *Initial treatment* is dilution with water (***Hydrotherapy).*** It is done using 15-20 litres of running tap water.
- *Neutralisation with antidote should never be done at initial phase* of treatment as it creates exothermic reaction which aggravates the tissue damage. Late neutralisation is done, if required by 0.2% acetic acid in alkali burns, sodium bicarbonate, calcium gluconate 10% gel, topical ziphrin solution in acid burns.
- Treatment should always be with hospitalisation.
- Mannitol diuresis, haemodialysis, calcium gluconate IV, pain relief, serum electrolyte management, TPN, ventilator support are systemic management required.
- Late treatment is reconstruction of the face.
- Oesophageal dilatation or colonic transposition is done for oesophageal stricture due to alkali burn.
- Gastrojejunostomy is done for acid induced pyloric stenosis.
- *Tar burns* are treated additionally with neosprin which contains Tween-80 emulsifier of tar.

- *Cement* is calcium oxide alkali. Its burn is due to hydroxyl ion which is often deep. Treatment is removal of cloth, irrigation with water, keeping pH below 8. Often it may form eschar.

Medicolegal and ethical aspects in burns

- Police should be informed whenever a female, pregnant patient arrives with burns
- Consent for high-risk should be taken especially in burns > 30%
- Burns should be assessed whether it is accidental or homicidal
- Relatives should be informed about the duration of stay, problems, repeated surgeries
- In patients with severe burns who is likely to die, when suspected, dying declaration should be arranged
- Cost of therapy, long duration of stay and cosmetic problems should be informed to the relatives

J. Trauma

CHAPTER OUTLINE

- Triage
- Spinal Injury
- Neck Injuries
- Bullet Injuries
- Blast Injuries
- Penetrating Injuries
- Abdominal Trauma
- Blunt Trauma of Abdomen
- Duodenal Injury
- Pancreatic Injury
- Small Bowel Injury
- Colonic Injury
- Liver Injury
- Splenic Injury
- Renal Injury
- Urinary Bladder Injury
- Abdominal Compartment Syndrome
- Seatbelt Injuries

Trauma is the major public health problem in all countries.

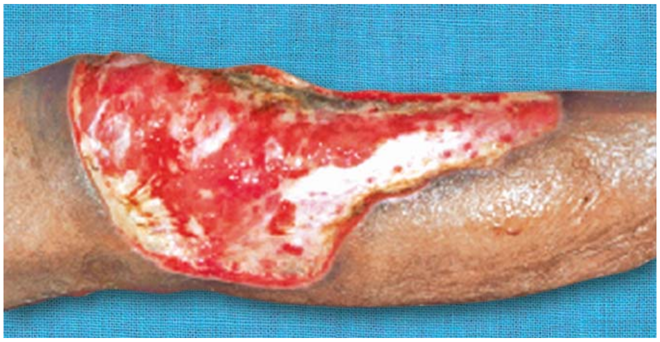

Fig. 1.260: Trauma causing large tissue defect exposing the bone.

TRIAGE

Triage means *"To sort"* in French.

Triage is a system to attend trauma patients, formulated by *Committee of Trauma of the American College of Surgeons.*

Advanced trauma life support (ATLS) is essential for first hour care of an injured patient.

Pre-hospital trauma life support (PHTLS) is to prevent deaths while injured patients are transported to the hospital.

Types of Triage System

- *Multiple casualties:* Staff and facilities are sufficient but priority is given to life-threatening injuries.
- *Mass casualties*: Staff and facilities are not sufficient to manage. Here those who are likely to have highest chance of survival are given priority.

Assessment of four components

i. Physiologic response
ii. Anatomical injury
iii. Biomechanical injury
iv. Comorbid factors

Triage algorithm

Step One (Assess physiological impact)

Measure vital signs and level of consciousness

- By Glasgow coma scale
- Systolic blood pressure
- Respiratory rate
- Revised trauma score. It is based on airway, laryngeal injury, spine injury, maxillofacial injury

Step Two (Assess anatomical impact)

All penetrating injuries to head, neck, thorax, major burns, fracture bones, pelvic fractures, paralysis

Step Three (Assess mechanism)

Automobile accidents, crash or blast injuries, high energy injuries, fall from more than 20 feet. Bullet injury

Step Four (Assess history)

- Patient's age below 5 years or age more than 55 years
- Cardiac diseases, respiratory and metabolic diseases
- Pregnancy
- Patients with bleeding disorders
- Immunosuppressed individuals

BASED ON THESE STEPS CONSIDER TO SHIFT THE PATIENT TO TRAUMA CENTER and TRAUMA TEAM SHOULD BE KEPT ALERT.
It is important in multiple and mass casualities (fire, blasts, automobile accidents, train accidents).

Management

- Initial evaluation of the patient.
- Physiologic stabilisation.
- Control of haemorrhage.
- Management of thoracic and abdominal injury.
- Management of cranial injury.

I. Primary Management

- **A**irway management (blocked by food, vomitus, clot, fallen tongue).

During an injury, unsuspected lesions of the spinal cord may cause the most excruciating abdominal pain.
—Theodore Schrire

- **B**reathing.
- **C**irculation.
- **D**isability and level of consciousness assessment by Glasgow coma scale.
- **E**xposure of the patient from head to toe for final assessment.
- **F**ingers and tubes: Finger evaluation, Foley's catheterisation.

Goals

- Identify life-threatening conditions.
- Decide and implement appropriate treatment to the area of trauma.
- First think to salvage the life, then think to salvage the limb.
- Rapid assessment, rapid resuscitation, rapid stabilisation.
- Optimum, complete care.
- Transport efficiently to higher trauma centre.

Categorise the patient

I: Deceased
II: Walking wounded
III: Immobile wounded
IV: Trapped wounded

Tag the patient accordingly

Red colour	: Immediate treatment is required
Yellow colour	: Urgent treatment is required
Green colour	: Delayed treatment is required
Blue colour	: Expectant treatment is required
Black colour	: Deceased

AIRWAY
Chin lift
Jaw thrust
Nasal airway
Oral airway
Endotracheal intubation
Tracheostomy
(assess airway patency)

BREATHING
100% oxygen
Assess bilateral chest raise
Assess breath sounds
Use pulse oximetry
Treat flail chest, pneumothorax
Intercostal tube drainage

CIRCULATION
Monitor vitals
Heart sounds
ECG
IV fluids blood transfusion
Treatment of shock
Control of external bleed
Use two IV lines—14G/16G

DISABILITY EVALUATION
Neurological examination
Glasgow coma scale
Pupillary reaction

EXPOSE THE PATIENT FULLY
Undress the patient
Hypothermia assessment
Assess injuries
Examine joints, bones, abdomen, other systems
Look for identification marks

FINGERS AND TUBES
Examine all orifices like P/R, P/V, etc.
Use required tubes like catheter, Ryle's tube

II. Investigations

- X-ray spine, chest, pelvis, extremities.
- CT scan.
- Blood group and cross-matching.
- Arterial blood gas analysis.
- Serum electrolytes.
- U/S abdomen.

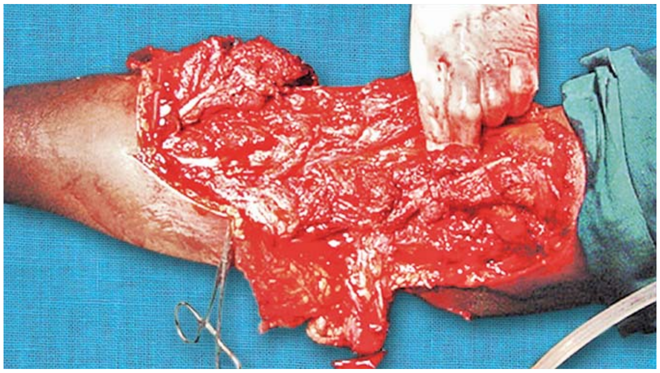

Fig. 1.261: Crush injury leg due to road traffic accident.

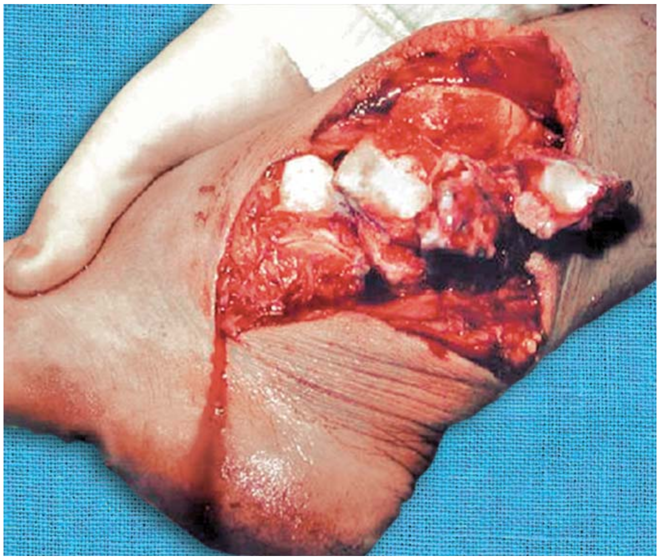

Fig. 1.262: Ankle injury with open wound.

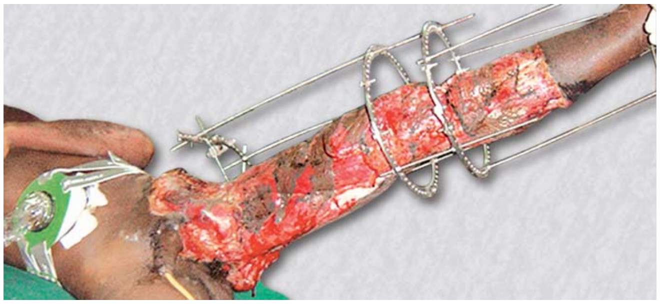

Fig. 1.263: Degloving injury involving entire left lower limb, perineum, and left groin. Patient has lost scrotum and both testes. There were no internal injuries and vessels and nerves were intact. Patient underwent wound excision extensively and colostomy was done to promote healing of perineal wound and prevent contamination.

III. Secondary Survey

Re-evaluate the patient completely again.

IV. Definitive Care

(All discussed under individual topics.)

Mechanism of Trauma

- Blunt trauma—direct or indirect blunt injury can occur. Seat belt reduces the blunt injury in vehicles.
- Penetrating injury—severity depends on the extent of deeper injury.
- Blast injury.
- Crush injury—earthquake, industrial accidents, and train accidents—causes crush syndrome; compartment syndrome.
- Burn injury.
- Injury in alcohol patients.

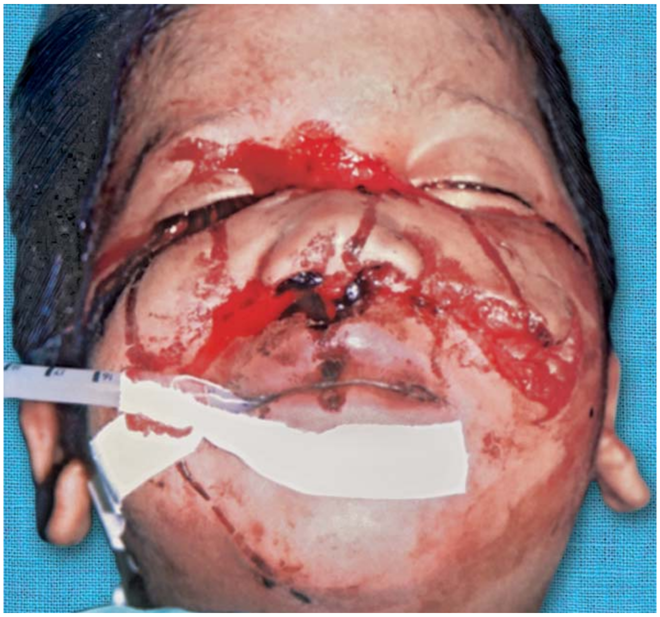

Fig. 1.264: Class III dog bite on face in a boy.

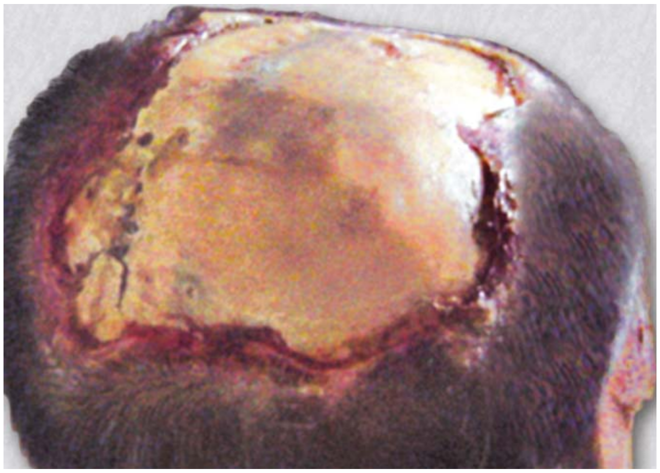

Fig. 1.265: Degloving of scalp with bone periosteum exposing the skull bone. Outer table was actually dead and required bone removal and graft after healthy granulation tissue was formed.

Concepts in Trauma Management

- Concept of '*golden hour*' to treat the trauma patient is important.
- Multidisciplinary approach.
- Planning, setting up, organizing, team work.
- Assess respiratory system; circulation; bleeding areas—as priority.
- Assess also whether patient is haemodynamically stable or unstable.
- Arrange fluids, blood, catheters, ventilator, etc.
- Further definitive therapy depending on severity and site of injury.

Damage Control Surgery

- Resuscitation and early therapy in operation theatre itself.
- Minimum but essential surgery to control bleeding and prevent contamination.
- Secondary definitive surgery at a later period to have final control.

SPINAL INJURY

- Assess the type, extent and severity of the injury.
- Careful first aid and transfer to prevent further damage to the spinal cord.

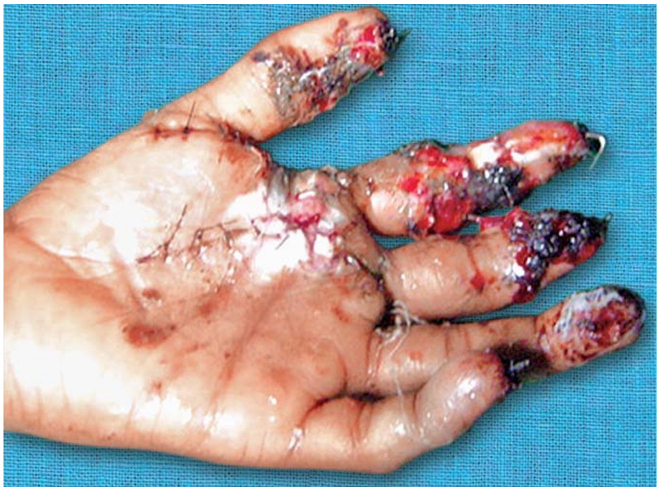

Fig. 1.266: Hand injury involving all fingers.

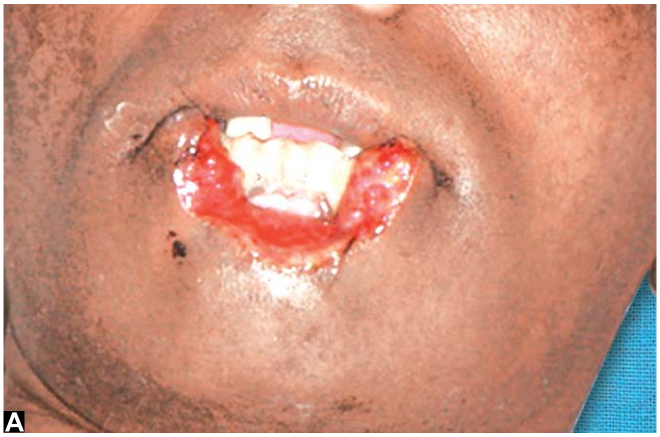

Fig. 1.267A

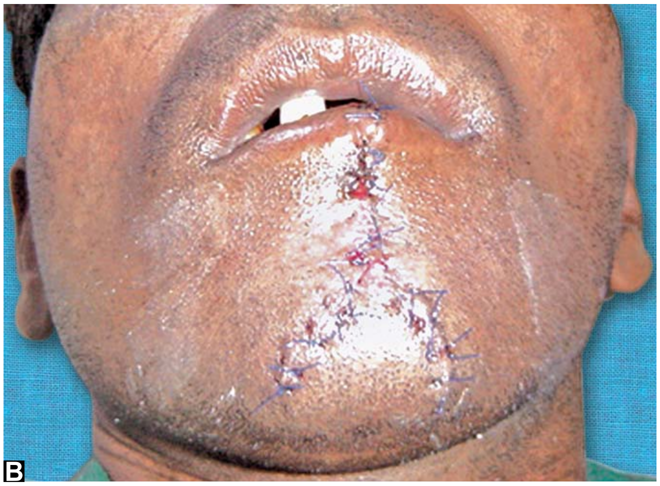

Fig. 1.267B

Figs 1.267A and B: A person has bitten this patient during a fight and removed the central part of lower lip. It is repaired primarily by Y repair. ***Remember human bites are most dangerous***.

- Assess the sensory loss or motor loss properly.
- Assess fractures clinically, by X-ray and MRI.
- *Central cord syndrome* is common and is due to hyperflexion or hyperextension of the neck in an injured patient causing ischaemia of spinal column due to interfering of spinal artery blood flow.
- *Brown-sequard syndrome:* It is due to partial transection of the cord causing ipsilateral motor function loss and contralateral sensory function loss.
- High dose of steroid is very useful to prevent further damage.
- Rest, traction to neck.
- Decompression of spinal canal surgically by removing bone, disc, haematoma is useful.
- Spinal stabilisation.

NECK INJURIES

Neck is divided into zones for managing neck injuries
- *Zone I*: From clavicle to cricoid cartilage
- *Zone II*: From cricoid cartilage to angle of the mandible
- *Zone III:* Above the angle of the mandible

Indications for Neck Exploration in Injuries

- Expanding haematoma.
- Uncontrolled external haemorrhage.
- Decreased carotid pulse.
- Stridor, hoarseness, dysphonia, haemoptysis.
- Severe dysphagia, odynophagia.
- Blood in oropharynx.

Treatment

- The neck is explored with adequate incision under general anaesthesia.
- The injured structure like vessels, oesophagus, trachea, muscles are sutured.
- Antibiotics.

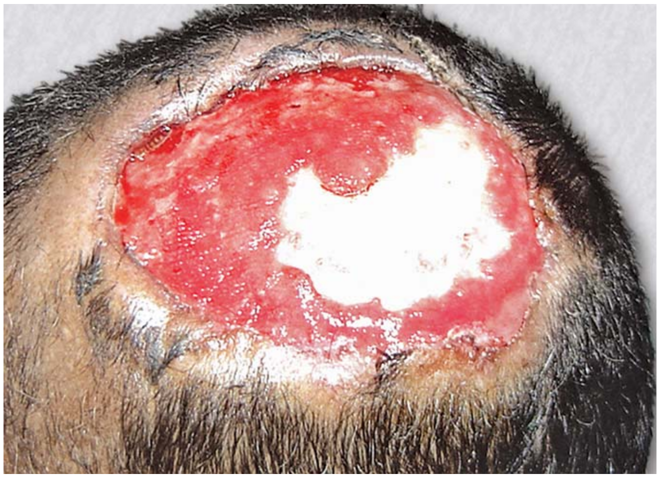

Fig. 1.268: Degloving scalp trauma and exposed bone is granulating well.

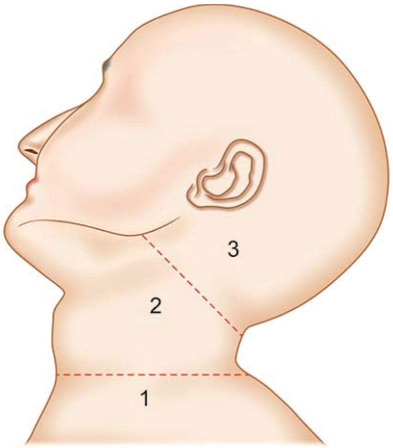

Fig. 1.269: Neck trauma zones.

- Blood transfusion is given as required.
- Ryle's tube for 5-7 days.

Other injuries like head, thorax, abdomen, maxillofacial area are discussed in respective chapters.

BULLET INJURIES

Bullet injury has *wound of entry* and *wound of exit*. Extent of damage is not related to the external wounds. It is related to the travel of bullet inside and extent of blast or cavitation effect inside caused by the bullet. It causes burn damage.

It can damage vessels, organs like liver, spleen, kidneys, bowel, lungs, heart, cranial structures, soft tissues, bones and joints.

Management

- The wounds are explored properly under general anaesthesia.
- All dead tissues and dead muscles are excised.
- Skin is generously and adequately incised.
- Injured nerves are cleaned and silk marker stitches are placed to identify for later secondary suturing (Nerve should not be sutured primarily in bullet injury).
- All foreign bodies are removed.

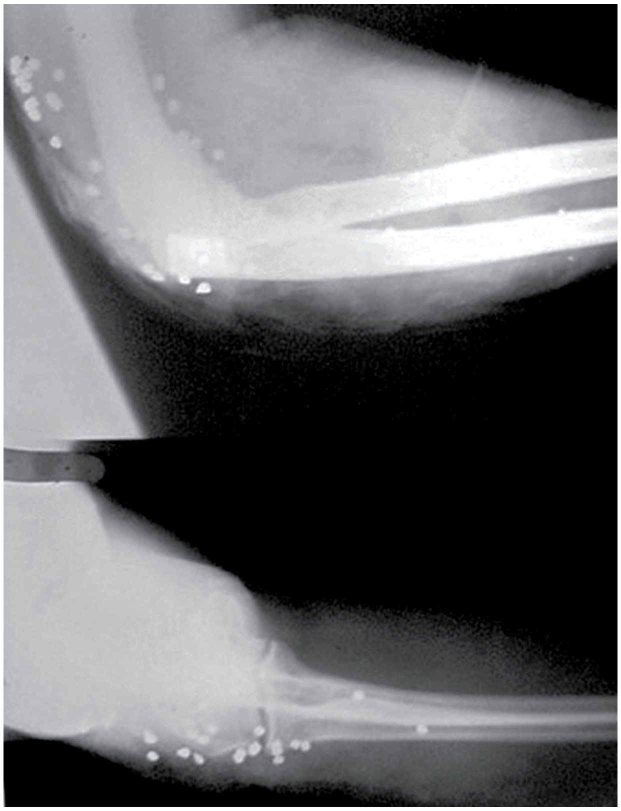

Fig. 1.270: Multiple pellets over elbow region after gunshot.

- Tendon repair should not be done primarily.
- Wound should not be closed. It should be left open.
- Adequate blood transfusion and antibiotics coverage should be given.
- Major artery or vein are sutured. Vein graft can be used. But *synthetic graft should never be used.*
- Thorough inspection, irrigation and debridement of injured joints is done.
- Immobilisation is done.
- Tetanus toxoid, antitetanus globulin (3000 units IM), antigas gangrene serum is given.
- Second look surgeries at a later period is done once patient has been stabilised.
- Delayed primary closure in 4-7 days or secondary closure in 14 days is done.
- Depending on extent of defect, skin grafting or flaps are used.
- Laparotomy, thoracotomy, craniotomy are done depending on the site of the injury.

BLAST INJURIES

Here extent of damage is much more than bullet injuries.

It creates ***complex blast wave*** which contains ***blast pressure wave*** and ***mass movement of air.***

This explosion pressure wave is more than 1000 pounds per square inch. This pressure wave has got incident pressure and reflected pressure. Both will cause severe damage.

Factors causing the damage

- High pressure wave
- Mechanical injury
- Chemical injury
- Thermal injury
- Inhalation of toxic gases and smoke

Organs Affected

- Ear drums, lungs.
- GIT, brain.
- Skeletal system.

Individual becomes ***deaf*** after blast and so rescue work may be delayed.

Management

- Critical trauma care.
- Management of shock and triage primary manage-ment.
- Urgent surgeries like laparotomy, thoracotomy, craniotomy.
- Massive blood transfusion.

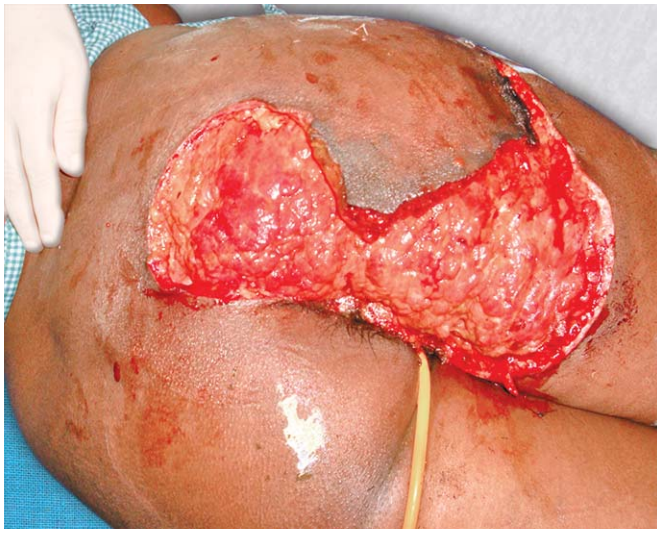

Fig. 1.271: Degloving injury buttock. It needs colostomy to protect the wound from sepsis. Later once skin coverage is done colostomy is closed.

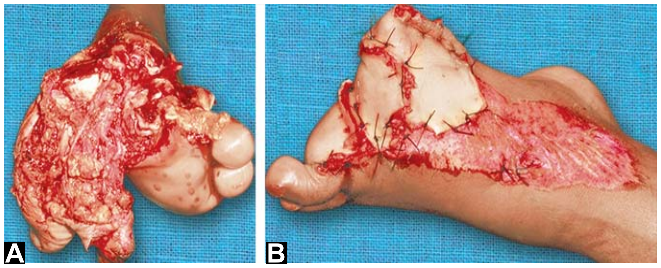

Figs 1.272A and B: Crush injury foot with retained skin from plantar aspect was rotated forward *(Courtesy*: Dr Mayur Rai, Orthopaedician, KMC, Mangalore).

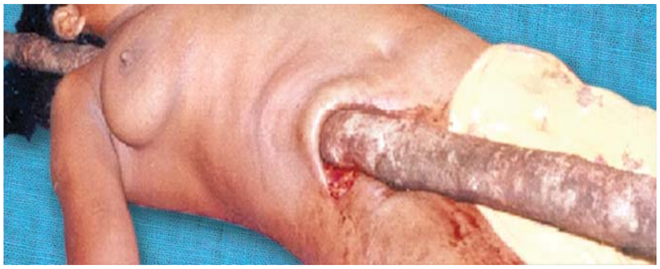

Fig. 1.273: Penetrating injury. Pole missed all the major vessels. Miraculously the patient survived, after a Marathon surgery to tell her tale to her children.

- Antibiotics.
- Ventilator support.
- Management of specific organs like eye, ear.

PENETRATING INJURIES

- It can occur in abdomen, thorax, cranial cavity.
- It causes haemorrhage, damage to internal organs like liver, bowel vessels, lung, pericardium and heart, brain.

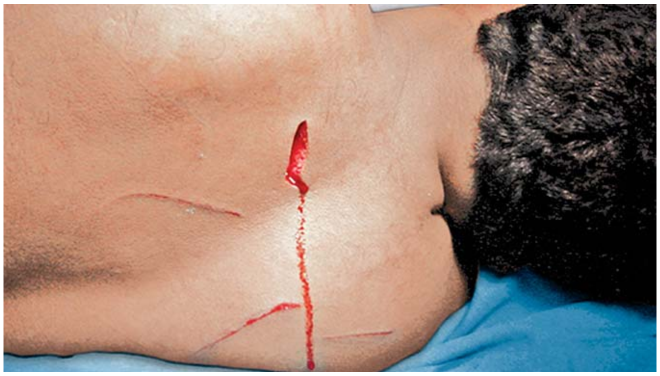

Fig. 1.274: Stab wound on the back communicating into the thoracic cavity. Wound was explored and sutured, with an ICT inserted into the thoracic cavity.

- It is life-threatening and immediate surgical intervention is the only treatment. Patient requires adequate amount of blood transfusion, antibiotics, shock management.

ABDOMINAL TRAUMA

It can be:
- Blunt trauma
- Stab injury
- Abdominal wall injury

Abdominal trauma is a major surgical emergency which most surgeons face. It is often associated with head injuries, chest, pelvic and bone injuries. Often patient is unconscious causing difficulty in diagnosing the condition. Often more importance is given to other system injuries like of head, thorax and bones whereas abdominal injury is not addressed properly causing life-threatening consequences. When patient is conscious, history related abdominal trauma is useful. Abrasion over the abdominal skin suggests the possibility of internal injury *(London's sign)*. Distension, tenderness, rebound tenderness, fullness and dullness in the flank when present one should presume internal injury. Tachycardia, hypotension, shock may be evident when there is significant haemoperitoneum. Injuries may be organ injury like of liver, spleen, kidney, pancreas, etc. or bowel injury or retroperitoneal injury which is often under diagnosed or missed. Major vessel injury like of inferior vena cava, mesenteric vessels can cause real threat to life unless it is identified and managed early. 25% of entire trauma patients need surgical exploration of the abdomen.

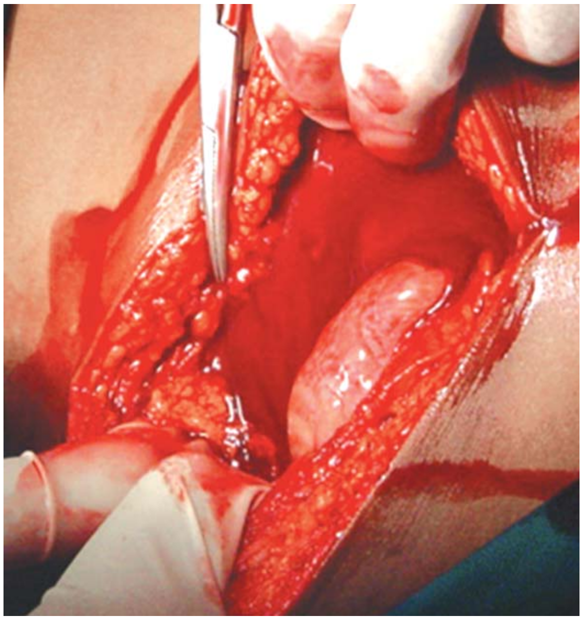

Fig. 1.275: Traumatic haemoperitoneum.

Abdominal trauma can be *blunt or stab/penetrating or abdominal wall injuries*. Spleen is the most common organ involved in *blunt trauma*. Often in blunt trauma first part of the jejunum or ileocaecal junction gives way (*blow out effect*) due to traction often causing complete transection of bowel horizontally close to the junction. It is due to force of the mobile part of the bowel over the fixed part. Liver is the most common organ involved in *penetrating injuries*.

Injuries of the abdomen may be *closed injuries, compression injuries and penetrating injuries*. Penetrating injuries may be *low velocity injury* like stab injuries or *high velocity injury* like gunshot injuries. Penetration of blunt weapon causes less deep trauma than sharp weapon (sickle, knife). In sickle injury tip and sharp edge moves in curved pattern and so it is often difficult to predict the depth, track and organs injured.

Routinely followed indications for exploration in abdominal trauma are—hypotension without any other cause; bleeding through wound; continuous bleeding in nasogastric tube; evisceration of abdominal content through the open wound except in case of protruded omentum without any hypotension and features of peritoneal irritation; air under diaphragm in blunt abdominal injury (not in penetrating injury as external air gets sucked into the peritoneal cavity through the wound).

Types

- Liver injury.
- Spleen injury.
- Gastric/small bowel/colonic injuries.

- Duodenal injuries.
- Pancreatic injuries.
- Injuries to kidney/bladder/urethra.
- Mesenteric injury.
- Vascular injuries.
- Associated injuries like of diaphragm, lungs.
- Abdominal compartment syndrome.
- Gunshot or blast injuries.

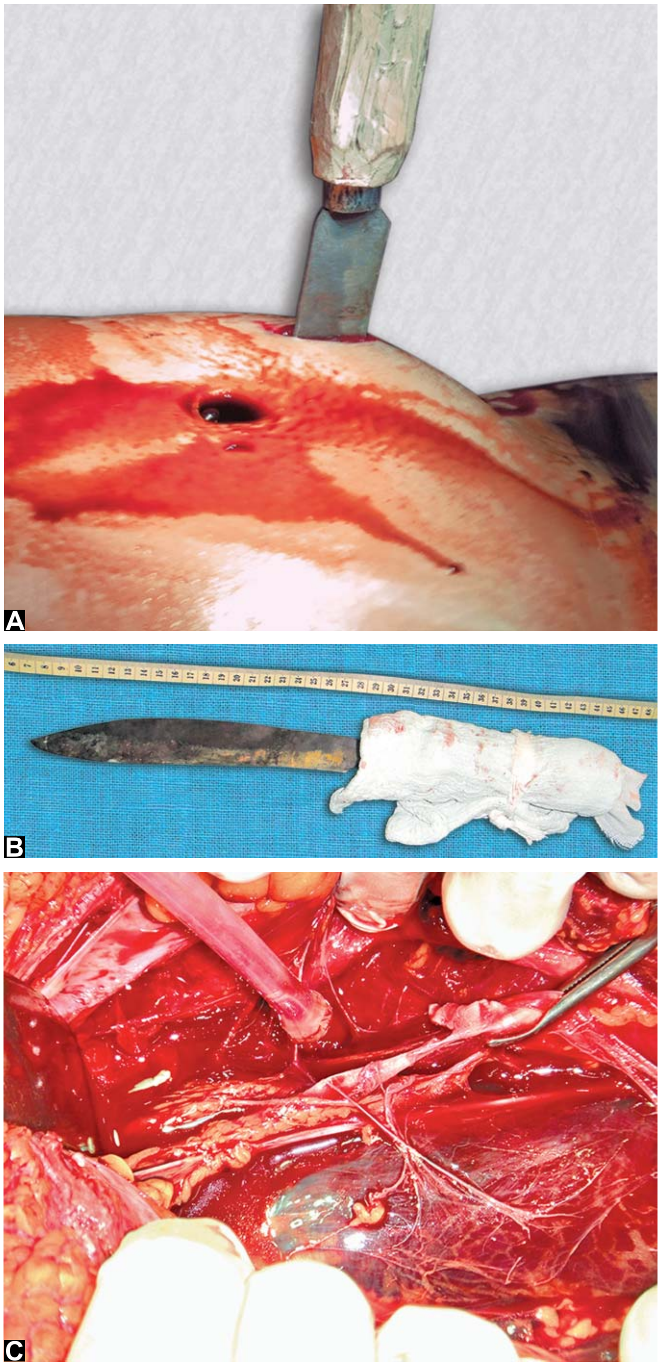

Figs 1.276A to C: Stab injury to LIF by an angry husband to his wife causing left common iliac artery transection injury with aortic partial injury. In spite saphenous vein graft reconstruction patient could not survive (*Courtesy*: Professor Yogishkumar, MS, KMC, Mangalore).

General Clinical Features

- Features of shock—pallor, tachycardia, hypotension, cold periphery, sweating, oliguria.
- Abdominal distension.
- Pain, tenderness, rebound tenderness, guarding and rigidity, dullness in the flank on percussion.
- Respiratory distress, cyanosis depending on the amount of blood loss.
- Bruising over the skin of the abdominal wall.
- Features specific of individual organ injuries.

Investigations

1. Ultrasound abdomen. FAST is *Focused Abdominal Sonar Trauma*: It is rapid, noninvasive, portable bedside method of investigation focusing on pericardium, splenic, hepatic and pelvic areas. Blood more than 100 ml in cavities can be identified. It is not reliable for bowel or penetrating injuries. It often needs to be repeated.
2. **Diagnostic peritoneal lavage (DPL):** It is done in case of blunt injury abdomen. Through a subumbilical lavage catheter one litre of normal saline/Ringer's lactate is infused into the peritoneal cavity. Patient is changed to different positions and side-to-side. Fluid content is aspirated from the abdomen for assessment.
   It has got 98% accuracy rate.

One of the criterias signifies positive lavage

- 10 ml or more of gross blood
- RBC count more than 1,00,000/cumm
- WBC count more than 500/cumm
- Amylase level in the fluid more than 175 IU/dl
- Presence of bile, bacteria, food particles or foreign body

It is the procedure of choice in physiologically unstable patient with blunt abdominal injury (like with spinal injury, unconscious patient).

Contraindications for DPL

- When laparotomy is definitely indicated
- Previous laparotomy
- Pregnancy
- Obesity

3. CT scan is indicated in assessing retroperitoneum, solid organ injuries. It is noninvasive and highly specific.
4. Diagnostic laparoscopy (DL) is valuable method in stable abdominal trauma patient.

Treatment

Emergency laparotomy.

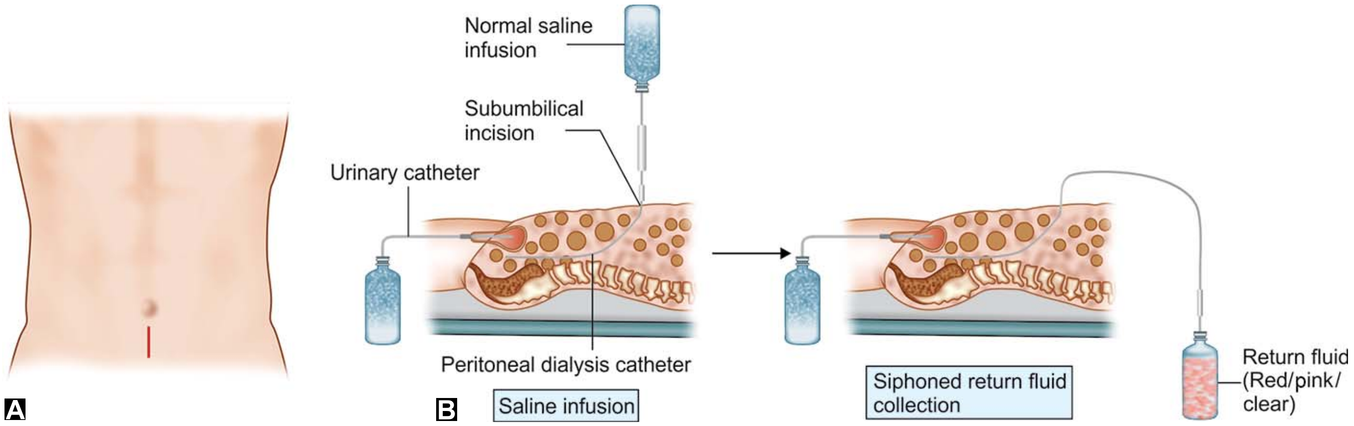
Normal saline infusion
Subumbilical incision
Urinary catheter
Peritoneal dialysis catheter
Saline infusion
Siphoned return fluid collection
Return fluid (Red/pink/ clear)
A
B

Figs 1.277A and B: *Diagnostic peritoneal lavage*—incision and technique. 10 French polyvinyl catheter is used. Urinary bladder is emptied by passing a catheter. After injecting xylocaine local anaesthesia into subumbilical region, 2-3 cm vertical subumbilical midline incision is made. Skin, linea alba is incised. Local anaesthesia is infiltrated into the peritoneum again. Peritoneum is held with two haemostats and a purse string suture is placed using polyglactic acid absorbable suture material. Peritoneum is incised for 3 mm length. Catheter (standard peritoneal dialysis catheter) is introduced into the peritoneal cavity. If blood enters the catheter immediately, it means early laparotomy is needed and carried out without continuing the peritoneal lavage. Otherwise, one litre of normal saline is infused into the peritoneal cavity rapidly in few minutes through the catheter using a drip set with elevation of the fluid bottle/bag. Patient is moved well to mix the fluid in all four quadrants. Now bag is lowered below so that fluid from the peritoneal cavity reenters/siphoned into the bag. Collected fluid is analysed for red cells, leukocytes, etc. DPL may not be useful in bowel injury, retroperitoneal injury, diaphragmatic injury, organ haematoma (subcapsular splenic haematoma). If patient is decided for observation catheter can be left *in situ* for repeat DPL after 6 hours. One has to remember that DPL is not a substitute for clinical assessment and monitoring. In *Lazarus-Nelson* approach Teflon catheter with a guide wire is used.

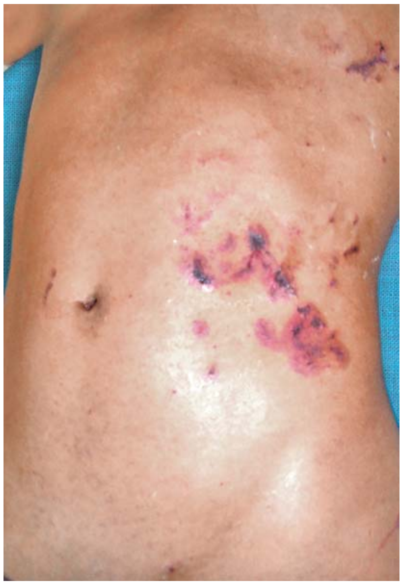

Fig. 1.278: Blunt injury abdomen. Note the bruising over abdominal skin—*London's sign.*

Indications for laparotomy

- Frank haemoperitoneum
- Significant diagnostic peritoneal lavage
- Haemodynamically unstable patient
- U/S or CT scan shows significant intra-abdominal injuries

BLUNT TRAUMA OF ABDOMEN

It is common in accidents. It is often missed or lately diagnosed. Ultrasound/CT abdomen or diagnostic peritoneal lavage (DPL) is useful. In many cases on clinical grounds direct exploratory laparotomy is done. Plain X-ray abdomen may show gas under diaphragm.

Difficulty arises in *deciding about the need for laparotomy* in abdominal trauma in *unconscious* patients. If severity of external injury is out of proportion to the existing severe shock then exploratory laparotomy is indicated in an unconscious patient. It is also often difficult to diagnose bowel injury in such patients. If it is suspected laparotomy should be undertaken in such patients. Associated spinal injury masks the abdominal findings.

Injuries may be of liver, spleen, GIT, pancreas, mesentery, vascular or diaphragm. Associated chest, pelvis, skeletal and head injuries should be remembered.

Features of Blunt Trauma

- Features of profound shock, progressive distension of abdomen, pain, tenderness, guarding, rigidity, rebound tenderness, dull flank.
- Features specific of individual organ injury like obliteration liver dullness in bowel injury.
- Bruising of skin over the abdomen—*London's sign*.
- Respiratory distress, cyanosis.
- Repeated clinical examination is a must in blunt trauma.

Evaluation

Ultrasound Abdomen

- It is very useful, simpler, noninvasive method of evaluating the abdomen. Negative ultrasound means no immediate further intervention is needed and also conservative treatment can be undertaken.
- *Advantages of ultrasound:* There is no danger of radiation; it can be done bedside; it can be repeated many times; it

is cost-effective. Its sensitivity is 90%; specificity is 98%. Focused abdominal sonar (ultrasound) for trauma *(FAST)* is very useful method.

- *Disadvantages:* It is less useful in obesity, with interposition of gas, when fluid is less than 500 ml; retroperitoneal injuries and bowel injuries.
- *Focused abdominal sonar trauma (FAST)*: It is rapid, noninvasive, portable bedside method of investigation focusing on pericardium, splenic, hepatic and pelvic areas. Blood more than 100 ml in cavities can be identified. It is not reliable for bowel or penetrating injuries. It often needs to be repeated.

Diagnostic Peritoneal Lavage (DPL) (by Perry)

It is useful in *blunt injury* abdomen. It is not very useful in penetrating injury, bowel injury, retroperitoneal and pelvic injuries.

CT Scan of Abdomen

It is most commonly used and better investigation for abdominal trauma. It is useful in blunt/penetrating trauma, suspected pancreas, spleen, liver, duodenal, retroperitoneal injuries. Smaller injuries, early haemoperitoneum are better detected. It is noninvasive, highly specific, highly accurate (96%), with low false-positive/low false-negative, noninvasive.

Other Investigations

a. *Abdominal diagnostic paracentesis* (Drapanas and McDonald): Here 18 G short bevel spinal needle is inserted into the peritoneal cavity after injecting local anaesthetic agent into the abdominal wall. With continuous suctioning through syringe, needle is passed at various sites. Positive tap means return of minimum of 0.1 ml of nonclotted blood. False-positive result occasionally can occur due to needle puncture of abdominal wall vessels. Needle should not be inserted close to previous abdominal scar as bowel may be adherent underneath the scar. Change of direction of needle is done by withdrawing the needle tip outer to peritoneum and again puncturing the peritoneum. Puncture by 18 G needle of nondistended bowel will seal without any leakage. Peritoneal tap should be avoided if bowel is distended. *Bilateral flank tap/four quadrant tap* is also done with similar result. Rectus sheath haematoma and false-negative results are the problems.

b. *Diagnostic laparoscopy* is very useful. It can be done under local anaesthesia. Haemoperitoneum, solid organ and diaphragmatic injuries are well assessed. But bowel and retroperitoneal injuries are more likely to be missed.

c. *Arteriography* through Seldinger technique is useful in suspected cases of renal arterial injury (thrombosis/spasm); intimal tears, traumatic aneurysm and aortic occlusion (after seat belt injury) are well diagnosed with arteriography. Often it can be therapeutic also. Pelvic bleed extending into retroperitoneum is not uncommon which can be assessed by arteriography and also the bleeding vessel can be identified. But venous bleed cannot be assessed by this.

d. *Doppler assessment of major vessels* may be beneficial especially for IVC, aorta, iliac vessels, and portal system; but with haemoperitoneum visualisation window may be poor and vessels can be better identified by contrast CT scan.

Management Concepts in Abdominal Trauma

- *Evaluation* of extent of the injury; number of organs injured and severity of injury; haematocrit assessment (haemoglobin drop up to 6 gm% is tolerated well with adequate tissue oxygenation. Rapid drop of hemoglobin needs adequate number of blood to be kept ready for transfusion, like 5/10/bottles or more); central line for volume replacement; urinary catheterisation; administration of systemic antibiotics.
- *Autotransfusion of blood* is very useful as a life-saving procedure in such situation. Blood from the cavity is sucked out into a sterile bottle which contains 150 ml of 3.8% sodium citrate dextrose solution. This blood is strained/filtered through gauze and re-transfused. If there is colonic and small bowel injuries auto transfusion is not possible for fear of sepsis due to contamination.
- *Upper midline incision* extending down across the left of the umbilicus is the preferred incision. But surgeon should not be hesitant to extend the incision into the thorax or do horizontal T or extend as needed depending on the internal organ injury.
- *First priority* after opening the abdomen *is immediate control of profuse bleeding using finger compression or mop* or pressure. Later once the field is clean; area is assessed for the extent of injury without releasing the compressed finger on the bleeding site. A vascular clamp or bulldog clamp is helpful in such situation. Once it is applied over the site of bleeding, compressing finger can be removed. Vascular suturing using 4 zero or 6 zero polypropylene/resection of the tissue; reconstruction of the area; persistent pressure mop *in situ* with closure of the abdominal wall with an option of second look surgery in 48 hours are the different options. Individual organs are assessed and graded for injuries and managed accordingly.
- During laparotomy entire abdomen should be inspected/palpated carefully for any additional missed injuries. Lesser sac, retroperitoneum, duodenum, pancreas and diaphragm should be checked. Often peritoneum on the margin of the duodenum and right side colon is incised, duodenum and colon is reflected medially to visualise the retroperitoneum. Pelvic structures need special attention. Rectum, urinary bladder injuries are likely to be missed if proper attention is not given. On catheterization, if urine is clear it means urinary system is normal. Portal venous system should be assessed.
- *Resection or repair* should be decided later once haemostasis is maintained. Whether the injury is to the bowel or organs

(liver/spleen/kidney, etc.) resection or repair approach is decided depending on the severity of individual organ injury (based on scale or grade).

- *Mesenteric tear* may be the cause for haemoperitoneum. Tear may be perpendicular or parallel to the bowel. If it is perpendicular, haemostasis and approximation of the mesentery is sufficient; if it is parallel tear, then blood supply to corresponding bowel may be compromised and resection of that part of the bowel is indicated. Mesenteric haematoma is left alone if small and nonprogressive. Whether there is any colour changes in the adjacent bowel should be observed. If haematoma is large; if it is progressive; if it causes compromised blood supply to the adjacent bowel, then it should be gently evacuated. Mesenteric leaf is opened using curved scissor; clot is evacuated using gentle finger dissection; bleeder is identified and ligated. If there is compromised bowel function, it should be resected. Bleeding from the major vein like superior mesenteric vein is disastrous as tear may not be localised but may be extensive; and even gentle dissection may cause more tear. It is carefully mobilised; vascular clamps are applied and repaired using 5 zero polypropylene sutures.

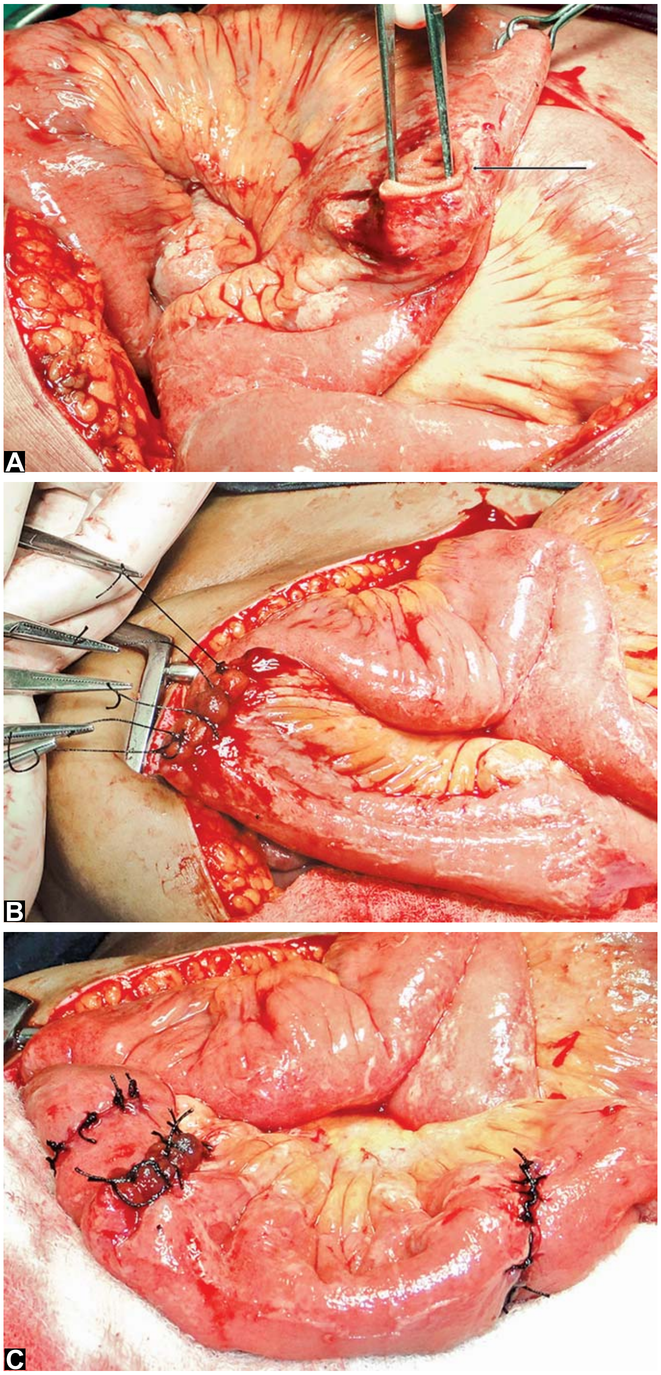

Figs 1.279A to C: Traumatic blunt injury abdomen causing small bowel injury which is sutured using interrupted horizontal silk sutures.

- **Aortic clamping**: Catastrophic bleeding found after opening the abdomen which cannot be controlled and bleeding with profound hypotension are the indications for aortic clamping. Profuse intraperitoneal bleed comes under control temporarily by tamponade effect of tense abdominal wall and it itself temporarily helps the patient. The moment abdomen is opened; tamponade effect is released causing further rapid bleed leading into critical catastrophe. If such event is expected prior to opening the abdomen very quick rapid thoracotomy (*prelaparotomy thoracotomy*) through left 5th intercostal space is done; left lung is deflated and displaced; pleura over the thoracic aorta is incised; aorta is dissected using finger; vascular clamp or soft intestinal occlusion clamp is applied to occlude the thoracic aorta. Later laparotomy is performed to go ahead with management of the bleeding. If profound bleeding is observed after laparotomy necessitating the aortic clamping, it is done by applying the clamp in infradiaphragmatic part of the aorta. Peritoneum is incised on the right of the abdominal oesophagus in infradiaphragmatic area; aorta is dissected using finger high up close to diaphragm to avoid injury to celiac plexus; clamp is applied across (infradiaphragmatic aortic occlusion).
- *Usually drainage* using tube drains on either side of the abdomen is used even though it is controversial. ICT should be placed if thoracotomy is also undertaken.
- *Jejunostomy* for enteral nutrition is ideal in all major abdominal injuries. Often gastrostomy is also done along with jejunostomy in case of duodenal and pancreatic injuries.
- *Management of individual organs* after grading its severity of the injury—duodenum, pancreas, liver, spleen, bowel, kidney, etc. ***(Please refer individual chapters for detail—highlights of individual organ injury is given below).***
- Management as critical care (ICU with intensivist); multiple blood transfusions; management of sepsis, maintenance of respiration, management of electrolyte changes, treatment of renal failure, provision of nutrition, prevention of DVT, management of DIC are very essential part of postoperative treatment.

DUODENAL INJURY

- Its severity depends on the type and extent of the injury.
- It can be haematoma or lacerations.
- Lacerations can cause duodenal disruption, may be < 50% or > 50% or 75% or more.
- Laceration may extend into the ampulla, distal CBD, pancreas or with duodenal devascularisation.

Grading of duodenal injury	
Grade I – Haematoma – Laceration	Involving single portion of the duodenum. Partial thickness injury without perforation.
Grade II – Haematoma – Laceration	Involving more than one portion. Disruption less than 50% circumference.
Grade III – Laceration	Disruption of 50-75% of the circumference of 2nd part of the duodenum; disruption 50-100% of the 1st, 3rd or 4th part of the duodenum.
Grade IV – Laceration	Disruption more than 75% of 2nd part of the duodenum and involving the ampulla or distal common bile duct.
Grade V – Laceration – Vascular	Severe disruption of duodenopancreatic complex. Duodenal devascularisation.

Management

- CT scan is more relevant investigation.
- Associated other injuries should be managed accordingly.
- Haematoma without extension is managed conservatively with nasogastric aspiration, antibiotics and IV fluids.
- Lacerations are sutured surgically with a stenting or often with bypass like gastrojejunostomy.
- ERCP stenting or CBD bypass is also often required.

Complications

- Infection, duodenal leak.
- Peritonitis, haemorrhage.

PANCREATIC INJURY

- It can be in the head or body and tail of the pancreas.
- It may be associated with injury to duodenum or portal or superior mesenteric veins.
- It can be contusion or severe lacerations.

Management

- High resolution CT scan is diagnostic.
- Distal pancreatectomy for injuries distally.
- Conservative treatment is useful with antibiotics, IV fluids.
- Whipple's operation or total pancreatectomy is done as a last resort.
- Drainage of the pancreatic bed is simple and often useful method.

Complications

- Pancreatitis, septicaemia.
- Pancreatic fistula, pancreatic abscess formation.
  Pancreatic injury has got high mortality (> 45%).

SMALL BOWEL INJURY

- It can be blunt injury or stab injury.
- Blunt injury causes disruption of either duodenojejunal region or at ileocaecal region.
- Presentation is like haemoperitoneum or features of peritonitis.
- *Monks localising zones* in the abdomen signify the location of the small bowel injury.
- Presence of pattern bruising over the abdominal wall signifies the small bowel injury and its site. It is called as *London's sign.*

Management

- Plain X-ray abdomen shows gas under abdomen with ground-glass appearance.
- U/S abdomen is useful.
- Laparotomy and closure of the perforation if it is small.
- In presence of extensive bowel injury or multiple injuries, resection and anastomosis is done.
- Any associated injuries should be dealt with accordingly.

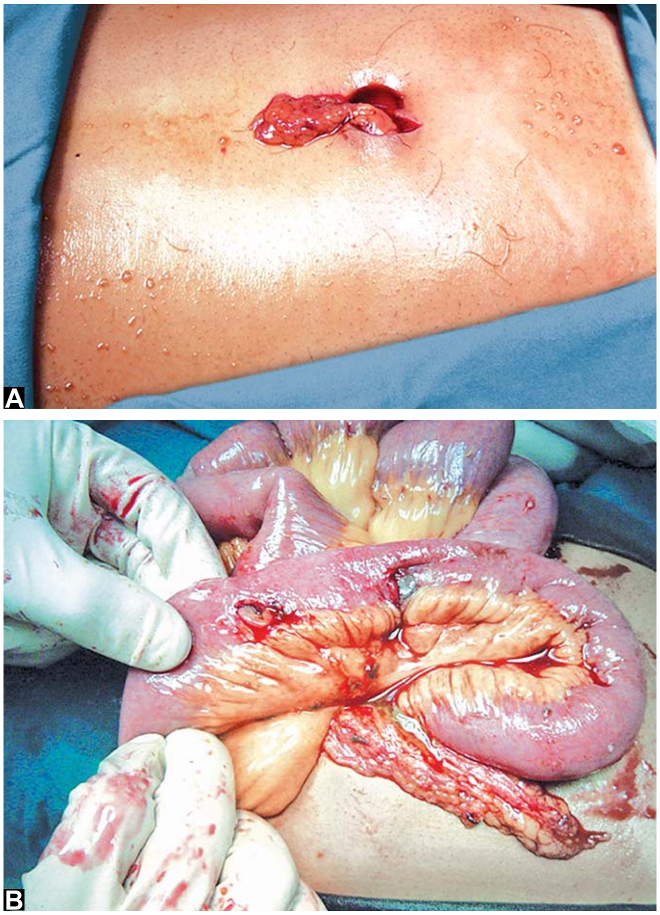

Figs 1.280A and B: Assault causing stab injury abdomen. On opening the abdomen, multiple perforations were found in the small bowel and was sutured. Patient recovered well.

COLONIC INJURY

- Left sided injury is treated with proximal colostomy with closure of the wound if it is small, or resection and anastomosis if it is wider area. Closure of colostomy is done at later stages after 3-6 months.
- Small wound over right sided colon can be sutured primarily.
- Ileostomy alone or ileostomy with ileo-transverse anastomosis or right hemicolectomy with ileostomy is indicated in following situations:
    - Extensive peritoneal contamination.
    - Colonic vascular injuries.
    - Haemodynamically unstable patients.
    - Long-term hypotension after trauma.

LIVER INJURY

It can be subcapsular haematoma, lacerations, deeper injuries, lacerations with disruption of hepatic lobes or segments or liver injury with vascular injuries like of inferior vena cava or hepatic veins.

Present with features of haemorrhagic shock, distension of the abdomen, tenderness, rebound tenderness, guarding, rigidity.

- CT is diagnostic tool
- Liver injury is graded depending on involvement of hepatic veins, portal system, biliary system and duodenum
- Often high grade liver injury also can be managed nonoperatively
- *Push* (direct compression); *Pringle* (occluding portal triad at foramen Winslow with fingers temporarily); *plug* by embolisation; *pack* the liver bed; repair of vena cava or portal vein; stenting of biliary tree and hemihepatectomy—are the treatment strategies

Management

- Small tear is sutured.

  For larger tears:
    - Deep sutures.
    - Packing.
    - Debridement.
    - Haemocoagulants.
- Liver resection is not done (not advisable) usually for injuries.
- ***Pringle manoeuvre***—by compressing the porta near foramen Winslow—to control bleeding (not more than 30 minutes).
- Blood transfusions.
- Treatment of associated injuries like of diaphragm, lung, duodenum, colon.
- Antibiotics.

Complications of Liver Injury

- Haemorrhage, septicaemia, bile leak.
- Liver failure, haemobilia.
- Subphrenic abscess, CBD stricture.

SPLENIC INJURY

It can be subcapsular haematoma, laceration or hilar injury.

It can be associated with other organ injuries like left kidney, left lobe of the liver, splenic flexure of the colon or pancreas.

It can cause torrential haemorrhage and shock.

It is the most common organ injured in blunt injury abdomen.

Management

- U/S abdomen, diagnostic peritoneal lavage are the investigations.
- Blood transfusions.
- Splenorrhaphy is done in selected patients so as to save the spleen.
- Splenectomy.
- Management of associated injuries.

Complications of Splenectomy

- Left lung atelactasis.
- Overwhelming postsplenectomy infection (OPSI).
- Pancreatitis and pancreatic fistula.
- Gastric bleeding.
- Subphrenic abscess.

RENAL INJURY

- It is commonly managed conservatively.
- IVU is the investigation of choice in renal injury.
- Surgery is indicated when there is hilar injury, progressive bleeding, failure of conservative treatment or perinephric abscess formation.

URINARY BLADDER INJURY

Intraperitoneal bladder injury occurs in distended bladder. It is treated always by surgical exploration through transabdominal approach. Bladder tear is sutured with keeping a suprapubic cystostomy using Malecot's catheter.

Extraperitoneal injury can be treated conservatively by placing a Foley's catheter for 2-3 weeks.

ABDOMINAL COMPARTMENT SYNDROME

- Normal intra-abdominal pressure is 2-12 mmHg. Abdominal compartment syndrome is increased intra-abdominal pressure *more than 12 mmHg*. It is often sudden, rapidly progressive decreasing the venous return to heart.
- It is common in *multiple traumas*. Ileus, bowel oedema are the factors causing it. It is also seen in retroperitoneal haemorrhage, pancreatitis, long-standing hernia after reduction into the peritoneal cavity.
- Upward displacement of the diaphragm, increased peak inspiratory pressure, peripheral resistance, intrapleural pressure, CVP and PCWP; hypoxia, hypercapnia, acidosis; compression of IVC, decreased venous return to heart, cardiac output and right atrial pressure, decreased visceral and renal blood flow and glomerular filtration; mesenteric venous hypertension; bowel wall oedema and ischaemia—are the effects. Oliguria, respiratory failure, cardiac arrest ensures if abdomen is not decompressed.

Abdominal compartment syndrome (ACS)

Causes	***Features***	***Management***
Multiple trauma and ICU patients—common	Hypoxia, hypercarbia	Bladder pressure assessment
Postoperative ileus	Decreased urine output	Ryle's tube aspiration
Acute abdomen	Hypotension	Resuscitation
Acute gastric dilatation	Tense abdomen—distended	ICU care
Laparoscopic procedures	Decreased venous return	Surgical decompression
Intestinal obstruction	Bowel ischaemia	
	Cardiac arrest	

- Intra-abdominal pressure is measured using a *urinary catheter in the urinary bladder*. Pressure is graded *(Busch)* as I—10-15 cm of H_2O; II—16-25 cm of H_2O; III—26-35 cm H_2O; IV—more than 36 cm H_2O.
- Beyond grade III immediate decompression is needed. Initial *volume preload is essential* otherwise sudden decompression may cause cardiac arrest in asystole due to reduced preload, sudden influx of high potassium, acid and other metabolic by products into the heart.
- Condition is a *surgical emergency*.

Intra-abdominal pressure grading (Busch) in cm of water

I—10-15 cm of H_2O
II—15-25 cm of H_2O
III—25-35 cm of H_2O
IV—more than 35

SEATBELT INJURIES

- In an individual with seatbelt, during impact *violent deceleration* of human body occurs. Seatbelt impinges heavily on its point of contact with trunk and viscera continue to move forward. It leads into severe contusion of abdominal contents; *detachment of bowel from its mesentery* due to free forward rapid mobility of the bowel over a relatively fixed mesentery. Solid organ injury occurs only occasionally.
- *Two point anchorages* causes' solid organ injuries like of liver/spleen. *Lap-belt* causes contusion and bowel injury commonly.
- It is often difficult to identify the injuries due to presence of more obvious other injuries. CT chest and abdomen diagnostic peritoneal lavage (DPL) are very useful.
- Petechiae around iliac crest or costal margin are signs wherein one can suspect seatbelt injuries.
- Distraction fracture of lumbar spine (*chance fracture*) with hyperaesthesia of T12 and L1 level is often associated. 10% of such fractures are associated with intra-abdominal injuries.
- *Treatment* is immediate laparotomy and proceed—bowel suturing/resection/suturing of the organ injuries/splenorrhaphy/splenectomy.

K. Hand and Foot

CHAPTER OUTLINE

- Hand
- Hand Infections
- Acute Paronychia
- Chronic Paronychia
- Apical Subungual Infection
- Terminal Pulp Space Infection
- Infection of Web Spaces
- Deep Palmar Space Infection
- Space of Parona Infection
- Acute Suppurative Tenosynovitis
- Compound Palmar Ganglion
- Orf
- Milker's Nodes
- Hand Injuries
- Dupuytren's Contracture
- Volkmann's Ischaemic Contracture
- Syndactyly
- Mallet Finger
- Heberden's Nodes
- Spina Ventosa
- Foot
- Callosity
- Corn
- Plantar Fasciitis
- Ingrowing Toe Nail
- Onychographosis
- Onychomycosis
- Athlete's Foot
- Hallux Valgus

HAND

Surgical Anatomy of the Hand

Flexor Retinaculum

- It extends medially from pisiform and hook of hamate, laterally to scaphoid tubercle and trapezium crest as a strong fibrous band so as to bridge carpus to create a carpal tunnel.
- Ulnar nerve and vessels, palmar cutaneous branches of median and ulnar nerves, palmaris longus muscle are superficial to the carpal tunnel.
- Median nerve, tendons of flexor digitorum superficialis, profundus and pollicis longus, radial and ulnar bursa are deep to flexor retinaculum.

Palmar Aponeurosis

It is a thickened, modified deep fascia in the palm with its apex pointing proximally (as continuation of palmaris longus) and base distally which in turn gets divided into four parts. They extend over deep transverse ligament into lumbrical tunnel.

Blood Supply of the Hand

- Superficial palmar arch is mainly formed by ulnar artery and completed by superficial palmar branch of radial artery. It gives four digital branches to medial three fingers.
- Deep palmar arch is formed by radial artery and is completed by deep branch of ulnar artery. It gives three palmar metacarpal arteries which communicate with superficial palmar arch. It also gives communicating, perforating branches to dorsal metacarpal arteries.

Muscles of the Hand

- ***Thenar muscles***: Abductor pollicis brevis, flexor pollicis brevis, opponens pollicis and adductor pollicis.
- ***Hypothenar muscles***: Palmaris brevis, abductor digiti minimi, flexor digiti minimi and opponens digiti minimi.
- ***Lumbricals*** are four in number—named from lateral to medial.
- Four ***palmar interossei***.
- Four ***dorsal interossei.***

Nerve Supply

- Abductor pollicis brevis, flexor pollicis brevis, opponens pollicis and 1st and 2nd lumbricals are supplied by median nerve (5 muscles).
- Rest of the muscles in hand are supplied by ulnar nerve (15 muscles).

Modified Verdan Zone System in the Hand (Tendon Zones)

Zone I

From the fingertip up to the attachment of flexor digitorum superficialis (middle of middle phalanx). It contains tendon of flexor digitorum profundus.

Zone II

It begins proximal to metacarpophalangeal joint at distal palmar crease and extends up to the attachment of flexor digitorum superficialis at the middle of the middle phalanx. It is called as "***No-Man's-Land.***" Here flexors are tightly enclosed within a fibro-osseous tunnel. It is the ***most dangerous zone*** in hand injuries (***critical zone***).

Zone III

It begins at the distal end of flexor retinaculum (base of the palm) and ends at the transverse crease of the palm. It contains lumbricals attached to flexor digitorum profundus.

Zone IV

It begins at the proximal end of the flexor retinaculum and ends at its distal end.

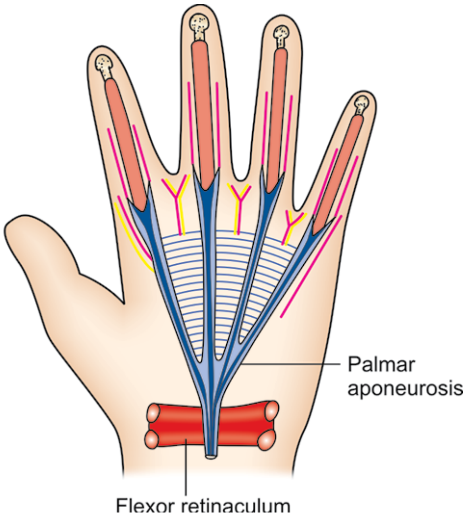

Fig. 1.281: Anatomy of flexor retinaculum and palmar aponeurosis.

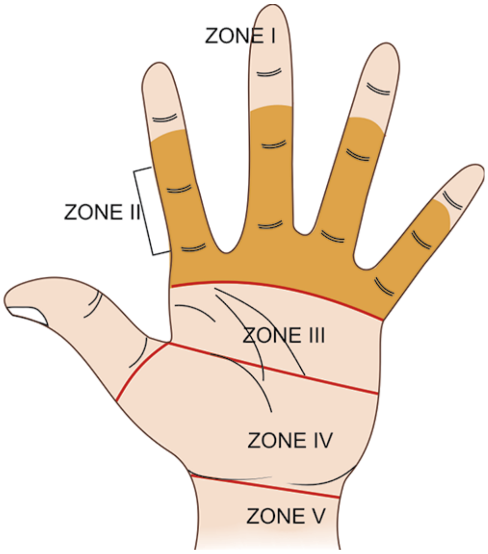

Fig. 1.282: Modified Verdan zone system in the hand.

Zone V

It extends from the proximal end of flexor retinaculum up to distal third of the forearm.

HAND INFECTIONS

Hand is a compact actively functioning unit. It contains neurovascular bundles, muscles, bones and ligaments.

Infection may be due to minor injuries or blood borne.

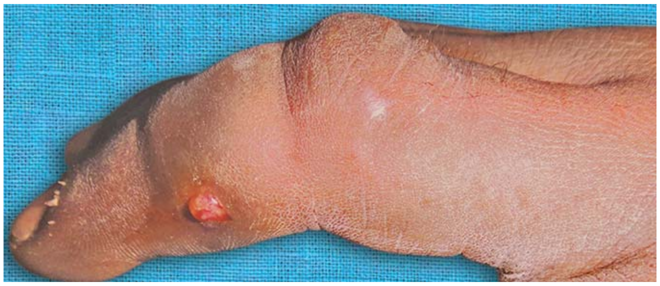

Fig. 1.283: Infection of finger at middle phalanx level. Note the oedema and position of the finger.

Precipitating causes

- Diabetes
- Immunosuppression
- Trauma
- HIV infection
- Steroid therapy
- Vascular diseases

Common organisms

- *Staphylococcus aureus—most common*
- *Streptococcus*
- Gram-negative organisms like *E. coli, Klebsiella, Pseudomonas*
- Occasionally fungal infection causing chronic paronychia, Madura hand due to Nocardia group of fungi, viral infection like *orf* can occur

General Features of Hand Infection

- Infection spreads faster in all areas.
- Causes oedema over the dorsum of hand due to lax skin and more lymphatic network even though infection per se is more over the volar aspect. It looks like ***frog hand.***
- Restricted movements of fingers and hand. The hand functions like ***hook***, ***pinch***, ***grip***, ***grasp*** are lost.
- Severe pain and tenderness, with fever.
- Tender palpable axillary lymph nodes are often present.

Different Types of Hand Infections

- Acute paronychia.
- Chronic paronychia.
- Terminal pulp space infection ***(felon)***.
- Subungual infection.

- Web space infection.
- Mid-palmar space infection.
- Thenar space infection.
- Deep palmar abscess.
- Acute suppurative tenosynovitis.
- Chronic tenosynovitis of flexor tendon sheath of palm and forearm—compound palmar ganglion.
- Lymphangitis of the hand.
- Arthritis of hand joints.
- Subcuticular abscess

Hand infection can be superficial or deep; it can be localised or spreading.

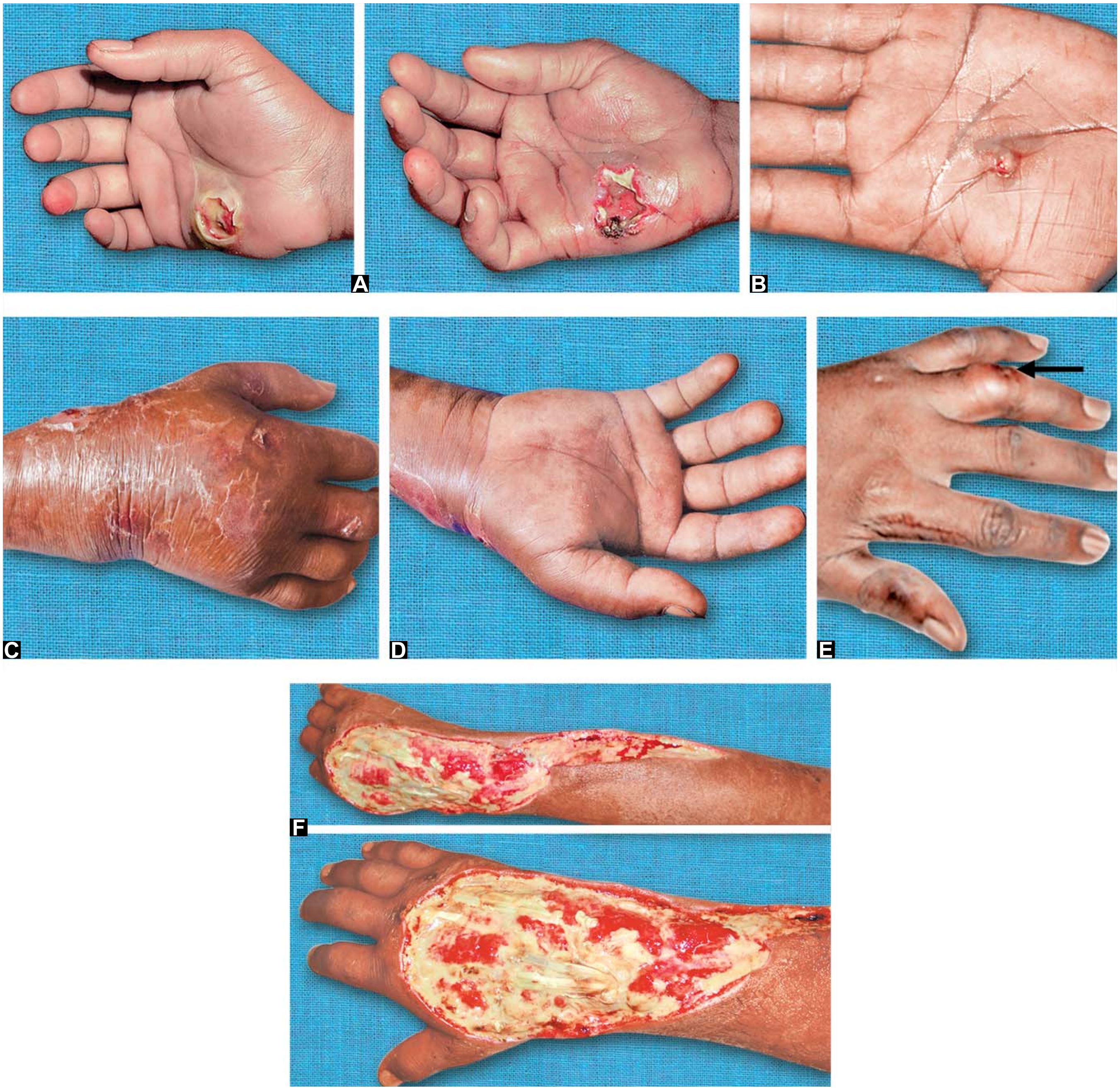

Figs 1.284A to F: Different types of hand infections. Note the oedema of hand even on dorsal aspect. Small infective focus can aggravate rapidly and so early proper drainage from deeper plane is important in managing the hand infections. Often it may cause extensive destruction exposing the tendons.

Investigations

- Pus for culture and sensitivity.
- Blood sugar.
- Urine sugar and ketone bodies.
- X-ray of the part.
- Arterial Doppler of the hand if needed.

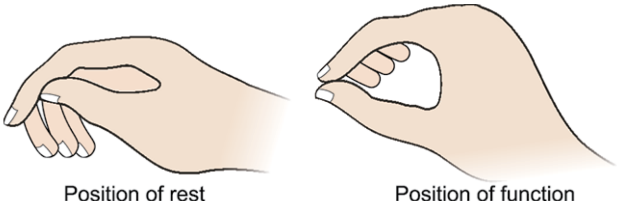

Fig. 1.285: Hand positions.

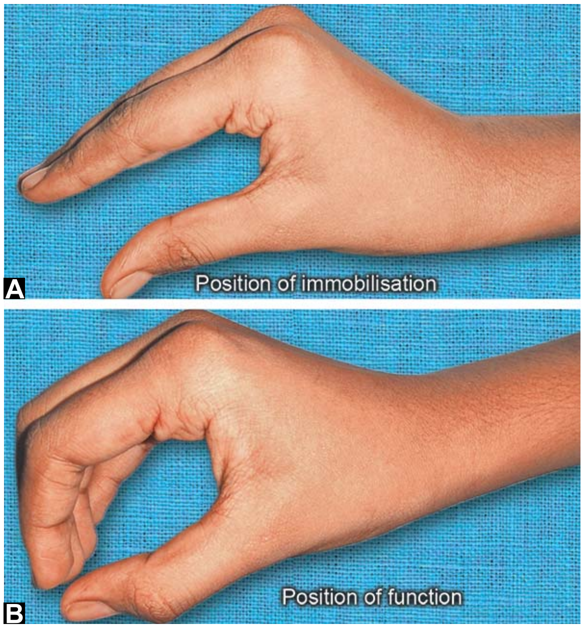

Figs 1.286A and B: Hand positions in immobilisation and function.

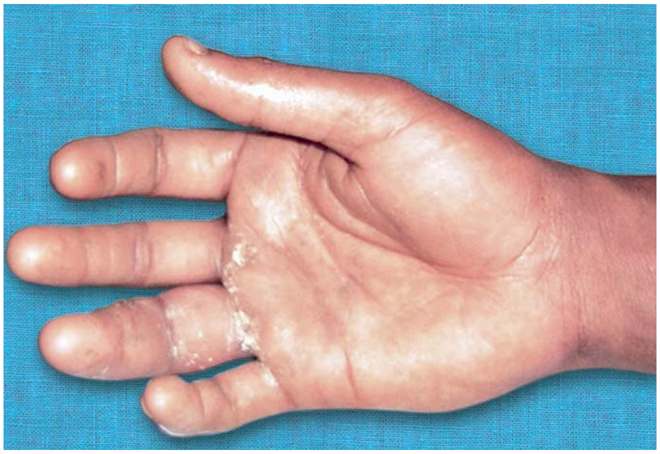

Fig. 1.287: Hand infection. Infection of ring finger extending into the palm.

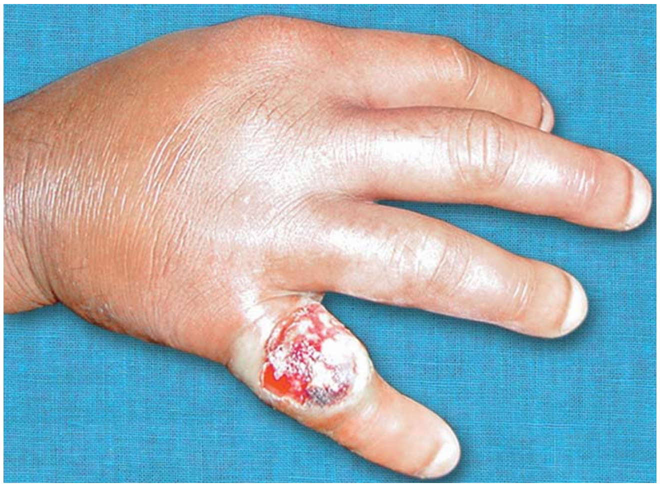

Fig. 1.288: Infection of little finger with dorsal oedema.

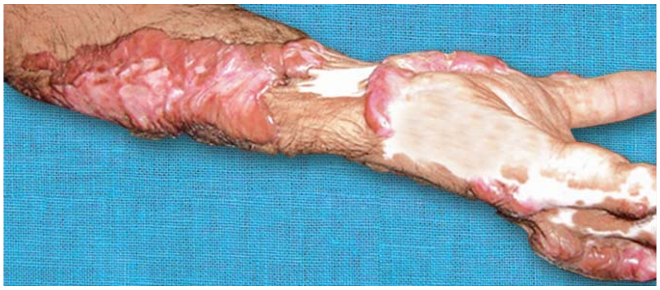

Fig. 1.289: Hypertrophic scar and keloid in hand and forearm after burn contracture.

General Principles of Managing Hand Infections

- Antibiotic therapy.
- **Position of rest** with wrist slightly abducted and extended, thumb and index fingers away (glass-holding position). *Position of function* is in which thumb and index fingers are pinching firmly with wrist extension.
- Elevation of hand reduces the oedema, increases perfusion, promotes healing.
- Early recognition of localised **pus.** Once localised, **Incision** and **Drainage** is done ideally under general anaesthesia or regional block **(not local anaesthesia).** Draining incision should not cross the palmar crease. Incision should have adequate length and adequate depth (deep to palmar fascia, otherwise evacuation of pus is inadequate). Care should be taken not to injure neurovascular bundles. Pus should be sent for culture and sensitivity. **Slough,** if present should be excised thoroughly. Gauze drain is placed. Regular dressings are done with continuation of antibiotics. Communications into other areas of hand should also be drained.
- **Bloodless field** (using tourniquet) is better to drain pus from hand.
- Proper measures must be taken after treatment. Initial rest, elevation of hand and later proper physiotherapy and regular exercise of hand and fingers are encouraged to restore normal function.

> *Early signs of Volkmann's contracture are pain, pallor, puffiness, pulselessness and paralysis.*
> **—David L Griffiths**

Complications of hand infections

- Stiffness of digits and hand (ankylosis)
- Deformity and disability
- Bacteraemia and septicaemia
- Osteomyelitis of bones depending on the location of abscess like metacarpal bones, terminal phalanx
- Suppurative arthritis of joints
- Paralysis of median nerve

Remember

- Hand should be flexible and strong; sensitive and pain free and coordinated to show all fine and powerful functions
- *Pinch* (picking a small object); *power grip* (holding a hammer); *key grip* (holding a key); *chuck grip* (holding a pen); *hook grip* (carrying a bag)—are the functions of hand
- Hand should be properly examined clinically for tendon functions; neurological problems—sensations (sweat test, two point discrimination test); for circulation (Allen's test); joint movements; examination of entire upper limb; opposite hand; axillary lymph nodes and other relevant systemic examinations
- Nerve conduction studies; electrophysiology; MRI hand; radioisotope bone scan; selective angiograms; X-ray hand are the relevant investigations other than systemic investigations
- Principles of treatment—elevation to reduce oedema; splintage to prevent contracture; early movements once inflammation subsides; early exploration of wound or surgical drainage of infective area; regional anaesthesia; usage of tourniquet; incisions when done across the flexor creases, should be at 45° angle

ACUTE PARONYCHIA

- It is the most common hand infection.
- It occurs in subcuticular area under the **eponychium**.
- Minor injury to finger is the common cause.
- Suppuration occurs very rapidly.
- It tracks around the skin margin and spreads under the nail causing **hang nail** or **floating nail**.
- Organisms are *Staphylococcus aureus* and *Streptococcus pyogenes*.
- Quantity of pus is very less around 0.5 ml but it should be drained to relieve sympotoms.

Clinical Features

Severe throbbing pain and tenderness (dependent throbbing) with visible pus under the nail root. Nail on touch is very tender (paronychia means "**Run around"**).

Treatment

- Pus is sent for culture and sensitivity.
- Antibiotics like cloxacillin, amoxycillin.
- Analgesics.
- The pus is drained by making an incision over the eponychium. Digital block using xylocaine 2% plain (*without adrenaline* as end artery supply to digits can develop arterospasm) is given as anaesthesia.
- If there is a floating nail, then the nail is dead and it has to be removed.
- Recovery is fast.

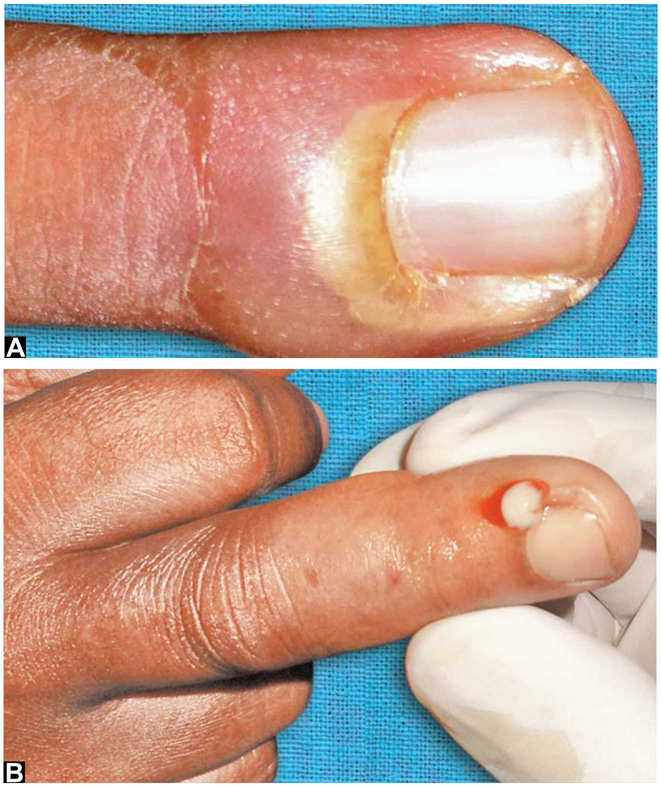

Figs 1.290A and B: Pointing pus in acute paronychia. Quantity of pus is very less usually around 0.5 ml.

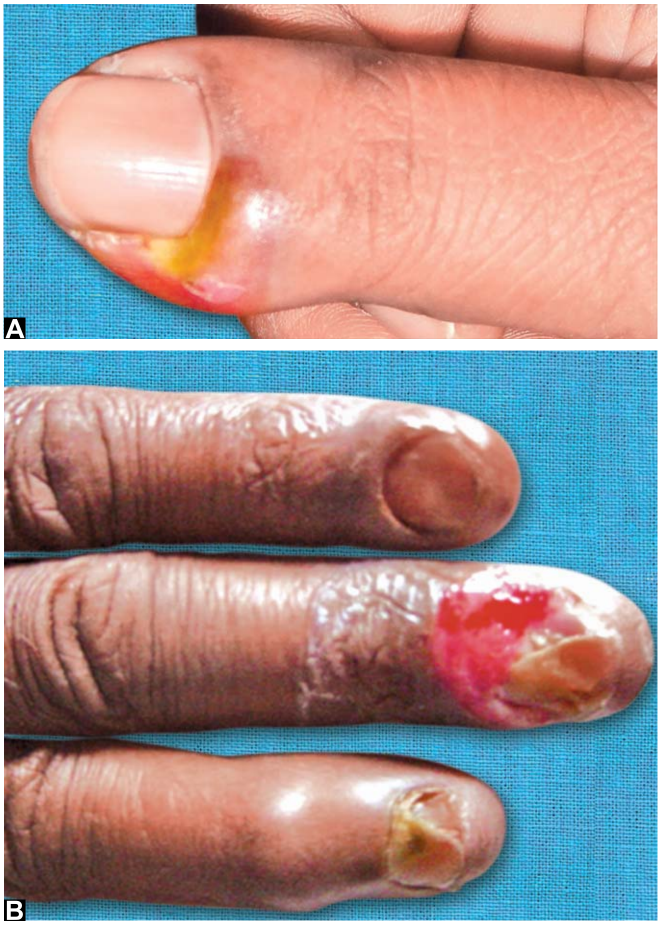

Figs 1.291A and B: Paronychia showing pointing pus in one picture and sloughed area granulating in another picture.

CHRONIC PARONYCHIA

It is commonly due to fungal infection—due to candida infection commonly.

Clinical Features

- It is common in females.
- Nail is diseased with ridges and pigmentation.
- Itching in the nail bed.
- Recurrent pain, discharge
- Secondary bacterial infection may supervene.

Investigation

Culture of scrapings for fungus and other causative agents.

Treatment

- Long-term antifungal therapy—local and systemic.
- Antibiotics for secondary infection.
- In severe cases removal of nail is required.

APICAL SUBUNGUAL INFECTION

- It is infection of the space between subungual epithelium and the periosteum.
- It occurs after minor trauma or rarely after formation of subungual haematoma.
- Beneath the free edge of the nail, pus comes to the surface.
- Excruciating tenderness with small visible pus under the tip (summit) of the nail is the feature.
- Drainage with 'V' incision over the summit is the treatment along with antibiotics.
- Osteomyelitis is not common.

TERMINAL PULP SPACE INFECTION (FELON)

- It is the second most common hand infection (25%).
- Index and thumb are commonly affected.
- Usually by a minor injury like finger prick.

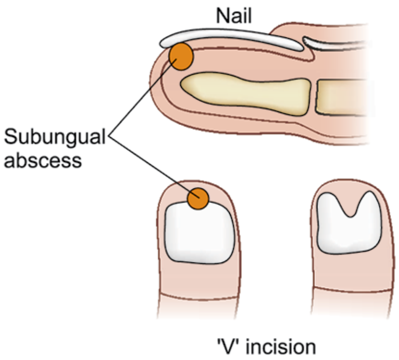

Fig. 1.292: Apical subungual infection and pus formation. Note the 'V' shaped incision to drain the same.

Surgical Anatomy

- Terminal pulp space contains fat and is partitioned by septae which is attached from periosteum of terminal phalanx to skin.
- Proximally deep fascia is attached to the periosteum distal to the base of terminal phalanx, i.e. distal to the attachment of flexor tendon.
- So, terminal space is a closed compartment, as the result of which pressure increases when there is infection, compressing terminal artery leading to thrombosis, resulting in osteomyelitis of terminal phalanx.

Bacteria

- *Staphylococcus*—most common.
- *Streptococcus,* Gram-negative organisms.

Clinical Features

- Pain, tenderness, swelling in the terminal phalanx.
- Fever.
- Tender axillary lymph nodes.

Often suppuration is severe, forming **collar stud abscess** which eventually may burst.

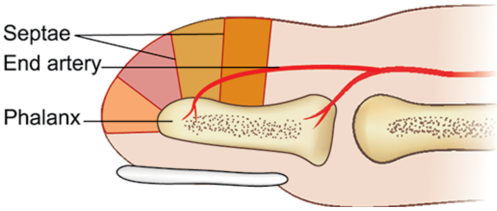

Fig. 1.293: Anatomy of the terminal pulp space.

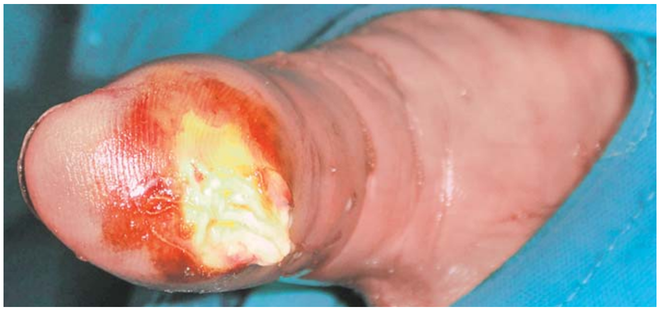

Fig. 1.294: Pulp space infection.

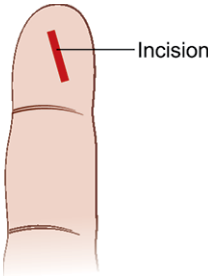

Fig. 1.295: Incision for pulp space drainage.

Dupuytren's contracture typically affects the ring finger and years later the little finger becomes implicated. In 30% cases little finger is primarily affected. **—Peter F Early**

Causes of collar-stud abscess

- Tuberculous cold abscess
- Terminal pulp space infection (Felon)
- Deep palmar space infection

Investigations

- X-ray of the part is required often to rule out osteomyelitis of terminal phalanx.
- Pus for culture and sensitivity.

Treatment

- Antibiotics and analgesics.
- **Drainage** of terminal pulp space by an oblique deep incision.
- If there is osteomyelitis of the terminal phalanx, it has to be amputated.

Complications

- **Osteomyelitis** of the terminal phalanx.
- **Pyogenic arthritis** of distal interphalangeal joint and tenosynovitis of flexor sheath.
- **Septicaemia**—in immunosuppressed individuals.

INFECTION OF WEB SPACES

Surgical Anatomy

There are three triangular web spaces filled with fat between the dorsal and volar skin. When the space is filled with pus it straddles the deep transverse ligament. Even though pus is volar, it points out dorsally.

It originates from:

- Abrasion.
- Infection of proximal volar space of finger.
- Callosities.
- Infection of proximal spaces.
- Trauma.
- Spread from other palmar spaces and from flexor sheaths through lumbrical canal.

Bacteria

- *Staphylococcus.*
- *Streptococcus.*
- Gram-negative organisms.

Clinical Features

- Fever.
- Pain and tenderness.
- Oedema of dorsum of hand.
- **Maximum tenderness** is on the volar aspect.
- **'V' sign—**separation of fingers.
- If untreated, infection may spread into other web spaces and hand spaces.

Treatment

- Elevation of hand.
- Antibiotics and analgesics.
- **Drainage** under regional or general anaesthesia. A **horizontal** incision is placed on volar skin of the web and deepened to reach the space by dividing fibres of palmar fascia. Pus is drained and sent for culture and sensitivity. If other web spaces are involved they should be drained through a separate incision. Edges of the wound are cut to leave a *diamond shaped* opening in front. Often counter-incision over dorsal skin of web is needed.

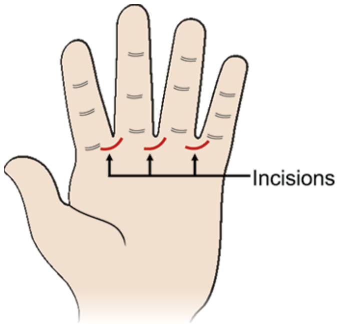

Fig. 1.296: Incision for web space drainage.

DEEP PALMAR SPACE INFECTION

Surgical Anatomy

Two deep palmar spaces are present

- **Midpalmar space.**
- **Thenar space.**

Midpalmar space is bound in front by palmar aponeurosis, behind by medial three metacarpals, laterally by a vertical line from lateral margin of the middle finger. It contains flexor tendons, neurovascular bundles and lumbricals. It is the common site of the infection.

Thenar space is located anterior to lateral two metacarpals. Infection here is usually due to extension from midpalmar space.

Midpalmar Space Infection

Causes

- Trauma.
- Spread from infection of finger spaces and web spaces.
- Haematogenous spread.
- Spread from tenosynovitis.

Clinical Features

- Pain and tenderness in the palm.
- Oedema of dorsum of hand **(frog hand).**
- Loss of concavity of palm.
- Painful movement of metacarpophalangeal joint (but interphalangeal joint movements are normal and painfree).
- Fever.

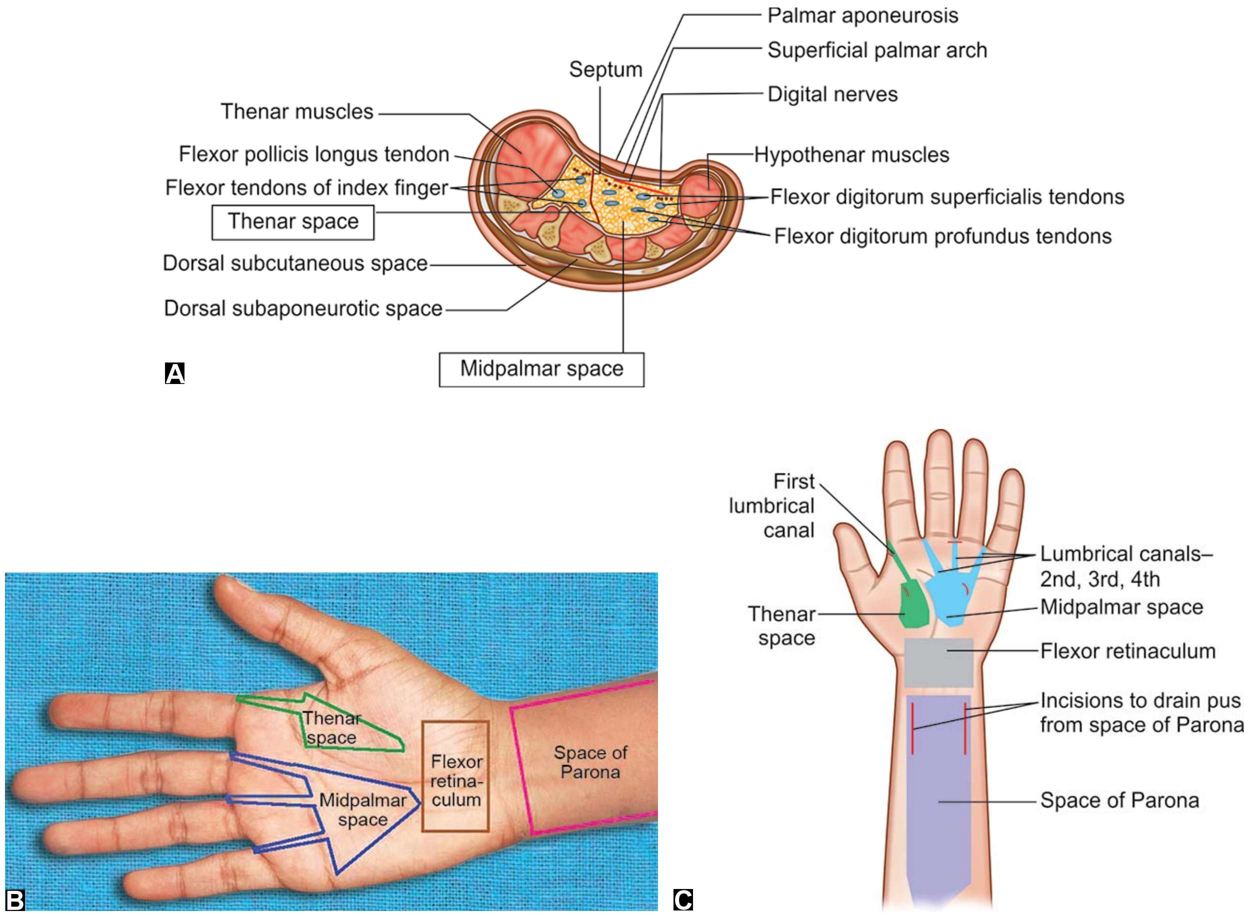
Palmar aponeurosis
Superficial palmar arch
Septum
Digital nerves
Thenar muscles
Flexor pollicis longus tendon
Hypothenar muscles
Flexor tendons of index finger
Flexor digitorum superficialis tendons
Thenar space
Flexor digitorum profundus tendons
Dorsal subcutaneous space
Dorsal subaponeurotic space
A
Midpalmar space
First lumbrical canal
Lumbrical canals–2nd, 3rd, 4th
Thenar space
Midpalmar space
Flexor retinaculum
Thenar space
Flexor retina-culum
Space of Parona
Incisions to drain pus from space of Parona
Midpalmar space
Space of Parona
B
C

Figs 1.297A to C: Anatomy of palmar spaces of the hand and forearm. Midpalmar space is on the medial aspect; thenar space is on the lateral aspect. *Space of Parona* is on the lower forearm.

- Palpable tender axillary lymph nodes.
- Eventually pus may come out of palmar aponeurosis forming collar stud abscess and later sinus formation.
  X-ray of the part is required.

Treatment

- Elevation of the affected limb.
- Antibiotics and analgesics.
- **Drainage:** It is *drained* under regional/general anaesthesia by placing horizontal/oblique incision parallel to the palmar crease. One should avoid crossing the crease line as much as possible. Palmar aponeurosis is carefully incised vertically to avoid injury of the neurovascular bundles. Alternatively one of the *interdigital web spaces* is incised horizontally*; lumbrical canal* (3rd or 4th) is opened to reach the deep palmar space. Pus is drained and sent for culture and sensitivity. Thorough saline irrigation is very essential. Drain is placed through the wound.

Complications

- ❖ Osteomyelitis of metacarpals
- ❖ Stiffness of hand
- ❖ Suppurative arthritis
- ❖ Extension of infection into other spaces

Thenar Space Infection

Thenar space (triangular shape) is located anterior to the lateral two metacarpals and fascia over transverse head of adductor pollicis; behind the short muscles of thumb, flexor tendons of index finger and 1st and 2nd lumbricals. Thenar muscles and flexor pollicis longus are lateral to it; fibrous vertical septum from palmar aponeurosis to 3rd metacarpal bone is medial to it. It is on the outer half of the hollow of the palm. Proximally it extends from flexor retinaculum; distally it extends to transverse palmar crease. It communicates to fascial sheath of 1st lumbrical. It is often associated with midpalmar space infection.

When a hand is seriously inflamed it takes up the position of greatest ease, which is, in fact, the position of rest.

— Frederic Wood Jone

It is *drained* similarly by placing incision on the lateral aspect of the palm or through the first web space incision is done along the *first lumbrical canal* on the radial side of the index finger. Often incision is made parallel to cleft between index and thumb on the posterior aspect.

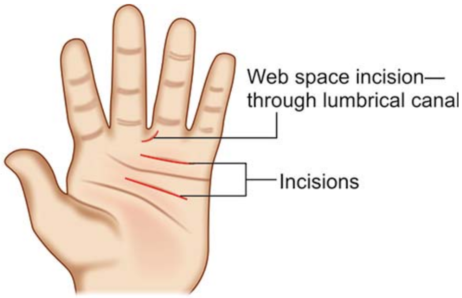

- Skin incision parallel to crease line
- Never cross the crease line
- Palmar aponeurosis is opened vertically
- Space can be approached through lumbrical canal through horizontal web space incision

Fig. 1.298: Incisions to drain midpalmar space infection.

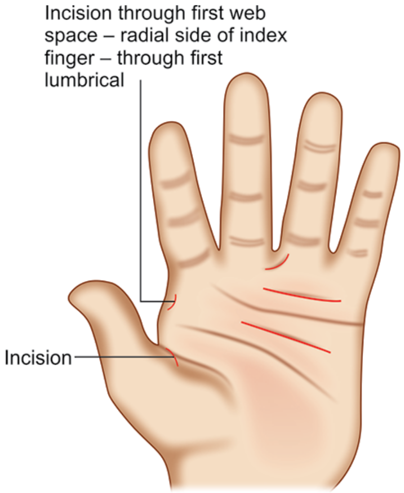
Incision through first web space – radial side of index finger – through first lumbrical
Incision

Fig. 1.299: Incision and drainage of thenar space abscess.

SPACE OF PARONA INFECTION

Forearm space of Parona is a rectangular space situated in the lower part of the forearm above the wrist, in front of pronator quadrates and deep to long flexor tendons. Above it extends up to oblique origin of flexor digitorum superficialis, below up to flexor retinaculum communicating with midpalmar space. Flexor tendon sheath proximally extends into this space. Pus in this space is drained through lateral incisions in the lower part of the forearm.

ACUTE SUPPURATIVE TENOSYNOVITIS

It is the bacterial infection of flexor tendon sheaths.

Surgical Anatomy

- **Radial bursa** is synovial sheath of flexor tendon of thumb which extends to the digit.
- **Ulnar bursa** is synovial sheaths of medial four flexor tendons of hand which extends into the digit of the fifth (little) finger.

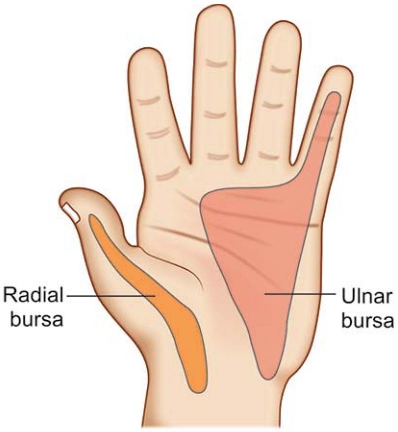

Fig. 1.300: Bursae hand.

- Extensor tendons are devoid of sheaths.
- Radial and ulnar bursa communicate with each other in 80% of cases.

Common bacteria: Staphylococcus aureus, Streptococcus pyogenes.

Clinical Features

- Symmetrical swelling of entire finger.
- Flexion of finger—**Hook sign**.
- Severe pain on extension.
- Tenderness over the sheath.
- Oedema of whole hand, both palm and dorsum (due to lymphatic spread).
- As ulnar bursa extends into the little finger its infection results in pain and tenderness extending up to little finger but not much to other fingers.

Kanavel signs

- Swollen finger held in flexion
- Exquisite pain on passive extension
- Tenderness precisely over the tendon sheath
- *Area of greatest tenderness over the part of ulnar bursa lying between transverse palmar creases*

- In infection of radial bursa thumb is swollen with pain and tenderness over the sheath of the flexor pollicis longus and there is inextensibility of interphalangeal joint.
- Swelling just above the flexor retinaculum is common.

Treatment

- Elevation of the affected limb.
- Antibiotics and analgesics.
- Position of rest.
- **Drainage** under general anaesthesia. Incisions are placed over the site of maximum tenderness and flexor sheath should be opened up. Many a times multiple incisions are required.

It is *drained through two incisions*—one over the proximal part of the sheath; other over the distal part of the sheath in the digit—along the crease lines. A fine catheter is passed into the sheath from proximal incision and irrigated with normal saline through this catheter. This catheter is left *in situ* for further regular irrigations, splinting of hand is necessary with boxing glove dressing.

Complications

- Spread of infection proximally into forearm—to *space of Parona*
- Stiffness of fingers and hand
- Suppurative arthritis
- Osteomyelitis
- Median nerve palsy
- Bacteraemia and septicaemia

COMPOUND PALMAR GANGLION

- It is chronic tenosynovitis of flexor tendon sheaths due to tuberculosis (tuberculous tenosynovitis) or rheumatoid arthritis.
- It can be unilateral or bilateral.

Pathology

- Flexor tendon sheath on either side of the wrist is involved, i.e. both in the volar surface of palm and lower forearm.
- Swelling contains fluid with typical ***melon seed bodies***.
- Condition is often bilateral in case of rheumatoid arthritis.

Clinical Features

- Swelling in the palm and lower forearm which is smooth, soft, nontender, fluctuant and also *cross-fluctuant* across flexor retinaculum, nontransilluminating.
- Wasting of hand and forearm muscles are seen.
- Matted axillary lymph nodes may be palpable.
- Primary focus may be present in lungs.

Investigations

- ESR, chest X-ray.
- FNAC of axillary lymph node and swelling itself.

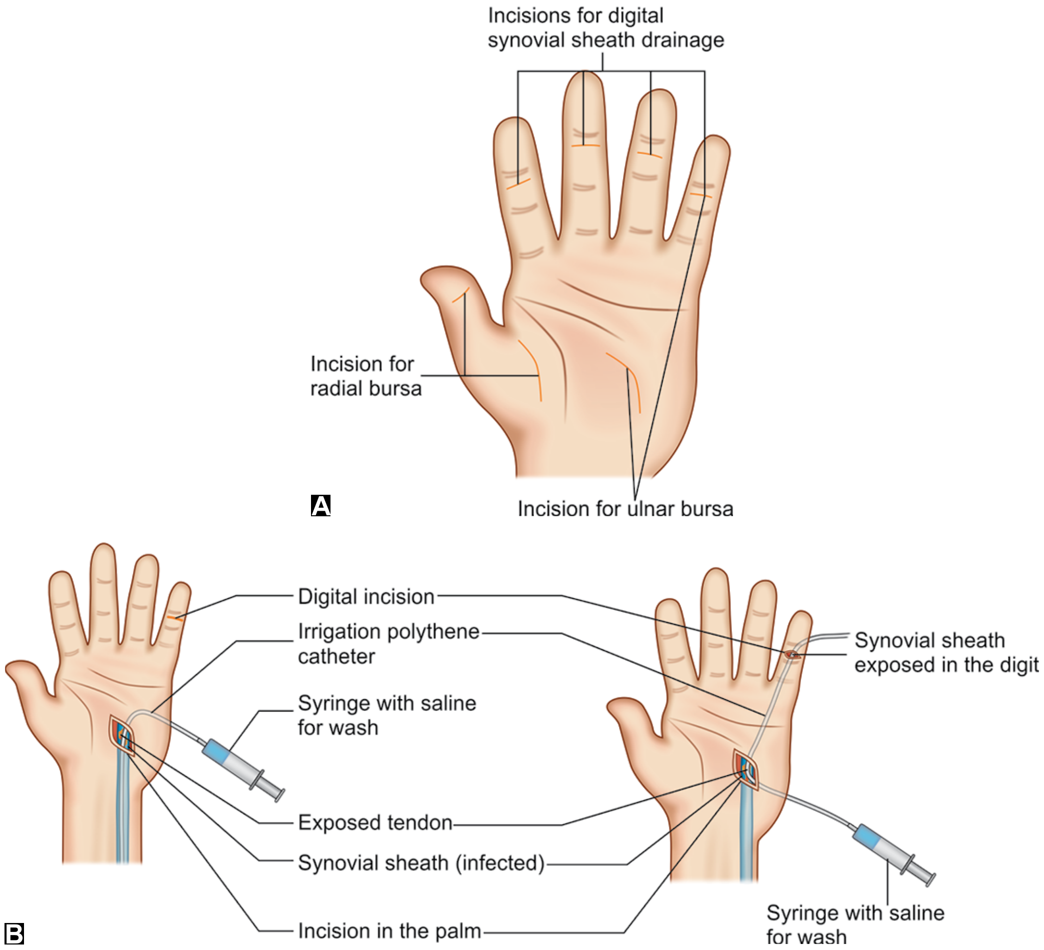
Incisions for digital synovial sheath drainage
Incision for radial bursa
Incision for ulnar bursa
A
Digital incision
Irrigation polythene catheter
Syringe with saline for wash
Exposed tendon
Synovial sheath (infected)
Incision in the palm
Synovial sheath exposed in the digit
Syringe with saline for wash
B

Figs 1.301A and B: Suppurative tenosynovitis is drained through incision at proximal part and another at digital sheath. Often by placing fine polythene catheter into the sheath, saline wash is given into the area.

Disease is the fate of poor, but also punishment of rich.— **Ivo Andrick**

Treatment

- Start antituberculous drugs: INH, rifampicin, ethambutol and pyrazinamide for 9 months.
- Excision of flexor tendon sheath is done along with scraping of caseating material, tubercles, melon seed bodies.
- Care should be taken not to injure median and ulnar nerves.

ORF

It is contagious pustular dermatitis of the hand due to a parapox virus infection.

MILKER'S NODES

It is similar viral infection of the hand seen in cow handlers.

HAND INJURIES

Classification

- ***Tidy injuries***: They are clean incised wounds and are usually treated by primary suturing but depends on the tissues involved like nerves, tendons and muscles.
- ***Untidy injuries***: They are lacerated wounds. Treated by debridement and later by delayed primary or secondary suturing.
- ***Compartment injuries.***
- ***Degloving injuries***
- ***Indetermined injuries*** which could not be assessed.

Assessment of Injury

It should include: Number, extent, depth, deformity and disability, neurovascular injuries, tendon injuries, muscle injuries bone and joint injuries.

Principles of Treatment

- Haemostasis.
- Use of tourniquet.
- Wound debridement and cleaning.
- Antibiotics and antitetanus treatment (toxoid and antitetanus globulin).
- Primary suturing if it is a incised wound or delayed primary suturing if there is odema.
- Skin grafting or flaps for skin loss.
- Tendon suturing or tendon graft for tendon injuries.
- Rest and elevation of the affected parts.
- Management of fractures by splint, wiring.

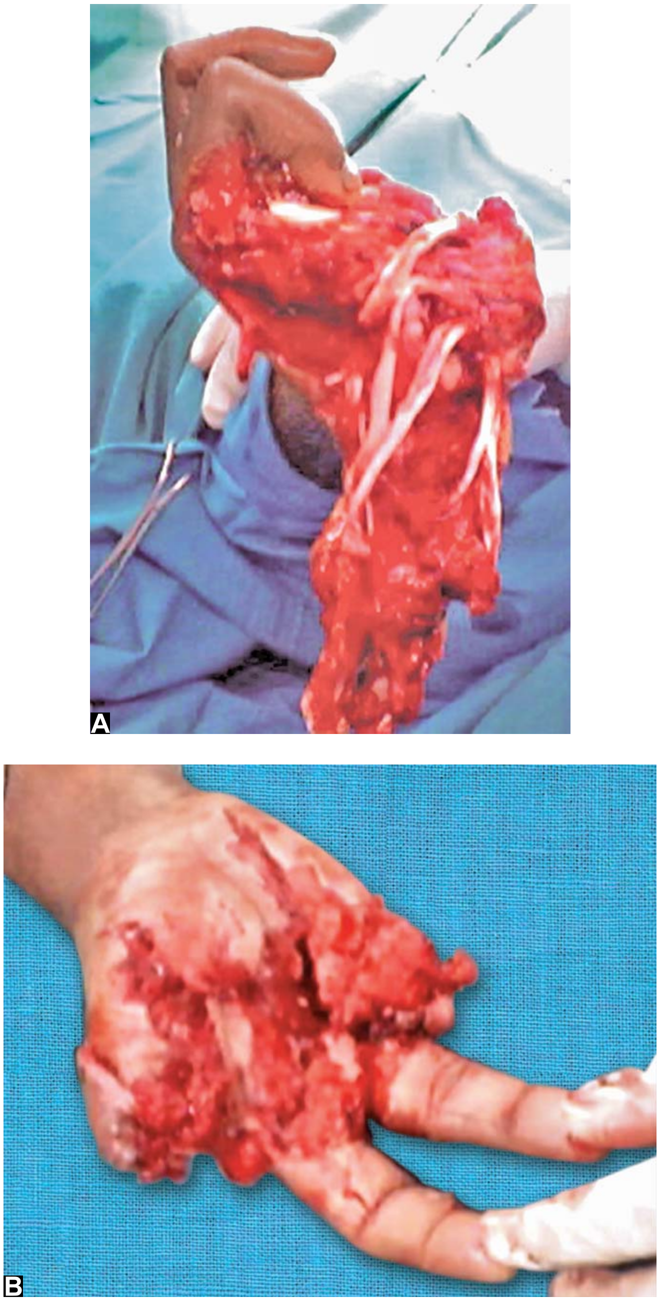

Figs 1.302A and B: Indeterminate and untidy hand injuries.

- Nerve repair for nerve injuries.
- Immobilisation up to 21 days.
- Later physiotherapy with warm, exercise, was bath active movements.
- Microsurgical restoration of digits. Reimplantation of the digits.

In hand

Do's	***Don'ts***
Do examine hand carefully	Do not incise every infected digit
Do think of other diagnosis	Do not make puncture incisions or over pads
Do wait for abscess to localise	Do not injure the digital nerves or vessels
Do place adequate length and depth of incisions	Do not place incisions crossing the crease line
Do immobilise, elevate the hand	Do not close human bites or lacerated wounds
Do give antibiotics and proper dressings	Do not forget to send pus for culture and sensitivity

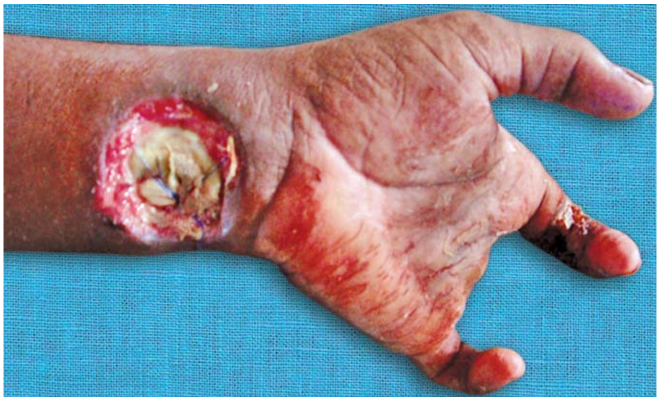

Fig. 1.303: Hand injury exposing tendons. Note the marker stitch in the tendon. It needs local transposition flap or groin flap to cover. Skin grafting is not possible over tendons.

- Amputation of digits or metacarpals only when inevitable.

Primary repair of tendons and nerves are of lesser priority in untidy injuries. Priority is wound debridement/ wound excision and early skin cover. Cut ends of nerves and tendons are tagged with coloured stitches for future identification purpose.

Complications and morbidity of hand injuries

- Infection
- Osteomyelitis
- Arthritis of joints
- Stiffness
- Loss of function due to disability

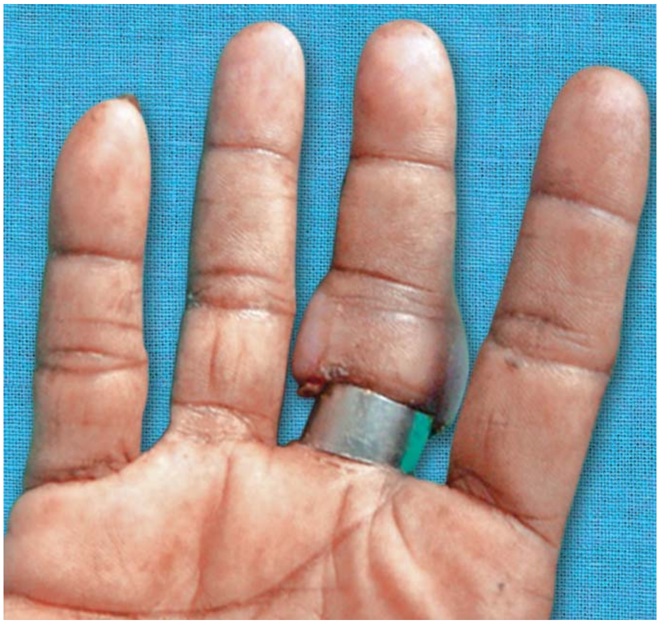

Fig. 1.304: Stuck finger by a ring. It is removed by applying soap, fat, and wax. String method is winding and unwinding a thread under and across the stuck finger. Sawing is done only when every method fails.

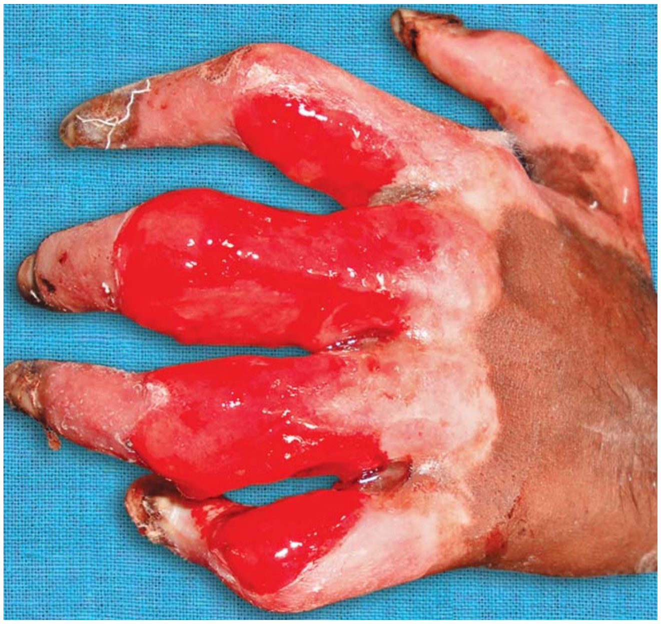

Fig. 1.306: Hand injury which is healing but with severe deformity of fingers.

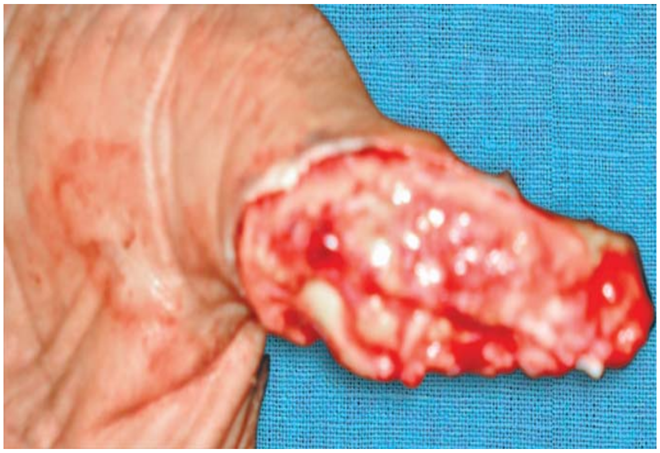

Fig. 1.305: Avulsion injury of thumb causing raw area. This needs proper skin cover to get the function adequately.

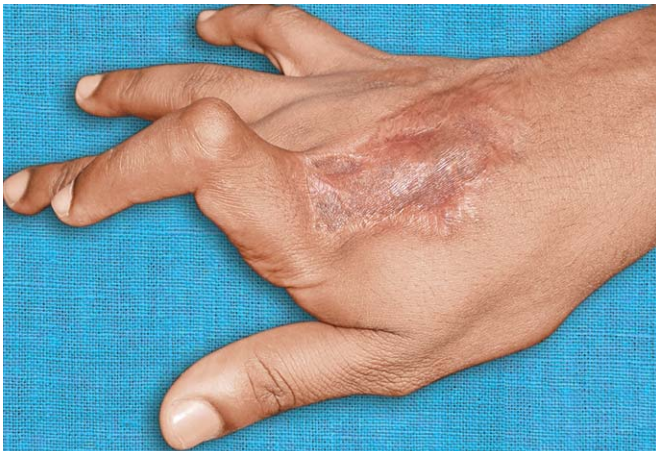

Fig. 1.307: Typical deformity of finger which needs correction for proper function.

"Impossible" is a word found only in the dictionary of fools.—Napoleon

DUPUYTREN'S CONTRACTURE

> *Individuals pre-disposed to the affection we are describing, observe that it is more difficult to extend the fingers of the affected hand… The first (interphalangeal joint) is flexed at nearly a right angle…the most powerful efforts are insufficient to extend it…. Hence, it was natural to conclude, that the commencement of the disease was in the unusual tension of the palmar aponeurosis.*
>
> —**Guillaume Dupuytren,** 1833

It refers to localised thickening of palmar aponeurosis and later formation of nodules with severe permanent changes in metacarpophalangeal and proximal interphalangeal joints. Terminal interphalangeal joint is *not involved* as palmar aponeurosis does not extend to terminal phalanx. It is common in *males* (10:1).

- It starts in ring and little fingers, with flexion of ring and little fingers. Later involving all fingers.
- There is thickening and nodule formation in the palm with adherent skin.
- It is often familial and bilateral 45%.
- Pads (of fat) develop in knuckles and are called as ***Garrod's pads*** (in proximal IP joints).

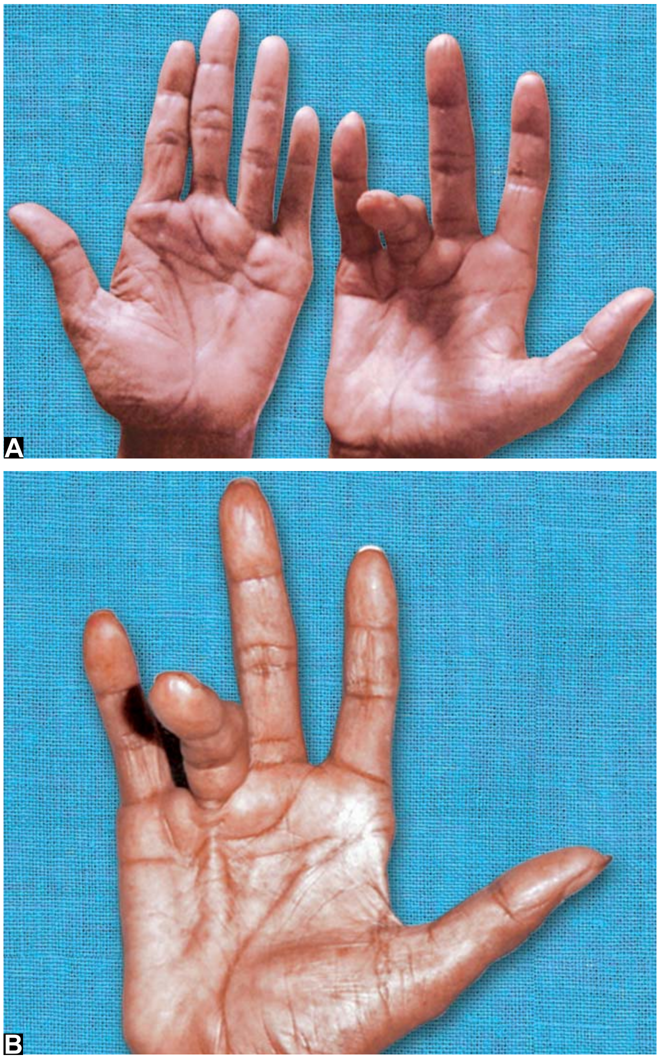

Figs 1.308A and B: Dupuytren's contracture in the hands. Note the involvement of the ring finger. Note the fibrous band on the other hand as early finding.

Conditions often associated with:

- Plantar fasciitis 5%—Ledderhose's disease
- Mediastinal and retroperitoneal fibrosis
- Peyronie's disease of penis 3%
- Nodules in the face and ear
- Pellegrini-Stieda's disease

Aetiology

- Repeated minor trauma, use of vibrating tools.
- Cirrhosis, alcoholism, smoking,
- Epileptics on treatment with phenytoin sodium.
- Diabetics, pulmonary tuberculosis, AIDS.
- Other metabolic conditions.
- Familial—autosomal dominant.

Galezia triad

- Dupuytren's contracture
- Retroperitoneal fibrosis
- Peyronie's disease of penis

Complications

- Restriction of hand function and so disability.
- Arthritis of MCP and proximal IP joints.

Treatment

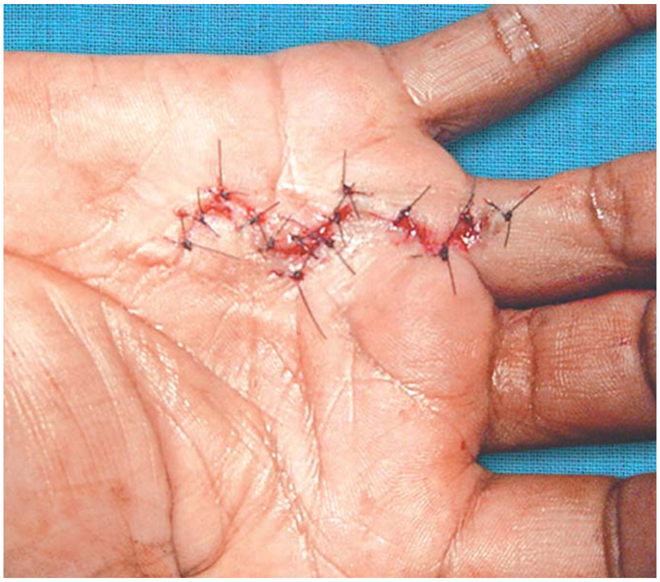

Fig. 1.309: Z plasty done for Dupuytren's contracture.

- Fasciotomy of palmar aponeurosis and later physiotherapy, Z plasty.
- In severe cases fasciectomy partial or complete.
- Treatment of the cause.

Recurrence can occur in 5-50% cases.

VOLKMANN'S ISCHAEMIC CONTRACTURE

It is a vascular injury leading to muscular infarction and subsequent contracture.

Causes

- Supracondylar fracture of the humerus.
- IV chemotherapy.
- Burns.
- Closed forearm crush injuries.
- Tight plaster after reduction of fracture.

Pathogenesis

Injury of brachial artery (tear, contusion, spasm, compression)

↙ ↘

Results in infarction of forearm flexor muscle | Injury to median nerve (mainly) and ulnar nerve both by ischaemia and infarction

↓

Aseptic muscle necrosis

↓

Fibrosis of flexor muscle of forearm

↓

Contracture

Clinical Features

Acute phase:

- Pain (persistent pain in forearm, hand, fingers—ominous symptom).
- Pallor.
- Puffiness (due to oedema).
- Pulseless (absence of radial pulse; but its presence does not rule out the onset of impending contracture).
- Paresis.

Late phase: Deformity

Deformity (due to injury to median nerve):

- Wrist joint extended.
- Extended metacarpophalangeal joints.
- Flexed interphalangeal joints.

Volkmann's sign:

In early stage, the fingers can be extended at the interphalangeal joints, only when the wrist is flexed fully. The fingers tend to flex if any attempt to extend the wrist is made.

Treatment

In acute phase:

- Removal of plastic cast applied after fracture reduction.
- Correction of fracture.
- Exposure of brachial artery and application of 2.5% papaverine sulphate to relieve the spasm if any.
- Suture of arterial tear if present, often with placement of arterial graft.
- Lateral incision over the deep fascia of forearm is placed to decompress the oedema.

In late phase (once deformity occurs):

- Physiotherapy
- Dynamic splints.
- ***Max-Page operation***: Release of flexor muscles (forearm muscles) from their origins from the bone and allowing it slide down until full extension.
- Excision of fibrous tissue and damaged muscles along with tendon transfer.
- Arthrodesis.

SYNDACTYLY

It is *webbing* or *fusion* of fingers.

Causes

- Congenital and hereditary—common.
- Traumatic like burns.

Types

- Cutaneous—simple.
- Fibrous.
- Bony—complex.
- It can be unilateral or bilateral.
- Often all four limbs may be involved with webbing of toes.
- It may be associated with polydactyly or visceral anomalies.
- If bony type is suspected, X-ray of the part should be taken.

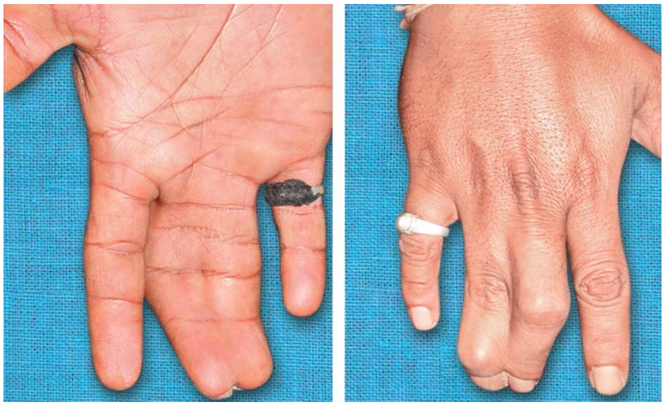

Fig. 1.310: Syndactyly.

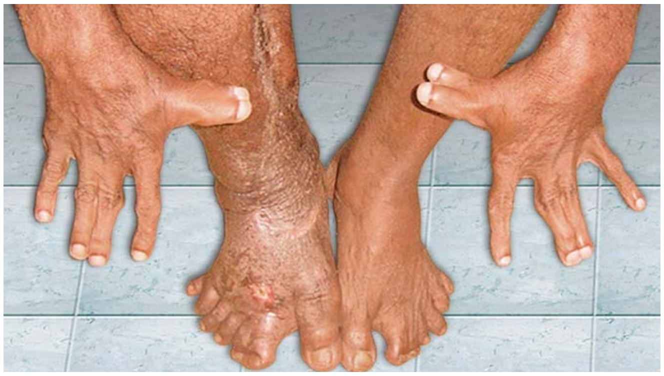

Fig. 1.311: Syndactyly and polydactyly of both hands and both feet.

You become successful by helping others to become successful.

Treatment

- If cutaneous, release of web is done as a staged procedure along with "Z" plasty or skin grafting.
- If fibrous, release can be done.
- If bony type, release is difficult because blood supply may be compromised which leads to gangrene of the digit.

MALLET FINGER (Base Ball Finger)

The terminal phalanx can not be extended because of tear at insertion of extensor tendon or avulsion fracture of the base of the terminal phalanx.

HEBERDEN'S NODES

These are seen in osteoarthritis, occurring behind the distal interphalangeal joints of ***index,*** middle, little and ring fingers.

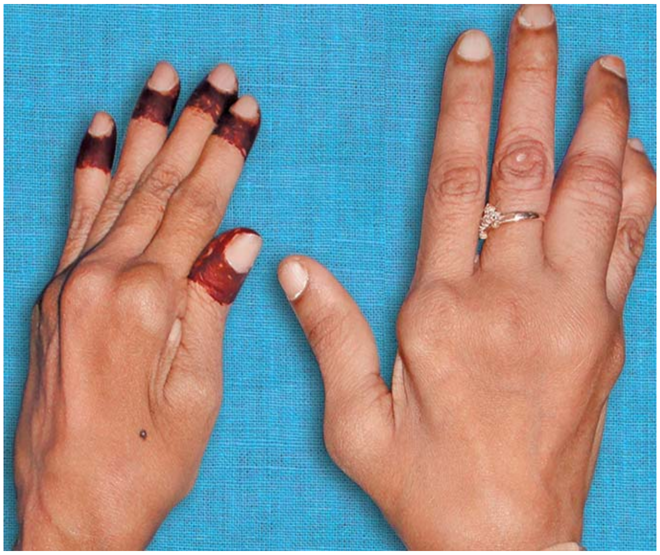

Fig. 1.312: Arthritis of joints of hands.

SPINA VENTOSA

Refers to phalangeal tuberculosis (***Tuberculous dactylitis***). It is called as spina ventosa because of its appearance as "air filled balloon".

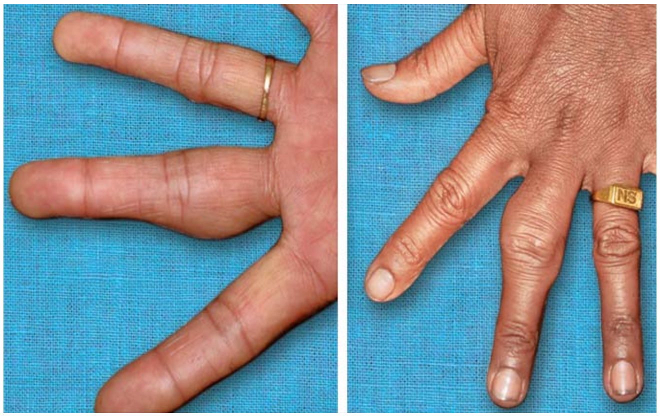

Fig. 1.313: Spina ventosa.

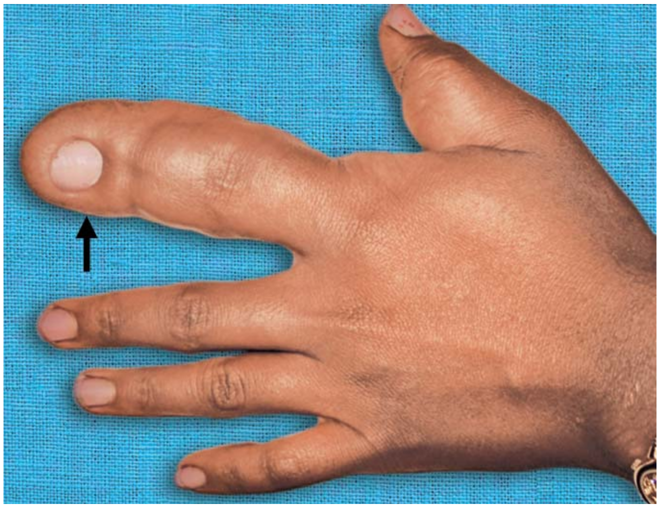

Fig. 1.314: Localised gigantism of index finger, thumb and part of the medial aspect of the hand.

FOOT

Foot contains 7 tarsal bones, 5 metatarsals, 14 phalanges (total 26 bones). Two sesamoid bones of 1st metatarsal bone are common. There are 4 layers of muscles in foot. Ligaments, muscles, joints, maintain the stability of foot complex. Blood supply is by anterior tibial, posterior tibial and peroneal arteries. Nerve supply is by saphenous, sural, posterior tibial, superficial and deep peroneal nerves.

CALLOSITY

- It is a hard, thickened skin occurs as a protective measure seen in wider area usually over heel and heads of metatarsals.

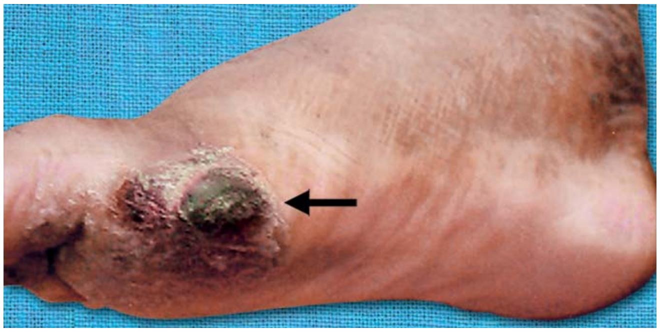

Fig. 1.315: Callosity in the foot.

- A callosity protrudes outwards from the skin.
- It is greyish-brown, raised, *protruded outwards*, thickened, hypertrophic skin occurs due to occupation and skeletal structure. It is *painless*. It is *wider* lesion. Paring the top layer exposes the shiny translucent dead skin beneath.
- It is grayish brown hypertrophic raised thick protective phenomenon which occurs commonly in areas of wear and tear like hands and feet. Top roughened layer when peeled off, shiny, translucent, homogenous dead skin can be exposed beneath.
- It is painless; it can get rubbed easily to create a sore.
- As it is a protective phenomenon it is best left alone.

CORN

Types

- Hard corn.
- Soft corn.

Hard Corn

- It is localised area of thickening over a bony projections like heads of metatarsals.
- Histologically it differs from callosity by having severe keratoses with a central core of degenerated cells and cholesterol.
- It presses over the adjacent nerves causing pain. It can get infected causing severe pain and tenderness with inability to walk.
- It is *smaller* lesion which is *pushed deep into* the skin forming a *localised* palpable *painful/tender* nodule with a *central yellow-white core* of dead cornified skin.
- Corn is common if there is deformity or by wearing tight fitting shoes/foot wears.
- Corn is narrow, deep and painful/tender.
- It is common in females.
- Corn is usually white/gray/yellow coloured, deep seated lesion.
- Infection, abscess formation and ulceration can occur especially if patient is diabetic.
- Corn may be associated with bursae causing bursitis.
- Corn often recurs after excision.

Treatment

- Excision.
- Local application of salicylic acid preparations or mixture of salicylic acid/lactic acid/collodion may be helpful. Skin softening agents are also tried.
- Eliminating the pressure is very important to prevent recurrence.
- Avoid excision of corn unnecessarily in diabetic (especially with neuropathy) and in ischaemic foot.

Soft Corn

It usually occurs between 4th and 5th toes due to friction of bases of adjacent proximal phalanges (See Fig. 1.318).

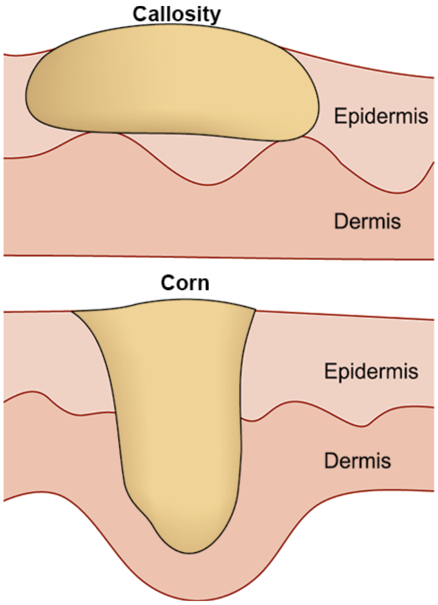

Fig. 1.317: Diagram showing differences between corn and callosity.

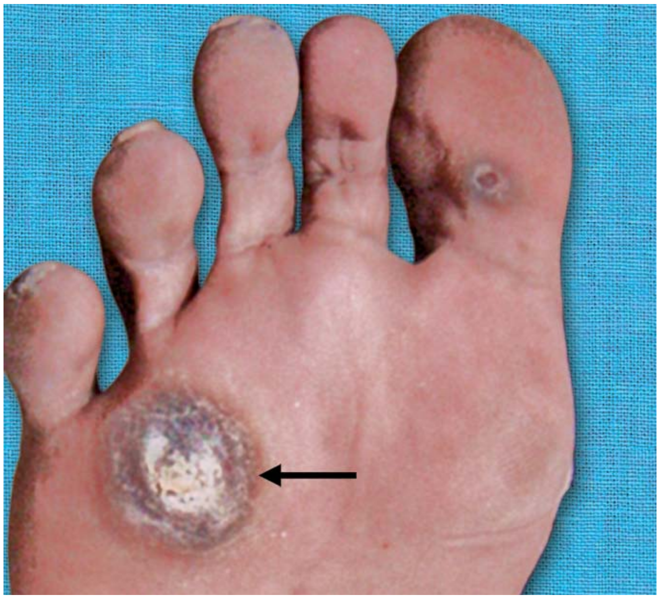

Fig. 1.318: Corns in the plantar aspect of the foot.

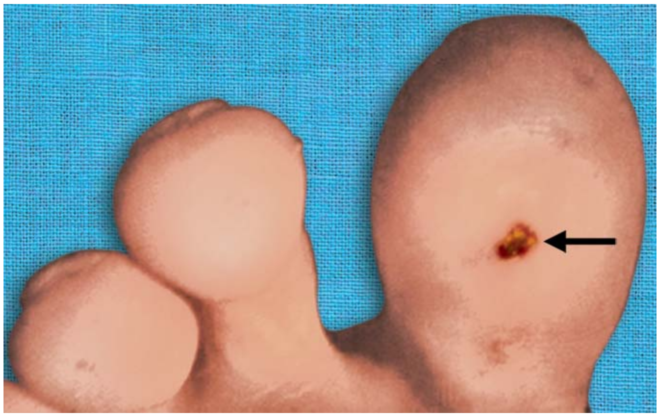

Fig. 1.316: Corn under great toe.

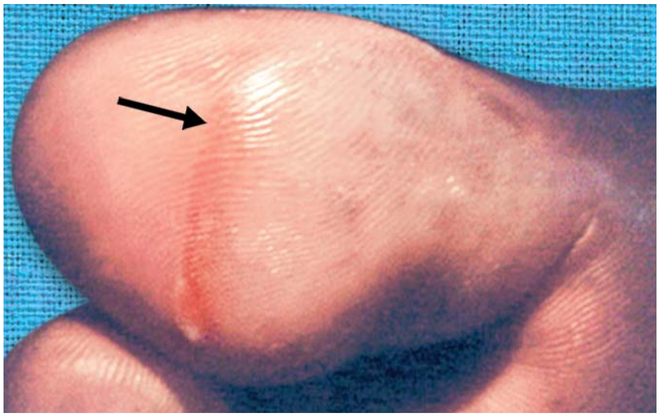

Fig. 1.319: Subcuticular abscess over great toe.

PLANTAR FASCIITIS (Policeman's Heel)

It occurs due to friction or tear of the ossified posterior insertion of the plantar fascia which is common in people who stand or walk for long-time.

Treatment

Analgesics, rest, steroid injections to the site.

INGROWING TOE NAIL (Onychocryptosis)

- It is also called as embedded toe nail.
- It is due to curling of the side of nail inwards, causing it to form a lateral spike resulting in repeated irritation and infection of overhanging tissues in the nail fold.

Causes

- Tight shoes.
- Improper cutting of nails (very short and convex).

Clinical Features

- It is common in ***great toe*** and is often ***bilateral***.
- Both medial and lateral sides of the toe can be involved.
- Recurrent attacks of acute and subacute paronychia occurs.
- Pain, tenderness, swelling of margins of the toe, often along with granulation tissue and foul smelling discharge.

Treatment

- Regular dressing and packing.
- Antibiotics. Discharge is sent for culture and sensitivity.
- Nails should be cut concavely or straight without leaving lateral spikes towards soft tissues.
- ***Zadik's*** or ***Fowler's operation***: Skin in lateral margin and root is incised so as to expose the lateral spike and germinal matrix. ***Infected tissues with pus and germinal matrix of ingrown part of the nail is excised or often entire nail with its root and germinal matrix is excised.***

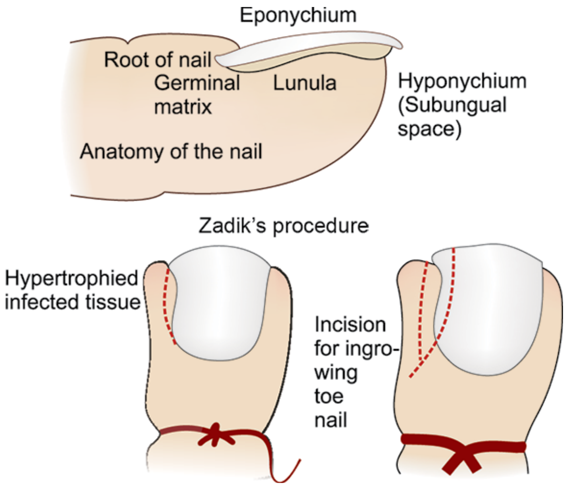
Eponychium
Root of nail
Germinal matrix
Lunula
Hyponychium (Subungual space)
Anatomy of the nail
Zadik's procedure
Hypertrophied infected tissue
Incision for ingrowing toe nail

Fig. 1.320: Zadik's or Fowler's operation.

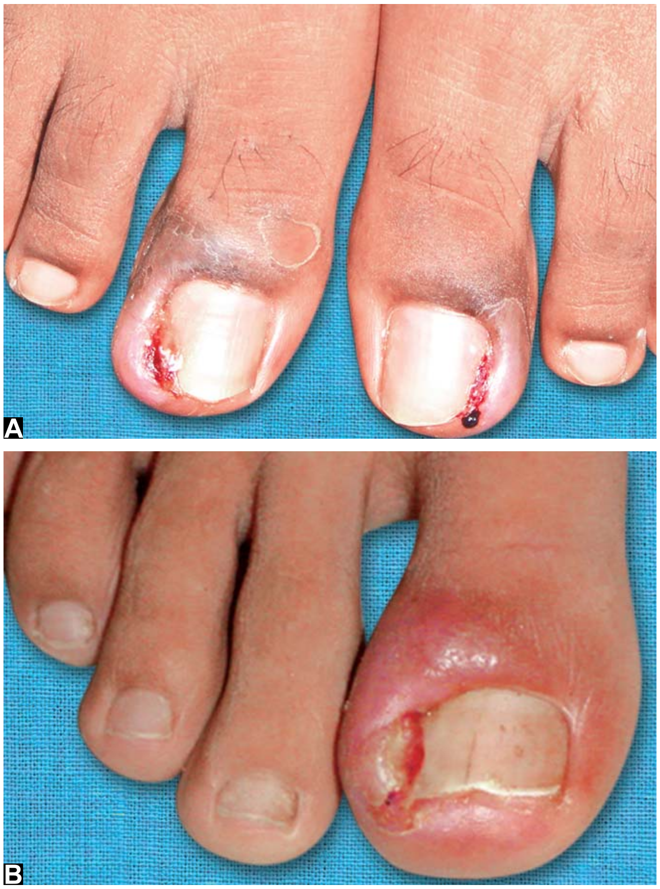

Figs 1.321A and B: Ingrowing toe nail. Note the granuloma caused by repeated infection and inflammation.

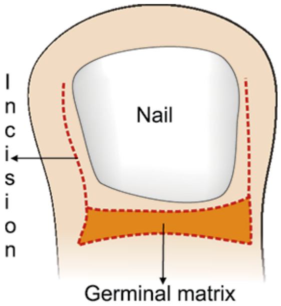
Incision
Nail
Germinal matrix

Fig. 1.322: Incision for nail excision. Note the germinal matrix. Removal of nail, entire or partial with adjacent germinal matrix is called as radical nail excision.

ONYCHOGRYPHOSIS

- It is curving of nail upwards (***Ram's Horn Nail***).
- It occurs due to repeated trauma or fungal infection.

ONYCHOMYCOSIS

It is fungal infection of the nail.

ATHLETE'S FOOT

- It is the fungal infection of the skin between the toes—*Tinea pedis.*

- Fungi enter through cracks; survive due to moisture in between toes.
- Skin is swollen, red, with sticky fluid, macerated with blisters.
- Itching, deep cracks, pain and discharge are common.

Treatment

- Part should be kept dry. Cotton, clean socks should be worn.
- Oral antifungals, antihistaminics and topical antifungals are used.
- Condition is contagious.

HALLUX VALGUS

- Here great toe is deviated laterally at first metatarsophalangeal joint. There is outward deviation of great toe with medial deviation of first metatarsal head.
- It may be due to persistent lateral force or occasionally hereditary.
- Condition is often *bilateral.*
- It is common in *females.*
- Thick walled bursa *(bunion)* over medial aspect of the head of the first metatarsal bone is common.
- Undue prominence of head of first metatarsal bone is typical often forming an *exostosis* at this point. *Osteoarthritis* of 1st metatarsophalangeal joint can occur.
- Lateral deviation of proximal phalanx over 2nd toe causing *crowding* of the toes.
- Initially it is painless; but eventually pain and tenderness develops with infection of bunion and *splaying* of forefoot.
- X ray shoes deviation with often osteoarthritis of the metatarsophalangeal joint.

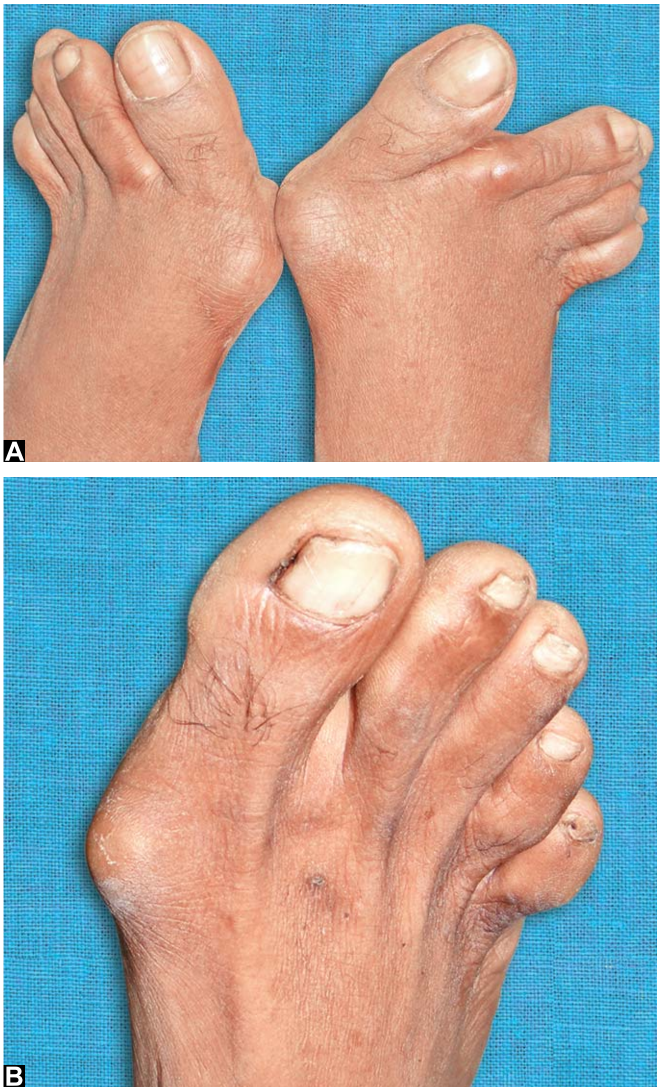

Figs 1.323A and B: Hallux valgus deformity.

Treatment

Keller's operation: Proximal 1/3rd of the proximal phalanx of great toe and medial part of head of 1st metatarsal bone is excised through medial curved incision. Soft tissue interposition is done.

Mayo's procedure: Medial part of base of the proximal phalanx of great toe and head of 1st metatarsal bone is excised – opposite of Keller's.

Simmond's procedure: Varus osteotomy at the base of 1st metatarsal bone with reinsertion of adductor hallucis tendon is done.

McBride procedure: Transfer of adductor hallucis tendon and lateral head of flexor hallucis brevis from proximal phalanx of great toe to the lateral part of head of 1st metatarsal bone.

Arthrodesis of metatarsophalangeal joint is done to relieve pain.

Excision of bunion, deformity correction, osteotomy, muscle transfers are also done as a **combined approach.**

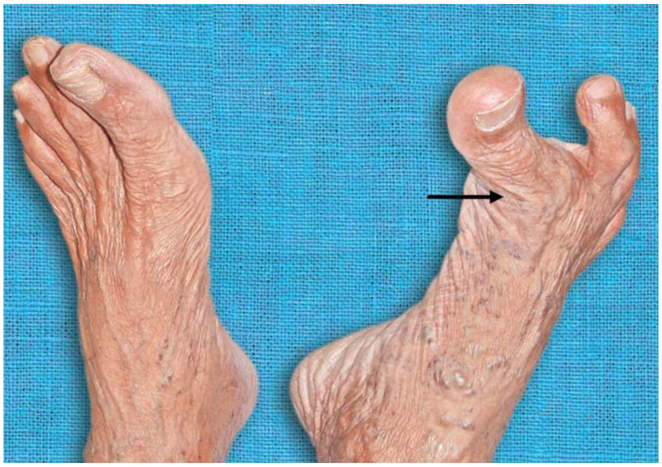

Fig. 1.324: Hallux varus deformity. It is opposite of hallux valgus deformigy.

Today well lived makes every yesterday, a dream of happiness and tomorrow, a vision of hope.

L. Arterial Diseases

CHAPTER OUTLINE

- Surgical Anatomy of Thoracic Outlet
- Arteries of Upper Limb
- Arteries of Lower Limb
- Arterial Diseases
- Intermittent Claudication
- Rest Pain
- Limb Ischaemia
- Pregangrene
- Gangrene
- Different Levels of Arterial Obstruction
- Other Features of Poor Circulation
- Investigations for Arterial Diseases
- Diseases of the Arteries
- Atherosclerosis
- Thromboangiitis Obliterans
- Takayasu's Pulseless Arteritis
- Raynaud's Phenomenon
- Temporal Arteritis
- Treatment of Arterial Diseases
- Subclavian Steal Syndrome
- Acute Arterial Occlusion
- Traumatic Acute Arterial Occlusion
- Embolism
- Reperfusion Injury
- Saddle Embolus
- Embolectomy
- Fat Embolism
- Air Embolism
- Therapeutic Embolisation
- Caisson's Disease or Decompression Disease
- Aneurysm
- Mycotic Aneurysm
- Abdominal Aneurysm
- Abdominal Aortic Aneurysm
- Peripheral Aneurysm
- Carotid Artery Aneurysm
- Dissecting Aneurysm
- Erythromelalgia/Erythralgia
- Livedo Reticularis
- Polyarteritis Nodosa
- Scleroderma/Systemic Sclerosis
- Acrocyanosis
- Gangrene
- Diabetic Foot and Diabetic Gangrene
- Trophic Ulcer
- Bedsores
- Frostbite
- Ainhum
- Endovascular Surgeries
- Upper Limb Ischaemia
- Arterial Substitutes

SURGICAL ANATOMY OF THORACIC OUTLET

Thoracic outlet is bounded by manubrium sternum in front, spine posteriorly, and the first rib laterally. At the superior aperture of thorax subclavian vessels, brachial plexus traverse the cervicoaxillary canal to reach the upper limb.

Cervicoaxillary canal is divided into proximal *Costoclavicular space* and *distal axilla (*divided by first rib).

Costoclavicular space is bounded superiorly by clavicle, inferiorly by first rib, anteromedially by the costoclavicular ligament, and posterolaterally by scalenus medius muscle along with long thoracic nerve.

Scalenus anticus muscle divides the costoclavicular space into two compartments, the anterior one containing subclavian vein and the posterior one containing subclavian artery and brachial plexus.

This posterior compartment is called as *Scalene triangle* bounded by scalenus anticus anteriorly, scalenus medius posteriorly, and the first rib inferiorly.

Cervical rib narrows this triangle and causes compressive features of the C_8, T_1 nerve roots and subclavian artery. Anything that narrows costoclavicular space causes *Thoracic outlet syndrome*.

Causes of thoracic outlet syndrome

- Cervical rib
- Long C_7 transverse process
- Anomalous insertion of scalene muscles
- Scalene muscle hypertrophy
- Scalene minimus
- Abnormal bands and ligaments
- Fracture clavicle or first rib
- Exostosis
- Tumours in the region

Differential diagnosis for thoracic outlet syndrome

- Carpal tunnel syndrome
- Cervical spondylosis
- Spinal canal tumours
- Shoulder myositis
- Angina
- Raynaud's disease
- Spinal stenosis
- Ulnar nerve compression

ARTERIES OF UPPER LIMB

Right subclavian artery begins from brachiocephalic trunk (innominate artery) whereas left subclavian artery arises directly from the arch of aorta. From underneath the sterno-clavicular joint artery arches over the pleura and apex of lung about 2.5 cm above the clavicle and then reaches the lateral border of first rib to continue as axillary artery.

Subclavian artery is divided into three parts by scalenus anterior muscle.

Axillary artery is divided into three parts by pectoralis minor muscle.

At the lower border of teres major muscle it enters the arm and continues as brachial artery.

About 2.5 cm below the crease of the elbow joint, it bifurcates into radial and ulnar arteries which run in the forearm.

Ulnar artery forms the superficial palmar arch which is completed by superficial palmar branch of radial artery.

Radial artery after passing through the anatomical snuff box enters the dorsum of hand and first intermetacarpal space to form deep palmar arch. It is completed by deep palmar branch of ulnar artery and is 1 cm proximal to superficial palmar arch.

ARTERIES OF LOWER LIMB

Abdominal aorta bifurcates at the level of fourth lumbar vertebra (corresponds to the level of the umbilicus in anterior abdominal wall) into two common iliac arteries.

Common iliac artery is about 5 cm in length passes downward and laterally; and at the level of lumbosacral intervertebral disc, anterior to sacroiliac joint, it divides into external and internal iliac arteries. Internal iliac artery supplies pelvic organs.

External iliac artery continues as common femoral artery at the level of inguinal ligament.

About 5 cm below the inguinal ligament common femoral divides into superficial femoral and deep femoral (Profunda femoris) arteries.

Deep femoral provides collateral circulation around the knee joint and also communicates above with gluteal vessels to maintain collateral circulation around the gluteal region.

Superficial femoral artery at the hiatus in the adductor magnus, continues as popliteal artery up to the inferior angle of the popliteal fossa where it divides into anterior and posterior tibial arteries.

Anterior tibial artery supplies anterior compartment of leg and ankle, continues as dorsalis pedis artery which forms dorsal arterial arch of the foot.

Posterior tibial artery supplies posterior compartment of leg and ends as medial and lateral plantar arteries which forms plantar arterial arch of the foot.

Posterior tibial artery gives peroneal artery which runs close to fibula supplying calf muscles.

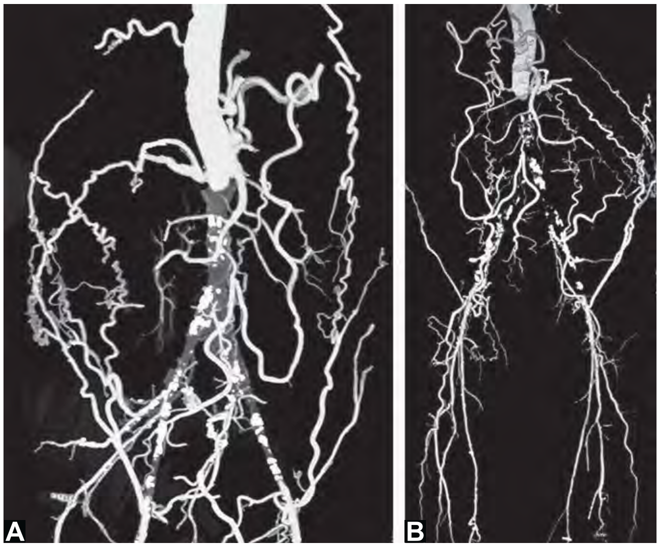

Figs 1.325A and B: CT angiogram of aortoiliac segment showing aortoiliac block due to atherosclerosis. Collaterals are also well-developed.

ARTERIAL DISEASES

- **Stenosis** due to trauma, atherosclerosis, emboli. It may be:
  - In the brain causing transient ischaemic attacks.
  - In the limbs causing claudication and rest pain.
  - In the abdomen causing pain, bloody stool.
  - In the kidney causing haematuria.
- **Dilatations are aneurysms.**
- **Arteritis**.
- **Small vessel abnormalities.**

INTERMITTENT CLAUDICATION

Claudio means *"I limp"* a Latin word. It is a crampy pain in the muscle seen in the limbs. Due to arterial occlusion, metabolites like lactic acid and substance P accumulate in the muscle and cause pain.

The site of pain depends on site of arterial occlusion.

- The most common site is calf muscles.
- Pain in foot is due to block in lower tibial and plantar vessels.
- Pain in the calf is due to block in femoropopliteal segment.
- Pain in the thigh is due to block in the superficial femoral artery.
- Pain in the buttock is due to block in the common iliac or aortoiliac segment, often associated with impotence and is called as **Leriche's syndrome.**

Pain commonly develops when the muscles are exercising. Cause for pain is accumulation of substance P and metabolites. During exercise increased perfusion and increased opening of collaterals wash the metabolites.

Boyd's classification of claudication

- Grade I: Patient complains of pain after walking, and distance in which pain develops is called as 'claudication distance'. If patient continues to walk, due to increased blood flow in muscle and opening of collaterals metabolites causing pain are washed away and pain subsides
- Grade II: Pain still persists on continuing walk; but can walk with effort
- Grade III: Patient has to take rest to relieve the pain

Claudication

- *Arterial*—typically develops after walking for certain distance and resolves rapidly within 5 minutes once walking is stopped
- *Neurogenic*—pain develops in standing or walking and disappears immediately after stopping walk; normal feeling pulses without ischaemic changes are present. It is usually due to narrow lumbar canal (spinal canal stenosis)
- *Venous*—it is rare but definitely occurs. It is observed in chronic pelvic venous obstruction as a mechanical high venous pressure. It is usually due to iliac vein thrombosis. Peripheral pulses are normal

Note:

- Beta blockers may aggravate claudication.
- Claudication is not that common in upper limb but can occur during writing or any upper limb exercise.

REST PAIN

- It is continuous aching in calf or feet and toes or in the region even at rest depending on site of obstruction.

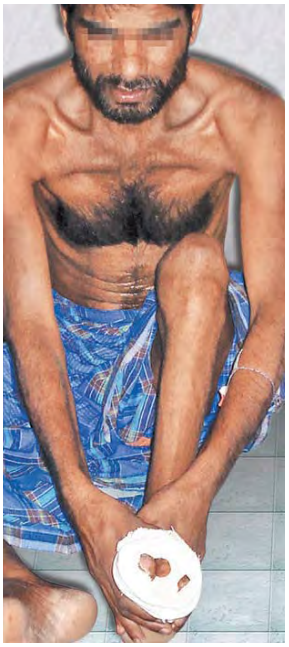

Fig. 1.326: Rest pain in a patient suffering from TAO. Holding the foot reduces the pain slightly.

- It is '*cry of dying nerves*' due to ischaemia of the somatic nerves. It signifies severe decompensated ischaemia. Pain gets aggravated by elevation and is relieved in dependent position of the limb.
- Pain is more in the distal part like toes and feet. It gets aggravated with movements and pressure.
- Hyperaesthesia is common association with rest pain.
- Rest pain is increased in lying down and elevation of foot; it may be reduced on hanging the foot down.
- Rest pain is worst at night and so patient is sleepless at night.
- Rest pain is apparently reduced by holding the foot with hand, probably due to suppression of transmission of pain sensation.

Features of arterial stenosis/block in limbs

- Intermittent claudication
- Rest pain
- Cold periphery, numbness, paraesthesia
- Color changes, ulceration, gangrene
- Altered sensation and decreased function/movements
- Diminished/absent arterial pulsation
- Thrill/bruit over the stenosed artery
- Altered venous filling—normally it is in few seconds; it is delayed in arterial stenosis; it is rapid AV fistula

LIMB ISCHAEMIA

Causes

- Atherosclerosis.
- Embolism (acute).
- Arteriopathies—Buerger's disease, Raynaud's disease, Takayasu's disease.
- Diabetes.
- Scleroderma.
- Physical agents—trauma, tourniquet, radiation injury.

Classification of Limb Ischaemia

Fontaine classification of limb ischaemia

- Stage 1: No clinical symptoms
- Stage 2: Intermittent claudication
  - 2a: Well compensated—can walk > 200 metres
  - 2b: Poorly compensated—walk only < 200 metres
- Stage 3: Rest pain
- Stage 4: Gangrene, ischaemic ulcer

Rutherford classification

Grade	*Clinical feature*
0	Asymptomatic
1	Mild claudication
2	Moderate claudication
3	Severe claudication
4	Ischaemic rest pain
5	Minor tissue loss
6	Major tissue loss

Limb ischaemia is also classified as

- *Functional* ischaemia.
- *Critical* ischaemia.

Functional Limb Ischaemia

Here flow of blood is normal when limbs are at rest; but will not be increased during exercise. It presents as claudication. It is defined as *"Muscle discomfort in the limb reproducibly produced by exercise and relieved by rest within 10 minutes."*

Critical Limb Ischaemia

It is persistently recurring ischaemic rest pain for 2 weeks, which requires regular analgesics for > 2 weeks or ulceration or gangrene of the foot or toes with an ankle systolic pressure < 50 mmHg or toe systolic pressure < 30 mmHg. Ankle brachial pressure index (ABPI) *will be less than 0.3*.

One should check blood pressure in all 4 limbs. ABPI is checked in supine position; systolic blood pressure in upper and lower extremities (two upper and two lower) is checked and higher value of each extremities is taken; ABPI is calculated.

Results—>**0.90 is normal; 0.70-0.89 is mild disease; 0.50-0.69 is moderate; <0.50 is severe.**

Note:
Calcification can alter the ABPI.

PREGANGRENE

It is the changes in tissue which indicates that blood supply is inadequate to keep the tissues alive and presents with rest pain, colour changes, oedema, *hyperaesthesia* with or without ischaemic ulceration.

GANGRENE

It is macroscopic death of tissue *in situ* with or without putrefaction.
(For detail see page no. 211, 212 and 213).

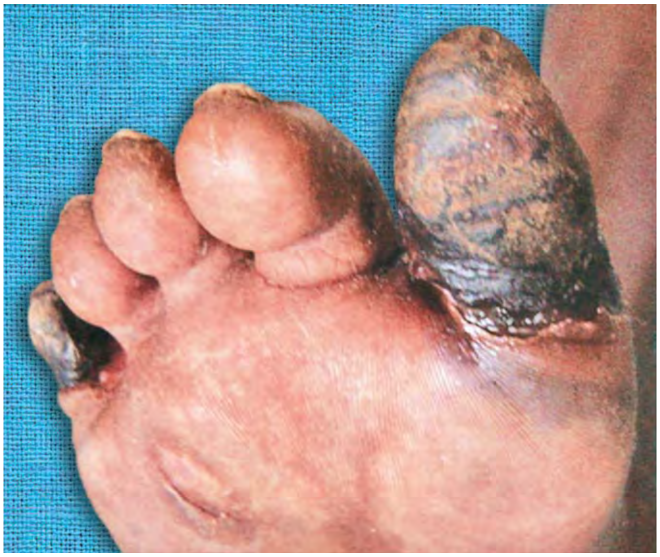

Fig. 1.327: Gangrene of great and little toes in a patient with vascular compromise.

Dry Gangrene

It is dry, desiccated, mummified tissue caused by gradual slowing of bloodstream. There *is a line of demarcation* and is localised.

Wet Gangrene

It is due to both arterial and venous block along with superadded infection and putrefaction. It spreads proximally and there is *no line of demarcation*. It spreads faster.

Organs in which gangrene can develop—appendix, bowel, gallbladder, testis, pancreas.

Necrosis*:* It is microscopic cell death.

Sequestrum is dead bone *in situ.*

Slough is dead soft tissue.

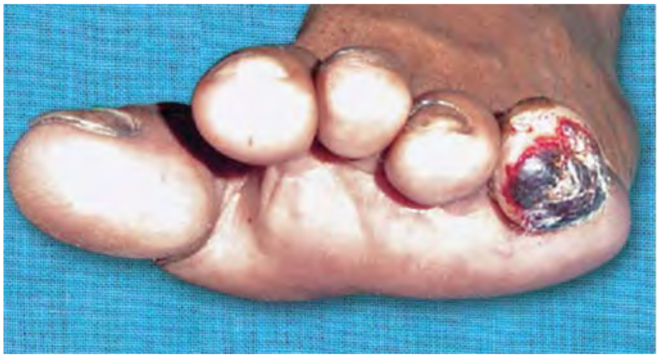

Fig. 1.328: Gangrene of little toe with line of demarcation.

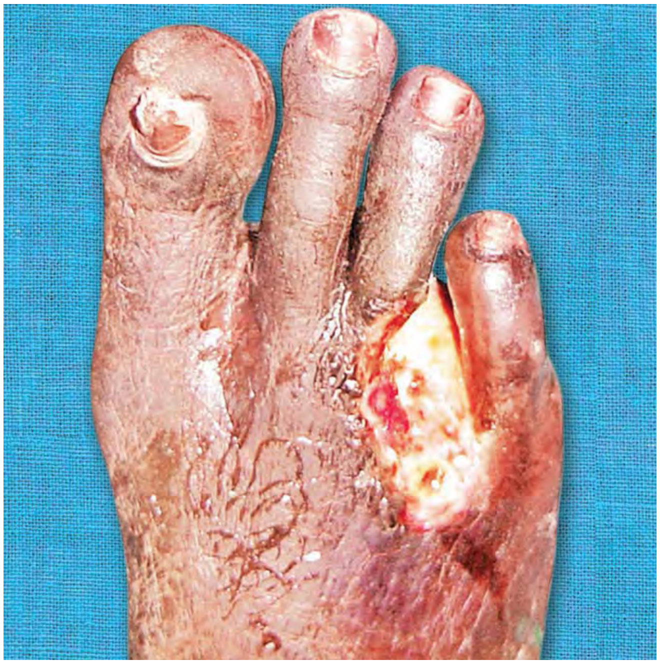

Fig. 1.329: Ischaemic ulcer foot in a diabetic patient.

Severely diseased distal aorta in atherosclerosis on arteriography is called as "shaggy aorta".

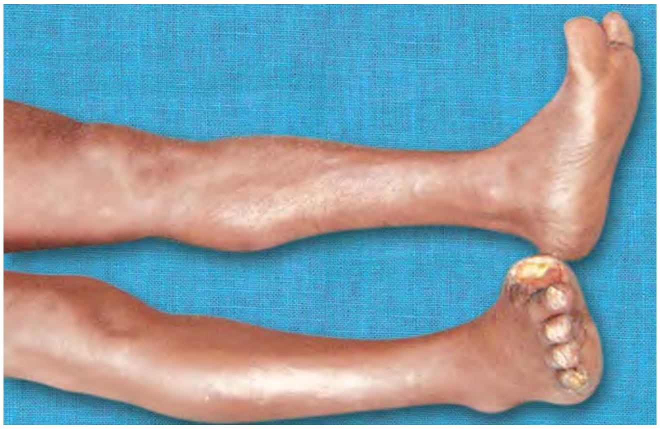

Fig. 1.330: Ischaemic right leg. Note the wasting and hair loss.

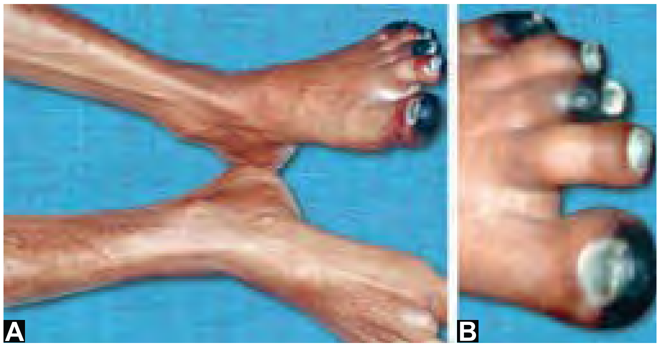

Figs 1.331A and B: Gangrene of all toes at their distal phalanges. All ischaemic features are obvious.

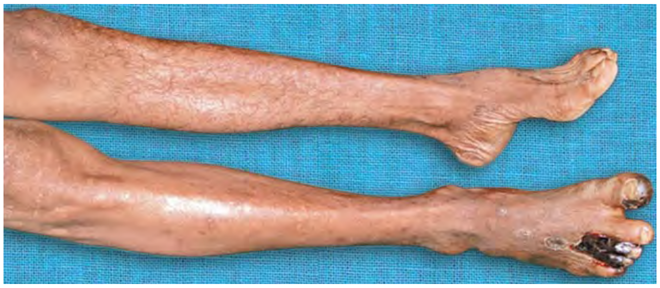

Fig. 1.332: Ischaemic changes in the right leg. Third and fourth toes are gangrenous with line of demarcation. Great and little toes are partly gangrenous. There are ischaemic features in the right foot and leg like hair loss/skin changes/wasting.

Line of Demarcation

- It is a line between viable and dying tissue indicated by a band of hyperaemia. It also indicates that disease is getting localised.
- *Final separation* between healthy and gangrenous tissue occurs by development of a layer of granulation tissue in between.
- It is hyperaesthetic due to *exposed nerve endings.*

Types of Separation

- Separation by *aseptic* ulceration is seen in dry gangrene.
- Separation by *septic* ulceration is seen in infected condition and *wet* gangrene.

Features of ischaemia

- Marked pallor, purple blue cyanosed appearance
- Thinning of skin
- Diminished hair
- Loss of subcutaneous fat
- Brittle nails, with transverse ridges
- Ulceration in digits
- Wasting of muscles
- Tenderness and temperature (cold)

DIFFERENT LEVELS OF ARTERIAL OBSTRUCTION

- *Aortoiliac block* causes claudication in both buttocks, thighs, and calves; absence of femoral and distal pulses bruit over aortoiliac region. *Impotence* occurs due to defective perfusion through internal iliac arteries and so into the penis causing erectile dysfunction (*Leriche's syndrome*)**.**
- *Iliac artery obstruction* causes claudication in thigh and calf; bruit over iliacs with absence of femoral and distal pulses.
- *Femoropopliteal obstruction* causes claudication in calf with absence of distal pulses but with palpable femoral.
- *Distal obstruction* shows absence of ankle pulses with palpable femoral and popliteal pulses.
  - *Paraesthesia* over the skin of the foot is due to shunting of blood from the skin to muscles in deeper plane.
  - *Sensation* in gangrenous area is absent. But, at the line of demarcation, skin is *hyperaesthetic*.

OTHER FEATURES OF POOR CIRCULATION

- The affected part is *cold* with *numbness*, *paraesthesia* and *colour changes*. Observing for *colour change* is important as, if it is not present one may have to think that numbness could be of neurological origin. On elevation the part blanches, on dependency part becomes purple. *Temperature sensation* on the ischaemic limb is lost apart from its coldness.
- *Ulceration, gangrene, decreased sensation and movements* are the features to be checked.
- **Delayed capillary filling**: Blanched nails or pulp of fingers, on pressure, will show delay in refilling (to turn pink) after release of pressure.
- **Delayed venous refilling**: Two fingers are placed over the vein. Finger nearest to heart is moved away so as to empty the vein. Distal finger is released to observe the venous refilling. Delay in filling is called *Harvey's sign*, signifies ischaemia. Venous filling is increased in AV fistulas.
- **Crossed leg test (*Fuchsig's test*):** Patient is asked to sit with the legs crossed one over the other so that the popliteal fossa of one leg will lie against the knee of other leg. Oscillatory movements of foot can be observed synchronous with the popliteal artery pulsation. This movement is absent with blockage of popliteal artery.
- **Disappearing pulse syndrome:** Exercise the limb after feeling the pulse. Pulse disappears once patient develops claudication. It is due to vasodilatation and increased vascular space occurring as the result of exercise wherein arterial tension can not be kept adequately and so pulse will disappear (*unmasking the arterial obstruction*).

- ***Buerger's postural test:*** Patient lying down on his back is asked to raise the leg above. In normal individuals, limb (plantar aspect of foot) remains pink even after raising above 90°. Ischaemic limb, when elevated shows marked pallor and empty veins. The angle in which pallor develops is called as *Buerger's angle of vascular insufficiency.* If this angle is < 30°, it indicates severe ischaemia.
- **Systolic bruit** may be heard over stenosed artery like subclavian artery, femoral artery, carotid artery, iliacs, renal artery.
- ***Adson's test (Scalene manoeuvre):*** In a patient sitting on a stool, the radial pulse is felt. The patient is then asked to take a deep breath (to allow the rib cage to move upwards so as to narrow the cervicoaxillary channel) and turn the face to same side (to contract scalenus anterior muscle so as to narrow the scalene triangle). If the radial pulse disappears or become feeble it signifies cervical rib or scalenus anticus syndrome.
- **Elevated arm stress test (EAST), or *modified Roos test*:** With both the arms kept in 90° abduction and external rotation position, patient is asked to make a fist and release repeatedly for 5 minutes. In normal side, patient will continue to do the manoeuvre whereas in diseased (Thoracic outlet syndrome) side patient gets pain and paraesthesia with difficulty in continuing the manoeuvre. Patient drops the arm down to relieve the symptoms.
- **Costoclavicular compression manoeuvre (*Falconer test*):** Radial pulse becomes absent when patient draws his shoulders backwards and downwards in excessive military position. This is because at this position, subclavian artery is compressed between first rib and clavicle, leading to feeble or absent radial pulse. *Halstead manoeuvre another similar text*.
- **Hyperabduction manoeuvre (*Wright test*):** When affected arm is hyperabducted, radial pulse becomes absent or feeble due to compression of artery by pectoralis minor tendon.
- ***Allen's test:*** It is done to find out the patency of radial and ulnar arteries. Both the arteries are compressed near the wrist and allowed to blanch completely in one minute (In the mean time patient closes and opens the fist several times for further venous outflow). Palm appears pale and white. One of the arteries is released and colour of hand is noted. Normally hand will become pink and flushed in no time; whereas in obstruction, the area will still remain pale. Other artery is also released and looked for changes in hand. Often test has to be repeated to get proper information.
- Abdomen should be examined for the presence of abdominal *aortic aneurysms*. It presents as pulsatile mass above the umbilicus, vertically placed, smooth, soft, nonmobile, not moving with respiration, resonant on percussion. Expansile pulsation is confirmed by placing the patient in knee-elbow position.
- Auscultation for *arterial bruit* over femoral artery, abdominal aorta, subclavian and carotid arteries is done.

Palpation of Blood Vessels

- *Dorsalis pedis artery* is felt just lateral to the extensor hallucis longus tendon at the proximal end of first web space, felt against the navicular and middle cuneiform bones. It is absent in 10% cases.
- *Posterior tibial artery* is felt against the calcaneum just behind the medial malleolus midway between it and tendon Achilles.
- *Anterior tibial artery* is felt anteriorly in the midway between the two malleoli against the lower end of tibia just above the ankle joint, lateral to extensor hallucis longus tendon.
- *Popliteal artery* is difficult to feel. It is palpated better in prone position with knee flexed about 40-50°, to relax the popliteal fascia. It is felt in the lower part of the fossa over the flat posterior surface of upper end of tibia. In upper end of the fossa, artery is not felt as there is no bony area in intercondylar region.
- *Femoral artery* in the groin is felt just below the inguinal ligament midway between anterosuperior iliac spine and pubic symphysis. Often hip has to be flexed for about 10-15° to feel it properly.
- *Radial artery* is felt at the wrist on the lateral aspect against lower end of the front of radius.
- *Ulnar artery* is felt at the wrist on the medial aspect against lower end of the front of ulna.
- *Brachial artery* is felt in front of the elbow just medial to biceps brachii tendon.
- *Axillary artery* is felt in apex of the axilla against shaft of the humerus.
- *Subclavian artery* is felt against first rib just above the middle of the clavicle.
- *Facial artery* is felt against body of mandible at the insertion of masseter.
- *Common carotid artery* is felt medial to sternomastoid muscle at the level of thyroid cartilage against carotid tubercle (*Chassaigne tubercle*) of transverse process of 6th cervical vertebra (in carotid triangle).
- *Superficial temporal artery* is felt just in front of the tragus of the ear against zygomatic bone.

INVESTIGATIONS FOR ARTERIAL DISEASES

- ***Segmental pressure measurements:*** Segmental BP is measured at multiple levels (upper and lower thigh, upper calf and ankle); pressure reductions between levels help to localise the occlusion; normally pressures increase as one moves further down the leg (>20 mmHg gradient abnormal); test is inaccurate in calcified artery walls.
- Blood tests: Hb%, blood sugar, lipid profile, peripheral smear, platelet count.
- ***Doppler*** to find out the site of block—hand held Doppler can be used (**Doppler**: *Christian Johann Doppler*, Austrian physicist).
  - *Duplex scan:* It is combination of B mode ultrasound and Doppler study. Difference in transmitted beam of the ultrasound and reflected beam is called as *Doppler shift* which is assessed and converted into audible signals. It is used to study the site, extent, severity of block, and also about collaterals. Audible sound is heard with normal flow, and sound is important. Turbulence is heard when there is stenosed partially blocked artery. Audible sound will be absent if there is complete block. Using Doppler probe blood pressure at various levels can be assessed. Pulse wave tracing along the artery is also important.

Half of us are blind, few of us feel and we are all deaf.—**William Osler**

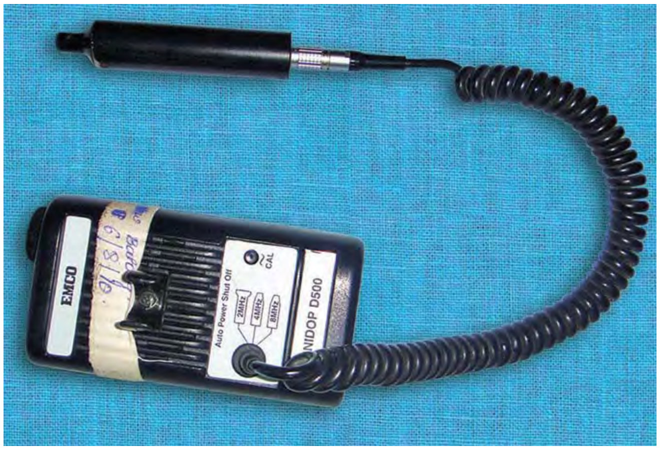

Fig. 1.333: Hand held Doppler is simpler way to identify and assess the artery.

  - Arterial diameter, blood flow rate, velocity of flowing blood, assessment of stenosed segment is properly done using Doppler.
- ***Treadmill test/ECG/echocradiography*** to assess cardiac/coronary status.
- ***Ultrasound abdomen*** to see aneurysm/aorta and its anatomical changes/other vessels in the abdomen/other organs.
- ***Plethysmography:*** It measures the blood flow in limbs. Water filled volume recorder; air filled volume recorder; mercury in silastic gauze is used after occluding the venous outflow. It is a noninvasive method. Segmental plethysmography using occlusion cuffs of 65 mmHg pressure is placed at thigh, calf and ankle levels and then quantitative measure of pulsation is done.
- ***Ankle-brachial pressure index***: Normally it is 1. If it is less than 0.9, it means ischaemia is present. If it reaches 0.3 or below then it signifies severe ischaemia with gangrene.
- ***Angiography:***
  - *Angiography* is *the appropriate* investigation for arterial diseases.

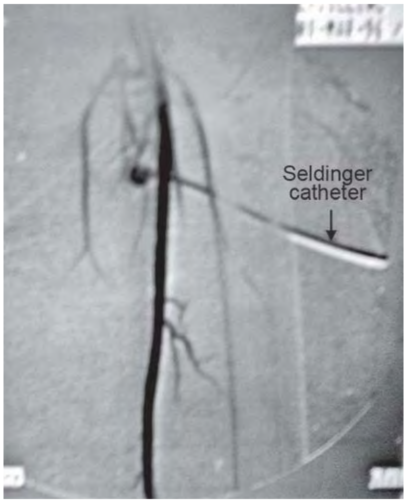

Fig. 1.334: Seldinger angiogram—retrograde femoral approach. (*Courtesy:* Dr Vasudeva Rao, Vascular Surgeon, Manipal Hospital, Bengaluru).

  - *Retrograde transfemoral Seldinger* (Sweden) *angiography*:
    - It is commonly done angiogram. It is done only when femorals are felt. If femoral pulsation is not felt then angiogram is done either *transbrachially (left brachial artery),* or *transaortic.*

Indications for angiogram

- TAO
- Atherosclerosis
- Raynaud's phenomenon
- AV fistulas
- Haemangiomas
- Thoracic outlet syndrome (e.g. cervical rib)
- Aneurysms
- Neoplastic conditions

    - Other angiograms are carotid angiogram, coeliac angiogram, superior mesenteric angiogram, coronary angiogram.

Seldinger technique (Steps)

- Arterial cannula is passed into the artery, e.g. femoral artery
- Needle is removed and guidewire is passed through the cannula
- Cannula is removed
- Dilator is passed over the guidewire
- Dilator is removed and arterial catheter (5 French sized) is passed over the guidewire
- Guidewire is removed

    - Femoral artery is used because it can be easily felt and cannulated to pass an arterial catheter.
    - Water soluble iodine dye (Sodium diatrizoate) is injected. X-rays are taken to see the block, its extent in the affected limb.
    - In TAO *cork screw* appearance is characteristic**.** *Distal run off* through collaterals is also important.
    - If catheter is passed still proximally angiogram of opposite side is possible.
    - Seldinger technique can also be used (to study) to do renal angiogram to detect renal artery stenosis, renal carcinomas, renal anomalies (vascular). But a *caution* should be remembered that angiogram in limb may precipitate further rapid thrombus formation, worsening ischaemia and precipitating gangrene.

Complications of retrograde angiogram

- Bleeding, hypotension
- Dissection of the vessel wall, pseudoaneurysm
- Haematoma formation
- Embolic blue toe syndrome
- Thrombosis, AV fistula
- Infection, osmolarity discomfort
- Anaphylaxis—4%

  - ***Direct aortic angiogram***, practiced earlier, is discouraged at present because of the risk of *aortic dissection* and paraplegia due to blockage/spasm of anterior spinal artery.
  - ***Digital subtraction angiography (DSA)***
    - Here vessel (artery) is delineated in a better way by eliminating other tissues through computer system. AV fistulas, haemangiomas, lesion in circle of Willis, vascular tumours, other vascular anomalies are well made out.

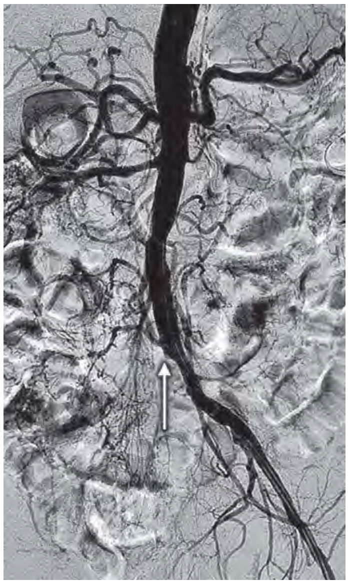

Fig. 1.335: DSA showing total block of right common iliac artery due to thrombosis.

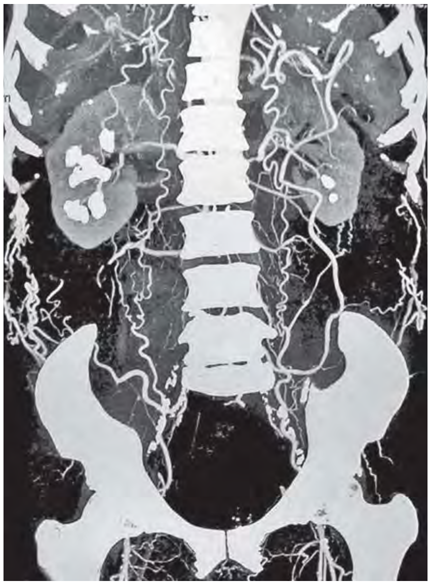

Fig. 1.336: CT angiogram showing abdominal aorta, branches and renal vessels.

  - Dye is injected either to an artery or vein. Injecting into a vein is technically easier but larger dose of dye is required. Injecting into an artery is technically difficult but small dose of dye is sufficient.
  - *Advantages:* Only vascular system is visualised; other systems are eliminated by computer subtraction. Small lesion, its location and details are better observed with greater clarity.
  - *Disadvantages:* Cost factor and availability**.**
  - *Complications:* Anaphylaxis, bleeding, thrombosis.
- ***CT angiogram/MR angiogram.***
- ***Brown's vasomotor index:*** Specific nerve of the ischaemic limb is anaesthetised like posterior tibial nerve or ulnar nerve (local anaesthetia or spinal anaesthesia is given to anaesthetise entire limb). If the ischaemic disease is at vasospasm stage (like in TAO), nerve block will relieve the sympathetic vasospasm and skin temperature rises. It is compared to mouth temperature of the patient.

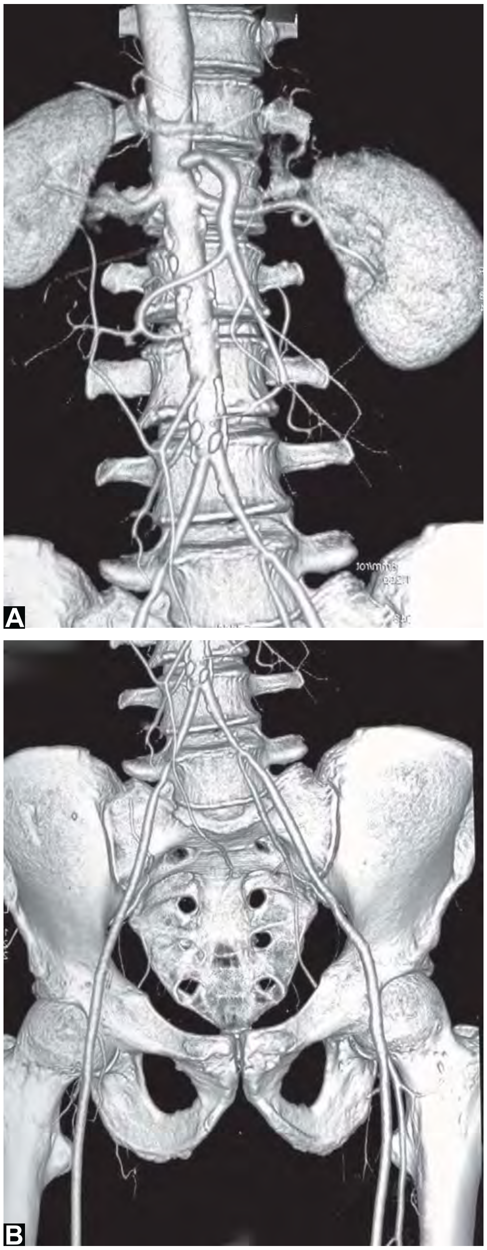

Figs 1.337A and B: Aortic and aortoiliac CT angiogram.

It is to assess the degree of vasospasm which is used as a predictor of the efficacy of sympathectomy.

(Rise in skin temperature—rise in mouth temperature) divided by rise in mouth temperature is called as *Brown's vasomotor index*. If it is more than 3.5, it is due to vasospasm, and can be relieved by sympathectomy. If less than 3.5, sympathectomy is not beneficial.

- ***Transcutaneous oximetry:*** By placing polarographic electrodes over the skin over thigh, leg and foot of oxygen tension (tcPO_2) can be measured which is reflection of underlying tissue perfusion. Normal tcPO_2 in the foot is

*The peroneal artery replaces the anterior tibial artery in 5% of cases.***—Solly M Cohen**

50-60 mmHg. Level less than 40 mmHg shows inadequate wound healing. Level below 10 mmHg suggests critical ischaemia with complete failure of wound healing.

- ***Other tests***
  - *Xenon 133 isotope method* is used to study muscle blood flow. Xenon 133 after mixing with normal saline is injected IM to study its clearance.
  - Two electromagnetic electrodes are placed in contact with arterial wall in opposite directions which pick up moving blood force to feed into electronic amplifier. But it is invasive as artery has to be dissected to place electrodes.

Treatment Plan for Arterial Diseases

- Stopping smoking; supervised exercises; regular controlled walk, diet (carbohydrate and lipid free diet); care of limbs.
- Control of hypertension, diabetes; antilipid drugs like atorvastatin 10 mg or pravastatin 40 mg; low dose aspirin 75 mg; clopidogrel 75 mg; vasodilators; ticlopidine; dipyridamole; cilostazole 100 mg are different drugs used.
- Percutaneous transluminal balloon angioplasty (PTA) done mainly to iliac arteries, subclavian arteries, renal artery, carotid and occasionally leg arteries and mesenteric or gastrointestinal arteries. PTA with stenting using expandable stents to the arteries is also often done to get a better result.
- Bypass graft surgeries—aortofemoral; femorofemoral; iliofemoral, etc. Dacron, human umbilical vein (3 mm), saphenous vein; PTFE (Polytetraflouoro ethylene)—are different grafts used.
- Endarterectomy, atherectomy, thrombectomy, profundaplasty, etc.

DISEASES OF THE ARTERIES

- Atherosclerosis.
- Thromboangiitis obliterans (TAO) (Buerger's disease).
- Raynaud's disease.
- Conditions causing Raynaud's phenomenon: Like scleroderma, rheumatoid arthritis, SLE, granulomatosis, vasculitis of other causes.
- Embolus.
- Aneurysms.
- Other causes: Fibromuscular dysplasia, radiation, Takayasu's arteritis.

ATHEROSCLEROSIS

It is a chronic, complex inflammatory condition of elastic and muscular arteries, involving as systemic and segmental. It begins in childhood as fatty streaks.

Risk factors
Definitive
• Hypercholesterolaemia, and hyperlipidaemia (cholesterol > 200 mg%; high LDL (>100 mg%); low HDL (< 35 mg%).
• Cigarette smoking
• Hypertension
• Diabetes mellitus
Relative
• Age—elderly
• Common in males
• Hypertriglyceridaemia
• Sedentary life, obesity
• Family history

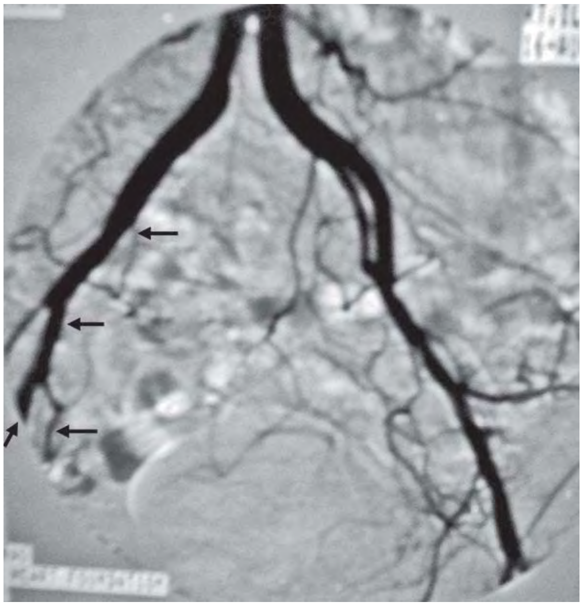

Fig. 1.338: Angiogram showing atherosclerosis.

Pathogenesis

- All risk factors cause initial endothelial injury, both mechanical as well as toxic. This reduces significantly *normal atheroprotective features* of endothelium (*barrier function; antiadhesive effect; antiproliferative effect* on smooth muscles of arterial wall). Progressive atheromatous plaque formation, thrombosis, migration and proliferation of vascular smooth muscle cell occur. *Migrated smooth muscle cells* into intima act as neointima and this migration is stimulated by PDGF (*platelet derived growth factor* released by endothelial smooth muscles, platelets); which (this migrated smooth muscle) newly becomes secretory to produce large quantity of matrix of the plaque. Lipid (LDL) gets oxidised to release factors which promote inflammation and coagulation and factors which prevent production of *protective nitric oxide*. Macrophages stabilise the plaque.
- *Pathology* constitutes of atherosclerotic plaque which contains smooth muscle cells, connective tissue matrix, macrophages and lipid (the feature of atherosclerosis). Ulceration and calcification occurs in these plaques. *Ulcerated plaque* is highly thrombogenic causing thrombosis and further critical block of the vessel leading to tissue ischaemia and infarction distally.
- Plaques are more at the *dividing junctions* of the artery where stress and shear force of the blood flow is more. Plaques are dynamic in nature with progression and regression phases.
- Plaque progression has got a unique ability of *adaptation* so that as the plaque progresses, lumen caliber is been tried

to be preserved until critical stage occurs. Stenosis more *than 40%* is said to be *critical*. Beyond this, compensatory mechanism fails causing rapid progression and further stenosis of lumina. Stenosis more than 40% causes *atrophy of tunica media* making arterial wall mechanically unstable leading into dilatation and aneurysm.
- ***Common arteries involved are***—infrarenal part of abdominal aorta, coronary arteries, iliofemoral vessels, carotid bifurcation, popliteal arteries. It is less common in upper limb arteries, common carotid, renal and mesenteric arteries.

Features and Evaluation

- It is common after 50 years, but can occur at earlier age group.
- It occurs in males and females. Family history is common.
- Smoking, hypertension, diabetes, raised cholesterol are common causes.
- Veins are not diseased. Arterial wall is thickened on palpation.
- Thrill and bruit over femoral, renal, carotid arteries may be felt/heard. It suggests localised stenosis with turbulence of blood flow.
- Features of ischaemia in the affected limb seen. Absence/feeble pulses including of main arteries of the limb—femorals. Abdomen should be examined for aortic aneurysm.
- Transient ischaemic attacks, chest pain, eye problems, mesenteric ischaemia, altered renal function may be associated.
- Blood sugar, fasting lipid profile, Doppler, angiogram (CT / DSA), US abdomen, ECG, echocardiography are essential investigations. Angiogram shows typical narrowed artery, site, extent, percentage of stenosis, and collaterals.

Management

- *Risk factor modification*: Avoid smoking; control of hypertension, diabetes, hypercholesterolaemia; weight reduction by diet, and exercise.
- *Drugs*: Antiplatelet agents (aspirin 75 mg, clopidogrel 75 mg); cilostazol 50 mg bd; atorvastatin to reduce cholesterol; pentoxiphylline.
- *Percutaneous transluminal angioplasty (PTA*) is very useful for iliac blocks and lower limb blocks.
- *Surgeries:*
    - Thrombectomy, endarterectomy, profundaplasty.
    - Reverse/saphenous vein graft.
    - By pass grafts—iliofemoral, aortofemoral, iliopopliteal, femorofemoral grafts.
    - Amputations if limb is gangrenous—toe/below knee, above knee. Forefoot and Syme's amputations are not feasible in vascular conditions.

Note:

Lumbar sympathectomy and omentoplasty are not much useful in atherosclerotic limb. Omental vessels as such are often poorly perfused in atherosclerotic patients due to involvement of coeliac trunk.

Aortoiliac Occlusive Disease

Common site of symptomatic atherosclerotic occlusive arterial disease of lower limb is infrarenal abdominal aorta and iliac arteries. Aortic bifurcation is the most common site of occlusion. Often disease may also extend into infrainguinal level.

Types of Aortoiliac Occlusive Disease

Type I: Disease localised to distal abdominal aorta and common iliac arteries.
Type II: Wide spread aortic and iliac disease.
Type III: Multiple level diseases along with infrainguinal diseases.

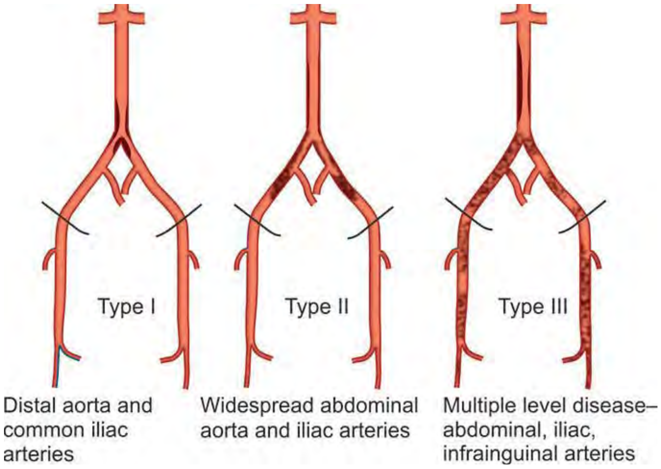
Type I
Type II
Type III
Distal aorta and common iliac arteries
Widespread abdominal aorta and iliac arteries
Multiple level disease–abdominal, iliac, infrainguinal arteries

Fig. 1.339: Types of aortoiliac occlusive disease.

Features

- Common in 5th and 6th decades. Common in males.
- Claudication in buttock, Leriche syndrome with impotence, distal ischaemia are the features.
- Femoral artery pulsations below are absent. Systolic bruit over aorta and iliac arteries may be heard suggesting stenosis.
- Atheromatous plaque may dislodge and may cause embolus causing acute presentation.
- Aortic angiogram is diagnostic.

Management

- Treatment for diabetes, hyperlipidaemia, etc.
- Surgical treatment is the mainstay.
    - Direct *anatomical* reconstruction
        - *Aortoiliac endarterectomy* is reboring/disobliteration procedure useful for type I disease. Diseased intima, plaque with thrombus is removed by arteriotomy along the entire length which is closed later using 4 zero/5 zero polypropylene continuous sutures (open endarterectomy). In lengthy disease, after making two small arteriotomies at proximal and distal diseased parts, endarterectomy loop is passed to remove the intima with diseased plaque (*semi-closed endarterectomy*). *Advantages:* It avoids prosthetic graft and its complications. *Problem* is—reocclusion and restenosis.

Dacron is polyethylene terephthalate.

  - *Aortofemoral bypass* graft is the gold-standard surgical procedure for type I and II disease. Long-term patency rate is 70-80%. Woven Dacron graft is used. *Complications are*—bleeding, thrombosis, embolisation, graft blockage, graft failure, graft infection, graft leak, aortovenacaval/aortoduodenal fistula, mesenteric ischaemia (colonic), impotence.
- Indirect *extra anatomical* bypass

  It is quicker and technically easier and is suitable to patients who cannot tolerate anatomical bypass. Axillo-bifemoral graft is used, only in such occasional situation.
- Nonoperative *catheter based* endovascular procedure

  If stenosis is less than 5 cm *percutaneous transluminal angioplasty* (PTA) with or without intravascular stents can be done. It is useful for single or multiple short focal stenoses. It is now proved that long-term patency of PTA is equal to surgical intervention.

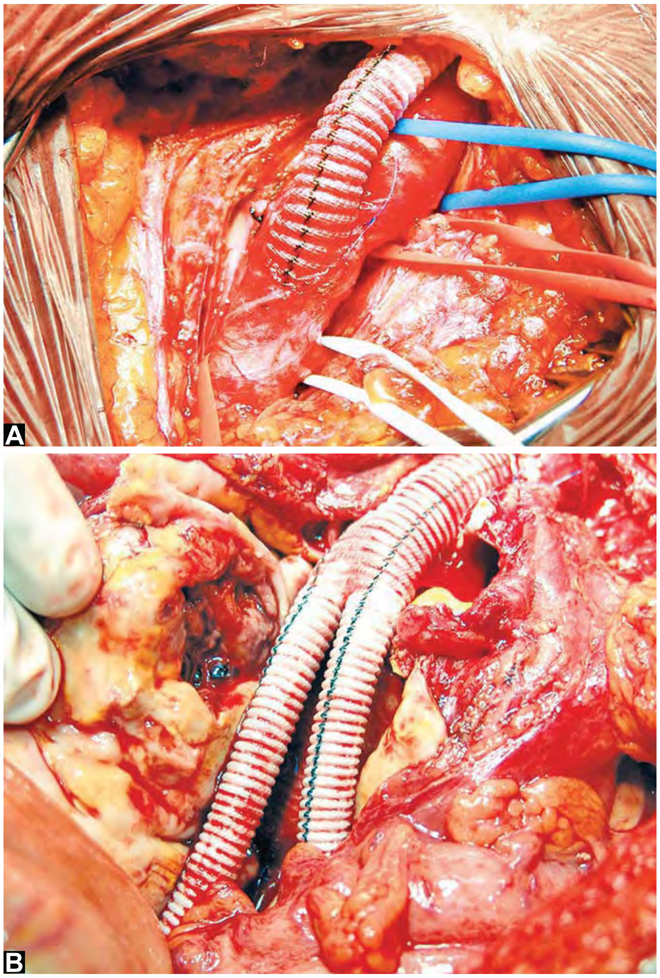

Figs 1.340A and B: Aortofemoral bypass graft.

Infrainguinal Arterial Occlusive Disease

It is either part of type III aortoiliac disease (aortoiliac femoral) or femoropopliteal tibial disease. Superficial femoral artery is most commonly involved. Involvement of long segment of the artery is common. Occasionally a short stenotic segment may be present.

Management

- If the popliteal artery below knee is patent *femoropopliteal bypass* is the ideal procedure used. Otherwise one of the patent branches is used for bypass. *In situ saphenous vein graft* is ideal; reverse saphenous vein graft or synthetic femoropopliteal graft can also be used.
- *Profundaplasty* may be done to improve the collateral circulation through profunda femoris (deep femoral).

Note:

PTA in infrainguinal blocks is occasionally useful, only when stenosis is short and well-localised; otherwise it is *not a good* option. Angioplasty with laser drilling is often tried.

THROMBOANGIITIS OBLITERANS (TAO)
Syn. Buerger's Disease

> *The disease (occurs) in young adults between the ages of twenty and thirty-five or forty years.... Upon examination we see that one or both feet are markedly blanched, almost cadaveric in appearance, cold to the touch, and that neither the dorsalis pedis nor the posterior tibial artery pulsates.... After months... trophic disturbances make their appearance.... Even before the gangrene, at the ulcerative stage, amputation may become imperative because of the intensity of the pain.*
>
> —**Leo Buerger**
> 1908 (Professor of Urology, 1879 to 1943)

- It is a disease exclusively seen in males of young age group (Not usually seen in females due to genetic reason). Seen only in smokers and tobacco users.
- Almost always starts in lower limb, may start on one side and later on the other side. Upper limb involvement occurs only after lower limb is diseased. Only upper limb involvemnt can occur (not uncommon) but it is *rare.*
- It is nonatherosclerotic inflammatory disorder involving medium sized and distal vessels with cell mediated sensitivity to type I and type III collagen.
- It is common in Jewish people; it is rare even in female smokers.
- Hormonal influence, familial nature, hypersensitivity to cigarette, altered autonomic functions are probable different causes.
- Lower socioeconomic group, recurrent minor feet injuries, poor hygiene are other factors.
- It is segmental, progressive, occlusive, inflammatory disease of small and medium sized vessels with superficial thrombophlebitis often may present as Raynaud's phenomenon with microabscesses, along with neutrophil and giant cell infiltration, with skip lesions.

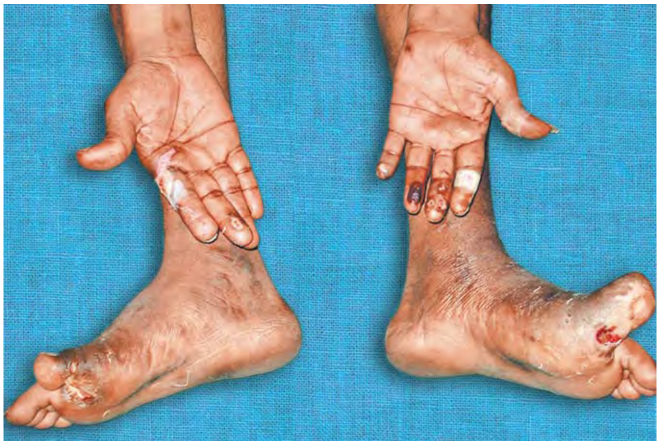

Fig. 1.341: Ischaemic features in upper and lower limbs (four limbs).

Pathogenesis

Smoke contains *carbon monoxide* and *nicotinic acid*
↓ ← Carboxyhaemoglobin
Causes initially vasospasm and hyperplasia of intima
↓
Thrombosis and so obliteration of vessels occur, commonly medium sized vessels are involved.
↓
Panarteritis is common
Usually involvement is *segmental*
↓
Eventually artery, vein and nerve are together involved
↓
Nerve involvement causes rest pain
↓
Patient presents with features of ischaemia in the limb
↓
Once blockage occurs, plenty of collaterals open up depending on the site of blockage either around knee joint or around buttock
Once collaterals open up, through these collaterals, blood supply is maintained to the ischaemic area
↓
It is called as *compensatory peripheral vascular disease*.
↓

If patient continues to smoke, disease progresses into the collaterals, blocking them eventually, leading to severe ischaemia and is called as *decompensatory peripheral vascular disease*. It is presently called as ***critical limb ischaemia.*** It causes rest pain, ulceration, gangrene.

***Note*:**
There is vasospasm → intimal hyperplasia → thrombosis → panarteritis → obliteration; tender, cord like veins with *superficial migratory thrombophlebitis* (30%); with nerve involvement due to vasa nervorum block/spasm. Arterial lumen is *blocked* but not thickened like atherosclerosis.

Smoking index (SI) =
[Number of cigarettes smoked per day] × [Number of years of smoking]
SI > 300 is a risk factor

Pack Years Index (PYI) =
[Number of years of smoking] × [Number of *packets of* cigarettes per day]
PYI > 40 is a risk factor

Shianoya's criteria for Buerger's disease

- Tobacco use. Only in males
- Disease starts before 45 years
- Distal extremity involved first without embolic or atherosclerotic features
- Absence of diabetes mellitus or hyperlipidaemia
- With or without thrombophlebitis

Classification of TAO

Type I: Upper limb TAO—rare.
Type II: Involving leg/s and feet—crural/infrapopliteal.
Type III: Femoropopliteal.
Type IV: Aortoiliofemoral.
Type V: Generalised.

Clinical Features

- Common in male smokers between the 20-40 years of age group. It is a *smoker's disease*.

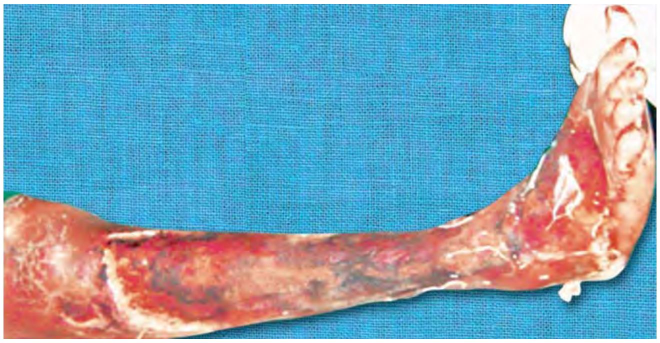

Fig. 1.342: Gangrene of leg and foot.

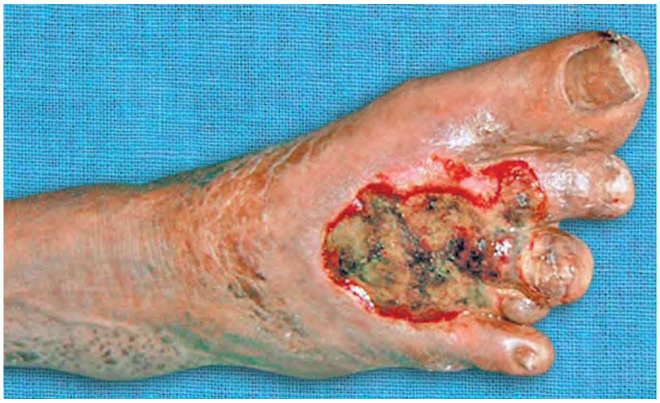

Fig. 1.343: Ischaemic ulceration in the foot.

The art of seeing invisible things is called as 'clinical vision'.

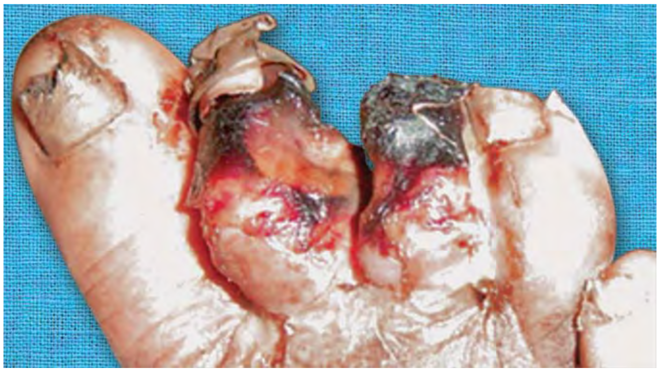

Fig. 1.344: Gangrene of mid-toes—distal part.

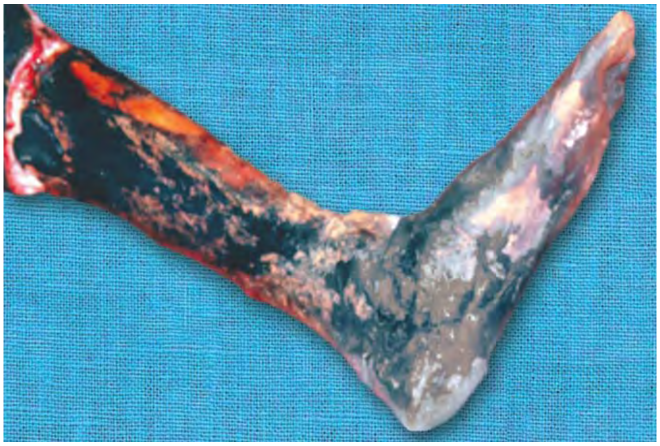

Fig. 1.345: Dry gangrene of leg. Patient needed above knee amputation.

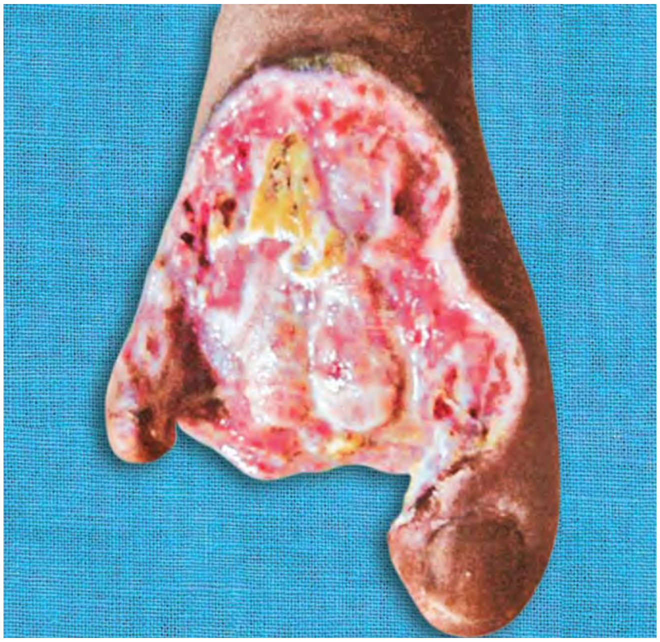

Fig. 1.346: Ischaemic ulcer foot. Note the poor blood supply and unhealthy granulation tissue.

- Intermittent claudication in foot and calf progressing to rest pain, ulceration, gangrene.
- Recurrent migratory superficial thrombophlebitis.
- Absence/feeble pulses distal to proximal; dorsalis pedis, *posterior tibial,* popliteal, femoral arteries.
- May present as Raynaud's phenomenon.

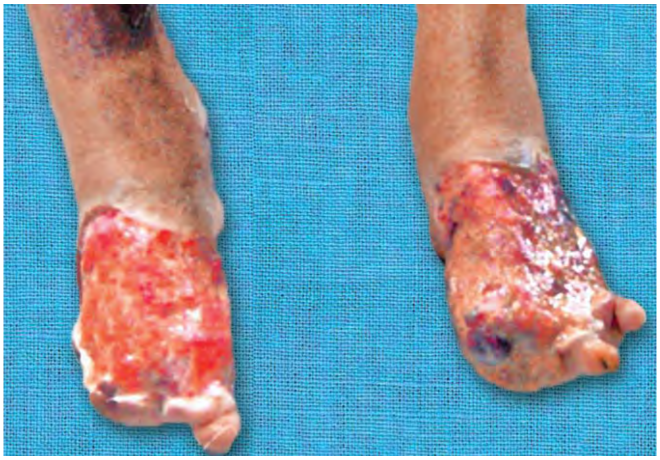

Fig. 1.347: Bilateral TAO in a young male with large ischaemic ulcers. Patient underwent amputation of both limbs (Below-knee).

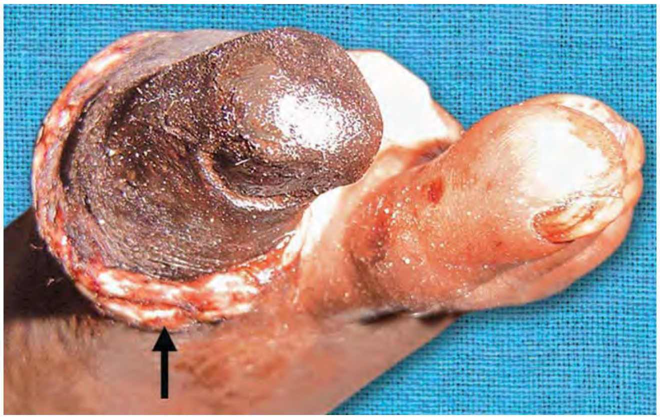

Fig. 1.348: Great toe gangrene. Note the clear line of demarcation.

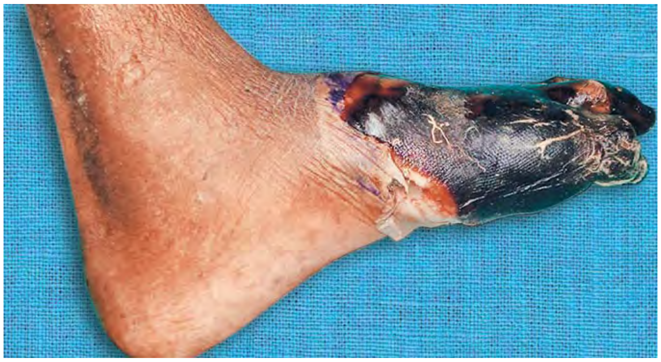

Fig. 1.349: Gangrene foot.

Investigations

- Hb%. Blood sugar.
- Arterial Doppler and Duplex scan (Doppler + B mode U/S).
- ***Transfemoral retrograde angiogram*** through Seldinger technique:
    - Shows blockage—sites, extent, and severity.
    - *Cork screw* appearance of the vessel due to dilatation of vasa vasorum.
    - Inverted tree/spider leg collaterals.
    - *Severe vasospasm causing corrugated/rippled artery.*
    - *Distal run off* is amount of dye filling in the main vessel distal to the obstruction through collaterals. If distal run off is good then ischaemia is compensated. If distal run off is poor then ischaemia is decompensated.

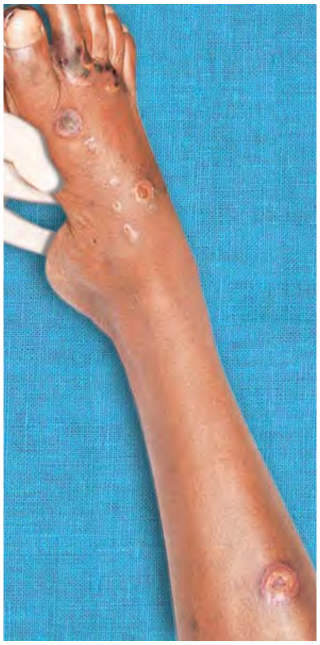

Fig. 1.350: Skip ischaemic ulcers are common in vascular diseases. It suggests severe ischaemia up to most proximal ulcer level.

- ***Transbrachial angiogram*****:** If femorals are not felt, then transbrachial angiogram (through left side brachial artery—left subclavian artery—and so to descending aorta) should be done.
- Ultrasound abdomen to see abdominal aorta for block/aneurysm.
- Vein, artery, nerve biopsy.

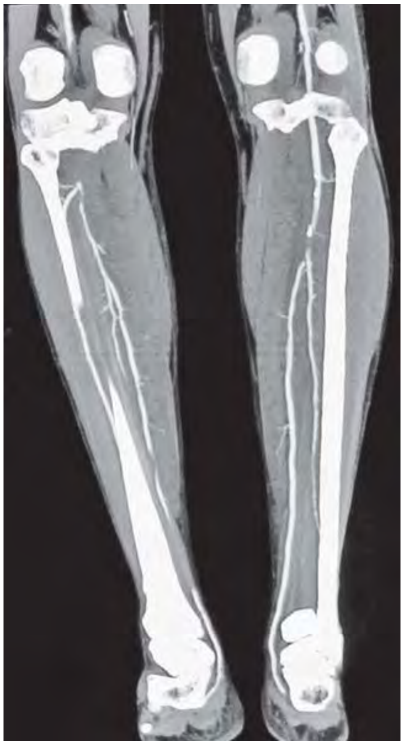

Fig. 1.351: CT angiogram of lower limb (leg area) showing segmental block.

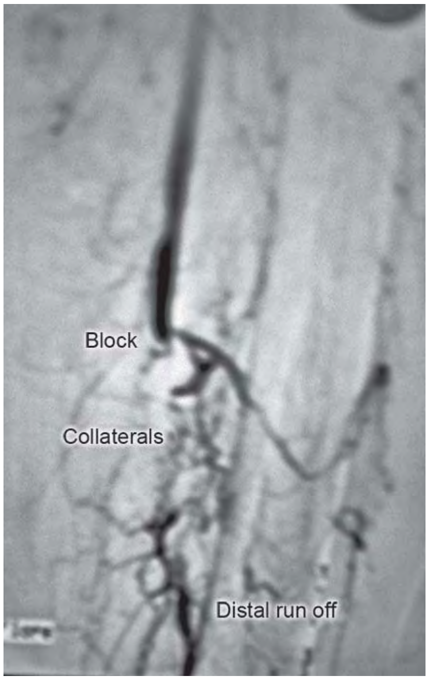

Fig. 1.352: Angiogram showing block in main vessel with opened up collaterals and adequate distal run off.

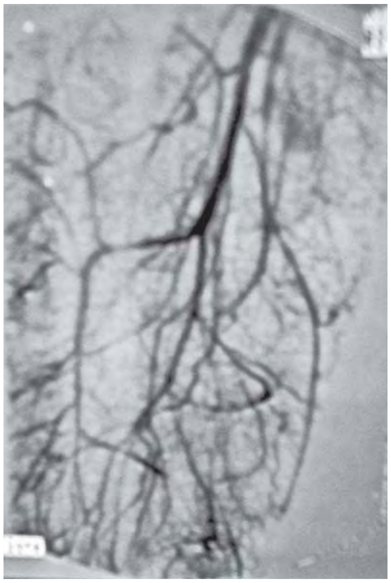

Fig. 1.353: Angiogram showing adequate collaterals.

Treatment

Stop smoking. “Opt for either cigarette or limb, but not both.”

Drugs

- Vasodilators, e.g. nifedepine.
- Pentoxiphylline increases the flexibility of RBC’s and helps them reach the microcirculation in a better way so as to increase the oxygenation. Its efficacy is more in *venous ulcer* than arterial diseases (now).
- Low dose of aspirin 75 mg once a day—antithrombin activity.

- Prostacyclins, ticlopidine, praxilene.
- Clopidogrel 75 mg; atorvastatin 10 mg; parvostatin 40 mg; *cilostazole* 100 mg bid—is a phosphodiesterase inhibitor which improves circulation (ideal drug).

All drugs act *at the collateral level* than on the diseased vessel.

- Analgesics, often sedatives, antilipid drugs like atorvastatin may be needed. Complamina retard (*xanthine nicotinate*) tablet which was used daily once earlier, is presently not in use. However, graded injection of xanthine nocotinate 3000 mg from day 1 to 9000 mg on day 5 is often practiced to promote ulcer healing, helps to increase claudication distance as a temporary basis. Low molecular dextran may be also used.

Care of the Limbs

- *Buerger's position and exercise*—regular graded exercises up to the point of claudication improves the collateral circulation.
  In *Buerger's position*, head end of bed is raised; foot end of bed is lowered to improve circulation. In *Buerger's exercise* leg is elevated and lowered alternatively, each for 2 minutes for several times at time.
- *Care of feet (Chiropady):* Exposure of feet to more cold and warm temperature should be avoided; trauma even minor like nail paring or pressure at pressure points in feet should be avoided. Dryness of feet and legs should be avoided by applying oil to the feet and legs. *Footwea*r should be selected carefully. It is better to wear socks with footwear. *Heel raise* by raising the heels of shoes by 2 cm decreases the calf muscle work to improve claudication.

Chemical Sympathectomy

- Sympathetic chain is blocked to achieve *vasodilatation* by injecting local anaesthetic agent (xylocaine 1%) paravertebrally beside bodies of L 2, 3 and 4 vertebrae in front of lumbar fascia, to achieve temporary benefit. Long time efficacy can be achieved by using *5 ml phenol* in water. It is done under C-Arm guidance. Feet will become *warm immediately* after injection. *Problems* are—possible risk of injecting phenol into IVC/aorta, spinal cord ischaemia.

Surgery

- *Omentoplasty* to revascularise the affected limb.
- *Profundaplasty* is done for blockage in profunda femoris artery so as to open more collaterals across the knee joint (It often makes better perfusion to the knee joint and flap of below-knee amputation).
- *Lumbar sympathectomy* to increase the cutaneous perfusion so as to promote ulcer healing. But it may divert blood from muscles towards skin causing muscle more ischaemic.
- *Amputations* are done at different levels depending on site, severity and extent of vessel occlusion. Usually either below-knee or above-knee amputations are done.
- *Ilzarov method* of bone lengthening helps in improving the rest pain and claudication by creating neo-osteogenesis and improving the overall blood supply to the limb.

Gene Therapy

- Intramuscular injection of vascular endothelial growth factor (VEGF) which is an endothelial cell mitogen that promotes angiogenesis.

TAKAYASU'S PULSELESS ARTERITIS (Takayashu, 1938—Ophthalmologist, Japan)

- It is progressive, initially symptomless panarteritis involving aortic arch and branches of aorta of unknown etiology, probably immunological.
- It is common in young females (85%); common in Japan; commonly subclavian artery is involved (85%); involves all layers of arteries of upper limb and neck; often bilateral. It remains unnoticed for long time.

Features

- Fever, myalgia, arthralgia, upper limb claudication.
- Absence pulses in upper limb/limbs, neck; hypertension.
- Fainting on turning the neck or change in position; atrophy of face.
- Thrill/bruit along major arteries of upper limb and neck are the features.
- Optic nerve atrophy without papilloedema.
- Weakness and paraesthesia of upper limb.
- Cerebral softening, convulsions, hemiplegia can occur.
- Occasionally it can be life-threatening. Myocardial infarction; embolism, ischaemia are other *complications.*
- DSA; MR angiography and Doppler are the *investigations.*

Treatment

- To suppress immunity prednisolone 50 mg/day and cyclophosphamide daily is given.
- Vascular reconstruction.

RAYNAUD'S PHENOMENON

> *Under the influence of a very moderate cold, and even at the height of summer, she (case 1) sees her fingers become ex-sanguine, completely insensible, and of a whitish yellow colour....One might indeed have suspected that the local asphyxia was connected with a spasmodic state of the vessels,...a functional trouble localised to the arterioles immediately contiguous to the capillaries.*
>
> —**Maurice Raynaud,** 1862 (French Surgeon)

It is an episodic vasospasm, i.e. arteriolar spasm. It leads to sequence of clinical features called as Raynaud's syndrome.

Raynaud's syndrome

Sequence of clinical features due to arteriolar spasm.

- *Local syncope:* It is due to vasospasm, causing white cold palm and digits along with tingling and numbness
- *Local asphyxia*: It is due to accumulation of deoxygenated blood as the result of vasospasm causing bluish discolouration of palm and digits with burning sensation (due to accumulated metabolites)

- *Local recovery:* It is due to relief of spasm in the arteriole, leading to return of blood to the circulation causing flushing and pain in digits and palm (pain is due to increased tissue tension)
- *Local gangrene*: If spasm persists more than ischaemic time (more than one hour in upper limb), then digits go for ulceration and gangrene. Does not occur regularly but is an occasional phenomenon in the cycle

Coffman criteria for Raynaud's syndrome—"episodic attacks of well-demarcated reversible self-limiting colour changes for 1-20 minutes on exposure to cold/emotional stimuli and is symmetrical/bilateral lasting for 2 years".

Causes for Raynaud's Phenomenon

- ***Raynaud's disease***:
  - It is seen in females, usually bilateral.
  - It occurs in upper limb with normal peripheral pulses.
  - It is due to upper limb (hand) arteriolar spasm as a result of abnormal sensitivity to cold. Patient develops blanching, cyanosis and later flushing as in Raynaud's syndrome. Occasionally if spasm persists it results in gangrene.
  - Symptoms can be precipitated and observed by placing hands in cold water.
- Working with ***vibrating tools***: Like pneumatic road drills, chain saws, wood cutting, fishermen travelling in machine boats—*vibration white finger*.

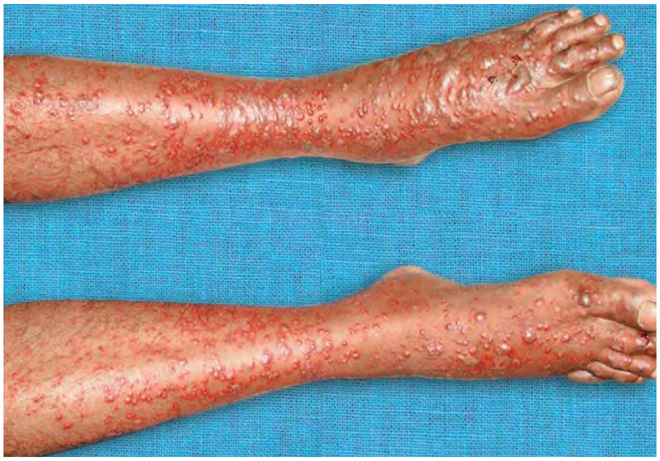

Fig. 1.354: Vasculitis can cause arterial insufficiency.

- ***Collagen vascular diseases***: Like scleroderma, rheumatoid diseases causing vasculitis (all autoimmune diseases).
- ***Other causes***: Cervical rib, Buerger's disease, Scalene syndrome.

It is often associated with CREST syndrome (**C**alcinosis cutis**, R**aynaud's phenomenon, **E**sophageal defects, **S**clerodactyly, **T**elangiectasia).

Types of Raynaud's phenomenon

- Vasospastic
- Obliterative

Raynaud's can be:

- *Primary Raynaud's* is an idiopathic vasospastic disorder without underlying identifiable causes. Usually there is no significant pain in primary type. Primary is probably due to increased sensitivity of alpha 2 receptors to nonepinephrine; decreased nitric oxide and endothelin 1 in endothelial cells; increased serotonin and thromboxane. It is common in females and younger age group. Usually it is bilateral involving all digits.
- *Secondary Raynaud's* is vasospasm due to some underlying cause. Significant pain will be present especially during rewarming stage. There are positive autoantibodies; equal in both sexes; occurs at any age group; need not be bilateral.

Features

- Commonly *bilateral.*
- Common in young *females*.
- Raynaud's disease is common in *w*estern *w*hite *w*omen.
- Usually medial *four digits* and palm are involved. Thumb is *spared*.
- Features of pallor/*blanching* (syncope), *dusky cyanosis* (asphyxia), rubor/painful *red engorgement* (recovery) are the presentation. Occasionally if vasospasm becomes longer, gangrene or ischaemic ulceration supervenes along the tips of the fingers.
- Peripheral pulses (radial/ulnar) are *normally* felt. These pulses will be absent in upper limb TAO.
- Repeated *attacks* are common.

Investigations

- Type is identified by angiogram of hand (DSA/MR angiogram), arterial Doppler/Duplex scan.
- Other investigations required are X-ray of the part, antinuclear antibody (ANA assay) tests specific for different conditions
- Assessment of segmental blood pressure gradient from brachial-forearm-wrist-fingers; finger tip thermography; *cold recovery time* (normal is less than 10 minute, but in Raynaud's it is more, often up to 30 minutes); *reactive hyperaemia time* (pneumatic cuff is inflated and kept for 5 minutes and released to observe hyperaemia); nail fold capillary microscopy; Laser Doppler flux to assess microvascular perfusion of finger skin—are *special methods* of evaluations.
- Other routine investigations for arterial diseases like blood sugar/lipid profile/hypercoagulability status.

Treatment

- Treat the cause.
- Avoid precipitating factors—protect from cold/proper dress/hand warmer electrical or chemical/hand gloves. Avoid smoking even though it is not direct etiological cause (other than upper limb TAO), but it may possibly aggravate the disease. Avoid vibrating tools.

- Vasodilators/pentoxiphylline/low dose aspirin (75-100 mg per day). Calcium antagonist (nifedipine 20 mg) is useful. Steroids may be useful in case of secondary Raynaud's.
- ACE inhibitors, nitrates, endothelin inhibitors (bosentan), epoprostenol—prostaglandin a potent vasodilator and antiplatelet drug (continuous intravenous infusion can be given), iloprost—prostacycline analogue, PG E1, misoprostol (oral PG E 1)—are all tried at different stages of the disease.
- Cervical sympathectomy—is used for nonhealing digital ulceration. Not very beneficial to Raynaud's syndrome.

TEMPORAL ARTERITIS

- There is localised inflammatory giant cell infiltration of arterial wall (*giant cell arteritis*) involving superficial temporal, facial, retinal, upper limb, coronary and vertebral arteries.
- It is common after 50 years. Common in females (2:1).
- Claudication of facial muscles, ischaemic severe headache, tender, thrombosed superficial temporal artery and its branches are the features.
- Retinal ischaemia leading into irreversible blindness is dangerous feature. Involvement of coronary artery may cause myocardial infarction.
- Temporal artery biopsy is diagnostic – shows giant cell granuloma with CD4+ T lymphocytes.
- High dose long-term *prednisolone 80 mg* / day is needed. In involvement of retinal artery IV hydrocortisone / *methylprednisolone* may be needed initially.

TREATMENT OF ARTERIAL DISEASES

a. Medical

General Measures

- Stop smoking, reduction of weight, exercise.
- Change in life style, care of feet.
- Control of diabetes and hypertension.
- Buerger's position and exercise.

Drugs

- Nifedipine, praxilene, pentoxiphylline, low dose aspirin, prostacycline, dipyridamole, ticlopidine.
- Clopidogrel (75 mg).
- Cilostazol (type III phosphodiesterase inhibitor) 100 mg BD—inhibits platelet aggregation.
- Oral anticoagulants are used *only* if there is history of embolism or atrial fibrillation.
- Prostaglandins, growth factors, vascular endothelial growth factor (VEGF), E_2Fdecoy (blocks intimal and smooth muscle cell proliferation), mesoglycan (breaks blood clot), testosterone, herbals like garlic (reduces viscosity of blood) are other newer drugs under use and trial.
- B vitamins and folic acid reduces homocysteine level (which is a risk factor).
- Inositol, L-carnitine (1500 mg), magnesium 500 mg (not in renal failure or with diarrhea), vitamin E and C are other agents often used to improve walking distance.
- Heparin is used *only* in acute phase or embolism.

b. Surgery

- ***Percutaneous transluminal balloon angioplasty (PTA):*** Through transfemoral Seldinger approach, initially angiogram is done. Then under guidance (fluoroscopic) stenosed area is approached. First guidewire is introduced through which balloon catheter is passed. Balloon of the angioplasty catheter is inflated at stenosed area *for one minute* and repeated if required. Plaques should rupture. Catheter is withdrawn. It is useful in cases of localised stenosed areas.

Note:

- Often *nonexpandable or self-expandable stents* are used if stenosed segment is not dilated adequately through balloon—PTA with stenting.
- PTA for carotid artery stenosis is risky and not ideal as there will be possible release of microemboli during dilatation procedure which can precipitate stroke. Specialised balloon catheters with umbrella tip which can trap the microemboli may be used in these places.

Types

- Conventional: Here balloon is inflated along the lumen to break the plaque circumferentially.
- Subintimal: Here balloon is inflated after passing subintimal plane to break the plaque.

Complications

- Thrombosis, bleeding, sepsis.
- Embolism, dissection, retroperitoneal haematoma.
- Pseuodoaneurysm formation.

Advantages

- It is done under local anaesthesia.
- Procedure can be repeated if needed.
- Stent can be placed at a later stage if needed.
- It is done when stenosis is *less than* 5 cm. In ideal indications its efficacy is equal to surgery.

Disadvantages

- It is less useful for lengthy blocks or stenosis more than 5 cm.

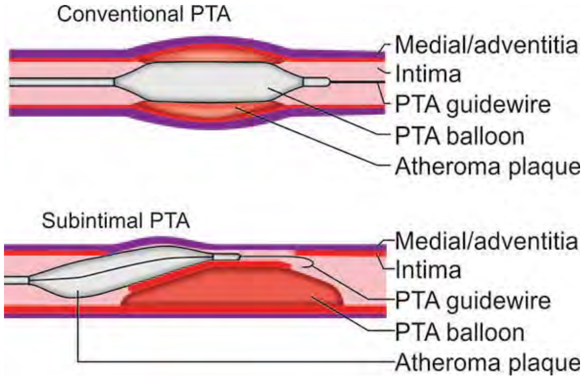

Fig. 1.355: Conventional and subintimal types of PTA.

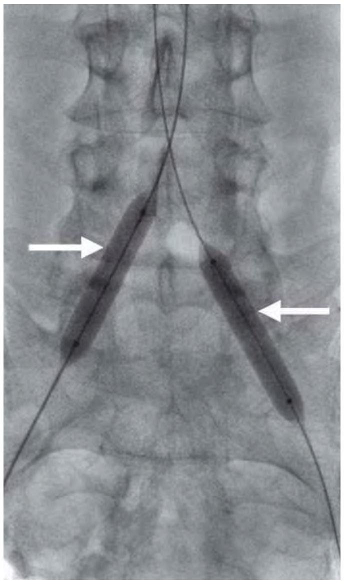

Fig. 1.356: Percutaneous transluminal balloon angioplasty (PTA). Note the inflated balloons on both side iliac arteries.

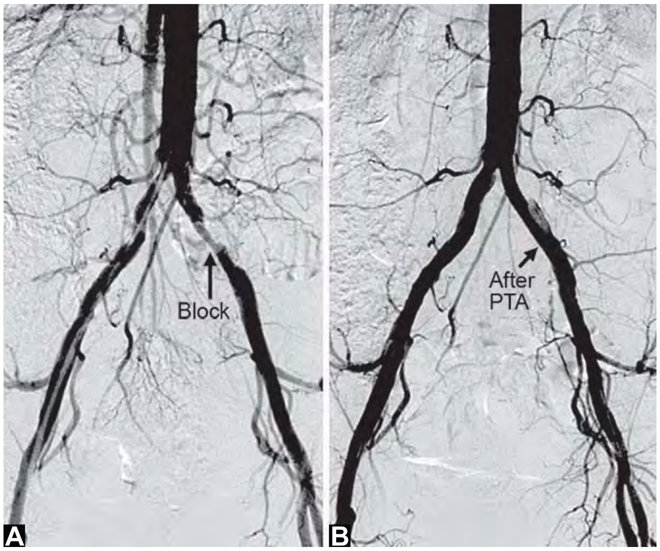

Figs 1.357A and B: DSA showing left sided aortoiliac block and correction after doing PTA.

- It is dangerous to do in stenosis of carotid artery where endarterectomy is ideal.

◆ ***Atherectomy:***

It is removal of atheroma either through open surgery or by percutaneous route from the wall of the vessels.

◆ ***Thrombectomy:***

It is removal of thrombus through an arteriotomy of larger vessels. Done in aortoiliac, femoropopliteal region.

◆ ***Endarterectomy:***

- It is removal of thrombus along with diseased intima through an arteriotomy. Endothelium of the vessel is removed, hence the name.

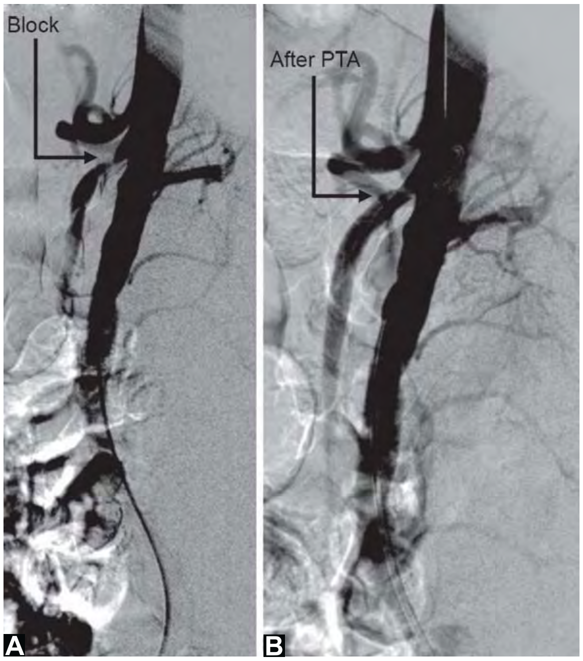

Figs 1.358A and B: DSA showing superior mesenteric artery stenosis. It is corrected by PTA.

- It is done in *carotid,* aortoiliac and occasionally aortofemoral blocks. It is also called as *disobliteration/reboring*. There are three methods—(1) *Open method*—Arteriotomy is done along the entire diseased segment; endarterectomy is done by removing thrombus, diseased intima with plaque along the plane of media. Arteriotomy is closed using 5 zero polypropylene suture and patient is heparinised. (2) *Semiclosed method*—Here two arteriotomies are done on either ends of the level of obstruction; loop endarterectomy stripper is passed from one end to complete the endarterectomy; two arteriotomies are closed. (3) *Wiley's eversion endarterectomy*—Here artery is cut transversely at the junction of diseased and normal nondiseased segment; diseased intima with plaque is circumferentially dissected; artery is everted out to extract the diseased intima like a tube; everted artery is reduced and sutured to normal end of the artery. *Advantages are*—it avoids prosthetic graft and its complications. *Problem* is—reocclusion and restenosis.

◆ Placement of ***intraluminal stent for localised stenosis***.

◆ ***Profundaplasty:***

- It is done when there is localised block in opening of profunda femoris (deep femoral). Profunda femoris is opened, thrombus if present, is removed. Opening is widened using either venous or synthetic (Dacron or PTFE) grafts. This procedure allows collaterals across the knee joint to open through profunda femoris and so gives good blood supply below-knee level and may prevent patient going in for above-knee amputation.

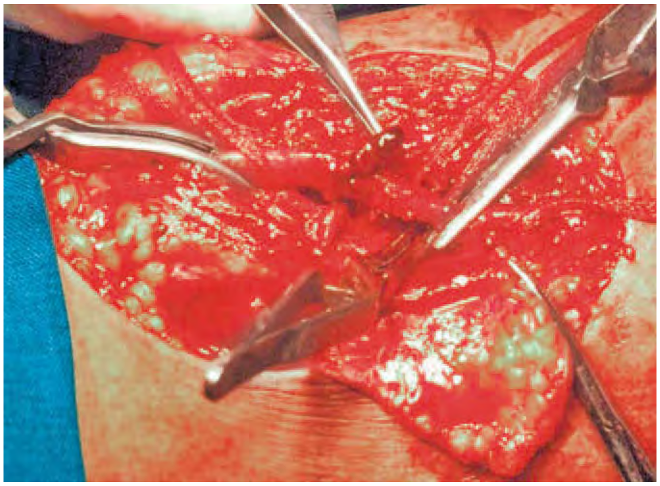

Fig. 1.359: Thrombectomy.

(May be able to save knee joint with below-knee amputation with better prosthesis.)

- Lateral angiogram view is needed to identify the orifice of profunda femoris. Disease involves *invariably only at the orifice* without extending distally towards 1st perforator branch. Endarterectomy at the junction and closure with a venous patch widens the opening adequately.

- ***Reverse saphenous vein graft:*** In case of femoropopliteal block, saphenous vein is dissected out, reversed and sutured above to the femoral artery and below to popliteal segment so as to bypass the blood through reverse saphenous vein graft. Saphenous vein is reversed to nullify the action of valves so as to allow easy flow of blood.

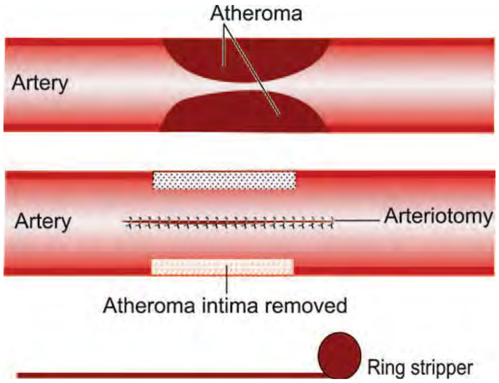
Atheroma
Artery
Artery
Arteriotomy
Atheroma intima removed
Ring stripper

Fig. 1.360: Endarterectomy technique. Both thrombus and diseased intima are removed through an arteriotomy. Often ring stripper is used to remove the atheromatous plaque.

- ***In situ saphenous vein graft:*** It is arterialisation of saphenous vein. Saphenous vein intact in same position is sutured above and below the blocked femoropopliteal region to bypass the blood across. Venous valves are removed through valvulotomy instrument so as to allow the blood to pass. Here nutrient supply of vein is left intact with proximal part of the vein sutured to wider part of the femoral artery, narrow distal part is sutured to narrow part of the artery.

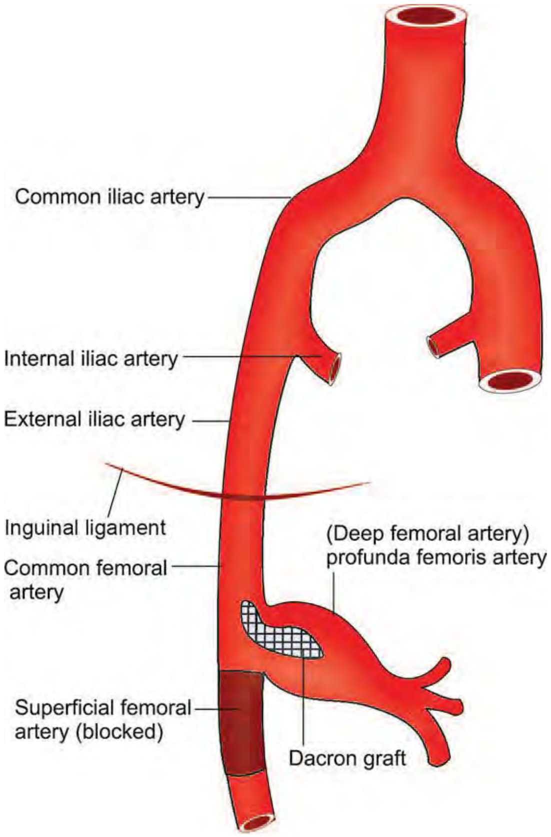
Common iliac artery
Internal iliac artery
External iliac artery
Inguinal ligament
Common femoral artery
(Deep femoral artery) profunda femoris artery
Superficial femoral artery (blocked)
Dacron graft

Fig. 1.361: Profundaplasty for deep femoral block.

- ***Arterial/venous grafts:***

Synthetic:
- Dacron woven graft
- Dacron knitted graft
- PTFE—polytetrafluoroethylene graft

Natural:
- Internal mammary artery (ideal one)
- Long saphenous vein either reverse or *in situ*
- Umbilical vein graft (cryopreserved)—3 mm vein is the minimum diameter required
  Grafts of different length and size are available

Different procedures

- Aortofemoral bypass graft (end to side)—5% mortality.
- Ileofemoral bypass graft.
- Femorofemoral bypass graft
- Femoropopliteal graft.
- Femorodistal graft.
- Axillofemoral graft.

Problems with grafts: Leak, infection, thrombosis, cost factor, availability, reblock.

Note:
Angioscope is used to visualise the valves in saphenous vein or to visualise the completion of the by pass grafts like femorodistal graft.

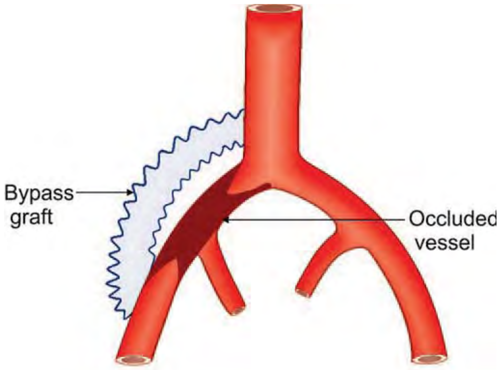
Bypass graft
Occluded vessel

Fig. 1.362: Bypass graft (Aortoiliac).

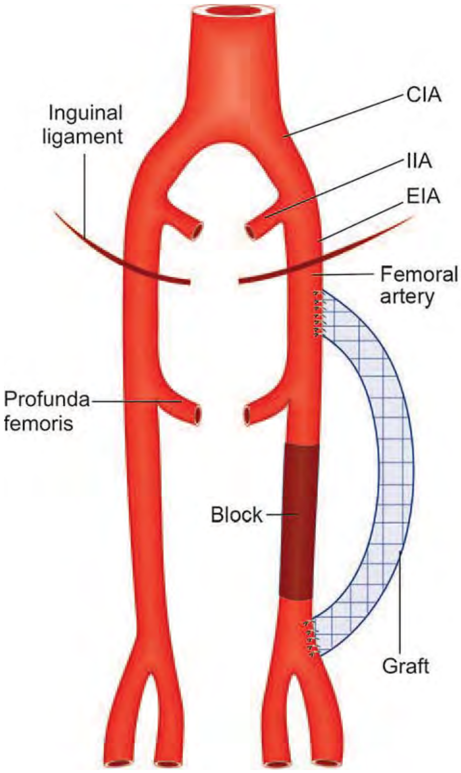
Inguinal ligament
CIA
IIA
EIA
Femoral artery
Profunda femoris
Block
Graft

Fig. 1.363: Femoropopliteal bypass graft.

- ***Cervicothoracic preganglionic sympathectomy:***
  It is removal of 2nd and 3rd thoracic ganglia which contains cells of postganglionic fibres supplying the upper limb. Preganglionic white rami communicantes fibres from 2nd and 3rd sympathetic nerves enter thoracic T1 ganglion and supplies head and neck region through upper part of the stellate ganglion. Preganglionic sympathetic nerve entering the 2nd and 3rd ganglia from below, supplies sympathetic fibres for upper limb through the lower part of the stellate ganglion. In cervical sympathectomy for upper limb ischaemia, lower part of stellate ganglion with Kuntz nerve is divided. For head and neck hyperhidrosis entire stellate ganglion should be removed which leads to development of Horner's syndrome. For hyperhidrosis of axillary area, along with stellate ganglion upper four thoracic ganglia has to be removed.

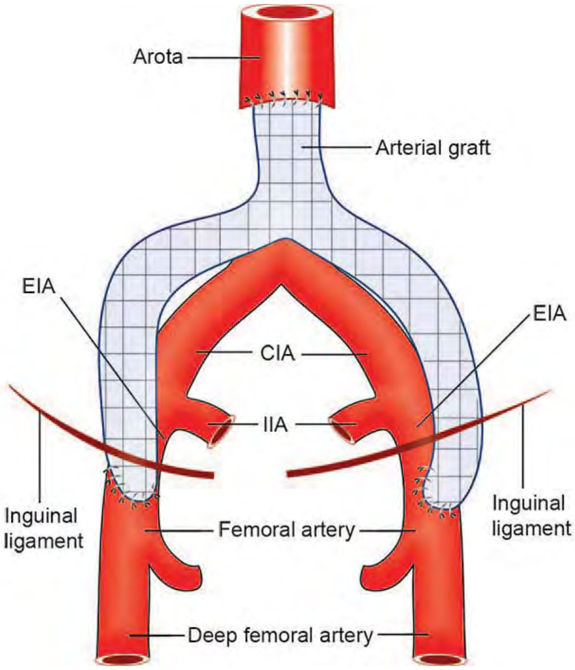
Arota
Arterial graft
EIA
EIA
CIA
IIA
Inguinal ligament
Femoral artery
Inguinal ligament
Deep femoral artery

Fig. 1.364: Aortofemoral bypass graft (End to side).

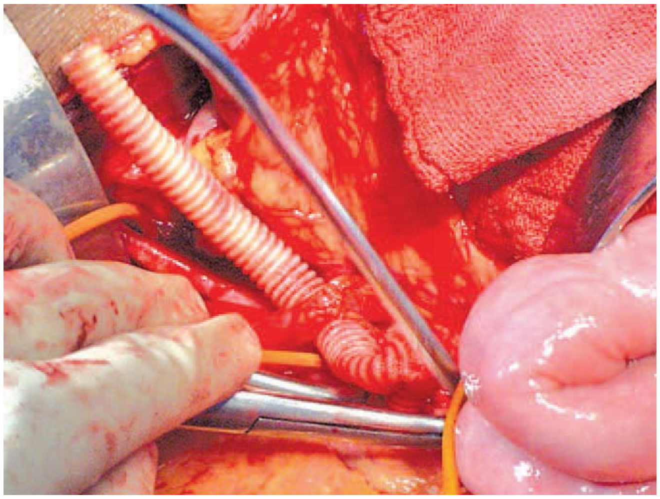

Fig. 1.365: On table picture of aortofemoral arterial graft.

Indications

- ❖ Cervical rib with vascular manifestations—useful
- ❖ Raynaud's phenomenon—useful
- ❖ Hyperhidrosis—very useful
- ❖ Upper limb vasospasm due to other causes—useful
- ❖ Acrocyanosis—useful
- ❖ Causalgia—very useful
- ❖ Sudeck's osteodystrophy

Approaches

- *Supraclavicular approach:*
  Through a supraclavicular incision sternomastoid, (omohyoid is retracted or divided) scalenus anterior

Einstein died of ruptured aortic aneurysm. Nissen did cellphone wrapping of the aneurysm in Einstein muscles, are divided. Phrenic nerve is displaced medially; subclavian artery is pushed downwards; thyrocervical trunk is identified and ligated securely, suprapleural membrane is depressed, stellate ganglion is identified in the neck of the first rib. All rami communicantes from second and third ganglia are divided. Grey ramus from second ganglion to first thoracic nerve called as *Kuntz nerve,* is also divided.

Complications

- Bleeding
- Injury to subclavian artery and nerves
- Pneumothorax and haemopneumothorax
- Horner's syndrome with ptosis, miosis, anhydrosis, enophthalmos
- Chylous fistula, chylocele
- Post-sympathetic neuralgia

- *Transthoracic/axillary approach (Hedley atkins)*: This gives better visibility and easier removal of rami, lower down compared to supraclavicular approach. Patient is placed in lateral position; transverse incision is made just below the hair bearing line; intercostobrachial nerve is preserved. Thorax is opened at 2nd space. Sympathetic chain is identified at the neck of 1st rib.
- *Thoracoscopic sympathectomy* is the choice, and popular approach at present.

  Advantages are better visibility with magnification, less trauma of access (wound), faster recovery, and precise.

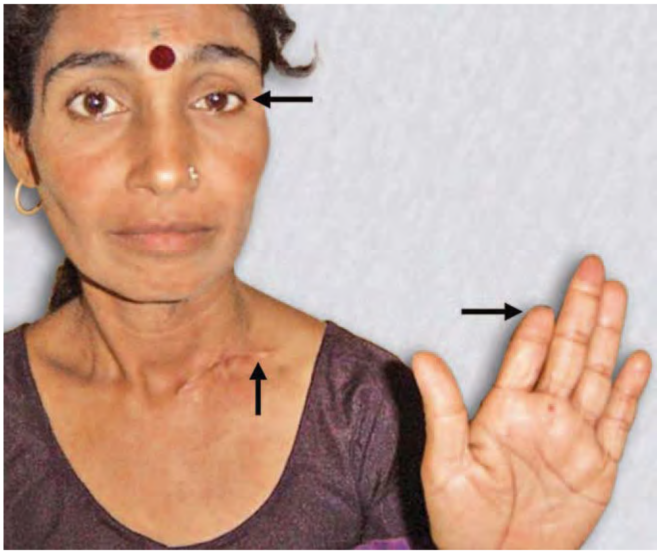

Fig. 1.366: Cervical sympathectomy scar in the neck and patient has developed Horner's syndrome. Healed ulcer over tip of index finger on left side is seen. Patient is asymptomatic after sympathectomy.

◆ ***Lumbar sympathectomy***:

Indications:

- Peripheral vascular disease like TAO.
- To promote healing of cutaneous ulcers.
- To change level of amputation and to make flaps to heal better after amputation.
- Causalgia of lower limb (it is common in upper limb).
- Hyperhidrosis of lower limb is rare.

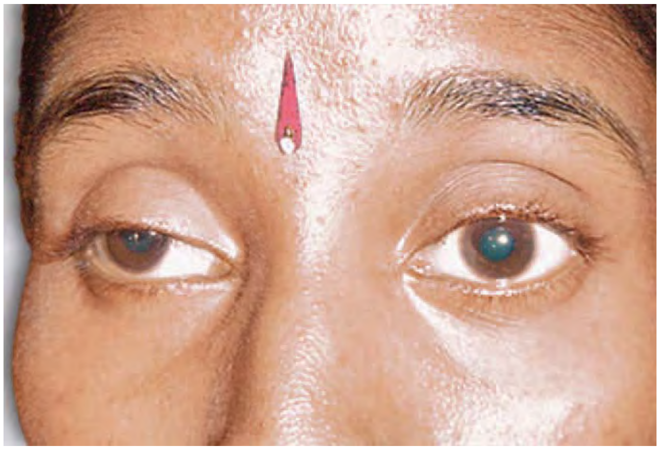

Fig. 1.367: Ptosis of right eye in Horner's syndrome.

Principle:

- It increases the cutaneous blood supply thereby promoting healing of ulcer and skin flaps in amputation. It is a preganglionic sympathectomy. Ganglion L2 and L3 supplies legs below knee level. L1 supplies upper part of thigh and buttock region. L1 lies under the crus of diaphragm. L4 lies under the common iliac vessels below.
- It increases the blood flow for 2-4 weeks by abolishing constriction of arterioles and precapillary sphincters (basal and reflux). It produces transient small increase in distal perfusion; increases the nutritive perfusion to *promote ulcer healing*; alters the pain perception and pain impulse transmission temporarily.

Procedure:

Under general or spinal anaesthesia, ganglia are approached through a transverse incision in the loin at the level of umbilicus, through extraperitoneal approach, by dividing external oblique, and internal oblique, and splitting transverse abdominis muscles. Inferior vena cava on right side, aorta on left side are identified. Sympathetic chain is identified by its rami, over transverse processes of lumbar spines. L_2, L_3, L_4, L_5 ganglia are removed. L_2 is identified by its size (**Larger**) and more number of rami. L_1 is retained on one side in bilateral cases. If both are removed it will lead to failure of ejaculation and so sterility *(Dry ejaculation).*

Complications

- Injury to IVC or aorta
- Bleeding lumbar veins
- Spinal vessel spasm and so ischaemia of spinal cord and paraplegia, dry ejaculation
- Injury to bowel and ureter
- Wound infection and abscess formation
- Post-sympathetic neuralgia
- Paradoxical gangrene of opposite leg and foot

- Its effects are only temporary (3-4 weeks). Long-term results are doubtful. It can be combined with omentoplasty.
- It can also be done along with below-knee amputation to increase the blood supply of skin flap so as to have better healing.

- Limb will become warmer immediately after sympathectomy.

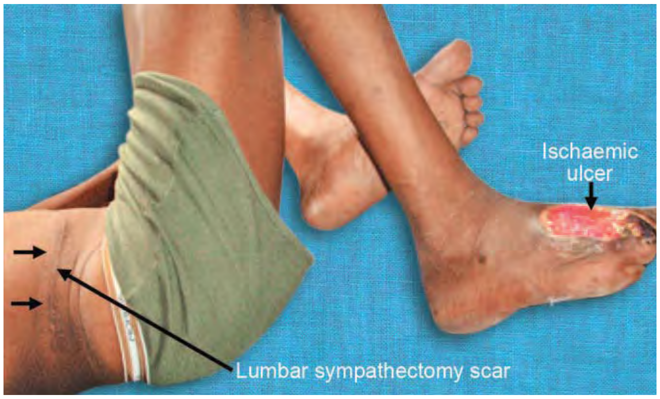

Fig. 1.368: Lumbar sympathectomy scar with ischaemic ulcer foot showing healing sign.

***Note*:**
Lumbar sympathetic chain may be mistaken for lymph nodes, fat, tendon of psoas muscle, genitofemoral nerve.

- ***Chemical sympathectomy*:**
  - It is done in lateral position using a long spinal needle under local anaesthesia. Position is confirmed by injecting dye under fluoroscopy. Later 5 ml of *phenol* in water or absolute alcohol is injected lateral to the vertebral bodies of fourth and second lumbar vertebrae. Care should be taken to see that the needle does not enter IVC or aorta.
  - Procedure is contraindicated in patients with *bleeding disorders* and in patients who are on *anticoagulants.*
- ***Omentoplasty:***

Indications

- Peripheral vascular disease—to improve circulation
- For lymphoedema, it helps by providing lymphatics and so to drain lymph from the limb
- It is also tried for revascularisation of pharynx, cranial cavity

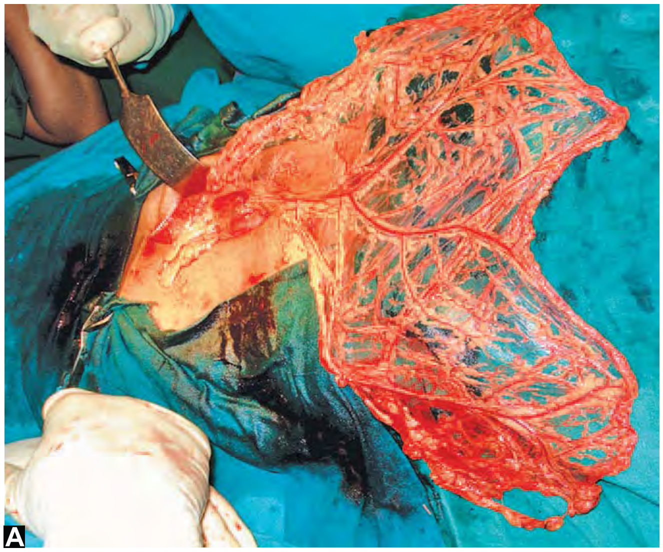

Fig. 1.369A

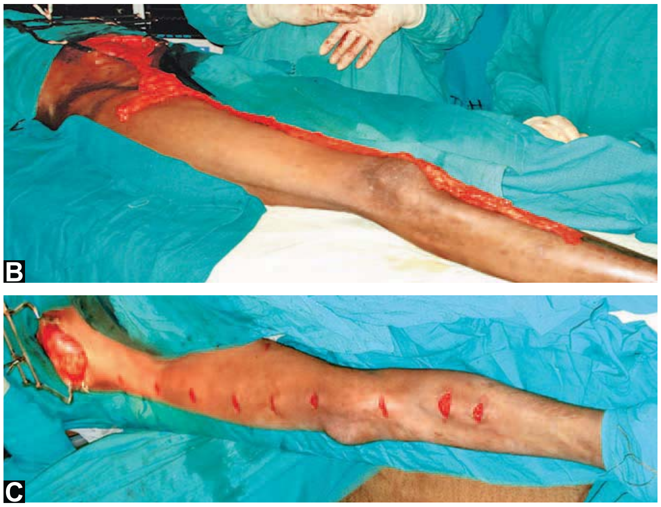

Figs 1.369B and C

Figs 1.369A to C: Omentoplasty procedure for revascularisation of the ischaemic limb.

Omentum is supplied by omental vessels

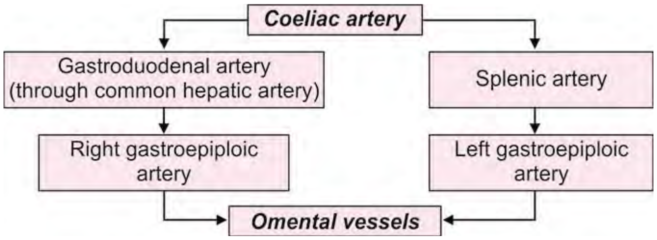
Coeliac artery
Gastroduodenal artery (through common hepatic artery)
Splenic artery
Right gastroepiploic artery
Left gastroepiploic artery
Omental vessels

Four layers of omentum has got omental arcades of vessels. Omentum is also rich in lymphatics. It has got adhesive property. By retaining one of the pedicles, omentum can be mobilised so as to reach the limb to maintain the circulation. It can also be mobilised up to the ankle. It promotes ulcer healing, reduces the pain and controls the features of ischaemia. It can also be used in upper limb ischaemia. But, if patient continues to smoke, disease spreads to these omental vessels also. Often it can be mobilised to both lower limbs in bilateral diseases.

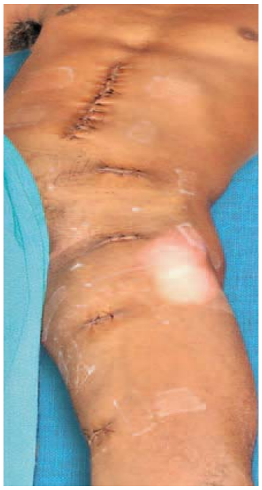

Fig. 1.370: Incisions for omentoplasty in the abdomen (upper midline) and in lower limb.

> *In early painful stage the ischemic foot (gangrene threatened) is nearly always pink, the skin being atrophied as though it were stretched tightly over underlying structure*
> **—Wilfrid G Oakley**

Complications of omentoplasty:

- Abdominal sepsis.
- Incisional hernia, where omental pedicle is tunneled into the limb from the abdomen.
- Adhesions and intestinal obstruction.

Procedure: Under general anaesthesia, abdomen is opened with upper midline incision. Omental vessels are identified. Omentum with its blood supply is carefully mobilised to get an adequate length. Lengthened, mobilised omentum is brought into the subcutaneous plane through abdominal wall, lateral to the lower part of rectus muscle. Later this pedicle is mobilised in the subcutaneous tunnel across the leg, burried in the deep fascia.

- ***Other treatment methods*:**
  - *Amputations* are done at different levels depending on extent of gangrene, site of block, amount of collaterals.

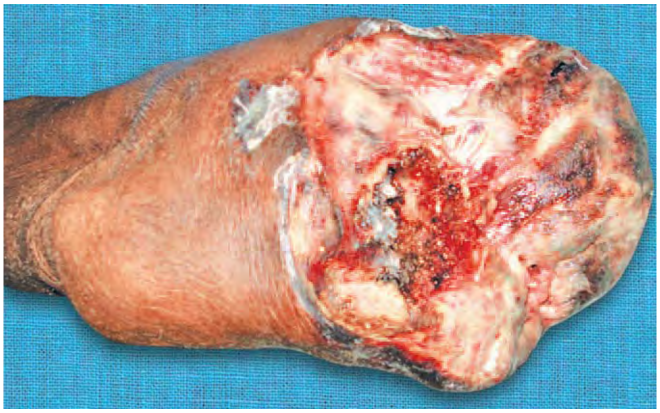

Fig. 1.371: Below knee amputation stump which is also ischaemic. Patient underwent below knee amputation for gangrene foot. But patient might need above knee amputation.

SUBCLAVIAN STEAL SYNDROME

Following obstruction of the first part of subclavian artery, vertebral artery provides collateral circulation to the arm by reversing its blood flow. This causes *cerebral ischaemia with syncopal attacks,* visual disturbances, diminished blood pressure in the affected limb.

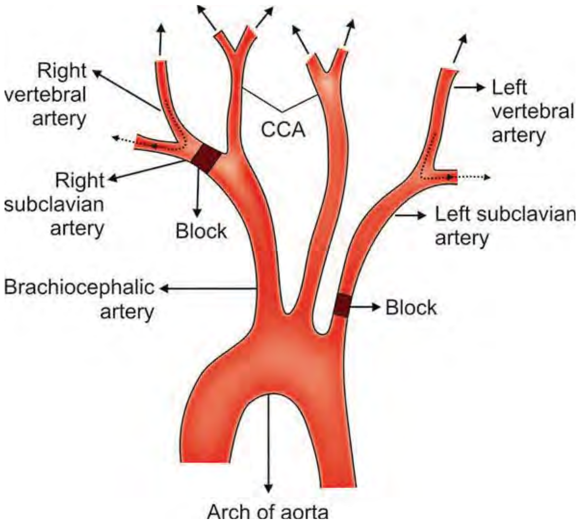

Fig. 1.372: Subclavian steal syndrome.

- It is more common on left side.
- Vertebrobasilar symptoms like dizziness, syncope, visual disturbances, vertigo can occur.
- Pain, heaviness, paraesthesia and fatigue in the arm which is aggravated by exercise.
- Radial pulses on both sides are asymmetrical.
- Blood pressure on the diseased side will be 20 mmHg less compared to normal side.
- *Javid test:* Here compression of carotid artery makes ipsilateral radial pulse feeble.

Investigations

Duplex scan and angiogram. DSA is useful.

Treatment

- Transluminal balloon angioplasty.
- Endarterectomy or bypass graft.
  (Common carotid—subclavian graft).

ACUTE ARTERIAL OCCLUSION

It is a condition of acute lack of tissue perfusion due to sudden cessation of circulation. Main axial artery of the limb is blocked presenting within minutes to hour after occlusion.

It is *common* in lower limb, upper limb; but can occur in mesenteric, cerebral, coronary arteries.

Causes

- *Embolism is the most common* cause in developing country.
- *Trauma.*
- *Thrombosis* of an artery: Normal artery can develop sudden acute thrombosis in certain special situations with hypercoagulable status like malignancy, leukaemia, antiphosholipid antibody syndrome, protein C/protein S/antithrombin deficiency; polycythaemia rubra vera, thrombocytosis. It is commonly observed in external iliac artery, profunda femoris artery and popliteal artery.
- *Thrombosis of a bypass graft* is common cause in western countries which occurs at the site of anastomosis.

Pathophysiology

Distal ischaemia begins immediately after acute obstruction. Most sensitive peripheral nerves are first involved, and then muscles, subcutaneous tissue and skin are affected in order. Irreversible ischaemia occurs in 6 hours. Golden period is 1-6 hours. Ischaemia may get aggravated by—propagation of thrombus below and above the block occluding the orifices of collaterals, fragmentation of embolus, associated thrombosis, acute compartment syndrome.

Acute ischaemia causes endothelial injury of capillaries, arterioles and venules with luminal obliteration. Raised capillary permeability causes fluid leakage into extravascular space forming massive tissue oedema deep to deep fascia which by raising the intracompartmental pressure further reduces the perfusion leading into *acute compartment syndrome*.

Features

- *Pain* which is continuous, severe, steady, bursting.
- *Pallor* of the distal part with extreme cold limb.
- *Pulselessness*—sudden loss of earlier palpable pulse.
- *Paraesthesia*—sensory disturbances like tingling, numbness or complete loss of sensation.
- *Paresis*—damage to motor nerve and muscle leading into paralysis as a late grave feature.
- *Poikilothermia*.

Pain, paraesthesia, paresis are due to ischaemia of peripheral nerves which are sensitive to hypoxia.

TRAUMATIC ACUTE ARTERIAL OCCLUSION

Causes

- Thrombus due to trauma.
- Subintimal haematoma.
- Acute compartment syndrome.
- During femoral or brachial arterial catheterisation for either diagnostic or therapeutic procedures.

Clinical Features

History of trauma, *pain, swelling at the site, pallor, pulselessness, cold limb.*

Investigation

Duplex scan, angiogram.

Treatment

- Wound is explored and tear in the artery is identified. It is sutured using nonabsorbable monofilament material, *polypropylene 6-0*. Often venous or dacron graft is required for interposition.
- Proper antibiotics and heparin are required to prevent thrombosis of the vessel. Later patient is advised to take oral warfarin for maintenance.
- ***Compartment syndrome*** is common in anterior compartment of leg and in front of forearm. Here because of the closed compartment, pressure increases following fracture, haematoma which compresses over the vessel. It leads to blockade of vessel causing acute ischaemia of the limb presenting with severe pain, pallor, pulselessness.
    - *Treatment: Immediate decompression by longitudinal fasciotomy*, is the treatment of choice, wherein deep fascia is cut adequately to relieve the compression. Otherwise limb may go for severe ischaemia, gangrene and may land in amputation.
- Associated fractures, haematoma, vessel tear has to be managed accordingly.

EMBOLISM

(‘Embolus’ means in *Greek*—peg; first this term was used by Virchow in 1854)

It is due to a solid, liquid or gaseous, material which is floating and travelling in the bloodstream, eventually blocking the vessel on its pathway.

- ***Arterial emboli*:**

Sources

- *Cardiac source* (80%):
    - Due to mural thrombus following mitral stenosis and atrial fibrillation (50%); myocardial infarction (25%); others (5%) like prosthetic valves, endocarditis, intracardiac tumours (atrial myxoma)
- *Noncardiac (10%)*:
    - Aneurysms (5%); atheromatous plaque in proximal artery, paradoxical (1%)
- *Idiopathic* is 10%
- Others (4%) like:
    - Cervical rib causing poststenotic dilatation of subclavian artery can cause emboli
- *Cryptogenic*—an unknown source (5%)—after investigations source is not found

- ***Venous emboli*** are due to DVT causing pulmonary embolism.
- ***Venous-arterial paradoxical emboli:*** Seen in intra-cardiac shunt (ASD) or intrapulmonary shunts (A-V malformations) *(Osler-Weber-Rendau syndrome)*
- ***Fat embolism***.
- ***Air embolism***.

Effects of Arterial Embolism

- *Brain:* Blockage at *middle cerebral artery* causes hemiplegia, transient ischaemic attacks (TIA), visual disturbances
- Blockage at *central retinal artery* causes amaurosis fugax or permanent blindness
- Blockage at *mesenteric vessels* causes intestinal gangrene
- Blockage at *renal artery* leads to haematuria, loin pain
- Blockage at *limb vessels* causes pain, pallor, pulseless, paraesthesia, paresis, ulceration, gangrene
- Most common site of arterial emboli is *common femoral artery*

Sites of Lodging of Emboli

The most common site is lower limbs (75%). 10% brain; 10% upper limb; 5% superior mesenteric and renal arteries.

In the lower limb the most common site is at the bifurcation of common femoral artery (40%); popliteal artery (15%); common iliac artery (12%); aortic bifurcation (10%).

Features of Embolism

- Earlier history of claudication is absent but history suggestive of disease for source of emboli will be present.
- Sudden, dramatic, rapid development of pain with numbness.
- Limb becomes rapidly cold and mottled with blebs.
- Loss of sensation and movements.
- Absence of distal pulses but forcible, expansile, prominent proximal pulse. For example—prominent femoral artery pulsation with embolic bock at popliteal level.
- Toxic features.
- Collapsed veins, cold limb distal to the level of block, oedema and presence of blebs distally.
- Muscle which is soft normally while palpating will feel doughy initially but later becomes stiff. Once stiffness of muscle is found embolectomy benefit is bleak.

Differences between embolism and thrombosis

Features	*Embolism*	*Thrombosis*
Source	Present	Not present
Pulse	Proximal and contralateral pulses normal	Same and opposite side pulses may be absent
Temperature	Severely cold	Cold or normal
Angiography	Sharp cut off sign	Diffuse tapered disease
Collaterals	Very few—not well-developed	Well-developed

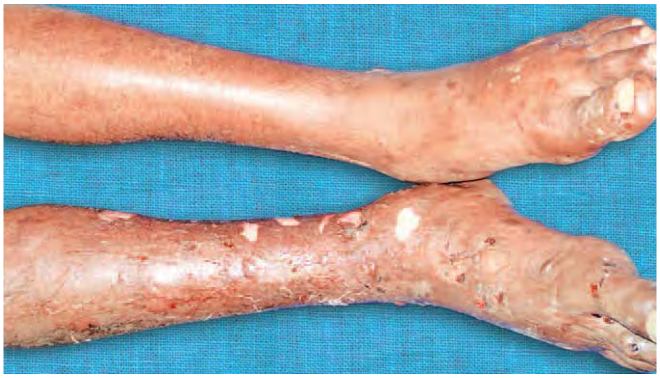

Fig. 1.373: Right leg showing features of acute embolism.

Investigations for Arterial Embolism

- Emergency Doppler angiogram, ECG and echocardiography. Angiogram is *gold standard* in all acute limb ischaemia. It differentiates between embolism and thrombosis; status of vessel proximally and distally. Angiography should ideally be done from contralateral limb or through left brachial.
- Relevant tests for origin of emboli. Prothrombin time, APTT, BT, CT, platelet count should be done.

Note:
Once embolism occurs irreversible changes occur distally in 6 hours, so ideal period for intervention is within 6 hours.

Classification of Severity of Acute Limb Ischaemia

Class I: Viable—no pain; no neurological deficit; Doppler shows audible signal. Venous flow present.
Class IIa: Marginally threatened—no pain; numbness/paraesthesia; no audible Doppler signal. Venous flow present.
Class IIb: Immediately threatened—persistent pain; sensory and motor loss; no Doppler signal. Venous flow present.
Class III: Irreversible—paralysis and anaesthesia. No venous flow.

Ischaemia up to class IIb with normal venous flow is called as **Early**; ischaemia which is class III, with muscle rigor, marbled skin and without any venous flow is **Late**. This late ischaemia is more likely to land with amputations even though revascularization can be tried.

Treatment

Treatment of Embolism and Thrombosis of Acute Limb Ischaemia

Immediate infusion of 5000-10,000 units of IV heparin and relief of pain are needed first.

Surgical

- *Embolectomy* (*surgical exploration and removal of clot*) is the choice for embolus. It is done either by interventional balloon 5 French (Fogarty, 1963) embolectomy or open method. It is the standard treatment for arterial embolism. It can be repeated several times until adequate bleeding occurs.
- For acute thrombosis causing acute limb ischaemia, open thrombectomy with or without bypass may be the surgical treatment; but it is not the standard treatment for acute thrombosis (Standard is thrombolysis, Dotter and co, 1974).

Endovascular therapy

- Intrarterial thrombolysis using *urokinase*.
- Percutaneous mechanical thrombectomy—it is done either by suctioning clot via catheter or dissolution of thrombus by pulverization and aspiration by high speed motors or fluid jets.
- Ultrasound accelerated thrombolysis using catheter based or transdermal using acoustic cavitation to ablate thrombus.

Embolectomy

- It is done as early as possible as an emergency operation.
- Under fluoroscopic guidance, *Fogarty catheter* (interventional radiology) is passed beyond the embolus and balloon is inflated. Catheter is withdrawn out gently with embolus. Procedure has to be repeated until embolectomy is completed and good back bleeding occurs. Angiogram is repeated to confirm the free flow.
- Postoperatively initially heparin and later oral anticoagulants are used. Procedure is done under general anaesthesia or local anaesthesia.
- *Open arteriotomy and embolectomy* can be done by *direct approach* and later the arteriotomy has to be sutured. Postoperatively anticoagulants and antibiotics are given.

***Intra-arterial thrombolysis using fibrinolysins*:** After passing arterial catheter, angiogram is done and agents are injected intraarterially through the arterial catheter.

Drugs used are:

- Streptokinase (here lysis occurs in 48 hours).
- Urokinase. It is commonly used for thrombolysis. It converts plasminogen to plasmin which breaks fibrin clots. Initial bolus of 2,50,000 IU is given followed by an infusion of 4,000 IU/min for 4 hours, later continuous infusion of 2,000 IU/min to complete the lysis. Even though controversial, it is of usual practice to infuse 1000 IU/hour of heparin to prevent new thrombus formation. Check angiography should be done during therapy. Multiholed catheter (5 French) is used for infusion.
- Tissue plasminogen activator (TPA)—Alteplase, Reteplase—here lysis occurs in 24 hours. TPA is *better and ideal*; but it is costly. It has very less side effects.
- TPA pulse-spray method—here lysis occurs in 6 hours.

Contraindications for thrombolysis

- Recent stroke
- Recent major surgery or major bleed like of varices
- Recent eye surgery
- History suggestive of or confirmed active duodenal/gastric ulcers
- Pregnancy
- Uncontrolled hypertension or coagulation disorders

***Note*:**
Heparin should not be used concomitantly with fibrinolytics.

Anticoagulant therapy to prevent recurrent emboli.

Immediate infusion of heparin 5000 units intravenously to prevent further extension of thrombus is needed.

Advantages of thrombolysis	***Disadvantages of thrombolysis***
• Gentle angiographic clot removal	• Useful only for class I and IIa acute ischaemia
• Survival and limb salvage is equal	• Bleeding at the site and elsewhere is possible
• It is mainly useful for acute thrombus	• 25% rate of failure
• For embolus it is used often as an adjunct along with embolectomy	
• It avoids surgery	

Complications of Revascularisation in Acutely Ischaemic Limb

- *Reperfusion* injury
- '*No reflow' phenomenon*: It is due to severe capillary oedema causing poor peripheral tissue hypoperfusion in spite of major vessel revascularisation.
- *Acute compartment syndrome* can occur due to massive ischaemic oedema especially of skeletal muscles deep to deep fascia which compress on venules exceeding tissue interstitial pressure causing further compromise in tissue perfusion. Compartment pressure when measured using transducer needles will be more than 40 mmHg or > 30 mmHg for 3 hours or above the mean arterial pressure. Presents with muscle weakness, sensory changes, leg pain which is aggravated by dorsiflexion of toes.
- Other complications are—sepsis, reblock, bleeding and catheter related complications.

REPERFUSION INJURY

- It occurs after reestablishment of arterial flow to an ischemic tissue bed which further leads to tissue death causing specifically *peripheral muscle infarction*. It is due to sudden release of oxygen free radicals which blocks the microcirculation, with release of high levels of potassium (hyperkalaemia) and myoglobin (myoglobinaemia and myoglobinuria). Haemodynamically patient becomes unstable with lactic acidosis, intracellular changes, interstitial oedema and cardiac dysfunction. It is often life-threatening.
- *Haimovici triad* of revascularisation injury (1960)—(1) Muscle infarction; (2) Myoglobinuria; (3) Acute renal failure.
- Severe ischaemia causes oedema in the muscular compartment with raise in compartment pressure more than the essential capillary perfusion pressure causing *acute compartment syndrome*. It is common in the anterior compartment of the leg. It is basically in the skeletal muscles deep to deep fascia. Compartment pressure when measured using transducer needles will be more than 40 mmHg or > 30 mmHg for 3 hours or above the mean arterial pressure. Muscle weakness, sensory changes, leg pain which is aggravated by dorsiflexion of toes.
- *'No re-flow' phenomenon* due to tissue oedema causes capillary perfusion block. Even though compartment syndrome and 'no reflow' phenomemenon are separate entities they are always seen together along with reperfusion injury.
- Metabolic acidosis, acute tubular necrosis causing acute renal failure and cardiac arrhythmias may set in and become life-threatening.
- **Features are**—toxaemia; oliguria; persistent pain and oedema in the leg with muscular tenderness; raised blood urea and serum creatinine with features of acute ischaemia in the limb. Raised creatinine level (renal failure), creatine kinase (muscle lysis) are typical.
- Treatment:
  - Mannitol to prevent renal failure; fluid therapy.
  - *Fasciotomy* to reduce raised compartment pressure. All four compartments of lower limb should be decompressed surgically. Long vertical lateral deep fasciotomy incision in the calf behind the fibula along the deep fascia and its fibular attachments is a must. Bleeding is common after fasciotomy as patient is heparinised. Infection of the wound can occur. Later, once the patient is stabilised and oedema subsides with healthy wound, secondary

suturing or skin grafting is done. If after fasciotomy, patient survives then it is with eventual development of *Volkmann's ischaemic contracture*.

- Antibiotics and supportive therapy.

SADDLE EMBOLUS

It is an embolus blocking at *bifurcation of aorta.*

Causes

- Mural thrombus after myocardial infarction.
- Mitral stenosis with atrial fibrillation.
- Aortic aneurysm.

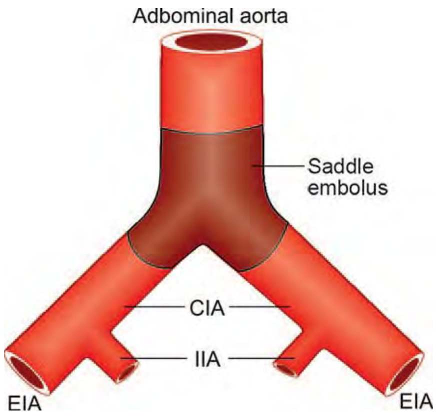

Fig. 1.374: Saddle embolus blocking the bifurcation of abdominal aorta. It causes severe, rapid, dramatic symptoms.

The embolus which blocks at aortic bifurcation is usually large.

Clinical Features

- Features of sudden, rapidly progressive ischaemia in both lower limbs.

Note:
In aortic bifurcation thrombus, there is earlier history of claudication in the buttock often with *Leriche's syndrome*. Symptoms are slow and gradual but not dramatic. Collaterals between aorta and iliac arteries have well-formed and so sudden, rapid development of gangrene will not occur.

- Gangrene of both lower limbs.
- Features of associated infection.

Investigations

- Arterial Doppler, aortic angiogram.
- U/S abdomen.

Treatment

- Initially, heparin is injected intravenously—10,000 units and later 5,000 units subcutaneously 8th hourly.
- Embolectomy can be done using Fogarty's catheter.
- Open arteriotomy and embolectomy can also be tried.
- Antibiotic prophylaxis is given to prevent infection.

EMBOLECTOMY

Indications

- Acute embolic blockade of artery commonly seen in common femoral, cranial vessels, mesenteric vessels.
- It should be done within 6 hours as after 6 hours irreversible changes occur.
- It is usually done under local anaesthesia under C arm guidance with anaesthetist monitoring the patient. It can be done under spinal or general anaesthesia.

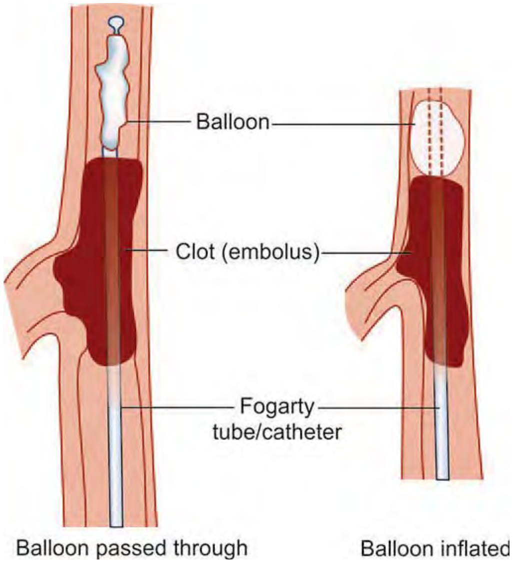

Fig. 1.375: Embolectomy technique.

Methods

- *Interventional method* is usually employed using Fogarty's catheter. Good back bleed signifies completeness of embolectomy.
- *Open arteriotomy method* is done directly over the artery followed by suturing the artery.

Note:
- *Intraoperative arteriogram* is a must to confirm the adequacy of blood flow and completion.
- *Intraoperative thrombolysis* as an adjunct to save the limb using urokinase 2,50,000 IU for minutes into distal artery may be beneficial.
- *Prophylactic fasciotomy* is needed in delayed cases to prevent reperfusion injury.
- *Postoperative systemic heparin* and later oral anticoagulant is given.
- *Treatment* for atrial fibrillation, atherosclerotic stenosis and other causes is needed.

Complications

- Bleeding
- Sepsis
- Thrombosis
- Narrowing
- Incomplete removal

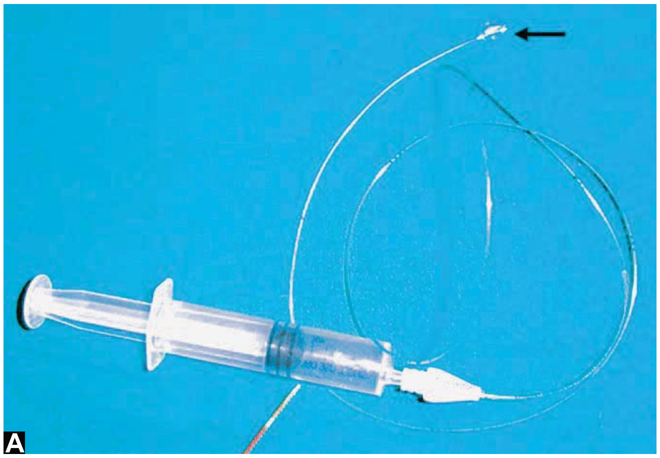

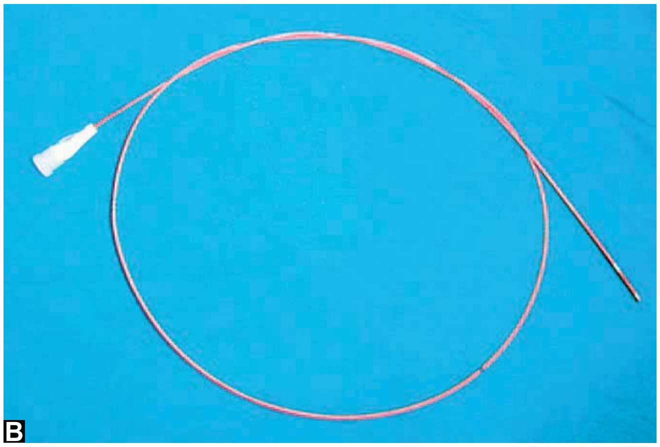

Figs 1.376A and B: Fogarty's catheter. It is 80 cm in length with 4 to 7 French size. It is used for embolectomy. Note the inflated balloon at the tip.

After Embolectomy

Patient is placed in ICU care.
Monitoring with—PTT, thromboplastin time.

FAT EMBOLISM

It is commonly seen after *fracture femur, tibia,* or *multiple fractures* and occasionally following electroconvulsive therapy, usually occurs in 72 hours.

It is due to aggregation of chylomicrons, derived from bone marrow, causing fat embolism.

It is often a fatal condition.

Clinical Features

- *Cerebral*: Drowsy, restlessness, disoriented, constricted pupils, pyrexia, coma.
- *Pulmonary:* Cyanosis, tachypnoea, right heart failure, froth in mouth and nostrils, fat droplets in sputum, eventually respiratory failure.
- *Cutaneous*: Petechial haemorrhages in the skin.
- *Retinal artery emboli* is the earliest sign to appear, causing *striae haemorrhages, fluffy exudates* confirmed on fundoscopic examination.
- *Kidney*: Blockage in renal arterioles results in fat droplets in urine.

Treatment

- Oxygen.
- Heparinisation.
- Low molecular weight dextran.
- Ventilator support and ICU management.

AIR EMBOLISM

Causes

- Through venous access like IV cannula, most common cause.
- During artificial pneumothorax.
- During surgeries of neck and axilla.
- Traumatic opening of major veins sucking air inside, causing embolism.
- During fallopian tube insufflation.
- During illegal abortion.
    - Amount of air required to cause air embolism is 15 ml**.**
    - When the air enters the right atrium, it gets churned up forming a foam which enters the right ventricle and blocks the pulmonary artery.
    - ***Mill-Wheel murmur*** heard over the precordium through a stethoscope is diagnostic.
    - During open heart surgery/therapeutic pneumothorax, by accidental pulmonary vein puncture or in atrial septal defect (ASD) air may enter left side of the heart (*paradoxical air embolism*) causing coronary block or cerebral air embolism.
    - Through paravertebral veins also air embolism to brain can occur.

Treatment

Patient is placed in Trendelenburg left side up position. By passing a needle, the air has to be aspirated from the right ventricle. Often requires life-saving open thoracotomy to aspirate the excess air causing the block.

THERAPEUTIC EMBOLISATION

Indications

- Haemangiomas.
- AV fistulas.
- Malignancies like renal cell carcinoma, hepatoma.
- Cerebrovascular problems.
- To arrest haemorrhage from GIT, urinary and respiratory tract.
    - In bleeding duodenal ulcer or gastric ulcer, embolisation is done to occlude gastroduodenal artery or left gastric artery respectively.
    - It is also useful in bleeding oesophageal varices, secondaries in liver (mainly due to carcinoids), hepatoma.

Materials used for therapeutic embolisation

- Blood clot
- Gel foam
- Balloons
- Quick setting plastics
- Stainless steel coils
- Human dura
- Plastic microspheres
- Ethyl alcohol
- Wool

CAISSON'S DISEASE OR DECOMPRESSION DISEASE

It occurs due to rapid decompression from high altitude, aircraft, compressed air chambers, deep sea divers causing bubbling of nitrogen which blocks the small vessels.

- In joints and muscles it causes excruciating pain *(bends).*
- Spinal cord ischaemia causing neurological deficits.
- Lungs may be affected causing choking with chest pain, tightness and dry cough.

Treatment

- Oxygen therapy.
- Recompression and gradual decompression in special chamber.

ANEURYSM

> *There is no disease more conducive to clinical humility than aneurysm of the aorta.*
>
> —**William Osler,** Circa 1900

It is an abnormal permanent dilatation of localised segment of arterial system. Diameter will be *50% more* than expected normal diameter of that artery in aneurysm. Atherosclerosis which is the most common (90%) facilitating cause of aneurysm is due to destruction and *loss of stability* of tunica media.

True aneurysm contains all three layers of artery.

False aneurysm contains single layer of fibrous tissue as wall of the sac and it usually occurs after trauma.

Types

- *Fusiform*—uniform dilatation of entire circumference of arterial wall
- *Saccular*—dilatation of part of circumference of the arterial wall
- *Dissecting*—through a tear in the intima blood dissects between inner and outer part of tunica media of the artery

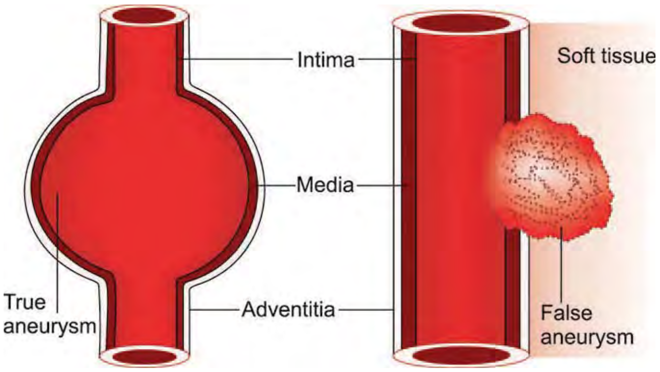

Fig. 1.377: True and false aneurysms. In true type, all layers are intact. In false type all layers breached with haematoma having a false capsule.

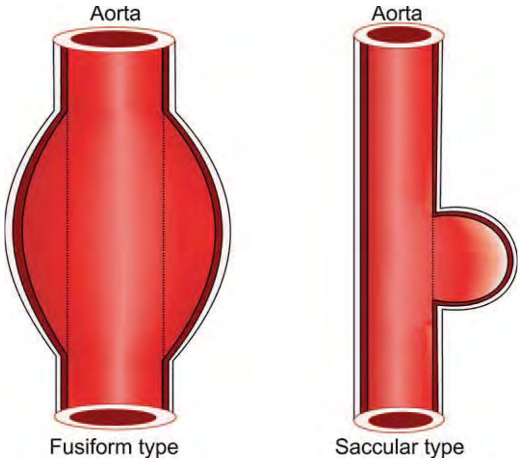

Fig. 1.378: Fusiform and saccular types of aneurysms.

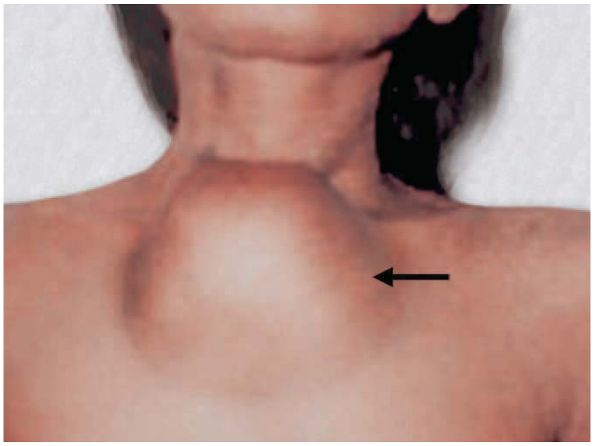

Fig. 1.379: Thoracic aortic aneurysm.

Causes

- ***Acquired:***
  - *Degenerative:* Atherosclerosis (*most common cause*); mucoid degeneration of intima and media (in South African young Negroes).
  - *Traumatic*: Direct; indirect like in post-stenotic dilatation by cervical rib; traumatic AV aneurysmal sac; aneurysm due to irradiation (due to dryness and destruction of vasa vasorum causing weakening).
  - *Infective*: Syphilis; mycotic; tuberculosis (in lung); arteritis; acute sepsis.
  - *Collagen diseases* like Marfan's syndrome, polyarteritis nodosa, Ehler-Danos syndrome.
- ***Congenital:***
  - Berry aneurysm; cirsoid aneurysm; congenital AV fistula.

Sites

- Aorta.
- Femoral.
- Popliteal.
- Subclavian.
- Cerebral, mesenteric, renal, splenic arteries.

The most common is *true, fusiform, atherosclerotic, aortic aneurysms.*

Berry aneurysms are multiple aneurysms occurring in circle of Willis.

Effects and complications of aneurysm

- Thrombosis and distal ischaemia
- Release of emboli causing acute arterial occlusion
- Pressure effects on bone (erosion); skin; veins (oedema); nerves (pain, paraesthesia); stomach (erosion—haematemesis); oesophagus (dysphagia)
- Rupture
- Infection of aneurysm

Clinical Features of Aneurysms

- Swelling at the site which is *pulsatile (expansile)*, smooth, soft, warm, compressible, with thrill on palpation and bruit on auscultation. Swelling reduces in size when pressed proximally.
- Distal oedema due to venous compression.
- Altered sensation due to compression of nerves.
- Erosion into bones, joints, trachea or oesophagus.
- Aneurysm with thrombosis can throw an embolus causing gangrene of toes, digits, extending often proximally also.

Differential Diagnosis

- Pyogenic abscess: Abscess has to be always confirmed by aspiration; especially in axilla, popliteal region, groin.
- Vascular tumours.
- Pulsating tumours: Sarcomas, pulsating secondaries.
- Pseudocyst of pancreas mimics aortic aneurysm.
- AV fistula.

Investigations

- Doppler study, duplex scan, angiogram, DSA.
- Tests relevant for the cause, like blood sugar, lipid profile, echocardiography.

Treatment

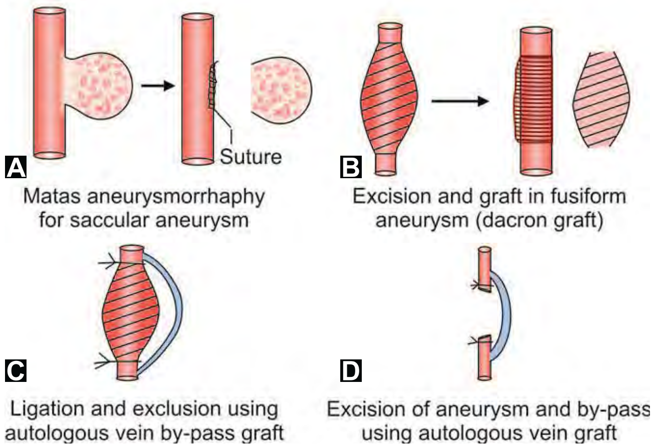

Matas aneurysmorrhaphy for saccular aneurysm

Excision and graft in fusiform aneurysm (dacron graft)

Ligation and exclusion using autologous vein by-pass graft

Excision of aneurysm and by-pass using autologous vein graft

Figs 1.380A to D: Different methods of aneurysm repair. (A) Matas aneurysmorrhaphy for saccular aneurysm. (B) Excision and Dacron grafting. (C) Ligation and exclusion of the aneurysm using autologous vein graft. (D) Excision of aneurysm and bypass using autologous vein graft.

- Reconstruction of artery using arterial grafts.
- Arterial endoaneurysmorrhaphy—MATAS. It is done usually for peripheral saccular aneurysm. Matas aneurysmorrhaphy may be restorative or endo-obliterative or reconstructive.
- Therapeutic embolisation.
- Clipping the vessel under guidance (e.g. cranial aneurysms).
- Older methods which are now not used but popular earlier were—wiring of the aneurysmal sac/wrapping of the aneurysmal sac/ligatures at different levels (ligation just proximal to aneurysmal sac—*Anel's;* ligation proximally proximal to an arterial branch—*Hunter's;* ligation just distal to aneurysmal sac—*Brasdor's*; ligation distally distal to an arterial branch—*Wardrop's*; ligature one proximal and another distal to aneurysmal sac—*Antylus'*).

MYCOTIC ANEURYSM

- It is a ***misnomer***.
- It is not due to fungus but due to bacterial infection.
- Common bacteria are gram-positive organisms *like Staphylococcus aureus (most common)* and *Streptococcus.*
- Common aetiology is bacterial *endocarditis* but could be any infective site.
- Common vessels involved are aorta, visceral, head and neck and intracranial.
- Commonly it is saccular, multilobed, with a narrow neck.
- Patient presents with fever, toxaemia and tender pulsatile mass if it is in the periphery.

Investigations

- Leucocytosis. Positive blood culture, MR or CT angiogram are relevant.

Treatment

- Broad spectrum antibiotics
- Resection of aneurysm; debridement and drainage of the infected aneurysm with adequate blood transfusions.
- Extra-anatomic bypass through uninfected tissue planes to avoid contamination of the graft.
- Long-term antibiotic therapy is necessary.
- It has got 25% mortality.

Note:

Microbial arteritis with aneurysm is a different entity is due to bacteraemia occurring in an atherosclerotic vessel due to *Salmonella* infection).

ABDOMINAL ANEURYSM

- Abdominal aortic aneurysm is the *most common* aortic aneurysm.
- Splenic artery aneurysm is the *2nd most common* type.
- Incidence is 2%. It is more common in males.
- Transverse diameter of aorta in an aneurysm should be 3 cm or more.

- Common in elderly; common in males (4:1); chance of getting aneurysm in genetically related first degree relatives is 10 times more.
- Common in smokers (8:1with nonsmokers); in 55% of patients *Chlamydia pneumoniae* is identified.

***Note*:**
Smoking is an important factor.

ABDOMINAL AORTIC ANEURYSM

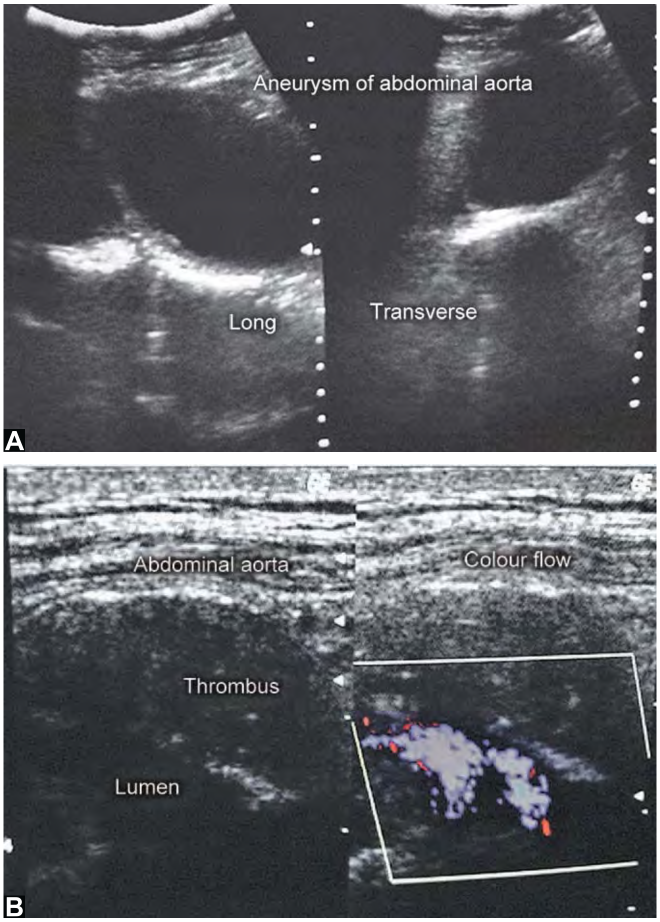

Figs 1.381A and B: Abdominal aortic aneurysm (A). Thrombosis of the aneurysm (B).

Causes

- *Atherosclerosis* (as degenerative process) is the *most common facilitating* cause (95%)—aortic wall contains smooth muscle cell matrix, elastin, collagen; *elastin* (in tunica media) is the main load bearing part with *collagen* (in adventitia) as safe net in the wall to provide tensile strength preventing aneurysm formation. *Elastin in medial layer* of aorta is *degraded* and reduced significantly in infrarenal aorta in relation to collagen, absence or less vasa vasorum in infrarenal aorta and atherosclerotic unstability of the medial wall of aorta cause infrarenal aorta more prone to develop aneurysm. Increased proteolytic activity of aortic medial wall due to increased *matrix metallo proteinases* (MMP) (derived from aortic smooth muscle cells and macrophages) cause elastin and collagen degradation and increase in diameter of aneurysm. *Collagen degradation in adventitia* causes rupture.
- *Familial* aortic aneurysm (associated with 25% of AAA) is more prevalent in females to reduce male to female ratio to 2:1. It is related to decrease in type III collagen, α1 antitrypsin and lysyl oxidase. Marfan's, Ehler Danlos syndromes are related genetically.
- Others: Syphilis, dissection, trauma, collagen diseases, infection, arteritis, cystic medial necrosis, association with *Chlamydia pneumoniae* (55%).

Classification I

- Infrarenal—most common 95%.
- Suprarenal—5%. Isolated suprarenal type is rare; it is usually associated with thoracic and or infrarenal types.

Classification II

- Asymptomatic.
- Symptomatic.
- Symptomatic ruptured.

Asymptomatic Type

- It is found incidentally either on clinical examination or on angiography or on ultrasound.
- Repair is required if diameter is over 5.5 cm on ultra sound.
- It is identified during routine abdominal palpation or while assessing or operating for some other abdominal conditions.

Symptomatic without Rupture **(Clinical features/ presentations)**

- It presents as *back pain, abdominal pain, mass abdomen* which is smooth, soft, nonmobile, not moving with respiration, vertically placed above the umbilical level, pulsatile both in supine as well as knee-elbow position with same intensity, resonant on percussion.
- Common in males (4:1); common in smokers.
- GIT, urinary, venous symptoms can also occur.
- Hypertension, diabetes, cardiac problems should be looked for and dealt with.
- In infrarenal type upper border is clearly felt.
- Lower limb ischaemia and embolic episodes can occur.
- 5% present as *inflammatory aneurysm* adherent to ureters, left renal vein, inferior vena cava and duodenum. Expanding aneurysm blocks lymphatics causing inflammation and fibrosis; or it may be due to infection and fibrosis of earlier localised ruptured abdominal aortic aneurysm. Such chronically inflamed aneurysm will not rupture further; but it is always symptomatic with fever and severe pain in abdomen and back. It needs surgical repair through retroperitoneal approach.

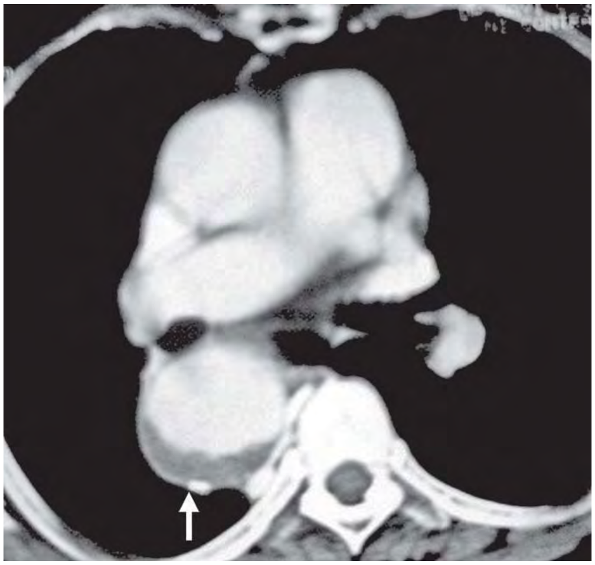

Fig. 1.382: CT scan picture showing right arch aorta (anomaly) with thrombosed aneurysm.

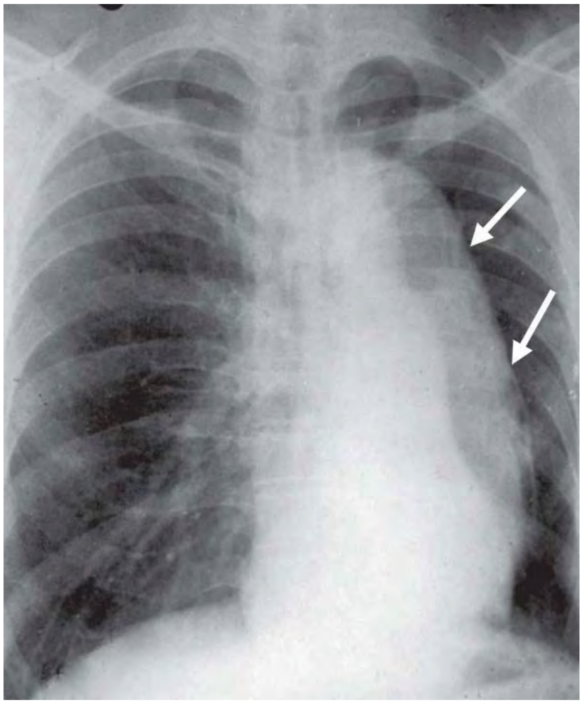

Fig. 1.383: Chest X-ray showing aortic aneurysm.

- *Aortocaval fistula,* presents as high output cardiac failure with continuous bruit in abdomen and severe lower limb ischaemia (*steal phenomenon*).
- *Aortoenteric fistula* is due to erosion of aneurysm into 4th part of duodenum presenting as GI bleed, malaena, shock. It is treated by duodenal closure, aortic ligation, aneurysmal exclusion with extra-anatomic bypass graft with gastrojejunostomy. Contamination is the major threat here.
- Being a retroperitoneal mass *back pain is common*—may be due to retroperitoneal stretching, nerve irritation or vertebral erosion.
- Aneurysm in a patient with horseshoe kidney which is anterior to aorta is difficult to manage. Left retroperitoneal approach is needed. EVAR is not possible.

Investigations

- Blood urea, serum creatinine.
- U/S (*most widely used* noninvasive test; but neck of the aneurysm, dimensions and relation to renal arteries are difficult to assess), aortogram, DSA, CT scan (*most precise*). *U/S is an effective screening tool*. Screening is done in cardiovascular patients in men (60-85 years), in women (60-85 years); men and women above 50 years with family history; annually in asymptomatic AAA with 4.0-4.5 cm size, with size > 4.5 cm once in every 6 months.
- CT angiogram, MR angiogram.
- Blood sugar, lipid profile, other relevant investigations like ECG, echocardiography, cardiac and pulmonary assessment.

Note:

X-ray will show *eggshell* calcification. CT scan is *more reliable and precise* investigation of choice—gives better information regarding extent on sides/neck, size, dimensions, size and site of the thrombus, calcification, relation of renal arteries, inflammation and fibrosis and adjacent tissues. MRI may be better only in renal failure patients.

Complications of abdominal aortic aneurysm

- Rupture, infection
- Thrombosis, embolism
- Distal ischaemia/gangrene
- Aortocaval fistula formation
- Aortoenteric fistula
- Erosion of vertebra
- Spinal cord ischaemia when thrombosis develops

Differential Diagnosis

- Retroperitoneal mass, pseudocyst of pancreas, retroperitoneal cyst mimic abdominal aortic aneurysm especially when it is thrombosed.
- Mesenteric ischaemia, acute pancreatitis, perforated duodenal ulcer may mimic ruptured aneurysm.
- Other conditions causing back pain like disc prolapse, sciatica.

Treatment

Conservative/Medical Treatment

- It is done in *low-risk* abdominal aortic aneurysm (age below 70 years; active physically without cardiac, respiratory, renal impairment and noninflammatory aneurysm); if aneurysm size is < 5 cm; if growth rate is < 0.5 cm/year.
- It includes risk factor modifications; stopping smoking; control of blood pressure (propranolol), cholesterol; usage of drugs—alpha blockers, elastase inhibitors (NSAID—indomethacin), matrix metalloproteinases (MMP) inhibitor (doxycycline).

- Periodic size measurement of an aneurysm using ultrasound once in 6 months to find out growth rate is essential during conservative treatment.

Surgical Treatment

- ***Indications for surgery***
    - Asymptomatic aneurysm more than 5.5 cm.
    - Growth rate more than 0.5 cm/year.
    - Painful, tender aneurysm.
    - Thrombosed aneurysm, aneurysm with distal emboli.

Low-risk abdominal aortic aneurysm—age below 70 years; active physically without cardiac, respiratory, renal impairment and noninflammatory aneurysm. Here surgical mortality is < 3%.

Intermediate risk—sedentary; stable coronary disease; mild COPD; creatinine 2-3 mg%; inflammatory/suprarenal aneurysm. Here surgical mortality is 3-7%.

High-risk—restricted daily works; significant coronary disease; dyspneic COPD; creatinine > 3 mg%; liver failure status. Here surgical mortality is > 10%.

- ***Open surgical repair***
    - It is called as *endo-aneurysmorrhaphy with intraluminal graft placement* (Crawford, 1960). It is done under GA with epidural support. Major challenges during anaesthesia are—blood loss, haemodynamic control, problems during clamping and declamping of aorta, temperature control, renal hypoperfusion, left ventricular strain.
    - Incision is commonly lengthy midline *transperitoneal* or supraumbilical transverse. *Retroperitoneal appro*ach is used in horseshoe kidney, abdominal wall stoma, inflammatory aneurysm, suprarenal extension, peritoneal dialysis, hostile abdomen. Retroperitoneal approach favors rapid control of proximal aorta but prevents visualisation of abdomen.
    - After laparotomy, duodenum and small bowel are retracted laterally and above; left renal vein which is in front of aorta is dissected and retracted; occasionally it may require to be ligated and it is safer provided left gonadal and left suprarenal veins are intact. Distal arterial clamps are applied first along common or external and internal iliac arteries on both sides. Proximal aortic clamp is applied at infrarenal level. Aorta is opened longitudinally midline towards right to avoid injury to orifice of inferior mesenteric artery. Atheroma, thrombus is removed until adequate backbleeding occurs. Lumbar vessels are ligated from the luminal side. Knitted Dacron graft after preclotting or woven Dacron graft or ePTFE tube graft is used. Graft is anastomosed above and below using polypropylene sutures (4 zero). Inferior mesenteric artery can be reimplanted. Clamps are released first below. Colonic and limb perfusion is checked for adequacy. Graft is covered with aneurysmal sac.
    - *Minimal incision aortic surgery (MIAS)* is done in thin individual with midline abdominal incision 12 cm in length with its 9 cm part above the umbilicus. Specialised retractors and vascular clamps are used for this. Advantages are less postoperative pain, ileus and incisional hernia.

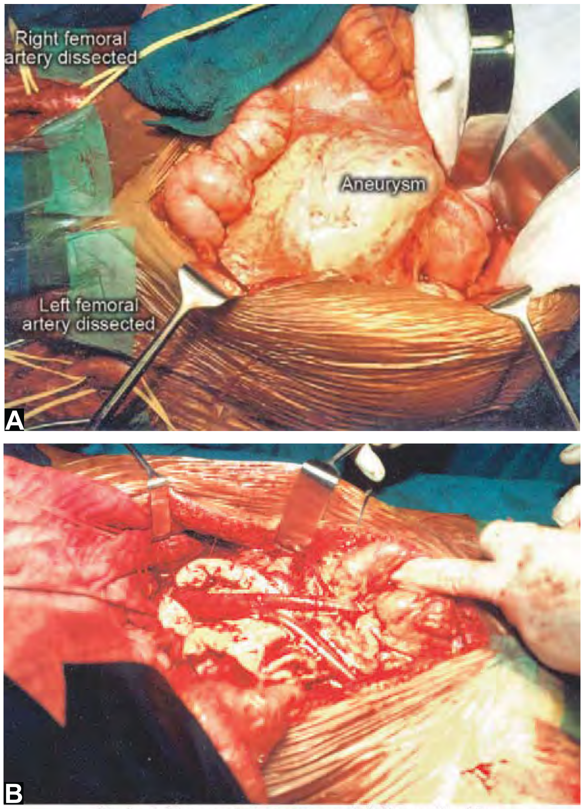

Aorto bifemoral dacron graft (Woven) with right femoral exposed

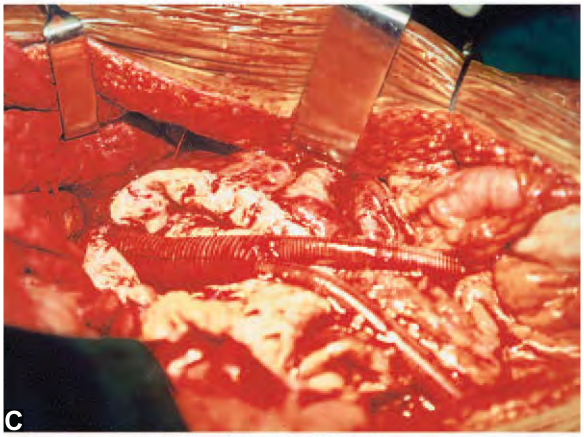

Graft in place in retroperitoneum

Figs 1.384A to C: Abdominal aortic aneurysm with aortofemoral graft placement (Courtesy: Dr Ashok Shetty, mch, Cardiothoracic Surgeon, Mangalore).

- ***Endovascular aneurysm repair (EVAR)***
    - In 1991 Juan *Parodi* and Julio Palmaz first did EVAR. It is less invasive, less morbid with less mortality rate and shorter hospital stay. It is basically aneurysm exclusion method. It is useful in old age and patients who are not fit for surgery. EVAR is basically a prophylactic procedure. EVAR is indicated if aneurysm is less than 5.5 cm in men and less than 5.0 cm in women. It is usually done in patients after 65 years.

- It is endoluminal stent graft placement into the aneurysmal segment of aorta using interventional radiology with Seldinger's technique approach through femoral artery.
- Dacron or ePTFE with integral metallic stent for support and firm attachment is used as stent/endovascular prosthesis.
- One aortic and iliac (of one/same side) stent is commonly used together which is passed through same side common femoral. Other iliac is maintained with a separate stent approached through opposite common femoral.
- Procedure can cause endoleak, thrombosis, embolism, malposition/displacement of stent, sigmoid ischaemia, renal failure, failure of stent function causing recurrence and infection.

Symptomatic Ruptured Aortic Aneurysm

- Risk of rupture is 1%, if diameter is within 5.5 cm in size. Risk increases to 20% once the diameter = 7 cm.

It may be *anterior rupture* (20%) into the free peritoneal cavity causing severe shock and *death* very early; or *posterior rupture* (80%) with formation of *retroperitoneal haematoma* of large size causing severe back pain, hypotension, shock, absence of femoral pulses and with a palpable mass in the abdomen.

- **Management of Ruptured Aortic Aneurysm**
  - Immediate diagnosis by ultrasound.
  - Resuscitation.
  - Massive blood transfusions (10-15 bottles).
  - Emergency surgery is the only life-saving procedure in these cases.

Patient has to be shifted to the operation theatre. Abdomen is opened. *Vascular clamps* or *bull dog* clamps are applied to the aorta above and below the aneurysm. Adventitia is opened and the clot is removed. Aneurysm is excised and the *arterial graft* either *PTFE (Polytetrafluoroethylene), knitted dacron graft,* or *woven dacron graft* is placed. The graft is sutured to the vessel above and below using monofilament, nonabsorbable suture material, polypropylene 5-zero.

Complications of surgery

- *MI* is the most common cardiac complication in perioperative and in first 2 days of postoperative period
- Haemorrhage and haemodynamic complications
- Renal failure—is most common noncardiac complication
- Colonic ischaemia—10% due to poor IMA circulation
- Sexual dysfunction
- Aortoduodenal fistula
- Aortovenacaval fistula
- Spinal cord ischaemia—paraplegia
- Paralytic ileus
- Distal thromboembolism—blue toe syndrome
- DVT
- Graft leak, graft thrombosis, graft failure
- Anastomotic disruption, pseudoaneurysm formation
- Prosthetic infection/migration

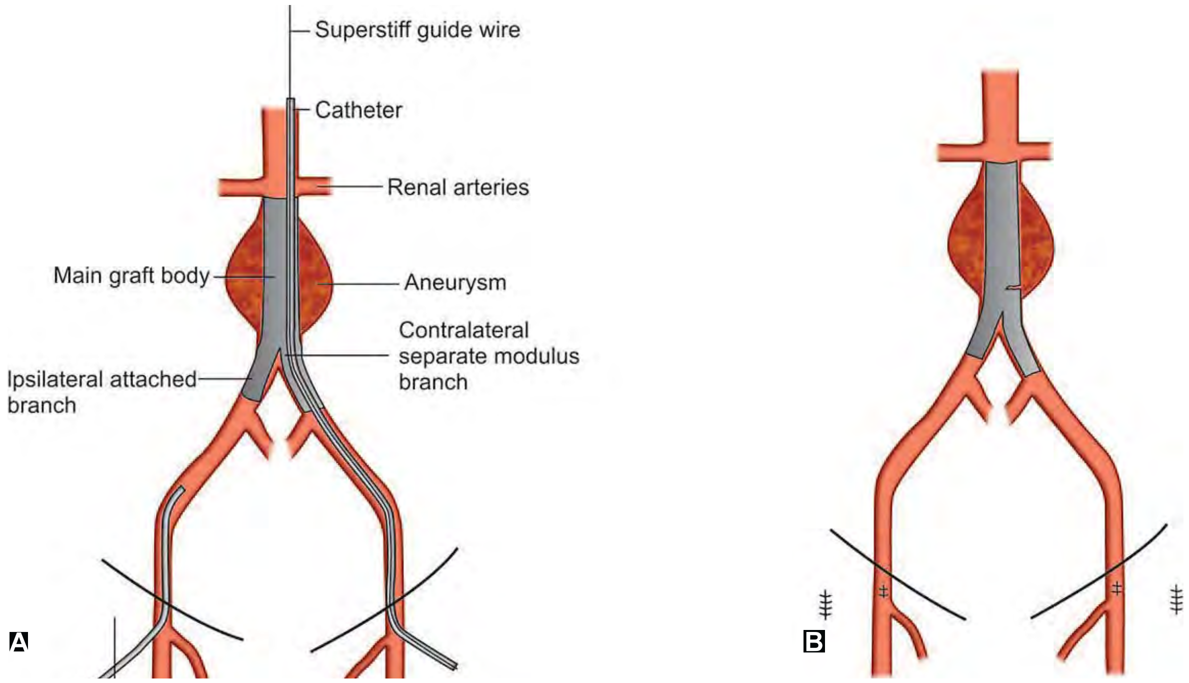

- Ipsilateral aortic and iliac single unit modulus is passed through same side CFA.
- Opposite side (contralateral side) branch CIA graft is deployed through superstiff guide wire.
- Through opposite CFA.

Completion of procedure with graft in place. Arteriotomy is closed and skin incision is closed.

Figs 1.385A and B: Endovascular aneurysm repair (EVAR). Aortic and one side iliac stent is used as one unit modulus which is passed through same side common femoral artery; opposite side iliac part is inserted as separate modulus through opposite CFA.

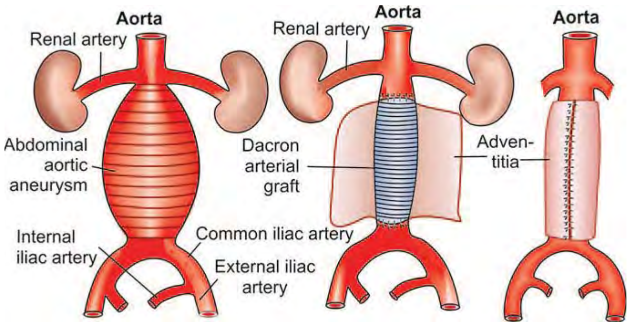

Fig. 1.386: *Infrarenal aortic aneurysm repair*. It is the most common site of aortic aneurysm. Adventitia of aorta is opened; aneurysm is excised; graft is sutured above and below; adventitia is wrapped around.

Remember

- Pulsation of an aneurysm is expansile. Pulsation may be absent if it is thrombosed
- Abdominal aneurysm of any size which is painful or tender should be operated
- Abdominal aortic aneurysm of any size causing embolus should be operated
- Abdominal aortic aneurysm more than 5.5 cm should be operated
- In ruptured abdominal aortic aneurysm emergency surgery is the only choice operation with rapid resuscitation; immediate opening and repair using graft. Systolic pressure in this patient should be just adequate to maintain the cardiac function but should not be more than 100 mmHg as it will cause more bleeding
- Anterior rupture is more dangerous than posterior rupture
- Endoluminal stenting is becoming popular

PERIPHERAL ANEURYSM

- *Peripheral aneurysms* are less common compared to aneurysms in the cavity. Such surface aneurysms are easily visible and better amenable for clinical examination. But same time it may be mistaken for abscess and inadvertent wrong attempt of incision and drainage can occur leading to disastrous consequences.
- Popliteal type is the *most common* one. Brachial, radial, femoral and axillary are other rarer sites. *Expansile pulsation* which is confirmed using two finger placement with thrill and bruit is typical. Infection, thrombosis make it less pulsatile mimicking an abscess.
- Erosion into adjacent bone and skin, rupture are known to occur. Distal emboli may lead into digital gangrene.
- Pressure on the affected artery proximally reduces the size, and eliminates the thrill/bruit; pressure distal to aneurysm increases the prominence of the aneurysm swelling with bounding pulsation.
- X ray, arterial Doppler, angiogram, echocardiogram are needed.
- *Treatment* is open repair using arterial graft or endovascular stenting.

Popliteal Aneurysm

- Is most common (70%).
- 65% are bilateral.
- 25% cases are associated with abdominal aortic aneurysm.
- 75% cause complications in 5 years.

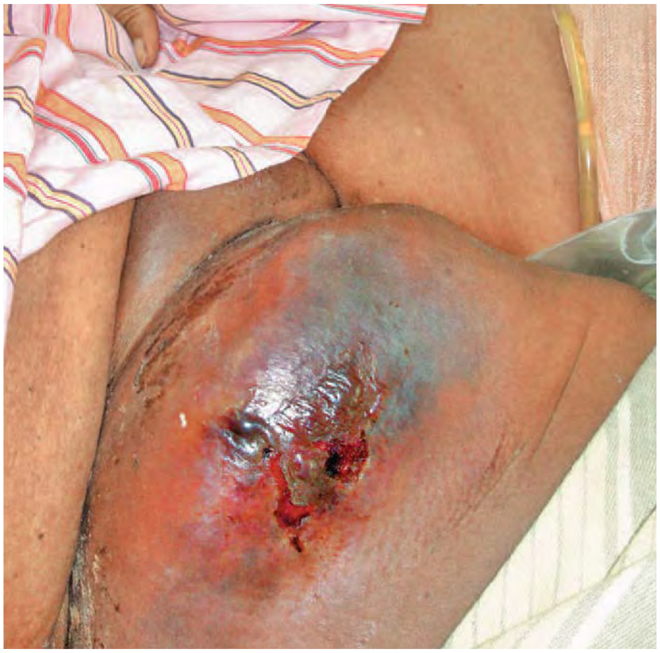

Fig. 1.387: Femoral artery aneurysm with impending rupture—needs emergency surgical intervention. It is rare type.

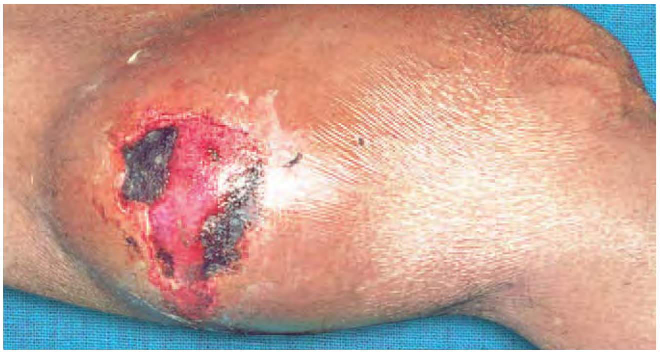

Fig. 1.388: Popliteal aneurysm about to rupture. It is the most common peripheral aneurysm.

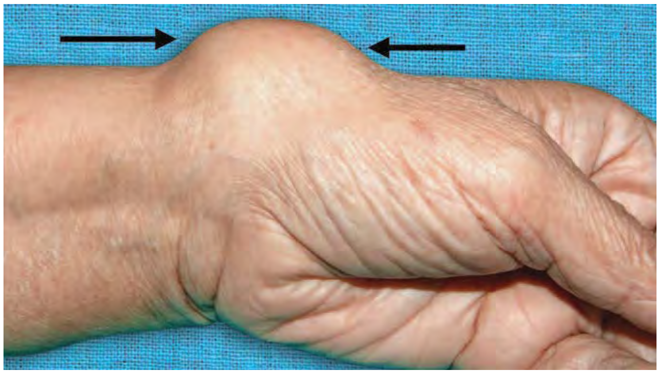

Fig. 1.389: Radial artery aneurysm.

Presentations

- Swelling in popliteal region which is smooth, soft, pulsatile, well-localised, warm, compressible, often with thrill and bruit. *It may mimic a pyogenic abscess*.
- Thrombosis and emboli from popliteal aneurysm can cause distal gangrene which may spread proximally and may lead to amputation.
- Rupture may cause torrential haemorrhage.

Investigations

- Duplex scan, Angiogram.
- CT scan, MRI.

Treatment

- Aneurysmorrhaphy.
- Repair with arterial graft using PTFE, Dacron.
- Endoluminal stenting.

CAROTID ARTERY ANEURYSM (EXTRACRANIAL)

Incidence is less than 4% of peripheral aneurysms.
Most common site: Common carotid artery bulb, often extends into the internal carotid artery.

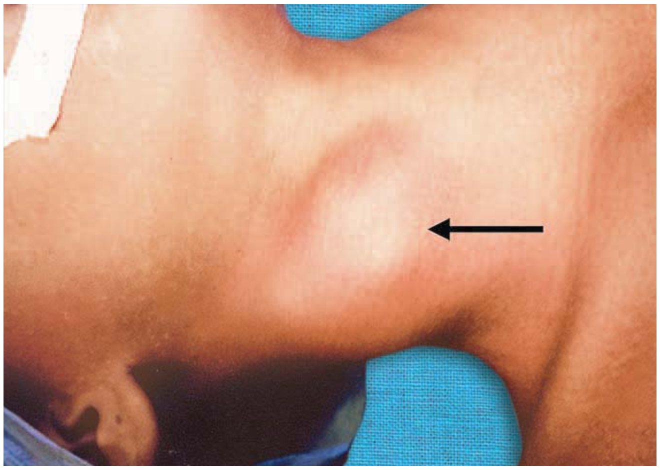

Fig. 1.390: Carotid aneurysm.

Causes

- Atherosclerosis, trauma.
- Syphilis, Marfan’s syndrome.
- Ehler-Danlos syndrome.
- Congenital.

Clinical Features

- 10% bilateral.
- Swelling in the neck at the level of the thyroid cartilage, below the angle of mandible.
- Pulsatile (expansile pulsation).
- Smooth, soft, nontender, horizontally mobile.
- Bruit felt.
- Neurological features due to embolic episodes (50%).
- Hoarseness of voice.
- Horner’s syndrome.
- Dysphagia.
- Swelling extending into the tonsillar bed.

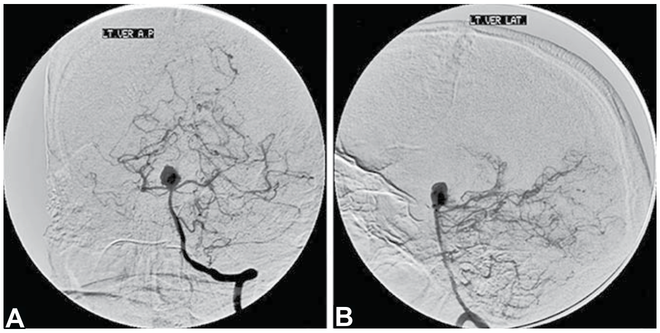

Figs 1.391A and B: Basilar artery aneurysm—angiogram (*Courtesty:* Dr Muralidhar Pai, MCh, Mangalore).

Differential Diagnosis

- Carotid body tumour.
- Neurofibroma arising from the vagus.
- Abscess in neck.

Complications

- Rupture.
- Thrombosis.
- Hemiplegia.

Investigations

- Doppler of neck, carotid angiogram.
- DSA, CT scan.

Treatment

- Reconstruction of the artery using vascular graft.
- Ligation of the bulb as a life-saving procedure, but results in hemiplegia.
- Intravascular stents.

DISSECTING ANEURYSM

It is a misnomer. It is not an aneurysm, only an aortic dissection.

It is the dissection of media of the aorta after splitting through intima creating a channel in the *media* of the vessel wall.

Causes

- Hypertension (It is associated in 80% of dissecting aneurysms).
- Cystic medial necrosis.
- Marfan’s syndrome and collagen diseases.
- Trauma.
- Weakening of the elastic layers of the media due to shear forces.

Today well lived makes every yesterday, a dream of happiness and tomorrow, a vision of hope.

Features

- It is always seen in thoracic aorta, common in ascending aorta (70%).
- It is uncommon in other parts of aorta or other vessels.
- It can occur in aortic arch or thoracic descending aorta.
- This dissected aortic channel gets lined by endothelium, often reopens distally into the aorta causing *double-barrelled aorta* which, in fact, prevents complications.
- It is commonly associated with aortic insufficiency. Atherosclerosis is *not a usual cause* for dissecting aneurysm.

I. Classification (DeBakey's)

- *Type I:* Dissection begins in ascending aorta extends into descending thoracic aorta (70%)
- *Type II:* Dissection originates in ascending aorta and extends only up to the origin of the major vessels. It is safer type with less complications
- *Type III*: Dissection begins in the descending thoracic aorta beyond the origin of the left subclavian artery

II. Stanford classification

- Proximal—includes DeBakey's Type I and II
- Distal—includes DeBakey's Type III

III. Dissecting aneurysm can be:

- Acute
- Chronic
- Healed dissecting aneurysm which communicates distally again to aorta as double barrelled aorta

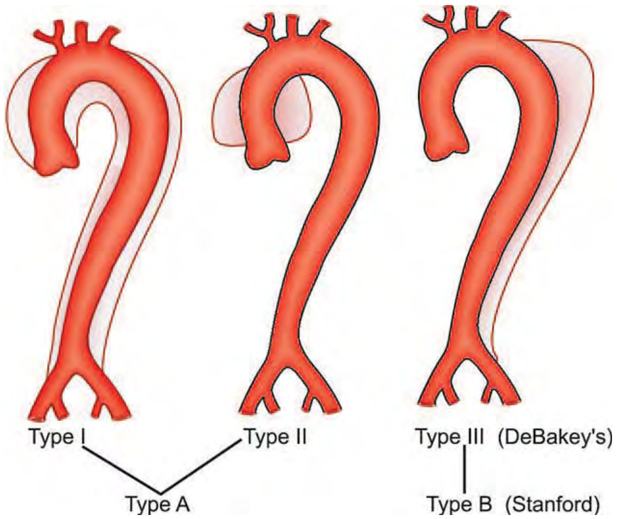

Fig. 1.392: Dissecting aneurysm.

Complications

- *Acute:* Rupture into the pericardium or pleura—*dangerous* type
- *Chronic:* Blockage of coronary vessels and major vessels like carotid and subclavian arteries with aortic insufficiency

Clinical Features

- Pain in the chest, back which is excruciating.
- Features of ischaemia due to blockage of different vessels.

Investigations

- Chest X-ray shows mediastinal widening
- Arterial Doppler
- Angiogram

Treatment

Antihypertensives.

Surgery: Using Dacron graft reconstruction of aorta has to be done with cardiopulmonary bypass.

Indications for surgery

- Progressive disease
- Significant ischaemia
- Impending rupture
- Type A aortic dissection

ERYTHROMELALGIA/ERYTHRALGIA

- It is severe burning pain and redness in the feet. Sensation of heat is so severe that patient keeps the feet in cold water to reduce it.
- It presents as episodic attack.
- There will be flushing in feet; prominent veins; warmness in the skin; severe hyperaesthesia is typical; even touching can be painful.
- It can be primary or secondary. *Secondary* is which is not uncommon is observed in arterial obliterative conditions, erythrocyanosis frigida, polycythaemia, gout and frostbite. *Primary* is due to unknown etiology; it is very rare.
- Vasodilators and sympathectomy may be beneficial.

LIVEDO RETICULARIS

It is a condition often associated with Systemic lupus erythematosus presenting as features of *arteriolar spasm and dilatation of venules* which is worsened by cold.

POLYARTERITIS NODOSA

- It is a necrotising inflammatory reaction with commonly microscopic polyarteritis and nodule formation, often of small and medium sized arteries (not capillaries), causing ischaemia of lower and upper limb.
- Visceral arterial (mesenteric) involvement (70%) can cause abdominal pain, GI bleed; mucosal ulceration and perforation of small bowel. Massive hepatic infarction, cholecystitis can develop.
- Renal artery can cause loin pain, haematuria, and renal hypertension.
- Coronary artery also can get involved causing myocardial infarction.
- Disease is common at bifurcation of medium/small sized arteries leading to localised aneurysms.
- It is common in males (3:1); fever, weakness, myalgia, arthralgia are early features.
- Presents with localised small aneurysms, like multiple 5-10 mm nodules, palpable along the course of the artery.
- In late stage presents with myocardial infarction, renal failure, sepsis, GI bled.
- HBsAg is positive in 40% patients of polyarteritis nodosa.

- Angiogram of renal, mesenteric, peripheral arteries will show aneurysms at branching points.
- Biopsy of tender nodule, tender muscle is useful for diagnosis.
- Treatment is prednisolone 60 mg daily with cytotoxic drugs.
- Prognosis is poor with rapid death in early years.

SCLERODERMA/SYSTEMIC SCLEROSIS

- It is *a progressive* disease causing fibrosis of skin, GI tract, lungs, heart and kidney.
- It is common in females (4:1) at 4th/5th decade.
- It is considered as vasculitis even though earlier considered as collagen disease.
- Pathology consists of cytotoxic endothelial injury causing interstitial oedema, severe fibroblast proliferation causing fibrosis of affected vessels, and dilatation and proliferation of remaining capillaries as telangiectasis.
- Thin epidermis, thick dermis with more collagen with absence of appendages and rete pegs are typical.
- Lower 2/3rd oesophagus is sclerosed (50%) with increased collagen in submucosa with atrophied mucosa and muscularis. Dysphagia is common.
- Diffuse interstitial fibrosis, thickening of alveolar membrane and pulmonary hypertension occurs.
- Synovial thickening causes arthritis; fibrosis of skeletal muscles; interstitial myocardial fibrosis causes bundle branch block, pericardial effusion.
- Glomerulosclerosis in kidney is common (50%). Renal failure is common.
- Fibrosis of thyroid, periodontal membrane can occur. Malabsorption syndrome is common due to small bowel involvement.
- Involvement of digital arteries present as Raynaud's phenomenon.
- Calcinosis, Raynaud's, oesophageal hypomotility, sclerodactyly, and telangiectasia are the presentation of CREST syndrome.
- ***Investigations***—anaemia, raised ESR, elevated IgG, presence of antinuclear antibodies and anticentromere antibodies (in CREST)—are different laboratory findings. Skin and peripheral arterial biopsy is confirmative.
- ***Treatment***—is difficult. Drugs like D pencillamine, colchicines, p amino benzoic acid, vitamin E, dimethyl sulfoxide, ranitidine are tried at various levels. Vasodilators, warming and massaging skin, avoiding detergent soaps, oil and hydrophilic ointment application are used.
  Steroids, oxygen therapy for irreversible pulmonary fibrosis; haemodialysis for renal failure; digitalis and other drugs for cardiac failure are needed later.
- Death is due to cardiac/pulmonary/renal failure.

ACROCYANOSIS (CRURUM PUELLARUM FRIGIDUM)

- It is persistent, painless cyanosis seen in fingers and often in legs with paraesthesia and chilblains affecting young females.
- It is chronic persistent arteriolar constriction with slow rate of blood flow.
- Trophic changes and ulcerations are not seen.
- Cyanosis *which is persisting* may aggravate on exposure to cold.
- It may be associated with endocrine dysfunction.

Treatment

- Vasodilators.
- Cervical sympathectomy (effective).

Raynaud's phenomenon	*Acrocyanosis*
• Episodic	• Persistent
• Painful	• Painless
• Acute arteriolar spasm	• Chronic constriction
• Ischaemic changes are common	• Ischaemic changes are not seen

GANGRENE

> *If the tips of (the patient's) fingers are falling off and are black, he will die.*
>
> **—(Anonymous),** circa 2000 BC

It is *macroscopic death* of tissue *in situ* (in continuity with adjacent viable tissue) with or without putrefaction.

It can occur in sites like:

- Limbs
- Appendix
- Bowel
- Testes
- Gallbladder

Causes

- Secondary to arterial occlusion like atherosclerosis, emboli, diabetes, TAO, Raynaud's disease, ergots.
- Infective: Boil, carbuncle, gas gangrene, Fournier's gangrene, cancrum oris.
- Traumatic: Direct, indirect.
- Physical: Burns, scalds, frostbite, chemicals, irradiation, electrical.
- Venous gangrene.

Clinical Features

- *Colour changes*: Pallor, greyish, purple, brownish black due to disintegration of haemoglobin to sulphide.
- *Absence of pulse*, loss of sensation, loss of function.
- *Line of demarcation* between viable and dead tissue by a band of hyperaemia and hyperaesthesia along with development of a layer of granulation tissue.

Differences between dry gangrene and wet gangrene

Dry gangrene	*Wet gangrene (Moist)*
Clear line of demarcation is seen	Line of demarcation is vague
Dry, shriveled, mummified	Oedmatous, putrified, discoloured (H_2S)
Slow, gradual loss of blood supply	Sudden loss of blood supply
Separation is by aseptic ulceration	Septic ulceration causes separation
Limits to the demarcation	Can extend proximally rapidly
Causes are atherosclerosis, TAO	Emboli, trauma are the causes
Limited amputation is sufficient	Major higher amputation is often needed

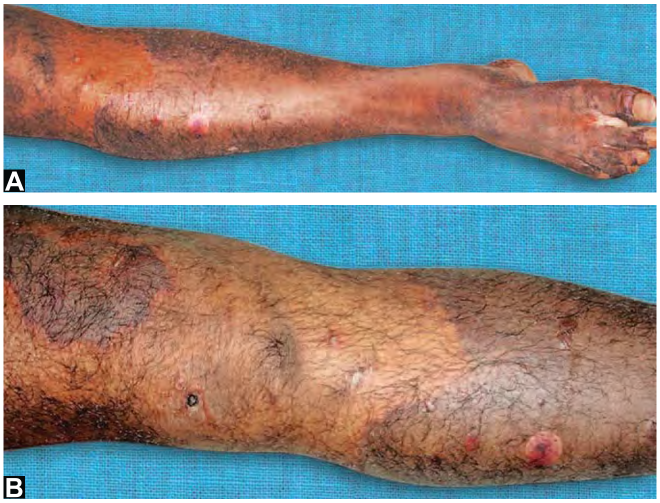

Figs 1.393A and B: Gangrene of foot and leg. Note the line of demarcation. But skip lesion proximally shown signifies more proximal spread of gangrene which may need above knee amputation.

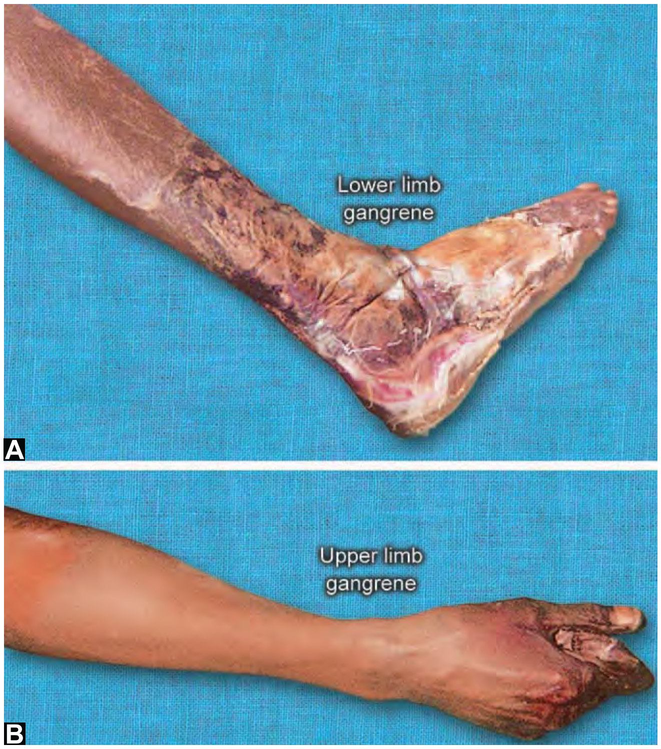

Figs 1.394A and B: Gangrene of lower and upper limbs in two different patients.

In *dry gangrene* the separation occurs by aseptic ulceration with minimum infection and the gangrene is dry and mummified**.**

In *wet gangrene,* separation takes place by septic ulceration. Often demarcation is vague with *skip lesions* more proximally and so landing with higher level of amputations. Even after amputation skin flap may show *die back* process, leading to failure of taking up of flap of amputation and so requiring still higher level of amputation.

- *Proximal ischaemic features* may be present with rest pain, colour changes, hyperaesthesia—pregangrene.

Types of gangrene

- *Dry gangrene* is due to slow, gradual loss of blood supply to the part causing dry, desiccated, wrinkled, mummified part with proper line of demarcation from the viable adjacent tissues
- *Wet gangrene* is due to infection with putrefaction, causing oedematous, swollen, discoloured part, spreading proximally, with vague line of demarcation from the adjacent viable tissues

Investigations

- Hb%, blood sugar.
- Arterial Doppler, angiogram (Seldinger technique).
- U/S abdomen to find out the status of aorta.

Treatment

***Limb saving methods*:**

- Drugs: Antibiotics, vasodilators, pentoxiphylline, praxilene, dipyridamole, small dose of aspirin, ticlopidine.
- Care of feet and toes:
  - The part has to be kept dry.
  - Any injury has to be avoided.
  - Proper footwear is advised (Microcellular rubber footwear, MCR).
  - Measures for pain relief is taken.
  - Nutrition supplementation is done.
  - The limb should not be warmed.
  - Pressure areas has to be protected.
  - Localised pus has to be drained.
- Cause is treated.
- Diabetes is controlled.

- Surgeries to improve the limb perfusion: Lumbar sympathectomy, omentoplasty.

Profundaplasty, femoropopliteal thrombectomy or endarterectomy, arterial graft bypass are done according to the need.

Life-saving procedures: *Amputations may have to be done occasionally.*

Level of amputation is decided on skin changes, temperature, line of demarcation, Doppler study.

- *Below-knee amputation* is a better option as *BK prosthesis* can be fitted better and also the movements of knee joint are retained. There is no need of external support and limp is absent.
- In *above-knee amputation* range of movements are less, limp is present, and often requires third (stick) support to walk.

Different amputations done are *Ray amputation*, below-knee amputation *(Buerger's amputation*), *Gritti-Stokes transgenial amputation*, above-knee amputation**.**

Lisfranc's, Chopart's, Symes', Modified Symes' amputations are not commonly used in ischaemic limb as flaps will not survive.

DIABETIC FOOT AND DIABETIC GANGRENE

Foot is a complex structure with many layers of muscles, ligaments, joints, arches, fat, thick plantar fascia, vascular arches, neurological system which maintains weight-bearing, gravity, normal walk, stability and gait (swing and stance phases).

Problems in diabetic foot

- Callosities, ulceration
- Abscess and cellulitis of foot
- Osteomyelitis of different bones of foot like metatarsals, cuneiforms, calcaneum
- Diabetic gangrene
- Arthritis of the joints

Meggitt's classification of diabetic foot

- Grade 0: Foot symptoms like pain, only
- Grade 1: Superficial ulcers
- Grade 2: Deep ulcers
- Grade 3: Ulcer with bone involvement
- Grade 4: Forefoot gangrene
- Grade 5: Full foot gangrene

Pathogenesis of Diabetic Foot/Gangrene

- High glucose level in tissues is a good culture media for bacteria. So infection is common.
- *Diabetic microangiopathy* causes blockade of microcirculation leading to hypoxia.
- *Diabetic neuropathy*: Due to sensory neuropathy, minor injuries are not noticed and so infection occurs. Due to motor neuropathy, dysfunction of muscles, arches of foot and joints occurs. And loss of reflexes of foot occurs causing more prone for trauma and abscess. Due to autonomic neuropathy, skin will be dry, causing defective skin barrier and so more prone for infection.

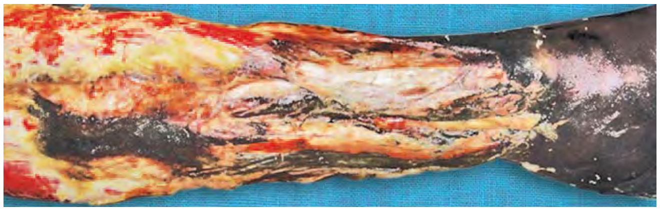

Fig. 1.395: Diabetic with extensive sepsis and gangrene. It requires amputation.

- *Diabetic atherosclerosis* itself reduces the blood supply and causes gangrene. Thrombosis can be precipitated by infection causing infective gangrene. Blockage occurs at plantar, tibial, and dorsalis pedis vessels.
- *Increased glycosylated haemoglobin* in blood causes defective oxygen dissociation leading to more hypoxia. At tissue level there will be increased glycosylated tissue proteins, which prevents proper oxygen utilisation and so aggravates hypoxia.

Clinical Features

- Pain in the foot.
- Ulceration.
- Absence of sensation.
- Absence of pulsations in the foot (Posterior tibial and dorsalis pedis arteries).
- Loss of joint movements.
- Abscess formation.
- Change in temperature and colour when gangrene sets in.
- Patient may succumb to *keto acidosis, septicaemia or myocardial infarction*.

Investigations

- Blood sugar, urine ketone bodies.
- Blood urea and serum creatinine.
- X-ray of part to look for osteomyelitis.
- Pus for culture and sensitivity.
- Doppler study of lower limb to assess arterial patency.
- Angiogram to look for proximal blockage.
- Ultrasound of abdomen to see the status of abdominal aorta.
- Glycosylated haemoglobin estimation.

Treatment

- ***Foot can be saved only if there is good blood supply.***
  - Antibiotics—decided by *pus C/S*.
  - Regular dressing.
  - *Drugs:* Vasodilators, pentoxiphylline, dipyridamole, low dose aspirin.
  - Diabetes is controlled by *insulin only*.
  - Diet control, control of obesity.
  - Surgical debridement of wound.

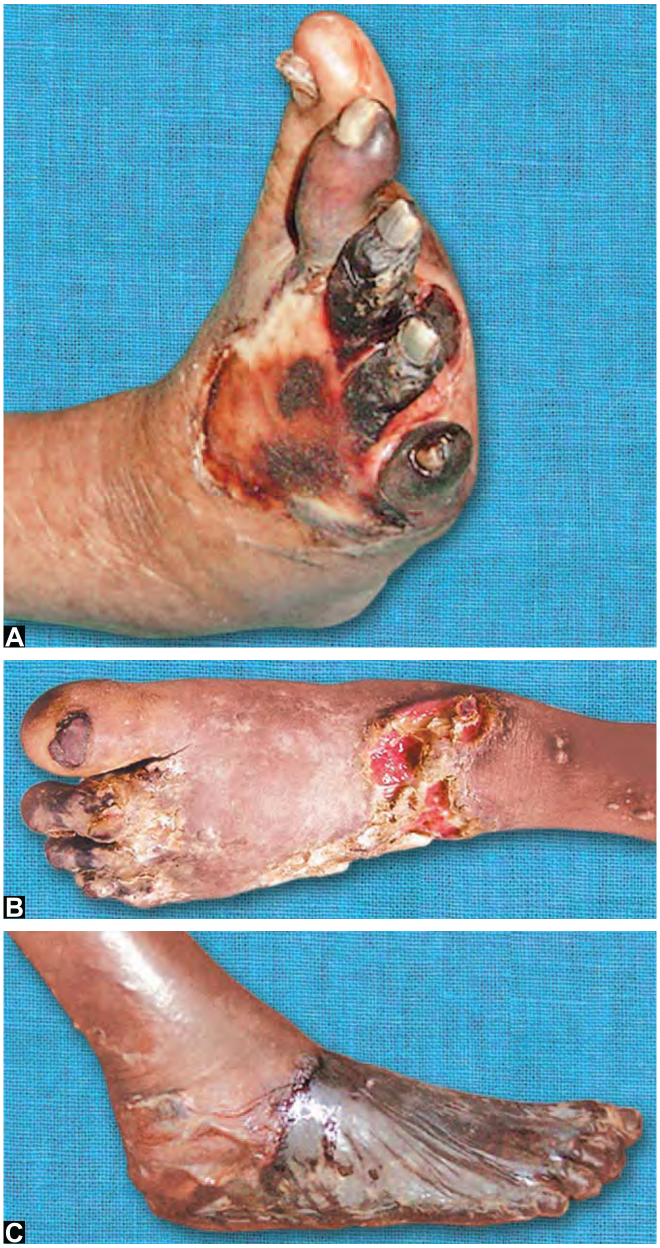

Figs 1.396A to C: Diabetic foot with ischaemic ulceration and gangrene. Note in one of the photos distal part of foot is gangrenous and proximal part is pregangrenous. This patient underwent above knee amputation.

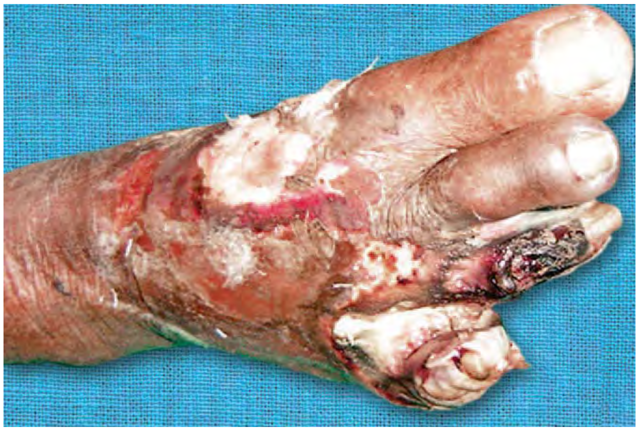

Fig. 1.397: Gangrene of 3rd and 5th toes in a diabetic. Patient already underwent amputation of 4 the toe earlier for gangrene.

- Amputations of the gangrenous area. Level of amputation has to be decided by skin changes and temperature changes or Doppler study.
- Care of feet in diabetic:
  - Any injury has to be avoided.
  - MCR footwears must be used (Microcellular rubber).
  - Feet has to be kept clean and dry, especially the toes and clefts.
  - Hyperkeratosis has to be avoided.

TROPHIC ULCER

Aetiology

Diabetic neuropathy, spinal injury and diseases, other neuropathies.

Causes

- ❖ Pressure
- ❖ Anaemia
- ❖ Malnutrition
- ❖ Injury
- ❖ Moisture

Common Sites

Heel, heads of metatarsals, sacrum, ischium, occipital region. *Bedsore is a trophic ulcer*.

Clinical Features

They are deep, punched out, nonmobile ulcers, with bone as its base.

Investigations

X-ray spine, nerve conduction studies, blood sugar, Hb%.

Treatment

- ♦ Nutrition is improved.
- ♦ Anaemia is treated.
- ♦ Diabetes is controlled.
- ♦ Regular dressing is done.
- ♦ Pressure and injuries has to be avoided.

BEDSORES (Decubitus Ulcer) (Pressure Sores)

Bedsore is a *trophic ulcer* with bone as the base.
It is nonmobile, deep, punched out ulcer.

It is common in		
Old age Bedridden	Tetanus	Patients with orthopaedic
Diabetic Paraplegic	Comatose	and head injuries

Sites of bedsore are occiput, heel, sacrum, ischium, scapula.

Factors: Malnutrition, pressure, anaemia, sensory loss, moisture.

Treatment

- Change of positions is always encouraged.
- Use of water bed, ripple bed is advised.
- Moisture has to be avoided.
- Soaking by urine, sweat, pus, and faeces has to be taken care off.
- Good nursing, regular dressing, good nutrition are necessary.
- Antibiotics, blood transfusions are very essential.
- Excision of dead tissue followed by skin grafting or local rotation flaps may have to be done.
- Rehabilitation.

FROSTBITE

- It is due to exposure to cold wind or high altitude.
- It is common in old age during cold spells.
- Damage to vessel wall occurs causing oedema, blistering, gangrene formation.
- Part is painless and waxy.
- Cells get frozen at – 5°C. Initially redness and oedema (1st degree); blister formation (2nd degree); skin necrosis (3rd degree); gangrene (4th degree develops gradually.

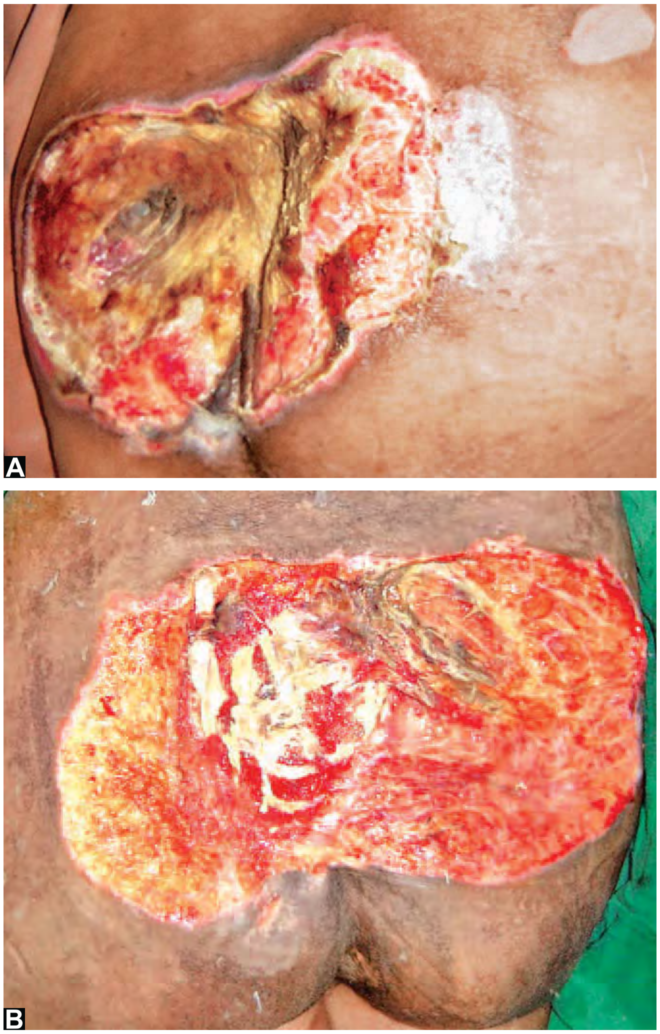

Figs 1.398A and B: Large bedsore over the sacrum in a abedridden patient. Note the size and slough on the surface.

Treatment

- *Gradual warming is done*. Part should be wrapped with cottonwool and rested. Warming is gradually done with 44°C in 30 minutes with warm water. Limb elevation is done to reduce oedema. Intra-arterial vasodilators may help.
- Warm drinks, analgesics, paravertebral injections to sympathetic chain, *hyperbaric oxygen* are effective.
- If gangrene develops, amputation is needed.

AINHUM

- Commonly affects males (can also occur in females).
- Common in blacks, in Negroes.
- History of running barefoot in childhood is common.
- *Fifth toe* is commonly affected.
- A fissure develops at the interphalangeal joint which becomes a fibrous band, that encircles the digit causing necrosis (Gangrene of little toe).
- Often it can be bilateral.

Treatment: It is early “Z” plasty. Amputation is often required later. Most often autoamputation occurs.

***Note*:**
Yoruba people of Nigeria named *ainhum*.

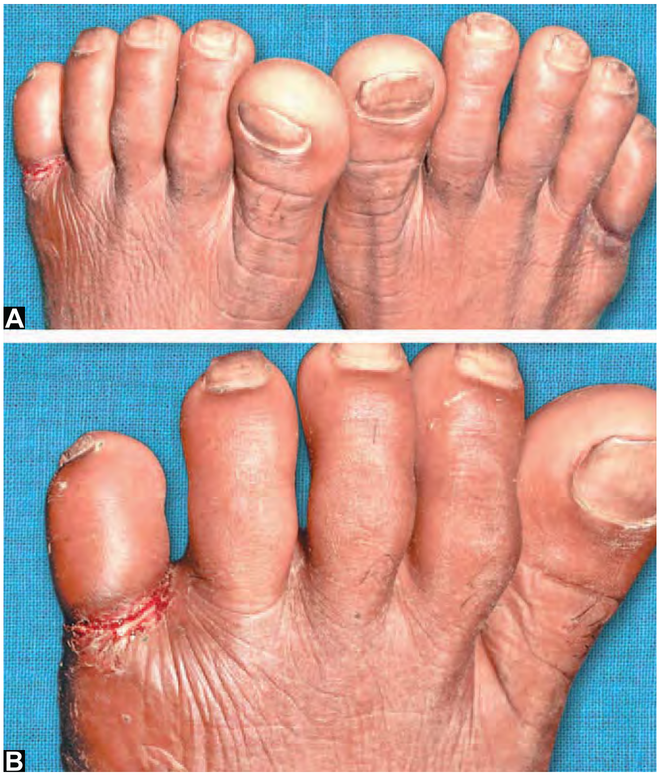

Figs 1.399A and B: Typical Ainhum. Note it is bilateral. Note the constriction ring in the little toe.

ENDOVASCULAR SURGERIES

It is mainly used in peripheral vessels like femoropopliteal, renal, coronary, cerebral vessels.

Types

- *Balloon angioplasty:* It is useful in short segment stenosis in large vessles like renal vessels, iliofemoral, coronary vessels. It is less effective compared to open surgery.
- *Intravascular stenting:* Balloon expandable and self-expanding stents are used at stenosed area.
- *Endovascular grafts* (PTFE, DACRON).
- Endovascular atherectomy.
- *Angioscopy:* Flexible, small, fibreoptic scopes to visualise vessel wall with sufficient irrigation to avoid opacification by blood.
- *Intravascular ultrasound:* To evaluate the vessel wall morphology.

Indications

- Aneurysm for stenting and grafting
- Aortoiliac constrictive disease
- Renal artery stenosis
- Carotid occlusive disease
- AV fistulas
- Management of pseudoaneurysm

Complications

- Thrombosis.
- Rupture.
- Sepsis.
- Fluid overload.
- Air embolism.

UPPER LIMB ISCHAEMIA

It is a rare uncommon entity compared to lower limb ischaemia but important because of its difficulty in managing. *Higher-level amputations are rare in upper limb ischaemia.*

Its incidence is rare (5%) due to abundant collateral supply, infrequency of atherosclerosis, decreased metabolic demand and smaller muscle mass.

It mostly affects distal small arteries (90%).

Symptoms are usually delayed.

Types of Upper Limb Ischaemia

- Acute.
- Chronic.

Acute type

Causes:

- **Embolism—common.** 30% of the peripheral emboli lodge in upper extremity. The most common site is at the bifurcation of brachial artery (40%); next common is at axillary artery (12%). Embolism is due to:
  - Cardiac origin (70%)—valvular lesions (atrial fibrillation, endocarditis), IHD, paradoxical.
  - Others—aneurysms, *thoracic outlet syndrome*, plaque.
- **Trauma—most common.** Brachial artery injury is seen in 30% of civilian trauma with arterial injuries, blunt injuries, fractures and dislocations, penetrating injuries.
- Iatrogenic.
- Post AV fistula 'Steal syndrome'.
- Aortic dissection.

Symptoms of acute ischaemia:

- Pain, pallor, poikilothermia, paraesthesia, paralysis.

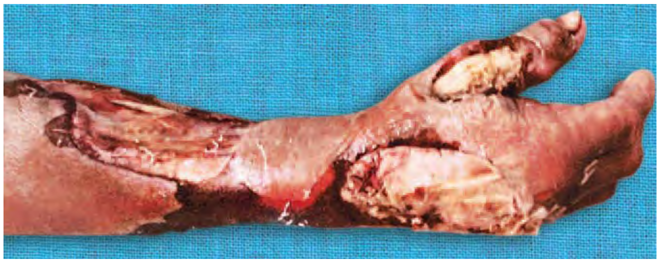

Fig. 1.400: Upper limb ischaemia with gangrene extending proximally towards elbow joint.

Chronic Type

Causes:

- Arteritis—aortoarteritis, Takayasu arteritis, giant cell arteritis, connective tissue disease/vasculitis—scleroderma, SLE, RA, PAN, etc.
- Atherosclerosis—most common cause in USA.
- TAO of upper limb.
- Others—fibromuscular dysplasia; postirradiation—lung, breast; occupational injuries; vibration injury, hypothenar hammer syndrome; hypercoagulable states; APLA, polycythemia, cold agglutinins.

Symptoms of chronic ischaemia:

- Upper limb 'Claudication'.
- Weakness and wasting.
- Digital ischaemia—ulcer, gangrene in finger tips.
- Raynaud's phenomenon.

Signs of chronic ischaemia:

- Wasting of arm, forearm and hand muscles.
- Ischaemic changes in skin; tapering of finger tips.
- Drop in systolic pressure >20 mmHg.
- Proximal thrill or bruit.
- Mass in the neck, thrill and bruit in the neck in supraclavicular region.
- Adson test, hyperabduction (Halsted) test, Roos test, Allen`s tests are important.

Raynaud's Phenomenon

- It is *episodic vasospasm*. It is common in upper limb.
- *Raynaud's disease*: It is *primary* Raynaud's phenomenon—no cause could be demonstrated (Idiopathic).
- *Raynaud's syndrome*: It is *secondary* Raynaud's phenomenon—secondary to a demonstrable lesion like SLE, scleroderma, TAO or atherosclerosis.
- *Symptoms*: Pain, discolouration, sensation of cold and numbness, pronounced on exposure to cold, and under stress.

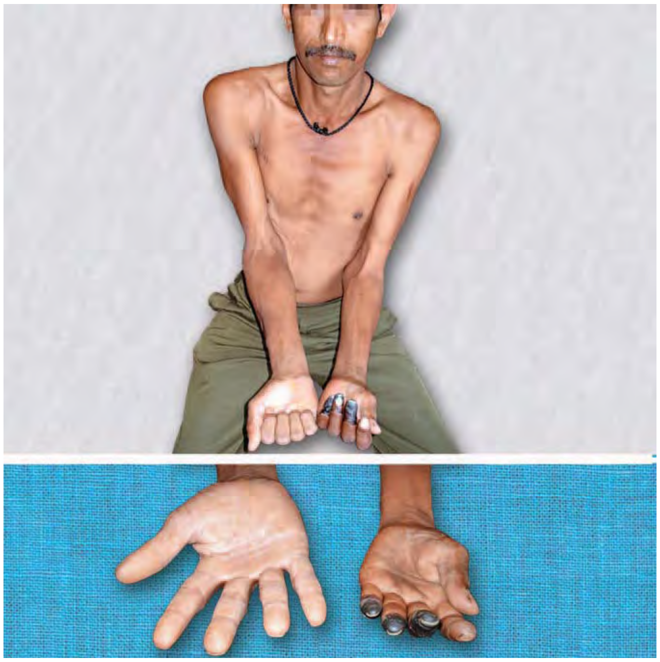

Fig. 1.401: Left upper limb ischaemia with gangrene of three fingers and ischaemic changes in hand, forearm and arm.

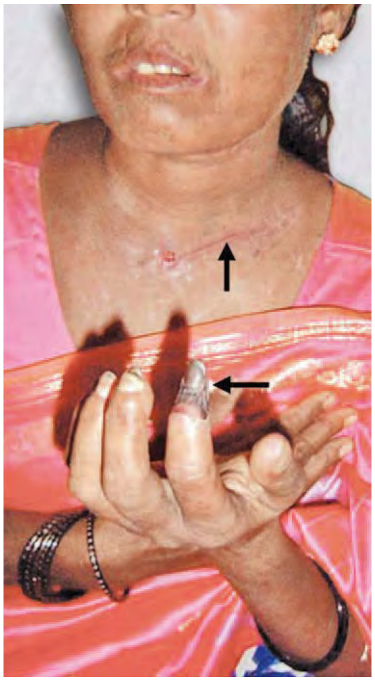

Fig. 1.402: Gangrene of left index finger. Patient has undergone cervical sympathectomy.

- *Signs*: Cyclical colour changes—*Pallor*, *Cyanosis* and *Rubor* with swelling; seen in fingers, toes, nose, ear lobes and lips.
- *Pathology of Raynaud's phenomenon*: Exaggerated vasomotor response to stress; more common in females; no structural changes in the vessels, except in late stages; recurrent attacks can lead to atrophy of skin, subcutaneous tissue and muscles, ischaemic ulcers and gangrene.

Investigations in Upper Limb Ischaemia

- Lab tests for vasculitis, hypercoagulable states, and atherosclerotic risk factors.
- X-rays—for cervical rib; clavicular and first rib fractures; fractures and dislocations in extremity; pulmonary lesions of connective tissue disorders.
- Arterial Doppler study.
- Angiogram (Subclavian angiogram)—CT/MR; conventional.
- CT scan neck and thorax.
- Blood sugar, lipid profile, cardiac evaluation.

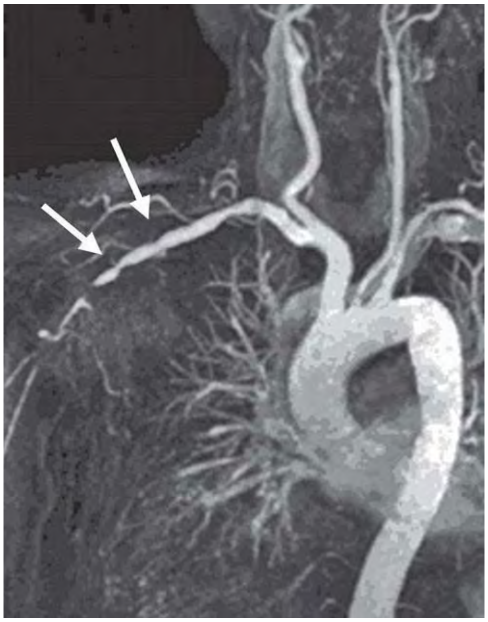

Fig. 1.403: Upper limb angiogram showing blocks in subclavian artery.

Management of Upper Limb Ischaemia

- Treatment of the cause.

Treatment of Embolus

- Time since the first symptom is very important.
- Clinical assessment of extent of ischaemia, immediate anticoagulation with heparin, Doppler study and angiogram of the arterial system, evaluation for the source of embolus—are the protocols.
- Embolectomy
  - *Brachial embolectomy*: Local/regional/general anaesthesia is used. Longitudinal incision in the arm is used for proximal embolus; Lazy S-shaped incision across the elbow is done for embolus extending into the bifurcation and to expose the branches.

Treatment in Trauma

- General evaluation and resuscitation.

- Control of bleeding in open wound: Pressure bandage/ manual compression (DO NOT USE TOURNIQUET).
- Time since the event and clinical assessment of limb perfusion.
- Stabilisation of fractures and dislocations
- Doppler study of arterial system, angiogram if required.
- Arterial repair; bypass graft either venous or synthetic.

Treatment of Chronic Ischaemia

- Medical management
  - Risk modification—diabetes, hypertension, dyslipidemia, smoking, homocystinaemia, exercise training.
  - Antiplatelets—aspirin/ticlopidine/clopidogrel.
  - Anticoagulants—heparin/warfarin.
  - Xanthines/pentoxiphylline/cilostazol.

  Cilostazol—suppresses cAMP phosphodiesterase III rise in cAMP levels with antiplatelet, antithrombotic effects; induces vasodilatation; increases plasma HDL cholesterol; decreases plasma triglycerides.
- Catheter based interventions
  - Atherectomy.
  - Angioplasty ± stenting by conventional or subintimal approach.
  - Stent grafts.
  - Cryoplasty.
- Surgery
  - Endarterectomy.
  - Bypass surgery.
  - Sympathectomy, extraperiosteal resection of the cervical rib.

Bypass Surgeries in Upper Limb Ischaemia

- Conventional bypass
  - Aorto-subclavian/axillary bypass.
  - Subclavian: Axillary/brachial bypass.
  - Brachiodistal bypass.
- Extra anatomical bypass
  - ***Carotid: Subclavian/Axillary bypass.***

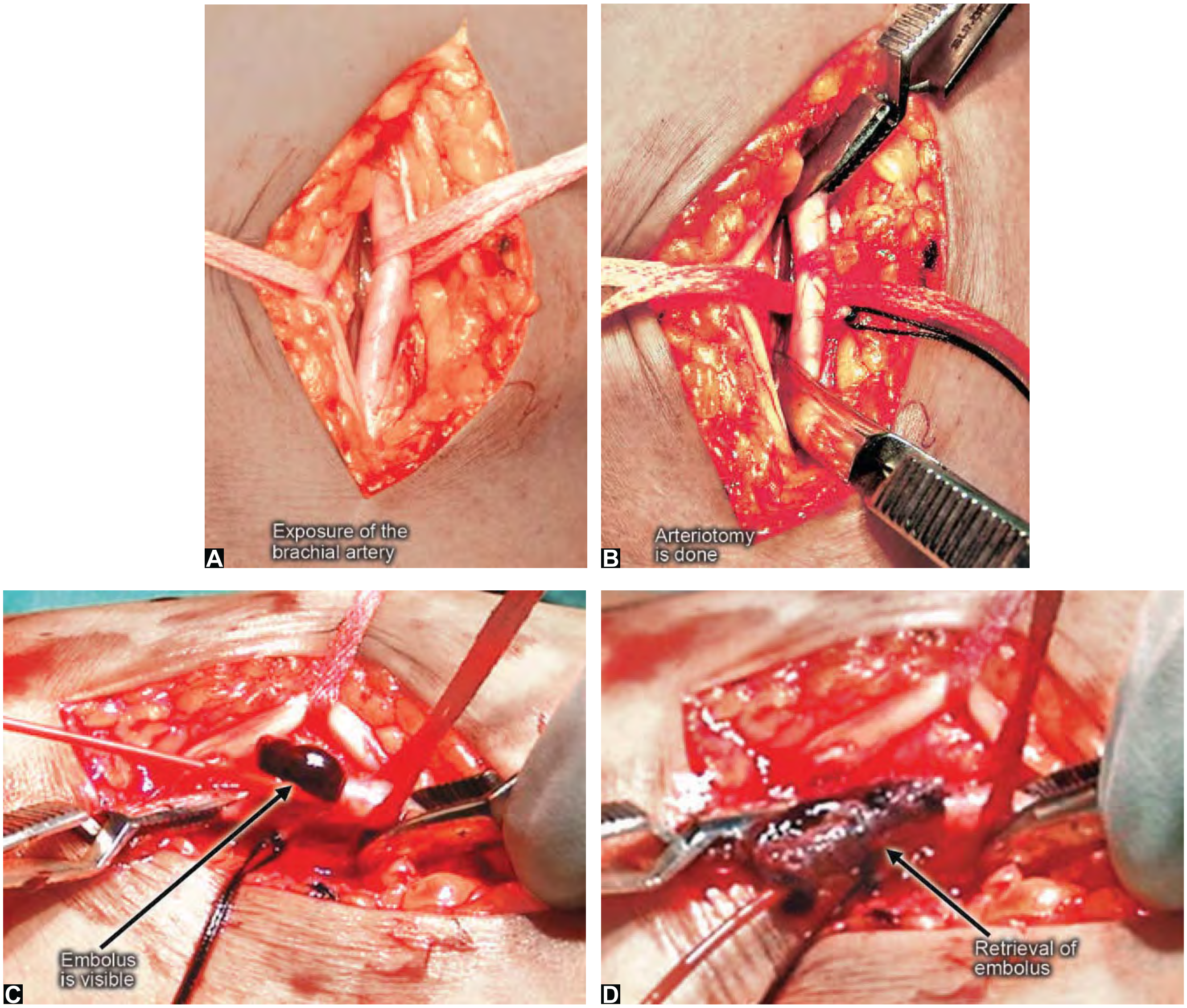

Figs 1.404A to D

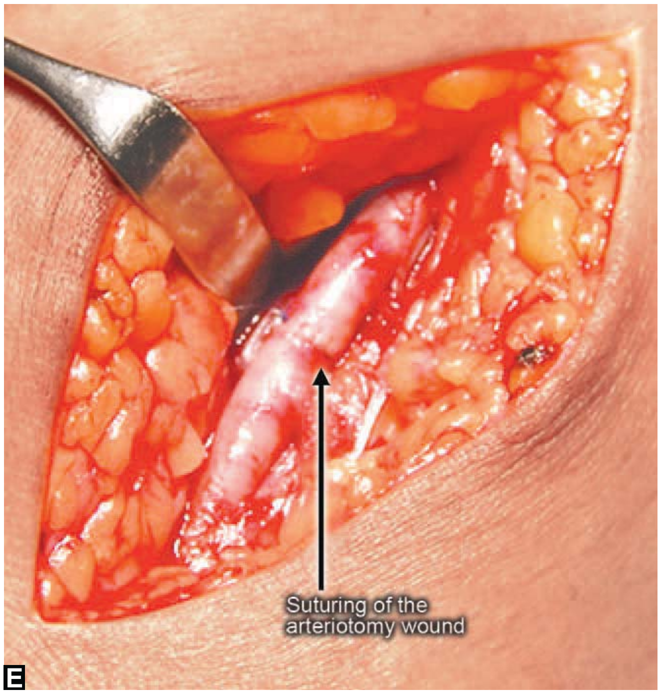

Fig. 1.404E

Figs 1.404A to E: Brachial embolectomy (upper limb)—open method.

  - Subclavian: Subclavian bypass.
  - Axillary: Axillary bypass.
- Subclavian Carotid transposition.

Treatment of Raynaud's Phenomenon

- Avoiding triggering agents
- Drugs (vasodilators)—calcium channel blockers, angiotensin II receptor blockers, alpha-1 adrenergic blockers, Sildenafil, prostaglandin E1.
- Surgery—sympathectomy.

Note:
Individual topics about causes of upper limb ischaemia are discussed in different places.

Upper limb ischaemia

- Trauma/cervical rib are the common causes
- Opposite limb, lower limbs should be examined
- Cardiovascular system should be examined
- Neck should be examined
- Wasting/girth should be checked
- All relevant clinical methods are equally significant
- Auscultation over neck/axilla/carotids for bruit are important
- Doppler; angiogram; nerve conduction studies; CT neck and thorax are essential investigations
- Arterial repair; therapy for cervical rib; scalenotomy; cervical sympathectomy are the different modalities of treatment
- Digital amputation may be required

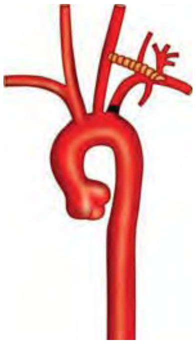

Fig. 1.405: Carotid subclavian bypass.

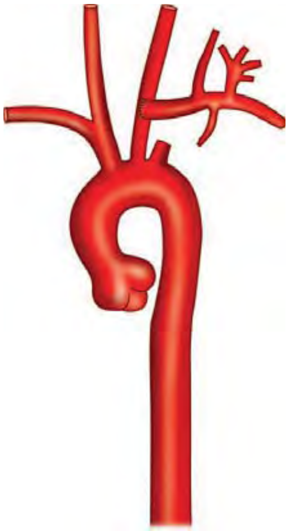

Fig. 1.406: Subclavian carotid transposition.

You my say you KNOW, but by your actions it is KNOWN.

ARTERIAL SUBSTITUTES

Ideal arterial substitute is not yet developed. Ideal arterial substitute should be strong, durable for patient's life, biocompatible, nonthrombogenic, should be resistant for infection, easily available, should have a long-term patency rate, and should have elastic property of normal artery.

Features of ideal graft	
Strong	Should be leak proof on restoration of blood flow
Durable for patient's life	Should not chemically or physically degenerate
Non thrombogenic	Should not cause any abnormal reaction to surrounding tissues
Biocompatible	Should not occlude when flexed
Resistant to infection	Should not damage blood contents
Flexible	
Should maintain long-term patency	
Should have elastic property of normal artery	

Carrel and Guthrie first did venous autograft into arteries of dogs. They did extensive histological study of viable and nonviable grafts. Lexer in 1907 used saphenous vein for axillary artery repair. Murray started to use intraoperative heparin. Enaz Moniz and dos Santos originated technique of arteriography. Gross and his colleagues in 1948 started to use viable arterial allografts. Later it was found that, as of graft is considered tissue viability is not essential for success of graft uptake.

Classification of Arterial Substitutes

- Arterial allograft—not used.
- Arterial autograft—internal mammary artery (common), internal iliac artery.
- Arterial xenograft—bovine carotid artery graft—not used.
- Venous autograft—long saphenous vein (common), small saphenous vein, basilic vein, cephalic vein.
- Venous allograft—umbilical vein graft.
- Prosthetic grafts
  - Textile grafts
    - Dacron graft—knitted or woven or crimping or velour types.
    - Teflon graft—knitted or woven crimping or velour types.
  - Non textile semi-inert polymer graft:
    - ePTFE graft—expanded polytetrafluoroethylene graft.

Preclotting the ***noncoated knitted*** or woven Dacron graft is done to seal the graft and to prevent leak, and to create a smooth lining at graft—blood interface. This step is not necessary for PTFE or newer grafts.

Complications of Graft

- *Neointimal fibrous hyperplasia* at suture lines of the graft is due to surgical trauma, PDFG, arterial smooth muscle proliferation.
- *Graft infection*—incidence is 2%. It is more in lower limb graft than abdominal graft. Peroperative cephalosporin administration reduces the rate of graft infection. If infection occurs graft should be removed and revascularisation is achieved using a saphenous vein graft.
- *Graft failure* is rare but can occur. It is due to fiber degeneration, manufacturing defect, diffuse dilatation of graft (is due to expansion of the knit in knitted Dacron).
- *Anastomotic false aneurysm* (3%) occurs just adjacent to suture line towards host artery. It is tearing of the artery adjacent to suture line due to mismatched graft artery compliance, improper suture placement, and arterial degeneration. There will be partial or total separation of the graft from the host with blood collection in a covering of fibrous capsule. Eventually it will rupture/may cause thrombosis and embolism. Treatment is graft—artery reanastomosis with insertion of additional piece of graft.

M. Vascular Lesions and Hamartoma

CHAPTER OUTLINE

- Vascular Anomalies
- Haemangioma
- Vascular Malformations
- Cirsoid Aneurysm
- Arteriovenous Fistula
- Campbell De Morgan Spots
- Parry-Romberg Disease
- Hamartomata

VASCULAR ANOMALIES

- It is *a collective term* used for haemangioma and vascular malformations.
- *Haemangioma* is a benign tumour containing hyperplastic endothelium with cellular proliferation with increased mast cells. Growth in tissue culture is observed. It is absent at birth, seen by 1 month in 30%. It usually shows biphasic growth phase with slow involution. 95% of cases achieve spontaneous involution. Fast growing type can cause platelet trapping and thrombocytopenia. Associated skeletal changes are not common but can occur. But bone erosion by the lesion can occur. It is common in girls (3:1).
- *Vascular malformations* are single layer endothelium lined spaces derived from arterial, capillary, venous or lymphatic system. There is no growth in tissue culture. Raise in mast cells is not seen. 90% cases are seen at birth; only few at later period. It is equal in both sexes (1:1). Quiescent endothelium with vessels showing progressive ectasia is the feature. Intravascular coagulation and mild thrombocytopenia can develop. Skeletal changes and overgrowth are common. Spontaneous involution is not common.
- Disfigurement, tissue destruction, deformity, dysfunction, telangiectasia, skin scarring are common.
- *Szilagyi classification*—(1) Cavernous haemangioma; (2) Microfistulous AV communications; (3) Macrofistulous communications; (4) Anomalous mature vascular channels.
- *Humburg classification*—(1) Predominantly arterial/venous/lymphatic defects with aplasia or obstructive dilatation which is limited or infiltrative; (2) Predominantly AV shunting defects with deep/superficial limited or infiltrating lesions; (3) combined vascular defects—arterial, venous and haemolymphatic which may be limited or infiltrating.
- *Diagnosis* is made clinically and by radiological imaging—coloured Doppler, DSA, MRI. *MRI* is *better* than CT to identify the flow (MR angiogram is ideal). Haemangiomas show intense parenchymal staining; vascular malformations show ecstatic vessels without much parenchyma; AVM shows rapid venous shunting.

HAEMANGIOMA

- It is the most common tumour in children (in 10% of term deliveries).
- It shows cellular endothelial hyperplasia with increased mast cells.
- Onset is few weeks after birth with biphasic growth showing initial rapid growth with gradual involution over 5-7 years.
- It is benign vascular endothelial tumour, common in girls (3 : 1).
- It is commonly seen in skin and subcutaneous tissue but can occur anywhere in the body like in liver, brain, lungs or other organs.

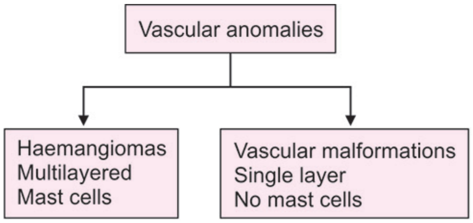
Vascular anomalies
Haemangiomas
Multilayered
Mast cells
Vascular malformations
Single layer
No mast cells

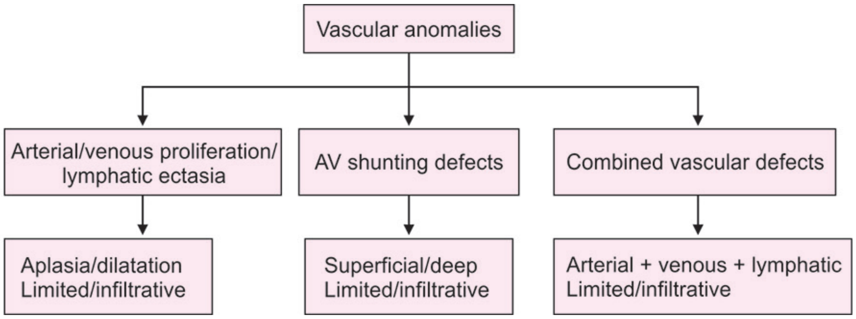
Vascular anomalies
Arterial/venous proliferation/
lymphatic ectasia
AV shunting defects
Combined vascular defects
Aplasia/dilatation
Limited/infiltrative
Superficial/deep
Limited/infiltrative
Arterial + venous + lymphatic
Limited/infiltrative

You may be disappointed if you fail, you are doomed if you don't try.

- It grows rapidly in first year and 70% involutes in 7 years.
- Early proliferative lesion is bright red, irregular; deep lesion is bluish coloured. Involution causes colour fading, softness, shrinkage leaving crepe paper like area.
- Commonly it is central; common in head and neck region (60%).
- Often large haemangiomas may be associated with visceral anomalies. Head and neck haemangioma is associated with ocular and intracranial anomalies; sacral with spinal dysraphism. Multiple cutaneous haemangiomas may be associated with haemangioma of liver causing hepatomegaly, cardiac failure (CCF), anaemia.
- Ulceration, bleeding, airway block and visual disturbances are common complications.
- A definitive even though rare, but important life-threatening complication is platelet trapping and severe thrombocytopenia presenting as ecchymosis, petechiae, intracranial haemorrhage massive GI bleed.
- Raised angiogenic (fibroblastic) growth factor which is secreted in patient's urine is useful lab investigation to differentiate it from vascular malformations.

Classification

- *Capillary*
  - Salmon patch (stork bite)
  - Strawberry haemangioma
  - Port-wine stain (naevus flammeus)
- *Cavernous*

Capillary Haemangioma

- ***Salmon patch (stork bite):***
  - It presents at birth. It commonly occurs in nape of the neck (50%), face, scalp and limbs. It usually involves wide area of skin. It is caused by an area of persistent fetal dermal circulation. With age, it goes for spontaneous regression and disappears completely (usually in one year). Hence masterly inactivity is the treatment.
- ***Strawberry haemangioma:***
  - It may start at birth or child is normal at birth; between one to three weeks it appears as red mark which rapidly increases in size in 3 months to form strawberry/raspberry haemangioma. It contains immature vasoformative tissues. There will be eventually intravascular thrombosis, fibrosis and mast cell infiltration.
  - It is a *true capillary* haemangioma.
  - It is 20 times more common than port wine stain.
  - It is common in *white girls*. Male to female ratio is 1 : 3.
  - It is common in head and neck region.
  - It is clinically *compressible,* warm with bluish surface.
  - Bleeding can occur after minor trauma and also ulceration.
  - It involves skin, subcutaneous tissues and often muscles also.
  - After one year of age, it slowly begins to disappear, and *completely* in 7-8 years (70% in 7 years).
  - It is the most common haemangioma.

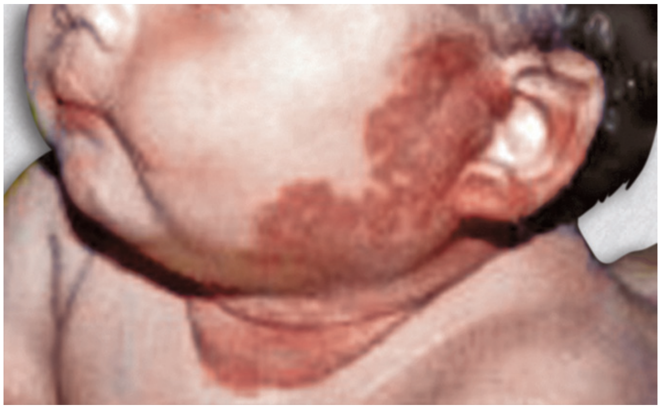

Fig. 1.407: Salmon patch.

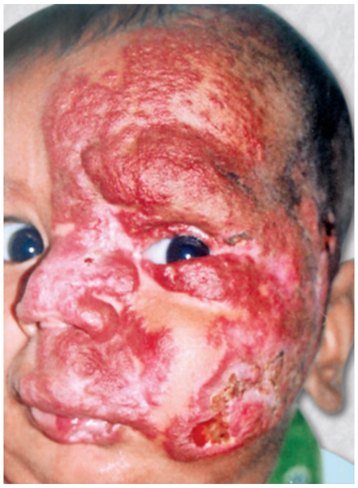

Fig. 1.408: Strawberry haemangioma.

- Haemangioma in periorbital region obstructs the vision in newborn with amblyopia and if it persists for 7 days causes permanent visual damage. Astigmatism also can occur
- Haemangioma in nasal area in newborn may obstruct nasal airway seriously (as newborn cannot breathe through mouth—obligatory nasal breathing)
- Skin ulceration may cause haemorrhage
- Infection can occur which may lead into sepsis, necrosis or rarely septicaemia
- Systemic steroid for 3 weeks induces involution
- Usually there is no role for surgery. Surgery is done only for retained tissue after involution

- ***Port-wine stain (Naevus flammeus)***:
  - It presents at birth and persists throughout life without any change. Spontaneous regression will not occur. It presents as smooth, flat, reddish blue/intensely purple area; common in head, neck and face; often with maxillary and mandibular dermatomes of 5th cranial nerve. Eventually surface becomes nodular and keratotic.
  - It persists throughout life.
  - It is actually a *capillary malformation* even though considered under haemangioma. It results from defect

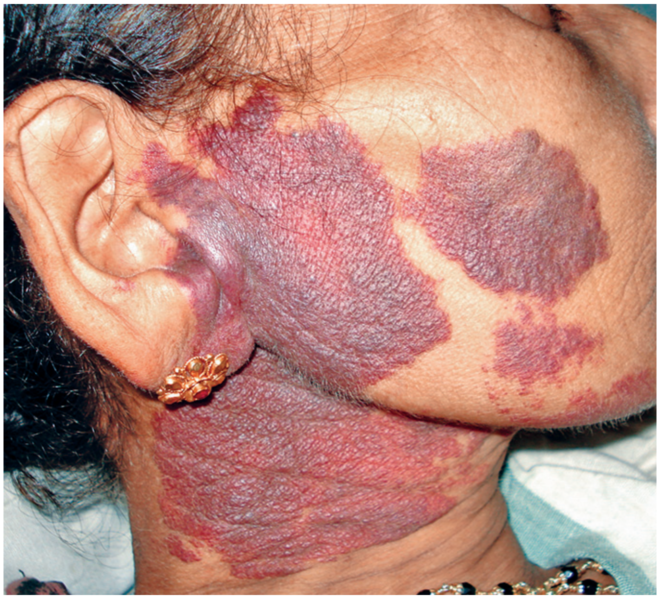

Fig. 1.409: Port-wine stain (Naevus flammeus).

in maturation of sympathetic innervation of skin causing localised vasodilatation of intradermal capillaries.

- It is often associated with Sturge-Weber syndrome, Klippel-Trenaunay-Weber syndrome and Proteus syndrome.
- It needs treatment—laser (pulsed dye/diode); excision and grafting; cosmetic coverage. Expected result by treatment is not possible many times.

Indications for surgery or intervention

- Uncontrolled growth
- Functional impairment like vision or hearing
- Accidental haemorrhage

Treatment

- They are treated by *wait and watch policy* commonly—allowed for *spontaneous regression*.
- Pulsed dye laser (diode laser), surgical excision and reconstruction. *Feeding vessels* may need to be ligated after wide exposure before achieving complete extirpation. Sclerotherapy/cryotherapy/CO_2 snow therapy cause *unpleasant scarring*.
- *Preoperative embolisation* facilitates surgical excision and reduces the operative blood loss. When once embolisation done, surgery should be done *as early as possible* otherwise recurrence occurs and much more worried formation of enlarged collaterals can occur.
- Materials used are—foam, plastic spheres, stainless/platinum steel coils, ethanol, polyvinyl alcohol foam of 1000 μ meters size, and rapidly polymerizing acrylic. Problems are tissue necrosis, reaction, normal tissue embolisation. Procedure is done with interventional radiology under C-Arm guidance.
- *Rapidly growing* haemangioma may need systemic/oral and intralesional *steroid* therapy.
- Antiangiogenic *interferon 2a* may be useful.
- Life-threatening *platelet trapping* may be controlled by *cyclophosphamide* chemotherapy.
- Haemangioma with *drug resistant CCF* can be treated with *radiotherapy*.

Associated syndromes

- ***Klippel-Trenaunay-Weber syndrome:*** Naevus flammeus + osteohypertrophy of extremities (soft tissue and bone hypertrophy) + varicose veins of lower limbs. If there is an association of arteriovenous fistula (AV fistula), it is called as ***Parkes-Weber syndrome***
- ***Kasabach Merritt syndrome:*** Capillary haemangioma + DIC (Disseminated intravascular coagulation) with thrombocytopenia
- ***Sturge-Weber syndrome:*** Haemangiomas (Naevus flammeus) + hemiplegia and Jacksonian epilepsy (calcified vascular cerebral and meningeal deposits) + glaucoma
- ***Maffucci syndrome:*** Cavernous haemangioma + dyschondroplasia
- ***Proteus syndrome:*** Naevus flammeus + regional gigantism with lymphaticovenous malformation (asymmetrical hypertrophy)
- ***Osler-Rendu-Weber syndrome:*** Haemangioma of skin and lip with gastrointestinal tract haemangioma (hereditary haemorrhagic telangiectasia), (autosomal dominant)

Cavernous Haemangioma

- It is present at birth and consists of a multiple venous channels.
- Its size increases gradually and may cause problems.
- It often contains feeding vessels which is of surgical importance.
- *Sites*: Head, neck, face, limbs, tongue, liver and other internal organs.
- Large or multiple cavernous haemangiomas can cause *congestive heart failure* (hyperdynamic) due to shunting of large quantity of blood.
- Cavernous haemangioma with dyschondroplasia is called as Maffucci syndrome.

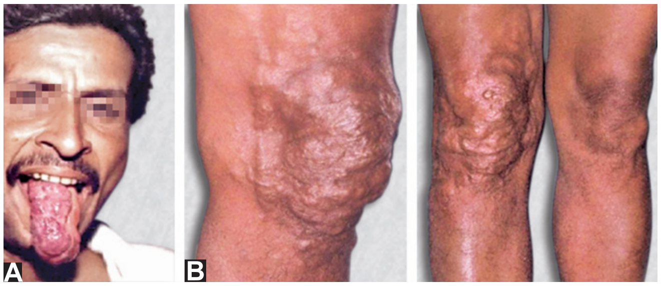

Figs 1.410A and B: Cavernous haemangioma in (A) tongue and (B) knee.

When I was young, I observed that nine out of ten things I did were failures. So I did ten times more work.
—George Bernard Shaw

- Cavernous haemangioma is often mixed with lymphatic component also (mixed vascular and lymphatic).

Clinical Features

- It is smooth, soft, well-localised, warm, fluctuant, compressible, nonpulsatile swelling with bluish surface occurring in skin and subcutaneous tissue (often in mucosa like oral cavity) without any transillumination.
- *Compressibility and bluish surface* is diagnostic. When swelling is pressed it reduces partially/often completely but when pressure is released it slowly attains its original size and shape. Vascular and lymphatic malformations are compressible.
- It is usually nontender unless it gets infected or undergoes thrombosis or causes haemorrhage.

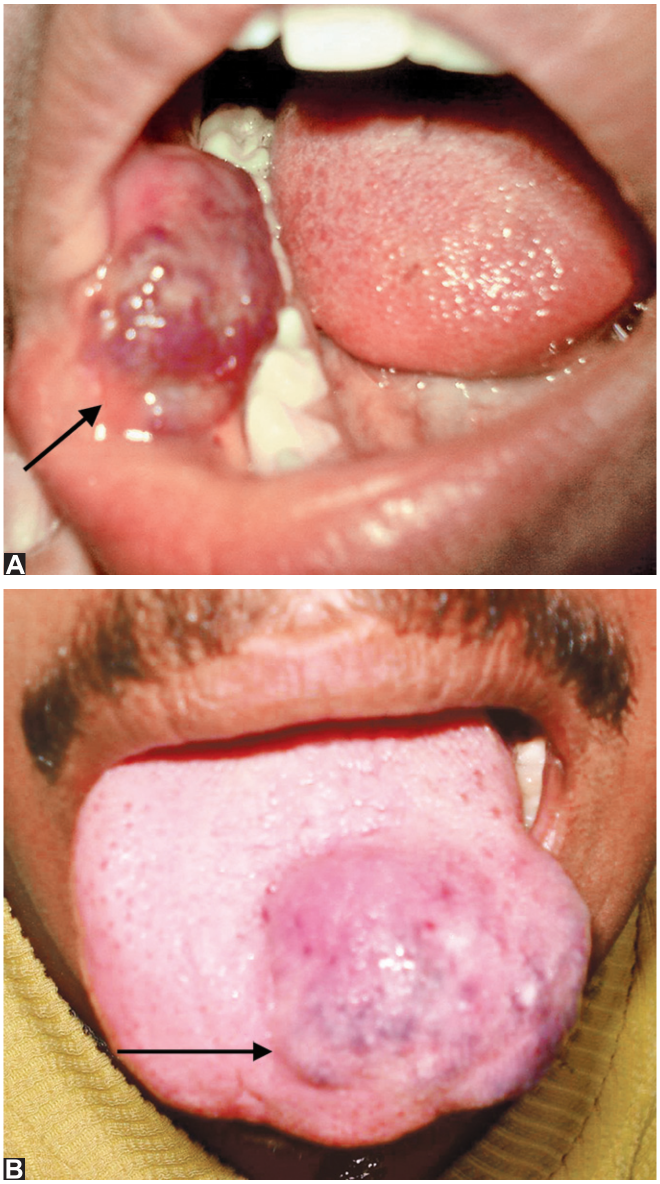

Figs 1.411A and B: Cavernous haemangioma in the cheek near angle of the mouth and in the tip of the tongue. Haemangioma tongue is one of the causes of macroglossia.

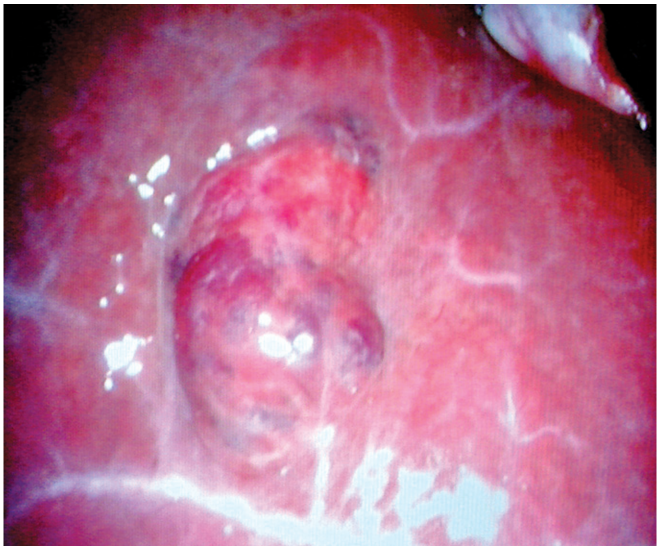

Fig. 1.412: Laparoscopic view of cavernous haemangioma of liver. It is the most common benign tumour of the liver.

Differential Diagnosis

- Lymphangioma: It is brilliantly transilluminant unless it is infected or fibrosed.
- Lipoma, cold abscess, lymph cyst—clinically it is easier to differentiate.

Complications

- Haemorrhage.
- DIC.
- Thrombosis.
- Infection, ulceration and septicaemia.
- Erosion into the adjacent bone.
- High output cardiac failure.

Investigations

- Ultrasound, Doppler.
- Angiogram to find out feeding vessel.
- Platelet count.
- MRI/MR angiogram to see feeding vessels and deeper extension.

Treatment

- Sclerosant therapy: It is the initial first line of therapy. It causes aseptic thrombosis and fibrosis of the cavernous haemangioma with less vascularity and smaller size. It is directly injected into the lesion. Sodium tetradecyl sulphate/hypertonic saline are used. Often multiple injections are needed to achieve complete required effect. Later excision of the lesion is done.
- Ligation of feeding artery and often at later stage excision is done once haemangioma shrinks.
- Therapeutic embolisation.
- If small and located in accessible area, excision is the initial therapy.

- Laser ablation—diode pulsed laser is becoming popular because of good control of bleeding. CO_2/Nd:YAG laser is also equally effective.

VASCULAR MALFORMATIONS

- Secondary to defect in development of vascular components, in 8th week of intrauterine period.
- Single layer endothelium lined spaces derived from arterial, capillary, venous or lymphatic system showing ectasia.
- There is no growth in tissue culture. Raise in mast cells is not seen.
- Associated with many syndromes.
- Can be located in skin or in deeper planes.
- Present at birth and grows in proportion to child's growth. Pale skin which later darkens over the age or faint blue mass is the presentation.
- Spontaneous involution is not common.
- Capillary malformation (CM) type is due to lack of sympathetic control.
- Venous malformation (VM) type is *most common* vascular malformation which shows hypoplasia, hyperplasia or aplasia of superficial or deeper system. It is seen in subcutaneous plane as faint blue compressible mass with morning pain and stiffness of the area.
- Lymphatic malformation (LM) type can be microcystic (lymphangioma) or macrocystic (cystic hygroma). It can cause lymphoedema, soft tissue and bony hypertrophy, asymmetry (face), macrochelia, macroglossia, macrotia, cellulitis.
- *Low/slow flow* malformations can cause skeletal hypoplasia; *high/fast flow* malformations can cause hypertrophy. AVM is high flow type.
- Consumption coagulopathy (DIC) can occur.
- It is equal in both sexes (1:1).
- Doppler is commonly used investigation; but MRI (*MR angiogram is ideal*) with contrast is ideal to identify and to differentiate low and high flow types.
- *Treatment*—conservative with compression garments and sclerotherapy. Laser photocoagulation is the choice for superficial malformations; multiple sittings may be needed; complete clearance may not be achieved. Surgical excision can be done. Preoperative embolisation may be needed.

Vin rose patch:

It is a congenital intradermal pale pink vascular malformation with dilatation of vessels in subpapillary dermal plexus. It may be associated with haemangiomas; AV malformations in limbs; congenital lymphoedema.

CIRSOID ANEURYSM

- It is a rare variant of capillary haemangioma occurring in skin, beneath which abnormal artery communicates with the distended veins.
- Commonly seen in **superficial temporal artery** and its branches.
- Often the underlying bone gets thinned out due to pressure.
- Occasionally extends into the cranial cavity.
- Ulceration is the eventual problem which will lead to uncontrollable haemorrhage.

Clinical Features

Pulsatile swelling in relation to superficial temporal artery, which is warm, compressible, with arterialisation of adjacent veins and with *bone thinning* (due to erosion). It feels like a *'pulsating bag of worms'*.

Investigations

- Doppler study, CT scan.
- Angiogram, X-ray of the part.

Treatment

- Ligation of feeding artery and excision of lesion, often requires preliminary ligation of external carotid artery.
- Intracranial extension requires formal neurosurgical approach.

ARTERIOVENOUS FISTULA (AVF)

It is an abnormal communication between an artery and vein.

Types

- Congenital—is arteriovenous malformation.
- Acquired (Trauma is common cause).

Congenital Arteriovenous Malformation (Fistula)—AVM (AVF)

During developmental period AV communications occur.

It is high flow type of vascular malformation. It is 30% of all vascular malformation. 90% of AVM contains both arterial and venous components. Shuting of blood with thrill and bruit with hyperdynamic circulation is common.

Sites

- Limbs, either part or whole of the limb is involved. It may be localised to toes or fingers.
- Lungs.
- Brain—in circle of Willis.
- Other organs like bowel, liver.

Clinical Features

Structural changes in the limb:

- Limb is **lengthened** due to increase in blood flow since developmental period.
- Limb **girth** is also increased.
- Limb is **warm.**

> *When stethoscope is applied aneurysm is either silent or a systolic bruit can be heard, but an arteriovenous fistula emits a continuous murmer throughout the systole and diastole.*
> **—Charle G Rob**

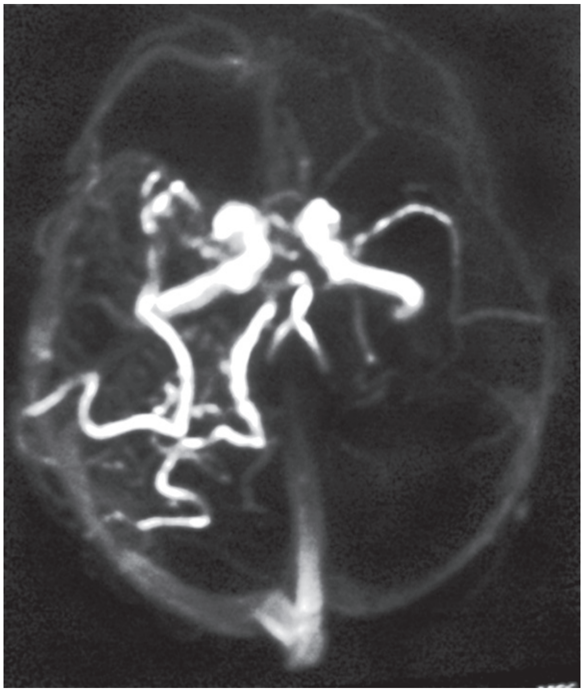

Fig. 1.413: MR angiogram of AV malformation in cranial cavity.

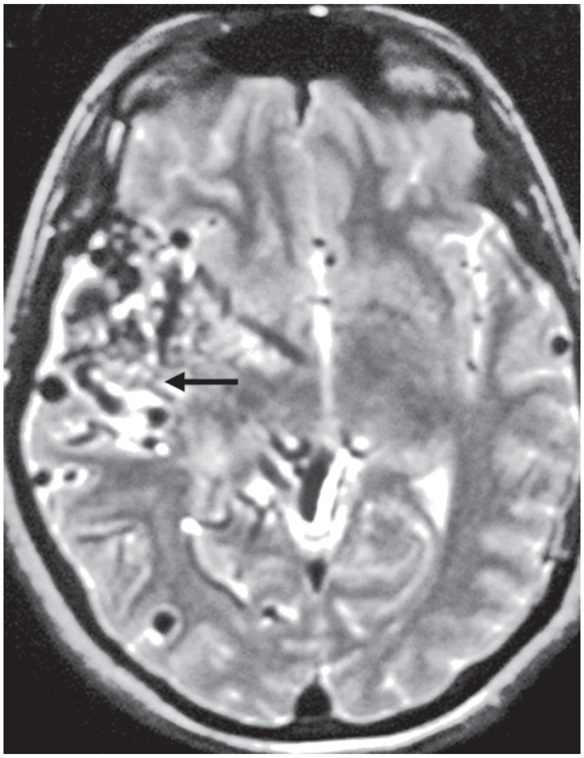

Fig. 1.414: MRI picture of AV malformation.

- **Continuous thrill and continuous machinery murmur** all over the lesion.
- **Dilated arterialised varicose veins** are seen due to increased blood flow and also due to valvular incompetence.
- Often there is **bone erosion** or extension of AVF into the bone as such.

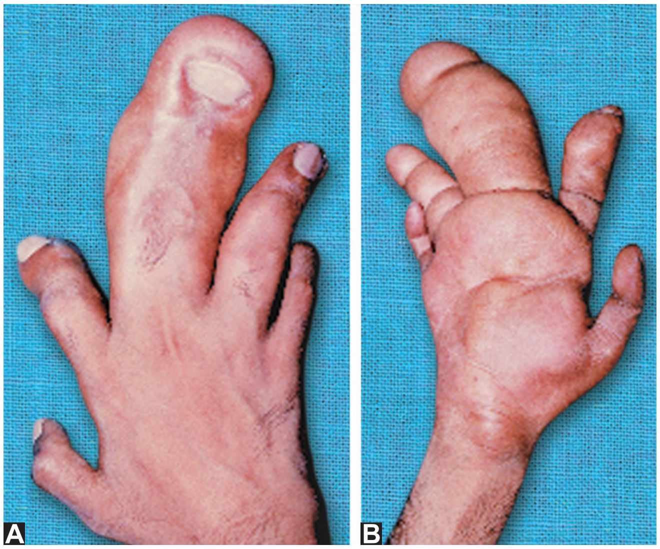

Figs 1.415A and B: Hypertrophic changes due to AV malformation.

Physiological changes

Because of the hyperdynamic circulation, there is **increased cardiac output** and so often **congestive cardiac failure**.

Complications

- Haemorrhage
- Thrombosis
- Cardiac failure (CCF)

Investigations

- Angiogram—MR angiogram is ideal.
- Doppler study.
- X-ray of the part.
- ECG, echocardiography.

Treatment

- *Conservative*—sclerotherapy, compression, avoiding injury.
- **Indications for intervention**
  - *Absolute*: Haemorrhage, ischaemia, CCF.
  - *Relative*: Pain, functional disability, cosmesis, limb asymmetry.
  - *Emergency*: Torrential bleeding usually after trauma (example—road traffic accidents).

Interventions

- Surgical ligation of *feeding vessels* and complete excision of the lesion. Often if lesion is extending into deeper planes it is technically difficult; but with *usage of tourniquet*, careful meticulous dissection and ligation of all vessels will lead into successful excision of entire lesion.
- Therapeutic *embolisation*/preoperative embolisation hasten the proper surgical excision.
- In *emergency bleeding*, adequate transfusion of blood , tourniquet usage, intraoperative embolisation and then excision of entire lesion is done. Occasionally when extensive AVM is present often involving the entire limb, amputation is the final option left as a life-saving procedure.

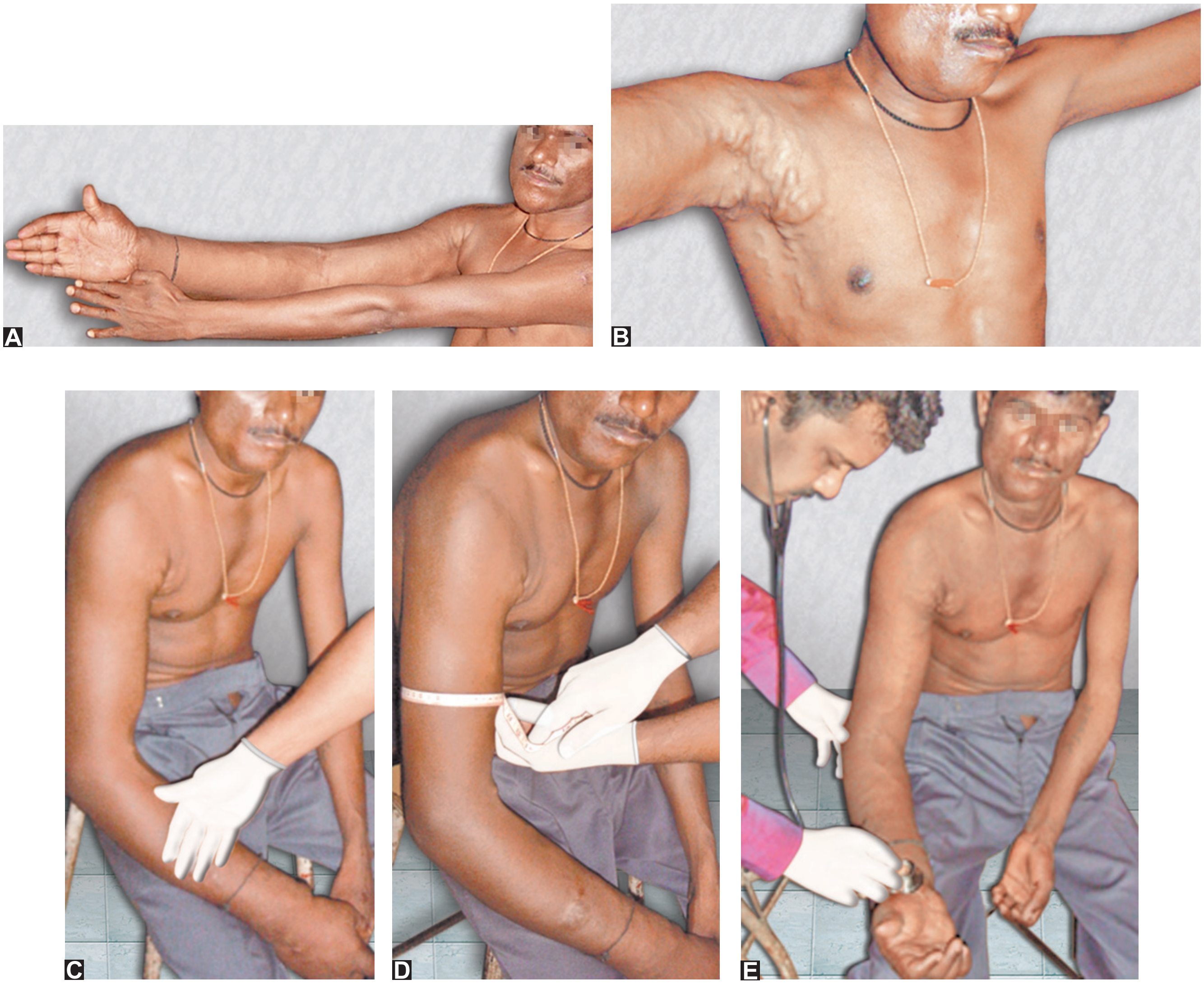

Figs 1.416A to E: Congenital arteriovenous malformation involving right upper limb, axillary region and part of right half of neck and face. Note the limb lengthening and widened girth. Limb is warmer. Limb should be measured at each level. Auscultation reveals continuous bruit over the limb. Note the dilated tortuous veins due to hyperdynamic circulation.

Acquired Arteriovenous Fistula (AVF)

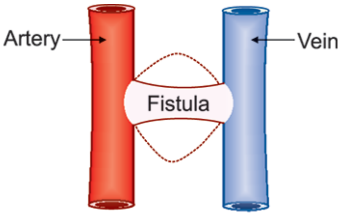

Fig. 1.417: AV fistula.

Causes

- Trauma in (most common cause):
  - Femoral region.
  - Popliteal region.
  - Brachial region.
  - Wrist.
  - Aorta—vena caval.
  - Abdomen.

  It may be following road traffic accidents, penetrating wounds, cock-fight injury (common in South India).
- After surgical intervention of major vessels.
- Therapeutic: For renal dialysis, AVF is created (**Cimino fistula**) to achieve arterialisation of veins and also to have hyperdynamic circulation. It is done to have easy and adequate venous access for long time **haemodialysis.** Common sites are **wrist, brachial, and femoral** region.

Pathophysiology

- *Physiological changes: Cardiac failure due to hyperdynamic circulation***.**
- *Structural changes:*

Changes at the Level of Fistula

Blood flows from high pressure artery to low pressure vein causing diversion of most of the blood. Between the artery and vein, at the site of fistula, dilatation develops with formation of fibrous sac called as **aneurysmal sac.** This presents as warm, pulsatile, smooth, soft, compressible swelling at the site with **continuous thrill and continuous machinery murmur**.

Changes Below the Level of the Fistula

Because of diversion of arterial blood **distal part becomes ischaemic**. Because of high pressure arterialisation of veins and valvular incompetence occurs causing **varicose veins**.

Changes Proximal to the Fistula

Hyperdynamic circulation causes cardiac failure.

If pressure is applied to the artery proximal to the fistula, swelling will reduce in size, thrill and bruit will disappear, pulse rate and pulse pressure becomes normal. This is called as **Nicoladoni's sign** or **Branham's sign.**

Cardiac failure may be very severe in traumatic AVF (often resistant to drug therapy).

Investigations

- Doppler, angiogram.
- ECG, echocardiography.

Treatment

- **Excision** of fistula and **reconstruction** of artery and vein with graft.

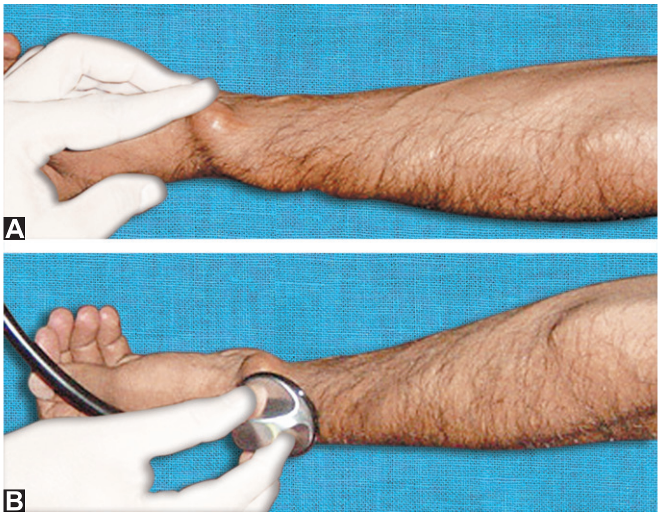

Figs 1.418A and B: Acquired arteriovenous fistula in the wrist over radial vessels. It should be palpated for compressibility and thrill. It should be auscultated for bruit.

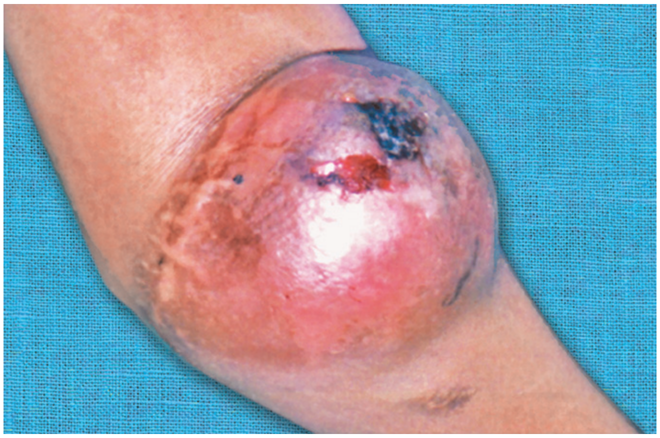

Fig. 1.419: AV fistula created for treating chronic renal failure has formed an aneurysm. It may rupture to cause severe haemorrhage. Thrombosis or sepsis also can occur in this.

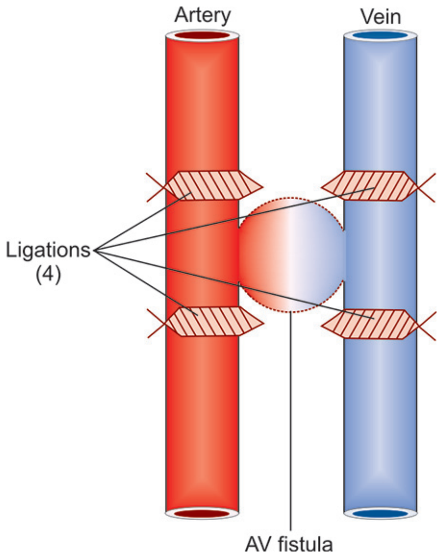

Fig. 1.420: Quadruple ligation of AV fistula.

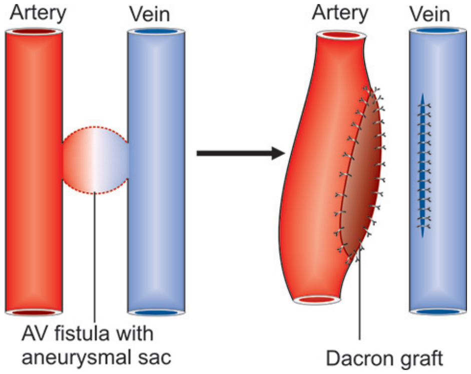

Fig. 1.421: Reconstruction of AV fistula using graft.

Done in early stages—larger vessels. Venous or Dacron graft is used.

- In emergency situation, **quadruple ligation**, i.e. both artery and vein above and below are ligated without

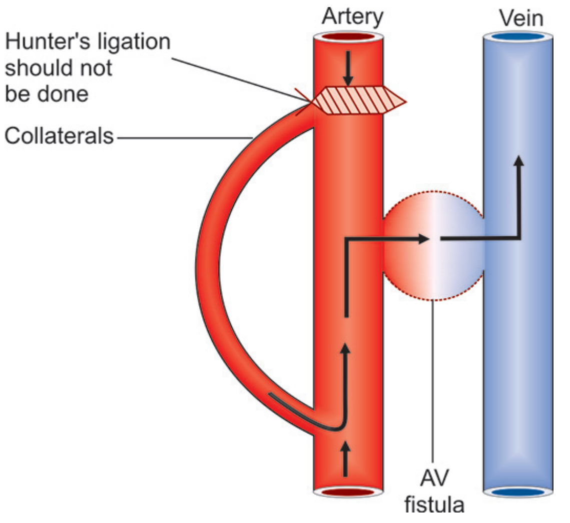

Fig. 1.422: Hunter's ligation is ligation of artery proximal to AV fistula. It should not be done as it causes diversion of all blood from periphery leading to gangrene of distal part.

touching the fistula and sac. Patient recovers well from cardiac failure.

- Therapeutic embolisation may be tried.

Hunter's ligation should be avoided. It is used as life-saving measure because it invariably causes limb ischaemia and gangrene even though patient recovers from cardiac failure. It is ligation of both artery and vein proximally so as to make cardiac function normal. But it invariably steals the blood from the limb leading to gangrene.

CAMPBELL DE MORGAN SPOTS

- It is usually smaller (0.2 to 6 mm) in size, circular, elevated and bright red swelling.
- It is common in trunk. Common in elderly.
- Also called as *cherry angiomas.*
- Usually it does not require treatment.
- When needed, excision or electrodesiccation or laser removal is done.

PARRY-ROMBERG DISEASE

- It is hemifacial atrophy of skin, soft tissue and bone.
- Common in females.
- It usually begins at twenties.
- Atrophy of skin, fat, muscle, cartilage and bone causing *coupe de sabre* deformity.
- It is a self-limiting disease.
- Aesthetic reconstruction is offered when severe deformity develops.

HAMARTOMATA

Hamartano means *'I miss' (Greek).* Or 'fault' or 'misfire' or 'error'—(missing the mark in spear throwing).

Note:
Presently this terminology is not very much in use.

- It is a benign lesion with aberrant differentiation producing a mass of disorganised but mature specialised cells or tissue indigenous to the particular site.
- It is tumour like overgrowth of tissue or tissues proper to that part.
- It may be single lesion or multiple lesions.
- Haemangiomas, lymphangiomas, AV malformations, neural malformations are the examples.

Problems with hamartomas

- Pressure symptoms locally
- Bleeding
- Infection
- Gigantism
- Cosmetic problem

Treatment

- Depends on site, type, extent.
- Cryotherapy, ligation of feeding vessels, sclerotherapy, excision or laser therapy.

N. Venous Diseases

CHAPTER OUTLINE

- Anatomy of Veins of Lower Limb
- Physiology of Venous Blood Flow in Lower Limb
- Deep Vein Thrombosis
- Varicose Veins
- Venous Ulcer
- Compression Therapy for Varicose Veins
- Thrombophlebitis
- Klippel-Trenauny Syndrome
- Anticoagulants
- Heparin
- Oral Anticoagulants
- Direct Thrombin Inhibitors
- Antiplatelet Drugs
- Pulmonary Embolism

ANATOMY OF VEINS OF LOWER LIMB

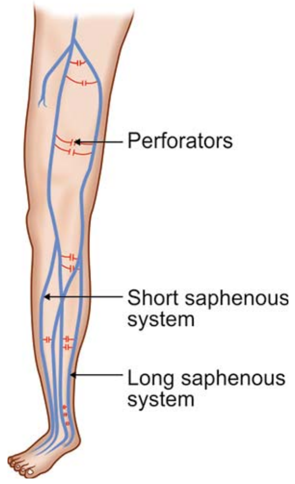
Perforators
Short saphenous system
Long saphenous system

Fig. 1.423: Anatomy of lower limb veins.

Deep Veins

- Tibial, popliteal, femoral veins are called as "*veins of conduits"* which drain blood into iliac veins and then to IVC.
- *Pumping veins*: They are venous sinuses existing in the calf muscles which pump blood towards major veins. They are better termed as *musculovenous pumps*. They are also called as the *peripheral heart*.

Superficial Veins

- *Long/great saphenous vein*: It is a subcutaneous vein over the inner aspect of the leg and thigh, joins into femoral vein at fossa ovalis. Tributaries of long saphenous vein are posterior arch vein, anterior vein of leg, anterolateral vein, posteromedial vein and sometimes accessory saphenous vein.
- *Short/small saphenous vein*: It is over the lateral and posterior aspect of the leg enters the deep fascia in the upper calf region and later joins popliteal vein at variable distance.
- Posterior arch vein of '*Leonardo*' (from medial ankle to the long sephanous vien below the knee).
- Anterior arch vein to peroneal veins.

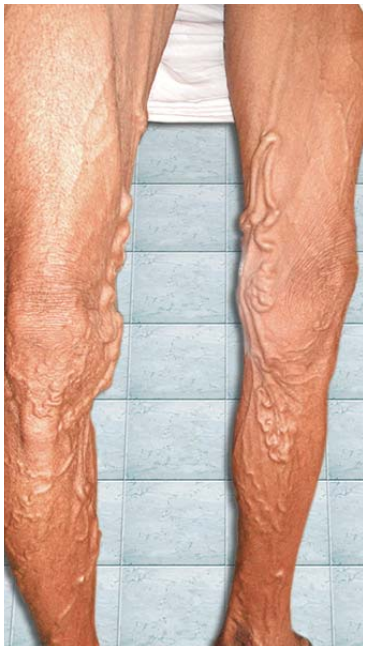

Fig. 1.424: Bilateral varicosity of great saphenous veins.

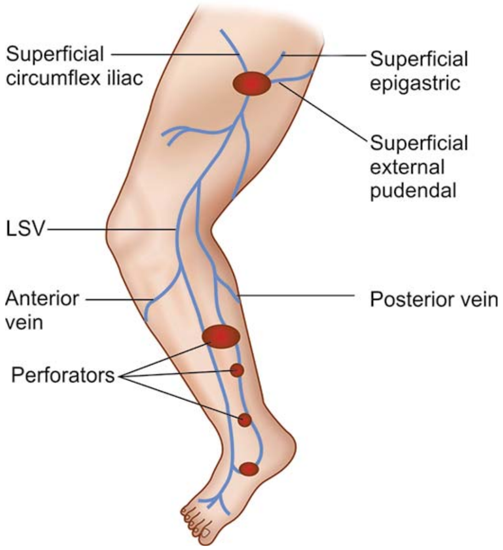
Superficial circumflex iliac
Superficial epigastric
Superficial external pudendal
LSV
Anterior vein
Posterior vein
Perforators

Fig. 1.425: Long saphenous vein anatomy and its tributaries.

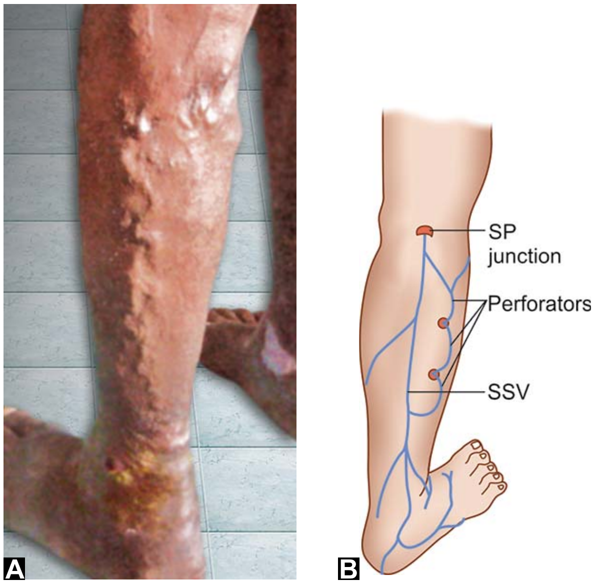

Figs 1.426A and B: Short saphenous vein varicosity.

Superficial veins have got multiple valves which facilitates blood flow towards heart. Superficial veins usually drain about 10% of lower limb blood, i.e. from skin and subcutaneous tissues.

Perforator Veins

They are the veins which connect superficial to deep veins at various levels. They travel from superficial fascia through an opening in the deep fascia before entering the deep veins. The direction of blood flow here is from superficial to deep veins. These perforators are also guarded by valves so that the blood flow is unidirectional, i.e. towards deep veins. Reversal of flow occurs due to incompetence of perforators which will lead to *varicose veins*.

Types

- Ankle perforators (May or Kuster)
- Lower leg perforators: I, II, III (of Cockett)
- Gastrocnemius perforators (of Boyd)—below knee
- Mid-thigh perforators (Dodd)
- Hunter's perforator in the thigh

There are no valves for IVC and iliac veins. Long saphenous vein has got 10-14 valves; whereas short saphenous vein has usually one valve. Veins can accommodate large quantity of blood and hence called as *capacitance* vessels.

PHYSIOLOGY OF VENOUS BLOOD FLOW IN LOWER LIMB

Veins are thin-walled vessels with collapsible walls, that assume an elliptical configuration in collapsed state and circular configuration in the filled state.

Venous valves are abundant in the distal lower extremity and the number of valves decreases proximally, with no valves in superior and inferior vena cava.

Venous Return

- Arterial pressure across the capillary increases the pumping action of vein.
- *Calf musculovenous pump*: During contraction phase of walking, pressure in the calf muscles increases to 200-300 mmHg. This pumps the blood towards the heart. During relaxation phase of walking, pressure in the calf falls and so it allows blood to flow from superficial to deep veins through perforators. Normally while walking, pressure in the superficial system at the level of ankle is 20 mmHg.
- During walking, foot pump mechanism propels blood from plantar veins into the leg.
- Gravity.

***Note*:**
Pressure in *arteriolar end of the capillary* is 32 mmHg; *venular end of capillary* is 12 mmHg.

Factors responsible for venous return

- Negative pressure in thorax
- Peripheral pump—calf muscles
- Vis-a-tergo of adjoining muscles
- Nonrefluxing valves in course of veins

DEEP VEIN THROMBOSIS (DVT)

It is called as *phlebothrombosis*. It is *semisolid clot* in the vein which has got high tendency to develop pulmonary embolism and sudden death. Common site of beginning of thrombus is soleal veins which later propagate proximally, often getting detached to cause acute massive pulmonary embolism or moderate sized emboli can cause pyramidal/wedge shaped pulmonary infarcts.

Aetiology: Factors

Virchow's triad

- Stasis
- Hypercoagulability
- Vein wall injury

Causes

- Following childbirth.
- Trauma—to leg, ankle, thigh, pelvis.
- Muscular violence.
- Immobility: Bed ridden patients, individuals on long duration air or bus travel ***(Traveller's thrombosis*).**
- Debilitating illness, obesity, immobility, bed rest, pregnancy, puerperium, oral contraceptives, estrogens.
- Postoperative thrombosis *(Most common cause)*: Common after the age of 40 years. Incidence following surgeries is 30%. In 30% of cases both legs are affected. Usually seen after prostate surgery, hip surgery, major abdominal surgeries, gynaecological surgeries, cancer surgeries. Bedridden for more than 3 days in the postoperative period increases the risk of DVT.

You may be disappointed if you fail, you are doomed if you don't try.

- *Spontaneous thrombosis* is common in visceral neoplasm like carcinoma pancreas or carcinoma stomach. It is often migrating type.
- Thrombus may start in a venous tributary which eventually may extend into the main vein causing DVT.
- ***Axillary vein thrombosis***
  - It can occur spontaneously, following compression by cervical rib, by various causes of thoracic inlet syndrome, or arm being in the hyperabduction state for prolonged period (e.g. painting the ceiling), after axillary lymph node block dissection, after radiotherapy to axilla, occasionally as a complication of venous cannulation.
  - Upper limb DVT is rare compared to lower limb DVT (5% of all DVT). It may be axillary or subclavian vein or both.
  - But 30% of upper limb DVT can cause pulmonary embolism.
  - Primary upper limb DVT is *Paget-Schroetter syndrome,* is due to subclavian vein compression that occurs in thoracic outlet syndrome. It may be precipitated by exertion of arms, swimming, exercise, etc.
  - Idiopathic upper limb DVT is rare. Occult underlying malignancy should be thought of.
  - Secondary upper limb DVT is due to CVP line, pacemaker thrombocytosis, malignancy, surgeries, radiotherapy, etc.
  - Unilateral arm, forearm swelling with bluish discolouration, pain, pitting oedema, often with skin blebs are the features.
  - Investigations are—Duplex scan, MR venography (as clavicle obscures proper duplex evaluation), BT, CT, PT, APTT, platelet count estimation.
  - *Treatment* is similar, with heparin/LMWH/warfarin, thrombolysis using tissue plasminogen activator, elevation of the arm, using compression stockings.
- Polycythaemia vera, thrombocytosis.
- Deficiencies of antithrombin III, protein C, protein S, factor V of Leiden, thrombophilia.
- Recent myocardial infarction, heart failure, nephrotic syndrome.
- Thrombosis can occur in individuals who sit with computers for long time—'***ethrombosis'.***

Sites

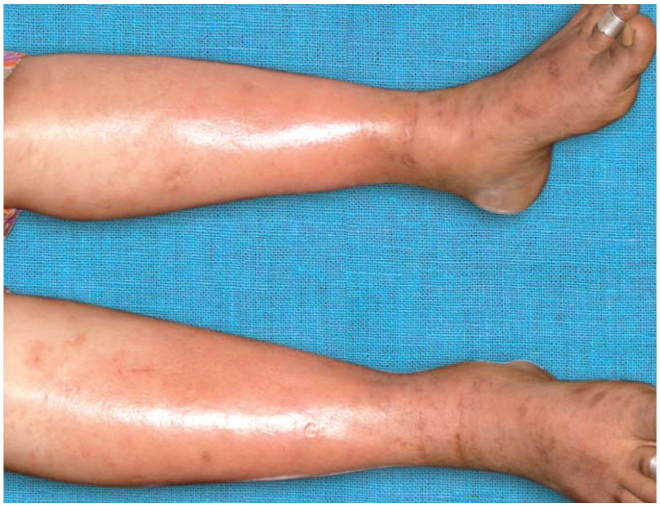

Fig. 1.427: Deep vein thrombosis (DVT) in both legs. 30% cases of DVT are bilateral.

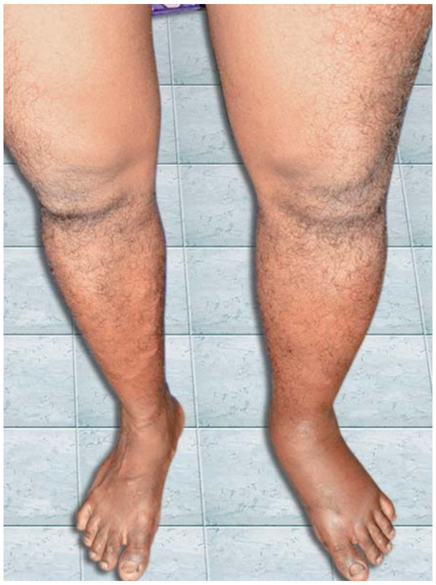

Fig. 1.428: Unilateral left leg DVT. Note the oedema and colour changes.

a. Pelvic veins—not uncommon; involves internal iliac veins. It is more common in PID (pelvic inflammatory disease) in females. In males prostatic veins may be the site of origin. It is difficult to identify clinically even though rectal or vaginal examination may help.
b. Leg veins—very common in veins of soleus muscle in calf; femoral vein/iliofemoral vein thrombosis can occur along with calf veins or independently without calf vein thrombosis which shows adductor canal tenderness. Iliofemoral vein thrombosis is common on left side due to its lengthy course/compression by right iliac artery/often due to presence of web at its entry into IVC. Incidence of bilateral leg DVT is 30% which should be differentiated from bilateral pedal oedema due to other causes like hypoproteinaemia, renal failure and cardiac causes.
c. Upper limb veins—not uncommon (Axillary vein thrombosis).

Phlegmasia Alba Dolens

It is DVT of *femoral vein* (deep femoral vein commonly) causing painful congestion and oedema of leg, with lymphangitis, which further increases the oedema and worsens the situation (*white leg*).

Phlegmasia Caerulea Dolens

It is extensive DVT of *iliac and pelvic veins* causing *blue leg* with either venous gangrene or areas of infarction.

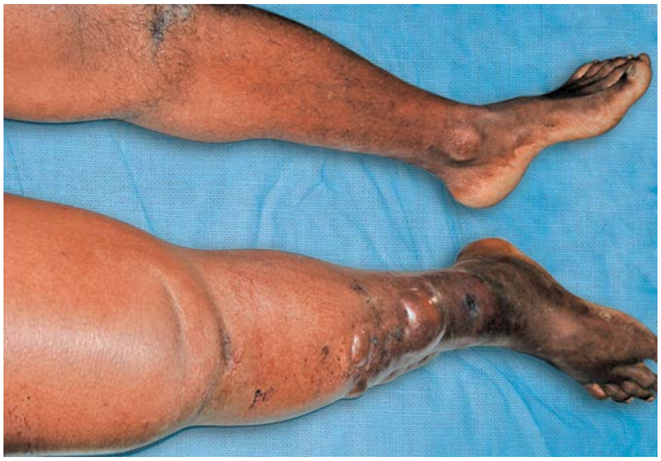

Fig. 1.429: Right leg venous gangrene. Note the discolouration, blebs and oedema.

Features of DVT

- Commonly it is asymptomatic—60%
- Fever—most common
- Tense, tender, warm, pale/bluish, shiny swelling calf
- Positive Homan's, Mose's or Neuhof's signs
- Inverted champagne bottle sign
- Features of pulmonary embolism

Clinical Features

- Fever—earliest symptom.
- Pain and swelling in the calf and thigh (often). Pain is often so severe that the patient finds it difficult to flex (or move) the leg.
- Leg is tense, tender, warm, pale or bluish with stretched and shiny skin.
- *Positive Homan's sign*: Passive forceful dorsiflexion of the foot with extended knee will cause tenderness in the calf.
- *Mose's sign*: Gentle squeezing of lower part of the calf from side-to-side is painful. Gentleness is very important otherwise it may dislodge a thrombus to form an embolus.
- *Neuhof's sign*: Thickening and deep tenderness elicited while palpating deep in calf muscles.
- Most often, DVT is *asymptomatic* and presents suddenly with features of *pulmonary embolism* like *chest pain, breathlessness* and *haemoptysis.*
- After applying tourniquet at saphenofemoral junction, patient is made to walk and without removing the tourniquet, limb is elevated—persisting prominent superficial veins will be observed in DVT—*Linton's test.*

Differential diagnosis for DVT

- Ruptured Baker's cyst
- Ruptured plantaris tendon
- Calf muscle haematoma
- Cellulitis leg
- Superficial thrombophlebitis

Investigations

- Venous Doppler.
- ***Duplex scanning:*** It shows noncompressible vein which is wider than normal. On compression over calf muscles, it does not show any augmentation of flow. Normal venous sound at the area of femoral vein which disappears during inspiration is conspicuously absent in DVT.
- ***Venogram:*** Contrast material is injected into venous system to get detailed idea of the veins after applying tourniquet into superficial system. Occlusive and nonocclusive thrombus can be differentiated by this. But as it is invasive one, it is not commonly done at present. MR venogram is under trial at present. Impedance plethysmography is used to measure the rate of venous emptying. Vein occlusion is done using cuff around upper thigh which is confirmed by flat electrical wave pattern. When cuff is released rapid flow of wave is observed in normal; sluggish flow of wave is seen in DVT.
- ***Radioactive I^{125} fibrinogen study:*** Sodium iodide 100 mg orally is given to the patient 24 hours before the test to block the thyroid activity. I^{125} labelled fibrinogen 100 μ curies is injected intravenously. First radioactivity of heart is measured by placing the scintillation counter over the precordium. Reading obtained by this is adjusted as 100%. After that legs are elevated using adjustable stands and to prevent venous pooling, scintillation counter is placed over the calf. Counting in the leg is done from below upwards at 5 cm intervals. Procedure is done in preoperative period; on 1st, 3rd and 6th postoperative days. A 20% or more raise in percentage value suggests deep vein thrombosis in leg. I^{125} labelled fibrinogen is used (earlier I^{131} labelled fibrinogen was used) because it has got softer radioaction; its detectability is with much lighter and mobile apparatus.
- Haemogram with platelet count; *D-dimer test*/analysis of fibrin degradation products (FDP) are relevant tests used. D-dimer test is measurement of cross-linked degradation products which interprets the plasmin activity on fibrin. Negative D-dimer test is of more value.
- Ventilation—perfusion scanning with mismatched defects; pulmonary artery CT scan with filling defect; pulmonary angiography are the investigations to confirm the pulmonary embolism.

Treatment

- Rest, elevation of limb, bandaging the entire limb with crepe bandage.
- Anticoagulants: Heparin/low molecular weight heparin, warfarin, phenindione.
- ***For fixed thrombus:***
    - Initially high dose of heparin of 25,000 units/day for 7 days is given. Then later patient is advised to continue warfarin for 3-6 months. Dose is controlled by assessing *Activated Partial Thromboplastin Time (APTT).* Duration of heparin treatment is usually for 7-10 days. *Dose*

Oral contraceptives increases the risk of DVT by 3-5 times.

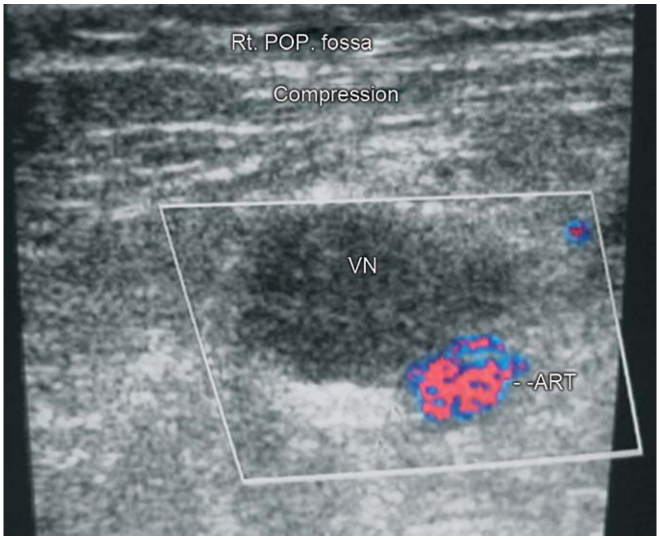

Fig. 1.430: Doppler study showing DVT.

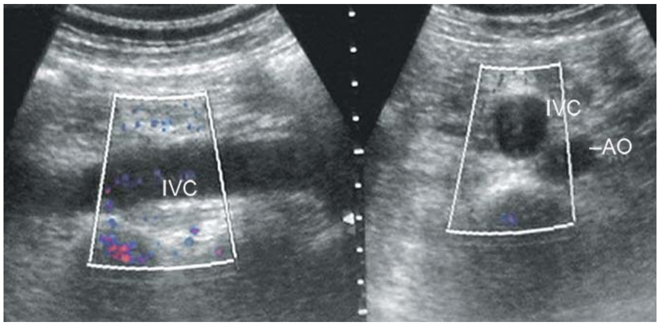

Fig. 1.431: U/S showing thrombosis of the IVC.

of heparin is often calculated as— 80 units/kg bolus of heparin followed by 15 units/kg of infusion.

  - Low molecular weight heparin is preferred to heparin.
  - Warfarin should be started as early as possible (*same day* of heparin therapy). Day one and day two—10 mg each day; day three—5 mg. On day three prothrombin time should be done. Warfarin is given for 3-6 months with regular monitoring, depending on the cause, risk group, and severity of DVT. INR should be maintained between 2.0 to 3.0.
  - Oral anticoagulants being teratogenic *cannot be used* during pregnancy. LMWH is the *drug* of *choice* used during pregnancy and postpartum period.

- ***For free thrombus*:**

Free thrombus

- Fibrinolysins
- Thrombectomy using Fogarty's catheter
- IVC filter

  - *Fibrinolysins*: Streptokinase, 6 lakhs to start with and later one lakh hourly. It is commonly infused directly into the affected vein through a venous catheter. Urokinase or tissue plasminogen activator may also be used to dissolve thrombus (it should not be given when patient is on heparin).

Fibrinolysins

- Tissue plasminogen activator—directly into the thrombus through popliteal/femoral vein
- Urokinase—1,20,000 to 2,50,000 units/hour; derived from human urine
- Reptilase—0.5-1 unit/hour
- Streptokinase—6 lakh to start with later one lakh hourly; derived from streptococci

  - Venous thrombectomy is done using Fogarty venous balloon catheter.
  - Thrombotic emboli is prevented from reaching the heart by filtering it at IVC level using intracaval filters—Kim ray Greenfield filter, suture sieve plication, stapler plication, vena caval ligation, Mobin Uddin umbrella filter.
  - Special thrombectomy device of 7-9 French sheath is passed through the thrombosed segment to have partial mechanical thrombectomy and through that thrombolytics (tissue plasminogen activator) are infused. Thrombus can be removed nowadays through balloon angioplasty tube. Open venotomy and thrombectomy also can be done.

- *Palma operation*: In iliofemoral thrombosis, common femoral vein below the block is communicated to opposite femoral vein through opposite long saphenous vein.
- *May-Husni operation*: When blockage is in popliteal vein, popliteal vein below the block is anastomosed to long saphenous vein (end-to-end) so as to bypass the blood across the popliteal block.

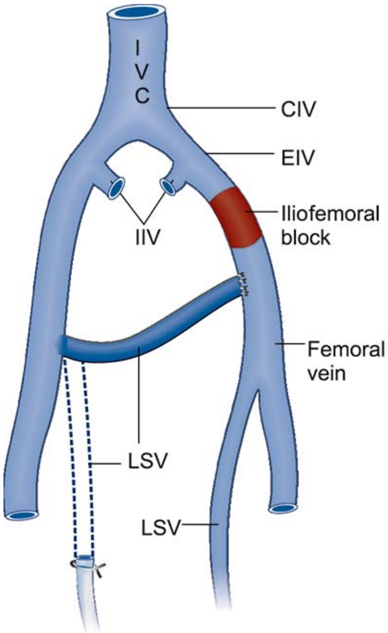

Fig. 1.432: Palma operation for iliofemoral block. Using opposite saphenous vein femoral vein is connected to other femoral vein.

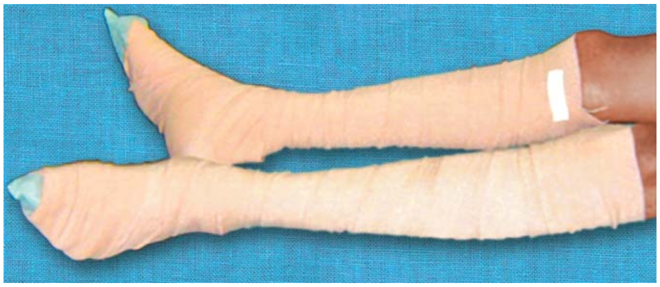

Fig. 1.433: Crepe bandages applied to both legs in bilateral varicose veins and DVT.

Prevention of DVT

- Categorise the patient as low/moderate and high risk. *Low-risk*—young patients undergoing surgery for less than 30 minutes. *Moderate-risk*—patients above 40 years of age undergoing major surgery. *High-risk*—one who had existing cardiac diseases, stroke, previous history of DVT, suffering from malignancy
- Mechanical methods—elastic compression bandage; elevation; external pneumatic bandage
- Pharmacological—low molecular weight heparin—once a day

- Care has to be taken to see for proper positioning of legs with no pressure on the calf muscles.
- Pressure bandage to the legs has to be applied during major surgeries, laparoscopic surgeries. During postoperative period, elevation, massaging, pressure bandage, early ambulation, maintaining hydration are essential measures.
- Low dose heparin is given in suspected cases, in major surgeries and continued during postoperative period till the patient is ambulated. 5,000 units is given subcutaneously 2 hours before surgery. Low molecular weight heparin is preferred drug than heparin as it can be used once a day; it does not require monitoring; it does not cause thrombocytopenia (like heparin); and there is lesser risk of bleeding.
- Various measures like graduated static compression, elastic stockings, electrical stimulation of calf muscles, pneumatic compression are used to prevent sluggish flow of blood.
- Dextran 70, intravenously 500 ml during surgery and another 500 ml postoperatively in 24 hours can also be used to prevent DVT.
- Smoking increases the viscosity of blood and so should be stopped.
- Patients on oral contraceptives or oestrogens should stop the drug 6-8 weeks prior to any elective surgery.

Effects and sequelae of DVT

- Pulmonary *embolism*—15%
- Infection; venous gangrene
- Partial recanalisation, chronic venous hypertension around the ankle region causing venous ulcers—*chronic venous insufficiency*—CVI
- Recurrent DVT—30%
- Propagation of thrombus proximally—20-30%

VARICOSE VEINS

They are dilated, tortuous, elongated veins in the leg. There is reversal of blood flow through its faulty valves.

It is *permanently* elongated, dilated vein/veins with tortuous path causing *pathological* circulation. *Risk factors* being heredity; female sex; occupation that demands prolonged standing; immobility; raised intra-abdominal pressure like in sports, tight clothing, pregnancy, raised progesterone level and altered estrogen-progesterone ratio, chronic constipation, high heels. *Prevalence* of varicose veins is 35%; severe varicose veins is 10%; chronic venous insufficiency (CVI) is 8%; ulcer is 2%.

Classification I

- Long/great saphenous vein varicosity.
- Short/small saphenous vein varicosity.
- Varicose veins due to perforator incompetence.

Classification II

- *Thread veins* (or dermal flares/telangiectasis/spider veins/ Hypen veins are 0.5-1 mm in size): Are small varices in the skin usually around ankle which look like dilated, red or purple network of veins (*Venulectasia*). Spider naevi/*venous flares* are common in females.
- *Reticular varices* (1-4 mm in size): Are slightly larger varices than thread veins located in subcutaneous/subdermal region.
- *Varicose veins*. Dilated palpable subcutaneous veins more than 4 mm in diameter (specifically located in saphenous compartment).
- *Combination of any of the above*.

Small varicose vein is < 4 mm in diameter. Large varicose vein is > 4 mm in diameter.

Corona phlebectatica are blue telangiectasias on the medial aspect of the *foot below the malleolus* around ankle level. More than 5 such lesions are the best independent predictor of the skin changes.

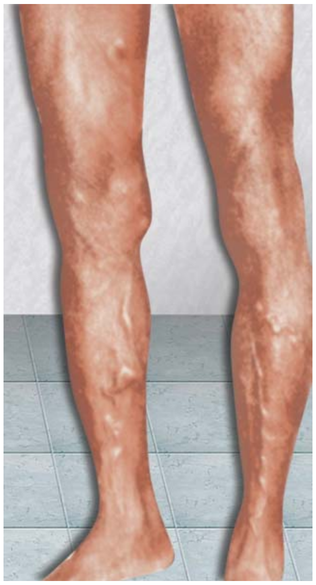

Fig. 1.434: Varicose veins in lower limbs.

Pulsatile varicose veins in lower limb is seen in Klippel Trenauny syndrome.

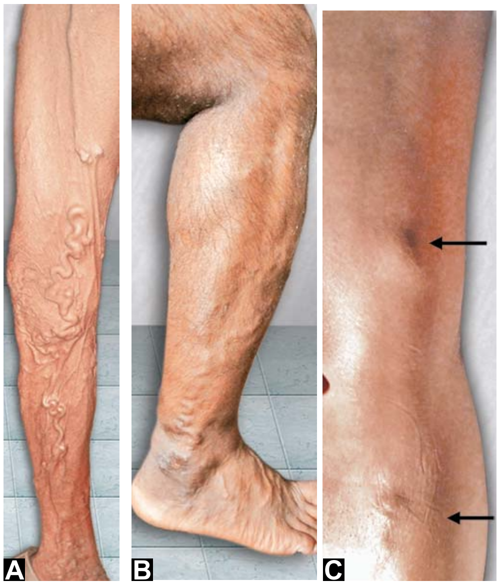

Figs 1.435A to C: (A) Great saphenous vein varicosity. (B) Small saphenous vein varicosity. (C) Perforator incompetence (*blow outs*).

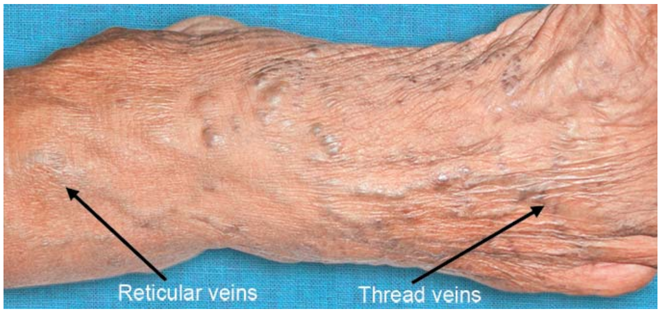

Fig. 1.436: Thread veins are up to 1 mm diameter; reticular veins are 1-4 mm in diameter.

Classification III

CEAP Classification of Lower Limb Varicose Veins (2004)

CEAP classification

C—*Clinical* signs (grade 0-6); (A) for asymptomatic or (S) for symptomatic presentation

E—*Etiological* classification: Congenital (Ec), Primary (Ep), Secondary (Es), No venous etiology (En)

A—*Anatomic* distribution: Superficial (As), Deep (Ad) or Perforator (Ap), No venous location identified (An)

P—*Pathophysiologic* dysfunction: Reflux (Pr), Obstructive (Po), Both, or No pathophysiology identified (Pn)

Grading of clinical signs (C)

0—No visible or palpable signs of venous diseases
1—Telangiectases, reticular veins or malleolar flare
2—Varicose veins
3—Oedema without skin changes
4—Skin changes due to venous diseases like pigmentation, eczema or lipodermatosclerosis 4a—pigmentation; 4b—lipodermatosis, atrophia blanche
5—Skin changes as above with healed ulceration
6—Skin changes as above with active ulceration

Anatomical distribution (A)

As—superficial system:
1–Telangiectases, reticular veins
2–Great saphenous vein above the knee—ostial and preterminal
3–Great saphenous vein below the knee
4–Small saphnous vein
5–Nonsaphenous—43%

Ad—deep system:
From 6 to 15

Ap—perforator system:
17–Perforator vein (PV) of the thigh
18–Perforator (PV) of the calf and leg

An—no anatomical lesion identified

Pathogenesis

Two theories

- Fibrin cuff theory
- White cell trapping theory

Incompetence of venous valves
↓
Stasis of blood
↓
Chronic ambulatory venous hypertension
↓
Defective microcirculation
↓
RBC diffuses into tissue planes
↓
Lysis of RBC's
↓
Release of haemosiderin
↓
Pigmentation
↓
Dermatitis
↓
Capillary endothelial damage
↓
Prevention of diffusion and exchange of nutrients
↓
Severe *anoxia*
↓
Chronic venous ulceration ***(Fibrin cuff theory*).**

- Inappropriate activation of trapped leucocytes release proteolytic enzymes which cause cell destruction and ulceration—***White cell trapping theory***. Fibrin deposition, tissue death, scarring occurs together, called as *lipodermatosclerosis***.**
- *Secondary valvular failure* → venous reflux → venous wall dilatation → effects. Weakening of the venous endothelial wall and valves occur due to raised venous wall tension by—(1) Shearing stress pressures of blood flow, (2) Increased matrix metalloproteinases (MMPs) activity on endothelium and smooth muscle cells reducing structural integrity of

venous wall with decreased elastin content in the media of the vein, (3) Changes in normal venous constriction and relaxation properties, (4) Recurrent inflammation.

- *Venous system* in the lower limb is maintained by—(1) *Valvular competence,* (2) *Venous patency,* (3) *Calf muscle pump* which is venous channel/plexus within the soleus muscle. Any change in any of these systems can cause venous insufficiency.
- *Chronic venous insufficiency (CVI)* is a syndrome resulting from continuous chronic venous hypertension/ambulatory venous hypertension [AVP] (> 80 mmHg venous pressure at ankle) in the erect posture either on standing or exercise (in normal people venous pressure in superficial system falls during calf contraction). CVI consists of postural discomfort, varicose veins, oedema, pigmentation, induration, dermatitis, lipodermatosclerosis and ulceration. CVI patients may be having superficial vein incompetence (30%) with or without perforator incompetence or deep vein incompetence (30%) or having previous DVT with complete obliteration or partial recanalisation with incompetence called as *post-thrombotic syndrome* (30%).
- Varicose vein is a condition of *progressive deterioration* even often with interventions.

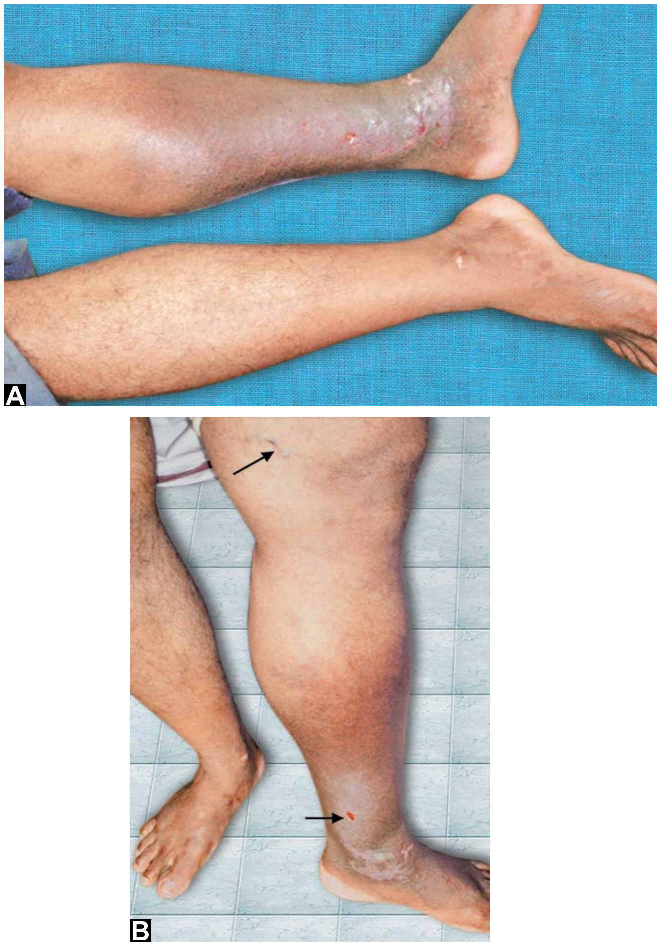

Figs 1.437A and B: Lipodermatosclerosis is pigmentation, thickening, and induration of the skin due to venous diseases.

Aetiology of Varicose Veins

Varicosities are more common in lower limb because of erect posture and long column of blood has to be supported which can lead to weakness and incompetency of valves. Incidence is 5% of adult population.

- *Primary varicosities* due to:
  - Congenital incompetence or absence of valves.
  - Weakness or wasting of muscles—defective connective tissue and smooth muscle in the venous wall.
  - Stretching of deep fascia.
  - Inheritance (family history) with FOXC2 gene.
  - Klippel-Trenaunay syndrome, avalvulia, Parkes-Weber syndrome. Here varices are of atypical distribution.
- *Secondary varicosities*:
  - Recurrent thrombophlebitis.
  - Occupational—standing for long hours (traffic police, guards, sportsman).

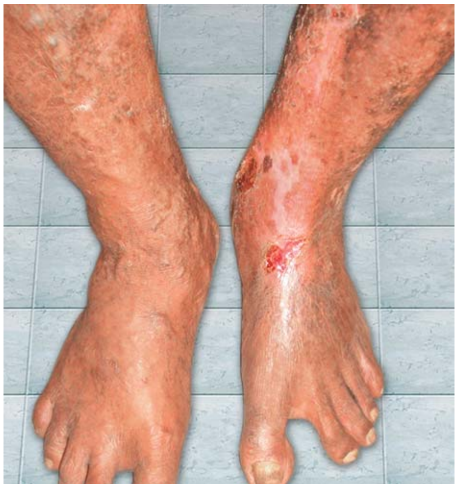

Fig. 1.438: Bilateral varicose veins with pigmentation, ulcer and lipodermatosclerosis.

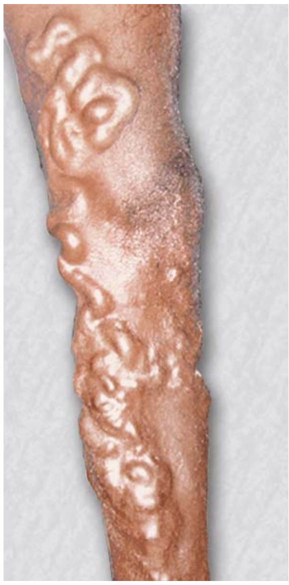

Fig. 1.439: Long saphenous vein varicosity.

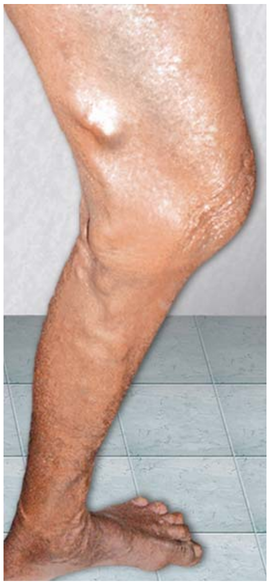

Fig. 1.440: Long saphenous vein varicosity on the medial aspect.

Venous claudication is acute bursting pain on ambulation due to chronic venous insufficiency.

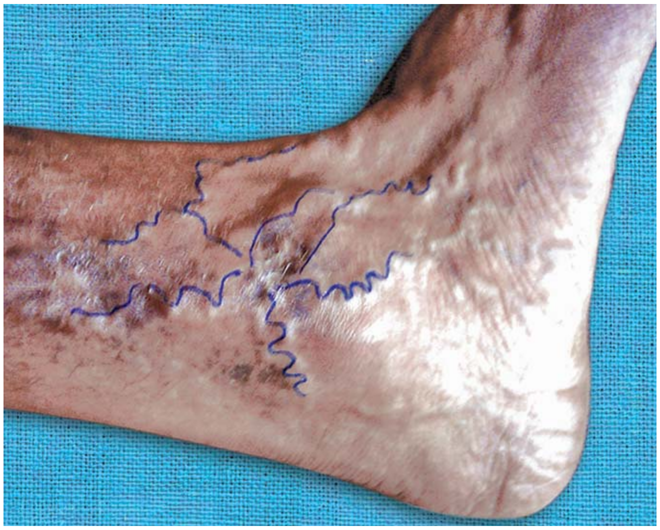

Fig. 1.441: Ankle flare seen around ankle, which eventually leads into complications of varicose veins.

- Obstruction to venous return like abdominal tumour, retroperitoneal fibrosis, lymphadenopathy, ascites.
- Pregnancy (due to progesterone hormone), obesity, chronic constipation.
- AV malformations—congenital or acquired.
- Iliac vein thrombosis.
- Tricuspid valve incompetence.

Sites where varicosities can occur

- Lower limb
- Pampiniform plexus of veins
- Vulva, perineum
- Sites of portosystemic anastomosis

Predisposing factors for varicose veins are—age, sex, race, obesity, height, left > right, occupation, family history, erect posture.

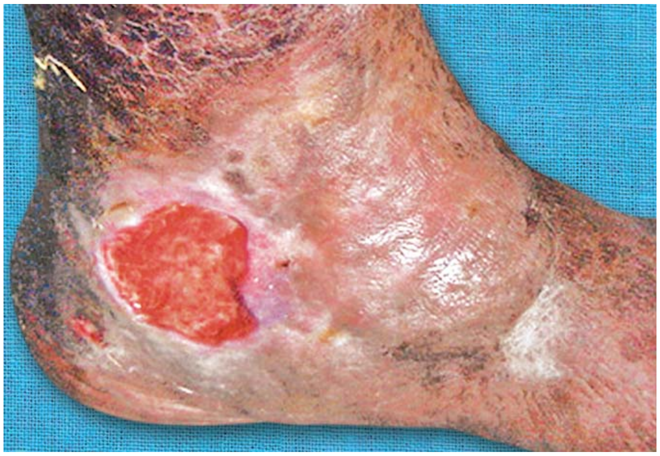

Fig. 1.442: Typical site of venous ulcer. Note the pigmentation and chronicity.

Clinical Features

Symptoms in varicose veins

- Dragging pain, postural discomfort
- Heaviness in the legs
- Night time cramps—usually late night
- Oedema feet, itching (feature of CVI)
- Discolouration/ulceration in the feet/painful walk

Causes of pain/cramps in varicose veins/venous diseases

- Increased venous wall tension—chronic venous hypertension
- Hypoxia of tunica media of the venous wall due to altered function of vasa vasorum
- Increased capillary pressure
- Hyperviscosity of red cells—haemorrheological disorders
- Platelet hyperaggregation
- Reduction in capillary permeability causing *capillary functional disorder*
- Altered cutaneous microcirculation due to leukocyte adhesion and accumulation into the venous wall; release of free radicals cause *microvascular lesional disease*

Oedema in venous diseases

- Can be localised or generalised
- Localised oedema is due to ankle flare or dilatation of medial marginal vein
- Cellulitis and lymphangitis association causes oedema
- Scarring and thickening of dermal and subdermal tissues—*lipodermatosclerosis* (brawny induration)
- Ankle becomes narrower due to contraction of skin and subcutaneous tissues but calf remains prominent—*champagne bottle appearance*
- Pale atrophic skin with white patches surrounded by dilated capillaries and pigmentation—*atrophic blanche*

It is more common in females (10 : 1). Often it is familial. *Familial varicose veins* begin in younger age group, seen bilaterally, involves all veins including deep veins.

Signs

- Visible dilated veins in the leg with pain, distress, nocturnal cramps, feeling of heaviness, pruritus.
- Pedal oedema, pigmentation, dermatitis, ulceration, tenderness, restricted ankle joint movement.
- Bleeding, thickening of tibia occurs due to periostitis.
- Positive cough impulse at the saphenofemoral junction.
- *Saphena varix*—a large varicosity in the groin, which becomes visible and prominent on coughing.
- **Brodie-Trendelenburg test**: Vein is emptied by elevating the limb and a tourniquet is tied just below the saphenofemoral junction (or using thumb, saphenofemoral junction is occluded). Patient is asked to stand quickly. When

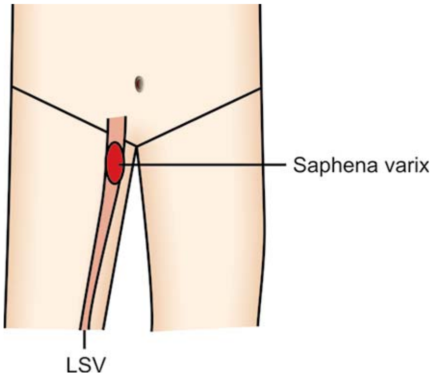

Fig. 1.443: Site of saphena varix—in the lower part of groin just below the inguinal ligament.

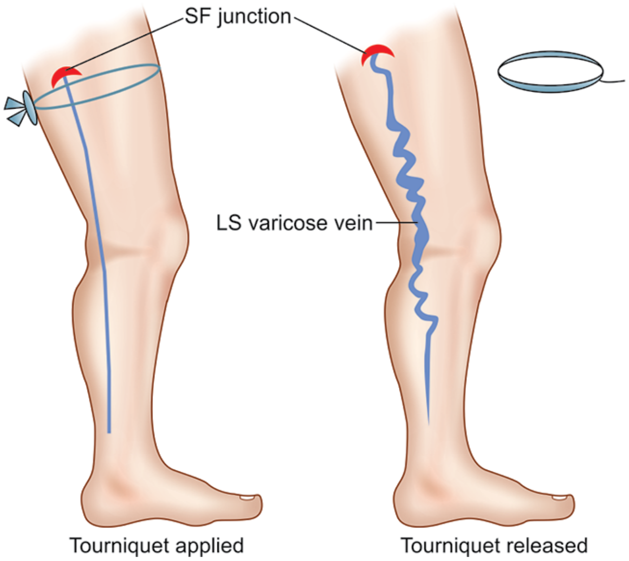

Fig. 1.444: Brodie-Trendelenburg test. Note the reversal of blood flow while releasing the tourniquet.

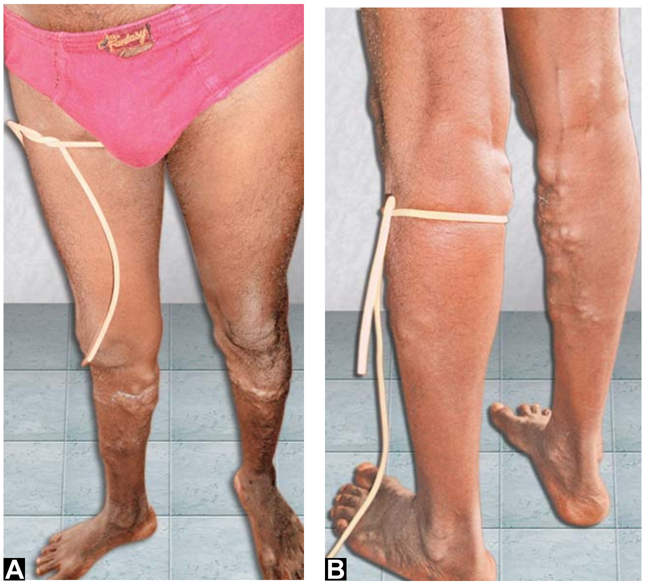

Figs 1.445A and B: Tourniquet tests for long saphenous vein and short saphenous vein.

tourniquet or thumb is released, rapid filling from above signifies saphenoemoral incompetence. *This is Trendelenburg test I.*

In *Trendelenburg test II*, after standing tourniquet is not released. Filling of blood from below upwards rapidly can be observed within 30-60 seconds. It signifies perforator incompetence.

- ***Perthe's test:*** The affected lower limb is wrapped with elastic bandage and the patient is asked to walk around and exercise. Development of severe cramp like pain in the calf signifies DVT.
- ***Modified Perthe's test***: Tourniquet is tied just below the saphenofemoral junction without emptying the vein. Patient is allowed to have a *brisk walk* which precipitates *bursting pain* in the calf and also makes superficial veins more prominent. It signifies DVT.

  DVT is contraindicated for any surgical intervention of superficial varicose veins. It is also contraindicated for sclerosant therapy.
- ***Three tourniquet test***: To find out the site of incompetent perforator, three tourniquets are tied after emptying the vein.
  - At saphenofemoral junction.
  - Above knee level.
  - Another below knee level.

  Patient is asked to stand and looked for filling of veins and site of filling. Then tourniquets are released from below upwards, again to see for incompetent perforators.
- ***Schwartz test:*** In standing position, when lower part of the long saphenous vein in leg is tapped, impulse is felt at the saphenous junction or at the upper end of the visible part of the vein. It signifies continuous column of blood due to valvular incompetence.
- ***Pratt's test***: Esmarch bandage is applied to the leg from below upwards followed by a tourniquet at saphenofemoral junction. After that the bandage is released keeping the tourniquet in the same position to see the "*blow outs*" as perforators.
- ***Morrissey's cough impulse test:*** The varicose veins are emptied. The leg is elevated and then the patient is asked to cough. If there is saphenofemoral incompetence, expansile impulse is felt at saphenous opening. It is a venous thrill due to vibration caused by turbulent backflow.
- ***Fegan's test***: On standing, the site where the perforators enter the deep fascia bulges and this is marked. Then on lying down, button like depression (crescent like) in the deep fascia is felt at the marked out points which confirms the perforator site.
- ***Ian-Aird test***: On standing, proximal segment of long saphenous vein is emptied with two fingers. Pressure from proximal finger is released to see the rapid filling from above which confirms saphenofemoral incompetence.
- Examination of the abdomen has to be done to look for pelvic tumours, lymph nodes, which may compress over the veins to cause varicosity

Everybody is in favour of progress. It is the change they don't like.

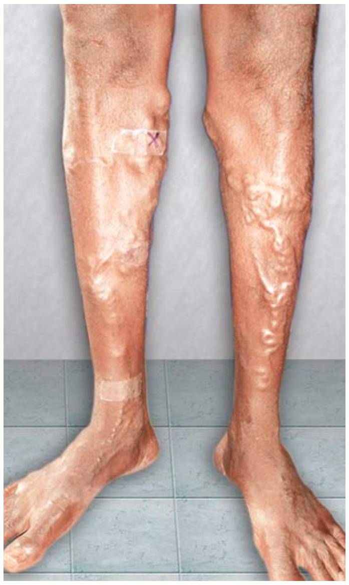

Fig. 1.446: Bilateral varicose veins. Note the tortuous long saphenous vein.

Venous disability scoring system

Score 0	Asymptomatic
Score 1	Symptomatic but able to carry out activities without any therapy
Score 2	Symptomatic—can do activities only with compression/limb elevation
Score 3	Symptomatic—unable to do daily activities even with compression or limb elevation

Venous segmental disease score (venous clinical scoring system) is done based on different symptoms/signs/ulcer activity/compression therapy with 10 parameters with each having 3 scores as mild/moderate/severe.

Complications of Varicose Veins

- *Haemorrhage:* Venous haemorrhage can occur from the ruptured varicose veins or sloughed varicose veins, often torrential, but can be controlled very well by elevation and pressure bandage.
- *Pigmentation, eczema and dermatitis.*
- *Periostitis* causing thickening of periosteum. It delays healing of ulcer due to poor perfusion of ulcer bed.
- *Venous ulcer.*
- *Marjolin`s ulcer*—due to *unstable scar of long duration*—very well differentiated squamous cell carcinoma.
- *Lipodermatosclerosis.*
- *Ankylosis* of the ankle joint is due to fibrosis of soft tissues around ankle joint—fibrous ankylosis.
- *Talipes equinovarus*, wherein patient walks on the tip of toes like horse.
- *Deep venous thrombosis* per se due to varicose vein is rare but can occur if there is associated deep vein disease or recurrent thrombophlebitis.
- *Calcification of the wall* of varicose veins or of sclerosed soft tissue.
- *Recurrent thrombophlebitis*, clot formation on the superficial system often at perforator level which often get infected causing fever and tenderness over the spot.

Investigations

- ***Venous Doppler:***
  - With the patient standing, the doppler probe is placed at saphenofemoral junction and later wherever required. Basically by hearing the changes in sound, venous flow, venous patency, venous reflux can be very well-identified.
  - *Doppler test:* When a *hand held Doppler* (continuous wave 8 MHz flow detector) is kept at SFJ, typical

Fig. 1.447: Doppler machine to assess venous system and its problems.

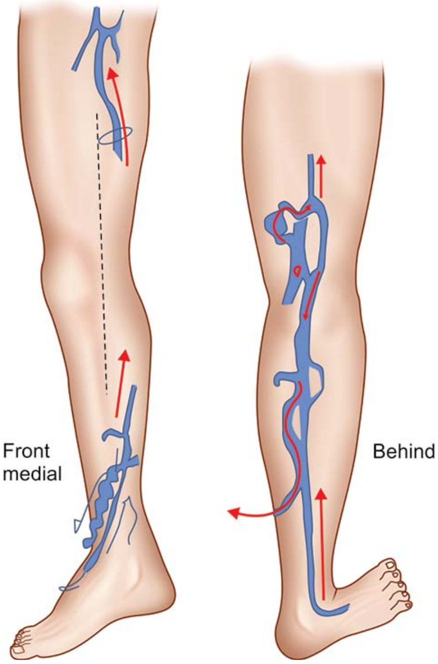

Fig. 1.448: Venous haemodynamic mapping (VHM) of the lower limb.

audible, '*whoosh signal'* > 0.5 sec while performing Valsalva manoeuvre is the sign of reflux at SFJ. It is also used at SPJ and at perforators. *Note:* All clinical/phlebological tests mentioned above have been superseded by Doppler test. *Doppler test is considered to be a clinical method.*

- ***Duplex scan:***
  - It is a highly reliable U/S Doppler imaging technique (here high resolution B mode ultrasound imaging and Doppler ultrasound is used) which along with direct visualisation of veins, gives the functional and anatomical information, and also colour map. Examination is done in standing, lying down position and also with Valsalva manoeuvre. *Hand-held Doppler* probe is placed over the site and visualised for any block and reversal of flow. DVT is very well-identified by this method.
  - *Venous haemodynamic mapping*/VHM/Cartography is essential prior to surgery.

***Note*:**
Proper *venous haemodynamic mapping (VHM)* is essential.

Venous Doppler in varicose veins

- To find out DVT—very important
- To find out saphenofemoral, saphenopopliteal incompetence
- To find out perforator incompetence
- Uniphasic signals signify flow in one direction—normal
- Biphasic flow signifies reversal flow with incompetence

  - **Digitally coded free flow (B flow) USG** (Professor Feder Lurie of Hawaii) allows simultaneous visualisation of flowing blood/blood cells and surrounding stationary structures to give proper *haemodynamic imaging* with functioning mechanism of venous valves, valve leaflets and flow across leaflets. This may be the *ideal* tool of investigation in future.

- ***Venography:***
  - *Ascending venography* was very common investigation done before Doppler period.

    A tourniquet is applied above the malleoli and vein of dorsal venous arch of foot is cannulated. Water soluble dye injected, flows into the deep veins (because of the applied tourniquet). X-rays are taken below and above knee level. Any block in deep veins, its extent, perforator status can be made out by this.

    It is a good reliable investigation for DVT.

    If DVT is present, surgery or sclerotherapy are contraindicated.
  - *Descending venogram* is done when ascending venogram is not possible and also to visualise incompetent veins. Here contrast material is injected into the femoral vein through a cannula in standing position. X-ray pictures are taken to visualise deep veins and incompetent veins.

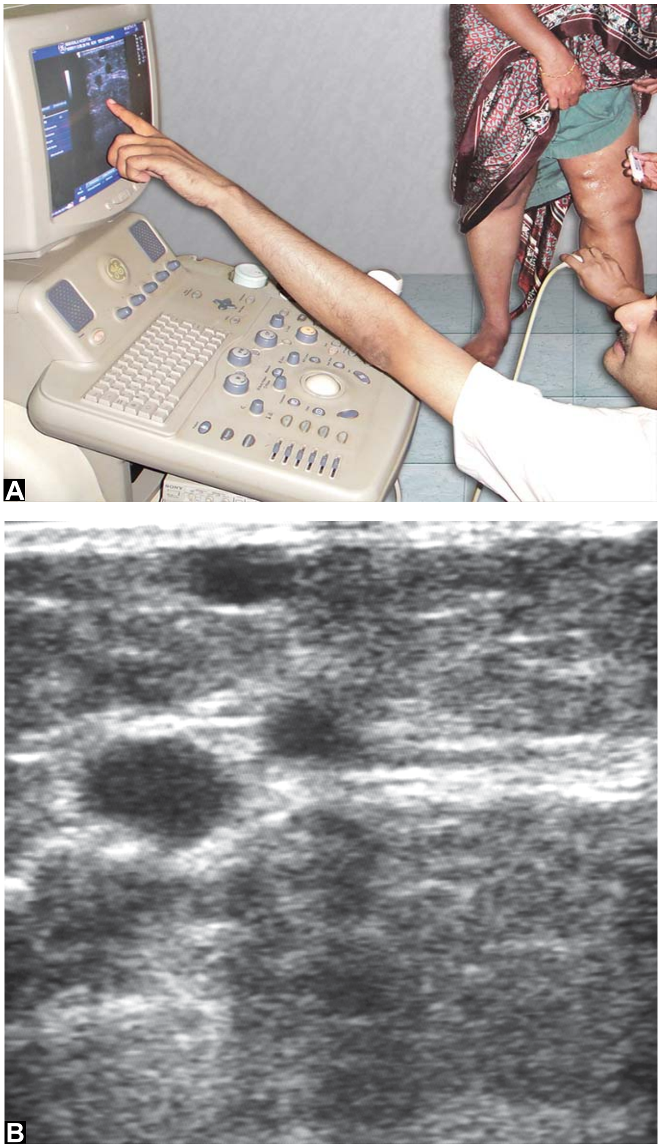

Figs 1.449A and B: Duplex scan for venous diseases of lower limb should also be done in standing position.

Phlebography

- Ascending phlebography defines obstruction
- Descending phlebography identifies valvular incompetence
- Regularly not required to be done

- ***Plethysmography*:**
  - It is a noninvasive method which measures volume changes in the leg. It gives *functional* information on venous volume changes and *calf muscle pump* insufficiency.
  - ***Photoplethysmography***: Using probe transmission of light through the skin, venous filling of the surface venules which reflects the superficial venous pressure is measured. Initially patient performs dorsiflexion at ankle for 10 times to empty the venules and pressure tracing falls in photoplethysmography. Patient takes rest and refilling occurs. In normal people, it occurs through

arterial inflow in 20-30 seconds. In venous incompetence filling also occurs by venous reflux and so refilling time is faster than normal.

Disadvantage: Site of reflux cannot be localised by this method.

- ***Air plethysmography***: Patient is initially in supine position with veins emptied by elevation of leg. Air filled plastic pressure bladder (inflatable polyurethane cuff filled with air) is placed on calf to detect volume changes. Minimum volume is recorded. Patient is turned to upright position and venous volume is assessed. Maximum venous volume divided by time required to achieve maximum venous volume gives the *venous filling index (VFI)*. VFI is a measure of reflux. *Ejection fraction* is volume change measured prior and after single tiptoe manoeuvre which is a measure of calf pump action. *Residual venous fraction* is an index of overall venous function which is venous volume in the leg after ten toetip manoeuvres divided by venous volume prior to manoeuvre. *Increased VFT and diminished ejection fraction in a patient will benefit from surgery.*

- ***Ambulatory venous pressure (AVP):***
  It is an invasive method. Needle inserted into dorsal vein of foot is connected to transducer to get its pressure which is equivalent to pressure in the deep veins of the calf. Ten tiptoe manoeuvres are done by the patient. With initial rise in pressure, pressure decreases and eventually stabilises with a balance. Pressure now is called as *ambulatory venous pressure (AVP)*. After stopping exercise, veins are allowed to refill with return of pressure to baseline. Time required for pressure to return to 90% of baseline is called as *venous refilling time (VRT).* Raise in AVP signifies venous hypertension. Patients with AVP more than 80 mmHg has got 80% chances of venous ulcer formation.
- ***Arm-foot venous pressure:***
  Foot pressure is not more than 4 mmHg above the arm pressure.
- ***U/S abdomen***, peripheral smear, platelet count, other relevant investigations are done depending on the cause of the varicose veins.
- If venous ulcer is present, then the discharge is collected for culture and sensitivity, biopsy from ulcer edge is taken to rule out Marjolin's ulcer.
- Plain X-ray of the part is taken to look for periostitis.
- ***Varicography***:
  Here nonionic, iso-osmolar, nonthrombogenic contrast is injected directly into the variceal vein to get a detailed anatomical mapping of the varicose veins. It is used in recurrent varicose veins.

Differential diagnosis

- Lymphoedema
- AV malformation
- Orthostatic oedema
- Renal and cardiac disease
- Hepatic causes
- Vasculitis
- Metabolic diseases like gout, myxoedema, and morbid obesity
- Chronic infections like tuberculosis, syphilis

Treatment

- ***Conservative treatment:***
  - *Elastic crepe bandage application* from below upwards or use of pressure stockings to the limb—pressure gradiant of 30-40 mmHg is provided.
  - *Diosmin therapy* which increases the venous tone.
  - *Elevation of the limb*—relieves oedema. Two short times, during day and full night, elevation of foot with feet above the level of heart and toes above the level of nose is the method.
  - *Unna boots*—provide nonelastic compression therapy. It comprises a gauze compression dressings that contain zinc oxide, calamine, and glycerine that helps to prevent further skin break down. It is changed once a week.
  - *Pneumatic compression method*—provide dynamic sequential compression.
  - These methods *reduce the AVP*, reduce transcapillary fluid leakage by increasing SC pressure and improve cutaneous microcirculation.
- ***Drugs used for varicose veins:***
  - Calcium dobesilate—500 mg BD. Calcium dobesilate improves lymph flow; improves macrophage mediated proteolysis; and reduces oedema.
  - Diosmin—450 mg BD.
  - Diosmin 450 mg + Hesperidin 50 mg (DAFLON 500 mg). Mainly used in relieving night cramps but not to improve healing of ulcers.
  - Toxerutin 500 mg BD, TID. Antierythrocyte aggregation agent which improves capillary dynamics.
  - Diosmin is micronized purified flavanoid fraction. It protects venous wall and valve, and it is anti-inflammatory, profibrinolytic, anti-oedema, lymphotropic.
  - Benzopyrones, saponins, plant extracts, Ruscus (venular α 1 adrenergic receptor partial agonist) coumarins are different drugs used.

  Benefits of all these drugs are doubtful.
- ***Injection—sclerotherapy:***
  - *Fegan's technique:* By injecting sclerosants into the vein, complete sclerosis of the venous walls can be achieved.

  Indications
  - Uncomplicated perforator incompetence.
  - In the management of smaller varices—reticular veins, thread veins (telangiectasis).
  - Recurrent varices.
  - Isolated varicosities.
  - Aged/unfit patients.

Sclerosants used are:

- Sodium tetradecyl sulphate 3% (STDS)—commonly used
- Sodium morrhuate
- Ethanolamine oleate
- Polidocanol

Mechanisms of action

- Causes aseptic inflammation
- Causes perivenous fibrosis leading to block
- Causes approximation of intima leading to obliteration by endothelial damage
- Alters intravascular pH/osmolality
- Changes surface tension of plasma membrane

  - A 23 gauge needle is inserted into the vein (3-8 mm sized) and vein is emptied. 0.5-1 ml of sclerosant is injected into the vein and immediately compression is applied on the vein (prevent the entry of blood which may cause thrombosis, which later gets recanalised, further worsening the condition) so as to allow the development of sclerosis and to have proper endothelial apposition.
  - Usually injection is started at the ankle region and then proceeded upwards along the length of the veins at different points. Later pressure bandage is applied for six weeks. Often injections may have to be repeated at 2-4 weeks intervals for 2-4 sessions. Technique is called as *macrosclerotherapy.*
  - Entrapped blood may require to be evacuated after 14 days which is often essential to prevent recanalisation.
- ***Microsclerotherapy:***

  Very dilute solution of sclerosing agent like STDS, (0.1% of 0.1 ml—dilute) Polidocanol is injected into the thread veins and reticular veins followed by application of compression bandage (30 G needle). Dermal flare will disappear well by this method.
- ***Transillumination microsclerotherapy*** *(vein—lite):* It is better imaging of the veins using light generated by halogen bulb with high quality fibre illumination over the skin uniformly and passing 30 gauge needle for microsclerotherapy.
- ***Foam sclerotherapy by Tessari:***
  - STDS taken in a syringe is passed rapidly into another syringe which contains air to result in formation of foam. 1 ml of STD is mixed with 4 ml of air to make 5 ml of foam which is injected to vein. Total 6 ml maximum of STDs with 30 ml foam can be used. This foam in much larger quantity is injected into the superficial vein. Air get absorbed between foam and endothelial lining is destroyed. Foam minimises thrombosis by pushing the blood out of the site of the vessel where action is needed. Polidocanol/STDs is used for foam sclerotherapy.
  - *Advantages:* Cheap, technically easy, easily available, OPD procedure, can be repeated many times, anaesthesia is not needed, can be used along with other procedures for varicose veins.
  - *Complications:* Headache, transient blindness, stroke, air embolism, thrombophlebitis, pain over injected site, pigmentation.
  - *Contraindications:* Peripheral arterial disease, DVT.
- ***Echosclerotherapy:***

  Sclerotherapy is done under duplex ultrasound image guidance.
- ***Catheter directed sclerotherapy:***

  It is devised at Miami vein clinic with specific catheter for sclerotherapy. This catheter has got side holes all around the specific length for uniform contact of venous wall with the foam. It also has got a balloon at the tip

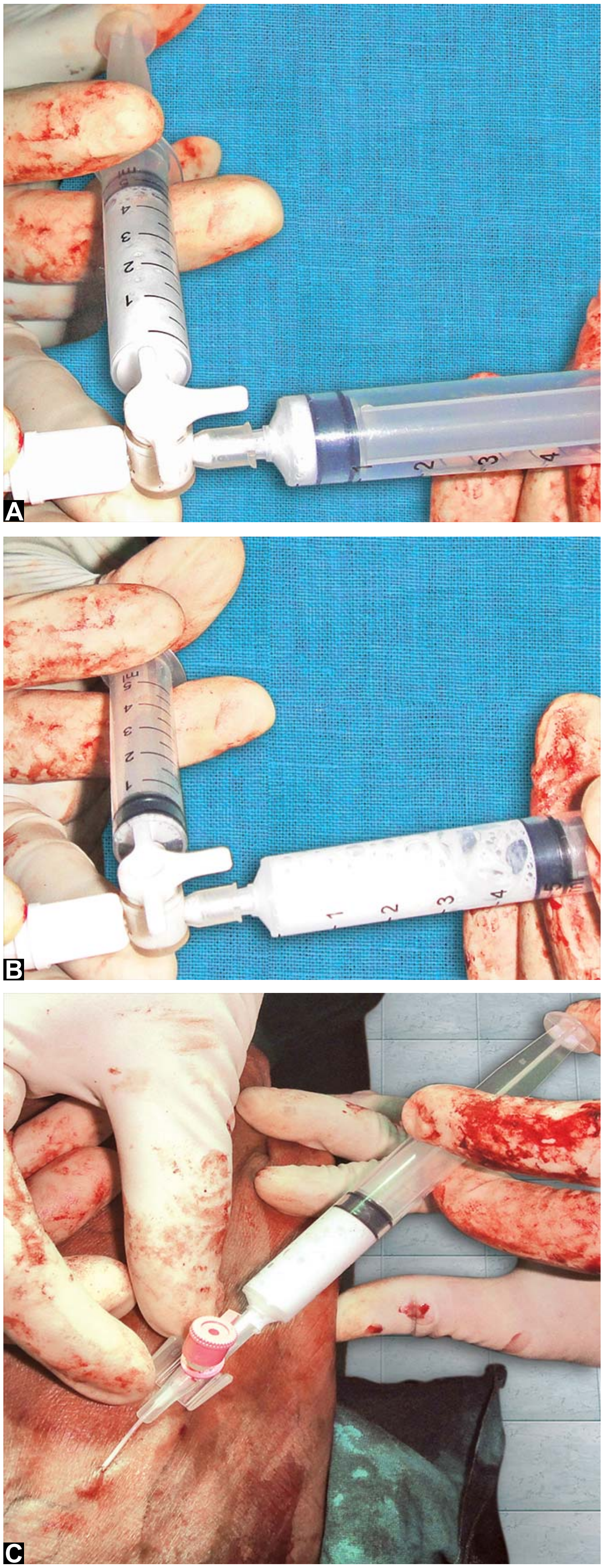

Figs 1.450A to C: Technique making foam and injecting into the vein. 4 ml air with 1 ml STD is mixed vigorously using 3 way stopcock and two 5 ml syringes. Created foam is injected into the vein immediately. Total of 6 ml STD (30 ml foam) can be injected.

INR (International Normalised Ratio): Ratio of measured PT to a mean lab control PT corrected for the sensitivity of thromboplastin

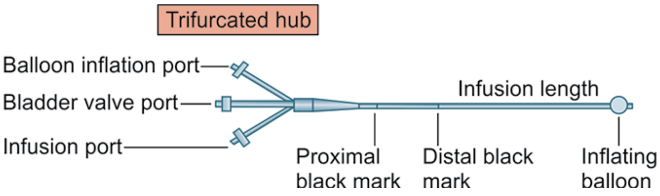

Fig. 1.451: Catheter directed sclerotherapy.

which after inflation blocks the SFJ thus preventing embolization of foam. It has got three external ports one for balloon inflation; one for bladder valve port; one for injection. This technique is under trial.

Contraindications for sclerotherapy

- Saphenofemoral incompetence
- Deep venous thrombosis
- Huge varicosities—may precipitate DVT
- Peripheral arterial diseases
- Hypersensitivity/immobility
- Venous ulcer—relative contraindication

Advantages of sclerotherapy

- It can be done as an outpatient procedure.
- It does not require anaesthesia.

Disadvantages of sclerotherapy

- Inadvertent subcutaneous injection can cause skin necrosis or abscess formation.
- Anaphylaxis, vasovagal shock, allergy.
- Hyperpigmentation.
- Thrombophlebitis.
- Deep venous thrombosis can occur.
- Inadvertent intra-arterial injection—serious complication.
- Intravenous haematoma.
- Temporary ocular disturbances.
- Skin staining, injection ulcers, persistent local pain.

Remember about sclerotherapy

- Current place of sclerotherapy is mainly for recurrent varicosities and thread/telangiectatic veins
- Sodium tetradecyl sulphate (STDS) is most commonly used sclerosant
- Hyperpigmentation is common after STDS
- Anaphylaxis is common after sodium morrhuate
- Anaphylaxis is least with polidocanol
- Extravasation (presence of pain/irritation/burning) should be avoided as it will cause skin necrosis
- Postsclerotherapy walking immediately after injection for 30 minutes with elastic bandage in place prevents/minimises the chances of DVT
- Compression bandage should be worn for minimum period of 6 weeks
- Sclerotherapy can be—macrosclerotherapy; microsclerotherapy; echo (ultrasound guided) sclerotherapy; foam sclerotherapy; transilluminated sclerotherapy

Indications for interventional procedures: *To relieve* **Complaints** which are relevant like pain, discomfort; to reverse **Complications** like skin changes, ulcers, bleeding, superficial thrombophlebitis, lipodermatosclerosis; **Cosmesis; prophylaxis** (not well accepted)**.**

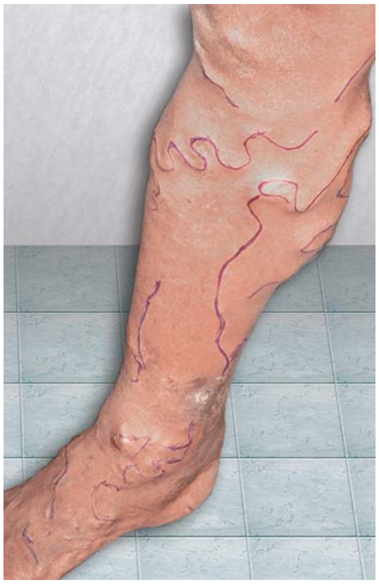

Fig. 1.452: Veins should be marked just prior to surgery using marking ink.

- ***Surgery:***
  - ***Trendelenburg operation (Crossectomy)***
    It is juxtafemoral flush ligation of long saphenous vein (i.e. flush with femoral vein), after ligating named (superficial circumflex, superficial external pudendal, superficial epigastric vein), deep external pudendal vein and unnamed tributaries. All tributaries should be ligated, otherwise recurrence will occur. Double saphenous vein is the most common anomaly occurring near saphenovenous junction.
  - **Stripping of vein**
    - Using *Myer's stripper* vein is stripped off. Stripping from below upwards is technically easier. Immediate application of crepe bandage reduces the chance of bleeding and haematoma formation. Stripping avulses the vein as well as obliterates the tributaries. Babcock's stripper and rigid metal pin stripper can also be used.
    - ***Two methods of stripping*** are used: (1) *Extraluminal collision technique* using Myer's stripper (Acorn head stripper, 73 cm long) is practiced since long time; but it damages the adjacent tissue, causes infection, postoperative pain, discomfort and haematoma along the stripped tract with possibility of revascularisation of the tract haematoma. (2) *Invagination technique* is better with less damage to adjacent tissue. *Codman's stripper* is used for invagination technique.
    - Stripping of the short saphenous vein is done from ankle below upwards after passing stripper from above downwards. It obliterates the mid calf perforator vein which is the common reason for recurrences.
    - ***Complications of stripping:*** *Saphenous neuralgia* (1%) due to saphenous nerve injury/avulsion; numbness and tingling along femoral nerve distribution; haematoma; infection; ulceration; *recurrence of the*

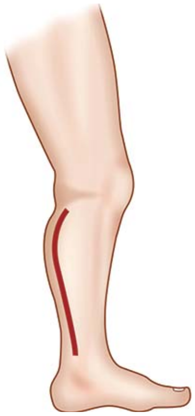

Fig. 1.453: Linton's vertical approach for subfascial perforator ligation in the leg.

disease is common (30%) which is not due to any technical default but due to progression of disease itself by *neo-angiogenesis and re-vascularisation*.

- *Stripping is not usually done for the veins in the lower part of the leg in LSV*.
- *Stripping of the vein is more effective than just ligation at the junction.*

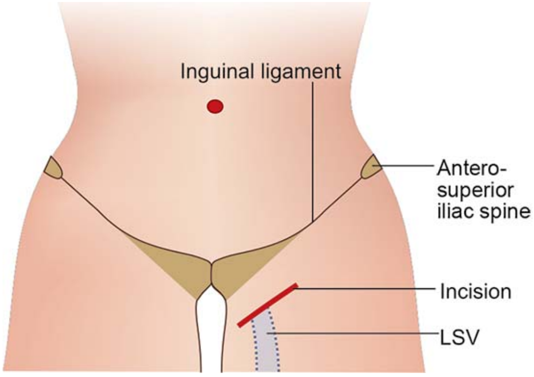
Inguinal ligament
Antero-superior iliac spine
Incision
LSV

Fig. 1.455: Incision for juxtafemoral flush ligation of long saphenous vein (Trendelenburg operation).

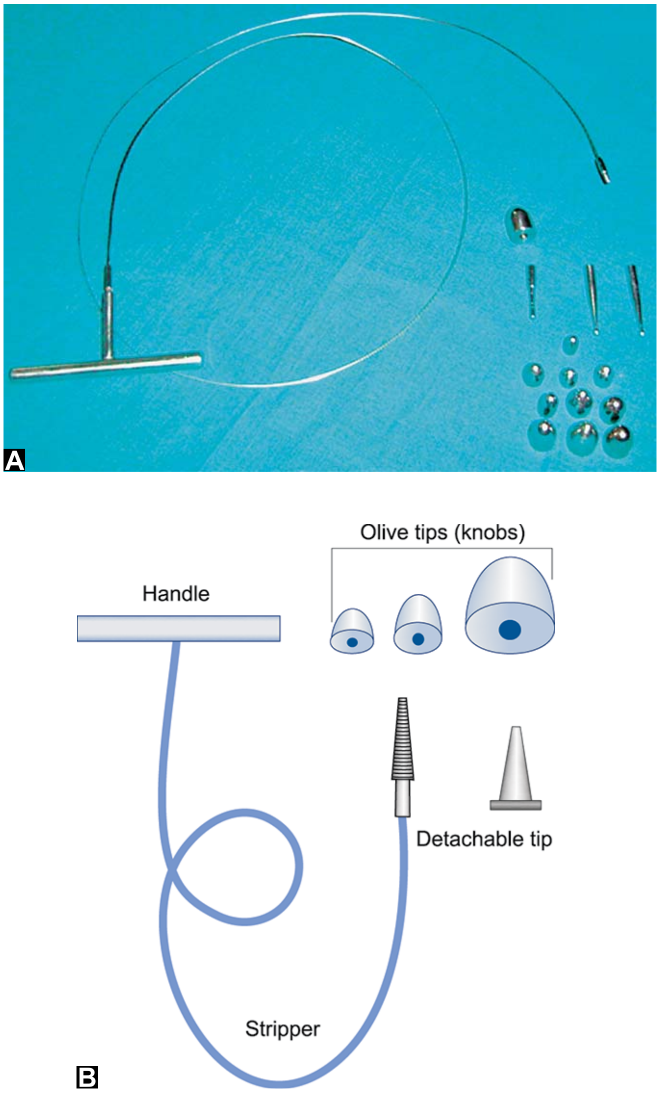
A
Olive tips (knobs)
Handle
Detachable tip
Stripper
B

Figs 1.454A and B: Photo of Myer's stripper with Olive tips and also diagrammatic look of the same.

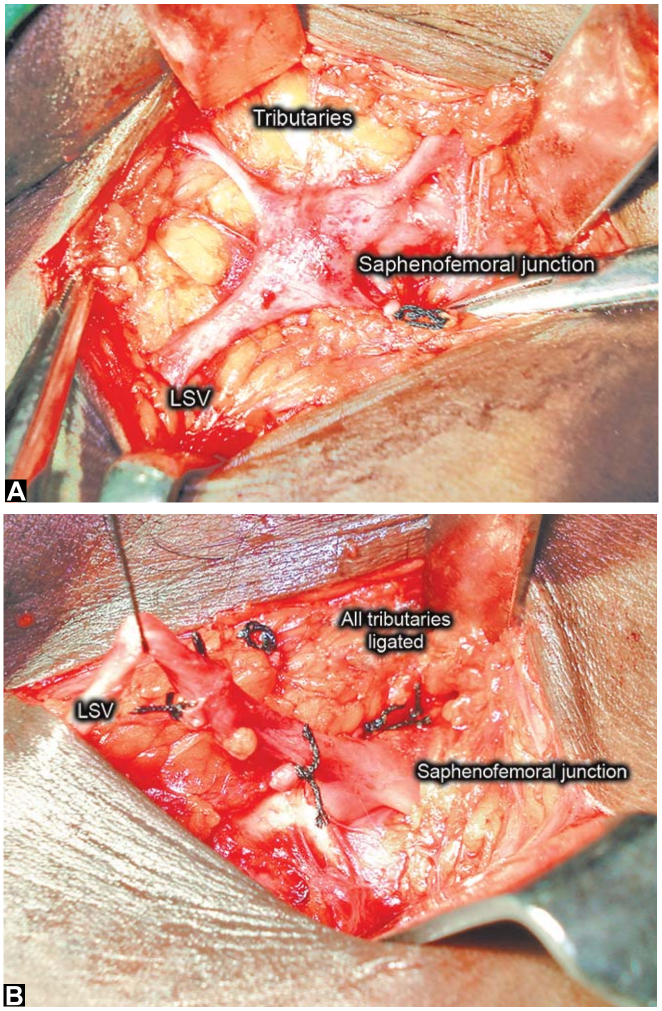
Tributaries
Saphenofemoral junction
LSV
A
All tributaries ligated
LSV
Saphenofemoral junction
B

Figs 1.456A and B

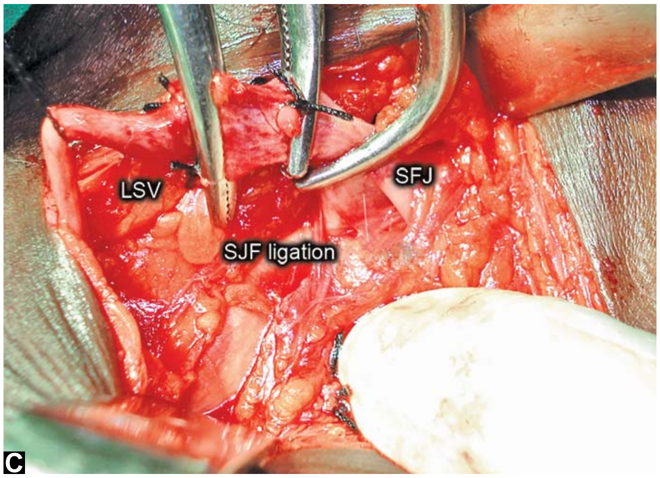

Fig. 1.456C

Figs 1.456A to C: Saphenofemoral junction and its ligation. Tributaries are well seen.

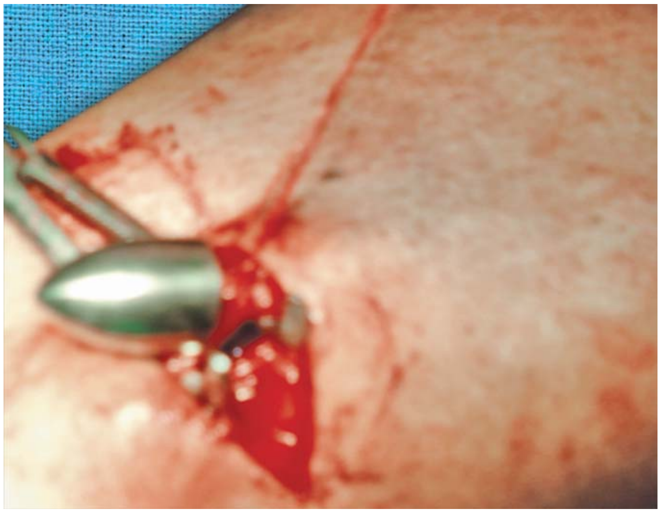

Fig. 1.457: Stripper knob on the distal part of the vein before stripping of the vein.

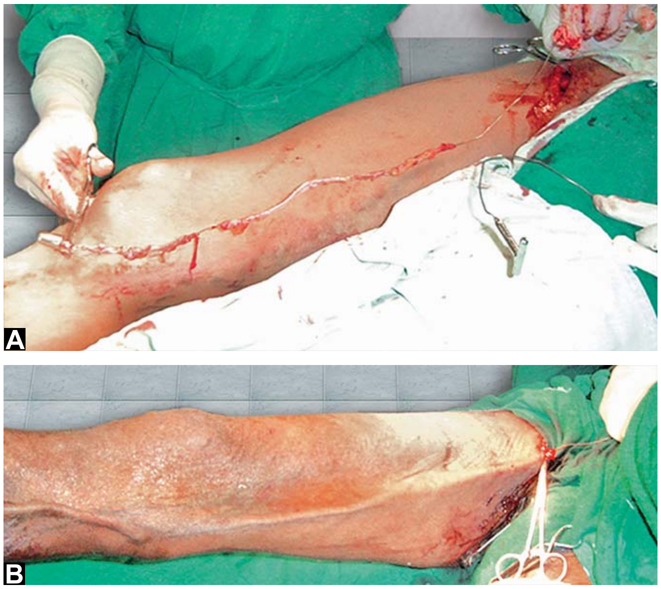

Figs 1.458A and B

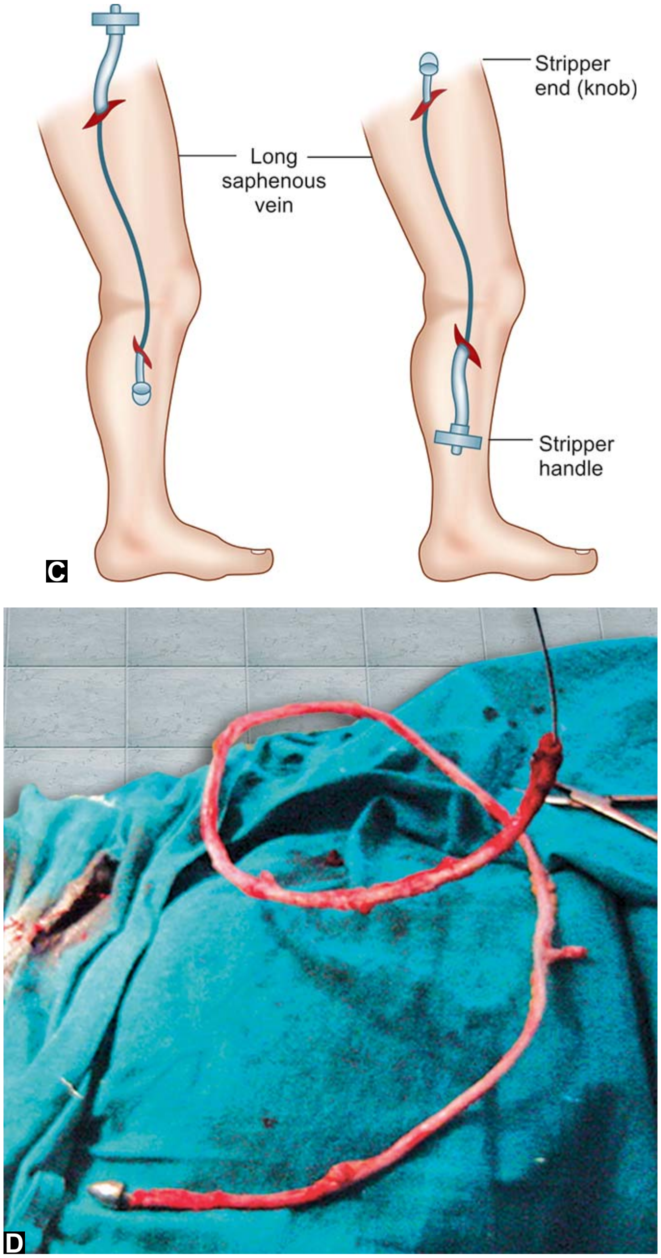

Figs 1.458C and D

Figs 1.458A to D: Long saphenous vein stripping. Note the stripped vein. Stripping is better than just ligation at the junction.

  - ***'Inverting or invaginating stripping'*** using rigid *Oesch pin stripper* is better as postoperative pain and haematoma is less common and also tissue damage. Vein should be very firmly fixed to the end of the stripper and pulled out to cause the inverting of the vein.
  - *Stripping of short saphenous vein is more beneficial than just ligation at saphenopopliteal junction*. It is done from above downwards using a rigid stripper to avoid injury to sural nerve.
- ***Subfascial ligation of Cockett and Dodd*:**
  - Perforators are marked out by Fegan's method. Perforators are ligated deep to the deep fascia through incisions in anteromedial side of the leg.

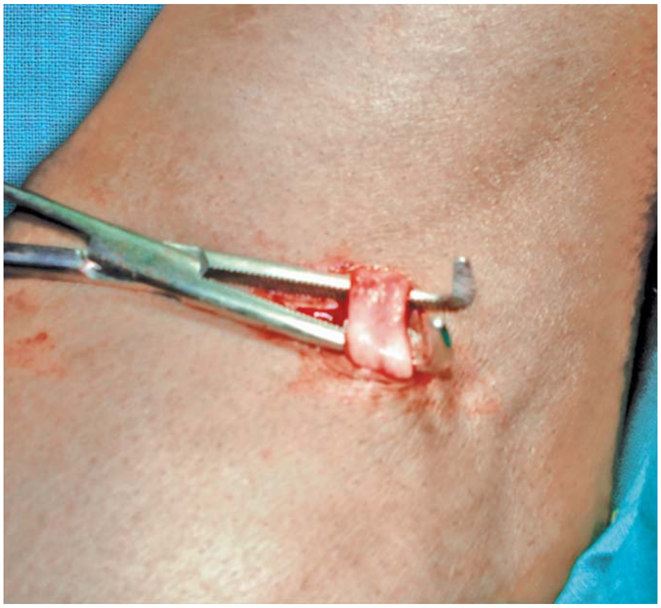

Fig. 1.459: Typical perforator ligation.

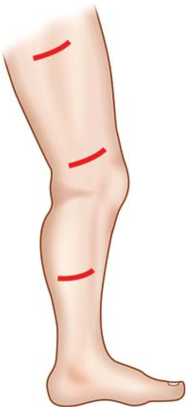

Fig. 1.460: Cockett and Dodd subfascial ligation of perforators using multiple small horizontal incisions.

- ***Ligation of short saphenous vein*** at saphenopopliteal junction. It is done in prone position with horizontal incision. Variations in SP junction are common. But stripping is better.
- ***Linton's vertical approach***—subfascial ligation of perforators.
- ***Stab avulsion of varicose vein and perforators (Fig. 1.461):*** Avulsion is done using mosquito forceps or avulsion hooks—*hook phlebectomy*. It is popular method, also used along with other minimal invasive techniques like EVLA, RFA. Multiple incisions are made and veins are carefully and gently avulsed/teased to clear it. Postoperative compression bandage is a must.

♦ **Minimally invasive methods:**

a. Subfascial endoscopic perforator ligation surgery (SEPS)

- A special telescope is introduced deep to deep fascia through a single small vertical incision at proximal leg selecting healthy skin. Potential space between muscle and deep fascia with loose areolar tissue is easy to dissect using endoscope. Technique is done under tourniquet 300 mmHg pressure. Endoscope is advanced down along the medial border of the tibia. Perforators travelling in subfascial plane are identified and fulgurated using bipolar cautery or clips can be applied into the perforators. It is recommended in chronic venous insufficiency (CVI). But its limitation is difficulty in getting 'lift off' skin in cases with severe lipodermatosclerosis to identify the perforators.

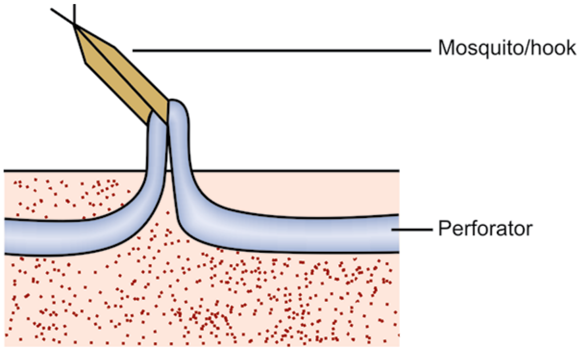

Fig. 1.461: Avulsion of the perforator using hook or mosquito forceps through a small skin incision.

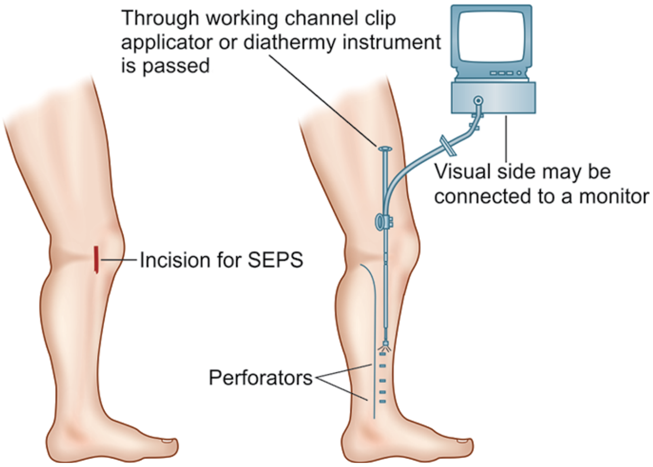
Through working channel clip applicator or diathermy instrument is passed
Visual side may be connected to a monitor
Incision for SEPS
Perforators

Fig. 1.462: Subfascial endoscopic perforator ligation surgery (SEPS).

b. Radiofrequency ablation (RFA) method (VNUS closure method) (VNUS medical technologies Inc; Sunnyvale, CA, USA) (by Goldman 2000):

This procedure is done under general or regional anaesthesia. A RFA catheter is passed into long/ short saphenous vein near saphenofemoral or sapheno-popliteal junction under guidance. 85°C temperature is used for longer period of time to cause endothelial damage, collagen denaturation and venous constriction. Phlebectomy is done while withdrawing the catheter. Wall of the vein is destroyed through its full thickness. Vein forms a cord, which gets dissolved by macrophages and immune cells.

c. ***Trivex method*:**
Under subcutaneous illumination and local anaesthesia, a large quantity of fluid is injected percutaneously to identify the superficial veins under. Tumescent anaesthesia created causes hydrodissection. Trivex resector and illuminator are placed under the skin. Resector gently extracts veins by suction and morcellation. Further stages of tumescence flushes all blood and delivers vasoconstriction solutions. Solution is passed through 18 gauze needle to clear all blood underneath. Method is cosmetically acceptable; removes all sized veins; achieves good pain relief; with minimal complications like bruising, induration which gets resorbed eventually; and can be used when there are venous ulcers.

d. ***Endovenous laser ablation (EVLA):***
It is done as an outpatient procedure or as day-care surgery. Patient lies supine with diseased leg flexed, hip externally rotated and knee flexed. With aseptic precaution, under U/S guidance LSV is cannulated above the knee and a guide wire is passed beyond SFJ and 5- French catheter is passed over guide wire and tip is placed 1 cm distal to the junction. 200 ml of 0.1 % lignocaine (crystalloid with local anaesthetic) is infiltrated along the length of the LSV. Laser fibre is inserted up to the tip of the catheter and catheter is withdrawn for 2 cm and laser fibre protrudes for 2 cm. Laser fibre is fired step by step using diode laser (810-1470 nm diode laser energy), one mm withdrawal in 2 seconds. Once procedure is over catheter is removed and pressure bandage is applied for 2 weeks. Heat produced (729-1000°C at tip) by the laser produces steam bubbles with thermal damage of endothelium leading into occlusion of the vein. Laser energy acts on the blood within the vein rather directly through the wall and heats the blood and in turn heats the vein wall. *Drawback of laser* therapy is inability to create flush occlusion allowing tributaries to open up to cause possible recurrence.
Complications of EVLA: Pain; ecchymosis, haematoma, skin burns, difficulty in cannulating the unsuitable vein if selected; DVT; sensory disturbances, infection.

Note:
EVLA has got 95% efficacy. *Tumescent anaesthesia* is prepared by mixing 5000 ml of normal saline, 30 ml of xylocaine 1% with adrenaline, 30 ml of 8.4% sodium bicarbonate. It is injected using long needle along the length adjacent to vein to cause tamponade and to prevent heat burn on the surface.

e. ***Other methods:***
- *Transilluminated phlebectomy* is done by passing transilluminating light under the skin and passing a rotating blade through another small incision. Veins are grasped and removed by rotating movements.
- *Ambulatory phlebectomy* is done through tiny small incisions using special phlebectomy instruments.
- *Electrodessication* using weak electric current through a fine needle directly into the spider veins (telangiectasis) is also used.

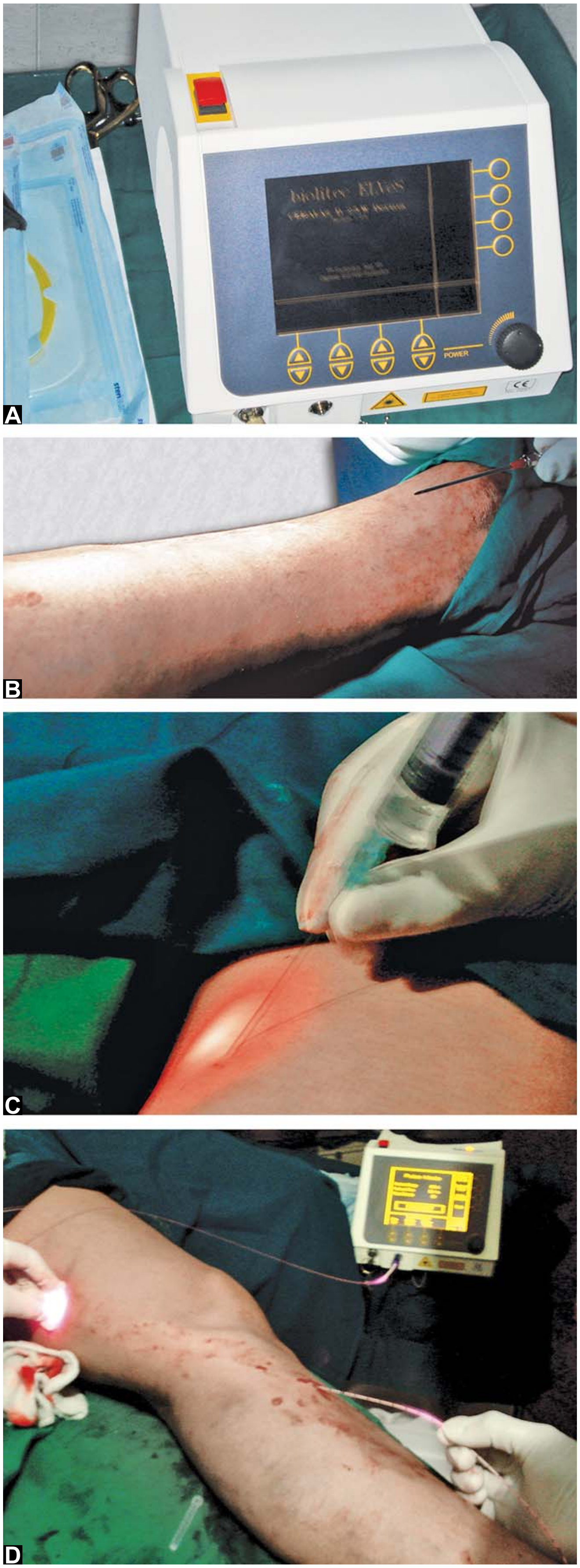

Figs 1.463A to D: Endovenous laser ablation (EVLA) for varicose veins.

Problems in varicose vein surgery

- Infection—10%
- Haematoma formation
- DVT—0.01%
- Saphenous neuralgia, sural nerve injury
- Recurrence

***Note*:**
Contraindication for surgery is deep vein thrombosis (DVT).

VENOUS ULCER (Gravitational Ulcer)

It is the complication of varicose veins or deep vein thrombosis.

Pathogenesis of Venous Ulcer

Varicose veins or DVT which are recanalised, eventually causes *chronic venous hypertension* around ankle

↓

Causes haemosiderin deposition in the subcutaneous plane from lysed RBC's

↓

Eczema

↓

Dermatitis

↓

Lipodermatosclerosis

↓

Fibrosis

↓

Anoxia

↓

Ulceration

Ambulatory venous hypertension is the *prime cause* of venous ulcer formation. Venous hypertension may be *gravitational* which is due to hydrostatic pressure by weight of blood column from the right atrium (*hydrostatic reflux*) which is maximum at foot and ankle **OR** *dynamic* which is due to muscular contraction across the incompetent perforator with a high pressure up to 200 mmHg (*hydrodynamic reflux*). There is a peculiar recycling of blood from deep veins → femoral vein → spillage of blood across incompetent SFJ into LSV / GSV → passage of same blood across perforators into the deep veins to reach femoral vein → again to enter the LSV as spillage.

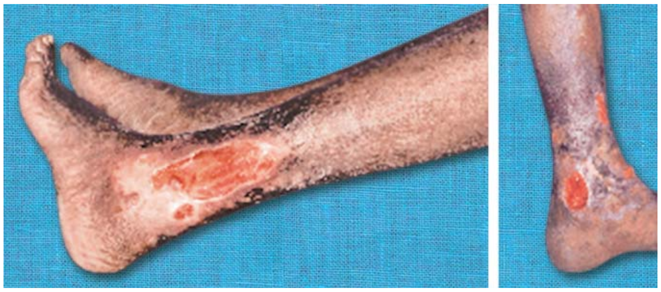

Fig. 1.464: Venous ulcers in the lower limb.

White cell trapping' theory and 'fibrin cuff' theory; release of free radicals; increased matrix metalloproteinases (MMPs); abnormal fibroblast activity; inhibition of growth factors; are other causes of venous ulcer formation.

- *Area where venous ulcer commonly develop, is around and above the medial malleoli* because of presence of large number of perforators which transmit pressure changes directly into superficial system. This area is called as *Gaiter's zone*. It can also be on both malleoli.
- Ulcer is often large, nonhealing, tender, recurrent with secondary infection. Vertical group of inguinal lymph nodes are usually enlarged and tender.

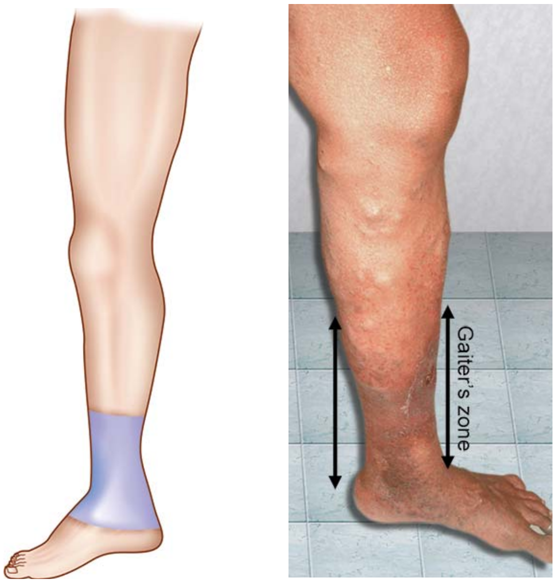

Fig. 1.465: Gaiter's zone. It is handbreadth area around malleoli where complications of venous disease occurs. *Word gaiter* (French) is a leather/cloth covering for lower leg and ankle.

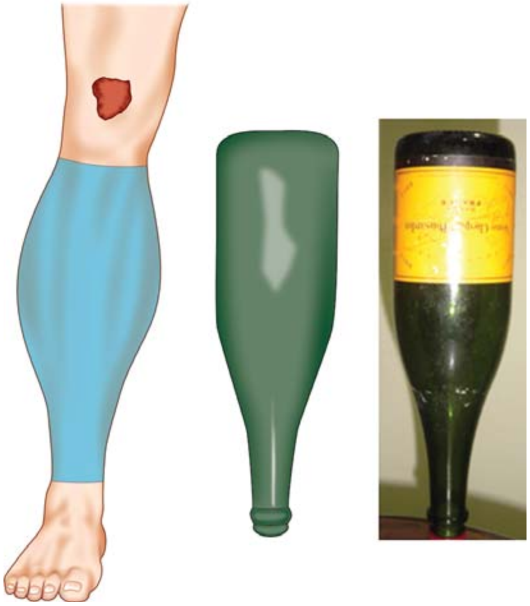

Fig. 1.466: *Champagne bottle sign/inverted beer bottle sign* is seen in lipodermatosclerosis due to prominent calf with narrow ankle contracted skin and subcutaneous tissue. Sign is often observed in DVT also.

Everything has its beauty but not everyone sees it.

- Often it leads to scarring, ankylosis, Marjolin's ulcer formation. Slough from the ulcer bed may give way causing venous haemorrhage.
- *Periostitis is common* which also prevents ulcer from healing.
- Most of the venous ulcers have surrounding lipodermatosclerosis. *Lipodermatosclerosis is pigmentation, thickening, chronic inflammation and induration of the skin in calf and around ankle.*
- Due to regular walking on toes so as to relieve the pain causes contraction and extra-articular fibrosis of achilles tendon. Proper exercise is the remedy for—*talipes equinovarus.*

Note:
70-80% of leg ulcers are venous ulcers.

Differential Diagnosis

- Ischaemic ulcer, diabetic ulcer.
- Rheumatoid ulcer.
- Traumatic ulcer.
- Neuropathic ulcer.
- Neoplastic ulcer.

Investigations

- Discharge from the ulcer for culture and sensitivity.
- X-ray of the area to look for periostitis.
- Biopsy from the ulcer edge to rule out Marjolin's ulcer.
- Investigations to rule out other causes of leg ulcers like arterial; neurological; diabetes; sickle cell disease and other haemolytic diseases.
- Erythrocyte sedimentation rate; C-reactive protein, peripheral smear; red cell counts.
- Doppler—venous and often arterial.

Treatment

- ***Bisgaard method*** of treating venous ulcer:
  - Measures to reduce oedema, increase venous drainage, so as to promote ulcer healing.
    - Elevation.
    - Massage of the indurated area and whole calf.
    - Passive and active exercise.
  - Care of ulcer by regular cleaning with povidone iodine, H_2O_2.
  - Dressing with EUSOL.
  - Four layer bandage (45 mmHg pressure) technique to achieve high compression pressure. It is changed once a week.
  - Antibiotics depending on culture and sensitivity of the discharge.

Different bandages used are:

- *Charing*—Cross (hospital London) elastic multilayered compression bandage, once a week
- Low compression bandaging
- Graduated elastic compression stockings with a leak proof absorbent dressing beneath
- Unna boots
- Crepe bandage/stockings

- Once ulcer bed granulates well, split skin graft (SSG) is placed *(Thiersch Graft),* or pinch graft.
- ***Specific treatment for varicose veins*** should be undertaken—Trendelenburg operation, stripping of veins, perforator ligation.
- Fifty percent of venous ulcer occurs as a result of recanalisation of DVT, and the leg is commonly called as ***postphlebitic limb (leg).*** It presents with all complications of venous diseases like eczema, ulceration, lipodermatosclerosis and venous ulcers. Here surgery for superficial varicose veins are contraindicated. Venous valve repair ***(Kistner's valvuloplasty)*** or *valve transplantation* or drugs like ***Stanazolol,*** which reduces the fibrous tissue thereby increasing the oxygenation are beneficial.

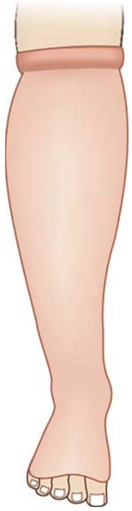

Fig. 1.467: Patient should wear stockings regularly in postoperative period for 3-6 months. It will reduce the recurrence rate and symptoms.

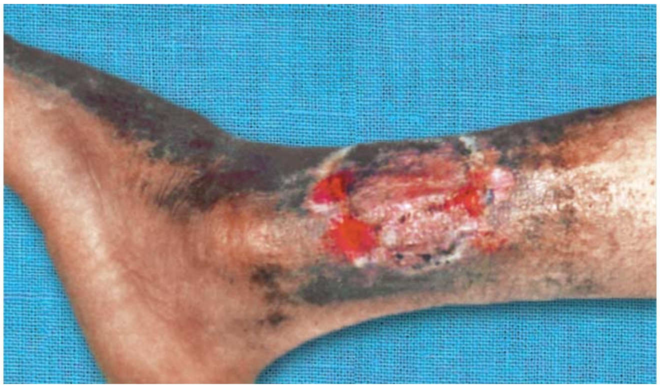

Fig. 1.468: Skin grafting over a venous ulcer after formation of healthy granulation tissue.

Complications of venous ulcers

- Haemorrhage
- Marjolin's ulcer (in unstable scar of long duration)
- Infection
- Talipes equinovarus
- Periostitis is common over the tibia
- Disability
- Calcification

(**EUSOL** is Edinburgh University solution of lime containing *boric acid, sodium hypochlorite, calcium hydroxide.)*

***Note*:**

- Present concept is to treat the ulcer first by compression bandage; regular dressing; skin grafting. Once ulcer has healed definitive procedure for varicose veins is done. Studies show that rapidity of healing of ulcer perse is not dependent on the surgery for varicose veins.
- Recurrence rate of venous ulcer after proper therapy is 30%. Reulcer formation is more in post-phlebitic/ thrombotic limb.

COMPRESSION THERAPY FOR VARICOSE VEINS

- Compression reduces the venous wall tension; prevents reflux; controls the venous over-distension.
- Compression diverts the blood towards deep veins through perforating veins; prevents the outward flow of blood in perforator incompetence; improves the efficacy of calf muscle pump. Compression reduces the oedema and improves the venous and lymphatic drainage; improves venous elasticity; improves the microcirculation and more important is it prevents further damage of the venous wall.
- Compression may be elastic/inelastic/combination of elastic and inelastic (*Unna boot*)/multilayered (four layered) compression system which can provide sustained high compression for several days—usually up to a week/intermittent pneumatic compression. *Unna boot* is three-layered paste gauze compression dressing containing calamine, zinc oxide, glycerin, sorbitol, gelatin and aluminium silicate which has mainly inelastic inner component with partly elastic outer layer wrap.

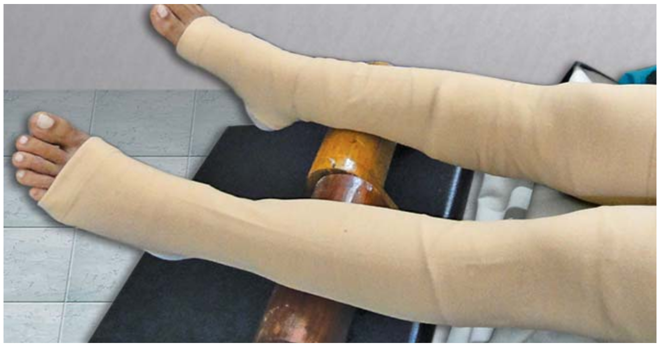

Fig. 1.469: Compression stockings should be worn in varicose vein disease even after intervention to reduce the chances of recurrence.

- Recommended pressure in mild varicose veins, pregnancy and postoperative period is 20 mmHg; in symptomatic varicose veins and after sclerotherapy is 30 mmHg; in venous ulcer and post-phlebitic leg it is 40-45 mmHg; in phlebolymphoedema (venous oedema with lymphatic oedema which is more dangerous) it is more than 45 mmHg.
- Skin maceration, excoriation, dryness, infection, ulceration and failure are the complications of compression therapy.

Remember

- Compression improves ulcer healing rate
- Supplementation of phlebotonic drugs in conjunction with compression therapy is accepted
- Inelastic compression causes more improvement (reduction) in venous filling index (VFI) than elastic compression. But elastic compression is more commonly used because it is better accepted
- Multilayered bandage system is most effective
- Ankle-arm pressure index less than 0.6 is contraindication for compression therapy as it may precipitate ischaemic ulcer formation
- Exercise may improve the muscle—pump action

Compression garments standard

British standard

- Class I: 14-17 mmHg
- Class II: 17-24 mmHg
- Class III: 24-35 mmHg

International (European) standard

- Class I: 20-30 mmHg
- Class II: 30-40 mmHg
- Class III: 40-50 mmHg
- Class IV: 50-60 mmHg

Compression Bandages

- *Type I*: Light weight confirming stretch bandages. These comprise light weight elastomer with high elasticity but little power. It is used to retain dressings.
- *Type II* (short stop): Light support bandages. Minimal stretch. Exhibit limited elasticity but tends to lock out on minimal extension. In ambulant patient CVI they form an essentially inelastic covering to the leg which will exert pressure during calf systole but not during diastole. They are unsuitable for control of oedema.
- *Type III* (Long stop): These are extensible elastic and powerful to a varying degree.

THROMBOPHLEBITIS

It is the inflammation of veins, usually superficial veins due to different causes. It is actually superficial vein thrombosis with inflammation (slight).

Attitudes are more important than facts.

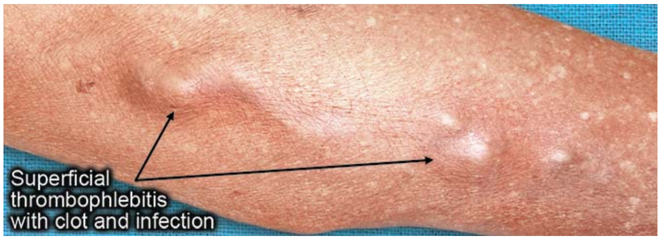

Fig. 1.470: Superficial thrombophlebitis with clot and infection.

Types

a. *Acute***:** Due to IV cannulation, trauma, minor infections, hypercoagulability.
b. Recurrent.
c. *Spontaneous:* Polycythaemia vera, polyarteritis, Buerger's disease.
d. *Thrombophlebitis migrans (Trousseau's sign, 1876*): It is spontaneous migrating thrombophlebitis seen in visceral malignancy like pancreas, stomach.
e. Mondor's disease.

Clinical Features

- Pain, redness, tenderness, cord like thickening of veins, fever.
- It can be seen either in upper limb or lower limb.

Complications

- Destruction of venous valves resulting in varicose veins.
- DVT, embolism, infection.

Treatment

- Elevation.
- Anti-inflammatory drugs, antibiotics.
- Application of crepe bandage.

KLIPPEL-TRENAUNY SYNDROME

It is a nonfamilial mesodermal anomaly with skin naevus, varicose veins, soft tissue and bone hypertrophy. Deep veins are often aplastic. It is usually managed with compression bandages. If patient is undergoing surgery for some other condition, then LMWH should be started. Condition itself occasionally can be treated with EVLA for varicose veins if only deep veins are normal; bone length discrepancy correction of leg is done.

Parkes Weber syndrome is a differential diagnosis. PW syndrome presents with varicose veins, multiple AV fistulas, chronic venous hypertension, high output cardiac failure and ulceration.

ANTICOAGULANTS

These are the agents used to prevent and treat thrombosis and thromboembolic events.

HEPARIN

- It is a natural anticoagulant, a mucopolysaccharide.
- It prevents clotting of blood both *in vivo* and *in vitro* by acting on all three stages of coagulation. It *prolongs clotting time* and *activated thromboplastin time* in specific (by 1.5-2.0 times the control).
- Heparin also causes hyperkalaemia, thrombocytopenia.
- Commercial heparin is derived from lung and intestinal mucosa of pigs and cattle.
- The onset of action is immediate after administration, lasting for 4 hours.
- It is metabolised in the liver by heparinase.
- It does not cross placental barrier and is not secreted in breast milk.

Indications

- As prophylaxis in major surgeries, postoperative period, puerperium.
- As therapy in DVT.

Dose

- For prophylaxis: 5,000 units/subcutaneously 8th hourly.
- For therapy: 10,000 units/IV 6th or 8th hourly. Later changed to subcutaneous dose.
- In severe cases, 5,000 units to 20,000 units is given daily through IV infusion at a rate of 1,000 units per hour. Daily dose should not exceed 25,000 units.

Heparin should not be given intramuscularly and should not be combined with streptokinase or urokinase. Heparin is not given orally.

Heparin administration should always be monitored with APTT.

Complications

Allergy, bleeding, thrombocytopenia.

Danaparoid is an antifactor Xa, heparinoid, is an anticoagulant used in patients where heparin is contraindicated.

LOW MOLECULAR WEIGHT HEPARIN (LMWH)

It is a commercially prepared heparin with a molecular weight of 4,000 to 6,500.

- Enoxaparin.
- Dalteparin.
- Parnaparin.
- Reviparin.
- Fraxiparin.

Advantages

- Have a longer duration of action—once a day
- Have a better anticoagulant effect
- Less interaction with platelets
- Less antigenic
- Usage is easier and more acceptable
- Monitoring is not necessary

Disadvantages

They are expensive.

Presently LMWH are becoming very popular.

Heparin antagonist: 50 mg of 1% protamine sulphate solution is given slow intravenous. 1 gm reverses 100 units of heparin. It is given only after doing activated thromboplastin time. Overdosing or infusion without indication may itself precipitate bleeding.

ORAL ANTICOAGULANTS

They are given orally and are slow acting.

Types

- *Coumarin derivatives*: Bishydroxycoumarin (Dicoumarol): First coumarin drug derived from sweet clover.
  Warfarin sodium: Most common oral anticoagulant used.
- *Indandione derivative*: Phenindione, anisindione.

Mode of Action of Oral Anticoagulant Therapy

- By suppressing synthesis of prothrombin, factors VII, IX and X.
- By inhibiting vitamin K mediated carboxylation of glutamic acid.
- Oral anticoagulant does not have *in vitro* action.
- They are slow acting, and long acting.
- Control of oral anticoagulant therapy is by *monitoring prothrombin time*.
- PT comes to normal only 7 days after cessation of the drug.
- *They cross placental barrier and are known to cause teratogenicity when given in 1st trimester.*
- *They are secreted in breast milk.*

Indications

- In DVT after cessation of heparin for maintenance therapy.
- After valve replacement surgery.

To achieve adequate anticoagulant effect and to prevent thromboembolic episodes the *INR* has to be maintained within 2-3.

Side Effects

- Bleeding—it may require blood transfusion/FFP or vitamin K injection intramuscular or oral to control.
- Cutaneous gangrene.
- Fetal haemorrhage and teratogenicity.
- Alopecia, urticaria, dermatitis.
- Drug interactions: with NSAIDS, cimetidine, omeprazole, metronidazole, cotrimoxazole, erythromycins, barbiturates, rifampicin, griseofulvin.

WARFARIN

WARFARIN (**W**iskonian **A**lumini **R**esearch **F**oundation + coum**ARIN** derivative) *SODIUM* is the most common drug used. It has got lesser side effects. It has got cumulative action and so given in tapering dose.

Dose is 5 mg, once a day.

It should be discontinued 7 days before any surgery like tooth extraction and prothrombin time should return to normal level. During surgery, if excess bleeding occurs, fresh frozen plasma may be given.

The effects of warfarin sodium is reversed by injection vitamin K; the dose depends on INR and emergency of reversal (takes 24 times to reverse).

IN VITRO ANTICOAGULANTS: Oxalates, citrates, EDTA (Ethylene diamine tetra-acetic acid).

Thrombolytic agents

- Streptokinase
- Urokinase
- Anistreplase
- Alteplase

Differences between oral anticoagulants and heparin

	Oral anticoagulant	***Heparin***
	Slow acting	Immediate
	Long acting	Short acting
	Only *in vivo* action	Both *in vitro* and *in vivo* action
Monitored by:	Prothrombin time	Partial thromboplastin time
	Crosses the placental barrier	Does not cross the placenta
	Secreted in milk	Not secreted in milk
Administration:	Orally	Intravenously/ subcutaneously

Contraindications for anticoagulant therapy

- Ongoing bleeding
- Recent surgery/invasive procedure
- Severe trauma
- Bleeding tendency (clotting factor deficiency)
- Intracranial haemorrhage
- Pericarditis/pericardial effusion
- Patient prone to fall

DIRECT THROMBIN INHIBITORS

a. Recombinant hirudin and hirudin analogues—derived from leeches, are direct inhibitors of thrombin.

b. Argatroban—synthetic direct thrombin inhibitor.

ANTIPLATELET DRUGS

- Small dose aspirin—inhibits platelet synthesis of thromboxane A2.
- Ticlopidine (125 mg BD)—alters platelet membrane, thereby platelet aggregation.
- Clopidogrel—action similar to ticlopidine.

- Dextran—decreases platelet aggregation.
- Abciximab—glycoprotein IIb/IIIa inhibitors, block platelet aggregation, and platelet adhesion to fibrin.
- Dipyridamole—xanthine oxidase inhibitor.

PULMONARY EMBOLISM

- It is commonly due to lower limb DVT (15% of lower limb DVT). It can also occur after pelvic vein DVT or upper limb DVT (30% of upper limb DVT).
- Chest pain, cough, haemoptysis, dyspnoea are the features.
- Often site of DVT may be asymptomatic. When symptomatic, fever, pain, tense, tender calf, with positive Homan's sign may be evident.
- Massive embolism causes sudden cardiac arrest and death due to pulmonary artery block. Moderate embolism causes *pyramidal wedge shaped* infarcts in lungs.
- Duplex scan of limb, CT angiogram of thorax, pulmonary angiogram (gold standard), X-ray chest, ventilation perfusion scan, ECG, echocardiography are useful investigations.
- *Treatment* is thrombolysis, heparin/LMWH, compression bandage.
- Occasionally surgical removal of clot from pulmonary artery is done if possible.
- *IVC filter* placement is essential in recurrent DVT with anticoagulation, DVT with contraindication for anticoagulation, pulmonary hypertension. *Greenfield IVC filter* is ideal with 95% patency rate. *Complications* are—bleeding, haematoma, migration of filter into pulmonary artery, thrombosis at filter level, IVC perforation.
- *Retrievable IVC filters* are newer method used to prevent long-term filter complications. It is used in young patients who are at risk of DVT and embolism, for short specified period only like—high-risk trauma with orthopaedic injuries, extensive iliofemoral thrombosis, during thrombolytic therapy. *Recovery filter, Gunthur-Tulip filter, OptEase filters* are used. They are deployed through IJV or femoral vein under angiographic or intravascular US guidance. Recovery and Gunthur-Tulip types are recovered through right IJV. OptEase is recovered from right femoral vein. *Complications* are same as nonretrievable IVC filters. Retrieval failure, retrieval site thrombosis and embolism are specific complications.
- DVT prophylaxis is a must in all these patients.

O. Lymphatics

CHAPTER OUTLINE

- Surgical Anatomy
- Lymphangiography
- Isotope Lymphoscintigraphy
- Lymphoedema
- Lymphomas
  - Hodgkin's Lymphoma
  - Non-Hodgkin's Lymphoma
- MALT Lymphomas
- Burkitt's Lymphoma
- Cutaneous T Cell Lymphoma
- Chylous Ascites
- Chylothorax
- Chyluria
- Sarcoidosis

SURGICAL ANATOMY

Primordial lymphatic system begins to develop during 6th week of development adjacent to jugular vein as lymph sacs. Peripheral lymphatic systems develop from these primordial lymph sacs. Lymphatic system has three components. *Terminal lymphatic capillaries,* which have high porosity absorb lymph, macromolecules, cells and microbes from tissues into the system*; lymphatic vessels* which collect and transport lymph; *lymph nodes* which are interposed in the lymphatic pathway filter lymph and maintain immunity of the body. Lymphatic vessels run adjacent to main blood vessels reaching the major lymphatic channels. *Cisterna chyli* is formed in the abdomen, continues as thoracic duct (formed at 9th week of gestation) in the thorax which has got initial main course towards right side of the mediastinum; but later towards left side entering the internal jugular vein at its joining point of the subclavian vein. In the periphery there is hardly any lymphovenous communications. Lymphovenous communications occur at lymph node level; iliac, subclavian and jugular levels. Lymphatics are absent in epidermis, cornea, CNS, cartilage, tendon and muscle.

Great lymph ducts are—the thoracic duct—single; right lymph duct—single; subclavian, bronchomediastinal and jugular trunks on both sides. These ducts contain valves to prevent backflow.

Cisterna chyli is formed by joining of right and left lumbar lymphatic trunks and intestinal lymphatic duct. Lumbar trunks are short lymph vessels arising from para-aortic lymph glands. It receives lymph from lower limb, pelvis and pelvic viscera, kidney, adrenal and deep lymphatics of abdominal wall. Left lumbar trunk is behind the aorta. Intestinal lymph duct arises from preaortic nodes. It joins the cisterna chyli from front. It receives lymph from stomach, intestines, liver (except most convex surface which drains into right lymph duct), spleen and pancreas. Cisterna chyli is a lymph sac lying in front of the L_1 and L_2 vertebrae between aorta and crus of the diaphragm. From its upper end it continues as thoracic duct. *Thoracic duct* passes through the aortic orifice of the diaphragm, runs medial to azygos vein and right of the aorta in posterior mediastinum. In front it is related to oesophagus, diaphragm and pericardium; behind right intercostal arteries, hemiazygos and accessory hemiazygos vein. At the level of 7th thoracic vertebra it crosses towards left side behind the oesophagus obliquely reaching left side at 5th thoracic vertebral level. It runs upwards between left margin of oesophagus, medial part of left pleura, and behind left subclavian artery. In the neck it passes in front of vertebral system (vertebral vessels and sympathetic chain) and behind carotid system (Common carotid artery, internal jugular vein, vagus nerve), crossing scalenus anterior, phrenic nerve, transverse cervical and suprascapular arteries ending as a single vessel at the junction of internal jugular vein and subclavian vein with a valve. *Tributaries of thoracic duct* are—trunk from lateral intercostal nodes from lower six spaces; efferents from posterior mediastinal nodes, lateral intercostal nodes of upper six spaces, left jugular lymph trunk from head and neck region, left subclavian lymph trunk from left upper limb, left bronchomediastinal trunk from left side of the thorax. Single termination of duct is common (77%); but double/triple/quadruple terminations are known to occur. Occasionally it may end in left subclavian vein, left vertebral vein, right internal jugular vein, right subclavian vein. Thoracic duct is 45 cm in length and 5 mm wide (wider at both ends; narrow in the middle).

Right lymph duct is 2.5 cm in length, formed by right jugular, right subclavian and right bronchomediastinal trunks; runs on the scalenus anterior joining the junction of right internal jugular vein and subclavian vein.

There are about total 450-600 lymph nodes in the body. Around 200 in the neck; around 100 in the thorax; around 50-60 in the axilla; around 250 in the abdomen and pelvis; around 50 in the groin area.

Lymphatic Watersheds of Skin

Lymph from the dermis and appendages drain into a plexus in deep fascia which in turn drains into respective lymph nodes. *There are six watershed areas in the body for lymphatic drainage*. One vertical midline divides into right and left. Two horizontal lines on each side divide the area into three zones. First lies above the line of clavicle; second between line of clavicle and line at umbilical level; third below the level of umbilical line. First drains into head and neck lymph nodes; second drains into axillary nodes; third drains into inguinal/

> *In very early oedema, during pinching, there is a resistance that is not present on normal site, owing to thickening of dermis and subcutaneous tissue.*
>
> **—Sidney S Rose**

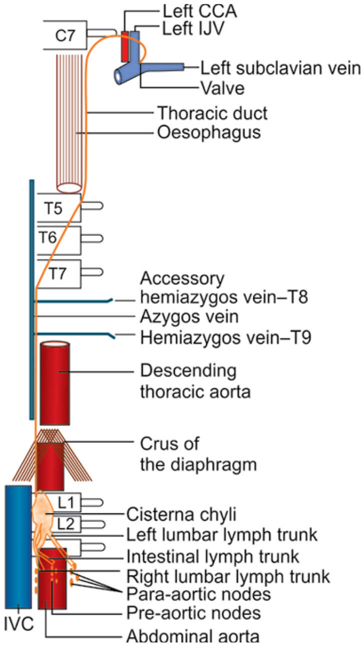
Left CCA
Left IJV
C7
Left subclavian vein
Valve
Thoracic duct
Oesophagus
T5
T6
T7
Accessory hemiazygos vein–T8
Azygos vein
Hemiazygos vein–T9
Descending thoracic aorta
Crus of the diaphragm
L1
L2
Cisterna chyli
Left lumbar lymph trunk
Intestinal lymph trunk
Right lumbar lymph trunk
Para-aortic nodes
Pre-aortic nodes
IVC
Abdominal aorta

Fig. 1.471: Thoracic duct anatomy; cisterna chyli; tributaries of thoracic duct.

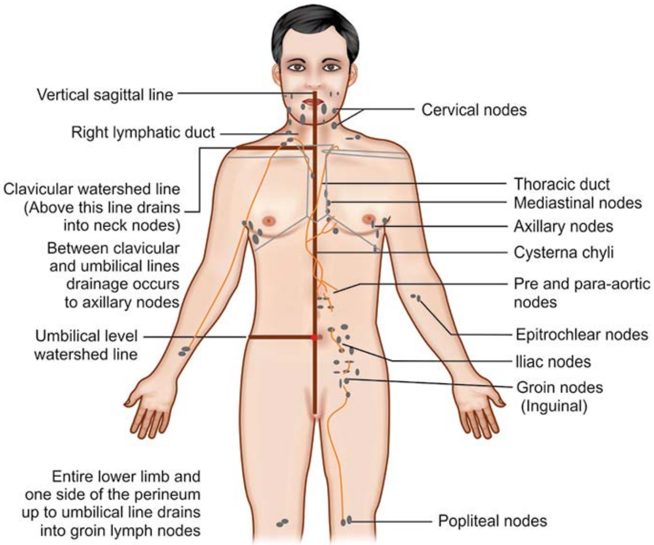
Vertical sagittal line
Right lymphatic duct
Clavicular watershed line (Above this line drains into neck nodes)
Between clavicular and umbilical lines drainage occurs to axillary nodes
Umbilical level watershed line
Entire lower limb and one side of the perineum up to umbilical line drains into groin lymph nodes
Cervical nodes
Thoracic duct
Mediastinal nodes
Axillary nodes
Cysterna chyli
Pre and para-aortic nodes
Epitrochlear nodes
Iliac nodes
Groin nodes (Inguinal)
Popliteal nodes

Fig. 1.472: *Watershed zones/areas of lymphatic drainage*. Vertical sagittal midline; clavicular horizontal line; umbilical line—are used to divide areas into three zones on each side. Above clavicular line drainage occurs to cervical nodes; between clavicular line and umbilical line drainage occurs into axillary nodes; below umbilical line drainage occurs to inguinal lymph nodes. At the level of (lines) drainage can occur to lymph nodes on either side.

groin nodes. Malignancy drains into their respective nodes depending on the location. Lesion on the line can spread to both territory lymph nodes. In skin and appendageal cancers, deep fascia also should be cleared.

Microanatomy of Lymph Node

Lymph node contains three regions—*cortex; paracortex and medulla*. Cortex contains mainly follicles. It may be rounded lymphocytic aggregations of primary follicles or lymphocytic aggregation with germinal centres of secondary follicles due to antigenic stimulation. It contains B lymphocytes, macrophages, dendritic reticulum cells. Germinal centre is surrounded by small B lymphocytes. Both cortex and medulla are associated with humoral immunity. Proliferation of germinal centres suggests active humoral immunity with antibody production. Central medulla contains mainly lymphatic sinuses, arteries and veins, plasma cell and B lymphocytes. Paracortex is located in a zone between cortex and medulla. It contains T lymphocytes, related to cell mediated immunity. Post-capillary venules with high endothelial cells and lymphocytes in the wall are typical. In cell mediated immunity, paracotex expansion occurs. Afferent lymph vessels enter the node through the capsule. It enters the marginal sinus, communicates with intranodal sinus, merging as efferent lymph vessels which enter the hilum. Intranodal sinus lining is highly phagocytic containing littoral cells and sinus lining histiocytes. Main artery and veins pass through the hilum to enter the medulla, paracortex and inner part of cortex. Superficial cortex is supplied by direct capsular vessels.

Function of Lymphatics

Most of the *intravascular proteins are daily filtered through lymphatics* and returned to the circulation again. Macromolecules (albumin, globulin and fibrinogen) and microbes are also filtered at the nodal level as first immune system. From GIT fat is absorbed directly through lymphatics. Lymph shows centripetal flow. Cholesterol, long chain fatty acids, fat soluble vitamins are transferred through lymphatics into cistern chyli directly from GIT bypassing the liver. Transport is mainly due *to intrinsic contractility of the lymphatic vessels* which contain valves for effective forward flow. To a lesser extent other factors like muscle contraction, arterial pressure, thoracic pressure, respiratory movements play role. *8 litres of lymph is produced daily; once it reaches to lymph nodes it is concentrated to 4 litres which* enters the venous circulation. *Protein concentration in lymph is very high (25 grams/litre)*.

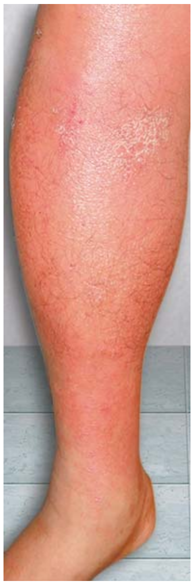

Fig. 1.473: Acute lymphangitis leg. Usually regional lymph nodes are enlarged and tender.

LYMPHANGIOGRAPHY

Indications

- Congenital lymphoedema like aplasia, hypoplasia, hyperplasia.
- Lymphomas show *reticular* pattern. It is also useful to assess the response to treatment.
- Secondaries in lymph nodes, especially iliac and para-aortic lymph nodes.

Technique

Patent blue dye or 1 ml isosulphan blue is injected subcutaneously between toes. Dye is taken up by lymphatics which will be visualised clearly. After making incision, one of the lymphatic vessels is dissected and 30 G needle is passed. Ultra-fluid lipiodol which is an oily contrast medium is injected slowly using pressure pump at a rate of 1 ml in 8 minutes (total quantity is 7 ml). Slowly in 24 hours, it passes through the lymphatics and reaches the iliac and para-aortic lymph nodes. Radiographs taken will help to visualise both lymphatic vessels as well as lymph nodes.

Secondaries in lymph nodes causes filling defects. Lymphomas shows enlarged nodes which have foamy or reticular appearance.

Disadvantages

- Technically difficult.
- Extravasation of dye can occur.
- Dye may not reach the required area.
- Time consuming and invasive procedure.

Lymphangiographic classification of lymphoedema **(Browse classification):** (Norman Browse)

- *Congenital hyperplasia (10%):* Congenital; common in males; entire leg is involved; one or both sides; with often family history; progressive; involves increased number of lymphatics and lymph nodes associated with chylous ascites, chylothorax and protein losing enteropathy
- *Distal obliteration (80%):* Common in females; starts at puberty; calf and ankle are involved; often bilateral; with often family history; slow progression; aplasia/hypoplasia of lymphatics
- *Proximal obliteration (10%):* Occurs at any age; equal in both sexes; leg and thigh involved; unilateral; rapid progression; proximal aortoiliac nodal block

ISOTOPE LYMPHOSCINTIGRAPHY

- Radioactive technetium labelled sulphide colloid particles, or radioiodinated human albumin are injected into the web space using fine needle. These particles are specifically taken up by lymphatics.
- Using gamma camera, limb and inguinal region is exposed to visualise the lymphatics and inguinal lymph nodes.
- Radioactivity in inguinal nodes is measured at 30 and 60 minutes. Normal uptake is 0.6-1.6%; if it is < 0.3% in 30 minutes it is diagnostic of lymphoedema. If it is > 2% it suggests rapid abnormal clearance due to oedema as the result of venous disease.
- In 3 hours, it reaches the para-aortic lymph nodes, other abdominal lymph nodes and liver.
- Later thoracic duct also can be visualised. It can be compared to the take up on the other limb.

Advantages

- It is more sensitive.
- Technically easier and faster compared to lymphangiography.
- Thoracic duct, other lymph nodes and liver can be visualised.
- It is the test of choice. It is simple and safe.
- It has 90% sensitivity; 100% specificity.

LYMPHOEDEMA

It is accumulation of fluid (lymph) in extracellular and extravascular fluid compartment, commonly in subcutaneous tissue. It is primarily due to defective lymphatic drainage.

It is *increased* protein rich interstitial fluid.

Classification

Kinmonth classified lymphoedema as:

- Primary without any identifiable lymphatic disease.
- Secondary is acquired due to definitive cause. ***Most common form***.

Primary type

It affects commonly females.

- It is common in lower limb and left side.

It can be:

- Familial
- Syndromic (Turner's / Klinefelter's / Down's / Klipple Trenauny Weber)

It can be:

- Lymphoedema congenital (10%) — Present at birth—< 2 years — Familial type is called as ***Nonne-Milroy's disease.*** It is type I familial, autosomal dominant—chromosome 5 related; bilateral upper and lower limbs, genitalia and face may be involved. Incidence is 1:6000 of live births.
- Lymphoedema praecox — Present at puberty—up to 2-35 years (80%). — Familial type is called as ***Letessier-Meige's syndrome.*** It is type II familial. It occurs between puberty and middle age.
- Lymphoedema tarda — Present in adult life—after 35 years

It can be radiologically (lymphangiography):

- Hypoplasia 70%
- Aplasia 15%
- Hyperplasia (varicose lymphatics) 15%

Don't consume your tomorrow's feeling on your yesterday.

Pathophysiology of Lymphoedema

Decreased lymphatic contractility, lymphatic valvular insufficiency, lymphatic obliteration by infection, tumour or surgery causes all effects and pathology of lymphoedema. This leads to lymphatic hypertension and dilatation causing lymph stasis, accumulation of proteins, glycosamines, growth factors, and bacteria. There is more collagen formation, deposition of proteins, fibroblasts, ground substance causing fibrosis in subcutaneous and outside deep fascia. Muscles are normal *without* any oedema but may get *hypertrophied*.

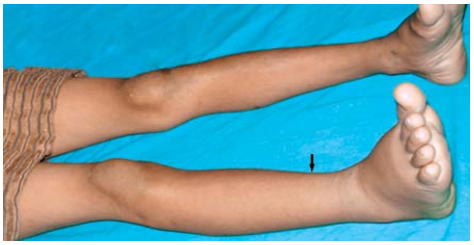

Fig. 1.474: Right side congenital limphoedema in a girl.

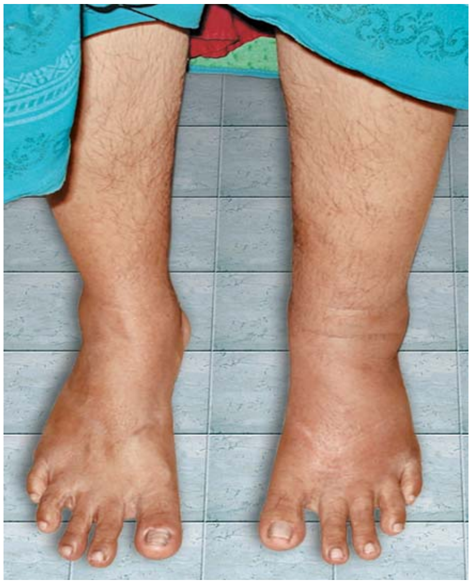

Fig. 1.475: Early lymphoedema left side—pitting type.

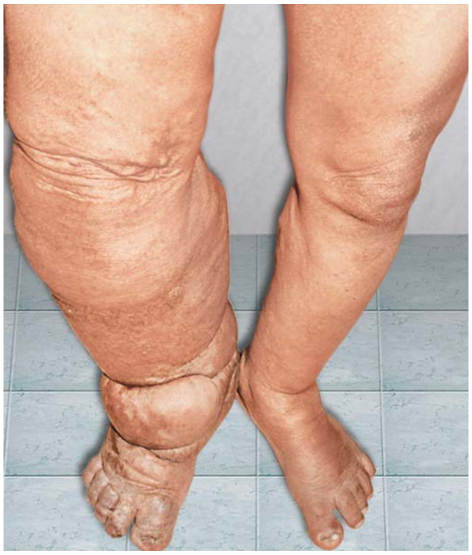

Fig. 1.476: Late lymphoedema—grade 2.

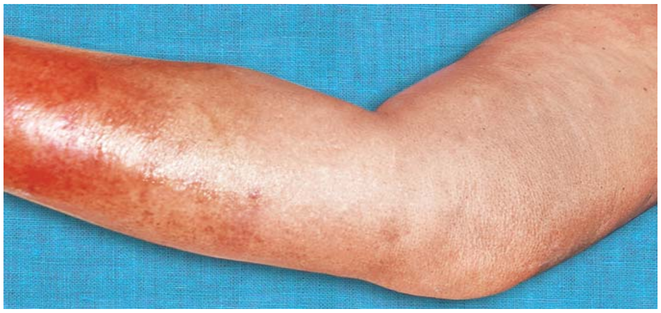

Fig. 1.477: Lymphoedema leg extending into the thigh with lymphangitis.

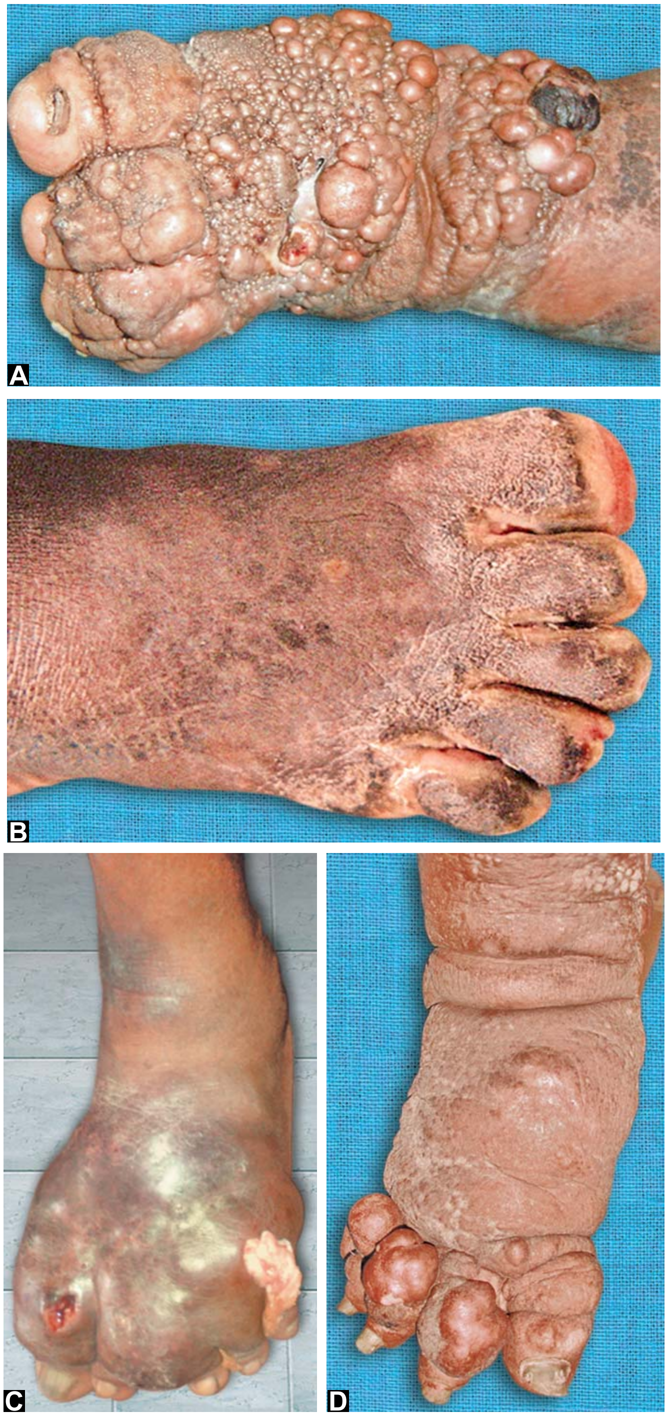

Figs 1.478A to D: Lymphoedema foot in different patients—severe with vesicles/oedema/skin changes/fissures.

Causes of secondary lymphoedema

- Trauma
- Surgery—inguinal block or axillary block dissection/postmastectomy with axillary clearance
- Filarial lymphoedema due to *Wuchereria bancrofti*—common cause in coastal region
- Tuberculosis
- Syphilis
- Fungal infection
- Advanced malignancy—hard, fixed lymph nodes in axilla or in inguinal region
- Postradiotherapy lymphoedema
- Bacterial infection
- Rare causes: Rheumatoid arthritis, snake and insect bites, DVT, chronic venous insufficiency

Note:
Secondary lymphoedema develops rapidly.

Filariasis

It is caused by a parasite *Wuchereria (Brazil) bancrofti (Australia*). It was also called as Malabar leg in 1709, by Clarke, Cochin. Female adult worm is longer 7-10 cm than male worm (4 cm). *Microfilaria* is colourless, translucent, 300 µ length and 10 µ thick. It has head, body and tail. Microfilaria circulates in blood. In India, Asian countries and China they show nocturnal periodicity (from 10 PM to 4 AM). It is related to night biting habits of the vector, *Culex fatigans* mosquito and sleeping habits of the host.

Man is the definitive host; animal or reservoir host is not known. *Female mosquito* is intermediate host (in India and China—Culex fatigans). Development or multiplication of microfilaria will never occur in human blood. Life span of microfilaria in human blood is 3 months. Microfilaria is infective to female mosquito. A density of 15 microfilaria/drop of blood are needed to make it infective.

Life Cycle

Microfilaria from carrier human blood → enters the stomach of female Culex mosquito when it bites human carrying the parasite in blood → ex-sheathing of microfilaria in stomach of mosquito in 6 hours → penetrate the stomach wall → migrate to thoracic muscles of mosquito in 12 hours → metamorphosis into sausage shaped first stage larva in 48 hours → second stage larva in one week → elongated actively motile *third stage infective larva* in one more week (one microfilaria forms one infective larva; microfilaria never multiplies in mosquito nor in human) → enters the proboscis of mosquito to become infective to human during the mosquito bite; it takes 20 days for microfilaria to develop into infective 3rd stage larva in mosquito [*extrinsic incubation period*]→ enters human skin while biting → many larvae get destroyed in human skin by immunity, few enters lymphatics → enters regional lymph nodes in inguinal or axillary or abdominal nodes → develop into adult worm in lymph nodes → mating of female and male worms takes place → gravid female worm releases up to 50,000 microfilariae/day into lymph circulation → thoracic duct → subclavian vein → microfilaria in human circulation → infective to female mosquito. Time from 3rd stage infective larva entering human skin and forming adult worm and later releasing microfilaria into blood is called as *biological incubation period* (12 months); time from 3rd stage infective larva entering human skin to appearance of first clinical feature is called as *clinical incubation period* (16 months).

Effects of Wuchereria Bancrofti Infection

- *Carrier stage* having circulating microfilaria but *asymptomatic.*
- *Immune and allergic reactions* by adult worm causing macrophage and lymphocyte infiltration, endothelial hyperplasia, lymphatic vessel wall thickening, lymph stasis, dilatation, further reaction, fibrosis, further blockage, calcification, recurrent streptococcal infection, filarial lymphoedema.
- *Filarial fever, utricaria, pruritus, epididymo-orchitis as acute* presentation.
- *Occult filariasis* where microfilaria is not demonstrable in blood but identified in lungs (by biopsy confirmation) causing eosinophilia, bronchospasm, nocturnal cough, fever, wheeze, weight loss, arthritis, thrombophlebitis, tenosynovitis.
- *Lymphadenitis, lymphangitis*.
- *Lymphangiovarix, lymphorrhagia, lymph scrotum, lymphocele,* chyluria, chylous diarrhoea, retroperitoneal lymphangitis, chylous ascites, chylothorax.
- *Blood smear—night time (thick and thin)*, lymph node biopsy, skin test, eosinophilia, serological tests, DEC provocation test (by giving 100 mg DEC) *are different investigations*.

Pathology *(Commonly in Filarial Lymphoedema)*

Recurrent lymphangitis causes obliteration of lymph vessels
↓
Dermal lymphatic backflow
↓
Retrograde obliteration (or die back of lymphatics)
↓
Oedema, initially pitting but later nonpitting
↓
Recurrent cellulitis—thickening of skin
↓
Accumulation of proteins, growth factor, glycosaminoglycans
↓
Activation of collagens and keratinocytes
↓
Protein rich lymphoedematous tissue formation
↓
Deposition of ground substance, subdermal fibrosis
↓
Dermal thickening and dermal proliferation, Fissuring → Cracks—Ulceration—Abscess formation
↓
Stout leg with unbearable weight
↓
Elephantiasis.

Rarely it causes protein losing diarrhoea, chylous ascites, chylothorax, chyluria, lymphorrhoea. Recurrent lymphadenitis occurs in the region which aggravate the condition.

Disease in the limb is confined to skin and subcutaneous tissue, i.e. often, only superficial lymphatics are involved by the disease, deep lymphatics are not. *Superficial and deep lymphatics are not communicating* with each other (Unlike the veins in the limb where superficial and deep veins are freely communicating with each other).

Sites of lymphoedema

- Lower limb—most common
- Upper limb
- Scrotum, penis (Ram's horn penis)
- Breast—requires reduction mammoplasty
- Labia
- Eyelid
- Localised lymphoedema

Clinical Features

- Swelling in the foot, extending progressively in the leg—*tree trunk* pattern leg.
- Loss of normal perimalleolar shape—*tree trunk pattern* leg.
- *Buffalo hump* in the dorsum of the foot.
- Squaring of toes.
- Skin over the dorsum of foot cannot be pinched because of subcutaneous fibrosis—*Stemmer's sign.*
- Initially pitting oedema occurs, which later becomes nonpitting.
- Dull ache/severe pain/burning/bursting/cramps; 50% patients will have pain.
- Debility/immobility/obesity/muscle wasting.
- Athlete's foot with joint pain and disability.
- Eczema, fissuring, papillae formation, ulceration, lymph ooze, elephantiasis are other features.
- Fever, malaise, headache.
- Recurrent abscess formation.
- Psychological and social discomfort causing severe morbidity.
- *Endemic elephantiasis/podoconiosis* is common in Africa; seen in barefoot workers; due to destruction of lymphatics by silica derived from soil containing alkaline volcanic rocks.

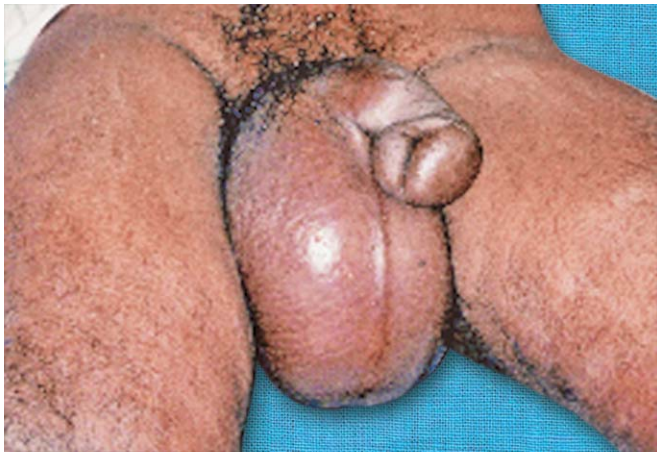

Fig. 1.479: Ram's horn penis.

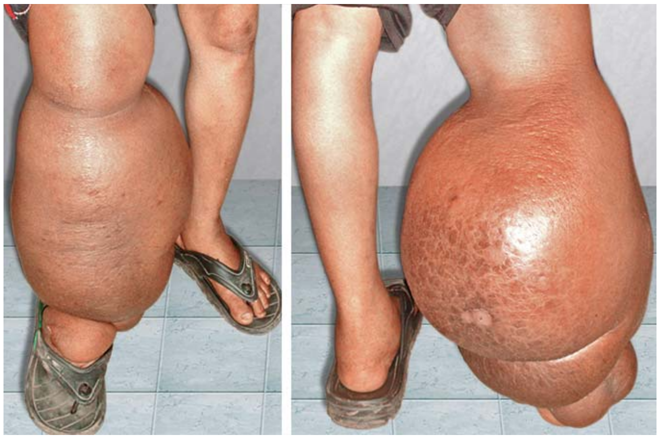

Fig. 1.480: Typical elephantiasis leg with late lymphoedema.

Brunner's grading of lymphoedema

Latent—subclinical: No clinically apparent lymphoedema.

Grade I	Pitting oedema which more or less disappears on elevation of the limb—is due to excess deposition of interstitial fluid
Grade II	Nonpitting oedema occurs which does not reduce on elevation
Grade III	Oedema with irreversible skin changes like fibrosis, papillae, fissuring

Lymphoedema can be:

- *Mild* lymphoedema—< 20% of excess limb volume.
- *Moderate* lymphoedema—20-40%.
- *Severe* lymphoedema—> 40%.

Differential Diagnosis

- Cardiac causes, hypoproteinaemia, malnutrition, nephrotic syndrome, liver failure.
- Myxoedema.
- Trauma.
- Venous diseases like DVT.
- Lipodystrophy and lipoidosis (lipoedema). Lipoedema occurs exclusively in females; begins in puberty; bilateral and symmetrical; trunk may be involved; feet are not involved; not pitting; not related to elevation/compression; MRI shows only fat without fluid.
- Arterial diseases including AV malformations.
- Gigantism.
- Drug induced—steroids, estrogens, nifedipine.
- External compression of veins as caused by abdominal tumours.

Investigations

- For the cause.
- ESR, peripheral smear.
- Lymphangiography.
- Isotope-lymphoscintigraphy.
- MRI, CT scan to identify the cause.

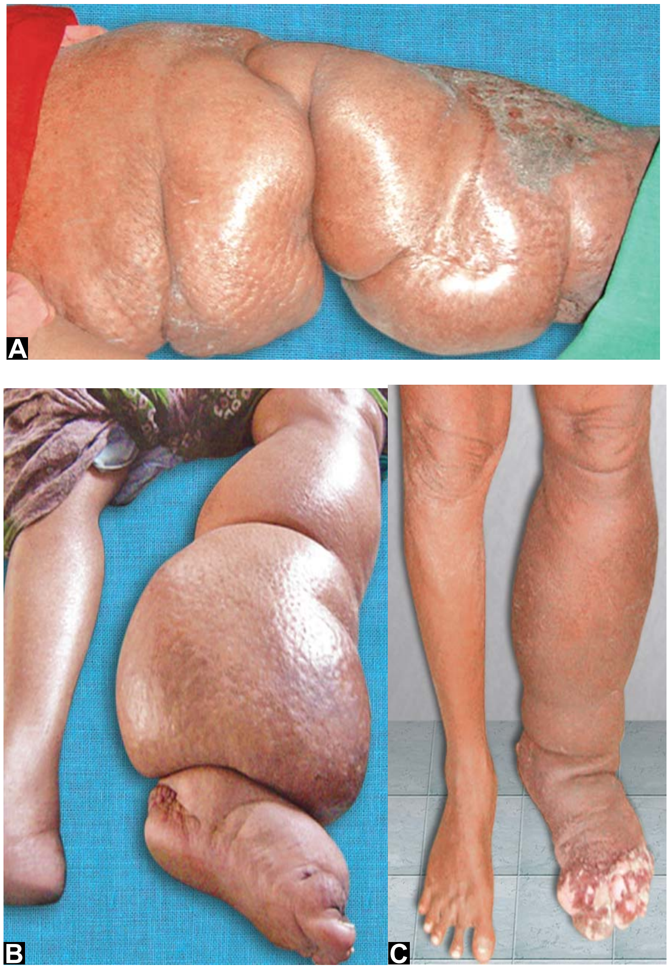

Figs 1.481A to C: Different lymphoedema pictures. Note the oedema, fissuring, cracks. All these make it more vulnerable for infection.

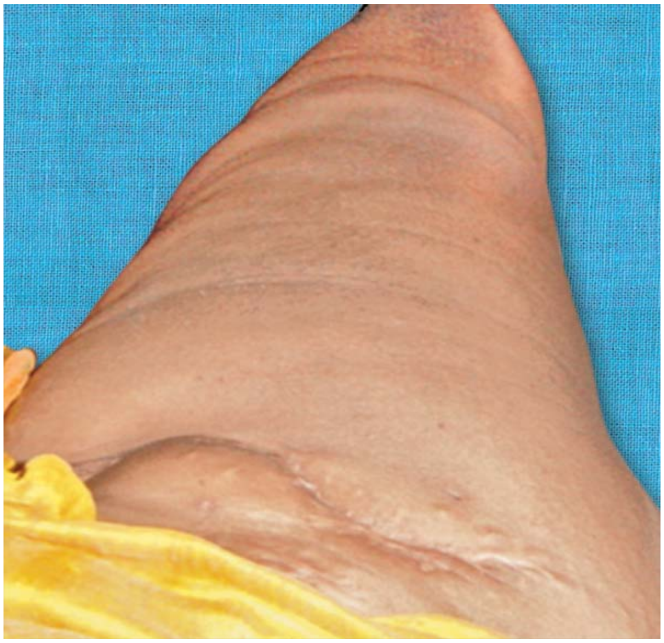

Fig. 1.482: Lymphoedema of lower limb developed after ilioinguinal block dissection for nodal secondaries from melanoma.

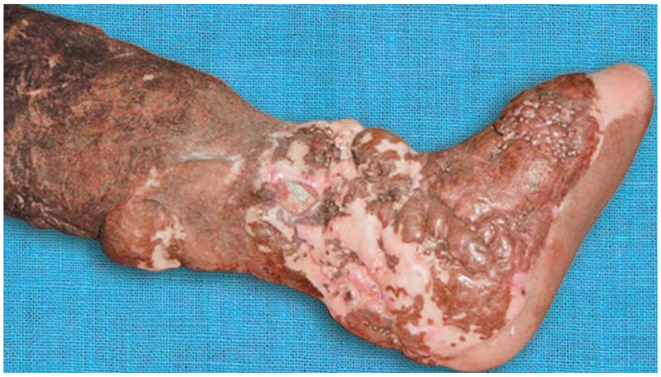

Fig. 1.483: Recurrent filarial leg with nodules and ulceration.

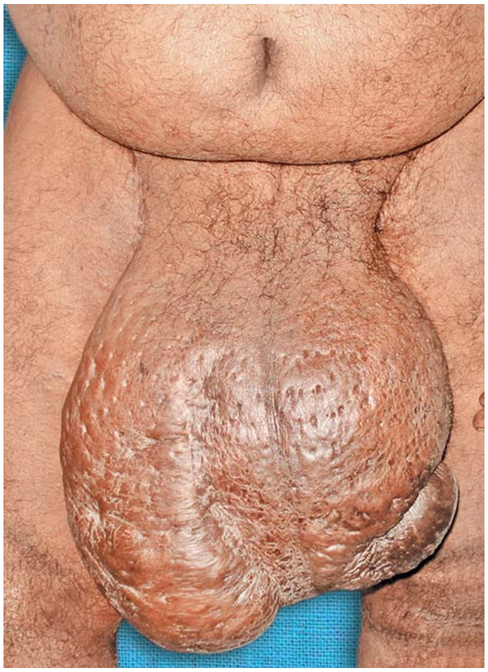

Fig. 1.484: Extensive scrotal lymphoedema of filarial origin. Patient underwent scrotal reduction.

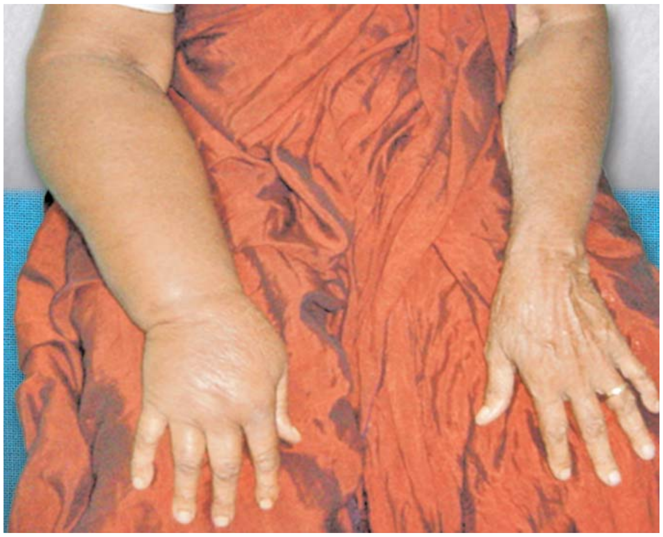

Figs 1.485: Lymphoedema right arm and forearm. Common causes for upper limb lymphoedema are filarial and post-mastectomy with axillary nodal clearance.

Indecisive people are like a blind man looking in a dark room for a black cat that isn't there.

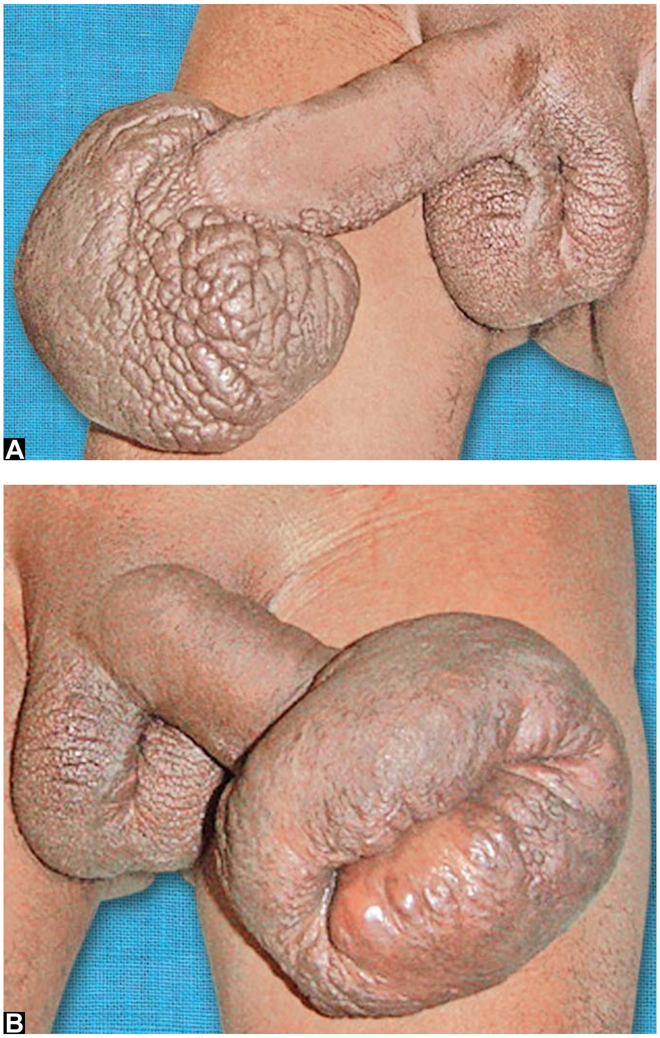

Figs 1.486A and B: Lymphoedema of penis. It needs reduction and reconstruction otherwise it may cause urinary problem and sexual dysfunction.

Complications

- Skin thickening, abscess and maggot's formation.
- Recurrent cellulitis, nonhealing ulcers, septicaemia.

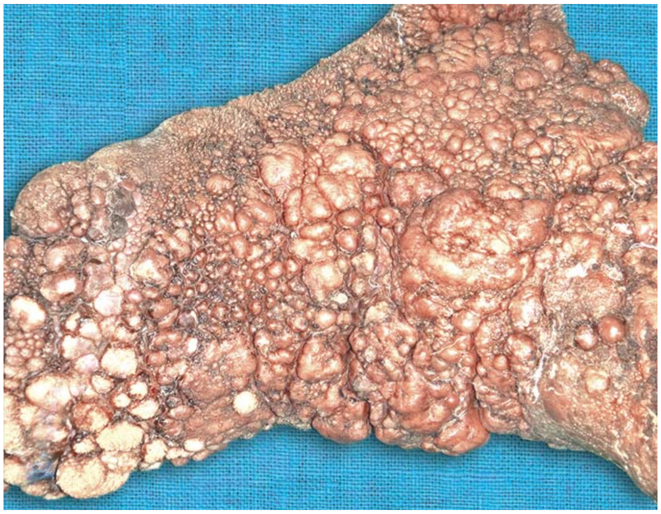

Fig. 1.487: Severe lymphoedema foot with vesicles and thickening.

- Lymphangiosarcoma; *Stewart Treves syndrome* (0.5%, occurs after 10 years), which presents as multiple bluish satellite nodules in the skin of the limb often with ulceration and haemorrhage. Skin/nodule biopsy is confirmative. Chemotherapy, radiotherapy, later even though radical amputation is the treatment, it carries very poor prognosis. This syndrome is usually seen in *upper limb after mastectomy*.
- Recurrent streptococcal infection.

Treatment

Conservative

a. Elevation of the limb, exercise, weight reduction.
b. Static isometric activities like prolonged standing or carrying weights should be avoided; rhythmic isotonic movements like swimming/massaging should be encouraged.
c. Diuretics to reduce the oedema is controversial. It more often causes eletrolyte imbalance than being beneficial.
d. Benzopyrones are protienolytic agents/*lympedim.* They are coumarin (I, 2 benzopyrones) derivatives with no anticoagulant effect but increase the lymphatic peristalsis and pumping mechanism along with proteolysis.
e. Daily wearing of below knee stockings.
f. Avoid trauma and infection.
g. Intermittent pneumatic compression devices (Pressure > 50 mmHg); multilayered lymphoedema bandaging (MLLB)—nonelastic type is preferred method; graded stockings.
h. Antibiotics—flucloxacillin, erythromycin, long acting penicillins.
i. Topical antifungal 1% clotrimazole and systemic griseofulvin 250-1000 mg.
j. Regular washing and keeping the limb clean is very important.
k. Diethyl carbamazine citrate (DEC) 100 mg TID for 3 weeks.
l. Pain relief—by suitable means.
m. Skin care:
    - Keratolytics like salicylic acid 5%; bland emollient; soft/liquid paraffin.
    - Avoidance of skin sensitisers like some soaps.
    - Topical steroids.
    - Control of allergy.
    - Control of fungal infection by drugs like fluconazole.
    - 3% benzoic acid ointment to prevent Athlete's foot.
    - Control of lymphorrhoea.
    - Prevention/control of skin infections.
n. **Complex decongestive therapy** is a comprehensive two phase program of elevation, exercise, massaging, and compression wraps. First phase is *intensive therapy* and second phase is *maintenance therapy*.

    Compression wraps may be high stretch wraps or low stretch wraps. Low stretch wraps are better accepted. It should be worn initially for 24 hours. Compression wraps are used in initial intensive phase of therapy. *Graduated elastic compression garments* are used in maintenance phase which provides maximum pressure of 50 mmHg at the distal part with gradual reduction of pressure in proximal portion.

    Manual lymphatic drainage is a specialised technique to stimulate the contractility of lymph collecting vessels and enhance fluid and protein transport by gentle, light, superficial massaging of the skin so as to open up new lymphatic vessels. Technique is done first on the opposite normal side; then trunk, same side trunk; same side proximal; same side distal and later same side distal to proximal fashion, so as to redirect the lymph towards functioning lymphatic territories.

Surgery

Surgeries for lymphoedema has been classified as:

a. *Excisional*
  - Charle's operation.
  - Homan's operation.

b. *Physiological*
  - Omentoplasty.
  - Nodovenous shunt (Neibulowitz).
  - Lymphovenous shunt (O'Brien's).
  - Ileal mucosal patch.

  Here either communication between superficial and deep lymphatics are created or new lymphatic channels are mobilised to the site.

  Omentoplasty (Omental pedicle): As omentum contains plenty of lymphatics, omental transfer with pedicle will facilitate lymph drainage.

c. *Combined:* Both excision + creation of communication between superficial and deep lymphatics.
  - Sistrunk operation.
  - Thompson's operation.
  - Kondolean's operation.

d. *Bypass procedure*:
  - *Handley's* (1908) silk threads/nylon threads/perforated polyethylene tubes placement as burial from ankle to mid-abdominal level, kept for one year. Procedure is only of historical interest.
  - Skin bridge across the thigh and abdomen (Gillies).
  - Nodovenous shunt.
  - Lymphovenous shunt using microscope.
  - Ileal mucosal patch (Kinmonth). Segment of ileum with pedicle is isolated and opened to expose the mucosa; mucosa is denuded and this mucosa is placed in the thigh as burial to communicate with lymphatics to drain into abdominal lymphatics across ileum.
  - Baumeister lymphatic grafting.
  - Autotransplantation of free lymphatic flap from opposite side—done in post-mastectomy lymphoedema (Trevidic and Cormier).

e. *Limb reduction surgeries*:
  - ***Sistrunk operation:*** Along with excision of lymphoedematous tissue, window cuts in deep fascia is done, so as to allow communication into normal deep lymphatics.
  - ***Homan's operation:*** Excision of lymphoedematous tissue is done after raising skin flaps. Later skin flaps are trimmed to required size and sutured primarily. Medial and lateral sides of the limb are done at separate sittings with 6 months interval.
  - ***Thompson's operation:*** Lymphoedematous tissue is excised under the skin flaps. Epidermis and part of

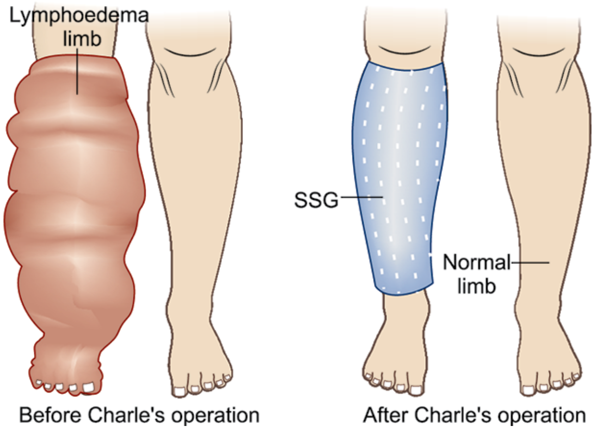

Before Charle's operation

After Charle's operation

Fig. 1.488: *Charle's excisional surgery*. Here after excising lymphoedematous tissue, area is covered with skin graft.

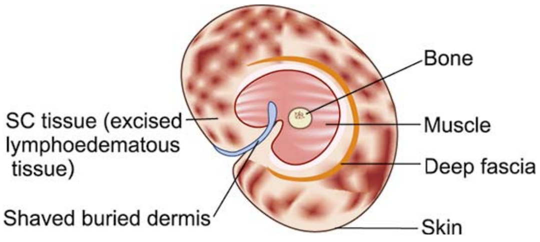

Fig. 1.489: *Thompson's Swiss-roll operation*. Here after removal of lymphoedematous tissue, deep fascia is opened to expose the muscle. Epidermis abraded using skin graft knife. This shaved dermis is buried into the muscle to get communication into the deeper lymphatics.

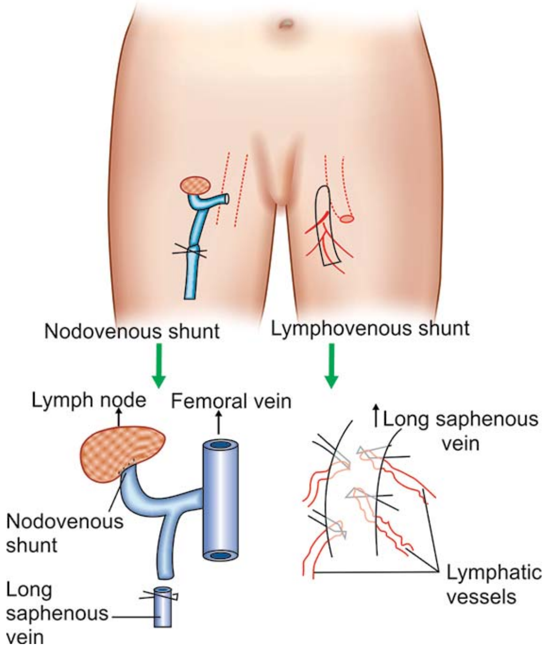

Fig. 1.490: Diagram showing right side groin nodovenous shunt between iguinal lymph node and long saphenous vein. Left side showing lymphovenous shunts between dilated lymphatics and long saphenous vein. At least 4 lymphatics should be anastomosed using 7 zero/11 zero prolene—using operative microscope.

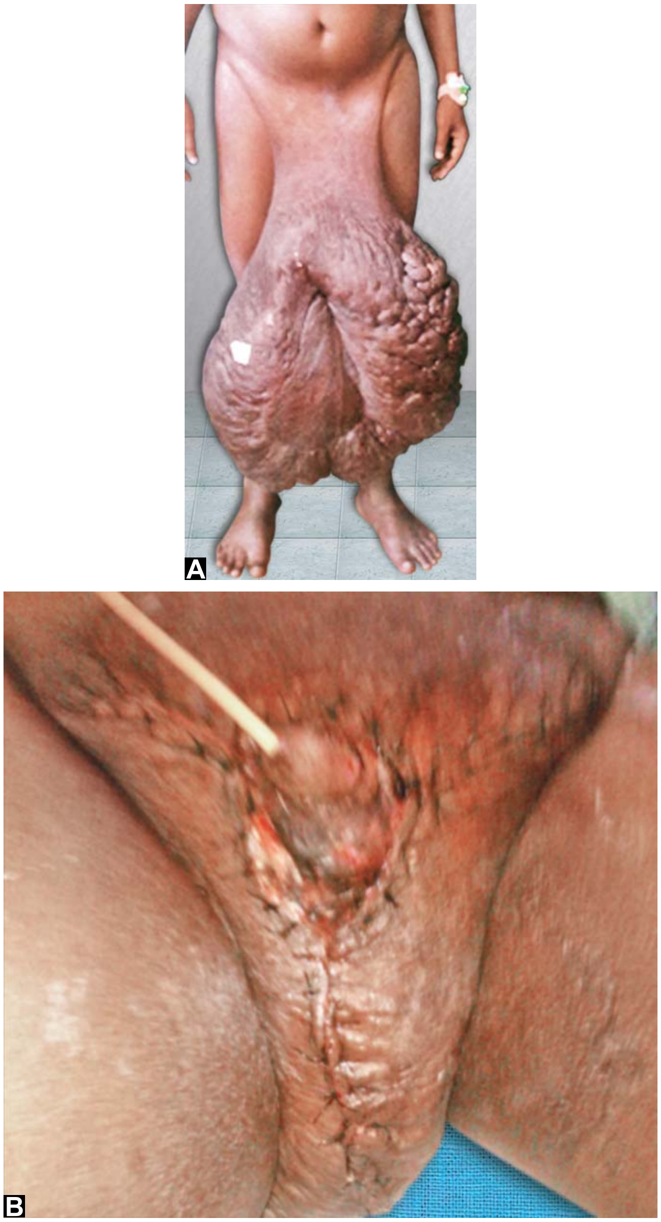

Figs 1.491A and B: *Believe it or not!* Severe scrotal lymphoedema reaching almost up to feet; after surgical excision it weighed 40 kg; postoperatively patient went home with – 40 kg weight (*Courtesy:* Professor Shivananda Prabhu, MS, KMC, Mangalore).

the dermis of one of the skin flaps is shaved off using Humby's knife. It is buried under opposite flap, deep to the deep fascia like a *swiss roll (Swiss roll operation* or *buried dermal flap operation).*

Problems here are formation of epidermal cysts and sinus.

- ***Kondolean's operation:*** Along with excision of lymphoedematous tissue, vertical strips of deep fascia is removed so as to open the deep lymphatics which creates communication between superficial and deep lymphatics.
- ***Macey's operation:*** Here skin and subcutaneous tissue are peeled back with deep fascia and split skin grafting is done over the denuded area. Overlying pad of tissue is sutured back temporarily and after 10 days, it is trimmed away.
- ***Miller's procedure*:** It is excision of subcutaneous tissues under the skin flap with deep fascia in two stages. First stage is done over the medial aspect of the limb; second stage done after two months over lateral aspect of the limb.
- ***Charle's (1912) operation*:** Done in severe lymphoedema with elephantiasis. Along with excision of lymphoedematous tissue, skin grafting is done. It reduces the size and weight of the limb. Patient becomes ambulatory. Wound sepsis, graft failure, dermatitis, hyperkeratosis are the complications.
- Reduction surgeries are done for lymphoedema of scrotum, penis, labia and eyelid.

In severe type, occasionally amputation may be required.

LYMPHOMAS

They are progressive neoplastic condition of lymphoreticular system arising from stem cells.

Lymphomas are the 3rd most common malignancy among children comprising 15% of paediatric cancers.

Aetiology

- Genetic predisposition.
- Sjogren's syndrome—30 fold increase of NHL.
- HIV infection.
- Wiskott—Aldrich syndrome.
- Ataxia—telangiectasia.
- Bloom's syndrome.
- Virus etiology—Epstein-Barr virus.
- Celiac sprue—intestinal T cell lymphoma.
- *H. pylori* may be associated with MALT lymphoma.
- Occupation causes—hair dye workers; herbicide exposure.
- Ionising radiation.
- Smoking; alcohol consumption; tobacco usage.
- Lymphomas are more common in western countries than in Asia.

Types

- *Hodgkin's lymphoma (HL).*
- *Non-Hodgkin's lymphoma (NHL).*

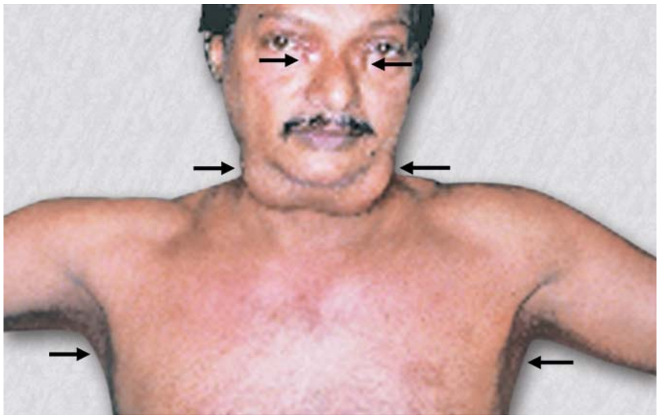

Fig. 1.492: Enlarged, B/L, rubbery neck LNs of lymphoma. Note the involvement of lacrimal sac.

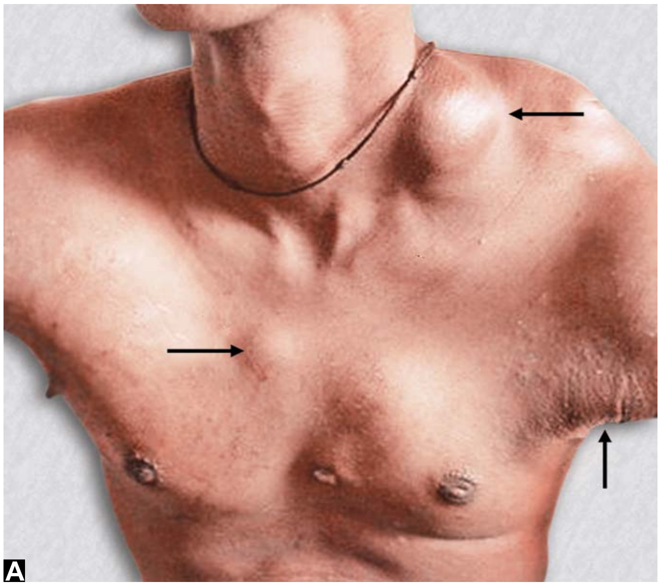

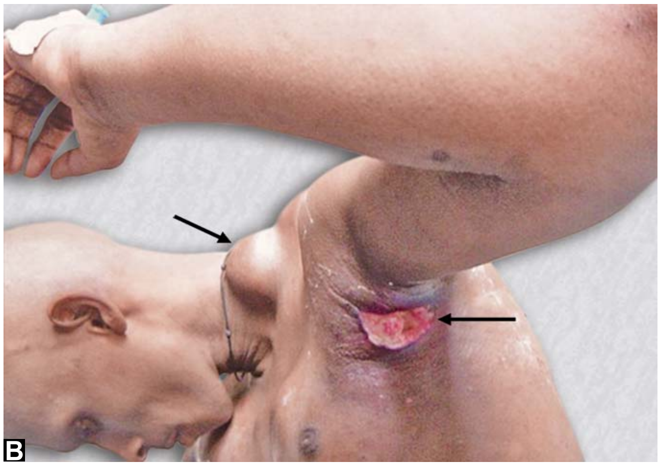

Figs 1.493A and B: Stage IV lymphoma with neck nodes/sternal swelling/axillary nodes which has ulcerated (ulceration is not common in lymphoma).

WHO modified **REAL** classification (**R**evised **E**uropean **A**merican **L**ymphoma) of lymphoma:

- B-cell neoplasms
  - Precursor B cell neoplasm—ALL, LBL.
  - Peripheral B cell neoplasm—it includes all B cell related non-Hodgkin's lymphomas
- T-cell and putative NK cell neoplasms
  - Precursor T cell neoplasms—ALL and LBL T cell related
  - Peripheral T cell and NK cell neoplasm—it includes all T cell related non-Hodgkin's lymphomas
- Hodgkin's lymphoma
  - Predominant HL—nodular lymphocyte type
  - Classical HL
    - Nodular sclerosis
    - Lymphocyte rich
    - Mixed cellularity
    - Lymphocyte depletion

Note: ALL is acute lymphoblastic leukaemia. LBL is lymphoblastic lymphoma.

Older Classifications (Not practiced now)

- Jackson and Parker (1944): Paragranuloma; granuloma; sarcoma.
- Lukes and Butler (1966): Lymphocytic and histiocytic nodular; lymphocytic and histiocytic diffuse; nodular sclerosis; mixed cellularity; diffuse fibrosis; reticular.

HODGKIN'S LYMPHOMA (HL)

> *(The lymph glands of the neck) exhibited a firm cartilaginous structure of a light colour and very feeble vascularity, but with no appearance of softening or suppuration. Glands similarly affected accompanied the vessels into the chest, where the bronchial and mediastinal glands were in the same state and greatly enlarged.... The spleen was enlarged to at least four times its natural size... presenting the same structure as the enlarged glands.*
>
> —**Thomas Hodgkin,** 1832

- It is the most common type of lymphoma.
- Infectious mononucleosis, Epstein-Barr virus, HIV infection and genetic monoclonal B cell disorder (90%) may be the aetiologies.
- Grossly lymph nodes are fleshy, pinkish grey, and rubbery in consistency.
- Microscopically contains cellular infiltration with lymphocytes, reticulum cells, histiocytes, fibrous tissue and ***Reed-Sternberg*** cells: (Reed-Sternberg cells are *giant cells with two large mirror image nuclei)*.

Rye's classification

- Lymphocytic predominance. Has got *good prognosis*
- Mixed cellularity
- Nodular sclerosis
- Lymphocytic depletion. Has got *poor prognosis*

(Reed-Sternberg cells are also seen occasionally in certain other conditions like glandular fever).

Clinical Features

- It is more common in males.
- It has got bimodal presentation. It is seen in young and adolescents (20-30 years) as well as in elderly (> 50 years).
- Painless progressive enlargement of lymph nodes. They are smooth, firm, nontender, typically *India rubber* consistency.

Site

- Cervical lymph nodes most common—82% (lower deep cervical group and in posterior triangle).
- Others include axillary, mediastinal, inguinal, abdominal.

Specific Features

- Nodular sclerosis is most common type.
- Consecutive group of lymph nodes are involved.
- *Splenomegaly* is very common (45%).
- Hepatomegaly with jaundice—jaundice is due to haemolysis or due to diffuse liver involvement.

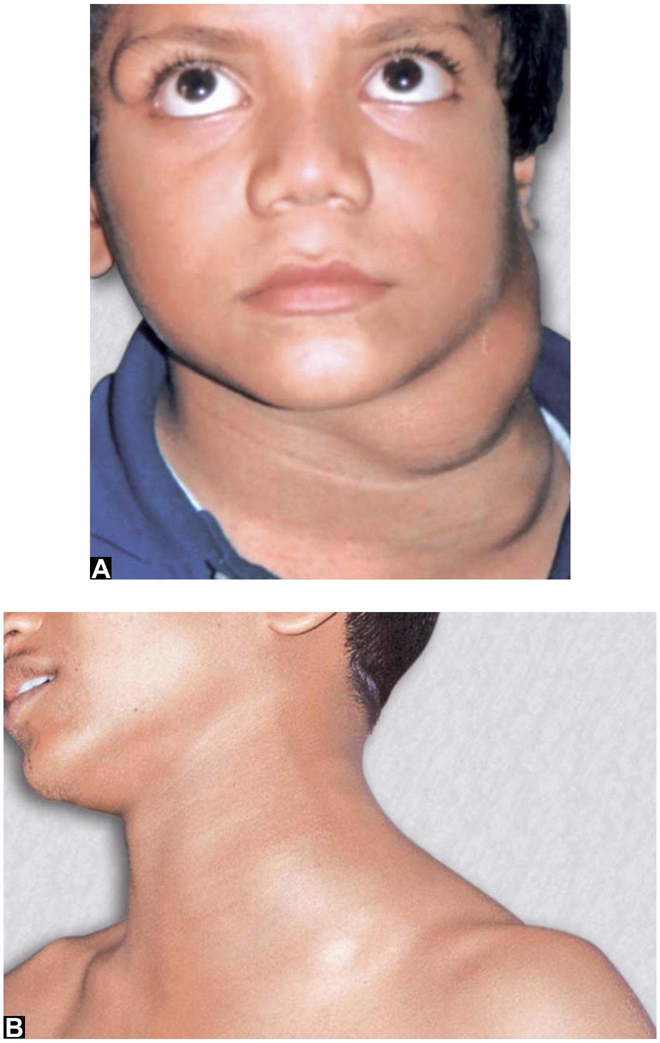

Figs 1.494A and B: Lymphoma in a boy and in an adolescent with neck nodes. Note the diffuse nature. Nodes are usually *India rubber* consistency.

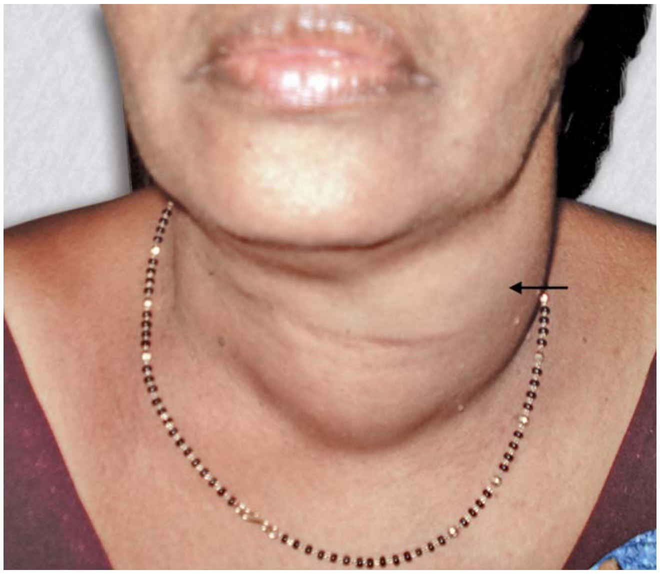

Fig. 1.495: Hodgkin's lymphoma in neck region.

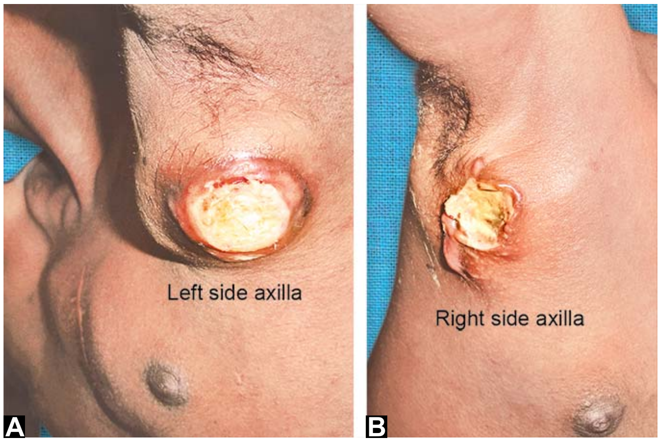

Figs 1.496A and B: Fungating lymphoma in both axilla of a patient.

- *Para-aortic lymph nodes* may be enlarged and often palpable as vertically placed mass at or just left of the midline, which does not move with respiration, does not fall forward, nonmobile, resonant, smooth, firm mass often with transmitted pulsation from aorta (secondaries are hard, nodular usually primary from GI, melanoma, testis). Ascites is not a common feature.
- Rarely alcohol may induce pain in enlarged node; classical *Pel Ebstein fever* described early is also seen in other infections like brucellosis.
- *Constitutional symptoms like fever, pruritus, weight loss may be present which signifies stage "B", which has got poor prognosis. Stage "A" is absence of these symptoms which signifies better prognosis.*

  Stage 'B' Symptoms

  - Weight loss, more than 10% in 6 months.
  - Fever, earlier called as *Pel Ebstein fever* is actually is due to brucellosis.
  - Pruritus—25%. It may be the only presenting symptom. It is usually seen in nodular sclerosis type.
  - Anaemia.
  - Bone pain.
- Mediastinal lymph node involvement may cause compression features like SVC obstruction. Mediastinal lymphoma is the most common mediastinal malignancy which usually occurs in *anterior* mediastinum. Occasionally presentation may be difficulty in breathing, chest pain, dysphagia and SVC obstruction (*Pemberton's sign may be positive*). It may be asymptomatic also. If ratio between maximum transverse diameter of mediastinal mass to maximum transverse intrathoracic diameter *(MMR*) is more than 0.33 in chest X-ray or more than 0.35 in CT chest, then it carries worst prognosis.
- Occasionally bone like vertebrae may get involved.
- Anaemia, pancytopaenia.

Ann-Arbor clinical staging

Stage 1: Confined to one group of lymph node
Stage 2: More than one group of lymph nodes on one side of the diaphragm
Stage 3: Nodes involved on both sides of the diaphragm
Stage 4: Extranodal involvement like liver, bone marrow
Suffix 'S'—Spleen involved
Suffix 'B'—Presence of constitutional symptoms
Suffix 'A'—Absence of constitutional symptoms

Note: In modification, following additions are there

- Single extralymphoid site is I E
- An extralymphoid site with one or more lymph nodes same side of diaphragm is II E
- An extralymphoid site with lymph nodes on both sides of diaphragm III E.
- An extralymphoid site with spleen and lymph nodes on both sides of diaphragm III SE.
- Spleen with lymph nodes on both sides of diaphragm is III S.

N—Nodes H—Liver S—Spleen L—Lung
M—Marrow P—Pleura O—Bone D—Skin
Stage III (1) is nodes above renal vein level and
Stage III (2) is below it.

Differential Diagnosis

- Tuberculosis.
- NHL.
- HIV.
- Chronic lymphatic leukaemia.
- Nonspecific lymphadenitis.
- Sarcoidosis.
- Secondaries in lymph nodes.

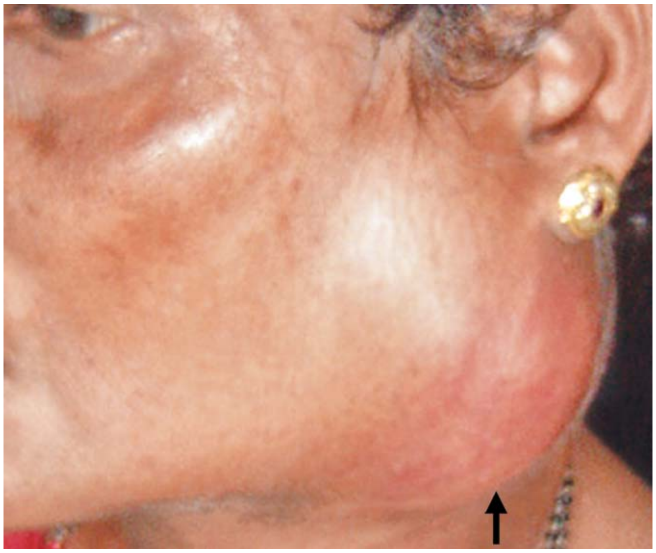

Fig. 1.497: Secondaries in neck lymph nodes. It is hard, large probably fixed and nonmobile making it advanced and inoperable.

Investigations

- *Blood:* Hb%, ESR, peripheral smear, blood urea, serum creatinine. Serum alkaline phosphatase and calcium may elevated.
- *FNAC of lymph nodes*.
- *Excision biopsy of lymph nodes*. Full lymph node is excised to retain the architecture of the lymph node. It is important to grade the tumour. It is better to have *immunohistochemistry* of tumour tissue. As cell mediated immunity is decreased, CD4/CD 8 ratio will be decreased.
- *Chest X-ray*—to look for mediastinal lymph nodes, pleural effusion.
- *U/S abdomen*—to look for the involvement of liver, spleen, abdominal lymph nodes.
- *CT scan of mediastinum/chest*, abdomen and pelvis is better, ideal and essential. *CT is* used ideally *to stage* the disease.
- *Chamberlain's mediastinoscopy* and biopsy of mediastinal lymph node is done if peripheral nodes are not available for biopsy. *Laparoscopy* and biopsy of different abdominal lymph nodes is also a good option.
- *Lower limb lymphangiography* to look for the pelvic and retroperitoneal lymph nodes. It shows reticular pattern in node; it can be used to assess the therapeutic response and prognosis. Lymphoscintigraphy is better and acceptable. It is prognostic tool also.
- *Bone marrow biopsy* to stage and also to see the response to treatment.
- *Staging laparotomy*:
    - The abdomen is opened. Splenectomy is done mainly to remove the tumour bulk, as spleen is commonly involved and also to avoid irradiation to splenic area which often causes unpleasant pulmonary fibrosis. Biopsies are taken from both lobes of the liver (needle biopsy) from para-aortic, celiac mesenteric, iliac nodes. In females ovaries are fixed behind the uterus to prevent radiation oophoritis *(oopheropexy/ovarian translocation)*.
    - Staging laparotomy is not routinely done now. It is done only if it benefits the patient to have better plan of treatment or better result.
    - It is done in stage I/IIA lymphoma (HL) in selected patients.

***Note*:**
Staging laparotomy, splenectomy (requires for immune function), lymphangiogram, IVU are no longer/very rarely done now for HL.

Problems in HL

- Pleural effusion—respiratory discomfort.
- SVC obstruction.
- Spine if involved—not very common but can occur.
- Opportunistic infection—mycobacteria, cytomegalovirus, herpes zoster.
- Bronchopneumonia, septicaemia.
- Immunosuppression and its effects.
- Risk for other malignancies in later life.

Man cannot discover new oceans unless he has courage to lose sight of the shore—**Anonymous**

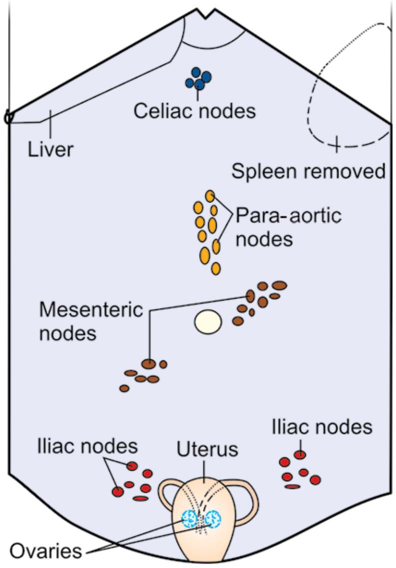

Fig. 1.498: Staging laparotomy for Hodgkin's lymphoma. Needle and wedge biopsies from liver/nodal biopsies from para-aortic, celiac, mesenteric, iliac nodes/splenectomy/ovarian translocation/iliac crest biopsy are the components of staging laparotomy. This can be very well-achieved through laparoscopy also now. Staging laparotomy is not commonly done now.

Treatment for HL

Treatment strategy for HL

- Early favourable stage—extended field radiation only
- Early unfavourable stage—extended radiation plus chemotherapy
- Advanced stage—extensive chemotherapy often with local radiation

Unfavourable signs in early HL

- Large mediastinal mass
- Extranodal disease
- Elevated ESR
- Four or more involved regions
- Presence of B symptoms
- Anaemia < 10 gram % low serum albumin level < 4 gram %
- Age more than 50
- Male gender

Note: First five parameters are most important.

- Stage I and II:
  - Mainly ***radiotherapy***—external high cobalt RT.
    - Above the diaphragm—*"Y" field therapy*, covering cervical, axillary, mediastinal lymph nodes. It may cause carcinoma of breast.
    - Below the diaphragm, *mantle* or *inverted "Y" field therapy*, covering para-aortic and iliac nodes. It may cause infertility.
  - Chemotherapy is also given.
- Stage III and IV: Mainly ***chemotherapy***.

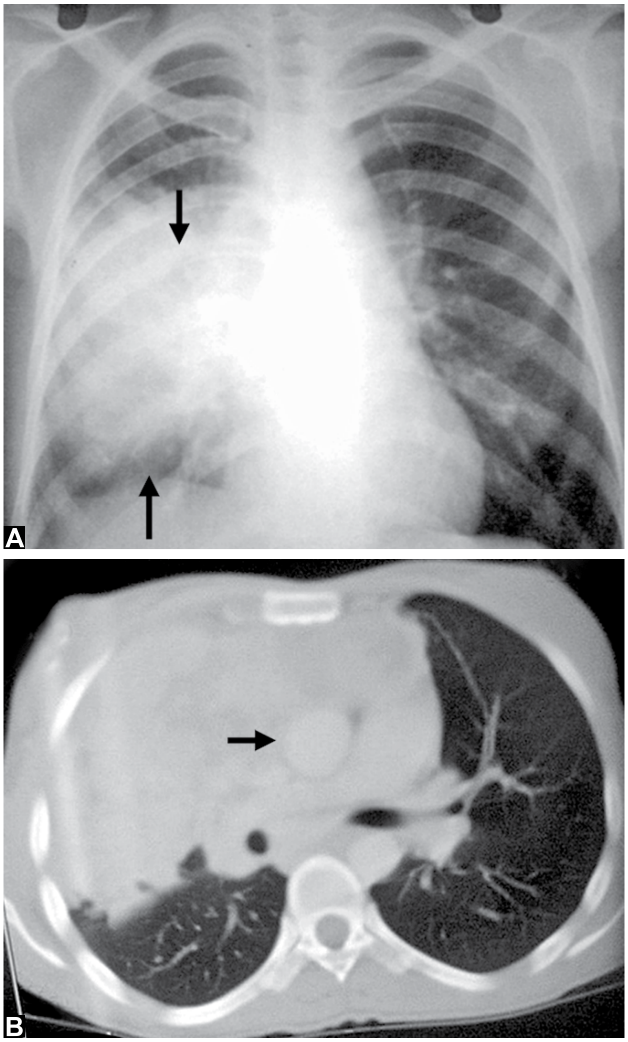

Figs 1.499A and B: Mediastinal nodes involved in lymphoma as seen in chest X-ray and chest CT scan.

Differences between HL and NHL

	HL (more common)	*NHL*
Age:	Young and elderly	Middle age and elderly
Pattern of involvement:	Symmetrical and consecutive	Asymmetrical
Cervical lymph node:	Commonly involved	Any group can be involved
Splenomegaly:	Common	Not common
Peripheral lymph node involvement (e.g. epitrochlear nodes)	Not common	Common
Treatment:	Mainly radiotherapy Chemotherapy (MOPP regime)	Mainly chemotherapy
Prognosis:	Better	Poor

Drugs used include:

❖ Mustine.	**M.** (Mechloroethamine). 6 mg/sq meter on 1st and 8th day.
❖ Oncovine.	**O.** (Vinca alkaloids). 1.4 mg/sq meter on 1st and 10th day.
❖ Procarbazine.	**P.** 100 mg orally daily for 10 days.
❖ Prednisolone.	**P.** 45 mg orally daily for 10 days.

Other regimens available—MVPP, ABVD

ABVD is becoming more popular and commonly used now. Adriamycin—30 mg/sq meter (cardiotoxic); Bleomycin—10 mg/sq meter (pulmonary fibrosis); vinblastine—6 mg/sq meter (bone marrow suppression); DTIC/Dacarbazine 350 mg/sq meter.

Treatment for Relapses

- Autologous stem cell transplantation.
- High dose chemotherapy.
- High dose chemotherapy with autologous stem cell transplantation.
- MOPP/ABV hybrid regime.
- Single dose vinorelbine (new vinca alkaloid); gemcitabine; immunotherapy; tumour vaccination; gene therapy.

Prognosis

- Stage I and II—80%.
- Stage II IA—70%.
- Stage III B and stage IV—< 40%.

Prognostic factors

- Stage I and II has got better prognosis
- Lymphocytic predominance has got better prognosis
- Stage "A" without constitutional symptoms has got better prognosis
- Anaemia < 10 g %; hypoalbuminaemia < 4 g %; lymphocyte count < 600/mm^3; WBC count > 15,000/ mm^3; age > 50 years; involvement of bone and liver are other poor prognostic factors

NON-HODGKIN'S LYMPHOMA (NHL)

- It occurs in middle aged and elderly. It is more aggressive than HL.
- It involves asymmetrical group of lymph nodes.
- General condition is poor.
- Inner Waldeyer ring, epitrochlear lymph nodes, peripheral lymph nodes are commonly involved.
- Spleen is not commonly involved.
- Hepatomegaly is common.
- Vertebral involvement is common; paraplegia can occur. (40%).
- Secondary infection, cachexia and immunosuppression is more common.
- Inner Waldeyer ring and oropharynx lymphoma (NHL-B cell) may be associated with MALT lymphoma of stomach; so gastroscopy is indicated in these patients.
- Pyoderma gangrenosum may the presentation.
- *Small bowel lymphomas* are usually NHL type. *Western type* is annular, ulcerative multiple B cell type, presents as obstruction, bleeding, weight and appetite loss, perforation. *Celiac disease related lymphoma* is usually primary T cell type presents with severe unresponsive diarrhoea, PUO, obstructive features. *Mediterranean* lymphoma of small bowel is seen in North America and Middle East which is associated with alpha chain disease.
- Sarcoma, carcinomas/secondaries are the differential diagnosis for NHL.

Types

- Nodular (Follicular)
- Diffuse lymphocytic
- Undifferentiated
- Histiocytic type

Rappaport classification	*Working classification*
• Nodular	• Low grade
• Diffuse	• Intermediate grade
	• High grade

Pathological Classification

- *B cell type*—small lymphocytic, follicular, Burkitt's lymphoma.
- *T cell type*—cutaneous, mycosis fungoides, Sezzary syndrome, lymphoblastic lymphoma.

Note:
Many different classifications are there for NHL. Students have to refer pathology book when needed.

Histochemistry and tumour markers are very sensitive and essential tools other than biopsy; CT chest, abdomen, pelvis; MRI spine, peripheral smear and bone marrow biopsy. **Serum LDH,** calcium, CD5, CD23 counts are very useful prognostic indicators. Raised LDH carries poor prognosis and also patients require aggressive chemotherapy.

Treatment

Mainly **chemotherapy**
Various regimens available include:

- ChOPP—Chlorambucil, Oncovin, Procarbazine, Prednisolone.
- ABVD—Adriamycin, Bleomycin, Vincristine, Dacarbazine.
- ABVP—Adriamycin, Bleomycin, Vincristine, Prednisolone.
- Combinations of above.
- *Rituximab* may be used with chemotherapy regimes.
- *Role of radiotherapy* in NHL: When vertebra is involved.
- Prognosis is poor compared to HL.

MALT LYMPHOMA (MALToma)

- It is lymphoma arising from mucosa associated lymphoid tissue. It is usually primary GI lymphoma (4% of gastric lymphoma); of Non-Hodgkin's B cell type.
- The most common site is stomach. *Helicobacter pylori* infection is the causative agent commonly.
- It can be low grade or high grade.
- *Low grade* occurs in patient with chronic gastritis due to *Helicobacter* infection. Diffuse mucosal thickening with ulceration is seen. Antihelicobacter therapy is very useful in this type.
- *High grade* is aggressive one. It presents with features similar to gastric carcinoma often with smooth, firm gastric mass in epigastric region.
- Systemic features like fever, night sweats, and weight loss can occur in 50% cases. Bleeding with haematemesis with obstruction is not uncommon. Locoregional lymph nodes (only) may get involved.
- Endoscopic biopsy is essential to diagnose. Endosonography, bone marrow biopsy, CT scan of chest, abdomen and pelvis is a must to confirm primary GI/gastric lymphoma and to identify or to rule out extragastric lymph node NHL.
- *Radical subtotal gastrectomy with adjuvant chemotherapy* is the treatment along with treatment for *Helicobacter* infection. Primary chemotherapy and RT is also useful. Palliative gastrectomy is needed in case of bleeding and obstruction.

BURKITT'S LYMPHOMA (Malignant Lymphoma of Africa)

> *Thirty-eight cases of a sarcoma involving the jaws of African children are described. This is a syndrome which has not previously been fully recognised. It is by far the most common malignant tumour of childhood seen at Mulago Hospital.*
>
> —**Denis Parsons Burkitt,** 1958

- It is common in South Africa and New Guinea.
- *Epstien-Barr virus* may be the etiological agent. It is common in children.
- It is associated with infectious mononucleosis.
- It is common in malaria endemic area.
- The tumour is multifocal, rapidly growing, painless.
- Different groups of lymph nodes can also be affected.

Microscopy

Primitive lymphoid cells with large clear histiocytes—*starry night (starry sky)* pattern.

Sites

- It is common in jaw—either lower or upper.
- Abdominal presentation and renal involvement is common (75%).
- Renal involvement often may be bilateral.
- In females, ovaries are commonly affected.

Types

a. Endemic (African)—commonly occurs in jaw.
b. Nonendemic (sporadic)—commonly occurs in abdomen.
c. Aggressive lymphoma—occurring in HIV patients.

Investigation

- FNAC and biopsy confirms the diagnosis.
- Jaw X-ray shows osteolytic lesions.
- U/S abdomen is done to look for involvement of kidneys.
- Blood urea and serum creatinine estimation is done.

Treatment

- Radiotherapy.
- Chemotherapy: Cyclophosphamide, methotrexate, orthomelphalan.
- Surgery is usually not indicated unless it is localised or in case of involvement of ovaries.

Prognosis is good.

CUTANEOUS T CELL LYMPHOMA

- Cutaneous T cell lymphoma comprises mycosis fungoides, Sezzary syndrome, reticulum cell sarcoma of skin and other cutaneous lymphocytic dysplasias. Mycosis fungoides is the most common among them.
- Cutaneous T cell lymphoma can be indolent (commonly mycosis fungoides); aggressive (Sezzary syndrome); provisional (granulomatous/panninculitis like T cell lymphoma).
- Initial macular patch/plaque phase slowly changes into tumour phase with painful, pruritic erythroderma often with visceral spread. Alopecia mucinosa and follicular mucinosis are common in mycosis fungoides. Lymph nodes may get involved. Tumour cells in peripheral smear are also important in deciding therapy and prognosis.
- Multiple skin biopsies/peripheral smear/node biopsy/immunohistochemistry/pheo or genotyping are important investigations
- Prognosis depends on extent of skin involvement (more the 10% body surface area carries poor prognosis)/nodal spread/blood spread.
- *Treatment:* Localised external beam radiotherapy; topical chemotherapy (bexarotene gel/carmustine ointment); phototherapy; total skin electron beam therapy; extracorporeal photochemotherapy. Bexarotene is a type of retinoid.
- ***Sezzary syndrome*** is a type of cutaneous T cell lymphoma with skin lesions with special Sezzary cells having cribriform nucleus. It is often associated with leukaemias. It is treated like any other cutaneous T cell lymphoma.

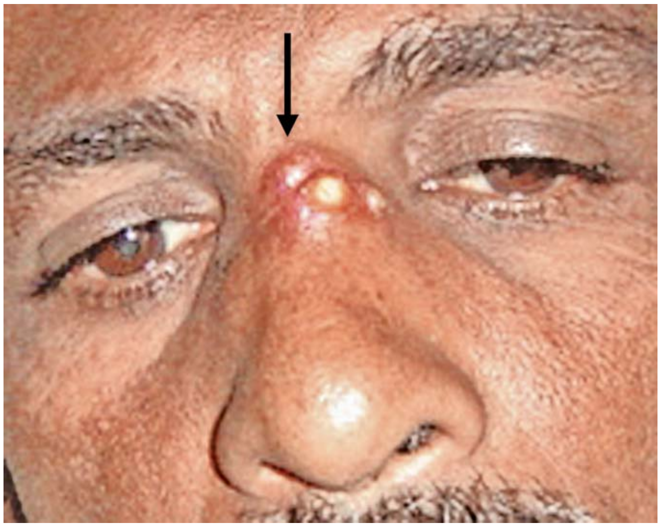

Fig. 1.500: Cutaneous T cell lymphoma.

CHYLOUS ASCITES

It is collection of lymph in the peritoneal cavity. It is due to obstruction of intestinal lymphatics and subsequent leak.

Causes

- *Most common cause* is—*congenital* lymphatic abnormality in children (*megalymphatics* with lymphoedema) and in adults lymph node malignancy, either primary or secondary.
- Filarial lymphoedema causing obstruction.
- Malignancy either nodal secondaries or nodal primary causing obstruction.
- Tuberculosis causing blockage of lymph drainage and rupture and leak into the peritoneal cavity.
- Post-surgical cause.

Features

- Ascites, often massive.
- Severe malnutrition and protein deficiency.
- Features specific to the cause.
- Triglycerides > 110 mg/dl in ascitic fluid is diagnostic.
- Ascitic fluid aspiration is chalky white in colour and it shows chylomicrons. It should be studied for fat globules, proteins, AFB and malignant cells.
- Laparoscopy and biopsy is necessary when lymphoma/secondaries are suspected.
- CT scan and CT guided biopsy may be needed.
- Lymphangiography to find out the site of leak.

Treatment

- Control of infection.
- Antituberculous and antifilarial drugs.
- Fat free, protein rich diet.
- Nutritional support—TPN/enteral.
- Medium chain triglycerides can be given as it directly gets absorbed into the blood rather into the lymphatics.
- Repeated tapping, peritoneovenous shunts are often required.
- Surgical ligation of leaking lymphatic duct.

CHYLOTHORAX

It is accumulation of lymph in the pleural cavity.

It is *common on right side* because of long course of thoracic duct towards right side.

Causes

- Injury due to trauma/surgeries in neck or chest. Surgical trauma is the *most common cause*—may be oesophageal surgeries, pneumonectomy, cervical sympathectomy, neck dissections or aortic surgeries.
- Tuberculosis, lymphoma or secondaries in the mediastinum.
- Carcinoma lung or oesophagus.

Features

- Chest pain, dyspnoea, pleural effusion.
- Protein loss and malnutrition
- Pleural tap will show chalky white fluid rich in chylomicrons. Triglycerides more than 110 mg/dl in the pleural fluid.
- Chest X-ray, CT chest are needed.

Treatment

- ICT placement.
- Fat free protein rich diet.
- Antibiotics therapy for the cause.
- Often pleurodesis using bleomycin, talc, tetracycline or pleural stripping is needed.
- Thoracic duct ligation is beneficial in traumatic/iatrogenic cases.
- Through thoracoscopic approach, ligation of the thoracic duct is done if leak persists *beyond one week.* If oral cream is given to the patient 6 hours before surgery, leaking site will be better identified. Thoracic duct is ligated *above and below* the leak either through thoracoscopy or thoracotomy.
- ICT drainage, oral diet of medium chain triglycerides which is absorbed directly into blood not through lymphatics; with TPN, is the usual earlier way of management.

CHYLURIA

- It is passage of milky white chylous urine, which is aggravated after fatty meal.
- It may be due to obstruction of intestinal lymphatic vessels leading to high lymphatic pressure causing diversion of lymph into renal lymphatics or it often may be due to rupture

of intestinal lymphatics into renal pelvis or ureter leading into a lymphourinary fistula.

- *The most common cause is filarial* (2% of filarial cases): Other causes are tumour, tuberculosis, malaria and ascariasis infestation.
- Urinary infection, protein loss is common.
- It mimics bacterial/tuberculous pyuria or phosphaturia.
- Clot colic due to *lymph clot* in urinary system may be the presentation.
- Urine study, culture, IVU, lymphangiography, U/S abdomen is needed.
- Treatment is low fat, protein rich diet, antibiotics, DEC, plenty of oral fluid intake, ligation of dilated lymphatics through laparotomy or sclerosing the lymph vessels.
- Condition causes severe psychological and nutritional problem.

SARCOIDOSIS

It is a differential diagnosis for lymph node mass. It is basically a granulomatous condition of unknown cause with bilateral hilar lymphadenopathy; with involvement of lungs, liver, spleen, lymph nodes, lacrimal glands, parotid glands, CNS, hypercalciuria, acute onset of erythema nodosum in the skin. Fever and loss of weight are not common. It shows noncaseating epithelioid granuloma with positive Kveim-Siltzbach skin test (80%); high levels of serum angiotensin converting enzyme (SAGE). Investigations needed are—CT chest; mediastinoscopy; nodal biopsy; slit lamp examination of eye; often abnormal immunoglobulins in the circulation. It is treated by corticosteroids with good response. It should be differentiated from other causes of lymphadenopathy especially Hodgkin's lymphoma.

Note:

For topics tuberculous lymphadenitis and lymph node secondaries please refer chapter 'Neck'.

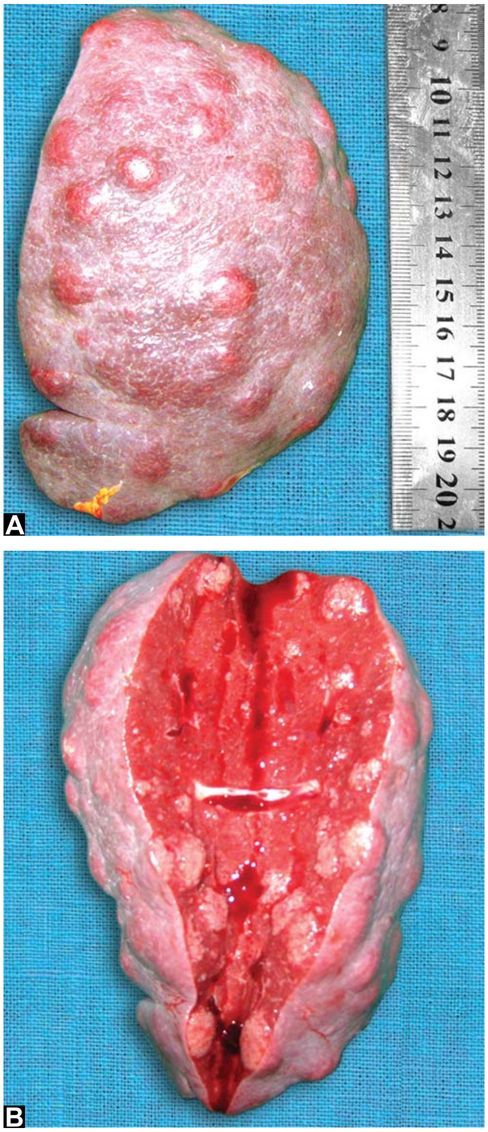

Figs 1.501A and B: Gross feature of sarcoidosis of spleen (*Courtesy:* Dr Arunkumar, MCh, Gastroenterologist).

P. Peripheral Nerves

> *The power of moving in every part of the body by means of the muscles which obey the will, or by means of others the actions of which are involuntary; the various perceptions by the five external senses; and lastly those mental powers named memory, imagination, attention, and judgement, together with the passions of the mind; all these seem to be exercised by the ministry of the nerves; and are impaired, disturbed, or destroyed, in proportion to any injury done to the brain, the spinal marrow, and nerves, not only by their peculiar diseases, of which we know little, but by contusions, wounds, ulcers, and distortions, and by many poisons of the intoxicating kind.*
>
> **—William Heberden,** *1802*

CHAPTER OUTLINE

- Peripheral Nerve Injuries
- Tinel's Sign
- Brachial Plexus Injuries
- Causalgia
- Median Nerve Injury
- Carpal Tunnel Syndrome
- Ulnar Nerve Injury
- Claw Hand
- Radial Nerve Injury
- Common Peroneal Nerve Injury
- Foot Drop
- Medial Popliteal Nerve Injury
- Axillary Nerve Injury
- Long Thoracic Nerve Injury
- Meralgia Paraesthetica

PERIPHERAL NERVE INJURIES

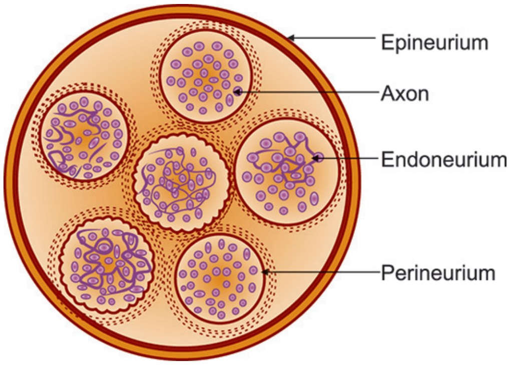

Fig. 1.502: Cross-section of a nerve.

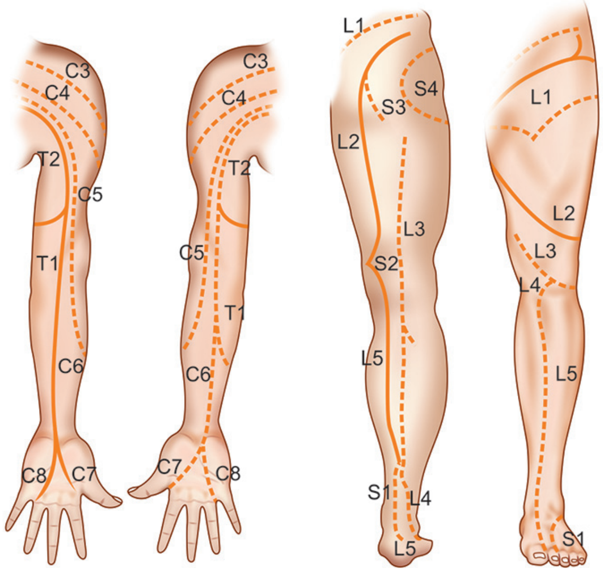

Fig. 1.503: Dermatomes.

Classification

Seddon's Classification

- *Neuropraxia:* It is temporary physiological paralysis of nerve conduction. Here recovery is complete. There is no reaction of degeneration.
- *Axonotmesis:* It is division of nerve fibres or axons with intact nerve sheath. There is reaction of degeneration distally with near complete recovery. Patient can present with sensory loss, paralysis of muscles or causalgia.
- *Neurotmesis:* Here complete division of nerve fibres with sheath occurs. Degeneration occurs proximally up to the first node of Ranvier as well as distal to the injury. Recovery is incomplete even after nerve suturing. There is complete loss of motor and sensory functions with loss of reflexes. If the nerve is mixed type other than pure motor or sensory recovery is still poorer.

It's how you deal with failure determines how you achieve success.

Injuries may be incised or lacerated or crushed one.

Cut end of the nerve forms proximally *neuroma* and distally *glioma.*

Neuromas may be:

- True neuroma or false neuroma.
- End neuroma or side neuroma.

Sunderland's classification

I: Conduction block—temporary neuronal block
II: Axonotmesis but endoneurium is preserved
III: Axonotmesis with disruption of endoneurium, but perineurium is preserved
IV: Here disruption of endo and perineurium has occurred but epineurium is intact
V: Neurotmesis with disruption of endo, peri and epineurium has occurred

Clinical Features

- Loss of sensory, motor, autonomous and reflex functions.
- Secondary changes in the skin and joint.

Primary nerve suturing is done if it is a clean incised wound.

Secondary nerve suturing is done after 3 weeks if it is a crushed wound.

Management

- Associated injuries like fracture, vessel injury, injuries in other systems should be looked for.
- Assessment of nerve injury is done by checking sensation, muscle power, reflexes.
- Nerve conduction studies.
- Investigations relevant for associated injuries.
- Exploration of the wound.

Debridement of the area is done. If injury is incised one, then nerve is sutured with 8-0 to 10-0 nonabsorbable interrupted sutures (polypropylene).

Types of Nerve Suturing

Usually microscope or loup is used for nerve suturing.

i. *Epineurorrhaphy*: Only epineurium is sutured using interrupted sutures.
ii. *Epi-perineurorrhaphy*: Initially perineural sheath and then epineurium is sutured.

Nerve suturing can be:

- ***Primary repair*: ** It is done immediately after injury. Nerve ends are minimally trimmed very close using a blade. All fascicles of the nerve are oriented correctly. Two stay sutures are placed to keep the orientation properly. Usually epineural suturing is done using 8 zero polypropylene interrupted sutures. It needs magnification. 6-8 sutures are placed for large peripheral nerve like median or ulnar nerve. For small nerve like digital nerve, only 2-3 sutures are placed.
- ***Secondary repair*:** It is done at a later period. It is done in a preexisting scar tissue. Here first nerve ends, both proximal and distal are identified, carefully dissected adequately. Proximal neuroma and distal glioma are trimmed for 1 cm to expose the normal fascicles of the nerve ends. Often guide sutures of silk may be present which were placed earlier during exploration of the trauma. Once nerve ends are clean, it is sutured alike primary suturing with stay sutures, with proper alignment of fascicles, followed by epineural suturing. Here as epineurium is thicker, suturing is easier.

If nerve is lacerated, then *marker stitches (using silk)* are placed at the cut end site to identify the nerve for suturing at a later period.

If nerve suturing fails or if could not be done, then tendon transfer is done at a later period after 4-6 months.

Incomplete injury usually does not require any suturing.

Easier suturing is achieved by following methods:

- Relaxing incisions.
- Transpositioning of the nerve.
- Shortening of the bone.
- Nerve graft—usually sural nerve is used for nerve graft.
- Positioning of the limb.
- Initially *neurolysis* (release of the scar tissue adjacent to injury) is done in case of secondary suturing.

Prognostic factors in healing of the nerve injury:

- Higher the lesion worser the prognosis.
- More the gap between the cut ends worser the prognosis.

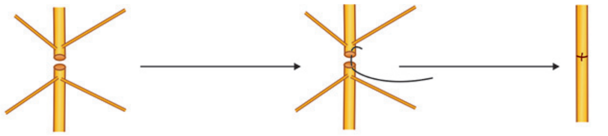

Severed nerve ends are trimmed just close to the ends. Guide/stay stitches are placed on both sides of each ends.

Orientation of fascicles are done in both cut ends. Epineural sutures are placed with knots on outside. Polypropylene interrupted sutures are used.

Completed epineurorrhaphy.

Fig. 1.504: Primary nerve repair.

If cut ends are without any undue tension (tension free), then ends are sutured with orientation of fascicles.

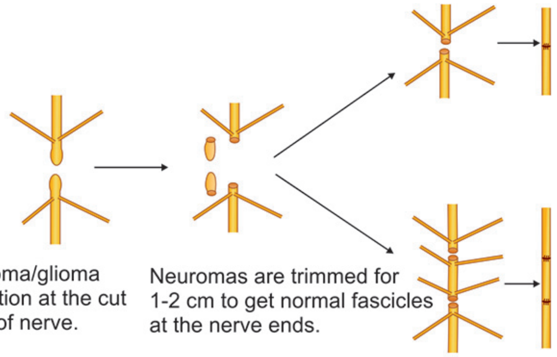

Neuroma/glioma formation at the cut ends of nerve.

Neuromas are trimmed for 1-2 cm to get normal fascicles at the nerve ends.

If there is tension then nerve graft is necessary using sural nerve.

Fig. 1.505: Secondary nerve repair.

- Associated injuries alter the prognosis.
- Children do better with nerve injury.
- Type of the injury also decides the prognosis.

The rate of growth of nerves after peripheral nerve suturing is 1 mm/day.

TINEL'S SIGN

It is the clinical sign (prognostic indicator) used to assess the level of regeneration. It is elicited 3 weeks after the nerve injury (Regeneration begins after the completion of nerve degeneration).

Tapping over the course of the nerve is done *from distal to proximal* to elicit a sensation of "*pins and needles*" or *hyperaesthesia.*

If sensation is felt at the site as well as distally along the distribution of the nerve, that means good recovery can be expected. If sensation is felt only at the site of tapping, then result is equivocal. If no sensation is felt it means no recovery.

Causes of peripheral nerve lesions

- Traumatic: Either closed or open injury
- Inflammatory: Leprosy, herpes zoster, diphtheria
- Compression neuropathies
- Lead poisoning
- Arsenical poisoning
- Alcoholism
- Diabetes mellitus
- Vitamin B_1 deficiency
- Porphyria
- Neurofibroma and other neural tumours
- Idiopathic

BRACHIAL PLEXUS INJURIES

It can be:

Supraclavicular injury	65%
Infraclavicular injury	25%
Combined	10%

It can also be:

- Upper plexus injury.
- Lower plexus injury.

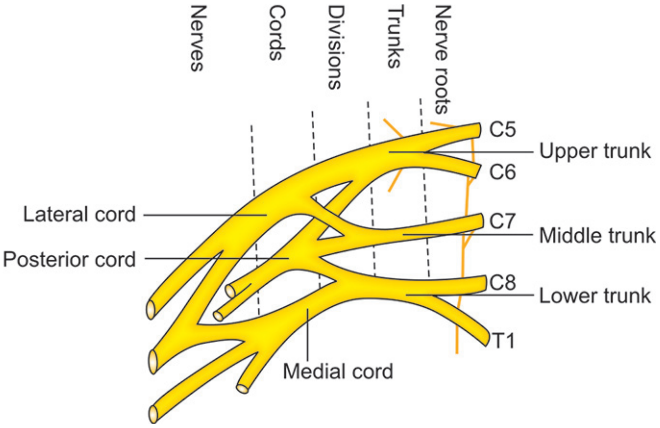

Fig. 1.506: Brachial plexus anatomy.

It can be:

Pre-ganglionic injury	Post-ganglionic injury
Avulsion injury	Usually less severe
More dangerous	Better recovery
Extends into the spinal cord	

Investigations

- Nerve conduction studies.
- CT/MRI.
- Electromyogram.
- X-ray cervical spine and part.

Treatment

- Conservative, nerve repair.
- Tendon transfer, physiotherapy.

Osteotomy of coracoid process proximal to the attachment of pectoralis minor, short head of biceps and coracobrachialis is done to improve abduction—*Sever's operation.*

CAUSALGIA

It is severe burning pain and hyperaesthesia in the distribution of a peripheral nerve due to *incomplete injury* to the nerve.

Features of upper and lower plexus injuries

	Upper plexus injury (Erb-Duchenne paralysis)		***Lower plexus injury (Klumpke's paralysis)***
1.	It is due to depression of shoulder by trauma	1.	Forcible hyperabduction of shoulder causes this injury
2.	After difficult labour in newborn	2.	In newborn it result due to difficult breech delivery
3.	Here C_5 and C_6 roots are injured	3.	Here C_8 and T_1 are injured
4.	Muscles affected are deltoid, biceps brachioradialis and supinator	4.	Intrinsic muscles of the hand are involved
5.	*Effects are:*	5.	*Effects are:*
	a. Elbow will be extended, pronated and upper limb is internally rotated (*Policeman receiving tip)*		a. Combined median and ulnar claw hand
	b. Sensory deficit over the lateral aspect of arm and upper part of the lateral forearm		b. Horner's syndrome
			c. Sensory deficit over the medial aspect of forearm, hand, and medial 1½ finger

In lesions of S_1 there is no dropped big toe.—**Gosta Norlen**

Sites

Common in upper limb.

Commonly seen in median nerve, also often in brachial plexus injuries. In the lower limb it is seen in sciatic nerve or tibial nerve injuries.

Pathology

Incomplete nerve injury produces abnormal impulse towards sensory nerve ending causing vasomotor instability and pain.

Clinical Features

- Hyperaesthesia with severe disabling and burning pain.
- Skin becomes red, shiny and glossy which sweats profusely—*Weir-Mitchell's skin.*
- Eventually skin becomes atrophic, cyanotic, cold and blotchy.
- Skin is less sensitive to heat, cold, and pin prick, but hypersensitive to touch and tender to pressure.
- Nails are rigid, brittle with change in colour.

Investigation

Nerve conduction studies.

Treatment

- Anti-inflammatory drugs, steroids, physiotherapy.
- IV guanethidine regionally.
- If not improved, sympathectomy—cervical for upper limb, lumbar for lower limb.

MEDIAN NERVE INJURY

- Median nerve arises from lateral ($C_{5,\,6,\,7}$) and medial cord (C_8 and T_1) of the brachial plexus. It is initially lateral to the axillary artery and becomes medial in the lower part of the arm and in the cubital fossa. It passes through the pronator teres, descends in relation to flexor muscles and enters the palm through the carpal tunnel at the wrist.
- It supplies pronator teres, flexor carpi radialis, palmaris longus and flexor digitorum superficialis. Anterior interosseous branch of the median nerve supplies pronator teres, lateral half of the flexor digitorum profundus, flexor pollicis longus and pronator quadratus.
- In the wrist, it supplies abductor pollicis brevis, flexor pollicis and opponens pollicis of thenar eminence and lateral two lumbricals. It gives sensory supply to lateral three and half fingers of the hand.

Median nerve is affected in:

- Injuries
  - Supracondylar fracture of the elbow
  - Fracture—dislocation of the elbow
  - Direct cut injuries
- Leprosy
- Carpal tunnel syndrome
- As a part of brachial plexus injury

Clinical Features of Median Nerve Palsy

In high median nerve palsy

- Wasting of the thenar eminence. Loss of sensation on lateral three and half fingers.
- Ochsner's clasping test shows pointing index because of the inactivity of lateral two divisions of the profundus.
- "*Ape or Simian thumb deformity"* is due to overaction of the adductor pollicis which is supplied by the deep branch of ulnar nerve. As all other thenar muscles are paralysed, thumb comes in the same plane of the metacarpals.
- "*Pen test"*: In median nerve injury, pen held in front of the hand cannot be touched by thumb as abduction is not possible due to paralysis of the abductor pollicis brevis.

In low median nerve palsy profundus is not paralysed and so *pointing index* is not seen.

Investigations

- Nerve conduction studies.
- X-ray of the part in case of fracture.
- Electromyogram.

Treatment

- Nerve suturing or nerve graft.
- Tendon transfer.
- Treat the cause like carpal tunnel syndrome.

CARPAL TUNNEL SYNDROME

- It is the compression neuropathy of median nerve in the carpus, deep to flexor retinaculum.
- Flexor retinaculum (transverse carpal ligament) maintains the concavity of wrist and extends laterally from trapezium and scaphoid to pisiform and hook of the hamate medially.

Carpal tunnel is formed by carpal bones behind and flexor retinaculum in front. It contains median nerve and long flexor tendons of fingers and thumb. Ulnar nerve lies superficially, not in the carpal tunnel.

Median nerve gets compressed if space of the carpal tunnel gets reduced.

Causes

- Lunate dislocation, malunited Colle's fracture.
- Radiocarpal arthritis, flexor tendon tenosynovitis.
- Myxoedema, acromegaly, pregnancy.

Clinical Features

- Common in females.
- Tingling, numbness, paraesthesia and burning sensation in the lateral three and half fingers supplied by median nerve. Burning sensation gets aggravated at night.
- Ape thumb deformity, wasting of thenar muscles, weakness of opponens pollicis and abductor pollicis brevis, i.e. features of low median nerve palsy.
- When BP cuff is inflated patient feels the typical pain in the fingers.

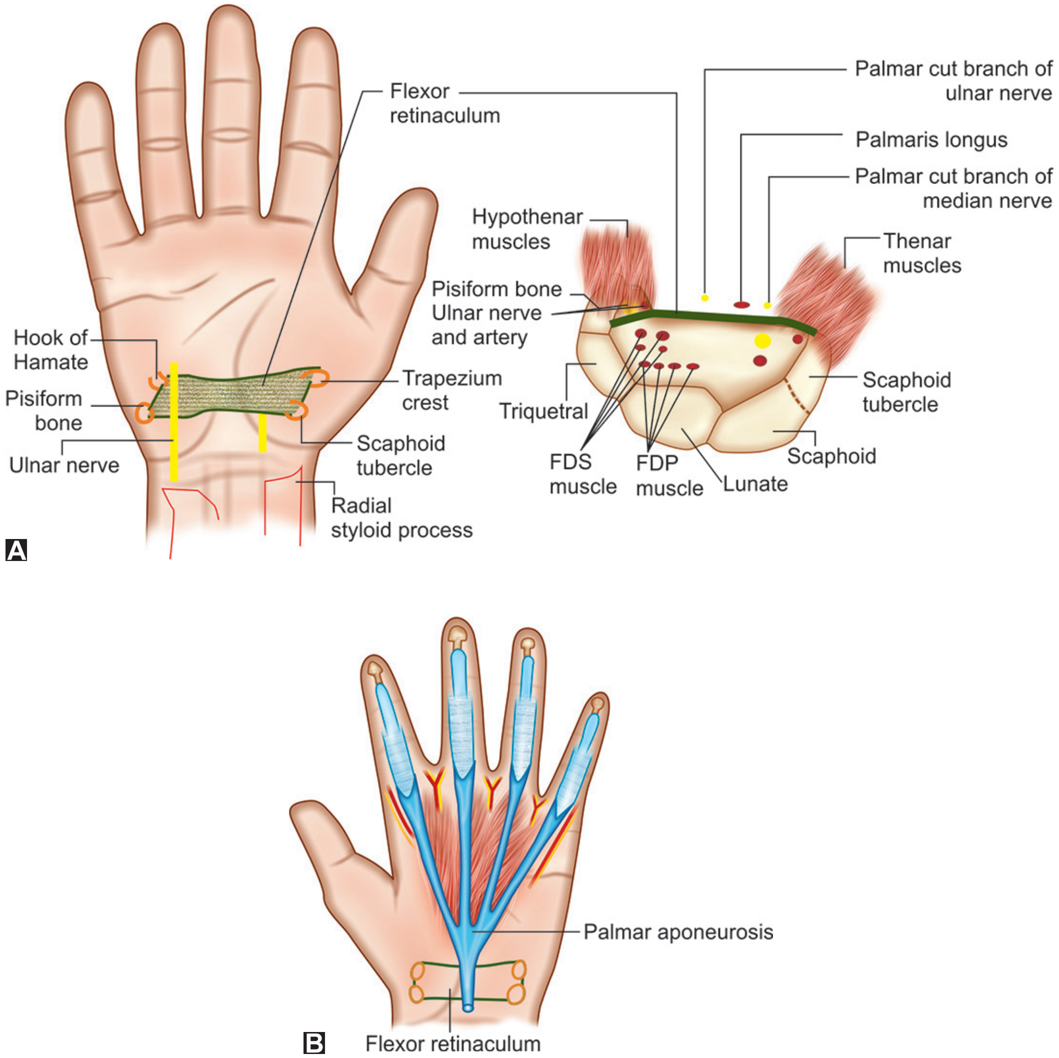
Flexor retinaculum
Hook of Hamate
Pisiform bone
Ulnar nerve
Trapezium crest
Scaphoid tubercle
Radial styloid process
A
Palmar cut branch of ulnar nerve
Palmaris longus
Palmar cut branch of median nerve
Hypothenar muscles
Thenar muscles
Pisiform bone
Ulnar nerve and artery
Triquetral
Scaphoid tubercle
Scaphoid
FDS muscle
FDP muscle
Lunate
Palmar aponeurosis
B
Flexor retinaculum

Figs 1.507A and B: Anatomy of the carpal tunnel and flexor retinaculum.

- Tapping the median nerve at the distal end of forearm with the wrist held in extension aggravates the symptoms.
- Condition is often bilateral.

Phalen's Test (Wrist Flexion Test)

Flexion of the wrist causes exacerbation of the symptoms within 1 minute and the symptoms will disappear as the wrist is straightened.

Differential Diagnosis

- Cervical spondylosis.
- Cervical rib syndrome.

Diagnosis

Nerve conduction studies.

Treatment

- Surgical decompression of median nerve by cutting both superficial and deep part of flexor retinaculum completely, by 'S' shaped incision.
- Surgery is usually done under *local anaesthesia*. General or brachial block can be used. Tourniquet is commonly used. *Vertical crease incision* is made in the proximal part of the palm with convexity of the incision towards the ulnar side.
- Skin incision is deepened. Palmar cutaneous branch of the median nerve should be preserved. Incision is deepened to identify the flexor retinaculum. *Entire length, both superficial and deep parts* should be cut properly. It is cut towards ulnar side of the wound. *Only skin* is sutured using interrupted nonabsorbable 3 zero polypropylene or polyethylene sutures.
- *Complications* are incomplete fasciotomy and recurrence, nerve injury.

To know how to laugh whole heartedly, you have to know how to weep also.

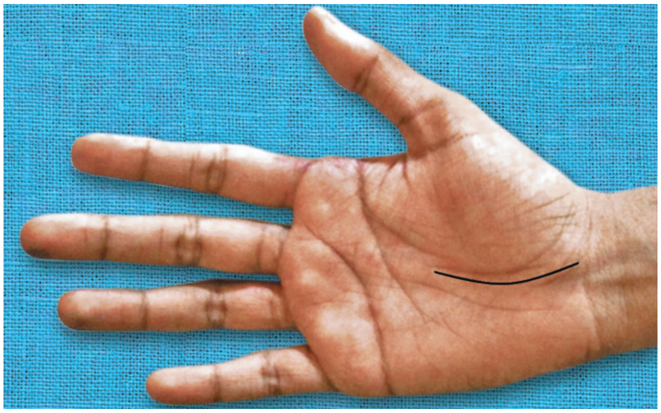

Fig. 1.508: Incision for carpal tunnel syndrome.

- Using small proximal incision, endoscopy can be passed to visualise and cut the entire flexor retinaculum—*minimal access surgery.*
- Postoperatively **good physiotherapy** is required.
- Condition is permanently curable.

ULNAR NERVE INJURY

- After arising from the medial cord of the brachial plexus (C_8 and T_1), it runs on the medial aspect of the axillary artery up to middle of the arm. Then it enters the posterior compartment in relation to triceps muscle. After passing behind the medial epicondyle and through two heads of flexor carpi ulnaris, it runs in front of the flexor digitorum profundus (FDP) in the forearm. It reaches the hand in front of the flexor retinaculum through *"Guyon's canal"*. Here it divides into superficial and deep branches.
- Ulnar nerve supplies flexor carpi ulnaris, medial half of flexor digitorum profundus, all muscles of the hypothenar eminence (palmaris brevis, abductor digiti minimi, opponens digiti minimi, flexor digiti minimi), adductor pollicis of the thenar eminence and all interossei of the hand. It also gives sensory supply to medial part of the hand, medial one and half fingers.

Ulnar nerve is affected in:

- Supracondylar fracture
- Injury to the medial epicondyle
- Tardy ulnar palsy
- Leprosy
- Cubitus valgus deformity

Clinical Features

- *Claw hand* deformity.
- Weakness of all the muscles supplied by the ulnar nerve.
- "*Card test"*: A card is placed between the two fingers of the patient to grasp. As the palmar interossei are weak, patient cannot grasp [palmar interossei are adductors of the fingers (*PAD*)].
- Abduction of fingers are checked [dorsal interossei are abductors (*DAB*)].
- *Froment's sign*: A book is placed to grasp between fingers and thumb of the patient. Normally thumb will be straight because of the action of adductor pollicis muscle. As it is paralysed in ulnar palsy, grasp is achieved by the action of flexor pollicis longus and there will be flexed thumb.
- Loss of sensation over medial one and half fingers and hand.

Investigations

- Nerve conduction studies.
- Electromyogram.

Treatment

- Nerve suturing or nerve grafting.
- Tendon transfer.

Intrinsic minus deformity: It is due to loss of intrinsic muscle power, i.e claw hand.

Intrinsic plus deformity: It is due to muscle contracture and fibrosis.

Ulnar paradox: In ulnar palsy, higher the lesion, lesser the deformity, lower the lesion more the deformity. In higher lesion, FDP is also paralysed. In lower lesion FDP is intact and so FDP causes more flexion (overaction) and so aggravates the claw hand.

CLAW HAND

- It is the hyperextension of the metacarpophalangeal joint with flexion of the interphalangeal joints of the hand.
- Extension of MCP joint is due to unopposed action of extensor digitorum.
- Flexion of MCP joint and extension of interphalangeal joints are by extensor hood of interossei and lumbricals. So extensor hood is functioning mainly by ulnar nerve and also by median nerve. In ulnar or median nerve palsies, these actions are paralysed and so patient develops claw hand.
- It is actually *intrinsic minus deformity.*

Causes

- Leprosy
- Trauma
- Entrapment neuropathies
- Tardy ulnar palsy
- Klumpke's palsy

Clinical Features

- Typical claw hand.
- Loss of sensation along the distribution of the nerve.
- Inability to grasp card between the fingers.
- While holding the book between the thumb and fingers, thumb will be flexed in ulnar claw hand (*positive Froment's test*).

Types

- ***Ulnar claw hand****:* Only medial two fingers are involved.
  - *Low ulnar palsy*: Here lesion is in the wrist (at Guyon's canal). Here deformity is more because of the overaction of the FDP.

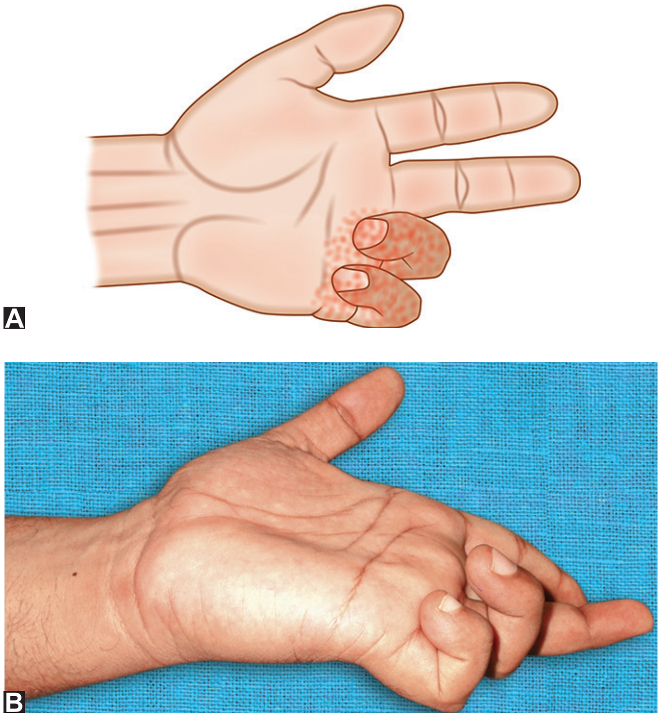

Figs 1.509A and B: *Ulnar claw hand* with hyperextension of metacarpophalangeal joints and flexion of proximal and distal interphalangeal joints in medial two fingers due to ulnar nerve palsy.

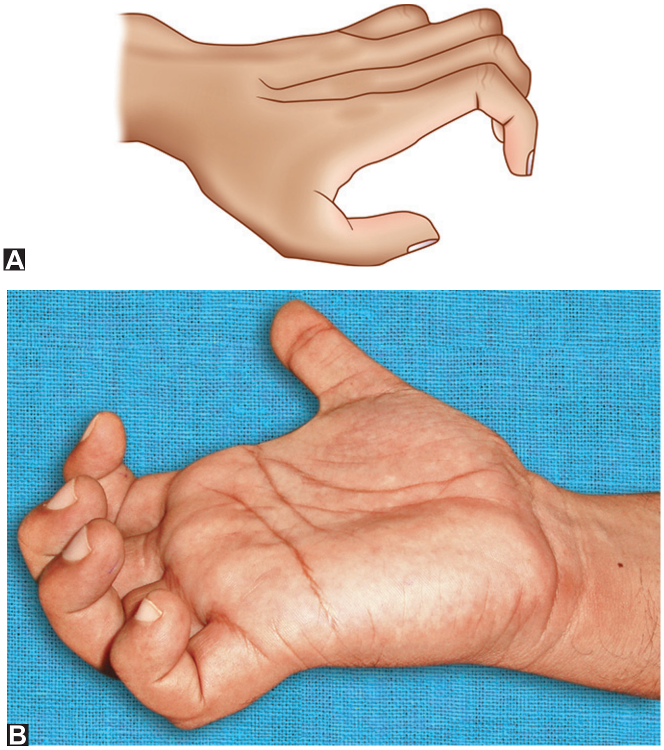

Figs 1.510A and B: *Combined claw hand* involving all fingers due to both ulnar and median nerve injuries.

- *High ulnar palsy*: Here FDP is also paralysed and overaction is not there. So lesser deformity occurs.

  ***Ulnar paradox*: **Higher the lesion lesser the deformity, lower the lesion more the deformity.

- ***Median claw hand*:** Only lateral two fingers are involved. It is less common.
- ***Combined median and ulnar claw hand*:** Here all four fingers of the hand are involved.

Investigations

- Electromyogram.
- Nerve conduction studies.

Treatment

- ***Paul Brand's operation*:** Extensor carpi radialis longus or brevis (ERCB) is transferred with a graft to the extensor hood through the lumbrical canal. Graft is taken from palmaris longus or plantaris muscle.
- ***Stye-Bunnell's operation*:** Flexor digitorum superficialis of index finger is used (only in ulnar claw hand) to transfer to extensor hood.
- *Fowler's operation*: Extensor digitorum is used to transfer to extensor hood.
- *Riordan operation*: Flexor carpi radialis is used for tendon transfer.
- *Anterior transpositioning* of the ulnar nerve in case of tardy ulnar palsy.

RADIAL NERVE INJURY

- Radial nerve is derived from the posterior cord of the brachial plexus ($C_{5, 6, 7, 8}$ and T_1). It descends behind the axillary artery in front of the subscapularis, latissimus dorsi and teres major. It passes through the medial and lateral heads of the triceps muscle, winds round the humerus through the radial groove and enters the forearm in front of the lateral epicondyle in relation to brachioradialis, brachialis and extensor carpi radialis longus muscles.
- In the arm it supplies triceps, anconeus, brachioradialis, extensor carpi radialis longus and part of brachialis. It gives posterior and lower lateral cutaneous nerves of the arm and posterior cutaneous nerve of the forearm.
- *Superficial branch* of the radial nerve from the elbow runs in the forearm in relation to supinator and brachioradialis and ends by forming five digital nerves which gives sensory supply to lateral three and half fingers on the dorsal aspect—except skin over the distal phalanges.
- *Deep branch also* called as posterior interosseous nerve winds round the radius supplying supinator and extensor carpi radialis brevis. It gives 3 short branches to extensor digitorum, extensor digiti minimi and extensor carpi ulnaris. It also gives two long branches—one to abductor pollicis longus and extensor pollicis brevis; another to extensor pollicis longus and extensor indicis.

Conditions where radial nerve is affected

In the ***axilla***

- *Crutch palsy.* It is neuropraxia
- Fracture upper end of the humerus
- Bony or soft tissue growth

In the ***radial groove***

- Pressure on the arm from the edge of the operating table
- *Saturday night palsy*—an individual with excessive alcohol consumption compresses his arm over the chair or by fall. It is neuropraxia
- Prolonged tourniquet application—*tourniquet palsy*
- Fracture of the shaft of the humerus
- *Rarely* intramuscular injection of drugs can cause radial nerve palsy

In the ***elbow***

- Dislocation or fracture neck of the radius

Clinical Features

- *Wrist drop* because of inability of extending the wrist.
- Inability to extend metacarpophalangeal joint, but extensions of the interphalangeal joints are normal.
- Inability to extend the forearm.
- Inability to extend the thumb.
- Flexion of the elbow against resistance with forearm in mid-prone position is difficult because of the weakness of the brachioradialis muscle.
- Loss of sensation in back of the arm, forearm, hand and lateral three and half fingers.

Posterior interosseous nerve is purely motor and so sensation is intact when it gets injured. *It causes only dropped fingers.*

Investigations

- X-ray of the part.
- Nerve conduction studies.

Treatment

- Nerve suturing or nerve graft.
- Tendon transfer.

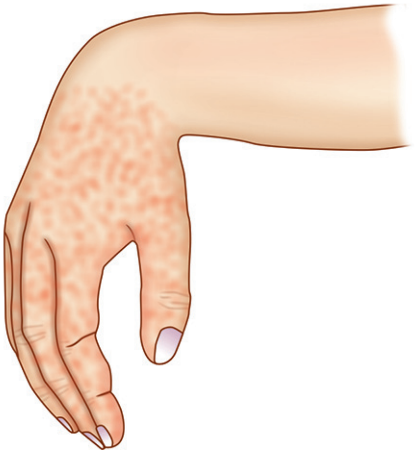

Fig. 1.511: *Wrist drop*—due to radial nerve palsy.

COMMON PERONEAL NERVE INJURY

This nerve supplies the extensor and peroneal group of muscles and sensory supply to the skin over the front and lateral aspect of the leg and dorsum of the foot.

Common peroneal nerve is affected in:

- Fracture neck of the fibula
- Leprosy
- Lead poisoning
- Iatrogenic

Clinical Features

- Foot drop with high stepping gait.
- Talipes equinovarus deformity.
- Loss of sensation on the lateral side of the leg and dorsum of the foot.

Management

- Treating the foot drop.
- MCR chappals.

FOOT DROP

Inability to dorsiflex and evert the foot due to paralysis of the peroneal and extensor group of muscles, as a result of common peroneal nerve injury.

Causes

- Fracture neck of the fibula
- Leprosy
- Lead poisoning
- Iatrogenic
- Direct incised wound

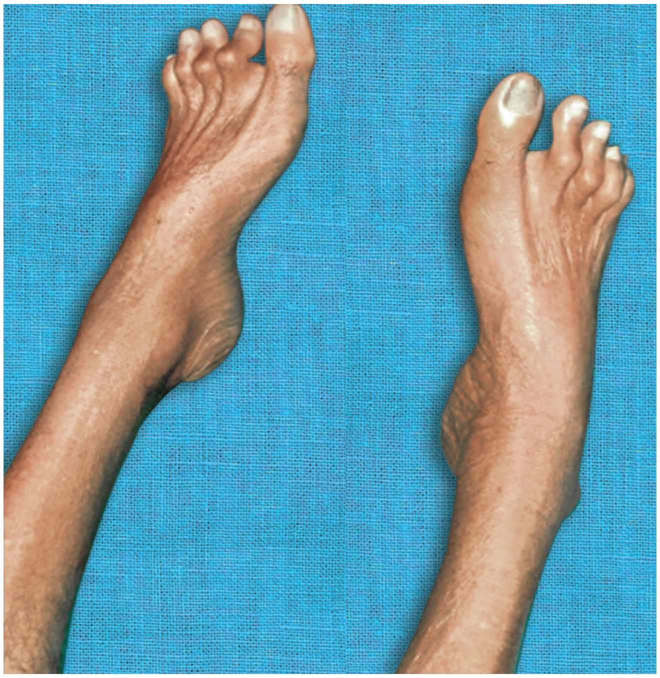

Fig. 1.512: *Foot drop* with claw toes. It is due to peroneal nerve injury (Lateral popliteal nerve).

Clinical Features

- High stepping gait.
- Loss of sensation over lateral and dorsum of the foot.

Treatment

Tendon transfer using tibialis posterior muscle. Tendon of the muscle is detached from its navicular insertion and with a tendon graft (from plantaris) it is transferred to cuboid and cuneiform bones to get dorsiflexion and eversion.

a. Ober's procedure.
b. Barr's procedure.

MEDIAL POPLITEAL NERVE INJURY

It supplies the soleus, gastrocnemius, popliteus, plantaris, tibialis posterior, flexor digitorum longus and flexor hallucis longus.

Medial popliteal nerve is rarely involved by any disease process. Trauma can cause medial popliteal nerve palsy.

Clinical features

- Inability to plantar flex the foot
- Claw toes
- Sensory loss in the sole of the foot

AXILLARY NERVE INJURY

Axillary nerve supplies the deltoid and teres minor muscle and also sensory supply to the skin over the upper lateral aspect of the arm.

Axillary nerve is affected in:

- Fracture neck of the humerus.
- Dislocation of humeral head.
- Following IM injection into the deltoid.

Clinically there will be loss of abduction of the shoulder and anaesthesia of the skin over the lateral part of the arm.

LONG THORACIC NERVE INJURY (NERVE OF BELL)

It supplies serratus anterior muscle. It arises from $C_{5,\,6,\,7}$ cervical roots.

The nerve is injured commonly in malignancy, during breast, axillary or chest wall surgeries.

Clinically, when outstretched (elbow extended) arm is pushed against the wall, the inferior angle of the scapula will become prominent (*Winging of the scapula*).

MERALGIA PARAESTHETICA

(Meralgia – Greek – thigh)

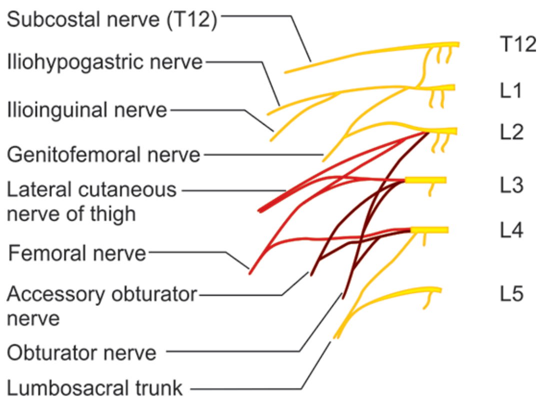

Fig. 1.513: Lumbar plexus showing origin of lateral cutaneous nerve of thigh.

- It is *entrapment neuropathy* of lateral cutaneous nerve of thigh. Nerve gets *compressed* while passing through the inguinal ligament. It arises from posterior divisions of lumbar plexus ($L_{2,\,3}$); runs over the quadrates lumborum and iliacus muscles; emerges behind the lateral part of the inguinal ligament; divides into anterior and posterior branches, supplying skin over anterolateral part of the thigh and anterior part of the gluteal region.
- It causes hyperaesthesia, tingling over upper lateral aspect of the thigh along the distribution of the nerve. Symptoms get worsened on standing or walking; it is relived by sitting.
- It *mimics* disc prolapse or Hansen's disease or neuropathies.
- It is *treated* by reassurance; carbamazepine; steroids. Often *release* of inguinal ligament fibres that are compressing the nerve is needed.

Q. Neoplasm

CHAPTER OUTLINE

- Definition
- Dysplasia
- Carcinoma *In Situ*
- Aetiologic Factors
- Spread of Malignant Tumours
- Grading of Tumour
- Staging of the Tumour
- Paraneoplastic Syndromes
- Investigations for Neoplasm
- Management Strategy for Cancers

DEFINITION

Willis defined neoplasm as "*it is an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues and persists in the same excessive manner even after cessation of the stimuli.*"

Neoplasia is:

- **P**rogressive
- **P**urposeless
- **P**erverted
- **P**ersistent
- **P**ervasive
- **P**roliferative mass of tissue.

Classifications

Benign	*Malignant*
Well-differentiated	Lack of differentiation
Structures are typical of tissue/cell of origin	Atypical structure with anaplasia
Smooth, slow, progressive rate of growth	Erratic, rapid growth
Normal mitotic figures	Abnormal mitotic activity
Well localised and capsulated	Not localised. Not capsulated
Do not infiltrate surrounding normal tissues	Infiltrate the surrounding tissues
No metastasis	Metastasise through lymphatics or blood
Curable	May not be completely curable
Few benign tumours after a long time may turn into malignancy	
Treatment is simple	Treatment is complex and complicated
No recurrence	Recurrence can occur

Sarcoma	*Carcinoma*
Arising from mesenchymal tissues	Arising from epithelial cells derived from any of the three germ layers
'Sar' means flesh (Greek), *Oma* means tumour	*Carc* means crab like
Smooth, firm or hard swelling	Hard, proliferative, with everted edge
Warm and vascular with dilated veins over the surface	
Spreads mainly through blood commonly to lungs, e.g. liposarcoma, fibrosarcoma,	Spreads through lymphatics as well as blood, e.g. squamous cell carcinoma, renal cell carcinoma, adenocarcinoma

- **Neoplasms which are only locally malignant**: No blood spread. No lymph node spread—Marjolin's ulcer, Rodent ulcer, Verrucous carcinoma, Adamantinoma.
- **Neoplasms which are loco-regionally malignant**: Spread only to regional lymph nodes is observed—Squamous cell carcinoma, Papillary carcinoma thyroid.
- **Neoplasms which are systemic** and spreads through blood and often also to lymph nodes.
- Melanoma, carcinoma breast.

Components

Parenchyma: It contains proliferating neoplastic cells.
Stroma: It contains supporting connective tissues and blood vessels.

Features of anaplasia

- Lack of cellular differentiation
- Pleomorphism—variation in size and shape
- Hyperchromatism—dark staining nuclei
- Anisocytosis
- Anisonucleosis
- Abnormal mitotic activity

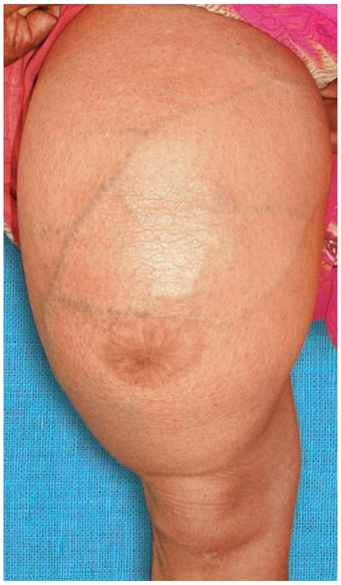

Fig. 1.514: Soft tissue sarcoma thigh. Note the dilated vassels on the surface.

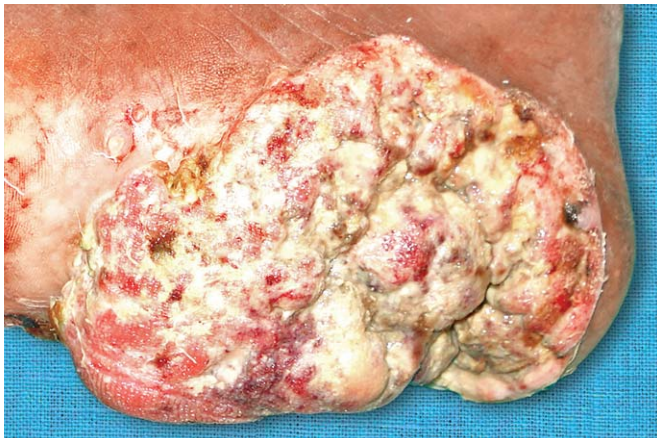

Fig. 1.515: Carcinoma heel region. Note the everted edge with proliferative cauliflower like look.

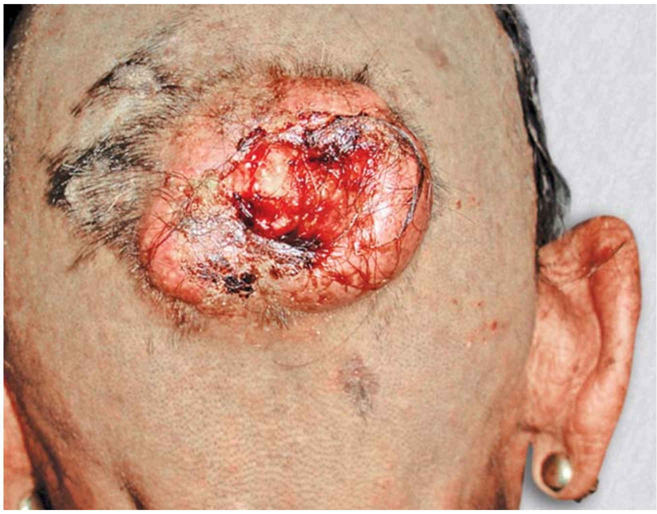

Fig. 1.516: Squamous cell carcinoma scalp. Note the everted edge. It is ulceroproliferative lesion.

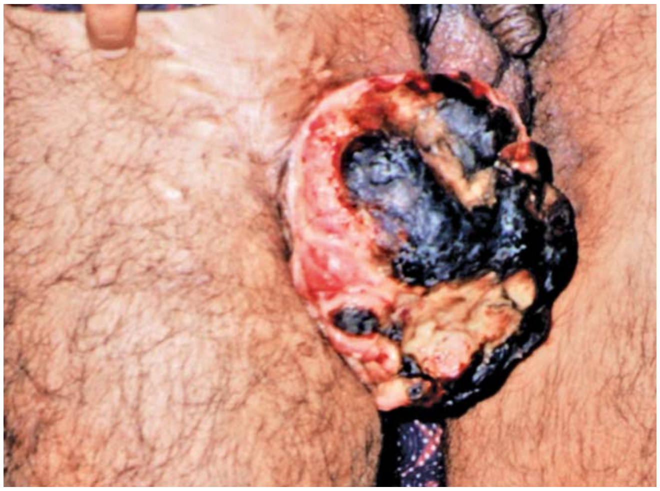

Fig. 1.517: Soft tissue sarcoma thigh.

Features of malignant tissues

- Altered differentiation, anaplasia
- Rapid rate of growth
- Local invasion—locally, lymphatic, vascular and perineural
- Metastasis
- Not capsulated
- Increased vascularity
- *Microscopic changes*—necrosis, numerous atypical mitotic activities, nuclear changes like pleomorphism/enlargement/hyperchromatism/clumping of chromatin/enlargement and multiplication of nucleoli

DYSPLASIA

It means “disorderd growth.” There is loss in the uniformity of the cells with pleomorphism and hyperchromatism, as well as loss in their architectural orientation.

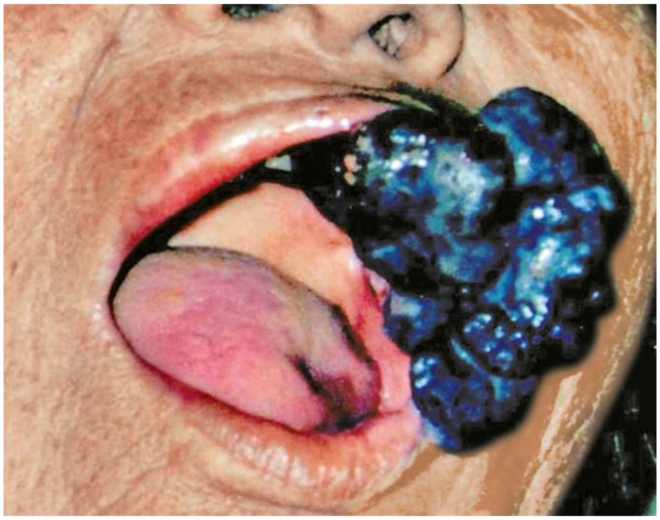

Fig. 1.518: Melanoma lip. Note the extension.

The creation of a thousand forests is in one corn.—**Ralph Waldo Emerson**

CARCINOMA *IN SITU*

Here dysplasia involves the entire thickness of the epithelium, and is preinvasive. Basement membrane is intact in carcinoma *in situ.*

AETIOLOGIC FACTORS

Age

It is more common in elderly. But it is variable.

Heredity

- Familial: Familial polyposis of colon
- MEN syndrome
- Neurofibromatosis
- von-Hippel-Lindau syndrome
- Familial breast and ovarian cancers

Genetic

- Xeroderma pigmentosa
- Ataxia telangiectasia
- Fanconi anaemia, Bloom syndrome

Tumour suppressor gene p53 plays important role in prevention of the cancers. Its function is prevention of replication of damaged DNA. If damaged DNA replicates there is high chances of abnormal mitotic activity and cancer transformation of the tissue. *Loss or reduction of ability/function of p53 suppressor gene* can lead into cancers. Common cancers associated with p53 loss are cancers of breast, colorectum, retina, bone, brain, soft tissues, blood and familial related type.

Acquired causes

- Chronic atrophic gastritis
- Solar keratosis
- Leukoplakia of oral cavity
- Ulcerative colitis

Chemical carcinogens

- Alkylating agents
- Hydrocarbons
- Smoking—lung, aerodigestive system, bladder cancer
- Asbestos—lung cancer
- Alcohol—liver cancer
- Amides, Azo dyes—bladder cancer
- Aflatoxin B_1
- Arecoline, collagenases and tannins (present in Betel nuts)
- Nitrosoamines, vinyl chloride, insecticides

Geographical variation

Japan—Carcinoma stomach
China, France—Carcinoma oesophagus
Hongkong—Nasopharyngeal carcinoma
Australia—Melanoma
India—Carcinoma oral cavity, gallbladder carcinoma
Newzealand—Small bowel tumours

Radiation carcinogens

- UV rays, ionising radiation.

Microbial carcinogens

- Human papilloma virus—carcinoma cervix
- Epstein-Barr virus—Burkitt's lymphoma, nasopharyngeal carcinoma
- Hepatitis 'B' virus—liver cancer
- Human T cell leukaemia virus Type I
- *Helicobacter pylori* can cause carcinoma of stomach and is associated with lymphomas [Mucosa Associated with Lymphoid Tissue (MALT)]

SPREAD OF MALIGNANT TUMOURS

1. **Local spread**: Into adjacent structures like soft tissues, vessels, bone.
2. **Lymphatic spread**
   - ***By permeation*****:** Here malignant cells proliferate through lymphatic vessels up to lymph node level. For example, in carcinoma breast malignant cells permeate into axillary lymph nodes.
   - ***By embolisation:*** Here cells get dislodged from lymphatic vessels and freely travel to spread into further level of lymph nodes. In carcinoma breast by embolisation, spread occurs from axillary lymph node to supraclavicular lymph node.
   - ***Retrograde lymphatic spread*** occurs once lymph vessel get blocked by malignant infiltration. In carcinoma breast retrograde spread occurs to opposite breast, opposite axilla, or to mediastinum. In melanoma, through dermal lymphatics and retrograde spread '***in transit nodules***' occur in the skin.
3. **Blood spread**

   Occurs through veins, as veins are thin-walled and infiltration is easier (Arteries contain elastic fibres in their wall, which resist malignant infiltration).

   Both by ***permeation*** (e.g. in renal cell carcinoma permeation through renal vein is common) and by ***embolisation*** (in other malignancies).

   Blood spread is commonly to lungs, bone (upper end of femur and humerus, ribs, skull), liver, brain, adrenals and other organs.

   In carcinoma prostate, due to increased pressure and venous blockade, retrograde venous spread occurs through vertebral venous plexus which causes **osteoblastic secondaries** in pelvic bones and vertebrae.
4. **Seedling:** For example
   - From lower lip cancer to upper lip as **kiss cancer.**

Kiss cancer

- Lip
- Bladder
- Cervix
- Vocal cord
- Vulva

   - Recurrence in the scar after surgery for malignancy, e.g. deposition of malignancy in the scar of SPC from bladder tumour.
   - Seedling in the peritoneal cavity from abdominal malignancy is common causing intractable ascites.

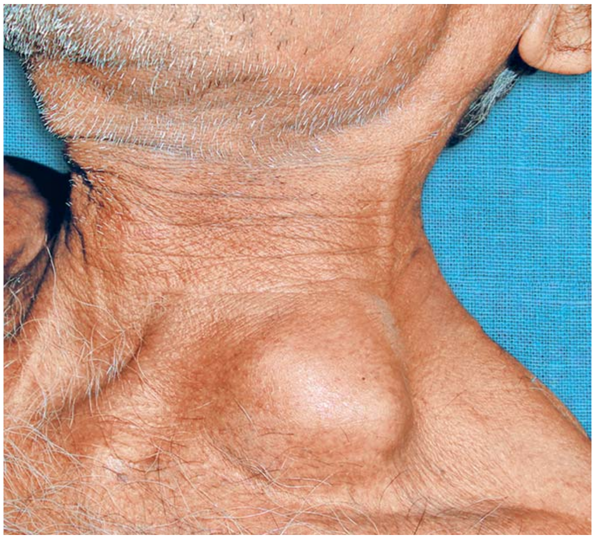

Fig. 1.519: Secondaries in neck. FNAC is the investigation to diagnose.

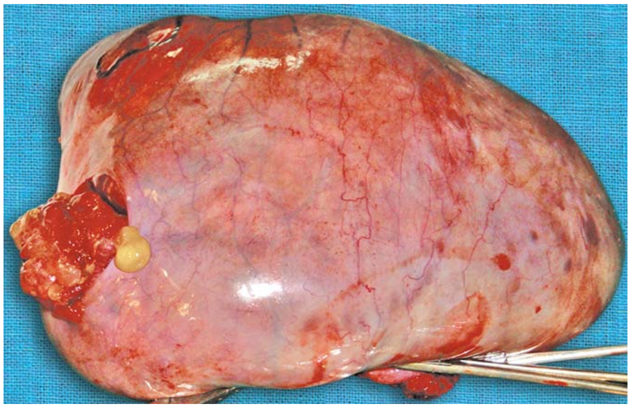

Fig. 1.520: Carcinoma ovary specimen. It was malignant teratomas of ovary. Other ovary was normal. So chemotherapy is used as an adjuvant. Patient was 23 years young female.

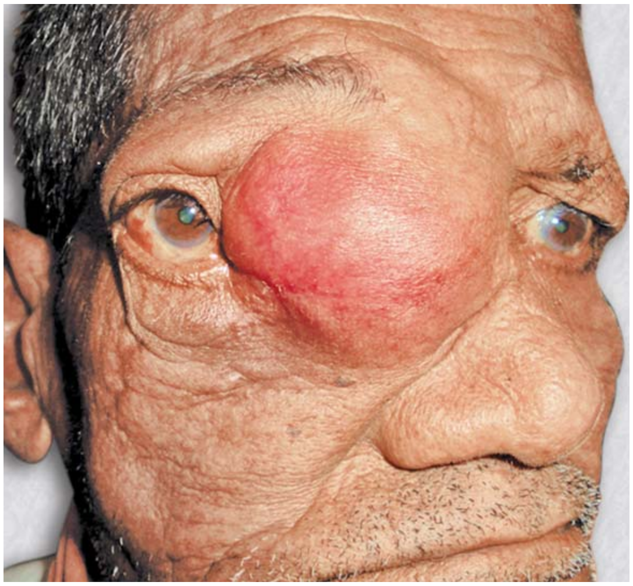

Fig. 1.521: Ethmoidal sinus tumour. Biopsy and CT scan is needed to confirm and assess the extent.

5. **Transcoelomic spread:** Spillage or dislodge of malignant cells occurs from primary site resulting in seedling on other organ, e.g. in carcinoma stomach secondaries in ovary (***Krukenberg tumour***) occurs due to transcoelomic spread. Here cells get deposited on the raw surface of ovary during ovulation (So it occurs in menstruating age group only).

GRADING OF TUMOUR

It signifies aggressiveness of tumour.

It is based on differentiation of tumour cells and mitotic activity.

Broder's grading of squamous cell carcinoma: Based on keratin/ epithelial pearls
Grade I: Well differentiated—>75% epithelial pearls
Grade II: Moderately differentiated—50-75% epithelial pearls
Grade III: Poorly differentiated—25-50% epithelial pearls
Grade IV: < 25% epithelial pearls

Special Features of Malignant Cells

- Independent and autonomous nature without any control or check.
- Normal telomeric shortening is altered. Normal cell divides for 40-60 times. Later telomere gets shortened and cell division stops. It is altered in malignancy with infinite cell division.
- Reduced cellular apoptosis. Normal cell has capacity to die once its function is over. It is normal programmed death of a cell. It is activated by tumour suppressor gene p53. If this gene loses its function, tumour will form.
- Tumour has got capacity to create and construct blood supply called as *angiogenic* capacity.
- Malignant cells invade through the basement membrane and spread through blood and lymphatics. It is through secretion of special enzymes, increased interstitial pressure, capacity to have motility, release of collagenases/proteases/integrins.
- Spread occurs by proliferation and permeation as well as embolisation.
- Creation of instability and alteration of existing tissue function (not obeying normal tissue function of that particular site).
- *Gompertzian growth*—rapid growth of a tumour occurs prior to its clinically detectable stage. Once it becomes clinically detectable which needs thirty generation of cell divisions (1 mm tumour is 10^9 cells), its growth slows down later. This late slow growth is due to decreased oxygen to tumour/by competition between host cells and tumour cells and also between tumour cells themselves.
- Tumour mutation occurs at different generation of divisions stabilising the tumour further.

STAGING OF THE TUMOUR

It is based on the size of the primary tumour, nodal spread and blood spread.

A man carries success or his failure with him; it does not depend on outside conditions.

It is called as 'TNM' staging.

'TNM' staging is more relevant than grading in managing and predicting prognosis.

PARANEOPLASTIC SYNDROMES

Certain symptom complexes which are not specifically explained by the tumour, but are relevant, often problematic and life-threatening are called as paraneoplastic syndromes.

Incidence: Seen in 10% of patients.

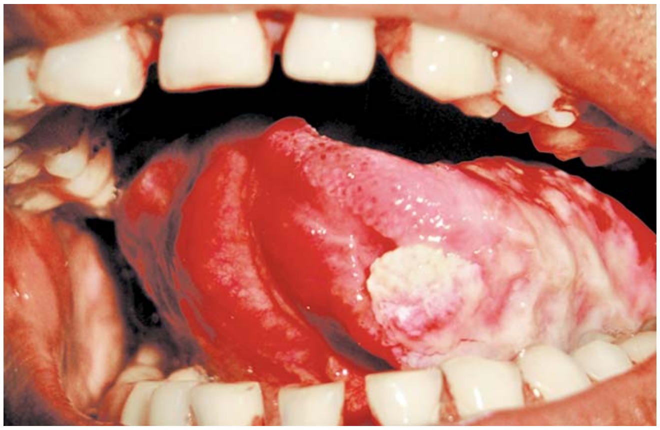

Fig. 1.522: Carcinoma tongue is more common in lateral margin. Incision biopsy from the margin is done here.

INVESTIGATIONS FOR NEOPLASM

I. *BIOPSY*

'Bio' means life or tissue. *Opsis* means vision or microscopy.

Biopsy means study of tissues using microscopy.

OR *life by vision*

OR *vision for life* (to save life).

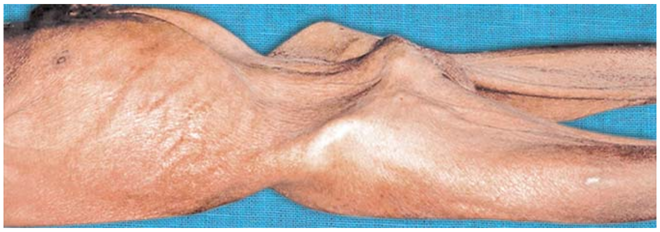

Fig. 1.523: Typical *malignant cachexia* observed in GIT malignancies. It is totally malnourished incapacitated situation.

Paraneoplastic syndromes

Syndrome	***Cause***	***Type of tumour***
Cushing's syndrome ACTH	or ACTH like substance	Small cell carcinoma lung Neural tumour Pancreatic tumour
Syndrome of inappropriate	ADH	Small cell carcinoma lung
ADH secretion		Brain tumour
Carcinoid syndrome	Serotonin, bradykinin	Bronchial adenoma
		Gastric and pancreatic carcinoma
Hypoglycaemia	Insulin	Fibrosarcoma
	Insulin like substance	Hepatoma
Hypercalcaemia	PTH related peptide	Renal cell carcinoma Carcinoma breast, SCC lung Ovarian carcinoma, leukaemia
Polycythaemia	Erythropoietin	Renal cell carcinoma, hepatoma Cerebellar haemangioma
Myasthenia	Immunologic	Lung cancer
Acanthosis nigricans	Epidermal growth factor	Lung cancer, uterine cancer, gastric carcinoma
Dermatomyositis	Immunologic	Lung carcinoma, breast carcinoma
Clubbing and hypertrophic		Ca lung
Osteoarthropathy of fingers		
Migrating thrombophlebitis	Tumour product mucin	Ca pancreas, lung
(Trousseau phenomenon)	It activates clotting	

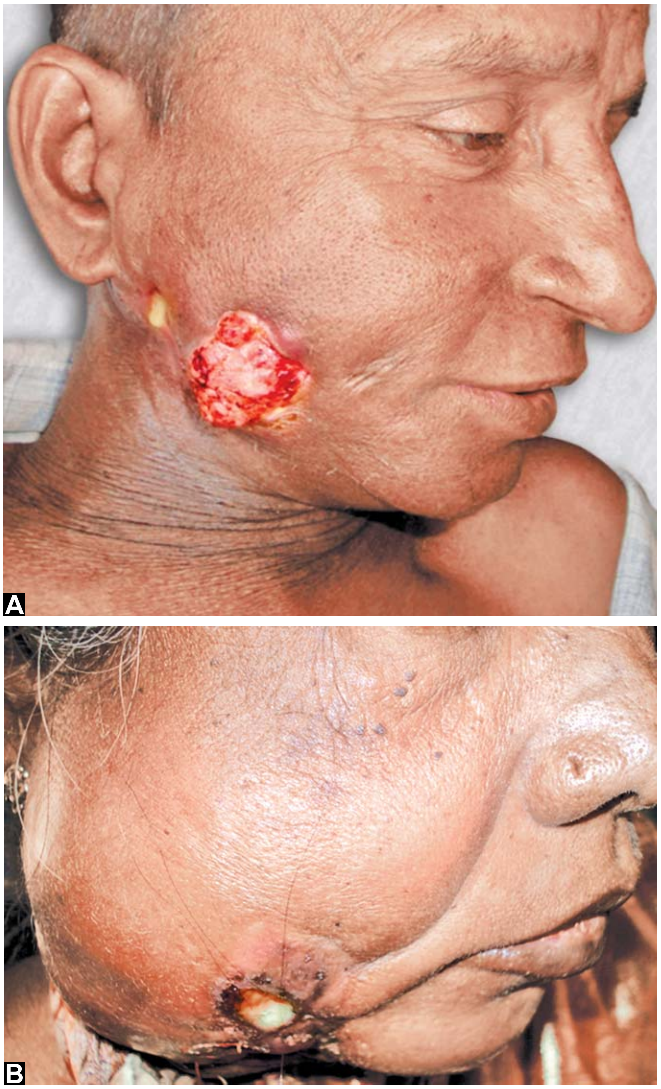

Figs 1.524A and B: Carcinoma cheek—advanced. It needs incision biopsy/orthopantomogram X-ray/CT of the part.

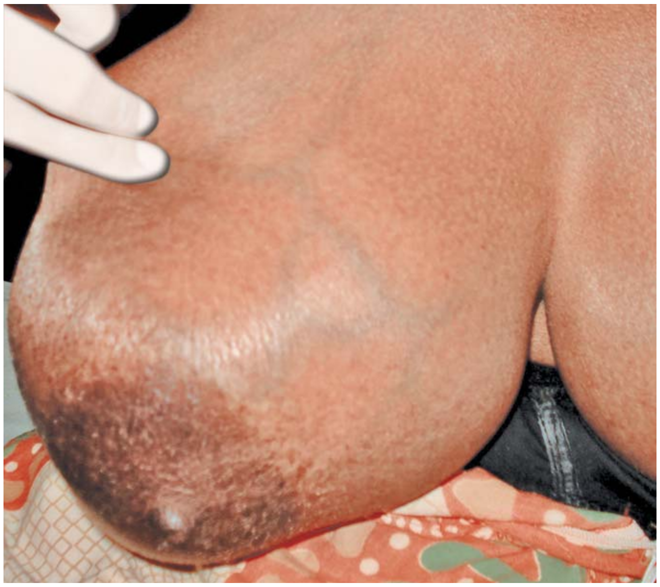

Fig. 1.526: Carcinoma breast. Note the dilated veins on the surface. FNAC should be done to confirm the diagnosis.

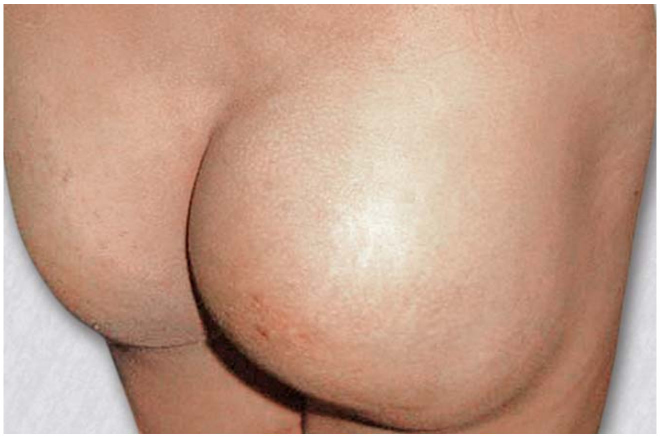

Fig. 1.527: Soft tissue tumour of buttock. It could be sarcoma.

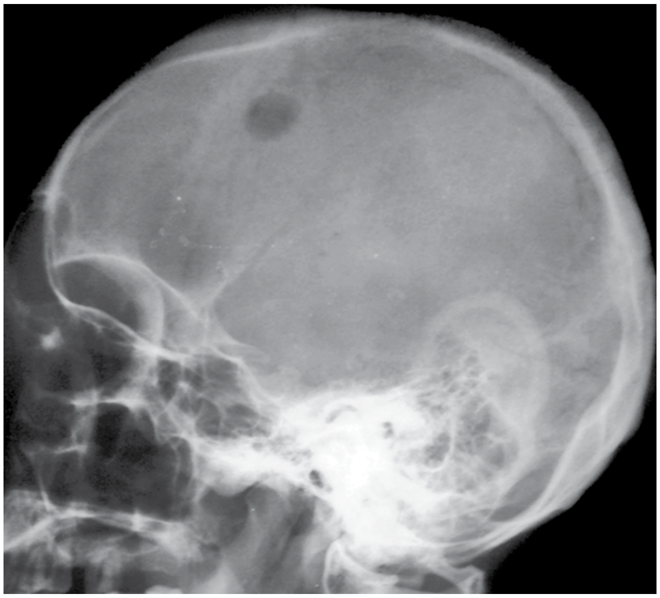

Fig. 1.525: Secondaries in skull—solitary.

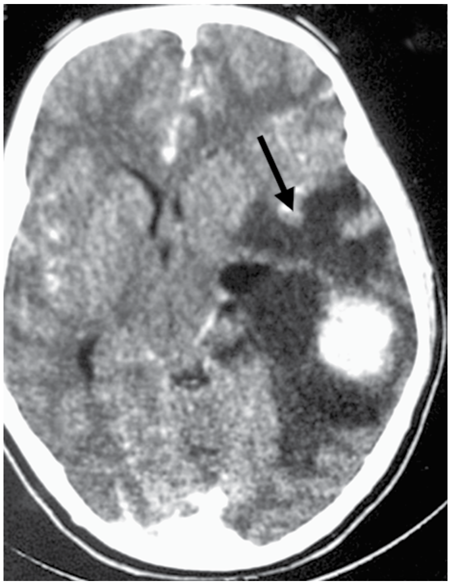

Fig. 1.528: CT picture of brain tumour—astrocytoma.

The pathological fracture is a transverse one, and although it unites readily, refracture is not unusual.

—Max E Lake

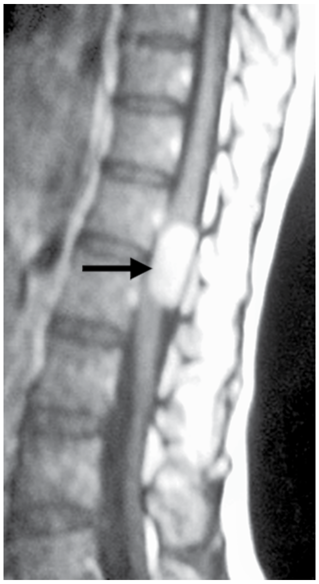

Fig. 1.529: Contrast MRI of spinal column showing extradural schwannoma.

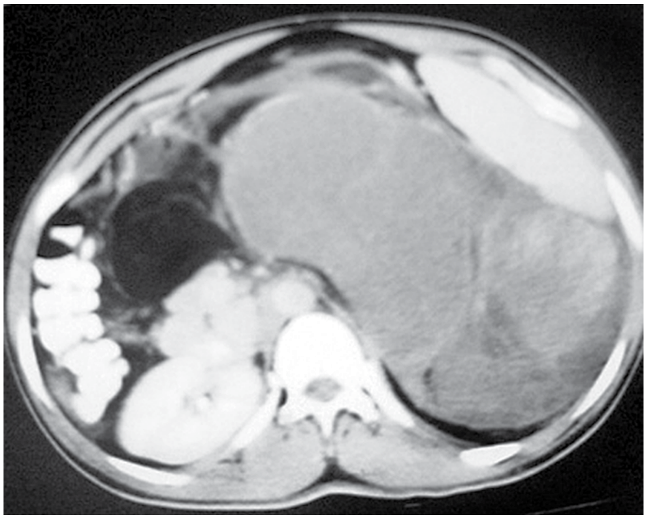

Fig. 1.530: CT scan picture of retroperitoneal leiomyosarcoma.

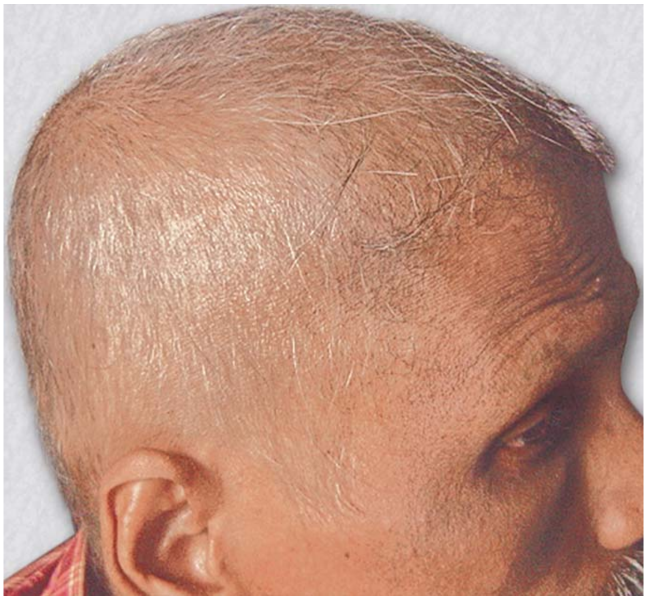

Fig. 1.531: Alopecia (hair loss) is common after chemotherapy for cancers.

Types of biopsies

- Incision biopsy
- Excision biopsy
- Trucut biopsy
- Pap smear
- FNAC
- Frozen section biopsy
- Punch biopsy
- U/S guided biopsy
- Laparoscopic biopsy
- CT guided biopsy
- Thoracoscopic biopsy
- Endoscopic biopsy (gastroscopic or colonoscopic or through ERCP or through cystoscopy)
- Proctoscopic biopsy
- Open biopsy either laparotomy or thoracotomy or craniotomy using Dandy's brain cannula

- Biopsy is sent in 10% formaldehyde solution. In special occasions like, to assess receptors/to do histochemistry, biopsy is sent in low temperature; in normal saline or in specialised ingredients. Tissues are kept in formalin for 24 hours and then taken for 'cut up' by pathologists. Sections are done after making tissue blocks using microtome up to 5 microns thickness. These sections are placed on a slide for staining with haematoxylin and eosin. This is studied under microscopy by pathologist to give the report. Detailed history, findings, markings of the specimen should be done by the surgeon prior to sending the specimen. Bone specimens are decalcified in hydrochloric acid for 7-21 days before sectioning and staining. So its report will be delayed usually.
- Inadequate sampling and improper reports may be due to—limited tissue sent; tissues are nonviable; too superficial biopsy done; cautery use on the tissues; crushing effect on the tissues (gentle meticulous biopsy is important without rough handling and without over use of cautery on the required tissues).
- Specimens are kept for 6 weeks. Blocks are kept for 30 years. Slides are kept for 10 years.
- Risk of false positivity in malignancy should be remembered. It is probably due to contamination, interchanged specimens, wrong interpretation, ulceration, etc. Re-biopsy/reinterpretation/repeat sectioning may be needed in such situations.
- Often additional methods like deeper sections; extra blocks; special stains are needed. Special stains are PAS (Periodic Acid Schiff) for glycogen, mucin and fungi; D (diastase) PAS for mucin; Perls Prussian blue for iron in haemochromatosis; reticulin for fibrous tissue; elastin stains for fibrosis; congo red for amyloid; Ziehl-Neelsen for mycobacteria; Grocott for fungi; Giemsa stain for protozoa; Warthin-Starry stain for spirochaetes.

Incision Biopsy

- It is taken from the edge of the lesion as in ulcer, not from the centre as there is necrosis.
- Usually *two biopsies* are taken in ulcerative lesion from the edge—edge/wedge biopsy (edge is a line which limits

two parts; wedge is an area at two meeting parts like thick to thin, etc.

- ♦ Incision biopsy is contraindicated in a case of melanoma where excision biopsy is preferred.
- ♦ In secondaries in lymph nodes, FNAC is preferred. If it fails to give information, incision biopsy is done.

Note:

Biopsy taken from the centre of the lesion is only in post-radiation ulcer and syphilitic gumma. Because of irradiation there is no blood supply in the margin and tumour proliferates in the centre of the lesion.

Excision Biopsy

In small lesions excision biopsy is done, e.g. lymph node biopsy is done in case of lymphoma.

Trucut Biopsy

It is done using a specialised device wherein gun with trucut tip is inserted into the surface tissue/organ and gun is fired to close the punching tip of the needle to catch and cut adequate tissue. It is done in prostate, breast and surface tumours. It cannot be done to deeper tissues or tissues which are close to major vessels/structures.

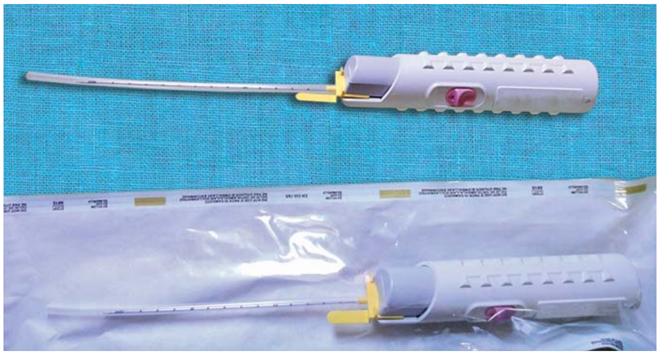

Fig. 1.532: Trucut biopsy gun used for trucut biopsy. It gives histological diagnosis; it is also helpful to assess histochemistry.

FNAC (Fine Needle Aspiration Cytology)

It is cytological study of tumour cells to find out the disease and also to confirm whether it is malignant or not.

Procedure: It is done using ***23*** or ***24*** gauge needle fixed to specialised syringes which creates negative pressure for aspiration and contents are smeared on the slides. Dry slides as well as slides fixed with 100% methanol are used for study.

- ♦ It is done in parotid, thyroid, lymph node, breast and all other surface lesions.
- ♦ In follicular carcinoma of thyroid it is not very useful, as angioinvasion and capsular invasion which are specific cannot be detected.
- ♦ In lymph nodes it is useful for detecting secondaries and tuberculosis.
- ♦ U/S guided or CT guided FNAC are popular at present—especially when it is done from liver, lungs, kidneys, etc. (from deeper structures/organs).

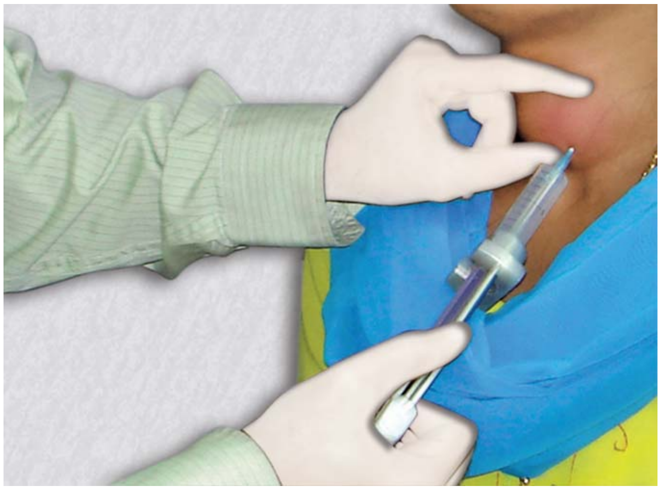

Fig. 1.533: Fine needle aspiration cytology (FNAC) of thyroid. It is also done in breast, lymph node and other swellings. It is not done for testicular tumour.

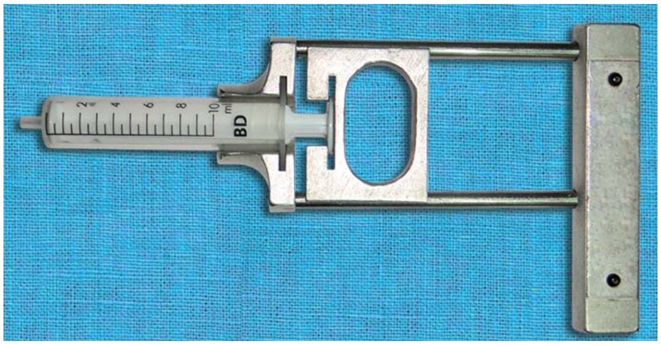

Fig. 1.534: FNAC vacuum creator with loaded syringe [FNAC gun; *Cameco syringe* (Sweden) holder]. (*Courtesy:* Dr Krishna Upadhya, MD, Nandikoor Laboratory, Mangalore)

It is absolutely contraindicated in testicular tumour. Because tough tunica albugenia usually prevents tumour spread and once it get disrupted by FNAC, spread can occur.

Cytological study is done after Papanicolaou stain; Giemsa staining or Romanowsky staining.

Note:

- In 1934, Martin and Ellis did **FNAC** using 18 G needle with syringe with local anaesthesia use.
- Kline and Neal in 1973 did syringe needle aspiration using 18 G needle making dry smears.
- In 1982, A John Webb detailed standard FNAC using 21 G needle. Needle is passed obliquely into the lesion; constant suction is applied using ***Comeco syringe or braced thumb method***. With suction persisting needle is moved in 4 directions. Suction is gently released and needle is withdrawn. Needle is detached from syringe to draw air which is reattached to withdrawn needle to blow the content over a dry slide.
- In 1987, A Zajdello et al fine needle sampling of the tissue in question *without aspiration*. 25 G needle is used without a syringe. Needle is moved in various directions to detach the cells by sharp end of the needle; by capillary force, cells are conducted into the needle lumen. Air filled syringe is attached to needle to express cells into a glass slide. Trauma to tissue is said to be less; and feel of tissue consistency is better through held needle. It is called as Fine Needle Sampling Without Aspiration / Fine

To different minds, the same world is a hell, and a heaven.

Needle Non-aspirating Cytology (*FNNAC*)/Fine Needle Capillary Sampling (*FNCS;* as capillary action causes conduction of cells into needle).

- US guided/MRI guided/CT guided/Mammographic guided FNAC are available now which are more precise versions of localization of lesions to be aspirated.
- Ether with alcohol in equal proportion is used for wet fixation to stain Papanicolaou or H & E stains. Grace in 1994 used modified ultra Pap stain 3 changes – by rehydration of air dried smears using normal saline to cause RMBC haemolysis to restore cell transparency; by using mixture of 4% formaldehyde and 65% ethanol to reduce fixation time from minutes to few seconds; by using Richard Allen Haematoxylin 2 and cyto stain. Total time of staining in this modification is 90 minutes.

Advantages

- Very sensitive.
- Done on OP basis.
- Least invasive, safer, fast and cheaper.
- Anaesthesia is not required.
- Tumour dissemination through the track is not present (except in testicular tumour where it is not done).
- Presently DNA study of the cell can detect very early malignancy.

Disadvantages

- Negative result cannot rule out malignancy.
- Tissue study is not possible.
- Further studies not amenable.

***Note*:**

FNNAC—Fine Needle Non-aspirating Cytology is done in some tissues (in some centres). Here needle is passed into the tissue and material collected from the needle (without any aspiration) is studied for cytological analysis.

Frozen Section Biopsy

Frozen section is done whenever biopsy report is needed at the earliest. It is usually done in a pathology set up existing adjacent to the operation theatre. An unfixed fresh tissue is frozen (using CO_2) in a metal and sections are made and stained. It is technically difficult; processing and staining is of inferior quality and often it is difficult to give accurate results. But advantage is it is quick and surgeon can decide the further steps of procedure in same sitting like nodal clearance/type of resection to be done, etc.

- It is done in carcinoma breast or in follicular carcinoma of thyroid when FNAC fails.
- During surgery after resection of the tumour to look for (on table) the clearance in the margin and depth and also to study the lymph nodes for their positivity.

II. *IMMUNOHISTOCHEMISTRY*

It is detection of specific antigen using an antibody. Antibody labelled with a dye, binds to antigen in a section of tissue causing specific colours like brown, and determines its presence and distribution in the tissues. It is safe, specific and quick.

It is used in:

- *Neoplasia*
  - To confirm, to find differentiation, to detect metastases; to plan type of therapy.
  - To categorise leukaemias or lymphomas; either of 'T' cell or of 'B' cell type.
  - To find out site of origin of metastatic tumours, e.g. by doing prostate specific antigen in prostate cancer or doing thyroglobulin in thyroid carcinomas or clear cells in the lung in metastatic renal cell carcinoma.
  - Detection of receptors or molecules, e.g. estrogen receptors [ER] in breast cancer. ER positive has got better prognosis. Presence of onco-protein product c—erbB2 in breast cancer signifies poor prognosis.
  - Gastrointestinal stromal tumour (GIST) can be treated with imatinib if it shows CD 117 expression.
  - Endocrine tumours are assessed on ki67 proliferative index.
  - Cytokeratin marker for epithelial tumors like carcinoma; S100 and actin for glandular malignancy; S100 for melanoma; CD31 and CD34 for vascular malignancy; CA 125 for ovarian tumours; PSA for carcinoma prostate; CD3 and CD20 for lymphoma.
- *Infections*
  - Detection of specific antibodies to the antigens of certain specific infective agents like Epstein-Barr virus, cytomegalovirus, human herpes virus 8, herpes simplex virus, etc.
- *To identify abnormal accumulation* of various proteins like ά-1-antitrypsin; deposition of amyloid.
- *Assessment of immunoglobulins*.
- *Screening for mutations*.

III. *TUMOUR MARKERS*

- Tumour markers are biochemical indicators of presence of a tumour. They are not used for primary pathological diagnosis.
- They are of prognostic value.
- Presence of tumour marker signifies recurrence or residual tumour.

IV. *OTHER METHODS*

Electron Microscopy

- It is visualisation of tissue in very high magnification of 1000 × 500 000 in difficult deciding cases. But it is time consuming and expensive.

In Situ Hybridisation

- It is determination of presence or absence of a specific gene and its location in a fresh/fixed tissue sections using an oligonucleotide probe, targeting at specific DNA or RNA sequence.

Polymerase Chain Reaction (PCR)

DNA is amplified using this special method and detected by technique like electrophoresis. PCR can be done on blood or nontissue samples. It is highly sensitive, fast and safe. Disadvantages are—expensive and risk of DNA contamination.

Uses

- Detection of mutations in congenital conditions like von Hippel-Lindau disease (vHL gene); haemochromatosis (HFE gene); colorectal carcinoma (APC gene).

Tumour markers

Category 1	
***Hormonal*:**	
Human chorionic gonadotrophin (HCG)	— Trophoblastic tumour, nonseminomatous testicular tumour
Calcitonin	— Medullary carcinoma of thyroid
Catecholamines and VMA	— Pheochromocytoma
Ectopic hormones	— In tumours of paraneoplastic syndromes
Category 2	
Isoenzymes:	
Prostatic acid phosphatase	— Carcinoma prostate
Neuron specific enolase	— Small cell carcinoma lung, neuroblastoma
Category 3	
Oncofetal antigens:	
Alpha fetoprotein	— Liver cancer, nonseminomatous germ cell tumour
Carcinoembryonic antigen (CEA)	— Carcinoma colon (common). Carcinoma pancreas, lung, stomach and breast
Category 4	
Mucin and other proteins:	
CA—125 (Carbohydrate antigen)	— Ovarian cancer
CA—15-3	— Breast cancer
CA—19-9	— Pancreatic and colon cancer

- Gene rearrangements in different conditions like lymphoma or other diseases can be detected by PCR (clonality).
- Assessment of loss of heterozygosity in tumours like oligodendroglioma (1p, 16q).
- PCR study for infections like tuberculosis.
- Fluorescence *in situ* hybridisation (FISH)
- It is a type of cytogenetics (study of chromosomes) that can be done in fixed or fresh tissues; used in haematological malignancies or trisomy 23. It is safe and quick.

MANAGEMENT STRATEGY FOR CANCERS

- *Diagnosis* is confirmed by biopsy from primary usually.
- *Evaluations* for staging—metastatic work up.
- *Approaches for primary*—surgery/radiotherapy/chemotherapy.
- *Approaches for secondary.*
- *Palliation* in advanced stage—palliation of *distressing* symptoms.

Treatment of tumours

- Surgery—wide excision/radical
- Radiotherapy—curative/palliative
- Chemotherapy
- Immunotherapy
- Gene therapy

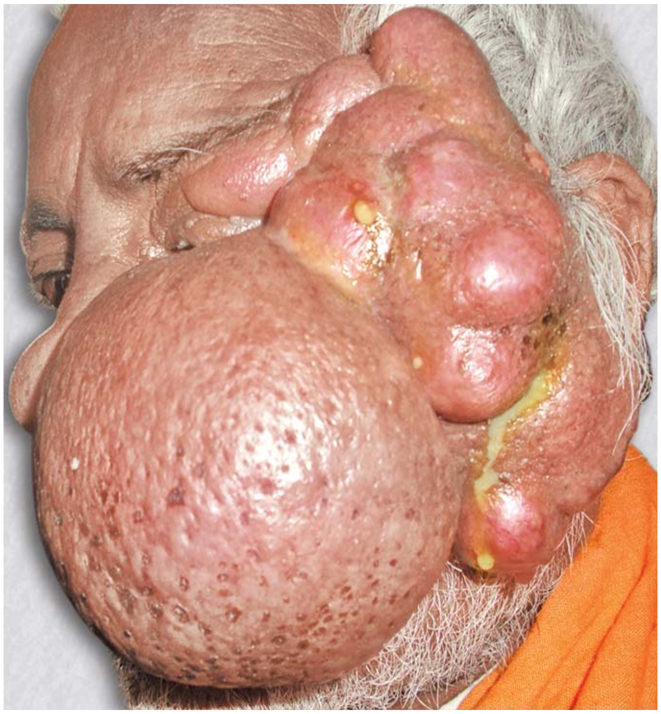

Fig. 1.535: Advanced aggressive malignancy face may be carcinoma or sarcoma.

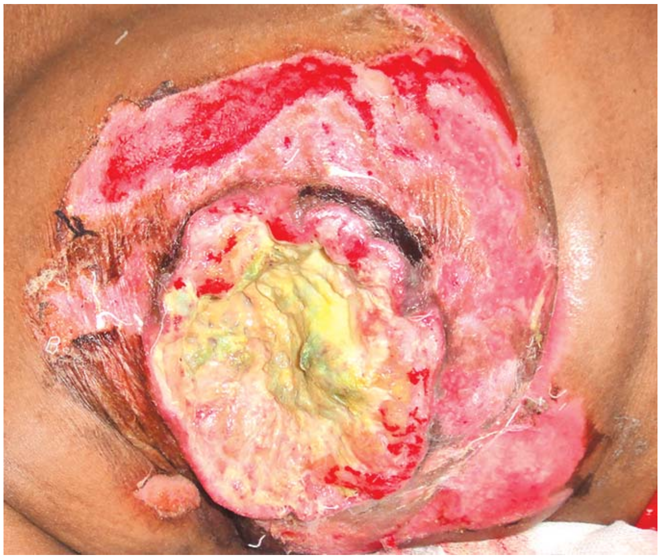

Fig. 1. 536: Advanced fungating carcinoma of breast.

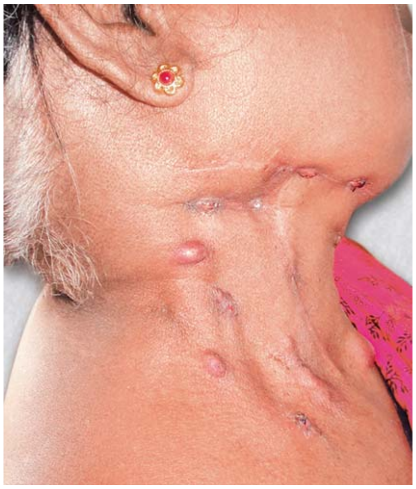

Fig. 1.537: Recurrent secondaries in neck after radical neck dissection.

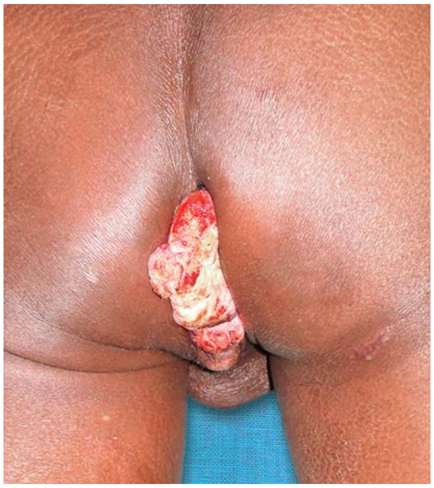

Fig. 1.538: Anal canal carcinoma (SCC).

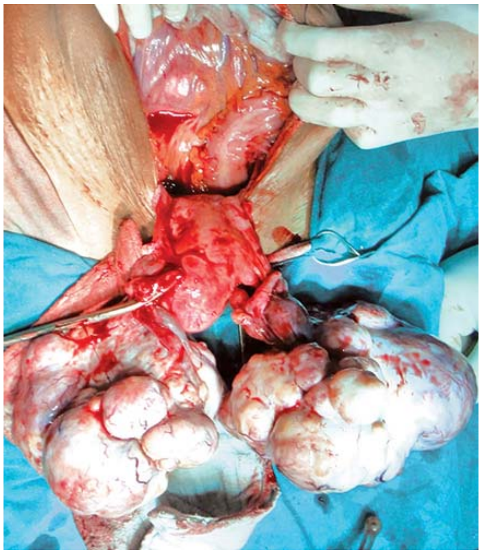

Fig. 1.539: Secondaries in ovaries—Krukenberg tumour.

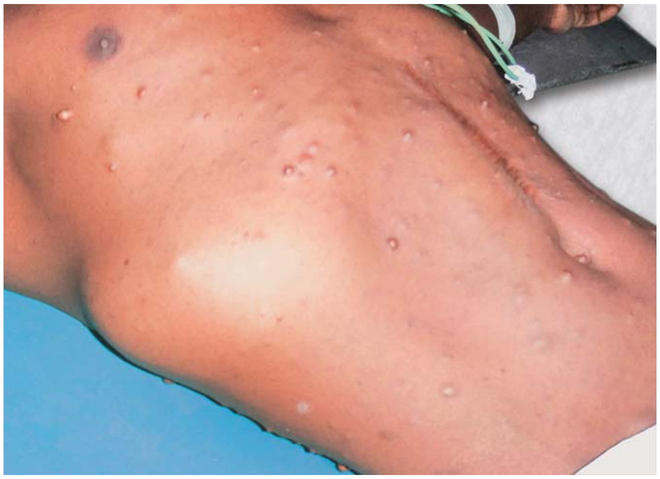

Fig. 1.540: Malignant primary nerve sheath tumour (MPNST).

Malignancies which are curable

- Basal cell carcinoma
- Adamantinoma
- Verrucous carcinoma
- Marjolin's ulcer
- Papillary carcinoma of thyroid
- Carcinoma colon

Sarcomas which spread to lymph nodes

- Synovial sarcoma
- Ewing's sarcoma
- Lymphosarcoma
- Kasposi's sarcoma
- Rhabdomyosarcoma
- Angiosarcoma

R. Skin Tumours

CHAPTER OUTLINE

- Anatomy
- Classification of Skin Tumours
- Skin Adnexal Tumours
- Dermatofibroma
- Dermatofibrosarcoma Protuberans
- Keratoacanthoma
- Rhinophyma
- Seborrhoeic Keratosis
- Squamous Cell Carcinoma
- Marjolin's Ulcer (1828)
- Basal Cell Carcinoma
- Turban Tumour
- Naevi
- Melanoma

ANATOMY

Epidermis: Layers

- Stratum corneum—1/3 of epidermis
- Stratum granulosum—1-3 layers of cells
- Stratum spinosum—3-5 layers of cells
- Stratum basale—columnar cells
- Stratum lucidum is seen in palms and sole.

Melanocytes—one melanocyte is seen for every 10 basal cells
Langerhan's cell—a clear cell in stratum spinosum.
Basement membrane is seen at dermo-epidermoid junction.

Epidermis is avascular.

Epidermis is 5% of the total skin. It is thickest (0.5-1 mm) in palm, sole, back and buttocks. It is thinnest in eyelids (0.05-0.09 mm).

Dermis—contains collagen fibres, elastic fibres, capillaries, venules, arterioles, lymphatics, nerves, erector pilorum muscle, sweat glands (eccrine, apocrine), sebaceous glands. Merkel cells, Meissner and Pacinian corpuscles are receptors in dermis. Dermis constitutes 95% of the skin. Dermis is 15-40 times thicker than epidermis.

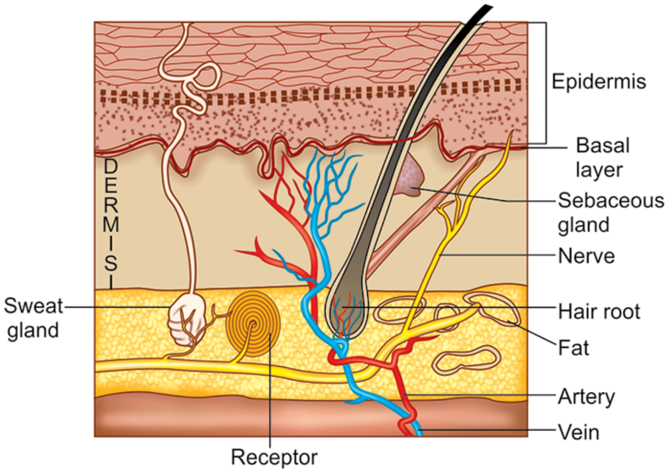

Fig. 1.541: Skin anatomy.

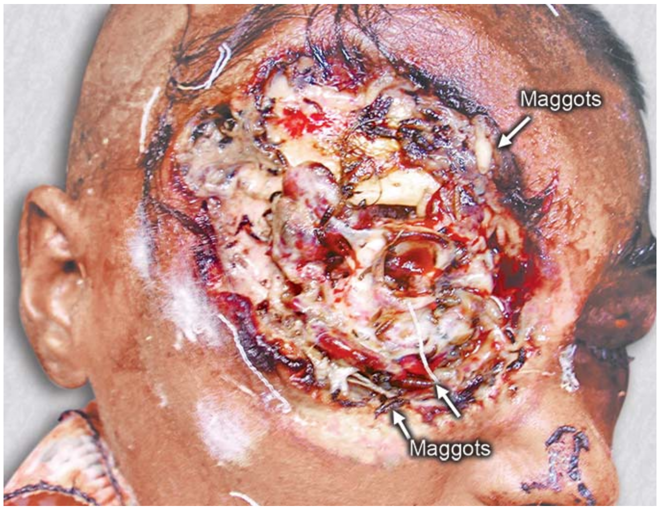

Fig. 1.542: Extensive infective destruction of the eyelids, face exposing bone and deeper structures. Maggots are seen on the surface. It mimics carcinoma but it was of infective origin. *Phagedena* is infective destruction without proliferation. It often mimics malignancy.

CLASSIFICATION OF SKIN TUMOURS

Benign

Epidermal

- Seborrhoeic keratosis.
- Trichilemmal tumour.
- Sebaceous adenoma.
- Sebaceous epithelioma.
- Hydrocystoma, syringoma, spiradenoma.

Probability is the rule of life, especially under the skin. Never make a positive diagnosis.—**William Osler**

Dermal tumour

- Neurofibroma.
- Dermatofibroma.

Malignant

- Squamous cell carcinoma.
- Basal cell carcinoma.
- Melanoma.
- Malignant skin adnexal tumour.
- Secondaries in the skin. Sister Joseph nodules around umbilicus.

Classification of skin tumours

Epidermal
Benign—papilloma, seborrhoeic keratosis
Malignant—BCC, SCC
Melanocytic
Benign—all types of naevi
Malignant melanoma
Skin adnexal tumour
Benign—syringoma, hidradenoma, sebaceous adenoma, trichofolliculoma, trichilemmoma
Malignant—hidradenocarcinoma, sebaceous carcinoma
Dermal tumours
Neurofibroma, dermatofibroma, dermatofibrosarcoma protuberans

- *Skin cancers* (SC) *are also classified as melanotic (MSC) or nonmelanotic (NMSC)*. NMSC are *commonest*. Patient who had BCC/SCC has higher risk to develop 2nd new skin cancer (35% in 3 years; 50% in 5 years). *NMSC can be low risk or high risk groups*. Lesion more than 2 cm in trunk and limbs; more than 1 cm in forehead and neck; more than 6 mm in central face; poorly defined margin; recurrent type; moderate or poor differentiation; perineural/vascular invasion; presence of immunosuppression; previous RT – are *high risk* lesions.

Note:
- Multiple sebaceous adenomas with visceral malignancy are called as *Torre-Muir syndrome*.
- Secondaries in skin can occur like ***Sister Joseph nodules*** around the umbilicus.

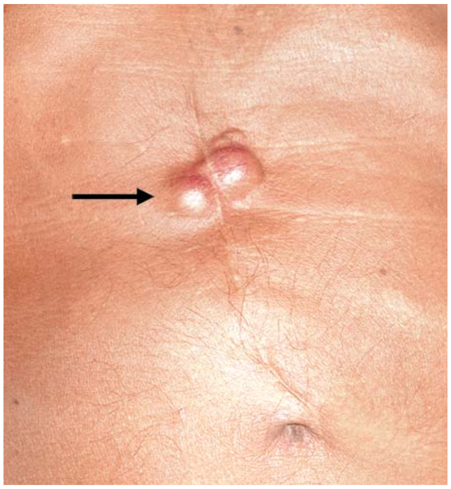

Fig. 1.543: Abdominal wall cutaneous secondaries; primary is from carcinoma stomach.

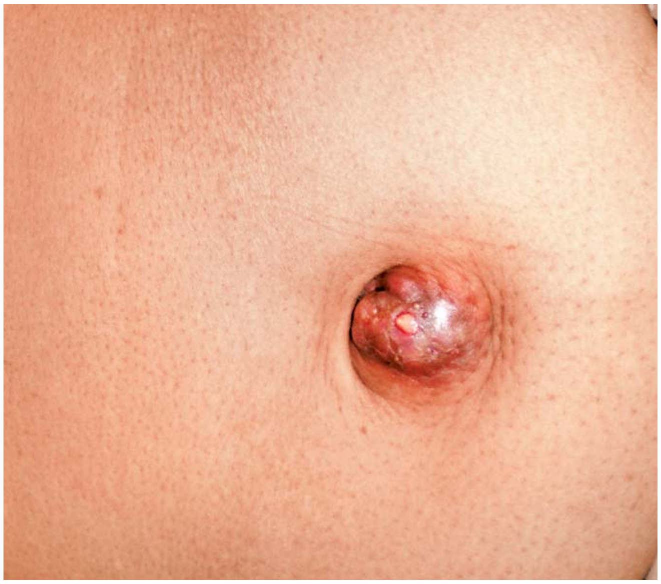

Fig. 1.544: Sister Joseph secondary nodule in umbilicus.

SKIN ADNEXAL TUMOURS

Classification

- *Eccrine gland tumours*: Syringoma, hidradenoma, syringocystadenoma.
- *Hair tumours*: Trichoepithelioma, tricholemmoma.

Features

- They are tumours arising from accessory skin structures like sebaceous glands, sweat glands, hair follicles.
- It is not uncommon. It may be **benign** or **malignant**.
- It presents as protruding well-localised swelling in the skin.
- *Trichoepithelioma* is usually seen in nasolabial fold; mimics BCC; presents as small cutaneous nodule.
- *Tricholemmoma* is also called as *naevus sebaceous of Jadassohn*. It is a hamartoma from hair follicle which can turn into BCC in 10% of cases.

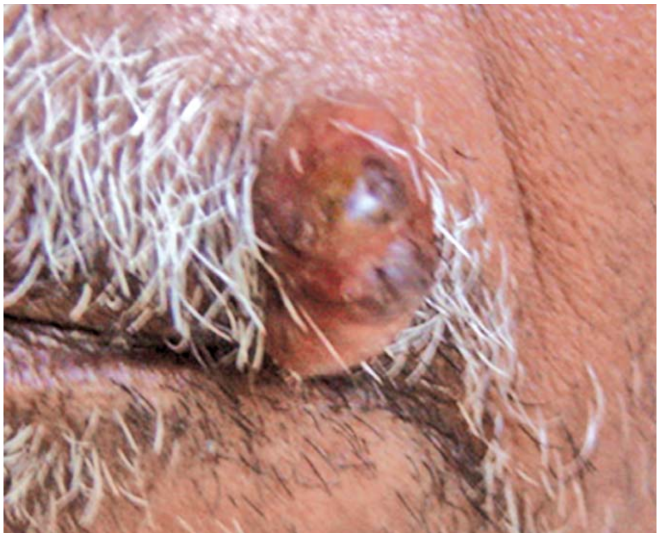

Fig. 1.545: Skin adnexal tumour (benign) in the face.

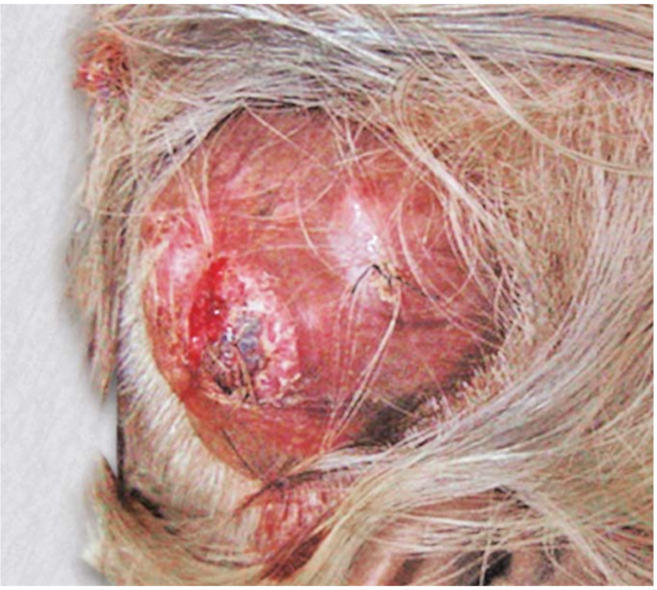

Fig. 1.546: Skin adnexal tumour (malignant) turned out to be of hair follicle origin.

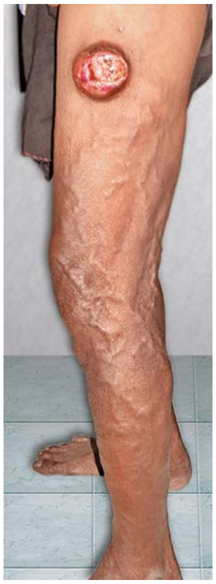

Fig. 1.548: Skin adnexal tumour. Vascularity is increased and it is malignant.

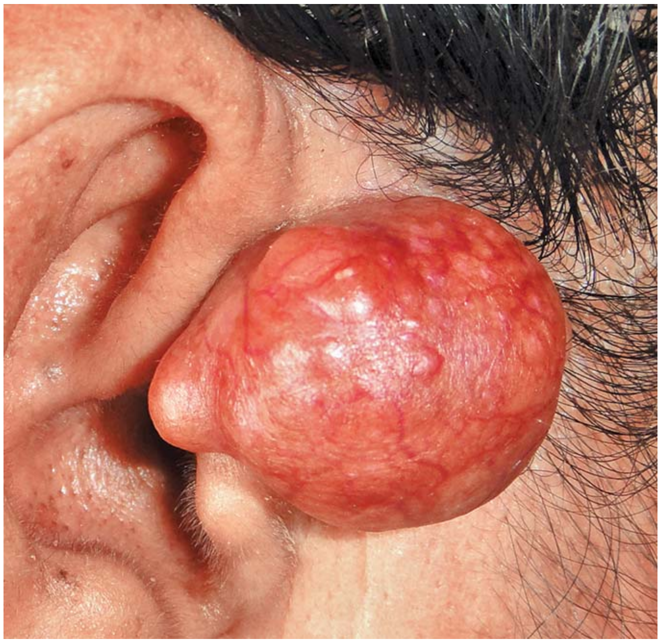

Fig. 1.547: Skin adnexal tumour—preauricular region.

- *Adenoma sebaceum* is seen in tuberous sclerosis as red papules in face which appears below 10 years. It is often called as *Bourneville's disease*.
- *Calcifying epithelioma of Malherbe/pilomatrixoma* is benign hair matrix cell tumour seen below the age of 10 years, containing basaloid and eosinophilic ghost cells with calcification.
- *Malignant skin adnexal tumour* forms a nodular, hard, indurated swelling in the skin, often with involvement of regional **lymph nodes** which are hard and nodular**.**
- It mimics **squamous cell carcinoma of skin.**

Differential Diagnosis

- Squamous cell carcinoma of skin.
- Dermatofibrosarcoma protuberans.

Diagnosis

- Biopsy—excision or incision type.
- FNAC of lymph node.

Treatment

- Benign tumour—excision.
- For malignant tumour—wide excision and regional lymph node block dissection when required.

Prognosis is good. It is better than squamous cell carcinoma.

Note:

- **Merkel cell carcinoma** is aggressive malignant condition arising from neuroendocrine receptor cells of the skin (dermis) which mimics histologically oat cell carcinoma. It is common in white elderly females (4 : 1) may be due to UV rays. Treatment is wide excision with radiotherapy.
- **Adenocarcinoma arising from the apocrine glands** of skin is called as *extramammary Paget's* disease of skin (intraepidermal adenocarcinoma) commonly observed in perianal region.
  - It can occur in genitalia or in axilla (more apocrine glands). In 25% of cases, the condition is associated with an *underlying in situ* or invasive carcinoma. *Presentation* is like red plaque/white/depigmented areas/crusts/scales mimicking dermatitis, eczema, fungal infections. Condition is often associated with GI or urinary malignancies (40%). Biopsy of lesion, CT evaluation for other malignancies, wide local excision and radiotherapy—are the management principles.

DERMATOFIBROMA (Sclerosing Angioma or Subepithelial Benign Nodular Fibrosis)

- It is a benign tumour arising from skin from dermal dendritic cells.
- It is formation of firm, single or multiple nodules occurring commonly in extremities (limbs).
- It can be red, brownish yellow (due to lipid), or bluish black (due to haemosiderin).

  Histologically, spindle cells are arranged in 'mat like' or 'cart-wheel' pattern.
- It is also called as ***dermal histiocytoma.***

Differential Diagnosis

- Squamous cell carcinoma of skin.
- Melanoma.
- Basal cell carcinoma.
- Skin adnexal tumour.

Treatment

- Excision.

DERMATOFIBROSARCOMA PROTUBERANS

- It is a low grade fibrosarcoma which grows slowly but persistently. It arises from *dermal fibroblasts*.
- Occurs in head and neck, limbs, abdominal wall and back–Trunk is commonest site (50%).
- It is not a rare entity, often attains a large size with multiple, nodular, hard, swelling with often involvement of lymph nodes. Malignant spindle cells are seen histologically.
- Rarely it spreads into lungs through blood.
- It mimics squamous cell carcinoma of skin and skin adnexal tumour.
- Positive for CD34 and ring chromosome.
- With melanin pigmentation it is called as *Bednar's tumour*.

Diagnosis

- Biopsy of the lesion.
- Chest X-ray, CT Scan.
- FNAC of the lymph node.

Treatment

- Wide excision and follow up.
- Recurrence is common—50%.
- Prognosis is good.

KERATOACANTHOMA (MOLLUSCUM SEBACEUM)

- It is an overgrowth and subsequent spontaneous regression of ***pilosebaceous glands*** with proliferation of squamous cells protruding out of the duct which are common in adult males (3:1) and places where more sebaceous glands are found.
- Cause is unknown. It may be self limiting, benign neoplasm of viral origin (papilloma virus).

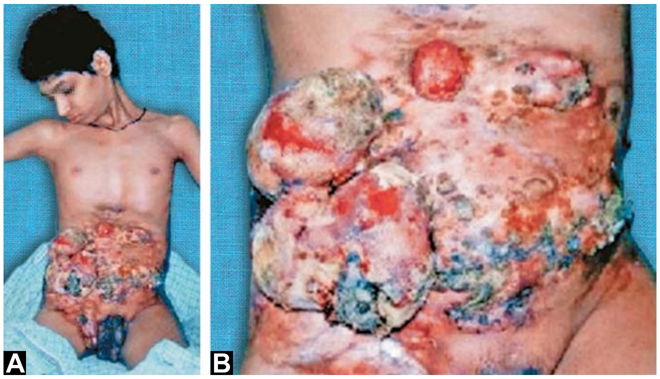

Figs 1.550A and B: Dermatofibrosarcoma.

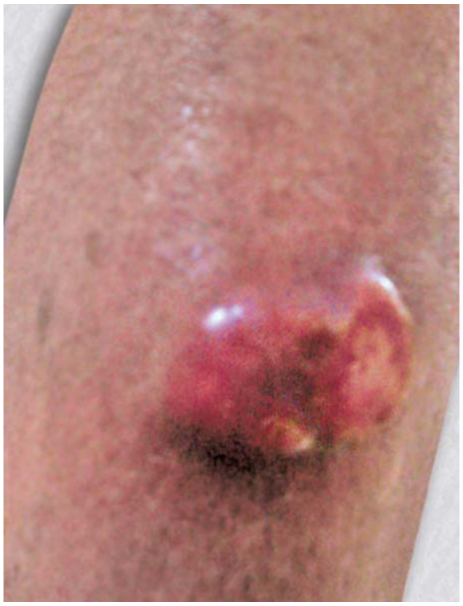

Fig. 1.549: Dermatofibroma in leg.

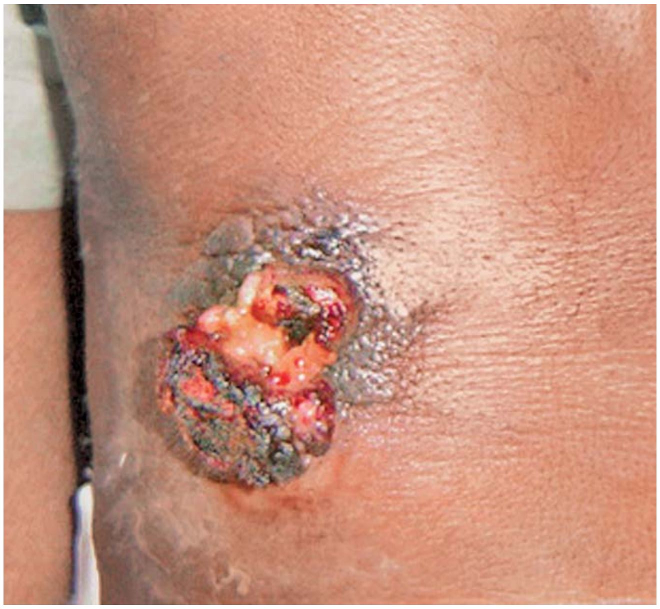

Fig. 1.551: Malignant skin tumour with ulceration of the skin over chest wall. Axillary nodes were enlarged in this patient.

- It presents as a rapidly growing, painless, single swelling in the skin with central brown area.
- It grows usually for 4 weeks and later shows spontaneous regression in 4 months. It is a pseudomalignancy.
- During regression phase, central area separates from the lesion leaving a deeply seated scar.

Clinical Features

- Mobile, hard, painless, nontender, lump with a *central brownish volcano* like area. It is common in face. It can be recurrent in lips and fingers.
- Lymph nodes are not enlarged.
- *It is totally benign.*

Differential Diagnosis

- Squamous cell carcinoma.

Treatment

- Excision. The tissue is sent for histopathological study.

RHINOPHYMA (Potato Nose) (Bottle Nose)

- It is a **glandular form of acne rosacea** causing immense thickening of **distal part of skin of nose** with visible openings of sebaceous follicles. Nose is bluish red in colour with dilated capillaries.
- It is due to hypertrophy and adenomatous changes in sebaceous glands.
- Male to female ratio is 12:1. 3% cases may have occult BCC in it. But rhinophyma itself will not cause BCC.

Treatment

- Excision of excess tissue and reconstruction.

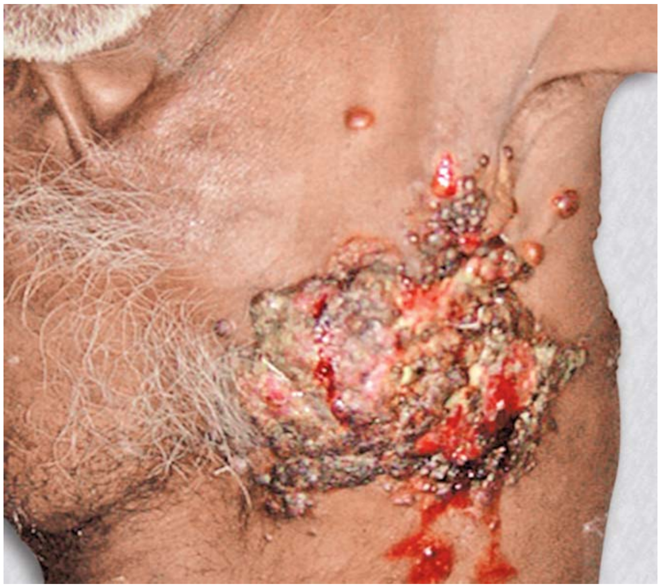

Fig. 1.552: Dermatofibrosarcoma protuberans.

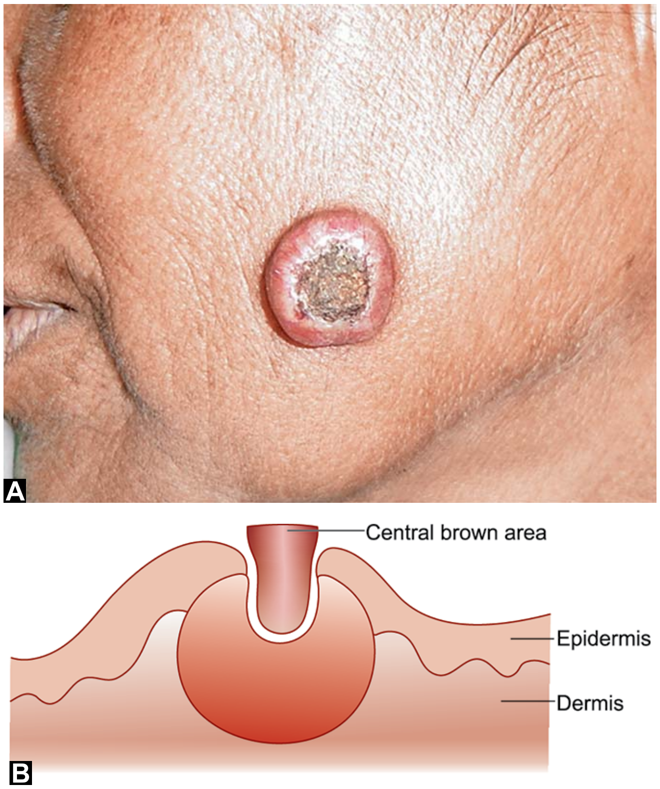

Figs 1.553A and B: Typical keratoacanthoma with central brownish area. It is self-limiting disease. It never turns into malignancy but looks like SCC.

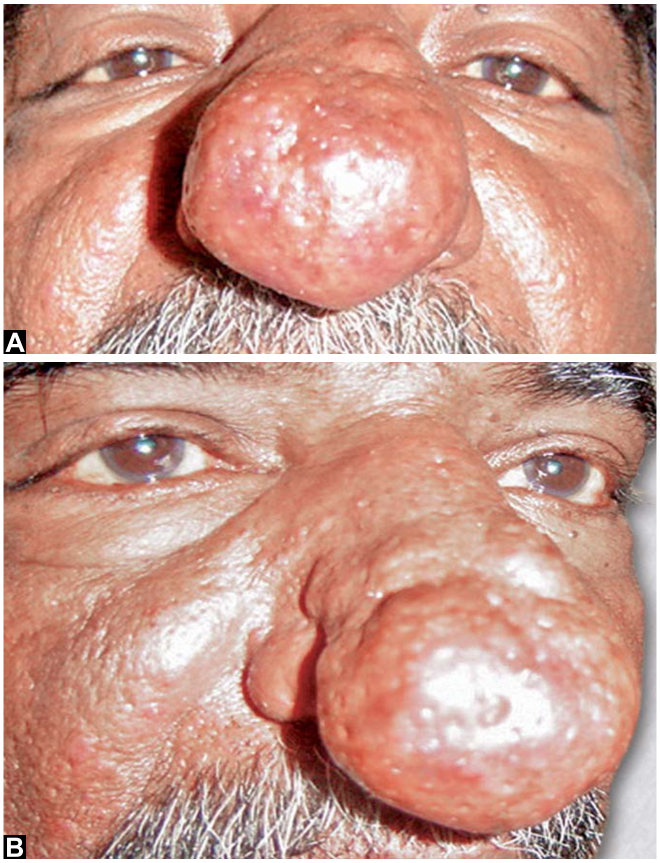

Figs 1.554A and B: Rhinophyma. Typical look and site.

SEBORRHOEIC KERATOSIS (Seborrhoeic Wart, Basal Cell Papilloma)

It is a benign overgrowth of the basal layer of epidermis with excess of small darkly stained basal ce1lls, which protrudes from the surface of the epidermis to give oily appearance.

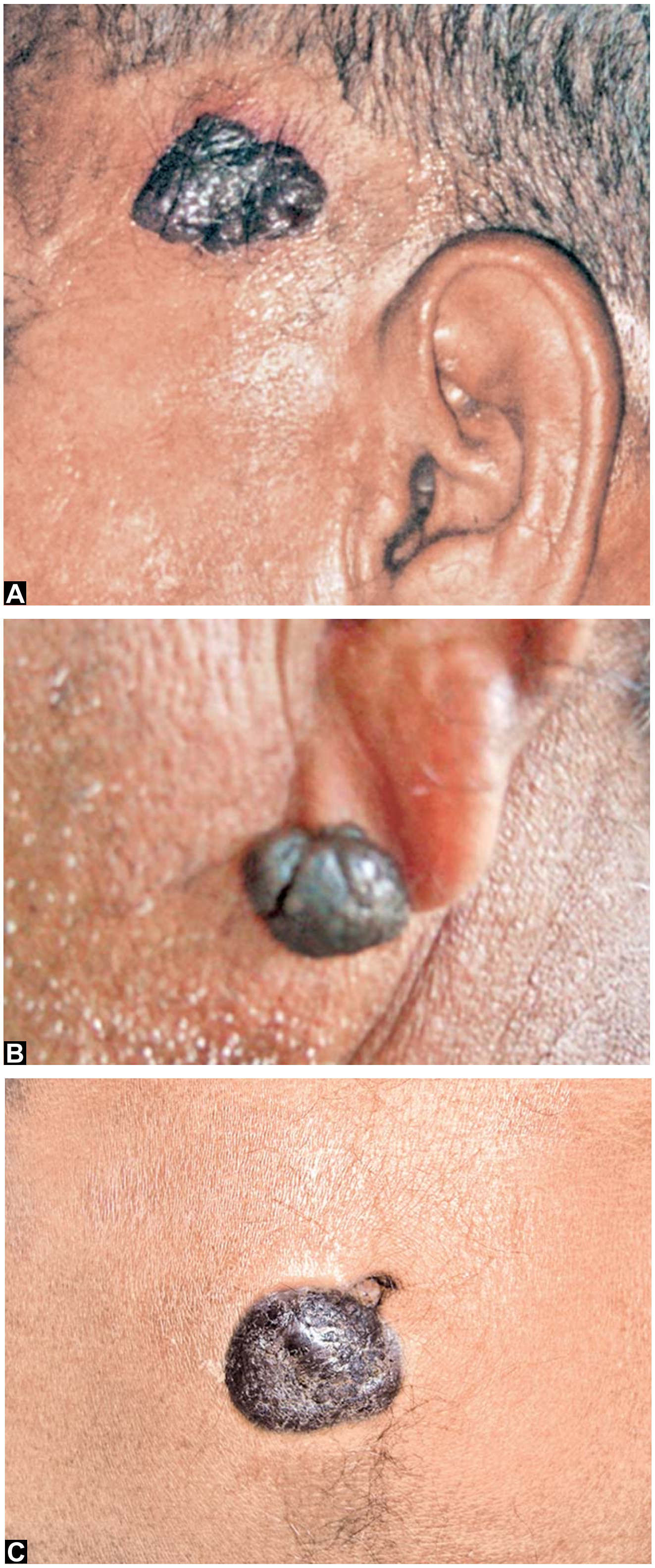

Figs 1.555A to C: Seborrhoeic keratosis in face, ear and in periumbilical region. It is not a premalignant condition.

Features

- It is common in elderly. Common sites are the back, face, neck.
- It grows slowly with widening in area without altering in thickness.
- It often gets infected but uncommon to bleed on touch.
- It is pigmented due to melanin and so mimics naevus or melanoma.
- It is common Caucasians. It is familial—autosomal dominant gene related.
- Often when it falls off, it leaves a pale pink patch on the skin with visible small surface capillaries.
- It is not a premalignant condition (Note: Solar keratosis is a premalignant condition).
- It is hard and stiffer than normal skin.
- Lymph nodes are not involved.
- It does not occur in palms and soles.
- It can be picked off from the skin.
- *'Stuck on' appearance is characteristic.*

Differential Diagnosis

- Melanoma, Pigmented BCC, Naevus,

Treatment

- Excision cures the condition.
- Shave excision or curettage can be done.

PREMALIGNANT CONDITIONS OF THE SKIN

- *Bowen's disease* of skin: It is an intradermal precancerous condition. It presents as brownish induration with a well-defined edge. Microscopically it contains large clear cells. Eventually, it will turn into carcinoma (10%). Entire epidermis is disorganised and irregular. It shows parakeratosis, acanthosis, hyperkeratosis. Chronic solar exposure; arsenic; human papilloma virus 16 are the aetiologies. Topical 5 fluorouracil or imiquimod; 4 mm margin surgical excision; MOHS; laser are the therapeutic options

  Erythroplasia of Queyrat is Bowen's disease occurring over glans penis
- *Paget's disease* of nipple
- Leukoderma
- *Senile* or *solar keratosis:* It is multiple, dry, hard, scaly, lesions on the face and back of hands due to exposure to sunlight, occurs after middle age. This is sunray induced hyperkeratosis with irregular, firm, raised or flat patch. Squamous cell carcinoma develops later after 10 years. Lesion in such situation becomes non-healing, indurated with central crust with everted edge; hard enlarged regional lymph nodes may be palpable
- *Radiodermatitis,* arsenic dermatitis
- *Chronic scars* develop into Marjolin's ulcer
- Albinism
- *Xeroderma pigmentosa* wherein there is defective DNA excision repair mechanism. It turns into malignant melanoma, BCC, SCC
- *Chronic lupus vulgaris*
- *Prolonged irritation of skin* by various chemicals like dyes, tar, soot

SQUAMOUS CELL CARCINOMA (Epithelioma) (Khangri Cancer) (Chimney Cancer)

- It occurs in premalignant conditions like Bowen's disease, Paget's disease, leukoplakia, chronic scars, chemically induced chronic irritation, radiodermatitis, senile keratosis, e.g. Khangri cancer in Kashmir, Chimney scrotal cancer, Kang cancer of Tibetans.
- It arises from squamous layer of the skin. Usually it occurs in a pre-existing predisposing lesion; occasionally can develop in *de novo* skin.
- It can be grossly proliferative/ulcerative/ ulceroproliferative/ red plaque like. Proliferative type is cauliflower like.
- It expresses cytokeratins one and ten.
- It is the 2nd (20%) most common skin cancer.
- It is common in males.
- *Exposure to UV B rays* (ultraviolet rays are A,B,C types) causes SCC by direct carcinogenic effects on keratinocytes, unrepaired mutations, decreased immune surveillance response, inhibition of tumour rejection, mutation of p53 suppressor gene (seen in 90% SCC).

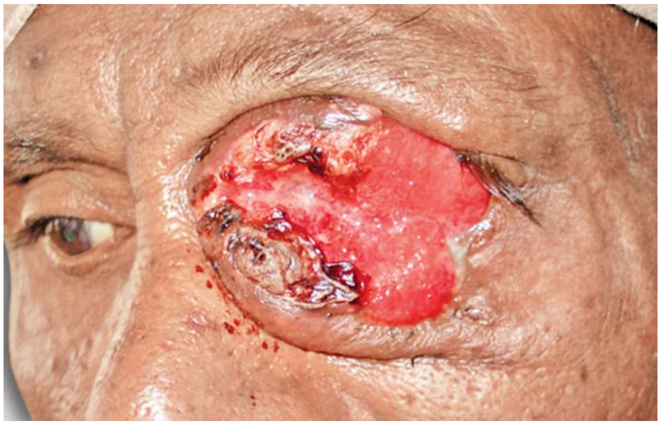

Fig. 1.556: Squamous cell carcinoma eyelid. Note the involvement.

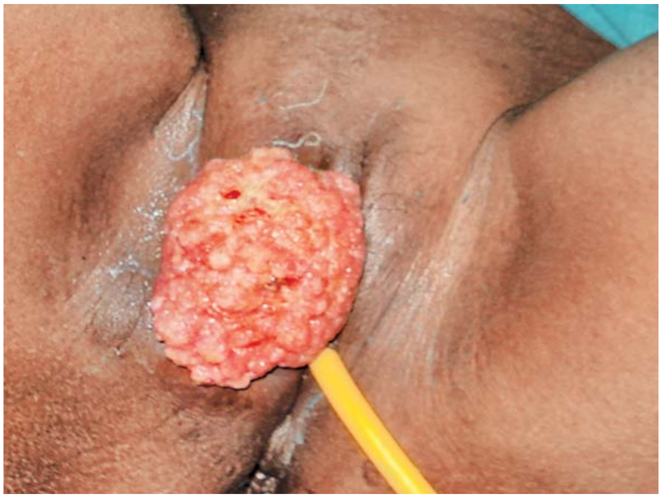

Fig. 1.557: Squamous cell carcinoma on the labia.

Aetiology of SCC

- Bowen's disease. Exposure to UV B rays
- Chronic scars and sinuses (burns; osteomyelitis; venous ulcer)
- Lupus vulgaris
- Solar keratosis—20% chances of SCC
- Senile keratosis
- Xeroderma pigmentosa
- Tobacco use
- Viral cause—human papilloma virus (HPV) five and sixteen
- Chemically induced chronic irritation
- Radiodermatitis
- ***Kangri cancer***—it is due to constant placing of the hot charcoal pot (kangri) over the abdominal wall to control cold. Seen in Kashmir
- ***Kang cancer*** is seen in buttocks and heel of Tibetans due to sleeping over oven bed to control cold
- ***Chimney sweep cancer*** is observed in scrotum due to constant irritation by tar
- *SCC is more common in immunosuppressed individuals*, immunosuppressive drugs like azathioprine, cyclosporine, prednisolone; risk becomes 10% in 10 years of intake of these drugs and 40% in 20 years.

Common Sites are:

- Dorsum of hand, limbs, face, and skin of abdominal wall SCC can occur in external genitalia, mucocutaneous junction, oral cavity, respiratory system, oesophagus, gallbladder, in urinary bladder as metaplasia from transitional cell lining.

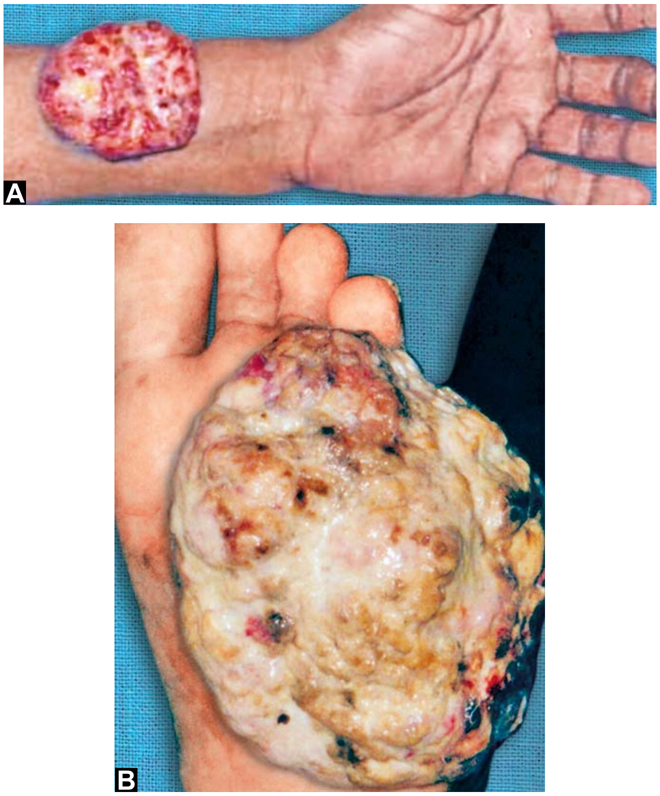

Figs 1.558A and B: Squamous cell carcinoma in the forearm and foot. It is proliferative cauliflower like lesion.

Finding fault is easier. Finding remedy is the one which is more difficult.

Clinical Features

- An ulcerative or ulceroproliferative lesion.
- **Raised** and **everted edge**.
- **Indurated** base and edge.
- Bloody discharge from the lesion.
- Regional lymph nodes are commonly involved, which are hard, nodular, initially mobile but eventually fixed to underlying structures.
- Usually blood spread does not occur.

Variants

- *Marjolin's ulcer* which occurs in chronic scar is a type of squamous cell carcinoma without lymph node spread.
- *Verrucous carcinoma* is a squamous cell carcinoma, commonly occurring in mucous membrane or mucocutaneous junction without lymph node spread. It is dry, exophytic, warty, indurated growth. It has good prognosis. It is a curable malignancy.
- A rare multiple self healing SCC is observed usually in face as familial autosomal dominant (Ch 9q) disease in Western Scotland-*Ferguson-Smith syndrome.*

Histology

- Malignant whorls of squamous cells with epithelial or **keratin pearls** are characteristic feature.
- Spindle cells, invasion, deep and peripheral margin clearance.

Broder's classification

I: Well differentiated: 75% or more keratin pearls
II: Moderately differentiated: 50-75% keratin pearls
III: Poorly differentiated: 25-50% keratin pearls
IV: Undifferentiated/anaplastic: <25% keratin pearls. It is seen in 20% of SCCs.

Differential Diagnosis

- BCC.
- Melanoma.
- Keratoacanthoma.
- Skin adnexal tumours.
- Actinic keratosis.
- Pyogenic granuloma.

Investigations

- Edge biopsy.
- FNAC from lymph node.

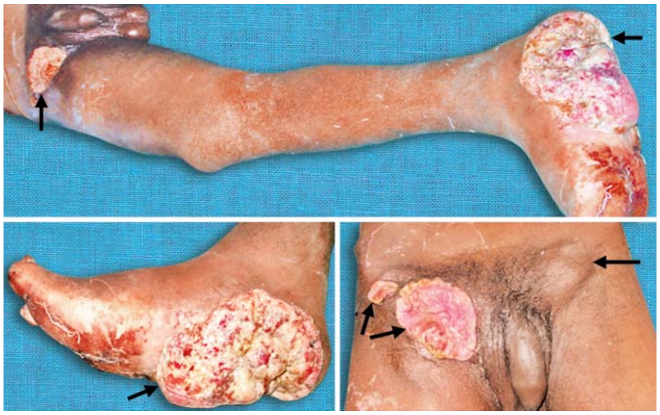

Fig. 1.559: Squamous cell carcinoma foot—ulceroproliferative lesion with fungating inguinal node secondaries. It is an advanced disease. Opposite inguinal nodes are also significantly enlarged.

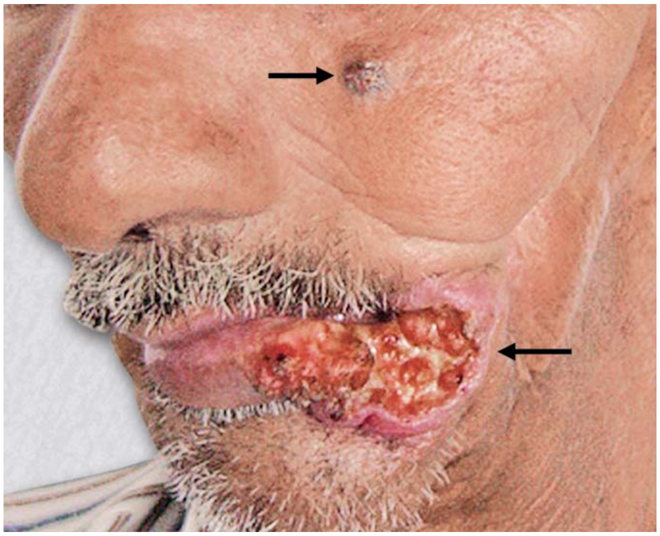

Fig. 1.560: Squamous cell carcinoma and BCC in face of a patient.

TNM staging for skin cancer other than melanoma

Tumour	***Node***	***Metastasis***
T 0 no tumour found	N 0–No nodes	M 0–No distant spread
T is Tumour *in situ*	N 1–Regional nodes ++	M 1–Distant spread ++
T 1 Tumour < 2 cm		
T 2 Tumour 2-5 cm	G1 – Low grade	
T 3 Tumour > 5 cm	G2 – Moderate grade	
T 4 Spread to cartilage, muscle or bone.	G3 – High grade	

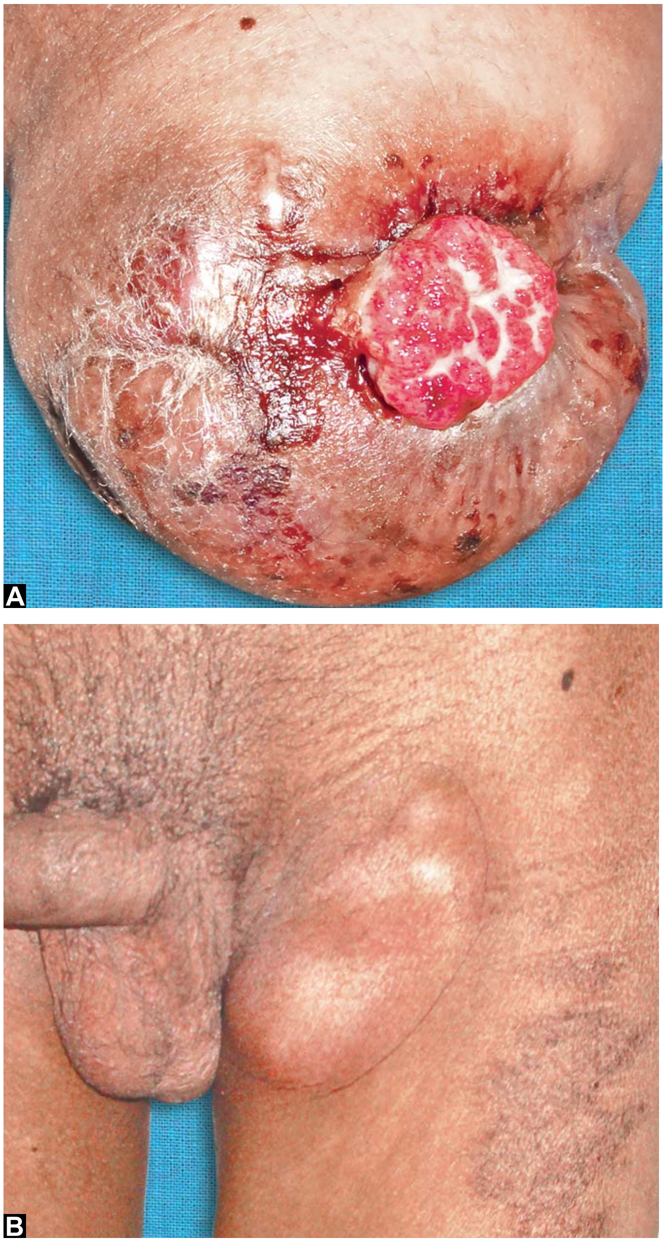

Figs 1.561A and B: Recurrent SCC in amputated stump. Note the enlarged significant lymph nodes in the groin.

Treatment

- **Radiotherapy** using radiation needles, moulds, etc. is given.
- **Wide excision**, *2 cm* clearance followed by skin grafting or flaps**.**
  (Presently for tumour less than 2 cm, 4 mm clearance and for tumour more than 2 cm, 1 cm clearance is sufficient). Wide excision should show *clearance* both at margin as well as in the depth. If muscle, fascia, cartilage are involved, it should be cleared. Reconstruction is usually done by primary split skin grafting (*SSG/Thiersch*). Delayed skin grafting also can be done once wound granulates well. Often flaps of different types may be needed depending on the site of lesion.
- **Amputation** with one joint above.
- For lymph nodes, block dissection of the regional lymph nodes is done.
- Curative radiotherapy (RT) is also useful in tumours which are not adherent to deeper planes or cartilage as SCC is radiosensitive. It is also useful in recurrent SCC and in patients who are not fit for surgery. A dose of 6000 cGy units over 6 weeks; 200 units/day is used. Recurrence after RT is treated by surgical wide excision.
- In advanced cases with fixed lymph nodes, **palliative external radiotherapy** is given to palliate pain, fungation and bleeding.
- **Chemotherapy** is given using methotrexate, vincristine, bleomycin.
- Field therapy using cryo probe or topical fluorouracil or electrodessication.

Prognostic factors in SCC

- Tumour size > 2 cm is worse
- Tumour border—ill-defined border is worse
- Associated immunosuppression is worse
- Differentiation—poorly differentiated is worse
- Perineural involvement has worse prognosis
- Invasion; depth < 2 mm has got better prognosis; depth more than 16 mm has got worst prognosis
- Local recurrence rate is 20%. Recurrence period is 5 years, not beyond

Verrucous Carcinoma (Fig. 1.563)

- Dry, exophytic, warty growth.
- No lymph node spread.
- No blood spread.
- Surgery is the treatment—wide excision.
- No radiotherapy.
- Examples:
  - Giant condyloma acuminatum (Buschke-Lowenstein tumour, Verrucous carcinoma of genitalia).

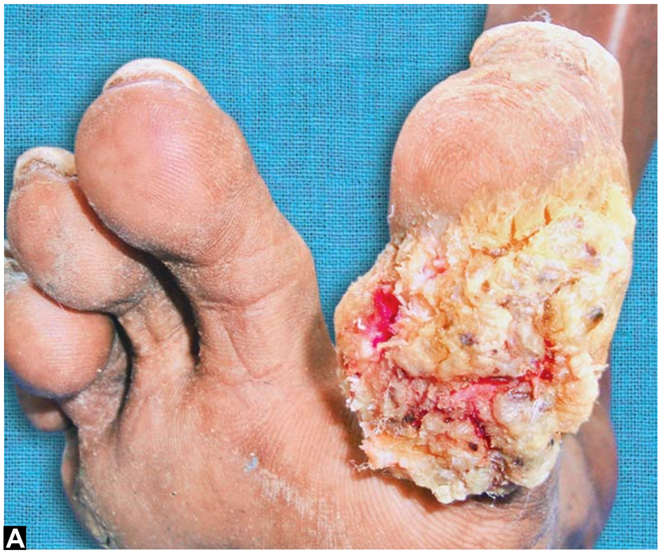

Fig. 1.562A

Moles must never be cauterised or curetted.

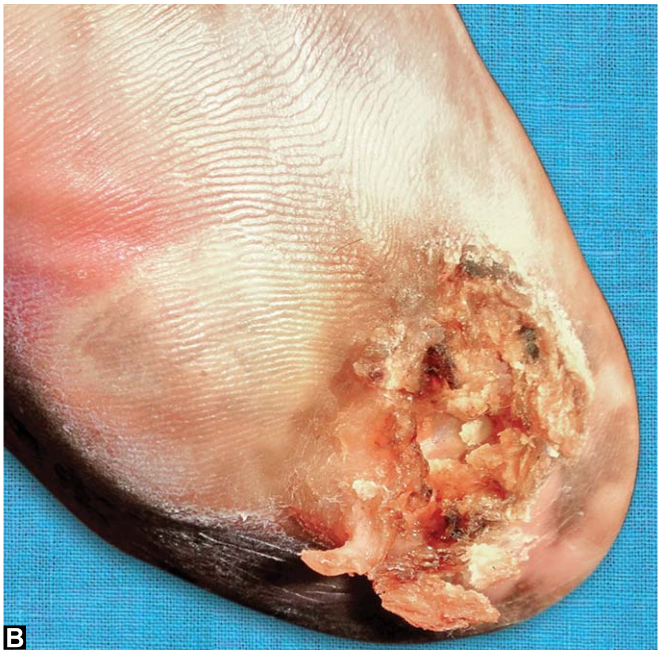

Fig. 1.562B

Figs 1.562A and B: *Carcinoma cunniculatum*—in heel and under great toe which is a low grade verrucous type of SCC.

- Oral florid verrucous carcinoma.
- Verrucous carcinoma of foot (plantar aspect)—***carcinoma cuniculatum.***

MARJOLIN'S ULCER (1828)

- It is a well-differentiated squamous cell carcinoma which occurs in chronic scars like burn scar, scar of venous ulcer.
- As it develops in a scar due to chronic irritation and there are **no lymphatics** in scar tissue, **it will not spread to lymph nodes**.
- As scar is relatively avascular it grows slowly. As scar does not contain nerves, it is painless.
- Once it reaches the normal skin it may behave like any other squamous cell carcinoma, i.e. it will spread to lymph nodes. It occurs in *unstable scar of long duration*.

Clinical Features

- History of pre-existing venous ulcer or burn scar.
- Indurated, painless, nontender, ulcer with raised and everted edge.
- Biopsy from the edge **confirms the diagnosis.**

Treatment

- **Wide excision**.
- In case of large ulcer, **amputation** is required.
- ***Radiotherapy should not be given*** as it may turn into poorly differentiated squamous cell carcinoma.
  It is a curable malignancy.

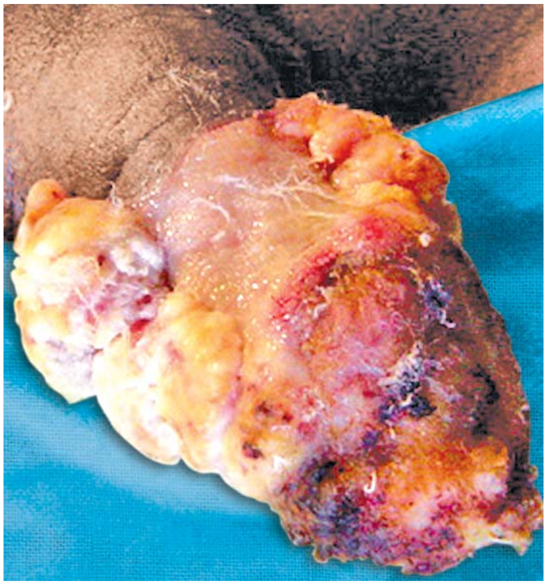

Fig. 1.563: Verrucous carcinoma of penis. It is an exophytic, dry, warty, locally malignant lesion (Squamous cell carcinoma).

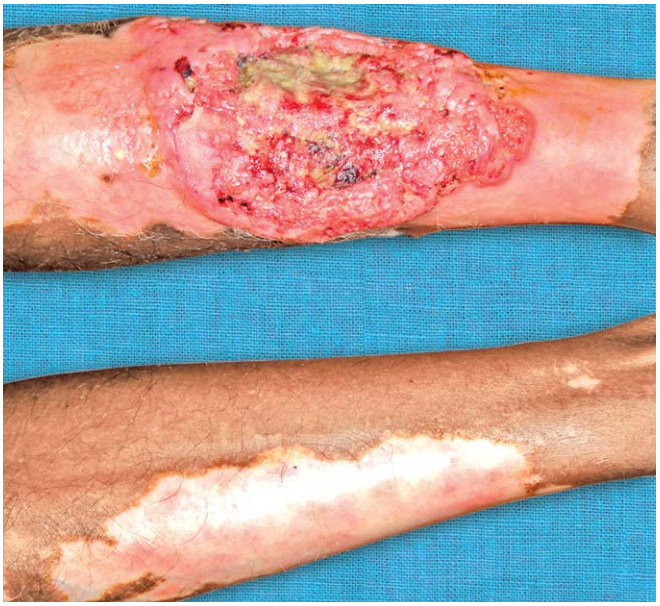

Fig. 1.564: Marjolin's ulcer in a burn scar.

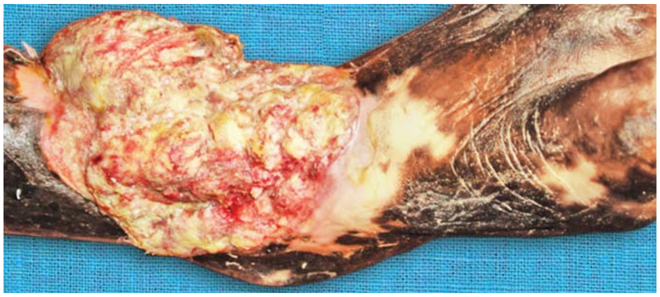

Fig. 1.565: Marjolin's ulcer foot in a pre-existing burns scar. It is a well-differentiated squamous cell carcinoma occurring in an unstable scar of long duration.

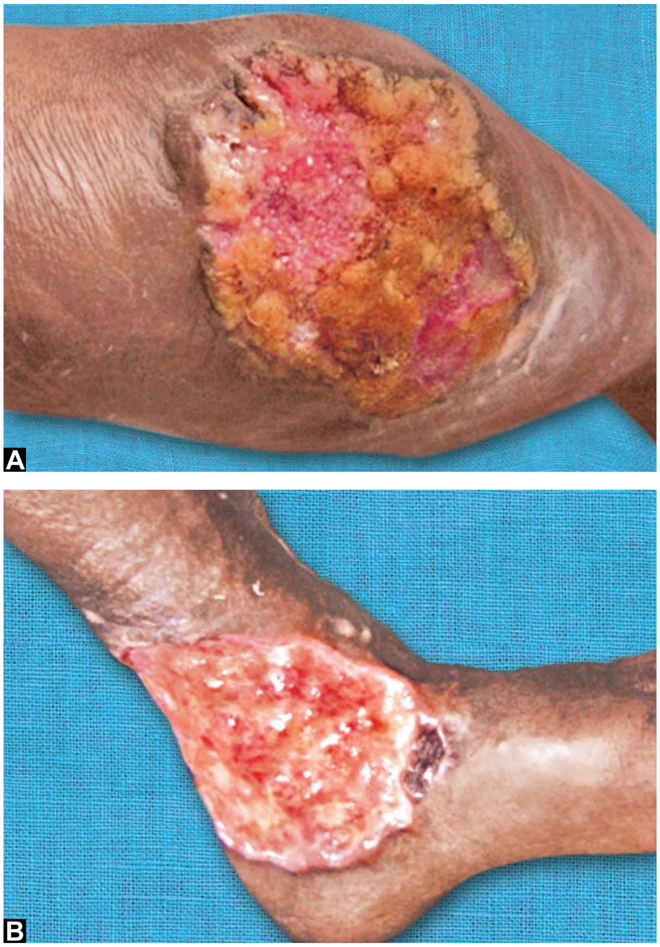

Figs 1.566A and B: Marjolin ulcers over elbow and over foot. First one occurred over elbow in a pre-existing snakebite scar. Second one occurred in a long-standing venous ulcer.

BASAL CELL CARCINOMA (Rodent Ulcer)

It is a low grade, locally invasive, carcinoma arising from basal layer of skin (or adnexal basal layer of hair follicle) or mucocutaneous junction. ***It does not arise from mucosa.***

- It is the commonest (70%) malignant skin tumour.
- It is more common in white-skinned people than blacks.
- Common in places where exposure to *UV light* is more (***Australia).***
- Other causes are—arsenics, coal tar, aromatic hydrocarbons, skin tumour syndromes.
- It is common in males, common in middle-aged and *elderly*.
- Common site is ***face***—above the line drawn between angle of mouth and ear lobule (90%)—*Onghren's line*.
- It is called as ***tear cancer*** because it is commonly seen in area where tears roll down.
- Often it can occur in muco-cutaneous junctions.
- *Basal cell naevus syndrome* (Gorlin syndrome) with BCC; medulloblastoma; bifid ribs.
- ***It is only locally malignant***. It does not spread through lymphatics nor through the blood. But it erodes deeply into local tissues including cartilages, bones causing extensive local destruction. Hence the name **"*rodent ulcer*".**

Types

1. Nodular.
2. Cystic/nodulocystic.
3. Ulcerative.
4. Multiple, often associated with syndromes and other malignancies.
5. Pigmented BCC—mimics melanoma.
6. Geographical or field fire or forest fire BCC is wide area involvement with central scabbing and peripheral active proliferating edge.
7. **Basisquamous**—behaves like squamous cell carcinoma which spread into lymph nodes. BCC which has not been treated for long time can turn into Basisquamous carcinoma.

Note:
Nodulocystic and noduloulcerative is the commonest form (70-90%).

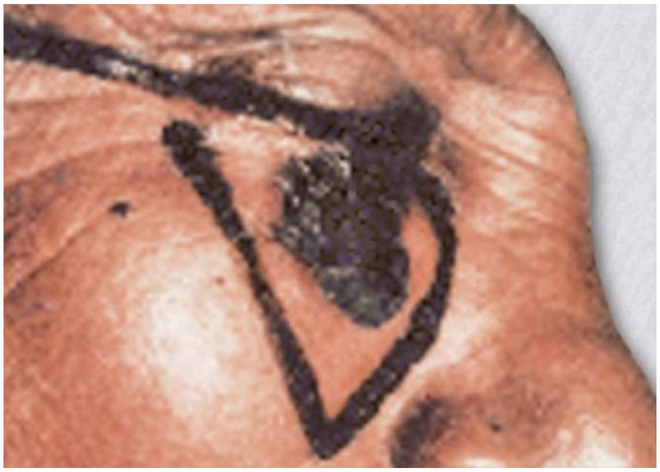

Fig. 1.567: Basal cell carcinoma.

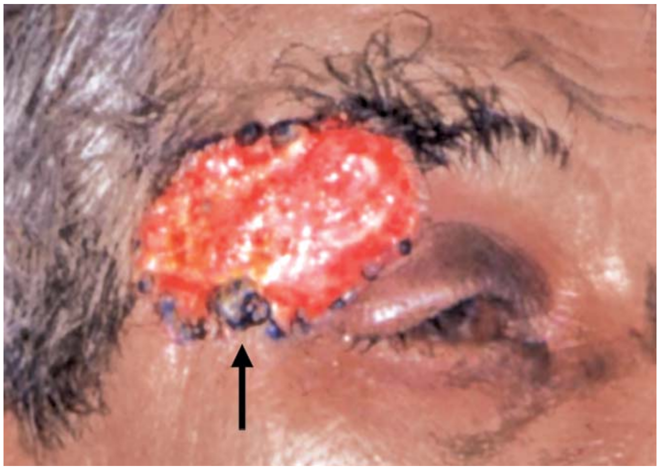

Fig. 1.568: Rodent ulcer lateral aspect of the eyebrow. Note the typical beaded edge.

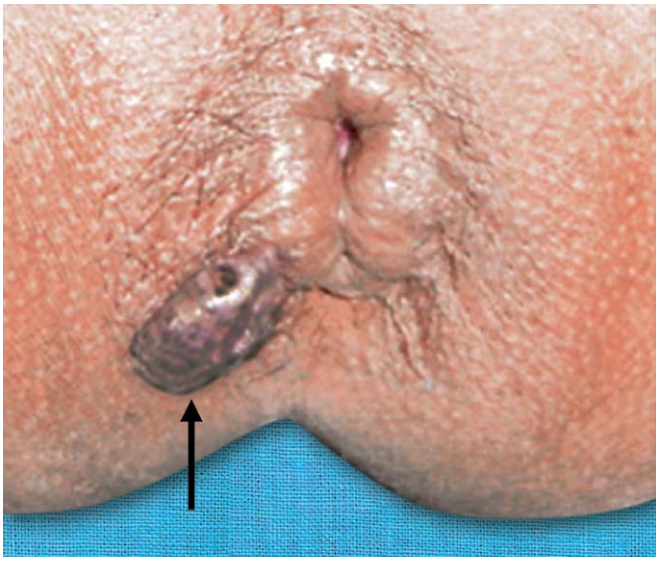

Fig. 1.569: BCC in perianal region. It is nodular type. Wide excision is required.

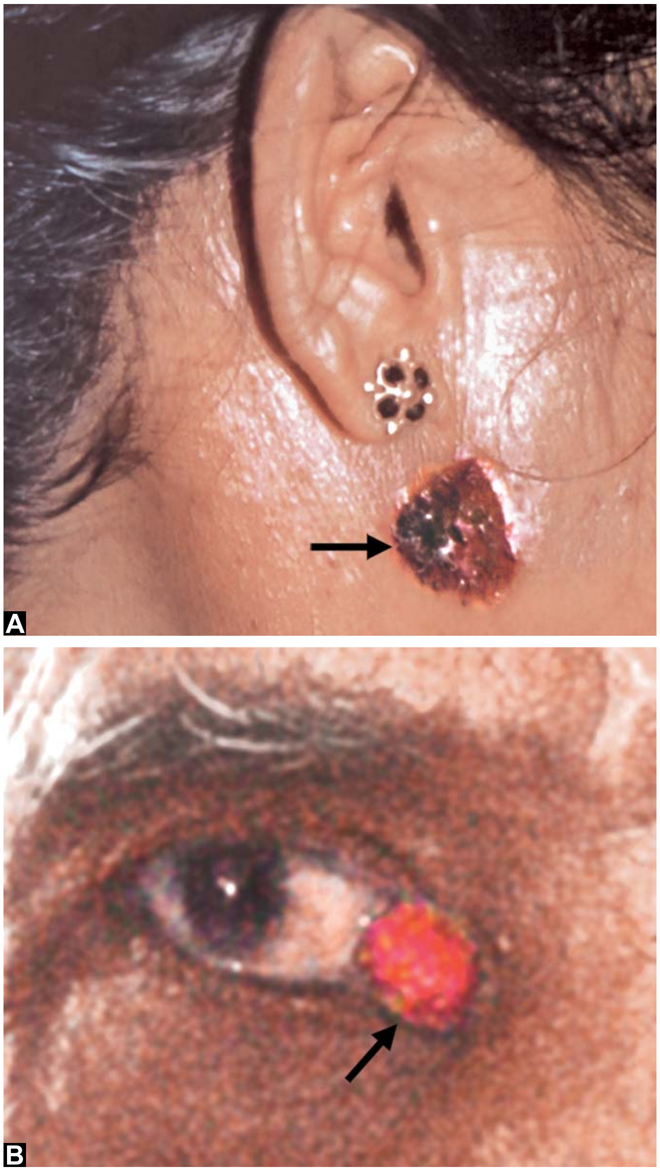

Figs 1.570A and B: BCC in different places—below the ear and in the inner angle of the eye.

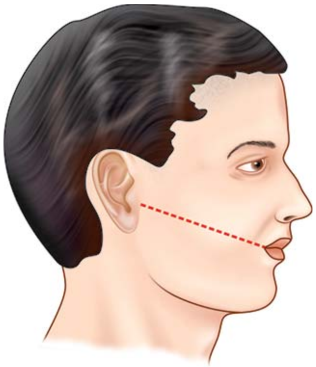

Fig. 1.571: Note the common site of BCC—in the face above the line drawn between angle of mouth and ear lobule—*Onghren's* line.

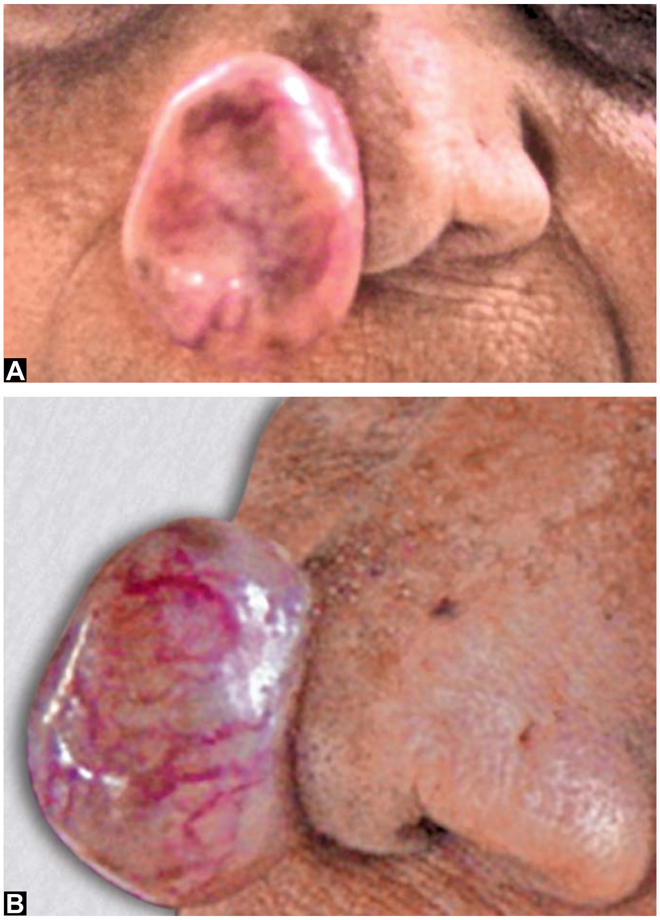

Figs 1.572A and B: Nodular BCC in the nose.

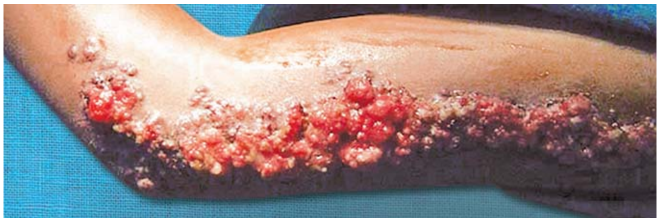

Fig. 1.573: *Sebaceous epidermal naevus*. It is common in females, often extensive, begins in childhood. It needs surgical excision and skin grafting. It has got 10% chances of turning into BCC.

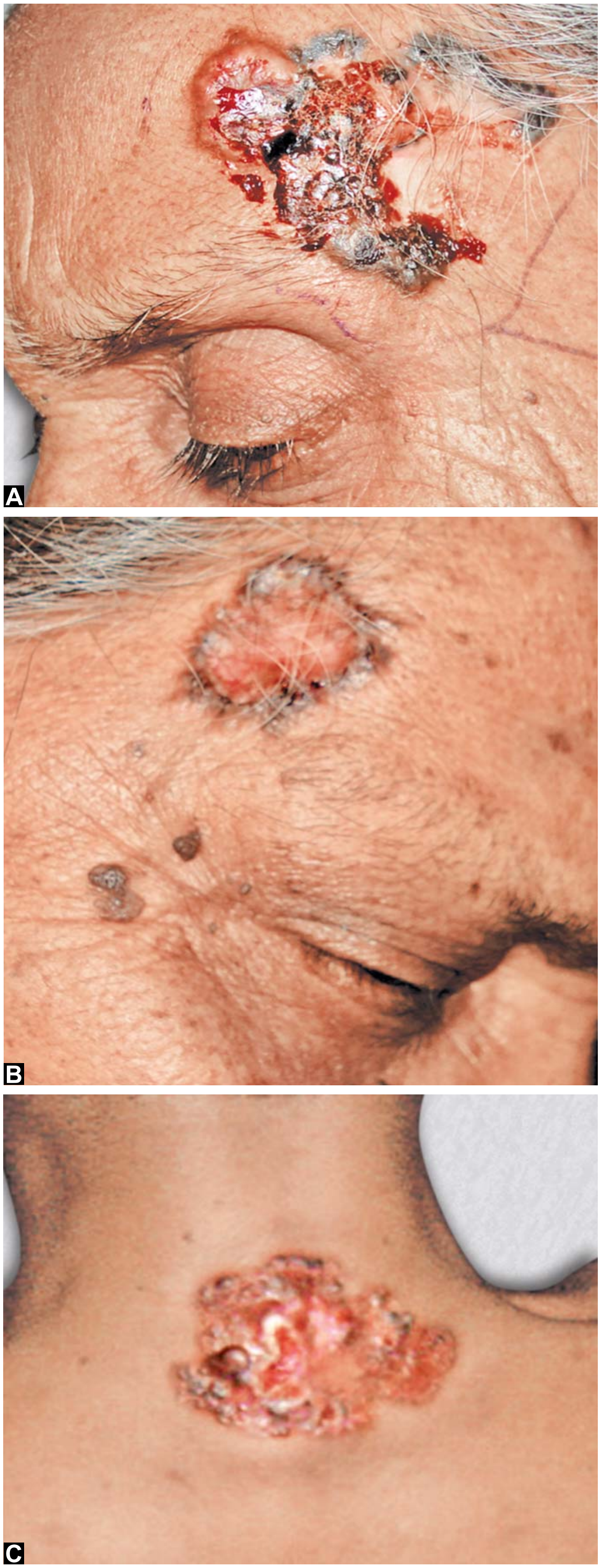

Figs 1.574A to C: Basal cell carcinoma in different locations. Note the nodularity and ulceration. Note the beaded edge.

Clinicopathological Types

a. *Superficial type*—small buds of tumour masses.
b. *Morpheic type*—dense stroma with basal cells and type IV collagen; spreads rapidly.
c. *Fibroepithelioma type* of Pinkus shows elongated cords of basaloid cells with mesh work.

It contains outer palisading columnar cells with central polyhedral cells but no *prickle cells* or *keratinisation*.

Clinical Features

- Ulcer on the face in a middle-aged man which is nontender, dry, slowly growing, nonmobile, with raised and beaded edge with central scab, often with central depression or umbilication.
- Site of beading signifies the area of active proliferating cells. In between beaded areas dormant nonactive cells are present.
- No lymph node or blood spread occurs. Due to large sized tumour cells/tumour cluster, it does not spread through lymphatics.

High-risk BCC

- BCC can be low risk or high-risk
- Size > 2 cm
- Near the eye/nose/ear
- Ill-defined margins
- Recurrent tumours
- Immunosuppressed individuals

Differential Diagnosis

1. Squamous cell carcinoma.
2. Melanoma.
3. Keratoacanthoma.
4. Seborrhoeic keratosis.

Investigations

Edge biopsy, X-ray of the part, CT scan.

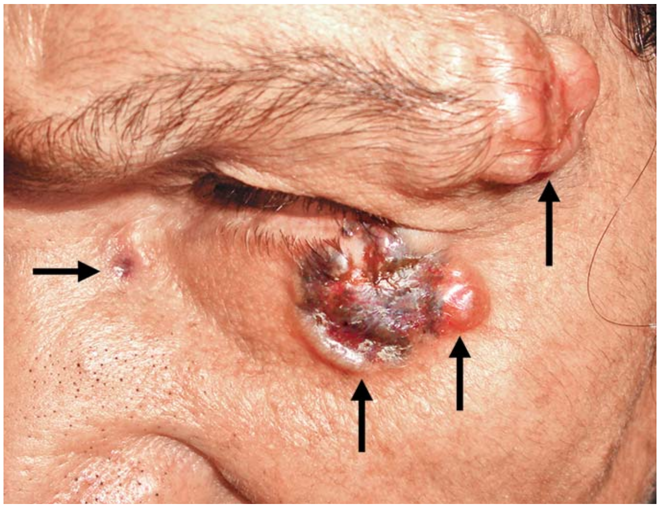

Fig. 1.575: BCC lesions in different places (multiple).

Treatment

- It is ***radiosensitive***. If lesion is away from vital structure (like away from eyes), then curative radiotherapy can be given. **Radiotherapy is not given, once it erodes cartilage or bone.** RT is *not given* to BCC of ear and close to lacrimal canaliculi.

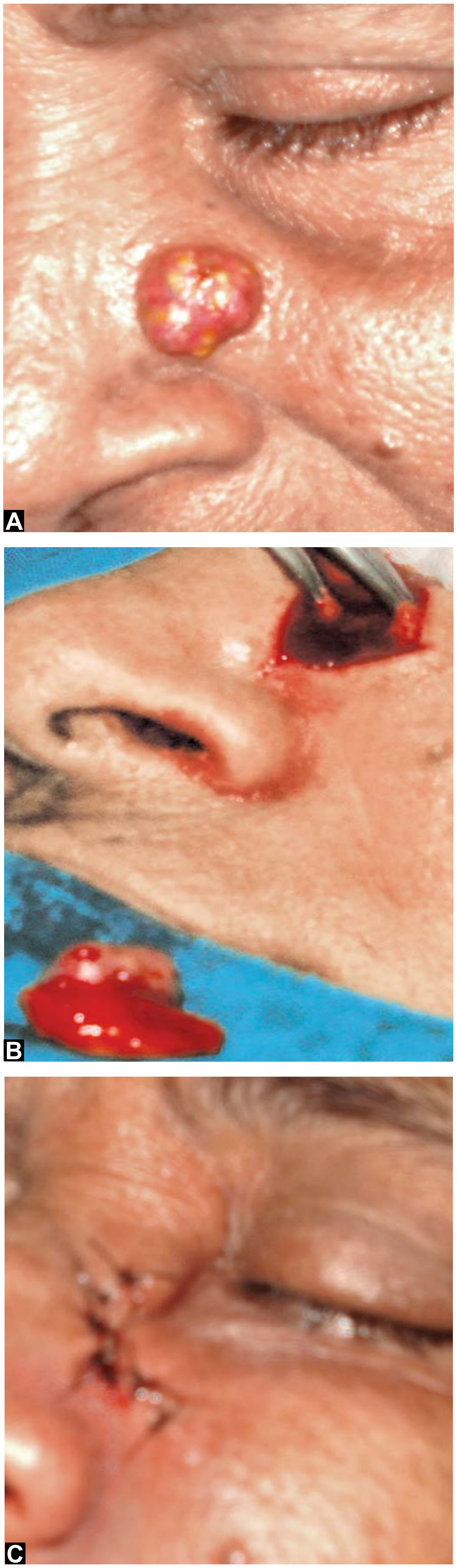

Figs 1.576A to C: Surgical wide excision in BCC.

- *Surgery:*
  - Wide excision (1 cm clearance) with skin grafting, primary suturing or flap (Z plasty, rhomboid flap, rotation flap) is the procedure of choice.
  - Laser surgery.
  - Cryosurgery.
  - **MOHS** (*Microscopically Oriented Histographic Surgery*) (Federic E Mohs, American Surgeon) is useful to get a clearance margin and in conditions like BCC close to eyes, nose or ear, to preserve more tissues. MOHS is becoming popular in BCC/dermatofibrosarcoma protuberans/melanoma. Procedure is done by dermatological surgeon along with a histotechnician/histologist. Under local anaesthesia, a saucerised excision of the primary tumour is done and quadrants of the specimen are mapped with different colours. Specimen is sectioned by histotechnician from margin and depth, and it is stained using eosin and haematoxylin. It is studied by MOHS surgeon or histologist. Residual tumour from relevant mapped area is excised and procedure is repeated until clear margin and clear depth are achieved. Clearance must be complete and proper in BCC otherwise there will be very high chance of recurrence (70%).

Indications for surgery

- Rodent ulcer eroding into cartilage or bone
- BCC close to the eye
- Recurrent BCC

TURBAN TUMOUR

It is a descriptive term wherein entire scalp looks like a turban because of multiple scalp swellings. It can be due to *multiple cylindroma; multiple hidradenomas; subcutaneous neurofibromas; nodular multiple basal cell carcinoma.*

- *Multiple cylindroma* is usually considered disease under this term. Cylindroma is a variant of eccrine spiradenoma (skin adnexal tumor). Multiple firm pinkish nodules in the scalp are the presentation in multiple cylindroma. They are rare, often locally malignant, grows slowly over the span of many years to cover entire scalp with reddish lobulated lesion.
- Hidradenoma is a rare benign sweat gland tumour. Multiple tumours commonly look like a turban in the scalp. They are painless, disfiguring, cosmetically problematic, soft, boggy, non-fluctuant, non-compressible cutaneous swellings; commonly observed in middle age group.
- Multiple sebaceous cysts over the scalp mimic the same.
- *Management* is initial biopsy to find out the cause; then wide excision with skin grafting.

NAEVI

- It is a *hamartomata* of melanocytes due to excessive stimulation.
- It may present during birth or appear later in life.

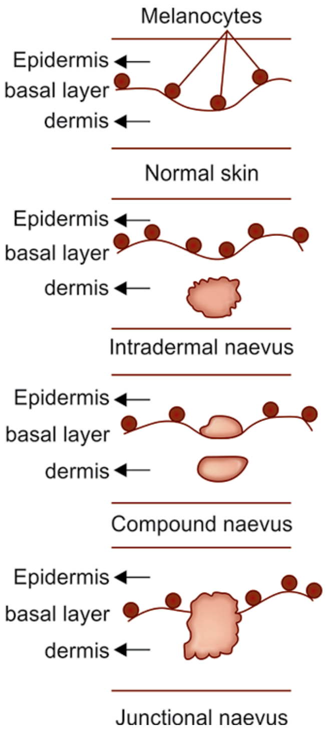

Fig. 1.577: Different types of naevi.

Types

1. ***Hairy mole*** is a mole with a hair growing on its surface.
2. ***Nonhairy mole.***
3. ***Blue naevus.*** It is seen in children. It is located deep in the dermis, hence appears blue. It is common in buttock (Mongolian spot), hand, feet.
4. ***Junctional naevus***. It lies centred in the junctional layer (basal layer) of the epidermis as clusters. It is immature, unstable and premalignant. Microscopically there is proliferation of melanocytes at the epidermal junction. Features of malignant transformation are—change in the size, colour, bleeding, ulceration, crusting, satellite spots.
5. ***Compound naevus***. It is combination of intradermal and junctional naevus. Intradermal part is inactive but junctional part is potentially malignant.
6. ***Juvenile melanoma (Spitz naevus)*** (It is a misnomer). It appears as junctional like mole before puberty. It is seen in children on face. They present as brownish red nodular lesion which needs always excision.
7. ***Hutchinson's freckle***. It is seen in elderly with large area of dark pigmentation. In the macular stage it is smooth and brown. In the tumour stage it is dark and irregular. It can turn into melanoma commonly.
8. ***Halo naevus*****:** Halo of depigmentation around the pigmented naevus. This halo is due to antibody response to melanocytes. Halo naevus is often seen along with vitiligo. Similar halo may develop around a malignant melanoma lesion.
9. ***Intradermal naevus:*** Cluster of dermal melanocytes is seen without junctional component. Common in face.
10. ***Spindle cell naevus*****:** It is dense, black pigmented lesion containing spindle cells and atypical melanocytes at dermo-epidermal junction; seen in females on high with malignant potential.
11. ***Naevus spilus***: It is hyperpigmented speckles throughout, also called as *speckled lentiginous naevus*. Malignant potential is rare.
12. ***Naevus of Ota*** is dermal melanocytic hamartoma seen in distribution area of trigeminal nerve (ophthalmic /maxillary). It is seen in oriental and African race adolescent females (thigh) with a hormonal influence.
13. ***Naevus of Ito*** is similar lesion occurring in shoulder region.
14. ***Dysplastic naevus*** is proliferation of atypical melanocytes from epidermal basal layer having variegated irregular look; it is usually > 5 mm in size; can be familial; 10% cases may turn into superficial spreading melanoma.

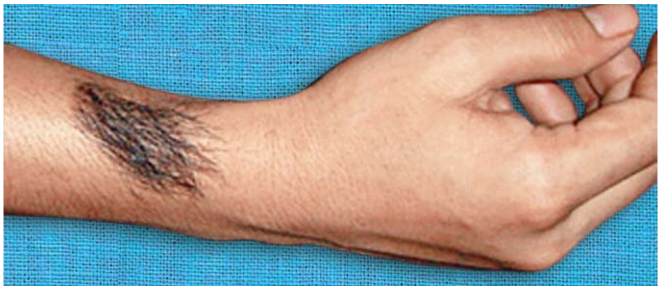

Fig. 1.578: Hairy naevus.

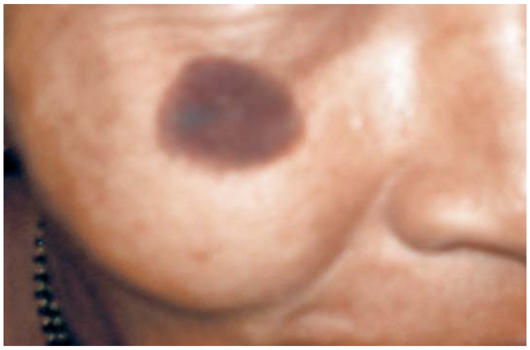

Fig. 1.579: Congenital naevus.

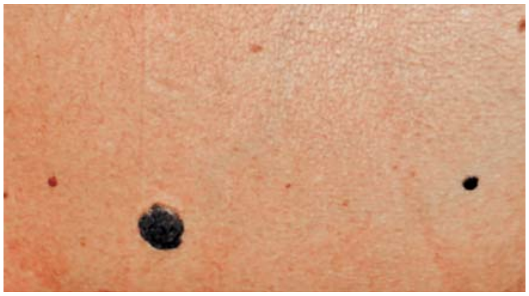

Fig. 1.580: Compound naevus.

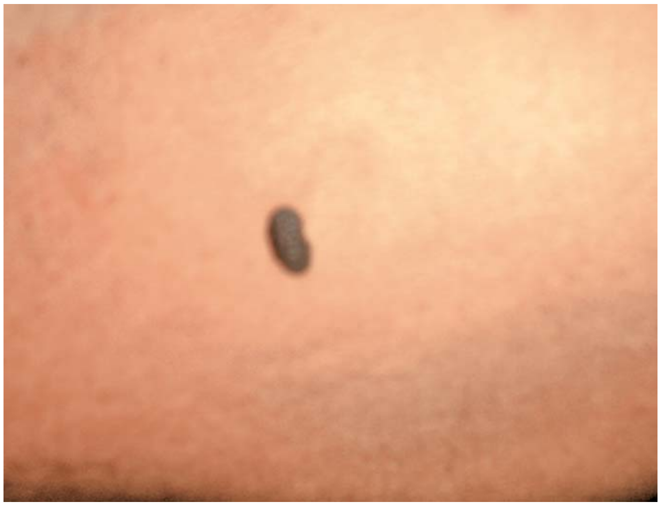

Fig. 1.581: Junctional naevus.

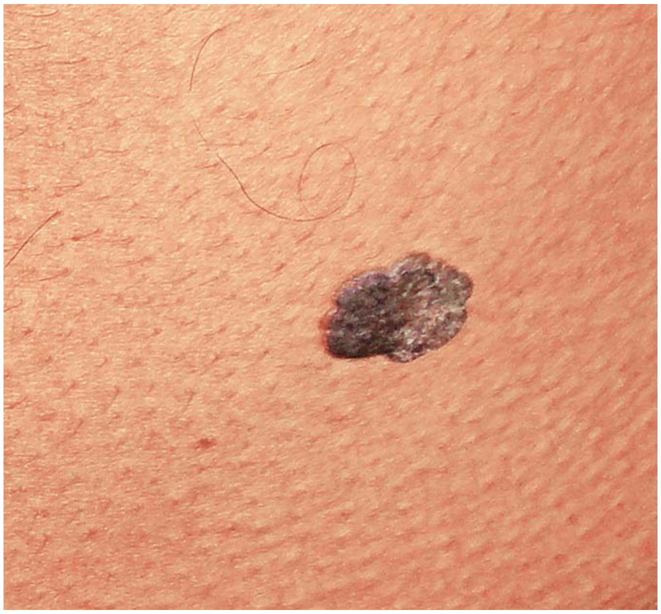

Fig. 1.582: Intradermal naevus.

Treatment

Excision. Always should be sent for histopathology.

Note:

- *Giant naevus* is naevus more than 1% of body surface area or more than 20 cm in size. Giant congenital pigmented naevus (GCPN) often shows dermatomal distribution. Pigment cells spread from epidermis to fat and muscle often. It may turn into melanoma (5% risk). Such melanomas are usually axial; is usually associated with retroperitoneal / intracranial melanosis. Curettage, dermabrasion, laser, excision and SSG are the treatment options.
- *Mongolian spot* is a blue brown/grey pigmented macular area which is seen in sacral region during birth and after initial intense pigmentation regresses fully in 7 years.
- Normal number of melanocytes releasing abnormally higher number of melanin granules is called as *freckle*/ephilis.

MELANOMA

- It is a malignant tumour arising from epidermal melanocyte which is *most aggressive* cutaneous malignant tumour.

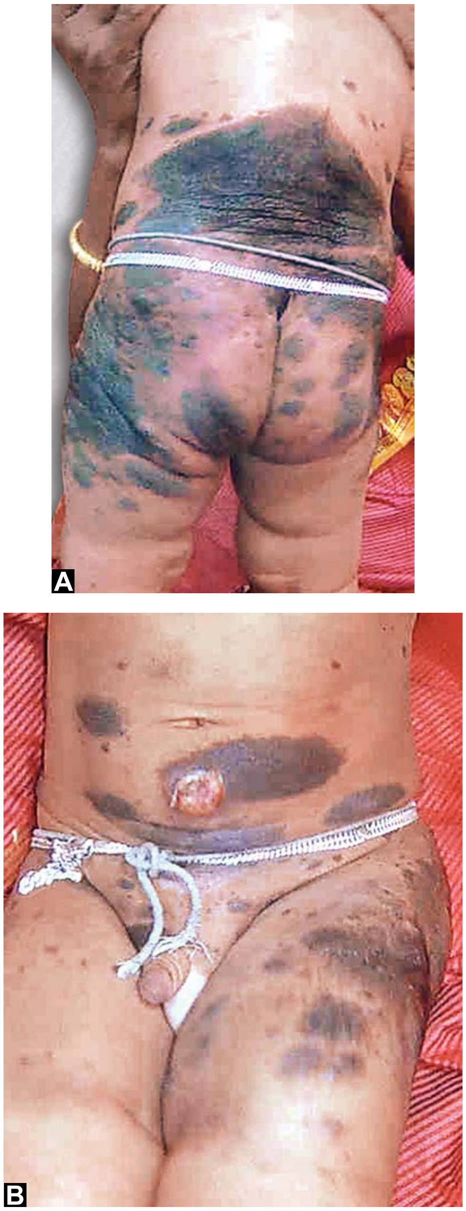

Figs 1.583A and B: Giant congenital pigmented naevus (*Courtesy:* Dr Murali Keshav, MD, Paediatrician, KMC, Mangalore).

- It is of *neural crest* (ectodermal) origin.
- It is 20 times more commonly seen in *whites* than blacks.

- *Melanoblast*: Primitive cell derived from the neural crest.
- *Melanocyte*: The cell which synthesises melanin is located in the basal layer (melanocyte : basal cell : : 1 : 10).
- *Melanophores* are pigment melanin carriers through dendrites into the epidermis.
- *Melanophages* are macrophages having melanin pigment.
- Melanoblasts and melanocytes contain DOPA oxidase enzyme and synthesise melanin. They show +ve DOPA reaction.
- Melanophores and melanophages show –ve DOPA reaction.
- Melanin synthesis is controlled by melanocyte stimulating hormone, ACTH and sex hormones.

DOPA reaction:

$$\text{Tyrosine} \xrightarrow[\text{Tyrosinase}]{} \text{DOPA} \xrightarrow[\text{Oxidase}]{} \text{Melanin}$$

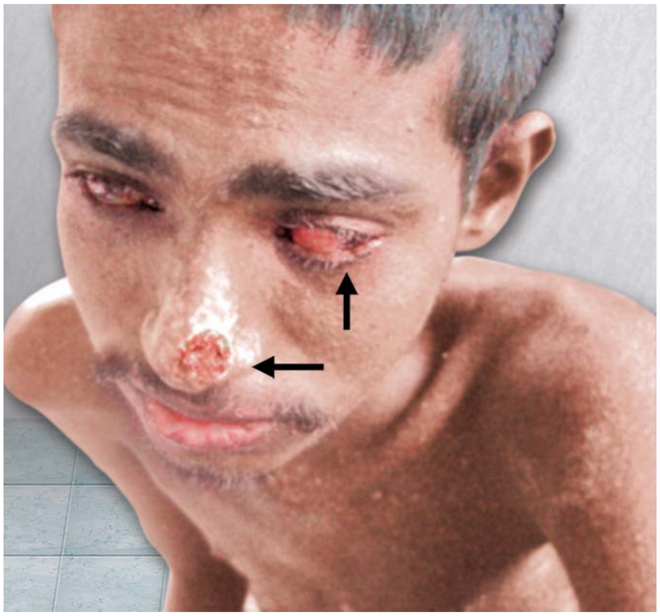

Fig. 1.584: Xeroderma pigmentosa with BCC in the nose. Such patients are prone for other skin malignancy like melanoma also. There is defective DNA excision repair.

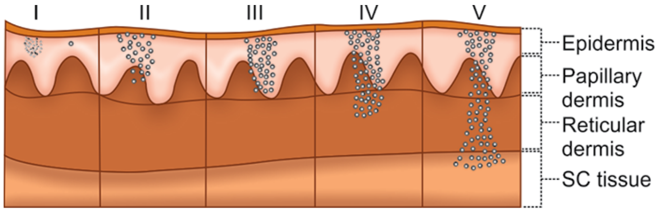

Fig. 1.585: Clark's level.

- Its incidence is equal in both sexes.
- Its incidence is increasing over the years.
- 5 % of skin cancers are melanomas.
- It is most common in Queensland, Australia. Auckland, New Zealand.

Sites

Sites of melanoma

- Head and neck—25%
- Trunk—25%
- Lower limb—25%
- Upper limb—11%
- Other sites—14%

- In females, *leg* is the commonest site.
- In males, the *front* or *back* of the trunk.
- In the Bantu tribe, *sole* is the commonest site.

Other sites:

Eyes (iris, ciliary body, choroids), mucocutaneous junction (anorectal region, genitalia), head and neck (meninges, oropharynx, nasopharynx, paranasal sinuses).

Risk Factors

- Exposure to sunlight (exposure to UV light; more common in white-skinned—20 times).
- Ethnic factors, socioeconomic status (high society), lifestyle, climate.
- Albinism.
- *Xeroderma pigmentosa* – RR is 1000 (By Kaposi in 1874): It is an autosomal recessive (Ch 9q) disease with defect in DNA excision repair mechanism causing formation of aberrant nucleotide causing ultraviolet rays' intolerance, erythema, pigmentation, photophobia, premature skin ageing, multiple malignancies with 60% mortality at the age of 20.
- Junctional naevus.
- Familial dysplastic naevus syndrome.
- Sporadic dysplastic naevi. 10% risk.
- Large congenital naevi (larger than 20 cm).
- Family history of melanoma (10% through chromosomes 1p, 6q, 7 and 9). They often present with multiple primary melanomas; associated with dysplastic naevus.
- History of earlier skin cancers other than melanoma.
- Patients who are on immunosuppressive drugs or after renal transplantation or NHL (RR - 30).

Malignancies which spread from mother to foetus

- Melanoma
- Lymphosarcoma

Classifications

Breslow's classification (1970): Based on thickness of invasion measured by optical micrometer—most important prognostic indicator until nodal spread

I: Less than 0.75 mm
II: Between 0.76 to 1.5 mm
III: 1.51 mm to 4 mm
IV: more than 4 mm

Relation of Tumour Thickness to Nodal Spread—Based on AJCC Classification

Lesion	*Tumour thickness*	*Nodal spread*
Thin	< 1 mm	< 10%
Intermediate	1-4 mm	20-25%
Thick	> 4 mm	60%

You will soon break the bow if you keep it always stretched.

Clark's levels (Fig. 1.585)

Level 1: Only in epidermis
Level 2: Extension into papillary dermis
Level 3: Filling of papillary dermis completely
Level 4: Extension into reticular dermis
Level 5: Extension into subcutaneous tissue

Differential diagnosis for melanoma (other pigmented lesions of the skin)

- Seborrhoeic keratosis, dermatofibroma
- Pigmented BCC, pigmented SCC
- Naevus, sebaceous epidermal naevus
- Kaposi's sarcoma, mycosis fungoides
- Cutaneous haemangioma
- Certain skin adnexal tumours
- Solar keratosis
- Pyogenic granuloma
- Cutaneous angiosarcoma

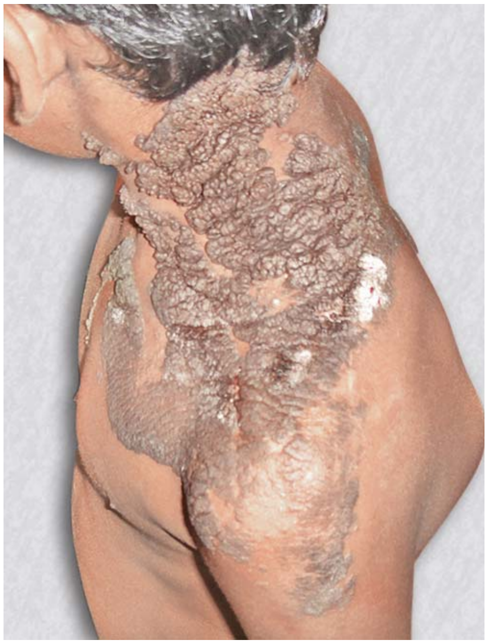

Fig. 1.586: Extensive pigmented lesion of the skin.

Types

- Cutaneous melanoma
- Extra cutaneous—10% (ocular is common site)
- Occult (Unknown primary)—2-7%

Clinical Types

1. ***Superficial spreading:***
    - **Most common**. 64-70%. Occurs in any part of the body with variegated irregular look. It has more radial growth and better prognosis. It commonly arises from a pre-existing naevus.
2. ***Nodular melanoma:*** 12-25%.
    - **More aggressive**. It is common in younger age group, occurring in any part of the body. It has more vertical growth. Common in mucosa and mucocutaneous junc-

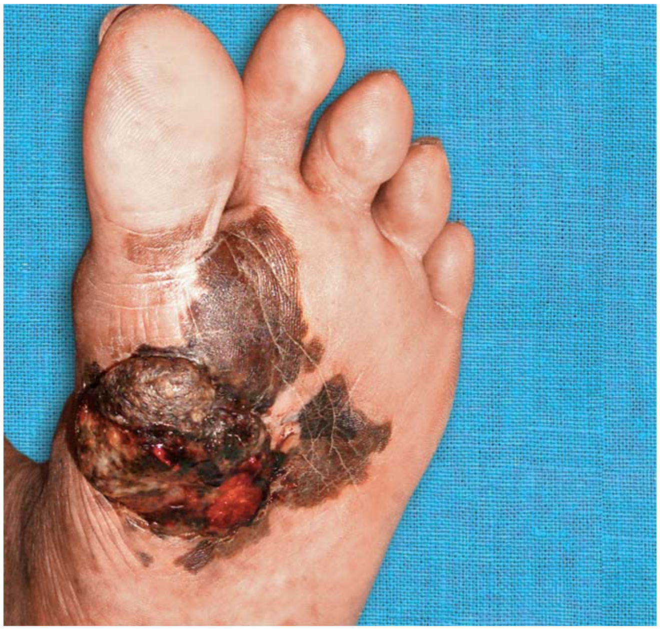

Fig. 1.587: Superficial spreading melanoma (70%).

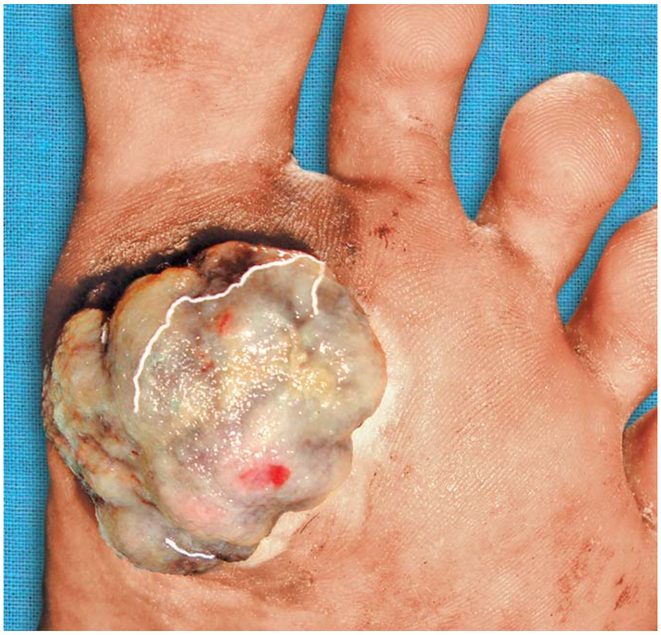

Fig. 1.588: Nodular malignant melanoma.

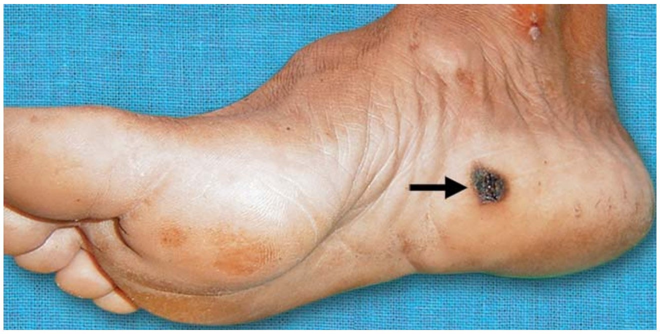

Fig. 1.589: Melanoma in the sole. Often such pigmentation may be unnoticed.

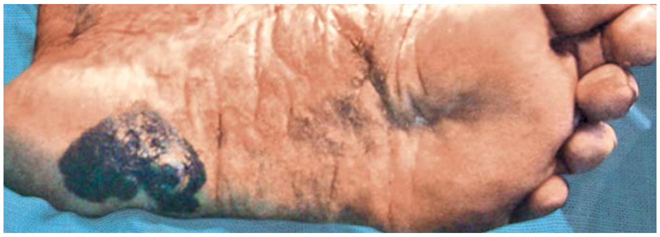

Fig. 1.590: Large melanotic lesion in the heel.

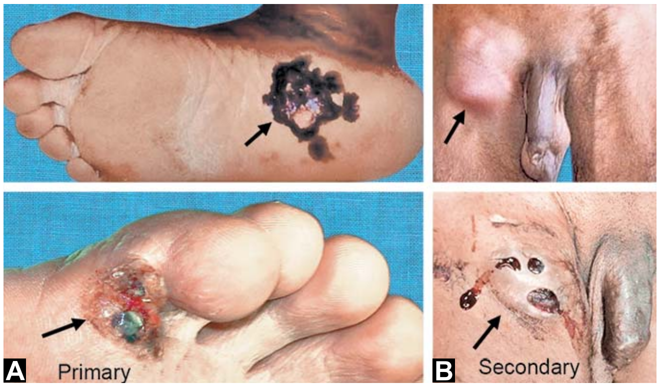

Figs 1.591A and B: Primary melanoma with lymph node secondaries in two different individuals. Note the pigmented ulcerated secondaries in one.

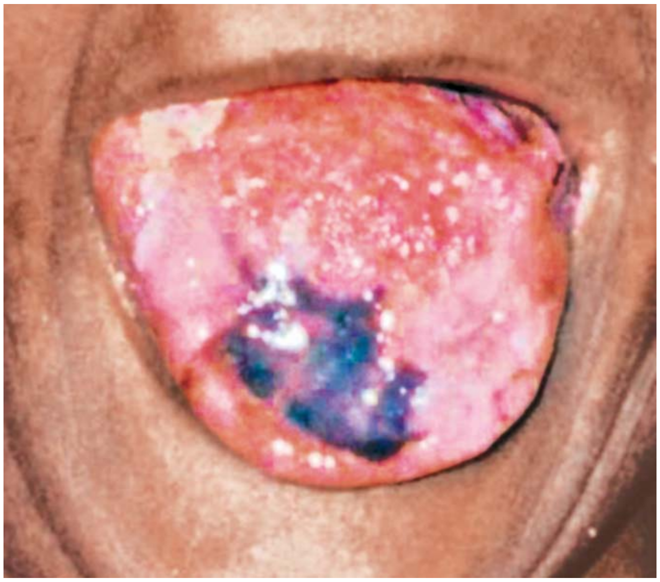

Fig. 1.592: Melanoma in tongue.

tion; uniform; nodular; more vertical growth; nodal spread is common; has got poor prognosis. It usually appears as *de novo*. Common in men; common in trunk, head and neck.

3. ***Lentigo maligna melanoma:***
   - 7-15%. **Less common, least malignant**. Occurs in old age and common in face (**Hutchinson's melanotic freckle**). It is slow growing, variegated, brown macule/lentigo; common in face/neck/hands; common in elderly women. Lentigo maligna is *in situ* type.

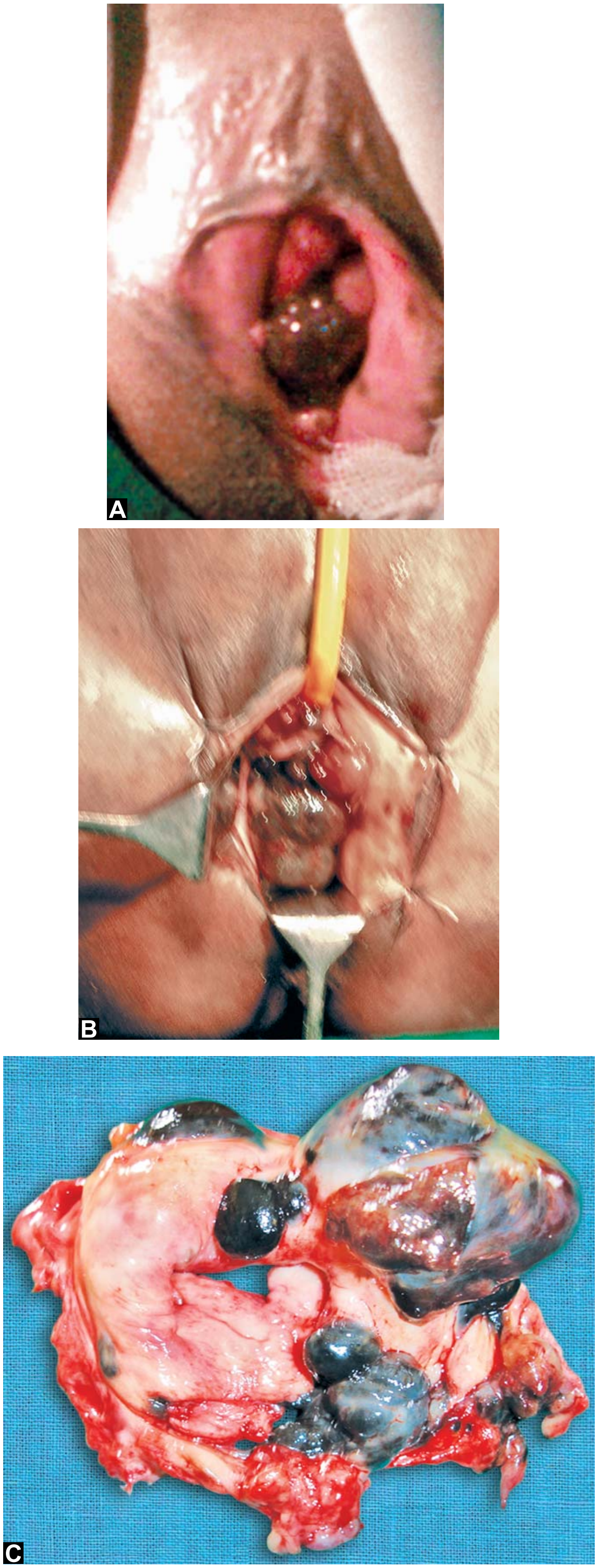

Figs 1.593A to C: Melanoma in vagina—in mucocutaneous junction. On table and excised specimen of melanoma.

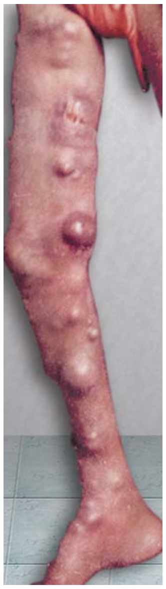

Fig. 1.594: 'In-transit' nodules in melanoma. They are secondary depositions in dermal lymphatics.

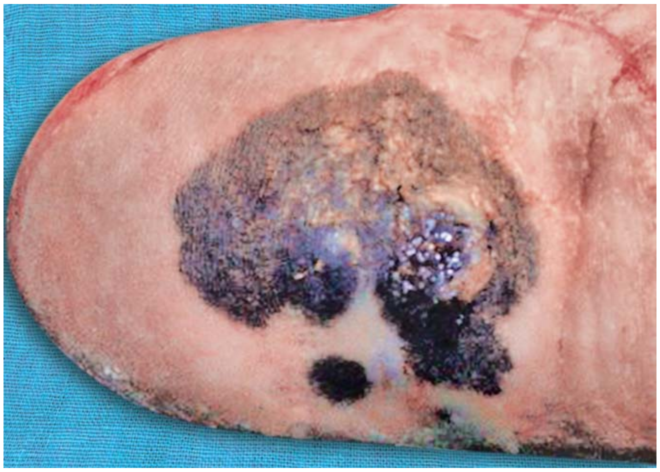

Fig. 1.595: Satellite nodule in melanoma occurs within 2 cm of primary. Note the primary in the heel and satellite nodule adjacent to it.

4. ***Acral lentiginous melanoma:***
    - 5%. Occurs in palms, soles and subungual region. Common in **Japan**. It has got a **poor prognosis**. It is **least common**. Usually attains large size; nodular type with more vertical growth phase. It is common in Africa and Asia. It is less common in whites. It is often flat, irregular macule. It mimics fungal infection/pyogenic granuloma.
5. ***Amelanotic melanoma:***
    - This is the **worst type.** Because of the undifferentiation, tumour cells loose their capacity to synthesise melanin. It presents as rapidly progressive pinkish fleshy tumour. It may mimic soft tissue sarcoma. It needs markers like S100, HMB45 for diagnosis.
6. ***Desmoplastic melanoma*** has high affinity for perineural invasion; is common in head and neck with higher recurrence rate. It is *amelanotic melanoma with thicker lesion* carrying poor prognosis due to neural invasion.

***Note*:**

Subungual melanoma which was earlier thought of acral lentiginous type is now considered as superficial spreading type. It is involvement of *nail fold matrix* (not nail plate). Triangular, macular, progressively widening pigmentation of nail fold with nail dystrophy is typical—*Hutchinson's sign*. It should be differentiated from benign racial melanonychia which are dark streaks under the nail. Biopsy of *nail matrix* should be done here.

Ocular melanoma: It is the commonest malignancy arising in eye. It may arise from retina, iris, ciliary body, choroid. It rarely metastasize or only at late stage as it is devoid of lymphatics. Ocular melanoma commonly shows its distant spread to liver. Massive hepatomegaly is typical and is often seen many years after the treatment of primary ocular lesion. Condition is treated with enucleation, radiation, photocoagulation.

Clark's concept—Two phases of growth: Initial radial growth phase occurs horizontally, later vertical growth phase occurs with invasion.

Melanoma

- 5% of all skin cancers—incidence
- 20 times more common in whites than blacks
- Mucosal melanoma has got poor prognosis
- Can spread from mother to foetus
- Multiple melanomas are 1% common
- Melanoma in choroid will not cause lymph node involvement, as it has no lymphatic drainage. But late massive liver secondaries even after 10-20 years is known to occur
- 10% of melanomas are familial—in whites
- Satellite nodules are secondary skin nodules within 2 cm of primary
- In transit' nodules are secondary skin nodules beyond 2 cm of primary any where up to lymph node region
- Melanoma may present as secondary (in liver, lungs, bone, brain) with occult primary when primary is situated in anus, genitalia, eye, external auditory canal, adrenal gland, nail bed and scalp—7%
- Pigmentation is not *mandatory* to diagnose melanoma even though it is commonly present.

Clinical Features

- It can stat in a pre-existing **naevus** (commonly junctional naevus)—50-60% or as **de novo** in a normal skin—40-50%.
- Melanoma is **unknown before puberty.**
- ***No induration is seen in melanoma***.
- Pigmentation with irregular surface and margin with rapid growth.
- **Ulceration, bleeding, itching, change in the colour.**

Note:

When a mole turns malignant, following changes should be observed (*Glasgow* criteria:

- ***Major signs*:** Change in *size* (diameter more than 6 mm), *shape* and *colour*
- ***Other changes***:
    - Inflammation, crusting, bleeding, itching
    - Nodularity, ulceration, halo around a mole
    - Satellite lesions
    - Doppler positive pigmented lesions using hand held Doppler (> 1 mm thick lesion)

Five most important features of melanoma

- **A**symmetry
- **B**order irregularity
- **C**olour variation
- **D**iameter > 6 mm
- **E**levation

Spread

- **Through lymphatics** it spreads to regional lymph nodes either by permeation or by embolisation.
- **In-transit nodules or satellite nodules** are seen in the skin between the primary lesion and regional lymph node area, and is due to retrograde spread to dermal lymphatics.
- **Through blood**: To lungs, liver (huge liver), brain, skin, bones. Secondaries are typically black.

Blood spread in melanoma

- *Brain*—convulsions, localising features and raised intracranial pressure
- Lung—cannon ball secondaries, pleural effusion-haemoptysis, chest pain and cough
- Liver (massive liver), ascites
- Skin—cutaneous nodules often pigmented
- Bones—bone pain, pathological fracture. Paraplegia/neurological deficits in spine metastasis

Extensive visceral involvement causes ***melanuria***

- Melanoma **in choroid** has got better prognosis, because as there are no lymphatics, spread is delayed.
- Sometimes primary is very small and unnoticed (in anus, subungual region). They present with features of secondaries only.

***Note*:**

Infiltration into deep fascia by melanoma is rare in initial stages as deep fascia acts as a strong barrier.

Staging of Malignant Melanoma (It is older staging system)

IA: Thickness less than 0.75 mm.
IB: Thickness between 0.76 to 1.5 mm.
IIA: Thickness between 1.51 to 4.0 mm.
IIB: Thickness more than 4.0 mm.
IIIA: Any of the above + nodes less than 3 cm.
IIIB: Any of the above + nodes more than 3 cm.
IV: Any of the above + any node + M1 (distant spread).

Occult melanomas (primary unknown) are common in

— Anus
— Scalp
— External auditory canal
— Nail bed
— Genitalia
— Eye
— Adrenal medulla

Tumour markers for melanoma

- MELAN-A
- S 100
- HMB 45 (Hydroxy Methyl Bromide)
- LDH

Investigations

- **No incision biopsy.** It can cause early blood spread.
- Excision biopsy of primary. It is done with 2 mm margin with deeper fatty tissue. One should avoid using cautery

Newer TNM staging (2002) for melanoma

Tumour	***Node***
T 0 No tumour	N 0 No nodes
T is *In situ* tumour	N 1 a—one node micrometastasis
T 1a < 1 mm level II, level III no ulceration	N 1 b—one node macrometastasis
T 1 b < 1 mm, level IV with ulceration	N 2 a—2 or 3 nodes micrometastasis
T 2 a 1-2 mm no ulceration	N 2 b—2 or 3 nodes macrometastasis
T 2 b 1-2 mm with ulceration	N 2 c—no nodes but satellite or in transit lesions
T 3 a 2-4 mm no ulceration	N 3-4 or more nodes; nodes with satellite or in transits
T 3 b 2-4 mm with ulceration	Stage 0 – TisN0M0
T 4 a > 4 mm no ulceration	Stage IA – T1aN0M0; IB – T1b/T2aN0M0
T 4 b > 4 mm with ulceration	Stage IIA – T2b/T3aN0M0; IIB – T3b/T4aN0M0; IIC – T4bN0M0
M—Metastasis	Stage IIIA – T1-4aN1a/N2aM0; IIIB – T1-4bN1a/N2aM0, T1-4aN1bN2bM0, T1-4a/bN2cM0; IIIC – T1-4bN1b/N2bM0, any TN3M0
M 0—no blood spread	Stage IV – Any T, Any N, M0
M 1 a—Skin, subcutaneous tissue, distant node	
M 1 b—Lung spread	
M 1 c—Other viscera or distant spread and increase in LDH	

and avoid crushing the tissues as much as possible. Instead of excision biopsy, punch biopsy is done in case of large primary tumour very close to pinna, eye, nose. Punch biopsy assesses the depth/thickness of the lesion. Punch should be done at the most elevated part of the lesion to get proper depth.

- FNAC of lymph node.
- U/S abdomen to look for liver secondaries (usually huge hepatomegaly occurs).
- Chest X-ray to look for secondaries in lung ("cannon ball" appearance). HRCT of chest is ideal.
- Relevant other methods depending on site and spread, e.g. CT scan of head, chest, abdomen, pelvis.
- Urine for melanuria signifies advanced disease.
- Sentinel lymph node biopsy (**SLNB).**

Treatment

Surgery is the main treatment.

For Primary:

a. Handley's ***wide local excision (WLE)*** is wide excision with clearance of margin as well as depth. Clearance margin used in olden days is 3-5 cm.

Present recommendation is— *in situ* melanoma needs 0.5 cm clearance; melanoma < 1 mm thickness needs 1.0 cm clearance; 1-2 mm thickness needs 1-2 (1.5) cm clearance; 2 cm/3 cm clearance is sufficient for > 2.0 mm thickness. Procedure can be done under regional or local anaesthesia. Evidence says that more than 2 cm clearance will not show any additional advantage in treating primary tumour.

Primary closure or SSG or local flaps are used to cover the defect.

b. If primary area is wide, then amputation with one joint above is done.

c. In fingers and toes, disarticulation is required.

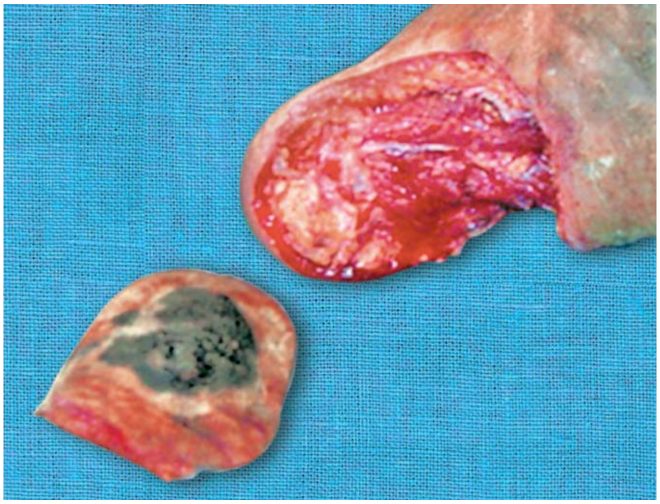

Fig. 1.596: Wide excision of melanoma from the heel. Note the clearance margin and depth of dissection.

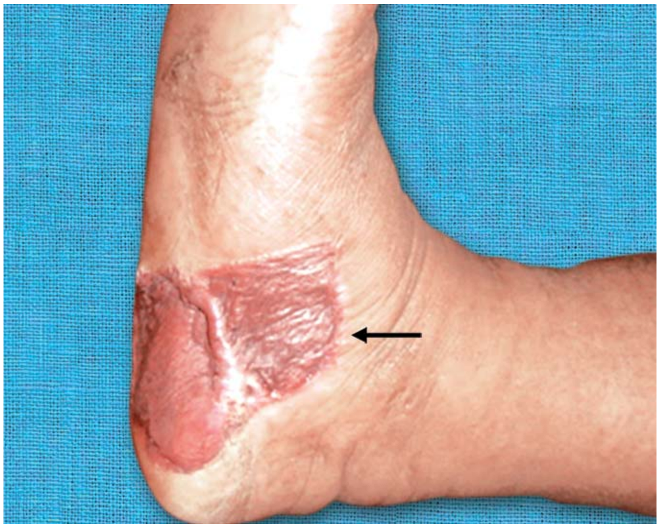

Fig. 1.597: Wide excision and skin grafting done for melanoma.

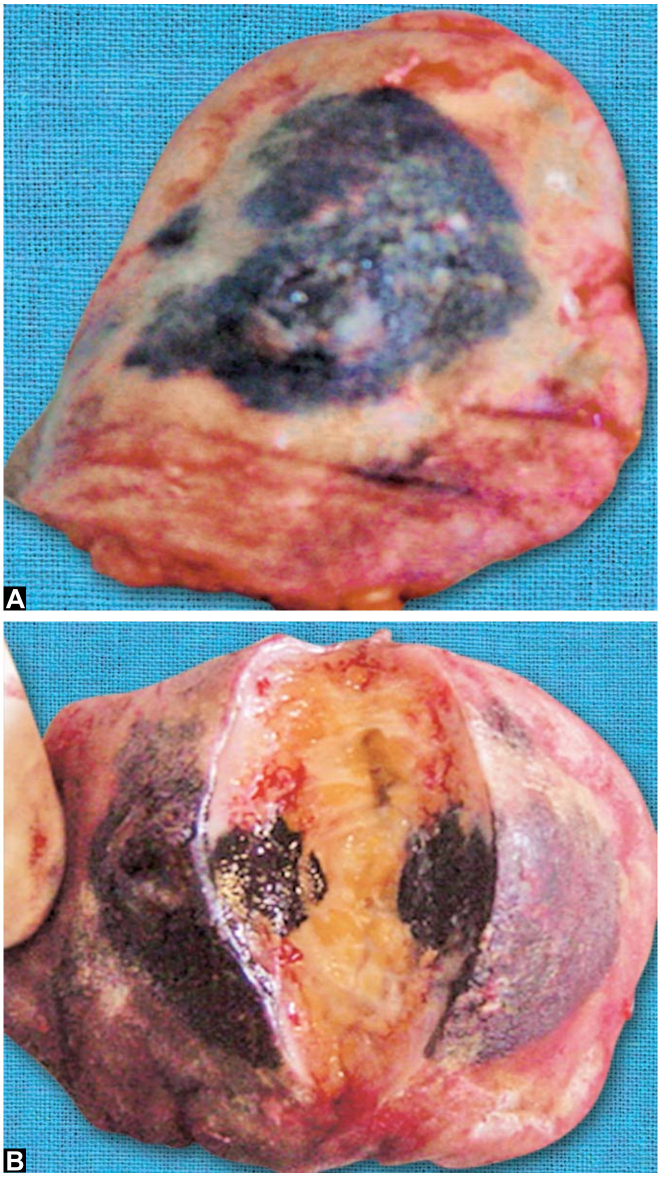

Figs 1.598A and B: Widely excised melanoma specimen from heel. Cut section of widely excised specimen of melanoma from heel shows the depth of the tumour.

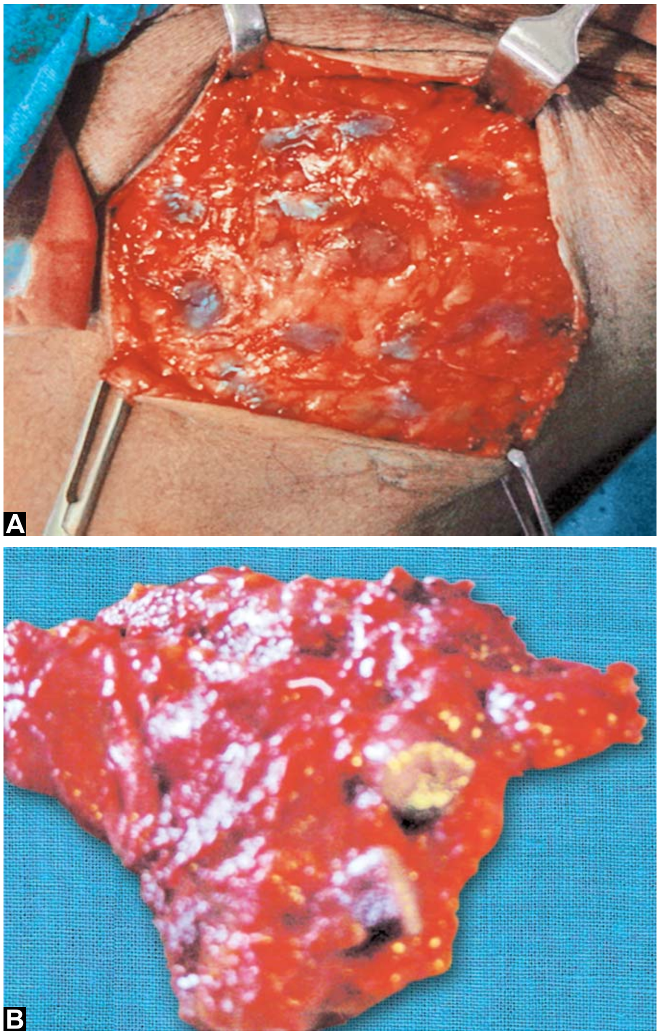

Figs 1.599A and B: Secondaries in inguinal lymph nodes from primary melanoma sole-on table finding. Specimen of ilioinguinal block is also shown. Pigmented nodes are observed.

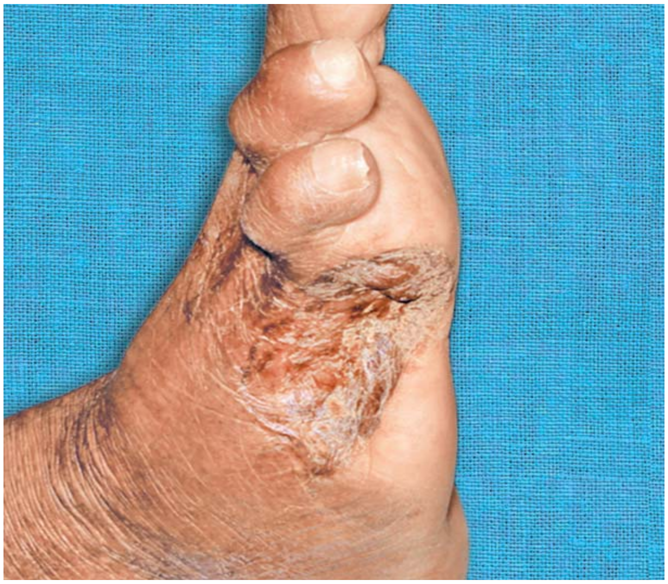

Fig. 1.600: Disarticulation of lateral two toes for interdigital melanoma.

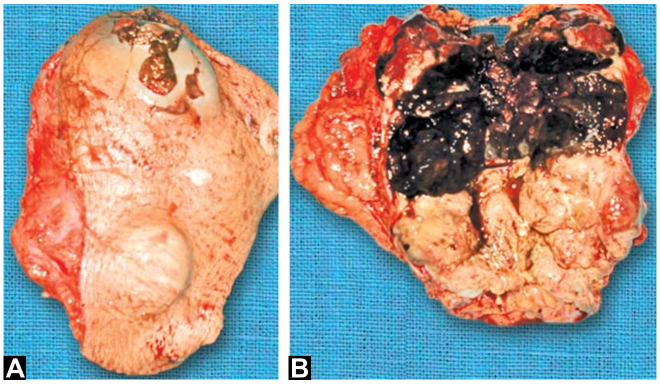

Figs 1.601A and B: Melanoma secondaries in groin nodes after block dissection. Black lesions are observed in cut section.

d. Melanoma in anal canal may require abdominoperineal resection.
e. Enucleation in case of melanoma in eye.
f. *Melanoma in pregnancy* is treated with termination of pregnancy and specific therapy for melanoma. Pregnancy should be postponed for 2 years.

For lymph node secondaries:

1. In a clinically palpable lymph node, FNAC of lymph node is done. In case of spread, then regional block dissection, i.e. ilioinguinal or axillary or neck is done. Once FNAC shows positive lymph node 5-year survival rate reduces to 50%.
2. In a fixed lymph node, only chemotherapy is the treatment because it is inoperable.
3. *Lymphatic mapping and sentinel node biopsy* (Dr Donald Morton, 1970)*:* Radioactive colloid is injected around primary site and lymphoscintigraphy is done using hand held gamma camera to visualise the micrometastasis in the nodal field. If there are micrometastasis, then regional block dissection (therapeutic LND) is done.

   SLN (Sentinel lymph node) can be identified in 95% or more of groin and axillary nodes and in 85% cases of head and neck melanomas. Often both blue dye and technetium sulfur colloid is used together to identify the SLN. SLNB is useful for melanoma with thickness more than 1 mm depth. Less than 1 mm thickness is considered as *low-risk* for metastases; between 1-4 mm thickness is considered as *intermediate-risk* for metastases mainly of nodal spread; more than 4 mm thickness will be considered as *high-risk* for both nodal as well as blood spread. SLNB is the investigation of choice for staging in intermediate thickness melanoma.
4. Prophylactic regional block dissection which was previously advocated is now controversial. But still used in many centres. Elective lymph node dissection (ELND) is done when tumour thickness is 1-4 mm.
5. *Management in unknown primary* (2%) presenting as nodal secondaries is by nodal radical dissection at the region with adjuvant chemotherapy. They have better prognosis

than with known primary. Patient should be monitored and evaluated to identify the primary site during every follow up. Once primary is identified it is treated accordingly depending on its location.

For Loco Regional Recurrent Melanoma:

- *Local recurrence* is one which recurs within 5 cm radius of the primary tumour in skin or subcutaneous tissues. Risk of local recurrence is 0.2% if primary tumour is less than 0.75 mm; 2% if it is between 0.75-1.5 mm; 6% if it is 1.5-4 mm; 12% if it is more than 4 mm.
    - *Isolated limb perfusion* (Creech et al, New Orleans, 1958) using cytotoxic agents like Melphalan (M for M), interleukin 2, tumour necrosis factor (TNF). Concentration used here is 15-25 times more than that is used for systemic therapy. Melphalan dose is 10 mg / l perfusion solution in leg (13 mg/L in arm).

      Melphalan is injected at high temperature of 41°C with a pump and oxygenator through cannulas in femoral artery and vein with a proximal tourniquet *in situ.* Hyperthermia and oxygenation increase the metabolic activity of tumour cells to make it more vulnerable to melphalan. Procedure controls the local disease well with preserving the functioning limb. *Complications* like DVT, bleeding, sepsis can occur. It is used in local recurrence or 'In-transit' deposits.

      It shows 80% response rate with 15% complete response; but only of short period.
    - Laser ablation of multiple small cutaneous lesions.
    - *Isolated limb infusion* (Thompson, 1993): Vascular catheters are passed and placed across femoral artery and vein through opposite femoral vessels or through arm vessels. The limb is warmed; patient is anaesthetized 2 hours later and also heparinized; papaverine is injected into arterial catheter and tourniquet is applied in the thigh/arm. Melphalan 7.5 mg/L, actinomycin D 75 µg/L in 400 ml saline (10 mg and 100 µg/L in 300 ml NS in upper limb) is infused into the isolated limb for 6 minutes which is pumped around the limb repeatedly for 30 minutes with a hypoxia in limb; drugs are washed out using a Hartmann's solution; tourniquet is removed and normal circulation is regained with removal of vascular catheters. Protamine is given to reverse heparin action.

For Distant Spread:

- Brain, *lung and liver* are the *most common* sites; skin, bone, GI are less common sites. But melanoma is one of the commonest tumours to spread to *intramural GIT* to present as intussusception.
- Distant spread when found or suspected, CT scan of head, chest, abdomen and pelvis are needed. PET scan and tumour markers are very useful.
    - Chemotherapy and immunotherapy is the main treatment.
    - Isolated lung or liver metastasis can be resected.
    - Radiotherapy is useful for bone and brain secondaries. Stereotactic program using gamma knife is better for brain secondaries.

Chemotherapy for Melanoma

Indications:

a. Secondaries in lungs, liver, bones.
b. After surgery for melanoma. Usually it is given intravenously.

Drugs are:

a. DTIC: Diethyl Triamine Imino Carboxamide.
b. *Melphalan* (Phenyl alanine mustard) (Melphalan for melanoma).
c. Carboplatin, vindesine.
d. *CVD regime*—is cisplatin, vinblastine and dacarbazine.

Immunotherapy/Biological Therapy

- It is done using specific tumour antibodies, BCG, levamisole, *Corynebacterium parvum*, alpha interferon, interleukins and tumour vaccines is tried with some success rate up to 40% in advanced melanomas.
- *Biochemotherapy* is combination of CVD with interferon α and interleukin 2.
- Interferon α is a cytokine which is antiangiogenic and stimulator of natural killer NK cells. Dose is 20 mU/m^2 IV 3 times a week for 4 weeks then maintenance dose of 10 mU/m^2 subcutaneously 3 times a week for one year. Severe myelodepression and fulminant liver necrosis are the toxicity and so often dose is reduced to 3 mU/m^2 three times weekly for 2 years.
- GM2 ganglioside based vaccine (stimulates production of IgM antibodies), Melacine (contains melanoma cell lysates) and cancerVax are three vaccines under trial at present.

Endolymphatic Therapy

- It can be done to control disease in the nodes using radioactive I_{131} or P_{32} with ultrafluid lipiodol along with lymphangiography.

***Note*:**

- There is not much role for radiotherapy.
- Radiotherapy is beneficial only in secondaries in brain and bones.

Prognosis for Melanoma

- It is not good since it is a very aggressive tumour.
- Old age has worse prognosis.
- Females show better prognosis.
- Extremity melanoma has better prognosis than head and neck.

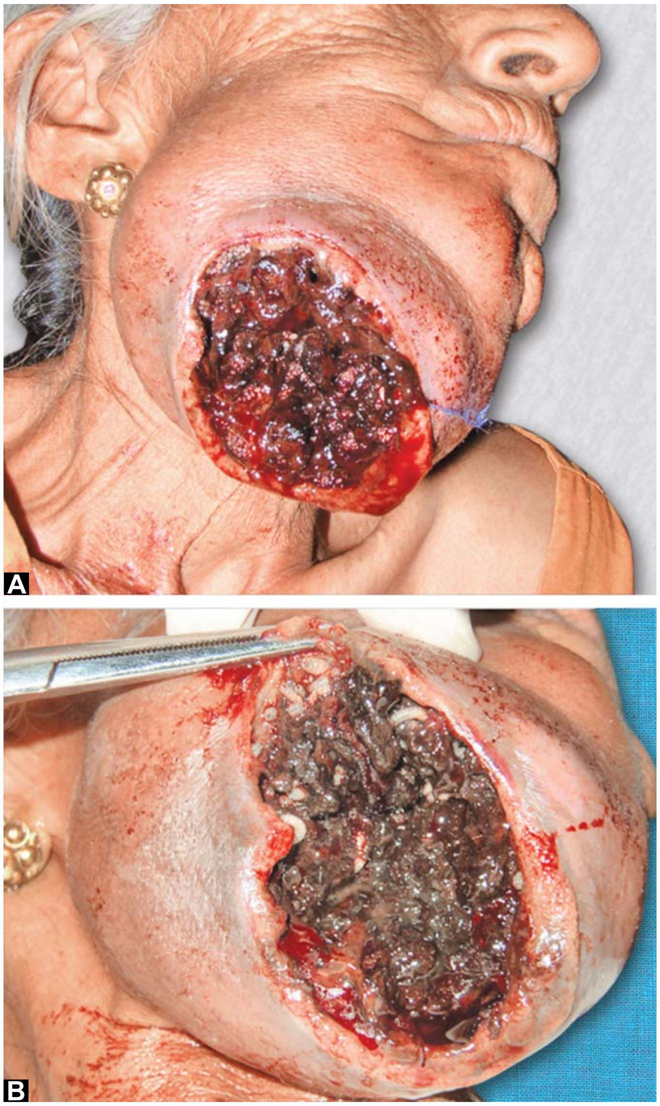

Figs 1.602A and B: Melanoma involving face extensively with destruction. Note the maggots over the surface of tumour. Melanoma is *most aggressive* cutaneous malignancy.

Prognostic factors—overall poor	***Staging as prognostic factor***
• **Tumour thickness**—very important factor	Stage I: > 90% prognosis Stage II: 70%
• Nodal spread—once regional nodes are positive, 85% of patients will have occult distant spread. Number of positive nodes is also important.	Stage III: 35% Stage IV: < 2%
• Ulceration—poor	
• Angiogenesis, vascular invasion—poor	
• In-transit nodules—poor	
• Vertical growth—poor prognosis	
• Metastatic disease—poor	
• Staging	
• Mitotic activity	

Follow-up After Therapy in Melanoma

- In stage I and II disease after treatment, follow up is done at 6 months interval for 3 years. It is done by clinical examination, LDH assay, USG abdomen and chest X-ray.
- In stage III disease, PET scan CT of head, chest and abdomen are indicated.

Sentinel node biopsy

a. Carcinoma breast
b. Melanoma
c. Carcinoma penis

Note: Melanoma also can occur in fishes, dogs and horses.

Happiness is the golden thread that ties the heart of all.

S. Sarcomas

CHAPTER OUTLINE

- Sarcoma
- Liposarcoma
- Fibrosarcoma
- Malignant Fibrous Histiocytoma
- Leiomyosarcoma
- Rhabdomyosarcoma
- Chondrosarcoma
- Haemangiosarcoma
- Lymphangiosarcoma
- Synovial Sarcoma
- Malignant Peripheral Nerve Sheath Tumour
- Kaposi's Sarcoma

SARCOMA

Sarcomas can arise from bone or soft tissue. *Osteosarcoma* is the commonest of all sarcomas. *Soft tissue sarcoma* (STS) which is arising from mesenchyma is the one which will be discussed in this chapter. STS are aggressive tumours which needs multimodality therapeutic approach.

Sarcoma occurs in younger age group compared to carcinoma with fish fleshy gross look with haemorrhage and necrosis. It shows rapid growth, with tendency to show early blood spread. Lymphatic spread is rare even though few sarcomas can spread through lymphatics along with blood spread. Blood spread commonly occurs to lung but liver, skin, brain can also get involved.

Soft tissue is tissue that connect, surround and support the skeletal system which is non-epithelial, extra skeletal tissue excluding reticuloendothelial system, glial tissue but it also includes peripheral nervous tissue by convention. Embryologically it is mainly from mesoderm but few from neuroectoderm.

STS are named based on their tissue which it resembles. Liposarcoma—fat; fibrosaroma—fibroblast; malignant fibrous histiocytoma—mesenchyma/histiocytes; leiomyosarcoma—smooth muscle; rhabdomyosarcoma—skeletal muscle; chondrosarcoma—chondroblast; angiosarcoma—blood vessels.

Features

- Sarcomas are much lesser in incidence compared to carcinomas.
- It occurs in younger age group compared to carcinomas.
- They can arise from bone (osteosarcoma) or from any soft tissues (soft tissue sarcomas) (Mesenchymal tissue).
- They are much more aggressive compared to carcinomas.
- They are rapidly growing tumours with fleshy appearance.
- They are not encapsulated but often are having pseudocapsule.
- They spread through blood especially to lungs often also to other organs.
- Lymphatic spread is not common with certain exceptions.
- Main method of treatment is surgery, i.e. wide excision, amputation.
- In inoperable cases debulking is the accepted method of treatment.
- Chemotherapy is the adjuvant therapy.

Important features of sarcoma

- More aggressive
- Rapidly spreading
- Not very much radiosensitive
- Blood spread
- Painless soft tissue mass is the presentation
- Very vascular

Soft tissue sarcoma

- 1% of adult malignancy
- 15% of paediatric malignancies
- Incidence:
  - 35% occurs in lower limb (commonest site)
  - 15% upper limb, 15% retroperitoneum
  - 10% trunk, 15% viscera, 10% other areas
- **Soft tissue sarcoma (STS)** arises from pluripotent mesenchymal stem cell without in situ changes. Transformation from benign lesion to sarcoma is not observed/disproved now except MPNST (Malignant peripheral nerve sheath tumour)
- Ratio of soft tissue sarcoma to bone sarcoma is 3:1
- Soft tissue sarcoma is more common in males (4:1)
- Ratio of benign soft tissue tumour to malignant soft tissue tumour is 100:1. Most of the sarcomas arise as *de novo*. Per se benign will not turn into malignancy, as it is now found that cell of origin to begin with multiples with anisocytosis in soft tissue sarcoma. In olden days it was accepted to consider benign (precursor tumour) turning into malignancy like lipoma turning into liposarcoma which is disproved now. But this argument does not hold good for nerve sheath tumour and probably lymphoedema turning into lymphangiosarcoma (post mastectomy)
- 50% of STS occurs in extremities called as extremity STS; 35% lower and 15% upper limb
- 3% of sarcoma spreads to lymph nodes
- Soft tissue tumour > 5 cm should be biopsied in suspicious of sarcoma

- Commonest sarcoma of bone is *osteosarcoma* (Note: commonest malignancy of bone is secondaries).
- Commonest soft tissue sarcoma is *liposarcoma / malignant fibrous histiocytoma* (MFH – 25%) overall; in the extremities both *MFH and liposarcoma*; in the retroperitoneum it is *liposarcoma*.
- Commonest visceral (GIT) sarcoma is leiomyosarcoma.
- In genitourinary system leiomyosarcoma is commonest in adults and rhabdomyosarcoma in *paediatric age group*; in uterus leiomyosarcoma; in myocardium angiosarcoma; in hand and foot synovial sarcoma; in skin Kaposi's sarcoma; in head and neck region angiosarcoma.

Note:

- Many liposarcomas arise at sites devoid of adipose tissue. Most rhabdomyosarcomas arise in locations that lack voluntary muscle.
- 40% of soft tissue sarcomas are more than 10 cm in size at the time of presentation; 30% are 5-10 cm and 30% are less than 5 cm.

Incidence of STS

In adults

- MFH and liposarcoma—35-45%.
- Rhabdomyosarcoma—10%.
- Leiomyosarcoma—9-15%.
- Synovial sarcoma—7%.

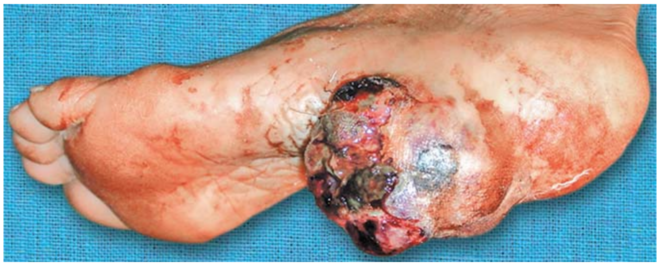

Fig. 1.603: Soft tissue sarcoma foot. Note the vascularity.

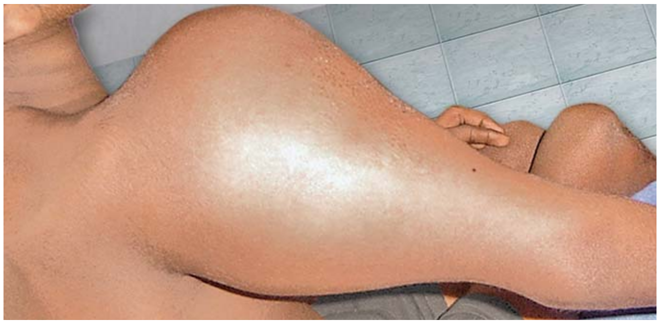

Fig. 1.604: Osteosarcoma of the upper end of humerus in an adolescent boy.

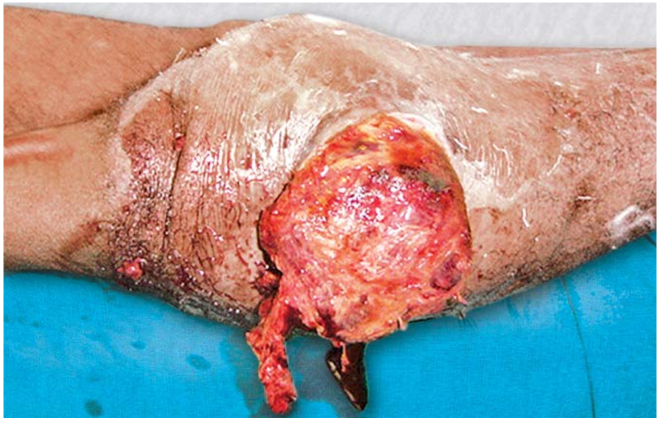

Fig. 1.605: Osteosarcoma upper tibia. Note the size of the tumour and ulceration. It is very vascular.

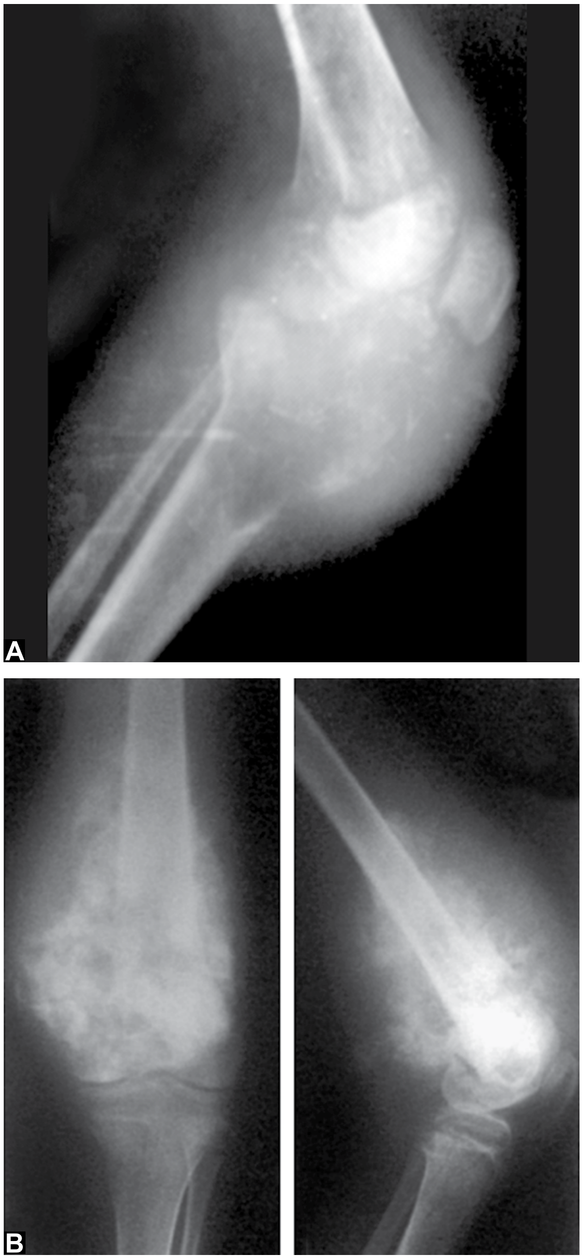

Figs 1.606A and B: X-ray pictures of osteosarcoma in tibia and femur.

- Malignant peripheral nerve sheath tumour (MPNST)—6%.
- Fibrosarcoma—5%.

In children

- Rhabdomyosarcoma, neuroblastoma are common tumors.

Aetiology

- *Genetic*
  - von Recklinghausen disease
  - Gardner's syndrome
  - Tuberous sclerosis
  - Basal cell naevus syndrome
  - Li-Fraumeni syndrome
- *Chemicals*—PVC, tetrachlorodibenzodioxin, arsenic
- *Viral*—HIV in Kaposi's sarcoma, cytomegalovirus
- *Ionising radiation*—malignant fibrous histiocytoma (p53)
- Lymphangiosarcoma in post-mastectomy lymphoedema—*Stewart—Treves syndrome.*
- Osteogenic sarcoma in Paget's disease of bone/exposure to radium
- Retinoblastoma associated sarcoma
- Gorlin's syndrome
- Thorotrast, vinyl chloride, arsenic, pesticides

Clinical Features of Soft Tissue Sarcoma

- Painless swelling of short duration with progressive increase in size—*soft tissue mass*. 30% of patients may present as *pain* during first evaluation.
- Compression of adjacent structures
- Smooth, firm/hard, warm and vascular
- Features of secondaries in lung—cough, haemoptysis and chest pain. *Lung* is the commonest site of secondary.
- Secondaries in *liver* as a principal site especially in visceral STS.
- There are no reliable findings to distinguish benign from malignant swellings.
- One has to maintain a high index of suspicion in any soft tissue mass deep to deep fascia, any soft tissue mass > 5 cm, any new enlarging or symptomatic soft tissue mass.

Investigations

Preoperative evaluation of STS is done:

- To determine the "exact extent" of the tumour.
- To obtain a tissue diagnosis.
- To evaluate metastatic disease.

Tissue Diagnosis

- ***Incision biopsy*** is the *most reliable method* of diagnosis. It provides adequate tissue sample.
- *Trucut biopsy*/core needle biopsy is an *acceptable first diagnostic* step as it is technically easier, not costly, with fewer complications. But it is not useful in visceral STS. Again if it is inadequate one should not be hesitant to go for incision biopsy at the earliest. It is done using 14 gauge needle; often US / CT guide is used.
- Excision biopsy is done *only* if the tumor size is < 3 cm which is cutaneous or subcutaneous wherein wide local re-excision is possible. Otherwise excision biopsy *should be avoided* as it may contaminate the tumour bed and restricts the therapeutic options.

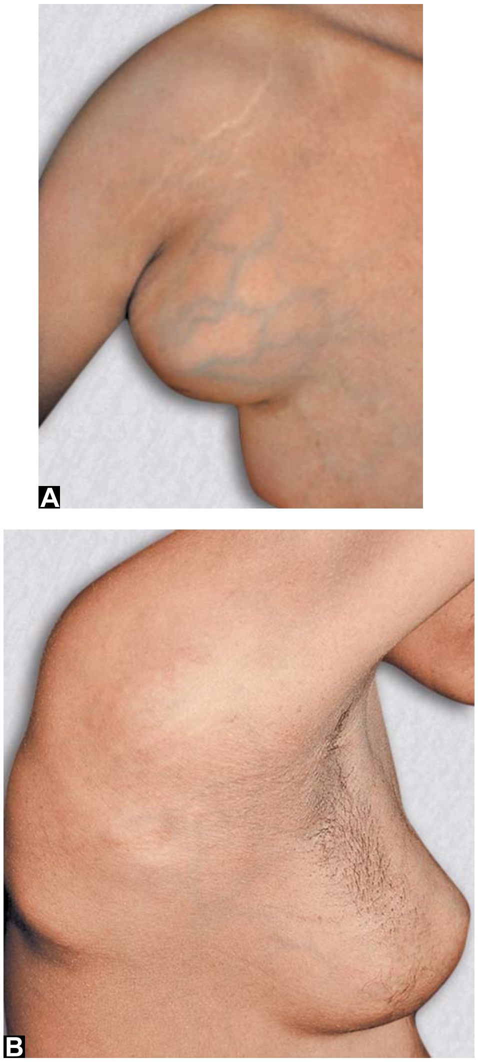

Figs 1.607A and B: Soft tissue sarcoma upper chest wall. Note the vascularity and extent.

***Note*:**

FNAC is of *less value* in STS. It is significant only if it positive. It is useful in local or distant recurrences in documented sarcoma patients, or to evaluate nodal status if enlarged. CT guided FNAC is useful for retroperitoneal/intra-abdominal sarcomas.

Assessment of the Extent of Tumour

Imaging in STS provides a 3–dimensional extent of the tumour and helps in accurate planning of surgical procedure. But it does **not** reliably distinguish between a benign and malignant process. It assesses macroscopic and **not** microscopic extent of the disease. Imaging is essential for metastatic work up.

- *MRI is the investigation of choice* as it determines the vascularity, relation to vessel and fascial planes (extent and invasion). *Advantages of MRI* are—imaging of choice in STS, excellent soft tissue delineation, without radiation, multiplanar imaging possible. Images of skip metastases are possible. *Disadvantages of MRI* are—cost, bone involvement is poorly delineated, claustrophobia, not possible in presence of metal implants and pacemakers.

- CT scan also can be used to see the extent and invasion but not equivalent to MRI. CT scan helps in identifying presence and extent of the soft tissue mass, status of the adjacent structures, with mandatory contrast enhancement isodense masses, vasculature are better delineated. SPECT—3 dimensional reconstruction feasible. *Advantages of CT* are—easy availability, relatively cost effective, best to demonstrate bony involvement, useful for guided biopsies. *Disadvantages of CT* are—cross-sectional imaging, ionizing radiation, inferior soft tissue detail.
- *US is less sensitive* investigation. It is useful in extremity lesions to assess vascular system. It is initial/first investigation done in GI leiomyosarcoma/retroperitoneal sarcoma. It is useful for serial studies, guided trucut biopsy.
- X ray of part is not necessary; it is only used in initial phases to differentiate STS from bony lesion.
- *Angiogram* is traditionally used to delineate adjacent vasculature. It is upstaged by CECT/MRA. It is not necessary to do regularly in all cases but when there is a need for accurate assessment of vasculature it is ideal. It is used in intra-arterial chemotherapy for unresectable tumours.
- CT abdomen is *better* in GI/retroperitoneal sarcomas.

Evaluation of Metastatic Disease

- Chest X-ray is done to look for secondaries.
- CT chest is ideal to see early lung secondaries. It is done in all deep seated, high grade and tumour more than 5 cm in size.
- US abdomen is sufficient to check liver secondaries. But CT abdomen may be better choice. Often CT pelvis is also added.

Other Investigations

- Radionuclide scintigraphy (Gallium-67).
- -p-MRS (p-Magnetic Resonance Spectroscopy) and FDG (Fluor-2-Deoxy Glucose) PET are done to assess the metabolic activity of tumour.

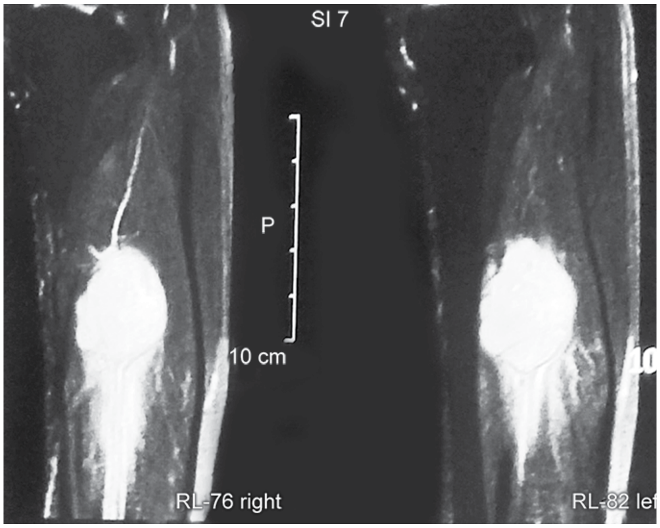

Fig. 1.608: MRI of STS leg.

- Immunohistochemistry and *FISH* (fluorescence *in situ* hybridization).
- Tumour markers.
- Haematocrit, peripheral smear, ESR, serum alkaline phosphatase, serum creatinine.

Incision biopsy for soft tissue sarcoma

- It is the ***ideal tool*** to conclude sarcoma histologically
- Incision should be placed in such a way that it can be included in wide tumor excision at a later period—as biopsy track is always contaminated
- One should achieve absolute / adequate haemostasis to avoid haematomas as tumours are vascular
- Proper site of incision biopsy should be decided
- One should use shortest possible route to tumour while taking incision biopsy; should not violate more than one compartment; should avoid neurovascular bundle. (Injury to vessels and nerves should be avoided)
- Smallest longitudinal incision is used to provide adequate specimen (Incision should be longitudinal in limbs)
- Transverse incision is *contraindicated* in the limbs except over the flexures
- Minimal tissue disturbance and avoiding raising of flaps are crucial (Flaps should not be undermined)
- It is better to use cold knife
- One should avoid crushing/distorting the specimen
- Frozen section/Imprint specimen can be used to avoid sampling error.
- Drains are *not used* routinely in incision biopsy. If used exit *near*/close to the wound and not away/distant from the incision biopsy wound.
- Excise the biopsy tract and drain site *enbloc* during the definitive procedure.
- Immunohistochemistry and cytogenetics are possible

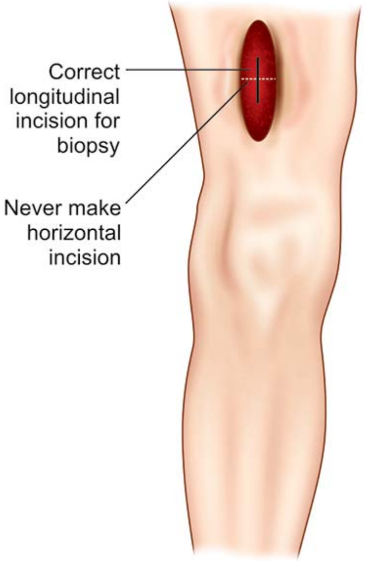
Correct longitudinal incision for biopsy
Never make horizontal incision

Fig. 1.609: Incision for biopsy in STS should be longitudinal so that it can be included in future surgical wide excision. Never place transverse incision for biopsy.

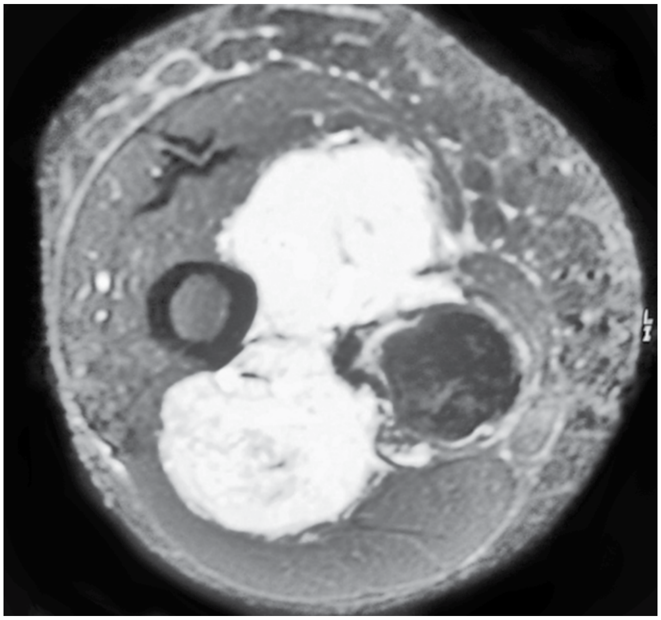

Fig. 1.610: MRI picture of the soft tissue tumour thigh encasing vessels partially and eroding the bone. Patient needed high level, above knee amputation.

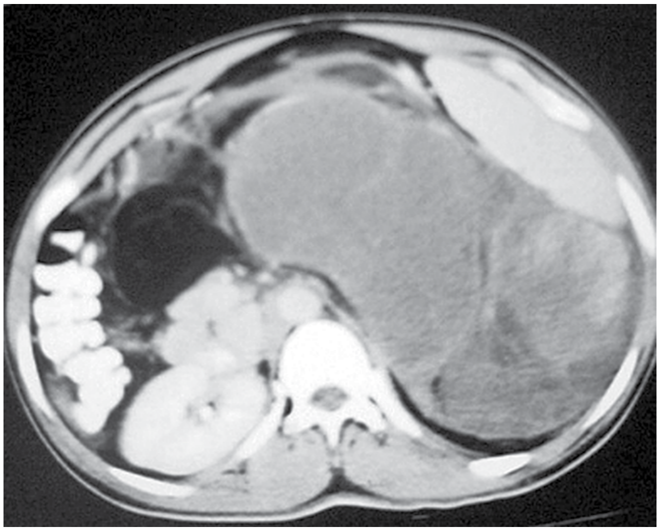

Fig. 1.611: CT picture of retroperitoneal leiomyosarcoma.

Staging

Staging of the soft tissue tumour is done depending on the tumour size, nodal status, metastasis and histological grading of the tumour **(GTNM staging).**

TNM staging of soft tissue sarcoma

Grade (G)	*Tumour (T)*
Gx Cannot be assessed	Tx Cannot be assessed
G1 Well-differentiated	T0 No primary
G2 Moderately differentiated	T1 Size < 5 cm or 5 cm (maximum dimension)
G3 Poorly differentiated	T1a: Superficial
	T1b: Deep tumour
G4 Undifferentiated	T2 Size > 5 cm
	T2a: Superficial
	T2b: Deep tumour
Regional lymph nodes (N)	*Distant spread (M)*
Nx Nodes cannot be assessed	Mx Cannot be assessed
N0 No nodes	M0 No distant spread
N1 Nodes present	M1 Distant spread present
Note: Superficial tumour is outside the superficial fascia. Deep tumour is deep to superficial fascia or any tumour invading the superficial fascia.	

Stage I – G1, 2; T1a, 1b, 2a, 2b; N0; M0.
Stage II – G3, 4; T1a, 1b, 2a; N0; M0.
Stage III – G3, 4; T2b; N0; M0.
Stage IV – Any G; any T; N1; M0. OR Any G; Any T; N0; M1.

Grading of sarcomas

Low	Desmoid, dermatofibrosarcoma, liposarcoma. Here < 15% chances of metastases
High	Synovial, Rhabdomyosarcoma, Angiosarcoma Malignant fibrous histiocytoma, Extraosseous chondrosarcoma. Here > 50% chances of metastases
Undetermined	fibrosarcoma, Leiomyosarcoma

Note:

- STS is an aggressive, invasive, destructive growth with high recurrence and distant metastases rate. Tumour shows *fish flesh* cut appearance.
- *Basis used for classification of STS*—STS being highly heterogeneous group of tumour is classified based on adult tissue it resembles; i.e the *type of tissue formed* rather than from the type of origin.
- ***Grade*** is the single most important factor in staging. It denotes the "biological aggressiveness" of the sarcoma. It predicts the likelihood of metastases.
- Nodal metastases are rare in STS (3%). It has the same prognosis as M1 disease hence staged as Stage IV.
- M1 is – Distant metastases.

Differential diagnosis for soft tissue sarcoma

- Haematoma
- Abscess
- Aneurysm
- Myositis

Treatment

Principles of Treatment

- ***Surgery*** is the *main treatment* modality. Amputation rate for STS has come down drastically from 50% in 1960 to 5% at present. It is also because of proper ***adjuvant radiotherapy*** following function/limb sparing complete excision, application of microvascular surgeries. Neoadjuvant chemotherapy, perioperative/postoperative RT also play a major role.
- In *low grade tumour* without any spread—functional/limb sparing complete wide excision is sufficient without any adjuvant therapy. If microscopic margin is positive for tumour then postoperative External Beam RT (EBRT) is given.
- In *high grade tumour < 5 cm* size, function sparing wide excision with more than 1 cm clearance margin is sufficient. If clearance margin is less than 1 cm or shows microscopic

positive margin, then perioperative brachytherapy OR postoperative EBRT is given.

- In *high grade tumour which is between 5-10 cm* size, function/limb sparing complete wide excision with perioperative brachytherapy OR postoperative EBRT is given.
- In *high grade tumour more than 10 cm* in size, initially neoadjuvant chemotherapy; then functional/limb sparing complete wide excision with postoperative brachytherapy and EBRT should be given.
- All limbs should be *conserved* if possible but with curative intent.

Surgery

Enneking classification of surgical procedures

- *Intralesional excision*—done inside pseudocapsule very high recurrence 100%—not used.
- *Marginal excision*—en bloc resections through the reactive zone-high recurrence rate 70%
- *Wide excision* means en bloc resection done through normal tissues beyond the reactive zone; it means if the margin is less than 5 cm; tumour is never visualised during surgery; it has local recurrence rate of 30%. Wide margin is classified as adequate if margin is at least beyond 1cm outside the reactive zone or inadequate if margin is within 1 cm.
- *Radical excision*—if the margin is more than 5 cm outside the reactive zone. It is like compartment excision with very low recurrence rate.

Other procedures

- Compartmental excision; function/limb sparing.
- Vascular resections with vascular reconstruction.
- Amputation.

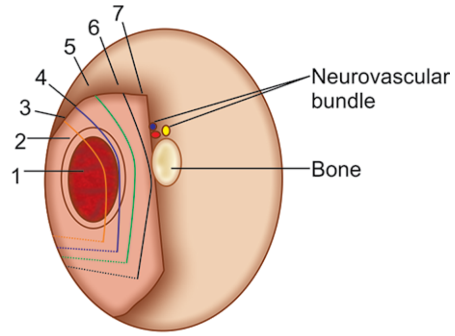

1. Tumour mass.
2. Reactive zone.
3. Intralesional excision.
4. Marginal excision.
5. Wide excision–clearance up to 5 cm.
6. Radical excision–clearance > 5 cm.
7. Compartment excision.

Fig. 1.612: Different surgical approaches for STS.

A thin barrier is considered to be a 2 cm thickness of normal tissue; a thick barrier is 3 cm thickness; and joint cartilage is said to be equivalent to a 5 cm thickness margin. A surgical margin that is outside a barrier, with normal tissue between the barrier and the reactive zone of the tumour, is considered to be curative.

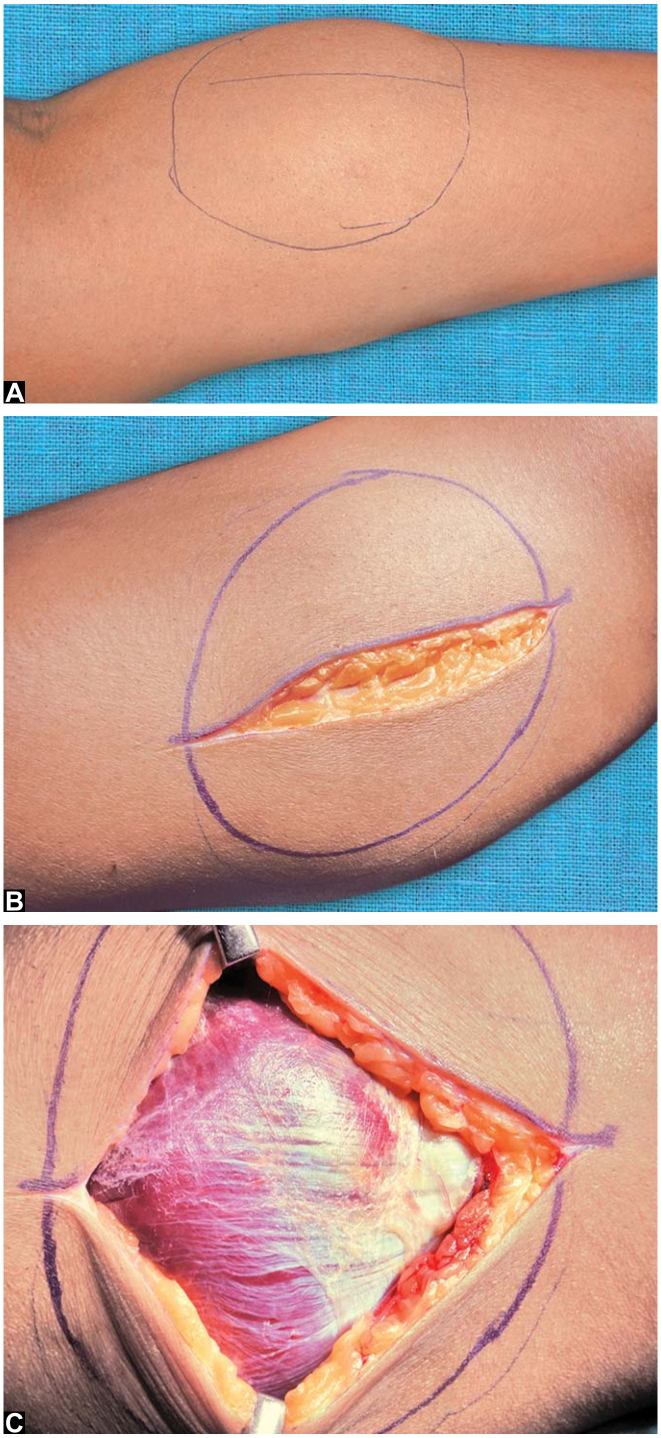

Figs 1.613A to C

People can be divided into three groups: (1) Those who make things happen, (2) Those who watch things happen, (3) Those who wonder what's happening.

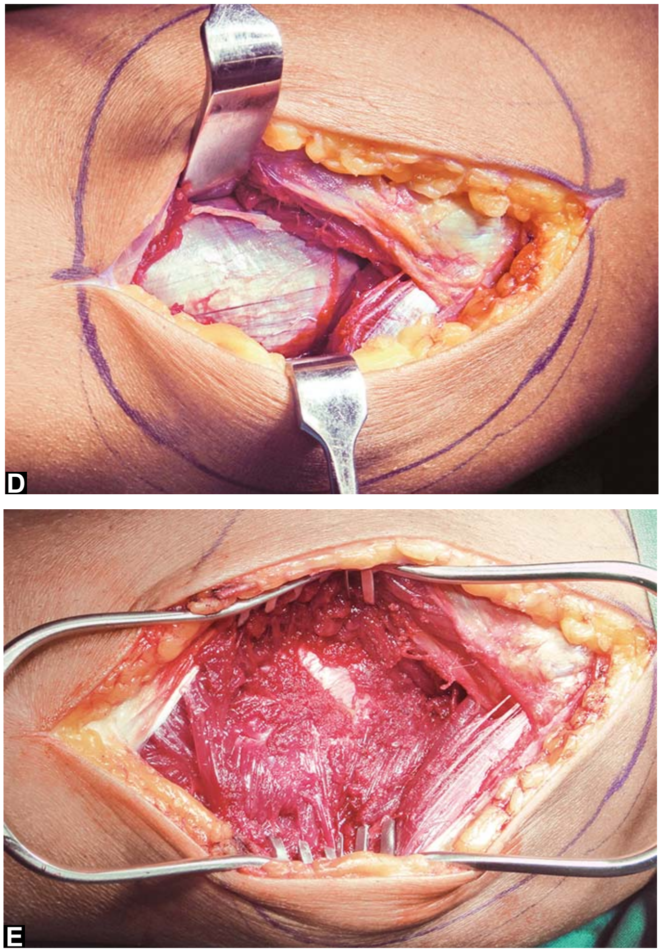

Figs 1.613D and E

Figs 1.613A to E: Surgical approach to a deep seated sarcoma in leg—deep to gastrocnemius and soleus.

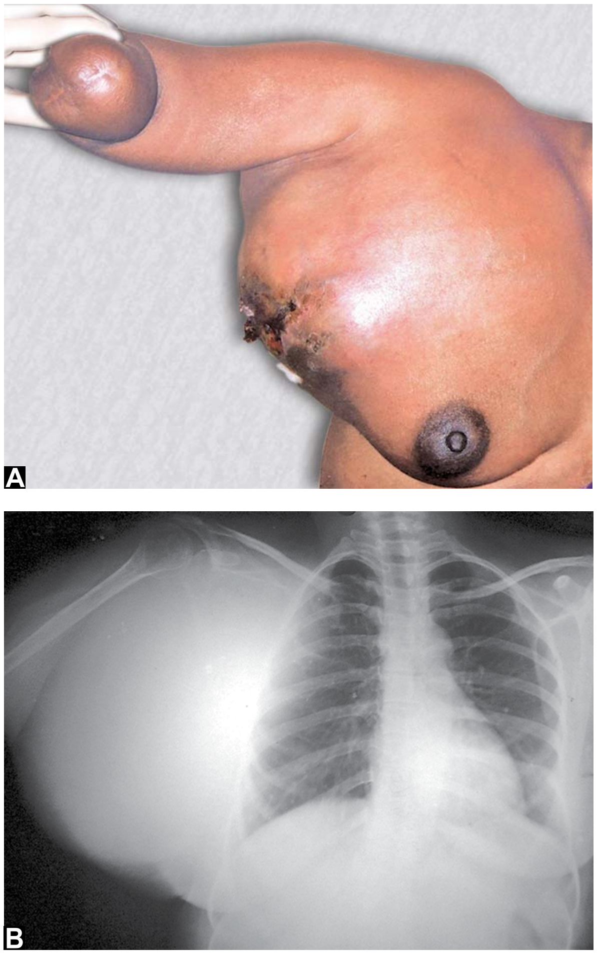

Figs 1.614A and B: Recurrent chest wall tumour, could be secondaries, as patient was earlier amputated for soft tissue tumour hand (forearm amputation). X-ray of the same patient shows large tumour mass.

- **Wide local excision** with clearance of 2 cm (minimum need is 1 cm) or more with preservation of function is needed. Depth clearance is also important. 3-5 cm clearance even though was practiced in olden days, is not necessary.
- **Compartment resection** is a radical limb saving procedure. Here muscle group of one compartment (anterior, posterior or medial) is resected entirely from its origin to insertion with the tumour. It is done only when tumour is intracompartmental. It is not suitable when tumour is extracompartmental or many compartments are involved or encased to major neurovascular bundle.
  Basis for compartmental excision is—STS rarely penetrate anatomical barriers unless it is very advanced.
- Amputation is done in large tumours of upper or lower limbs.

Indications for amputations in soft tissue sarcoma

- Major neurovascular encasement
- Bone involvement
- Multiple compartment involvement
- Limb itself is diseased like lymphoedema
- Recurrence with multicentricity

  - *Radical amputation* is done as disease has not spread systemically which should be confirmed by CT chest, abdomen and pelvis. In metastatic disease there is no need to do amputation as long term survival is not possible except if primary is fungating and distressing.

Tumor site	*Level of amputation*
Foot	Below knee
Leg	Above knee
Thigh-middle and lower third	Hip disarticulation
Thigh proximal third	Hemipelvectomy
Buttock	Hemipelvectomy
Hand and wrist	Below elbow
Forearm	Above elbow
Distal arm and elbow	Shoulder disarticulation
Axilla and shoulder girdle	Forequarter amputation

- Debulking surgery is useful in large advanced tumours like retroperitoneal sarcomas.

Radiotherapy

- ***Preoperative radiotherapy*** followed by wide excision—neoadjuvant RT.
- ***Postoperative radiotherapy*** is commonly used because of less tumour burden and less wound problems. Titanium clips are placed during surgery at high risk areas to identify the site to concentrate proper RT.
  - *Brachytherapy* is very effective in local control of the tumour. Initially precise mapping of the area is done in the

operation theatre. Loading catheters are placed in surgical field peroperatively. Later these catheters are loaded with iridium 192. Dose is 45 Gy to tunbour bed for 6 days.
  - Permanent radioactive sources can also be placed to the area.
  - *Postoperative external beam radiotherapy* (EBRT)—it is quiet effective and used in high grade tumour more than 5 cm often with brachytherapy. Dose is 70 Gy-25 fractions.
- Palliative external radiotherapy can be given to prevent bleeding, fungation and to reduce pain in advanced cases. It is also used in secondaries in brain, bone.
- *Primary radiotherapy* alone (radical) is of less beneficial in soft tissue sarcoma, but now it is used with more favorable results.

Radiotherapy in soft tissue sarcoma

- Brachytherapy is given in high grade tumour
- External beam radiotherapy is used in low grade tumour
- All tumours more than 5 cm need adjuvant radiotherapy (external beam)
- Preoperative radiotherapy is also beneficial
- Deep seated tumour; high grade; size more than 5 cm needs chemoradiation

Chemotherapy

- Chemotherapy drugs—*VAC (Vincristine, Adriamycin, Cyclophosphamide*) are commonly used. Other drugs *ifosfamide*, *dacarbazine* are used in combination with above drugs. *Mesna* is used as a protection for haemorrhagic cystitis. Chemotherapy is used when tumour is more than 5 cm or high grade. Usually *postoperative chemotherapy* is given. MAID regime is used especially in recurrent STS. Mesna, Adriamycin, Ifosfamide, Dacarbazine drugs are used. Its response rate is 50%. But survival benefit is controversial.
- *Neoadjuvant chemotherapy* is used to make the primary tumour better operable. It makes eventual surgery better; provides early treatment for micrometastasis; gives idea about the response for chemotherapy. Drugs used are adriamycin and ifosfamide.
- *Isolated limb perfusion using cytotoxic drugs and tumour necrosis factor* with hyperthermia is also often used.
- Chemoradiation is a good alternate adjuvant therapy used.

Distant Spread

- *Pulmonary metastasis* can be treated with wedge resection, segmentectomy, lobectomy, pneumonectomy. Surgery is done only when primary is well-controlled. Radiotherapy and chemotherapy are also tried. More than three number metastases in lung signify poor prognosis.

Prognostic factors in STS

- Size > 5 cm—important factor
- High grade
- More than one compartment involvement
- Deep tumours and multicentric
- Neurovascular invasion
- Lung secondaries
- Clearance margin

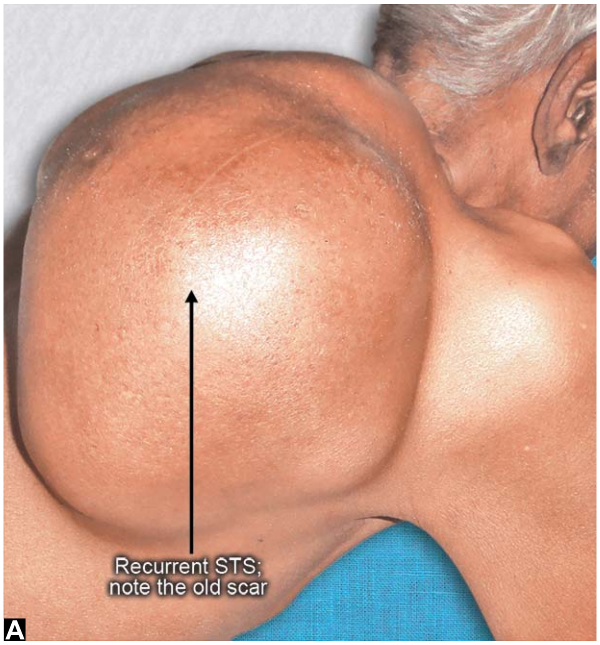

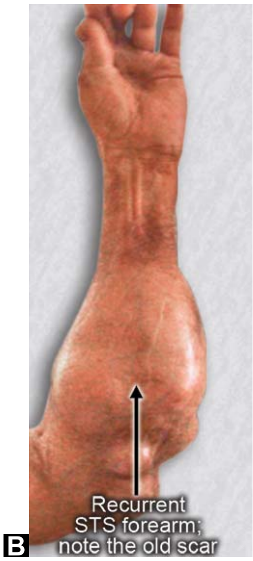

Figs 1.615A and B: Recurrent soft tissue sarcoma over scapular region and forearm in two different patients. Old surgical scar is seen.

LIPOSARCOMA

- It is the ***commonest type*** of soft tissue sarcoma arising from the fat cells (of primitive mesenchymal cells).
- Lipoma as a precursor tumour for liposarcoma which was thought earlier is *disproved*/not accepted now.
- It is 20% of all soft tissue sarcomas.

Common Sites

1. Thigh—commonest site.
2. Retroperitoneum.
3. Back.
4. Shoulder.

Small minds are first to condemn great ideas.

Types

1. *Well-differentiated*—common between 50-70 years age group; common in extremities (75%); metastasis is rare.
2. *Dedifferentiated*—common between 50-70 years age group; common in retroperitoneum (75%); metastasis is high.
3. *Myxoid*—common at 25- 45 years age; common in extremities (75%); high rate of metastasis.
4 *Round cell type*—common at 25- 45 years age; common in extremities (75%); high rate of metastasis.
5. *Pleomorphic* (5%)—poor prognosis.

Microscopically, it contains lipoblasts with 'signet ring' malignant cells. It is low grade type.

Spread is to lungs.

Treatment is wide excision or radiotherapy with surgical debulking in places where complete removal of tumour is not possible like in retroperitoneal liposarcoma.

FIBROSARCOMA

- It can arise from ***the bone*** or from ***soft tissues.***
- It is common between 30-55 years; common in deep soft tissues of lower extremities with intact overlying skin.
- It is arising from fibroblasts. Intramuscular and intermuscular fibrous tissue, fascial envelops, aponeurosis and tendons are common origin.
- Commonest site is thigh.
- Spindle fibroblasts with *'herring bone'* pattern are typical on microscopy.

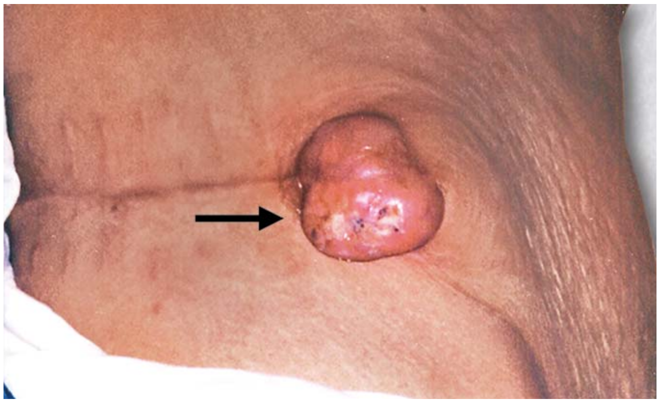

Fig. 1.616: Recurrent dermatofibrosarcoma protuberans.

Types

1. *Well-differentiated.*
2. *Poorly differentiated.*
3. *Dermatofibrosarcoma protuberans*. It is common in trunk. *DFSP* is an intermediate grade fibrohistiocytic tumor with nodular cutaneous mass; common in adult male; common in trunk and proximal extremities having slow growth without deep muscle invasion. Skin is taut, nodular and commonly ulcerated.
4. *Aggressive fibromatoses* are variant of fibrosarcoma which are locally malignant in which desmoid tumour is also included.

Fibrosarcoma is slow growing tumour which attains large size.

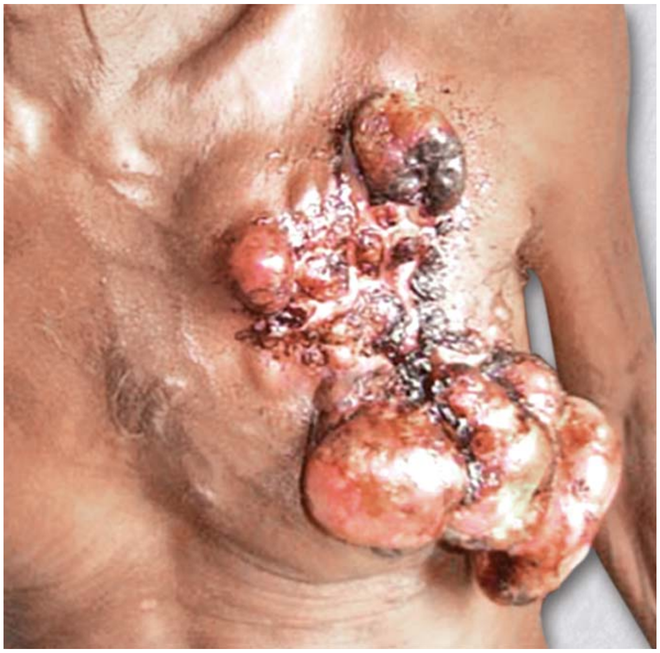

Fig. 1.617: Malignant dermatofibrosarcoma left side of the chest wall.

Clinically it is smooth, hard, warm, localised tumour.

It compresses or infiltrates the adjacent structures like neurovascular bundle.

MALIGNANT FIBROUS HISTIOCYTOMA (MFH)

- It is group of malignant soft tissue tumours with a fibrohistiocytic appearance.
- MFH is one of the most common STS in adult.
- 70% occur in skeletal muscles.
- Lower extremity is the common site.
- It presents as solitary, multilobular lesion.
- It spreads along the fascial planes or between muscle fibres which is the probable reason for local recurrence.
- Myxoid/giant cell/inflammatory/angiomatoid/pleomorphic are the types. It is common in adults and elderly.

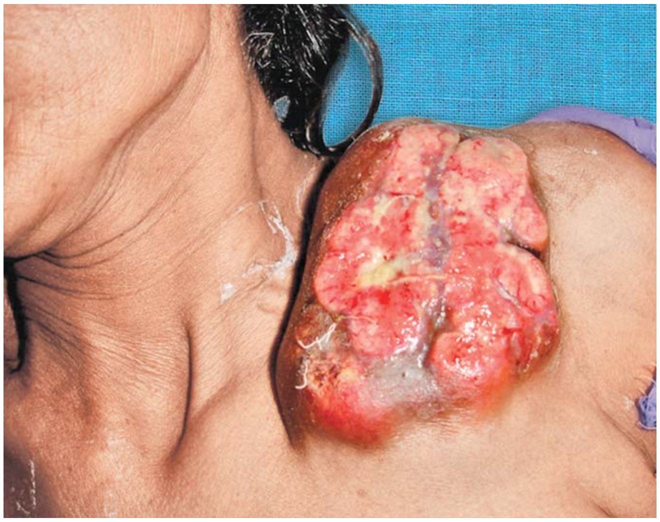

Fig. 1.618: Dermatofibrosarcoma.

LEIOMYOSARCOMA

- It arises from smooth muscle. Cut section shows whorled appearance.
- It constitutes 10% of STS; common after 60 years.
- Two third occurs in women.
- It is undetermined grade—aggressive.
- It is common in retroperitoneum and viscera, but can occur in limbs and skin. Uterus is also common site.
- Recurrence is common. It has got poor prognosis.
- It can occur in the piloerector muscle of skin; inferior vena cava; pulmonary artery.
- Desmin and actin are the most common immunohistochemical stains.

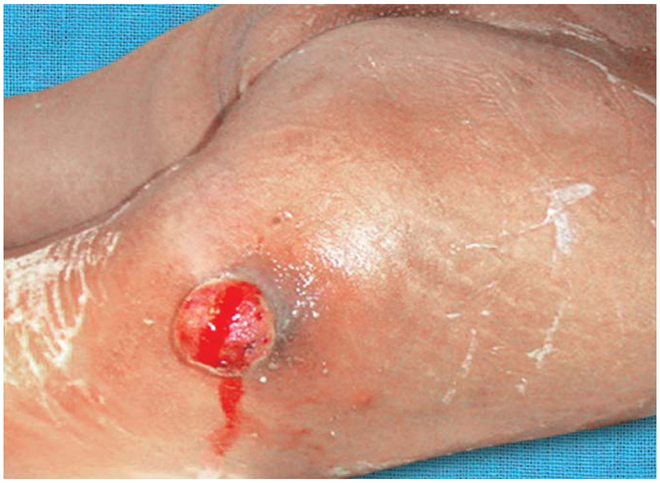

Fig. 1.619: Large soft tissue tumour over gluteal region with ulceration and vascularity. It was confirmed as malignant fibrous histiocytoma. Patient was successfully operated.

RHABDOMYOSARCOMA

- It arises from striated muscle.
- It is common in *head and neck*, upper thigh and arm.
- *It is commonest sarcoma in children.*
- It can occur in retroperitoneum, pelvis and genitourinary tract.
- It is common in males.

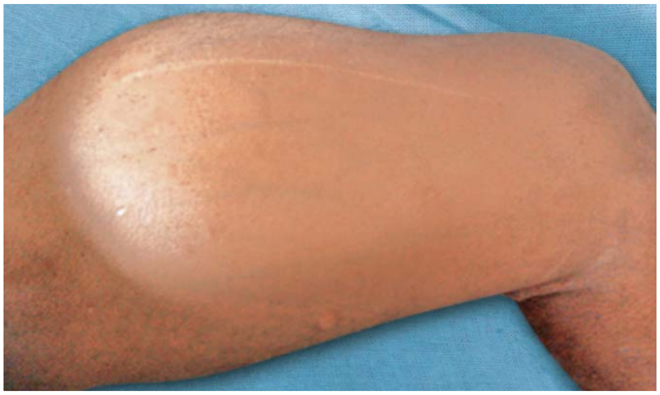

Fig. 1.620: Recurrent soft tissue tumour thigh. Note the scar of old surgery.

Types

1. *Pleomorphic*—msot common type of rhabdomyosarcoma. It is common in adult, aggressive with poor prognosis.
2. *Embryonal*—common in infants and children—*is seen in viscera like urinary bladder.*
3. *Botryoidal*—slow growing polypoidal—respond well to chemotherapy and radiotherapy.
4. *Alveolar*—chromosomal translocation is common.

It is more aggressive tumour with poorer prognosis (High grade).

It also metastasises to lymph nodes.

CHONDROSARCOMA

- It arises from chondroblasts.
- It attains large size with slow growing nature.
- Common sites are ribs, flat bones.

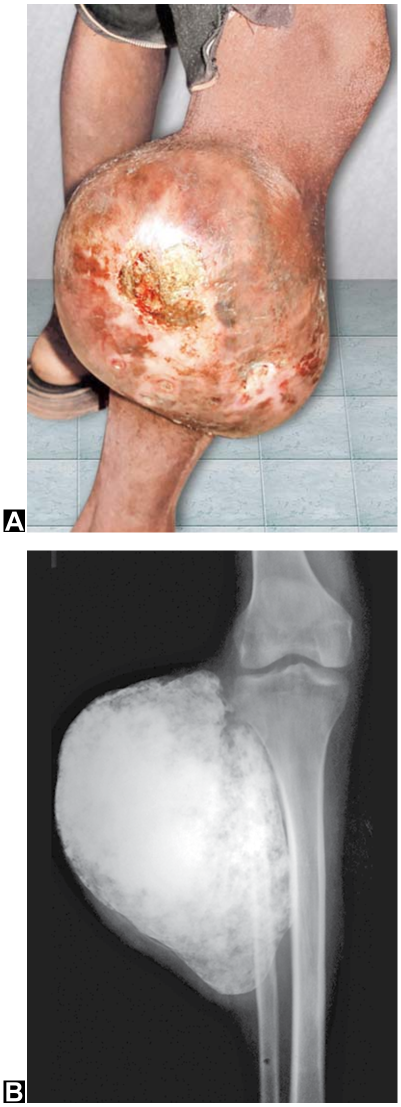

Figs 1.621A and B: Chondrosarcoma leg lateral aspect. X-ray shows calcified tumour. It was successfully removed as limb salvage procedure.

HAEMANGIOSARCOMA

It originates from blood vessel endothelium.

Types

Malignant haemangioendothelioma.
Malignant haemangiopericytoma.

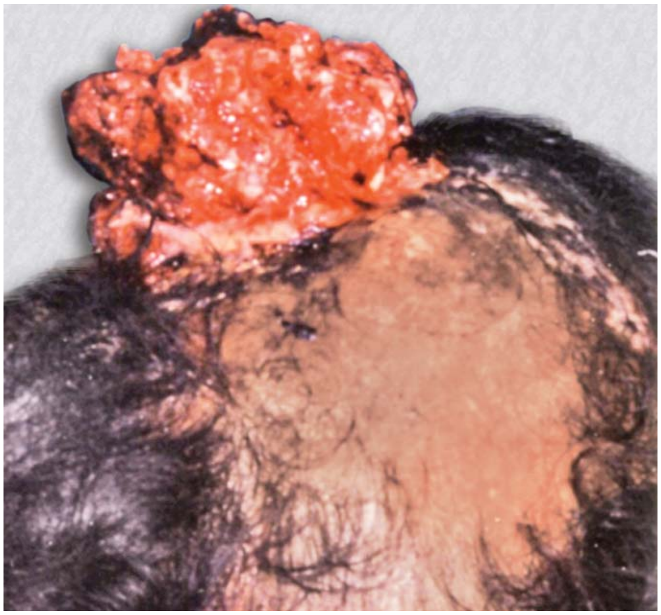

Fig. 1.622: Aggressive haemangiosarcoma of the scalp.

LYMPHANGIOSARCOMA

- It arises from lymph vessel endothelium.
- It commonly occurs after radical lymph node dissection *Stewart-Treves syndrome*.

SYNOVIAL SARCOMA

- It is the 4th commonest type. It occurs 15-40 years of age.
- Origin need not be from synovium.

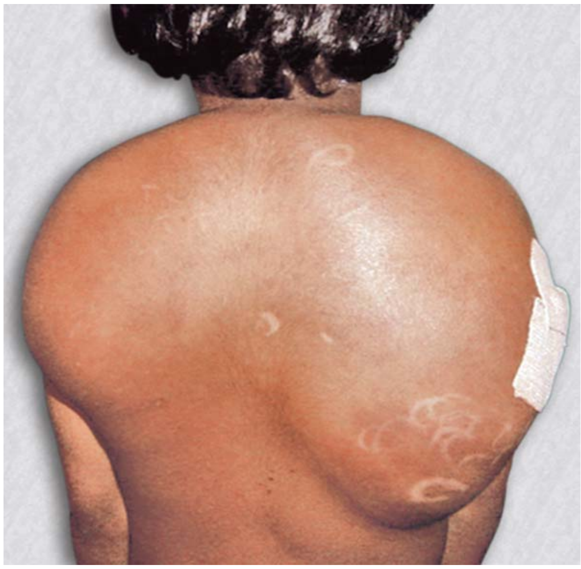

Fig. 1.623: Synovial sarcoma.

- 85-90% occurs in lower limb; head, neck and shoulder are the next common site.
- It is common in thigh, leg, shoulder, hand and foot.
- Occasionally it can occur in the abdominal wall and retroperitoneum.
- It is common in young individuals.
- It occurs adjacent to joint but uncommon to involve the synovial sheath of the joint.
- It spreads both through blood as well as through ***lymph nodes (20%).***
- It is very aggressive soft tissue sarcoma (High grade).
- Calcification with or without ossification is common—10%.
- Synovial sarcoma may be biphasic or monophasic.
- It possess specific chromosomal translocation (96-100%) – t(X; 18)(p11.2; q11.2).

MALIGNANT PERIPHERAL NERVE SHEATH TUMOUR (MPNST)

- It arises from peripheral nerves.
- It shows differentiation along nerve elements.
- MPNST replaced olden terms—malignant schwannoma, neurofibrosarcoma, neurogenic sarcoma, malignant neurilemmoma.
- 20-50% arises from neurofibromatosis type 1 (latent period 15-20 years). MPNST is an exception to sarcoma will not arise from benign precursor.
- Common sites are major/proximal nerve trunks.

KAPOSI'S SARCOMA

- It is malignant blood vessel tumour of multicentric origin arising from vascular smooth muscle or pericytes.
- It is seen commonly in HIV patients due to immunosuppression.
- Primary tumour commonly occurs in skin, mucous membrane, lymph nodes or viscera.
- It is linked with Human Herpes Virus 8(HHV8) as causative agent.

Types

1. *European Kaposi's sarcoma*: It is common in old age. It is first described by Kaposi in 1862. It mainly involves skin especially lower extremity. Visceral involvement is rare.
2. *African Kaposi's sarcoma*: It occurs commonly in children and young individual. It involves skin and lymph nodes commonly. It resembles lymphoma.
3. *Transplant associated Kaposi's sarcoma*: It is due to drug induced immunosuppression. It involves mainly skin and often regresses once immunosuppression is discontinued.
4. *AIDS associated Kaposi's sarcoma*: It occurs in 30-40% of AIDS patients. It is common in homosexuals. It has got wide, disseminated involvement with metastases. It is very aggressive. It is often associated with lymphoma and other malignancies.

Note:
Kaposi's sarcoma is not found in transfusion related AIDS.

Clinical Features

- Multiple reddish-blue nodules in the skin with ulceration over the nodule.
- Lymph node enlargement.
- Koebner phenomenon is common in areas of trauma.

Differential Diagnosis

- Lymphomas.
- Cutaneous angiomatoses.
- Mycobacterial infection of skin.

Investigations

- Biopsy from the skin lesion.
- Tests for HIV infection.

Sarcomas which also spread to lymph nodes

1. Synovial sarcoma
2. Lymphangiosarcoma
3. Rhabdomyosarcoma—alveolar type
4. Ewing's sarcoma
5. Angiosarcoma
6. Epithelioid sarcoma
7. Carcinosarcoma of uterus

Treatment

- Irradiation.
- Chemotherapy. Drugs used are adriamycin, bleomycin and vinblastine.
- Antiretroviral therapy.
- Interferons.

Peace is so hard to find because it is under your nose.

T. Amputations

> *"Amputation is one of the meanest yet one of the greatest operations in surgery, i.e. mean—when resorted to where better may be done. Great—as the only step to give comfort and prolong life."*
>
> **—Sir William Ferguson**

CHAPTER OUTLINE

- Amputation
- Complications of Amputations
- Prosthesis

AMPUTATION

Indications

- Gangrene due to atherosclerosis, embolism, TAO, diabetes, ergots.
- Trauma: To save life in crush injuries.
- Neoplasms: Osteosarcomas, Marjolin's ulcer, melanomas.
- Gas gangrene.
- Severe sepsis.
- Occasionally severe elephantiasis, madura foot, when all other methods have failed to help.
- **Dead, dying, devitalised tissues.**
- Severe deformity congenital or acquired.

Types

It can be:
- *Non-end bearing/side bearing*—Weight is taken up by the joint.
- *End bearing/cone bearing*—Weight is taken up by the body.

It can be:
- Weight bearing.
- Non-weight bearing.

It can be:
- *Provisional* amputation with flap – later final formal amputation my be required.
- *Guillotine* amputation which always requires revision formal amputation.
- *Formal* amputation – is definitive one.

Types of Flaps

- *Long posterior flap* in below-knee amputation.
- *Equal flaps* in above-knee amputation.

Ideal Stump

- Should heal adequately by 1st intention.
- Should have rounded, gentle contour, with adequate muscle padding.
- Should have sufficient length to bear prosthesis.
  - For B-K 7.5 (minimum) to 12.5 cm from tibial tuberosity
  - For above and below elbow 20 cm stump.
  - For A-K 23 cm from greater trochanter.

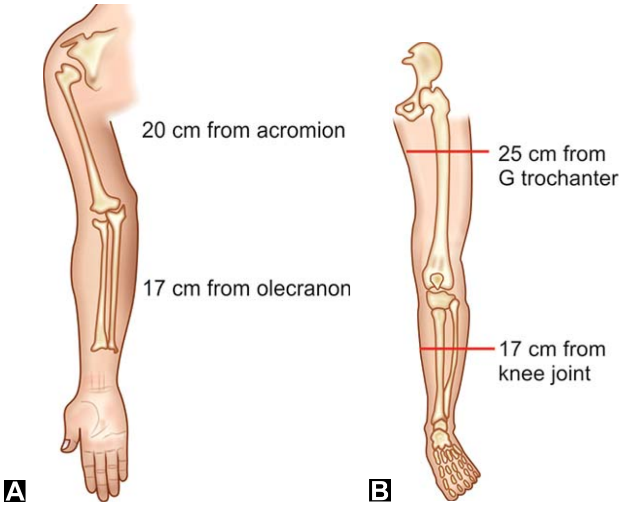

Figs 1.624A and B: Different levels of amputation in upper limb and lower limb.

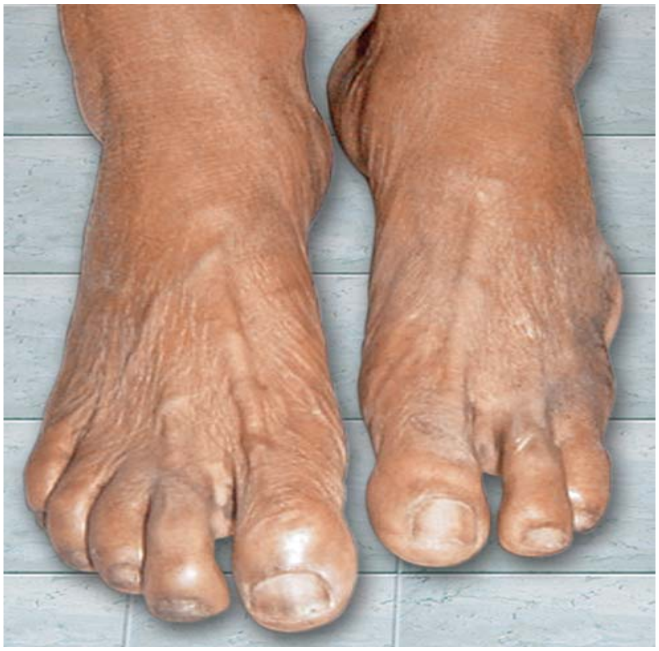

Fig. 1.625: Little toe amputation for gangrene. Wound has healed well.

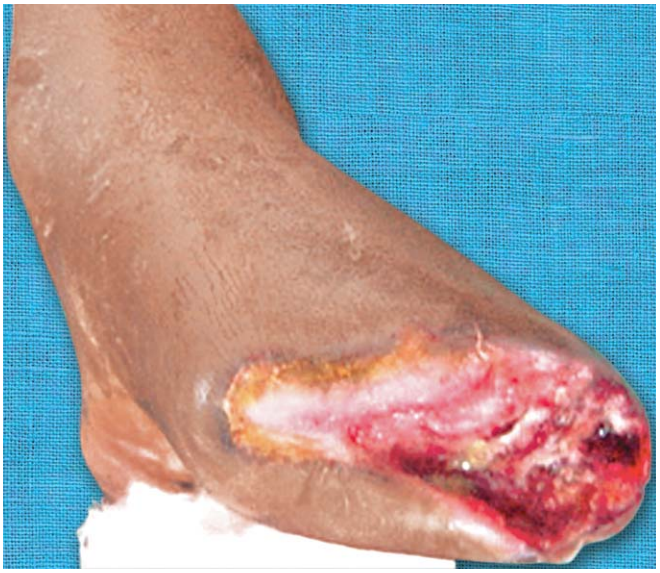

Fig. 1.626: Forefoot amputation.

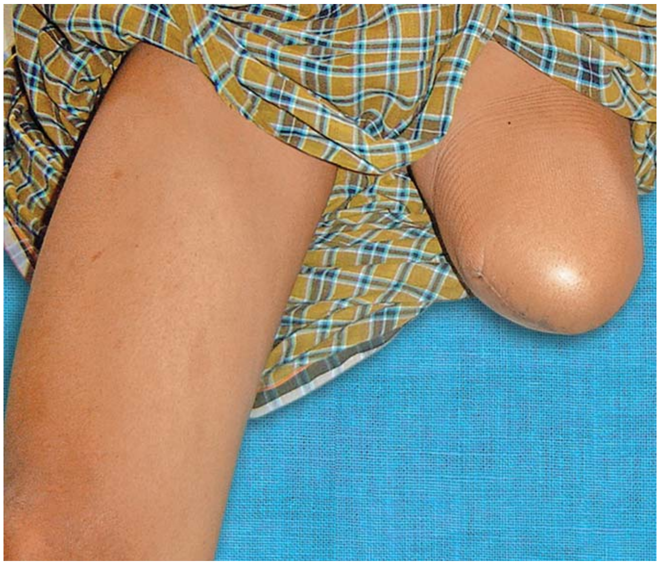

Fig. 1.627: Above-knee amputation done for osteosarcoma of the upper end of the tibia. Here equal flaps are used.

- Should have thin scar which does not interfere with prosthetic function.
- Should have adequate adjacent joint movement.
- Should have adequate blood supply.
- Scar should be in a place where it is not exposed to pressure.
- Scar should be freely mobile over underlying tissues. Skin and scar should be freely mobile over the underlying bone. It is achieved only if deep fascia is closed properly. Scar and skin should be free to achieve free movement of the prosthesis. Socket of prosthesis with mobile skin creates a piston to bone to move like a joint.
- Skin should not be infolded.
- Redundant soft tissue should not be there.
- Stump should be free from tenderness and conical.

Conical bearing

- Here healing is by primary intention
- Bone should not be projecting
- Myoplastic
- No neuroma
- Scar should not be tender
- Proximal joint should be supple

Evaluation of the Patients who need Amputation

- Haematocrit, control of anaemia by transfusing blood/packed cells.
- Control of infection using antibiotics.
- Decision of level of amputation by skin temperature, arterial Doppler.
- Informed consent should be taken.
- Plan for prosthesis and rehabilitation by physiotherapist and rehabilitation team.

Different incisions for amputation

- *Circular incision amputation*—skin and muscles are divided circularly at a lower level than that of bone
- *Elliptical/oval incision* amputation
- *Racquet incision amputation*—for digital disarticulation
- *Amputation using flaps*—it may be of equal flap (above knee amputation) or with long posterior single flap (below knee amputation). Total length of single flap or combined length of two flaps should be equal to 11/2 times the diameter of the limb at the line of amputation. Flap should be semicircular to get a conical stump, not rectangular

Principles in Amputation

- Adequate blood supply of the flap should be maintained.
- Proper marking of the skin incision is essential.
- Tourniquet *should not be used* if amputation is done for vascular diseases.
- Proximal part of the flap contains muscle component but distal part should contain only skin and deep fascia.
- Flap length should be adequate; not short. It should be ideally semicircular not rectangular to get a conical stump.
- Nerve should be pulled down and cut using a sharp knife and allowed to retract into the soft tissue otherwise neuromas may develop.
- In crush injury/entrapment injury/sepsis—guillotine amputation is done. Later skin is pulled down by using skin traction, eventually to have better skin coverage.
- Bone should be cut with beveling and all sharp margins should be rounded.
- Postoperatively regular dressings are done. Patient is mobilised using axillary crutches. After 3 months, once scar has matured and stump has become supple, proper prosthesis is fit. *Berlamont* first started immediate postoperative fitting

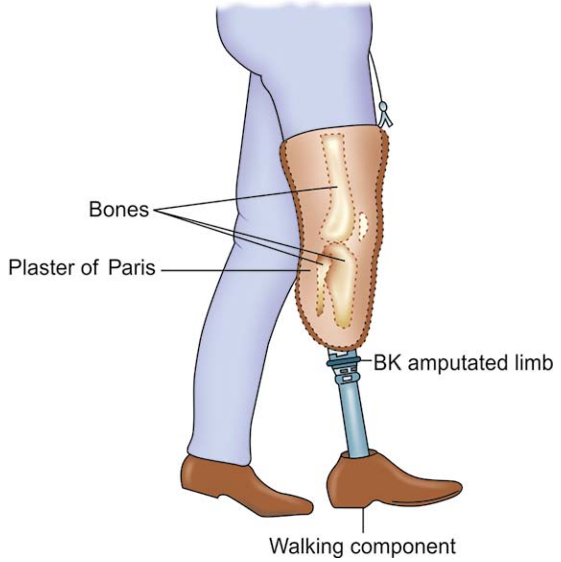

Fig. 1.628: Pylon is plaster cast with walking component used for immediate mobilisation after below-knee amputation.

of prosthesis to leg for early mobilisation. *Plaster pylon* is applied to the stump and a prosthetic extension is fit to facilitate partial weight bearing immediately after surgery. It has got more stump complications and so it has not become popular.

- Stumps can be side bearing (sutures are on the side); end bearing/conical (sutures are on the end) or cylindrical.
- Postoperatively active exercise should be given to the proximal joint so that prosthesis can be fit to it properly.
- If there is sepsis especially in gangrene limb, flaps should be left open or loosely sutured otherwise flap necrosis occurs.
- Proper anatomy of muscles and neurovascular bundle around should be known in all amputations.

Different Amputations

- ***Ray amputation***
  Amputation of toe with head of metatarsal or metacarpals.

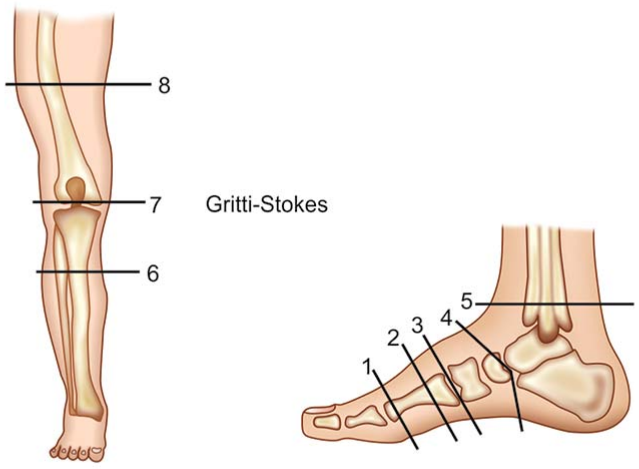

1. Ray's
2. Transmetatarsal (Gillies)
3. Tarsometatarsal (Lisfranc's)
4. Midtarsal (Chopart's)
5. Syme's
6. Below-knee (Burgess)
7. Transcondylar
8. Above-knee

Fig. 1.629: Different levels of amputation in lower limb.

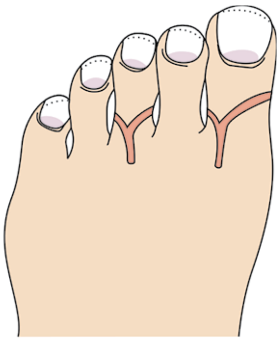

Fig. 1.630: Ray amputation for toes—racquet incision.

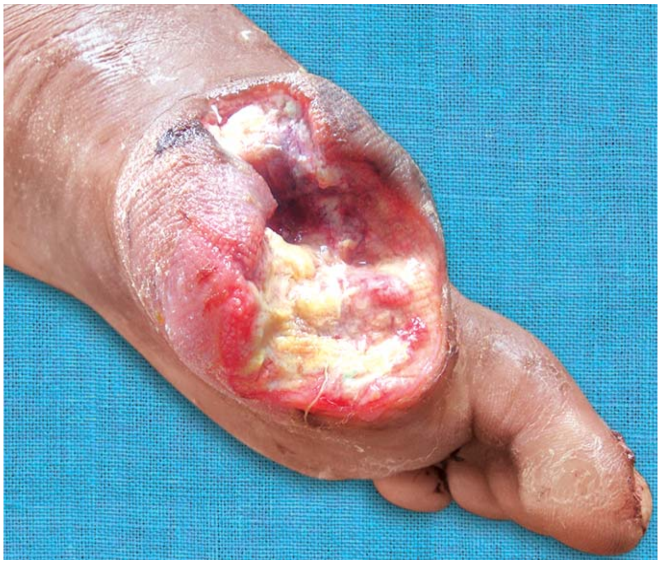

Fig. 1.631: Ray's amputation done for gangrene of great toe. It is usually not sutured.

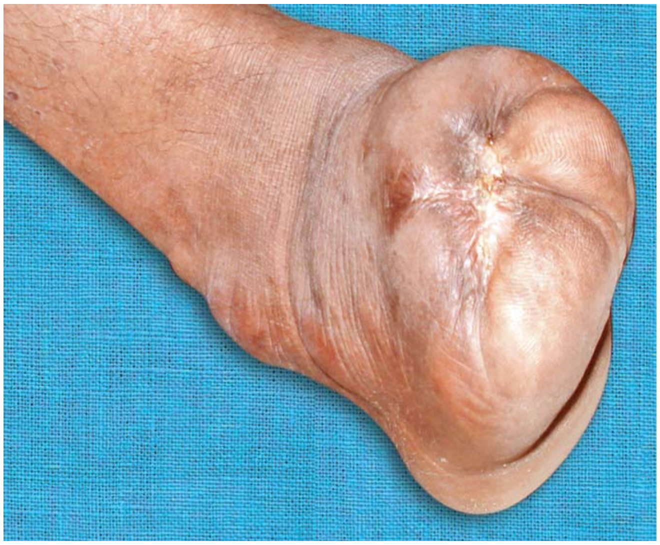

Fig. 1.632: Forefoot amputation.

- ***Transmetatarsal amputation (Gillies')***
  Here amputation is done proximal to the neck of the metatarsals, distal to the base.
- ***Lisfranc's amputation*** (***Tarsometatarsal amputation***)

Here tarsometatarsal joint is disarticulated with a long volar flap. It needs a surgical boot. But there is inevitable development of equinovarus deformity. So stabilisation of midtarsal and ankle joints is needed. In Hey's modification, 2nd metatarsal is cut at base instead of disarticulation.

- ***Chopart's amputation (Midtarsal amputation)***

  Here talonavicular joint and calcaneo cuboid joints are disarticulated. Tibialis anterior muscle is sutured to drilled talus bone. A long volar flap is used and immobilised for 6 weeks after surgery.

- ***Syme's amputation***
  - It is removal of the foot with calcaneum and cutting of tibia and fibula just above the ankle joint with retaining heel flap (dividing both malleoli). Heel flap is supplied by medial and lateral calcaneal vessels (branches of posterior tibial artery). *Elephant boot* is used for the limb after Syme's which is inexpensive. Many patients walk well with Syme's stump without difficulty. It is presently mainly used in trauma (crush injury) and malignancies in distal part of the foot. In vascular diseases, calcaneal vessels may not be adequate to maintain the viability of the flap. While raising the flap, knife should be very close to the calcaneum so as to avoid injury to calcaneal vessels and to maintain the viability of the flap.
  - In *Wagner's method* deep fascia of heel is sutured to drill holes made on the anterior edge of tibia and fibula. *Above knee cast* is essential.
  - *Advantages are*—it is an end bearing stump having good proprioception. Patient can walk without prosthesis. Low energy consumption ambulation is possible.
  - *Disadvantages*—posterior displacement of heel pad; poor cosmesis.
  - *Boyd's amputation*—anterior part of the calcaneum is excised (osteotomy) just distal to the peroneal tuberosity and calcaneotibial arthrodesis is created. It is done to prevent posterior migration of heel pad.

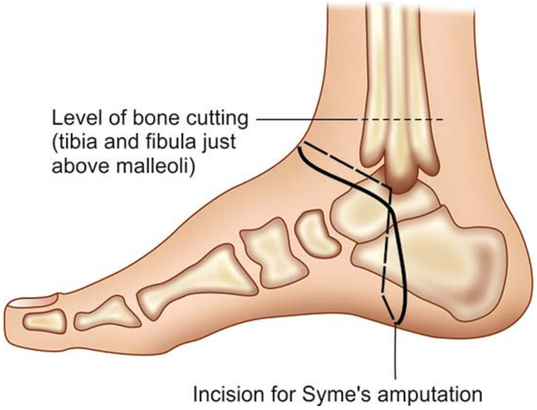

Fig. 1.633: Syme's incision and level of amputation.

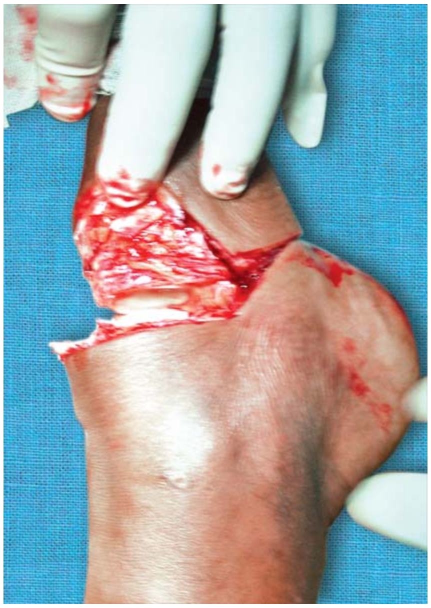

Fig. 1.634: Incision for Syme's amputation.

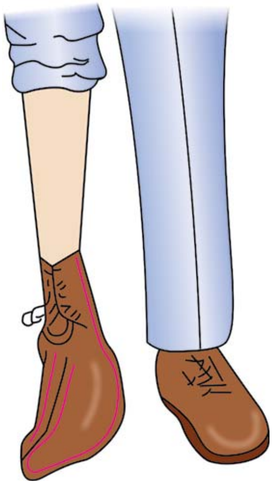

Fig. 1.635: Elephant boot used after Syme's amputation.

- ***Modified Syme's amputation***

  Here heel flap is elliptical. Tibia and fibula are divided slightly higher. But variation here is the elliptical flap.

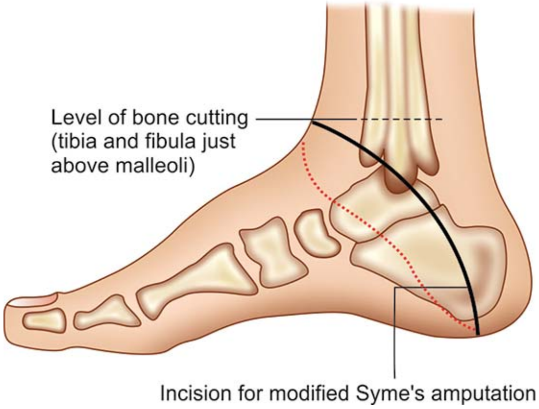

Fig. 1.636: Modified Syme's incision is elliptical one.

Imagination is the highest kite one can fly.

♦ ***Pirogoff‘s amputation***

It is like Syme’s amputation except posterior part of the calcaneum is retained along with heel flap. It provides longer stump than Syme’s amputation.

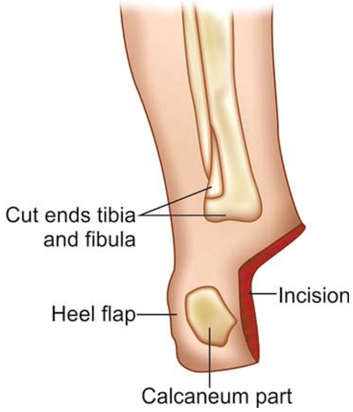

Fig. 1.637: Pirogoff’s amputation.

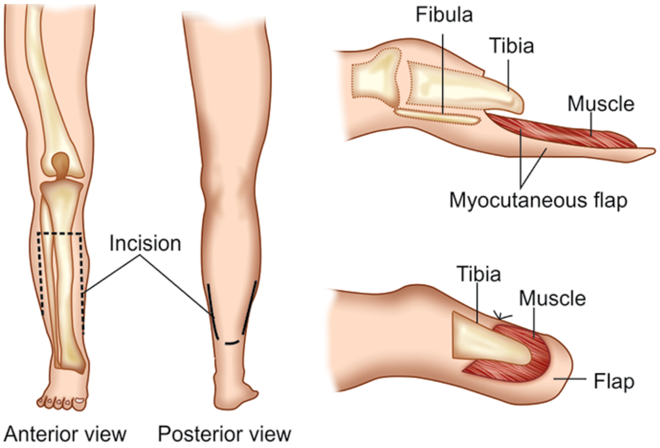

Fig. 1.638: Below-knee amputation technique—incision, flap and levels of cutting bones.

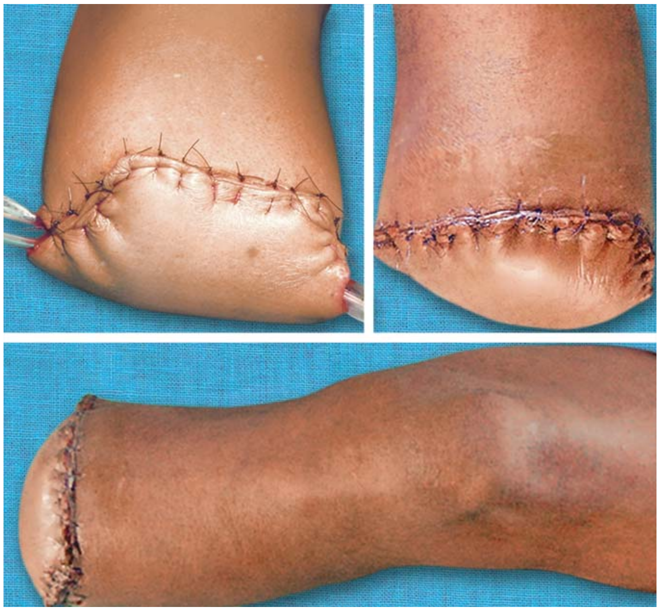

Fig. 1.639: Below-knee amputation—long posterior flap. Note the placing of the drain.

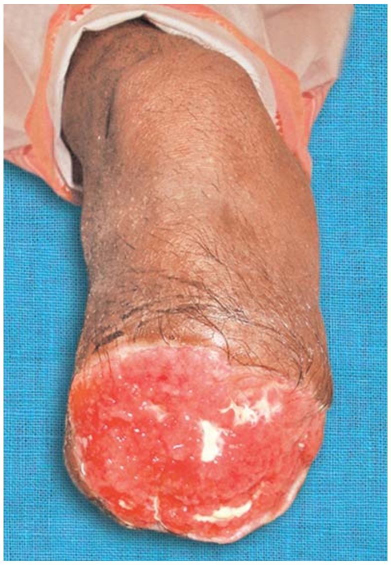

Fig. 1.640: Below-knee amputation flap necrosis has occurred which eventually granulated with repeated dressings. Area is ready for skin grafting.

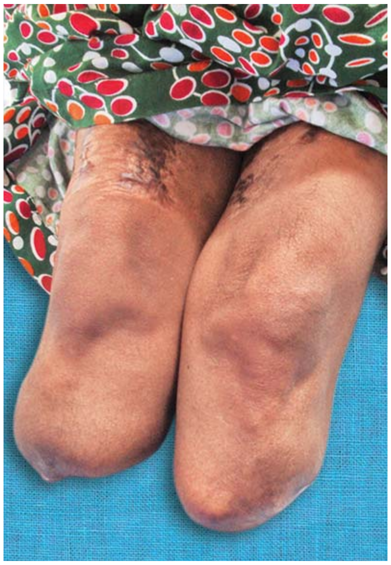

Fig. 1.641: Bilateral below knee amputations done for vascular causes with diabetes.

♦ ***Below-knee amputation***

Here we use a long posterior flap with scar placed over anterior aspect is used. Prosthesis placement is better here with greater range of movements without limp and without support. It is also called as *Burgess amputation*. Fibula should be divided first, higher than the proposed site of cut of tibia otherwise its sharp end will press on the skin flap. Tibial stump should be beveled anteriorly. Posterior muscles are sutured across the bone end, to the periosteum in front. In more proximal type of below knee amputation, fibula often is removed to allow the proper use of flap. Length of the flap should be 11/2 times the circumference of the site (around 12 cm). Stump length is 14-17 cm from knee joint. Minimum length required for prosthesis is 8 cm. If need to extend more proximally, it is better to do above knee amputation. Modern artificial limbs like *suction socket prosthesis* are used now.

- ***'Peg-leg' amputation***
    - It is amputation 5 cm below the knee level – proximal most below knee amputation. It is *not practiced* nowadays. Here anterior flap is rotated posteriorly like a hood and patient kneels and bears weight which is well-accustomed to pressure. It is done whenever prosthesis cannot be used probably due to economic causes (in developing countries).

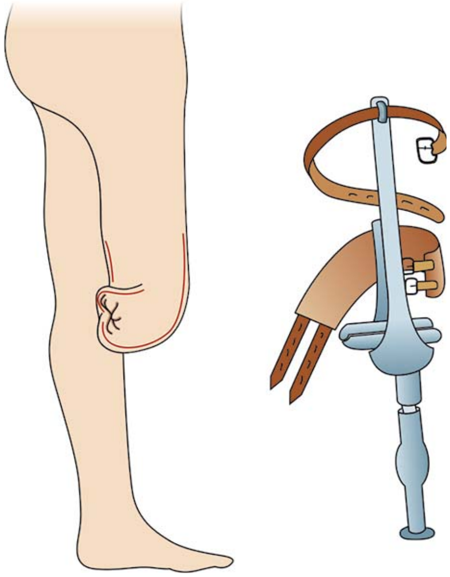

Fig. 1.642: 'Peg-leg' amputation is below-knee amputation done close to the knee joint (5 cm stump) with folding flap and metallic support down. It is not used now.

- ***Transcondylar-Gritti-Stokes amputation*** with long posterior flap.
    - Femur is divided just above the articular surface and patella is anchored to the divided femur. There is risk of nonunion between patella and femur. This procedure is *no longer performed.*
- ***Above-knee amputation***
    - Usually equal anterior and posterior flaps are used. Lower third and middle third level amputations are done. Ideally the required length of the femur as stump is 25 cm from the tip of the trochanter. Femur length lesser than 10 cm is not possible and here one needs to do hip disarticulation. In children as growing epiphysis of femur is in lower end, it is essential to preserve as much length of femur as possible. It is done in ischaemia, trauma, sepsis, gangrene which is spreading above. Often patient might earlier have undergone below knee amputation but now as indicated need above knee amputation. It is usually contraindicated in children (done only in undue definitive indication) or if stump is less than 7.5 cm.
    - *Advantages* are technically easy, healing chances are better and faster. *Disadvantages* are cosmetically poor, rehabilitation is difficult, and fitting of prosthesis is not proper, patient needs a third support for walk with often a limp.

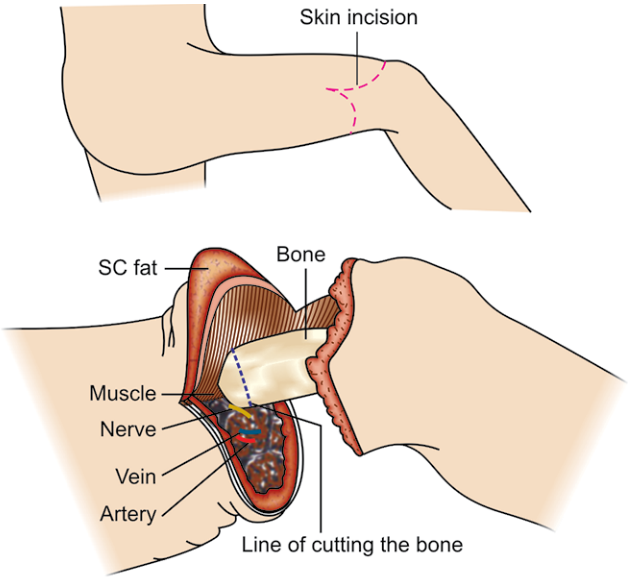

Fig. 1.643: Above-knee amputation, incision and approach.

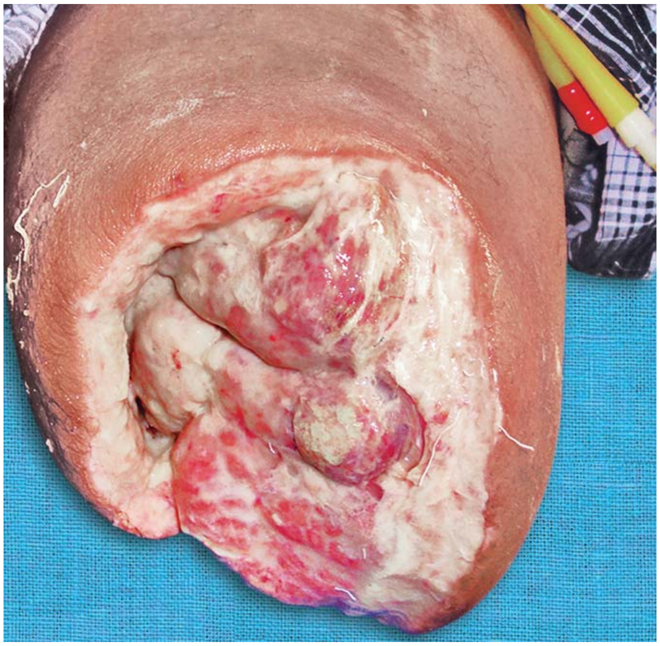

Fig. 1.644: Above-knee amputation wound which is infected in a diabetic patient.

- ***Hip disarticulation***
  It is done whenever it is not possible to save the minimum 10 cm length of the femur. Incision used is either single posterior flap—Solcum's approach (better) or anterior racquet incision—Boyd's approach.
- ***Hind quarter amputation***
  Inter innominate abdominal amputation (*Sir Gordon Taylor's* amputation): Removal of one side pelvis with innominate bone, pubis, muscles and vessels. Original ligation of common iliac artery is modified to individual ligations of external and internal iliac vessels. Internal iliac artery is ligated beyond the origin of the superior gluteal artery to keep the large posterior flap viable. *Now hind quarter name*

The most popular Jaipur foot was designed by PK Sethi.

is replaced by hemipelvectomy. It may be *standard hemipelvectomy* with classic gluteal flap; *extended hemipelvectomy* with removal of posterior part of the sacrum; *conservative hemipelvectomy* with retaining part of the pubis, ilium bones on that side. *Internal hemipelvectomy* is new method wherein *hemipelvectomy* is done with preserving the limb.

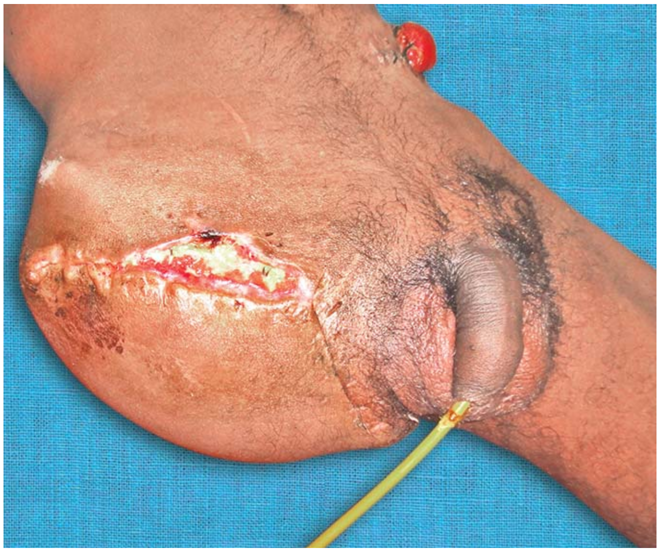

Fig. 1.645: Hindquarter amputation done for run over of vehicle over pelvis, right thigh. Patient also underwent colostomy. Patient survived with severe morbidity.

- ***Krukenberg's amputation***
  Done in upper limb following any trauma. Here forearm amputation is done in such a way that it creates a gap between radius and ulna like a claw to have a hold or grip.
- ***Interscapulothoracic amputation*** *(Forequarter amputation)* (*Littlewood's* posterior approach or Berger's anterior approach):

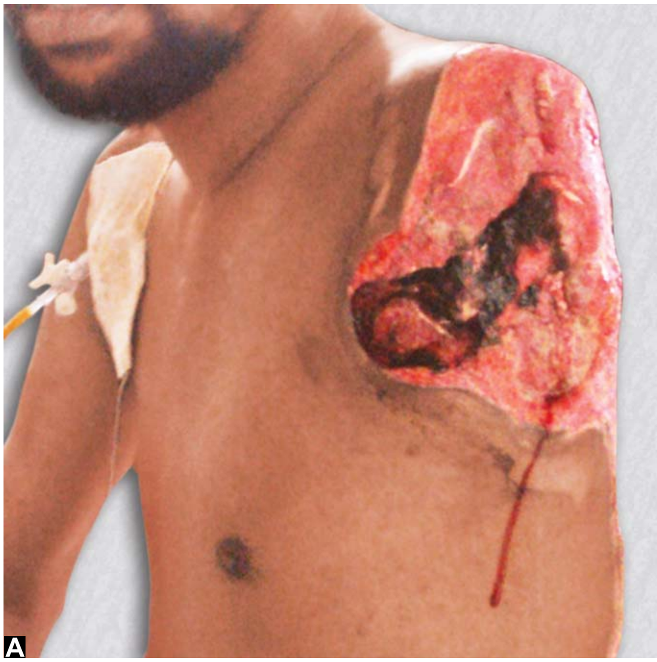

Fig. 1.646A

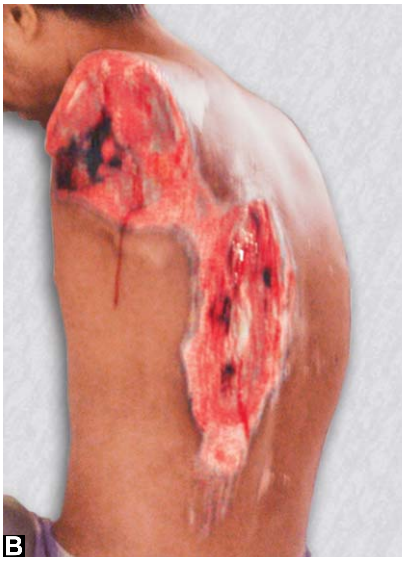

Fig. 1.646B

Figs 1.646A and B: Forequarter amputation done for electric burn which caused extensive damage to upper limb.

It is removal of entire upper limb with scapula and lateral 2/3rd of the clavicle and muscles attached to it. It is done in malignancies involving scapula, upper part of humerus and near shoulder joint.

- In emergency conditions like severe sepsis, gas gangrene and machinery entrapment, ***Guillotine amputation*** is done without suturing. Suturing is done at later period. All tissues are divided at same level.

Postoperative Period

- Physiotherapy is advised.
- Regular dressings are done.
- Crutch is used initially, after 3 months prosthesis is placed.
- Rehabilitation is important.

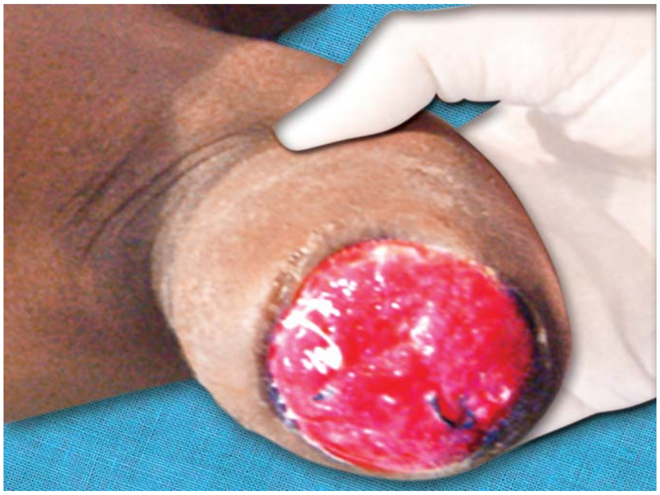

Fig. 1.647: Upper limb *Guillotine amputation*—above elbow done for trauma induced gas gangrene.

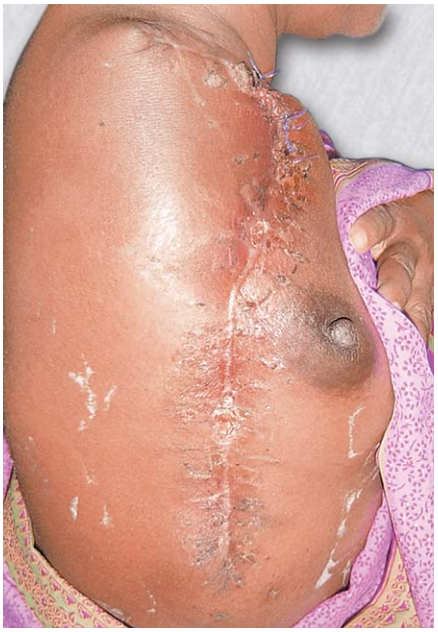

Fig. 1.648: Forequarter amputation done for sarcoma proximal to shoulder joint.

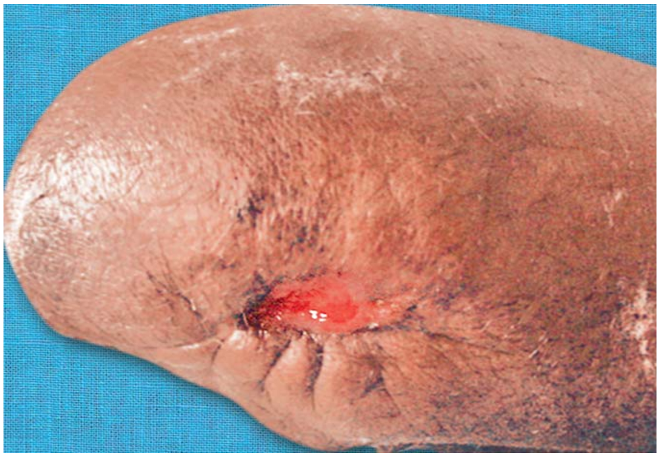

Fig. 1.649: Above-knee amputation stump having small sinus.

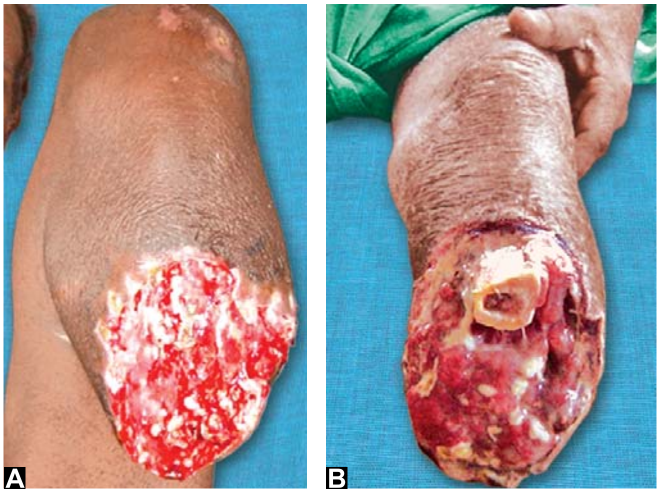

Figs 1.650A and B: (A) Below-knee amputation stump infected. (B) Tibia is exposed.

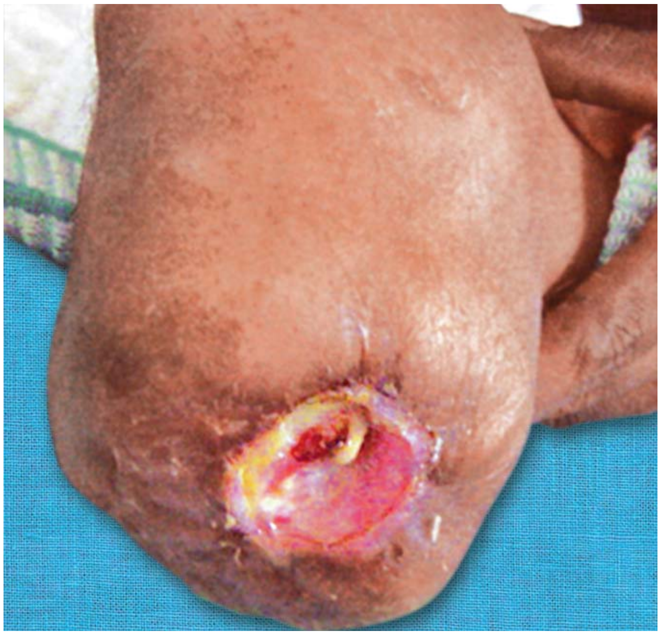

Fig. 1.651: Ulcer in the below-knee amputated stump.

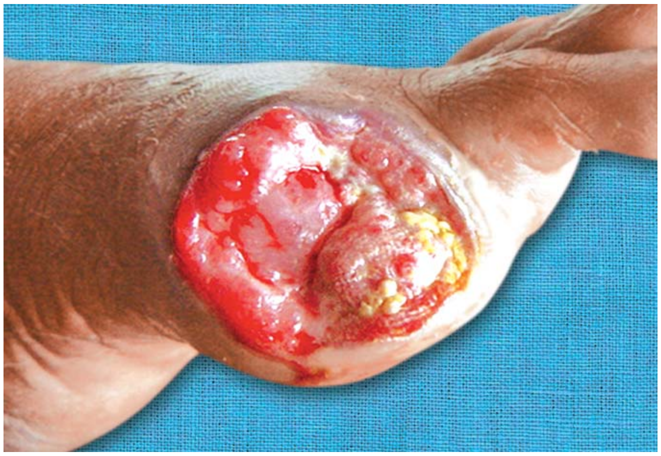

Fig. 1.652: Ulcer after great toe amputation done for vascular gangrene.

COMPLICATIONS OF AMPUTATIONS

Early

Haemorrhage, haematoma, infection.

Late

Pain, ulceration of stump, ring sequestrum formation, flap necrosis, painful scar, *Phantom limb*—feeling of amputated part in toto or partially with pain over it.

Haematoma

It is identified by pain, swelling over the stump underneath the flap. It is aspirated using a wide bore needle. Haematoma may delay healing; may precipitate infection or flap necrosis due to pressure. After aspiration, pressure dressing is needed. If haematoma reforms after 2-3 aspirations, it should be drained by opening the wound on one corner and inserting haemostat into the wound.

Infection of the Stump

It may cause abscess formation, delay in wound healing, flap necrosis, giving way of the wound. Removing few or all sutures to relieve pressure and draining the pus underneath is needed. Infection may also lead into poor scar, adherent scar which causes difficulty in placing the prosthesis.

Flap Necrosis

It is a *common* complication. Main causes for flap necrosis are poor blood supply, infection, haematoma underneath, inadequate length of the flap causing stretching of flap. Small area of necrosis can be excised. Wider area needs laying opening of the wound or revision of the stump or higher level amputation. Anaemia, poor nutrition, nutritional deficiencies, diabetes mellitus, immunosuppression, smoking, old age are other factors causing flap necrosis.

Stump Neuroma

It can occur due to proliferation of the nerve fibrils beyond the point of nerve division and is usually due to failure of cutting of the nerve more proximal to the level of division of the bone. It causes pain and tenderness over the stump. It is usually relieved by analgesics, reassurance and prosthesis. Occasionally it may require reexploration of the wound, excision of end neuroma and also cutting nerve more proximally.

Stump Pain after Amputation

It is *a common problem*. Causes are—infection, poor blood supply, causalgia, stump neuroma, phantom pain/limb, deep vein thrombosis, adherent scar, formation of spurs and osteophytes at amputated bone end. Scar adhesion to bone is prevented by keeping adequate length of deep fascia underneath intact. Spurs and osteophytes are confirmed by X-ray and needs removal using bone nibbler after appropriate skin incision.

Phantom Limb

It is *typical* awareness of sensation that as if amputated part is still present *partly or in toto*; often such part may be painful or disturbing or hyperaesthetic. Exact cause is not known; but it is probably due to presence of severe pain at the amputated part just prior to amputation making brain area for that part in alert situation causing phantom limb. Reassurance, prosthesis, analgesics help to control the condition. It is said that it can be prevented by proper pain control for 24 hours prior to amputation; but it is often difficult. It is common in upper limb.

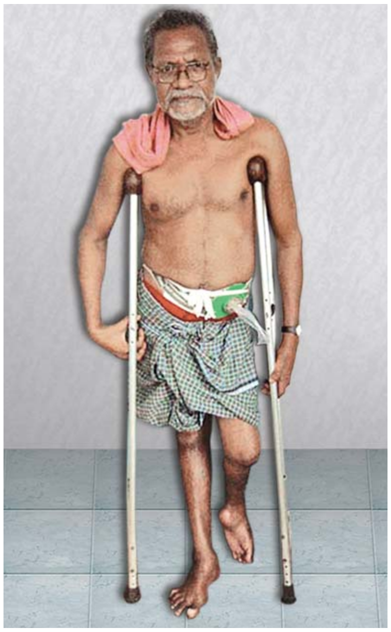

Fig. 1.653: Patient with axillary crutches.

Ulceration Over the Stump

It is *not uncommon*. It is due to necrosis, infection, lengthy bone stump pressing on the summit of the flap, prosthesis, nutritional deficiencies, diabetes mellitus, ischaemia. Ulcer may be small/large; superficial/deep. Callous chronic ulcer at the end of the stump is called as *Douglas ulcer*. Small ulcer is later treated by regular dressings and suturing. Large ulcer needs flap to cover the defect. *Osteomyelitis of the stump* should be ruled out in chronic stump ulcer. *Ring sequestrum* may be typical in such situation. Revision amputation is needed for the stump.

Contracture of the Joint

Contracture of the joint proximal to the amputated stump is *common.* It is mainly due improper positioning after amputation due to pain, poor exercise and occasionally due to inflammation of surrounding soft tissues. Contracture interferes with proper fitting and functioning of the prosthesis and delays rehabilitation. Proper positioning, passive stretching and exercises, strengthening exercises with help to correct it. Rarely needs surgical release of the contracture.

Other Complications

- Scar hypertrophy, skin thickening, hyperkeratosis, papilloma, eczema,lymphoedema, boils, bursae over bony point can occur which are treated accordingly.
- Spur, osteophyte formation, causalgia, jactitation of the stump, stump aneurysm, stump fracture— are other complications.

PROSTHESIS

It is the substitution to a part of the body to achieve its optimum function.

(Orthroses are supplement for limb function.)

In Lower Limb

- *Syme's amputation*: Elephant boot, Canadian Syme's prosthesis.
- *Below-knee amputation*: Patellar tendon bearing (PTB) prosthesis and solid ankle cushion heel (SACH).
- *Above-knee amputation:* Suction type prosthesis. It is placed above the stump. It is better and well-tolerated.
- *Nonsuction type prosthesis*: It is placed at the ends. It requires additional support.
- *Hind-quarter amputation:* Tilting table prosthesis (TTP) or Canadian prosthesis is used here.
- *Patellar tendon bearing prosthesis (PTB prosthesis)*: Here all the weight bearing is done below knee; movement is controlled by his own knee joint. Patellar tendon is the main key weight bearing area within the socket; stump posteriorly up to the popliteal fossa is also important to provide counter pressure so that patellar tendon is kept in place. Medial and lateral paratibial areas and medial condylar flare also bear significant weight. Head of the fibula, tibial tubercle, cut ends of tibia and fibula are pressure intolerant areas. Socket should have proper relief areas to these intolerant parts.
- *Computer-assisted designing and computer-assisted manufacturing (CAD – CAM) socket* is an automated processing method which fulfills all above criteria accurately with modifications. It is more comfortable and is made up of thermoplastic or laminated plastic with closed cell polyethylene foam.
- Suspension for below knee amputation prosthesis is leather cuff strap above femoral condyles. Exo or endoskeleton is used. For athletics endo skeleton is preferable.
- *SACH (Solid Ankle Cushion Heel) foot* is used. It is multiaxial, optimizes gait, and facilitates walking on rough ground. It needs minimal maintenance. It is preferred in old people. Energy storing foot is often used wherein ankle joint is replaced by a plastic spring.

In Upper Limb

1. Above elbow prosthesis is a high technology prosthesis. It is sophisticated device with harness, socket, elbow joint unit, control cable, forearm and wrist device.
2. Below elbow prosthesis.
   Krukenberg's amputation does not require any prosthesis.

Advantages of Prosthesis

- Cosmetic.
- Function of the part relatively can be got.
- Ambulation in lower limb prosthesis.

Disadvantages

- Infection.
- Pressure ulcers.
- Joint disability.

Prosthesis Types

- Exoskeletal prosthesis.
- Endoskeletal prosthesis with modular system.

Internal prostheses are one used inside. They are placed by open surgery. They are nonreactive, long durable materials, e.g. hip prosthesis in total hip replacement.

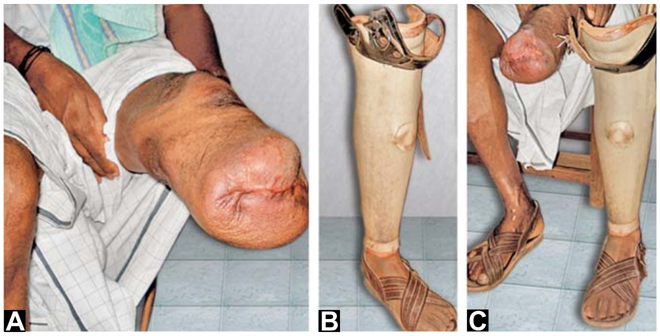

Figs 1.654A to C: Below-knee amputee with prosthesis. It gives good range of movements, normal walk and comfort.

U. Reconstruction

CHAPTER OUTLINE

- Graft
- Skin Grafts
- Flaps
- Tendon
- Tendon Repair
- Tendon Transfer
- Tendon Graft

GRAFT

Graft: It is transfer of tissue from one area to other without its blood supply or nerve supply.

Autograft: It is tissue transferred from one location to another on the same patient.

Isograft: It is tissue transfer between two genetically identical individuals, i.e. between two identical twins.

Allograft: It is tissue transfer between two genetically different members, e.g. kidney transplantation (Human to human) (Homograft).

Xenograft: It is tissue transfer from a donor of one species to a recipient of another species (Heterograft).

SKIN GRAFTS

> *We bring back, refashion, and restore to wholeness the features which nature gave but chance destroyed, not that they may charm the eye but that they may be an advantage to the living soul, not as a mean artifice but as an alleviation of illness, not as becomes charlatans but as becomes good physicians and followers of the great Hippocrates. For although the original beauty of the face is indeed restored, yet this is only accidental, and the end for which the physician is working is that the features should fulfill their offices according to nature's decree.*
>
> —**Gaspare Tagliacozzi,** 1597

Skin Grafting: It is transfer of skin from one area (donor area) to the required defective area (recipient area). It is an autograft.

Types

1. PARTIAL THICKNESS GRAFT (Split-thickness skin graft—SSG)

Also called as *Thiersch graft*, is removal of full epidermis + part of the dermis from the donor area.

It may be

- *Thin SSG.*
- *Intermediate SSG*, all depends on the amount of thickness of dermis taken.
- *Thick SSG.*

Indications

1. Well granulated ulcer
2. Clean wound or defect which can not be apposed
3. After surgery to cover and close the defect created.
   For example:
   After wide excision in malignancy
   After mastectomy
   After wide excision in squamous cell carcinoma
   Graft can survive over periosteum or paratenon or perichondrium

Prerequisite

1. Healthy granulation area
2. β-haemolytic streptococci load less than 10^5 per gram of tissue, otherwise graft failure will occur

Contraindications

SSG can not be done over bone, tendon, cartilage, joint.

Technique

Donor area: Commonly *thigh,* occasionally arm, leg, forearm.

- Knife used is *Humby's Knife.*
- Blade is *Eschmann blade, Down's blade.*
- Using Humby's knife graft is taken, punctate bleeding is observed which says that proper graft has been obtained.

Different instruments used to harvest the skin graft

- Humby's knife
- Watson modification of Humby's knife
- Power dermatome is also used (Brown)
- Sterilised razor blade can be used with a specialised device to harvest small grafts under local anaesthesia

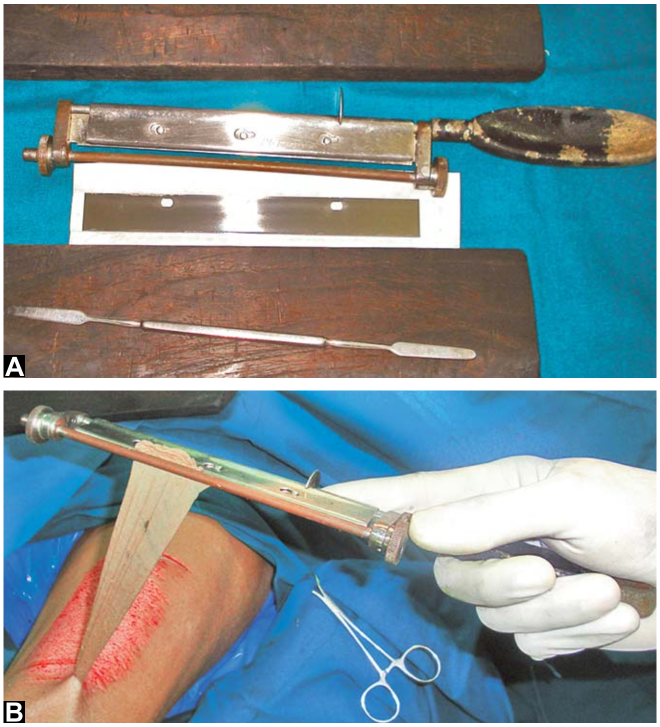

Figs 1.655A and B: Split skin grafting knife and set; and harvesting.

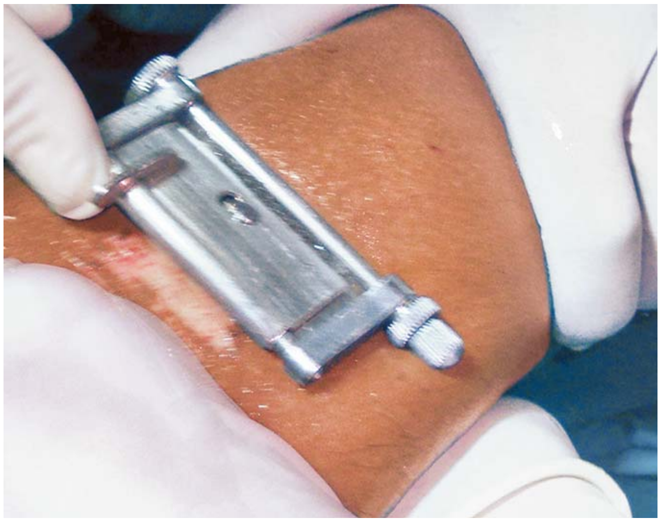

Fig. 1.656: Miniature SSG knife to take stamp sized small grafts. Sterile razor blade is used in this.

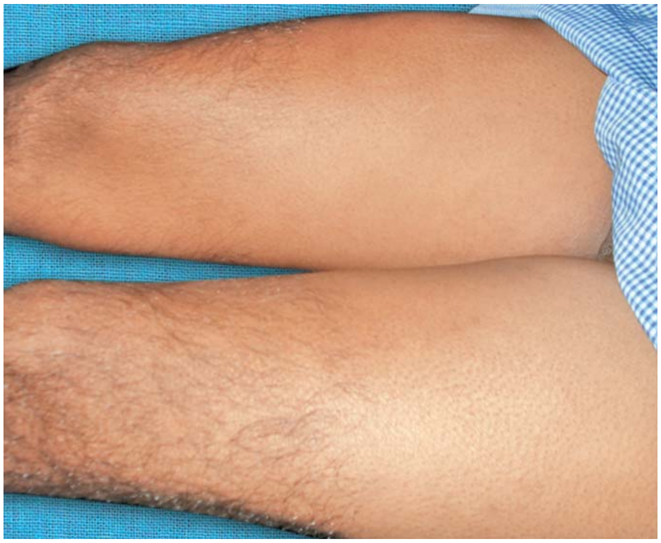

Fig. 1.657: Commonest donor area for spilt skin grafting is thigh.

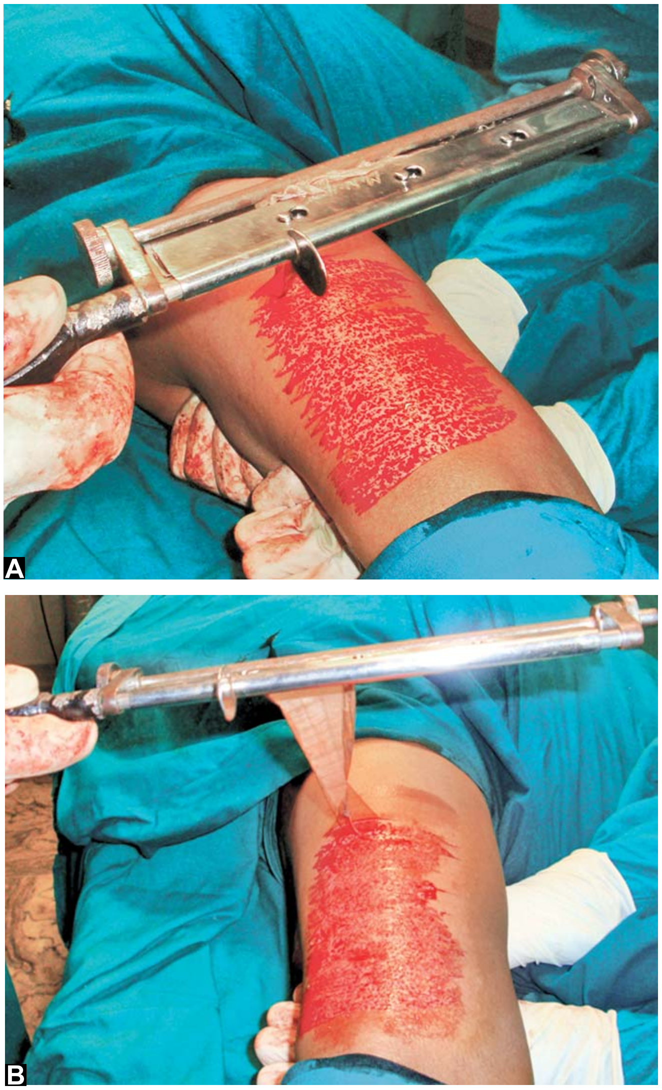

Figs 1.658A and B

Donor area is dressed and dressing is opened *after 10 days*, not earlier.

Recipient area is scraped well and the graft is placed after making window cuts in the graft to prevent the development of seroma. Graft is fixed and *tie-over* dressing is placed. If graft is placed near the joint, then the part is immobilised to prevent friction which may separate the graft. On *5th day,* dressing is opened and observed for graft take up. *Mercuro chrome* is applied over the recipient margin to promote epithelialisation.

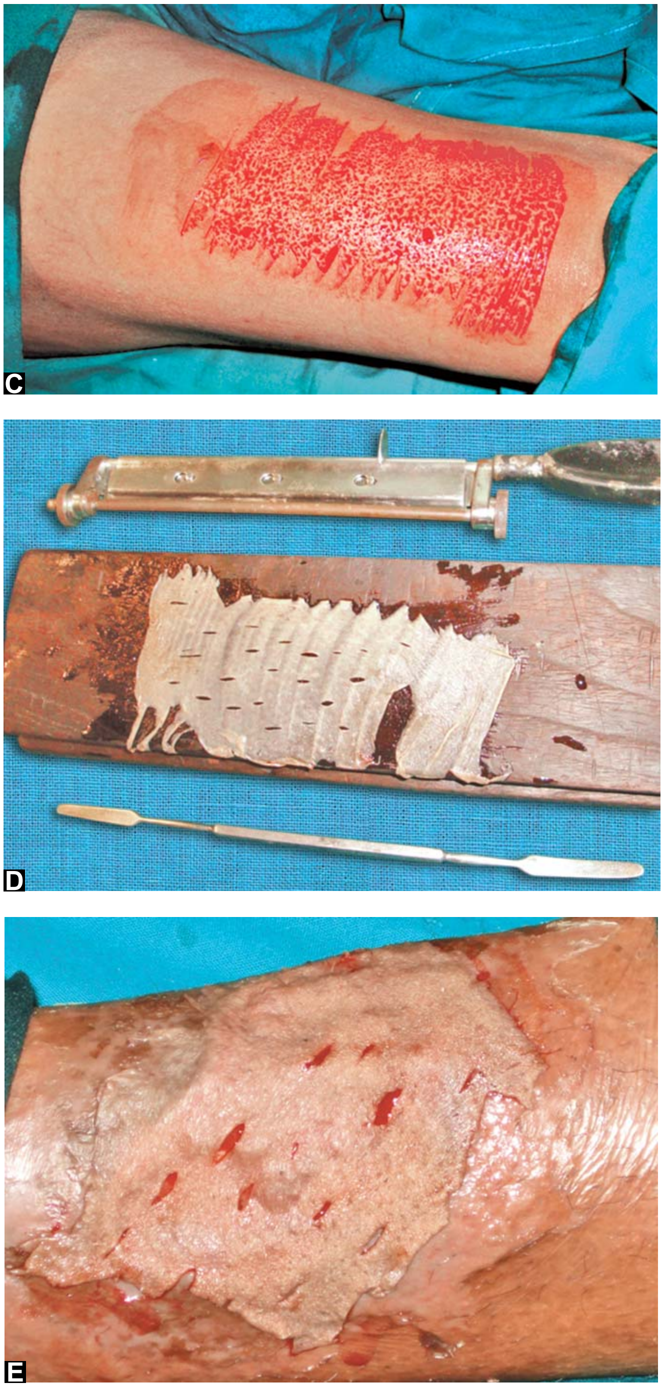

Figs 1.658C to E

Figs 1.658A to E: Technique of split skin grafting.

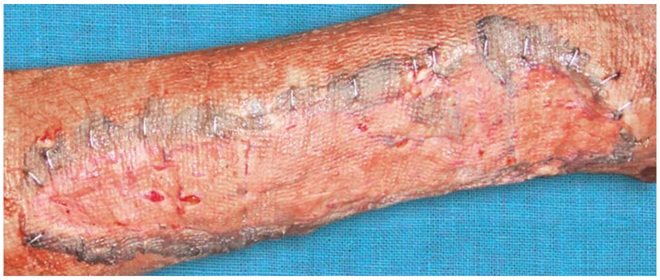

Fig. 1.659: Skin stapler can be used to fix the SSG to the margin of the recipient bed.

Stages of Graft Intake

1. *Stage of plasmatic imbibition*: Thin, uniform, layer of plasma forms between recipient bed and graft.
2. *Stage of inosculation*: Linking of host and graft which is temporary.
3. *Stage of neovascularisation*: New capillaries proliferate into graft from the recipient bed which attains circulation later.

***Note*:**

Graft is stored at low temperature of 4°C for not more than 21 days.

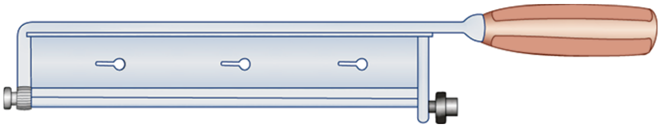

Fig. 1.660: Humby's knife with Eschmann blade.

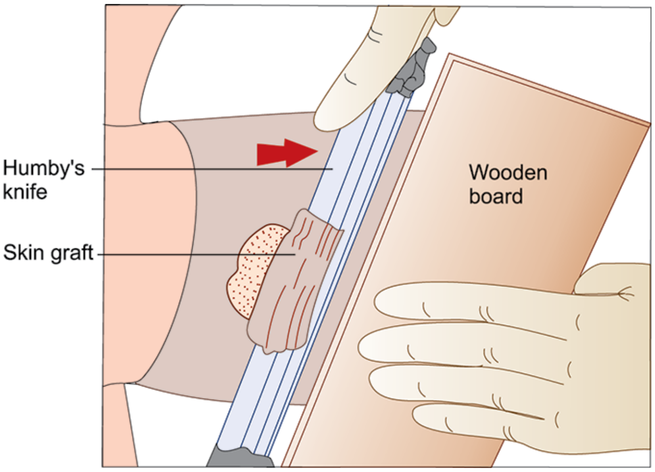

Fig. 1.661: Harvesting a skin graft.

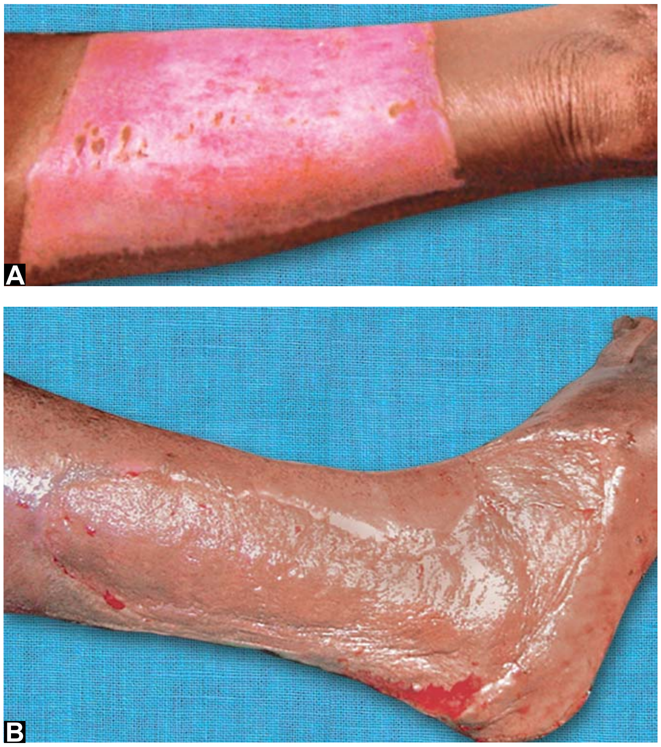

Figs 1.662A and B: Donor area of split skin graft in the thigh and graft placed over raw area in the leg.

Disadvantages of SSG

1. Contracture of graft. Two types:
   A. *Primary contracture* means SSG contracts significantly once graft is taken from donor area (20-30%). Thicker the graft more the primary contracture.
   B. *Secondary contracture* occurs after graft has taken upto recipient bed during healing period, due to fibrosis. Thinner the graft more the secondary contracture.
2. Seroma and haematoma formation will prevent graft take up.
3. Infection.
4. Loss of hair growth, blunting of sensation.
5. Dry, scaling of skin due to nonfunctioning of sebaceous glands. So after healing, oil (coconut oil) should be applied over the area.
6. Graft failure.

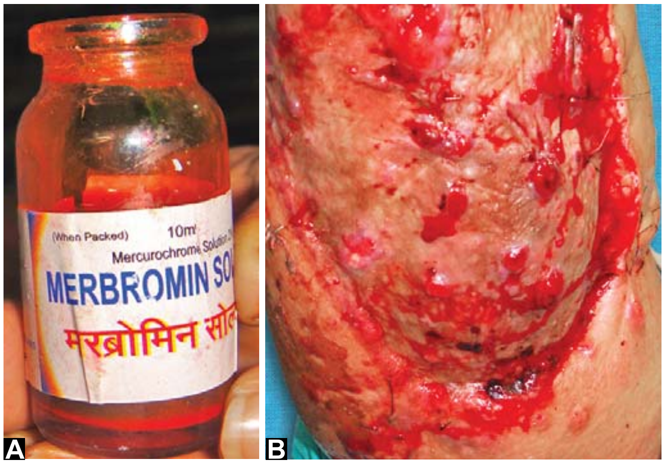

Figs 1.663A and B: Mercurochrome is often used to apply on the SSG recipient bed margin to promote epithelialisation. It is applied once a day.

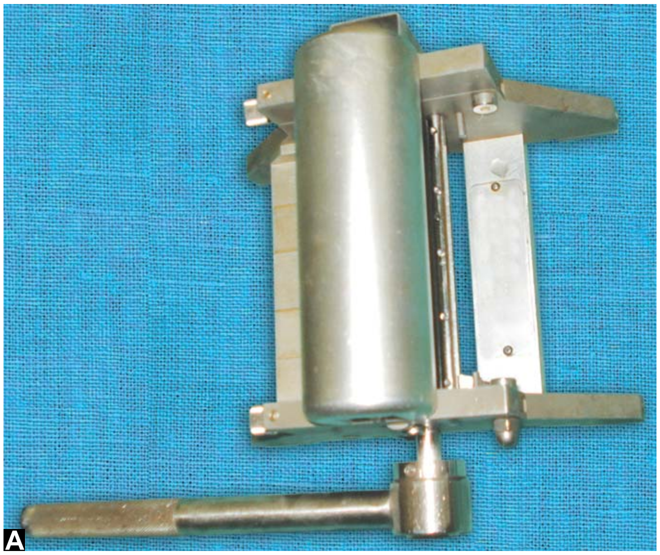

Fig. 1.664A:

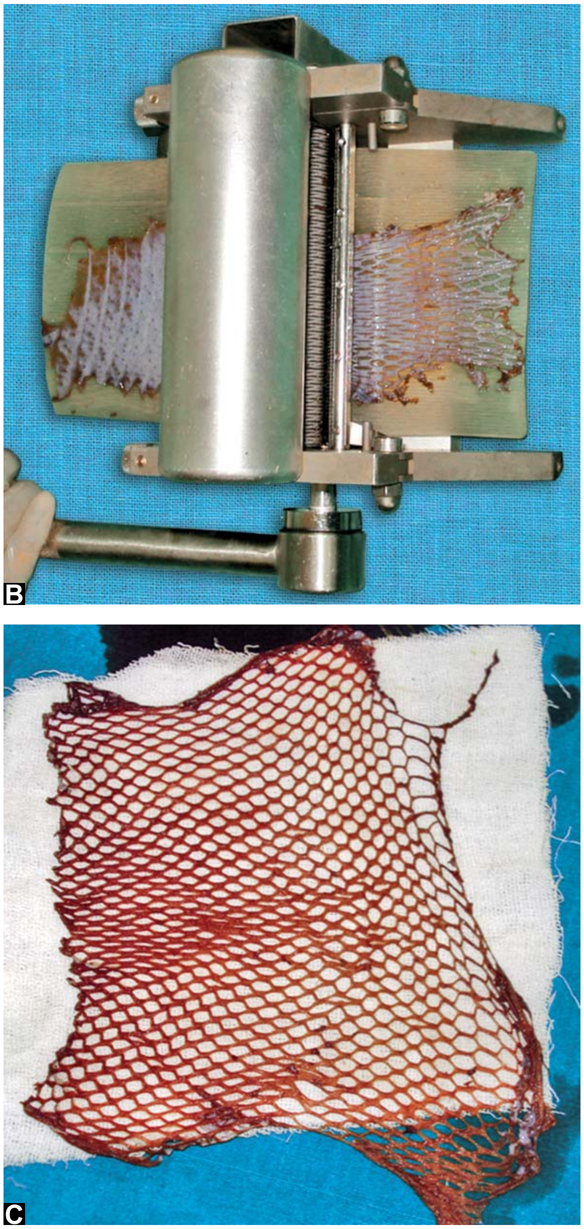

Figs 1.664B and C

Figs 1.664A to C: Mesher used in split skin grafting to increase its surface area to cover wider area like burns wound. A large defect can be covered by this. It can cause expansion of skin upto six times.

Advantages

1. Technically easier.
2. Wide area of recipient can be covered. To cover large area like burns wound, graft size is increased by passing the graft through a ***Mesher*** which gives multiple openings to the graft, which can be stretched on the wider area like a net. It can cause expansion upto 6 times.
3. Graft take up is better.
4. Donor area heals on its own.

Goodness is the only investment that never fails.

5. Mercurochrome/merbromin is used as a local applicant to the edge of the grafted area (SSG) and small raw areas to promote epithelialisation. It is applied once a day.

2. FULL THICKNESS GRAFT (Wolfe Graft)

- It includes both epidermis + full dermis.
- It is used over the face, eyelid, hands, fingers and over the joints.
- It is removed using scalpel blade. Underlying fat should be cleared off properly. Deeper raw donor area is closed by primary suturing. If large area of graft is taken, then that donor area has to be covered with SSG which is a disadvantage in full thickness graft.

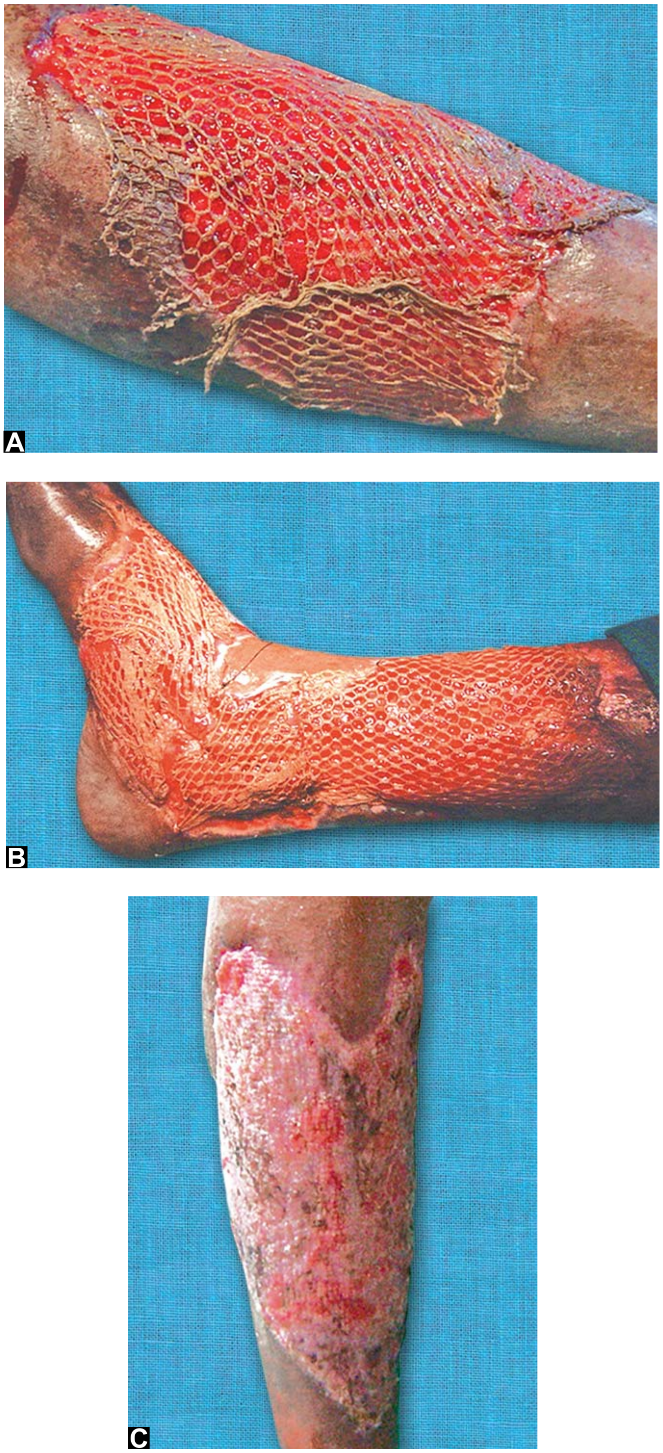

Figs 1.665A to C: Mesh graft placed over donor area and photo of it, once it takes up.

Common sites of donor area

1. Post-auricular area
2. Supraclavicular area
3. Groin crease area

Advantages

1. Colour match is good. Especially for face.
2. No contracture (unlike in SSG).
3. Sensation, functions of sebaceous glands, hair follicles are retained better compared to SSG.
4. Functional and cosmetic results are better.

Disadvantages

1. It can be used only for small areas.
2. Wider donor area has to be covered with SSG to close the defect.
3. Can not be used to cover ulcers.

Other Grafts

1. Composite graft which includes skin + fat + other tissues like cartilage.
2. Tendon graft.
3. Bone graft.
4. Nerve graft.
5. Venous graft.
6. Corneal graft.
7. Combined graft (allograft + autograft).

FLAPS

It is transfer of donor tissue with its blood supply to the recipient area.

Parts of Flaps

Base, pedicle, tip of flap. Vasculature is usually through the pedicle in the centre of the flap. Tip is the place where often flap goes for necrosis.

Indications

- To cover the wider, deeper defects
- To cover over bone, tendon, cartilage
- If skin graft repeatedly fails

Types

1. *Random pattern flaps*: Here vascular basis is subdermal plexus of blood vessels. No known blood vessel is supplying it. Rectangular flap with length to width ratios 1 : 1 or less than 1.5 : 1.

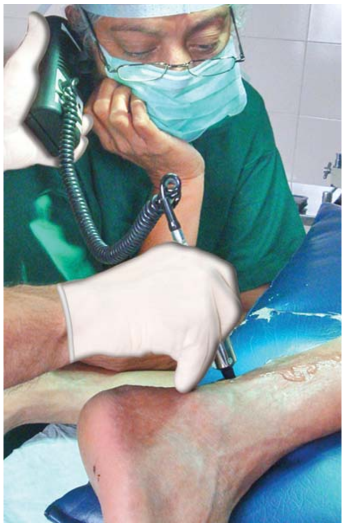

Fig. 1.666: Hand held Doppler to hear audible signal of perforator to do rotation flap in the leg.

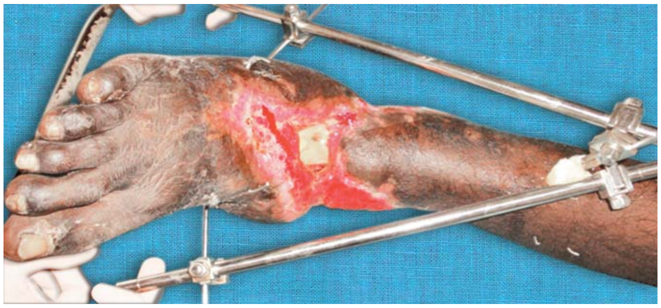

Fig. 1.667: Traumatic exposure of the bone. Flap is needed to cover this defect. It could be cross leg flap or rotation flap. Skin grafting is not possible in this situation.

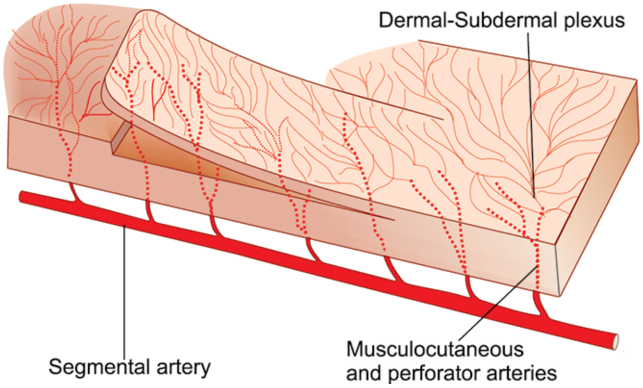
Dermal-Subdermal plexus
Segmental artery
Musculocutaneous and perforator arteries

Fig. 1.668: Anatomy and blood supply of a skin flap.

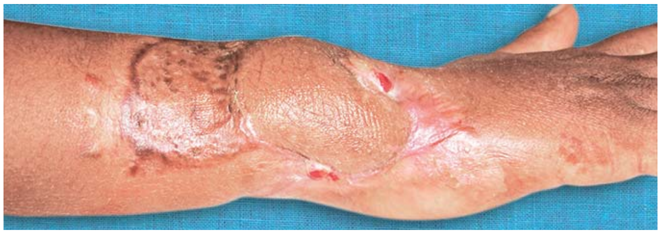

Fig. 1.669: Flap which has taken up well-placed over the defect on the bone with osteomyelitis.

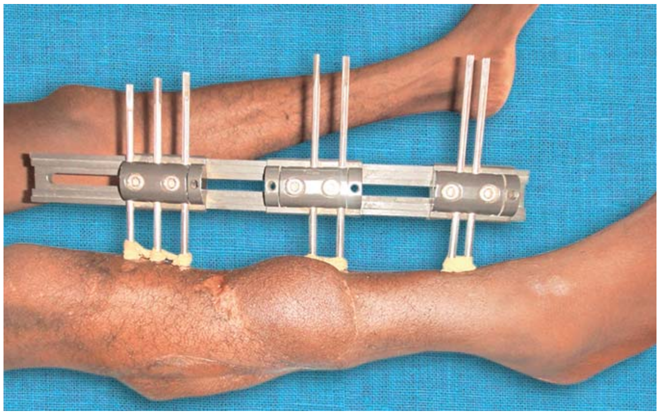

Fig. 1.670: Different methods of immobilisation is needed for flap and SSG. Fixation is often used in limbs.

2. *Axial pattern flaps*: Here superficial vascular pedicles pass along their long axes, e.g. forehead flap, deltopectoral flap, groin flap. Anatomically a known blood vessel is supplying it. It is long lengthy flap.

Anatomical types depending on the types of tissue in the flap:

1. *Cutaneous flap*: Forehead flap, deltopectoral flap.
2. *Fasciocutaneous flap*: Radial forearm flap, scapular flap, lateral arm flap.

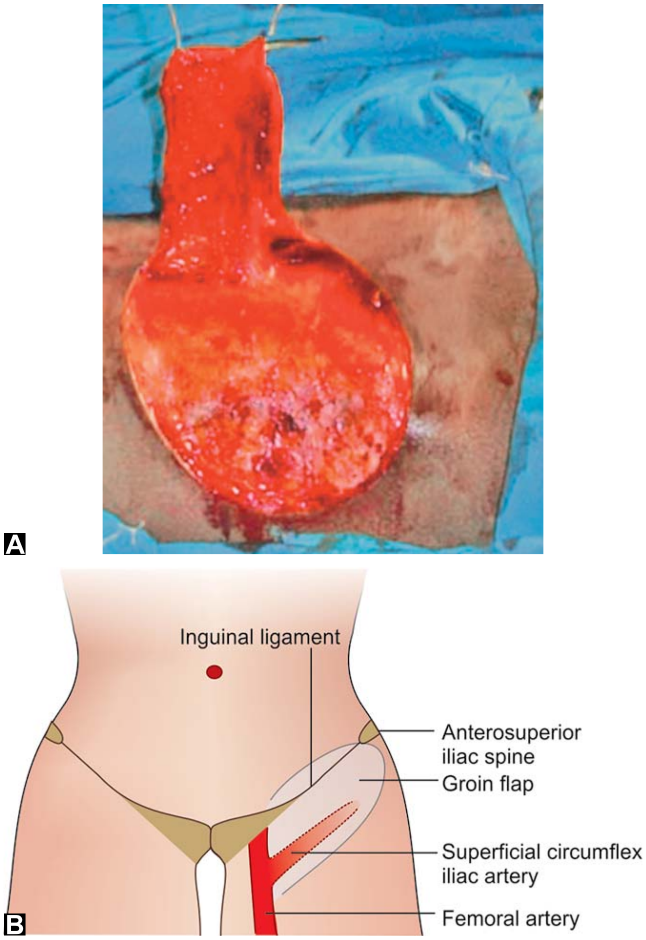
A
Inguinal ligament
Anterosuperior iliac spine
Groin flap
Superficial circumflex iliac artery
Femoral artery
B

Figs 1.671A and B: Groin flap is based on superficial circumflex iliac artery. It can be used in defects in hand and forearm. It is a cutaneous flap.

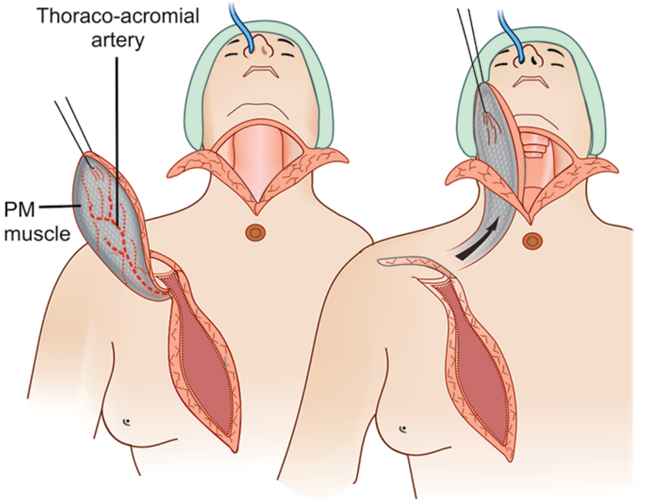
Thoraco-acromial artery
PM muscle

Fig. 1.672: Pectoralis major myocutaneous flap.

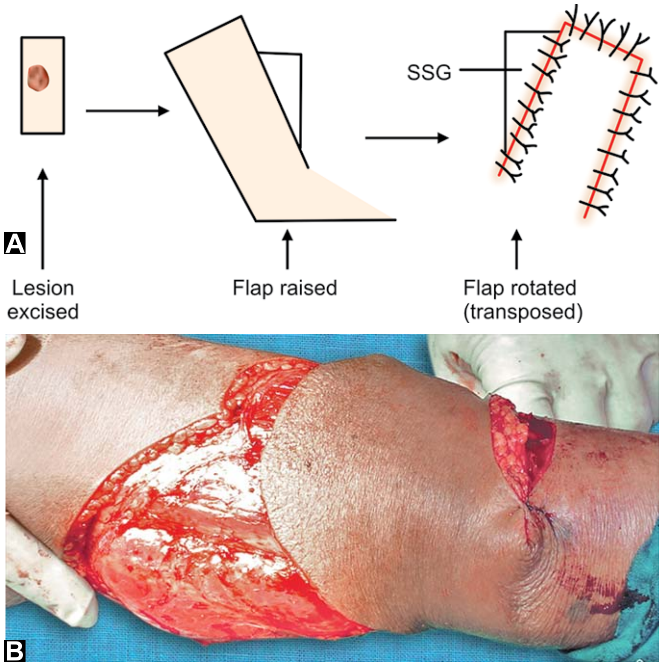
A
SSG
Lesion excised
Flap raised
Flap rotated (transposed)
B

Figs 1.673A and B: Transposition flap.

3. *Muscle flap*: Gluteus maximus muscle flap, gracilis flap, tensor fascia lata muscle flap.
4. *Myocutaneous flap*: Pectoralis major myocutaneous flap, latissimus dorsi flap - composite flap.
5. *Osteomyocutaneous flaps*: Radius with brachioradialis and skin, rib with intercostal muscles and skin—composite flap.
6. *Local rotation flaps*, *transposition flaps*: When the flap moves laterally it is called as transposition flap. When the flap rotates laterally towards defect it is called as rotation flap. Transposition flap is squarely designed which moves laterally to close the defect creating a larger area on its original place which has to be covered with split skin graft.
7. ‘*Z’ plasty*: It is a procedure which involves transposition of two inter-digitating triangular flaps. There is change in

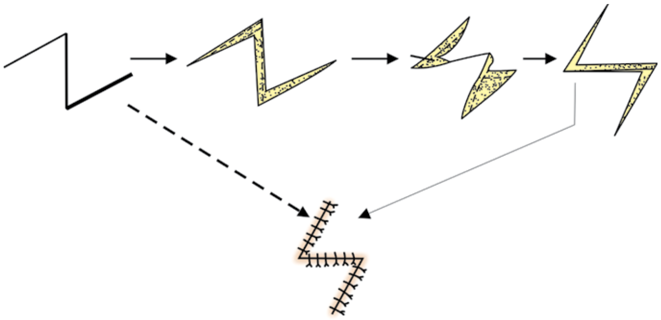

Fig. 1.674: Technique of ‘Z’ plasty. It is used in contracture release. Dupuytren’s contracture and pilonidal sinus.

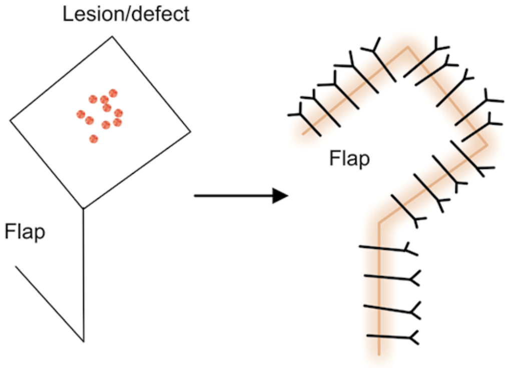
Lesion/defect
Flap
Flap

Fig. 1.675: Rhomboid flap

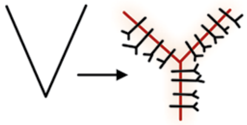

V-Y plasty

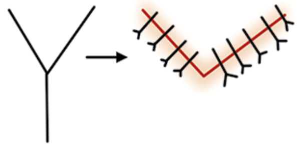

Y-V plasty

Fig. 1.676: Technique of V-Y plasty and Y-V plasty.

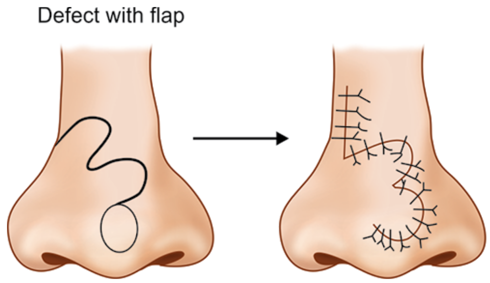
Defect with flap

Fig. 1.677: Bilobed flap is used in lesions of the nose commonly.

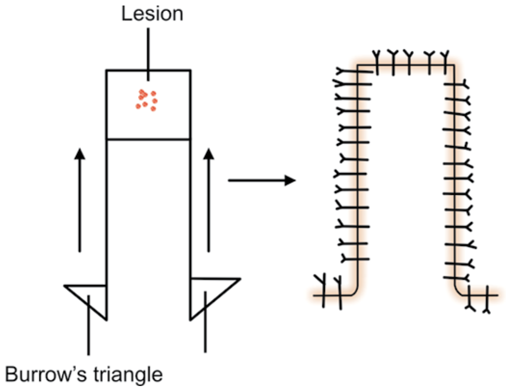

Fig. 1.678: Advancement flaps, used in distal defects.

direction as well as gain in length of the common limb of Z. Angle size and length of the limb are the most important factors. It is used in managing contracted scars, facial scars, Dupuytren's contracture and to cover the excised defects like pilonidal sinus (example). There should be transverse skin slack available equal to the length between the axes of Z. It can be single or multiple Z plasty. Complications are flap necrosis near the angle tip, infection, failure.

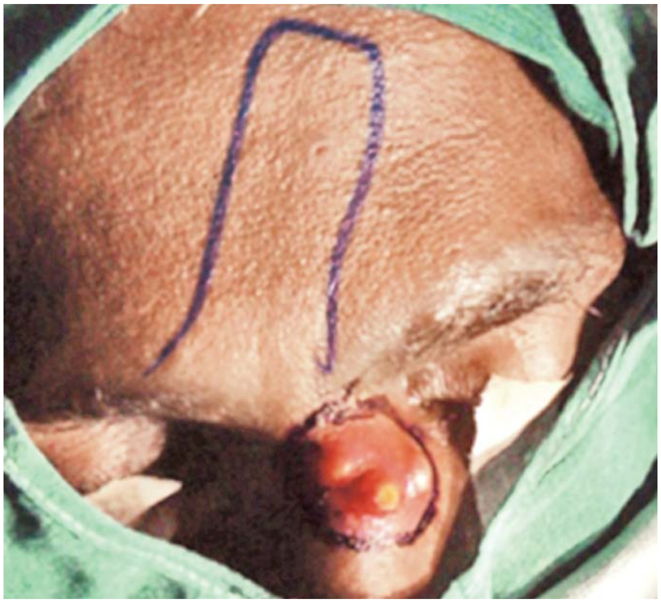

Fig. 1.679: On table marking of glabellar flap to place after excising a lesion in the root of the nose.

8. *Free flaps*: Vascular pedicle of the flap, both artery and vein are anastomosed to recipient vessels using operating binocular microscopes.
9. Omental flaps.
10. *Island flap*: Localised flap is swung around a stalk from the donor area to the recipient area often with the pedicle buried underneath the skin bridge in between. Pedicled flap is also an island flap.

Areas where flaps are commonly used: Oral cavity, neck, breast, limbs (leg), buttock, bedsores.

Flaps mobilised from donor area with its pedicle is placed and sutured to recipient area. Once flap takes up usually in 3-6 weeks, base of the flap is cut and sutured to recipient area.

Saltatory flap is mobilising the flaps in stages from distant donor area towards recipient area. It requires many staged surgeries and long term hospitalisation.

Waltzing is a technique wherein flap is moved from donor area and attached adjacent to the recipient defect area. Later in 2nd stage, it is moved towards the defect formally. It reduces the tension on the flap and increases the success rate.

Advantages of Flaps

1. Good blood supply, good take up.
2. Gives bulk, texture, colour to the area.
3. Allows required movements in the recipient area. For example, jaw movements after pectoralis major flap after wide excision with hemimandibulectomy for carcinoma cheek.
4. Cosmetically better.

Disadvantages of Flaps

1. Long-term hospitalisation.
2. Infection.
3. Kinking, rotation and flap necrosis.
4. Staged procedure.

Positioning of the patient for long time is important to have a good flap take up which is a real discomfort to the patient.

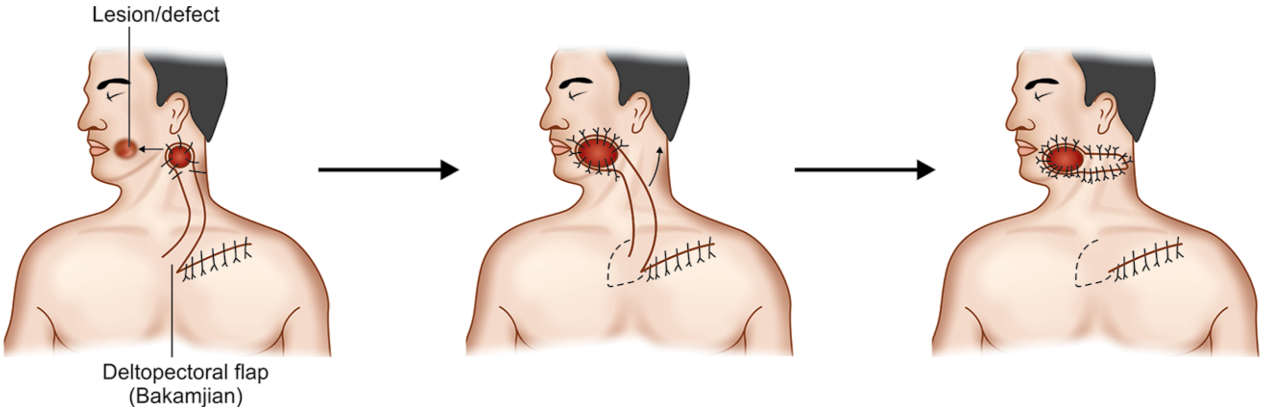

Fig. 1.680: Waltzing of flap.

Note:

Delaying of the flap after mobilisation is done to reorient its blood supply so as to decrease flap necrosis and increase flap survival. Delaying period is 10-14 days. Flap is raised properly as required and resutured to same place so as to allow the formation of reorientation of vasculature near the tip of the flap. This delayed flap is raised again in 10-14 days to place in required area.

Different Flaps Used are:

Forehead Flap

It is fasciocutaneous flap from forehead based on *anterior branch of superficial temporal artery.* Superficial temporal artery is terminal smaller branch of external carotid artery. It begins under the parotid behind the neck of mandible, runs vertically upwards, crossing the root of zygoma at preauricular point; 5 cm above the zygoma it divides into anterior and posterior branches. Anterior branch anastomoses with supraorbital and supratrochlear branches of ophthalmic artery. Artery supplies scalp of temple region and side, parotid, ear, facial muscles. Superficial temporal artery gives transverse facial artery which runs from anterior margin of parotid, and middle temporal artery which runs deep to temporalis muscle.

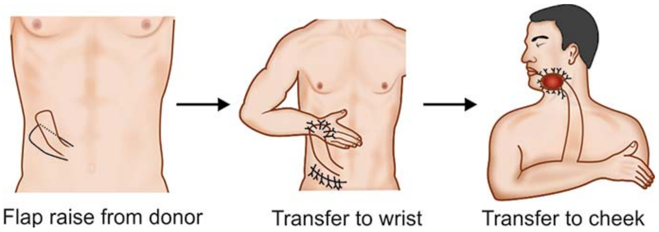

Fig. 1.681: *Saltatory flap*. Flap raised from the abdomen is reached to the cheek in stages—from abdomen to wrist; from wrist to cheek.

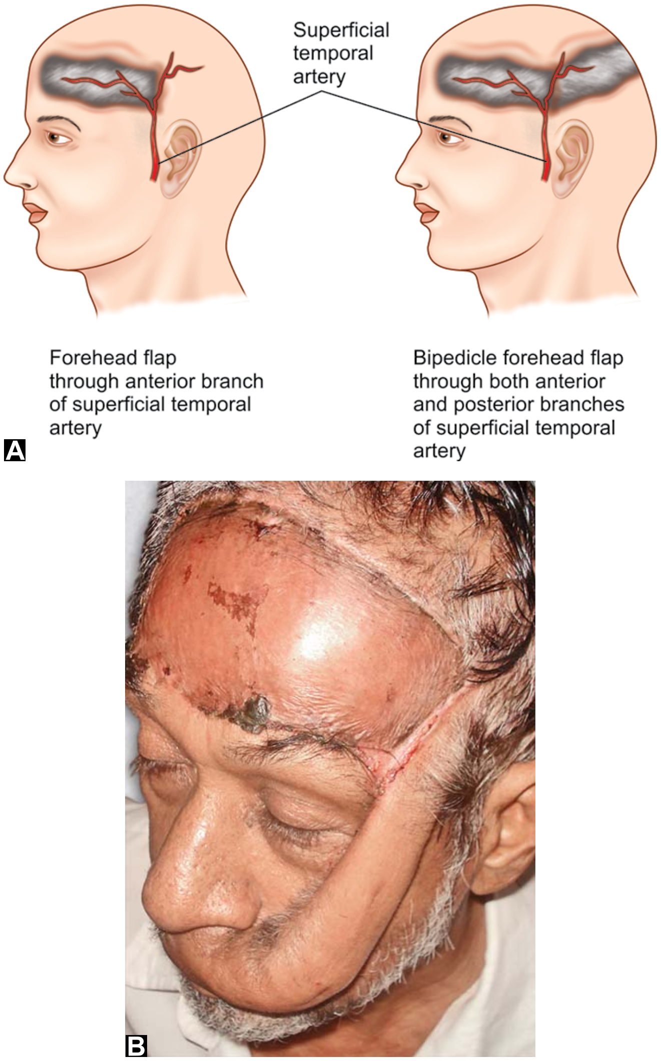
Superficial temporal artery
Forehead flap through anterior branch of superficial temporal artery
A
Bipedicle forehead flap through both anterior and posterior branches of superficial temporal artery
B

Figs 1.682A and B: Forehead flap based on superficial temporal artery. It is fasciocutaneous flap.

It is used for defects in cheek (carcinoma, cancrum oris) and nasal reconstruction.

Standard forehead flap is taken from forehead above the level of eyebrow starting from the opposite side of the midline with base just above the zygoma. Width of the flap is usually about 3-4 cm. Flap is dissected from the distal end of the marked area going deep up to epicranium, raising the flaps using scissor dissection. Flap is held up using skin hooks. Flap is rotated towards the defect area in the cheek. Inner area of the flap is covered with split skin graft. Donor flap area is covered with another split skin graft. After 3 weeks, base of the flap is disconnected; remaining proximal part of the flap can be replaced into forehead donor area. SSG over donor flap area takes up well. This flap often can be rotated under (deep to) the zygoma also.

A lined forehead flap can be used. After flap elevation, under surface of the flap is lined by split skin graft prior to rotation. This grafted lined flap is resutured into the donor area for 2 weeks until undersurface of graft takes up well; after 2 weeks flap is rotated towards the defect area (cheek).

Delaying of the flap is often done in forehead flap. Flap after elevation, is replaced into the original position to have optimum vascular reorientation; after 2 weeks it is again rotated towards defect. Delayed flap reduces the flap necrosis chances.

Often bipedicled forehead flap (*Narayanan's flap*) is used taking both from anterior and posterior branches of the superficial temporal artery. This flap is moved to defect in cheek with anterior branch part staying outside and posterior part will line the mucosal area.

A different type—*middle forehead flap* is used for nasal reconstruction.

Problems—Poor color match, contraction of flap and donor area is cosmetically nonacceptable.

Deltopectoral flap (Bakamjian flap)

It is based on first three perforating branches of the internal mammary artery (mainly 2nd perforator). Flap runs horizontally across the chest wall anteriorly towards shoulder tip from its base over the sternal border. Its upper border is along the line of the clavicle; its lower border is along the line of anterior axillary fold line. Raw area often requires a spilt skin grafting. It is usually rotated upwards often with waltzing. It is tubed and attached above. Tube is drained to prevent any collection to occur. Rotation angle is important to prevent any kinking in the pedicle. It is usually used to cover the defects in cheek, chin, mastoid and parotid region. Often flap is delayed to get adequate length.

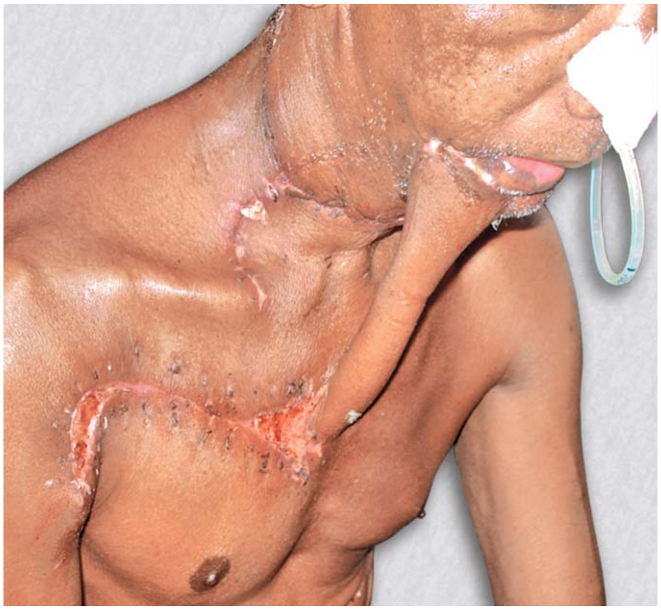

Fig. 1.683: Deltopectoral cutaneous flap.

Groin flap (Fig. 1.684)

It is based on superficial circumflex iliac artery which is 2-3 cm below and parallel to the inguinal ligament. Artery originates from femoral artery over medial border of the sartorius and ends at anterior superior iliac spine. 1:1 rectangular flap with deep fascia is used. Secondary defect can usually be closed with sutures. It is used mainly for defects in wrist/forearm where positioning is easier.

Latissimus dorsi muscle/myocutaneous flap

It is based on thoracodorsal artery, a branch of subscapular artery. Skin over the upper and anterior border of Latissimus dorsi is used for transfer. It is commonly used to cover the defect after mastectomy. But it does not give the bulk. It is technically easier. It can be used as muscle flap also. It helps as skin cover. Prosthesis is needed to place underneath to provide bulk in post-mastectomy defect.

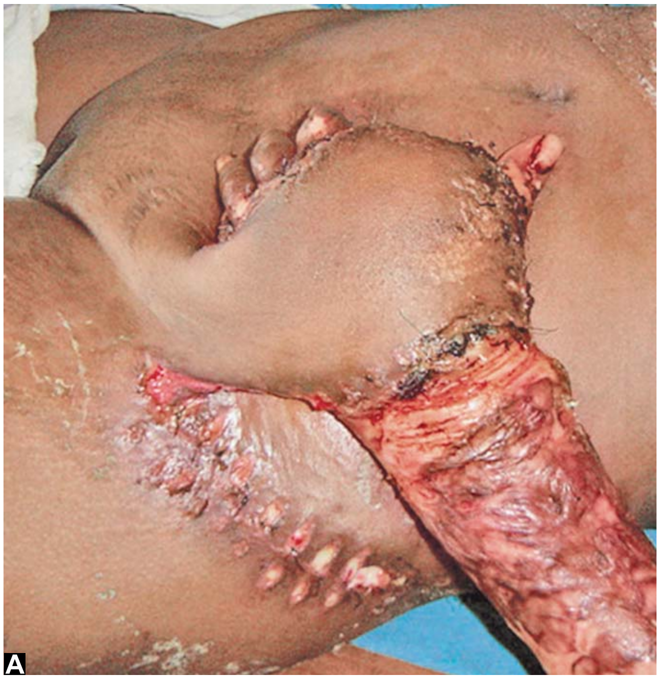

Fig. 1.684A

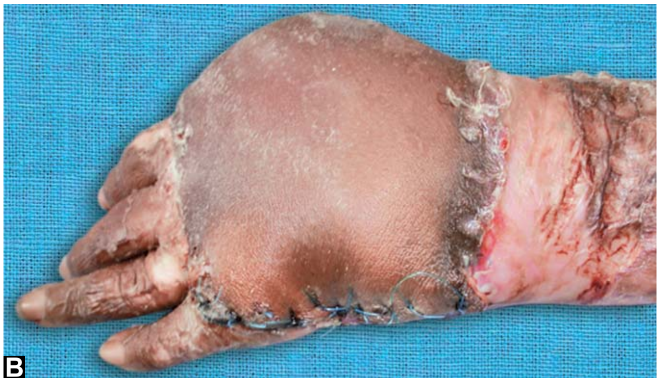

Fig. 1.684B

Fig. 1.684A and B: Groin flap used for burns defect in the hand.

Pectoralis major myocutaneous flap

It is based on the pectoral branches of thoracoacromial artery. Usually skin below and medial to nipple over the muscle is used. Muscle pedicle is made as broad as skin. It used to cover the defect over the cheek/neck/pharynx/intraoral lesions after wide excision with removal of skin over the tumour. Vessel marking is 2 cm medial to coracoid process, obliquely below the clavicle at the junction between middle third and outer third. Skin with muscle is dissected from the deeper structures like ribs, intercostal muscles and pectoralis minor. Flap is raised upwards up to the coracoid. Lateral pectoral vessels if possible are retained, otherwise can be sacrificed. Defect

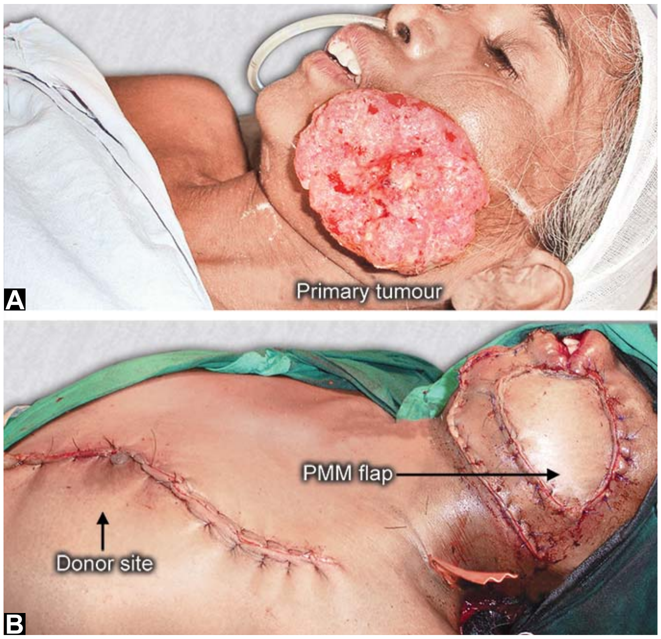

Figs 1.685A and B: Pectoralis major myocutaneous flap used for carcinoma cheek.

below is usually closed primarily with sutures. Often it needs spilt skin grafting. Pectoralis major flap can be used along with deltopectoral flap with proper planning.

Gastrocnemius muscle flap

It is either medial or lateral and is commonly used to cover the upper part of the tibia and knee joint. It is technically easier and functional deficit occurring at donor area is insignificant. It is rarely used as myocutaneous flap.

Transverse rectus abdominis muscle flap (TRAM flap)

It is either superior pedicle based on the superior epigastric vessels or inferior pedicle based on the inferior epigastric vessels. Superior pedicle based flap is used to cover postmastectomy area or chest wall defect. To cover post mastectomy area opposite side superior pedicle is used to reduce the arc through which flap has to rotate (but rotating from opposite side is technically difficult). Inferior pedicle flap is used to cover the defects in groin and thigh. Proper marking of the flap is essential. Skin incision is made like an ellipse. Anterior rectus sheath is cut in the line of incision and is raised upwards carefully of the rectus muscle up to the xiphisternum. Muscle is gently separated of the posterior rectus sheath with care not to injure the epigastric vessels. Once dissection is complete lower part of the muscle is cut in superior pedicle to rotate upwards carefully. In the upper part again anterior rectus sheath is opened to pass the flap towards the defect in subcutaneous plane. TRAM flap gives bulk and contour to the defect. But it is technically difficult. Usually opposite side of the defect is taken as flap as it is easier to rotate from opposite side. Defect in the abdomen usually needs mesh to support and close. It is not possible to do this flap in obese individuals and if patient has undergone laparotomy earlier (with a lengthy scar). Inferior cut end of the inferior epigastric artery in superior pedicle flap can be anastomosed to a vessel in recipient bed to improve the perfusion *(supercharging)*. Inferior epigastric artery in inferior pedicle can be additionally perfused using opposite inferior epigastric artery (*recharging*).

Radial forearm flap

It is perfused from the radial vessels and raised on the flexor aspect of the forearm. Perforating branches of these vessels supply deep fascia and skin over it. Flap can be fasciocutaneous or osteofasciocutaneous/osteomyofasciocutaneous if radial bone is also used as part of the flap. Radial forearm free flap is commonly used for mandible defects. It is technically easier and safer. Flap is raised along with skin, segment of the radius along its intermuscular septum through which vessels pass and brachioradialis as components. Care is taken in dissecting vessels of the bed and not to injure the radial nerve. In free flap artery is sutured to the recipient artery like facial artery using microscope. Other similar flaps are—ulnar forearm flap, scapular flaps, and vascularised fibular transfer.

Limberg flap

It is a type of rhomboid flap used in pilonidal sinus with base at gluteal skin.

Cross leg flap

It is commonly used to cover the defect in the foot/leg from opposite leg.

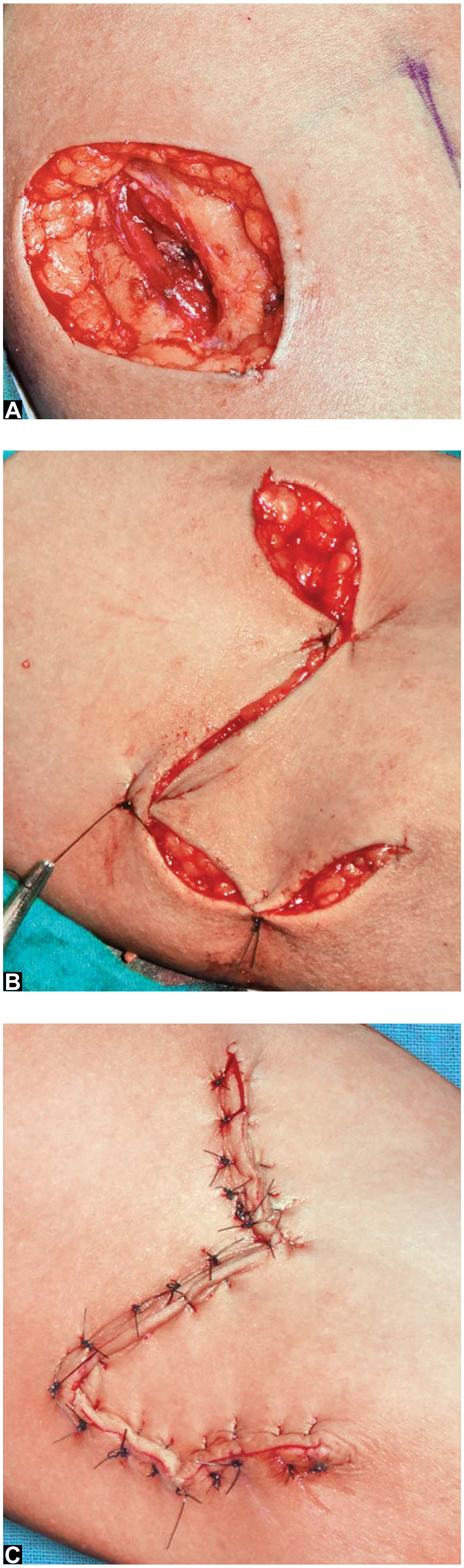

Figs 1.686A to C: Rhomboid flap used to cover the defect after excising a lesion in the arm.

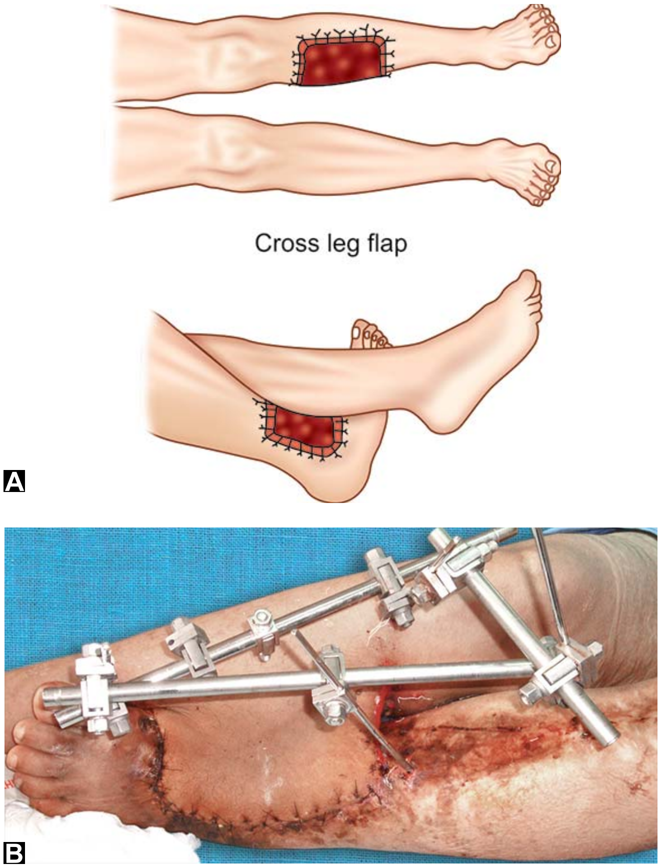

Figs 1.687A and B: Cross leg flap.

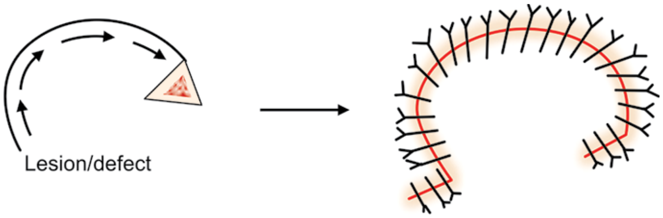

Fig. 1.688: Rotation flap. It is commonly used in gluteal region in managing bedsores.

TENDON

- Tendon is the continuity of the muscle to have its action at the site especially in hand, foot and digits.
- It is covered by synovial sheath with a thin layer of fluid in between which allows smooth gliding of the tendon.

Tendon after injury heals by:

1. **Intrinsic healing method** occurs through synovial fluid when tendon is not under stress.
2. **Extrinsic healing method** occurs through proliferation of fibroblasts across epitenon. It occurs when tendon is under stress. It forms a mass of fibrous tissue at the site called as "*tenoma*". It may interfere with the proper gliding of the tendon.

Stages of tendon healing

- By 21 days, weak healing occurs and contraction of muscle is possible
- By 6 weeks, mild traction can be applied to the tendon
- By 3 months, moderate stress can be used
- By 8 months, full tensile strength is recovered

Tendon injuries may be cut wound, lacerations, injury associated with nerve or vessel injury.

TENDON REPAIR

1. ***Primary repair*** is done within 24 hours.
2. ***Delayed repair*** within a week after 24 hours.
3. ***Secondary repair*** anytime after one week.

Types of Suturing the Tendon

1. ***Kessler method***. Here knot comes in the cut part of the tendon.
2. ***Goldner method***. Here knot comes away from the cut ends of the tendon.

Suture material used is monofilament nonabsorbable suture material (polypropylene, 3 or 4 zero).

- Continuous sutures are used for suturing.
- Epitenon can be later apposed with interrupted sutures.
- Postoperative immobilisation is advised for 3-4 weeks.
- Later passive and active exercise is done with the help of a physiotherapist.

Complications of tendon suturing are infection, adhesion, stiffness and failure.

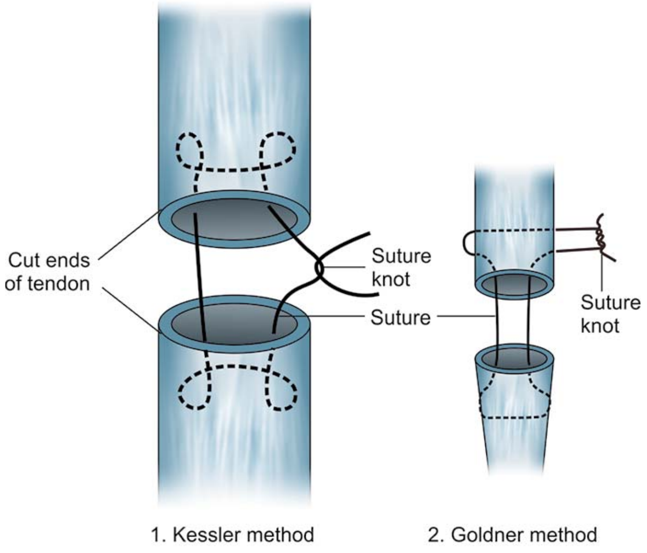

Fig. 1.689: Types of tendon suturing

Every job is self portrait of the person who did it.

TENDON TRANSFER

- It is the transfer of one tendon from its existing site to another site where its function is required to have a function required at the newer site.
- Function of the transferred tendon should be maintained by other tendons.
- Tendon should be able to acquire the function at the newer site properly.
- For example, in ulnar claw hand, the tendon of flexor digitorum superficialis (FDS) (of index finger) is transferred to lumbrical canal of the digits to have flexion at MCP joint and extension of proximal IP joint.

TENDON GRAFT

- When tendon suturing or transfer is not possible because of inadequate length, tendon of a muscle which is not of much help functionally, is taken as a graft to obtain required length.
  Common grafts used are:
  - Palmaris tendon in forearm.
  - Plantaris tendon in leg.
  - For example, extensor carpi radialis brevis (ECRB) is lengthened using tendon graft to transfer to lumbrical canals in claw hand.

Problems in tendon graft: Infection, adhesions, graft failure, stiffness of the part.

V. Transplantation

CHAPTER OUTLINE

- Preoperative Evaluation
- Organ Procurement
- Renal Transplantation
- Immunosuppressive Agents
- Liver Transplantation
- Bone Marrow Transplantation
- Pancreatic Transplantation
- Small Bowel Transplantation
- Dialysis
- Cimino Fistula

PREOPERATIVE EVALUATION

- General evaluation: Pulmonary, cardiac, GIT, renal status and cancer screening.
- Immunologic evaluation: Serology for hepatitis, HIV, cytomegalovirus.
- Placing the organ in the same position is called as orthotopic transplantation, e.g. liver.
- Placing the organ in new position is called as heterotopic transplantation.

Donor Criteria

a. *Cadaver donor*
  - Individuals with severe brain injury resulting in brain death.
    Brain death is defined as "complete irreversible cessation of all brain functions".

Criteria for brain death

- Absence of spontaneous respiration
- Absence of cranial reflexes
- Absence of response to stimuli
- Irreversible causes
- Absence of cerebral blood flow
- Isoelectric EEG
- Sustained apnoea with elevated CO_2

Other criteria for cadaver donor:

- Normothermic patient.
- No respiratory effort by the patient.
- Brain—dead donor.
- The heart is still beating.
- No depressant drugs intake should be there while evaluating the patient.
- Individual should not have any sepsis, cancer (except brain tumour).
- Not a HIV or hepatitis individual.

b. *Living donor*
  - Living unrelated donor.
  - Living related donor.
  - Individual should have normal health.

Requirements

- ABO typing.
- Serology tests.
- Angiogram.
- Intravenous urography.
- HLA typing.

Human Leucocyte Antigens (HLA)

They are cell surface molecules which are highly antigenic. They play main role in graft rejection and hence called as *major histocompatibility complex/antigens* MHC) (Dausset, 1958). *Class I* are located in nucleated cells. *Class II* are located in dendritic cells/macrophages/B cells, etc.

Major histocompatibility complex (MHC)

- MHC is located in chromosome number 6
- Individual will have two HLA (*Human Leucocyte Antigens*) genes inherited from each parents, which together called as HLA types
- Class I— HLA is HLA-A, HLA-B and HLA-C. Intracellular antigens
- Class II—HLA is HLA-DP, HLA-DQ and HLA- DR. Extracellular antigens
- HLA-DR is most important in kidney transplantation

Principles of donor organ retrieval

- Adequate exposure is needed
- Control of the vessels above and below the organs to be removed is done
- Initiation of preservation *in situ*
- Removal of the organs, and separation.
- Completion of preservation
- Removal of iliac vessels for vascular reconstruction of pancreas and liver grafts
- Organ packaging

The winner always has a programme; the loser always has an excuse.

ORGAN PROCUREMENT

Principles of Organ Procurement

- Once brain death has been confirmed in cadaver donor, after giving inotropic support drugs (T3 and argipressin) various organs are surgically removed carefully with preservation of their vessels. After removal, organs are flushed with chilled preservative solution (in specific organ procurement like of kidney, *in situ* perfusion of organs (kidneys) is done by placing double balloon catheter into the abdominal aorta and a balloon catheter into the femoral vein) and placed in sterile bags containing saline and organ preservative solutions which are then immersed in 0-4°C box containing ice. Donor specimens are transported to the site of the recipient centre. *Wisconsin and Euro-Collins solutions* are commonly used.
- *Non-heart beating donors (NHBD)*—here organs are procured from individuals who are just dead on arrival to the hospital or who have died in the hospital in spite of resuscitation. *Category I*—Dead on arrival; *Category II*—Unsuccessful resuscitation; *Category III*—'Awaiting cardiac arrest' after support withdrawal; *Category IV*—Cardiac arrest with brain death.
- *University of Wisconsin solution* contains—potassium lactobionate; sodium phosphate; magnesium sulphate; adenosine; allopurinol; glutathione; raffi-nose; hydroxyethyl starch; insulin; dexamethasone; potassium; sodium; with 320 mosmol/l osmolality and pH of 7.4.
- *Living donor's organs* are used commonly in kidney transplantation from genetically related individuals. It can be used from genetically unrelated donors after proper MHC match. Donor nephrectomy is done through loin incision (commonly from left side). Laparoscopic donor nephrectomy has become popular and safe.

Technique of Organ Procurement

Positioning the donor on the operating table: Supine position, with arms abducted on boards and legs laid flat and uncrossed. The neck is extended by placing a sandbag under the shoulders (as during thyroidectomy).

Midline sternotomy incision is used extending up to the pubic symphysis with a supraumbilical horizontal part. Abdomen and thorax are exposed properly. Retroperitoneal right sided mobilisation is done (*Cattel-Braasch manoeuvre*). Cannulation of inferior mesenteric vein is done. Aortic cannulation is done. Perfusion of preservative solution is done.

Sequence of the thoracic organ procurement: First the heart (care must be taken to leave enough supradiaphragmatic IVC for both organs such as the liver and the heart); then the lungs separately or together are procured. Cooling of the abdominal organ has to be continued until the last thoracic organ is procured.

Sequence of abdominal organ procurement: The small bowel is the most sensitive organ for ischemia; therefore, it is retrieved first. The second organ to be procured is the pancreas followed by the liver. Liver and the pancreas could also be retrieved *en block* and split on the back table. Finally, the kidneys are the last organs to be procured.

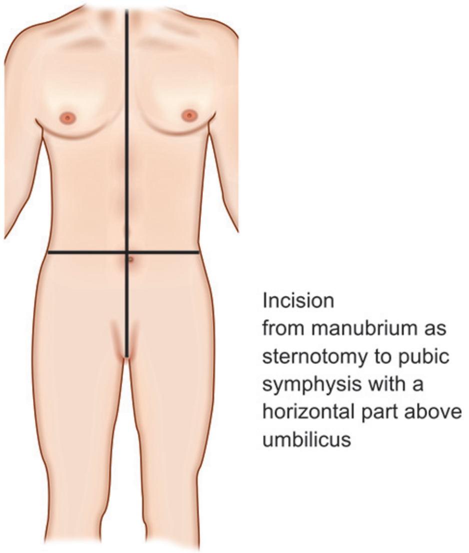

Fig. 1.690: Incision for organ procurement.

The most popular tool kit comprises the iliac vessels, which consists of common, external and internal artery and the vein.

Organ Packing in Steps (Eurotransplant or National Transplant Organisation)

The first bag is filled with a cold preservative solution. The procured organ must be completely covered by the preservative solution, and the bag must be closed (well tied) without any air. The second bag or a wax-impregnated fibre container is filled with cooled saline or Ringer lactate solution. The first tied bag must be completely covered in one of the solutions and closed (well tied) without air. Third bag dry, without air, well tied and sometimes covered with a sterile drape is also used. Finally the organ is placed in an icebox and well covered with nonsterile melting ice, the box is firmly closed to make the organ ready for transportation.

RENAL TRANSPLANTATION

> *An animal which has undergone a double nephrectomy and grafting of both kidneys from another animal can secrete almost normal urine with his new organs, and live in good health for at least a few weeks.*

The Criteria for an Ideal Deceased Kidney Donor

- Normal renal function.
- Without hypertension requiring treatment; without diabetes mellitus.
- No malignancy other than a primary brain tumor or treated superficial skin cancer.
- No generalised viral or bacterial infection.
- Acceptable urinalysis; age between 6 and 50 years.
- Negative assays for syphilis, hepatitis, HIV, and human T-lymphoproliferative virus.

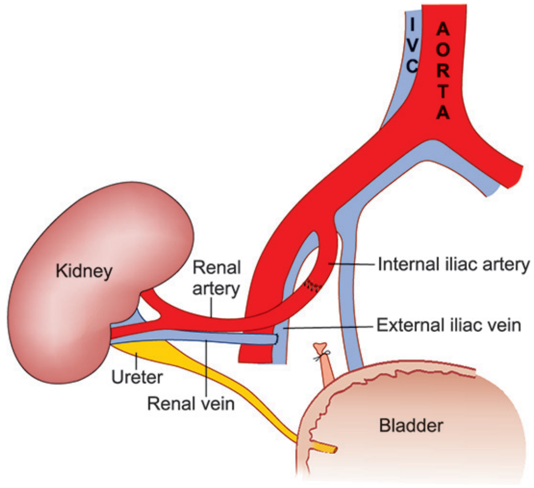

Fig. 1.691: Renal transplantation.

Three types of donors:

1. Living related donors.
2. Living nonrelated donors.
3. Cadaver donors.

Highest chances of success in any transplant (renal) is seen when the donor is the identical twin.

Compatibility should be checked by tissue typing, i.e. ABO blood group system and major histocompatibility complex.

Evaluation of living donor:

- Tissue typing, ABO, MHC typing.
- Renal function—blood urea, serum creatinine.
- IVP.
- Selective renal angiogram. Single vessel is better.
- HIV/hepatitis evaluation.

Usually *left kidney* is taken for *transplantation* because of *long left renal vein.* It is placed in right iliac fossa with ureter connected to the urinary bladder; renal artery to internal iliac artery (end-to-end); renal vein to external iliac vein (end-to-side). Renal artery (end) to external iliac artery (side) through Carrel stitch is also used for arterial continuity.

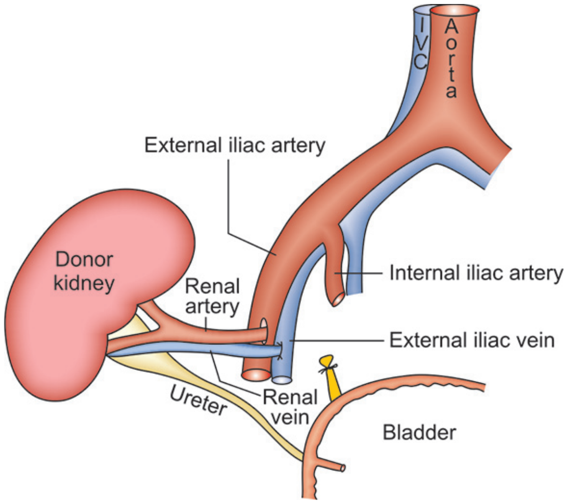

Fig. 1.692: Another method of getting arterial perfusion of the transplanted kidney by suturing (*Carrel stitch*) external iliac artery to renal artery side-to-end.

Technique

Before the removal of kidney, the donor receives IV mannitol to prevent kidney ischaemia, diuretics and IV heparin. After removal of donor kidney, protamine sulphate is given to the donor. Removed donor kidney is perfused with cold perfusion fluid at 4°C and cold intravascular electrolytes. Solutions used to preserve the donor kidney are *Euro-Collins* solution or *University of Wisconsin (UW)* (both contains inert sugar which prevents swelling of cells). First renal vein, then renal artery and at the end ureter is anastomosed.

Azathioprine, cyclosporine and prednisolone should be started 3 days prior to surgery. Diuretics and mannitol should be continued as required.

Bilateral nephrectomy in recipient is required only in:

- Polycystic kidney disease.
- Haematuria.
- Severe hypertension.

Kidney will stand cold ischaemia for 72 hours.

Note: Overall survival is 90% in one year and 80% in five years.

Postoperative Management

- Immunosuppression: By cyclosporine, azathioprine, prednisolone, antithymocytic globulin and antilymphocytic serum.
- Proper fluid balance has to be taken care of.

Complications

- Acute tubular necrosis.
- Rejection (Rejection is identified by radioisotope study and percutaneous kidney biopsy). Chronic rejection is the common cause of graft failure.
- Obstruction of the collecting system.
- Infection.
- Urine leakage—5%.
- Secondary haemorrhage.
- Renal infarction.
- Hazards of immunosuppression:
  - Infection by unusual organisms like cytomegalovirus, herpes, *Pneumocystis carinii*, varicella and other bacterial infections, candidial infection.
  - Changes in the cellular component of blood.
  - Uncommon malignancies of CNS, skin.
  - Nephrotoxicity, GIT bleeding and perforation.
  - Hirsutism, delayed wound healing, cataract formation.
  - Renal artery stenosis—10%.
  - Renal vein thrombosis—5%.
  - Lymphocele.

IMMUNOSUPPRESSIVE AGENTS

Induction Therapy

- It is used *immediately* after transplantation (up to 3 weeks).
- *Antilymphocytic globulin (ALG)*—human lymphocytes are injected into rabbit/horse to develop antisera—ALG. It acts

If there were no obstacles, there would have been no achievements.

against T cells. It inhibits cell mediated immunity, allograft rejection, graft versus host reaction. It is used in transplantation of kidney, pancreas, heart, small bowel. Anaphylaxis, anaemia, thrombocytopenia, allergy (*serum sickness*) are the reactions. Rabbit serum is better than horse serum in terms of preventing the graft rejection (acute). Rabbit serum has antibodies against CD2,3,48,11a,18,25; HLADR, HLA class I.

- *Monoclonal antibody*—OKT3 (*1975 by Kohler, Milstein – got Nobel prize*) is developed by hybridoma technique, is a monoclonal antibody which acts / blocks at TCR complex (T cell receptor complex CD3) affecting the function of native T cell and cell mediated immunity. Within 60 minutes of IV injection of OKT3 it blocks T cell function removing the circulating T cells. Problem with OKT3 is its immune reactions. So often it is combined with steroid or indomethacin. It can cause acute *cytokine release syndrome*.
- *Interleukin 2 receptor (IL2 R) inhibitors*—α chain of IL 2 R is related to activated T cells. Basiliximab (chimeric) and daclizumab (human) are two anti-CD25 monoclonal antibodies which bind with α chain of IL 2R. These drugs are well tolerated through a peripheral IV line; stops acute rejection; with less risk of infection and malignancy. It does not cause serum sickness or cytokine release syndrome. But all alone its effects are inadequate and so it should be used only in concomitant with other immunosuppressants.
- *Rituximab* is an anti-CD20 monoclonal antibody as a depleting agent on B cell. It is used to control humoral mediated rejections. It is used in heart transplantation.
- *Alemtuzumab*—an anti-CD52 human monoclonal antibody which is expressed in B cell, T cell, monocyte, macrophage. It causes prolonged lymphocyte depletion for 6 months. It is used along with tacrolimus in kidney transplantation. Steroid can be spared while using his. This drug is also used in lymphoma, multiple sclerosis and rheumatoid arthritis.
- *IVIG* (intravenous immunoglobulin) derived from pooled plasma neutralises the circulating autoantibodies, blocks the T cell cytokines, lymphocyte proliferation and apoptosis.

Maintenance Therapy

All transplant patients require maintenance therapy for prolonged period. Different drugs are used. Steroid and azathioprine are old drugs commonly used.

- ***Prednisolone:*** It inhibits cytokines, binding of IL 2 to receptors, blocks macrophage migration, and inhibits delayed hypersensitivity reaction. But it causes Cushing's syndrome, hypertension, peptic ulcer, cataract, diabetes, osteoporosis, muscle wasting. Often methyprednisolone may be used.
- **Antiproliferating agents:** They inhibit differentiation and division of lymphocytes by blocking the purine, pyrimidine and folic acid metabolism as a structural analogue.
  - *Azathioprine* is a purine analogue containing 6 mercaptopurine with a labile sulfhydryl chain. Conversion of inosine nucleotide to adenosine and guanosine is blocked. Drug inhibits both humoral and cell mediated immunity. Toxic effects are—bone marrow suppression, hepatotoxicity.
  - *Mycophenolate mofetil* (MMF): It inhibits inosine monophosphate dehydrogenase of purine metabolism blocking lymphocyte proliferation. Dose is 2 grams/day. It is now more commonly used than azathioprine. Diarrhoea and bone marrow suppression are the side effects.
  - *Leflunomide*: It blocks pyrimidine synthesis in lymphocytes by inhibiting dihydroorotate enzyme. This drug is used in rheumatoid arthritis. FK 778 is its analogue which is under trial in renal transplant.
- **T cell directed immunosuppressants:**
  - *Cyclosporine* (Borel, 1972) is extracted from fungus *Tolypocladium inflatum*. It is selective inhibitor of TCR mediated activation suppressing T cells. It inhibits formation of mature CD4 and CD8 T cells in thymus. It contains 11 amino acids with molecular weight 1202. It is a very good immunosuppressant. But it *does not* cause myelosuppression. It is metabolised in the liver by cytochrome P-450. *Side/toxic effects* are—*nephrotoxicity*, hypertension, hirsuitism, gingival hyperplasia, hyperkalaemia, neurotoxicity, tremor, hepatotoxicity, risk if infection (CMV, *candida*, *Pneumocystis carinii*, secondary infections) and malignancy (lymphoma, skin and CNS). IV dose given initially is 4 mg/kg in 500 ml of saline; later changed to oral therapy as 12 mg/kg daily; after few weeks dose is tapered to 5 mg/kg/day. Regular serum cyclosporine estimation twice weekly with estimation of haematocrit, blood urea and serum creatinine is needed.
  - *Tacrolimus* (FK 506, 1984, Japan) is derived from fungus *Streptomyces tsukubaensis*. It is macrocyclic lactone (MW 822) causes similar effect like cyclosporine. Drug binds with intracellular binding protein immunophilin forming complexes which block the phosphatase activity of calcineurium so that inhibits IL 2 production. It is used in liver transplant, in acute rejection of kidney. Side effects are similar as cyclosporine but will not cause hirsuitism and gingival hypertrophy. It causes alopecia and diabetes.
  - *Sirolimus* (rapamycin) is a macrolide antibiotic which inhibits signal transduction from IL2 R to nucleus. It inhibits allograft rejection. It is used along with other immunosuppressants like cyclosporine. It causes anaemia, thrombocytopenia and proteinuria. Nephrotoxicity is less with sirolimus.
- **Costimulation blockage agent** like *belatacept* which is a fusion protein of extracellular part of CTLA4. It s a good immunosuppressant used monthly or bimonthly.

Risks Related to Immunosuppression

They are mainly infection and malignancy. *Individual drug related toxicity* also can occur.

Infection

- Opportunistic infections are common with immunosuppression. Cytomegalovirus (CMV) infection is common causing pneumonia, hepatitis, pancreatitis, GI problems. *Pneumocystis carinii* infection is also common.

- Trimethoprim, sulfamethoxazole, acyclocvir, ganciclovir, valganciclovir, clotrimazole, pneumococcal vaccine, hepatitis B vaccine, pentamidine nebulizer are different drugs used to prevent sepsis.
- BK virus associated nephropathy is controlled by cidofovir, IV immunoglobulin.

Malignancy

- Malignancy potentiality increases by 10 times.
- Skin cancers and carcinoma cervix are common.
- Virus mediated tumours like carcinoma cervix (HPV); hepatoma (Hepatitis B and C); Kaposi's sarcoma (herpes virus 8); lymphoma (EBV) are common. EBV related lymphoma is post-transplant lymphoproliferative disorder (PTLDs). Rituximab, anti-CD20 monoclonal antibody reduces B cells is effective in these patients. Hyper CMS IG is used as prophylaxis in high-risk groups.

LIVER TRANSPLANTATION

> *An ideal treatment for several kinds of liver disease would be removal of the diseased organ and orthotopic replacement with a hepatic homograft.*
>
> —**TE Starzl, TL Marchioro, KN von Kaulla,** et al, 1963

Indications

- Primary biliary atresia, metabolic liver disease.
- Cirrhosis.
- Malignant disease of the liver.

Children respond better for liver transplantation. Tissue typing and cross-matching are not that necessary and do not influence the results.

If the transplantation is done at the same site after doing hepatectomy, it is called as ***orthotopic liver transplantation***. If it is placed in a different site it is called as ***ectopic*** or ***heterotopic liver transplantation***. Success rate in liver transplantation is better.

Liver transplantation is the choice for end stage liver disease. Liver transplant is an accepted and effective treatment. Many tumours can be treated by transplantation. Cirrhosis, hepatitis, sclerosing cholangitis, biliary atresia, tumours are indications.

CTP (Child Turcote Pugh) scoring system (ascites, encephalopathy, bilirubin, albumin, PT- INR); *Model for end stage liver disease (MELD score)* consists of total bilirubin, INR, creatinine; *Paediatric end stage liver disease scoring* (PELD); are different scorings used to assess the patients.

- *MELD Score* = 0.957 x Log$_e$(creatinine mg/dL) + 0. 378 x Log$_e$(bilirubin mg/dL) + 1.120 x Log$_e$ (INR) + 0.643

Contraindications for liver transplantation

- Active sepsis, SBP and HIV.
- Extrahepatic malignancies.
- Large hepatocellular carcinoma and cholangiocarcinoma.
- Unfit for surgery - severely advanced cardiopulmonary disease.
- Active alcohol or substance abuse.
- Inability to comply with immunosuppression protocols because of psychosocial situations.

Donor criteria

They are related to donor liver function, hepatitis screening, history of consumption of alcohol and toxic substances. Marginal donor and expanded donor criteria are in use in places due to very high demand for donor for transplantation. ABO compatibility is important. HLA matching is not necessary unlike in renal transplantation.

Donor operation

It is by a midline incision from suprasternal notch to pubic symphysis. Initially dissection is done with a beating heart; but later with a cold preservation into aorta and portal vein and local ice application. UW solution extends cold ischaemic time for 24 hours but usually within 10 hours transplantation is

DDLT (deceased donor liver transplantation) donor	***LDLT (living donor liver transplantation) donor***
• 50 years or younger	• Young, Non-obese, tall
• No hepatobiliary disease	• No infection, malignancy, No fatty liver
• Hemodynamic and respiratory stability (systolic blood pressure >100 mmHg, and central venous pressure >5 cm/H$_2$O); an acceptable PaO$_2$ and haemoglobin level	• Enough liver to donate
• No severe abdominal trauma, systemic infection, or cancer	• Enough residual liver
• Normal kidneys—diuresis greater than 50 mL/h and normal creatinine;and finally, a dopamine requirement less than 10 µg/kg/min	• Suitable anatomy of liver
• Simple technically	• Technically complex procedure
• Recipient complication is low (of bile duct)	• Recipient complication is higher
• Graft quality is average	• Graft quality is good
• Donor's complication—nil	• Donor's complication—25% morbidity; 0.4% mortality

Success is never final and failure is never fatal; its courage that counts.

done. IVC segment is removed with hepatic veins. Portal vein is transected. Celiac artery with branches along with hepatic artery is dissected and transected at celiac artery origin. Bile duct is transected.

- Remnant liver—≥30% of the original liver volume with complete venous drainage is safe for donor survival.

Recipient Operation

Bowel preparation, prophylactic antibiotics, administrative immunosuppressive agents are needed before surgery. Bilateral subcostal incision is used with midline upward extension. Falciform ligament, left lateral ligaments, hepatogastric and hepatoduodenal ligaments are divided. Right and left branches of hepatic arteries are ligated. CBD and cystic ducts are divided with ligatures. Portal vein is dissected. Infra and suprahepatic vena cava are exposed. Portal vein and suprahepatic vena cava are clamped. Venovenous bypass between portal and femoral veins and internal jugular vein is created with a flow of blood through it more than 2.5 ml/minute. Warming circuits and ultrafiltration is needed. Suprahepatic vena caval cuff is created at the opening of the right, middle and left hepatic veins. Donor IVC is sutured to recipient IVC first above and then below as end-to-end. Recipient IVC can be preserved entirely and donor IVC is sutured to recipient IVC as end-to-side. Preservative solution is passed. Portal vein is sutured. Reperfusion injury can occur at this stage with hyperkalaemia and acidosis. Arterial reconstruction is done between donor celiac artery to recipient gastroduodenal artery. Bile ducts are sutured as end-to-end with or without T tube. Often choledochojejunostomy is also done.

Graft should be:

- 40% of the estimated standard liver volume.
- 0.8 to 1% of the body weight of the recipient is necessary for the recipient recovery.

Segmental/Lobar/Split/Living Donor Liver Transplantation

It is newer beginning but becoming preferred one. It is based on segmental anatomy and liver regeneration capacity. Minimum liver mass required is more than 1% of graft to body weight ratio. Right lobe is better. It is basically used in children; but useful in adults also. MELD score in donor should be less than 20 for living donor transplantation. Regeneration occurs very significantly in recipients in first 2 weeks reaching standard in 1 month. Regeneration is slow in living donor liver in 1 year. Graft to standard liver ratio should be more than 40%. Function of liver in living donor should be optimum. Proper imaging of the donor for entire anatomy including vasculature is important. Individual branches of portal vein, hepatic artery and bile duct are dissected and isolated. Liver is carefully separated from IVC; small branches of hepatic vein are ligated. Main hepatic vein is isolated; liver parenchyma is dissected with finger/Kelly fracture technique. After completion of parenchymal division, vessels are transected; dissected liver is perfused with cold preservative solution. Diseased liver of the recipient is removed completely. Graft is anastomosed in usual manner hepatic vein, portal vein, hepatic artery, bile duct. Living donor transplantation has got higher chances of morbidity, complications compared to cadaveric transplantation.

Postoperative Management

- Antirejection drugs to be given.
- Liver shows low immunogenicity and high regeneration capacity causing long-term outcome. Initially cyclosporine or tacrolimus with MMF or azathioprine with prednisolone is given. ICU care is needed.
- Electrolyte and fluid management, sepsis management, prevention of encephalopathy, observation for complications are important.

Complications and Problems of Liver Transplantation

- *Bleeding* on table and postoperatively can occur especially when recipient liver is cirrhotic. FFP, platelet and blood are needed. Reexploration is needed if bleeding persists postoperatively.
- *Hepatic artery thrombosis* is the common vascular complication leading into graft failure, CBD necrosis and anastomotic dehiscence.
- *Immediate graft failure* as primary nonfunctional graft can occur (5% of liver grafts).
- *Bile leak* due to CBD ischaemia which can be identified by HIDA scan, ERCP, revision anastomosis is needed.
- *Infection* is common especially enterococci, staphylococci, gram negative organisms, candida, aspergillus, etc. Antibiotic and antifungal therapy is needed.
- *Acute rejection* (30%) of T cell mediated is seen within 10 days commonly but can occur up to 6 months. Liver biopsy confirms the rejection. Higher steroid therapy, polyclonal anti-T cell antibodies are used.
- *Chronic rejection* is seen after 6 months; appears gradually with liver cell dysfunction and hyperbilirubinaemia. Liver biopsy shows very few numbered biliary radicles—*vanishing bile duct syndrome*. It is due to humoral immunity. It is difficult to control; eventually needs retransplantation.
- *Recurrence* of the earlier disease to the transplanted liver like hepatitis, biliary sclerosis, sclerosing cholangitis. Hepatitis can be controlled by lamivudine (antiviral DNA polymerase for HBV), interferon α, ribavirin for HCV.
- *Complications of immunosuppressive drugs* like hypertension, hyperglycaemia, hyperlipidaemia, osteoporosis, malignancy, infection, bone marrow suppression are often difficult to manage.

Survival

- 5-year survival is 70% in adults and 80% in children.
- It depends on age, general condition, earlier disease, associated problems, MELD scoring.
- The risk of rejection is highest (40%) during the first 3-6 months after transplantation and decreases significantly thereafter.

BONE MARROW TRANSPLANTATION

Indications

- Leukaemias.
- Aplastic anaemias.
- Immune deficiencies, etc.

Recipients are initially treated with total body irradiation. As bone marrow is an active immune system, *proper tissue typing* is essential. *Infant bone marrow* is better marrow as a *donor*. Marrow aspirated from donor's bone is transplanted by *intravenous injection* to the recipient.

Immunosuppression with cyclosporin-A is always needed. It will take few weeks to show the response.

Problems

1. Graft rejection.
2. *Graft-Versus-Host* disease (GVH) is more dangerous.

PANCREATIC TRANSPLANTATION

- It is used in diabetic patients, taken from cadaver donor to replace insulin.
- It can be combined with kidney transplantation to patients who have diabetes with end stage kidney disease.
- Donor criteria are individuals without pancreatitis.
- Entire pancreas with duodenum (Lillehei) or part of the pancreas (body and tail of the pancreas) can be transplanted.
- It is placed in the ***right iliac fossa*** with anastomosis done between portal vein and iliac vein, duodenum fixed to bladder.
- Graft take up is 60-70%.

Complications

Graft pancreatitis, pancreatic leak, bleeding, urinary infections like cystitis, failure.

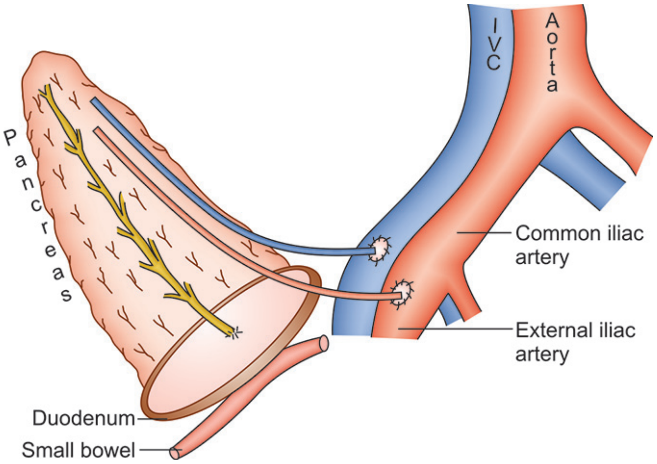

Fig. 1.693: Pancreatic transplantation with duodenum and small bowel to achieve adequate exocrine function. Exocrine function is achieved immediately; endocrine (insulin) function is achieved after few days.

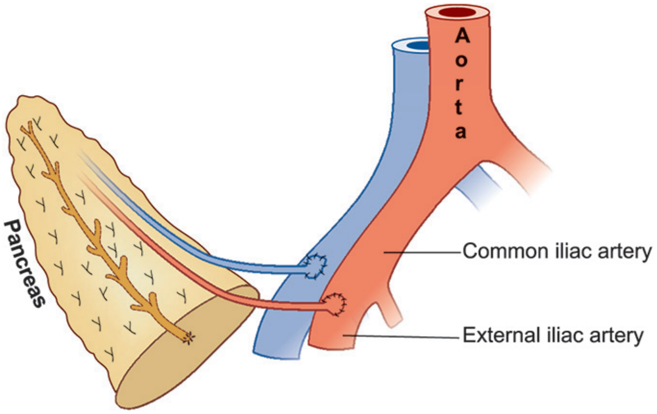

Fig. 1.694: Pancreatic transplantation.

Isolated Pancreatic Islet Transplantation

Islets of Langerhans are obtained by mechanical disruption of pancreas by injecting collagenase into the pancreatic duct. Tissue disrupted is collected and purified by density gradient centrifugation. These islet cells are injected into the liver through portal vein. Islet cell rejection is prevented by covering them with semipermeable membrane which prevents antibodies reaching islet cells but allowing insulin to get secreted. By this method animal pancreatic islet cells transplantation is also under trial.

SMALL BOWEL TRANSPLANTATION

- Indication is short bowel syndrome following massive resection, atresia, necrotising enteritis, Crohn's disease.
- Bowel anastomosis with a stoma (ileostomy) is usual method.
- As small bowel is rich in lymphoid tissue, graft versus host reaction is a major problem. So graft take up is poor.
- It is an immunological challenge even though technically easier.

Graft rejection

- **Hyperacute** rejection is due to antibody reaction which occurs within minutes of vascularisation of the graft
- **Acute** rejection results due to cell mediated immunity, occurring in a few days to first 3 months of transplant. It responds well to steroids and immunosuppressants
- **Chronic rejection** occurs in 12 months after transplant, is not very common. It is due to late vasculitis. When it occurs, responds very poorly to steroids or other drugs. It is common type

DIALYSIS

It is technique for the removal of waste product of metabolism, normalisation of plasma electrolytes and removal of plasma water.

Everytime something good happens to you, make something good happen to someone else.

Types

- Peritoneal dialysis.
- Haemodialysis.

Peritoneal Dialysis

Indications

- Acute renal failure until renal function recovers.
- Chronic renal failure until long term dialysis is instituted.

Contraindications

- Abdominal surgery.
- Hypercatabolic states.
- Infection.
- Pre-existing malignancy.

Insertion of catheter

A rigid plastic catheter is inserted through the abdominal wall into the peritoneal cavity using a trocar through a small cut made in the skin under L/A. Catheter is bound to abdomen.

Dialysis is done using sterile dialysate solution instilled into and drained out of abdomen.

Complications

- Leakage
- Pain—short lasting, mild
- Peritonitis
- Membrane failure
- Hernia

Continuous ambulatory peritoneal dialysis (CAPD) is the preferred method of dialysis in some centre.

Haemodialysis

Indications

- Acute renal failure.
- Chronic renal failure.
- Acidosis.
- Electrolyte disturbance.
- Intoxication.
- Uraemia (pericarditis and polyneuropathy due to uraemia are absolute indications).

Haemodialysis requires access to circulation which is achieved by creating a fistula between the radial artery and the cephalic vein at the wrist *(Cimino fistula).* Here dialysis occurs in a dialysing machine across a semipermeable membrane (usually cellulose membrane).

Complications

- Access site—arterial and venous stenosis, thrombosis, infection.
- Hypotension.
- Dyspnoea.
- Bleeding (due to dysfunctional platelets due to uraemia and due to heparin use).
- Disequilibrium syndrome.

Even though repeated dialysis (in CRF, at least twice a week is needed) helps, patient needs repeated blood transfusions, which will lead into haemochromatosis/iron overload. Patient should be given erythropoietin injection 3,000 units twice weekly to prevent repeated blood transfusions. But it is expensive. Transplantation is the best answer eventually, in all these individuals.

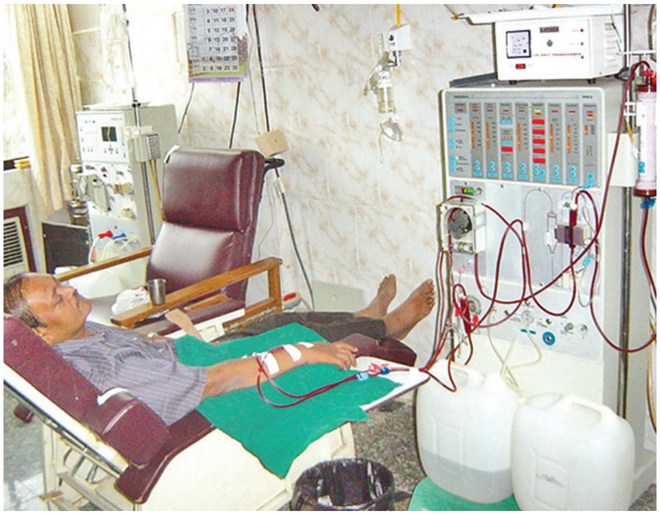

Fig. 1.695: Patient with CRF undergoing haemodialysis using a dialysis unit.

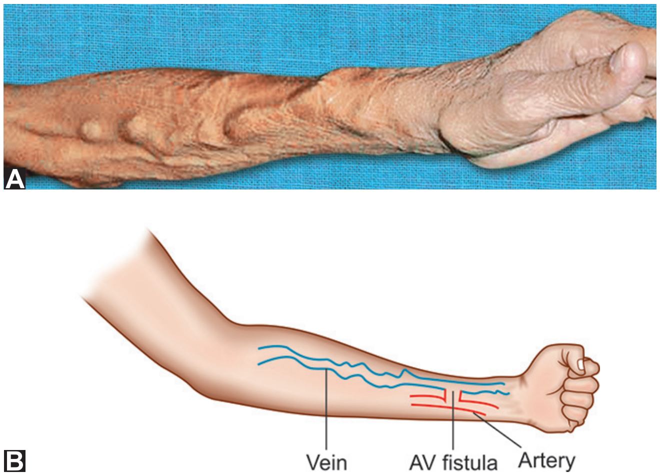

Figs 1.696A and B: Arteriovenous fistula between radial artery and cephalic vein at wrist is done to attain increased venous engorgement and arterialisation so as to achieve rapid blood flow which is necessary for haemodialysis. Patient with chronic renal failure who requires regular haemodialysis needs this.

CIMINO FISTULA (CIMINO-BRESCIA)

It is an arteriovenous fistula created for *haemodialysis.* Usually at wrist, radial artery is anastomosed to cephalic vein side-to-side and a created good fistula shows continuous thrill and bruit, with increased venous engorgement along with hyperdynamic circulation. At the ankle, often fistula is created between posterior tibial artery and saphenous vein; in the thigh between femoral artery and long saphenous vein. Distal gangrene is not common**.**

Complications

Infection, bleeding, hyperdynamic circulation.

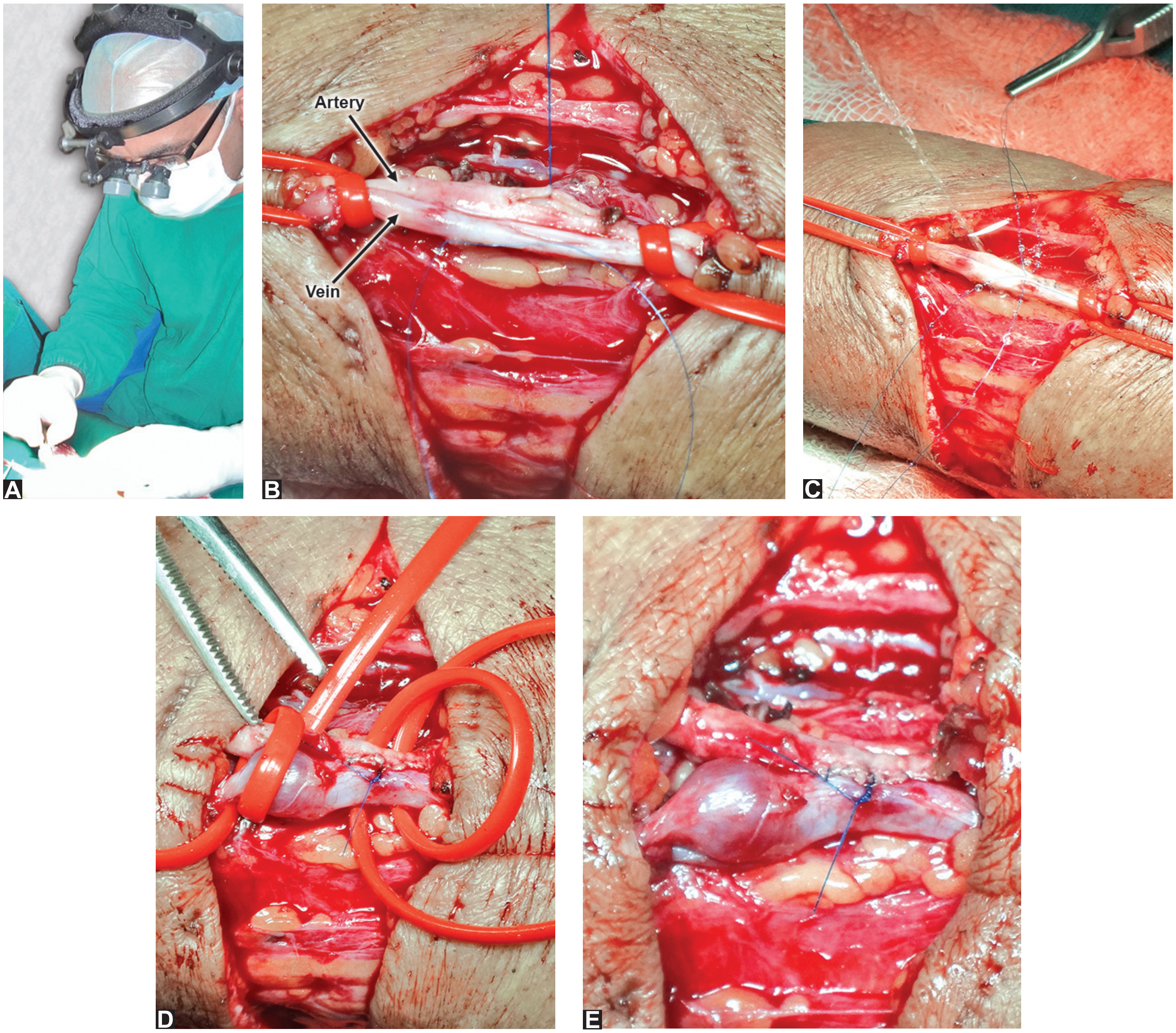

Figs 1.697A to E: Creation of Cimino AV fistula using radial artery and cephalic vein (magnifying operating loupe is used during anastomosis using 6 zero polypropylene) (*Courtesy:* Ashok Pandit, MCh, Urologist).

I am a great believer in luck and I find the harder I work, the more I have of it.

W. Stings and Bites

CHAPTER OUTLINE

- Snake Bite
- Spider Bite
- Bee Bite
- Mammalian Bite

SNAKE BITE

Snakes belong to Crotalidae, Elapidae and Viperidae family. Common poisonous snakes are cobra, krait and viper.

Common snakes are cobra, Russell's viper, saw scaled viper, pit viper, krait.

It causes renal failure, pulmonary oedema, cardiac complications and neurological problems.

Clinical Features

Local

- Burning pain, oedema and erythema.
- Swelling, echymoses and haemorrhagic bullae.
- Tissue necrosis, ulceration and gangrene.

Systemic

- Weakness, perioral paraesthesia, muscle twitching.
- Shock.
- Pulmonary oedema.
- Renal failure.
- Neurological manifestations.
- Bleeding tendency.

Vipers affect multiple organs and soft tissues. Cobra and coral snakes are neurotoxic.

Investigations

- Blood count.
- Coagulation studies like bleeding time, clotting time and prothrombin time.
- Blood urea and serum creatinine.
- Serum electrolytes.
- Creatine phosphokinase [CPK].
- Urine analyis for RBCs, albumin, myoglobin.
- PO_2 and PCO_2 assessment.

Management

First Aid

- Reassurance.
- Immobilisation.
- The site has to be incised and cleaned.
- Tourniquet to *occlude lymphatics* only but not venous or arterial circulation.
- The snake has to be identified.
- The patient must be transferred to proper medical centre as early as possible.
- The bite wound is identified and assessed.
- It is thoroughly cleaned with debridement.
- *Polyvalent antisnake venom* (against cobra, krait, vipers) should be given. It should be given earliest within 4-24 hours. Dose is 20-150 ml depending on the type, severity and age of the individual. It is dissolved in normal saline and given as IV infusion in 500 ml saline with 20 drops/minute as flow rate. If the snake has been identified monovalent serum is better and more potent.
- Tetanus toxoid.
- IV fluids, blood transfusion and plasma.
- Antibiotics.
- Urine output measurement.
- Monitoring by regular checking of blood urea, serum creatinine and bleeding and clotting time.
- In cobra bite, there is neuromuscular blockade and paralysis occurs. So *neostigmine* should be given 0.5 mg IV every half hourly and later repeated as required. It is given along with 0.6 mg of atropine.
- In viper bite DIC is common. So heparin is given as 10,000 to 15,000 units loading dose and later 5,000 units as maintenance dose 8th hourly.
- Human fibrinogen, whenever required.

Complications of Snake Bite

- Cellulitis and gangrene of the part.
- Deep venous thrombosis [DVT].
- Pancytopenia.
- DIC and haemorrhage.
- Neurological complications.
- Septicaemia.
- Renal failure.
- Marjolin's ulcer.

SPIDER BITE

Types

1. *Black Widow Spider Bite*
    - It causes neurotoxicity.
    - It causes muscle spasms, pain and abdominal cramp and rigidity.
    - Hypertension, tachycardia and diaphoresis are the other features.
    - Its effects last for 2-3 days.

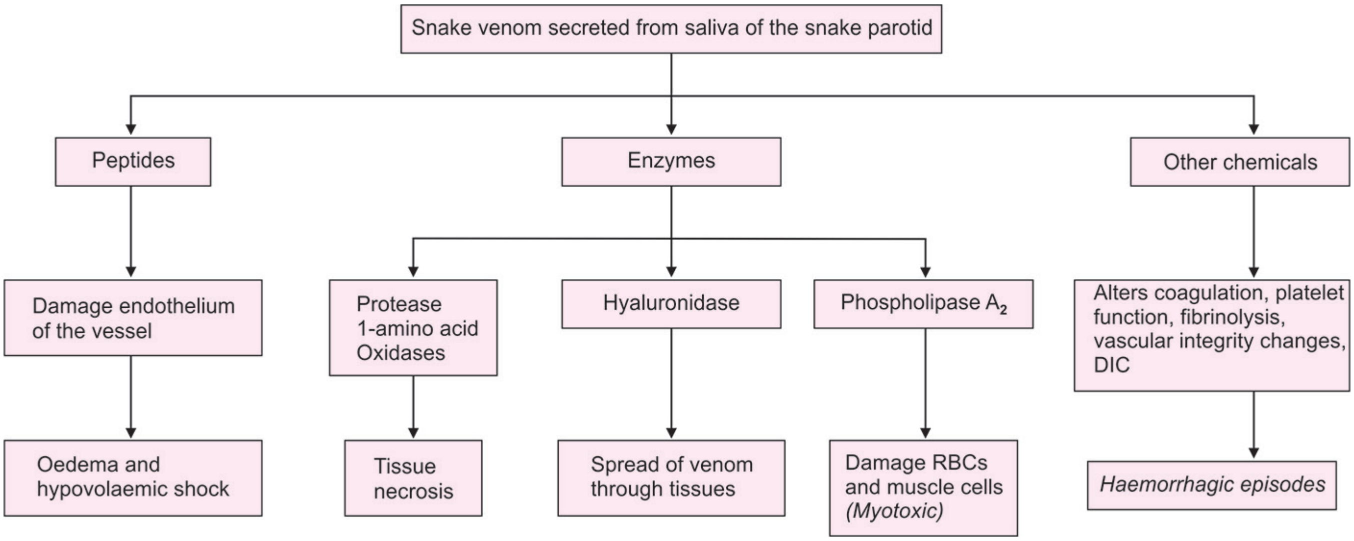
Snake venom secreted from saliva of the snake parotid
Peptides
Enzymes
Other chemicals
Damage endothelium of the vessel
Protease 1-amino acid Oxidases
Hyaluronidase
Phospholipase A_2
Alters coagulation, platelet function, fibrinolysis, vascular integrity changes, DIC
Oedema and hypovolaemic shock
Tissue necrosis
Spread of venom through tissues
Damage RBCs and muscle cells (Myotoxic)
Haemorrhagic episodes

- Cleaning the wound, antibiotics, antihistamines, specific antivenins are the line of treatment.

2. *Brown Spider Bite*
    - It releases sphingomyelinase-D which causes necrosis of the skin and haemolysis.
    - Local effects include skin rash, blister formation, necrosis of the skin and extensive ulceration.
    - Systemic effects are anaphylaxis, arthralgia, thrombocytopenia, haemolysis and renal failure.
    - Often it can be life-threatening especially in children and elderly.
    - The effects last for 2-3 weeks.
    - Wound debridement, antibiotics, antihistamines, steroid therapy, antivenins and management of systemic complication.

BEE BITE

Honey bee has got barbed stinger with two lancets. These lancets get attached to human skin to release the venom. Bee dies after bite.

Wasps: Yellow jacket wasps are more aggressive.

Bee venom contains dopamine, histamine, neurotoxin and toxic peptides.

Clinical Features and Management

- Allergic reactions.
- Anaphylaxis.
- Pain in the local region, oedema, pruritus, flushing.
- Hypotension, laryngeal oedema, bronchospasm.
- Muscle spasm, convulsions.
- Renal failure in severe cases.
- Soda bicarbonate is used to neutralise the bee venom.
- Antibiotics and antihistamines.

MAMMALIAN BITE

- It includes human bites also.
- Infection rate is more in mammalian bites.
- Proper wound toileting is very important.
- Within 12 hours, incised wound is closed primarily.
- All lacerated wounds and wound which is seen after 12 hours is left open. Wound is closed secondarily.
- Antibiotics are must in all mammalian bites.
- Human bite is very dangerous bite.

Awareness of change in important and equally acceptance of it also.

X. Pain

CHAPTER OUTLINE

- Gate Control Theory

Pain is a subjective one and is difficult to assess and quantify. Pain perception varies from person-to-person and from time-to-time.

- It can be physical or mental
- It can be localised or diffused
- It can be acute or chronic
- It can be intermittent or persistent
- It can be mild or severe

Pain Pathway

Pain receptors in the skin
↓
Neurotransmitters of pain like substance P or peptides are activated
↓
Sensory nerves
↓
Posterior horn of spinal cord
↓
Spinothalamic tract
↓
Thalamus
↓
Cerebrum
↓
Pain perception

GATE CONTROL THEORY

Gate control system is located at the junction of first and second neuron. Large diameter 'A' fibre is stimulated by temperature and touch. Fine 'C ' fibre is stimulated by pain. If 'A' fibre once gets stimulated, blocks the gate mechanism, then pain from 'C' fibre cannot pass through the gate to reach the brain for perception.

Pain modulators like endorphins and opioid peptides in brain and spinal cord inhibit the release of substance 'P'.

Causes of Pain

- Inflammatory causes due to any infection or infestations.
- Hypoxia due to poor blood supply like in myocardial infarction, peripheral vascular disease.
- Trauma.
- Obstruction like intestinal obstruction.
- Colicky pain like ureteric, biliary, intestinal.
- Compression over nerve roots like in inter vertebral disc prolapse.
- Advanced malignancies cause severe distressing pain, which requires proper pain control.
- Ulcers, perforation, peritonitis, abscess formation are all other causes.

Clinical Assessment of Pain

- Its severity, nature, cause should be assessed.
- Cause should be thoroughly analysed by doing all investigations like haematocrit, sinology, CT scan, culture of the fluid like pus, blood.

Pain is a most common symptom which patient complains to a clinician. Latin word *'poena'* means penalty/punishment. Pain is the one patient feels; tenderness (sign) is one surgeon/clinician elicits.

Specific Points in History in Relation to Pain to be Asked are:

- ***Original site of pain*** is very important. In acute appendicitis original site of pain is in umbilicus; but later it is referred to right iliac fossa. Shift of pain towards other site.
- ***Time and mode of onset of pain***—it is sudden onset and rapidly progressive in acute appendicitis; it is of insidious onset and of long duration with episodic nature in chronic peptic ulcer.
- ***Type/nature of pain***—superficial/deep; dull ache or sharp severe/pricking/bursting/vague aching (continuous mild pain), throbbing, scalding (burning sensation particularly felt during urination in cystitis, pyelonephritis, urethritis), pins and needles pricking sensation in peripheral nerve injury or irritation, shooting pain (seen in intervertebral disc prolapse and sciatica—pain shoots along the course of nerve), stabbing (sudden, severe, sharp, episodic—seen in perforated duodenal ulcer), distension pain (a feeling of restricted or distended like in paralytic ileus or intestinal obstruction), colicky pain is due to muscular contraction in a hollow tube in an attempt to obviate the obstruction by forcing the content out—gripping, episodic pain with vomiting and sweating (seen in intestinal colic, ureteric colic of stone, biliary colic of stone), twisting pain of bowel volvulus/twisted ovarian cyst/torsion testis, constricting pain around the chest by angina, etc.

- ***Severity of the pain:*** In acute conditions like peritonitis and abscess pain will be severe compared to chronic one.
- ***Progression of pain:*** It may be persistent and progressive; or initially mild gradually increases, later gradually subsides; fluctuation in intensity whether increases and decreases in intensity at regular intervals; quickly reaches maximum and remains like that.
- Duration of Pain
- ***Periodicity of pain:*** Pain appears, persists for few weeks and then disappears for few weeks; again reappear. Such periodicity is often observed in chronic peptic ulcer; trigeminal neuralgia.
- ***Precipitating/aggravating factors:*** Abdominal pain may get worsened by taking food like in gastric ulcer. Pain due to appendicitis, ureteric stone aggravates by change of position, walking, jolting. Pain of urinary bladder stone aggravates in standing position. In reflux oesophagitis pain increases while scooping. Pain in pancreatitis increases while lying down. Pain in intervertebral disc prolapse aggravates by lifting the weight.
- ***Relieving factors of pain:*** Pain reduce by certain methods and patient uses that method to relieve the pain. Hunger pain of early morning in duodenal ulcer is relieved by taking food. Pain of pancreatitis is relieved in sitting and bending forward. Propped up position relieves pain of reflux oesophagitis. In acute peritonitis, pain reduces temporarily by lying still.
- ***Associated symptoms***: Acute pain may be associated with pallor, sweating and vomiting. Migraine pain with vomiting and visual disturbances; intestinal/ureteric colic with sweating, vomiting and cold periphery; acute pyelonephritis and urinary infections with chills/rigors and fever; ureteric colic with haematuria; biliary colic with jaundice and pale stool are other examples of such association.
- ***Time of occurrence of pain*** is often important in diagnosing the condition. In duodenal ulcer, hunger pain occurring in early morning or later evening is typical. Migraine occurs in early morning; frontal sinusitis induced headache occurs a few hours after getting up.
- ***Pain may move from one place to other:*** *Radiation of pain* It is extension of pain from original site to another site with persisting of pain at original site. This radiating pain is of same character of original site. Penetration of duodenal ulcer posteriorly causes pain both in epigastrium and back—is an example. Pain of pancreatitis radiates to back. *Referred pain:* Pain is not felt at the site of the disease but felt at distant site. Diaphragmatic irritation causes referred pain at the tip of shoulder through same segmental supply of diaphragm (phrenic nerve C4, C5) and shoulder (cutaneous supply C4, C5). Hip joint pathology may cause referred pain in knee joint—through articular branches of femoral, obturator and sciatic nerves. Other *examples*—referred ear pain from carcinoma tongue through lingual and auriculotemporal nerve; referred pain in the epigastrium from the heart; referred pain in the abdomen from pleura; referred pain over the testis from the ureter. *Shifting/migration of pain:* Origin of pain is one site; later pain shifts to another site and pain at original site disappears. Pain when begins in viscera, it is felt at the same somatic segmental area in the body; but once parietal layer is involved by inflammation/pathology pain is felt at the anatomical site. Example is pain of acute appendicitis where original visceral pain is at the umbilicus (T9 and T10 segments supply both umbilicus and appendix) shifts later to right iliac fossa when once the parietal peritoneum of that area is inflamed.

Types of pain:

- *Superficial pain*: It is sharp usually localised pain, due to irritation of peripheral nerve endings in superficial tissue by chemical/mechanical/thermal/electrical injury. *Segmental pain:* It occurs due to irritation of particular nerve trunk/root; located in particular dermatome of the body supplied by the sensory nerve trunk or root.
- *Deep pain:* It is due to irritation of deeper structures like muscles/tendons/bones/joints/viscera. It is vague and diffuse when compared to superficial pain. It is often referred to common segmental areas of representation. Often spasm of skeletal muscle of same spinal cord segment can occur.
- *Psychogenic pain:* It may be functional/emotional/hysterical.
- *Other pain:* like due to thalamic/spinothalamic diseases/ *causalgia* (intense burning pain along the distribution of the partially.

Grading of pain is done using pain scale. It is compared to a 10 cm line numbered 0 to 10. This is called as *visual analogue scale* (VAS). Minimum is 0 means no pain. 10 is the worst excruciating pain. 2 is mild; 4 is discomforting; 6 is distressing; 8 is intense.

Reasons to control postoperative pain/acute pain

- Uncontrolled pain causes tachycardia, hypertension and vasoconstriction
- Abdominal (upper abdominal mainly) and thoracic wound pain restricts the respiration causing tachypnoea, altered respiration, coughing, chest infection, pneumonia
- Persisting pain causes restricted movements, deep venous thrombosis and its problems, bed sores
- Pain delays the recovery and also causes psychological trauma to the patient

Management of Pain

- Correct the cause like removal of renal stone, cholecystectomy for gallstones.
- Analgesics.
- Surgical removal of tumour.
- Injection of phenol or alcohol.
- Electric stimulation, massaging, infrared therapy, wax bath.
- Proper physiotherapy.
- Hormone therapy.
- Injection to ganglion like in trigeminal neuralgia.

Acute pain	*Drugs/opioids/epidural anaesthesia/drug infusion*
Chronic pain due to malignancy	First level—simple analgesics (aspirin, paracetamol, NSAIDs, tricyclic antidepressants)
	Second level—intermediate opioids codeine, tramadol, dextropropoxyphene
	Third level—strong opioids like oral morphine, intravenous morphine, subcutaneous diamorphine, epidural diamorphine. Analgesic infusion is also useful. Neurolytics are used whenever there is limited life-expectancy. Other methods are - Subcostal phenol injection for rib secondaries/celiac plexus block using alcohol/intrathecal hyperbaric phenol/percutaneous anterolateral cordotomy/pituitary ablation/hormone ablation/ palliative radiotherapy/steroids/flecainide therapy
Chronic pain in benign disease	Local anaesthetics/steroid injections/nerve stimulation like acupuncture/nerve decompression/pain due to sympathetic overactivity is corrected by—IV pentolamine (alpha adrenergic block), stellate ganglion block, block using guanethidine, percutaneous chemical lumbar sympathectomy using phenol. Other methods for chronic pain are—paracetamol/NSAIDs/antidepressants/opioids/carbamazepine/ neuroablative surgeries

- Chemotherapy for malignancies.
- Radiotherapy.
- Sympathectomy for vascular diseases, causalgia.
- Cordotomy for severe pain in case of advanced tumours.
- Mental relaxation.
- Attending pain clinic.
- Continuous epidural anaesthesia/analgesia using opioids.
- *Patient controlled analgesia* (PCA) is injecting opioids through epidural route or intravenous route by patient himself after training him.
- Intravenous infusions of the analgesia.

Pain may be:

- Acute pain.
- Chronic pain due to malignancy.
- Chronic pain due to benign disease.

Drugs for Pain

Drugs can be given orally, intramuscularly, few intravenously, intrathecally, per rectally as suppositories, sublingually.

Narcotic Analgesics

- Morphine 10-15 mg. Very useful in intractable pain. It can cause nausea, constipation, and respiratory depression. Its action is neutralised by naloxone.
- Pethidine 50-100 mg IM.
- Diamorphine 5-10 mg. It is used only in intractable pain.
- Pentazocine 30-60 mg.
- Dihydrocodeine.
- Codeine phosphate is commonly used after intra-cranial surgery. It should not be given intravenously as it may precipitate severe hypotension. Codeine also causes constipation.
- Bupremorphine.

Non-narcotic Drugs

- Aspirin.
- Ibuprofen.
- Naproxen.
- All these NSAIDs cause gastric irritation and may risk the gastric bleeding. NSAIDs are non-specific cyclooxygenase inhibitors. So gastric protection and platelet function are lost. Rectal diclofenac as suppository is good analgesic.
- Paracetamol is effective analgesic and antipyretic agent. It is relatively less common to cause gastric bleeding.

***Note*:**
Specific COX2 inhibitors are contraindicated in ischaemic heart disease.

Other Drugs

- Anticonvulsants.
- Antidepressants, anxiolytics.
- Carbamazepine for neuralgia.

Chapter 2 Faciomaxillary Diseases

CHAPTER OUTLINE

- Diseases of the Palate
- Orthopantomogram
- Cleft Lip and Cleft Palate
- Maxillofacial Injuries
- Primary Care (Early Care) in Maxillofacial Injuries
- Fracture Middle Third Area
- Zygomatic Complex Fracture
- Fracture of the Mandible
- Dislocation of the Mandible
- Preauricular Sinus
- Jaw Tumours
- Epulis
- Ameloblastoma
- Dentigerous Cyst
- Dental Cyst
- Osteomyelitis of Jaw
- Alveolar Abscess
- Fibrous Dysplasia of Bone/Jaw
- Cherubism

DISEASES OF THE PALATE

- Cleft palate.
- *Torus palatinus*—a bony hard swelling in the centre of the hard palate.
- Nasopalatine cysts.
- *Epstein's pearls*—at the junction of soft and hard palates, in the midline, in infants due to retained developmental cell rests.
- Apical cyst or abscess.
- Minor salivary gland tumour—commonest site is palate.
- Maxillary tumour extending into the palate.
- Squamous cell carcinoma of the palate.
- Gummatous perforation in the middle of the palate seen in congenital syphilis.
- Perforation of the palate anywhere in carcinoma palate.

ORTHOPANTOMOGRAM (OPG)

It is a plain X-ray of the mandible which shows the entire mandible in a single plane. It is better than X-ray mandible lateral view as it highlights proper dentition, inner and outer plates of mandible and joints. It is like a rotational tomogram.

Indications

a. Jaw tumours—Adamantinoma, dental cyst, dentigerous cyst, osteoclastoma.
b. Osteomyelitis of the mandible.
c. Fracture mandible.
d. Carcinoma oral cavity to check *infiltration* into the mandible.

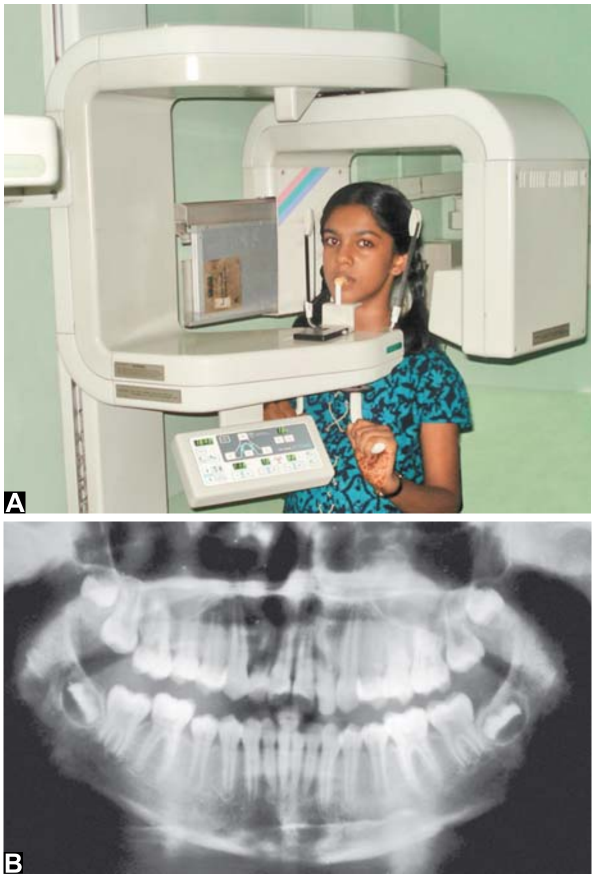

Figs 2.1A and B: Orthopantomogram is being taken and also picture showing OPG X-ray.

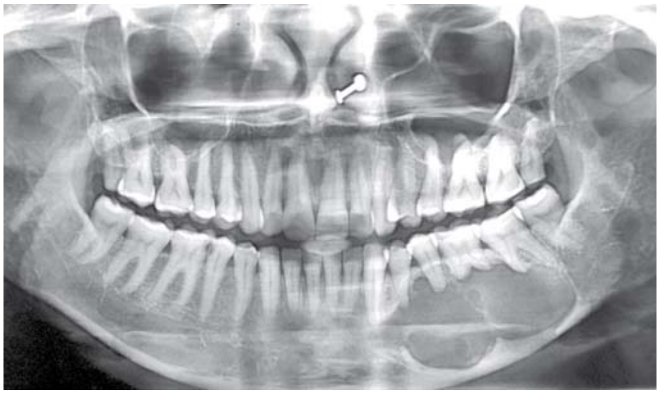

Fig. 2.2: OPG showing jaw tumour probably adamantinoma.

CLEFT LIP AND CLEFT PALATE

Development of Face

Face develops from median nasal process, lateral nasal process, maxillary process, mandibular arch, globular arch, olfactory pit and eye. Any change in the development or fusion of these arches leads to formation of different types of cleft lip or cleft palate.

Aetiology

- Familial—more common in cleft lip or combined cleft lip and palate (Risk is 1:25 live births).
- Protein and vitamin deficiency.
- Rubella infection.
- Radiation.
- Chromosomal abnormalities.
- Maternal epilepsy and drug intake during pregnancy (steroids/eptoin/diazepam).

Classification

I. *Cleft lip alone*: Unilateral.
Bilateral.
Median.
II. *Cleft of primary palate (in front of incisive foramen) only:*
    a. Complete—means absence of pre-maxilla.
    b. Incomplete—means rudimentary pre-maxilla.
        i. Unilateral.
        ii. Bilateral.
        iii. Median.
III. *Cleft of secondary palate (behind the incisive foramen) only:*
    a. Complete—nasal septum and vomer are separated from palatine process.
    b. Incomplete.
    c. Submucous.
    It can be - Cleft with soft palate involvement.
    - Cleft without soft palate involvement.
IV. *Cleft of both primary and secondary palates.*
V. *Cleft lip and cleft palate together.*

Defect is often associated with other congenital anomalies of cardiac, gastrointestinal, neurological system, *Pierre-Robin syndrome* (most commonly associated syndrome with features of isolated cleft palate, retrognathia, posteriorly displaced tongue), *Klippel-Feil syndrome*, *Stickler's syndrome* (eye, skeletal, muscular, cleft disorder), *Shprintzen's syndrome* (cardiac and cleft disorder), *Down's syndrome, Treacher-Collin's syndrome, Apert's syndrome* and trisomy.

Cleft lip

- *Central*—rare. In upper lip. Between two median nasal processes. (*Hare lip)*
- *Lateral*—maxillary and median nasal process, *commonest*; can be unilateral or bilateral
- *Incomplete* cleft lip does not extend into nose
- *Complete* cleft lip extends into nasal floor
- *Simple* cleft lip is only cleft in the lip
- *Compound* cleft lip is cleft lip with cleft of alveolus

LAHS classification of cleft disorders

- 'L' for lip, 'A' for alveolus, 'H' for hard palate, 'S' for soft palate
- *Capital* 'LAHS' for 'complete' type
- *Small letters* 'lahs' for 'incomplete type'
- *Asterisks* 'lahs' for microclefts
- 'LAHSHAL' for bilateral clefts

Incidence

- Common in Caucasians.
- In 75% of cases it is unilateral. Commonly occurs on the left side (60%).
- In 50% of cases it is combined cleft lip and palate. Incidence is 1:600 live births. Common in boys.
- In 15-25% of cases it is cleft lip alone.
- In 25-40% of cases it is cleft palate alone. Incidence is 1:1000 live births. More common in girls.

Problems in Cleft Disorders

- Difficulty in sucking and swallowing. This is commonly observed in cleft palate than in cleft lip.

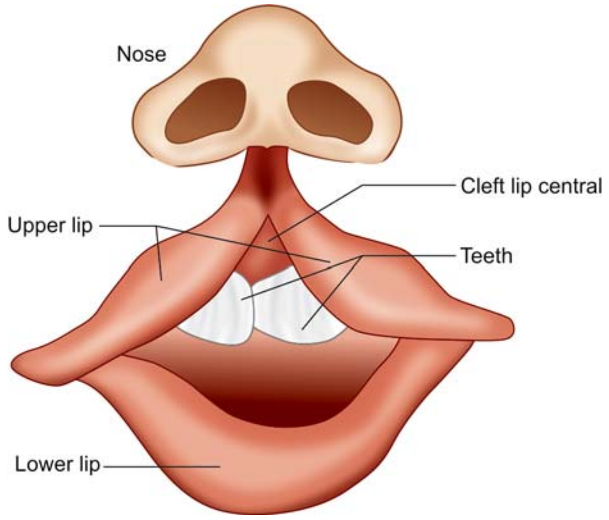

Fig. 2.3: Central cleft lip (Hare lip, Type I cleft lip—It is rare).

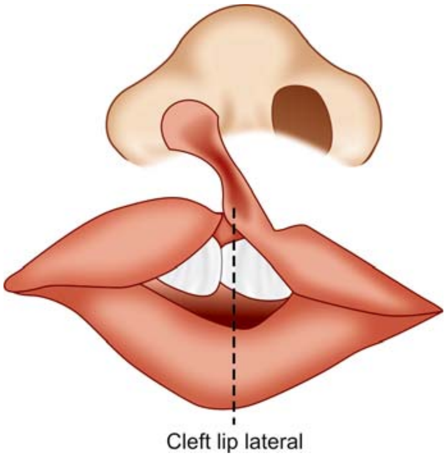

Fig. 2.4: Lateral type of cleft lip (Type II variety—it is commonest). It is due to imperfect fusion of maxillary process and median nasal process. It can be unilateral or bilateral.

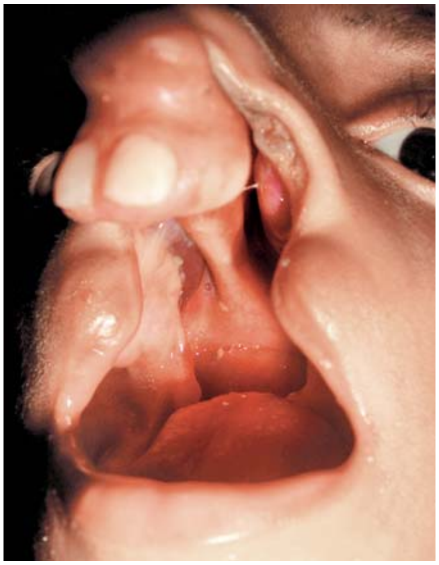

Fig. 2.5: Wide cleft palate and lip—complete type in a child.

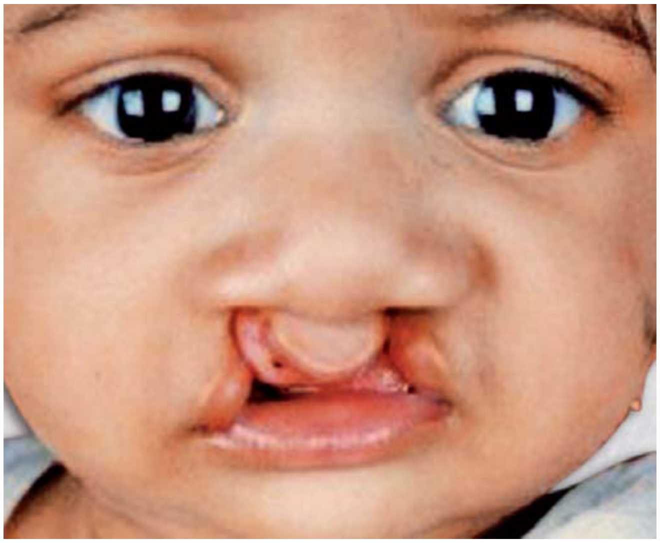

Fig. 2.6: Bilateral cleft lip.

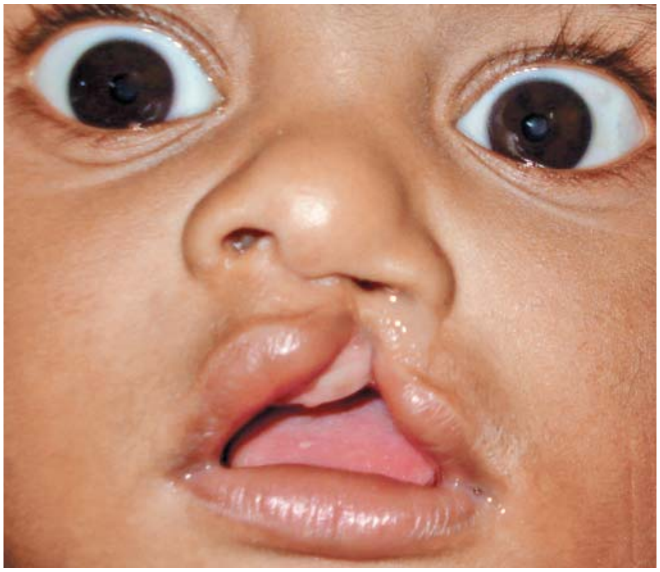

Fig. 2.7: Unilateral cleft lip, lateral type which is commonest.

- Speech is defective especially in cleft palate, mainly to phonate B, D, K, P, T and G.
- Altered dentition or supernumerary teeth.
- Recurrent upper respiratory tract infection.
- Respiratory obstruction (in *Pierre Robin syndrome*)
- Chronic otitis media, middle ear problems.
- Cosmetic problems.
- Hypoplasia of the maxilla.
- Problems due to other associated disorders.

Treatment for Cleft Lip

Millard criteria is used to undertake surgery for cleft lip.

Millard criteria (Rule of '10')

- 10 pound in weight
- 10 weeks old
- 10 gm % haemoglobin

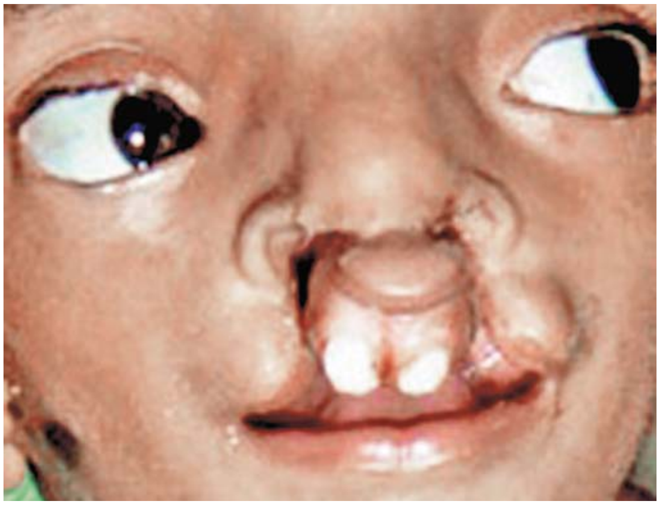

Fig. 2.8: Bilateral cleft lip involving palate.

Bleeding gums in uraemia are not so spongy as in scurvy: in fact they may look nearly normal.—**Frederic J Wright**

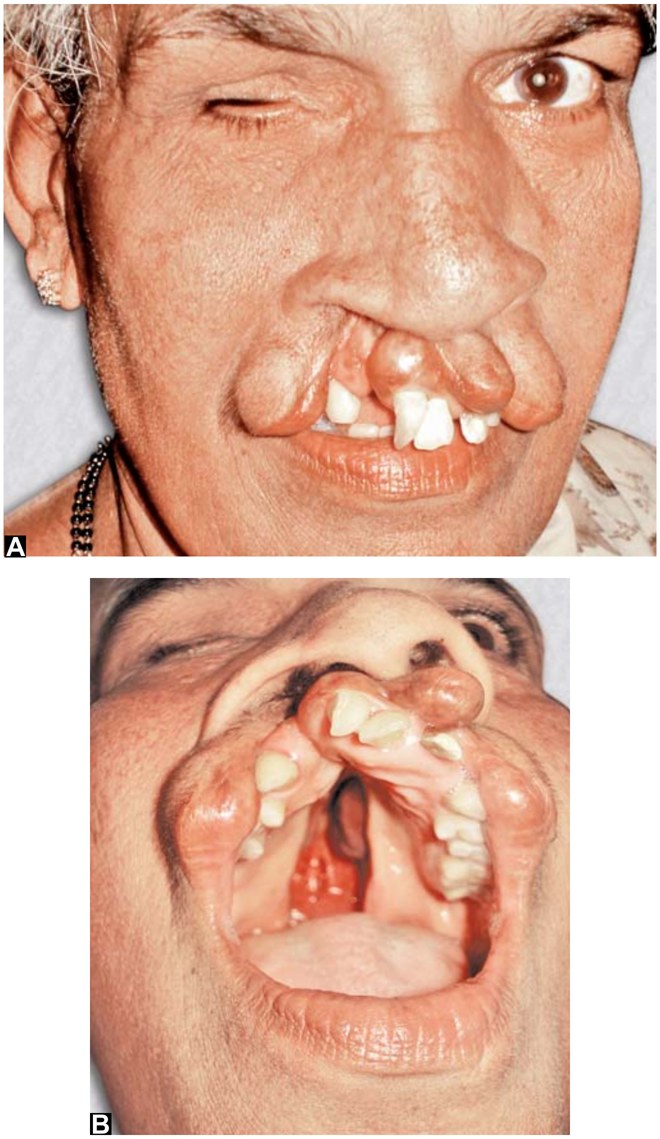

Figs 2.9A and B: Cleft lip and cleft palate in an adult.

a. *Millard cleft lip repair* by rotating the local nasolabial flaps.
b. Management of associated primary or secondary cleft palate deformity.
c. Proper postoperative management like control of infection, training for sucking, swallowing and speech.
d. Tenninson's 'Z' plasty (*Tenninson-Randall triangular flap*).

Note: *Delaire timing of the cleft surgery* – Unilateral/ bilateral cleft lip alone, in one stage operation done in 4-6 months. For cleft palate alone involving only soft palate, in one stage surgery is done in 6 months. For cleft palate alone but involving both soft and hard palates – soft palate in 6 months; hard palate in 18 months. In combined cleft lip and palate, unilateral or bilateral, in two stages – cleft lip and soft palate in 6 months; hard palate in 18 months.

Cleft Palate

- It is due to failure of fusion of the two palatine processes.
- Defect in fusion of lines between premaxilla (developed from median nasal process) and palatine processes of maxilla one on each side.

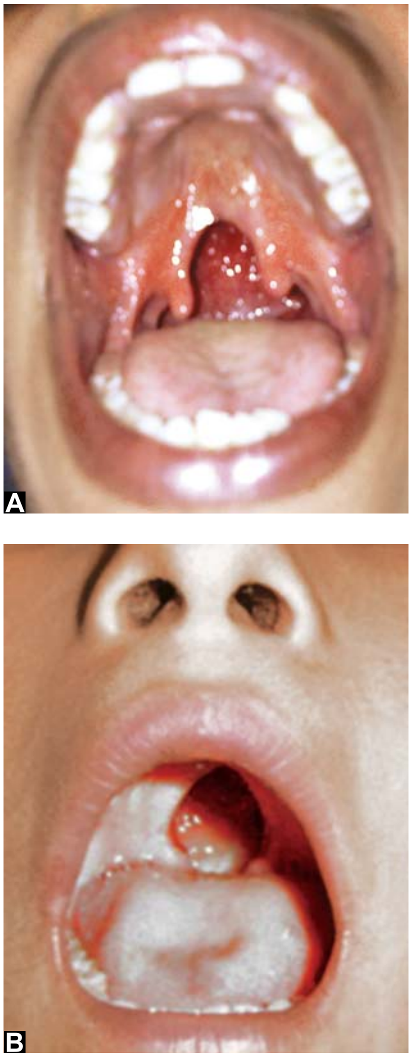

Figs 2.10A and B: Cleft palate only. Lip is normal. Pre-maxilla is not involved.

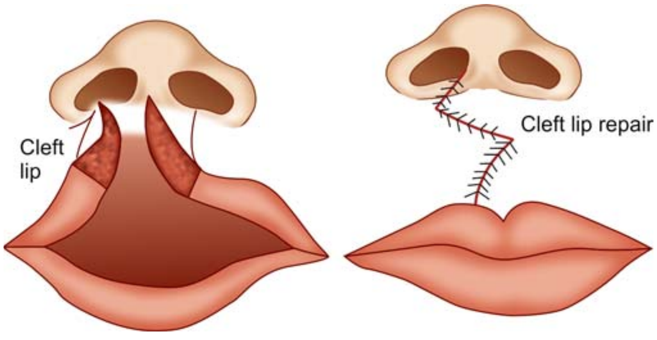

Fig. 2.11: Steps of cleft lip repair.

Principles of cleft lip repair

- "Rule of 10' should be fulfilled
- Before 6 months it should be operated
- Infection should not be present
- *Millard* advancement flap is commonly used for unilateral cleft lip repair
- Bilateral cleft lip repair can be done either in single or two stages (with 6 months gap between each stage)

- One stage bilateral cleft lip repair is done using *Veau III method/Millard's single stage/Black* method
- Proper markings are made prior to surgery and incision should be over full thickness lip
- Often 1 : 2,00,000 adrenaline injection is used to achieve haemostasis
- Three layer lip repair should be done (mucosa, muscle and skin)
- Cupid's bow should be horizontal
- Continuity of white line should be maintained
- Vermilion notching should not be there

- When premaxilla and both palatine processes do not fuse, it leads into complete cleft palate (Type I cleft palate).

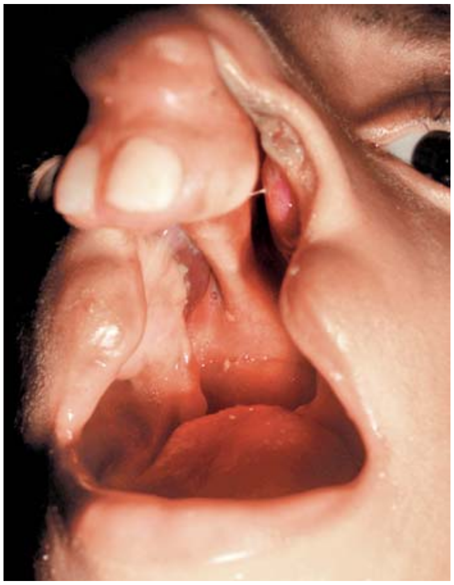

Fig. 2.12: Cleft palate with cleft lip in a child. Child also is suffering from congenital cardiac defects.

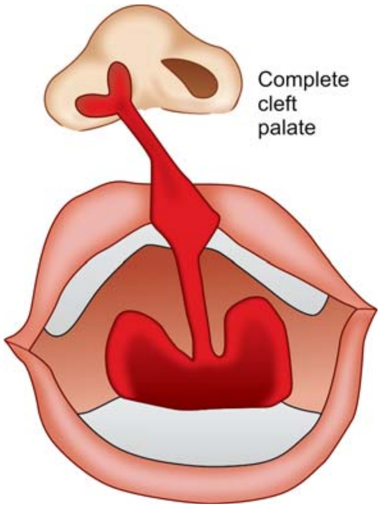

Fig. 2.13: Complete cleft palate Type I.

- Incomplete fusion of these three components can cause incomplete cleft palate beginning from uvula towards posteriorly at various lengths. So it could be Type II a–bifid uvula, Type II b–bifid soft palate (entire length) or Type II c –bifid soft palate and posterior part of hard palate (but anterior part of hard palate is normal).

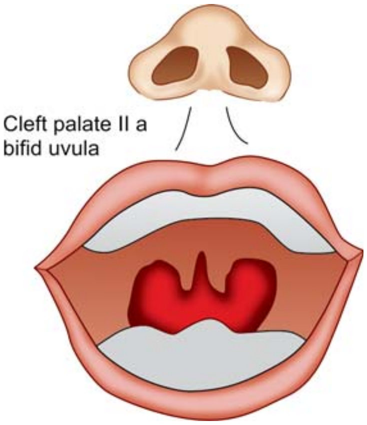

Fig. 2.14: Cleft palate Type II a—bifid uvula.

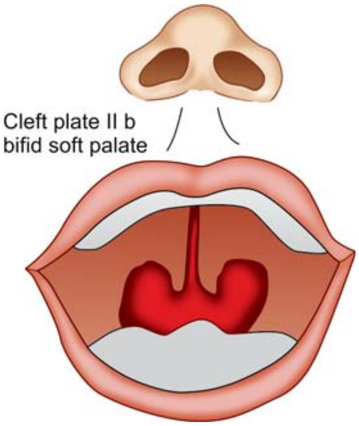

Fig. 2.15: Cleft palate Type II b—bifid soft palate (entire soft palate).

- Small maxilla with crowded teeth, absent/poorly developed upper lateral incisors.

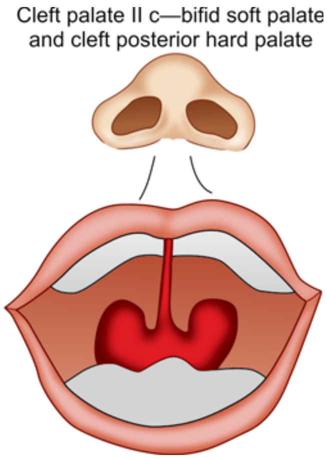
Cleft palate II c—bifid soft palate
and cleft posterior hard palate

Fig. 2.16: Cleft palate Type II c—bifid soft palate entire length with cleft of posterior hard palate (anterior palate is normal).

Preauricular ulcer is a late stage of an abscess due to a congenital preauricular sinus. It refuses to heal for infection is maintained from the sinus.
— Francis AR Stammers

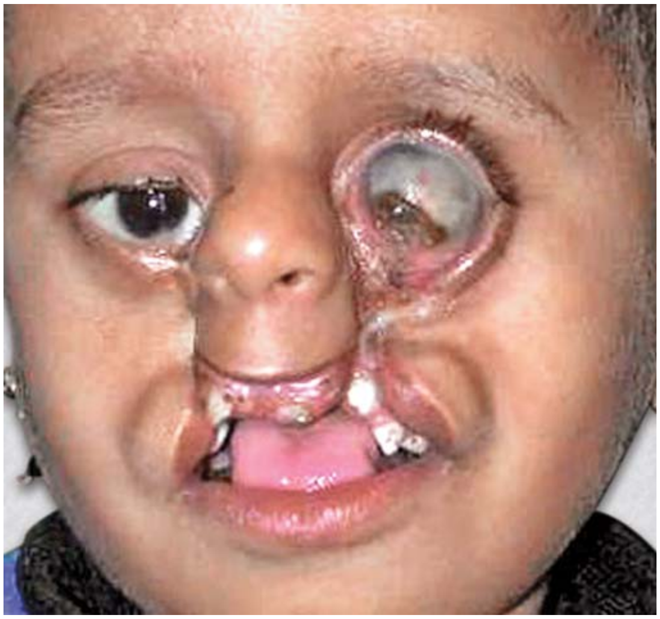

Fig. 2.17: Bilateral congenital craniofacial cleft (*Courtesy*: Dr Sathish Bhat, Plastic Surgeon, Mangalore).

- Bacterial contamination of upper respiratory tract with recurrent infection is common.
- Chronic otitis media with deafness may occur.
- Swallowing difficulties to certain extent and speech problems can occur.
- Cosmetic problems can occur.

Treatment for Cleft Palate

Criteria for surgery

- 10 kg weight
- 10 months of age (10-18 months)
- 10 gm % haemoglobin

- Cleft palate is usually repaired in 12-18 months. Early repair causes retarded maxillary growth (probably due to trauma to growth center and periosteum of the maxilla during surgery if done early). Late repair causes speech defect.
- Both soft and hard palates are repaired.
- Abnormal insertion of tensor palati is released. *Mucoperiosteal flaps* are raised in the palate which is sewed together. If maxillary hypoplasia is present, then *osteotomy of the maxilla* is done. With *orthodontic help teeth extraction and alignment* of dentition is done.
- Regular examination of ear, nose and throat during follow up period.
- Postoperative speech therapy.
- Whenever complicated problems are present, staged surgical procedure is done.
- *Wardill- Kilner push back operation*—by raising mucoperiosteum flaps based on greater palatine vessels.
- *Secondary management:*
  - Hearing support is given using hearing aids if defect is present; control of otitis media.
  - Speech problems occur due to velopharyngeal incompetence; articulation problems also can occur – speech therapy is given. It is corrected by pharyngoplasty, veloplasty, speech devices.
  - Dental problems like uneruption, unalignments are common. They should be corrected by proper dentist opinion, and reconstructive surgery.
  - Orthodontic management with alveolar bone graft, maxillary osteotomy – done in 8-11 years of age.
  - Veloplasty, dental implants, rhinoplasty, orthognathic surgeries, etc.

Principles of palatoplasty

- Timing is between 10-18 months
- Mucoperiosteum flap is raised
- Palatal defect is closed using 3 layers—nasal, muscle and oral layers
- Hook of pterygoid hamulus is fractured to relax tensor palate muscle to relieve tension on suture line

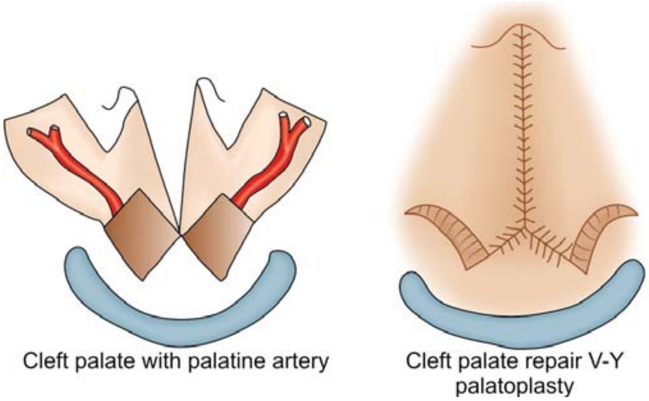
Cleft palate with palatine artery
Cleft palate repair V-Y palatoplasty

Fig. 2.18: V-Y palatoplasty.

MAXILLOFACIAL INJURIES

It may be due to road traffic accidents, assaults, bullet injuries or sport injuries.

Classification

Fracture in maxillofacial region can be grouped as:

- Fracture lower third that comprises mandible.
- Fracture middle third that comprises maxilla, zygoma and nose.
- Fracture upper third of the face involving part of the orbit, frontal bones.

Maxillofacial fracture also can be grouped as:

- Fractures of the face which do not involve the dental occlusion—fractures of zygoma and nose.
- Fracture which involves the dental occlusion – fracture mandible and maxilla.

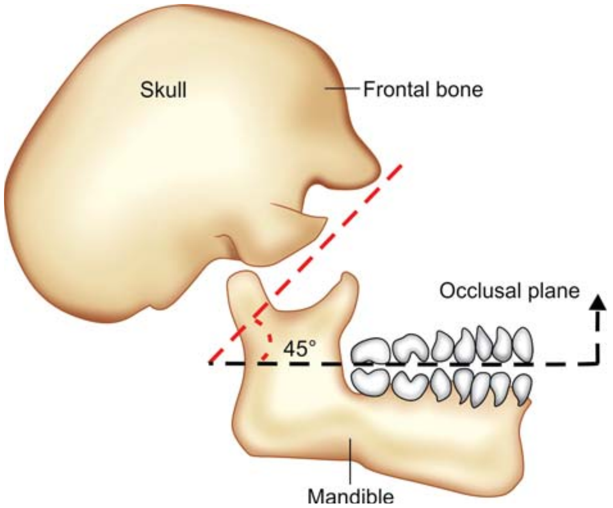

Fig. 2.19: Relation between middle third and cranium in 45° plane.

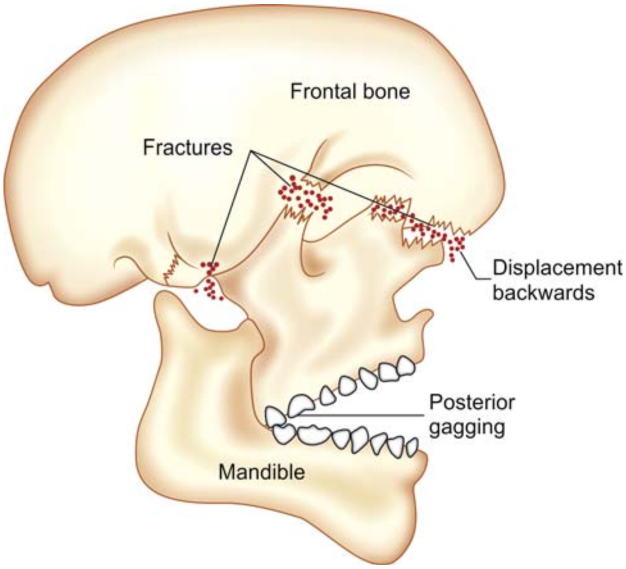

Fig. 2.20: Posterior gagging of occlusion due to backward displacement of fracture segment in middle third fracture.

Soft Tissue Injuries

- Lacerations, contusions, cut wounds, etc.
- Eyelid injuries with black eyes.
- *Facial nerve injury*: Primary repair is required.
- *Parotid duct injury*: Here primary anastomosis of the injured duct is done, with a fine polythene cannula is kept as a stent inside the duct which will be removed in 14 days.
- *Lacrimal apparatus injury*: Here the duct is sutured with a fine nylon thread in the canaliculus which is kept for 3 months.

Injuries to the Facial Bones

- *Fracture nose*: Nasal bones are most commonly injured bones in face. Patient presents with pain and swelling in the nose with deviation and displacement. Here reduction of the fractured nasal bones and nasal septum under general anaesthesia is done. Later position is maintained by nasal packs from inside (which is removed in 7 days) and by a nasal plaster from outside (which will be kept for 14 days). Procedure is done using Walsham's and Asch's forceps.
- Injuries to the *maxilla.*
- *Zygomatic* bone injuries.
- *Mandibular* bone fracture and mandibular dislocation.
- *Orbital bone fracture*: Presents with diplopia, enophthalmous, sensory loss in the area of infraorbital nerve.
- Infraorbital ecchymosis of the orbit is called as *Panda sign*.

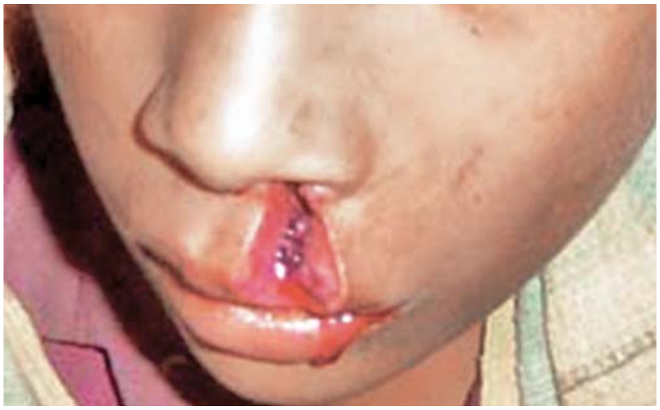

Fig. 2.21: Soft tissue injury face involving upper lip.

Clinical Features

- Localised swelling due to haematoma.
- Facial oedema.
- Bleeding with open wounds.
- Asymmetry which is clinically confirmed by observing supraorbital ridges, nasal bridge.
- Localised tenderness.
- Step deformity.
- Trismus.
- Diplopia.
- Features of associated injuries like intracranial, abdominal or thoracic injuries.

Investigations

- X-ray face.
- CT scan of head and jaw.

General treatment for faciomaxillary injuries

- Suturing of soft tissues
- Airway maintenance
- Control of bleeding
- Pain relief
- Control of infection
- Treating the individual fractures

Ankylosis of mandible joint causes receding of chin giving a characteristic shrew mouse profile.
—Leon Dufourmentel

PRIMARY CARE (EARLY CARE) IN MAXILLOFACIAL INJURIES

Injury can be isolated single bone fracture or multiple bone fractures. Real primary care is usually not required except when there is mechanical respiratory block causing airway obstruction.

Respiratory Obstruction

Causes

- Oronasal airway block can occur by blood, clot, vomitus, foreign body, dentures, teeth, saliva, bone pieces, etc.
- Backward falling of tongue can cause obstruction of the nasopharynx and oropharynx. It is common in bilateral mandibular fracture.
- Occlusion of the nasopharynx and oropharynx can occur in fracture maxilla with posterior and inferior displacement.
- Haematoma in floor of the mouth or posterior oral cavity can cause airway block.
- Oedema of larynx/tongue/posterior third of oral cavity/ pharynx.
- Surgical emphysema.

Treatment

- Cleaning of the oral and nasal cavities to remove obstructing agents like clot, dentures, teeth or bone. Gauze swabbing and suction.
- Fallen tongue should be placed forward using finger and often temporary alignment of the occlusion may be needed.
- Maxillary disimpaction is done when needed in fracture maxilla.
- Positioning of the patient is important. Prone/semiprone position with head towards one side is the safest position. If this is not possible, then patient may be placed in sitting position which also improves the breathing. Placing the patient flat on his back in supine position should be avoided as much as possible.
- Tracheostomy should be done when needed without delay as it will be life saving by facilitating the easy airway and breathing.

Control of Pain

Analgesics like NSAIDs are used to control pain. Morphine and analogues are not used as they may suppress the respiration. They may mask the pain of alarming severe injury in chest, abdomen or other areas, or they may interfere with pupillary reaction and neurological signs in the presence of intracranial injuries.

Control of Infection

Antibiotics are needed. Tetanus toxoid and often antitetanus globulin (ATG 3000 units IM) are required.

Haemorrhage in Maxillofacial Injuries

Haemorrhage in maxillofacial injuries is usually not life threatening. But it should be identified and controlled properly. In association with other internal injury, such haemorrhage may be important to cause the circulatory failure.

Haemorrhage may be due to:

- Soft tissue bleeding.
- Bleeding from inferior alveolar artery, palatine vessels.
- Nasal bleeding.

Control of Bleeding

- Blood transfusion, IV fluids, resuscitation.
- Nasal packs.
- Fracture correction.
- Ligation of the bleeder.
- Cauterisation.
- Packing the area.
- Under running the bleeding field.
- Embolisation.
- External carotid artery ligation above the level of the origin of the superior thyroid artery.

Fractures not involving occlusion	***Fracture involving occlusion***
Central	***Dentoalveolar***
• Fracture nasal bones and/ or nasal septum	*Subzygomatic* • Le Fort-I-low level either unilateral or bilateral • Le Fort II-pyramidal either unilateral or bilateral
• Fracture of frontal process of maxilla	
• Fractures of above two extending into ethmoid-nasoethmoid	
• Fractures above three which extends into frontal bone – fronto-orbito-nasal dislocation	*Suprazygomatic* • Le Fort III – high level • Craniofacial disjunction– unilateral or bilateral
Lateral	
• Fractures involving zygomatic bone, arch and maxilla excluding the dentoalveolar component	

Le Fort classification	
(*Rene Le Fort*—French surgeon classified these fractures by dropping rocks on the face of the cadavers and later dissected the area for study and research and published paper in 1911)	
Types	***Features***
Le Fort I (Guerin's fracture-low level) (floating fracture, horizontal fracture of maxilla)	• Bleeding from nose • Posterior gagging of occlusion • Upper lip swelling • Palatal ecchymosis • Occlusion derangement • Floating maxilla
• It runs horizontally above the floor of the nasal cavity involving lower third septum, palate, alveolar process of maxilla and lower third of pterygoid plates of maxilla	

Le Fort II (pyramidal fracture) • From the nasal bones at top-most, fracture runs laterally towards lacrimal bones, medial wall of orbit, infraorbital margin, through ormedial to infraorbital foramen and backwards below the zygomatico maxillary area through lateral wall of maxillary sinus and pterygoid plates. Zygoma is intact with skull base	• Oedema of middle third face • Both sides circumorbital and subconjunctival ecchymoses • Nasal bleeding/ obstruction/deformity • Deformity of face (dish face), diplopia • Retroposition of maxilla with posterior gagging • Limitation of ocular movements, CSF rhinorrhoea • Tenderness and separation of infraorbital margin
Le Fort III (craniofacial disjunction, high level) • Here fracture runs parallel to skull base. It passes through the nasal bone, lacrimal bone, ethmoid bone, optic foramen, inferior orbital fissure, pterygomaxillary fissure and lateral orbital wall with frontozygomatic suture with zygomatic arch	• Lengthening of face • Enophthalmos, ocular level depression • Hooding of eyes, occlusal plane tilting • Entire facial skeleton moves as a single block • Tenderness and separation of suture line • Diplopia • Trismus, teeth malalignment *Guerin sign*: Haematoma at greater palatine foramen

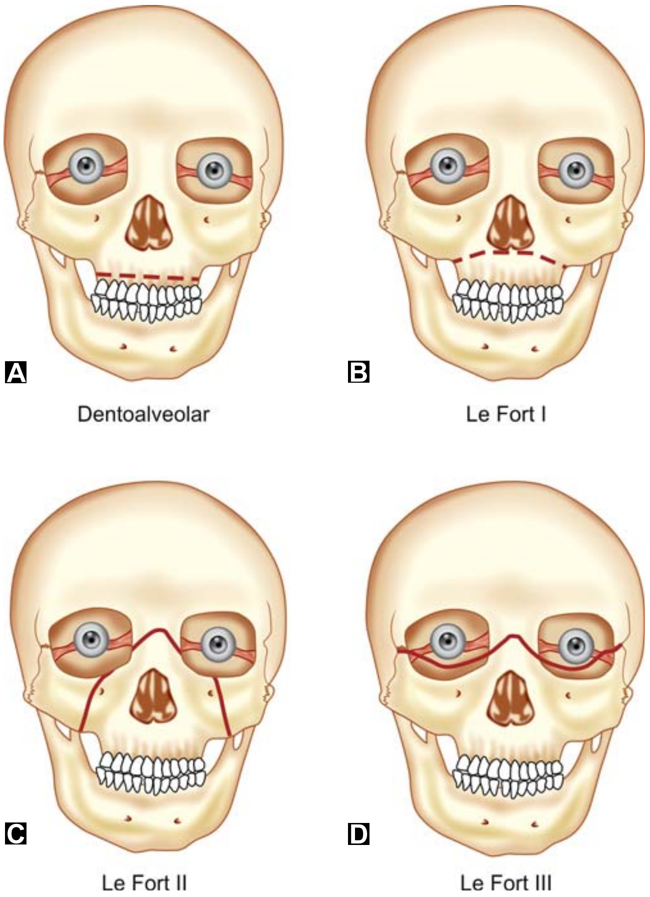

Figs 2.22A to D: Le Fort classification–different types and also dentoalveolar fracture (*Refer table for details*).

Associated Injuries

All associated injuries should be assessed properly and individually. On priority basis it should be treated.

- Soft tissue injuries.
- Cranial injuries.
- Orbital injuries.
- Intra-abdominal/thoracic/pelvic injuries.

FRACTURE MIDDLE THIRD AREA

It includes-

- Maxillae, zygomatic bones, palatine bones, nasal bones, lacrimal bones, inferior conchae (one on each side).
- The vomer, ethmoid and its attached conchae, pterygoid plates of sphenoid.

Fracture middle third includes fracture maxilla, zygoma and nasal bones.

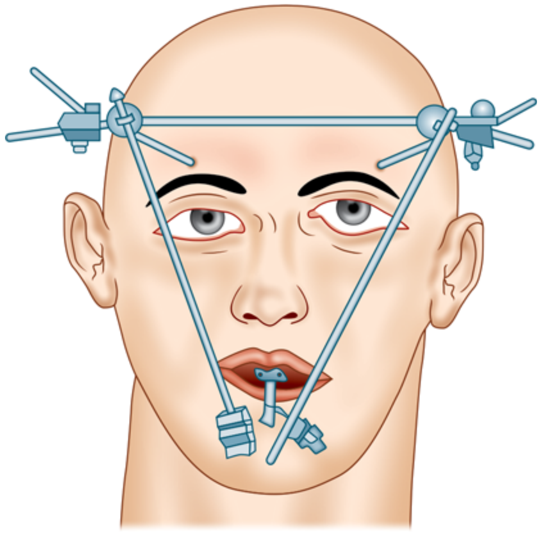

Fig. 2.23: Fixation of the splint to skull. It is often used with gunning splints to fix it to skull.

Clinical Features

- Oedema face, subconjunctival haemorrhage, ocular ecchymosis.
- Bleeding from the nose.
- Diplopia due to trapping of the extra-ocular muscles in the fracture segments.
- Anaesthesia of the cheek.
- Trismus and malalignment of teeth.
- *Guerin's sign*: Haematoma at greater palatine foramen.
- Always patient should be examined and observed for CSF leak and intracranial injuries.

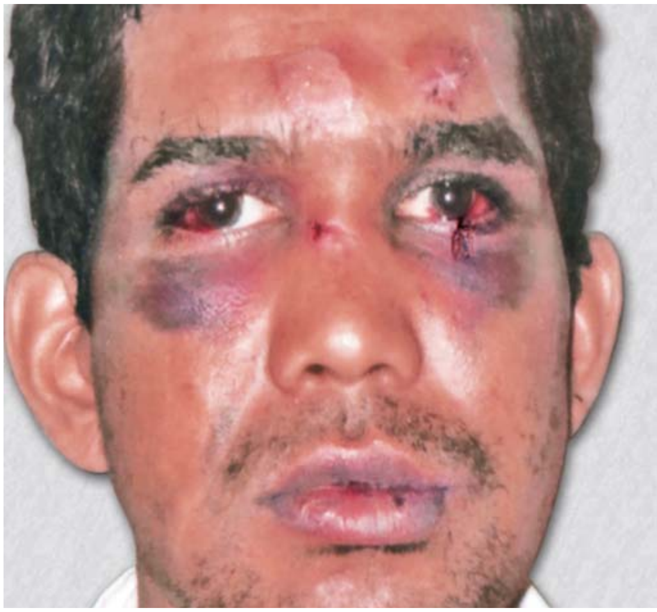

Fig. 2.24: Photo showing bilateral maxillary fracture.

Investigations

- CT scan head.
- X-ray skull.

Treatment

- It should be managed in a center for maxillofacial injuries.
- Antibiotics.
- Tracheostomy.
- Associated zygoma and nasal fractures are reduced first.
- Direct wire suturing of the zygomaticofrontal region.
- Fixation of teeth in occlusion using eyelet wires, bars or cap splints.
- Once reduced, fracture bones are immobilised using extra-oral rods called as *Mount Vernon box frame.*
- Initially intravenous fluids and blood transfusions are required. Later Ryle's tube feeding is done.
- Proper ophthalmic consultation is necessary when there are orbital injuries.

ZYGOMATIC COMPLEX FRACTURE

Classification

- *Simple fracture which is stable and undisplaced*— here fracture line passes across the infraorbital foramen downwards over anterior wall of the antrum.
- *Simple fracture which is displaced medially*. It may be associated with rotation / tilt in vertical axis, either medial tilt or lateral tilt. Infraorbital nerve may get compressed or branches of superior dental nerve may get torn.
- *Unstable fracture with rotation around horizontal axis* with medial tilt or lateral tilt.
- *Comminuted fracture* extending into the floor of the orbit.
- *Fracture of the zygomatic arch* causes a localised depression of the arch which displaces medially and tends to impinge on the coronoid process of the mandible.
- '*Blow-out' fracture* of the orbit is due to direct blunt trauma on the eyeball causing depressed comminuted fracture of the orbital floor with herniation of the orbital fat into the antrum (Fig. 2.23C).
- *Enbloc dislocation of zygomatic bone* medially/ inferiorly/ posterolaterally.

Clinical Features

- Swelling and bruising in the cheek with subconjunctival haemorrhage.
- Flattening of the cheek prominence.
- Step in the margin of the bony orbit at the infraorbital foramen.
- Sensory loss over the supply of the branches of the superior orbital nerve— teeth on the affected area are anaesthetic on percussion.
- Sensory loss over the supply of the infraorbital nerve usually over infraorbital region, upper lip and alar region of the nose—common.
- Enophthalmos is due to herniation of the orbital fat across the fracture floor of the orbit into the antrum
- Diplopia is due to entrapment of the inferior rectus muscle preventing upward rotation of the eyeball while looking up.
- Trismus with marked restriction of the lateral movements.
- Epistaxis, lowering of pupil level.
- Infraorbital ecchymosis of the orbit is called as *Panda sign.*

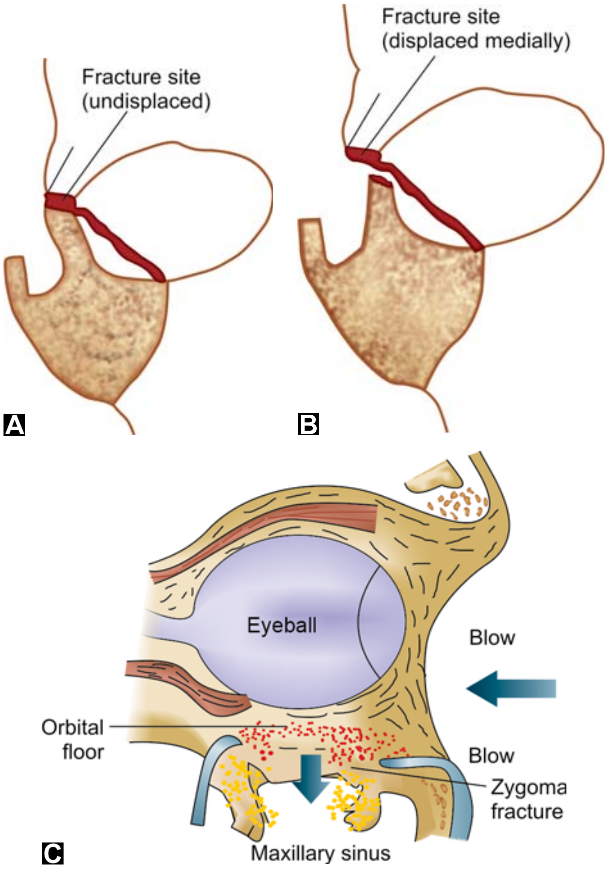

Figs 2.25A to C: (A and B) Diagrams showing different types of zygomatic fractures, (C) Blow-out fracture.

Investigations

- 30° occipitomental X-ray is used commonly but often obliquity of X-ray may be increased to 60°. In X-ray, findings observed are:
  - Fracture line near infraorbital foramen, zygomatic arch and lateral wall of the antrum.
  - Orbital floor line for fracture.
  - Opacity in the antrum due to blood.
- CT scan is done to see orbital depression and herniation of orbital fat.

Treatment

- *Every patient with zygoma fracture need not require surgical correction.*
- Need for surgery is decided based on clinical features.

Indications for surgery are:

- Infraorbital anaesthesia, trismus
- Diplopia, enophthalmos
- Flattening of the cheek
- Undisplaced fracture with infraorbital anaesthesia

Surgical Approaches

1. *Closed reduction of the zygomatic arch through Gillies temporal approach*:

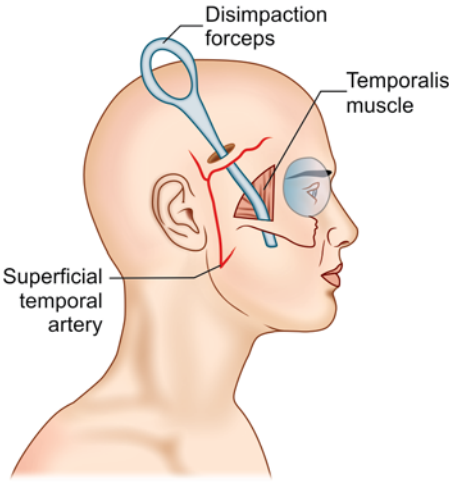

Fig. 2.26: Technique showing method of temporal reduction using disimpaction forceps.

An oblique skin incision of 2 cm length temporal is made between the two branches of the superficial temporal artery. Care is taken to avoid injury to artery. Whitish glistening temporal fascia is identified and incised. Zygoma elevator is introduced beneath the zygoma and fracture fragments are manipulated and elevated into proper position. An audible snap is heard when fracture gets reduced into position. Reduced, disimpacted fracture is always stable. Additional corrections in other parts can be done by different leverage actions of the elevator. Orbital rim, zygomatic arch are palpated for completion of correction. Skin wound is closed with sutures.

Elevators used are Bristow's periosteal elevator, Rowe's zygomatic elevator.

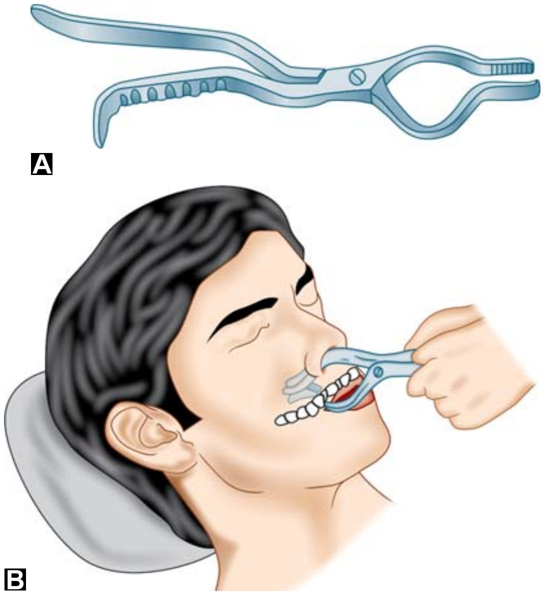

Figs 2.27A and B: Disimpaction of maxilla using Rowe's disimpaction forceps by downward leverage action.

2. *Internal fixation by open reduction and fixation is needed*

- When fracture is unstable or
- Comminuted or
- Zygoma fracture with middle third fractures.

By proper incisions infraorbital and zygomaticofrontal fracture sites are exposed; after open reduction they are fixed using wires/plates and screws.

3. *Exploration of the orbital floor is necessary whenever there is*

- Comminuted fracture in orbital floor.
- Orbital fat herniation.
- Diplopia with entrapment of the inferior rectus muscle.

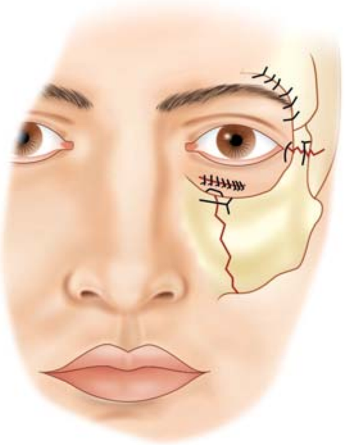

Fig. 2.28: Fracture zygoma showing open reduction and fixation using wires directly. Two types of incisions are shown depending on the site of the fracture.

FRACTURE OF THE MANDIBLE

Types

I. At the *neck of the condyle* (35%), as it is the weakest point. The condyle is displaced in front and medially often with dislocation. Painful jaw movement is the clinical features. It may be unilateral or bilateral.
II. At the *angle of the mandible*: If fracture is upwards and inwards, it is impacted and undisplaced. So it is a favourable fracture. If fracture is downwards and outwards, it gets displaced and so it is an unfavourable fracture. It needs open reduction using wires.
III. Fracture *near the mental foramen through the canine fossa*. This fracture causes displacement. Such bilateral fractures can cause pull on digastric and geniohyoid muscles precipitating fall of *tongue backwards which will block the airway.*

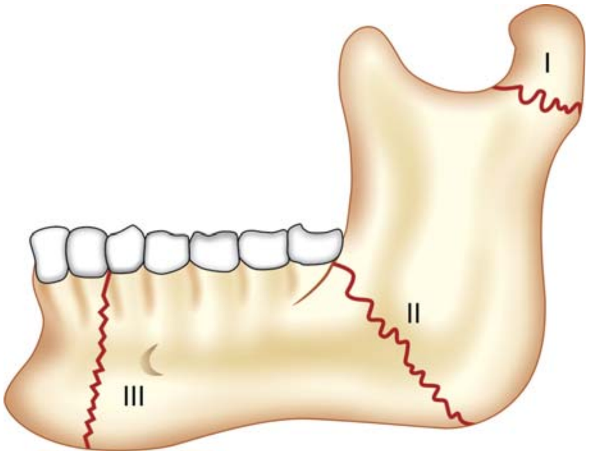

Fig. 2.29: Types of mandibular fractures.

Other Classifications

Classification of the fracture mandible

Depends on the type	***Depending on the anatomical site***
• Simple	• Dentoalveolar fracture
• Compound	• Condylar fracture
• Comminuted	• Coronoid fracture
• Pathological	• Fracture ramus of the mandible
• Green stick fracture in children	• Fracture angle of the mandible
	• Fracture in the body of the mandible
	• Symphyseal region fracture

Guardsman fracture is direct fracture of symphysis and indirect fractures of both the condyles of the mandible. In olden days guards of the queen who are in attention position used to faint and fall forward to get these fractures.

Dentoalveolar fracture

Features	***Management***
• Horizontal fracture below the alveolar margin	• Look for other injuries in face
• Dentoalveolar segment will be freely mobile	• X-ray face to see injuries
• Tooth may get split vertically/horizontally	• Dentoalveolar segment reduction and placing jaws in central occlusion position
• Derangement in occlusion and alignment	• Stabilisation using inter-dental wires or arch bars
• Gingival laceration	• Liquid diet for 3-4 weeks
• Bleeding	
• Infection and late osteomyelitis of mandible	

Clinical Features

- Pain and tenderness in the lower jaw with bruising over the surface.
- Haematoma in the floor of the mouth is called as *Coleman's sign.*
- Difficulty in opening the mouth, speech and swallowing.
- Anaesthesia of the lower lip due to compression of inferior dental nerve.
- Deranged dental occlusion.
- Step deformity.

Investigations

- X-ray of the mandible.
- Orthopantomogram (OPG).

Treatment

- Antibiotics to prevent formation of osteomyelitis of the mandible.
- Open fixation of the fracture segments using silver wires for 4-6 weeks.
- Fixation by:
  - Interdental wiring.
  - Using arch bars.
  - Silver alloy or plastic caps.
- Only fluid diet for 6 weeks.
- Irrigation wash to the oral cavity to maintain the hygiene.

Complications of fracture mandible

- Obstruction of the airway
- Osteomyelitis of the mandible
- Trismus
- Speech disturbances

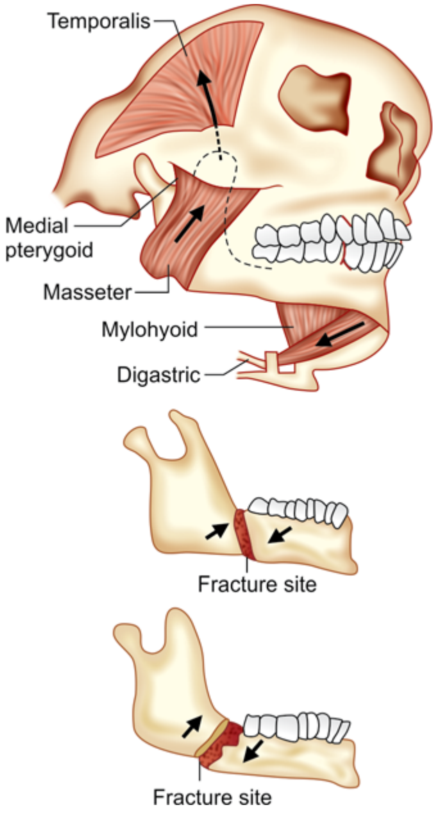
Temporalis
Medial pterygoid
Masseter
Mylohyoid
Digastric
Fracture site
Fracture site

Fig. 2.30: Muscle actions in mandible fracture causing different displacements.

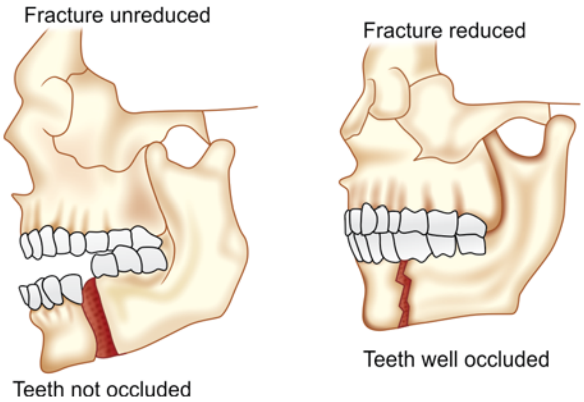
Fracture unreduced
Fracture reduced
Teeth not occluded
Teeth well occluded

Fig. 2.31: Unreduced and reduced fracture mandible.

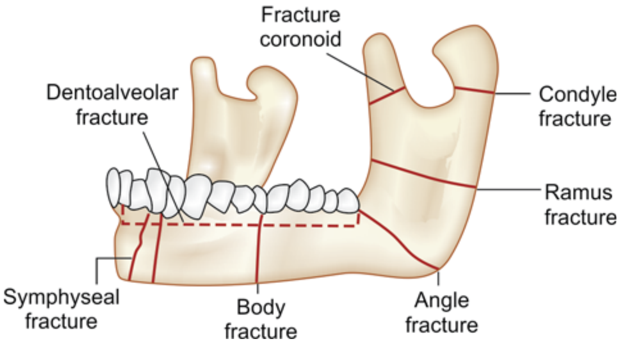
Fracture coronoid
Dentoalveolar fracture
Condyle fracture
Ramus fracture
Symphyseal fracture
Body fracture
Angle fracture

Fig. 2.32: Different sites of fracture mandible.

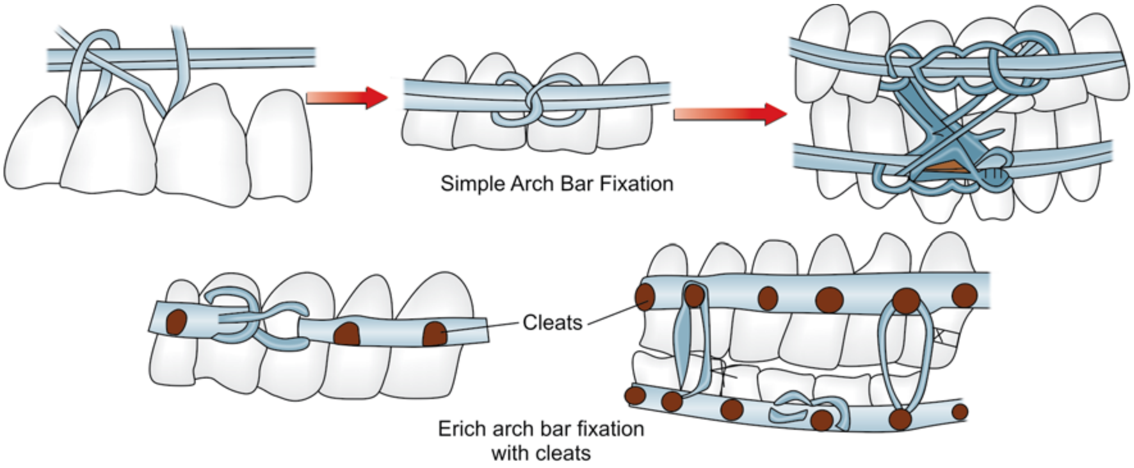
Simple Arch Bar Fixation
Cleats
Erich arch bar fixation with cleats

Fig. 2.33: Arch bar wiring. Figure shows both simple and Erich arch bar wiring with cleats to pass wire.

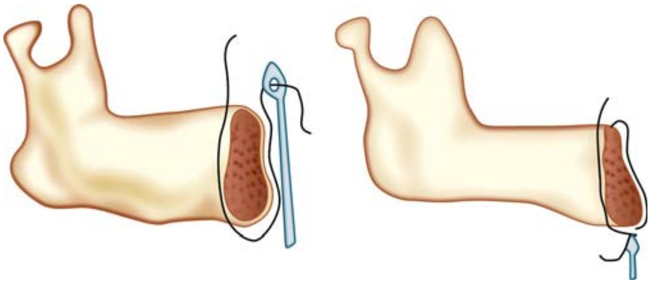

Fig. 2.34: Circummandibular wiring used in gunning splints.

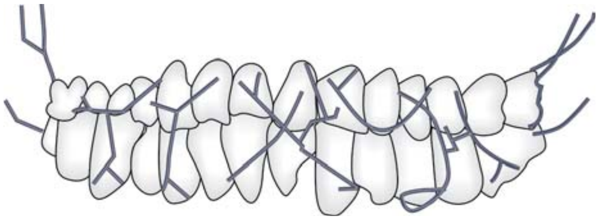

Fig. 2.35: Interdental wiring. It is commonly used and accepted method of wiring.

Rudeness is a weak imitation of strength.

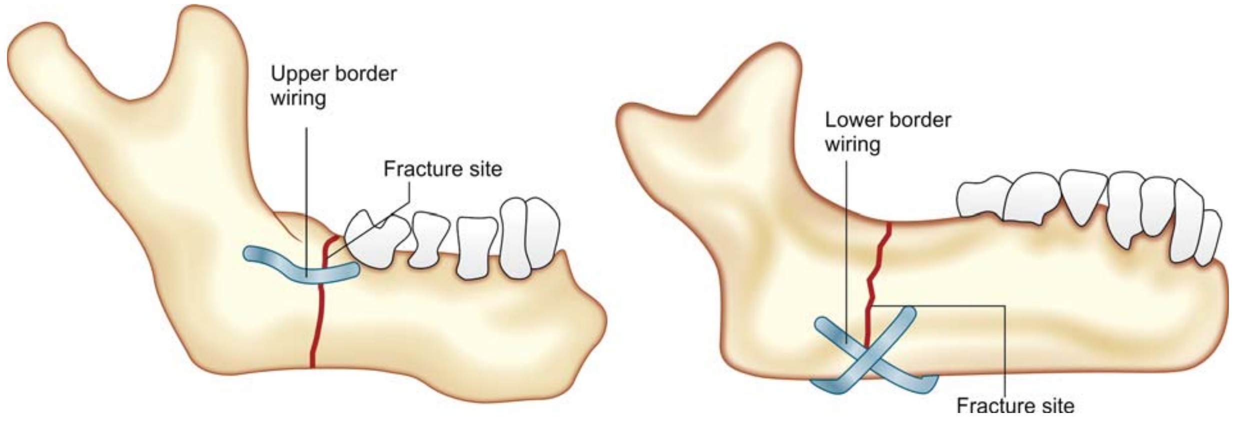
Upper border wiring
Fracture site
Lower border wiring
Fracture site

Fig. 2.36: Upper and lower border wiring. It is used to fix the mandibular fractures and is often done together with other fracture fixations in the face.

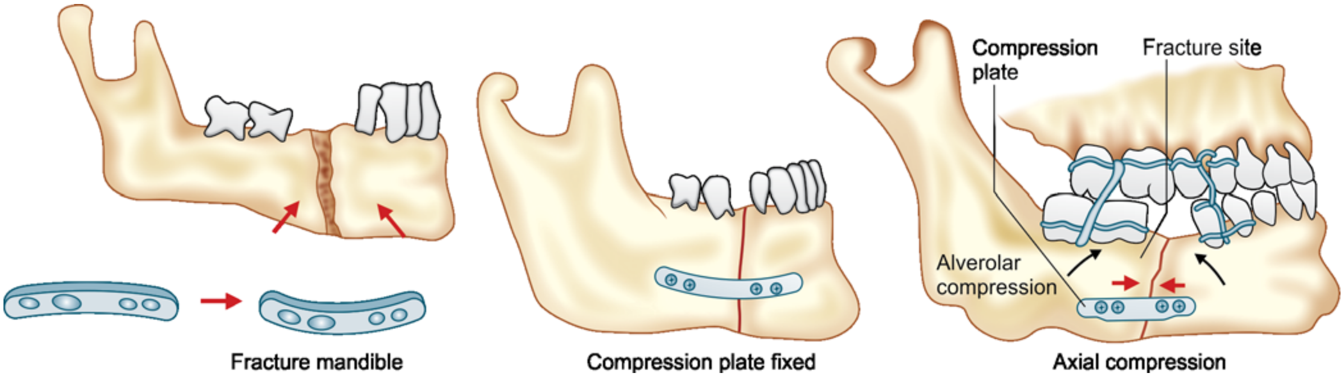
Fracture mandible
Compression plate fixed
Compression plate
Fracture site
Alverolar compression
Axial compression

Fig. 2.37: Compression plating of a mandibular fracture. Note the different methods.

DISLOCATION OF THE MANDIBLE

- It occurs at *temporomandibular joint.*
- Unilateral dislocation after trauma is common.
- Bilateral dislocation occurs during yawning and it is recurrent.

Clinical features are difficulty in opening the mouth with pain and tenderness over the joint.

Treatment

- Reduction of dislocation under general anaesthesia.
- If there is associated fracture mandible it should be dealt accordingly.

PREAURICULAR SINUS

- It is due to failure of fusion of anterior tubercles of the auricle creating a sinus.
- Often sinus opening gets sealed forming a preauricular cyst which gets infected forming an abscess.
- Sinus can get infected repeatedly, discharging pus through its opening.
- It is often multiple.

Investigations

- Sinusogram.
- Discharge study.

Differential Diagnosis

- Cold abscess.
- Sebaceous cyst.

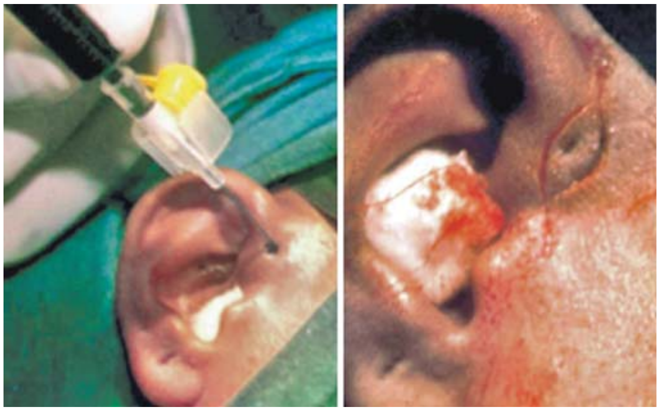

Fig. 2.38: Preauricular sinus. During excision, methylene blue is injected into the track initially and later it is excised using elliptical incision.

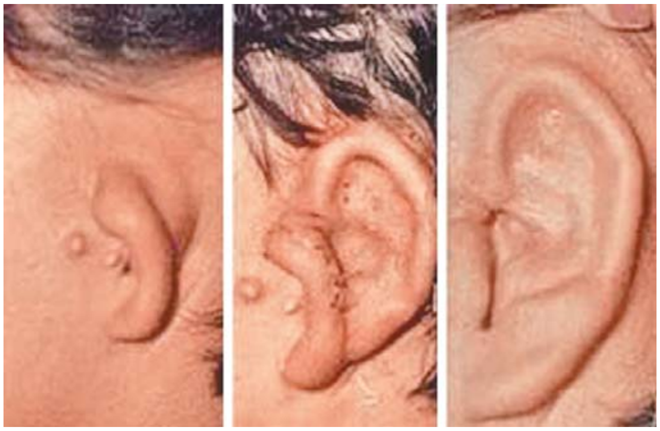

Fig. 2.39: Microtia which is corrected later (*Courtesy*: Dr Sathish Bhat, Plastic Surgeon, Mangalore).

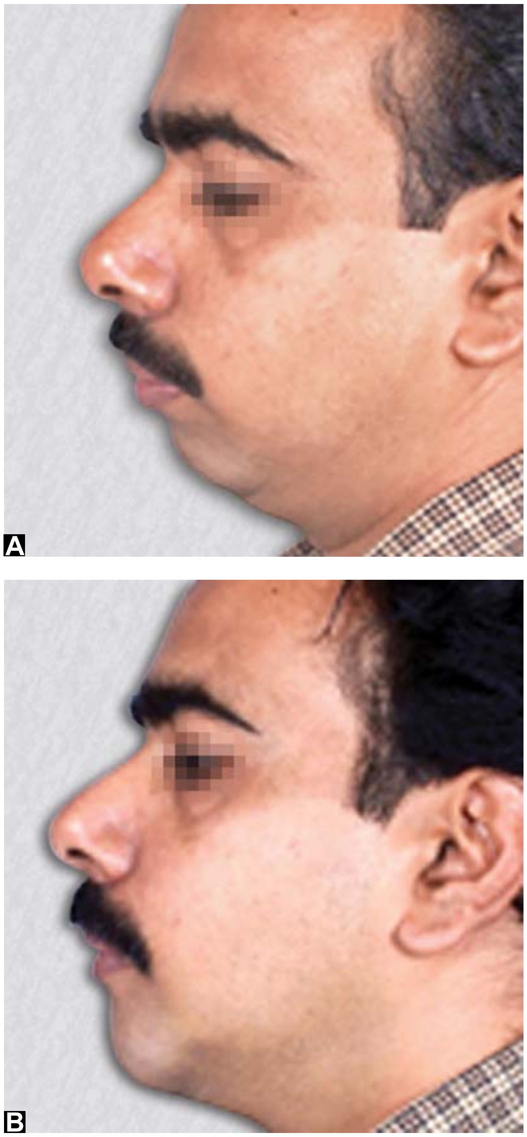

Figs 2.40A and B: Retruded chin is corrected by plastic reconstruction (*Courtesy:* Dr Sathish Bhat, Plastic Surgeon, Mangalore).

Treatment

Complete excision of the sinus with entire track.

Treacher-collins syndrome

- Mandibulofacial dysostosis
- Hypoplasia of the zygomatic bone and mandible
- Antimangoloid slant to the palpebral fissure
- Coloboma of lower eyelid
- Low ear lobule with deficient middle ears
- Familial—3rd arch syndrome (Mandibulofacial dysostosis)

JAW TUMOURS

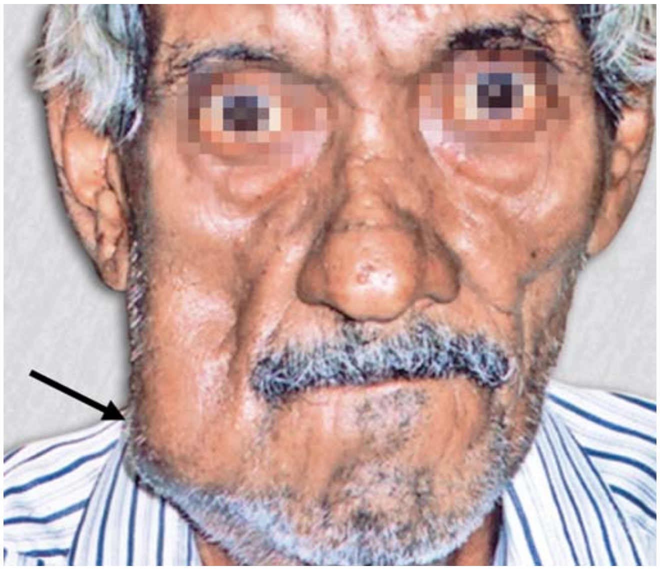

Fig. 2.41: Jaw tumour (Right sided).

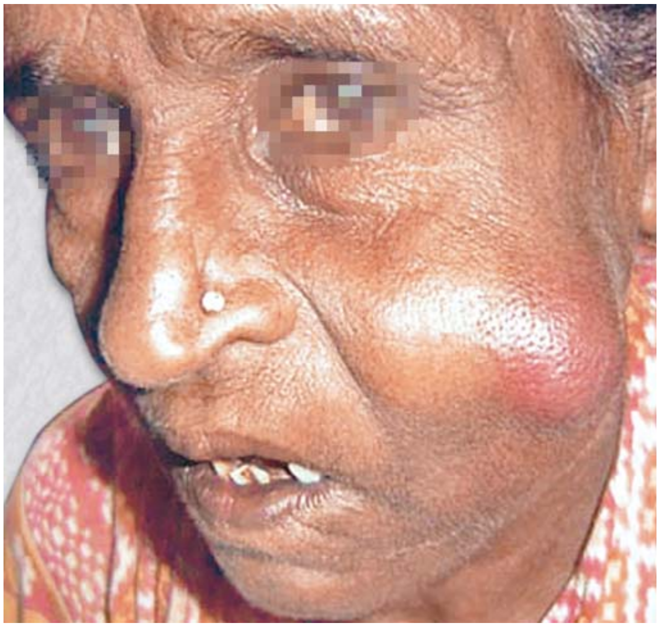

Fig. 2.42: Upper jaw tumour (Left sided).

Classification

I. ***Swelling arising from the gums*** (*Epulis*):
  - Congenital epulis.
  - Fibrous epulis.
  - Pregnancy epulis.
  - Giant cell epulis.
  - Myelomatous epulis.
  - Sarcomatous epulis.
  - Carcinomatous epulis.

II. ***Swelling arising from the dental epithelium*** (*Odontomes*):
  - Ameloblastoma.
  - Compound odontome.
  - Enameloma.
  - Cementoma.
  - Dentinoma.
  - Odontogenic fibroma and myxoma.
  - Radicular odontome.
  - Composite odontome.

Cysts arising in relation to dental epithelium:
- Dental cyst.
- Dentigerous cyst.

III. ***Swelling arising from the mandible or maxilla:***
  - Osteoma and osteoblastoma.
  - Torus palatinus and mandibularis.
  - Fibrous dysplasia.
  - Osteoclastoma (Common in mandible).
  - Osteosarcoma.
  - Secondaries.
  - Giant cell reparative granuloma.

IV. ***Surface tumours:***
  - Tumours from the surface which extend into the jaw.
  - Ossifying fibroma.
  - Osteofibrosis of maxilla.
  - Ivory osteoma of jaw.
  - Leontiasis ossea (diffuse osteitis).
  - Carcinoma extending into the jaw.

EPULIS (Greek—means upon gum)

Swelling arising from the mucoperiosteum of gums.

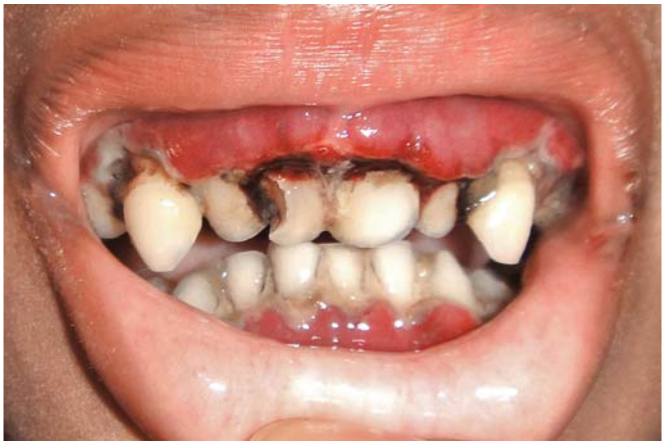

Fig. 2.43: Epulis.

Congenital Epulis

- It is a benign condition seen in a *newborn* arising from gum pads.
- It is a variant of granular cell myoblastoma originating from gums.
- It is more common in *girls*. It is more common in *upper jaw*, common in canine or premolar area.
- It is not a malignant condition.

Clinical Feature

Well localised swelling from the gum which is firm and bleeds on touch.

Treatment

Excision.

Fibrous Epulis

- It is a benign condition, can occur in any individual.
- It is red, firm/hard, sessile/pedunculated.
- It is commonest type.
- It is fibroma arising from periodontal membrane.

Clinical Features

Painless, well localised, hard, non-tender, grey pink swelling in the gum which bleeds on touch.

Differential Diagnosis

Squamous cell carcinoma from the gum.

Investigations

- X-ray jaw.
- Orthopantomogram
- Biopsy from the lesion.

Treatment

Excision with extraction of the adjacent tooth. Recurrence can if root is not removed properly.

Pregnancy Epulis

- It occurs in *pregnant women* due to *inflammatory gingivitis.*
- Usually during 3rd month of pregnancy.
- *Clinically* it resembles fibrous epulis or pyogenic granuloma.
- It usually resolves after delivery. Otherwise it should be excised.

Myelomatous Epulis

- It is seen in *leukaemic* patients.

Investigated for leukaemia by peripheral smear, bone marrow biopsy.

Treatment: For leukaemia.

Granulomatous Epulis

It is a mass of granulation tissue in the gum around a caries tooth. It forms a localised soft/firm/fleshy mass in the gum which bleeds on touch.

Giant Cell Epulis

Osteoclastoma causing ulceration and haemorrhage of gum.

Carcinomatous Epulis

Squamous cell carcinoma of the alveolus and gum presenting as localised, hard, indurated swelling with ulceration.

Fibrosarcomatous Epulis

Fibrosarcoma arising from fibrous tissue of the gum.

Epulis

- ❖ Congenital
- ❖ Fibrous-commonest
- ❖ Granulomatous
- ❖ Pregnancy
- ❖ Carcinomatous
- ❖ Myelomatous
- ❖ Fibrosarcomatous

Odontogenic tumours

Epithelial tumours

- ❖ Ameloblastoma
- ❖ Calcifying odontogenic tumour
- ❖ Odontogenic adenomatoid tumour
- ❖ Composite odontoma, which may be either complex or compound. It is odontogenic hamartoma contains all 4 layers, dentin, enamel, cementum and pulp

Mesodermal tumours

- ❖ Odontogenic fibroma, myxoma
- ❖ Cementoma, dentinoma

Malignant odontogenic tumours

- ❖ Malignant ameloblastoma
- ❖ Fibrosarcoma

AMELOBLASTOMA (Adamantinoma, Eve's Disease, Multilocular Cystic Disease of the Jaw)

- ♦ It arises from the dental epithelium probably from the *enamel/dental* lamina.
- ♦ It occurs commonly in *mandible* or maxilla.
- ♦ Occasionally it is seen in the base of the skull in relation to Rathke's pouch or in tibia.
- ♦ Histologically it is a variant of *basal cell carcinoma*.
- ♦ It is a *locally malignant tumour.*
- ♦ It neither spreads through lymph node nor through blood. Hence it is *curable.*

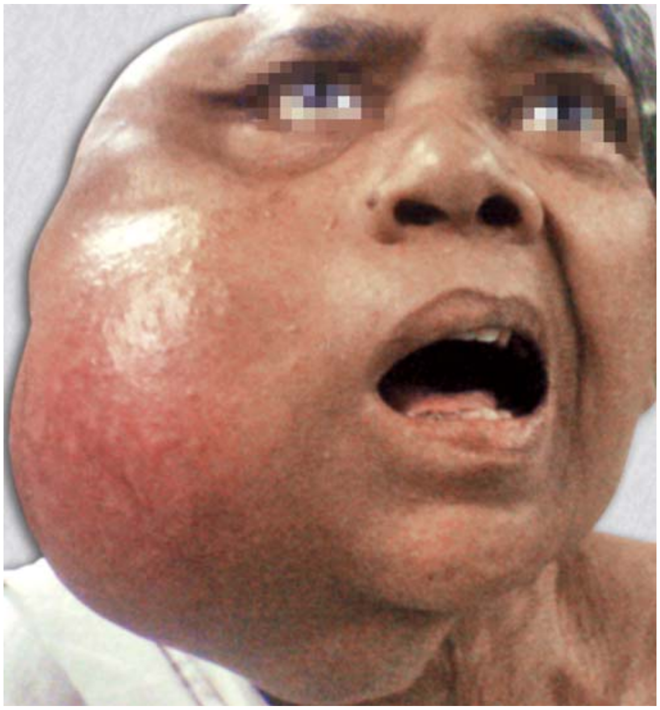

Fig. 2.44: Large jaw tumour—could be adamantinoma/ osteoclastoma.

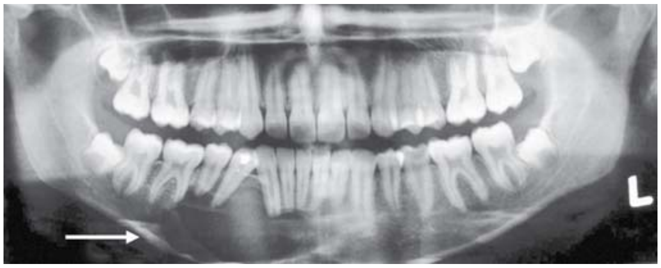

Fig. 2.45: Orthopantomogram (OPG) X-ray showing unilocular odontogenic tumour in lower jaw.

- ♦ It is usually unilateral.
- ♦ It can occur in a pre-existing dentigerous cyst.
- ♦ It is usually multilocular cystic spaces but can be unilocular also.
- ♦ Histologically cords of odontogenic epithelium, connective tissue, stellate reticulum like cells with columnar ameloblast like cells.

Clinical Features

- ♦ Swelling in the jaw usually in the mandible near the angle which attains a large size, extending to vertical ramus—*Eggshell crackling.*
- ♦ It is a gradually progressive, painless swelling which is smooth and hard with intact inner table (enlarges externally).
- ♦ Lymph nodes are not enlarged.
- ♦ Outer table expansion.
- ♦ It is common in males, common in 4th to 5th decades.

Differential Diagnosis

- Osteoclastoma of the mandible: Here inner table is not intact.
- Dentigerous cyst.
- Dental abscess.
- *Giant cell reparative granuloma (Jaffe's tumour)* – It is a swelling which occurs due to haemorrhage within the bone marrow. It contains vascular stroma, collagen and connective tissue cells. It is common in women. It causes painless enlargement of jaw. It can be treated by calcitonin (100 units/0.5 mg subcutaneously daily for 12 months) or surgical curettage.

Investigations

- Orthopantomogram (OPG) shows multiloculated lesion—*Honeycomb* appearance.
- Biopsy from the swelling.
- CT scan of the region.

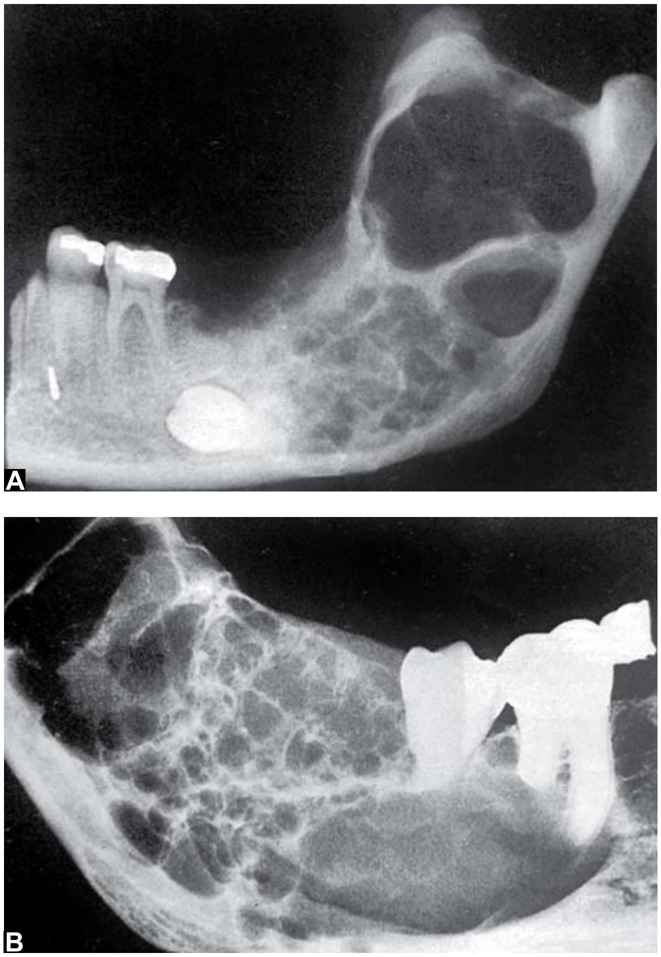

Figs 2.46A and B: X-ray (two different X-rays) showing typical honeycomb/multiloculated features of adamantinoma (*Courtesy:* Dr Veena Jagadish, MDS).

Treatment

- Segmental resection of the mandible. OR
- *Hemimandibulectomy with reconstruction of the mandible.*

Note:

Curettage and bone grafting should not be done. It is a curable condition.

Recurrent adamantinoma can spread through blood into lungs.

DENTIGEROUS CYST (Follicular Odontome)

- It is a unilocular cystic swelling arising in relation to the dental epithelium from an unerupted tooth.
- Common in lower jaw, but can also occur in upper jaw.
- It occurs over the crown of unerupted tooth. Commonly seen in relation to premolars or molars.
- It causes expansion of outer table of the mandible.

Clinical Feature

Painless swelling in the jaw which is smooth and hard.

Differential Diagnosis

- Adamantinoma.
- Dental cyst.
- Osteoclastoma.

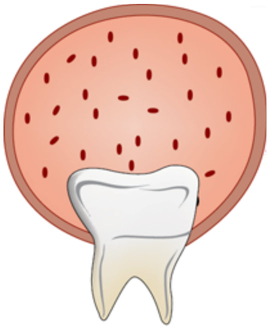

Fig. 2.47: Dentigerous cyst.

Complication

It can turn into *adamantinoma.*

Investigation

Orthopantomogram. Tooth within the cyst, which is well-defined.

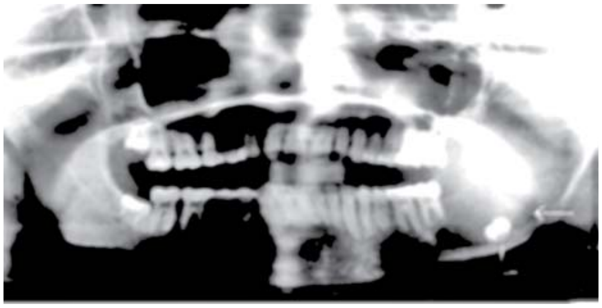

Fig. 2.48: Orthopantomogram showing dentigerous cyst.

Treatment

- If it is small, excision of the cyst is done.
- If it is large, initial marsupialisation and later excision is done.
- *Unerupted tooth should be extracted.*

DENTAL CYST (Radicular Cyst, Periapical Cyst)

- It occurs under the root of the chronically infected dead erupted tooth.
- It is lined by squamous epithelium derived from epithelial debris of *Mallassez.*

Clinical Feature

As a smooth, tender swelling in the jaw in relation to caries tooth which causes expansion of the jaw bone.

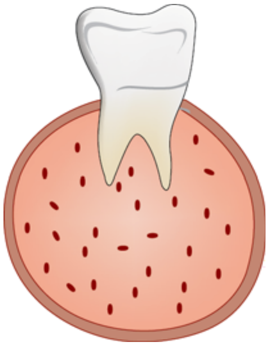

Fig. 2.49: Dental cyst.

Complication

It can cause osteomyelitis of the jaw.

Differential Diagnosis

Dentigerous cyst.

Investigation

Orthopantomogram.

Treatment

- Antibiotics.
- Drainage or excision of the cyst with extraction of the infected tooth is done.

Differences between dental cyst and dentigerous cyst

	Dental cyst	*Dentigerous cyst*
a. Site of occurrence	Erupted tooth under the root	Over the crown of an unerupted tooth
b. Infection	Common	Not common
c. Complication	Osteomyelitis	Adamantinoma
d. Treatment	Excision and extraction of tooth	Marsupialisation, excision and then extraction of tooth

Curable malignancies

- Adamantinoma
- Basal cell carcinoma
- Verrucous carcinoma
- Papillary carcinoma thyroid
- Marjolin's ulcer
- Carcinoma colon

OSTEOMYELITIS OF JAW

It is an inflammatory process in jaw; acute or chronic. It can be in the maxilla or mandible.

Causes

- Alveolar abscess leading into osteomyelitis.
- Recurrent dental infection.
- Trauma.
- After dental extraction; surgeries of the jaw.
- Postradiotherapy osteomyelitis (osteoradionecrosis).

Types

- *Acute* is common in children; maxilla or mandible may get involved; swelling, redness, fullness is the features; pus may trickle through nostril if it is in maxilla.

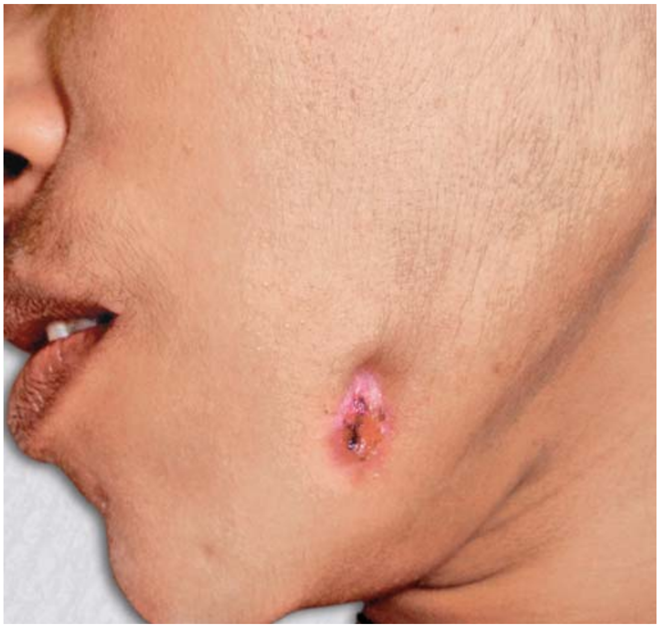

Fig. 2.50: Osteomyelitis of the mandible. Discharging sinus is obvious.

- *Subacute* type is the commonest type; common in adult; apical sepsis, endarteritis, bone necrosis is the pathology; common in mandible; rare in maxilla due to existing network vasculature which prevents endarteritis. Compression over inferior dental nerve causes numbness in chin in area of distribution of mental nerve. Pain, swelling, tenderness, irregularity, bone thickening are typical.

- *Chronic type* is also common in mandible; apical abscess, alveolar abscess, trauma, radiation, chemicals like phosphorus, tuberculosis, syphilis, actinomycosis are the causes. Pain, bone thickening, irregularity, discharging sinus, sequestrum in the discharge, discomfort are the features. Infection from lower incisor causes median mental sinus. X-ray shows features of osteomyelitis with new bone formation and sequestrum.

Management

- X-ray jaw; CT scan of jaw; discharge study; ESR are essential investigations. Biopsy from the sinus is needed often.
- It is often difficult to treat. In acute phase, antibiotic coverage, treatment of cause is done. In chronic type, sequestrectomy, mandibulectomy is needed.

Actinomycosis of Jaw

Faciocervical is the commonest type; lower jaw is commonly involved; infection begins at carious tooth; indurated gums → nodules → abscess → multiple sinuses → discharging sulphur granules with normal X-ray (*Ray fungus*). *Actinomycosis israelii* is the causative agent. It is treated by penicillins.

ALVEOLAR ABSCESS (Dental Abscess)

It is due to spread of infection from root of the tooth into the periapical tissue. Initially, it forms periapical abscess which later spreads through the cortical part of the bone into the soft tissues around forming an alveolar abscess.

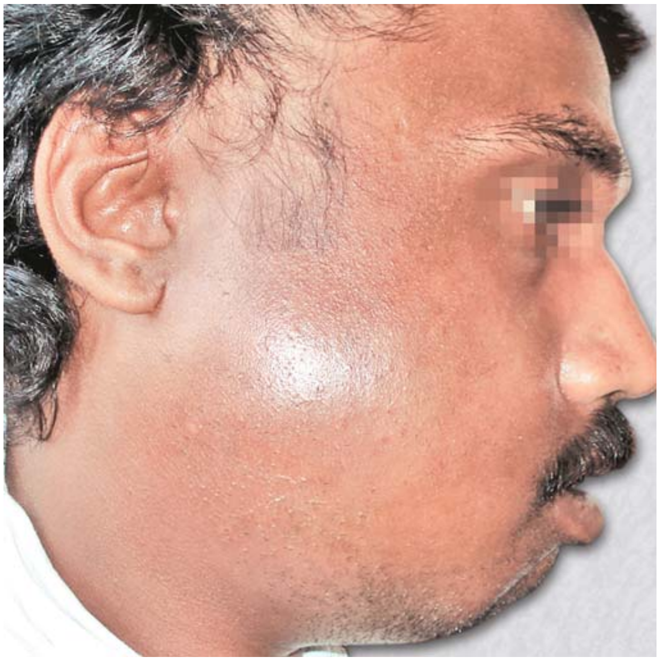

Fig. 2.51: Tooth infection causing large dental abscess. It needs antibiotics, proper drainage and tooth extraction. Patient may develop trismus/retropharyngeal infection/chronic osteomyelitis of the mandible.

Disease begins in the pulp of tooth → pulpitis → spread to root → localized osteitis → abscess formation → spread into soft tissues outside in cheek → initially diffused later localised swelling in the jaw with redness and oedema of gum. Initial dull continuous pain later becomes severe excruciating pain.

Bacteria: Staphylococci, streptococci, anaerobic-bacteria and gram-negative organisms.

Clinical Features

- Deep, throbbing pain in the jaw and adjacent oral cavity with diffuse swelling over the cheek.
- Tender soft tissue swelling in the jaw which eventually bursts spontaneously leading to sinus formation.
- Oedema, pain and tenderness in the floor of the mouth.
- Trismus and dysphagia.
- Fever and features of toxaemia.
- Tender palpable lymph nodes in the neck.

Investigations

- X-ray of the mandible or maxilla.
- Discharging pus for culture study.

Complications

- Septicaemia.
- Spread of infection into other spaces like parapharyngeal spaces; sublingual and submandibular spaces causing Ludwig's angina; oedema of epiglottis and respiratory distress; spread to pterygoid space and along pterygoid muscles through emissary vein → cavernous sinus thrombosis; upper canine tooth abscess → medial corner of eye → angular vein thrombophlebitis → cavernous sinus thrombosis; submasseteric abscess.
- Lower incisor abscess can cause abscess in the chin and later median mental sinus; chronic osteomyelitis of the jaw with discharging sinuses. Osteomyelitis is common in mandible – *horizontal process near the mentum*, presenting with pain, swelling, discharging sinuses, bone thickening, loose tooth, and trismus. Sequestrum is commonly seen. It is treated by antibiotics, sequestrectomy, mandibulectomy.

Treatment: Antibiotics, sequestrectomy, mandibulectomy.

Treatment

- Antibiotics.
- Drainage of the abscess under general anaesthesia.
- Extraction of the tooth at a later period.
- Excision of the sinus whenever required.

FIBROUS DYSPLASIA OF BONE/JAW

It is benign self limiting non-capsulated lesion of bone wherein normal bony architecture is replaced by collagen, fibroblasts, osteoid and calcified tissue. It is often classified as benign tumour with localized developmental arrest, with bone being not differentiated into a mature bone tissue.

It is seen in childhood and adolescents.

Types

It may be *polyostotic or monostotic*. Condition can occur in long bones, ribs and jaw bones, either mandible or maxilla. Disease is either metaphyseal or in the shaft, never in epiphysis.

1. Monostotic (70%)

- It is equal in both sexes. It occurs in children and adolescents; stops once growth plate is closed.
- Femur is the commonest bone involved; tibia, ribs, jaw bones, skull and humerus can get involved.
- It can present as asymptomatic diffuse hard bony swelling or can be painful due to fracture. Discrepancies of the part with asymmetry are common.
- Monostotic will not turn into polyostotic type.
- Monostotic will not turn into sarcoma.

2. Polyostotic Fibrous Dysplasia (27%) without Endocrine Dysfunction

- It begins in earlier age group than monostotic.
- It is common in femur, skull, tibia, humerus, ribs, fibula, radius, ulna, mandible and vertebrae. Craniofacial bones are involved in more than 50% of patients.
- It may continue to grow in adulthood (progressive).
- There is no evidence of hyperparathyroidism. It should be differentiated from primary hyperparathyroidism of bone.
- Involvement of shoulder and pelvis causes severe deformity.
- Severe involvement of femur causes '*shepherd crook*' deformity.
- Recurrent spontaneous fractures are common.
- Polyostotic occasionally turns into sarcoma.

3. Polyostotic Fibrous Dysplasia with Endocrinopathies (3%)

- *Polyostotic fibrous dysplasia with skin pigmentation* (*Café au lait*, on same side of the disease in neck, chest, back, shoulder, pelvis, larger) with sexual precocity in females (McCune Albright's syndrome); often with hyperthyroidism, growth hormone secreting pituitary adenoma and primary adrenal hyperplasia is 3% common.
- It is due to mutation of guanyl nucleotide binding protein gene (GNAS gene).

- Fibrous dysplasia is most common in *femur—shepherd crook deformity; metaphyseal*
- In the jaw, mandible is the common site, vertical ramus, outer table expansion
- Monostotic is more common
- Polyostotic is more problematic—discrepancies, pathological fracture, sarcoma changes
- Monostotic ceases with cessation of growth
- Surgery should never be done during growing period

Fibrous Dysplasia of Jaw

- In the jaw, it can occur in maxilla or mandible; but mandible is more common site.
- It presents with diffuse swelling of vertical ramus of the mandible or maxilla. Gritty white, hard cartilages with cysts are the pathology. Diffuse hard, painless swelling which causes asymmetry is the usual presentation. It progresses with the growth of the bone.
- It is commonly monostotic but can be polyostotic. Monostotic ceases once bone develops completely. Polyostotic may continue to grow.
- Teeth are normal.
- Expansion is towards outer cortex of the mandible.
- Polyostotic occasionally turns into sarcoma (but not monostotic).

Complications of Fibrous Dysplasia

- Deformity and cosmetic problems.
- Pathological fractures.
- Sarcomatous transformation in polyostotic type only.

Differential Diagnosis

- Osteoclastoma, adamantinoma.
- Osteitis fibrosa cystica of primary hyperparathyroidism.

Investigations

- X-ray is diagnostic showing ground glass/smoke screen appearance.
- Serum alkaline phosphatase may be slightly elevated.
- Biopsy may be needed to confirm the condition and to rule out other conditions.
- Parathormone assay, serum calcium estimation in suspected parathyroid pathology.

Treatment

It should not be operated during growing period as if intervened there may be chances that it may turn into osteosar-

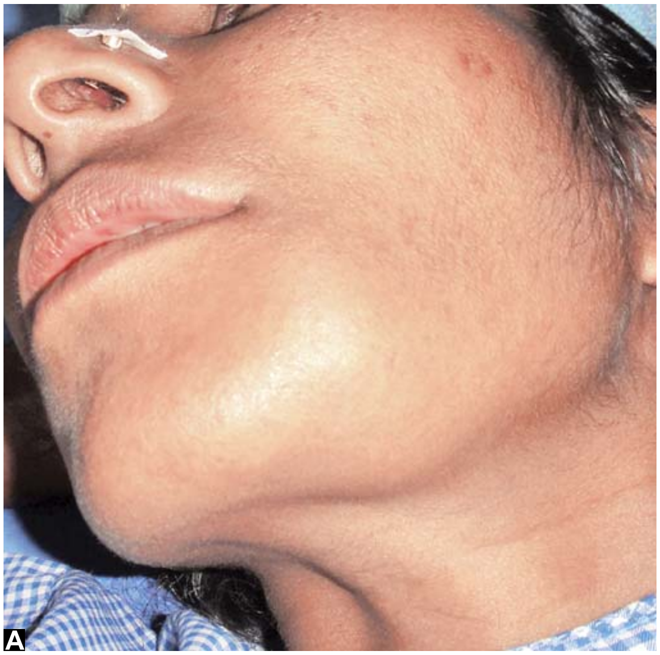

Fibrous dysplasia mandible

Fig. 2.52A

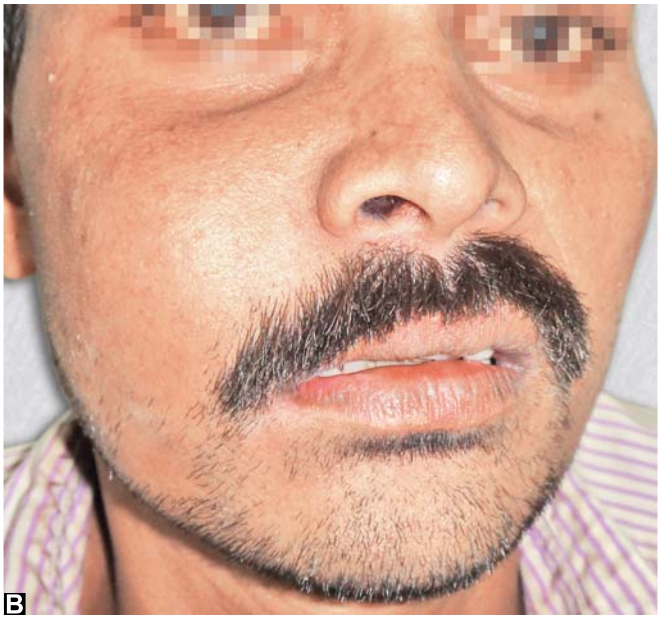

Fibrous dysplasia maxilla
Fig. 2.52B

Figs 2.52A and B: Fibrous dysplasia of mandible and maxilla in two different patients. Mandible is common site in jaw. Overall femur is the commonest site.

coma. As it is a self limiting disease it can be left alone once the growth stops or can be corrected by restorative excision to maintain facial contour. Thorough curettage with grafting of cancellous bone may be done. Bisphosphonates are often used to relieve pain.

CHERUBISM (Cherub—Angelic Being)

It is an autosomal dominant condition that occurs in first year of life.

Pathology

- Giant cell granuloma with fibrous tissues in the jaw.
- It is commonly *bilateral.*
- Commonly seen in the *angles of the mandible* and also in maxilla.
- It is familial fibrous dysplasia of jaw commonly involving both halves of the mandible with bulging outwards near the angle of the jaw causing ‘*winged face*’ appearance of angelic babies.

Clinical Features

- Diffuse enlargement of maxilla and both sides of the mandible.
- Bulging of the cheek causes pull of the lower eyelid. Hence, child appears like, as if looking upwards.
- Interference with the development and eruption of the teeth.

Treatment

It is a self-limiting disease. Often requires dental care and treatment for proper dentition.

Pierre-Robin syndrome

- Congenital condition
- Cleft palate alone
- Mandibular hypoplasia
- Cyanotic episodes
- Deficiency in transforming growth factor
- Defective sucking and tongue falling backwards in infants
- Cryptorchidism

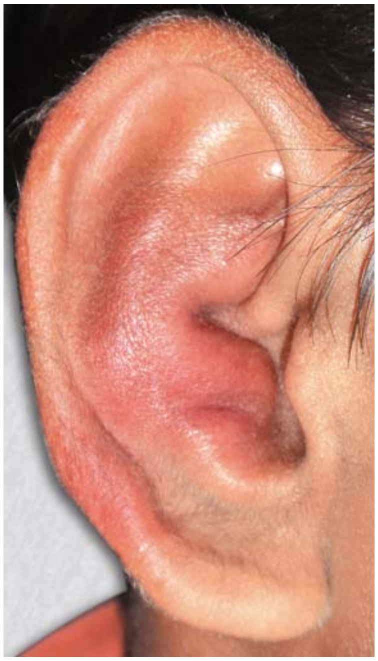

Fig. 2.53: Earlobe infection.

Chapter 3
Oral Cavity

CHAPTER OUTLINE

- Ranula
- Sublingual Dermoids
- Stomatitis
- Cancrum Oris
- Syphilitic Lesions of Oral Cavity
- Leukoplakia
- Erythroplakia
- Oral Submucosal Fibrosis
- Premalignant Conditions of Oral Cavity
- Oral and Upper Aerodigestive Cancers
- Cheek
- Carcinoma Cheek
- Lip
- Neoplasm of Lip
- Carcinoma Lip
- Tongue
- Tongue Ulcers
- Benign Tumours of Tongue
- Tongue Fissure
- Glossitis
- Tongue Tie
- Carcinoma Tongue
- Carcinoma of Posterior One-Third/Base of the Tongue
- Nasopharyngeal Carcinoma
- Maxillary Tumours
- Malignant Tumours of Tonsil
- Carcinoma Hard Palate
- Laryngeal Tumours
- Malignant Tumours of Larynx
- Trismus

RANULA

(*Rana* = *Frog*, Ranula looks like belly of frog, hence the name—Latin).

- Ranula is an *extravasation cyst arising from sublingual gland* or *mucous glands of Nuhn* or *glands of Blandin* in the floor of the mouth. Occasionally it can occur in submandibular salivary gland also.
- Initially there is blockage of the duct (of sublingual gland) causing retention cyst, which causes rupture of the acini due to increased pressure leading into extravasation cyst.

Clinical Features

- Presents as a bluish smooth, soft, fluctuant, brilliantly transilluminant swelling in the lateral aspect of the floor of the mouth.
- It often extends into the submandibular region through the deeper part of the posterior margin of mylohyoid muscle and is called as *plunging ranula*. It is intraoral ranula with cervical extension. It is cross fluctuant across mylohyoid. It can arise from both submandibular and sublingual salivary glands as a mucus retention cyst initially, which reaches neck by passing across the mylohyoid muscle presenting as soft, fluctuant, nontender, dumbbell shaped swelling in the submandibular region. It is bidigitally palpable. US and/or MRI is diagnostic. It is treated by surgical excision through neck approach along with excision of submandibular and sublingual salivary glands. Small plunging ranula is often excised per orally along with excision of sublingual salivary gland.
- Ranula has a delicate fibrous capsule and is lined by a layer of macrophages. It contains clear fluid.
- It may damage the Wharton's duct.
- It may rupture; it may get infected.
- It may interfere with speech and swallowing occasionally.

Clinical features of ranula

- Bluish swelling in the floor of the mouth
- Laterally placed
- Nontender
- Fluctuant and cross fluctuant (plunging ranula)
- Brilliantly transilluminant

Differential Diagnosis

Lymph cyst.
Sublingual dermoid.

> *Leukoplakia tongue looks as though it has been covered with white paint that had hardened, dried and cracked.*
> **—Sir Henry T Buutlin**

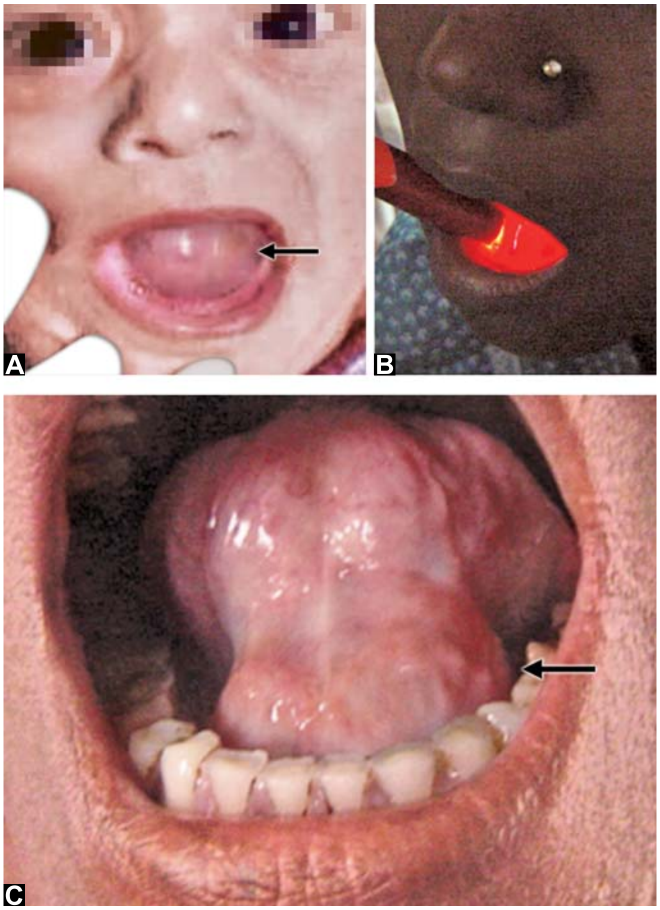

Figs 3.1A to C: Ranula. Note the bluish discolouration, brilliantly transilluminant and the location.

Treatment

- *Marsupialisation* can be done initially, and later once the wall of the ranula is thickened it is excised (*Marsupial* means pouch where baby is kept, carried and sucked on the mother's belly, like in Kangaroo).
- If ranula is small it can be excised without marsupialisation.
- Excision of sublingual salivary gland is often needed. In plunging ranula submandibular salivary gland needs to be excised.

SUBLINGUAL DERMOIDS

They are *sequestration dermoids* lined by squamous epithelium containing keratin.

It is smooth, soft, fluctuant, nontransilluminant bidigitally palpable swelling.

Types

1. ***Median sublingual dermoid:*** It is derived from epithelial cell rests at the level of fusion of *two mandibular arches*. It may be supramylohyoid or inframylohyoid. It is located between two genial muscles, in relation to mylohyoid muscle. It is a midline swelling which is smooth, soft, cystic, nontender, nontransilluminant.

   Treatment is excision through oral approach.

   Complication is abscess formation.
2. ***Lateral sublingual dermoid:*** It develops in relation to submandibular duct, lingual nerve and stylohyoid ligament. It is derived from *first branchial arch*. It forms a swelling in the lateral aspect of the floor of the mouth.

   Treatment: Small one is removed per orally. Larger one, through submandibular incision.

STOMATITIS

- It is inflammation of oral mucosa by trauma, radiotherapy, chemicals, nutritional deficiency or infection.
- *Traumatic stomatitis* may be due to dentures, teeth bite, and brushing of teeth harshly which presents as painful thin covering of furr with increased salivation. Proper mouth wash will cure the condition.
- *Aphthous stomatitis* is seen in malnutrition, debility, steroid usage. Present as multiple hyperaemic painful vesicles later forming deep round painful ulcers. It is treated with mouth wash and if needed by antibiotics. Recurrent aphthous stomatitis with ulcers is often familial, more common in women, common in lip, cheek, tongue which are very painful with more salivation. It heals spontaneously. But during active period, it interferes with speech, swallowing distressfully. It is treated by many drugs like levamisole, antibiotics, vitamin B and C, local applications of anaesthetics (xylocaine)/choline salicylate/benzalkonium chloride.
- *Candida stomatitis* (Monilial thrush) is due to fungal infection, *Candida albicans* which is seen in diabetics, individuals on steroid therapy, long-term antibiotics, patients who are bedridden, on prolonged ICU care, in infants, and debilitated patients. Initially multiple red spots which are painful appear in the tongue and buccal area which later turn into curdy white patches. Often it extends into pharynx and oesophagus causing dysphagia. It is treated with antifungal drugs like clotrimazole or flucanazole. *Rhagades* occur at corners of mouth in congenital syphilis leaving radiating scar and furrow.
- *Vincent's ulcerative stomatitis* (*Vincent's angina/trench mouth)* is due to infection by Gram –ve anaerobic bacteria *Borrelia vincentii and Fusiformis fusiformis*. It is common in adolescents and young adults below the age of 35 years. Presents with fever, excessive salivation, red swollen gums with painful ulcers covered with yellow slough (*pseudomembrane*) which can be removed like *membrane – ulcerative gingivitis*. From the gums it spreads to cheek, palate, and pharynx. Tongue involvement is uncommon. Tender neck lymph nodes are palpable. *Musty foetor oris* is typical. Edentulous patients will not develop this infection. Infection in tonsillar crypts is called as *Vincent's angina*. It is confirmed by swab culture. It is treated by antibiotics (penicillin group); peeling of membrane, mouth wash, supportive measures, vitamin B and C.
- *Nutritional stomatitis* is due to—(1) vitamin B deficiency like nicotinic acid (pellagra), riboflavin deficiency. It is

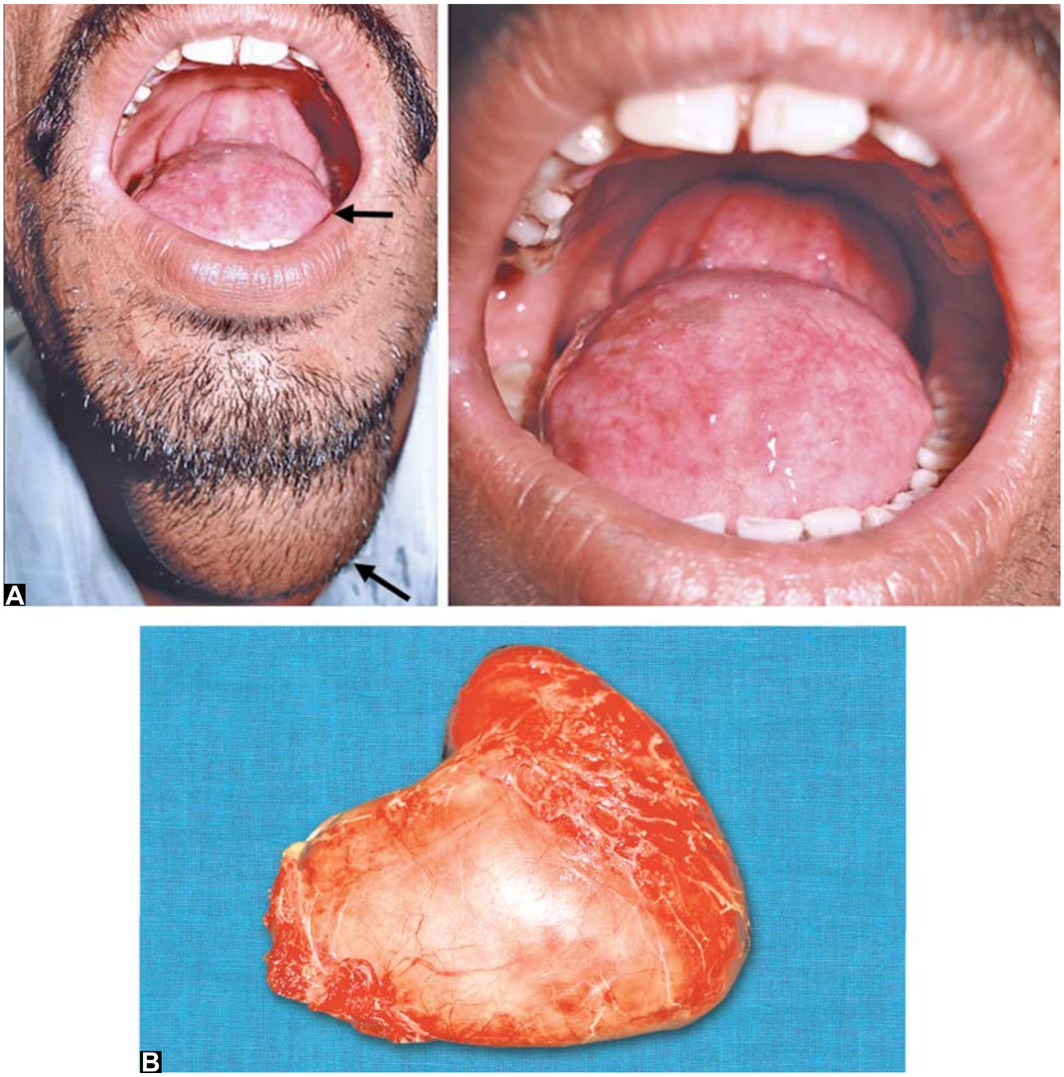

Figs 3.2A and B: Sublingual dermoid. It is usually midline. It is not transilluminant. Note the extension into the neck in submental region. Excised specimen is also shown.

common in tongue presenting as red area with atrophy of papillae. (2) Vitamin C deficiency is commonly seen as bleeding gums and loosening of teeth. (3) Iron deficiency anaemia causes superficial glossitis mainly in females.

- *Angular stomatitis* is superficial lengthy red brown fissures/ulcers in and around the angle of the mouth with cracks. Candida and streptococci infections are common. It is often called as *cheilosis/perleche*. It is treated with vitamin B, C, iron and protein supplements with adequate oral hygiene. Perleche is seen in children who suck their finger.

CANCRUM ORIS (NOMA)

- '*Noma*' means—'to devour' in Greek—'eat greedily' or 'consume destructively'.
- It is an infective gangrene, a severe form of Vincent's acute ulcerative gingivitis and stomatitis.
- Seen in poorly nourished, ill-child due to *Borrelia vincentii* and Fusiformis bacteria.

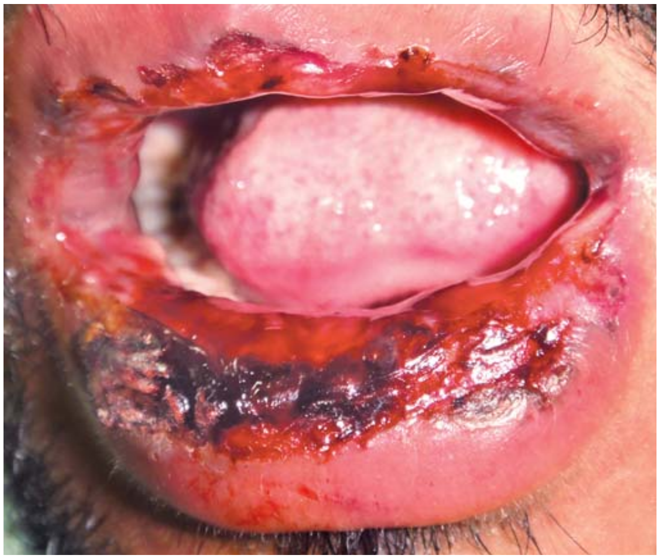

Fig. 3.3: Severe infection of lip.

> *In the case of malignancy of upper jaw, depression of the angle of the mouth on the affected side is also a typical sign.*
> **—Charles P Wilson**

- It starts in lips later extends to gums, spreads into cheek, bone, soft tissues and skin causing extensive tissue loss with severe toxaemia.
- Extensive necrosis of the mucus membrane of the oral cavity with destruction of deeper soft tissue and often bone.
- In children it may follow after an attack of measles, gastro-enteritis, typhoid, bronchopneumonia. Malnutrition is a predisposing factor.
- Excessive salivation, fetid odour with destruction, discharge and toxic features.
- *Borrelia vincentii* can be cultured.
- X-ray part shows bone destruction.
- Condition has got high mortality.
- Secondary infections may also coexist.

***Note*:**

Phagedena is a destructive ulceration with gangrene seen in cancrum oris, chancroid.

(Phagedena is destruction without proliferation but malignancy is destruction with proliferation).

Treatment

1. Systemic antibiotics, high dose penicillins, metronidazole,
2. High protein and vitamin rich diet, through nasogastric tube.
3. Wound irrigation and liberal excision of the dead tissue.
4. Blood transfusion, TPN.
5. Later patient requires flaps to cover the defect.

SYPHILITIC LESIONS OF ORAL CAVITY

Chancre in lip: It is highly contagious primary chancre, presents as painless macule later forming painful superficial ulcer. Ulcer eventually heals with a scar. It can be on both upper and lower lip—*primary syphilis*.

Mucous patches which are greyish white contagious patches seen on lip, cheek and fauces. Mucous patches fuse together to form *linear snail track ulcers* in fauces, pillars—*secondary syphilis*. *Hutchinson's contagious condyloma* in midline tongue can occur.

Gummatous painless ulcer is seen in anterior 2/3rd of tongue, palate and nasal septum (causes perforation and collapse of nasal bridge).

Syphilitic chronic glossitis is seen in *tertiary syphilis* which is a precancerous condition.

LEUKOPLAKIA

It is a white patch in the mucosa of the oral cavity that cannot be characterised clinically or pathologically to any other disease. It is a premalignant condition.

Types

1. *Homogenous*.
2. *Nodular*—more potentially malignant.
3. *Speckled*—more potentially malignant–highest.

Clinically the lesion appears as white or greyish coloured, well-localised patch in the cheek, tongue, palate or other areas of the oral cavity.

Hairy leukoplakia is dense pigmented friable lesion seen in tongue; it is more common in patients with AIDS. There are no hairs (misnomer).

Common causes

- Smoking
- Spirit
- Sepsis
- Superficial glossitis
- Syphilis
- Spices
- Sharp tooth
- Susceptibility
- Pan chewing using areca, tobacco, slaked lime
- Chronic hypertrophic candidiasis (long-standing candida infection)

- ***Incidence of*** leukoplakia in those who smoke or chew pan is 20%, whereas incidence in non-smokers is 1%.
- ***Incidence of*** its turning into malignancy is 2-4%. It increases with age, duration of the pan chewing, smoking.

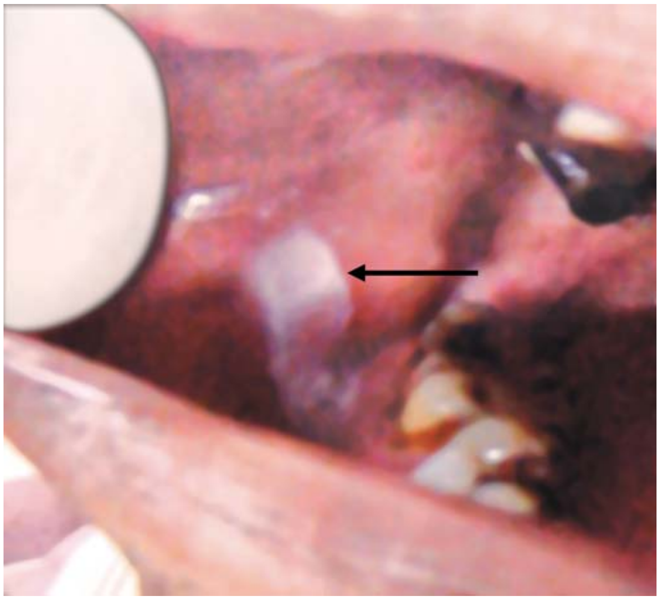

Fig. 3.4: Leukoplakia right cheek.

- Buccal mucosa and oral commissures are most common sites. Common in males (3:1).
- Leukoplakia of long duration; leukoplakia in elderly; leukoplakia in younger females; leukoplakia in floor of the mouth and tongue; leukoplakia with induration, cracks and fissures are more likely to turn into malignancy.
- ***Histology:*** *Parakeratosis with widening of rete pegs.*

Histological staging

- Acanthosis—elongation of rete pegs—smooth, white, dry patch
- Parakeratosis
- Widening of rete pegs
- Dyskeratosis—keratin cell layer formation deep to epidermis
- Dysplasia
- Carcinoma *in situ*

- Biopsy confirms the diagnosis as well as rules out the carcinoma.

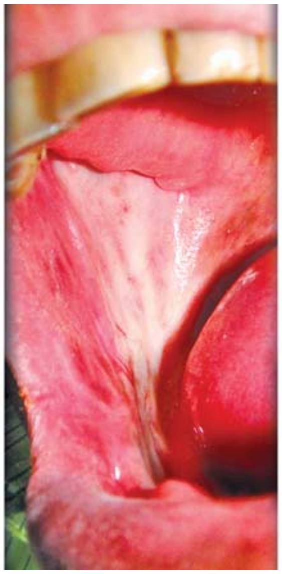

Fig. 3.5: Leukoplakia cheek.

Treatment

- Pan chewing and smoking has to be stopped.
- Excision, if required skin grafting has to be done.
- Regular follow-up is necessary.
- Isoretinoin is helpful. Beta-carotene, tocoferol are also used.
- CO_2 laser excision.

ERYTHROPLAKIA

- It is red velvety appearance of the mucosa which cannot characterise any recognised condition.
- It is 17-20 times more potentially malignant than leukoplakia.
- Histologically *parakeratosis with severe epithelial dysplasia* is the typical feature.

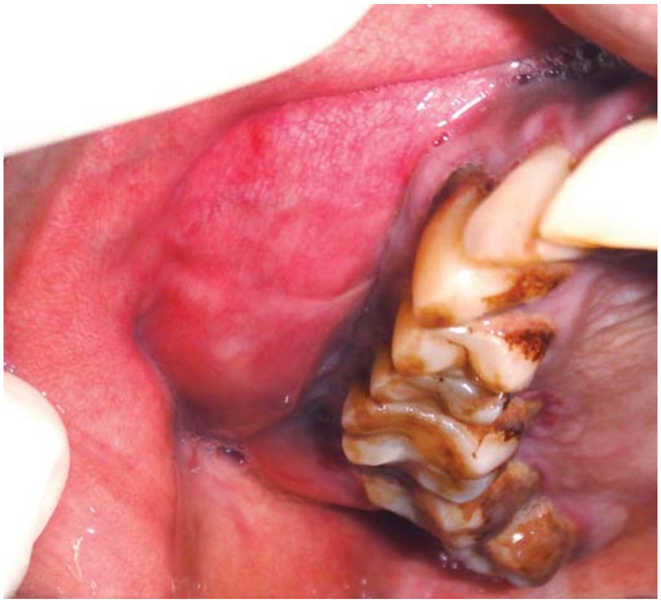

Fig. 3.6: Erythroplakia of cheek.

- Red color is due to decreased keratin causing shining and prominence of submucosal red vascularised connective tissue.
- It is equal in both sexes.
- It is common in lower alveolar mucosa, gingivobuccal sulcus and floor of the mouth.
- It can be homogenous/speckled/granular or erythroplakia interspersed with leukoplakia.
- Diagnosis is done by biopsy.
- *Treatment*: Biopsy and surgical excision.

ORAL SUBMUCOSAL FIBROSIS

- It is a progressive fibrosis deep to the mucosa of the oral cavity which causes *trismus* and *ankyloglossia*.
- The mucosa of cheek, gingivae, palate and tongue shows *a mottled/marbled pallor*.
- It is common among Asians and Indians.
- *Aetiology*: Hypersensitivity to chilli, betelnut, tobacco and vitamin deficiencies probably alter the collagen metabolism leading to juxtaepithelial fibrosis, epithelial atrophy and dysplasia.
- 4.5-7.6% of oral submucosal fibrosis turns into malignancy (Paymaster—1956 study shows 30-33% incidence—very high).
- *Treatment*: Precipitating factors has to be avoided.
- Surgical excision when required, followed by skin grafting, has to be done.

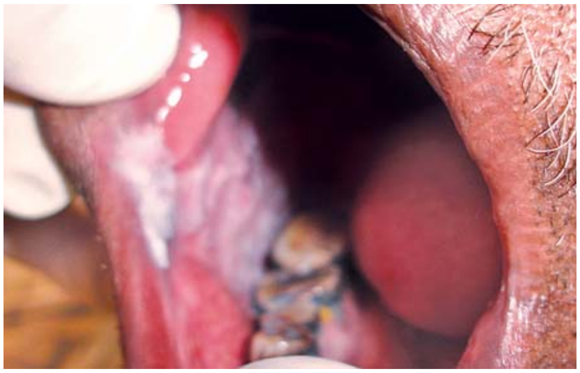

Fig. 3.7: Submucosal fibrosis of right cheek. Note the stiff fibrotic bands and scarring.

Submucosal fibrosis

- Due to:
  - Prolonged local irritation by chilies, tobacco (pan/quid), areca due to arecoline
  - Dietary causes—deficiencies of vitamin A, B complex (riboflavin) and iron
  - Localised collagen disorder
- Racial: It is common among Indians / Asians and people of Indian origin
- Prevalence in India is 5 per 1000
- Incidence is 4-7 %
- Common in middle age; equal in both sexes

- ❖ Soreness and burning in mouth which is more during meals; vesicular eruptions; trismus; difficulty in protruding the tongue. Initial red area turns into superficial ulcers which later forms *stiff, fibrotic bands* and scarring.
- ❖ Common in soft palate, faucial pillars; buccal mucosa
- ❖ Disease is progressive, even after cessation of causative factor like areca use/smoking
- ❖ It shows epithelial atrophy, hyperplasia, dysplasia and fibrosis
- ❖ *Treatment*—local injection of dexamethasone (4 mg) with hyalase (1500 units) biweekly for 10 weeks; avoidance of irritants; vitamin supplements; correction of anaemia; surgical wide excision and skin grafting

PREMALIGNANT CONDITIONS OF ORAL CAVITY

High risks—lesions with definite risk of malignant change

- ♦ Leukoplakia.
- ♦ Erythroplakia.
- ♦ Chronic hyperplastic candidiasis: It is common in commissures of the mouth and tongue. Dense plaque of leukoplakia is common with curdy white patches due to *Candida albicans* infection. It often may not respond to drugs, surgery or laser. Immunodeficiency is often associated with this. It is treated by topical or systemic antifungal drugs/surgical excision or laser therapy.

Medium risks—premalignant but not associated with higher incidence of carcinoma

- ♦ Oral submucosal fibrosis.
- ♦ Syphilitic glossitis.
- ♦ *Sideropenic dysphagia* (*Sideropenia* is iron deficiency without anaemia); or Plummer-Vinson syndrome. Sideropenia is common in Scandinavian females. It causes atrophy of epithelium and becomes potentially malignant. Proper iron therapy controls the disease and reduces the risk.

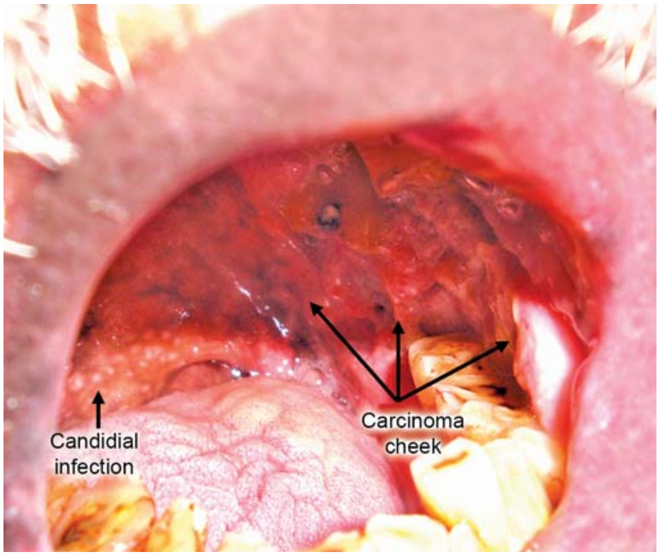

Fig. 3.8: Carcinoma cheek with candida (curdy white dots) infection.

Equivocal risk lesions

- ♦ Oral lichen planus.
- ♦ Dyskeratosis congenital—reticular atrophy, nail dystrophy, leukoplakia in oral cavity.
- ♦ Discoid lupus erythematosus.

***Note*:**

- Oropharyngeal cancer is the most common cancer—40% in Indian subcontinent. In western countries, it accounts for 4% only.
- *Risk factors*—tobacco and related products; alcohol; areca nut; human papilloma virus; Epstein Barr virus; *Paterson Kelly syndrome*; nutritional deficiency.
- Patient may develop a second primary (15%) in the oropharynx in different site at same time or within 6 months of the existing primary (*synchronous*—4% prevalence; 20% of second primaries) or after 6 months of first primary (*metachronous*—80% of second primaries). Metachronous second primary is more common than synchronous second primary and it usually occurs in 2 years.

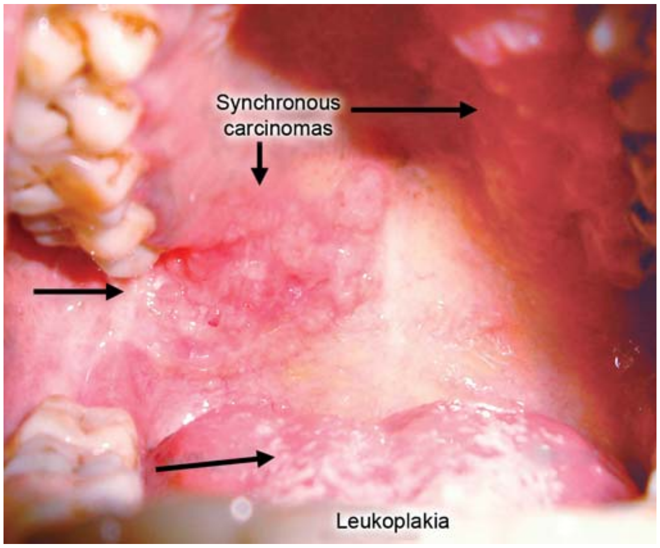

Fig. 3.9: Synchronous oral carcinoma with leukoplakia.

Premalignant conditions of oral cavity

- ❖ Leukoplakia
- ❖ Erythroplakia
- ❖ Chronic hyperplastic candidiasis
- ❖ Oral submucosal fibrosis
- ❖ Syphilitic glossitis
- ❖ Sideropenic dysphagia
- ❖ Oral lichen planus
- ❖ Discoid lupus erythematosus
- ❖ Dyskeratosis congenita

ORAL AND UPPER AERODIGESTIVE CANCERS

- ♦ It is one of the commonest cancer in Asian countries and India (40%).
- ♦ All 'S' mentioned probably are the causative agents. Smoking, quid of chewing pan are important causes. Tobacco, betel nut, alcohol, human papilloma virus (present in 80% of oral cancers; present in 40% normal individuals), EB virus, vitamin A deficiency, Plummer-Vinson syndrome,

bad dental hygiene, denture irritation—are etiologies. Risk is 8 times in tobacco chewers; 10 times with quid users; 30 times with night quid users.

- Alcohol increases the solubility of carcinogens and suppresses the DNA repair.
- Incidence of oral cancer in India is 28 per 1, 00,000 population. Commonest oral cancer in India is of buccal mucosa (more than 70%).
- In oral cavity, in West, it is common in tongue (50%), buccal mucosa (25%), floor (15%), gums and others (10%).
- Leukoplakia (*commonest*), erythroplakia, chronic hyperplastic candidiasis are *precancerous lesions*; submucosal fibrosis, syphilitic glossitis, sideropenic dysphagia are *precancerous conditions*. Oral lichen planus, discoid lupus, dyskeratosis congenita are doubtful associated lesions. Precancerous lesion is one where cancer is more likely to occur; precancerous condition is one where there is increased risk of cancer.
- Upper aerodigestive cancers include that of oral cavity, larynx and pharynx. Depending on anatomical location they present with different features other than common features—trismus, ear pain, hoarseness of voice, dysphagia, ankyloglossia.
- Usually they are locoregional disease with high affinity to involve regional lymph nodes. Distant spread is rare except in nasopharyngeal carcinoma. Tongue has highest incidence of nodal spread, then floor of the mouth, lower alveolus, cheek, upper alveolus, palate.
- Multiple synchronous (at same time, 10%) de novo sites and or metachronous (at different periods, 15%) multiple sites
- Cancers in posterior third of tongue and floor of the mouth is often missed on clinical examination—*coffin corner or sump area*.
- Primary may be very small to be detected clinically in places like fossa of Rosenmuller, pyriform fossa, nasopharynx, posterior third tongue but present clinically as hard lymph node secondaries in neck called as *secondaries with unknown primary*. Hard secondaries in neck confirmed by FNAC but all investigations including blind biopsies, CT head and neck region and endoscopies could not identify primary lesion creates a situation called as secondaries in neck nodes with an *occult primary (30%)*.
- *Trismus* (pterygoid muscle involvement), *ear pain* due to auriculotemporal nerve involvement, *eye pain* in nasopharyngeal carcinoma, *dysphagia* due to tongue involvement mainly the posterior third, *hearing loss* due to spread to eustachian tube can occur.
- Bronchopneumonia, aspiration are common problems.
- Biopsy, endoscopy, CT neck, MRI, chest X-ray are different investigations needed depending on anatomical location of lesion.
- Staging will help to plan the treatment and predict prognosis.
- *Surgical wide excision and radiotherapy are main modalities of treatment. Chemotherapy is used as an adjuvant*. Curative treatment in early growth with preservation of functions like swallowing, speech, cosmesis; but with adequate oncological clearance is the principle of surgical approach. Radiotherapy is also used as curative therapy.
- Palliative chemotherapy, radiotherapy and surgery can be done depending on location of the lesion.
- Involvement of mandible, neck nodes—number, size and fixity alters the prognosis and treatment schedule.
- Outcome also depends on the anatomical location of the malignancy. Lip carries better prognosis; tongue has poor prognosis.
- General principles used in approaching oral cancers are as follows (however it depends on grading and staging of the tumour):
  - If only primary is present which is mucosal with size less than 2 cm without nodal spread, then wide local excision with supraomohyoid block dissection of same side is done (N0); primary may also be treated with curative brachytherapy or external beam teletherapy. If nodes are histologically positive then radical neck dissection is done.
  - Larger mucosal primary with similar features are also treated similarly; but postoperative RT or/and chemotherapy is added depending on grading of the tumour.
  - In all these types of lesions, if there are positive mobile neck nodes which is confirmed by FNAC, then radical neck dissection should be done.
  - If primary lesion extends into adjacent soft tissue with mandibular involvement then mandibular resection (marginal mandibular/segmental/partial/hemimandibulectomy) is needed. Part is reconstructed using plates or bone graft taken from iliac crest or opposite 11th rib. 2.4 mm reconstruction plate with PMMF or non-vascularised bone graft (iliac crest cancellous chips) or vascularised bone graft from fibula/iliac crest/scapula are the present recommendations. Skin covering is done by split skin graft inside to mucosa or by appropriate flaps depending on the need and feasibility of donor area (PMMF/DP flap/forehead flap). Neck is addressed similarly. Postoperative EBRT and chemotherapy is needed either concurrent or sequential.
  - If primary is advanced then chemotherapy with EBRT is used. If lesion reduces in size and becomes operable it is then operated accordingly.
  - In fixed primary or secondary, RT with chemotherapy is used for palliation to relieve pain, fungation, sepsis.
- In advanced stage terminal events may be severe malnutrition, bleeding, sepsis, and bronchopneumonia.
- Posterior lesions has got poor prognosis than anterior lesions. Lip carries best. Prognosis depends on anatomical location, grading, lymph node status, soft tissue involvement and response of therapy.

Carcinoma of Gingivobuccal Complex

- It is cancer (SCC) involving *buccal mucosa and gingiva* etiology of which is keeping tobacco quid in gingivobuccal sulcus.

- It is often called as *Indian oral cancer* as it is most commonly seen in India.
- Buccal mucosa extends from upper to lower alveolus; from commissure in front to retromolar region behind.
- Features, management are same. Marginal mandibulectomy/ segmental resection is commonly needed.
- Adjuvant RT and chemotherapy is useful.

CHEEK

Anatomy of Cheek

- They are fleshy flaps on either side of the face. The demarcation between the lips and cheek is nasolabial fold.
- It is composed of skin, superficial fascia with parotid duct, buccinator muscle, submucosa with buccal glands and mucous membrane.
- Lymphatics: Submandibular and pre-auricular nodes.

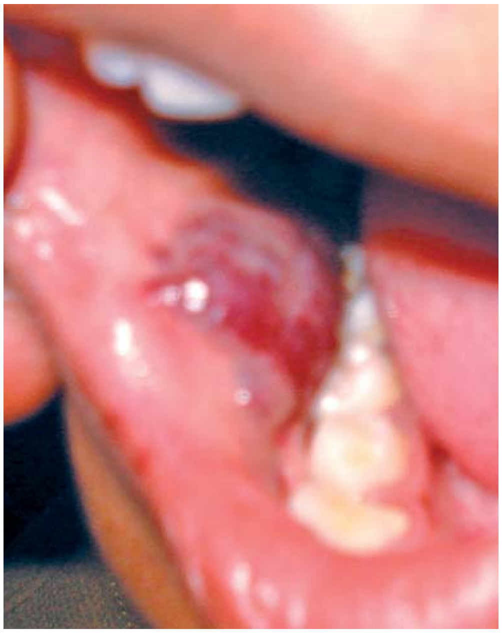

Fig. 3.10: Haemangioma cheek.

Retromolar Trigone

- It is the mucosa on the anterior surface of the ascending ramus of the mandible. It is triangular in shape with the base being superior and apex inferiorly behind the third molar tooth.
  - Malignancy in this area commonly invades the ascending ramus of the mandible, often to soft palate and tonsillar fossa. It is also said that lymphatics here communicates with pharyngeal lymphatics in this region. Lip split or mandibulotomy approach is needed in these patients.

CARCINOMA CHEEK/BUCCAL MUCOSA

Squamous cell carcinoma is the most common type of carcinoma of the cheek.

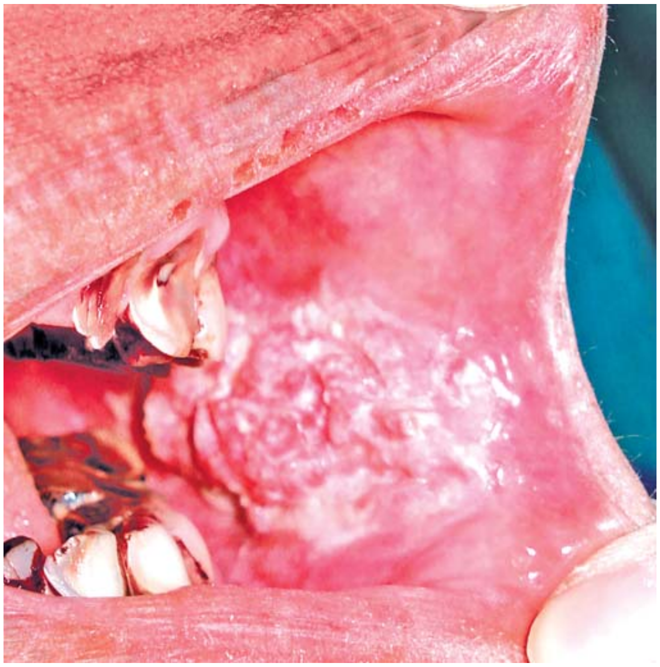

Fig. 3.11: Carcinoma buccal mucosa. It is the most common oral cancer in India.

- Occasionally it can be adenocarcinoma arising from the minor salivary glands or mucous glands. Rarely it can also be melanoma.

Malignancies of the oral cavity

- *Squamous cell carcinoma—commonest*
- Minor salivary gland tumours
- Melanomas
- Adenocarcinomas—rare
- Sarcomas—rare (Sarcomatoid)
- Basaloid SCC

Sites of carcinoma in oral cavity in order

In India	*In Western countries*
Cheek—commonest	Tongue
Tongue	Floor of the mouth
Floor of the mouth	Lip
Palate	Cheek
Lips	

Precipitating Factors

All 'S'—Smoking, spirit, syphilis, sharp tooth, sepsis, spices.

Incidence of oral cancer is six times more in smokers than nonsmokers.

Premalignant lesions and conditions

- Leukoplakia
- Erythroplakia
- Chronic hyperplastic candidiasis
- Oral submucosal fibrosis
- Sideropenic dysphagia
- Syphilitic glossitis

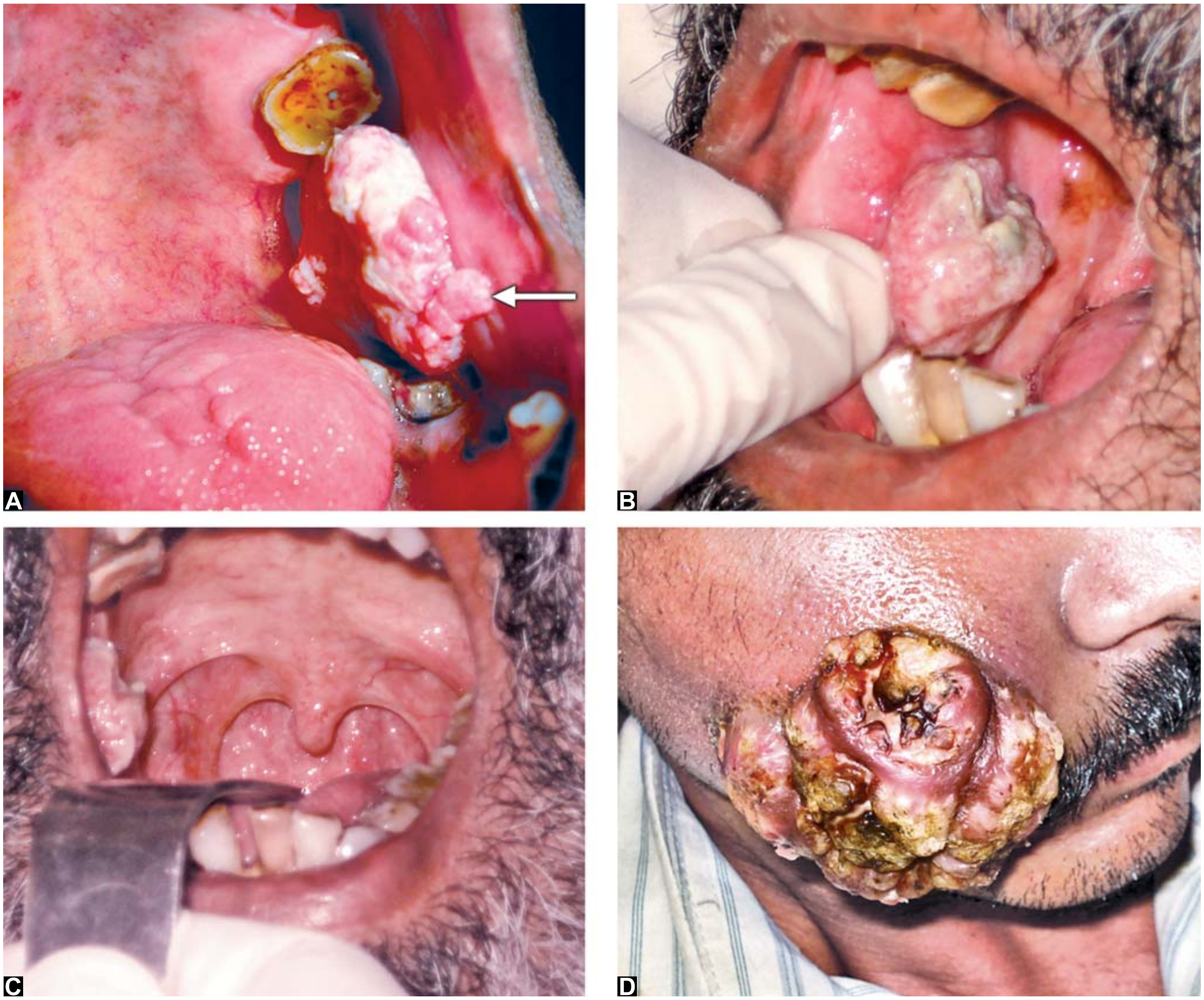

Figs 3.12A to D: Carcinoma cheek in different patients. Note the usage of tongue depressor for proper visualisation of the lesion.

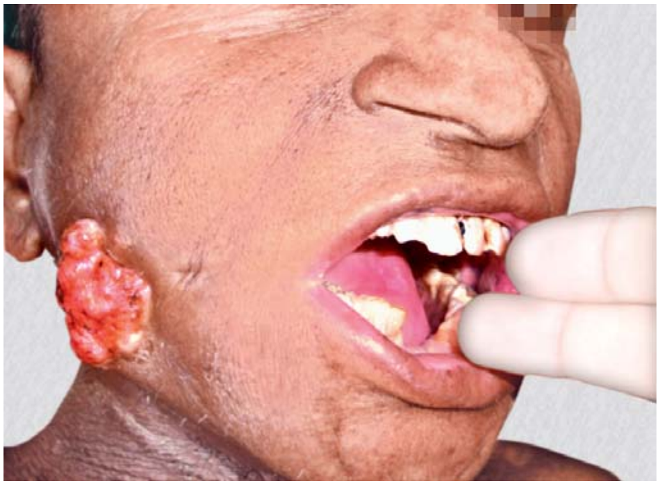

Fig. 3.13: Trismus in a patient with advanced carcinoma cheek (T4 lesion). Patient is on chemotherapy and so developed alopecia.

Betel nut chewing (*Pan*, with *pan quid* kept in cheek pouch for a long time) is an important causative factor of carcinoma cheek.

Betel/areca nut, betel leaf, slaked lime, and tobacco (often with catechu and condiments) wrapped in betel leaf is repeatedly chewed after putting into the mouth; and the *quid* formed is kept for long duration in gingivolabial sulcus; which is said to be highly carcinogenic. Tobacco is the main carcinogenic component, followed by arecoline (stimulate collagen synthesis and fibroblast proliferation) and tannin (stabilizes collagen) alkaloids of areca.

Types

- Ulcerative
- Proliferative (exophytic)
- Verrucous

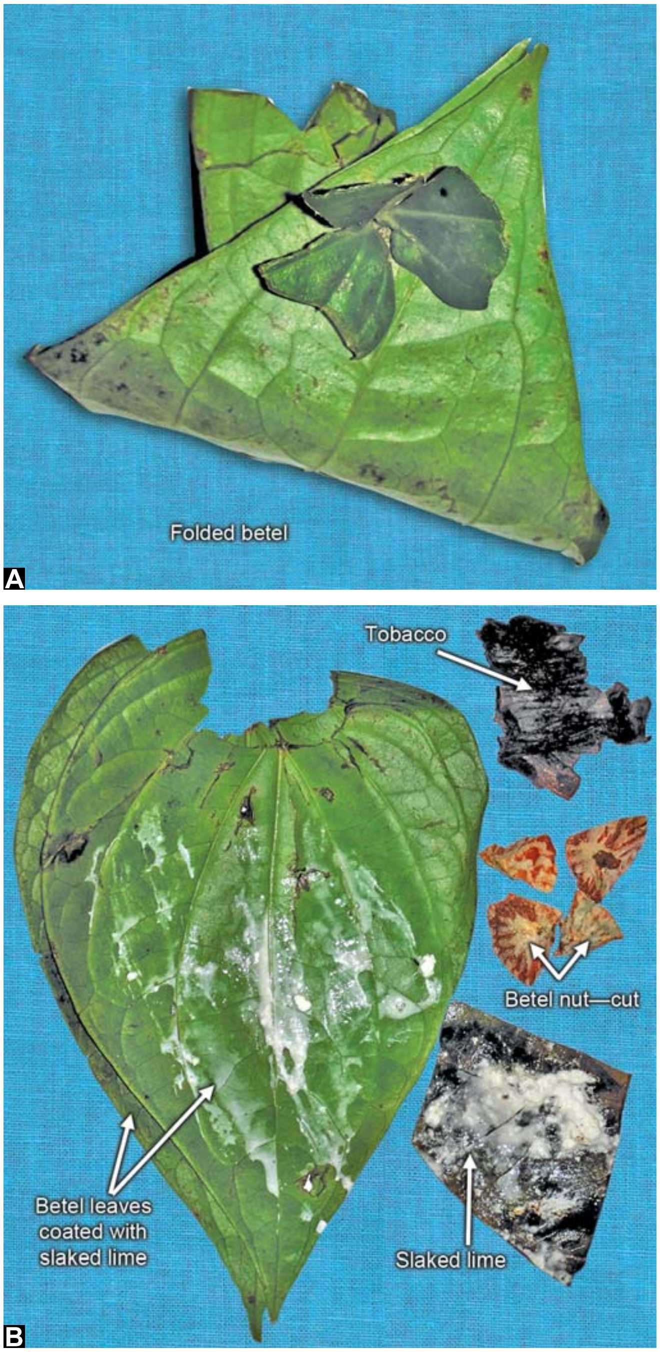

Figs 3.14A and B: Betels commonly used to prepare quid which is kept in gingivoalveolar groove. It is common cause for carcinoma oral cavity.

Verrucous Carcinoma

- It occurs as a superficial proliferative exophytic lesion with minimal deep invasion, often multiple.
- Lesion has white, dry, velvety or warty, keratinised surface. It is common in females.
- It is of low grade, very well-differentiated squamous cell carcinoma, which is locally malignant without any lymphatic spread.
- It is a curable malignancy.
- After biopsy treatment is *wide excision. Radiotherapy* is not given as it may lead to poorly differentiated carcinoma.

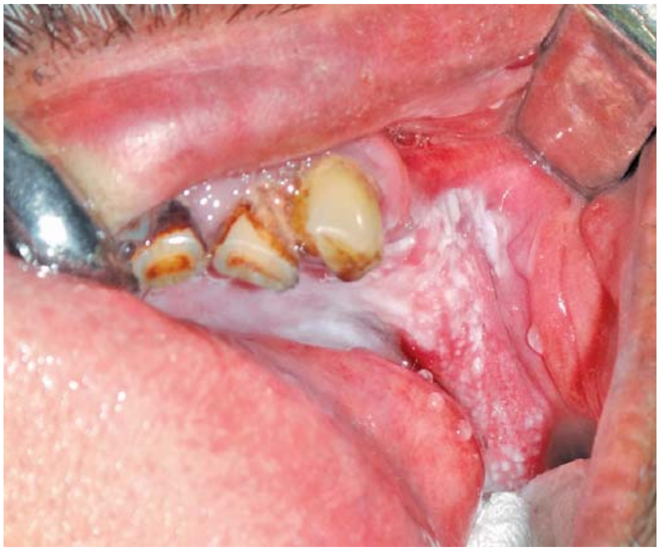

Fig. 3.15: Verrucous carcinoma of oral cavity at gingiobuccal sulcus with leukoplakia.

Biological Behaviour of Carcinoma Cheek

- Carcinoma is common in posterior half of cheek than anterior.
- It spreads into the deeper plane to involve buccinator, pterygoids; into the retromolar trigone, base of the skull, pharynx.
- It spreads outwards to involve the skin causing fungation, ulceration, orocutaneous fistula formation.
- Mandible is commonly involved either by direct extension or through subperiosteal lymphatic plexus which communicates freely with oral lymphatics.
- Lymph nodes commonly involved are submental, submandibular, deep cervical and often lateral pharyngeal groups. Nodal spread is seen in 50% of cases.
- Infection of the tumour area and soft tissues around is common, causing fever, foul smelling ulcer, halitosis.
- Respiratory infection is common in these patients.
- Once tumour extends into the retromolar region, soft palate and pharynx, dysphagia will occur.
- Lesion will later spread to involve alveolus.

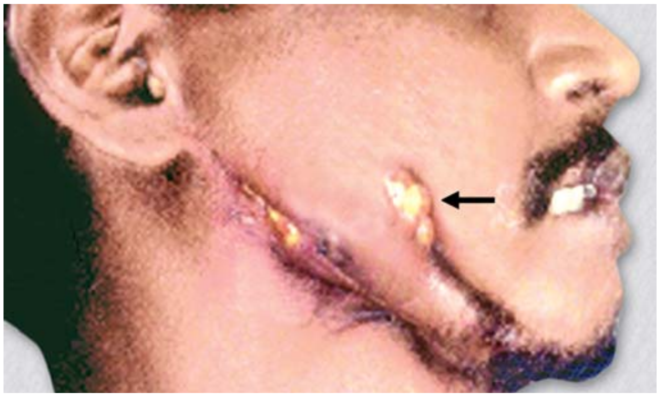

Fig. 3.16: Recurrent carcinoma cheek in an operated case.

Clinical Features

- *Ulcer* in the cheek which gradually increases in size in a patient with history of chewing pan and smoking is the commonest presentation and initially it is painless.
- *Pain* occurs when it involves the skin, bone or if secondarily infected. Referred pain to the ear signifies involvement of lingual nerve. Lingual and auriculotemporal nerves arise from mandibular division of trigeminal nerve.
- *Halitosis* which is bad odour breath is common in many oral cancers. It is due to necrosis of tumour, release *of mercaptan*, butyric acid and ammonia.
- *Involvement of retromolar trigone* indicates that it is an advanced disease, as the lymphatics here communicate freely with the pharyngeal lymphatics.
- *Everted edge, induration* are the typical features of the ulcer.

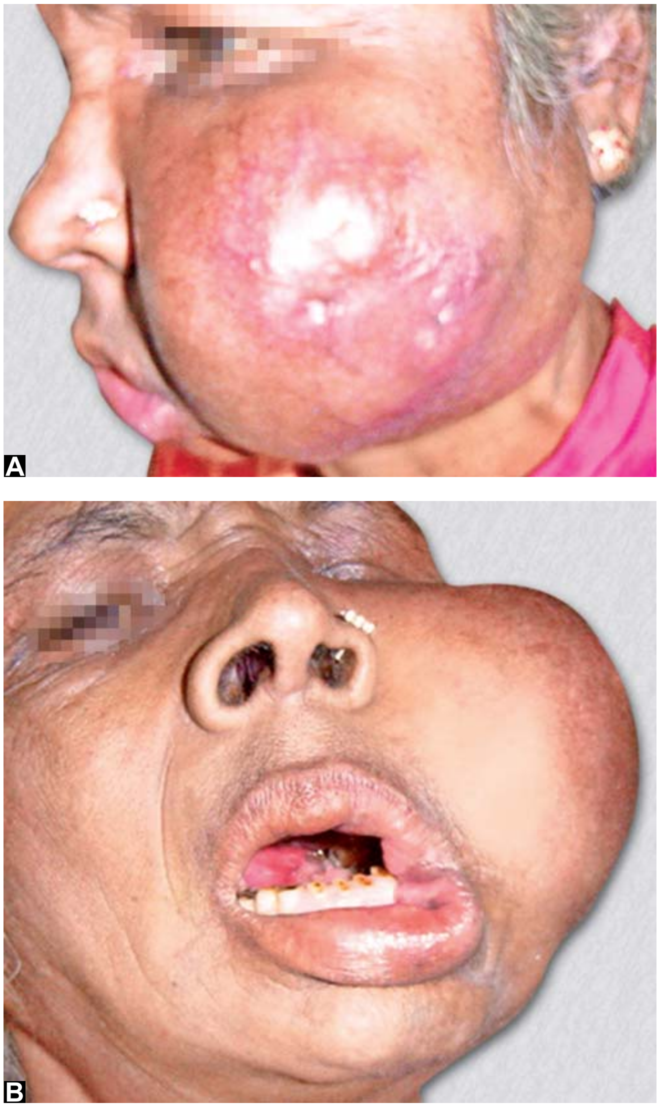

Figs 3.17A and B: Advanced carcinoma cheek presenting as swelling outside. There is an ulcer inside over the mucosa.

- Mandible is examined bidigitally, for thickening, tenderness, irregularity and sites of fracture. Mandible may get involved by direct extension, through mandibular canal, or through periodontal membrane. Loss of central part of mandible due to destruction by tumour will cause pouting of lower lip with drooling of saliva—*Andy Gump* deformity. Mandibular canal is close to occlusive alveolar surface in *elderly and edentulous* patients to cause early mandibular spread in carcinoma.
- *Trismus and dysphagia* signify involvement of pterygoids or posterior extension.
- Occasionally it may extend into the *upper alveolus and to the maxilla* causing swelling, pain and tenderness.
- Once involvement of soft tissue occurs, it may come out through skin as fungating lesion often with orocutaneous fistulas with saliva dribbling through fistula.
- Submandibular lymph nodes and upper deep cervical lymph nodes are involved which are hard and nodular; initially mobile and later get fixed to each other and then to deeper structure.
  - Once lymph nodes get fixed it may infiltrate into hypoglossal nerve (tongue will deviate towards the same side), spinal accessory nerve (defective shrugging of shoulder) and cervical sympathetic chain (*Horner's syndrome*).
  - Compression over external carotid artery leads to absence of superficial temporal artery pulsation.
  - Eventually it causes fungation and bleeding from major vessels—-***carotid blow out.***

***Note*:**

Node involvement may be due to infection. So often trial antibiotic is given initially.

Staging (for all oral cancers)

TNM Staging

T_0	No evidence of primary tumour
Tis	Carcinoma in situ
T_1	Tumour size < 2 cm—greatest dimension
T_2	Tumour size 2-4 cm
T_3	Tumour > 4 cm
T_4	Tumour is of any size involving bone, soft tissues, muscles
Nx	Nodes cannot be assessed
N_0	No lymph node spread
N_1	Lymph node size < 3 cm, same side
N_2	Lymph node size 3-6 cm and single (*N2a*); multiple lymph nodes 3-6 cm size on same side (*N2b*); bilateral or opposite side nodes up to 6 cm size (*N2c*)
N_3	Lymph node spread, more than 6 cm size

Stages

Stage I – T1,N0,M0.
Stage II – T2,N0,M0.
Stage III – T3,N0,M0; T1-3,N1,M0.
Stage IV – T4,N0,M0; T1,N2-3,M0; T0,N0,M1.

> *Fissuring of tongue due to ariboflavinosis is longitudinal and the bottom of fissure is beefy red.*
>
> **—Dimitri Afonsky**

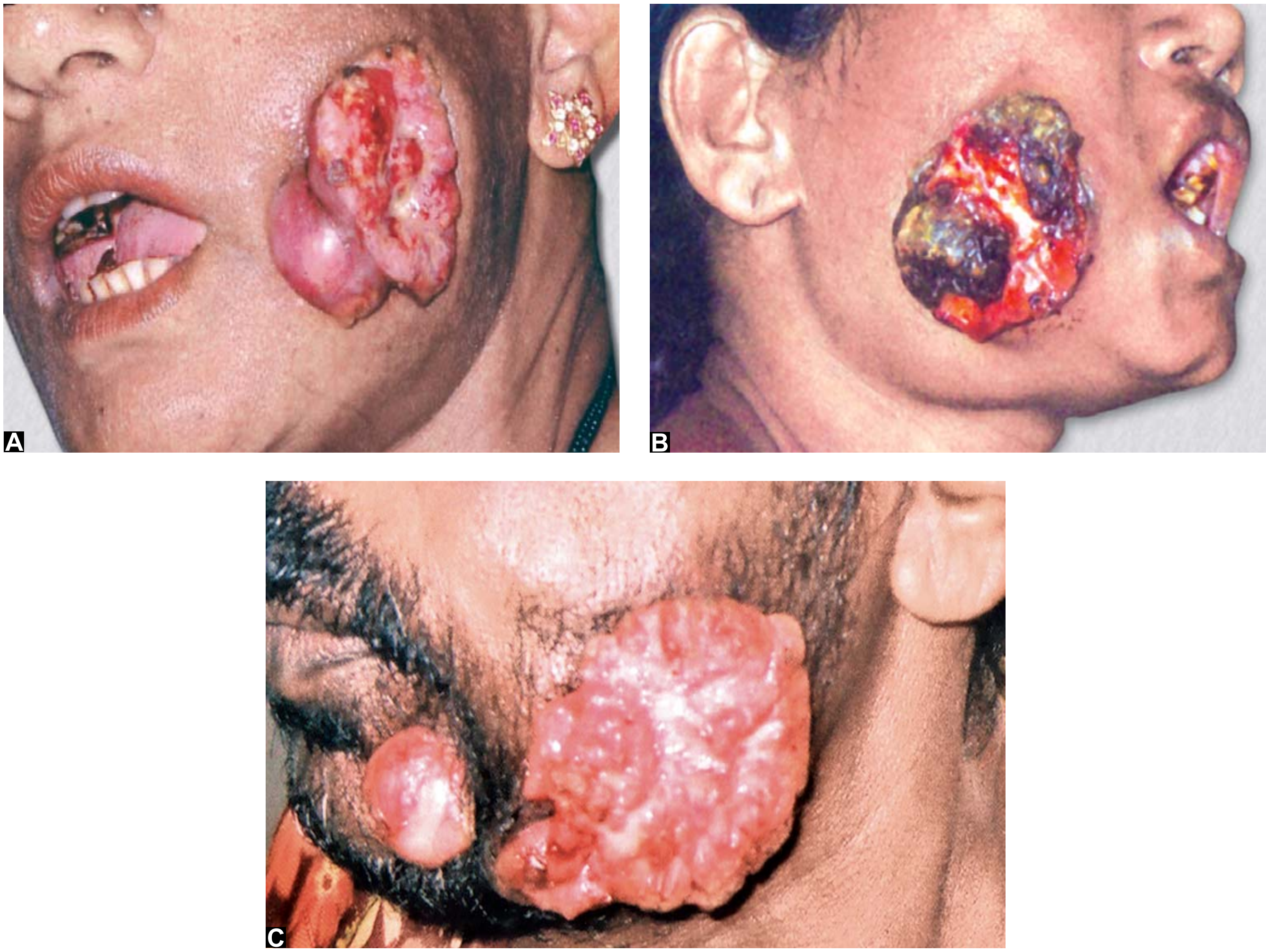

Figs 3.18A to C: Carcinoma cheek fungating outside extensively. Note the pigmentation in one of the pictures.

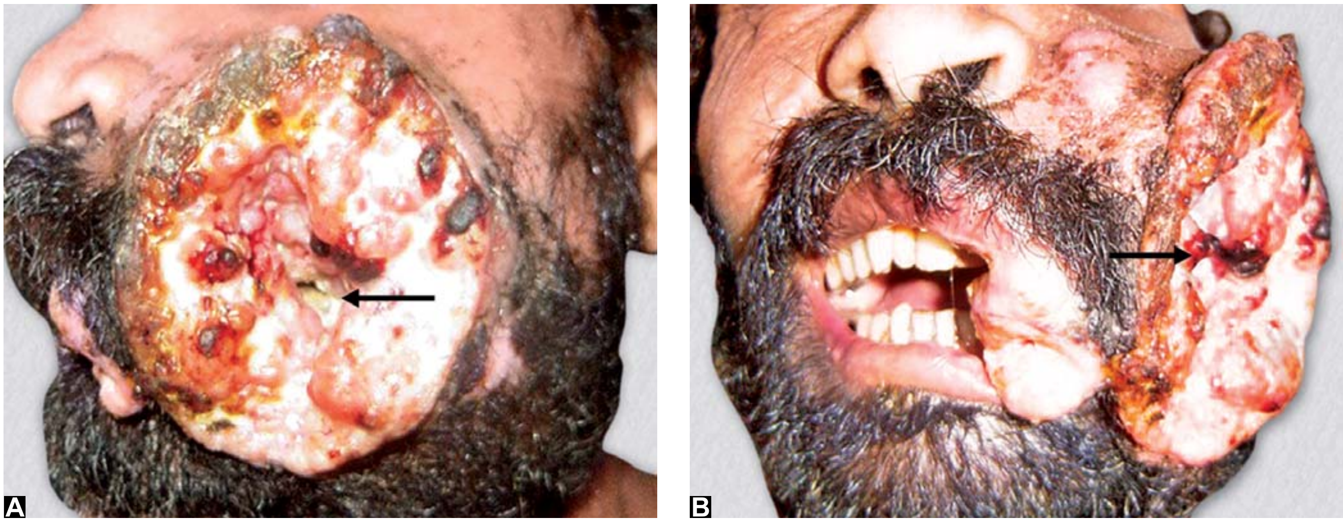

Figs 3.19A and B: Advanced carcinoma cheek with fungation and orocutaneous fistula.

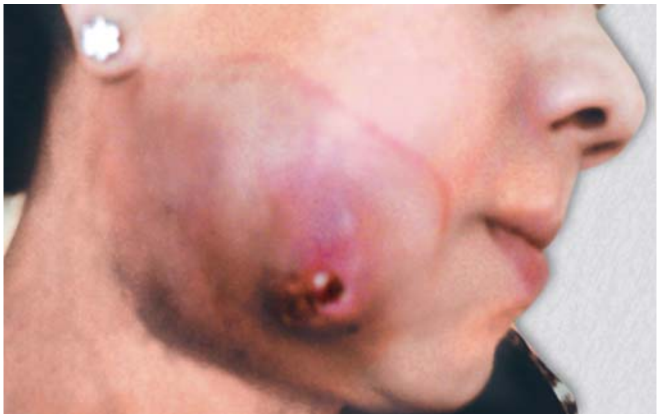

Fig. 3.20: Oral cavity carcinoma with mandibular involvement, with T4 lesion.

Features of advanced carcinoma cheek

- Involvement of retromolar trigone
- Extension into the base of skull and pharynx
- Fixed neck lymph nodes
- Extension to the opposite side

Problems with oral carcinomas

- Upper airway obstruction and bronchopneumonia
- Feeding difficulties and severe malnutrition
- Immunosuppression
- Secondary sepsis, uncontrollable bleeding
- Fixity of secondaries, fungation and disability
- Psychological trauma
- When once trismus develops, patient is unable to take adequate food and eventually leads into cancer cachexia. Trismus may develop by tumour infiltration or after RT.
- Orocutaneous fistula causes salivary dribbling which is distressing.

Investigations

1. *Edge biopsy*, usually taken from two sites. Biopsy has to be taken from the edge as it contains active cells; not from the centre as it is the area of necrosis. *Malignant squamous cells with epithelial pearls (Keratin pearls)* are the histological features.

***Note*:**
Biopsy from the centre is taken only from postradiotherapy ulcer and ulcerated minor salivary gland tumours.

Broder's histological grading

a. Well-differentiated: > 75% epithelial pearls
b. Moderately differentiated: 50-75% epithelial pearls
c. Poorly differentiated: 25-50% epithelial pearls
d. Very poorly differentiated: < 25% epithelial pearls

2 FNAC from lymph nodes.
3. CT scan is used to assess the extent of tumour into mandible, pterygoid region, in patient with trismus, with neck lymph nodes, with carotid involvement by lymph nodes.
MRI is very useful in assessing the soft tissues, base of skull, and perineural spread.

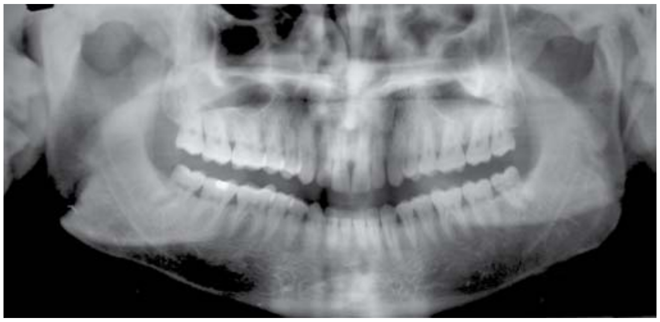

Fig. 3.21: Normal orthopantomogram. OPG is a must in all oral carcinomas to stage the disease and to plan the treatment.

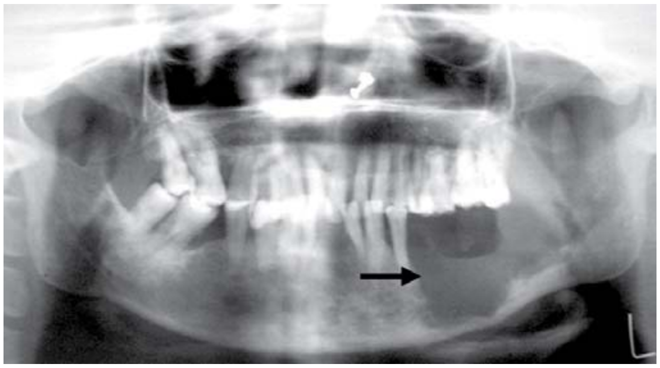

Fig. 3.22: Orthopantomogram showing secondaries in mandible.

4. Orthopantomogram to look for the involvement of mandible—destruction and fracture sites. Symphysis menti and lingual plate are not clearly appreciated. So often OPG may be supported with dental occlusion and intraoral X-rays.
5. Opinion and help of dentist is taken for teeth extraction, dental care, for prosthesis preparation especially for palate, alveolus, dentures.

***Note*:**
Open biopsy should be avoided in case of secondaries in lymph nodes as it may aggravate the spread to further level of lymph nodes.

Treatment

Treatment may be ***curative*** or ***palliative***.

Treatment Strategy

- ***Surgery***. Wide excision, hemimandibulectomy, neck lymph nodes block dissection.
- ***Radiotherapy:*** Curative or palliative; external or brachytherapy.
- ***Chemotherapy:*** Intraarterial, IV or orally.

a. ***Early growth without bone involvement:***
   1. *Curative radiotherapy* using ^{137}Caesium needles or ^{192}Iridium wires, i.e. brachytherapy.
      Advantages:
         i. Surgery is avoided.
         ii. No surgical mutilation.

Aphthous ulcer is rare after the age of 50. **—Wilfred Sircus**

Indications for surgery	*Advantages*	*Disadvantages*
• Early tumour • Tumour spreading to mandible bone/ alveolus • Fungation, haemorrhage due to erosion • Recurrence of tumour after RT • Multiple sites • Soft tissue spread • Locally advanced but amenable to surgical resection	• Pain is controlled • Often curative • Fungation, ulceration, bleeding is prevented	• Mortality (5%) • Morbidity, sepsis, flap necrosis, fibrosis • Cosmetic problem • Loss of anatomy and its function

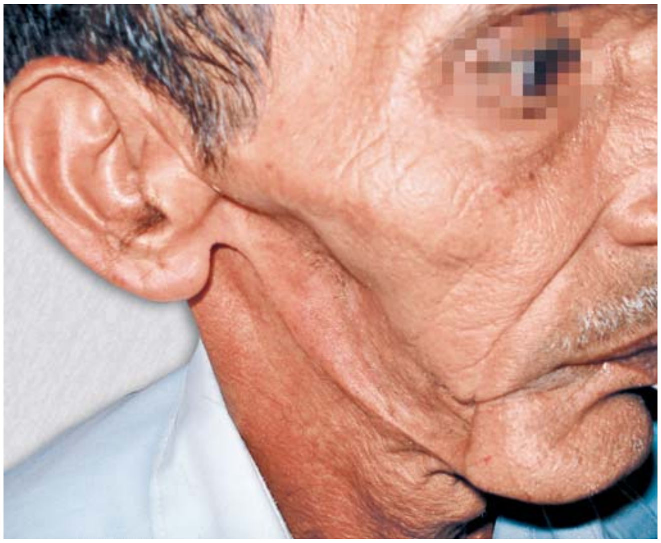

Fig. 3.23: Patient who has undergone hemimandibulectomy with primary closure. Note the defect.

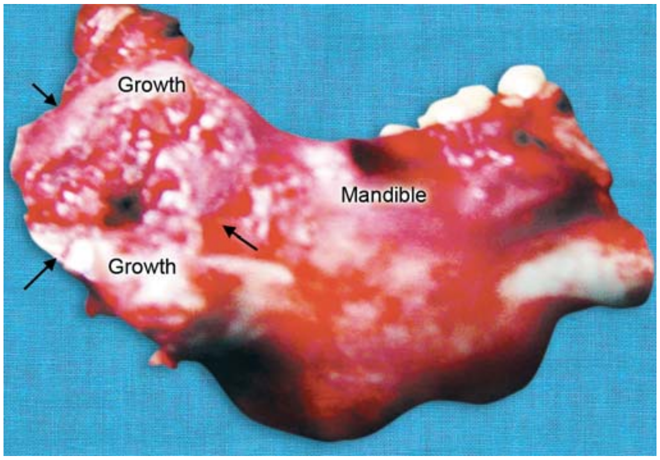

Fig. 3.24: Hemimandibulectomy, resected specimen (*Courtesy:* Dr. Jagadish Chandra, MDS).

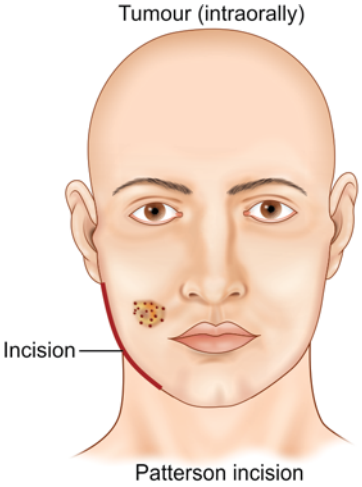

Fig. 3.25: *Patterson incision* to approach carcinoma cheek by raising skin flaps from lower part of face adjacent to lower margin of the mandible.

   iii. Parts are retained.
   iv. As it is a squamous cell carcinoma, primary is radiosensitive—90% cure rate.

2. Other option is *wide excision with 1-2 cm clearance*. Often, the approach to the tumour is by raising the cheek flap (outside). After the wide excision, the flap is placed back ***(Patterson operation)***.
3. Presently advanced technology in radiotherapy, facilitates the use of external radiotherapy also. The incidence of dreaded complication like *osteoradionecrosis of mandible* has been reduced due to better RT methods.

b. ***Growth with mandible involvement:***
   Here along with wide excision of the primary tumour, hemimandibulectomy or segmental resection of the mandible or marginal mandibulectomy (using rotary electric saw) is done.
c. ***Operable growth with mandible involvement and mobile lymph nodes on the same side*** (confirmed by FNAC):
   Along with wide excision of the primary, hemimandibulectomy and radical neck lymph node dissection is done (like commondo operation).
   Wide excision of primary lesion, hemimandibulectomy with radical neck node dissection is called as *composite resection*.
d. ***Operable growth with mandible involvement; mobile lymph nodes on same side and opposite side:***
   Along with wide excision of the tumour, hemimandibulectomy, radical neck lymph node dissection on same side and functional block dissection on opposite side are done, retaining the internal jugular vein, sternomastoid, spinal accessory nerve.
e. ***Operable primary tumour with mobile lymph nodes on same side but without mandibular involvement:***
   Wide excision of primary tumour and radical neck lymph node dissection on same side are done. Mandible is not removed.

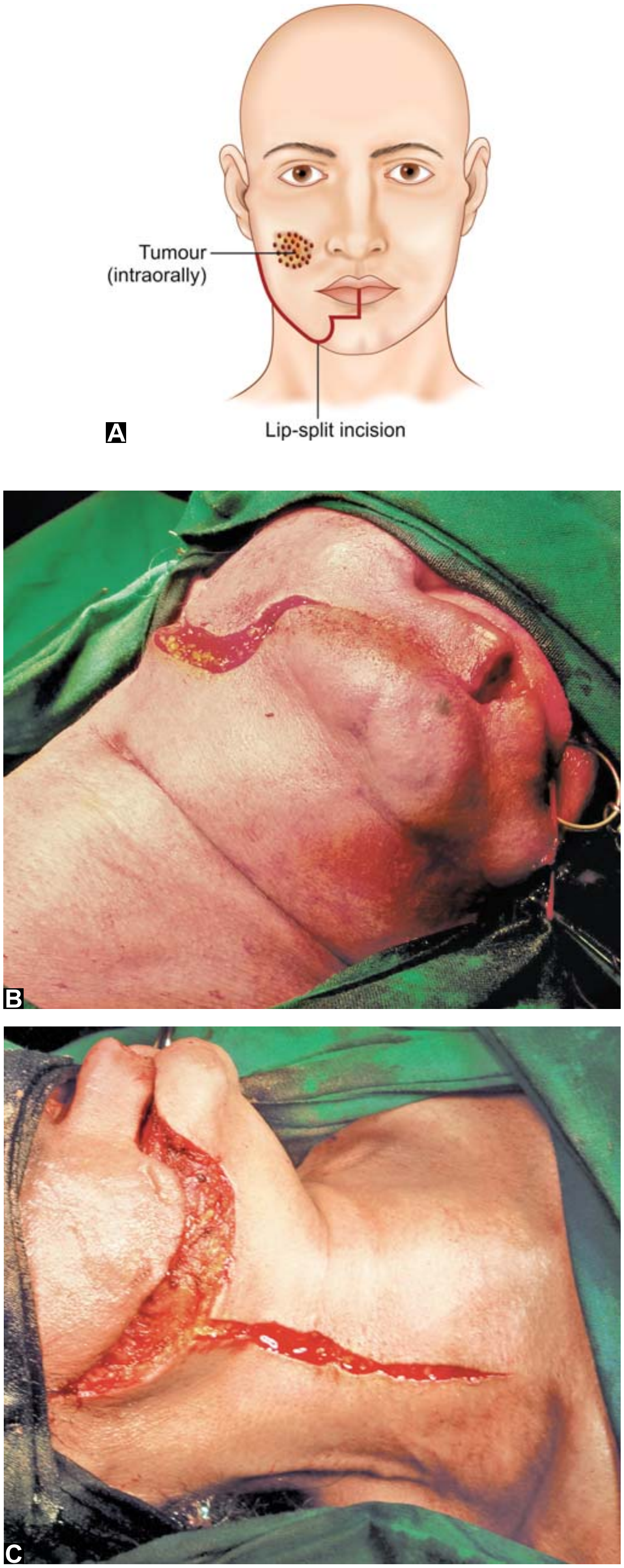

Figs 3.26A to C: Lip-split incision approach for carcinoma cheek. It gives adequate exposure of the cheek. Lower extension is for neck block dissection.

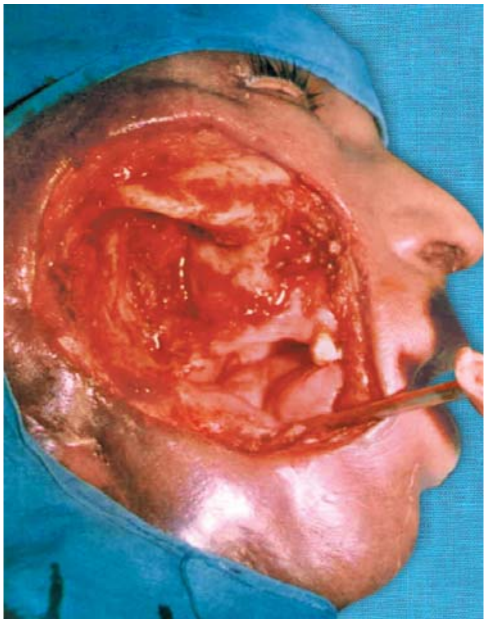

Fig. 3.27: Defect in the cheek after wide excision for carcinoma cheek. Defect needs coverage using PMMF with skin graft inside (*Courtesy*: Professor Kishore Chandra Prasad, ENT Surgeon and Head of the Department and Dr Sampath, ENT Surgeon, KMC, Mangalore).

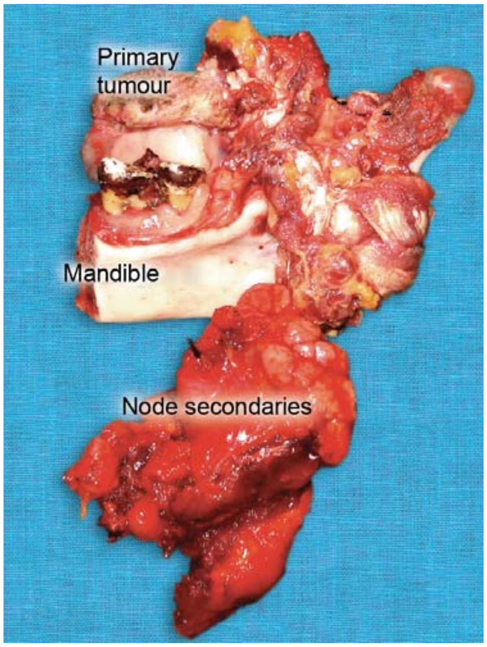

Fig. 3.28: Specimen showing primary tumour in cheek with mandible (hemimandibulectomy) and block dissection nodes.

f. ***Fixed primary tumour or advanced neck lymph node secondaries:***
Only palliative *external radiotherapy* is given to palliate pain, fungation and to prevent anticipated torrential haemorrhage.

g. ***Preoperative radiotherapy is often used in fixed lymph nodes to downstage*** the disease so as to make it operable.

Halitosis is better breath than no breath at all !!!

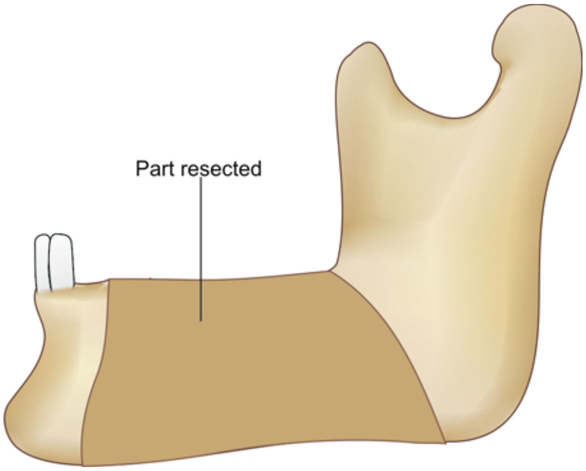

Fig. 3.29: Segmental resection of the mandible (angle of mandible to mental foramen).

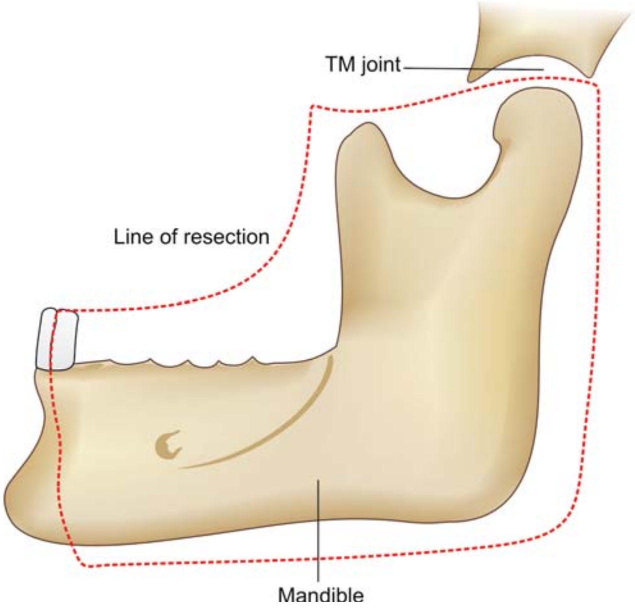

Fig. 3.30: Hemimandibulectomy.

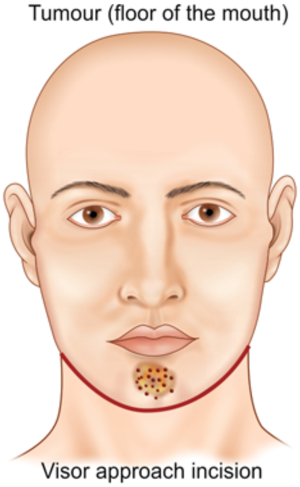

Fig. 3.31: *Visor approach* for oral malignancy. Visor approach for anterior mandible, floor of the mouth and tongue. Here skin over the anterior curved margin of the mandible is incised to approach the floor of the mouth for needed procedure.

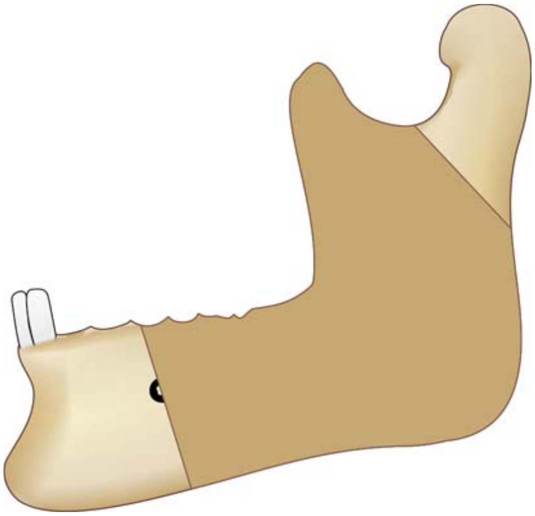

Fig. 3.32: Partial mandibulectomy is removal of mandible one side from mental foramen to line of coronoid process including coronoid process but leaving condylar process.

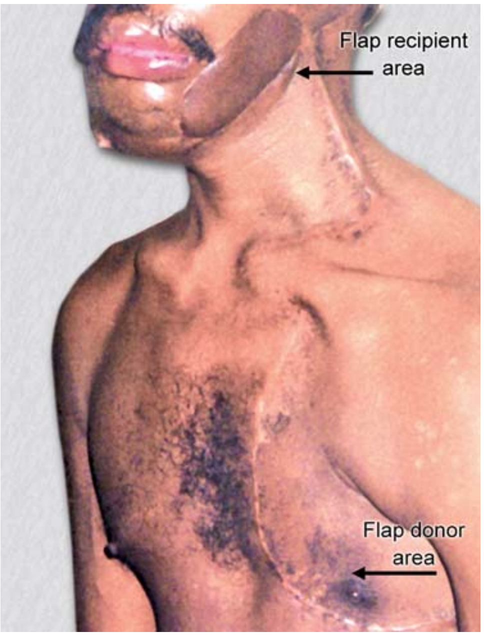

Fig. 3.33: Carcinoma cheek operated with radical neck dissection of same side lymph nodes. Reconstruction done using pectoralis major myocutaneous flap.

h. ***Postoperative radiotherapy is given in T_3 and T_4 tumours:*** N_2 and N_3 nodal status to reduce the recurrence and to improve the prognosis (in multiple nodes and nodes with extracapsular spread).

i. ***Prophylactic block dissection has become popular in N_0 diseases.***

Reasons are—even though clinically, lymph nodes are negative, there may be microscopic involvement of lymph nodes (25-65%).

- Clinically detectable disease in lymph nodes of the patient signifies extracapsular spread which has got poor prognosis.
- Recurrence rate is less after prophylactic block compared to block dissection with clinically positive nodes because

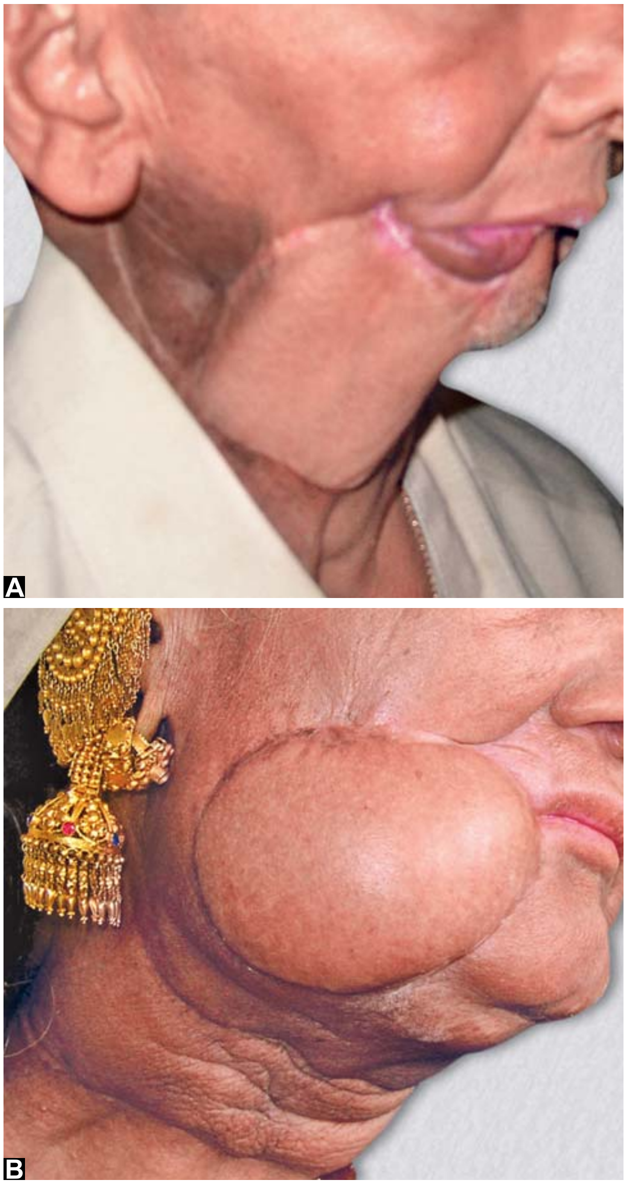

Figs 3.34A and B: Pectoralis major myocutaneous flap to cover the defect in the cheek after wide excision (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and Head of the Department and Dr Sampath,ENT Surgeon, KMC, Mangalore).

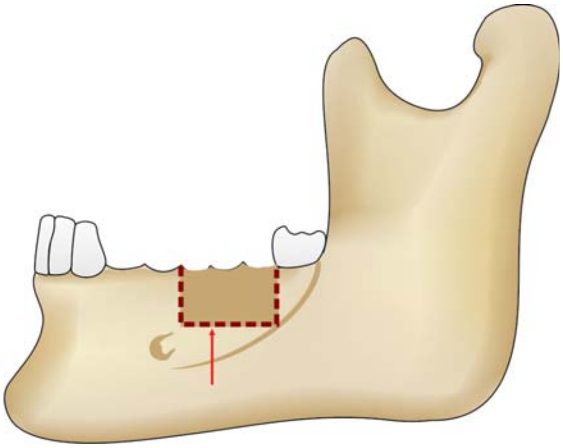

Fig. 3.35: Marginal mandibulectomy.

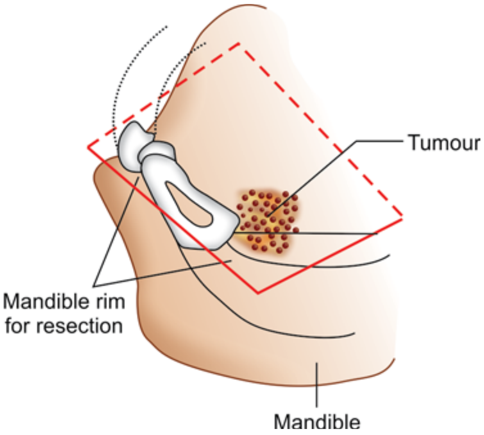

Fig. 3.36: Rim resection of mandible done for carcinoma floor of the mouth.

there is no extracapsular spread in the former even if there is microscopic spread of tumour in many cases.

- Block dissection is an acceptable surgery as there is negligible mortality and less morbidity.
- It is advocated in T3, T4 lesions, carcinoma alveolus or floor of the mouth.

j. ***If growth is extending to upper alveolus:***
Partial maxillectomy or total maxillectomy may be required.

k. ***Role of chemotherapy:***
Drugs used are methotrexate, cisplatin, vincristine, bleomycin, adriamycin. Often it is given intraarterially through external carotid artery using arterial pump or by increasing the height of the drip more than 13 feet, so as to attain a pressure more than systolic pressure. Chemotherapy can also be given IV or orally - postoperatively.

l. Initial chemotherapy to downstage the tumour followed by surgery and later again end with chemotherapy.

m. Chemoradiotherapy is used in unresectable tumours – as consecutive therapies.

Reconstruction after Surgery

Flaps used for reconstruction after oral surgery

- ❖ Forehead flap based on superficial temporal artery
- ❖ Deltopectoral flap based on 1, 2 and 3 perforating vessels from internal mammary vessels
- ❖ Pectoralis major myocutaneous flap (PMMF) based on thoracoacromial artery
- ❖ Free microvascular flaps may be from radial artery forearm flap
- ❖ For small defects—tongue flap, buccal flap, palatal mucoperiosteal flap

- ♦ Split skin graft.
- ♦ Deltopectoral cutaneous flap.
- ♦ Forehead flap, radial artery forearm flap.
- ♦ Pectoralis major myocutaneous flap.
- ♦ Mandible reconstruction by cortical bone graft or rib, fibula or synthetic material like titanium, stainless steel plate.

In great attempts, it is glorious even to fail.

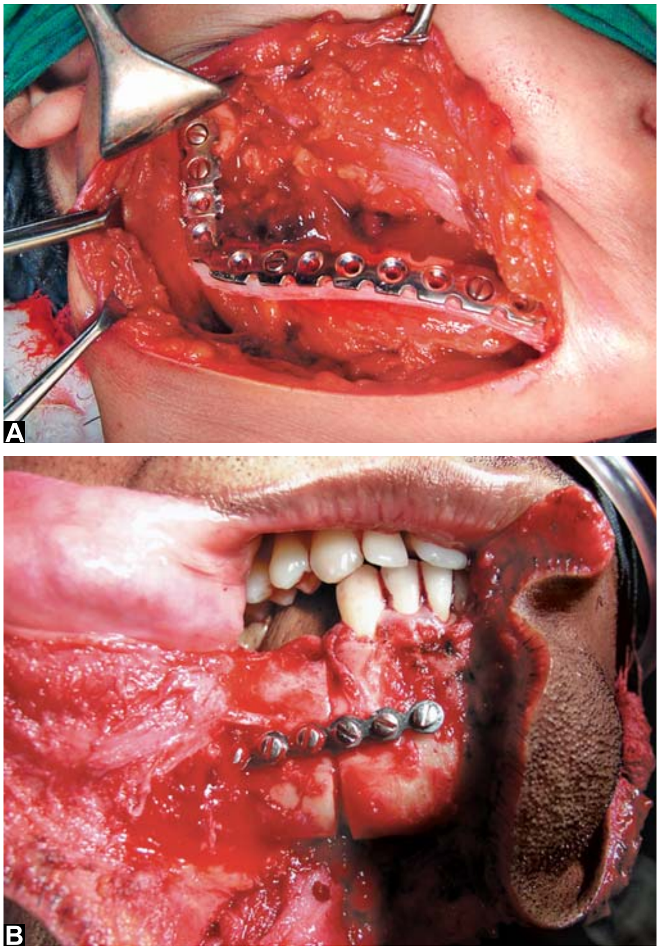

Figs 3.37A and B: Reconstruction of the mandible after segmented mandibulectomy using plate and opposite rib fixation (*Courtesy*: Jagadish Chandra, MDS).

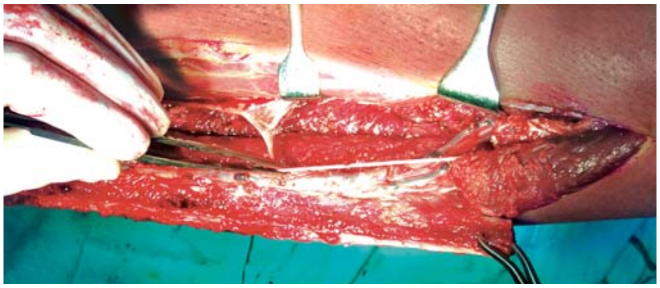

Fig. 3.38: Harvesting fibular flap for mandibular reconstruction. (*Courtesy*: Dr Satish Bhat, MCh, Plastic Surgeon, Mangalore).

Approaches to Carcinoma Cheek

- Transoral/intraoral approach.
- Lip split incision.
- Patterson approach.
- *Visor approach* for anterior mandible, floor of the mouth and tongue. Here skin over the anterior curved margin of the mandible is incised to approach the floor of the mouth for needed procedure. (Visor is French derived word which means mobile lower part of the helmet which covers the chin).

Different mandibular resections

- Marginal mandibulectomy
- Partial mandibulectomy
- Segmental resection
- Hemimandibulectomy
- Rim resection of mandible
- Resection of the anterior mandible through visor approach

Marginal Mandibulectomy

- It is in continuity wide excision of tumour with gingival and adjacent margin of the mandible.
- Superior margin is removed using electric saw after proper marking leaving at least 1 cm of inferior margin.
- It is done in lesion which is close to mandible but not invading. Invasion needs segmental resection. Invasion is invariably through dental root leading into mandibular canal and cancellous bone.
- Mandibular involvement either clinically or radiologically, previous RT, retromolar lesion (as pterygoid clearance is needed here) are the contraindications.

Problems with surgery

- Mutilation (surgical)
- Anaesthesia complications
- Bleeding
- Infection
- Flap necrosis
- Requirement for reconstruction
- Mortality
- Morbidity—stiffness; contracture; cosmetic problem; cutaneous anaesthesia; speech and swallowing problems
- Problems of neck dissection—hypoglossal, accessory, phrenic nerve injuries; thoracic duct injury, carotid blow out; oedema neck and face

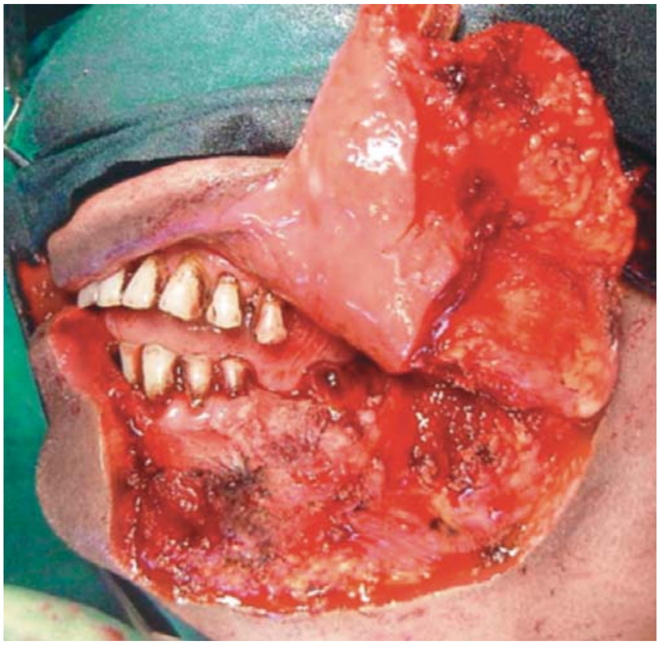

Fig. 3.39: Carcinoma cheek, on table wide excision. It requires pectoralis major myocutaneous flap for reconstruction.

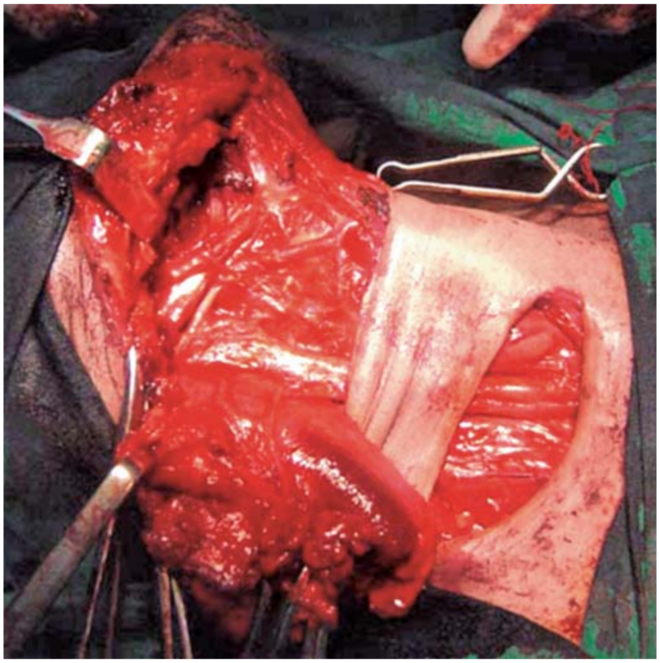

Fig. 3.40: Radical neck dissection for secondaries in neck through MacFee incision.

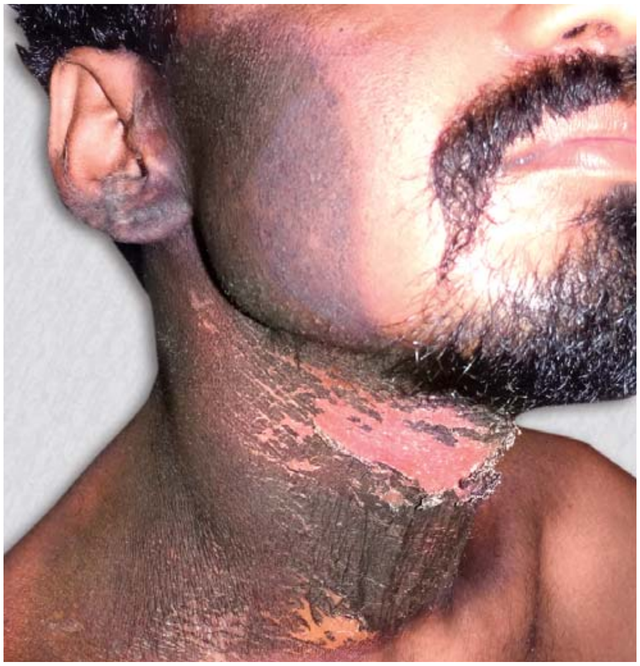

Fig. 3.41: Radiodermatitis after RT to tonsillar growth.

Problems with radiotherapy

- When mandible is irradiated, chances of the dreaded problem, osteoradionecrosis is high which requires the removal of mandible
- Loss of taste sensation and dryness, xerostomia
- Infection, mucositis, dental diseases
- Skin excoriation, hair loss
- Trismus may get aggravated
- Can itself cause dysphagia, laryngeal oedema
- Hypothyroidism if neck is irradiated
- Radiation neuritis causing severe pain
- Carotid artery atherosclerosis
- Visual impairment
- Shoulder and neck dysfunction
- Soft tissue fibrosis

Problems with chemotherapy

- Bone marrow suppression
- Megaloblastic anaemia
- GIT symptoms
- Hepatotoxicity and renal toxicity
- Alopecia
- Nausea, vomiting and severe stomatitis

Prognostic factors in oral carcinomas

- Stage of the disease
  - Stage I and II has got 80% 5 years survival
  - Stage III and IV has got less than 20% 5 years survival rate

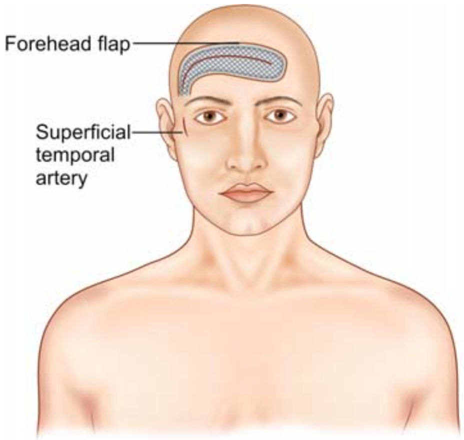

Fig. 3.42: Forehead flap is used after wide excision of carcinoma cheek. It is based on superficial temporal artery—anterior branch.

- T3 and T4 lesions has got poor survival rate
- Carcinoma lip has got best prognosis
- Carcinoma posterior 1/3rd tongue has got worst prognosis
- Cheek, floor of the mouth and palate has got intermediate prognosis
- Perineural invasion and angioinvasion carries poor prognosis
- Histologically positive nodes decrease the survival rate by 50%
- Level III and IV, node > 3 cm, bilateral nodes extracapsular nodal spread are poor prognostic factors
- Grading (differentiation) of the tumour
- Tumour thickness > 6 mm has got poor prognosis
- Exophytic tumour is better than infiltrating type

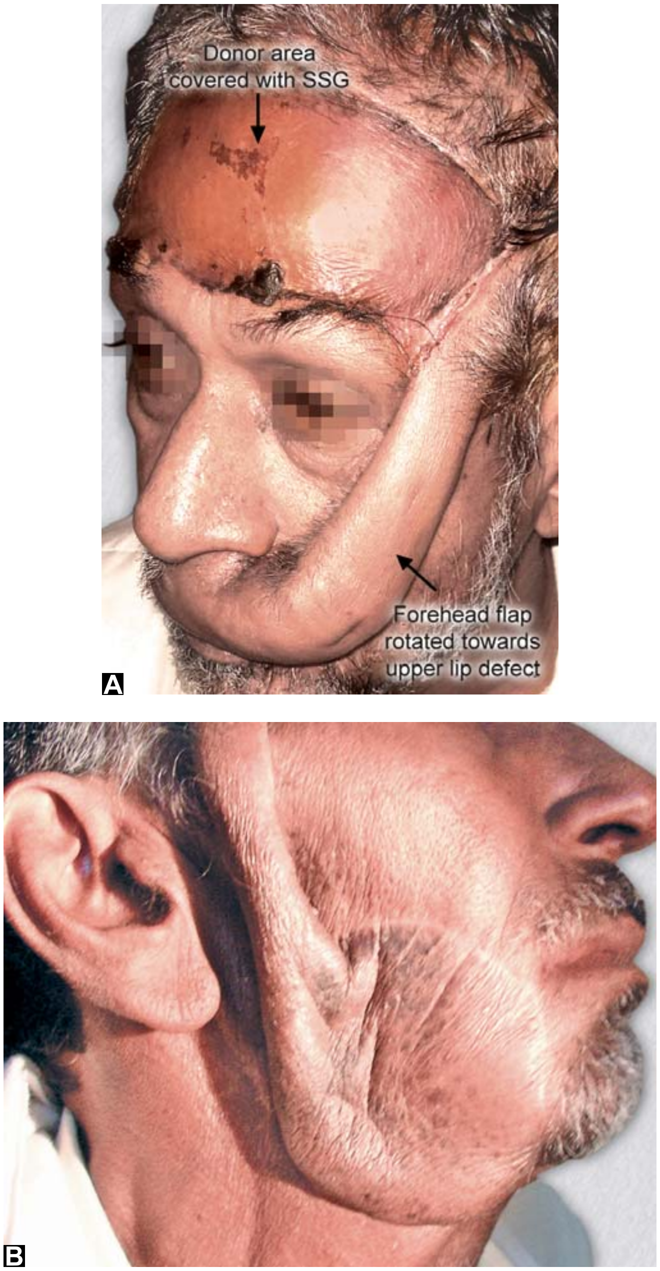
Donor area covered with SSG
Forehead flap rotated towards upper lip defect
A
B

Figs 3.43A and B: Forehead flap used to cover the defect after wide excision for carcinoma lip or cheek (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and Head of the Department and Dr Sampath, ENT Surgeon, KMC, Mangalore).

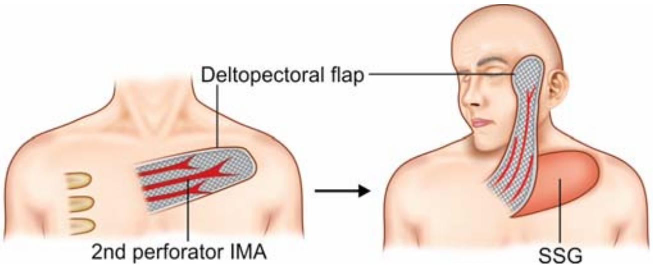

Fig. 3.44: Deltopectoral flap used to cover defect after surgical clearance in carcinoma cheek. It is based on 2nd perforator, a branch of internal mammary artery (*Bakamjian—1965*).

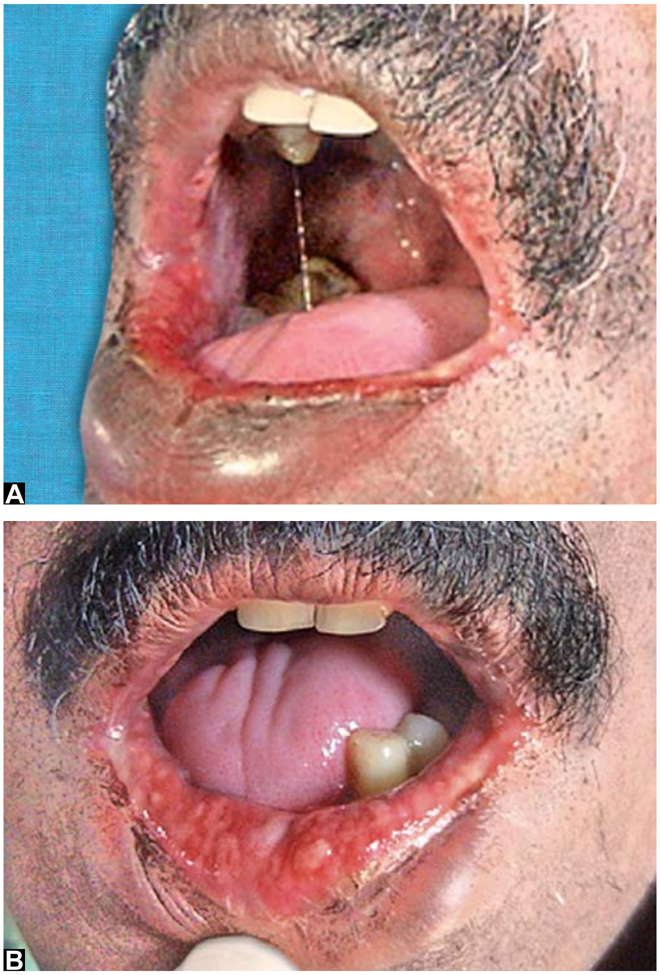

Figs 3.45A and B: Mucositis in cheek and lip in a patient who is receiving chemotherapy for carcinoma.

LIP

Lip begins at vermilion border. It has got upper lip, lower lip and oral commissure. SCC is the commonest lip cancer (90%). SCC is common in lower lip; BCC is common in upper lip. Other cancers in lip are spindle cell carcinoma, adenoid squamous carcinoma, malignant melanoma, minor salivary gland tumour.

Mucous cyst of lip is a common condition. It can occur in upper or lower lip. It is a retention cyst derived from mucous glands of the lip. It presents as bluish, soft, fluctuant often transilluminating well localized swelling. It may resolve on its own. If it does not it should be excised under local anaesthesia. Absorbable suture is used to appose the wound starting from vermilion margin. Usually vertical elliptical incision is used to excise. Sutured wound heals rapidly with very limited scar.

Macrocheilia is enlargement of lip mass in other than neoplastic conditions. Lymphangioma is the common cause. Haemangioma, inflammatory conditions also can cause macrocheilia.

Papilloma, lipoma, pyogenic granuloma, keratoacanthoma, minor salivary tumours are other swellings which can occur in lip.

Cheilitis often associated with stomatitis is common in vitamin deficiency, malnutrition, sepsis, drug induced, RT induced presents as redness, pain, diffuse swelling. In chronic cases, there will be linear ulcers especially at commissure. *Actinic cheilitis* is a premalignant lesion. Cause is treated.

Pigmentation of lip occurs in Peutz-Jeghers syndrome (brown), Addison's disease (black).

Herpes labialis is formation ulcers in lip due to herpes simplex virus. Repeated multiple ulcers develop in lip. It is contagious by kissing. So kissing should be avoided including children. Touching repeatedly can transfer the virus to eye causing herpes keratitis.

Cleft lip is a common congenital condition.

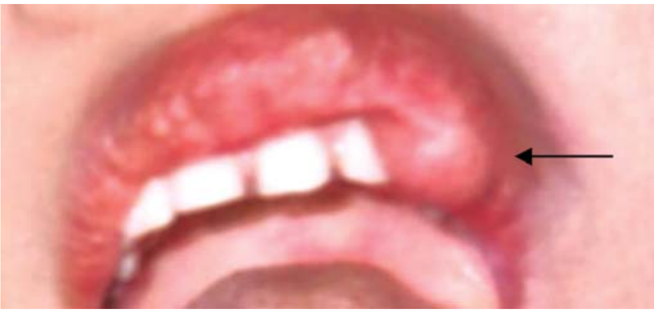

Fig. 3.46: Mucous cyst upper lip. It requires simple excision. It is a retention cyst.

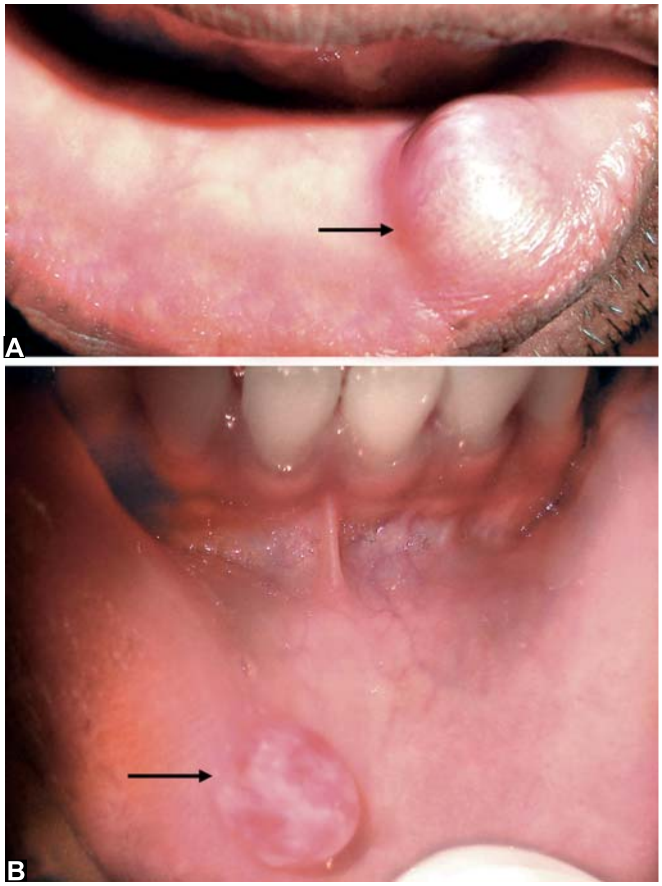

Figs 3.47A and B

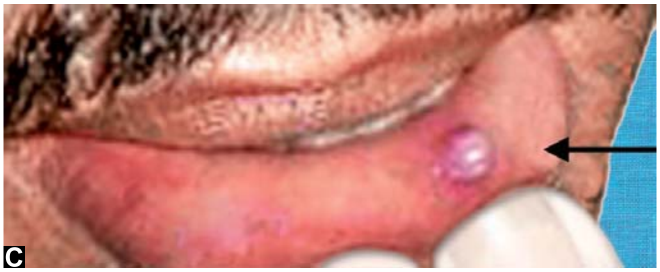

Fig. 3.47C

Figs 3.47A to C: Mucous cyst seen in lower lip. It is bluish, localised, smooth, nontender, soft, fluctuant swelling, arising as a retention cyst from mucous glands of lip. It needs excision.

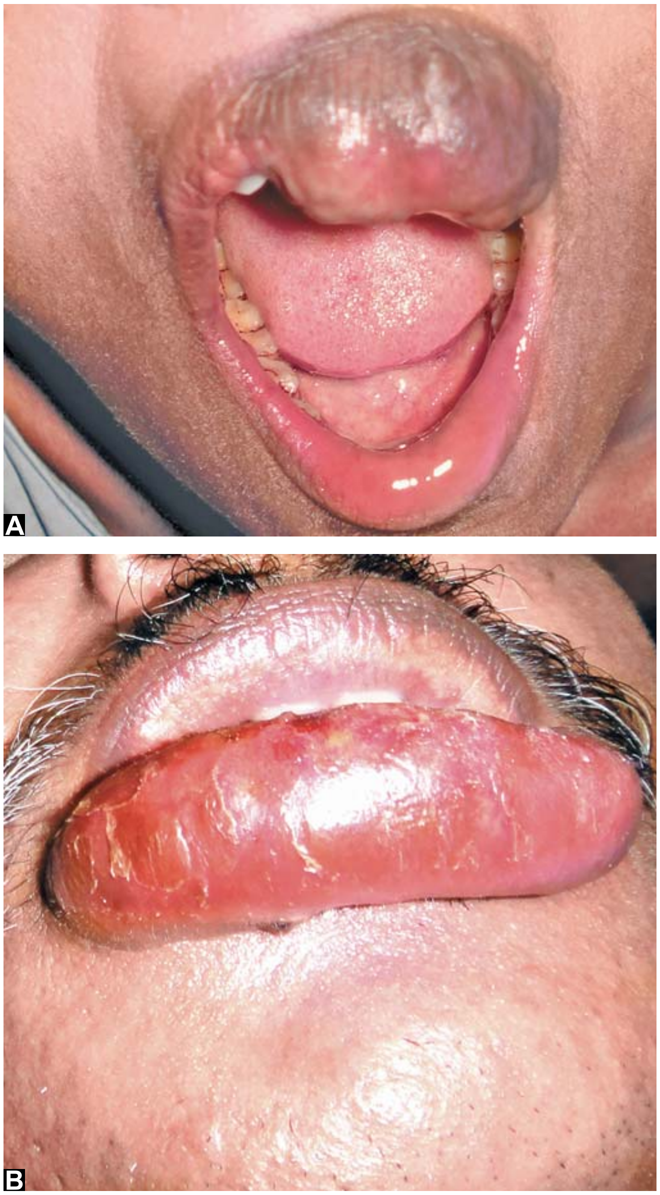

Figs 3.48A and B: (A) Haemangioma lip causing macrocheilia. (B) Macrocheilia due to inflammatory cause.

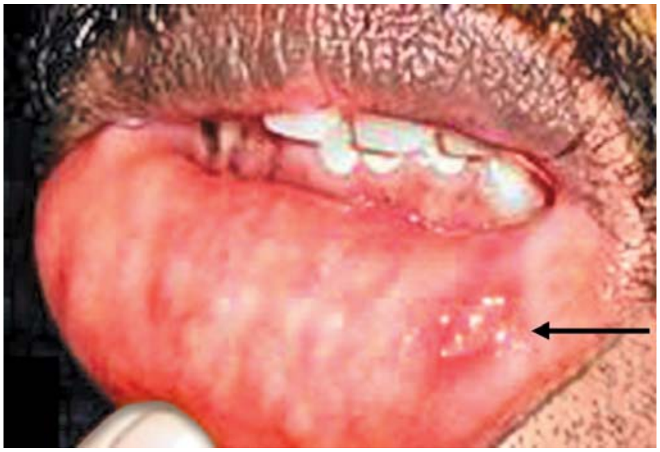

Fig. 3.49: Aphthous ulcer lower lip. It is painful, self limiting ulcer. It can occur in tongue, lip, cheek and other parts of the oral mucosa.

NEOPLASM OF LIP

> *One may…cure cancers of the lip without applying caustics or any similar thing….Pass a threaded needle through the cancer so the thread held in the left hand can lift and control the cancer without any of its escaping. One can then cut to good flesh with scissors in the right hand; and so cut that a layer of good flesh of the lip remains to serve as a base and foundation for regeneration of flesh in place of the portion amputated, supposing the cancer has not taken root and spread from top to bottom.*
>
> —**Ambroise Paré,** 1585

*Minor salivary gland tumour*s are common in upper lip. They are usually pleomorphic adenomas.

CARCINOMA LIP

- Incidence of carcinoma lip is 15% of head and neck cancers and 1% of all cancers.
- It may arise from vermilion surface or mucosa part of the lip.
- It is common after 40 years of age. In younger age group even though it is rare, carries poor outcome.
- It is more common in whites; rare in blacks. It is common in *Caucasians*.
- Khaini, a mixture of tobacco and lime kept under the lip called as *khaini chewers* are more susceptible for carcinoma lip.
- All'S's and irradiation are predisposing factors.
- *Verrucous carcinoma* can occur in lip. It is well differentiated SCC with exophytic, warty, dry surface. It usually will not invade the deeper tissue and lymph node spread is rare. It carries good prognosis.
- It is common in men (15:1). Common in *lower lip* (90%); upper lip 5-10%. Upper lip is not exposed to direct actinic radiation.
- Commonly due to exposure to sunlight (ultraviolet rays). Common in pipe smokers.
- Initially starts as a red, granular dry lesion which eventually gets ulcerated and forms an ulceroproliferative lesion. Occasionally it occurs at the angle of mouth.
- It spreads to submental nodes and later to other neck nodes on both sides.

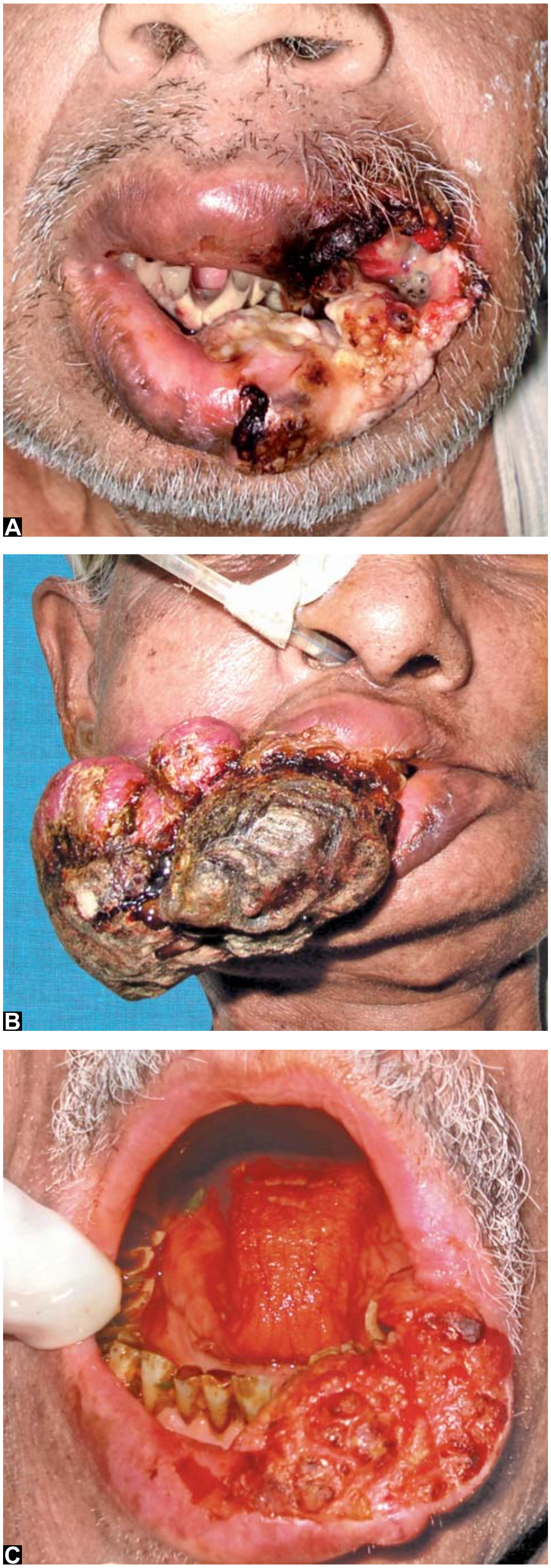

Figs 3.50A to C: Carcinoma of lip involving extensively.

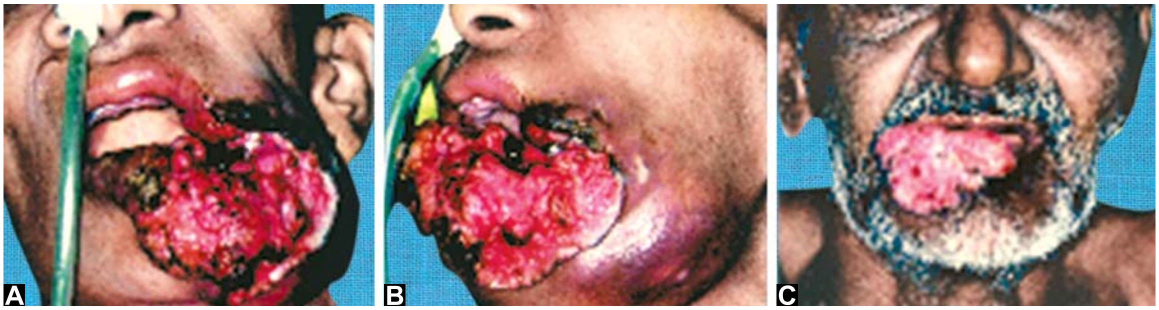

Figs 3.51A to C: Carcinoma lip.

- Usually it is a *well-differentiated squamous cell carcinoma*.
- Commissure tumours are 2% of all lip cancers but has higher rate of lymph node spread.
- Initially it is well localized. A depth of more than 5 mm will spread to lymph nodes mainly submental and submandibular (level I).
- Painless ulcerative lesion is the most common presentation. Pain develops once necrosis occurs or tumour infiltrates nerve, periosteum of bone underneath.
- Pigmented SCC can occur. It mimics melanoma.
- *Anaesthesia of chin* can occur if mental nerve is invaded.
- Staging is same as other oral cancers.

Clinical features of carcinoma lip

- Non-healing progressive ulcer, painless to begin with
- Everted edge with indurations
- Growth moves with the lip
- Submental, submandibular and upper deep neck nodes may get enlarged.
- Tender firm lymph node may be due to infection; hard initially nontender node is due to carcinoma spread.
- In half of the cases lymph nodes are enlarged due to infection or as reactive process
- Fungation, bleeding, halitosis

Predisposing Factors

- Cheilitis—actinic type
- Solar keratosis.
- Papilloma.
- Leukoplakia.
- Smoking, U-V rays, pipe smokers, reverse smoking.
- Tobacco chewing, Khaini chewers (tobacco+lime).
- Agriculturists who are commonly exposed to sunlight get carcinoma lip and is called as *countryman's lip*.

Differential Diagnosis

- Keratoacanthoma.
- Basal cell carcinoma. BCC occurs only in upper lip.
- Minor salivary gland tumours.
- Often carcinoma of lip is an extension from carcinoma of cheek.
- Pyogenic granuloma in early cases only.
- Malignant melanoma in case of *pigmented SCC*.

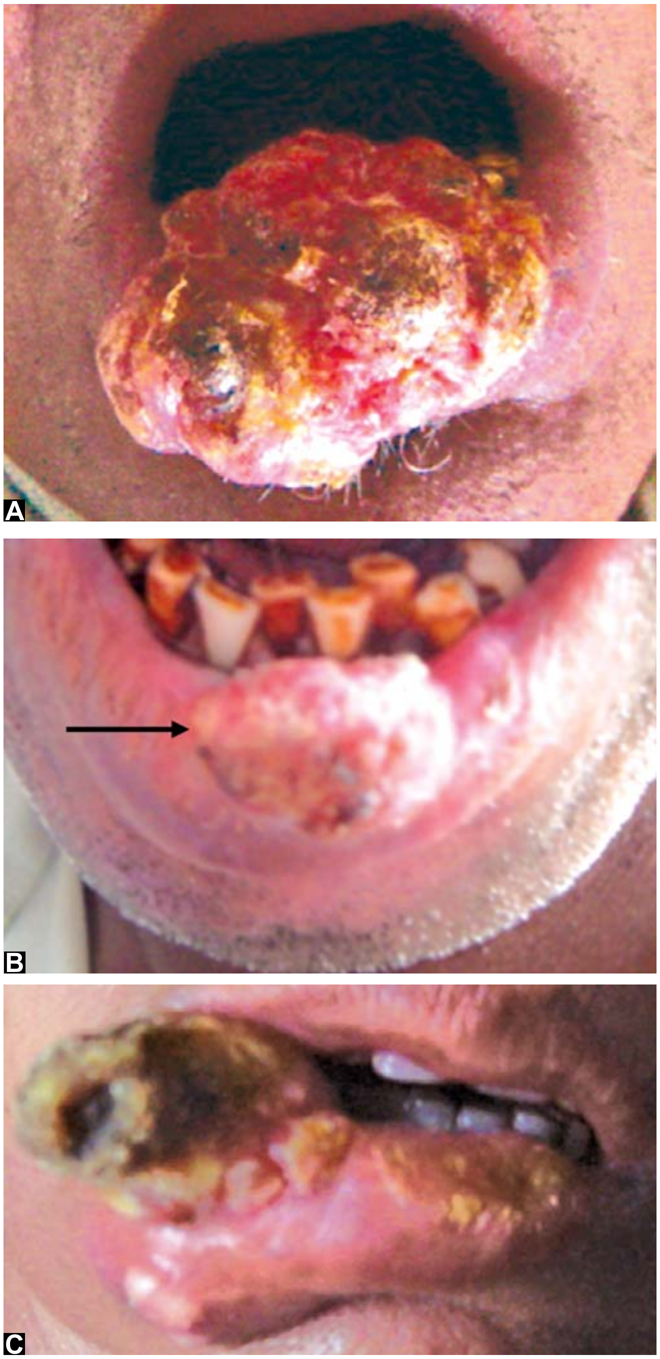

Figs 3.52A to C: Carcinoma lip involving different parts. Commonly they are squamous cell carcinomas.

Diagnosis

Edge biopsy, FNAC of lymph nodes.

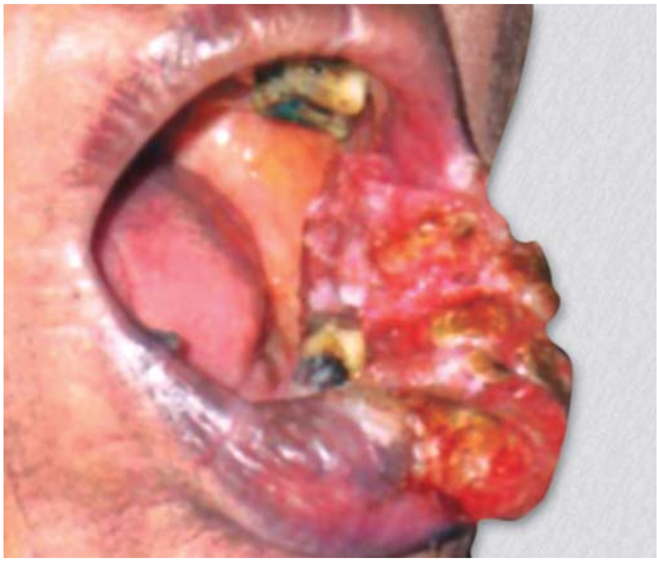

Fig. 3.53: Carcinoma lip extending into the floor of the mouth.

Treatment

- If lesion is less than 2 cm, then curative radiotherapy, either brachytherapy or external beam radiotherapy. It gives a good cure.
- If tumour is more than 2 cm, wide excision is done. Excision of lower lip up to one-third can be sutured primarily, in layers keeping vermilion border in proper apposition without causing any microstomia.
- Excision of more than one-third of the lip requires reconstruction using different flaps.

Methods

1. *Abbe-Estlander's rotation flap* used for either upper or lower lip lesions (of less than ½ of lip) located *at the angle* based on labial artery. Here base at a later stage need not be disconnected unlike in Abbe lip.
2. *Fries' modified Bernard facial flap*—reconstruction using lateral facial flaps. It is used when defect is more than ½ of lip and midline.

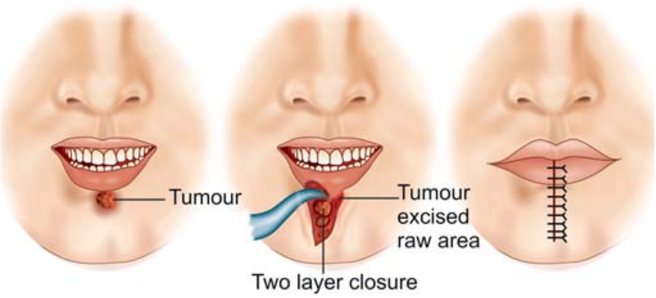

Fig. 3.54: Primary repair of lip after wide excision of small tumour. One-third of the lip can be sacrificed. Lip is sutured in layers. First layer, muco-muscular layer with absorbable suture. Second layer is skin with non-absorbable monofilament sutures like polypropylene.

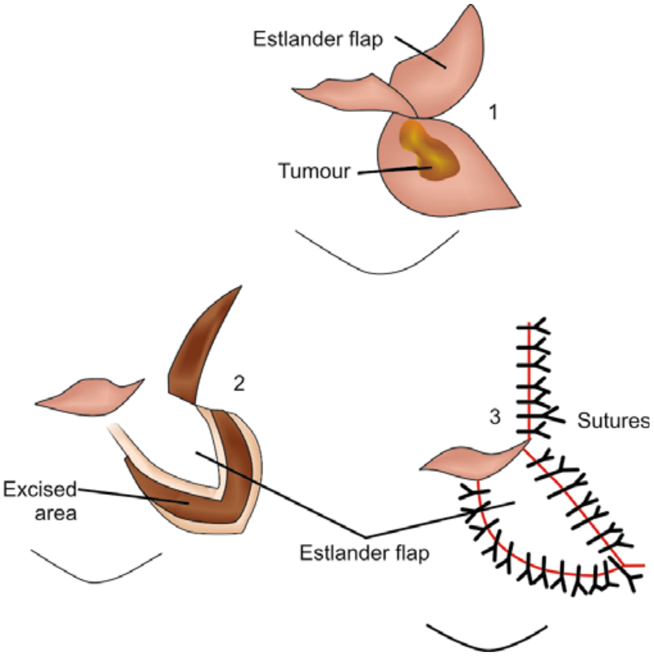

Fig. 3.55: Lower lip tumour after excision when primary suturing not possible, then upper lip flap based on upper labial artery can be used – *Abbe-Estlander flap.*

3. Microvascular flaps.
4. *Nasolabial flap*: It is used when defect is more than ½ of lip laterally or defect is in the floor of the mouth.
5. Cheek flap.
6. Free radial artery flap.
7. *Abbe flap* (switch flap) is used for upper or lower lip lesions at the *middle* or at the site *other than angle* based on the labial artery. Here at a later *2nd stage* base of the flap should be released once flap takes up.
8. *'W' flap plasty*: It is done for lower lip middle tumour which is less than 1/3rd of the lip.
9. *Gillies fan flap*: It is a cheek flap usually bilateral but can be unilateral. Incision is full thickness around commissures extending into nasolabial fold and upper lip upto upper lip vermilion border. Flap which is based on labial vessels is advanced towards the defect. Vermilion is reconstructed with tongue mucosal flap which is divided in 3 weeks.
10. *Karapandzic flap*: It is modified version of the Gillies flap used for lower lip defect with less angulation towards upper lip. Reverse Karapandzic flap is used for upper lip defect.
11. *Johansen 'stepladder'* procedure is used for extensive carcinoma of lower lip.
12. Other regular flaps like forehead flap, deltopectoral flap also can be used.
    - Lymph nodes are dealt with by radical neck dissection on one side and functional block or supraomohyoid block dissection on other side. For central tumour N_0 disease, bilateral elective (prophylactic) supraomohyoid dissection is done. For lateral tumour N_0 disease, elective ipsilateral supraomohyoid dissection is done.
    - Postoperative radiotherapy is given if tumour is large or if lymph nodes are involved.
    - When mandible is involved, segmental resection is done.

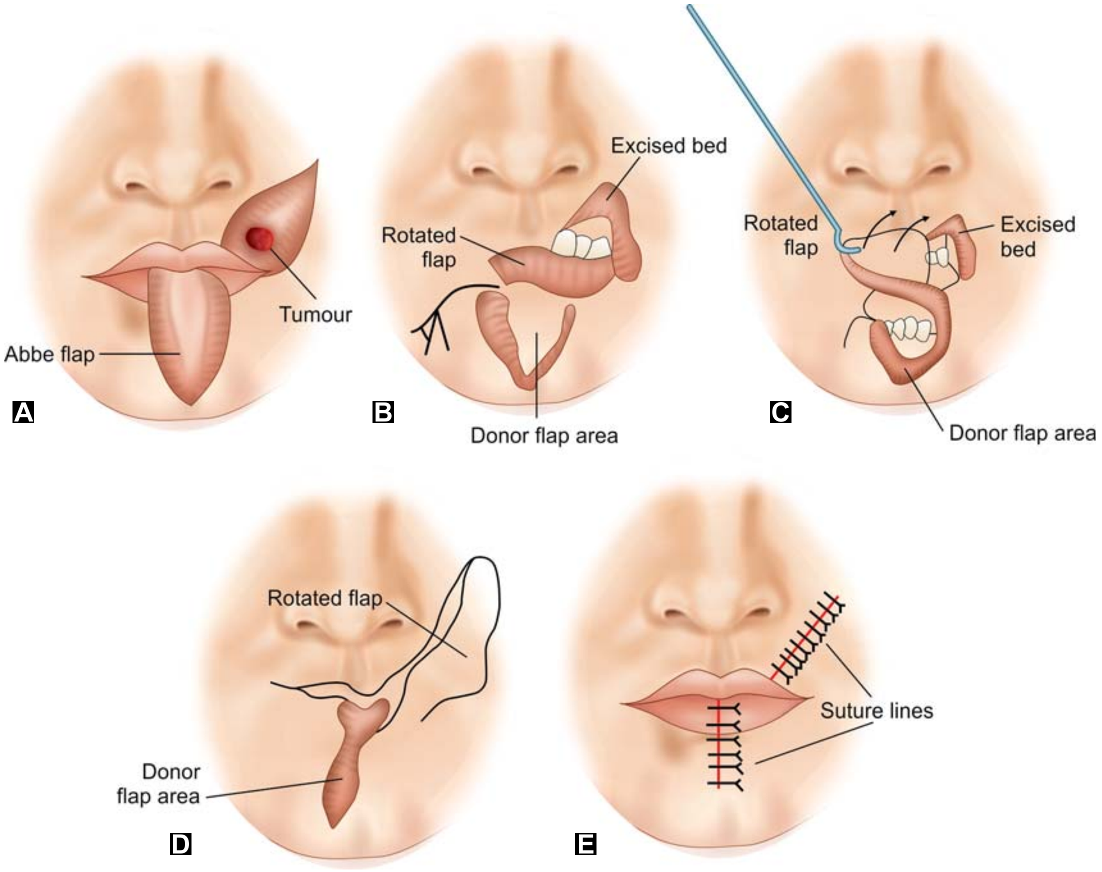
Tumour
Abbe flap
A
Excised bed
Rotated flap
Donor flap area
B
Rotated flap
Excised bed
Donor flap area
C
Rotated flap
Donor flap area
D
Suture lines
E

Figs 3.56A to E: For upper lip tumour when primary repair is not possible, then lower lip flap based on inferior labial artery can be rotated to upper lip—*Abbe flap*.

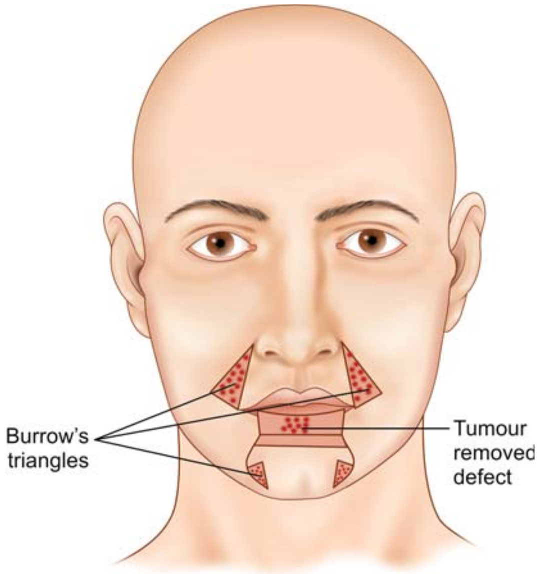
Burrow's triangles
Tumour removed defect

Fig. 3.57: It is used for central lip tumours. Angles of the lower part of the defect and upper part of the angles of the mouth are rotated inwards with Burrow's triangles at the site of rotation—*Bernard flap.*

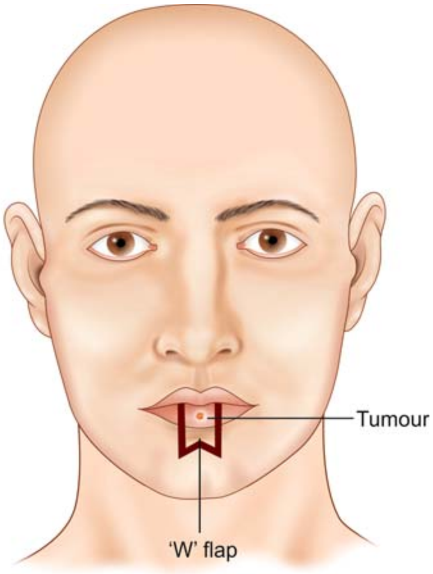
Tumour
'W' flap

Fig. 3.58: It is done for lower lip middle tumour which is less than 1/3rd of the lip—*'W' flap*.

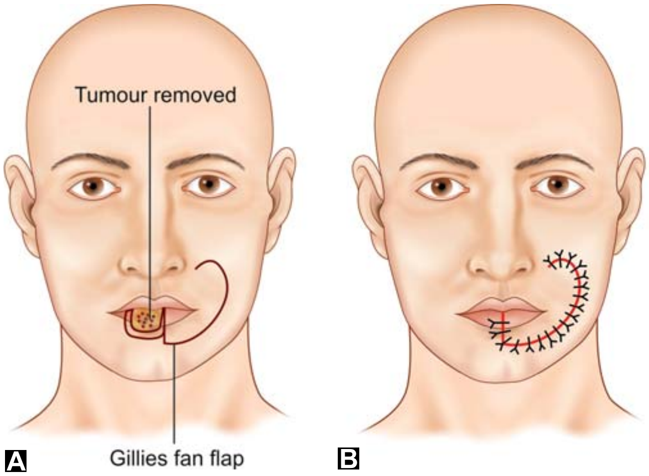

Figs 3.59A and B: *Gillies fan flap*.

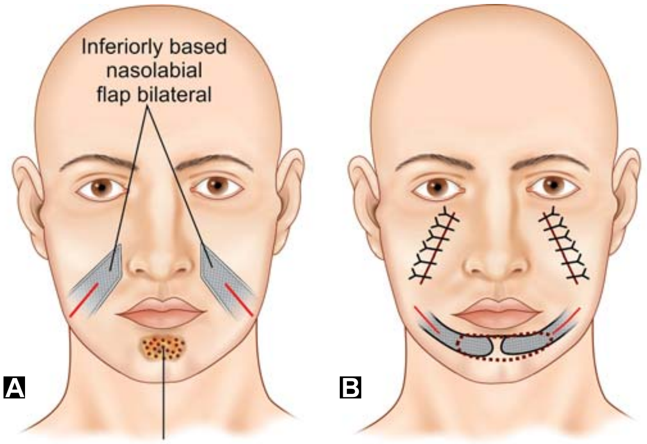

Tumour in the floor of the mouth removed leaving a defect

Flap rotated into the defect in the floor of the mouth

Figs 3.60A and B: Nasolabial flap used in extensive lower lip defect or over the defect after wide excision of a tumour in the floor of the mouth.

Prognosis

- Prognosis is good, 5 years survival is 70%.
- Lip has best prognosis.
- Nerve involvement, fixation, nodal spread, upper lip or commissure lesions, age less than 40 years—are poor prognostic factors.

TONGUE

Anatomy of Tongue

Tongue is a muscular organ located in the floor of the mouth.

Parts

1. *Tip*: Anterior free end lies behind the upper incisor teeth.
2. *Root*: Attached to the mandible above and hyoid bone below.
3. *Body*: Dorsal surface is rough due to papillae; and is divided into anterior 2/3rd (oral part) and posterior 1/3rd (pharyngeal part) by *sulcus terminalis*. Ventral surface is smooth, has a median fold, *'frenulum linguae'* and deep lingual vein on either side.

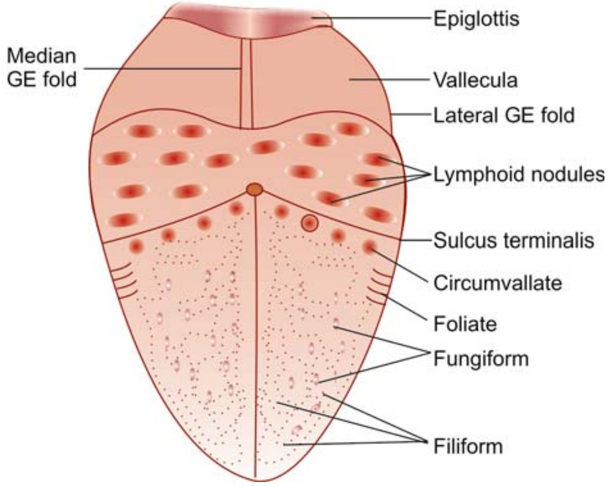

Fig. 3.61: Anatomy of the tongue showing parts and papillae.

Papillae

1. *Vallate*—large, located in front of sulcus terminalis.
2. *Fungiform*—lies over the tip and margin of the tongue.
3. *Filiform*—lies over the dorsum of tongue, gives the velvety appearance—commonest.
4. *Foliate*—over the margin.

Muscles of Tongue

a. Intrinsic muscle: Superior and inferior longitudinal, transverse and vertical.
b. Extrinsic muscle: Genioglossus, hyoglossus, styloglossus, palatoglossus.

Blood supply is from lingual artery, a branch of external carotid artery. Venous drainage by deep lingual vein which drains into fascial vein or internal jugular vein.

Movements of the tongue

- Forward protrusion—genioglossus—most commonly affected by carcinoma
- Backward movement—styloglossus
- Elevation of tongue—palatoglossus
- Depression of tongue—hyoglossus

Lymphatic Drainage of the Tongue

- Tip of tongue drains into submental lymph nodes.
- Lateral margin drains to submandibular lymph nodes and into upper deep cervical lymph nodes. Many lymphatic vessels pass as subperiosteal lymphatics of mandible. So carcinoma can involve the bone through this route.
- Lymphatics in the midline of tongue freely cross communicate with each other and so spread of malignancy can occur to both side neck lymph nodes.

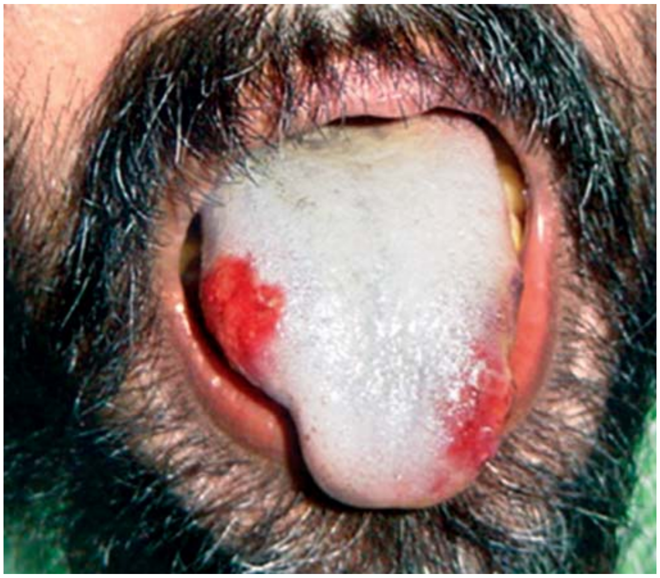

Fig. 3.62: Severe superficial glossitis.

- Lymphatics from posterior third of tongue drain into pharyngeal group of lymph nodes, as well as to the upper deep cervical lymph nodes. Early spread to the pharyngeal lymph nodes from carcinoma of posterior third of tongue has a poor prognosis.

Lymphatic vessels are named as:

1. Apical vessels.
2. Central vessels.
3. Marginal vessels.
4. Basal vessels.

Development and Nerve Supply of the Tongue

- *Anterior 2/3rd* develop from first branchial arch through two lingual swellings and one tuberculum impar. It is supplied by lingual nerve for general sensation and by chorda tympani for taste sensation.
- *Posterior 1/3rd* develops from third arch from cranial half of hypobranchial eminence. It is supplied by glossopharyngeal nerve for both general and taste sensations.
- *Posterior most part* develops from the fourth arch. It is supplied by vagus nerve (internal laryngeal nerve).
- *Muscles of the tongue* are derived from occipital myotomes and are supplied by hypoglossal nerve except palatoglossus, which is supplied by cranial part of accessory nerve.

TONGUE ULCERS

Differential diagnosis for tongue ulcers

- Dental ulcers—painful
- Aphthous ulcers—painful
- Ulcers in lichen planus—painless
- Syphilitic ulcers—painless
- Tuberculous ulcers—painful
- Malignant ulcers—painless

Dental Ulcer

It is common on sides of tongue due to sharp tooth, denture, and broken tooth. Usually it is acute painful, self limiting ulcer. Occasionally repeated trauma form an indolent chronic ulcer which mimic carcinoma; it should be excised to rule out carcinoma and to cure the ulcer.

Aphthous Ulcer

It can be (1) *Minor* aphthous ulcer, common in menstruating women as a crop with painful, round, yellow based ulcer with red margin. It regresses spontaneously in 2 weeks. (2) *Major* aphthous ulcer, large and deep which often becomes chronic and takes more time to subside with a scar. Chlorhexidine gluconate, local application of triamcinolone acetate, choline salicylate gel are different local applicants used to promote healing. (3) *Behcet's syndrome* is genital ulcer, conjunctival ulcer and multiple oral ulcers. *Reiter's syndrome* is urethritis, arthritis, periarteritis nodosa, conjunctivitis, and oral ulcers. (4) Herpetiform aphthous ulcer is not due to herpes simplex. They are small, 1-2 mm diameter ulcers in crops which heal by usual drugs mentioned above.

Syphilitic Ulcer

Extragenital chancre often occurs in tongue in *primary syphilis* which is painless with shotty, submental and submandibular lymph nodes. *In secondary syphilis*, multiple shallow snail track ulcers in the margins and undersurface; mucous patches on the tongue and fauces; *Hutchinson's condyloma* wart in midline of tongue can occur. *In tertiary syphilis*, gummatous ulcer occurs in anterior 2/3rd of tongue as a deep punched out painless ulcer as gumma with wash leather slough. Endarteritis is the cause for the punched out look. *Interstitial glossitis* with loss of papillae causes longitudinally fissured bald lobulated tongue in tertiary syphilis. In carcinoma arising from syphilitic ulcer, RT is *questionable* as blood supply is precarious due to endarteritis; RT further compromises it leading to tongue necrosis.

Tuberculous Ulcer

It is undermined shallow, often multiple, painful ulcer. Ulcer can occur in margins, tip or anterior 2/3rd of tongue. Neck nodes may be involved. Associated tuberculous larynx and lung may be present.

Herpetic Lingual Ulcer

It is involvement of lingual nerve presenting as acute neuralgia with vesicles which form multiple superficial painful ulcers.

Other Ulcers

Multiple ulcers in smokers due to glossitis (*smoker's ulcer*), ulcers due to vasculitis, eosinophilic granuloma. Post-pertussis ulcer in whooping cough occurs on upper part of frenum linguae and under the tip of tongue.

Macroglossia (Megaloglossia/Pachyglossia)

It is a disorder in which the tongue is larger than normal. It is commonly painless, diffuse enlargement of the tongue. Macroglossia is *usually* caused by an increase in the amount (volume) of tissue on the tongue, rather than by a growth, such as a tumour. It is often seen in haemangioma, lymphangioma, muscular macroglossia (in cretins), acromegaly, *Beckwith-Wiedemann syndrome* (hypoglycaemia, abdominal wall defects, Wilms' tumour, macroglossia, adrenal tumour), Down's syndrome, mucopolysaccharidoses, primary amyloidosis, occasionally plexiform neurofibromatosis. Often it causes functional and cosmetic problems.

Causes

1. *Lymphangioma*—soft, painless enlarged tongue with ulcers—bilateral; prevents closure of lip and jaw.
2. *Haemangioma*—soft, fluctuant, compressible, bleeding, red/blue lesion
    i. Both are treated by sclerotherapy (ethonalamine oleate)/partial excision
    ii. Angiogram/MR angiogram is a must in haemangioma
    iii. Ligation of lingual artery/ECA on both sides may be needed in large lesions

    Note:
    - Varicosities of sublingual veins mimic haemangioma.
    - Often causes *unilateral* enlargement of tongue.
    - Pressure on alveolus causes spacing of teeth and incisor deformity.
    - Combination of haemangioma and lymphangioma can occur.
3. *Neurofibroma*—partial excision is done
4. *Tongue muscular hypertrophy*: Partial excision is done.

- Elongation is corrected by wedge resection of anterior part of extra tongue from midline.
- Vertical thickening is rectified by slice cutting of lateral margins without injuring lingual artery and nerve.

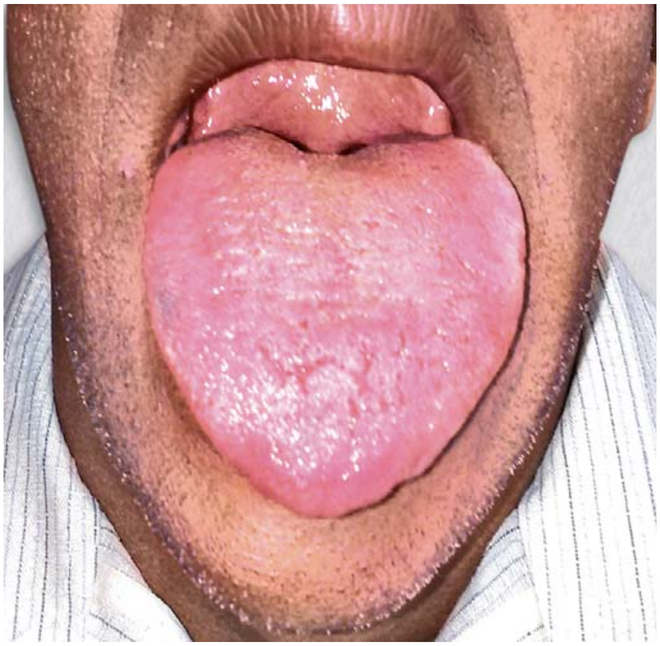

Fig. 3.63: Macroglossia in Down's syndrome patient (*Courtesy*: Professor Kishore Chandra Prasad, ENT Surgeon, Dr Sampath, ENT Surgeon, Mangalore).

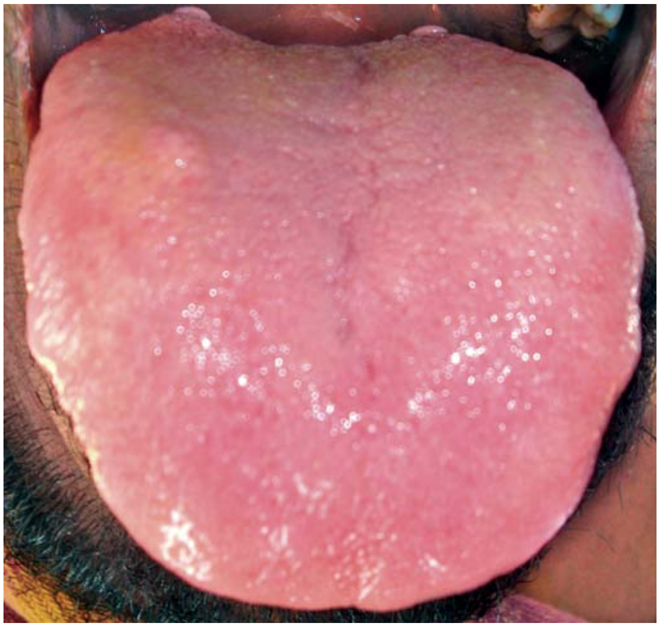

Fig. 3.64: Macroglossia.

BENIGN TUMOURS OF TONGUE

Benign tumours of the tongue

- Papilloma
- Fibroepithelial polyp
- Haemangioma and lymphangioma
- Neurofibroma
- Lipoma
- Granular cell myoblastoma

Fibroepithelial Polyp

It is due to repeated trauma at one place may be due to teeth (incisor) forming a thickened submucous scar which gets pulled out like a stalked polyp due to sucking and swallowing mechanism. It commonly enters the gap between the lower incisor teeth.

Granular Cell Myoblastoma

It is a benign noncapsulated firm mobile mass in the tongue showing pseudoepithelial hyperplasia of mucosa of tongue with eosinophilic granular cells in deeper plane. It is often mistakenly diagnosed as carcinoma of tongue. It is treated by excision.

TONGUE FISSURE

- Congenital fissures are transverse which run laterally from midline with normal papillae in between. Candida infection can occur on this.
- Syphilitic fissures are deep bald and longitudinal.

GLOSSITIS

Median Rhomboid Glossitis

- It is smooth, lobulated, triangular firm patch anterior to foramen caecum of tongue in midline with deeper colour.
- Candida infection can occur in it.

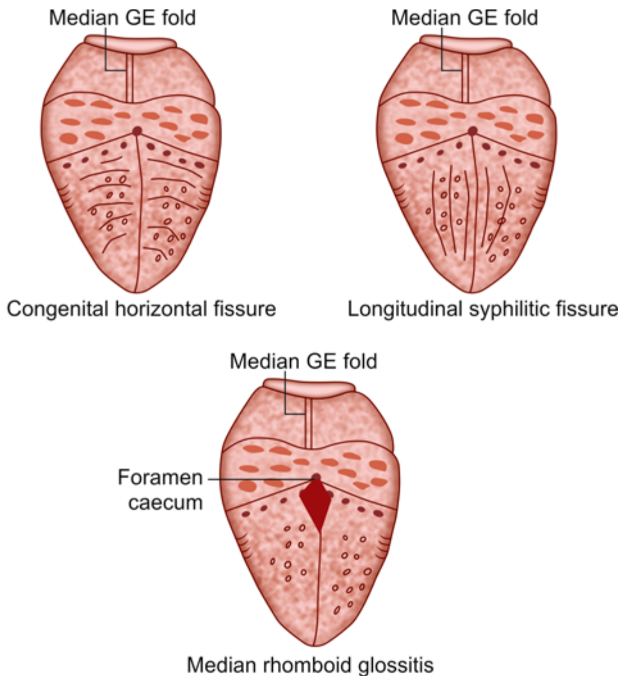
Median GE fold
Median GE fold
Congenital horizontal fissure
Longitudinal syphilitic fissure
Median GE fold
Foramen caecum
Median rhomboid glossitis

Fig. 3.65: Congenital fissures of tongue are transverse; syphilitic fissures are longitudinal. Median rhomboid glossitis is persistent tuberculum impar.

- It mimics carcinoma. Biopsy rules out malignancy. Carcinoma is uncommon in midline.

Glossitis Migrans (Geographic Tongue)

- It begins as benign small red patches with white furred margin which spread and recede in an irregular way to appear as fresh patches.
- White margin contains keratinized epithelium and inflammatory cells over filiform papillae.
- It is often seen in patients with congenital heart diseases and acute gastrointestinal diseases. Etiology is unknown.

Other Glossitis

- *Hunter's glossitis* is seen in pernicious anaemia.
- *Hairy tongue* is overgrowth of filiform papillae with black/brown stain on it due to bacteria, fungi, tobacco or drugs. There are no hairs. It is a misnomer. Cessation of causative agent, mechanical scraping, cleaning are the treatment methods.
- Agranulocytosis glossitis.
- Nonspecific glossitis.
- *Pellagra glossitis*, due to deficiency of nicotinamide (B12).
- Chronic superficial glossitis in malnutrition, iron and vitamin B deficiencies.

TONGUE TIE

- It is short, thick, fibrous frenum linguae.
- During protrusion lateral margin and tip of the tongue is everted with dorsal mid part heaping.
- It causes speech defect, difficulty in cleaning the inner part of lower teeth.
- It is treated surgically under local anaesthesia (or general in child). Tongue lifted upwards with a stay suture at the tip; fibrous frenum is divided using fine scissor; linear wound is closed longitudinally using fine catgut from the tip of the tongue towards the margin of the floor of the mouth.

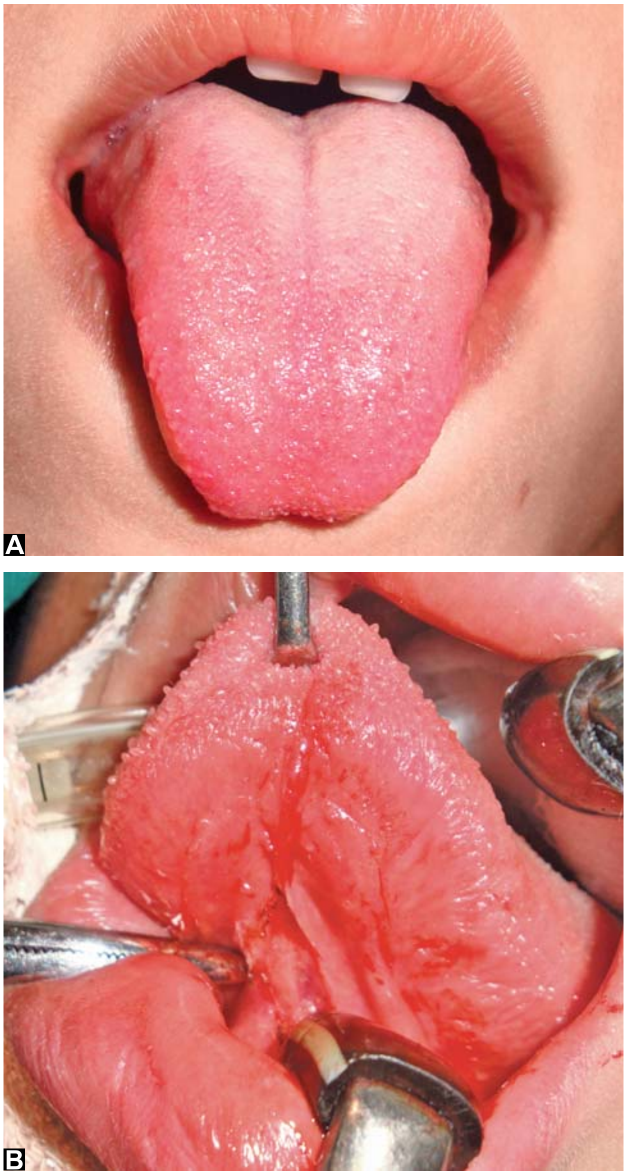

Figs 3.66A and B: Tongue tie—on table photo.

CARCINOMA TONGUE

Incidence is equal in both sexes. Presently its incidence is increasing in females due to increase in number of female smokers.

Aetiology

- Leukoplakia.
- Erythroplakia.
- All 'S's (as mentioned in leukoplakia).
- Premalignant conditions mentioned earlier.

Hope sees the invisible, feels the intangible and achieves the impossible.

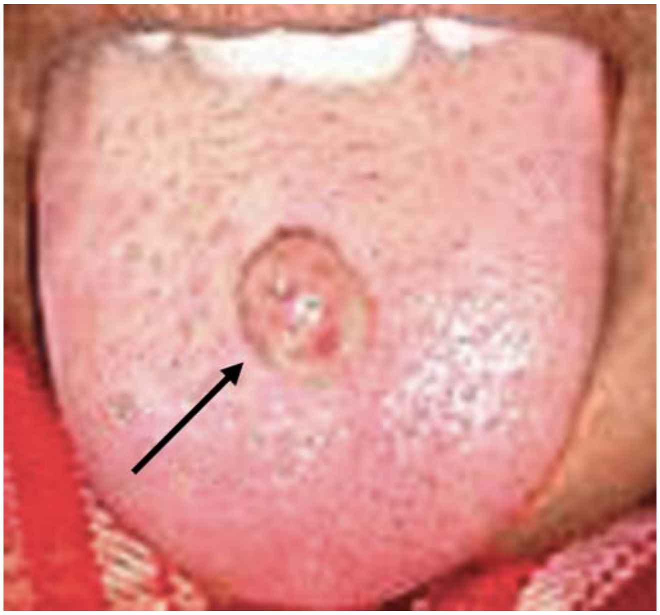

Fig. 3.67: Papilloma tongue. It is a premalignant condition. It requires surgical excision.

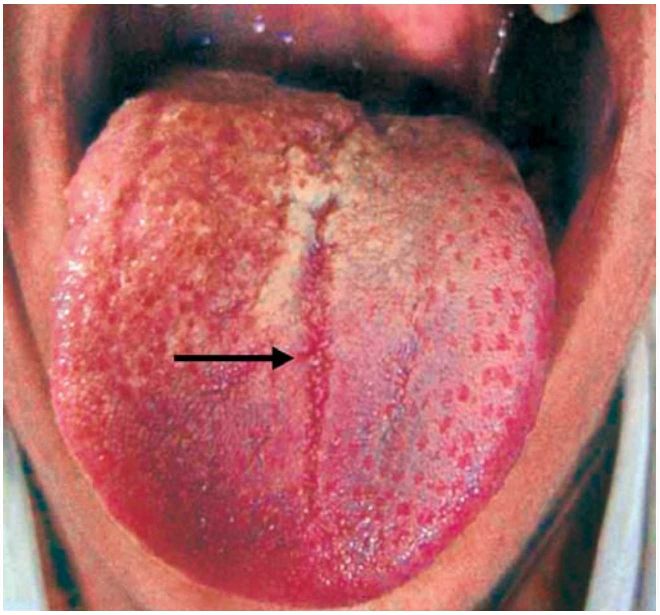

Fig. 3.69: Tongue showing fissuring. Fissuring is one of the presentations of carcinoma tongue.

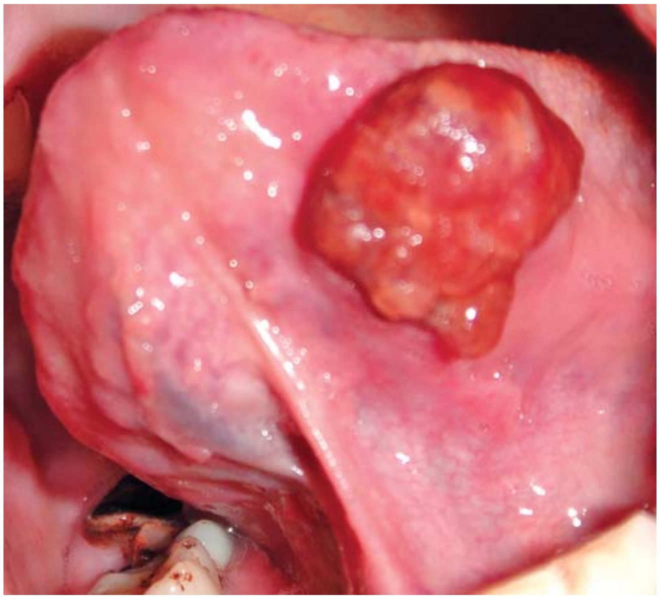

Fig. 3.68: Pyogenic granuloma of tongue under surface.

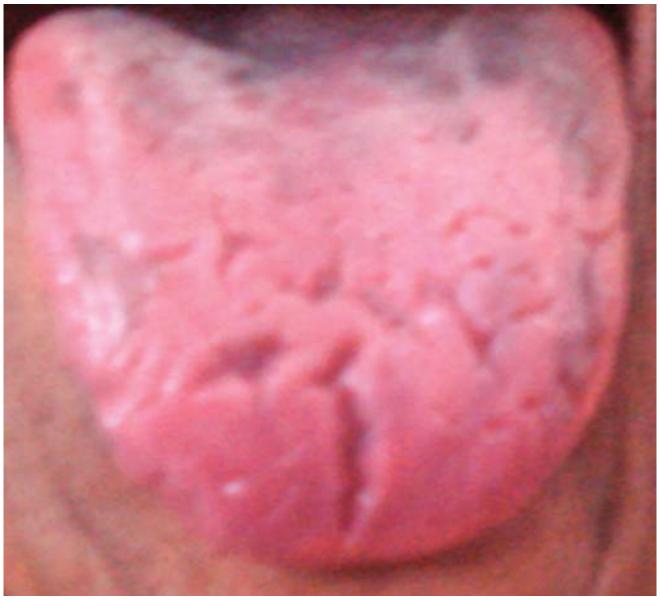

Fig. 3.70: Horizontal fissures of tongue.

Types

Gross

1. Papillary.
2. Ulcerative or ulceroproliferative 60%.
3. Fissure with induration.
4. Lobulated, indurated mass—*frozen tongue.*

Histologically

1. Squamous cell carcinoma—commonest.
2. Adenocarcinoma, may arise from minor salivary glands or mucous glands.
3. Melanomas.
4. Transitional cell carcinoma and lymphoepithelioma rarely can occur in posterior 1/3rd of tongue.

Sites

1. Lateral margin—commonest—47-50%.
2. Posterior third—20%.
3. Dorsum—6.5%.
4. Ventral surface—9%.
5. Tip—10%.

Clinical Features

- Painless ulcer/swelling in the tongue which later may become painful. *Pain* in the tongue due to infection or

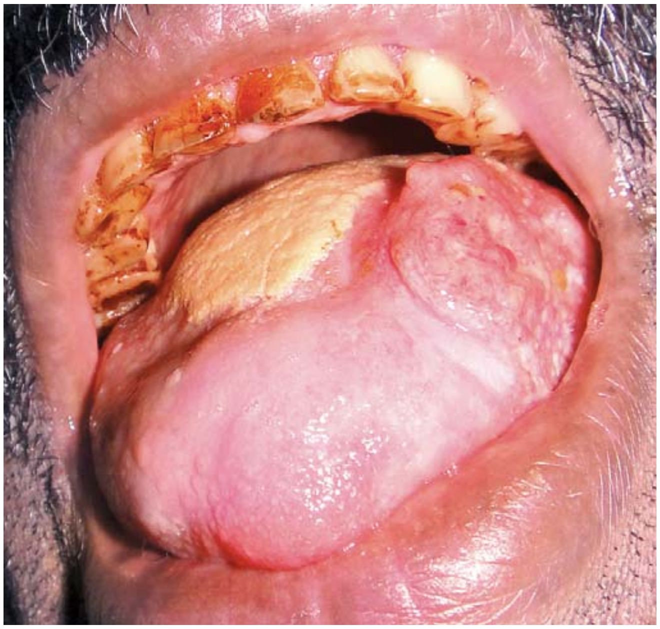

Fig. 3.71: Carcinoma tongue with leukoplakia.

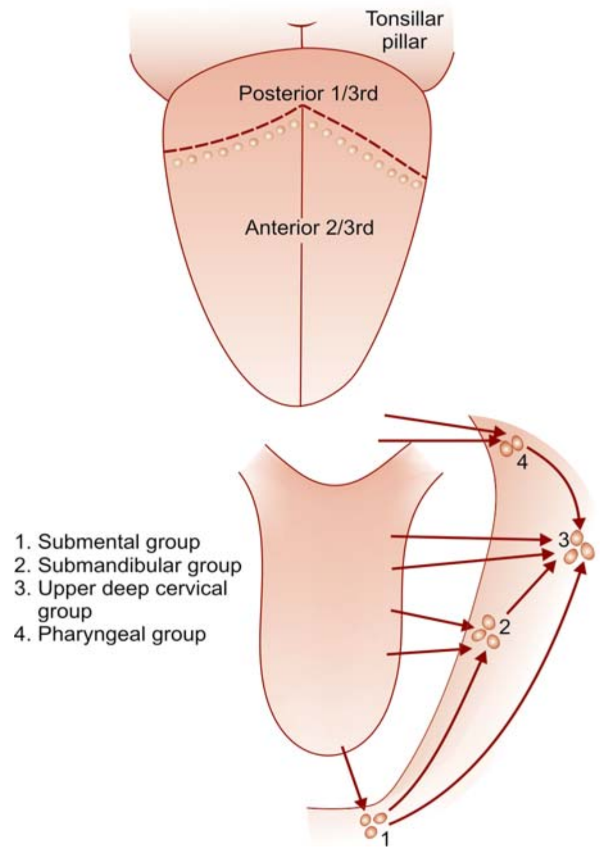
Tonsillar pillar
Posterior 1/3rd
Anterior 2/3rd
1. Submental group
2. Submandibular group
3. Upper deep cervical group
4. Pharyngeal group
4
3
2
1

Fig. 3.72: Anatomy and lymphatic drainage of tongue.

ulceration or due to the involvement of lingual nerve (pain is referred to ear). Pain on swallowing, in case of carcinoma of posterior third of tongue.

- *Excessive salivation*. Saliva is often blood stained.

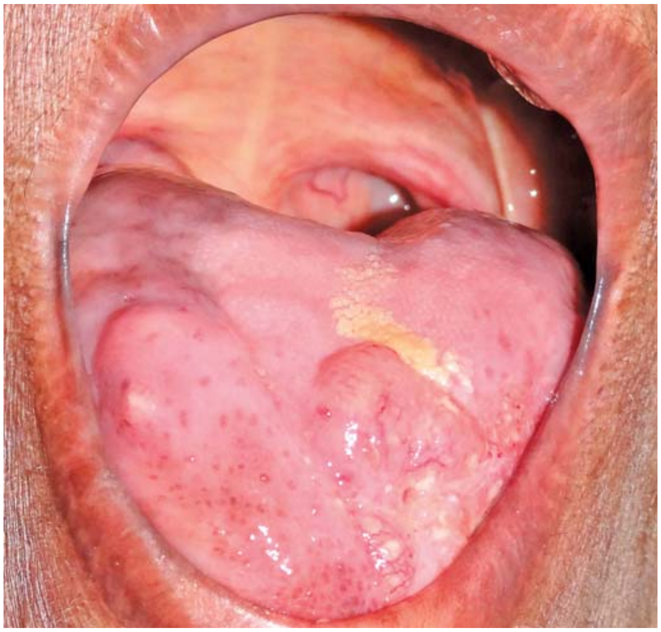

Fig. 3.73: *Frozen tongue* is feature of carcinoma tongue with ankyloglossia.

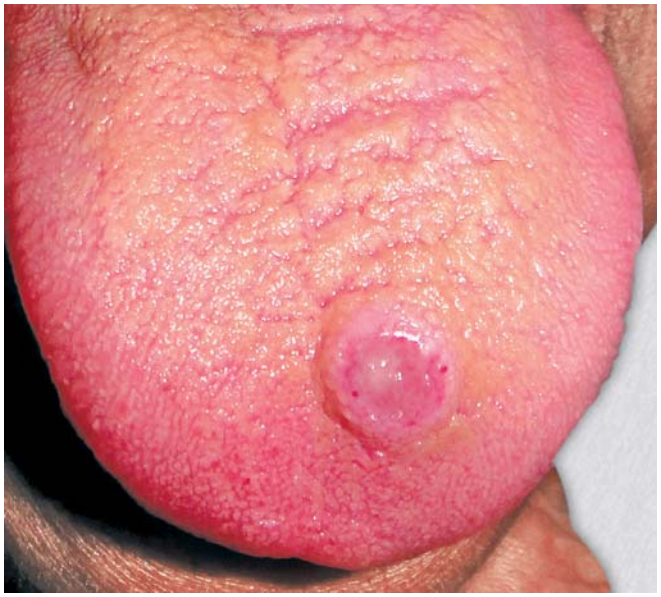

Fig. 3.74: Carcinoma tip of tongue. Well-circumscribed raised edge is seen.

- *Dysphagia* either due to fixed tongue or due to the involvement of genioglossus or growth in the posterior third of the tongue.
- *Visible ulcer* in anterior two-thirds of tongue. Ulcer can bleed on touch; edge, base and surrounding areas are indurated. Often indurated area is much more extensive than the primary tumour (it is also common in carcinoma penis). Edge is everted commonly. Ulcer may cross the midline; may extend into the floor of the mouth/alveolus/mandible. Growth or ulcer in posterior third, is usually not visible.

Efforts and energies should be consistent, constructive and compassionate.

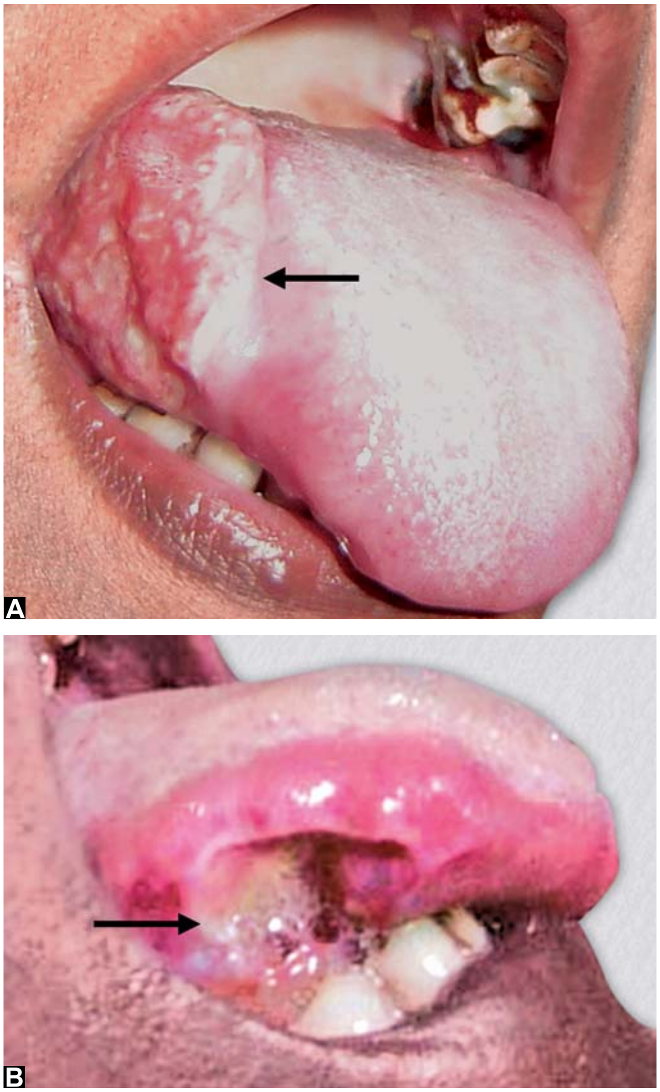

Figs 3.75A and B: Carcinoma tongue lateral margin (commonest site).

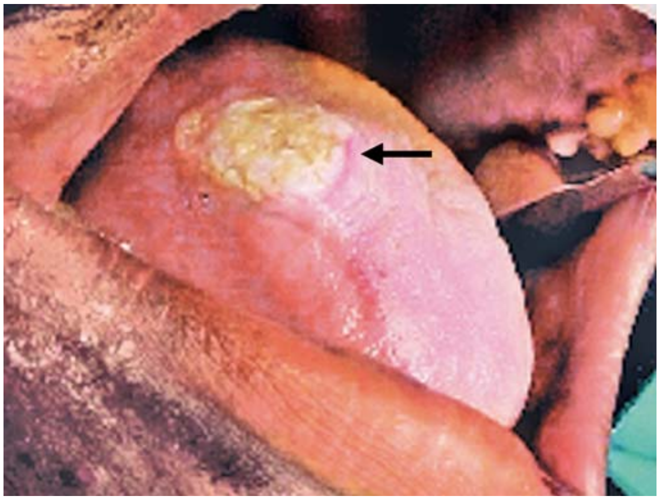

Fig. 3.76: Carcinoma tongue.

- *Ankyloglossia*—involvement of muscles of the tongue.
- Movements of the tongue especially forward protrusion is commonly affected.
- *Inability to articulate*.
- *Foetor* (Halitosis). Due to infection and necrosis in the oral cavity. It is due to release of ammonia, butyric acid and mercaptan by tumour cells.
- *Change in voice*. Occurs in posterior third tumours. Tumour in posterior third area is more aggressive.
- Palpable *lymph nodes* in the neck which are hard, nodular and get fixed to underlying tissues in advanced stages.
- Features of bronchopneumonia—due to aspiration during lying down/sleeping mainly to lower segment of lung.

Spread of Carcinoma Tongue

Local spread:

In case of anterior two-thirds of tongue, the spread occurs to genioglossus muscle, floor of the mouth, opposite side and mandible. In case of posterior third of tongue it spreads locally to tonsil, side of pharynx, soft palate, epiglottis, larynx and cervical spine.

Lymphatic spread:

From tip of tongue it spreads to submental nodes. From lateral margin it spreads to submandibular lymph nodes and later to deep cervical lymph nodes. Lymphatics in the tongue are freely communicating, and so involvement of bilateral neck lymph nodes is common. From posterior third it spreads to pharyngeal nodes and upper deep cervical lymph nodes.

Note:

- Any tongue lesion more than 4 mm depth has 30% metastasis to lymph nodes.
- Among oral cancers, carcinoma tongue is more aggressive, rapidly growing tumour with high potential for lymph node spread.
- Bilateral neck lymph node spread is common due to crossing of lymph vessels in tongue.
- Clearance of lingual lymph node which is located between tongue and submandibular lymph glands has to be done which may contain micrometastasis.
- Carcinoma tongue has got highest incidence of nodal spread.

Investigations

1. Edge biopsy.
2. FNAC of lymph nodes.
3. Indirect and direct laryngoscopy to see posterior third growth.

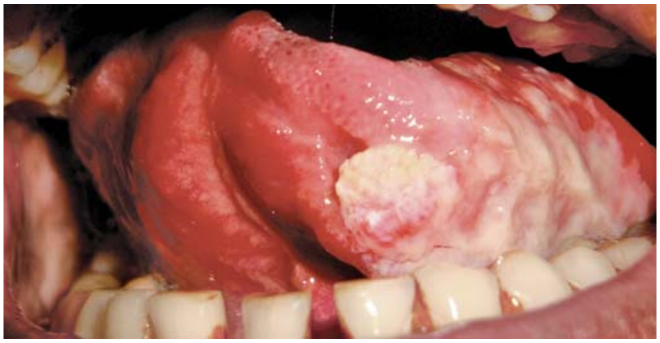

Fig. 3.77: Carcinoma left lateral margin of the tongue—ulcerative lesion. Note the proliferation. Induration extends for beyond the visible lesion.

4. CT scan to see the extension of posterior third growth, or to see the status of advanced secondaries. *MRI* is also very useful to assess the extent of primary tumour.
5. Chest X-ray to see bronchopneumonia.
6. Orthopantomogram.

***Note*:**
Staging is same as carcinoma cheek.

Treatment

Surgery, radiotherapy, chemotherapy.

Surgery

- *Wide excision* with 1 cm clearance in margin and depth is done in tumour less than 1 cm in size or in carcinoma in situ. Laser (CO_2 / diode) can be used.
- Tumour between 1-2 cm in size, *partial glossectomy* is done with 2 cm clearance from the margin with removal of 1/3rd of anterior two-third of the tongue.

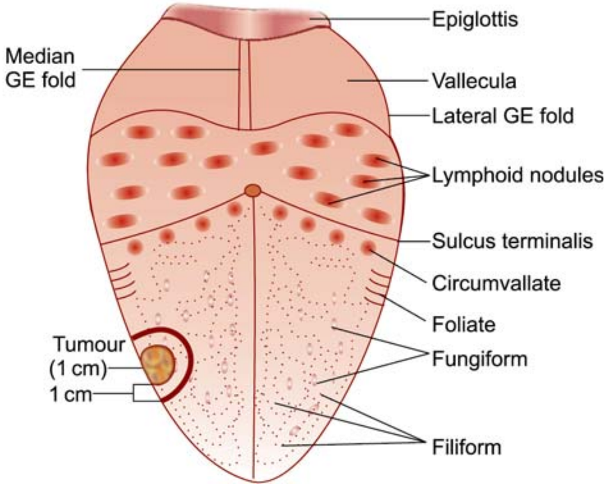
Epiglottis
Median GE fold
Vallecula
Lateral GE fold
Lymphoid nodules
Sulcus terminalis
Circumvallate
Foliate
Fungiform
Tumour (1 cm)
1 cm
Filiform

Fig. 3.78: Wide excision is done in small lateral margin tumour of 1 cm size with 1 cm clearance.

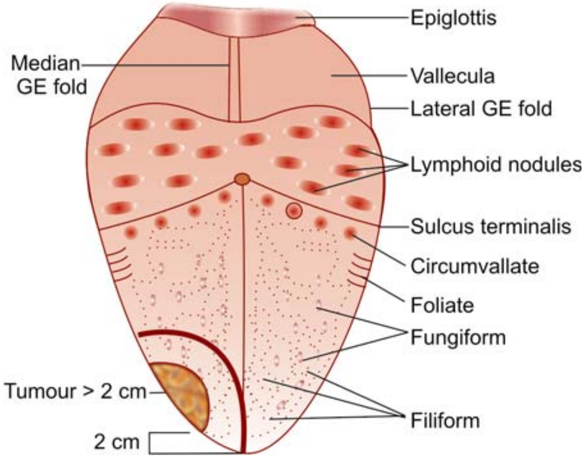
Epiglottis
Median GE fold
Vallecula
Lateral GE fold
Lymphoid nodules
Sulcus terminalis
Circumvallate
Foliate
Fungiform
Tumour > 2 cm
Filiform
2 cm

Fig. 3.79: Tumour between 1-2 cm size is treated by partial glossectomy with 2 cm clearance.

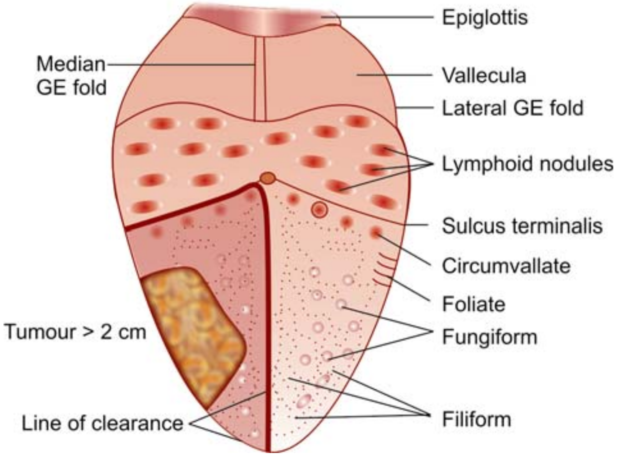
Epiglottis
Median GE fold
Vallecula
Lateral GE fold
Lymphoid nodules
Sulcus terminalis
Circumvallate
Foliate
Tumour > 2 cm
Fungiform
Line of clearance
Filiform

Fig. 3.80: Tumour larger than 2 cm requires hemiglossectomy.

- Tumour larger than 2 cm, *hemiglossectomy* is done with removal of anterior 2/3rd of tongue on one side up to sulcus terminalis.
    - *Raw area* in these procedures can be left alone when area is wide allowing it to granulate and heal by epithelialisation. If area is small like in wide excision it can be closed by primary suturing. Wide raw area can also be covered with PMMF or *quilted split-skin-graft*.
- Larger primary tumour can be given preoperative radiotherapy, then later hemiglossectomy is done.
- Same side palpable, mobile lymph nodes are removed by *radical neck block dissection*.
- *Bilateral mobile lymph nodes are dealt with one side radical block and other side functional block dissection with essentially retaining internal jugular vein* (on opposite side) to maintain the cerebral venous blood flow. Other option is doing same side radical neck dissection and on opposite side supraomohyoid block dissection.
- *Wide excision* is done when growth is in *the tip* of the tongue.
- Posterior third growth can be approached by lip split and mandible resection, so as to have *total glossectomy—Kocher's approach*. It is not done commonly as it carries significant morbidity and mortality due to difficulty in speech, swallowing, aspiration, sepsis.
- When mandible is involved *hemimandibulectomy* is done.
- The procedure that involves wide excision or hemiglossectomy, hemimandibulectomy and radical neck dissection together is called as *Commando Operation*.
- Reconstruction of tongue and other area after surgery: By deltopectoral flap, forehead flap, pectoralis major muscle flap, skin grafting.
- Prophylactic block dissection is becoming popular at present.

Postoperative management

- ❖ Control of infection and oedema
- ❖ Regular mouth wash
- ❖ Maintaining the airway
- ❖ Prevention of aspiration
- ❖ Nutrition (through nasogastric tube commonly/TPN often)

He who believes, is strong; he who doubts, is weak.

Radiotherapy

1. In small primary tumour—curative radiotherapy (Brachytherapy using caesium or iridium^{192} needles).
2. Large primary tumour—initial radiotherapy is given to reduce the tumour size so that the resection will be better later.
3. Advanced primary as well as secondaries in the neck can be controlled by palliative external radiotherapy.
4. Postoperative radiotherapy is given in large tumours to reduce the chances of relapse.
5. In case of growths in the posterior third of tongue, radiotherapy is of curative as well as palliative mode.

Complications of radiotherapy

- Loss of sensation like taste
- Trismus and ankyloglossia
- Infection
- Pharyngeal and laryngeal oedema
- Dermatitis and skin infection

Chemotherapy

Given in postoperative period and also for palliation.

Price-Hill regimen is commonly used. Drugs are methotrexate, vincristine, adriamycin, bleomycin and mercaptopurine.

It is either given intraarterially, as regional chemotherapy through external carotid artery using arterial pump or through IV. It can also be given orally.

Complications of chemotherapy

- Megaloblastic anaemia
- Bone marrow suppression
- Alopecia
- Sepsis

For melanoma, Melphalan and DTIC are used.

Anterior chemotherapy (preoperative) is becoming popular to downstage the tumour.

Terminal events

- Inhalational bronchopneumonia
- Haemorrhage from erosion of lingual artery. In posterior third of the tongue, erosion of internal carotid artery can occur
- Cancer cachexia
- Asphyxia due to pressure on air passages or due to oedema glottis

Prognosis

Five-year survival for females is 50%, for males is 25%.

Nodal prognostic factors

- Positive histology in node reduces the survival
- Level III and IV has poor prognosis
- Bilateral/ contralateral nodes carry poor prognosis
- Extracapsular spread/size > 3 cm carry poor survival
- > 3 in number of nodes involved is poor sign

Prognostic Factors

- Size of the tumour > 4 cm caries poor prognosis.
- Site of tumour (posterior third has got poor prognosis).
- Tumour crossing the midline.
- Lymph nodes status.
- Differentiation.
- Bone involvement.

CARCINOMA OF POSTERIOR ONE-THIRD/BASE OF THE TONGUE

- Lesion may remain asymptomatic for long time.
- Clinically may be missed easily.
- Earlier symptoms are features mimicking sore-throat and throat discomfort.
- Dysphagia and change in voice (*hot potato voice*) occurs later.
- Referred pain in the ear, bleeding from mouth, visible mass in posterior third of tongue is late local features.
- Induration on palpation in posterior third tongue is diagnostic of the carcinoma.
- As posterior third tongue has got abundant lymphatics which cross communicates on either side, lymph node spread is common (70%). Bilateral nodal spread is common. Massive nodes and involvement of jugulodigastric node are also common.
- Infiltration into the tongue muscles like genioglossus, epiglottis, pre-epiglottic space, *tonsillar pillars* and hypopharynx are common.
- Carcinoma posterior third of the tongue is often poorly differentiated and so caries poor prognosis.
- Blood spread can occur into bones, liver and lungs in posterior third cancers.
- Palpation *under anaesthesia* gives better idea about the tumour, its spread and also allows the biopsy.
- Presentation as unknown/occult primary and often with blood spread can occur.
- CT scan/MRI is always needed to plan the staging and therapy.
- T_1, T_2, N_0 and N_1 diseases are treated by surgical wide excision or often by total glossectomy using midline mandibulotomy incision (mandible split) with neck dissection on both sides (MRND one side). Postsurgery radiotherapy is needed if it is a poorly differentiated type or nodal status is more than N_1.
- Advanced lesions need palliative radiotherapy or chemotherapy.
- T_4 lesions are often treated by total glossectomy with laryngectomy and neck dissection but overall outcome is not good.
- In many centers primary curative radiotherapy is used.
- Lymphoepithelioma and transitional cell carcinoma can occur in posterior third tongue (rarely).

Carcinoma floor of the mouth

- It is usually aggressive tumour
- It is rare in India
- It is 2nd common site of oral carcinoma (SCC) in western countries
- It invades hyoglossus, mylohyoid, genioglossus and anterior mandible early
- Bilateral neck nodes are commonly involved
- Rim resection of mandible with wide excision of tumour with muscles and soft tissues and bilateral neck dissection is necessary
- Often visor anterior approach with anterior mandible resection followed by proper reconstruction with bone graft and plates is needed
- Postoperative radiotherapy and later chemotherapy is used to prevent recurrence
- Prognosis is poor and also has poor cosmetic results

Carcinoma alveolus

- It is squamous cell carcinoma arising from gums
- It is common in males
- It is common in India
- It is commonly due to tobacco/pan chewing
- Features and precipitating factors are similar to other oral carcinomas
- There will be invariable bone involvement by direct extension
- Nodal spread is also common
- Wide excision with mandibulectomy and block dissection of neck is the treatment

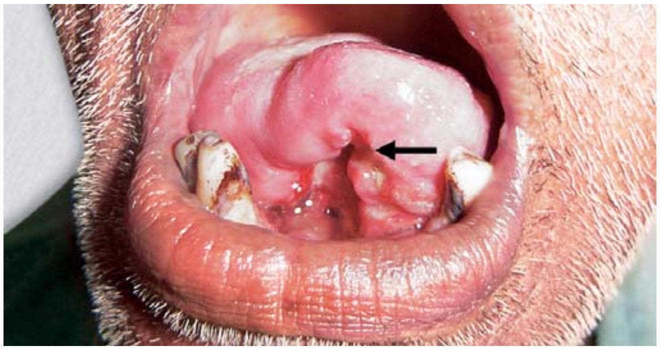

Fig. 3.81: Carcinoma of floor of the mouth. It has got poor prognosis.

NASOPHARYNGEAL CARCINOMA

- Nasopharynx lies above the level of the soft palate which divides it from oropharynx below.
- It is also called as post-nasal space or epipharynx. Eustachian tube opens on its anterolateral wall. Fossa of Rosenmuller is located above and behind the opening of the Eustachian tube as a small depression.
- Nasopharyngeal carcinoma is common in China and Mongolia. In India it is common in North-East region. It is commonly squamous cell carcinoma (85%). Lymphoma, minor salivary tumours and sarcoma are other malignancies that can occur rarely in nasopharynx.
- It can be of proliferative, ulcerative, and infiltrative types. Commonest site is fossa of Rosenmuller in lateral wall of pharynx. It is three times common in males.
- *HO's triangle* in supraclavicular fossa (bounded by medial and lateral ends of clavicle and point where neck meets the shoulder) is the site where metastatic nodes commonly exist in nasopharyngeal carcinoma.
- In 50% of cases nodal involvement is bilateral. Often cervical lymphadenopathy may be the first presentation.
- *Clinical features* may be nasal, otogenic, ophthalmoneurogenic (involving most of the cranial nerves with facial pain, squint, diplopia, exophthalmos, and ophthalmoplegia), jugular foramen syndrome (cranial nerves IX, X, XI spread), nodal spread and distant spread to bones, lungs and liver.
- Unilateral serous otitis media may be the only presentation.

Clinical Features

1. Epistaxis, nasal speech, post-nasal discharge and nasal obstruction.
2. Pain in the ear with unilateral deafness due to compression of Eustachian tube with fluid collection in the middle ear.
3. Elevation and immobility of soft palate on the same side.
4. Pain in the area of distribution of trigeminal nerve due to direct infiltration of the nerve at foramen lacerum.
5. Palpable secondaries in upper deep cervical lymph nodes (70%).

Trotter's triad

- Unilateral deafness
- Immobile elevated soft palate
- Pain in the distribution of trigeminal nerve

Differential Diagnosis

- Lymphoma.
- Lymphoepithelioma.
- Minor salivary gland tumour.

Investigations

- Biopsy from the primary site.
- FNAC from the neck lymph nodes.
- X-ray of the skull to visualise erosions.
- CT scan skull.

Histological Type

Squamous cell carcinoma.

Treatment

- External irradiation for primary. RT is the main modality of treatment.
- Radical block dissection of cervical lymph nodes.
- In N2a, N2b and N2c contralateral neck dissection is needed. Spinal accessory nerve is never preserved while doing block dissection in nasopharyngeal carcinoma.

- Chemotherapy: Methotrexate, Vincristine.
- Skull base surgeries are useful.

MAXILLARY TUMOURS

They are rare.

- Maxillary sinus is the commonest site for malignancy of paranasal sinuses. Ethmoids, frontal and sphenoids are next in order.
- It is common in people working in furniture industries, mustard gas industries, and leather industries. It is common in Bantus in South Africa where snuff with nickel and chromium is commonly used.

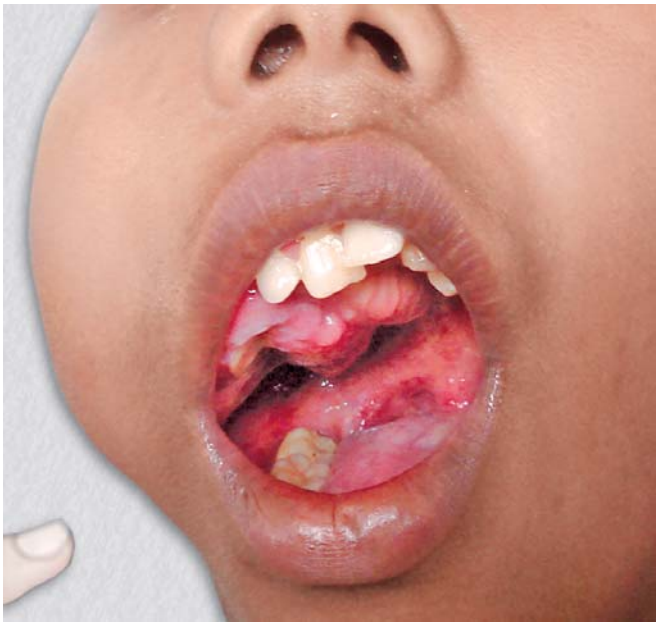

Fig. 3.82: Carcinoma maxillary sinus in a boy. Note the extension into the palate.

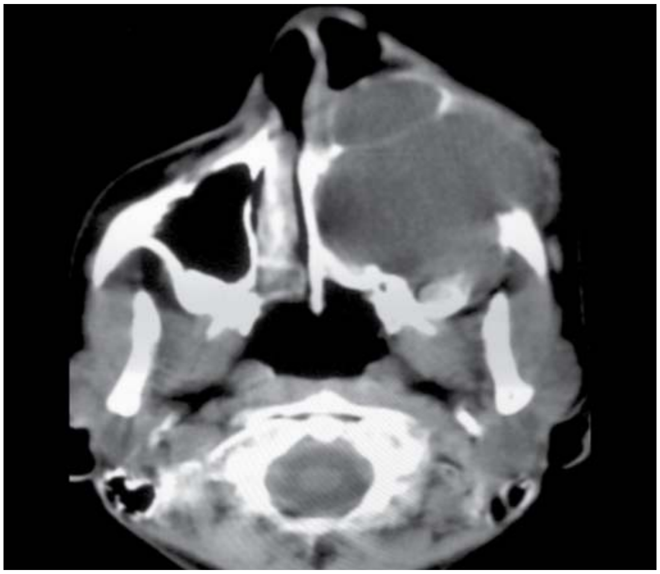

Fig. 3.83: CT picture of carcinoma of maxillary sinus (left-sided).

Types

1. Squamous cell carcinoma 80%.
2. Adenocarcinoma.
3. Transitional cell carcinoma.
4. Salivary gland tumours.
5. Sarcomas and melanoma.
6. Burkitt‘s lymphoma.

Behaviour and Presentation

1. Initially may be symptomless or may present with *epistaxis* or *features of chronic sinusitis*.
2. When it spreads to the floor, loosening of the teeth, necrosis, antro-oral fistula can occur.
3. Extension medially causes nasal block, fungation, nasal discharge, blockage of nasolacrimal duct (*epiphora*).
4. Extension anteriorly causes pain, anaesthesia and swelling in the cheek, ulceration and fungation in the skin of cheek.
5. Spread above into the orbit causes epiphora, diplopia, proptosis.
6. *Posterior spread is most dangerous* as it is not revealed easily. It causes postnasal discharge, pain, trismus, limitation of temporomandibular joint movement.
7. Involvement of upper deep cervical lymph nodes in later stage is common.

Differential Diagnosis

Chronic sinusitis.

Classification

1. ***Ohngren's classification***

An imaginary plane is drawn extending between medial canthus of eye and the angle of mandible. Growth situated above this plane is called as *suprastructural* which has got poor prognosis. Growth below this plane is called as *infrastructural* and has got better prognosis.

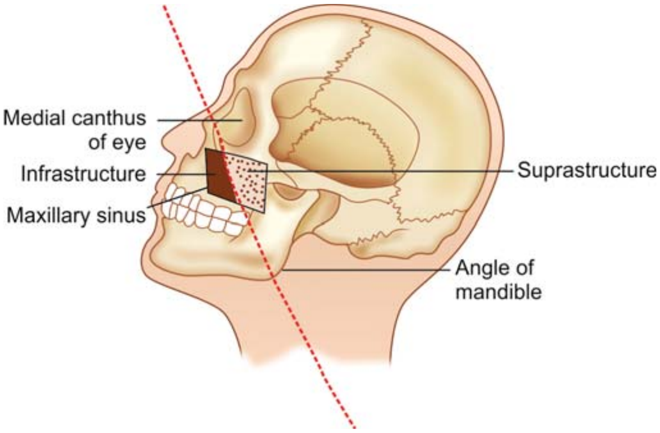

Fig. 3.84: An imaginary plane is drawn extending between medial canthus of eye and the angle of mandible and line in this plane is called as *Ohngren's line*.

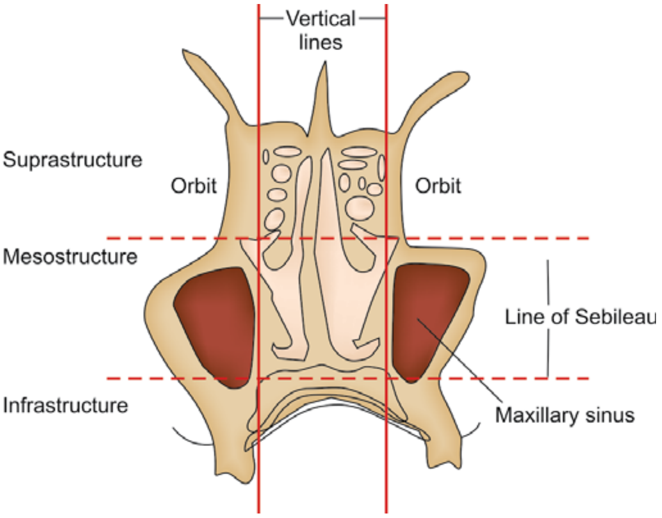

Fig. 3.85: Diagrammatic representation of Lederman's classification.

2. ***Lederman's classification***

Two horizontal lines are used, one passes through the floor of the orbit, another passes through the floor of the antra. These lines are called as *line of Sebileau*.

- *Suprastructure type*: In this type olfactory area of nose, ethmoidal, sphenoid, and frontal sinuses are involved.
- *Mesostructural type*: This involves maxillary sinus and nasal respiratory part.
- *Infrastructural type*: This type involves alveolar process.

Lederman's classification is further divided by two vertical lines over medial walls of the orbit to separate ethmoid sinuses and nasal fossa from maxillary sinuses.

TNM staging	*Staging*
• T_1 *Tumour* limited to antral mucosa • T_2 *Tumour* causing bone erosion/destruction including extension into hard palate or middle meatus of nose. • T_3 *Tumour* invading bone of posterior wall of maxillary sinus/skin of cheek/medial wall of orbit/infratemporal fossa/pterygoid plates/ethmoid sinuses. • T_4 *Tumour* invading orbital contents beyond the floor or medial wall including orbital apex/cribriform plate/base of skull/nasopharynx/sphenoid or ethmoidal sinuses	Stage I $T_1 N_0 M_0$ Stage II $T_2 N_0 M_0$ Stage III $T_3 N_0 M_0$, $T_1/T_2/T_3 N_1 M_0$ Stage IVA $T_4 N_0 M_0$, $T_4 N_1 M_0$ Stage IV B Any T $N_2 M_0$, Any T $N_3 M_0$ Stage IVC Any T Any N M_1

Diagnosis

- X-ray of the part—pacity of the involved sinus with destruction of bony walls is seen.
- CT scan—ideal method.

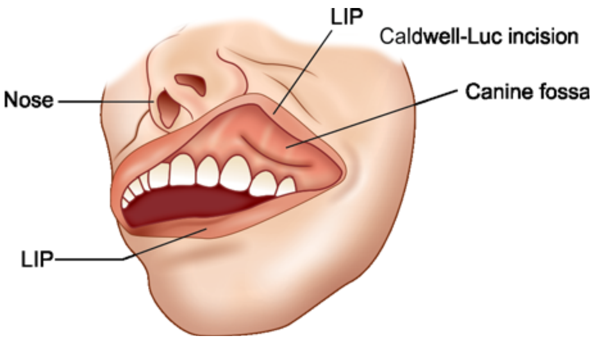

Fig. 3.86: Incision for *Caldwell-Luc operation* which is used for taking biopsy from maxillary tumour. Incision is not used for definitive therapy for carcinoma maxilla. Incision also used in benign conditions to approach maxillary sinus. Gingivobuccal mucosa is incised and mucoperiosteum is raised. Bone of canine fossa is cut to reach the maxillary antrum.

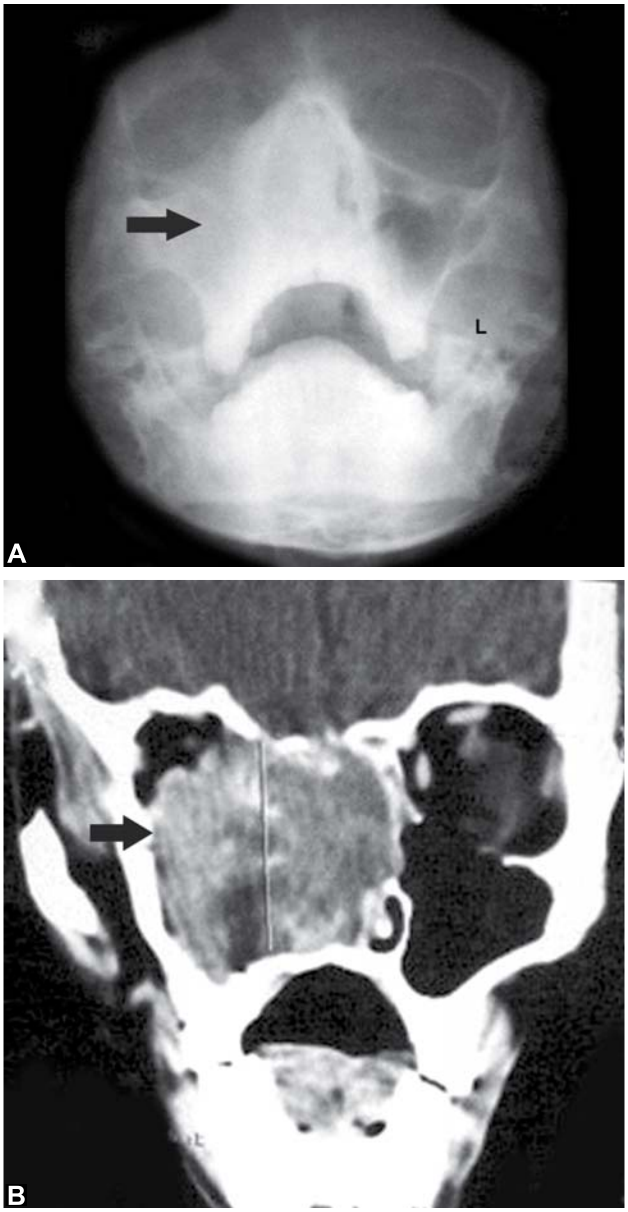

Figs 3.87A and B

Don't wait for your ship to come; swim out to it.

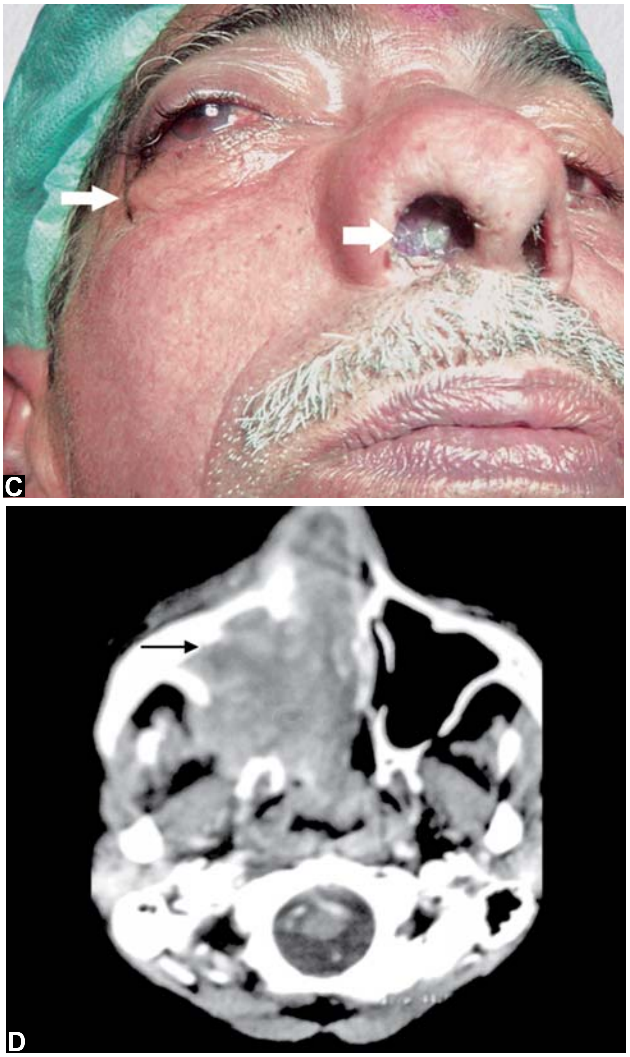

Figs 3.87C and D

Figs 3.87A to D: Carcinoma maxilla extending into the nose with X-ray and CT pictures. It is a fairly advanced growth.

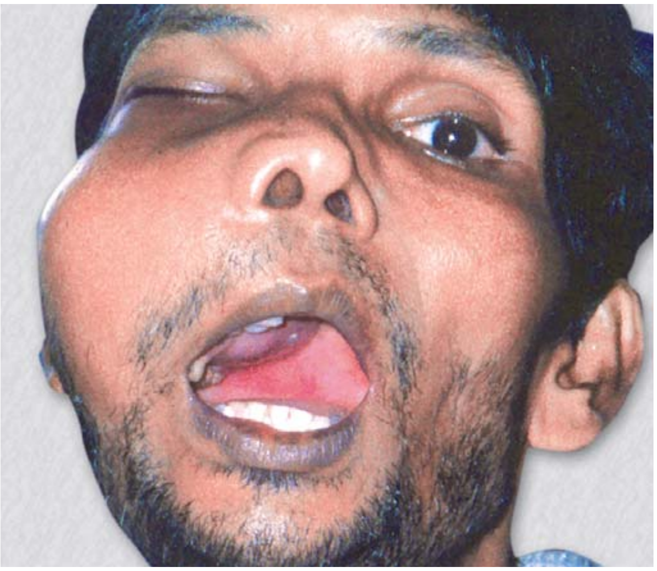

Fig. 3.88: Carcinoma maxillary antrum extending outwards on to all walls. An advanced malignancy.

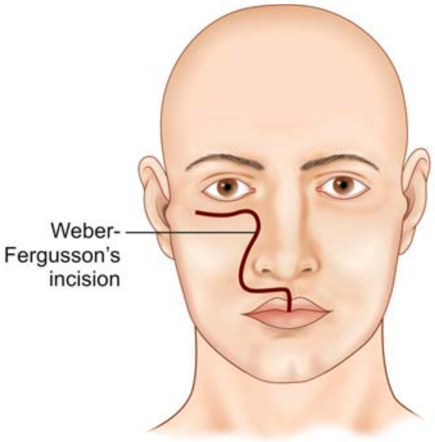

Fig. 3.89: *Weber-Fergusson's incision* used for total maxillectomy.

- Biopsy is done through nasal/oral or on early stage through Caldwell-Luc operation.
- Sinus endoscopy for detailed examination of sinus and for biopsy.

Treatment

- *Preoperative megavoltage radiotherapy* is given. After six weeks, *total maxillectomy* is done. Reconstruction of maxilla along with dental reconstruction is required.
- When lymph nodes are involved radical neck lymph nodes dissection is done.
- Postoperative radiotherapy and chemotherapy is given as an adjuvant therapy.
- Overall prognosis is 30-40%.

MALIGNANT TUMOURS OF TONSIL

- **Carcinoma of tonsil:** It is squamous cell carcinoma, similar to carcinoma cheek, but more aggressive with poor prognosis.
- **Carcinosarcoma** of tonsil.
- **Lymphoma**—NHL type.

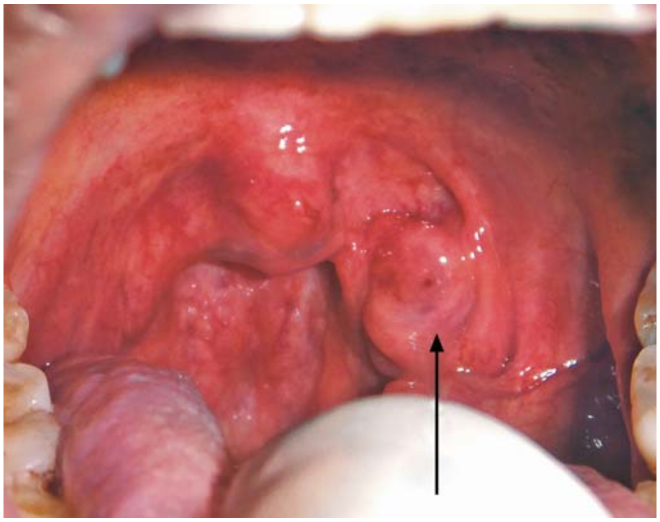

Fig. 3.90: Lymphoepithelioma of tonsil.

CARCINOMA HARD PALATE

- Minor salivary gland tumours are more common in palate.
- In males, *in reverse smokers* (*Churat* is rolled tobacco leaf) squamous cell carcinoma is seen in palate due to repeated thermal injury.
- Malignant tumours may spread to periosteum, bone, maxilla, sinus, or nose.
- Salivary gland tumours are commonly malignant and are of adenoid cystic type. Other types also can occur. It presents as a single, solid, smooth swelling with ulcer over the summit.
- Squamous cell carcinoma is ulcerative with raised and everted edge.
- Upper deep cervical lymph nodes are involved in 25% of patients.
- Investigations are edge biopsy, FNAC lymph node and CT scan to see extensions.
- *Treatment* is wide excision with removal of the underlying palatal bone. Often partial or total maxillectomy (*Weber-Fergusson* incision) may be required. Myocutaneous flap with dental prosthesis is essential to reconstruct after surgery. Postoperative radiotherapy and neck block dissection are often required.

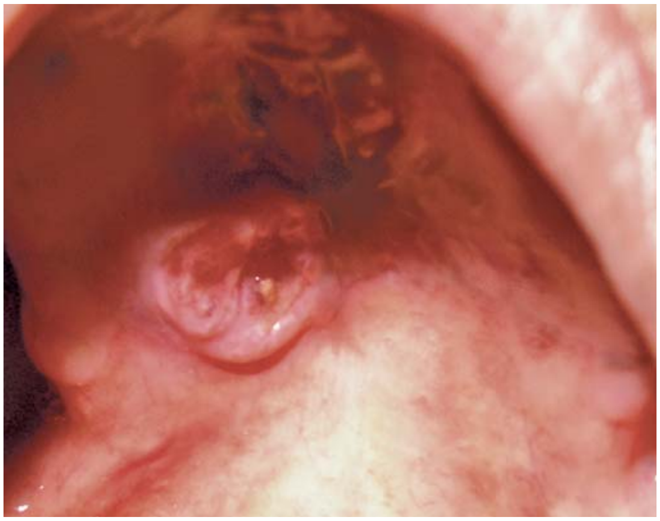

Fig. 3.91: Minor salivary gland tumour in posterior half of the palate. Note the nodular lesion with ulceration over the summit (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and Head of the Department and Dr Sampath, ENT Surgeon, KMC, Mangalore).

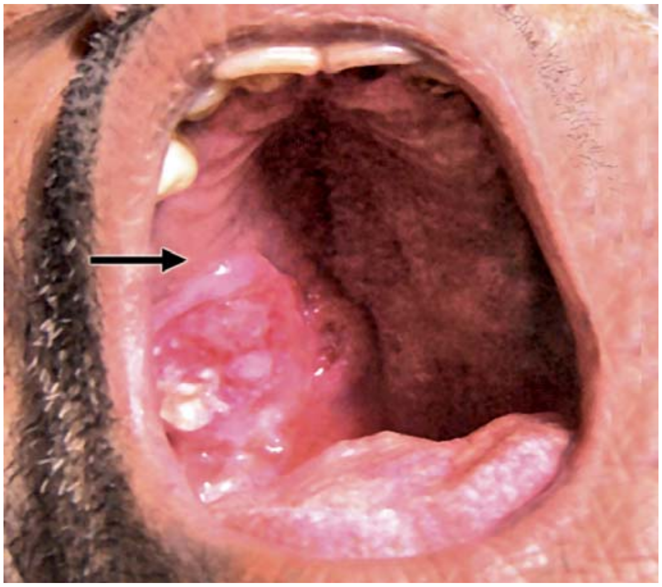

Fig. 3.92: Carcinoma hard palate extending laterally.

LARYNGEAL TUMOURS

Benign

- *Epithelial* can be papilloma, vocal nodule or vocal polyp. Papilloma is usually single in adult, multiple in children.
- *Connective tissue tumours* like fibroma, myxoma, angioma. Indirect laryngoscopy (ILS) or direct laryngoscopy confirms the diagnosis.

Childhood papilloma	***Adult papilloma***
Multiple, of viral origin (Papova)	Single, of neoplastic origin
Can occur in glottic, supra and infraglottic	Glottic only
Not premalignant	Premalignant
Recurrence common	Not common after complete excision
Excision is difficult	Easier removal
Causes more dyspnoea, *stridor*, cough May require tracheostomy	Hoarseness only

***Note*:**
Stridor can be inspiratory, expiratory or biphasic.

Treatment

- Endoscopic removal. Application of podophyllum
- Cryosurgery, Laser surgery.

MALIGNANT TUMOURS OF LARYNX

Aetiology

- Smoking, tobacco.
- Alcohol intake.
- Occupational/industrial exposure to chemicals like mustard gas, asbestos, benzopyrones, petroleum products.
- Previous radiation.
- Genetic: Russians develop familial laryngeal cancers.
- Papillomavirus, keratosis, malnutrition.

Incidence

- Squamous cell carcinoma is commonest (95%).
- Common in males (10 : 1).
- Common in 5th/6th decade.

He who attempts the absurd can achieve the impossible.

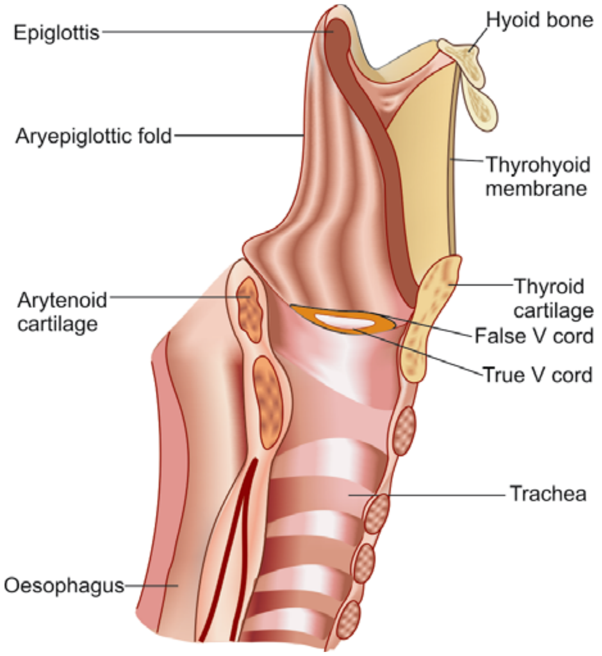

Fig. 3.93: Anatomy of larynx showing supraglottic, glottic and subglottic regions.

Types

- Ulcerative.
- Proliferative.

Anatomical Types

- **Supraglottic (25%):** It arises from infrahyoid part of epiglottis, ventricles, and arytenoids. It spreads to neck lymph nodes early (40%) due to rich lymphatics in this area. Throat pain, dysphagia, palpable neck nodes and referred pain are common features. Hoarseness of voice, loss of weight, respiratory obstruction, and halitosis are late features. Carcinoma in epiglottis causes bilateral nodal spread. Local spread occurs to vallecula, base of tongue and pyriform fossa.
- **Glottic (65%):** It is the commonest type. It begins from upper part or free edge of vocal cords (mid or anterior) often extending 10 mm below. Lymphatic spread is slow (only 4%) as this area has got least lymphatics. Opposite vocal cord can involve as *kiss cancer*. Vocal cord mobility is unaffected in early cases. Vocal cord fixation signifies spread to thyroarytenoid which is a poor prognostic sign. It presents very early due to hoarseness of voice. Eventual cord fixation causes stridor. Locally it spreads anteriorly to anterior commissure, posteriorly to vocal process and arytenoids, above to ventricle and false vocal cords, below to subglottis.
- **Subglottic (2%)** is less common involving under-surface of true vocal cords and subglottic space. It spreads to deep cervical and paratracheal nodes (20%). Upward spread is rather late and so hoarseness is not an early symptom in this type. It can spread through cricothyroid membrane or thyroid gland.

Note:

- In Indian subcontinent supraglottic tumours are more common than glottic. Glottic type is common in Western countries.
- Fixation of cords is due to involvement of thyroarytenoid muscle or cricoarytenoid joint.

Clinical features	***Investigations***
• Hoarseness of voice	• ILS (Indirect laryngoscopy
• Pain and discomfort	• Direct laryngoscopy and biopsy
• Cough. dyspnoea, stridor, dysphagia in late cases	• CT neck—very useful investigation
• Bloody sputum	
• Palpable neck nodes, which eventually get fixed	• Chest X-ray
	• FNAC of lymph node
• Absence of laryngeal crepitus	• Microlaryngoscopy in small lesions to identify and to have proper biopsy
• Common in males—10:1	• Toluidine blue staining to stain early superficial cancers which facilitate the accurate biopsy
	• Hopkin's endoscopy
	• Flexible, fibreoptic laryngoscopy

Staging

Tumour

T1 Tumour confined to one anatomical site in larynx with normal cord mobility.
T2 Tumour confined to one anatomical region within larynx.
T3 Tumour spreads beyond one anatomical region within larynx with cord fixation.
T4 Tumour spreads beyond larynx.

Nodal staging is similar to any other oral carcinomas.

Treatment

- *Supraglottic*: Stage I—curative radiotherapy is the choice. In stage II and III total laryngectomy and block dissection of neck nodes.
- *Glottic:* Radiotherapy is the choice as nodes are commonly not involved. Endoscopic laser surgery or open partial laryngectomy can be done.
- *Subglottic:* Total laryngectomy is the treatment of choice with nodal block when needed.
- *In advanced stage IV* carcinomas surgery and radiotherapy both are not possible. Here chemotherapy is given using cyclophosphamide, cis platin and methotrexate.
- *In advanced resectable tumours, induction chemotherapy for 3 cycles with methotrexate and cisplatin 100 m/sq metre BSA (at 0, 22. 43 days) and total laryngectomy is under trial. Concurrent chemotherapy and radiotherapy are also under trial in these cases.*

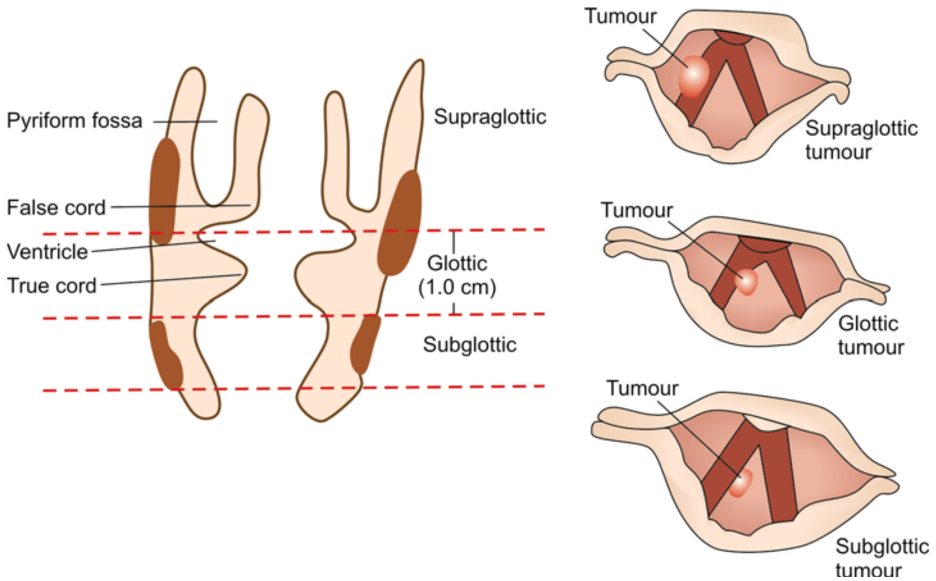
Pyriform fossa
False cord
Ventricle
True cord
Supraglottic
Glottic
(1.0 cm)
Subglottic
Tumour
Supraglottic tumour
Tumour
Glottic tumour
Tumour
Subglottic tumour

Fig. 3.94: Laryngeal carcinoma, types. Note the typical sites. Glottic is the commonest site. Next is supraglottic. Subglottic is rare.

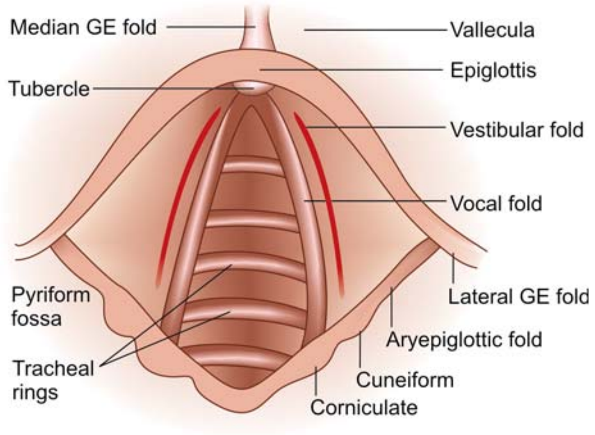
Median GE fold
Vallecula
Epiglottis
Tubercle
Vestibular fold
Vocal fold
Pyriform fossa
Lateral GE fold
Aryepiglottic fold
Tracheal rings
Cuneiform
Corniculate

Fig. 3.95: View of larynx as seen through a laryngoscope.

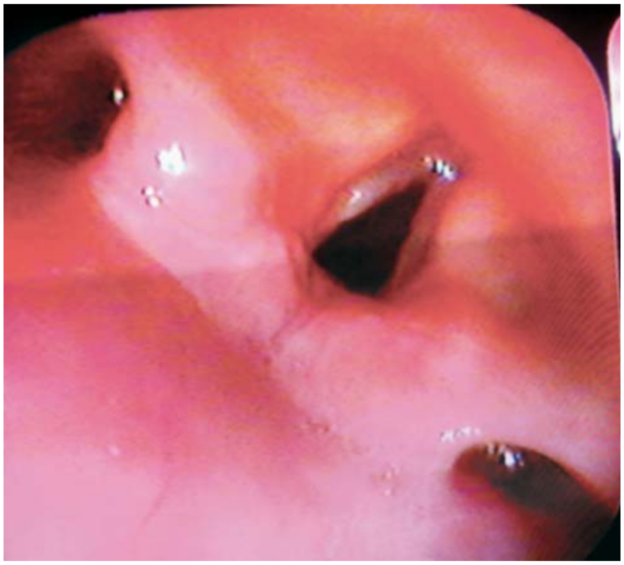

Fig. 3.96: Endoscopic view of vocal cords, pyriform fossa and aryepiglottic fold.

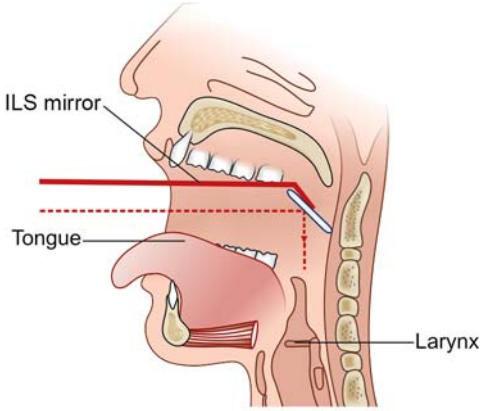
ILS mirror
Tongue
Larynx

Fig. 3.97: Indirect laryngoscopy to visualise larynx and its parts.

Role of Radiotherapy in Laryngeal Cancer

- In early growth with no impairment in motility curative RT is very useful with 90% cure rate with preservation of voice
- It is commonly used in superficial exophytic lesions, growth in tip of epiglottis and aryepiglottic folds.
- In subglottic extension, fixed growths, and in presence of nodes, radiotherapy is less effective.

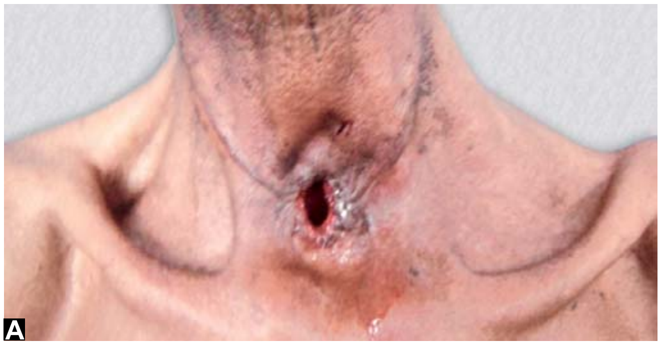

Fig. 3.98A

The greatest happiness of life is the conviction, that we are loved.

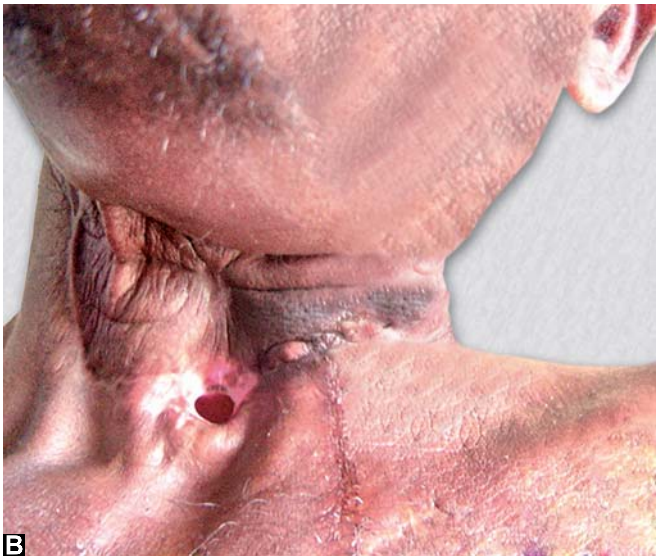

Fig. 3.98B

Figs 3.98A and B: Carcinoma larynx—patients underwent total laryngectomy with permanent tracheostomy.

Conservative laryngeal surgery

Indications	*Advantages*
• Early growth without fixation especially glottic type	• Permanent tracheostomy is avoided Voice is retained
Types	*Disadvantages*
• Cordectomy through laryngofissure–excision of vocal cord after splitting of larynx	• Inadequate clearance • Recurrence of the disease
• Partial frontolateral laryngectomy–excision of vocal cord and anterior commissure	
• Partial horizontal laryngectomy–excision of supraglottis (epiglottis, aryepiglottic folds, false vocal cords and ventricle)	

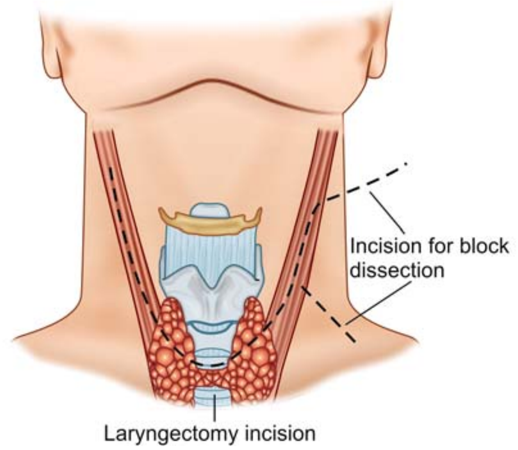

Fig. 3.99: *Gluck-Sorenson's* laryngectomy incision. Extension lines sideward can be used for adding radical neck dissection.

Total laryngectomy

Indications	*Technique*
• T_3 lesions with cord fixation • All T_4 lesions • Bilateral arytenoids spread • Thyroid/cricoid spread • Transglottic cancers involving ventricle with fixation of the cord • Posterior commissure disease • Failure of conservative surgery or RT *Problems* • Mortality of surgery • No speech • Having permanent tracheostomy and its problems	• Entire larynx with hyoid bone, pre-epiglottic space, strap muscles, one or more rings of trachea are removed • Pharyngeal wall is repaired • Lower tracheal stump is sutured to the skin as permanent tracheostomy • Often laryngo-oesophagectomy is done when pharyngeal spread is present. Gastric pull-up is done to maintain GI continuity • Technique may be combined with neck nodal dissection both sides

Care after total laryngectomy

- Speech therapy by pseudoglottis creation, battery operated artificial larynx or Singer-Blom prosthesis or Panje's prothesis. Tracheo-oesophageal prosthesis (TOP) is ideal
- Social and job rehabilitation
- Care of permanent tracheostomy by avoiding immersion in the water, care during bath, shower use and swimming. Shower covers are available for this purpose
- Often along with total laryngectomy, total thyroidectomy and removal of parathyroid glands are required. Patient then needs supplementation of thyroxine and calcium for life time

TRISMUS

- It is inability to open the mouth adequately.
- Causes are—submucosal fibrosis, carcinoma buccal and gingivobuccal complex, post-radiotherapy sequelae, infection like tetanus, parotitis, dental or peritonsillar abscess.
- In carcinoma, invasion of tumour to pterygoids, buccinator, masseter, temporalis causes trismus.
- *Grading of trismus:* Interincisor distance more than 3.5 cm is—normal; Grade I is between 3.0–3.5 cm; Grade II is between 2.0—3 cm; Grade III is less than 2 cm.
- It is often clinically assessed by placing fingers perpendicularly between two jaws at incisor level. More than 3 fingerbreadth is considered as normal.
- *Problems with trismus* are—inability to put fingers or spoon into mouth; difficulty in cleaning the mouth; infection; difficulty in assessing the tumour / pathology clinically. During surgery, intubation is difficult and so tracheostomy may be needed in these patients.
- *Management*—treating the cause; release of soft tissue in case of fibrosis; draining the abscess and antibiotics for infection.

Chapter 4 Salivary Glands

> *(Three weeks after arriving at the anatomy laboratory of Gerhard Blasius in Amsterdam) fortune so favoured me that in the first sheep's head, which I…was dissecting alone in my room, I found a duct which, so far as I knew, had been described by no one before.*
>
> **—Niels Stensen,** 1661

CHAPTER OUTLINE

- Anatomy
- Saliva
- Sialography
- Salivary Calculus and Sialadenitis
- Sialosis
- Sialectasis
- Recurrent Childhood Parotitis
- Parotid Abscess
- Parotid Fistula
- Sjögren's Syndrome
- Mikulicz Disease
- Salivary Neoplasms
- Pleomorphic Adenoma
- Adenolymphoma
- Oncocytoma
- Basal Cell Adenoma
- Mucoepidermoid Tumour
- Adenoid Cystic Carcinoma
- Acinic Cell Tumour
- Malignant Mixed Tumour
- Adenocarcinoma of Salivary Glands
- Squamous Cell Carcinoma of Salivary Glands
- Submandibular Salivary Gland Tumours
- Management of Malignant Salivary Tumours
- Minor Salivary Gland Tumours
- Parotid Lymphoma
- Parotidectomy
- Frey's Syndrome
- Facial Nerve Injury

ANATOMY

Parotid Gland (Para—around, otis—ear)

It is the largest of the salivary gland, situated below the acoustic meatus between the ramus of mandible and sternomastoid muscle. The deep cervical fascia splits to form a capsule (parotid capsule) to enclose the gland. The superficial layer is thickened and adherent to the gland.

It is deep to parotid fascia, superficial to masseter. So parotid swelling occupies *below, behind, in front of the ear lobule*, obliterating the normal hollow below the ear lobule. When patient opens the mouth, parotid fascia stretches and swelling may become less preeminent but this test is difficult to elicit. When patient clinches his teeth, masseter contracts and parotid becomes more prominent.

Parts of the Parotid Gland

- Superficial part (80%)—lies over the posterior part of the ramus of mandible.
- Deep part (20%)—lies behind the mandible and medial pterygoid muscle; in relation to mastoid and styloid process.

Accessory Parotid

It is prolongation of the gland above the parotid duct (*socia parotidis*).

Parotid Duct

Parotid (*Stensen's*) duct is 2-3 mm in diameter, 5 cm in length, emerges from anterior surface of the gland, runs over the surface of the masseter muscle, passes through the buccinator

Both over-salivation and under-salivation are not advised. Only optimum salivation is accepted and ideal.

muscle, and opens into the oral mucosa opposite to the crown of upper second molar tooth. It is palpated *bidigitally*, one fingerbreadth below the zygomatic bone with index finger inside and thumb outside, in front of masseter which is felt during clinching of teeth. Transverse facial artery is *above* the parotid duct. Duct is 1 cm below the zygomatic arch.

Facial Nerve

It emerges from the stylomastoid foramen lying between external auditory meatus and mastoid process. It passes around the neck of the condyle of mandible and becomes superficial, later dividing into *temporofacial and cervicofacial branches* which in turn divides into many branches. Some of these may be interconnected *as pes anserinus* (goose foot). *Branches are* – temporal (auricularis anterior and superior part of frontalis), zygomatic (frontalis and orbicularis oculi), upper buccal and lower buccal (buccinator, orbicularis oris, elevators of the lip) mandibular (lower lip muscles) and cervical (platysma).

While exiting the skull through stylomastoid foramen (it is accompanied by stylomastoid branch of posterior auricular artery which enters the same foramen), it gives posterior auricular nerve and motor nerves to posterior belly of digastric and stylohyoid. Trunk is initially 1cm from posteromedial surface (*extraglandular*) and *intraglandular* for 1 cm before giving divisions.

Gland is often called as *parotid sandwich* due to transverse of facial nerve.

Blood Supply

It is from external carotid artery; and venous drainage is by external jugular vein.

Nerve Supply

It is from autonomic nervous system, parasympathetic is secretomotor from auriculotemporal nerve, sympathetic is vasomotor from plexus around the external carotid artery. *Faciovenous plane of Patey* is of surgical importance. Facial nerve is superficial to this plane which contains retromandibular vein and posterior facial vein. External carotid artery dividing into superficial temporal artery and maxillary artery is deeper to venous plane.

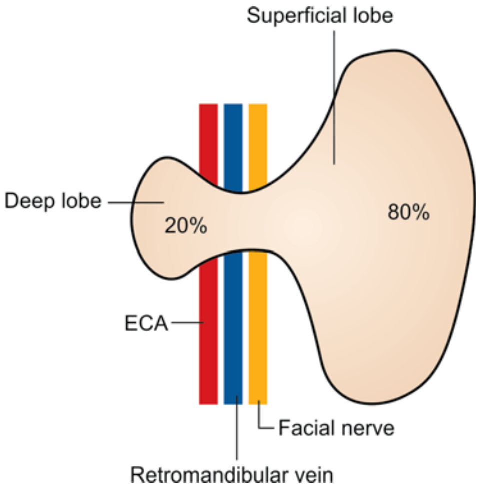

Fig. 4.1: Superficial lobe (80%) and deep lobe (20%) are separated by external carotid artery, retromandibular vein and facial nerve from deep to superficial.

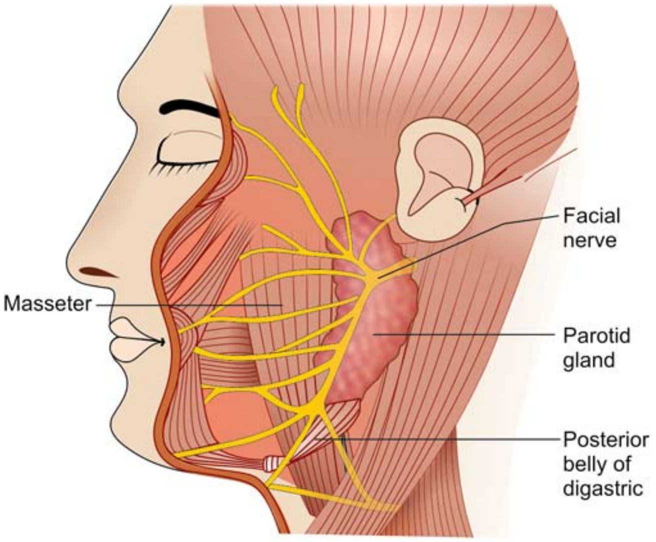

Fig. 4.2: Anatomical relations of the parotid gland.

Lymphatic Drainage of Parotid

It drains into parotid lymph glands which are partly intraglandular and partly extraglandular (preauricular and infraauricular). Mainly intraglandular nodes are involved which later drains into deep cervical lymph glands. Parotid lymph nodes also drain from temple, side of scalp, lateral part of auricle, external acoustic meatus, upper part of cheek, parts of eyelids and orbit.

Structures within the parotid gland from deep to superficial

- External carotid artery, maxillary artery, superficial temporal artery, posterior auricular artery
- Retromandibular vein (by maxillary and superficial temporal veins)
- Facial nerve with its branches

Great auricular nerve (cutaneous sensory around angle and lower part of the ear lobule) and auriculotemporal nerve which is from mandibular division of trigeminal nerve (secretomotor to parotid gland) are other nerves present in relation to parotid gland.

Submandibular Salivary Gland

It is a 'J' shaped salivary gland situated in the anterior part of the digastric triangle.

Parts

- *Superficial part:* Lies in submandibular triangle, superficial to mylohyoid and hyoglossus muscles, between the two bellies of digastric muscle.
- *Deep part* is in the floor of the mouth and deep to the mylohyoid.

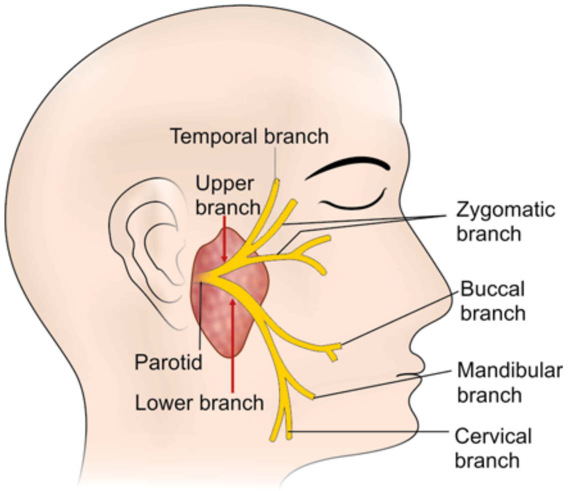
Temporal branch
Upper branch
Zygomatic branch
Buccal branch
Parotid
Mandibular branch
Lower branch
Cervical branch

Fig. 4.3: Pes anserinus.

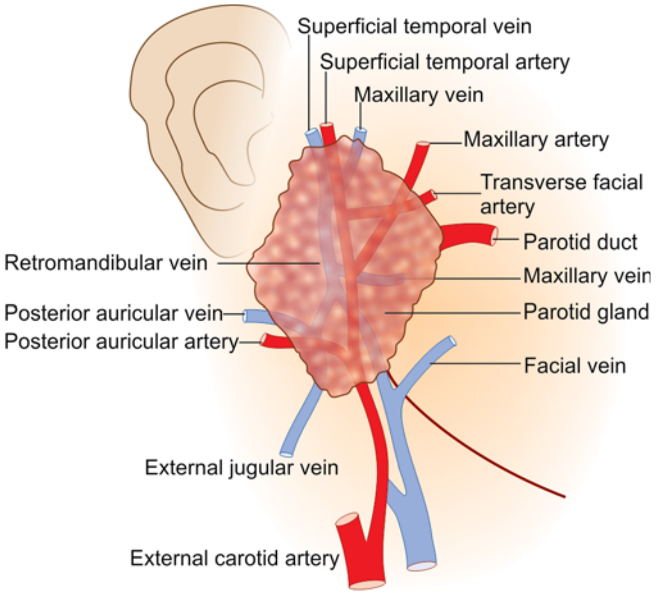
Superficial temporal vein
Superficial temporal artery
Maxillary vein
Maxillary artery
Transverse facial artery
Parotid duct
Retromandibular vein
Maxillary vein
Posterior auricular vein
Parotid gland
Posterior auricular artery
Facial vein
External jugular vein
External carotid artery

Fig. 4.4: Vascular plane of parotid.

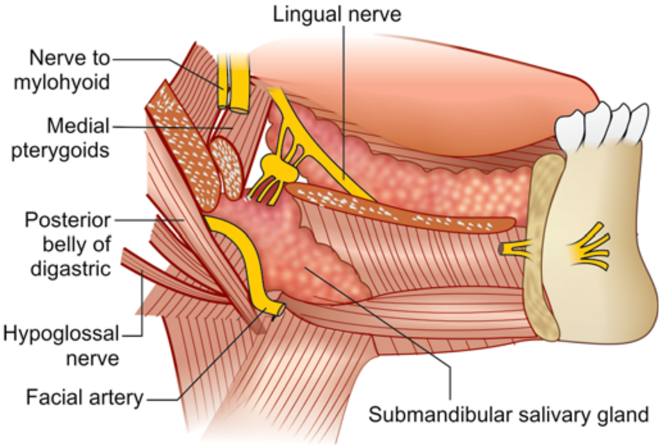
Lingual nerve
Nerve to mylohyoid
Medial pterygoids
Posterior belly of digastric
Hypoglossal nerve
Facial artery
Submandibular salivary gland

Fig. 4.5: Anatomical relations of the submandibular salivary gland.

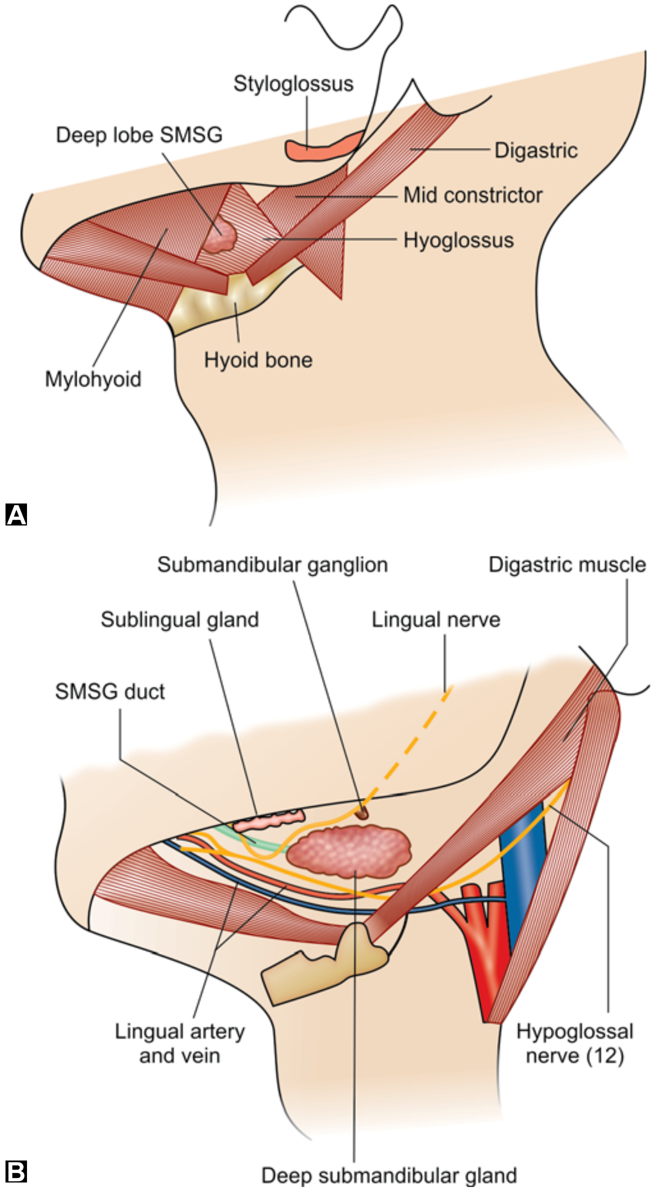
Styloglossus
Deep lobe SMSG
Digastric
Mid constrictor
Hyoglossus
Hyoid bone
Mylohyoid
A
Submandibular ganglion
Digastric muscle
Sublingual gland
Lingual nerve
SMSG duct
Lingual artery and vein
Hypoglossal nerve (12)
B
Deep submandibular gland

Figs 4.6A and B: Relations of the submandibular salivary gland and its deep lobe.

Submandibular (*Wharton's*) duct (5 cm), emerges from the anterior end of the deep part of the gland, enters the floor of the mouth, on the summit of papilla beside the frenulum of the tongue.

Lingual nerve and submandibular ganglion are attached to upper pole of the gland. Hypoglossal nerve is deep to the gland.

Facial artery emerges from under surface of the stylohyoid muscle, enters the gland from posterior and deep surface, reaching its lateral surface crossing the lower border of mandible to enter the face.

Venous drainage is to anterior facial vein.

Nerve supply: Branches from the submandibular ganglion.

Resting salivary flow usually arises from the submandibular salivary gland. *Sialorrhoea* is increased salivary flow often seen due to drugs, in cerebral palsy, physically handicapped person, children, and psychiatry patients. Intractable sialorrhoea can be

How a person masters his fate is more important than what his fate is.

corrected by different surgeries to submandibular salivary gland like duct repositioning to excision of the gland.

Xerostomia is decreased salivary flow. It is seen in postmenopausal women, depression, dehydration, use of antidepressant drugs; anticholinergic drugs, Sjogren's syndrome, radiotherapy to head and neck region.

Minor Salivary Glands

There are around 450 minor salivary glands which are distributed in lips, cheeks, palate and floor of the mouth. Glands also may be present in oropharynx, larynx, trachea and paranasal sinuses. They contribute to 10% of total salivary volume.

Sublingual glands are minor salivary glands one on each side; located in the anterior aspect of the floor of the mouth in relation to mucosa, mylohyoid muscle, body of the mandible near mental symphysis. Gland drains directly into mucosa or through a duct which drains into submandibular duct. This duct is called as *Bartholin duct*.

Minor salivary glands are not present in gingivae and anterior portion of the hard palate.

Ectopic Salivary Gland

Ectopic salivary gland also called as aberrant salivary gland is nothing but ectopic lobe of the juxtaposed salivary gland. It is commonly seen in relation to submandibular salivary gland.

Commonest ectopic salivary tissue is *Stafne bone cyst* (Edward C Stafne, Dental Surgeon, Mayo Clinic). It is invagination of the juxtaposed submandibular salivary gland into the mandible bone on its lingual aspect.

X-ray shows radiolucent area due to the cyst below the angle of the mandible, lower to inferior dental vessels and nerve.

SALIVA

- 1500 ml of saliva is secreted per day. pH of resting saliva is less, 7.0; active saliva is 8.0.
- Saliva contains lingual lipase secreted from tongue glands, α amylase from salivary glands.
- Saliva contains mucin, glycoproteins, immunoglobulin IgA, lysozyme, lactoferrin which binds iron, proline rich proteins that protect enamel and bind toxic tannins.
- Parotid saliva is 20% of total secretion of saliva per day and is serous and watery; submandibular is 70% and is mucous and moderately viscous; sublingual is 5% and is mucous and viscous. Minor salivary and other oral glands—5%.
- Saliva facilitates swallowing, keeps mouth moist, serves as solvent for taste buds, facilitates speech, keeps oral cavity rinsed and clean, antibacterial, and neutralizes gastric acid content in regurgitation to relieve *heartburn*.

SIALOGRAPHY

Indications

1. Salivary fistulas.
2. Sialectasis.
3. Congenital conditions.

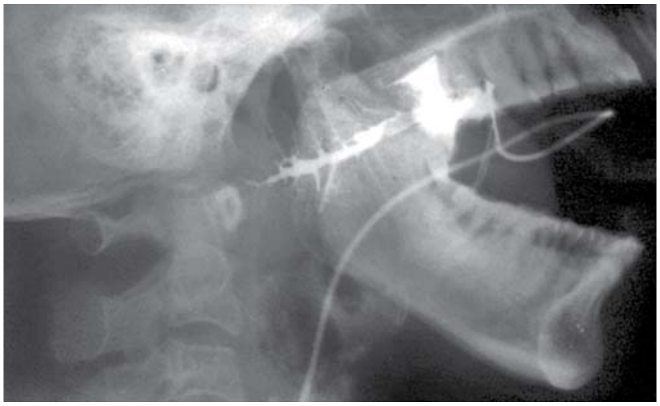

Fig. 4.7: Sialogram X-ray done to visualise the ductal pattern in parotid.

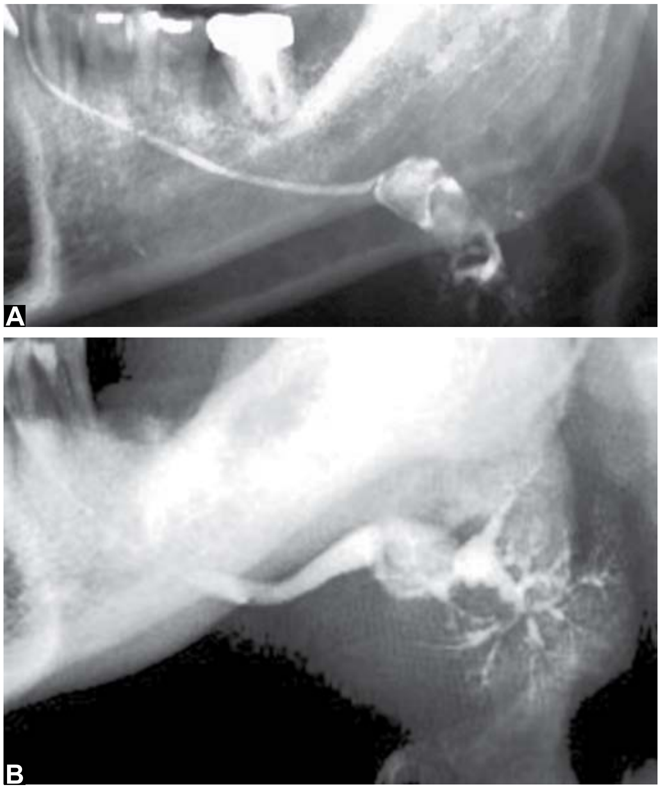

Figs 4.8A and B: Submandibular gland sialogram pictures (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and HOD, KMC, Mangalore; Dr Sampath, MS ENT, KMC, Mangalore).

4. Extraglandular masses.
5. Parotid duct stones.

Dye used is lipiodol or sodium diatrizoate (Hypaque).

24-gauge cannula is passed into either Stensen's duct or Wharton's duct and one ml of the dye is injected into the duct and X-ray is taken.

Findings

- Narrowing (Stricture)
- Grape-like cluster appearance (Sialectasis)
- Dilatations
- Communications (Fistulas)
- Mass lesions

Precautions

1. *Sialography should never be performed in acute inflammation.*
2. Only one ml of dye should be injected, if more dye is injected it causes *extravasation and chemical sialadenitis*.

SALIVARY CALCULUS AND SIALADENITIS

- 80% Submandibular.
- 80% Radio-opaque.
- It is commonly calcium phosphate and calcium carbonate stones.
- Calculi are more common in submandibular gland, because the gland secretion is viscous, contains more calcium and also, its drainage is nondependent, causing stasis.
- Secretion from parotid is serous, contains less calcium and so stones are not common.

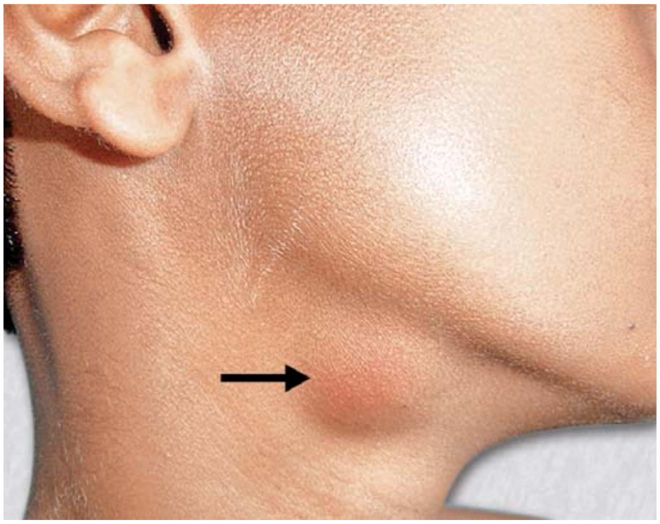

Fig. 4.9: Submandibular sialadenitis in a young boy who required excision of gland.

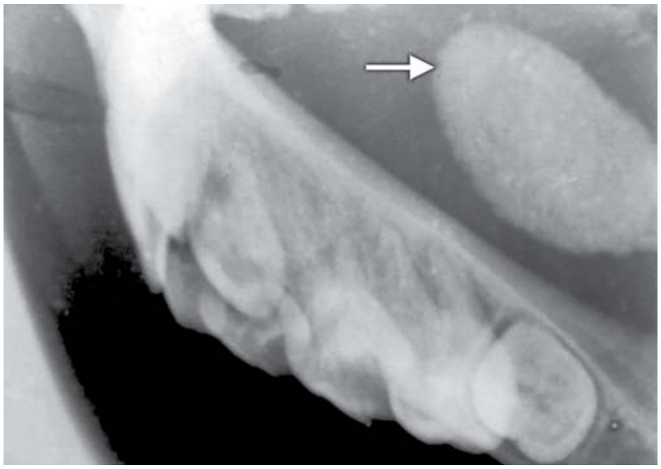

Fig. 4.10: Stone in the submandibular salivary gland as seen in X-ray (*Courtesy:* Dr Jagdish, Mangalore).

Salivary calculi

Submandibular gland	*Parotid gland*
80% common	20% incidence (*Rare*)
80% Radiopaque	Radiolucent
Seen in plain X-ray (intraoral)	Not seen in plain X-ray
Sialogram is not needed	Identified by sialogram

Calculi are common in submandibular salivary gland, because:

- Viscous nature and mucin content
- Calcium content
- Nondependent drainage
- Stasis
- Hooking of nerve by submandibular duct

Causes for submandibular sialadenitis	***Types of sialadenitis***
• Bacterial—more common. It is usually due to obstruction and stasis • Trauma over duct causing oedema/stricture and stasis • Viral—mumps—rare (Paramyxovirus)	• *Acute* – Bacterial—occurs following submandibular salivary ductal obstruction (Wharton's) or in parotid gland. In parotid, suppuration can occur leading into parotid abscess – Viral—common in parotid • *Chronic*—common after partial obstruction of submandibular gland duct or due to stones in submandibular gland or hilum proximal to the level of crossing of the lingual nerve over the duct

Presentation

Acute Sialadenitis—Features

- Pain, swelling, tenderness is seen in submandibular region and floor of the mouth.
- Dysphagia, trismus, fever.
- Double chin appearance due to spreading of oedema downwards.
- Duct is inflamed and swollen.

Chronic Sialadenitis—Features

- Pain is more during mastication due to stimulation (*Salivary colic* which can be induced by meals, lemon juice, etc). Salivary colic is pain induced by obstruction to the outflow of saliva may be ductal stone. During salivation size of the swelling will decrease again 2 hours after meal/stimulation.
- Salivary secretion is more during mastication causing increase in gland size.

Diagnosis means finding the cause of the disorder, not just giving its name.—**Sydney Walker**

- Firm/rubbery tender swelling is palpable bidigitally.
- When stone is in the duct, it is palpable in the floor of the mouth as a tender swelling with features of inflammation in the duct. Pus exudes through the duct orifice. (Irritation of the lingual nerve, which is in very close proximity to submandibular salivary duct, causes referred pain to tongue—*lingual colic*).
- In submandibular salivary gland, often the stones are multiple, with chronic inflammation of gland (*sialadenitis*).
- Often *acute on chronic sialadenitis* can occur.
- *Kuttner tumour* is chronic sclerosing sialadenitis of submandibular salivary gland.

Differential Diagnosis

- Submandibular lymphadenitis.
- Salivary neoplasm.

Investigations

- Intraoral X-ray (dental occlusion films) to see radiopaque stones (80%).
- FNAC of the gland to rule out other pathology.
- Total count and ESR in acute phase.
- USG will demonstrate stone *with posterior acoustic shadow*.

Note:

Radiological demonstration of stone/stones is called as *sialolithiasis*.

Treatment

- If the stone is in the duct, removal of the stone is done intraorally, by making an incision in the duct. Incised duct is not sutured as it may result in stricture. Laying open allows free drainage of saliva. Procedure is usually done under *local anaesthesia*.

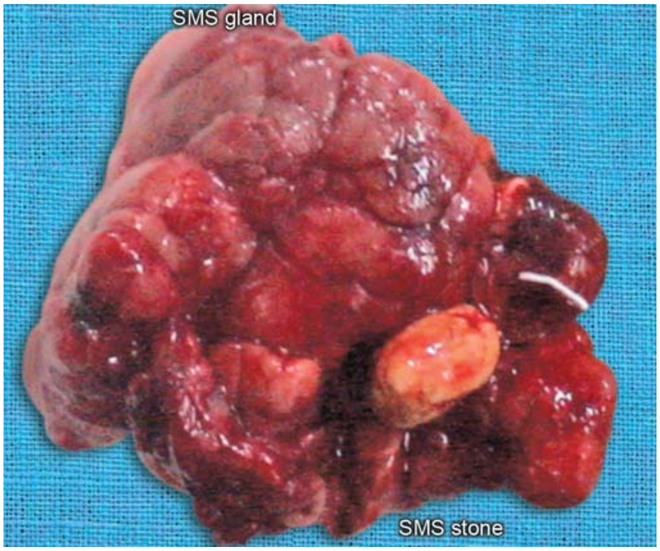

Fig. 4.12: Excised specimen of submandibular salivary gland with stone in the gland.

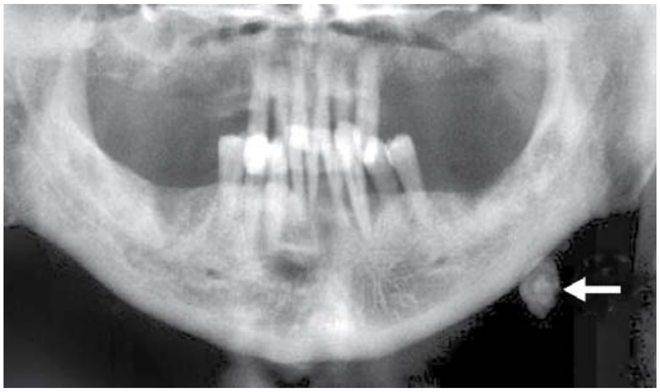

Fig. 4.13: X-ray (OPG) showing left sided submandibular salivary stone.

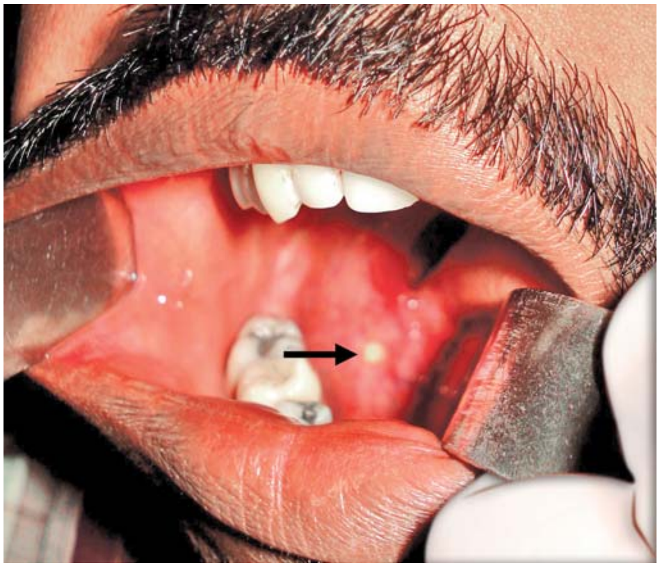

Fig. 4.11: Stone in the duct of submandibular salivary gland (Wharton's duct).

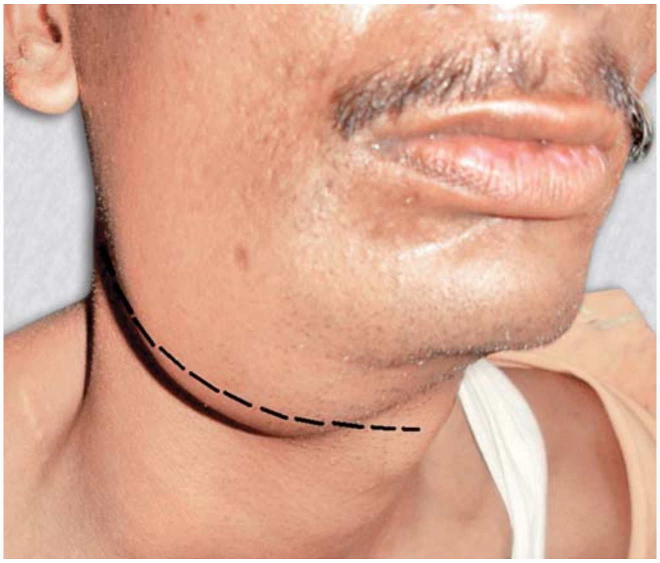

Fig. 4.14: Incision for removal of the submandibular salivary gland.

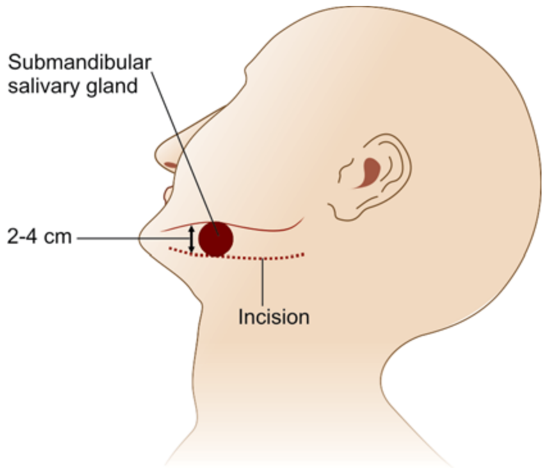

Fig. 4.15: Incision for excision of submandibular salivary gland. It should be 2-4 cm below the margin of the mandible to avoid injury to marginal mandibular nerve.

- If stone is in the gland, *excision of submandibular* gland is done—*sialadenectomy*. It is always done under *general anaesthesia*.

Approach is from submandibular region (outside). An incision is made on the skin in submandibular region, about 5-8 cm length, parallel to and 2-4 cm below the mandible. Incision is deepened through the deep fascia until the gland is visualised without raising the flaps (so as to avoid injury to marginal mandibular nerve, branch of facial nerve). Facial artery is ligated twice. Lingual nerve and hypoglossal nerves are taken care of. Mylohyoid is retracted so as to remove the deep portion of the gland. Drain is placed after removal of the gland.

Antibiotic coverage is a must.

Facial artery lies in the groove on the deeper aspect of the gland; often embeds in the gland or runs around the gland with a variable course and so artery has to be ligated twice above anteriorly and below posteriorly.

Marginal mandibular nerve is in subplatysmal plane in neck, so incision should be deepened across the deep fascia without raising subplatysmal plane to avoid injury to this nerve.

Complications of Surgery

1. Haemorrhage.
2. Infection.
3. Injury to marginal mandibular nerve, lingual nerve, hypoglossal nerve.
4. Injury to nerve to mylohyoid causing anaesthesia over submental skin.

Note:

Stone in parotid duct can be removed by opening the duct longitudinally.

Stone in the gland or in collecting duct is treated by parotidectomy.

Indications for submandibular salivary gland excision	***Steps in submandibular salivary gland excision***
• Chronic sialadenitis • Submandibular salivary tumours *Pathway of secretomotor fibres of the submandibular and sublingual salivary glands* Superior salivary nucleus ↓ Preganglionic fibres ↓ Facial nerve, Chorda tympani nerve ↓ Lingual nerve ↓ Submandibular ganglion—post-ganglionic fibres ↓ Submandibular and sublingual salivary glands	• Anaesthesia—general
	• Position—neck extension with chin to opposite side
	• Incision—2-4 cm (3 cm) below and parallel to the margin of the mandible 5-8 cm in length (6 cm)
	• Mobilisation of the gland–intracapsular in sialadenitis; extracapsular in tumours with ligation of anterior facial vein
	• Facial artery ligation proximally and distally as artery is in the gland or in the groove posteriorly
	• Dissection of deeper lobe from mylohyoid muscle
	• Identification of lingual and hypoglossal nerves
	• Duct identification and ligation
	• Wound closure with a suction drain

Rule of 2 in submandibular salivary gland

- 2 parts divided by mylohyoid muscle
- 2 conditions affect it—tumour and stone
- 2 superficial nerves—cervical and mandibular branch of facial nerves
- 2 deep nerves—lingual and hypoglossal nerves
- Incision—2-4 cm below the mandible
- Ligate facial artery at 2 places

SIALOSIS

It is enlargement of the salivary gland due to fatty infiltration, as a result of various metabolic causes like diabetes, acromegaly, obesity, liver disease, alcoholism, bulimia, idiopathic, drug induced (sympathomimetics, carbimazole, thiouracil).

Clinical features: Bilateral diffuse enlargement of parotids, which is smooth, firm, nontender.

Treatment: The cause is treated.

SIALECTASIS

- It is an aseptic dilatation of salivary ductules causing *grape-like (cluster-like)* dilatations.

By perseverance the snail reached the ark.—**Charles Haddon Spurgeon**

- It is a disease of unknown aetiology with destruction of parenchyma of gland accompanied by stenosis and cyst formation in the ductules.
- It is common in parotids; often bilateral.
- Presents as a smooth, soft, fluctuant, nontransilluminating swelling which increases in size during mastication. It is tender initially. It lasts for many days with a long symptom-free period of the disease.
- Sialogram is diagnostic (grape or cluster-like dilatations).
- Treatment is conservative (nonsurgical).

RECURRENT CHILDHOOD PAROTITIS

It is a recurrent, rapid enlargement of one or both parotids with fever and malaise in children of age group between 3-6 years without any known aetiology. Recurrent episodes with a quiescent period in between are typical. Sialogram shows snowstorm punctate sialectasis. Low dose antibiotics for long period may be required. Occasionally patient may need total conservative parotidectomy especially if it occurs late in adolescent period.

PAROTID ABSCESS (SUPPURATIVE PAROTITIS)

- It is a result of an *acute bacterial sialadenitis* of the parotid gland.
- It is an ascending bacterial parotitis, due to reduced salivary flow, dehydration, starvation, sepsis, after major surgery, radiotherapy for oral malignancies and poor oral hygiene.
- Parotid fascia is densely thick and tough and so parotid abscess does not show any fluctuation until very late stage.
- Causative organism are *Staphylococcus aureus* (commonest), *Streptococcus viridans* and often others like gram-negative and anaerobic organisms.

Causes of acute parotitis (Differential diagnosis of suppurative parotitis)

- Viral—Mumps, Coxsackie virus A and B, parainfluenza 1 and 3, Echo and lymphocytic choriomeningitis
- Bacterial—*Staphylococcus aureus*
- Allergic
- HIV infection
- Radiotherapy, postoperative period
- Specific infections like syphilis
- Sjogren's syndrome often causes bilateral parotitis

Clinical Features

- Pyrexia, malaise, pain and trismus.
- Red, tender, warm, well-localised, firm swelling is seen in the parotid region (brawny induration).
- Tender lymph nodes are palpable in the neck.
- Features of bacteraemia are present in severe cases.
- Pus or cloudy turbid saliva may be expressed from the parotid duct opening.

Investigations

- U/S of the parotid region.
- Pus collected from duct orifice is sent for culture and sensitivity.
- Needle aspiration from the abscess is done to confirm the formation of pus.

Sialogram is contraindicated in acute phase, as it will cause retrograde infection leading into bacteraemia.

Note:
In suppurative parotitis patient may develop severe laryngeal or pharyngeal oedema and may require steroids, tracheostomy and critical care.

Complications of suppurative parotitis and parotid abscess

- Septicaemia
- Severe trismus
- Rupture into the external auditory canal

Treatment

1. Antibiotics are started depending on culture report.
2. When it is severely tender and localised, incision and drainage is done under G/A.

Areas where one should not wait for the abscess to form

- Parotid
- Breast
- Ischiorectal fossa
- Thigh

Skin is incised in front of the tragus vertically and then parotid sheath (pyogenic membrane) is opened horizontally. Pus is drained using sinus forceps and sent for C/S. Antibiotics are continued (***Blair's incision***).

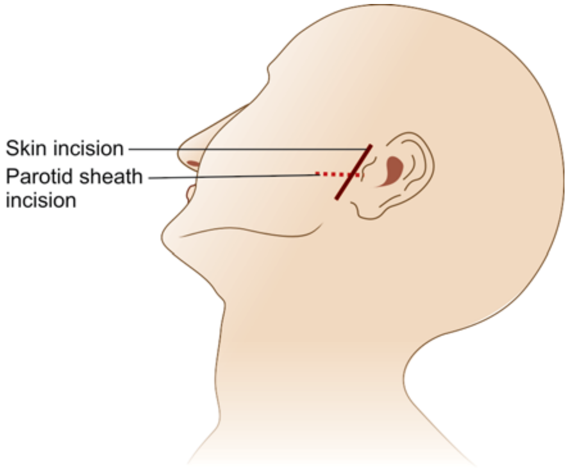

Fig. 4.16: *Blair's incision* used to drain the parotid abscess.

3. Proper hydration, mouth wash using povidone iodine/potassium permanganate solutions, nutrition. Often patient with parotid infection needs admission and treatment.

PAROTID FISTULA

Parotid fistula may arise from *parotid gland* or *duct* or *ductules*. It may open inside the mouth as internal fistula; or open

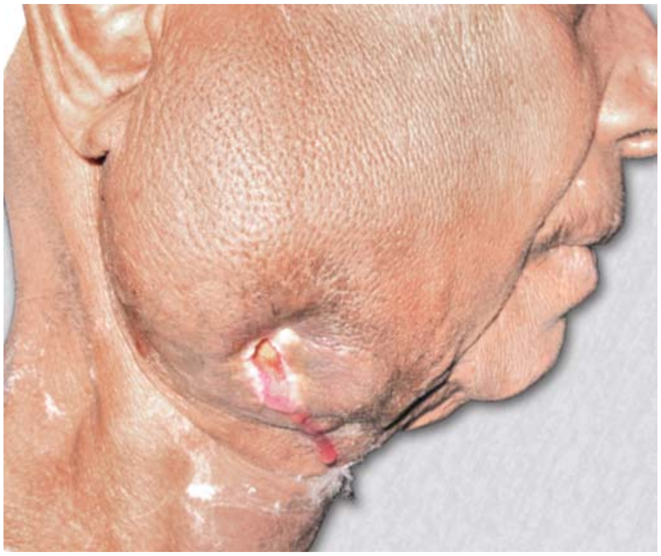

Fig. 4.17: Parotid fistula with discharging saliva and pus.

outside onto the skin as external fistula. Fistula from the duct has profuse discharge. Fistula from the gland often shows only minimal discharge.

Incidence of salivary fistula is 0.2-3%.

Types

1. *Duct fistula* forms after superficial parotidectomy. It is profuse and often persisting. So duct should be ligated using nonabsorbable suture as far as possible, anteriorly to allow normal saliva drainage from deep lobe. If common duct is ligated deep lobe atrophies without causing any fistula.
2. *Gland fistula* occurs from the raw surface after superficial parotidectomy. It is mild and symptom subsides in a month with anticholinergic drugs. Jacobsom tympanic neurectomy completely stops the secretion from the fistula in this type.

Causes

- After superficial parotidectomy.
- After drainage of parotid abscess, ruptured abscess.
- After biopsy.
- Trauma.
- Recurrence of malignant tumour.

Clinical Features

- Discharging fistula in the parotid region of the face, and discharge is more during eating.
- Tenderness and induration.
- Trismus.

Diagnosis

- Sialography to find out the origin of the fistula whether from parotid gland or duct or ductules.
- Fistulogram or CT fistulogram.
- Discharge study.
- MRI.

Treatment

- Anticholinergics—hyoscine bromide (probanthine).
- Radiotherapy.
- Often exploration of fistula is required.
- Repair or reinsertion of the duct into the mucosa.
- *Newman Seabrock's operation*—a probe is passed into the parotid duct through the opening in the mouth. Another probe is passed through the fistula. Duct and fistula are dissected over the probe. After removal of the fistula track severed duct ends are identified; and ends are trimmed. Probes are removed. A tantalum wire is passed into the duct across the severed ends and duct is sutured over it using 4 zero vicryl. Tantalum stent is removed after 3 weeks.
- If still persists, auriculotemporal nerve which supplies secretomotor component of parotid is cut.
- If there is stenosis at the orifice of the Stenson's duct, papillotomy at the orifice may help.
- Total conservative parotidectomy is done in failed cases.

Remember about salivary fistula

- Commonly from the parotid
- It can be *internal* draining into the mouth or can be *external* draining outside
- It is *acquired commonly* but rarely can be congenital
- It can be due to surgery, trauma or due to sepsis
- Fistula arising from the *gland parenchyma* drains through suture line but usually closes spontaneously. Leakage will be more during meals. Saliva is confirmed by its high amylase content compared to seroma/serous fluid
- Fistula due to *ductal disruption* leaks profusely and invariably needs surgery to close it
- Submandibular gland fistula commonly closes spontaneously, rarely if not requires complete removal of the gland
- Anticholinergics, irradiation, denervation of the gland, duct ligation are done to reduce saliva production
- Excision of fistula, repair of the duct, diversion of the duct into the mouth are other options
- In severe intractable cases, removal of the gland/total conservative parotidectomy is needed

SJÖGREN'S SYNDROME

It is an autoimmune disease causing progressive destruction of salivary and lacrimal glands, leading to *keratoconjunctivitis sicca* (dry eyes), and *xerostomia* (dry mouth).

Types

1. Primary.
2. Secondary.

> *Tumours of ectopic salivary glands are of low grade malignancy (cylindroma) may ultimately metastasize to regional lymph nodes, the viscera and the skeleton.*
> **—Kenneth Harrison**

Primary Sjögren's Syndrome (Primary Glandular Sicca Syndrome)

- Severe dry mouth.
- Severe dry eyes.
- Widespread dysfunction of exocrine glands.
- Incidence of developing lymphomas is high.
- There is no association of connective tissue disorders.

Primary extra glandular sicca syndrome

- Dry mouth, dry eyes.
- Hyperglobulinaemic purpura, vasculitis.
- Raynaud's phenomenon or B cell lymphoma.

Secondary Sjögren's Syndrome

- Dry mouth.
- Dry eyes.
- With association of connective tissue disorders like
  - Primary biliary cirrhosis (near 100%).
  - SLE (30%).
  - Rheumatoid arthritis (RA) (15%).
- Female to male ratio is 10 : 1.

Clinical Features

- It is common in middle aged females who present with dry eyes, dry mouth, enlarged parotids and enlarged lacrimal glands.
- Often they are tender.
- Superadded infections of the mouth, *Candida albicans* is common.

Investigations

1. Autoantibody estimation—rheumatoid factor, antinuclear factor, salivary duct antibody.
2. Sialography.
3. Estimation of salivary flow.
4. Slit lamp test of eyes.
5. Schirmer test – to detect lack of lacrimal secretion.
6. FNAC of parotids and lacrimal glands.
7. ^{99}Technetium pertechnetate scan for gland function.

Treatment: *It is conservative.*

1. Artificial tears.
2. Artificial saliva.
3. Frequent drinking of water.
4. Treat the cause.

MIKULICZ DISEASE

- It is a clinical variant of Sjögren's syndrome
- It is an autoimmune disorder of salivary and lacrimal glands, resulting in infiltration of the glands with round cells.
- It may be due to or associated with sarcoidosis, Sjögren's syndrome, leukaemia, lymphoma.

Triad

- Symmetrical and progressive enlargement of all salivary glands
- Narrowing of palpebral fissures due to enlargement of the lacrimal glands
- Parchment-like dryness of the mouth

SALIVARY NEOPLASMS

Aetiology

- *Genetic*—loss of alleles of chromosomes in 12q, 8q, 17q. Eskimos are more prone for salivary neoplasm.
- *Infective*—mumps, Epstein-Barr virus, chronic sialadenitis may be the cause; but not proved emphatically. Recurrent inflammation can cause duct dysplasia and carcinoma.
- *Radiation*—it is more common in survivors of atomic bomb explosion; mucoepidermoid carcinoma is more in these patients.
- *Smoking*—adenolymphoma of Warthin's shows 40% risk in smokers.
- *Sex*—benign tumours and many malignancies are common in females; Warthin's and some malignancies are common in males.
- *Environment and diet*—Arctic-Eskimos show dietary deficiency of vitamin A and develop salivary tumour. Industrial agents like nickel, cadmium, hair dyes, silica, preservatives may increase the risk of salivary tumours.

Classification (WHO)

a. ***Epithelial*** (90%):
   1. *Adenomas*
      - Pleomorphic adenoma.
      - Monomorphic adenomas.
        - Adenolymphoma (Warthin's tumour).

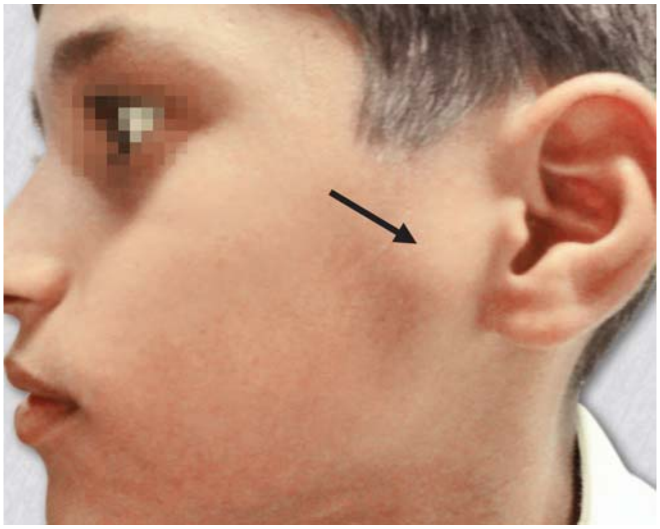

Fig. 4.18: Partoid enlargement in a boy of 11-year-old.

  - Oncocytoma (oxyphil adenoma) seen in elderly; seen in parotid gland.
  - Basal cell adenoma—it is a rare benign tumour.

2. *Carcinomas*
   - Mucoepidermoid carcinoma—most common malignancy.
   - Acinic cell carcinoma—1%.
   - Adenoid cystic carcinoma—very aggressive—10%; common minor salivary glands.
   - Adenocarcinoma.
   - Squamous cell carcinoma—2%.
   - Carcinoma in ex pleomorphic adenoma.
   - Undifferentiated carcinoma.

Note:

Carcinoma can be:

Low grade (acinic cell; adenoid cystic; low grade mucoepidermoid) or *High grade* (adenocarcinoma; squamous cell carcinoma; high grade mucoepidermoid) mesenchymal.

b. ***Nonepithelial:***

- Haemangioma—*commonly* seen in infants, usually in parotids. Spontaneous regression is common. Most common benign salivary gland tumour in paediatric age group.
- Lymphangioma.
- Neurofibromas and neurilemmomas.

c. ***Malignant lymphomas—NHL type:***

- Common in parotids.
- Common with HIV, Sjögren's syndrome (44 times more chances than normal people).

d. Secondary tumours from head and neck region; bronchus and skin.

e. Lymphoepithelial tumours:
   - Benign—it is 5% of all benign salivary tumours (*Godwin's tumour*). It is common in females; can be bilateral. Benign lymphoepithelial lesion (BLEL) is of unknown etiology characterized by replacement of salivary parenchyma with lymphoid tissue. It may be diffuse/focal/capsulated/unencapsulated. It may be associated with Mikulicz's disease or Sjögren's syndrome.
   - Malignant—it is rare tumour occurs in parotid and submandibular glands (ESKIMOMA).

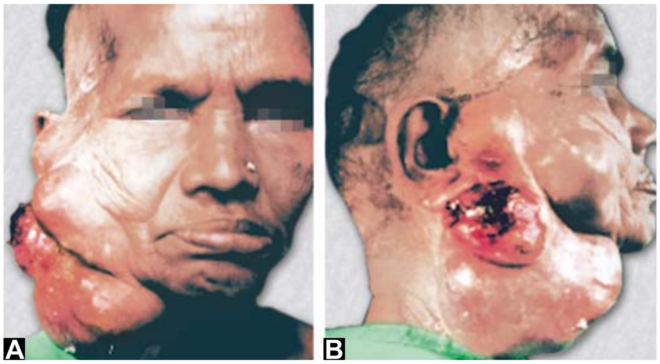

Figs 4.19A and B: Neoplasm of parotid gland.

Incidence

- Eighty per cent salivary neoplasms are in the parotids of which 80% are benign; 80% of these are pleomorphic adenomas.
- Fifteen per cent of salivary tumours are in the submandibular salivary gland, of which 50% are benign. 95% of these are pleomorphic adenomas.
- Ten percent of salivary neoplasms are in the minor salivary glands—palate, lips, cheeks, and sublingual glands. Of these only 10% are benign.

Note:

- Parotid tumours are common but only 20% are malignant.
- Submandibular tumours are uncommon but 50% of them are malignant.
- Minor salivary gland tumours (other than sublingual glands) are rare and 90% of them are malignant.
- Sublingual salivary tumours are very rare but almost all sublingual salivary tumours are malignant.
- Incidence of malignancy in salivary glands is inversely related to size of the gland; in parotid it is 15%; in submandibular it is 50%; in sublingual it is 85%.

PLEOMORPHIC ADENOMA (Mixed Salivary Tumour)

- Commonest of the salivary gland tumour in adult.
- It is 80% common.
- More common in parotids (80%). 10% in submandibular salivary gland; 0.5% in sublingual salivary gland.
- It is mesenchymal, myoepithelial and duct reserve cell origin.

Grossly it contains cartilages, cystic spaces, solid tissues.

Microscopically it is biphasic in nature with epithelial and stromal components. Benign tumours will usually not show necrosis.

Histologically it shows:

- Epithelial cells
- Myoepithelial cells
- Mucoid material with myxomatous changes
- Cartilages/pseudocartilages

- Even though it is capsulated, tumour may come out as *pseudopods* and may extend beyond the main limit of the tumour tissue.
- When disease occurs in parotid, commonly it involves superficial lobe or superficial and deep lobe together.
- But sometimes only deep lobe is involved and then it presents as swelling in the lateral wall of the pharynx, soft palate and posterior pillar of the fauces. There may not be any visible swelling in the preauricular region. It is called as '*dumbbell tumour*'. This tumour is in relation to styloid process, mandible, stylohyoid, styloglossus, stylopharyngeus muscles.

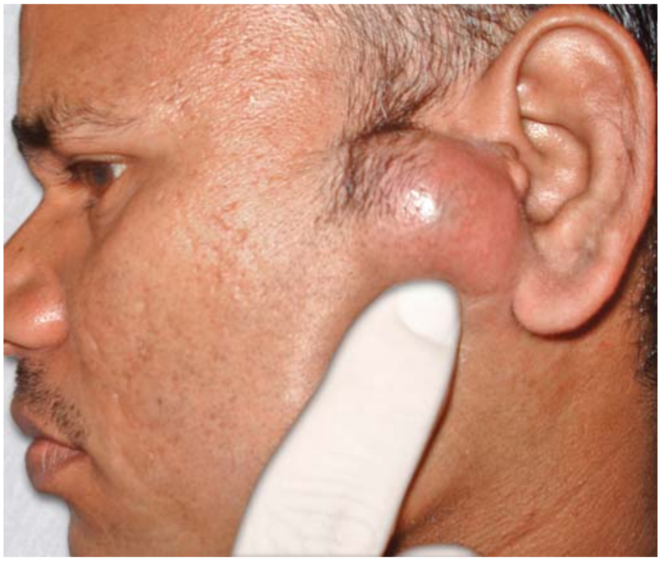

Fig. 4.20: Pleomorphic adenoma showing *curtain sign*.

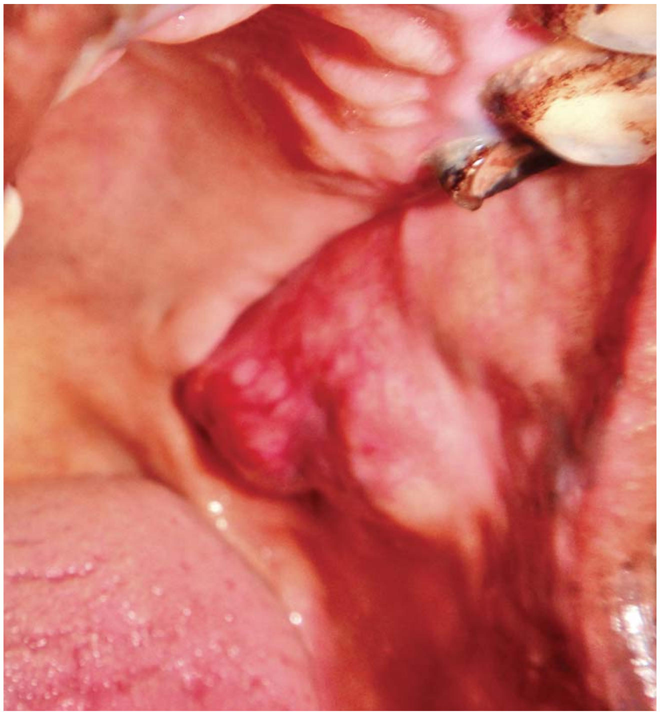

Fig. 4.21: Deep lobe tumour of parotid is examined per orally.

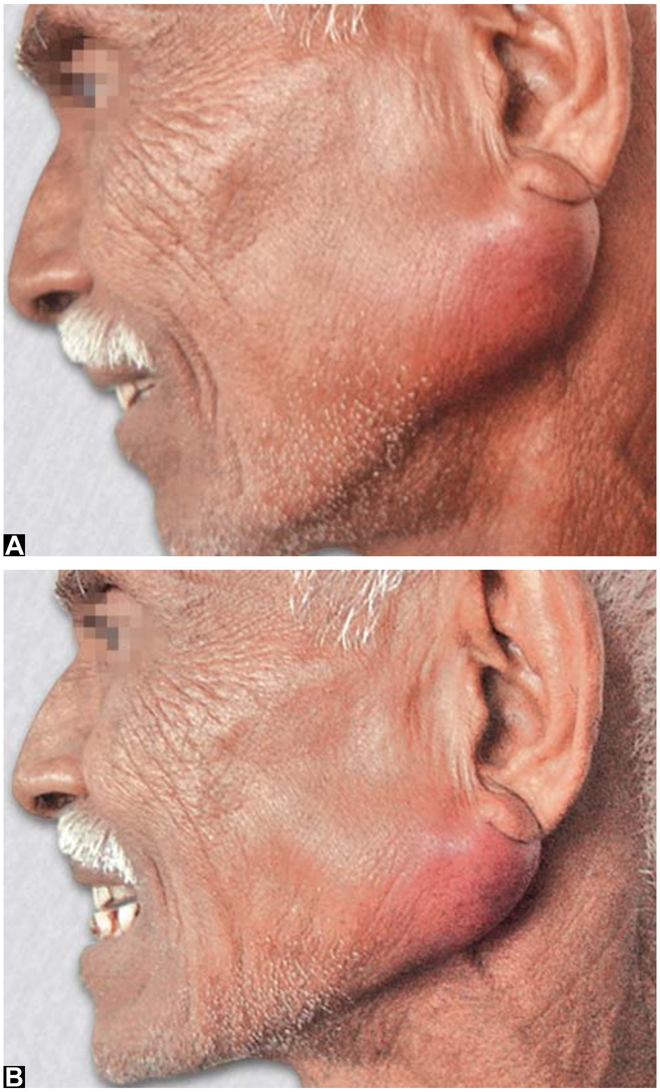

Figs 4.22A and B: Typical parotid swelling with ear lobe raised. Facial nerve should be tested by clenching the teeth.

Clinical Features

Clinical features of parotid tumour

- Swelling, pain, ulceration, dysphagia (if deep lobe is involved)
- Raised ear lobule
- Cannot be moved above the zygomatic bone—curtain sign
- Deviation of uvula and pharyngeal wall towards midline in case of deep lobe tumour
- Facial nerve, masseter, skin, lymph node and bone involvement eventually occurs in case of malignancy

- 80% common.
- Common in females (3:1).
- Occurs in any age group. But common in 4th and 5th decade
- Usually unilateral.
- Present as a single painless, smooth, firm lobulated, mobile swelling in front of the parotid with *positive curtain sign* (As the deep fascia is attached above to the zygomatic bone, it acts as a curtain, not allowing the parotid swelling to move above that level. Any swelling superficial to the deep fascia will move above the zygomatic bone).
- Obliteration of retromandibular groove is common.
- The ear lobule is lifted.
- When deep lobe is involved, swelling is commonly located in the lateral wall of pharynx, posterior pillar and over the soft palate—10%. Deep lobe tumour passes through *Patey's stylomandibular tunnel* pushing tonsils, pharynx, soft palate often without any visible swelling or only small swelling when only deep lobe tumour is present; it also presents as dysphagia. Bidigital palpation of parotid is significant in such occasion with one finger inside mouth.
- Facial nerve is not involved.

Long-standing pleomorphic adenoma may turn into carcinoma (carcinoma in ex pleomorphic adenoma). Its features are:

- Recent increase in size
- Pain and nodularity
- Involvement of skin, ulceration
- Involvement of masseter
- Involvement of facial nerve—lower facial nerve palsy—(Difficulty in closing eyelid, difficulty in blowing and clenching teeth)
- Involvement of neck lymph node
- Restriction of jaw movements

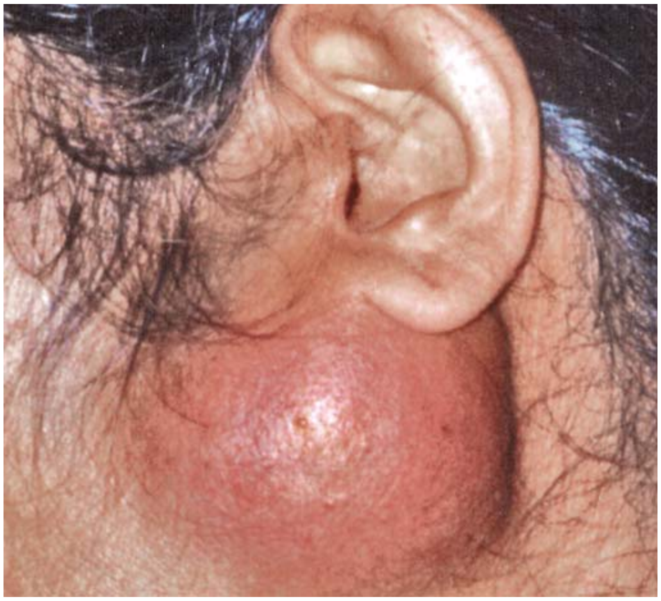

Fig. 4.23: Typical parotid tumour—a pleomorphic adenoma.

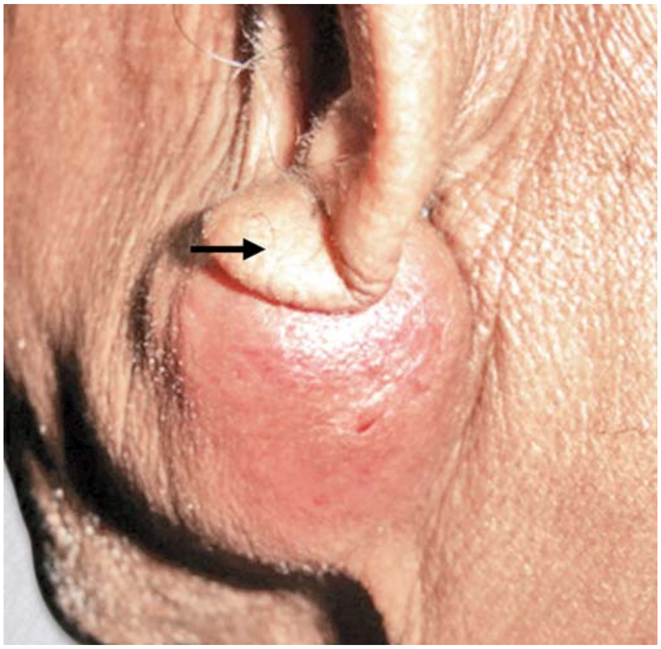

Fig. 4.24: Parotid tumour showing typical raise in earlobe.

Pain in salivary tumours

- Benign tumours are usually painless
- Sudden onset of pain denotes malignant transformation
- Pain is dull boring at primary site or referred to ear through auriculotemporal nerve
- *Pain is due to:*
  - Capsular distension by tumour
  - Obstruction to free flow of saliva
  - Nerve infiltration
  - Inflammation like in Warthin's
  - Tumour necrosis

Complications

- Recurrence—5-50%.
- Malignancy.
  - 3-5% in early tumours.
  - 10% in long duration (15 or more years) tumours.

Investigations

- FNAC is very important and diagnostic.
- CT scan to know the status of deep lobe, local extension and spread.
- MRI is better method.

Note:

Incision biopsy of parotid tumour is contraindicated as chances of *seedling* and *recurrence* are high and also there is a chance of *injuring the facial nerve*, and chance of developing *parotid fistula* while doing the biopsy.

Open biopsy is contraindicated in parotid tumours due to:

- Chance of injury to facial nerve
- Seedling and high chance of recurrence
- Chance of parotid fistula formation

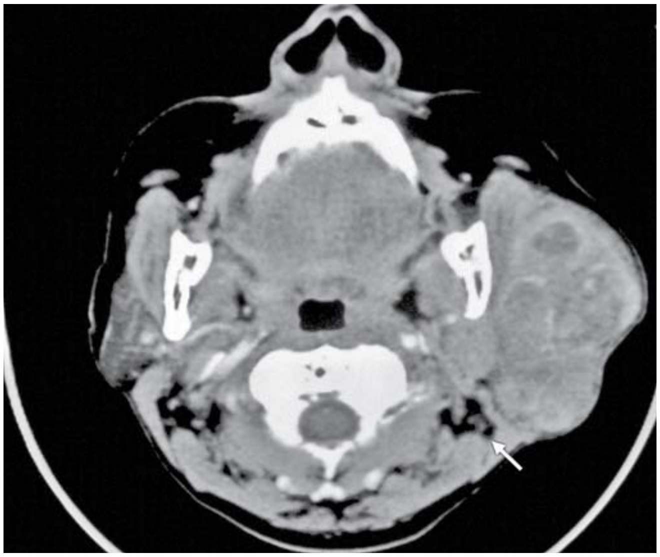

Fig. 4.25: CT picture of pleomorphic adenoma.

Anything the mind of man can conceive and believe, it can achieve.

Treatment

- Surgery—first line treatment.
- If only superficial lobe is involved, then superficial parotidectomy is done wherein parotid superficial to facial nerve is removed.
- If both lobes are involved, then total conservative parotidectomy is done by retaining facial nerve.

Note:

- Enucleation is avoided as it causes high recurrence due to extension of *tumour outside as pseudopods* across the capsule.
- Incomplete excision, 10% of tumours which are highly cellular are other causes for recurrence.
- RT is given after surgery even though it is benign.
- Inexplicable metastasis can occur even though it is benign.
- Tumour may implant due to spillage while surgical removal into retained residual parotid (deep lobe in superficial parotidectomy).

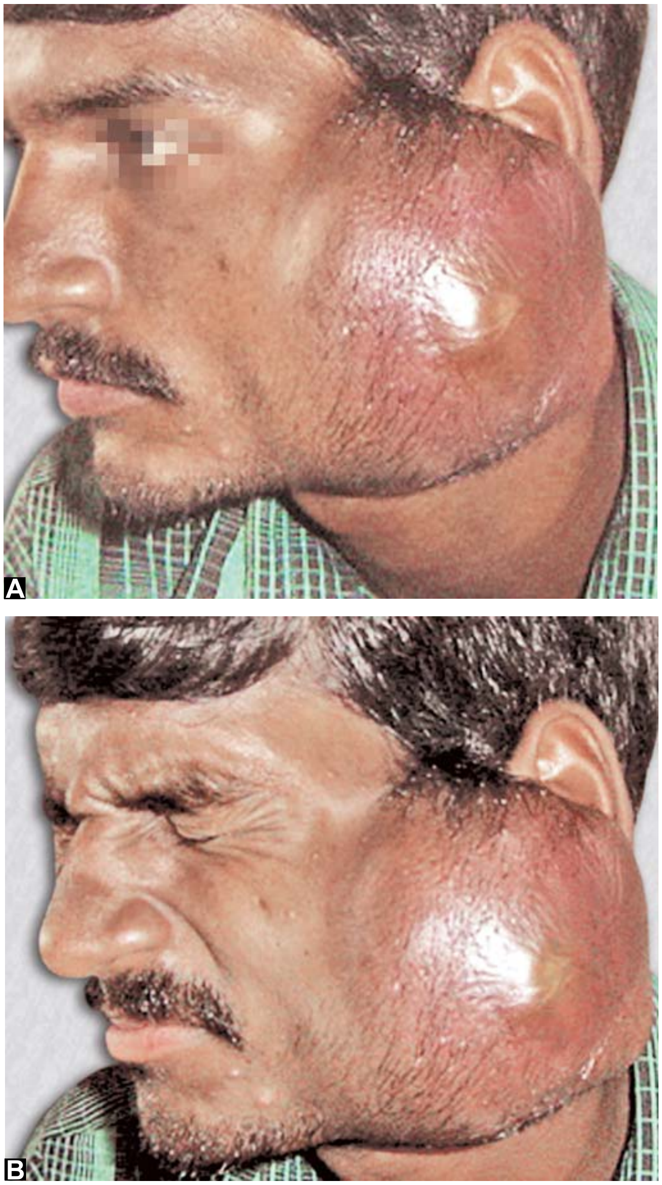

Figs 4.26A and B

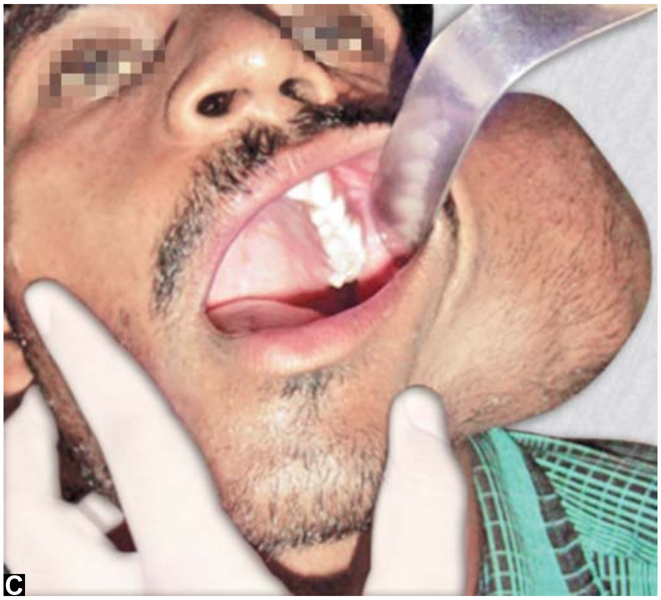

Fig. 4.26C

Figs 4.26A to C: Recurrent parotid tumour which has attained large size. Note facial nerve is intact. Duct orifice should be inspected using retractor.

- *Recurrence after parotidectomy* in pleomorphic adenoma is 5%. It is due to spillage, improper technique, inadequate margin, retained pseudopods, multicentricity. Recurrent tumour is multinodular without any capsule. *Expression of MUC1/DF3* in the tumour is marker to predict recurrence.

ADENOLYMPHOMA (Warthin's Tumour, Papillary Cystadenolymphomatosum)

- It is a misnomer. It is not malignant, it is not lymphoma.
- It is a benign tumour that occurs only in parotid, usually in the lower pole.
- It is said to be due to trapping of jugular lymph sacs in parotid during developmental period.

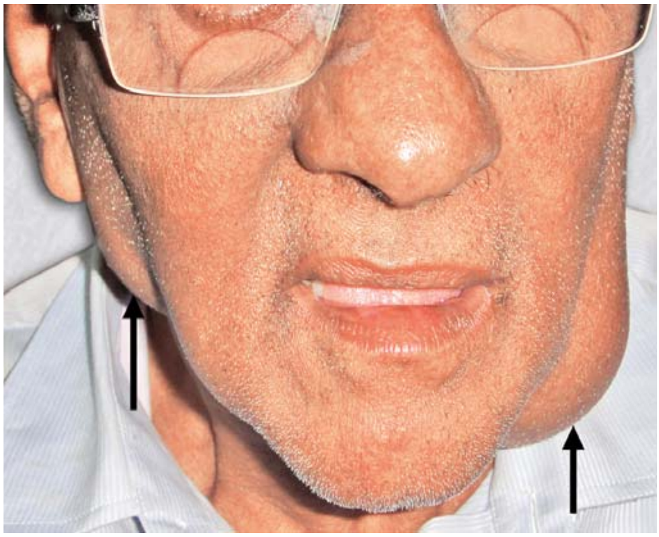

Fig. 4.27: Warthins tumour of parotids. It is common in males; often bilateral; common in elderly.

- It composed of *double layer of columnar epithelium*, with papillary projections into cystic spaces with lymphoid tissues in the stroma.
- It usually involves only superficial lobe of parotid gland. It may also be *multicentric*.
- Smoking (40%/8 times more risk than nonsmokers) and radiation exposure may be the cause.

Clinical Features

- It presents as a slow growing, smooth, soft, cystic, fluctuant swelling, in the lower pole, often bilateral and is nontender.
- It is common in males—4:1. Common in smokers.
- Common in old people – 60 years.
- Its incidence is 10%. Common in Whites.
- It is 2nd most common benign tumour.
- It is often bilateral—10%.

Investigations

- Adenolymphoma produces a "hot spot" in ^{99}Technetium pertechnetate scan—*it is diagnostic* (Due to high mitochondrial content).
- FNAC.

Adenolymphoma does not turn into malignancy. But occasionally it can simultaneously be associated with pleomorphic adenoma, carcinoma or lymphoma of parotid.

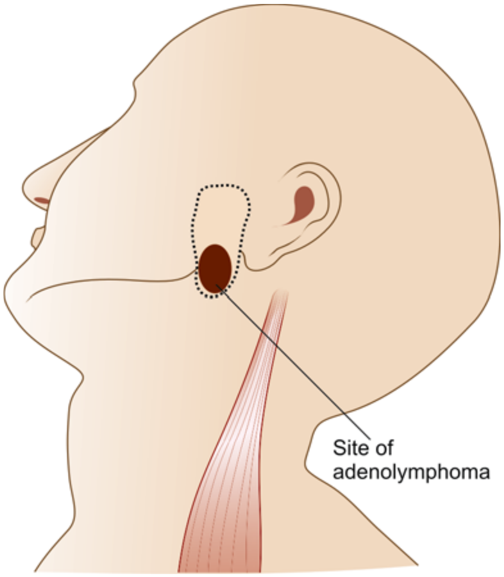

Fig. 4.28: Warthin's tumour. It is seen in parotid gland usually in superficial lobe, in lower pole.

Treatment

Superficial parotidectomy.

ONCOCYTOMA (Oxyphil Adenoma)

- It is < 1% of salivary tumours.
- Usually benign, originating from *oncocytes* (oxyphilic cells).
- Radiation and occupational hazards are the causes.
- Common in parotid; but rarely can occur in submandibular salivary gland.
- Gross - small, *tan coloured*, well circumscribed encapsulated solid tumour.
- Microscopy - *large oncocytes* with *swollen granular cytoplasm* due to abundant mitochondria. Tyrosine crystals are present in glandular spaces.
- Predilection for ^{99}Tc with hotspots and FNAC are the investigations.

BASAL CELL ADENOMA

- It is rare, benign, now classified under monomorphic adenoma containing isomorphic basaloid cells with basal layer and basement membrane.
- It is common in minor salivary glands; in major salivary glands it is multicentric.
- Grossly it looks like lymph node.
- *Microscopy*—isomorphic basaloid cells with solid/trabecular/tubular/membranous pattern.
- *Canalicular adenoma* is its variant with bilayered ribbons of columnar cells separated by vascular stroma.

MUCOEPIDERMOID TUMOUR

- It is the commonest malignant tumour in parotid.
- It is 2nd common malignant tumour in submandibular and sublingual salivary glands.
- It is commonest malignant tumour of parotid in childhood.
- Incidence is 9% of salivary tumours; 20% of malignant salivary tumours.
- It occurs both in major as well as minor salivary glands. Parotid is the commonest site; palate is the commonest minor salivary gland site (In the palate adenoid cystic carcinoma is common).

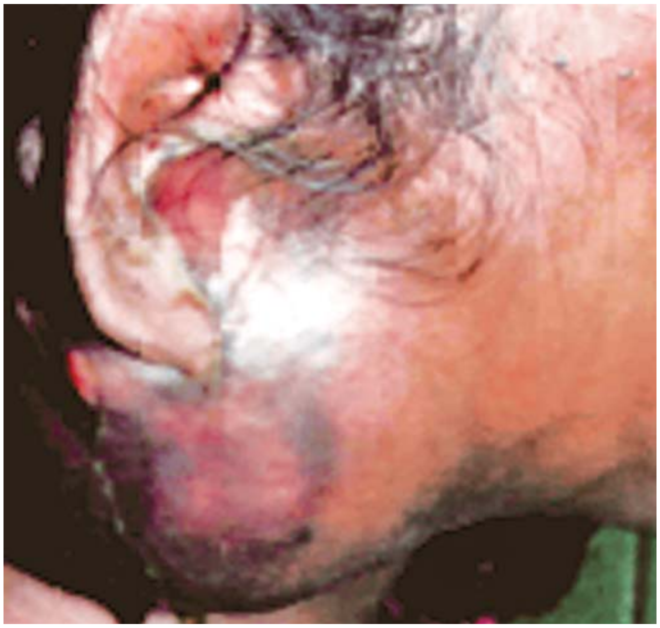

Fig. 4.29: Mucoepidermoid tumour.

- Radiation is the commonest etiological factor.
- *Gross*—unencapsulted solid tumour with cystic spaces.
- *Microscopy*—biphasic with *mucin secreting* (+ve for PAS, - ve for diastase) low grade and *epidermoid* with high grade; *clear cell* with intermediate type.
- It is slowly progressive, often attains a large size and spreads to neck lymph nodes.
- It contains malignant epidermoid and mucous secreting cells.

Types

- Low grade—*mucous cells* mainly—spreads to regional nodes.
- Intermediate—*clear cell* variety
- High grade—*epidermoid cells* mainly—spreads to regional nodes and also shows high propensity for distant spread.

Facial nerve involvement is late in mucoepidermoid carcinoma of parotid.

Clinical Features

- Swelling in the salivary (parotid or submandibular) region, slowly increasing in size, eventually attaining a large size, which is hard, nodular, often with involvement of skin and lymph nodes.
- Common in females (3:1); slow growing.
- Pain, skin and facial involvement are not common unless it is high grade.

ADENOID CYSTIC CARCINOMA (10% of Salivary Tumours)

- It is *most common* tumour in submandibular and sublingual salivary glands. 50% of cases occur in minor salivary glands—*palate*.
- It is also called as *cylidromatous carcinoma*.
- It is 2nd most common malignant salivary tumour; but it is rare in parotid (2% of parotid tumours, 15% of malignant parotid tumours.
- It is common in females (3:2). Common in 5th and 6th decades.
- It is s*low growing but highly malignant* with remarkable capacity for recurrence. But it is classified under *low grade* malignancy.
- *Microscopy*: Cribriform, tubular, and solid are 3 types. *Cribriform* type shows cells in nests separated by round or oval spaces—'*Swiss-cheese*' pattern. Myo and duct epithelial cells with lace like pattern are also seen. It *invades* periosteum and bone medulla early and spreads extensively.
- It has got high affinity for *perineural spread* (both axially and circumferentially; antegrade and retrograde fashion) along mandibular and maxillary divisions of trigeminal (*common*) nerve and facial nerve. It infiltrates nerve more proximally for long distance. Tumour may reach Gasserian trigeminal ganglion, pterygopalatine ganglion and cavernous sinus.
- Blood spread can occur to *lungs*, bones and liver. Lung secondaries may remain *dormant for many years* and so is not a contraindication for surgery of primary tumour. Blood spread can occur decades after removal of primary tumour.
- Radical parotidectomy/wide or radical excision of submandibular and sublingual glands with neck nodal dissection and postoperative RT is the treatment of choice.
- Positive margin, perineural spread, solid type on microscopy carry poor prognosis. Lung metastasis will not affect the prognosis.
- Local recurrence is common. 5-year survival is 70%.
- Regional nodal spread can occur but rare.

ACINIC CELL TUMOUR

- It is a rare, slow growing tumour that occurs almost always in parotid and is composed of cells alike serous acini.
- It is more common in women. It occurs in adult and elderly.
- It is 3% of salivary tumours; 90% occurs in parotid.
- Microscopically it can be microcystic (commonest), papillary, follicular, medullary etc.
- It can involve facial nerve or neck lymph nodes.
- Clinically it is of variable consistency with soft and cystic areas.
- It is low grade malignant tumour.

MALIGNANT MIXED TUMOUR (MMT)

- It is 10% of salivary malignancy in incidence with epithelial and mesenchymal elements.
- It carries worst prognosis.

Types

- *Carcinoma ex pleomorphic adenoma*: It is the commonest type. Previous long standing parotid swelling shows rapid change, fixity, facial nerve spread, neck nodal involvement are typical. Transformation is 2% in tumour of 5 years duration; 10% in 15 years tumour. It is the most aggressive salivary malignancy. Radical parotidectomy is the treatment.
- *Primary malignant mixed tumour*: It is also called *as carcinosarcoma* which arises as *de novo*. It shows components of both carcinoma and sarcoma with metastatic potential both through lymph nodes and blood.
- *Metastasizing mixed tumours*: It contains structures typical of benign mixed tumour both at original and at metastases sites.
- *In situ non-invasive carcinoma* in pleomorphic adenoma: There is no evidence of capsular invasion. Lesion with less than 8 mm invasion in depth shows 100% of 5-year survival; > 8 mm invasion carries 5-year survival < 50%.

ADENOCARCINOMA OF SALIVARY GLANDS

- It is 3% of parotid and 10% of submandibular and minor salivary gland tumours.
- It is equal in both sexes.
- It is common in children
- It can be tubular, papillary and undifferentiated.
- 20% involve facial nerve clinically.
- Undifferentiated type is aggressive.

SQUAMOUS CELL CARCINOMA OF SALIVARY GLANDS

- It is rare in salivary glands.
- In salivary glands, parotid is the common site.
- It is almost never seen in minor salivary glands.
- It is classified as high grade tumour.
- It is common in men (3:1).
- It occurs in 6th or 7th decade. It is aggressive nonencapsulated tumour arising from ductal system
- It grows rapidly causing pain, facial palsy, skin fixity, ulceration.
- It spreads commonly to neck nodes.
- It carries poor prognosis.
- Radical parotidectomy and RT is the treatment of choice.

SUBMANDIBULAR SALIVARY GLAND TUMOURS

Benign tumours:

- They are commonly pleomorphic adenomas, are smooth, firm or hard, bidigitally palpable, without involving adjacent muscles or hypoglossal nerve or mandible bone.
- Diagnosis is by FNAC, Orthopantomogram (OPG) and CT scan.
- Excision of both superficial and deep lobes of the gland is done.

Malignant tumours of submandibular salivary gland:

- They are hard, nodular, often get fixed to skin, muscles, hypoglossal nerve and mandible.
- Diagnosis is by FNAC of primary tumour and of lymph nodes when involved, CT scan and OPG.
- *Treatment:*
    - Wide excision, with removal of adjacent muscle, soft tissues and mandible.
    - If lymph nodes are involved, block dissection of neck (Classical neck dissection) is done.

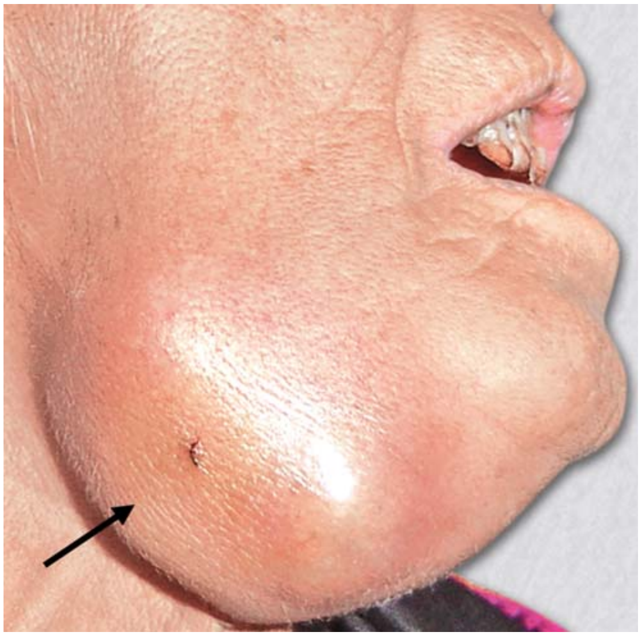

Fig. 4.30: Submandibular salivary gland tumour which is malignant. Patient underwent wide excision.

General features of malignant salivary tumours

- Fixation
- Resorption of adjacent bone
- Pain and anaesthesia in the skin and mucosa
- Muscle paralysis
- Skin involvement and nodularity
- Involvement of jaw and masticatory muscle
- Nerve involvement (facial nerve in parotid or hypoglossal nerve in submandibular salivary gland)
- Blood spread when occurs, commonly to lungs
- Mandibular branch of 5th cranial nerve may be involved when tumour tracks along the auriculotemporal nerve to the base of the skull causing severe pain in the distribution area

MANAGEMENT OF MALIGNANT SALIVARY TUMOURS

TNM Staging of Malignant Salivary Tumours (AJCC, 7th Edition, 2010, Clinical Staging)

T—Tumour
Tx—Tumour cannot be assessed.
T0—No evidence of primary tumour.
T1—Tumour < 2cm without extraparenchymal spread.
T2—Tumour 2-4 cm, without extraparenchymal extension
T3—Tumour >4 cm.
—or with extraparenchymal spread.
—but no facial nerve spread.
T4—Spread to facial nerve, skin, mandible, ear canal
—T_{4a}.
—or spread to base of skull, pterygoid plates, encased external carotid artery – T_{4b}.

Staging

Stage I – T1N0M0
Stage II – T2N0M0
Stage III – T3N0/N1Mo
Stage IVA – T4aN2N1N0M0
Stage IVB – T4b any NM0; any T N3M0
Stage IVC – Any T any N M1

Note:

Pathological TNM staging is also present which differs from clinical in many points. Example—T3 is tumor of size 4-6 cm; T4 is size > 6 cm.

N—***Lymph node***

Nx — Nodes not assessed.
N0 — Regional nodes not involved.
N1 — Single ipsilateral node < 3 cm.
N2a — Single ipsilateral node 3-6 cm.
N2b — Multiple ipsilateral nodes < 6 cm.
N2c — Bilateral or contralateral nodes < 6 cm.
N3 — Single node spread > 6 cm.

In time of test, family is best.

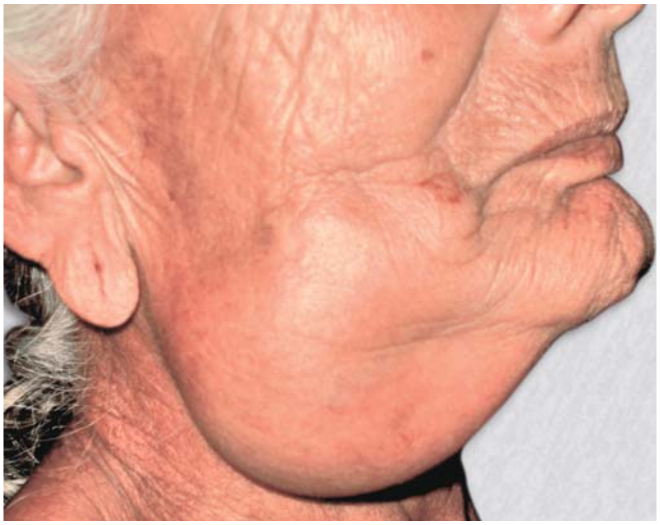

Fig. 4.31: Submandibular salivary gland tumour in old female.

M—***Metastases***

M0 — No blood spread.

M1 — Blood spread present.

Specific Investigations

- FNAC.
  - FNAC also confirms possibility of (35% of patients) lymphoma/inflammatory masses.
  - FNAC allows preoperative counselling regarding nature of tumour, likely extent of resection (conservative/radical), management of facial nerve (high grade adenoid cystic), and likelihood of neck dissection (high grade).
  - But FNAC may cause haematoma leading to difficulty in surgical dissection. Sampling problem, difficulty in finding histological type.
  - Conservative approach for benign tumours in high risk surgical candidates may be thought of.
- CT scan to see the deep lobe of the parotid: the involvement of bone, extension into the base of the skull; relation of tumour to internal carotid artery, styloid process, deep lobe tumour, parapharyngeal space extension. Neck nodes are better assessed. Bony changes in the foramina, erosions and sclerotic margins in fissures or canals, enlarged diameter of canal and fissure are the CT features of perineural spread.
- *MRI is very useful* to find out perineural spread, bone marrow involvement, skull bones, internal architecture and intracranial extensions, recurrent tumours. Replacement of perineural fat with tumour, contrast (gadolinium) enhancement, increased size of the nerve are features of perineural spread.
- OPG.
- Blood grouping and cross matching; required amount of blood is kept ready.

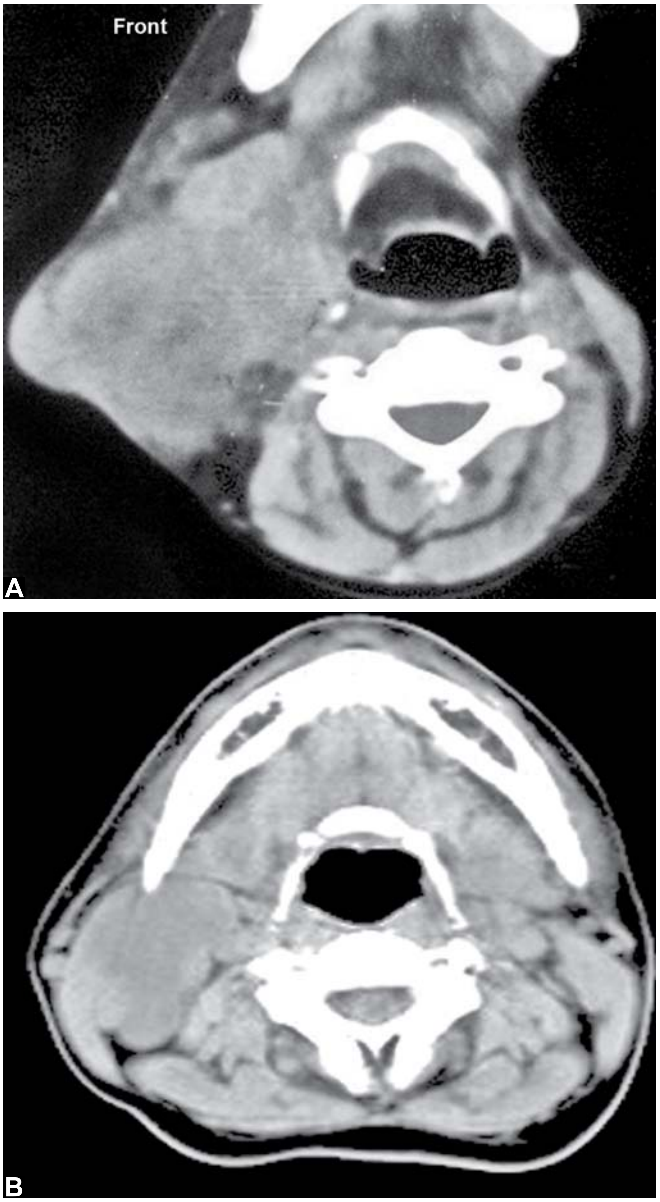

Figs 4.32A and B: CT scan pictures of salivary tumours.

- FNAC of lymph node.
- MRI shows better soft tissue definition than CT scan. *Sialogram is not useful in assessment of tumour.*

Treatment

In Parotid

- ***Surgery***

Indications for surgery

- T1, T2, T3 tumours of low grade—total conservative parotidectomy
- T4 tumours, high grade tumours, SCC—radical parotidectomy

It includes facial nerve sacrifice, may involve resection of skin, mandibular ramus, masseter muscle, infratemporal fossa dissection, subtotal petrosectomy.

Note: In T1 low grade, superficial parotidectomy is often practiced; but not ideal.

a. *Radical parotidectomy* is done which includes removal of both lobes of parotid, soft tissues, part of the mandible with facial nerve.
   - It is done in high grade malignant tumours and squamous cell carcinoma.

Indications for facial nerve sacrifice

- Preoperative weakness/paralysis of nerve
- Intraoperative evidence of gross invasion even in presence of normal preoperative function
- Tumours transgressing through facial nerve from superficial to deep lobe
- Nerve stump is checked for frozen section for negative margins, if positive mastoidectomy and nerve dissection in temporal bone is needed

   - Facial nerve is reconstructed using greater auricular nerve, or sural nerve. All branches except buccal branch are repaired using cable graft. Nerve graft is not a contraindication for future RT.
   - Often lateral tarsorrhaphy or temporal sling reconstruction is done.

b. *Total conservative parotidectomy* is becoming popular in many parotid malignancies.

Complications of surgery

- Haemorrhage
- Infection, flap necrosis
- Fistula
- Frey's syndrome
- Facial nerve palsy
- Facial numbness
- Numbness in ear lobule due to injury to great auricular nerve
- Sialocele

♦ ***Postoperative radiotherapy***

It is quite useful to reduce the chances of relapse. Usually, external radiotherapy is given. It is given in all carcinomas, but more useful in adenoid cystic and squamous cell carcinomas.

Indications for radiotherapy

- T3, T4 tumours
- High grade tumours
- Perineural spread
- Adenoid cystic carcinomas
- Deep lobe tumour
- Vascular involvement
- Close clearance margin
- Multiple neck nodes involvement
- Recurrent malignant tumours
- Recurrent pleomorphic adenoma
- Spillage after surgery for pleomorphic adenoma
- Residual/refractory tumours/nerve involvement
- Inadequate clearance margin

  Note:
  - It is given in 3-6 weeks after surgery. Dose is 50-70 Gy; 1.5- 2.0 Gy in 5-8 weeks
  - RT is delayed for 6 weeks if nerve grafting is done
  - Neuron beam therapy is used

Complications

- Xerostomia
- Osteoradionecrosis of temporal bone/mandible
- Skin ulcers, mucositis
- Fibrosis of optical apparatus, brainstem injury
- Trismus due to fibrosis of masseter, pterygoids and TM joint
- Otitis media, localized hair loss

♦ ***Chemotherpy***

It is also given. Drugs given here depends on tumour type. Intra-arterial chemotherapy is beneficial. But overall efficacy of chemotherapy is very less compared to RT. 5 FU, cisplatin, doxorubicin, epirubicin, cetuximab are used.

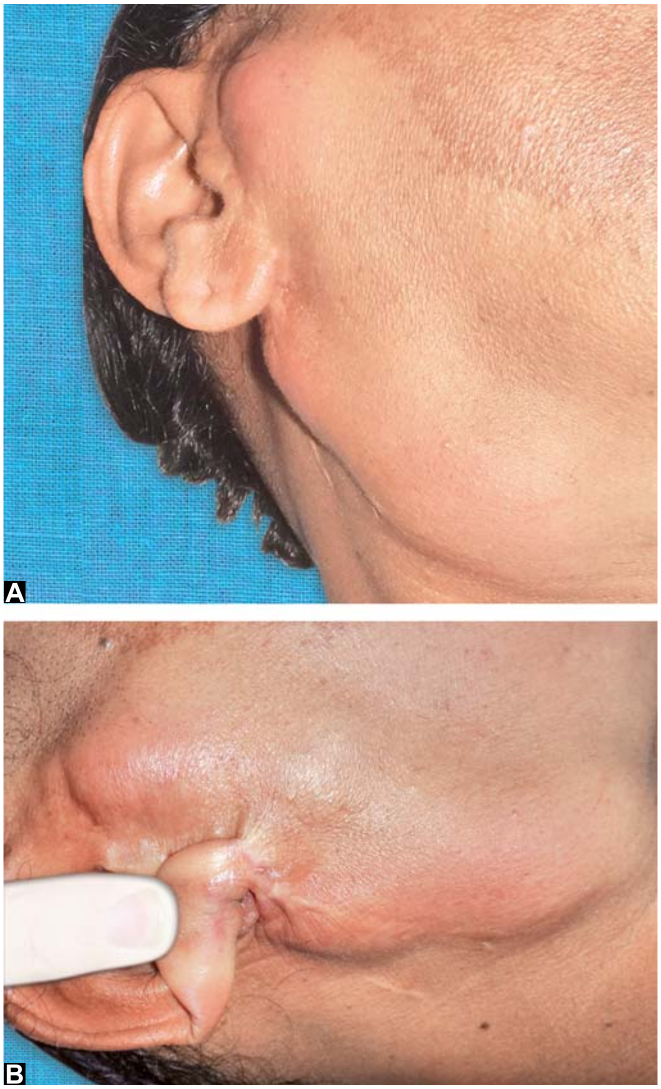

Figs 4.33A and B: Recurrent parotid tumour right side. Note the scar of previous surgery. It was adenoid cystic carcinoma. Recurrent has occurred after 6 years. Patient underwent radical parotidectomy. Neck nodes are involved in this patient even though it is rare in adenoid cystic carcinoma.

You will soon break the bow if you keep it always stretched.

- ***Preoperative radiotherapy***
  It is given in large tumours to reduce the size and make it better operable, i.e. to down stage the disease.
- If lymph nodes are involved, which is confirmed by FNAC, radical neck dissection is done. It is also done in N0 with high-grade tumour or T3/T4 tumours.

In Submandibular Salivary Gland

Wide excision is done, with removal of mandible, and soft tissues around—extraglandular excision. If lymph nodes are involved, then block dissection of the neck is done.

MINOR SALIVARY GLAND TUMOURS

- It is 10% of salivary tumours.
- It is common in — *palate* (40%).
  — *lip*.
  — *cheek*.
  — *sublingual glands*.
- *Palate* is the commonest site.
- 10% are benign—*commonly pleomorphic adenomas*.
- 90% are malignant—*commonly adenoid cystic carcinomas*.
- They present as swelling with ulcer over the summit.
- If it is malignant, then extension into the palate, maxilla, pterygoids can occur often with involvement of the lymph node.

Differential Diagnosis

Squamous cell carcinoma of oral cavity.

Investigations

1. Incision biopsy.
2. CT scan.
3. X-ray maxilla.
4. FNAC of lymph node.

Treatment

- Wide excision often with palatal excision or maxillectomy is done—for malignancy.
- If the tumour is *less than 1 cm* in size excision biopsy is done with 1 cm clearance margin. If the tumour is *more than 1 cm* in size, initially incision biopsy is done and then wide excision is done. Even in larger defect in hard palate region, it usually re-epithelialise if once left open to allow it to granulate. If bony palate is infiltrated then that part of palate bone is removed to get a clearance; area is reconstructed by moulds, synthetic materials.
- Reconstruction by dental plates, skin grafting, or flaps are done.
- Lymph node block dissection of the neck is done if involved.
- Excision with primary closure is done for benign tumours.

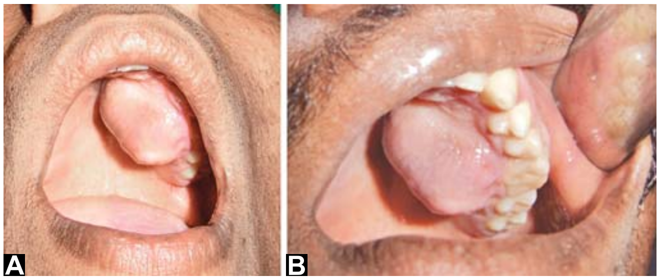

Figs 4.34A and B: Minor salivary gland tumour in the palate. Palate is the common site of minor salivary gland tumour.

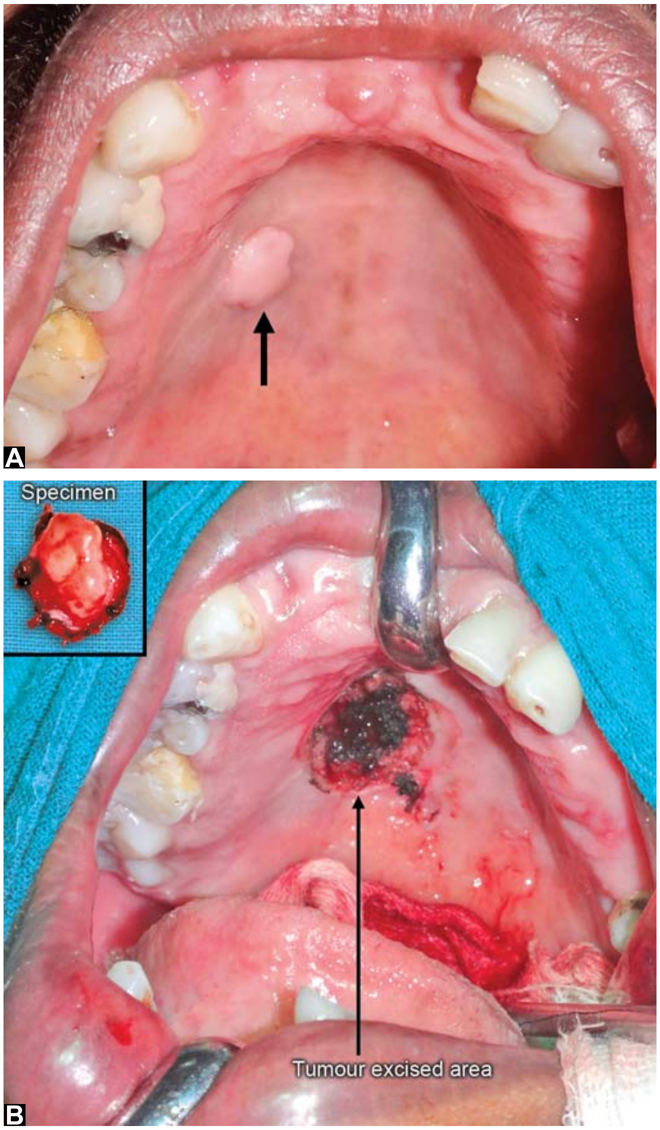

Figs 4.35A and B: Minor salivary gland tumour of hard palate. It is excised with 1 cm margin; area is allowed to granulate and epithelialise.

Points to be remembered

- ❖ Salivary gland tumours are usually benign in an adult
- ❖ It is rare in children but when it occurs, it is commonly malignant
- ❖ Clinical and FNAC are diagnostic methods
- ❖ Open biopsy is contraindicated
- ❖ Sialogram is not useful in salivary tumours
- ❖ CT scan or MRI are often needed
- ❖ Nerve should be preserved in benign lesions
- ❖ Nerve can be sacrificed to achieve clearance in malignancies

PAROTID LYMPHOMA

- Parotid lymphoma can occur from the lymph nodes in the gland or from parotid parenchyma.
- It can occur in HIV patients; lymphoepithelial diseases and in Sjögren's syndrome.
- Common in elderly.
- Disease may be confined to parotid gland or may involve other nodes in neck, mediastinum.
- When it is confined to parotid total parotidectomy with radiotherapy and later chemotherapy is the treatment.
- When many other nodes are involved chemotherapy is the choice therapy.

Note:
Lymphoma occasionally can occur in other salivary glands also (10% of all salivary lymphomas).

PAROTIDECTOMY

Types

1. *Superficial parotidectomy*: It is the removal of superficial lobe of the parotid (superficial to facial nerve). Done in case of benign diseases of superficial lobe of the parotid.

Steps in parotidectomy

- Lazy 'S' incision—modified Blair's/Sistrunk's approach and raising the skin flaps
- Mobilisation of the gland
- Flap is reflected in front *just up* to anterior margin of the parotid; never beyond. After identification of sternocleidomastoid great auricular nerve is identified and can be sacrificed. Posterior belly of digastric is identified.
- Location of *stylomastoid branch* of posterior auricular artery is anterior to facial nerve trunk which enters the stylomastoid foramen.
- *Thrust* the mosquito haemostat 5 mm in front of facial nerve; *open* the blades for 5 mm; *lift* the blades for 5 mm.
- One should worry about the nerve not small bleeding and haemostasis.
- Identification of facial nerve trunk
- Dissection of the gland off the facial nerve using bipolar cautery
- Removal of parotid—superficial/both
- Distilled water (hypertonic) irrigation to kill spilled tumour cells
- Haemostasis and closure with a suction drain

2. *Total conservative parotidectomy*: It is done in benign diseases of parotid involving either only deep lobe or both superficial and deep lobes. Here both lobes are removed with preservation of facial nerve. Here initially superficial parotidectomy is done and facial nerve and its branches are retracted gently and deep lobe is removed.
3. *Radical parotidectomy*: Both lobes of parotid are removed along with facial nerve, fat, fascia, muscles (masseter,

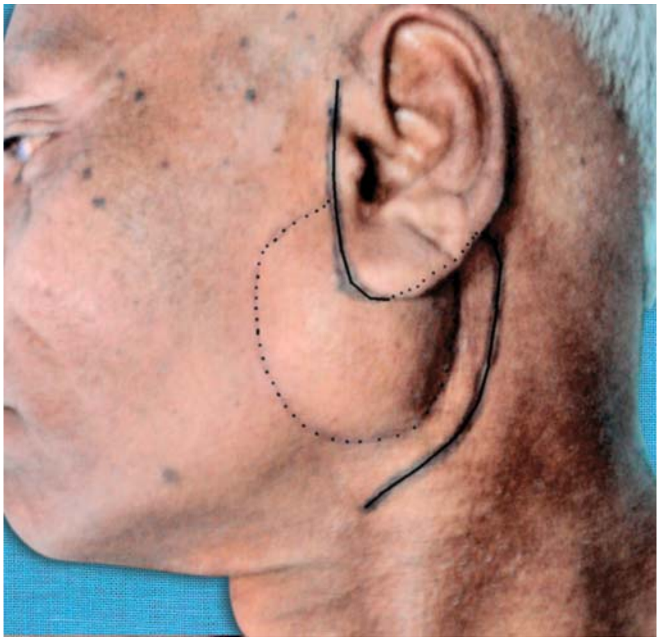

Fig. 4.36: Incision for parotidectomy. Typical lazy 'S'.

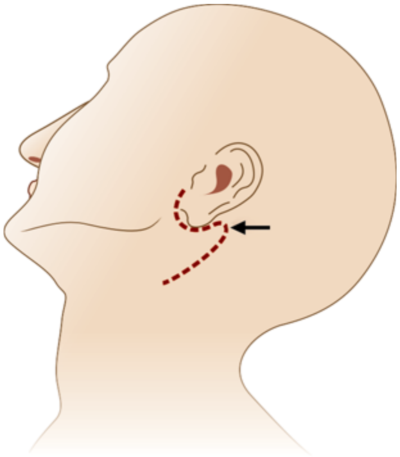

Fig. 4.37: 'S' shaped incision for parotidectomy.

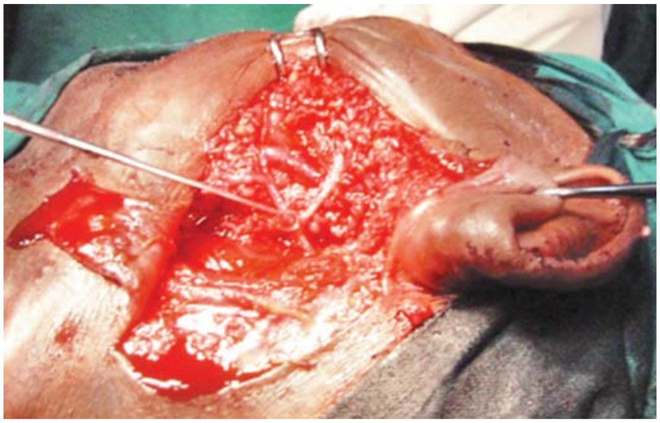

Fig. 4.38: On table exposure of facial nerve branches after superficial parotidectomy for a benign tumour.

pterygoids and buccinator), lymph nodes. It is done in case of carcinoma parotid. Later facial nerve reconstruction is done using great auricular nerve graft.

Conservative surgeries are becoming popular for malignancy but they are not universally accepted.

4. *Suprafacial parotidectomy* is done in lower pole parotid tumours wherein all branches of the facial nerve need not be dissected.

Identification of facial nerve

- Facial nerve is 1 cm deep and below the tip of the inferior portion of the cartilaginous canal—*Conley's point*
- By nerve stimulator
- It is inferomedial to tragal point
- Deep to digastric muscle and tympanic plate
- Nerve is just lateral to the styloid process
- Tracing branch from distal to proximal (Hamilton-Bailey technique)

Complications of parotidectomy

- Facial nerve injury
- Haemorrhage
- Salivary fistulas
- Infection—Flap necrosis is common
- Frey's syndrome
- Sialocele
- Numbness over the face and ear - due to injury to great auricular nerve. Female patients find difficult to wear ear rings. Spontaneous recovery may occur in 1½ to 2 years.

FREY'S SYNDROME (Auriculotemporal Syndrome, Gustatory Sweating); (Lucie Frey—Polish Surgeon—1932)

- Occurs in 10% of cases.
- It is due to injury to the auriculotemporal nerve, wherein post-ganglionic parasympathetic fibres from the otic

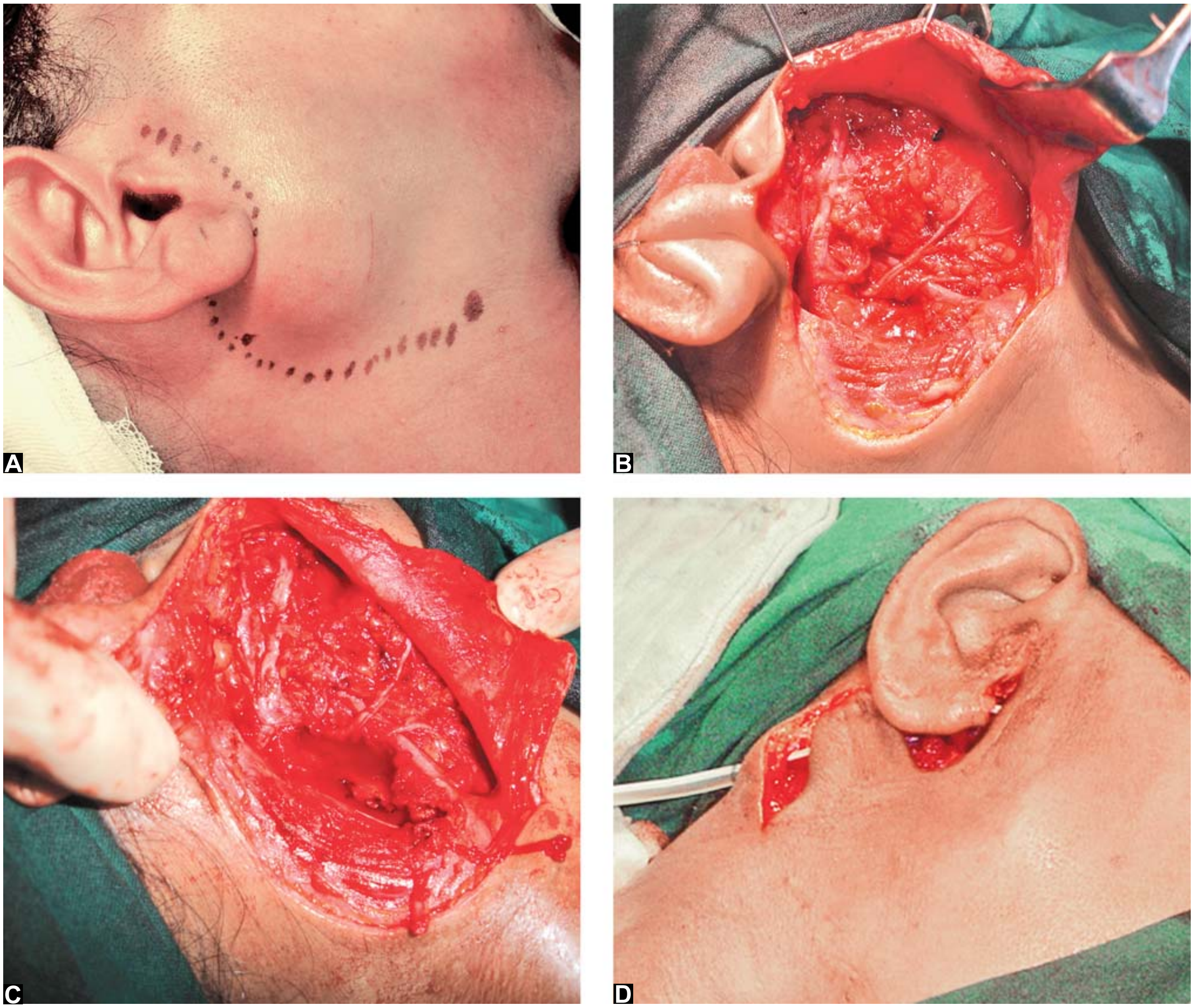

Figs 4.39A to D: Steps in parotidectomy and demonstration of facial nerve. Note the placement of drain after surgery (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and HOD, KMC, Mangalore; Dr Sampath, MS, ENT, KMC, Mangalore).

ganglion become united to sympathetic nerves from the superior cervical ganglion (Pseudosynapsis). *There is inappropriate regeneration* of the damaged parasympathetic autonomic nerve fibers to the overlying skin.

- Auriculotemporal nerve has got two branches. Auricular branch supplies external acoustic meatus, surface of tympanic membrane, skin of auricle above external acoustic meatus. Temporal branch supplies hairy skin of the temple. Sweating and hyperaesthesia occurs in this area of skin.

Causes

1. Surgeries or accidental injuries to the parotid.
2. Surgeries or accidental injuries to temporomandibular joint.

Clinical Features

- Flushing, sweating, erythema, pain and hyperaesthesia in the skin over the face innervated by the auriculotemporal nerve, whenever salivation is stimulated (i.e. during mastication).
- Condition causes real inconvenience to the patient.
- Involved skin is painted with iodine and dried. Dry starch applied over this area will become blue due to more sweat in the area in Frey's syndrome—***Starch iodine test.***

Treatment

- Initially conservative and reassurance. Most often they recover without any active treatment in 6 months. Antiperspirants, anticholinergics like scopolamine 3%, glycopyrrolate 1%, methyl sulfate, radiation 50 Gy are used.

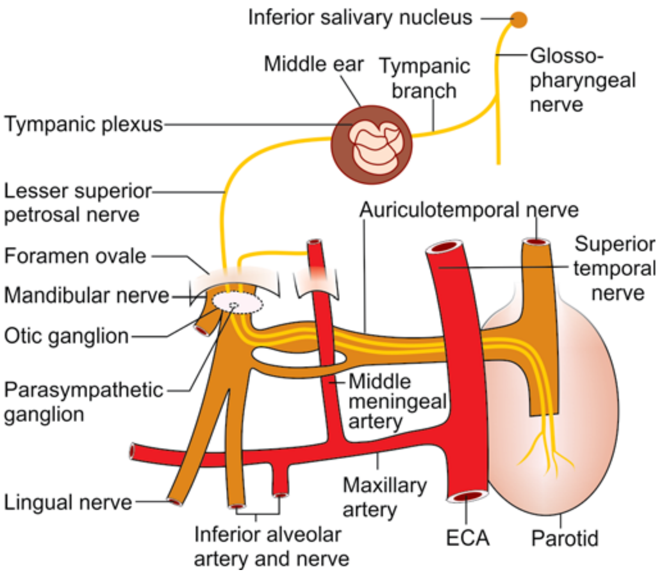

Fig. 4.40: Parasympathetic and sympathetic supply of the parotid gland. Otic ganglion located just below the foramen ovale close to mandibular nerve relays the parasympathetic supply from glossopharyngeal nerve through tympanic branch, tympanic plexus in middle ear and then lesser superficial petrosal nerve. Postganglionic fibres run in auriculotemporal nerve along with unrelayed sympathetic fibres from middle meningeal plexus.

- Occasionally (10%) they require surgical division of the tympanic branch of the glossopharyngeal nerve below the round window of middle ear [i.e. intratympanic parasympathetic (Jacobsen nerve) neurectomy].
- Dermal/fat graft; avulsion of auriculotemporal nerve; interposition of temporal fascia, fascia lata, sternomastoid muscle, acellular human dermal collagen; alcohol injection—are all tried.

Note:

Incidence is higher with a flap elevation superficial to platysma. Elevation of thick flap or insertion of fat flap under skin during surgery may reduce the chance of Frey's syndrome.

Treatment of Frey's syndrome

- Reassurance
- Jacobsen neurectomy (tympanic)
- Injection of botulinum toxin to the affected skin
- Antiperspirants like aluminium chloride
- Syndrome can be *prevented* on table by placing muscle (sternomastoid) or fascial (temporalis) flaps or artificial membranes over parotid bed, under the skin

FACIAL NERVE INJURY (Lower Motor Nerve Lesion, Surgically Related)

Causes

- Trauma
- Surgery—parotidectomy, drainage of parotid abscess
- Compression of facial nerve—Bell's palsy.
- Incidence of temporary/transient facial nerve palsy after parotidectomy is 30%. Recovery occurs usually in 12 weeks.
- It is due to transection of trunk or branches or excessive traction or over use of nerve stimulator.

Clinical Features

- Inability to close the eyelid.
- Difficulty in blowing and clenching.
- Drooping of the angle of the mouth.
- Obliteration of nasolabial fold.
- Loss of forehead wrinkles.
- Wide palpebral fissure.
- Epiphora.

Treatment

- Nerve grafting using greater auricular nerve, sural nerve, lateral cutaneous nerve of thigh or hypoglossal nerve.
- Suspension of angle of mouth to zygomatic bone using temporal fascia sling.
- Lateral tarsorrhaphy—to prevent corneal ulceration.
- Medial canthus reconstruction—to reduce epiphora.
- Cross facial nerve transplantation from opposite side using its insignificant branches.
- Dynamic neurovascular muscle graft.
- Upper lid gold weights to protect cornea.

Surgeries for facial nerve palsy

- *Static*
  - Suspension surgeries using temporal fascia
  - Correction of medial canthus
  - Lateral tarsorrhaphy—to prevent exposure keratitis due to widened palpebral fissure
  - Upper lid weights
- *Dynamic*
  - Muscle transfer—temporal to masseter
  - Free muscle graft. Gracilis muscle neurovascular transfer
  - Cross facial nerve transplant from opposite facial nerve to injured facial nerve using sural nerve
  - Nerve grafts

Heerfordt's syndrome is sarcoidosis of parotid swelling; anterior uveitis; facial palsy and fever.

Remember

- **Taste sensation** and general sensation (lingual nerve) should be checked. Patient is not allowed to speak but *asked to write* in a paper. Taste material is instilled on the surface of the diseased side first and then normal side. Prior to each instillation patient should wash his mouth with warm water. Usually four substances are used. After 10 seconds patient should identify the substance and write. *Facial nerve serves 3 tastes—salt* (rock salt) on the tip of tongue; sweet using syrup on the tip of the tongue; *sour* using lemon juice on the lateral aspect of the tongue. Bitter taste is mediated by glossopharyngeal nerve and is tested using quinine on posterior third of the tongue.
- **Secretomotor fibres of parotid:** Secretomotor preganglionic fibres from *inferior salivary nucleus* → glossopharyngeal nerve → tympanic branch → tympanic plexus → lesser superficial petrosal nerve → otic ganglion → post ganglionic fibres → auriculotemporal nerve, branch of mandibular division of trigeminal nerve → parotid gland.
- **Secretomotor fibres of submandibular salivary gland:** Preganglionic fibres from *superior salivary nucleus* → facial nerve → chorda tympani nerve → lingual nerve → Langley's submandibular ganglion → postganglionic fibres → submandibular and sublingual salivary glands.
- Parotid gland is *serous*. Submandibular gland is *mixed* (major is mucous). Sublingual is *mucous*. Minor salivary glands are *mucous* except *von Eber's glands* which empty into the circumvallate papillae and glands in the tongue tip.
- *70% of total saliva* is from submandibular salivary gland. *25% of saliva* is from parotids. 10% of total salivary volume is from minor salivary glands.
- Normal salivary secretion per day is 1500 ml. It is hypotonic fluid with pH 7.0. It contains α-amylase.

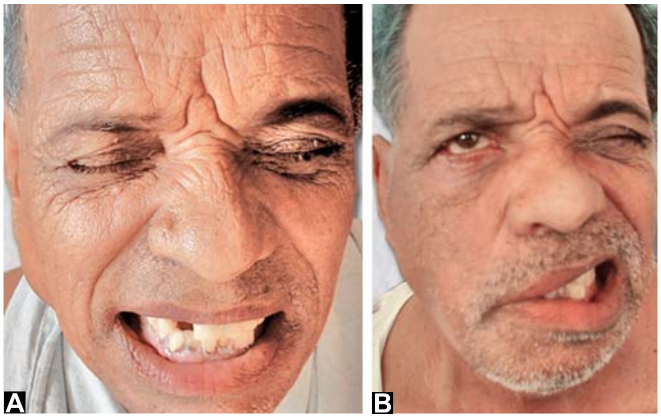

Figs 4.41A and B: Features of post-parotidectomy facial nerve palsy.

Note:

Accessory parotid tumour—Accessory parotid tumour is very rare tumour arising from accessory parotid usually above the parotid duct level in front of the master muscle. It can be benign or malignant. Pleomorphic adenoma is common. Tumour behaves like tumour from main parotid gland. Incidence is 1% of all parotid tumours. 30% of them are malignant. It is located in a line at central 1/3rd of the line joining the middle of the tragus to a point between the ala of the nose and vermilion border of upper lip. FNAC confirms the diagnosis. CT scan confirms the anatomical location. Main parotid gland is usually normal. Surgery is the treatment. Standard parotidectomy incision or direct cheek approach can be used. But direct cheek approach can cause higher incidence of buccal and zygomatic branches of facial nerve injuries (40%). Standard parotidectomy approach has got less chance of nerve branch injuries. Lymph node spread in the neck is dealt with radical dissection. Excision cures the benign disease.

Chapter 5 Neck

CHAPTER OUTLINE

- Anatomy of Lymphatics of Head and Neck
- Thoracic Outlet Syndrome
- Cervical Rib
- Branchial Cyst
- Branchial Fistula
- Pharyngeal Pouch
- Laryngocele
- Cystic Hygroma
- Ludwig's Angina
- Parapharyngeal Abscess
- Retropharyngeal Abscess
- Subhyoid Bursitis
- Carotid Body Tumour
- Torticollis
- Sternomastoid Tumour
- Tuberculous Lymphadenitis
- Cold Abscess
- Secondaries in Neck Lymph Nodes
- Chemotherapy for Head and Neck Cancers

ANATOMY OF LYMPHATICS OF HEAD AND NECK

Waldeyer's Lymphatic Ring (Inner)

It consists of adenoids above, lingual tonsils below and two palatine tonsils laterally one on each side.

Outer Circular Chain of Nodes (Outer Waldeyer's Ring)

Occipital, postauricular, preauricular, parotid, facial, submandibular, submental, superficial cervical and anterior cervical.

Facial nodes are:

a. Superficial
  - Upper—infraorbital.
  - Middle—buccinator.
  - Lower—supramandibular.

b. Deep groups—in relation to pterygoids.

Submandibular lymph nodes drain

- The side of the nose.
- The cheek.
- Angle of the mouth.
- Entire upper lip.
- Outer part of the lower lip.
- The gums.
- Side of the tongue.

Submental lymph nodes: Drain from the central part of the lower lip, floor of the mouth and apex of the tongue.

Superficial cervical nodes: They lie on outer surface of the sternomastoid around the external jugular vein. They drain the parotid region and lower part of the ear.

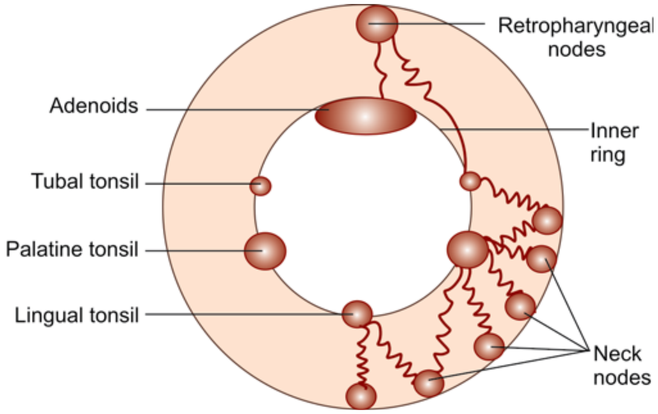

Fig. 5.1: Inner and outer Waldeyer's ring anatomy.

Deep cervical lymph nodes: Upper deep cervical lymph nodes—jugulodigastric nodes Below the digastric and in front of IJV.

Lower deep cervical lymph nodes—jugulo-omohyoid nodes—Above the omohyoid and behind the IJV.

They drain the ipsilateral half of head and neck, finally form a jugular lymph trunk from lower deep cervical nodes to join thoracic duct on the left side, and the junction of right subclavian and right jugular vein on right side.

Rule of 7 in the neck

- 7 days—inflammation
- 7 months—neoplasm
- 7 years—congenital defect

Note: The Rule of 7 provides a probable diagnosis of the neck mass based on the average duration of the patient's symptoms

Do not worry and hurry so much rather walk this earth lightly and leave your mark.—**Dr Saroj Kanna**

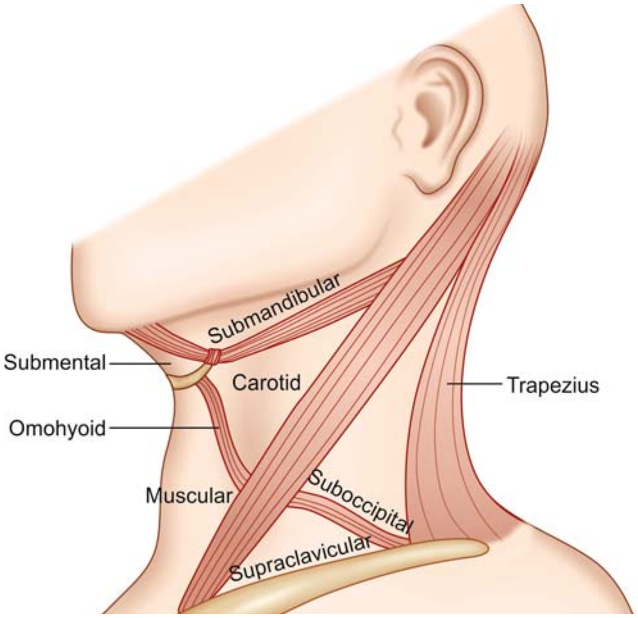

Fig. 5.2: Triangles of the neck.

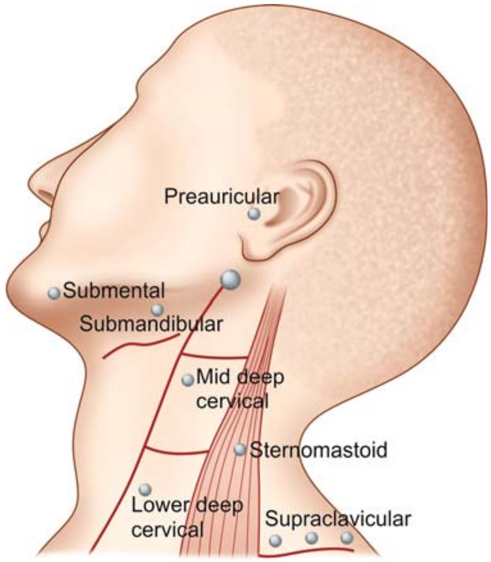

Fig. 5.3: Lymphatic drainage of the neck.

THORACIC OUTLET SYNDROME (TOS)

It is syndrome complex due to neurovascular bundle compression in the thoracic outlet.

Thoracic outlet has got two main spaces:

- *Scalene triangle* is bound by scalenus anterior, scalenus medius and first rib. It contains subclavian artery and brachial plexus.
- *Costoclavicular space* is bound by clavicle, first rib, costoclavicular ligament and scalenus medius. It contains subclavian artery and vein and brachial plexus.

Causes

- Cervical rib.
- Long C_7 transverse process.
- Anomalous insertion of scalene muscles.
- Scalene muscle hypertrophy.
- Scalene minimus.
- Abnormal bands and ligaments.
- Fracture clavicle or first rib.
- Exostosis.
- Tumours in the region.
- Brachial plexus trauma and diseases.

Differential Diagnosis of TOS

- Carpal tunnel syndrome.
- Cervical spondylosis.
- Spinal canal tumours.
- Shoulder myositis.
- Angina.
- Raynaud's disease.
- Spinal stenosis.
- Ulnar nerve compression, epicondylitis.

Clinical Features

Neurological symptoms

- Paraesthesia.
- Pain in shoulder, arm, forearm and fingers.
- Occipital headache as referred pain from tight scalene muscles.
- Weakness in forearm, hand.

Vascular symptoms

- Claudication, ischaemic ulcers, gangrene.

Signs

- Scalene muscle tenderness.
- Pulsatile swelling in supraclavicular region with thrill and bruit (25%).
- Bony mass above clavicle.
- Adson's test (+ve).
- Roos test (+ve).
- Elevated arm stress test (+ve).
- Costoclavicular compression maneuver.
- Hyperabduction maneuver.

} reveals feeble/absence of radial pulse due to compression of the subclavian artery.

- Poor capillary refilling.
- Absence or feeble pulse.

(*Please refer Chapter on Arterial Diseases for details of tests*).

Investigations

- X-ray neck and cervical spine.
- Doppler.
- Subclavian angiogram, CT angiogram, CT neck.
- Nerve conduction studies, electromyography.

Treatment

- *Conservative*—if nerve velocity is > 60m/second
- *Surgical*—if nerve velocity is < 60m/second

Conservative treatment for TOS

- Exercises—neck stretching, postural and breathing exercises
- Drugs—analgesics, muscle relaxants, antidepressants
- Avoid weight lifting
- Physiotherapy

Surgical treatment of TOS

- Transaxillary (ROOS)—mainly for first rib excision and also cervical rib
- Supraclavicular approach for cervical rib and soft tissue excision, scalenotomy, neurolysis, arterial reconstruction
- Cervical sympathectomy may be needed

CERVICAL RIB

Definition

- It is an extension of costal element (anterior part) of transverse process of C$_7$ vertebra more than 2.5 cm (normal).
- Syndrome caused by it is called as *cervical rib syndrome, thoracic-inlet syndrome, thoracic-outlet syndrome, scalene syndrome*.
- It is 0.46% common; common in females; more frequent on right side.
- It can be *unilateral or bilateral (50%)* can be *asymptomatic* or *symptomatic*.

Types (Refer Fig. 5.4)

1. *Complete bony*: Cervical rib is radio-opaque, anteriorly ends over the first rib or manubrium.
2. *Complete fibrous*: Cannot be demonstrated radiologically.
3. *Combined*: Partly bony partly fibrous.
4. Partial bony: With free end expanding as *bony mass, which is felt in the neck.*

Pathology

- Cervical rib narrows the scalene triangle (bounded by scalenus anterior, scalenus medius and first thoracic rib below).

Compression of subclavian artery; C$_8$ and T$_1$ nerve roots due to cervical rib.
↓
Angulation of subclavian artery occurs.
↓
Causes constriction of artery at the site where artery crosses the cervical rib.
↓
'Eddie's current' created in the blood flow causes sudden release of pressure distal to the narrowing.
↓
Poststenotic dilatation → *Venturi phenomenon*.
(due to vessel wall ischaemia)↓
Stasis of blood occur.
↓
Thrombosis → *Embolus*.
↓ ↓
Features of ischaemia in the hand and forearm.
Later digital gangrene occurs.

- Rarely thrombus may extend proximally into the subclavian artery causing vertebrobasilar insufficiency.

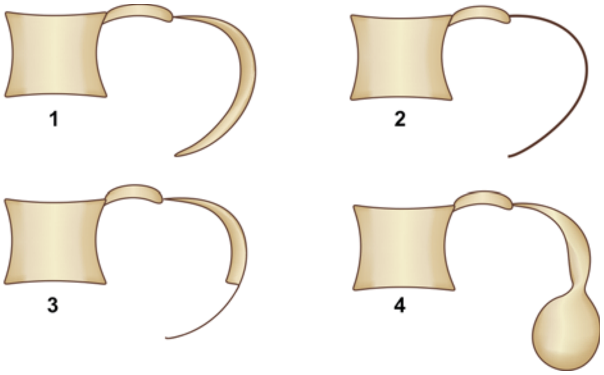

Fig. 5.4: Cervical rib types.

- Compression of C$_8$ and T$_1$ causes tingling and numbness along its distribution, i.e. in the little finger, medial side of hand and forearm.
- ***Paget-Schroeter syndrome*** is subclavian vein compression by cervical rib. It is rare.

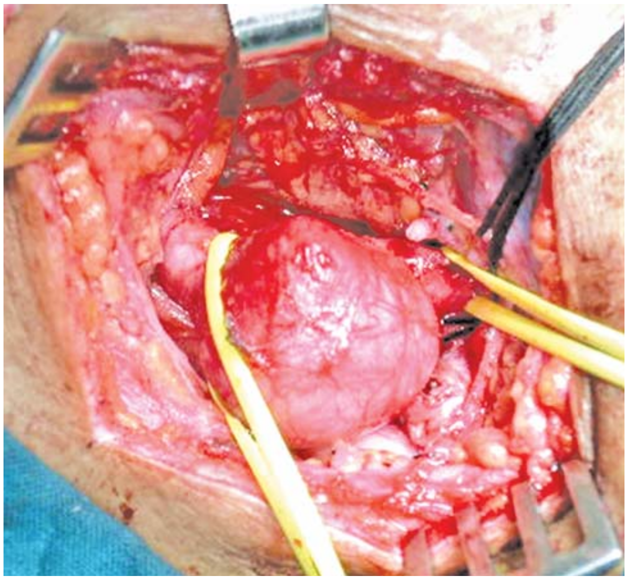

Fig. 5.5: Post-stenotic dilatation of subclavian artery in cervical rib *(Courtesy:* Dr Ashok Shetty, MCh, Cardiothoracic Surgeon, Mangalore*)*.

Clinical Features

Majority of patients are asymptomatic—(80%).

1. ***Vascular manifestations:***
    - *Pain* is due to ischaemia in the muscle. It is more during work, exercise, and is relieved by rest.
    - *Upper limb claudication* usually is observed in forearm and arm more obvious after usage of limb. Pain (dull pain) in posterior triangle of neck may be due to presence rib mass. Ischaemic pain in the digits and hand may be present.
    - *Vasomotor changes* with cyanosis, cold fingers, excessive sweating may be observed.
    - *Roos test* is raising the arm above the shoulder. The side where cervical rib is present, patient cannot continue and so drops the hand down.

Nothing dies quicker than a new idea in a closed mind.

- *EAST—Elevated Arm Stress Test (Modified Roos test)*: Arm is elevated above the shoulder, with elbow stretched fully. Rapid movements of fingers will cause fatigue on the side where cervical rib is present.
- *Adson's test*: The hand is raised above after feeling the radial pulse. The patient is asked to take a deep breath and turn the head to the same side. Any change in pulse, i.e. either becoming feeble or absent, is noted.
- *Modified Adson's test* is same as Adson's, but neck is turned towards the opposite side.
- *Costoclavicular compression manoeuvre*: While palpating the radial pulse of the patient he is asked to move his shoulder backwards and downwards (*exaggerated military position*) which may cause absence/feeble radial pulse and a bruit may be heard while auscultating the supraclavicular region—*military attitude test.* This is due to compression of subclavian artery between clavicle and first rib. Similary *Halstead manoeuvre* is done by 45° abduction and extension of arm with downward pushing of the shoulder with neck turned opposite side to cause radial pulse feebly palpable.
- *Hyperabduction manoeuvre (Wright's test):* While palpating the radial pulse, arm on the diseased side is passively hyperabducted causing feeble or absence of radial pulse. This is due to compression of artery by pectoralis minor tendon (*pectoralis minor syndrome*). An axillary bruit may be heard on auscultation.
- *Allen's test*: It is used in hand to find out the patency of radial and ulnar arteries. Both radial and ulnar arteries of the patient is felt and pressed firmly at the wrist. Patient clinches his hand firmly (often repeated clinching) and holds it tightly. After 1 minute clinch is released to open the palm of the hand which looks pale. Pressure on radial artery in the wrist is released to see area of distribution of the radial artery. Normally it becomes flushed with pink color. If there is block in radial artery, the area will remain white. Test is repeated again. This time pressure on the ulnar artery is released to check the patency of ulnar artery. Area will be pale and blanched after releasing in case of ulnar artery block. Otherwise it becomes pink after release in normal individual.
- ***Wasting*** of thenar, hypothenar and forearm muscles.
- Often digital gangrene, ischaemic ulcers in digits, oedema of fingers and hand are observed.
- Limb is colder and paler than the opposite side.

2. ***Neurological features*** is due to compression of T_1 and C_8 causing tingling and numbness in the little finger, medial side of hand and forearm.
   - Pain, on the medial side, weakness on the medial side of hand and anaesthesia may be evident.
   - *Card test* for interossei muscle weakness, *Froment's sign*/test to detect weakness in adductor pollicis are positive. Wasting may be due to neurological cause also.
3. ***Features in the neck:***
   - Hard, fixed, bony mass in the supraclavicular region.
   - *Palpable thrill* above the clavicle in the subclavian artery.
   - *Bruit* on auscultation.

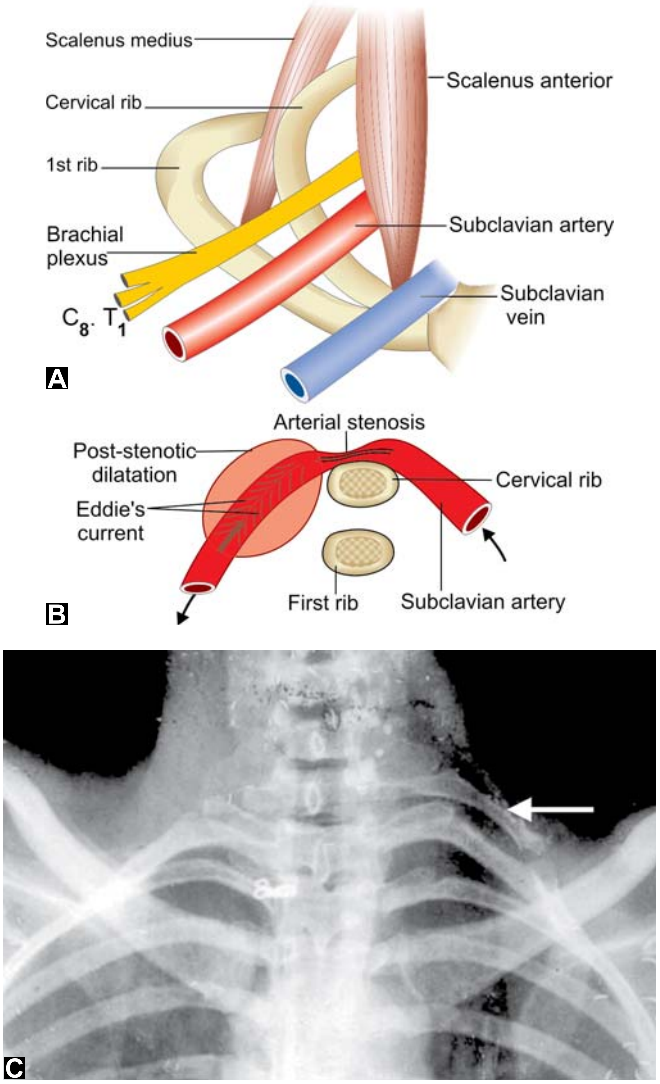

Figs 5.6A to C: Anatomy of cervical rib and its relation to subclavian artery and vein and brachial plexus. Note the Eddie's current, post-stenotic dilatation due to arterial compression by cervical rib. X-ray picture shows left sided complete cervical rib.

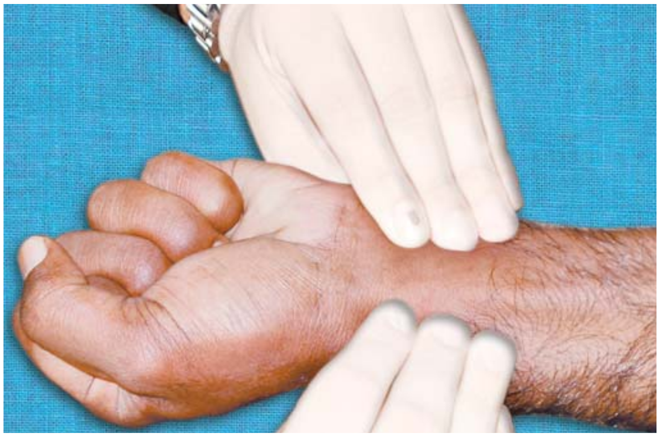

Fig. 5.7: *Allen's test* done upper limb ischaemia to look for perfusion of hand.

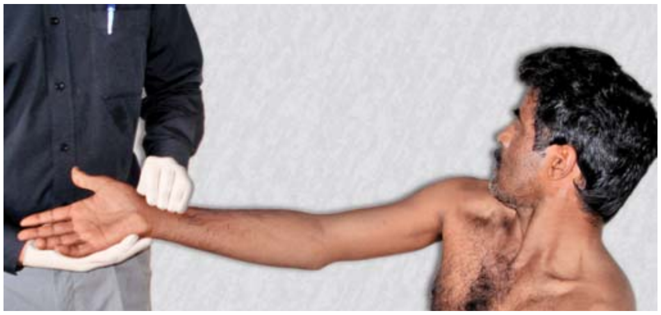

Fig. 5.8: Adson's test—after palpating radial pulse, elbow stretched hand is slightly raised; neck is turned to same side; deep breath is taken; feeble or pulse becoming absent may be obvious in positive Adson's.

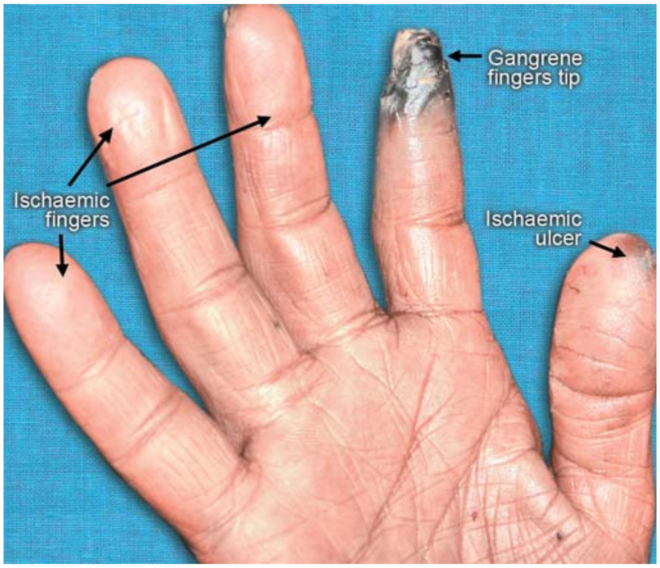

Fig. 5.9: Upper limb ischaemia showing ischaemic ulcers and gangrene of finger tips.

Note:

- Most common presentation is *neurological.*
- Most problematic presentation is vascular which requires surgery.

Differential Diagnosis

1. Cervical spondylosis—to differentiate, X-ray neck—lateral view should be taken.
2. Carpal tunnel syndrome.
3. Tumours or swellings compressing over the vessel or nerves in the neck.
4. Other causes of digital gangrene like atherosclerosis, Raynaud's syndrome, collagen diseases, diabetes mellitus, and embolism.
5. Syringomyelia, motor neuron disease.
6. Pancoast tumour.

Investigations

- Chest X-ray PA view and lateral view including neck—only (radio-opaque) bony rib can be identified.
- Nerve conduction studies to confirm neurological compression and also to rule out carpal tunnel syndrome or cervical spondylosis.
- Arterial Doppler of subclavian artery and of the upper limb.
- Subclavian angiogram.
- Other relevant investigations like blood sugar, lipid profile, cardiac assessment.
- CT scan neck and thorax and CT angiogram of subclavian artery are ideal investigations.

Treatment

Note:
For ***conservative treatment*** refer TOS above.

Surgical

1. In symptomatic cervical rib without arterial compression (subclavian artery), along with scalenotomy (cutting scalenus anterior muscle), extraperiosteal resection of cervical rib and often resection of first rib is done to increase the thoracoaxillary channel and so as to reduce arterial compression.
2. In symptomatic cervical rib with significant subclavian artery compression along with scalenotomy, extraperiosteal resection of cervical rib, resection of first rib, subclavian artery reconstruction with or without a graft is done.
3. Along with scalenotomy, extraperiosteal resection of cervical rib, resection of first rib, reconstruction of subclavian artery, cervical sympathectomy is also done to improve the circulation to the ischaemic upper limb.
4. Amputation of the gangrenous toe.
5. In pectoralis minor syndrome, pectoralis minor tendon is released from its insertion to coracoid process.

Approaches

- Supraclavicular approach—mainly when there is need for vascular reconstruction
- Transaxillary approach—through axillary crease, rib is approached and removed
- Thoracotomy approach

BRANCHIAL CYST

It arises from the remnants of second branchial cleft. Normally 2nd, 3rd, 4th clefts disappear to form a smooth neck. Persistent 2nd cleft is called as *cervical sinus (of His)* which eventually gets sequestered to form *branchial cyst.*

Epithelial infusion within lymph node may be the other cause as branchial cyst contains lymphoid tissues in their wall.

Six branchial arches with five pharyngeal pouches (endoderm lining) inside and five pharyngeal clefts (ectoderm lining) outside are present.

Features

- Swelling in the neck beneath the anterior border of upper third of the sternomastoid muscle. It is smooth, soft, fluc-

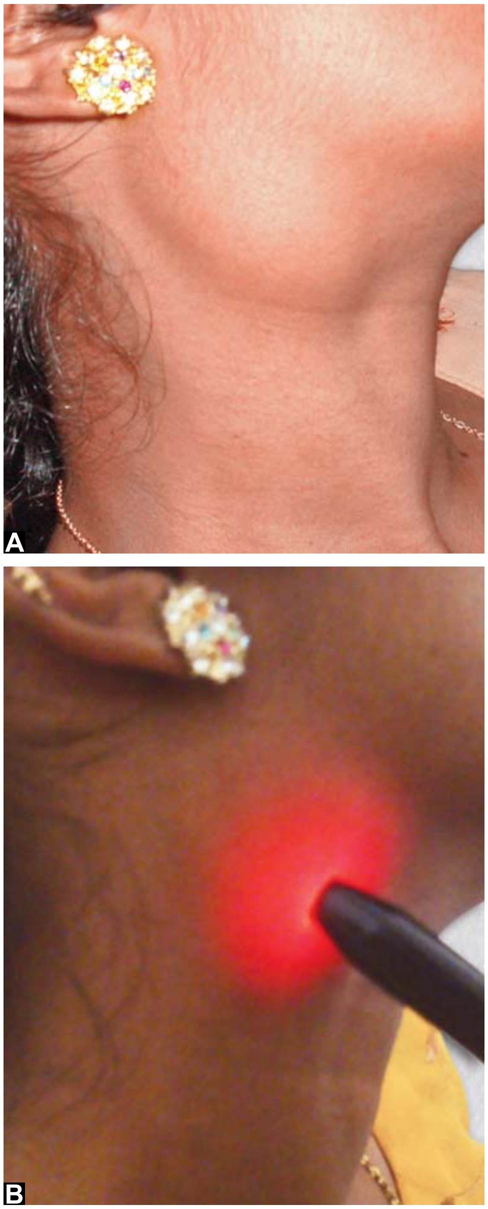

Figs 5.10A and B: Branchial cyst. It is only occasionally transilluminant.

tuant, often transilluminant with a sensation of '*Half-filled double hot water bottle*'.

- It is equal in both sexes. Even though congenital, it is seen in late adolescents and early 3rd decade.
- Usually painless unless it is infected.
- It contains *cholesterol crystals* which is from the lining of mucous membrane which contains sebaceous gland. *Cheesy toothpaste like* material is typical
- Histologically, it is lined by *squamous epithelium*. Occasionally it contains ciliated columnar epithelium. Cyst wall shows plenty of *lymphoid* tissue.
- It may get infected to form an abscess.
- FNAC shows *cholesterol crystals*.

Complications

- Recurrent infection.
- Rupture may cause *acquired branchial fistula* at upper third of sternocleidomastoid muscle.

Differential Diagnosis

1. Cold abscess, lipoma neck.
2. Lymph cyst.
3. Chronic lymphadenitis.

Treatment: *Excision* under G/A.
Cyst is in relation to carotids, hypoglossal nerve, glossopharyngeal nerve, spinal accessory nerve, posterior belly of digastric and pharyngeal wall. Medially it is close to the posterior pillar of tonsils. During dissection, all these structures should be taken care of.

Complications of Surgery

- Injury to major structures.
- Infection.
- Recurrence/fistula formation due to incomplete removal of the track.

Cholesterol crystals are seen in:

- Branchial cyst
- Dentigerous cyst
- Hydrocele

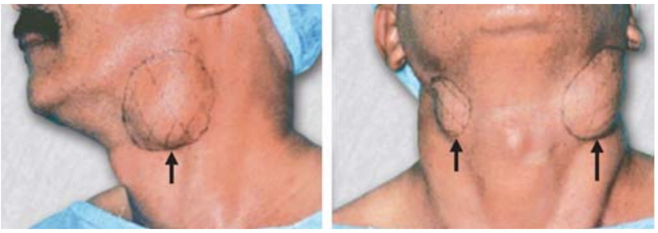

Fig. 5.11: Branchial cyst—both sides.

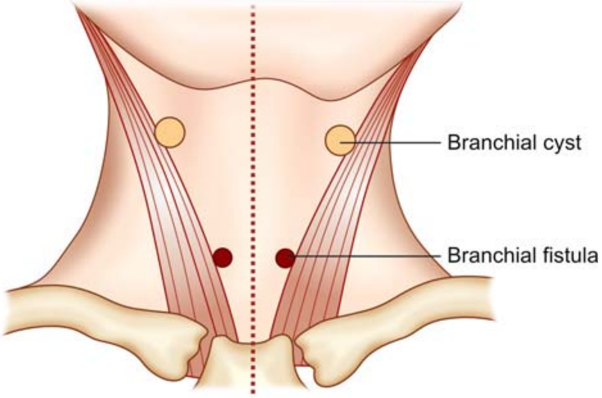

Fig. 5.12: Positions of branchial cyst and fistula.

BRANCHIAL FISTULA

Branchial fistula is ***commonly a congenital lesion***. It is due to *persistent precervical sinus* between 2nd branchial cleft and 5th branchial cleft having opening in the skin at lower 1/3rd of neck on the inner margin of sternocleidomastoid muscle, often ends as a *sinus* just proximal to the posterior pillar of fauces behind tonsil which is also the site of inner opening when presents as *fistula*. Fistula runs between the structures related to 2nd and 3rd branchial arches (2nd arch artery is ECA, nerve is facial; 3rd arch artery is ICA, nerve glossopharyngeal). From external opening at skin below, it runs in subcutaneous plane to pierce deep fascia at level of thyroid cartilage; to travel between ECA and ICA; behind posterior digastric belly and stylohyoid; outer to IJV, stylopharyngeus, hypoglossal and glossopharyngeal nerves; perforates superior constrictor to reach the internal opening.

Occasionally ***acquired branchial fistula*** can occur due to rupture of or after drainage of infected branchial cyst or incomplete excision of the cyst track. This type of fistula is located outside at skin at the level of upper third of sternomastoid muscle.

Features

- It is a persistent second branchial cleft with a communication outside to the exterior. It is commonly a congenital fistula. Occasionally, the condition is secondary to incised, infected branchial cyst.
- Often it is bilateral (30%).
- *External orifice* of the fistula is situated in *the lower third of the neck* near the anterior border of the sternomastoid muscle.
- *Internal orifice* is located on the anterior aspect of the posterior pillar of the fauces, just behind the tonsils.
- Sometimes fistula ends internally as blind end.

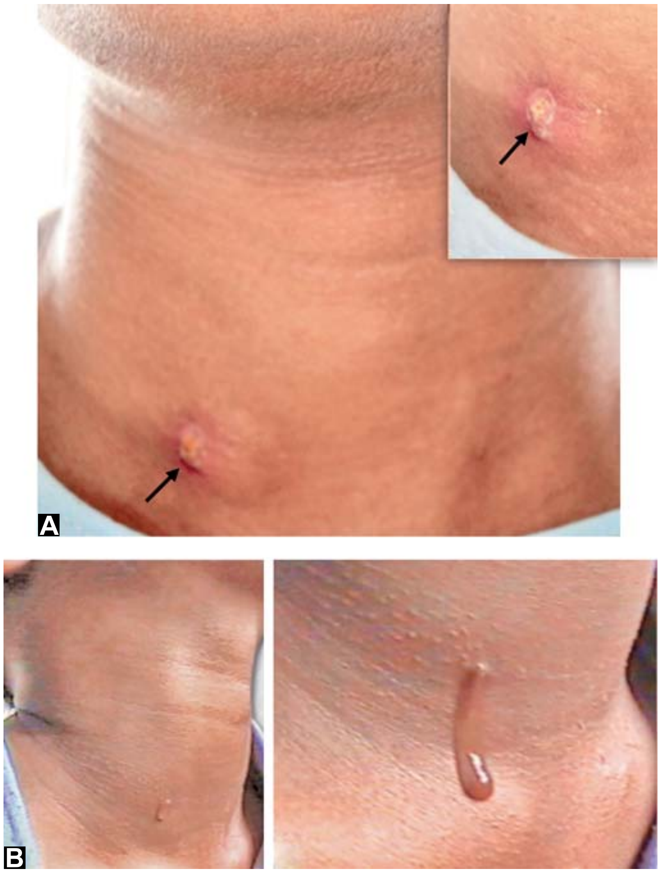

Figs 5.13A and B: Branchial fistula in two different patients. Note the location and discharge (*Courtesy:* Professor Kishore Chandra Prasad, ENT Surgeon and Head of the Department and Dr Sampath ENT Surgeon, KMC, Mangalore)

- Track is lined by *ciliated columnar epithelium* with patches of lymphoid tissues beneath it, causing recurrent inflammation.
- It usually presents at birth. It is common in *children* and early adolescent period. Equal in both sexes.
- *External orifice* is very small with a *dimple* which becomes more prominent on dysphagia with *tuck in appearance*.
- *Discharge* is mucoid or mucopurulent.

Investigations

Discharge study, fistulogram, MR/CT fistulogram.

Treatment

Always surgery:

- Under general anaesthesia, methylene blue is injected into the track. Probe is passed into the fistulous track. Through circumferential/elliptical incision around the fistula opening,

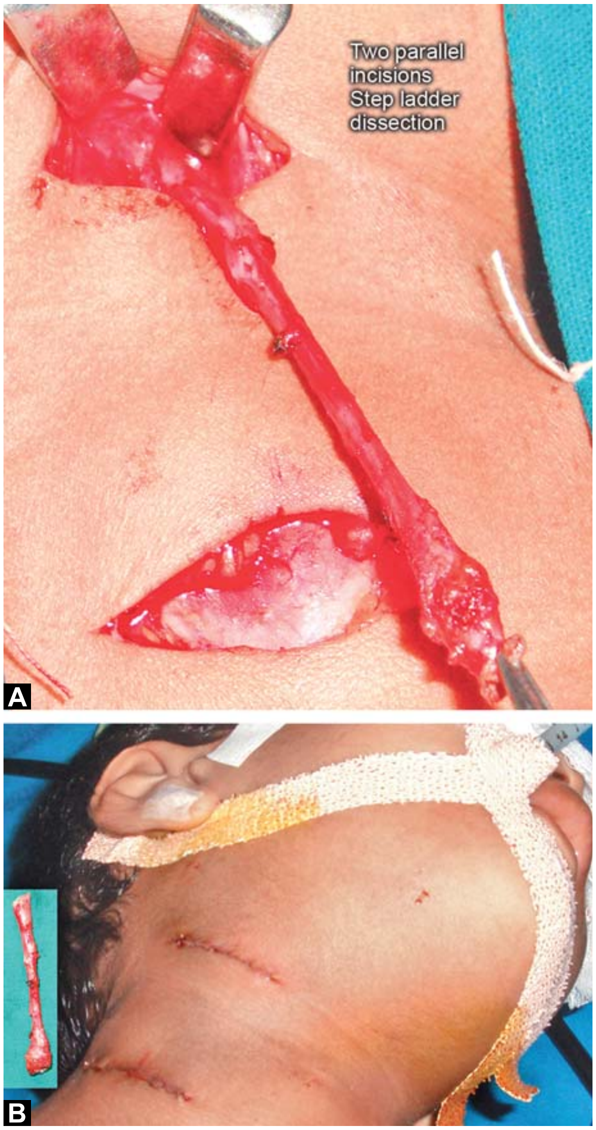

Figs 5.14A and B: Branchial fistula is usually operated using two transverse parallel incisions with step ladder dissection (*Courtesy:* Br Ganesh Pai, MCh)

There is only one degree difference between hotwater and steam.

entire length of the track is dissected until the internal orifice. Care should be taken to safeguard carotids, jugular vein, hypoglossal nerve, glossopharyngeal nerve and spinal accessory nerve. *Entire track should be excised.*

- *Step ladder dissection* is done using two parallel incisions one below at lower part another above at upper part of the neck, will make dissection easier and complete.

PHARYNGEAL POUCH

- It is a protrusion of mucosa through *Killian's dehiscence*, a weak area of the posterior pharyngeal wall between *thyropharyngeus (oblique fibres)* and cricopharyngeus (*transverse fibres*) of the *inferior constrictor muscle* of the pharynx.
- Thyropharyngeus is supplied by pharyngeal plexus from cranial accessory nerve. Cricopharyngeus is supplied by external laryngeal nerve.
- Pharyngeal pouch is a pulsion diverticulum. It starts in the midline of posterior pharyngeal wall. Once it expands and reaches the vertebra, it deviates towards left side of the neck because of resistance of vertebra.
- Imperfect relaxation of the cricopharyngeus increases the pressure in the pharynx, mainly during swallowing which leads to protrusion of mucosa through the Killian's dehiscence causing pharyngeal pouch.
- *The protrusion is usually towards left.*

Stages

1. Small diverticulum pointing towards vertebra. It is asymptomatic and incidentally diagnosed by barium meal X-ray. Foreign body sensation in pharynx may be present.
2. Large, globular diverticulum with vertical mouth/opening causing regurgitation, violent cough, dysphagia, respiratory infection. Regurgitation is more after meals and while turning the neck.
    - Large pouch which is visible in the neck as a globular swelling often tender, smooth and soft. Swelling is below the level of the thyroid cartilage and behind sternocleidomastoid muscle and can be emptied on pressure. Opening of the pouch is not vertical but horizontal. They present with dysphagia, features of respiratory infection like pneumonia and lung abscess, weight loss and cachexia. Pouch may itself get infected and may form an abscess. Often the pouch may descend downward and enter the superior mediastinum. Gurgling sound in the neck is observed.

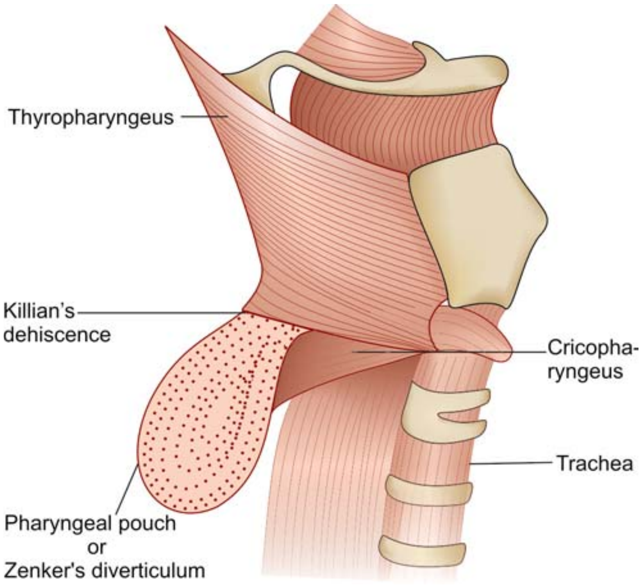

Fig. 5.15: Pharyngeal pouch (*Zenker's diverticulum*).

Problems in pharyngeal pouch

- ❖ Progressive dysphagia
- ❖ Respiratory problems like pneumonia, lung abscess
- ❖ Abscess in the neck due to infection in the pouch
- ❖ Weight loss and cachexia

Differential Diagnosis

- Branchial cyst.
- Cold abscess in the neck.
- Lymph cyst.
- Haemangioma neck.

Clinical Features

- Pain, dysphagia, recurrent respiratory infection, swelling in the neck on the left side which is smooth, soft and tender.
- Regurgitation during night while turning neck, smooth, soft, tender swelling in the posterior triangle of the left side of the neck; typical gurgling noise while swallowing—are typical features. It is common in males.
- Swelling is deep to sternocleidomastoid muscle below the level of thyroid cartilage; initially soft and emptying; impulse on coughing may be evident unless opening of the pouch is blocked due to recurrent inflammation.

Investigations

- *Barium swallow*—lateral view shows pharyngeal pouch.
- Chest X-ray shows pneumonia.
- CT neck is very useful.

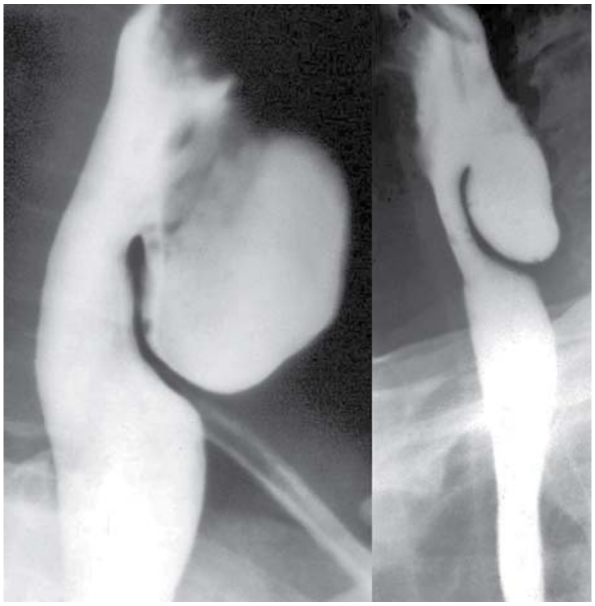

Fig. 5.16: Barium study showing pharyngeal pouch.

***Note*:**
Oesophagoscopy should be gentle or avoided as scope may enter the friable pharyngeal pouch and can cause perforation and life threatening mediastinitis.

Treatment

- Antibiotic is started to control infection.
- *Pharyngeal pouch is excised* by an oblique neck incision (approach from neck). As there is cricopharyngeal spasm, *cricopharyngeal myotomy* (i.e. cutting of cricopharyngeal circular muscle fibres without opening mucosa) is done to prevent the recurrence.

Indications for surgery

- Progressive symptoms.
- Recurrent respiratory complications.
- Dysphagia.

Technique

- Under anaesthesia endoscope is passed to identify the opening of the pouch.
- Pouch is packed with acriflavine soaked gauze for identification of the pouch.
- Under anaesthesia nasogastric tube is passed into the oesophagus under visualisation.
- Pouch is approached through neck with oblique incision along the anterior margin of sternocleidomastoid muscle for proper removal.
- Myotomy of cricopharyngeus and upper circular muscle of the oesophagus is done to relieve the spasm, at posterior midline.
- *Dohlman's operation* is endoscopic excision of the pharyngeal pouch.

Complications of surgery

- Infection, either mediastinitis or lung infection (Pneumonia or lung abscess)
- Pharyngeal fistula
- Abscess in the neck
- Oesophageal stenosis and recurrence

LARYNGOCELE

- It is a unilateral narrow necked, air containing diverticulum resulting from herniation of laryngeal mucosa.
- It occurs in professional *trumpet players, glass blowers and in people with chronic cough.*

Types

a. **External:** It is situated in the anterior third of the laryngeal ventricle, between the false cords and thyroid cartilage, herniates through the *thyrohyoid membrane* where it is pierced by superior laryngeal nerve.

b. **Internal:** Confined within the larynx, presents as a distention of false cords.

c. **Combined**.

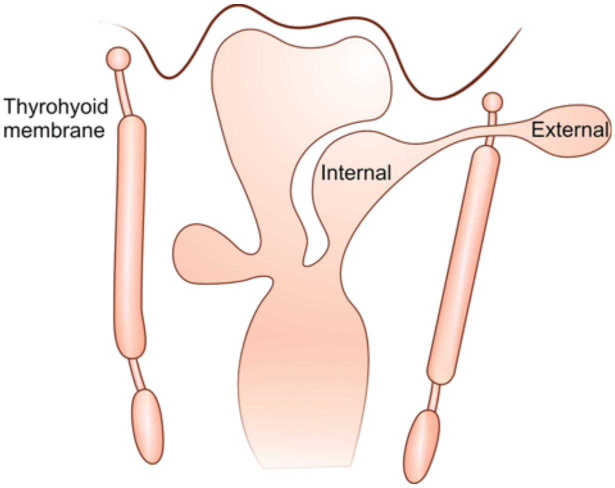

Fig. 5.17: Laryngocele.

Clinical Features

- Swelling in the neck in relation to larynx, adjacent to thyrohyoid membrane which is smooth, soft, *resonant* and is more prominent while blowing, coughing and Valsalva manoeuvre.
- It moves *upwords* during swallowing with *expansile impulse* on coughing.
- Infection is quite common in the sac of laryngocele, leading to the blockade of opening of the sac causing an abscess.
- Pus often may be discharged into the pharynx repeatedly.
- Hoarseness and cough.
- If large, causes obstruction to larynx.

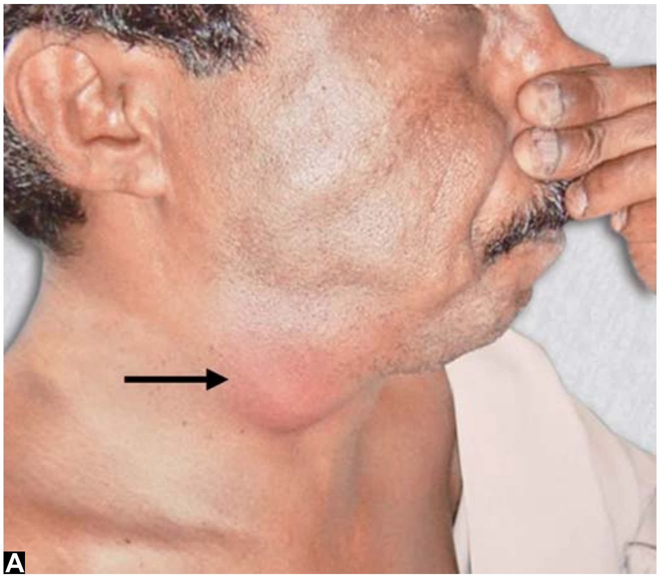

Fig. 5.18A

One-third of congenital torticollis is due to sternomastoid tumour and two-third due to abnormal position in utero which recovers spontaneously in a few weeks. **— Kenneth F Hulbert**

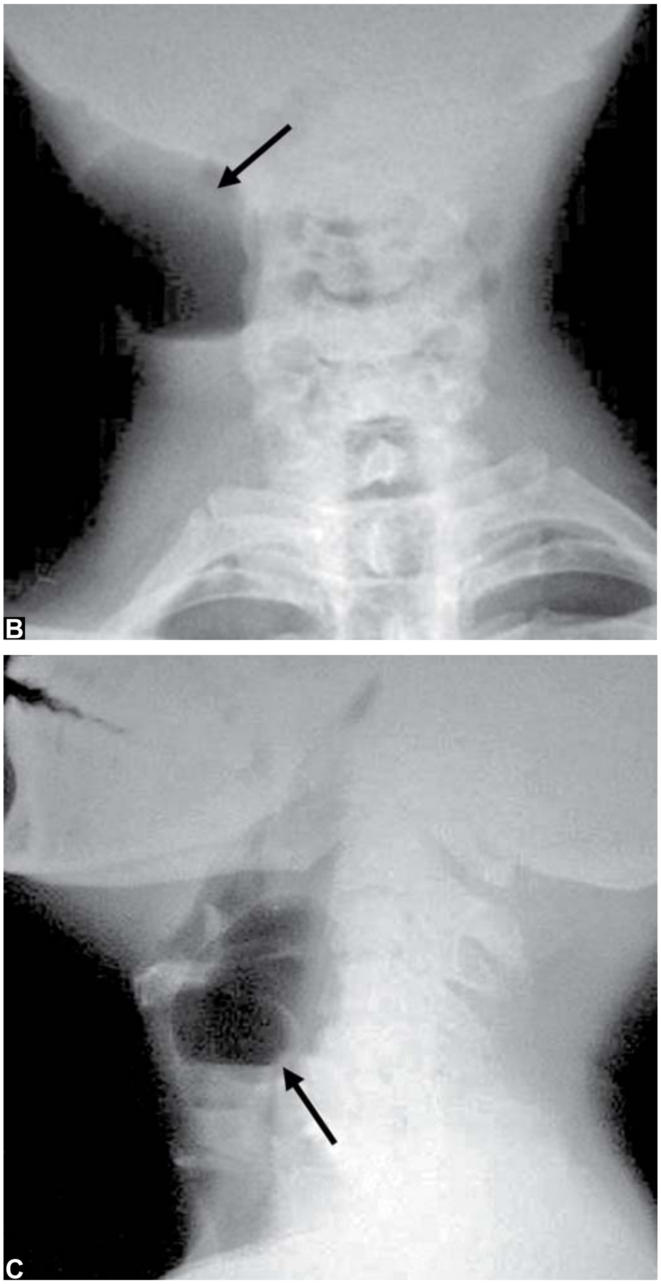

Figs 5.18B and C

Figs 5.18A to C: Typical laryngocele in the neck which becomes prominent after blowing. X-rays (AP and lateral) show radiolucent air in the neck.

Diagnosis

Clinical features, X-ray neck, laryngoscopy, CT scan.

Treatment

External laryngocele: Excision through neck incision. Neck of the sac should be ligated. Thyrohyoid membrane is repaired using 3 zero nonabsorbable polypropylenes sutures.

Internal laryngocele: Marsupialisation, with the help of laryngoscope.

CYSTIC HYGROMA (Cavernous Lymphangioma)

- It is a cystic swelling due to sequestration of a portion of jugular lymph sac from the lymphatic system, during the developmental period *in utero*.
- Present at birth and so may cause obstructed labour. Occasionally present in early infancy.
- It is also called as ***hydrocele of the neck***.
- Cyst will not communicate with normal lymphatics and so existing lymph gets absorbed and cyst will be filled with clear watery mucous derived from endothelial lining of the cyst wall. Cyst even though is subcutaneous; it commonly extends into deeper planes across many anatomical planes and barriers. Cyst is *multilocular*. Often extension may occur across two or more lymphatic regions; example—involvement of both neck and axilla.
- *Lymphangioma circumscripta* (< 5 cm), *lymphangioma diffusum* (> 5 cm) and *lymphangioma ab agne* (reticulate pattern) are different variants. For detail refer chapter 'Swelling'.

Sites

1. *Posterior triangle of the neck*—75%—most common site. Eventually may extend upwards in the neck
2. Axilla—20%
3. Cheek
4. Tongue—lymphangiogenetic macroglossia
5. Groin
6. Mediastinum
7. Often multiple sites

Pathology

- It contains aggregation of cysts looking like *soap bubbles*. Cysts have *mosaic appearance* with larger cysts near the surface and smaller cysts in the deeper planes. Each cyst

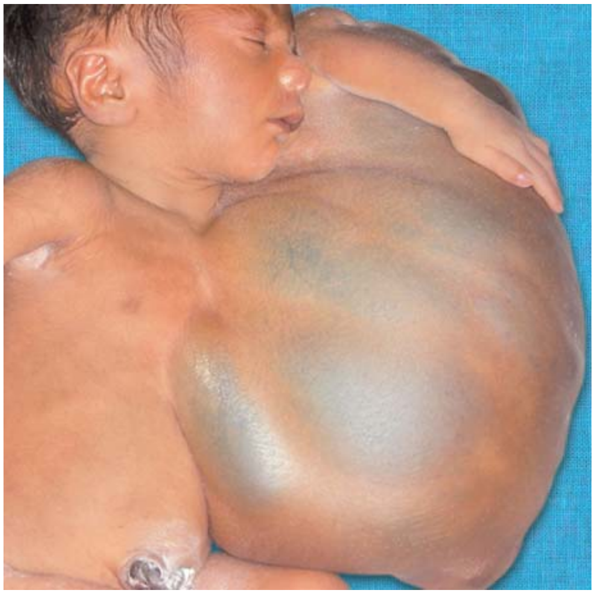

Fig. 5.19: Cystic hygroma in newborn baby.

contains clear lymph with endothelial lining. Fluid *does not* coagulate.

Clinical Features

- Swelling is *present at birth* in the *posterior triangle of neck* causing obstructed labour.
- Swelling is smooth, soft, fluctuant (cystic), *partially compressible*, *brilliantly transilluminant*. It is *not reducible* completely.
- During crying swelling often increases in size.
- *Disfigurement* of face of the child which is more worrying factor for the parents.
- Swelling may rapidly increase in size causing *respiratory obstruction*—dangerous sign.
- It may get infected forming an *abscess* which is a tender, warm, soft swelling. It may cause septicaemia which may be life threatening.
- Rupture with lymph ooze can occur.

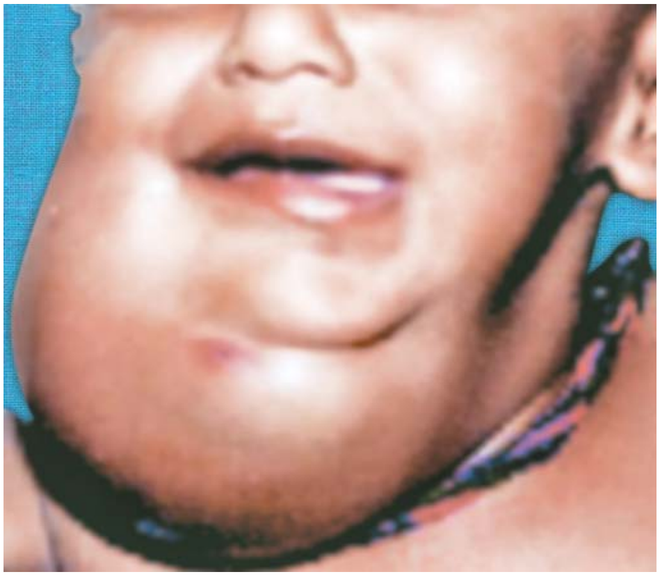

Fig. 5.20: Cystic hygroma.

Treatment

- Aspiration of the contents. Later once the sac or capsule gets thickened by fibrous tissue, it is excised.
- Care should be taken to have meticulous dissection across all planes including deeper muscular one to *clear entire cyst wall*; otherwise recurrence will occur.
- When it causes respiratory obstruction, *aspiration* and *tracheostomy* is done.
- Under proper antibiotics coverage, drainage of abscess is done. Later sac is excised.
- Preoperative injection of sclerosants (OK 432, PICIBANIL recently used) and later once fibrosis develops excision of entire aggregation of cysts. In olden days injecting boiling water into the cyst as a sclerosant used to be popular.

Complications

- Respiratory distress
- Infection Abscess Septicaemia
- Surgery itself may cause torrential haemorrhage

LUDWIG'S ANGINA

- It is an inflammatory oedema of submandibular region and floor of the mouth, commonly due to streptococcal infection.
- It causes diffuse swelling and brawny oedema of the submandibular region. It is common in severely ill and in advanced malignancy, causing *trismus, laryngeal oedema*. Extension of infection into *parapharyngeal space* may lead to dreaded *internal jugular vein thrombosis*.
- As the infection is deep to the deep fascia in a closed fascial plane, it spreads very fast causing dangerous complications.

Precipitating factors

- Caries teeth
- Oral or other malignancy
- Submandibular salivary infection/calculi
- Chemotherapy
- Chronic diseases like diabetes mellitus
- Cachexia of any cause

Special features

- Brawny oedema/diffuse swelling of submandibular and submental region
- Intraoral oedema in floor of the mouth
- Putrid halitosis
- Severe toxicity
- Dyspnoea, dysphagia
- May cause laryngeal oedema, septicaemia, extension of sepsis into other spaces in the neck

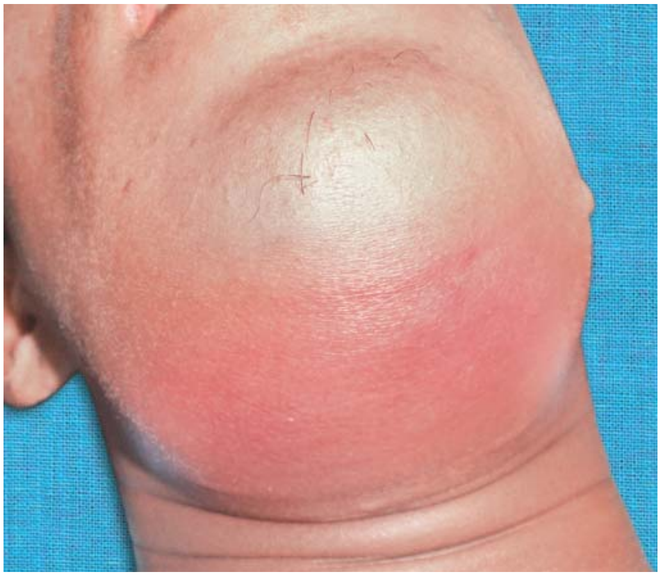

Fig. 5.21: Typical Ludwing's angina.

It is better to aim at good things and miss it , than to aim at a bad thing and hit it.

Clinical Features

- Fever, toxicity, diffuse swelling, dysphagia, dyspnoea, trismus.
- Intraoral oedema is common.
- Brawny swelling in submandibular region.
- Putrid halitosis.

Treatment

- Antibiotics.
- IV fluids.
- *Decompression of the submandibular region is done,* by making a deep incision extending into the deep fascia and also *cutting both the mylohyoid muscles*. Either it is left open and *delayed suturing is done* (better option) or it is loosely sutured.

Complications

- Laryngeal oedema and respiratory distress, may require tracheostomy
- Septicaemia
- Extension of infection into parapharyngeal space

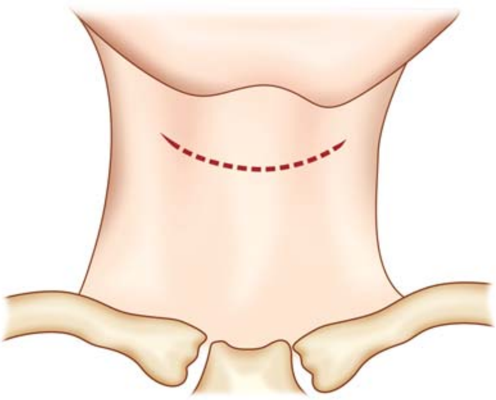

Fig. 5.22: Incision for draining Ludwig's angina.

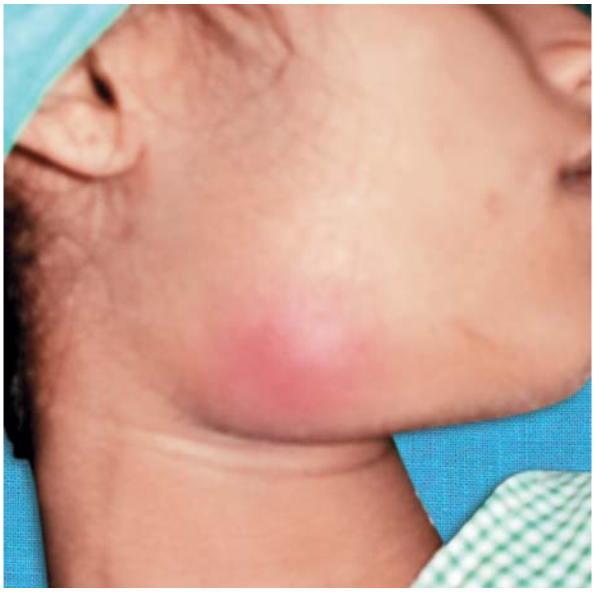

Fig. 5.23: Abscess in submandibular region. Note the redness and swelling. It needs incision and drainage under general anaesthesia.

PARAPHARYNGEAL ABSCESS

- It is *infection of pharyngomaxillary space*.
- This is a cone-shaped space; base is formed by the base of skull; apex is formed by the greater cornu of hyoid bone; medial wall by the superior constrictor; lateral wall is formed by the lateral pterygoid, angle of mandible and below by submandibular salivary gland.
- Usually, infection arises from the tonsils, after tonsillectomy and from the submandibular space.

Clinical Features

- It causes diffuse swelling in the upper neck, trismus, fever, toxicity.
- Swelling lies behind the posterior pillar with oedema of soft palate.

Complications

- Thrombosis of internal jugular vein.
- Erosion into the internal carotid artery causing torrential bleeding.
- Septicaemia.

Treatment

- Under G/A, drainage is done by making incision (deep incision) between angle of the mandible and hyoid bone. Early drainage is needed.
- Antibiotics are given.
- Causative pathology should be addressed.

RETROPHARYNGEAL ABSCESS

Surgical Anatomy

The wall of the pharynx has got 5 layers. Mucosa, submucosa, pharyngobasilar fascia, muscular layer (contains 3 constrictors and stylo, salpingo, palatopharyngeus muscles) and buccopharyngeal fascia covers outer part of constrictors and extends over buccinator. Buccopharyngeal fascia is adherent to prevertebral fascia posteriorly in the midline. Retropharyngeal lymph nodes are located between buccopharyngeal fascia and prevertebral fascia in paramedian (eccentric) position (not midline).

Types

- Acute.
- Chronic.

Acute Retropharyngeal Abscess

- It is infection and suppuration of retropharyngeal lymph nodes due to staphylococci or streptococci organisms.
- Commonly from tonsils or pharynx.
- Common in infants and children.

Clinical Features

- It presents as *lateral (paramedian, eccentric) smooth, tender swelling* in the pharynx with dysphagia, dyspnoea, cough, toxic features and neck rigidity.
- Diagnosis is obvious on proper clinical examination.

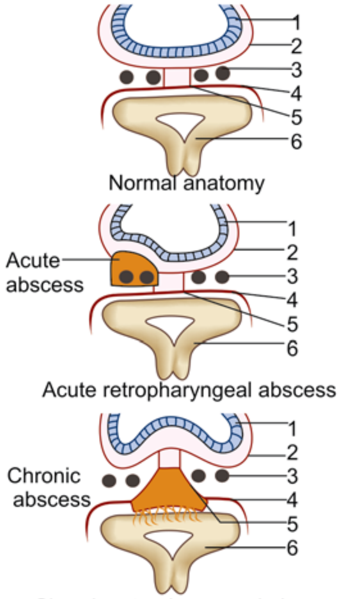
1
2
3
4
5
6
Normal anatomy
Acute abscess
Acute retropharyngeal abscess
Chronic abscess
Chronic retropharyngeal abscess

1. Pharyngeal wall; 2. Buccopharyngeal fascia; 3. Retropharyngeal LN; 4. Prevertebral fascia; 5. Median septum; 6. Vertebra

Fig. 5.24: Note acute and chronic retropharyngeal abscess. Normal anatomy is also shown. Acute is eccentric and is due to suppuration of retropharyngeal lymph nodes. Chronic is central, midline and is due to tuberculosis of the cervical vertebra.

Treatment

- Antibiotics intravenously.
- Drainage is done usually *through per oral incision* under careful general anaesthesia. Only occasionally drainage may be done through a neck incision. Pus should be sent for culture.

Chronic Retropharyngeal Abscess

- It is invariably due to tuberculosis of cervical spine.
- Abscess is in the *midline behind the prevertebral fascia.*
- There is destruction of the body of the vertebra due to tuberculosis.

Clinical Features

- It is *midline swelling in the posterior pharyngeal wall,* which is smooth and nontender.
- Features of tuberculosis of cervical spine will be observed.
- Often abscess may point in the neck in relation to sternomastoid.
- Neurological manifestations may occur in severe disease.

Investigations

X-ray spine, chest X-ray, ESR, MRI of cervical spine are essential investigations.

Treatment

- Antitubercular drugs.
- Drainage of the abscess should be done *through neck approach (never intraoral approach).*
- Decompression of the vertebra and stabilisation is also often required.

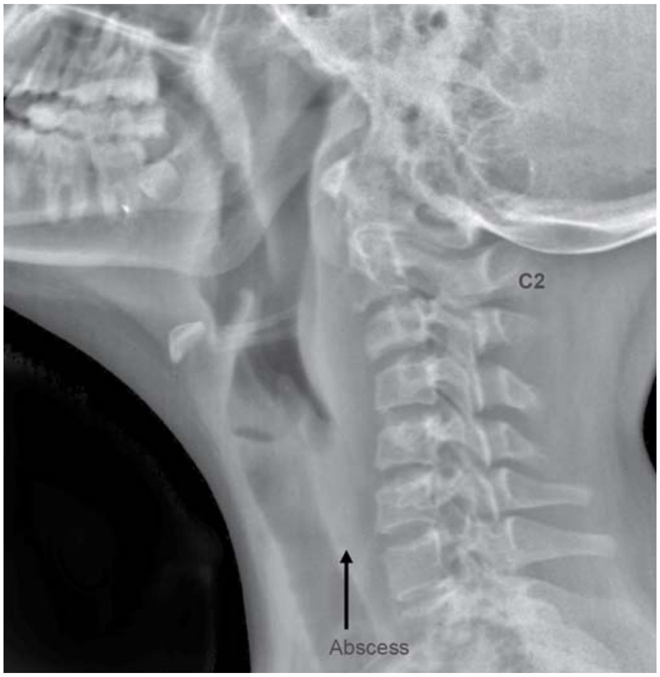

Fig. 5.25: X-ray neck showing retropharyngeal abscess with tuberculosis lesion involving the C_2 cervical spine (*Courtesy:* Dr Navin Chandra Shetty, MD, Radiologist, Professor and Head, Department of Radiodiagnosis, KMC, Mangalore).

SUBHYOID BURSITIS

- Subhyoid bursa is space between posterior surface of the body of hyoid bone and thyrohyoid membrane. It lessens friction between these two structures during swallowing.
- Due to constant friction inflammatory fluid collects in the bursa leading to bursitis, which presents like a *horizontally placed* midline swelling between lower part of the hyoid bone and thyrohyoid membrane.

Features

- Smooth, soft, cystic, fluctuant, nontransilluminating swelling which moves upwards with deglutition but *not while protruding the tongue out*.
- It should be differentiated from thyroglossal cyst and pretracheal lymph nodes.
- It contains turbid fluid often may get infected to make the swelling tender or to form an abscess.

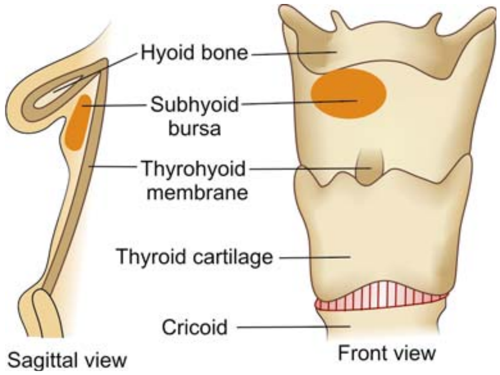

Fig. 5.26: Sagittal and front view showing location and relation of subhyoid bursa.

Treatment

- Excision under general anaesthesia.

CAROTID BODY TUMOUR (Potato Tumour, Chemodectoma, Nonchromaffin Paraganglioma)

- It arises from the carotid body, which is located at the bifurcation of the common carotid artery.
- The tumour is situated in the adventitia of the artery.
- They are benign or locally malignant tumours (10%), but in 20% cases spread can occur to the regional lymph nodes and lungs.
- Blood supply to the tumour is from *external carotid artery*. Tumour does not secrete epinephrine or any endocrine substances. Blood supply comes through *Meyer's ligament* on the posteromedial wall of the carotid at bifurcation.
- They can be *familial*.
- Common in high altitude area, common in females.
- Pathologically, it is well-encapsulated, hard creamy yellowish tumour with dense fibrous tissue. Carotid body tumour cells are not hormonally active.

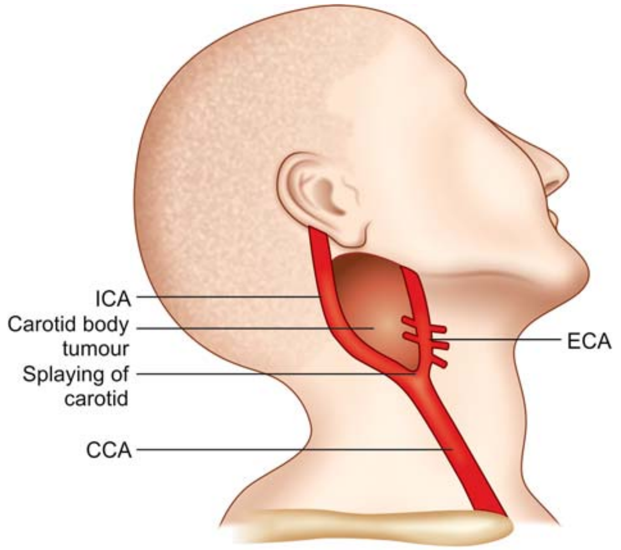

Fig. 5.27: Carotid body tumour—note the site and location. *Splaying* of the carotid is common.

Carotid body

- ❖ Normal carotid body is 3-5 mm sized flat brownish nodule in the adventitious of common carotid artery
- ❖ It consists of chief cells (contains catecholamine granules) and supportive cells
- ❖ It gets its nerve supply from glossopharyngeal nerve
- ❖ These chemoreceptors are sensitive to changes in pH and temperature in the body especially in hypoxia, help in autoregulation of respiration and circulation
- ❖ Carotid body hyperplasia can occur in people residing in high altitudes who are exposed to chronic hypoxia
- ❖ Other chemoreceptors in the body are—aortic bodies in the arch of aorta; glomus jugular in the bulb of the internal jugular vein; glomus intravagale in relation to ganglion nodosum of the vagus nerve and others like pulmonary (near pulmonary artery) and myocardial (near coronary artery origin) receptors

Clinical Features (0.5%—Incidence)

- Usually unilateral.
- More common in middle age.
- Swelling (75%) in the carotid region of the neck which is smooth, firm, pulsatile (*transmitted pulsation*—due to pulsatile carotid vessel overlying its surface) and moves only side to side but not in vertical direction.
- It can often compress over oesophagus and larynx.
- Headache, neck pain (35%), dysphagia, syncope are other presentations.
- It can present with unilateral vocal cord palsy; can cause Horner's syndrome.
- Features of transient ischaemic attacks due to compression over the carotids, *"carotid body syncope."*
- Thrill may be felt and bruit may be heard.
- It is located at the level of hyoid bone deep to anterior edge of the sternomastoid muscle in anterior triangle, vertically placed, round, firm 'potato' like swelling.
- Often tumour may extend into the cranial cavity along the internal carotid artery as *dumbbell tumour*.

Sites where dumbbell tumours are seen

- ❖ Parotid tumour
- ❖ Spinal cord tumour
- ❖ Carotid body tumour

Shamblin classification of carotid body tumor

- ❖ Type I: Localised, easily resectable (26%)
- ❖ Type II: Adherent, partially surrounding the carotids (46%)
- ❖ Type III: Adherent, encased carotids completely (27%)

Investigations

- Doppler.
- Angiogram to see the *'tumour blush'*—DSA. Widening/splaying of the carotid artery with tumour blush in an angiogram is called as *Lyre sign*.
- CT scan, MRI, MR angiography.

No FNAC: No partial excision.

Differential Diagnosis

- Carotid artery aneurysm.
- Soft tissue tumour (Sarcoma).
- Lymph node enlargement.
- *Neurofibroma of the vagus nerve* presents as swelling in the carotid triangle in the region of thyroid as vertically placed, oval, hard swelling. On palpation of the swelling, patient often develops bradycardia and dry cough. It does not move with deglutition and has only transverse mobility. As the tumour lies behind the carotid it can stretch the carotid in front causing transmitted pulsation.

Treatment

- If it is small, it can be excised easily as the tumour is situated in the adventitia.
- When it is large, as commonly observed, *complete excision has to be done followed by placing a vascular graft.*

- During resection a *temporary shunt* (diversion of blood) is placed between common carotid below and internal carotid above to safeguard cerebral perfusion; external carotid artery is ligated. Venous or prosthetic graft is placed between common carotid and internal carotid arteries.
- Carotid body tumour is not *radiosensitive* (controversial).

Complications of the Surgery

- Bleeding.
- Blockage of common carotid artery, leading to contralateral side hemiplegia (3%). This can be prevented by *stenting the common carotid artery* towards internal carotid artery and is done *during surgical excision* of the tumour.
- Cranial nerve injury X and XI (40%).

Cystic swellings in the neck

- Cold abscess
- Cystic hygroma
- Branchial cyst
- Thyroglossal cyst
- Laryngocele
- Pharyngeal pouch
- Subhyoid bursa

TORTICOLLIS (Wry Neck)

It is turning of the neck to one side with chin pointing towards opposite side.

Causes

- Sternomastoid tumour.
- Trauma—spinal injury, disc prolapse, spondylosis.
- Inflammatory: Lymphadenitis either tuberculous or suppurative; tuberculosis of cervical spine.
- Spasmodic—due to spasm of sternomastoid muscle of same side or spasm of posterior cervical muscles of opposite side.
- Reflex.
- Rheumatic—after exposure to cold/draught.
- Burns—causing contracture.

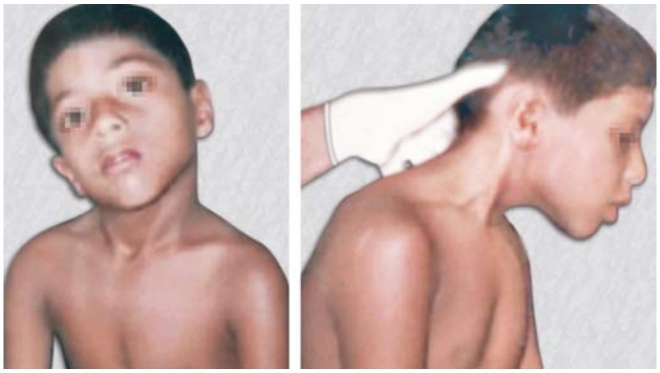

Fig. 5.28: Boy having right-sided torticollis.

- Ocular causes.
- Compensatory due to scoliosis.

Clinical Features

- Restricted neck movements.
- Chin pointing towards opposite side.
- Squint.
- Features relevant of the causes.

Treatment

The cause is treated.

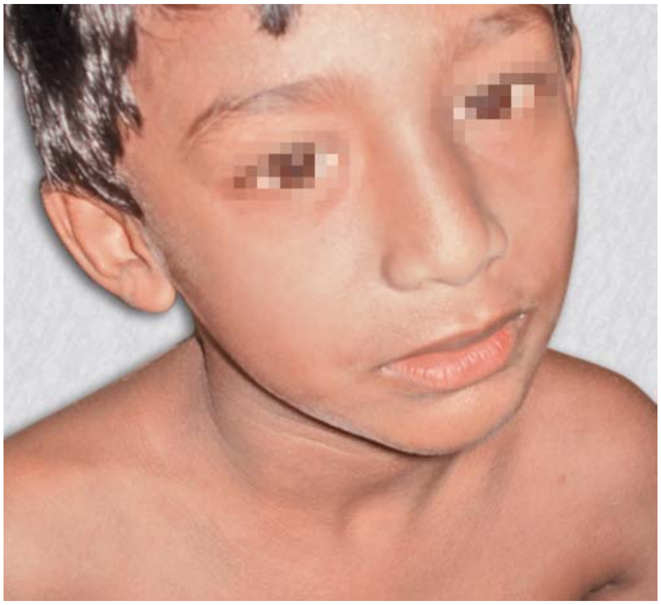

Fig. 5.29: Torticollis (right-sided) with chin towards opposite side.

STERNOMASTOID TUMOUR

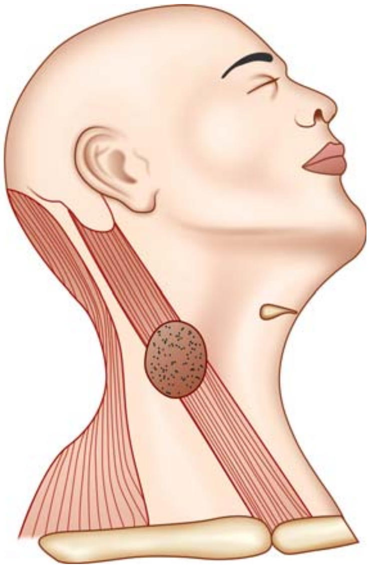

Fig. 5.30: Sternomastoid tumour—typical location.

Direction is a matter of fact; ideas are matter of opinion.

It is due to birth injury to the sternomastoid muscle. It is a *misnomer*. It is not a tumour.

Pathogenesis

- During child birth, injury to the sternomastoid muscle causes *haematoma* in the muscle which gets organised to form sternomastoid tumour.
- Common in breech delivery.

Clinical Features

- It is seen in infants of 3-4 weeks age.
- Swelling of about 2 cm size, in the sternomastoid muscle which is smooth, hard, nontender and adherent to the muscle—in the middle part.
- Chin pointing towards opposite side. Head towards same side *(Scoliosis capitis)*.
- In later age groups it causes *hemifacial atrophy* due to less blood supply as a result of compression of the external carotid artery by sternomastoid tumour and due to kinking by position of neck. Distance between the outer canthus of eye to angle of mouth is reduced, with less arched eyebrow, flat or less filled cheek and flat nose compared to opposite side.
- Compensatory cervical scoliosis.
- Compensatory squint.

Differential Diagnosis

Other causes for torticollis.

Treatment

- Division of the lower end of the sternomastoid muscle or excision of the muscle. Both sternal and clavicular heads of sternocleidomastoid muscle should be divided under general anaesthesia using horizontal incision. One should not injure IJV, carotid, vagus, spinal accessory nerve. Additional all fibrous bands are also cut. Usually over correction is done. Physiotherapy exercise and toricollis harness is used for 6-12 months.
- Exercise and active stimulation of muscles in early cases.

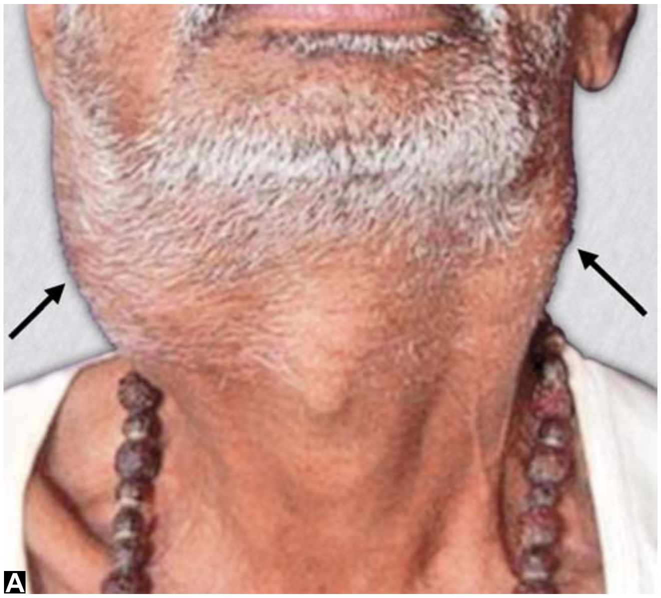

Fig. 5.31A

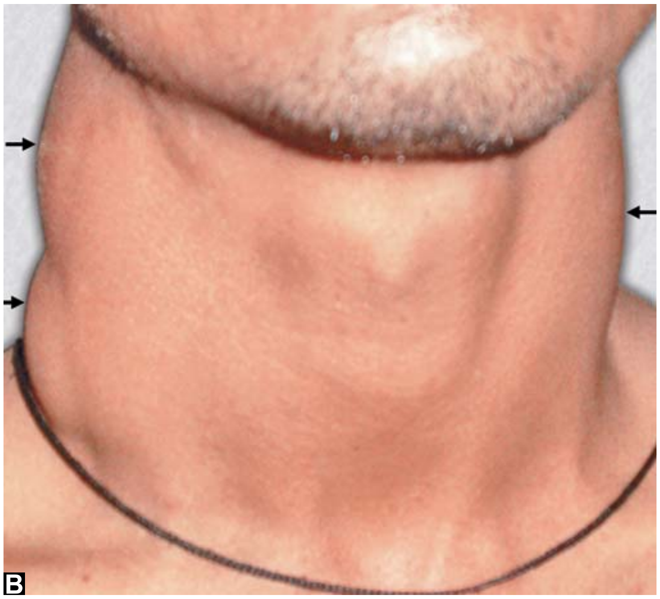

Fig. 5.31B

Figs 5.31A and B: Lymphoma neck involving both sides. Lymphomas are smooth, nontender, firm/India rubber consistency.

Differential diagnosis for neck lymph node enlargement	
• Tuberculous lymphadenitis	• Secondaries in lymph nodes
• HIV infection	• Lymphomas
• Chronic lymphatic leukaemia	• Nonspecific lymphadenitis
• Infectious mononucleosis	• Sarcoidosis
• Actinomycosis	• Brucellosis
• Toxoplasmosis	

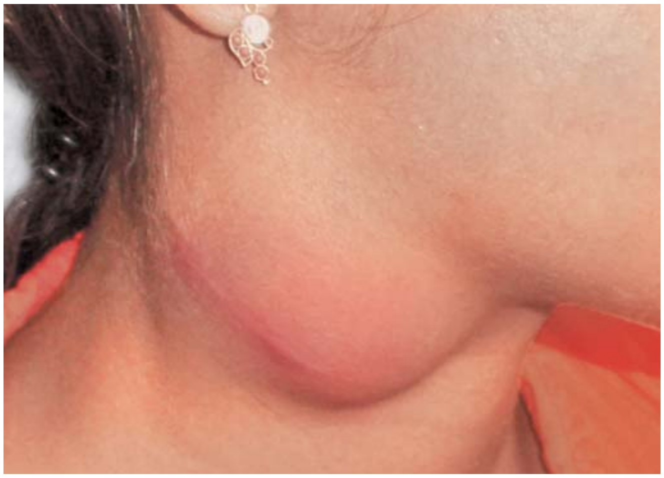

Fig. 5.32: Swelling in the neck well-localised. Note the scar of previous biopsy/excision. It could be lymph node enlargement due to lymphoma or tuberculosis.

TUBERCULOUS LYMPHADENITIS

Causative organism: *Mycobacterium tuberculosis* (not *M. bovis*).

Site

- Common in neck lymph nodes.
- Common in *upper deep cervical (jugulodigastric—54%) lymph nodes.*

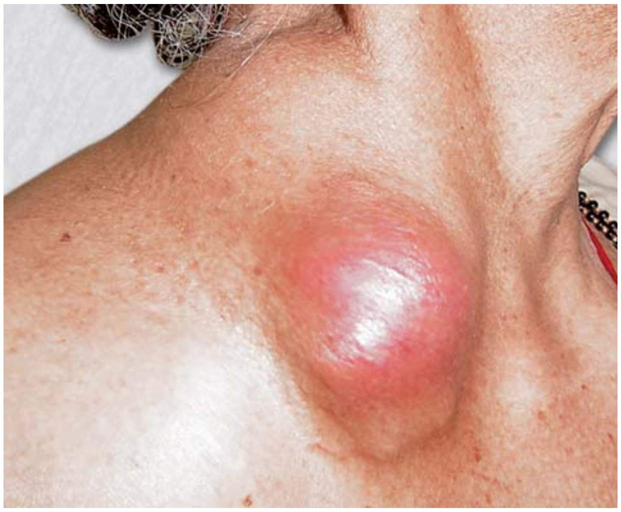

Fig. 5.33: Cold abscess in the neck which has formed collar stud abscess. It eventually leads into sinus formation.

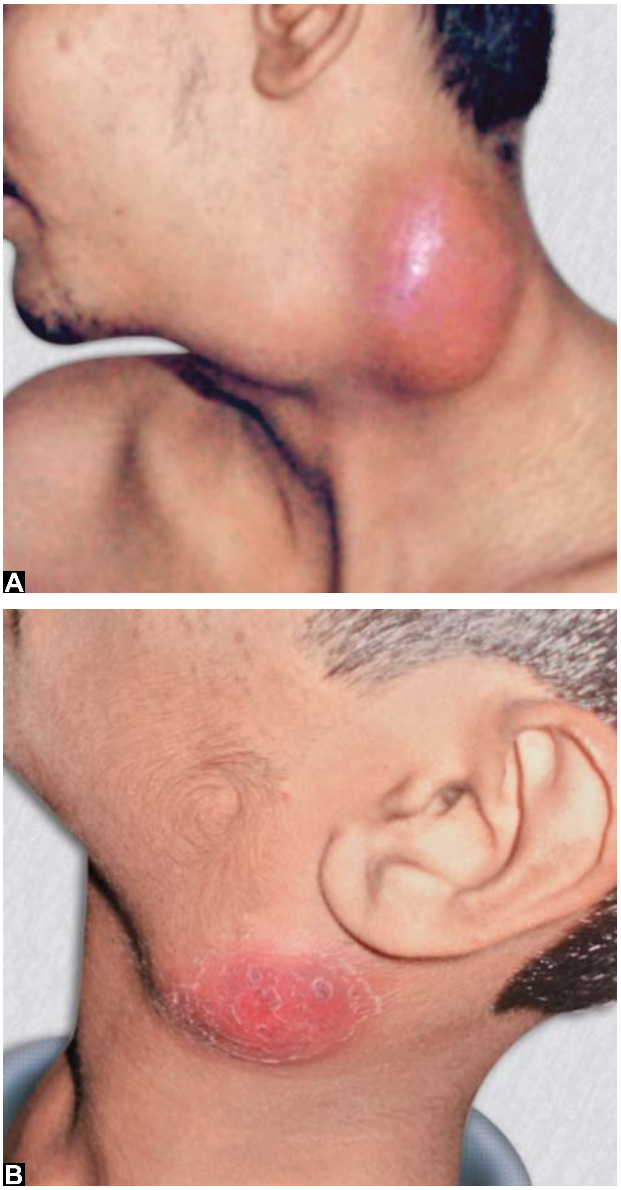

Figs 5.34A and B

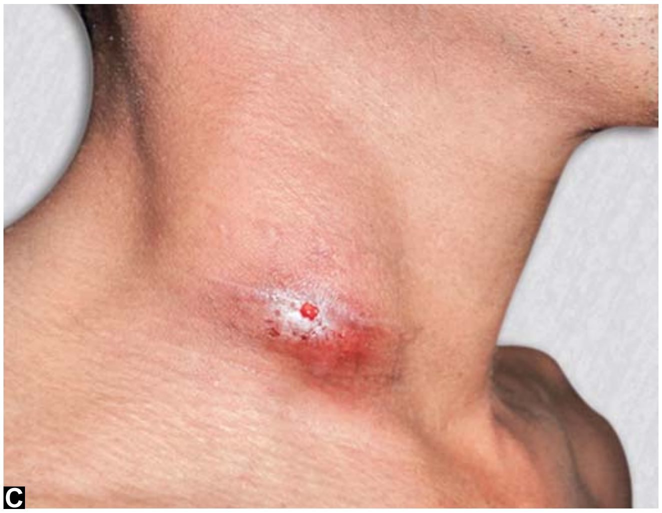

Fig. 5.34C

Figs 5.34A to C: Cold abscess due to caseating tuberculous lymphadenitis in neck. Collar stud abscess in the neck. Tuberculous sinus formation after drainage.

- Next common is posterior triangle lymph nodes (22%).
- Disease can also occur in other lymph nodes like, axillary lymph nodes, para-aortic lymph nodes, mesenteric lymph nodes, inguinal lymph nodes.
- Disease may be associated with HIV infection, lymphomas.

Mode of Infection

- Usually through the *tonsils*, occasionally through blood from lungs. *Tonsillar* infection shows multiple tubercles on

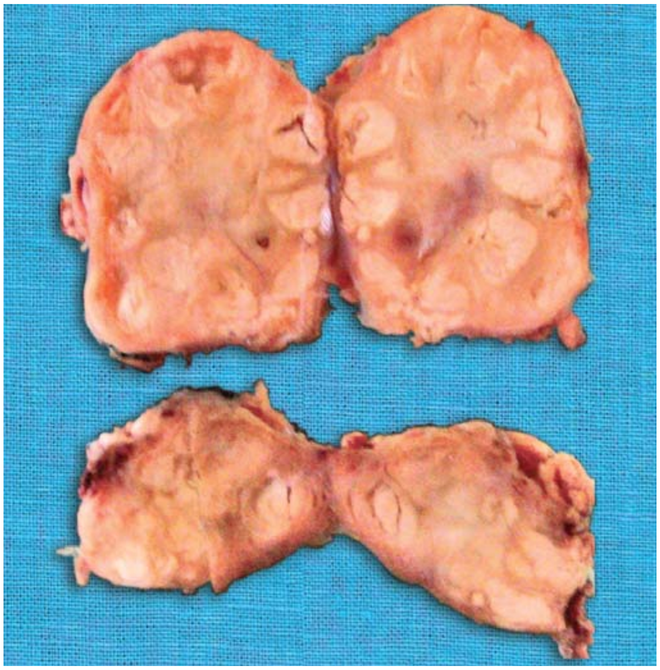

Fig. 5.35: Tuberculous lymphadenitis—cut section of the specimen. Note the yellowish caseating material with periadenitis (matting). Caseating type is more common than hyperplastic type.

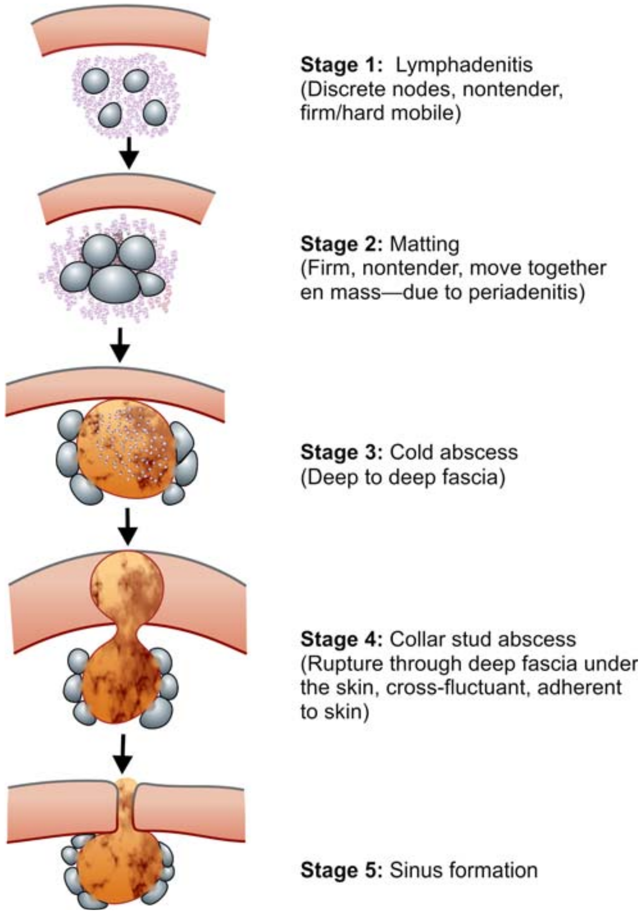
Stage 1: Lymphadenitis
(Discrete nodes, nontender, firm/hard mobile)
Stage 2: Matting
(Firm, nontender, move together en mass—due to periadenitis)
Stage 3: Cold abscess
(Deep to deep fascia)
Stage 4: Collar stud abscess
(Rupture through deep fascia under the skin, cross-fluctuant, adherent to skin)
Stage 5: Sinus formation

Fig. 5.36: Stages of tuberculous lymphadenitis (*Hogarth's Rake*).

its surface; from here infection spreads into jugulodigastric nodes (anterior triangle nodes) then to other nodes. Infection reach lymph node first into subcapsular space/sinus then to lymph node cortex which contains plenty of lymph follicles. Matting is due to periadenitis involving subcapsular sinus/space of lymph node. In children infection to neck node can come from either tonsils or adenoids or both. When it occurs from *adenoids*, lymph nodes in posterior triangle are involved through retropharyngeal lymphatics.

- It may be associated with pulmonary tuberculosis or renal tuberculosis. Through blood infection reaches medullary cords of lymph node and so medulla of lymph node.
- Rarely spread can occur from tuberculous lesion of the apex of lung through suprapleural Sibson's fascia/membrane to supraclavicular nodes.

Stages of tuberculous lymphadenitis

1. Stage of *infection, and lymphadenitis*
2. Stage of *periadenitis with matting*
3. Stage of *caseating necrosis and cold abscess* formation
4. Stage of formation of *collar stud abscess*
5. Stage of formation of *sinus* which discharges yellowish caseating material

- Often fibrosis and calcification can occur with or without treatment.

Gross Pathology

- Firm, matted, lymph node, with cut section showing yellowish caseating material.

Microscopic Features

- Epithelioid cells with caseating material are seen along with Langhan's type of giant cells.

Clinical Features

- Swelling in the neck which is firm, matted.
- Cold abscess is soft, smooth, nontender, fluctuant, without involvement of the skin. It is not warm.
- As a result of increased pressure, cold abscess ruptures out of the deep fascia to form *collar stud abscess* which is adherent to the overlying skin.
- Once collar stud abscess bursts open, discharging sinus is formed. It can be multiple, wide open mouth, often undermined, nonmobile with bluish color around the edge. It is usually not indurated.
- *Tonsils may be studded with tubercles* and so clinically should always be examined.
- Associated pulmonary tuberculosis should also be looked for. In 20% cases of tuberculous lymphadenitis, there may be associated pulmonary tuberculosis or it may be a primary focus.
- Cervical spine is examined for tuberculosis.

Types

1. *Hyperplastic*			2. *Caseating*	
a.	20% common	——	a.	80% common
b.	*Discrete*, firm or hard	——	b.	*Matted* due to periadenitis
c.	Occurs in the cortex of lymph node	——	c.	Involves medulla with periadenitis
d.	Host immunity is good	——	d.	Body resistance is not adequate
e.	Drugs act better	——	e.	Drugs do not reach in proper concentration and may not be effective
f.	Drug resistance is uncommon	——	f.	Drug resistance is common
g.	No cold abscess or sinus formation	——	g.	Cold abscess or sinus are common
h.	Blood spread	——	h.	Spread from tonsils

- Axillary nodes, when involved, is due to retrograde lymphatic spread from neck nodes or blood spread.
- Inguinal lymph nodes are involved occasionally through blood.
- Bluish hyperpigmented involved overlying skin is called as *scrofuloderma*.
- Tuberculous pus with *caseating cheesy creamy* material is *infective* as it contains multiplying organisms.
- Atypical mycobacterial tuberculosis can occur occasionally. Such disease may be *resistant* to drug therapy.
- *Sinus may persist due to*—fibrosis, calcification, secondary infection, inadequate reach of drug to maintain optimum concentration in caseation.

Clinical types

- *Acute type:* seen in infants and early childhood below 5 years
- *Hyperplastic type:* Lymphoid hyperplasia is typical; it is seen in patients with good resistance; hard discrete mobile lymph nodes; 15-20% common
- *Caseating type:* 75-80% common; matted nodes often with cold abscess; poor body resistance; seen in young adults
- *Atrophic type:* Rare type; seen in elderly; small lymph nodes but caseating type with atrophied nodes

Differential Diagnosis

1. Nonspecific lymphadenitis.
2. Lymphomas, and chronic lymphatic leukaemia.
3. Secondaries in the neck.
4. Branchial cyst mimics cold abscess.
5. Lymph cyst mimics cold abscess.
6. HIV with lymph node involvement.
7. When there is discharging sinus—actinomycosis.

Cold abscess

- Deep to deep fascia
- No evidence of signs of inflammation
- Not warm, nontender, smooth, soft and fluctuant, non-transilluminating
- Not adherent to skin (skin is free); no redness
- Contains cheesy caseating material
- It is seen in caseating tuberculous lymphadenitis due to caseation necrosis
- It may form collar stud abscess and later sinus
- FNAC, AFB, culture are useful investigations
- Differential diagnosis are branchial cyst, lymph cyst
- Treated by
  - Antituberculous drugs
  - 'Zig-zag' aspiration by wide bore needle in non dependent area to prevent sinus formation
  - Drainage using nondependent incision; later closure of the wound without placing a drain

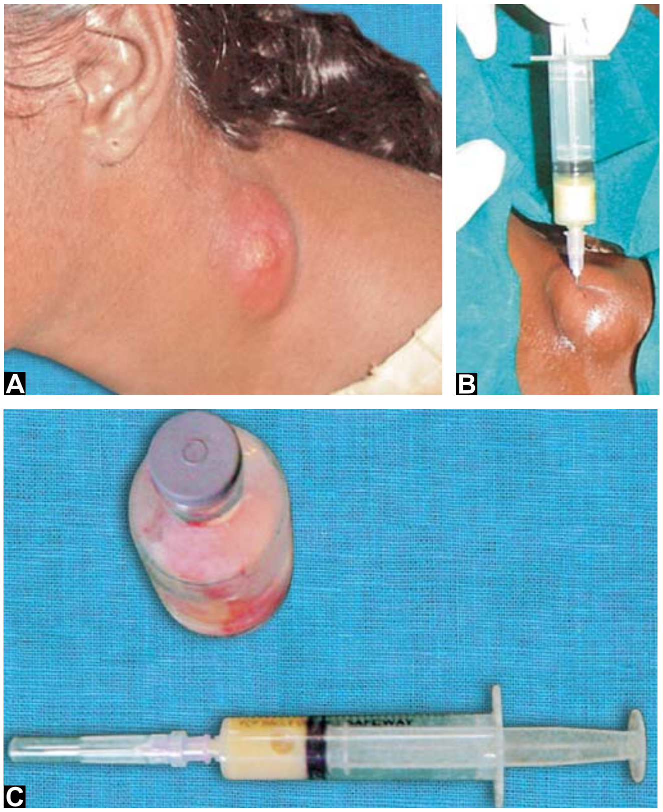

Figs 5.37A to C: Typical cold abscess in the neck in which pus is aspirated. Pus should be sent for cytology (for epithelioid cells), staining (Ziehl-Neelsen—AFB) and culture.

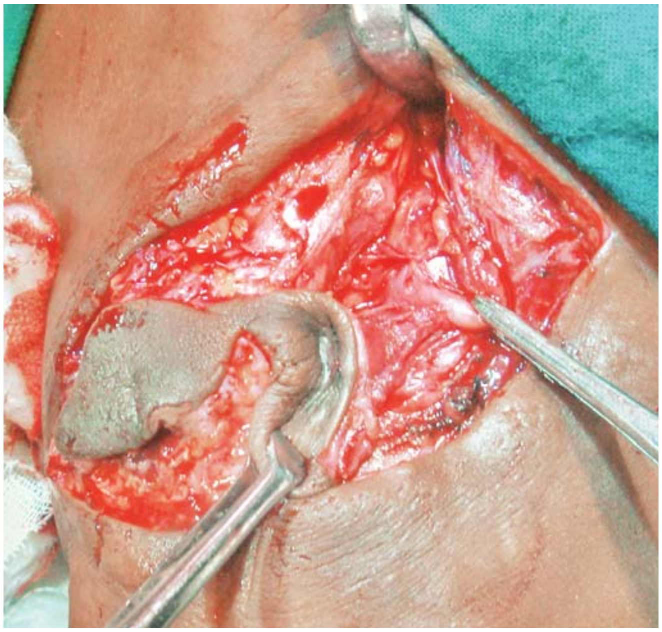

Fig. 5.38: Tuberculous sinus excision with excision of diseased lymph nodes.

Investigations

- Haematocrit, ESR, peripheral smear.
- FNAC of lymph node and smear for AFB and culture. FNAC is very useful but not as superior as open node biopsy. False negative, false positive results and altering the node architecture, and so eventual need of open biopsy are the problems. *Epithelioid cells* (modified histiocytes/macro-

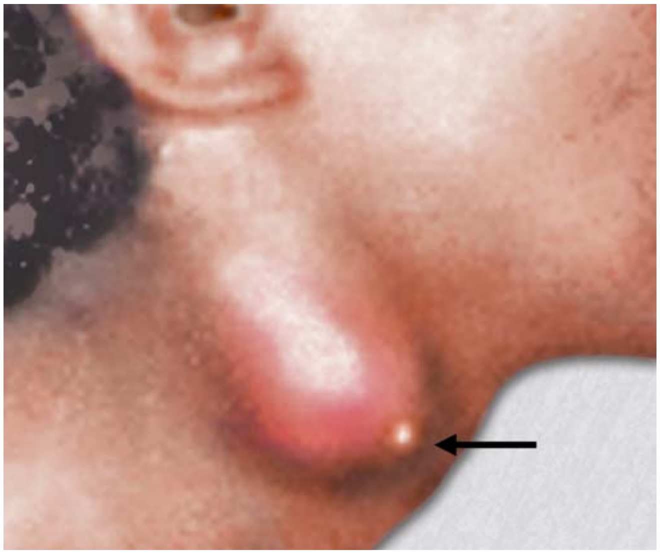

Fig. 5.39: Cold abscess with impending rupture.

phages) are diagnostic. *Langhans giant cells*, lymphocytes, plasma cells are other features.

- *Open biopsy* when FNAC is inconclusive. Open biopsy is more *reliable* for tuberculosis (and also in lymphoma; but it is contraindicated in node secondaries); entire node ideally two nodes if possible has to be taken intact; one in formalin for pathology, other in normal saline for microbiology (AFB).
- HIV test (ELISA and western blot).
- *Lowenstein Jensen media* is used for culture which takes 6 weeks to give result; *so selinite media* is often used which shows growth in 5 days.
- *Mantoux test* may be useful; but not very reliable.
- Chest X-ray to look for pulmonary tuberculosis.
- Polymerase chain reaction (PCR) is very useful method.

Treatment

- Drugs

Antitubercular drugs has to be started:

1. Rifampicin 450 mg OD on empty stomach. It is bactericidal. It discolours urine red. It is also hepatotoxic.
2. INH: 300 mg OD. It is bactericidal. It causes intolerance of GIT, Neuritis, Hepatitis (INH).
3. Ethambutol 800 mg OD. It is bacteriostatic. It causes GIT intolerance, retrobulbar neuritis (green colour blindness).
4. Pyrazinamide 1500 mg OD (or 750 mg BD). It is bactericidal. It is hepatotoxic, also causes hyperuricaemia and increases psychosis.

Duration of treatment is usually 6-9 months.

- Aspiration

When there is cold abscess, initially it is aspirated. (Wide bore needle is introduced into the cold abscess in a nondependent site along a *"Z" track* (in zig-zag pathway) so as to prevent sinus formation.)

- *Incision and drainage*

If it recurs, then it should be drained. Drainage is done through a nondependent incision. After draining the caseating material, wound is closed without placing a drain.

- *Surgical removal*

Surgical removal of tubercular lymph nodes are indicated when

1. There is no local response to drugs or
2. When sinus persists.

It is done by raising skin flaps and removing all caseating material and lymph nodes. Care is taken not to injure major structures.

- *Excision of the sinus track* is often essential when sinus develops.

COLD ABSCESS

Cold abscess is common in neck. It can also occur in groin, intercostal space, loin or any site where tuberculous caseating material with cheesy content can get collected and localised.

Sites of Origin

Cold abscess may originate from tuberculosis of spine (thoracic or cervical spines), lymph node, internal organs, bone, etc.

In the Neck

- *Tuberculous lymphadenitis* is common cause. Here cold abscess is commonly seen in *anterior* triangle.
- *Tuberculosis of cervical spine* is also an important cause. Commonly here cold abscess occurs in *posterior* triangle. Caseating material from the cervical spine collects *in front of* the vertebra behind the prevertebral fascia which eventually ruptures either anteriorly or posteriorly.
    - *Anterior rupture* allows passage of caseating material *below* and behind the prevertebral fascia reaching superior mediastinum; *laterally* behind the prevertebral fascia and carotid sheath to form cold abscess in posterior triangle; in midline upper part, protruding *forwards* from behind the prevertebral fascia in *midline* presenting as chronic retropharyngeal abscess; in midline lower part protrudes into oesophagus; caseation runs along the axillary sheath and neurovascular plane to reach axilla and arm to cause cold abscess in axilla and arm/cubital fossa.
    - *Posterior rupture* occurs towards spinal canal facilitating the passage of caseation along the cervical nerves towards posterior triangle and brachial plexus and so axilla and arm.

Features

- It is common in young but can occur in any age group. Equal incidence in both sexes.
- Swelling in the neck, which is smooth, nontender, soft, fluctuant, nontransilluminating, with restricted mobility but is not adherent to skin.
- Neck pain, neck rigidity, restricted movements of cervical spine in case of cervical spine tuberculosis. With every change of position and often when patient is seated he supports his head with his hands and forearm—*Rust's sign* (Jan N Rust Surgeon, Poland).
- Evening fever, loss of weight and appetite, anaemia.

- Features of systemic disease if present like of pulmonary tuberculosis—cough, haemoptysis.
- Matted lymph nodes adjacent to cold abscess may be palpable.
- Oral cavity, tonsils, chest should be examined.
- Raised ESR, positive Mantoux test, anaemia, lymphocytosis, chest X-ray may show pulmonary tuberculosis, aspiration of cold abscess (FNAC) to see microscopically *epithelioid cells*. Acid fast bacilli may be identified from the aspirated fluid using Ziehl-Neelsen stain.
- X-ray neck in case of cervical spine tuberculosis to identify reduced joint space, vertebral destruction, soft tissue shadow.
- MRI of cervical spine, US/CT scan neck are needed to confirm the anatomical location, number of lesions.

Sequelae of Cold Abscess

- Secondary infection of the cold abscess making it tender.
- Formation of collar stud abscess, once pressure increases inside the cold abscess which will give way through the deep fascia to reach the subcutaneous plane to get adherent to skin.
- Sinus formation.
- Spread of disease to multiple lymph nodes and other organs.

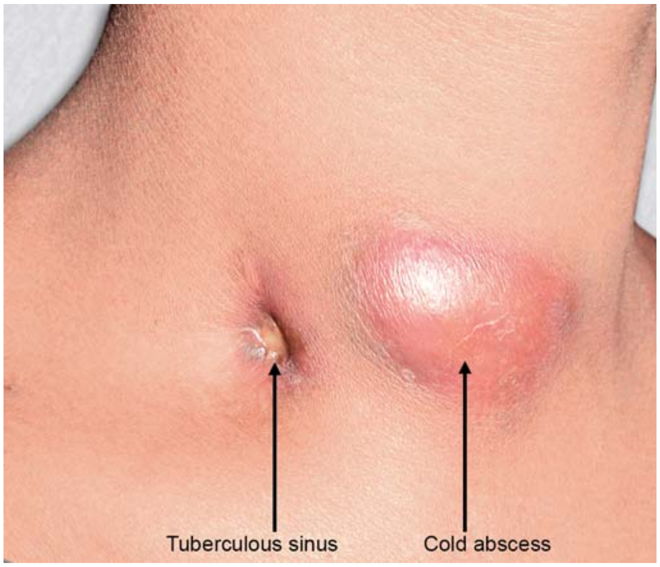

Fig. 5.40: Tuberculous cold abscess and sinus in the neck.

Differential Diagnosis

- Branchial cyst and other cystic swellings in neck.
- Secondaries in neck lymph nodes.
- Secondaries in cervical spine.

Treatment

- Antituberculous drugs.
- Nondependent aspiration of the cold abscess.
- Excision of the diseased neck nodes.
- Immobilization of cervical spine by plaster jacket/collar for 4 months. Cervical spine fusion by open surgical method if diseased spine is unstable.

SECONDARIES IN NECK LYMPH NODES

Levels in Neck Nodes (Memorial Sloan—Kettering Cancer Centre Levelling of Neck Nodes)

Level I—Submental (Ia) and submandibular (Ib) lymph nodes.

Level II—Lymph nodes in upper deep cervical region. (It extends from base of skull to hyoid bone and from lateral margin of sternohyoid to posterior margin of sternomastyoid muscle).

Level III—Lymph nodes in middle cervical region (from hyoid bone to omohyoid muscle or cricothyroid membrane).

Level IV—Lymph nodes in lower cervical region (from omohyoid muscle to clavicle).

Level V—Lymph nodes in posterior triangle including supraclavicular region from posterior border of sternocleidomastoid muscle to anterior border of trapezius muscle.

Level VI—Lymph nodes in the midline neck—pretracheal and prelaryngeal from hyoid bone above to suprasternal notch below, medial border of carotid sheath on either side.

Level VII—Lymph nodes in the mediastinum. inferior to suprasternal notch to innominate artery below.

Note:

Level II and V are now subdivided into Level IIa/Level IIb and Level Va/Level Vb; depending whether these nodes are above the level (Level IIb/Level Va) of the spinal accessory nerve or below (Level IIa/Level Vb). *Level I* node from oral cavity, lip, salivary gland, skin; *level II* node from oral cavity, oropharynx, nasopharynx, salivary gland; *level III* node from oral cavity, oropharynx, hypopharynx, larynx, thyroid; *level IV* node from oropharynx, hypopharynx, larynx, thyroid, cervical oesophagus; *level V* node from nasopharynx, scalp, GIT, breast, lungs.

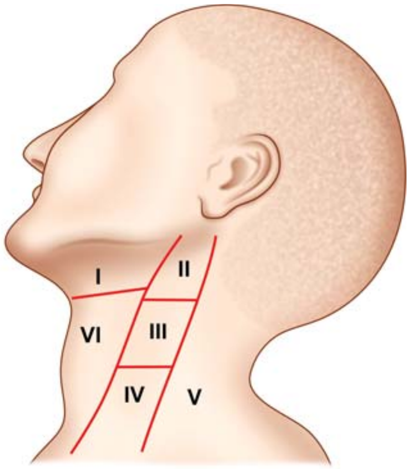

Fig. 5.41: Levels of neck nodes.

Common sites of primary

- Oral cavity, tongue, tonsils
- Salivary glands
- Pharynx—nasopharynx
- Larynx
- Oesophagus
- Lungs
- GIT
- Thyroid

Example isn't the best way to teach, it's the only way.

- ♦ It is commonly from squamous cell carcinoma, but can also be from adenocarcinoma or melanoma.
- ♦ Squamous cell carcinoma is mainly from oral cavity, pharynx.

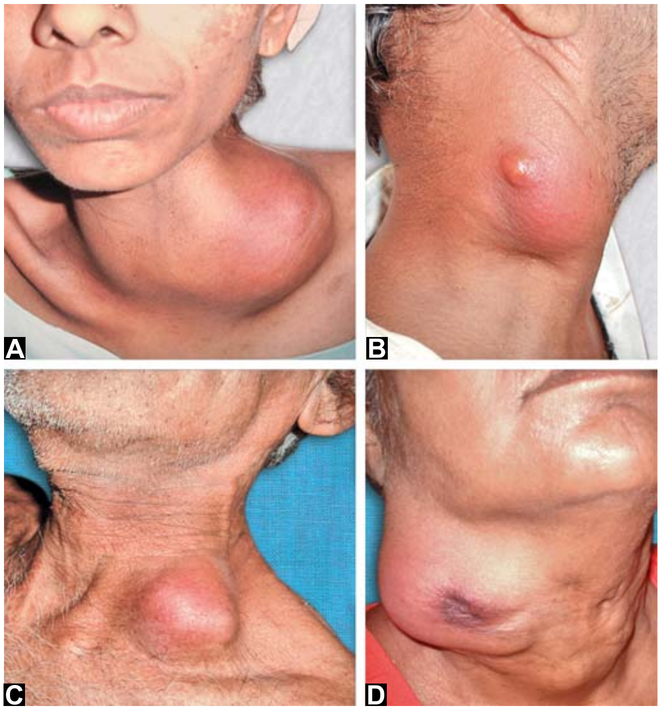

Figs 5.42A to D: Typical secondaries in the neck. Note the different levels involved in different patients. Note the skin involvement and sinus formation in few photos.

- ♦ Adenocarcinoma is usually from GIT, commonly involving left supraclavicular lymph nodes.
- ♦ Breast, lungs, abdominal viscera are other areas where primary may cause secondaries in neck which should be examined when suspected.

Note:

- 20% of metastatic squamous cell carcinoma is from aerodigestive tract.
- Source of primary from head and neck region usually have frequency like this—nasopharynx, tonsil, base of tongue, thyroid, larynx, floor of the mouth, cheek, palate, pyriform fossa. Non-head and neck source of primary are—bronchus, oesophagus, breast, stomach.
- Histologically it is SCC (80%) or non-SCC. In non-SCC it may be poorly differentiated carcinoma (10-15%), adenocarcinoma (5-10%) or others like melanoma, poorly differentiated neoplasm (5%).
- In poorly differentiated neoplasm immunohistochemistry/immunoperoxidase staining, electron microscopy and chromosome analysis are needed to rule out lymphoma, neuroendocrine tumours, undifferentiated sarcomas.
- 3 years after treatment of one primary, if recurrence occur it is called as *new primary* in aerodigestive tract.
- Lower lip, tongue, soft palate and supraglottis can cause bilateral secondaries in neck.
- Soft palate, retromolar region, nasopharynx, hypopharynx posterior and lateral oropharynx can involve retropharyngeal nodes.
- Nasopharyngeal carcinoma spreads to level II-V, retro and parapharyngeal nodes.
- 15% of secondaries are from *infraclavicular primaries*—lung, pancreas, oesophagus, stomach, breast, ovaries, testis, prostate.
- Nodal spread below the lower border of cricoid (level IV and V; **L**ower) carries poor prognosis compared to nodes above cricoid (**U**pper).

Branchiogenic Carcinoma

It is a primary squamous cell carcinoma arising from remnants of branchial cleft or arch. It is a differential diagnosis for secondaries in neck. It is common in men. It is common near the area of carotid bifurcation. Histologically it contains malignant squamous cells with lymphoid tissues around. It spreads to lymph nodes and can infiltrate into adjacent soft tissues.

Treatment: Wide excision.

Features of Secondaries in Neck

- ♦ Common in adult/elderly male (Male to female ratio is 4:1), presents as painless rapidly increasing localised swelling in the neck.
- ♦ Nodular surface and hard in consistency, often fixed when it is advanced.
- ♦ Secondaries from *papillary carcinoma of thyroid* can be soft, cystic with brownish black fluid.
- ♦ Secondaries can infiltrate into carotids, sternomastoid, posterior vertebral muscles, spinal accessory nerve (shrugging of shoulder is affected), hypoglossal nerve (tongue will deviate towards the same side), cervical sympathetic chain *(Horner's syndrome)*.
- ♦ Secondaries spread into adjacent soft tissues and also to the skin causing fungation and ulceration. Often because of tumour necrosis, softer area develops in the hard node. Skin fold prominence due to infiltration of the platysma is typical.
- ♦ In advanced cases tumour may infiltrate into the major vessels like carotids, or branches of external carotid artery causing torrential haemorrhage.
- ♦ Dysphagia, dyspnoea, haemoptysis, hoarseness of voice, ear pain, deafness are other features depending on the primary site.

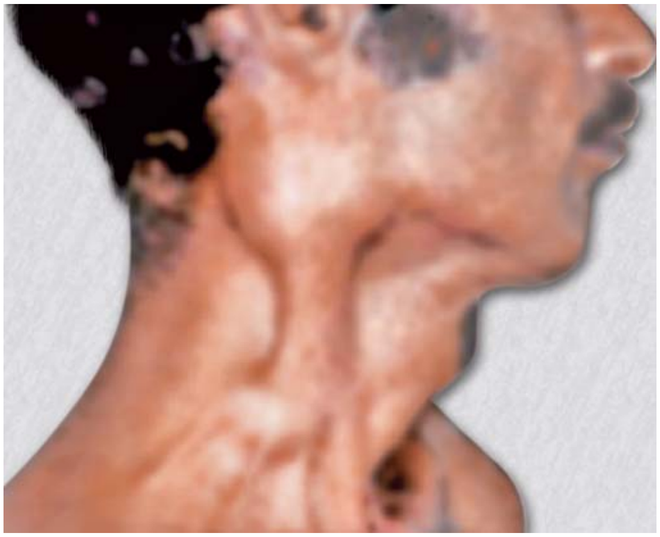

Fig. 5.43: Secondaries in the neck nodes from laryngeal carcinoma.

- *Hoarseness*—carcinoma larynx, thyroid
- *Dysphagia*—carcinoma posterior 1/3rd of the tongue, pharynx, oesophagus
- *Haemoptysis, cough, dyspnoea*—carcinoma lung
- *Ear pain, deafness*—nasopharyngeal carcinoma
- *Spinal accessory nerve*—shrugging of shoulder is difficult
- *Hypoglossal nerve*—tongue deviates to same side with wasting
- *Sympathetic chain*—Horner's syndrome with miosis, anhidrosis, upper eyelid droop (pseudoptosis), enophthalmos, loss of spinociliary reflex

Types of Secondaries in the Neck

1. Secondaries in the Neck with Known Primary

- Here secondaries are present and primary has been identified clinically in the oral cavity, pharynx, larynx, thyroid or other areas.
- Biopsy from the primary and FNAC from the secondaries are done.
- Primary is treated accordingly either by curative radiotherapy or by surgery (wide excision).
- Secondaries, when mobile are treated by radical lymph node block dissection in the neck.

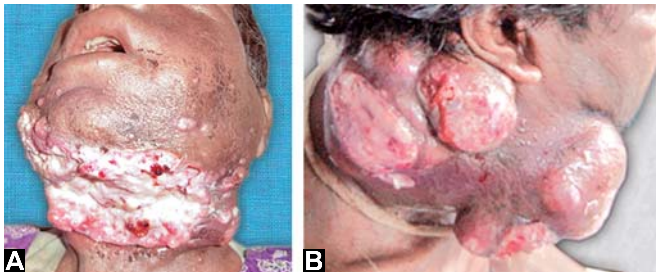

Figs 5.44A and B: Advanced secondary carcinoma neck with ulceration and fungation.

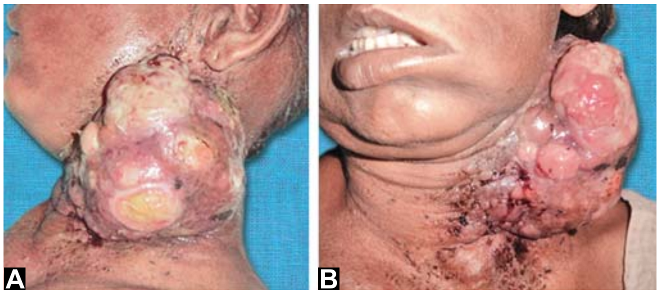

Figs 5.45A and B: Large secondaries in neck with extensive fungation.

2. Secondaries in the Neck with Clinically Unidentified Primary

- Hard neck lymph nodes are the secondaries, but primary has not been identified clinically.
- FNAC of the neck node is done and secondaries is confirmed. Then search for the primary is done by various investigations.

They are:

a. Panendoscopy
   - Nasopharyngoscopy.
   - Laryngoscopy.
   - Oesophagoscopy.
   - Bronchoscopy.

b. Blind biopsies are taken from fossa of Rosenmuller, lateral wall of pharynx, pyriform fossa, tonsillar bed, base of tongue, subglottic region (larynx). It is called as *surveillance biopsy* and is done to reveal unknown primary in 15% of cases of secondaries in neck. If this surveillance biopsy is negative, then *ipsilateral tonsillectomy* may be needed.

c. FNAC of thyroid and suspected areas.

d. CT scan.

Once the biopsy confirms the primary, it is treated either by surgery or by *curative radiotherapy*.

Secondaries in the neck is treated by *radical neck dissection*.

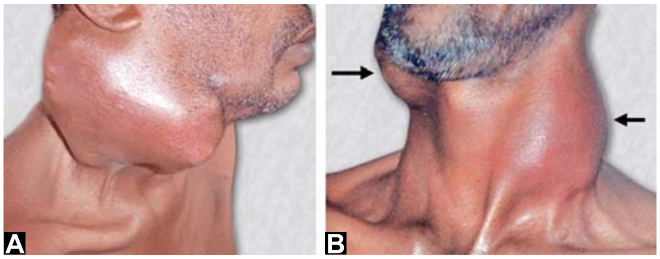

Figs 5.46A and B: (A) Large secondaries in neck with nodularity, with involvement of platysma below. (B) Bilateral secondaries are observed. **Skin fold prominence is obvious**.

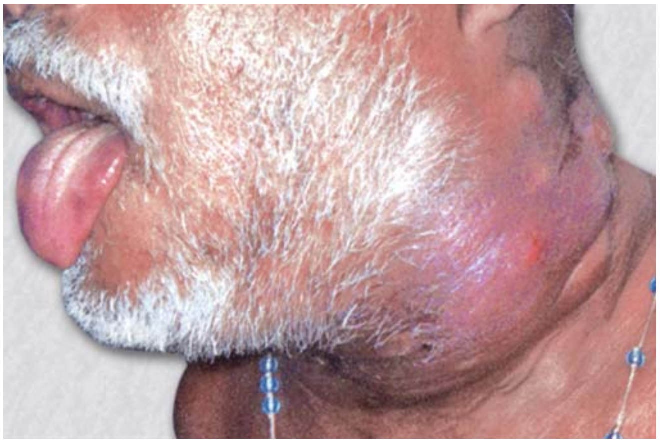

Fig. 5.47: Secondaries in neck with nodules and ulceration. Note the deviation of tongue towards same side due to hypoglossal nerve involvement.

3. Secondaries in the Neck with an Occult Primary

Occult primary sites which can cause secondaries in neck

- Fossa of Rosenmuller
- Lateral wall of pharynx
- Posterior third of the tongue
- Thyroid
- Paranasal sinuses
- Bronchus
- Oesophagus

Listen to the faintest sound; for opportunities knock only once.

Histologically secondaries in neck with occult primary may be of squamous cell carcinoma or of nonsquamous cell carcinoma, i.e. adenocarcinoma/poorly differentiated tumours (lymphoma/sarcoma/melanoma). In upper and midcervical region 80% are due to squamous cell carcinomas. In lower cervical and supraclavicular region 40% can be adenocarcinomas. Common sites of primary here (for adenocarcinoma) are thyroid, breast, gastrointestinal tract, salivary glands, lungs, prostate and kidney.

- Here secondaries in the neck lymph nodes are confirmed by FNAC, *but primary has not been revealed clinically and by any available investigations*.
- When all the investigations mentioned above are done, do not show any evidence of primary, only then it is called as *occult primary*.
- Primary tumor is not identified at the time when definitive therapy has started.
- 70% of occult nodes occur in jugulodigastric group.
- Differential diagnosis for secondary with occult primary is lymphoma and *primary branchiogenic carcinoma*.
- Reasons for primary lesion being occult—too small a primary to detect; possibility of immunological spontaneous regression of primary and inability of the present diagnostic tools to detect the primary.
- FNAC is the tool to confirm the occult secondary. If FNAC is inconclusive, only then open biopsy (incision/excision) is done to confirm. Open biopsy helps in high suspects of lymphomas or poorly differentiated carcinomas. It facilitates tissue study, immunohistochemistry, and special stains. Many studies prove that risk of seedling, survival and prognosis will not alter by open biopsy. But at present it is proposed only when FNAC fails or special methods are mandatory to type the disease. After open biopsy, frozen section confirmation and immediate neck dissection has to be done.
- Immunoperoxidase staining can be done in FNAC specimen or formalin fixed paraffin tissue using monoclonal or polyclonal antibodies. Immunoperoxidase is the most commonly used tool. It is mainly useful in lymphomas/neuroendocrine tumours. Electron microscopy is superior to immunohistochemistry as ultrastructure details can be assessed. But it is costly. Chromosomal analysis for tumour specific genes is used in B, T and germ cell lymphomas.

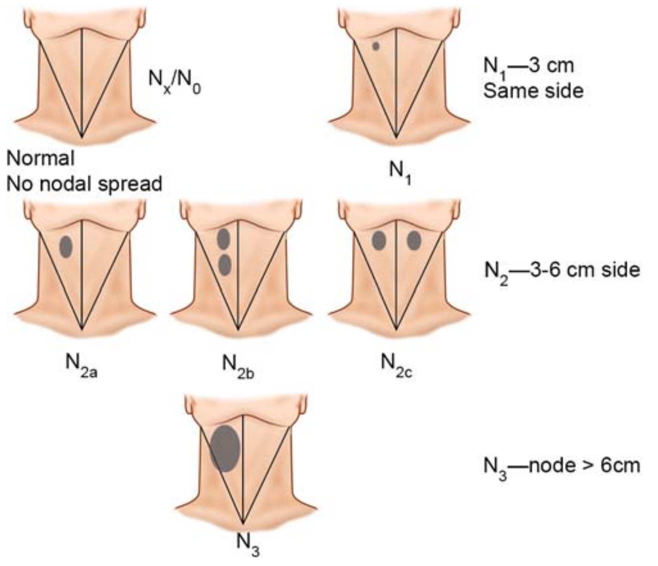

Fig. 5.48: Diagrammatic representation of neck lymph node staging in secondaries.

- Initially the secondaries in the neck are treated by *radical neck dissection*, then *regular follow-up* is done (at three monthly intervals) until the primary reveals.
- Once primary is revealed it is *confirmed by biopsy* and treated accordingly, either by *curative radiotherapy* or by *wide excision* depending on location of revealed primary.
- This type is usually less aggressive and has got better prognosis.

Nodal staging in secondaries

N_x—nodes cannot be assessed
N_0—no nodal metastasis
N_1—single node same side < 3 cm in greatest dimension
N_{2a}—single node same side 3-6 cm
N_{2b}—multiple nodes same side < 6 cm
N_{2c}—bilateral/contralateral nodes < 6 cm
N_3—node > 6 cm

Investigations for Secondaries in Neck

- *FNAC of secondary:* Open incision biopsy is not advised here. It destroys the fascial barriers and causes the spread of tumour faster and earlier into next level nodes or other soft tissues. Eventual neck dissection technically becomes difficult. Recurrence rate in neck will be higher after open biopsy. If FNAC of node and all investigations for primary become negative, then open biopsy of node following confirmation with frozen section and immediate neck dissection is undertaken. In such situation if neck dissection is delayed after open biopsy confirmation, chances of cure will be reduced.

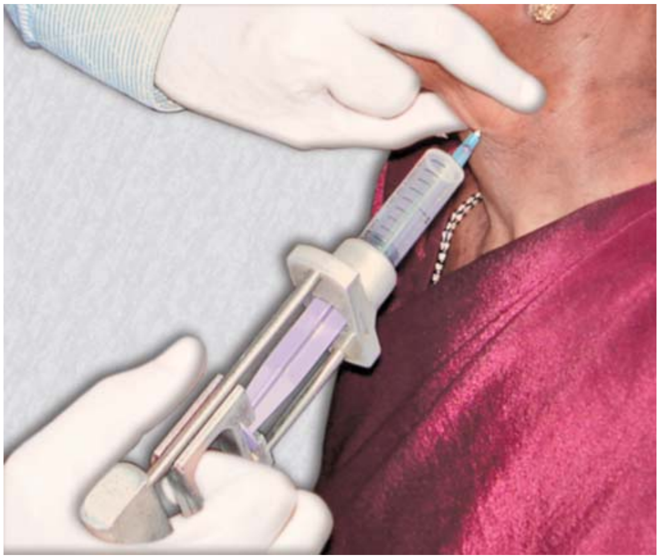

Fig. 5.49: FNAC of neck lymph node. It is very useful method in secondaries in neck node and tuberculosis. Its use is equivocal in lymphoma where open biopsy of the node is preferred method. *(Courtesy:* Dr Krishna Upadhya Pathologist, Nandikoor Laboratory, Mangalore).

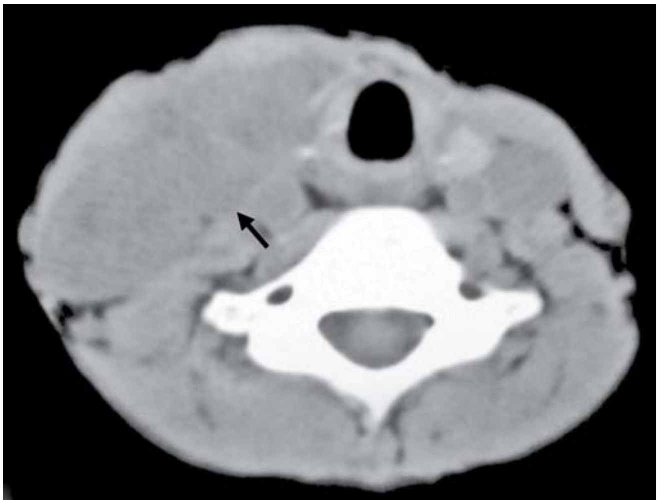

Fig. 5.50: CT neck showing secondaries.

- *Biopsy from primary:* Incision biopsy is the choice here.
- Blind biopsies from suspected areas.
- Nasopharyngoscopy, laryngoscopy, bronchoscopy, oesophagoscopy—*panendoscopy* with examination under anaesthesia.
- CT scan is to see the base of skull, paranasal sinuses, nasopharynx, extension of primary tumour/secondary deposits; CT scan of chest and abdomen.
- Chest X-ray to visualise primary or secondaries in case melanomas or mediastinal nodes.
- MRI scan or PET scan in conjunction with CT scan or MRI. MRI identifies soft tissue extension/changes; guided primary biopsy is possible; extension into bone is identified.
- CT chest and abdomen in case of infraclavicular primaries or to assess nodes.

Open incision biopsy is avoided in lymph node secondaries.

***Triple endoscopy** includes direct/indirect laryngoscopy, oesophagoscopy and bronchoscopy.*

Differential Diagnosis

1. Lymphomas.
2. Tuberculous lymphadenitis.
3. Nonspecific lymphadenitis.
4. HIV.
5. Chronic lymphatic leukaemia.

In secondaries neck with occult primary

- FNAC of node/open biopsy to confirm
- Proper clinical methods to identify the location of primary
- CECT is the investigation of choice to look for primary
- Other methods are—MRI, triple endoscopy, examination under anaesthesia, blind biopsies from fossa of Rosenmuller, pyriform fossa, base of tongue, subglottic area and tonsils, FNAC of thyroid, ipsilateral tonsillectomy if surveillance biopsy and other methods are negative
- Once occult primary is confirmed and if node is less than 3 cm in size, it is treated as N_1 disease with RND/MRND
- Postoperative RT is essential if disease is N_{2a}; N_{2b} (more than 2 nodes are positive); if 2 or more levels are involved; if extracapsular spread is present
- In fixed lymph node/nodes, initially radiotherapy is given followed by surgery (RND) if nodes become mobile and later adjuvant chemotherapy
- In N_{2c} bilateral neck dissection (one side IJV preservation) with bilateral RT is given
- N_3 stage has got poor prognosis. RND and RT with later chemotherapy are used. In fixed nodes, RT and then chemotherapy is used
- Proper follow-up at regular intervals is essential with all diagnostic tools to identify the possible site of primary which may get revealed during follow up period

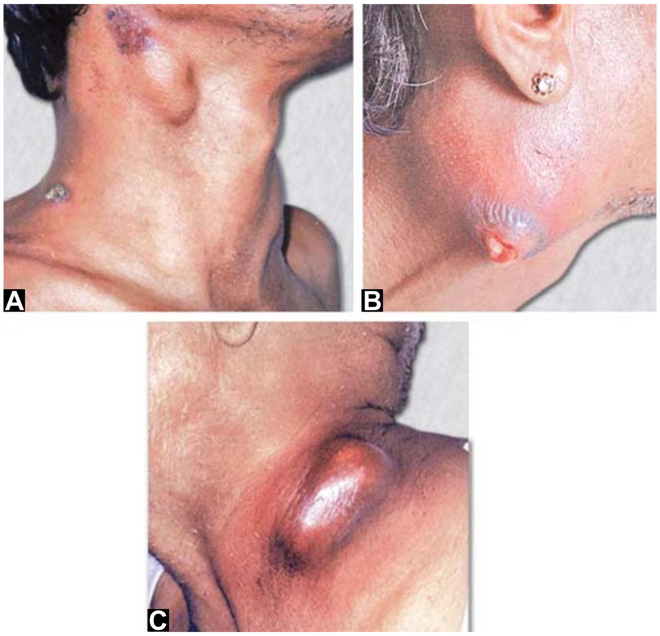

Figs 5.51A to C: (A) Secondaries in neck but no skin involvement is seen. (B and C) Secondaries with skin involvement is obvious.

Treatment

- Primary is treated depending on the site, either by *wide excision (surgery)* or by *curative radiotherapy*. Then the secondaries are treated.
- Secondaries when *mobile*, are treated by *radical neck dissection*.
- When *fixed* it is *inoperable. Palliative external radiotherapy* is given to palliate pain and to prevent the anticipated bleeding.
- Sometimes *initially, external radiotherapy* is given to *downstage* the disease so that it becomes operable and later classical block dissection can be done.

'Excellence' is an outcome, of good intentions and the right ways to do a work.

- *Postoperative RT* is given after neck dissection when—more than two lymph nodes are positive for metastases; nodes show metastases at two or more level; extracapsular spread in lymph node. Suspected occult primary is included/covered in the RT field. RT is also given to contralateral neck nodes in nasopharyngeal carcinoma. Level II lymph node *alone* from an occult is more likely to be from nasopharyngeal carcinoma and RT is preferred in such situation covering nasopharyngeal area; later RND is done.

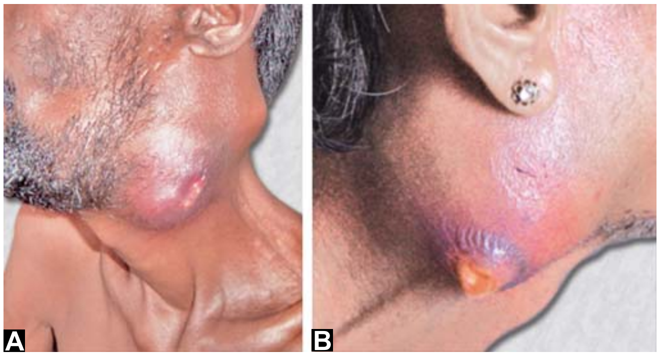

Figs 5.52A and B: Secondaries in neck with ulcer over the summit.

Types of Block Dissection

1. Classic Radical Neck Dissection

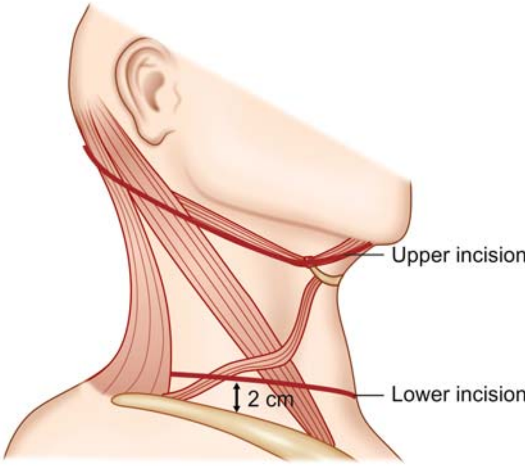
Upper incision
2 cm
Lower incision

Fig. 5.53: MacFee incision for radical neck dissection. Upper incision is from mastoid process along the line of digastric to hyoid bone point, then upwards to chin. Lower incision is parallel to clavicle 2 cm above from anterior margin of trapezius to midline.

It is resection of lymph nodes (level I to V), fat, fascia, sternomastoid muscle, omohyoid muscle, internal jugular vein, external jugular vein, accessory nerve, sub-mandibular salivary gland, lower part of parotid, prevertebral fascia—*"en-block"(Criles' operation)*.

Incision that is commonly made is MacFee incision which are two parallel incisions, one at submandibular region, another at supraclavicular region. Blood supply of the flap remains intact and so healing will be better without flap necrosis.

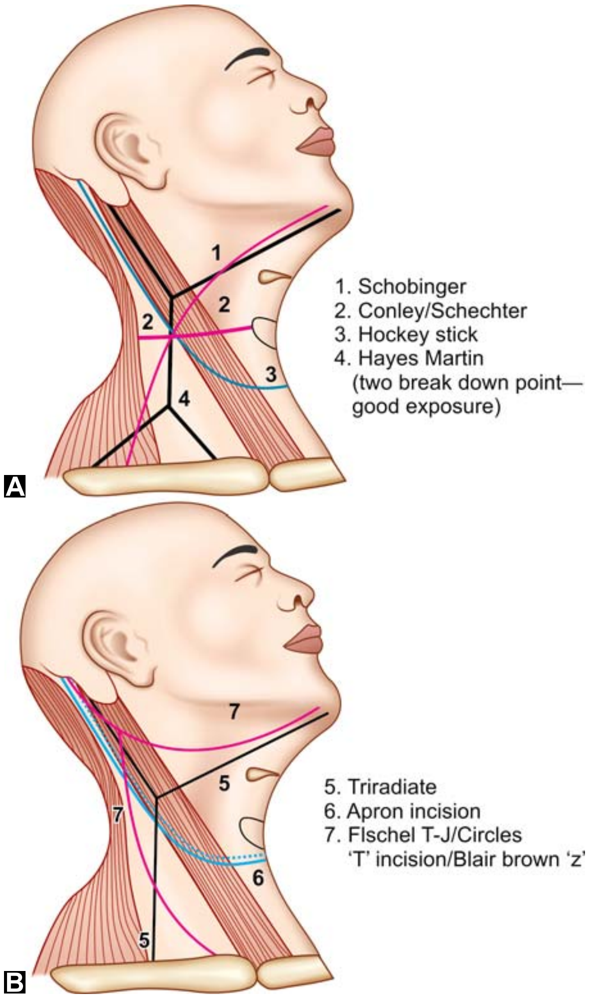
1
2
2
3
4
1. Schobinger
2. Conley/Schechter
3. Hockey stick
4. Hayes Martin
(two break down point—good exposure)
A
7
5
7
6
5
5. Triradiate
6. Apron incision
7. Flschel T-J/Circles
'T' incision/Blair brown 'z'
B

Figs 5.54A and B: Incisions/different approaches for radical (RND/MRND) neck dissections. There are many incisions mentioned. Few of them are shown in diagram.

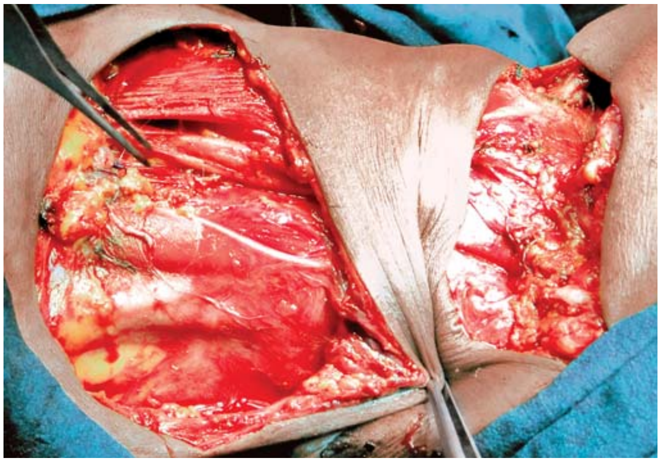

Fig. 5.55: Radical neck dissection, on table.

2. Conservative Functional Block Dissection (After BOCCA Surgeon): (Modified Radical Neck Dissection—MRND)

It is done only in selected cases where tumour is very well-differentiated and less aggressive like in papillary carcinoma

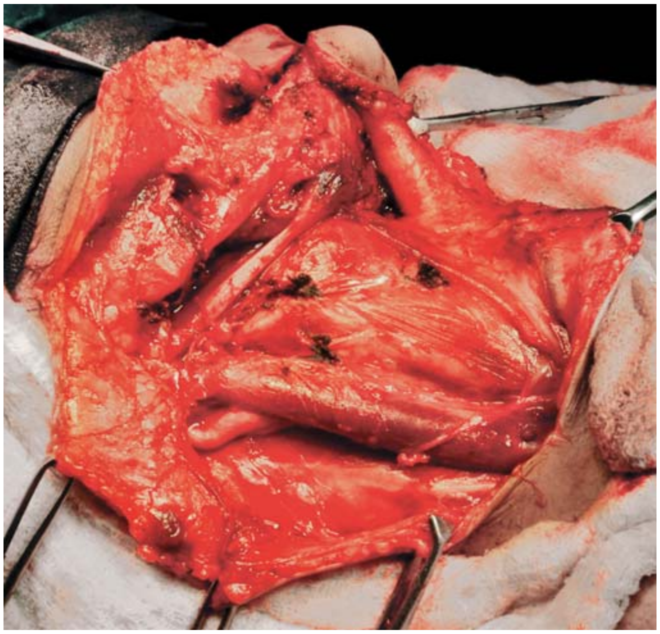

Fig. 5.56: Radical neck dissection on table photo. Note the internal jugular vein, digastric muscle, strap muscles, lower margin of the mandible.

of thyroid with lymph node secondaries. Structures preserved here are sternomastoid muscle, internal jugular vein and spinal accessory nerve.

- Only spinal accessory nerve is preserved—MRND type I (most important). **N**-Preserved.
- Accessory nerve and sternocleidomastoid are preserved—MRND type II. **NM**-Preserved.
- Accessory nerve, sternomastoid and internal jugular veins are preserved—MRND type III. **NMV**-Preserved. It is called as *functional neck dissection.*

3. Supraomohyoid Block

Removal of only fat, fascia, lymph nodes, muscles, submandibular salivary gland, with dissection above the omohyoid muscle is done. Done only in selected individuals with well-differentiated tumour and involvement of few submandibular lymph nodes *(Levels I, II, III are removed)*. Done in N0 lesions.

4. Bilateral Neck Dissection

Here internal jugular vein is preserved on one side. *Always the side where the vein is preserved, is operated first* (If both the jugulars are ligated, cerebral congestion occurs leading to cerebral oedema which is dangerous. Jugular veins are ligated as an inevitable procedure during surgery, the patient is kept in propped-up position; antibiotics, diuretics, steroids, mannitol infusion are given, repeated CSF taps are done to control the cerebral oedema).

Ligating one IJV increases the ICP by 3 fold*; both IJV ligations* increase ICP by 5 fold. ICP gradually falls over 8-10 days. For this reason pressure dressing should be *avoided* over the wound of neck dissection after surgery.

5. Commando Operation (Combined Mandibular Dissection and Neck Dissection)

It is *en-block removal*, which includes wide excision of primary tumour with hemimandibulectomy and neck block dissection, e.g. in tongue.

6. Lateral Neck Dissection (Anterolateral/Jugular/ALND)

It is done in laryngeal and pharyngeal primaries with clinically negative nodes. Levels II, III, IV are removed bilaterally.

7. Anterior (Central) Dissection

Level VI (pre, paratracheal) nodes are removed.

8. Posterolateral Dissection

Levels II, III, IV, V are removed for cutaneous malignancies, with suboccipital nodes.

9. Extended Radical Dissection

Additional nodes in the mediastinum are cleared (level VII). Nodes like level VI or parapharyngeal, retropharyngeal, external carotid artery, hypoglossal nerve, vagus nerve, parotid gland, mastoid tip—are addressed.

Complications of block dissection

- Haemorrhage
- Infection
- Lymph ooze
- Carotid blow out
- Seroma and flap necrosis
- Frozen shoulder is common
- Rarely pneumothorax and chylous fistula
- Drooping of shoulder due to paralysis of trapezius in radical neck dissection

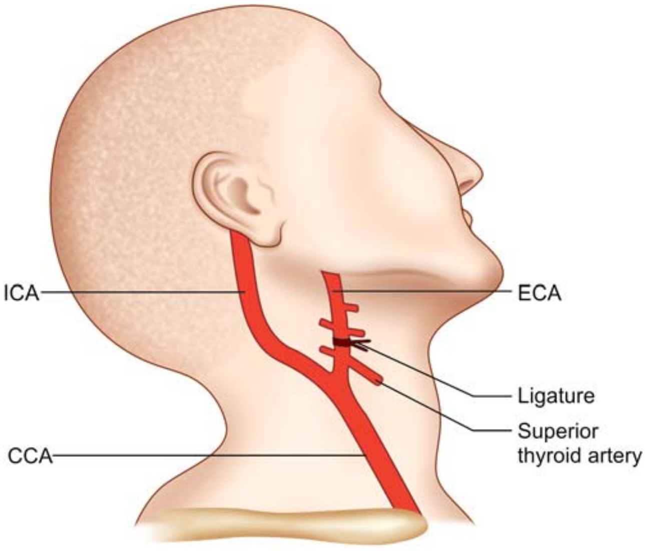

Fig. 5.57: Ligation of external carotid artery should be done when needed distal to the origin of the superior thyroid artery.

If you have the best products, you won't need much advertising.

Type of neck dissection	*Nodes removed*	*Structures preserved*
Comprehensive:		
Radical neck dissection (RND)	Level I-V	None
MRND type I	Level I-V	Spinal accessory nerve
MRND type II	Level I-V	Spinal accessory nerve; sternomastoid muscle
MRND type III	Level I-V	Spinal accessory nerve; sternomastoid muscle; IJV
Selective node dissection:		
Supraomohyoid neck dissection (N_0)	Level I-III	Spinal accessory nerve; sternomastoid muscle; IJV
Extended supraomohyoid dissection (N_0)	Level I-IV	Spinal accessory nerve; sternomastoid muscle; IJV
Anterolateral neck dissection (N_0)	Level II-IV	Spinal accessory nerve; sternomastoid muscle; IJV
Posterolateral neck dissection (N_0)	Level II-V with suboccipital, retroauricular nodes	Spinal accessory nerve; sternomastoid muscle; IJV
Anterior/central dissection	Level VI	Spinal accessory nerve; sternomastoid muscle; IJV

Note:

- To control haemorrhage in head and neck cancers and during head and neck surgeries (like radical parotidectomy, commando operation, maxillectomy) often ligation of external carotid artery is required. ***Ligation should be done distal to the origin of the superior thyroid artery***. It should never be ligated below the origin of the superior thyroid artery as this will lead to formation eddy current and thrombus at the carotid bifurcation and intracranial embolism. Ligation of ECA to control bleeding is usually done in continuity (Fig. 5.57).
- Middle of the neck, laterally over CCA is the poorly vascularised area of skin which can lead into skin necrosis. So vertical incision or three point junctions at this point should be avoided. Horizontal incisions are better in neck.
- Carotid artery should be protected by *muscle flap or free dermal graft*. *Levator scapulae* muscle flap is commonly used.

Carotid blow out is most dangerous complication. It is due to sepsis, wound breakdown, stripping of arterial adventitia, necrosis and drying of the artery. Ligation of the carotid is done to save the life of the patient but procedure itself has got 20% mortality and 50% morbidity (hemiplegia).

CHEMOTHERAPY FOR HEAD AND NECK CANCERS

It may be used alone or as multimodality therapy.

Types

- *Adjuvant chemotherapy*: Chemotherapy is used before, during or after main therapeutic modality (surgery or radiotherapy). When used before it is called as anterior/ *nduction* chemotherapy. It reduces the burden, downstages the tumour, reduces the chance of micrometastasis that may occur during surgery. When used with radiotherapy (concurrent) it is used as radio sensitiser. When used after surgery/ radiotherapy it is called as *posterior* chemotherapy.
- *Palliative chemotherapy*: It is used in advanced/ recurrent/ metastatic cancers to relieve symptoms like pain/dysphagia/ dyspnoea or to prevent chances of bleeding or fungation.

It may be single drug therapy like methotrexate/bleomycin/ cisplatin/5-fluorouracil or multidrug combination chemotherapy. Combination chemotherapy is more beneficial which increases efficacy.

- Methotrexate—40 mg/m^2 IV weekly. It causes mucositis, bone marrow suppression with liver and kidney toxicity. Hydration and alkalisation of urine before and after therapy is beneficial to reduce toxicity.
- 5 - Fluorouracil—10-15 mg/m^2 IV daily for 5 days. Complications are bone marrow suppression and gastrointestinal symptoms.
- Bleomycin—10-20 mg/m^2 IV weekly. Pneumonitis and *pulmonary fibrosis* are the complications.
- Vincristine—1-2 mg/m^2 IV monthly. *Neurotoxicity* (sensory and motor neuropathy), constipation and alopecia are the side effects.
- Cisplatin—80-120 mg/m^2 IV infusion once in 3 weeks. Complications are neurotoxicity, bone marrow suppression, renal toxicity, ototoxicity. Adequate hydration and mannitol diuresis may be needed.
- Cyclophosphamide—60-120 mg/m^2 IV for 5 days at regular 3 weeks cycles. Alopecia, bone marrow suppression and *cystitis* (haemorrhagic) are the toxicities (ABC). Barbiturates should be avoided during therapy.
- Adriamycin—60-90 mg/m^2 IV. It is *cardiotoxic* and hence cardiac monitoring is needed while infusion.
- Paclitaxel and carboplatin as empiric chemotherapy are used in neck secondaries with occult primary.

Mode of Administration

- It can be given by *intra-arterial route, through external carotid artery* (Never through internal carotid as it will cause cerebral damage). Site of arterial catheter should be confirmed by Doppler or angiogram. Drug is usually administered through an arterial pump. Other method is to increase the height of the drip stand to get a pressure above the level of the systolic pressure of the patient (i.e. more than 13 ft).

- Drugs can also be given intravenously or orally (methotrexate).

General toxicity of chemotherapeutic agents: Mucositis, alopecia, stomatitis, nausea and vomiting, diarrhoea, bone marrow suppression.

Often transfusions of blood/FFP/platelet are needed before or after chemotherapy depending on parameters.

Monitoring the Patient on Chemotherapy

- Clinical assessment—pulse, blood pressure, alopecia, urine output, jaundice, skin changes like rashes, fever, pallor.
- Biochemical parameters—total count, platelet count, blood urea and serum creatinine, liver function tests.
- Specific relevant tests—chest X-ray, endoscopy, CT scan, tumour marker if specific in some types of carcinomas.

Rule of 80 in the neck

- 80% of nonthyroid neck masses are neoplastic
- 80% of neoplastic neck masses are seen in males
- 80% of neoplastic neck masses are malignant
- 80% of malignant neck masses are metastatic
- 80% of metastatic neck masses are from primary sites above the clavicle

***Note*:**
For carotid artery aneurysm—refer Chapter on Arterial Diseases.

Chapter 6 Thyroid

> *(The thyroid gland) has no very evident mechanical or local office to fulfill…. Yet we may one day be able to show, that a particular material principle is slowly formed, and partially kept in reserve; and that this principle (serves) important subsequent functions in the course of the circulation…. Something analogous to a reservoir function obtains in this part.*
>
> —**Thomas Wilkinson King,** 1836

CHAPTER OUTLINE

- Development
- Anatomy
- Physiology
- Congenital Anomalies
- Thyroid Function Tests
- FNAC of Thyroid
- Classification of Goitre
- Diffuse Hyperplastic Goitre
- Nodular Goitre
- Discrete Thyroid Nodule
- Solitary Thyroid Nodule
- Retrosternal Goitre
- Thyrotoxicosis and Hyperthyroidism
- Radioactive Iodine
- Thyroid Neoplasms
- Differentiated Thyroid Carcinoma
- Papillary Carcinoma of Thyroid
- Follicular Carcinoma of Thyroid
- Anaplastic Carcinoma of Thyroid
- Medullary Carcinoma of Thyroid
- Malignant Lymphoma
- Hashimoto's Thyroiditis
- De-Quervain's Subacute Granulomatous Thyroiditis
- Riedel's Thyroiditis
- Thyroid Incidentaloma
- Thyroidectomy
- Emil Theodor Kocher
- Kocher's Test
- Hypothyroidism
- Recurrent Laryngeal Nerve Palsy

DEVELOPMENT

The thyroid gland develops from a median down growth of a column of cells from the pharyngeal floor between the first and second pharyngeal pouches (subsequently marked by the foramen caecum of the tongue).

The canalised column becomes the *thyroglossal duct* which is displaced forward by the developing hyoid bone and then below the hyoid, lies slightly to one side, more commonly to the left. The duct bifurcates to form the thyroid lobes and a portion of the duct forms the pyramidal lobe.

ANATOMY

It is located in the anterior triangle of the neck. It weighs about 20 grams.

Parts

- *Right and left lateral lobes* located in a space (thyroid fossa) between trachea and oesophagus medially and carotid sheath laterally. Each lobe is 5 × 3 × 1.5 cm in size, extends from the middle of thyroid cartilage to 6th tracheal ring.
- *Isthmus* is the connecting part between two lateral lobes in midline extending from 2nd to 4th tracheal rings.
- *Pyramidal lobe* is upward extension as fibrous strands or muscular strands from the junction of the isthmus and left lateral lobe.
- Gland is invested by *pretracheal fascia*.
- *Berry's ligament* is a strong condensed vascular connective tissue between the lateral lobe and cricoid cartilage on each side.

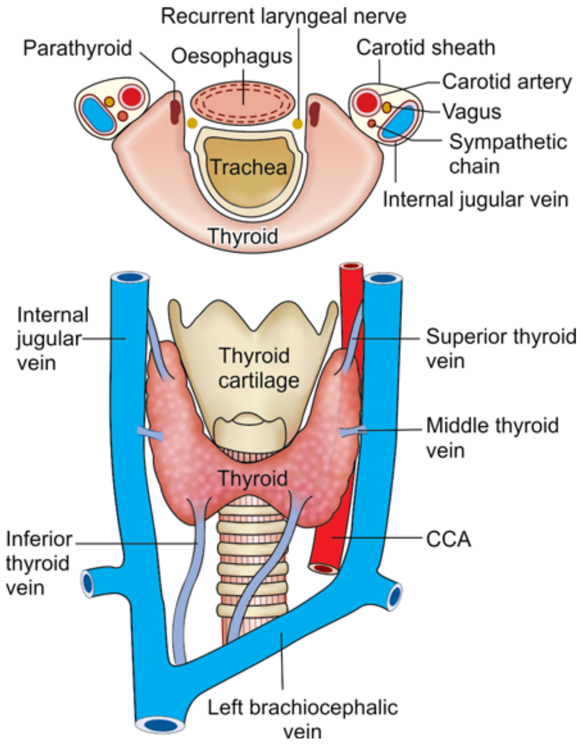
Recurrent laryngeal nerve
Parathyroid
Oesophagus
Carotid sheath
Carotid artery
Vagus
Sympathetic chain
Trachea
Internal jugular vein
Thyroid
Internal jugular vein
Thyroid cartilage
Superior thyroid vein
Middle thyroid vein
Thyroid
Inferior thyroid vein
CCA
Left brachiocephalic vein

Fig. 6.1: Relations of the thyroid.

Blood Supply

- *Superior thyroid artery* is first anterior branch of external carotid artery enters the gland near superior pole as a larger anterior superficial branch and a smaller posterior branch.
- *Inferior thyroid artery*, a branch of thyrocervical trunk of subclavian artery passes behind the carotid sheath running medially reaching the posterolateral aspect of the gland.
- *Thyroidea ima artery*, a branch of aorta or brachiocephalic artery enters the isthmus or lower pole of one of the lateral lobes (10%).
- Tracheal and oesophageal branches serve blood supply to retained thyroid gland after thyroidectomy.

Venous Drainage

- *Superior thyroid vein.*
- *Middle thyroid vein* is short and drains into the internal jugular vein. It is first to be ligated in thyroidectomy.
- *Inferior thyroid veins* are many in number.
- *Kocher's vein* may be present which drains lower or middle thyroid.

Lymphatic Drainage

- *Primary*:
  - Tracheo-oesophageal nodes.
  - Prelaryngeal nodes (*Delphian nodes*). Formerly purpose of these lymph nodes were uncertain hence the name. *Delphi* is a place in Greece where Pythia, snake women after sulphurous fume inhalation uttered meaningless jargon purpose of it was unclear.
  - Mediastinal nodes.
- *Secondary*:
  - Deep cervical nodes.
  - Supraclavicular nodes.
  - Occipital nodes.
    - Ascending medial lymph vessels from the upper border of isthmus drain to prelaryngeal nodes which are located in the cricothyroid membrane.
    - Ascending lateral vessels from the upper pole of the gland along the superior thyroid artery drain into deep cervical nodes.
    - Descending medial vessels begins at lower part of the isthmus to reach pretracheal lymph nodes.
    - Descending lateral vessels run from the deep surface of the thyroid to recurrent laryngeal chain nodes.

Important Relations of Thyroid Gland

- Recurrent laryngeal nerve lies in the tracheooesophageal groove, in relation to Berry's ligament.
- Superior laryngeal nerve which gives a branch, external laryngeal nerve supplies cricothyroid muscle. It accompanies superior thyroid artery.
- Parathyroid glands—four in number, two on each side embedded in thyroid.

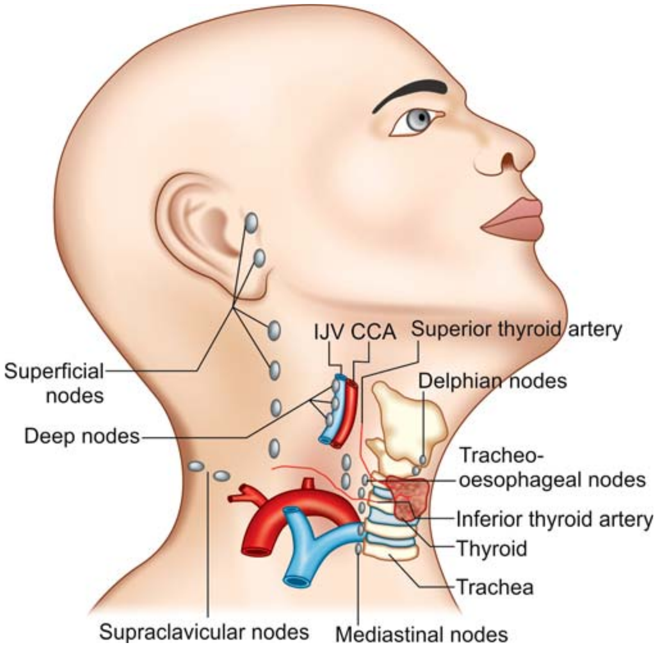
IJV CCA
Superior thyroid artery
Superficial nodes
Delphian nodes
Deep nodes
Tracheo-oesophageal nodes
Inferior thyroid artery
Thyroid
Trachea
Supraclavicular nodes
Mediastinal nodes

Fig. 6.2: Lymphatic drainage of thyroid.

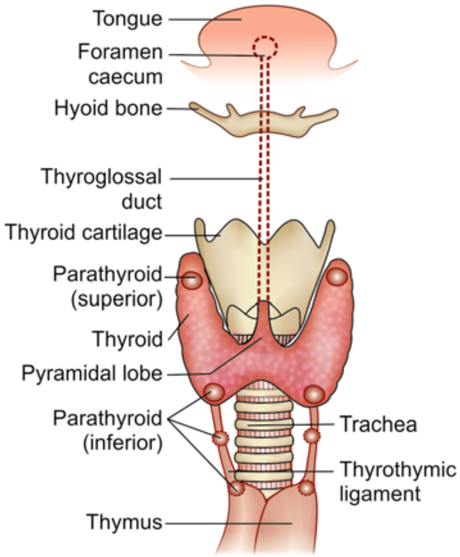

Fig. 6.3: Anatomy of thyroglossal duct showing its pathway.

PHYSIOLOGY

Thyroid gland has two secretory cells:

1. *Follicular cells*—secretes thyroid hormones (Thyroxine (T_4), Tri-iodothyronine (T_3).
2. *Parafollicular cells* (‘C’ cells)—secretes calcitonin.

I_2 Metabolism

Ninety per cent of body iodide uptake is in the thyroid gland, whose uptake into the follicular cells is regulated by TSH and follicular iodide content.

Thyroid Hormone Synthesis

Iodothyronins (MIT, DIT) are formed in follicular cells by the coupling of inorganic iodide with tyrosine. These are biologically inert molecules. T_4 (Thyroxine) is formed by coupling of 2 DIT molecules and T_3 (Tri-iodothyronine) is formed by coupling of 1 MIT and 1 DIT molecules. Both are bound to thyroglobin which is the primary component of colloid matrix.

The hypothalamus-pituitary-thyroid axis regulates thyroid hormone production and releases in a classic feedback system. TRH is a regulatory hormone from hypothalamus and TSH is a regulatory hormone from anterior pituitary.

CONGENITAL ANOMALIES

Athyreosis is total absence of lateral lobes and isthmus with hypothyroidism.

Ectopic Thyroid

It is residual thyroid tissue along the course of thyroglossal tract.

Ectopic thyroid tissue may lie anywhere along the line of descent. Whole of the thyroid gland or residual thyroid lies in an abnormal position either in the posterior part of the tongue, or in the upper part of the neck in midline, or intrathoracic region. Carcinoma develops more commonly in ectopic thyroid tissue than normal thyroid. Radioisotope scan, CT scan for intrathoracic thyroid will confirm the diagnosis.

Lingual Thyroid

It is a thyroid swelling in the posterior third of tongue, at the foramen caecum, presenting as rounded swelling. It may be the only existing thyroid tissue which may cause:

a. Dysphagia, pain, speech impairment.
b. Respiratory obstruction, haemorrhage.
c. 70% present as hypothyroidism, 10% as cretin.
d. Common in females (3:1).

Bare tracheal rings in midline with absence of normal thyroid tissue may often be evident.

Any diseases which can occur in normal thyroid can also occur in lingual thyroid, i.e. nodularity, toxicity, malignancy (If turns malignant, *follicular carcinoma* is more common, papillary carcinoma is very rare).

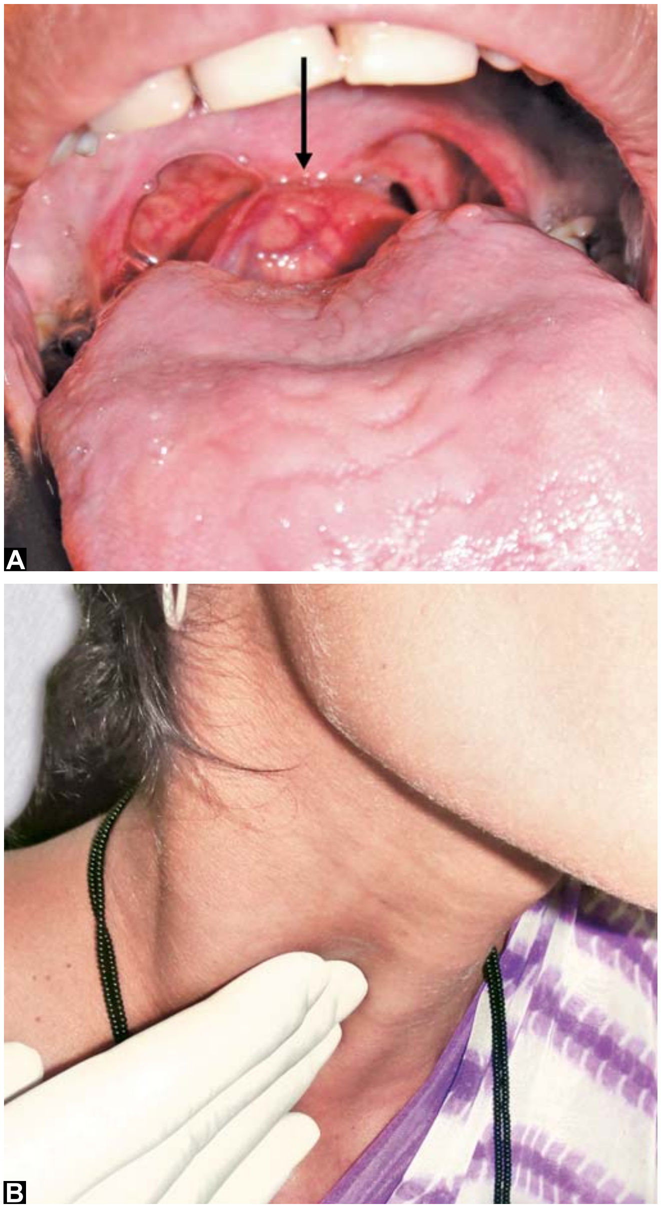

Figs 6.4A and B: Lingual thyroid visible on the posterior 1/3rd of the tongue. Neck should be palpated for the presence or absence of normal thyroid.

Differential diagnosis for lingual thyroid

- Carcinoma of posterior third of tongue
- Angiofibroma
- Sarcoma
- Hypertrophied lingual tonsil

Diagnosis

- *Radioisotope study* shows the uptake of iodine by the lingual thyroid and also shows the status of the thyroid in normal fossa.
- U/S neck has to be done to see the absence of thyroid in normal location.

Treatment

- L-thyroxine is given daily orally.
- Often requires surgical excision and is technically easier.
- Radioisotope therapy for ablation is also often used.

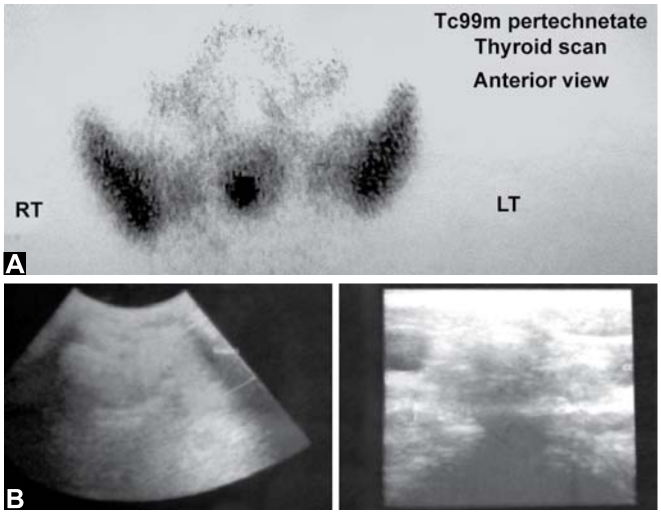
Tc99m pertechnetate
Thyroid scan
Anterior view
RT
LT
A
B

Figs 6.5A and B: Technetium 99 radioisotope scan in lingual thyroid. USG of neck is done to assess lingual thyroid and normal thyroid region.

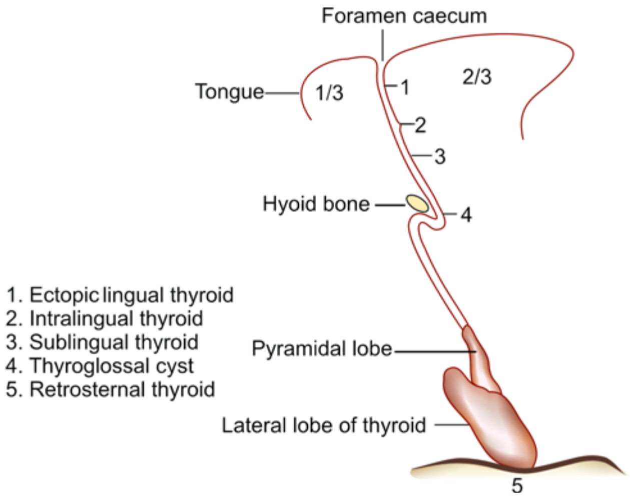
Foramen caecum
Tongue
1/3
2/3
1
2
3
Hyoid bone
4
1. Ectopic lingual thyroid
2. Intralingual thyroid
3. Sublingual thyroid
4. Thyroglossal cyst
5. Retrosternal thyroid
Pyramidal lobe
Lateral lobe of thyroid
5

Fig. 6.6: Ectopic sites of thyroid.

Thyroglossal Cyst

- Thyroglossal cyst is a swelling occurring in the neck in any part along the line of thyroglossal tract.

Possible sites for thyroglossal cyst

a. Beneath the foramen caecum
b. In the floor of mouth
c. Suprahyoid
d. Subhyoid—commonest site
e. On the thyroid cartilage—2nd common site

- It is usually congenital wherein there will be degeneration of a part of the tract causing cystic swelling. Normal thyroid may be present in the normal location (fossa). Sometimes, thyroid may not be present in the normal site but may be present in the wall of the thyroglossal cyst.
- It is a tubulodermoid type of cyst.
- It is lined by pseudostratified, ciliated columnar epithelium.

Clinical Features

a. Swelling in the midline, towards the left.

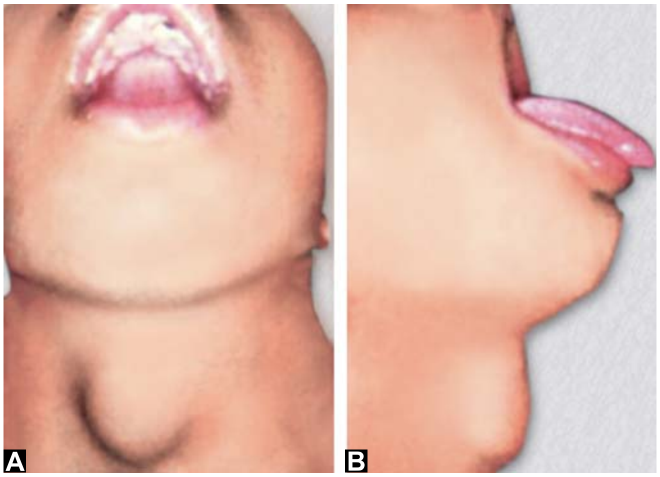

Figs 6.7A and B: Thyroglossal cyst.

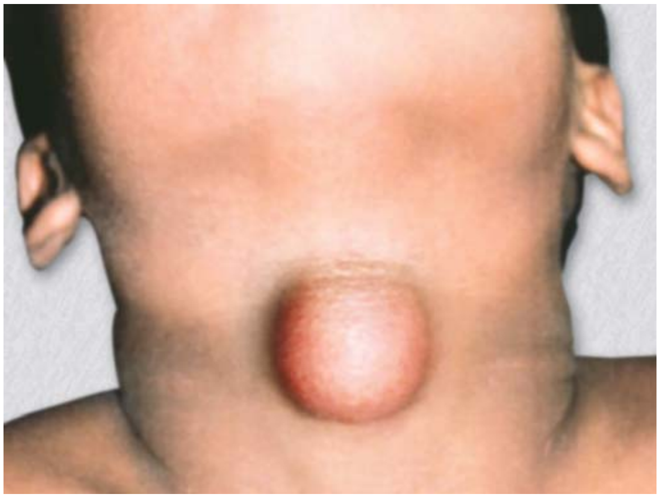

Fig. 6.8: Thyroglossal cyst in a boy. It is a midline swelling with a 'TUG' feeling while protruding the tongue out.

The person who really wants to do something finds a way: the other finds an excuse.

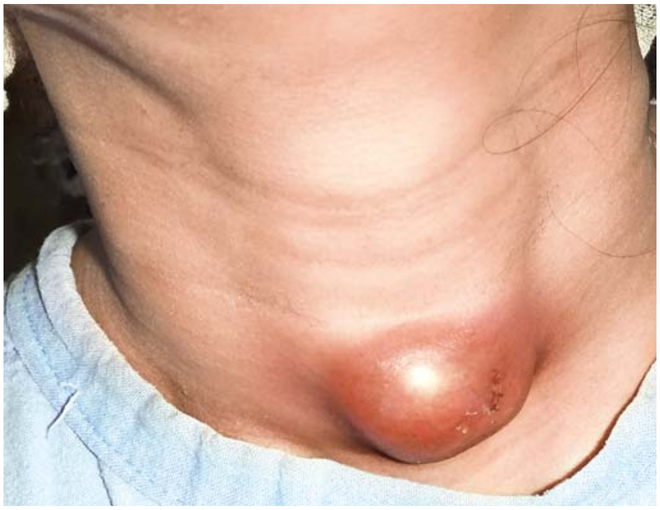

Fig. 6.9: Infected thyroglossal cyst.

b. Moves with deglutition as well as with the protrusion of tongue. Patient is asked to open the mouth and keep the lower jaw still. Examiner holds the cyst between the thumb and forefinger. When the patient is asked to protrude the tongue, a "*tugging sensation"* can be felt.
c. Swelling is smooth, soft, fluctuant (cystic), nontender, mobile, often transilluminant.
d. Thyroid fossa is empty, if there is no thyroid in normal location.
e. Thyroglossal cyst can get infected and may form an abscess. Cyst wall contains lymphatic tissue and so infection is common.
f. Malignancy can develop in thyroglossal cyst (papillary carcinoma)—1%. Cyst will be harder, fixed, with palpable neck nodes. More often it is difficult to suspect and will be confirmed by histology after excision. If biopsy report is papillary carcinoma, then completion thyroidectomy is *indicated* if the remaining is nodular thyroid shows cold nodule or these are enlarged neck nodes, or with history of neck irradiation. If neck nodes are present, node dissection, and radioactive iodine therapy with suppressive dose of L thyroxine 0.3 mg OD is needed.
g. Incidence is equal in both sexes.

Investigations

a. Radioisotope study.
b. U/S neck.
c. FNAC from the cyst.

Differential diagnosis for thyroglossal cyst

- Subhyoid bursa
- Pretracheal lymph node
- Dermoid cyst
- Solitary nodule of thyroid—isthmus
- Submental lymph node
- collar stud abscess

Treatment

a. ***Sistrunk operation:***
Excision of cyst and also full tract upto the foramen caecum is done along with removal of central part of the hyoid bone, as the tract passes through it.

Technique

Through transverse neck incision placed over the cyst, skin flap is raised above along with platysma. Care should be taken not to open the cyst. Cyst with surrounding tissues is dissected up to the hyoid bone. Sternohyoid and thyrohyoid muscles are divided. Central part of the hyoid bone of 1 cm width is resected along with intact track within it. Geniohyoid and mylohyoid muscles are divided off from the hyoid. Track with adjacent tissues is dissected above upto the foramen caecum. Adjacent tissues also should be removed because of possibility of multiple tracks which otherwise lead to recurrence or fistula formation. Often anaesthetist is asked to apply digital pressure over the base of tongue near foramen caecum to facilitate the dissection and to confirm the reach up to the foramen caecum. Track is ligated at foramen caecum and removed. Recurrence rate by removal of only track without central hyoid is 25%. Removal of track with central hyoid o*f 1 cm width* reduces the recurrence rate to 5%. Complications are recurrence, thyroglossal fistula formation, haemorrhage/haematoma formation and infection. Low lying cyst often requires two parallel incisions to remove entire track up to foramen caecum.

b. If there is no normal thyroid gland after the surgery, maintenance dose of L-thyroxine 0.1 mg OD is given life long.

***Note*:**
If tract is not completely excised, it will result in thyroglossal fistula.

Thyroglossal Fistula

- *It is not a congenital condition.*
- It either follows infection of thyroglossal cyst which bursts open or after inadequate removal of the cyst.
- It is lined by columnar epithelium, discharges mucous and is a seat of recurrent inflammation. "*Hood sign*" is characteristic. Opening of fistula is indrawn and been overlaid by a fold of skin as '*hood*'.
- Peculiar crescentic appearance is called as *semilunar* sign.
- It secretes mucous discharge.
- Site of the fistula is just below the hyoid bone commonly; in infants it may be much lower.

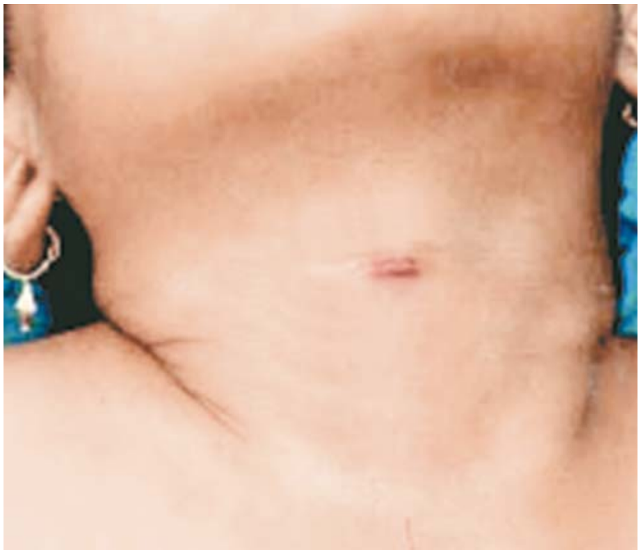

Fig. 6.10: Thyroglossal fistula.

- **Investigations:** Radioisotope study and fistulogram.
- **Treatment:** Sistrunk operation.

Note:
(One more Sistrunk operation is done in case of lymphoedema.)

Lateral Aberrant Thyroid

- It is at present considered as a misnomer.
- It is the metastasis into cervical lymph node from a papillary carcinoma of thyroid.
- FNAC has to be done and treated as papillary carcinoma of thyroid.

Agenesis

- Total agenesis of one thyroid lobe may occur. This is rare but can be clinically important. It leads to confusion in diagnosis, especially in the toxic gland, where it could be diagnosed as a secreting nodule.

Dyshormonogenesis

- It is an autosomal recessive condition wherein there is either deficiency of thyroid enzymes (either peroxidase or dehalogenase) or inability to concentrate or to bind or to retain iodine.
- It may be familial and patient presents with large diffuse vascular goitre involving both lobes.
- They respond very well to L-thyroxine and may not require surgery at any time.
- Condition may be associated with congenital deafness which is being called as *Pendred's syndrome*. CT scan of temporal bone shows abnormal bony labyrinth. Condition shows thyroid peroxidase deficiency.

THYROID FUNCTION TESTS

1. *T_3* (Serum tri-iodothyronine): 1.2-3.1 nmol/litre. 80% of T_3 is from deiodination of T_4 at periphery in liver, muscle, kidney and pituitary. T_3 is 4 times more potent with half life of 24 hours.
2. *T_4* (Serum thyroxin): 55-150 nmol/litre. It reflects the output of the gland (Edward Calvin Kendall 1915).
3. TSH 0-5 IU/ml of plasma. TSH is secreted from anterior pituitary; its secretion is inversely related to circulating thyroid hormones. TSH secretion is regulated by TRH from hypothalamus.
4. *PBI* (Protein bound iodide)—8 ug/100 ml.
5. *Free* T_3 is 0.3% (3-9 nmol/litre). It is the best single test in assessing hyperthyroidism.
6. *Free* T_4 is 0.03% (8-26 nmol/litre).
7. *RA I^{123} scan* can show either cold nodule, hot nodule, or warm nodule.
8. *TRH stimulation test* for hypothalamic—pituitary axis: Intravenous TRH (200 μg) shows rise in serum TSH level in 20 minutes (from basal 1 μ unit/ml to 10 μ unit/ml) and reaches to normal in 2 hours. Patients with pituitary insufficiency develop a subnormal response; patient with hypothyroidism will show enhanced TSH response; in hyperthyroidism there will be no response. This test is useful in doubtful hyperthyroidism, hypothyroidism, T_3 thyrotoxicosis, ophthalmic Graves' disease.
9. Serum creatinine is increased in hyperthyroidism; decreased in hypothyroidism.
10. Serum cholesterol is increased in hypothyroidism and decreased in hyperthyroidism.
11. BMR is increased in hyperthyroidism.
12. *Thyroid autoantibodies* are also useful to evaluate the function (LATS). TSH receptor antibodies (TSH RAb) has got long acting potential and are reasons for all primary thyrotoxicosis.
13. *Werner's T_3 suppression test*: Initial isotope uptake study is done. 40 μg of T_3 is given to the patient orally 8th hourly for 5 days. Uptake study is repeated. In normal uptake suppression up to 80% is noted. In toxic goitre suppression is 10 - 20%. It is used in patients with antithyroid drugs for primary thyrotoxicosis to assess the remission status.
14. *Thyroglubulin estimation*: Normal value is 0.5-50 μg/L. It is used during follow up period especially in follicular carcinoma of thyroid. But if thyroglobulin antibodies are raised in patient, then it is of no use. Thyroglobulin is produced only by the thyroid tissue. After total thyroidectomy, its level drastically reduces. It is the ideal follow up marker in well differentiated thyroid carcinoma after thyroidectomy. Sudden raise in its serum can occur in thyroiditis, primary or secondary toxic goiter.

FNAC OF THYROID

- It is the investigation of choice in most the thyroid diseases to conclude pathological diagnosis.
- It is useful in papillary/medullary (amyloid)/anaplastic carcinomas, lymphomas, colloid nodule, thyroiditis.
- 23G needle is used. Suspicious solitary/multiple nodules/dominant nodules should be aspirated.
- *Karolinska hospital* (Lowhagen) at Sweden pioneered this method.
- Minimum 6 aspirations should be done. An adequate FNAC smear should have six aspirations with six groups of cells with each group containing 20 cells. USG guided aspiration is better.
- Diagnostic accuracy of FNAC is 95%; sensitivity 85%; specificity 94%.
- Aspiration is *graded* as: Thy 1—Nondiagnostic; Thy 2—Nonneoplastic; Thy 3—Follicular; Thy 4—Suspicious of malignancy; Thy 5—malignancy.
- In a cyst of thyroid FNAC may be less reliable; if cyst recurs after 3 aspirations, surgery is needed.
- Malignancy rate in a simple cyst is 5%; in a complex cyst it is 75%.
- FNAC is *not reliable* at present in follicular carcinoma of thyroid as capsular and vascular invasions cannot be found. But by newer technique it is possible to identify the differences—Benign is polyploidy, malignant is aneuploidy; benign are monoclonal, malignant are polyclonal; MR spectroscopy and thyroimmunoperoxidase estimation are useful to differentiate.
- FNNAC is fine needle nonaspiration cytology—is said to be more reliable.

> *In the case of retrosternal goitre, hemoptysis unmixed with sputum is due to rupture of an engorged tracheal vein.*
> **—Peter Burgess**

USG in thyroid

- To identify nodules, number, size, vascularity, echogenicity
- To do USG guided FNAC
- To identify neck lymph nodes
- To find out solid or cystic nature
- Benign lesion is *hyperechoic*, often cystic with well-defined margin; shows peripheral *egg shell calcification* with sonolucent rim (halo) around nodule
- Malignant lesion is *hypoechoic* with poorly defined margin, with high vascularity, with microcalcification without any halo around

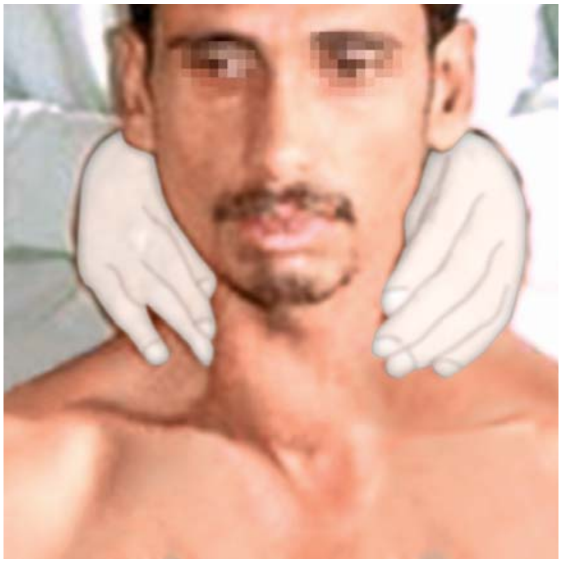

Fig. 6.11: Examination of the thyroid gland from behind for any nodules. Thyroid enlargement in male is rare. Nodule in a male could be malignant.

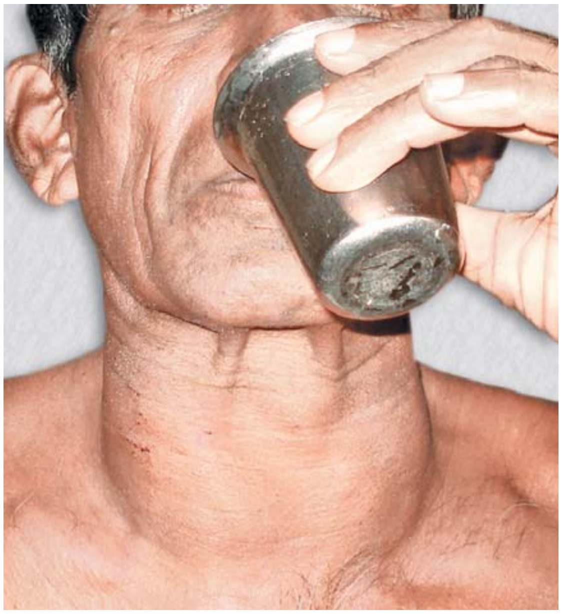

Fig. 6.12: Thyroid swelling moves on deglutition because of attachment of pretracheal fascia through Berry's ligament to cricoid cartilage and to constrictor muscles.

CLASSIFICATION OF GOITRE

Goitre is enlargement of thyroid gland. (*'goiter'*-*latin*-guttur-throat)

1. *Simple nontoxic*
   a. Diffuse hyperplastic:
      – Physiological
         - Puberty.
         - Pregnancy.
      – Primary iodine deficiency (Endemic; dietary iodine intake less than 100 μg/day).
      – Secondary iodine deficiency:
         - Goitrogens of Brassica family, e.g. cabbage, soya bean. Common in hill stations.
         - Excess dietary fluoride.
         - Drugs: PAS, lithium, phenylbutazone, thiocyanates, potassium perchlorate, antithyroid drugs, radioactive iodine.
         - Dyshormonogenetic goitre.
   b. Colloid goitre.
   c. Nodular goitre (Multinodular).
   d. Solitary nontoxic nodule.
   e. Recurrent nontoxic nodule.
2. *Toxic*
   a. Diffuse (Primary)—*Graves' disease.*
   b. Multinodular (Secondary)—*Plummer's disease*.
   c. Toxic nodule (solitary) (Tertiary).
   d. Recurrent toxicosis.
3. *Neoplastic*
   a. Benign—adenomas: follicular, Hurthle cell.
   b. Malignant:
      – Carcinomas: Papillary, follicular, medullary, anaplastic.
      – Lymphomas.
4. *Thyroiditis*
   a. Hashimoto's autoimmune thyroiditis.
   b. de-Quervain's autoimmune thyroiditis.
   c. Riedel's thyroiditis.
5. *Rare causes:* Bacterial (suppurative), amyloid.

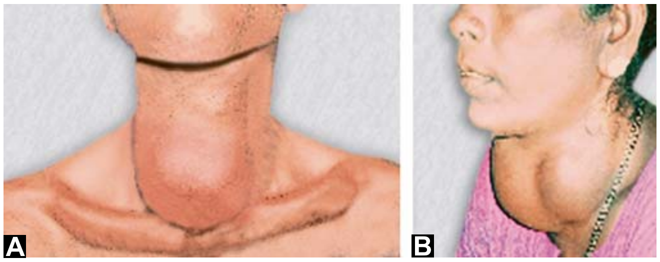

Figs 6.13A and B: Note the goitres in (A) male and (B) female patient.

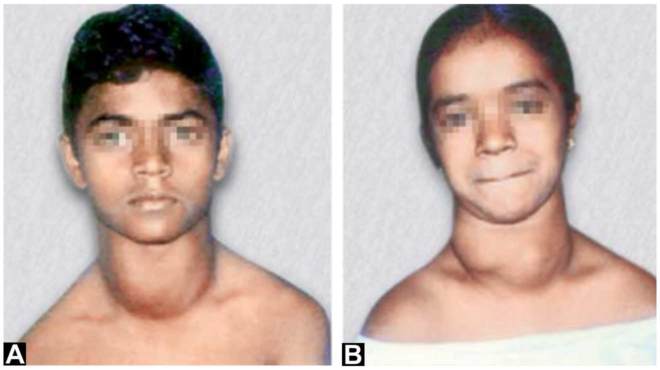

Figs 6.14A and B: Familial goitre. Goitres in brother and sister.

WHO grading of goitre

Grade 0: No visible or no palpable goitre.
Grade 1: Palpable thyroid/goitre but not visible in normal positioned neck.
Grade 2: goitre which is visible in normal positioned neck.

In a case of thyroid disease following things should be made very clear

- Functional status—hyperthyroid/euthyroid/ hypothyroid
- Compression on trachea/recurrent nerve
- Neck lymph nodal status
- Tracheal deviation
- Carotid infiltration
- Retrosternal extension
- Systemic features like toxicity or malignant spread to different organs like bone/liver/lungs

DIFFUSE HYPERPLASTIC GOITRE

Initial persistent increase in TSH level causes diffuse active lobules. In late stages of diffuse hyperplasia,TSH stimulation decreases and many follicles become inactive, get filled with colloid and it is called *as colloid goitre*.

*As diffuse hyperplastic goitre is a reversible stage, L-thyroxine is beneficial***.**

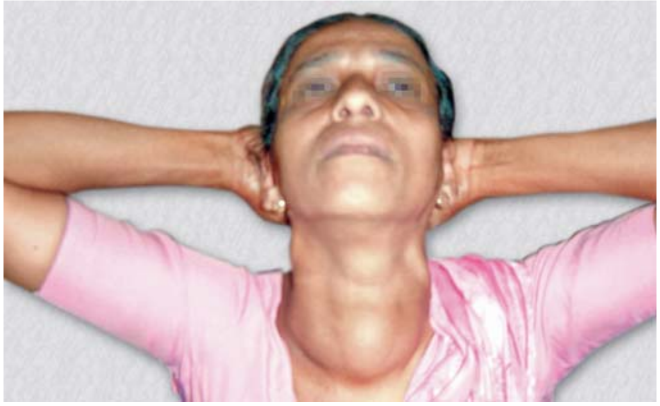

Fig. 6.15: *Pizzillos method* to make the thyroid more prominent while examining the thyroid.

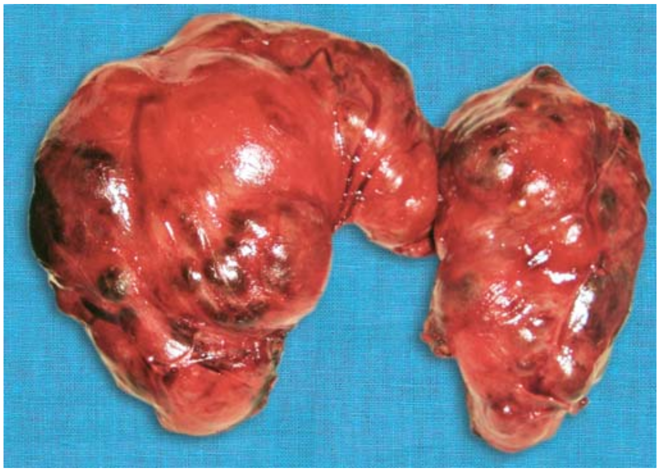

Fig. 6.16: Multinodular goitre surgical specimen—total thyroidectomy was done.

MULTINODULAR GOITRE (MNG)

- MNG is discordant growth with functionally and structurally altered thyroid follicles presenting as multiple nodules in thyroid.
- It may be due to mainly *fluctuation in TSH level*; other causes may be iodine deficiency, goitrogens, hereditary, dyshormonogenesis.

Pathogenesis

Persistent TSH stimulation.
↓
Diffuse hyperplasia of gland (all active lobules).
↓
Later with fluctuation of TSH level.
↓
Mixed areas of active and inactive lobules develop. It is also probably due to *increased sensitivity* of follicular cells to TSH.
↓
Active lobules become more vascular and hyperplastic.
↓
Haemorrhages occur with necrosis in the centre.
↓
Nodule formation.
↓
Centre of the nodule is inactive and only margin is active, i.e. internodular tissue is active.
↓
Formation of many nodules.
↓
Multinodular goitre (MNG).

Other factors involved are growth stimulating immunoglobulins and growth prone cell clones.

Stages of multinodular goitre formation

- Stage of hyperplasia and hypertrophy
- Stage of fluctuation in TSH
- Stage of formation of nodules (inactive); (inter-nodular tissues are active)

Colloid goitre is a goitre due to long standing iodine deficiency with localised accumulation of significant colloid in the gland.

Clinical Features

- More common in middle aged females (10:1).
- It is a slowly progressive disease with many years of history.
- Multiple nodules of different sizes are formed in both lobes, also in isthmus, which is firm, nodular, nontender, moves with deglutition.

Some people speak from experience; others, from experience, don't speak.

- Recent increase in size signifies malignant transformation or haemorrhage.
- Positive Kocher's test is due to compression over trachea (tracheomalacia/scabbard trachea) in a long standing MNG.
- Nodule when calcified becomes harder; necrosis softens the nodule.

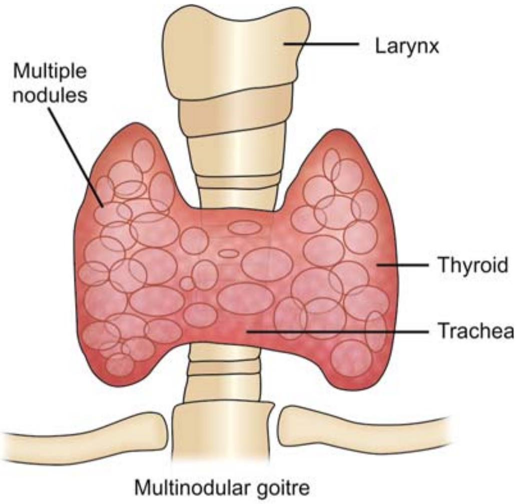

Fig. 6.17: Multinodular goitre.

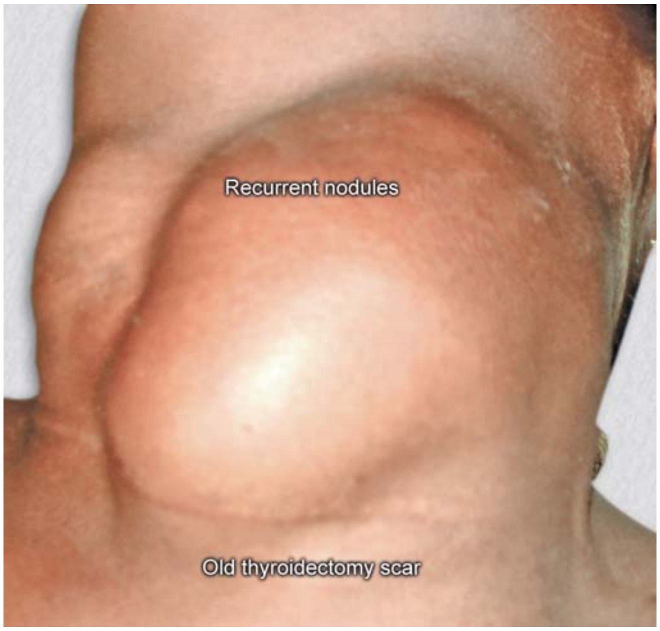

Fig. 6.18: Recurrent nodule thyroid. Note scar of previous surgery.

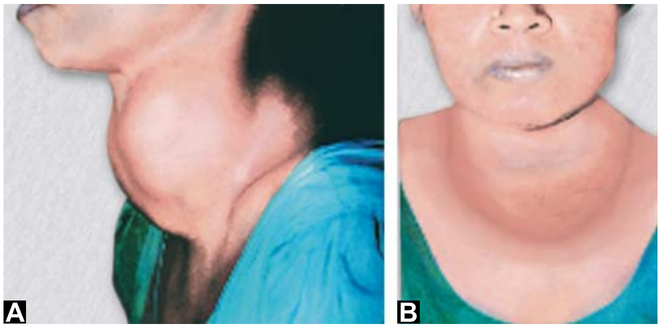

Figs 6.19A and B: Multinodular goitre.

Note:

Thyropharyngeus and cricopharyngeus parts of inferior constrictor muscle are attached to thyroid and cricoid cartilages respectively; during swallowing these muscles will contract to move thyroid and cricoid cartilages upwards. Thyroid is attached to larynx (cricoid) through condensed pretracheal fascia, *Berry's* ligament. So thyroid moves *upwards* with deglutition.

Complications of MNG

- Secondary thyrotoxicosis (30%)
- Follicular carcinoma of thyroid (10%)
- Haemorrhage in a nodule
- Tracheal obstruction, calcification
- Cosmetic problem

Investigations

- T_3, T_4, TSH, U/S neck, FNAC. FNAC is done from most *dominant* and *suspicious* nodule. FNAC from more than one

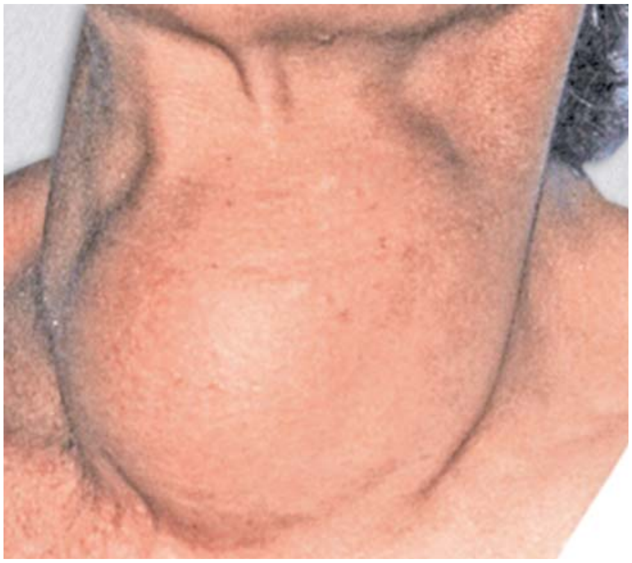

Fig. 6.20: Goitre neck, left sided.

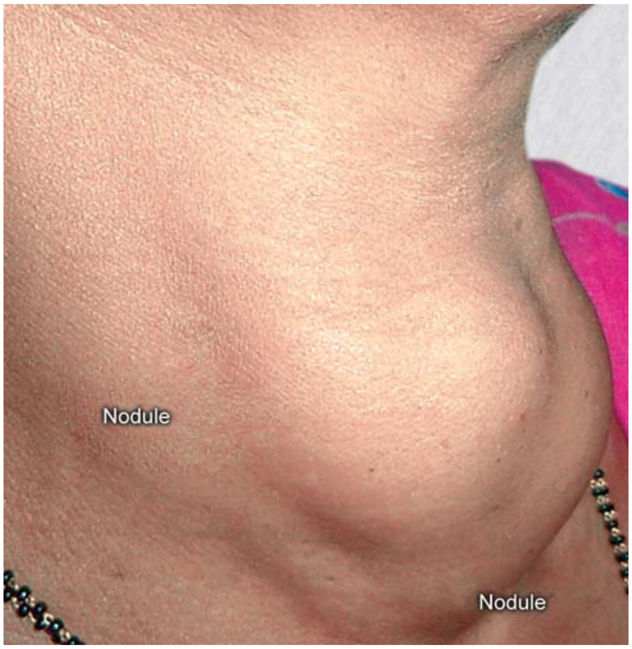

Fig. 6.21: Multinodular goitre. Note the nodules marked.

nodule is better; US guided FNAC is more reliable. High resolution US identifies impalpable nodules, number, nature of nodule, vascularity of nodule. Nodule less than 0.3 cm is identifiable in US.

- X-ray neck shows ring or rim calcification; also reveals the position (displacement) and compression of trachea.
- Indirect laryngoscopy to see vocal cords prior to surgery (This is mainly for documentation and legal purpose as even in individual with normal voice, one of the vocal cords may have been paralyzed by viral infection like mumps, probably during childhood and have compensated)—occult *recurrent laryngeal nerve* (RLN) palsy.
- Radioisotope iodine scan—in selected patients when indicated only.
- Routine blood investigations, serum calcium.
- CT scan/MRI are routinely not indicated. It is done in retrosternal extension.

Treatment

Usually surgery is preferred. Reason for doing surgery in nodular goitre is—it is an irreversible stage and chances of complications like development of toxicity, haemorrhage and follicular carcinoma is high and also for cosmetic reason.

- When entire gland is diseased *total thyroidectomy* is a better option.
- *Subtotal thyroidectomy* is done depending on the amount of gland involved, amount of normal gland existing and location of nodules—commonly done procedure in multinodular goitre. Eight grams of thyroid tissue is retained in each lateral lobe.
- Often partial thyroidectomy or *Hartley Dunhill operation* (isthmus + one entire lateral lobe and opposite side subtotal or partial) is also done depending on the amount of diseased gland and normal tissues behind. Partial thyroidectomy is not well approved now.
- Postoperative L thyroxine is often given to prevent any fluctuation in TSH level which may cause recurrent nodule formation.

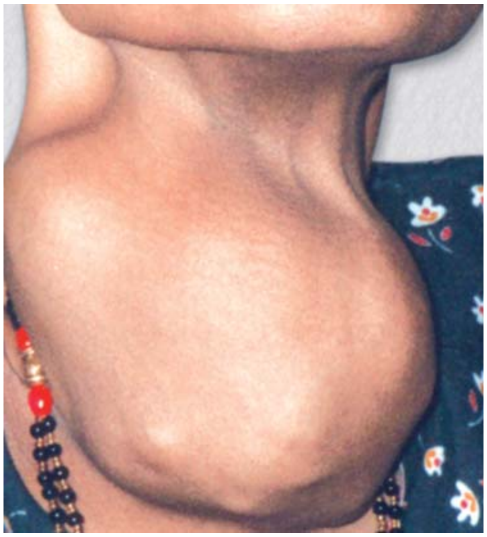

Fig. 6.22: Large multinodular goitre.

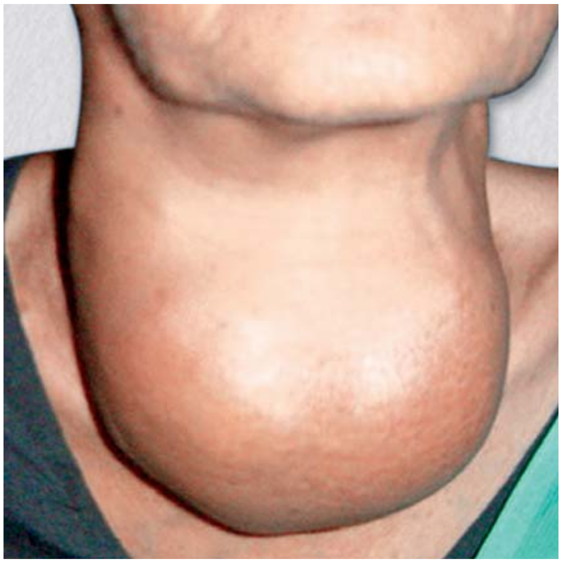

Fig. 6.23: Nodular thyroid involving both lobes in a female.

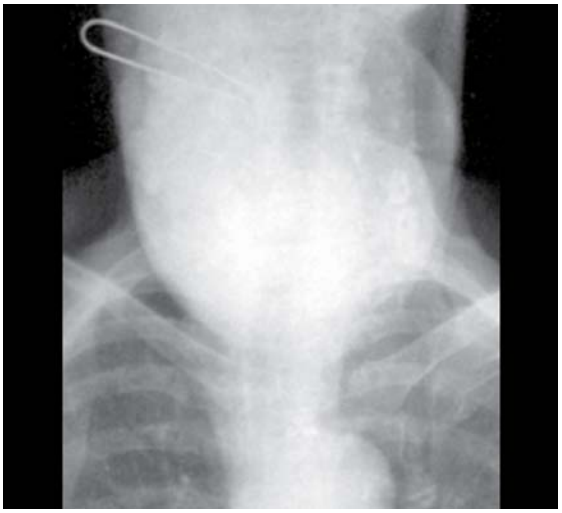

Fig. 6.24: X-ray neck showing calcification in thyroid. *Fine* calcification is seen in papillary carcinoma. *Coarse* calcification is seen in multinodular goitre.

- Prevention of multinodular goitre is possible by supplementing with L-thyroxine (0.1-0.2 mg) when patient develops goitre in puberty. Formation of nodular goitre can be prevented by correcting iodine deficiency by using iodine rich diet like eggs/seafood/milk or iodized salts and also avoiding goitrogenic drugs and diet.
- Suppressive dose of L-thyroxine alone may be used occasionally in small nodules with limited results. TSH level should be suppressed consistently below 0.5 mIU/L. Problems are need of periodic monitoring with TSH estimation; hormone insensitive part of thyroid tissue continue to grow; indefinite period of treatment; high recurrence after stopping L-thyroxine. So it is not ideally accepted therapy. It is found that TSH suppression is of no use in treating residual/recurrent MNG.

A genius is someone who shoots at a target which no one else sees and hits it.

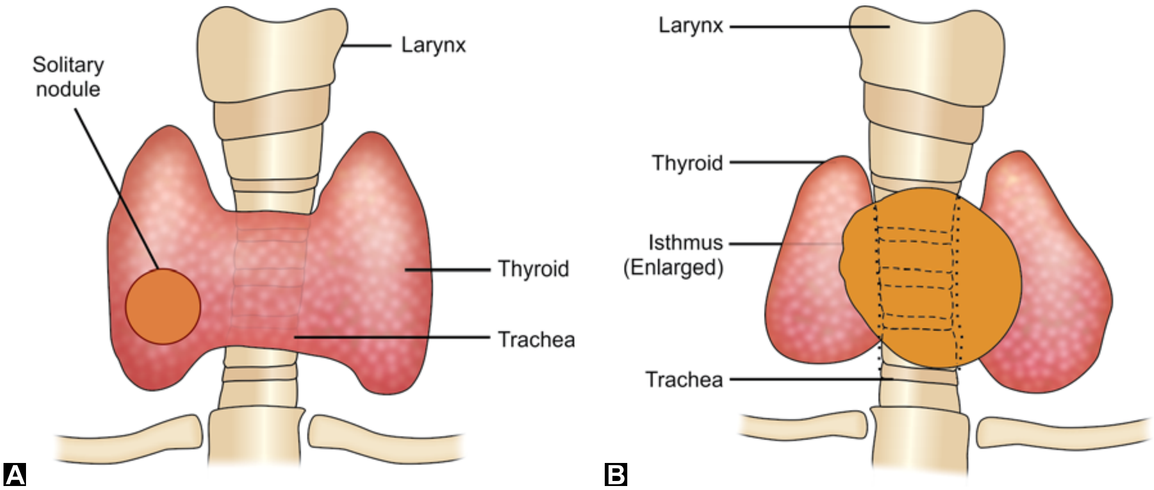

Figs 6.25A and B: Solitary nodule thyroid in one lateral lobe and isthmus.

Note:
Adequate thyroidectomy is essential, either total or subtotal, complete removal of diseased thyroid without retaining is considered, in view of chances of high recurrence in 10 or more years due to incomplete removal. Re-surgery for recurrent MNG is difficult due to loss of anatomical plane and fibrosis; hence total thyroidectomy is often advised.

DISCRETE THYROID NODULE

- ***Discrete thyroid nodule*** (4% of adult population) is *a* clearly palpable nodule in thyroid. It can be ***solitary (isolated) nodule*** (70%) if clinically only *a* nodule is felt without palpable remaining gland *OR* it can be ***dominant nodule*** (30%) if clinically *a* nodule is felt in a palpable remaining one or both lateral thyroid lobes.
- A *discrete thyroid nodule* is common in females. Discrete nodule may be solid or cystic. USG, CT, MRI may confirm the diagnosis.
- 15% solitary/isolated nodule may be malignant; 40% may be follicular adenoma. Other causes are thyroid cyst, thyroiditis or colloid degeneration.
- Incidence of malignancy in *dominant nodule* is 50% less than that of solitary nodule.
- Risk factors for malignancy in *discrete thyroid nodule* are—(1) ***solitary*** is 2 times more risk than dominant; (2) ***male*** is 4 times more than female (in solitary—48% to 12% in solid, 24% to 6% in cystic; in dominant 24% to 6% in solid, 12% to 6% in cystic); (3) ***solid nodule*** is 2 times more risk than cystic—Rule of 12.

SOLITARY THYROID NODULE

It is a single palpable nodule in thyroid on *clinical examination*, in an otherwise normal gland. *Rest of the gland is impalpable*.

Causes

1. Thyroid adenomas—20%.
    a. Follicular—colloid (commonest); embryonal; fetal.
    b. Hurthle cell.

Note:
Papillary adenoma earlier called, was actually papillary carcinoma).

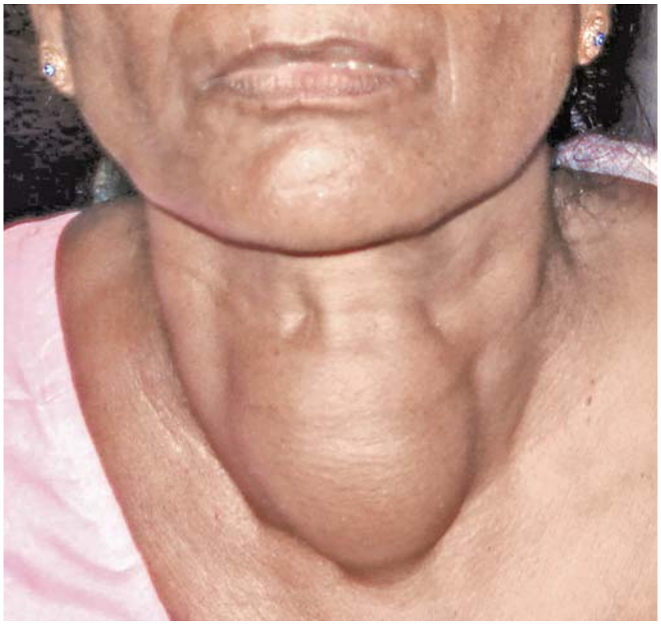

Fig. 6.26: Solitary nodule thyroid involving isthmus.

2. Papillary carcinoma of thyroid—20%
3. Only one nodule may be clinically palpable in an underlying multinodular goitre—50%.
4. Thyroid cyst—10%.

Types

1. Toxic solitary nodule—3-5% of solitary nodules of thyroid
2. Nontoxic solitary nodule.
   Based on radioisotope study:
    - *Hot*: Means autonomous toxic nodule. Normal surrounding thyroid tissue is inactive and so will not take up isotope. Nodule is overactive. It is 5% common of which only 5% can be malignant.
    - *Warm*: Normally functioning nodule. Nodule and surrounding normal thyroid will take up the isotope (active). It is 10% common of which 10% can be malignant.
    - *Cold*: Nonfunctioning nodule; may be malignant (need not be always). Nodule will not take up isotope (underactive). It is 80% common of which 20% are malignant.

Features

- Single nodule palpable in one or other lobes of the thyroid which is usually smooth and firm.
- Lahey's test does not show any other nodules in posterior part of the gland.
- *Hot or warm in* ^{99m}Tc *scan but cold in* I^{123} *scan (discordant nodule)* commonly they are malignant.
- Thyroid nodule in children and elderly can be malignant.
- Rapid enlargement of thyroid nodule can be malignant.
- Tracheal deviation towards opposite side is common– confirmed by *trail sign*, three-finger test, auscultation and X-ray neck.
- 30% of solitary nodules are cystic.
- 20% of cold nodules are malignant. Cold nodule may be due to malignancy, thyroiditis, thyroid cyst or haemorrhage, benign adenoma.
- Commonest site of a nodule is at the junction of isthmus with one of the lateral lobes.
- Solitary thyroid nodule is the most common thyroid surgical disease.

Possible features of suspected malignancy in solitary nodule thyroid

- Any nodule can be malignant whether it is hard/ firm/cystic/ small/large/asymptomatic
- Rapid onset/rapid recent increase in size
- Hoarseness of voice/dysphagia/stridor/dyspnoea
- Fixity of the nodule
- Palpable significant neck nodes
- Nodule in a male patient
- Nodule in a child (50% are malignant)
- Nodule in extremes of age group

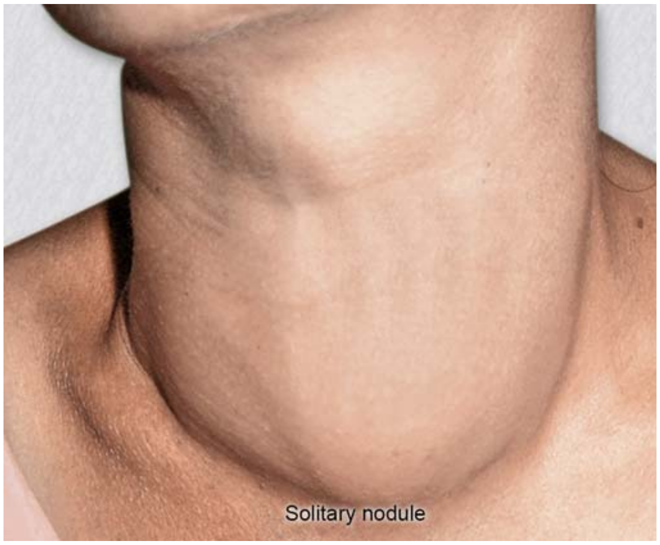

Fig. 6.27: Solitary nodule thyroid, left-sided.

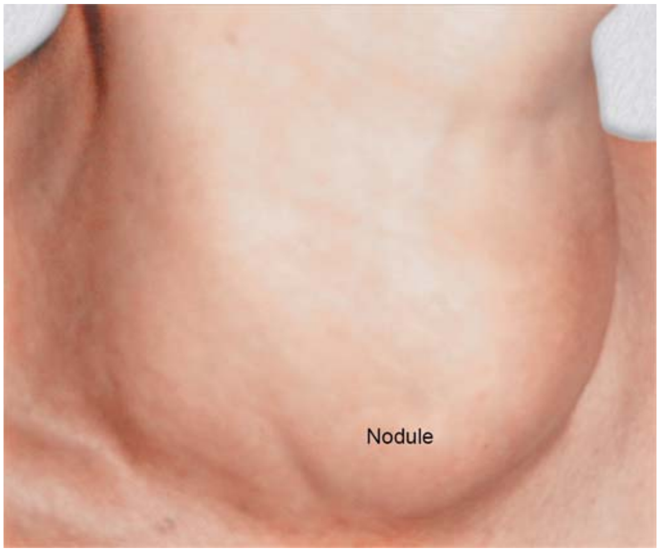

Fig. 6.28: Solitary nodule thyroid.

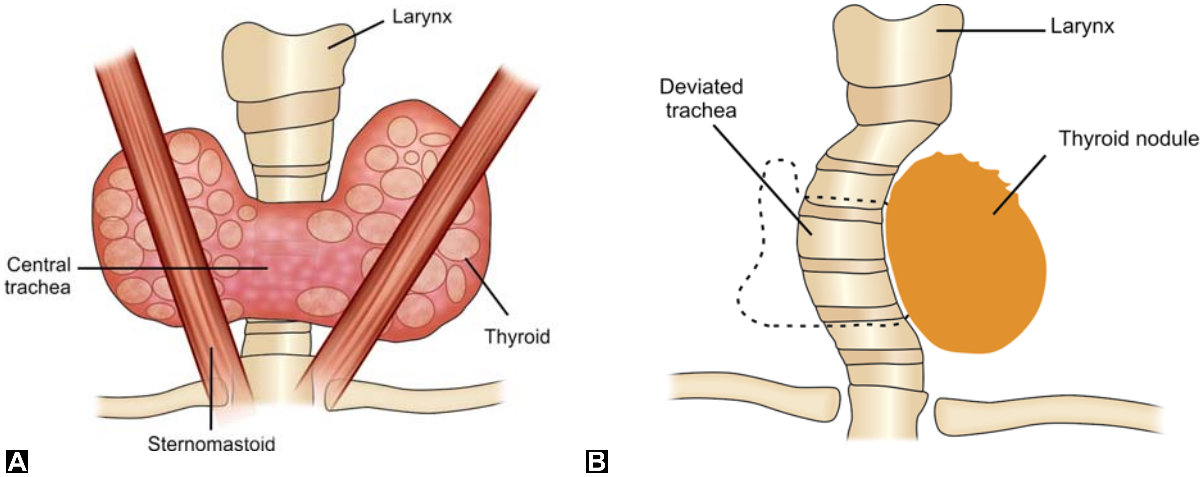

Figs 6.29A and B: Trachea should be clinically examined to find out whether it is central or deviated. In goitre involving both lobes, it is central. One lateral lobe enlargement usually causes deviation of trachea towards opposite side. Tracheal compression should be checked by *Kocher's test*. It confirms the *scabbard trachea*.

Success consists of a series of little daily efforts.

Investigations

- U/S neck (very useful).
- FNAC.
- T_3, T_4, TSH.
- *Power Doppler* is done to know the vascularity of the gland. Vascularity is described in ***resistive index (RI)*** (Harley De Nicola 2005). Normal RI is 0.65-0.7; if RI is more than 0.7 it indicates malignancy in that nodule. Malignant nodule shows anarchical angiogenesis. Flow pattern are: Type 0 – no flow; Type 1: only peripheral flow; Type 2: peripheral with small central flow; Type 3: peripheral with extensive central flow; Type 4: only central flow.
- Radioisotope study ($I^{123}/I^{131}/^{99m}Tc$).
- Serum calcitonin estimation if FNAC confirms medullary carcinoma.
- CT scan or MRI neck is not done routinely, but only in selected cases. Large swelling/to see vascularity/retrosternal extension are the indications.
- X-ray neck to see tracheal deviation.

Note:
True incidence of solitary nodule will come down to 50% from its original clinical diagnosis after investigations and surgical exploration.

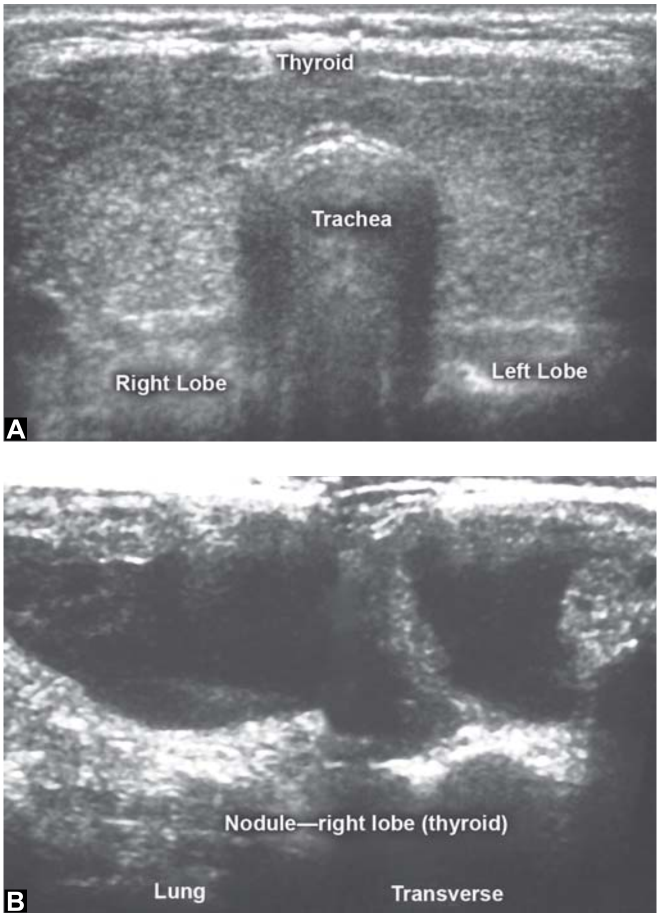

Figs 6.30A and B: Ultrasound pictures of thyroid. It is often used to see nodules, content—solid/fluid, size and extent.

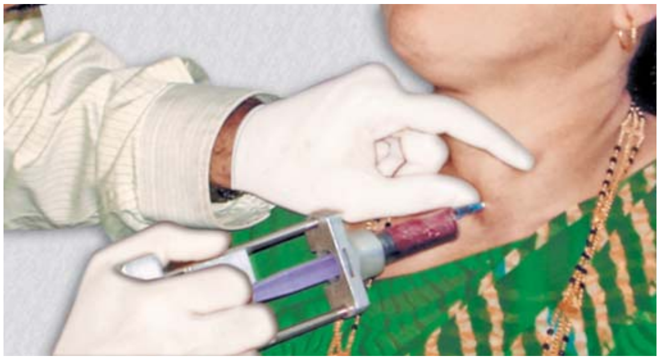

Fig. 6.31: FNAC of thyroid (*Courtesy:* Dr Krishna Upadhya, Pathologist, Nandikoor Laboratory, Mangalore).

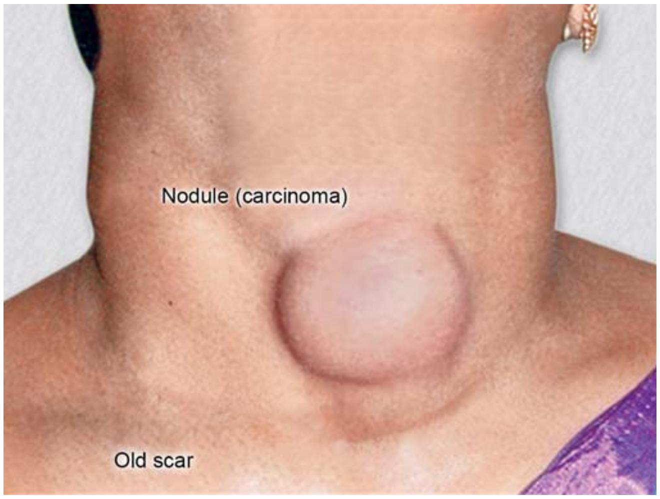

Fig. 6.32: Patient earlier operated for nodular goitre. Note recurrent nodule and scar of previous surgery.

Treatment

Indications for surgery in solitary nodule thyroid

- Malignant nodule
- Follicular neoplasm
- Toxic nodule in young
- Nodules with obstruction
- Recurrent cystic nodule
- Complex cyst (both solid and cystic components)
- Cosmetic reason

- If it is a *nontoxic nodule due to any cause, hemithyroidectomy* with complete removal of lateral lobe and whole of the isthmus is done.
- If it is *papillary carcinoma thyroid*, then *near total thyroidectomy* is done along with suppressive dose of L-thyroxine given 0.3 mg OD daily.
- If it is a *toxic nodule, radioiodine therapy,* I^{131}—5 milli curie is given orally, if the age of the patient is more than 45 years.
- If age is less than 45 years, then initially toxicity has to be controlled by antithyroid drugs, always followed by *surgery—Hemithyroidectomy*.

- If FNAC report says follicular adenoma, then hemithyroidectomy is done. If histology report says follicular carcinoma (capsular and vascular invasion) then completion total thyroidectomy is done. *Completion thyroidectomy* is done usually within 7 days or after 3 weeks. If frozen section biopsy proves carcinoma then total thyroidectomy is done.
- If there is a nodule in the isthmus, isthmectomy is done with excision of part of adjacent lateral lobes.
- If FNAC report says medullary carcinoma of thyroid, then total thyroidectomy with bilateral neck nodal dissection including central compartment is done.
- Colloid nodule may respond for conservative drug treatment using thyroxine orally in 50% cases. If nodule reappears/enlarges progressively and significantly/causing cosmetic problem then hemithyroidectomy is indicated in colloid nodule.

Thyroid cyst

- It is thyroid swelling which is cystic in nature eliciting positive fluctuation
- But tensely cystic swelling can be hard (*thyroid paradox*—with cellular tumour of thyroid can be soft also)
- Common cause is colloid degeneration—50%. There will be absence of epithelial lining
- Involution in follicular adenomas present like a cyst; 15% of such cysts may be malignant
- 30% of solitary nodules are cystic
- 15% cystic swellings in thyroid are malignant
- Cyst formation is common in papillary carcinoma of thyroid
- A cyst if contains both solid and cystic areas is called as complex cyst which is more likely to be malignant
- FNAC may cause regression of simple cyst. Even after three repeated aspirations if recurrence occurs, surgery is needed
- Surgery is indicated in complex cyst and if cyst is more than 4 cm in size

RETROSTERNAL GOITRE

Retrosternal goitre is defined as having > 50% goitre below the suprasternal notch.

- *Primary* is rare—1%. Primary retrosternal goitre arises from ectopic thyroid tissue from mediastinum. It gets its blood supply from mediastinum itself, not from the neck. And also it is not related to the existing thyroid in the neck.
- *Secondary* is common. It is extension from the enlarged thyroid from the neck. Usually arises from the lower pole of a nodular goitre. Commonly seen in short neck or obese individuals. Due to negative intrathoracic pressure, nodule gets drawn into the superior mediastinum.

Types

1. *Substernal type*: Part of the nodule is palpable in the lower neck.
2. *Plunging goitre*: An intrathoracic goitre is occasionally forced into the neck by increased intrathoracic pressure.
3. *Intrathoracic goitre* itself. Neck is normal.

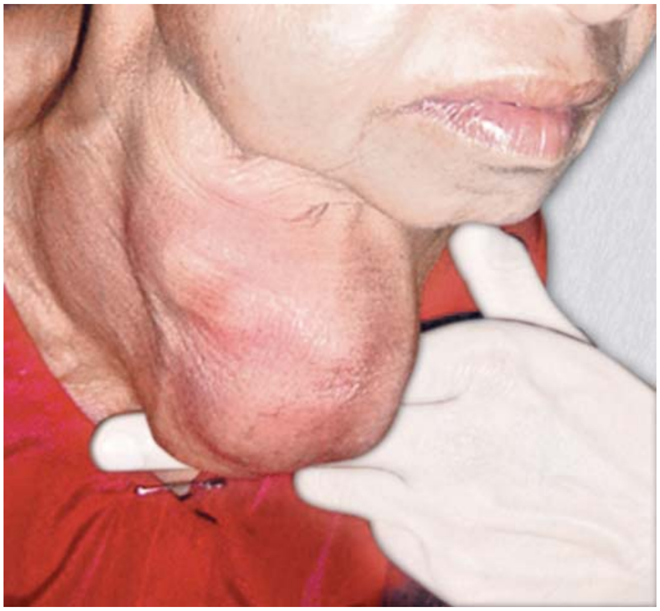

Fig. 6.33: Lower border should be checked to confirm retrosternal extension whether present or not.

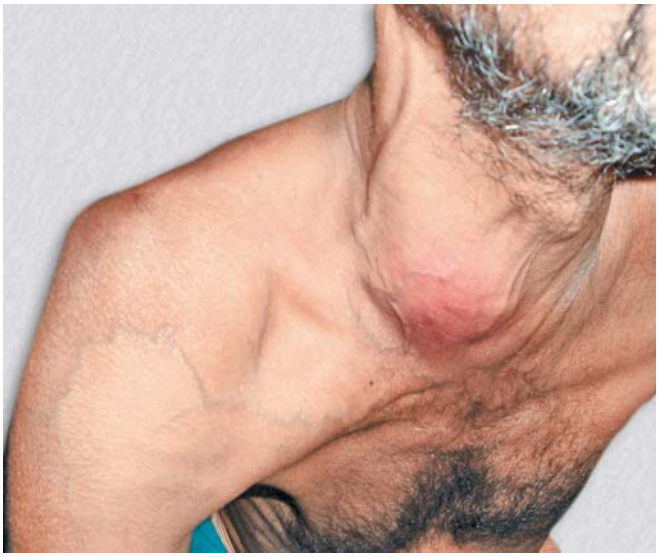

Fig. 6.34: Superior vena cava obstruction due to thyroid extending into the retrosternal area. Note the dilated veins over neck, chest wall and arm.

Clinical Features

- Dyspnoea at night during lying down or neck extended.
- Cough and *stridor* (stridor is harsh sound on inspiration).
- Dysphagia.
- Engorgement of neck veins and superficial veins on the chest wall.
- Lower border is not seen on inspection and not felt on palpation.
- *Pemberton's sign* is positive. The patient is asked to raise the arm above the shoulder level. Dilated veins are seen over neck and upper part of chest wall. Stridor and rarely dysphagia may occur. (When patient raises the arm above the shoulder level, retrosternal goitre compresses over the

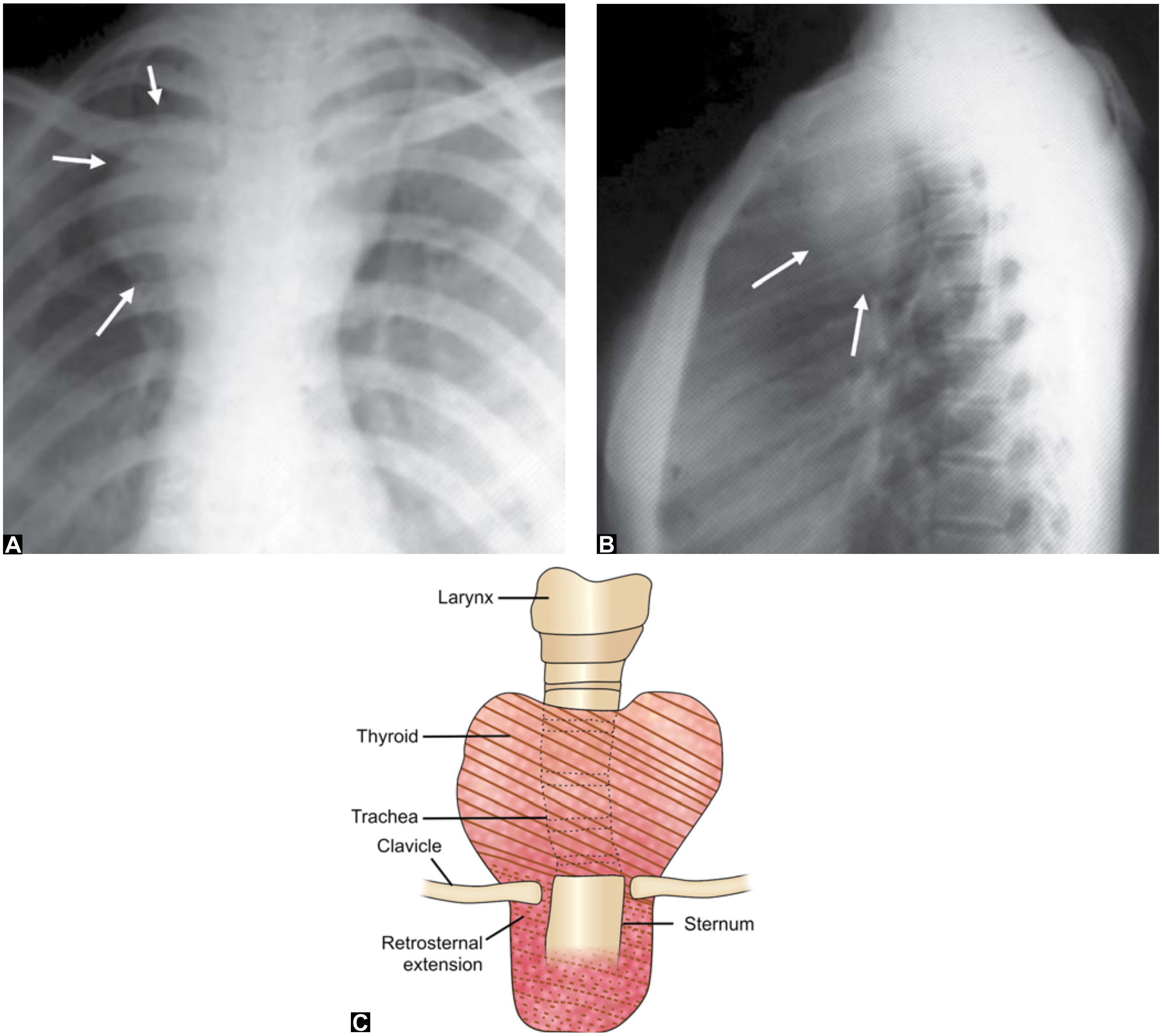

Figs 6.35A to C: (A and B) X-rays (PA view and lateral view) showing retrosternal goitre. (c) Diagrammatic representation of the retrosternal goitre.

easily compressible structures like SVC and trachea causing dilated veins and dyspnoea respectively).
- Dull note over the sternum on percussion.
- Retrosternal goitre can be either nodular, toxic or malignant.
- Rarely recurrent nerve palsy can occur.

Differential Diagnosis

- Mediastinal tumours.

Investigations

- Chest X-ray shows soft tissue shadow under the sternum.
- Radioactive iodine study is diagnostic.
- CT scan is useful.

Treatment

- Surgical removal of retrosternal thyroid is done. Commonly it can be removed through an incision in neck (as blood supply of retrosternal goitre is from neck), but in case of large retrosternal extension or in malignant type median sternotomy is required (rarely).

***Note*:**

- *Radioiodine therapy is not accepted in retrosternal goitre.*
- Stridor due to compression of tracheobronchial tree by retrosternal goitre is very dangerous because it is often not possible to clear the airway either by intubation or by tracheostomy.
- Surgical removal should be complete because recurrent retrosternal goitre is very difficult to re-operate.

Breathing difficulties in thyroid swelling

- Retrosternal goitre—positive Pembertone's sign
- Multinodular goitre of long duration—positive Kocher's test—compressive stridor
- Secondary toxic goitre—congestive cardiac failure
- Carcinoma infiltrating the trachea—stridor on rest—without compression with fingers

THYROTOXICOSIS AND HYPERTHYROIDISM

> I have lately seen three cases of violent and long-continued palpitations in females, in each of which the sample peculiarity presented itself—viz., enlargement of the thyroid gland. . . . A lady, aged twenty, became affected with some symptoms which were supposed to be hysterical. . . . It was now observed that the eyes assumed a singular appearance, for the eyeballs were apparently enlarged, so that when she slept, or tried to shut her eyes, the lids were incapable of closing.
> **—Robert James Graves,** 1835 (Irish physician)

Thyrotoxicosis is *symptom complex due to raised levels of thyroid hormones***.**

Thyrotoxicosis refers to biochemical and physiological manifestations of excessive thyroid hormones. *Hyperthyroidism* is the term used for overproduction of the hormones by thyroid gland. In hyperthyroidism pathology is in thyroid gland itself. Hyperthyroidism is one of the causes of *thyrotoxicosis*. *Thyrotoxicosis* can also occur due to other causes other than hyperthyroidism.

Other causes of thyrotoxicosis without hyperthyroidism are—ectopic functioning thyroid, struma ovarii, functioning metastatic follicular carcinoma, trophoblastic tumours, thyrotoxicosis factitia.

Types

1. Diffuse toxic goitre—(Graves' disease, *Basedow's disease*. Primary thyrotoxicosis).
2. Toxic multinodular goitre (Secondary thyrotoxicosis) (*Plummer disease*).
3. Toxic nodule *(Goetsch's disease)*.
4. Thyrotoxicosis due to rarer causes:
    a. *Thyrotoxicosis factitia*—drug induced. Due to intake of L-thyroxine more than normal.
    b. *Jod Basedow thyrotoxicosis*—because of large doses of iodides given to a hyperplastic endemic goitre.
    c. *Autoimmune thyroiditis or de Quervain's thyroiditis*.
    d. Occasionally *carcinoma thyroid.*
    e. *Neonatal thyrotoxicosis*. It subsides in 3-4 weeks as TsAb titres fall in the baby's serum.
    f. Struma ovarii.
    g. Drugs like amiodarone—an antiarrhythmic agent.
    h. Very rarely, well-differentiated carcinoma can cause thyrotoxicosis.

[***Wolf-Chaikoff effect***—Iodides inhibit the further release of hormone causing hypothyroidism *(Hokkaido goitre)*.]

Graves Disease

Graves disease is an autoimmune disease with increased levels of specific antibodies in the blood (TSH receptor antibodies). It is often associated with vitiligo. It is often familial. Thyroid stimulating immunoglobulins (TSI)/thyroid stimulating antibodies (Ts Ab) and long acting thyroid stimulator (LATS) cause pathological changes in the thyroid. Histologically there is acinar cell hypertrophy and hyperplasia with absence of normal colloid in the tall columnar epithelium (normal is flat epithelium with colloid). As cells are empty, they look vacuolated. Tissues are highly vascular. Exophthalmos producing substance (EPS) causes Graves ophthalmopathy.

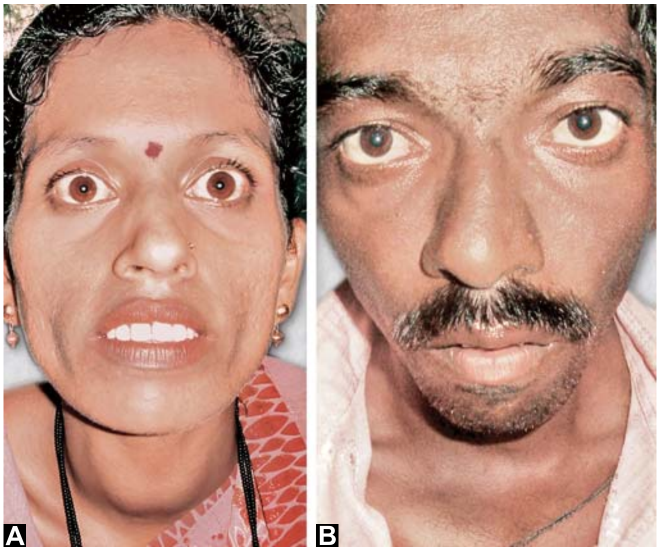

Figs 6.36A and B: Graves disease in a female and male patient. Note the exophthalmos with clearly visible lower sclera.

- Diffuse goitre, thyrotoxicosis and autoimmune manifestations like infiltrative ophthalmopathy, dermopathy, myopathy are essential components of Graves, disease. Thyroid stimulating immunoglobulins (TSI) are produced against thyroid antigen in Graves disease which is directed to TSH receptor acting as TSH receptor antibody. *This TSHR Ab is observed only in Graves disease*.
- Puberty, pregnancy, emotion and infection are the precipitating factors for primary thyrotoxicosis.
- Familial/genetic cause is also attributed in Graves disease (50%); both identical twins can develop Graves disease.
- There is *hyperplasia and hypertrophy* of entire thyroid due to prolonged continuous action by binding of abnormal thyroid stimulating antibodies to TSH receptor sites.

Toxic Adenoma

- It is benign functioning monoclonal thyroid tumour, usually more than 3 cm in size.
- It usually presents as *functioning (toxic) solitary nodule* of thyroid.
- It is autonomous functioning tumour; not TSH responding.
- Toxic adenoma secretes large quantity of thyroid hormones suppressing the function of the remaining normal thyroid tissue.
- There are no eye signs and other features of Graves disease.
- It commonly shows higher T_3 levels than T_4.
- TSH receptor or G protein genes show somatic mutation.
- US neck, T_3, T_4, TSH and radioisotope scan (shows *hot* nodule)—are the relevant investigations.

Treatment

- Initial control of toxicity with antithyroid drugs; later hemithyroidectomy is done after 6 weeks. Once thyroidectomy is done, suppressed remaining normal thyroid tissue starts functioning to secrete normal level of thyroid hormones.
- In patients after the age of 45 years radioactive iodine therapy can be used which selectively concentrates and ablates the thyroid adenoma; later remaining normal thyroid starts functioning.

T_3 Toxicosis

Here T_3 alone is raised; TSH is decreased; T_4 is normal.
Free T_3 estimation is important.

Subclinical Hyperthyroidism

- Here serum TSH is low; free T_4 is normal; there are no symptoms.
- Its incidence is 1% of hyperthyroidism.
- It is one of the causes for infertility often missed (both subclinical hyper- and hypothyroidism can cause infertility in females).
- It needs radioisotope study; US neck.
- It requires treatment—antithyroid drugs, often thyroidectomy.

Struma Ovarii

- Ovarian teratoma with thyroid differentiation will secrete T_3 and T_4 and suppress TSH.
- Function of normal thyroid in neck is suppressed.
- Radioisotope scan shows uptake in pelvis with no or less uptake in neck.

Note:

The River *Struma* arises in Bulgaria and flows into Aegean Sea. Endemic goitre exists in area along its banks; '*Struma*' means goitre.

Hashitoxicosis

- It is due to autoimmune Hashimoto's thyroiditis.
- Mild toxic features develop during initial stage of hyperplasia.
- Already formed thyroid hormones are released by inflamed gland causing toxicity. It eventually leads into euthyroid and later hypothyroidism in Hashimoto's disease.

Thyrotoxicosis Factitia

Intake of L thyroxine *without indications* to lose weight or over dose intake causes toxicity.

Postpartum Hyperthyroidism

It is exacerbation of previously confirmed or undiagnosed hyperthyroidism during pregnancy or after delivery due to increased autoimmune factors.

It is associated with HLA DR3 and HLA DR5.

Neonatal Thyrotoxicosis

It is seen in infants born to mother with Graves' disease due to crossing of the thyroid stimulating antibody (TSH RAb) across placental barrier. Infant will be toxic for 3-4 weeks which subsides gradually.

Trophoblastic Thyrotoxicosis

HCG secreted from vesicular mole, choriocarcinoma or metastatic embryonal carcinoma in females, acts like TSAb causing toxicity.

Jod Basedow Thyrotoxicosis

Patient with hyperplastic endemic goitre takes large doses of iodide, which will be taken up by hyperplastic gland in large quantity causing temporary hyperthyroidism called as *Jod Basedow thyrotoxicosis*. It differs from Basedow disease which is the other name for Graves disease ('*Jod*'—in German means iodine).

Apathetic Hyperthyroidism

- It lacks all usual clinical features of toxicity.
- It is commonly observed in old people.
- Thyroid gland is not enlarged.
- Patient presents with behavioral problems; often considered as psychiatry patient with decreased appetite.
- Such lethargic individual also often show features of recent angina and atrial fibrillation.
- Unless serum T_3, T_4 and TSH are done diagnosis is masked.

Clinical Features of Thyrotoxicosis

- It is eight times more common in females.
- Occurs in any age group.
- Primary type is seen commonly in younger age group.
- Secondary type is common in older age group.
- Graves disease often presents without any obvious thyroid swelling in the neck. Whenever there is unexplained behavioural problem, insomnia, myopathy, unexplained diarrhoea or loss of weight, tachycardia, Graves disease should be suspected and evaluated.

A. Symptoms of Hyperthyroidism/Toxicosis

Gastrointestinal system

- Weight loss in spite of increased appetite.
- Diarrhoea (due to increased activity at ganglionic level).

Cardiovascular system

- Palpitations.
- Shortness of breath at rest or on minimal exertion.
- Angina.
- Irregularity in heart rate.
- Cardiac failure in the elderly (CCF).

Differentiating points between primary and secondary hyperthyroidism

Primary thyrotoxicosis	*Secondary thyrotoxicosis*
1. Symptoms appear first, then thyroid swelling	1. Thyroid swelling appears first
2. Goitre is diffuse, smooth, firm or soft, both lobes are involved with thrill and bruit	2. Swelling is large nodular, obvious
3. Features are much more severe compared to that of secondary toxicosis of short duration	3. Symptoms appear after long time, which is less severe and slowly progressive compared to primary toxicosis
4. Eye signs and exophthalmos are common	4. Eye signs are not common
5. As it is an autoimmune disease, there may be hepatosplenomegaly	5. Cardiac features are more common
6. It occurs in younger age group	6. It occurs in adult and elderly individual
7. Entire gland is overactive	7. Internodular tissues are overactive
8. There is no preexisting goitre	8. Occurs in a long standing preexisting multinodular goiter
Histologically, there is hyperplasia of acini, lined by columnar epithelium, often containing vacuolated colloid scalloping with pseudopapillary formation	

Neuromuscular system

- Undue fatigue and muscle weakness.
- Tremor.

Skeletal system

- Increase in linear growth in children.

Genitourinary system

- Oligo- or amenorrhoea.
- Occasional urinary frequency.

Integument

- Hair loss, gynaecomastia.
- Pruritus.
- Palmar erythema.

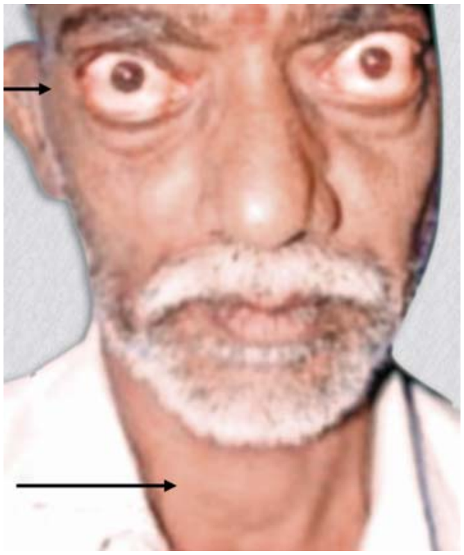

Fig. 6.38: Primary thyrotoxicosis. Note the exophthalmos.

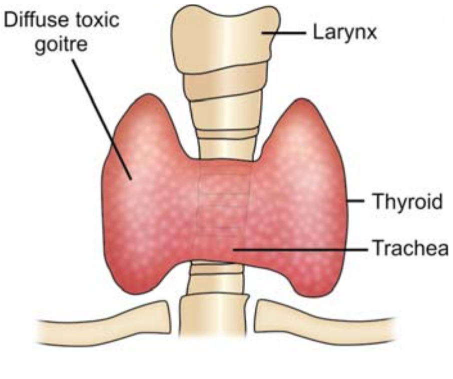

Fig. 6.37: Diffuse toxic goitre (primary) involves both lobes with hyperfunctioning acini.

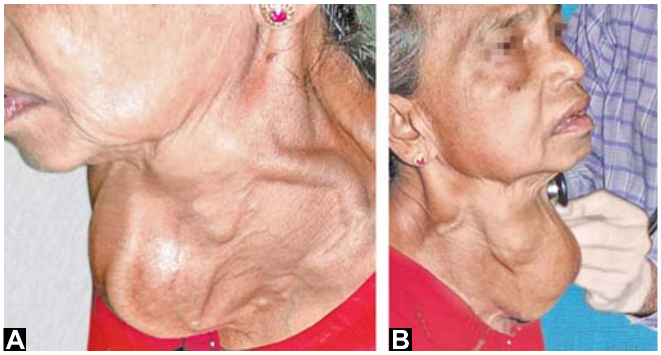

Figs 6.39A and B: Note the dilated veins in an enlarged vascular thyroid gland. Auscultation is important to hear a bruit.

Psychiatry

- Irritability.
- Nervousness.
- Insomnia.

Sympathetic overactivity

- It causes dyspnoea, palpitation, tiredness, heat intolerance, sweating, nervousness, increased appetite and decrease in weight.
- Because of the *increased catabolism* they have increased appetite, decreased weight and so also increased creatinine level which signifies *myopathy* (due to more muscle catabolism).

Fine tremor is due to diffuse irritability of grey matter.

B. Signs of Hyperthyroidism/Toxicosis

1. Eye Signs in Toxic Goitre

Eye signs are common in primary thyrotoxicosis. Lid lag, lid spasm can occur in secondary thyrotoxicosis also.

1. *Lid retraction:* Here upper eyelid is higher than normal; lower eyelid is in normal position. It is due to *sympathetic over activity* causing spasm of involuntary smooth muscle part of the levator palpebrae superioris (*Muller's muscle*). It is *a sign of* thyrotoxicosis, *not a sign* of exophthalmos.
2. *von Graefe's sign* (Lid Lag's sign): It is inability of the upper eyelid to keep pace with the eyeball when it looks downwards to follow the examiners finger.
3. *Dalrymple's sign*: Upper eyelid retraction, so visibility of upper sclera.
4. *Stellwag's sign*: Absence of normal blinking—so *staring look*. First sign to appear. It is due to widening of palpebral fissure due to lid retraction and also due to contraction of voluntary part of levator palpebrae superioris muscle.
5. *Joffroy's sign*: Absence of wrinkling on forehead when patient looks up (frowns).
6. *Moebius sign*: It is lack of convergence of eyeball. Defective convergence is due to lymphocytic infiltration of inferior oblique and rectus muscles in case of primary

Wayne's diagnostic indices (clinical)

	Symptoms	*Present*	*Absent*
1.	Dyspnoea on effort	+1	
2.	Palpitation	+2	
3.	Tiredness	+2	
4.	Preference to heat		–5
5.	Preference to cold (Heat intolerance)	+5	
6.	Excessive sweating	+3	
7.	Nervousness	+2	
8.	Appetite increased	+3	
9.	Weight decreased	+3	
	Signs	*Present*	*Absent*
1.	Bruit over thyroid	+2	
2.	Exophthalmos	+2	
3.	Lid retraction	+2	
4.	Lid lag	+1	
5.	Hyperkinetic movements	+4	–2
6.	Fine finger tremors	+1	
7.	Hands hot	+2	–2
	Moist	+1	–1
8.	Atrial fibrillation	+4	–3
9.	Pulse rate 80/minute	0	
	80-90/minute	+3	
	More than 90/minute	+3	
10.	Palpable thyroid		
	< 11 points—nontoxic goitre	11-19 equivocal	> 19 points—toxic

thyrotoxicosis. There will be diplopia. It may be an early sign of eventual ophthalmoplegia.

7. *Jellinek's sign*: Increased pigmentation of eyelid margins.
8. *Enroth sign*: Oedema of eyelids and conjunctiva.
9. *Rosenbach's sign*: Tremor of closed eyelids.
10. *Gifford's sign*: Difficulty in everting upper eyelid in primary toxic thyroid. Differentiates from exophthalmos of other causes.
11. *Loewi's sign*: Dilatation of pupil with weak adrenaline solution.
12. *Knie's sign*: Unequal pupillary dilatation.
13. *Cowen's sign*: Jerky pupillary contraction to consensual light.
14. *Kocher's sign*: When clinician places his hands on patient's eyes and lifts it higher, patient's upper lid springs up more quickly than eyebrows.
15. *Naffziger's sign*: With patient in sitting position and neck fully extended, protruded eyeball can be visualized when observed from behind.

Order of appearance of signs

- Stellwag's sign—mild; first sign to appear
- von Graefe's sign—mild
- Joffroy's sign—moderate
- Moebius sign—severe

EXOPHTHALMOS

- It is proptosis of the eye, caused by infiltration of the retrobulbar tissues with fluid and round cells, with *lid spasm* of upper eyelid (Lid spasm is spasm of levator palpebrae superioris muscle which is partly innervated by sympathetic fibres).
- Sclera can be seen clearly below the limbus of the eye.
- Proptosis can be measured by *exophthalmometer.*
- Exophthalmos is often self limiting, but not always. Sleeping in propped up position and lateral tarsorrhaphy will help to protect the eye.

Severe Exophthalmos

- Eyelid oedema, chemosis, conjunctival injection.
- Diplopia, ophthalmoplegia (complete weakness of all extraocular muscles and so no movements possible).
- Corneal ulceration.
- Papilloedema soon develops.
- Finally it may also cause loss of vision.

It is called as ***malignant exophthalmos*** (It is misnomer even though it is not malignant nor related to any malignancy).

Treatment is emergency, i.e. large doses of systemic steroids (Prednisolone) are given along with *orbital decompression*, systemic antibiotics, steroid drops, antibiotic drops.

- **Remember:**
    - *Antithyroid drugs may worsen exophthalmos and the patient should be observed once antithyroid drugs are started as steroid supplementation may be required.*
    - Ophthalmopathy may worsen by thyroidectomy or radioiodine therapy also.

Treatment of severe exophthalmos

- Steroids intravenously
- IV antibiotics
- Guanethidine, steroid, antibiotic drops
- Lateral tarsorrhaphy
- Orbital decompression by surgical removal of lateral wall or roof of orbit
- Diuretics
- Dark spectacles, protective eye patches
- Eyelid surgeries, extraocular muscle surgeries
- Local radiation therapy to orbital space
- 1% methylcellulose eye drops to prevent corneal ulceration.
- Sleeping with head end elevation

- **Lid retraction** is higher upper eyelid with normal lower eyelid with visible sclera adjacent to upper eyelid
- **Lid lag** is inability of the upper eyelid to keep pace with the eyeball when it looks downwards to follow the examiners finger
- **Exophthalmos** is visible sclera *first below* (lower part) the lower edge of the iris and later eventually upper part of sclera will be visible. It is due to pushing of the eyeball forwards due to fat, oedema fluid, cells like macrophages in retrobulbar space

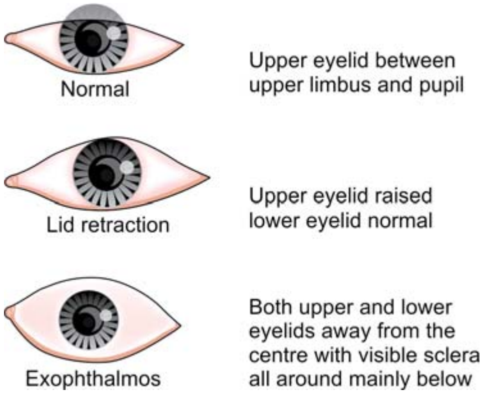
Normal
Upper eyelid between upper limbus and pupil
Lid retraction
Upper eyelid raised lower eyelid normal
Exophthalmos
Both upper and lower eyelids away from the centre with visible sclera all around mainly below

Fig. 6.40: Diagram showing normal relation of eyelid and pupil in lid retraction and exophthalmos.

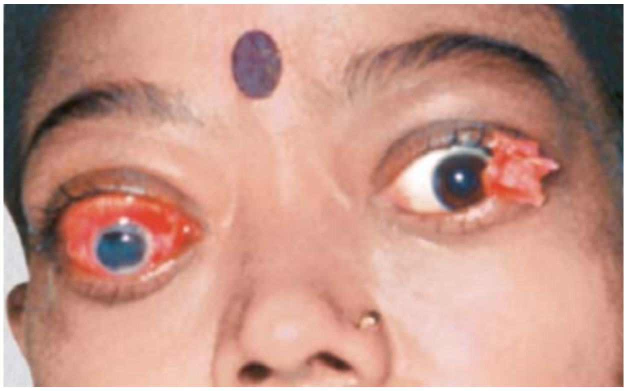

Fig. 6.41: Malignant exophthalmos.

You don't drown by falling in the water; you drown by staying there —**Edwin Louis Cole**

Important signs to be remembered

- *Visible lower sclera—sign of exophthalmos*
- Naffziger's sign
- von Graefe's sign-upper lid lag—contraction/overactivity of the involuntary part of the levator palpebrae superioris muscle—Muller's muscle
- Joffroy's sign
- Moebius sign—*most important—early sign of ophthalmoplegia*

Thyroid ophthalmopathy in Grave's disease-Werner's abridged classification of ocular changes with van Dyke's modification

Class-grade	*Definition*
0	No signs and symptoms
1	Eye signs only—refer table below for eye signs
2	Soft tissue involvement
3	Proptosis more than 22 mm
4	Extraocular muscle involvement
5	Corneal involvement- ulceration
6	Loss of sight/vision due to optic nerve and corneal involvement

Eye signs only

- Resistance to retro-displacement of eye
- Oedema of conjunctiva and caruncle
- Lacrimal gland enlargement
- Injection of conjunctiva
- Oedema and fullness of lids

Grading of exophthalmos

- **Mild:** Widening of palpebral fissure due to lid retraction
- **Moderate:** Orbital deposition of fat causing bulging with positive Joffroy's sign
- **Severe:** Congestion with intraorbital oedema, raised intraocular pressure, diplopia and ophthalmoplegia
- **Progressive:** In spite of proper treatment progression of eye signs is seen with chemosis, corneal ulceration and ophthalmoplegia

2. *Cardiac Manifestations*

1. Tachycardia is common.

   As per **Crile's grading**

   Sleeping pulse rate is usually checked for three consecutive nights and average is taken as the value.

Pulse rate

- Grade I: < 90/mt
- Grade II: 90-110/mt
- Grade III: >110/mt

2. Ectopic.
3. Pulsus paradoxus.
4. Wide pulse pressure.
5. Multiple extrasystoles.
6. Paroxysmal atrial tachycardia.
7. Paroxysmal atrial fibrillation.
8. Persistent atrial fibrillation (not responsive to digoxin).

3. *Myopathy*

1. Weakness of proximal muscles occurs, i.e. the front thigh muscles, arm muscles.
2. Weakness is more when muscle contracts isometrically either while getting down steps, or lifting a full bucket.
3. Often when it is severe it resembles myasthenia gravis. Once hyperthyroidism is controlled recovery occurs.

4. *Pretibial Myxoedema*

It is a misnomer. *Pretibial myxoedema* is often a feature of primary thyrotoxicosis:

- Is usually bilateral, symmetrical, shiny, red thickened dry skin with coarse hair in the feet and ankles.
- In severe cases skin of entire leg below the knee with involvement of foot and ankle can occur.
- It is due to deposition of myxomatous tissues (mucin like deposits) in skin and subcutaneous plane. Glycosaminoglycans (hyaluronic acid) deposition occurs.
- It might or might not regress completely after treatment for toxicity.
- It is associated with exophthalmos with high levels of thyroid stimulating antibodies.
- Skin becomes cyanotic when cold. Skin changes in toxicosis are called as *thyroid dermopathy*. They include—pretibial myxoedema, pruritus, palmar erythema, hair thinning, Dupuytren's contracture (fascial).

5. *Thyroid Acropachy*

Thyroid acropachy is clubbing of fingers and toes in primary thyrotoxicosis. Hypertrophic pulmonary osteoarthropathy can develop.

6. *Others*

- *Thrill is felt* in the upper pole of the thyroid and also bruit on auscultation. It is because in upper pole, superior thyroid artery enters the gland superficially and so thrill and bruit can easily be felt. In lower pole inferior thyroid artery enters the gland from deeper plane and so thrill cannot be felt.
- Hepatosplenomegaly.

TOXIC NODULE

- Is a solitary overactive nodule.
- There is an autonomous hypertrophy and hyperplasia of a part of the gland where there is a nodule. [It is not due to Thyroid stimulating antibody (Ts Ab)].
- Here high levels of circulating thyroid hormones suppress TSH secretion and so, normal thyroid tissue surrounding the nodule is itself suppressed and inactive.
- Once patient becomes euthyroid by drugs, surgery (hemithyroidectomy) or radioactive iodine therapy I^{131} in a therapeutic dose of 5 m curie is given orally.
- Because normal gland is inactive, radioactive iodine affects only the autonomous nodule, allowing the normal gland to remain intact which later gets activated and functions normally.

Causes of exophthalmos	*Causes of pulsating exophthalmos*
	Carotid-cavernous sinus A-V fistula
Endocrinal	Cavernous sinus thrombosis
• Thyrotoxicosis—common	Orbital vascular neoplasm
• Cushing's syndrome, acromegaly—rare	Orbital haemangioma
	Ophthalmic artery aneurysm
Congenital deformities of skull	
• Craniostenosis, oxycephaly, hypertelorism	
Primary tumours	
• Periorbital meningioma	
• Optic nerve glioma	
• Orbital haemangioma	
• Lymphoma	
• Osteoma	
• Pseudo tumour—granuloma	
Secondary tumours	
• Antral carcinoma, neuroblastoma	
Inflammatory	
• Orbital cellulites, frontal sinusitis	
Vascular causes	
• Cavernous sinus thrombosis/A-V fistula	
• Ophthalmic artery aneurysm	
Other eye causes	
• Severe myopia	
• Severe glaucoma—buphthalmos	

- Drugs are used initially, only for a temporary period to make the patient euthyroid.
- Before the age of 45, surgery is preferred.
- After 45 years age, radioiodine therapy is used.

Investigations

- *Thyroid function tests*
  Serum T_3 and T_4 levels are very high. TSH is very low or undetectable. Sometimes, only T_3 level is increased and is called as *T_3 toxicosis*. Here in T_3 toxicosis, *free T_3 estimation is important*. Free T_3, free T_4 estimation is done as total T_3 and total T_4 levels will vary depending on the amount of thyroid binding globulin (TBG). TBG will be raised in pregnancy, cirrhosis, hyperestrogenism. It decreases in conditions with high androgen level, hypoproteinaemia, acromegaly. Free T_3 and free T4 are measured using radioimmunoassay. Normal free T_3 is 3.0-9.0 pmol/L; free T_4 is 8-26 nmol/L.

Type of disease	*T_4*	*T_3*	*TSH*
Conventional hyperthyroidism	Increased	Increased	Undetectable
T_3 hyperthyroidism	→	Increased	Undetectable
Subclinical hyperthyroidism	→	→	Undetectable

- *Radioisotope study*
  - *Radioisotope study by I^{131}* (Diagnostic dose -5 microcurie is used) shows more uptake, i.e. *hot nodules* or hot areas. This is very useful in autonomous solitary toxic nodule.
  - I^{131} causes more irradiation and its half life is 8 days. So *intravenous ^{99m}Tc* is used for diagnostic purpose. 99m technetium has become isotope of choice for diagnosis as it is cheap, less radiation, scanning is done 20 minutes after IV injection of ^{99m}Tc (half life is 6 hours). Drawback of technetium is that it concentrates in carcinoma, so forms hot nodule (means hot nodule need not be benign in Tc scanning). Warm nodule in Tc scan may appear as cold nodule in RAI scan and so is called as *discordant nodule* which suggests malignancy. If radioactive iodine is used for diagnosis then I^{123} is better as it has got short half life (13 hours).
  - Autonomous toxic nodule is absolute indication for radioisotope scan in toxic thyroid showing *hot nodule*. Graves disease shows *diffuse* over activity (*uniform*); hypofunctioning cold nodule in Graves disease could be malignant. In secondary thyrotoxicosis internodular tissues are overactive (*heterogenous* activity). Nonhyperthyroid toxicosis shows increased uptake in nonthyroid areas of toxicity like struma ovarii in pelvis.
- TRH estimation.
- ECG—to look for cardiac involvement; if required opinion from cardiologists is taken and cardiac problems are managed.

Dissatisfaction and discouragement are not caused by the absence of things but the absence of vision.

- Total count and neutrophil count are very essential base line investigations before starting antithyroid drugs (as it may cause agranulocytosis).
- Thyroid antibodies estimation—antithyroglobulin antibody, TSH receptor antibody.

Treatment

- Antithyroid drugs
- Surgery
- Radioiodine therapy

1. ANTITHYROID DRUGS

Indications for antithyroid drugs:
- Toxicity in pregnant women—Propylthiouracil is preferred
- Toxicity in children and young adults
- Before subtotal thyroidectomy, to make the patient euthyroid
- Soon after starting radioactive I^{131} therapy for 6 to 12 weeks (Effects of radiotherapy start only in 6 to 12 weeks)

a. Carbimazole:
- Is the commonest drug used.
- Dose is 5-10 mg, exactly 8th hourly (as T1/2 of carbimazole is 8 hours). Each tablet is 5 mg.
- Usually given for 12-18 months.
- Peak plasma level should be maintained in optimum concentration to have a proper benefit.
- Often tri-iodothyronine 20 microgram 4 times daily or Thyroxine 0.1 mg daily are given in combination with antithyroid drugs, to prevent iatrogenic thyroid insufficiency or to prevent the increase in size of goitre.
- It acts by blocking thyroid hormone synthesis. Carbimazole also suppresses the autoimmune process in thyroid in Grave's disease.
- Carbimazole causes fever, rashes, arthralgia, myalgia, neuritis, lymph node enlargement, liver cell dysfunction, psychosis, agranulocytosis.

b. Methimazole: Alike carbimazole. Dose is 5 to 20 mg daily. long acting, once a day dose.

c. Propyl thiouracil:
- It acts by blocking thyroid hormone synthesis as well as by blocking peripheral conversion of T_4 to T_3.
- It also decreases the thyroid autoantibody levels.
- It can be given for hyperthyroidism in children and in pregnancy, lactation.
- Dose is 200 mg 8th hourly.

PTU causes dose unrelated hepatoxicity; agranulocytosis; antineutrophilic cytoplasmic antibody in 20% of patients after long term usage.

Antithyroid drugs are continued during and after surgery, for 7-10 days. It has to be given after starting radioactive iodine therapy for 6 weeks to 12 weeks.

Response to treatment and possibility of relapse in primary thyrotoxicosis can be assessed by studying HLA status and TsAb level.

d. Propranolol:
- It is a beta blocker, which is used as an antithyroid drug.
- Dose is 40 mg tid.
- It reduces the cardiac problems and also blocks the peripheral conversion of T_4 to T_3, as it is the T_3 which is the principle active agent in periphery.
- Contraindications are bronchial asthma, heart block, cardiac failure.
- Long acting nadolol 160 mg OD can also be used.

e. Lugol's iodine (5% iodine + 10% potassium iodide):
It decreases the vascularity of the gland and makes it more firm and easier to handle during surgery. Dose is 10-30 drops/day (minims) for 10 days prior to surgery. Potassium iodide tablets 60 mg tid also can be given instead of Lugol's iodine. But its use at present is disqualified.
(One minim = one drop. One ml = 16 drops).

Note:
Lugol's iodine prevents the release of hormone from the gland – **thyroid constipation**. After 2 weeks, effect of Lugol's iodine is lost causing thyroid escape from iodine control.

f. Block and replacement therapy:

It is giving high dose of carbimazole to inhibit T_3 and T_4 production compeltely with a maintenance dose of 0.1 mg of L thyroxine. It reduces the iatrogenic thyroid insufficiency.

g. High dose of glucocorticoids impair peripheral conversion of T_4 to T_3 and also lowers serum TSH level; hence can be used in severe resistant/refractory cases.

Advantages of Antithyroid Drugs

1. Avoids surgery and its complications.
2. Avoids radioiodine therapy.

Clinical improvement occurs in 2 weeks. Biochemical improvement occurs in 6 weeks.

Remission is confirmed by TsAb assessment, which will be low.

Permanent remission rate is very less in adults and is 20% in children.

Disadvantages

1. Prolonged course of treatment for 18 months, and in spite of this cannot predict the remission or relapse. Relapse rate is 40%. Large gland/severe disease/abnormal TSH receptor antibodies (TSH-RAbs) are likely to lead into high recurrence.
2. Size of swelling may not regress.
3. It may lead to *agranulocytosis and thrombocytopenia*, liver damage, hair loss (except propranolol).
Sore throat is the earliest presentation of agranulocytosis. If it is so, drug is stopped, total count is done. If it is less, agranulocytosis is confirmed. High doses of injection benzyl penicillin 10-20 lac, 6th hourly, IV is started to prevent infection. If required, blood transfusion is done. Patient usually recovers by this. To control toxicity, Tab. Propranolol 40 mg tid is started. Rarely they need bone marrow transplantation.

Other Antithyroid Drugs which are not Commonly Used are:

- Potassium perchlorate inhibits iodide transport.
- Iopanoic acid 50 mg is used in severe unresponsive cases. . It inhibits peripheral conversion of T_4 to T_3.

- Lithium carbonate 300 mg 6th hourly.
- Guanethidine 40 mg orally 6th hourly.
- Reserpine 5 mg IM.
- Dexamethasone 2 mg orally 6th hourly. It inhibits peripheral conversion of T_4 to T_3. It is used in thyroid storm.
- 2 mg IV propranolol in thyroid storm.

2. SURGERY

Indications

- Failure of drug treatment in primary thyrotoxicosis in young patients
- Autonomous toxic nodule
- Nodular toxic goitre
- When malignancy cannot be ruled out
- Graves disease in children, Graves with nodules
- Need for antithyroid drugs for more than 2 years
- Large goitre, substernal/intrathoracic goitre
- Pressure symptoms, Graves ophthalmopathy
- Amiodarone induced thyrotoxicosis.

- Surgery done is *subtotal thyroidectomy*—Both lobes with isthmus are removed and a tissue equivalent to pulp of finger is retained at lower pole of the gland on both sides (5-8 grams).
- In autonomous nodule, *hemithyroidectomy* is done. Here entire lateral lobe with whole of isthmus is removed.
- It is now observed that *total thyroidectomy* may be a better option in Graves disease to achieve lowest relapse rate and successful stabilisation of thyroid ophthalmopathy as it clears the antigenic focus in thyroid completely.

Advantages

- Rapid and high cure rate.
- Problems of radioiodine therapy are avoided.
- Surgery provides tissue for biopsy, removes the occult malignant foci.
- Surgery is better option for ophthalmopathy due to thyrotoxicosis.
- It is the option for women planning for child.
- Coexisting parathyroid carcinoma can be removed.
- For intrathoracic retrosternal toxic thyroid, surgery is the choice. Antithyroid drugs and also radioactive iodine may increase the goitre size.

Patient should be made euthyroid before doing surgery. (It should be confirmed by repeated estimation of serum T_3, T_4 and TSH levels).

Disadvantages

- Recurrent thyrotoxicosis (5%).
- Thyroid insufficiency (20-45%). It is revealed in 6 months to 2 years and is confirmed by estimating T_3, T_4 and TSH levels. Hypothyroidism is better than recurrent thyrotoxicosis. It is treated by tab. L-thyroxine 0.1 mg daily (OD) for life-long.
- Complications of thyroid surgery itself.

3. RADIOIODINE THERAPY

Indications

- Primary thyrotoxicosis after 45 years of age
- In autonomous toxic nodule
- In recurrent thyrotoxicosis

- Radioiodine destroys the cells and causes the complete ablation of thyroid gland. It is given only after the age of 45 years, as the chances of genetic mutation (damage), leukaemia, carcinomas are high in younger individual.
- *Usual dose is 5 to 10 millicurie, or 160 microcurie/gm of thyroid.*
- Patient is made euthyroid using antithyroid drugs; drug is discontinued for 5 days; I^{131} 300-600 MBq is given orally; antithyroid drugs are started after 7 days and are continued for 8 weeks. In 30% of patients additional 2 or 3 doses may be required.

Choice of therapies

	Condition	*Age*	*Treatment*
1.	*Diffuse toxic goitre*:		
	a. Small goitre	Over 45 years	Antithyroid drugs for 18 months
	b. Large goitre	Under 45 years	Radioiodine therapy Surgery (*Subtotal thyroidectomy*)
2.	*Toxic nodular goitre*		Surgery (Subtotal thyroidectomy)
			Initially antithyroid drugs are given to make the patient euthyroid before surgery
3.	*Toxic solitary nodule*	Over 45 years	Radioiodine
		Under 45 years	Surgery (Hemithyroidectomy)
4.	*Recurrent thyrotoxicosis after surgery*	Under 45 years	Antithyroid drugs
		Over 45 years	Radioiodine therapy
5.	*Failure of antithyroid drugs or radioiodine therapy*		Surgery

Don't spend a dollars' worth of time for ten cents worth of results.

- It takes 3 months to get full response and so until then, the patient has to take antithyroid drugs. Often additional one or two doses of radioiodine are required to have complete ablation. Eventually they go for hypothyroidism and so require maintenance dose of L-thyroxine 0.1 mg daily.
- To give therapeutic dose, patient should be admitted and isolated for 7 days (Half life) to prevent irradiation. It is given orally soon after getting from the manufacturer without much delay to have optimal efficacy.

Advantages

1. No surgery.
2. No prolonged drug therapy.
3. Cure rate 90%.

Disadvantages

1. Availability of facilities.
2. Proper follow-up is essential.

TOXIC THYROID IN PREGNANCY

- *Radioiodine therapy is absolutely contraindicated* in pregnancy (High-risk to foetus).
- Antithyroid drugs can be administered carefully.
- But, the problem here is that both TSH and antithyroid drugs cross the placental barrier and baby born may be hypothyroid and goitrous.
- Propylthiouracil is preferred in pregnancy.
- Subtotal thyroidectomy can be done in second trimester.

TOXIC THYROID IN CHILDREN

- Radioiodine therapy is *absolutely contraindicated* in children because of the high-risk of developing thyroid carcinoma.
- Recurrence rate is also very high after surgery.
- So proposed treatment is, initially antithyroid drugs are given until adolescent period and then subtotal thyroidectomy is done.

THYROCARDIAC

Severe cardiac damage (partly or wholly) resulting from hyperthyroidism, usually secondary type, requires proper opinion from cardiologists and treatment with propranolol. Subtotal thyroidectomy is the treatment.

IN A PATIENT WITH THYROTOXICOSIS, WITH RECENT ONSET OF PROPTOSIS

Early thyroidectomy has to be avoided, because early surgery may precipitate malignant exophthalmos. Here the patient has to be treated initially with antithyroid drugs and if required with steroids, until the proptosis has remained static for six months. Then subtotal thyroidectomy is done.

Since half-life of L-thyroxine is 7 days, propranolol and antithyroid drugs has to be continued for 7 days after thyroidectomy.

T_3 Thyrotoxicosis should be suspected if the clinical picture is suggestive of toxicosis, but routine tests for thyroid function are within normal range.

RADIOACTIVE IODINE

It is used both as a diagnostic as well as a therapeutic agent.

1. **I^{131}**—*is used for radioactive iodine therapy* (β-rays are used).
2. **I^{123}**—*is used for diagnostic studies* (γ-rays are used).

For diagnostic purpose I^{123} is given orally in empty stomach on previous day *(dose—5 micro curie; $T_{1/2}$ (half life) of I^{123} is 13 hours* and so it is suitable for diagnostic purpose). Patient should not take L-thyroxine for 6 weeks prior to radioisotope study.

Note:
If patient is on T_3 (60 µg/day) medication, the drug is stopped 10 days before radioisotope scan.

Injection TRH, if given, radioisotope scan can be done in 24 hours.

Thyroid treats this I^{123} similar to inorganic I^{127}. This I^{123} enters the thyroid from the circulation and gets incorporated into T_3, T_4 and later released into circulation as protein bound iodide (PBI). *Normal value of PBI is 8 µgm%.*

Using *Gieger Muller's gamma ray counter,* scanning of thyroid gland is done to visualize the gland.

- *Hot* area suggests more uptake,
- *Warm* area suggests normal uptake,
- *Cold* area suggests no uptake.

I^{123} radioisotope can be safely used in children and pregnancy for diagnostic purpose only (5 microcurie) as the dose is low.

Note:
- Presently I^{131} is also becoming popular as a diagnostic tool. For treating malignancy β-rays are more used than γ-rays.
- Technetium 99 scan is used for diagnostic purpose at present as it is effective and faster (in 20 minutes). IV administration is used.

Indications for diagnostic radioactive iodine study

- Doubtful toxicity
- Ectopic thyroid
- Autonomous toxic nodule
- After total thyroidectomy, to look for secondaries in follicular carcinoma thyroid
- Retrosternal thyroid

Radioisotope study is done to look for secondaries by doing whole body scanning (total body scintygraphy). For diagnostic radioactive study Technetium 99 pertechnetate can also be used.

Therapeutic Uses

1. In primary thyrotoxicosis after 45 years.
2. In autonomous toxic nodule after 45 years, it is useful as remaining gland still will function adequately after radiotherapy (As during radiotherapy radioisotope will not be taken up by this retained normal gland as it is suppressed in the presence of toxic nodule which will function later adequately).
3. In follicular carcinoma of thyroid, after total thyroidectomy, if there are secondaries elsewhere in the body, as in bones

or lungs, then radioiodine therapy is given. I^{131} is given as its half-life is 8 days. Patient should be isolated for this period and careful disposal of urine has to be done during this period. It is given orally in a dose of 5 millicuries (160 microcurie/gm of thyroid).

Problems in Radioactive Iodine Therapy

- Permanent thyroid failure with hypothyroidism.
- Genetic damage, leukaemia, damage to foetus.
- Effects will be seen only after 3 months.
- Ophthalmopathy and dermopathy will be worsened.
- There is risk of malignant transformation of thyroid.
- It may induce hyperparathyroidism.

Dose of radioactive iodine

Diagnostic	*Therapeutic*
• For thyroid-5-50 microcurie	• Residual thyroid ablation—50 mCi
• For whole body iodine scan 5-10 millicurie in 72 hours	• Bone secondaries from FCT—100-120 mCi
	• Lung secondaries from FCT—180 mCi

Contraindications for Radioactive Iodine Therapy

- Pregnancy/females desiring to have pregnancy within 1 year/lactating mothers.
- Children/adolescents.

Radioactive isotopes used in thyroid

Isotopes	*Route of administration*	*Half life*	*Type of rays*
I^{123}	Oral	13 hr	γ rays
I^{125}	Oral	60 days	γ rays
I^{131}	Oral	8 days	β, γ rays
I^{132}	Oral	2.3 hr	
Tc 99 scan*	IV	6 hr	

* Used mainly for malignancy in thyroid itself. It is sensitive, convenient, low radiation exposure, inexpensive, with good images but nonspecific and not used for therapy.

Lithium is also used as isotope for diagnosis in thyroid diseases. Technetium is better to identify nonfunctioning secondaries.

Note

- High dose of retinoic acid will make I^{131} to concentrate in tumour cells (70 mg/daily for 2 weeks)
- Fertility should be avoided for 1 year after I^{131} therapy
- Avoid contrast CT in thyroid diseases as much as possible because I^{131} study in later period will be difficult
- MRI is ideal when radioiodine therapy is needed

THYROID NEOPLASMS

A. *Benign*

a. Follicular adenoma—colloid, embryonal, fetal.
b. Hurthle cell adenoma.
Colloid adenoma (commonest adenoma).
Papillary adenoma—its existence is doubtful. It is invariably low-grade papillary carcinoma.

B. *Malignant* *(Dunhill classification).*

a. *Differentiated—80%*
1. Papillary carcinoma (60%).
2. Follicular carcinoma (17%).
3. Papillofollicular carcinoma behaves like papillary carcinoma of thyroid.
4. Hurthle cell carcinoma behaves like follicular carcinoma.

b. *Undifferentiated—20%*
Anaplastic carcinoma (13%).
c. *Medullary carcinoma* (6%).
d. *Malignant lymphoma* (4%)
e. *Secondaries* in thyroid (rare)—from colon, kidney, melanoma, breast.

Annual incidence of thyroid cancers is 3.7 per 1,00,000 population. It is common in females (3:1).

Papillary carcinoma mainly spreads through lymphatics; follicular through blood; anaplastic through lymphatics and blood.

Aetiology of Thyroid Malignancy

1. Radiation either external or radioiodine can cause papillary carcinoma thyroid. There is increased incidence of thyroid carcinoma among children following exposure to ionising radiation after the *Chernobyl nuclear disaster* in Ukraine in 1986.

   Earlier irradiation was practised to head and neck region to treat benign conditions like tonsillitis, adenoids, thymus enlargement, acne vulgaris, haemangiomas during first two decades of life. As a consequence papillary carcinoma of thyroid became common in these individuals. Radiotherapy received in adolescent period for Hodgkin's lymphoma may predispose to papillary carcinoma.
2. Pre-existing multinodular goitre. It turns into follicular carcinoma of thyroid.
3. Medullary carcinoma thyroid is often familial.
4. Hashimoto's thyroiditis may predispose to NHL/papillary carcinoma of thyroid.
5. Familial.
6. Elevated TSH is observed in papillary carcinoma of thyroid.
7. Genetic—*Cowden syndrome* is differentiated thyroid carcinoma, carcinoma breast, multiple hamartomas. It is due to germ cell mutation of *PTEN tumour suppressor gene*. Oncogenes—C myc, C erb, , C fos, Ras are associated thyroid neoplasms.

DIFFERENTIATED THYROID CARCINOMA (DTC)

- DTC is a spectrum of disease derived from follicular cells. Both papillary and follicular carcinomas are grouped under this. 90% of thyroid malignancies are differentiated one. Papillary, follicular and Hurthle cell carcinomas are DTCs.

- AGES (Mayo Clinic, Hay); AMES (Lahey Clinic); MACIS; *Sloan Kettering* scoring—are different scoring systems used for DTCs. Sloan Kettering scoring includes low, intermediate and high risk groups. First three scoring systems have low and high risk groups.
- Papillary carcinoma spreads to lymph nodes; follicular carcinoma through blood. FCT causes pulsatile vascular secondaries.
- Incidence of thyrotoxicosis in DTCs is 2%.
- Galectin—3, RET/PTC rearrangements, CD44 are the probable tumour markers under evaluation.
- *TNM staging for DTCs are*: Tx—primary tumour size unknown; T1—< 2 cm; T2—2-4 cm; T3—> 4 cm limited to thyroid or only minimally extrathyroid extension (to sternothyroid soft tissue); T4a—invasion to subcutaneous tissue, larynx, trachea, oesophagus and recurrent laryngeal nerve; T4b—invasion to prevertebral fascia, carotid encasement. N0—no nodes; N1a—spread to level VI (pre/paratracheal, prellaryngeal); N1b—neck nodal spread same or opposite or both sides and superior mediastinal nodes. M0—no distant spread; M1—distant spread present. *Age* is included in staging.

Postoperative AJCC staging for DTCs for predicting risk of death

Age less than 45 years

Stage I – Ant T, any N, M0; Stage II – any T, any N, M1.

Age after 45 years

Stage I – T1, N0, M0; Stage II – T2, N0,M0;
Stage III – T3,N0,M0 or T1,N1a,N0; T2,N1a, M0; T3,N1a,M0
Stage IVA – T4a,N0,M0; T4a, N1a, M0; T1,N1b,M0; T2,N1b,M0; T3,N1b,M0; T4a,N1b,M0
Stage IVB – T4b, any N; M0
Stage IVC – Any T, any N, M1

American Thyroid Association for predicting the risk of recurrence

Low risk—90% of patients

- Microscopic tumour clearance
- No spread to node/blood
- No capsular/vascular invasion
- Extrathyroid I_{131} uptake is not found
- 25 year mortality is 2%

Intermediate risk

- Microscopic perithyroid invasion
- Vascular invasion

High risk

- Incomplete tumour removal
- Macroscopic invasion
- Distant spread
- Extrathyroid bed I131 uptake present
- 25 years mortality is 46%

- Total/near total thyroidectomy, central compartment dissection ± MRND, suppressive L-thyroxine radioactive iodine therapies are different modalities of treatment.
- They usually carry good prognosis.

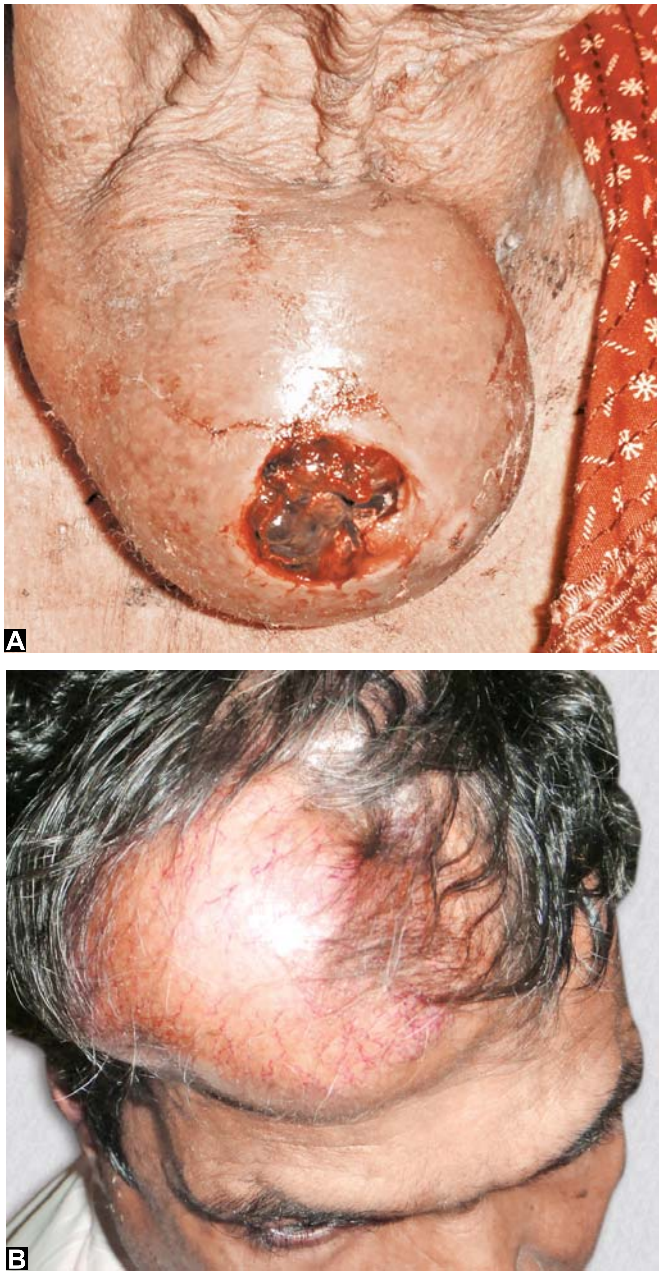

Figs 6.42A and B: (A) Fungating follicular carcinoma thyroid in a female. (B) Pulsatile vascular skull secondaries from FCT in a male patient.

- Tall/columnar cell, trabecular, scirrhous, solid types of papillary carcinoma; oxyphilic, insular types of FCT carry poor prognosis.

PAPILLARY CARCINOMA OF THYROID (PCT)

- It is 60% common.
- Common in females and younger age group.

Aetiology

- Radiation either external or radioactive iodine therapy.

TSH levels in the blood of these patients are high and so it is called as *hormone dependent tumour.*

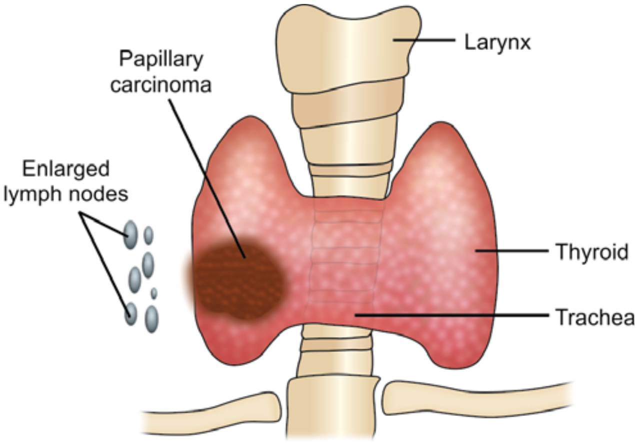

Fig. 6.43: Papillary carcinoma of thyroid with nodal infiltration—diagrammatic representation.

Woolner Classification

Types

i. Occult primary (< 1.5 cm).
ii. Intrathyroidal.
iii. Extrathyroidal.

Micropapillary carcinoma is a tumour, clinically not detectable or less than 1 cm.

- *Encapsulated variant of papillary carcinoma,* like adenoma shows capsule with local invasion, commonly nodal spread; it often mimics hyperplastic nodule; but this type carries good prognosis.
- *Diffuse sclerosing variant* is seen in children which is very aggressive type with lymphocytic infiltration; it shows near 100% nodal spread; it carries poor prognosis.
- *Encapsulated papillary variant* and follicular carcinoma may be found together as *papillofollicular (Lindsay tumour)*; it behaves like papillary carcinoma with good prognosis.

Gross

It can be soft, firm, hard, cystic. It can be solitary or multinodular. It contains brownish black fluid.

Microscopy

- It shows cystic spaces, papillary projections with *psammoma* bodies, malignant cells with '*Orphan Annie eye'* nuclei (intranuclear cytoplasmic inclusions).
- *Orphan Annie eye nuclei* are identified in histology (paraffin section of formalin tissue). It is not seen in FNAC. Orphan Annie is strip cartoon character with empty circled eyes.
- Tall cell type of papillary carcinoma (10% of papillary carcinomas) is very aggressive type seen in elderly, 30% show capsular and vascular invasion, with 25-30% 5 year survival rate.
- Columnar type is seen only in males with near 100% mortality in 5 years.

Spread

- It is a slowly progressive and less aggressive tumour.
- It is commonly multicentric.
- It spreads within the gland through intrathyroidal lymphatics to other lobe, comes out of the capsule and spreads to cervical lymph nodes.
- Usually there is no blood spread. Extrathyroidal disease—invasion into thyroid capsule can cause blood born secondaries occasionally.

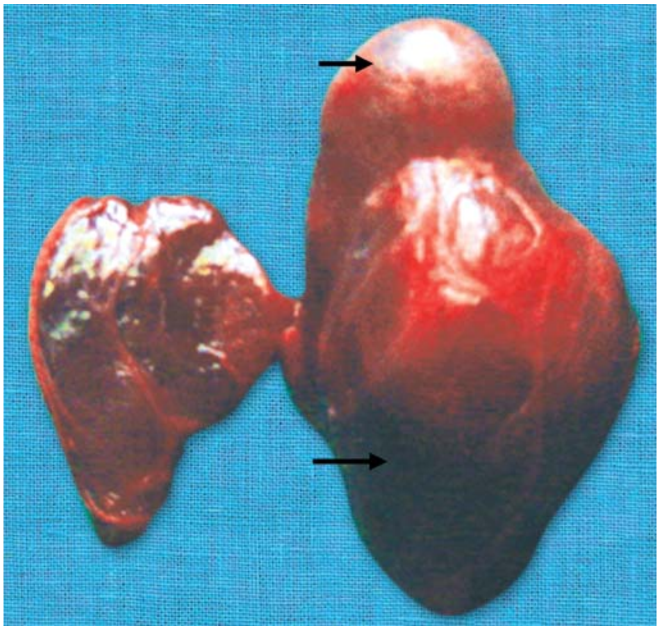

Fig. 6.44: Papillary carcinoma thyroid, operated specimen.

Clinical Features

- Soft or hard or firm, solid or cystic, solitary or multinodular thyroid swelling.
- Compression features are uncommon in papillary carcinoma thyroid.
- Often discrete lymph nodes in the neck (40%) are palpable.
- May present with secondaries in neck lymph nodes with occult primary.

Diagnosis

- *FNAC* of thyroid nodule and lymph node.
- Radioisotope scan shows cold nodule.
- TSH level in the blood is higher.
- Plain X-ray neck shows fine calcification whereas nodular goitre shows coarse—ring/rim calcification.
- U/S neck or CT scan neck (better) to identify non- palpable nodes in neck.

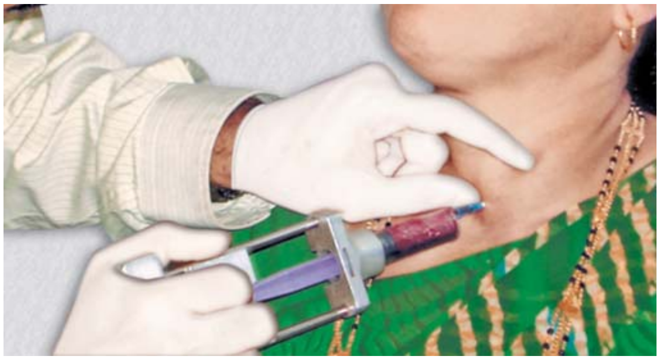

Fig. 6.45: FNAC of thyroid *(Courtesy:* Dr Krishna Upadhya Pathologist, Nandikoor Laboratory, Mangalore).

Treatment

- *Total or near total thyroidectomy*, with central node compartment dissection (level VI).
- *Suppressive dose* of L-thyroxine 0.3 mg OD life long. TSH level should be < 0.1 m U/L.
- *MRND (modified radical neck dissection)* type III (functional neck dissection), is required if lymph nodes are involved.
- Occasionally if small lymph nodes are present, '*Berry picking'* may be done (universally not accepted). (Not done now).
- Present concept is extrathyroidal type also responds well to radioactive I^{131} therapy.

***Note*:**

- If tumour is < 1.0 cm, solitary, low grade, probably unicentric, hemithyroidectomy is done with proper follow up at regular intervals.
- Suppressive dose of L-thyroxine can cause osteoporosis and so often needs calcium and Vitamin D supplementation.

Prognosis

Prognosis is good and it is one of the curable malignancies.

AMES scoring
A : Age. Age less than 40 years has got better prognosis
M : Distant metastasis
E : Extent of the primary tumour
S : Size of the tumour. Size less than 4 cm has got better prognosis

AGES scoring
A : Age less than 40 years has got better prognosis
G : Pathologic grade of the tumour
E : Extent of the primary tumour
S : Size of the primary tumour. Size less than 4 cm has got better prognosis

Note:

- Lymph node status does not alter the prognosis of papillary carcinoma of thyroid.
- MACIS scoring system is metastases, age, and completeness of excision, invasion, and size.
- All scoring systems categorise the patients as *high risk* for death—40% in 20 years; low risk for death—1% in 20 years. 80% of patients are in low risk for death. Low death risk is achieved by complete clearance of macroscopic tumour without any possible retaining of the disease during first surgery.
- *Follow-up:* Regular estimation of TSH levels, helps to monitor recurrence.

Psammoma bodies are seen in:

- Papillary carcinoma thyroid
- Meningioma
- Serous cystadenoma of ovary

Berry's in thyroid

- Berry ligament
- Berry sign
- Berry picking

Thyroid paradox: *Cellular tumours are soft, and cystic tumours are firm or hard (tensely cystic). It is observed in papillary carcinoma of thyroid.*

Role of ultrasound (U/S) in thyroid diseases

- To detect number, size, nature of the nodules (cystic/solid/complex) (complex means cystic and solid together—more suspicious of carcinoma). Size up to 2 mm can be detected
- U/S guided FNAC is very useful
- U/S at regular intervals is advisable to observe a small nodule in thyroid
- To detect recurrent nodule
- To find out the invasion/spread/vascularity/status of carotid artery and internal jugular vein
- To find out enlarged lymph nodes in neck

Role of FNAC in thyroid swelling

- Highly sensitive in papillary carcinoma of thyroid and also its nodal spread
- Useful to differentiate between benign and malignant
- Should be done in all thyroid diseases especially when there is a nodule or multiple nodules
- Useful in lymphoma/anaplastic carcinoma/medullary carcinoma thyroid/Hashimotos thyroiditis
- It is not very useful in follicular carcinoma as it is difficult to differentiate it from follicular adenoma as main feature in follicular carcinoma is capsular invasion/vascular invasion which are not made out in FNAC

FOLLICULAR CARCINOMA OF THYROID (FCT)

- It is 17% common.
- It is common in females.
- It can occur either *de novo* or in a pre-existing multinodular goitre.
- Thyroglobulin immunostaining is positive.

Types

a. Noninvasive—blood spread not common.

b. Invasive—blood spread common.
Typical features: Capsular invasion and angioinvasion.

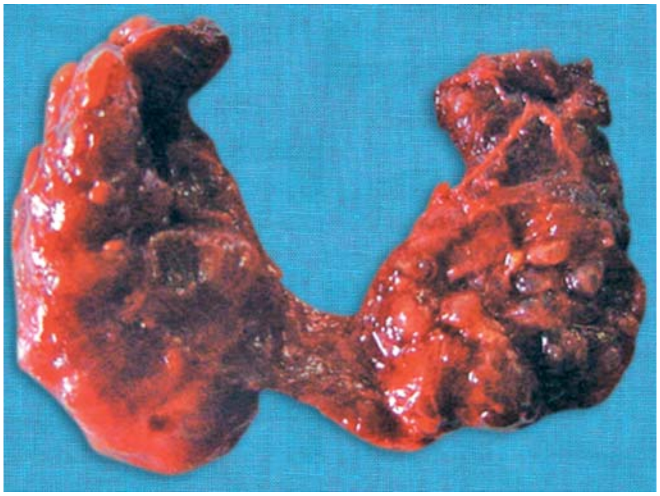

Fig. 6.46: Total thyroidectomy specimen—done for carcinoma thyroid.

Spread

- It is a more aggressive tumour.
- It spreads mainly through blood into the bones, lungs, liver.
- Bone secondaries are typically vascular, warm, pulsatile, localised, commonly in *skull,* long bones, ribs.
- It can also spread to lymph nodes in the neck occasionally (10%).

Clinical Features

1. Swelling in the neck, firm or hard and nodular.
2. Tracheal compression/*infiltration* and stridor.
3. Dyspnoea, haemoptysis, chest pain when there are lung secondaries.
4. Recurrent laryngeal nerve involvement causes hoarseness of voice, +ve '*Berry's sign*' signifies advanced malignancy (infiltration into the carotid sheath and so absence of carotid pulsation).
5. Pulsatile secondaries in the skull and long bones.

Investigations

- Most often FNAC is inconclusive, because capsular and angioinvasion, which are the main features in follicular carcinoma, cannot be detected by FNAC.
- Frozen section biopsy was said to be useful earlier; but it is questionable now. In 15% cases frozen section biopsy may be inconclusive or facility for frozen section biopsy may not be available in many places, then initial hemithyroidectomy is done.
- U/S abdomen, chest X-ray, X-ray bones are the other investigations required.
- *Trucut biopsy* gives tissue diagnosis, but danger of haemorrhage and injury to vital structures like trachea, recurrent laryngeal nerve, vessels are likely. It may be useful in lymphoma and anaplastic carcinoma; but it is *not very well accepted.*

Treatment

- *Total thyroidectomy is done* along with central node compartment dissection (level VI).
- *Functional neck node dissection* is done with preservation of sternocleidomastoid, spinal accessory nerve and IJV if nodes are present clinically or by imaging, FNAC shows positive node, and on table frozen section biopsy of lymph node shows positive for tumour.
- Maintenance dose of L-thyroxine 0.1 mg OD or T_3 80 μg/day is given lifelong. Immediate thyroxin supplementation is often not started following surgery to keep TSH level raised

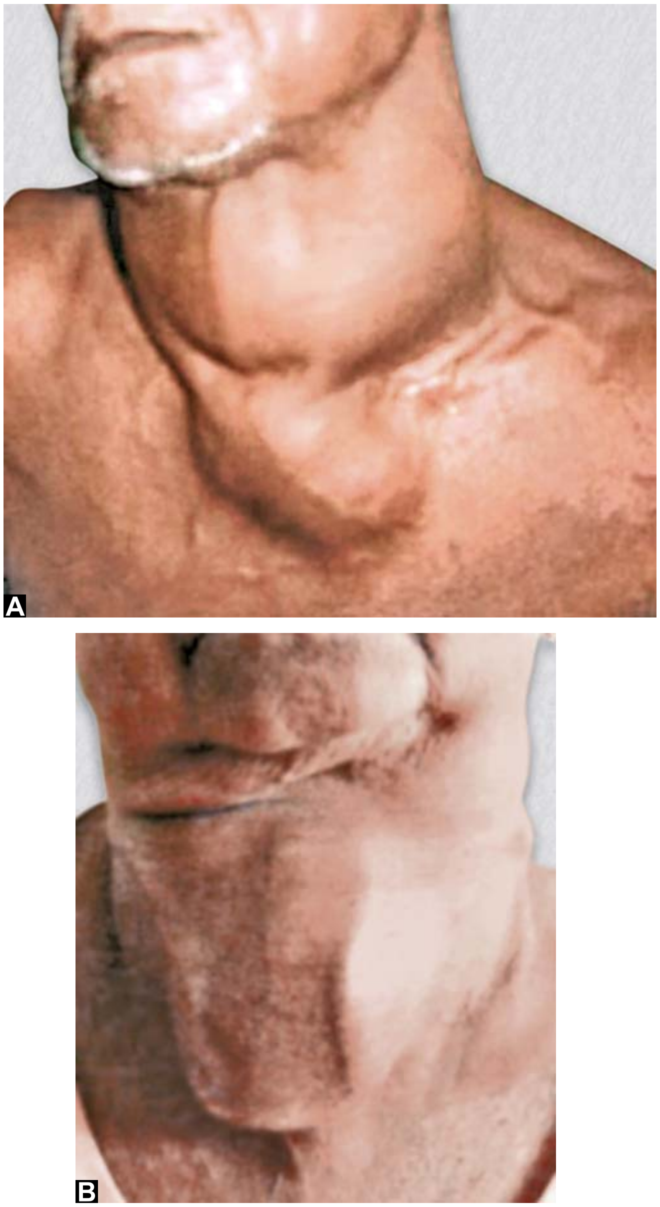

Figs 6.47A and B: Thyroid swelling which is highly vascular and malignant in two different patients (male and female).

Voice can say both: What you have done and what you would not have done also !!

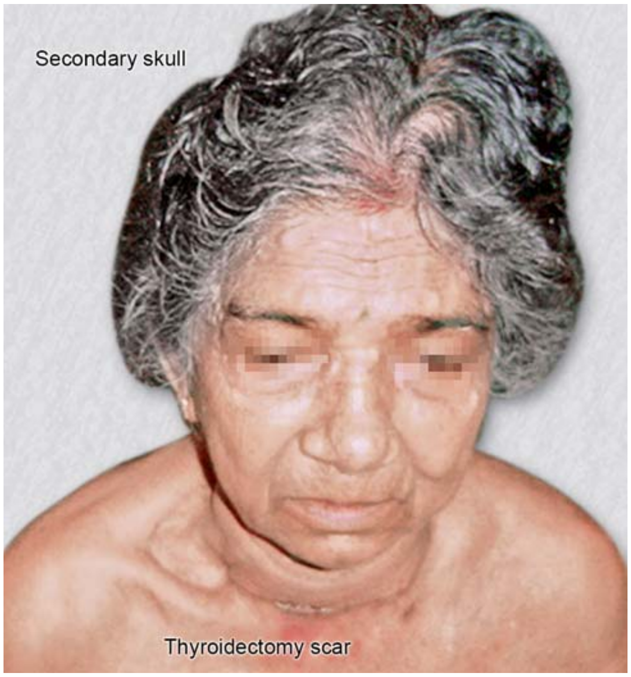

Fig. 6.48: Secondaries in the skull in a patient who underwent thyroidectomy earlier for follicular carcinoma thyroid.

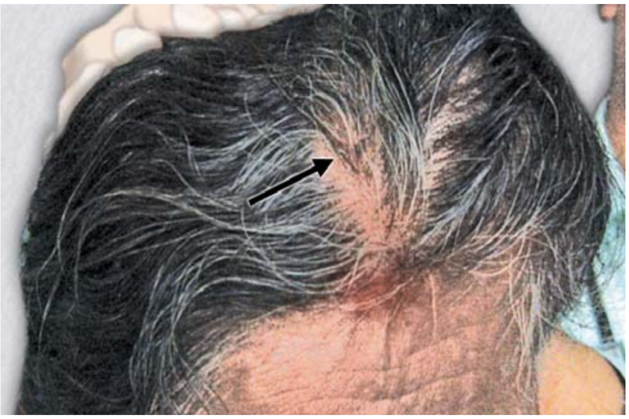

Fig. 6.49: Secondaries in the skull from thyroid primary. Note well-localised, warm, pulsatile, vascular tumour with underlying bone erosion. It is common in flat bone like skull/ribs/sternum/scapula. Most common in frontoparietal part of skull.

so that all extrathyroidal tissues that takes up iodine will also take up radioiodine to achieve optimum radioablation. TSH level should be more than 30 mIU/L for this.

- *On table frozen section biopsy* is useful in negative FNAC but doubtful cases. Definitive procedure is undertaken once frozen section report comes on table. But in frozen section biopsy itself, 15% of follicular carcinoma report may be inconclusive or negative which causes difficulty in taking decision. In such occasion hemithyroidectomy is done and once histology report of follicular carcinoma is obtained completion thyroidectomy is done usually immediately within a week. If biopsy report is delayed then completion thyroidectomy is done after 6-12 weeks.

Follow-up

- It is by radioisotope I^{123} scan done at regular intervals (6 months) to look for secondaries.

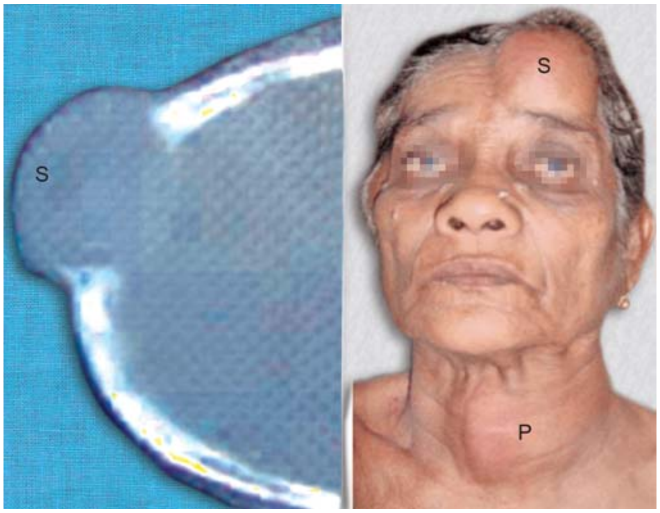

Fig. 6.50: CT skull showing secondaries from follicular carcinoma thyroid.

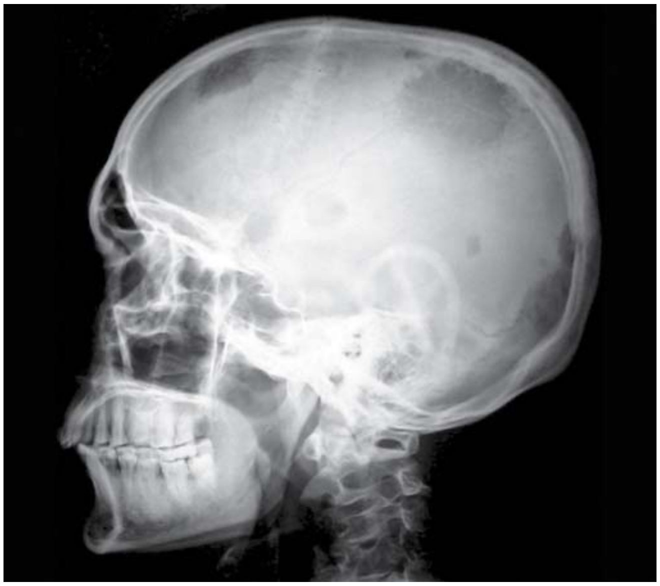

Fig. 6.51: X-ray skull showing secondaries with lytic lesions from follicular carcinoma of thyroid as primary.

- ***Thyroglobulin estimation*** is a good follow-up method to decide for Radioisotope study. Normal value (0.5-50 µg/L). TG > 50 µg/L is abnormal. It should be done once in 3-6 months. Its raise signifies recurrent/metastatic disease. Serum thyroglobulin level estimation is of no value in preoperative assessment. After thyroidectomy thyroglobulin secretion is stopped, hence its level should not be traceable after total thyroidectomy.
- U/S neck or MRI neck to identify early relapse. MRI neck is better.

Further Treatment

- If secondaries are detected *therapeutic dose* Ra I^{131} is given orally. L-thyroxin has to be stopped 6 weeks prior to RT, then required dose of Ra I^{131} is given (50-150 m curie).
- Radio remnant ablation (RRA)
  - Initially thyroid radioisotope scan is done. If patient is on 300 µg suppressive dose of L-thyroxine, it should

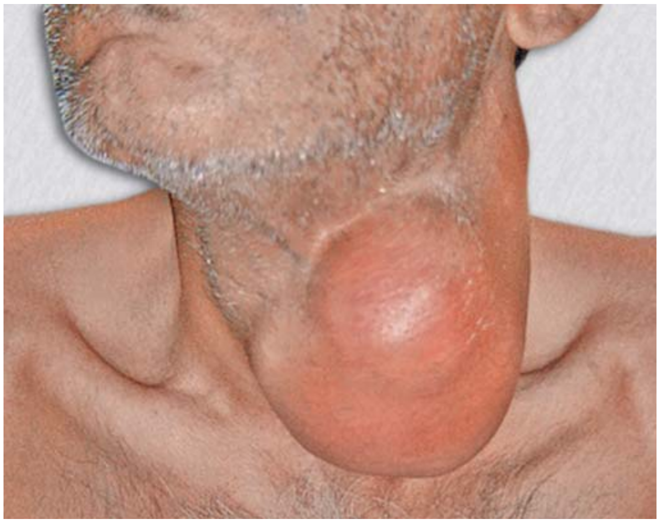

Fig. 6.52: Follicular carcinoma of thyroid in a male patient. It involved mainly left lobe but it is extensive and spreading to soft tissues adjacent to it.

be stopped for 3-6 weeks so as to achieve serum TSH level above 30 M IU/l or two intramuscular injections of 0.9 mg of recombinant human TSH is given.

- If iodine isotope scan detects remnant disease, then 30-100 mCi of radio ablation dose of I_{131} is given orally. RR will destroy occult microscopic disease. Post-ablation isotope scanning is done in 8 days. Later in intermediate and high risk patients, follow up body scan is done once in 6 months. Serum thyroglobulin estimation is done for 6-12 months, later once in 6 months; neck US at 6, 12 months and annually for 5 years. Raise in thyroglobulin level indicates for further whole body isotope scan. To detect metastases radioisotope I_{131} 100-200 mCi is given and 250-300 mCi is given for bone.
- *Prior to radioactive iodine therapy following precautions* should be taken—low iodine diet for 10 days; isolation when dose is more than 30 mCi; increase intake of oral fluid to maintain high urine output to avoid radio iodine induced bladder injury; sucking of lemon to avoid sialadenitis; laxatives; reduced sperm count for 6-8 months; pregnancy should be avoided for 6-12 months.

- Secondaries in bone are treated by *external radiotherapy.* Internal fixation should be done whenever there is pathological fracture.
- There is *no role* of chemotherapy for follicular carcinoma thyroid.

Hurthle Cell Carcinoma

Hurthle cell carcinoma is a variant of follicular carcinoma of thyroid which contains *abundant oxyphill cells*. It spreads more commonly to regional lymph nodes than follicular carcinoma of thyroid.

Note:

- Hurthle cell carcinoma does not take up I^{131}
- It secretes thyroglobulin
- It has got poorer prognosis than follicular cell carcinoma
- 30% multicentric
- *99mTc sestamibi scan* is very useful for Hurthle cell carcinoma
- Regional nodes are commonly involved than follicular carcinoma
- 20% show distant spread
- Abundant oxyphill cells (*Askanazy*) are specific
- Total thyroidectomy, MRND and TSH suppression is the treatment

Central Node Compartment Neck Dissection

It is removal of paratracheal, tracheooesophageal, pretracheal and prelaryngeal nodes along with thymus and thyroid enbloc *extending* from hyoid bone above, brachiocephalic vein below, carotids on both sides.

Completion Thyroidectomy

- It is done after hemithyroidectomy if histology confirms as differentiated thyroid cancer, either papillary or follicular.
- It is indicated in—high-risk patients, age more than 45 years, family history, size more than 1 cm, opposite side thyroid nodule, nodal or distant spread, multifocal disease, extrathyroidal spread, capsular and vascular invasion.
- Reasons for completion total thyroidectomy are:
  - In differentiated thyroid cancer (for follicular carcinoma of thyroid) radioiodine I_{131} is the treatment for blood spread secondaries. It is only possible if entire thyroid gland is removed to make radioiodine to concentrate on tumour tissue to achieve needed efficacy.
  - In many indications papillary carcinoma of thyroid also needs radioablation like extrathyroid spread, size more than 4 cm etc.
  - During follow up, in DTCs/FCT thyroglobulin as tumour marker estimation at regular intervals (6 months; to suspect tumour bed recurrence/metastases) can be done if thyroid tissue is removed entirely.
  - Papillary carcinoma of thyroid are multicentric with intrathyroid spread and so completion thyroidectomy is better.
- Timing of completion thyroidectomy is within 7 days of first surgery or after 6-12 weeks to allow the settling of inflammatory response (dissection and to get surgical planes is difficult during this period).

ANAPLASTIC CARCINOMA OF THYROID

- It is an *undifferentiated* very aggressive carcinoma (1%) occurs in elderly females.

Only in advanced cases of hypothyroidism there is a substantial loss of hair.—**Richard A J Asher**

- It is a very aggressive tumour of short duration, presents with a swelling in thyroid region which is rapidly progressive causing:
  - i. Stridor and hoarseness of voice due to tracheal obstruction.
  - ii. Dysphagia.
  - iii. Fixity to the skin.
  - iv. *Positive Berry's sign*—involvement of carotid sheath leads to absence of carotid pulsation.
- Swelling is hard, with involvement of isthmus and lateral lobes.
- FNAC is diagnostic.
- Tracheostomy and isthmectomy has got a role to relieve respiratory obstruction temporarily.
- Treatment is *external radiotherapy*, as usually thyroidectomy is not possible.
- Adriamycin as chemotherapy.
- However prognosis is poor. Life span is counted in few weeks to months only.

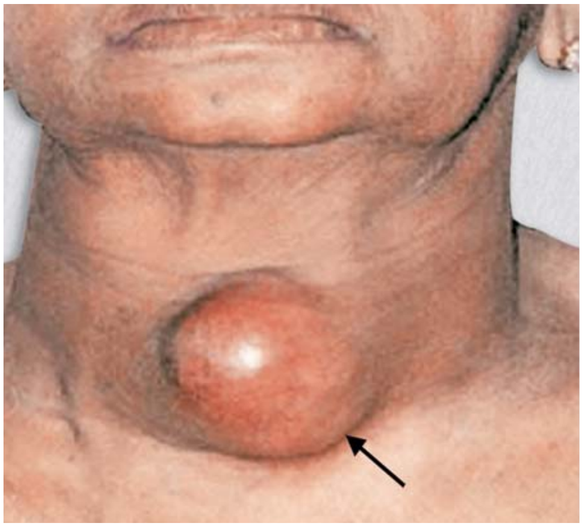

Fig. 6.55: Anaplastic carcinoma of thyroid.

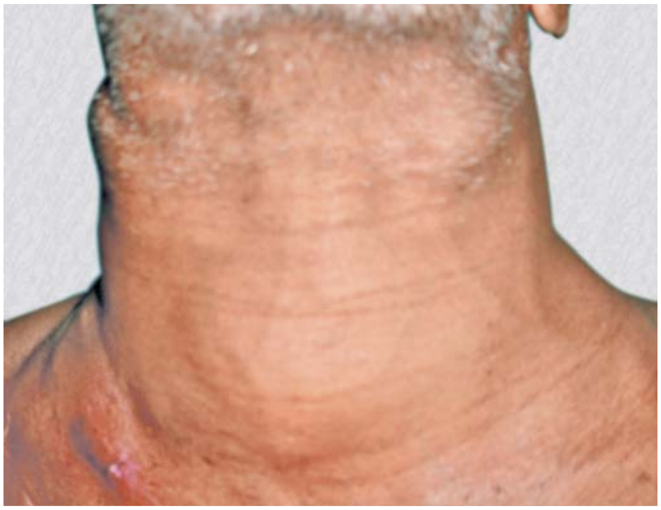

Fig. 6.53: Carcinoma of thyroid gland. It is anaplastic carcinoma. It carries poor prognosis.

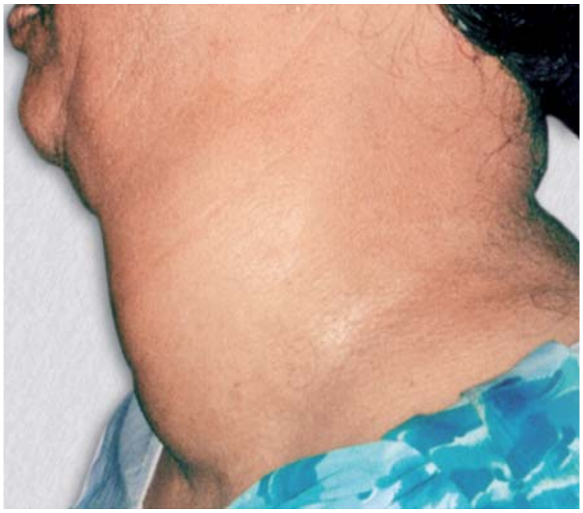

Fig. 6.56: Anaplastic carcinoma thyroid in a female patient.

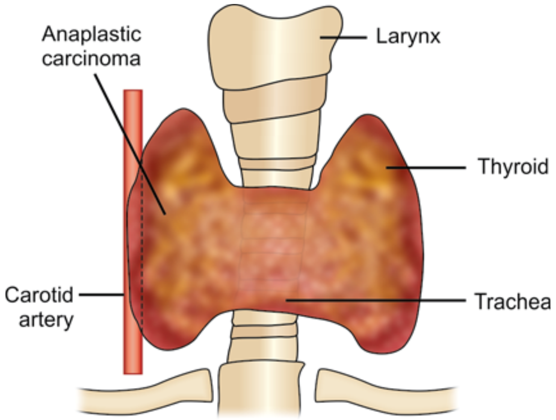

Fig. 6.54: Anaplastic carcinoma of thyroid with infiltration through the capsule. It commonly encases the carotid artery causing Berry's sign positive.

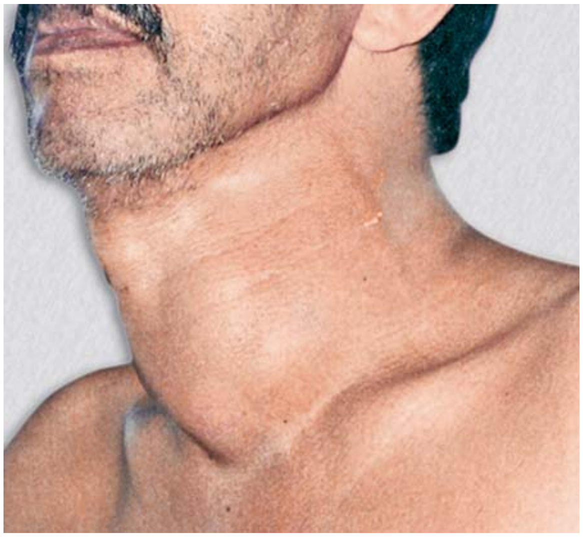

Fig. 6.57: Anaplastic carcinoma thyroid, an aggressive tumour with poor prognosis.

Follow-up in differentiated thyroid carcinoma

- Proper clinical examination in the neck for residual/nodal disease and for distant spread
- Whole body radioisotope scan after one week of surgery to see residual tumour in the neck or metastases
- Estimation of thyroglobulin at regular intervals is very important—once in 6 months
- Follow up whole body radioisotope scan at 3-6 months intervals. Thyroxine should be stopped for 6 weeks. But it is commonly done if thyroglobulin level in the blood is significantly high

STAGE GROUPING FOR PAPILLARY OR FOLLICULAR THYROID CARCINOMA (DIFFERENTIATED THYROID CANCER)

Patients younger than 45 years:

Stage I: The tumour can be any size and may or may not have spread to nearby lymph nodes. It has not spread to distant sites.

Stage II: The tumour can be any size and may or may not have spread to nearby lymph nodes. It has spread to distant sites.

Patients 45 years and older:

Stage I: The tumour is less than 2 cm across; no lymph nodes/distant spread.

Stage II: The tumour is 2 to 4 cm across; no lymph nodes/distant spread.

Stage III: The tumour is larger than 4 cm or slightly outside the thyroid; no lymph nodes/distant spread. OR The tumour is any size and has spread to lymph nodes around the thyroid in the neck (cervical nodes) but not to distant sites.

Stage IVA: The tumour is any size and has grown beyond the thyroid gland to invade nearby tissues of the neck. Cervical nodes ± but no distant spread. OR The tumour is any size and may have grown outside of the thyroid gland. It has spread to lymph nodes in the side of the neck (lateral cervical nodes) or upper chest (upper mediastinal nodes) but not to distant sites.

Stage IVB: The tumour is any size and has grown either back to the spine or into nearby large blood vessels. Nearby lymph nodes ± but no distant spread.

Stage IVC: The tumour is any size and may or may not have grown outside the thyroid. It may or may not have spread to nearby lymph nodes. It has spread to distant sites.

Stage grouping for medullary thyroid carcinoma

It is same as of differentiated thyroid cancer older than 45 years.

Stage grouping for anaplastic/undifferentiated thyroid carcinoma

All anaplastic thyroid cancers are considered ***stage IV***

Stage IVA: Tumour is still within the thyroid and may be *resectable* (removable by surgery). Nearby lymph nodes ± but no distant spread.

Stage IVB: Tumour has grown outside of the thyroid and is *not resectable*. Nearby lymph nodes ± but no distant spread.

Stage IVC: The tumour is any size and may or may not have grown outside of the thyroid. Nearby lymph nodes ± but distant spread *always present*.

Features of thyroid carcinoma

Any thyroid can be malignant of any size, of any texture—solid/cystic, of any number—single/multiple, in any age group.

Features of infiltration

- Infiltration of strap muscles often with sternomastoid muscle
- Infiltration of laryngotracheal complex causing stridor and often haemoptysis
- Infiltration of recurrent laryngeal nerve causes hoarseness of voice
- Infiltration of oesophagus causes dysphagia/ odynophagia (painful swallowing)
- Infiltration into carotid sheath causing absence of carotid pulsation—*Berry's sign*
- Infiltration of cervical sympathetic chain causing Horner's syndrome
- Rarely infiltration into cranial nerves or brachial plexus can occur

Features of lymph nodal spread

- Discrete neck node involvement can occur commonly in papillary carcinoma of thyroid, often in medullary carcinoma and occasionally in follicular carcinoma. Lymph node is often cystic (20%) and contains brownish— black material in papillary carcinoma
- Central neck (level VI) and mediastinal nodes often gets involved in thyroid malignancy. Primary nodes may be involved but clinically may not be palpable. Superior mediastinal nodes (level VII) can cause compression of SVC, recurrent laryngeal nerve with often dullness on percussion over the sternum. These nodes can get involved without palpable neck nodes
- In the neck commonly palpable nodes are levels—II, III, IV and occasionally level V. Secondary nodes—clinically palpable
- Only palpable neck node may be the presentation without clinically palpable thyroid—occult secondary with primary (papillary) thyroid carcinoma. FNAC of the node concludes the diagnosis
- Central node dissection is the common practice while doing total thyroidectomy in carcinoma thyroid especially in medullary carcinoma of thyroid

Features of blood spread

- Follicular carcinoma commonly spreads through blood to bone, lungs, and liver. Bone secondary is typical. It is well-localised, smooth, soft/hard, warm, nonmobile, vascular and pulsatile. It is common in the skull bone—frontal/parietal bone. It can occur in other bones also
- Lung secondaries present with chest pain, dyspnoea and haemoptysis
- Liver secondaries cause hepatomegaly and jaundice
- Blood spread also can occur in medullary carcinoma of thyroid

MEDULLARY CARCINOMA OF THYROID (MCT)

- It is uncommon (5%) type of thyroid malignancy.
- It arises from the *parafollicular 'C' cells* which is derived from the ultimobranchial body (neural crest). C cells are more in upper pole of the thyroid gland.
- It contains characteristic '*amyloid stroma'* wherein malignant cells are dispersed. Immunohistochemistry reveals calcitonin in amyloid.
- In these patients blood levels of *calcitonin* both basal as well as that following calcium or pentagastrin stimulation is high, a very useful tumour marker.
- Tumour also secretes 5-HT (serotonin), prostaglandin, ACTH and vasoactive intestinal polypeptide (VIP).
- It spreads mainly to lymph nodes (60%).

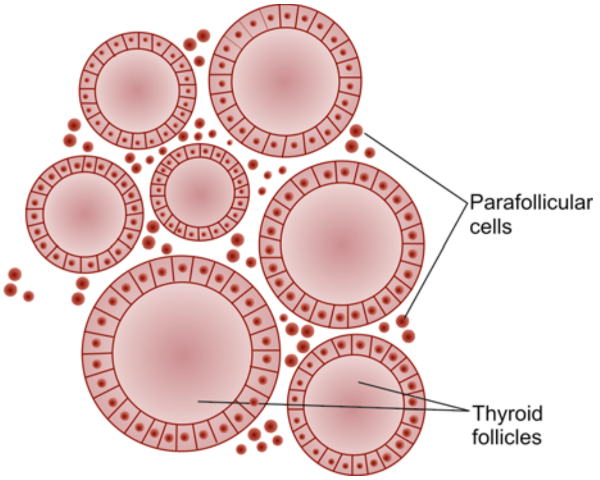

Fig. 6.58: Note the parafollicular cells. Cell of origin of medullary carcinoma thyroid.

- It may be associated with MEN II syndrome and pheochromocytoma with hypertension. MCT associated with MEN type II B with phaeochromocytoma (Sipple's disease) is most aggressive.
- There may be mucosal neuromas in lips, oral cavity, tongue, eyelids with marfanoid features.
- MCT is not TSH dependent and does not take up radioactive iodine.

Clinical Features

- Thyroid swelling often with enlargement of neck lymph node.
- Diarrhoea, flushing (30%).
- Hypertension, phaeochromocytomas and mucosal neuromas when associated with MEN II syndrome.
- Sporadic and familial types occur in adulthood whereas cases associated with MEN syndrome II occur in younger age groups.
- Paraneoplastic syndrome like Cushing's and carcinoids.

Note:
Consider MCT in thyroid swelling with significant diarrhoea and hypertension.

Types

1. Sporadic. Usually solitary—70%.
2. MCT with MEN II syndrome. It is more aggressive, bilateral, multifocal/multicentric, affects younger age group. MCT with MEN type IIb is most aggressive, often involves infants and children, with marfanoid features.
3. Familial MCT (20%). Commonly multicentric, autosomal dominant in chromosome no. 10. Familial non-MEN MCT is least malignant, occurs in 4th and 5th decades, shows extra and intracellular cysteine codon.

Investigations

- FNAC: shows amyloid deposition with dispersed malignant cells and "C" cell hyperplasia.
- Tumour marker: Calcitonin level will be higher. Unstimulated serum calcium more than 100 pg/ml is suggestive of MCT.
- Increased levels of calcitonin after injection of calcium 2 mg/kg or pentagastrin 0.5 µg/kg.

Note:
Calcitonin is undetectable in the serum of normal individuals (< 0.08 ng/L).

- U/S abdomen.
- U/S neck to see neck nodes. CT neck and chest is preferred method to identify neck and mediastinal nodes.
- Urinary VMA, urinary catecholamines, urinary metanephrine (24 hour estimation), serum calcium, serum parathormone estimation.
- 111 Indium octreotide scanning is useful in detecting medullary carcinoma thyroid (70% sensitivity). It is also useful in postoperative follow up to find out residual/metastatic disease.
- 50% of MCT shows rise in CEA.

Treatment

Surgery is the main therapeutic modality.

- *Total thyroidectomy with central node dissection* (level 6) in all patients even if there are no nodes in the neck + maintenance dose of L-thyroxin.
- Neck lymph nodes block dissection if lymph nodes are involved (bilateral modified radical dissection of neck nodes should include levels—II, III, IV, V and VI). Later regular U/S neck is done to detect early neck nodes.
- No role of suppressive hormone therapy or radioactive iodine therapy.
- External beam radiotherapy for residual tumour disease.
- Somatostatin/octreotide for diarrhoea.
- Adriamycin is the drug used as chemotherapy with limited results.
- If there is associated phaeochromocytoma it should be treated surgically by adrenalectomy first and later only total thyroidectomy is done.
- *All family members of the patient should be evaluated for serum calcitonin and if it is high they should undergo prophylactic total thyroidectomy* (Can also be assessed by genetic evaluation). If there is positive RET proto-oncogene in MCT

with MEN II A and familial MCT types, prophylactic total thyroidectomy is done at the age of ***5 years***. In positive RET proto-oncogene in MCT with MEN II B, prophylactic total thyroidectomy is done at the age of ***one year***.

- MCT with associated parathyroid hyperplasia (30%) in MEN IIA, total thyroidectomy with central nodal dissection with total parathyroidectomy and autotransplantation of half of gland in sternomastoid or nondominant forearm brachioradialis muscle is done.

Prognosis

- Sporadic MCT and MCT with MEN II are aggressive.
- Familial MCT not associated with MEN II has got better prognosis.

Calcitonin

- It is a polypeptide of 32 amino acids. It is derived from ultimobranchial body. It is secreted from C cells of thyroid (parafollicular cells).
- It lowers the plasma calcium and phosphorus levels. It blocks the PTH induced bone resorption. Calcium from the circulation is shunted into the bone. It increases the excretion of calcium, phosphorus, sodium and potassium. It rapidly lowers the serum calcium.
- Normally it is less than 0.08 ng/L (undetectable).
- It is increased in medullary carcinoma of thyroid. It is very good tumour marker for MCT. It confirms the relapse/metastases/residual disease. Increased levels in family members confirm the genetic relation and such relatives should undergo prophylactic total thyroidectomy.
- Calcitonin level will further increase after injection of calcium 2 mg/kg or pentagastrin 0.5 µg/kg.
- In disease free individual, after therapy for MCT calcitonin level decreases.
- Calcitonin as a therapeutic agent is used in hypercalcaemia, Paget‘s disease, bone pain of neoplastic diseases, menopausal osteoporosis. Calcitonin (pork), 4 units/kg is given SC or IM three times a week.
- Salmon calcitonin SC/IM/nasal spray—50-400 units can be given three times a week or daily depending on need.

MALIGNANT LYMPHOMA

- It is NHL type. Occurs in a pre-existing Hashimoto's thyroiditis (Not proved well).
- FNAC is useful to diagnose the condition (Often trucut biopsy).
- Chemotherapy and radiotherapy is the main treatment.
- Rarely total thyroidectomy is done to enhance the results.

Degroot's staging of carcinoma thyroid*

Stage 1: Malignancy is intrathyroidal
Stage 2: Malignancy is intrathyroidal with mobile lymph nodes
Stage 3: Malignancy is intrathyroidal with fixed lymph nodes
Stage 4: Distant metastasis

*This staging is applicable to all thyroid malignancy but mainly used in follicular carcinoma of thyroid

Differential diagnosis for carcinoma thyroid

- Multinodular goitre
- Solitary nodule of other causes
- Riedel's thyroiditis

HASHIMOTO'S THYROIDITIS (Struma Lymphomatosa)

- Also called as diffuse non-goitrous thyroiditis.
- It is an autoimmune thyroiditis—common in women (15 times more common).
- There is hyperplasia initially, then fibrosis, eventually infiltration with plasma cells and lymphocytic cells.
- *Askanazy cells* are typical (like Hurthle cells).
- The river Struma arises in Bulgaria and flows into Aegean Sea. Struma means goitre. Banks of this river are endemic area for goitre.

Features

- Painful, diffuse, enlargement of usually both lobes of thyroid which is firm, rubbery, tender and smooth (occasionally one lobe is involved).
- Initially they present with toxic features but later, they manifest with features of hypothyroidism.
  Hyperplasia → Hyperthyroid—*Hashitoxicosis*
  → Euthyroid.
  Fibrosis → Hypothyroid.
- There may be hepatosplenomegaly,
- It is often associated with other autoimmune diseases.
- In 85% cases significant rise in the thyroid antibodies (microsomal, thyroglobulin, or colloid antibodies) is observed.
- Often condition may be associated with or may predispose to *malignant lymphoma*. It is at present not well-proved.
- Common in perimenopausal females.
- Occasionally it can predispose to papillary carcinoma of thyroid also.

Investigations

FNAC, T_3, T_4, TSH. Thyroid antibodies assay. Usually ESR is very high (over 90 mm/hour).

Treatment

1. L-thyroxine therapy.
2. Steroid therapy often is helpful.
3. If goitre is large and causing discomfort, then subtotal thyroidectomy is done.

DE-QUERVAIN'S SUBACUTE GRANULOMATOUS THYROIDITIS

It is due to viral aetiology either mumps or coxsackie viruses causing inflammatory response with infiltration of lymphocytes, neutrophils, multinucleated giant cells.

Features

- Painful diffuse, swelling in thyroid which is tender
- Commonly seen in females.
- Initially there is transient hyperthyroidism with high T_3 and T_4 but poor radioiodine uptake.
- FNAC is useful.
- It is usually a self limiting disease.
- Prednisolone 20 mg for 7 days helps.

RIEDEL'S THYROIDITIS (0.5% Common)

- A very rare benign entity wherein thyroid tissue is replaced by fibrous tissue which interestingly infiltrates the capsule into surrounding muscles, paratracheal tissues, carotid sheath.
- ('*Woody Thyroiditis*', '*Ligneus Thyroiditis*').
- It is often associated with retroperitoneal and mediastinal fibrosis and sclerosing cholangitis.
- There is both intrathyroidal as well as extrathyroidal fibrosis.
- It also encroaches parathyroids and recurrent laryngeal nerves.
- It may be unilateral or bilateral.

Clinical Features

- Swelling with *irregular* surface, *stony hard* consistency, stridor, with positive *Berry's* sign (absence/*impalpable* carotid pulsation); small goitre; common in males.

Differential Diagnosis

- Anaplastic carcinoma of thyroid.

Investigations

- T_3, T_4 may be low due to hypothyroidism.
- Radioisotope scan will not show any uptake.
- FNAC to rule out carcinoma.

Treatment

Isthmectomy is done to relieve compression on the airway. They require L-thyroxine replacement later, as hypothyroidism is common.

- High dose of steroid often used.
- Thyroidectomy is not necessary.

Causes of dyspnoea/stridor in thyroid diseases

- Carcinoma thyroid infiltrating recurrent laryngeal nerve/ trachea
- Large, long standing goitre causing tracheomalacia
- Retrosternal goitre
- Congestive cardiac failure in thyrotoxicosis

Recent rapid increase in thyroid swelling is due to:

- Previous MNG undergoing malignant transformation
- Haemorrhage into a nodule
- Anaplastic carcinoma of thyroid

Remember

- *Goitre* is enlargement of the thyroid gland
- *Solitary nodule* is single palpable nodule on clinical examination without palpable rest of the gland
- *Dominant nodule* is single nodule with palpable enlargement of the remaining thyroid gland
- *Thyroid swelling* is confirmed by its *movement with deglutition* due to attachment of enclosed pretracheal fascia to inferior constrictor muscle which is attached to trachea and cricoid cartilage and so moves with deglutition
- *Berry's ligament* is condensed vascularised pretracheal fascia postero-supero-medially. It is important as it is close to recurrent laryngeal nerve
- Any thyroid swelling *can be malignant* unless proved otherwise
- U/S neck, FNAC, estimation of T3, T4, TSH are essential investigations
- CT scan neck is needed in large goitre and fixed or malignant thyroid
- Radioisotope study I^{123} is done only in selected cases like borderline toxicity, ectopic thyroid, retrosternal goitre and after thyroidectomy in follicular carcinoma thyroid to see secondaries during follow-up period
- Normal thyroid gland is usually not palpable
- A rare entity called as *black thyroid* shows lipofuscin deposition in thyroid in a patient who is on longstanding tetracycline therapy which may interfere with thyroid function

Occasions wherein thyroid swelling may not move upwards with deglutition

- Anaplastic carcinoma thyroid—often
- Carcinoma thyroid with extensive local infiltration into soft tissues, trachea/larynx and posterior muscles
- Intrathoracic retrosternal extension with infiltration/impaction
- Riedel's thyroiditis with encasement of trachea
- Massive thyroid wherein upward movement is difficult to observe and appreciate

Narrowing of trachea is seen in:

- Scabbard trachea in longstanding MNG
- Retrosternal goitre
- Carcinoma of thyroid
- Riedel's thyroiditis

THYROID INCIDENTALOMA

Clinically unsuspected and impalpable thyroid swelling which is identified while doing *imaging of head and neck for nonthyroid head and neck diseases is called as thyroid incidentaloma*. It is usually left alone with yearly review both clinically and sonologically. Such nodule is usually less than 1 cm.

THYROIDECTOMY

Types

1. *Hemithyroidectomy*: Along with removal of one lobe, entire isthmus is removed. It is done in benign diseases of only one lobe. It is also done in follicular neoplasm involving only one lobe. Solitary toxic or nontoxic nodule, thyroid cyst are other indications.
2. *Subtotal thyroidectomy* commonly done in toxic thyroid either primary or secondary and also often for nontoxic multinodular goitre. Here about 8 grams, or a tissue, size of pulp of finger is retained on lower pole, on both sides and rest of the thyroid gland is removed. It is also done in MNG.
3. *Partial thyroidectomy* (By Thomas) is removal of the gland in front of trachea after mobilisation. It is done in nontoxic multinodular goitre. Its role is controversial.
4. *Near total thyroidectomy*: Here both lobes except the lower pole (***one*** or both sides) which is very close to recurrent laryngeal nerve and parathyroid is removed (To retain blood supply to parathyroids). It is done in case of papillary carcinoma of thyroid. Here less than 2 grams of thyroid tissue is left behind near its lower pole *on one side usually opposite side of the diseased*, occasionally on both sides.
5. *Total thyroidectomy*: Entire gland is removed. It is done in case of follicular carcinoma of thyroid, medullary carcinoma of thyroid.
6. *Hartley Dunhill operation* is removal of one entire lateral lobe with isthmus and partial/subtotal removal of opposite lateral lobe. It is done in non-toxic multinodular goitre. 4 grams of tissue is left behind only on one side.

Preoperative preparation

- Blood grouping and cross matching. Keep the required blood ready
- Indirect laryngoscopy. Patient is asked to tell' E' to check the abduction of vocal cord
- Serum calcium estimation
- T_3, T_4, TSH
- Thyroid antibodies
- ECG and cardiac fitness especially in toxic goitre
- Lugol's iodine 10 days prior to surgery to make gland firm and less vascular.

Procedure

- *Position*:
  - Under general anaesthesia patient is put in supine position with neck hyperextended by placing a sand bag under shoulder—with table tilt of 15 degree head up to reduce venous congestion.
- *Incision*:
  - Horizontal crease incision is done, two finger breadth above the sternal notch, from one sternomastoid to the other (Kocher's thyroid incision) (Posterior margin of sternomastoid).
- *Procedure:*
  - Skin and platysma are incised (Subplatysmal plane)—upper flap raised up to thyroid cartilage, lower flap up to sternoclavicular joint. Deep fascia is opened vertically in the midline.

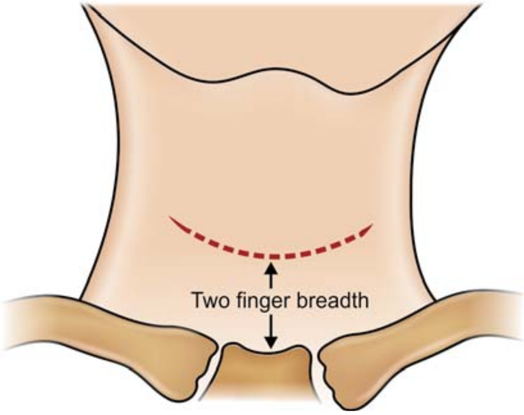

Fig. 6.59: Note the incision for thyroid surgery.

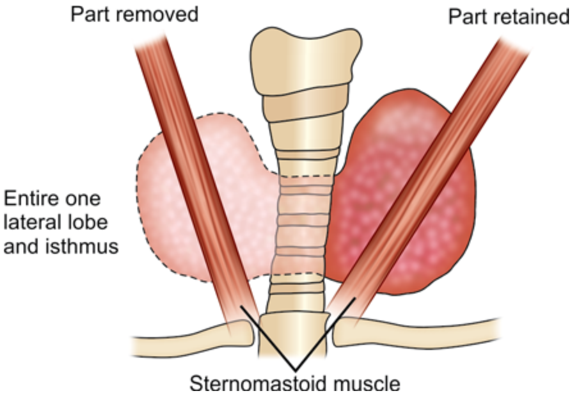

Fig. 6.60: *Hemithyroidectomy*: Entire one lateral lobe and entire isthmus are removed retaining entire opposite lateral lobe. It is done in solitary nodule/toxic or nontoxic adenoma in one lobe.

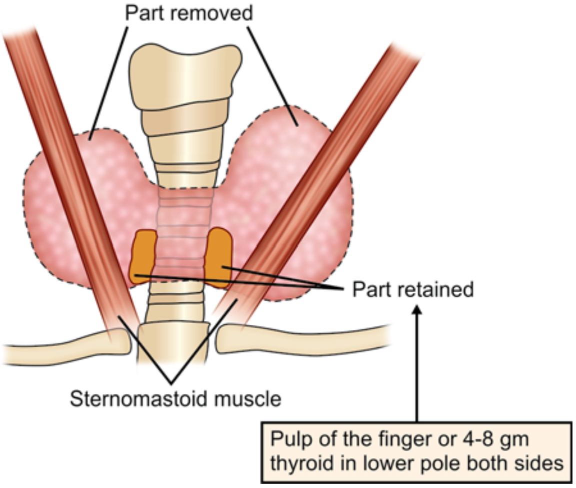
Part removed
Part retained
Sternomastoid muscle
Pulp of the finger or 4-8 gm thyroid in lower pole both sides

Fig. 6.61: *Subtotal thyroidectomy*—it is done in toxic/nontoxic multinodular goitre. Most of the gland except lower pole (4-8 gram) on both sides is removed.

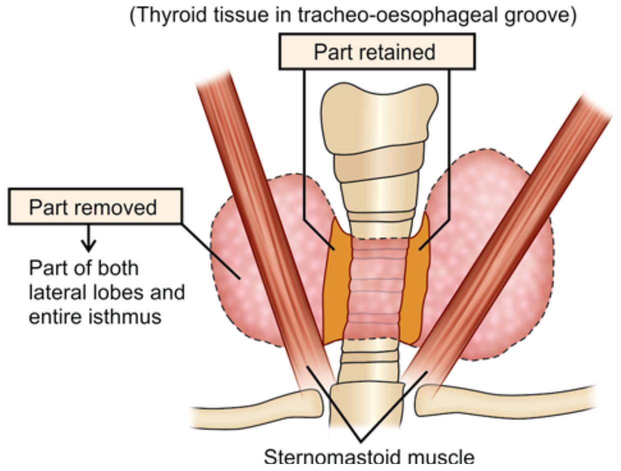
(Thyroid tissue in tracheo-oesophageal groove)
Part retained
Part removed
Part of both lateral lobes and entire isthmus
Sternomastoid muscle

Fig. 6.62: *Partial thyroidectomy*—it is done in nontoxic nodular goitre if there is adequate normal gland posteriorly. Tissue in the tracheo-oesophageal groove is retained. Isthmus and gland with nodules in front is removed. It is not commonly done now.

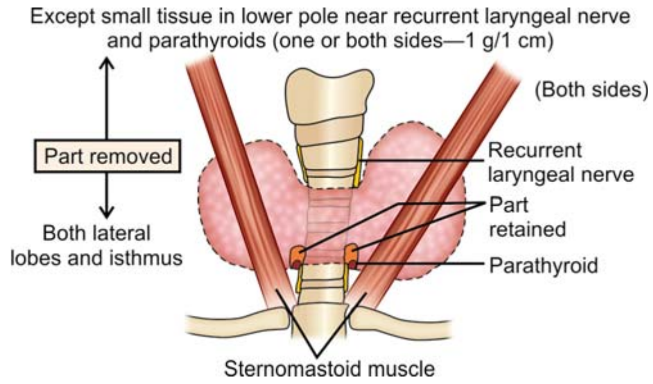
Except small tissue in lower pole near recurrent laryngeal nerve and parathyroids (one or both sides—1 g/1 cm)
(Both sides)
Part removed
Both lateral lobes and isthmus
Recurrent laryngeal nerve
Part retained
Parathyroid
Sternomastoid muscle

Fig. 6.63: *Near total thyroidectomy* is done in papillary carcinoma of thyroid. Here most of the gland except lower small tissue of 1 gram on one side usually is retained to safeguard recurrent laryngeal nerve and parathyroid gland. Tissue is retained either in one or both sides.

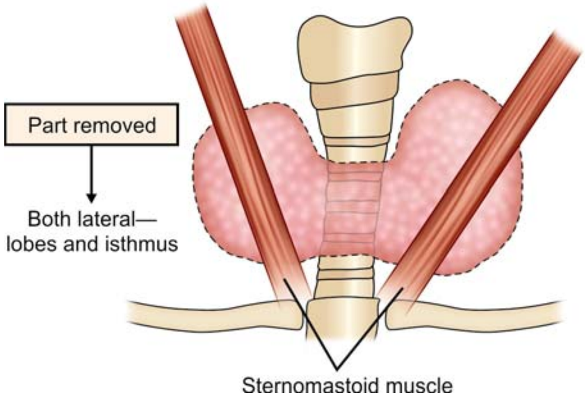
Part removed
Both lateral—lobes and isthmus
Sternomastoid muscle

Fig. 6.64: *Total thyroidectomy* is done for follicular carcinoma and medullary carcinoma of thyroid.

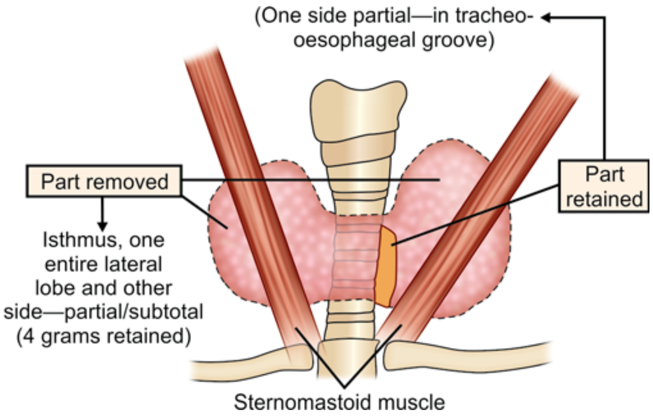
(One side partial—in tracheo-oesophageal groove)
Part removed
Isthmus, one entire lateral lobe and other side—partial/subtotal (4 grams retained)
Part retained
Sternomastoid muscle

Fig. 6.65: *Hartley Dunhill procedure:* Here one entire lateral lobe, isthmus, and most part of the opposite lateral lobe except small quantity of tissue in the lower pole/tracheo-oesophageal groove—subtotal/partial/one gram is retained.

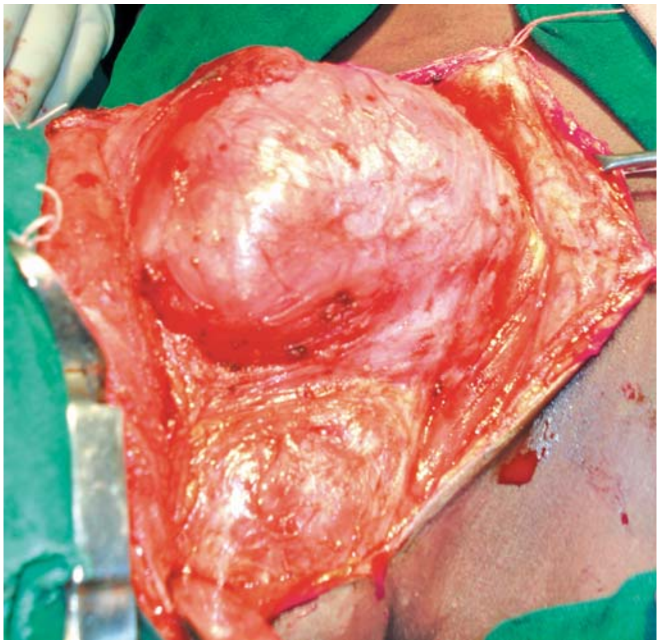

Fig. 6.66: Skin flaps raised in thyroidectomy. Thyroid covered with pretracheal fascia is seen.

- *Strap muscles* are retracted; in large goitre they are often divided in upper part to retain their nerve supply *ansa cervicalis*. Often anterior jugular veins need to be ligated using 3 zero vicryl.
- *Pretracheal fascia* is opened vertically to expose thyroid gland. Short stout middle thyroid vein which enters the IJV is ligated immediately (first vessel to ligate; it is vein without accompanying artery) using vicryl or silk and divided. Avulsion of this vein from its junction to IJV will cause bleeding. Gland is mobilised medially using peanut dissection and bipolar cautery.
- *Superior pedicle* is dissected; artery and vein are individually ligated and divided. In olden days mass liga-

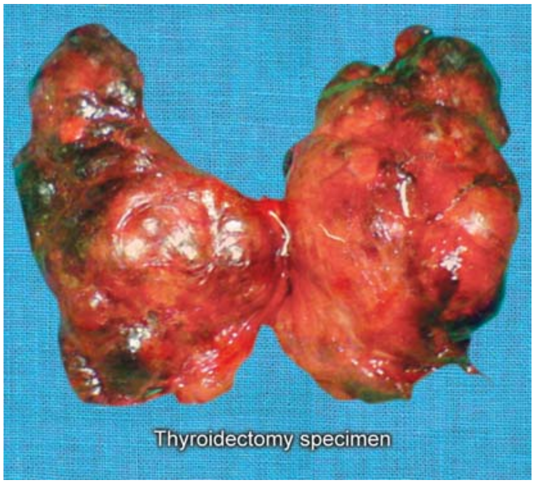

Fig. 6.67: Total thyroidectomy specimen done for follicular carcinoma of thyroid.

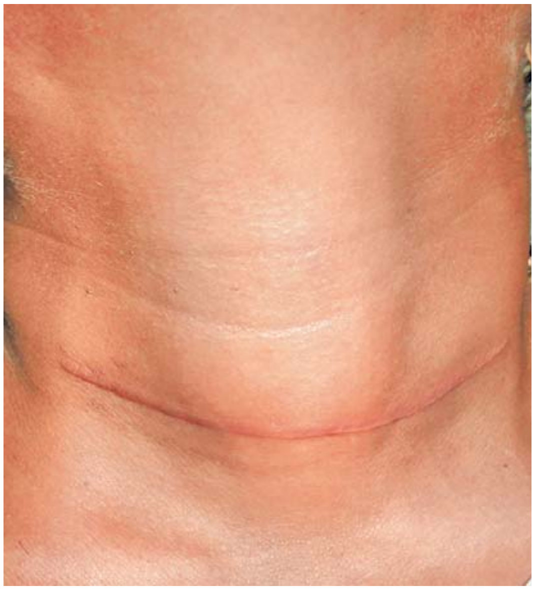

Fig. 6.68: Typical thyroidectomy postoperative scar.

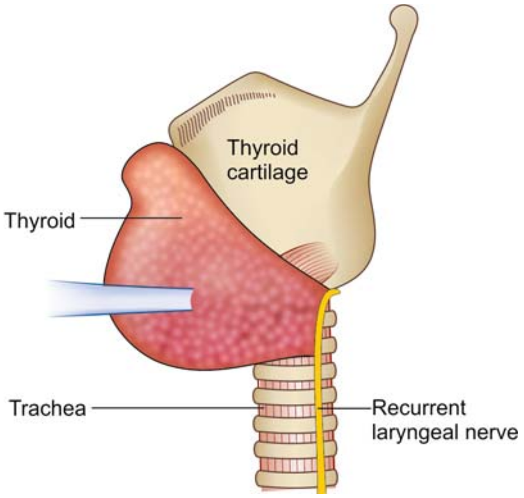

Fig. 6.69: Note the relation of recurrent laryngeal nerve.

tion close to gland at superior pole was the practice. Chances of injuring external laryngeal nerve and AV fistula may happen in mass ligation. It is also always better to identify external laryngeal nerve entering the cricothyroid. Dissection is done in an avascular plane between cricothyroid and gland.

- *Parathyroids* both superior and inferior are identified. They are 6 × 4 × 2 mm in size weighing 50 mg with yellowish brown/orange brown colour. Superior (parathyroid IV) and inferior (parathyroid III) glands are identified and dissected. Both glands receive their blood supply from inferior thyroid artery and through an anastomotic branch. Superior parathyroid is *above and behind* the junction of RLN and inferior thyroid artery; inferior parathyroid is *below and in front* of this junction.
- *Recurrent laryngeal nerve* should be identified with careful dissection through its entire course. *Riddle's triangle* is between inferior thyroid artery above, carotid artery laterally trachea medially. From this area nerve runs upwards to enter the larynx at greater cornu of thyroid cartilage. Many branches of nerve and the variations should be remembered while dissecting here. One should not use monopolar cautery here; only bipolar cautery should be used carefully. Nerve usually crosses the inferior thyroid artery from deeper aspect; but variations are common. Posterior extension of lateral thyroid lobes close to Berry's ligament is called as *Zuckerkandl tubercle* which is seen in 40% of cases. Nerve runs upwards in a fissure between Zuckerkandl tubercle and trachea or main thyroid gland. *Nonrecurrent laryngeal nerve* (0.5%) is occasionally seen in right side due to failure of development of 4th aortic arch; it directly arises from vagus enters the larynx at the level of inferior horn of thyroid cartilage. Recurrent laryngeal nerve is in close contact with *suspensory ligament of Berry*.
- *Inferior thyroid artery* which is a branch of thyrocervical trunk ascends upwards reaching gland at its lower pole after turning towards midline behind the carotid artery. Here ligation is done at capsular level by identifying every small branch entering the gland (*capsular ligation of inferior thyroid artery*). This retains the blood supply of parathyroids which is very important. In olden days, ligation of inferior thyroid artery was done away from the gland often in continuation using absorbable suture material, is now no longer in practice.
- Mobilized gland is removed. *Critical points of recurrent laryngeal nerve injury* are—at the entry of inferior thyroid artery and crossing the nerve, at suspensory ligament of Berry, at lower pole of the gland.

Thyroid steal: In thyrotoxicosis, patient is taken to operation theatre daily for few days before doing surgery, so as to reduce the anxiety of the patient.

MIVAI*—Minimally Invasive Video-Assisted Thyroidectomy* is becoming popular for small nodules and gland without thyroiditis. But it is costly.

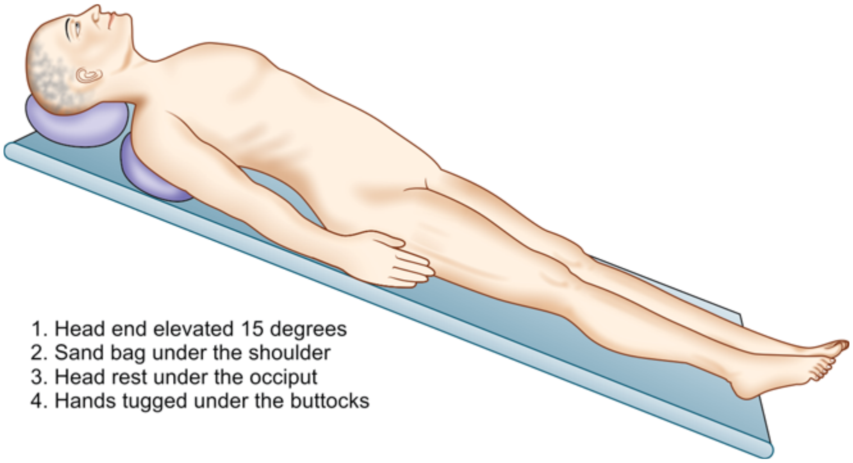
1. Head end elevated 15 degrees
2. Sand bag under the shoulder
3. Head rest under the occiput
4. Hands tugged under the buttocks

Fig. 6.70: Position of the patient for thyroid surgery. Head end elevated, sand bag under the shoulder.

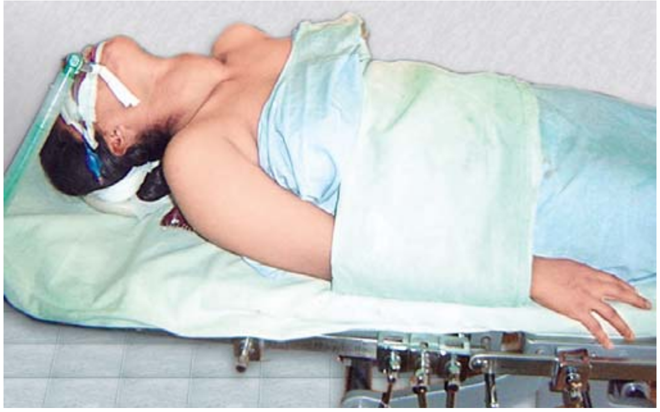

Fig. 6.71: Patient under anaesthesia with proper position for thyroidectomy surgery.

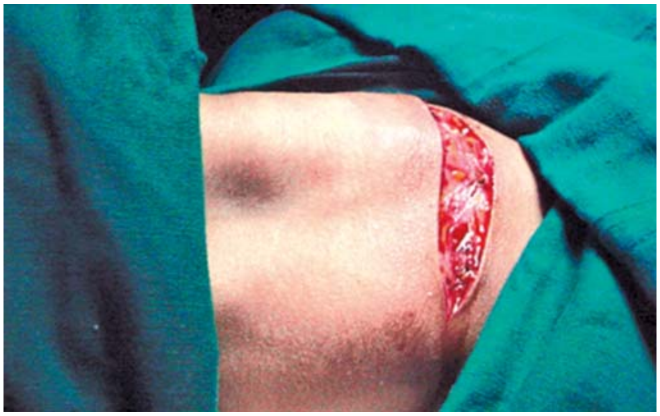

Fig. 6.73: Kocher's thyroid incision for thyroidectomy.

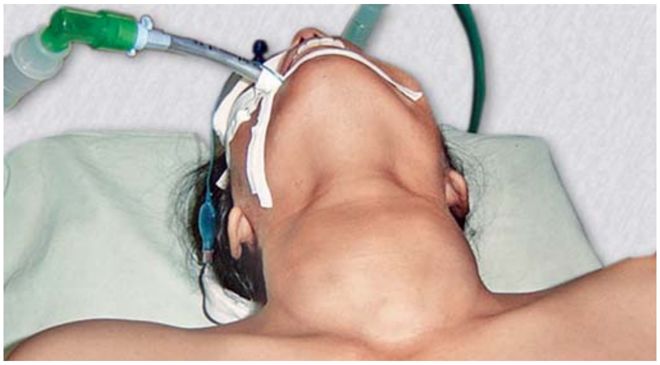

Fig. 6.72: Position of neck in thyroid surgery.

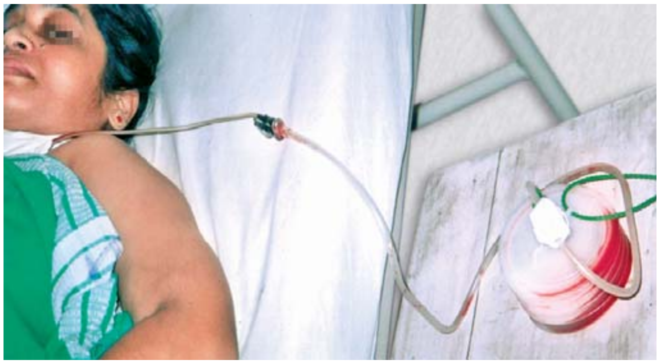

Fig. 6.74: Suction drain kept after thyroidectomy.

Complications of Thyroidectomy

1. *Haemorrhage:*

   May be due to slipping of ligatures either of superior thyroid artery or other pedicles or small veins. It causes tachycardia, hypotension, breathlessness and compression over the trachea may cause severe stridor, respiratory obstruction due to tension haematoma under strap muscles. As a first aid, immediate release of sutures including that of deep fascia has to be done and pressure over the trachea is released. Then patient is shifted to operation theatre and under general anaesthesia exploration is done and bleeders are ligated. Blood transfusion may be required.

2. *Respiratory obstruction*:

   It may be due to haematoma (if it is so, the haematoma has to be evacuated), or due to laryngeal oedema, or due to tracheomalacia or bilateral RLN palsy *(emergency endotracheal intubation* is done along with steroid injections). Often emergency tracheostomy may be required as a life-saving procedure.

   Laryngeal oedema is the commonest cause may be due to haematoma with tension along with intubation injury and surgical trauma. In such situation even though haematoma should be evacuated immediately to avoid further compression, airway obstruction will not be relieved due to laryngeal oedema and so endotracheal intubation is also

Complications of Thyroidectomy

Metabolic	***Nerve injuries***	***Other complications***
Hypoparathyroidism • Temporary hypoparathyroidism • Temporary hypocalcaemia without hypoparathyroidism (hungry bone syndrome) • Permanent hypoparathyroidism • Spurious hypoparathyroidism (total calcium is less but ionized calcium is normal) **Thyroid crisis** **Hypothyroidism / thyroid failure / myxoedema**	• External laryngeal nerve injury – pitch of the voice is lost • Recurrent laryngeal nerve injury **Vascular complications** • Haemorrhage – primary and reactionary • Haematoma formation • Compromised tracheooesophageal blood supply	• Respiratory obstruction • Skin flap necrosis – rare • Infection • Discharge and sinus formation • Wound granuloma • Keloid formation
Intraoperative	***Immediate postoperative***	***Late postoperative***
• Haemorrhage • Nerve injuries	• Haematoma • Laryngeal oedema • Respiratory obstruction – stridor • Hypoparathyroidism • RLN injury and its problems	• Hypothyroidism • Recurrent nodule • Recurrent toxicity • Wound infection • Granuloma / keloid

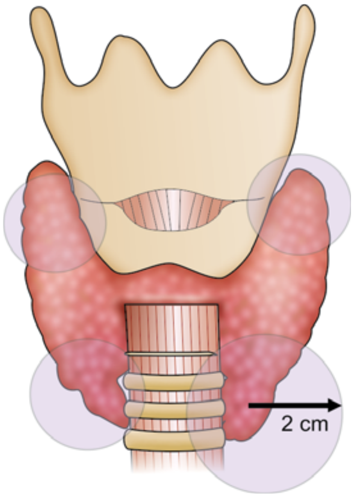

Fig. 6.75: Note the location of parathyroid glands.

needed immediately; in difficult situation if it is impossible to intubate, a needle tracheostomy using 12 gauge needle (2.3 mm diameter) is done temporarily.

3. *Recurrent laryngeal nerve palsy:*
   It can be transient or permanent. Transient is 3% common. They usually recover in 3 weeks to 3 months. Often they require steroid supplement and speech therapy. Permanent paralysis is rare. It presents with hoarseness of voice, aphonia, aspiration, ineffective cough. Unilateral injury can be well compensated. It is advisable to do routine postoperative ILS on 5th postoperative day. Later ILS is done in 4 weeks and 12 weeks.
4. *Hypoparathyroidism:*
   It is rare 0.5% common. Mostly it is temporary due to vascular spasm of parathyroid glands, occurs in 2nd-5th postoperative day. Present with weakness, +ve Chvostek's sign, carpopedal spasm, convulsions. Serum calcium estimation is done and then 10 ml of 10% calcium gluconate is given IV 8th hourly. Later supplemented by oral calcium carbonate 500 mg 8th hourly. After 3-6 weeks, patient is admitted, drug is stopped and serum calcium level is repeated.

   Note:
   Earliest symptom of hypocalcaemia is muscle weakness.
5. *Thyrotoxic crisis 2% (Thyroid storm):*
   Occurs in a thyrotoxic patient inadequately prepared for thyroidectomy and often a thyrotoxic patient presents in a crisis following an unrelated operation or stress. Other causes are—infection, trauma, preeclampsia, diabetic ketosis, emergency surgery, stress.
   Features: They present in 12-24 hours after surgery; with *severe dehydration*, *circulatory collapse, hypotension, hyperpyrexia*, tachypnoea, hyperventilation, palpitation, restlessness, tremor, delirium, diarrhoea, vomiting and cardiac failure; later coma.
   Bayley's symptom complex of thyroid storm are—insomnia, anorexia, diarrhea, vomiting, sweating, emotional instability, fever, tachycardia, aggravated toxic features, multiorgan dysfunction.
   Treatment: Injection hydrocortisone, oral antithyroid drugs, tepid sponging of whole body, beta blocker injection, oral iodides, large amount of IV fluids for rehydration, digitoxin, cardiac monitor, often ventilator support, with close observation. It has got high mortality rate with critical period of 72 hours. Fluid and electrolyte management, cardiac management are the important aspects to be monitored and treated.

Treatment of thyrotoxic crisis

❖ Injection hydrocortisone high dose—500-1000 mg IV; Tepid sponging; antipyretics, cooling blankets
❖ Rehydration with adequate IV fluids to control circulatory collapse

- Injection sodium iodide IV-1 gram; oral iodides—Lugol's iodine 10-15 drops 8th hourly
- Injection propranolol—2 mg; esmolol; reserpine 500 μg/kg loading dose, then 5 μg/kg/minute; heart rate should be < 100/min
- Oral antithyroid drugs—propylthiouracil, 200 mg; carbimazole 20 mg 4th hourly ithium carbonate 300 mg 6th hourly
- Digitoxin; heparin; plasmapheresis
- Fluid and electrolyte management, antibiotics
- ICU care, ventilator support and cardiac monitor
- Exchange transfusion; charcoal plasma perfusion
- Heamodialysis

6. *Injury to external laryngeal nerve* causes weakness of cricothyroid muscle leading to alteration in pitch of voice, voice fatigue, breathy voice, frequent throat clearing. ILS reveals short hyperaemic vocal cords, lower level affected vocal cod, oblique glottis chink due to rotation of the posterior commissure to the paralyzed one.
7. *Hypothyroidism.* Revealed clinically after 6 months.
8. Wound infection, stitch granuloma formation.
9. Keloid formation.
10. Recurrent thyrotoxicosis is 5% common. It is due to retaining of more thyroid tissue during thyroidectomy for toxic thyroid. It is difficult to manage. It is treated with *antithyroid drugs, radioiodine therapy* or re-excision surgically which is technically demanding (so often avoided).

Remember in thyroidectomy

- Middle thyroid vein is present only in 30% of cases
- Bipolar cautery or harmonic scalpel (harmonic wave) are better to control bleeding and to prevent injury to RLN
- RLN should be identified in both sides. Variations should be remembered
- Parathyroids should be retained and also their blood supply
- Parathyroid autotransplantation may be needed if their blood supply is compromised
- Rapid influx of serum calcium into bones in immediate postoperative period may cause severe hypocalcaemia—***hungry bone syndrome*** which is initially corrected by IV calcium gluconate—10 ml of 10%. It is more commonly observed in patients with beta blockers. It is due to sudden drop in PTH level after surgery
- Tension haematoma under strap muscles is very dangerous and should be relieved by removing sutures from the skin and strap muscles
- Permanent hypoparathyroidism is rare (0.5-1%) whereas temporary hypoparathyroidism and hypocalcaemia is common (25%) which is corrected well with calcium for certain period of 3-6 months.
- External laryngeal nerve injury is more common than of RLN. There is loss of tension of vocal cords causing reduced power and range in voice (pitch).
- Thyroid insufficiency develops in 25-45% cases. It is confirmed by doing serum T_3, T_4 and TSH after 6 months.

EMIL THEODOR KOCHER

> *We have arrived at the point where we generally recommend the radical operation for goitre as the surest and simplest method of treatment.... Surgeons have simply assumed that the thyroid gland has no function whatever.... Patients with total excision... all show more or less severe disturbances in their general condition.... We cannot fail to recognise (the) relation to idiocy and cretinism.... We see no objection for the time being, to the use of the name cachexia strumipriva.*
>
> —**Emil Theodor Kocher**, 1883

He is the first surgeon to get Nobel prize. He did extensive work on thyroid surgeries and designed present technique of thyroid surgeries.

He was from Switzerland. He was the founder of:

- Kocher's vein.
- Kocher's forceps (has got tooth in the tip).
- Kocherisation (Duodenal mobilisation).
- Kocher's incision (Right subcostal for cholecystectomy).
- Kocher's thyroid incision.
- Kocher's test.
- Kocher's method for reduction of shoulder dislocation.

He died in 1917. *He is the father of thyroid surgery.*

Other surgeons who got Nobel prizes are—Alexis Carrell, for his work on vascular anastomosis; Christian Bernard, for heart transplantation, Charles Huggins, urologist, for management of carcinoma prostate.

KOCHER'S TEST

In case of a huge or long-standing thyroid enlargement, both lateral lobes of thyroid are pushed posteromedially with fingers. The test is positive, if it produces **stridor.** This signifies weakened tracheal (rings) cartilage because of constant pressure by the thyroid gland. But then trachea is kept patent by the thyroid itself. Once the gland is removed there is no support to trachea and so it collapses causing stridor (*scabbard trachea*). Such patient *requires temporary tracheostomy for 2-3 weeks*. By then trachea regains its strength back.

HYPOTHYROIDISM

Causes

- Agenesis or dysgenesis
- Enzyme deficiency
- Iodine deficiency
- Hashimoto's thyroiditis
- Antithyroid drugs
- Radioiodine
- *Drugs:* Lithium, amiodarone
- After thyroidectomy, common cause

Primary—is due to thyroid diseases or removal of thyroid.
Secondary—is due to hypopituitarism.
Tertiary—is due to hypothalamic diseases.

Cretinism

- It is fetal/infantile hypothyroidism due to inadequate thyroid hormone production during fetal and neonatal period. It may be due to agenesis, inborn error of thyroid metabolism, dyshormonogenesis or dietary deficiency (endemic).
- Typical hoarse cry, macroglossia, umbilical hernia, thickened skin are the features.
- TSH will be raised; T_3 and T_4 will be low.
- Incidence is 1:4000 live births.
- It is treated with L thyroxine once a day morning orally.

Clinical Features

- *General:* Tiredeness, weight gain, cold intolerance, goitre, hyperlipidaemia.
- *Cardiovascular*: Bradycardia, angina, cardiac failure, pericardial effusion.
- *Haematological*: Anaemia.
- *Dermatological*: Dry skin, vitiligo, Alopecia, erythema.

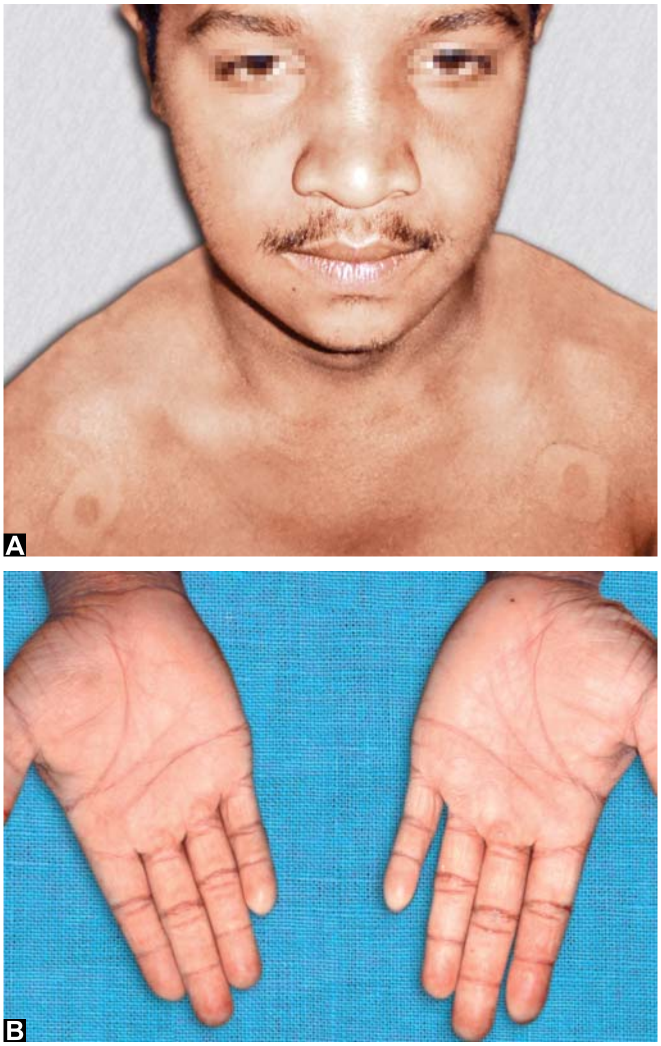

Figs 6.76A and B: Typical myxoedema patient.

Myxoedema coma		
Characteristic features	***Precipitated by***	***Treatment***
Hypothermia	Cold	Drug of choice—T_3, (Levothyronine) 10 µg given IV 6th hourly
Hypotension	Infection trauma	Levothyroxine—300 µg IV/1000 µg orally (large dose)
Hyponatraemia		Glucocorticoids
Hypoventilation		Antibiotics
Hypoglycaemia		Slow warming
Bradycardia		Electrolyte management
Skin is deadly cold like of a todd		
Coma		
Note: Myxoedema is severe thyroid failure.		

- *Reproductive:* Infertility, menorrhagia, galactorrhoea.
- *Gastrointestinal*: Constipation, ileus.
- *Developmental*: Growth and mental retardation, delayed puberty.
- *Other features*: Carpal tunnel syndrome, myalgia, hoarseness, deafness, ataxia, depression, psychosis (Myxoedema madness).

Investigations

- T_3, T_4 estimation.
- TSH level estimation which is higher.

Treatment

- Replacement with L-thyroxine 100 to 150 µg/day. In old patients with ischaemic heart disease initial therapy is with 25-50 µg/day and then gradually increased upto the required dose.
- Initial rapid response can be achieved by giving L iodothyronine 20 µg tid.

RECURRENT LARYNGEAL NERVE PALSY

Anatomy and Relations of Nerves with Thyroid Gland

- *External laryngeal nerve* is a branch of superior laryngeal nerve. Vagus gives superior laryngeal nerve branch at inferior ganglion of vagus which is at the level of greater cornu of hyoid bone, which in turn divides into internal laryngeal and external laryngeal nerves. Internal laryngeal nerve pierces the thyrohyoid membrane and gives sensory supply to larynx above the vocal cord and hypopharynx. External laryngeal nerve runs close to superior thyroid vessels to supply cricothyroid muscle which is tensor of vocal cord.

Don't judge each day by the harvest you reap but by the seeds you plant.

It may get injured while ligating superior pedicle causing defective pitch of the voice. So upper pedicle of thyroid should be ligated close to gland.

- *Recurrent laryngeal nerve* is a branch of vagus which hooks around ligamentum arteriosum with arch of aorta on the left and right subclavian artery on the right side. It runs in the tracheo-oesophageal groove near the posteromedial surface close to thyroid gland. Nerve lies in between branches of inferior thyroid artery. So inferior thyroid artery should be ligated away from the gland. Anomalies of the nerve should be remembered. In 5% of cases nerve passes through the gland. Nerve may be much more away from the gland. Nerve may be closely adherent to gland posteriorly. Nerve may lie within the ligament of Berry (25%). Non-recurrent laryngeal nerve may be present in 1 in 1,000 cases with a horizontal course. RLN supplies all muscles (abductors, adductors) of the larynx except cricothyroid and also gives sensory supply of larynx below the vocal cords.

Muscles of the Larynx

- Cricothyroid: It is the only muscle which is located on the external aspect of the larynx. It is supplied by external laryngeal nerve. It is tensor and mild adductor of the vocal cord.
- Abductors of the vocal cord—Posterior cricoarytenoids.
- Adductors of the vocal cord—Lateral cricoarytenoids, transverse arytenoid, thyroarytenoids and cricothyroids.
- Relaxant of vocal cords—Thyroarytenoids and vocalis.
- Muscles which close the laryngeal inlet—Oblique arytenoids and aryepiglottic.
- Muscles which open the laryngeal inlet—Thyroepiglotticus.

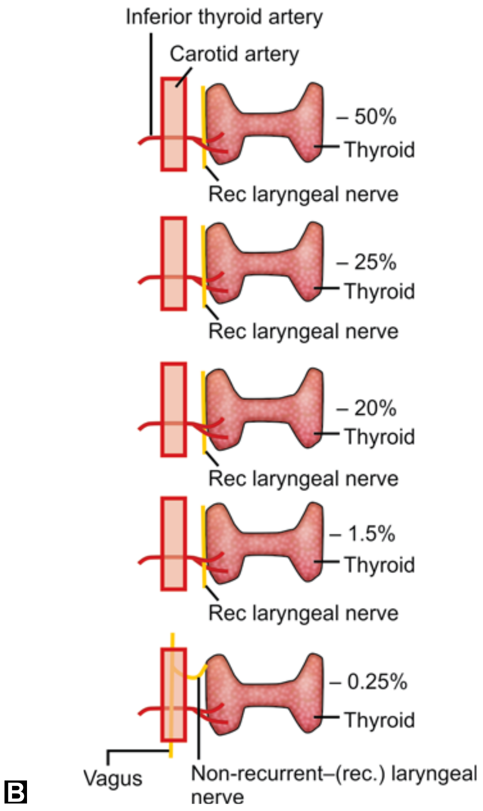
Inferior thyroid artery
Carotid artery
– 50%
Thyroid
Rec laryngeal nerve
– 25%
Thyroid
Rec laryngeal nerve
– 20%
Thyroid
Rec laryngeal nerve
– 1.5%
Thyroid
Rec laryngeal nerve
– 0.25%
Thyroid
Vagus
Non-recurrent–(rec.) laryngeal nerve
B

Fig. 6.77B

Figs 6.77A and B: Anatomy of superior and recurrent laryngeal nerve. Superior laryngeal nerve divides into internal and external laryngeal nerves. Right recurrent winds round subclavian artery and left winds around arch of aorta.

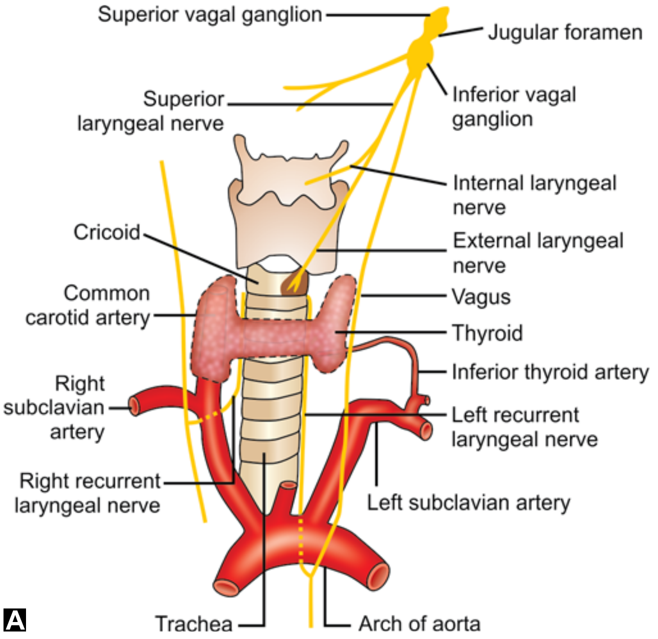
Superior vagal ganglion
Jugular foramen
Superior laryngeal nerve
Inferior vagal ganglion
Internal laryngeal nerve
Cricoid
External laryngeal nerve
Common carotid artery
Vagus
Thyroid
Inferior thyroid artery
Right subclavian artery
Left recurrent laryngeal nerve
Right recurrent laryngeal nerve
Left subclavian artery
A
Trachea
Arch of aorta

Fig. 6.77A

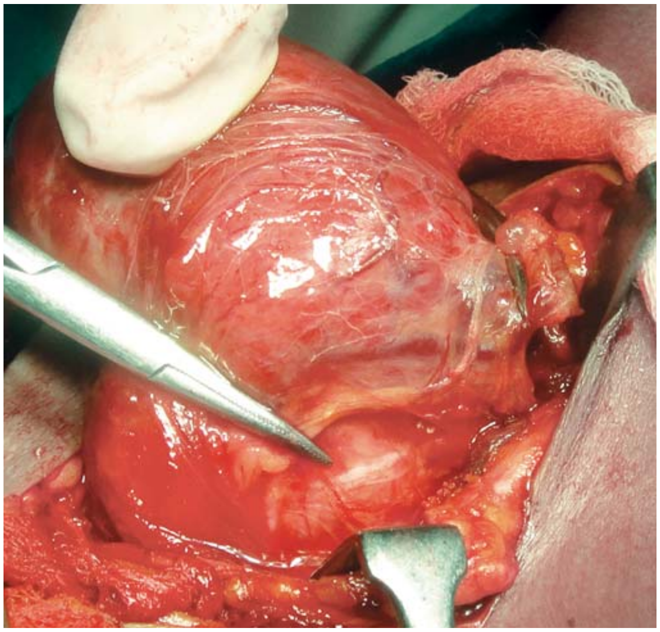

Fig. 6.78: Recurrent laryeal nerve on table identification is a must during thyroidectomy.

Positions of Vocal Cord

Median; paramedian (1.5 mm); normal; cadaveric/neutral (3.5 mm); gentle abduction (7 mm); completely abducted (9.5 mm).

Assessment of voice change

- Pitch of the voice—whether raised/lowered or pitch locked
- Breath support during speaking is adequate or not
- Ability to alter the rapidity of speech—slow/fast/ medium
- Altered laryngeal and neck muscle tension
- *Indirect laryngoscopy*—with tongue pulled out using gauze, warmed ILS is placed into the oral cavity to see vocal cords. Patient is asked to say 'e' to see the vocal cord movements

Types

1. Unilateral recurrent laryngeal nerve palsy.
2. Bilateral recurrent laryngeal nerver palsy.
3. Unilateral, combined recurrent laryngeal and superior laryngeal nerve palsy.
4. Bilateral, combined recurrent laryngeal and superior laryngeal nerve palsy.

1. Unilateral Recurrent Laryngeal Nerve Palsy

All intrinsic muscles of the larynx except cricothyroid are paralysed. Vocal cord becomes median or paramedian position. Reasons are:

a. Only retained cricothyroid is a weak adductor of vocal cord causing median or paramedian position of vocal cords (*Wagner-Grossman hypothesis*).
b. Abductor fibres which are phylogenetically newer (to posterior cricoarytenoid) of the recurrent laryngeal nerve is more susceptible and paralysed than adductor fibres (*Semon's law).*

Causes for unilateral recurrent laryngeal nerve paralysis

- Infiltration from carcinoma thyroid
- Infiltration from carcinoma oesophagus
- Infiltration from carcinoma bronchus (left-sided)
- Mediastinal tumours (left-sided)
- Aortic aneurysms
- Trauma
- Idiopathic
- Thyroidectomy
- Neck or mediastinal lymph node mass

Clinical features

- Asymptomatic in 33% cases.
- Some change in voice, which gradually becomes normal with speech therapy.
- Aspiration never occurs.
- Airway obstruction never occurs.

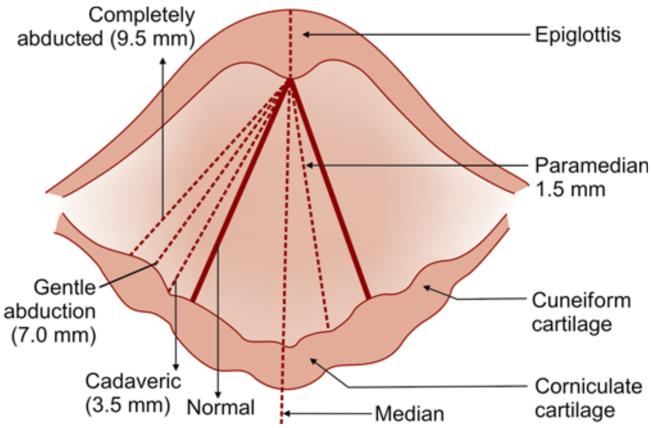

Fig. 6.79: Vocal cord in recurrent laryngeal nerve palsies

Treatment

No specific treatment is required.

But steroid should be started. Prednisolone 20 mg tid for 10 days orally after food is given with gradual tapering in another 10 days.

2. Bilateral Recurrent Laryngeal Nerve Palsy

Both side intrinsic muscles are paralysed. So both vocal cords lie in median or paramedian position due to unopposed actions of the both side cricothyroid.

Clinical features

- Change in voice.
- Severe dyspnoea and stridor (more during exertion) leading to airway block and respiratory arrest.

Treatment

- Emergency tracheostomy.
- Lateralisation of the cord by:
  - Arytenoidectomy by open surgery or through an endoscope.
  - Vocal cord lateralisation through endoscope.
- Excision of vocal cord through an endoscope or laser cordectomy.
- Implantation of sternohyoid.
- Thyroplasty.

3. Unilateral, Combined Recurrent Laryngeal Nerve and Superior Laryngeal Nerve Palsy

All the muscles on one side are paralysed. Vocal cord is in cadaveric position, 3.5 mm from the midline.

Clinical features

- Hoarseness of voice.
- Aspiration through ineffective glottis.
- Ineffective cough.

Have, goals, they give direction, purpose and meaning to lofe.

Treatment

- Speech therapy (Commonly healthy opposite cord moves to opposite side of the midline so as to compensate for the paralysis of the opposite side).
- Injection of Teflon to the paralysed cord.
- Muscle or cartilage implant to the paralysed cord.
- Arthrodesis of cricoarytenoid joint.

4. Bilateral, Combined Recurrent Laryngeal Nerve and Superior Laryngeal Nerve Palsy

Causes

- All (total) paralysis of intrinsic muscles of larynx
- Total laryngeal anaesthesia

Clinical features

- Aphonia (no voice).
- Aspiration due to severe glottis incompetence and laryngeal anaesthesia.
- Absence of cough.
- Retention of secretions in the chest.
- Respiratory arrest.

Treatment

- Emergency tracheostomy.
- Fixing epiglottis over the arytenoids to prevent aspiration.
- Plication of vocal cords to prevent aspiration.
- Total laryngectomy.

❖ Unilateral recurrent laryngeal nerve palsy	– Not dangerous
❖ Bilateral recurrent laryngeal nerve palsy	– Most dangerous
❖ Unilateral combined palsy	– Not dangerous
❖ Bilateral combined palsy	– Dangerous

Chapter 7 Parathyroids and Adrenals

> *About three years ago I found on the thyroid gland…a small organ, hardly as big as a hemp seed, which was enclosed in the same connective tissue capsule as the thyroid, but could be distinguished therefrom by a lighter color. A superficial examination revealed an organ of a totally different structure from that of the thyroid, and with a very rich vascularity…. (I) suggest the use of the name Glandulae parathyreoideae; a name in which the characteristic of being bye-glands to the thyroid is expressed.*
>
> **—Ivar Victor Sandström,** 1879

CHAPTER OUTLINE

- Anatomy
- Calcium
- Hyperparathyroidism
- Parathyroidectomy
- MEN Syndrome
- Apudomas
- Hypoparathyroidism
- Tetany
- Adrenals
- Adrenal Cortical Tumours
- Adrenocortical Carcinoma
- Cushing's Syndrome
- Conn's Syndrome
- Virilising Syndrome or Adrenogenital Syndrome
- Neuroblastoma
- Pheochromocytoma

ANATOMY

- Parathyroids are endocrine glands situated behind the thyroid gland. They are four in number two on each side.
- Two upper glands are constant in position. Superior is *behind* recurrent laryngeal nerve. They develop from 4th pharyngeal pouch hence called as parathyroid IV.
- Two lower glands are variable in position. They develop from endoderm of 3rd pharyngeal pouch hence called as parathyroid III. It is usually *in front* of the recurrent laryngeal nerve, lower part.
- Each gland weighs *40-50 mg*. It is brownish (khaki coloured) firm gland, which sinks in the fluid unlike fat which floats. It is usually adjacent to the anastomosis between superior and inferior thyroid arteries posteriorly.
- Glands (chief cells) secrete parathormone (PTH) which controls the calcium metabolism.
- Variations are common in inferior parathyroids. It can be located in thyrothymic ligament, superior pole of thyroid, tracheooesophageal groove, behind oesophagus, carotid sheath. There may be more than two glands on one side called as supernumerary gland (2.5-22%).
- Both superior and inferior glands receive their blood supply from inferior thyroid artery and its anastomotic branch. Blood supply is through *an end arterial* branch.
- Gland contains *chief and oxyphil* (water clear) cells.

Parathormone—PTH

- Increases absorption of the calcium from the gut
- Mobilizes calcium from the bone
- Increases the calcium reabsorption from the renal tubules
- Half life of PTH is 4 minutes.

CALCIUM

Total calcium in plasma in ionised and nonionised (unbound and bound) form is 8.5-10.2 mg/dl (2.2-2.5 mmol/L). 55% is bound; 45% is unbound ionised free active part (4.5-5.0 mg/dl). Commonest protein part of bound calcium is albumin (80%); remaining is beta globulin, non-protein molecules. Ionised calcium is inversely related to pH of blood; raise in pH by 1 will reduce the ionized calcium by 0.36 mmol/L. Levels of calcium is controlled by PTH, calcitonin and vitamin D acting on bone,

Any adrenal tumour more than 50 gm is usually malignant.

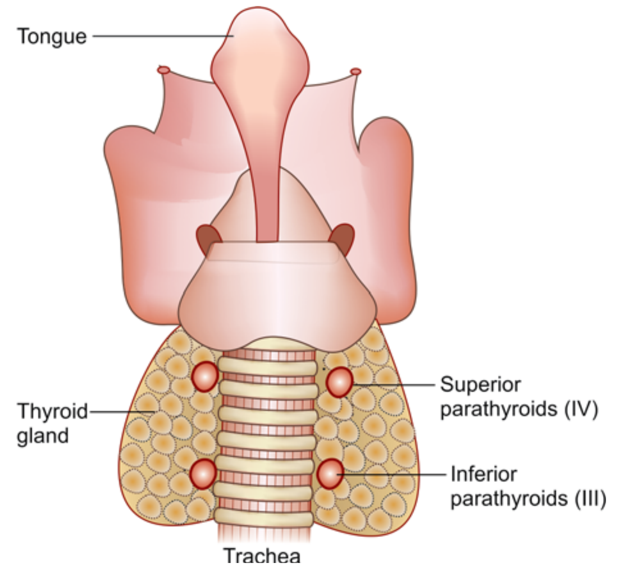

Fig. 7.1: Anatomical location of parathyroids.

kidney and GIT. PTH (84 amino acids) is secreted by chief cells whenever serum calcium level falls. PTH stimulates osteoclasts for bone resorption; kidney to promote calcium reabsorption and production of 1,25 dihydroxy vitamin D; GIT to promote absorption of calcium and phosphorous. Vitamin D hydroxylation at carbon 25 occurs in liver, and later at carbon 1 occurs in kidney. It is important for blood coagulation, neuromuscular activity, cellular activity, bone integrity.

HYPERPARATHYROIDISM (HPT)

Types

1. Primary.
2. Secondary.
3. Tertiary.

Primary is unstimulated inappropriate high PTH secretion due to hyperplasia or adenoma causing hypercalcaemia**.**

Secondary is due to chronic renal failure or due to malabsorption, vitamin D deficiency rickets, i.e. due to chronic hypocalcaemia**.** There is always parathyroid hyperplasia. CRF causes hyperphosphataemia and reduced conversion of 25 DHCC to 1, 25 DHCC. So absorption of calcium from GIT is reduced causing chronic hypocalcaemia which stimulates PTH secretion and parathyroid hyperplasia.

Tertiary is due to autonomous reactive hyperplasia of parathyroids often seen like after renal transplantation.

Primary hyperparathyroidism

- Adenoma (75-90%)
- Hyperplasia (20-24%)
- Carcinoma, rare (1%)

Carcinoma of parathyroid is rare—1%. It is aggressive. It presents as nodule, palpable neck lymph nodes along with HPT. It is invasive to soft tissues, thyroid, perineural spaces. Often blood spread can occur. It needs wide tumour excision, hemithyroidectomy, neck node dissection and later radiotherapy to neck and mediastinum. With therapy 5-year survival rate is 85%.

Primary HPT is 3rd most common endocrine disease after diabetes and thyroid disease. Increased PTH causes hypercalcaemia.

Clinical Features

Clinical vignette of hyperparathyroidism—**“*Bones, stones, abdominal groans, and psychic moans.*”**

1. Hyperparathyroidism is common in middle aged women.
2. Presentation may be *asymptomatic*—> 50% cases.
3. *Nonspecific symptoms and psychiatric symptoms*: They are most often named as neurotics.
4. Behavioural problems.
5. In the bone there will be *osteitis fibrosa cystica* **(*von Recklinghausen disease*) (*5%*)***,* which shows single or multiple cysts or pseudotumour in the jaw, skull or middle phalanges.
   Lamina dura of tooth is the first bone to show changes.
6. In the kidney**,** there may be bilateral, multiple *renal stones* or *nephrocalcinosis* (may go for renal failure).
7. It may be associated with the *peptic ulcer, pancreatitis, MEN I syndrome*.
8. They are more prone for skin necrosis, band keratopathy, pseudogout, myalgia, arthralgia, polyuria, glycosuria, hypertension (33%).

- Primary HPT is commonly sporadic than familial
- Histological difference between adenoma and hyperplasia is difficult to assess
- On table gross look is preferred method
- Adenoma is usually single. Multiple adenomas occur in elderly (5%)
- Hyperplasia involves all four parathyroids
- Parathyroid cyst may be developmental/secondary to degeneration of nodule or adenoma. Cyst aspiration shows clear fluid with high PTH level in the content fluid
- Preoperative IV methylene blue 5 mg/kg in 500 ml dextrose saline which stains parathyroids is used often to localise glands
- 50 percentage of one gland is sufficient to maintain function
- Persistent HPT is one in which HPT persists immediately even after initial surgery
- Recurrent HPT (initially after first surgery HPT is corrected but recurs 12 months after surgery) is due parathyromatosis, development of new adenoma, and hyperplasia of transplanted parathyroid. *Parathyromatosis* is due to rupture and spillage of parathyroid tissue in neck, mediastinum forming functioning nodules
- Re-exploration and removal in persistent/recurrent HPT needs on table PTH assay. On table PTH level drop more than 50% of preoperative value indicates the success of surgical removal
- CT, MRI, PET scan and selective angiography are usually done only for recurrent diseases

Acute Hyperparathyroidism (Crisis)

Causes

- Sudden increase in PTH level due to rupture of parathyroid cyst or bleeding in the parathyroid tumour.

- Severe dehydration precipitates crisis.
- Secondaries in bone (primary may be from breast).

Acute hyperparathyroidism is rare but dangerous presentation (crisis) wherein patient presents with abdominal pain, vomiting, dehydration, oliguria, muscular weakness and death.

Serum Ca^{++} is very high > 12 mg% (3.5 mmol/L).

Treatment of Crisis

- Forced diuresis using 3-5 litres of saline with frusemide.
- Rehydration using normal saline 300 ml/hour.
- Steroids, inhibits effects of vitamin D. dose is 400 mg/day IV for 5 days.
- Clodranate Na^{+}, Pamidronate (90 mg IV slowly in 4 hours). Zoledronic acid 4 mg IV initially and later 8 mg is also used.
- Drugs to reduce Ca^{++} level, i.e. *mithramycin, calcitonin, prednisolone, biphosphonates*. *Cinacalcet* is calcium receptor agonist which reduces the serum calcium level. Gallium nitrate inhibits osteoclast resorption of calcium at a dose of 200 mg/m^2/day for 5 days.
- Condition has high mortality rate.
  Patient develops hyperchloraemic metabolic acidosis.

Investigations for HPT

- High serum Ca^{++} —> 10 mg/100 ml.

Note:
Serum albumin should also be assessed to get accurate calcium level.

- Decreased serum phosphorus.
- Increased urinary Ca^{++} —> 250 mg/24 hr.
- Increased serum alkaline phosphatase.
- Increased PTH level in the serum is diagnostic—> 0.5 mg/L.
- X-ray skull shows *salt-pepper appearance*.
- X-ray phalanges and jaw are specific.

X-ray features in hyperparathyroidism

- Pepper lesions in the skull
- Sub-periosteal erosion of radial side of middle phalanx
- Calcification in bones

- U/S abdomen to find out problems in kidney, pancreas.
- U/S neck and CT/MRI scan neck and mediastinum.
- Selective venous sampling for PTH is also very useful.
- Thallium-Technetium scan shows *hot spots* which is diagnostic of parathyroid adenoma.
- *Technetium-99m labelled Sestamibi isotope* scan is better and sensitive (80%) than Thallium-Tc scan. As it is very expensive it is used in parathyroid re-exploration. It is often combined with *single photon emission computerised tomography* (SPECT).
- Urinary cAMP level increases in 90% cases.
- Angiography, venous sampling, SG guided biopsy are other methods.

Differential Diagnosis

- Secondaries in the bone—due to secretion of PTH related polypeptide by tumour. Actual PTH is suppressed.
- Multiple myeloma.
- Vit D intoxication.
- Sarcoidosis.
- Functioning carcinoma.
- *Familial hypocalciuric hypercalcaemia* is an autosomal dominant disease with mild raise in serum calcium and PTH levels secondary to mutation in the cell membrane calcium receptor. Urinary calcium excretion is low. It does not require parathyroidectomy. Calcium creatinine clearance

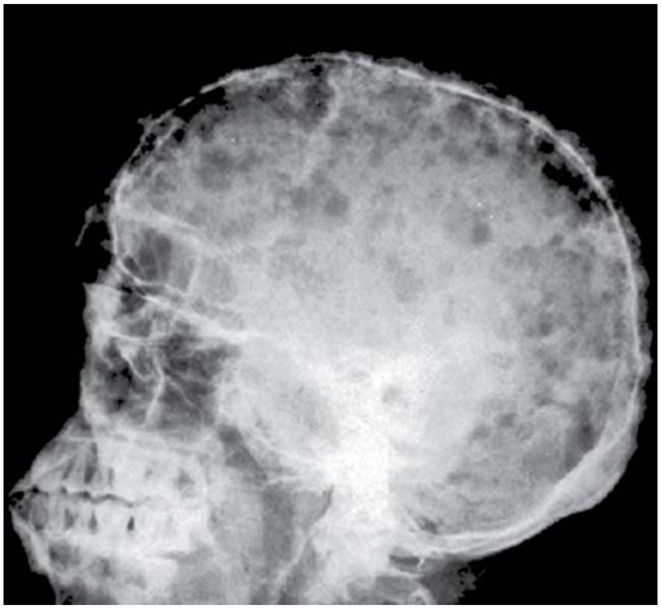

Fig. 7.2: Plain X-ray skull. Note the characteristic salt-pepper appearance of the skull bone.

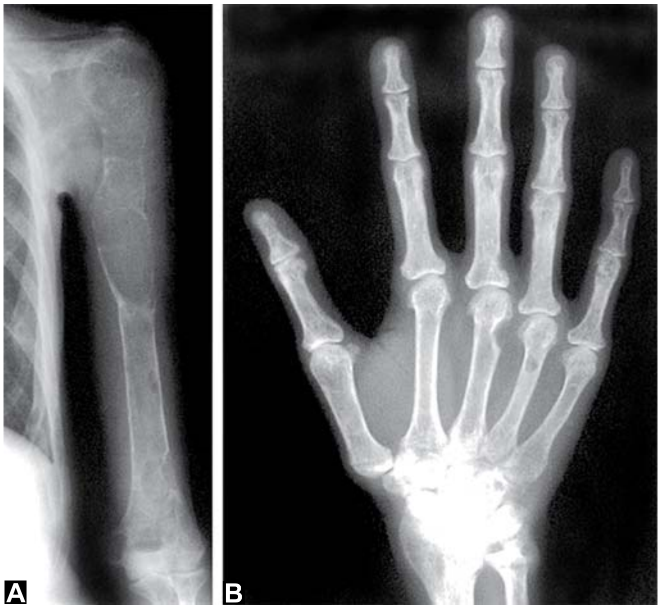

Figs 7.3A and B: X-ray of humerus bone and hand bones showing bone features—**brown tumour**—osteitis fibrosa cystica.

The fact which Addison has published seem to lead to the conclusion that these little organ are essential to life.
—Charles Edward Brown-Sequard

ratio is less than 0.01 in this condition whereas it is > 0.02 in primary HPT.

Indications for parathyroidectomy

- Severe symptoms
- Young age group
- Markedly reduced bone density
- Serum calcium more than 11 mg%
- Urinary calculi
- Neuromuscular presentations
- Urinary calcium more than 400 mg /24 hours

Problems in parathyroidectomy

- Permanent hypoparathyroidism
- Persistent hyperparathyroidism—5%
- Recurrent hyperparathyroidism—hypercalcaemia recurs 12 months after first parathyroid surgery
- Recurrent laryngeal nerve injury—1%
- Often needs additional thyroidectomy
- Variations in positions of the gland especially lower-may be in mediastinum
- Sudden drop in calcium level after surgery due to increased absorption of calcium by bones—***hungry bone syndrome***

Treatment

- Surgical removal of the glands and implantation of fragments of the gland in forearm muscle mass (brachioradialis) or neck (sternomastoid). Marker stitch is placed at the transplantation site. 1/3rd of one gland or 100 mg of parathyroid gland is autotransplanted—subtotal parathyroidectomy.
- If it is carcinoma, additional hemithyroidectomy with postoperative radiotherapy is required.
- Adenoma when occurs in one gland with normal other glands, removal of that gland with adenoma may be sufficient.
- When all four glands are diseased, transcervical thymectomy is also added along with total parathyroidectomy to reduce persistent and recurrent disease.
- In familial and MEN syndromes, total parathyroidectomy is better.
- If it is mediastinal parathyroid adenoma, after proper localisation thoracoscopic removal may be sufficient.
- *Medical treatment of primary hyperparathyroidism* is usually ineffective and not popular. However occasionally as initial therapy and acute crisis it is being advocated. Estrogens, progestogens, raloxifene (estrogen receptor modulator), mithramycin, calicitonin are the few drugs used. *Mithramycin is* used *once a week* but it is hepatotoxic and causes thrombocytopenia.

PARATHYROIDECTOMY

Primary hyperparathyroidism—*Criteria (2002) for surgical intervention are*—raise in serum calcium level more than 1 mg/dl of upper limit of the normal calcium range; 24 hours urinary calcium if more than 400 mg; creatinine clearance when reduced more than 30%; bone density greater than two standard deviations below peak bone mass in lumbar spine/hip/lower end of radius; age below 50 years; when medical therapy is not possible.

Secondary HPT is an indication for removal of all four glands with autotransplantation of parathyroid only in severe cases or with renal osteodystrophy.

Preoperative Preparation

Vocal cords should be assessed by preoperative indirect laryngoscopy.

High calcium levels preoperatively may require treatment with hydration; diuresis; steroids (prednisolone 20 mg TID for 5 days before surgery); 100 mmol phosphate infusion in 6 hours; 200 units calcitonin subcutaneous injection for 5 days twice daily before surgery; diphosphanate—etiodronate disodium 7.5 mg/kg daily as slow IV infusion for 3 days; mithramycin 25 µg/kg as single dose.

Anaesthesia and Position

General anaesthesia is used with neck hyperextension by placing rolled sheet under the shoulder blades. Head is placed on the head ring; head end of the table is raised to semi-erect position (*Semi-fowler position).*

Incision and Dissection

It is same as for thyroidectomy. Flaps are raised in similar way. Strap muscles are separated after opening the deep fascia in the midline. Thyroid gland is mobilised to identify the parathyroid adenoma. Parathyroid having adenoma is mobilised which is close to recurrent laryngeal nerve. End artery of the parathyroid is identified and ligated. Adenoma is separated from adjacent thyroid tissue using gauze dissection. Either on table venous sampling for PTH assay is done or venous sample from cubital vein is done for PTH assay.

Parathyroid may be confirmed by frozen section biopsy or on table aspiration of parathyroid tissue which is analysed for PTH assay which will be more than 1500 pg/ml (confirms that removed tissue is parathyroid).

Total parathyroidectomy is done for parathyroid hyperplasia by removing all four glands and ⅓ of one gland is autotransplanted into the forearm muscle (brachioradialis) or sternocleidomastoid muscle with marker stitch. Gland to be transplanted is sliced into 1 mm pieces and around 18 pieces are embedded in decided muscle with a marker stitch or clip. If in postoperative period patient still presents with features of primary HPT; transplanted area is re-explored and further reduction in parathyroid tissue is done.

Wound is closed with proper haemostasis.

Complications

- Haemorrhage, recurrent laryngeal nerve palsy, hypocalcaemia and *hungry bone syndrome* are known to occur.
- *Persistent HPT* (serum calcium does not normalise immediately after surgery leading to total failure) or *Recurrent HPT* (serum calcium after surgery becomes normal but in 6-12 months, it again increases) may be a problem.

- Hypoparathyroidism with severe hypocalcaemia is a problem when all glands are removed (3½) with bilateral neck exploration. 10 ampoules of calcium gluconate is diluted with one liter of normal saline and given as continuous infusion at a rate of 30 ml/hour—initial method of management. Hypomagnesemia should also be corrected.
- Migration or inability to identify the transplanted parathyroid is often a problem in autotransplantation of parathyroid.
- **Hungry bone syndrome**
    - It occurs usually in patients with preoperative hyperthyroidism. They have increased bone breakdown in their hyperthyroid state. When a patient's thyroid hormone level drops acutely after surgery, stimulus to break down bone is removed. The bones are now "hungry" for calcium, remove calcium from the plasma rapidly.
    - It usually occurs after parathyroidectomy; thyroidectomy for toxic thyroid; prostate cancer patients on estrogen therapy. Sudden cessation of existing increased bone breakdown makes bones to absorb calcium, magnesium and phosphorus rapidly. Calcium levels in blood prior to operation cannot predict hungry bone syndrome.
    - Hypocalcaemia, hypophosphataemia, hypomagnesemia and hyperkalaemia are four typical features in these patients. Estimation and correction of all these four factors is essential. ECG changes can occur. Hyperkalaemia should be treated judicially. Magnesium infusion is needed. There is an unusually high need for calcium, with a low calcium excretion in the urine.
    - Bone-specific alkaline phosphatase (ALP) continues to rise in the first few weeks indicating increased bone reconstruction.
    - Supplementation of vitamin D and elemental calcium is needed during discharge for 6 months.
    - Postoperatively they need calcitriol (gradually increased to 16 mg in 1 month, then gradually reduced) with 2 gm calcium supplement.
    - Monitoring is done by evaluating serum calcium, albumin, magnesium, phosphorus and bone specific alkaline phosphatase.
- Injury to recurrent laryngeal nerve, oesophagus can occur.

Effects of Surgery

- Among neuromuscular symptoms of primary HPT, proximal muscle weakness responds better than respiratory muscle weakness by parathyroidectomy.
- Among psychiatric illnesses, depression and spatial learning and processing improve well by surgery.
- Bone mineral density in hip and lumbar spine becomes better.
- Nephrocalcinosis is improved by surgery; but hypertension and renal excretion will not improve much.
- Half life of PTH is 4 minutes. On table serial PTH assays are done at several intervals—before dissection of the gland, during dissection and after dissection. 50% reduction in PTH level from baseline in post-removal sample is a 96% predictor of complete removal.
- Operative failure rate is 1.5 to 6%. Intraoperative PTH assay improves the success rate very much (76-94%). Cure rate is defined as normocalcaemia in 6 months postoperative period.
- Radio guided parathyroidectomy using intravenous injection of 20 mCi of 99mTc sestamibi 2 hours before surgery is not routinely practiced but may be useful for removal of adenoma appropriately. Hand held quantitative gamma counter is used intraoperatively on the neck over all parathyroid tissues.

Surgical Approaches

Classic Approach (Traditional Approach)

It is under general anaesthesia exploring bilateral neck to remove parathyroid tissue which is confirmed by frozen section biopsy. It shows 95% cure rate with 2% complication rate.

Minimally Invasive Parathyroidectomy (MIP)

It is done in case of *single adenoma* of parathyroid under regional cervical block anaesthesia usually in ambulatory set up. It is actually unilateral parathyroidectomy with removal of adenoma and involved entire that particular one parathyroid gland using limited neck exploration (small incision of 2-4 cm). Procedure is not useful in multiple adenomas or hyperplasia. Preoperative localisation with sestamibi combined with SPECT is a must. Intraoperative PTH assay is a must to confirm drop in PTH level to required level. Cervical block is given by infiltrating the xylocaine (1%) with adrenaline along posterior and deep to sternocleidomastoid muscle and also along the neck incision. Direct field block along the anterior border of the sternocleidomastoid and over the incision is also used. Postoperative serum calcium and PTH level should be assessed until 7th day.

Median Sternotomy (3%) Extension

Median sternotomy is often needed when parathyroid is in anterior mediastinum along with thymus. Often parathyroids may be 5, 6 or 7 in numbers instead of four.

Video Assisted Parathyroidectomy (Paolo Miccoli)

It is done in *localised single adenoma* using multiple ports on one side with intermittent CO_2 insufflation and suction irrigation. 1.5 cm incision 1 cm above the sternal notch helps in tactile assessment and dissection. Strap muscles are retracted laterally and thyroid is retracted medially. Conversion rate is 11%; RLN palsy is 2%.

Endoscopic Parathyroidectomy

Entire parathyroidectomy is done using laparoscopy. It is first used (1994) for a mediastinal parathyroid adenoma. In 1996, it was used for neck parathyroid hyperplasia by Gagner for removing 3½ glands. Now technique is *limited to single adenoma* to remove tumour and gland. Low pressure insufflations with 5 mm four trocars are used. Placement of trocars is dependent on the need of the operating surgeon. Many place trocars on one side only; but few prefer to place working trocars on opposite side.

Remedial Parathyroidectomy

It is done for persistent HPT (serum calcium is not normalised immediately after surgery leading to total failure) or recurrent HPT (serum calcium after surgery becomes normal but in 6-12 months, it again increases). It is done through lateral approach (*Feind*) between anterior margin of sternocleidomastoid and strap muscles so that scar tissues are avoided from initial dissection. On table localisation methods, on table PTH assay are needed. Medial sternotomy often may be needed to explore the mediastinum for complete removal of the gland. Parathyroids in such situations may be embedded in the tissue of thyroid or thymus. So removal of these glands also may be a need in such situations.

Subtotal Parathyroidectomy

It is indicated in hyperplasia or secondary HPT wherein 3½ glands are removed, retaining ½ of one gland.

Total Parathyroidectomy with Parathyroid Autotransplantation

It is done alternatively in hyperplasia wherein all four glands are removed; ⅓ or ½ of one gland is transplanted into sternocleidomastoid or brachioradialis muscle with a marker. Disease control will not occur if more part of the gland is transplanted. It is removing all four glands and½ or ⅓ of one gland is autotransplanted into the forearm muscle (brachioradialis) or sternocleidomastoid muscle with marker stitch. Transplanting gland is sliced into 1 mm pieces and around 18 pieces are embedded in decided muscle with a marker stitch or clip. If postoperatively patient still presents with features of primary HPT; transplanted area is re-explored and further reduction in parathyroid tissue is done. Migration or inability to identify the transplanted parathyroid is often a problem in autotransplantation of parathyroid.

MEN SYNDROME (MEA SYNDROME)

- Commonly inherited as an *autosomal dominant*.
- Cells involved has got common features of *APUD cells (APUDomas*).

Note:
- APUD—**A**mine **P**recursor **U**ptake **D**ecarboxylation.
- **MEN**—**M**ultiple **E**ndocrine **N**eoplasia; **MEA**—**M**ultiple **E**ndocrine **A**denomatosis.

Types

- ***Type—I*:** Parathyroid hyperplasia or adenomas; pituitary tumour; pancreatic tumour [Endocrine—Insulinoma, gastrinoma, glucagonoma, vipoma]. It is also called as *Werner's syndrome*. Here the defect is in chromosome 11.
- ***Type—II***: Also called as *Sipple's disease.*
  1. II a includes medullary carcinoma of thyroid+ phaeochromocytoma + parathyroid hyperplasia (50%). Here the defect is in chromosome 10.
  2. II b includes medullary carcinoma of thyroid + phaeochromocytoma mucosal neuromas in lips and eyelids with bumpy-lumpy lesions, with marfanoid face, megacolon.

Evaluation and management of MEN syndrome: Family history; biochemical screening with parathormone, serum calcium, prolactin, growth hormone, blood sugar, insulin, proinsulin, pancreatic polypeptide, glucagons, gastrin levels, calcitonin, urinary catecholamines estimation; genetic screening from isolated DNA from peripheral blood white cells for ret proto oncogene. Specific surgical treatment of conditions is needed.

APUDOMAS

APUD (Amine Precursor Uptake Decarboxylation) *cells* are cells having specific cytochemical characteristics. *They are*:
- High amine content.
- Capacity of amine precursor uptake.
- Property of decarboxylation of these precursors to form amines.

- Initially it is thought that APUD cells are derived from *neural crest cells* (Pearse) but now found to be from endoderm. Cells share *similarities* in structure, properties, histological, histochemical, immunocytochemical and electron microscopic appearance. *Neuron specific enolase* enzyme is specific for these cells.
- These cells have got capacity to *synthesize peptides* which has got different modes of action. They are:
  1. *Endocrine action* where peptides get secreted into circulation to have distant target actions.
  2. *Paracrine action* where peptides get secreted locally to have action at local sites.
  3. *Neurocrine action* where peptides act as neurotransmitter at neuronal synapses.
  4. *Neuroendocrine action* where peptides stimulate release of peptide product of the neuron into the circulation.
- Tumours arising from these cells are grouped as ***Apudomas.*** Many of parathyroid tumours, pancreatic tumours are under this group. Their presentations are commonly due to increased secretions of these neuroendocrine hormones. Commonly *presentation* is like syndromes. Insulinoma, glucagonoma, gastrinoma, VIPoma are different examples. Tumours are ***entopic type*** if they secret hormones normal to the tissue like insulinoma/glucagonoma. They are ***ectopic type*** if they produce hormones which are not normal to the tissue of origin like gastrinoma/VIPoma.
- APUDOMAS are commonly associated with ***MEN syndrome*** (commonly type I). Radioimmunoassay, MRI abdomen, CT neck are *useful investigations.*
- ***Treatment*** is of individual diagnosed components of the condition.

HYPOPARATHYROIDISM

Types

Temporary	More common (2-50%) average 10%	Usually lasts for 2 months maximum up to 6 months. Decrease in calcium; increase in phosphorous
Permanent	Less common (0.4-13%) average 1%	Permanent—continues beyond 6 months. Decrease in calcium; increase in phosphorous
Hungry bone syndrome	Common (5-13%)	Severe, rapid begins in immediate postoperative period. Decrease in calcium; decrease in phosphorous

- Inadequate production of PTH leads to hypocalcemia. Hypoparathyroidism and the resulting hypocalcemia may be permanent or transient. The rate of permanent hypoparathyroidism is 0.4-13%.
- *Causes* are - direct trauma to the parathyroid glands, devascularization of the glands, or removal of the glands during surgery.
- The rate of *temporary hypocalcemia* is reportedly 2-50%. The cause of transient hypocalcemia after surgery is due to temporary hypoparathyroidism caused by reversible ischaemia to the parathyroid glands, hypothermia to the glands, or release of endothelin-1. Endothelin-1 is an acute-phase reactant which suppresses PTH production. Other hypothesis is—calcitonin release and hungry-bone syndrome. Calcitonin is produced by the thyroid C cells which inhibits bone breakdown while stimulating renal excretion of calcium (opposite of PTH).
- An effective method of evaluation of parathyroid function is to assess ionized calcium (or total calcium and albumin) levels in the perioperative period. A normal postoperative PTH level can accurately predict normocalcemia after thyroid surgery. Immediate postoperative PTH level if <1.5 pmol/L and morning serum calcium <2.0 mmol/L, then patient is at risk of hypocalcemia.

Features of Hypoparathyroidism

- Circumoral tingling, numbness, paraesthesia.
- Carpopedal spasm, laryngeal stridor.
- Respiratory muscle spasm, suffocation.
- Convulsions, blurred vision due to intraocular muscle spasm.
- Cataract formation as late feature.

Treatment

- Patients who have *symptomatic hypocalcemia* in the early postoperative period or whose calcium levels continue to fall rapidly need treatment. In symptomatic patients, intravenous calcium gluconate 10 ml 10% solution (1 gm) is administered over 10 minutes. A calcium infusion is started at a rate of 1-2 mg/kg/hr if symptoms do not resolve. Rate of infusion is decided by repeated analysis of serum calcium and clinical features.
- One to two grams of elemental oral calcium should be given each day. Calcium carbonate 1250 mg provides 500 mg of elemental calcium; the patient should take 2500-5000 mg/day.
- The patient needs vitamin D supplementation with calcitriol 0.25-1 mcg/day. The calcium supplementation should be divided 4-5 times per day rather than in a single dose to maximize absorption by the GI tract.
- Intravenous/intramuscular magnesium 0.5 gm/4 mEq/kg body weight for 5 days; later magnesium gluconate tablets 500 mg orally.
- In 2 months, trial weaning of oral calcium can be made to identify whether the hypoparathyroidism is temporary. Need for calcium supplementation longer than 6 months, indicates *permanent* hypoparathyroidism.

Prevention of Hypoparathyroidism

- Proper preoperative parathyroid localisation.
- Maintaining blood supply of parathyroids.
- Capsular ligation of the thyroid vessels at lower pole to retain parathyroid end arteries.
- Parathyroid autotransplantation if all 4 parathyroids are removed.

Spurious Hypocalcemia

- It is decrease in total calcium, albumin and haematocrit in the first two postoperative days following any surgery including non-thyroid. Surgical stress releases ADH causing water retention and haemodilution. Albumin bound total calcium is decreased. Free ionized calcium is normal unlike in true hypocalcaemia.

TETANY

Decreased level of calcium in blood causing its effects.

Causes

- After thyroidectomy (It is decreased level of parathormone in the blood causing hypocalcaemia). It is usually temporary, lasts for 4-6 weeks. It is the commonest cause of hypoparathyroidism.
- Other causes of hypoparathyroidism are neck dissection, haemochromatosis, Wilson's disease, DiGeorge's syndrome.

- Severe vomiting, hyperventilation associated with respiratory alkalosis.
- Metabolic alkalosis.
- Rickets, osteomalacia.
- Chronic renal failure.
- Acute pancreatitis.

Decreased PTH causes decrease in calcium level in the blood leading to:

- *Circumoral paraesthesia*, paraesthesia of neck, fingers and toes.
- Twitching and weakness of tongue muscles, muscles of forearm, hand, foot and digits—*carpopedal spasm.* Fingers are extended except at the MCP joints and thumb is strongly adducted (obstetrician' / accoucheur's hand); there is extension of feet to develop carpopedal spasm.
- *Chvostek's sign*—Tapping above the angle of the jaw to stimulate branches of facial nerve causes the twitching of the angle of mouth and eyelids.
- Applying the sphygmomanometer to the arm and inflating the pressure more than systolic pressure (200 mm of Hg) of the patient for three minutes can demonstrate carpal spasm *(Trousseau's sign).*
- *Stridor* and difficulty in breathing due to paralysis of respiratory muscles.
- Generalised weakness and twitching all over the body in severe cases mimicking *convulsions.*
- Prolonged QT interval and QRS complex in ECG.

Management

- Serum calcium estimation is done. It will be less than 7 mg%.
- IV calcium gluconate 10% 10 ml sixth to eighth hourly is given.
- Later oral calcium (1 gram three times daily) with vitamin D supplementation (1-3 µg daily).
- Follow up at regular intervals by doing serum calcium level.
- Magnesium sulphate supportive therapy is also often needed—10 ml 10% magnesium sulphate intravenously.

DiGeorge's syndrome

- Absence parathyroids
- Thymic aplasia with immunodeficiency
- Cardiac defects

ADRENALS

> *The facts which Addison has published seem to lead to the conclusion that these little organs are essential to life.*
>
> —**Charles Edouard Brown-Séquard,** 1856

Adrenal cortex contains three layers:

1. Zona glomerulosa secretes mineralocorticoids.
2. Zona fasciculata secretes glucocorticoids.
3. Zona reticularis secretes sex hormones.
    - Cortex—85%; mesoderm; produces and secretes aldosterone, cortisol, DHEA.
    - Medulla—15%; neuroectoderm; produces and secretes norepinephrine, epinephrine.

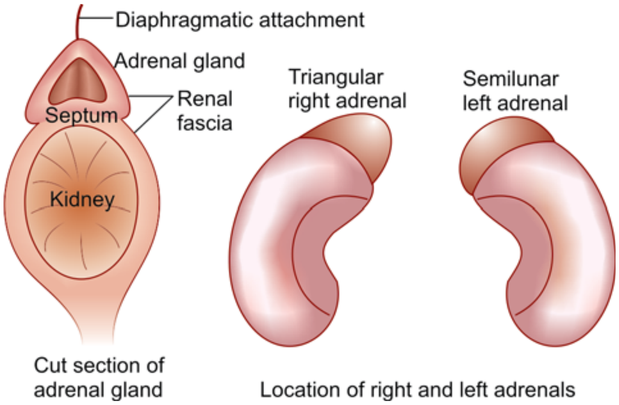

Fig. 7.4: Anatomy of adrenal gland.

ADRENAL CORTICAL TUMOURS

- Are usually adenomas.
- Any tumour measuring 6 cm or more are likely to be malignant, or have high risk of turning into malignancy. So requires surgical resection.
- *Adenocarcinoma* is the commonest adrenal cortical malignancy (1%). It is very aggressive tumour.
- Cortical tumours may be functioning or nonfunctioning.
- Functioning tumours secrete mineralocorticoids, glucocorticoids or sex hormones or combinations of these.

Investigations

- U/S abdomen.
- CT scan, Hormone evaluation.

Treatment

Adrenalectomy.

Incidentalomas (3-5%)

- Incidentalomas are adrenal tumours incidentally identified either through U/S; CT scan; MRI; or any other methods done for other reasons.
- When incidentalomas are identified a proper biochemical work up for hormones is essential. Overnight dexamethasone suppression test; 24-hour urinary cortisol excretion assay; 24-hour urinary excretion of catecholamines; serum potassium, renin and aldosterone analysis are done in these patients. MRI in suspected malignancy is better.
- If secondaries are suspected adrenal mass biopsy is done. Otherwise it should not be done. Non-functioning adenomas (commonest—78%), Cushing's adenoma, adrenocortical carcinoma, phaeochromocytoma, secondaries, Conn's adenoma are different causes of adrenal incidentalomas. Adenoma with Cushing's syndrome is better than adenoma with Conn's syndrome. Tumour with adrenogenital syndrome is commonly malignant.
- Non-functioning adrenal tumour more than 4 cm should be operated. Smaller tumours that increase in size over specified

period should be removed. All functioning tumours should be operated.
- Non-functioning tumour less than 4 cm should be followed up at regular intervals by hormone tests and CT/MRI.

ADRENOCORTICAL CARCINOMA

- It is a rare malignancy with 1 case in 1000000 population. It is common in females with bimodal presentation.
- ***Features are***—tumour size; necrosis; haemorrhage; capsular invasion; vascular invasion. Immunohistochemistry is specific.
- ***Presentations are***—abdominal pain; back pain; Cushing's syndrome (60%); hyperaldosteronism; virilisation.
- It is *least common* malignant endocrine tumour.
- One of the *most malignant* endocrine tumour after anaplastic carcinoma of the thyroid.
- Can be functioning or nonfunctioning tumour.
- Larger tumours (>6 cm) are more likely to be malignant.
- Cytologic criteria alone are not diagnostic but should find capsular infiltration and vascular invasion.
- Often presents with no symptoms or only vague symptoms.
- May present with mass effect and compression of adjacent structures.
- Increased secretion of one or more steroid hormones can occur.
- ***Diagnosis is by***—hormone evaluation, CT/MRI; MR angiography to identify IVC tumour thrombus.
- Secondaries occur commonly in lungs. So HRCT of lungs is usually done.
- ***Mc Falane staging*: **Stage I—tumour < 5 cm; Stage II—tumour > 5 cm; Stage III—tumour with local invasion; Stage IV—distant spread. In stage I and II, 5 year survival is 25%; in stage III and IV, 5-year survival is 5%.
- ***Treatment*** is en block adrenalectomy; removal of tumour thrombus from vena cava if present; chemotherapy using mitotane—Op-DDD (choice) with cis platin, etoposide and doxorubicin. Adjuvant radiotherapy may be used to prevent recurrence *(little role)*.
    - *Recurrence* is treated by debulking and chemotherapy.
    - Laparoscopic adrenalectomy is not advisable in adrenocortical carcinoma.

Secondaries in adrenal gland

- Common entity
- Commonly from breast, lungs, kidney, melanoma
- 1/3rd of above primaries show secondaries in adrenal gland
- Commonly bilateral
- In unilateral secondaries adrenalectomy can be done
- Bilateral secondaries are treated by chemotherapy depending on type of primary

Approches for adrenals (surgical)

- Posterior lumbotomy
- Anterior transabdominal
- Abdominothoracic
- Laparoscopic
- Retroperitoneoscopic

CUSHING'S SYNDROME

- Results from lengthy and inappropriate exposure to excessive glucocorticoids; iatrogenic administration of steroids (commonest cause); endogenous Cushing can also occur.
- More common in women.

Corticotropin (ACTH) Dependent—80%

- ACTH secreting pituitary adenoma.
- Ectopic ACTH secreting tumour.

Corticotropin Independent—20%

- Adenoma.
- Carcinoma.
- Macronodular adrenal hyperplasia.
- Primary pigmented nodular adrenal disease.

Features

- Centripetal obesity; redistribution of fat centrally.
- Moonface; fullness of supraclavicular pad of fat.
- Buffalo hump; proximal muscle wasting.
- "*Lemon on toothstick* appearance".
- Skin changes.

Diagnosis

- Loss of normal diurnal or circadian rhythm of serum cortisol. Values more than 4 fold the upper limit of normal is diagnostic.
- 24 hour urinary free cortisol is elevated > 90% of Cushings syndrome.
- *Low dose dexamethasone suppression test*: Dexamethasone is a potent glucocorticoid that suppresses adrenal production of corticosteroids in normal persons but not in persons with Cushing's syndrome. 1mg of dexamethasone is injected; normal—< 1.8 mcg/dl; in Cushings—high > 10 mcg/dl.
- *Late night salivary cortisol*—Cortisol concentration in saliva correlates with free plasma cortisol; it is independent of salivary flow rate; it shows high sensitivity and specificity.
- Negative urinary free cortisol and Low dose dexamethasone suppression test rules out Cushing's syndrome.
- *Measurement of corticotropin (ACTH*): If low—corticotropin independent Cushing's syndrome; if high—corticotropin dependent Cushing's disease.
- *Localising methods/imaging*: **Adrenal**—CT scan/MRI abdomen; Iodocholesterol scan. **Pituitary -** MRI pituitary. It can be incidentaloma in pituitary. Inferior petrosal sinus sampling of blood for ACTH shows central to peripheral ratio of ACTH > 3.

Treatment

Medical: Metapyrone; ketaconazole; aminoglutethamide; mitotane.

Surgical

For Pituitary Disease

- Transsphenoidal microsurgery.
- Pituitary irradiation—conventional fractionated therapy; Sterotactic radiosurgery.
- Bilateral adrenalectomy.

For Adrenal Disease

- Unilateral adrenalectomy in adenoma, carcinoma.
- Bilateral adrenalectomy in macronodular adrenal hyperplasia and primary pigmented nodular adrenal disease.

CONN'S SYNDROME (Jerome Conn, 1954)

It is primary hyperaldosteronism with excessive secretion of aldosterone from the adrenal gland associated with suppression of plasma renin activity.

Aldosterone secretion is related to angiotensin I and II and plasma rennin with angiotensin converting enzyme.

Causes

Aldosterone producing adrenocortical adenoma—65%; Idiopathic hyperaldosteronism—30%.

Features

- ***Hypertension; hypokalaemia.***
- Hypernatremia or normal sodium; metabolic alkalosis
- In primary hyperaldosteronism—hypertension of early onset which is difficult to control and is with hypokalaemia.

Diagnosis

- *Hypokalemia*; increase in urinary potassium excretion.
- *Elevated* plasma aldosterone concentration (PAC); *suppressed* plasma renin activity (PRA); PAC: PRA > 30; it is confirmed by suppression test by oral or IV salt loading.
- *Localising tests*—CECT abdomen; Selective adrenal venous sampling—*Gold standard* to differentiate between unilateral versus bilateral aldosterone hypersecretion.

Treatment

- Adenoma—unilateral adrenalectomy.
- Idiopathic hyperaldosteronism—medical treatment with spiranolactone.

VIRILISING SYNDROME OR ADRENOGENITAL SYNDROME

- Virilising tumours/syndromes. Such tumours are excised.
  - *In female*: Virilism, ambiguous external genitalia, clitoral enlargement.
  - *In male*: Precocious puberty, premature fusion of epiphysis, short stature.
- Congenital adrenal hyperplasia (adrenogenital syndrome) is—enzyme 21 hydroxylase deficiency; autosomal recessive. It is treated by—replacement of deficient steroids. Adrenal hyperplasia does not require surgical intervention but the genital manifestations of excess androgen production, particularly in women, may require specialised surgery.

NEUROBLASTOMA

- Commonest childhood tumour.
- It is a tumour of adrenal medulla.
- An aggressive malignant tumour in childhood usually below the age of 5 years.
- Incidence is equal in both sexes.
- A reddish-grey tumour gets invaded early into kidney, pancreas and adjacent tissues. Can also cause distant spread to liver, bones (skull), orbit.

It can occur anywhere in sympathetic chain but common in adrenal gland (40%).

Types

1. *Pepper type is right side* adrenal neuroblastoma with *liver secondaries.* Common in infants.
2. *Hutchinson's type is left side* adrenal neuroblastoma with *secondaries in orbit and skull*. Common in late childhood.

Secondaries in the skull mimics spicular osteogenic sarcomas.

Risk Groups

- Low risk groups—Stage I disease; Stage II disease with single N myc value; Stage II with favourable Shimada histology.

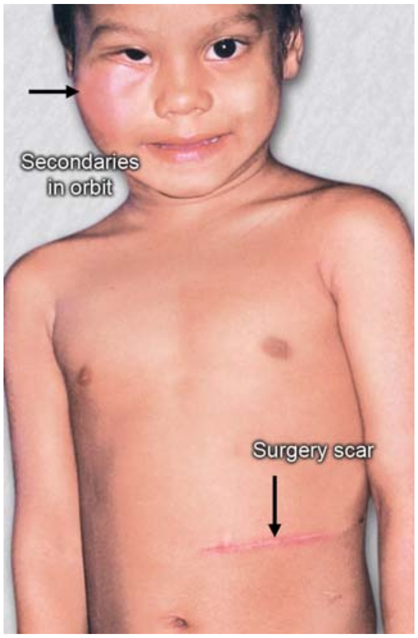

Fig. 7.5: Left-sided adrenal neuroblastoma with secondaries in orbit. Note scar of left sided adrenalectomy.

- Intermediate risk groups—Stage III without N myc amplification; Stage III with favourable Shimada's histology.
- High-risk groups—all patients with N myc amplification; stage IV neuroblastoma.

Pathology

- *Gross*—tumour with vascularity, necrosis, haemorrhage and often calcifications.
- *Histologically* it contains uniform round cells with hyperchromatic speckled nucleus with Homer-Wright rosettes with central fibrillar core. PAS stain is negative and NSE stain is positive. Often histochemistry is needed to differentiate from other tumours.
- *Shimada's histological classification:*
    - Stroma rich tumour: Presence of Schwannian spindle cell stroma.
    - Stroma poor tumour: Absence of Schwannian spindle cell stroma.

Clinical Features

- Child presents as a huge mass per abdomen, in the loin which is non-mobile. Not moving with respiration. Knobby (nodular) surface, crosses the midline.
- *Dancing eye syndrome and opsomyoclonus.*
- *Racoon's eye sign* is infraorbital ecchymosis due to secondaries in retroorbital region.
- Hypertension, fever, weight loss, anaemia, flushing (due to catecholamine release) and sweating.
- Diarrhoea, hypokalaemia due to release of Vasoactive Intestinal Polypeptide (VIP). Other hormones like ACTH are also released.

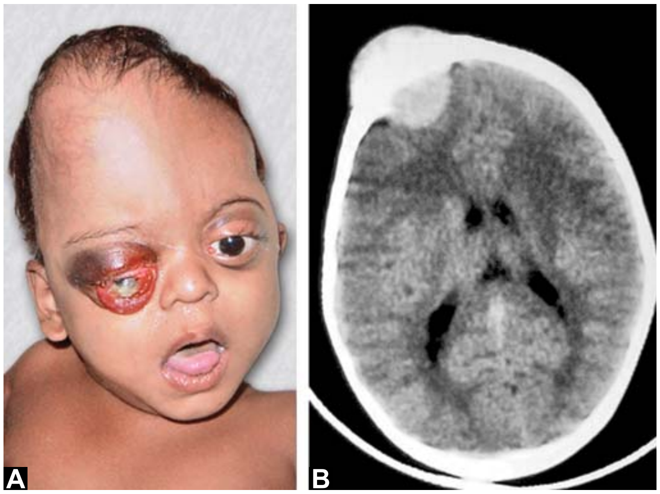

Figs 7.6A and B: Figure shows typical secondaries in skull and orbit with primary in adrenal gland. Such patients carry poor prognosis. *Racoon's eye sign* is infraorbital ecchymosis due to secondaries in retroorbital region. Dancing eye syndrome and opsomyoclonus are other eye features. CT picture of the same patient shows obvious secondaries.

Staging of neuroblastoma

- Stage I: Localised tumour with gross complete excision. Same side representative nodes are negative on microscopy
- Stage II A: Localised tumour with incomplete gross excision. Same side representative nodes are negative on microscopy
- Stage II B: Localised tumour with or without complete gross excision. Same side nonadherent representative nodes are positive but contralateral nodes are (may be enlarged) negative on microscopy
- Stage III: Unresectable unilateral tumour infiltrating across midline/localised tumour with opposite side node spread/bilateral nodal spread/midline tumour with bilateral infiltration
- Stage IV: Any primary tumour with spread to distant nodes, bone, bone marrow, liver, skin and other organs
- Stage IV S: Localised primary tumour (stage I, IIA, IIB) with spread to skin, liver or bone marrow—limited to infants younger than one year

Differential Diagnosis

Wilm's tumour, which is mobile, with smooth surface, moves with respiration, does not cross the midline.

Diagnosis

- U/S, CT to assess the mass and secondaries in liver. MRI is better than CT.
- Plain X-ray abdomen shows stippled calcification.
- Urinary VMA and Homovanillic acid (HVA) estimation in 90% cases.
- MIBG scan.
- Bone marrow biopsy may be positive in 60% cases.

Treatment

- Adrenalectomy (Complete surgical excision).
- In inoperable cases debulking is beneficial.
- Postsurgical radiotherapy and folic acid supplements are useful.
- Chemotherapy is useful adjuvant. Drugs used are carboplatin, doxorubicin, cyclophosphamide and etoposide.
- Rarely spontaneous regression of tumour is known to occur.
- Low risk groups are treated by surgery. Intermediate risk groups are treated by surgery and multidrug chemotherapy. High risk groups are treated by high dose multidrug chemotherapy and later surgery.

Prognosis

- It depends on staging of the neuroblastoma—1, 2A, 2B, 3, 4, 4S.
- Low risk has got 3 years survival—90%; intermediate risk—70% and high risk group has got 30%.
- Factors are—age of the child; stage of the disease; Shimada's histology; N myc amplification status (more means high risk); DNA ploidy; neurotrophin receptor Trk A (increased favourable); neutrophin receptor Trk B (increased unfavourable).

One only gets to the top rung of the ladder by steadily climbing up one at a time.

PHEOCHROMOCYTOMA

- It is a tumour arising from chromaffin cells, commonly from the adrenal medulla but occasionally can arise from extraadrenal chromaffin tissues (*Organ of Zuckerkandl*).
- It is catecholamine secreting tumours that arise from chromaffin cells of sympathetic origin derived from neural crest representing a potentially curable form of hypertension.
- *Incidence* is 0.005% to 0.1% of general population; 0.1% to 0.2% of adult hypertensive population.
- It is a soft, brownish grey pink tumour, mainly *secretes noradrenaline* or other catecholamines.
- It may also secrete calcitonin, ACTH, VIP (vasoactive intestinal polypeptide), PTH related polypeptide.
- Prevalence of phaeochromocytoma is 0.05%. In patients with hypertension it is up to 0.6%. 4% of incidentalomas are pheochromocytoma.
- Often it is difficult to differentiate between benign and malignant types. Necrosis, haemorrhage, high Ki-67 positive cells, size of the tumour, increased pheochromocytoma of adrenal gland scale score (PASS), capsular invasion and vascular invasion, nuclear DNA ploidy and increased neuron specific enolase (NSE) level are possible features of malignant phaeochromocytoma.
- Currently mutations in at least six distinct genes predispose to pheochromocytomas—RET, NF1, VHL, SDHB, SDHC, SDHD.
- Extra-adrenal pheochromocytoma—10 % common; occurs in organ of *Zuckerkandl*, urinary bladder, paravertebral or para-aortic area, thorax, neck. It secretes norepinephrine rather than epinephrine because they lack the enzyme PNMT.
- Commonly benign (90%).

Tumour is

- 10% malignant
- 10% extraadrenal
- 10% bilateral
- 10% familial
- 10% childhood
- 10% multiple
- 10% not associated with hypertension
- 10% calcified

Clinical Features

- Clinical manifestations are due to increased secretion of epinephrine and norepinephrine.
- Commonest presentation is severe *headache.*
- Palpitation, dyspnoea, weakness, pallor, blurred vision and other symptoms of sympathetic overactivity.
- They may present as persistent or paroxysmal hypertension (90%).
- As an abdominal mass which is nonmobile, smooth, does not move with respiration, crossing the midline, palpitation may cause fluctuation in BP.
- It may precipitate hypertensive encephalopathy, cardiac arrhythmias or cerebral haemorrhage.
- Panic attacks and sudden death are known to occur.
- It may be associated with ***MEN-IIa*** or ***MEN-IIb*** syndromes which includes medullary carcinoma of thyroid and mucosal neuromas.
- It may be associated with familial multiple neurofibromatosis with cafe au lait spots in the skin ***(von Recklinghausen disease***). Or with ***von-Hippel-Lindau syndrome*** (cystic lesion of pancreas, non-functioning islet cell tumour, phaeochromocytoma), RCC, CNS and Retinal haemangioblastoma.
- *Familial paraganglioma syndrome* may be an association with carotid body and extraadrenal paraganglioma.

Differential diagnosis

- Hyperthyroidism
- Anxiety status
- Cardiac conditions
- Carcinoids (functioning)

Investigations

- *VMA excretion* in urine in 24 hours will be >7 mg/24 hr in pheochromocytoma.
- U/S abdomen, IVU, CT scan.
- MRI is preferred to CT as contrast used for CT scan can precipitate paroxysms.
- Measurement of *plasma free metanephrines* is the recommended *test of choice* for excluding or confirming diagnosis of pheochromocytoma.
- *Urinary normetadrenaline* or other catecholamines estimation.
- Arteriography.
- Iodine labelled metaiodobenzylguanidine *(I, MIBG*). MIBG is useful to find out extraadrenal involvement—SPECT scan is very useful. I^{131} MIBG scan is safe, noninvasive with 100% sensitivity and 95% specificity.
- Measurement of plasma free metanephrine and normetanephrine has the *highest sensitivity and specificity* and appears to be the *best initial test* for screening patients with pheochromocytoma.

Treatment

Adrenalectomy

- Before surgery, BP is controlled initially by α-blocking agent, phenoxybenzamine or doxazosin; then by β-blocking agent, propranolol. Then during surgery, sodium nitroprusside IV infusion is used.
- Careful anaesthetic management with a good postoperative care is very essential. In case of bilateral presentation opposite side can be operated in later period.
- Intraoperative hypovolemia and postoperative hypotension is a demanding situation for anaesthetist to manage.
- Handling of adrenal tumour on table must be careful and gentle.
- Adrenal vein should be ligated first. CVP, arterial lines should be present for monitoring.
- Rupture and spillage of tumour should be prevented.

Note:

- *Laparoscopic adrenalectomy is becoming popular*. It is choice of approach for benign functioning or non-functioning adrenal tumours that are less than 6 cm in size. It is contraindicated in malignant tumours and tumour more than 6 cm.
- *Recent trends are*—cortical sparing adrenalectomy; thoracoscopic transdiaphragmatic approach; Robotic assisted laparoscopic adrenalectomy.
- Specimen should be sent for *bichromate staining which stains the specimen brow*n.

Remember

- ❖ Pheochromocytoma is rarely malignant in MEN II
- ❖ Pheochromocytoma under 40 years of age suspect MEN 2A, VHL
- ❖ Beta-blocker is given only after patient is fully alpha blocked with phenoxybenzamine (20-60 mg/day) or doxazocin
- ❖ Alpha-blocker is given 4 weeks prior to surgery to control hypertension and beta-blocker is given one week before surgery to control tachycardia and arrhythmias
- ❖ Tumours, those secrete dopamine exclusively has got high malignant chances
- ❖ 5-year survival for malignant pheochromocytoma is 50%. It needs additional chemotherapy using vincristine, dacarbazine and cyclophosphamide.
- ❖ Pheochromocytoma in pregnancy has got 50% maternal mortality
- ❖ Vaginal delivery is contraindicated in pregnancy with pheochromocytoma
- ❖ Adrenal vein should be ligated first
- ❖ Avoid breach in the capsule of tumour during surgery
- ❖ Careful handling and haemodynamic monitoring is a must
- ❖ Sodium nitroprusside may be required on table to control the hypertension—10 μg/kg/minute

Malignant pheochromocytoma is 10% common. It is more common in extraadrenal site. It commonly spreads to lymph nodes, bone and liver. Adrenalectomy in early tumour and debulking in advanced cases with α blockers, mitotane and I^{131}MIBG therapy and combination chemotherapy are the therapeutic choices. Overall prognosis is not good with 5-year survival being < 50%.

Chapter 8 Breast

CHAPTER OUTLINE

- Anatomy
- Mammography
- Aberration of Normal Development and Involution (ANDI) of the Breast
- Fibroadenoma
- Fibrocystadenosis
- Sclerosing Adenosis
- Phyllodes Tumour
- Mastalgia
- Traumatic Fat Necrosis
- Galactocele
- Mastitis
- Antibioma
- Duct Ectasia
- Mondor's Disease
- Tuberculosis of the Breast
- Breast Cysts
- Galactorrhoea
- Gynaecomastia
- Duct Papilloma
- Zuska-Atkins Disease
- Mammary Fistula of Atkins
- Carcinoma Breast
- Management of Early Carcinoma Breast
- Advanced Carcinoma Breast
- Prognostic Factors in Carcinoma Breast
- Prophylactic Mastectomy
- Carcinoma of Male Breast
- Breast Reconstruction
- Breast Implants
- Nipple Retraction

ANATOMY

Breast is a modified sweat gland derived from ectoderm, as branching epithelial cords which form lactiferous ducts. About 15-20 lobes develop during puberty, each of which drains into a single lactiferous duct. True secretory alveoli develop during pregnancy and lactation under the influence of oestrogen, progesterone and prolactin.

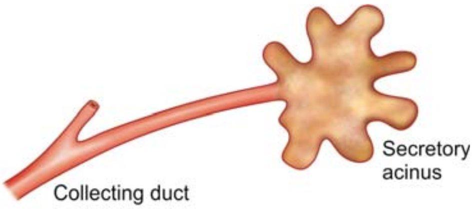

Fig. 8.1: Secretory apparatus of the breast.

Secretory Apparatus of Breast

The breast is composed of acini which make up lobules, aggregation of which form the lobes of the gland. The lobes are arranged in a radiating fashion like the spoke of a wheel and converge on the nipple, each lobe is drained by a lactiferous duct.

Different portions of the duct system are associated with different diseases.

1. Larger ducts are sites of *duct papilloma and duct ectasia*.
2. The distal smaller ducts are the sites of *fibroadenoma* during development of the breast; and *cyst formation and sclerosing adenosis* during the involution period.

- *Estrogen*—ductal proliferation
- *Progesterone*—glandular proliferation
- *Prolactin*—milk secretion

Topography

Vertically—it extends from the second to the sixth rib in the mid-clavicular line and lies over pectoralis major, serratus anterior and external oblique muscles.
Horizontally—from the side of sternum to the mid-axillary line.

2/3rd of the breast rests upon pectoralis major.
1/3rd rest upon serratus anterior.

At its lower medial quadrant the gland rests on the external oblique aponeurosis, which separates it from the rectus abdominis.

Breast lies in superficial fascia. Deep to breast, structures related are—retromammary space containing loose areolar tissue, pectoral deep fascia, muscles (pectoralis major, serratus anterior, external oblique), chest wall.

Retromammary bursa/space is located between deep layer of superficial fascia and pectoral (deep) fascia allowing free mobility of breast.

Nipple is located at the level of 4th intercostal space just below the center/summit of the breast. It contains circular and longitudinal muscles to make nipple stiff or flat. It is pierced by 15-20 lactiferous ducts. Each duct independently opens into the nipple. It has rich sensory nerve endings. It also contains modified sweat and sebaceous glands. Nipple is supplied by 4th intercostal nerve.

Areola is circular pigmented area around the nipple. It is rich in modified sebaceous glands which enlarge during pregnancy and lactation as *Montgomery tubercles*. They secrete oily lubricant to nipple and areola. Areola and nipple do not contain hair and fat beneath.

Breast parenchyma contains 15-20 lobes. Each lobe contains alveoli, lactiferous sinus and lactiferous duct (2-4 mm in diameter). Alveolus is lined by cuboidal (in rest) and columnar (in lactation) epithelium; smaller duct is by single layer of columnar epithelium; larger ducts by many layered columnar; lactiferous duct is by stratified squamous epithelium. *Myoepithelial cells* lie between epithelium and basement membrane from alveoli to duct.

Axillary Tail of Spence

This is a prolongation from the outer part of the gland which passes up to the level of the 3rd rib in the axilla through a defect in the deep fascia (*Foramen of Langer*) where it is in direct contact with the main lymph node of the breasts (anterior axillary nodes).

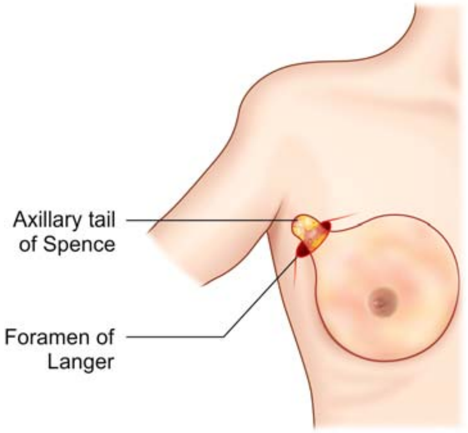

Fig. 8.2: Axillary tail of *Spence*.

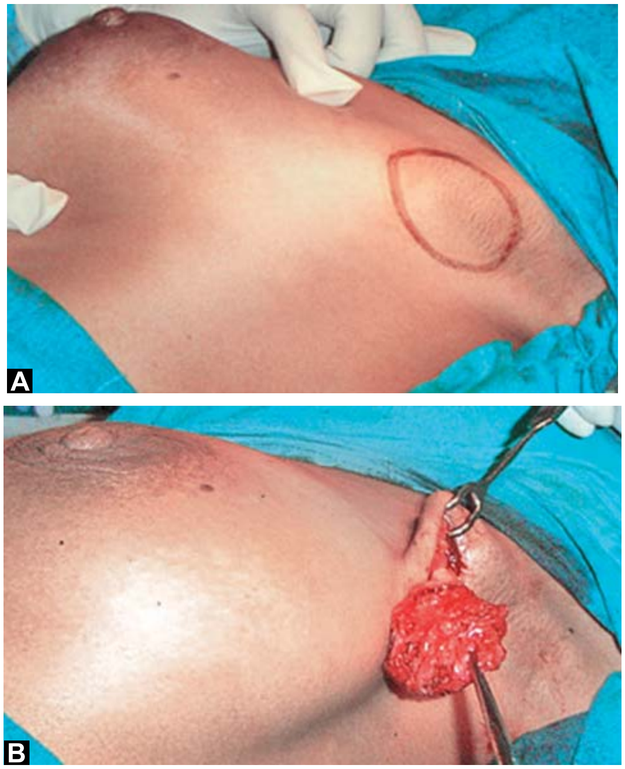

Figs 8.3A and B: Axillary tail of *Spence*–exact location and excision of hypertrophied axillary tail.

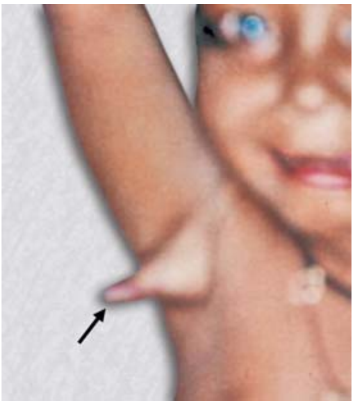

Fig. 8.4: Hypertrophy of the axillary tail.

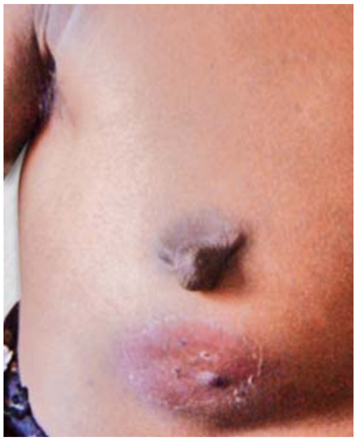

Fig. 8.5: Accessory nipple.

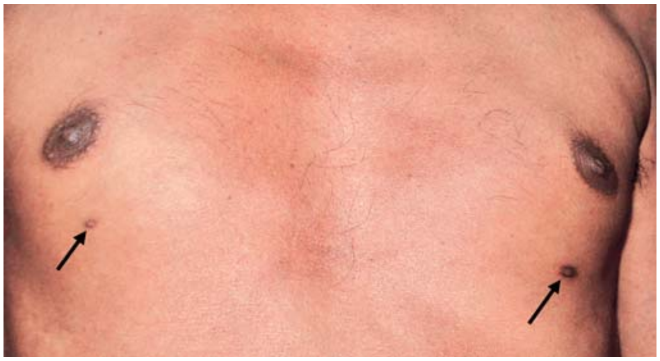

Fig. 8.6: Accessory nipple is not an uncommon condition. Often it can be bilateral.

This process of the breast gets into axilla through an opening in the deep fascia, known as *foramen of Langer.* When it enlarges it is ofen mistaken for a lipoma. Axillary tail of Spence is deep to deep fascia.

Ligament of Cooper

The breast is anchored to the overlying skin and to the underlying pectoral fascia by bands of connective tissue called *ligament of Cooper.*

1. In cancer, the malignant cells may invade these ligaments and consequent contraction of these strands may cause dimpling of the skin or attachment of the growth to the skin, which in turn cannot be pinched off from the lump.
2. If the cancer grows along the ligament of Cooper binding the breast to the pectoral fascia, the breast gets fixed to the pectoralis major. It then cannot be moved along the long axis of the muscle.

Polands syndrome

- ❖ Absence of breast—Amazia
- ❖ Absence of sternal portion of pectoralis major
- ❖ Common in males

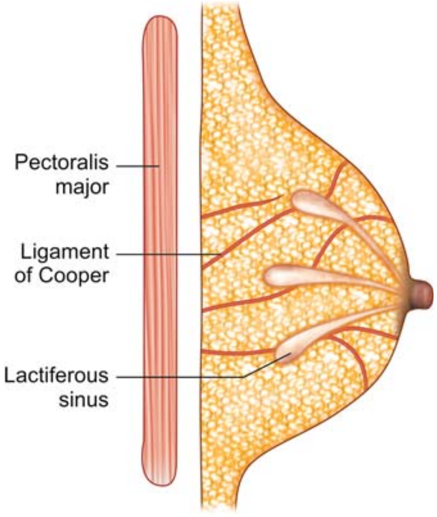

Fig. 8.7: Anatomy of the breast.

Note:

Amazia is very rare entity with congenital absence of breast.
Polymazia is accessory breasts along milk line like in axilla (most common site), groin, thigh and buttock.
Athelia is absence of nipple; accessory nipples are called as *polythelia* or supernumerary nipples.
Micromastia is small sized breast due to less hormonal stimulation due to ovarian deficiency.

Blood Supply to the Breast

- ♦ The lateral thoracic artery, from the 2nd part of the axillary artery—30%.
- ♦ The perforating cutaneous branches of internal mammary artery to the 2nd, 3rd and 4th intercostal spaces—60%.
- ♦ The lateral branch of the 2nd, 3rd and 4th intercostal arteries.
- ♦ Pectoral branches of acromiothoracic artery.
- ♦ Superior thoracic artery.

Venous Drainage

- ♦ The superficial veins from the breast characterised by their proximity to the skin drain to the axillary, internal mammary, and intercostal vessels.
- ♦ Phlebitis of one of these superficial veins feel like a cord immediately beneath the skin—***'Mondor's disease'.***
- ♦ Through posterior intercostal veins, venous drainage communicates with paravertebral venous plexus (*Batson's venous plexus*). So secondaries in vertebrae, is common in carcinoma of breast.

Nerves-related (During MRM)

- ♦ Long thoracic nerve—*Bell* supplies serratus anterior.
- ♦ Thoracodorsal nerve supplies latissimus dorsi.
- ♦ Medial pectoral nerve (from medial cord of brachial plexus) which lies *lateral,* runs and winds from lateral margin of pectoralis minor.
- ♦ Lateral pectoral nerve arises from lateral cord passes through the pectori either middle or medial part.
- ♦ Intercostobrachial nerve is communicating nerve between lateral cutaneous branch of 2nd intercostal nerve and medial cutaneous nerve of arm; denervation of this nerve causes sensory loss of skin over upper medial and inner aspect of the arm, apex and lateral axilla.

Nerve supply of breast is by anterior and lateral cutaneous branches of 4th to 6th intercostal nerves.

Milk secretion is brought by prolactin hormone from anterior pituitary—not by nerves.

Lymphatic Drainage of the Breast

Commonly into the axillary lymph nodes—75%.

1. Anterior group (pectoral, external mammary)—along lateral thoracic vessels. Main drainage node.
2. Central group—next common node. It is the node most easily properly clinically palpable in axilla.
3. Posterior group (subscapular)—rare to involve in carcinoma.

4. Lateral group—along axillary vein; rare to involve in carcinoma.
5. Interpectoral node *(Rotter's node)*—signifies the retrograde spread of tumour. It lies between pectoralis major and minor.
6. Apical. They are 4-6 nodes, also called as subclavicular or *Halsted nodes*. It lies most superior and deep to pectoralis minor medial to axillary vessels.

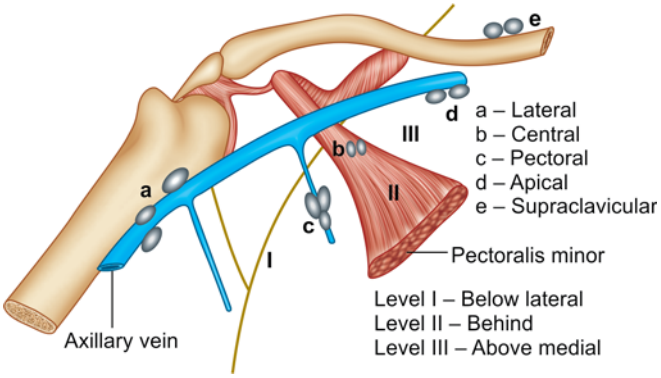

Fig. 8.8: Surgical levels of lymph nodes in the axilla draining the breast.

Later they drain into supraclavicular lymph nodes. 25% drains mainly from medial half of the breast into 2nd, 3rd and 4th intercostal space internal mammary lymph nodes. *Internal mammary nodes* are located in retrosternal intercostal spaces 1-2 cm lateral to the sternal margin; it is vertically placed parallel to internal mammary vessels in relation to endothoracic fascia; its efferent ends in subclavicular nodes.

Drainage into contralateral axilla and opposite lymph nodes also occur.

The concept of Sappey's subareolar drainage of lymph centripetally is not well accepted now.

Axillary reverse mapping (ARM): Here blue dye is injected into upper part of inner aspect of the arm to achieve a mapping of axillary lymphatic drainage.

Levels of the axillary nodes (Berg's levels)

- **Level I :** Below and lateral to the pectoralis minor muscle—anterior, lateral, posterior
- **Level II :** Behind the pectoralis minor muscle—central
- **Level III :** Above and medial to pectoralis minor muscle—apical

Note in carcinoma breast

- Spread restricted to level I nodes carries better prognosis
- Spread to level II has poor prognosis
- Spread to level III indicates worst prognosis

MAMMOGRAPHY

It is a plain X-ray of soft tissue of breast using low voltage and high ampearage X-rays. Two films are taken.

1. *Craniocaudal* from above downward.
2. *Mediolateral* from side to side.

Dose of radiation is 0.1 Gy, a low dose. So it is a safe and effective procedure.

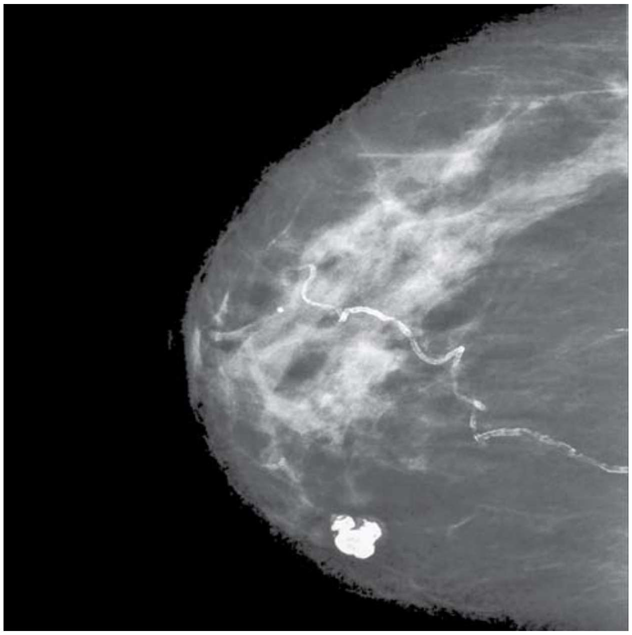

Fig. 8.9: Mammography showing calcified lactiferous duct with a benign lesion.

Findings

- Microcalcifications signify malignancy. They are < 0.5 mm, pleomorphic/heterogenous, grouped, fine, linear and branching.
- Soft tissue shadow may be smooth, round/oval and regular in benign conditions or irregular in carcinomas. Lobulated mass is also suggestive of carcinoma.
- Size and location of mass lesion.
- Spiculations, duct distortion—features of malignancy.
- Margin may be circumscribed (benign lesion), microlobulated, obscure, ill-defined, spiculated or stellate.

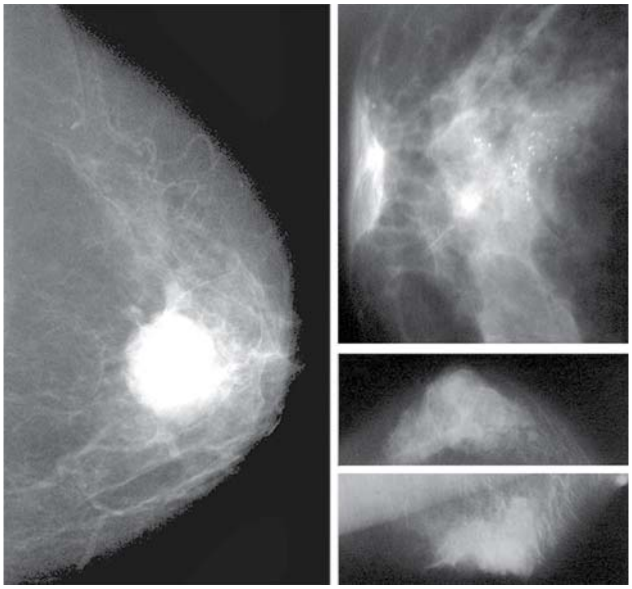

Fig. 8.10: Different mammography pictures, craniocaudal and mediolateral films showing mass lesion, microcalcification (*Courtesy:* Dr Ganesh, MDRD, Mangalore).

Breast imaging reporting and data system (BIRADS)

Grade	Feature
0	Need further imaging
1	Negative
2	Benign—repeat mammography 1 year
3	Probably benign—mammography after 6 months
4	Suspicious of carcinoma—biopsy
5	Highly suggestive of carcinoma—biopsy
6	Known carcinoma

Indications

- *Screening mammography:* For screening purpose it is done after 40 years. Early screening is indicated when there is family history of carcinoma breast or histological risk factor. Mammography before 35 years of age is usually not done unless there is a suspicious lump or a strong family history. Screening is done in asymptomatic female where even cancer is suspected.
- *Diagnostic mammography* is done in a patient to evaluate the existing symptoms/features of breast disease like breast lump, nipple discharge; or in patient who had breast conservative surgery earlier.
- In obese patients.
- Whenever conservative breast surgeries are planned.
- To find out spread or de novo tumour in the opposite breast.
- Mammography guided biopsy can be done.
- Evaluation and follow-up in benign breast disease with malignant potential.
- Follow-up mammography after conservative breast surgery (of operated side and opposite side when needed.
- Mastalgias.

***Note*:**

- 5% of breast cancers are missed in total mammograhic screening.
- Breast imaging reporting and data system (BIRADS) have got its own categories, assessment (0-6) and recommendations.
- Digital mammography is computerised electronic image of the breast with enhanced magnified pictures.
- Digital spot-view mammography allows faster and more accurate stereotactic biopsy.
- The condition when lump is clinically not palpable but mammogram shows identifiable carcinoma is ideal for breast conservative surgery like quadrantectomy/QUART therapy.

Xeromammography is same as above, but here a photoconductor is used to produce a final image on a Selenium paper rather than on X-ray film.

Advantages: Edge enhancement effect, therefore, useful in dense breasts.

Disadvantage: Exposure to high radiation dose and selenium plates are needed.

ABERRATION OF NORMAL DEVELOPMENT AND INVOLUTION (ANDI) OF THE BREAST

ANDI includes variety of benign breast disorders occurring at different periods of reproductive periods in females—early, matured and involution phase of reproductive age group. It was first coined at Cardiff breast clinic in 1987 by LE Hughes. All conditions under ANDI should be carefully clinically examined and often mammography and FNAC/core cut biopsy should be done to rule out malignancy. ANDI includes different aberrations and diseases.

It is based on change in *normal three phases* of physiology of breast—(1) Lobular development; (2) Cyclical hormonal modifications; (3) Involution.

In early reproductive age group (15-25 years):

- Normal lobule formation may cause aberration as *fibroadenoma*. If it is more than 5 cm it is called *as giant fibroadenoma* as a *diseased status*. It is AND of a lobule.
- Normal stroma may develop juvenile hypertrophy as aberration and multiple fibroadenoma as *diseased status*.

In mature reproductive age group (25-40 years):

- Normal cyclical hormonal effects on glands and stroma get exaggerated by aberration causing generalised enlargement. Its *diseased status* is cyclical mastalgia with nodularity also called as *fibrocystadenosis.*

Involution age group (40-55 years):

- *Lobular involution* with microcysts, fibrosis, adenosis, apocrine metaplasia and eventual aberrations as macrocysts and cystic disease of breast. Macrocyst is an aberration of normal involution (ANI). Sclerosing adenosis is also a type of aberration.
- *Ductal involution* may cause ductal dilatation and nipple discharge as aberration. Later disease status develops with periductal mastitis, bacterial infection, nonlactational breast abscess and mammary duct fistula. Periductal fibrosis may cause partial nipple retraction.
- *Epithelial changes* leads into epithelial hyperplasia and atypia.

FIBROADENOMA

- It is a benign encapsulated tumour occurring commonly in young females of 15-25 years age group.
- Presently it is considered as hyperplasia of a single lobule of the breast (classified under ***ANDI****)*.
- It is the most common benign tumour of the breast below 30 years of age in females.
- It is aberration in normal development (AND) of a lobule.
- It shows similar hormonal activities of normal breast like lactation, perimenopausal involution.
- Incidence is 15% of palpable breast lumps. It is common in blacks and Negroes.
- It is bilateral in 20% of cases. 20% are multiple.

- *Juvenile fibroadenoma* occurs in adolescent girls, rarely (variant). Even though it shows rapid growth with stromal and epithelial hyperplasia, it does not show any alteration in stromal epithelial balance or cellular atypia or periductal cellular concentration. It may clinically mimic phyllodes tumour. But it does not turn into phyllodes tumour or carcinoma.
- *Complex fibroadenoma* (*Dupon et al)* is a condition (variant) having typical fibroadenoma with fibrocystic changes like apocrine metaplasia, cyst formation, sclerosing adenosis. 15% of proven fibroadenomas are complex. It occurs in older age group. *Occasionally* it may turn into malignancy unlike usual fibroadenomas. *Core biopsy* is needed to confirm the condition.
- 30% of fibroadenomas may disappear or reduce in size in 2-4 years.
- 10 -15% will increase in size progressively. It does not occur after menopause unless women are on hormones.
- Fibroadenoma *does not* turn into malignancy.

Types

Gross:

1. Soft—common after 30 years; more cellular; often bilateral.
2. Hard—common below 30 years; more fibrous.
3. Giant (> 5 cm in size)—common in Africa.

Microscopy:

1. Intracanalicular—large and soft—mainly cellular. Stroma with distorted duct.
2. Pericanalicular—small and hard—mainly fibrous. Stroma with normal duct.

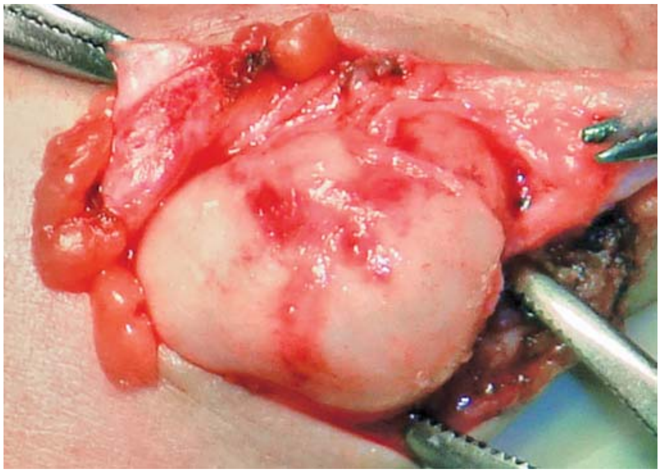

Fig. 8.11: Fibroadenoma, on table look—well-capsulated neoplasm. It is part of ANDI.

Clinical Features

- It presents as a painless swelling in one of the quadrants, which is *smooth, firm, nontender, well-localised and moves freely* within the breast tissue (mouse in the breast).
- Axillary lymph nodes are not enlarged.

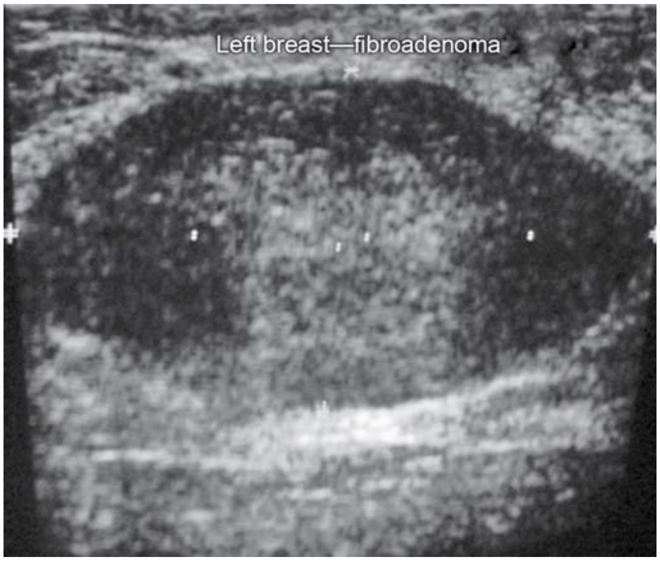

Fig. 8.12: US picture showing fibroadenoma left breast.

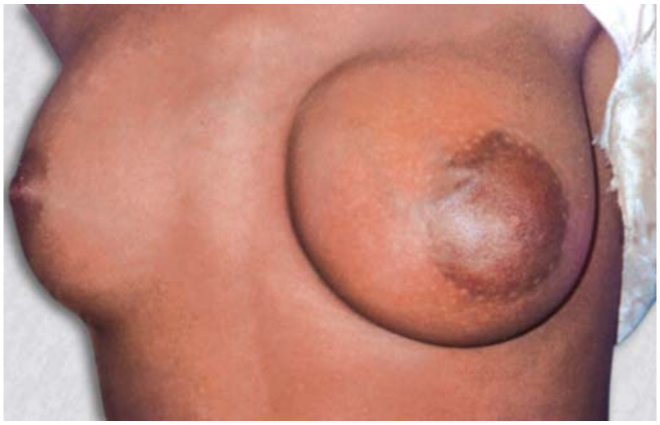

Fig. 8.13: Large fibroadenoma left breast in a 14-year-old female.

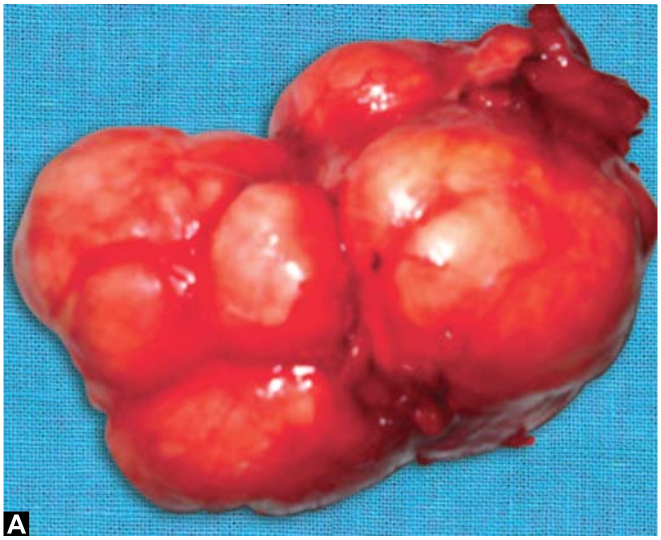

Fig. 8.14A

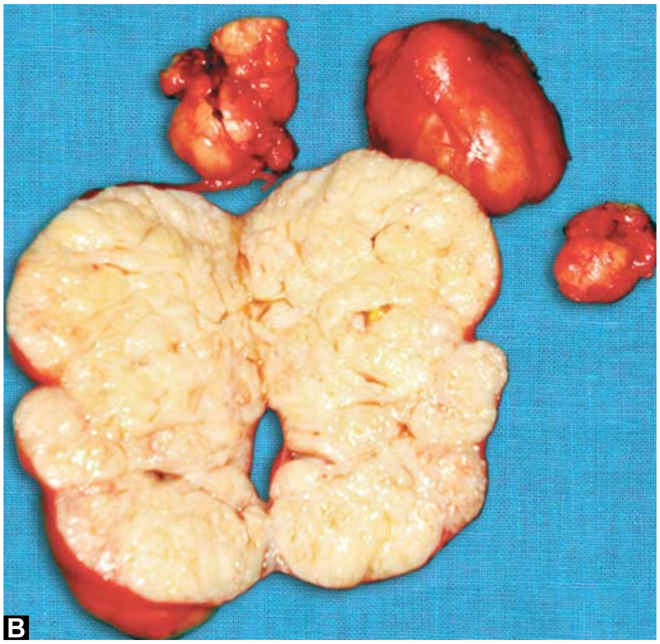

Fig. 8.14B

Figs 8.14A and B: Multiple fibroadenomas.

Investigations

- Mammography (well-localised smooth regular shadow).
- FNAC.
- Ultrasound (to confirm solid nature).

Treatment

Excision through a circumareolar incision (*Webster's*) or submammary incision (*Galliard Thomas incision*) is done under general anaesthesia.

Fibroadenoma which is small (< 3 cm)/single/age < 30 years can be left alone with regular follow-up with USG at 6 monthly interval. But anxiousness of patient and parents find difficult for this conservative approach.

Indications for surgery are:

- Size > 3 cm.
- Multiple.
- Giant type.
- Recurrence.
- Cosmesis.
- Complex type.

Points to be remembered

- *Size of giant duodenal ulcer is > 2 cm*
- *Size of giant gastric ulcer is > 3 cm*
- *Size of giant fibroadenoma is > 5 cm*
- *Diameter of transverse colon in toxic megacolon is > 6 cm*
- *Size of giant naevus is > 20 cm*

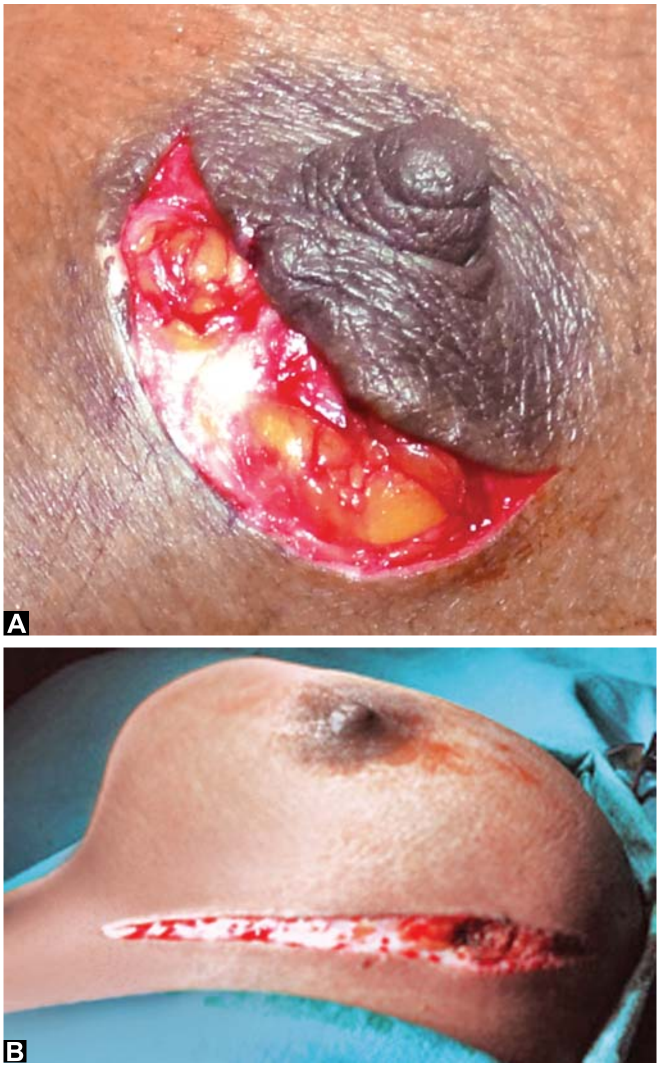

Figs 8.15A and B: Subareolar and submammary incision to excise benign lesions of the breast.

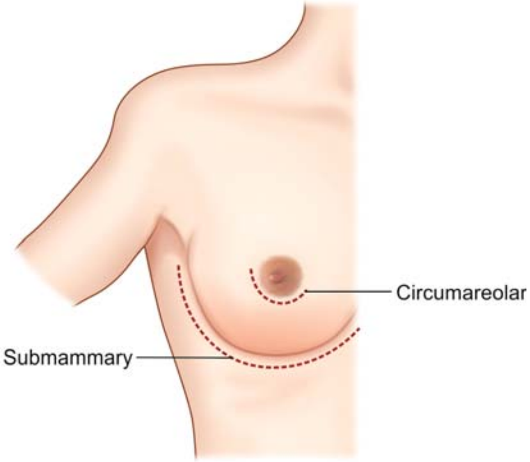

Fig. 8.16: Diagram showing circumareolar and submammary (*Galliard Thomas*) incisions.

FIBROCYSTADENOSIS (FIBROCYSTIC DISEASE OF THE BREAST/MAMMARY DYSPLASIA/CYCLICAL MASTALGIA WITH NODULARITY)

- It is due to **A**berration of **N**ormal **D**evelopment and **I**nvolution (***ANDI***) of breast causing.
- It is presently called as *cyclical mastalgia with nodularity.*

Microscopic changes

- Stromal fibrosis
- Microcyst formation
- Glandular proliferation (Adenosis)
- Hyperplasia (Epitheliosis)—in ducts and acini
- Papillomatosis—within the ducts, often with apocrine metaplasia

- It is an estrogen dependent condition. One of the cysts may get enlarged to become a clinically palpable, well-localised swelling—*bluedome cyst of Bloodgood.* It is fluctuant, transilluminant, nontender, often tensely cystic swelling (macrocyst) with thin bluish capsule. It should be aspirated initially. Surgical excision is done if it persists or recurs even after two aspirations; if it is blood stained; if there is residual lump after aspiration.
- When diffuse small, multiple cysts are the main component, it is called as *Schimmelbusch's disease.*
- *It is the most common breast disease.*
- It is common in upper and outer quadrant.
- It is an exaggerated response of breast stroma and epithelium to hormones and growth factors.
- It is rare in nulliparous/ovulating/OCP taking women.

Classification

- *Nonproliferative—moderate hyperplasia of ductal luminal cells*—no life time risk for cancer.
- *Proliferative without atypia (severe hyperplasia).*
- *Proliferative with atypia (Atypical ductal/lobular hyperplasia)*—is risk factor for breast cancer—often mimics *in situ* carcinoma. RR is 4.5.

Usual hyperplasia is presence of 3 or more layers of proliferating epithelial cells without atypia above the basement membrane in a lobular or ductal unit. It can be mild, moderate or florid. It is seen in 20% of biopsies. Risk of developing cancer is 1.5 times. Observation of such patient is the advice.

Atypical ductal hyperplasia (ADH) is defined as a lesion that has got some but not all features of DCIS; a lesion that has all features of DCIS but less than 2 mm in size; lesion with all features of DCIS but involving less than 2 duct spaces. It is incidence is 30% of all mammographies and 5% of benign breast disease biopsies. Risk of developing cancer is 5 times the normal women of that age.

Atypical lobular hyperplasia (ALH) is LCIS with half of the acini of lobular unit is involved. Developing invasive carcinoma in 10 years is 20%.

Stages

I—Stromal proliferation or hyperplasia.
II—Adenosis (increased glands).
III—Cyst formation.

Pathophysiology of fibrocystadenosis

- Oestrogen predominance over progesterone is considered causative
- Serum levels of oestrogen ≥
- Luteal phase is shortened
- Progesterone level decreased to 1/3rd normal
- Corpus luteum deficiency/anovulation in 70%
- Patients with premenstrual tension syndrome more likely to develop fibrocystadenosis
- Prolactin levels are increased in 1/3rd of women with fibrocystadenosis. It is probably due to oestrogen dominance on pituitary
- Thyroid in suboptimal levels sensitize mammary epithelium to prolactin stimulation
- Methylexanthiones derived from increased intake of coffee, tea, cold drinks, chocolate is associated with development of fibrocystadenosis
- Oestrogens stimulate proliferation of connective and epithelial tissues
- Fibrocystadenosis entails simultaneous progressive and regressive change

Clinical Features

- Presentation is during menstruating age group as a *bilateral, painful, diffuse, granular, tender, swelling* which is better felt with palpating fingers (poorly felt with palm).
- Common in upper outer quadrant.
- Pain and tenderness are more just prior to menstruation *(cyclical mastalgia).*
- It subsides during pregnancy, lactation and after menopause.
- Discharge from the nipple when present will be serous or occasionally greenish.
- Occasionally shotty enlargement of axillary lymph nodes can occur (20%).
- Not fixed to skin, muscle or chest wall.

Investigations

- FNAC (Epitheliosis, when florid is undoubtedly premalignant).
- Ultrasound.
- Mammography.

ANDI was coined by Cardiff breast clinic.

Treatment

I. *Conservative line of management* is preferred.
  1. *Reassurance,* avoid caffeine, chocolate, salt.
  2. *Medical (Drugs)*
    - Goal:
      - To stop progression.
      - To relieve pain.
      - To reverse changes.
      - To soften breast tissue.
    - Indicated when:
      - Fibroadenosis is not increasing in size.
      - No nipple discharge especially blood.
      - No psychological effect.

Drugs are:

- *Oil of evening primrose* used in moderate pain—*drug of choice*. It contains gamolenic acid which reverses saturated to unsaturated fatty acids. 1000-3000 mg/day for 4-6 months—but costly. It also contains 7% of linolenic acid and 72% of linoleic acid.
- Gamolenic acid—120 mg/day.
- Danazol—interferes with FSH and LH (gonadotrophin releasing hormone inhibitor); *most effective drug*; but *second drug of choice*; used in severe cases; 200 mg/day; very effective but causes acne, hirsutism, weight gain and amenorrhoea. It is teratogenic and so cannot be used if patient is planning for pregnancy.
- Bromocriptine—lowers prolactin—2.5 mg/day for 3 months.
- Tamoxifen—10 mg BD is an antiestrogenic drug.
- LHRH agonist (Goserelin) is *reserved for refractory cases*. It shows 96-99% success. But it causes reversible postmenopausal symptoms.
- Vitamin E and B_6 are tried.
- NSAIDs—oral and topical.
- Diuretics even though used by many—*not effective.*

II. *Surgery*:

- Subcutaneous mastectomy with prosthesis placement—only in severe, persistent disease.
- Excision of the cyst or localised excision of the diseased tissue.

Indications for surgery

- Intractable pain
- Florid epitheliosis—on FNAC
- Blood good cyst
- Persistent bloody discharge
- Psychological reason

***Note*:**

- *Subcutaneous mastectomy* is removal of entire breast with retaining skin over the breast, areola and nipple. It is done through a submammary Galliard Thomas incision. Adequate skin flap containing subcutaneous fat is raised which maintains the blood supply of the flap and prevents flap necrosis. After haemostasis drain is placed. Breast implant can be placed in subcutaneous/submuscular plane either immediately or as delayed reconstruction. Indications for subcutaneous mastectomy are—fibrocystadenosis with epitheliosis, sclerosing adenosis, persistent nodules, gynaecomastia and DCIS.
- *Macrocysts* (>1 cm) is an advanced form of fibrocystic disease; occurs in women in their forties and pericystic fibrosis develops later making cyst harder.

SCLEROSING ADENOSIS

Known cause of breast lump.

Features

- Occurs in 30-50 years of age.
- Patient may present as breast lump or mastalgia.
- On palpation—it is smooth, relatively mobile mass.
- It can mimic carcinoma clinically, radiologically and histologically.
- It contains proliferative terminal ductules and acini with proliferation of stroma often with deposition of calcium.
- Number of acini per terminal duct is increased more than double the number of normal lobule.
- *Complex type* is with papilloma and epithelial hyperplasia. Radial scar is variant of this.
- No risk of malignancy.
- It is included under ANDI.
- Treatment is like ANDI.

Histology

- Lobular enlargement and distortion.
- Fibrous stromal proliferation.

PHYLLODES TUMOUR (CYSTOSARCOMA PHYLLODES/SEROCYSTIC DISEASE OF BRODIE)

- They are not simply giant fibroadenoma.
- They show a wide spectrum of activity, varying from almost a benign condition (85%) to a locally aggressive and sometimes metastatic tumour (15%).
- Depending on mitotic index and degree of pleomorphism they are graded as low grade to high grade tumours.
- When malignant (sarcoma) spreads to lungs or bone.

Gross: Large capsulated area with cystic spaces and cut surface shows soft, brownish, cystic areas.

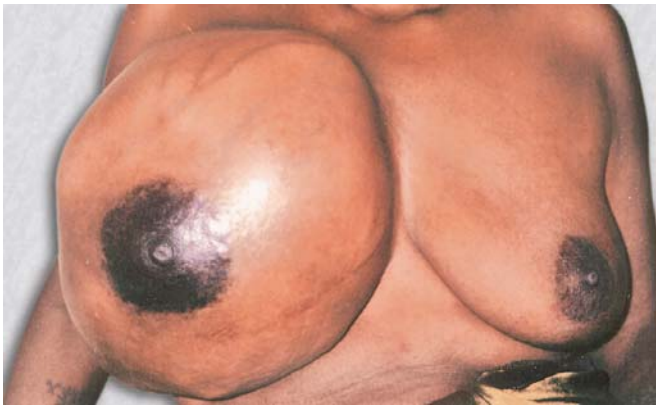

Fig. 8.17: Cystosarcoma phyllodes of right breast. Note the dilated veins. Tumour occupies the entire breast.

Microscopy:

- It contains cystic spaces with leaf like projections, hence the name (*Phyllodes*—Greek—leaf-like).
- Cells show hypercellularity and pleomorphism.
- It may be a variant of intracanalicular fibroadenoma of breast (*Giant type*).

Clinical Features

- They occur in premenopausal women (30-50 years).
- It is usually unilateral, grows rapidly to attain a large size with *bosselated* surface.
- Swelling is *smooth, nontender, soft, fluctuant with necrosis of skin* over the summit due to pressure.

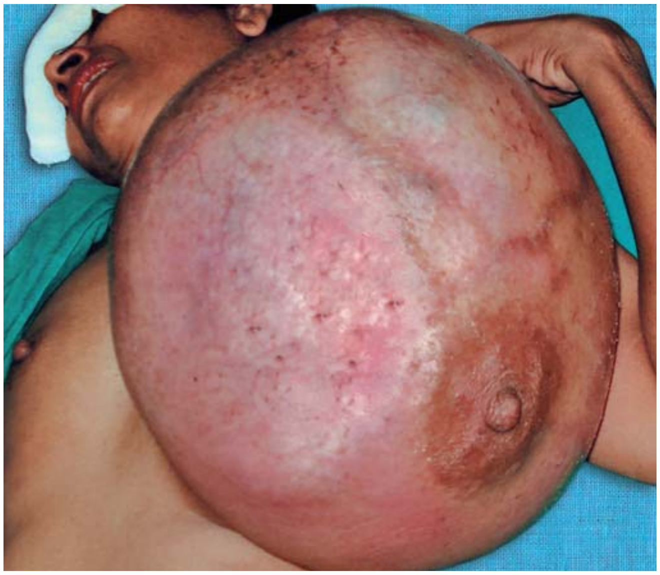

Fig. 8.18: Phyllodes tumour of left breast (Cystosarcoma phyllodes of left breast).

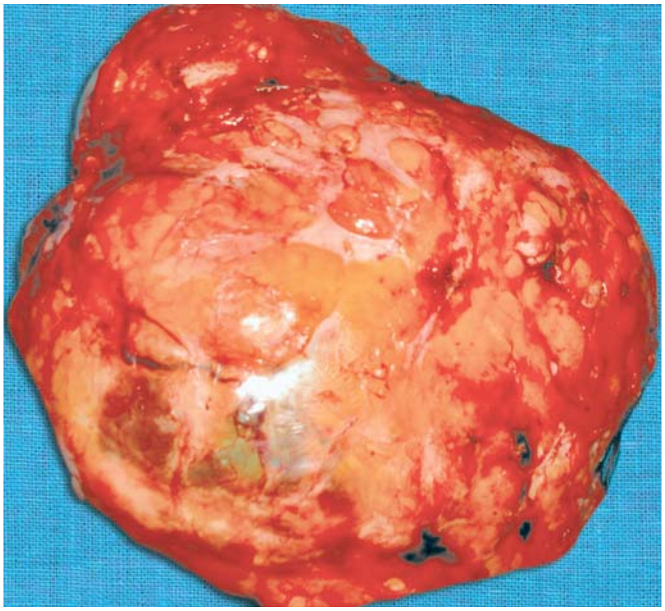

Fig. 8.19: Operated specimen of cystosarcoma phyllodes.

- Skin over the breast is stretched, red and with dilated veins over it. Tumour is warmer, not fixed to skin or deeper muscles or chest wall. Nipple retraction is absent. Lymph nodes are usually not involved. These are the *differentiating features* from carcinoma.
- Tumour grows rapidly; undergoes necrosis at various places; causes cystic areas.
- Recurrence is common.

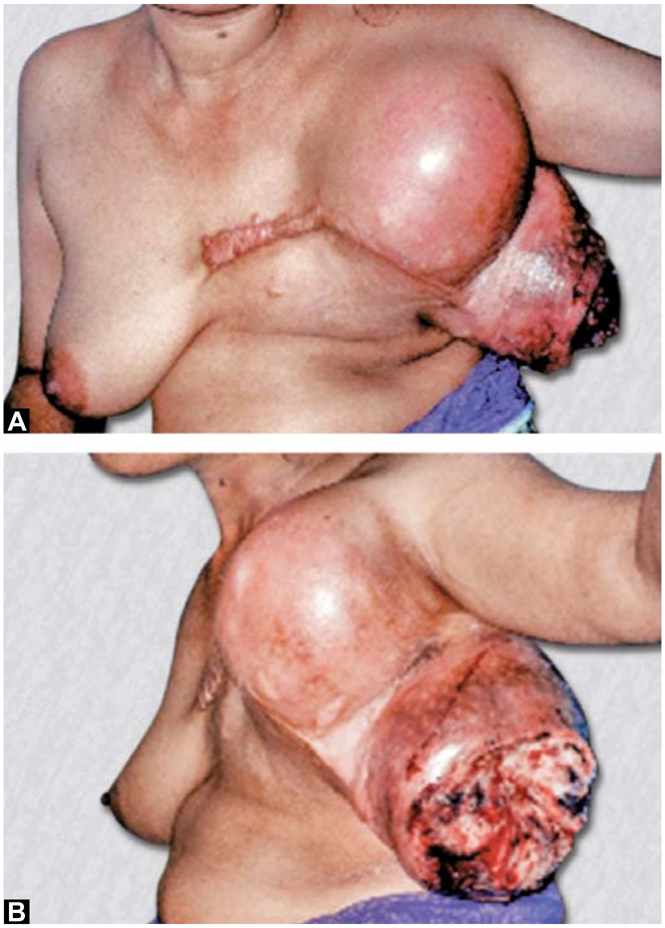

Figs 8.20A and B: Recurrent cystosarcoma phyllodes.

Investigations

- U/S.
- FNAC, core biopsy
- Mammography.
- Chest X-ray CT chest in malignancy to see secondaries.

Treatment

- Excision or subcutaneous mastectomy is done.
- If malignant (sarcoma), total mastectomy is indicated.

MASTALGIA (“Pain in the Breast”)

- 45% of women report breast pain, 21% severe.
- An entity that is ubiquitous; has an unknown aetiology, and a poorly understood.
- Mastitis, carcinoma presenting with only mastalgia (8%).
- Patients who are on HRT, caffeine, tobacco, large pendulous breasts, etc.

Types

- Cyclical—65%.
- Noncyclical—30%.
- Chest wall pain—5%.

Saying ‘no’ to something is actually much more powerful than saying ‘yes’. — **Tom Hanks**

Cyclical

- Pain related to menstrual cycles.
- Usually seen in ANDI like fibrocystadenosis.
- Present in women of menstruating age group.
- Pain is more during menstruation.
- It is bilateral, diffuse with "heavy feeling".

Treatment:

- Evening primrose oil 325 mg BD.
- Gamolenic acid 120 mg BD.
- Danazol (100-200 mg BD)—antigonadotrophin agent.
- Bromocriptine (2.5 mg BD)—prolactin inhibitor.
- Tamoxifen (20 mg daily).
- GnRH analogue 3.6 mg injection depot-monthly.
- Testosterone undecanoate 40 mg BD.
- Vit B_6, B_{12}.
- Analgesics.

Noncyclical

- Other causes of breast pain are periductal mastitis, malignancy, cervical root pain, musculoskeletal pain, previous surgery, Tietze's syndrome.
- It is unilateral, chronic, burning or dragging in nature, occurs both in pre- and postmenopausal age group.
- 5% of breast cancers present as pain during first presentation.

Treatment:

- Cause has to be identified.
- Malignancy has to be ruled out.
- Avoid coffee and stress.
- Proper support to breasts.

Tietze's syndrome:

Costochondritis of second costal cartilage, commonly seen in females, mimics mastalgia.

TRAUMATIC FAT NECROSIS

It may be due to either direct or indirect trauma (trauma may not be noticed many times).

Pathogenesis

Capillary ooze causes triglyceride in the fat to dissociate into fatty acids. It combines with calcium from the blood resulting in *saponification* which causes inflammatory reaction and later presents as a nonprogressive swelling in the breast.

Clinical Features

- Painless swelling in the breast which is *smooth, hard, nontender and adherent* to breast tissue.
- It is nonprogressive, nonregressive.

Investigations

- FNAC shows chalky fluid with fat globules.
- Mammography to rule out malignancy.

Differential Diagnosis

- Carcinoma breast.

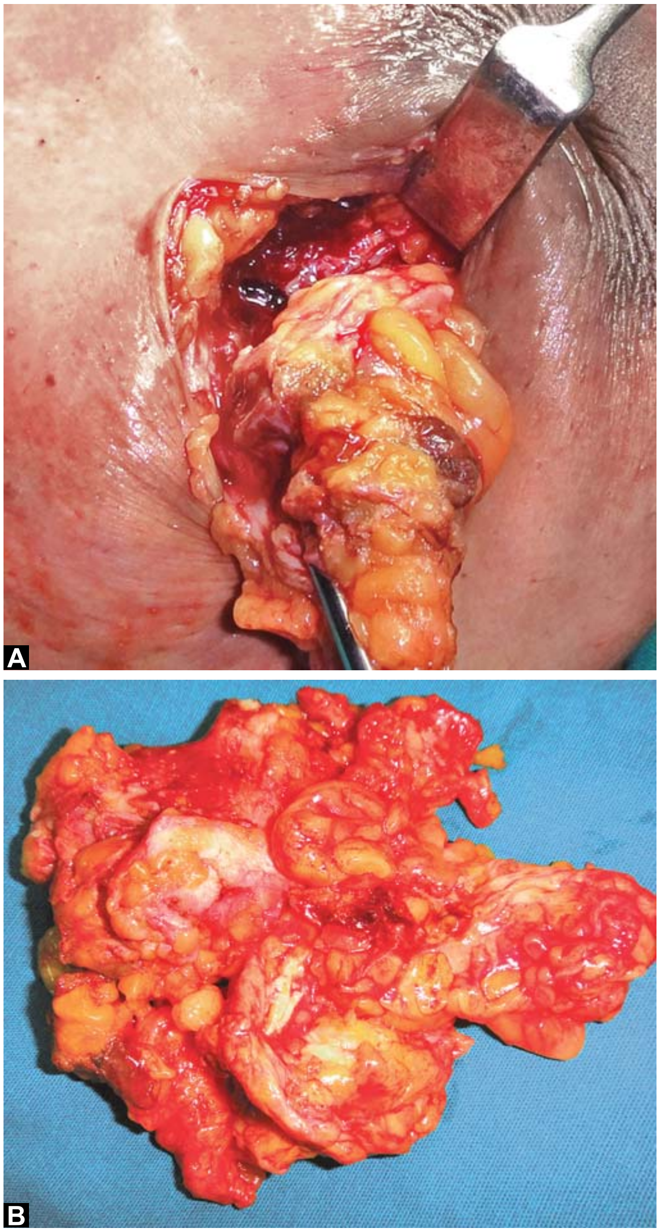

Figs 8.21A and B: Traumatic fat necrosis.

Treatment

- Excision.

GALACTOCELE

- Seen in lactating women.
- Occurs during cessation of lactation. Often up to 10 months after lactation.
- It is due to the blockage of lactiferous duct resulting in enormous dilatation of lactiferous sinus.
- It contains milk and epithelial debris within.
- It is a retention cyst in subareolar region attaining large size.

Clinical Features

- Lump in the lower quadrant of the breast which is usually *unilateral, large, soft, fluctuant, with smooth surface.*
- It is usually nontender.

- It may get *precipitated, inspissated* or get *calcified*. When it is calcified it mimics carcinoma breast. If it gets infected it will form an abscess.
- When it is cystic other cystic swelling in the breast should be ruled out.

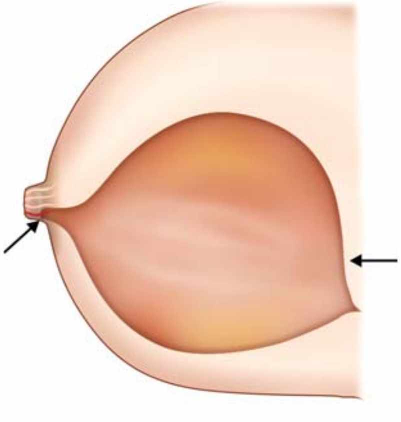

Fig. 8.22: Galactocele.

Investigations

- U/S.
- FNAC. Aspiration shows thick, creamy, greenish/brown fluid.

Treatment

- Aspiration of the content.
- Excision (submammary incision).
- Abscess when formed should be drained under general anaesthesia under cover of antibiotics.

MASTITIS

Types

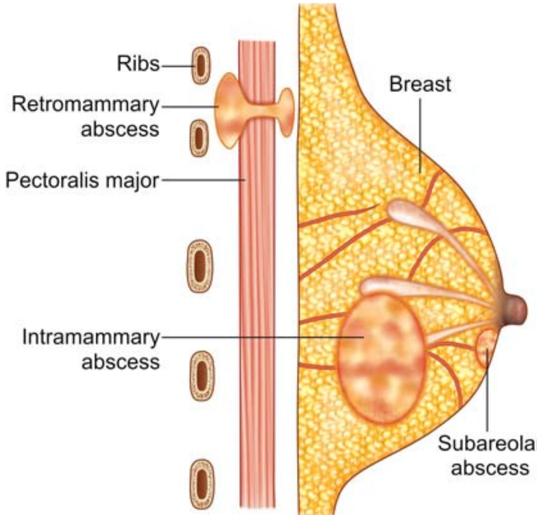

Fig. 8.23: Diagram showing subareolar, intramammary and retromammary abscess.

1. Subareolar.
2. Intramammary.
3. Retromammary (submammary).

Subareolar Mastitis

- It is the infection under the areola due to cracks in the nipple or areola. It results from an infected gland of Montgomery or a furuncle of the areola.
- Often it is associated with duct ectasia—causing formation of abscess, sinus and fistula.
- It is common in nonlactating women.

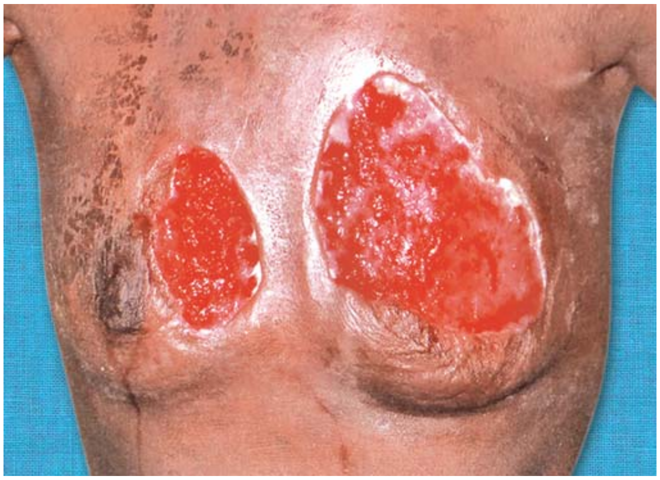

Fig. 8.24: Severe infection caused sloughing of the skin and breast tissue, eventually granulating and is ready for graft.

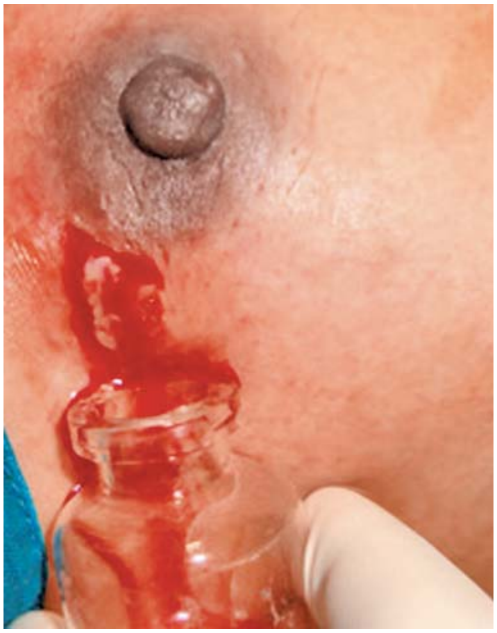

Fig. 8.25: Subareolar abscess.

Clinical features

- Red, inflamed, edematous areola with a tender swelling underneath.
- Nipple retraction may develop.

Differential diagnosis

Paget's disease of the nipple.

When arms are raised fully above the head, visible signs of breast carcinoma frequently becomes more apparent.
—Hugh Auchincloss

Treatment

- Under cover of antibiotics pus is drained by making a subareolar incision.

Intramammary Mastitis (Breast Abscess)

a. Lactational abscess of the breast:

Commonly seen in lactating women. Usually up to 6 months of lactation period. It occurs in 3% of breastfeeding mothers.

Precipitating factors

- Cracked nipple
- Retracted nipple
- Improper cleaning of the nipple
- Inadequate milk sucking by baby or milk expression causing stasis
- Infection from the mouth of the baby
- Haematoma getting infected

Mode of infection:

Bacteria (*Staph. aureus*—most common) enters the breast during sucking through the cracked nipple. Occasionally it can be from haematogenous spread. Gram-negative and other bacterial infections can supervene later.

Staphylococcus aureus causes clotting of milk in the blocked duct and multiply. Duct initially gets blocked by epithelial debris or by retracted nipple. Initially it begins in one quadrant but later involves entire breast.

Clinical features:

- Continuous throbbing pain in the breast and high grade fever.
- Diffuse redness, tenderness, warmness and *brawny induration* in the breast.
- Purulent discharge from the nipple.
- Entire breast may get involved eventually.
- Occasionally tender fluctuant swelling (10%) may be felt; ulceration and discharge can occur at a later period.
- It is difficult to differentiate initial *stage of mastitis* (stage of cellulitis) from *stage of breast abscess* formation. When it is treated by a ntibiotics without incision and drainage eventually it may get organised to form a nontender, hard breast lump with sterile pus inside—*stage of antibioma* formation.

Differential diagnosis:

Inflammatory carcinoma of breast.

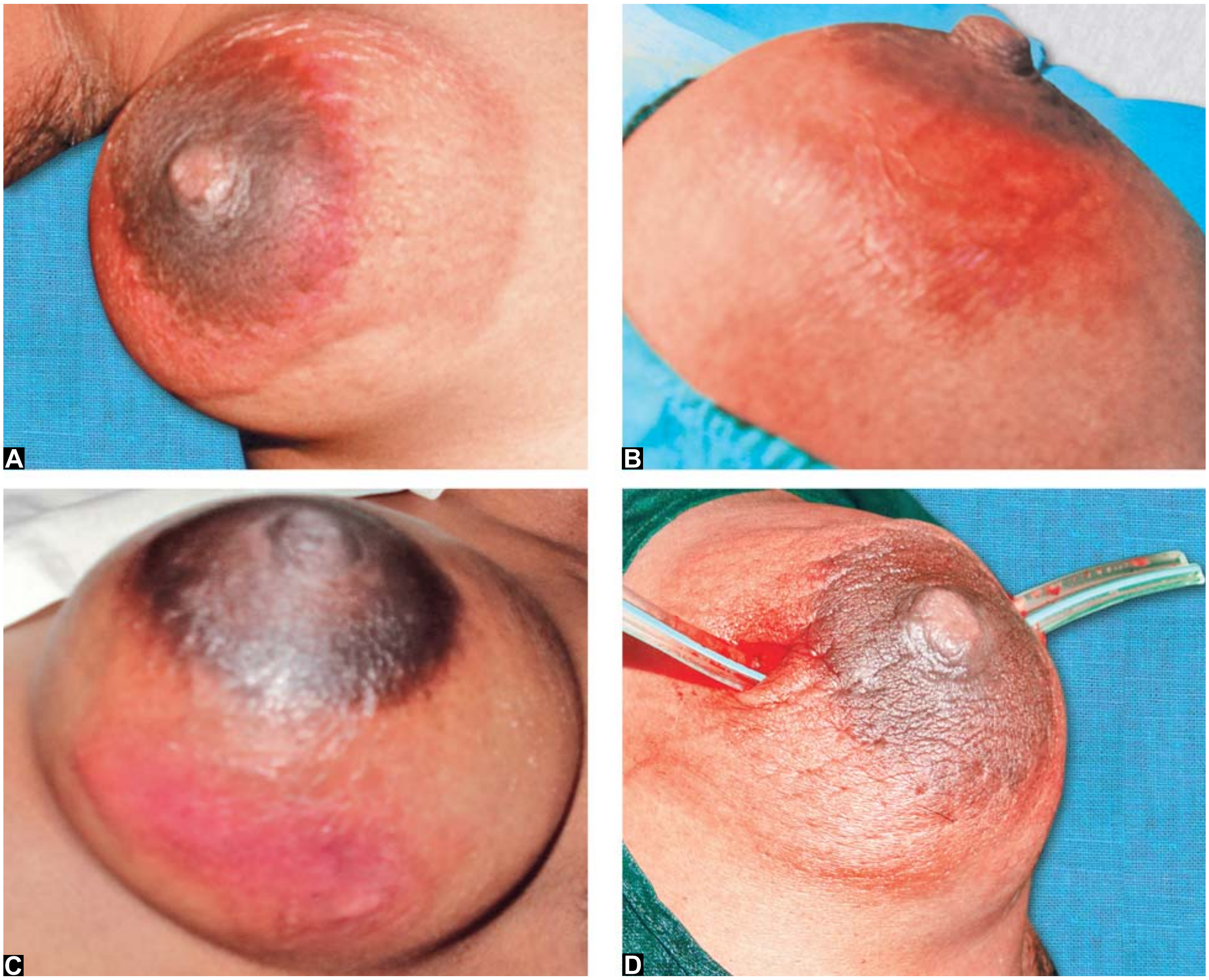

Figs 8.26A to D: Typical breast abscess with features of acute inflammation. Drain should be placed after incision and drainage in such abscesses. Often gauze drain can be used.

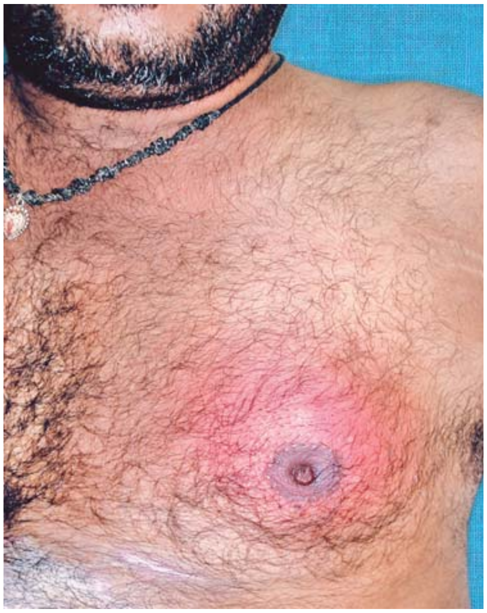

Fig. 8.27: Breast abscess in a male patient. Breast abscess eventhough is uncommon in males, it can occur in puberty and middle age.

Treatment:

- Antibiotics—cephalosporins, flucloxacillin and amoxicillin.
- *Repeated US guided aspirations* can be tried which avoids surgery and scar.
- Drainage under general anaesthesia, *a counter incision* may be needed.

It is not advisable to wait till the formation of abscess.

Often takes very long time to heal after surgery causing distress to patient and surgeon as well.

Indications for drainage in mastitis/breast abscess

- Mastitis not resolving with antibiotics in 48 hours
- Persistent fever and progression of mastitis
- Brawny induration

Do not wait for abscess to form (fluctuation to develop)

Complications:

- Antibioma formation.
- Sinus formation, skin necrosis, fistula formation.
- Recurrent infection, bacteraemia, septicaemia.

Suppression of lactation is often required by giving Bromocryptine 2.5 mg BD for 2 weeks.

***b. Nonlactational abscess of the breast*:**

It commonly occurs in duct ectasia and periareolar infections. Common organisms are *bacteroides, anaerobic streptococci*, enterococci and gram negative organisms. It is *commonly recurrent* with tender swelling under the areola.

Treatment:

- Antibiotics.
- Repeated aspirations.
- Drainage and later cone excision of the duct is done.

Retromammary Mastitis

- It is due to tuberculosis of the intercostal lymph nodes or ribs beneath or suppuration of the intercostal lymph nodes.
- Breast is normal.

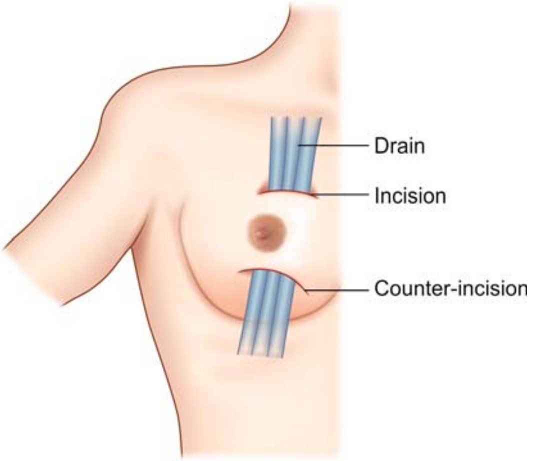

Fig. 8.28: Incision and counter-incision for breast abscess.

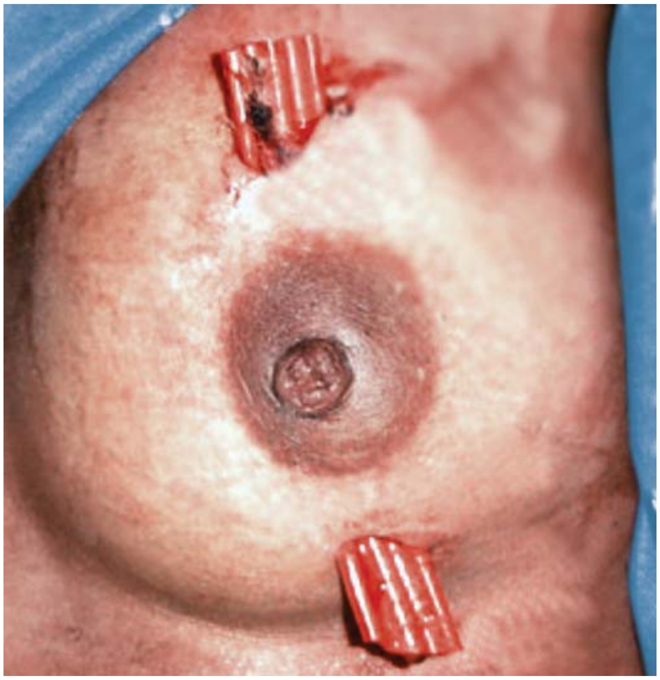

Fig. 8.29: Photograph showing placement of drain in breast abscess.

Causes

- Tuberculosis of intercostal lymph nodes
- Tuberculosis of ribs beneath
- Suppuration of intercostal lymph nodes
- Empyema necessitans
- Infected haematoma

Transillumination in a case of duct papilloma may reveal an opacity in the line of the duct due to pent-up blood.
—William J Moore

Investigations:

- Chest X-ray, FNAC, ESR.
- Peripheral smear.
- U/S of breast and chest wall.
- Often CT scan chest may be needed.

Treatment:

- Cause has to be treated.
- Drainage through submammary/retromammary incision.

***Note*:**

- Mastitis of infancy (*Witch's milk*) is due to maternal hormone in infant blood. It usually subsides but may cause suppuration.
- Mastitis of infancy affects both sexes equally.
- Mastitis of puberty is common in boys, is invariably unilateral with tender, swollen and inflamed breast.
- Mastitis of mumps is usually unilateral and can occur in both sexes.
- Bacterial mastitis is seen in adult women, commonly lactating and is due to staphylococci infection.
- Subareolar mastitis is due to infection of gland of Montgomery or due to areolar furuncle.

ANTIBIOMA

If intramammary mastitis is not drained but only treated by antibiotics, pus localises and becomes sterile (flaques) with a thick fibrous tissue cover and it is called as antibioma.

Clinical Features

- Previous history of mastitis treated with antibiotics.
- Swelling which is painless, smooth, nontender, hard, fixed to breast tissue without involving the pectorals and chest wall.

Differential Diagnosis

Carcinoma breast (Scirrhous carcinoma breast).

Investigations

- FNAC; mammography; U/S breast.

Treatment

Excision (Submammary incision). Later antibiotics are given. It should be sent for histology.

DUCT ECTASIA

- It is dilatation of lactiferous ducts due to muscular relaxation (myoepithelial relaxation) of duct wall with *periductal mastitis*.
- It is also called '*plasma cell mastitis*' as periductal inflammation contains plasma cells.
- Commonly many ducts are involved.
- Hormonally induced myoepithelial relaxation with poor ductal absorption of secretions and desquamated cells causing obstruction are the probable other causes and features.

Clinical Features

- Greenish discharge or *creamy/paste* like from the nipple.
- Indurated mass under the areola which is often *tender*.
- Retraction of nipple which occurs at later stage of the disease. *Slit like retraction of nipple* due to fibrosis occurs.
- Eventually it forms an *abscess* and *fistula*.
- Often they are bilateral and multifocal.
- *More common* in smokers—in relation to arterial pathology.
- Common in multiple pregnancies, *perimenopausal age*, hyperprolactin status.
- May present as mastalgia.
- Axillary nodes may be palpable as nonspecific.
- Secondary bacterial infection (anaerobic) is common.

Differential Diagnosis

Carcinoma breast.

Investigations

- Discharge study, FNAC.
- Mammography.

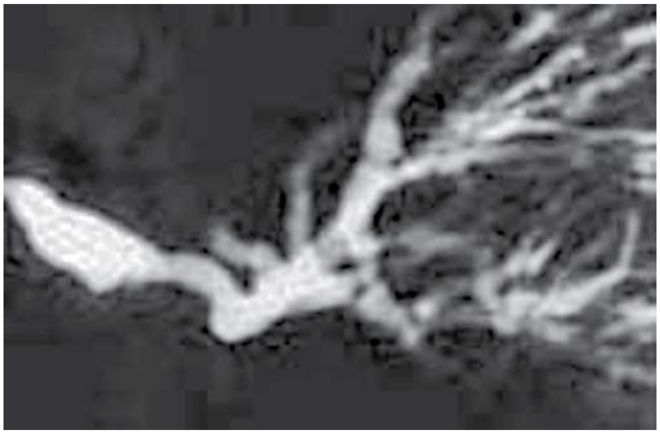

Fig. 8.30: Duct ectasia—ductography.

Treatment

- It is important to stop smoking.
- Cone excision of involved major ducts (*Adair-Hadfield operation*).
- Antibiotics.
- *Melhem Novel modified* breast ductal system excision.

MONDOR'S DISEASE (Henri Mondor—Paris, 1939)

- Mondor's disease is *spontaneous* thrombophlebitis of the superficial veins of the breast and anterior chest wall.
- Cause is not known. History suggestive of injury or infection is not observed.
- Presents as a thrombosed subcutaneous cord (2-3 mm sized) which is attached to the skin.
- On raising the arm above, a narrow, shallow subcutaneous groove appears alongside the cord like thrombosed vein.
- The thoracoepigastric vein, the lateral thoracic vein, and the superior epigastric vein—are involved. The upper, inner portions of the breast are never involved.
- Trauma, infection, surgery may be the cause but clearly not proved and so controversial.

- Rarely penile Mondor's disease and Mondor's in the arm are observed.
- It is often a self-limiting disease without any recurrence, complication or deformity.
- It often mimics the lymphatic permeation of carcinoma breast.
- Restriction of arm movements, brassiere support and anti-inflammatory drugs may be needed. Occasionally refractory cases need surgical excision of involved segment of vein.

TUBERCULOSIS OF THE BREAST

- It is relatively rare. Even though it is rare, often seen in India (4% of benign breast diseases). It may be due to high resistance offered by mammary gland tissue to the survival and multiplication of the tubercle bacillus, a resistance similar to that offered by spleen and skeletal muscle.
- Usually associated with active pulmonary tuberculosis.
- Infection reaches through blood or retrograde lymphatic spread from lymph nodes of axilla.
- Common in lactation.
- Nipple and areola are *not commonly* involved.
- *Lump*—irregular ill-defined; peaud' orange; discharge, sinus, matted axillary nodes often with sinus are the features.
- Presents as a swelling in the breast with cold abscess, sinuses and a typical bluish appearance of surrounding skin with matted lymph nodes in the ipsilateral axilla.

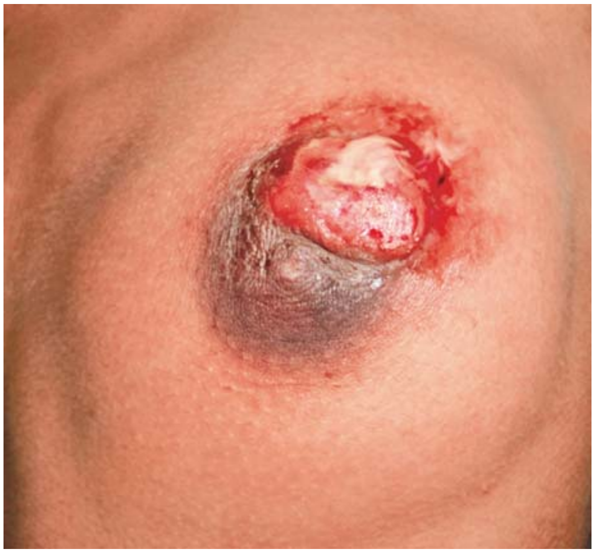

Fig. 8.31: Tuberculosis of breast showing undermined lesion.

Differential Diagnosis

Carcinoma breast.

Investigations

- FNAC.
- Frozen section biopsy is useful to differentiate from carcinoma.
- Excision biopsy.

Treatment

- Antituberculous drugs—INH, rifampicin, ethambutol, pyrazinamide.
- Drainage of cold abscess.

Note:

- Mastectomy is not done.

BREAST CYSTS

- They are cavities lined by epithelium in the breast containing fluid. It arises from destruction and dilatation of breast lobule and terminal ductules. It is due to nonintegrated stromal and epithelial involution.
- Cyst may be microscopic or macroscopic. It contains straw coloured or green or opaque fluid.
- Incidence is very high (1 in 14 females). It is common after the age of 35 years up to menopause. It is uncommon after menopause. Hormone replacement can cause cyst formation in old women.
- Cyst size varies with menstruation due to influence of ovarian hormones.
- Cysts can be multiple (50%). Often bilateral.
- Cysts can be recurrent (50%).
- Risk of breast cancer in breast cyst is very less (0.1%). But incidental associated carcinoma may be present in 3% of breast cysts.
- *Clinically*—smooth, soft, fluctuant often transilluminating well-localised swelling may be felt.
- *Differential diagnoses* are—*Bloodgood cyst*, haematoma, cystic necrosis in a carcinoma, Brodie's disease, galactocele, lymph cyst, hydatid cyst.
- *Investigations*—US of breast; FNAC. Mammography to rule out associated carcinoma.

Treatment

- Aspiration for *two* times.
- Surgical excision is done if cyst *recurs* after two aspirations or if there is *bloody* discharge or *residual mass* if felt after aspiration.

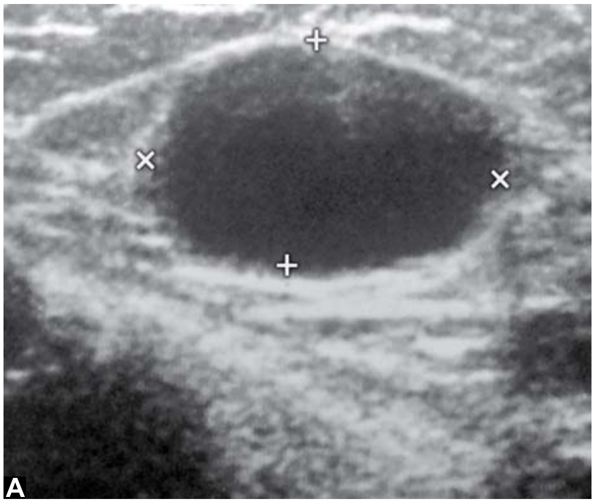

Fig. 8.32A

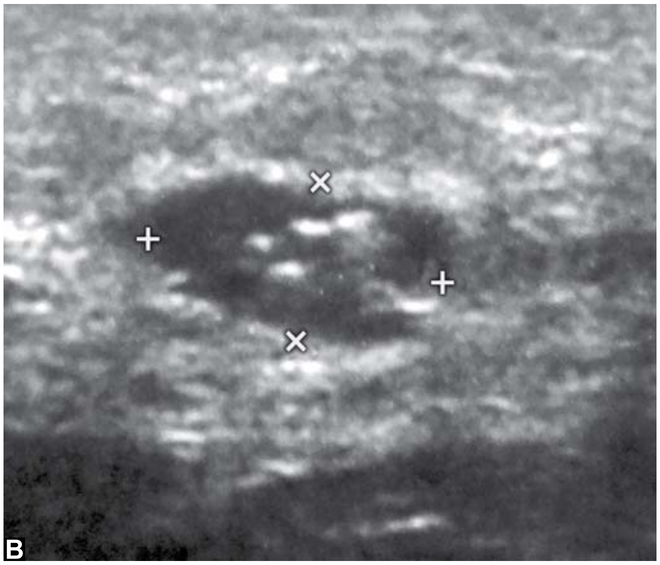

Fig. 8.32B

Figs 8.32A and B: US look of breast cysts—simple and complex type.

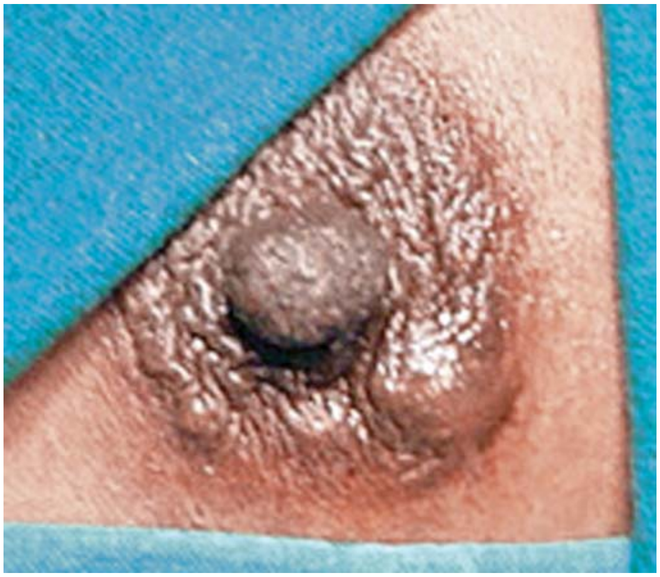

Fig. 8.33: Retention cyst in areola.

Cystic swellings of the breast

- Bloodgood cyst
- Breast abscess
- Hydatid cyst
- Galactocele
- Serocystic disease of Brodie
- Cystic necrosis in carcinoma breast
- Lymph cyst
- Haematoma in breast

GALACTORRHOEA

- It is secretion of milk not related to pregnancy or lactation.
- It is always bilateral.

Primary galactorrhoea is due to:

- Stress and other factors. It is physiological in puberty or menopause. Reassurance is the treatment.

Secondary galactorrhoea is due to:

- Dopamine receptor blocking agents like haloperidol, methyldopa, chlorpromazine, metoclopramide or by hyperprolactinaemia due to pituitary tumours. It enhances the prolactin activity.
- Hypothyroidism.
- Drugs like oral contraceptives, atenolol, clonidine, ranitidine.
- Ectopic prolactin secreting tumours usually from lungs (bronchogenic carcinoma).
- Chronic renal failure.

Management

- Estimation of serum prolactin, T3, T4, TSH, CT/MRI head.
- Bromocriptine therapy.
- Treatment of cause. Causative drug should be stopped and different drugs should be used for the needed condition.

Note:

Witch's milk is secretion of milk in both male and female infants due to maternal hormonal effects in foetus. It lasts up to three weeks after child birth.

GYNAECOMASTIA (Greek—Women Breast)

- It is hypertrophy of male breast more than usual due to increase in ductal (epithelial) and connective tissue (stromal) elements often attaining features of female breast.
- It can be unilateral or bilateral. Bilateral can be symmetrical or asymmetrical.

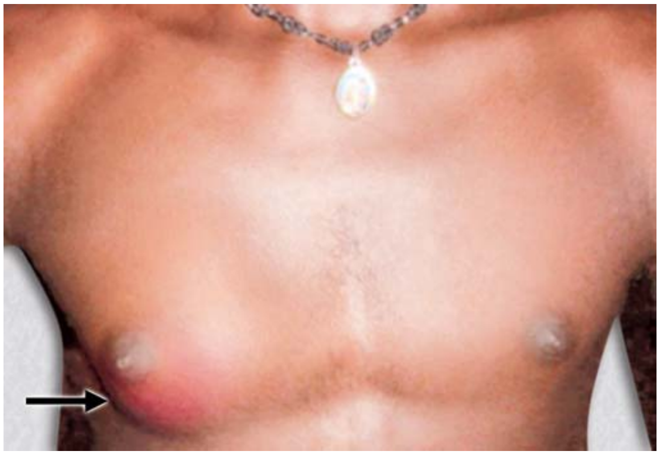

Fig. 8.34: Gynaecomastia—right side breast (diffuse type). Compare with opposite side to note the difference in size.

- It is nonphysiological *stimulatory* oestrogen excess or *inhibitory* androgen deficiency. There is nonphysiological hypertrophy of male breast attaining the size of female breast with increase in connective and ductal tissues.
- It can be *diffuse*—involving all quadrants or small *well localised* hard subareolar nodule.

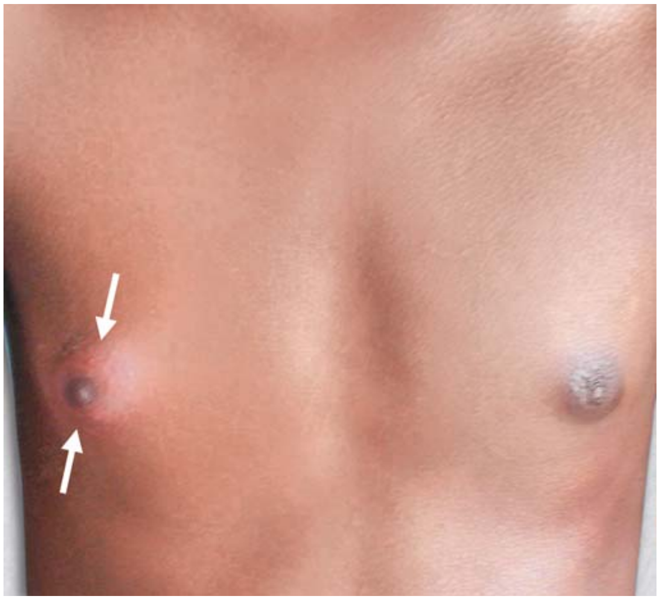

Fig. 8.35: Right-sided gynaecomastia (well-localised—puffy nipple).

Etiology

- *Oestrogen excess*—increased estradiol secretion due to testicular tumour (*Leydig* cell) or nontesticular tumours from adrenal cortex, lung, liver; hyperthyroidism (increases conversion of androgen to estrogen), liver diseases, oestrogen therapy for prostate cancer.
- *Androgen deficiency*—aging, *Klinefelter's/Kallmann syndromes*, eunuch, ACTH deficiency.
- *Secondary testicular failure*—cryptorchidism, orchitis, trauma, CRF, etc.
- *Drugs* (25%):
  - Increases estrogen activity (digitalis, anabolic steroid)/ estrogen synthesis (reserpine, theophylline, frusemide).
  - Inhibition of testosterone synthesis—cimetidine, phenytoin, spironolactone.

Initially there will be *florid proliferative stage* of ductal epithelium with oedematous stroma without acini (1 year); later there will be *quiescent stage* with ductal dilatation and stromal fibrosis.

- Always one should examine genitalia and liver.
- It is physically *embarrassing;* psychologically *devastating*.

Types

- *Neonatal* gynaecomastia is due to action of placental estrogen
- *Pubertal* (25%) in young boys is due to excess estrogen level in relation to testosterone. It is usually unilateral. Here breast tissue will be more than 2 cm in diameter in nonobese young male
- *Senescent* in old is due to fall circulating testosterone causing elative hyperestrinism
- *Prepubertal*—may also be seen in girl child

It can be:

- Puffy nipples
- Pure glandular
- Adolescent hereditary
- Adult—most common
- Androgenic

Simon's classification of gynaecomastia:
Group 1 – Minor visible type without skin redundancy
Group 2 A– Moderate without skin redundancy
Group 2 B – Moderate with minor skin redundancy
Group 3 – Gross pendulous breast

Differential Diagnosis

- Pseudogynaecomastia—adipose tissue deposition as a part of obesity; needs liposuction.
- Pectoral muscle hypertrophy.
- Lipoma, dermoid, haematoma.

Presentations

- Diffuse enlargement of breast occupying all quadrants.
- Often well-localised, small, firm or hard nodule under the areola which is often painful and tender.

Causes

- Idiopathic—25%
- Teratoma testis—3%
- Ectopic hormonal production in bronchial carcinoma
- Anorchism, after castration
- Adrenal and pituitary disease
- Leprosy, because of bilateral testicular atrophy
- Drugs (25%): Stilbestrol, digitalis, cimetidine, spironolactone, INH, phenothiazides
- Liver diseases and liver failure—10%
- Klinefelter's syndrome (XXY Trisomy), Kallman syndrome
- Primary or secondary hypogonadism
- Hyperthyroidism
- Renal diseases, dialysis induced (1%)

Investigations

Relevant to the cause, e.g. liver function tests, DNA study, hormone assay, FNAC, USG breast.

Treatment

- Well-localised type which is symptomatic or patient is psychologically worried is treated by surgical excision using circumareolar incision.
- Large diffuse gynaecomastia needs *Galliard Thomas* submammary incision for excision.
- *Reduction mammoplasty;* nipple reduction surgeries.
- *Drugs*—tamoxifen, clomiphene, androgens (dihydrotestosterone), danazol, testolactone as aromatase inhibitor—but less useful. Drugs are given only for 6 months. Gynaecomastia more than 2 years duration is less likely to show benefit from drugs.

A man is not the best because he works hard, he works hard because he is the best.

- Breast irradiation to prevent gynaecomastia in prostatic cancer patients who are on oestrogen therapy.
- Causative drugs should be stopped.
- Testicular tumours, hyperthyroidism should be treated if they are the causes.
- Testosterone can be given if hypogonadism is the cause.
- Endoscopic assisted subcutaneous mastectomy is done and under trial mainly for cosmetic purpose.

***Note*:**

- Rarely gynaecomastia may turn into malignancy. Gynaecomastia is a differential diagnosis for male breast cancer.
- 80% of Klinefelter's syndrome shows gynaecomastia; it has got very high-risk (20 times) of developing carcinoma.

DUCT PAPILLOMA

- It is the most common cause of bloody discharge from nipple
- It is usually single, from a single lactiferous duct
- It blocks the duct causing ductal dilatation

- They are epithelium lined true polyps of breast lactiferous ducts.
- Usually it is < 1 cm in size often with a small lump under areola. But can attain large size.
- Vascular stalk is present usually.
- Rarely a cystic soft swelling may be present underneath which is probably due to obstruction of the duct by papilloma.
- Papilloma may often project out like a pedunculated mass.

Types

- Subareolar—common.
- Peripheral—occasional.
- Solitary—common.
- Multiple.
- Unilateral—common.
- Bilateral.

Clinical Features

- Papilliferous swelling (projection), usually seen near the nipple orifice.
- Blood stained discharge from the nipple is common.
- But serous or serosanguinous discharge can also occur.
- Single papilloma is not premalignant.
- But multiple papillomas in many ducts can be premalignant.
- Peripheral papilloma should be differentiated from invasive papillary carcinoma.

Investigations

- Discharge study (FNAC).
- Injection of contrast into the duct (Ductogram).
- Mammography may show dense lesion under the areola.

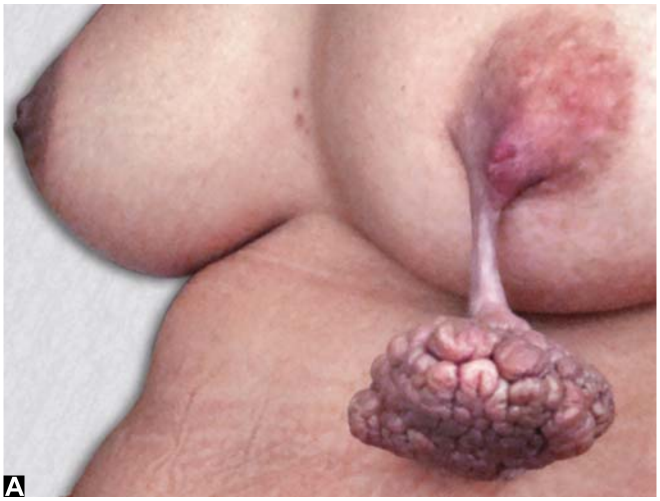

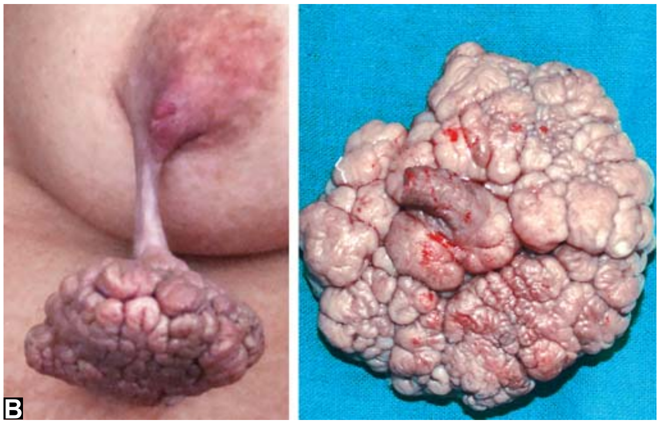

Figs 8.36A and B: Papilloma breast which is large. Usually it is small intraductal.

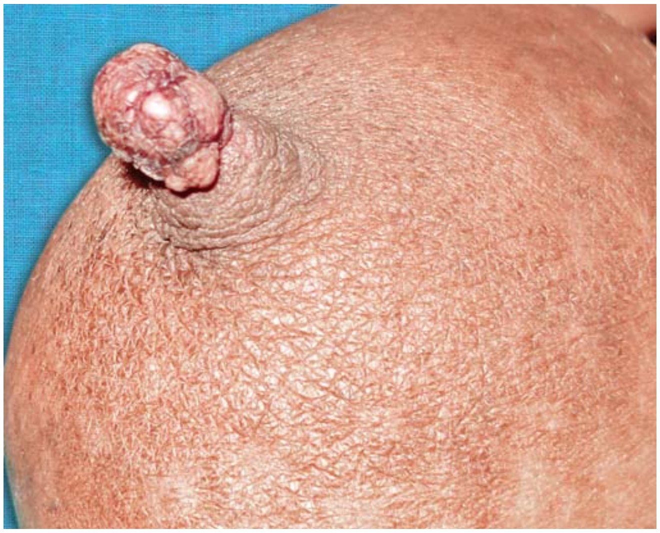

Fig. 8.37: Breast papilloma. It usually arises from single duct. It projects near the nipple orifice.

Treatment

Microdochectomy: Probed lactiferous duct is opened, and the papilloma is excised using *tennis racquet* incision.

Probing of lactiferous duct is done up to 5 cm deep and excised with part of the areolar skin.

Note:

- Intraductal papillary carcinoma arises usually not from a preexisting duct papilloma but as ***de novo***.
- *Papillomatosis* is hyperplasia of the duct lining, filling the duct like a polyp without a pedicle/stalk (not a true polyp); it is not duct papilloma but part of fibrocystadenosis.
- Axillary nodes are not involved.

ZUSKA-ATKINS DISEASE

- A condition considered mainly by pathologists. Squamous metaplasia causes plugging and obstruction of the ducts; subsequent inflammatory reaction and infection.
- Nipple retraction, recurrent episodes of infection and presence of painful nodules under the areola in a nonlactating woman is seen.
- The presence of a sinus draining milky discharge in the areola is characteristic.
- Seen in smokers—men or women. Vitamin A deficiency is also the cause.
- It can mimic carcinoma.
- Treatment is antibiotics and surgical excision like for duct ectasia.

MAMMARY FISTULA OF ATKINS

- It is a fistula of lactiferous duct which opens on to the areola.
- It causes recurrent abscess that points on to the areola which discharges and continues to discharge for a week.
- It is a differential diagnosis for carcinoma breast.

Treatment

- Antibiotics.
- Excision of fistula along with lactiferous duct.

CARCINOMA BREAST

> *When cancer occurs in the breast of women under forty, it is more rapid in its progress than when the patient is older, and also more extensive; remote sympathy likewise takes place more readily in them than in the old, so that the operation succeeds better in the latter on this account.*
>
> —**John Hunter**, 1728-1793

Aetiology

- Carcinoma breast is more common in developed, western countries.
- In African-American women, it is more aggressive. It is less common in Japan, Taiwan.
- It is second most common carcinoma in females. Incidence is 19-34%. Median age is 47 years.
- Carcinoma in one breast increases the risk of developing carcinoma on opposite breast by 3-4 times. Incidence of bilateral carcinoma is 2%.
- It is more common after middle age, but can occur at any age group, after 20 years.
- It can be familial in 2-5% cases.
- Mutation of tumour suppressor genes BRCA1/BRCA2 is thought to be involved with high-risk of breast carcinoma. BRCA1 mutation is having more risk (35-45%) than BRCA2 mutation. It is located in long arm of chromosome 17, whereas BRCA2 is located in long arm of chromosome 13. *BRCA1* more commonly shows *ER negative* status, high grade, aneuploid with raised S fraction than *BRCA2* which shows *ER positive* status. BRCA1 is associated with increased risk in males. Lifelong risk of breast cancer in BRCA1 and BRCA2 mutations is 50-70%. Both are associated with high-risk for ovarian cancer.
- Occasionally mutation of BRCA3 and p53 suppressor gene is also involved.
- *Li-Fraumen's syndrome* (LFS) is autosomal dominant condition with breast cancer inheritance (90%) along with sarcoma, leukemia, brain tumours, adrenocortical tumours.
- Diet low with phytoestrogens and high alcohol intake have high-risk of breast cancer. Vitamin C reduces the risk.
- It is more common in nulliparous woman.
- Attaining early menarche and late menopause have high-risk of breast malignancy.
- Early child-bearing and breastfeeding reduces the chances of malignancy. Early 1st child birth reduces the risk; late first child birth after 35 years increases the risk.
- It is more common in obese individuals.
- Breast cancer relative risk is qualified as relative risk (RR). If RR is 2.0 means, risk is twice the normal population. If RR is 0.5 means, risk is 50% less than normal population.
- Risk is 3-5 times more if 1st degree relative is having breast cancer. Risk is more if 1st degree relative is younger or premenopausal or having bilateral breast cancers.
- In males, occasionally gynaecomastia turns into carcinoma—not proved.
- Benign breast diseases with atypia, hyperplasia and epitheliosis has got higher risk in a patient with family history. RR in nonproliferative fibrocystic disease is 1.0; proliferative without atypia is 1.5; proliferative with atypia is 4.0 (with family history 6.5, premenopausal 6.0).
- *Cowden's syndrome*—it is an autosomal dominant condition, with cutaneous facial lesion (100%), bilateral breast lesion (50%), GI polyps, brain, thyroid tumours.
- It is often associated with ataxia telangiectasia.
- Previous *therapeutic* radiation (*thoracic*) may predispose carcinoma breast especially when RT is given at younger age mainly for Hodgkin's lymphoma.
- Radial scar may predispose the carcinoma. It is a *complex sclerotic* condition of breast with microcyst, epithelial hyperplasia, adenosis, central sclerosis with lesions less than 1 cm in size. It mimics carcinoma clinically and mammographically.
- It is more common in individuals who are on oral contraceptive pills and hormone replacement therapy (HRT) for more than 5 years.

- Presently carcinoma breast is considered as a systemic disease. ***Halsted concept*** of spread is *sequential spread*. Breast—axillary lymph node—systemic spread. ***Fischer concept*** is early to begin with itself, there is distant blood spread because of *micrometastasis* without nodal disease. Only tumour lesser than 1 cm size can be sequential. ***Spectrum concept*** is new one where disease spreads loco-regionally as well as systemically which makes it to aim at both locoregional disease control as well as systemic disease control.
- Prior diagnosis of uterine/ovarian/colonic cancers.

Incidences in carcinoma breast

- 30% of all female cancers
- 20% of cancer related deaths in females
- 2-4% bilateral
- 2-5% hereditary
- Lump in the breast—most common presentation (75%)
- 10% presents with pain
- 35-45% with mutation of BRCA1 gene
- 70% blood spread occurs to bones

Modified Gail risk assessment model

- Age, age at menarche, age at 1st live child birth, race
- Number of 1st degree female relatives having breast cancer
- Number of previous breast biopsies
- Proliferative lesion with atypia

These risk factors are translated into risk scores by specified calculations

But model does not include genetic factor

Pathology

- Breast carcinoma arising from lactiferous ducts is called as ***ductal carcinoma.***
- Breast carcinoma arising from lobules is called as ***lobular carcinoma.*** It is 10% common.

Risk factors for breast carcinoma

- Breast carcinoma in 1st degree relative
- Breast carcinoma in contralateral breast
- BRCA1/BRCA2 gene mutation
- Obesity and alcohol intake
- Nulliparity
- Early menarche and late menopause

- ***In situ carcinoma*** is preinvasive carcinoma which has not breached the epithelial basement membrane. It may be:
  - *Ductal in situ carcinoma (Ductal Carcinoma In Situ, DCIS)* or
  - *Lobular in situ carcinoma (Lobular Carcinoma In Situ, LCIS*).
- ***Invasive carcinoma*** can occur eventually.

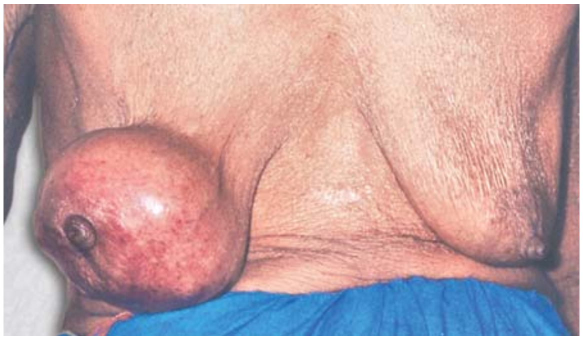

Fig. 8.38: Carcinoma right breast in an elderly—an atrophic scirrhous carcinoma.

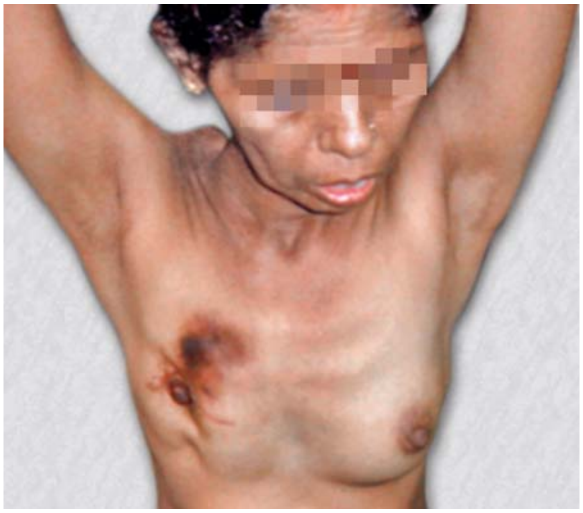

Fig. 8.39: Examination with both arms raised above the shoulder.

Classifications

I. Ductal carcinoma.
 Lobular carcinoma.
II. (a) *In situ* carcinoma (Noninvasive)
 – DCIS (Ductal carcinoma *in situ*).
 – LCIS (Lobular carcinoma *in situ*).
 (b) Invasive.
 – Invasive ductal carcinoma—most common type. Adenocarcinoma with *no special type* (80%) is

Risk factors classification

Slight to moderate risk	*Moderate to high-risk*	*Very high-risk*
Florid hyperplasia	Age > 60 years	Therapeutic radiation
Solid duct papilloma	ATD / ALS/ LCIS	Family history of breast cancer in two 1st degree relatives
Obesity, alcohol, HRT	History of DCIS	Family history of breast and ovarian cancer
Nulliparity	Cancer on one side breast	BRCA1 and BRCA2 mutation carrier or 1st degree relative with mutation
Early menarche, late menopause		

more common. 60% of this will show micro or macroscopic spread to axillary nodes. Invasive type can be special type or no special type (*NST*) [not otherwise specified/*NOS*].
- Invasive lobular carcinoma. It is commonly multifocal and often bilateral.

III. Unilateral.
Bilateral—2-5% common.

IV. Unifocal.
Multifocal—tumour tissues within the same quadrant at multiple foci.
Multicentric—tumour tissues within the same breast but in different quadrant.

Foote Stewart original classification of invasive breast cancer

- Paget's disease of nipple
- Invasive ductal carcinoma (adenocarcinoma with productive fibrosis [scirrhous/simplex/no special type]—80%; medullary—4%; colloid—2%; papillary—2%; tubular and invasive cribriform—2%)
- Invasive lobular carcinoma—10%
- Rare other types—adenoid cystic, squamous cell, apocrine type

Classification of primary breast cancer

Noninvasive epithelial
- LCIS
- DCIS (intraductal)—papillary, solid, cribriform, comedo

Invasive epithelial
- Invasive lobular—10%
- Invasive ductal
  - Invasive ductal with NST (no special type)/NOS (not otherwise specified)—70%
  - Tubular—2%
  - Colloid—2%
  - Medullary—5%
  - Medullary variant—basal like
  - Invasive cribriform—2%
  - Invasive papillary—1%
  - Adenoid cystic—1%
  - Metaplastic—1%

Mixed connective tissue and epithelial
- Phyllodes, angiosarcoma, carcinosarcoma

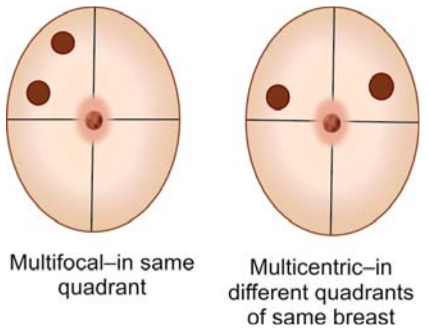
Multifocal–in same quadrant
Multicentric–in different quadrants of same breast

Fig. 8.40: *Multifocal* is occurrence of 2nd *in situ* breast cancer whithin the same breast quadrant as the primary *in situ* cancer: *multicentricity* is occurrence of 2nd *in situ* breast cancer outside (other) the breast quadrant as the primary *in situ* cancer in the same breast.

DCIS (Ductal Carcinoma *In Situ*)

- It is intraductal carcinoma (proliferation of malignant mammary ducal epithelial cells) without any invasion into the basement membrane.
- It is 5-20% common.
- It can be:
  - Solid—*high grade.*
  - Comedo with necrosis is *high grade* with increased chances of microinvasion.
  - Cribriform—*low grade.*
  - Papillary—*low grade.*
  - Micropapillary.
- It is associated with high expression of C – erb2 gene (80%).
- It can be high grade DCIS or low grade DCIS.
- It can be comedo DCIS (more malignant and more likely to be invasive later) or noncomedo DCIS (less malignant).
- In 20% of cases synchronous invasive carcinoma in duct is seen.
- Untreated DCIS becomes invasive in > 50% cases (5 fold).
- 5% of *male* breast cancers are DCIS.
- Minor ductal epithelial proliferation is the typical histology.
- Invasive ductal cancer forms in the same breast and same quadrant of DCIS unlike LCIS. DCIS is an *anatomical precursor* of invasive DC.
- Presence of > 25% of DCIS component is present in the *main invasive tumour* or if DCIS is present elsewhere in the surrounding breast tissue, it is called as *extensive in situ component*.

Modern concept of breast cancer classification based on molecular markers

Type	*Marker*	*Treatment*
• ER positive	• ER positive; HER 2 negative	• Endocrine treatment
• HER 2 positive	• ER positive or negative; HER 2 positive	• Endocrine treatment + Standard – surgery/RT/CT?? trastuzumab
• Basal like	• ER negative; HER 2 negative	• Standard – surgery / RT / CT??

Accountable, dependable, reliable are qualities that make a person responsible.

- Nipple discharge and often small swelling are main presentations.
- U/S assisted FNAC and mammography are the needed investigations.
- Risk of lymph node spread in DCIS is less than 4%. So axillary dissection is not necessary.
- *Sentinel Lymph Node Biopsy* and proceed is the preferred method.

Van Nuy's prognostic index for DCIS

Score	***1***	***2***	***3***
Size in mm	< 15 mm	15-40 mm	> 40 mm
Clearance in mm	> 10 mm	1- 10 mm	< 1 mm
Grade and necrosis	not high grade	not high grade	high grade
	No necrosis	Necrosis present	Necrosis present
Total score is 9			
Score 3-4 conservative breast surgery (wide local excision)			
Score 5-7 conservative surgery + Radiotherapy			
Score 8-9 total mastectomy			

Nottingham Prognostic Index (NPI): (0.2 × Tumour size in cm) + Lymph node stage + Tumour grade

NPI score—< 3.4	Good prognosis with 80% survival (15 years)
NPI score—3.4 –5.4	Moderate prognosis with 40% survival
NPI score—> 5.4	Poor prognosis with 15% survival

Management of DCIS

- FNAC confirms the disease but will not differentiate from DCIS and invasive carcinoma
- Mammography, U/S breast, MRI breast
- Routine metastatic work up
- *Breast conservative surgery with RT* to breast and axillary dissection after SLNB (if +ve)
- *Hormone therapy* (tamoxifen) prevents both local recurrence and development of new primary breast carcinoma
- *Total mastectomy* is done in DCIS when—positive margin after wide local excision; two or more primary tumour; when radiation to breast is not possible (collagen diseases); tumour/breast size ratio is not appropriate for conservative surgery and DCIS in pregnancy. *Skin sparing mastectomy* (SSM) may be ideal in such occasions with immediate TRAM/LD flap

Types of Carcinoma Breast

1. ***Scirrhous carcinoma:*** It is 60% common. It is hard, whitish, or whitish yellow, noncapsulated, irregular, with *cartilaginous* consistency. It contains malignant cells with fibrous stroma.
2. ***Medullary carcinoma (5%):*** Also called as 'encephaloid type' because of its brain like consistency. It contains malignant cells with dispersed lymphocytes.
   - *Medullary variant with some features of pure form* shows uniformly high grade aggressive tumour cells with negative ER, PR, HER$_2$ NEU cell surface receptors (triple negative). They express molecular markers of basal/myoepithelial cells and so now termed as *basal-like breast cancers.*

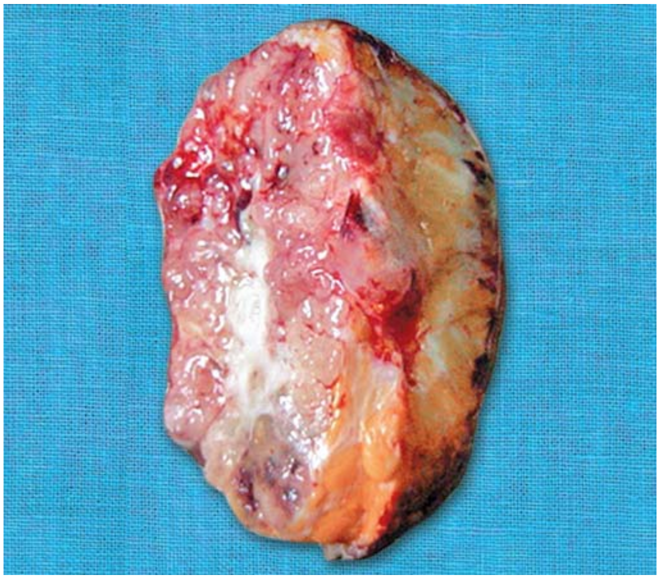

Fig. 8.41: Specimen of carcinoma breast. Cut section showing tumour tissue of scirrhous type (most common type).

3. ***Inflammatory carcinoma/Lactating carcinoma/Mastitis carcinomatosis:***
   - Most aggressive type of carcinoma breast.
   - It is 2% common.
   - It is common in lactating women or pregnancy.
   - It mimics acute mastitis because of its short duration, pain, warmth and tenderness.
   - Clinically, it is a rapidly progressive tumour of short duration, diffuse, painful, warm often involving whole of breast tissue with *occurrence of peau d'orange*, often extending to the skin of chest wall also.
   - More than 1/3rd of skin over the breast is involved; diffuse lymphoedema is due to tumour emboli within *dermal lymphatics*. Underlying localised palpable mass is not evident clinically. It should be differentiated from other LACB with skin involvement where underlying palpable mass is well evident.
   - Mammography *may not* show any finding except skin thickening. Inflammatory carcinoma of breast is a *clinical diagnosis*.
   - *Ductal or lobular* type with cancer cells in dermal lymphatics is the histology.
   - It rapidly metastasises to chest wall, bone and lungs.
   - *It is always stage IIIB carcinoma (T_{4d}).*

     FNAC confirms the diagnosis—it contains undifferentiated cells. Punch biopsy is ideal and better which shows undifferentiated cells.
   - Total count is normal.

   Treatment
   - External radiotherapy and chemotherapy.
   - Salvage surgery whenever possible.
   - It has got *worst prognosis*.

   Differential diagnosis

   Acute mastitis—total count is increased here.

4. ***Colloid carcinoma:*** It produces abundant mucin, both intra- and extracellularly carrying better prognosis.
5. ***Paget's disease of the nipple***

> *I believe it has not yet been published that certain chronic affections of the skin of the nipple and areola are very often succeeded by the formation of scirrhous cancer in the mammary gland.... In the majority (the affection) had the appearance of a florid, intensely red, raw surface...like the surface of very acute diffuse eczema.*
>
> —**James Paget,** 1874

- It is superficial manifestation of an intraductal carcinoma. The malignancy spreads within the duct up to the skin of the nipple and down into the substance of the breast. It mimics eczema of nipple and areola.

Differences between Paget's disease and eczema of nipple

Paget's disease		***Eczema***	
1.	Unilateral	1.	Bilateral
2.	Edges are distinct	2.	Edges are indistinct
3.	Itching absent	3.	Itching present
4.	Seen in menopausal women	4.	Occurs during the time of lactation
5.	Vesicles absent	5.	Vesicles present
6.	Nipple is usually destroyed	6.	Nipple is usually intact
7.	Underlying lump is usually present	7.	No underlying lump

- In Paget's disease, there is a hard nodule just underneath the areola, which later ulcerates and causes destruction of nipple. Histologically, it contains large, ovoid, clear *Paget's cells* with malignant features. *Paget's hyperchromatic cells* are located in rete pegs of epidermis containing intracellular *mucopolysaccharides* as clear *halo* in cytosol.
- It is 2% common. 90% is invasive ductal carcinoma. 70% shows mass underneath nipple and areola.
- Breast conservation surgery (BCS) is difficult here; hence MRM is needed.

6. ***Tubular, papillary, cribriform*** are other types of duct carcinomas.
7. ***Atrophic scirrhous carcinoma:***
   - Seen in elderly females. It is a slow growing tumour which has got better prognosis.
   - FNAC is diagnostic.
   - Mastectomy or curative brachytherapy (using breast moulds) is the treatment of choice.
   - It is curable.
8. ***Lobular carcinoma in situ:***
   - It originates in terminal duct lobular unit *only of female breast* showing its distension and distortion.
   - It is 12 times more common in *white females*. Predominantly *perimenopausal.*
   - It is 3-5% common. High chance to predispose to invasive cancer. 35% of LCIS may develop invasive lobular carcinoma either in same or contralateral breast; 65% may develop invasive ductal cancer (same side/opposite side/both sides). LCIS is a *marker/predictor* of increased risk of invasive breast cancer; not an anatomical precursor unlike DCIS. It is now *advocated as a risk factor* for developing breast cancer.
   - It is multifocal, bilateral (50%).
   - It is an incidental pathological entity. *Classical type* carries better prognosis; *pleomorphic type* does not so; occasionally *mixed* ductal and lobular *in situ* may be seen.
   - Clinically it does not form a lump.
   - Need not be detected by mammography, as it does not provoke calcification.

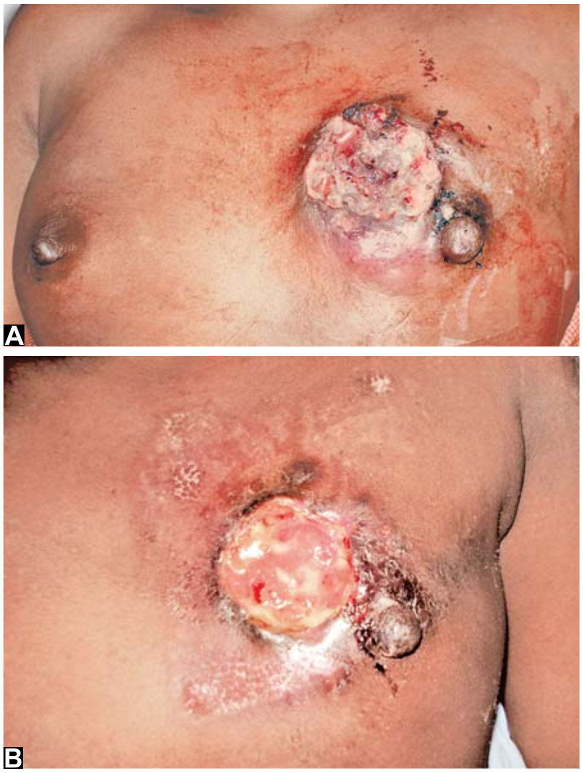

Figs 8.42A and B: Paget's disease of the breast.

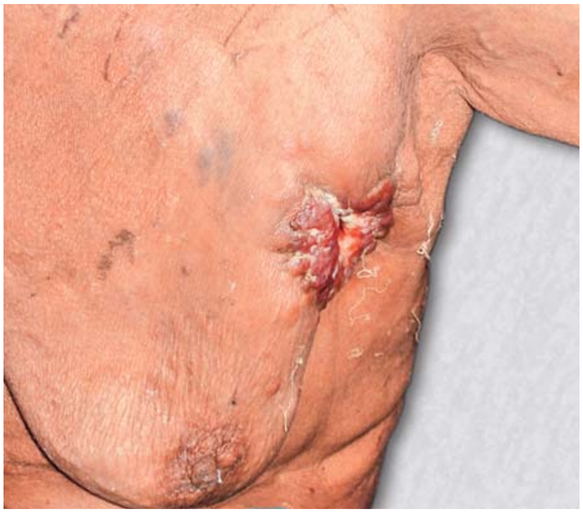

Fig. 8.43: Carcinoma breast involving axillary tail of Spence.

- 50% cancers can develop in the contralateral breast.
- Immunohistochemistry using e-cadherin antibody shows positive reaction in lobular carcinoma.
- It has poor prognosis due to bilateral, multifocal nature and difficulty in identifying it.

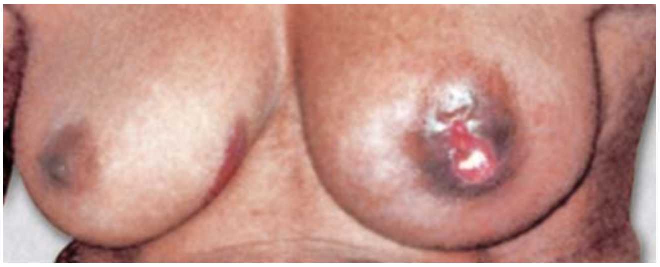

Fig. 8.44: Subareolar carcinoma with destruction of nipple-areolar complex.

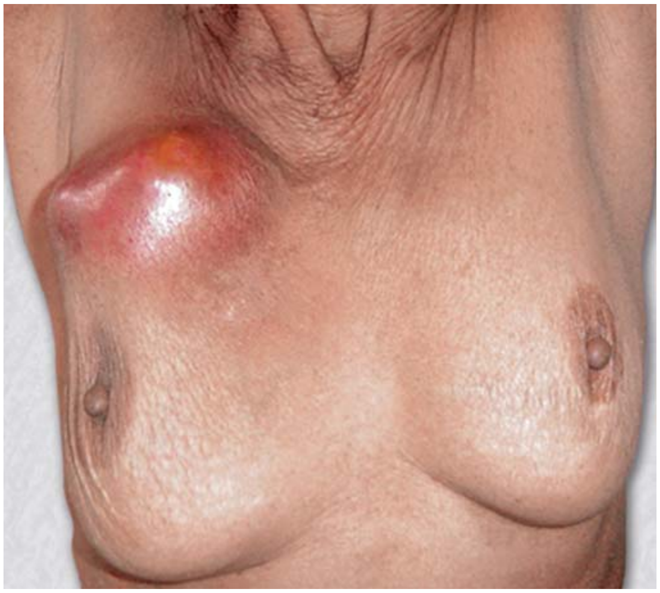

Fig. 8.45: Carcinoma breast over the most common site—upper outer quadrant—more visible on raising the arm.

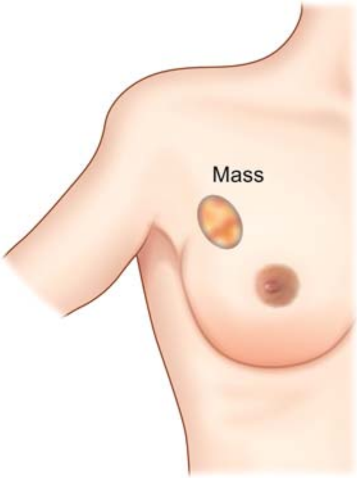

Fig. 8.46: Mass in the upper outer quadrant—is the most common site of the carcinoma breast.

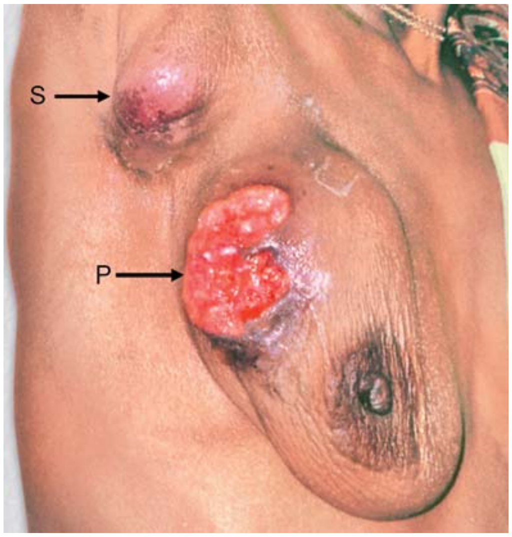

Fig. 8.47: Carcinoma right breast with ulceration in the primary (P) with axillary lymph node secondaries (S).

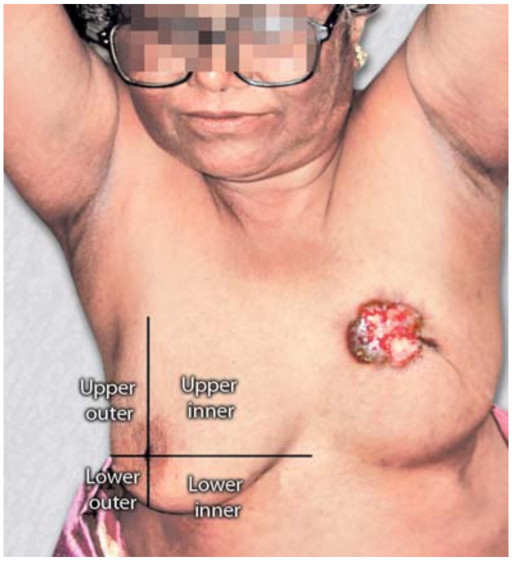

Fig. 8.48: Quadrants of breast. Carcinoma is more common in upper outer quadrant as more breast tissue is located in this quadrant.

- Tamoxifen (risk reduction in premenopausal) or raloxifen (postmenopausal) often with bilateral total mastectomy is the treatment.

9. ***Disease of Reclus:*** It is a rare intracystic papilliferous carcinoma of breast presenting as a cystic swelling with bloody discharge from the nipple.

Grading of the Tumour

- It is based on nuclear pleomorphism; tubule formation; mitotic rate.
- It can be—well-differentiated (grade 1); moderately differentiated (grade 2) and poorly differentiated (grade 3).

Tumour doubling occurs in 6 months with reaching 1 cm in size in 30 doublings. Breast cancer more than 1 cm has its own blood supply and so high chances of systemic spread and so systemic therapy is a must.

Elston-Ellis Modified Bloom-Richardson Grading System

Parameters used

- **Nuclear pleomorphism**: Score 1—relatively small uniform nuclei; score 2—intermediate pleomorphic nucleoli; score 3—relatively large prominent nucleoli
- **Mitotic count**: Score 1—< 10% mitoses in 10 HPF; score 2—10-20% mitoses; score 3—>20% mitoses
- **Tubule formation**: Score 1—>75% cells in tubule forms; score 2—10-75% cells in tubule forms; score 3—< 10% cells in tubule forms

Grade	Score
Favourable	1
Unfavourable	Up to 3
Grade I	3-5
Grade II	6-7
Grade III	8-9

- Well differentiated low grade—grade 3, 4, 5
- Moderately differentiated intermediate grade—grade 6, 7
- Poorly differentiated high grade—8, 9

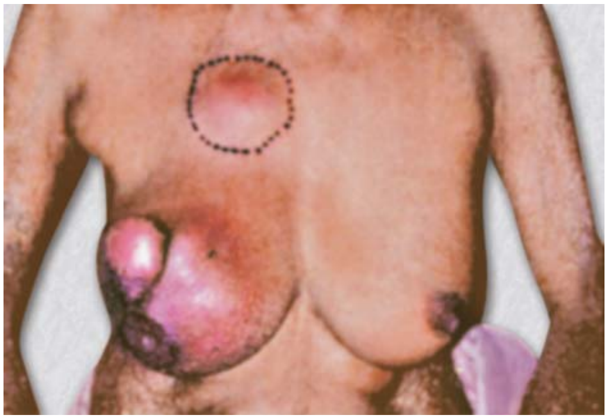

Fig. 8.49: Carcinoma breast with rib secondaries.

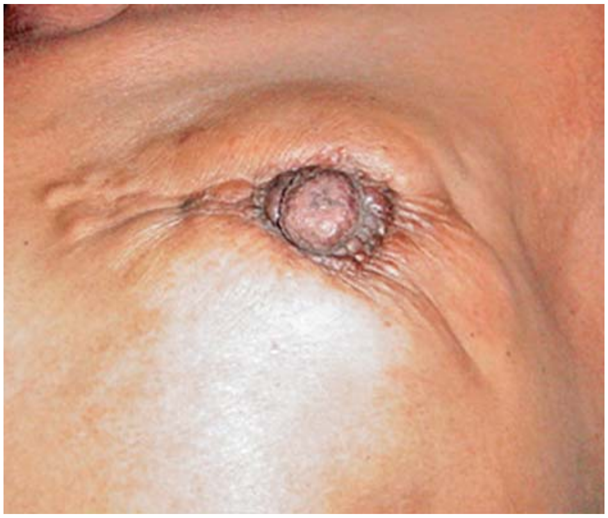

Fig. 8.50: Carcinoma breast with extensive skin involvement.

Biological Behaviour and Clinical Features of Carcinoma Breast

Most common site is *upper outer quadrant (60%) because breast tissue is more in this quadrant.*

Cutaneous Manifestations of Carcinoma Breast

- *Peau d'orange*: Due to obstruction of dermal lymphatics, openings of the sebaceous glands and hair follicles get buried in the oedema giving rise to orange peel appearance.
- *Dimpling of skin* due to infiltration of ligament of Cooper.
- *Retraction of nipple* due to infiltration of lactiferous duct.
- *Ulceration, discharge* from the nipple and areola.
- Skin ulceration and fungation.
- *Cancer-en-cuirasse*: Skin over the chest wall and breast is studded with cancer nodules appearing like an armour coat.
- *Tethering* to skin.

Spread into the Deeper Plane

- Into pectoralis major muscle (is confirmed by observing the restricted mobility of the swelling while contracting the PM muscle).
- Into latissimus dorsi muscle (extending the shoulder against resistance).
- Into serratus anterior (by pushing the wall with hands without flexing the elbow).
- Into the chest wall (breast will not fall forward when leaning forward, and while raising the arm above the shoulder, breast will not move upwards as it is fixed to the chest wall).

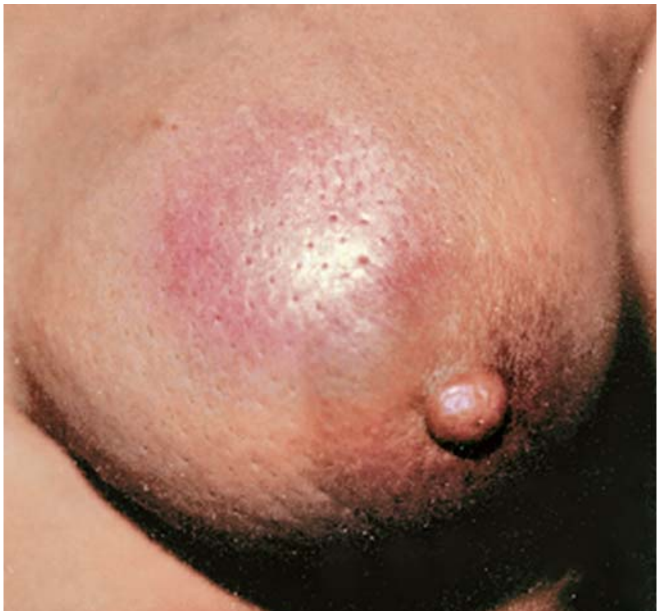

Fig. 8.51: *Peau d'orange* appearance of skin.

> *Mammary fistula presents as a recurrent abscess that points and discharges on to the areola and continues to discharge for weeks at a time.*
> **—Sir Headly Atkins**

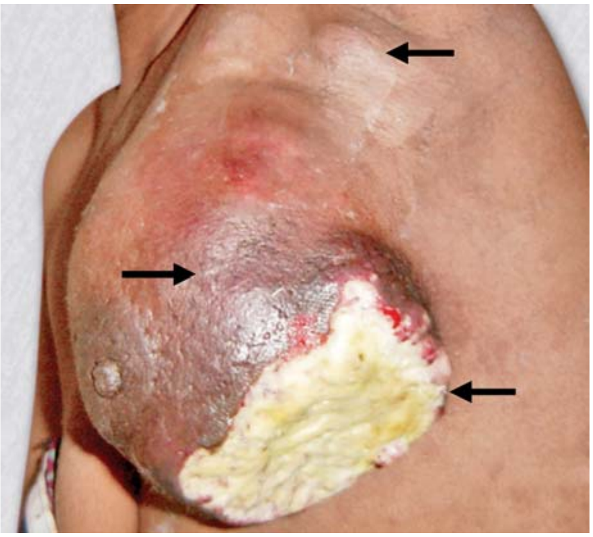

Fig. 8.52: Ulcerated carcinoma breast with peau d'orange. Note the lymph node enlargement.

Lymphatic Spread

- It occurs through:
    - Subareolar Sappey's lymphatic plexus (presently its Significance is discounted).
    - Cutaneous lymphatics.
    - Intramammary lymphatics.

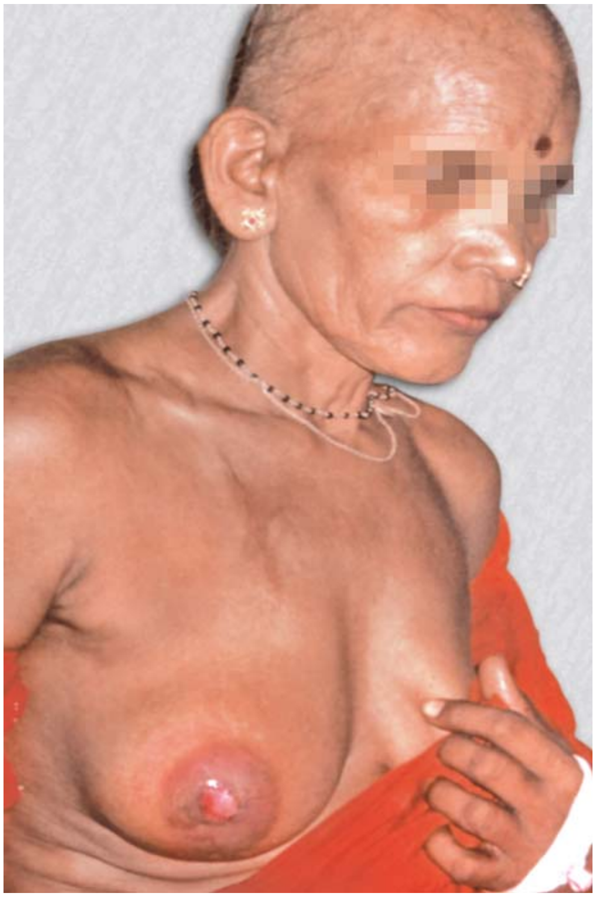

Fig. 8.53: Patient with carcinoma breast on chemotherapy having alopecia.

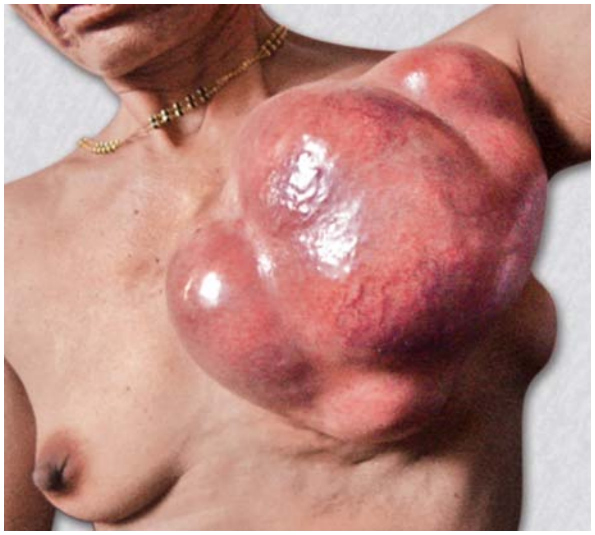

Fig. 8.54: Advanced carcinoma left breast involving entire breast with nodularity.

- Lymphatic drainage of the breast is predominantly through axillary (75%) and internal mammary lymph nodes.

Axillary group of nodes are:

1. *Anterior* along lateral thoracic vessels (Pectoral).
2. *Central* embedded in fat in the centre of the axilla.
3. *Posterior* along subscapular vessels.
4. *Lateral* along axillary vein.
5. *Apical* lies above pectoralis minor tendon in continuity with the lateral nodes and receive efferents from all the groups. Spread to these lymph nodes occur by *lymphatic permeation.*

- *Interpectoral,* lies between pectoralis major and minor muscle (***Rotter's nodes***). Presently involvement of these lymph nodes are considered due to retrograde spread.

  These lymph nodes are cleared during *Patey's mastectomy.*

Groups 1st and 2nd are commonly involved.
Groups 3rd and 4th are rarely involved.
Finally Group 5th is involved.

- From axillary lymph nodes spread occurs to supraclavicular lymph nodes by *lymphatic embolisation.*
- Through dermal lymphatics, it may spread to opposite breast or to opposite axillary lymph nodes.
- Spread restricted to level I nodes carries better prognosis.
  Spread to level II has poor prognosis.
  Spread to level III indicates worst prognosis.
- Spread may occur into internal mammary lymph nodes of same side and then to mediastinal lymph nodes.
- Contralateral internal mammary lymph nodes can also get involved by retrograde spread.
- *Fixed enlarged axillary nodes* can cause lymphoedema due to lymphatic block; venous thrombosis and venous oedema due to venous block; and severe excruciating pain along the distribution of the median and ulnar nerves (rare in radial nerve) with often significant sensory and motor deficits due to tumour infiltration of the cords of brachial plexus (medial cord often lateral cord).

Once axillary lymph nodes get fixed, it can result in ***lymphoedema of the upper limb.*** It develops gradually.

Compression of nodes on axillary vein can cause sudden onset of ***venous oedema of upper limb*** (Venous edema develops faster, and is more proximal with bluish discolouration of the skin of the upper limb. It may lead to venous gangrene and so the compression has to be relieved early by radiotherapy).

Causes of lymphatic block in carcinoma breast	***Treatment of lymphoedema of arm***
• Involvement and fixation of the axillary nodes level I, II and III	• Elevation of limb • Elastic stockings • Pneumatic compression
• After levels I, II and III dissection	• Drugs like diuretics and benzopyrones
• After radiotherapy to axilla	
• Inoperable fixed nodes in axilla	
• Recurrent axillary disease	
• May be associated with *cancer-en-cuirasse*	
• Secondary infection	
Effects of lymphatic obstruction	
• Peau d'orange	
• Brawny oedema of arm—indurated, painful, non-pitting—occurs in fixed nodes in axilla	
• Elephantiasis chirurgens—after radical mastectomy or radiotherapy to axilla	
• *Cancer-en-cuirasse*—seen in locally advanced carcinoma of breast. Skin of chest wall is studded with hard fixed nodules like armour coat (of soldiers)	
• Lymphangiosarcoma after radical mastectomy or MRM (Stewart-Treve's)	

Breast is in subcutaneous plane, but its extension, *axillary tail of Spence* passes through an opening in the deep fascia (*foramen of Langer)*. Often it is difficult to differentiate between lymph node in pectoral region and tumour invasion of axillary tail. Mobility will be independent if it is a lymph node, but if it is an axillary tail tumour, it is along with the primary tumour in the breast.

Haematogenous Spread

- ***Bone (most common)*** **(70%)**
  - Lumbar vertebrae, femur, ends of long bones, thoracic vertebrae, ribs, skull, in order.
  - They are *osteolytic* lesion often with pathological fracture.

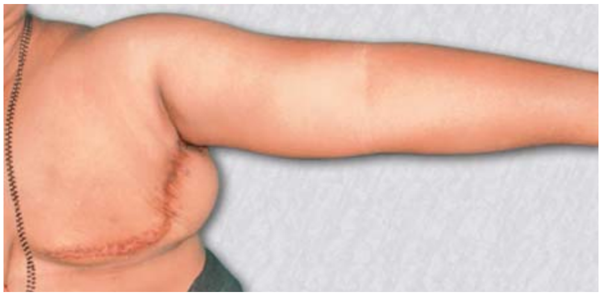

Fig. 8.55: Carcinoma breast operated, with upper limb lymphoedema postoperatively.

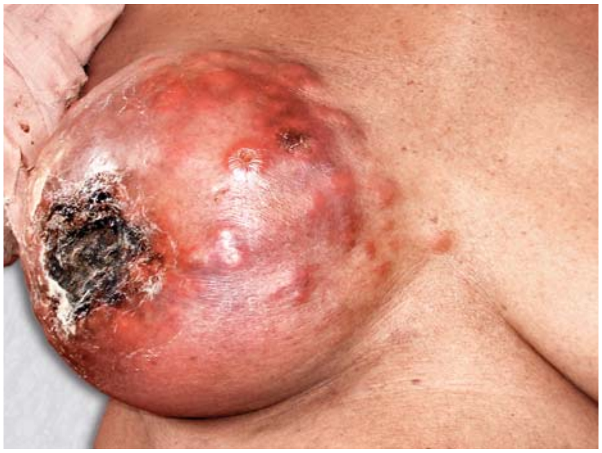

Fig. 8.56: *Cancer-en-cuirasse* with malignant nodules in the chest wall and skin.

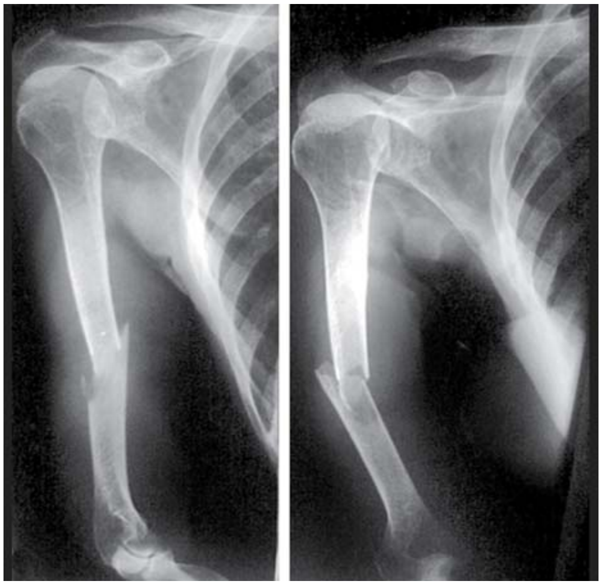

Fig. 8.57: Pathological fracture of the humerus due to metastatic secondaries from the carcinoma breast.

  - Presents with painful, tender, hard, nonmobile swelling, with disability.
  - *70% of secondaries in bone in a women is due to carcinoma breast.*
  - Spine secondaries can cause paraplegia.

Wise man has got the art of knowing what to overlook.

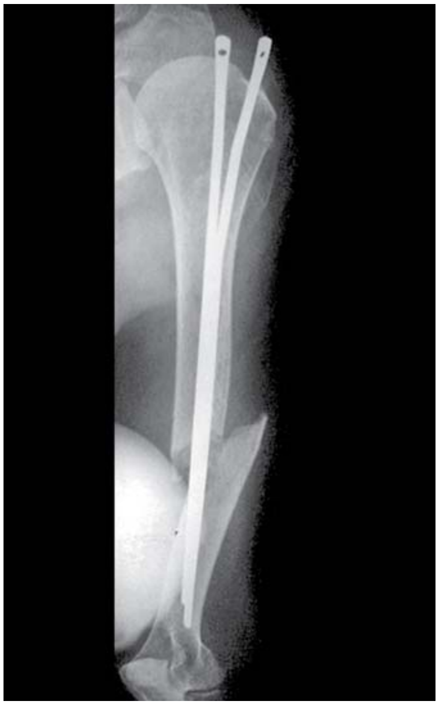

Fig. 8.58: Intramedullary nailing as a treatment for pathological fracture.

Features of bone secondaries

- Pain
- Swelling
- Pathological fractures

- ***Liver***—either through blood, occasionally through transcoelomic spread.
- ***Lung***—causes malignant pleural effusion and 'cannon ball' secondaries.
- ***Brain***—causes increased intracranial pressure, coning.
- ***Adrenals and ovaries***.

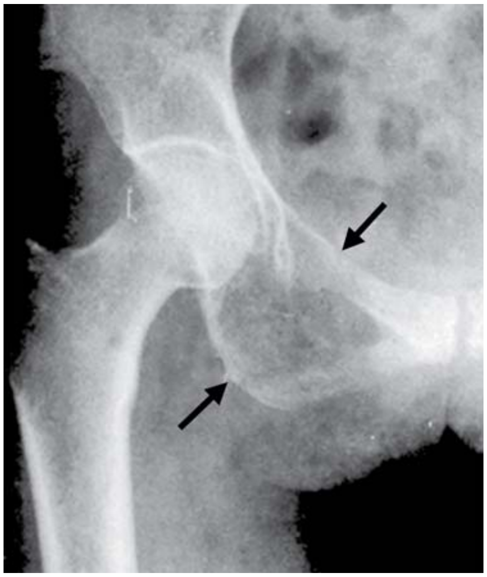

Fig. 8.59: Osteolytic secondaries involving ischium and pubic bones.

Common sites of distant spread in carcinoma breast

- Bones—70% (lumbar vertebrae, pelvic bones, long bones)
- Lungs and pleura—20-30%
- Soft tissues—5-15%
- Liver—10-12%
- Brain—2-5%
- Adrenals—2-5%

Transcoelomic Spread

Through mediastinal lymph nodes, it may spread into peritoneal cavity causing secondaries in liver, peritoneum, ovary (*Krukenberg secondaries*—occurs in menstruating age groups. During ovulation, cells get attached over the ovarian capsule).

***Note*:**

Present concept of Krukenberg tumour is haematogenous and lymphatic modes of spread. Older concept of transcoelomic spread is no longer well-accepted.

Presentation of carcinoma breast

- Lump in the breast which is hard, painless (most common). At least tumour should become 1 cm to clinically palpable (Fig. 8.62)
- Nipple discharge is the second common presentation
- Ulceration and fungation
- Axillary lymph node enlargement; supraclavicular lymph node enlargement
- Chest pain and haemoptysis
- Bone pain, tenderness, and pathological fracture
- Pleural effusion, ascites
- Liver secondaries, secondary ovarian tumour
- Pain in the lump in 10% cases

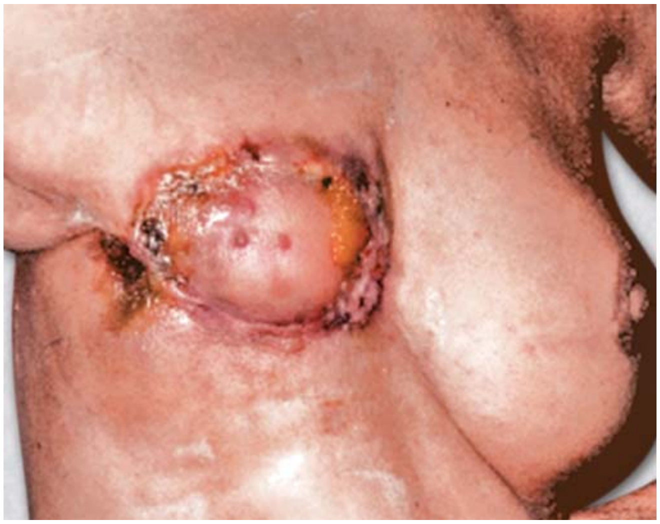

Fig. 8.60: Recurrent carcinoma breast.

Note:

Dominant breast mass is three-dimensional, distinct from surrounding tissue of same breast, asymmetric in relation to other breast.

Breast Self-examination (BSE)

Breast self-examination plays a major role in early detection and intervention of breast carcinoma. This underlines the importance of advocating *self-examination of the breast.*

Ideally done once a month just after the menstruation, as during this time breasts are less engorged. In postmenopausal age group it is done at monthly regular intervals.

- Examine both breasts
- Remind the patient that 90% of breast lumps are not cancer.
- Better way is in lying down position with arm raised with a mattress support behind
- Palpation should be using the fingers over all quadrants of the breast
- If any doubtful swelling is palpable, consult the surgeon
- American Cancer Society recommends monthly BSE after 20 years of age
- Nursing mother should perform BSE just after feeding the baby

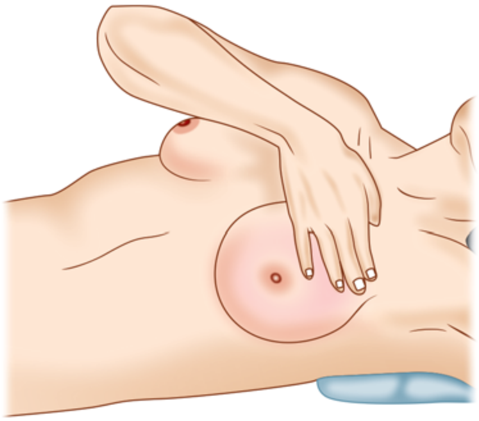

Fig. 8.61: Self-examination of breast is done in lying down position.

Staging of Carcinoma Breast (Manchester and TNM Staging)

Manchester Staging

1. Tumour in the breast, not involving pectoral or deeper plane. Skin involvement if present, it is lesser than the size of tumour. Lymph nodes are not palpable.
2. Same as stage I but with *mobile, discrete lymph nodes* palpable in the ipsilateral axilla.
3. Tumour fixed to pectoral muscle, or skin involvement more than the tumour size or ipsilateral axillary lymph nodes adherent to each other or fixed.
4. Tumour fixed to the chest wall, 'cancer-en-cuirasse', skin involvement wider than that of the breast, involvement of ipsilateral or contralateral supraclavicular lymph nodes or opposite breast or opposite axillary lymph nodes, spread to bone, lung, liver or inflammatory carcinoma of breast.

TNM Staging (AJCC Cancer Staging Manual, 2002, Sixth Edition)

Tumour:

T_x – Tumour cannot be assessed.

T_0 – No evidence of primary.

Tis – Carcinoma *in situ* (DCIS or LCIS)

Tis Paget's – Paget's disease of nipple with no tumour (with tumour underneath is staged according to size)

T_1 mic – Microinvasion < 0.1 cm.

T_1— Tumour size < 2 cm in greatest diameter (T_{1a}—0.1-0.5 cm; T_{1b}—0.5-1.0 cm; T_{1c}—1-2 cm).

T_2 – Size 2-5 cm.

T_3 – Size > 5 cm.

T_4 – Tumour fixed to chest wall or skin (T_{4a}—fixed to chest wall, T_{4b}—fixed to skin, T_{4c}-T_{4a} + T_{4b}, T_{4d}—inflammatory carcinoma breast).

Node:

N_X – Nodes cannot be assessed.

N_0 – No nodes.

N_1 mic – Node with micrometasiasis.

N_1 – Axillary nodes—ipsilateral, mobile, discrete.

N_2 –

N_{2a} – Axillary nodes—ipsilateral fixed to one another and other structures.

N_{2b} – *Clinically apparent** and ipsilateral internal mammary nodes in the absence of clinically palpable axillary nodes.

* *Clinically apparent* means nodes detected by imaging/clinically/pathologically.

N_3 –

N_{3a} – Spread to ipsilateral infraclavicular lymph nodes with or without axillary nodes.

N_{3b} – Spread to ipsilateral internal mammary nodes and axillary nodes.

N_{3c} – Spread to ipsilateral supraclavicular lymph nodes with/without axillary or internal mammary nodes.

Metastasis:

M_X – Metastases cannot be assessed.

M_0 – No metastasis.

M_1 – Distant Metastases.

Stage I : $T_1N_0M_0$
Stage IIa : $T_0N_1M_0$;$T_1N_1M_0$;$T_2N_0M_0$.
Stage IIb : $T_2N_1M_0$;$T_3N_0M_0$
Stage IIIa : $T_0N_2M_0$;$T_1N_2M_0$;$T_2N_2M_0$;$T_3N_1M_0$;$T_3N_2M_0$
Stage IIIb : $T_4N_0M_0$;$T_4N_1M_0$;$T_4N_2M_0$
Stage IIIc : Any TN_3M_0
Stage IV : Any T, any N, M

Early breast cancer—Stage I and II; T1N1, T2N1; T3N0
Locally advanced breast cancer (LABC)—Stage IIIA, IIIB
Metastatic breast cancer—Stage IV

Bilateral Breast Cancer

- It is separate primary breast cancer on the opposite side in a patient who is having or treated for ipsilateral breast cancer.
- It can be synchronous or metachronous.
- There is 5 fold increase in 2nd breast cancer on opposite side in a female who has had breast cancer on ipsilateral side.
- LCIS, multifocal breast cancer, family history are other risk factors.

If you don't know where you're going, you may miss it when you get there.

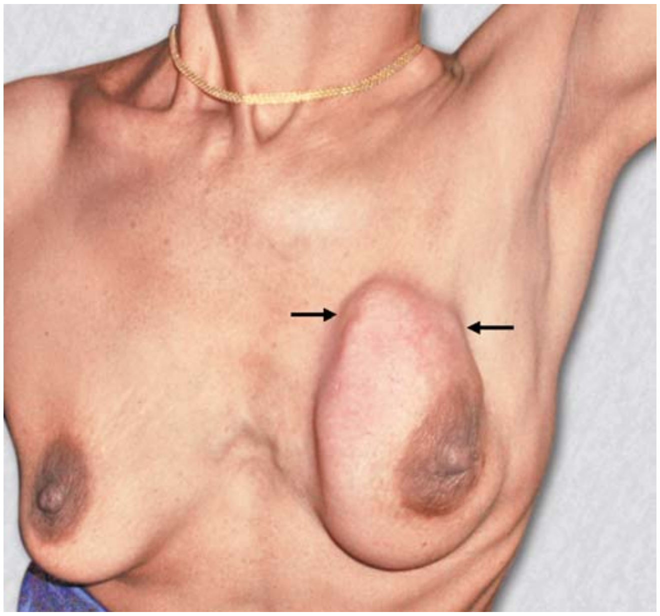

Fig. 8.62: Carcinoma breast presenting as lump in the breast. It is the most common presentation.

- *Choudary Millis criteria* that help to differentiate primary contralateral breast cancer from metastatic—*in situ* change, different histological pattern, different differentiation shown in opposite breast, clinically or by evaluation ipsilateral tumour has not shown any features of spread.
- It is investigated like any breast cancer, staged accordingly and treated.
- If tumour on the opposite side is metastatic then it will show separate histology and differentiation with mammography showing less infiltrative, diffuse, without microcalcifications but with oedema. If it is metastatic it is stage IV disease.

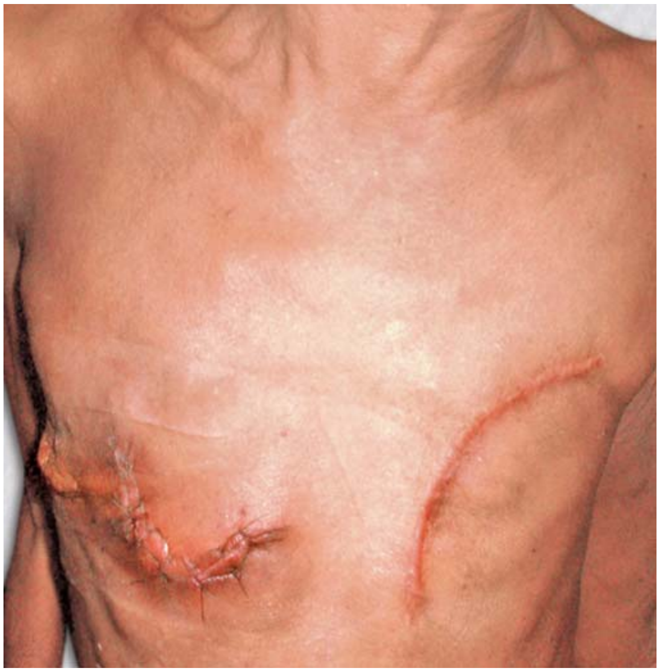

Fig. 8.63: Bilateral breast cancer operated.

Differential diagnosis for carcinoma breast

- Fibroadenosis
- Traumatic fat necrosis
- Tuberculosis of breast
- Blood good cyst
- Filariasis breast
- Mastitis
- Antibioma
- Galactocele
- Mondor's disease
- Cystosarcoma phyllodes

Investigations in Carcinoma Breast

- ***Mammography***.

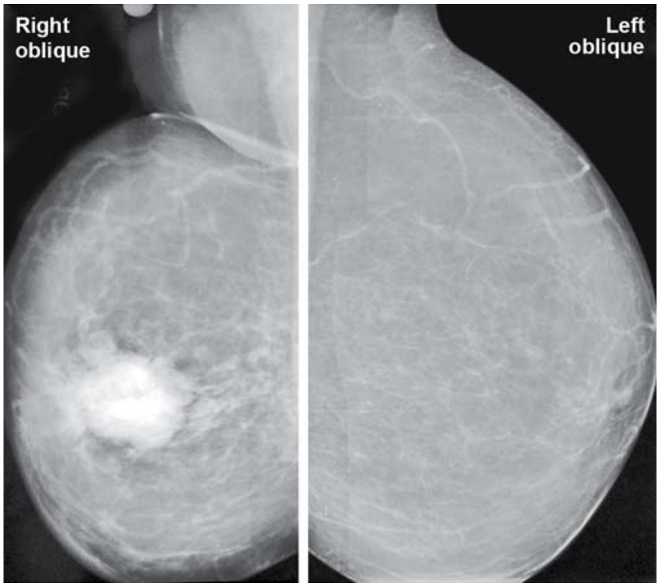

Fig. 8.64: Mammography in carcinoma breast.

Findings

- Size and location of mass lesion
- Microcalcifications signify malignancy
- Soft tissue shadow is irregular
- Spiculations

- Bilateral mammography is done to identify multicentricity, to have guideline for assessing eventual chemotherapy or RT in LABC
- Stereotactic mammography guided biopsy is very useful
- LCIS may be missed in mammography
- 50% of breast cancers can be seen on mammography before they are palpable
- It is noninvasive with less radiation exposure
- 5% false-positive rate in mammography; hence biopsy is a must
- Mammogram is used during follow-up period after conservative breast surgery and of opposite side
- Ideally specimen mammography is a must after conservative breast surgery to assess the completion; and after core biopsy to confirm the sample tissue
- ***Ultrasound of breast:***
  - To find out whether the lesion is solid or cystic.

Ultrasound of breast

- To look for whether the lesion is solid or cystic, margin of the lesion, internal echoes, retro-tumour acoustic shadowing, compressibility, dimensions
- *Irregular margin, irregular internal echoes, irregular posterior shadowing, noncompressibility, ratio between anteroposterior to width (lateral/horizontal) dimensions more than 1 are the features of carcinoma.* Doppler will show high frequency signals with continuous flow. It is hypoechoic with more vertical taller growth
- Benign lesions are smooth, rounded with well-defined margins with weak internal echoes and compressibility. It can be elliptical, hyperechoic/hypoechoic smooth lesion
- Cyst is anechoic, oval/round, well circumscribed lesion
- Disadvantage is lesions less than 1 cm may not be identified
- FNAC can be done under U/S guidance
- It is cheaper, easily available and there is no risk of radiation
- It is preferred method of screening in young females where mammography is not done and in pregnancy and early lactation
- US of axilla also can be done to assess axilla and to do guided FNAC of node

- ***FNAC*** (Martin and Ellis 1930): It is very useful in diagnosing the carcinoma breast. U/S guided FNAC is also used. But negative results are difficult to interpret because it may be due to sampling errors and so requires further diagnostic methods. FNAC of opposite breast, lymph nodes, opposite axillary lymph nodes are also often required.

It is done with 23 gauge needle using FNAC aspiration special syringe (aspiration gun). With the lump held firmly, the needle is passed into the lump and with negative pressure continuous aspiration is done until adequate material comes through the needle (suction pressure of 40 cm of H_2O is created into the syringe). Needle with syringe is removed without negative pressure. Material is collected on a slide; a smear is made using 100% alcohol. Cytology is studied after staining it under microscopy.

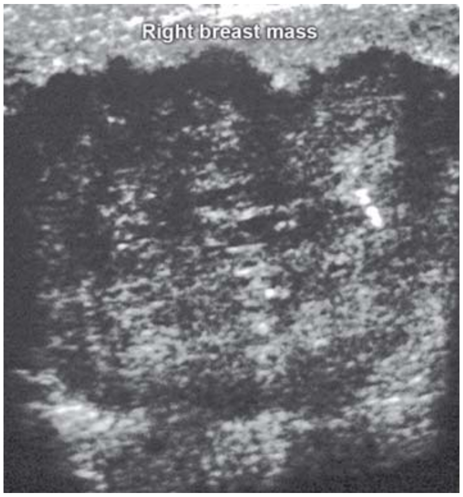

Fig. 8.65: US picture of right breast showing breast mass.

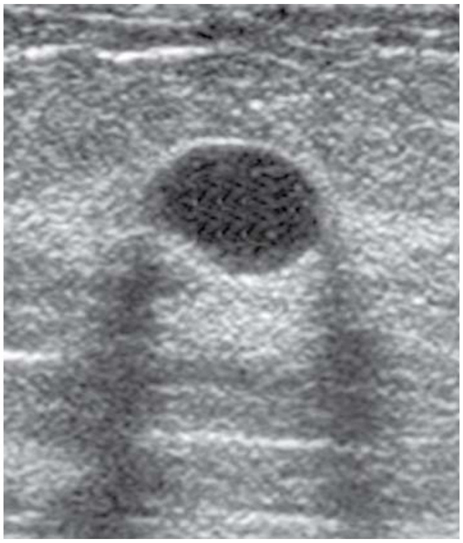

Fig. 8.66: US breast is basic noninvasive simple investigation to assess breast lump.

The Columbia classification (Haagsen, Cooley and Stout)

GRAVE SIGNS:	
• Oedema of skin	Stage A—only tumour. No Grave signs
• Skin ulceration	Stage B—tumour + axillary lymph nodes < 2.5 cm
• Fixity to chest wall	Stage C—tumour + anyone of five grave signs
• Axillary lymph nodes >2.5 cm	Stage D—two or more Grave signs
• Fixed axillary lymph nodes	Supraclavicular lymph node involvement Distant metastasis
Stage A—	No skin oedema, ulceration, or fixation to chest wall. Axillary nodes are not clinically involved.
Stage B—	Clinically involved axillary nodes less than 2.5 cm in diameter. Not fixed.
Stage C—	Grave signs of comparatively advanced carcinoma. Oedema of skin, skin ulceration, fixation to chest wall. Massive axillary involvement with nodes > 2.5 cm in diameter. Axillary fixation.
Stage D—	Advanced carcinoma including two or more signs in stage C. In addition satellite nodules, supraclavicular nodes. Inflammatory cancer, arm oedema or distant metastasis.

It is better to light a small candle than to curse the darkness.

- Minimum six aspirations are done.
- Giemsa, Papanicolaou, hematoxylin and eosin stains are used.
- Repeat FNAC can be done for further 2 times.
- But it gives only cytological diagnosis. Receptor study cannot be done.
- It is difficult to differentiate between *in situ* and invasive breast cancer by FNAC.

Advantages

FNAC is least painful, can be done on OP basis, reliable and cheaper. Malignant deposits will not occur along FNAC track (only contraindication for FNAC is testicular tumour).

***Note*:**
FNNAC is **F**ine **N**eedle **N**on-**A**spirating **C**ytology.

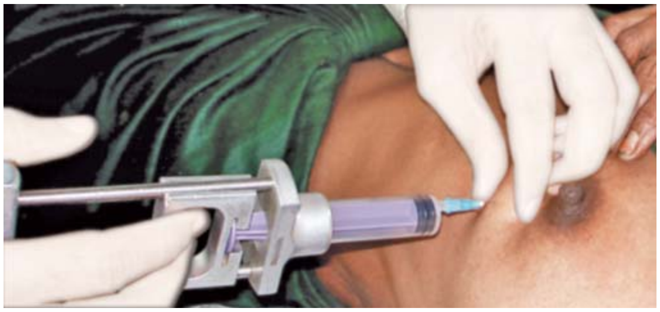

Fig. 8.67: FNAC of breast lump.

Reliability of FNAC and mammography

	FNAC	*Mammography*
Sensitivity (true positivity)	90-98%	90%
Specificity (without false-positive)	98-100%	90%
False-negative	2-10%	10%
False-positive	Near 1-5%	10%

FNAC scoring

C_0 : No epithelial cells
C_1 : Scanty epithelial cells, benign
C_2 : Benign cells
C_3 : Atypical cells
C_4 : Suspicious cells
C_5 : Malignant cells

- ***Frozen section biopsy***: If FNAC fails even after two trials or in cases of negative FNAC, then on table frozen section biopsy is done for diagnosis. Frozen section biopsy also has got drawbacks. It has got 20% false-negative results. So its use at present is under debate eventhough it is practiced in few oncocenters. It is *not ideal method*.
- ***Corecut/Trucut biopsy:***
  - It is done under local anaesthesia. It gives clear histological evidence and also confirms DCIS (FNAC can not confirm DCIS). This allows proper neoadjuvant/primary chemotherapy, receptor status of the tumour. Wide bore needle biopsy with vacuum is also used.
  - 14-18 gauge spring loaded needle is used. Multiple punctures are needed. US guided biopsy is also done.
  - Large core biopsy is done using 6-14 gauge needle; single puncture is sufficient; large single sample of tissue is obtained.
  - Vacuum-assisted core biopsy is also done.
  - Stereotactic mammographic/MRI/US guided core biopsy are also used in small/impalpable lesions. Stereotactic core biopsy is done in prone position with compressed breast. Under local anaesthesia, 3 mm incision is made and 11 gauge core needle is passed under digital mammographic guidance. Multiple core biopsies are taken using vacuum. A clip may be placed under guidance at the site of the lesion as marker. Mammography of core biopsy specimen is done to confirm the sample tissue. If stereotactic biopsy is inconclusive then a wire localised surgical excision should be done.
  - Core needle biopsy is the *method of choice*.

Image guided biopsies

It is done when lump is not clearly palpable
- US guided core needle biopsy
- Stereotactic mammographic core needle biopsy
- Mammography guided wire localisation using needle sheath over the tumour and through an incision under local anesthesia hook is reached and biopsy is done. It is used if core needle biopsy fails in localising nonpalpable lesion
- MRI guided core needle biopsy

- ***Excision biopsy:***
  - It is done only when FNAC is inconclusive and a facility for frozen section is not available. Incision should be planned in such a way that it will be included in eventual mastectomy.
- ***Chest X-ray***:
  - To look for pleural effusion, cannon ball secondaries in lungs, mediastinal lymph nodes, secondaries in rib.
- ***CT chest:***
  - It is more reliable method to see lung secondaries.
- ***Ultrasound abdomen:***
  - To look for liver secondaries, ascites, 'Krukenberg' tumour.
- ***X-ray spine*** shows osteolytic secondaries.
- ***Oestrogen receptor study*:**
  - They are oestrogen sensitive receptors, which are cytosolic, glycoprotein present in the breast and tumour tissue. It is an important indicator of prognosis of carcinoma breast.
  - Tissue for receptor study is sent at low temperature in ice flasks. It is assessed by quantitative analysis (Frozen –70°).
  - If value is more than 10 units (f/mols) per n gram of tissue it is called as *ER +ve status*. ER positivity is common in postmenopausal women (60%) compared to premenopausal women (30%). If value is less than 10 units per gram of tissue it is called as *ER –ve status*.

  In ER +ve status
  - Prognosis is good.
  - Hormone therapy including tamoxifen is more beneficial.
  - Response to treatment is better.

In ER –ve status
  - Prognosis is poor.
  - Hormone therapy is not very beneficial (but used) as compared to ER +ve patients.
  - Response to treatment is not good.

- ***Progesterone receptor (PR status)*** study or *Her 2 Neu receptor status* or *cErb B2* (growth factor receptor study) are other studies done at present to plan the therapy and assess the prognosis.
  - *Her 2/Neu receptor* (Human epidermal growth receptor 2 Neu oncogene) is a tyrosine kinase receptor and is associated with ER negative, high grade, tumours. It carries poor prognosis. Over expression of Her 2/Neu shows good response to adriamycin.
- ***Study of discharge*** from the nipple.
  - Nipple discharge is usually unilateral in carcinoma breast. *Ductal lavage* may be useful in some patients. Microcatheter of 1 cm length is introduced gently into the ductal opening. 10 ml saline is infused through the catheter. Fluid is withdrawn into the syringe and cytological analysis is done.
- ***MRI of breast:***
  - To differentiate scar from recurrence.
  - To image breasts of women with implants.
  - To evaluate the management of axilla and recurrent disease.
  - It is useful in screening females with high-risk group and *young women*.
  - T1 and T2 weighed images are taken.
  - Irregular mass with spiculations, changes in skin and nipple, lymphoedema are the findings in carcinoma breast.

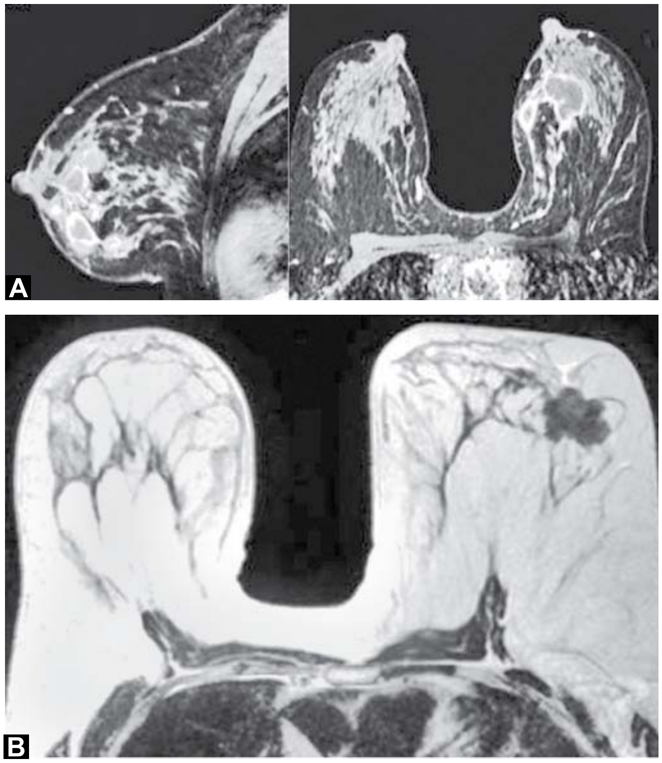

Figs 8.68A and B

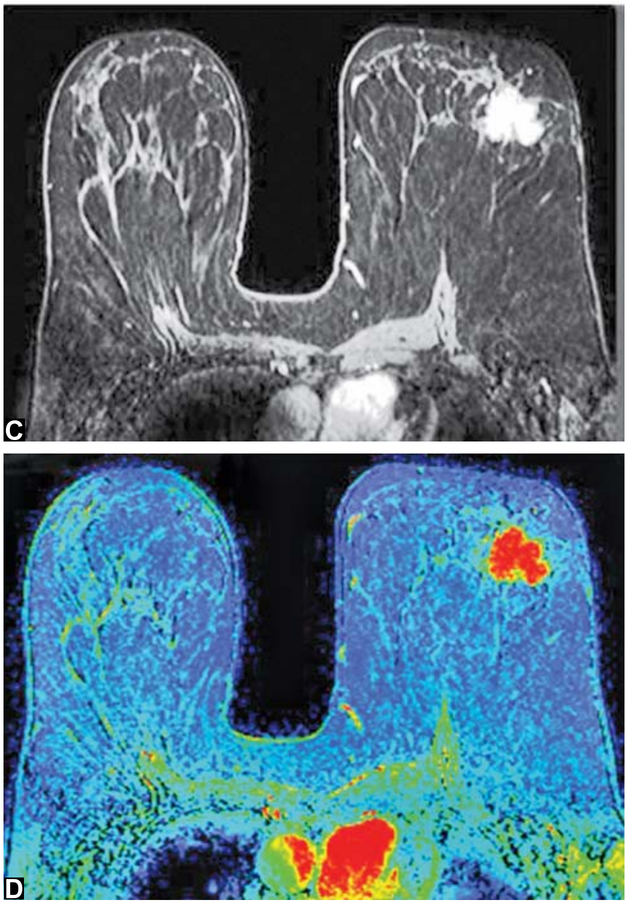

Figs 8.68C and D

Figs 8.68A to D: MRI of breast is ideal method to assess or identify recurrence after BCS or after breast implants. (A) T1 weighted with contrast; (B) T1 weighted; (C) T1 weighted with contrast and fat suppressed; (D) MRI perfusion image (*Courtesy:* Dr Raghavendra Bhat, MD, Radiologist, Balmatta Scan and Research Centre, Mangalore).

  - Lesion indeterminate by US/mammography is assessed by MRI.
  - Ionising radiation is not there with MRI.
  - It is the method *choice of imaging* breasts in pregnancy.
  - It is better in dense breasts.
  - Patient lies in *prone position* with breasts placed over breast coils; both precontrast and postcontrast (*gadolinium*) MRI are taken.
  - *Disadvantages are*—it is costly; nonavailability; not sensitive for premalignant lesions. It cannot be done in patients who are having incompatible metal prosthesis in the body. It is not accurate if done within 9 months of the radiotherapy for carcinoma breast.
- ***Edge biopsy:***
  - Done only when there is ulceration and fungation. Diathermy should be avoided in incision biopsy as it may distort the histology of tumour and study of hormone receptor status may not be possible.
- ***Tumour markers:*** CA 15/3 (normal value < 40 U/ml of serum) are used mainly during follow-up period. CEA, CA 15-3, CA 27-29 may be useful.

The place to improve this world is first in one's own heart, head and hands.

- ***MRI spine/pelvis*** show osteolytic secondaries in the bone like vertebra and pelvic bones.
- ***Radioisotope bone scan*** to look for secondaries in bone in advanced cases. It is not done routinely in early carcinoma of breast.

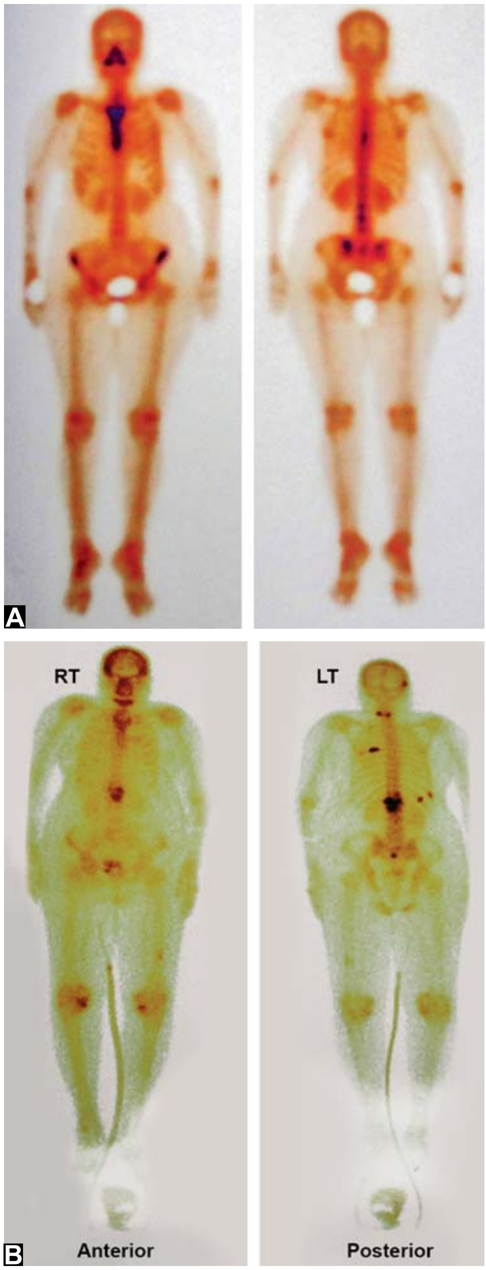

Figs 8.69A and B: Isotope bone scan is a must in LABC/T3 disease.

Indications for whole body bone scan in carcinoma breast

- T3, T4 advanced disease
- Advanced nodal disease
- Bone pain, bone swelling, pathological fracture
- Chest/liver secondaries

- ***PET scan:***
  - It may be an effective single scan for bone, soft tissue or visceral metastases in patients with symptoms or signs of metastatic disease.
- ***Sentinel lymph node biopsy (SLNB):***
  - The first axillary node draining the breast (by direct drainage) is designated as the *sentinel node (SLN)*. SLN is first node involved by tumour cells and presence or absence of its histological involvement, when assessed will give a predictive idea about the further spread of tumour to other nodes. The incidence of involvement of other nodes without SLN is less than 3% and so if SLNB is negative nodal dissection can be avoided but regular follow-up is needed. SLNB is done in all cases of early breast cancers, T1 and T2 *without clinically palpable node.*
  - It is not done in clinically palpable axillary node as there is already distortion of lymphatic flow due to tumour. It is also not done in multifocal and multicentric tumours, as there is involvement of many lymphatic trunks from different places of breast, and chances of false-negative is high.
  - Sentinel node is localised by preoperative (within 12 hours prior) or peroperative injection of patent blue (Isosulfan vital blue dye 2.5-7.5 ml) or 99m TC radioisotope labelled albumin (one mCi on previous day)/sulphur colloid (6 hours before) near the tumour (peritumour area) or into subdermal plexus around the nipple. Marker will pass through the sentinel node which can be visually detected as blue staining or with a hand held gamma camera; and is biopsied with a small incision made directly over it. Frozen section biopsy or touch imprint cytology is done for presence of malignant cells. If there is no involvement of sentinel node by tumour then further axillary dissection is not required as *skip lesions* (skipping sentinel node) occur only in less than 3% cases.
  - Detection rate of sentinel node for blue dye and radioisotope is 90% and 98%, respectively. Subdermal/subareolar injection of radioisotope *has got better* sentinel node localisation than peritumour injection. But better imaging is obtained by peritumour injection and so *peritumour injection is usually* practiced. Radioisotope tracer injection done in the early morning of the day of surgery into peritumour area and perioperative injection of patent blue dye in subareolar region—as a *combined method* is often used in many centers. After injection of patent blue, breast is massaged continuously to enhance the uptake. *Incision* is made after 5-7 minutes between pectoralis major and latissimus dorsi to identify blue stained lymphatics which are traced to 2-3 blue lymph nodes. Hand-held radioprobe is used to identify the sentinel node which is later excised. Often 2-3 nodes are removed.
  - *Paraffin section histology is better* than frozen section to identify positive sentinel lymph node. If report comes negative immunohistochemistry test is done to confirm that lymph node is negative for tumour. Sentinel lymph node biopsy *should be done before* wide local excision of the primary tumour.
  - Wide local excision of the primary tumour is done after SLNB in the same sitting.
  - SLNB is *less invasive* than axillary dissection. It is ideal in early invasive carcinoma. Positive SLNB is again classified as macrometastasis (> 2 mm) or micrometastasis (< 2 mm).
  - SLNB is *contraindicated* in patients who are allergic to vital blue dye or radio-colloid, in pregnancy and in inflammatory carcinoma of breast.

- *Complications* of SLNB are rare. They are—blue tattooing of skin which gradually fades; blue—green urine and stool for short period; allergic reactions; anaphylaxis (0.1%); seroma formation. Blue dye binds to oxyhaemoglobin and so artificially registers in oximeter as hypoxaemia.

Note:

Facility for SLNB is not available in many centres.

SLNB is done in:

- Carcinoma breast
- Carcinoma penis
- Malignant melanoma

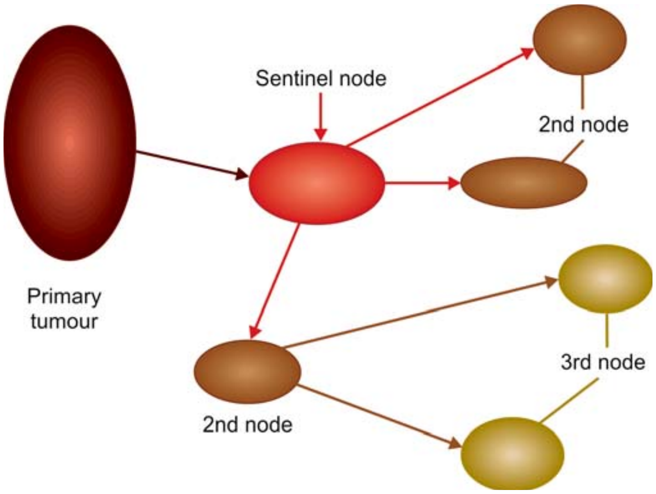

Fig. 8.70: Sentinel lymph node biopsy (SLNB) of breast. Note: Spread by skipping the sentinel node is less than 3%.

- ***Axillary sampling*:**
  - It is often done with an adequate axillary incision. 10-15 nodes are removed for sampling. It is not commonly practiced now (Minimum 10 nodes should be removed-level I nodes).
- ***CT scan of chest, abdomen and brain*** whenever needed.
  - CT is said to be more useful to detect secondaries in these regions.
- ***Ductography*:**
  - It is contrast study of ducts of breast in case of unilateral nipple discharge. Fine cannula is passed under vision carefully through the duct opening into the duct and 0.2 ml of dilute water-soluble contrast media is injected into the duct. Craniocaudal and mediolateral X-ray films are taken. Contrast irregular filling defect may be observed.
- ***Breast ductal endoscopy:***
  - It is useful in direct visualisation of the tumour in DCIS and invasive ductal carcinomas. But it is technically difficult and demanding.
- ***Thermography*:**
  - It is not very sensitive test (50%). Malignant tumours are hypervascular and so transmitted temperature is detected through different thermographic methods.
- ***Blood count, complete liver function tests*** are needed.

Triple assessment includes:

- Clinical assessment
- Radiological imaging US/MRI/ mammography (after 40 years)
- Cytological or histological analysis FNAC/core biopsy

Newer investigation modalities for carcinoma breast

- Stereotactic core biopsy using computer mammography
- Vacuum-assisted biopsy using 11-guage biopsy probe (becoming popular)
- Needle localised biopsy under mammographic guidance
- I^{125} – seed localisation biopsy

Treatment

Treatment of carcinoma breast

It is usually through a combined approach

- *Surgery*
- *Radiotherapy*
- *Hormone therapy*
- *Chemotherapy*

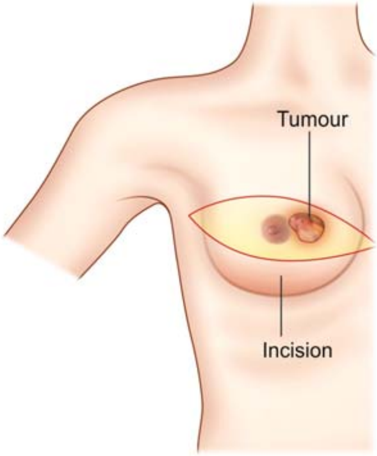

Fig. 8.71: Horizontal incision for total mastectomy.

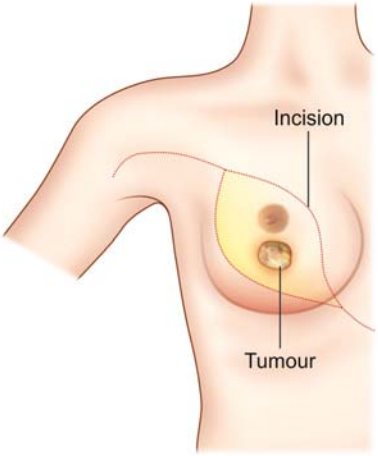

Fig. 8.72: Oblique incision extending into the anterior axillary fold for radical mastectomy.

Failure is the only opportunity to begin again more intelligently.

A. Surgeries for Carcinoma Breast

- ***Total (simple) mastectomy*:**
    - Along with the tumour, entire breast, areola, nipple, skin over the breast, including axillary tail are removed. There is no axillary dissection. Often the patient is subjected to radiotherapy later (External) to axilla. *Pectoral fascia* is removed along with breast specimen; breast tissue superficial to axillary fascia is removed.

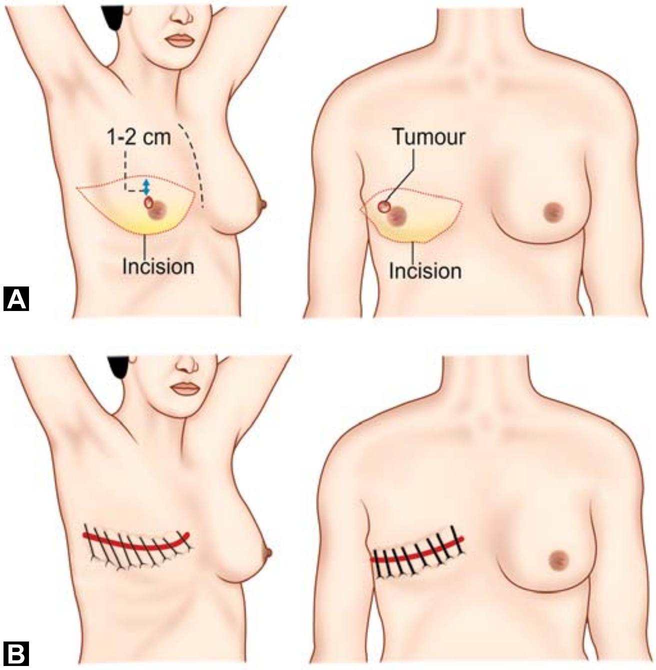
1-2 cm
Incision
Tumour
Incision
A
B

Figs 8.73A and B: *Stewart incision* for mastectomy.

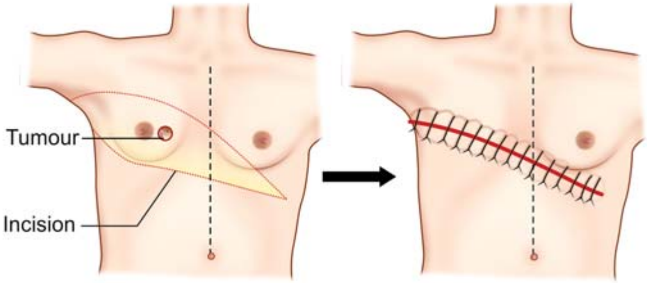
Tumour
Incision

Fig. 8.74: *Gray incision* for mastectomy in carcinoma breast which extends to opposite side.

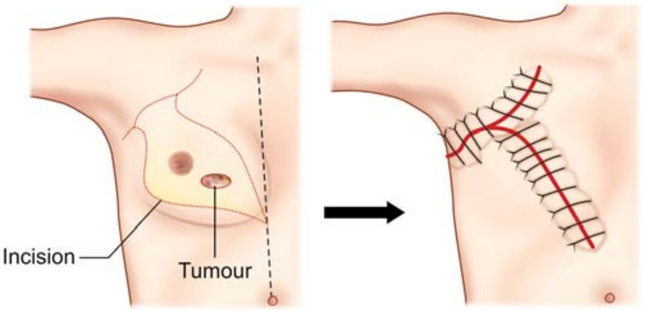
Incision
Tumour

Fig. 8.75: *Greenough incision* for mastectomy.

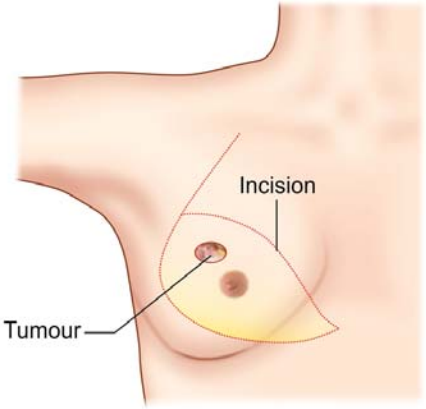
Incision
Tumour

Fig. 8.76: *Kocher's incision* for mastectomy.

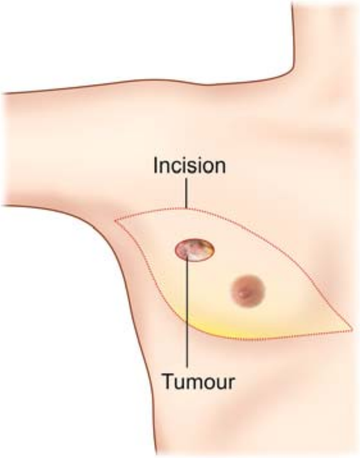
Incision
Tumour

Fig. 8.77: *Orr incision* for mastectomy.

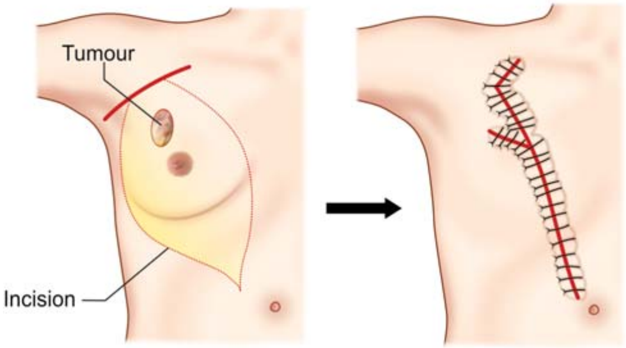
Tumour
Incision

Fig. 8.78: *Rodman incision* for mastectomy.

- ***Total mastectomy with axillary clearance*:**
    - Commonly used procedure. Total mastectomy is done along with removal of axillary fat, fascia and lymph nodes. *Level I and II axillary nodes* are removed along with total/simple mastectomy.

- ***Modified radical mastectomy [MRM]:***
    - ***Patey's operation*:** It is total mastectomy along with clearance of all levels of axillary nodes and removal of pectoralis minor muscle. It is enblock dissection of breast and axilla. An elliptical incision is made from medial aspect of the second and third intercostal space enclosing the nipple, areola and tumour extending laterally into the axilla along the anterior axillary fold. Upper and lower skin flaps are raised. Breast with tumour is raised from the medial aspect of the pectoralis major muscle. Dissection is proceeded laterally with ligating pectoral vessels. Once dissection reaches axilla, lateral border of pectoralis major muscle is cleared with level I nodes. Pectoralis minor is divided from coracoid process to clear level II nodes. Medial and lateral pectoral nerves should be preserved (otherwise atrophy of pectoralis minor muscle occurs). Later from the apex of axilla level III nodes are cleared. Nerve to serratus anterior, nerve to latissimus dorsi, intercostobrachial nerve, axillary vein, cephalic vein and pectoralis major muscle are preserved. Wound is closed with a suction drain.

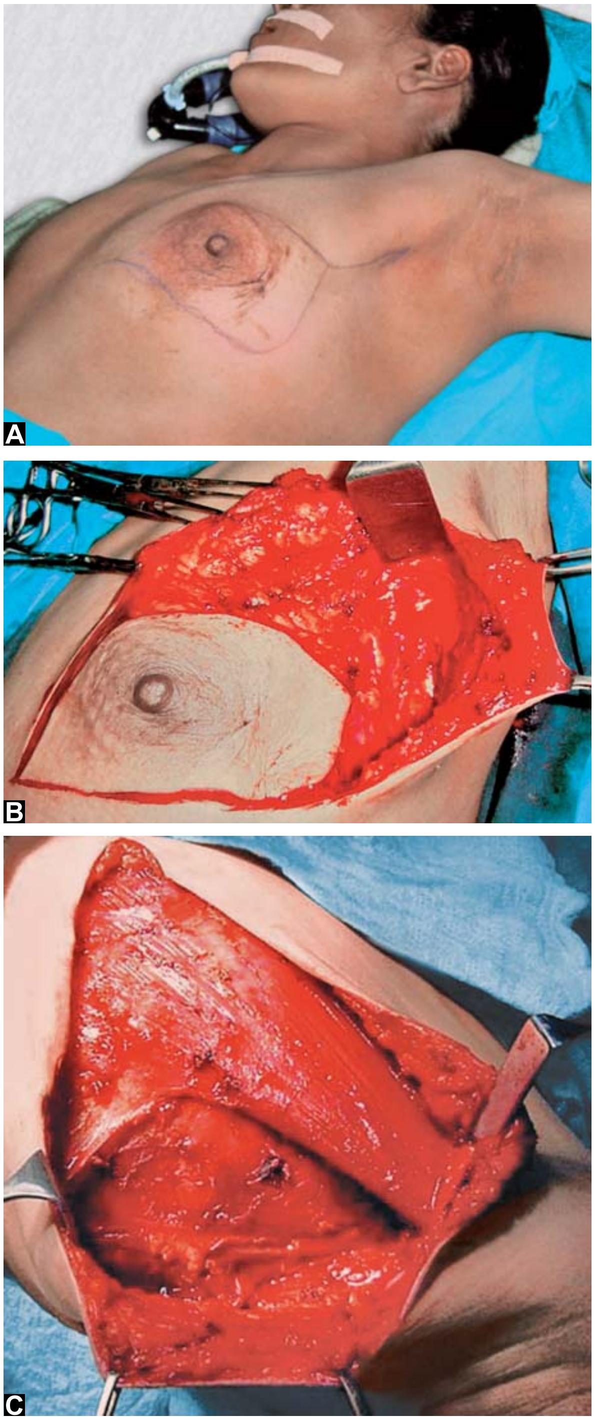

Figs 8.79A to C

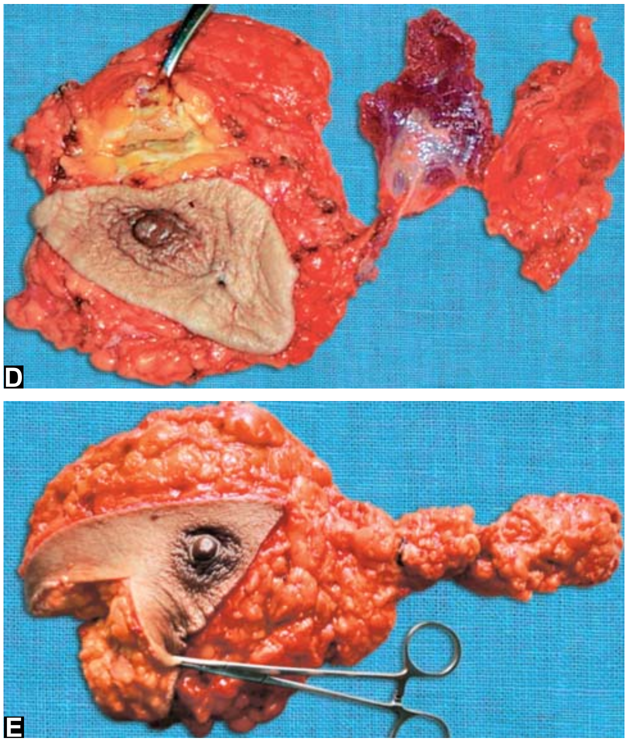

Figs 8.79D and E

Figs 8.79A to E: Different steps of Patey's mastectomy with incision, flap raise, dissection in breast and axilla and specimen after surgery.

    - ***Scanlon's operation:*** Is a modified Patey's operation wherein instead of removing pectoralis minor, it is incised to approach the affected level III lymph nodes.
    - ***Auchincloss modified radical mastectomy:*** Here pectoralis minor muscle is left intact and level III lymph nodes are not removed—*commonly done now*.

Mastectomy specimen should be carefully inspected and sent to pathology department

- Specimen is sent in formalin for histology.
- It is sent in saline in low temperature for ER/PR/Her 2 neu status study (histochemistry).
- Specimen mammography is often done.
- Tumour grading, tumour clearance, nodal involvement—its number and capsular breach are assessed histopathologially.

He who is afraid of doing too much always does too little.

- ♦ ***Halsted Radical Mastectomy**** (*Complete Halsted*) (***RM***):
  Structures removed are:
  - Tumour.
  - Entire breast, nipple, areola, skin over the tumour with margin.
  - Pectoralis major and minor muscles.
  - Fat, fascia, lymph nodes of axilla.
  - Few digitations of serratus anterior.

Structures retained are *(ABC)*

- ❖ **A**xillary vein
- ❖ **B**ells nerve (nerve to serratus anterior)
- ❖ **C**ephalic vein

Complications are lymphoedema and eventual lymphangiosarcoma (after 3 to many years later) of the limb.

- ♦ ***Conservative breast surgeries:***
  - **a. Wide local excision:** It is removal of unicentric tumour with 1 cm clearance margin. Incision is made directly over the tumour. Skin flaps should not be raised. Normal breast tissue of *1 cm clearance* with excision of tumour is done. Pectoral fascia is usually not opened in wide local excision unlike in total mastectomy. The specimen is marked after placing in orientation grid and mammography of the specimen is done followed by frozen section biopsy to look for clearance. At least 1 mm clearance is needed for adequacy. Margins where clearance is less than 1 mm need re-excision at that particular margin. If margins show no clearance then patient requires probably total mastectomy. So prior consent for mastectomy should be taken. Drain is not placed; deeper cavity is usually not closed/obliterated as small seroma gets absorbed without any problem. Skin is closed cosmetically. It is often called as *lumpectomy or partial mastectomy*; but better term is *wide local excision*. It is ideal breast conservative surgery. Along with this, axillary dissection through separate incision and RT to breast and chest wall area is given.
  - **b. Quadrantectomy:** It is removal of entire segment/quadrant with ductal system with 2-3 cm normal breast tissue clearance. It is not advocated now as it is proved that there is no outcome benefit between wide local excision and quadrantectomy. Quadrantectomy is done as a part of *QUART therapy* (by Umberto Veronesi from Milan, Italy) along with axillary dissection (level I and II) through separate incision and RT to breast area.

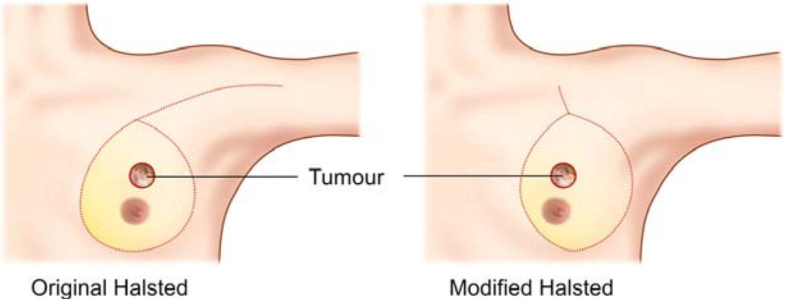

Fig. 8.80: Note the original Halsted and modified Halsted incision.

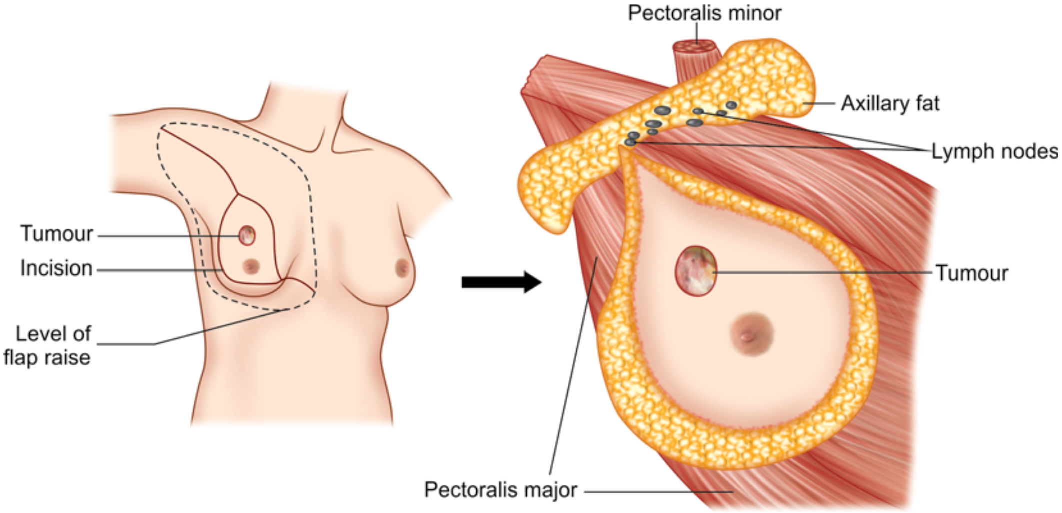

Fig. 8.81: Complete Halsted operation—radical mastectomy with removal of fat, lymph nodes, pectoralis major, pectoralis minor, entire breast with tumour and skin over the breast, nipple and areola.

***Not commonly done at present.** If at all, it is done in early stage I carcinoma breast.

- ***Toilet mastectomy*:**
  - In locally advanced tumour, tumour with breast tissue and whatever possible is removed to prevent further fungation. But its use and significance is under question. It is often done after giving chemotherapy.
- ***Extended radical mastectomies*:**
  - It includes radical mastectomy + removal of internal mammary lymph nodes of same side with or without opposite side. **It is not done at present.**
- ***Skin sparing mastectomy*** (SSM/Key hole mastectomy) is becoming popular with different approaches.
- ***Lumpectomy/partial mastectomy is better called as wide local excision*** which includes removal of tumour with gross rim (1 cm) of normal breast tissue.
  - Small (< 0.5 cm), unicentric, low grade tumour of breast may be treated with lumpectomy with RT. But it is technically difficult to assess such indications. Stereotactic wire localisation, lumpectomy with clearance, specimen mammography and proper follow-up are the needed principles in such situations. One should achieve negative margin of 1 mm on frozen section and with post-excision mammography of retained breast to confirm adequacy of excision. Presently term 'lumpectomy' is ***rarely*** used to such procedure which basically is wide excision with 1 cm clearance. Patient always needs postoperative adjuvant RT and chemotherapy if poor prognostic factors are present.

Complications of MRM/mastectomy

- Injury/thrombosis of axillary vein
- Seroma—50-70%
- Shoulder dysfunction 10%
- Pain (30%) and numbness (70%)
- Flap necrosis/infection
- Lymphoedema (15%) and its problems
- Axillary hyperaesthesia (0.5-1%)
- Winged scapula
- Occasionally if on table injury occurs to axillary vein, it should be repaired by vascular suturing using 5 zero polypropylene
- Numbness over the medial upper part of the arm can occur due to intercostobrachial nerve injury
- Pectoral muscles atrophy if medial and lateral pectoral nerves are injured
- Weakening of internal rotation and abduction of shoulder occurs due to injury to thoracodorsal nerve

Lymphangiosarcoma (Stewart-Treve's syndrome) of upper limb can develop in patients who have developed lymphoedema after mastectomy with axillary clearance. Usually it occurs *3-5 years* after development of lymphoedema. Such patient may require fore-quarter amputation. It has got poor prognosis. It presents as multiple subcutaneous nodules.

B. Radiotherapy in Carcinoma Breast

Indications:

- Patient who undergo conservative breast surgery, breast is irradiated after surgery.

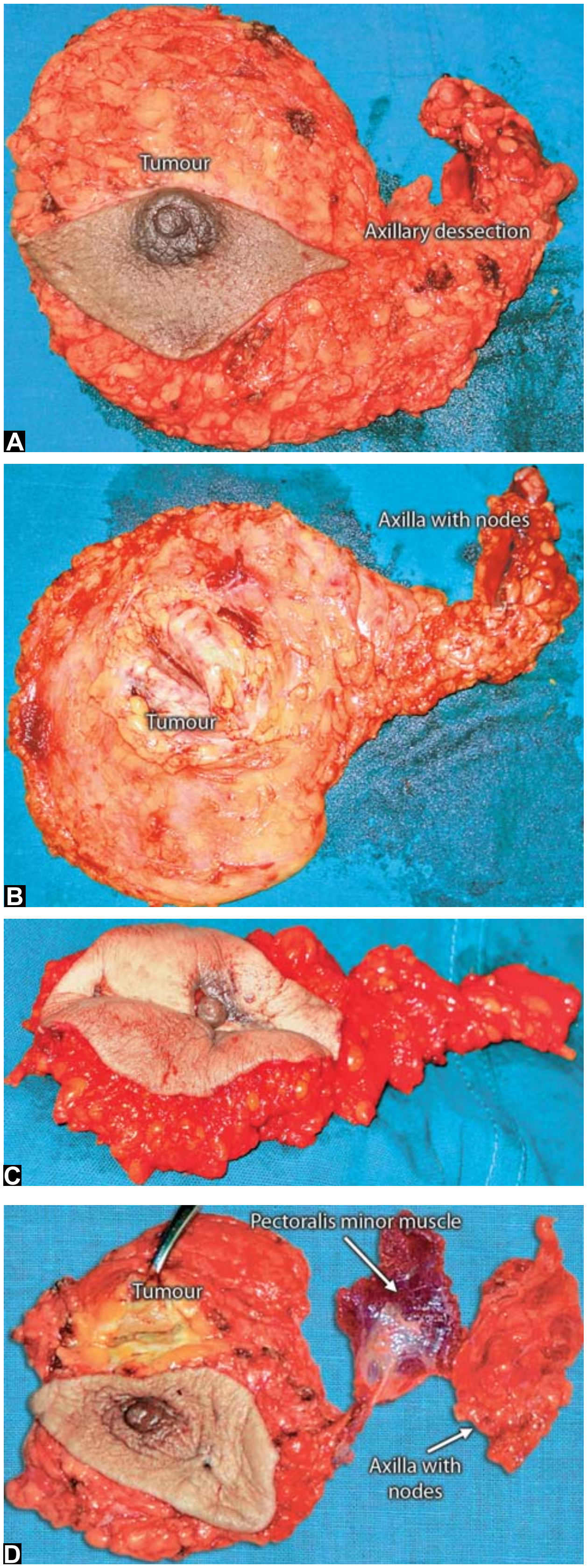

Figs 8.82A to D: (A to C) Mastectomy specimen with dissected axillary nodes. (D) Organ specimens removed during Patey's operation. Note the pectoralis minor, axillary area and tumour with breast tissue.

What lies within us is more important than what lies before or behind us.

Radiotherapy in carcinoma breast

To chest wall:	*To axilla:*	• *RT is a must* after conservation of breast
• T3 tumour > 5 cm	• 4 or more nodes positive	
• Residual disease-LABC	• Extranodal spread	• Local as well as to axilla
• Positive margin/close surgical margin of < 2 cm	• Axillary status not known/not assessed	• Tangential fields 50 Gy/25 fractions/5 weeks
• After conservative surgery		• Another 10 Gy to tumour bed
• Higher risk group • Inflammatory carcinoma		• Internal mammary and supraclavicular area may be included in radiation field

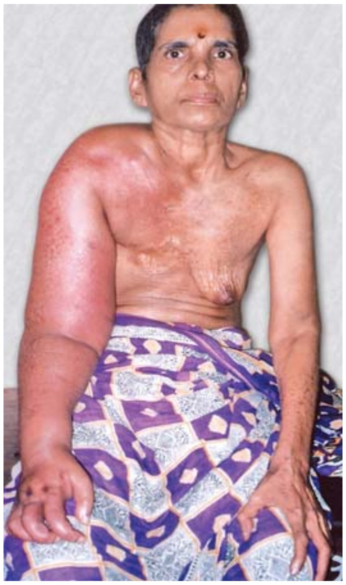

Fig. 8.83: Postmastectomy lymphoedema. Note the mastectomy scar.

- After total mastectomy, external irradiation is given to axilla.
- Patients with higher risk of local relapse after surgery:
  - a. Invasive carcinoma.
  - b. Extensive *in situ* carcinoma.
  - c. Patients under 35 years.
  - d. With multifocal disease.
- In bone secondaries, to palliate pain and swelling. If there is pathological fracture in the bone, internal fixation has to be done along with external irradiation.
- Inflammatory carcinoma of breast.
- In atrophic scirrhous carcinoma of breast, as a curative radiotherapy.
- As preoperative radiotherapy, to reduce the tumour size and downstage the tumour, so that the operability is better.
- More than 4 positive lymph nodes in the axilla, pectoral fascia involvement, positive surgical margins, extranodal spread or in patients with axillary status not known/not assessed.

External radiotherapy is given over the breast area, axilla (in selected patients like if axillary dissection is not done or more than 4 positive axillary nodes), internal mammary and supraclavicular area

- Total dosage 5000 cGY units
- 200-cGY units daily 5 days a week for 6 weeks

C. Hormone Therapy in Carcinoma Breast

Principles:

- It is used in ER/PR positive patients in all age groups (earlier it is used in perimenopausal age groups).
- It is relatively safe, easy to administer.
- It gives prophylaxis against carcinoma of opposite breast.
- It is useful in metastatic breast carcinoma.
- Hormone therapy (tamoxifen) is useful in breast cancer in elderly (positive ER) after wide local excision or occasionally as tamoxifen alone.
- It is now not used in ER negative patients.
- Menopausal status; nodal status; chemotherapy used are not factors to defer the use of hormone therapy.
- Hormone therapy reduces the recurrence rate and so probably improves the life span and quality of life.

Includes:

- Oestrogen receptor antagonists—*tamoxifen*.
- Ovarian ablation by surgery (Bilateral oophorectomy) or by radiation.
- LHRH agonists (Medical oophorectomy).
- Oral aromatase inhibitors for postmenopausal women.
- Adrenalectomy or pituitary ablation.
- Progesterone receptor antagonist.
- Androgens—Inj testosterone propionate 100 mg IM three times a week.
- Aminoglutethimide—blocks the synthesis of steroids by inhibiting conversion of cholesterol to pregnenolone—*medical adrenalectomy*.
- Progestogens, e.g. medroxyprogesterone acetate.

Tamoxifen

- It is an antioestrogen. It blocks cytosolic oestrogen receptors.
- Dose is 10 mg BID or 20 mg OD for 5 years.
- Half life of tamoxifen is 7 days; it takes 4 weeks to show its benefits. It reduces the cholesterol and also cardiovascular morbidity.

Adverse effects:

- *Tamoxifen flare*—flushing, tachycardia, sweating, genital itching, vaginal atrophy and dryness (premenopausal), vaginal discharge (postmenopausal), fluid retention, weight gain.
- Occasionally it causes bone pain which should be differentiated from pain due to bone metastasis. It is due to loss of bone density in premenopausal women.
- It increases the incidence of endometrial cancer.
- DVT (3%), pulmonary embolism, CVA, TIA, cataract, fractures.

Side effects and endometrial cancer are less in selective drugs like raloxifene.

Advantages:

- It reduces the recurrence rate by 25%.
- It improves the prognosis.
- It is used presently in all age group, ER +ve patients.

Drugs used in hormone therapy

Drug	*Mechanism of action*	*Dose*	*Adverse effects*
Tamoxifen	Antioestrogen	20 mg	Nausea, weight gain, hot flushes, vaginal bleeding, bone pain, amenorrhoea
Medroxyprogesterone	Progestogen	400 mg	Nausea, flushing, vaginal bleeding
Aminoglutethimide	Aromatase inhibitor	250 mg QID	Rash, dizziness, lethargy, Cushing facies
Arimidex	Aromatase inhibitors	1 mg	Lethargy, GI upset
Letrozole	Aromatase inhibitor	2.5 mg OD	Vaginal dryness, hot flushes,vaginal bleeding, cardiovascular problem
Zoladex (Goserelin)	LHRH agonist	3.6 mg monthly	Amenorrhoea, nausea. It is expensive
Diethylstilbestrol	Oestrogen	15 mg daily	Fluid retention, vomiting, thrombosis, hypercalcaemia
Fluoxymestrone	Androgen	30 mg daily	Masculinization, nausea
It is more beneficial in ER +ve patient; perimenopausal patient with node +ve; postmenopausal patient. Presently ***Tamoxifen and Oophorectomy*** are the commonly used methods.			

- Cheap, easily available, less toxic effects, very effective.
- It is equally effective in carcinoma male breast.
- It is presently also **under trial** for certain benign diseases of breast (*ANDI, cyclical mastalgia*).

Selective oestrogen antagonists

- Do not cause endometrial hyperplasia or endometrial carcinoma.
- Drugs include droloxifen, toremifen, raloxifene.

Tamoxifen used in:

- Carcinoma breast
- Benign diseases of the breast like fibroadenosis
- Infertility in males
- Desmoid tumour
- Risk reduction therapy in high-risk females in premenopausal age

Letrozole

- It is a nonsteroidal competitive inhibitor of the enzyme 'aromatase'. This enzyme converts adrenal androgens to oestrogen (aromatization). So it is an aromatase inhibitor.
- Other aromatase inhibitors are *anastrozole and exemestane* (in postmenopausal). It is expensive but more effective than tamoxifen. It is also used in recurrent disease.
- Letrozole is used as an adjuvant endocrine therapy in post-menopausal women with hormone sensitive breast cancer (In premenopausal women this will cause rise in gonadotrophins and ovarian aromatase is not well suppressed). It can also be used in metastatic and recurrent cases. It slows down and stops the growth of oestrogen sensitive breast tumours. It reduces oestrogen level by 98%. Its half life is 45 hours. It decreases the bone density.
- Dosage of letrozole is 2.5 mg once daily.
- It is given for 5 years or for 2 years following 3 years of tamoxifen.
- Side effects of letrozole are vaginal dryness, night sweats, hot flushes, vaginal bleeding, cardiovascular problems and osteoporosis.

Transtuzumab (Herceptin)

- It is a monoclonal antibody that blocks HER-2/Neu receptors thereby preventing growth of cancer cells. It is a new drug. It is presently marketed as herceptin. It is c-ErbB2 (growth factor receptor) inhibitor. It is a newer biological agent. Her 2/Neu receptor is tyrosine kinase receptor.
- It has very little effect on HER-2/Neu negative cancers.
- It is useful only in HER-2/Neu positive cancers. It is currently approved by FDA for use only in metastatic disease. It is given as intravenous infusion.
- Studies have shown substantiate improvement in disease free and overall survival rate.
- Dose is 4 mg/kg as loading; 2 mg/kg as maintenance for 1 year.
- It has got cardiac side effects.

Note

- *Tamoxifen* interferes with oestrogen receptors
- *Letrozole* interferes with oestrogen production
- *Transtuzumab* (Herceptin) interferes with HER-2 Neu receptors
- *Bevacizumab*—vascular growth factor receptor inhibitor—newer agent
- *Lapitinab*—combined growth factor receptor inhibitor—newer agent

Hormone therapy for carcinoma breast

In premenopausal women

- Tamoxifen—antioestrogen
- Ovarian ablation by surgery/by Goserelin/by radiation
- Progestogens—medroxyprogesterone 400 mg
- Androgens—fluoxymestrone

In postmenopausal women

- Tamoxifen
- Aromatase inhibitor like letrozole 2.5 mg OD
- Progestogens
- Androgens
- Medical adrenalectomy using aminoglutethimide (Mitotane)—as major source of oestrogen after menopause is adrenal gland. Cortisone supplement is also needed to prevent feedback rise of ACTH which may block the effect of aminoglutethimide

Surgical endocrine ablations

- Bilateral oophorectomy (Beatson—1896) in premenopausal age
- Bilateral adrenalectomy
- Pituitary ablation

Note:

Present changing concept—is letrozole and tamoxifen are continued even after five years (probably life long) to reduce the chance of late recurrence.

D. Chemotherapy in Carcinoma Breast

- *Adjuvant chemotherapy* refers to administration of cytotoxic drugs to women after breast cancer surgery in the hope of eliminating clinically undetectable distant spread. Reduced recurrence rate and improved survival rate is observed in all women with *invasive* breast cancer. Chemotherapy is usually not indicated in DCIS.
- *Neoadjuvant chemotherapy* refers to administration of cytotoxic drugs in a large operable tumour to reduce the loco-regional tumour burden to make it better amenable for surgical resection, usually after 3 doses. It downstages the disease; may achieve breast conservative surgery in *selected patient*s (only); early systemic control is achieved.
- *Palliative chemotherapy* is used in advanced and metastatic carcinoma breast.

Indications:

- All node positive patients.
- Primary tumour more than 1 cm in size.
- Presence of poor prognostic signs of any tumour—vascular and lymphatic invasion; high nuclear and histologic grade; Her 2/Neu overexpression; negative hormone receptor status.
- In advanced carcinoma breast as a palliative procedure.
- In postoperative period after simple mastectomy in stage III carcinoma breast with fixed axillary nodes.
- In inflammatory carcinoma of breast.
- In stage IV carcinoma breast with secondaries in bone, lungs, liver.
- In premenopausal age group with poorly differentiated tumours.

Drugs used

CMF regime	*CAF regime*	*MMM regime*
Cyclophosphamide	Cyclophosphamide	Methotrexate
Methotrexate	Adriamycin	Mitomycin-C
5-Fluorouracil	5-Fluorouracil	Mitozantrone

Toxic effects are: Alopecia, bone marrow suppression, cystitis, megaloblastic anaemia, GIT disturbances, nephritis.

- ***CMF and CAF*** are commonly used with monthly/3 weeks cycles for 6 months.
- Other anthracyclines like doxorubicin or epirubicin should be used often for better result.
- ***Taxanes***: They are newer chemotherapeutic drugs which act by G2/M phase of cell cycle. It is commonly used in metastatic carcinoma of breast. Drugs are **paclitaxel and docetaxel**. Taxanes have no cross-resistance with anthracyclines and so can be used sequentially or concurrently with anthracyclines.
- Gemcitabine is also used often in selected cases for better results.

Chemotherapy regimes

- *1st line drugs*—anthracyclines—FEC regimen (better); CAF; CMF
- *2nd line drugs*—taxanes
- *3rd line*—gemcitabine

***Note*:**

- Neoadjuvant (primary) chemotherapy is useful in large operable primary tumour to make it amenable for conservative breast surgeries. Postoperative radiotherapy and chemotherapy is a must in this situation.
- High dose chemotherapy with stem cell rescue is not found to be beneficial by trials in patients with heavy large nodal spread.
- Hormone therapy is given after completion of chemotherapy for additive effect. Concurrent therapy of both (together) is not used at present. It is sequential—chemotherapy then hormone therapy. Chemotherapy gives quicker and better response initially and so it is used early and later only hormone therapy is continued.
- Adriamycin shows better response with Her-2 Neu overexpression patients.
- Paclitaxel may be combined with trastuzumab in Her-2 Neu (Human epidermal growth receptor 2) over expression patients.
- Chemotherapy and radiotherapy can be given concurrently or as sandwich therapy.
- Presently one day dose of all drugs of the regimen used as a *standard* at 3 weekly cycles for 6 cycles for stage III—CAF regime. For metastatic paclitaxel single day dose cycles or etoposide and cisplatin 3 days cycle—6 cycles.

Complications of chemotherapy—bone marrow suppression, alopecia, GI side effects, cardiac toxicity. Monitoring with haematocrit, blood urea, serum creatinine, LFT are needed. Cytoprotective agents like GMCF (granulocyte macrophage colony stimulating factor), GCSF (granulocyte colony stimulating factor) are given to reduce severity and duration of neutropenia. It is given in 24 hours as 30 million units.

MANAGEMENT OF EARLY CARCINOMA BREAST

- Early breast cancer is the one diagnosed by mammography or stage I carcinoma breast.
- Usually chest X-ray, blood count and liver functions are sufficient to screen the early breast cancer patients.

Aims of Treatment

- To achieve possible '*cure*'.
- Control of local disease in breast and axilla.
- Breast conservation, i.e. breast form and function.
- Prevention of distant metastasis.
- To prevent local recurrence.

Breast conservative surgery

Indications	*Contraindications*
• Lump < 4 cm	• Tumour > 4 cm
• Clinically negative axillary nodes	• Positive axillary nodes > N1
• Mammographically detected lesion	• Tumour margin is not free of tumour after breast conservative surgery needs MRM
• Well-differentiated tumour with low S phase	
• Adequate sized breast to allow proper RT to breast	• Poorly differentiated tumour
• Breast of adequate size and volume	• Multicentric tumour
• Feasibility of axillary dissection and radiotherapy to intact breast	• Earlier breast irradiation
	• Tumour/breast size ratio is more (centeral tumour)
	• Tumour beneath the nipple
	• Extensive intraductal carcinoma

Modalities of Treatment

- ***Breast conservative surgery (BCS)***—ideally done as *wide local excision* with axillary dissection with RT to breast and chest wall. Quadrantectomy as a part of QUART therapy may be used only in selected patients. RT is given to the entire breast with 4500 cGy dose.
- *Patey's operation* or simple mastectomy with axillary clearance.
- Postoperative radiotherapy in high-risk patients.
- Hormone therapy, i.e. tamoxifen 10 mg BID or 20 mg OD.
- Sentinel node biopsy when required.
- Regular follow-up often with radioisotope bone scan and CEA tumour marker.

In early breast cancer, breast conservative surgery like *wide local excision*/quadrantectomy, axillary dissection (level I and II) and postoperative radiotherapy (to the breast) is used which prevents the disfigurement and psychological trauma of mastectomy to the patient. Tumour is removed with a rim of normal tissue. Wide excision and QUART therapy are different procedures. *Wide local excision* is ideal and better where clearance margin is 1 cm. In *quadrantectomy* entire segment with ductal system with 2-3 cm clearance margin is achieved. But it is not advocated now as there is no benefit in outcome (survival/recurrence rate) by quadrantectomy over wide local excision.

Principles of Conservative Breast Surgery (Refer Table Above)

- Curvilinear nonradial incisions (radial incisions should not be placed, because if there is a need to convert into total mastectomy, then incision plan may be difficult).
- Separate incision for axillary dissection.
- Undermining of the skin flap must be avoided.
- Confirm tumour clearance by frozen section. It may be often difficult and so tumour is cut and only margin which is close and doubtful is advocated for frozen section.
- Radiotherapy is a must to breast and chest wall region (locally).

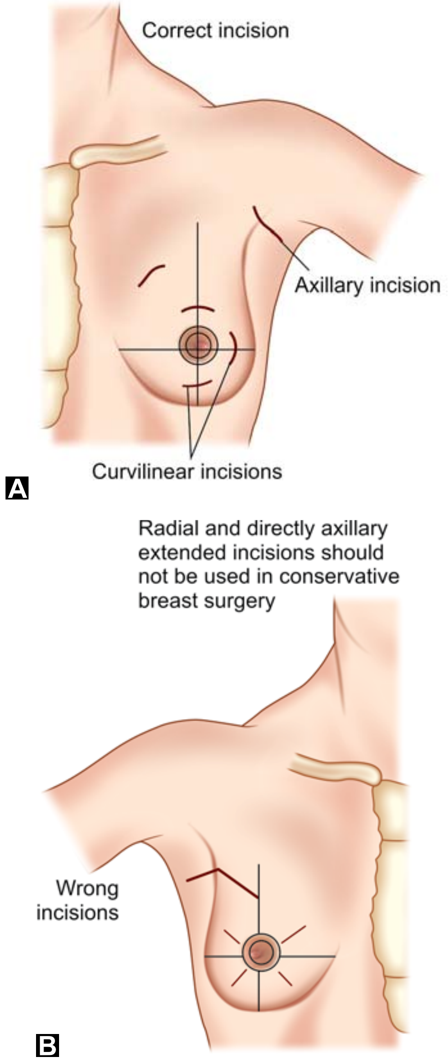
Correct incision
Axillary incision
Curvilinear incisions
A
Radial and directly axillary extended incisions should not be used in conservative breast surgery
Wrong incisions
B

Figs 8.84A and B: Curvilinear incision should be placed in conservative breast surgeries. Never place wrong radial incisions because if conversion to total mastectomy is needed then placement of incision will be difficult. In conservative breast surgery for axillary dissection separate incision in the axilla should be placed.

In time of test, family is best.

QUART Therapy

It is *quadrantectomy, axillary dissection* of level I and II nodes with separate axillary incision and *postoperative radiotherapy* to breast (5000 cGy) and axilla (1000 cGy). First it was started by *Umberto Veronesi* from Milan. Now Quart therapy is only occasionally used.

Skin Sparing Mastectomy (SSM)

- It is like a key-hole surgery of breast.
- Skin sparing/limited skin excision (5-10%) will not alter/affect the recurrence rate.
- *Indications are*—central tumour/multicentral/extensive intraductal/T1/not feasible for conservation.
- Excision of nipple-areola complex with very limited skin removal.
- Marginal skin excision over the tumour/biopsy site.
- Total glandular mastectomy.
- Axillary dissection using either same (extension of SSM incision) or separate incision in the axilla.

Indications for Total Mastectomy in Early Breast Cancer

- When tumour is more than 4 cm.
- Multicentric tumour.
- Poorly differentiated tumour—high grade.
- Tumour margin is not clear of tumour after breast conservative surgery.

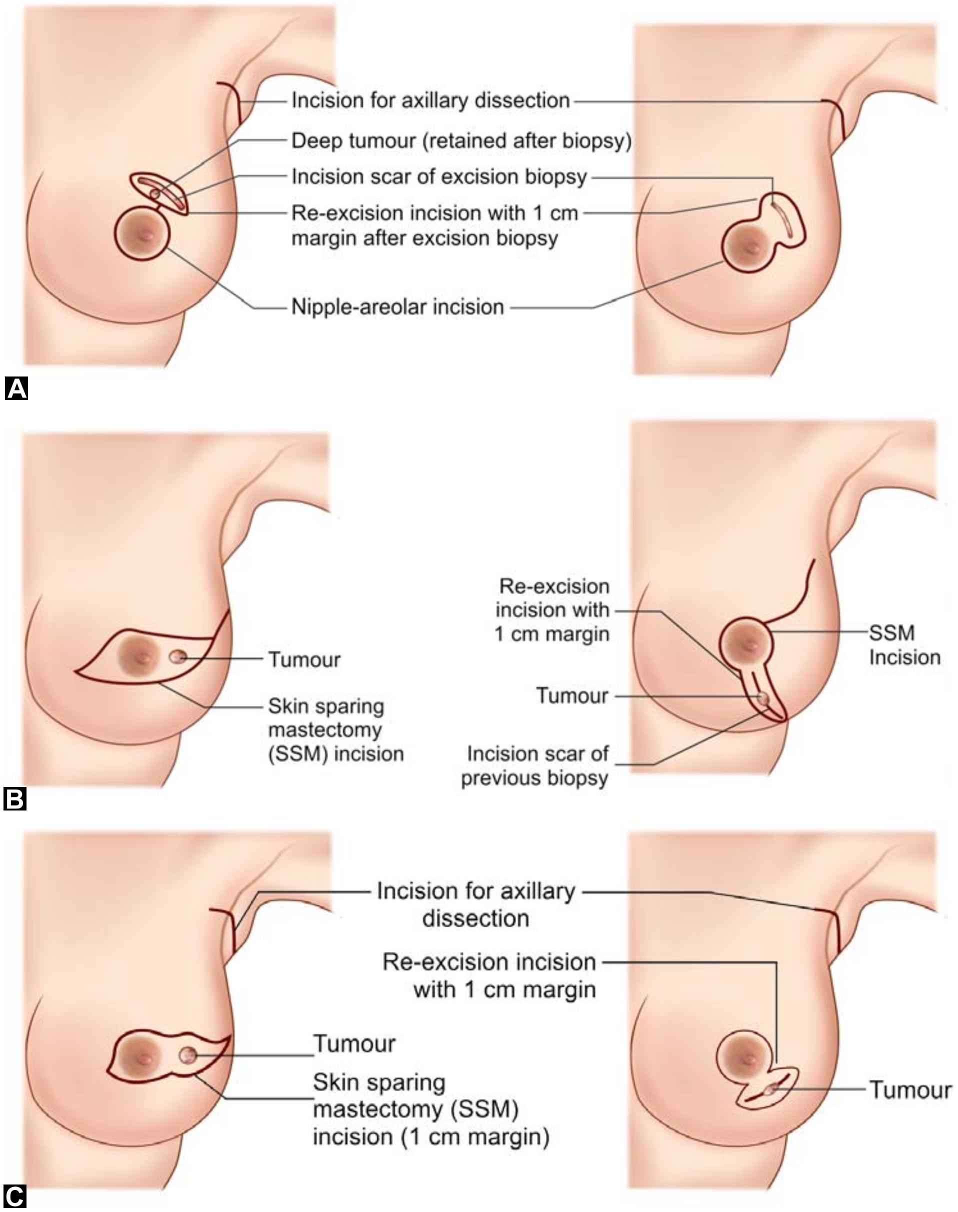
Incision for axillary dissection
Deep tumour (retained after biopsy)
Incision scar of excision biopsy
Re-excision incision with 1 cm margin after excision biopsy
Nipple-areolar incision
A
Tumour
Skin sparing mastectomy (SSM) incision
Re-excision incision with 1 cm margin
Tumour
Incision scar of previous biopsy
SSM Incision
B
Incision for axillary dissection
Re-excision incision with 1 cm margin
Tumour
Skin sparing mastectomy (SSM) incision (1 cm margin)
Tumour
C

Figs 8.85A to C: Skin sparing mastectomy for carcinoma breast—different approaches. Skin sparing mastectomy (SSM) does not affect the recurrence rate.

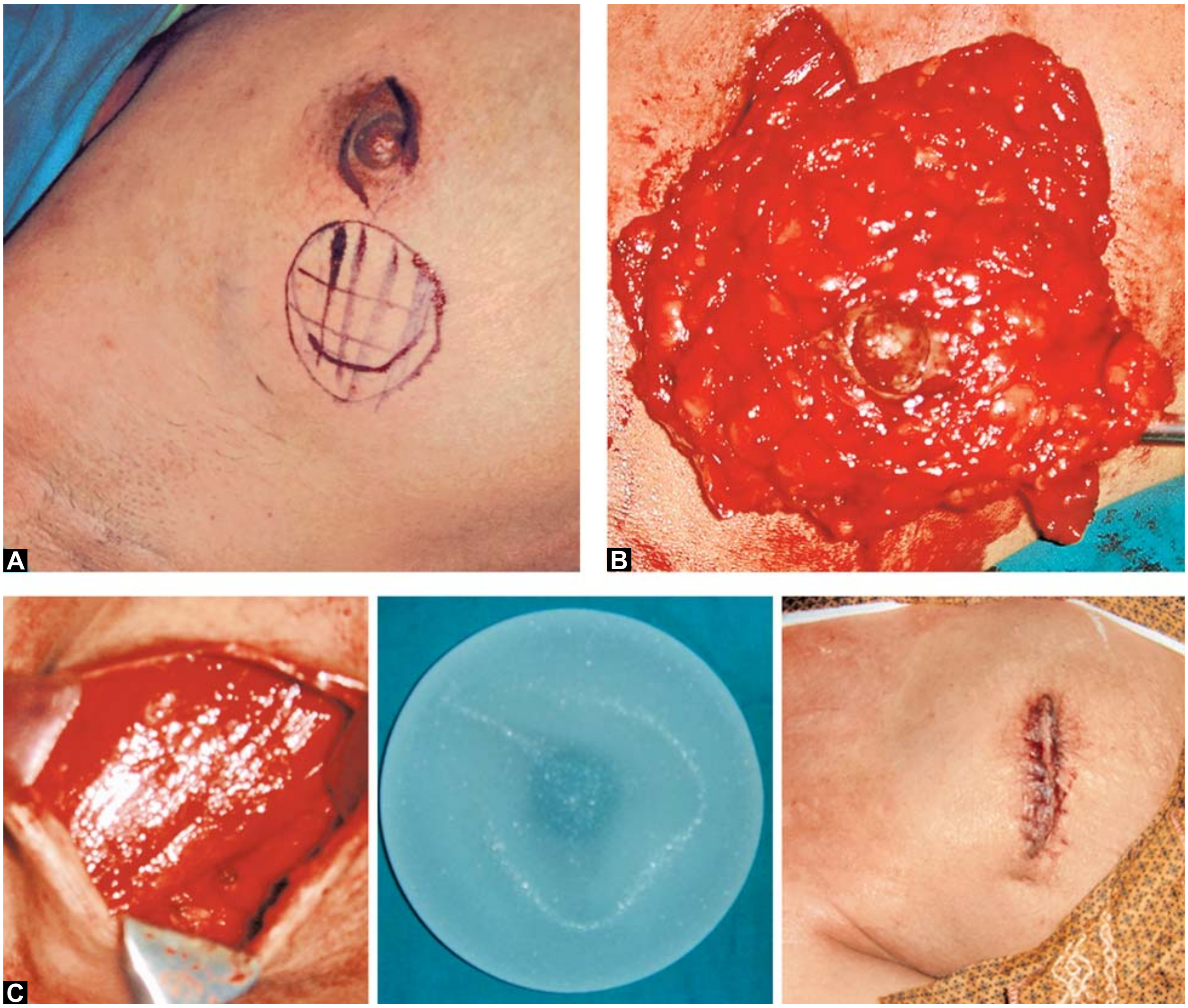

Figs 8.86A to C: Skin sparing mastectomy incision with prosthesis implant for placement.

Management of Axillary Nodes when Clinically Not Palpable

- *Sentinel lymph node biopsy (SLNB)* is done. If node is positive for tumour then axillary dissection is done. But facility for SLNB is not available in most of the centres and so axillary dissection is done as a routine. Usually level I and II nodes—below the axillary vein are dissected.
- *Axillary sampling* is done by separate curved incision between the outer border of pectoralis major and latissimus dorsi 6 cm below the apex of axilla. About 10-15 nodes (level I) are sampled. Aim is to remove *largest nodes* in axilla which are likely to be involved. At least *4 lymph nodes* should be removed. All 4 nodes should be sent *separately* for histology. First level I nodes are removed; if they are not palpable then only level II or III are sampled. Axillary sampling is said to detect *skip metastasis* in level II and III. It is not commonly advocated now.

Principles of Axillary Dissection/ Axillary Clearance

- *Indications*: Axillary dissection is done when there is clinical involvement of nodes; when FNAC of lymph node or sentinel node biopsy proved malignancy; when in a large tumour where reconstruction is needed so that 2nd eventual axillary surgery is difficult at a later period.
- It is *removal* of fat, fascia, nodes in the axilla. Usually level I and II nodes are removed. Nodes below the axillary vein are only removed. Level III dissection increases the chances of lymphoedema. Risk of lymphoedema further rises if RT is given later along with level III dissection.
- Dissection is done through an individual crease transverse incision in the axilla, when it is advocated with breast conservative surgery.
- When it is advocated along with total mastectomy it is done like modified radical mastectomy—*Auchincloss* is commonly used.

Hope sees the invisible, feels the intangible and achieves the impossible.

Principles of axillary dissection/axillary clearance

It is removal of different levels of axillary nodes; with fat and axillary fascia

Why axillary dissection is done?	***Which levels?***	***Technical principles***
• For staging	• Level I—60%	• Any incision but caudal hair-line incision is preferred
• To assess the prognosis—number of nodes/size of the node	• Level I, II—20-25% • Level I, II, III—15-20%	• Nerve to serratus anterior/thoracodorsal nerve should be safeguarded
• As a treatment—regional control of		• Medial and lateral pectoral nerves should be retained when the dissection is done with the mastectomy
• To plan adjuvant therapy irradiation/chemo/hormone	• Level I and II dissection Low axillary dissection less chances of lymphoedema	• Drain should be kept to the area • After surgery specimen should be properly dissected for all lymph nodes, labelled and sent for histological study. Usual number of nodes in axilla are 20-32. So minimum of 10 nodes should be dissected, identified and sent for study. Nodes can be easily identified if fat dissolving agents are used over them.

- In *skin sparing mastectomy* also axillary dissection is done using an extension of SSM incision or by separate axillary incision.

Adjuvant therapy after surgery in early breast cancer

- Radiotherapy
- Chemotherapy—CMF, CAF regime commonly used. Taxanes are also used
- Endocrine manipulation:
  - Ablation
  - Tamoxifen (receptor antagonist)—20 mg/day for 5 years
  - Aromatase inhibitors—blocks oestrogen production. Letrozole 2.5 mg OD
  - LHRH agonists—Goserelin 3.6 mg/28 day's cycle for 2 years

ADVANCED CARCINOMA BREAST

Refers to:

- Locally inoperable (adherent to chest wall) tumour.
- Inflammatory carcinoma of breast.
- Fixed axillary lymph nodes, or supraclavicular lymph nodes, or opposite axillary nodes.
- Bilateral carcinoma breast.
- Metastatic carcinoma of breast, i.e. spread to bones, liver, lungs, brain.

Locally Advanced Carcinoma of Breast (LABC)

- It means locally advanced tumour with muscle/chest wall involvement, extensive skin involvement or fixed axillary nodes. It will be T3, T4a, T4b, T4c or T4d or N2 or N3.
- It is investigated by FNAC of tumour/core needle biopsy/incision biopsy/mammography of opposite breast, chest CT, CT abdomen or whole body bone scan. Biopsy is needed to assess receptor status.
- *Bilateral mammography* is done to assess tumour size and multicentricity which is needed to check the chemotherapy response at a later period. FNAC of axillary node is required. If bone scan is positive it becomes metastatic carcinoma of breast not LACB. Only when 60% of bone is dimeneralised in metastatic bone disease, plain X-ray bone will detect the lesion. Even though X-ray may be normal in stage III disease, in 30% of these patients bone scan will be positive making the disease stage IV metastatic.
- Treatment of LACB is always palliative by simple mastectomy/toilet mastectomy, chemotherapy and hormone therapy using tamoxifen.
- Initial neoadjuvant chemotherapy; surgery; radiotherapy; adjuvant chemotherapy are other therapeutic plan.
- Palliation is to control pain, to prevent fungation or bleeding.
- In inoperable fixed tumour chemotherapy is given initially. Later, after 3-4 cycles of chemotherapy, when tumour size reduces and becomes operable, total mastectomy is done.
- Postoperative radiotherapy is given to breast field and axilla.
- Usually axillary dissection is not necessary in LACB.
- Only chemotherapy and radiotherapy to breast and axilla (without palliative mastectomy) also can be done in LACB.
- There is usually no role of breast conservative surgery for LACB.
- 5-year survival is 40% and 10-year survival is less than 25%.

Present strategy of treatment for LABC

- *Initial neoadjuvant* (anterior) chemotherapy is given to downstage and achieve cytoreduction, to target possible micrometastases first, to assess chemosensitivity. *FEC*, CMF, CAF regimes are used.
- *Response to chemotherapy* is assessed by *complete responders* without palpable tumour; *partial responders* with > 50% decrease in tumour size; *nonresponders* with < 50% decrease in tumour size; *progressive disease* with > 25% increase in size
- Patients with nonresponders and progressive disease are treated by RT to breast, chest wall, axilla and supraclavicular region; taxanes; hormone therapy; surgery if operable
- *Responders* are later treated with total mastectomy/MRM, occasionally BCS. After surgery remaining 3 or 4 cycles of chemotherapy are completed. Later hormone therapy should be given for 5 years (tamoxifen 20 mg OD). ER +ve (> 10 fmol/mg) patients show 75% positivity
- Stage IIIA patients are *classified* as operable and inoperable

Inflammatory Carcinoma of Breast

Inflammatory carcinoma is T4d LACB (Stage IIIB). It is also called as mastitis carcinomatosis or lactating carcinoma of breast. It is 2% common. It is observed in younger age group usually in pregnancy or lactating period. There will be extensive skin involvement with pain. It often mimics mastitis of lactation. FNAC or incision biopsy concludes diagnosis. It is initially treated by chemotherapy or radiotherapy and later if tumour reduces in size total mastectomy with axillary clearance can be done. But most often it is inoperable. After surgery, chemotherapy and tamoxifen is given. 5-year survival for inflammatory carcinoma of breast is 25-30%.

Metastatic Carcinoma of Breast

It is blood spread into different places like bone, lungs and pleura, liver, soft tissues, brain and adrenals. It is evaluated by FNAC/incision biopsy, chest CT, LFT, U/S abdomen, CT abdomen, whole body bone scanning, CT brain, tissue study for ER/PR/HER-2 Neu receptor status. It is *stage IV* disease.

- *Bone* is the most common site of metastasis. Spread to vertebra is through posterior intercostal vein and *Batson's venous plexus (valveless)*. Vertebrae, ribs, upper end of humerus and femur are common bones involved.
- *Lungs and pleural* spread causes 'cannon ball' secondaries, effusion, consolidation, chest wall secondaries.
- *Liver* secondaries may develop either by haematogenous or through lymphatics across diaphragm from lower inner quadrant of breast.
- *Brain* secondaries present with headache, vomiting, convulsions, localising features.
- Soft tissue secondaries has got *better prognosis*; visceral secondaries has got worst prognosis.
- Response to treatment decreases with number of organs involved with secondaries.
- Receptor negative secondaries are more aggressive.
- Median survival time for metastatic breast cancer is 24 months.

Treatment concept in metastatic carcinoma of breast

- To improve quality of life.
- To relieve pain of secondaries like bone, lungs.
- To relieve neurological problems like convulsions, space occupying cranial problems.
- Other symptomatic relief.

Treatment strategy in metastatic carcinoma of breast

- *Chemotherapy*—CMF, CAF, Taxanes in combination. Chemotherapy in metastatic breast cancer is useful (indicated/commonly used) in rapidly spreading visceral and skin secondaries, lymphangitis or when predicted disease free survival is less than 2 years; if response is negative for first line hormone therapy; in receptor negative status.
  - High dose of chemotherapy using cyclophosphamide, cisplatin, carmustine, melphalan is tried in view to get high response rate of 55-70% along with bone marrow transplant. But toxic effects are often life-threatening. Its advantages are now doubtful when compared to toxic effects.
  - Haemopoietic growth factor is also used along with chemotherapy to enhance the cell kill with less bone marrow toxicity. It may also allow multiple high dose chemotherapy to increase the response rate.
- *Radiotherapy* is used in bone metastasis, brain secondaries, to prevent paraplegia in spine involvement, and advanced axillary nodes. It is also used in painful bone secondaries, chest wall secondaries.
- *Hormone therapy* has got important role. Tamoxifen, oophorectomy, adrenalectomy, androgens, progestogens, aminoglutethimide are used. It is useful in slow growing soft tissue or bone secondaries; or predicted disease-free survival is more than 2 years or age more than 35 years.
- Blockage of over expression of epidermal growth factor (EGF)/transforming growth factor alpha (TGF-alpha) which are related to ErbB1/ErbB2 receptors in relation to aggressive carcinoma factor is tried.
- Palliative surgeries done are total/toilet mastectomy, fixation of bones in case of pathological fractures, spinal cord decompression to prevent paraplegia, lung resection in case of localised secondaries, bilateral oophorectomy.
- Trastuzumab (herceptin) is monoclonal antibody used in cancers with good results. It blocks the Her-2/Neu and ErbB2 receptors. Bevacizumab and lapitinab are other newer biological agents of different actions.

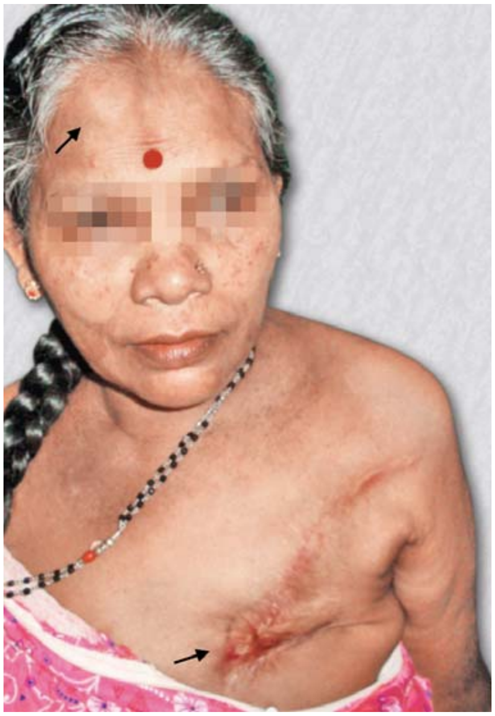

Fig. 8.87: Carcinoma of breast operated earlier with skull secondaries. It is metastatic carcinoma of breast. Skull secondaries are not common. Common site of bone secondaries are lumbar vertebrae, pelvis, upper part of femur, ribs, etc.

- Metabolic complications like *hypercalcaemia* may be life-threatening emergency in metastatic breast cancer. It is treated with hydration, steroids, biphosphonates (palmidronate, clodronate) 90 mg intravenously once a month. It also reduces the demineralisation of bone, fracture, pain and paraplegia.

Bone secondaries in carcinoma breast

- Most common site of blood spread (70%)
- Lumbar vertebrae, femur, pelvis
- Pathological fracture can occur
- Spinal compression and paraplegia
- Radiotherapy, internal fixation, spinal decompression is required
- Biphosphonates 1600 mg/day

Malignant pleural effusion as secondaries from Ca breast

- It signifies terminal event
- It has got poor prognosis
- Respiratory distress and failure is the main feature
- HRCT is ideal diagnostic tool
- *Treated by*
  - Intercostal tube drainage
  - Pleurodesis using talc/tetracycline (1.5 gm)/bleomycin (60 units)
  - Chemotherapy

Causes of death in carcinoma breast

- Secondaries in lung—haemoptysis, respiratory failure
- Spine involvement—quadriplegia
- Secondaries in brain
- Cancer cachexia

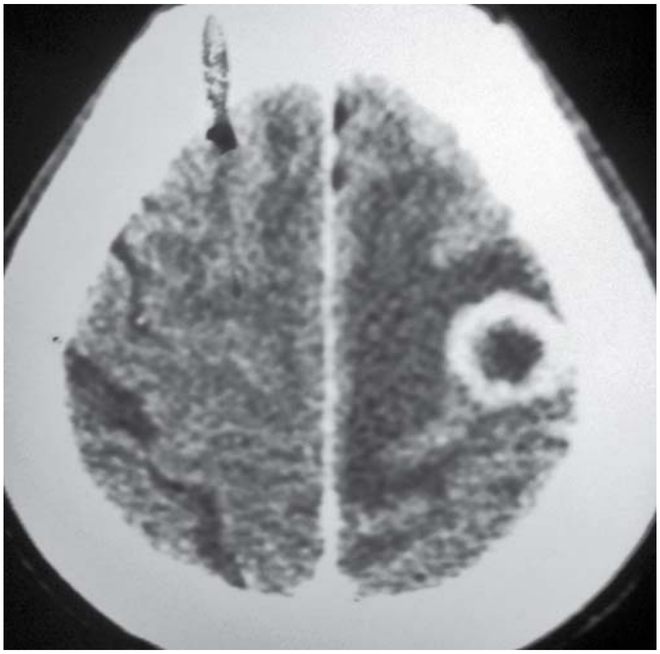

Fig. 8.88: Secondaries in brain—primary is from breast.

PROGNOSTIC FACTORS IN CARCINOMA BREAST

Stage	*5 yr survival*
I	90%
II	70%
III	40%
IV	20%

- **Spread to axillary nodes** is the most important prognostic indicator.

Lymph node—as a prognostic factor

- Number of nodes: > 2 carries poor prognosis
- Location of nodes
- Capsular invasion
- Extranodal status
- Size of the node: > 2.5 cm has poor prognosis
- More than 4 nodes/level III (apical nodes) involvement has got worst prognosis (5-year survival is 30%) and also decides for radiotherapy to axilla

- **Age:** Younger the age worser the prognosis.
- **Sex:** Carcinoma male breast has got worser prognosis compared to female breast. Because of early spread in carcinoma male breast.
- **Stage I and II** has got better prognosis.
- **Atrophic scirrhous** has got best prognosis.
- **Medullary carcinoma** has got better prognosis than scirrhous carcinoma because of lymphocytic infiltration.
- **Invasive carcinoma** has got worser prognosis.
- **Inflammatory carcinoma** breast has worst prognosis.
- **ER +ve** tumours has got better prognosis.
- **Differentiation** also decides prognosis.
- **Presence of elastic fibres** in histology has got better prognosis.
- Tumour proliferation stages, growth factor and oncogene factors.
- Tubular, colloid, papillary types has got better prognosis.
- **Tumour grade**, growth factor and oncogene factors. ErbB2—Her-2/Neu positive has got poor prognosis. ErbB1 with overexpression of epidermal growth factor (EGF), TGF alpha and cathepsin D has got poor prognosis.
- **DNA flow aneuploid** status has got poor prognosis. Low S phase fraction (< 5%) has got good prognosis.
- **Size of the tumour**—tumour size less than 1 cm has got better prognosis.
- **p53 tumour suppressor gene** (guardian gene) shows bad prognosis.

Prognostic factors are classified as:

- *Host factors:* Age; sex; menopausal status; family history; previous breast cancer; immunosuppression; nutrition.
- *Tumour factors:* Tumour size; nuclear grade; staging; receptor status; DNA nature; spread, etc.

Breast cancer group	*Features*	*Prognosis—5-year survival*
Very low-risk primary	DCIS; tubular	> 90%
Low-risk primary	Node negative; favourable microscopy	70-90%
High-risk primary	Node positive; unfavourable microscopy	< 70%
Locally advanced	Inflammatory ca/large primary (T4)	< 30%
Metastatic	Blood spread	Poor

PROPHYLACTIC MASTECTOMY

It is removal of *entire breast including axillary tail of Spence* prophylactically as risk reduction procedure in high-risk patients with indications. It reduces the developing of invasive cancer by 90% in high-risk group and also decreases the chances of death due to the disease. Axillary dissection is not done.

Modified Gail model and evaluation of mutation carriers are used to assess the risk group for risk reduction surgeries.

Indications

- DCIS (noninvasive).
- LCIS.
- Florid changes in fibrocystadenosis.
- Suspicious lesion in mammogram.
- Strong family history with 1st degree relatives having carcinoma breast.
- BRCA1 and BRCA2 mutation carriers.

Risk Reduction Surgeries

- Total mastectomy.
- Subcutaneous mastectomy.
- Skin sparing mastectomy.

Mastectomy is done bilaterally in LCIS and in strong family history. In suspicious unilateral lesion only unilateral mastectomy may be done.

***Note*:**
Other risk reduction methods are—life style modifications, avoiding alcohol and HRT, *tamoxifen*.

CARCINOMA OF MALE BREAST

- It is less than 1% of cases of breast cancers. Mean age is 65 years.
- Commonly associated with BRCA2 gene mutation.
- Other *risk factors* are—radiation, excess oestrogen, cirrhosis, Klinefelter's syndrome, liver schistosomiasis, family history, testicular disease, infertility, obesity.
- Commonly it is *infiltrating duct carcinoma* (90%), DCIS (10%). Lobular carcinoma is *not seen* in male breasts.
- Unilateral hard breast lump often eccentric with nipple retraction and discharge, skin dimpling and axillary lymphadenopathy are the clinical features.
- Differential diagnosis—gynaecomastia, breast abscess, sarcoma, secondaries.
- Presentation, spread, behaviour are same as carcinoma of female breast. Investigations and treatment are same as carcinoma female breast.
- MRM (as pectoralis major is involved early) with RT to breast area and axilla is the treatment.
- *Tamoxifen* is very useful in carcinoma male breast.
- Carcinoma male breast has poorer prognosis when compared to carcinoma breast in females.
- 80% cases are ER positive. 75% PR +ve; 35% overexpress HER 2/Neu.
- LHRH agonists are also often used. It works at hypothalamo-pituitary axis level via tachyphylaxis. It produces reversible "*chemical castration*". Goserelin 3.6 mg/28 days for 2 years is the current recommendation.
- Bilateral orchidectomy which was earlier used is no longer commonly advocated. Orchidectomy may be only useful in metastatic male breast cancer.

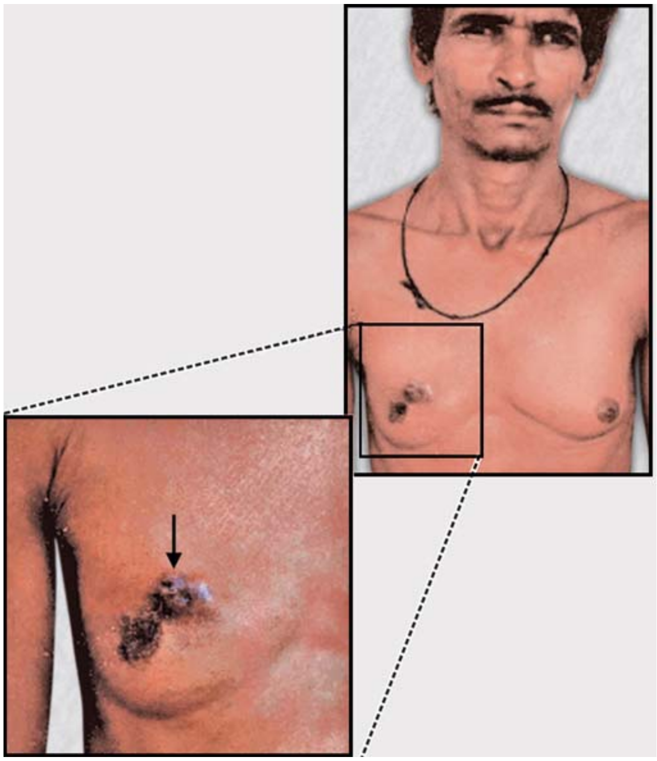

Fig. 8.89: Carcinoma male breast in a pre-existing gynaecomastia. Note the gynaecomastia on opposite side.

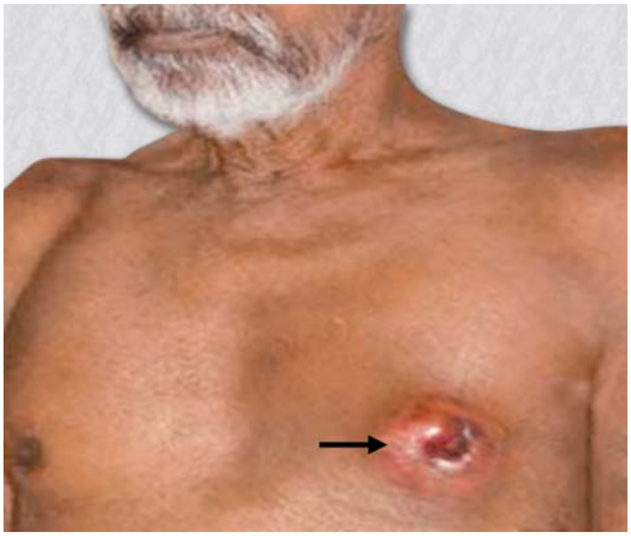

Fig. 8.90: Carcinoma male breast showing mass under the areola, puckering and nipple retraction.

Carcinoma breast in pregnancy

- Incidence is 3%
- *Breastfeeding* should be avoided during chemotherapy. Lactation should be suppressed by giving bromocriptine 2.5 mg OD for 5 days
- *Hormone treatment* during pregnancy is potentially teratogenic
- During pregnancy breast shows increased vascularity, density, water content and cellularity. So US and mammogram may show false-positive results. Chest X-ray can be done with abdominal shield
- Alkaline phosphatase is elevated during pregnancy and so it is also not significant
- Bone isotope scan cannot be done during pregnancy
- *MRI is the* investigation of *choice* for primary in breast and also for metastatic work up
- Treatment is modified radical mastectomy (MRM)
- Chemotherapy can be given in 2nd trimester with care
- Radiotherapy has no role
- As commonly ER negative, hormone therapy is not used
- When distressing secondaries are present termination of pregnancy may be required
- *Women with breast cancer can become pregnant 2 years after the completion of therapy, as recurrence is more common in 2 years*

Management in carcinoma breast during pregnancy

1st trimester	*2nd trimester*	*3rd trimester*
MRM	MRM	MRM
If axillary node is +ve then pregnancy is terminated; chemotherapy is given	Chemotherapy carefully	After delivery, chemotherapy with suppression of lactation

Follow-up after Therapy for Carcinoma Breast

- Clinical examination in detail at regular intervals.
- Yearly/two yearly mammography of treated and contralateral breast is a must.
- Bone scan/CT chest, abdomen/tumour marker are done only if there is clinical suspicion of spread/metastases. It is not a regular routine follow-up method at present.

BREAST RECONSTRUCTION

- **Immediate reconstruction.**
  - In early stages of malignancies as well as in selected more advanced stages where the response to neoadjuvant chemo therapy has been good. Not advised in locally advanced disease.
  - *Advantages of immediate reconstruction* are—maximum amount of breast skin is preserved for reconstruction. It will not affect the long-term survival or recurrence rate.
- **Delayed reconstruction** (3-9 months after surgery).

Indications

- Locally advanced disease.
- Radiation needed in postoperative period.
- Patient unfit for prolonged surgical procedure.

Advantages

- Allows for postoperative radiation without prosthesis exposure.
- Avoids fibrosis and fat necrosis where TRAM flap is used.

Factors deciding the reconstruction

- Amount of skin retained
- Stage of the carcinoma
- Earlier radiotherapy
- Type of flap used

Methods of Reconstruction

- Oncoplastic techniques.
- Insertion of breast implants or expanders. They are best for small breasts. Healthy nonirradiated overlying skin is the requirement.
- Flap with implant or expanders.
- Flap reconstruction:
  - Pedicled flap—TRAM flap.
  - Free flap—free TRAM flap into internal mammary artery.
  - Muscle preserving perforator abdominal flap.

Types

- Silicon gel implant under pectoralis major muscle.
- Expandible saline prosthesis with prior tissue expansion.
- If there is less skin or after radiotherapy latissimus dorsi musculocutaneous flap (*LD flap*) or contralateral transversus abdominis muscle flap *(TRAM flap)* is used.
- Superior gluteal flap based on superior gluteal vessels.

- *Ruben's flap* using soft tissue pad overlying the iliac crest based on deep circumflex iliac vessels.

External breast prosthesis which fits within the bra is a simpler cosmetic method.

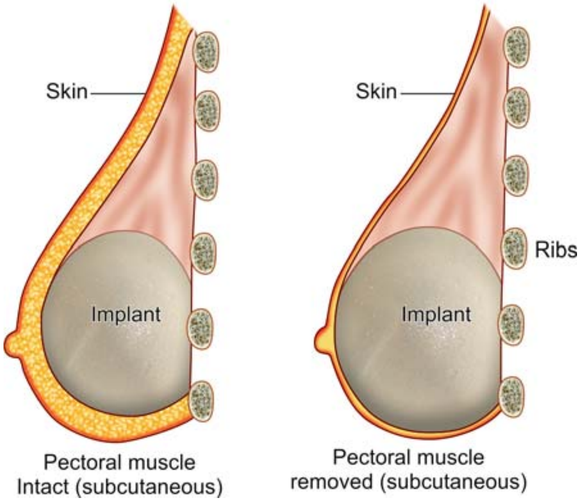
Skin
Implant
Pectoral muscle
Intact (subcutaneous)
Skin
Ribs
Implant
Pectoral muscle
removed (subcutaneous)

Fig. 8.91: Breast implant.

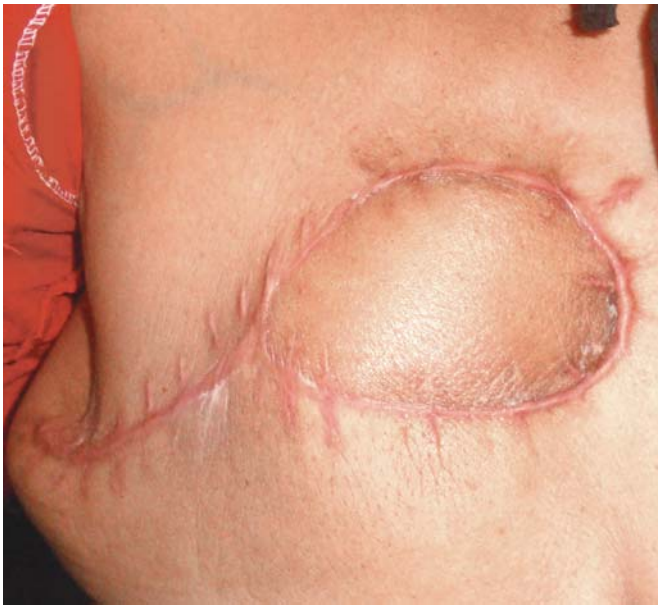

Fig. 8.92: LD flap.

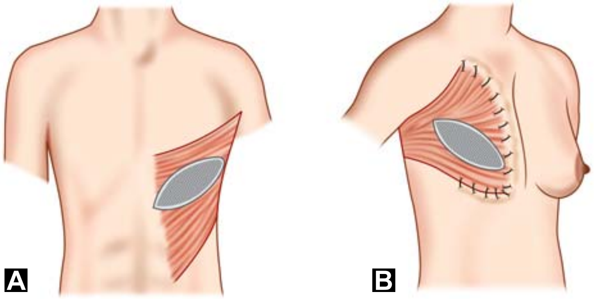

Figs 8.93A and B: LD flap (A) Donor area; (B) Recipient area.

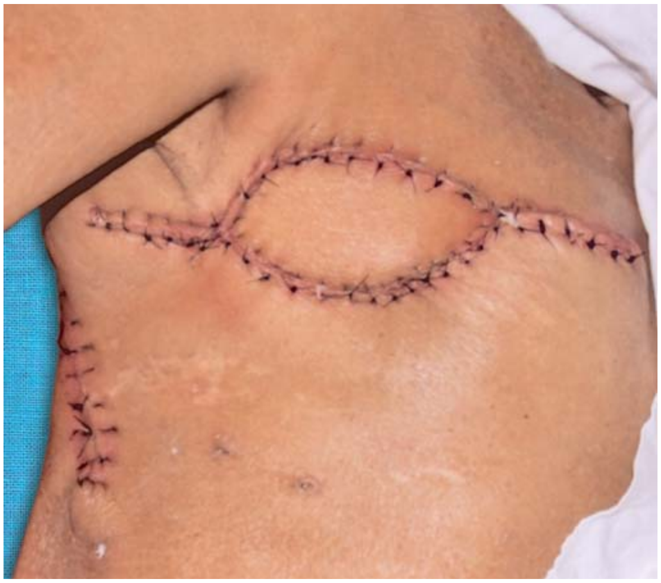

Fig. 8.94: Latissimus dorsi flap (LD flap) placed after mastectomy. It is based on thoracodorsal artery.

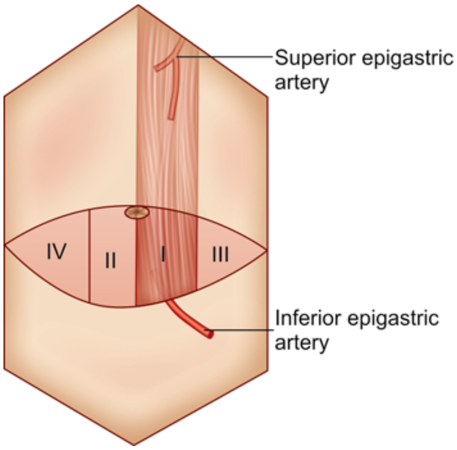
Superior epigastric artery
IV
II
I
III
Inferior epigastric artery

Fig. 8.95: Transverse rectus abdominis myocutaneous flap (TRAM). Note the blood supply of the flap.

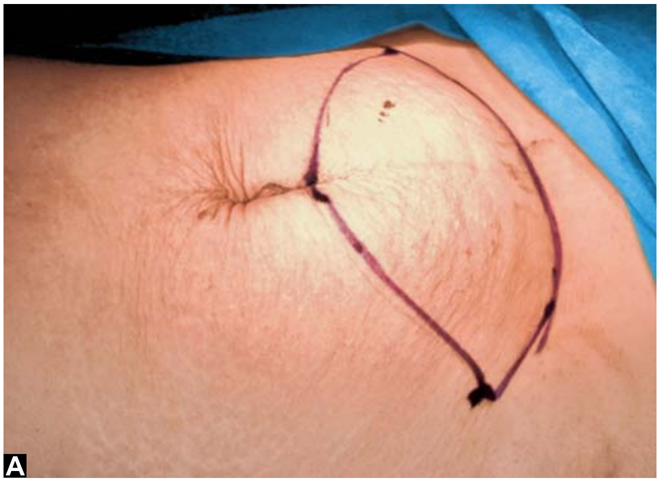

Fig. 8.96A

LD flap	***TRAM flap***
Myocutaneous flap based on subscapular artery	Transverse rectus abdominis myocutaneous flap based on
Easy to perform	superior epigastric artery. ipsilateral or contralateral TRAM flap is used
Reliable flap, well vascularised	It gives the bulk needed for reconstruction and so implant is not needed
Can be placed over prosthesis	Donor site morbidity and fat necrosis can occur
Low complication rate	Free TRAM flap into internal mammary/thoracodorsal axis can be done
But causes unsightly donor area on the back	Limitations to position, shape and rotate; often needs mesh placement in the abdomen wound to prevent development of incisional hernia

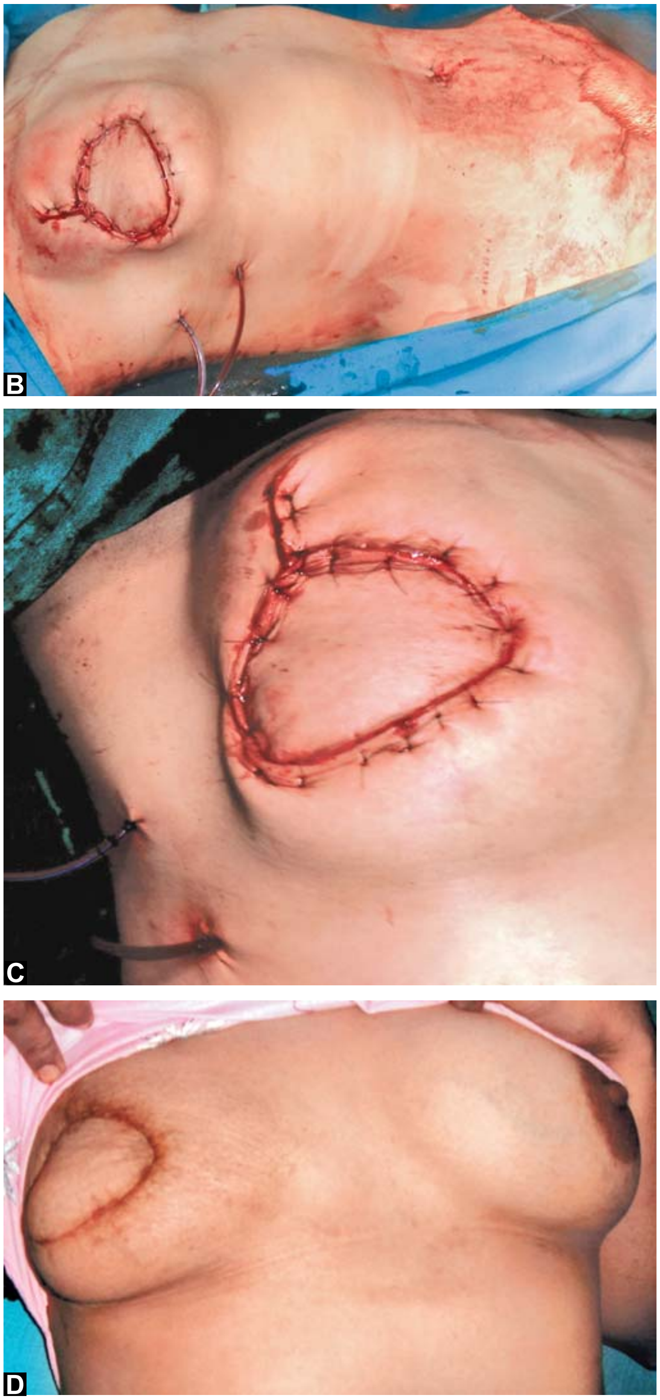

Figs 8.96B to D

Figs 8.95A to D: Different operative steps in TRAM flap.

Note:

- Breast reconstruction is done in young patients with early stage disease.
- Skin sparing mastectomy with removal of nipple areola complex may be better for reconstruction.
- Symmetry is the most important factor in breast reconstruction.

Nipple is created using:

- Local breast flaps 3 months after breast reconstruction.
- Nipple sharing from contralateral nipple using composite graft.
- *Skate flap*: Local flap with de-epithelialised donor site around the periphery over which a full thickness graft is applied.
- Nipple prosthesis can be fitted.

Areola pigmentation is created using *(it is done 3 weeks after nipple creation):*

- Full thickness skin graft from non-hairy skin lateral to labia majora, as the pigmentation of this graft matches that of the areola.
- From contralateral areola if reduction mammoplasty is done on that side.
- Tattooing—colour tends to fade with time and may need to be repeated.
- SSG from retroauricular area or from thigh.

BREAST IMPLANTS

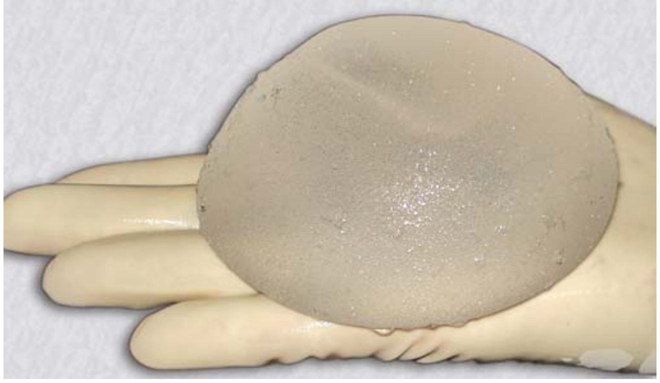

Fig. 8.97: Typical breast implant and its placement. It can be placed in subcutaneous or submuscular plane.

- Technically simple.
- Achieve symmetry easily.

- Implant in *submuscular plane* is better whenever muscle is not removed during surgery.
- If muscle is removed as in radical mastectomy, then *subcutaneous implant* is placed.
- Silicon gel implants are used.

Complications of breast implants

- Pain, exposure of implant and rupture
- Displacement, extrusion
- Infection
- Capsular contraction

Causes of massive enlargement of breast

- Benign hypertrophy—usually bilateral
- Giant fibroadenoma
- Serocystic disease of Brodie
- Sarcoma
- Colloid carcinoma
- Filariasis of breast

Changes that can occur in nipple

- Destruction
- Depression—retraction
- Discolouration
- Displacement
- Deviation
- Discharge
- Duplication

Causes of hard swellings in the breast

- Carcinoma breast
- Antibioma breast
- Traumatic fat necrosis
- Calcified haematoma
- Fibroadenoma—hard variety

NIPPLE RETRACTION

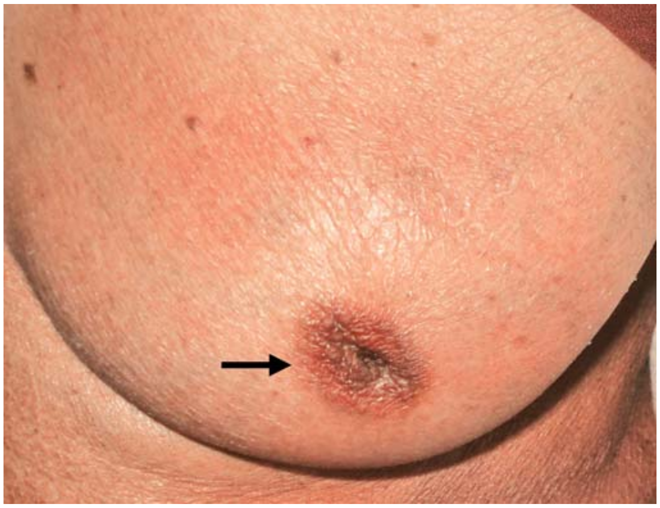

Fig. 8.98: Typical nipple retraction in carcinoma breast. It is circumferential retraction in carcinoma and partial slit like in duct ectasia.

- 25% of retractions are bilateral.
- Congenital nipple retraction is simple nipple inversion. It causes problems in breastfeeding, recurrent infection, collection of secretions.
- Slit like retraction is due to duct ectasia and periductal mastitis.
- Circumferential retraction is usually due to carcinoma.
- Any nipple retraction of recent onset should be evaluated.

Note:

- *Nipple inversion* is congenital, developmental one; it is seen in 25% of females; can be bilateral; does not interfere with infant feeding; can predispose periductal mastitis. Surgical correction is not usually required and is technically difficult. Suction device is often tried mechanically.
- *Nipple retraction* occurs in a nipple which was normal earlier and is now pathologically retracted. Duct ectasia, carcinoma, previous surgeries may be the cause.

Nipple discharge

Classifications	*Significant nipple discharge*	*Types of nipple discharge*	*Discharge which suspects' carcinoma*
It can be: • Physiologic – Bilateral, involves multiple ducts. – Usually not bloody. – Nonspontaneous	• Spontaneous. • Persistent. • Nonlactating patient. – Incidence of carcinoma in nonlactating nipple discharge is 4%.	Blood: • Papilloma/malignancy. *Purulent* • Infection, usually related to lactation; often in malignancy.	• Spontaneous discharge. • Age ≥ 40. • Unilateral discharge. • Bloody or guaiac-positive discharge. • Palpable lump/mass.
• Pathologic – Unilateral. – Spontaneous.	Causes of nipple discharge • Physiological is—during lactation and pregnancy.	*Milky:* • After childbearing up to one year; hypothyroidism,	*Management:* • Discharge study for cytology, culture.

Everytime something good happens to you, make something good happen to someone else.

– Can be often bloody. – Most common cause intraductal papilloma. *It can be*: • Spontaneous. • Nonspontaneous. *It can be due to:* • Benign cause—88%. • Malignant cause—12%.	• Galactorrhoea is milk secretion not related to pregnancy either primary or secondary. • Fibrocystic disease of breast—serous or greenish discharge. • Duct papilloma—bloody discharge. • Duct ectasia—thick, creamy discharge. • Carcinoma—bloody discharge. • Idiopathic—here cause of discharge is not identified in spite of all investigations. Patient can be reassured with regular follow-up. It may be improper reabsorption of the normal ductal secretions (Normally all ductal secretions which occur continuously are reabsorbed completely).	prolactinomas, galactorrhoea. • Medications: Oral contraceptives, tricyclic antidepressants, dopamine agonists. • In milk fistula; rarely in galactocele. *Grey, brown, green, sticky:* • Duct ectasia. Common in 5th decade, with nipple tenderness and pain. *Creamy discharge:* • Duct ectasia. *Serous*: • Fibrocystic disease (most common; but can be serous/greenish/greenish black), ectasia. *Serosanguinous:* • Carcinoma, infection. *Greenish:* • Fibrocystadenosis (commonly), duct ectasia. *Black:* • Duct papilloma/carcinoma.	• Mammography, US of breast and axilla, FNAC/ core biopsy. • Ductography is specific for duct ectasia. • Treat the identified cause accordingly.

Chapter 9 Peritoneum

> *Among the inflammations of the abdominal region, I have given a place in our Nosology to the Peritonitis, comprehending under that title, not only the inflammations affecting the peritoneum lining the cavity of the abdomen, but also those affecting the extensions of this membrane in the omentum and mesentery.*
>
> **—William Gullen,** 1792

CHAPTER OUTLINE

- Anatomy
- Physiology
- Acute Peritonitis
- Spontaneous Bacterial Peritonitis
- Sclerosing Peritonitis
- Biliary Peritonitis
- Postoperative Peritonitis
- Other Forms of Peritonitis
- Pelvic Abscess
- Subphrenic Spaces and Subphrenic Abscess
- Mesenteric Cysts
- Mesenteric Panniculitis
- Acute Mesenteric Lymphadenitis
- Mesenteric Malignancy
- Mesenteric Trauma
- Peritoneal Malignancy
- Omental Cyst
- Omental Torsion
- Omental Tumour

ANATOMY

Peritoneum

It is a serous membrane lining the abdominal cavity. It is composed of outer fibrous tissue layer (which gives strength) and inner mesothelial cell layer (secretes fluid giving lubricating function to the peritoneum).

Parts of Peritoneum

1. *Parietal peritoneum*: It lines the inner surface of the abdominal wall, under surface of diaphragm and pelvic wall. It is loosely attached to the overlying walls and can be easily stripped off. It is innervated by the somatic nerves, so pain sensitive. Anterior peritoneum is most sensitive when compared to pelvic peritoneum.
2. *Visceral peritoneum*: It lines the outer surface of the abdominal viscera, firmly adherent, cannot be stripped off. It is innervated by autonomic nervous system; hence not pain sensitive.

Peritoneal Cavity

It is the potential space between the parietal and visceral peritoneum. Normally it contains 100 ml of clear, straw coloured fluid secreted by the mesothelial cells. Its quantity and quality varies in pathological conditions. It has got lubricating function, allowing frictionless movements of adjacent peritoneal surfaces.

Spaces in peritoneal cavity: Peritoneal cavity being largest cavity in the body is divided into *different spaces* by ligaments and mesenteries. Eleven ligaments are—coronary, gastrohepatic, hepatoduodenal, falciform, phrenicocolic, splenorenal, gastrosplenic, duodenocolic, gastrocolic; mesenteries are—transverse mesocolon and bowel mesentery. Intra-abdominal spaces are nine in number; they are—right and left subphrenic, subhepatic, lesser sac, supramesenteric, inframesenteric, right and left paracolic gutters, and pelvic. Peritoneal cavity is surgically divided basically into *supracolic and infracolic* compartments.

PHYSIOLOGY

Surface area of peritoneum is 2 m^2 equal to surface area of skin. It has property of bidirectional transfer of substances. It can absorb from or transfer to peritoneal cavity, fluid and electrolytes. This property is used in peritoneal dialysis. The

Attack the problem, not the person.

mesothelial cells facilitate rapid healing following any injury, which takes place very fast in matter of hours. Adhesions occur due to delayed or incomplete peritoneal healing.

Diaphragmatic defense: Movement of peritoneal fluid and lymph in the peritoneal cavity depends on the diaphragmatic movement and phases of respiration. During expiration altered intrathoracic and intra-abdominal pressure makes all fluid, bacteria and particles to move/circulate upwards towards diaphragm; during inspiration, it facilitates the entry of these contents into lymph to clear it. There are fine intercellular pores in the peritoneum called as *diaphragmatic stomatas* covering the inferior surface of the diaphragm which leads into diaphragmatic lymphatics and so to regional lymph nodes and thoracic duct. During expiration intra-abdominal pressure is reduced and peritoneal fluid through capillary action travels upwards towards diaphragm. Peritoneal macrophages promote leukocyte migration into the peritoneal cavity. Peritoneal mast cells release vasoactive products causing vasodilatation on peritoneal surface and release of complements and immunoglobulins into the peritoneal cavity. Complements cause bacterial opsonisation and phagocytosis. Release of fibrin causes adhesions and localisation of infection to one place preventing generalised peritonitis; this often leads into localised abscess formation.

ACUTE PERITONITIS

Types

- *Primary*
- *Secondary*
- *Tertiary*

PRIMARY PERITONITIS

- It is commonly due to pneumococci, and can occasionally be due to streptococci, haemophilus, gonococcus (rare now) and other gram-negative (*Escherichia coli*) organisms. It is common in young girls between 3-9 years of age group and also in women.
- Here there is no documented source of infection. Infection usually spreads from lower genitals through fallopian tubes, from upper respiratory tract infection or from middle ear in males.
- It is uncommon after 10 years of age. It is common in malnourished child and child with nephritis. Child is toxic, severely ill and develops septicaemia very early.
- It is also seen in ascites, patient with indwelling catheter for peritoneal dialysis, patients with peritoneovenous shunt.
- It can also be due to Chlamydial, fungal or mycobacterial infection.
- TC is very high, > 30,000/mm^3.

Treatment

- Diagnostic tapping, tube peritoneal drainage and laparoscopic drainage and wash are useful methods.
- Laparotomy and peritoneal toilet.
- Broad spectrum antibiotics including combination of aminoglycosides, cephalosporins and metronidazole.
- Local instillation of antibiotics, into the peritoneal cavity to achieve quick and effective results.

Mortality is high.

SECONDARY PERITONITIS

It is secondary to any bowel or other visceral pathology, e.g. perforation, appendicitis. *Escherichia coli* (70%) is the most common organism involved. Other bacteria are—aerobic and anaerobic streptococci, *Clostridium welchii*, bacteroides, staphylococci, *Klebsiella, Salmonella typhi*.

TERTIARY PERITONITIS

- It occurs after any abdominal surgeries, which is usually severe and the patient may go in for SIRS or MODS early.
- Tertiary peritonitis is defined as persistent or recurrent intra-abdominal infection after an adequate treatment for primary or secondary peritonitis—usually after 48 hours.
- It is common in immunosuppressed individual with ineffective peritoneal host defenses against microbes. Infection due to *E. faecalis*, *E. faecium*, *S. epidermidis*, *P. aeruginosa*, *C. albicans* are common in such patients. Virulence and resistance to drugs are other factors.
- It is difficult to diagnose clinically causing delay in therapy. CT abdomen, total and platelet count, LFT, monitoring of renal functions, hourly urine output assessment, chest X-ray are *required investigations*.
- *Treatment* is aggressive antibiotic therapy, antifungal therapy, TPN, maintaining of haemodynamic stability, exploration of abdomen, thorough wash, colostomy/ileostomy or exteriorisation of bowel segment. FFP, packed cells, platelet transfusions may be required. Ventilator, ICU care is often needed.
- *Mortality rate* is > 50%. Problems are—DIC, septicaemia, uraemia (haemodialysis may be needed), haemorrhage, pneumonia, ARDS.

Bacteria Causing Peritonitis

a. *Bacteria from GIT—E. coli,* aerobic streptococci, *Streptococcus faecalis, Staphylococcus*, anaerobic streptococci, anaerobes (bacteroides), *Klebsiellla, Cl. welchii.*

b. *Bacteria not from GIT*—Gonococcus, pneumococcus, are from fallopian tubes, commonly occurs in young females. *Chlamydia*, β-haemolytic streptococci, *Mycobacterium* are few other bacteria which can cause peritonitis.

a.	*Primary*	Monomicrobial, extraperitoneal source, blood spread
b.	*Secondary*	Most common, polymicrobial, intra-peritoneal source
c.	*Tertiary peritonitis*	Due to superadded infection following treatment of secondary/primary

Most common bacteria during the phase of peritonitis is *E. coli* and during abscess formation is *B. fragilis. Mortality for diffuse peritonitis is 10%.*

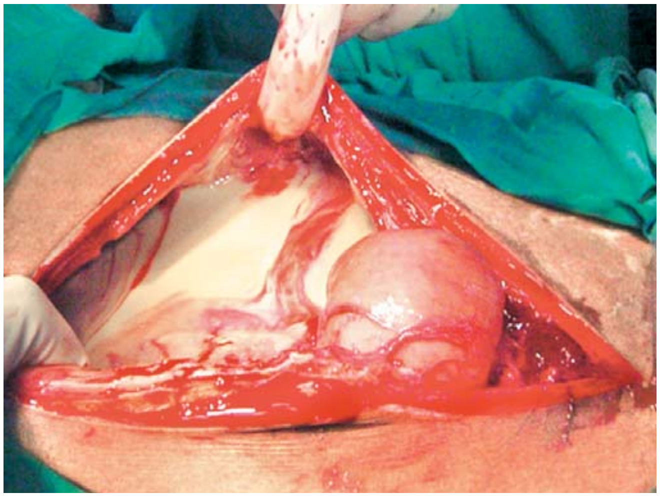

Fig. 9.1: On table finding in a case of peritonitis, where the peritoneal cavity is filled with pus.

Mode of Infection

- Perforation of the GIT—duodenal/gastric/enteric/colonic ulcers; Meckel's diverticulitis perforation.
- Penetrating or blunt trauma.
- Surgery.
- Drains.
- Dialysis.
- Foreign body.
- Appendicitis, cholecystitis, diverticulitis.
- Intestinal obstruction with strangulation.
- Via fallopian tubes.
- Through blood spread—in septicaemia.
- Transmural spread.
- Following uterine perforation/injury during abortion or termination of pregnancy.

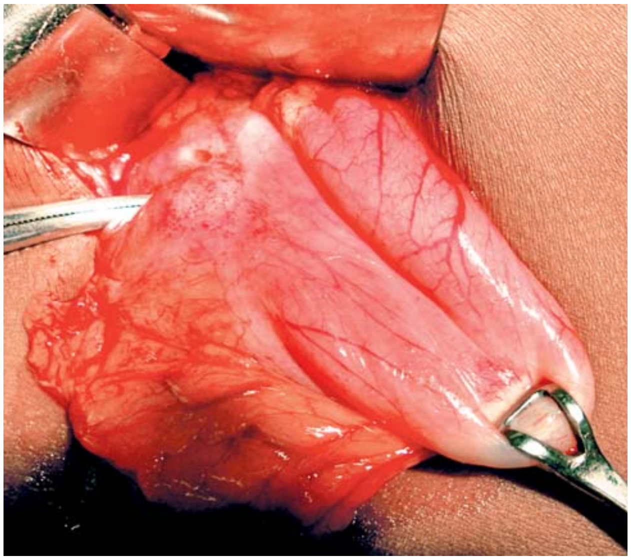

Fig. 9.2: Duodenal ulcer perforation causing peritonitis.

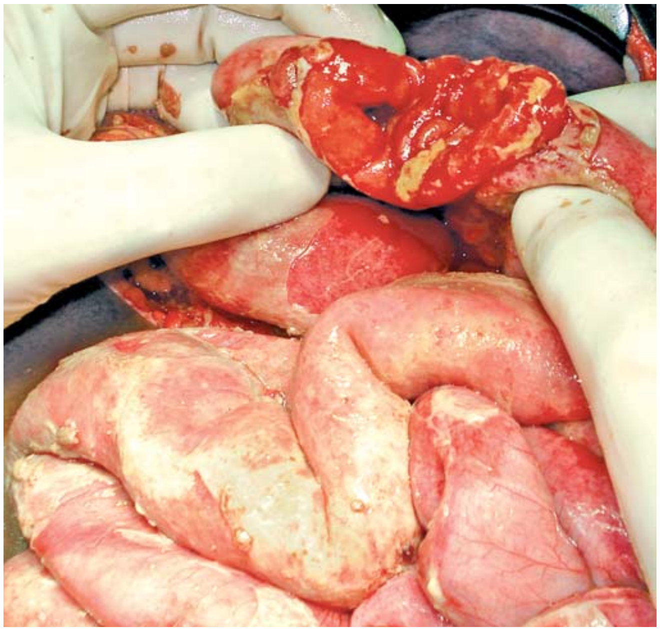

Fig. 9.3: Jejunal perforation due to trauma. Trauma is the common cause for jejunal perforation. It is commonly blunt abdominal injury due to shearing force (traction injury) causing jejunal tear near duodenojejunal junction.

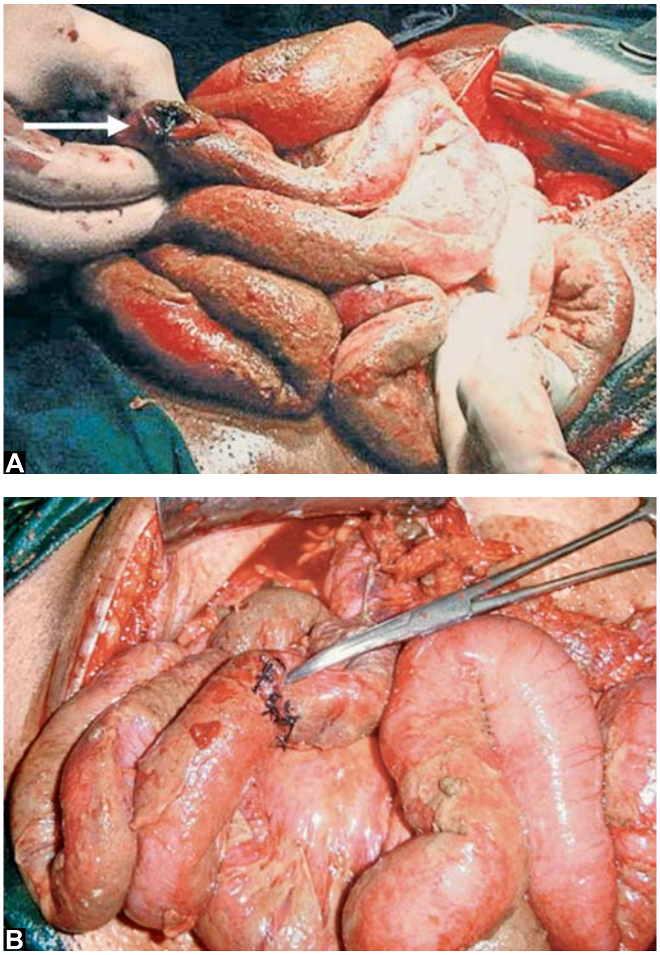

Figs 9.4A and B: Ileal perforation with severe peritonitis. Note the closure done horizontally. It could be typhoid perforation.

Better three hours too soon, than one minute too later.

Factors affecting the spread of the infection in peritonitis

- Rapidity by which the pus is gushed into the peritoneal cavity, e.g. burst appendix, perforations
- Amount of peristalsis (more the peristalsis more the spread)
- Virulence of the organism
- Localising action of the omentum (in children localisation is poor as omentum is small)
- Immunosuppression—HIV, steroids
- Anatomical nature of the peritoneal cavity
- Age, associated diseases like malignancy, malnutrition, anaemia

Pathogenesis

- Lot of fluid is secreted into the peritoneal cavity which is infected, containing bacteria and toxins causing shock, toxaemia and its subsequent effects. Fibrinogen forms fibrin which tries to localise the infection. Bowel gets adhered to each other with fluid collecting between the loops. Thick flakes are formed adhering to the surface of the bowel. Peritoneum is thick, oedematous, becomes velvety and reddish with loss of its glistening appearance. Omentum is thickened, adherent. Often site of perforation may be identified by the location of the end of the omentum. Dilated bowel loops with site of obstruction/gangrene may be found. Pus, often with pockets may be present in subphrenic, paracolic and pelvic spaces.
- Peritoneal contents are initially sterile but eventually become infected in certain situations, as in acute pancreatitis, in haemoperitoneum, and in ruptured urinary bladder. It is due to transmural migration of bacteria.
- In perforated duodenal or gastric ulcers, contents in the peritoneal cavity are initially sterile but later get infected to form bacterial peritonitis.
- *Localised peritonitis*—initially localisation of peritonitis occurs based on the anatomical factors like supracolic/infracolic compartments; greater omentum; paracolic gutters; dilated small bowel, etc. Pathological factors are thickened peritoneum, fibrin deposition, omental adhesions, reduced bowel peristalsis. Localised peritonitis may resolve by proper therapy. It may form abscess—pelvic/subphrenic. If it progresses it may form generalised peritonitis.

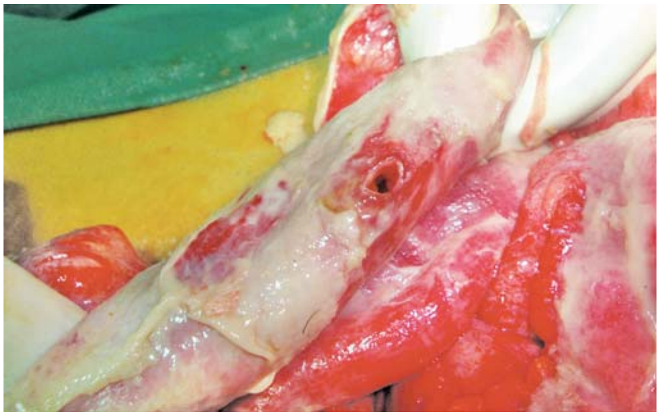

Fig. 9.5: Small bowel perforation with peritonitis with thick plaques on the bowel surface.

- *Diffuse peritonitis* sets in due to poor localisation, rapid peritoneal contamination, violent peristalsis, virulent organisms, immunodeficiency status, and small omentum in children (not properly developed in children). Diffuse peritonitis may occasionally get localised to form abscess. Mortality in diffuse peritonitis is 10%.

Clinical Features

- Sudden onset of pain which is *severe*.
- Fever, vomiting.
- Tenderness—initially localised later becomes diffused.
- Rebound tenderness—*Blumberg sign.*
- Guarding and rigidity, dull flanks on percussion.
- Tachycardia, tachypnoea.
- Tenderness on P/R examination.
- Distension with silent abdomen.
- Eventually leading to Hippocrates facies, septicaemic shock and loss of consciousness.
- Bowel sounds are absent due to paralytic ileus.
- Fever may be absent in severe peritonitis due to loss of pyrogenic reaction. Total count also may be very low in severe peritonitis.

Investigations

- Plain X-ray abdomen (in erect posture)—will show *ground-glass appearance* along with gas under diaphragm in the presence of perforation.
- Total count is increased.
- U/S abdomen—shows fluid in the abdominal cavity.
- Also clinches the other causes like haemoperitoneum, pancreatitis.
- Electrolyte study.
- Blood urea and serum creatinine.
- Serum amylase if *four times the normal value*, it is significant. LFT also should be done. Platelet count, bleeding time, clotting time, prothrombin time are to be assessed in severe peritonitis.
- Four quadrant abdominal tap—reveals pus or infected fluid. In suspected pancreatitis fluid should be analysed for amylase level (which will be high). Often US guided aspiration is better and more accurate.

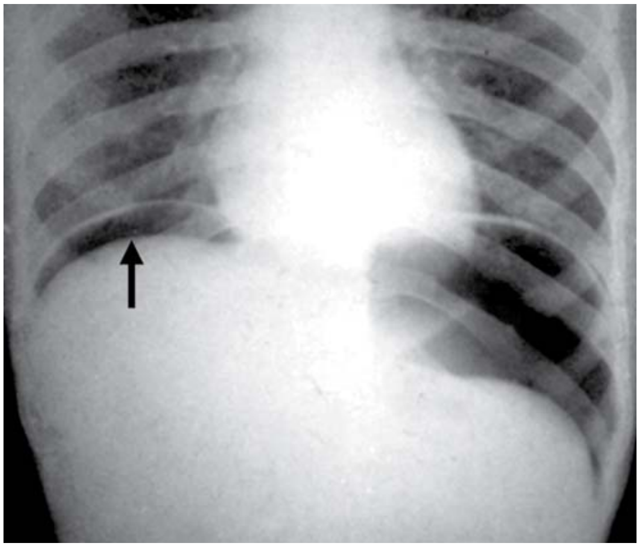

Fig. 9.6: Plain X-ray abdomen showing gas under diaphragm. Perforation is the most common cause of peritonitis.

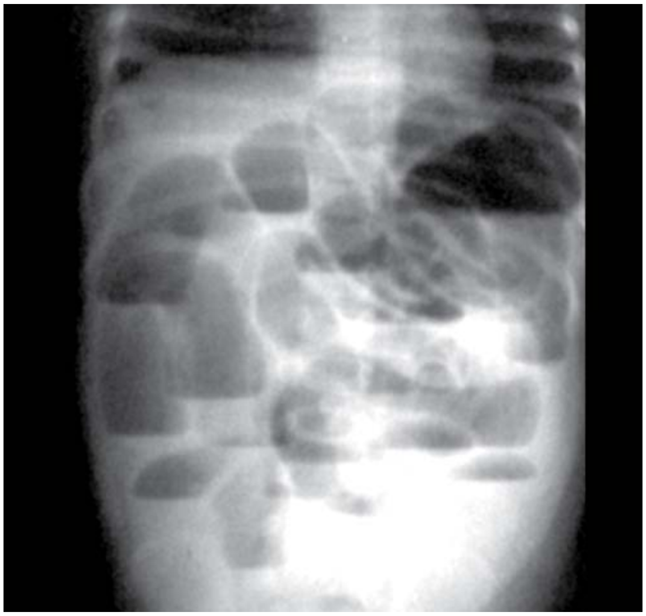

Fig. 9.7: Multiple air fluid levels in plain X-ray abdomen—feature of intestinal obstruction.

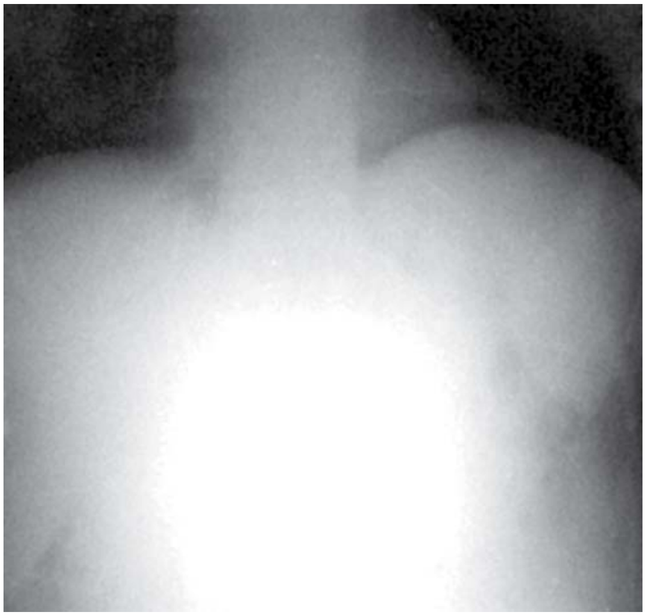

Fig. 9.8: Plain X-ray abdomen showing ground glass appearance—a feature of peritonitis.

- ***Diagnostic laparoscopy. It can also be used as therapeutic for duodenal ulcer perforation or primary peritonitis to give peritoneal wash.***
- CT scan is often useful to confirm the cause or to rule out conditions like pancreatitis. CT scan detects bowel ischaemia, gangrene, perforation, internal hernias and quantity of pus/fluid in the peritoneal cavity.

Differential Diagnosis

- *Pancreatitis:* In pancreatitis back pain is common. But it is often difficult to differentiate clinically from acute peritonitis. Serum amylase is increased. CT abdomen may help in differentiating from peritonitis.
- *Intestinal obstruction*: There will be distension, vomiting and pain. Plain X-ray shows dilated bowel loops. CT scan confirms intestinal obstruction and its cause.
- *Ruptured ectopic pregnancy*: Urine pregnancy test will be positive. History of amenorrhoea is significant. Palor, tachycardia, lower abdominal pain and distension are common presentations. US abdomen and pelvis confirms the diagnosis.
- *Acute pyelonephritis*: Often clinically acute pyelonephritis mimics acute peritonitis. Guarding, tenderness, rigidity may be present in pyelonephritis also. Urinary symptoms, tender renal angle, urine microscopy revealing pus cells, US abdomen may confirm the diagnosis.
- *Acute mesenteric ischaemia*: Bloody diarrhoea, colicky pain, tender abdomen, toxicity are the features of mesenteric ischaemia. CT abdomen and angiogram are useful investigations.
- *Diabetic acute abdomen*: Patient with diabetes mellitus may present as acute abdomen. Free fluid in the flank may not be present. Patient may be ketotic also.

Treatment

- Intravenous fluids for resuscitation are essential. It improves the tissue perfusion, corrects the hypotension and also improves the urine output. Normal saline, Ringer's lactate are usually used. Usual requirement is 2 ml/kg/hour.
- Nasogastric tube aspiration—to decompress bowel; to reduce toxic fluid; to prevent aspiration.
- Total parenteral nutrition, CVP line to perfuse and to monitor.
- Blood transfusion, FFP, platelet transfusions if indicated.
- Catheterisation with maintenance of adequate urine output (30 ml/hour) (0.5 ml/hour/kg).
- Antibiotics—ampicillin, gentamicin, metronidazole, ceftazidime, cefoperazone, cefotaxime, tazobactum, piperacillin, meropenem, linezolid, etc.
- Analgesics to relieve pain by drugs, epidural analgesia.
- Sitting propped up position, early mobilisation, exercise, respiratory physiotherapy, prevention of DVT using heparin/low molecular heparin are essential.
- Surgical correction of underlying cause.
- Use of dopamine, steroids, and management of shock.
- Often ICU and ventilator support is required during postoperative period.
- Monitoring the patient using PO_2, PCO_2, electrolytes, and pulse oximeter.

Laparotomy and proceed

  - On laparotomy, the cause for peritonitis is identified and corrected.
- In bowel perforation—perforation closure.
- In intestinal obstruction—resection and anastomosis.

Eyes are more accurate witness than ears.—**Hiraclitus.**

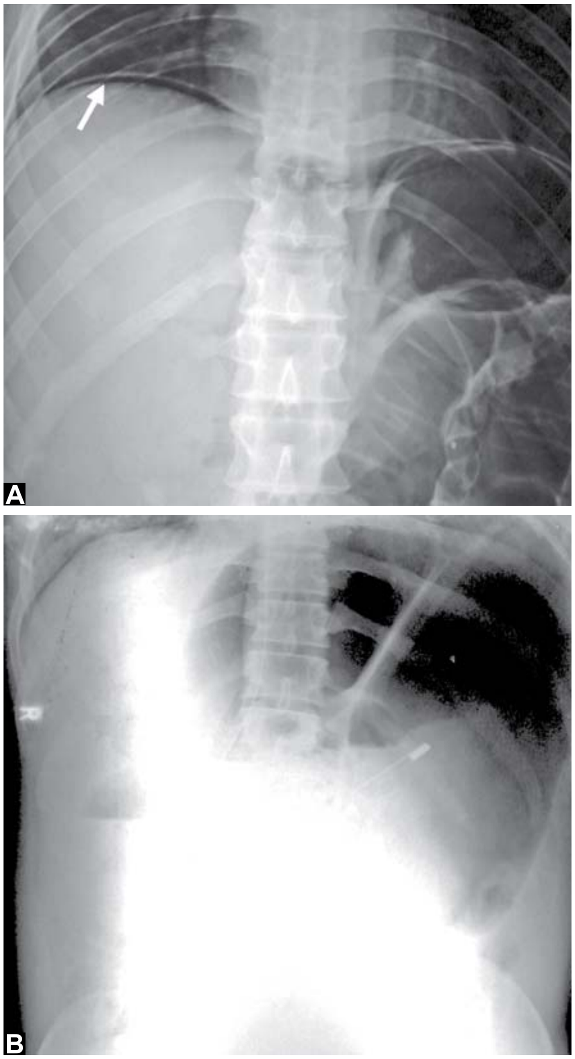

Figs 9.9A and B: X-ray showing gas under diaphragm due to perforation. Ground glass appearance of peritonitis is seen in another X-ray.

- In appendicitis—appendicectomy.
  - Proper peritoneal toilet (wash with 10 litres of saline) is given. A drain is placed (tube).
  - Pus should be sent for culture and sensitivity.
  - Tension sutures are placed when required to prevent burst abdomen.

  Laparostomy—may be *open or closed method*. In *open method,* (Ettapan's lavage) abdominal wall/fascia and peritoneum are not sutured but kept open. It will reduce the chances of abdominal compartment syndrome and also allows easier and faster drainage of the infected fluid from the peritoneal cavity. But disadvantages are it causes lot of fluid loss and electrolyte imbalance and open area itself may allow infection to get into the peritoneal cavity. In *closed method,* abdominal fascial wound is closed using a zip. It facilitates the regular washing of the peritoneal cavity with normal saline without contaminating the cavity from outside and reduces the unnecessary fluid loss. But it may cause abdominal compartment syndrome.

Technique of emergency laparotomy

Abdomen is opened with midline incision (with wide exposure), often extending both above and below the umbilicus, with incision inclined towards the left side at umbilicus. Once peritoneum is opened gushing pus is removed using suction. Pus is collected for culture. Abdominal cavity is inspected thoroughly for site of origin of sepsis—appendicitis/perforation/gangrene, etc. Definitive procedure is done depending on the cause. Proper wash is given using normal saline (10 litres). Drain should be placed. Wound is closed with tension sutures using nonabsorbable monofilament sutures (zero or one number).

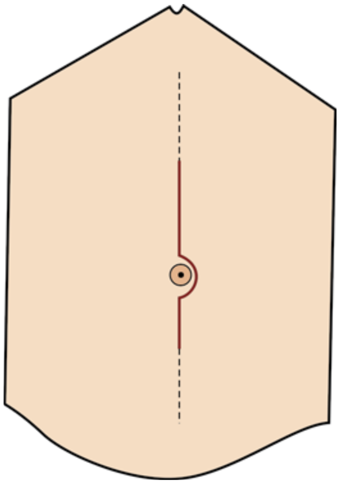

Fig. 9.10: Vertical midline incision is used for emergency laparotomy in peritonitis. Lengthy incision with wide exposure of the peritoneal cavity is advisible.

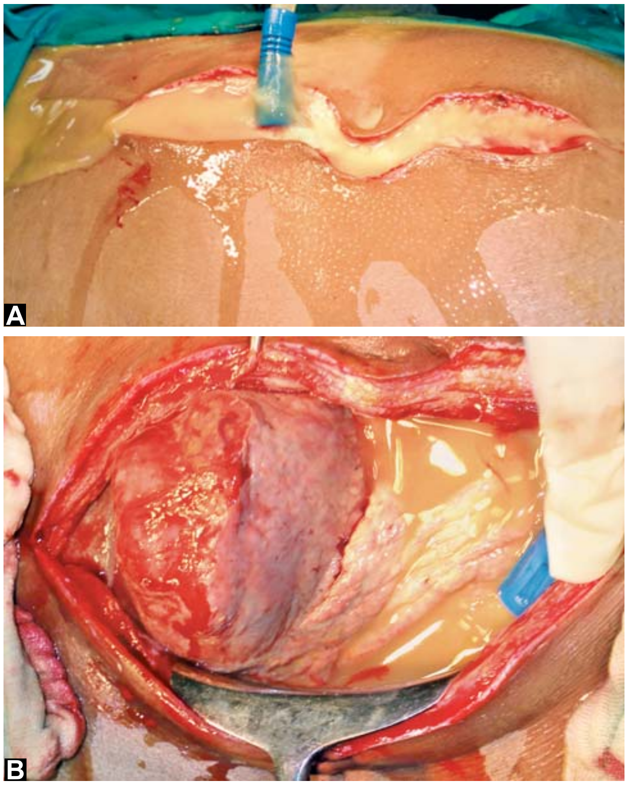

Figs 9.11A and B: Severe peritonitis showing pus in the peritoneal cavity.

Postoperative management

- Proper critical care (ICU) postoperatively.
- Ventilatory support; monitoring with urine output, temperature, breathing, pulse rate, blood pressure, arterial blood gas

analysis, total count estimation (both very high and very low counts signify severe sepsis/septicaemia), neutrophil count estimation, platelet count, prothrombin time analysis, blood urea and serum creatinine, liver function tests (altered liver function signifies septicaemia). These investigations need to be repeated as and when required.

- Proper fluid and electrolyte management.
- Total parenteral nutrition (TPN).
- Prevention of deep vein thrombosis by leg exercises, low molecular weight heparin.
- Prevention/identification of ARDS and its management.
- Prevention of bedsore.

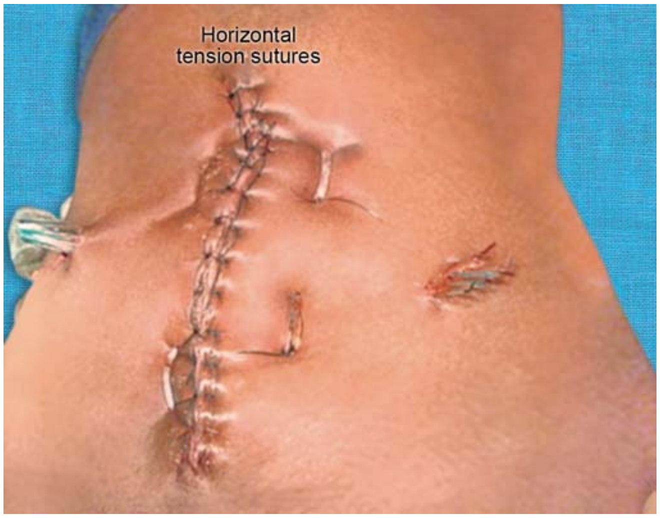

Fig. 9.12: Abdomen is closed with tension sutures in case of peritonitis using monofilament nonabsorbable sutures to prevent postoperative burst abdomen. Drain also should be placed into the peritoneal cavity.

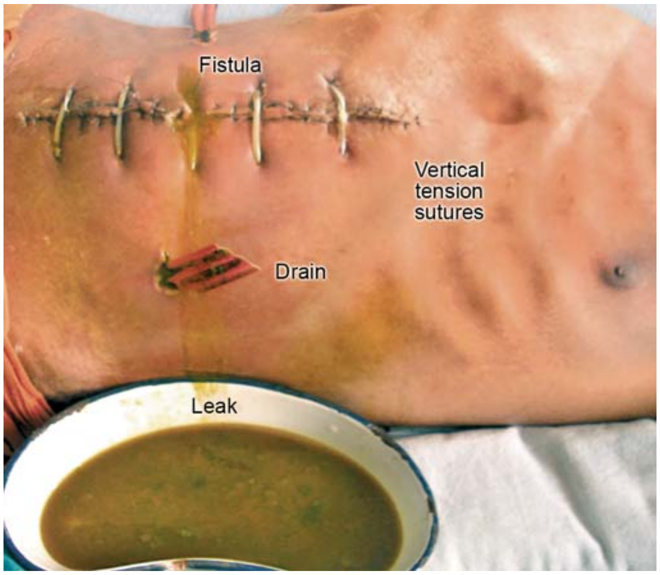

Fig. 9.13: Postoperative patient with tension sutures with fistula and leak. Patient has undergone resection and anastomosis for gangrene bowel.

Terminal Features in Peritonitis

a. *Hippocratic facies*—bright, hollow eyes; pale, pinched face; cold perspiration in eye brows; dry, fissured tongue; blue lips. Patient will be in severe shock. It signifies terminal status with high mortality rate. It is a typical facial sign of *severe end stage peritonitis.*

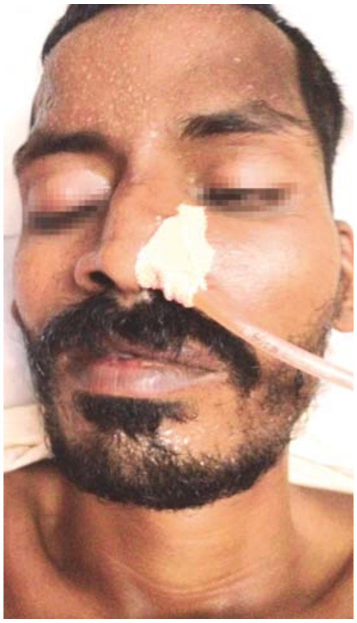

Fig. 9.14: *Hippocratic facies* in severe end stage peritonitis.

b. Septic shock (cold stage).
c. Systemic inflammatory response syndrome (SIRS).
d. Multiorgan dysfunction syndrome (MODS).

Complications and Sequelae of Peritonitis

- Septicaemia.
- Paralytic ileus, adhesions, intestinal obstruction.
- Respiratory infection.
- ARDS, bronchopneumonia.
- Electrolyte imbalance, DVT.
- Renal failure.
- MODS.
- DIC.
- Formation of subphrenic abscess, pelvic abscess and other intraperitoneal abscess.
- Burst abdomen; biliary/gastric/enteric/faecal fistula formation; later incisional hernia.

Note:

Complications may be systemic (septicaemia, respiratory, DIC, DVT, renal failure) or abdominal (ileus, intra-abdominal abscess, fistula, burst abdomen, portal pyaemia, adhesions).

SPONTANEOUS BACTERIAL PERITONITIS (SBP)

SBP is defined as bacterial infection of ascitic fluid in the absence of an intra-abdominal surgically treatable source of infection.

It is seen in:

- Infants and children commonly.
- In nephritic syndrome.
- In adults, with cirrhosis, most common cause. In cirrhotic patients, there is low Fc γ receptor on macrophage surface causing poor phagocytosis; opsonic action of peritoneal surface cells in presence of ascites is low causing poor control of bacterial load.
- In patients who has undergone splenectomy.
- Patients with ascites due to any cause.
- Malnutrition and malignancy.

Causative Organisms

- Gram-negative bacilli (70%)—*E.coli*, *Klebsiella*.
- Gram-positive cocci (20%)—pneumococci, streptococci and enterococci.
- Anaerobes (10%).
- In girls it is commonly through fallopian tubes. In males it is always blood born from respiratory tract infection or otitis media.

Features

- Features of peritonitis but initially with dull aching pain.
- Features of shock.
- Common in girls.
- High fever with neutrophilia is characteristic.
- Profuse diarrhoea is common.
- Features relevant of cause.
- Total count, LFT, U/S is useful tool.
- SBP is rare in ascites with high protein content like in peritoneal carcinomatosis. More than 250 neutrophils/cu mm^3 of ascitic fluid are diagnostic.

Note:

Scoring of the severity of the peritonitis is done based on age, sex, duration, cause, origin, content, general condition, haematocrit, electrolytes, O_2 saturation, blood urea, serum creatinine.

1. Mannheim peritonitis index.
2. Acute physiology and chronic health evaluation score (APACHE score).

Management

- Usually conservative with proper higher generation antibiotics. But condition has got high mortality.
- Tapping of the purulent fluid from the peritoneal cavity is done only with atypical needed cases.
- Laparoscopy is useful method.
- Often peritoneal purulent fluid tube drainage may be beneficial.
- Laparotomy and peritoneal wash is done occasionally if secondary peritonitis and toxicity develops.
- Plasma albumin is helpful in SBP with cirrhosis to reduce hepatorenal syndrome and to decrease mortality.

Note:

- Exudate in pneumococcal peritonitis is odourless, turbid and sticky.
- Exudate in streptococcal peritonitis, is odourless, thin, blood stained with flecks of fibrin. Greenish watery diarrhoea is common. It is common in infants and children. Upper respiratory tract infections, tonsillitis, pharyngitis are the causes. It is treated with penicillins, amoxycillins, cephalosporins.
- Staphylococcus aureus peritonitis is common in those who use intravaginal tampons causing toxic shock syndrome and DIC.

SCLEROSING PERITONITIS

It is usually Practolol (β blocker) induced peritonitis. It is fibrinous peritonitis causing thickening of bowel and other contents of abdomen. It leads to intestinal obstruction and malnutrition.

BILIARY PERITONITIS

Causes

- Trauma.
- Postoperative leak after surgery for gallbladder, CBD, duodenum, ERCP, pyloroplasty, post-gastrectomy duodenal stump leak/blow out, leak from choledochojejunal anastomosis.
- Perforation of gallbladder or CBD.
- Gangrenous cholecystitis and rupture.
- Idiopathic.
- Leakage from accessory bile duct after surgery; leakage from displaced T tube or displaced percutaneous stent.

Clinical Features

- Features of peritonitis—abdominal distension, guarding and rigidity.
- Jaundice: It may be due to CBD injury, liver biliary ductile injury, sectorial duct injury, cystic duct leak or by absorption of bile from free peritoneal cavity.
- Toxicity.
- Features of shock.

Investigations

- X-ray abdomen.
- Diagnostic peritoneal lavage.

Treatment

- Laparotomy, wash, drainage.
- Cause to be treated. US guided percutaneous drainage; ERCP stenting; duodenal stump exploration, wash and drainage or jejunal serosal patch over it; ligation of cystic or accessory duct; choledocho/hepaticojejunostomy, etc. depending on cause.
- It has got high mortality.
- It may get localise and form abscess or may lead into biliary fistula.

POSTOPERATIVE PERITONITIS

- It is the development of peritonitis in postoperative period, either due to anastomotic leak, or biliary leak; and there is collection of pus in the peritoneal cavity including interloop areas.
- It is often difficult to differentiate it from postoperative urinary infection, respiratory infection like bronchopneumonia, cardiac causes in old people, and infection in

other areas like intravenous cannula site (thrombophlebitis). Symptoms may mask if patient is on steroid therapy.

- In all abdominal surgeries patient initially develops paralytic ileus which lasts for 3-5 days. If it persists or progresses one should suspect postoperative peritonitis. Other suspicious features are—persisting toxaemia in spite of antibiotics, oliguria, pyrexia, altered liver functions, altered total WBC count (very high as well as very low and neutrophil count is more specific) and increasing leak from drain site either fecal/intestinal/biliary, increase in abdominal distension, pain and tenderness. Leak may be due to sepsis, hypotension in pre/per and postoperative period, hypoxia, blood loss, ischaemia of bowel and friability of bowel wall, and specific causes like tuberculosis, malignancy or Crohn's disease. Often multiple areas of the bowel become necrotic which eventually causes leak. Leak from multiple sites is difficult to manage and has got poor prognosis.

Management

- Immediate US or CT abdomen to find out the abdominal cause, total count, assessment of urinary parameters like blood urea and serum creatinine, liver function tests.
- Immediate exploration is the only way to save the life of the patient. Proper exploration with wide exposure of the abdomen, is done to identify the cause for infection is done. Resuturing of the anastomotic site can be done but chances of releak are more likely. It can be done, if leak site is small and less friable. Otherwise either ileostomy or colostomy is the better choice. Often continuous tube irrigation and drainage is needed to clear the infective material. Laparostomy using zip technique is also beneficial. Multiple cavities with multiple abscesses are often found which should be drained and irrigated adequately with normal saline. Extensive bowel necrosis is the finding often observed which needs massive bowel resection that leads to short bowel syndrome which is difficult to manage on long-term basis. Higher antibiotics like piperacillin, meropenem are required. Ventilator support, total parenteral nutrition (TPN) haemodialysis support and critical care management is required. Postoperative peritonitis has got high mortality and morbidity.

OTHER FORMS OF PERITONITIS

1. Meconium peritonitis.
2. Chronic bacterial peritonitis—tuberculous.
3. Chronic nonbacterial peritonitis—granulomatous peritonitis due to talc or starch (from surgical gloves)—starch peritonitis.
4. Gonococcal peritonitis with perihepatitis—*Curtis-Fitz-Hugh syndrome.*
5. ***Familial periodic peritonitis (Familial mediterranean fever):*** A rare entity gets recurrent pain abdomen often after appendicectomy. It is familidal, common in Arabs, Americans and Jewish populations. It is common in children; common in females; associated with MEFV gene mutation. Specific gene produces a protein—pyrin from neutrophils in these patients. Peritoneum adjacent to spleen and gallbladder are inflamed. But these two organs are normal. It is common in children and common in girls/females. Pain is usually in upper quadrant, left side. There is localised peritoneal inflammation. It is treated conservatively using colchicines.
6. ***Parturition/abortion peritonitis***: It occurs after delivery and abortion conducted without asepsis and instrumentation, incomplete instrumental abortion, uterine perforation with or without bowel injury. It causes pelvic peritonitis commonly, generalised peritonitis often. It is due to anaerobic bacteria, anaerobic streptococci, and gram-negative bacteria like *Escherichia coli*. It is often rapidly progressive causing septic shock. It needs emergency surgery. Complications are ARDS, intestinal obstruction, faecal fistula and intraperitoneal abscess formation.

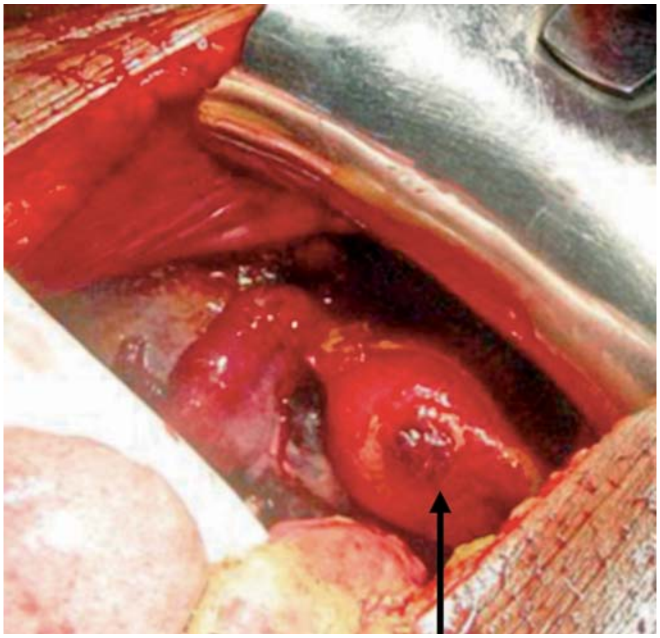

Fig. 9.15: Perforation in the fundus of the uterus due to attempted abortion. It caused severe peritonitis.

PELVIC ABSCESS

- It is the most common intraperitoneal abscess (50-60%).
- It is the collection of pus in rectovesical or rectouterine pouch (pouch of Douglas).

Causes

- Appendicitis.
- Pelvic infections.
- Sequelae of diffuse peritonitis.
- Postoperative and other abdominal causes.

Bacteria: Bacteroides fragilis, *E. coli,* anaerobic streptococci.

Note:

Waltman-Walters syndrome: It is postcholecystectomy fluid collection in subphrenic space causing compression over IVC. It mimics coronary thrombosis.

The mother's heart is child's schoolroom.—**Henry Ward Beecher.**

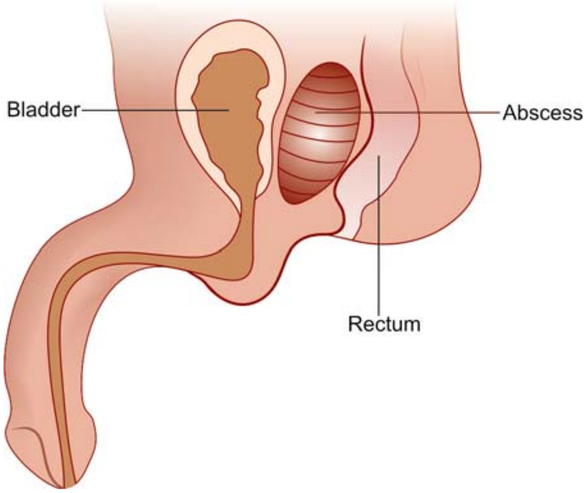
Bladder
Abscess
Rectum

Fig. 9.16: Pelvic abscess.

Clinical Features

- Diarrhoea.
- Mucus discharge per rectum.
- High temperature with chills and rigors.
- Lower abdominal pain and distension.
- Frequency and burning micturition.
- P/R shows a soft, boggy, and tender swelling in the anterior wall of the rectum.

Investigations

- TC is raised.
- U/S is diagnostic—shows pus in rectovesical or pouch of Douglas.
- CT scan to find out the size and extent.

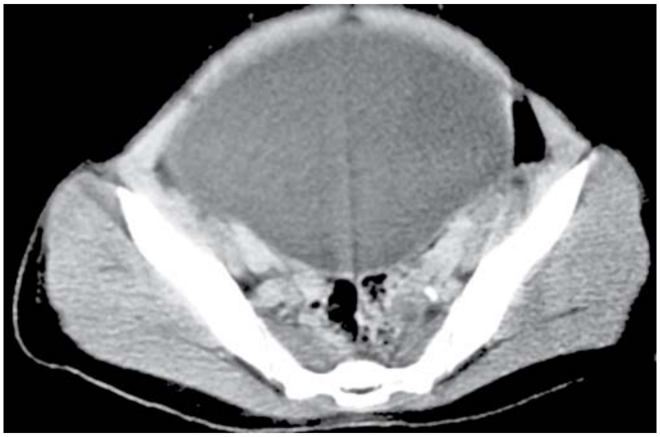

Fig. 9.17: CT pelvis showing pelvic abscess.

Treatment

- After starting antibiotics, under G/A, abscess is drained **per rectally**, through the boggy area. Catheter is passed to the bladder before draining the pus.
- In females, *posterior colpotomy* is done to drain the *pus*.
- Occasionally spontaneous rupture of the abscess into the rectum occurs leading to natural regression.
- When abscess is very large, when it is progressing into the general peritoneal cavity or when in doubt, laparotomy through lower abdomen incision is done to drain the pus and to correct the cause. Often in such cases, drainage tube can be placed through the abdomen under U/S guidance.
- CT or US guided insertion of drainage tube into the abscess cavity per rectally or per vaginally or percutaneously.
- But most often laparotomy is not necessary.

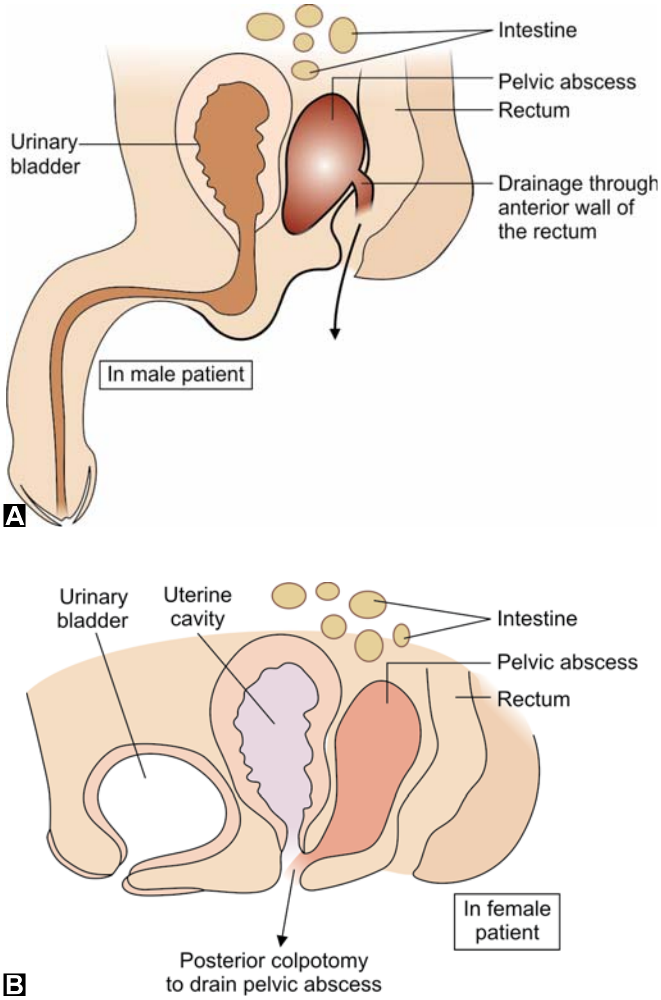
Intestine
Pelvic abscess
Rectum
Urinary bladder
Drainage through anterior wall of the rectum
In male patient
A
Urinary bladder
Uterine cavity
Intestine
Pelvic abscess
Rectum
In female patient
Posterior colpotomy to drain pelvic abscess
B

Figs 9.18A and B: In male, pelvic abscess is drained through an incision over the anterior aspect of the rectum. In female, it is drained through posterior colpotomy.

SUBPHRENIC SPACES AND SUBPHRENIC ABSCESS

Surgical Anatomy (By Boyd)

There are four intraperitoneal and three extraperitoneal spaces.

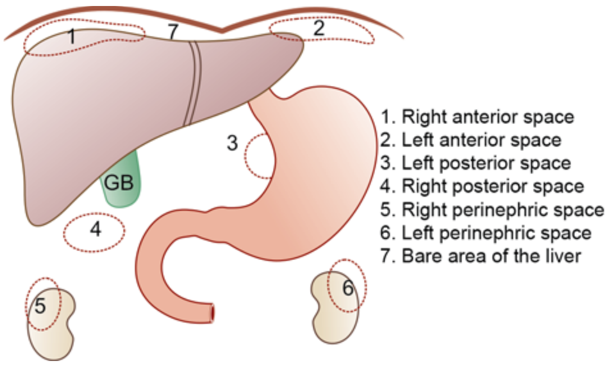
1
7
2
3
GB
4
5
6
1. Right anterior space
2. Left anterior space
3. Left posterior space
4. Right posterior space
5. Right perinephric space
6. Left perinephric space
7. Bare area of the liver

Fig. 9.19: Subphrenic spaces.

Intraperitoneal Spaces

Right anterior intraperitoneal space (Right subphrenic space): It is bounded by right lobe of the liver and diaphragm, posteriorly by anterior layer of coronary and right triangular ligament, and to the left by falciform ligament.

Causes: Abscess here occurs due to cholecystitis, perforated duodenal ulcer, postoperative, appendicitis, duodenal cap blow out.

Right posterior intraperitoneal space (Right subhepatic space): (*Rutherford Morison's kidney pouch*) is bounded in front by the liver and gallbladder, above by the liver, behind by the right kidney and diaphragm, below by the transverse colon and hepatic flexure, to the left by foramen of Winslow and duodenum. It is large and deepest space of all. *It is the most common site of subphrenic abscess.*

Causes: Appendicitis, cholecystitis, postoperative, perforated duodenal ulcer, intestinal obstruction.

Left anterior intraperitoneal space (Left subphrenic space): It is bounded above by the diaphragm, behind by left lobe of liver and left triangular ligament, gastrohepatic ligament and anterior surface of the stomach, to the right is the falciform ligament.

Causes for abscess here are surgeries of the stomach, tail of the pancreas, spleen, colon (splenic flexure), diverticulitis.

Left posterior intraperitoneal space: It is bounded by stomach, pancreas, greater omentum, liver, transverse colon (Lesser sac).

Most common *cause* here is pseudocyst of pancreas. Rarely perforated gastric ulcer.

Extraperitoneal Spaces

Right extraperitoneal space is right perinephric space **and left extraperitoneal space** is left perinephric space.

Causes: Abscess here are due to tuberculosis, trauma, haematoma.

Midline extraperitoneal is *bare area* of the liver. Pus collects here commonly due to ruptured amoebic liver abscess and pyogenic abscess of the liver.

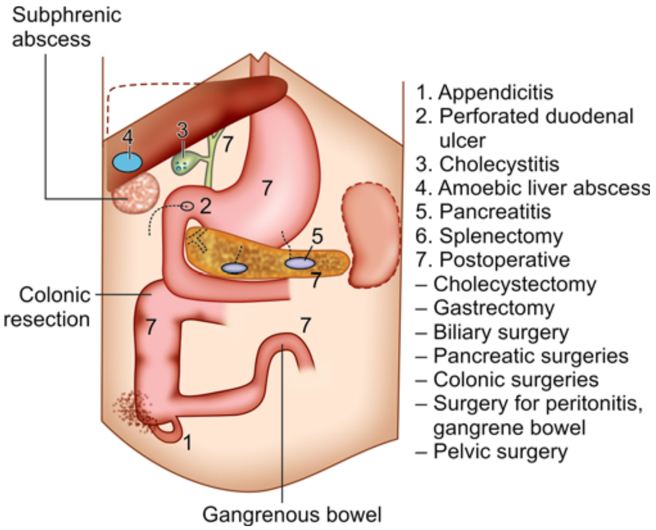
Subphrenic abscess
Colonic resection
Gangrenous bowel
1
2
3
4
5
7
1. Appendicitis
2. Perforated duodenal ulcer
3. Cholecystitis
4. Amoebic liver abscess
5. Pancreatitis
6. Splenectomy
7. Postoperative
– Cholecystectomy
– Gastrectomy
– Biliary surgery
– Pancreatic surgeries
– Colonic surgeries
– Surgery for peritonitis, gangrene bowel
– Pelvic surgery

Fig. 9.20: Common causes of subphrenic abscess both right and left. Subphrenic abscess is common on right side.

Pathology of Subphrenic Abscess

- During expiration, intra-abdominal pressure (especially in subphrenic area) is reduced and so this with capillary action and upward movement of the diaphragm make the peritoneal fluid to move upwards towards the diaphragm. This is the reason why there is higher incidence of subphrenic abscess and *subphrenic abscess is the second most common type of intra-abdominal abscess,* first one being the pelvic abscess.
- Subphrenic abscess is more common on the right side than on left (left anterior) because infective conditions are more common on the right side (like appendicitis/perforation/liver abscess/cholecystitis); and right paracolic gutter is wide and deep without any barrier (unlike left side paracolic gutter which is narrow and limited above by colophrenic ligament). Left posterior space is lesser sac in which infected fluid commonly collects following pancreatitis. Right subphrenic abscess causes elevation of the diaphragm. Pus collects in front or posterior aspect of the liver (right anterior or right posterior spaces) or often in both. Pus may be in a single cavity or may be multiloculated. Occasionally pus may get thickened to form a thick fibrous wall around leading into a chronic stage which causes generalised ill health. Pus gets well localised with a thick pyogenic membrane pushing pleura away and above and peritoneum away and below making subphrenic abscess essentially *extraserosal*. Because of the congestion/hyperaemia of the pleura and elevation of the diaphragm, there will be *sympathetic pleural effusion* on the side of the abscess. In severe acute subphrenic abscess, empyema thorax can also occur.
- If infective organism is gas forming one, abscess will contain pus with gas. So it is resonant above the liver dullness; dull above it due to collapsed lung/atelectasis; then again resonant on topmost due to normal lung (From below upwards—dull liver—resonant gas in abscess cavity—dull collapsed lung—resonant normal lung topmost). But it is rare presently.

A superior man is modest in his speech, but exceeds in his actions.

- *Postoperative surgeries*—mainly of gastric/biliary/colonic and emergency surgeries (peritonitis) are *becoming most common cause* of the subphrenic abscess.

Clinical Features

Barnard's aphorism (Harold Barnard):
"*Pus somewhere, pus nowhere else, pus under diaphragm*".
History relevant of the specific causes; history of any previous surgery:

- Fever with chills and rigors.
- Pain in right hypochondrium, epigastrium or lower thorax.
- Tenderness in right hypochondrium.
- Sympathetic right sided pleural effusion due to congestion and hyperaemia of the diaphragm.
- Collapse of the lung/basal atelectasis—right side may be evident. Tachypnoea and respiratory distress can be observed.
- Pain in the right shoulder due to irritation of phrenic nerve.
- Hiccough, tachycardia.
- ***Hoover's sign***: Scoliosis towards same side in subphrenic abscess.
- Wasting and anorexia is common.
- Occasionally tender mass in the right hypochondrium may be palpable.
- Tenderness over 11th intercostal space may be evident in right subhepatic space abscess.

Bacteria causing:

- *E. coli.*
- *Klebsiella.*
- Streptococci.
- Anaerobic organisms.

Differential diagnosis

- Amoebic liver abscess
- Pylephlebitis
- Empyema
- Pulmonary collapse

Investigations

- Plain X-ray chest and abdomen shows soft tissue shadow, pleural effusion, tenting of diaphragm, collapse of the lung. Fluoroscopy shows elevated right diaphragm with reduced right sided diaphragmatic movement.

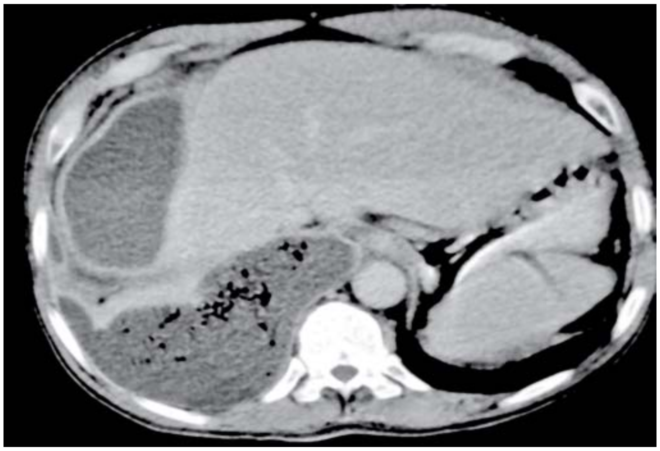

Fig. 9.21: CT scan of abdomen showing subphrenic abscess.

- U/S abdomen confirms the diagnosis.
- TC is high. There is raise in 'C' reactive protein.
- CT scan is very useful to find out the extent and relation. CT shows 'rind' sharp outer margin; septations; size, shape, extent, diaphragmatic position, pleural changes, displaced bowel and other organs and communications.
- Gallium 67 or iridium 111 isotope imaging.

Treatment

- **Antibiotics:** Ampicillin, metronidazole, gentamicin, cephalosporins.
- **Percutaneous drainage:** Initially ultrasound guided aspiration is useful. Here, under ultrasound or CT guidance, a catheter is placed and pus is drained. Catheters used are pigtail catheter, 16 French trocar catheter or sump catheter. 90% of subphrenic abscess are drained percutaneously.

Note:
Aspiration alone is not done in subphrenic abscess.

- **Open drainage is indicated** (Required in 10-20% of cases) in:

- When symptoms persist
- When abscess increases in size
- When there is thick pus or multiloculated pus or multiple abscesses
- Failure of/contraindications for catheter drainage
- Multiple abscesses
- Presence of foreign body/necrosis/haematoma/obstruction/fistula

  - Under G/A through anterior or posterior subcostal incision, abscess is approached extraperitoneally and extrapleurally (extraserosally), so as to prevent serosal contamination. Pyogenic membrane is opened, pus is collected for culture sensitivity. Pus is drained by breaking all loculi. Malecot's catheter is placed *in situ*.
  - Both *anterior and posterior extraperitoneal approaches* are used. Posterior approach allows dependent drainage; anterior approach (Clairmont approach for left side) is technically easier and allows proper assess.
  - *Transperitoneal approach* is needed whenever definitive procedure has to be undertaken like managing the existing fistula, treating the cause, removal of foreign body, multiple abscesses, active intra-abdominal pathology, etc.
- Antibiotics are continued. Malecot's catheter is removed once pus stops draining, usually after 14 days, which can also be confirmed by U/S abdomen or by X-ray with injection of dye.

Catheters used for drainage are:

- Malecot's catheter/wide Foley's catheter when open drainage is done
- Pigtail catheter
- Catheter placement through trocar under guidance (16 French trocar catheter)
- Sump catheter (double lumen tube which allows irrigation as well as suction drainage)

Note: Catheter is removed when pus drainage from the tube is <10 ml/day, normal total count, improvement in symptoms, U/S shows clearance of abscess cavity.

- Good nutritional supplement, respiratory physiotherapy are additional support required.

Complications

- Empyema.
- Respiratory arrest.
- Septicaemia.
- Sinus formation.
- Recurrence.
- Peritonitis.

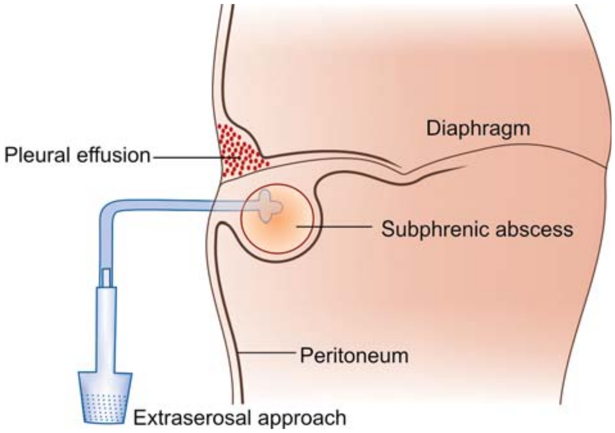

Fig. 9.22: Subphrenic abscess open drainage. Incision must be extraserosal. Malecot's tube is placed in the abscess cavity.

MESENTERIC CYSTS

Causes

1. Chylolymphatic.
2. Enterogenous.
3. Cysts of urogenital remnant.
4. Teratomatous dermoid cysts.
5. Other causes:
    - Traumatic haematoma and cyst formation.
    - Tuberculous cold abscess of mesentery.
    - Hydatid cyst of mesentery.

Chylolymphatic cysts are the most common one.

- It arises from congenitally misplaced lymphatic system. It does not have efferent communication into lymphatic system.
- Common in ileum, is a thin walled cyst with flat endothelium occasionally containing lymph or chyle, which is either milky or cream coloured.
- It is solitary and commonly unilocular, with loop of the bowel in front.
- It has got independent blood supply, i.e. not from the adjacent bowel loop. So enucleation can be done without resecting bowel.
- They are often called as *non-neoplastic mesothelial cysts* as it is lined by mesothelium. It is common in small bowel mesentery (60%); but can occur in colonic mesentery (40%). It occasionally may contain chyle even though serous fluid is the common content. It is common in females.
- Cyst is *enucleated* surgically (*standard therapy*); aspiration should not be done; in a large cyst internal drainage of cyst into the peritoneal cavity is advocated by few. Cyst wall should be sent for histology to rule out neoplastic condition.

Enterogenous type arises as a diverticulum or duplication from the adjacent bowel.

Hence it is a thick walled cyst (contains all layers of the bowel) and receives its blood supply from the adjacent bowel (not independent). So resection of the adjacent bowel along with the cyst is essential. Enucleation is contraindicated.

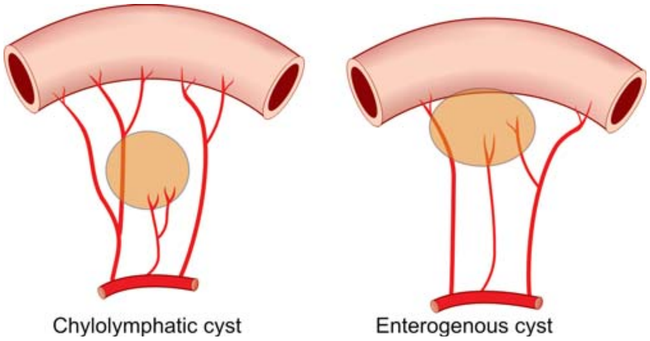

Fig. 9.23: Mesenteric cyst.

Clinical Features

- It presents as a painless abdominal swelling in umbilical region, smooth, fluctuant, not moving with respiration.

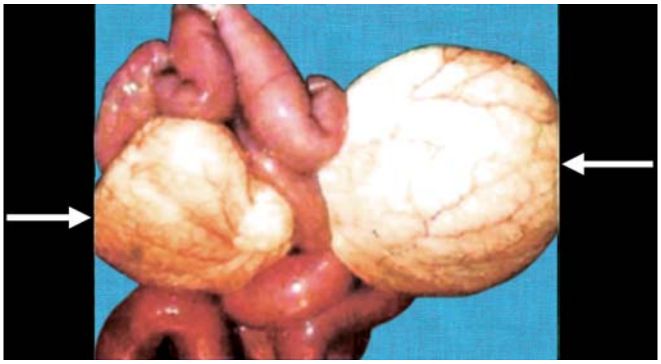

Fig. 9.24: Mesenteric cyst.

- It is common in 2nd decade, often in childhood.
- It is freely mobile, mobile in the direction perpendicular to the line of mesentery.
- There is a band of resonance in front of the cyst.
- Presentation as complications.

Tillaux's triad

- Soft, smooth swelling in the umbilical region
- Freely mobile in a direction perpendicular to mesentery
- Zone of resonance all around

Complications of mesenteric cysts

- *Torsion of cyst* can lead to volvulus of the adjacent bowel
- *Rupture of the cyst*
- *Haemorrhage* into the cyst
- *Infection*—patient presents with acute painful swelling in umbilical region

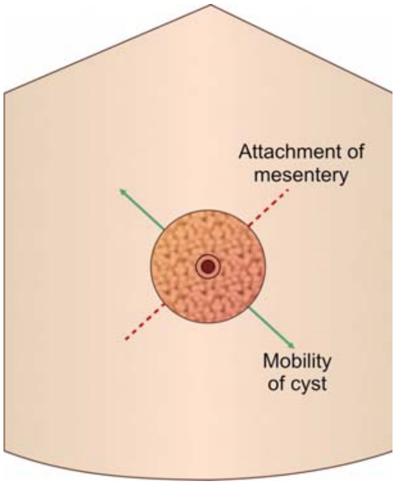

Fig. 9.25: Mobility of mesenteric cyst.

Investigations

- U/S abdomen.
- Barium meal X-ray.
- Plain X-ray abdomen.
- CT scan of the abdomen is diagnostic.

Differential Diagnosis

- Hydronephrosis.
- Omental cysts.
- Tuberculosis.

Treatment

- In chylolymphatic cyst—*enucleation.*
- In enterogenous type, removal of the cyst with resection of adjacent bowel is done, because blood supply is from adjacent bowel.

MESENTERIC PANNICULITIS

- It is an inflammatory disease of mesentery with mesenteric *fat necrosis, acute/chronic inflammation and fibrosis*.
- Enormous thickening of the mesentery near its root with multiple discrete nodules are typical. Venous and lymphatic obstruction of mesentery develops, causing bowel oedema, congestion and ischaemia.
- It commonly involves small bowel mesentery; but can involve colonic mesentery also. Spontaneous resolution is known to occur.
- It is common in middle aged men. Recurrent abdominal pain, features of obstruction, distension of abdomen, tender resonant mass abdomen are the features.
- ESR and CRP is elevated. CT is useful with features like—pseudocapsule with well delineated fat margin; normal adipose tissue surrounding the mesenteric vessels—*fat ring sign*; no vascular deviation; intra-abdominal mass displacing bowel without infiltration. Laparoscopic biopsy of mesentery is conclusive.
- *Differential diagnoses* are—peritoneal carcinomatosis, carcinoid tumour, mesenteric or retroperitoneal sarcoma, mesenteric tuberculosis and intestinal obstruction.
- *Treatment:* Usually conservative using corticosteroids and immunosuppressive drugs. Laparotomy is done only when diagnosis is in doubt or intestinal obstruction is suspected.

ACUTE MESENTERIC LYMPHADENITIS

- It is nonspecific acute inflammation of the mesenteric lymph nodes (*Nurse's syndrome*) usually due to *Yersinia enterocolitica* infection; often in 30% of cases respiratory infection will be present. Occasionally virus, parasite, fungal infections may also be the cause. Other regional lymph nodes (axilla, groin, neck) are also often involved and enlarged.
- Mesenteric nodes in ileocaecal region are inflamed, enlarged often clinically palpable in right iliac fossa. Lymph nodes are red but not attached to mesenteric leaves or peritoneum. Appendix is normal.
- It is self-limiting recurrent disease.
- It is common in children and young females.
- Pain in right iliac fossa, slight fever, vomiting, abdominal colicky, but normal bowel habits; tender nodes may be palpable in RIF.
- Leucocytosis is common but regresses rapidly. US abdomen will be normal.
- Condition often mimics acute appendicitis.
- *Treatment:* Bed rest, re-assurance, analgesics, follow-up. Laparoscopy and appendicectomy is often needed to diagnose and to prevent missed appendicitis.

MESENTERIC MALIGNANCY

- The most common mesenteric malignancy is *secondaries*. GI adenocarcinoma is the most common site of origin. Carcinoid also can cause mesenteric metastasis. Mesentery is involved by either direct spread or through lymphatics or transperitoneal route. Distortion, fixation, desmoplastic reaction with intestinal obstruction are the features.

Mesenteric Desmoid Tumour

- It is the most common primary tumour of mesentery.
- It is common in familial adenomatous polyposis and Gardner's syndrome.
- FAP shows 75% intra-abdominal desmoid with 75% of them in mesentery.
- Mesenteric desmoid is aggressive tumour causing infiltration of adjacent structures.
- *Complications* are—intestinal obstruction, bowel necrosis and perforation, hydronephrosis, aortic erosion and bleeding.
- 10-year survival rate is 65%. But recurrence rate is also high—75%.
- *Investigations:* CT abdomen, laparoscopy and biopsy, colonoscopy in case of FAP, to look for other features of Gardner's syndrome.
- *Treatment:* Even though surgery is the treatment, in view of high recurrence rate and inability to clear tumour surgically, drug therapy is becoming popular many times. Drugs used are—sulindac; antiestrogen drugs, imatinib mesylate orally (inhibits tyrosine kinase receptor for PDGF and c-kit).

Mesenteric Lymphoma

It is also *not uncommon* as mesenteric fold contains plenty of lymph nodes. It is commonly NHL type. Other regional lymph nodes like of para-aortic, axillary, neck region may be enlarged. Liver and spine may be involved. CT abdomen, peripheral smear, laparoscopy and node biopsy, CT chest, bone marrow biopsy are investigations done. Treatment is usually chemotherapy.

MESENTERIC TRAUMA

Mesenteric injury is commonly seen in blunt abdominal trauma; traction injury or seatbelt injury causes mesenteric tear. Penetrating injury can also cause mesenteric injury.

Presentations

Features of haemoperitoneum—shock, pallor, abdominal distension and pain, guarding and rigidity.

Types

1. *Transverse tear* in mesentery causes not only more bleeding but also causes adjacent bowel ischaemia. So laparotomy and resection of bowel is needed.
2. *Longitudinal tear* can be sutured using interrupted absorbable sutures (vicryl) after haemostasis.

Investigations

- Us abdomen; diagnostic peritoneal lavage.
- CT abdomen.
- Always evaluate for associated injuries—fractures, thoracic and head injuries.
- Haematocrit, electrolyte estimation, blood grouping.

Treatment

Emergency laparotomy and resection of bowel or suturing of the mesentery.

PERITONEAL MALIGNANCY

The most common peritoneal malignancy is *secondaries*. Primaries causing peritoneal secondaries are—*GIT*: stomach, colon, pancreas, small bowel, liver, lower oesophagus; *Genitourinary*: ovarian tumour commonly; *Extra-abdominal*: Breast.

Peritoneal carcinomatosis is diffuse metastatic deposits all over parietal and visceral peritoneum.

Primary Peritoneal Malignancy

The most common is malignant mesothelioma; sarcomas are rare type.

Malignant mesothelioma: It is very aggressive primary type with survival rate of 6-12 months. Presentation is abdominal pain, ascites, weight loss, and cachexia. Omentum is involved with malignant nodules presenting like a nodular abdominal mass. Peritoneal cavity is studded with nodular deposits. Entire peritoneal cavity is involved often involving the organs like

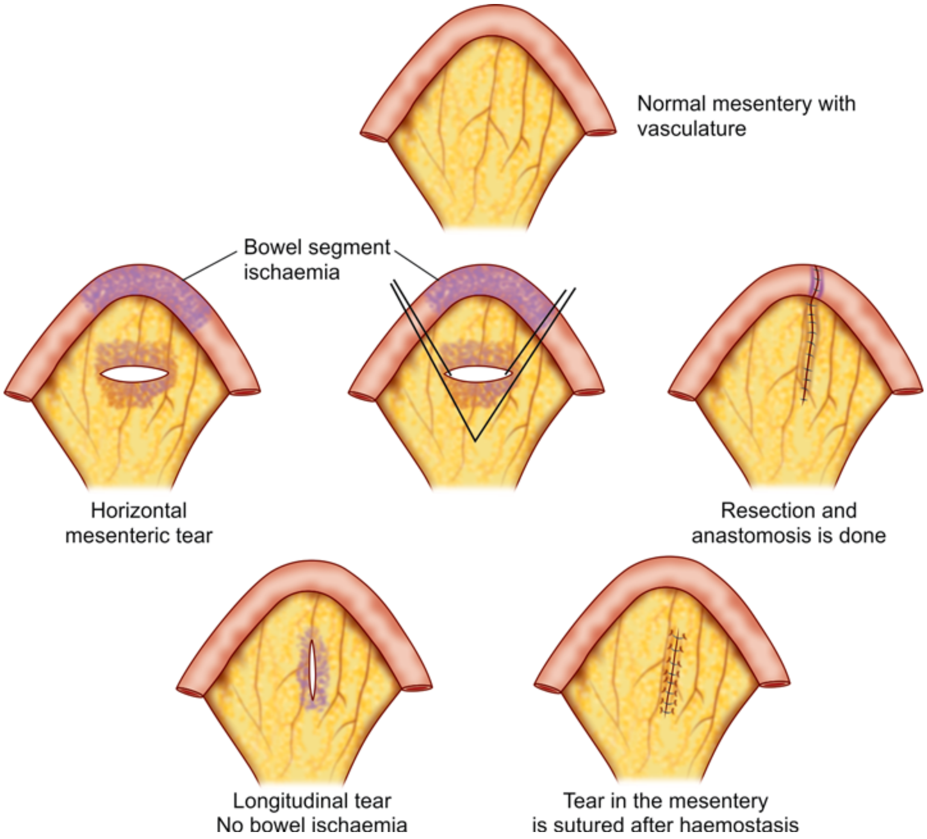
Normal mesentery with vasculature
Bowel segment ischaemia
Horizontal mesenteric tear
Resection and anastomosis is done
Longitudinal tear
No bowel ischaemia
Tear in the mesentery is sutured after haemostasis

Fig. 9.26: Mesenteric tears due to trauma can be transverse or longitudinal; transverse tear needs bowel resection; longitudinal tear needs just suturing.

I am a great believer in luck and I find the harder I work, the more I have of it.

liver or their surfaces. It is often difficult to differentiate it from peritoneal carcinomatosis. However, extra-abdominal spreads to lungs and bone is uncommon in malignant mesothelioma. CT abdomen, ascitic fluid analysis are the diagnostic methods. Chemotherapy, intraperitoneal hyperthermic perfusion of mitomycin C, intraperitoneal chemotherapy of cis platin, surgical peritonectomy—are *treatment* options.

Pseudomyxoma Peritonei

- It is a rare malignant condition that develops due to ruptured malignant ovarian cyst or adenocarcinoma of appendix. Peritoneum is coated with mucus secreting tumour tenacious gel like mucus with cystic masses. It is common in old aged females.
- Vague abdominal pain with distension and dullness which is not shifting are typical findings. CT scan shows bowel displacement with fluid content in the peritoneal cavity.
- *Treatment:* If origin is known then it is treated accordingly. In patients operated by appendicectomy, removal of all mucus gel with right hemicolectomy with omentectomy is done. If ovarian tumour is the cause then panhystrectomy with omentectomy is done. In cases with unknown cause, after laparotomy, panhystrectomy, right hemicolectomy, omentectomy are done. Intraperitoneal chemotherapeutic agents like 5 FU, mitomycin C, oxaliplatin; intraperitoneal instillation of mucolytic agents like dextran, urokinase after surgery is beneficial.
- It shows better prognosis compared to mesothelioma and carcinomatosis.

Peritoneal carcinomatosis (carcinoma peritonei)

- Entire peritoneal cavity is studded with carcinomatous nodules—both parietal and visceral peritoneum
- Haemorrhagic ascites, interloop bowel adhesions, hard studded nodules in the omentum are common. Intestinal obstruction can occur
- It may be primary peritoneal carcinomatosis (mesothelial) or secondary from GIT, ovary, breast, pancreas
- Liver secondaries are often present
- It mimics peritoneal tuberculosis, acute pancreatitis, and multiple lymphomatous nodules, multiple peritoneal hydatids
- As malignant ascitic fluid contains high immunoglobulin, peritonitis is not common
- Instillation of radioactive gold and chemotherapy are the treatment
- It has got poor prognosis

OMENTAL CYST

- It arises from obstruction of omental lymphatics either congenital or acquired.
- It is unilocular or multilocular cyst lined by endothelium.
- It is common in children.
- When it attains large it presents as mass in central or lower abdomen which is smooth well localised mobile often moves with respiration.
- It can cause torsion, rupture and infection. Extrinsic compression of bowel can occur.
- US/CT shows cyst with internal septations.
- Differential diagnoses are—mesenteric cyst, retroperitoneal cyst and desmoid tumour.
- Treatment is laparotomy and cyst excision which can also be done laparoscopically.

OMENTAL TORSION

- Torsion of greater omentum is twist of omentum along its longitudinal axis; it usually occurs on right side of omentum.
- Primary torsion is torsion occuring without any existing cause.
- Secondary torsion is occurring due to some cause like omentocele in a hernial sac or tumour or adhesion.
- Torsion is more common in men (2:1) in 5th decade.
- Sudden severe abdominal pain, mainly right sided often mimicking acute appendicitis, cholecystitis, ureteric stone or twisted ovarian cyst.
- Often signs of peritonitis may be present.
- CT abdomen is diagnostic. Total count will be raised.
- Treatment is laparotomy and omentectomy which can also be done laparoscopically.

OMENTAL TUMOUR

- Metastatic omental deposits are common; usually from GIT cancers.
- Primary neoplasm of omentum is rare. It may be soft tissue tumour arising from fat of omentum.

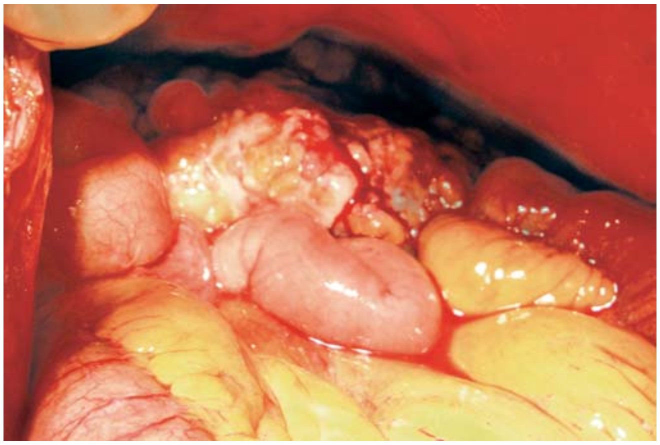

Fig. 9.27: Omental secondaries.

Chapter 10 Abdominal Tuberculosis

> *In man the balance between immunity and susceptibility to tuberculosis is delicately adjusted: There is a small factor of safety.*
>
> **—Milton Rosenau,** 1927

CHAPTER OUTLINE

- Abdominal Tuberculosis
- Ileocaecal Tuberculosis
- Ileal Tuberculosis
- Peritoneal Tuberculosis
- Tuberculous Mesenteric Lymphadenitis
- Ano-recto-sigmoidal Tuberculosis
- Tuberculosis of the Omentun

ABDOMINAL TUBERCULOSIS

It is common in India and developing countries.

It is the 6th most common type of extrapulmonary tuberculosis. Its incidence is high in HIV infected patients.

Types

1. *Intestinal*
   Ileocaecal region
   - Ulcerative—60%.
   - Hyperplastic.
   - Ulcero-hyperplastic.

   Ileal region, commonly:
   - Stricture type.
2. *Peritoneal tuberculosis*
   a. Acute.
   b. Chronic.
      i. Ascitic type.
      ii. Encysted (loculated) type.
      iii. Plastic (fibrous/adhesive) type.
      iv. Purulent type.
3. *Tuberculosis of mesentery and its lymph nodes.*
4. *Ano-recto-sigmoidal*—present as fistula, fissure, abscess, mass.
5. *Involvement of liver, spleen and other organs as a part of miliary tuberculosis.*
6. *Tuberculosis of the omentum.*
7. *Rare types:* Oesophageal (0.2% of abdominal tuberculosis)/gastroduodenal (1% of abdominal tuberculosis)/retroperitoneal tuberculosis. Because of hydrochloric acid (acid media) gastric tuberculosis is rare.

Note:

- Chronic peritoneal tuberculosis may be associated with pleural effusion and pericardial effusion.
- Tuberculosis is not common in stomach, duodenum and jejunum.
- *Diffuse tuberculous colitis* is less commonly seen (4%), mimics ulcerative colitis in every respect even in colonoscopy. Patient recovers well with antituberculous drugs.
- *Intestinal tuberculosis is called as* ***Koenig's syndrome*** *(1892).*

Modes of Spread of Abdominal Tuberculosis

- By ingestion
  - Ingestion of food contaminated with tubercle bacilli causing *primary intestinal tuberculosis*.

Site spectrum	***Ileocaecal and terminal ileum—most common site***
• Luminal GI tuberculosis 75%	Jejunum and colon—less common
• Abdominal solid organ tuberculosis—liver, spleen, pancreas	Stomach, duodenum, oesophagus—rare
• Peritoneal tuberculosis—*adhesive* (cocoon/doughy/rolled up omentum)/*ascitic* (wet)—loculated. It could be acute or chronic	*Pathology:* • In malnourished poor resistance patients—ulcerative and or stricture types
• Abdominal lymph node tuberculosis—mesenteric/periportal/para-aortic/retroperitoneal. It could be caseating/hyperplastic/calcified	• In well nourished with good resistance—hyperplastic and nodular types • Combination types—ulcerostricture type; ulcero-hyperplastic type

*Phthisis ia an ulcer of the lung that consumes the whole body.***—Bernard de Gordan,** 1495

Different clinical presentations of abdominal tuberculosis

Clinical presentation may be:

- Acute
- Acute on chronic
- Chronic

Constitutional symptoms:

- Low-grade fever, malaise, night sweats, anaemia, weight loss
- Observed in 30% of patients

Abdominal tuberculosis is 6th most common form of tuberculosis

It may be associated with HIV/lymphoma/carcinoma

Types:

- *Ulcerative*—diarrhoea and malabsorption
- *Stricture*—subacute or acute intestinal obstruction
- *Hyperplastic*—mass abdomen (RIF) and obstruction
- *Ascites*—generalised distension of abdomen
- *Peritoneal*—abdominal cocoon; vague abdominal pain, parietal peritoneal thickening as doughy abdomen; mass abdomen
- *Mesenteric*—tabes mesenterica, obstruction, mass
- Associated tuberculosis of other organs is seen in 30% of cases

Atypical presentations:

- Lower GI bleed, fistula-in-ano, PID like pain, gastric disease symptoms, dysphagia, GI fistulae, perforation

Tuberculosis of liver:

- Miliary type
- Granuloma/tuberculoma
- Like liver abscess/intra-hepatic calcification
- Obstructive jaundice
- PUO/altered LFT

Tuberculosis of pancreas:

- Like or part of miliary tuberculosis
- Common in immunocompromised
- Usually presents as acute or chronic pancreatitis
- Pancreatic mass or abscess may develop
- Can mimic malignancy

Tuberculosis of spleen:

- Disseminated or miliary form
- Can present as PUO with hepatosplenomegaly
- Can occur as multiple abscess

  - Ingestion of sputum containing tuberculous bacteria from primary pulmonary focus causing *secondary intestinal tuberculosis.*
- Haematogenous spread from tuberculosis of lungs.
- From neck lymph nodes (tuberculous cervical lymphadenitis—5-10%) through lymphatics.
- From fallopian tubes by retrograde spread to involve peritoneum (10%).

ILEOCAECAL TUBERCULOSIS

Most common site of abdominal tuberculosis due to presence of Peyer's patches; and stasis of luminal contents favoured by ileocaecal valve.

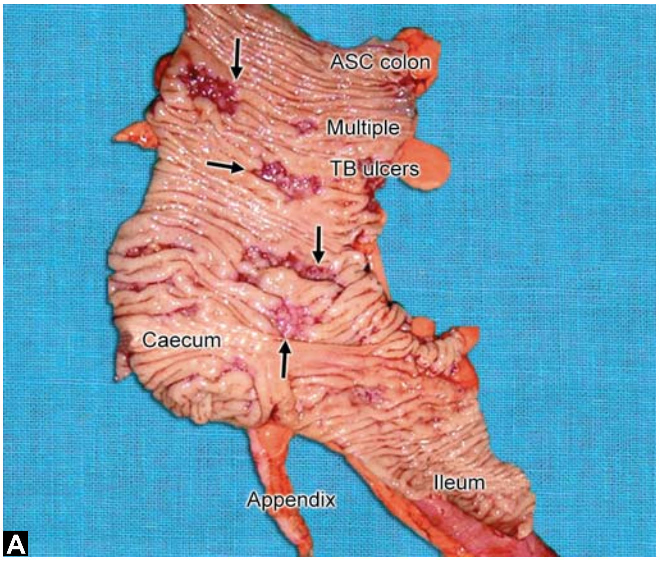

Fig. 10.1A

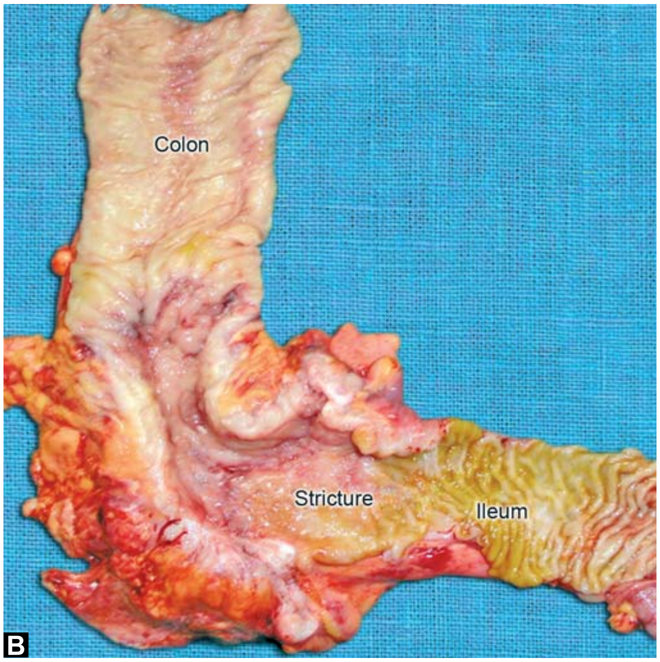

Fig. 10.1B

Figs 10.1A and B: Ileocaecal tuberculosis. (A) Note the multiple transverse undermined ulcers. (B) Note the ileocaecal tuberculosis with stricture.

Causative organism

Mycobacterium tuberculosis

- Acid fast 20% H_2SO_4
- Alcohol fast
- Gram neutral

Histology

- Epithelioid cells diagnostic
- Langhan's giant cells
- Features of granuloma
- Caseating necrosis

- *It is presently due to Mycobacterium tuberculosis,* earlier used to be due to *Mycobacterium bovis.* Mode of infection may be direct or blood spread, usually from lungs.
- Atypical mycobacteria can spread directly.
- *Mycobacterium avium* spreads through lymphatics.

Types

- *Ulcerative—most common 60%. Circumferential transverse often multiple 'girdle' ulcers*—with skip lesions. It is common in old, malnourished people. Long-standing ulcers cause fibrosis and later stricture formation. Stricture (*Napkin ring stricture*) is common in ileal part. Often related intestinal nodes are also involved with caseation, abscess (cold) formation. Bowel adhesions are common. Patient mainly presents with diarrhoea, blood in stool, loss of appetite and reduced weight.

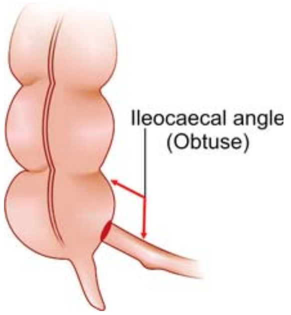

Fig. 10.2: Obtuse ileocaecal angle—ileocaecal tuberculosis.

- *Hyperplastic*: Fibroblast reaction in submucosa and subserosa causing thickening of bowel wall and lymph node enlargement, leading to nodular mass (tumour-like) formation. It is 10% common, less virulent, with adequate host resistance, seen in young well nourished individuals. It is common in caecal part. It causes extensive chronic inflammation, fibrosis, bowel adhesions, nodal enlargement, often presents with mass in the right iliac fossa. Caseation necrosis is not common. When present as a mass, it can cause subacute intestinal obstruction. It is commonly primary intestinal tuberculosis. There is no primary focus in the lungs.
- *Ulcerohyperplastic*—30%.

Clinical Features

- Abdominal pain is the most common symptom (90%). It is dull in mesenteric type; colicky in intestinal type.
- Common in 25-50 years age group. Equal in both sexes.
- Anaemia, loss of weight and appetite (80%).
- Diarrhoea—10-20%.
- Fever—50-70%.
- *Mass in right iliac fossa*, (35%) which is hard, nodular, nonmobile, nontender with impaired resonance, which may mimic carcinoma caecum. Subacute obstruction can occur.
- Ileocaecal tuberculosis can be associated with adenocarcinoma of caecum, or large bowel lymphoma or HIV.
- Often ileocaecal TB can cause intestinal obstruction (20%).

Ileocaecal region is common site due to:

- Stasis
- Abundant Peyer's patches—organisms get trapped in Peyer's patches
- Bacteria contact time with mucosa is more
- M cells in Peyer's patches phagocytose bacilli and transfer to host cells
- Liquid content in the region
- Increased rate of fluid and electrolyte absorption
- Minimal digestive activity

Differences between ulcerative and hyperplastic types of ileocaecal tuberculosis

	Ulcerative type (60%)		*Hyperplastic type (10%)*
1.	Secondary to pulmonary tuberculosis	1.	Primary GIT tuberculosis, could be due to bovine bacilli
2.	Virulent organism	2.	Less virulent organism
3.	Poor body resistance, old people	3.	Good body resistance, young individual
4.	Multiple transverse ulcers commonly in the ileum, often in the caecum	4.	Chronic granulomatous lesion in the ileocaecal region
5.	Clinically presents with diarrhoea, bleeding P/R, loss of appetite and reduced weight	5.	Presents as a mass in right iliac fossa
6.	Complications: Stricture, intestinal obstruction	6.	Complication: Subacute intestinal obstruction
7.	Chest X-ray shows primary lesion	7.	No primary lesion in chest X-ray
8.	Barium study shows ileal strictures with hypermotility	8.	Barium study—pulled-up caecum, obtuse ileocaecal angle

Whilst meagre Pthisis gives a silent bow, her strokes are sure, but her advances slow. No loud alarms, nor fierce assaults are shown. She starves the fortess first, then takes the town. **—Samuel Garth**

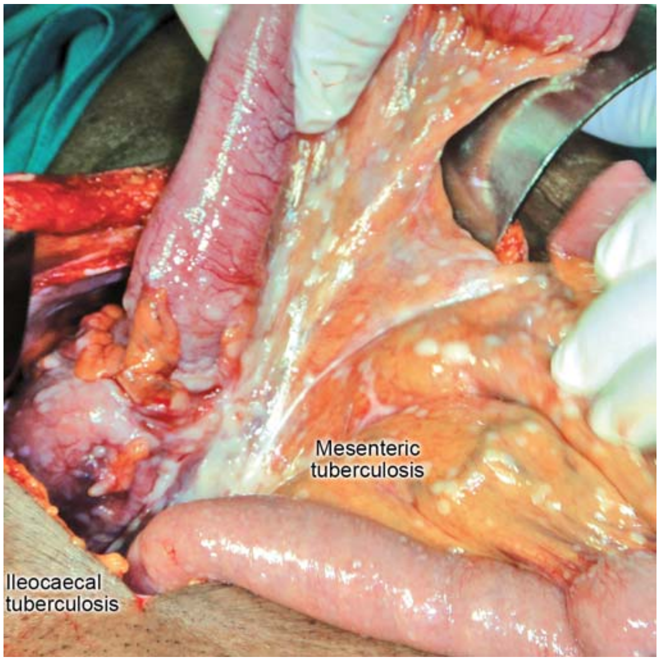

Fig. 10.3: Ileocaecal and mesenteric tuberculosis.

Differential diagnosis

- Carcinoma caecum
- Ameboma
- Appendicular mass
- Ectopic kidney
- Retroperitoneal tumour
- Lymph node mass
- Psoas abscess
- Crohn's disease

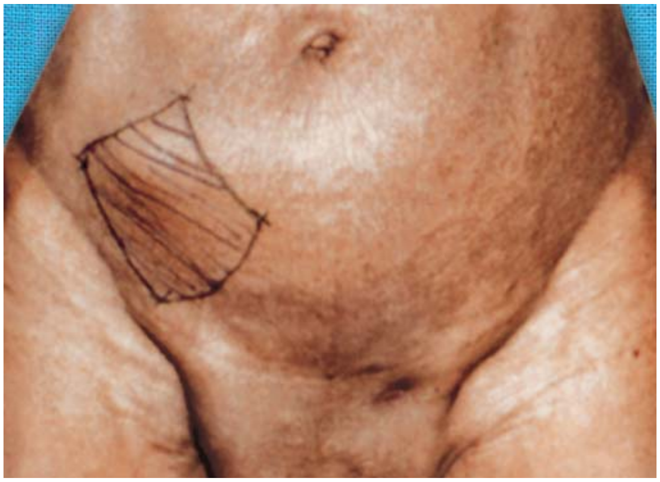

Fig. 10.4: Ileocaecal tuberculosis or tuberculous mesenteric lymphadenitis can present as mass in the right iliac fossa which is usually nontender, nonmobile, hard/firm resonant/impaired resonant.

Investigations

- Chest X-ray to find out primary focus.
- Mantoux test, ELISA (90%), SAFA (soluble antigen fluorescent antibody—80%), serum IgG.
- ESR is raised.
- U/S abdomen to see ascites, caecal thickening, nodal status and other organs.
- Plain X-ray abdomen, if presentation is of intestinal obstruction. It often shows calcification. It shows calcified lesion in the bowel; in the lymph node; in the liver (calcified granuloma). Perforation when it occurs (rare) shows gas under diaphragm.

Barium study X-ray (Enteroclysis followed by barium enema or barium meal follow through X-ray) (efficacy—75%)

- Pulled up caecum, conical caecum, pulled down hepatic flexure
- Obtuse ileocaecal angle
- Hurrying of barium due to rapid flow and lack of barium in inflamed segment (*Steirlin sign*)
- Narrow ileum with thickened ileocaecal valve *(Fleischner sign) (Inverted umbrella sign)*
- Calcifications
- Incompetent ileocaecal valve, ileocaecal spasm
- Ulcers and strictures in the terminal ileum and caecum—*Napkin lesions*
- Earliest signs are—increased transit time; hypersegmentation (chicken intestine); flocculation of barium
- Other signs are—persistent narrow stream (string sign); multiple strictures with enormous dilatation of proximal ileum (mega ileum); straightening of ileocaecal junction with 'goose neck' deformity

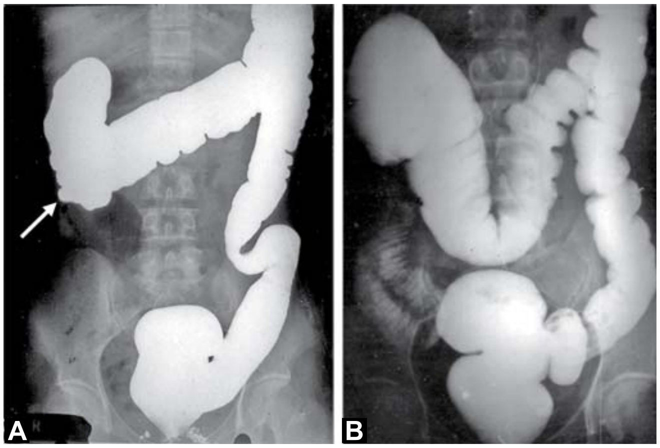

Figs 10.5A and B: Pulled-up caecum, obtuse ileocaecal angle—ileocaecal tuberculosis in barium study X-ray.

- Colonoscopy is of value to rule out carcinoma. It is *easiest and most direct method* in establishing the diagnosis. Colonoscopy shows mucosal nodules or ulcers; caecal and ileal strictures; deformed ileocaecal valve; mucosal oedema and pseudopolyps and occasionally diffuse colitis. Biopsy can be taken to confirm the diagnosis. Tissue culture or tissue PCR can be done.
- Capsule endoscopy is also useful to see small intestinal (tuberculous) pathology in difficult cases.
- FNAC of palpable mass.
- Laparoscopy is very useful method of investigation. Transabdominal peritoneoscopy is visualisation of the peritoneal cavity using endoscope through a small incision in the abdomen. It aids in visualisation, to collect ascitic fluid for analysis and to take biopsy. Biopsy can be taken from omentum, peritoneum, nodes and suspected areas. Ascitic fluid can be collected for analysis.

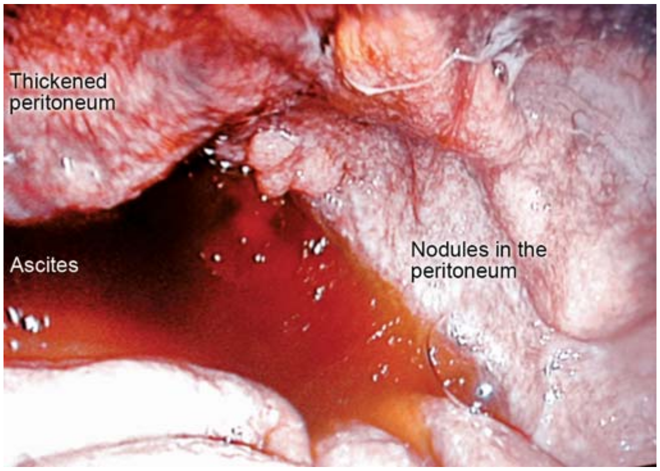

Fig. 10.6: Laparoscopic view of abdominal tuberculosis showing peritoneal nodules and thickening, ascitic fluid, bowel surface tubercles. Fluid for analysis, peritoneal, omental biopsy can be taken.

- Abdominal CT scan is better and more reliable.
- PCR assay of endoscopically biopsied tissue or of ascitic fluid. DNA—PCR can detect 1-2 organism or 8 fg of mycobacterial DNA. Positive PCR signifies infection but need not be active disease.
- Stool culture for AFB—not useful. Presence of tuberculous bacteia in stool does not signify disease.
- Blind percutaneous needle peritoneal biopsy using *Cope's/ Abraham's needle* is also practiced.
- Ascitic tap fluid analysis.
- Anticord factor antibody to differentiate it from Crohn's disease.
- Biochemical assay—proteins, ADA (Adenosine deaminase activity) and interleukins. ADA is specific and sensitive marker with 95-98% sensitivity and specificity. Its value more than 33 IU/litre in ascitic fluid and more than 42 IU/ litre in serum is significant. Gamma interleukin is two times more expensive than ADA.

Note:

Serum ADA > 54 U/l; asctic ADA > 33 U/l; ascitic fluid: serum ADA > 0.985.

- **Bactec MGIT broth culture:** It is *Mycobacterium* growth indicator tube culture system containing Middlebrook Broth with silicon base; fluorescent technology is used.

Method used to isolate tuberculous bacteria	***AFB/ml specimen***	***Time taken for reporting***	***Sensitivity***	***Specificity***
Ziehl-Neelsen stain microscopy	5,000-10,000/ml	1-2 hours	75%	85%
Lowenstein Jenson solid media culture	100/ml	4-8 weeks	80%	90%
Bactec MGIT broth culture	10/ml	2-6 weeks	94%	92%
COBAS Taq-man PCR	Less than 10/ml	24-48 hours	97%	99%

Ascitic fluid in abdominal tuberculosis

- Exudate with protein level > 2.5 g/dl
- Serum-ascitic fluid albumin gradient is < 1.1
- Lymphocyte predominant cells with count as high as 4000/ mm^3 > 250/cumm
- AFB in ascitic fluid is seen only < 3% cases
- ADA (Adenosine deaminase activity) in ascitic fluid (95% specificity and 98% sensitivity)
- Specific gravity >1.016
- Glucose < 30 mg
- Decreased pH
- LDH more than 90 units/litre

Ultrasound features observed in abdominal tuberculosis

- Thickened bowel wall, mesentery, omentum, peritoneum
- Loculated ascites with fine septae
- Interloop ascites with alternate echogenic and echofree areas—Club-sandwich appearance
- Bowel loop radiates from its mesenteric root—*stellate sign*
- Mesenteric thickness *more than 15 mm*
- Hepatosplenomegaly
- Lymph nodal enlargement matted
- Pulled up caecum presenting with a mass in subhepatic region and also sonologically visible—"pseudokidney sign"
- Concentric uniform mural thickening

CT scan in abdominal tuberculosis

- It is very useful and reliable investigation
- It is done with oral contrast—CT enteroclysis
- Findings are:
  - Thickened bowel wall, thickened peritoneum
  - Ileocaecal valve thickening
  - Enlarged/necrosed/matted mesenteric nodes often with cold abscess
  - Adhesions
  - Mesenteric thickening and nodules
  - Nodules in the peritoneum/solid organs like liver
  - Adhesions in the bowel/stricture/dilatations of the bowel/ features of obstruction
  - Loculated ascites
- CT guided FNAC/biopsy/aspiration of fluid can be done

Complications of abdominal tuberculosis

- Obstruction—20%
- Malabsorption, blind loop syndrome
- Dissemination of tuberculosis to other areas of abdomen as well as extra-abdominal sites
- Faecal fistula
- Cold abscess formation
- Haemorrhage, perforation (rare)

Note:

Perforation is **rare** in GIT tuberculosis **but can occur**.

A man carries success or his failure with him; it does not depend on outside conditions.

Treatment

Drugs

- INH; rifampicin; pyrazinamide; ethambutol—*first line drugs*. In drug resistance cases second line drugs are needed. WHO recommends 6-9 months course. Commonly patient presents with complications and late presentations are also common in many. So, more often treatment for one year may be required in these patients. Recurrent abdominal tuberculosis has got high mortality and difficult to manage.

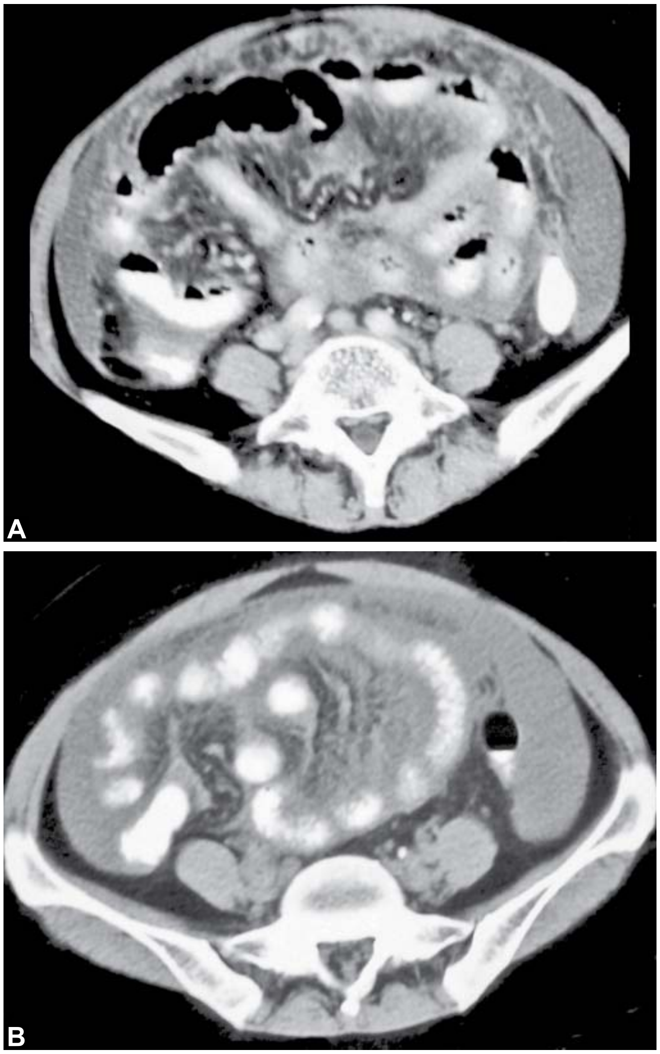

Figs 10.7A and B: CT scan of abdomen showing features of abdominal tuberculosis.

- As they are malnourished supportive therapy in the form of TPN, blood transfusion in preoperative as well as postoperative period is required. Often steroid is used to prevent adhesions along with antituberculous drugs.

Surgery

Indications for surgery

- Intestinal obstruction.
- Severe haemorrhage.
- Acute abdominal presentation like perforation.
- Intra-abdominal abscess formation or fistula formation.
- Uncertain diagnosis.

Surgeries are:

- *Limited ileocaecal resection* (with 5 cm margin) is the surgical therapy of choice for ileocaecal tuberculosis. This may be done in initial period depending on the obstructive and other presentations. Often during therapeutic period, healing with fibrosis causes stricture and obstruction in 3-6 weeks after drug therapy. Patient during this time needs limited ileocaecal resection.
- In single ileal stricture—stricturoplasty may be done. But if bowel wall is oedematous and friable then resection would be the ideal choice.
- In multiple strictures resection of ileum and anastomosis is done (ideal). Multiple strictures with long segment gaps between each can be treated by multiple stricturoplasty. But viability of the sutured area should be ensured. Resection is better option for stricture within 10 cm of ileocaecal valve.
- In perforation of ileal bowel, resection and anastomosis is done. Biopsy from perforation site and closure can be done in early perforations but chances of leak and fecal fistula formation is high (due to closure of perforation over a diseased bowel) and so resection is better option. In severely contaminated peritoneum, resection and exteriorisation is done. Bowel continuity is maintained after proper antituberculous chemotherapy and proper nutritional improvement.
- During therapy, if patient develops ileocaecal obstruction, ileotransverse colon anastomosis (bypass) can be done. But this is not a good procedure as it causes blind loop and tuberculous focus is retained. Now this bypass is used only in patient presenting with acute intestinal obstruction with poor general condition and in high-risk group patients as a life-saving procedure. A definitive second stage right hemicolectomy is needed in these patients at a later period.
- Adhesive obstruction may be released through laparoscopic adhesiolysis. It is often technically difficult to release dense adhesions even by open method.
- Drainage of intra-abdominal abscess, perianal abscess and treatment for tuberculous fistula-in-ano is done when necessary.

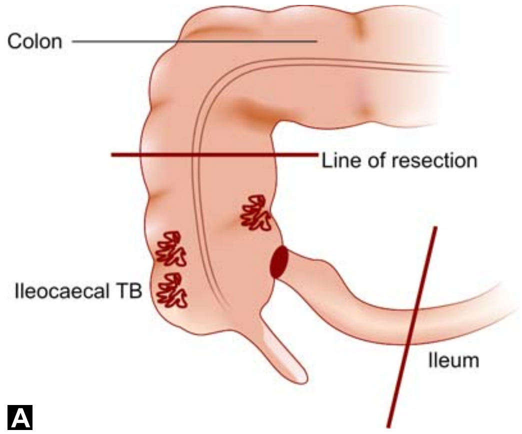

Fig. 10.8A

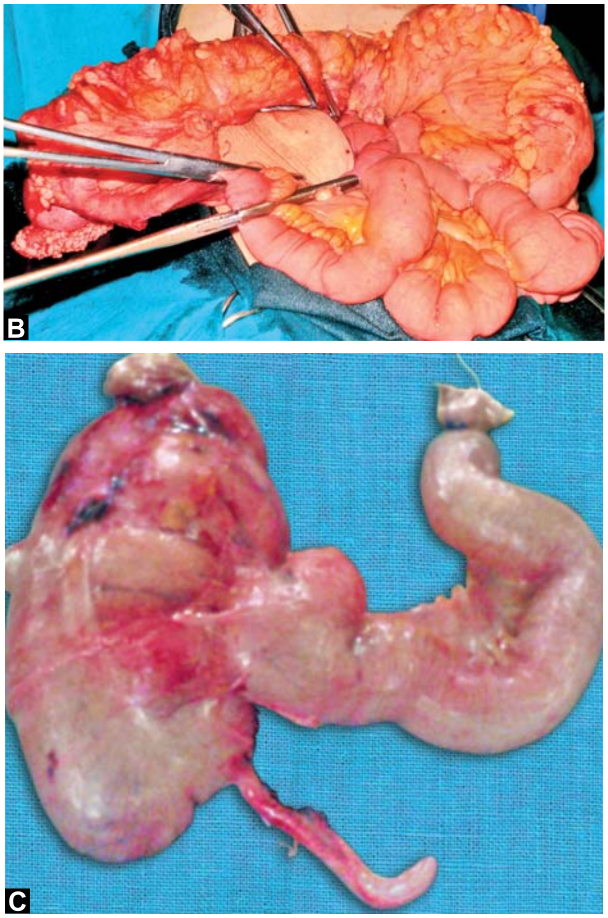

Figs 10.8B and C

Figs 10.8A to C: Ileocaecal limited resection with 5 cm bowel margin is the treatment of choice for ileocaecal tuberculosis.

ILEAL TUBERCULOSIS

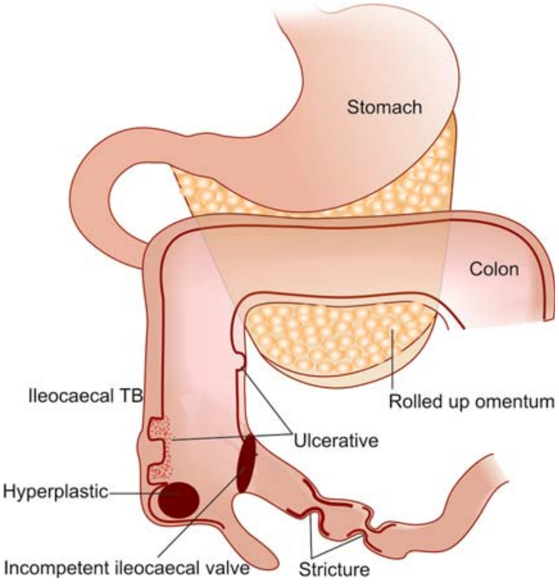

Fig. 10.9: Ileocaecal tuberculosis with rolled up omentum. Note the sites and tubercles in ileocaecal valve, which will be incompetent due to fibrosis.

- It is usually *stricture* type.
- It may be multiple.
- It presents usually with intestinal obstruction.
- Bowel adhesion, localisation, fibrosis, secondary infections are quiet common.
- Perforation (5%) though rare culminates in peritonitis.
- Plain X-ray shows multiple air fluid levels.
- Treatment *is resection and anastomosis and to continue anti-TB drugs*.
  Often *stricturoplasty* is beneficial.

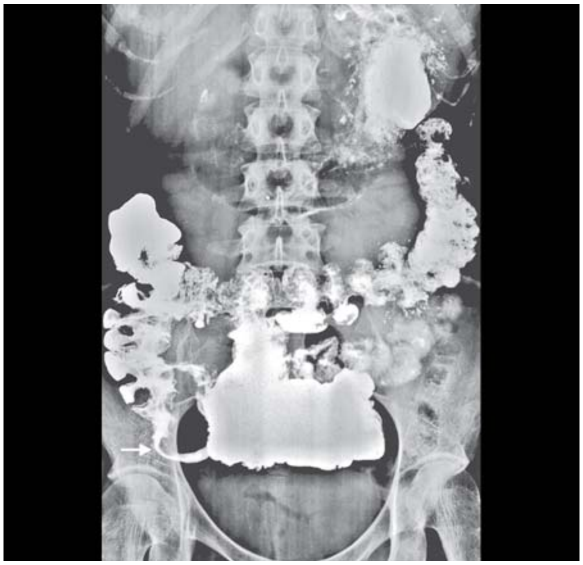

Fig. 10.10: Barium enema X-ray showing stricture in ileocaecal region due to ileocaecal tuberculosis.

PERITONEAL TUBERCULOSIS

Peritoneal tuberculosis

- Post-primary
- Becoming more common
- Activation of long-standing latent focus
- Blood spread
- Can develop from diseased mesenteric lymph node, intestine or fallopian tubes

Pathology in Peritoneal Tuberculosis

- Enormous thickening of the parietal peritoneum with multiple tiny yellowish tubercles.
- Dense adhesions in peritoneum and omentum with content inside as small bowel looking like ***abdominal cocoon***. It may precipitate intestinal obstruction.
- Multiple dense adhesions between bowel loops and between bowel and peritoneum and omentum.
- Thickening of bowel wall with adhesions.

Great things are done by a series of small things brought together.

Types

1. Acute Type—Mimics Acute Abdomen

Exploratory laparotomy reveals straw-coloured fluid with tubercles in the peritoneum, greater omentum and bowel wall. It is an on-table diagnosis. Fluid is evacuated and collected for AFB study and culture. Omental biopsy is taken. Abdomen is closed (without a drain) with tension sutures to prevent burst abdomen. ATD is started.

Acute peritoneal tuberculosis

- Rare
- On-table diagnosis
- Features of peritonitis
- Due to perforation or rupture of mesenteric tuberculous lymph nodes

2. Chronic Tuberculous Peritonitis

- Present as abdominal pain, fever, ascites, loss of weight and appetite, abdominal mass, ***doughy abdomen (10%).***
- Peritoneum is thickened with multiple tubercles. Omentum is thick, ***fibrosed, rolled up***.
- Infection is usually from mesenteric lymph nodes, ileocaecal tuberculosis, from fallopian tubes rarely blood born (from lungs).
- *Laparoscopy is very useful in this type to diagnose.*

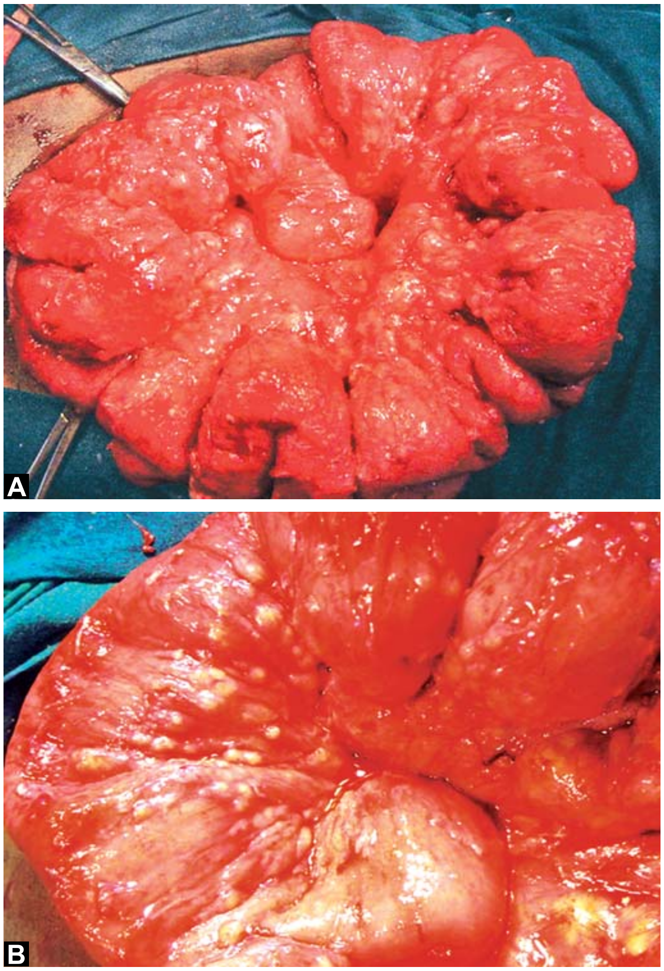

Figs 10.11A and B

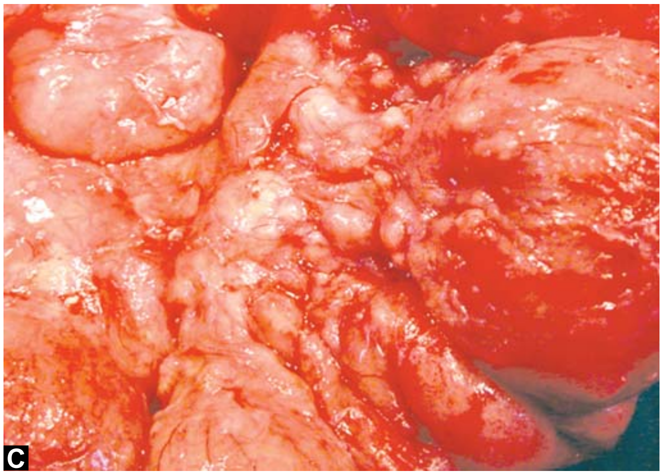

Fig. 10.11C

Figs 10.11A to C: On-table findings in intestinal tuberculosis, of extensive involvement with multiple tiny tubercles, thickening, adhesions and involvement of mesenteric lymph nodes.

Ascitic form

- **Ascitic form** shows enormous distension of abdomen with dilated veins.
- It presents with congenital hydrocele in male with patent processus vaginalis, umbilical hernia, rolled up omentum, shifting dullness, fluid thrill, and mass abdomen.
- ***Ascitic tap*** reveals straw coloured fluid from which AFB can be isolated. Fluid is pale yellow, clear, rich in lymphocytes, with high specific gravity.
- Chest X-ray, Mantoux test are other required investigations.
- ATDs for one year is required. Repeat tapping may be required initially as part of the treatment.

Note:

Ascites may be *free or loculated; clear or complex; debris or septae; 'sliced bread' appearance*.

Encysted (Loculated) ascites

- Ascites gets loculated because of the fibrinous deposition.
- Dullness, which is not shifting, is the typical feature.
- They may present as intra-abdominal mass, which may mimic ovarian cyst, retroperitoneal cyst or mesenteric cyst.
- Treatment is U/S guided aspiration along with ATD's.

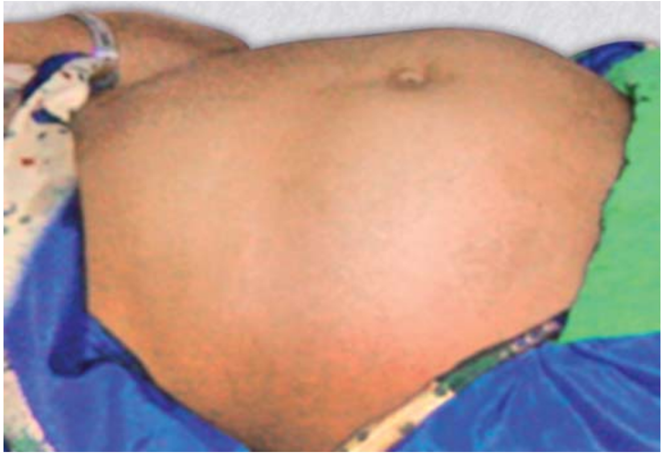

Fig. 10.12: Abdominal distension with ascites due to abdominal tuberculosis.

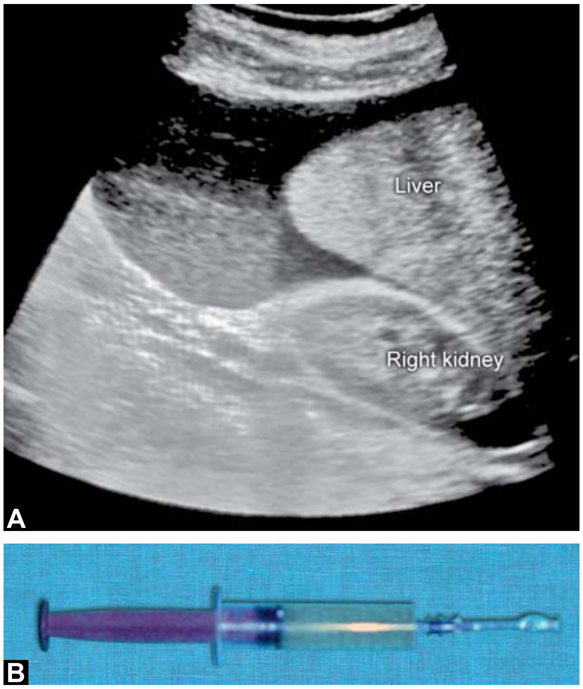

Figs 10.13A and B: US picture showing ascites. Straw coloured ascitic fluid tapping done under US guidance.

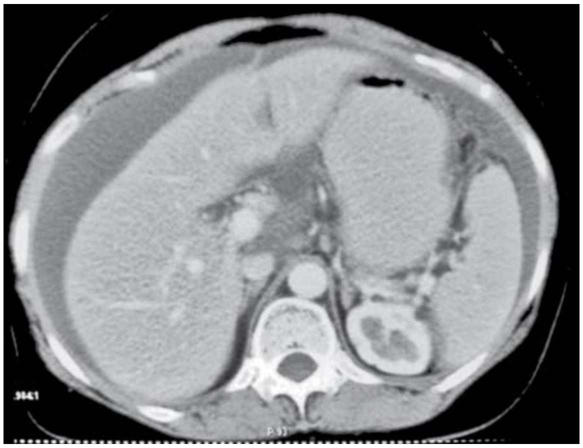

Fig. 10.14: CT scan abdomen showing ascites due to tuberculosis.

Plastic type

- Here there are wide spread adhesions between the coils of the intestine (ileum commonly), abdominal wall, omentum, with distension of the small bowel, leading to blind loop, ileus, intestinal obstruction (subacute, acute), thickened parietal peritoneum.
- They get recurrent colicky abdominal pain, diarrhoea, wasting, and loss of weight, mass abdomen, and *doughy abdomen*.
- *Differential diagnosis*: Peritoneal carcinomatosis. Open/laparoscopic peritoneal biopsy is very useful tool to diagnose.
- They respond well for drug treatment. Surgery is indicated if obstruction occurs.

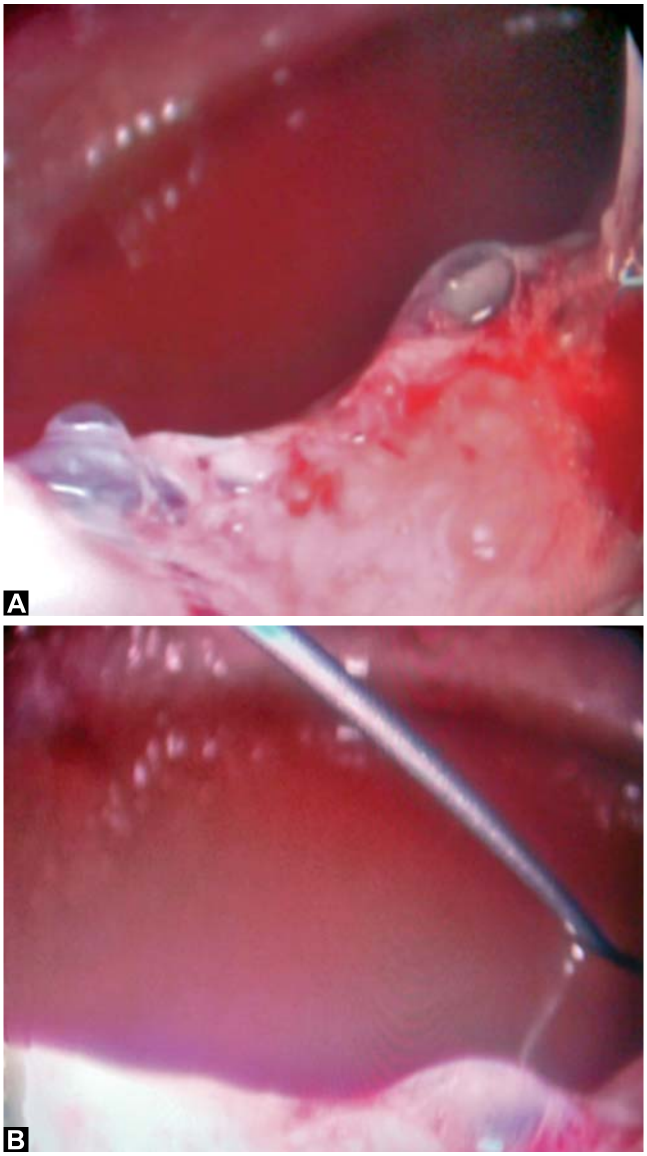

Figs 10.15A and B: Laparoscopic picture of loculated ascites due to abdominal tuberculosis. Fluid is getting aspirated using a needle under laparoscopic vision.

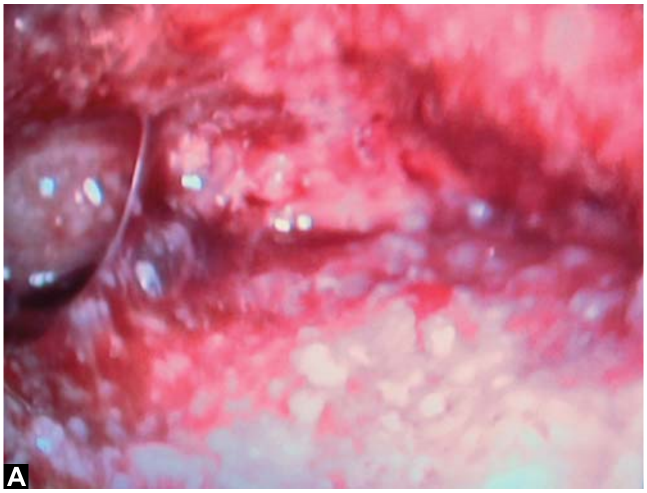

Fig. 10.16A

Example isn't the best way to teach, it's the only way.

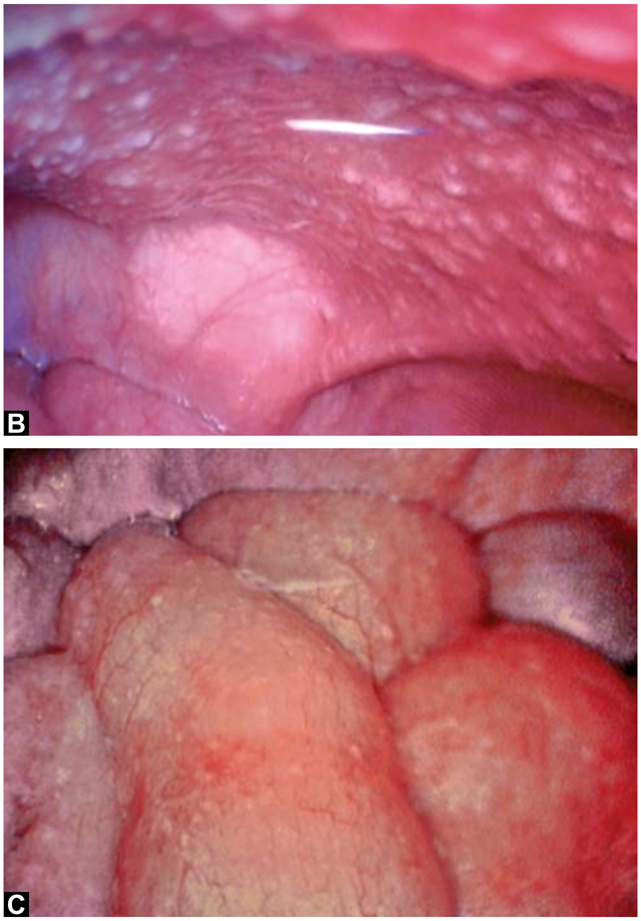

Figs 10.16B and C

Figs 10.16A to C: Laparoscopy showing plastic type of abdominal tuberculosis and multiple tubercles over parietal and visceral surface of the peritoneum.

Purulent form

- It is invariably due to tuberculous salpingitis, presenting as a mass in the lower abdomen containing pus, omentum, fallopian tubes, small and large bowel.
- Cold abscess gets adherent to the abdominal wall, umbilicus and may form an umbilical fistula.
- Patient commonly has got genitourinary tuberculosis.

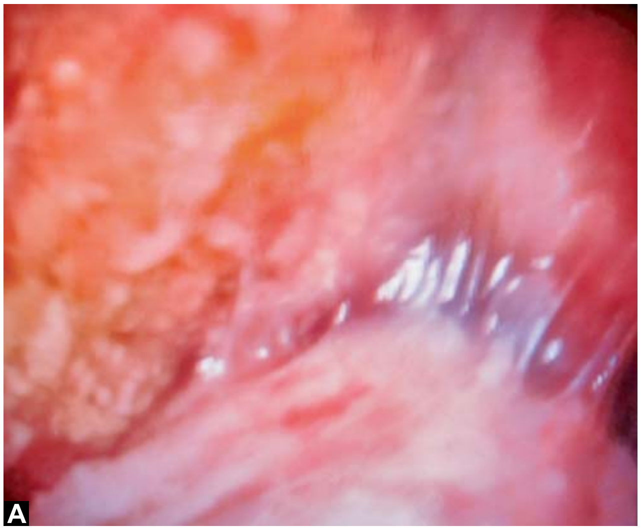

Fig. 10.17A

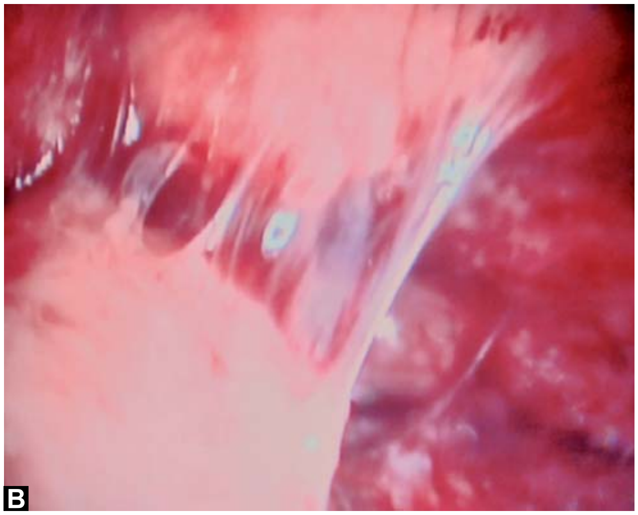

Fig. 10.17B

Figs 10.17A and B: Laparoscopy showing adhesions between bowel and abdominal wall due to tuberculosis.

- U/S, discharge study, X-ray abdomen and other investigations are useful.
- *Treatment:* ATD's are started exploration of umbilicus, exploration of fistula and bowel by pass is done.
- Prognosis is poor in this type.

Ascitic form	***Loculated form***
Insidious onset Fatigue, weight loss, fever, anorexia	Loculated abdominal swelling Ascites which is not shifting
Abdominal pain usually absent	Diagnosis—US guided aspiration
Abdominal distension usually without pedal oedema	Antituberculous drugs
Rolled up omentum is common association as mass	
Fibrous form	***Purulent form***
Widespread adhesion of bowel loops	Due to tuberculous salpingitis
Adhesions, matting and dilatation	Presenting as lower abdominal mass
Act as blind loop—malabsorption	Umbilical fistula, abdominal wall abscess can occur
Abdominal pain, loss of weight, *doughy abdomen* due to thickened parietal peritoneum, diarrhoea, often mass abdomen—are usual features	Cold abscess adherent to abdominal wall

TUBERCULOUS MESENTERIC LYMPHADENITIS

- Infection is usually through the Peyer's patches of the intestine (i.e. through oral cavity). Usually several lymph nodes are involved often causing massive lymph node enlargement. Commonly right-sided lymph nodes are involved, but left sided nodes can also get involved.
- It presents with general symptoms (fever, malaise, weight loss).

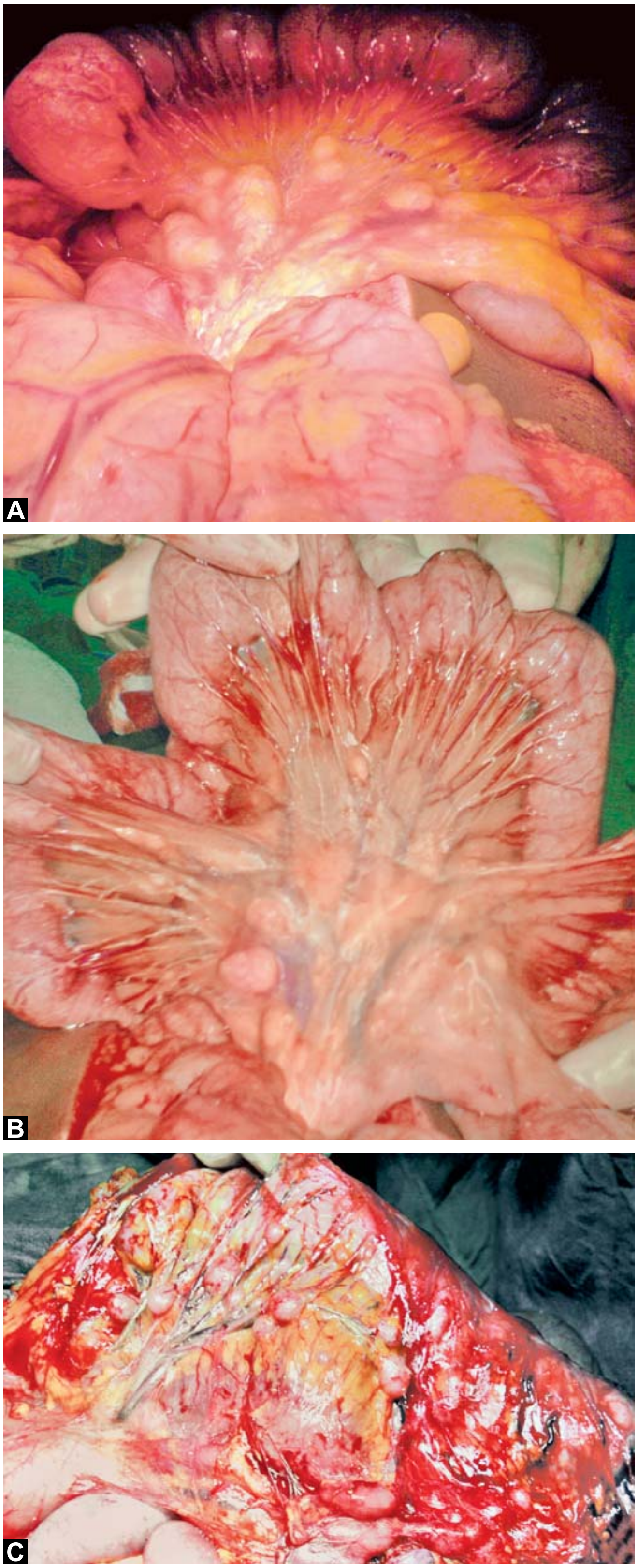

Figs 10.18A to C: Mesenteric tuberculous lymphadenitis on table findings.

- Pain in umbilical region and right iliac fossa, mass in right iliac fossa, which is matted, nonmobile.
- It may present with features of *acute appendicitis*.
- Often coils of intestine get adherent to the caseated mesenteric lymph nodes leading to intestinal obstruction.
- Most often caseating material may collect between the layers of the mesentery, forming a cold abscess, mimicking a mesenteric cyst (***Pseudomesenteric cyst).***
- *Massive enlargement of mesenteric lymph nodes due to tuberculosis is called as **tabes mesenterica**.*
- Mesenteric tuberculous adenitis is more common in children. Present with anaemia, fever, loss of appetite and reduced weight, failure to thrive, palpable mass in right iliac fossa which is firm and nodular.

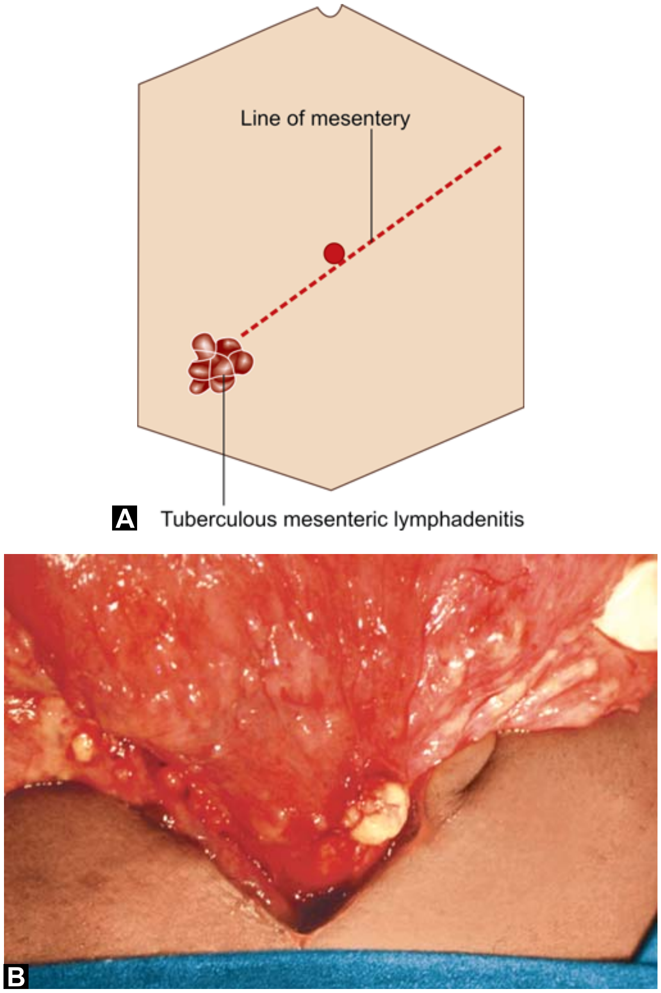

Figs 10.19A and B: Note the location of mesenteric tuberculous lymphadenitis in the rigth iliac fossa. But it can occur anywhere in the line of mesentery. On-table caseating tuberculosis lymphadenitis is also shown.

Presentations of mesenteric tuberculous lymphadenitis

- Mesenteric tuberculous adenitis—acute/chronic in right iliac fossa—mass
- Pseudomesenteric cyst
- Cold abscess within the mesentery
- Tabes mesenterica

The same beam of light can illuminate two objects and produce two different effects.

Differential Diagnosis

- Carcinoma caecum.
- Lymphoma.
- Retroperitoneal tumour.
- Nonspecific lymphadenitis.

(Acute nonspecific mesenteric lymphadenitis is called as **nurses' syndrome**).

Investigations

- X-ray abdomen shows calcification.
- U/S may confirm the diagnosis.
- Mantoux test may be positive.
- *Diagnostic laparoscopy*—is very useful in TB lymphadenitis. Mesenteric cold abscess can be drained safely through laparoscopy.

Treatment: ATD's; Laparoscopy and proceed. Prognosis is good.

ANO-RECTO-SIGMOIDAL TUBERCULOSIS

- It mimics carcinoma rectum.
- It presents as tenesmus, diarrhoea, and discharge from the fistula and occasionally as mass per abdomen. Rectal tuberculosis occurs usually within 10 cm of anal verge.
- Fistulas are *painful and characteristically not indurated.* Tuberculous fistulas are commonly multiple.
- Tuberculous anal ulcers when occur are shallow, bluish, with undermined edges.
- Sigmoidoscopy, U/S, discharge study, fistulectomy and biopsy confirms the diagnosis.
- Treatment is ATD's, fistulectomy, often *sigmoid resection.*

TUBERCULOSIS OF THE OMENTUM

- It usually occurs as a part of the other types of abdominal tuberculosis.
- *Rolled up omentum* with thickening is characteristic.
- Often cold abscess can develop *per se* in the omentum.
- If it is so it can be dealt with laparoscopy safely under the cover of ATDs.

In all abdominal tuberculosis, drug treatment is for one year In adrenal tuberculosis and severe adhesions steroids may be beneficial

Indications for steroids in tuberculosis

- Miliary tuberculosis
- Tuberculous meningitis
- Adrenal tuberculosis
- Tuberculous pericarditis
- Pleural and endobronchial tuberculosis
- In abdominal tuberculosis to prevent adhesions
- Tuberculosis of the mediastinal lymph nodes (Rapidly progressive)
- Genitourinary tuberculosis
- Ocular tuberculosis

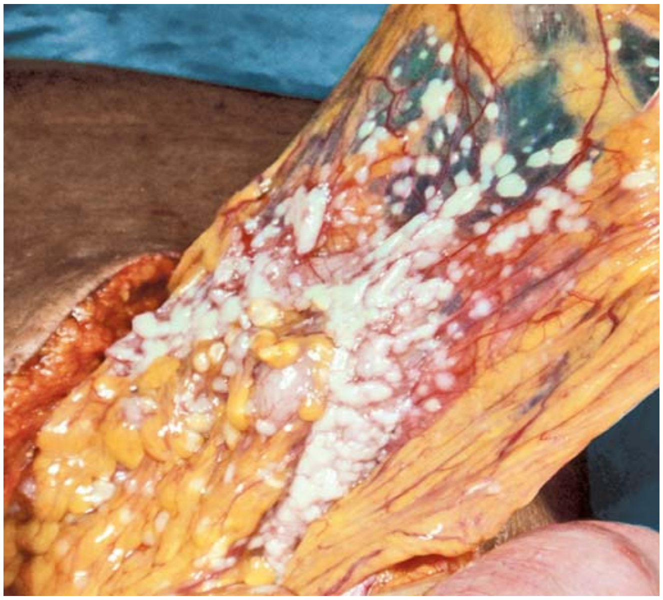

Fig. 10.20: Omental tuberculosis showing tubercles.

Remember

- AFB staining of fixed tissue in abdominal tuberculosis is possible only in 5% cases. *Ziehl-Neelsen* staining is done. (Carbol fuchsin—steam heat for 7 minutes—water wash-decoloraisation with 20% H_2SO_4—95% ethanol wash—counterstain with Loeffer's methylene blue/1% picric acid/0.2% malachite green. Tuberculous bacilli are bright red rods seen under oil immersion lens).
- Culture takes 8 weeks to get the result. 35% of abdominal tuberculosis shows positive in culture. *Lowenstein Jenson media* is used. Guinea pig inoculation is also useful method.
- *WHO now recommends* antituberculous drugs for 6 months. Uncomplicated cases—4 drugs for 2 months and 2 drugs for 4 months. Complicated cases—4 drugs for 2 months and 2 drugs for 7 months. But many patients need one year treatment. Because of complications and difficulty in managing recurrent cases of abdominal tuberculosis one year therapy is commonly used in developing countries.
- *First line drugs* are—isoniazid—5 mg/kg; rifampicin—10 mg/kg; ethambutol—15 mg/kg; pyrazinamide—25 mg/kg.
- *Second line drugs* are—amikacin; kanamycin; PAS (Paraamino sulphuric acid); ciprofloxacin; ofloxacin; clarithromycin; azithromycin; rifabutin.
- Follow-up and prognosis is monitored by—regular weight check to see for gain; improvement in appetite; reduction of abdominal pain and distension; absence of fever; normal bowel habits; normal haemoglobin; ESR becoming normal; and US abdomen shows improvement in sonological features.
- Patients who are not responding in 6 weeks should be reassessed again for—drug resistance; associated other diseases like malignancy (carcinomas or lymphoma), Crohn's disease, eosinophilic enteritis.
- During therapy patient who is responding for drug therapy can also go for intestinal obstruction due to fibrosis during healing stage. It needs surgical intervention.
- Repeated surgery in abdominal tuberculosis is difficult and dangerous as chances of developing faecal fistula, further adhesions are more likely.

Chapter 11 Liver

The liver…that great maroon snail…. No wave of emotion sweeps it. Neither music nor mathematics gives it pause in its appointed tasks.

—Richard Selzer, 1976

CHAPTER OUTLINE

- Surgical Anatomy of Liver
- Liver Function Tests
- Alpha-Fetoprotein
- Liver Biopsy
- Liver Injury
- Infections of Liver
  - Amoebic Liver Abscess
  - Pyogenic Liver Abscess
  - Portal Pyaemia
  - Hydatid Cyst of Liver
  - Actinomycosis of Liver
- Liver Tumours
  - Benign Tumours of the Liver
  - Primary Malignant Tumours of the Liver
  - Hepatocellular Carcinoma
  - Secondaries in the Liver
- Liver Cysts
  - Congenital Liver Cysts
  - Neoplastic Liver Cysts
  - Traumatic Liver Cyst
- Portal Hypertension
- Oesophageal Varices
- Portal Hypertensive Gastropathy
- Ascites
- Ascites in Portal Hypertension
- Budd-Chiari's Syndrome
- Hepatic Failure
- Hepatic Encephalopathy
- Hepatorenal Syndrome
- Hepatic Resection
- Portal Biliopathy

SURGICAL ANATOMY OF LIVER

Liver has four lobes—*right lobe, left lobe, quadrate lobe and caudate lobe.*

Supporting structures are *right triangular ligament, left triangular ligament and falciform ligament*.

Blood Supply

It is a unique organ with dual blood supply. Hepatic blood flow is around 1500 ml/minute of which *portal vein* contributes 80% and *hepatic artery* 20%.

Hepatic artery commonly originates from coeliac axis. There are many variations of hepatic artery, both in its origin and its pathway and is considered surgically important. Hepatic artery ligation is one of the surgical palliation for advanced hepatocellular carcinoma and secondaries in liver as tumour tissue is supplied exclusively by hepatic artery but normal tissue is also supplied by portal vein.

Superior mesenteric vein and splenic vein join dorsal to the neck of pancreas to form portal vein.

Hepatic artery, portal vein and bile duct are located in the free edge of the lesser omentum until it enters the liver.

Venous drainage is usually through the three major hepatic veins, right, left, and middle which drain into the IVC. Often inferior hepatic vein may be present.

The liver parenchyma is entirely covered by a thin capsule called '*Glissons capsule*' and by visceral peritoneum in all but the posterior surface of liver termed 'bare area'.

I prefer to learn and to teach anatomy, not from books but from dissections; not from the positions of philosophers but from the fabric of nature.

—William Harvey

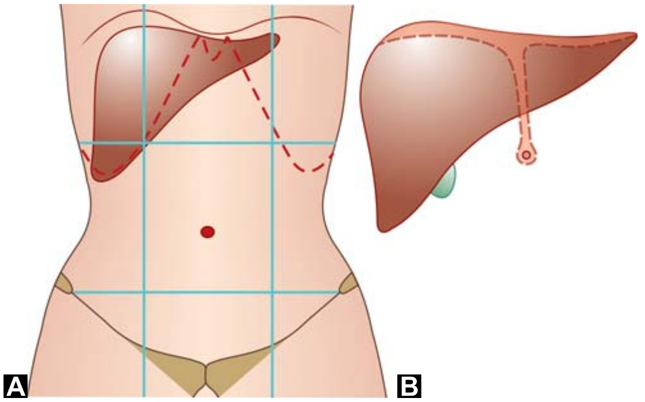

Figs 11.1A and B: Anatomical location of liver and its lobes.

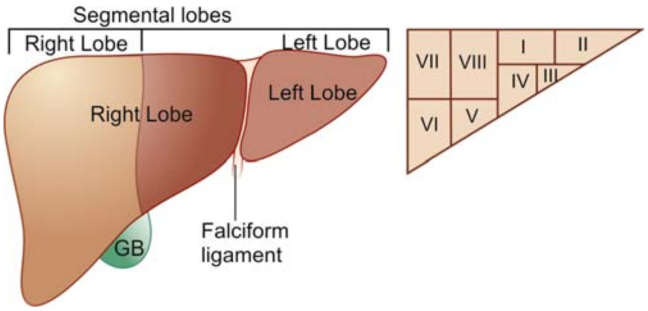

Fig. 11.2: Segmental anatomy of the liver.

Segmental Anatomy of Liver

Liver is divided into functional right and left lobes by a line passing from the left of the gallbladder fossa to the left of IVC—*Cantlie's line* creating *Couniaud's* segments. There are *eight segments*:

Segments I, II, III, and IV are of left lobe.

Segments V, VI, VII, VIII are of right lobe.

Segment I is the caudate lobe of the liver and has independent supply of portal and hepatic veins. This hepatic vein directly joins IVC.

Right lobe is having right hepatic artery, right branch of portal vein, and right hepatic duct.

Left lobe is having left hepatic artery, left portal vein, left branch of bile duct.

Functional unit is called as hepatic lobule and it contains central hepatic vein and portal triad (hepatic arteriole, portal venule, bile ductule).

Functions of the liver

- Removal of gut endotoxins and foreign antigens—liver acts as the first filter
- Drug and hormone metabolism
- Formation of bilirubin and its metabolism
- Formation of urea from protein catabolism
- Glucose metabolism, glycolysis, and gluconeogenesis
- Clotting factors synthesis
- pH balance and correction of lactic acidosis
- Maintaining body temperature
- Storage of vit. B12, vit. A, Cu, Fe

LIVER FUNCTION TESTS (LFT)

1. Serum bilirubin which includes both direct and indirect. Test is known as *van den Bergh's test*.
2. Serum albumin, globulin and A : G ratio; serum albumin is the indicator of chronic liver disease.
3. Prothrombin time: Normal value is *12-16 seconds*. Difference between control and test more than 4 seconds or test being more than 1½ times the control is significant. When it is altered it is corrected by injecting vitamin K, 10 mg IM for 5 days or by fresh frozen plasma (FFP)—*For cell synthesis.*
4. Alkaline phosphatase—*Secretary function.*
5. Aspartate amino transaminase *5-40 IU/litre* (AST, SGOT)—*signifies inflammation.*
6. Alanine transaminase *5-40 IU/litre* (ALT, SGPT), liver specific.
7. 5 nucleotidase.
8. Gamma glutamyl transpeptidase (GGT) 10-48 IU/L.
9. Immunological tests: Antimitochondrial or antinuclear antibodies.
10. AFP.
11. Specific tests: (a) For haemochromatosis: Serum iron, total iron binding capacity, serum ferritin. (b) Wilson's disease: Serum copper, urinary copper, serum ceruloplasmin.
12. Fluorodeoxyglucose-positron emission tomography (FDG-PET): It is to find out the uptake of labelled glucose which varies in different diseases of liver, i.e. benign, malignant and inflammatory.
13. Technetium-99m labelled radioisotope scan shows the uptake and excretion of bile.
14. A sulphur colloid liver scan shows specifically Kupffer cell activity. Sulphur colloid will not show uptake in adenoma and haemangioma as Kupffer cells are absent in these lesions.
15. Urine for bile salts (*Hay's test*), for bile pigments (*Fouchet's test*) and for urobilinogen (*Ehrlich's aldehyde test*).

Hepatocyte function	***AST, ALT***
Synthetic function and metabolism	PT-INR, factor V. VII, albumin, bilirubin
Biliary canalicular function	ALP, 5' nucleotide, gamma glutamyl transferase, bilirubin

Other investigations for liver diseases

- U/S abdomen
- Angiography
- CT Scan
- PTC
- ERCP
- MRI
- Laparoscopy and laparoscopic U/S
- Liver biopsy

ALPHA-FETOPROTEIN (AFP)

- It is a normal component of plasma protein of the human foetus secreted by embryonal hepatocytes which disappears 2-3 weeks after birth.
- Normal value is up to 10 ng/ml of plasma.
- It is an important tumour *marker for HCC and more often HCC with cirrhosis* (between 100-1000 ng/ml).
- It also increases in *hepatoblastoma, testicular tumours and ovarian tumours.*
- It is a prognostic indicator.

- It also increases slightly in benign conditions of the liver (only upto 100 ng/ml if at all).
- It increases significantly in non-seminomatous testicular tumours and ovarian tumours (yolk sac tumour).
- AFP is a useful marker to suspect, relapse or residual tumour.

LIVER BIOPSY

Prerequisites:

- Prothrombin time should be normal before doing liver biopsy, otherwise severe bleeding can occur.
- Normal value of prothrombin time is 12-16 seconds (It is compared to control and difference should be less than 4 sec between test and control). It is corrected by giving inj. vit K 10 mg IM, for 5 days. After that if repeat PT is still high then fresh frozen plasma (FFP) is required to correct it.

Indications

1. Cirrhosis, chronic hepatitis, haemochromatosis, Wilson's disease.
2. Hepatoma, secondaries in liver.

Complications

- Haemorrhage.
- Bile leak and biliary peritonitis.
- Infection.

Contraindications

- Hydatid disease, where it will precipitate anaphylaxis.
- Haemangioma, bleeding disorders.
- Ascites.

Procedure

Needle used here is either *Vim-Silvermann needle* or *Menghini needle* (better option).

Needle is usually punctured through right midaxillary line in 9th intercostal space or under U/S guidance into any place as required.

Presently, guided liver biopsy is commonly recommended and popular, i.e. U/S guided, CT scan guided, laparoscopy guided.

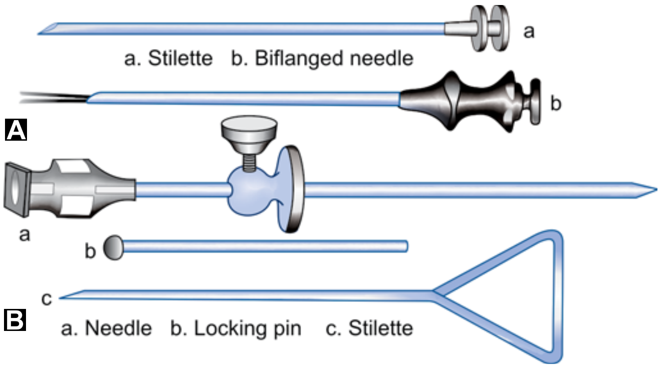
a
a. Stilette b. Biflanged needle
b
A
a
b
c
B
a. Needle b. Locking pin c. Stilette

Figs 11.3A and B: Liver biopsy needles. A. Vim-Silvermann, B. Menghini.

LIVER INJURY

Causes

It can be due to blunt injury, stab, gun shot injury.

Types

It can be contusion, laceration, avulsion, extension into thorax and biliary tree may be associated with other organ injuries (spleen, kidney, duodenum, bowel, IVC) and with fracture ribs.

Liver injury can be:

- *Penetrating*: It often requires surgical intervention. After laparotomy with rooftop incision, laceration is assessed, clots and blood in the peritoneal cavity is removed. Using *Pringle manoeuvre* bleeding is temporarily controlled. Often suprahepatic IVC and infrahepatic IVC control may be required. Liver wound is sutured using specialized needle with vicryl using *co-opting sutures*. Gelfoam or Surgicel is placed over it. Other injuries like of diaphragm, biliary system, bowel should be looked for. Adequate amount of blood, FFP should be kept for transfusions. Postoperative assessment of BT, CT, PT, platelet count, observation for sepsis is needed. Ideal antibiotic coverage using 3rd generation cephalosporins or higher are needed.
- *Blunt trauma*: It is assessed by CT scan. It is usually treated conservatively. Indications for surgical intervention are—progressive deterioration and bleeding, grade 5 liver injury on CT scan, associated bowel injury. Commonly used procedure after laparotomy is keeping packs between liver and diaphragm which is removed in 48-72 hours. Occasionally venovenous bypass, hepatic resection may be required.

Clinical Features

- Features of shock due to severe torrential bleeding (pallor, hypotension, tachycardia, sweating).
- Distension of abdomen with dull flank, guarding, tenderness and rigidity.
- Oliguria.
- Tachypnoea, respiratory distress and often cyanosis.
- Rupture of right lobe is more common than left lobe leading to *haemoperitoneum.*
- Occasionally can cause localized haematoma which may form an abscess.
- Bile leak from the injured site can lead to *biliary peritonitis*.

Investigations

- Chest X-ray to look for rib fractures.
- U/S abdomen, CT scan of chest and abdomen (CT is ideal).
- Diagnostic peritoneal lavage.
- Hb%, PCV, blood grouping and crossmatching. Adequate amount of blood (5-10-15 bottles of blood) must be kept ready for transfusion, i.e. requires massive transfusion.
- Arterial blood gas analysis.

Regeneration of liver after resection is complete in 4-6 months. 80% of liver can be resected without compromising its function.

CT grading of liver injury

Grade and type	*Injury*
I – Haematoma – Laceration	Subcapsular haematoma less than 10% surface area. Capsular laceration less than 1 cm in depth.
II – Haematoma – Laceration	Subcapsular haematoma 10-50%; intraparenchymal less than 10 cm diameter. Capsular laceration 1-3 cm in depth; less than 10 cm length.
III – Haematoma – Laceration	Subcapsular haematoma more than 50%; intraparenchymal more than 10 cm diameter—expanding. Capsular laceration 3 cm in depth.
IV – Laceration	Parenchymal disruption 25-75% hepatic lobe; 1-3 Couinaud segments.
V – Laceration – Vascular	Parenchymal disruption more than 75% or more than 3 Couinaud segments. Major hepatic veins or retrohepatic venacaval injuries.
VI – Vascular	Hepatic avulsion

- *Coagulation profile.*
- *Thromboelastography* is dynamic form of assessing the coagulation status on table.
- Arteriography to visualise the bleeding branch/vessel in the liver and embolisation can be tried.

Treatment

General measures

1. IV fluids, blood transfusion (massive), FFP.
2. Have both central venous access, and peripheral venous access.
3. Bladder catheterisation has to be done to measure the urine output.

Initial conservative nonoperative management

- It is done in—nonprogressive liver injuries in patients who are haemodynamically stable, low grade (I-III) liver injury, need of less than 2 units of transfusion, without peritoneal signs, normal mental status.
- However CT abdomen (absence of extravasation of contrast during arterial phase can be treated nonoperatively) and repeat CT or regular ultrasound follow up is a must.
- Replacement of lost blood; prevention of sepsis; regular monitoring by haematocrit, liver function tests, prothrombin time are needed.
- *Angiographic embolisation* increases the success rate of nonoperative therapy.
- Intensive care unit (ICU) management for 2-5 days; repeat CT scan after 5 days; bed rest to be continued; patient can have normal activity only after 3 months.
- When at any time patient's condition worsens, he should be taken up for laparotomy.

Specific treatment

- *Push (direct compression); plug* (plugging the deep track injuries using silicone tube or SB tube); *Pringle's* manoeuvre; *pack* (liver wound is directly packed with mop).
- Laparotomy is done through a large *bucket handle* abdominal incision or thoracoabdominal incision, and extent of liver injury and also other associated injuries are looked for.
- Small liver tear is sutured with *vicryl* or *PDS* mattress sutures with placing of gel foam to control bleeding.
- To control bleeding on table, from hepatic artery and portal vein, both are temporarily occluded using fingers, compressing at foramen of Winslow-*Pringle manoeuvre*. Often bull-dog clamp or vascular clamps can be used.

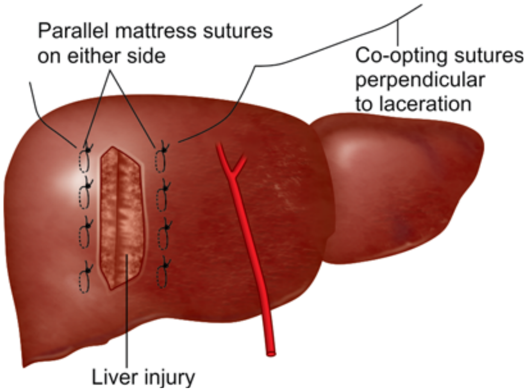
Parallel mattress sutures on either side
Co-opting sutures perpendicular to laceration
Liver injury

Fig. 11.4: Co-opting sutures are placed perpendicular to already placed parallel mattress sutures.

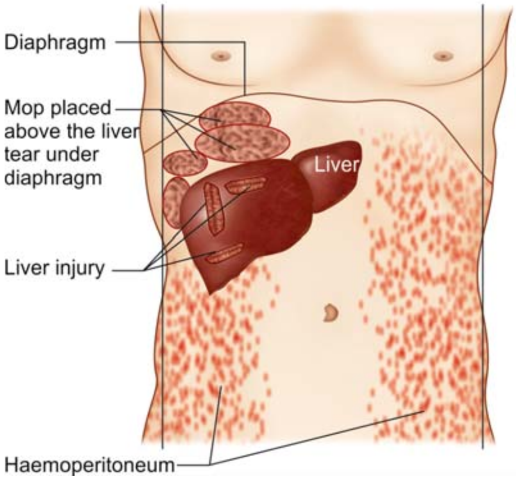

Fig. 11.5: By placing 3-5 mops between injured liver and diaphragm tamponade can be achieved to control bleeding from the liver. After 48-72 hours electively abdomen is re-explored for further intervention. By then commonly bleeding will be controlled and further management becomes easier. Sepsis should be taken care of by higher generation antibiotics.

- In deep severe injuries, following methods are used:
  - Hepatic artery ligation, not usual.
  - Segmental resection.
  - Hemihepatectomy, not commonly used.
  - Packing the liver temporarily with mops and closing the abdomen.
- Ligation of bleeding vessel individually on the cut surface is needed and is called as *tractotomy*. Biliary ductules are also ligated. As vessels are supported by *fibrous adventitia* ligation suture holds well on these vessels even though liver parenchyma is soft and friable.
- Liver can be wrapped with absorbable mesh *(mesh hepatorrhaphy)* to achieve haemostasis.

- Cholecystectomy and placement of ‘T’ tube in CBD.
- In associated IVC injuries it is very difficult to manage (has high mortality). A *veno-venous bypass* between femoral vein and SVC is done and then repair of IVC is carried out.
- ICT placement to thorax and repair of diaphragmatic injury.
- Management of other organ injuries accordingly.
- *Kous Netzoff aluminium needle* is used to suture liver tear.
- Postoperatively patient requires ventilator support, blood transfusion, electrolyte management, antibiotics like third generation cephalosporins, ceftazidime, cefoperazone.
- FFP, cryoprecipitate are also required.

Management of liver trauma

- Control of haemorrhage—packing; tractotomy; mesh hepatorrhaphy; debridement of devitalized tissue; segmentectomy, vein repair with veno—venous bypass; transfusions.
- Management of metabolic disturbances—prevention of acidosis, DIC, FFP; platelets.
- Monitoring—repeat US; repeat CT; urine output; LFT; prothrombin time; platelet count; haematocrit.
- Prevention of sepsis—proper debridement; irrigation of liver wound with warm normal saline; antibiotics.
- Identification and management of complications of liver injury

Porta hepatic injury

- It is life threatening even though rare. Penetrating trauma commonly cause this injury.
- It is commonly associated with organ injuries. Isolated porta injury is rare. It is difficult to manage with mortality 50-80%.
- Portal vein is sutured after veno venous bypass.
- Common bile duct injury involving less than half the circumference can be sutured over a T tube. Injury more than 50% or transection needs choledochojejunostomy.

Complications and sequelae of liver injury

- Shock and haemorrhage
- Liver abscess or septicaemia—7%
- Bile leak, biliary peritonitis, biliary fistulas—10%
- Disseminated intravascular coagulation
- Hepatic artery aneurysm, arteriovenous and arteriobiliary fistulas—haemobilia
- Intrahepatic haematoma
- Intraabdominal abscess—subphrenic, pelvic
- Complications of massive blood transfusion
- Electrolyte imbalance
- Respiratory complications
- Liver failure
- **Late sequelae** of liver trauma is *CBD stricture* causing obstructive jaundice. It can be managed by endoscopic stenting or by open Roux-en-Y hepaticojejunostomy
- Mortality is 10%; morbidity is 30%. Associated injuries increase the mortality from isolated liver injury of 3% to 20%.

INFECTIONS OF LIVER

AMOEBIC LIVER ABSCESS

- It is common in India and other tropical countries and it is caused by a parasite entamoeba histolytica.
- It is more common in *alcoholics and cirrhotic* patients**.**
- It is the commonest extraintestinal presentation of amoebiasis. It is often called *as tropical abscess*.
- Infection commonly occurs from the caecum after an attack of amoebic typhlitis (inflammation of the caecum) through the superior mesenteric vein and portal vein. Infection from sigmoid (rectosigmoid) colon spreads through the inferior mesenteric vein and portal vein to liver. Right lobe is commonly involved over posterosuperior surface (because of streamline effect and larger size of the right lobe). Trophozoites destroy the hepatocytes by releasing *histiolysin*, a cytolytic agent. It causes amoebic hepatitis with multiple microabscesses formation. It leads into liquefaction necrosis, thrombosis of blood vessels, release and, breaking of red cells. It causes formation of “*Anchovy sauce’* pus which is chocolate brown coloured and odourless (Anchovy sauce is sauce prepared from a type of fish). Pus may be green coloured if mixed with bile. Secondary infection is common (30%).
- In western countries pyogenic abscess is much more common. Amoebic abscess is much more common than pyogenic abscess in endemic areas like Indian sub-continent, and Africa.
- Amoebic abscess is usually sterile unless infected. Trophozoites are found in the wall of the abscess not in the content.

Life Cycle of Entamoeba Histolytica

Mature cyst in faeces → contaminate food/water → ingestion of cyst → pass through stomach undamaged → in alkaline medium, cyst wall lysis occurs by trypsin → excystation → release of quadrinucleate amoebae—metacyst trophozoites formed → Habitat in crypts of caecum commonly, often in sigmoid colon *as to form trophozoites.*

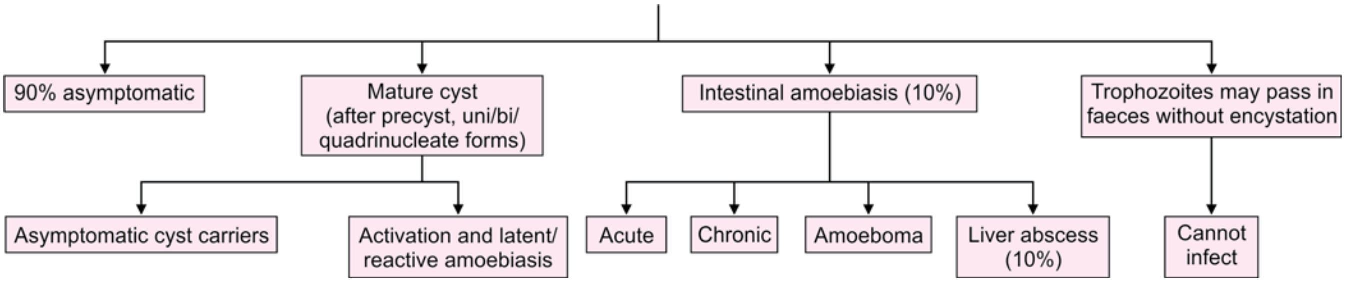
90% asymptomatic
Mature cyst (after precyst, uni/bi/ quadrinucleate forms)
Intestinal amoebiasis (10%)
Trophozoites may pass in faeces without encystation
Asymptomatic cyst carriers
Activation and latent/ reactive amoebiasis
Acute
Chronic
Amoeboma
Liver abscess (10%)
Cannot infect

> *Amoebic liver abscess may be present as bulge in the right lower intercostal spaces and may exhibit pitting edema.*
> **—Ajit K Basu**

Note:
- 10% of amoebic infection causes intestinal amoebiasis commonly in caecum; 10% intestinal amoebiasis leads into amoebic liver abscess.
- Malnutrition, patients on steroid therapy, immunosuppression, cirrhosis and alcoholism increases the infection rate and also rate of amoebic liver abscess.
- Amoebic liver abscess is not common in females especially in menstruating age group.
- Anchovy sauce pus is odourless unless infected secondarily.
- Size is the independent important *risk factor* for rupture.

Pathology

- Initially from infected rectosigmoid or ileocaecal region, amoebic trophozoites reach the liver through portal veins causing amoebic hepatitis, may be in the form of micro-abscesses all over the liver. This might resolve on its own or with antiamoebic drugs, but often leads to a localized *amoebic liver abscess*.
- In 70% of cases it is *single large abscess,* in 30% it is multiple, may involve both lobes. Problems and difficulties in treating, in addition to poor prognosis are more common in *multiple abscesses*.
- Amoebic liver abscess is more common in *right posterior-superior* region (80%) because of streamline effect, i.e. the portal vein is in direct continuation with the right branch. It can be *multiloculated* also.
- Pus is *chocolate coloured,* classically called as **anchovy sauce,** contains dead liver cells, RBCs, necrotic material. Pus may be green due to bile admixture.
- Often secondary infection by *E. coli*, staph, strepto may occur (30%) and so may present with features of pyogenic liver abscess. Because of perihepatitis, liver is fixed to diaphragm or abdominal wall, hyperaemia in the diaphragm causes sympathetic *pleural effusion on right side*.
- Commonly amoebic abscess presents as an *acute entity*, but it can also be present as *chronic type* where it is covered by a capsule, that remains dormant for a long period.
- Sometimes it can get calcified also**.**

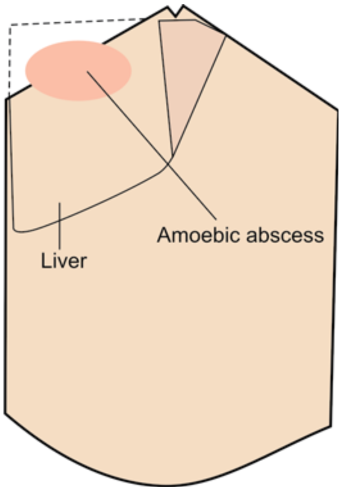

Fig. 11.6: Commonest location of amoebic liver abscess—right posterosuperior part of the liver.

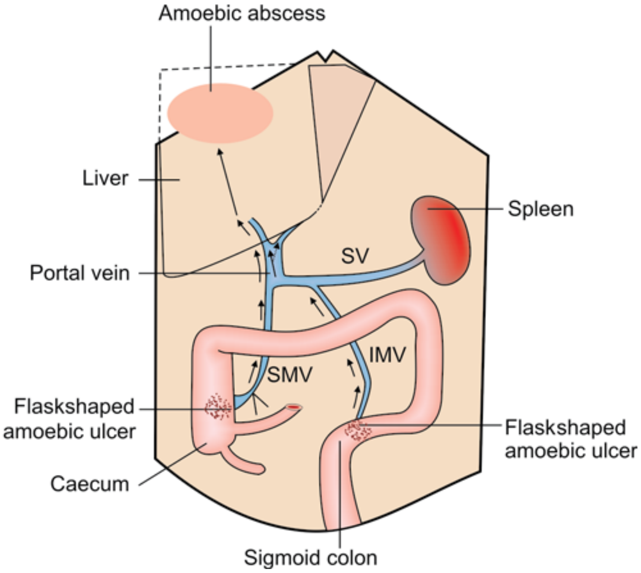

Fig. 11.7: Spread of trophozoites occurs from ileocaecal region and sigmoid through mesenteric vessels.

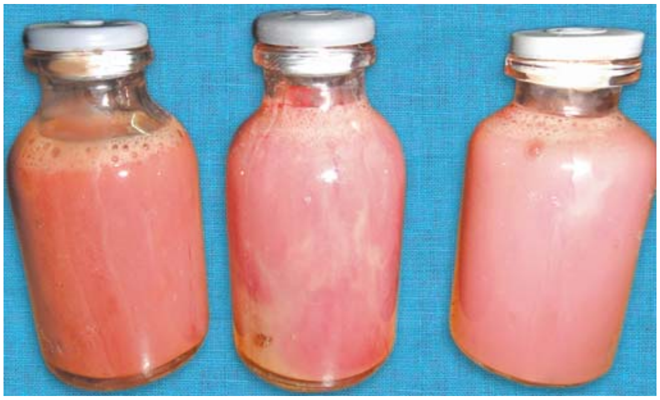

Fig. 11.8: Anchovy sauce pus seen in amoebic liver abscess.

Course and Sequelae of Amoebic Liver Abscess

- It can *rupture into lungs* leading to expectoration of chocolate-coloured sputum resulting in natural regression of abscess—commonest site of rupture.
- It can rupture into the peritoneum causing *peritonitis* which requires emergency laparotomy.
- It can rupture into pleural cavity leading to *empyema.*
- Rupture into bronchus can cause bronchopleural fistula leading into coughing out of Anchovy sauce pus.
- Rupture into bare-area of liver causing *retroperitoneal abscess*.
- Rupture into the intestines, or to the skin (*Amoebiasis cutis).*
- Most dangerous complication is rupture into pericardial cavity (*cardiac tamponade*) which has very *high mortality* (30%) requiring emergency thoracotomy and pericardial decompression.
- *Septicaemia and liver failure* can occur in a patient with amoebic liver abscess with *cirrhosis.*

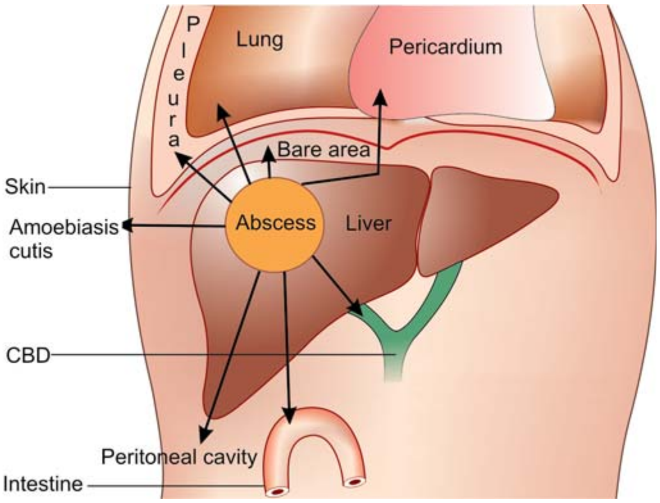

Fig. 11.9: Complications of amoebic liver abscess.

Clinical Features

- It is common in males (20:1), may be after an attack of amoebic dysentery or many months after the attack or history of dysentery may not be there at all.
- They present with fever, loss of weight, chills and rigors, non-productive cough, shoulder pain.
- Pain in the right hypochondrium—90%.
- Soft, tender, smooth, liver with increased liver span—70%.
- **Intercostal tenderness** is elicited which is a useful clinical sign.
- Right sided pleural effusion may be evident.
- Mild jaundice is not uncommon especially in cirrhotics and multiple abscesses which may signify poor prognosis—20%.
- Tenderness, rigidity and skin oedema in right hypochondrium may be present in acute cases.
- In chronic amoebic liver abscess, smooth, firm/hard, nontender liver may be palpable.

Amoebic liver abscess may be:

- *Acute*—present with high fever, chills, rigors, tender, soft palpable liver, with intercostal tenderness.
- *Chronic*—present with firm/hard, smooth, nontender palpable liver without acute features.

Features may be of:

- *Systemic*—present with fever, chills and rigors, loss of appetite, reduced weight, and jaundice.
- *Abdominal*—present with pain and tenderness, localised guarding and rigidity, mass in right upper abdomen (tender, soft liver), ascites, splenomegaly, abdominal wall oedema.
- *Thoracic*—present with dry cough, chest pain in right lower part, right shoulder pain, pleural effusion, and intercostal tenderness.
- *Features of complications*—rupture/infection/septicaemia/liver failure.

Differential diagnosis

For acute type

- Acute cholecystitis
- Acute presentation of hepatocellular carcinoma (HCC) due to haemorrhage or necrosis
- Subphrenic abscess
- Pyogenic liver abscess

Chronic amoebic liver abscess mimics hepatoma in every respect

Investigations

- Total count may be increased.
- Liver function tests may show altered bilirubin and albumin level.
- Prothrombin time may be widened and if it is so Inj. vit K 10 mg IM for 5 days should be given. Even with this if P.T. remains widened then fresh frozen plasma (FFP) is needed to rectify the P.T.
- Serum alkaline phosphatase, SGPT, SGOT levels are altered.
- U/S abdomen shows altered echogenicity (anechogenic, hypoechogenic), size, location, number of abscess, nature of the liver—90% sensitivity.
- Chest X-ray findings:

- Raised fixed diaphragm (tenting)
- Pleural effusion
- Soft tissue shadow

- Indirect haemagglutination test (95% positive rate), ELISA and gel diffusion precipitative test are reliable serological tests. Serological tests are reliable in non-endemic areas than endemic areas. Counter immunoelectrophoresis is more useful in active disease.
- CT scan-contrast study. CT scan shows raised diaphragm; abscess cavity (low density area)—its size, location, number; presence of effusion; changes in the lung—95% sensitivity.
- Sigmoidoscopy/colonoscopy are used to identify the active ulcers. Scrapings of the ulcer show trophozoites.

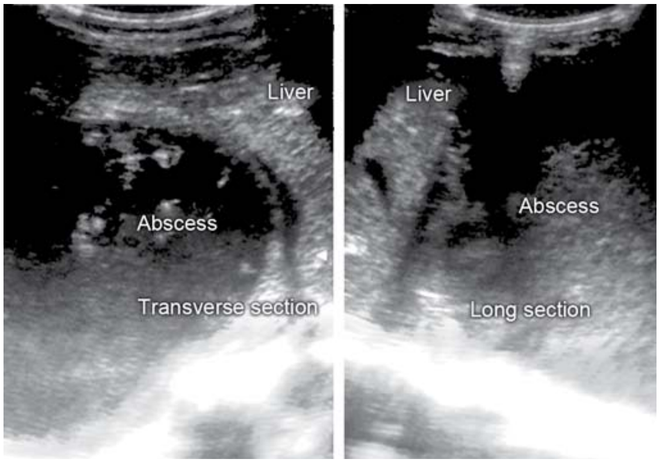

Fig. 11.10: US picture of amoebic liver abscess. US guided aspiration is very useful.

PIVKA II: Protein induced in the absence of vitamin K or antagonists-II.

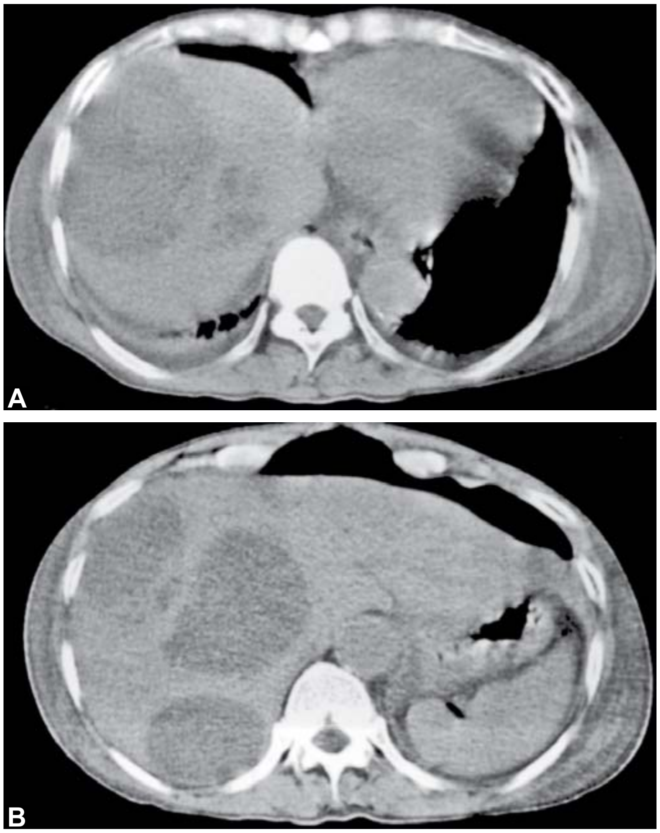

Figs 11.11A and B: CT scan pictures showing single, large and multiple liver abscesses.

Treatment

Drugs

- Tab. metronidazole 800 mg tid or Inj. metronidazole 500 mg IV tid for 10 days.
- Tinidazole 600 mg BD dose for 5 days.
- IV or oral antibiotics are essential to control secondary infection (cefotaxime, ciprofloxacin, amoxycillin) (Small abscesses < 3 cm respond to drugs).
- *Other drugs:*
  1. Injection dihydroemetine 1.5 mg/kg/day IM for 5 days should be given under cardiac monitoring.
  2. Chloroquine 250 mg BD given for 10-14 days.

Drugs used for amoebic infection

- Metronidazole, Tinidazole, Secnidazole, Ornidazole
- Dihydroemetine injection
- Chloroquine, tetracycline
- Diloxanate furoate, iodoquinol, paromonycin—as *cyst eradicators*

Note:
Metronidazole is not given in 1st trimester of pregnancy.

Aspiration

- In case of large abscess and infected abscess aspiration with a wide bore needle is done under U/S guidance after correcting the P.T.
- Previously without U/S, aspiration used to be done by passing needle in right 6th intercostal space in midaxillary line.

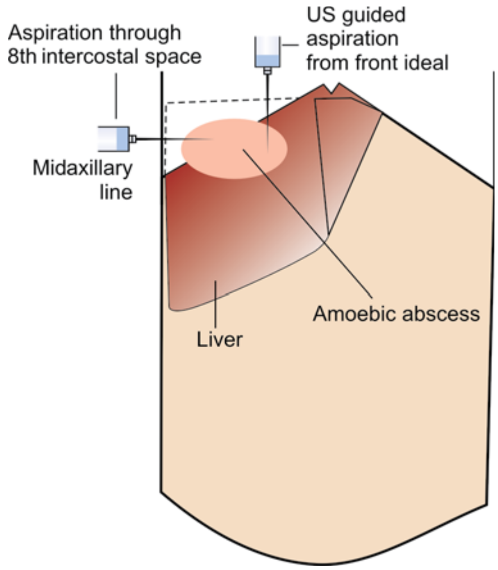
Aspiration through 8th intercostal space
US guided aspiration from front ideal
Midaxillary line
Amoebic abscess
Liver

Fig. 11.12: Aspiration of the amoebic liver abscess under US guidance or through right 8th intercostal space in midclavicular line.

Percutaneous Drainage

- Under U/S guidance pigtail catheter is placed into the abscess cavity percutaneously to drain the pus. Catheter tube

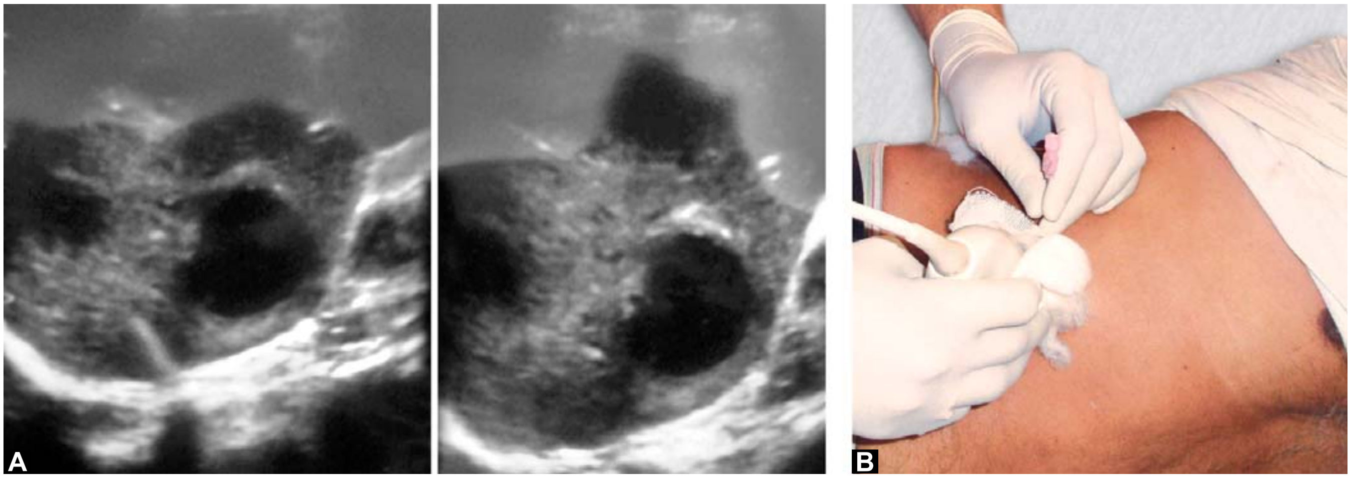

Figs 11.13A and B: Amoebic liver abscess US picture. US guided aspiration is being done.

and abscess cavity has to be washed and irrigated at regular intervals with normal saline. It may fail if there is thick pus, multiloculated abscess, and multiple abscesses. Procedure may cause bleeding and infection.

Note:

Aspirated fluid is sent for C/S, cytology and for study of trophozoites (last part of the aspirated fluid should be sent for this).

Surgery

Indications for surgery

- Even after repeated aspirations if abscess cavity fills again
- Thick pus
- Multiloculated abscess
- Left lobe abscess, because of danger of rupture into pericardial cavity
- Ruptured abscess
- Caudate lobe abscess
- Multiple abscesses

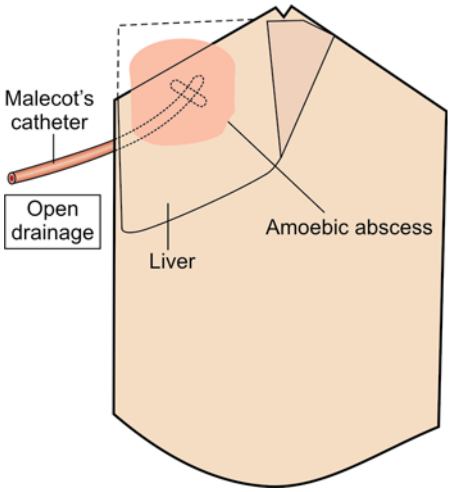

Fig. 11.14: Drainage of amoebic liver abscess transperitoneally and placing Malecot's catheter into the abscess cavity.

Procedure

- Through transperitoneal approach, abscess area is opened, pus is evacuated. Malecot's catheter is placed and brought out through a separate stab incision. The catheter is kept *in situ* until drainage stops completely. Complete drainage of pus will be confirmed by repeat U/S. During discharge, advice is given to avoid alcohol, Chloroquine 250 mg BD for 10 days and Diloxanide furoate 500 mg tid is given for 10-14 days.
- *Complications of surgery*—anaesthetic problems; bleeding; liver failure (in cirrhotic patients); intraperitoneal abscess formation; bile leak—bile peritonitis and fistula.

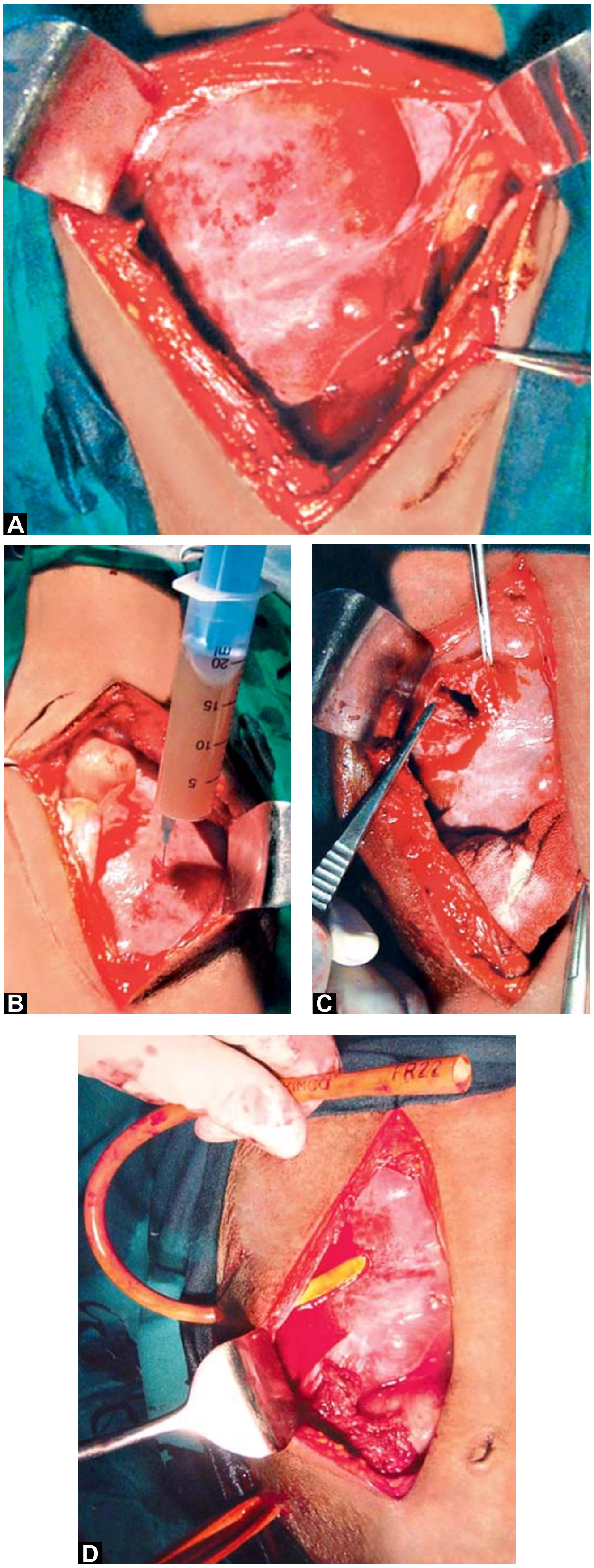

Figs 11.15A to D

Anything the mind of man can conceive and believe, it can achieve.

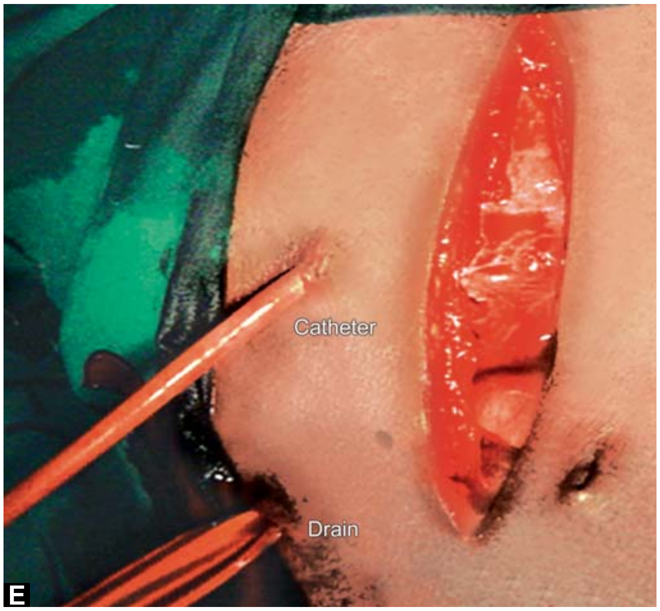

Fig. 11.15E

Figs 11.15A to E: Operative findings of amoebic liver abscess. Note the exposed abscess in the liver, its aspiration and drainage. Placement of Malecot's catheter in the abscess cavity and drain placed into the peritoneal cavity is also seen.

Follow-up

At regular intervals is important. Measures taken are proper counseling to avoid alcohol; repeat LFT; repeat ultrasound to confirm the complete resolution of the abscess cavity.

Prognosis

- *Mortality* in amoebic liver abscess is 4%. It rises with rupture especially pericardial (30%).
- *Poor prognostic factors* are—Rupture, serum bilirubin > 3.5 mg%, serum albumin < 2.0 g/dl, liver failure, cirrhosis, multiple abscesses, volume of abscess > 500 ml, anaemia, diabetes.

PYOGENIC LIVER ABSCESS

Aetiology

1. *Biliary sepsis* 35%; commonest route.
   a. Empyema gallbladder.
   b. Cholangitis.
   c. After biliary tract surgery.
   d. Instrumentation.
   e. Stone disease, Caroli's disease, biliary ascariasis, biliary enteric anastomosis.
2. *Portal vein sepsis:*
   a. Appendicitis.
   b. Diverticulitis.
   c. Inflammatory bowel disease, pancreatitis, perforation, PID, colorectal carcinoma.
   d. Omphalitis in newborn.
3. *Distant infections* (*through hepatic artery*):
   a. Pneumonia.
   b. Upper UTI.
   c. Endocarditis, osteomyelitis, bacteraemia.
4. *Super added infections:*
   a. Amoebic liver abscess.
   b. Hydatid cyst.
5. *Cryptogenic liver abscess*—No identified primary infection.
6. *Trauma* becoming common cause.
7. *Direct extension*:
   From suppurative cholecystitis, subphrenic abscess, perforation, perinephric abscess.

Causative bacteria

- *E. coli*—commonest
- *Klebsiella*
- *Proteus*
- *Pseudomonas*
- Clostridia
- Enterococci, streptococci viridians in polymicrobial infection

Pathology

- Due to laminar blood flow right lobe (75%) is commonly involved; left lobe (20%), caudate lobe (5%) are also often involved.
- Usually solitary—60%; occasionally it can be bilobar and multiple.
- Cavity contains pus with virulent organisms. Usually abscess is acute. In cryptogenic type chronic presentation is known to occur.
- Ascites and splenomegaly is not common.
- It is more common in diabetics.
- Male to female ratio is *2:1*. It is more common in old people after 55 years of age.
- *Blood culture* commonly shows positive for bacteria.

Clinical Features

- Pain in the right hypochondrium—60%.
- High fever, with rigors—90%.
- Weight loss.
- Jaundice—occasionally—20%.
- Intercostal tenderness.
- Tender, soft liver-60%.
- Features of toxicity.
- Constitutional symptoms like malaise, lethargy, vomiting.

Diagnosis

- Ultrasound abdomen, CT scan. Sensitivity is 90% for USG; 97% for contrast CT scan.
- LFT, total count.
- Ultrasound guided aspiration of pus after controlling PT.
- Chest X-ray shows elevated diaphragm often with right sided pleural effusion.
- Blood culture is very relevant.

Differential Diagnosis

- Amoebic liver abscess, hydatid cyst, subphrenic abscess.

Treatment

- Systemic antibiotics—combination of third generation cephalosporins and metronidazole.
- *Ultrasound guided aspiration/pigtail catheter—Percutaneous drainage* is the treatment of *choice* at present. Drainage tube/catheter are placed under US / CT guidance into the liver abscess. Pus should be sent for culture and sensitivity. Follow up USG, LFT, assessment of quantity of daily drainage—should be done to assess the response. 75% of pyogenic abscess is drained percutaneously. *Percutaneous aspiration* without drainage tube placement is also used; but repeated guided aspirations are required; otherwise success rate is similar to drainage.
- Occasionally open drainage is required. *Open drainage* is indicated in recurrent abscess, failure of percutaneous drainage, large abscess of size more than 5 cm. Open drainage is becoming *less common* at present; but in selected patients it may be a life saving essential therapeutic modality.
- Treating the primary causes is very essential.

Complications

- Septicaemia, liver failure.
- Rarely rupture and peritonitis can occur.
- Klebsiella hepatic abscess can cause dangerous endogenous endophthalmitis commonly in diabetic patients impairing vision.

PORTAL PYAEMIA (PYLEPHLEBITIS)

It usually follows after severe infection of areas drained by portal vein.

Causes

- Appendicitis.
- Diverticulitis.
- Any severe abdominal sepsis.

Presently it is becoming *rare* because of availability of good effective antibiotics.

Infective thrombus in the vein (draining the infective area)
↓
It dislodges as infective emboli
↓
Reaches the liver
↓
Causes multiple abscesses in liver parenchyma
↓
Pylephlebitis.

E. coli is the most common organism. Often staphylococci, anaerobes, or combined infections may be involved.

Note:

Gas and thrombus in portal vein is diagnostic of portal pyaemia.

Clinical features

- High fever with rigors
- Toxicity
- Drowsiness, jaundice
- Tender soft liver
- Ascites
- Blood culture is a useful investigation

Treatment

- *Antibiotics*: Combination of third generation cephalosporin + aminoglycoside, ceftriaxone sodium, cefoperazone, ceftazidime, amikacin, tobramycin, metronidazole.
- IV fluids, blood transfusion, FFP.
- Ventilator support.
- Mortality is very high and usually patients die of hepatic failure, septicaemia, MODS.
- It is better to prevent portal pyaemia by prior good antibiotics in suspected cases.
- Treatment of primary cause is very important.

HYDATID CYST OF LIVER

Word meaning is '*dew drop*' (Latin). In Greek it means 'watery vesicle'.

Caused by **Echinococcus granulosus (EG),** *dog tape worm*, a parasite.

Life-cycle: Infected offal of sheep
↓
Eaten by the dog (definitive host)
↓
EG released, develops in the dog's intestine into a parasite of 1 cm long with a head and three segments, last of which contains about 500 ova.
↓
Ova expelled from the dog's intestine to grass and vegetables
↓
Eggs are ingested by sheep, cattle or human beings *(intermediate host)*
↓
Through portal vein—liver—larva form
↓
Hydatid cysts (70% in liver).

It takes few years to evolve into a complete hydatid cyst.

Most commonly involved segment is segment VII—27%. Commonly right lobe—66%; both lobes in 16% and only left lobe is involved in 17%.

Pathology

It has got 3 layers

1. *Adventitia (pseudocyst)* is an inseparable fibrous tissue due to reaction of the liver to the parasite.
2. *Laminated membrane (ectocyst),* formed of the parasite itself is whitish, elastic, containing hydatid fluid, which can be peeled off readily from the adventitia.

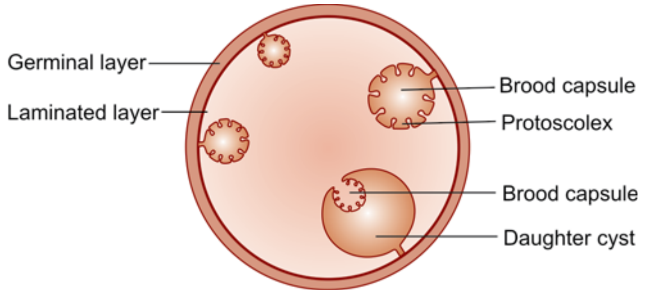

Fig. 11.16: Anatomy of hydatid cyst.

3. *Germinal epithelium* is the only living part, lining the cyst (*endocyst).*

This layer secretes *hydatid fluid*, *brood capsules with scolices* (heads of future worms).

Features of hydatid fluid

- Clear
- High specific gravity (1.005-1.009)
- Shows hooklets and scolices

Once brood capsules disintegrate, it grows into daughter cysts.

Course of the Disease

- The parasite may die and cyst eventually may get *calcified.*
- Commonly cyst *enlarges* and is palpable per abdomen.
- It may cause complications like *jaundice* due to pressure over biliary tree.
- *Rupture into the peritoneal cavity* causes anaphylactic reaction which may cause life-threatening shock, requiring proper management with steroids.
- *Rupture into biliary channels is commonest (60%)*. Rupture into bowel, pleural cavity can occur.
- *Secondary infection* causing suppuration and septicaemia.
- *Secondary cysts* in the lung, spleen, mesentery, retroperitoneum and other organs can occur**.**
- *Hepatic dysfunction*.
- *Disseminated abdominal hydatidosis* can occur after silent rupture.

Clinical Features

- Asymptomatic palpable liver with *classical thrill (hydatid thrill)* elicited by *three-finger test.*
- Jaundice and pain.
- Features of anaphylaxis.
- Discomfort in right upper quadrant area; dyspepsia; hydatid cachexia in children; weight loss; fatigue; vomiting.
- Occasionally splenomegaly, pleural effusion, cholangitis, allergic asthma, fever.
- *Camellotte sign*: Following intrabiliary rupture, gas enters into cyst causing partial collapse of the cyst wall.

Differential Diagnosis

- Hepatoma.
- Amoebic liver abscess.
- Cystic disease of the liver.

Complications of hydatid

- Anaphylaxis
- Rupture
- Obstructive jaundice— ERCP sphincterotomy may be needed
- Infection
- Calcification
- Liver failure

Investigations

- U/S is diagnostic. It reveals rosettes of daughter cysts, double contoured membrane of the cyst due to detachment of the cyst membranes, and calcification of cyst wall. Intraoperative ultrasound (IOUS) is very useful tool.
- X-ray often shows calcification.
- CT scan abdomen is more accurate in identifying cyst characteristics—*cart wheel* like—multivesicular rosette like.
- *Primary serological tests*—ELISA; indirect haemagglutination test; latex agglutination test; immunofluorescence antibody test; immunoelectrophoresis. 80-95% sensitivity for liver hydatid.
- *Secondary laboratory tests*—Detection of precipitation line—arc 5; immunoblotting; polymerase chain reaction (PCR). More specific, very useful in extrahepatic hydatid disease and calcified non-fertile liver hydatid.
- Liver function tests.
- Casoni‘s test (intradermal test—75% sensitive); complement fixation test—historical interest.
- MRI when there is jaundice to visualise biliary tree and its relation to hydatid cyst; to find out cystobiliary communication; biliary hydatids in bile duct and hepatic ducts. ERCP can also be done to find out the communications. Other method to find out the cystobiliary communications is intraoperative cholangiogram.
- ERCP to find out biliary communications but not as useful as MRI.

Aspiration of the cyst should not be done due to risk of anaphylaxis but presently PAIR (Puncture-Aspiration-Injection-Reaspiration) is done effectively.

Hassen Gharbi's ultrasound based classification of liver hydatid cysts (1981)

Type 1 – Pure fluid collection
Type 2 – Fluid collection with split wall
Type 3 – Fluid collection with septa
Type 4 – Heterogeneous appearance
Type 5 – Reflecting thick walls

WHO classification of liver hydatid cyst

Type – CL – Active; unilocular; no cyst wall; early stage; not fertile
Type – CE 1 – Active; cyst wall present; hydatid sand present; fertile
Type – CE 2 - Active; multivesicular rosette like cyst wall; fertile
Type – CE 3 – Transitional; detaching laminated membrane; *water-lily* sign; beginning of degeneration
Type – CE 4 – Inactive; degenerative contents; no daughter cysts; not fertile
Type – CE 5 – Inactive; thick calcified wall; not fertile

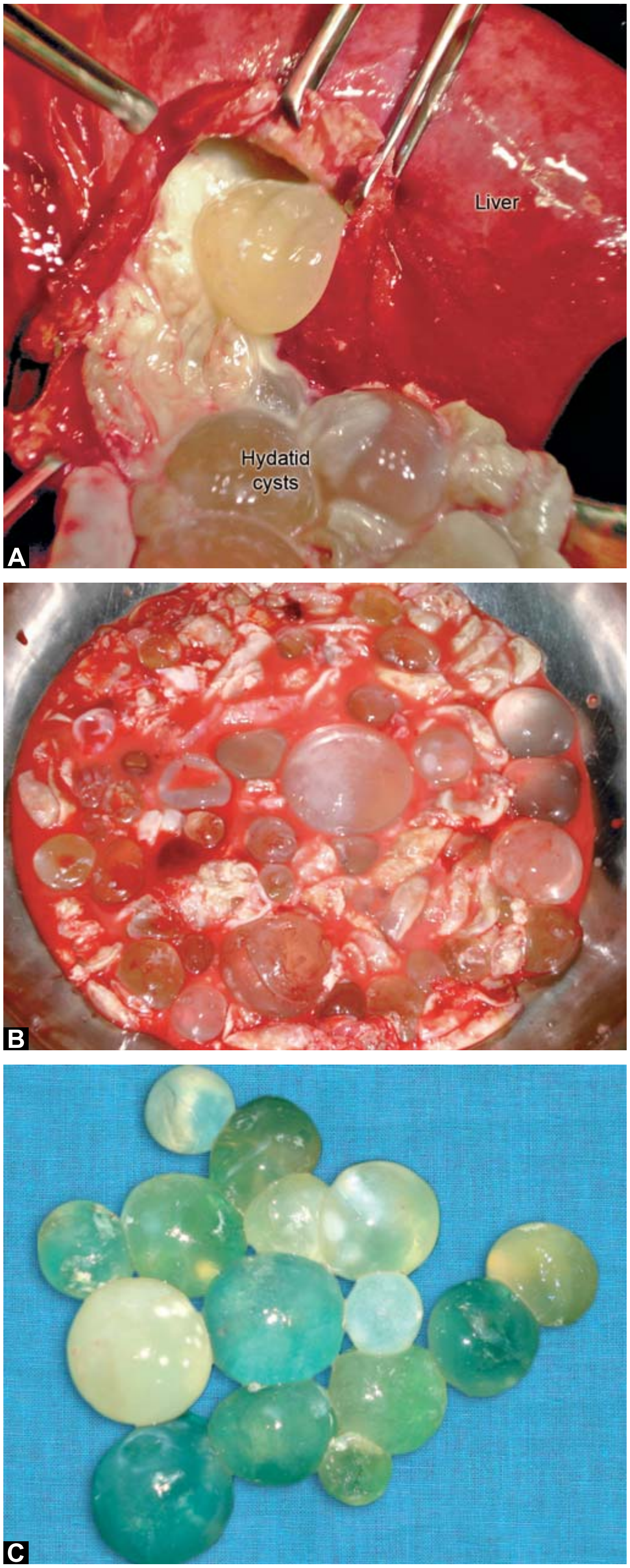

Figs 11.17A to C: Hydatid cyst of the liver—typical look.

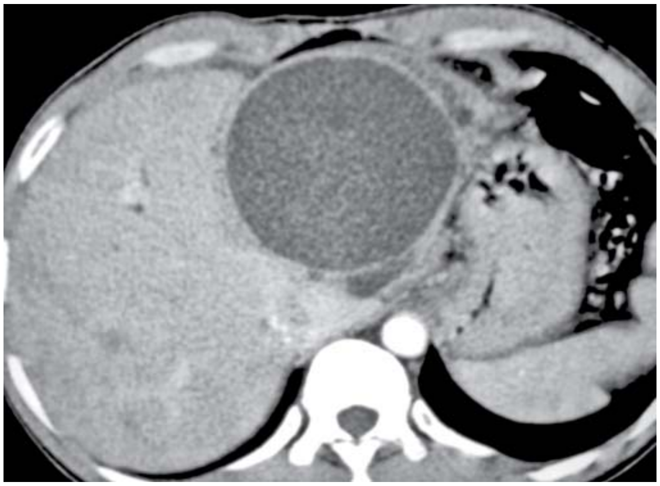

Fig. 11.18: CT scan of hydatid disease of the liver.

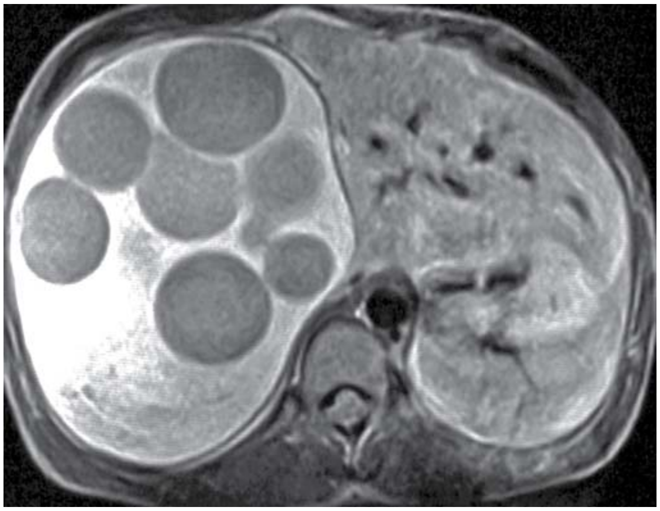

Fig. 11.19: MRI picture of multiple hydatid cyst of liver—MRI is very useful investigation in hydatid cyst of liver.

Treatment

Drugs

- *Indications for drug therapy*
  - 4 days prior to intervention and to continue it for 1 month (albendazole) to 3 months (mebendazole) after the intervention.
  - Inoperable cysts.
  - Multiple or multiorgan cysts.
  - Recurrent hydatids.
  - Surgically unfit patients.
  - Cysts in lungs, bone, brain, eyes.
- *Contraindications*
  - Large cysts.
  - Honeycomb cysts (with septae).
  - Infected cysts.

Venous hum, which is heard louder on inspiration in epigastrium, in portal hypertension is called as Kenawy's sign

  - Calcified cysts.
  - Pregnancy.
- *Drugs used are:*
  - Albendazole 4-week cycles with 2 weeks drug free interval. It is ovicidal/larvicidal/*vermicidal*. Commonly used.
  - Praziquantel—60 mg/kg along with albendazole for 2 weeks.
  - Mebendazole—600 mg daily for 4 weeks.

PAIR – (Puncture – Aspiration – Injection – Reaspiration) – Percutaneously Done [1980]

Indications for PAIR	*Results and problems of PAIR*
• Inoperable patients	• Complication rate—10-40%
• Patients who refuse surgery	• Mortality rate—0.9-2.5%
• CL, CE 1, CE 2 and CE 3 types/Gharbi type 1 and 2	• Fever—35% - disappears in 72 hours
• Relapse cysts	• Anaphylaxis—0.1-0.2%- Same as open surgery—but drugs should be kept ready for anaphylaxis
• Infected cysts	• Infection—10% well controlled by antibiotics
• In pregnant women; children less than 3 years	
• Cysts more than 5 cm in different liver segments	• Local recurrences—4% Repeat PAIR can be done
Contraindications for PAIR • Inaccessible cysts • Cysts with multiple septae—honeycomb cysts • Hyperechogenic cysts • Communicating cysts to bile duct • Calcified cysts • Cysts in the lung	

Technique of PAIR

- Done under US/CT guidance.
- Under local anaesthesia cyst is punctured using a cholangiography 22 gauge needle through thickest route/part of cyst wall.
- Cyst is entered through non-dependent wall and 50% of fluid is aspirated. All multiple/daughter cysts are aspirated. Radiopaque dye is injected to see if any communications are present. Scolicidal agents—15-20% hypertonic saline is injected into the cyst. After 20 minutes reaspiration is done. A sclerosant—alcohol is injected. If cyst is 6 cm or more, a drainage catheter is placed for 24 hours for complete drainage and later alcohol sclerosant is injected.
- Double—puncture aspiration and injection (D-PAI) or modified PAIR/PEVAC (Percutaneous EV Acuation of Cyst) are other procedures done.
- PAIR has gained wide acceptance as it is safe, less invasive, and easier to do, with low morbidity and mortality. Complications and results are same as open surgery.

Surgery

- Surgery is still the choice and *gold standard* therapy for hydatid disease. The abdomen is opened, and the peritoneal cavity is packed with mops [black or coloured mops are used to identify white scolices clearly so as to pick up all and prevent any spillage]. Fluid from the cyst is aspirated and scolicidal agents [cetrimide, chlorohexidine, alcohol, hypertonic saline (15%-20%), 10% povidone iodine or H_2O_2] are injected into the cyst cavity (formalin should not be used)**.** Hypertonic saline should be left within the cavity for 15-20 minutes to have effective scolicidal effect.

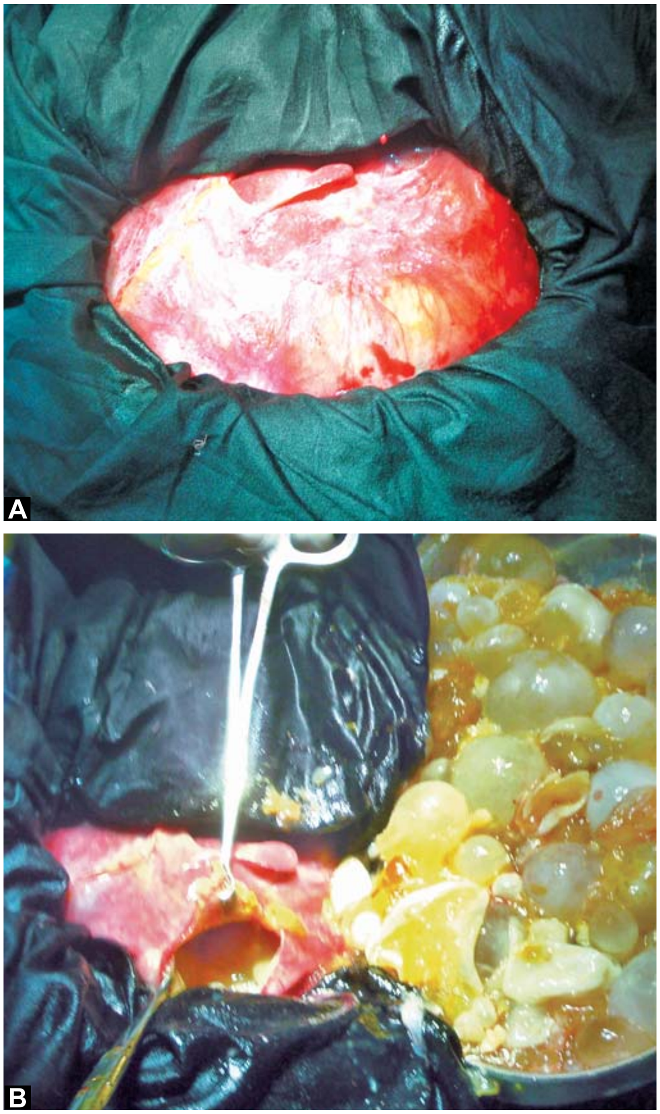

Figs 11.20A and B

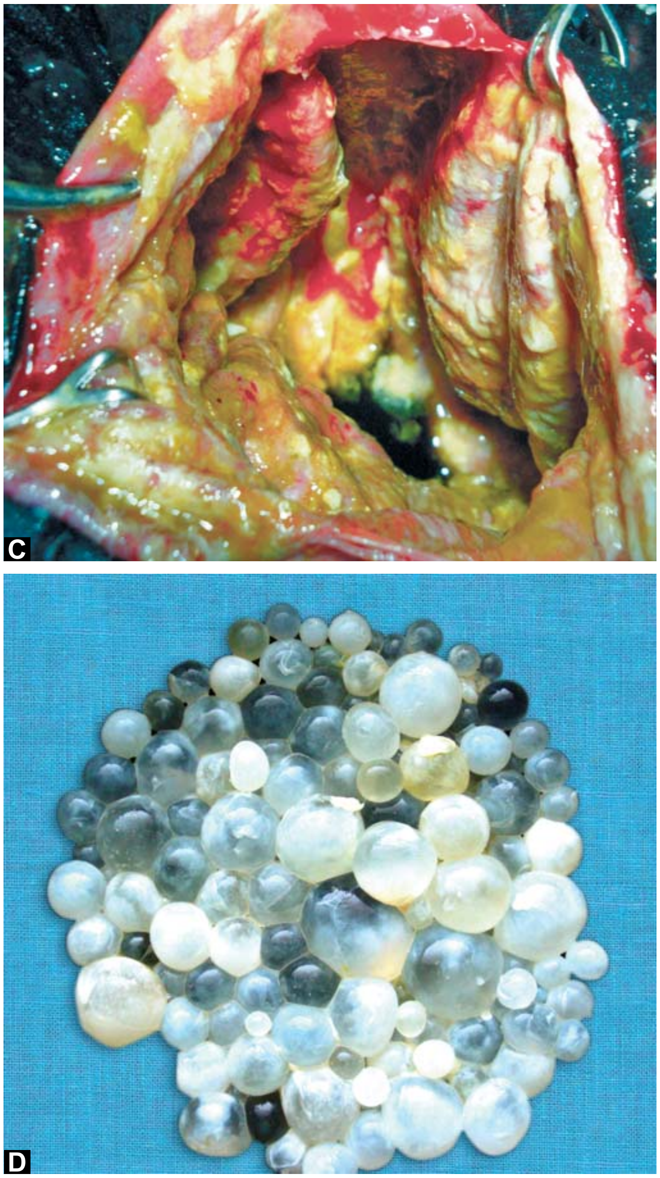

Figs 11.20C and D

Figs 11.20A to D: On table finding of hydatid cyst of liver. Black mops are used to identify and isolate scolices. Note the cavity containing bile suggestive of biliary communication.

- Detection of *cystobiliary communications* is very crucial as it may cause caustic sclerosing cholangitis when scolicidal agent like formalin is used. Communicating openings may be single or multiple. Cyst more than 10 cm is likely to have cystobiliary communications. Often clinical features of communications may not be present. Preoperatively there may be recurrent cholangitis; dilated bile duct. Factors important are—its size; number; site; involvement of hepatic/bile ducts; liver dysfunction. Finding of bile stained cyst on table during aspiration is highly suggestive of communication. White mops soaked with hypertonic saline are kept in the cyst cavity and gallbladder is gently squeezed to see for the bile staining of the mop in the cavity which confirms communication.
- *Laparoscopic pericystectomy* is becoming more popular. Contraindications are deeply situated cyst, densely adherent cyst, and inaccessible cysts; more than 3 cysts; calcified cysts and cysts in other organs. Main problem with laparoscopic pericystectomy is spillage and difficulty in preventing it.
- Liver resection—only occasionally segmental or hemihepatectomy is done.

Scolicidal agents

- Cetrimide—can cause acidosis
- Alcohol 80%—can cause cholangitis
- Hypertonic saline—hypernatraemia
- Sodium hypochlorite—hypernataraemia
- Hydrogen peroxide

In cases with biliary communication only hypertonic saline (15-20%) is used (not other agents)

Procedures used to correct the cystobiliary communications and to obliterate the cavity are:

- Suturing of the communication—simple suture using vicryl/PDS suture with T tube drain of bile duct.
- Pericystectomy (*Pericystectomy* is done by peeling off the cyst wall and abdomen is closed with or without a drain); marsupialization.
- Tube drainage of the cavity; *capitonnage*—spiral suturing of the bottom of the cavity upward from base of cavity to the edge of the cyst wall; *introflexion*—inverting the rim of the cyst edge without apposition; *omentoplasty.*
- Internal drainage procedures like choledochojejunostomy; transduodenal sphincteroplasty; Roux-en Y cysto/intracystic hepatico jejunostomy; Roux-en-Y hepaticojejunostomy.
- In cystobiliary communication (fistula), cystobiliary disconnection is achieved anatomically between cyst and fistula and a multiperforated transparenchymal drain is placed into the fistula through the cyst and another multiperforated drain is placed into the cyst cavity. A 'T' tube drain is placed into the bile duct (*perdomo procedure*) to achieve complete external drainage with facilitation of closure of the communication.
- Bipolar drainage—a drain is placed to cyst cavity and a T tube is placed into common bile duct.
- Reconstructive procedure like *pericysto jejunostomy*; bile duct repair.
- ERCP sphincterotomy.

Malignant hydatid disease

- It is a misnomer as it is a benign condition
- It is caused by *Echinococcus multilocularis (Alveolaris*). It presents with multiple small cysts in both lobes of the liver, all over
- It is difficult to treat, and mimics clinically and prognosis-wise to malignancy, hence the name
- They die of liver failure

Every job is self portrait of the person who did it.

ACTINOMYCOSIS OF LIVER

- *Actinomyces israelii* reaches the liver—via portal vein from ceacum (60%) or via hepatic artery from distant focus—from faciocervical region—30%.
- Disease progresses gradually and slowly to destroy the liver tissue causing multiple actinomycotic abscesses which produce a typical '*honey comb liver.*'

Complications

- Secondary infection.
- Liver failure.

Investigations

- Liver function tests.
- Ultrasound guided aspiration and staining of pus shows '*sun-rays*' appearance.

Treatment

Penicillins for 3-6 months.

LIVER TUMOURS

BENIGN TUMOURS OF THE LIVER

Benign tumours of the liver

- It is two times common than malignant liver tumor
- It is seen in 20% of population
- Haemangioma is the most common benign tumor
- CT and MRI are the essential investigations
- MRI shows 95% accuracy
- Diagnostic accuracy of liver biopsy is low—40%
- Hepatic adenoma is potentially malignant
- FNH and haemangiomas are not premalignant
- Primary (HCC) tumour or secondaries are differential diagnosis; AFP and CEA are helpful to differentiate

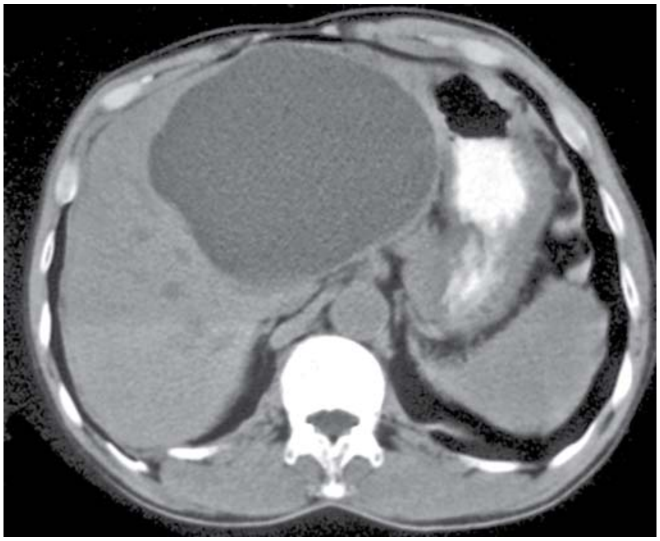

Fig. 11.21: Solitary benign cyst of the liver. It is a rare entity. Polycystic liver also can occur. Polycystic liver is associated with polycystic kidney disease. Condition is a rare differential diagnosis for hydatid cyst, hepatoma, and solitary secondary.

1. Haemangiomas

- They are the *commonest benign* tumour of the liver.
- It is usually solitary but can be multicentric.
- *Compressibility of tumour is diagnostic*.
- It is common in females (3:1) in 5th decade.
- Commonly they are asymptomatic.
- They can occur in right and left lobes.
- In children it is 10-15% of liver tumours. They are multifocal, often involves other organs also. Large one can cause AV shunting leading to CCF. Large childhood hepatic haemangiomas have got 70% mortality due to rupture, DIC, thrombocytopenia. It is confirmed by MRI. It is treated by embolization, radiotherapy and often resection.
- Cavernous haemangioma, consumption coagulopathy, thrombocytopenia is called as *Kasabach-Merritt syndrome*.
- LFT, AFP and CEA are normal.

Complications

- Bleeding rupture
- Thrombosis
- DIC
- Infection

- Complication usually occurs if haemangioma is *more than 8 cm* in size.
- ***Needle aspiration and biopsy are contraindicated.***
- U/S and CT scan are confirmative.
- Angiogram may be needed if surgical intervention is planned.
- MRI is diagnostic.
- Giant haemangioma is of size > 4 cm.

Treatment:

- Prior radiotherapy is given to reduce the size and then hemihepatectomy is done.
- Surgery is done when size is large, symptomatic, impending rupture or any other complications. Enucleation or anatomical resection with inflow control is needed.

***Note*:**
It has got very little or no malignant potential.

2. Hepatic adenomas

- They are common in females (10:1). They present as a solitary nodular lesions in the liver.
- It is said to be due to use of oral contraceptive pills (OCPs).
- It is uncommon in males.
- It is relatively rare compared to FNH and haemangioma.
- It is commonly solitary; but 20% of cases can be multiple; situation where > 10 adenomas are present is termed as *adenomatosis*. Multiple adenomas are not related to OCP intake and not more common in females.
- Upper abdominal pain (65%) is the commonest presentation.
- AFP is normal but rises if there is malignant transformation.
- They might turn into malignancy, hence resection is advised.
- U/S, CT scan are diagnostic but angiography is needed prior to resection, as it is vascular.
- MRI is ideal tool.
- Haemorrhage is more common.
- Hepatic adenoma cells are larger, contain glycogen and lipids. They are devoid of bile ducts whereas FNH contains bile duct components.

- Surgical resection is the needed treatment. Limited resection is needed. It is better to do resection prior to pregnancy. In adenomatosis either hemihepatectomy or transplantation may be required.
- OCP should be stopped.

Indications for surgical resection

- Rupture of adenomas—50%
- If there is possibility of turning into malignancy (large adenomas)

Note:

Kupffer cells are not present in hepatic adenomas.

3. Focal nodular hyperplasia (FNH)

- It is a benign condition of unknown aetiology, seen in females showing focal overgrowth of functioning liver tissue with fibrous stroma support.
- It contains hepatic cells as well as Kupffer cells which is characteristic.
- It is 2nd most common benign tumour.
- Usually it presents as solitary nodule.
- A sulphur colloid liver scan is diagnostic, shows 'hot spot' with a spoke wheel pattern—85%.
- It is a harmless condition.
- CT scan shows central scar with stellate distribution of the blood vessels.
- Telangiectatic FNH and hyperplastic with adenomatous FNH are nonclassic forms of FNH, which are 20% common and more common in men.
- It is totally benign tumour; but often difficult to differentiate from fibrolamellar type of HCC. AFP is normal.
- Vague pain abdomen may be the presentation.
- Rupture, bleeding are very rare.
- Histologically it can be typical (*with central scar*) or atypical (*without central scar*)—15%.
- Usually it does not require any treatment. OCP should be stopped.

Note:

Von Meyenburg's complexes are benign liver malformations that originate from embryonic bile ducts that fail to involute with cystic dilatation. It mimics metastatic liver disease. MRI is essential to diagnose.

PRIMARY MALIGNANT TUMOURS OF THE LIVER

1. Hepatocellular carcinoma/Hepatoma/HCC (80%).
2. Cholangiocarcinoma (20%).
3. Hepatoblastoma in infants and children.

Hepatoblastoma occurs within 2 years of life. *It is most common primary malignant tumour of liver in children*. It is common in male child. It is derived from fetal or embryonic hepatocytes. Serum AFP is *elevated* in 90% of cases. CT scan shows vascular mass with speckled calcification. It is highly sensitive to chemotherapy (vincristine, doxorubicin, 5FU). If it is resectable, it is the initial choice of therapy. Liver transplantation is also beneficial.

HEPATOCELLULAR CARCINOMA (HCC/HEPATOMA)

- Its incidence is rising.
- It is common in cirrhotics and hepatitis B and hepatitis C virus infection.
- It is common in Mozambique, South East Asia, tropical Africa, Taiwan. In Mozambique it is often seen in younger age group - below 30 years.
- Male to female ratio is 4:1.
- It is usually unicentric but occasionally can be multicentric.
- Right lobe is commonly involved.

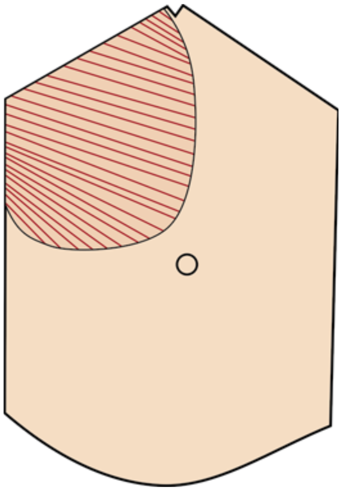

Fig. 11.22: Hepatocellular carcinoma (HCC/hepatoma) is usually solitary, common in right lobe, attains large size.

Aetiology

- Aflatoxin B_1, a product of fungus aspergillus. It is powerful carcinogen.
- Hepatitis B and hepatitis C virus infection. It is more common in individuals who have chronic positive status for HBs Ag and chronic carriers. Vertical transmission of virus at birth raises HCC rate. In Europe, Japan, USA, HCV infection is more common cause. It is not necessary to develop cirrhosis to cause HCC. It is the infection which is the cause.
- Alcoholic cirrhosis. It is co-carcinogen.
- Clonarchis sinensis infestation.
- Smoking.
- Haemochromatosis, α1 antitrypsin deficiency.
- Hepatic adenoma—potentially malignant.
- Environment related chemicals like DDT, nitrite and nitrate related food products; trichloroethylene (dry cleaning solvents), biphenyls, carbon tetrachloride, herbicides, solvents like dioxane.
- Tannins, griseofulvin, bush tea (pyrrolizidine), rice toxins.
- Anabolic steroids, polyvinyl chloride.

Pathology

Gross

It is highly vascular with indistinct margin. Often it is well demarcated by fibrous tissue. Haemorrhage and necrosis are common.

- *Hanging type of tumour*—tumour attached to normal liver by a small vascular stalk—always very well resectable.
- *Pushing type of tumour*—large, well demarcated, displaces normal vasculature—resectable.
- *Infiltrative type of tumour*—indistinct, infiltrative—often difficult to resect.

Histology

- Well/moderate/poorly differentiated tumour.

Okudas classification—multifocal; indeterminate; spreading; expanding.

Fibrolamellar Variant of HCC (FL HCC)

- FL HCC occurs in *younger* age group.
- Incidence is *equal* in both sexes.
- It does not show any elevated levels of serum AFP.
- *Tumour marker for FL HCC type is increase in serum vitamin B_{12} binding protein and increase in neurotensin levels.*
- It is not related to viral hepatitis or cirrhosis.
- Left lobe is commonly involved.
- It involves lymph nodes more commonly than HCC.
- CT scan shows characteristic central scarring.
- Fibrous stromas with thin hyaline bands are typical. Sheets of well differentiated hepatocytes are seen which are sandwiched between collagen and fibroblasts.
- It is fairly resectable (50-75%) with better prognosis than HCC.

Okuda staging system (1984) for HCC

Parameter		***Points***	***Stage***
Tumor size	< 50 %	0	Stage one—0 points
	> 50 %	1	Stage two—1-2 points
Ascites	Absent	0	Stage three—3-4 points
	Present	1	
Serum	< 3 mg %	0	
Albumin	> 3 mg %	1	
Serum	< 3 mg %	0	
Bilirubin	> 3 mg %	1	

Staging of HCC (AJCC)

T_0 – No tumour
T_1 – Solitary tumour without vascular invasion
T_2 – Solitary tumour with vascular invasion / multiple tumours equal or less than 5 cm
T_3 – Multiple tumours, more than 5 cm/tumour invading major branch of portal or hepatic veins
T_4 – Tumour with direct invasion of adjacent organs other than gallbladder or to visceral peritoneum
N_0 – No regional nodes are involved
N_1 – Regional nodes are involved
M_0 – No distant spread
M_1 – Distant spread is present

Cancer of liver Italian group scoring (CLIP scoring)

Parameters	***Values***	***Scores***
Child Pugh's stage	A B C	0 1 2
Tumour morphology—Gross	Uninodular less than 50% of extension Multinodular less than 50% of extension Massive—extension more than 50% of liver	0 1 2
AFP ng/dl	Less than 400 More than 400	0 1
Portal vein thrombosis	No Yes	0 1
Total scores 0-6 Early disease with potential long term survival 0-3 Advanced stage 4-6		

Chinese university prognostic index (CUPI) scoring system uses TNM staging, symptoms, ascites, bilirubin, alkaline phosphatase, AFP. It is used mainly for HBV induced HCC.

Clinical Features

- Painless mass in right hypochondriac region with loss of appetite and weight. Liver is hard, smooth/irregular and often massively enlarged. Cirrhotic liver may be nodular.
- Acute presentation is not uncommon, when the tumour undergoes necrosis and haemorrhage. Both types of presentations mimic amoebic liver abscess (Haemoperitoneum is also known).
- Jaundice when present is commonly due to hepatic dysfunction, but occasionally due to compression of bile ducts.
- Ascites (40%) often it is massive, splenomegaly and features of portal hypertension may be present.
- Hepatic thrill and bruit—25%.
- Fever (10-20%) may be present due to tumour necrosis.
- Dull aching pain in right upper quadrant is common.
- Features of chronic liver disease—jaundice, dilated veins, palmar erythema, gynaecomastia, testicular atrophy, etc.
- Liver failure sets in once tumour replaces the functioning liver parenchyma or portal vein gets occluded by tumour thrombus.
- Gastrointestinal bleeding may be the presentation in 10% of cases due to portal hypertension or portal vein invasion by tumour.
- Occasionally may present with paraneoplastic syndromes. 1%; hypercalcaemia, hypoglycaemia, hyperlipidaemia, hyperthyroidism, erythrocytosis.

Spread of Tumour

- *Lymphatic spread*: It can spread to other part of liver through lymphatics within the liver, to the lymph nodes in the porta hepatis and other abdominal lymph nodes later. Often spread occurs directly to cisterna chyli.
- *Blood spread*: To lungs, bones and adrenals often can occur.
- *Direct infiltration*: To diaphragm and neighbouring structures.

Differential diagnosis

- Secondaries in liver
- Polycystic disease of liver
- Amoebic liver abscess
- Hydatid cyst of the liver

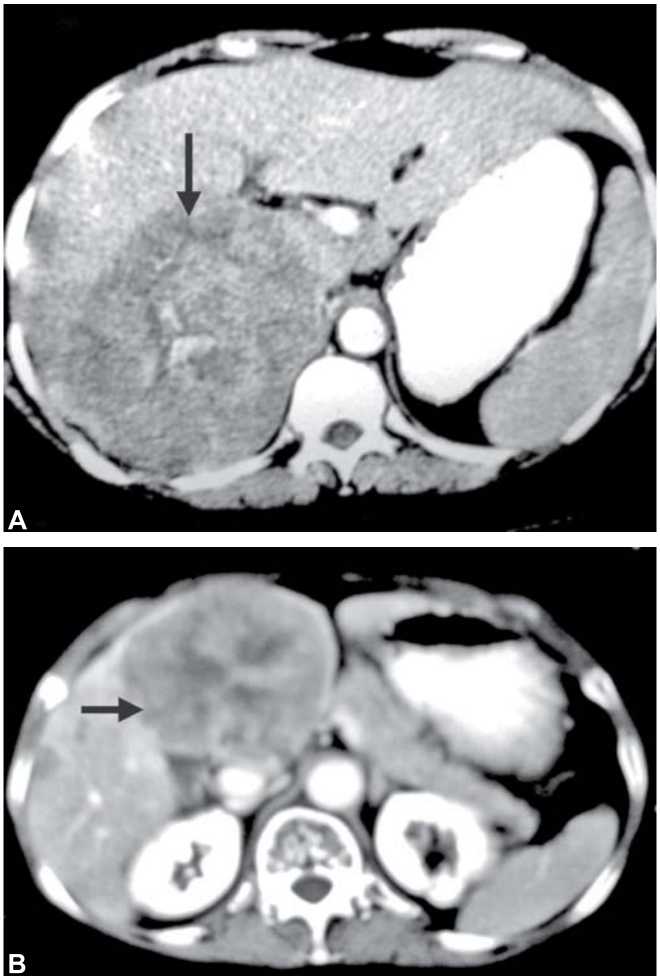

Figs 11.23A and B: CT pictures of hepatoma. In right lobe and left lobe.

Investigations

- U/S abdomen—very useful method. It shows hyperechoic mass; mosaic pattern with thin halo and lateral shadows. Extent, tumour thrombi extension can be made out.
- CT scan abdomen (CECT) more reliable and ideal (hypodense, mosaic, vascular lesion with irregular margin).
  - reveals the size, location and extent, vascularity, portal vein invasion, nodal status, portal vein thrombosis.
  - helps to assess operability.
- *Tumour markers—α feto protein (AFP).* AFP will be raised more than 100 IU; as high as 1000 IU is possible in HCC. PIVKA II is des γ carboxy prothrombin protein induced by vitamin K abnormality/antagonists 2. It is increased in 80% of HCC patients. AFP is usually more than 400 IU in 70% of HCC patients.
- Celiac angiography/CT angiography—it shows the vascularity of the tumour, tumour blush, arterial pattern of the liver and tumour; venous phase can show portal vein invasion or thrombosis and invasion/spread to IVC. Angiogram is essential while planning for hepatic resection.
- Liver function tests like serum bilirubin, albumin, enzymes (alkaline phosphatase, transaminase, 5' nucleotidase) including prothrombin time.
- Liver biopsy is done after controlling prothrombin time if it is elevated (by Inj vitamin K IM 10 mg for 5 days; FFP transfusion). US / CT guided core liver biopsy is useful and mandatory before clinical trials. It is better than FNAC as it reveals tissue architecture. Problems with the core liver biopsy are spillage of tumour and bleeding (due to hypervascularity, associated thrombocytopenia and reduced liver dependent clotting factors like prothrombin). Such problems may be minimal in FNAC. Patients with high suspicious of HCC by clinical and imaging methods and who are appropriate for surgery may be taken up for surgery without a preoperative biopsy. In case of portal vein infiltration by tumour a portal vein biopsy may be done to exclude the patient for surgery or transplantation.

Note:

Preoperative liver biopsy is *not recommended* now. Biopsy is relatively contraindicated in any candidate for resection, as there is a significant risk of haemorrhage and peritoneal implantation of tumour cells. In case of potentially resectable tumour one should proceed to surgical exploration without tumour biopsy. Biopsy is done only in inoperable advanced cases wherein palliative therapies are undertaken to establish the diagnosis.

- Contrast MRI/CT scan 1-2 weeks following intrahepatic infusion of lipiodol (ethiodized oil emulsion 5-15 ml) with contrast agent through hepatic artery is very useful.
- MRI—T_2 weighted studies are useful for small HCC. MR angiography is also done to see tumour thrombus in portal vein, hepatic vein and IVC.
- Ascitic tap when ascites is present for cytology.
- Laparoscopic evaluation and laparoscopic US is useful for proper assessment of the tumour.
- Investigations in relation to hepatitis B and hepatitis C virus infections.
- Metastatic work up—HRCT of chest is essential.

Treatment

Definitive Treatment

- When limited to one lobe, *hemihepatectomy i*s done (Removal of 80% liver is compatible with life). Laparoscopic resection is becoming popular. It is done only to Child's grade A and favourable Child's B group patients. It is the treatment of choice for operable HCC in noncirrhotic patients. There is no regenerative capacity in cirrhotic patients with poor coagulation status, portal hypertension, varices, and ascites.
- In cirrhotic patients with HCC, *total hepatectomy* with *orthotopic liver transplantation* is required. Now it is found

The best bridge between despair and hope is a good night's sleep.—**E. Joseph Cossman** that even resection candidates will do better by transplantation than resection. But waiting period for transplantation is long. To bridge this waiting period for transplantation radiofrequency ablation, transarterial embolisation, ethanol/acetic acid injections are used.

Note:
Liver mass which is confirmed by CT/MRI and serum AFP is > 500 ng/dl is almost diagnostic and resection treatment can be started without tissue diagnosis.

Indications for transplantation in HCC

- Patient who is not a candidate for resection
- Tumour less or equal to 5 cm
- Tumour less than 3 in number
- Tumour without portal/hepatic vein invasion
- Tumour without extra hepatic spread

Palliative Treatment

- *Radiofrequency ablation (RFA)*: It is thermal ablation of the tumour by passing 18G needle into the middle of the tumour and passing electric current of 500 kHz. This creates frictional heat causing sphere of necrosis. Maximum zone of necrosis which can be created is 7 cm for a 5 cm tumour. This causes adequate cyto-destruction of the tumour. Tumour close to vessels and biliary tree, tumour more than 3-5 cm and multicentric tumours are difficult to manage by radiofrequency ablation. One course ablation takes around 20 minutes. It can be done percutaneously under US or CT guidance or through laparoscopy. Percutaneously it can be done on outpatient basis. It can be repeated several times. Tumour less than 3 cm, tumour deep in parenchyma, tumour away from hilum is ideal for radioablation.
- *Percutaneous ethanol or acetic acid injection:* (*Rule of 3* is used here—indicated in HCC not >3 cm and not > 3 nodules). Ethanol injection is minimally invasive, can be injected using fine needle, and is cheaper. Local recurrence is lower with acetic acid injection than ethanol. But radiofrequency ablation has got better result than ethanol/acetic acid injection. RFA is costly and not available in many centers.
- *Intra-arterial chemotherapy* using Adriamycin/cisplatin/mitomycin through gastroduodenal artery. It is often given along with lipiodol to make drugs to stay longer time in the liver tissue so that better results are achieved. This regional chemotherapy is much more successful in HCC.
- *Intra-arterial embolisation* using gelfoam, microspheres, gelatin sponge, or chemoembolisation.
- *Microwave or cryo ablations* similar to RFA.
- As it is considered that tumour tissue exclusively gets its blood supply from hepatic artery, *ligation of hepatic artery* can be a good palliation (to achieve tumour necrosis) for HCC. Normal tissue will still have its blood supply from portal vein.
- Transarterial I^{131} lipiodol, Yttrium 90 microspheres.

Adjuvant Therapy

- Systemic chemotherapy using intravenous adriamycin (doxorubicin), cisplatin, carboplatin, mitomycin C, 5 fluorouracil.
- Octreotide is used along with other chemotherapeutic agents. It decreases the size of the tumour.
- Neoadjuvant (preoperative) chemotherapy also practiced to increase the operability/resectability of the tumour.
- Newer treatments are tried using tamoxifen, interferons, EGFR antibody, interleukins, I^{131} lipiodol, Yttrium^{90} microspheres, etc.

Neoadjuvant therapy in HCC

Large primary tumour—downstaging/*downsizing* may make resection possible

- Transarterial chemoembolisation (TACE)
- Chemotherapy and radiotherapy
- Hepatic artery infusion of chemotherapy (HAI)
- Radioimmunotherapy
- Fractioned radiotherapy
- Transarterial ^{90}Y microspheres

Treatment strategies for HCC

- Surgical resection, only 20% have resectable liver
- Liver transplantation
- Radiofrequency ablation
- Percutaneous ethanol/acetic acid injection
- Transarterial embolisation/chemoembolisation
- Microwave/cryoablation
- Transarterial radiotherapy
- Transarterial Yttrium^{90} microspheres/I^{131} lipiodol
- Hepatic artery ligation
- Adjuvant systemic chemotherapy
- Immunotherapy/hormone therapy/growth factors

Hepatoma/hepatocellular carcinoma/HCC

- Common etiologies are aflatoxins, hepatitis B and hepatitis C virus infection, alcoholic cirrhosis, haemochromatosis, smoking, hepatic adenoma, clonorchis sinensis, polyvinyl chloride
- Unicentric and right lobe involvement is more common
- Fibrolamellar variant is common in left lobe, not related to hepatitis or cirrhosis without AFP level raise. There are increased serum vitamin B_{12} binding capacity and neurotensin levels
- Invasion to portal veins is common and it has got negative prognosis with lesser chances of resection or transplantation. It may cause portal vein thrombosis and so splenomegaly
- It can be multifocal/indeterminate/spreading/expanding—*Okuda classification*
- Presents as large, smooth, hard liver mass—later jaundice, fever, pain and tenderness, ascites and bruit over mass
- Spreads by lymphatics, blood and direct extension
- Mimics amoebic liver abscess, secondaries, hydatid cyst, polycystic liver disease
- LFT, CT scan, raised AFP, liver biopsy are the investigations
- *Liver mass, confirmed by CT/MRI and serum AFP > 500 ng/dl is almost diagnostic* and treatment can be started without tissue diagnosis

- Liver biopsy is done (FNAC) only in patients with nondiagnostic level of AFP and not a candidate for curative surgery but only for palliation
- Hemihepatectomy is the treatment in early operable growth
- Liver transplantation is the good surgical option
- Hepatic artery ligation/intra-arterial chemotherapy/chemoembolisation/percutaneous ethanol or acetic acid injection/radiofrequency ablation/chemotherapy using adriamycin, carboplatin, gemcitabine—are palliative procedures

SECONDARIES IN THE LIVER

It is the commonest malignant tumour in liver. The incidence of primary : secondary : : 1 : 20.

Secondaries are common malignant tumours

- In bone
- In liver
- In brain

Causes

***Abdominal*:** Carcinoma in stomach, colon, pancreas, small bowel, kidney, abdominal oesophagus, rectum, carcinoids.

Extra-abdominal:

- Melanoma.
- Carcinoma breast, lung, thoracic oesophagus, bladder, prostate.
- Testicular and adrenal tumours.
- Follicular carcinoma thyroid (FCT).

Commonly secondaries in liver are multiple, multiple in one lobe or in both lobes. It can be solitary (rare).

It is often also classified as:

Colorectal

- Colorectal carcinoma has got special interest in relation to its secondaries in liver. It carries good prognosis.
- Secondaries can be *synchronous or metachronous.*
- Resection of secondaries in liver often gives good outcome.
- PET scan is done to identify occult extrahepatic metastases. CEA is reliable indicator of spread or recurrence.
- Resection, chemotherapy, intra-arterial chemotherapy are therapeutic modalities.
- *Contraindications for liver resection* are—extrahepatic spread, nodal involvement along with primary, short disease free interval less than 1 year, four or more secondary deposits, bilobar spread, CEA more than 200 ng/dl, secondary more than 5 cm in size, involved histologic margin.
- Colorectal liver secondaries show good prognosis when secondary are solitary, and with absence of all poor prognostic indicators.
- 5 FU with irinotecan or oxaliplatin, bavacizumab (antivascular endothelial antibodies) are also useful.

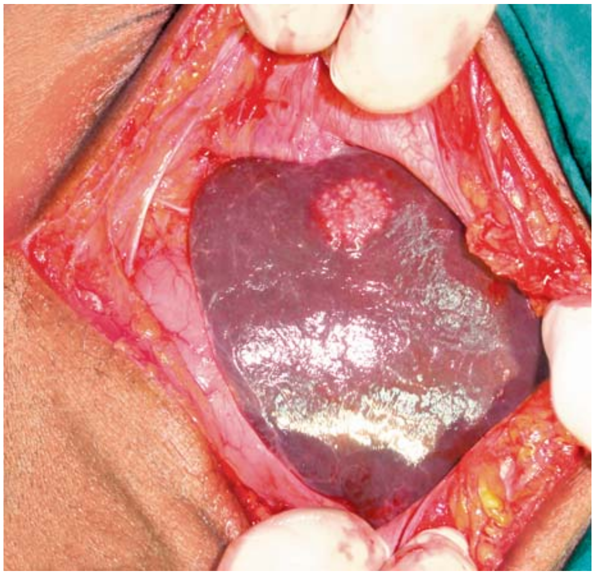

Fig. 11.24: Secondaries in liver—solitary type from colonic primary.

Neuroendocrine

Gastrinoma, glucagonoma, somatostatinoma, nonfunctioning neuroendocrine tumours can cause liver secondaries.

- These tumours are slow growing.
- Main symptoms are often due to secreting peptides causing different clinical syndromes.
- Hormonal assay, CT abdomen, CT angiogram are needed.
- Hormone controlling therapies, chemotherapy and cytoreduction using liver resection gives good 5-year survival up to 75% in these patients.

Non-colorectal and Non-neuroendocrine

- These are aggressive primary tumour causing liver secondaries like from stomach, pancreas, small bowel, oesophagus, gallbladder, melanoma, urinary bladder and prostate.
- They carry poor prognosis.
- Treatment is chemotherapy. Early primary like—in stomach with isolated liver secondaries without extrahepatic spread can be benefited with liver resection.
- But overall prognosis is guarded.

Other classifications

- *Precocious:* Secondaries are presented/identified before primary is suspected—carcinoid, colorectal cancer.
- *Synchronous*: Both primary and secondaries are identified at the same time—carcinoma stomach.
- *Metachronous*: Secondaries develop much later may be long time after the treatment for primary—melanoma, breast carcinoma.

Route of Spread

a. *Direct spread*: Stomach, colon, gallbladder, bile ducts**.**
b. *Hepatic artery*: Melanoma.
c. *Portal vein*: Carcinoid tumours, other GIT malignancies.
d. *Lymphatic spread*: Breast, lung.

A superior man is modest in his speech, but exceeds in his actions.

Clinical Features

- They are multiple, hard, nodular, with nodule showing umbilication due to *central necrosis.*
- Patient may have jaundice depending on the extent of liver involvement.
- Loss of appetite and weight.
- Ascites.
- Pain in liver secondaries is due to stretching of liver capsule/ tumour necrosis. Often pain may be of primary tumour.
- Rectovesical secondaries—'Blumer shelf'.
- Supraclavicular lymph node enlargement.

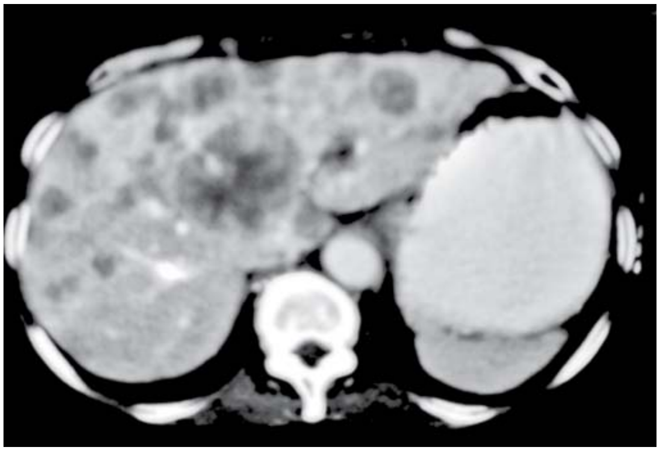

Fig. 11.25: Secondaries in liver. Note the multiple lesions (nodules) in both lobes of the liver.

- Clinical features of primary tumour, e.g. stomach, colon—
  - Stomach: Vomiting; visible gastric peristalsis; palpable stomach mass.
  - Colon: Constipation/diarrhoea; palpable mass; features of obstruction.
  - Pancreas: Jaundice; steatorrhoea; itching; palpable gallbladder.
  - History of earlier surgery: Like mastectomy/wide excision of the melanoma.
  - Testis: Fullness; hard mass; loss of testicular sensation.

Clinical features of secondaries in the liver

- Poor general condition
- Palpable liver which is hard, multinodular with umbilication
- Clinical features of primary
- Ascites ±
- Jaundice ±

Differential Diagnosis

- Multicentric hepatoma.
- Macronodular cirrhosis.
- Polycystic liver disease, hydatid cyst of liver.

Investigations

- Primary is identified by—gastroscopy, colonoscopy, contrast X-rays, CT scan.

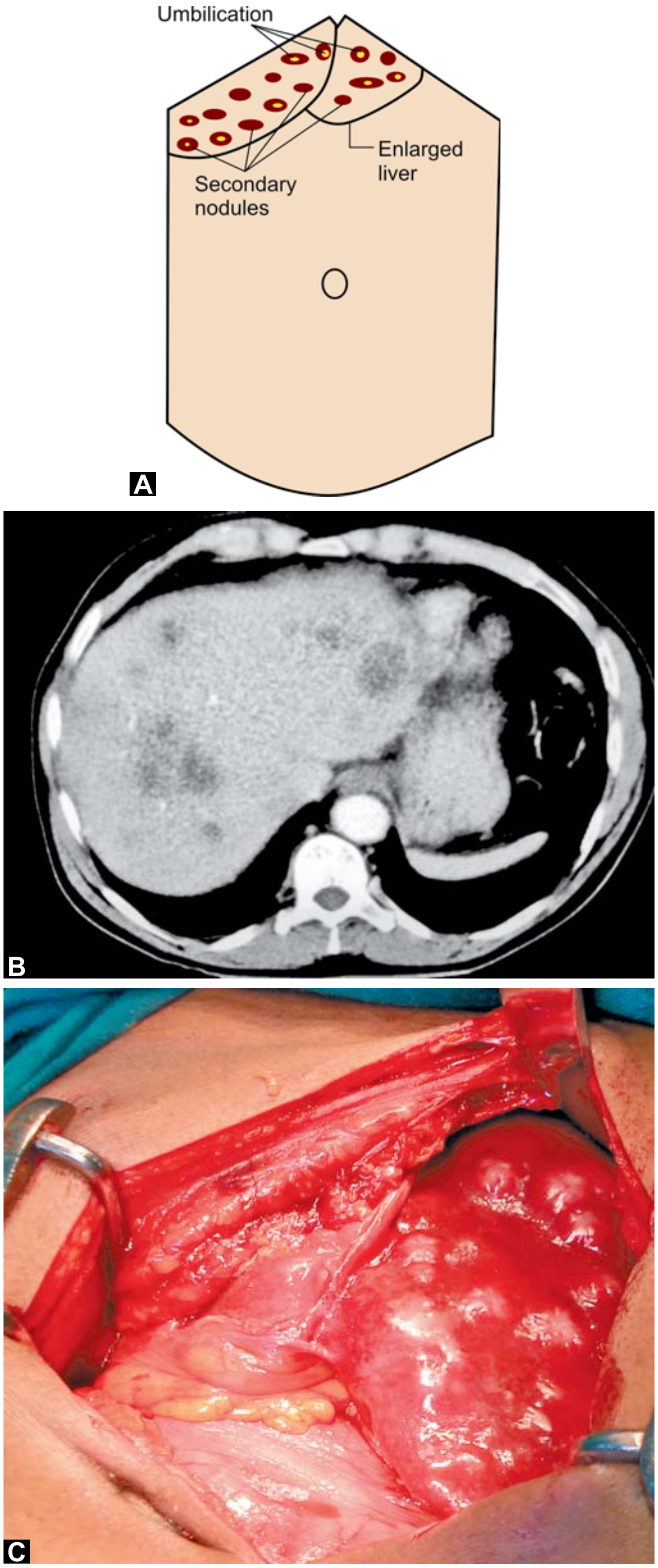

Figs 11.26A to C: Secondaries in liver are multiple and often with umbilication. CT picture shows multiple secondaries with ascites. On table photo of multiple liver secondaries with umbilication.

- Liver biopsy is done, only if primary is not identified by any of these methods.
- Liver function tests.
- U/S abdomen—multiple hypo-or hyperechoic lesions; site of possible primary; nodal status; ascites.

- Tumour marker: CEA.
- MRI/CT scan whenever resection is indicated.
- Laparoscopy often with laparoscopic US is very useful to detect small secondaries.

Treatment

- Chemotherapy is the treatment of choice.
- *Surgery has a definitive role if primary is from the colon, kidney.*
- If the primary growth is in the colon, then a palliative hemicolectomy is done. Along with that if a solitary secondary in the liver is present, it can be resected. If secondaries are in single lobe, hemihepatectomy is done.

Criteria for selection for liver resection in secondaries in liver, when primary is in colon are-
- Stage of the primary
- Type of primary resection
- Size of the largest liver secondary
- Number of secondaries
- Segments involved
- Associated liver disease like cirrhosis

- Palliative procedures for other primaries like—for carcinoma stomach—palliative partial gastrectomy/anterior gastrojejunostomy; for carcinoma pancreas—triple anastomosis/ERCP stenting.
- Depending on the site of primary, drugs vary.
  - For carcinoma stomach, mitomycin and 5FU are given.
  - For carcinoma colon, 5-FU, oraloplatinum/VEGF or EGFR antibodies used with folinic acid.
- Hepatic artery ligation or therapeutic embolization with clot/gel foam can be tried to relieve pain.
- Intra-arterial chemotherapy is tried with little success.

Methods other than resection available to ablate the tumour locally in secondaries in liver
- Microwave therapy
- Radiofrequency ablation
- Laser hyperthermia
- Ultrasound or electrolyte therapy

Overall outcome is poor in liver secondaries.
They die of malignant cachexia, infection, liver failure.

Causes of massive enlargement of liver
- Hepatoma
- Secondaries in liver when primary is from melanoma, carcinoid tumour
- Syphilitic enlargement of liver (hepar lobatum)

LIVER CYSTS

CONGENITAL LIVER CYSTS

Simple Liver Cyst

- It is 3% common. It is common in males (2:1).
- Solitary simple cyst is congenital; developing from abnormal intrahepatic bile duct *in utero*.
- 1/4th of simple cysts are symptomatic only at 4th or 5th decade.
- It is thin bluish walled cyst in one or other lobe containing clear fluid. Simple cyst will never communicate with biliary system. If communication is found it should be cystadenoma.

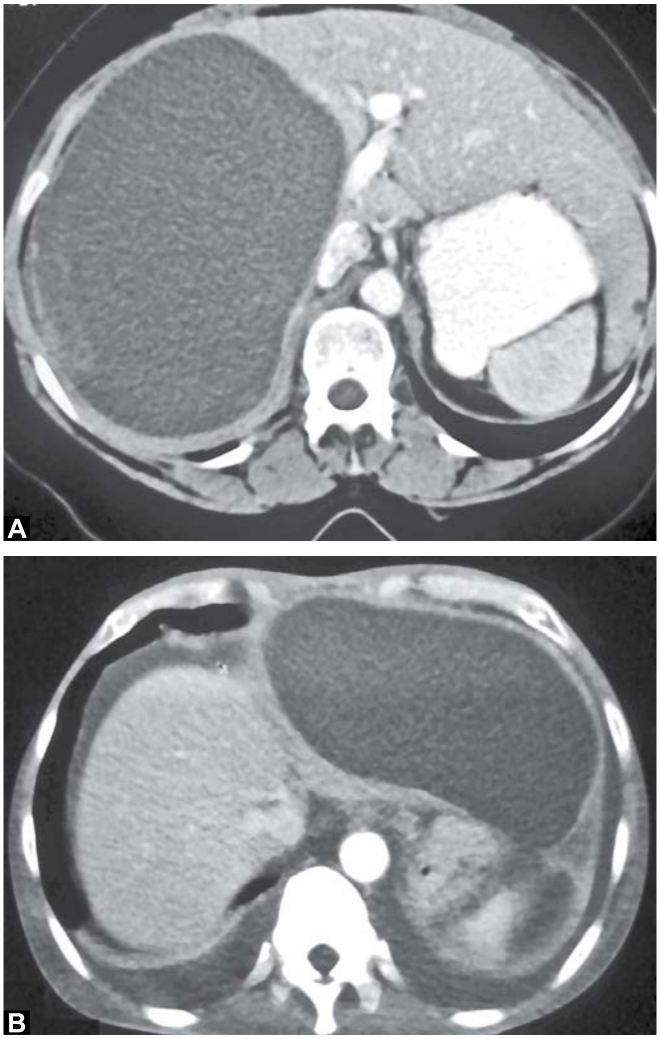

Figs 11.27A and B: Simple liver cysts—CT pictures in two different patients.

Venous hum in epigastrium observed in portal HT is called as Kenawy's sign.

- 50% present with abdominal pain. Hepatomegaly is common. Jaundice can be present.
- *Differential diagnoses* are—hydatid cyst, cystic neoplasms, solitary secondary with tumour necrosis.
- US shows anechoic lesion with smooth margin.
- CT scan shows nonenhancing thin lesion. MRI shows hypointense T_1 image and hyperintense T_2 image.
- Aspiration confirms the diagnosis. Aspiration will not cure the cyst as it recurs 100%.
- *Treatment* is—aspiration with sclerosant therapy where the recurrence rate is 20%. Deroofing with partial excision of cyst wall, by laparoscopic or open method has recurrence rate of 10%. Complete excision has very less recurrence rate but surgical mortality of 2% is present.

Polycystic Liver Disease

- It is autosomal dominant condition associated with polycystic kidney disease in 40% of cases.
- It arises from biliary epithelium but doesn't communicate with biliary system.
- Usually they are asymptomatic.
- Hepatomegaly, pain abdomen, jaundice and portal hypertension are the presentations.
- Haemorrhage, biliary obstruction, Budd-Chiari syndrome, rarely malignant transformation - are the *complications*.
- Liver function is well preserved and so liver failure will not occur.
- US, CT, MRI are investigations done.
- *Treatment*: Percutaneous aspiration with sclerosant injection, deroofing of all cysts, resection, liver transplantation.

NEOPLASTIC LIVER CYSTS

- They are *acquired* cysts.
- They are common in *females*.
- They attain large size; multiloculated with papillary projections.
- 10% of neoplastic cysts are malignant.
- Cystadenoma is benign containing *mucinous or bilious* fluid.
- Cystadenocarcinoma contains *haemorrhagic* fluid.
- AFP, CEA are normal; CA 19/9 is elevated.

Hepatobiliary (Hepatic) Cystadenoma

- Incidence is 5% of biliary intrahepatic cysts.
- 50% are in right lobe; 30% are in left lobe; 15% bilateral.
- Cystadenoma *without mesenchymal stroma* is equal in both sexes and occurs in 5th decade. It is not associated with cystadenocarcinoma.
- Cystadenoma *with mesenchymal stroma* is common in *young females* and is often associated with cystadenocarcinoma.
- It is septated, multiloculated, whereas if cavity contains debris it suggests malignancy. It is surrounded by a smooth and thick fibrous capsule. It contains internal septations and intraluminal papillary projections and lined by cuboidal or columnar biliary epithelium.

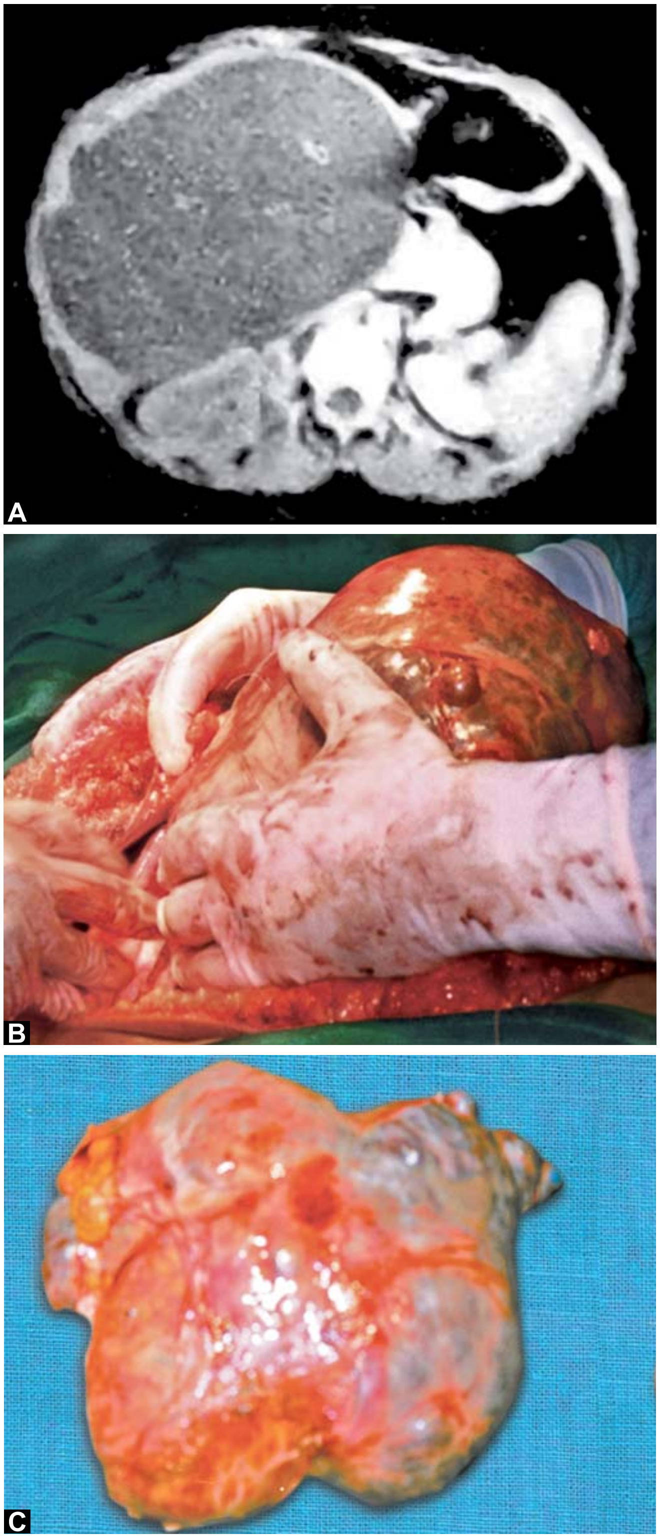

Figs 11.28A to C

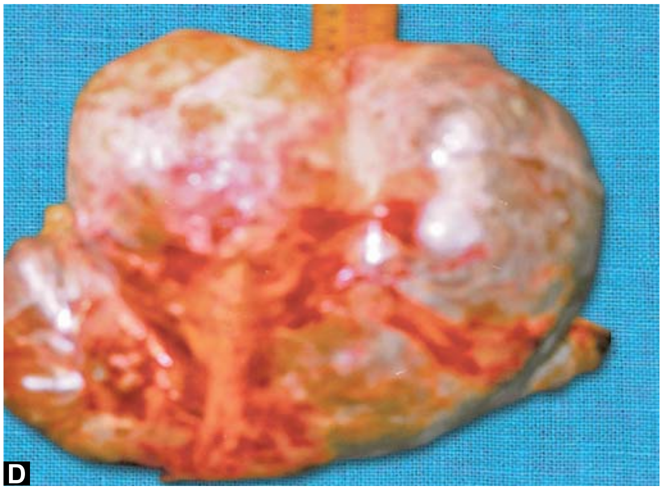

Fig. 11.28D

Figs 11.28A to D: Hepatic cystadenoma of right lobe of liver—CT picture and on table look (*Courtesy:* Dr Ashfaque Mohammed, DNB, KMC, Mangalore).

- It originates from—congenital aberrant bile duct/embryonic foregut cells/peribiliary endocrine cells.
- MRI is diagnostic. US/CT scan is also useful.
- ERCP is done to identify communication with biliary system; usually to proximal left hepatic duct.
- *Treatment*—enucleation of entire cyst with surrounding rim of normal liver tissue. Hepatic resection is also an option.
- *Note:* Aspiration/deroofing, partial excision, sclerotherapy *should not be d*one.

Hepatic Cystadenocarcinoma

- It presents as large palpable smooth liver with abdominal pain.
- It carries better prognosis than HCC or cholangiocarcinoma.

It can be:

  - Cystadenocarcinoma arising from preexisting cystadenoma with mesenchymal stroma, which occurs only in young females, carries good prognosis.
  - Cystadenocarcinoma without mesenchymal stroma not associated with cystadenoma, which is common in males and is more malignant.
  - Cystadenocarcinoma without mesenchymal stroma occurring in women.

- Resection is the treatment. It carries good prognosis.

TRAUMATIC LIVER CYST

- It is an *acquired cys*t of liver after liver injury in a blunt abdominal trauma. Biliary duct which is injured will cause bile leak into the existing intrahepatic haematoma causing mixture of bile and blood. Cyst is pseudocyst in liver without an epithelial lining.
- It may resolve spontaneously or may cause abdominal pain, jaundice. It may present months or years later.
- CT scan, old history of liver trauma may be helpful. MRI is diagnostic.
- *Treatment* is conservative. Often deroofing or cyst excision with ligation of biliary ductile may be needed but technically demanding.

PORTAL HYPERTENSION

- *Definition*: Sustained elevation of portal pressure more than 12 mm Hg (normal 8-12 mmHg).

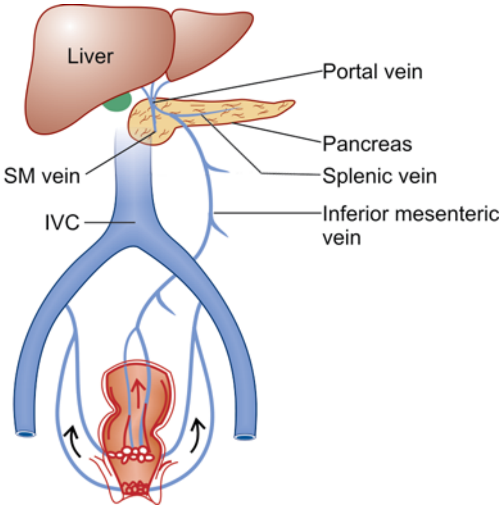

Fig. 11.29: Anatomy of the portal system.

- Left sided portal hypertension ***(sinistral)*** can be caused by isolated splenic vein thrombosis, which is often caused by adjacent pancreatitis.
- A rise in portal pressure stimulates portasystemic circulation.
- Portal vein carries 75% of blood flow to liver with all nutrients to maintain its integrity and gives 50% oxygen supply to liver. 25% hepatic arterial blood flow gives remaining 50% of oxygen supply to liver.
- Portal hypertension causes compensatory portosystemic venous collateral formation, altered intrahepatic circulation and increased splanchnic blood flow. High pressure portal blood is diverted via coronary (left gastric vein), short gastric and esophageal veins into azygos venous system.
- There is increased portal resistance and altered portal blood flow. Increased resistance may be presinusoidal, sinusoidal and postsinusoidal.
  - *Presinusoidal*
    - Extrahepatic—portal vein/splenic vein thrombosis.
    - Intrahepatic—schistosomiasis, sarcoidosis.
  - *Sinusoidal* – cirrhosis, haemochromatosis, Wilson's disease, congenital hepatic fibrosis.
  - *Postsinusoidal* – Budd-Chiari syndrome, conges-tive heart failure, veno-occlusive disease.
- There is *dysregulation* of compensatory increased hepatic arterial flow response in relation to decreased portal vein flow.
- 30% of varices patients will bleed; 30% of them will die of bleed; 30% of patients with cirrhosis will develop portal hypertension; 30% of them have variceal bleed in 2 years; 70% of patients who had bleeding once, will rebleed later.
- Variceal bleed accounts for 7% of upper GI bleed.

Glyceryl trinitrate should be given along with vasopressin to prevent coronary vasospasm.

- In western country more than 90% cases cirrhosis is the cause. In India extrahepatic portal vein obstruction (33%) and extrahepatic periportal fibrosis (33%) is becoming more common. As liver is normal in these two conditions, patients get better cure rate by shunt surgery.

Sites of Portosystemic Collateralisation

1. *Lower end of oesophagus*, between left gastric and short gastric veins with azygos vein resulting in *oesophageal varices*—commonest.

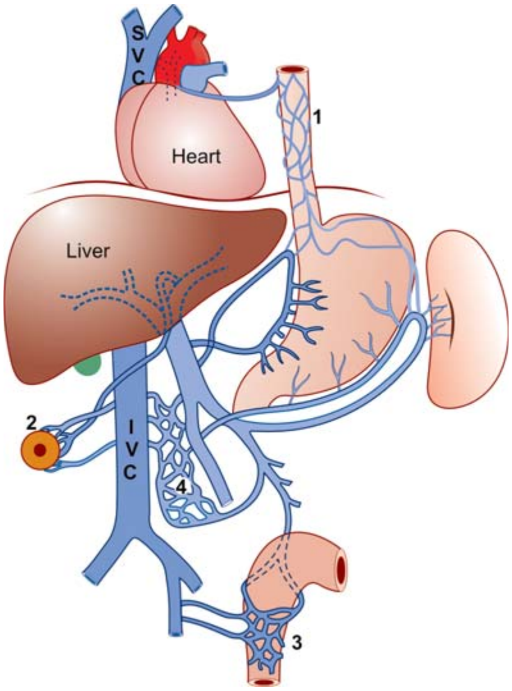

Fig. 11.30: Portosystemic collaterals.

2. *Umbilicus,* between paraumbilical vein and anterior abdominal vein resulting in *caput medusae.*
3. *Lower end of rectum,* between superior haemorrhoidal vein and inferior, middle haemorrhoidal vein resulting in *piles.*
4. *Retroperitoneum (*vein of Retzius).
5. *Bare area of the liver.*

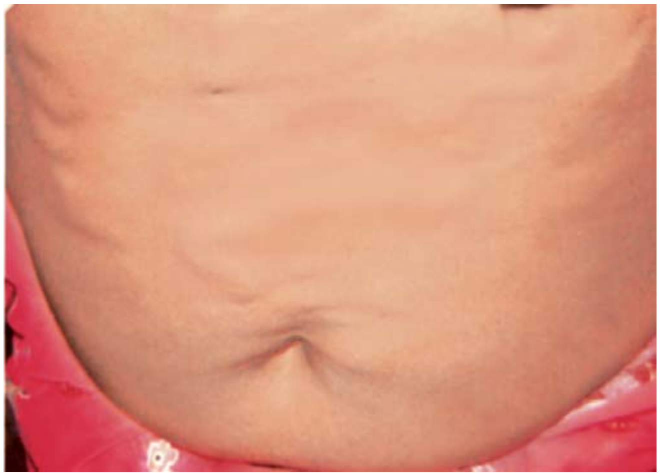

Fig. 11.31: Photo showing caput medusae.

Presentations

- *Triad of portal hypertension*
  - i. Oesophageal varices.
  - ii. Splenomegaly.
  - iii. Ascites.
- Jaundice.
- Features of encephalopathy.
- Recurrent infection.
- Coagulopathy.
- Hepatorenal syndrome.

Investigations

- Liver function tests.
- Ultrasound.
- Alfa-feto protein.
- *MELD scoring is Model for End-stage Liver Disease,* is assessed by specific equation using creatinine and bilirubin.
- *Child-Pugh score system.*
- Serological assessment—AFP, iron studies, α_1-antitrypsin, ceruloplasmin, autoantibodies, investigations for hepatitis.
- Doppler imaging to see vascular pattern, direction of blood flow, size of the vein.

Causes of portal hypertension

Prehepatic	*Hepatic (80%)*	*Posthepatic*
Portal vein or splenic vein thrombosis	Portal pyaemia	Budd-Chiari syndrome
Hypercoagulable status	Alcoholic cirrhosis	Constrictive pericarditis
Periportal inflammation	Idiopathic portal hypertension	Congestive cardiac failure
Trauma	Primary biliary cirrhosis	Veno-occlusive disease
Extrinsic compression from pancreas, stomach	Schistosomiasis	
Neonatal umbilicus sepsis	Hepatitis	
	Nodular regenerative hyperplasia	
	Wilson’s disease	
	Haemochromatosis	
	Congenital hepatic fibrosis	

Note: Commonest prehepatic cause of portal hypertension is portal vein thrombosis.

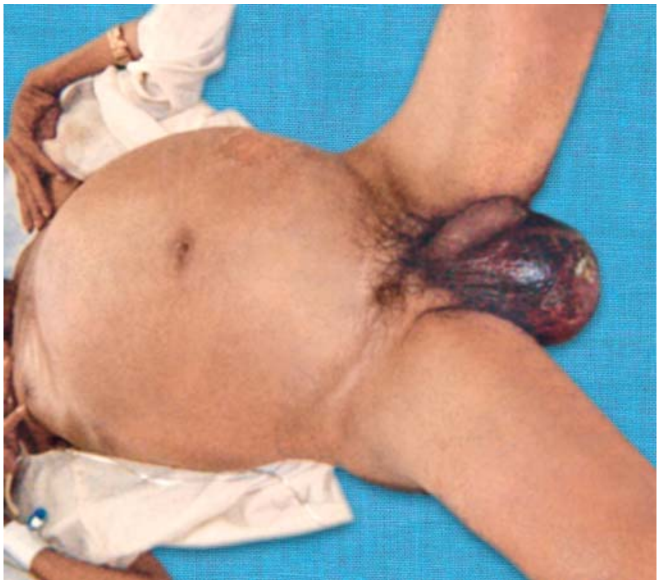

Fig. 11.32: Patient with portal hypertension showing massive ascites with oedema of limb and scrotum—end-stage liver disease.

- CT/MRI to look for cause, see collaterals and liver nodules. *Magnetic resonance venogram (MRV)* is very sensitive noninvasive method to study extrahepatic portal system mainly to confirm extrahepatic portal vein thrombosis.
- Hepatic venous pressure gradient (HVPG): It is done using a balloon catheter. Difference between free hepatic vein pressure and wedged hepatic vein pressure is HVPG. Normal HVPG is 8 mm Hg. It is more than 10 mm Hg in cirrhosis.
- *Splenoportography:*
  - Bleeding/clotting time and prothrombin time should be normal prior to the procedure.
  - Antibiotic prophylaxis is a must.
  - A fine lumbar puncture needle is passed into the spleen through left 9th intercostal space in midaxillary line. It is placed into the pulp of the spleen.
  - Splenic pressure is assessed by connecting to a manometer.
  - Water soluble iodine dye is injected under C arm guidance. Flow, anatomy of portal system, size of the splenic/portal vein, site of the obstruction, communications are assessed.
  - Complications are—bleeding, infection, left shoulder pain.
  - Its use is replaced by MR angiogram/celiac angiogram.
- GI scopy—to identify varices.
- Liver biopsy.
- Endosonography.

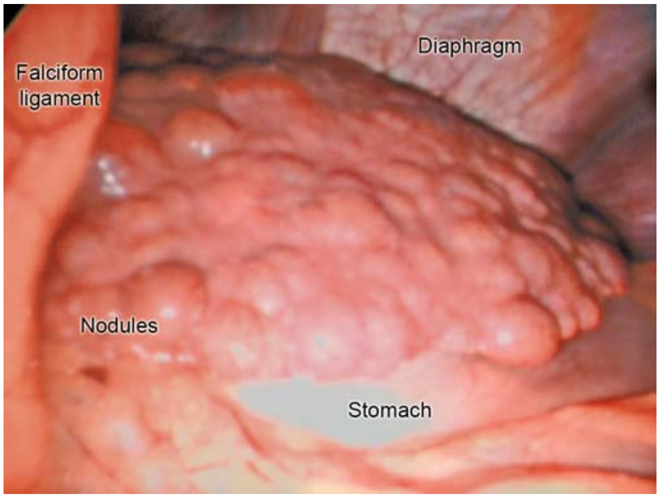

Fig. 11.33: Laparoscopic view of cirrhosis with nodules on the surface.

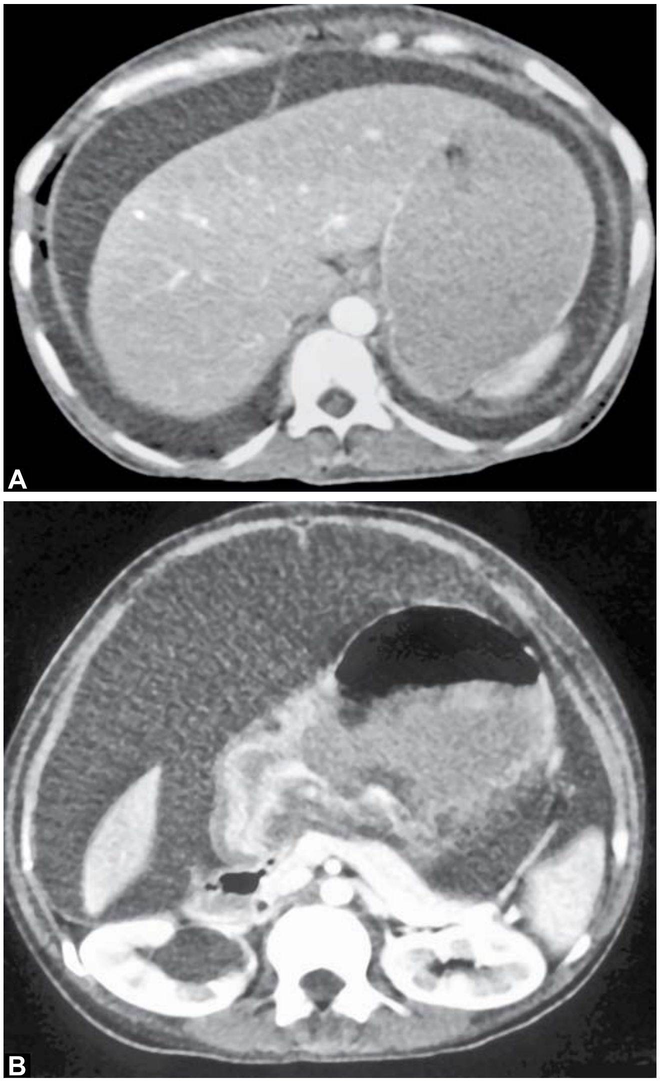

Figs 11.34A and B: CT scan pictures showing ascites, splenomegaly.

Management of Portal Hypertension

a. *General measures:*
  - Anaemia to be corrected.
  - Nutrition supplementation.
  - Inj. vitamin K—10 mg IM for 5 days.

b. *Specific measures:*
  - *Treatment of oesophageal varices.*
  - *Prevention of hepatic encephalopathy.*
  - *Treatment of ascites.*

- *Measures to reduce portal pressure:*
  - Surgeries—Portosystemic shunt.
    - Nonselective.
    - Selective.
  - TIPSS.
  - Drugs to reduce the portal pressure like propranolol, nadolol, isosorbide-5-mononitrate.
- *Liver transplantation*.

OESOPHAGEAL VARICES

- May be asymptomatic.
- May present with *haematemesis* or *melaena* or as *recurrent bleeding*. Varices begin to bleed when portal pressure exceeds 12 mmHg.
- When present with severe haematemesis, patient shows features of *shock*.
- Mortality in bleeding varices is 25-30%.
- Factors related to variceal bleed are—portal venous pressure; gastro-oesophageal reflux causing ulceration; variceal size; variceal wall tension.

Types of Varices

1. *Oesophageal* which is in the lower 1/3rd of the oesophagus, usually 3 or more in number, graded as I, II, III, IV (based on gastroscopic findings) (80%).
2. *Gastric* which is fundal or in upper part of the stomach (20%).

Gastric varices

Classification

A. Extension of oesophageal varices
  - I. Into lesser curve
  - II. Into fundus

B. Isolated gastric varices
  - I. In fundus common
  - II. In other parts—stomach, duodenum—rare

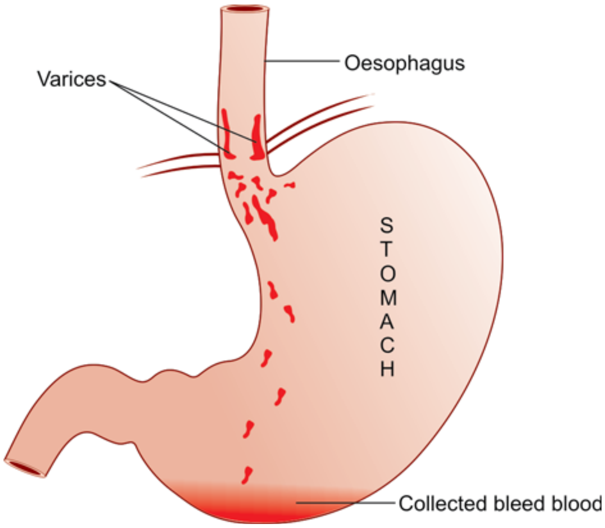

Fig. 11.35: Picture showing variceal bleed.

Gastric varices are *treated by endoscopic N butyl cyanacrylate glueing, sclerotherapy.* Banding is usually not done.

Differential diagnosis for oesophageal varices (Other causes for haematemesis)

Bleeding peptic ulcer	Carcinoma stomach
Mallory-Weiss syndrome	Acute erosive gastritis
Bleeding disorders	Dieulafoy's syndrome
Gastric polyps	Gastric leiomyoma
Gastric lymphoma	

Investigations

- Hb%, liver function tests, packed cell volume, blood grouping and crossmatching, prothrombin time, platelet count, blood urea, serum creatinine.
- Endoscopy—*Grading of varices:*

I. Minimal varices without luminal prolapse
II. Moderate varices, with luminal prolapse and with minimal obscuring of O-G junction
III. Large varices, with luminal prolapse and with moderate obscuring of O-G junction
IV. Very large varices, with luminal prolapse and with complete obscuring of O-G junction

"Cherry red spots" may be sign of impending rupture. *Endoscopic sclerotherapy or variceal band ligation* should be tried in acute bleed also, while doing initial endoscopy to confirm the variceal bleed. If facility for sclerotherapy is not available or sclerotherapy/banding fails or not possible due to massive bleed, then balloon tamponade should be used.

Occasionally patient may be having varices, but bleeding (upper GI bleed) is due to other causes like duodenal or gastric ulcer.

- Once acute bleed stops patient is investigated further by splenoportography, liver biopsy, and superior mesenteric or celiac angiography, MR angiogram.

Note:
Cruvielhier-Baumgarten syndrome is loud venous hum heard at umbilicus in case of portal hypertension through congenital patent umbilical vein to portal vein.

- U/S abdomen to see status of the liver and portal vein (size *more than 13 mm* is significant).

Treatment of Oesophageal Varices

- Treatment of varices prior to bleed (no history of earlier variceal bleed)—*primary prophylaxis.*
  - Drugs—propranolol (reduces the portal pressure by 20% with pulse rate below 55/minute or 25% of resting pulse); nadolol (long acting, less lipophilic, not metabolised by liver); isosorbide mononitrate. Drugs reduce bleeding by 40%.
  - Endotherapy.

***Note*:**
No prophylactic shunt surgeries.

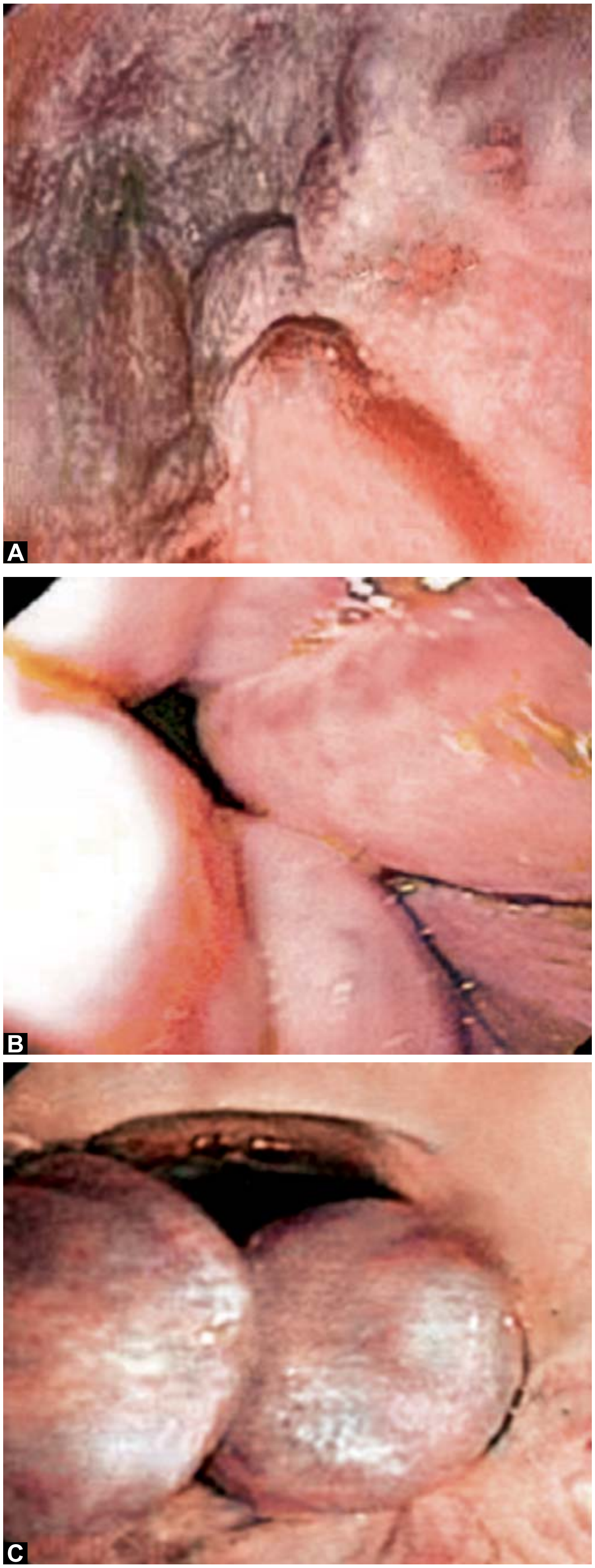

Figs 11.36A to C: Different grades of oesophageal varices—endoscopic views.

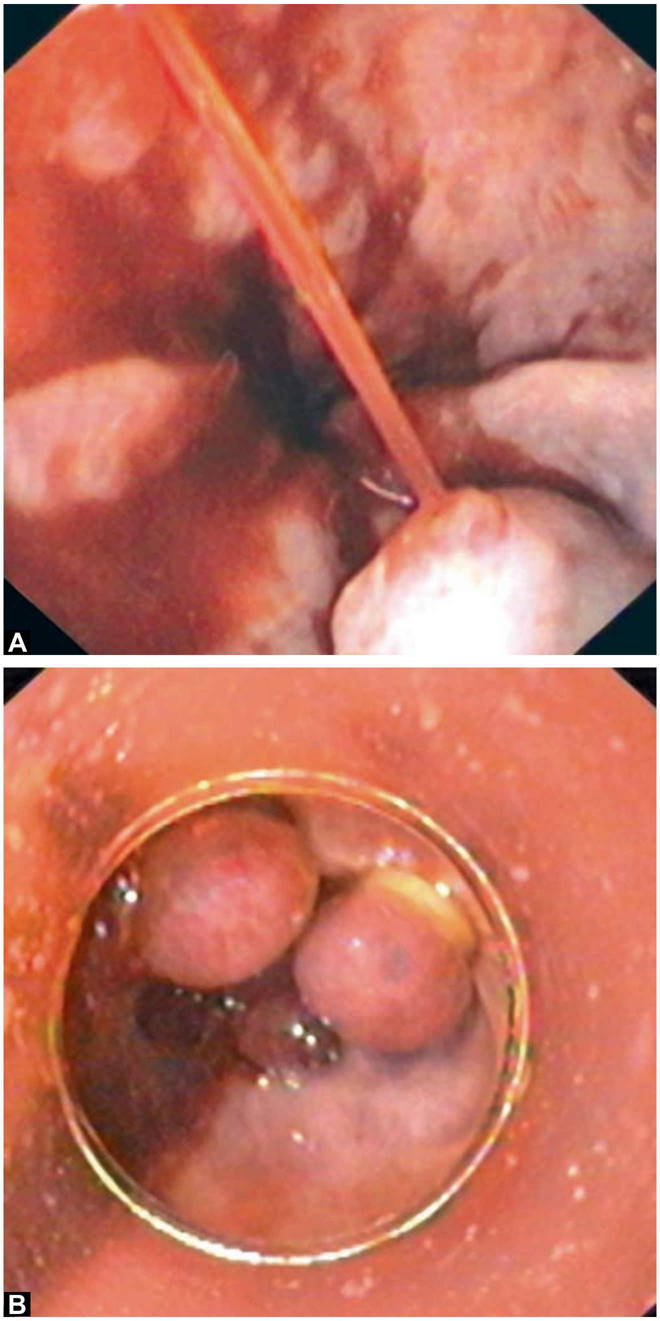

Figs 11.37A and B: Active spurting variceal vessel. Endoscopic variceal banding is gold standard for oesophageal varices. It is not very useful for gastric varix.

- Treatment of varices after one or more episodes of bleeding—*secondary prophylaxis*.
  - Endotherapy.
  - Shunt surgeries.
  - Drugs like propranolol. Drugs may prevent rebleed by 40%.

Treatment of varices is considered under two headings:

- Emergency management of bleeding varices.
- Definitive management of varices.

SAAG means Serum, Ascites, Albumin, Gradient.

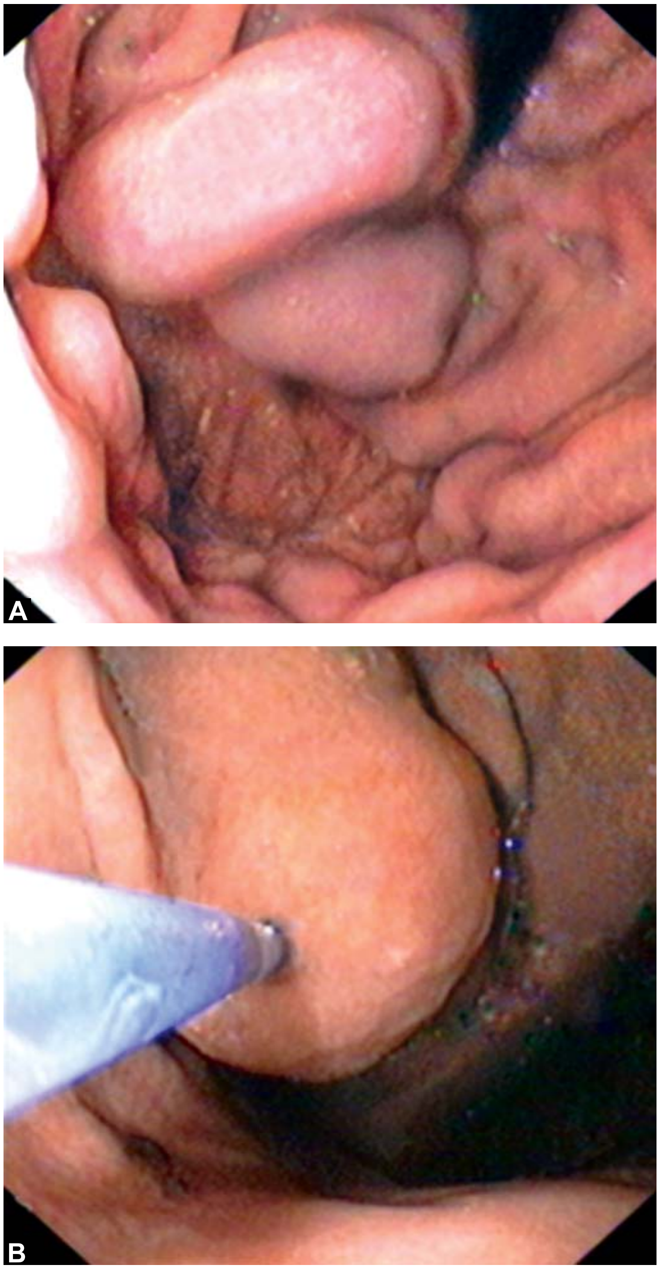

Figs 11.38A and B: Typical tortuous gastric varix and cyanoacrylate glue injection to the same.

EMERGENCY MANAGEMENT IN SEVERE HAEMORRHAGE

Control of acute bleed

- Initial important measures
  - Volume and electrolyte corrections; volume replacement using crystalloids with central line (CVP line placement with CVP maintained at 5 cm of water, sudden overload should be avoided which may precipitate pulmonary oedema/ascites/hyponatraemia)
  - Blood (massive transfusion of 5,10,15 bottles) and blood products (FFP, platelets) replacement endotracheal intubation, ventilator, critical care/ ICU management
  - Antibiotics, nutrition (TPN)/Vitamin K injection
  - Catheterisation and hourly urine output monitoring
  - Prevention of encephalopathy
  - Rest/sedation only if needed with care
- Pharmacotherapy
- Endoscopic banding/sclerotherapy
- Balloon tamponade
- TIPSS
- Devascularisation surgical procedures

Pharmacotherapy (Drugs)

- *Inj. Vasopressin 0.4* units/min for 24 hours is commonly used. 20 units in 100 ml 5% dextrose IV in 20 minutes also can be given. Vasopressin infusion is combined with glyceryl trinitrate tablets sublingually (to reduce cardiac side effects and increase the efficacy). Vasopressin is contraindicated in ischemic heart disease as it causes coronary vasoconstriction. Nitroglycerin prevents the coronary vasoconstriction and also reduces the portal pressure. Vasopressin can cause fluid retention and hyponatraemia.
  - Vasopressin *constricts* the splanchnic vessels thus *lowers the portal pressure* controlling the variceal bleed. It controls acute bleed in 80% of cases; it is often used along with balloon tamponade to increase bleeding control to 95%. Complications are—MI, arrhythmias, intestinal ischaemia, hyponatraemia and reduction in hepatic, renal blood flow.
- *Glypressin* (terlipressin), *desmopressin*. Glypressin 2 mg IV bolus 5th hourly is better than vasopressin. It has got longer half life with fewer side effects.
- *Somatostatin*—half life 2 minutes. Somatostatin does not cause systemic vasoconstriction like vasopressin but reduces the splanchnic and hepatic blood flow. It is given as continuous IV infusion 250 µg/hour after an initial bolus of 250 µg.
- *Octreotide*—half life 2 hours. 50 µg IV bolus and 50 µg/hour for 5 days.
- *Propranolol* decreases the portal pressure. It is used at later period as prophylaxis to prevent bleeding. Nadolol is also used which has got a longer half life.
- Metoclopramide (20 mg IV, it causes lower oesophageal sphincter constriction); isosorbide 5 mononitrate which is not metabolized in liver can be safely used in patients in liver failure.
- Acid inhibiting drugs (ranitidine IV, pantoprazole IV); and antifibrinolysins (tranexamic acid) are also used.

Endoscopic Therapy

Endotherapy—banding/sclerotherapy/glueing is used to control acute bleed as well as to prevent further bleeding. Pharmacotherapy is often combined with endotherapy.

1. *Endoscopic variceal banding (EVB)*
    - It is done for oesophageal varices. Banding/band ligation has become *gold standard and ideal* for oesophageal varices.
    - Multishoot banding device is used.
    - It is not commonly done for gastric varix as banding is difficult in retroflexed position.
    - Banded varices thrombose and slough off leaving small discrete ulcers.
    - It controls the bleeding in 90% cases. It is technically easier; chances of rebleeding are less; complications like perforation or stricture are less.
    - Earlier single band application device was used. It should be passed through an oesophageal overtube as it requires repeated endoscopy for shooting many bands. Multishoot banding device/applicator carries 5/6/8/10 bands. It does not require repeated removal and insertion of endoscope and so overtube can be avoided.

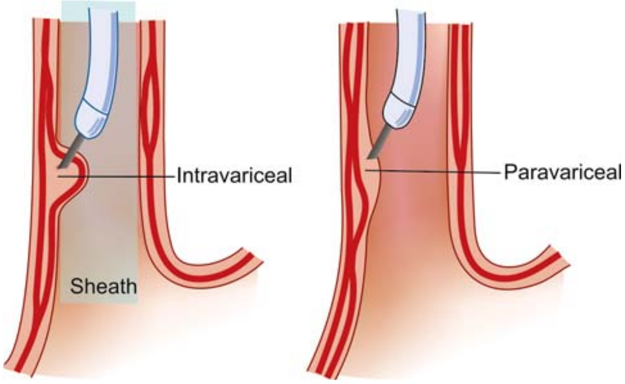

Fig. 11.39: Types of sclerotherapy.

2. *Endoscopic variceal sclerotherapy (EVS)*
    - Endoscopic sclerotherapy is also commonly used. It is technically easier, and cheaper.
    - Both intravariceal and paravariceal or only intravariceal injections are given.
    - In large varix lower part and upper part of the varix is sclerosed. Around 6 ml of sclerosant to lower part and 2-3 ml to upper part is injected. In small varix, injection is sufficient to only lower part.
    - In combined para and intravariceal injection, first paravariceal injection is done into submucosal plane followed by intravariceal injection.
    - Sclerosants used are—*Fatty acid derivatives*: ethanolamine oleate (5%); sodium morrhuate (2-5%); *Synthetic:* sodium tetradecyl sulphate; polidocanol (3%). Polidocanal is hydroxyl polyethoxydodecan (HPD/aethoxy skletrol 1-3%)
    - Varices should not be injected blindly.
    - In acute bleed it is better to insert an endotracheal tube to the patient and give 45° elevation to the head end of the table.
    - Sclerotherapy causes intimal injury, submucosal vessel thrombosis, ulceration, submucosal fibrosis. 5-6 sclerotherapies are needed to create total obliteration. If after sclerotherapy aspiration fluid is clear, then it is called as successful sclerotherapy. Success rate is 80%. Sclerotherapy during bleeding is called as acute sclerotherapy; at bleed free intervals it is called as chronic.

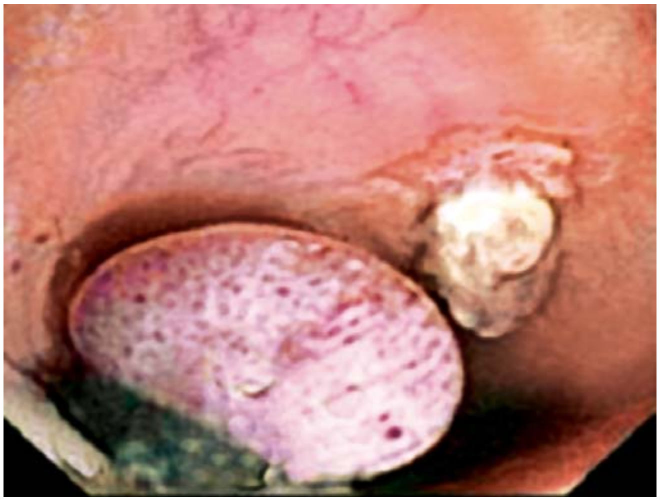

Fig. 11.40: Endoscopioc view of variceal banding. Endoscopic variceal banding (EVB) is done for oesophageal varices. Not commonly used for gastric varices.

    - *Complications*
        - Fever, retrosternal pain, massive bleed, perforation, mediastinitis, stricture oesophagus.
        - Atelectasis, pleural effusion, ARDS are known to occur.
    - Sclerotherapy is done weekly for 6-8 sessions until varices are obliterated. Later follow up endoscopy is done once in 3 months for 6 months; and once in 6 months for 2 years; then yearly for life time.
    - Sclerotherapy causes 70- 80% control of bleeding but rate of re-bleed is higher than banding.
3. *Endoscopic gluing using tissue adhesives*
    - It is commonly used for gastric varices.
    - Butyl cyanoacrylate is used. It hardens in 20 minutes. If properly injected into the varix it controls the bleeding immediately with 90% success rate.
    - Risk of equipment damage is high if it spills into the working channel and blocks it. Lipiodol is used along with cyanoacrylate to delay premature hardening. Silicone oil is applied to the tip of the instrument to prevent the damage.
    - Needle tip is correctly placed into the gastric varix and 0.5-1 ml of adhesive is injected. Instrument should be removed immediately.
4. *Endoscopic thrombin/dilute adrenaline injection into the varices.*

Balloon Tamponade

- *Four lumen Minnesota tube/Sengstaken-Blakemore tube/ Linton's tube/modified 4 lumen Pitcher tube* are used for the procedure. If bleeding does not stop by banding/sclerotherapy/gluing, balloon tamponade should be tried. Most of the time bleeding comes under control by balloon tamponade.
- Initially gastric balloon is inflated with 300-400 ml of *air*, later the oesophageal balloon is inflated with air at 30-40 mmHg pressure. This should be deflated for a short period in 12-24 hours so as to prevent necrosis of oesophageal mucosa. Inflation of gastric balloon is most important. Oesophageal balloon inflation is done if needed only.

During infancy until about the end of the third year the normal liver extends one to two finger breadth below the costal margin. During inspiration at this period of life, the extreme edge of normal spleen is also palpable.

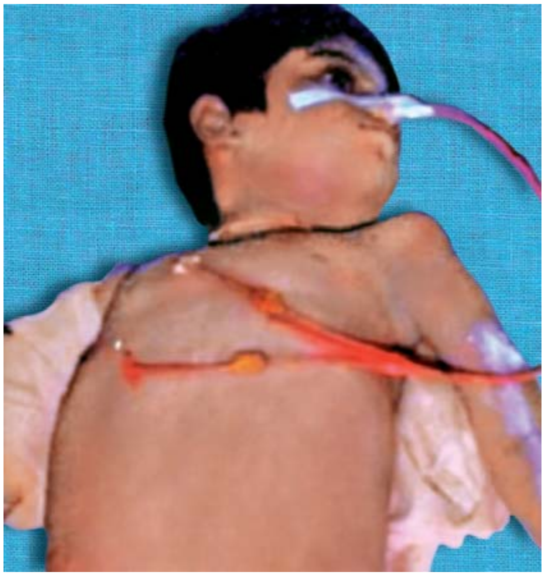

Fig. 11.41: Child with Sengstaken-Blakemore (SB) tube for haematemesis showing three channels.

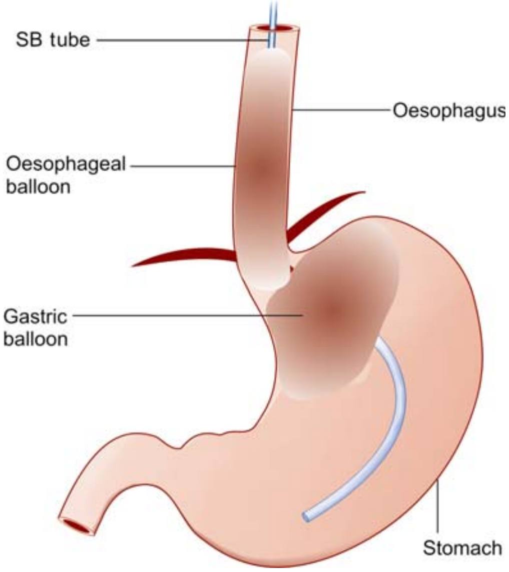

Fig. 11.42: Sengstaken-Blakemore tube showing how it controls the bleeding varices. Oesophageal balloon is inflated with 40 mmHg pressure of air, gastric balloon with 400 ml of air.

- X-ray should be taken to confirm the position of balloons.
- Usually within 24-72 hours bleeding comes under control and in mean time resuscitative measures has to be undertaken. Success rate is very good but only temporary. It should be followed by definitive procedures like banding/sclerotherapy. Otherwise rebleed rate is 60%. Balloon tamponade is only a rescue procedure and emergency bridging method for more definitive procedures.

TIPSS (Transjugular Intrahepatic Portosystemic Shunt/Stenting)

It is used if all other earlier methods mentioned have failed, in less than 10% of acute bleed patients. It commonly controls the uncontrolled acute bleeding and prevents further bleed and also acts as *a bridge* for *future transplantation*.

Surgeries for Acute Bleeding Varices

- *Open oesophageal stapler transection and oesophagogastric anastomosis (*Jhonston).
- *Borema—Crile operation*: The oesophagus is opened. Oesophageal varices are under-run using vicryl.
- *Sugiura-Futagawa* operation: Oesophagogastric transection and devascularization, vagotomy, pyloroplasty and splenectomy.
- *Milnes-Walker* thoracic oesophageal transection
- *Tanner's* abdominal oesophagogastric resection.
- *Hassab* operation.

Shunt surgeries are not commonly used to control bleeding in acute stage.

***Note*:**

Once a bleed, there will be always a rebleed without treatment and rebleed is much more severe than earlier bleed. Time gap between each rebleed shortens.

Other Measures

- *Liver biopsy* to confirm the cause for the portal hypertension (after controlling PT).
- *Correction of clotting mechanisms* by Vit. K, blood, plasma or platelet concentrate.

Control of encephalopathy
- Purgatives
- Protein free diet
- Lactulose syrup 30 ml tid
- Lactitol
- Neomycin orally—1.5 gm 6th hourly
- Parenteral nutrition

Sclerotherapy

Sclerosants used	***Method of injection***	***Problems with sclerotherapy***
Ethanolamine oleate (5%)	Both intravariceally (6 ml) as well as para variceally (2-3 ml) to create thrombosis as well as fibrosis	Oesophageal ulceration
Sodium tetradecyl sulphate (1-3%)		Oesophageal perforation
Sodium morrhuate 2-5%		Worsening of bleeding
Polidocanol 3%	Repeated sessions in 2-3 weeks	Aspiration pneumonitis
Absolute alcohol	Total 15 ml in one session	Difficult to inject gastric varices
		Chest pain and dysphagia are common

DEFINITIVE MANAGEMENT OF VARICES

- ♦ ***Sclerotherapy***
- ♦ ***Shunt surgery***

  Prophylactic shunt surgery should not be done.

  Shunt surgeries should not be done if the varices have not bled before.

Indications

Child's Grades A and B.

(Child's grading is used for selecting patients for surgery and predicting prognosis.)

Surgery is contraindicated in Child C.

1. **Non-selective shunts**:
   - *Porta-caval shunt* (end-side, side-to-side).
   - Mesentericocaval shunt with or without graft.
   - Mesenterico-renal shunt.
   - *Proximal splenorenal shunt—Linton's shunt.*

Non-selective shunts control and prevent further bleeding as well as reduce the ascites. But since it diverts the blood from liver *there is high incidence of encephalopathy.*

Note:

End to side porta caval shunt is called as ECK fistula.

2. **Selective shunts:**
   a. *Distal splenorenal shunt—Warren's shunt*. End of splenic vein is shunted to the side of the left renal vein. Splenic vein should be more than 10 mm in size (diameter). It selectively decompresses the portal bed near oesophago-gastric area (so that it prevents bleeding). Liver is also perfused adequately and so possibility of encephalopathy is less.
   b. *Inokuchi shunt* is a selective shunt between IVC and left gastric vein (using a graft).

- ♦ ***Splenectomy***

Left sided portal hypertension also called as *segmental portal hypertension* (sinistral portal hypertension) is due to splenic vein thrombosis (If partial channels are present within it, it is called as *splenic vein cavernoma*). Here liver is normal and *splenectomy* alone will correct the condition.

- ♦ ***TIPSS (Transjugular Intrahepatic Portosystemic Stenting/ Shunt)***
  - Size of stent must be 10 mm. It is a nonsurgical, interventional radiological method wherein a stent is placed in the liver between hepatic venule and portal venule through a guidewire following dilatation.
  - It is a simple and easier method to reduce portal pressure.
  - It can cause hepatic encephalopathy, bleeding, bile leak and infection.
  - It is contraindicated in encephalopathy, tumours and liver cysts.
  - It is quite useful as a temporary measure to reduce portal bed pressure, before doing orthotopic liver transplantation in cirrhotic patients.

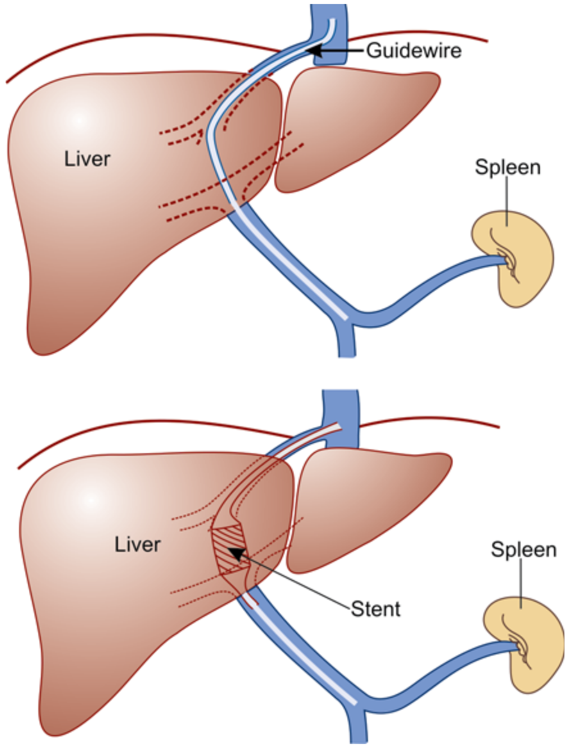

Fig. 11.43: TIPSS.

Child's grading of the severity of liver disease

Clinical and biochemical measurement	***Points scored***		
	1	2	3
Bilirubin (mg/dl)	< 2.0	2.0-3.0	> 3.0
Albumin (gm/dl)	> 3.5	3.0-3.5	< 3.0
Ascites	None	Controlled	Uncontrolled
Mental status	Normal	Disoriented	Coma
Nutrition	Very good	Good	Poor
Total Score	**< 6** *Child's 'A'* disease	**7-9** *Child's 'B'* disease	**10-15** *Child's 'C'* disease
(Pugh's modification)			
Prothrombin time seconds prolonged	Increase upto 3	Between 3-6	> 6

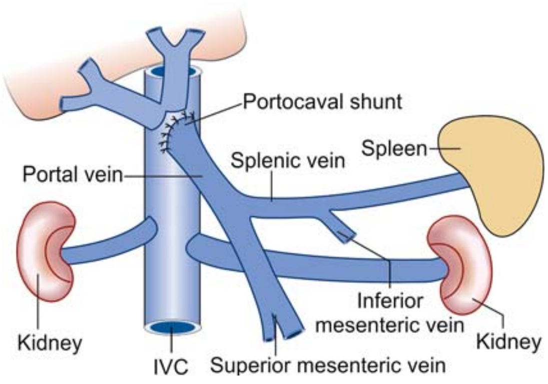

Fig. 11.44: End-to-side portocaval shunt (ECK fistula).

Efforts and energies should be consistent, constructive and compassionate.

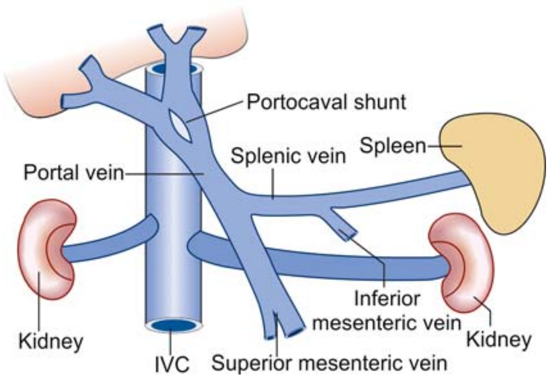
Portocaval shunt
Portal vein
Splenic vein
Spleen
Inferior mesenteric vein
Kidney
Kidney
IVC
Superior mesenteric vein

Fig. 11.45: Side-to-side portocaval shunt.

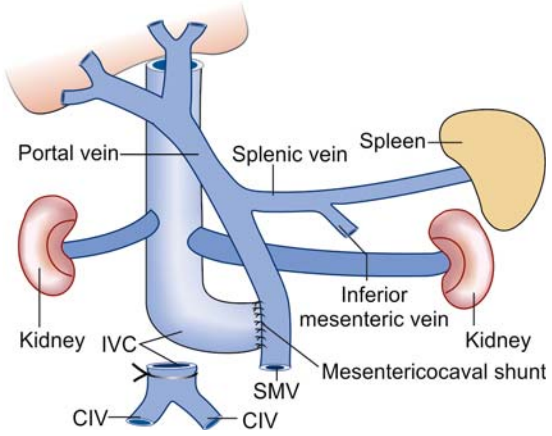
Portal vein
Splenic vein
Spleen
Inferior mesenteric vein
Kidney
Kidney
IVC
Mesentericocaval shunt
SMV
CIV
CIV

Fig. 11.46: Side-to-side mesenteriocaval shunt.

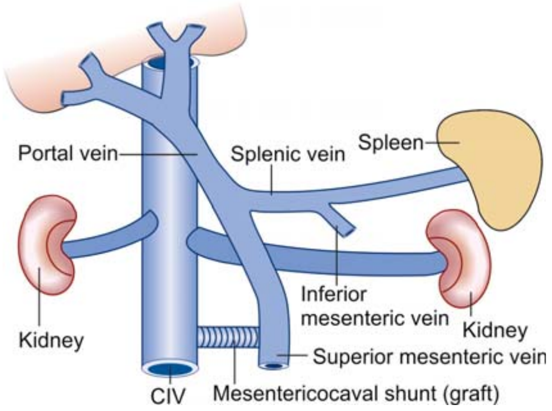
Portal vein
Splenic vein
Spleen
Inferior mesenteric vein
Kidney
Kidney
Superior mesenteric vein
CIV
Mesentericocaval shunt (graft)

Fig. 11.47: Side-to-side mesenteriocaval shunt using a graft.

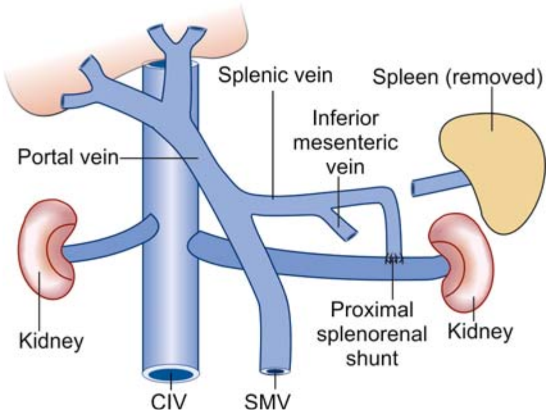
Splenic vein
Spleen (removed)
Inferior mesenteric vein
Portal vein
Kidney
Proximal splenorenal shunt
Kidney
CIV
SMV

Fig. 11.48: Proximal splenorenal shunt (Linton's)

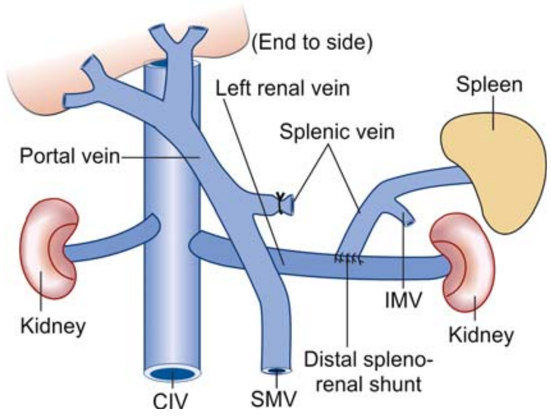
(End to side)
Left renal vein
Spleen
Splenic vein
Portal vein
Kidney
IMV
Kidney
Distal spleno-renal shunt
CIV
SMV

Fig. 11.49: Distal splenorenal shunt (Warren's, DSRS)—selective.

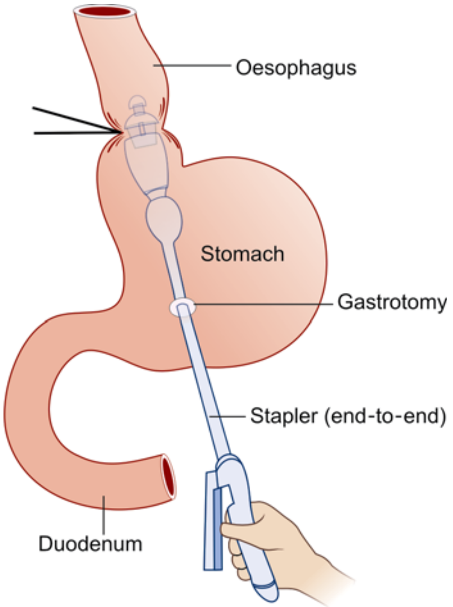
Oesophagus
Stomach
Gastrotomy
Stapler (end-to-end)
Duodenum

Fig. 11.50: Oesophageal open stapler transection.

Problems with TIPSS

- Post-stent encephalopathy—40%
- Stenosis (50%) and recurrent varices
- Bleeding
- Bile leak
- Sepsis

Surgeries for portal hypertension

- Devascularisation surgeries
- Decompression surgeries
- Orthotopic liver transplant (OLT)

Surgeries for oesophageal varices and portal hypertension

Nonselective shunts

- End-to-side portacaval shunt (ECK fistula)
- Side-to-side portacaval shunt
- Mesentericocaval shunt

- Mesentericocaval shunt with graft
- Proximal splenorenal shunt—Linton's shunt
- TIPSS

Selective shunts

- Distal splenorenal shunt—Warren's shunt
- Inokuchi shunt between left gastric vein and IVC through a graft.

Porta azygos disconnection

- Oesophageal stapler transection of Johnston
- Boerema-Crile operation—thoracic approach
- Milnes-Walker thoracic oesophageal transection
- Tanner's abdominal oesophagogastric transection
- Hassab operation. Devascularisation and splenectomy.
- Sugiura-Futagawa operation—vagotomy, pyloroplasty, devascularisation, splenectomy
- Splenectomy

Liver transplant

◆ ***Liver transplant (orthotopic)***

Liver transplant is becoming popular for cirrhosis with varices. It is ideal, final and best. But donor availability and cost is the problem. If patient is for orthotopic liver transplantation (OLT), open shunt surgeries should not be done as liver hilum should be kept virgin for effective transplantation. TIPSS can be done as a bridge in such patients.

PORTAL HYPERTENSIVE GASTROPATHY

- Portal hypertension causes vascular dilatation and ectasia in the stomach which appears pink, speckled, with red mosaic pattern in the gastric mucosa. It occasionally causes upper GI bleed (5% of total; 20% of portal hypertension bleed). Endoscopy is diagnostic.
- Mucosal changes are diffuse and so not amenable for endotherapy.
- Portal gastropathy is more common in cirrhotic patients.
- Commonly presents with chronic bleed causing anaemia; but acute bleed with haematemesis and melaena also can occur.
- Treatment is mainly to reduce the portal pressure using propranolol.

ASCITES

It is pathological collection of fluid in the peritoneal cavity.

In Western countries cirrhosis is the commonest cause (80%) of ascites. Ascites is the most common complication of cirrhosis. It is a poor prognostic factor. Portal hypertension, renin angiotensin aldosterone pathway causing renal sodium retention, increased hydrostatic pressure in hepatic sinusoids and splanchnic vessels cause ascites. In metastatic malignancy ascites develops due to liver secondaries, peritoneal carcinomatosis which secretes protein rich fluid and lymph. Leakage of pancreatic juice, bile, lymph also causes specific ascites.

Types

◆ Mild	— Up to 150 ml amount required to demonstrate radiologically
◆ Moderate	— 1500-2000 ml causes clinical dullness in flanks.
◆ Severe	— > 2000 ml.

Classification

- *Transudate* (Protein < 2.5 gm/ dl).
  - CCF—Commonest (SAAG > 1.1).
  - Hypoproteinaemia.
  - Anaemia.
  - Beri-beri.
  - Nephrotic syndrome.
  - Portal hypertension.
  - Polyserositis.
- *Exudate* (Protein > 2.5 gm/dl) (SAAG < 1.1).
  - Peritoneal diseases.
    - Tuberculosis.
    - Bacterial, fungal and parasitic infection.
    - Neoplasm.
  - Collagen disorder.
  - Eosinophilic gastroenteritis.
  - Granulomatous peritonitis.

Management of oesophageal varices

Emergency management in acute bleed	*Definitive management*
1. Balloon tamponade	1. Sclerotherapy, Banding, Glueing
2. Pharmacotherapy	2. Nonselective shunts: Porta caval
3. Oesophagogastric transection	Mesentericocaval
4. Borema-Crile operation	Mesentericorenal
5. Sugiura-Futagawa operation	3. Selective shunts: Distal-splenorenal shunt
6. Sclerotherapy	Inokuchi shunt
	4. Splenectomy
	5. TIPSS
	6. Liver transplant

Clinical Features

Symptoms

- Gradual distention of abdomen
- Severe cases of ascites may lead to respiratory embarrassment.
- Specific features related to the cause may be evident.
- Risk of spontaneous bacterial peritonitis.

Signs

Mild:
- *Puddle sign*—Tapping around the umbilicus in knee-elbow position elicits dullness.

Moderate:
- Positive shifting dullness in the abdomen.

Severe:
- Positive fluid thrill.
- *Tanyol sign*: Umbilicus is shifted upward in pelvic mass (ovarian cyst) downwards in ascites.
- Hamilton ruler test is also used to differentiate ascites from ovarian cyst.
- *Blaxland (Athelstan Blaxland) ruler test* shows pulsation in ovarian cyst not in ascites. A flat ruler is placed on the abdomen just above the anterior superior iliac spines and pressed firmly backwards. If the swelling is due to ovarian cyst, transmitted aortic pulsation can be felt across the ruler.
- *Smiling* horizontal *umbilicus*.

Investigations

- U/S abdomen.
- *Ascitic tap*—Always should be done after emptying the bladder by placing the needle below the umbilicus lateral to the rectus muscle. It may be *Diagnostic* or *Therapeutic tap.* Fluid is sent for analysis.
  - Ascitic fluid having cells less than 1000/mm^3 is clear; from 1000-5000/mm^3 is unclear; more than 5000/mm^3 is turbid/cloudy.
  - Non-traumatic blood in the ascitic fluid does not clot. Traumatic (trauma during tapping) blood will clot.
  - Lipid in the ascitic fluid is seen in chylous ascites.
  - In cirrhotic patients ascitic fluid without SBP contains less than 500 cells/mm^3 with 50% of them are neutrophils. If ascitic fluid contains more than 250 neutrophils/mm^3 it suggests acute inflammation.
  - ***Serum ascites albumin gradient (SAAG)*** is calculated by subtracting ascitic fluid albumin level from serum albumin level. If SAAG more than 1.1, it is called as high gradient which suggests portal hypertension. If SAAG is less than 1.1, it is called as low gradient which suggests absence of portal hypertension. SAAG is 98% accurate.
    - *High gradient SAAG* is seen in—cirrhosis, alcoholic, cardiac ascites, multiple liver secondaries, fulminant liver failure, Budd-Chiari syndrome, portal vein thrombosis.
    - *Low gradient SAAG* is seen in—peritoneal secondaries, peritoneal tuberculosis, pancreatic ascites, biliary ascites, nephrotic syndrome, lymph leak, connective tissue diseases causing ascites.

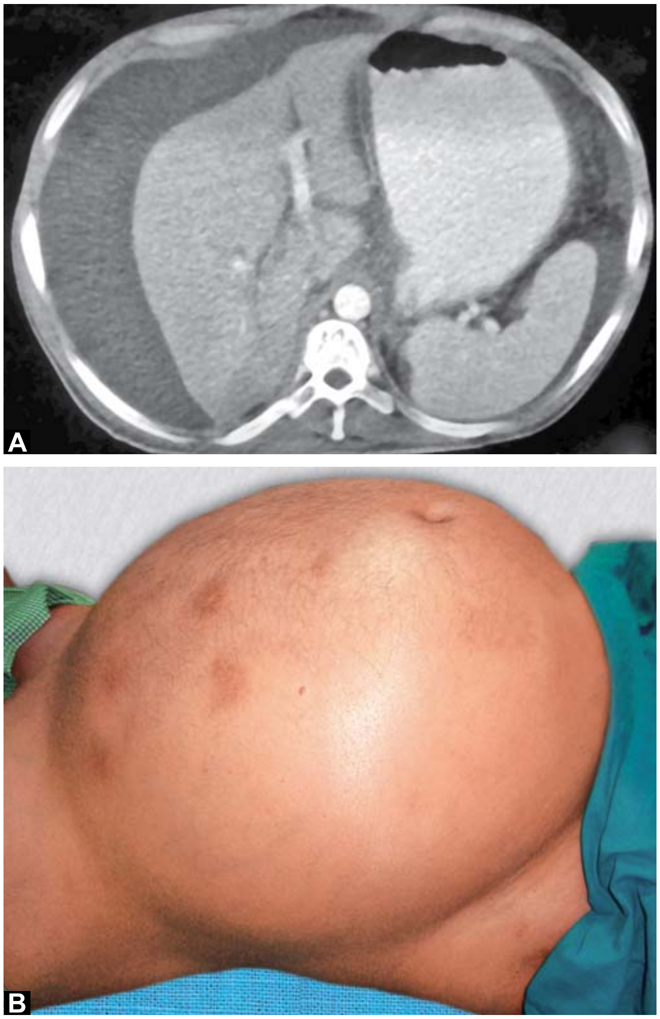

Figs 11.51A and B: Ascites on CT scan and gross clinical look of severe ascites.

Treatment

- The cause is treated.
- Therapeutic tap—It should be slow and gradual or staged tapping. Up to 5 litres can be tapped in 90 minutes.
- Ascitic shunt surgeries.
- Spironolactone, salt restriction.
- Salt (sodium) restriction in diet is important. It is assessed by measuring 24 hour urinary sodium excretion. If it is more than 78 mmol/day patient will loose weight with salt restricted dite.
- Spironolactone 100 mg/day and frusemide 40 mg/day promotes natriuresis and reduces the ascites. It is very useful after paracentesis to prevent re-accumulation of fluid.
- TIPSS.
- Liver transplantation.

Complications of ascites

- Respiratory embarrassment
- Umbilical hernia with erosion, ulceration and leak
- Spontaneous bacterial peritonitis

ASCITES IN PORTAL HYPERTENSION

Causes

- Hypoproteinaemia.
- Increased hydrostatic pressure.
- Decreased colloidal osmotic pressure.
- Lymphatic blockage.
- Altered aldosterone mechanism.

Treatment of ascites in portal hypertension

- Medical—spironolactone, salt restriction
- Abdominal paracentesis
- TIPS as a bridge to liver transplant
- Liver transplantation

Abdominal fluid

- *Straw coloured fluid*
  - Cirrhosis
  - Tuberculosis
  - Hypoproteinaemia
  - Nephrotic syndrome
  - Budd-Chiari syndrome (Hepatic vein thrombosis)
  - Constrictive pericarditis
  - Chronic pancreatitis
  - Ovarian tumour (Meig's syndrome)
  - Malignancy
- *Haemorrhagic fluid*
  - Traumatic tap
  - Tumour
  - Acute haemorrhagic pancreatitis
  - Bleeding disorder
- *Chylous fluid*
  - Parasitic infestation—filariasis
  - Tuberculosis
  - Malignancy of thoracic duct
  - Trauma to thoracic duct
  - Thrombosis of subclavian vein
- *Purulent fluid*
  - Abdominal infections
  - Penetrating wounds or infections
  - Pyaemia and septicaemia
  - Rupture/perforation of an organ

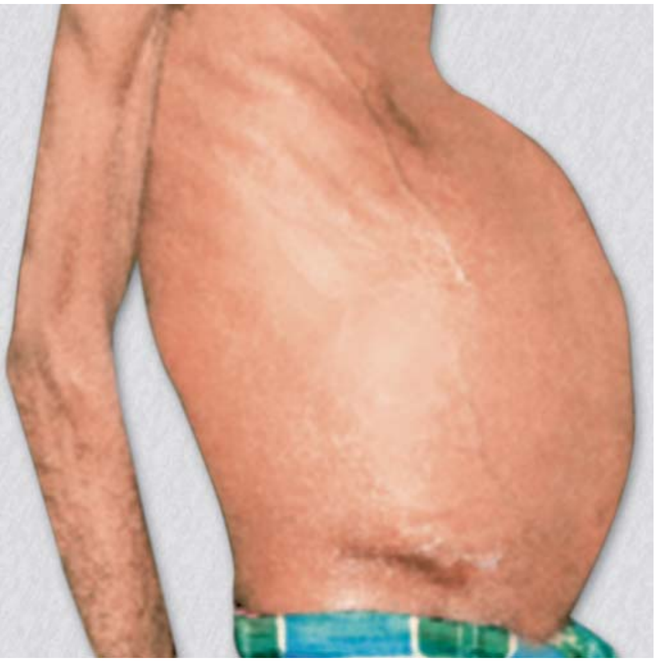

Fig. 11.52: Ascites. Observe the dilated abdominal veins.

Investigations

Liver function tests, U/S, CT scan, MRI, gastroscopy.

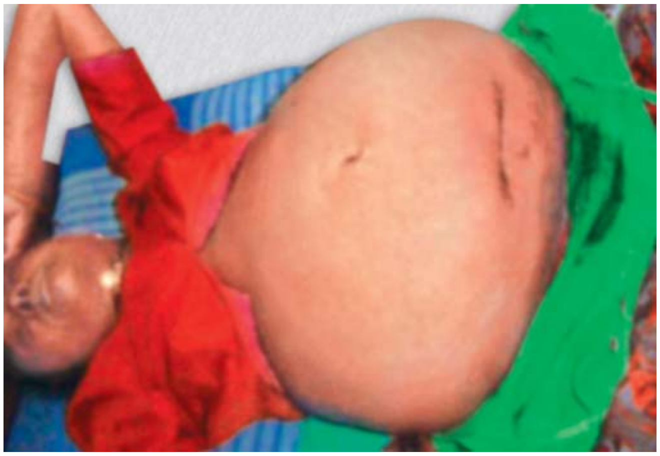

Fig. 11.53: Massive ascites in severe portal hypertension.

Treatment

- Correction of serum proteins, salt restriction.
- *Ascitic tap*—Site of tap is below the umbilicus, lateral to the rectus muscle. Bladder should be empty before tapping. Slow gradual tapping is important, otherwise patient goes in for fluid and electrolyte imbalance. Up to 5 litres can be tapped at a time.
- *Spironolactone* 25 mg orally BD and other diuretics are often used.

- ***Le Veen shunt*** is a one way valve *peritoneo-venous shunt* done in order to drain ascitic fluid from peritoneal cavity to internal jugular vein**.**
- ***Denver* shunt** is similar to Le Veen type with additional facility of having a chamber to evacuate the debris by digital pressure.
- *Liver transplantation* may be beneficial for ascites in liver diseases.
- The cause is treated.
- TIPSS.

Ascitic shunts are not used when there is

- Hepatorenal syndrome
- Bacterial peritonitis

Complications of Ascitic Shunt Procedure

- Bleeding, infection.
- Displacement, blockage.
- Thrombosis of the vein, volume overload.

BUDD-CHIARI'S SYNDROME

It is a syndrome due to obstruction of the hepatic veins.

Usually three major hepatic veins are involved by thrombosis or block. It is common in *females*.

Causes

1. Spontaneous thrombosis of hepatic veins (due to myeloproliferative diseases (50%) commonly seen in adults and IVC webs commonly seen in children).
2. Neoplastic encroachment of hepatic veins.

} Majority of cases

3. Polycythaemia.
4. Contraceptive pills.
5. A membranous web in suprahepatic portion of IVC—a curable type of Budd-Chiari syndrome.
6. Plant extracts of Senecio and Crotalaria used in herbal tea in Jamaica causes acute Budd-Chiari syndrome.
7. Protein C, S or antithrombin III deficiency, Behcet's disease.

Clinical Features

1. *In acute form:* Nausea, vomiting, severe pain in right hypochondrium, rapid enlargement of liver, hypotension and often death.
2. *In less sudden onset type:* Rapid development of ascites, hepatic insufficiency, oedema feet, features of IVC obstruction, hepatic coma can occur.
3. *Chronic type:* Present with hepatic cirrhosis, ascites, oesophageal varices, infection, which gradually leads to hepatic failure.
4. Jaundice, ascites and hepatomegaly is common.

Investigations

- Transcutaneous Doppler will show hepatic vein thrombosis and collaterals across retrohepatic IVC.
- CT shows caudate lobe (segment I) hypertrophy with nonhomogenous contrast enhancement.

Treatment

- Anticoagulation is the standard therapy.
- In membrane obstructing IVC, it can be cut by transatrial meatotomy, i.e. under guidance a catheter is passed, via right atrium to IVC and the web is cut. It is curable.
- In other types of portacaval or mesocaval shunts are tried but results are equivocal.
- Mesoatrial shunt between right atrium and SMV using Gortex graft.
- TIPSS is also useful.

HEPATIC FAILURE

Aetiology

- Cirrhosis.
- Congenital diseases.
- Terminal liver diseases.
- Drug induced.
- Fulminant hepatitis.

Precipitating Causes

- Infection.
- GIT bleeding.
- Surgery.
- General anaesthesia.
- Drugs—narcotics, analgesics.
- Old age.
- Electrolyte abnormality—hypokalaemia.

Clinical Features

- Weakness and jaundice.
- Cyanosis.
- Fever, fetor hepaticus.
- Features of hyperdynamic circulation: Bounding pulse, flushed extremities, fresh crops of spider naevi, palmar erythemas.
- Loss of axillary and pubic hairs.
- In male, gynaecomastia and testicular atrophy.

Fulminant Hepatic Failure

It occurs as a result of massive liver cell death due to:
1. Acute viral hepatitis.
2. Drugs like carbon tetrachloride, halothane, cytotoxic drugs.

Presentation

- Sudden onset of the disease
- Personality changes
- Delirium
- Altered behaviour
- Oliguria and kidney failure

Treatment

- Correction of precipitating factors, e.g. GIT bleed, infection, hypokalaemia.
- High carbohydrate diet.
- IV fluids or nasogastric feeding.
- Diuretics and morphine is avoided as it may precipitate coma.
- Correction of hypoglycaemia and electrolytes.
- Management of renal failure by haemodialysis.
- Correction of clotting factors.
- Temporary hepatic support like:
  - Exchange blood transfusion
  - Cross circulation between donor and patient
  - Extracorporeal perfusion of patient's blood through pigs liver.
  - Perfusion through columns of activated charcoal.

HEPATIC ENCEPHALOPATHY

It is a chronic neuropsychiatric disorder wherein blood bypasses the liver, as a result of which, detoxification of *toxic products* do not occur, which, in turn enters the brain to cause toxicity.

Toxic Products

1. Ammonia.
2. Aminobutyric acid.
3. Mercaptan.
4. Methionine.
5. Short chain fatty acids.

Collaterals developed in chronic liver disease like cirrhosis or after shunt surgeries, allow the blood with toxic products to enter the brain causing toxic effects (As toxic products of bacterial degradation, and nitrogenous material from the bowel is normally metabolized in the liver, these toxic products bypass the liver through collaterals to enter the cerebral circulation).

Encephalopathy is precipitated by:

- Large protein meal.
- GIT bleed.
- Hypokalaemia.
- Drug-induced diuresis.
- Uraemia.
- Narcotics.
- Infection.
- Dehydration.
- Constipation.

Clinical Features

- Disorientation, *flapping tremor*.
- Cogwheel rigidity of limbs, ankle clonus.
- Slurred speech, intellectual deterioration.
- Personality changes.

Grading of hepatic encephalopathy

- Decreased skills—Subclinical
- Confusion—Grade 1
- Drowsiness—Grade 2
- Somnolence, hyperreflexia—Grade 3
- Coma—Grade 4

Treatment

To reduce the entry of nitrogenous products and toxic products from the gut into the circulation.

- Restriction of dietary protein.
- Removal of blood from bowel by purgation or enemas.
- Antibiotics administration.

Oral neomycin to reduce bacterial load inside the bowel lumen. 1 gm 4-6 hours or metronidazole 250 mg every 8 hours.

- Oral lactulose 30-50 ml 4 times a day.
- Sedatives like benzodiazepines and opioids are avoided.
- Correction of vitamin and electrolyte deficiencies.
- Lactulose enema: 300 ml of lactulose in 700 ml of water can be given.

HEPATORENAL SYNDROME

- It is development of acute renal failure due to severe hepatic or biliary disease with jaundice.
- Patient develops oliguria, azotaemia and hyponatraemia.

Causes

- Bile salt sludging in the tubules.
- Absorption of toxins.
- Increased ADH release.
- Hypoperfusion and renal ischaemia.
- It may be precipitated by surgery, stress.

Treatment

- Mannitol infusion.
- Diuretics.
- The cause is treated.
- Dialysis.
- Liver transplantation.

HEPATIC RESECTION

Anatomy and Types of Resections

- Right hepatic artery divides into anterior and posterior segmental arteries. Left hepatic artery divides into medial and lateral segmental arteries.
- Portal vein divides at porta hepatis as right and left branches and these branches enter the respective lobes of the liver.
- Hepatic veins originate as central vein of the liver which is formed by liver sinusoids. Middle hepatic vein is located in main lobar fissure. Left hepatic vein is in left segmental fissure—upper portion; right hepatic vein is in right segmental fissure. Inferior part of middle and anterior segments are drained by middle hepatic vein. In 60% of cases middle and left hepatic veins join to form single trunk to enter IVC.
- Main portal fissure divides liver into two lobes by Cantlie's line from anteroinferior part of gallbladder fossa to left of the IVC with 75° angle with horizontal plane. It divides into right/left paramedian sectors/segments.
- Liver is divided into 8 segments (Couinaud segments). Right lobe is formed by segments V, VI, VII, VIII from front to

Nomenclature for liver resection

Segments	***Brisbane - 2000***	***Couinaud - 1957***	***Goldsmith and Woodburne—1957***
V-VIII	Right hemihepatectomy	Right hepatectomy	Right hepatic lobectomy
IV-VIII	Right trisectiontectomy	Right lobectomy	Extended right hepatic lobectomy
II-IV	Left hemihepatectomy	Left hepatectomy	Left hepatic lobectomy
II, III	Left lateral sectionectomy	Left lobectomy	Left lateral segmentectomy
II, III, IV, V, VIII	Left trisegmentectomy	Extended left hepatectomy	Extended left lobectomy

Indications	***Contraindications***
• Primary/Secondary malignancy from gallbladder or colorectum or carcinoids	• Acute hepatitis—viral/alcoholic
• Benign tumours of liver	• Severe chronic hepatitis
• Parasitic/non-parasitic cysts—hydatid cysts/simple liver cysts	• Poor liver reserve
• Refractory abscess/Caroli disease / RPC	• Severe portal hypertension
• LDLT	• Severe coagulopathy
• Klatskin's tumour	• Severe thrombocytopenia—platelet count less than 30,000/-
• Trauma to liver	

behind. Left lobe formed by segments II, III and IV. Segment I (Spigel lobe) is caudate lobe which has got independent blood supply.

- Resection along the main portal fissure (Cantlie's line) and removal on right side is *right hemihepatectomy* and removal on left side is *left hemihepatectomy*.
- Resection on right of the ligamentum teres is called as *right trisegmentectomy* (older—right lobectomy); and left of the ligamentum teres is called as *left lateral segmentectomy*-segment II and III (older left lobectomy).
- Removal of single segment is called as *unisegmentectomy.*
- *Plurisegmentectomy* is removal of two or more segments. Segments IV and V are removed in carcinoma of gallbladder.
- Removal of small portion of liver either within the segment or traversing segments is called as *wedge resection* of the liver.

Technical Principles

Inflow (portal vein, hepatic artery, bile duct) control; outflow (hepatic veins) control; parenchymal transection; preservation of adequate sized liver remnant with intact inflow, biliary drainage and venous outflow; mobilisation of left and right triangular ligament; ligation of branches of portal vein and hepatic artery and resection of hepatic duct branch of the side; ligation of hepatic vein at a later period; proper haemostasis and ligation of individual biliary ductules. Retrohepatic caval dissection includes—dissection of right adrenal vein, inferior hepatic veins and IVC ligament.

Types of Hepatectomy Techniques

- Preliminary vascular section (Lortat-Jacob).
- Primary parenchymal transection (Ton That Tung in 1965).
- Selective clamping.
- Total vascular exclusion.
- Pedicular clamping (Pringles).
- Suprahilar clamping (Glissonian approach).
- Intrahepatic portal control.

Procedure

- Proper preoperative preparation is essential. LFT, prothrombin time is done. Adequate blood, FFP, along with central line is kept ready.
- Subcostal or thoracoabdominal incision is needed.
- Proper mobilisation, identification and dissection of portal areas is essential; dissection of hepatic veins, usage of vascular clamp secured ligation of branches of portal vessels and major hepatic veins are essential steps.

Techniques of parenchyma dissection	***Intraoperative ultrasound is done to:***
• Kelly clamp and bipolar forceps	• Identifying tumour
• Water jet dissection	• Recognising the anatomy
• Ultrasonic dissection	• Unexpected contralateral tumour
• Ultrasound cutting	• Intrahepatic selective vascular control using a balloon catheter
• Dissecting sealer	• Guided biopsy

How much to cut:

- Remnant liver—≥30% of the original liver volume with complete venous drainage is safe

After completion of resection	***Complications***
• Check for bile leak by injecting dye through cystic duct	• Bleeding, infection, septicaemia
• Check for absolute haemostasis.	• Subphrenic abscess formation
• Cholangiogram to confirm contralateral biliary patency	• Bile leak, haemobilia
• Fix the left lobe	• Liver failure
	• Pleural effusion, empyema

- Liver is dissected by finger dissection, back of scalpel, Cavitron ultrasonic surgical aspirator (CUSA), laser ligation of biliary radicles and vascular branches individually using non-absorbable sutures are other methods.
- Control of bleeding is done by on table Pringle manoeuvre, gel foam, cautery, laser and vessel ligation.
- Proper antibiotic coverage is needed to prevent sepsis.
- Postoperative ICU care is needed.

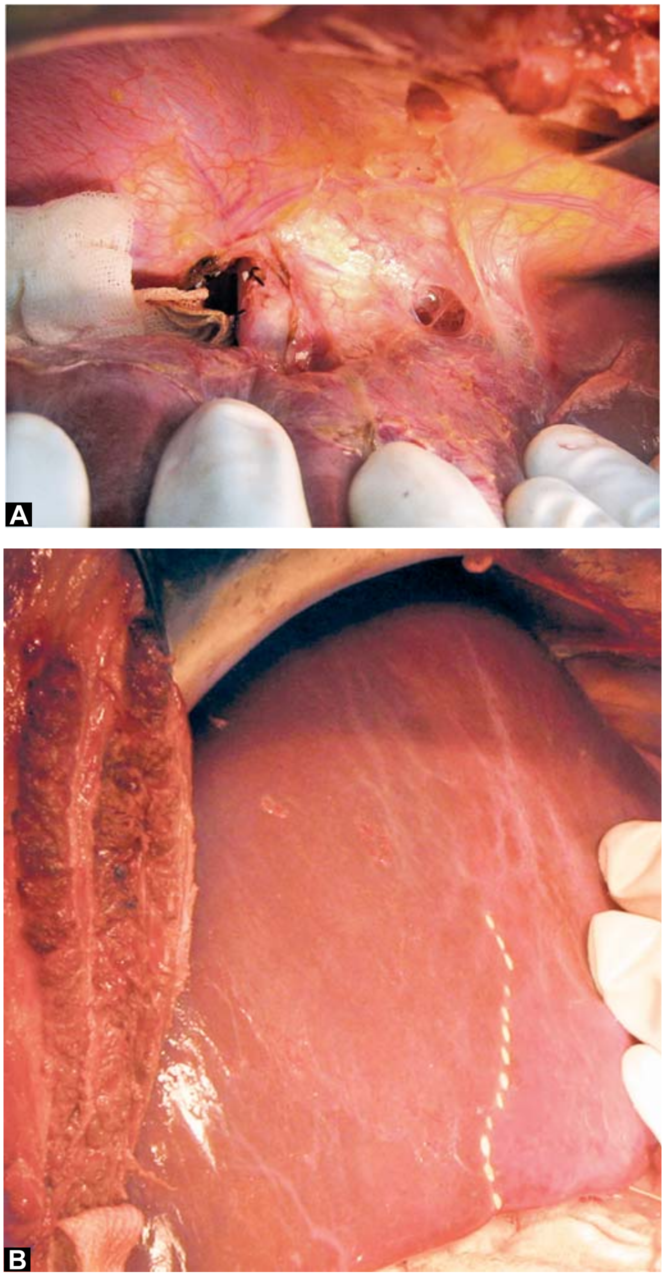

Figs 11.54A and B: Liver resection. Suprahepatic outflow control and parenchymal marking is shown (*Courtesy:* Dr Ravishankar, MCh, Gastroenterologist and Liver Transplantation Surgeon, Bengaluru).

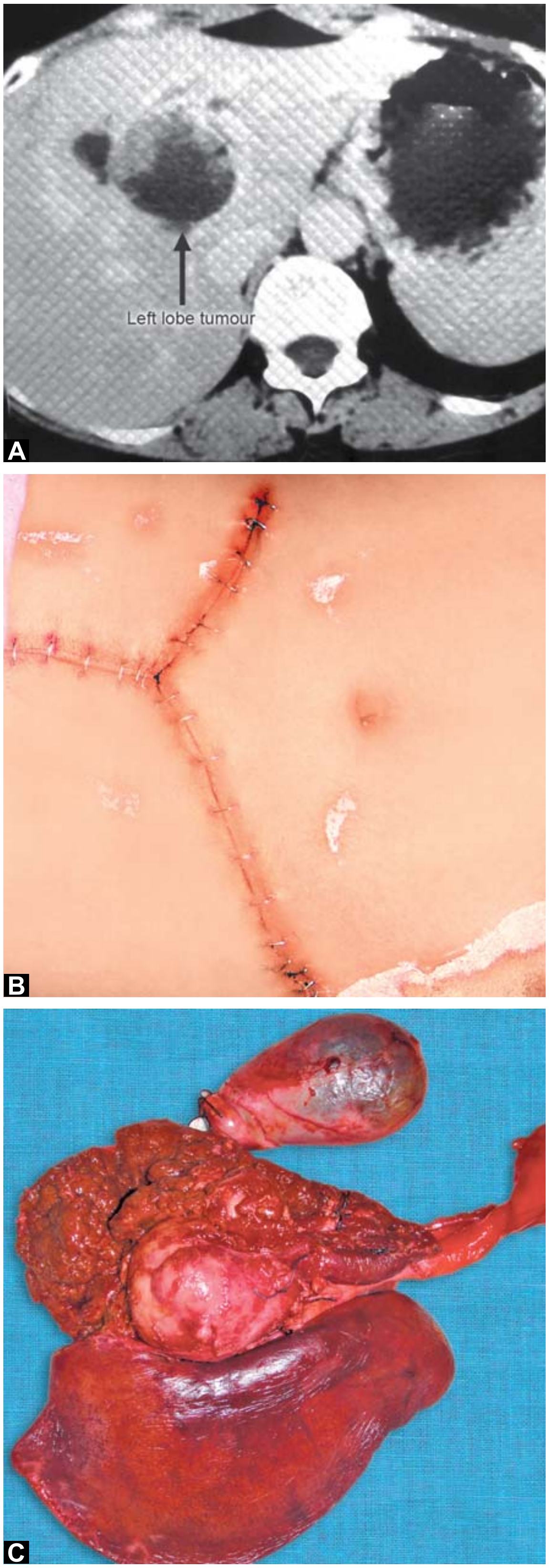

Figs 11.55A to C: Liver resection of left lobe tumour. CT picture, incision scar and specimen are also shown.

The best way to behave when crisis strikes is to be brave.

PORTAL BILIOPATHY

- In patients with portal hypertension, particularly with extrahepatic portal vein obstruction, *portal biliopathy* producing biliary ductal and gallbladder wall abnormalities are common. Portal cavernoma formation, *choledochal varices* and ischemic injury of the bile duct have been implicated as causes of these morphological alterations.
- Gallbladder varices, bile duct varices are typical. Often it may cause torrential haemorrhage per se or on table if patient is undergoing choledocholithotomy.
- While a majority of the patients are asymptomatic, some present with a raised alkaline phosphatase level, abdominal pain, fever and cholangitis.
- Choledocholithiasis may develop as a complication and manifest as obstructive jaundice with or without cholangitis.
- Endoscopic sphincterotomy and stone extraction can effectively treat cholangitis when jaundice is associated with common bile duct stones.
- Definitive decompressive shunt surgery is sometimes required when biliary obstruction is recurrent and progressive.
- Portal biliopathy is a specific entity which may cause severe technical difficulties especially bleeding during ERCP sphincterotomy or open biliary surgeries.

Differential diagnosis of space occupying lesions in the liver

- Hepatoma
- Amoebic liver abscess
- Secondaries in the liver
- Cysts of the liver
- Haemangioma

Chapter 12 Gallbladder

> *If the passages which convey the bile to the intestine, be obstructed from inflammation or scirrhous, the bladder gets over-distended, and the bile regurgitates; it therefore becomes mixed with the blood, and the blood, passing over the whole system, carries the bile to every part of the body, which acquires the appearance of bile. But the hardened faeces are white and clayey, as not being tinged with bile, because the bowels are deprived of this secretion.*
>
> **—Aretaeus the Cappadocian**

CHAPTER OUTLINE

- Surgical Anatomy
- Oral Cholecystogram
- IV Cholangiogram
- Endoscopic Retrograde Cholangiopancreatography
- Percutaneous Transhepatic Cholangiography
- Magnetic Resonance Cholangiopancreatography
- Radioisotope Scan Study
- Peroperative Cholangiogram
- Postoperative T-Tube Cholangiogram
- Congenital Anomalies of Gallbladder
- Choledochal Cysts
- Caroli's Disease
- Biliary Atresia
- Gallstones
- Acute Cholecystitis
- Acute Acalculous Cholecystitis
- Empyema Gallbladder
- Mucocele of the Gallbladder
- Chronic Cholecystitis
- Cholecystoses
- Dissolution Therapy for Gallstones
- Choledocholithiasis
- Sump Syndrome
- Courvoisier's Law
- Surgical Jaundice
- CBD Strictures
- Sclerosing Cholangitis
- Gallbladder Polyp
- Benign Biliary Papilloma
- Carcinoma Gallbladder
- Cholangiocarcinoma
- Klatskin Tumour
- Biliary Fistulas
- Gallstone Ileus
- Hemobilia
- White Bile
- Cholecystectomy
- Laparoscopic Cholecystectomy
- Single Incision Laparoscopic Surgery (SILS) in Cholecystectomy
- Bile Duct Injuries
- Post-cholecystectomy Syndrome
- Biliary Dyskinesia

SURGICAL ANATOMY

Gallbladder

It is a pear-shaped (size is 5-12 cm) reservoir, located in a fossa on the inferior surface of the liver.

Parts

- Fundus, body, infundibulum and neck. *Hartmann's pouch* is pathological one located in the infundibular region created by gallstones.

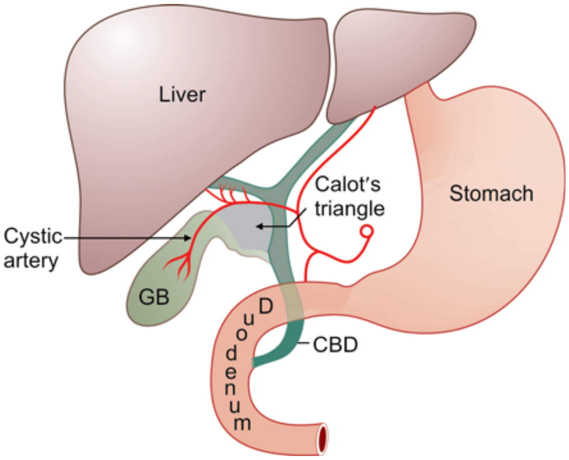

Fig. 12.1: Anatomy of gallbladder.

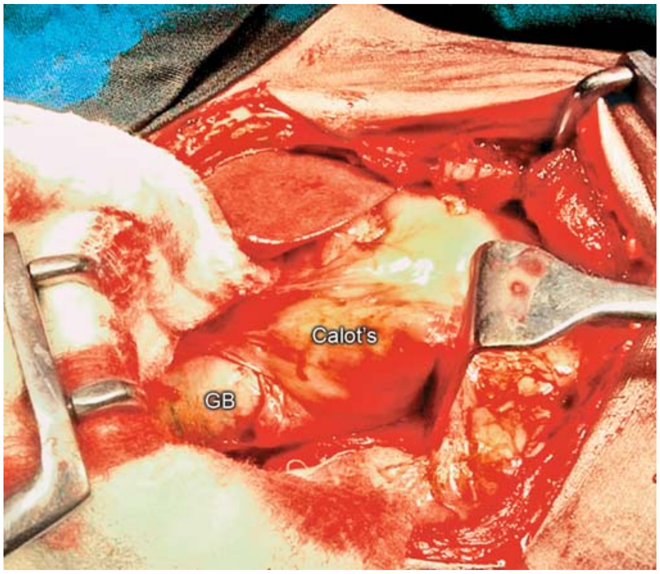

Fig. 12.2: On table finding of empyema gallbladder with inflamed areas near Calot's triangle.

- Gallbladder drains through cystic duct into common hepatic duct to form common bile duct.
- It is supplied by cystic artery, a branch of right hepatic artery.
- *Calot's triangle* is formed by common hepatic duct to the left, cystic duct below, and inferior surface of liver above. Cystic artery originating from right hepatic artery passes behind the common hepatic artery, enters the Calot's triangle to reach the gallbladder. It contains lymph node of '*Lund*' *(Fred Bates Lund).*
- Often cystic artery, hepatic artery, cystic duct have anomalous positions and anomalous origins. Both gallbladder neck and cystic duct contain mucosal fold called *valves of Heister.*

Extrahepatic Biliary Tree

- The left hepatic duct is formed by the ducts draining II, III, IV segments of the liver. The right hepatic duct is formed by the ducts draining V, VI ,VII, VIII segments of the liver. Both join to form the *common hepatic duct*, which joins with the cystic duct to form the *common bile duct*.
- Common bile duct is normally 10-12 cm in length and 6-8 mm in diameter. It joins the major pancreatic duct in the wall of the 2nd part of duodenum to form the *ampulla of Vater*. Intraduodenal part of common bile duct (CBD) is surrounded by smooth muscle fibres called as *sphincter of Oddi.*

Blood supply: From gastroduodenal, retroduodenal, posterosuperior pancreaticoduodenal arteries.

Bile

- Daily up to 1000 ml of bile is secreted from the liver which contains water (98%), bile salts, bile pigments, fatty acids, lecithin, cholesterol, and electrolytes (sodium, potassium, chloride, bicarbonate, calcium, magnesium) with a pH more than 7.0.
- Main function of gallbladder is to concentrate and store the bile. Capacity of gallbladder is 40-50 ml.
- Bile salts form micelle which makes cholesterol soluble. Bile salts (i.e. salts of cholic and chenodeoxycholic acids) with formed deoxycholic acid (in the bowel) enters the enterohepatic circulation to get resecreted in the bile.
- Bilirubin conjugated in the liver is secreted into the bile, which in the bowel is converted into urobilinogen by bacteria. Urobilinogen gets absorbed in the bowel, enters the liver for resecretion again and part of it is excreted in the urine. Absence of urobilinogen in the urine signifies obstructive jaundice.
- In the absence of gallstones or any other disease, bile is sterile. Symptomatic gallstone disease shows positive culture for bacteria, commonest being *E. coli* and *Klebsiella*.

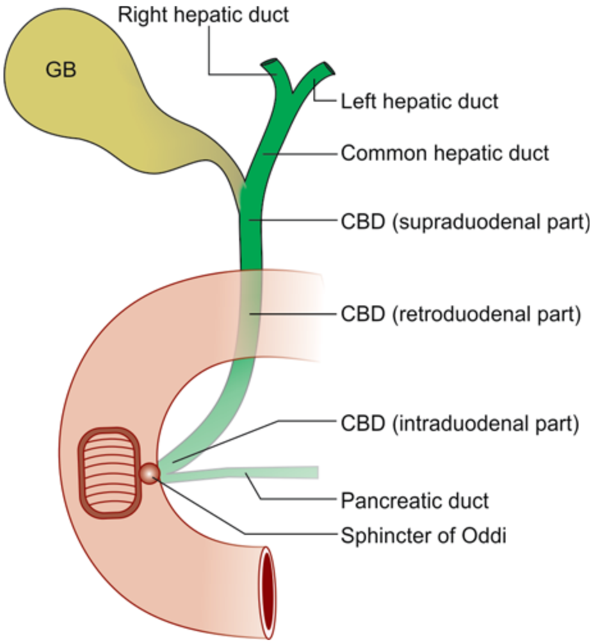

Fig. 12.3: CBD parts.

ORAL CHOLECYSTOGRAM (OCG; GRAHAM-COLE TEST)

- This test is done to study function of the gallbladder.
- Patient is advised to have fat free diet for 3 days. Previous night 6 tablets of *iopanoic acid* (Telepaque) is given orally. On the day of OCG, plain X-ray abdomen in erect posture is taken to visualise the gallbladder.

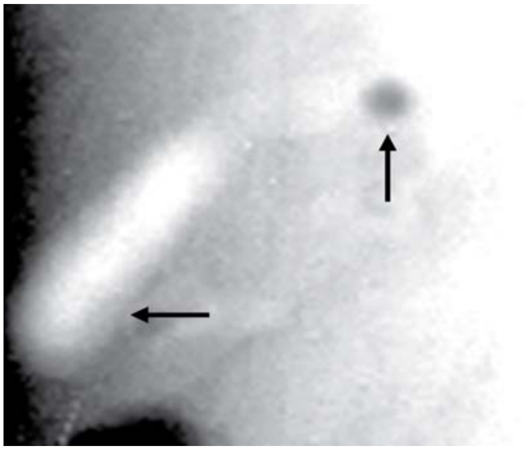

Fig. 12.4: Oral cholecystogram with smooth filling defect (Cystic duct stone).

- Later, fatty meal is given and one more X-ray is taken to see the change in the size of the gallbladder (which should be less in size compared to the earlier film, as the gallbladder contracts on stimulation with fat rich meal if it is functioning normally). Smooth filling defect signifies nonopaque stone.

Note:
- One tablet of Iopanoic acid is 500 mg.

Contraindications

- Patients with serum bilirubin > 3 mg%
- Acute cholecystitis
- Vomiting

IV CHOLANGIOGRAM

- It is to visualise bile ducts and biliary tree, by injecting IV *Meglumine ioglycamate* (Biligram) and taking X-ray abdomen.
- It can be combined with OCG.
- Problems with this method are poor visualisation, drug reaction. It is not very useful if serum bilirubin is > 3 mg%.

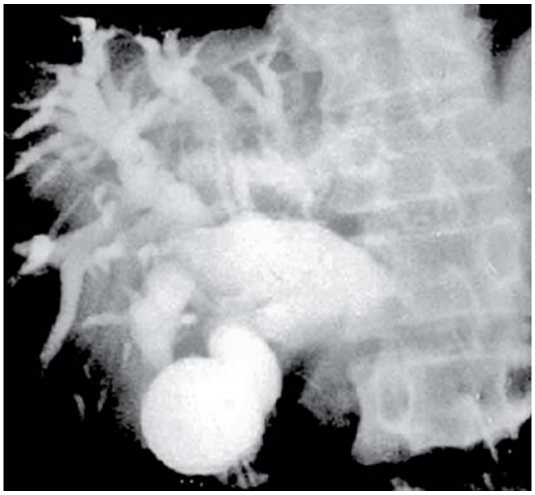

Fig. 12.5: Cholecystocholangiogram. Note the dilated CBD, complete block in lower CBD, dilated biliary radicles.

ENDOSCOPIC RETROGRADE CHOLANGIO-PANCREATOGRAPHY (ERCP)

Through a side viewing gastroduodenoscope, sphincter of Oddi is cannulated, and dye is injected. Biliary and pancreatic trees are visualised. It is done under C-ARM guidance and sedation like midazolam or propofol anaesthesia. Patient is placed in prone position with the head turned towards right. After passing

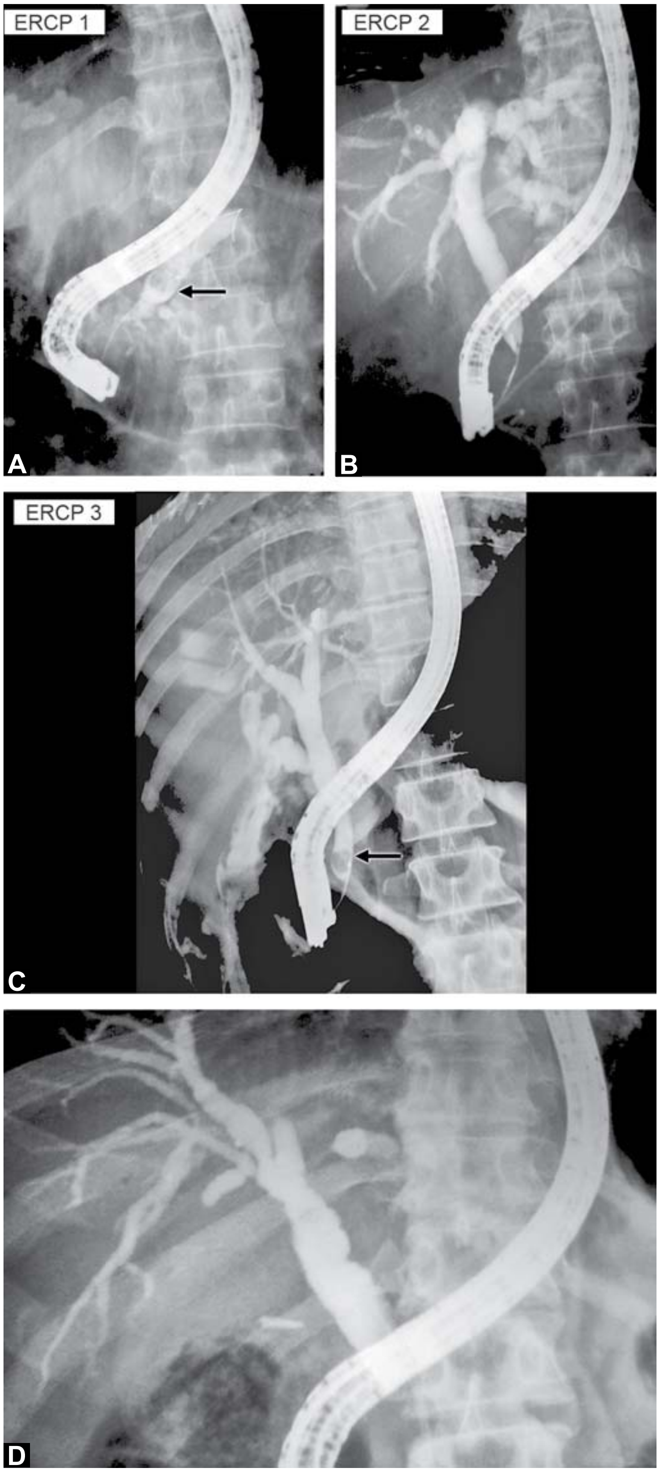

Figs 12.6A to D

There is little if any restriction to respiratory movements of the abdominal wall in pleuropneumonia.
—C Allan Birch

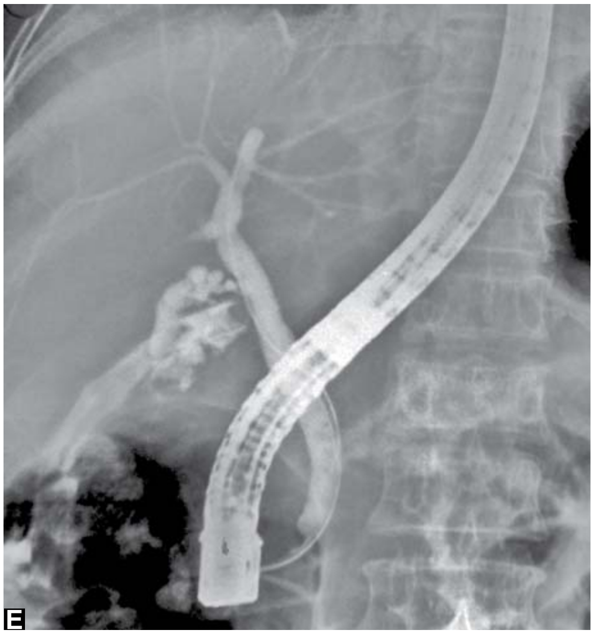

Fig. 12.6E

Figs 12.6A to E: ERCP being done. Note the gastroduodenoscope with injection of dye. (A) Shows filling defect in the CBD. (B) Shows dilatation of biliary radicles. (C) Shows radiolucent stone (smooth filling defect) in distal CBD which can be removed through ERCP. Antibiotics should be given to prevent cholangitis.

gastroduodenoscope, sphincter is identified and cannulated. Under visualisation 3 ml of water soluble iodine contrast is injected into the bile duct and pancreatic duct. When cannula goes upwards beside vertebra, it signifies that it is in bile duct; and if cannula goes across the vertebra it is in pancreatic duct.

Indications

- Malignancy—irregular filling defect.
- Chronic pancreatitis—*'chain-of-lakes'* appearance.
- Congenital anomalies.
- Stones.
- Stricture of biliary tree.
- Choledochal cyst.
- For sampling of biliary and pancreatic juices for analysis and cytology.
- Brush biopsy from tumour site.

Therapeutic uses

- Extraction of stone from biliary duct
- Nasobiliary drainage
- Stenting of tumour in the CBD or in the pancreas
- Dilatation of the biliary stricture
- Endoscopic papillotomy

Complications

- Pancreatitis.
- Duodenal injury, perforation.
- Cholangitis.
- Bleeding from pancreaticoduodenal artery.
- Sphincter stenosis.

Relative Contraindications

- Acute pancreatitis.
- Previous gastrectomy.

PERCUTANEOUS TRANSHEPATIC CHOLANGIOGRAPHY (PTC)

- It is done in case of severe obstructive jaundice under cover of antibiotics and after control of any bleeding tendency—by giving injection vitamin K or FFP transfusion.

Indications

- Failure of ERCP
- High biliary strictures
- Klatskin tumour
- High blocks when external caternal catheter drainage is needed
- Stenting in high tumours

- With the help of fluoroscopy (C-ARM)/US/CT guided, *Chiba* or *Okuda needle* which is long, flexible, thin, blunt, without a bevelled end (15 cm long and 0.7 mm in diameter) is passed into the liver through right 8th intercostal space in midaxillary line. Once needle is in the dilated biliary radicle, bile is aspirated (sent for culture, cytology, analysis) and then water soluble iodine dye is injected into the same so as to visualise the dilated biliary radicles, also the site and extent of the obstruction, i.e. tumour, stricture.
- Procedure can be used for therapeutic stenting across the biliary tree through the obstruction either in the hepatic ducts or in the CBD into the duodenum.
- It is used whenever ERCP fails, in high strictures (in CHD), in Klatskin tumour, in catheter drainage (external) in high blocks, in stenting high tumours.

Prerequisites for PTC

- Should have normal prothrombin time (or should be made normal by injection vitamin K or FFP).
- Should be a last and final inevitable method for evaluation or therapy.
- Blood for transfusion in case of bleed, consent for intervention in need, antibiotic prophylaxis—are a must prior to PTC.

Note:

- Chiba is an university in Japan. Okuda is the name of person.
- **PTBD** is percutaneous transhepatic biliary drainage. It can be *external or internal*. It is done through PTC.

Complications

- Bleeding
- Biliary leak and biliary peritonitis
- Septicaemia

MAGNETIC RESONANCE CHOLANGIOPANCREATOGRAPHY (MRCP)

- It is a noncontrast noninvasive imaging method, better than ERCP as diagnostic tool in biliary and pancreatic diseases.

RADIOISOTOPE SCAN STUDY

- *I^{131} Rose Bengal and Tc^{99} labelled imino diacetic acid (HIDA, PIPIDA)* are very useful in diagnosing acute cholecystitis and other biliary disorders, like biliary atresia.

Note:

HIDA is Hippuran Immuno Diacetic Acid.

PEROPERATIVE CHOLANGIOGRAM

- It is done during CBD exploration for stricture, residual CBD stones, atresia, choledochal cyst, cholangitis.
- Fine polythene catheter is passed into the CBD through cystic duct and dye is injected. *Under C-ARM image-intensifier*, any block or stricture can be identified and completion of the procedure can be confirmed.

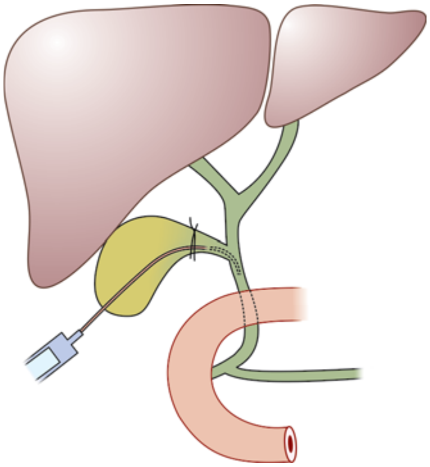

Fig. 12.7: Preoperative cholangiogram. Fine catheter is passed through cystic duct into the CBD. Care should be taken to avoid air getting into the CBD (it will mimic stone). Dye is injected under C-ARM guidance.

Complications

- Infection.
- Bile leak.

Precaution

Air should not be present in the syringe, as it may mimic stones.

POSTOPERATIVE T-TUBE CHOLANGIOGRAM

After choledochotomy, *Kehr's T-tube* is *placed in CBD for 14 days* and then water soluble dye is injected into the tube and X-ray is taken. Complete free flow of dye into the duodenum indicates that there is no blockage. T-tube can then be removed safely. Block indicates residual CBD stones.

Residual CBD stones are removed by:

- Dormia basket
- Fogarty's catheter
- Choledochoscope
- ERCP

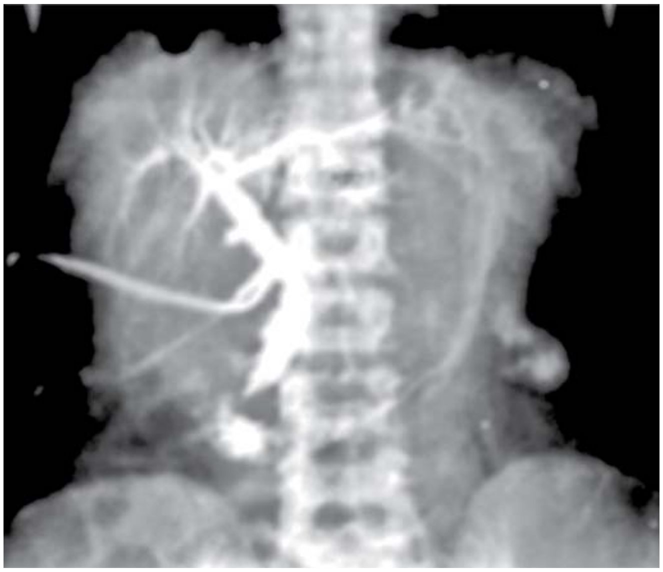

Fig. 12.8: T-tube cholangiogram.

CONGENITAL ANOMALIES OF GALLBLADDER

- *Absence of gallbladder*—rare
- *Phrygian cap*—cap like projection and bend over the fundus of the gallbladder (*Phrygia* is an ancient Asian country—Mongolia) (Liberte cap of French revolution). It is identified in cholecystogram. It is 6% common
- *Double/triple gallbladder*—additional one may be intrahepatic. Cystic ducts may join to form a common cystic duct which joins the CBD or cystic duct of each gallbladder may join separately into the CBD
- *Mobile/floating/mesenteric gallbladder*—gallbladder has a long mesentery attached. It may cause torsion. It can cause recurrent abdominal pain. Its removal is easier through laparoscopic or open method
- *Long cystic duct* with low insertion of the cystic duct into CBD near the ampulla. This is important in all pancreatic/biliary/laparoscopic biliary surgeries
- *Absence of cystic duct* with gallbladder directly entering the CBD through a wide opening near infundibulum
- *Accessory cholecystohepatic duct* (*duct of Luschka*) may be present(10%) which if not identified and ligated during cholecystectomy will cause biliary fistula/ peritonitis due to continuous leak
- *Cystic artery may originate anteriorly* (15%) from right hepatic artery/from common hepatic artery
- Very tortuous hepatic artery in front of the origin of the cystic duct is called as *Moynihan's caterpillar hump*. It is important cause of bleeding in cholecystectomy
- There may be *double cystic duct* or cystic duct may insert high into the common hepatic duct or right hepatic duct

If vomiting or distinct nausea preceeds pain, the case is not one of appendicitis.— **John B Murphy**

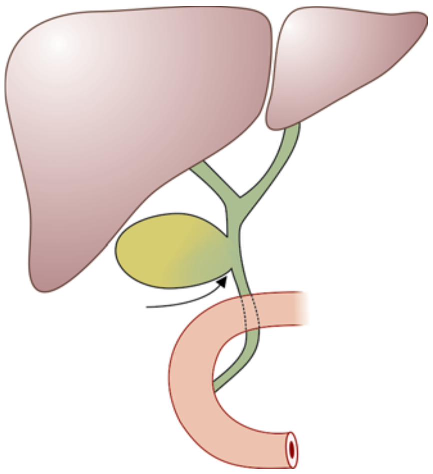

Fig. 12.9: Cystic duct is absent with wide communication into CBD. There is higher chance of CBD injury here.

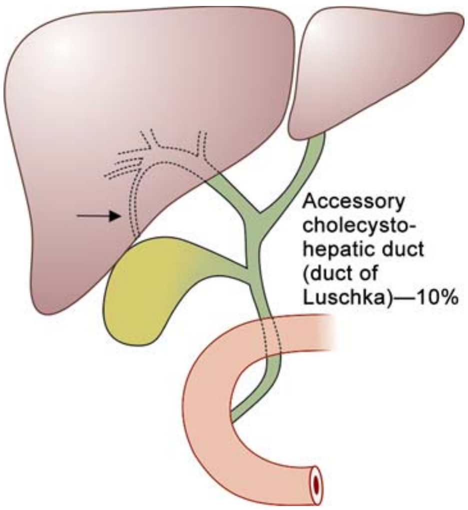

Fig. 12.10: Accessory cholecystohepatic duct (*Duct of Luschka*—10%). It may cause bile leak postoperatively if it is not ligated during cholecystectomy.

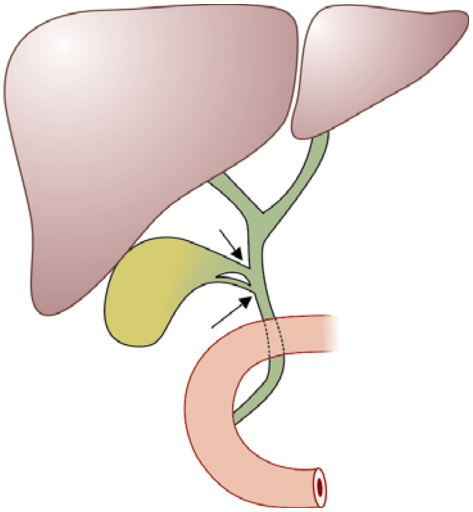

Fig. 12.11: Double cystic ducts from the gallbladder joining the CBD.

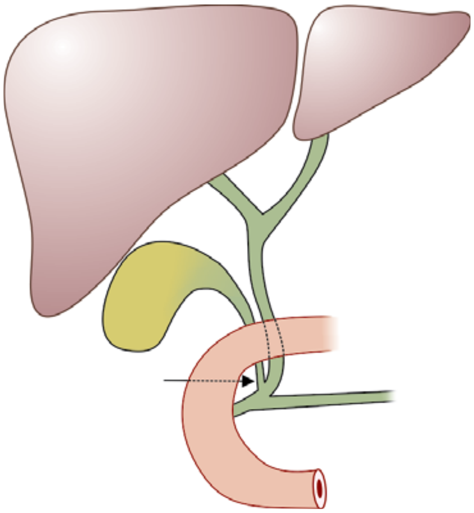

Fig. 12.12: Long cystic duct joining the CBD very low.

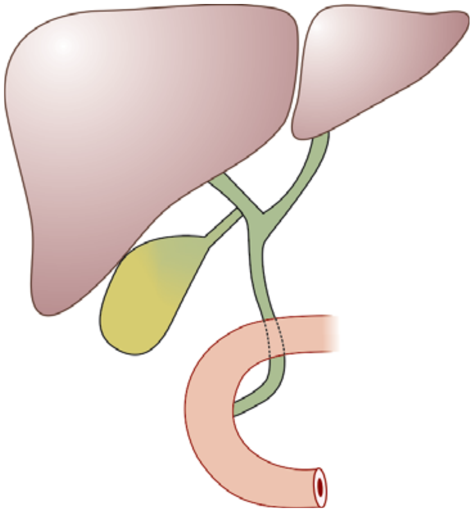

Fig. 12.13: Cystic duct joining the right hepatic duct—a variation.

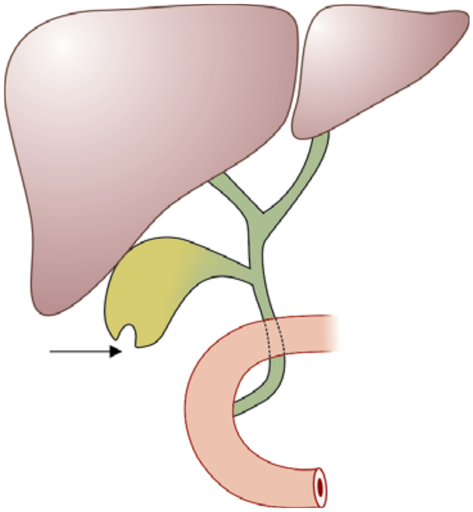

Fig. 12.14: *Phrygian (cap) gallbladder.*

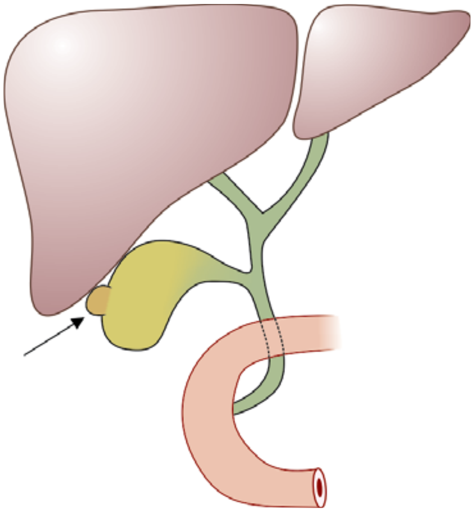

Fig. 12.15: Gallbladder diverticulum.

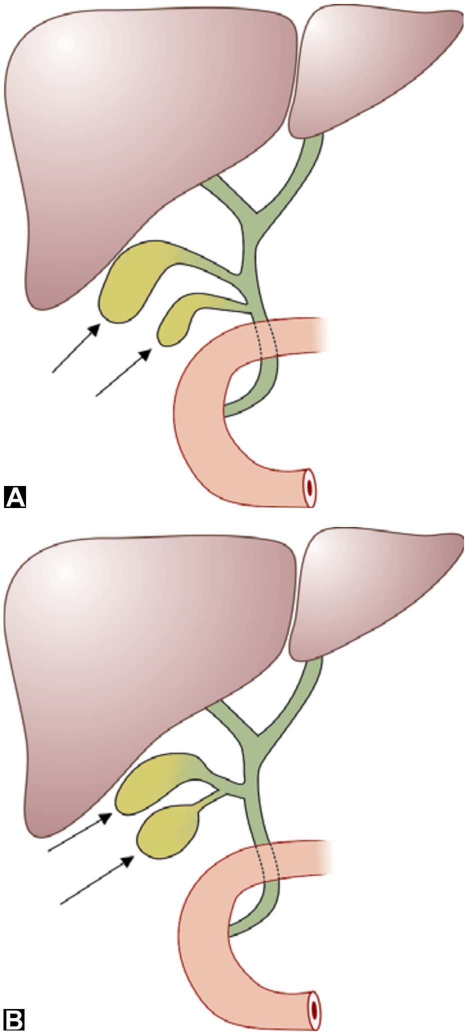

Figs 12.16A and B: Double gallbladder. One may be intrahepatic. It may have separate cystic ducts joining CBD or two cystic ducts join to form single duct to join the CBD.

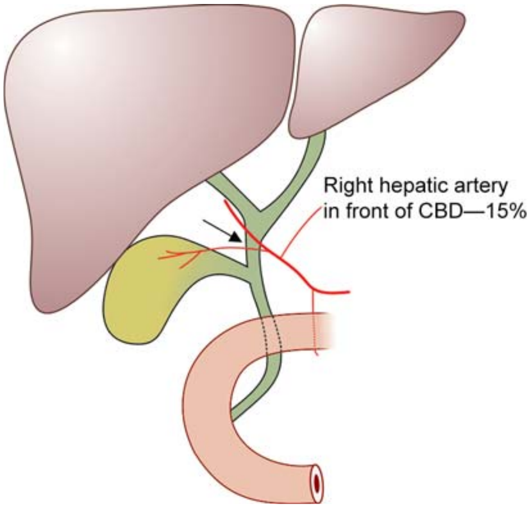

Fig. 12.17: Cystic artery originates from right hepatic artery which is anterior to common bile duct.

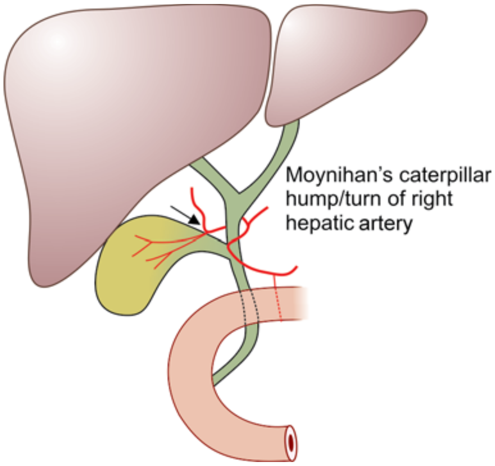

Fig. 12.18: *Moynihan's caterpillar turn/hump* is bend of right hepatic artery in front of/close to cystic duct. It may get injured during cholecystectomy causing torrential haemorrhage.

CHOLEDOCHAL CYSTS

- It is defined as isolated/focal or combined/diffuse congenital dilatation of extra or intrahepatic biliary tree.
- It is congenital cyst with partial or complete weakness of the CBD biliary wall.
- It may be single/multiple or extrahepatic or intrahepatic.
- It is more common in Asia. It is common in Japan. Its incidence is 1 in 1000.
- It is often associated with pancreaticobiliary maljunction. There is long common channel more than 2 cm. It causes reflux of pancreatic juice into the bile duct, enzymatic destruction of the bile duct wall, ductal wall weakening and dilatation—*Babbit theory*.
- During embryogenesis there is abnormal early canalisation of bile duct with distal obstruction causing increased proximal pressure, weakening and ductal dilatation.
- Reduced postganglionic autonomous neurons in the distal portion of the cyst causing its poor function distally.

Gangrenous cholecystitis is associated with thrombosis of cystic artery.

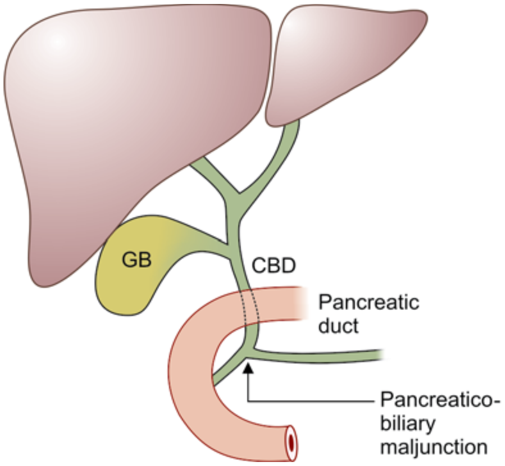

Fig. 12.19: Pancreaticobiliary maljunction. Here pancreatic duct joins CBD more proximally than normal causing reflux of pancreatic enzymes into the CBD. It leads into weakening, dilatation and choledochal cyst.

- Chronic inflammation, sparse mucin glands and metaplasia are the histological features. Histologically it is either glandular with normal cuboidal epithelium with cavities in the mucosa. Or fibrotic type with fibrous cyst wall with thickened bile duct.

Types of Choledochal Cysts (Todani Modification of Alonso-Lej Classification)

- Type I: Dilatation of extrahepatic biliary tree (60%).
    - Type Ia—cystic.
    - Type Ib—focal segmental (saccular).
    - Type Ic—fusiform.
- Type II: Diverticulum of extrahepatic biliary tree (5%).
- Type III: Choledochocele—cystic dilatation of intraduodenal part of CBD (5%).
- Type IV: Dilatation of extra- and intrahepatic or multiple parts of extrahepatic biliary tree (30%).
    - Type IVa: Dilatation of extrahepatic and intrahepatic biliary tree. It is 2nd most common.
    - Type IVb: Dilatation of multiple sections of the extrahepatic bile duct.
- Type V: Dilatation of the only intrahepatic biliary tree (*Caroli's disease*).

Presentations

- Common in females 4:1. Common in Japan.
- Present in infants, children and even in adults.
- Obstructive jaundice (80%), pain in right hypochondrium, cholangitis.
- Mass per abdomen—mass is to the right and above the umbilicus, smooth, soft, not moving with respiration, not mobile and resonant (30%).
- Failure to thrive.

Triad of choledochal cyst (seen in 10% of cases)

- Right upper quadrant pain
- Jaundice
- Palpable abdominal mass

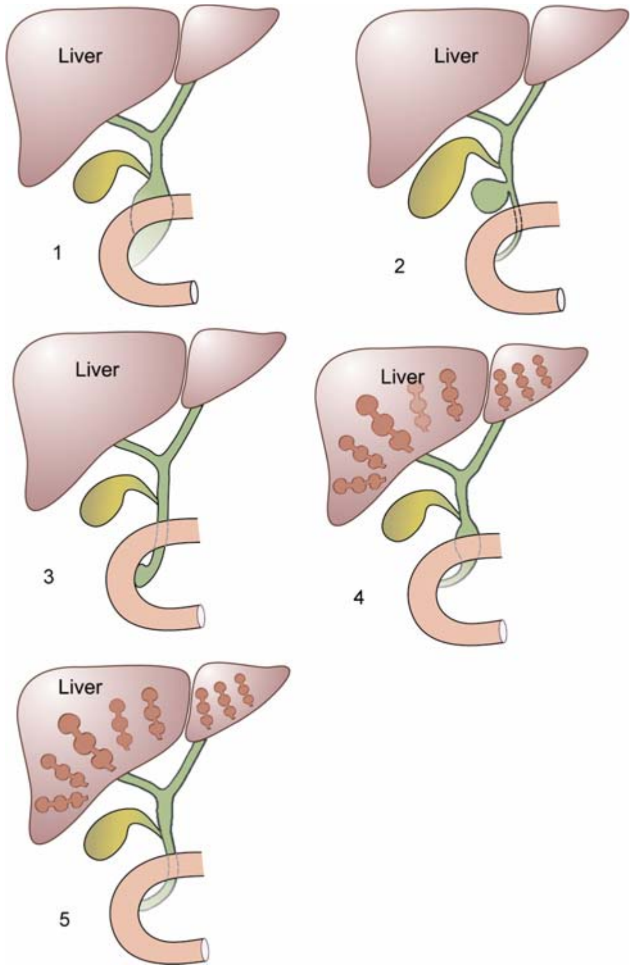

Fig. 12.20: Different types of choledochal cysts—(1) Fusiform type, (2) Saccular type, (3) Choledochocele, (4) Fusiform + Intrahepatic type, (5) Intrahepatic type.

Swellings which appear and disappear

- Hydronephrosis
- Intussusception
- Pseudo pancreatic cyst
- Choledochal cyst

Complications

- Pancreatitis, mainly in type III.
- Suppurative cholangitis.
- Gallstone and CBD stone formation.
- Biliary cirrhosis—secondary; portal hypertension.
- Rupture of cyst and peritonitis.
- Cholangiocarcinoma in CBD (30% of cases). Incidence of malignancy increases with age and is more common in type I and V. Reflux, stasis, super infection, pancreaticobiliary maljunction are the causes for malignant transformation. Malignancy can develop even in residual cyst or at anastomotic site.
- Malignancy is common in posterior wall.

Diagnosis

- U/S abdomen—unilocular cyst mainly in infants.
- CT scan—mainly to see intrahepatic biliary system.

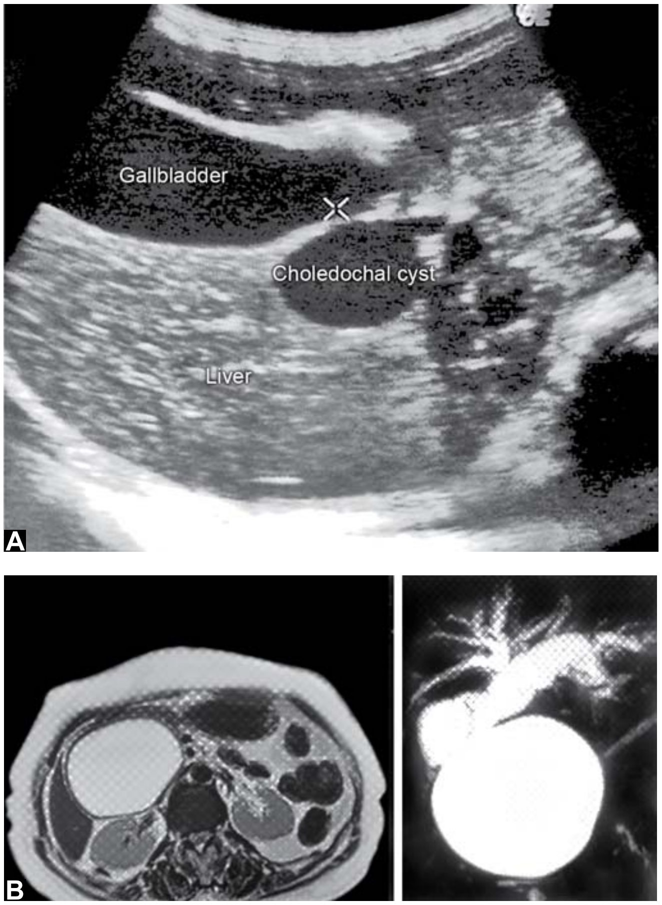

Figs 12.21A and B: CT picture of choledochal cyst (*Courtesy:* Dr Yogish Kumar, MS, Professor, KMC, Mangalore).

- Hepatobiliary nuclide scanning.
- ERCP, Cholangiography—to see ductal anatomy.
- MRCP—to see status of pancreatic and biliary system and pancreaticobiliary maljunction.
- Liver function tests.
- PTC to see intrahepatic biliary tree.

Treatment

- *Resection of extrahepatic biliary tree* with removal of choledochal cyst along with cholecystectomy and Roux-en-Y hepaticojejunostomy is the ideal treatment for choledochal cyst especially types I, II and IVb.
- In Type I—excision of cyst with its mucosa and reconstruction by Roux-en-Y hepaticojejunostomy.
- In Type II—excision of the diverticulum and suturing of the CBD wall is done.
- In type III—endoscopic sphincterotomy is done. Excision is also often needed.
- Intrahepatic dilatation is difficult to treat. If it is localised, hepatectomy is sufficient, but if it is diffuse, liver transplantation may be required.
- In type IV if cyst is adherent to portal vein posteriorly, that part of the cyst wall over the portal vein is left behind. But mucosa of the part should be removed (*Lily's operation*).

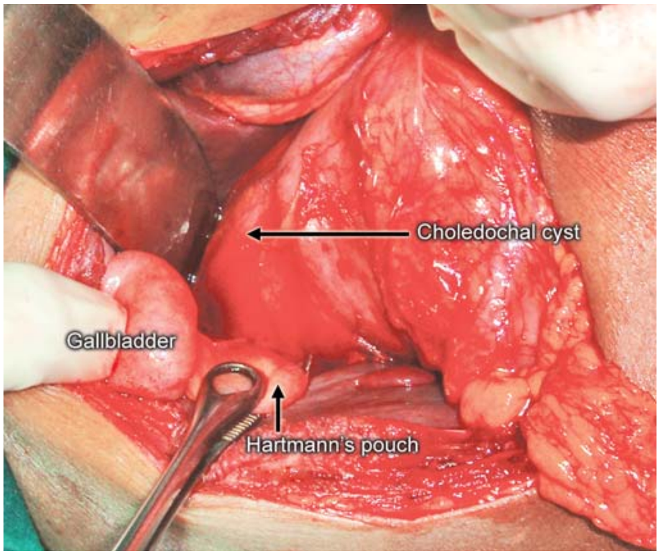

Fig. 12.22: Choledochal cyst on table look.

- Cholecystectomy is a must in all these types. Incidence of gallbladder carcinoma is also high in these patients.
- If cyst has already turned into malignancy—adenocarcinoma, then radical surgery and chemotherapy is given.

CAROLI'S DISEASE

- It is congenital, nonfamilial, multiple, irregular, dilatations of the intrahepatic ducts with stenotic segments in between. It is associated with *congenital hepatic fibrosis* and *medullary sponge kidney.*
- It can also be included under type V choledochal cyst.

Investigations

- ❖ CT scan
- ❖ ERCP
- ❖ LFT
- ❖ MRCP

Complications

- Recurrent cholangitis.
- Stones in biliary tree.
- Portal hypertension.
- Cholangiocarcinoma (7%).

Treatment

1. Liver transplantation.
2. Hepatectomy.

BILIARY ATRESIA

- It may be either due to viral infection or defective embryogenesis resulting in fibrosis of extra- and intrahepatic biliary tree.

Porcelain gallbladder is radiopaque gallbladder, due to calcification of the gallbladder wall following chronic cholecystitis which has got high malignant potential.

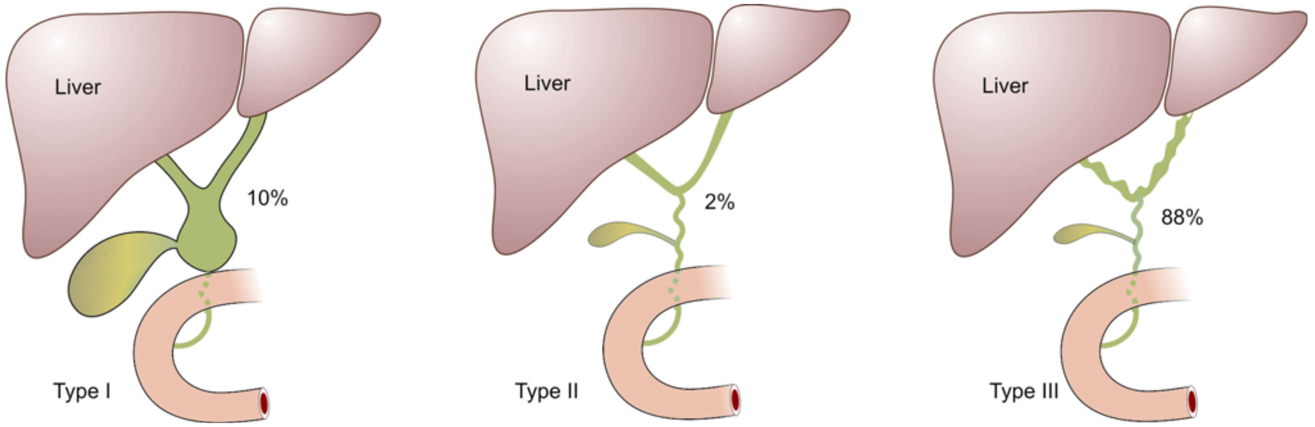

Fig. 12.23: Types of biliary atresia.

- Incidence is 1 in 10,000.
- It is obstructive cholangiopathy with inflammation, destruction and obliteration predominantly involving the extrahepatic biliary tree. There is bile duct proliferation, cellular and canalicular bile stasis, bile plugs in portal tract bile system, periportal inflammation and fibrosis.

Aetiology of biliary atresia

- Infections—cytomegalovirus, reovirus, rotavirus and human papilloma virus
- Immune and autoimmune diseases
- Abnormal development of the biliary system
- Toxins
- Vascular defects of the hepatic artery

Classification

- *Correctable—10%.*
- *Noncorrectable—90%.*

Biliary atresia

- Type I: Atretic CBD—10%
- Type II: Atretic CBD and common hepatic duct
- Type III: Atretic CBD, common hepatic duct and right and left hepatic ducts—88%

Biliary atresia can be:

- *Perinatal acquired* (70%), not associated with congenital anomalies.
- *Fetal/embryonic* (30%) is associated with congenital anomalies.

It may be associated with other congenital anomalies (20%) of the GIT—(malrotation, annular pancreas duodenal atresia) and congenital heart diseases, polysplenia, situs inversus, absent vena cava, preduodenal portal vein.

Clinical Features

- Progressive jaundice in a newborn.
- Steatorrhoea (Pale stool is common).
- Hepatomegaly.
- Splenomegaly with portal hypertension.
- Osteomalacia.
- Biliary rickets.
- Severe pruritus.
- Clubbing, skin xanthomas.

Differential Diagnosis

- Neonatal hepatitis.
- Choledochal cyst.
- Cholestatic jaundice.
- Sclerosing cholangitis.
- Metabolic diseases.
- Storage disorders.
- Alagille syndrome—hypoplastic biliary ducts.

Diagnosis

- Liver function test—conjugated hyperbilirubinaemia.
- US abdomen shows triangular cord, tubular echogenisity extending above the bifurcation of the portal vein which is more than 4 mm thick—diagnostic.
- MRCP—100% accuracy.
- Radioisotope study (HIDA or PIPIDA scan).
- On table cholangiogram.
- Liver biopsy (Percutaneous)—is usually first test obtained.
- Rose Bengal I^{131} study, if shows < 10% secretion it is biliary atresia. If secretion is > 10% it is more likely to be hepatitis.

Treatment

- *In correctable cases*: Roux-en-Y jejunal anastomosis.
- In noncorrectable cases, hepaticoportojejunostomy (*Kasai's operation*).
- Liver transplantation is becoming more popular in biliary atresia—ideal.
- Presently *Kasai (1974) porto enterostomy* is done (in 8 weeks) as a preliminary procedure followed by liver transplantation eventually. After opening the abdomen, gallbladder is cannulated and on table cholangiogram is done to confirm atretic segment. Liver biopsy should also be done at the same time. If biliary patency is found procedure is abandoned. In atretic biliary tree, gallbladder is dissected off the liver and by applying traction to it, fibrotic atretic biliary tree is felt and dissected upto intrahepatic segment (upto the segmental branches of portal vein). Portal plate is anastomosed to Roux-en-Y part of the small bowel. Postoperatively, antibiotics, intravenous corticosteroids initially, later orally (for 6-12 weeks), ursodeoxycholic acid (for 1 year), nutrition, medium chain and essential fatty acids, essential elements supplementation is given. Success of

procedure depends on age of the patient; presence of cholangitis; ductule size (more than 200 µm is better); bridging fibrosis; type of atresia; bilirubin level. Eventually cholangitis, portal hypertension, variceal bleeding will develop.

GALLSTONES

> *Thus with stone obstruction of the common duct, dilatation of the gallbladder is rarely observed; the organ has already undergone contraction; with obstruction from other causes, dilatation is to be expected.*
>
> —**Ludwig Courvoisier,** 1890

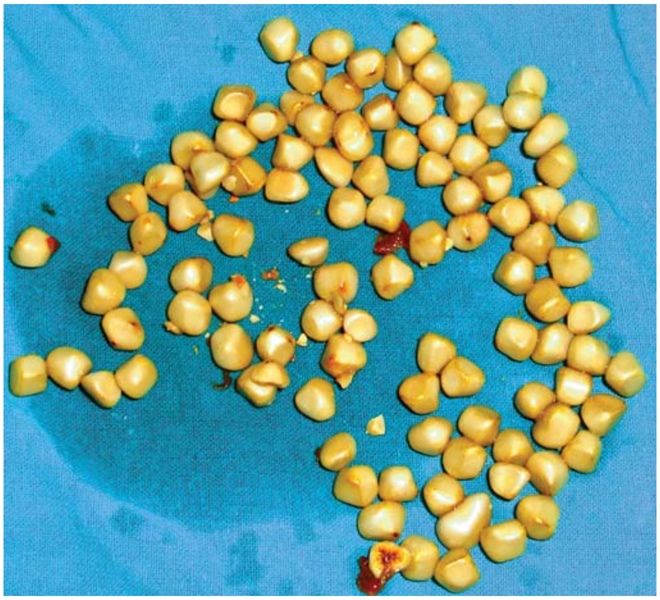

Fig. 12.24: Multiple gallstones removed along with gallbladder.

Types

1. *Cholesterol stones* (Cholesterol solitaire—radiating crystalline appearance) are 6% common, often solitary.

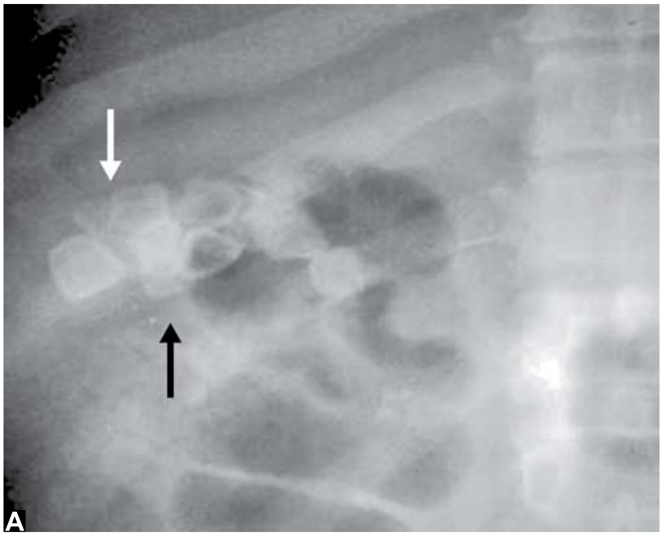

Fig. 12.25A

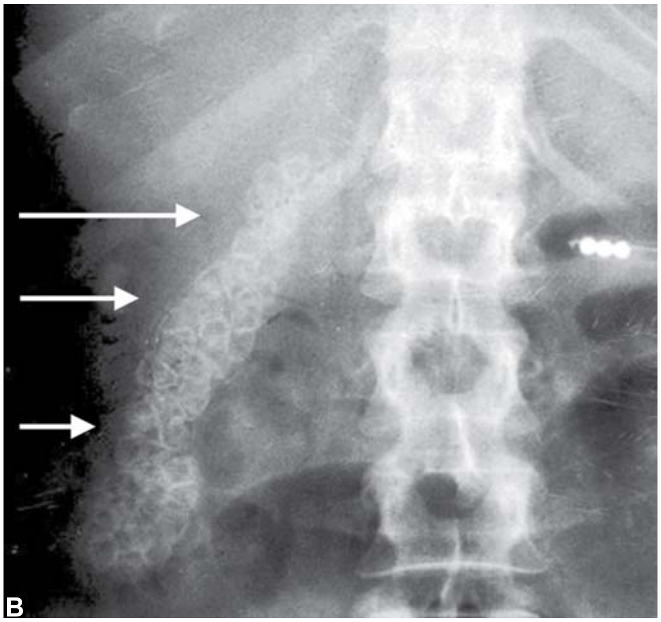

Fig. 12.25B

Figs 12.25A and B: Multiple gallstones in a plain X-ray. Only 10% gallstones are radio-opaque. Often they are faceted because of compactness and equal pressure exerted by each stone. Centre of the gallstone is often found radiolucent and is called as *Mercedes Benz sign/Seagull sign*.

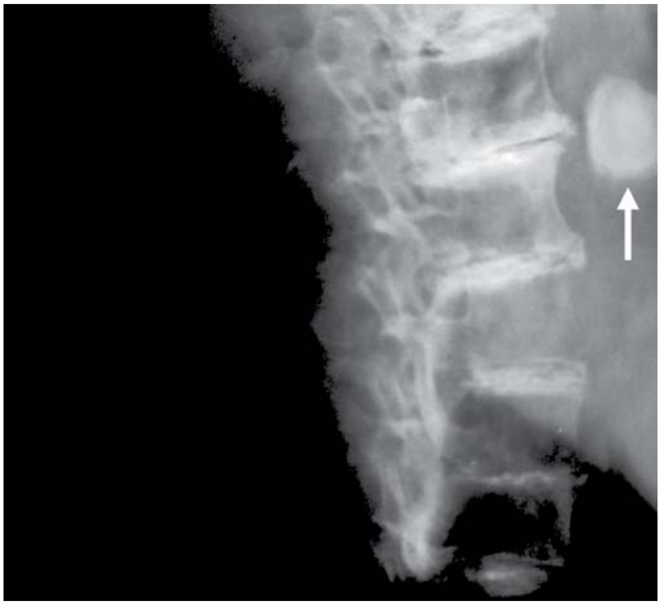

Fig. 12.26: Lateral X-ray of spine showing radio-opaque shadow in front of the vertebrae—could be radio-opaque gallstone. Note also the fused intervertebral disc spaces.

2. *Mixed stones* are 90% common. It contains cholesterol, calcium salts of phosphate carbonate, palmitate, proteins, and are multiple faceted.
3. *Pigment stones* are small, black or greenish black, multiple. Often they can be sludge like.

> *It frequently happens that gallstones are found in the gallbladder after death. Where there was not the least suspicion of existence during life.*
>
> — **Mathew Baillie**

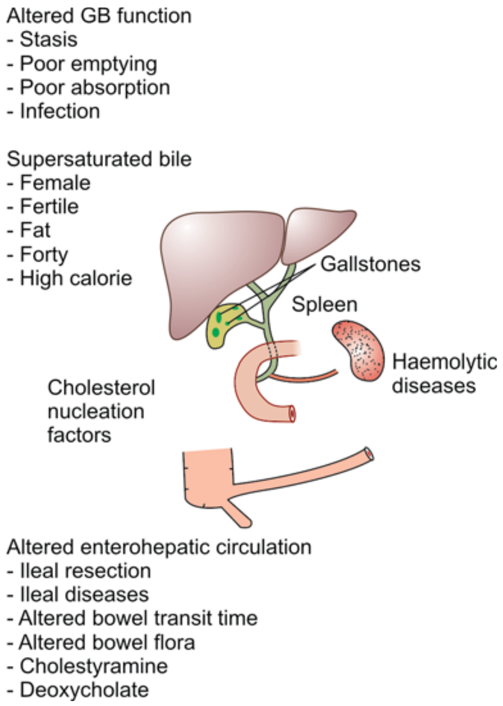
Altered GB function
- Stasis
- Poor emptying
- Poor absorption
- Infection
Supersaturated bile
- Female
- Fertile
- Fat
- Forty
- High calorie
Gallstones
Spleen
Haemolytic diseases
Cholesterol nucleation factors
Altered enterohepatic circulation
- Ileal resection
- Ileal diseases
- Altered bowel transit time
- Altered bowel flora
- Cholestyramine
- Deoxycholate

Fig. 12.27: Diagram showing different causes of gallstone formation.

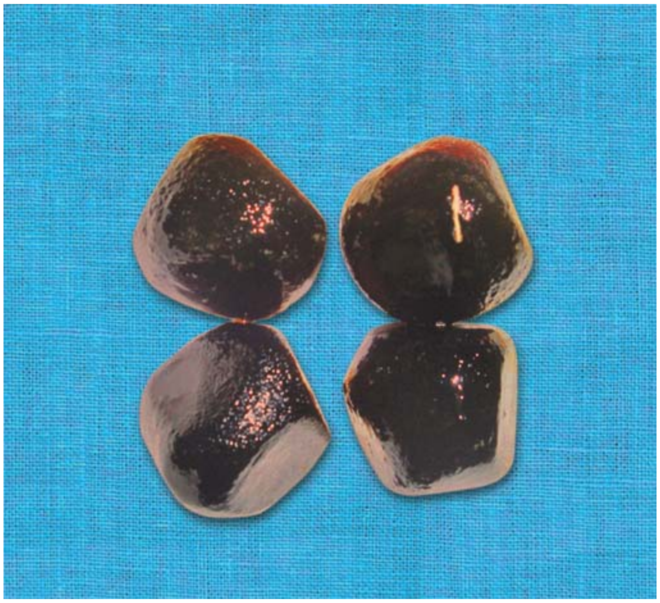

Fig. 12.28: Pigment stones showing faceted look.

Common in *"Fat, Fertile, Forty, Flatulent, Female"*. Common in western countries and in north India.

Pathogenesis

I. ***Metabolic:***

- Cholesterol is synthesised in liver. Its solubility is determined by relative concentration of cholesterol, bile salts and lecithin. Altered levels of cholesterol, lecithin, and bile salts in bile reduces the micelle concentration in the bile leading to precipitation of insoluble cholesterol, hence, the stone formation *(Lithogenic bile)*.
- Normal ratio of bile salt and lecithin to cholesterol is 25:1. Ratio below 13:1 leads to precipitation of cholesterol. Insoluble cholesterol is within the soluble micelle which is formed by lecithin and bile salts. If cholesterol component increases bile gets supersaturated and inadequate micelle makes insoluble cholesterol to undergo crystallisation and cholesterol monohydrate stone formation (*Admiron's triangular hypothesis*).
- Some cholesterol remains as bilayered lipid vesicles which are soluble. A specific heat labile glycoprotein in bile induces cholesterol monohydrate crystal formation in the vesicle and causes their aggregation. It is called as *nucleation*.
- Eventual precipitation and stone formation occurs by infection/infestation; pancreatic fluid reflux into CBD causing conversion of toxic lecithin to lysolecithin which is also toxic (causes supersaturated bile); bile stasis or altered enterohepatic circulation.
- Any condition which either increases the cholesterol secretion in the bile or reduces the bile salt concentration causes cholesterol stone formation. Old age; OCP; obesity; clofibrate may increase cholesterol secretion. Oestrogen, ileal resection and cholestyramine reduce the bile salt concentration.
- Chenodeoxycholic acid and ursodeoxycholic acid prevent cholesterol stone formation by maintaining bile acid pool; reducing cholesterol synthesis and secretion; converting supersaturated bile into normal bile.

Factors altering the cholesterol to bile salt ratio

- Obesity
- Drugs
  - Oral contraceptive pills
  - Clofibrate
  - Cholestyramine
- Ileal disease
- Ileal resection
- Altered enterohepatic circulation

II. ***Infections and Infestations:***
Bacteria like *E. coli, Salmonella*
Parasites like *Clonarchis sinensis* and *Ascaris lumbricoides* are often associated.
Moynihan's aphorism: "A gallstone is a tomb stone erected to the memory of the organism within it."

III. ***Bile stasis:*** Occurs due to estrogen therapy, pregnancy, vagotomy and in patients who are on long term intravenous fluids or TPN.

IV. ***Increased bilirubin production*** due to any of the causes of haemolysis as in hereditary spherocytosis, sickle cell anaemia, thalassaemia, malaria, cirrhosis. Here pigment stones are common.

Saint's triad

- Gallstones
- Diverticulosis of the colon
- Hiatus hernia

- Rarely centre of the stone contains radiolucent gas which is either triradiate (*Mercedes Benz sign*) or biradiate *(Seagull sign*).

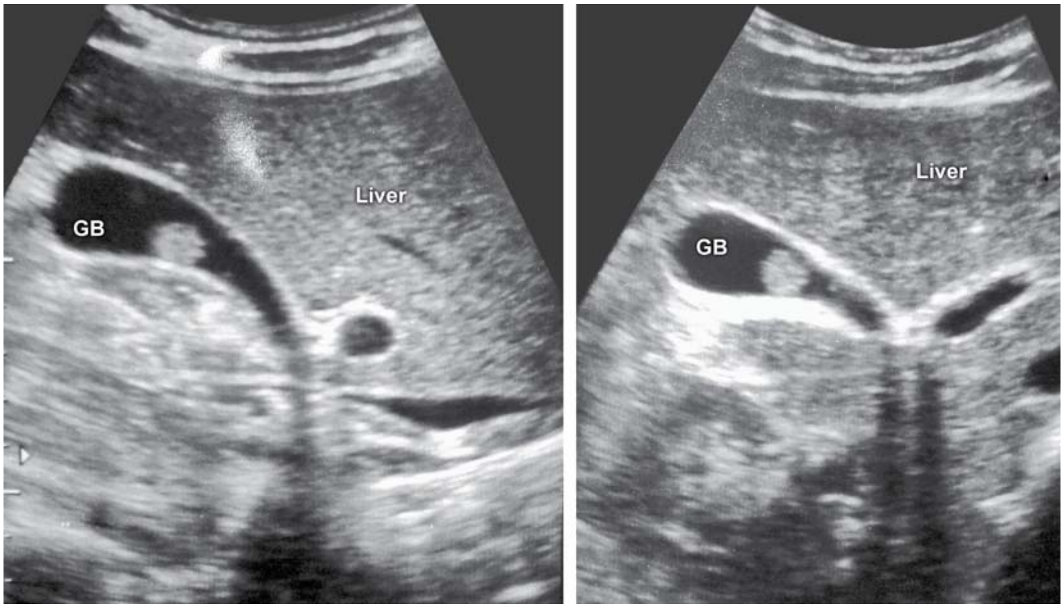

Fig. 12.29: Ultrasound gallbladder showing echogenic lesion. US should be done with change of position to find out movement of the lesion with posterior acoustic shadow to say it as gallstone. Otherwise it will be gallbladder polyp or sludge ball.

- Sometimes gallbladder may be filled by 'toothpaste like' material which is a mixture of calcium carbonate and phosphate, which on plain X-ray looks like an opacified gallbladder, so called *as limey gallbladder*.
- Only 10% of gallstones are radio-opaque, 90% are radiolucent.

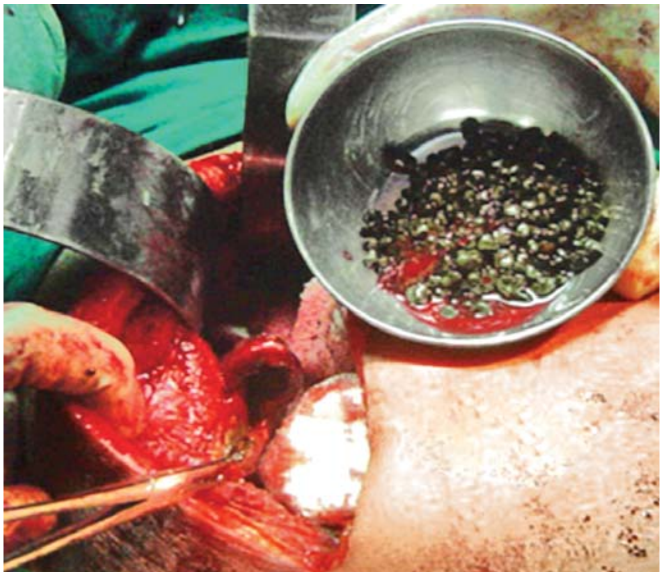

Fig. 12.30: Multiple gallstones collected during cholecystectomy.

***Note*:**

- Black pigment stones are common in gallbladder. It is usually calcium bilirubinate, calcium phosphate and bicarbonate stone with a matrix. It is common in haemolytic disorders. They are usually multiple, small black and hard in consistency. Mucin A and mucin C5 secreted by biliary glands may be the aetiology. Cholesterol component here is less than 30%. It is often seen in cirrhosis. They almost always form in gallbladder. They are common in Asia and Japan.
- Brown pigment stones are formed in biliary tree as primary biliary stones. It is commonly due to infection like *Escherichia coli* and bacteroides (98%) with bacterial nidus at the centre (often *Ascaris lumbricoides* or *Clonorchis sinensis* infestation or foreign body or stents). They secrete β glucuronidase to cause hydrolysis of soluble conjugated bilirubin to insoluble calcium bilirubinate. It also contains calcium palmitate, calcium stearate and cholesterol. They are brownish yellow, soft and mushy.

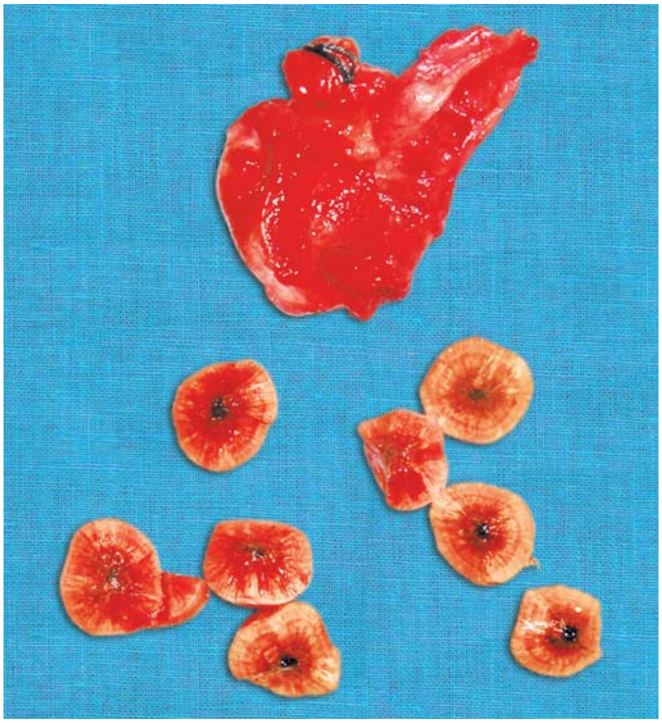

Fig. 12.31: Gallstones removed, on cutting showing central nidus.

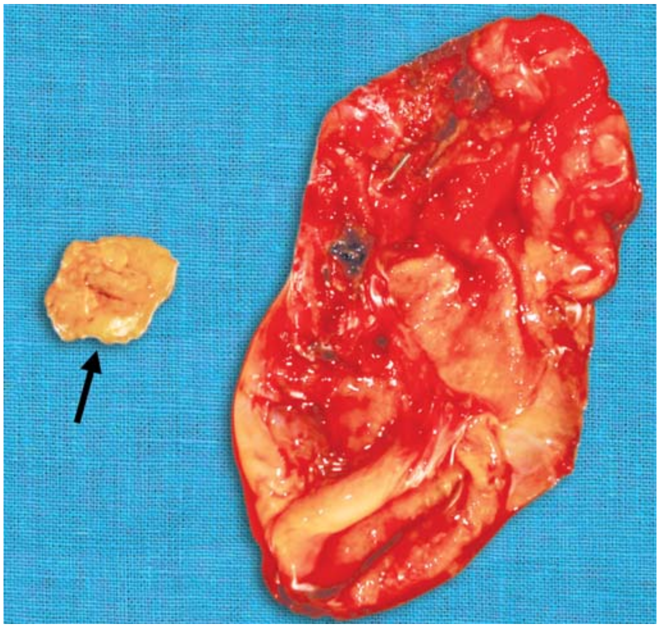

Fig. 12.32: Thickened gallbladder due to chronic cholecystitis with single cholesterol stone.

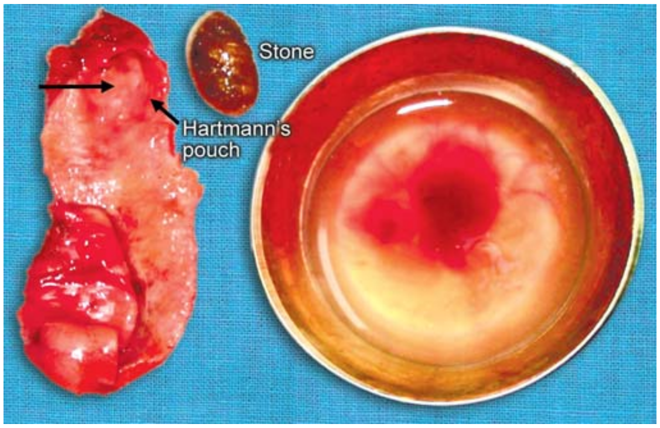

Fig. 12.33: Gallbladder with stone removed from Hartmann's pouch.

Effects of the Gallstones

a. ***In the gallbladder***

   i. Silent asymptomatic stones occurs in 10% of males and 20% of females.

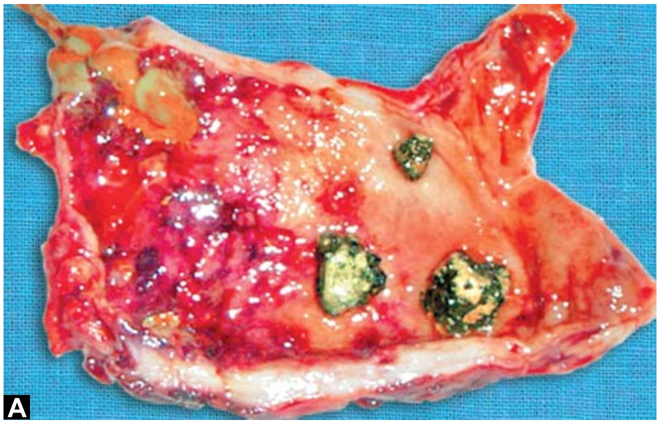

Fig. 12.34A

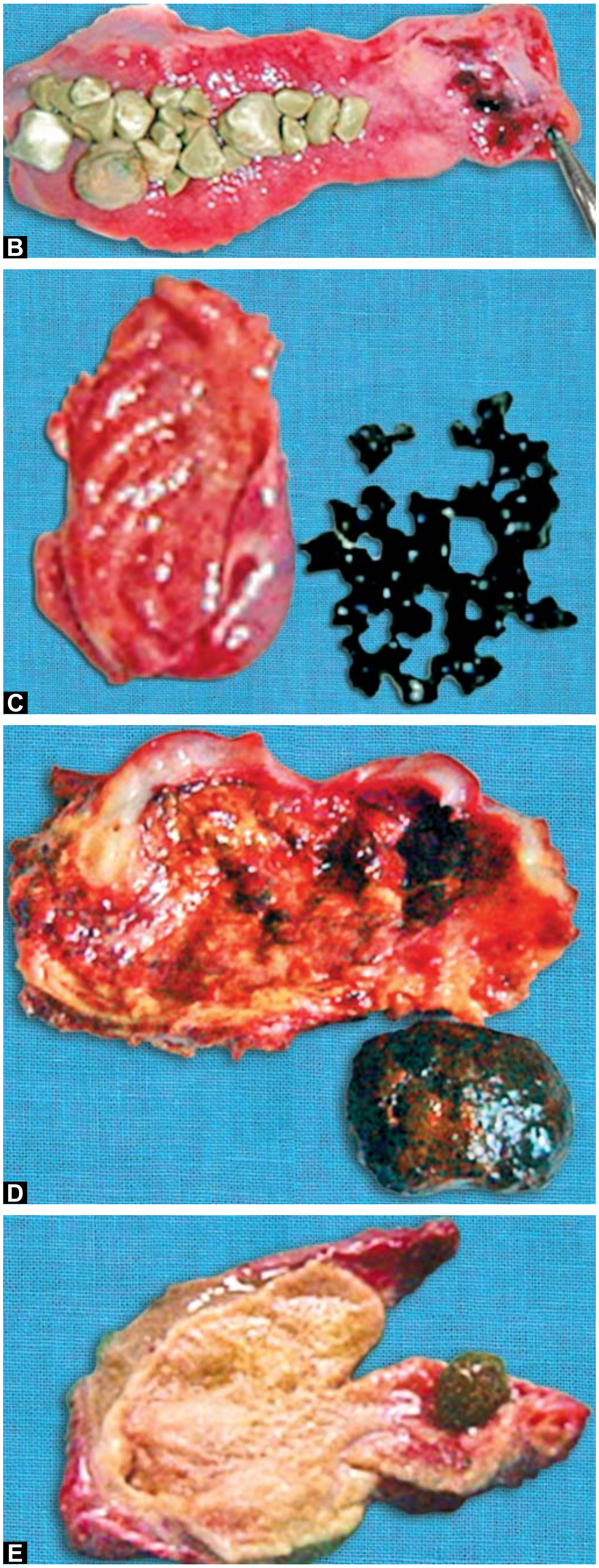

Figs 12.34B to E

Figs 12.34A to E: Different types of gallstones. Note the faceted stones; black stones; brown stones; green stones. Thickening of gallbladder is seen. Hartmann's pouch is obvious in one of the photos (with cholesteroses).

ii. *Biliary colic with periodicity*, severe within hours after meal (commonest presentation). Biliary colic is spasmodic pain often severe, in right upper quadrant and epigastrium radiating to chest, upper back and shoulder. It is self-limiting, recurs unpredictably, often precipitated by a fatty/heavy meal. Fever and increased WBC count may be observed.

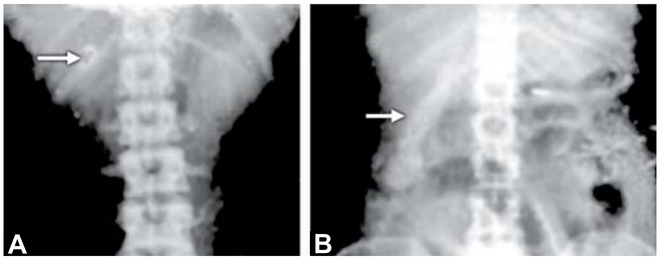

Figs 12.35A and B: Plain X-ray showing (A) Mercedes Benz sign, (B) Multiple faceted stones.

iii. Acute cholecystitis.
iv. Chronic cholecystitis.
v. Empyema gallbladder.
vi. Perforation causing biliary peritonitis or pericholecystitic abscess.
vii. Mucocele of gallbladder.
viii. Limey gallbladder.
ix. Carcinoma gallbladder.

b. ***In the CBD***

x. Secondary CBD stones (occurs in 10% of gallstones).
xi. Cholangitis.
xii. Pancreatitis.
xiii. *Mirizzi syndrome* (compression of CBD by stone from cystic duct or cholecysto-choledochal fistula).

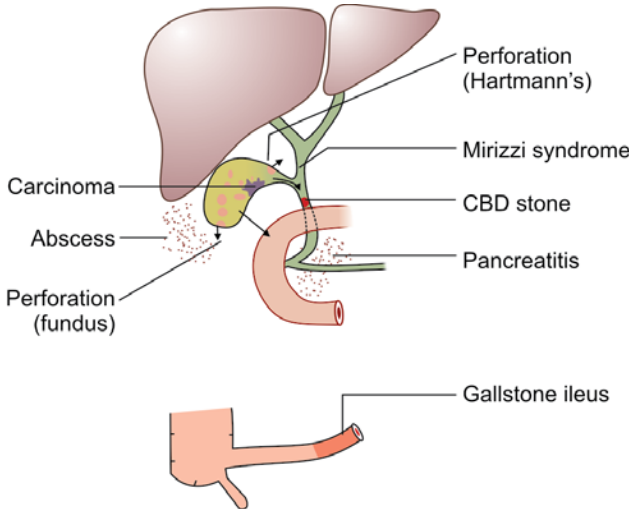

Fig. 12.36: Diagram showing complications of gallstones.

c. ***In the intestine***

xiv. Cholecystoduodenal fistula causing *gallstone ileus* and so intestinal obstruction.

Flatulent Dyspepsia

- It is discomfort in the abdomen, belching, heartburn, fat intolerance, sensation of fullness in the abdomen usually observed in *fatty, fertile, flatulent female.*

Gallstone Colic

- It is sudden, severe colicky abdominal pain in right upper quadrant which radiates to back and shoulder. This pain is due to sudden spasm of gallbladder wall when gallstone moves towards the neck of the gallbladder or cystic duct and gets impacted. Tachycardia and restlessness are common. Right hypochondrium is tender.
- It is precipitated by supine position while sleeping at night. It lasts for few hours and is episodic. It may precipitate acute cholecystitis or empyema gallbladder.
- There is reflex pylorospasm causing vomiting.

Mirizzi Syndrome (Physician from Argentina, 1948)

- In Mirizzi syndrome, gallstone impacts in the gallbladder wall and compresses it causing pressure necrosis which further gets adherent to CBD wall. It eventually causes compression and later occasionally leads into cholecystocholedochal fistula. It occurs either from Hartmann's pouch into CBD (common) or from fundus of gallbladder into the CBD.
- Mirizzi syndrome is suspected on CT scan, but usually identified on table. It needs cholecystectomy; on table cholangiogram; and exploration of CBD. It often needs choledochojejunostomy.
- Partial cholecystectomy with primary closure of CBD is done with a T tube insertion through a separate choledochotomy in type 2.
- Partial cholecystectomy with closure of gallbladder flaps is done with T tube insertion through a separate choledochotomy in type 3.
- Partial cholecystectomy with duodenal/jejunal anastomosis is done in type 4 is done in few centres.
- Open approach is ideal for Mirizzi even though laparoscopic approach is done in few centres.

Mirizzi syndrome

Type I	:	Compression of CBD without lumen narrowing
Type II	:	Compression causing CBD lumen narrowing
Type III	:	Compression causing CBD wall necrosis.
Type IV	:	Cholecysto-choledochal fistula

Silent gallstone

- *Asymptomatic stone in the gallbladder*
- Usually it is cholesterol stone, often single
- It is accidentally discovered by U/S
- It need not be treated unless:
  - Patient is diabetic/immunosuppressed
  - High chances of developing gallbladder carcinoma
  - Stone more than 2.5 cm/multiple stones
  - If gallbladder wall is thickened
  - If there is high risk for carcinoma GB

Surgery is to be done, not to have.

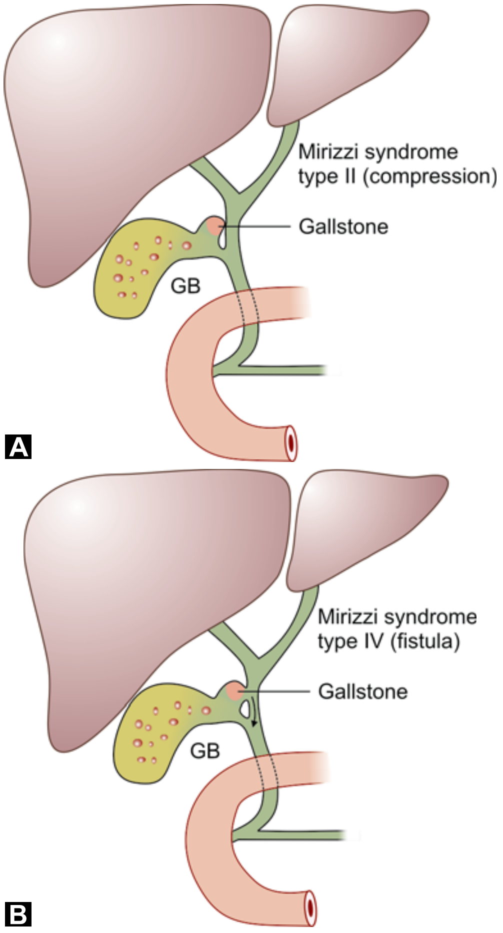

Figs 12.37A and B: Mirizzi syndrome type II and IV.

Management of Gallstones

- U/S abdomen (gallstones are seen with posterior acoustic shadowing); plain X-ray abdomen; LFT; total WBC count.
- Laparoscopic cholecystectomy ideal.
- Open cholecystectomy is done through right subcostal Kocher's incision. Open approach is used if patient is not fit for laparoscopic surgery (anaesthesia); in suspected CBD stones; Mirizzi syndrome, suspected carcinoma gallbladder. During laparoscopic cholecystectomy, if there is on table difficulty in dissection at Calot's triangle, then conversion into open approach may be required.
- Dissolution therapy for asymptomatic cholesterol stones using ursodeoxycholic acid can be tried. It is not very successful.

Differential diagnosis of radio-opaque shadow in X-ray

- Kidney stone
- Gall stone
- Calcified 12th rib tip
- Phlebolith
- Pancreatic stone
- Radio-opaque foreign body
- Faecolith
- Calcified lymph node
- Calcified renal tuberculosis
- Renal cell carcinoma—calcifications
- Adrenal tumour—calcification
- Teratomatous dermoid
- Calcification of atheroma in aorta
- Calcified lesion in liver—amoebic liver abscess/calcified hydatid cyst

ACUTE CHOLECYSTITIS

- Commonly it occurs in a patient with pre-existing chronic cholecystitis but often also can occur as a first presentation.
- Usual cause is impacted gallstone in the Hartmann's pouch, obstructing cystic duct.

Causative bacteria are:

- *E. coli*—most common
- *Klebsiella, pseudomonas, proteus*
- *Strep. faecalis*
- *Salmonella*
- *Clostridium welchii*

Classification

1. Acute calculous cholecystitis.
2. Acute acalculous cholecystitis.

Mode of Infection

- Haematogenous through hepatic artery—cystic artery.
- Portal vein.
- Through bile after filtering in the liver via portal circulation.

Pathogenesis of Acute Cholecystitis

- Stone causes obstruction at Hartmann's pouch or in cystic duct. Obstruction causes stasis, oedema of the wall, bacterial infection, acute cholecystitis and its effects.
- Impacted stone also causes mucosal erosion allowing bile salts to act over the submucosal tissues as bile is toxic to these tissues. It leads into necrosis, further infection and often perforation of the gallbladder usually at Hartmann's pouch.

Pathology of Acute Cholecystitis

- Gallbladder will be distended with oedematous friable wall. Wall contains dilated vessels.
- Areas of necrosis and patchy gangrene may occur in severe cases.
- Mucosa shows ulceration and necrosis.
- Lumen contains infected fluid/infected bile or frank pus.
- Histology shows features of acute inflammation with neutrophils, oedema, and areas of necrosis and cell death.

Complications of Acute Cholecystitis

Acute cholecystitis can lead to:

1. *Perforation*, which usually occurs in the fundus or in the neck (Hartmann's). It can cause cholecystoduodenal, cholecystointestinal or cholecystobiliary fistula.
2. *Peritonitis.*
3. Pericholecystitic abscess, empyema GB.

4. Cholangitis and septicaemia.
5. Empyema gallbladder, gangrenous gallbladder.

Emphysematous cholecystitis

- Seen in elderly male patients
- Common in diabetic and immunosuppressed individuals
- *Clostridium welchii* is the causative organism
- Gas is seen in the gallbladder
- Results in life-threatening septicaemia
- It causes severe fulminant cholecystitis
- Gangrene, perforation and peritonitis are common
- Emergency cholecystectomy is needed
- Absence of stones—observed in more than 50% of cases.

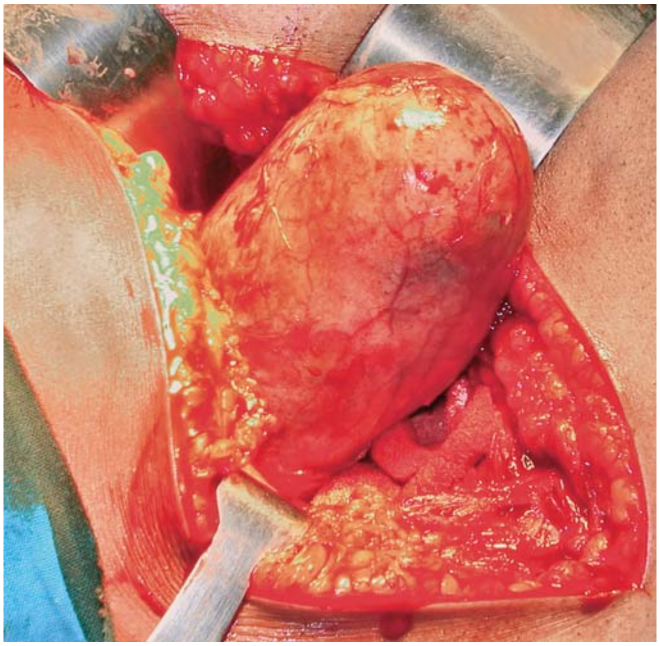

Fig. 12.38: On table look of gallbladder with acute cholecystitis.

Clinical Features

- Sudden onset of pain in the right hypochondrium, with tenderness, guarding, and rigidity.
- Palpable, tender, smooth, soft gallbladder.
- Area of hyperaesthesia between 9th and 11th ribs posteriorly on the right side (*Boas's sign)*.
- Jaundice may be present**.**
- Fever, nausea, palpable tender mass in GB region (25%).
- Tachycardia and toxic features.

Investigations

- Ultrasound abdomen—very useful, reveals presence or absence of gallstones; and thickening of gallbladder wall.
- Plain X-ray abdomen—10% of gallstones are radio-opaque; also rules out other causes of acute pain abdomen. Gas is seen in emphysematous GB.
- Total count shows neutrophilia.
- HIDA/PIPIDA radioisotope study—very useful. Non-visualisation of gallbladder is diagnostic.
- LFT is important. Increased serum bilirubin often signifies cholangitis or stone in the CBD.

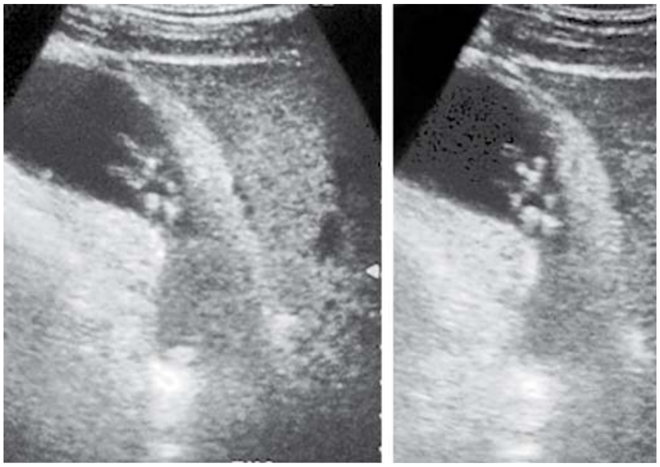

Fig. 12.39: Ultrasound picture of visualisation of multiple gallstones.

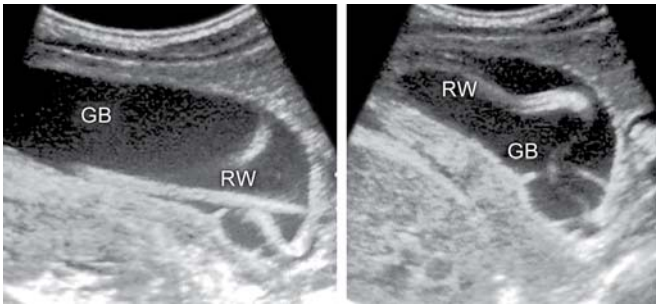

Fig. 12.40: Ultrasound showing roundworm in gallbladder.

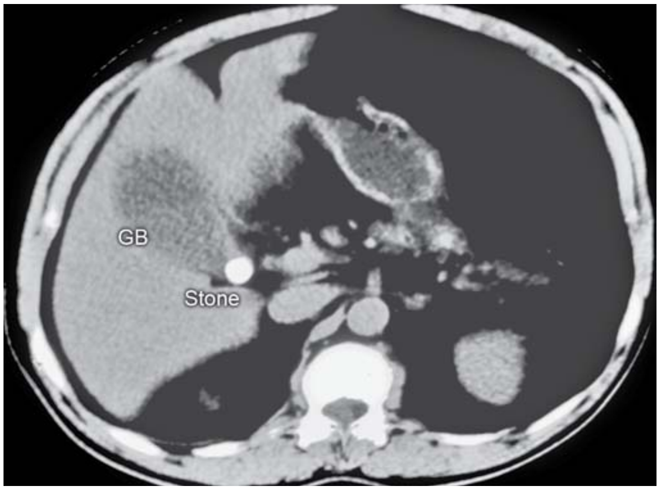

Fig. 12.41: CT scan showing stone in the neck of the gallbladder.

- Plain X-ray abdomen is not very relevant but is often important to rule out duodenal ulcer perforation, peritonitis. Only 10% of gallstones are radio-opaque. In emphysematous cholecystitis gas shadow may be seen in the region of gallbladder. Porcelain gallbladder may be seen as opacified area in gallbladder region.

Success is the ability to go from failure to failure without losing enthusiasm.

Differential diagnosis

- ❖ Duodenal ulcer perforation
- ❖ Acute pancreatitis
- ❖ Acute appendicitis
- ❖ Acute pyelonephritis
- ❖ Lobar pneumonia, myocardial infarction
- ❖ Ruptured ectopic pregnancy

Treatment

- ♦ Advised hospitalisation.
- ♦ Initially *(nonoperative) conservative treatment (95%)*:
  - ▸ Nasogastric aspiration.
  - ▸ IV fluids.
  - ▸ Analgesics and antispasmodics.
  - ▸ Broad spectrum antibiotics (cefoperazone, ceftazidime, ceftriaxone, cefotaxime + amikacin, tobramycin + metronidazole {antimicrobial}).
  - ▸ Observation.
  - ▸ Follow-up U/S scan.
- ♦ *Later after 3-6 weeks*, *elective cholecystectomy*, either by open method through right subcostal (Kocher's) incision or through laparoscopy is done.

Cholecystostomy is done immediately if patient is having:

1. Empyema gallbladder.
2. Persisting symptoms.
3. Progressing symptoms.

Here the gallbladder is opened and all stones and pus are removed. Either a Foley's or Malecot's catheter is placed in the gallbladder and is exteriorised.

After 3 weeks, elective cholecystectomy is done.

Indications for cholecystostomy or emergency cholecystectomy (5%)

- ❖ Empyema GB
- ❖ Persisting symptoms/failure of medication
- ❖ Emphysematous cholecystitis
- ❖ Perforation/peritonitis
- ❖ Elderly

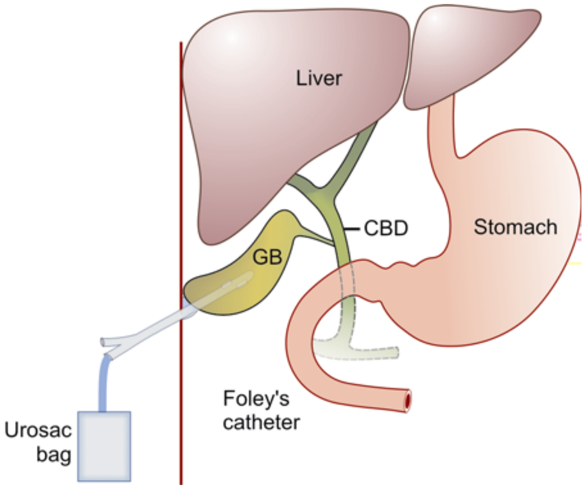

Fig. 12.42: A Foley's or Malecot's catheter is placed in severely ill-patient with acute cholecystitis to drain pus initially, and later elective cholecystectomy is done. Procedure is done in elderly, severely ill, empyema GB, emphysematous GB, perforation of GB (cholecystostomy).

Note:

- Cholecystostomy is better option than *emergency cholecystectomy* as chances of injuring adjacent structures are higher in emergency cholecystectomy. Other option is *partial/subtotal cholecystectomy* with removal of part of GB distal to the Calot's triangle.
- There is a *changing trend* now towards *early cholecystectomy* in acute cholecystitis. Laparoscopic cholecystectomy can be done in 48-72 hours in these patients. But conversion chances are high—5-6 times more than usual. Complications also may be higher due to difficulty in dissecting the Calot's triangle. It can be done in good set up with surgeons having adequate experience.
- When patient's condition is deteriorating, and acute cholecystitis is progressing, suspicious about forming empyema or necrosis and perforation; surgery should be done by abandoning conservative treatment. Usually open surgery, either cholecystectomy (may be partial) or cholecystostomy is done.

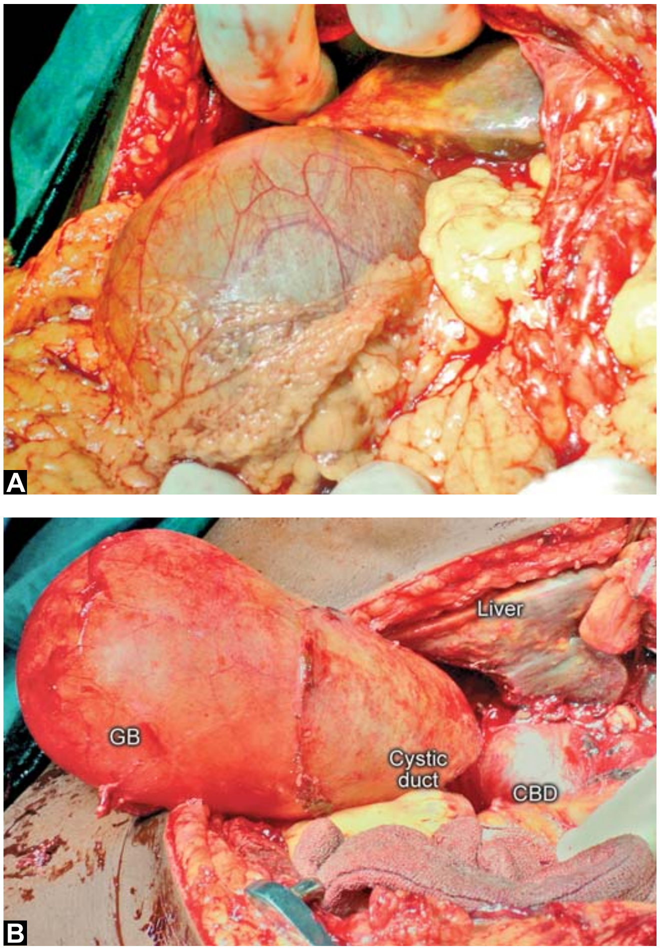

Figs 12.43A and B: On table photo of distended gallbladder. Note the Calot's triangle, liver and CBD.

ACUTE ACALCULOUS CHOLECYSTITIS (5%)

- ♦ It is not an uncommon entity, but can be commonly missed.
- ♦ It is common in patients who have undergone major surgeries, trauma, burns, or any other stress or in cases of *cholecystoses*.

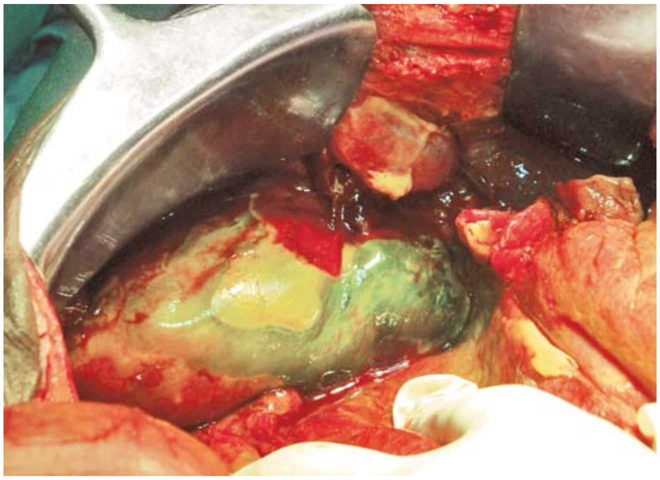

Fig. 12.44: Acalculous cholecystitis. Note the necrotic gallbladder wall.

- Common in ICU patients—in critically ill patients. It is due to bile stasis and ischaemia.
- Exact cause is not known. Gallbladder distension, release of activation factor VII may be the causes.
- Pathology is oedema and necrosis of the gallbladder wall with features of acute inflammation.
- Presentation is usually acute.
- *Investigations:* Isotope study (HIDA), U/S abdomen.
- Treatment: Cholecystectomy.

Percutaneous US guided/CT guided or open cholecystostomy initially, later cholecystectomy is the treatment of choice.

EMPYEMA GALLBLADDER

- It is a type of acute cholecystitis wherein the gallbladder is filled with pus. In 30% cases pus may be sterile.
- It also can occur in a pre-existing mucocele of the gallbladder where it gets infected.
- It is commonly observed in impacted stone, diabetic individual, immunosuppressed people like HIV, long time steroid therapy.
- It can perforate; can form abscess or can cause peritonitis—biliary and bacterial.
- Condition has got high mortality.

Clinical Features

- Fever, toxicity.
- Pain and tenderness in right hypochondrium.
- Tender, smooth, globular, gallbladder is palpable in right hypochondrium to the right of the right rectus muscle.

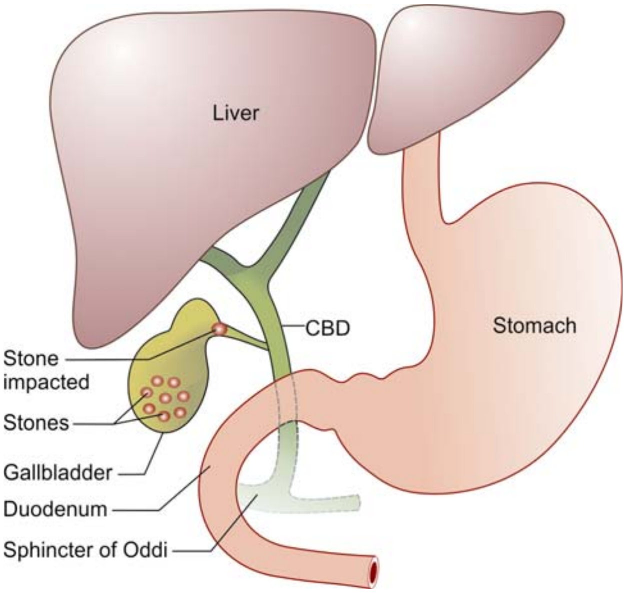

Fig. 12.45: Gallstone obstructing the neck of the gallbladder in Hartmann's pouch, may lead into empyema or mucocele of the GB.

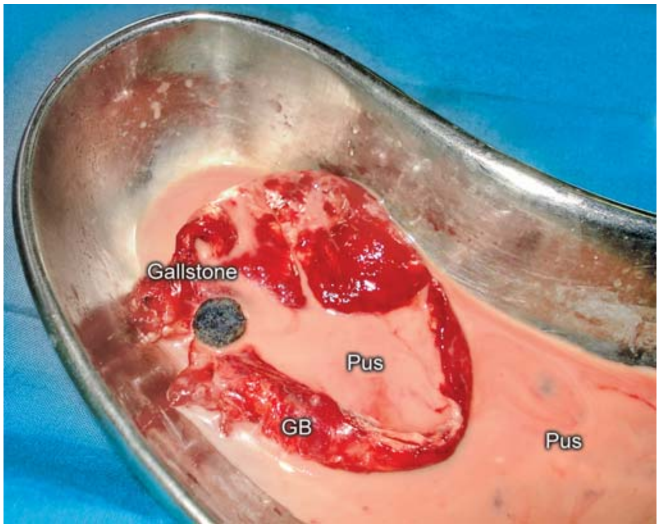

Fig. 12.46: Empyema gallbladder. Note the pus and stone.

Complications

- Septicaemia.
- Rupture and peritonitis.

Investigations

- U/S abdomen.
- Total count is raised.
- Radioisotope scan.

Treatment

- Antibiotics: Cefotaxime, quinolones, ceftriaxone.
- Cholecystectomy—an emergency procedure.

- Often initially cholecystostomy is done, with either Foley's or Malecot's catheter kept *in situ.*
  Later after 3-6 weeks, cholecystectomy is done.

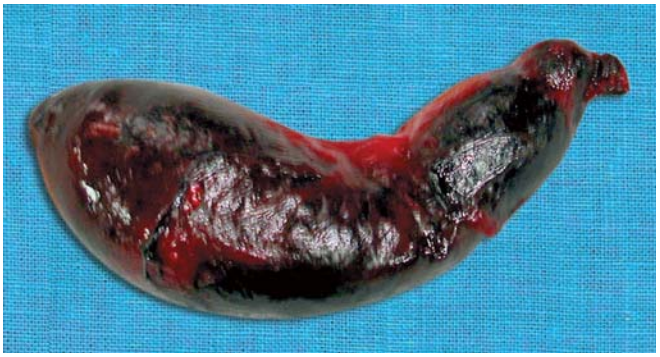

Fig. 12.47: Gangrenous gallbladder.

MUCOCELE OF THE GALLBLADDER

- It is due to obstruction of the cystic duct by a stone in the neck (Hartmann's pouch) of the gallbladder, without any infection or inflammation in the gallbladder. This causes absorption of all bile and secretion of mucous into the gallbladder allowing gallbladder to distend causing *mucocele* of the gallbladder (*Hydrops* of the gallbladder).
- Rarely cholangiocarcinoma occluding cystic duct may cause mucocele of gallbladder.
- Content is usually sterile.

Clinical Features

- Painless swelling in the right hypochondrium.
- Nontender, smooth, soft, globular, palpable gallbladder (content is sterile).
- Features of dyspepsia.

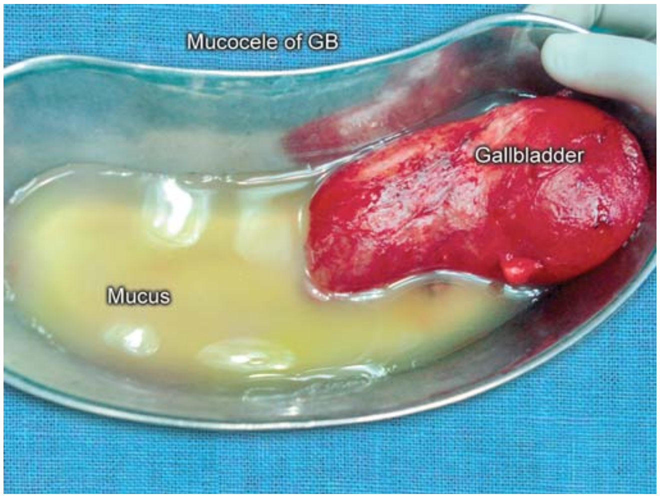

Fig. 12.48: Mucocele of gallbladder.

Complications

- If infected, can cause empyema gallbladder.
- Rupture can cause pseudomyxoma peritonei (but rare).

Investigations

- U/S abdomen.
- Liver function tests.

Treatment

Cholecystectomy, either laparoscopic or open method.

CHRONIC CHOLECYSTITIS

It is chronically inflamed, thickened gallbladder, which is nonfunctioning and nondistending.

Causes

- Gallstones.
- Cholecystoses.
- Chronic acalculous cholecystitis.

Organisms

- *Klebsiella.*
- *Streptococci.*
- *Salmonella.*

Pathology

- Gallbladder is shrunken, contracted, small, non-functioning and fibrotic with thickened gallbladder wall. Mucosa proliferates into the lumen creating deep clefts in the wall projecting into the muscular wall of the gallbladder being lined by epithelium. It is called as *Rokitansky-Ashchoff's sinuses.*
- Muscular wall is atrophied and is often replaced by fibrous tissue.
- Histologically it shows dense chronic inflammation with fibrous tissue.

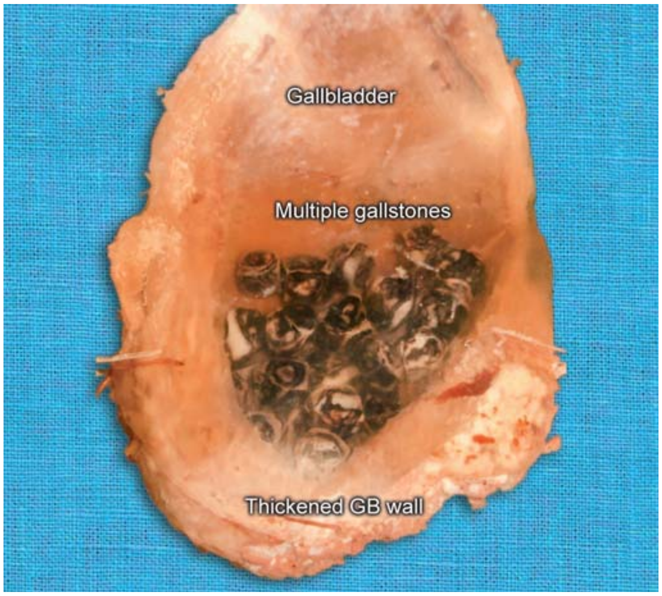

Fig. 12.49: Specimen of gallbladder showing thickened gallbladder wall with multiple gallstones. Faceting of stones is due to equal pressure.

Clinical Features

- Pain in right hypochondrium, may be colicky, or persistent.
- Positive *Murphy's sign,* where in sitting position during deep inspiration, while palpating in right hypochondrium, patient winces with pain at the summit of the inspiration. Same sign elicited in lying down position is called as *Moynihan's sign*. It may also be elicited in acute—on chronic cholecystitis.
- Flatulent dyspepsia.
- Intolerance to fatty meals.
- Biliary dyspepsia.

Complications

- CBD stone.
- Cholangitis.
- Pancreatitis.
- Mirizzi's syndrome.

Differential Diagnosis

- Peptic ulcer.
- Pancreatitis.
- Hiatus hernia.

Investigations

- U/S abdomen may show stone with posterior acoustic shadowing. Gallbladder wall will be thickened.

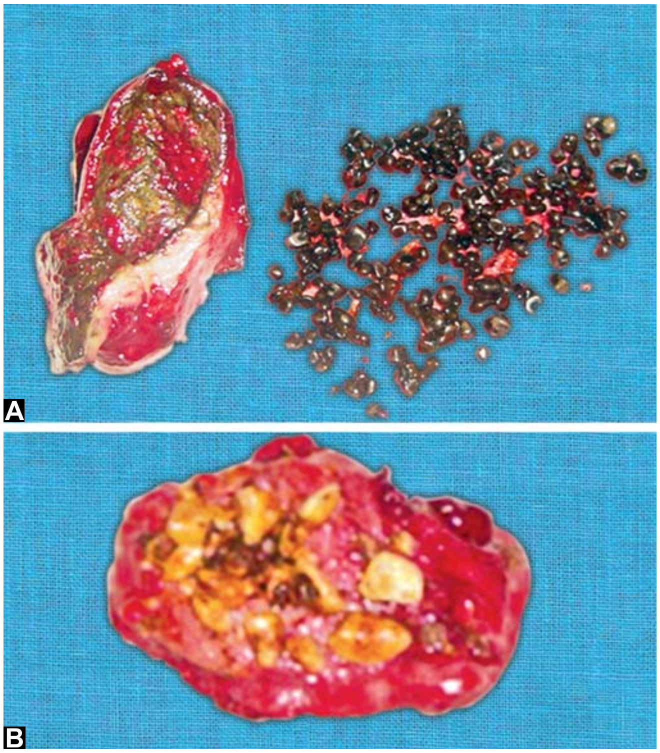

Figs 12.50A and B: Thickened GB due to chronic cholecystitis with multiple stones.

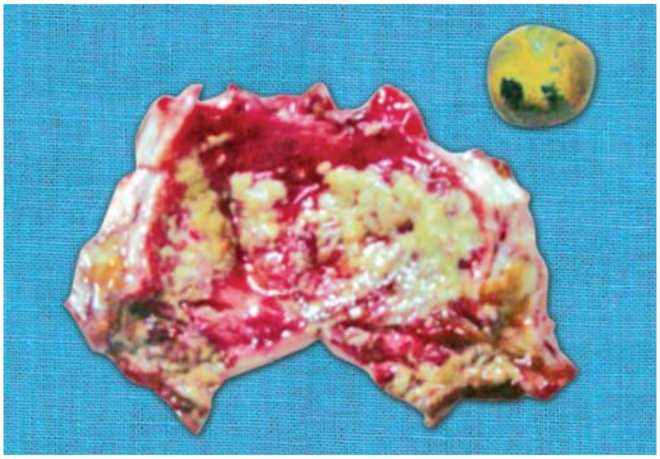

Fig. 12.51: Thick GB with single large stone.

- Isotope study may help to confirm the infection.
- LFT.
- Total count may be raised if there is an acute recurrent infection.

Treatment

Cholecystectomy (Laparoscopic or open method).

CHOLECYSTOSES

- Cholecystoses are chronic inflammatory conditions of gallbladder with cholesterol deposits.
- It is due to defective transport of the absorbed cholesterol which accumulates in mucosa. There is also increased absorption of cholesterol by gallbladder epithelium.
- Cholecystoses gallbladder is more prone for infection.
- It may precipitate stone formation.
- It is a premalignant condition.

Types

I. Aggregations of cholesterol crystals in the mucosa or submucosa—*cholesterosis* (*Strawberry gallbladder*). Lipoid contents are present in large foamy cells which has phagocytosed cholesterol. Here cystic duct is normal. Disease occurs only in gallbladder. It is a premalignant condition.
II. Cholesterol laden polypoid projections in the mucosa—*cholesterol polyposis* (Gallbladder polyp).
III. Granulomatous thickening and hyperplasia of the gallbladder—*cholecystitis glandularis proliferans.*
IV. Diverticula formation in the wall of the gallbladder—*diverticulosis of gallbladder*.
V. Gallbladder wall fistula.

Clinical Features

- Features of acalculous cholecystitis like dyspepsia, positive Murphy's sign.

***Typhoid Mary:** Salmonella typhi infection with chronic cholecystitis acting as a carrier.*
Mary was a cook who spread typhoid to many people.

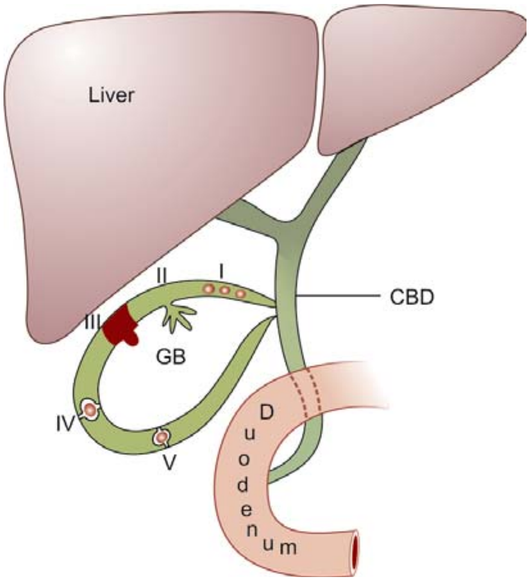

Fig. 12.52: Types of cholecystoses. (I) Cholesterosis, (II) Cholesterol polyposis, (III) Cholecystitis glandularis proliferans, (IV) Diverticulosis of gallbladder, (V) Fistula.

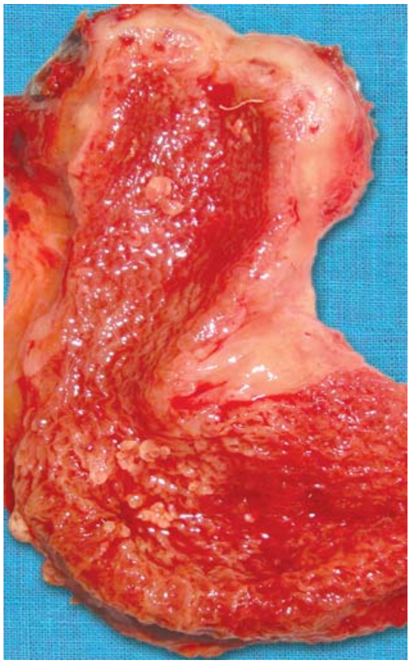

Fig. 12.53: Cholesterosis of gallbladder. It is a premalignant condition.

Investigations

- U/S abdomen.
- OCG.
- Isotope study.

Treatment

Cholecystectomy—always should be done whether it is symptomatic or not as it is potentially malignant.

DISSOLUTION THERAPY FOR GALLSTONES

Indications

- Functioning gallbladder with cholesterol stone.
- Single stone less than 1.5 cm.
- Radiolucent stone.
- Old age.
- Patients who are not fit for surgery.

Contraindications

- Nonfunctioning gallbladder
- Stone more than 1.5 cm
- Radio-opaque stone
- Multiple stones

Drugs Used

- Chenodeoxycholic acid (for 2 years).
- Ursodeoxycholic acid (15 mg/kg/day).

They inhibit absorption of cholesterol from the gut and synthesis of cholesterol in the liver. They inhibit HMG CoA—a rate limiting step in cholesterol synthesis. Ursodeoxycholic acid also inhibits absorption of cholesterol in GIT.

Other methods used

- Citrate
- Monoterpenes
- Percutaneous infusion of methyl-tertiary butyl ether (MTBE) into the gallbladder using a catheter
- Extracorporeal shock wave lithotripsy (ESWL)—not popular

Problems with Dissolution Therapy

- Drugs should be given for a long time.
- Results are not good.
- Expensive.
- Causes side effects like diarrhoea, pruritus.
- Hepatic dysfunction.

Overall results are not good by dissolution therapy.

CHOLEDOCHOLITHIASIS

It is stones in the CBD and biliary tree.

Classification

i. Primary—Rare—brown pigment stones.
ii. Secondary—Common—black pigment stones/cholesterol stones. It is seen in 15% of gallstone disease; 75% are cholesterol stones, 15% are pigment stones.

Primary stones

They are formed in CBD and biliary tree itself, and are multiple, often sludge like, commonly pigment or mixed type, extends into hepatic ducts (Brown pigment stones).

Causes:

1. Defective pathophysiology of biliary tree causing stasis, biliary dyskinesia
2. Congenital conditions like Caroli's disease, choledochal cyst.
3. Infections and infestations like clonorchiasis, ascariasis.
4. Others: Low protein diet, malnutrition, obesity, females, old age.

Secondary biliary stones

- They are from gallbladder (gallstones) passes through cystic duct to CBD. Here CBD and biliary tree are otherwise normal.
- *Secondary stones are better and easier to manage than primary stones.*
- Commonly gallstones get impacted in supraduodenal portion of CBD.

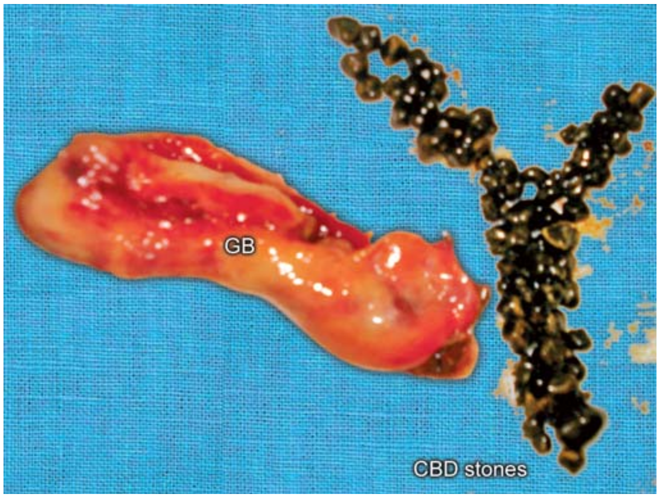

Fig. 12.54: Multiple stones removed from common bile duct and hepatic ducts. They cause cholangitis. After stone removal patient needs choledochoduodenostomy/sphincteroplasty/choledochojejunostomy.

Note:

- Black pigment stones are common in—haemolytic diseases, cirrhosis, ileal resection, fasting for long time, TPN for long time (home TPN).
- Intrahepatic stones are common—biliary sepsis, biliary stasis, parasitic infection, malnutrition.
- Retained stones means—if stones are found in CBD within 2 years of cholecystectomy.
- Recurrent stones means—if stones are found in CBD 2 years after cholecystectomy.

Clinical Features

- Incidental CBD stones along with jaundice/without jaundice.
- Pain: It may be biliary colic; nonspecific abdominal pain; pain of ascending cholangitis, pain of pancreatitis.
- Jaundice—most common clinical manifestation.
- Fever with chills and rigors.

Charcot's triad of ascending cholangitis

- Intermittent pain (may be colicky)
- Intermittent fever
- Intermittent jaundice

Stone moves proximally and floats, obstruction is relieved and symptoms subside (**Intermittent features**).

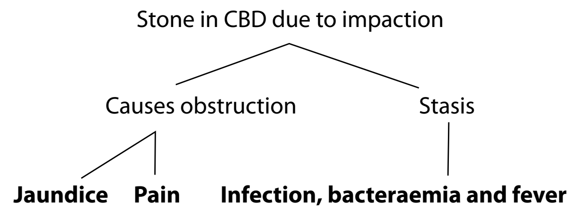
Stone in CBD due to impaction
Causes obstruction
Stasis
Jaundice
Pain
Infection, bacteraemia and fever

- ***Reynold's pentad*** *of acute obstructive cholangitis (suppurative cholangitis).*

Sometimes obstruction persists causing

- Persistent pain
- Persistent fever
- Persistent jaundice
- Shock (toxicity) and
- Altered mental status

Here biliary tree contains pus.

- Pain and tenderness in epigastrium and right hypochondrium.
- Steatorrhoea and darkening of urine.
- Pruritus.

Complications

- Liver dysfunction and biliary cirrhosis.
- White bile formation and liver failure.
- Suppurative cholangitis.
- Liver abscess.
- Septicaemia.
- Pancreatitis if CBD stone is near sphincter of Oddi blocking drainage of bile and pancreatic duct.

Differential diagnosis

Obstructive jaundice due to other causes:

- Carcinoma of head of pancreas
- Carcinoma of periampullary region
- Carcinoma of biliary tree
- Biliary stricture
- Viral hepatitis

Investigations

- U/S abdomen; CBD diameter > 1 cm indicates biliary obstruction.
- ERCP, gold standard for diagnosis.
- MRCP.
- Liver function tests.
- Endoultrasonography is useful (*EUS*). It is more accurate.
- PTC if only indicated.

Treatment

- Injection Vit. K 10 mg IM once a day for 5 days or FFP infusion to correct the prothrombin time.

Commitment is what transforms a promise into reality.

- IV antibiotics (cefoperazone, cefotaxime).
- *Correction of dehydration*.
- IV Mannitol daily 200 ml BD to prevent *hepatorenal syndrome*.
- **ERCP—Therapeutic,** i.e. endoscopic papillotomy (sphincterotomy) and stone extraction through Dormia basket or balloon catheter; or fragmenting the stone and extraction; or removal through baby endoscope. CBD stent is placed *in situ*.

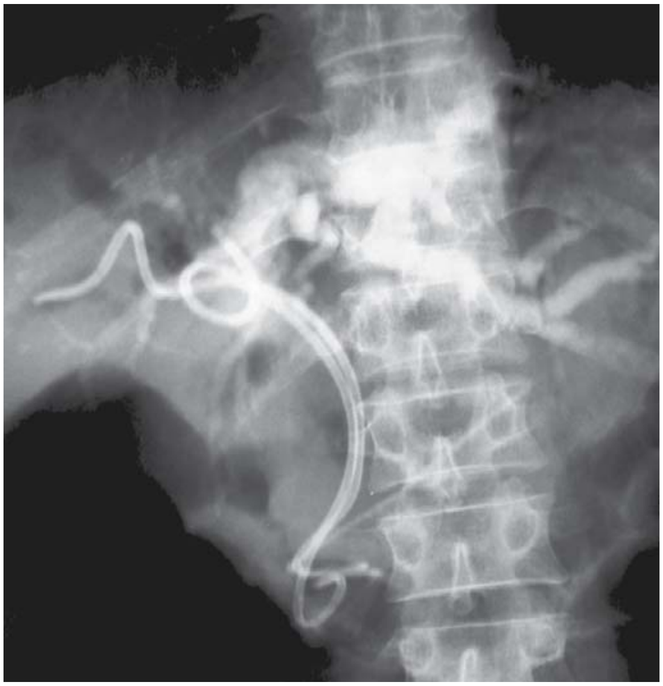

Fig. 12.55: Plain X-ray showing stents in CBD and pancreatic ducts. It was placed in a patient who presented with recurrent pancreatitis with block/stricture in both terminal CBD and pancreatic duct. Later patient underwent choledochojejunostomy and pancreaticojejunostomy.

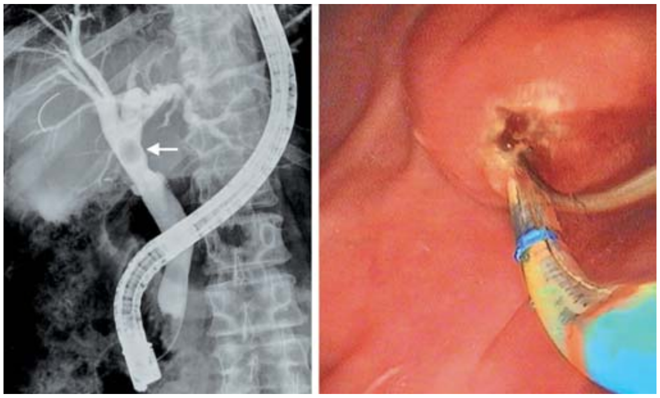

Fig. 12.56: ERCP showing CBD stones. Endoscopic view of stenting is also seen.

Complications of sphincterotomy

- Bleeding
- Sepsis
- Perforation
- Pancreatitis

Once CBD stones are extracted through ERCP, laparoscopic cholecystectomy is done.

- In open cholecystectomy
  - After the removal of gallbladder, on table cholangiogram is done through cystic duct using water soluble iodine dye to see any stones in CBD. Using stay sutures choledochotomy is done (opened longitudinally) to remove stones in CBD.

Indications for choledochotomy

(After open cholecystectomy)

- U/S shows stone in CBD
- Palpable stone in CBD
- CBD diameter >10 mm
- Recent history of jaundice, and with raised serum alkaline phosphatase level
- On table cholangiogram shows stone (done using C-arm image machine)
- Failure of stone extraction by ERCP prior to cholecystectomy.
- When in doubt

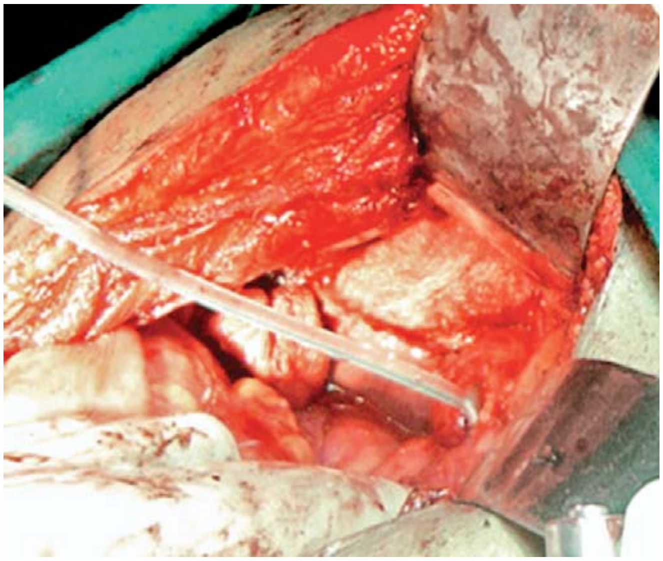

Fig. 12.57: Placing T-tube in CBD after choledocholithotomy. It is kept for 10-14 days.

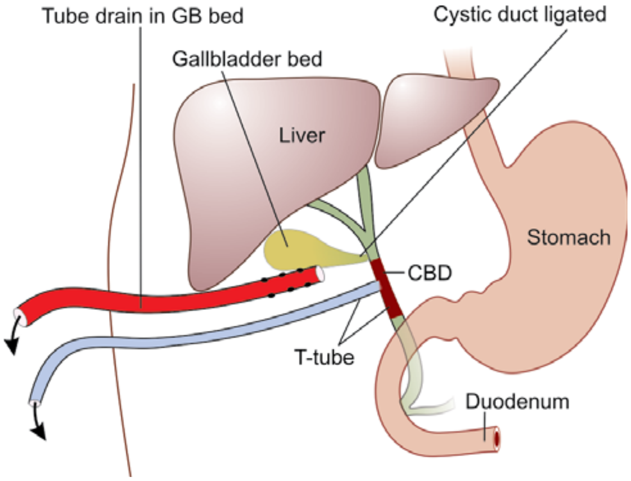

Fig. 12.58: Diagram showing placement of T-tube in CBD after choledocholithotomy. Note the separate drain placed in the cholecystectomy bed to drain fluid, blood, bile.

- After choledochotomy, stones are removed using Des jardin's choledocholithotomy forceps. Bake's CBD dilator is used to confirm the CBD patency.
- T-tube (Kehr's) is then placed in the CBD and kept for 14 days.
- After 14 days a postoperative T-tube cholangiogram is done to see for free flow of dye into the duodenum, so that T-tube can be removed.
- If T-tube cholangiogram shows persistent stone, it can be extracted after 6 weeks, through a basket (Dormia) or catheter (Fogarty) through the track or through a choledochoscope. Retained stones can also be removed through ERCP.

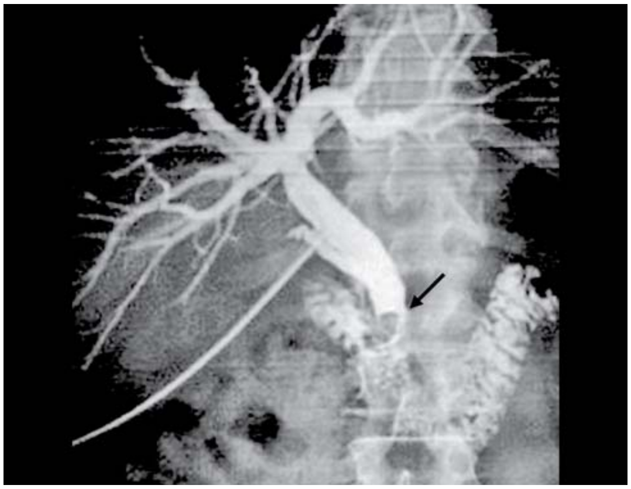

Fig. 12.59: T-tube cholangiogram showing retained stone in the CBD (*Courtesy:* Dr Navin Chandra Shetty, HOD, Department of Radiodiagnosis, KMC, Mangalore).

- After open cholecystectomy, on table ERCP can be tried if there are indications for CBD intervention case of secondary CBD stone/s and if ERCP was not done prior to cholecystectomy. If secondary stone is retrieved by this method, open exploration of CBD can be avoided.
- Laparoscopic CBD exploration is becoming popular with availability of expertise and high technology imaging and instruments.

Methods to confirm the removal of T-tube

- Clamp the T-tube, after 10-14 days and observe 48 hours for development of pain, jaundice and fever
- Confirm free flow of dye in T-tube cholangiogram

Management of retained CBD stones

- Small stones may *spontaneously* pass down.
- Heparinised saline or bile acid *flushing* through the T-tube. (Wash 250 ml of normal saline with 25,000 IV heparin for 5 days)
- Burhenne technique (Canada)—after 6 weeks once T-tube track gets matured, track if needed is dilated using graduated dilators. Either using Dormia basket or Fogarty catheter or choledochoscope, stone is removed through T-tube track under fluoroscopic guidance (C-ARM)
- *ERCP* and stone removal in 3 weeks
- *Reoperation* if everything fails, either transduodenal sphincteroplasty or choledochojejunostomy
- ESWL with endoscopic sphincterotomy/extraction/lavage/stenting
- Through percutaneous transhepatic route, cholangioscope is passed and CBD is visualised, stone is identified and removed using Dormia basket or Fogarty catheter
- Laparoscopic choledocholithotomy
- Open choledocholithotomy often with choledochojejunostomy

Treatment of Primary CBD Stones

- *Difficult and is associated with more complications.*

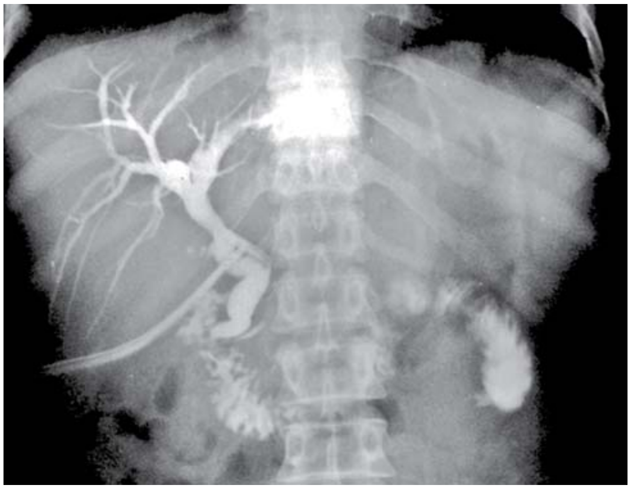

Fig. 12.60: T-tube cholangiogram showing dye in the duodenum and visualisation of normal biliary tree without any obstruction.

Preferred procedures:

- *Transduodenal sphincteroplasty (open method),* or
- *Open choledochoduodenostomy,* side-to-side, or
- *Open choledochojejunostomy*—Roux-en-Y method.
  - These procedures attain complete drainage of bile.
  - Postoperatively they may often require long-term antibiotics.

Indications for choledochoduodenostomy/choledochojejunostomy:

- Multiple CBD calculi with distal narrowing (*Funnel syndrome*).
- Papillary stenosis.
- Impacted calculi.
- Biliary sludge—symptomatic.
- Residual stones.

Prerequisite for choledochoduodenostomy:

- CBD should be more than 1.4 cm and stoma should be 2 cm.

Advantages of choledochoduodenostomy are:

- Bile leak is minimal/not there.
- Beneficial as a permanent remedy in multiple stones/sludge/stenosis/strictures/intrahepatic stones.

Problem with choledochoduodenostomy:

- Sump syndrome.

Note:

Roux-en-Y choledochojejunostomy is equally good as choledochoduodenostomy.

You are the sole master of your thought processes.

Remember

- Gallstones are commonly radiolucent (90%).
- Multiple stones are usually faceted because of exertion of equal pressure in a compact gallbladder.
- Plain X-ray shows radio-opaque lesion to the right side of the vertebra below rib cage. It should be differentiated from kidney stones. In lateral view X-ray, gallstone will be in front of the vertebra whereas kidney stone overlaps the vertebra. Often gallstone has got central radiolucent area—*sea gull sign/Mercedes Benz sign.*
- *Silent/asymptomatic gallstone* is one which is identified on routine investigation where there are no specific relevant symptoms related to gallstones. Chances of developing symptoms in a silent gallstone is 5% in 5 years and 20% in 15 years.
- Presently ultrasound is ideal investigation for gallstones.
- To see gallbladder function or confirm cholecystitis radioisotope HIDA/PIPIDA scan is ideal.
- Cholecystitis can cause jaundice due to cholangitis. But other causes of jaundice should be ruled out—CBD stones/Mirizzi syndrome.
- *Limey gallbladder* is gallbladder filled with toothpaste like mixture of calcium carbonate and calcium phosphate. Plain X-ray shows dense radio-opaque gallbladder shadow (*opacified gallbladder*).
- *Porcelain gallbladder* is one where gallbladder wall is calcified because of chronic cholecystitis. It is potentially malignant. It should be removed either by open or laparoscopic method.
- Cholesterol stone occurs when there is alteration in levels of cholesterol, lecithin and bile salts. This altered bile has got more cholesterol than adequate micelle and is called as *lithogenic bile*. Here bile is in supramicellar zone. Cholesterol stone is radiolucent but causes acoustic shadow in U/S
- Mixed stones are most common—90%.
- '*Gallstone is a tomb stone erected to the memory of the organism within it*'—Moynihan's aphorism.
- *Saint's triad*: Gallstones—colonic diverticulosis—hiatus hernia.
- *Complications of gallstones*: Acute cholecystitis, chronic cholecystitis, empyema gallbladder, mucocele of gallbladder, perforation and peritonitis, secondary CBD stones, cholangitis, pancreatitis, Mirizzi syndrome, gallstone ileus, pericholecystic abscess and carcinoma of gallbladder.
- Black pigment stones are more common in gallbladder; brown pigment stones are common in CBD.
- Cholesterol stones are common in Western countries; pigment/mixed stones are common in Asian countries.
- *Acute acalculous cholecystitis* is 5% common. It occurs after stress, major surgeries or in cholecystoses.
- *Xanthogranulomatous cholecystitis* is a rare pathological condition. Gallbladder is thickened with chronic inflammation. It is yellow xathogranulomatous in nature. It often extends to adjacent organs. It is due to reaction to penetrated bile.
- *Residual/retained bile duct stones* are one which is present in CBD within 2 years of initial surgery—cholecystectomy. They are usually missed secondary bile duct stones.
- *Recurrent bile duct stones* are one which is present 2 years after the initial surgery—cholecystectomy and CBD exploration. They are primary biliary stones.
- *Salmonella cholecystitis* (Typhoid Mary—a cook in New York transmitted typhoid through her infected faeces and urine) causes typhoid gallbladder. It can cause acute or chronic cholecystitis. Salmonella itself may predispose stone formation. Patient may be silent typhoid carrier.

SUMP SYNDROME

It is commonly observed after choledochoduodenostomy rarely after choledochojejunostomy. CBD distal to the choledochoduodenostomy acts as a reservoir with stasis of food particles, bile, stones and sludge. Often it causes cholangitis and narrowing of the stoma of choledochoduodenostomy. Conversion to end-to-end Roux-en-Y choledochojejunostomy is required.

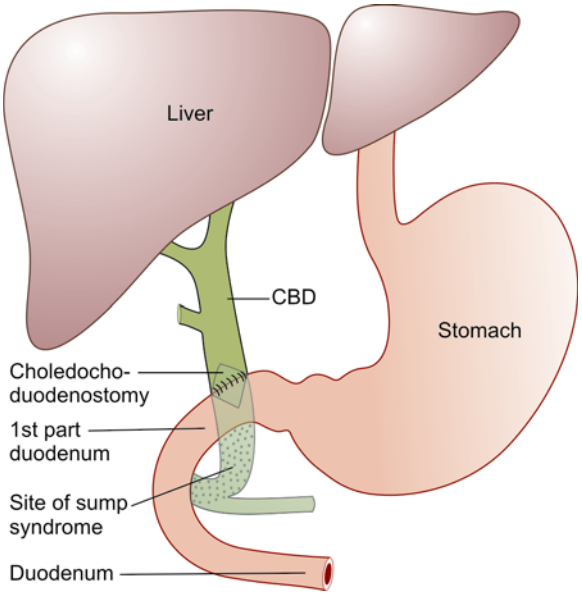

Fig. 12.61: Technique of choledochoduodenostomy for CBD stones. Note the site of possible occurrence of *sump syndrome*.

COURVOISIER'S LAW

- **'In a patient with jaundice if there is palpable gallbladder, it is not due to stones'**.

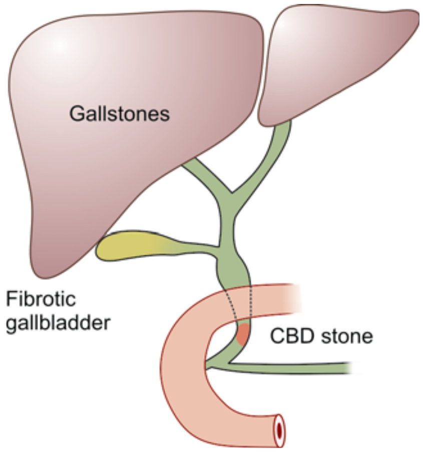

Fig. 12.62: Courvoisier's law—"In a patient with jaundice if there is palpable gallbladder it is not due to stones". In stone disease gallbladder is contracted and fibrotic and so nondistensible.

- In obstruction due to CBD stone, gallbladder does not distend because it is chronically inflamed, thickened, fibrotic, contracted and nondistensible.
- In malignancy, like carcinoma of head of the pancreas or periampullary carcinoma, gallbladder will be distended and palpable to the right of rectus muscle in the right hypochondrium, as nontender, globular, smooth, soft, dull mass which moves with respiration and with horizontal mobility.

Rule may not be useful in:
- Absence of gallbladder
- Intrahepatic gallbladder
- Previous cholecystectomy

Exceptions to the rule are:
- Double impacted stone—one in CBD and one in cystic duct, with mucocele of gallbladder.
- Large stone in Hartmann's pouch.
- Empyema gallbladder with CBD stone.

SURGICAL JAUNDICE (Obstructive Jaundice)

Definition

It is the jaundice that develops due to biliary obstruction, partial or complete or intermittent. It causes conjugated hyperbilirubinaemia. Normal serum bilirubin level is 0.2-0.8 mg%. Scleral icterus is visible when serum bilirubin level exceeds 2.5 mg%.

Effects of Obstructive Jaundice

- *In liver*: Enlarged green bile stained liver (*hydrohepatosis*) shows dilated intrahepatic biliary radicles. Once intraductal CBD pressure increases bile secretion from liver is reduced causing formation of 'white bile' in CBD. Biliary cirrhosis may develop later.
- *In the biliary tree*: Recurrent inflammation—cholangitis—fibrosis can occur.
- *In bowel*: Absence of bile from bowel impairs digestion, reduces fat absorption making faeces bulky and fatty. Vitamin K absorption is reduced causing fall in prothrombin level raising PT-INR.
- *Retention of bile salts and bile pigments* in blood and body fluids occurs.
- *Altered coagulation profile; hepatorenal syndrome and renal failure; sepsis.*

Causes

1. Biliary atresia.
2. Choledochal cyst.
3. CBD stones.
4. Ascending cholangitis.
5. Biliary strictures.
6. Sclerosing cholangitis.
7. Carcinoma of head and periampullary region of the pancreas.
8. Cholangiocarcinoma.

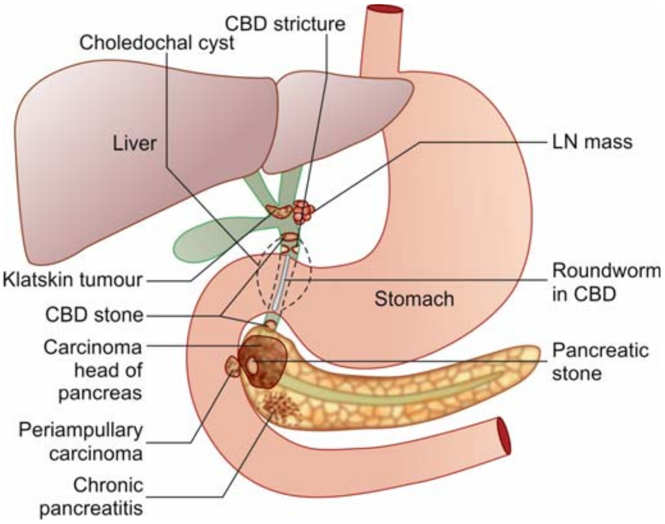

Fig. 12.63: Different causes of obstructive jaundice (surgical jaundice).

9. *Klatskin tumour* (Carcinoma at the confluence of hepatic ducts above the level of the cystic duct and so will cause hydrohepatosis without GB enlargement).

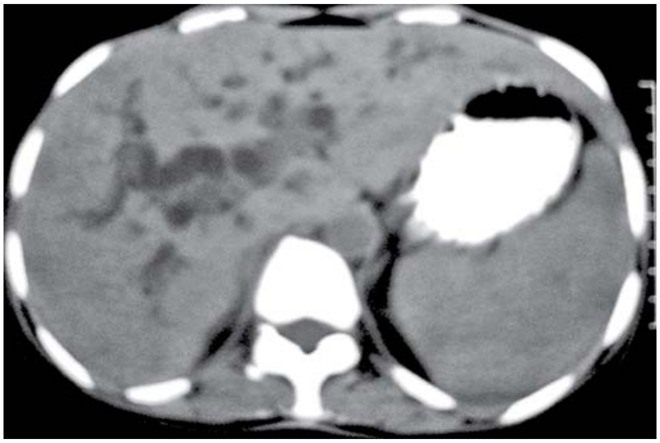

Fig. 12.64: CT scan showing hydrohepatosis due to obstruction below, in a patient with obstructive jaundice.

10. Extrinsic compression of CBD by lymph nodes or tumours.
11. Parasitic infestations.

Clinical Features

- Severe jaundice.
- Pruritus, more on the back and forearms.
- Fever, may or may not be present.
- Loss of weight.
- Loss of appetite.
- Pain in right hypochondrium, *palpable gallbladder*, hydrohepatotic palpable, smooth, soft, nontender liver are other features.
- Courvoisier's law may suggest inflammatory/neoplastic cause.
- Charcot's triad/Reynold's pentad as presentation in cholangitis.
- Steatorrhoea (more fatty stool) due to improper absorption of fat soluble vitamins.

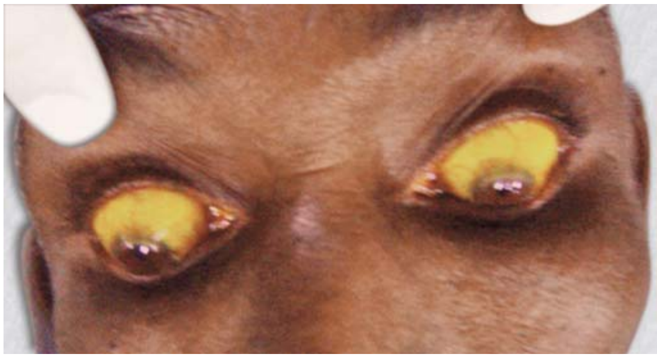

Fig. 12.65: Sclera of the patient with obstructive jaundice is greenish yellow in colour.

Classification of Causes of Obstructive Jaundice

1. *Congenital*: Biliary atresia, choledochal cyst.
2. *Inflammatory*: Ascending cholangitis, sclerosing cholangitis.
3. *Obstructive:* CBD stones, biliary stricture, parasitic infestation.
4. *Neoplastic*: Carcinoma of head or periampullary region of pancreas, cholangiocarcinomas, Klatskin tumour.
5. *Extrinsic compression* of CBD by lymph nodes or tumours.

Benjamin's (1983) classification of biliary obstruction

Type 1: Complete obstruction

- Tumours—pancreatic, cholangiocarcinoma
- CBD ligation—iatrogenic
- Primary/secondary liver tumours

Type 2: Intermittent obstruction

- Choledocholithiasis
- Periampullary tumour
- Choledochal cyst
- Bile duct papilloma
- Hemobilia
- Duodenal diverticula

Type 3: Chronic complete obstruction

- Bile duct stricture
- Congenital
- Traumatic
- Post-radiotherapy
- Chronic pancreatitis
- Cystic fibrosis

Type 4: Segmental obstruction

- Traumatic
- Sclerosing cholangitis
- Cholangiocarcinoma, intrahepatic biliary stones (hepatolithiasis)

Investigations for Obstructive Jaundice

- Serum bilirubin. Normal value is less than 1.0 mg%. Both direct and indirect bilirubin are assessed. Direct is increased in obstructive jaundice, i.e. conjugated hyperbilirubinaemia. van den Bergh's test is done.
- Serum albumin, globulin and A : G ratio. Normal S. albumin is more than 3.5 gm%.
- Prothrombin time. Normal value is 12-16 seconds. It is significant if it is more than 4 from the control or more than one and half times the control. It is corrected by injection vitamin K, 10 mg IM OD for 5 days or by FFP—5-10 units.
- Serum alkaline phosphatase, SGPT, SGOT, 5' nucleotidase.
- Total count may be raised with neutrophilia in inflammatory conditions.
- U/S abdomen.
- ERCP to visualise the site of obstruction, brush biopsy, bile sample for analysis.
- PTC to decompress, assess proximal dilated obstructed biliary system if ERCP fails; dine polythene catheter can be kept *in situ* to have biliary drainage; PTC—stenting across the obstruction can be done under image (C arm) guidance.
- MRCP—Noninvasive diagnostic tool. It shows 96% sensitivity; 99% specificity.
- CT scan in case of tumours to assess operability.
- *Tumour markers*: CA 19/9 is useful for carcinoma pancreas (more than 70 units/L) with 70% sensitivity and 90% specificity. But it may also increase in other causes of biliary obstruction and cystadenoma.
- *Endoscopic US (EUS*): It is done through endoscope. It is more accurate in assessing pancreatic mass, staging of the disease, to identify involvement of portal venous system, CBD stones. It is also useful in EUS guided FNAC, celiac axis neurolysis, EUS guided immunotherapy.
- *Intraductal US (IDUS):* It is very useful in assessing tumour stage, tumour margin in bile duct cancer. It is also used in assessing pancreatic duct to differentiate pancreatic cancer and chronic pancreatitis.
- *CT / MR angiogram or venogram* to assess vascularity and portal venous system in malignancy.
- Urine tests: Fouchet's test for bile pigments, Hay's test for bile salts and test for urobilinogen in urine.

Fouchet's test: 10 ml of urine + 5 ml of $BaCl_2$ + pinch of $MgSO_4$ causes formation of $BaSO_4$ which is filtered over a filter paper and few drops of Fouchet's reagent is added. Green or blue colour signifies presence of bile pigments in the urine.

Hay's test for bile salt: Sprinkle sulphur to 2 ml of urine. In presence of bile salts sulphur sinks to the bottom.

Ehrlich's test: 5 ml of freshly voided urine + 1 ml of Ehrlich reagent (p-dimethyl amino benzaldehyde) and wait for 5 minutes. Formation of red colour signifies presence of urobilinogen in urine. Normally it is present in traces, in obstructive jaundice it is absent and in haemolytic jaundice it is in excess.

Preoperative preparation of patient with obstructive jaundice

- Proper diagnosis and assessment
- Injection vitamin K IM 10 mg for 5 days
- Fresh Frozen plasma—often requires 6 bottles or more
- *Adequate hydration* is most important 5/10% dextrose
- Blood transfusion in case of anaemia
- Oral neomycin, lactulose
- Mannitol 100-200 ml BD IV to prevent hepatorenal syndrome

- Repeated monitoring by doing prothrombin time, electrolytes
- Antibiotics like third generation cephalosporins.
- Calcium supplements as calcium chloride IV
- *Preoperative decompression* is indicated if bilirubin is > 12 mg%, sepsis, hepatorenal syndrome, severe malnutrition or cardiopulmonary disease.
- Correction of coagulopathy, prevention of renal failure, infection, hepatic encephalopathy and electrolyte imbalance (correction of hypoglycaemia and dilutional hyponatraemia due to water retention; avoiding isotonic saline infusion).

Treatment of Obstructive Jaundice

- CBD stones—ERCP stone removal, choledocholithotomy, transduodenal sphincteroplasty, choledochojejunostomy or choledochoduodenostomy.
- Carcinoma periampullary or head of pancreas—Whipple's operation or triple bypass or ERCP stenting.
- Biliary stricture—Stenting, choledochojejunostomy, Roux-en-Y hepaticojejunostomy.
- Klatskin tumour—Radical resection or palliative stenting.
- Biliary atresia—Kasai's operation or liver transplantation.
- Choledochal cyst—Excision, hepaticojejunostomy, mucosal resection.
- *Management of pruritus:* Pruritus may be due to retention of bile salts which activates the release of histamine in skin, central mechanism or by release of endogenous opioids. It is often difficult to treat. Once cause is treated and obstruction is relieved, pruritus will regress. Drugs and therapies used are—cholestyramine (ion exchange resin binds bile salts in intestine inhibiting their absorption), rifampin, ondansetron, gabapentin, sertratine, ursodeoxycholic acid, antioxidants, phototherapy, plasmapheresis.

Postoperative Management

- Monitoring with prothrombin time, bilirubin, albumin, creatinine, electrolyte estimation.
- FFP or blood transfusion.
- Antibiotics.
- Observation for septicaemia, haemorrhage, pneumonia, pleural effusion, bile leak.
- Care of T-tube and drains.
- T-tube cholangiogram in 10-14 days.
- TPN, CVP line, nasogastric tube, urinary catheter.

CBD STRICTURES (BILIARY STRICTURES)

Causes

1. *Postoperative (80% common)*
    - After cholecystectomy [open or laparoscopic, more common following laparoscopic (0.8%) than open method (0.35%)].
    - Biliary surgery.
    - Gastrectomy.
    - Liver surgery.
    - Duodenal and pancreatic surgery.

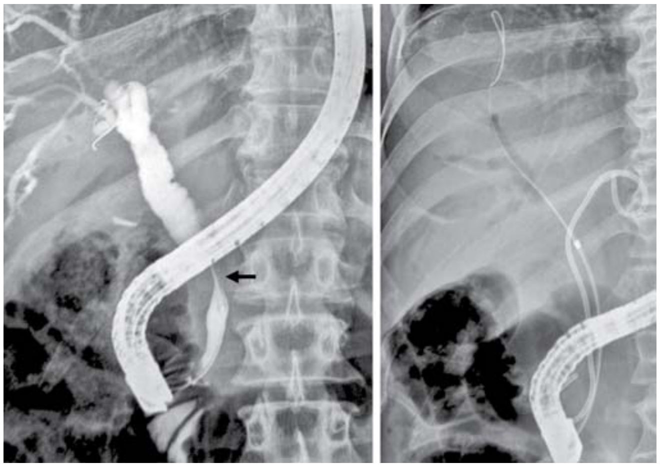

Fig. 12.66: ERCP picture showing stricture and passing stent into CBD.

2. *Inflammatory:* Stricture following recurrent attacks of cholangitis due to:
    - CBD stones.
    - Parasites—*Ascaris lumbricoids, Clonorchis sinensis.*
    - Primary sclerosing cholangitis.
3. *Malignant:* Due to cholangiocarcinoma.
4. *Traumatic.*

Bismuth classification of biliary stricture

I. Low CBD stricture with stump (proximal to stricture) > 2 cm
II. Middle CBD stricture with proximal stump < 2 cm
III. Hilar. Confluence of right and left hepatic ducts is intact
IV. Separated right and left hepatic ducts
V. Stricture involving intrahepatic ducts

Clinical Features

- Obstructive jaundice.
- Pain abdomen.
- Features of ascending cholangitis.
- Profuse persistent bile leak.

Causes of postoperative stricture

- Too much traction to the GB during ' Fundus first cholecystectomy'
- Blind application of artery forceps in Calot's to control bleeding
- Anomalies of biliary tree
- Inflammatory adhesions at Calot's triangle
- Injury during other surgery like gastrectomy

Note:
Hogarth Pringle's manoeuvre is ideal to control bleeding on table by applying pressure using thumb and fingers over the Foramen Winslow.

The mother's heart is child's schoolroom.—**Henry Ward Beecher.**

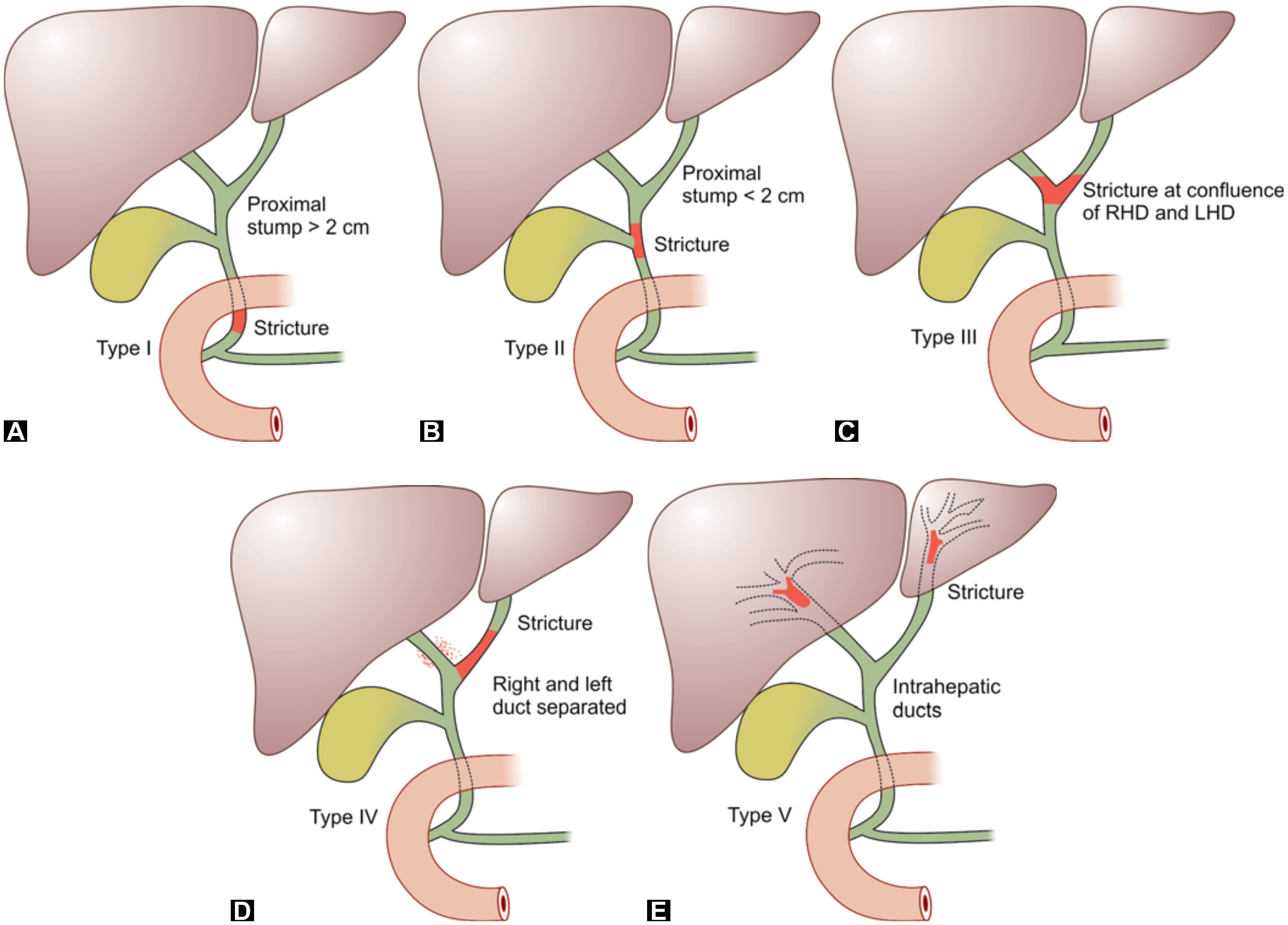
Proximal
stump > 2 cm
Stricture
Type I
A
Proximal
stump < 2 cm
Stricture
Type II
B
Stricture at confluence
of RHD and LHD
Type III
C
Stricture
Right and left
duct separated
Type IV
D
Stricture
Intrahepatic
ducts
Type V
E

Figs 12.67A to E: Bismuth classification of stricture CBD.

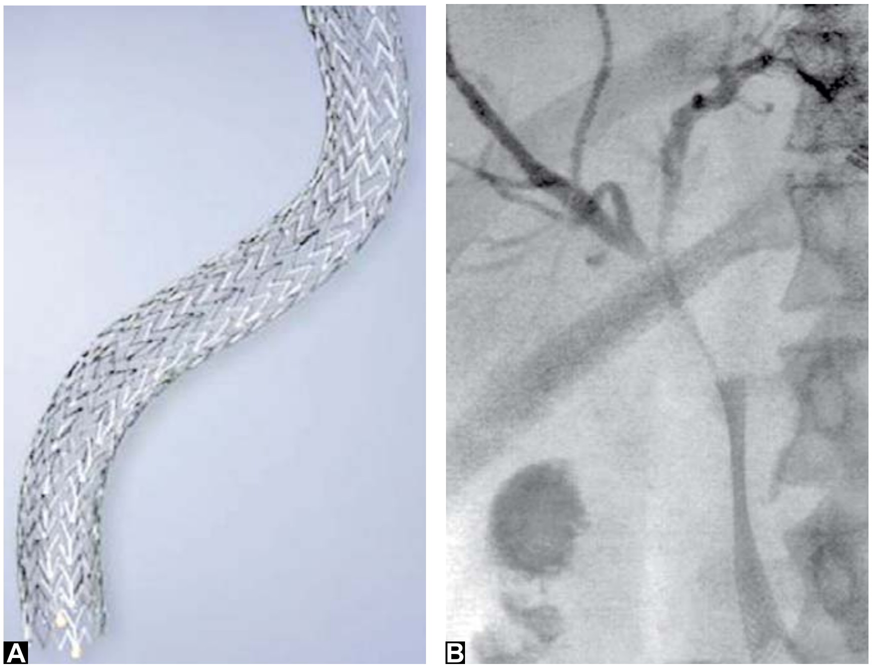

Figs 12.68A and B: CBD stent and imaging after placement.

Investigations

- U/S abdomen.
- Liver function tests.
- ERCP.
- On table cholangiography.
- MRCP.

Treatment

- ERCP stenting.
- Choledochoduodenostomy or jejunostomy.
- Roux-en-Y hepaticojejunostomy—ideal.

SCLEROSING CHOLANGITIS

- It is fibrous thickening of the CBD and biliary ductular wall, associated with multiple strictures with dilatation in between the strictures. Both extra- and intrahepatic ducts are involved.
- They have increased risk of developing cholangiocarcinoma.

Types

1. *Primary* sclerosing cholangitis is one wherein no cause is found and is associated with ulcerative colitis, Sjogren's syndrome, Crohn's disease, Grave's disease. It eventually leads to biliary cirrhosis. *Primary sclerosing cholangitis* (PSC) is an idiopathic, progressive, chronic, cholestatic pathology with diffuse inflammation, sclerosis, and obliteration of intra- and extrahepatic biliary systems. There will be multiple areas of strictures and dilatations (1%/year). It has high risk for cholangiocarcinoma. It could be an autoimmune disease. It is common in HLA/B8/DR3 halotype.
2. *Secondary* sclerosing cholangitis is due to stones, trauma, congenital lesions. AIDS, chemotherapy (5FU), transplantation, collagen diseases, sarcoidosis, histiocytosis.

Features

- Common in young men (70%).
- Intermittent jaundice.
- Abnormal liver functions.
- Weight loss, pain, fever, pruritus.
- Features of ulcerative colitis in case of PSC (70%).
- PSC may be associated with retroperitoneal fibrosis, mediastinal fibrosis, Riedel's thyroiditis, orbital pseudotumour.
- *Hepatic duct confluence* is most severely strictured segment in PSC.

Investigation

- ERCP shows *beaded appearance* of biliary tree.
- LFT is altered.
- Liver biopsy is needed to identify the severity of hepatic fibrosis.

Treatment

- Stenting.
- T-tube drainage.
- Large dose steroids.
- Immunosuppression therapy—methotrexate, azathioprine, tacrolimus, cyclosporine.
- Ursodeoxycholic acid is beneficial.
- Liver transplantation.
- *Median survival in PSC* after diagnosis is 10 years.

GALLBLADDER POLYP

- Its incidence is 5% in routine US abdomen; 10% in cholecystectomy tissue specimen.
- It is usually less than 10 mm in size, pedunculated appearance.
- Sessile polyps are often more than 10 mm in size.
- 30% are multiple.
- Gallbladder polyp more than 10 mm in size, associated with gallstones, age above 60 years, symptomatic, sessile polyp and multiple polyps—are indications for surgical intervention.
- CT scan should be done if malignancy is suspected on US.
- It is treated by laparoscopic cholecystectomy. Histology is a must to confirm benign nature and to rule out carcinoma.

BENIGN BILIARY PAPILLOMA

- It is the most common benign biliary tumour.
- 50% occurs close to ampulla presenting as obstructive jaundice.
- Benign adenoma (soft), benign inflammatory pseudotumour and cholangiocarcinoma—are differential diagnosis.
- ERCP, CT scan, EUS, LFT—are the investigations.
- Treatment—papillotomy, wide local excision.

CARCINOMA GALLBLADDER

- It is more common in *India (Patna)* and Asian countries. It is also common in Chile.
- It is common in *females* and elderly. Male : Female :: 1 : 3.

Aetiologies for Carcinoma of Gallbladder

- 3% of gallstones with cholecystitis will develop carcinoma of gallbladder.
- 90% of carcinoma of gallbladder is associated with gallstones. Risk of developing carcinoma in gallstone disease is 7-10 times more than general population. Relative risk is less if stone size is less than 2 cm; if stone size is 2-3 cm in size it is 2.5; it is 10 or more if stone size is more than 3 cm.
- Choledochal cyst, anomalous pancreaticobiliary duct junction (20%), cholesteroses of gallbladder, gallbladder polyp more than 1 cm in size or more than 3 in number or adenomatous polyp, PSC.

- Chronic typhoid carriers, carcinogens, inflammatory bowel disease, hepatitis B and hepatitis C virus infection.
- Porcelain gallbladder is more prone for malignant transformation (25%) and 90% of them are inoperable tumours.
- Nitrosamines.

Gross Types of Carcinoma Gallbladder

- Polypoid/papillary—better prognosis.
- Scirrhous/nodular.
- Proliferative/infiltrative.

Microscopy

- Commonly it is adenocarcinoma (90%); occasionally squamous cell carcinoma, adenosquamous or carcinoid tumour can occur.
- 25% show only localised disease; 35% have lymph node spread; 40% have distant spread at the time of first diagnosis.
- It is very *aggressive* tumour.

Spread of Carcinoma Gallbladder

- Direct spread to liver (segment IV and V), bile duct, duodenum, colon and kidney.
- Lymphatic—lymph node of Lund, periportal nodes, peripancreatic and periduodenal nodes.
- Blood spread—to liver, lungs and bones.
- Perineural spread is also known to occur.

Staging of carcinoma gallbladder

Nevin's staging:
Stage I – Intramural
Stage II – Spread to muscularis propria
Stage III – Spread to serosa
Stage IV – Spread to cystic lymph node of Lund
Stage V – Direct spread to adjacent organs/metastases

TNM staging

Tumour	***Nodal spread***
Tis – Carcinoma *in situ*	N0 – No nodes
T1 – Spread to mucosa or muscle layer	N1 – Spread to cystic/ nodes in porta/hepatoduodenal ligament N2 – Spread to peripancreatic/celiac/periduodenal/superior mesenteric nodes
T1a – Only mucosal involvement	
T1b – Spread to muscle layer T2 – Spread to subserosa but not beyond serosa	
T3 – Spread beyond serosa or one adjacent organ or < 2 cm to liver	*Metastases*: M0 – No distant spread M1 – Presence of distant spread
T4 – Spread > 2 cm to liver, 2 or more adjacent organs—CBD, stomach, duodenum, colon, omentum	

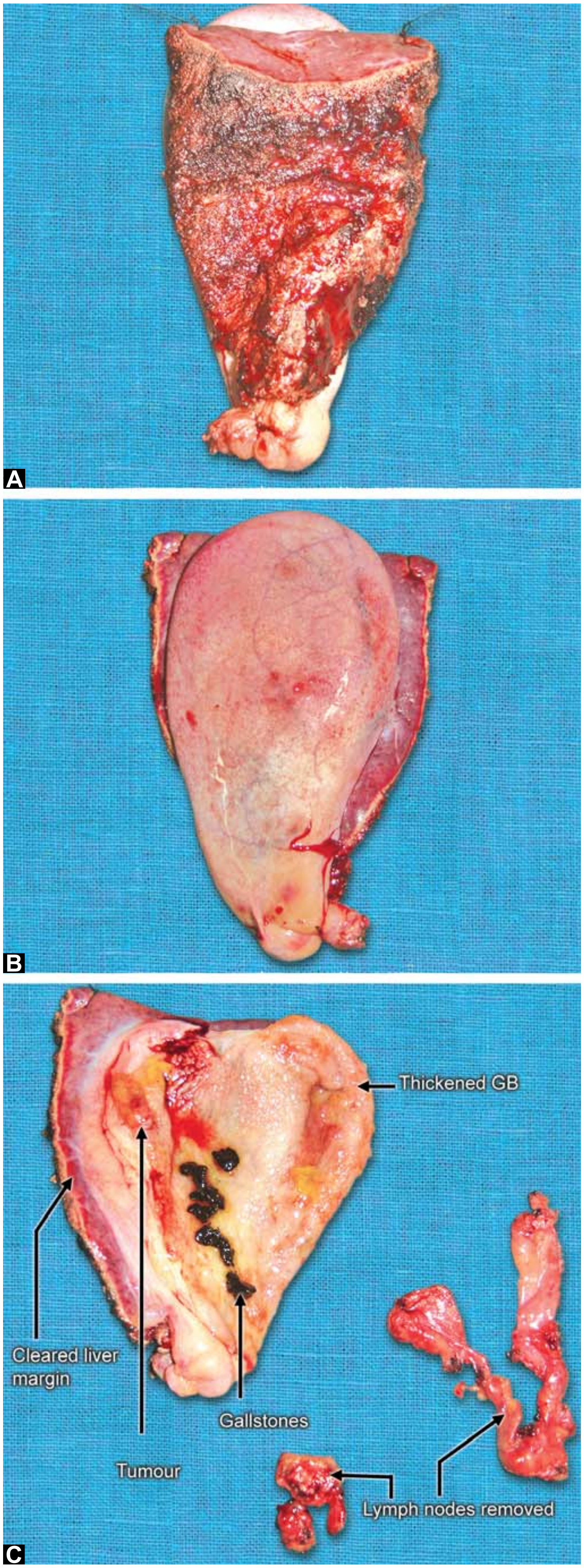

Figs 12.69A to C: Specimen of gallbladder after extended cholecystectomy. Note the liver margin cleared of carcinoma, stone in the gallbladder and lymph nodes cleared.

Features of Carcinoma of Gallbladder

- Pain in right hypochondrium, mass in right upper abdomen which is hard and nontender.
- Jaundice is common.
- Significant weight loss in short duration.
- Acute presentation of cholecystitis.
- Palpable nodular liver secondaries, ascites.
- It is common in places where there is more prevalence of gallstone disease—Patna, Bihar
- It is common in females.
- Incidentally confirmed as carcinoma gallbladder histologically after cholecystectomy for chronic cholecystitis.

Investigations

- Ultrasound abdomen.
- CT abdomen to see operability.
- U/S guided FNAC.
- Liver function tests.
- MRCP.
- Laparoscopy.
- CA 19-9 is elevated in 80% of cases.

Treatment

- Cholecystectomy with resection of liver segments IV and V—*extended cholecystectomy with perihepatic nodal clearance*. At least 2 cm margin in the liver from the gallbladder bed should be cleared. All pericholedochal lymph nodes should be removed. Frozen section biopsy from cystic duct stump should be done to identify for the existence of microscopic tumour. If present, CBD resection and hepaticojejunostomy is done. *Open approach* rather than laparoscopic is ideal for carcinoma gallbladder.
- Hemihepatectomy with cholecystectomy with nodal clearance.

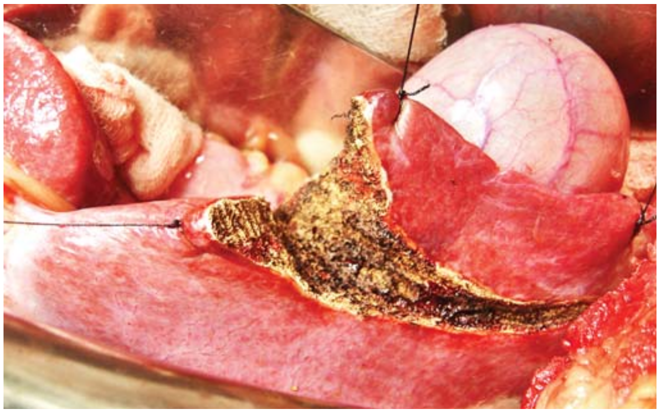

Fig. 12.70: Extended cholecystectomy technique.

- If patient has undergone laparoscopic cholecystectomy and histology confirmed carcinoma, then staging should be done. All port areas should be reexcised to prevent *port site recurrence*. Often extended resection of segment IV and lymph nodes may be needed. Spillage of bile during laparoscopic cholecystectomy is common (30% in nonmalignant GB, 50% in carcinoma GB).
- Chemotherapy either systemic or intraarterial, and adjuvant radiotherapy but with poor success rate.

Prognosis

- Overall prognosis for carcinoma gallbladder is poor due to early spread and aggressive nature of the tumour.
- 5-year survival is only 5%. Muscle invasion, nodal and distant spread carry poor prognosis. In stage T_1 simple and extended cholecystectomy will not make difference in prognosis. In T_2 stage extended cholecystectomy is very much beneficial which gives 60% 5-year survival rate. T_3 and T_4 carry poor prognosis.

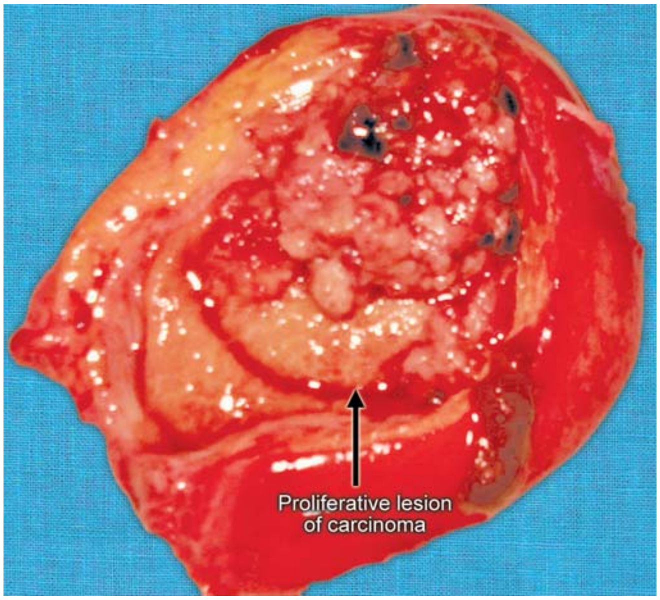

Fig. 12.71: Gallbladder showing proliferative carcinomatous lesion.

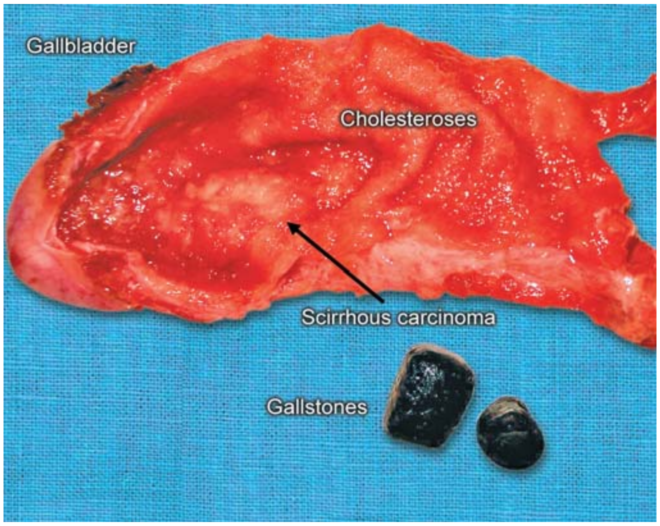

Fig. 12.72: Gallbladder showing scirrhous carcinoma with pre-existing gallstones and cholesteroses.

CHOLANGIOCARCINOMA (Bile Duct Carcinoma)

- It is associated with sclerosing cholangitis, clonorchiasis infestation, Caroli's disease or choledochal cyst.
- It is an aggressive adenocarcinoma which presents as obstructive jaundice.
- It commonly occurs at the hepatic duct confluence.

Risk Factors for Cholangiocarcinoma

- Primary sclerosing cholangitis (PSC): Here commonly it is extrahepatic, usually at the confluence.
- Choledochal cyst.
- Hepatolithiasis.
- Hepatitis B and C are risk factors for intrahepatic cholangiocarcinoma.
- Lynch syndrome II.
- Multiple biliary papillomatosis.
- Previous biliary enteric anastomosis.
- Clonorchis sinensis infestation.
- Thorotrast, nitrosamines, dioxin.

Classification I

- Intrahepatic—10%.
- Perihilar—65%.
- Distal—25%.

Classification II—Pathological

Sclerosing (80%); nodular; papillary.

Classification III

- *Upper third*—from porta to cystic duct – 50%.
- *Middle third*—from cystic duct to upper margin of first part of the duodenum—25%.
- *Lower third*—from upper margin of first part of duodenum to ampulla—20%.
- *Diffuse—5%.*

Classification of perihilar cholangiocarcinoma (Bismuth)

- Type I : Below the confluence of RHD and LHD
- Type II : Reaching the confluence of RHD and LHD
- Type III : Involving CHD and RHD/LHD
- Type IV : Involving confluence and both RHD and LHD or multicentric tumour

Features

- Main presentation is painless obstructive jaundice of short duration.
- Palpable liver either smooth and soft (hydrohepatotic) or hard nodular (secondaries).
- Weight loss and anorexia is typical and significant.

Investigations

- ERCP and choledochoscope.
- Liver function tests.
- PTC.
- U/S abdomen, CT scan.
- MR scan—Best investigation. MRCP to *see duct*; MRI to *see nodes*; MR angiogram to *see vascularity.*

Assessment should be done to check the operability of tumour by—assessing the spread in bilateral hepatic duct up to secondary radicles, both hepatic arterial spread, portal vein encasement, hepatic lobe atrophy, distant spread.

TNM staging of cholangiocarcinoma

Tumour	*Nodes*	*Staging*
Tis – Carcinoma in situ	N0 – No regional nodes	Stage 0 – TisN0M0
T1 – Invasion of subepithelial connective tissue	N1 – Hepatoduodenal ligament nodes	Stage I – T1N0M0
T2 – Invasion of perifibromuscular tissue	N2 – Peripancreatic, periduodenal, periportal, celiac and or SMA lymph nodes	Stage II – T2N0M0
T3 – Invasion of adjacent organs	*Metastases*	Stage III – T1/2N1/N2M0
	M1 – Distant metastases present	Stage IVA – T3 any N M0
	M0 – No metastases	Stage IVB – Any T anyNM1

Treatment

- When operable, portal region clearance with hemihepatectomy can be done. *Intrahepatic type* is treated with hemihepatectomy. *Perihilar* is treated with hemihepatectomy or extensive bile duct resection, nodal clearance, caudate lobe removal, cholecystectomy. Distal tumour is treated with Whipple's pancreaticoduodenectomy.
- Most often it is inoperable. Stenting can be done to relieve jaundice, through PTC or ERCP or on table.
- Chemotherapy—5 FU; Gemcitabine, cis platin.
- Chemotherapy, with external beam radiotherapy (ERBT).

KLATSKIN TUMOUR

- It is cholangiocarcinoma at the confluence of the hepatic ducts and common hepatic duct above the level of the cystic duct (20% of cholangiocarcinomas).
- *It causes obstructive jaundice with hydrohepatosis without enlargement of gallbladder.*
- Management is like cholangiocarcinoma.
- Hilar resection with hepaticojejunostomy.
- Segment III left hepaticojejunostomy—Blumgart's.
- Longmire's left hepaticojejunostomy after left liver lobe resection.
- Cattell's hepaticojejunostomy.
- Smith's mucosal graft hepaticojejunostomy.
- Palliative PTC with stenting/ERCP stenting, intraoperative stenting. Metal stents show long period of patency (12 months) than polyethylene stents (4 months).

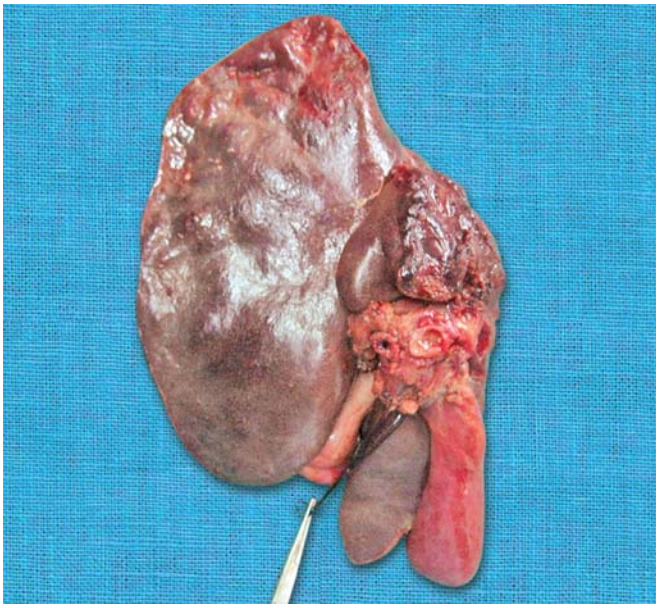

Fig. 12.73: Resected specimen of Klatskin tumour (*Courtesy:* Dr Arunkumar, MCh).

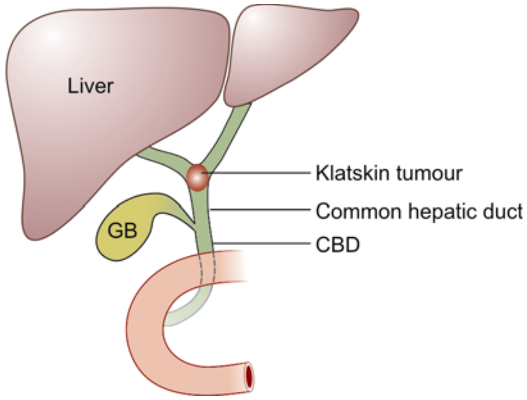

Fig. 12.74: Klatskin tumour.

Cholangiocarcinoma

- It is an uncommon tumour of elderly. 1 in 1,00,000 is the incidence
- Primary sclerosing cholangitis is the commonest association. 20 fold increased risk
- Choledochal cyst, Caroli's disease, pyogenic cholangiohepatitis, Clonorchis sinensis, opisthorchis viverrini infestation, hepatolithiasis, hepatitis C, Lynch syndrome II, bile duct adenoma and papillomatosis are other causes
- Chemical carcinogens like thorium, nitrosamines, diosmin also can cause
- Intrahepatic cholangiocarcinoma is 10%
- 60% cases are proximal to cystic duct CBD junction
- Features are jaundice, pain, hepatomegaly, cachexia, early satiety and weight loss
- It is adenocarcinoma type
- Lymph node spread is common
- Resection along with liver/resection along with pancreas are surgical options
- But surgical resection is possible only in 5% of cases
- Stenting/bypass are palliative options
- Doxorubicin, cis platin and I_{131} anti-CEA antibodies are adjuvant therapies
- Condition has got 90% mortality in one year

BILIARY FISTULAS

Types

a. *External:* Usually occurs as a complication of surgery following:
   - Gastrectomy.
   - Cholecystectomy open or laparoscopic.
   - CBD surgery.
   - Pancreatic surgery.

b. *Internal:*
   - Cholecystoduodenal fistula causing gallstone ileus.
   - Cholecystoenteric fistula.
   - Cholecystocholedochal fistula.

   External biliary fistulas are difficult to manage, often dangerous, but many a times resolve spontaneously.

Investigations

- Fistulogram.
- ERCP.
- Electrolyte estimation.
- Liver function tests.

Treatment

- Total parenteral nutrition (TPN).
- Antibiotics.
- Blood transfusion.
- Electrolyte management.
- Care of the fistula wound with regular dressing, using zinc oxide cream has to be done to protect the skin.
- Later continuity of biliary system has to be restored by open surgery or through ERCP and stenting.

GALLSTONE ILEUS

It is a type of acute intestinal obstruction, often seen in elderly and is due to blockage by a bolus or mass of gallstones which commonly enter the intestine through cholecystoduodenal fistula (75%) or rarely through cholecystointestinal or gastric fistulas.

The same beam of light can illuminate two objects and produce two different effects.

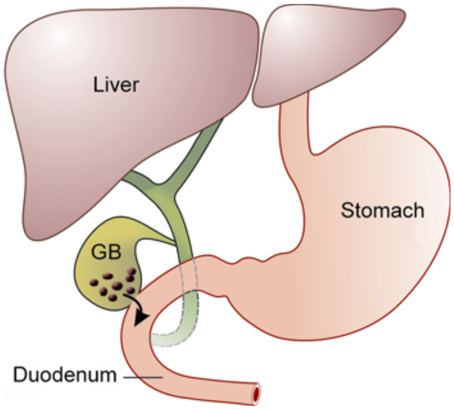

Fig. 12.75: Gallstone ileus.

Gallstones in the gallbladder
↓
Cholecystitis
↓
Suppuration and adhesion over the duodenal wall
↓
Communication of gallbladder into the duodenum
↓
Gallstones pass into the duodenum forms a *bolus ('Rolling stone gather mass')*
↓
Blocks narrow part in the ileum.
↓
Gallstone ileus

Clinical Features

- Pain abdomen and features of intestinal obstruction.
- Stones may perforate the ileum to cause peritonitis.
- It is 1% of all intestinal obstruction overall; 25% of obstruction in elderly.
- Recurrent episodic obstruction due to moving stone bolus is typical—*tumbling obstruction*.

Investigations

- Plain X-ray abdomen in erect posture shows *air in the biliary tract* (branching gas pattern) and multiple air fluid levels.
- U/S abdomen.
- CT is diagnostic.

Treatment

- Laparotomy, enterotomy, removal of gallstones and closure of enterotomy is done. Enterotomy is done not at the site of obstruction but more proximal to the site of obstruction and stones are milked towards the enterotomy site. If bowel is found ischaemic at the impacted area, resection and anastomosis is done.
- Laparotomy and crushing of stones with fingers to relieve the obstruction is only occasionally useful.
- Cholecystectomy, correction of fistula with T tube drainage can be done in same sitting if patient's general condition is good. Otherwise it is done after 12 weeks. If cholecystectomy and definitive corrective surgery is not done, recurrent gallstone ileus is likely to occur.

Causes for gas in the biliary tree

- Cholecystoduodenal fistula
- Choledochoduodenostomy
- Choledochojejunostomy
- Transduodenal sphincteroplasty
- Emphysematous cholecystitis

Bouveret's Syndrome

- A rare entity with cholecystoduodenal fistula causing gallstones from gallbladder to pass into the duodenum leading to *duodenal obstruction*.

HEMOBILIA

- It is bleeding commonly from the liver or occasionally from the gallbladder into the biliary tract.
- There is abnormal communication between a blood vessel and a bile duct or any part of the biliary tree.

Causes

- Accidental trauma, ***iatrogenic*** trauma (50%). In accidents hemobilia is more commonly caused by blunt trauma (3%) than by penetrating one.
- Percutaneous diagnostic and therapeutic procedures.
- Vascular diseases of the hepatic artery.
- Malignant liver diseases.
- Portal hypertension.
- Parasitic liver diseases like hydatid disease.
- Gallstones rarely erode into hepatic artery causing life-threatening hemobilia.
- Blood and necrotic material drains into the biliary tree causing gastrointestinal bleeding. As bile interferes with coagulation a fatal haemorrhage can occur.
- Arterial hemobilia is more common. Portal venous hemobilia is rare.
- Early rapid presentation can occur; but delayed, after weeks/months/years, presentations are known to occur which is often difficult to suspect and diagnose.
- These recurrent clots can cause pancreatitis, cholangitis.

Clinical Features

- Pain which is colicky in nature.
- Obstructive jaundice.
- Haematemesis and melaena.

Triad of Sandblom in hemobilia

- Jaundice—60%
- Pain—70%
- Upper GI bleed with melaena (90%), haematemesis (60%)

All 3 features of triad are seen only in 20% of cases.

Investigations

- LFT.
- U/S abdomen.
- Selective arteriography—test of choice in detecting bleeding site in 90% of cases.
- Upper GI scopy—useful only in 10% of cases.

Treatment

- *Aim: To stop bleeding and to relieve biliary obstruction.*
- Antibiotics.
- Blood transfusions.
- Selective arterial embolisation. Transarterial embolisation (TAE) shows 90% success rate. It is the *choice therapy* with least morbidity.
- Surgical intervention is *only rarely indicated* in failed TAE. Laparotomy, ligation of bleeding vessel or hepatic artery, excision of aneurysm, hepatic resection—are the procedures done.

  Hemobilia is often a fatal condition.

Note:
Bilhemia: It is different very rare entity wherein bile flows into the hepatic or portal venous system due to raise in intrabiliary pressure more than that of portal or hepatic venous pressure. It may be due to trauma, iatrogenic or gallstone eroding into the venous system. Large quantity of bile enters the venous system and later to lung becoming fatal. If flow is low which is common, subsides spontaneously. ERCP is diagnostic. Treatment is stenting. Septicaemia and mortality is common.

WHITE BILE

- It is double misnomer.
- It is *neither white nor bile.* It is opalescent.
- It is *mucous secreted by the lining of biliary tree*.
- It signifies severe obstruction due to stone (impacted in the CBD), or carcinoma head of pancreas or periampullary region.
- It is on table finding during surgery.
- It means liver is unable to secrete bile due to raised intraductal pressure, and so can anticipate hepatic failure.
- Indicates a poor prognosis.

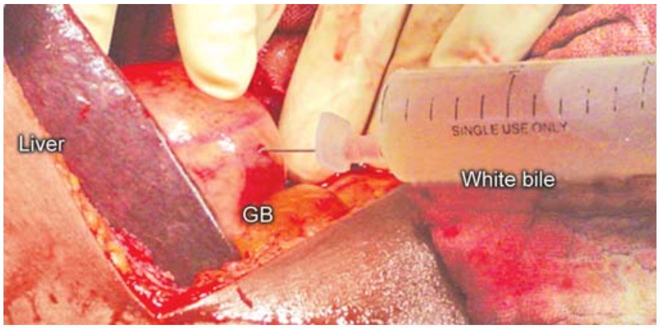

Fig. 12.76: Demonstration of white bile on table. It is not white but opalescent. It is not bile but it is mucous.

CHOLECYSTECTOMY

It is the surgical removal of gallbladder.

Prophylactic cholecystectomy is done in

- Diabetic patients
- Congenital haemolytic anaemia
- Patients who has underwent bariatric surgery

Indications

- Gallstones—symptomatic.
- Cholecystitis—acute, chronic.
- Acalculous cholecystitis.
- Empyema gallbladder.
- Mucocele gallbladder.

Approach

- *Open*
  - Right subcostal incision (Kocher's).
  - Right paramedian.

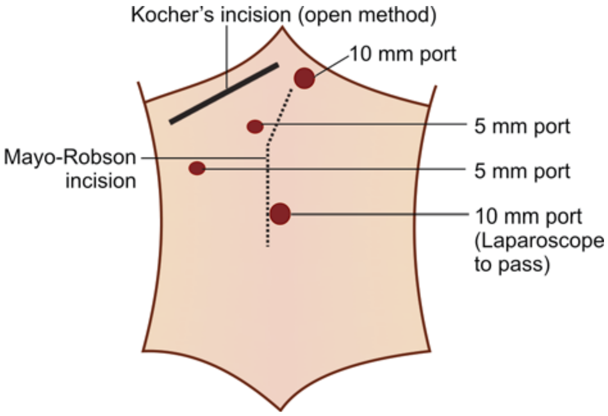
Kocher's incision (open method)
10 mm port
5 mm port
Mayo-Robson incision
5 mm port
10 mm port (Laparoscope to pass)

Fig. 12.77: Incisions for open cholecystectomy and laparoscopic cholecystectomy.

  - Horizontal incision.
  - Mayo-Robson incision.
- *Laparoscopic approach.*

Technique

After opening the abdomen, colon is pushed downwards and stomach medially.

Duct—first method: Here *Calot's triangle* is dissected. Cystic artery is identified and ligated. Cystic duct is ligated close to the gallbladder. Gallbladder is separated from gallbladder fossa and removed. Haemostasis is maintained.

Fundus—first method: It is done in difficult gallbladder due to dense adhesions. Fundus is separated from the liver bed. Dissection is carried proximally until cystic duct and cystic artery are identified, which are then ligated.

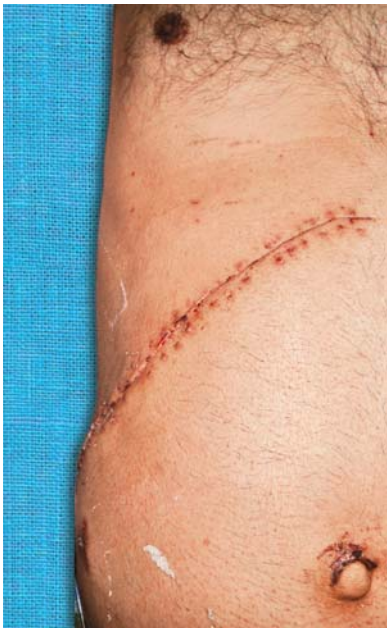

Fig. 12.78: Kocher's incision used for open cholecystectomy.

Drain is placed, which is removed after 72 hours.

On table cholangiogram is a must after cholecystectomy.

Complications of Cholecystectomy

- Complications can occur either in open method or in laparoscopic method.
- Open method is done through either right paramedian incision or Kocher's incision (right subcostal).

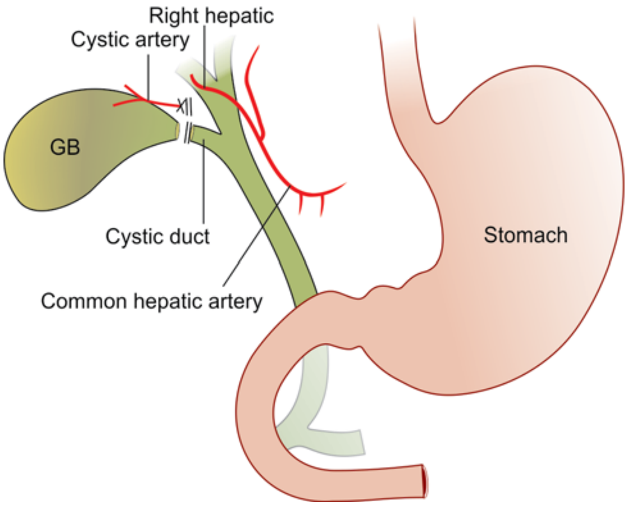

Fig. 12.79: Ligation of cystic duct and cystic artery while doing cholecystectomy.

Complications which can occur in both:

- Infection and subphrenic abscess
- Bleeding from cystic artery, and from liver bed
- Injury to CBD or hepatic duct
- Bile leak and biliary fistula formation
- Biliary stricture formation
- Injuries to colon, duodenum, mesentery

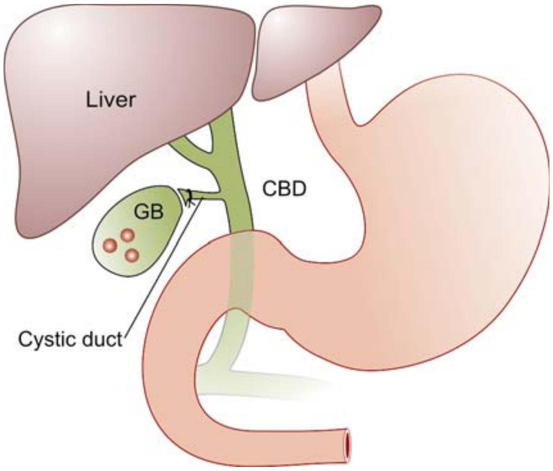

Fig. 12.80: Diagram showing the partial cholecystectomy wherein GB is removed at Hartmann's pouch, distal to Calot's triangle. It is done when there is dense adhesions,with difficult surgical plane.

LAPAROSCOPIC CHOLECYSTECTOMY

It is the most popular method to remove gallbladder. It is the gold standard treatment for gallstone.

Position

- Supine, head end up and right side up. *Anaesthesia*—general.

Ports

- 10 mm port in umbilicus to pass 10 mm telescope.
- 10 mm port in midline epigastrium as working channel.
- Two 5 mm ports at midclavicular and anterior axillary line in subcostal region.

Procedure

After creation of pneumoperitoneum with 12-14 mm pressure, 10 mm umbilical port is inserted. Telescope is passed. Under vision remaining ports are passed. With lateral 5 mm port, gallbladder grasper forceps is passed and fundus of gallbladder is held and pushed up towards the diaphragm. With middle 5 mm port grasper is passed to hold Hartmann's pouch. With 10 mm port dissector is passed using reducer. Calot's triangle is dissected. Cystic duct is identified. Adhesions are released. First posterior dissection is completed. Cystic artery is above and deep to cystic duct. Cystic duct is clipped or ligated. Cystic artery is also clipped. Gallbladder is dissected off the liver bed using cautery (hook/spatula)/harmonic scalpel. Gallbladder is removed through 10 mm working port with reducer or using a sterile bag. Any bleeding points are coagulated. If needed saline wash is given to the bed. If infected, if gallbladder is opened, if there are adhesions, if there is oozing from gallbladder bed, a tube drain is placed through lateral 5 mm port. All ports are removed. Umbilical port is sutured in layers. Other ports are sutured. Patient is asked to take oral food in 24 hours and can be discharged in 24-48 hours.

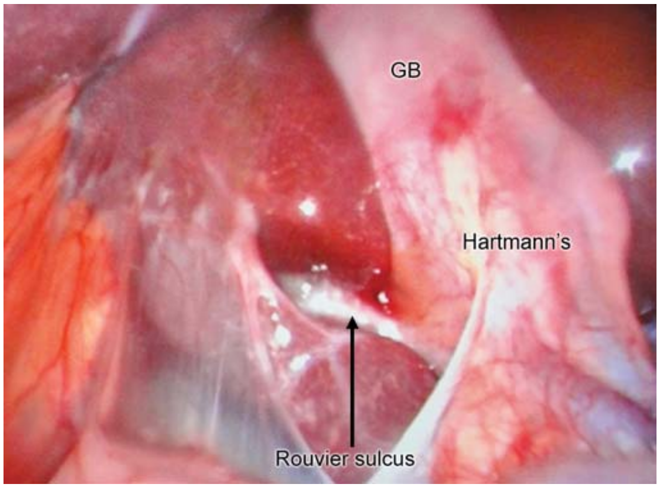
GB
Hartmann's
Rouvier sulcus

Fig. 12.81: Rouvier sulcus on the undersuface of liver is a guideline during laparoscopic cholecystectomy wherein dissection in Calot's should be always kept in front of it to avoid CBD/hepatic duct injury.

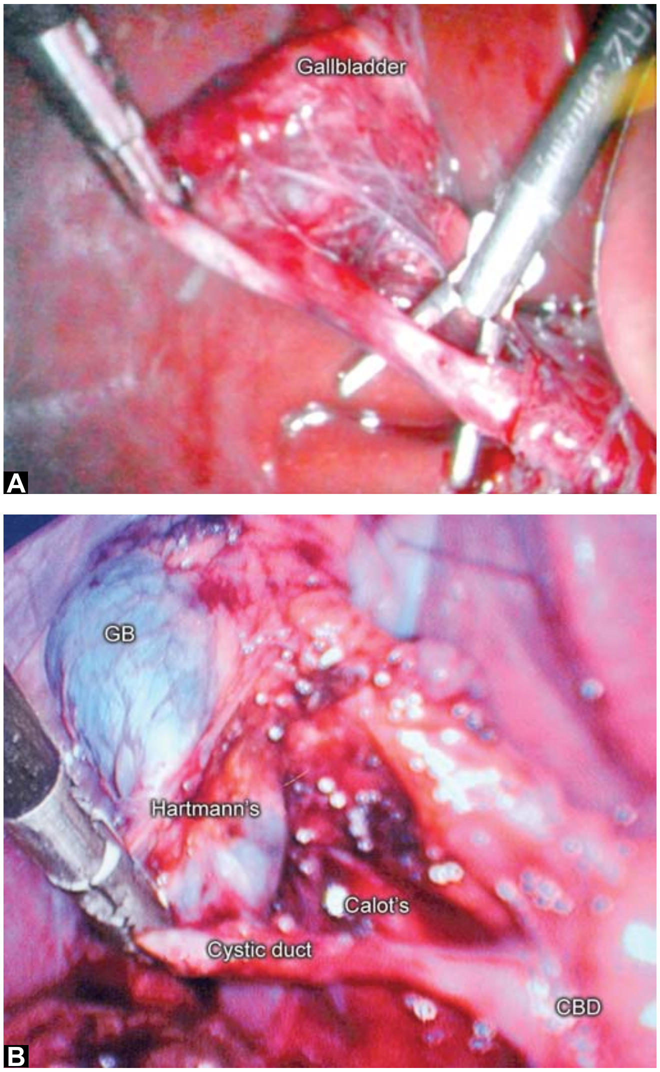
Gallbladder
A
GB
Hartmann's
Calot's
Cystic duct
CBD
B

Figs 12.82A and B

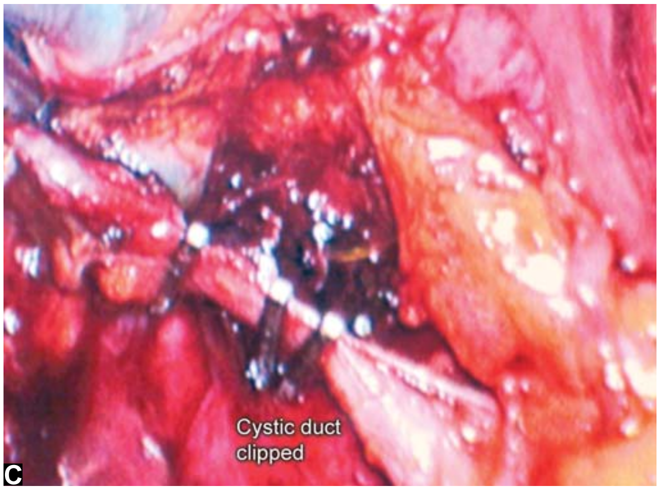
Cystic duct clipped
C

Fig. 12.82C

Figs 12.82A to C: Photos of laparoscopic cholecystectomy showing cystic duct, gallbladder, Calot's triangle and clipped cystic duct.

Problems

- Difficult Calot's triangle.
- Dense adhesions.
- Bleeding.
- Anomalies of cystic duct, cystic artery.

Complications

- Bile duct injury—0.8 %.
- Bleeding.

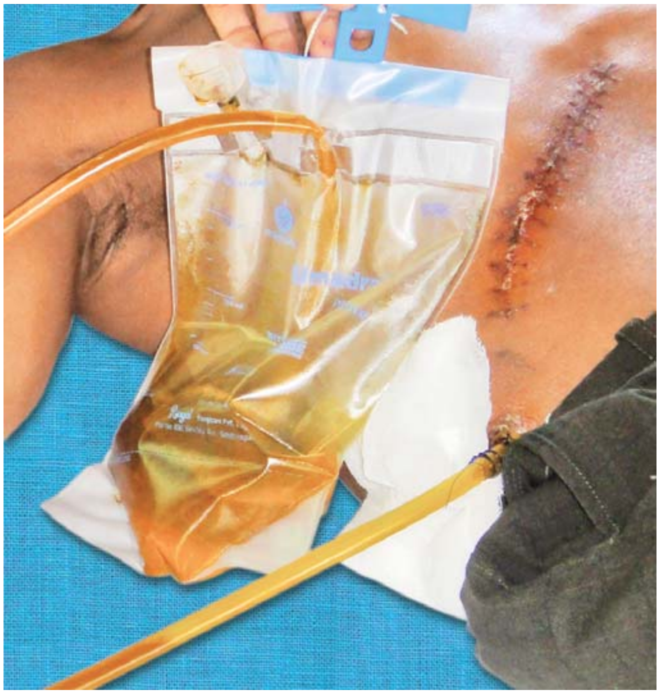

Fig. 12.83: Post-cholecystectomy bile leak with fistula. It is common and problematic complication after cholecystectomy.

Condition your mind and remain cool under all conditions.

- Bile leak.
- Infection, cholangitis, septicaemia.
- Subphrenic abscess formation.
- Injury to colon, duodenum, mesentery.

SINGLE INCISION LAPAROSCOPIC SURGERY (SILS) IN CHOLECYSTECTOMY

SILS is an advanced minimally invasive surgical procedure wherein surgeon operates exclusively through a single umbilical entry port. It is also called as single port access surgery (SPA), one port umbilical surgery (OPUS), and Single port incision less conventional equipment-utilising surgery (SPICES), natural orifice transumbilical surgery (NOTUS).

It needs general anaesthesia, specialised umbilical large trocars which accommodates working instruments along with, flexible laparoscope, rotatable instruments, articulating handles, harmonic scalpel.

Here through a large 2.5 cm umbilical vertical incision dissection is done by open method to reach peritoneal cavity. Specialised port in which one can pass 10 mm telescope and two 5 mm instruments for work is used. Instruments are angled and flexible to meet the ergonomic principles to certain extent. Dissection of gallbladder is done in similar fashion like four port technique. Specimen is easily retrieved through umbilical port as it is wide enough. If difficulty arises any time one can add additional ports as required.

Figs 12.84A and B: Single incision laparoscopic surgery (SILS) for cholecystectomy.

Advantages

- There is no visible scar like a traditional multiport.
- Faster recovery time, early return to work.
- Cosmetically better.

Disadvantages

- Expensive trocars and instruments—cost factor.
- Skilled work, learning curve.
- Dissection against normal surgical ergonomics.

Complications

- Umbilical wound pain, infection.
- Umbilical hernia.
- Because of limited visibility time consuming.
- During learning curve complications of cholecystectomy and conversion rate may be more.

***Note*:**
Waltman-Walter syndrome: It is compression of IVC due to collection of fluid in subphrenic space during post-cholecystectomy period. It mimics coronary thrombosis.

BILE DUCT INJURIES

Causes

- Cholecystectomy—open or laparoscopic.
- Trauma.
- Instrumentation—ERCP, baby endoscope, choledochoscope.
- Other surgeries—pancreatic, duodenal, gastric.
- Anomalous biliary system increases the risk of injury.

Classifications

Strasberg classification

- Class A: Bile leak from the cystic duct or an accessory duct showing continuity with the common bile duct
- Class B: Section/injury of an accessory duct/sectorial ducts aberrant right hepatic duct with no continuity with the common bile duct with occlusion
- Class C: Leak/open drainage from a sectorial duct/aberrant right hepatic duct with no continuity with the common bile duct
- Class D: Partial section/lateral injury of an extrahepatic duct bile duct with no complete loss of continuity with the rest of the bile duct system
- Class E: Complete section/circumferential injury of the bile duct with subtypes according to the length of the stump at various levels
  - E1- Stricture/injury at more than 2 cm distal to bifurcation.
  - E2 – Stricture/injury less than 2 cm distal to bifurcation.
  - E3 – Stricture/injury at bifurcation
  - E4 – Stricture/injury involving right and left hepatic ducts
  - E5 – Complete obstruction of entire bile duct

Note: Only right and left partial injuries are not included in this classification

Stewart-Way classification

This classification involves four strata based on the mechanism and anatomy of injury.

- Class I: It refers to the incomplete section of bile duct with no loss of tissue. It has a prevalence rate of 8%
- Class II: It is a lateral injury of the common hepatic duct that leads to stenosis or bile leak. It is the consequence of thermal damage and clamping the duct with surgical staples. It has a prevalence of 2% with a concomitant hepatic artery injury in 20% of cases. T-tube related injuries are included within this class
- Class III: It is the most common (60%) represents the complete section of the common hepatic duct. It is subdivided into type IIIa, remnant common hepatic duct; type IIIb, section at the confluence; type IIIc, loss of confluence; and type IIId, injuries higher than confluence with section of secondary bile ducts. Associated injury of right hepatic artery occurs in 25% of cases
- Class IV: It is right (65%) and accessory right (25%) hepatic duct injuries with associated injury of the right hepatic artery (60%). Occasionally it includes the common hepatic duct injury at the confluence (5%) besides the accessory right hepatic duct lesion (prevalence of 10%)

Hannover classification (2007)

It classifies injuries in relationship to the confluence and also includes vascular injuries.

- Type A: It refers to cystic and/or gallbladder bed leaks.
- Type B is a complete or incomplete stenosis caused by a surgical staple.
- Type C represents lateral tangential injuries.
- Type D refers to complete section of the common bile duct emphasizing their distance to the confluence as well as the concomitant injuries of hepatic artery and portal vein.
- Type E is late bile duct stenosis at different lengths to the confluence.

Note: Bismuth classification (for open surgery/pre-laparoscopic era) is discussed under biliary stricture earlier.

Presentations

- Fever, pain, features of bile stasis like cholangitis, jaundice.
- 10% of laparoscopic bile duct injuries are identified on table; 30% with bile leak; 25% with obstructive jaundice; 30% with cholangitis. Patients with bile leak present early; with formed biliary stricture presents with jaundice and cholangitis months or years later.

Complications

- Biliary fistula with electrolyte imbalance.
- Biliary peritonitis, septicaemia.
- Subphrenic abscess formation.
- *Biloma* formation.
- Severe malnutrition, deficiencies.
- Late—biliary cirrhosis, portal hypertension.

Investigations

- LFT, coagulation profile.
- US abdomen, fistulogram.
- Dynamic CT scan will delineate the anatomy of injury.
- ERCP is useful if injury is only partial where stenting can also be done.
- PTC is the *investigation* of choice to identify site, nature, extent of stricture. It also facilitates drainage and stenting.
- Hepatobiliary scintigraphy may be useful.
- MRCP may be useful to identify ductal anatomy.
- CT arteriography is useful to identify associated hepatic artery or portal vein injury which is 20% in association with CBD injury.

Management

- General: antibiotics, nutrition, TPN.
- Conservative: Small partial injury to CBD may resolve spontaneously or with the help of ERCP, sphincterotomy and stenting is done.
- *Management of biliary duct injury on table* (during ***laparoscopic cholecystectomy***/open): Conversion into open surgery with a lengthy subcostal/bucket handle incision. Intraoperative cholangiogram should be done. Partial injury of CBD less than 30% of circumference is treated with primary repair with a T-tube in place. Extensive injury more than 30% of the circumference or cautery injury or complete CBD transection should be treated with Roux-en-Y choledochojejunostomy. Isolated hepatic duct injury less than 3 mm in size should be ligated. More than 3 mm in size should be reimplanted or Roux-en-Y hepaticojejunostomy should be done. If facility for surgical repair is available, patient should be sent to a higher centre after placing a drain with abdomen closure.
- *Management of bile duct injury identified at a later period*: In such situation it is better to wait for 6 weeks for the inflammation to subside which facilitates easier identification of proximal CBD. After proper assessment of injury, Roux-en-Y choledochojejunostomy or *Roux-en-Y hepaticojejunostomy* is done.
- *Hepp-Couinaud approach*: Here hilar plate is meticulously dissected; left hepatic duct is anastomosed to Roux jejunal loop to create hepaticojejunostomy often with creation of proximal 'access loop' for future endoscopic approach. Complications are—recurrent cholangitis, bile leak from stoma, hemobilia, stenosis of biliary enteric anastomotic site (10%, occurs often after many years).

POST-CHOLECYSTECTOMY SYNDROME (15%)

- Recurrent, new or persistent symptoms after cholecystectomy in patients who have no demonstrable abnormality is called as post-cholecystectomy syndrome.

- It may be due to loss of reserving function of the gallbladder or continuous bile flow into the duodenum may be causing oesophagitis/gastritis or diarrhoea and colicky pain. 20% causes are other than due to hepatopancreatic biliary problems.
- However importance lies in ruling out the other causes of the symptoms like peptic ulcer, hiatus hernia, pancreatic diseases, residual CBD stone, cystic duct remnant, papillary stenosis.
- True post-cholecystectomy syndrome is treated with proper counseling, psychiatric evaluation and drug therapy. It should be evaluated thoroughly.

Causes of post-cholecystectomy syndrome

Gallbladder	Residual or reformed gallbladder Stump stones in cystic duct Neuroma in the stump Slipped stones into the peritoneum Clips applied to cystic duct
Biliary tree	Residual stone in CBD Biliary dyskinesia Stricture Dilatation without obstruction Fistula Cholangitis Malignancy
Liver	Fatty liver, hepatitis, cirrhosis Sclerosing cholangitis
Pancreas	Pancreatitis, stones Carcinoma pancreas Sphincter of Oddi dyskinesia
GIT	Oesophagitis / hiatus hernia/ achalasia cardia Duodenal diverticula IBS, constipation, incisional hernia Mesenteric ischaemia
Cardiac	Coronary heart disease
Others	Adrenal tumour Thyrotoxicosis Intercostal neuralgia Arthritis *Psychiatric diseases*

BILIARY DYSKINESIA

Biliary dyskinesia is *motility disorder* of either gallbladder or sphincter of Oddi. Patient presents with features of biliary colic without any evidence of gallstone disease.

It can be:

- **Gallbladder dyskinesia**:
    - Here CT scan, gastroscopy, ERCP are normal. CCK is injected intravenously after filling gallbladder with radionuclide labeled Tc99m, gallbladder ejection fraction is assessed after 20 minutes. If it is less than 35% then it is called as gallbladder dyskinesia.
    - It is treated with laparoscopic cholecystectomy.
- **Sphincter of Oddi dysfunction**:
    - It is also known as *biliary sphincter dyskinesia or pancreatic sphincter dyskinesia*. It is benign acalculous obstruction to bile and/or pancreatic flow across sphincter of Oddi causing cholestasis, pancreatitis and pain.
    - It may be due to trauma, congenital anomalies.
    - In occurs in 1% of cholecystectomy patients. Condition is common in females.
    - Normally CCK relaxes sphincter by decreasing its pressure. After IV injection of CCK, if CBD diameter becomes more than 12 mm or if there is increase in CBD diameter in response to CCK on US evaluation it suggests dysfunction of sphincter of Oddi. If basal sphincter pressure, on study increases more than 40 mm Hg it also suggests the same.
    - *Milwaukee's classification*: ***Type I***: *Any one of the* following features—unexplained biliary pain for more than 6 months after cholecystectomy, ERCP showing CBD > 12 mm, delayed bile drainage > 45 minutes, altered liver enzymes. ***Type II***: Unexplained biliary pain for more than 6 months after cholecystectomy + *with* any one of the other criteria. ***Type III:*** *Only* unexplained biliary pain for more than 6 months after cholecystectomy, without any of other criteria.
    - ERCP/MRCP/sphincter of Oddi manometry, IV cholangiography are the *investigations.*
    - *Treatment*: Endoscopic sphincterotomy, botulinum toxin injection into the sphincter, nifedipine, transcutaneous electric nerve stimulation to raise serum VIP level which decreases the sphincter pressure.

Chapter 13 Spleen

CHAPTER OUTLINE

- Surgical Anatomy
- Functions of the Spleen
- Splenunculi
- Splenic Injury
- Splenomegaly
- Hereditary Spherocytosis
- Immune Haemolytic Anaemia
- Thalassaemia
- Sickle Cell Disease
- Idiopathic Thrombocytopenic Purpura
- Thrombotic Thrombocytopenic Purpura
- Splenectomy
- Overwhelming Postsplenectomy Infection
- Splenic Artery Aneurysm
- Splenic Abscess
- Hypersplenism
- Splenic Cyst

SURGICAL ANATOMY

- It is a wedge-shaped organ lying mainly in left hypochondrium, along the long axis of 10th rib.
- Hilum of spleen transmits splenic vessels and nerves. The visceral surface is related to stomach, splenic flexure of colon, kidney.

Ligaments of Spleen

Spleen is suspended by two ligaments *(a) lienorenal ligament, (b) gastrosplenic ligament.*

- Lienorenal ligament transmits blood vessels to spleen. The tail of pancreas lies in this ligament, which can be damaged during splenectomy.
- The gastrosplenic ligament contains short gastric vessels which supply the left half of greater curvature of stomach.
- Phrenicocolic ligament comes in contact with lower pole of spleen, which may be damaged during mobilization of splenic flexure of colon.

Blood Supply

Splenic artery is the branch of coeliac artery but may arise from aorta or superior mesenteric artery, blood flow is 300 ml/mt. *Splenic vein* joins the superior mesenteric vein at right angle behind the neck of pancreas to form the portal vein.

Splenic parenchyma contains *white pulp* and *red pulp*. *White pulp* lies in centre surrounding the central artery, which is a branch of trabecular artery. It is made up of lymphatic nodules with germinal centres and periarterial lymphatic sheaths with a network containing lymphocytes and macrophages. White pulp is surrounded by marginal zone which contains end arteries from central and peripheral peniciliary arteries. Marginal zone contains marginal sinus which filters the materials from the white pulp. Immunoglobulins secreted by white pulp enter marginal zone and into main bloodstream.

Red pulp is located outer to marginal zone. Red pulp contains cords and sinuses. Central artery gives reticular branches which open into these sinuses and cords wherein particles are phagocytosed. Commonly central artery eventually ends in these cords and sinuses. Few end branches of the central artery directly enters the pulp vein. Cords and sinuses eventually drain into pulp vein. Blood circulating through these cords and sinuses are called as '*open*' circulation (90%). Blood passing through white pulp but not entering these cords and sinuses are called as '*closed*' circulation (10%).

Spleen is palpated under left costal margin during inspiration. It has to enlarge 2-3 times its normal size to become palpable.

FUNCTIONS OF THE SPLEEN

Spleen has got two groups of functions:

1. *Cellular function.*

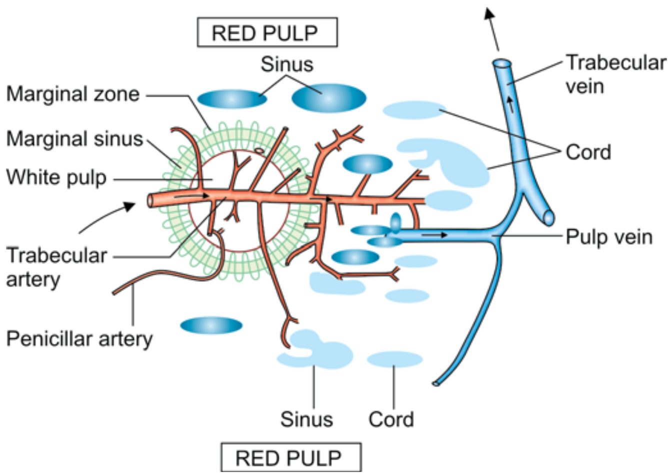

Fig. 13.1: Architecture of spleen showing red pulp (peripheral); white pulp (central); open and closed circulation; marginal zone and sinuses and trabecular artery and vein.

- Removal of non-deformable intracellular substances from deformable cells is called as "*pitting*". *Heinz bodies/ Howell-Jolly bodies, Pappenheimer siderotic bodies* are removed by this method from RBCs. Post-splenectomy patients will show these bodies in the RBCs in peripheral smear. In splenunculi, these bodies will be absent even after splenectomy.
- Removal of aged/abnormal red cells is done by a process of "*culling*". RBCs which loose osmotic balance and membrane integrity is called as non-deformable and they are removed by this method.
- Half life of platelets is 10 days. In splenomegaly 80% of platelets may be sequestered in spleen causing thrombocytopenia. Normally spleen is reservoir for platelets. 1/3rd of total platelet mass is present in spleen. In pathological status like immune diseases phagocytosis of platelets in spleen is accelerated by many folds.
- Splenomegaly and hypersplenism can cause neutropenia. Neutrophil half life is 6 hours.
- Bacterial clearance also occurs in spleen by phagocytosis. After splenectomy, patients are more prone for OPSI.

2. *Immunologic functions* like synthesis of antibody IgM; formation of lymphocytes; production of tuftsin, opsonins, properdin and interferons.

Functions of spleen

- Response to antigenic challenge—by secreting antibodies like IgM, tuftin, opsonins, properdin, interferons. These agents make bacteria and other organisms more vulnerable to phagocytosis. All bacteria including the capsulated types, virus and fungi are destroyed efficiently.
- Destruction or correction of abnormal cells like old RBCs, target cells, siderocytes, spherocytes or removal of inclusion bodies or parasites from the RBCs by culling and pitting.
- Phagocytosis of foreign substances.
- As platelet reservoir.
- Erythrocyte production: In fetal life, it is an important site of RBC production till 5th month of gestation. In adults erythropoiesis in spleen occurs only if marrow production is inadequate, e.g. in myelofibrosis.
- Iron reutilisation.

SPLENUNCULI (30%)

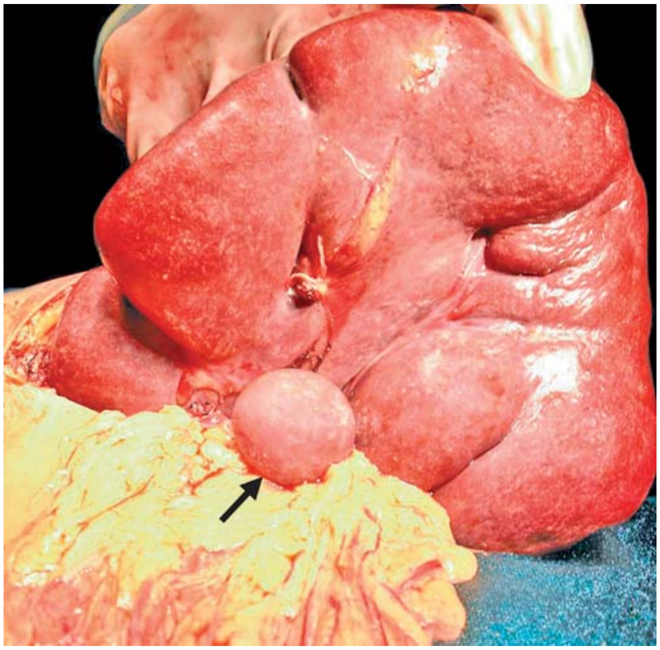

Fig. 13.2: On table photo of splenunculi with splenomegaly. Splenunculi should be removed otherwise disease will recur.

- These are single or multiple **accessory spleens**.

Sites

- Hilum of spleen (50%)
- Near splenic vessels
- Tail of the pancreas (30%)
- Splenic ligaments—gastrosplenic/splenorenal
- Mesocolon
- Greater omentum

- After splenectomy, they undergo hyperplasia and lead to recurrence of the disorder for which spleen was removed.
- Blood peripheral smear, in these patients with splenunculi *will not show* Howell-Jolly bodies. Pappenheimer siderotic bodies.
- Accessory spleen can also be occasionally present adjacent to left ovary or adjacent to left testis.

SPLENIC INJURY (RUPTURE SPLEEN)

Causes

- Splenic injury occurs commonly following road traffic accidents, other blunt injury or penetrating/stab injuries.
- Most often associated with fracture of left lower ribs, haemothorax, injury of liver (left lobe commonly, occasionally both lobes), bowel, tail of pancreas, left kidney.
- Injury is more common and severe in *enlarged spleen,* i.e. in malaria, tropical splenomegaly, infectious mononucleosis.
- Spontaneous rupture of spleen can occur in malaria and infectious mononucleosis.
- *Larang* was used to kill by the murderers in far east where malaria was endemic leading to splenomegaly, which ruptured more easily.

- Spleen is the most common solid organ injured in blunt abdominal trauma.

Types of Injury

1. *Splenic subcapsular haematoma:* After initial injury patient remains asymptomatic for a short period. But this haematoma ruptures later, may be after few days causing torrential haemorrhage.

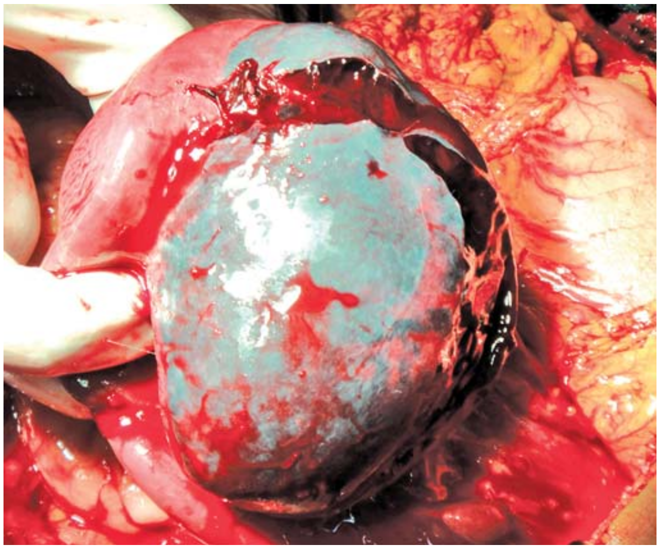

Fig. 13.3: Large subcapsular splenic haematoma.

2. *Clean incised wound over the surface*: This can be treated by splenorrhaphy.
3. *Lacerated wound.*
4. *Splenic hilar injury* causes torrential haemorrhage, may even cause death. So immediate surgical intervention and splenectomy is done.
5. Splenic injury associated with other injuries (left kidney, left colon, small bowel, pancreas, diaphragm, left lung).

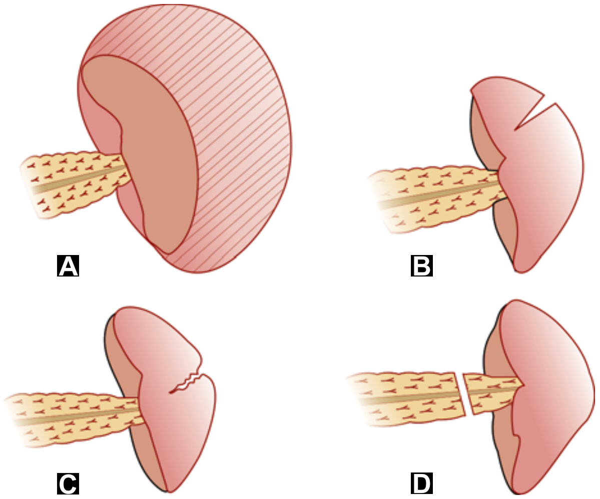

Figs 13.4A to D: Types of splenic injury. (A) Subcapsular haematoma, (B) Incised wound, (C) Lacerated wound and (D) Hilar injury.

Associated injuries

- Left lobe liver injury
- Tail of pancreas injury
- Left kidney, left colonic injury
- Small bowel injury
- Diaphragm and left lung injury
- Fracture lower ribs—left sided (30%)
- Left sided haemothorax

Presentation

Hilar injury presents with rapid development of *shock and deteriorates fast.* Even death can occur sometimes. Here *emergency surgery and splenectomy is mandatory*.

- In other types, features of shock (pallor, tachycardia, restlessness, hypotension), pain, tenderness and abdominal rigidity in left upper quadrant is seen.
- Later there will be abdominal distension due to haemoperitoneum.
- Dullness in the left flank which does not shift, as the collected blood gets clotted. *Dullness without shifting—**Ballance's sign***.
- Clot collected under the left side of the diaphragm irritates it and the phrenic nerve causing *refered pain in the left shoulder—**Kehr's sign***. There may be left sided haemothorax with fracture of ribs.

Delayed presentation is also possible due to formation of subcapsular haematoma which later gives way. Initially gets temporarily localized by greater omentum, later giving way leading to torrential bleeding. Blood clot temporarily seals off the bleeding which later gets dislodged causing severe bleeding. This time period in between is called ***'latent period of Bandet'.***

- Pseudoaneurysm and traumatic splenic arteriovenous fistula formation can also occasionally cause delayed life threatening haemorrhage.
- Features of other abdominal organ injuries may be present.

Signs in splenic rupture

- *Kehr's sign*—pain in left shoulder 15 minutes after foot end elevation
- *Ballance's sign*—left sided abdominal dullness which will not shift
- *Saegesser's tender point* between left sternomastoid and scalenus medius
- Latent period of Bandet—refer above

Splenosis: Autotransplantation of fragments of splenic tissue may occur within the peritoneal cavity following rupture of spleen

A smile is the shortest distance between two people.

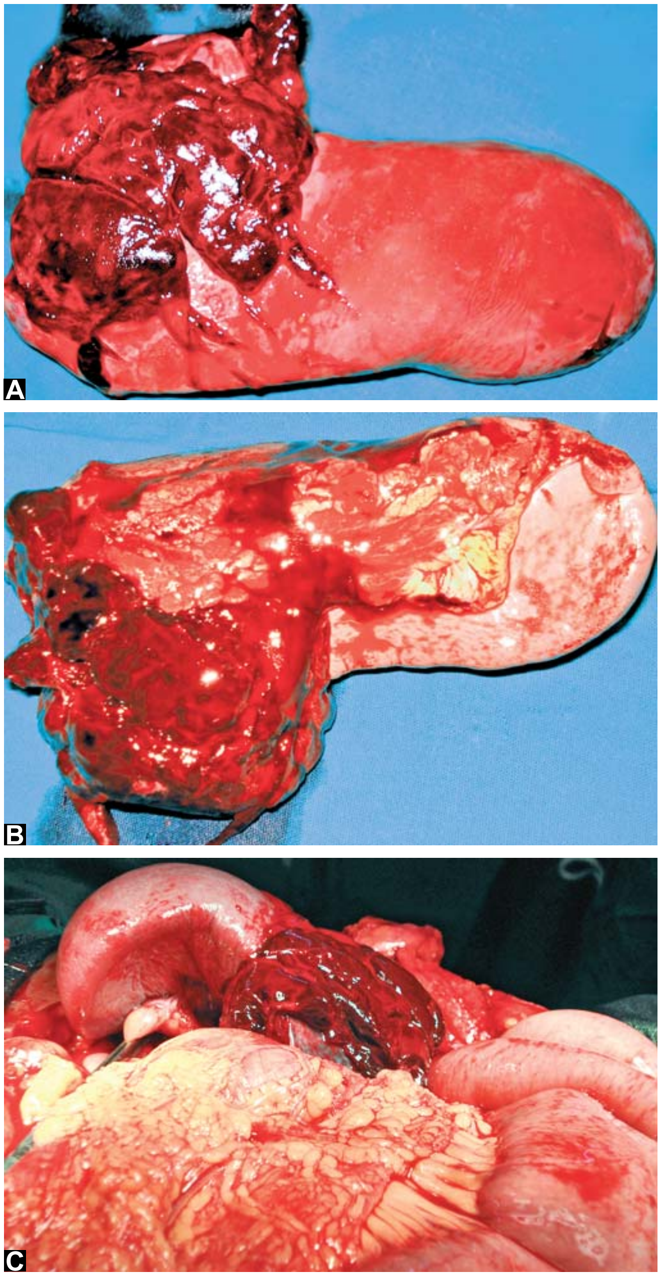

Figs 13.5A to C: Splenic injury with splenic haematoma.

Investigations

- U/S abdomen is the investigation of choice, as it is quicker, cheaper and non-invasive.
- Hb%, PCV, blood grouping and cross matching.
- *Adequate amount of blood must be kept ready for transfusion.*
- CT scan will show type of splenic injury and its class.

Plain X-ray abdomen. Findings are:

- Obliteration of splenic outline
- Obliterarion of psoas shadow (most important feature)
- Indentation of fundic gas shadow
- Fracture of lower ribs
- Elevation of left side of diaphragm
- Free fluid in the abdomen between coils of intestine

- ***Diagnostic peritoneal lavage (DPL):*** By subumbilical incision the peritoneal lavage catheter is introduced into the peritoneal cavity. One litre of crystalloid (normal saline) is introduced into the cavity. Patient is turned to both left and right side and fluid is collected back. It is sent for cytology, culture, microscopy and biochemical analysis.
  It is significant when the aspirated fluid contains:
  - Gross blood of 10 ml.
  - $>100000/\text{mm}^3$ of RBC.
  - $>500/\text{mm}^3$ of WBC.
  - Bile, bacteria or food fibres.
  - Amylase > 175 units/dl.

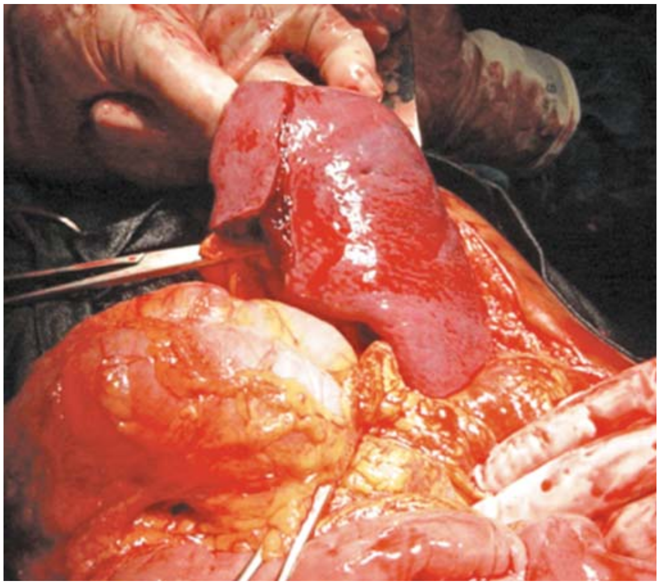

Fig. 13.6: Lacerated injury over the spleen. Patient underwent splenectomy.

- Angiogram, for diagnosis and often for angiographic embolisation to reduce the bleeding and to manage splenic injury non-operatively is also becoming routine in many centres.

Complications of Splenic Rupture/Trauma

- Blood loss.
- DIC.
- Sepsis.
- Splenic artery pseudoaneurysm.
- Splenic arteriovenous fistula.
- Problems of associated injuries like of pancreas.

Splenic organ injury scale (1994)

Grade I	Non-expanding subcapsular haematoma <10% surface area.
	Non-bleeding capsular laceration with <1 cm depth
Grade II	Non-expanding subcapsular haematoma 10-50% surface area.
	Non-expanding intraparenchymal haematoma < 5 cm in diameter

	Laceration—capsular tear 1-3 cm in depth which does not involve trabecular vessel
Grade III	Expanding subcapsular or > 50% surface area or ruptured bleeding subcapsular haematoma / intraparenchymal haematoma.
	Intraparenchymal haematoma > 5 cm or parenchymal laceration >3 cm depth involving trabecular vessels
Grade IV	Laceration involving segmental or hilar vessels with > 25% devascularization
Grade V	Shattered or avulsed spleen; hilar devascularization with entire spleen separation

Treatment

Initial Management

- Central venous line for perfusion and monitoring.
- Transfusion of blood as needed.
- Antibiotics coverage.
- Urinary catheterisation.
- Nasogastric tube aspiration.

Surgical Management

Emergency splenectomy

- It is done through midline/left subcostal incision. Thoraco-abdominal extension of incision may be needed for rapid control of bleeding for injury to a large tropical spleen with severe bleeding. Other associated injuries should be looked for and dealt with (injury to left lobe liver/pancreas/ intestine/ colon).

Partial splenectomy (upper/lower)

- It can be done by retaining either of the upper or lower polar branches of the splenic artery.

Splenorrhaphy

- In especially clean incised wound, spleen can be salvaged by suturing the wound carefully with placement of gel foam, topical thrombin, absorbable mesh wrap over the wound. Suture repair, oxidised cellulose, debridement of lacerated spleen—are other methods used. Temporary occlusion of splenic artery is often needed during splenorrhaphy. 10% of splenic injuries undergo splenorrhaphy. Its application is getting reduced due to nonoperative approach in such patients
- But in class IV or V injuries splenorrhaphy is not possible.

Advantages of splenorrhaphy

- It avoids OPSI
- It avoids dead space and so prevents potential space for subphrenic abscess
- Cellular components of blood are better maintained

Salutary effect

- In traumatic rupture of spleen, spleen is cut into pieces of 0.5 cm and placed in omental pockets.

Nonoperative Management

- When facilities are available, splenic injury can be managed conservatively by non-operative management.
- Clinically close observation, serial haematocrit evaluation, serial CT abdomen/U/S abdomen at regular intervals to assess the progress or regress of the bleeding spleen has to be done.
- Absolute bed rest, sedation, antibiotic coverage and proper monitoring are needed.
- It is often supported by angiographic embolisation to improve the splenic salvage rate. But it may cause pain, splenic abscess or infarction.
- *Indications for nonoperative treatment*
  - Only splenic injury—no other associated injuries.
  - Grade I, II and III injuries.

Note:

- If patient shows features of haemodynamic instability—blood pressure below 90 mmHg, pulse rate above 120/minute, if patient does not show any positive response after 2 litre of crystalloid solution infusion, then nonoperative management should be abandoned.
- Nonoperative approach prevents possibility of OPSI, thrombocytosis causing possible *DVT*.
- 80% of children (due to more elastin in spleen and complete coverage with ribs in childhood) and 50% of adult with splenic injury are treated by nonoperative method.
- Failure rate in nonoperative management up to grade II is 15-20% which is gradually reducing to 5% in many centres. It is more in patients with age more than 55 years.
- Delayed splenic rupture and formation of splenic pseudoaneurysm are problems in nonoperative approach.
- Normal activity of the patient can be resumed only after 3 weeks following nonoperative treatment in grade I and II injuries; after 8 weeks in grade III or beyond usually with a confirmation by CT scan.

Complications of splenectomy

- Haemorrhage and shock
- Haematemesis
- Pancreatitis
- Pancreatic fistula
- Gastric dilatation
- Left sided pleural effusion
- Left sided colonic injury
- Severe sepsis
- Changes in cellular component of blood
- DIC

The best bridge between despair and hope is a good nights sleep—**E. Joseph Cossman**

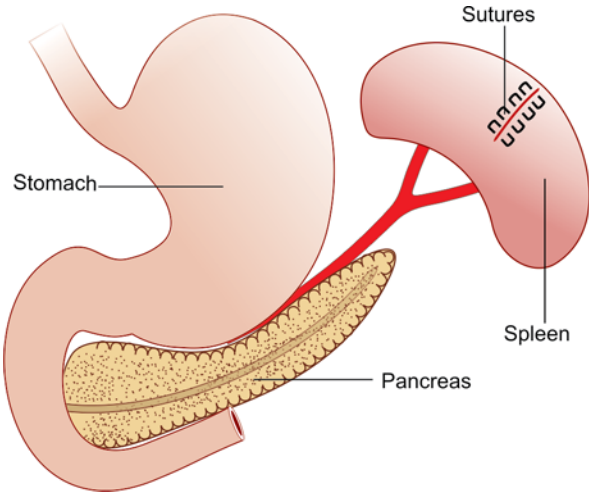

Fig. 13.7: Splenorrhaphy is done in incised wound.

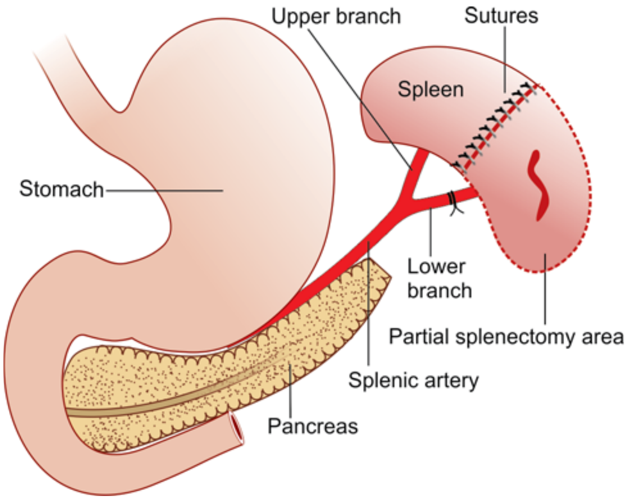

Fig. 13.8: Partial splenectomy is done when one of the branches of the splenic artery is ligated and other part is retained.

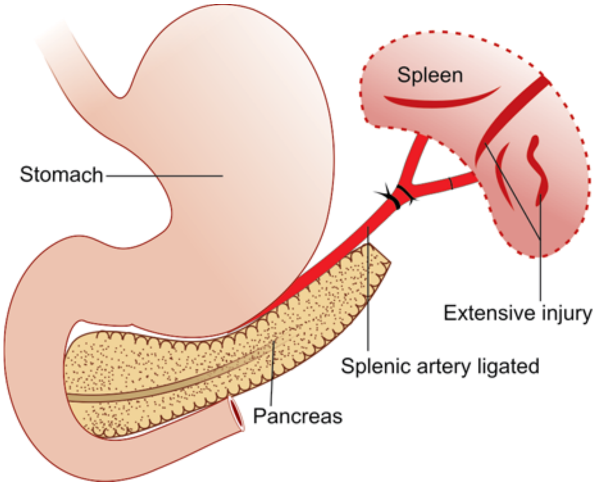

Fig. 13.9: Splenectomy is done in large wounds, lacerations, hilar injuries.

SPLENOMEGALY

Causes

1. *Infective:* Tuberculosis, splenic abscess, infectious mononucleosis, malaria, typhoid, kala azar.
2. *Blood diseases:* Chronic myeloid leukaemia, myelofibrosis, polycythaemia, hereditary spherocytosis, haemolytic anaemia, idiopathic thrombocytopenic purpura (ITP).
3. *Metabolic:* Gaucher's disease, amyloidosis, porphyrias,
4. *Portal hypertension.*
5. *Collagen diseases.*
6. *Malignancy*: Hodgkin's lymphoma, splenic sarcomas.

Masses which mimic enlarged spleen

- Kidney mass (Left sided)
- Retroperitoneal mass
- Left colonic mass
- Left adrenal mass

Clinical features of enlarged spleen

- Mass in the left hypochondrium
- Notch is felt
- Moves with respiration
- Dull to percuss
- Directed towards right iliac fossa
- Hook sign: Inability to hook under left costal margin
- Can not insinuate finger under left costal margin

HEREDITARY SPHEROCYTOSIS

It is an autosomal dominant disease effecting males and females equally.

Pathology

- Here there is an increase in red cell wall permeability to sodium. So sodium leaks into the red cells and it becomes spherical and more fragile. This leads to greater loss of membrane phospholipid resulting in weakening of the membrane with increase in energy and oxygen requirement. So these RBCs are destroyed in spleen causing haemolytic anaemia, haemolytic jaundice, unconjugated hyperbilirubinaemia, pigmented gallstones, cholangitis.
- Defect in components of red cell membrane like *spectrin (mainly),* ankyrin, band 3, or protein 4-2 causes weakness of red cell wall. This promotes the influx of sodium into the cell.
- Condition is common in North America.

Clinical Features

- Pallor, jaundice, recurrent fever, pain abdomen, splenomegaly, hepatomegaly, chronic leg ulcer. Gallstones are seen in 60% of cases.
- Spontaneous remission is uncommon.
- Patients are more prone for several infections.
- *Acute haemolytic crisis* precipitated by infection and stress; with features of acute pain abdomen, fever, vomiting and anaemia. It often mimics other acute abdominal conditions.

Investigations

- *Fragility test:* Here increased fragility of the erythrocytes is the typical feature. Haemolysis occurs in 0.6% or in even stronger solutions.

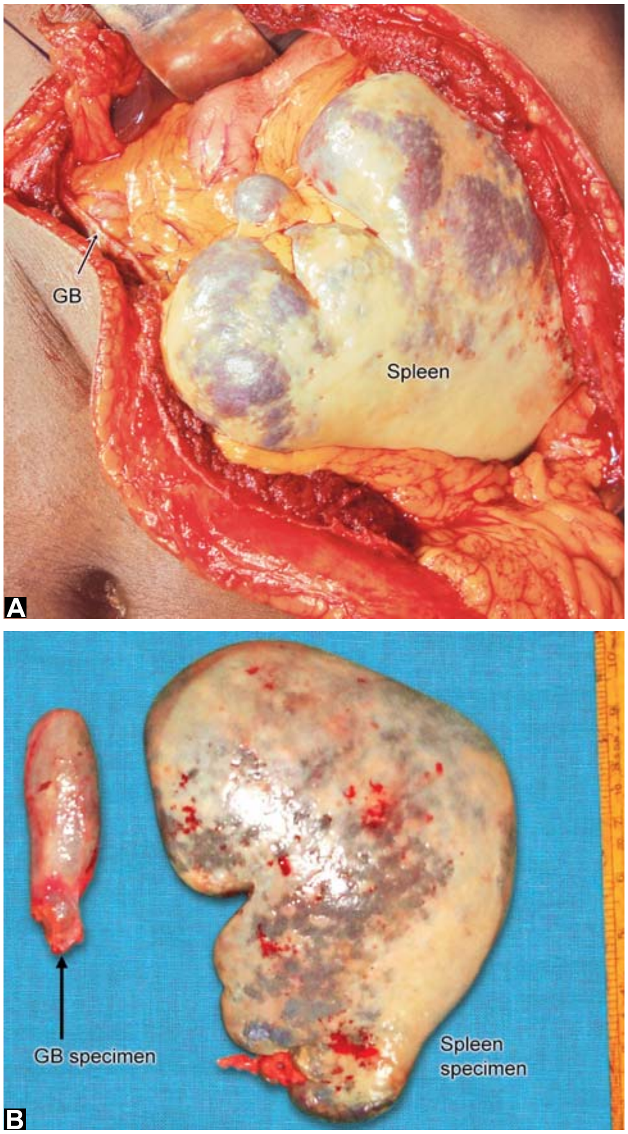

Figs 13.10A and B: Hereditary spherocytosis—on table and after surgery showing gallbladder and spleen specimens.

- LFT-↑ serum bilirubin.
- Reticulocyte count is increased significantly (25% increased).
- Faecal urobilinogen is increased.
- Labelled radioactive chromium shows faster red cell destruction.
- U/S abdomen is done to look for gallstones, spleen, liver, CBD.
- Peripheral smear shows spherocytes, haematocrit.

Problems with hereditary spherocytosis

- Haemolytic anaemia
- Haemolytic jaundice
- Pigmented CBD stones
- Cholangitis with Charcot's triad
- Gallstones (30%)
- Chronic leg ulcers

- Direct Coomb's test—negative.
- Radioactive chromium 51 labelled patients' red cells shows rate of sequestration of red cells in spleen.

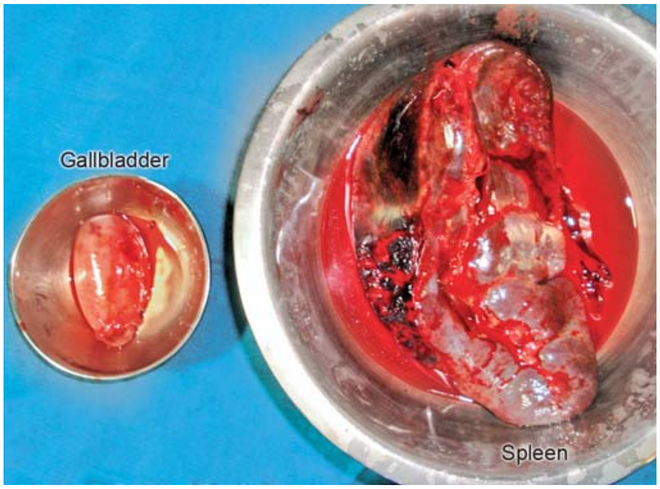

Fig. 13.11: Specimen of gallbladder and spleen (postsurgical) in case of hereditary spherocytosis.

Treatment

- Blood transfusion to improve the Hb status. Later splenectomy is done—Optimum age for splenectomy is 7 years.
- Accessory spleen should be removed.
- Splenectomy prevents early RBC destruction thereby controlling anaemia, jaundice and also prevents further formation of leg ulcers.
- Pneumococcal vaccine should be given to all these patients before elective splenectomy—3 weeks prior to surgery; 3 weeks after surgery and later once in 5-6 years.
- If there are gallstones, cholecystectomy is done.
- Both splenectomy and cholecystectomy can be done through *laparoscopy.*
- Splenectomy for hereditary spherocytosis is not done before the age of 6 years.

Note:

- *Hereditary elliptocytosis* is an autosomal dominant condition where 90% of RBCs are elliptical. Mostly they are asymptomatic, if symptomatic splenectomy cures the condition.

IMMUNE HAEMOLYTIC ANAEMIA

Here IgG or IgM antibodies bind to erythrocyte antigens causing red cell destruction through complement and RE systems.

It can be:

- ***Alloimmune haemolytic anaemia*:** It occurs after exposure to allogenic erythrocytes during transfusion, pregnancy or transplantation.
- ***Drug-induced immune haemolytic anaemia*:** It is drug-induced antibody reaction for drugs like α methyl dopa, penicillins or 2nd generation cephalosporins. It occurs

Autosplenectomy can occur in sickle cell anaemia.

via new antigen formation with positive Coomb's test. Cessation of causative drug and steroid therapy will cure the disease.

- *Autoimmune haemolytic anaemia (AIHA*): Here there is Development of IgG and IgM antibodies that bind to RBCs which are identified by Fc receptors of macrophages and RE cells causing phagocytosis and destruction of red cells. Two types of antibodies are:
  - *Common warm antibodies*: They react at 98.6°F. Often it is associated with CLL. In classical type, platelets are normal. Evan's syndrome is combination of AIHA and ITP. Coomb's test is positive (95% cases). Clinical features are of any haemolytic anaemias like palor, jaundice, splenomegaly (50%), and gallstones (25%). Longterm prednisolone is the treatment of choice. Failure to respond after long-term treatment is an indication for splenectomy. 80% of patients benefit with splenectomy.
  - *Less common cold antibodies (Cold agglutinin syndrome)*: It is treated by avoiding exposure to cold; folic acid; chlorambucil; cyclophosphamide and plasmapheresis. Splenectomy is not beneficial in this condition.

THALASSAEMIA (Mediterranean Anaemia/Cooley's Anaemia/Erythroblastic Target Cell Anaemia)

- It can be α, β and γ thalassaemia depending on defect in the synthesis of particular peptide.
- *Beta thalassaemia* is commonest. There is decrease in Hb-A and persistence of Hb-F. Homozygous, thalassaemia major is commonest type. Heterozygous, thalassaemia minor also can occur.
- Intracellular *Heinz bodies are typical*.
- Chronic anaemia, jaundice and splenomegaly are the features. Major type presents in first year of life with retarded growth, frontal bossing and prominent malar bones. Leg ulcers, recurrent infections, splenic infarction, pleural effusion, crisis, gallstones (25%) are other features.
- *Complications* are haemosiderosis, diabetes mellitus, pancreatitis, cirrhosis of liver, intractable sepsis and aplastic crisis.
- Haemoglobin electrophoresis, peripheral smear, LFT, bone marrow study and U/S abdomen are the investigations required.
- Regular blood transfusions, splenectomy are the treatment options.

SICKLE CELL DISEASE

- It is hereditary haemolytic anaemia seen commonly in blacks wherein HbA is replaced by *HbS* leading to formation of *crescent shaped* red cells which is deformed and more prone for trapping in spleen and destruction. Under low oxygen tension, HbS undergoes crystallisation causing elongated and deformed red cells. It increases the blood viscosity and causes stasis in circulation. Splenic micro infarcts are common causing initial splenomegaly and later autosplenectomy and calcification.
- Because of autosplenectomy patient is more prone for infection by pneumococcus and other organisms due to reduced production of antibodies.
- Patients are commonly asymptomatic.
- It is common in Africa. Patients in India have high HbF which protects them from developing symptoms.
- Anaemia, pain abdomen, leg ulcers are the usual features.
- Cerebral, pulmonary and mesenteric infarctions can occur which may be often life threatening.
- Peripheral smear and electrophoresis are diagnostic. U/S abdomen is needed to see spleen and for gallstones.
- *Treatment* is sodium cyanate (prevents sickling of HbS); hydration; partial exchange transfusion, antibiotics, symptomatic therapy by folic acid. Splenectomy has only limited role in initial sequestration stage.

IDIOPATHIC (IMMUNE) THROMBOCYTOPENIC PURPURA (ITP)

- It is development of *antiplatelet antibodies*, (IgG) which damage patient's own platelets. These platelets get sequestered earlier and faster in the spleen than normal causing thrombocytopenia. Spleen is the source of antiplatelet antibody production and destruction of formed platelet—antiplatelet antibody complex (by macrophage phagocytosis).
- Low platelet count, a normal bone marrow and absence of other causes of thrombocytopenia are the features. Bone marrow megakaryocytes are normal or increased but there is relative marrow failure in terms of compensatory platelet production for more destruction.
- In adult it is common in females (70%) who are young below the age of 40 years (70%). Chronic type is common in adults.
- In children it is equal in both sexes. Acute is common in children. Spontaneous remission (80%) is common. In girls it may become chronic continuing to adolescent period.
- Platelet count above 50,000/mm^3 usually will not cause spontaneous bleeding and undue bleeding is not common with invasive procedures. So specific therapy is not indicated. Regular platelet count estimation is needed.
- Regular careful observation is done clinically and haematologically for count between 30,000/- 50,000/mm^3 if there is no bleeding. If any spontaneous bleeding occurs or any invasive procedure is needed then platelet count should be corrected with steroids, platelet transfusion. With steroid/prednisolone therapy count becomes more than 50,000/mm^3 in 1-3 weeks.

Types

1. *Acute ITP:*
   - Initiated after viral infection.
   - Common in children (90%).
   - Spontaneous remission is common.
2. *Chronic ITP:*
   - Common in adults (90%).
   - Gradual, may be present for years.

Clinical features

- Common in young women (3:1 ratio)
- Purpuric patches in skin and mucous membrane (most common presenting sign)
- Epistaxis
- Menorrhagia
- Haematuria, haemarthrosis (rare)
- GIT bleeding
- Intracranial haemorrhage—2%; dangerous
- Tourniquet test is *positive* (*Hess test*)

- Purpuric and petechial patches are common in buttocks and limbs as these areas are more dependent having more vascular pressure. These patches and petechiae are mild or moderate.
- Splenomegaly is present in 25% of cases.
- After applying sphygmomanometer cuff and inflation (just below systolic pressure for 10 minutes), cubital fossa is observed for petechial haemorrhagic spots. More than 20 petechial spots in 3 cm circled area are significant—*Hess test*.

Differential Diagnosis

- Other causes for purpura—pregnancy, preeclampsia, drugs like heparin, quinidine, sulphonamides, HIV, rubella, infectious mononucleosis, myelodysplasia, congenital thrombocytopenia, thrombotic thrombocytopenic purpura, uraemia.

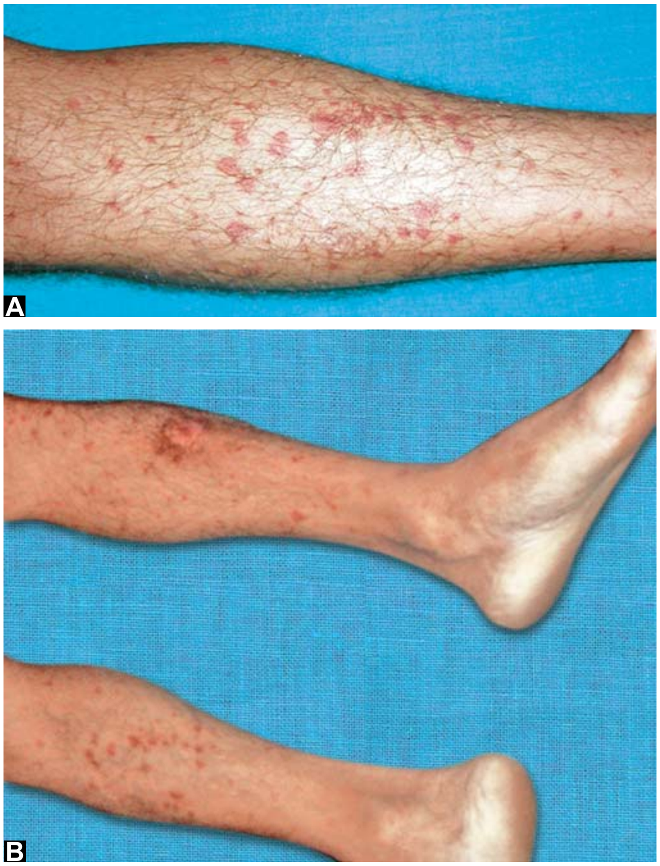

Figs 13.12A and B: Purpura spots in legs in case of ITP.

- Increased capillary fragility.
- Bone marrow suppression due to aplastic anaemia.
- Chemotherapy.
- DIC.
- Autoimmune diseases, SLE.

Investigations

- Bleeding time is increased times.
- Clotting and prothrombin time are normal.
- Platelet count is decreased.
- Bone marrow biopsy reveals increased megakaryocytes.
- Usually anaemia and neutropenia is not observed.
- U/S shows splenomegaly only in 25% cases.

Note:

Positive Hess test is formation of > 20 purpuric patches in an area of 3 cm diameter.

Treatment

Drugs

- In children, spontaneous regression occurs in 75% of cases after one attack.
- Short courses of *steroids* or *azathioprine* hastens the recovery.
- In adults, treatment is indicated when platelet count is < 30000/mm^3.
- Glucocorticoid—1 mg/kg is given. Response to treatment is seen within a week. Prednisolone 10 mg/day for 6 weeks is given. Response to steroid (prednisolone—1 mg/kg body weight) therapy showing complete remission is 25%.
- Danazol 200 mg TID decreases IgG receptor on phagocytes causing lesser degradation of platelets.
- Vincristine 2 mg per week for 6 weeks.
- IV immunoglobulin 0.4-1.0 g/kg × 5 days (action like danazol). It is very useful in acute bleed, in preparing the patient for surgery, in conducting delivery in pregnancy with ITP. Dose should be given for at least two days before taking up for surgery. It is also used in preparing ITP patient for therapeutic splenectomy.
- Anti-Rh D antibodies.

Indications for admission

- Platelet count less than 30,000/mm^3
- Platelet count less than 50,000/mm^3 showing bleeding episodes/mucosal bleed
- Risk factors association like existing peptic ulcer, hypertension.
- Life threatening haemorrhage

Indications for splenectomy

- Relapsed severe cases after treatment with steroids—for 6 weeks
- Girls attaining menarche
- ITP persisting even after 6-9 months refractory to treatment
- In 2nd trimester pregnancy, with platelet count <10000/mm^3 and severe bleeding problems

- ❖ Recurrence after stopping/tapering steroids
- ❖ Profound GI bleed and intracranial bleed
- ❖ ITP with persistent platelet count below 10,000/ mm^3 for 6 weeks whether bleeding present or not, ITP with only transient response to drugs for 3 months with platelet count less than 30,000/mm^3

Note:

- Splenectomy is not indicated in nonbleeding ITP for 6 months with platelet count above 50,000/ cu mm^3.
- Response rate is 80%; it is usually achieved in first 10 days of splenectomy.
- Relapse of ITP after splenectomy is 10%. It is probably due to hepatic sequestration of platelets in such patients or presence of accessory spleens (10% of such cases). Such patients are difficult to treat. They need prednisolone, azathioprine (single agent—needs 4 months to achieve initial response in such patients), cyclophosphamide (single drug).
- Presence of accessory spleen is confirmed by blood smear and radionuclide imaging but exact anatomical number and location of accessory spleen is often difficult to identify preoperatively. Thorough exploration for possible all sites of splenunculi should be done to achieve good success rate.
- 15% of HIV patients develop ITP; splenectomy response is good in these patients; splenectomy does not progress HIV status.

Surgery

- ♦ *During surgery and in emergency,* fresh blood transfusions or platelet transfusions are required. Sometimes anti-platelet immune response is temporarily blocked by IgG transfusion so to allow the platelet count to rise at the time of surgery or in cases of severe bleeding or in pregnancy.
- ♦ Accessory spleen should be identified and removed.
- ♦ In 7 days platelet count raises to more than 100000/ cu mm.
- ♦ Laparoscopic approach is ideal in elective splenectomy for ITP.

Treatment for ITP

- ❖ Methylprednisolone IV for 3-5 days only
- ❖ Oral prednisolone tablets—drug of choice for 6-12 weeks
- ❖ IV immunoglobulin 0.4-1.0 gm/kg/day for 5 days only. It competes with antibodies
- ❖ Vincristine 2 mg/week for 6 weeks
- ❖ Danazol 200 mg tid
- ❖ Anti-Rh D antibodies
- ❖ Azathioprine
- ❖ Splenectomy
- ❖ FFP, platelet concentrate, fresh blood transfusions.

THROMBOTIC THROMBOCYTOPENIC PURPURA (TTP)

It is an idiopathic disease associated with bone marrow transplantation, drugs like penicillin, mitomycin, cyclosporine.

Main features are (pentad:

1. Microangiopathy haemolytic anaemia.
2. Thrombocytopenia.
3. Fever.
4. Neurological disturbances.
5. Renal dysfunction.

Features and management

- ♦ Here only arterioles and capillaries of microcirculation are involved. Venules are not involved. Platelet microthrombi, endothelial proliferation are the pathological features.
- ♦ Anaemia, platelet trapping in the spleen and thrombocytopenia are typical.
- ♦ Altered mental status, neurological deficits and renal failure are common.
- ♦ Plasmapheresis improves the survival which is the main treatment.
- ♦ Laparoscopic splenectomy is done only to reserved patients.

Felty's syndrome is chronic rheumatoid arthritis, mild leukopenia and splenomegaly.

SPLENECTOMY

Indications for Splenectomy

1. Trauma.
2. In radical gastrectomy—removed "*en bloc*" with stomach.
3. Staging laparotomy in case of Hodgkin's lymphoma.
4. Hereditary spherocytosis and other haemolytic anaemias.
5. Idiopathic thrombocytopenic purpura (ITP).
6. Portal hypertension with variceal bleeding.
7. Tuberculosis of spleen.
8. Others:
   - Gaucher's disease.
   - Chronic myeloid leukaemia.
   - Felty's syndrome.
   - Carcinoma of tail of the pancreas during distal pancreatectomy.
   - Schistosomiasis.
   - Splenic artery aneurysm.
   - Splenic infarction.
   - Cystic disease of spleen.
   - Splenic tumours.

Indications for splenectomy

Benefit is maximum/very good	***Benefit is equivocal***	***Benefit may be there/low***
• Splenic injury • Hereditary spherocytosis • ITP • Splenic cysts • Splenic tumours • Hypersplenism • Segmental left sided portal hypertension	• AIHA (autoimmune haemolytic anaemia) • Tropical splenomegaly • Felty's syndrome	• Thalassaemia • CML • Gaucher's disease • Hodgkin's lymphoma for staging laparotomy • Sickle cell anaemia

Procedure

- ♦ Incision—left paramedian or left subcostal or in case of large adherent spleen left thoracoabdominal approach.

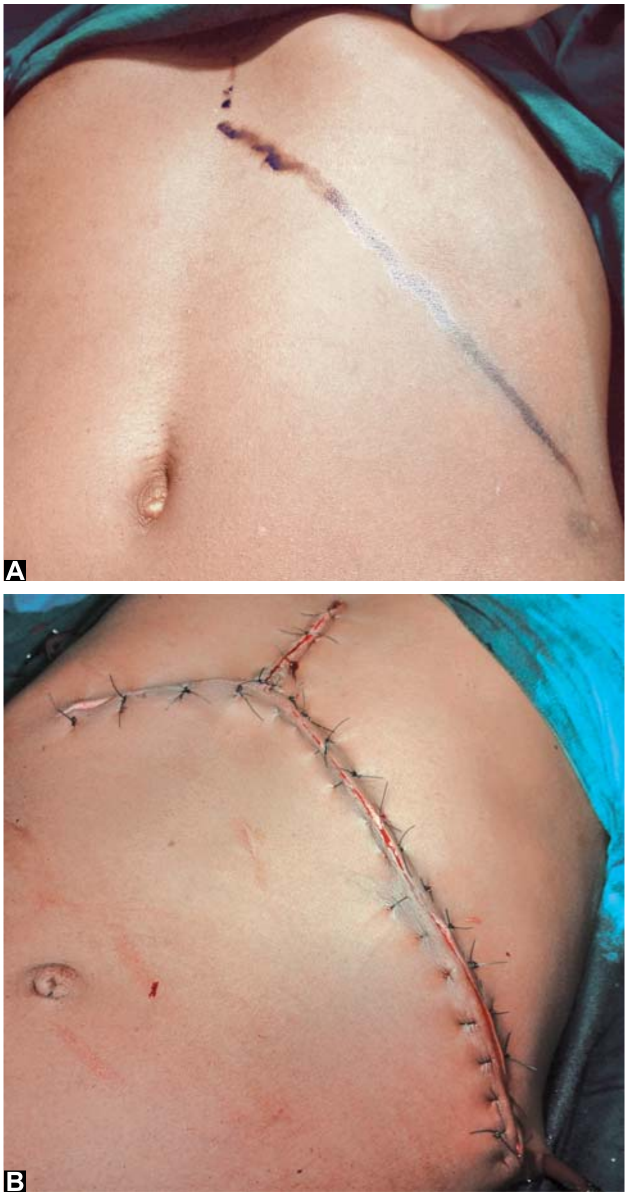

Figs 13.13A and B: Incision for open splenectomy – left subcostal. Often in a massive spleen or adherent spleen incision need to be extended to opposite side and towards xiphisternum.

- Lieno-renal ligament and short gastric vessels are ligated. Splenic vessels are ligated at the hilum of spleen. Spleen is removed. Abdomen is closed after placing a drain *in situ*. Presently, drain is *indicated* only if there is injury to tail of pancreas during splenectomy.
- *Presently laparoscopic splenectomy is becoming popular.*
- Preoperative splenic artery embolisation is beneficial in massive spleens.

Complications of Splenectomy

- Haemorrhage—mainly from short gastric vessels. Re-exploration and control of bleeding is needed.
- Gastric dilatation.

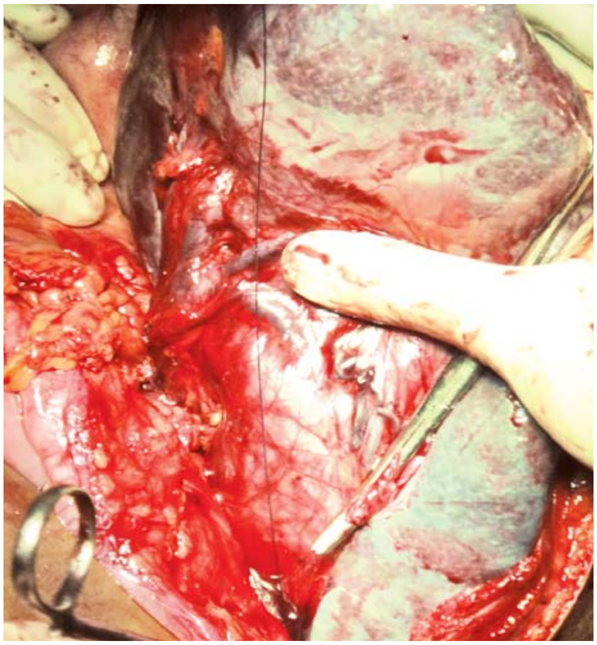

Fig. 13.14: Splenic vessels should be ligated individually securely with double ligatures in large spleen during splenectomy.

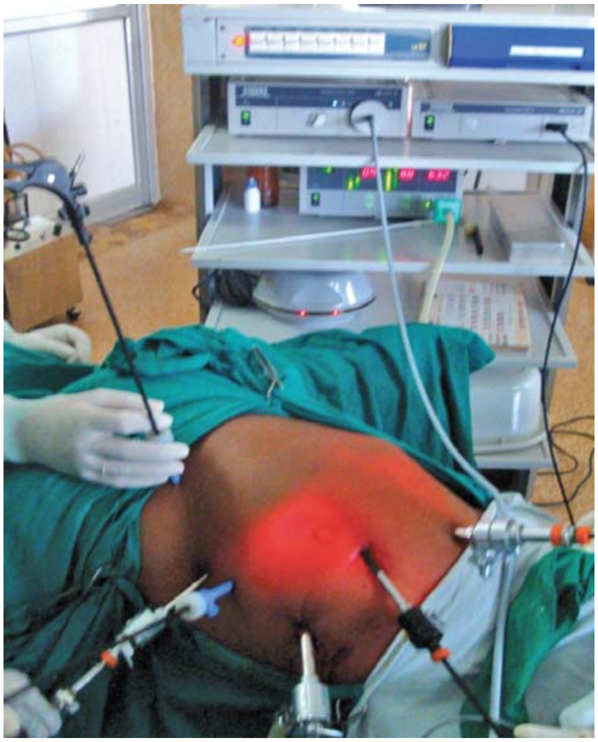

Fig. 13.15: Laparoscopic splenectomy is also a good alternative in many splenectomies when splenomegaly is not massive. It is not useful in massive spleens and trauma.

- Haematemesis due to mucosal congestion following ligation of short gastric vessels.
- Left sided pleural effusion and left basal atelectasis (left lobe atelectasis is the commonest complication) (30%).
- Left subphrenic abscess.

Efforts and energies should be consistent, constructive and compassionate.

- Pancreatitis following damage to tail of the pancreas and pancreatic fistula (3%).
- Haematologic changes: Rise in WBC and platelet count, rise in abnormal RBCs and RBC bodies.
- Gastric fistulas.
- Left sided colonic injury.
- Infection—Postsplenectomy septicaemia, OPSI—most dangerous.
- Portal vein thrombosis can occur.
- *DVT as a late sequelae* after splenectomy is often dangerous. Risk is 4 times more than in non-splenectomised people and so also pulmonary embolism. It is better to put them on long term small dose of *aspirin* to reduce the incidence of thromboembolism.

OVERWHELMING POSTSPLENECTOMY INFECTION (OPSI)

- As there is reduced IgM, tuftin, properdin and other antibodies, phagocytosis of encapsulated bacteria is defective. So the postsplenectomised patient is more prone for *Pneumococcal septicaemia* (commonest), *N. meningitides, H. influenzae, Babesia microti* infections. Risk is more in patients on chemotherapy, radiotherapy and haemolytic diseases.
- Incidence is 4%.
- Common in first two years after splenectomy. But it can occur any time. Risk persists for lifetime.

Clinical Features

- Prodromal phase—fever, chills, sore throat.
- Hypotension, shock.
- DIC.
- Respiratory distress, coma, death.

Mortality for fully developed OPSI—50-70%.

Prevention

- *Pneumococcal vaccine* should be given to all splenectomised patients. Polyvalent pneumo-vac is given 2-3 weeks prior to surgery and soon after recovery from surgery and it is repeated once in 5 years (Given to patients older than 2 years).
- Other vaccines advised are meningococcal vaccine (only to those who travel with high-risk), *H. influenzae* 'B' vaccine (to all whatever the age, once in 10 years).
- In malaria endemic areas, antimalarial prophylaxis is given for patients after splenectomy.

Treatment of OPSI

- ❖ Antibiotics like Cefoperazone, Ceftazidime, Amikacin
- ❖ Ventilatory support—ICU care
- ❖ Blood transfusion
- ❖ Immunoglobulin transfusion
- ❖ Nutrition (TPN) and maintaining of urine output

Prevention of OPSI

- ❖ Life long prophylaxis using benzathine penicillin 12-24 lac units—controversial in adults
- ❖ Pneumococcal vaccine given 2-3 weeks prior to splenectomy—70% protection
- ❖ *H. influenzae* B type vaccine
- ❖ Meningococcal vaccine is given only to high-risk groups, as its effects are short term. So it is not routinely given
- ❖ *Note: Protection by vaccine is not completely guaranteed*

Note:

- OPSI is more common when splenectomy is done for malignancy or haematological diseases or done in young individuals.
- Usage of prophylactic penicillin (benzathine) is controversial in adult even though it is beneficial in children up to 16 years.
- OPSI can occur at any time after splenectomy. Drugs like erythromycin, amoxicillin are also used as prophylactic agents.
- Mortality in fully developed OPSI (hypotension, DIC, MODS) is 60%—very high.
- Patients who survived after OPSI with therapy may develop residual problems like – gangrene of limbs, deafness, meningitis, mastoid osteomyelitis, heart valve destruction.
- 7% of OPSI cases have S pneumonia sepsis. Others are—*N. meningitidis*, *H. influenzae*, *Capnocytophaga canimorus* after dog bite.
- *PPV 23* is the commonest vaccine used. Revaccination is done once in 5 years; decline in pneumococcal antibody level may be the criteria for revaccination.
- Preservation of *25% of spleen* is sufficient to protect against pneumococcal infection. So partial splenectomy is definitely helpful.
- Splenic autotransplantation into the omental pouch is under trial at present to reduce the chance of OPSI.

SPLENIC ARTERY ANEURYSM

- Most common intraabdominal site for aneurysm after abdominal aorta.
- Most common in women, due to atherosclerosis.
- Congenital splenic artery aneurysm can occur.
- Splenic artery pseudoaneurysm (10%) can occur after an attack of acute pancreatitis which may cause life threatening rupture.

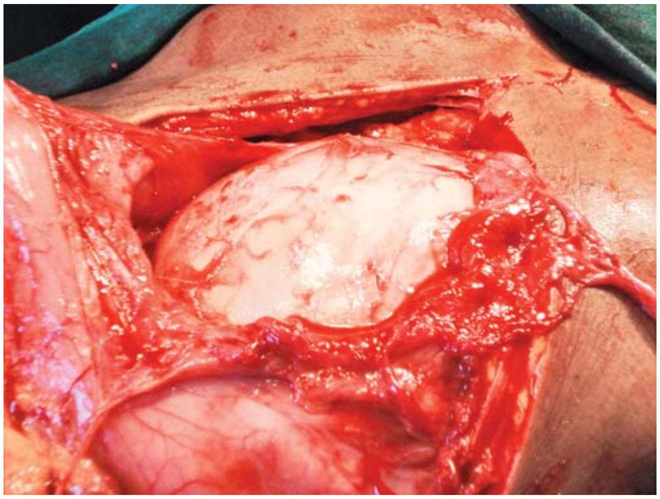

Fig. 13.16: Splenic artery aneurysm. It is the commonest site of aneurysm in the abdomen after aorta.

- Rupture of splenic artery may be intraperitoneal or into the stomach, colon and intestine. Calcified aneurysm rarely ruptures.
- Plain X-ray shows calcified lesion in 70% of cases.
- Contrast CT, angiogram are needed.
- Indication for surgery:
    - >2 cm diameter.
    - Symptomatic aneurysm.
    - Anticipating pregnancy.
    - Aneurysm identified during pregnancy should be operated before 3rd trimester.

Treatment

- Splenectomy with ligation of artery proximal to aneurysm.
- Splenic artery embolisation under radiological guidance is also beneficial.

SPLENIC ABSCESS

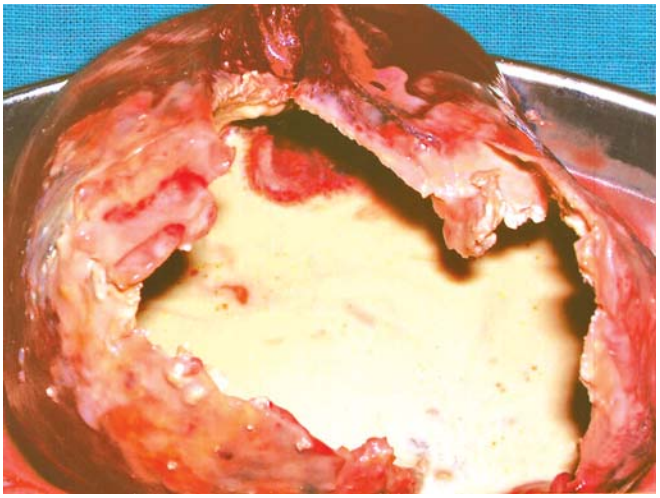

Fig. 13.17: Splenic abscess due to tuberculosis.

- It is uncommon but potentially fatal.
- Incidence is 0.7%.
- **Precipitating factors**:
    - Endocarditis is commonest focus.
    - AIDS.
    - Polycythaemia.
    - Malignancy.
    - Sickle cell disease.
    - Splenic vein thrombosis.
    - Trauma.
    - Drug abuse.
    - Typhoid, osteomyelitis, puerperal sepsis, pancreatitis, otitis media.

Organisms responsible: *Staphylococcus*, *Streptococcus*, *Enterobacter*, *E. coli, Proteus, Klebsiella, Candida albicans, Mycobacterium tuberculosis* in developing countries.

Clinical features

- Fever
- Pain in left hypochondrium
- Pain in left side of chest
- Left sided pleural effusion
- Spleen when palpable is tender
- Single or multiple abscesses
- Need not move with respiration

Investigation

Ultrasound, ESR, chest X-ray, CT scan, blood culture.

Treatment

- High dose antibiotics has to be started.
- ATT in case of tuberculosis.
- Ultrasound/CT scan guided aspiration of abscess is done.
- Splenectomy.

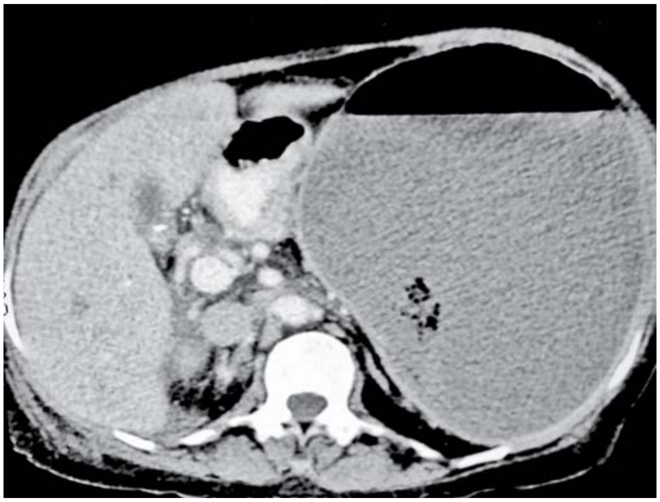

Fig. 13.18: CT scan showing abscess in the spleen.

Note:
- Splenic abscess can rupture causing peritonitis and left subphrenic abscess.
- Splenic abscess has 80% mortality in immunocompromised patients and 20% in an otherwise normal individual.
- 2/3rd of splenic abscesses in adults are single and 1/3rd are multiple; in children 1/3rd are single, 2/3rd are multiple.
- Solitary abscess responds well for guided aspirations; whereas multiple abscesses need splenectomy.

HYPERSPLENISM

It is overactivity of spleen causing pancytopenia with more of hypercellular bone marrow.

Causes

- Primary hypersplenism.
- Portal hypertension.
- Malaria, kala-azar.

- Schistosomiasis.
- Tuberculosis.
- Myeloproliferative disorders.

Hypersplenism often requires splenectomy.

Note:
- Commonest primary splenic neoplasm is haemangioma
- *Littoral cell angioma* is rare splenic tumour
- Lymphangioma, inflammatory pseudotumours are rare splenic tumours
- Hodgkin's lymphoma commonly involves spleen
- Secondaries in spleen are rare. Occasionally spread can occur from lungs, stomach, pancreas, breast, melanoma
- Primary malignant splenic tumours are exceedingly rare. Angiosarcoma is the most common primary tumour
- *Wandering spleen* is due to failure to form suspensory ligaments of spleen so that the spleen is *not* fixed in position. It is due to failure of dorsal mesogastrium to fuse with the posterior abdominal wall. It shows long splenic pedicle. It is prone for torsion and recurrent ischaemia. It is common in females. CT scan is diagnostic. If there is no ischaemia, *splenopexy* is done; if there is ischaemia which is confirmed by absence of perfusion in CT scan, splenectomy is done

SPLENIC CYST

Splenic cysts can be –

1. ***True primary cyst (20%).***
   It can be -
   - *Non-parasitic* [simple congenital cysts, epidermoid cysts (75% of non-parasitic cysts)]. It is common in 3rd decade, usually lined by squamous epithelium. Even though they are truly benign without malignant potential, they often show +ve for Ca 19/9 and CEA. Usually they are asymptomatic, but may present with splenic mass with pain, chest pain and left shoulder pain usually when cyst attains more than 8 cm in size. Occasionally sepsis, rupture and bleeding in a cyst can occur.
   - *Parasitic* (hydatid) – one should look for associated liver hydatids. CT abdomen confirms the diagnosis. It is treated with splenectomy; care should be taken not to spill or rupture during surgery.
2. ***Pseudocyst/secondary cysts*** (80%) are due to old trauma, malaria, tuberculosis, infectious mononucleosis, old parenchymal haematoma and syphilis. They are not epithelial lined; they are smooth, unilocular, thick walled. Initially treatment is guided percutaneous drainage or partial splenectomy. Splenectomy is done when entire spleen is involved with pseudocyst.

Chapter 14 Pancreas

CHAPTER OUTLINE

- Anatomy
- Serum Amylase
- Magnetic Resonance Cholangiopancreatography
- Pancreatitis
- Acute Pancreatitis
- Complications of Acute Pancreatitis
- Pseudocyst of Pancreas
- Chronic Pancreatitis
- Pancreatic Tumours
- Exocrine Pancreatic Tumours
- Carcinoma Pancreas
- Endocrine Pancreatic Tumours
- Insulinomas
- Gastrinomas
- Glucagonomas
- Zollinger–Ellison Syndrome
- Cystic Fibrosis
- Annular Pancreas
- Ectopic (Accessory) Pancreatic Tissue
- Pancreatic Divisum
- Pancreatic Calculus
- Pancreatic Ascites
- Pancreatic Fistulae
- Pancreatic Abscess
- Pancreatic Necrosis
- Pancreatic Trauma
- Cystic Lesions of Pancreas
- Exocrine Pancreatic Disease

Pancreas is also called as *abdominal tiger.*

ANATOMY

Pancreas is an elongated retroperitoneal organ; 15-20 cm in length; lies against L_1–L_2 vertebra. It lies posterior to stomach, separated by lesser sac.

Parts

It is divided into head, neck, body, tail.

- The head lies in the concavity of duodenum and tail reaches the hilum of the spleen.
- The posterior surface of the neck of pancreas is related to terminal part of superior mesenteric vein and beginning of portal vein.

Ducts of Pancreas

1. Main duct of pancreas (*Duct of Wirsung*): It begins in the tail of pancreas and runs on the posterior surface of the body and head of pancreas, receives numerous tributaries at right angle along its length (*Herring bone pattern*).

   It joins the bile duct in the wall of the second part of duodenum to form hepatopancreatic ampulla *(of Vater*) and opens on the summit of major duodenal papilla (8-10 cm from pylorus).

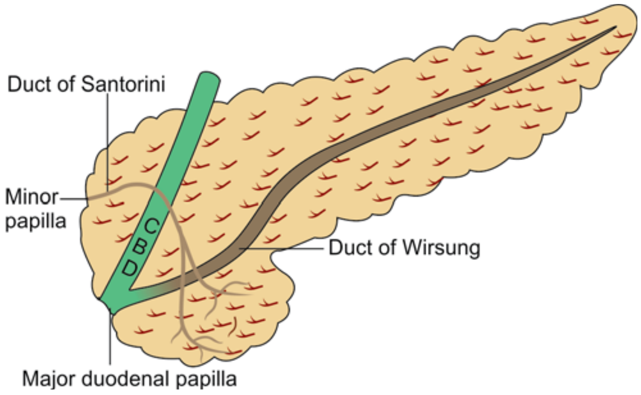

Fig. 14.1: Ducts of pancreas.

2. Accessory pancreatic duct *(Duct of Santorini):* It begins in the lower part of the head and opens into the duodenum at minor duodenal papilla (6-8 cm from the pylorus).

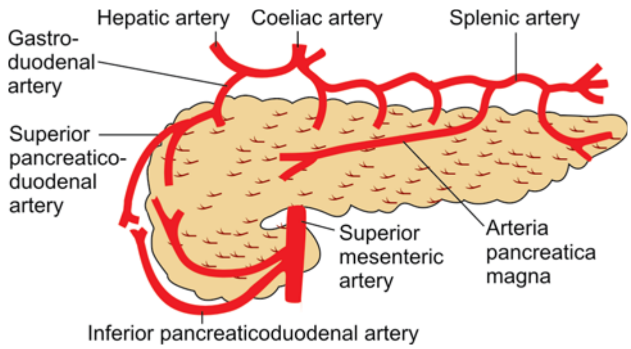
Hepatic artery
Coeliac artery
Splenic artery
Gastro-duodenal artery
Superior pancreatico-duodenal artery
Superior mesenteric artery
Arteria pancreatica magna
Inferior pancreaticoduodenal artery

Fig. 14.2: Arterial supply of pancreas.

Blood Supply of Pancreas

1. Pancreatic branches of splenic artery.
2. Superior pancreaticoduodenal artery.
3. Inferior pancreaticoduodenal artery.

Venous drainage is into portal vein.

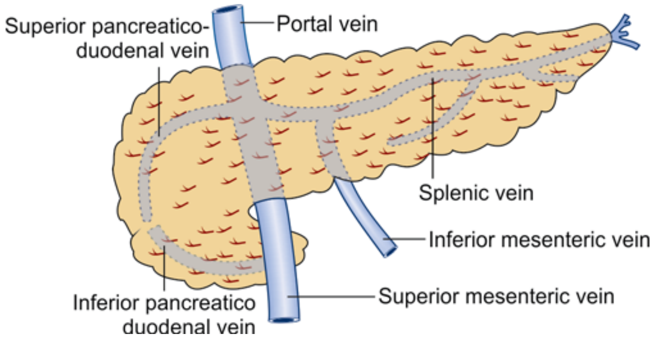

Fig. 14.3: Venous drainage of pancreas.

Nerve Supply

Parasympathetic supply is from vagus and sympathetic innervation is from splanchnic nerves.

Functions

- *Exocrine* part secretes pancreatic juice which helps in digestion of proteins, carbohydrates and fats.
- *Endocrine* part constitutes islets of pancreas which is distributed more numerous in tail of pancreas.
  β cells of islets secrete insulin.
  α cells secrete glucagon.

Normal pancreatic juice is colourless bicarbonate rich fluid containing around 15 gram of protein in total, 2.5 litres secretion/day. It alkalises duodenal content and helps digestion. Inactive proenzymes secreted into the duodenum from pancreatic juice are activated by trypsin in duodenum. Amylase and lipase can also be secreted in active forms. Basal secretion of these enzymes is low at rest; rapidly increases by hormonal and neural stimulation. It is controlled by secretin (cAMP) and cholecystokinin (phospholipase C, calcium). Protein part of the juice is secreted by *acinar cells*. *Duct cells* secrete fluid and electrolytes. Pancreatic secretion is very low in resting phase. During eating, cephalic phase stimulates 10% of pancreatic secretion through peripherally mediated acetylcholine; gastric phase stimulates further 15% of pancreatic secretion through gastrin release and vagal stimulation; finally in main intestinal phase, 75% of the pancreatic juice is secreted by release of secretin due to duodenal acidification, and by release of bile and cholecystokinin following the entry of fat and proteins in duodenum.

SERUM AMYLASE

- Normal value is up to 200-250 Somogyi units.
- It increases more than 1000 units or a rising titre is very significant in acute pancreatitis.
- Its half life is 24 hours.
- It is not very sensitive because it also increases in other conditions:
  - Salivary gland diseases.
  - Mesenteric ischaemia.
  - Ruptured aortic aneurysm.
  - Intestinal obstruction.
  - Ectopic gestation.
  - Salpingitis.
  - Perforated duodenal ulcer.

Types

- ***Amylase isoenzyme-P*** is specifically increased in pancreatitis.
- In other conditions, ***Amylase-S*** is commonly increased.

- It can be analysed with other proteolytic enzymes like lipase, elastase which are more relevant.
- In peritoneal tap, if amylase level of peritoneal fluid is more than serum amylase, then it is significant.

Note:
Amylase level in pseudocyst fluid is very high.

- Magnitude of amylase elevation is not related to severity of the pancreatitis. It increases once symptoms appear for 12-24 hours then decreases in 6 days. Persistent elevation suggests complications like pseudocyst, ascites and abscess formation.
- 10% cases of severe pancreatitis with necrosis show normal serum amylase level. This is due to necrosis of most of the acinar cells retaining hardly any functioning cells at time.
- Rise in pancreatic amylase will be very high in comparison with nonpancreatic causes of rise where it is only 2-3 times the normal.
- Raised urinary amylase level persists for longer period then serum amylase level.
- Amylase enters the lymphatics and circulation from the basal part of acinar cells and also, weakened intercellular adhesions in pancreatitis allow amylase to seep into the space to enter circulation.
- Circulating amylase inhibitors often when present, mask the true amylase level in pancreatitis.
- Occasionally amylase binds with large abnormal circulating albumin like protein which cannot be cleared normally causing false raise in serum amylase level (false positive) even though there is absence of pancreatitis. This *macroamylasaemia* is 0.5% common.

MAGNETIC RESONANCE CHOLANGIO-PANCREATOGRAPHY (MRCP)

It is a non-invasive diagnostic method using magnetic field and energy as resonance.

T_1 weighted images are used for pancreas.

T_2 weighted images are used for biliary tree.

Both images are used to see invasion and adjacent structures.

Advantages

- It is non-invasive.
- It gives equal or better imaging than ERCP.
- Pancreatic divisum or annular pancreas are identified easily by MRCP.
- Ductal dilatation in chronic pancreatitis and visualization of duct beyond stricture is possible.
- Islet tumours are better visualised and diagnosed by MRCP.

Disadvantages

- Availability.
- It is only diagnostic. Not therapeutic (ERCP is both diagnostic as well as therapeutic).
- Costly.

PANCREATITIS

> *We learn that pain was an early symptom in nearly one-half of the cases; that it was usually severe, and might be intense, and was to be found in the abdomen or lower chest.…The appearances found after death are conspicuously the hemorrhage within and near the pancreas.…It is evident that all treatment, at the onset, can be nothing but palliative.*
>
> —**Reginald Heber Fitz,** 1889

Pancreatitis is inflammation of the pancreas acute, chronic or relapsing which may lead into complications. It is one of the most devastating conditions in the abdomen.

Classification

I. *Marseilles' classification*:
   1. Acute pancreatitis.
   2. Acute relapsing pancreatitis.
   3. Chronic relapsing pancreatitis.
   4. Chronic pancreatitis.

   Acute means acute inflammatory changes which are reversible.

   Chronic means chronic continuous inflammatory changes which are morphologically irreversible.

II. *Trapnell's aetiological classification*:

Major causes

- Biliary tract disease 50%—stones—common cause
- Alcoholism 25%

Other causes

- Trauma
- After biliary, gastric, splenic surgery, ERCP
- Hyperparathyroidism
- Hypercalcaemia, hyperlipidaemia
- Diabetes
- Porphyria
- Drugs: Steroids(?), INH, diuretics (thiazides), septran, azathioprine, 5-amino salicylic acid, estrogens, tetracycline, valproic acid, petamidine – 5%
- Viral infections (mumps, coxsackie)
- Autoimmune diseases
- Vascular diseases
- Idiopathic
- Scorpion venom—Common in Trinidad
- Biliary ascariasis, Clonorchis sinensis
- Mycoplasma pneumoniae
- Infectious mononucleosis
- Pancreatic divisum

ACUTE PANCREATITIS

Acute pancreatitis is the development of acute inflammation of the normally existing pancreas. It may be *first* attack or *relapsing* attacks with an apparently normal gland in between. ***Biliary tract disease*** is the ***commonest*** cause for acute pancreatitis.

Pathogenesis

- *Present concept* is the biliary stone induced *pancreatic ductal* obstruction and ductal hypertension as etiological factor (*Pancreatic duct obstruction* occurs due to biliary tract stones (*commonest*), duodenal ulcer, duodenal Crohn's, periampullary diverticulum/tumour, trauma, pancreatic duct stricture, pancreatic divisum, ascariasis, Clonorchis sinensis).
- OR *alcohol* causes direct toxicity, hypersecretion of gastric and pancreatic juices, reflux, plugging of pancreatic proteins, injury by release of free radicals like superoxide and hydroxyl, spasm of Oddi and stimulates trypsinogen.

All these *aetiological* factors

↓

Either cause spasm of sphincter of Oddi and lead into the parenchyma (*common channel theory*), OR causes increased secretion of pancreatic enzymes

↓

Trypsinogen gets activated forming trypsin which activates other enzymes.

↓	↓
Proelastase to elastase.	*Prolipase to lipase.*
↓	↓
Causes capillary rupture.	Metabolises triglycerides to glycerol + fatty acids and fatty acids combine with calcium forming saponified fat.
↓	↓

Sequestered fluid, saponified fat, blood, toxins all together forms a ***chicken broth fluid.***

Surgery is a painful procedure to relieve the pain.

↓

Lecithinase, amylase and factors like lysolecithinase, prostaglandins, bradykinins, kallikrein, myocardial depressant factor, free radicals, cytokines, tumour necrosis factor, and platelet activation factor are also released to have local and systemic effects.

↓

Infection occurs causing bacteraemia, septicaemia.

↓

Large volume of fluid sequestration occurs causing *hypovolemic shock.*

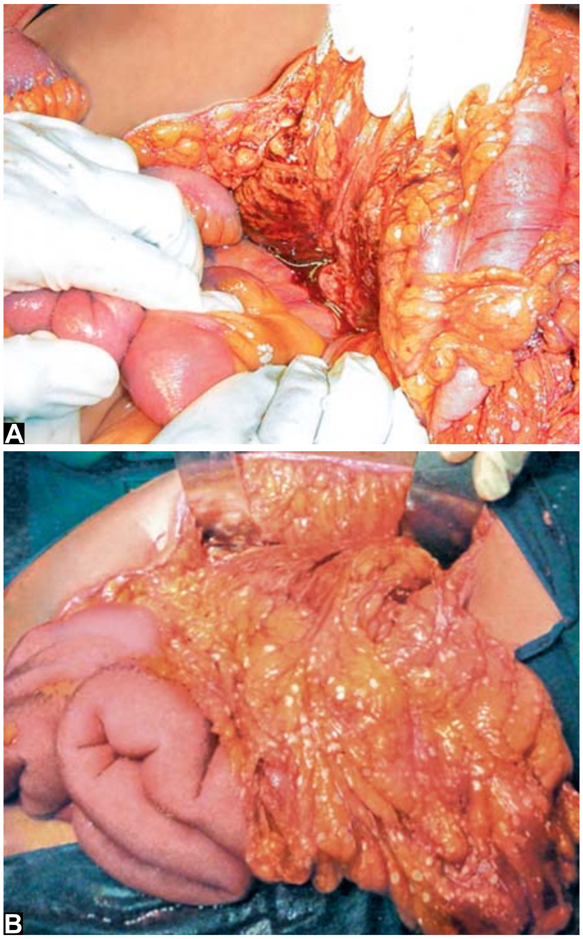

Figs 14.4A and B: Features of acute pancreatitis with necrosis in the omentum and over pancreas; with haemorrhage.

- Toxins released may lead to *acute tubular necrosis and so acute renal failure.*
- Left sided diaphragm gets elevated and *left sided pleural effusion* occurs.
- Lecithinase reduces the surfactant in the alveoli of lung, and infection leads to pulmonary insufficiency, *ARDS and respiratory failure.*
- Because calcium is utilised for saponification, *hypocalcaemia* sets in.
- Diffuse oozing in pancreatic bed occurs which utilizes platelets and causes *disseminated intravascular coagulation (DIC).*
- In severe cases, extensive necrosis with haemorrhage occurs causing *acute haemorrhagic necrotising pancreatitis (Fulminant pancreatitis), which has got a high mortality.*
- Here enzymes seep across the retroperitoneum causing haemorrhagic spots and ecchymosis in the flanks ***(Grey Turner's sign)***, or through falciform ligament causing discolouration around the umbilicus ***(Cullen's sign)***, umbilical black eye or below the inguinal ligament **(*Fox sign*).**
- SIRS and later MODS set in.
- *Sequence of events*:

Acute oedematous pancreatitis →→→→→→Severe acute pancreatitis → Acute sterile necrosis → Acute infected necrosis

↓ ↓ ↓ ↓

Resolution/recovery/complications (rare)
Less mortality

Local or systemic complications with high mortality.
(pseudocyst/abscess/ARDS/MODS/haemorrhage, etc.)

- **Moynihan described acute pancreatitis as** *"The most terrible of all the calamities that occur in connection with abdominal viscera".*

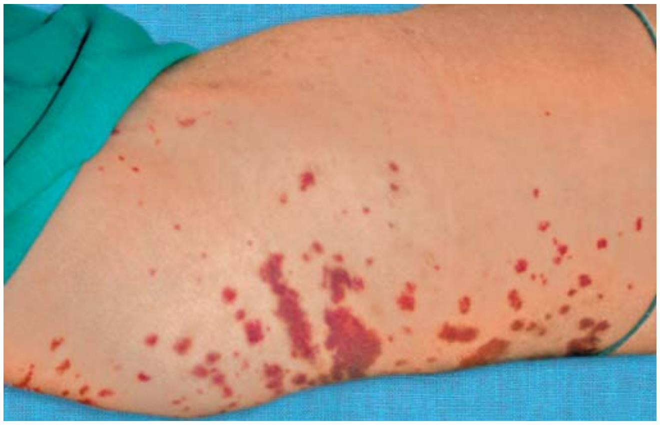

Fig. 14.5: *Grey-Turner's sign* in a patient with acute pancreatitis.

Effects of alcohol on pancreas

- Hypertriglyceridaemia, fatty acids and ethyl ester metabolites → pancreatic injury
- Formation of pancreatic juice which contains high enzymes but low enzyme inhibitors → enzyme activation within pancreas; precipitation of proteins, intraductal plug formation, ductal hypertension and obstruction → pancreatic injury
- Formation of O_2 free radicals inside the pancreas → pancreatic injury
- Direct pancreatic acinar injury
- Sphincter of Oddi spasm by alcohol leading into ductal obstruction
- Repeated subclinical acute pancreatitis causes fibrosis and chronic pancreatitis

Note:

- *Hereditary pancreatitis*: In normal individuals small amount of spontaneous trypsinogen activation into trypsin occurs in pancreas which is further prevented and protected by trypsin inhibitors. In hereditary pancreatitis genetic mutation causes defective trypsin inhibitors and trypsin becomes resistant to trypsin inhibitors leading into high concentration of intrapancreatic active trypsin which activates all other enzymes leading into pancreatitis. Here even though acute pancreatitis can occur; this is commonly subclinical leading into chronic pancreatitis.

- *Autoimmune pancreatitis*: It can be associated with primary sclerosing cholangitis, Sjögren's syndrome, biliary cirrhosis. *Lymphoplasmacytic autoimmune pancreatitis* is a condition which causes high levels of circulating IgG_4, pancreatitis, bile and pancreatic ductal strictures (*double duct sign* like mimicking the pancreatic cancer), pancreatic head mass.
- ERCP, major surgeries activate pancreatic enzymes causing acute pancreatitis.
- Hypercalcaemia, hyperlipidaemia type I, IV, V can cause pancreatitis. It is due to activation of digestive enzymes, blockage of pancreatic microcirculation causing pancreatic ischaemia.
- *Idiopathic pancreatitis (20%)*: Even though it is due to some unidentified cause, it is probably due to gallbladder *sludge* or microcrystals which often can be very well controlled or further prevented by cholecystectomy and sphincterotomy. It also can be due to malfunction of sphincter of Oddi which is treated by sphincterotomy and pancreatic septotomy. Some genetic cause is also postulated.
- *Colocalisation hypothesis*: Trypsinogen within the cytoplasmic vacuoles of acinar cell gets colocalised into the lysosomal hydrolases, commonly cathepsin B to activate into trpsin which in turn activates other intracytoplasmic digestive enzymes leading into intrapancreatic inflammation and pancreatitis.
- *Infection* occurs by *bacterial translocation across gut* due to altered mucosal barrier.

Atlanta classification

1. Acute oedematous pancreatitis (80%): Milder form, mortality is <1%
2. Acute necrotising pancreatitis (20%): Characterised by pancreatic and peripancreatic necrosis. Mortality is ≈ 15% - 30%

Present Classification of Acute Pancreatitis

- *Interstitial acute pancreatitis* with mortality less than 1%.
- *Necrotizing pancreatitis*.
  - Sterile necrosis with mortality 10%.
  - Infected necrosis with mortality 30-40%. It is confirmed by CT guided aspiration and Gram's stain. When confirmed, needs pancreatic necrosectomy.

Note:
In *Ulm classification* pancreatic abscess and pseudocysts are also included.

Organisms Involved

Infection is commonly polymicrobial (60%). It may be from gallbladder, colon or small bowel via transmural migration or by haematogenous spread. Infection rate in one week is 24% and in 3 weeks it is 70%.

- *E. coli* (35%).
- Klebsiella (25%).
- Enterococcus (25%).
- Others—staphylococci, *Pseudomonas*, *Proteus*, *Enterobacter*, *Anaerobes*, *Candida* (10%).

Clinical Features

- Sudden onset of upper abdominal pain which is referred to back. Pain is severe, agonizing and refractory. Pain may be relieved or reduced by leaning forward.
- Vomiting and high fever, tachypnoea with cyanosis.
- Tenderness, rebound tenderness, guarding, rigidity and abdominal distension, severe illness.
- Often mild jaundice (due to cholangitis). Jaundice may also be due to bile duct disease / obstruction or cholestasis.
- Features of shock and dehydration.
- Oliguria, hypoxia and acidosis.
- Grey-Turner's sign, Cullen's sign, Fox sign.
- Haematemesis/malaena due to duodenal necrosis, gastric erosions, decreased coagulability/DIC.
- Hiccough when present is refractory.
- Ascites may be present. Paralytic ileus is common.
- Pleural effusion (20%), pulmonary oedema, consolidation, features of rapid onset ARDS is often observed.
- Neurological derangements due to toxaemia, fat embolism, hypoxia, respiratory distress can occur. It may be mild psychosis to coma.
- Occasionally haematemesis or melaena can occur.

Fluid, Metabolic, Haematologic and Biochemical Changes

- *Hypovolemia* due to diffuse capillary leak and vomiting causing raised haematocrit, blood urea, serum creatinine levels.
- *Hypoalbuminaemia* which is more revealed after fluid correction.
- *Hypocalcaemia* is either due to decreased albumin level or specific loss of ionized calcium. Hypocalcaemia due to reduced ionised calcium carries poor prognosis. Response of calcium reserve in bone to PTH is also reduced.
- Total count is raised with *neutrophilia. Thrombocytopenia*, raised FDP, decreased fibrinogen, prolonged partial thromboplastin time and PT—are common. DIC can develop later.
- *Hypochloraemic metabolic alkalosis* is common due to repeated vomiting.
- Reduced insulin secretion, increased glucagon and catecholamine secretion causes *hyperglycaemia*, more so in diabetic patients.
- *Hyperbilirubinaemia* may be due to biliary stone/obstruction or cholangitis or non-obstructive cholectasis.
- *Hypertriglyceridaemia* is common especially in hyperlipidaemic patients.
- *Methemalbuminemia,* when it occurs in acute pancreatitis indicates poor prognosis.

Clinical features are classified as:

- Mild attack
- Severe attack
- Fulminant attack

Differential diagnosis

- Perforated duodenal ulcer
- Cholecystitis
- Mesenteric ischaemia
- Ruptured aortic aneurysm
- Ectopic pregnancy
- Salpingitis
- Intestinal obstruction
- Diabetic ketoacidosis

Investigations

- *Serum amylase* is very high (>1000 Somogyi units) or shows rising titre.
- *Amylase creatinine clearance* ratio is increased. It is - urine amylase/serum amylase X serum creatinine/ urinary creatinine X 100. Normal value is 1-4%. More than 6% signifies acute pancreatitis.
- *Serum lipase* more specific than amylase. Serum lipase level after rise persists for longer period than amylase. Pancreas is the only source unlike amylase, hence more specific.
- *Serum lactescence* (related to triglyceride metabolism)—Most specific in hereditary hyperlipidaemia or alcohol pancreatitis.
- Serum trypsin is the most accurate indicator but it is not commonly used.
- Trypsinogen activation polypeptide (TAP) assay in serum and urine reveals the severity of the acute pancreatitis. CRP (>150 mg/L) is also useful. Phospholipase A2, LDH levels are also often assessed.
- Liver function tests: Serum bilirubin, albumin, prothrombin time, alkaline phosphatase.
- Blood urea, serum creatinine.
- Blood glucose (hyperglycaemia is seen).
- Serum calcium level (hypocalcaemia occurs).
- Arterial PO_2 and PCO_2 level to assess pulmonary insufficiency (or ARDS).
- Urinary lipase estimation.
- Total count, haematocrit, platelet count, coagulation profile.
- Peritoneal tap fluid shows *high amylase* and protein level (very useful method). Lipase level in ascitic fluid is also useful.
- Plain X-ray shows
  - *'Sentinel loop'* of dilated proximal small bowel.
  - Distension of transverse colon with collapse of descending colon (*colon cut off sign).*
  - Air-fluid level in the duodenum.
  - Renal halo sign.
  - Obliteration of psoas shadow.
  - Localized ground glass appearance.
- U/S abdomen.
- *Spiral CT is* better—gold standard. It is done after 72 hours. To look for oedema, altered fat and fascial planes, fluid collections, necrosis (non-enhancement area > 30% or 3 cm), bowel distension, mesenteric oedema and haemorrhage. CT guided aspiration (fluid should be sent for Gram's stain and bacterial culture); pigtail catheter insertion can also be done. If fluid culture shows bacterial growth with CT showing necrosis, it means it is infected necrosis and needs early pancreatic necrosectomy. CT guided aspiration and Gram's staining may need to be repeated.
- MRI, MRCP—should be done at a later date.
- ERCP is usually not done in acute phase.
- Chest X-ray for effusion and ARDS.
- EUS—to see necrosis, calcifications and to assess CBD.

Balthazar CT Scoring System (Emil J Balthazar, 1990)

CT grade:

- A—Normal pancreas—0
- B—Oedematous pancreatitis—1
- C—B+ mild extrapancreatic changes—2
- D—Severe extrapancreatic changes with one fluid collection—3
- E—Extensive/multiple extrapancreatic collections or gas bubbles in or adjacent to pancreas—4

Necrosis score:

- None—0
- < 1/3rd—2
- More than 1/3rd less than half—4
- More than half—6

CT severity index— CT grade + necrosis score
0—3, 4—6, 7—10.

Treatment

Treatment of acute pancreatitis

- Conservative, 70-90%
- Surgical treatment when indicated, 10-30%
- Management of complications like acute pseudocyst, abscess, fistula, haemorrhage; systemic complications like ARDS, renal failure, MODS

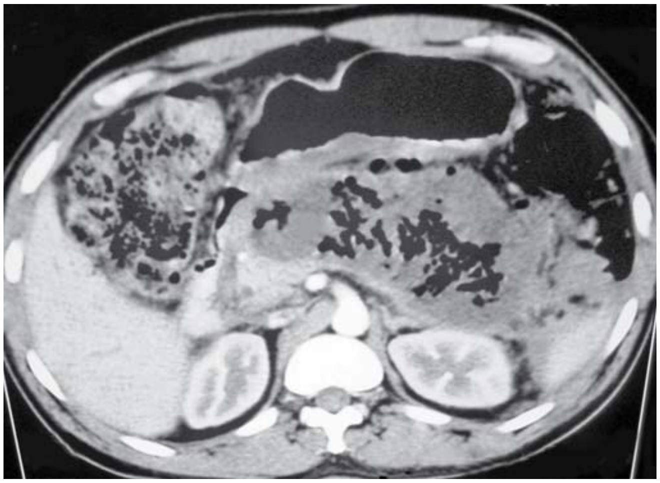

Fig. 14.6: CT picture showing features of acute pancreatitis.

Conservative Treatment

- Commonly opted.
- Rehydration is essential (250-400 ml / hour) as there is lot of fluid sequestration and 3rd space fluid loss. It is done by using ringer lactate, normal saline, dextrose saline, plasma and fresh blood transfusion/packed cells.
- Pain relief by pethidine and other analgesics. Morphine is not used as it causes spasm of sphincter of Oddi.
- In severe haemorrhagic episodes, fresh frozen plasma and platelet concentrate may be required in anticipation of DIC and haemorrhage.
- Nasogastric aspiration, urinary catheterisation to maintain and monitor urine output 50 ml hourly.
- Antibiotics like third generation cephalosporins, imipenem, meropenem, cefuroxime are used even though its role is not clear but it is commonly used to reduce the anticipated sepsis.

  Indications for antibiotics

  - In severe infected necrosis with proved culture.
  - Prophylactic antibiotic therapy, in severe pancreatitis.
  - In biliary pancreatitis with biliary stones and cholangitis.
  - Pancreatic abscess formation.
  - Clinically disease is rapidly progressing with deterioration.
  - No role of antibiotics in early pancreatitis.
  - Imipenem, cilastatin, cefuroxime are used.
- Calcium gluconate 10 ml 10% IV 8th hourly is given as patient will be hypocalcaemic.
- CVP line is essential to monitor, for rapid fluid therapy and for Total Parenteral Nutrition (TPN) using carbohydrate, amino acids, vitamins, essential elements.
- IV Ranitidine 50 mg 6th hourly or IV omeprazole 40 mg BD or IV pantoprazole 80 mg BD to prevent stress ulcers and erosive bleeding.
- Proper electrolyte management with monitoring is needed.

Mainly conservative treatment

1) (P) : **P**ain relief [pethidine, meperidine (no morphine)]
Protease inhibitors (aprotinin, antisnake venom, EACA etc.), **P**lasma
2) (A) : **A**ntibiotics (ceftazidime, cefoperazone, cefotaxime) imipenem
Anticholinergics (to reduce sphincter pressure)
3) (N) : **N**asogastric aspiration, **N**asal O_2
Nutritional support (TPN)
4) (C) : **C**alcium gluconate 10 ml 10% 8th hourly
Calcitonin
CVP line
5) (R) : **R**ehydration by IV fluids, plasma, blood
Ranitidine IV 50 mg 8th hourly
Respiratory support (ventilator)
Resuscitation when required
6) (E) : **E**ndotracheal intubation
Electrolyte management
7) (A) : **A**ntacids
8) (S) : **S**wan-Ganz catheter for CVP and TPN
Somatostatin and its analogue (*Octreotide*)

- It is always better to manage the patient in an ***intensive care*** set up so that when needed endotracheal intubation, ventilatory support, tracheostomy can be done as an emergency basis.
- In case of renal failure haemodialysis is required.
- Somatostatin/octreotide is often used to reduce pancreatic secretion.
- Anticholinergics, protease inhibitors and calcitonin are other agents used. Their efficacy is not sure.
- Steroid injection in initial phase of shock is beneficial. It is also useful in respiratory distress and ARDS.
- Nebulisation, bronchodilators are also tried.
- *Nasojejunal tube placement and feeding* should be started as early as possible once ileus subsides so that it reduces the infection rate by transmucosal migration of bacteria and it also improves nutritional status. During recovery period nasogastric feeding or jejunostomy feeding can be undertaken.
- Repeated calcium estimation, U/S examination to see the response is useful.
- Dopamine or low molecular weight dextran (to improve renal perfusion), somatostatin/octreotide (to reduce pancreatic secretion) are also often used.
- *Management of complications* like acute lung injury, atelectasis, renal failure, GI bleeding, metabolic encephalopathy, electrolyte deficiency as and when needed by repeated observation and evaluation.

Other measures:

Continuous monitoring, observation, maintaining good urine output 30 ml/hr. The required investigations are repeated to find out the response to treatment like serum calcium, total count, blood urea, creatinine, PaO_2, haematocrit, ECG.

Uses of octreotide

- Pancreatic surgeries
- Variceal bleeding
- Endocrine pancreatic tumours
- GI fistulas, pancreatic fistulas
- Dumping syndrome
- Carcinoid tumour, acromegaly
- Acute pancreatitis

Dose:

50 μg as loading dose IV

50 μg 1 hour in 5% dextrose as maintenance dose.

Surgery

Indications for Surgical Intervention (10% cases)

1. If condition of patient deteriorates in spite of good conservative treatment.
2. If there is formation of pancreatic abscess, or infected necrosis.

> *Each case has its lesson—a lesson that may be, but is not always, learnt for clinical wisdom is not the equivalent of experience.*
>
> — **William Osler**

3. When diagnosis is in doubt.
4. In severe necrotising pancreatitis as a trial to save the life of the patient which has got very high mortality.

Surgical management of acute pancreatitis

Surgery removes intra- and extrapancreatic necrotic materials, pancreatic fluid, and toxins. It permits preservation of viable pancreatic tissue. *Open surgery is the gold standard for infected pancreatic necrosis.*

- After opening the abdomen, all necrotic tissue, pus, infected fluid and toxins are removed. 10-12 litres of normal saline wash is given. Drainage tubes are placed liberally. Abdomen is closed in layers—*Conventional closed method* (necrosectomy, wide debridement, adequate drainage, cholecystectomy, closure). Re-laparotomy is done only on demand later.
- Laparotomy – necrosectomy – wide debridement – wash – wide packing. Wound is left open. Repeated wash and packings are done until healthy granulation develops – *open method.*
- Laparotomy – necrosectomy and closure with drain and re-laparotomy later—*semi-open* method**.**
- *Zip technique* can be used to give repeated wash to remove toxins and necrotic tissues until healthy granulation tissue develops in the pancreatic bed—*Bradley's repeated laparotomies and wash.*
- Continuous closed peritoneal lavage of the pancreatic bed and lesser sac is done with 10-12 litres of normal saline or hyperosmolar, potassium free dialysate fluid 2 litres/hour using multiple tubes to remove toxic material in the peritoneal cavity/retroperitoneal area until return fluid becomes clear—*Beger's lavage*. Procedure is done after initial surgical debridement.
- Extra peritoneal lavage through bilateral flank incisions is also used. But being a blind procedure it is technically difficult and one may not be sure about the adequacy of the procedure.
- Under laparoscopic visualisation, necrosectomy, wash and drainage can be done. But often it is difficult to create pneumoperitoneum in acutely ill patient. But it is increasingly becoming popular.
- A jejunostomy is often done along with these procedures to have early enteral nutrition.
- *Endoscopic necrosectomy* is often done in some centres. *Early endoscopic intervention* (within 48 hours) with ERCP, biliary stone removal and stenting in biliary pancreatitis is done and favored in many centres.
- *Further management* is important to prevent recurrence. Gallstones should be dealt by laparoscopic cholecystectomy in 2 weeks after acute attack during same admission period. Endoscopic sphincterotomy (ERCP) and often stenting may be needed if there are CBD stones.

Note:
GI fistula either enteric or pancreatic and incisional hernia are the late complications of these surgeries.

COMPLICATIONS OF ACUTE PANCREATITIS

- Shock—Hypovolemic and septic
- Respiratory failure and ARDS—Common in 7 days
- Septicaemia—Common after 7 days
- Hypocalcaemia
- Disseminated intravascular coagulation (DIC)
- Acute renal failure
- Pancreatic pleural effusion (left sided 20%)
- Pancreatic pseudo aneurysm
- Pancreatic ascites
- Colonic stricture
- Pseudocyst of pancreas
- Chronic pancreatitis
- Splenic vein thrombosis
- Pancreatic abscess

- *Acute fluid collection*
  - It is collection of fluid in or near the pancreas during an attack of acute pancreatitis with an ill-defined or lacking fibrin wall or granulation tissue.
  - It is 40% common; it usually occurs at peripancreatic area, occasionally intrapancreatic.
  - More than 50% of acute fluid collection regress spontaneously, remaining may form pseudocyst of pancreatic necrosis.
  - CT guided aspiration at one puncture site is often necessary to confirm that it has not formed abscess.
- *Acute pseudocyst*
  - It is collection of fluid with pancreatic juice in or near the pancreas localised by thin fibrin wall or granulation tissue.
  - Occurs in 2 weeks, usually resolves spontaneously.
  - Fluid can be removed percutaneously under guidance or through an endoscope.
- *Pancreatic pseudocyst*
  - It is collection of fluid in a false cavity which commonly contains brownish pancreatic enzyme rich fluid, *lined by granulation tissue* but *not true epithelium* with an organised thick fibrous covering.
  - It is commonly located in peripancreatic region, in lesser sac. It can be often intrapancreatic or in other places in the peritoneal cavity. It can be multiple also.
  - It usually forms 4 weeks after an attack of pancreatitis—*a chronic entity.*
  - It is usually sterile but can get infected.
  - Pseudocyst may rupture to form pancreatic ascites.
  - It may cause pancreaticopleural fistula when once it erodes into left side pleural cavity or it may erode into splenic vessels causing life threatening haemorrhage.
- *Pancreatic necrosis*
  - It is focal/diffuse area of non-viable pancreatic parenchyma with peripancreatic fat necrosis.
  - It is initially sterile but eventually gets infected.
  - It contains *paste/putty* like material.

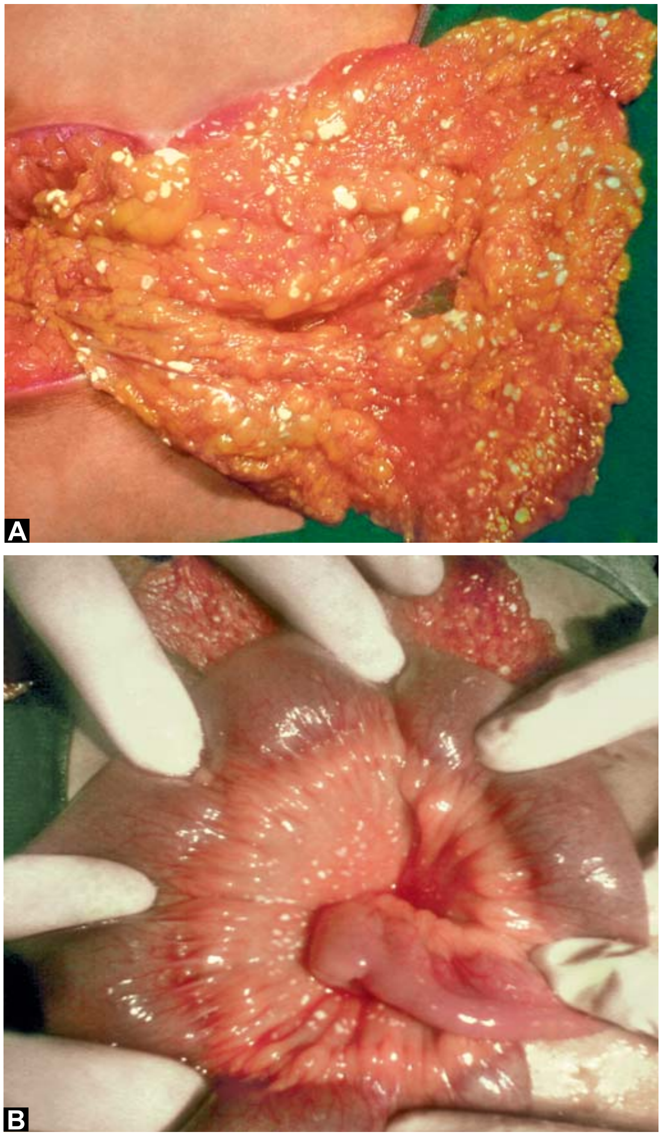

Figs 14.7A and B: On table photos of acute pancreatitis. Note the typical saponification in the omentum and mesentery.

- It may form pseudocyst, abscess or may be replaced by fibrous tissue during healing.
- It can be sterile or infected necrosis.
- It is confirmed by CT scan and CT guided aspiration usually at one site; collected fluid should be sent for culture.
- *Infection of necrotic area* usually occurs within 4 weeks. It is ideally *conventionally* treated with laparotomy, debridement is done by removing all necrotic tissue with adequate drainage for remaining viable exocrine function. Often repeat laparotomies are done once in 3 days until all necrotic materials are cleared and jejunostomy is done in final approach.
- Laparotomy and debridement done once with continuous lavage; antibiotics with percutaneous drainage/only surgical drainage without debridement/debridement using minimally invasive methods—are *unconventional* approaches.

- *Pancreatic abscess*
  - It is collection of pus in lesser sac (intraabdominal) in relation to pancreatic surface which contains mainly pus with only less or no necrotic pancreas.
  - It may slough off the pancreatic/splenic vessel wall to cause torrential haemorrhage. Abscess may be single or multiple (60%). It is commonly in head/body or tail. But often entire gland may be involved (25%).
  - Abscess may rupture into viscera or extend into other part of the abdomen. Features of sepsis, tender palpable mass in the epigastrium with leukocytosis are observed.
  - It is treated by antibiotics, percutaneous U/S or CT guided aspiration or drainage or open drainage.

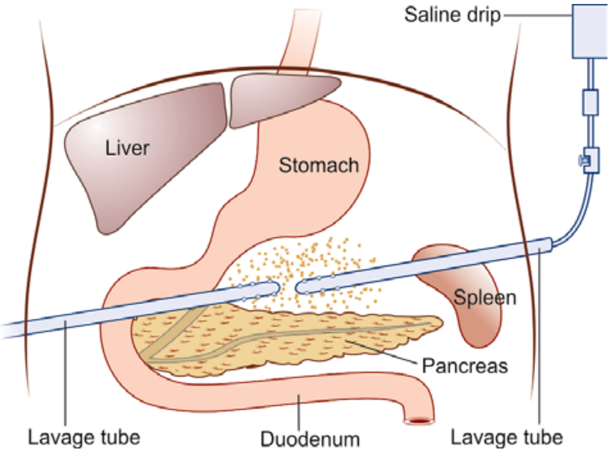

Fig. 14.8: *Beger's lavage* to pancreatic bed. Tubes are placed postoperatively. Continuous saline irrigation through the tube is done until irrigation fluid becomes clear.

- *Respiratory complications*
  - Thay are often severe and life-threatening. It is due to distension of abdomen, diaphragmatic elevation, pleural effusion, reduced surfactant (lecithin) activity in alveoli due to lecithinase, severe pain, pleural effusion (left), intravascular coagulation in lungs and ARDS. *Arterial blood gas analysis* should be done. Often it needs ventilator support.
- *Pancreatic pseudoaneurysm*
  - It is due to enzymatic digestion (elastase) of the wall and weakening and aneurysmal dilatation of the splenic (50%) or gastroduodenal vessels (15%) or inferior and superior pancreaticoduodenal arteries (10%).
  - It may rupture and cause life threatening haemorrhage or may rupture into stomach or duodenum to cause massive upper GI bleed or may rupture into pancreatic duct causing *haemosuccus pancreatitis*.
  - It is diagnosed by CT angiogram.
  - It is treated by critical care, blood transfusion, emergency angiographic embolisation or by open surgery and ligation of the involved vessel. It carries high mortality.
- *Pancreatic fistula*
  - It can occur due to ductal wall disruption and necrosis or after surgical intervention for acute pancreatitis (necrosectomy).

> *In case of chronic pancreatitis jaundice supervenes in only fifteen percent of cases, but diabetes is present on thirty percent.*
> **— Richard B Cattell**

    - It may be internal into bowel or external to skin.
    - Fistula may be low (< 200 ml) or high (> 200 ml) out put.
    - It can be straight or curved.
    - It is confirmed by biochemical analysis, ERCP, CT fistulogram.
    - If fistula persists for 6 months then sphincterotomy, resection of fistula with pancreatic resection, pancreaticojejunostomy is done.
- *Emphysematous pancreatitis*
    - It is gas in pancreatic parenchyma, a dangerous type and can be diagnosed by CT scan.

Note:

- Decrease serum calcium level is *worst prognostic indicator* of pancreatitis.
- These scoring systems differ for non-gallstone pancreatitis and gallstone pancreatitis.
- Other scoring systems which are often used are APACHE II (Acute Physiology And Chronic Health Evaluation II) and SAP (simplified acute physiology) scoring systems. Twelve variables are used to assess in APACHE II scoring.
- *C- Reactive Protein (CRP) > 210 mg* / L in first 4 days and > 120 mg/L in one week; phospholipase A2 assay; pancreas related protein; interleukin-6; polymorphonuclear elastase > 120 µg / L; serum ribonulcease assay—are other parameters used to assess the severity.

Criteria to Find Out Systemic Failure in Acute Pancreatitis

- *Cardiac*: Hypotension; pulse > 130/minute; arrhythmias, ECG changes.
- *Pulmonary*: $PaO_2 > 60$ mmHg; ARDS.
- *Renal*: Urine output < 40 ml/hour; increase in blood urea and serum creatinine.
- *Metabolic*: Falling serum calcium; magnesium and albumin.
- *Haematologic*: Fall in haematocrit; DIC.
- *Gastrointestinal*: Severe ileus; sequestration of fluid.
- *Neurologic*: Irritability; confusion; localising features.

PSEUDOCYST OF PANCREAS

- It is localized collection of sequestered pancreatic fluid, usually 3 weeks after an attack of acute pancreatitis.
- It can occur after trauma and recurrent chronic pancreatitis.
- Collection usually occurs in the lesser sac in relation to stomach, but can occur in relation with duodenum, jejunum, colon, splenic hilum.
- About 50% of acute pancreatitis leads to pseudocyst formation, but among that 20-40% will resolve spontaneously.

Sites of pseudocyst

- Lesser sac—commonest, i.e. between colon and stomach

In relation to:

- Duodenum
- Jejunum
- Colon
- Splenic hilum

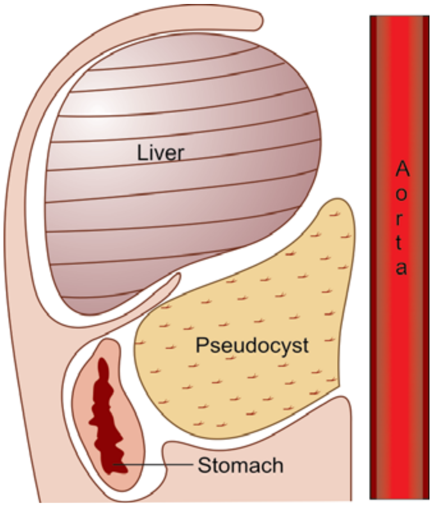

Fig. 14.9: Pseudocyst of pancreas.

Ranson's prognostic criteria in gallstone pancreatitis	***Ranson's prognostic criteria in Non-gallstone pancreatitis***	***Glasgow – Imrie prognostic criteria***	***Acute Physiology and Chronic Health Evaluation (APACHE II) score > 8 points predicts 11% to 18% mortality***
On admission: • Age > 70 years • TC > 18,000/cu mm • Blood sugar > 220 mg% • LDH > 400 IU/L • AST > 250 IU/100 ml	On admission: • Age > 55 years • TC > 16,000/cu mm • Blood sugar > 200 mg% • LDH > 700 IU/L • AST > 250 IU/100 ml	On admission: • Age > 55 years • TC > 15,000/cu mm • $PaO_2 < 60$ mmHg • Blood urea > 16 mmol/L (No response to IV fluids) • Blood sugar > 200 mg% (no H/O diabetes)	• Equation includes the following factors: age, rectal temperature, mean arterial pressure, heart rate, PaO_2, arterial pH, serum potassium, serum sodium, serum creatinine, hematocrit, white blood cell count, Glasgow coma scale score.
Within 48 hours: • Haematocrit drop > 10% • BUN rise > 2 mg% • Serum calcium < 8 mg% • Base deficit > 5 mEq/L • Fluid sequestration > 4 L	Within 48 hours: • Haematocrit drop > 10% • BUN rise > 5 mg% • $PaO_2 < 60$ mmHg • Serum calcium < 8 mg% • Base deficit > 4 mEq/L • Fluid sequestration > 6 L	Within 48 hours: • Serum calcium < 2 mmol/L • Serum albumin < 3.2 gm/dl • LDH > 600 U/L • AST / ALT > 600 U/L	APACHE II modified (1996) LFT is added in biliary pancreatitis APACHE – O (Toh 1996) Obesity is added

Types

Depending on whether it communicates with pancreatic duct or not it is classified as:

1. *Communicating pseudocyst.*
2. *Noncommunicating pseudocyst.*

It can be

- *Acute pseudocyst*.
- *Chronic pseudocyst*.

D'Egidio Classification of Pseudocyst

- *Type I*: After an attack of acute pancreatitis. Normal duct anatomy; no fistula/communication.
- *Type II*: After an attack of acute on chronic pancreatitis. Abnormal duct anatomy without stricture; 50% chances of fistula.
- *Type III*: After an attack of chronic pancreatitis, abnormal anatomy with stricture; always communicating. It appears like retention cyst.
- Initially cyst wall is thin *(unformed),* but later it gets fibrosed and thickened *(formed)* (matured).
- It is lined by fibrin layer but no endothelium, hence called as pseudocyst.
- It contains typically brownish fluid with sludge like necrotic material. It can get infected to form *infected pseudocyst* or *pancreatic abscess.*
- Amylase level in the cyst fluid is very high (> 5000 units/ml).

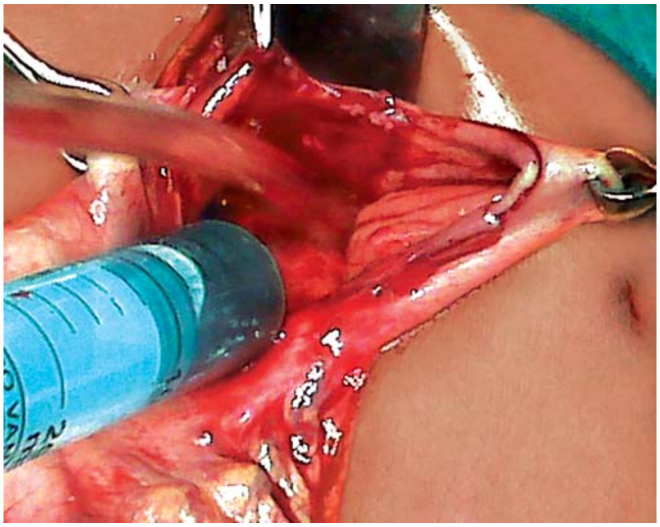

Fig. 14.10: Pancreatic pseudocyst on table aspiration. Note the brownish black colour of the fluid.

Note:
It is not a true cyst as there is no epithelial lining.

Clinical Features

- A swelling in the epigastric region which is hemispherical, smooth, soft, not moving with respiration, not mobile, upper margin is diffuse but lower margin well defined, resonant or impaired resonant on percussion, with transmitted pulsation confirmed by knee-elbow position.
- If it is infected, it will be tender mass and patient will be toxic with fever and chills.
- Because stomach is stretched towards the abdominal wall, Ryle's tube passed will be felt per abdominally *(Baid test)***.**

Differential Diagnosis

- Aortic aneurysm.
- Retroperitoneal cyst or tumour.
- Cystadenocarcinoma of pancreas**.**

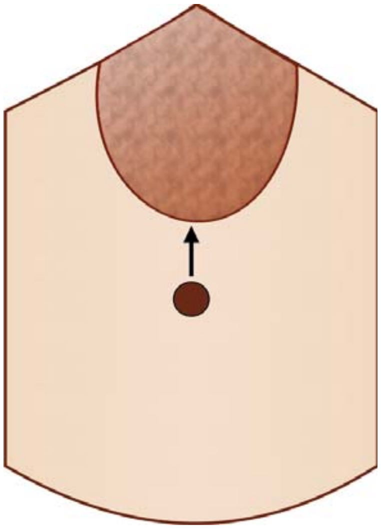

Fig. 14.11: Pseudocyst mass.

- Cyst of the liver.
- Mesenteric cyst.
- Hydatid cyst.

Investigations

- U/S abdomen (commonly done procedure), U/S reveals the size and thickness of the pseudocyst. Size less than 6 cm indicates that one can wait for spontaneous resolution. Endosonography (EUS) is very useful.
- CT scan is ideal and study of choice. It is two times more sensitive than U/S. It demonstrates size, shape, number, wall thickness, contents, pancreatic duct size, and extent of necrosis in pancreas, calcification and atrophy in chronic pancreatitis, regional vessels, pseudoaneurysm, splenic/portal vein thrombosis.
- MRCP delineates the ductal anatomy and its abnormality.
- ERCP can be done to find out the communication.
- Barium meal (lateral view) shows widened vertebrogastric angle with displaced stomach (Not usually done now).
- LFT, serum amylase.

Indications for surgery/intervention

- Size more than 6 cm
- Formed pseudocyst
- Infected pseudocyst
- Cyst persisting after 6 weeks/progressive cyst
- Multiple cysts/cyst due to trauma
- Communicating cysts/cyst with severe pain
- Thick-walled pseudocyst

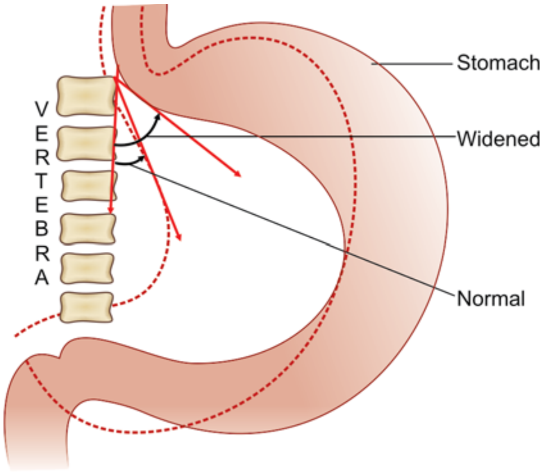

Fig. 14.12: Vertebrogastric angle.

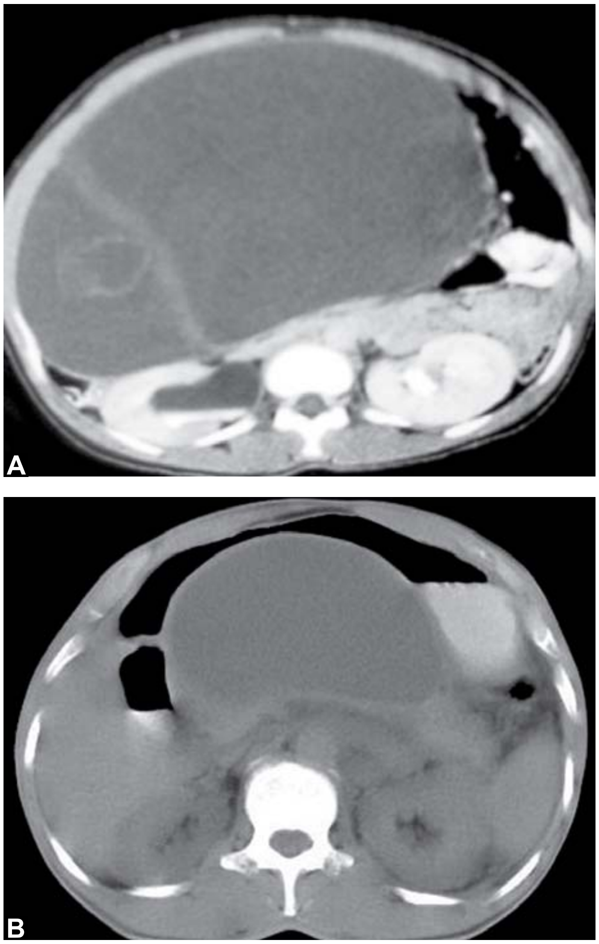

Figs 14.13A and B: CT scan showing typical pseudocyst of pancreas.

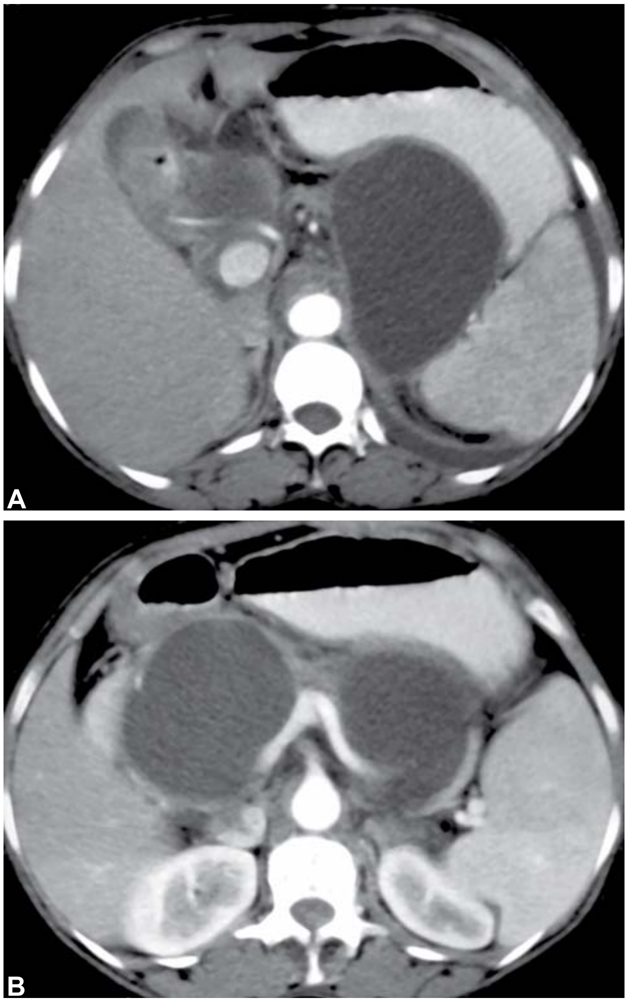

Figs 14.14A and B: CT scan showing pseudocyst of pancreas both single and multiple. Note the cyst cavity and the thick cyst wall. Patient needs drainage procedure either cystogastrostomy or cystojejunostomy.

Surgery

- *Cystogastrostomy:* At laparotomy, anterior wall and posterior wall of the stomach is opened. *Brownish fluid* is aspirated. The thick capsule of pseudocyst is opened. All fluid with necrotic material are sucked. Fluid should be sent for cytology, culture and sensitivity and amylase estimation. Cyst wall always should be biopsied. Cyst cavity should be washed with normal saline after breaking septae. Pseudocapsule is anastomosed to posterior wall of the stomach—*Jurasz operation.*

Note:
Roux-en-Y cystojejunostomy is better with lesser recurrence rate than cystogastrostomy.

Other procedures

- Cystoduodenostomy.
- Cystojejunostomy is done in large cyst, recurrent cyst.
- If infected, cystogastrostomy with external drainage is done using Malecot's catheter (*Smith operation)*.
- Endoscopic cystogastric stenting.
- ERCP and *transpapillary stenting* and drainage in communicating pseudocyst. There is loss of exocrine and endocrine functions with progressive pancreatic fibrosis.

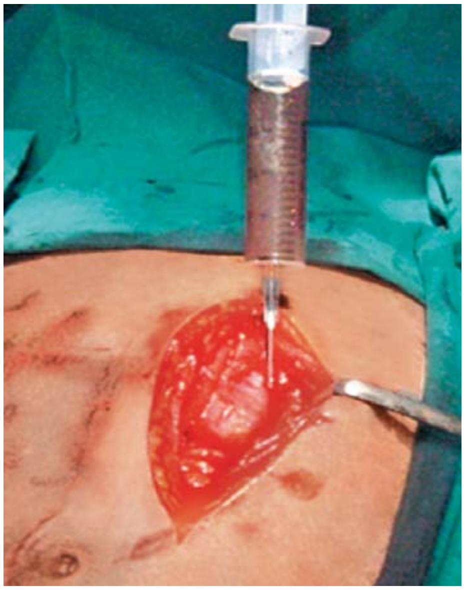

Fig. 14.15: Pseudocyst of the pancreas aspiration on table. Note the brownish-black colour of the fluid aspirated.

- Laparoscopic cystogastrostomy is becoming popular, effective and less invasive.
- U/S guided aspiration is useful in initial phases.
- Along with cystojejunostomy, pancreaticojejunostomy should be done if there is ductal stricture and dilatation and communication with pseudocyst.
- Distal pancreatectomy with pseudocyst removal if cyst is in distal part.

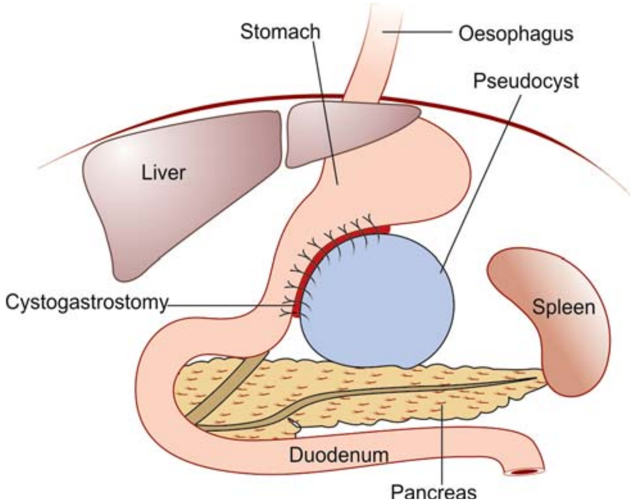

Fig. 14.16: Cystogastrostomy for pseudocyst of the pancreas.

Complications of pseudocyst

- Rupture—3%—into bowel or peritoneum
- Infection, commonest—20%
- Bleeding from the splenic vessels—7%
- Cholangitis
- Duodenal obstruction
- Portal/splenic vein thrombosis and segmental portal hypertension
- Cholestasis due to CBD block

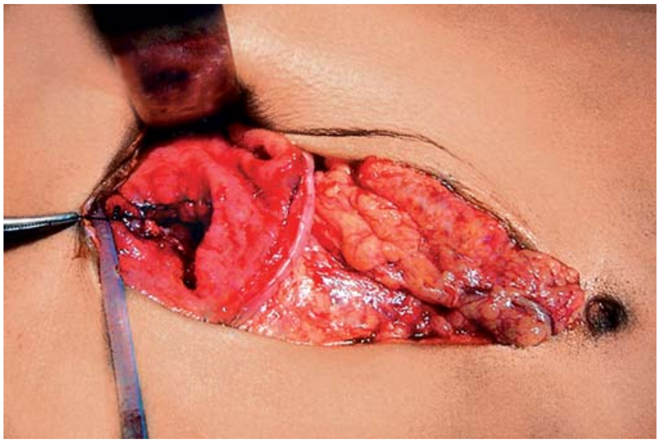

Fig. 14.17: Cystogastrostomy procedure, on table.

Pseudopseudopancreatic Cyst

It is a mass in epigastric region due to pancreatitis formed by bowel, omentum which clinically mimics a pseudocyst of pancreas. But fluid collection is absent.

Note:

- Small cyst can be drained through endoscopy. But debridement is not possible. Bleeding and cyst leakage are other problems.
- CT guided percutaneous catheter placement is done in critically ill patients, infected pseudocyst, unfit patients, cyst in pelvis/mediastinum. Improper drainage, recurrence and fistula formation are the problems.
- External drainage is done when cyst is infected, haemorrhagic or ruptured. Problem with external drainage is formation of fistula (20%).
- Therapy is decided based on thickness of the wall of pseudocyst; location of pseudocyst; contents of pseudocyst and ductal status.

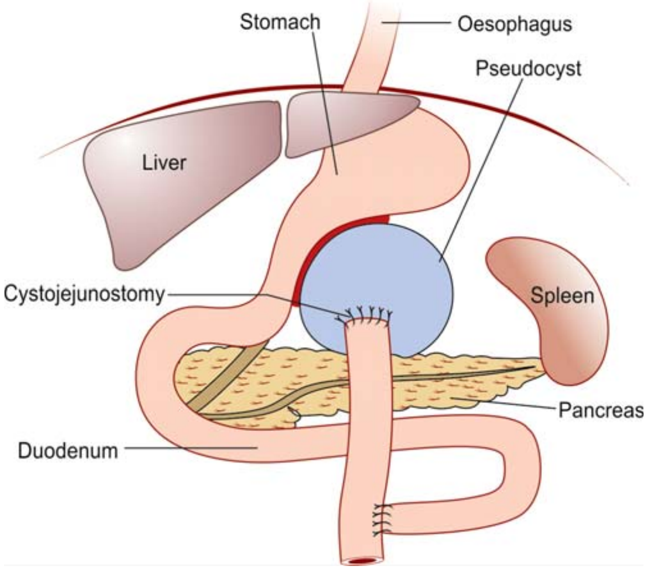

Fig. 14.18: Roux-en-Y cystojejunostomy for pseudocyst of the pancreas.

> *Carcinoma pancreas should be suspected when an elderly person develops diabetes and in spite of adequate treatment, continues to loose weight.*
> **— Robert D Lawrence**

CHRONIC PANCREATITIS

- It is persistent progressive irreversible damage of the pancreas due to chronic inflammation.

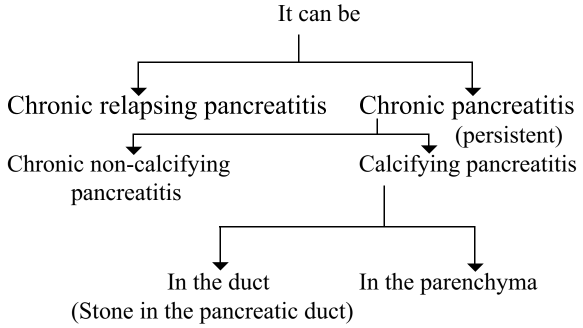
It can be
Chronic relapsing pancreatitis
Chronic pancreatitis (persistent)
Chronic non-calcifying pancreatitis
Calcifying pancreatitis
In the duct (Stone in the pancreatic duct)
In the parenchyma

- Chronic pancreatitis is more common in males, common in Kerala (induced by diet, rich in Tapioca).

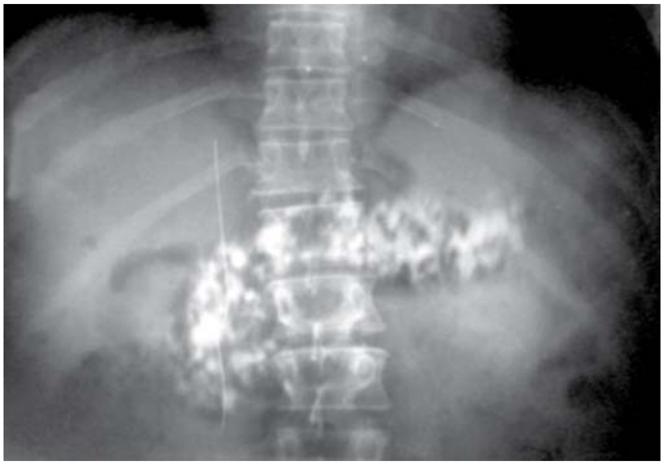

Fig. 14.19: Calcifications in the parenchyma of the pancreas.

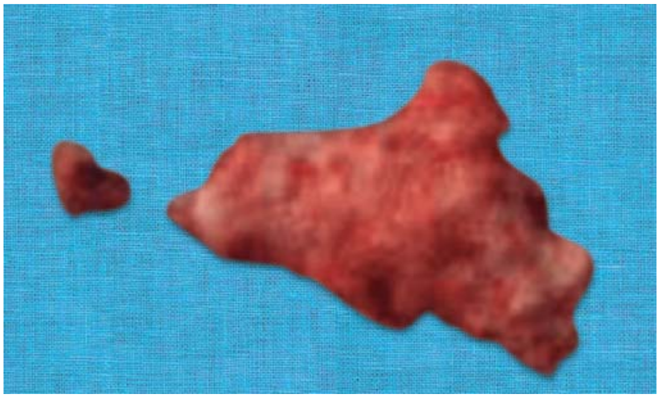

Fig. 14.20: Pancreatic ductal stone removed—specimen. Patient underwent pancreaticojejunostomy.

Aetiology

- Alcohol—80% main cause
- Stones in biliary tree—rare cause
- Malnutrition, diet
- Hyperparathyroidism
- Hereditary (familial hereditary pancreatitis)
- Idiopathic—20%—as mutation
- Trauma
- Congenital anomaly (Pancreatic divisum)
- Cystic fibrosis
- Autoimmune pancreatitis
- Hyperlipidaemia

TIGAR-O Risk Factor Classification 2001

T – Toxic - alcohol/tobacco/dietary/drug.
Metabolic-hypercalcemia/lipidemia/lipoprotein lipase deficiency.
I – Idiopathic—early/late onset/tropical.
G – Genetic mutations—CFTR/SPINK 1.
A – Autoimmune primary/with Sjogren/Crohn's disease.
R – Recurrent and severe acute/ischemic.
O – Obstructive—pancreas divisum//annular pancreas/stenotic papilla/duodenal obstruction/trauma/pancreatic ductal stones/choledochocele.

- Alcohol reduces pancreatic blood flow, alters cell viability, releases the free radicals, creates pancreatic ischaemia, and activates the pancreatic stellate cells which produce abundant extracellular matrix and collagen.
- Genetic predisposition may be the cause of idiopathic pancreatitis. Mutation in pancreatic secretory trypsin inhibitor causes activation of trypsin causing pancreatitis.
- Hereditary pancreatitis is an autosomal disorder with mutation in trypsinogen gene in chromosome 7. It causes recurrent painful episodes of acute pancreatitis in childhood, leading to chronic pancreatitis and pancreatic cancer in adulthood.

Theories and Concepts in Pathogenesis of Chronic Pancreatitis

- *Oxidative stress hypothesis*—Reactive by-products of hepatic mixed function oxidase activity damage the pancreas through chronic reflux of bile into the pancreatic duct.
- *The toxic-metabolic theory*—Alcohol is directly toxic to the acinar cell where it brings changes in intracellular metabolism causing pancreatic lipid accumulation, fatty degeneration, cellular necrosis, and eventual widespread fibrosis.
- *Stone and duct obstruction theory*—Alcohol increases the lithogenicity of pancreatic juice, leading to stone formation. Chronic contact of stones with duct epithelial cells produces ulceration, scarring, atrophy, fibrosis and chronic obstruction of the acini.
- *The necrosis-fibrosis theory*—Acute and chronic pancreatitis represents a spectrum of disease. Inflammation from acute pancreatitis leads to scarring, extrinsic compression of the pancreatic ductules, obstruction, stasis, atrophy and stone formation.
- *The genetic defect* of hereditary pancreatitis produces recurrent acute pancreatitis in early childhood, leading to chronic pancreatitis in early adulthood.
- *Cellular mechanisms* of pancreatic fibrogenesis is due to pancreatic *stellate cells* which are stimulated and activated by alcohol, oxidative stress, cytokines of acute pancreatitis; activated stellate cells migrate to the periacinar areas to deposit collagen and fibronectin. Transforming growth factor beta 1 is an important mediator of pancreatic fibrosis.
- *The sentinel acute pancreatitis event (SAPE) hypothesis*—An episode of acute pancreatitis, the sentinel event, sets the stage for the attraction of collagen-secreting stellate cells.
- *Heavy, prolonged alcohol use* is the most common cause of chronic pancreatitis. Alcohol-related chronic pancreatitis

is associated with more severe pain, more extensive calcification and ductal changes, and more rapid progression to endocrine and exocrine insufficiency. Often recurrent episodes of acute pancreatitis for several years are seen in these patients. Some cofactors amplify the effect of alcohol in these patients. Prevalence of some genetic mutations linked with pancreatitis like cystic fibrosis transmembrane regulator (*CFTR)*, serine protease inhibitor Kazal type-1 (*SPINK1*) has been noted in alcoholic pancreatitis. A high-fat diet and smoking affect pancreatic bicarbonate and water secretion adversely, inducing oxidative stress and increases the rate of pancreatic calcification.

- *Tropical pancreatitis* is endemic in some regions of India (Kerala), Africa, and South America. Episodic abdominal pain begins in childhood and is followed by rapid progression to endocrine and exocrine insufficiency with pancreatic calculi in non-alcoholic individuals. Dietary toxins (cyanogens in the cassava plant, tapioca) and micronutrients like zinc, copper, and selenium, vitamin A deficiencies, genetic factors, ductal abnormalities are the probable causes of tropical pancreatitis. *Linnamarin* and its methyl derivative, in acid pH of stomach releases hydrocyanic acid which is cytotoxic in presence of rhodanase releases thiocyanates causing depletion of methionine, damaging pancreas → pancreatitis. Gross look is small, firm fibrotic/adipose type.
- *Early-onset* idiopathic chronic pancreatitis manifests with severe abdominal pain in childhood, with relatively few structural and functional changes. *Late-onset* idiopathic chronic pancreatitis manifests in middle and late adulthood, often with minimal pain and pronounced exocrine insufficiency.

Pathology

- It shows atrophy of acini, hyperplasia of duct epithelium, interlobular fibrosis, calcifications, ductal dilatation, with strictures in the duct, focal necrosis.
- There is loss of exocrine function initially and endocrine functions eventually.
- Ductular metaplasia and acinar atrophy along with fibrosis and cyst formation develops.

Pathology of chronic pancreatitis

- Focal necrosis
- Segmental or diffuse fibrosis
- Parenchymal calcification or ductal stones
- Stricture or dilatation of the duct

Spectrum of Chronic Pancreatitis

- *Early* – pancreatic oedema – chronic inflammation – normal secretory function.
- *Moderate* – early fibrosis; only few acinar cells – exocrine dysfunction.
- *Late* – fibrosis – loss of secretory function – diabetes mellitus.
- *Complications* develop secondary to healing and fibrosis; deposition of inspissated proteinaceous material in the duct; over expression of CTGF and TGF-B1 that stimulate extracellular matrix.
- *Pain hypothesis include:*
  - Acute and chronic inflammation of the pancreas.
  - Increased pressure within the pancreatic ductal system and parenchyma, ductal dilatation, stasis.
  - Ischaemia of the parenchyma secondary to increased interstitial pressure.
  - Over expression of a particular protein in pancreatic nerve fibres.
  - Pancreatic nerve growth factor released by degenerating acinar cells and its high affinity receptor that promotes nerve growth and repair.
  - Pain may be due to perineural sheath destruction by toxins, ischaemia and also due to tissue acidosis.

Classification of Chronic Pancreatitis

I: Based on main duct dilated or not:

- Large duct disease—main pancreatic duct is dilated.
- Small duct disease—main pancreatic duct, is normal or smaller in size.

II: Staging/classification of chronic pancreatitis (Stages A, B, C)

A new classification of chronic pancreatitis, based on combination of clinical signs, morphology and function, is presented (2009 Büchler et al).

Specific definition of chronic pancreatitis stage A

- **Stage A chronic pancreatitis**

It is the early stage of chronic pancreatitis where complications have not yet appeared and the clinical exocrine and endocrine function is preserved. Subclinical signs (impaired glucose tolerance, reduced exocrine function but without steatorrhoea) might already be apparent.

Stage A is accepted under the following conditions: Pain of any type and degree and/or attacks of acute pancreatitis, no complications, no steatorrhoea, and no insulin-dependent diabetes mellitus.

Specific definition of chronic pancreatitis stage B

- **Stage B chronic pancreatitis**

It is the intermediate stage where chronic pancreatitis has led to complications but clinical exocrine and endocrine function is still preserved. The type of complication is specified (e.g. stage B, bile duct). Stage B is accepted under the following conditions: Patients with complications but without steatorrhoea or diabetes mellitus.

Specific definition of chronic pancreatitis stage C

- **Stage C chronic pancreatitis**

Stage C is the end stage of chronic pancreatitis, where pancreatic fibrosis has led to loss of clinical exocrine and/or endocrine pancreatic function (steatorrhoea and/or diabetes mellitus). Complications of chronic pancreatitis might or might not be present. The type of exocrine and/or endocrine pancreatic function loss is specified (e.g. stage C, steatorrhoea).

Stage C can be sub-classified into three categories:

C1: Patients with endocrine function impairment.

C2: Patients with exocrine function impairment.

C3: Patients with exocrine/endocrine function impairment and/or complications—as they are defined.

Stage C is accepted under the following conditions: Patients with clinical manifestation of end-stage functional impairment with or without complications.

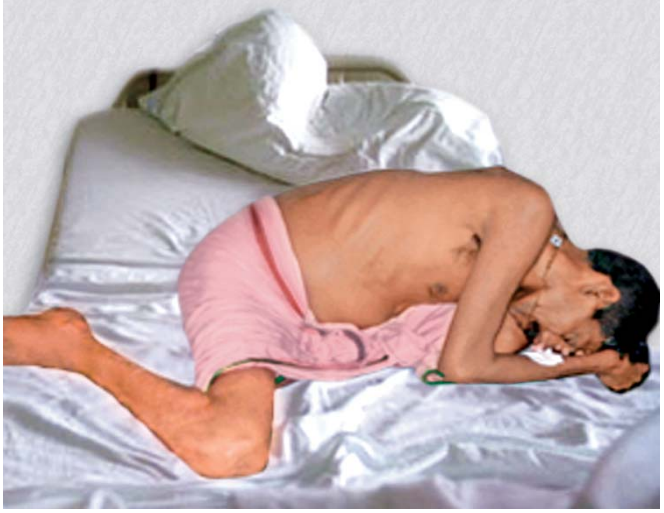

Fig. 14.21: Patient will occupy leaning forward position to relieve severe pain of pancreatitis.

Clinical Features

- *Pain in epigastric region* (80%)
  - It is persistent and severe, which radiates to back.
  - This pain is due to irritation of retropancreatic nerves, or due to ductal dilatation and stasis, or due to chronic inflammation itself.
  - *Two patterns of pain* have been described (Ammann and Muellhaupt). *Type A* pain is short relapsing episodes lasting days to weeks, with pain-free intervals. *Type B* pain is prolonged, severe, unrelenting pain.
  - Pain exacerbations *need not be always* associated with rise in amylase and lipase levels.
  - There is often a gradual diminish in pain over years due to "*pancreatic burnout*" by extensive calcifications, exocrine and endocrine insufficiency.
- *Exocrine dysfunction*: Diarrhoea, asthenia, loss of weight and appetite, steatorrhoea (signifies severe pancreatic insufficiency) (90%), malabsorption.
- *Endocrine dysfunction*: Diabetes mellitus. Pancreatic diabetes may often be *typically brittle* because of concomitant glucagon deficiency and *requires insulin*.
- *Mild jaundice* is due to narrowing of retropancreatic bile duct and cholangitis.
- *Mass per abdomen*, just above the umbilicus, tender, nodular, hard, felt on deep palpation, not moving with respiration, not mobile, resonant on percussion.

Triad of chronic pancreatitis

- Pancreatic calcification
- Steatorrhoea
- Diabetes mellitus

Triad is found in *less than one-third* of patients with CP

Mallet-Guys sign: In right knee-chest position, if left hypochondrium is palpated tenderness can be evoked in case of chronic-relapsing pancreatitis. In this position, bowel loops are being shifted to right so as to have a direct palpation of pancreas.

Chronic pancreatitis can lead to carcinoma pancreas.

Clinical presentations

- *Stage A*: 85%, recurrent / acute episodic pain with weight loss.
- *Stage B*: Severe prolonged progressive pain with impaired pancreatic function with cholestasis, pseudocyst, sinistral portal hypertension.
- *Stage C:* Severe exocrine / endocrine deficiency, less severe pain, complications like pseudocyst and obstruction.

Complications of chronic pancreatitis

- Pseudocyst of pancreas
- Pancreatic ascites
- CBD stricture due to oedema or inflammation
- Duodenal stenosis
- Portal thrombosis—segmental portal hypertension
- Peptic ulcer
- Carcinoma pancreas
- Pancreatic pleural effusion, pancreatic ascites
- Pancreatic fistula
- Splenic vein thrombosis
- Pancreatics enteric fistula

Differential Diagnosis

- Carcinoma of head of the pancreas.
- Retroperitoneal tumour.

Investigations

- Serum amylase—not very useful
- Blood sugar
- Liver function tests and prothrombin time
- U/S abdomen to see duct dilatation, stones, parenchyma, liver status, CBD (normally, diameter of pancreatic duct is < 3 mm)
- Endosonography
- CT scan—CT guided FNAC
- ERCP—*Chain of lake appearance*. Duct is dilated
- MRCP
- Plain X-ray shows calcification in 65% of patients
- Pancreatic secretory juice analysis
- *Pancreolauryl test*—pancreatic esterase cleaves fluorescein dilaurate after oral intake. Fluorescein is absorbed and quantified in urine after 2 days.
- Oral glucose tolerance test
- Faecal chymotrypsin and elastase analysis
- Bentiromide test

Investigations presently used are:

- *CT scan abdomen*: It is 95% reliable; to see pseudocyst, calcification, ductal stones, duct stricture and dilatation, vasculature, fibrosis, surrounding structures, CBD status. It shows 90% sensitivity; 95% specificity.
- *ERCP* is useful in chronic pancreatitis to see dilatations, strictures and altered ductal anatomy. It is mainly to assess *structural pathology* of the pancreas.

Cambridge grading of CP on ERCP

Grade	*Main pancreatic duct*	*Side branches*
I: Normal	Normal	Normal
II: Equivocal	Normal	<3 Abnormal
III: Mild	Normal	>3 Abnormal
IV: Moderate	Abnormal	>3 Abnormal
V: Severe	Abnormal with at least one of the following:	>3 Abnormal
	• Large cavity (>10 mm)	
	• Obstruction	
	• Filling defects	
	• Severe dilatation or irregularity	

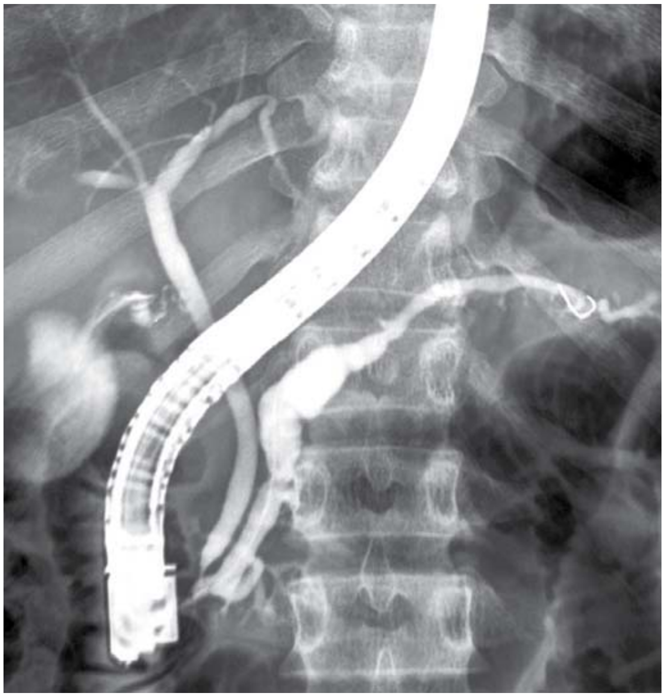

Fig. 14.22: ERCP showing pancreatic duct.

- *MRCP* is noninvasive method to see ductal anatomy.
- *Endosonography (EUS)* to see possible malignant site and to take FNAC. Site, duct status, stricture, stones, parenchyma, pseudocyst, CBD status, nodes are identified in EUS. Positive five parameters suggest chronic pancreatitis. It also helps in assessing operability.

EUS criteria in chronic pancreatitis—parenchymal	*EUS criteria in chronic pancreatitis—ductal*
• Gland atrophy	• Narrowing, dilation, irregularity
• Hyperechoic foci	• Calculi
• Hyperechoic stranding	• Side branch dilation
• Cysts, lobularity	• Hyperechoic walls

Rosemont EUS criteria

Major:
- A—Hyperechoic foci with shadowing and calculi in main pancreatic duct (PD)
- B—Lobularity with honeycombing

Minor:
- Cysts
- Dilated ducts ≥3.5 mm
- Irregular PD contour
- Dilated side branches ≥1 mm
- Hyperechoic duct wall, strands
- Nonshadowing hyperechoic foci
- Lobularity with noncontiguous lobules

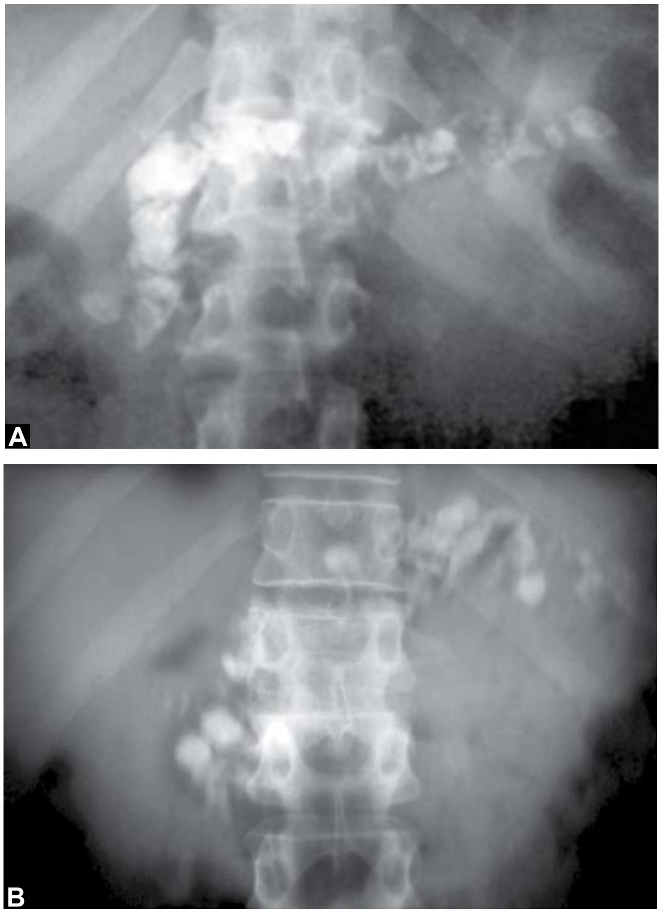

Figs 14.23A and B: Multiple stones in pancreatic duct with ductal dilatation. It needs removal of stones with pancreaticojejunostomy.

- *Secretin cholecystokinin test* is the gold standard for assessing pancreatic function. After over night fasting, double lumen tube is placed into the duodenum at the level of ligament of Treitz under C ARM guidance. Gastric and duodenal juices are aspirated. Continuous IV secretin 1 u/kg/hour and CCK are infused in 90 minutes. Sampling of duodenal juice is done in every 10 minutes for one hour for analysis of *volume, HCO_3*, amylase, lipase and proteases. It is to assess *functional deficiency*. In chronic pancreatitis volume is normal—> 2 ml/kg but less bicarbonate content —< 80 mEq/L. In malignancy volume is reduced due to obstruction—< 2 ml/kg; bicarbonate level is normal—> 80 mEq/L; enzyme levels are normal.

- *Pancreolauryl test*: After overnight fasting, blood sample is taken. IV secretin is injected 1 u/kg in 3 minutes. Test meal containing bread, butter, tea and fluorescein dilaurate is given to eat. IV metoclopramide 10 mg is injected. Blood samples are taken at 2, 2½, 3 and 4 hours to assess serum fluorescein concentration. It is also quantified in urine after 2 days.
- *LFT, prothrombin time* is needed to manage the patient.
- Serum trypsinogen < 20 ng/ml and fecal elastase < 200 mcg/g stool—severe exocrine insufficiency.

Treatment

Aim of treatment

- Control of pain
- Improvement in maldigestion and nutrition
- Management of complications

Conservative

- Avoid alcohol.
- Low fat, high protein, high carbohydrate diet; small and more frequent meals.
- Pancreatic enzyme supplements, vitamins and minerals, medium chain fatty acids.
- For pain—analgesics, splanchnic nerve or coeliac plexus block. *Pancreatic enzyme replacement therapy (PERT)* is used to relieve pain based on—negative feedback mechanism of pain; protease content responsible for pain relief (dose: >50,000 units). It may be beneficial only in females with idiopathic pancreatitis, small duct disease and pancreas divisum. Other drugs used are—antioxidants, amitriptyline, fluoxitine, octreotide. Novel pain therapies like NGF, antinociceptive agents, mast cell directed therapies are also used.
- Control of diabetes by oral hypoglycaemics or insulin.
- Somatostatin and its analogues. Its role is not clear.
- Repeated ascitic taps for pancreatic ascites.
- *Steatorrhoea* can be controlled by proton pump inhibitors with 1,50,000 units of oral lipase with low fat diet. PPI inhibits acid in stomach to prevent lipase getting inactivated in stomach.

Endoscopic Therapy in Chronic Pancreatitis

- It is used as it has got low morbidity.
- Main *indications* are—pain relief, ductal stones, main duct stricture, pseudocyst drainage, pancreatic ascites, effusion and fistula.
- It is quite useful for main pancreatic duct obstruction and pseudocyst. It is less useful for biliary stricture.
- *Procedures* are—pancreatic duct sphincterotomy, main ductal stone extraction using Dormia basket, main ductal stenting in strictures, stricture dilatation, ESWL of main duct stones, minor papillary sphincterotomy.
- *Pseudocyst* can be treated by *transmural stenting* across gastric wall if there is visible bulge, distance of cyst wall is less than 1 cm, no major vessel at puncture site. *Transpapillary stenting* is done if pseudocyst is communicating, if cyst is less than 6 cm, if there is no visible bulge in the stomach.
- *Complications* are—pancreatitis, stent displacement/occlusion, perforation, bleeding, sepsis, restenosis.
- *Advantages* of endotherapy are—less invasive, can be repeated, less severe complications, can be used as bridge therapy prior to surgery.
- *Disadvantages* are—long term benefits are less (50%). Definitive treatment is delayed. It is not useful in small duct disease; main duct dilatation decides the success rate.

Surgery

Indications for surgery

- Persisting pain—main indication
- Severe malabsorption
- Suspicion of malignant transformation
- Multiple relapses
- Complications like pseudocyst, segmental portal hypertension
- Biliary obstruction (*Wadsworth syndrome*)
- Pseudocyst
- Pancreatic ductal dilatation > 7 mm
- Pancreatic ascites/fistula
- Pancreatic ductal stenosis

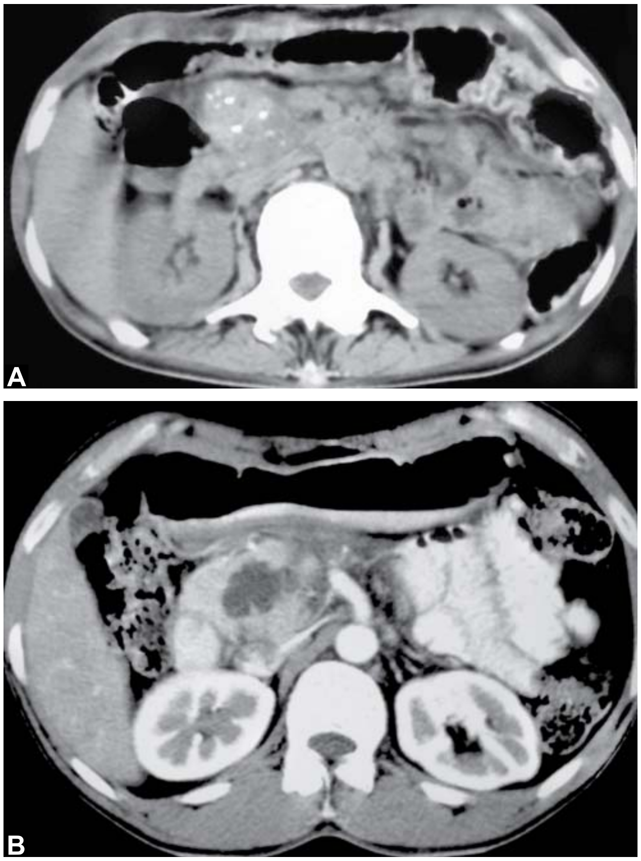

Figs 14.24A and B

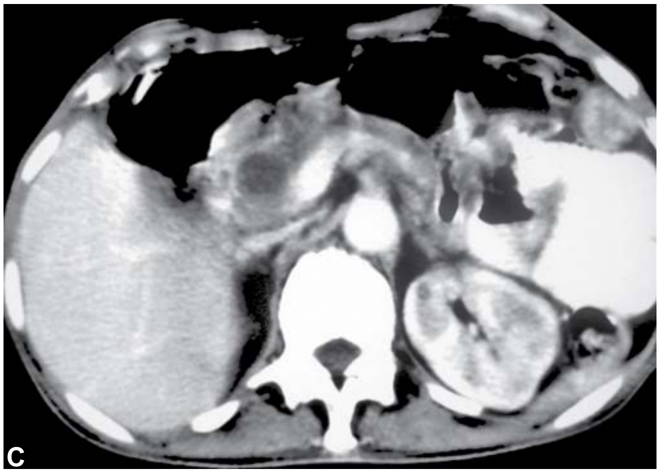

Fig. 14.24C

Figs 14.24A to C: CT scan abdomen showing features of chronic pancreatitis—calcification, ductal changes, and cyst formation.

Objectives of Surgery

- Pain relief when endotherapy has failed.
- Management of complications.
- Procedure is selected which shows less morbidity and mortality.
- Quality of life (QOL) should be preserved.

Problems and Limitations of Surgery

- Disease is irreversible parenchymal and ductal damage and so surgery *cannot cure* or reverse completely.
- Disease progression cannot be prevented or arrested by surgery.
- Risk of development of cancer continues.
- Complications of surgery are a drawback.

Benefits of Surgery in Relation to Pain Relief and Obstruction

- Relieves intraductal, interstitial, intracellular hypertension—through drainage procedures.
- Entrapped nerves are released and reduction in pancreatic enzyme reduces nerve irritation—resectional methods.
- *Pacemaker*—head of pancreas is removed in head coring procedure (Frey's) to relieve pain. It is less morbid and acceptable method.
- Resection is better than drainage procedure.

Principles of Surgery

- *Pancreatic duct decompression* (*drainage*) reduces the pain and retains the existing exocrine and endocrine functions. But chances of malignant transformation are still high. In significant number of patients recurrence of symptoms (50% recurrence of pain in 5 years) and progression of the disease pathology occurs and it does not give a complete cure. Normal diameter of pancreatic duct 4 mm in head; 3 mm in body; 2 mm in tail of the pancreas. *Pancreatic duct diameter more than 7 mm is an indication for surgery (pancreaticojejunostomy)*.
- *Pancreatic resection* (total pancreatectomy) is the actually ideal technique which relieves pain, removes entire diseased tissue. But technical difficulty; high surgical mortality (21%); and severe exocrine and endocrine deficiency (brittle diabetes) are the drawbacks. Head of pancreas is considered as pacemaker of the pancreatitis disease and so pancreaticoduodenectomy (*Whipple's*) is the other option to give adequate relief. It is preferred in patients who are having relatively normal body and tail of the pancreas. Occasionally when disease is limited to tail and body then *distal pancreatectomy* with or without splenectomy often with distal pancreatico jejunostomy may be done. But it is not beneficial if disease is extensive and diffuse. Resection is the main method of treatment if ductal dilatation is not adequate.
- A *combined resection and drainage procedures* are also done. Head is decored at various levels and decompressed duct is anastomosed to jejunum.

Surgeries

- 95% *subtotal pancreatectomy.*
- *Distal pancreatectomy*—Spleen, body and tail of the pancreas are removed—*Child's operation.*
- *Puestow's operation* [*Puestow-Gilesby* (1958)]—As the duct is dilated more than 8 mm, duct can easily be opened longitudinally. After removing all stones from the duct, it is anastomosed to the jejunum as *Roux-en-Y anastomosis.* In Puestow's operation spleen is removed.
- *Partington-Rochelle operation*: Here longitudinal pancreaticojejunostomy is done using almost entire laid open pancreatic duct. Spleen is retained in this procedure. This is now commonly done procedure.

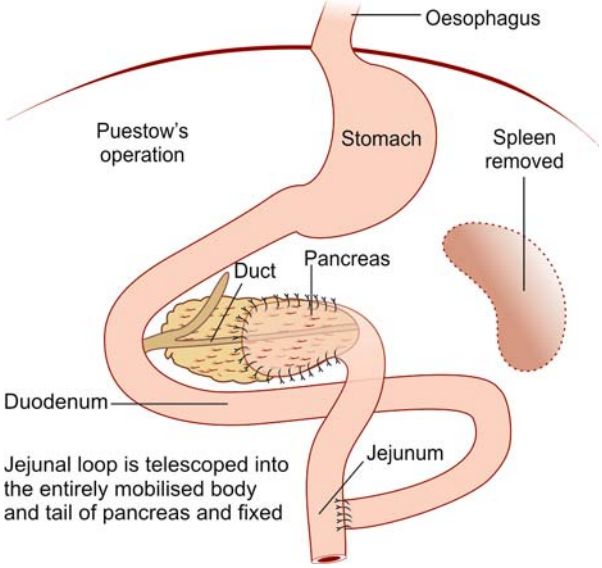
Oesophagus
Puestow's operation
Stomach
Spleen removed
Duct
Pancreas
Duodenum
Jejunum
Jejunal loop is telescoped into the entirely mobilised body and tail of pancreas and fixed

Fig. 14.25: *Puestow's pancreaticojejunostomy*. Here spleen is removed. Entire length of dilated duct is opened. Distal pancreas is mobilised up to superior mesenteric vessels. Pancreas with opened duct is telescoped into the Roux loop of jejunum and anastomosed.

Listen to the faintest sound; for opportunities knock only once.

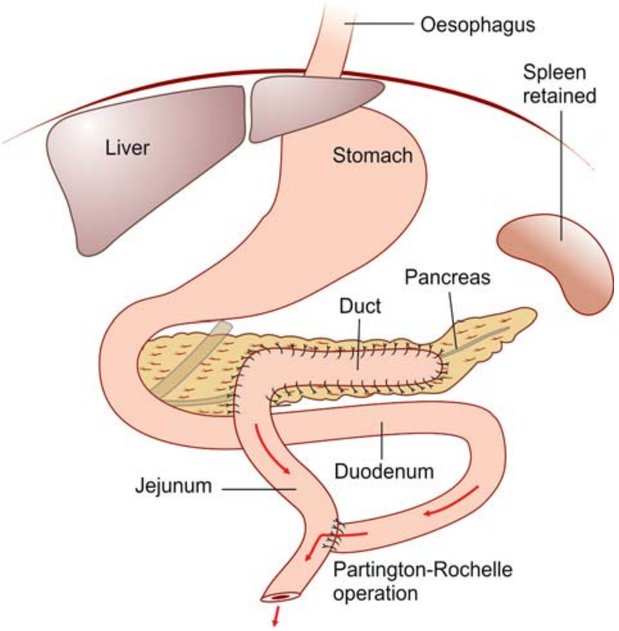
Oesophagus
Spleen retained
Liver
Stomach
Pancreas
Duct
Jejunum
Duodenum
Partington-Rochelle operation

Fig. 14.26: *Partington-Rochelle operation*. Here spleen is retained and longitudinal pancreaticojejunostomy is done to the opened duct from front.

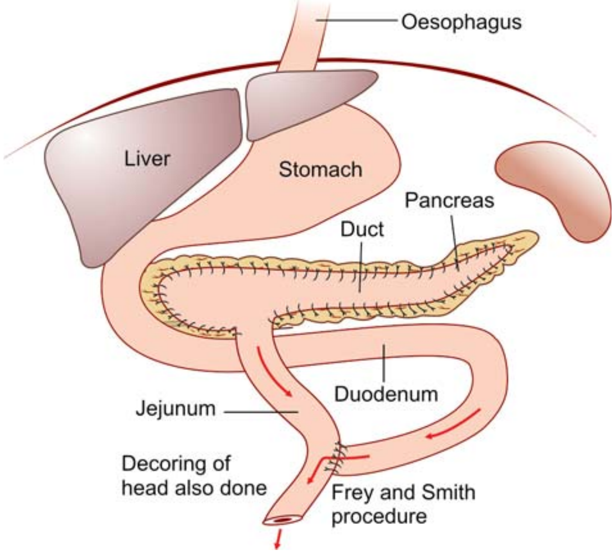
Oesophagus
Liver
Stomach
Pancreas
Duct
Jejunum
Duodenum
Decoring of head also done
Frey and Smith procedure

Fig. 14.27: *Frey's procedure*. Here head of pancreas is decored and opened duct with decored head is anastomosed to Roux loop of jejunum.

- *Total pancreatectomy* is indicated when entire gland is diseased. It relieves the pain and also prevents the diseased pancreas from turning into malignancy. Patient has to take insulin and oral pancreatic enzymes permanently (*brittle diabetes*).
- The cause is treated like cholecystectomy for gallstones.
- Therapeutic ERCP is useful in removal of stones in dilated duct.
- If disease mainly involves head of the pancreas, then pancreaticoduodenectomy can be done—*Whipple's procedure.*
- Duodenal preserving resection of head of pancreas in front of portal vein with jejunal loop anastomosis to transected neck of pancreas—*Beger procedure.* Here extensive resection of head (than Frey's) is done.

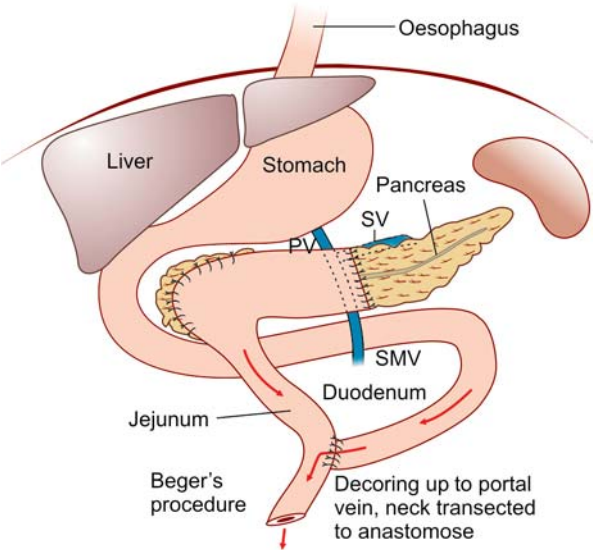
Oesophagus
Liver
Stomach
Pancreas
SV
PV
SMV
Duodenum
Jejunum
Beger's procedure
Decoring up to portal vein, neck transected to anastomose

Fig. 14.28: *Beger's procedure*. Here head is decored up to the visible portal vein. Neck is transected. Distal end of the pancreas with retained part is anastomosed to Roux loop of jejunum.

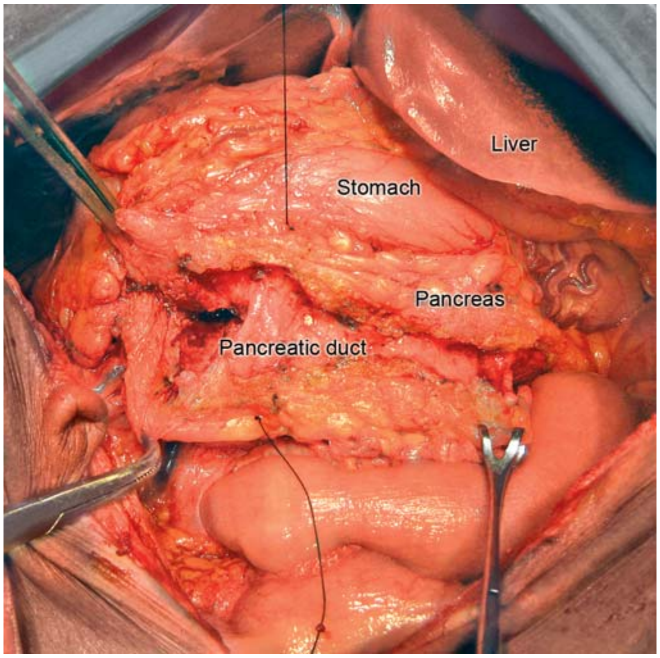
Liver
Stomach
Pancreas
Pancreatic duct

Fig. 14.29: Dilated pancreatic duct is laid open; multiple stones are removed and duct is ready for longitudinal pancreaticojejunostomy (LPJ).

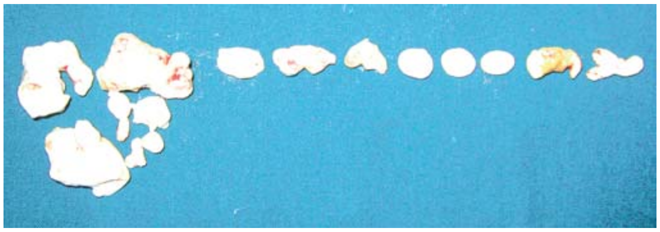

Fig. 14.30: Pancreatic duct stones—removed during pancreaticojejunostomy.

- *Longitudinal pancreaticojejunostomy* after excision of peri-pancreatic duct tissue—here superficial part of the head of pancreas is removed to achieve improved drainage—*Frey procedure*. It is done when ductal dilatation is not adequate; head is more than 4 cm thick. Head coring is done with retaining 5 mm thick tissue in front of veins, close to duodenum. It shows 75% pain relief in 3 years.

- Resection of tail of pancreas with retrograde pancreaticojejunostomy—*Duval procedure*.

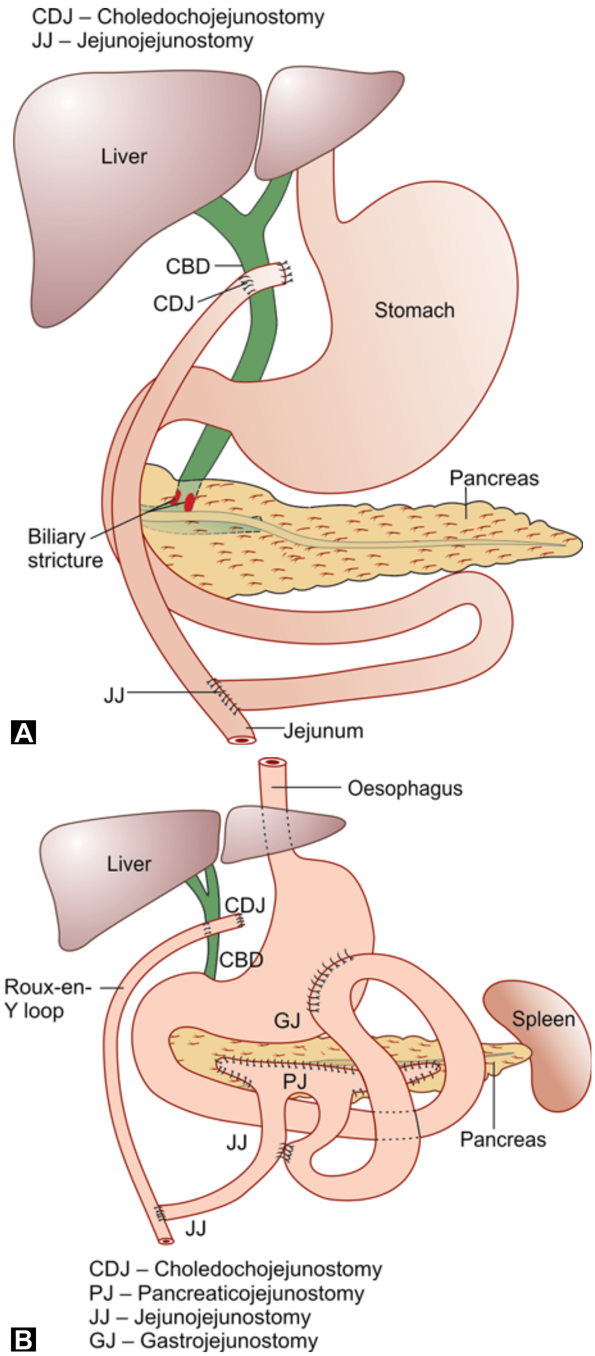
CDJ – Choledochojejunostomy
JJ – Jejunojejunostomy
Liver
CBD
CDJ
Stomach
Pancreas
Biliary stricture
JJ
Jejunum
A
Oesophagus
Liver
CDJ
CBD
Roux-en-Y loop
GJ
Spleen
PJ
JJ
Pancreas
JJ
CDJ – Choledochojejunostomy
PJ – Pancreaticojejunostomy
JJ – Jejunojejunostomy
GJ – Gastrojejunostomy
B

Figs 14.31A and B: *Choledochojejunostomy*. It is Roux-en-Y anastomosis. It is done whenever there is complication of biliary stricture after pancreatitis. Often triple anastomosis may be required—pancreaticojejunostomy, choledochojejunostomy and gastrojejunostomy. It is done whenever there is duodenal obstruction, biliary block and pancreatic duct dilatation (drainage procedure).

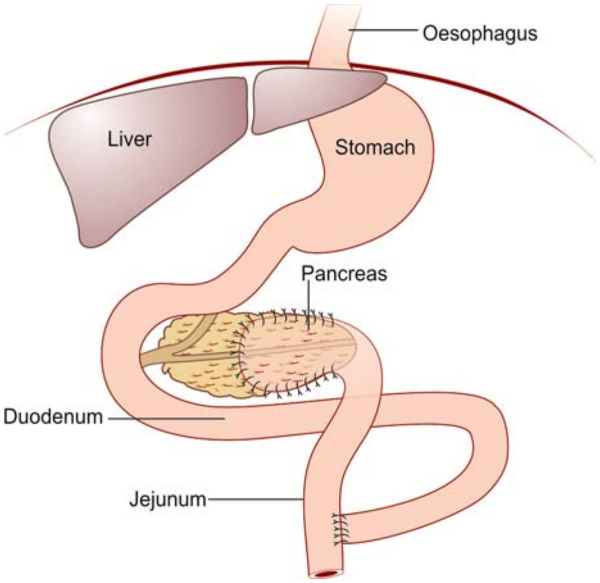
Oesophagus
Liver
Stomach
Pancreas
Duodenum
Jejunum

Fig. 14.32: *Retrograde pancreaticojejunostomy* after resection of the tail of the pancreas (Duval).

- When duodenal, biliary and pancreatic obstruction is present—choledochojejunal, pancreaticojejunal and gastrojejunal anastomosis may be needed—*triple anastomosis*. Only biliary stricture which can not be managed by ERCP needs choledochojejunostomy.

Complications of surgery:

- Pancreatic leak/fistula (10%); infection; bleeding; recurrence; brittle diabetes.
- *Pain relief* in chronic pancreatitis is achieved by drugs; decompression/drainage surgeries, resection surgeries, epidural analgesia, coeliac ganglion block, operative chemical splanchnicectomy, extra/intra peritoneal right and left splanchnicectomy, transhiatal bilateral splanchnicectomy, thoracoscopic splanchnicectomy.

Postoperative care

- Nutrition—TPN/jejunostomy feed
- Fluid and electrolyte management
- Prevention/control of sepsis
- Proper monitoring
- Octreotide on table and postoperatively—regular intervals or slow infusion—5 days

Surgery for chronic pancreatitis

Drainage	*Resection*
• Technically easier	• Technically demanding
• Less mortality < 2-5%	• Mortality 8-21%—very high
• Adequate pain relief	• Used when carcinoma is suspected or localised disease
• Recurrence of pain in 5 years—50%	• Head is the pacemaker of the disease
• Diseased tissue is left behind	• Whipple's is preferred
• Disease progression	• Total pancreatectomy causes high mortality, brittle diabetes; though ideal it is not done commonly
• Fear of existing occult carcinoma (17%) or later onset	

PANCREATIC TUMOURS

Classification

A. ***Exocrine tumours***

Benign: Benign cystadenoma. It is rare.

Malignant:

1. Adenocarcinoma in ampulla or periampullary region or head of pancreas. Periampullary carcinoma may arise from any of the component—duodenal mucosa or CBD or pancreatic duct component or all. Occasionally squamous cell carcinoma, or combination of adeno squamous can occur.
2. Cystadenocarcinoma of pancreas occurs commonly in body and tail of the pancreas, which usually attains a large size (5%).

B. ***Endocrine tumours***

1. Insulinoma (β cells)	Whipple's triad
2. Gastrinoma (G cells)	Peptic ulcer.
3. Glucagonoma (α cells)	Diabetes, necrolytic migratory erythema.
4. Vipoma—Pancreatic cholera (Verner-Morrison syndrome)	Watery diarrhoea, hypokalaemia achlorhydria. (WDHA syndrome)
5. Somatostatinoma. (S cells) δ cells	Diabetes, steatorrhoea, gallstones.

C. ***Lymphomas***

EXOCRINE PANCREATIC TUMOURS

Exocrine Pancreatic Tumours can be Solid (75%) or Cystic

Solid neoplasms

- *Infiltrating ductal adenocarcinoma* (85%) is common in head (yellowish white mass), neck and ampulla. There is evidence of ductal/glandular differentiation; with intense non-neoplastic desmoplastic stromal reaction; haphazard infiltrative growth pattern; growth along the nerve roots; and lymphovascular invasion. Initial early non-invasive epithelial proliferation is called as *pancreatic intraepithelial neoplasia* (PanIN). Immunohistochemistry shows CA 19-9; mucins (MUC1, 3, 4, 5); CEA; cytokeratins 9, 19. 5 years survival is < 5%. After surgery it may be 20%. *Signet ring, colloid, anaplastic types are variants.*
- *Acinar cell carcinoma* shows sheets of acinar structures with often metastatic fat necrosis, eosinophilia, polyarthralgia. Presence of zymogen granules; antibodies against trypsin / chymotrypsin and lipase are typical.
- *Pancreatoblastoma* are malignant tumour seen commonly in children.

Cystic neoplasms

- They mimic closely pseudocyst of pancreas.
- Cystic fluid analysis will commonly differentiate the problem. Pseudocyst is the fluid of low viscosity, with low levels of CEA, and tumour markers like CA 15/3, CA 72/4 but high levels of amylase with inflammatory cells on cytology. In cystic neoplasms fluid shows high viscosity, high CEA (mucinous) but low amylase and tumour markers with tumour cells on cytology.
- They are 1% of all pancreatic tumours; 10% of all cystic lesions of pancreas.
- They can be *mucinous* (commonest) or *serous cystadenoma*; *mucinous ductal ectasia of pancreas* (common in elderly males, common in head, can turn into malignancy); *papillary cystic tumours, teratomas,* etc. (Note: cystic islet tumours also cystic neoplasm but are endocrinal nonfunctioning with 30% malignant potential).

Types of cystic neoplasms

- ***Mucinous cystic neoplasms*:**
  - They are more common.
  - They are common in females (90%) and often seen in early age group.
  - They are common in *tail and body*. They do not communicate with main pancreatic duct.
  - It can be unilocular or multilocular. Mucous secreting columnar epithelium (mucin rich goblet ells) with dense ovarian like stroma is typical histology. It is not associated with extrapancreatic neoplasm.
  - It can be mucinous cystadenoma; borderline; *in situ*; mucinous cystadenocarcinoma. Mucinous cystadenoma can turn into malignancy. Increased mitosis, atypia, tumour markers like CA 15/3, CA 72/4 will be positive with over expression of EGFR.
  - Mucinous tumours may show high levels of CEA but low levels of amylase.
  - Presentations may be mass abdomen (75%, large epigastric mass, non-mobile, not moving with respiration), abdominal pain, back pain, weight loss.
  - *Investigations*—CT scan; EUS; EUS guided aspiration for cytology, fluid analysis for CEA (↑), amylase (↓), angiography shows hypervascularity.
  - *Treatment:* Depends on location of tumour either Whipple's operation; distal pancreatectomy; median pancreatectomy.
  - Survival rate for mucinous cystadenocarcinoma is 50%.

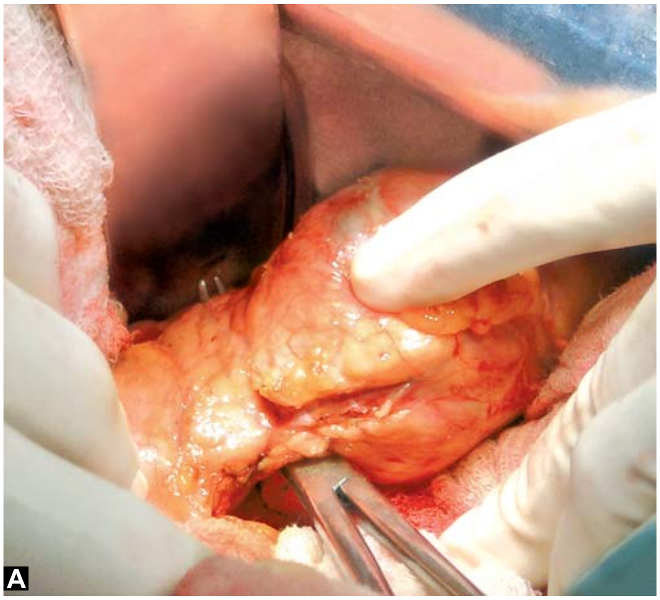

Fig. 14.33A

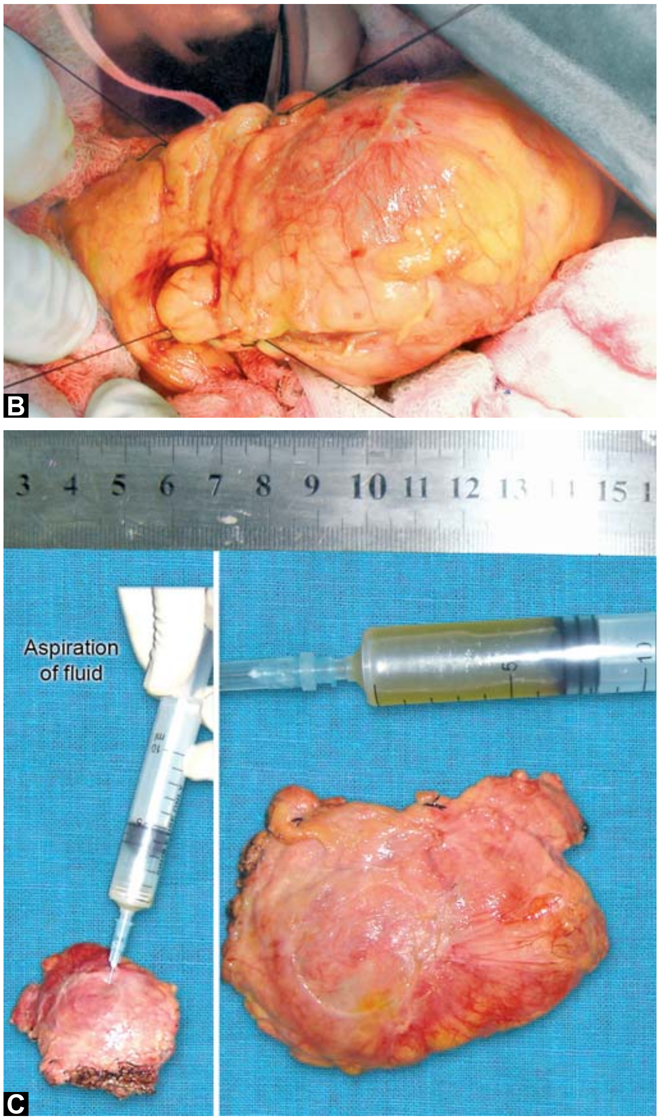

Figs 14.33B and C

Figs 14.33A to C: Mucinous cystadenoma of pancreas. Only cystadenoma is removed and specimen was aspirated for fluid content (*Courtesy:* Dr Arunkumar, MCh).

- ***Serous cystic neoplasms***
    - They are almost always benign and are common in females (2:1).
    - It can be *microcystic adenomas* (90%) or *oligocystic adenomas* (10%).
    - Extrapancreatic neoplasm can be associated (20%) with it.
    - It is more common in *von Hippel Lindau syndrome*.
    - Abdominal pain, backache, features of pancreatitis (5%), jaundice (10%) and palpable mass in epigastrium (65%) are the presentations.
    - It is common in *head of pancreas*, usually solitary cells which contain only glycogen but *no mucin*.
    - Microcysts of 3-6 mm size within the cyst is common; cyst is lined by cuboidal epithelium.
    - CT scan shows central calcification with enhancing septae. EUS is very useful investigation.
    - Treatment: Central or distal pancreatectomy.

Intraductal papillary mucinous neoplasms (IPMN) (Ohashi 1982):

- It is '*intraductal papillary mucinous* neoplasm with tall, columnar, mucin containing epithelium with or without papillary projections, involving main pancreatic duct and or the branch ducts'.
- It is common in 7th decade; equal in both sexes.
- It can be ***main duct IPMN*** with duct (75% proximal part) dilatation ≥ 1 cm containing mucous bulges through ampulla; ***branch—duct IPMN*** involving side branches of the non-dilated main duct but communicating to main duct mainly involving uncinate process, often head, neck and body, seen as multifocal disease. ***Combined IPMN*** involving both main duct and branch ducts is also possible.
- IPMN can be benign; borderline; malignant *histologically*. 70% main duct IPMN and 25% of branch duct IPMN is associated with *malignancy*. *Invasive type* is more common with main duct IPMN. Adenoma—carcinoma sequence is commonly observed.
- IPMN is also classified as—PAN IN 1 with absence of dysplasia; PAN IN 2 with moderate dysplasia; PAN IN 3 with carcinoma in situ.
- Main duct IPMN presents *with* abdominal pain, features of pancreratitis, steatorrhoea where as branch duct IPMN can be asymptomatic. Jaundice, recent onset/worsening diabetes occur once it is *malignant*.
- ERCP shows *triad of Ohashi—a bulging ampulla of Vater, mucin secretion and dilated main pancreatic duct*. *Dynamic MR* is very useful. Endosonography and helical CT are useful. EUS aspiration is also useful.
- IPMN should be *identified first from other cystic* diseases of the pancreas; then *type* should be assessed; later *invasiveness* should be evaluated.
- *Features of malignant IPMN are*—presence of jaundice/steatorrhoea/worsening diabetes/cyst > 3 cm; thick wall cyst; nodules/papillary lesions; main duct > 1 cm; *raised CEA* (120 ng/ml) in pancreatic juice/cyst fluid.
- **Treatment:** Main duct IPMN should be resected always. Branch duct IPMN < 3 cm without nodules can be observed carefully. Middle pancreatectomy is done when needed in branch duct IPMN. Multifocal disease needs total pancreatectomy or extended resections to achieve clearance. Nodal spread in malignant IPMN is 45%. Long-term survival is good for noninvasive type.

CARCINOMA PANCREAS

- Carcinoma pancreas is higher in men.
- It is common in African American males.
- 80% of pancreatic cancers are metastatic at the time of first diagnosis.
- It is common in Jewish heritage and native Hawaiians.
- Its incidence is 9 new cases per 100000 people. Mean age is 60-65 years.

Aetiology

- Smoking.
- High energy diet rich in fat.

Surgical Triad: *Measure thrice, think twice, cut once.*

- Chronic pancreatitis.
- Familial pancreatitis.
- Diabetes mellitus.
- Carcinogens like benzidine.
- Hemochromatosis with pancreatic calcification.
- Cirrhosis, obesity.
- Occupational exposure to carcinogens like DDT, benzidine.
- Previous cholecystectomy.
- 85% show mutant *K ras gene on codon 12*; 60% show mutation of p53 gene in chromosome 17; over expression of EGFR.
- *Peutz-Jegher syndrome*, *HNCC* (Hereditary Nonpolyposis Colonic Cancer—Lynch II type), ataxia telangiectasia, *hereditary breast and ovarian cancers, hereditary atypical multiple mole melanoma syndrome*, familial adenomatous polyposis (FAP).

Sites

- Head and neck region } 70%
- Ampullary and periampullary region } 70%
- Body and tail—30%

Periampullary Carcinoma

It is tumour arising at or near the ampulla.
It could be:

- Adenocarcinoma from head of pancreas close to the ampulla—50%.
- Tumour from ampulla of Vater—30%.
- Distal bile duct carcinoma—10%.
- Duodenal carcinoma adjacent to ampulla—10%.

Pathology

Gross

- Grayish white scirrhous nodular gritty tumour in the head is the usual gross look. More than 3 cm sized tumour shows nodal and distant spread commonly. Induration, areas of haemorrhage and necrosis are common.
- Pancreatic duct may be dilated due to obstruction.
- Carcinoma head of pancreas present earlier than carcinoma body and tail of pancreas.

Microscopy

- It shows malignant cells with variable degree of differentiation, high mitotic index, severe *desmoplastic reaction* with fibrosis, a *halo* around the tumour due to obstruction of the adjacent ducts caused by the tumour leading into chronic pancreatitis in close proximity.
- Along with vascular and lymphatic invasion tumour has got high propensity for *perineural spread* and so to neural plexus causing back pain.
- Colloid (mucinous noncystic), signet cell, squamous, adenosquamous, anaplastic, sarcomatoid are variants of duct cell carcinoma.

Spread

Local Spread

- To adjacent structures like duodenum, portal vein, superior mesenteric vein, retroperitoneum.
- Spread is more likely in carcinoma head of pancreas than in periampullary carcinoma.
- Spread is common once carcinoma head of pancreas becomes more than 3 cm in size.
- Perineural spread is common.

Nodal Spread

- Usually to perihepatic nodes around the duodenum and CBD, subpyloric, celiac nodes.
- Hard dark greenish nodes are typical. Often nodal enlargement may be due to just reactive hyperplasia.

Distant Spread

- To *liver* commonly as multiple secondaries in both lobes.
- Occasionally to lungs, adrenals, bran, bone, etc.

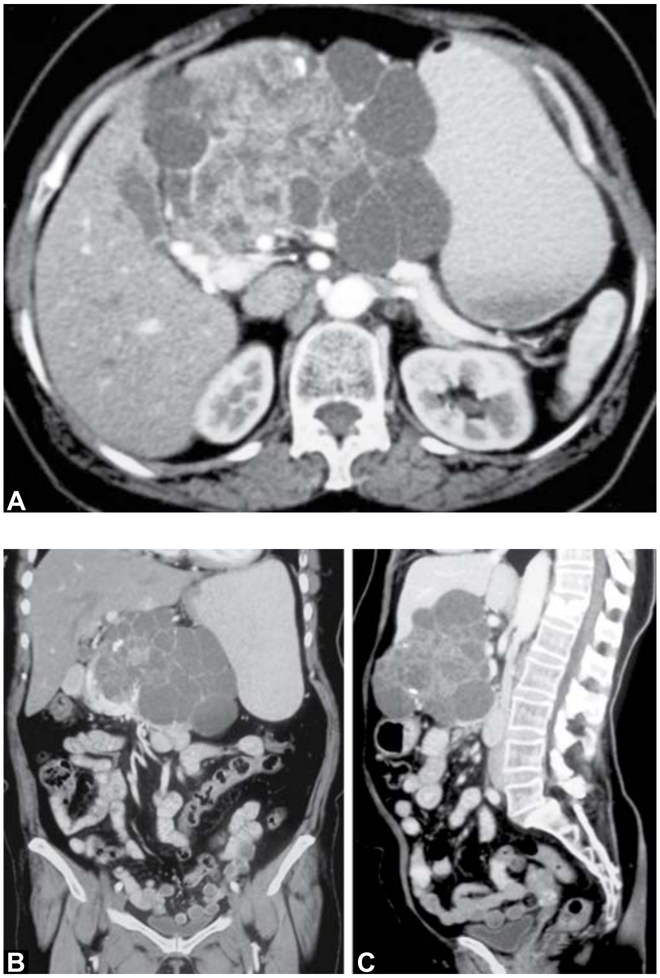

Figs 14.34A to C: CT scan pictures of cystadenocarcinoma of pancreas.

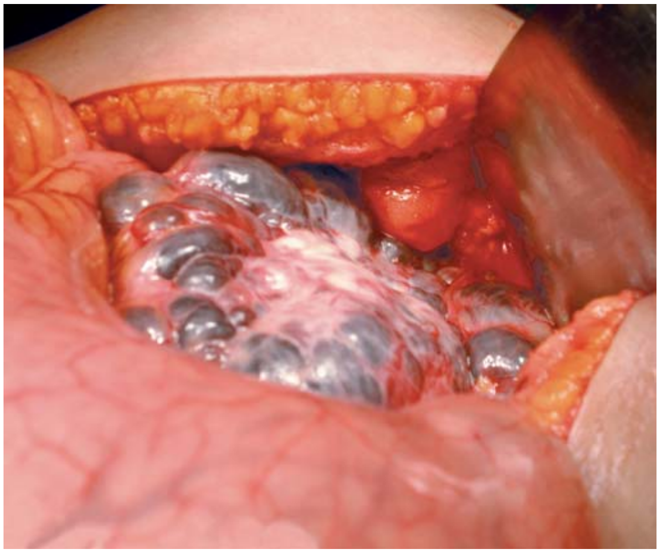

Fig. 14.35: On table finding of cystadenocarcinoma of pancreas.

Clinical Features

Presentations

a. Ampullary tumours mainly present with *jaundice and weight loss.*
b. Carcinoma of head and neck region present with weight loss and jaundice.
c. Cystadenocarcinoma of pancreas present with pain, weight loss and mass.

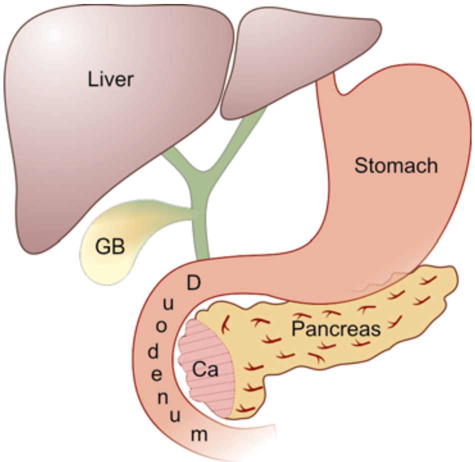

Fig. 14.36: Carcinoma head of the pancreas.

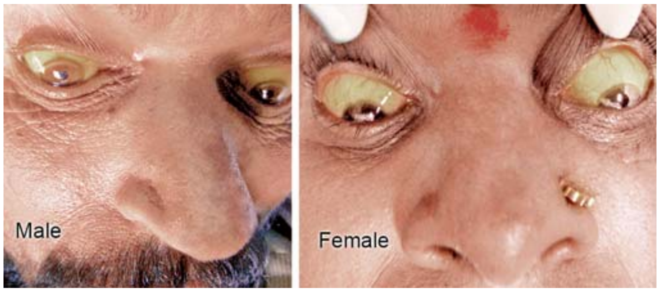

Fig. 14.37: Obstructive jaundice due to periampullary carcinoma in male and female patients.

Features

- *Jaundice is of obstructive nature* which is of short duration, severe, progressive, associated with pruritus (due to deposition of bile salts in the skin which releases histamine). Painless jaundice is seen in ampullary malignancies.
- In periampullary carcinoma sometimes necrosis of tumour occurs, as a result of which jaundice may reduce temporarily thus being *intermittent.*
- Pain in the right hypochondrium, epigastrium, or left hypochondrium depending on location of the tumour.
- Back pain, when present, is due to involvement of retropancreatic nerves, or pancreatic duct obstruction or stasis, disruption of nerve sheath by tumour. Pain is more at night, after food and on recumbency; it is relieved by leaning forward.
- Diarrhoea, steatorrhoea, acholic stools, tea colored urine.
- Silvery stool (due to mixing of undigested fat with metabolised blood derived from the ooze of periampullary growth).
- Loss of appetite and weight.
- Scratch marks on the back.
- *Migratory superficial thrombophlebitis—Trousseau's sign* (10%) is due to release of platelet aggregating factors from the tumour or its necrotic material (Trousseau himself died of carcinoma pancreas who had migrating thrombophlebitis).
- Ascites.
- Left supraclavicular palpable lymph node.
- Secondaries in rectovesical pouch (Blumer's shelf).
- Gallbladder may be palpable which is nontender, soft, globular, smooth, moving with respiration, mobile horizontally, dull on percussion (30% in carcinoma head of pancreas; 50% in periampullary carcinoma). *Courvoisier law* favours gallbladder enlargement.
- Liver is enlarged, smooth, firm, nontender, due to dilated bile filled biliary radicles—*Hydrohepatosis*. Liver can show multiple hard nodules due to secondaries (70%).
- Cystadenocarcinoma of pancreas can present with mass in epigastric region, which is nonmobile, not moving with respiration, smooth, soft, nontender.
- Splenic vein thrombosis with splenomegaly (10%) can occur.

Cystadenocarcinoma of the pancreas

- Common in body and tail of the pancreas
- Epigastric pain which is radiating to back is common
- Mass in epigastrium which is nonmobile, smooth, soft/hard but not moving with respiration. It is resonant, does not fall forward
- Jaundice is not observed
- May mimic pseudocyst of pancreas or retroperitoneal tumours
- 1% of pancreatic malignant tumours
- Papillary cystadenocarcinoma is common
- Mucinous cystic tumours show high levels of CEA
- CT scan is diagnostic
- Distal pancreatectomy is the procedure of choice. In large tumours total pancreatectomy is required
- It has got better prognosis

To overcome rigidity of abdomen in refractory cases, with the base of the left hand pressing upon the lower part of the sternum, thoracic respiration is impeded and the abdominal muscles relax. **—Neville J Nicholson**

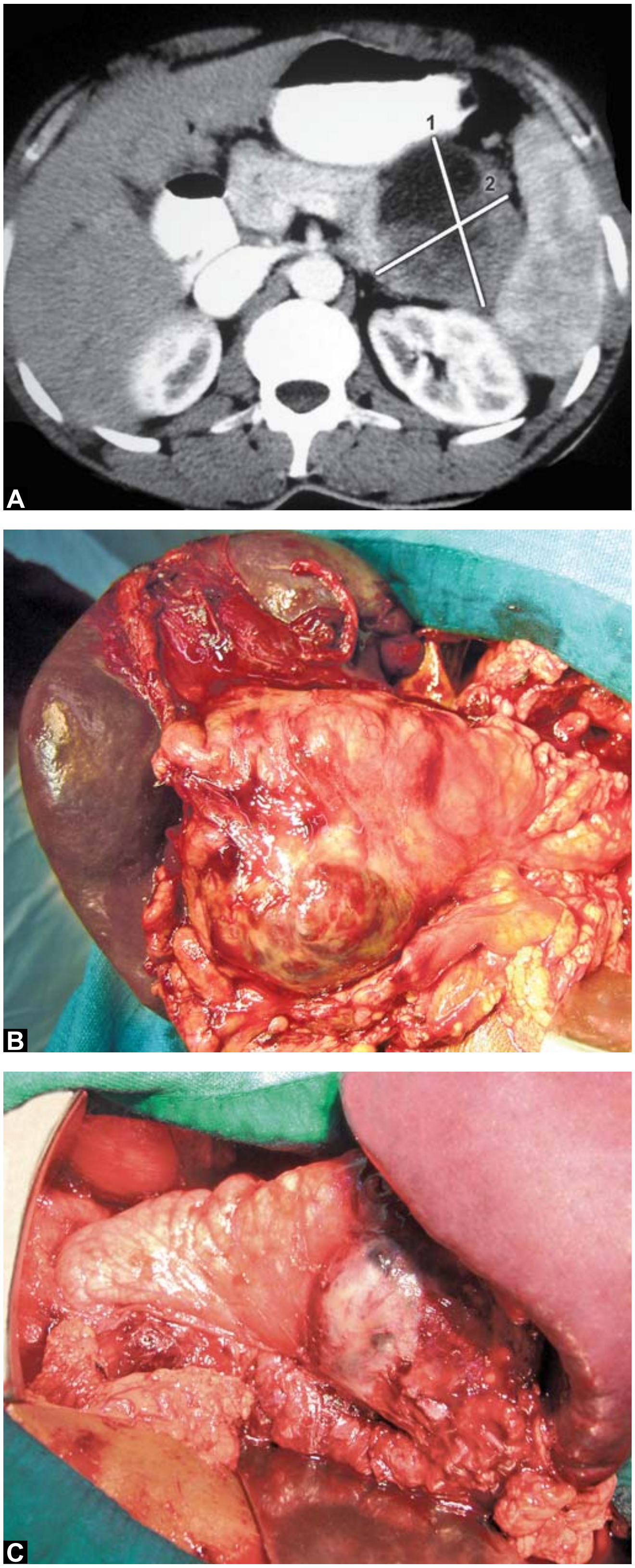

Figs 14.38A to C: Mucinous cystadenocarcinoma of distal body and tail of pancreas—CT picture and on table look. Distal pancreatectomy was done (*Courtesy*: Dr Arunkumar, MCh).

Differential diagnosis

- Retroperitoneal mass/tumour/lymph nodes
- Advanced adherent carcinoma of stomach
- Advanced carcinoma of transverse colon
- CBD stone
- Bile duct stricture
- Lymph node compressing CBD
- Cholangiocarcinoma of CBD
- Chronic pancreatitis

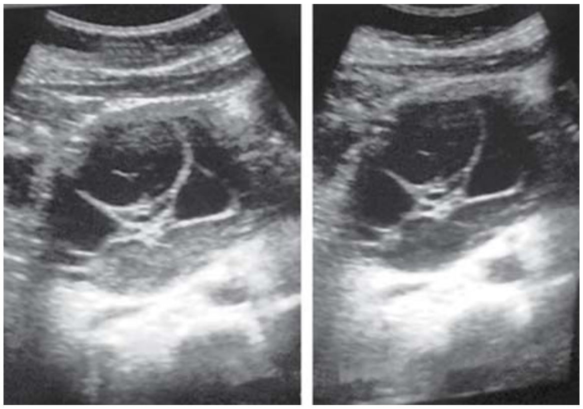

Fig. 14.39: Ultrasound showing mass lesion in the pancreas.

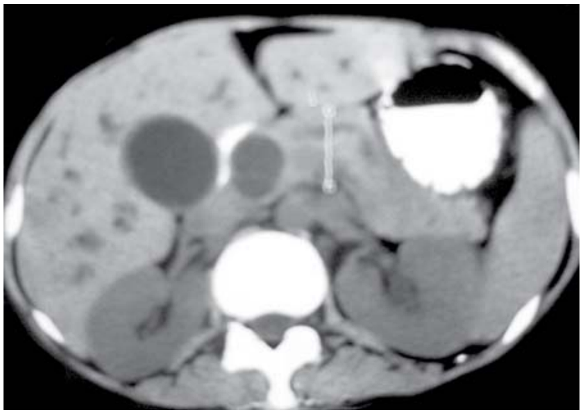

Fig. 14.40: CT picture of carcinoma pancreas showing dilated CBD due to obstruction.

Investigations

- Liver function tests: Serum bilirubin, direct component (conjugated) is increased (van den Bergh's test). Serum albumin is decreased with altered A : G ratio.
  Prothrombin time is widened. Serum alkaline phosphatase is increased.
- U/S abdomen to see gallbladder, liver, growth, CBD size (normal diameter is < 10 mm), lymph node status, portal vein, ascites.
- Barium meal shows widened duodenal "*C" loop—pad sign*. *Reverse 3 sign* is seen in carcinoma—periampullary region.

Differences between presenting features of carcinoma of head of pancreas and periampullary carcinoma of pancreas

		Carcinoma of head of pancreas	*Periampullary carcinoma*
1.	Pain and weight loss	Early features	Late features
2.	Jaundice	Persistent and progressive	Intermittent
3.	Occult blood in stool	Absent	Present; stools are silvery
4.	Endoscopic examination	Growth not visible	Growth visible
5.	Prognosis	Not good	Good

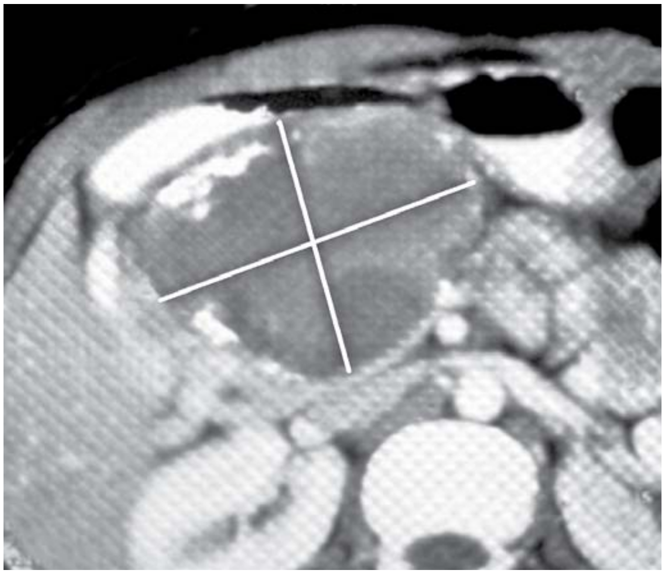

Fig. 14.41: CT picture showing cystadenocarcinoma of pancreas.

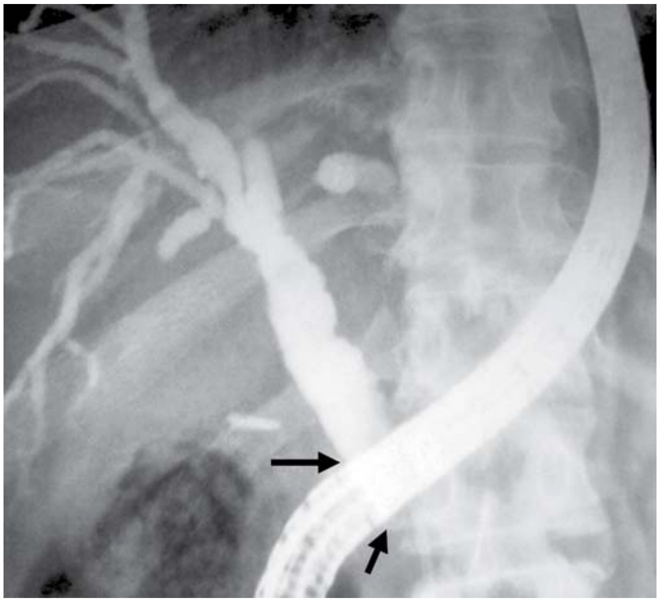

Fig. 14.42: ERCP showing growth in the distal CBD.

- Spiral CT scan (ideal) shows portal vein infiltration, retroperitoneal lymph nodes, size of the tumour.
- ERCP with pancreatic juice cytology or brush biopsy is useful. MRCP to see biliary tree.
- CA 19-9 (carbohydrate antigen) is a useful tumour marker. More than 37 units/ml is significant with 85% sensitivity and specificity. CA 494 is useful to differentiate it from chronic pancreatitis. CEA is also often done in suspected cystadenocarcinoma.
- *Coeliac* and superior mesenteric angiogram can be done to reveal tumour circulation and invasion.
- Gastroduodenoscopy reveals ampullary tumour and biopsy can also be taken.
- *Endosonography* is useful to see the invasion and size, to stage and to do EUS guided FNAC.
- Urine for bile salts (*Hay's test*), bile pigments (*Fouchet's test*), urobilinogen.
- *Role of biopsy*:
    - Trucut biopsy and other biopsies are not advisable as it may cause bleeding, infection, leak and negative biopsy will not rule out malignancy. Biopsy is not done in potentially resectable tumour; it is probably useful in inoperable cases to start palliative therapy (chemotherapy). Transduodenal/transgastric or EUS guided trucut biopsy may be useful in such situation.
- Laparoscopy to assess. Laparoscopy and laparoscopic US is useful for staging and to identify peritoneal deposits which prevents unnecessary laparotomy. Palliative laparoscopic bypass like choledochojejunostomy can also be done.
- CT angiogram to see the vascularity. It also provides idea about resectability of the tumour in 70% cases.
- *PTC is useful* if ERCP fails to detect and assess the site of lesion proximally; to decompress the obstructed biliary system through a fine catheter left *in situ* or to place an endobiliary prosthesis across the obstruction.

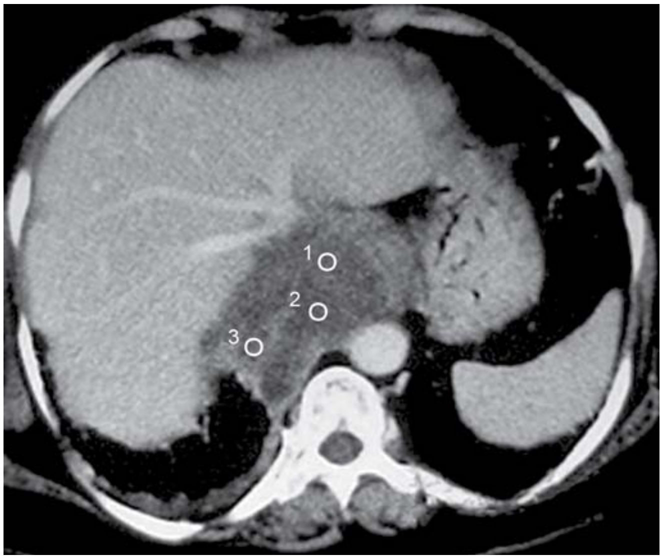

Fig. 14.43: CT scan abdomen showing features of carcinoma of pancreas.

ERCP signs	***Important investigations***
• Abrupt block of pancreatic duct with irregular stricture • Pancreatic duct encasement • *Double duct sign*—cut off of both pancreatic and bile duct • Parenchymal filling • Scrambled egg appearance	• Spiral CT/3D CT is ideal—to detect operability, portal vein invasion, size, extent, pancreatic and biliary tree and nodal status • ERCP to take biopsy/brush cytology, stenting, to see the luminal extent of the tumour • MRCP is non invasive—entire biliary tree is visualised properly • Endosonography (EUS)—to assess size of primary properly; to take endosonographic FNAC • PTC often with PBD if ERCP fails
Barium meal X-ray	***Investigations to stage the disease***
• Is rarely done in periampullary carcinoma • *Rose thorn* appearance in hypotonic duodenography • *Reverse 3 sign* in periampullary carcinoma • *Pad sign*—widened C loop of duodenum in carcinoma head of pancreas • Features of Gastric outlet obstruction	• 3D/spiral CT scan • EUS • Laparoscopy • ERCP

TNM staging of carcinoma pancreas	***Metastases—M***
Tumour—T	Mx – Distant spread cannot be assessed
Tx – Primary cannot be assessed	M0 – No distant spread
T0 – No evidence of tumour	M1 – Distant spread present
Tis –Carcinoma in situ	
T1 – Growth limited to pancreas < 2 cm	
T2 – Growth limited to pancreas > 2 cm	
T3 – Extension to duodenum or bile duct	
T4 – Extension to portal vein, SMV, SMA, stomach, spleen, colon, celiac axis	
Nodes—N	Staging
Nx – Regional nodes cannot be assessed	0 – Tis N0 M0
N1 – No nodal spread	IA – T1 N0 M0
N2 – Nodal spread present	IB – T2 N0 M0
	IIA – T3 No Mo
	IIB – T1,2,3 N1 M0
	III – T4 Any N M0
	IV – Any T AnyN M1

Resection status

R0 – No residual tumour found after resection
R1 – Microscopic residual tumour after resection
R2 – Macroscopic residual tumour after resection

Preoperative Preparation in Carcinoma Pancreas

♦ Adequate hydration is important in prevention of dehydration postoperatively. Dehydration is common in obstructive jaundice.

♦ Glycogen reserve in liver will be inadequate which should be replenished by giving preoperative glucose orally or intravenously.

♦ Patient is prone to develop hepatorenal syndrome postoperatively, leading into renal failure due to sludging of the bile salts, due to toxins and sepsis; and so mannitol should be given intravenously for 3 days prior to surgery to flush the kidney (200 ml IV twice a day).

♦ Injection vitamin K 10 mg IM for 5 days.

♦ ERCP stenting is done to drain bile if serum bilirubin is high (> 30 mg %). Surgery is done after 2-3 weeks once bilirubin level drops down adequately. If ERCP is not possible then percutaneous transbiliary drainage (PTBD) or *cholecystostomy* using Foley's or Malecot's catheter is done. Surgical resection (Whipple's) is done after 3 weeks. Present indications for ERCP stenting are severe albumin deficit and cholangitis.

♦ Antibiotics one day prior to surgery—cephalosporins/aminoglycosides.

♦ Often TPN may be required preoperatively which is also continued postoperatively.

♦ Pulmonary function study and respiratory physiotherapy to have adequate postoperative pulmonary function.

Note:

- In obstructive jaundice, serum bilirubin level will not go beyond 40 mg%. At this point raised intrabiliary pressure (Normal is 12-22 cm of H_2O) to 25 cm of H_2O, blocks the formation of bilirubin and secretion of cholesterol and bile acids completely. Maximum production of bilirubin per day possible is 250 mg%. Normal maximum excretion of bilirubin by kidney per day is 210 mg%. So maximum possible raise of serum bilirubin is 40 mg% in obstructive jaundice.
- Cholangitis in malignant obstruction is rare; but when it occurs it is rather severe form.
- Preoperative albumin therapy is usually *not done.*

Treatment

Only 10-15% of pancreatic carcinomas (head) are operable. 40-50% are locally advanced. Another 40-50% will have distant spread to liver or lungs.

Criteria for resection

- Tumour size less than 3 cm
- Periampullary tumours
- Growth not adherent to portal system

In operable cases:

- *Whipple's operation* is done by removing tumour with head and neck of pancreas, C loop of duodenum, 40% distal stomach, 10 cm proximal jejunum, lower end of the common bile duct, gallbladder, peripancreatic, pericholedochal, paraduodenal and perihepatic nodes. Continuity is maintained by choledochojejunostomy, pancreaticojejunostomy and gastrojejunostomy. Few advocate pancreaticogastrostomy into posterior wall of the stomach. Mortality in Whipple's operation is 2-8%. Original Whipple's operation (1935) was two staged procedure– initial bypass and a second stage resection with closure of pancreatic stump. In 1941, Trimble performed one stage pancreaticojejunostomy. *Complications* are delayed gastric emptying (19%); pancreatic fistula (14%); infection (intraabdominal abscess, wound infection, cholangitis, pancreatitis, pneumonia); bile leak. Mortality is 3%.
- *Traverso-Longmire pylorus preserving pancreaticoduodenectomy* (1978): Here duodenum is cut 2 cm distal to the pylorus and continuity is maintained by anastomosing with jejunum.
- *Fortner's regional pancreatectomy* (extended Whipple's) includes resection like Whipple's along with removal of segment of superior mesenteric vein and clearance of all regional nodes; and portal vein continuity is maintained by a synthetic vascular graft. Even though technically it gives adequate clearance, results are only equivocal.

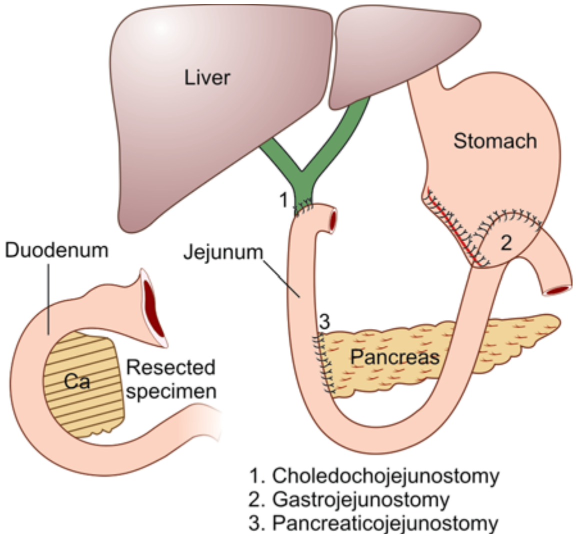

Fig. 14.44: Whipple's operation.

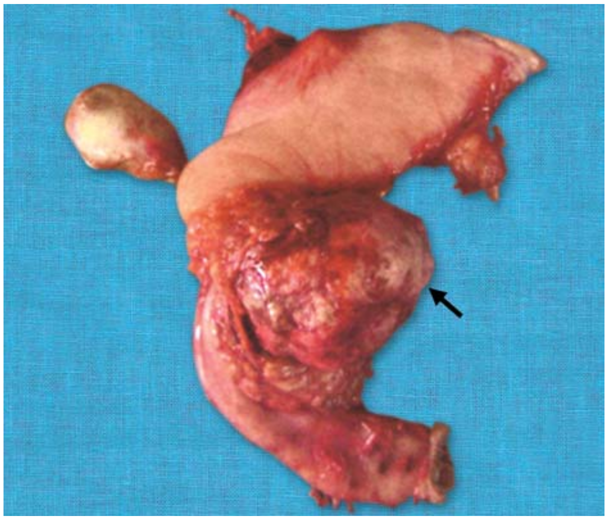

Fig. 14.45: Specimen of pancreaticoduodenectomy with tumour in the head of pancreas.

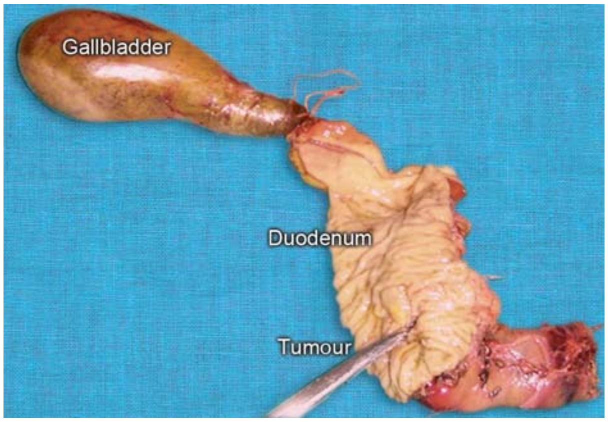

Fig. 14.46: Resected specimen (after Whipple's operation) of periampullary carcinoma of pancreas.

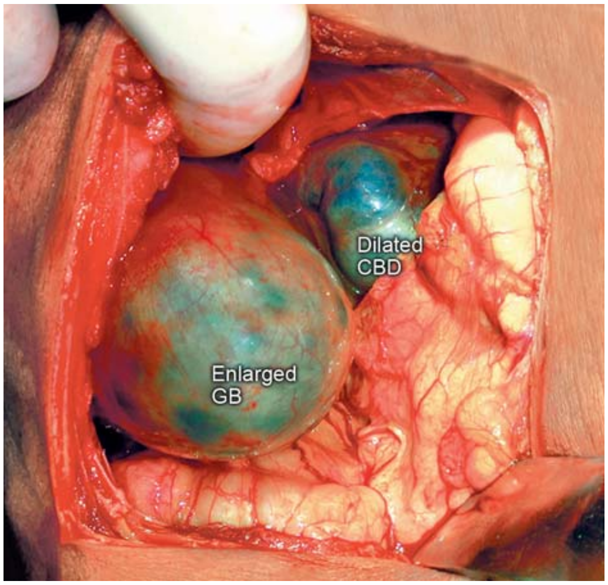

Fig. 14.47: On table finding in periampullary tumour showing dilated gallbladder and CBD.

A goal is a dream with a deadline.

- *Total pancreatectomy is presently said to be better*. It is done along with nodal clearance. Reasons are—possibility of multicentric nature of the disease, higher chance of recurrence after Whipple's operation, malignant cells may be present in pancreatic duct, morbidity by pancreatic fistula or pancreatitis after Whipple's operation is not seen here. Mortality in total pancreatectomy is higher 10-20%. Severe, resistant brittle diabetes mellitus may be seen after total pancreatectomy.

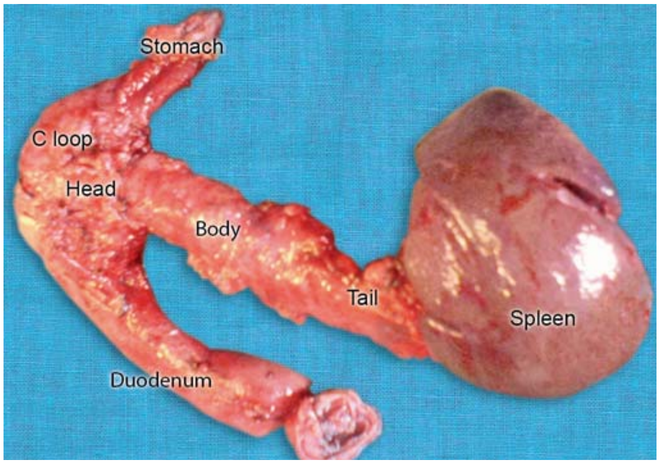
Stomach
C loop
Head
Body
Tail
Spleen
Duodenum

Fig. 14.48: Specimen of total pancreatectomy with C loop, head, body and tail of pancreas with spleen. Total pancreatectomy is done in refractory severe chronic pancreatitis and carcinoma pancreas.

- Distal pancreatectomy or central pancreatectomy or total pancreatectomy for cystadenocarcinoma of pancreas depending on the extent and size of the tumour.

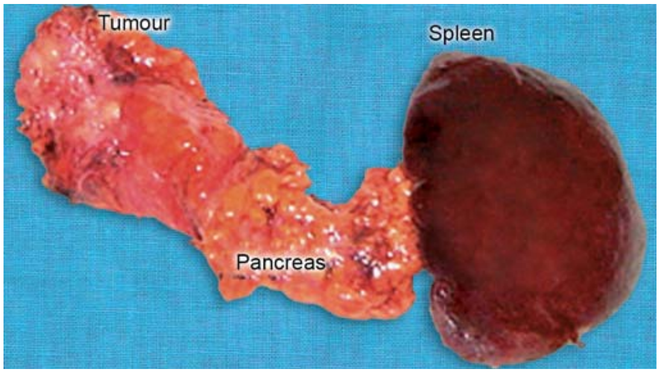
Tumour
Spleen
Pancreas

Fig. 14.49: Photo showing spimen of distal pancreatectomy.

In inoperable cases:

Here mainly to *palliate* obstructive jaundice, duodenal obstruction and pain.

- Roux-en-Y choledochojejunostomy is ideal palliative procedure along with gastrojejunostomy after doing cholecystectomy. Choledochojejunostomy is better than cholecystojejunostomy as CCJ may get blocked either by tumour infiltration or by bile sludge or by inflammation. 30% of periampullary carcinoma/carcinoma of head of pancreas eventually develop gastric (duodenal) outlet obstruction and so gastrojejunostomy is undertaken.
- ERCP and stenting is done to drain bile. Problem here are recurrent cholangitis, stent blockage and displacement, requirement of repeated stenting procedure. Duodenal stenting for duodenal obstruction is tried.
- Chemotherapy; neoadjuvant chemotherapy.
- Steatorrhoea is treated by enzymes. Control of diabetes is important.

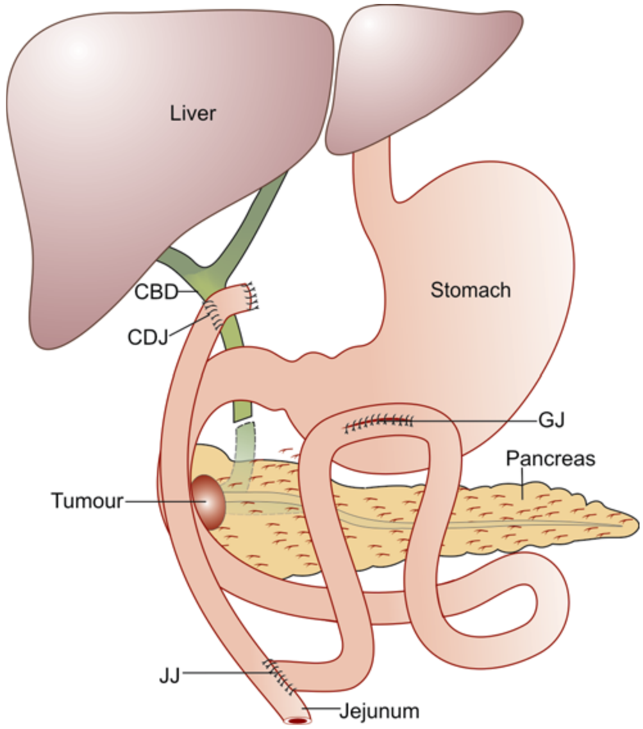
Liver
CBD
CDJ
Stomach
GJ
Pancreas
Tumour
JJ
Jejunum

Fig. 14.50: Choledochojejunostomy (CDJ) with gastrojejunostomy (GJ) and jejunojejunostomy (JJ). It is a Roux-en-Y anastomosis. CDJ should be 70 cm from the JJ.

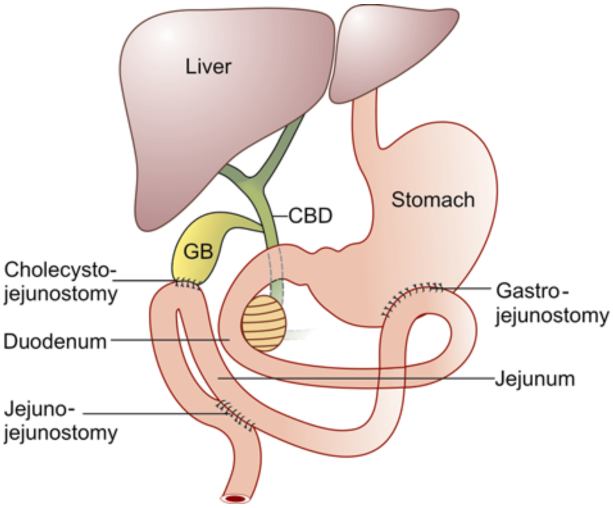
Liver
CBD
Stomach
GB
Cholecysto-
jejunostomy
Gastro-
jejunostomy
Duodenum
Jejunum
Jejuno-
jejunostomy

Fig. 14.51: Cholecystojejunostomy—is not a preferred method.

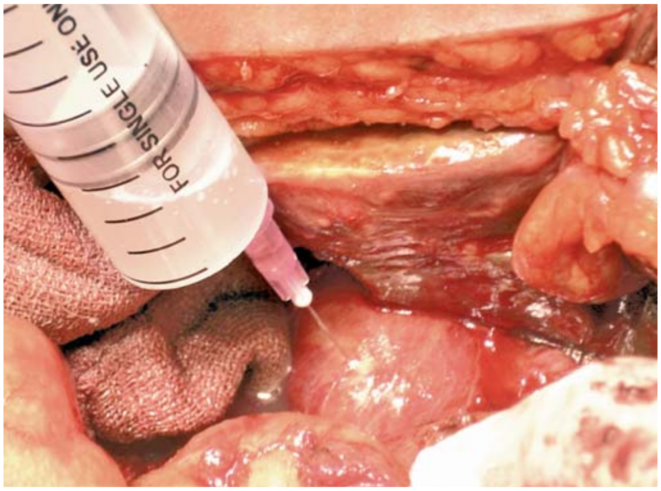

Fig. 14.52: White bile. It is neither white nor bile. It is opalescent; and is mucous. It suggests complete obstruction. It is usually observed in carcinoma pancreas. It is peroperative finding.

Adjuvant Therapy

- Adjuvant chemotherapy using gemcitabine—better but costly; dose is 1000 mg/m^2 surface area; 5-fluorouracil; mitomycin; vincristine, cisplatin, docetaxel, oxaliplatin are used along with gemcitabine.
- Radioactive iodine seeds I^{125} to the field are on trial. External radiotherapy 4000 cGy units to relieve pain and to reduce the tumour size.
- *Immunotherapy*—specific type to increase the effectiveness of chemotherapy and to improve the cure rate (allogenic tumour cell vaccine).

Neoadjuvant Chemoradiotherapy

Only 20% of pancreatic carcinomas are amenable for surgical resection. Even in these patients in spite of resection, overall outcome is poor. So neoadjuvant chemoradiotherapy is becoming popular [(50 Gy dose of *radiotherapy* for 5-6 weeks (25-30 fractions) with *infusion chemotherapy*, then *surgical resection*)] and under trial to improve the resectability and survival. It increases the R0 resection status and reduces the node positive rate. There is less chance of tumour hypoxia (which is more after surgery) by which increases the susceptibility of tumour to chemoradiotherapy; less chance of mucosal anastomotic toxicity. It reduces the tumour burden at surgery; decreases the local or regional recurrence; decreases the chances of pancreatic fistula. *But* there is no survival benefit.

Postoperative Management in Carcinoma Pancreas

- Maintenance of proper fluid and electrolyte balance.
- Observation for bleeding and its control by transfusion of blood, fresh frozen plasma (FFP), and prevention of DIC at initial period.
- Injection vitamin K 10 mg IM for 5 days.
- Respiratory care—ideally postoperative ICU care is better. Often ventilator is needed for 24 hours
- Maintaining adequate urine output—mannitol should be continued.
- Injection octreotide infusion for 5 days to suppress pancreatic secretion so as to prevent leak.
- Antibiotics, nasogastric aspiration.
- Continuous monitoring the patient with pulse/blood pressure/oxygen saturation/hourly urine output/inspection of drain site/abdomen distension/by doing HB%, LFT, serum creatinine, bilirubin, arterial blood gas analysis if needed, platelet count, prothrombin time.

Pain Control in Carcinoma Pancreas

- CT guided 50% of 20 ml ethanol injection into coeliac ganglion.
- Epidural anaesthesia.
- Opioids administration.
- Transthoracic splanchnicectomy—greater splanchnic nerve.
- Palliative radiotherapy—4000 cGy units.

Prognostic factors in carcinoma pancreas—poor prognosis

- Mean survival rate 6-9 months
- Growth more than 3 cm
- Nodal involvement
- Resection status—R0/R1/R2
- Portal vein infiltration
- Liver/lung secondaries
- Ascites/Trousseau's sign/left supraclavicular lymph nodal spread
- *White bile* on table carries poor prognosis
- Associated problems like pancreatitis, diabetes mellitus
- Liver dysfunction

ENDOCRINE PANCREATIC TUMOURS

- Most common is insulinoma.
  Others are gastrinoma, glucagonoma, vipoma, pancreatic polypeptidoma.
- They belong to group of tumours called "APUDOMAS".
- Endocrine tumours are associated with MEN syndrome (Type I Werner's syndrome).
- They are associated with:
  - Parathyroid adenoma.
  - Pituitary adenoma.
  - Peptic ulcer.
- They are usually multiple, small, 35-50% are malignant and presents with features of endocrine lesion. Commonly occurs in body and tail (except gastrinomas).
- Requires special method to identify.
- MRI, angiogram, hormone assay are the diagnostic methods required.
- Normal islet cells are < 2% of pancreatic mass. They contain α, β, δ, G, D2 and PP cells (pancreatic polypeptide).
- Incidence is 5 per 100000 per year.

Those who wish to sing always finds a song.

- 20% are nonfunctional.
- Insulinomas are commonly benign; glucagonomas are almost always malignant.

INSULINOMAS

- Insulinomas are commonest endocrine pancreatic tumour (60%) arising from B cells (β cells) of pancreas.
- They are commonly solitary but 10% can be multicentric and secrete insulin in excess causing profound hypoglycaemic features.
- Equal in both sexes.
- 90% of insulinomas are < 2 cm in size.
- 15% are malignant. 85% are benign.

Clinical Features

- Abdominal discomfort, trembling, sweating, hunger, dizziness, diplopia, hallucinations. Later epilepsy and unconsciousness. They are usually over weight.
- Weight gain is common due to overeating.
- Permanent neurological deficits can occur with behavioural changes.

Whipple's triad:

- An attack of hypoglycaemia in fasting state
- Blood sugar below 45 mg% during the attack
- Symptoms relieved by glucose

Diagnosis

- Insulin radioimmunoassay. An insulin level > 7 µU/ml; insulin/glucose ratio > 0.3 signifies insulinoma. Proinsulin level more than 24% of total insulin signifies insulinoma. Proinsulin level more than 40% of total insulin signifies malignant insulinoma. C peptide level > 1.2 µg/ml with glucose level < 40 mg% suggests insulinoma.
- MRI to localize the tumour.
- Angiogram (coeliac).
- Endosonography.
- On table U/S.
- Blood sugar estimation.
- Insulin provocation test using calcium gluconate or tolbutamide.

Radioimmunoassay in insulinoma

- Increase in insulin levels of plasma
- Increased ratio of proinsulin to insulin > 20%
- Increased circulating C-peptide
- Plasma insulin to glucose ratio > 0.3 is diagnostic

Differential Diagnosis

- Nesidioblastosis.
- Noninsulinoma pancreatogenous hypoglycaemia.

Treatment

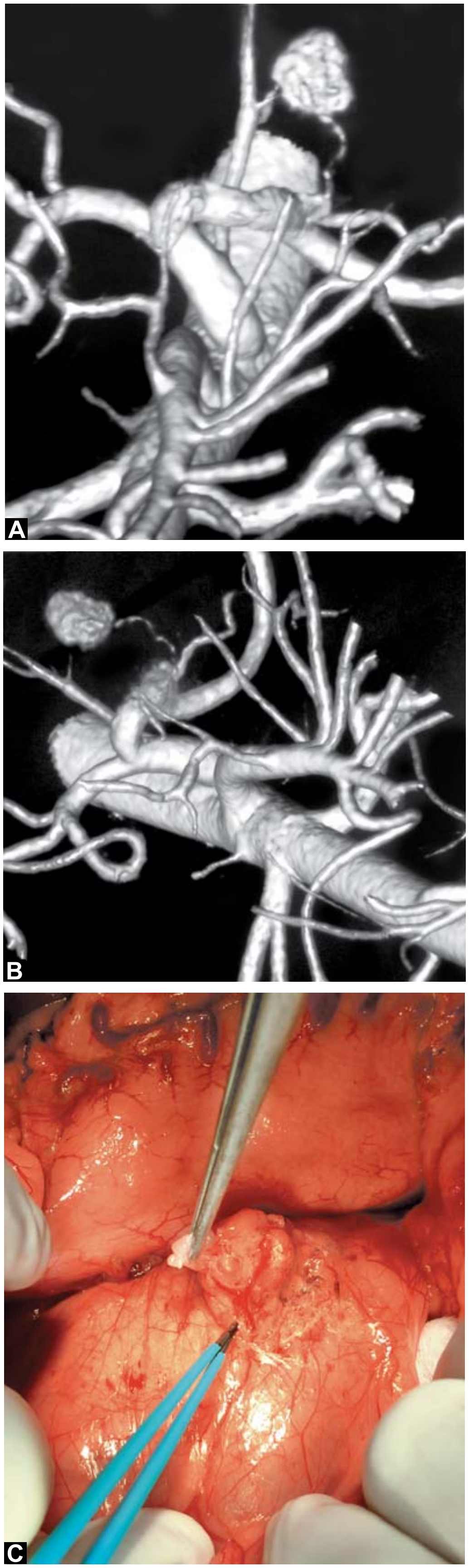

Figs 14.53A to C: Insulinoma imaging picture and on table look (*Courtesy:* Dr Arunkumar, MCh).

Treatment of insulinoma

- Enucleation commonly, often distal pancreatectomy
- Diazoxide, beta blockers, phenytoin, verapamil, steroids, growth hormone to control hypoglycaemia
- Octreotide decreases the insulin secretion
- Calcium channel blockers
- Streptozotocin whenever secondaries are present in liver or elsewhere

- Enucleation of all tumours.
- Often distal pancreatectomy is required (spleen; tail and body of the pancreas are removed) 90% are curable.

GASTRINOMAS

- They arise from non-beta cells (G cells) of the pancreas, which secretes high levels of gastrin.
- It is 2nd most common endocrine pancreatic tumour.
- It is the commonest endocrine pancreatic tumour seen in MEN I syndrome. In MEN II syndrome lesion is often seen in duodenum (30%).
- It is common in males.
- Common in gastrinoma triangle (*Passaros triangle*).

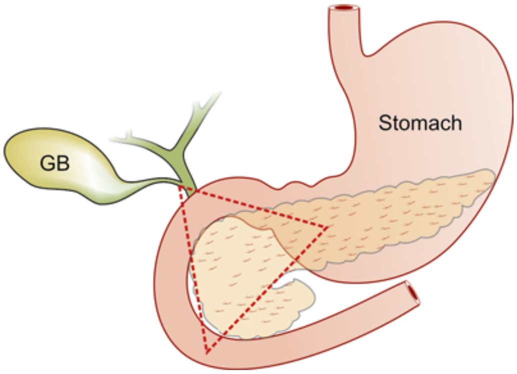

Fig. 14.54: Passaros triangle.

Three points forming Passaros triangle:

1. Junction between the head and neck of the pancreas
2. Junction of cystic duct with CBD
3. Junction between the 2nd and 3rd parts of the duodenum

- Half of them are multiple and malignant (50%).
- It causes severe, multiple peptic ulcers in the stomach, duodenum and jejunum.
- It is due to hypergastrinaemia causing high levels of acid secretion.
- Other features are diarrhoea, steatorrhoea, hypokalaemia.
- It causes ZES type II.
- When malignant commonly present with secondaries in liver (80%)/lymph nodes/lungs or bones.

Salient features of ulcers in gastrinomas

- Multiple ulcer
- Resistant ulcer/refractory ulcer
- Jejunal ulcer
- Recurrent ulcer
- Ulcer which is more prone for bleeding and perforation

Investigations

- Gastrin assay (normal level < 200 pg/ml).
- Gastroscopy.
- U/S, Endosonography.
- MRI (more sensitive) or CT scan to detect the site of the tumour.
- Angiogram.
- Increased gastrin level by calcium (injection IV) provocation test.
- Pentagastrin test (both shows very high basal and peak levels).
- Secretin provocation test.
- Somatostatin receptor scintigraphy (SRS) with radiolabeled octreotide is used to locate the site of tumour.

Note:

Normal level of gastrin is 100-150 pg/ml. It is more than 200 pg/ml in gastrinoma patients. It usually reaches upto or greater than 1000 pg/ml in gastrinoma. Other causes of hypergastrinaemia like G cell hyperplasia, incomplete antrectomy, low gastric acid level, achlorohydria, pernicious anaemia, acid controlling drugs like H2 receptor block or proton pump inhibitors and atrophic gastritis should be ruled out. Basal acid output in gastrinoma is > 15 mEq/hour which is specific of gastrinoma (ZES II). Gastric pH more than 3 rules out gastrinoma. *Secretin stimulation test* by injecting 2 units/ kg secretin IV and assessing blood samples before and every 5 minutes after injection of secretin for 30 minutes will show rise in gastrin level more than 50% of baseline.

Treatment

60% are curable.

- Enucleation of tumours.
- Distal pancreatectomy.
- Pancreaticoduodenectomy.
- Subtotal pancreatectomy.
- Often total gastrectomy may be required.

Remember

- Gastrinomas are often associated with MEN syndrome and are multiple
- Sporadic gastrinomas are often solitary and then are malignant
- Gastrinoma often secretes ACTH
- Most common malignant endocrine pancreatic tumour is gastrinoma
- Gastrinoma causes ulcer of unusual nature (refractory/resistant); unusual recurrence (repeated and multiple); unusual number (multiple); unusual sites (2nd/3rd part duodenum/jejunum); unusual age (young and aged)

It is better to aim at good things and miss it, than to aim at a bad thing and hit it.

Triad of somatostatinoma

- Diabetes mellitus
- Steatorrhoea
- Cholecystolithiasis

GLUCAGONOMAS

- Arise from α cells of pancreas.
- Common in body and tail of pancreas.
- It is common in females.
- It attains large size of 5-10 cm.
- It is commonly sporadic. 17% associated with MEN type II syndrome.
- It is commonly malignant (80%). 80% spreads to liver.

Clinical Features

- Necrolytic migratory erythema (65%).
- Diabetes (90%).
- Diarrhoea and weight loss.
- Stomatitis.
- Anaemia and features of amino acid deficiency.
- Vulvovaginitis is common.

Investigations

- MRI, CT scan.
- Endosonography.
- Angiogram.
- Increased serum glucagon level. Fasting glucagon level more than 50 p mol/litre is diagnostic.

Treatment

- Correction of anaemia, protein and amino acid deficiency by enteral/parenteral nutrition.
- Intravenous amino acid infusion to control necrolytic dermatitis.
- Prevention of DVT.
- Distal pancreatectomy.
- Occasionally Whipple's/total pancreatectomy is needed.
- Enucleation is rarely sufficient for glucagonoma.
- Dacarbazine is specifically effective in glucagonoma.

Vipoma

- Arising from D2 cells of pancreas
- Watery diarrhoea, hypokalaemia, achlorohydria (WDHA syndrome)
- Also called as *pancreatic cholera or Verner-Morrison syndrome, weak—tea syndrome*
- Usually malignant
- Secretes vasointestinal polypeptide > 150 pg/ml
- Common in body and tail—solitary
- Distal pancreatectomy is the treatment of choice
- Prednisolone controls the diarrhoea
- Octreotide is useful

ZOLLINGER-ELLISON SYNDROME

Type I: "G" cell hyperplasia with hypergastrinaemia and chronic peptic ulceration.

Treatment is partial gastrectomy with removal of "G" cell area.

TYPE II: Gastrinomas (see above).

Note:

- Somatostatinoma arises from δ cells of pancreas. It is exceedingly rare. Usually larger than 2 cm and single. It is common in head of pancreas and duodenum. Its association with MEN II syndrome is rare. It is often associated with neurofibromatosis type I and phaeochromocytoma. Presentations are steatorrhoea; diabetes mellitus; gallstones; hypochlorohydria and occasional jaundice. Tumour is commonly malignant and commonly present with nodal and liver spread. Elevated fasting somatostatin level is diagnostic. Treatment is pancreaticoduodenectomy.
- 20% of pancreatic endocrine tumours are nonfunctional. Nonfunctional means these tumours do not produce hormones or produced hormones will not produce any symptoms.
- Pancreatic polypeptide tumour (PP tumour) is classified under the non functioning endocrine tumour as it does not cause any symptoms even though there is high pancreatic polypeptide. It is commonly malignant, often with spread. PPoma occurs predominantly in head of pancreas. Pancreaticoduodenectomy is the treatment.

CYSTIC FIBROSIS

It is inherited as an autosomal recessive disease. It causes—

1. *Severe exocrine dysfunction*—due to blockage of pancreatic ducts as a result of precipitation of pancreatic enzymes leading to duct ectasia.
2. *Chronic pulmonary disease* due to blockage of bronchi and bronchioles.
3. *Elevated sodium and chloride*—in the sweat more than 90 mmol/L.
4. It is common in Caucasians. It is inherited as mutation in cystic fibrosis transmembrane conductance regulator gene in chromosome 7.

Presentations

- Chronic pulmonary disease with emphysema, bronchiectasis.
- Steatorrhoea.
- Meconium ileus with intestinal obstruction.
- Salty sweat (while kissing the baby).
- Poor growth due to malabsorption.
- Cirrhosis of liver due to bile plugging.
- Pancreatitis is common.
- Sialadenitis, choroiditis can occur.
- Absence of vas deferens in men.

Diagnosis

- DNA study.
- Vomitus when put on an exposed X-ray film will not digest the gelatin in half an hour (wherein in normal individual vomitus contains trypsin which will digest the gelatin of X-ray film and turns white).

- U/S abdomen.
- Sweat test for sodium and chloride.
- Pulmonary function tests.

Treatment

- Low fat and salt rich diet.
- Correction of malabsorption.
- Pancreatic enzyme supplements.
- Antibiotics.
- Respiratory physiotherapy.
- If meconium ileus is present, it might require surgery, i.e. Bishop-Koop operation.

Prognosis is poor.

Problems in cystic fibrosis

- Meconium ileus
- Recurrent respiratory infection and complications
- Sodium deficit
- Severe malnutrition
- Sialadenitis, choroiditis
- Cirrhosis of liver
- Usually die in childhood or early age group

ANNULAR PANCREAS

It is due to failure of complete rotation of ventral bud of pancreas, so that ring of pancreatic tissue encircles the 2nd part of the duodenum causing obstruction. It contains a duct that joins the main pancreatic duct.

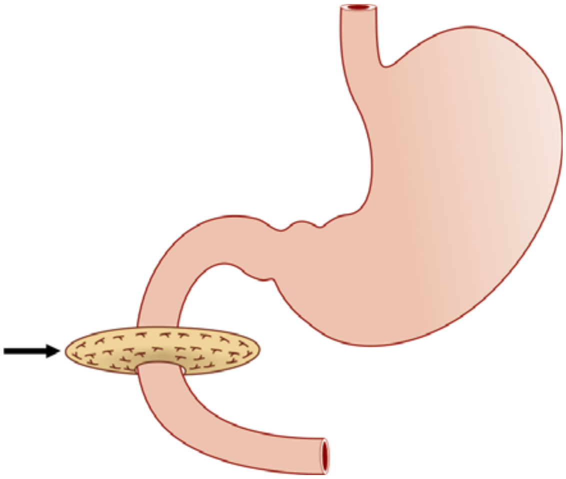

Fig. 14.55: Annular pancreas.

Clinical Features

- Distension of upper abdomen.
- Bilious vomiting.
- Visible gastric peristalsis.
- It is often associated with the Down's syndrome, and other congenital gut abnormalities.

Types

- *Neonatal type*—produces symptoms of intestinal obstruction.
- *Adult type*—presents after the age of 20. Presents with features of duodenal ulcer and bilious vomiting.

Differential Diagnosis

- Duodenal atresia.
- Wilkies syndrome.

Investigations

- Plain X-ray abdomen shows double-bubble appearance.
- ERCP and Radioisotope study.
- Barium meal shows obstruction at 2nd part of duodenum.

Treatment

- *Duodenoduodenostomy—Ideal.*
- Duodenojejunostomy.
- **Do not resect the ring.** Attempt at resection will lead to pancreatic fistula.
- G-J is not done.

ECTOPIC (ACCESSORY) PANCREATIC TISSUE

Sites

- Stomach wall.
- Small intestine.
- Meckel's diverticulum.
- Greater omentum.
- Splenic hilum.
    - In the bowel it may be in the submucous or intramuscular plane. It may precipitate intussusception.
    - About 30-33% contain islet of Langerhan cells. Often endocrine pancreatic tumour can arise from ectopic pancreas.
    - It may cause upper GI bleed.

PANCREATIC DIVISUM

- During development ventral pancreatic bud with distal bile duct is towards right and dorsal pancreatic bud is towards left side. Ventral pancreatic bud forms uncinate process and inferior part of the head of the pancreas. Dorsal bud forms body and tail of the pancreas. Ventral bud rotates clockwise towards left from behind the duodenum to join dorsal bud to form pancreas. Ventral pancreatic bud duct forms main pancreatic *duct of Wirsung*; dorsal bud duct forms accessory pancreatic *duct of Santorini*.
- *Pancreatic divisum* is failure of fusion of ventral and dorsal pancreatic ducts. So that dorsal pancreas drains through duct of Santorini (proximal minor papilla); ventral one drains through duct of Wirsung into *major papilla of Vater*.

- Its incidence is 10%.
- Here major pancreatic secretion drains through minor papilla causing partial obstruction precipitating pancreatitis.

PANCREATIC CALCULUS

Causes

- Chronic pancreatitis commonly due to dietary factors, familial, hyperparathyroidism.
- Commonly they are multiple and associated with or due to chronic pancreatitis.

Pathogenesis

Stricture of pancreatic duct and stasis of pancreatic secretion.

Types

a. **Calculi in the duct**:
  - Here parenchyma is still functioning and less severely affected.
  - Duct is dilated significantly (more than 3 mm) with multiple large stones in the duct along with multiple strictures.
  - Duct is having dilatation alternating with strictures—*chain of lake appearance.*
    - Plain X-ray abdomen shows stones placed horizontally at L_1-L_2 level.
    - U/S, ERCP, MRCP are other useful investigations.
    - Blood sugar should be checked.

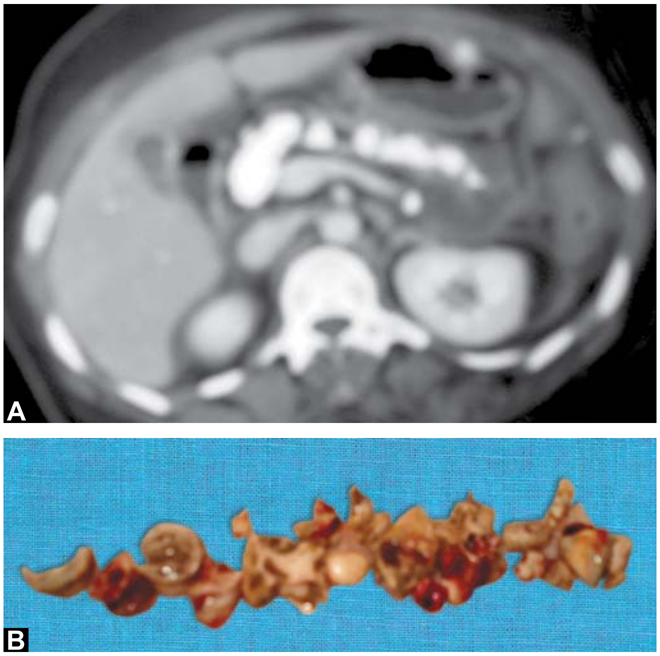

Figs 14.56A and B: CT scan picture showing pancreatic ductal stones. Stones that are removed from the same patient are also shown. Patient underwent longitudinal pancreaticojejunostomy.

Treatment

The duct is opened longitudinally and after removing all stones, side-to-side pancreaticojejunostomy is done.

b. **Calculi in the Parenchyma**
  - Here multiple calcifications occur in the pancreatic parenchyma with pre-existing chronic pancreatitis. Gland is fibrosed and severely affected with less functioning capacity.
  - Plain X-ray, U/S, ERCP, MRCP are the diagnostic methods.

Treatment

- *Total pancreatectomy.*
- It is more likely to turn into carcinoma pancreas.

PANCREATIC ASCITES

- It can occur in chronic pancreatitis or as a complication of acute pancreatitis or in ruptured pseudocyst of pancreas or after pancreatic trauma (10%, in children trauma is the cause).
- It may be associated with pleural effusion which can be both sided but commonly on left side. Often pleural effusion and ascites may be massive. It is due *to internal pancreatic fistula.*
- Incidence is 1%. If not treated mortality is 30%.
- It is due to leakage from disrupted pancreatic duct or ruptured pseudocyst. Anterior disruption causes pancreatic ascites. Posterior disruption causes fluid to travel along oesophageal and aortic hiatus resulting in pleural effusion.
- Pancreatic ascites and fistula are common in pancreatitis due to alcohol intake. It is less common in pancreatitis due to biliary tract disease.
- U/S, very high amylase and protein levels in ascites fluid, ERCP to identify leak, CT scan are the essential investigations.
- Repeated ascitic tap; pleural tap; TPN; somatostatin or octreotide to reduce pancreatic secretion are tried as conservative therapy for 3 weeks. 50% respond to this therapy.
- If there is no improvement beyond 3 weeks, ERCP stenting of pancreatic duct is done initially. Then resection (better option) or drainage (jejunal loop Roux-en-Y anastomosis directly over disrupted duct) surgery is done. Pancreatic pleural effusion disappears once abdominal cause or ascites is corrected.

PANCREATIC FISTULAE

Types

External and internal fistulae.

External Fistulae

Types

a. Low output fistulae < 200 ml.
b. High output fistulae >200 ml.
(*Note:* For other GI fistulae it is 500 ml)

Causes

- Whipple's operation.
- Pancreaticojejunostomy.

- Splenectomy.
- Colonic surgeries.
- Trauma.
- After external drainage of the infected pseudocyst of pancreas.
- After percutaneous drainage of the pseudocyst.

Clinical features

- Patient is in a catabolic state.
- Skin excoriation.
- Electrolyte imbalance.
- Malnutrition.
- Sepsis.

Investigations

- CT scan.
- ERCP.
- Fistulogram.
- Amylase estimation of the discharge.

Treatment

Conservative treatment:

- Total parenteral nutrition.
- Zinc oxide cream for skin excoriation.
- Correction of electrolyte imbalance.
- Antibiotics.
- Octreotide.

Surgery:

- Roux-en-Y anastomosis.
- Resection of fistula with pancreas.
- Endoscopic stenting of the pancreatic duct.

Internal Fistulae

- Communicating pseudocyst.
- Pancreatic ascites.
- Pancreatic pleural effusion.
- Pancreatic-enteric fistulae.

Complications of pancreatic fistulae

- Septicaemia
- Malnutrition
- Bleeding due to erosion of vessels
- Electrolyte imbalance
- Severe skin excoriation

PANCREATIC ABSCESS

- 5% common.
- It can occur immediately after an attack of acute pancreatitis or in an existing pseudocyst.
- Common in alcoholics and gallstone diseases.
- Mortality is 25%.
- *Pancreatic abscess* is collection of pus in lesser sac (intraabdominal) in relation to pancreatic surface which contains mainly pus with only less or no necrotic pancreas. It may slough off the pancreatic/ splenic vessel wall to cause torrential haemorrhage. Abscess may be single or multiple (60%). It is commonly in head/body or tail. But often entire gland may be involved (25%). Abscess may rupture into viscera or extend into other part of the abdomen.

Clinical Features

- Persistent pain abdomen.
- Tender epigastric mass.
  Commonest organism involved is *E. coli.*

Investigation

- Ultrasound and CT scan are diagnostic.
- CT guided FNA and Gram's staining and culture.

Treatment

- Antibiotics.
- Open surgical drainage with cystogastrostomy.
- U/S guided aspiration.
- *CT guided insertion of polythene catheter.*

Complications

- Septicaemia.
- Haemorrhage due to erosion of splenic vessels.
- Recurrent abscess formation.

PANCREATIC NECROSIS

It is diffuse or focal area of nonviable parenchyma of pancreas, occurs during an attack of severe pancreatitis that is associated with peripancreatic fat necrosis. It occurs in 20% of acute pancreatitis.

Types

- Sterile necrosis is 60% of total pancreatic necrosis. Mortality here is 10%.
- Infected necrosis is 40% of total necrosis. Mortality here is 30-40%, very high. Risk of infection is 25% in one week of acute attack; 35% at the end of 2nd week; and 70% at the end of 3rd week.
- Necrosis extends to retroperitoneal fat, mesentery, retrocolic and perinephric areas.

Pathogenesis

- Premature activation of proteolytic enzymes.
- Intrapancreatic vessel thrombosis.

- Pancreatic microcirculation failure.
- Release of inflammatory mediators.

Mode of Infection

- Haematogenous.
- Reflux through ampulla.
- Mucosal translocation of bacteria.
- From the biliary tree.
- Lymphatic route.
- Transperitoneal spread.

Bacteria

- 65% are polymicrobial.
- *E. coli* is commonest.
- Others are—*Proteus*, *Pseudomonas*, *Klebsiella*, staphylococci, *Streptococcus faecalis*, enterococci, anaerobes, *Clostridium welchii*.

Diagnosis

- Clinical features.
- CT scan abdomen and CT guided FNA for Gram's staining and culture.
- CRP value—> 120 mg/litre; interleukin level; elastase level; urinary TAP assessment.

Treatment

- Antibiotics—imipenem, cefuroxime, ofloxacin, meropenem.
- Percutaneous drainage.
- Laparotomy, necrosectomy, debridement, saline wash, closure with tube irrigations/lavage.

Complications

- Septicaemia, ARDS, renal failure, MODS.
- Pancreatic fistula after surgery.
- Intra-abdominal abscess.
- Abdominal dehiscence, incisional hernia after surgery.

PANCREATIC TRAUMA

- It is rare because of its anatomical location - retroperitoneum.
- Its injury is usually associated with injuries to liver/ duodenum/spleen/portal system/biliary system/ kidney.
- Deep force in epigastrium may cause crushing of body of pancreas against vertebra—closed injury.
- Penetrating injury may cause direct sharp injury of pancreas.

Types

- Parenchymal contusion/laceration without duct disruption.
- Parenchymal injury with duct disruption.
- Complete transection of pancreas.
- Massive destruction of pancreatic head.

Classification

Pancreatic injury

Class I Capsular damage; minor parenchymal injury

Class II Transection of duct in body or tail—partial or complete

Class III Major duct injury in pancreatic head or intrapancreatic CBD injury

Duodenal injury

Class I Contusion, hematoma, partial thickness injury

Class II Full thickness duodenal injury

Class III Full thickness duodenal injury with

1. More than 75% circumference injury
2. Extrapancreatic CBD injury

Combined pancreatico duodenal injury

Type I Class I injury to both organs/class I injury of one organ with class II injury to other

Type II Class II injuries of both organs

Type III Class III injury to one organ with less severe injury to other

Type IV Class III injury to both organs

Grading of pancreatic injury

Grade I – Minor contusion/laceration without duct injury

Grade II – Major contusion/laceration without duct injury

Grade III – Distal transection or injury with duct involvement

Grade IV – Proximal transection or injury with involving ampulla

Grade V – Massive disruption of pancreatic head

Features

- Pain in epigastrium.
- Features of associated injuries.
- Features of shock.
- Rise in serum amylase level is common.

Investigations

- CT scan is diagnostic.
- ERCP to confirm duct disruption.
- Assessment of blood loss and other injuries.

Treatment

- Commonly conservative with fluid management; blood transfusions; antibiotics; pain relief.
- Surgery is needed when there is:
  - Major ductal disruption.
  - Vascular injury.
  - Extensive injury to head.
  - Other organ injury.

Surgeries are

- Proper exploration and haemostasis.
- In neck transection, distal pancreatectomy with splenectomy.

- In injury to head, haemostasis and external drainage is done.
- Pancreaticoduodenectomy is done if there is major injury.
- Choledochojejunostomy is done if there is CBD injury.
- Different pancreatic drainage procedures are done to prevent sepsis and fistula.
- Portal system injury carries poor prognosis. It needs repair or intervening graft placement to control and maintain the continuity.

Prognosis

- Depends on type and class of injury and other associated injuries.

Complications

- Haemorrhage.
- Septicaemia.
- Pancreatitis—acute/recurrent.
- Pseudocyst formation.
- Pancreatic fistula.
- Pancreatic abscess.
- Pancreatic duct stricture at the site of duct injury as a delayed complication.

CYSTIC LESIONS OF PANCREAS

Types

1. True pancreatic cysts:
    - Congenital.
    - Not premalignant.
    - Asymptomatic.
    - May be associated with cystic diseases of liver, kidney.
    - Multiple true pancreatic cysts (polycystic disease of pancreas) are common than solitary true cyst. They are lined with cuboidal epithelium and are associated with *von Hippel Lindau disease* in 50% of patients. They don't require any treatment.
2. Pseudocyst of pancreas.
3. Cystic tumours of pancreas.
    a. Serous cystadenoma.
    b. Mucinous cystadenoma.
    c. Cystadenocarcinoma.
    d. Angioma.
    e. Cystic teratomas.
4. Hydatid cyst of the pancreas.

EXOCRINE PANCREATIC DISEASE

Causes

- Pancreatitis.
- Cystic fibrosis.
- Pancreatic resections.

Investigations

- *72 hours faecal collection* for estimation of daily faecal fat: In a person ingesting 100 gm of fat/day, if faecal excretion of fat is more than 7 gm/day, it signifies pancreatic insufficiency. It is the most sensitive and specific test (Presence of neutral fat suggests pancreatic disease whereas split fat suggests small bowel disease).
- *Secretin-cholecystokinin stimulation test* shows duodenal fluid normally contain more than 80 mEq/L HCO_3^- and HCO_3^- greater than 15 mEq/30 minutes.
- *PABA excretion in the urine* after ingestion of bentinomide, if less than 50% it signifies pancreatic insufficiency.
- Serum trypsin level is low in pancreatic insufficiency.
- Pancreolauryl test.
- Triolin breath test.
- *Lundh test*—Duodenal juice collected 2 hours after meal shows trypsin level > 8 mEq/ml. It is very low in chronic pancreatitis (Goran Lundh).
- Glucose tolerance test.

Treatment of Pancreatic Insufficiency

- Low fat diet.
- Pancreatin enteric coated tablets during meals.
- Cause is treated.

Chapter 15 Retroperitoneal Space

CHAPTER OUTLINE

- Anatomy of Retroperitoneum
- Retroperitoneal Fibrosis
- Retroperitoneal Swellings
- Retroperitoneal Tumours
- Psoas Abscess

ANATOMY OF RETROPERITONEUM

This is the space between the peritoneal cavity and posterior abdominal wall. It is bounded anteriorly by posterior parietal peritoneum; posteriorly by vertebral column, psoas muscles, quadratus lumborum muscle and tendinous portion of transversus abdominis muscles; superiorly by diaphragm; inferiorly by pelvic levator muscles (by Ackerman). It is real potential space.

Diseases of retroperitoneal spaces

- Retroperitoneal fibrosis
- Retroperitoneal cysts
- Retroperitoneal tumours
- Retroperitoneal lymphomas
- Retroperitoneal vascular diseases, e.g. aneurysms
- Retroperitoneal trauma and haematoma
- Retroperitoneal infection, e.g. psoas abscess, pyogenic abscess

Diseases of the specific retroperitoneal organs, e.g. adrenals, kidney.

RETROPERITONEAL FIBROSIS

- *Idiopathic type (70%) is called as* ***Ormond's disease*** which is associated with mediastinal fibrosis, Dupuytren's contracture, plantar fasciitis, Peyronie's disease, Riedel's thyroiditis, sclerosing cholangitis, mesenteric panniculitis, pseudotumour of the orbit and other fibromatosis. It is nonspecific inflammation of fibrofatty tissue in the retroperitoneum.
- Patient presents with severe and persistent *back pain.*
- Diffuse fibrosis in retroperitoneum can compress both ureters leading to bilateral hydronephrosis and renal failure.
- Extravasation of urine, ascending lymphangitis, and autoimmune cause are the aetiologies.
- In 30% cases it is bilateral. Hypertension, lower limb oedema (lymphatic or venous block), feeble lower limb arterial pulses are common.
- It is like woody fibrous tissue which narrows ureters, vessels and nerves. Ureteric obstruction commonly in lower 1/3rd causing dysuria, frequency, oliguria, renal failure. Venous and lymphatic oedema of limbs can occur.
- It may be localised or generalised. Generalised type (15%) may extend into duodenum, CBD and pancreas. It is not encapsulated but margin is well-demarcated.
- It can also cause vascular compression, both venous and arterial.
- Disease is progressive one.

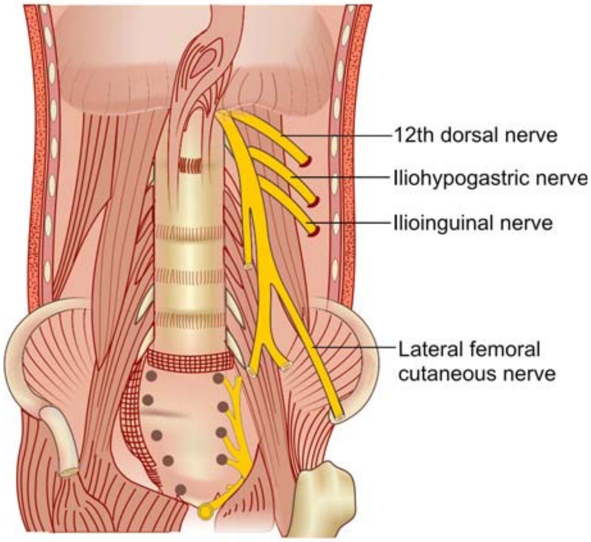

Fig. 15.1: Anatomy of retroperitoneal space.

Investigation

- U/S
- IVU—diagnostic:

Triad

- Deviation of middle 1/3rd of ureter medially
- Hydroureteronephrosis
- Extrinsic ureteral compression

- Blood urea and serum creatinine, raised ESR and CRP.
- CT scan.

Treatment

- *Drug therapy*: Methylprednisolone, azathioprine, penicillamine, tamoxifen.
- Cystoscopic stenting of the ureters to prevent renal failure.
- Symptomatic treatment: If stenting fails, bilateral nephrostomy is done.
- Ureterolysis.
- Lateral repositioning of ureters.
- Omental wrap after ureterolysis.
- Vascular bypass graft for vascular encasement.

Other causes for retroperitoneal fibrosis (30%)

- Methysergide and beta blocker drugs
- Urine leak and collection in the space
- Haematoma, blood collection
- Advanced malignancy, radiation.
- Ergot's alkaloids, dopamine agonists
- Retroperitoneal surgeries, haemorrhage
- Trauma, infection

RETROPERITONEAL SWELLINGS

Types of retroperitoneal (RP) tumours (0.3 to 3%)

Benign—20% of RP tumours

- Retroperitoneal lipoma
- Retroperitoneal neurofibroma, neurilemmoma
- Retroperitoneal leiomyoma
- Extra-adrenal chromaffinomas
- Paraganglioma
- Mucinous cystadenoma
- Haemangiopericytoma

Malignant—80% of RP tumours

- Retroperitoneal liposarcoma, leiomyosarcoma—50%
- Retroperitoneal lymphoma (commonly NHL)
- Malignant tumours from specific organs
- Germ cell tumours, chordomas
- RPLN secondaries with hard nodules

Note:

- Tumours of retroperitoneal organs like kidneys, ureters, pancreas, and adrenals are conventionally *not included* in retroperitoneal tumours.
- *Paragangliomas* are of neural crest origin that arises from paraganglionic tissues distributed along the major vasculatures (abdominal aorta in abdomen) or sympathetic chain, may be functioning or nonfunctioning. 20% secrete catecholamines. They may be multiple. Only 0.5% of cases present with hypertension. It can be familial, often associated with von Hippel Lindau disease and RET proto-oncogene. It may arise from *type I cells* as chromogranin A positive—NSE positive; or *type II cells* as S 100 positive with good prognosis.
- *Costello syndrome* is mental retardation, benign papilloma; develop embryonal rhabdomyosarcoma.
- *Retroperitoneal cysts* are—cysts arising from Wolffian remnants or urogenital tract; RP mesenteric cyst; teratomatous dermoid cysts; RP lymphatic cyst; parasitic cysts.

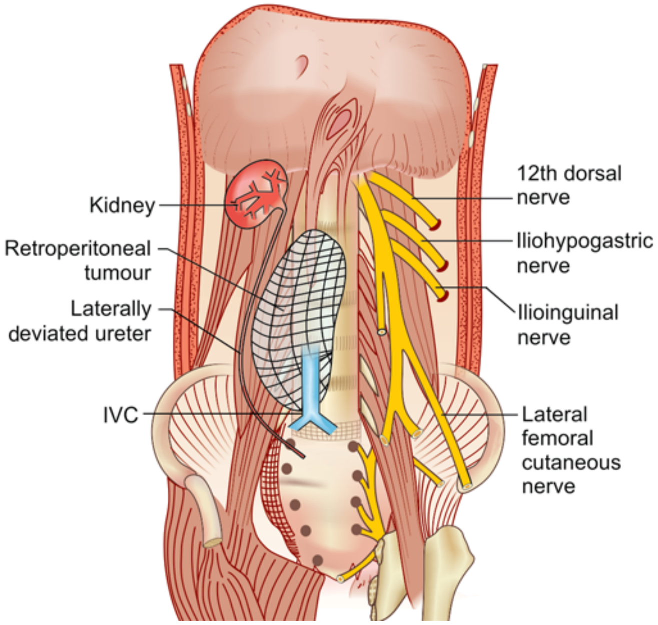

Fig. 15.2: Retroperitoneal tumour can attain very large size often. It causes deviation of ureter laterally often may invade the ureter (In retroperitoneal fibrosis ureter is deviated medially). It may invade or encase the IVC also causing IVC obstruction with features of bilateral oedema feet, dilated lateral abdominal wall vein with upward direction flow of blood.

Features of the Retroperitoneal Mass

- It is usually large.
- Not moving with respiration.
- Nonmobile.
- Does not fall forward (confirmed by knee-elbow position).
- Deeply placed.
- Resonant on percussion (because of the bowel in front).

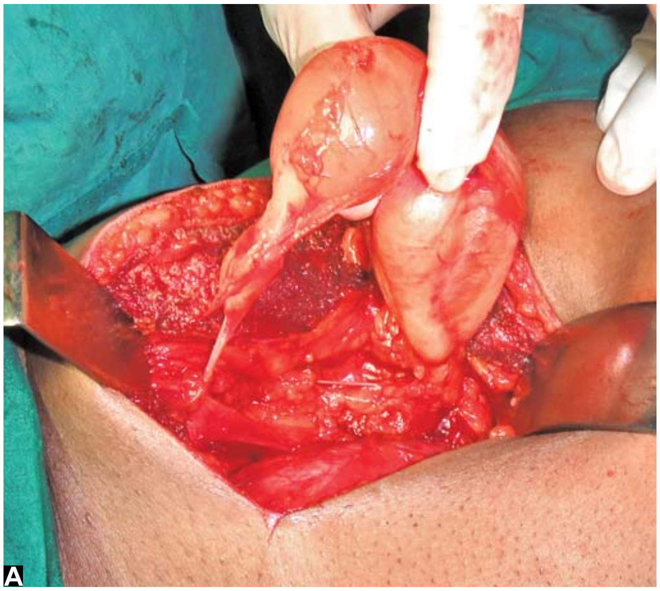

Fig. 15.3A

The lesser the indication, the greater the complication.

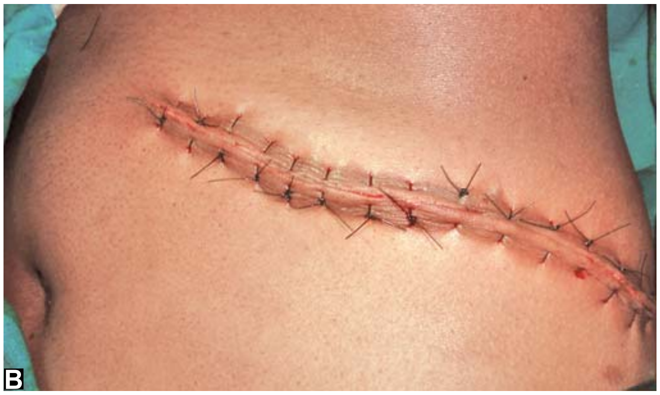

Fig. 15.3B

Figs 15.3A and B: Retroperitoneal tumour—benign lipoma was excised with a loin retroperitoneal approach.

- In retroperitoneal cyst and sarcoma the swelling is smooth.
- In lymphoma it is smooth and firm.
- In aneurysm it is pulsatile (expansile pulsation) which will persist even in knee-elbow position.
- Confirmation is by U/S, CT scan, MRI.
- Most common retroperitoneal benign tumour is lipoma.
- Sarcoma (*commonest is liposarcoma 40%*) which attains a very large size, goes for myxomatous degeneration very early and is aggressive.

RETROPERITONEAL TUMOURS

- Retroperitoneal tumour can be *benign or malignant* (refer table above); *solid or cystic*. Solid tumours are commonly malignant (85%); cystic tumours are usually benign.

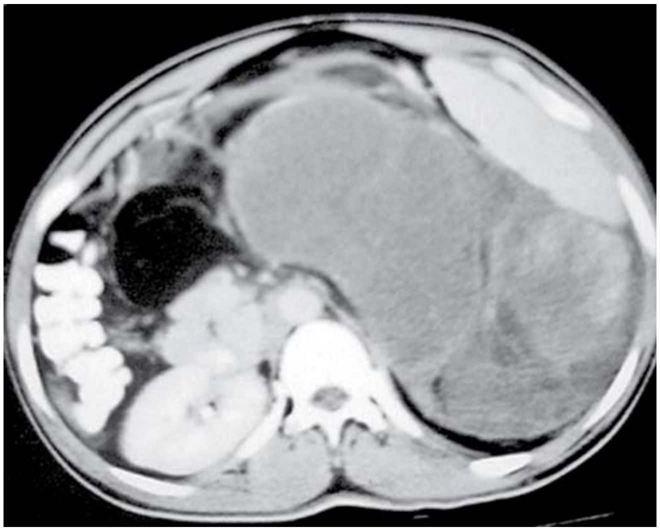

Fig. 15.4: Retroperitoneal leiomyosarcoma, CT picture.

- Malignant retroperitoneal tumour may be *mesodermal (75%),* neuroectodermal (24%) or embryonic remnant (1%) in origin. Rarely it can be urologic or metastatic also. Sarcoma is the most common malignancy; *in adult*—liposarcoma is the *most common* retroperitoneal sarcoma (50%); leiomyosarcoma is 30%; malignant nerve sheath tumour (5%); fibrosarcoma (5%); malignant fibrous histiocytoma (MFH—10%) can also occur. 20% show neurological deficit; 40% show distant blood spread. *In children* rhabdomyosarcoma is the most common one. 15% of all sarcomas are of retroperitoneal.

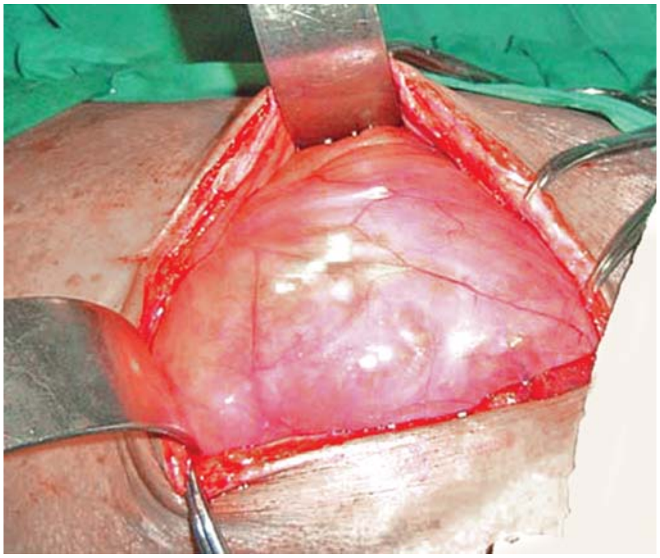

Fig. 15.5: Retroperitoneal tumour, on table finding. It was large and inoperable.

Aetiology

- Retroperitoneal tumors may be associated with radiation; exposures to chemicals like vinyl chloride/thorium dioxide, familial Gardner's syndrome, familial retinoblastoma or familial neurofibromatosis, Li Fraumeni syndrome, germline mutations of p53 in chromosome 17.

Features

- Tumor arises from fat, areolar tissue, vessels, nerves, skeletal and smooth muscle, fascia, lymph node and lymphatic systems and tissues of embryologic origin. Sarcomas can be commonly solid, cystic or mixed. Usually tumour develops a pseudocapsule. Liposarcoma often originates from (> 30%) perinephric fat.
- *Liposarcoma* can be well-differentiated (60%) or undifferentiated (dedifferentiated). It spreads along the fascial planes *enveloping* the organs. Infiltration of organs occurs only at late stage. It shows whorl like pattern in histology with myofibroblastic or osteoblastic differentiation. Insulin like production by tumour or rapid utilisation of glucose by active tumour causes intermittent *hypoglycaemia (paraneoplastic)*.
- Spread of retroperitoneal sarcoma mainly occurs through blood to lungs and often to liver. Nodal spread occurs rarely (5%).

Presentations

- Most patients have initial asymptomatic course with eventual presentation as *mass abdomen* (90%). Pelvic mass (20%), vague abdominal pain and discomfort (70%), anorexia, fatigue, vomiting, weight loss, fullness in abdomen, back pain due to compression over lumbar and pelvic nerves, leg pain, recurrent episodes of fever—are other features.

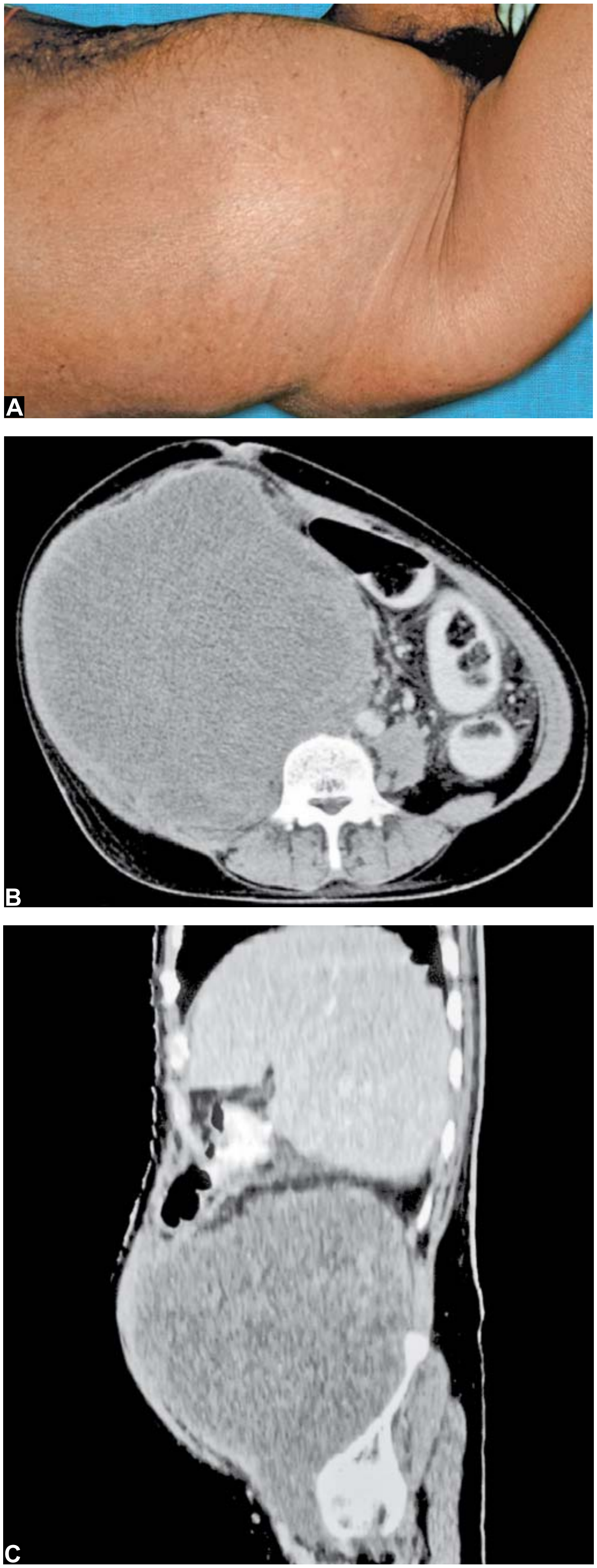

Figs 15.6A to C: Retroperitoneal tumour presenting as large mass abdomen with all features. CT pictures of the same patient.

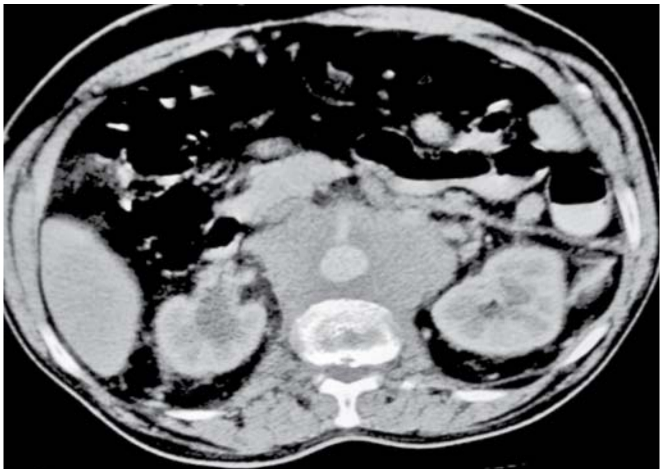

Fig. 15.7: Retroperitoneal tumour showing encasement of major vessels. It becomes inoperable. Ureter, duodenum, intestines, kidney are other structures which may get infiltrated by tumour.

- *Other* presentations are—GI bleeding; intestinal/urinary obstruction (10%); neurological manifestations (20%) by compression or invasion of neurological systems by tumor; iliac vein or IVC compression can cause lower limb varicosities, varicocele, and dilated abdominal veins with venous flow from below upwards, and oedema.

Investigations

- CT abdomen—helical contrast CT (commonly used) and pelvis and *CT chest* (to see metastases) are essential investigations.
- *CT guided core biopsy*; laparoscopic biopsy are needed. FNAC has got limited role.
- *MRI* is better than CT with near 100% accuracy. It identifies extent, desmoplastic reaction. Liposarcoma shows decreased signal in T1 and increased signal in T2. MFH shows heterogamous signals.
- Retroperitoneoscopy may be ideal to take biopsy.
- IVU shows laterally deviated ureter. Plain X-ray may show soft tissue shadow, obliterated psoas shadow, and often calcification.
- AFP, HCG and other tumour markers are done often. LFT, blood urea and serum creatinine is done.
- Histochemistry is essential in many suspicious types.

TNM staging (AJCC 6th edition)	***AJCC (6th edition) grading***
Tumour (T):	Low grade
Tx: Cannot be assessed	High grade
T0: No tumour	
T1: 5 cm or less	
T2: > 5 cm	**Staging (AJCC 6h edition):**
Nodal (N):	I: T_1/T_2; N_0; M_0; low grade
N_x: Nodes cannot be assessed	II: T_1; N_0; M_0; high grade
N_0: No nodes	III: T_2; N_0; M_0; high grade
N_1: Nodes present	IV: Any T; N_1; M_1
Metastasis (M):	*Note:*
M_x: Distant spread cannot be assessed	Nodal spread carries poor prognosis; it is equal to M_1 when outcome is considered
M_0: No metastases	
M_1: Distant spread present	

Time neither subtracts nor divides, but adds to such a pace it seems like multiplication.

Treatment

- *Surgery is the main modality* of therapy with wide excision—*complete enblock excision* (possible only in 30% of cases) is the ideal. Limited resection; debulking (controversial) are other options. Often it needs extensive surgical dissection; resection of adjacent bowel. Wide exposure is the key for successful surgery (lengthy midline/thoracoabdominal). Limiting factors are encasement of major vessels; invasion of ureters; extension into duodenum or pancreas; extensive involvement of mesentery and spread. Proper bowel preparation, colostomy, urinary diversion, adequate blood to replace the lost blood, *hypotensive* general anaesthesia are the *requirements* during surgery. Posterior dissection along the IVC, aorta and near spinal canal are the difficult areas of dissection. *Complications* are—injury to major structures, bleeding, sepsis, and leak.

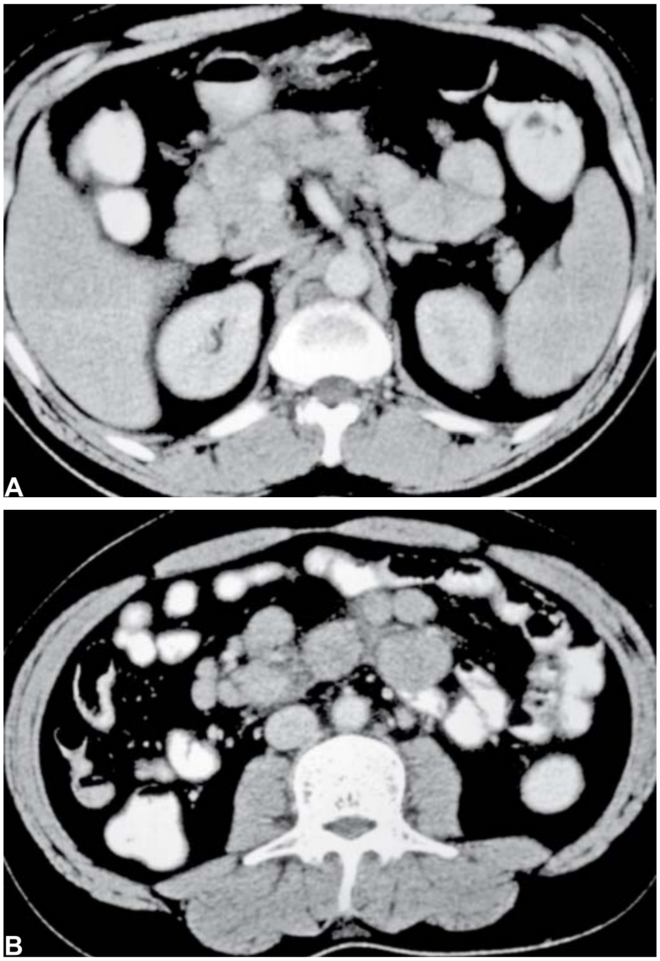

Figs 15.8A and B: Retroperitoneal lymph node mass. It could be tuberculosis/lymphoma/secondaries/autoimmune lymph node enlargement.

- Postoperative chemotherapy and radiotherapy is used as adjuvant method.
- *Radiotherapy* may be external beam RT (5000 cGy); or IORT (intraoperative RT during surgical resection) which reduces local recurrence and radiation enteritis. Problems with RT are radiation injury to kidney, bowel, liver and spinal cord. Preoperative intensity modulated RT (IMRT) is tried.
- Recurrent disease needs re-exploration but often difficult to do complete clearance.
- Mesna, *Adriamycin (main drug),* ifosfamide and dacarbazine (MAID) are the chemotherapeutic agents used.
- Germ cell tumour, lymphoma (NHL) are treated accordingly.
- *Prognosis* depends on:
  - Tumour grade—low or high;
  - Resectability;
  - Invasion into organs, veins, nerves;
  - Size of tumour < 5 cm or > 5 cm or > 10 cm;
  - Deep location;
  - Nonliposarcoma in histology—poor prognosis;
  - Recurrent disease;
  - Distant spread.
  - Tumour grade and completeness of resection without distant spread are important prognostic factors.
  - RP tumours has got overall 40% recurrence rate. Overall survival is 1½ to 2 years.
  - 5-year survival for well-differentiated tumour is 75%; < 40% for undifferentiated tumours.

Features of retroperitoneal sarcomas

- Mass per abdomen
- Pain in the back and compressive features
- GIT obstruction especially by leiomyosarcoma
- Compression over the ureter causing hydronephrosis
- Compression over the major veins causing varicose veins
- Arterial compression features are not common
- Liver and lung secondaries can occur
- CT guided biopsy is ideal method of investigations

Treatment of retroperitoneal tumours

- *Wide local excision.* Often requires removal of adjacent structures like colon, kidney, and spleen
- *Debulking operation* as a palliative
- *Enucleation* through pseudocapsule. It has got high recurrence rate—not done; may be useful in advanced cases
- *Radiotherapy* either brachytherapy or external beam radiotherapy or intraoperative RT
- *Chemotherapy* using adriamycin
- Retroperitoneal NHL is treated by *chemotherapy*

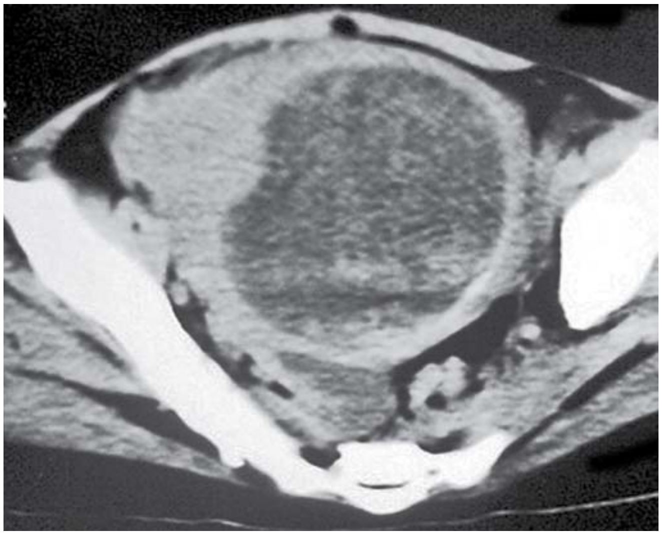

Fig. 15.9: Large pelvic retroperitoneal tumour.

Retroperitoneal sarcoma (RPS)

- Accounts for *15%* of all sarcomas
- Common *presentations* are—nontender palpable RP mass (90%); weight loss, pain and discomfort; neurologic (20%); urinary (10%); weight loss (20%); bowel obstruction; paraneoplastic syndrome (hypoglycaemia)
- 70% present with tumour more than 10 cm; 20% present with metastases
- Liposarcoma is the *most common* RP sarcoma; it can be well-differentiated, dedifferentiated, myxoid, round cell, pleomorphic
- *Histological grading* is important prognostic factor—*high grade* carries median survival of 2-3 years; *low grade* has got survival of 10 years
- CT scan, *CT guided core biopsy* through posterior approach, MRI (extremely accurate)—are the investigations needed. CT chest and abdomen, isotope renal scintigraphy or IVU to check the function of opposite kidney has to be done which may required to be removed
- A preoperative biopsy is *not usually* indicated in patients with resectable RPS; but is needed when preoperative therapy alters or lymphoma/germ cell tumours are suspected
- Open biopsy is also *not* done in a resectable tumour as it is potentially dangerous, can cause peritoneal contamination. It is done only if CT core biopsy is inconclusive
- FNAC in RPS is *not useful*
- PET scan do not have usual role in screening, diagnosis and staging. It *can be used* to quantify tumour activity; to assess the response of neoadjuvant chemotherapy and response of newer therapies targeted at specific molecular level to arrest the tumour growth
- **Dutch/memorial Sloan-Kettering cancer centre classification system of RPS (post-surgical)** is commonly used:
  - Stage I – Low grade, complete resection, no metastases
  - Stage II – High grade, complete resection, no metastases
  - Stage III – Any grade, incomplete resection, no metastases
  - Stage IV – Any grade, any resection, distant metastases
- ***Surgery is the choice of therapy***—monobloc complete surgical resection with dissection outside the pseudocapsule with adequate negative margin is possible only in 50% of cases
- Often associated visceral organ resection (in 50% cases) is needed—kidney and adrenals (45% cases), colon (25%), small bowel, pancreas (15%), spleen (10%). Need for removal of kidney arises when there is encasement of renal hilar vessels and fascia of Gerota, but not invasion of kidney
- **Adjuvant RT to retroperitoneum** is used to prevent local recurrence. It can be—preoperative, intraoperative (IORT), postoperative. It can be external beam (EBRT) or brachytherapy. *Advantages of preoperative RT* are—well demarcated gross tumour; radiosensitive viscera is displaced away by tumour itself without adhesions; absence of potential contamination of abdomen cavity; better oxygenated tumour bed. *IORT is given* as a single large dose (20 Gy) of RT to the retroperitoneum on table after displacing sensitive viscera away. *Intensity modulated RT (IMRT)* gives 45 Gy RT to planned gross tumour with 1.5 cm margin with intergrated booster doses of 65 Gy at 2.5 Gy per fraction to the margins at high-risk; IMRT has minimal local toxicity like damage to kidney, bowel (stricture), spinal cord. *Postoperative RT* has more disadvantages and toxicity but *allows* patient selection in terms of risk of local recurrence
- ***Chemotherapy (CT)***—no role of neoadjuvant CT. *Adjuvant CT* may be useful in synovial and myxoid types of liposarcoma. Palliative CT is used in unresectable and metastatic RPS—*MAID and AIM* regimes are used
- ***Specific targeted therapy on tumourogenesis***—antiangiogenetic agent, tyrosine kinase inhibitors, and mammalian target of rapamycin inhibitors (MTOR) are used. Trabectedin—synovial sarcoma, myxoid sarcoma, leiomyosarcoma
- ***Prognostic factors***—*tumour grade* and *completion of surgical resection* are the best prognostic factors (not tumour size)
- ***Local recurrence*** is common in 50% of cases. It is usually seen in high grade liposarcoma. Leiomyosarcoma, malignant peripheral nerve sheath tumours show distant spread commonly along with local recurrence

PSOAS ABSCESS

- Abscess in the psoas region in the iliac fossa, is commonly due to tuberculosis of spine (T 10) and lymph nodes, retroperitoneal structures, lower intercostal region.
- It can also be a pyogenic abscess.

Clinical Features

- Back pain, fever, spinal tenderness, paraspinal spasm and restricted spine movement.
- Swelling in iliac fossa which is smooth, nontender, nonmobile, not moving with respiration.
- When it extends across the groin and below the inguinal ligament, it is ***cross fluctuant*** across the inguinal ligament.
- Patient has typical *psoas muscle spasm* in which patient is able to flex his hip but finds it difficult to extend.

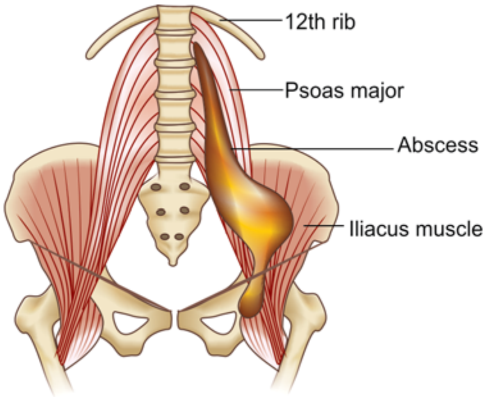

Fig. 15.10: Psoas abscess.

A man carries success or his failure with him; it does not depend on outside conditions.

Swellings which are cross fluctuant

- Psoas abscess
- Ranula
- Compound palmar ganglion
- Bilocular hydrocele

Differential Diagnosis

a. Iliac artery aneurysm.
b. Ovarian cyst in females.
c. Soft tissue tumours and cysts in the iliac region.
d. Hernia.

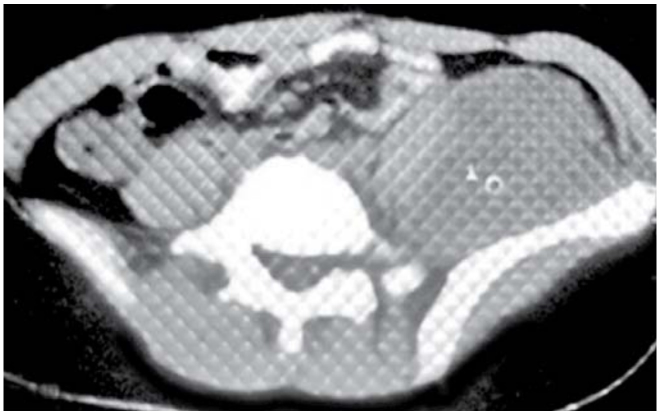

Fig. 15.11: CT picture of large psoas abscess.

Investigations

- X-ray spine and chest, CT scan.
- Mantoux test, ESR, peripheral smear.
- U/S abdomen.

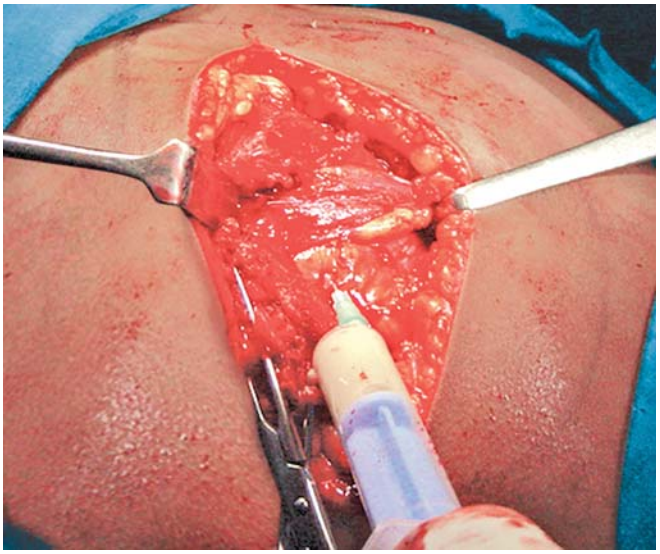

Fig. 15.12: On table finding of psoas abscess.

Treatment

- Antituberculous drugs are started—INH, ethambutol, rifampicin, pyrazinamide.

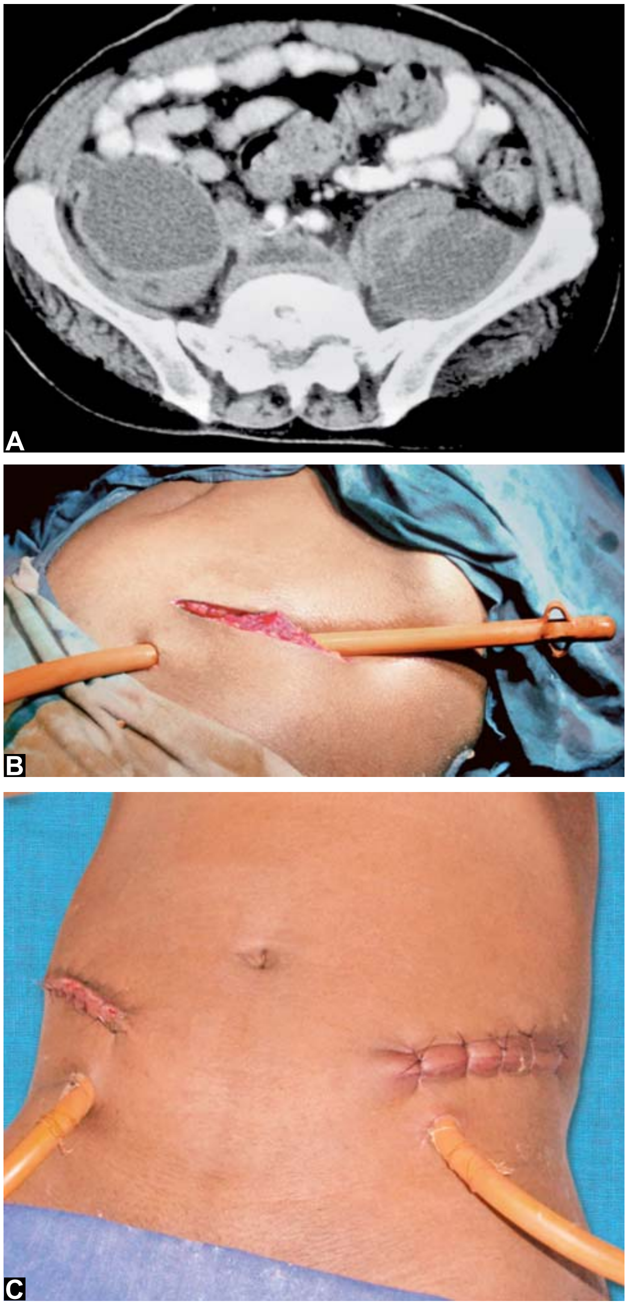

Figs 15.13A to C: Bilateral large psoas abscess—CT picture. Drainage and placement of Malecot catheters for the same. Note the location of incisions. It is obliquely placed incision above the lateral aspect of groin with extraperitoneal approach.

- Under G/A through lateral loin incision, the psoas region is reached extraperitoneally. The pus is drained and collected for culture and sensitivity, AFB, culture. The wound is closed. All caseating material and diseased parts of vertebra should be removed.
- Only lateral approach is advised. In thoracic region—*anterolateral decompression* with *costotransversectomy* with *bone grafting* is the treatment.

Note:
Posterior approach, i.e. laminectomy—**not advised.**

Chapter 16 Differential Diagnosis of Mass Abdomen

CHAPTER OUTLINE

- Mass in the Right Hypochondrium
- Mass in the Epigastrium
- Mass in the Left Hypochondrium
- Mass in the Lumbar Region
- Mass in the Umbilical Region
- Mass in the Right Iliac Fossa
- Mass in the Left Iliac Fossa
- Mass in the Hypogastrium

In a patient presenting with mass abdomen, generally following clinical features should be assessed carefully.

- *Pain:* Site, nature, aggravating or relieving factors, duration of pain, referred pain.
- *Vomiting*: Type, content, haematemesis, relation to food, frequency.
- *Jaundice*: It is an important factor in relation to liver, gallbladder or pancreatic masses.

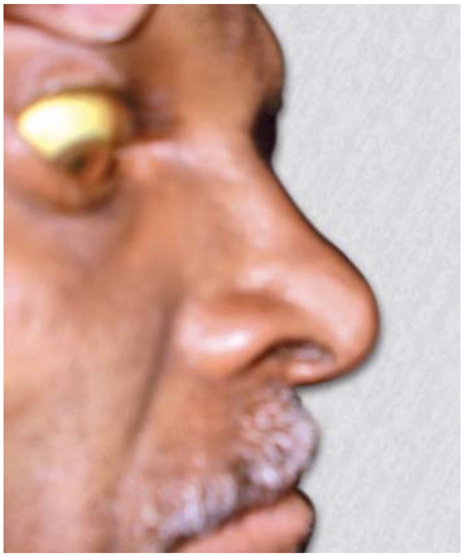

Fig. 16.1: Obstructive jaundice in a patient with carcinoma head of pancreas. Note the sclera for discolouration. Severe itching is common in these patients.

- *Bowel habits*: Constipation, diarrhoea, bloody diarrhoea, furious diarrhoea, tenesmus.
- Decreased *appetite* and *weight*.
- *Inspection of the mass*: Anatomical location, margin, surface, movement with respiration.
- *Palpation of the mass:* Site, extent, surface, tenderness, consistency, movement with respiration, mobility, borders, plane of the swelling (by leg raising test), presence of other masses.
- *Often mass needs to be examined for change of position*—in sitting, in standing, in side position, after a brisk walk, in knee elbow position for retroperitoneal mass and *for puddle sign* (but difficult to keep patient in this position).
- *Percussion* is an important aspect of examination in case of an abdominal mass. Percussion over the mass is important to predict the anatomical location of the mass. If mass is dull, then it is in the anterior abdominal wall or in front of the bowel intra-abdominally like liver, spleen, gallbladder. If the mass is with a impaired resonant note, then the mass is arising from the bowel like stomach, colon, small bowel. If the mass is resonant on percussion, then the mass is probably in the retroperitoneal region. Other than this, liver dullness, free fluid in the abdomen should be elicited during percussion.
- *Per rectal examination*: It is done to look for any secondaries in rectovesical pouch, primary tumour or relation of lower abdomen masses (pelvic masses).
- *Pervaginal examination* is done to assess pelvic masses.

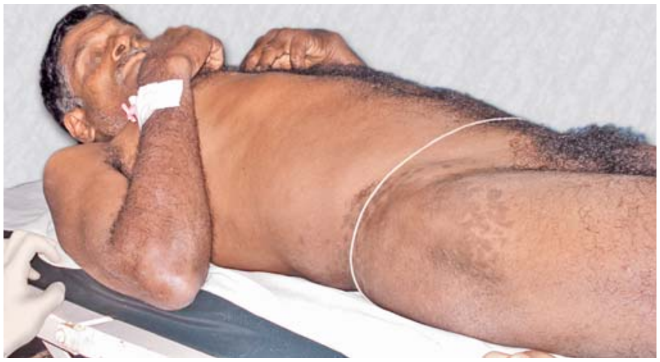

Fig. 16.2: Abdomen should be inspected by exposing from nipple level to knee level. Inspection is done from foot end and from right side.

Diagnosis of diseases os often easy; often difficult and often impossible. —**William Harvey**

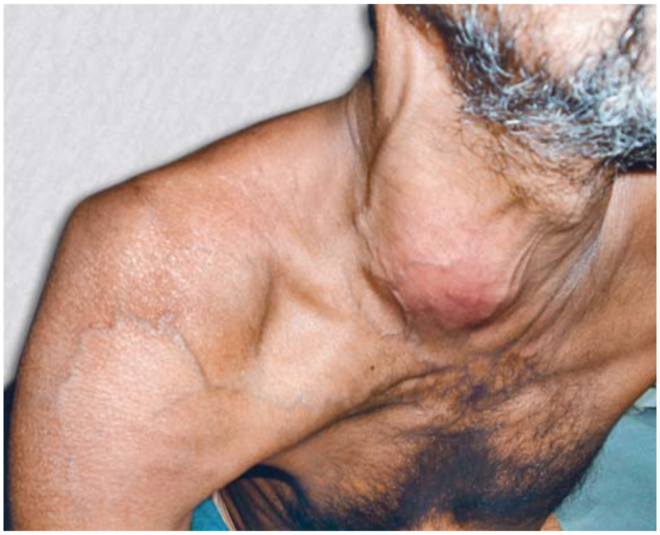

Fig. 16.3: Superior vena caval obstruction causing dilated veins in the neck, chest wall and shoulder. Note the neck swelling extending into the mediastinum.

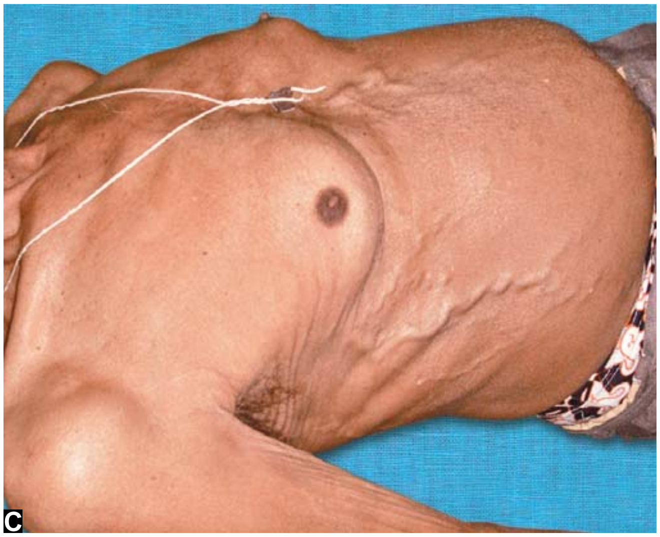

Fig. 16.4C

Figs 16.4A to C: Inferior vena caval obstruction causing dilated veins over the lateral aspect of the flank with flow of blood upwards.

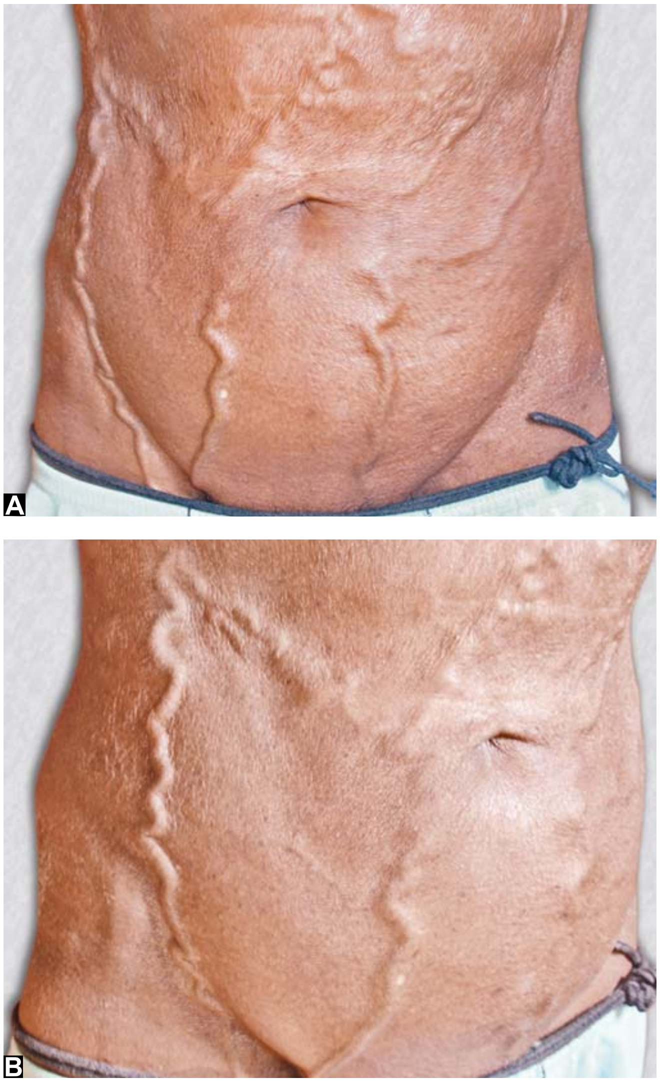

Figs 16.4A and B

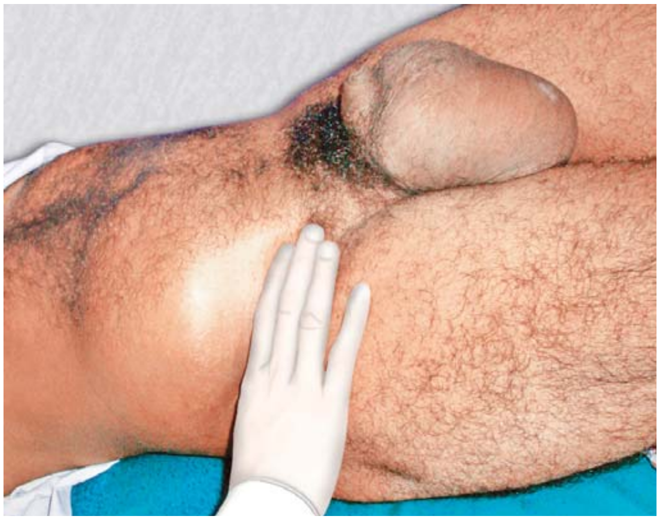

Fig. 16.5: Abdomen mass should be palpated clinically properly.

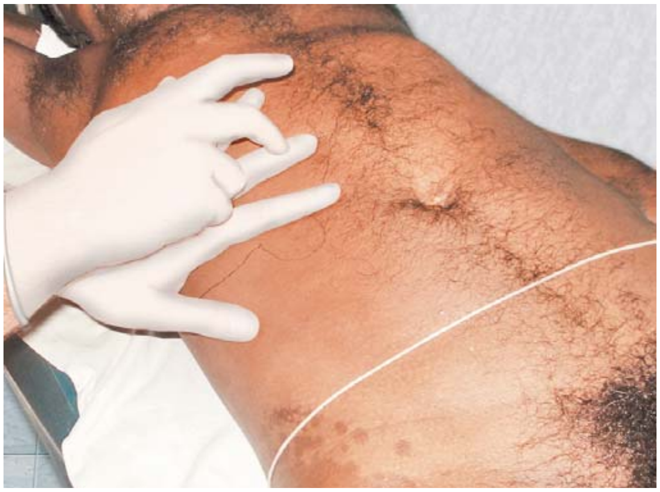

Fig. 16.6: Percussion over the mass is absolute need to find out the plane of the mass. Anterior masses are dull on percussion; posterior masses are resonant and mass from bowel has impaired resonance.

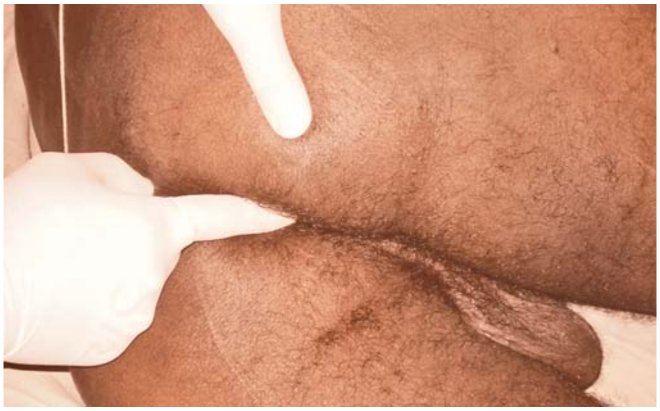

Fig. 16.7: Digital examination of rectum is a must in patients with mass abdomen to look for secondaries, possible primary and often to assess the mass itself.

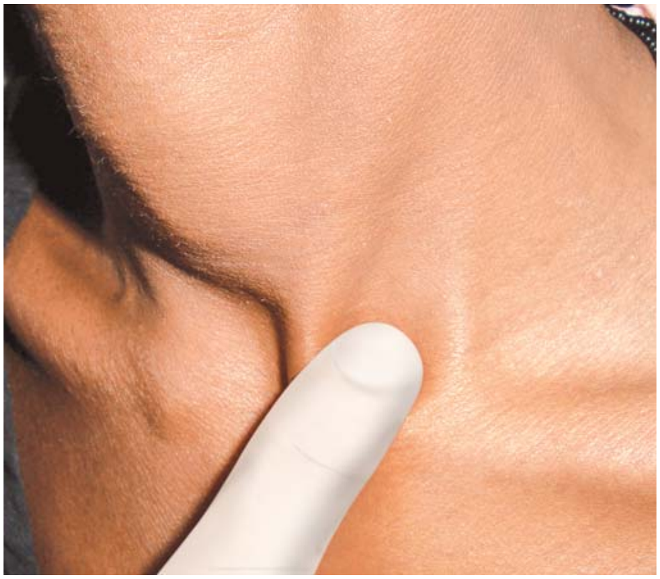

Fig. 16.8: Neck should be examined for left supraclavicular lymph node enlargement in all abdominal masses.

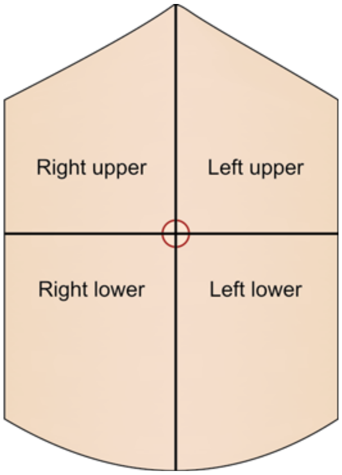

Fig. 16.9: Different quadrants in the abdomen. They are four in number formed by two lines—one is vertical midline through the umbilicus; another is horizontal line passing through the umbilicus. Quadrants are—right upper, right lower, left upper and left lower.

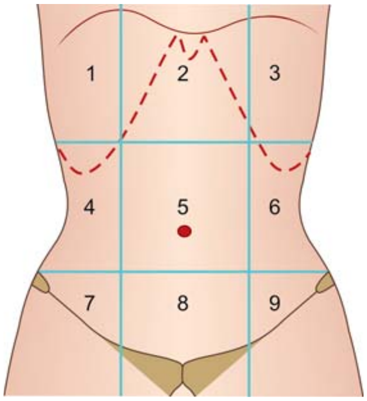

Fig. 16.10: Regions in the abdomen.

Regions in the abdomen

1. Right hypochondrium
2. Epigastric region
3. Left hypochondrium
4. Right lumbar region
5. Umbilical region
6. Left lumbar region
7. Right iliac fossa
8. Hypogastrium
9. Left iliac fossa

Abdomen is divided into nine regions by four lines.

1. Upper horizontal or *transpyloric line* is mid-way between the umbilicus and xiphisternum.
2. Lower horizontal line is *transtubercular line* at the level of two tubercles on the iliac crest.
3. Right vertical line is the line through the midpoint of right anterosuperior iliac spine and pubic symphysis.
4. Left vertical line is the line through the midpoint of left anterosuperior iliac spine and pubic symphysis.

MASS IN THE RIGHT HYPOCHONDRIUM

Liver Palpable as Mass in Right Hypochondrium

- It is horizontally placed.
- It usually moves with respiration.
- Upper border is not felt.
- It is dull on percussion (This dullness continues over liver dullness above).
- Fingers can not be insinuated under right costal margin.

Conditions where liver gets enlarged:

1. *Soft, smooth, nontender liver:*
   - *Hydrohepatosis*: It is due to obstruction of CBD causing dilatation of intrahepatic biliary radicles.
   - Congestive cardiac failure.
   - Hydatid cyst of the liver: Here mass is well-localised in the liver with typical hydatid thrill. *Three finger test:* Three fingers are placed over the mass widely. When central finger is tapped fluid movement is elicited in lateral two fingers.

> *As beds and couches vary in height, the examiner to palpate the abdomen, if need be, must sit on a suitable chair or even kneel upon the floor, no matter how undignified this may appear.* — **Charles P**

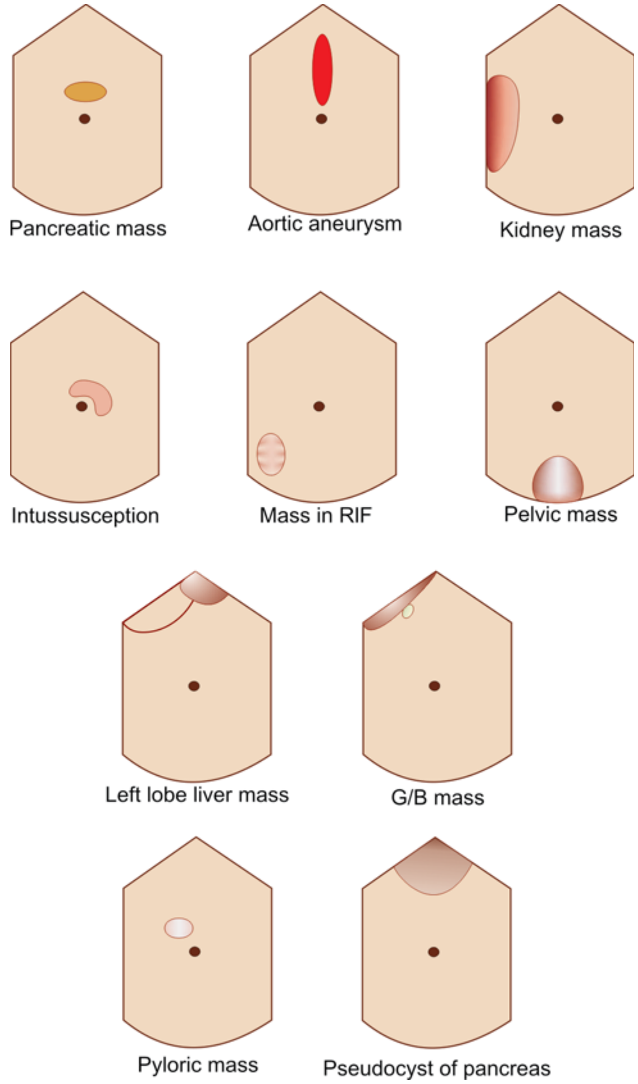

Fig. 16.11: Mass abdomen.

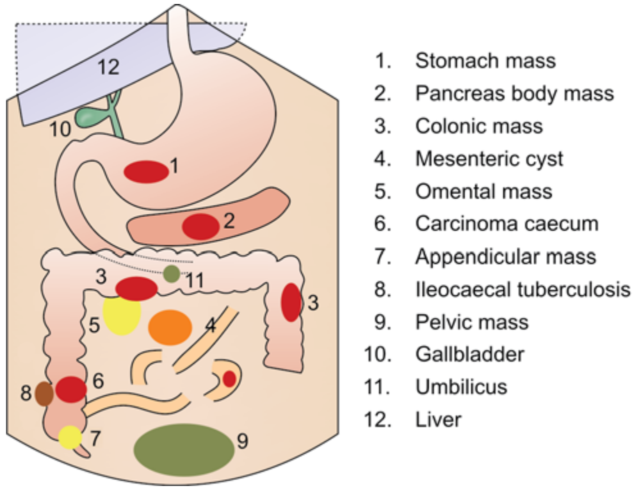

1. Stomach mass
2. Pancreas body mass
3. Colonic mass
4. Mesenteric cyst
5. Omental mass
6. Carcinoma caecum
7. Appendicular mass
8. Ileocaecal tuberculosis
9. Pelvic mass
10. Gallbladder
11. Umbilicus
12. Liver

Fig. 16.12: Different causes of mass abdomen.

2. *Soft, smooth, tender liver:*
   - *Amoebic liver abscess:* Here liver often gets adherent to the anterior abdominal wall and will not move with respiration. Intercostal tenderness, right sided pleural effusion are common.

Amoebic liver abscess

- It is due to entamoeba histolytica infestation
- It is more common in alcoholics and cirrhotics
- Single abscess is common—70%; common in right postero-superior lobe—80%
- Chocolate colored Anchovy sauce pus is classical
- Secondary infection can occur—30%—life-threatening due to septicaemia
- It can be acute or chronic; both mimics hepatoma
- Rupture into lungs—most common site of rupture
- Most dangerous rupture is into pericardium—left lobe abscess
- Liver failure can develop in cirrhotic patient

Features

- Common in males (20:1), fever, pain, intercostal tenderness, tender liver
- Mimics cholecystitis, subphrenic abscess, hepatoma
- Total count, LFT, prothrombin time, US abdomen are relevant investigations
- Chest X-ray may show left sided sympathetic pleural effusion
- CT scan to differentiate from hepatoma
- *Treatment*—drugs like metronidazole, injection dehydroemetine, chloroquine tablets, diloxanate furoate; U/S guided aspiration after controlling prothrombin time using inj vitamin K or FFP; if recurs percutaneous guided drainage using pigtail catheter, or open laparotomy and drainage with placement of Malecot's catheter

3. *Hard, smooth liver:*
   - *Hepatoma* (HCC): Here a large, single, hard nodule is palpable in the liver. But occasionally there can be multiple nodules when it is multicentric. Rapidly growing tumour can be soft also. Hepatoma often can also be tender due to tumour necrosis or stretching of the liver capsule. Vascular bruit may be heard over the liver during auscultation. It mimics amoebic liver abscess in every respect.
   - Solitary secondary in liver.

Hepatoma/hepatocellular carcinoma/HCC

- Common aetiologies are aflatoxins, hepatitis B and hepatitis C virus infection, alcoholic cirrhosis, haemochromatosis, smoking, hepatic adenoma, clonorchis sinensis, polyvinyl chloride
- Unicentric and right lobe involvement is more common
- Fibrolamellar variant is common in left lobe, not related to hepatitis or cirrhosis without AFP level raise. There are increased serum vitamin B_{12} binding capacity and neurotensin levels.
- It can be multifocal/indeterminate/spreading/expanding—*Okuda classification*
- Presents as large smooth hard liver mass—later jaundice, fever, pain and tenderness, ascites and bruit over mass
- Spreads to lymphatics, blood and direct spread
- Mimics amebic liver abscess, secondaries, hydatid cyst, polycystic liver disease
- LFT, CT scan, raised AFP, liver biopsy (only needed) are the investigations
- Hemihepatectomy in early operable growth is the treatment
- Hepatic artery ligation/intra-arterial chemotherapy/chemoembolisation/percutaneous ethanol or acetic acid injection/radiofrequency ablation/chemotherapy using adriamycin, carboplatin, gemcitabine—are palliative procedures

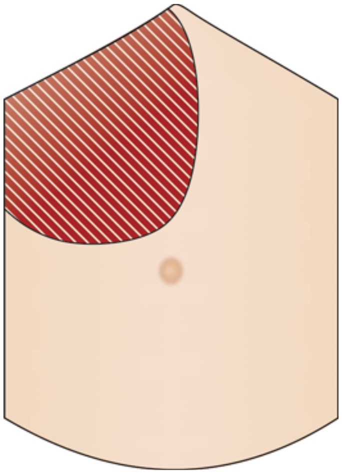

Fig. 16.13: Hepatocellular carcinoma/hepatoma. It is common in right lobe, unicentric, attains large size.

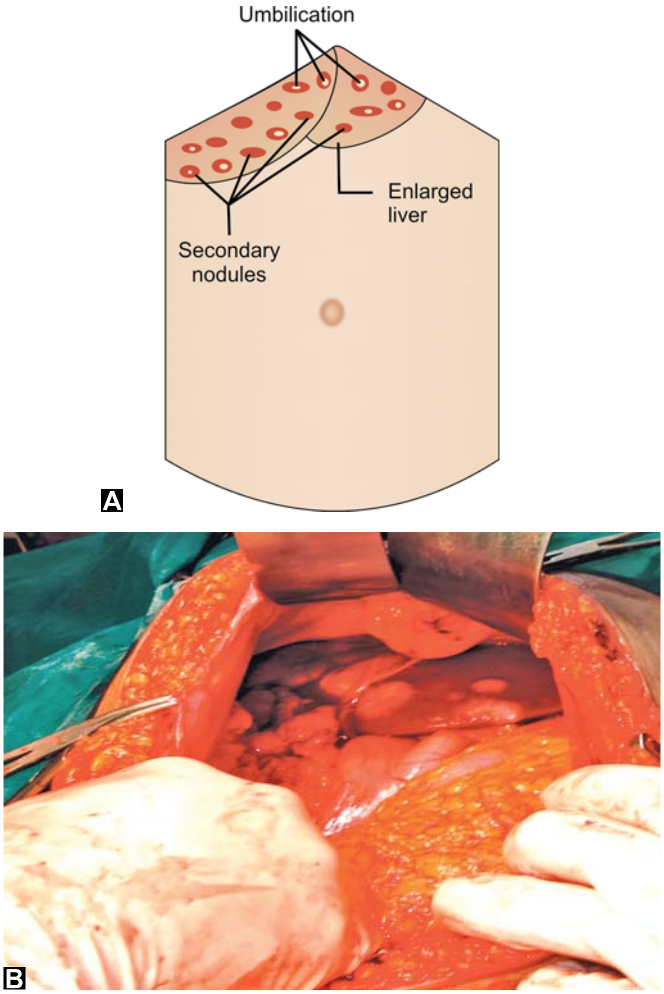

Figs 16.14A and B: Multiple secondaries in liver with umbilication. It is due to central necrosis. Secondaries are the most common malignant tumour of the liver. It could be from GIT or extra-gastrointestinal sites like from breast, lungs, melanoma, thyroid, prostate, kidneys, etc. Patient with liver secondaries has got poor general condition. It should be differentiated from multicentric hepatoma. It is usually treated by palliative chemotherapy. Solitary secondary from carcinoma colon can be removed by segmentectomy.

4. *Hard, multinodular liver:*
   - Multiple secondaries in liver: Here hard nodules show umbilication which is due to central necrosis.
   - Macronodular cirrhotic liver.

Palpable Gallbladder in Right Hypochondrium

- It is smooth and soft (except in carcinoma gallbladder).
- It is mobile horizontally (side-to-side).
- It moves with respiration.
- It is located right of the right rectus muscle, below the right costal margin or below the lower margin of the palpable liver.
- It is dull on percussion.

Conditions where gallbladder is palpable:

1. *Soft, nontender gallbladder*:
   - Mucocele of the gallbladder.
   - Enlarged gallbladder in obstructive jaundice due to carcinoma head of the pancreas or periampullary carcinoma or growth in the CBD.
2. *Hard gallbladder:*
   - Carcinoma gallbladder.
3. Tender gallbladder—empyema GB.

Other Masses in the Right Hypochondrium

- *Pericholecystic inflammatory mass*: It is tender, smooth, firm or soft, nonmobile, intra-abdominal mass often with guarding.
- *Kidney mass arising from upper pole of the kidney*: It may be due to renal cell carcinoma or hydronephrosis.

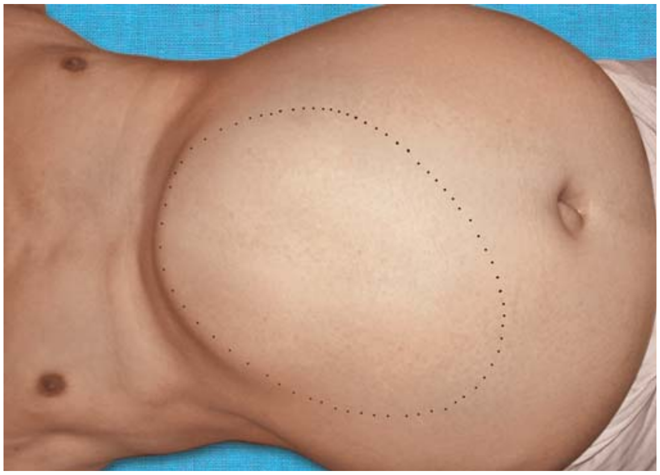

Fig. 16.15: Large upper abdomen mass—could be liver mass, pancreatic mass, retroperitoneal mass.

MASS IN THE EPIGASTRIUM

Palpable Left Lobe of the Liver

- It is in the epigastric region.
- Its upper border cannot be felt.
- It moves with respiration.
- It extends towards left hypochondrium.
- It is dull on percussion.

After every end is a new beginning and every beginning has an end.

Conditions where left lobe of the liver is palpable

- Hepatoma
- Amoebic liver abscess in left lobe
- Left lobe secondaries
- Hydatid cyst of the left lobe

Features of Stomach Mass

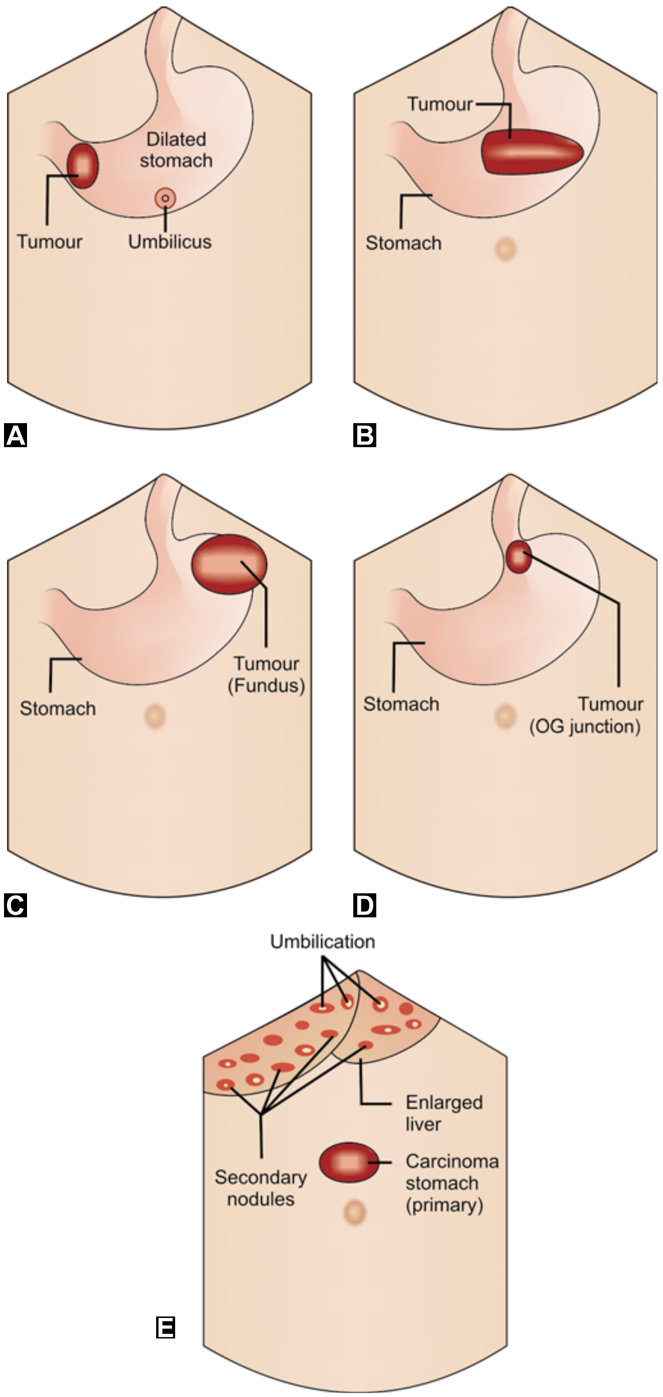

Figs 16.16A to E: Carcinoma pylorus causes gastric outlet obstruction with palpable mass above the umbilicus. Carcinoma body of stomach mainly presents as loss of appetite and decreased weight with horizontally placed stomach mass. Carcinoma from fundus of the stomach presents as mass abdomen with loss of appetite and weight. Carcinoma oesophagogastric (OG) junction presents as dysphagia. Carcinoma stomach is one of the common causes of secondaries in liver.

- It is located in the epigastric region.
- It moves with respiration. It is intra-abdominal.
- It is resonant or impaired resonant on percussion.
- Mass may be better felt on standing or on walking.
- Mass is often mobile, unless it gets adherent posteriorly.
- In pylorus mass, all margins are well felt which is mobile with features of gastric outlet obstruction.
- Mass from the body of the stomach is horizontally placed without any features of obstruction.
- Mass from the upper part of the stomach near the OG junction causes dysphagia.
- Mass from the fundus of the stomach is in the upper part of the epigastric region towards left side.
- Carcinoma stomach is nodular and hard. It is the most common cause for stomach mass.
- Leiomyoma of stomach is smooth and firm.

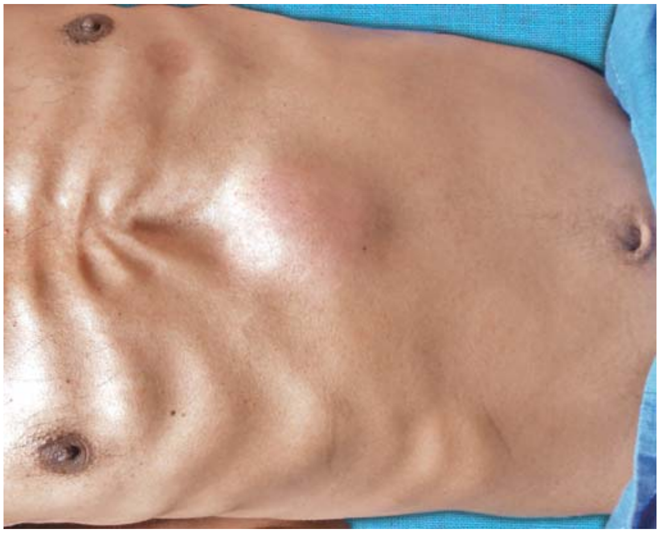

Fig. 16.17: Epigastric mass arising from carcinoma stomach.

Management of gastric carcinoma

- Early growth—pylorus—lower radical gastrectomy with removal of tumour, proximal 5 cm clearance, nodal clearance, greater and lesser omentum, distal pancreas and spleen (now not regularly removed; it is removed to clear splenic nodes—one of the node stations) and Billroth II anastomosis or Roux-en-Y anastomosis is done. Postoperatively adjuvant chemotherapy should be given—5 fluorouracil, mitomycin, epirubicin, cisplatin
- Growth in body, proximal growth, diffuse carcinoma and generalised linitis plastica are the indications for total radical gastrectomy with oesophagojejunal anastomosis
- Neoadjuvant chemotherapy in advanced gastric cancer prior to surgery and later gastrectomy
- Instillation of mitomycin C impregnated charcoal intraperitoneally to control lymphatic disease (Japan)
- Palliative procedures like palliative partial gastrectomy, anterior gastrojejunostomy, Devine's exclusion procedure, luminal stenting in proximal inoperable growths, chemotherapy are used in inoperable cases
- In early carcinoma proper lymph nodal clearance is important

Pseudocyst of the Pancreas

- Mass in the epigastric region. It is smooth, soft. It can be tender if it is infected.
- It does not move with respiration.
- It is not mobile.

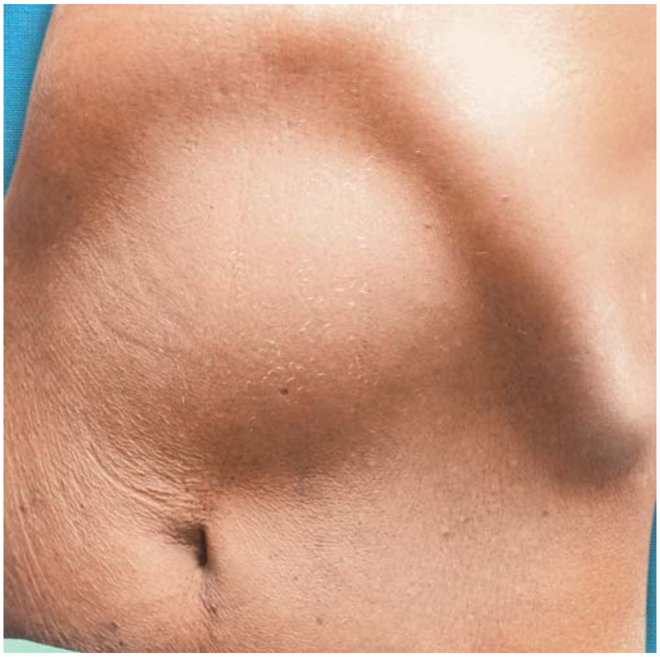

Fig. 16.18: Pseudocyst of pancreas presenting as epigastric mass.

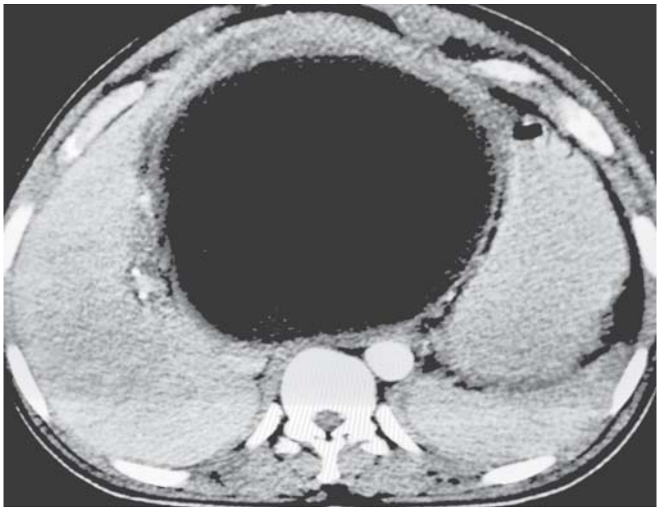

Fig. 16.19: Pseudocyst of pancreas—CT scan picture.

- It has got transmitted pulsation. It is confirmed by placing the patient in knee-elbow position.
- Lower border is well felt. Upper border is not clear.
- It is resonant on percussion.
- *Baid test*: As the stomach is pushed in front, Ryle's tube when passed, can be felt per abdomen on palpation.

Investigations for pseudocyst of pancreas	***Indications for intervention***
• Ultrasound—commonly done procedure	• Size more than 6 cm
• CT scan, ideal and choice	• Formed thick-walled pseudocyst
• LFT, serum amylase, prothrombin time	• Infected pseudocyst
• ERCP to find out communications	
• Barium meal—not done now—shows widened vertebrogastric angle	
Interventions	***Complications***
• Roux-en-Y cystojejunostomy is ideal	• Rupture—3% • Infection—20%
• Cystogastrostomy—Jurasz procedure—commonly done cystoduodenostomy	• Bleeding—torrential 7% • Cholangitis
• Cystogastrostomy with external drainage if infected—Smith operation	
• Endoscopic stenting	
• Laparoscopic cystogastrostomy—popular—safer	
• Guided aspiration helps but high recurrence rate of 70%	
• Acute fluid collection—just fluid collection	
• Acute pseudocyst with thin wall	
• Chronic pseudocyst—thick walled	
• Pseudopseudocyst—inflammatory mass of bowel, omentum, etc. after acute pancreatitis mimics pseudocyst	

Cystadenocarcinoma of the Pancreas

Mass is smooth, firm, does not move with respiration, nonmobile, resonant on percussion. Patient complaints of back pain.

Colonic Mass

- It is due to carcinoma of transverse colon.
- It is mobile, horizontally placed, nodular, hard mass which does not move with respiration. Caecum will be dilated and palpable.
- It is resonant or impaired resonant on percussion.
- Patient will be having bowel symptoms, loss of appetite and decreased weight.

Para-aortic Lymph Node Mass

- Mass in the epigastric region which is deeply placed, nonmobile, not moving with respiration.

- It is vertically placed, above the level of the umbilicus and resonant on percussion.
- Causes for enlargement are: *Secondaries, lymphomas* or *tuberculosis.*

Aortic Aneurysm

It is smooth, soft, pulsatile (expansile pulsation which is confirmed by placing the patient in knee-elbow position).

It is vertically placed above the level of the umbilicus, nonmobile, not moving with respiration and resonant on percussion.

MASS IN THE LEFT HYPOCHONDRIUM

Enlarged Spleen

- Spleen has to enlarge three times to be palpable clinically.
- It enlarges towards the right iliac fossa from left costal margin.
- It moves with respiration, mobile, obliquely placed, smooth, soft or firm, with a notch on the anterior edge which is directed downwards and inwards.

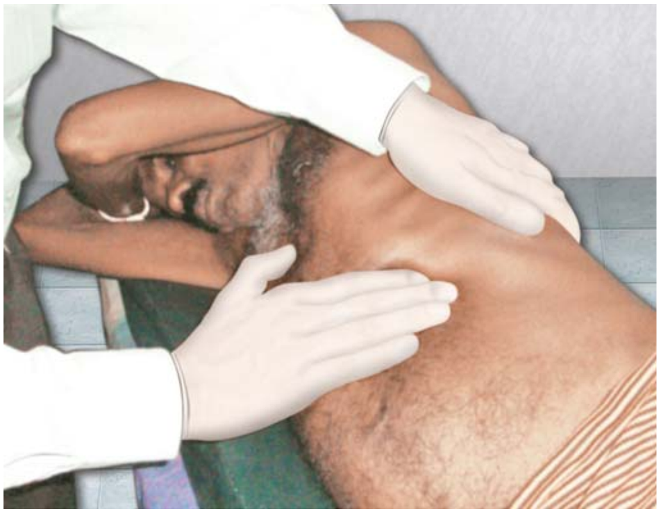

Fig. 16.20: Method of palpating spleen.

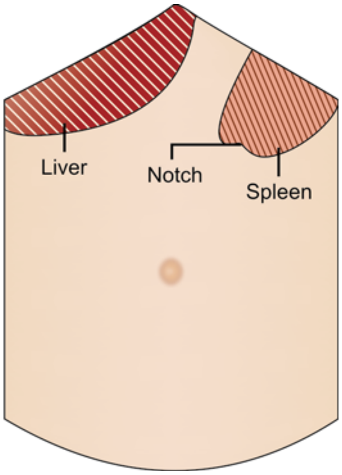

Fig. 16.21: Hepatosplenomegaly is the common condition (clinical entity). It is due to macronodular cirrhosis with portal hypertension, lymphoma, autoimmune diseases, congestive cardiac failure, hepatoma with portal hypertension, haemolytic diseases, etc. There may be ascites, supraclavicular palpable lymph node and pleural effusion (right-sided).

- Fingers cannot be insinuated over the upper border.
- "*Hook sign" is positive*, i.e. one cannot insinuate the fingers under the left costal margin.
- It is dull on percussion.

Left Sided Colonic Mass

- It is mobile, nodular, resonant.
- It does not move with respiration.
- It is commonly due to carcinoma colon.

Left Renal Mass from Upper Pole of any Cause

It has got features of renal mass.

Left Sided Adrenal Mass

- It does not move with respiration. It is not mobile.
- It is deeply placed mass. Often it crosses the midline.
- It is resonant on percussion. It mimics kidney mass.

Mass Arising from the Tail of the Pancreas

Clinical features are same as other pancreatic masses. Causes are pseudocyst in tail of the pancreas and cystadenomas.

MASS IN THE LUMBAR REGION

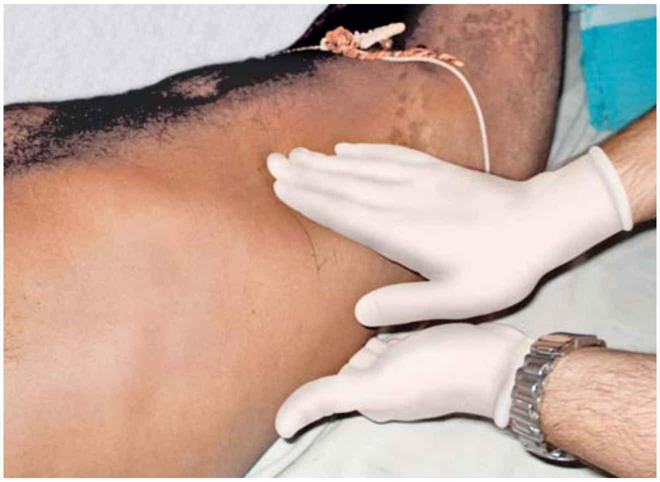

Fig. 16.22: Kidney should be palpated bimanually with two hands. Ballotability also should be checked.

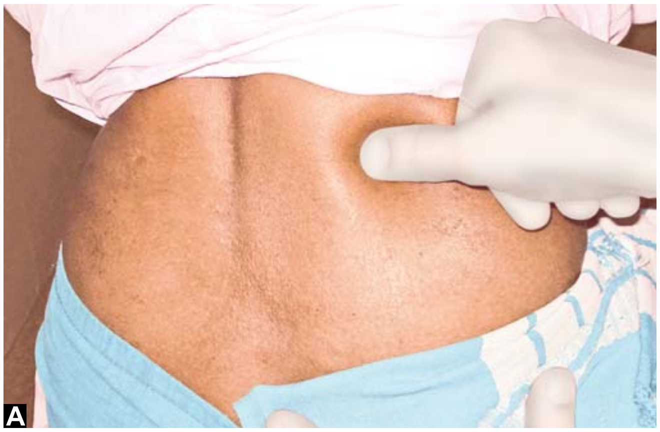

Fig. 16.23A

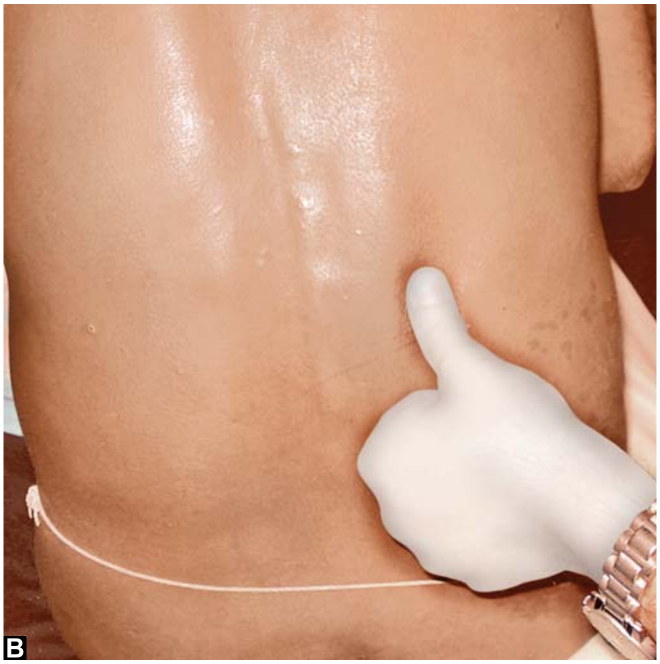

Fig. 16.23B

Figs 16.23A and B: Renal (Kidney) punch. Renal angle should be palpated for tenderness. Renal angle is junction of 12th rib and erector spinae. *On percussion it is normally resonant* as it is occupied by colon. It is replaced by kidney in kidney enlargement. Renal angle should be inspected, palpated and percussed in sitting position from behind.

Palpable Kidney Mass

- There is fullness in the loin which is better observed in sitting position.
- Mass moves with respiration. It is vertically placed.
- It is bimanually palpable. It is ballotable.
- Renal angle is dull on percussion (normally it is resonant due to colon).
- There is a band of resonance in front due to reflected colon.
- It does not cross the midline.

Conditions Where Kidney Gets Enlarged

Hydronephrosis:

- It is smooth, soft, lobulated, nontender mass, nonmobile.

Pyonephrosis:

- History of throbbing pain in the loin, pyuria and fever with chills.
- It is smooth, soft and tender kidney mass, nonmobile.

Polycystic kidney:

- History of loin pain and haematuria.
- Hypertension, anaemia and features of renal failure.
- Usually bilateral. But one side can present early than on the other side.
- Lobulated smooth surface.

Renal cell carcinoma:

- History of mass in the loin, haematuria, fever and dull pain.
- Mass is nodular and hard.
- It does not cross the midline.
- Initially mobile; eventually it infiltrates gets fixed and becomes nonmobile.

Mass from the Ascending Colon on Right Side or Descending Colon on Left Side

- History of altered bowel habits with decreased appetite and weight.
- Mass is nodular, hard which does not move with respiration and is not ballotable.
- It is resonant or there is impaired resonance on percussion.
- Renal angle is resonant.
- Proximal dilated bowel may be palpable.

Adrenal Mass

- It is nodular and hard.
- It does not move with respiration.
- It is not mobile and often crosses the midline.
- It is felt on deep palpation.
- It is resonant in front.
- It is not ballotable.

Retroperitoneal Tumours

- They are not mobile, resonant and do not fall forward in knee-elbow position.
- They are deeply placed mass which are usually smooth and hard.
- They may be retroperitoneal sarcomas or teratomas or lymph node mass.

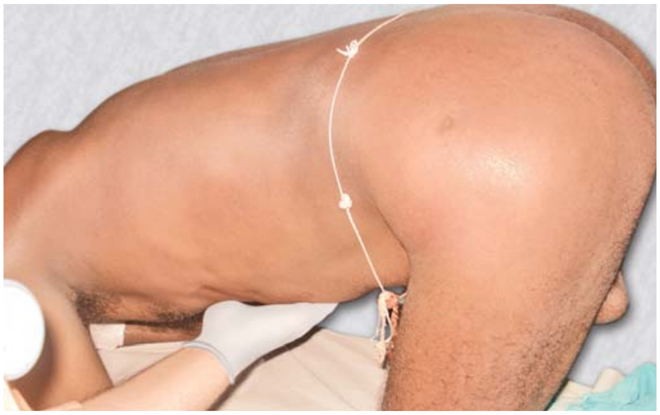

Fig. 16.24: Palpation of mass in knee-elbow position. Mobility and falling forward should be confirmed in knee-elbow position.

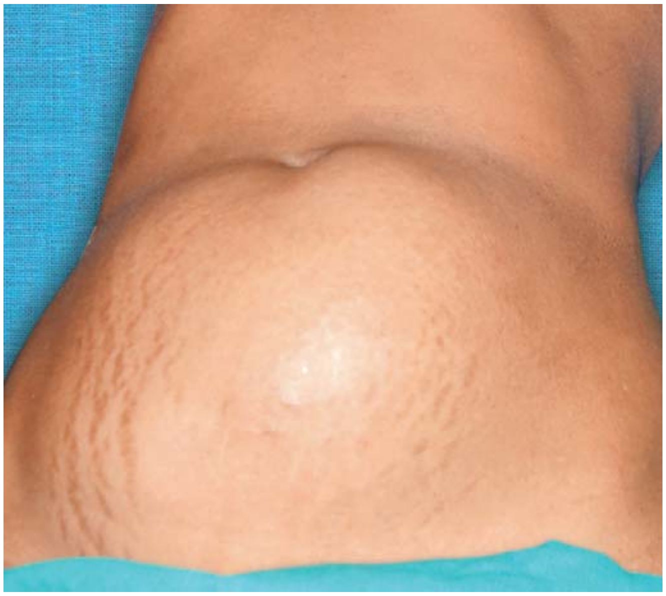

Fig. 16.25: Retroperitoneal tumour.

Retroperitoneal Cysts

They are *smooth and soft* with the same features as retroperitoneal tumours.

Cystic lesions in the abdomen

- Mucocele/empyema of gallbladder
- Pseudocyst of pancreas
- Hydatid cyst of liver
- Congenital nonparasitic cyst of liver
- Hydronephrosis
- Mesenteric cyst
- Ovarian cyst
- Omental cyst
- Aneurysm
- Retroperitoneal cyst
- Cystadenocarcinoma of ovary
- Loculated ascites

MASS IN THE UMBILICAL REGION

Usual masses are:

- Mesenteric cyst
- Omental cyst
- Ovarian cyst (pedunculated)
- Small bowel tumours
- Extension of masses from other region
- Transverse colon mass
- Mass in the body of pancreas
- Mesentery mass
- Lymph node mass—secondaries (primary from GIT, testis, ovary, melanoma)/lymphoma/tuberculosis
- Retroperitoneal tumour

Mesenteric Cyst

- ***Tillaux triad:***
  1. Soft intra-abdominal umbilical mass.
  2. Mobile in the direction perpendicular to the attachment of the mesentery.
  3. Resonant mass.
- May precipitate intestinal obstruction, volvulus.

Omental Cyst

- It is smooth, soft and nontender.
- It moves with respiration. It is mobile in all directions.
- It is dull on percussion.

Small Bowel Swellings

- Small bowel lymphomas.
- Small bowel carcinomas.
- Intussusception.

Intussusception

- Mass in umbilical region usually towards left and above the umbilicus.
- Occasionally towards right side.
- Mass is intra-abdominal which is *sausage shaped*, with concavity towards umbilicus, well-defined, smooth, firm and mobile.
- Mass does not move with respiration.
- Mass contracts under palpating fingers.
- Often mass disappears and reappears.
- Mass is resonant or there is impaired resonance on percussion.
- *"Red currant gelly"* stool with features of intestinal obstruction may be present.

MASS IN THE RIGHT ILIAC FOSSA

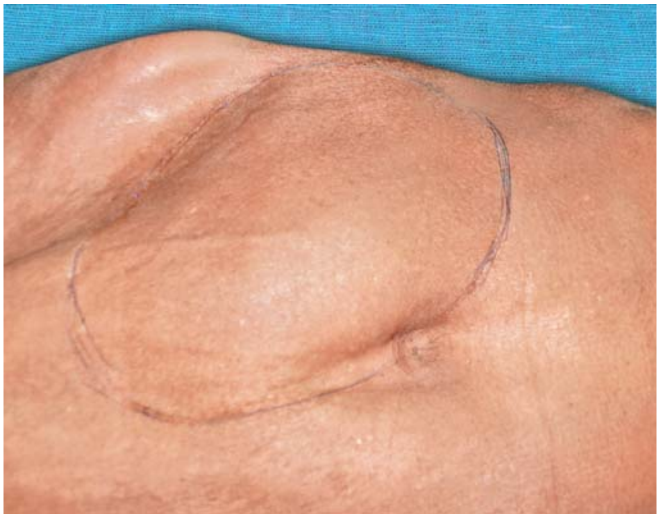

Fig. 16.26: Mass in right iliac fossa. All possible differential diagnosis should be considered and clinically analysed. Common masses are appendicular mass; carcinoma caecum; ileocaecal tuberculosis; lymph node mass; ameboma.

- Appendicular mass or abscess
- Carcinoma caecum
- Ileocaecal tuberculosis
- Amoeboma
- Psoas abscess
- Lymph node mass either mesenteric or external Iliac lymph nodes
- Bony swellings
- Ectopic kidney
- Undescended testis (Abdominal)
- Actinomycosis
- Crohn's disease
- Iliac artery aneurysm
- Ovarian swelling-ovarian cyst
- Tubo—ovarian mass
- Uterine mass like pedunculated fibroid

Appendicular Mass

- It is *smooth, firm, tender mass* in the right iliac fossa.
- It is not mobile. It does not move with respiration.
- It is *resonant on percussion*. It is well-localised mass with distinct borders.

Appendicular Abscess

It is *smooth*, *soft, tender and dull mass* in the right iliac fossa with indistinct borders.

Carcinoma Caecum

- It is nodular, hard, mass in the right iliac fossa.
- It does not move with respiration.
- It is mobile but mobility may be restricted once it gets adherent to psoas muscle.
- Mass is resonant or there is impaired resonance on percussion.
- Often features of intestinal obstruction may be present.

Ileocaecal Tuberculosis

- Mass in the right iliac fossa which is smooth, hard, resonant and nontender.
- It does not move with respiration and has restricted mobility.
- Caecum may be *pulled up* to lumbar region due to fibrosis.

Amoeboma

- History of dysentery with pain in the right iliac fossa.
- Smooth, hard, well-defined mass in the right iliac fossa which is nonmobile.
- It may or may not be tender.

Psoas Abscess

- It is localised, smooth, soft, nonmobile mass in the right iliac fossa.
- *Psoas spasm* (flexion of the hip joint) is typical.
- Spine may show *gibbus, tenderness, paraspinal spasm*. Spinal movements will be restricted.

MASS IN THE LEFT ILIAC FOSSA

- Carcinoma sigmoid or descending colon
- Bony masses
- Ovarian/uterine masses
- Psoas abscess
- Ectopic kidney
- Lymph node mass
- Undescended testis

MASS IN THE HYPOGASTRIUM

Bladder Mass

- It is in the midline. It is dull on percussion. Lower border is not felt.
- It can be mobile in horizontal direction. Mass reduces in size after emptying the bladder. It can be felt on per-rectal examination.
- It is either carcinoma bladder (common) or leiomyoma or sarcoma bladder.

Uterine Mass

- It is midline mass which is smooth, hard.
- Lower border is not felt which extends into the pelvis.
- It is felt on pervaginal examination.

Ovarian Mass

Pelvic soft tissue mass.

In all lower abdomen masses P/R and or P/V is a must.

*In all regions **parietal masses** can occur*:

- Benign and malignant soft tissue tumours.
- Common, is lipoma.
- Fatty hernia of linea alba.
- Desmoid tumour.
- Parietal wall abscess.

Blaxland ruler *test* (Athelstan J Blaxland): A flat ruler placed on the lower abdomen just above the anterosuperior iliac spines and pressed firmly backwards. In ovarian cyst aortic pulsation is transmitted to fingers through ruler; it is not so in ascites.

Carnett's Test

The patient, in lying down position, is asked to lift both legs off the bed with knee extended. This puts the abdominal muscles into contraction. Intra-abdominal mass becomes less prominent, parietal mass persists as same, but becomes less mobile.

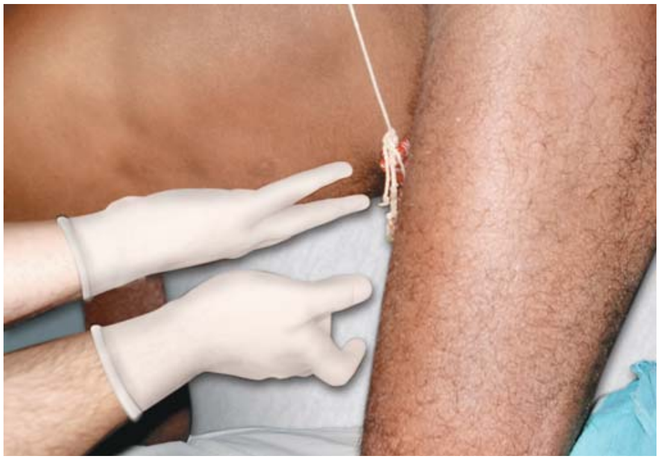

Fig. 16.27: Looking for minimal ascites in knee-elbow position—Puddle sign.

Investigations for Mass Abdomen

- Haematocrit, liver function tests, renal function tests, stool/urine examination.
- Ultrasound abdomen.
- Endoscopies-gastroscopy-colonoscopy-ERCP-MRCP.
- Barium studies-Barium meal-Barium enema-Barium meal follow through.
- CT scan-MRI.
- Endosonography.
- Ascitic tap.
- Diagnostic laparoscopy.
- U/S guided/CT guided biopsy.
- IVU/RGP/Cystoscopy/Isotope renogram.
- Exploratory laparotomy.

Note:

- Hard mass in the abdomen is commonly malignant
- Firm mass may be tuberculous, lymphoma or many benign conditions
- Soft masses are hydronephrosis, pseudocyst, mesenteric cyst, omental cyst, and loculated ascites
- In tuberculosis abdomen may be doughy due to thickened parietal peritoneum or omentum
- It is difficult to elicit fluctuation in the mass abdomen as mass cannot be fixed properly
- Plane of the swelling should be checked by leg raising/head raising test/Valsalva manoeuvre/knee elbow test
- Bimanual palpation, ballotability, renal angle inspection, palpation and percussion should be done in case of renal mass
- Intrinsic mobility should be checked. Different mobilities/movements are—gallbladder shows side to side; stomach lateral; ovarian mass—all over; mesenteric cyst—right angle to line of mesentery; transverse colon mass—vertical; small bowel mass all over; appendicular mass/pancreatic mass/nodal mass (para-aortic)/retroperitoneal mass do not show any mobility; cystadenocarcinoma of pancreas may show false mobility (*tree top mobility*)

Percussion is very important method of examination to find out the anatomical plane of the mass. Mass in front of the bowel like liver/spleen/gallbladder/parietal mass are dull on percussion. Mass from the bowel is resonant on percussion like from stomach, small bowel and colon. Retroperitoneal masses like cyst, sarcoma, nodal mass, aneurysm, pancreatic mass are resonant on percussion

- Succussion splash and auscultopercussion tests are done for gastric outlet obstruction in pyloric stenosis
- Digital examination of rectum (P/R) and left supraclavicular examination for nodes is a must. PR is done to see Blumer shelf secondaries in rectovesical pouch. Pervaginal examination should be done in pelvic mass in females
- Bladder should be emptied while examining pelvic mass
- Bimanual examination is done often under general anaesthesia in case of pelvic mass
- Auscultation for bruit depends on condition and location of mass—over epigastrium, over liver
- Other relevant systemic examination is a must in examination of mass abdomen—respiratory system, skeletal and neurological systems

Mass that appears and disappears

- Pseudocyst of pancreas (communicating)
- Hydronephrosis (intermittent)
- Choledochal cyst
- Intussusception

Srinivasan Costal Sign

"Srinivasan costal sign" is an inspectory finding to assess the acute distended abdomen secondary to surgical or post-operative status or medical causes under treatment. The bilateral costal margin is visualised tangentially on the either side of the abdomen in supine position to appreciate this costal sign.

"Prominence of costal margins will be lost in distended abdomen. During therapy, once distension reduces, prominence of costal margins will be visible (during recovery)".

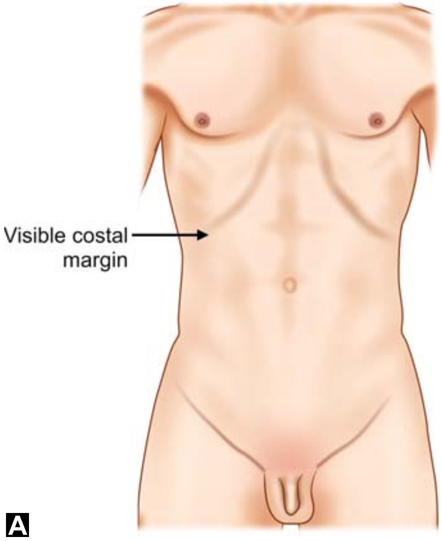

Fig. 16.28A

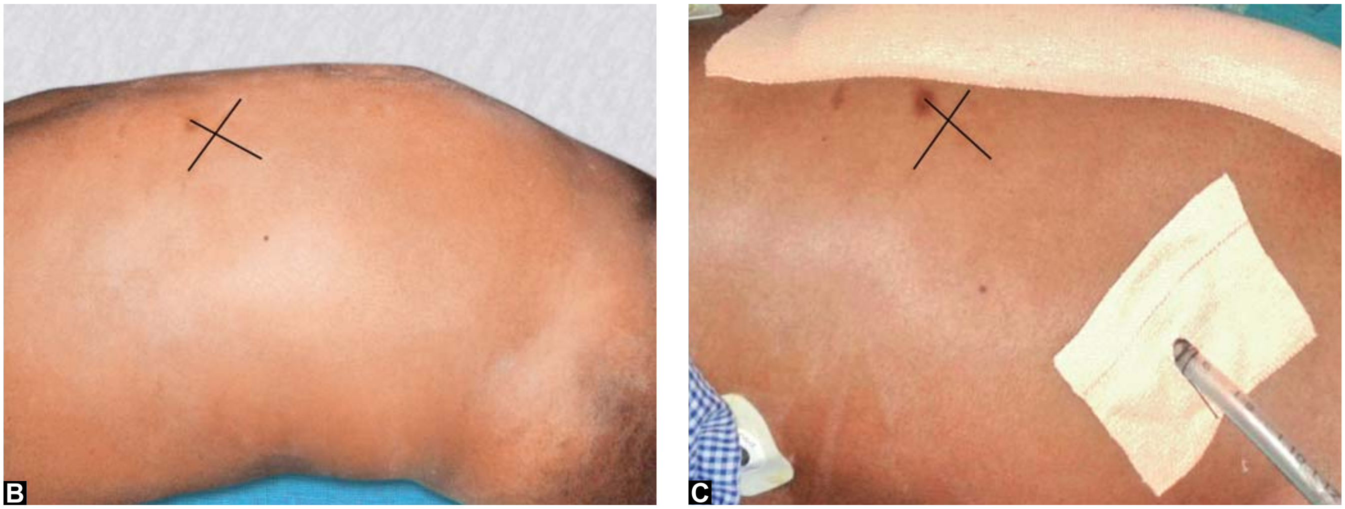

Figs 16.28B and C

Figs 16.28A to C: Srinivasan costal sign: This sign can be appreciated under inspection, the costal margin which is not visible or poorly visible in case of distended abdomen. Once the acute abdomen or distended abdomen secondary to medical cause start settling, the costal margins become visible and prominent, this holds good even in case of fatty protuberant abdomen.

Acknowledgements

With kind permission from Professor Narayanaswamy Srinivasan, MS, FRCS who was earlier worked as HOD, General Surgery, Bangalore Medical College; Director of Sanjay Gandhi Accident Hospital and Research Institute, Byrasandra, Bangalore; Dean and Director of Bangalore Medical College and Research Institute, Bangalore; Joint Director of Medical Education, Govt of Karnataka, India.

Open confession is good for the soul.

Chapter 17 Abdominal Wall and Umbilicus

> *Why do they meet this most unfortunate and lamentable fate? Have these vessels committed some great crime, that you, O Nature, a just and provident mother to all other parts of the body, have become a step-mother to these alone, so that they deserve to be visited with an affliction so dire, a punishment so severe as wasting and death? Did not the vein continuously supply blood to the fetus? Did not the arteries give life to it? Were they not the means of bestowing increase and nutrition upon the fetus?*
>
> **—Girolamo Fabrizio (Fabricius ab Aquapendente),** 1649

CHAPTER OUTLINE

- Diseases of the Umbilicus
- Omphalitis
- Umbilical Granuloma
- Anomalies of Vitellointestinal Duct
- Umbilical Sinus
- Umbilical Adenoma
- Umbilical Fistula
- Patent Urachus
- Burst Abdomen
- Abdominal Wall Tumours
- Desmoid Tumour
- Exomphalos
- Gastroschisis
- Rectus Sheath Haematoma
- Abdominal Wall Abscess
- Meleney's Progressive Synergistic Bacterial Gangrene of Abdominal Wall
- Diverication of Recti

DISEASES OF THE UMBILICUS

1. **Inflammations:**
   - Omphalitis.
   - Umbilical granuloma.
   - Pilonidal sinus.
2. **Fistulas:**
   a. *Faecal*
      - Patent vitellointestinal duct.
      - Neoplastic ulceration.
      - Tuberculosis of peritoneum.

   b. *Urinary*—Patent urachus.
3. **Neoplasms:**
   a. *Benign*
      - Adenoma (Raspberry tumour).
      - Endometrioma.

   b. *Malignant*
      - Primary (rare).
      - Secondary carcinoma—Sister Joseph's nodule through lymphatics of the round ligament, primary being in the stomach, colon, ovary, uterus, breast (often blood spread).
4. **Umbilical hernias.**
5. **Umbilical calculus (Umbolith).**

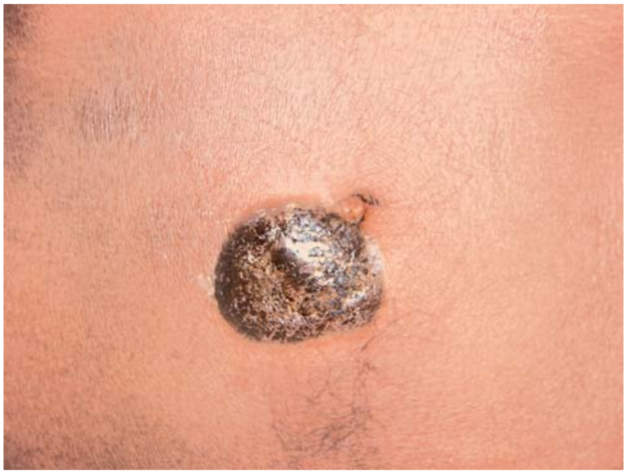

Fig. 17.1: Periumbilical pigmented lesion—could be melanoma.

OMPHALITIS

- It is infection and inflammation of umbilicus. It is usually seen in newborn babies.
- Umbilical cord which is severed during birth contains bacteria like staphylococci (50%), streptococci, *E. coli*, *Clostridium tetani* (occasionally). Proper aseptic technique

(by staff nurses and doctors) and cleaning of umbilical cord with chlorhexidine solution will reduce the umbilical infection.

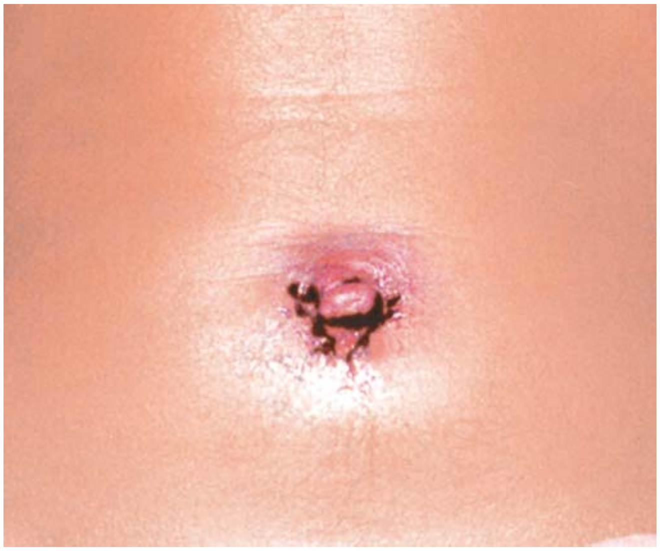

Fig. 17.2: Omphalitis (Umbilical infection).

Causes

- Ligation of the umbilical cord without aseptic precautions.
- Poor asepsis and umbilical hygiene during delivery and after birth.

Bacteria

- Staphylococci, streptococci.
- Gram-negative organisms.
- *Clostridium tetani* causing neonatal tetanus.

Features

- Infection may spread along the hypogastric vessels into the abdominal wall causing abdominal wall abscess.
- Infection may spread into the peritoneum causing peritonitis.
- In severe cases, septicaemia can occur.
- Umbilical granuloma can occur once infection gets localised.
- Purulent discharge with red, swollen umbilicus.

Investigations

- Culture study of discharge.

Treatment

- Antibiotics.
- Drainage of the abscess.
- Aseptic precautions while ligating the cord will prevent the chances of umbilical sepsis.
- Excessive granulation tissue must be removed surgically or by cautery (commonly) using trichloroacetic acid or silver nitrate or copper sulphate in small quantity.

Complications

- Abdominal wall abscess
- Extensive abdominal wall ulceration and skin and subcutaneous gangrene formation
- Septicaemia via umbilical vein
- Peritonitis carries poor prognosis
- Umbilical granuloma, umbilical hernia
- Neonatal tetanus
- Portal vein thrombosis and portal hypertension
- Neonatal jaundice due to cholangitis—6 weeks after childbirth due to spread of infection into liver across umbilical vein

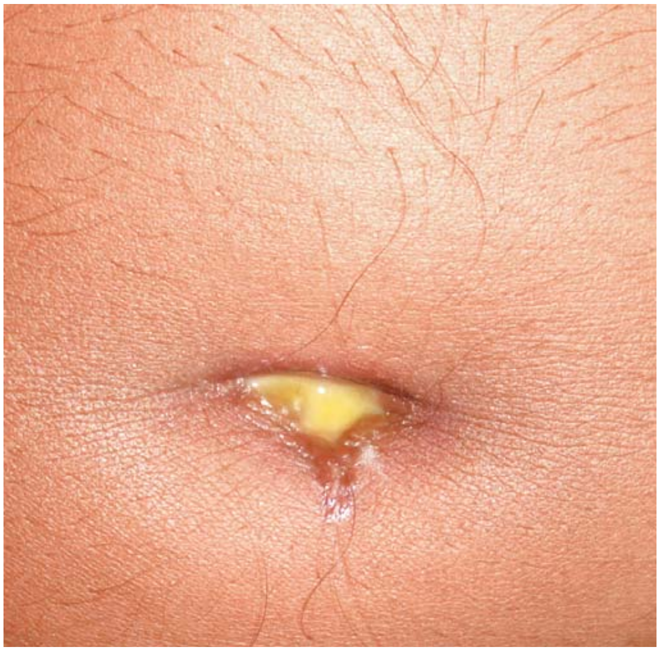

Fig. 17.3: Umbilical sepsis with profuse discharge.

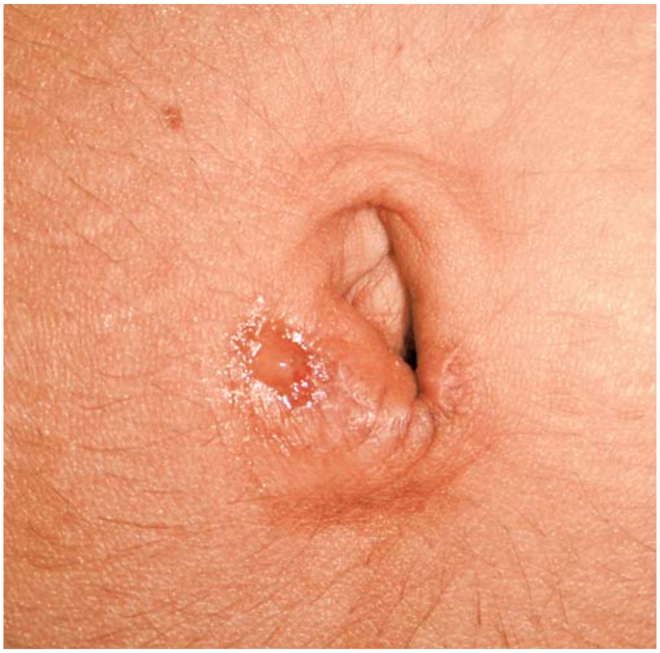

Fig. 17.4: Umbilical infection causing abdominal wall abscess.

Umbilical black eye is Cullen's sign in acute pancreatitis and ruptured ectopic pregnancy.

UMBILICAL GRANULOMA

- It is due to chronic infection of the umbilical cicatrix, causing the granulation tissue to pout, leading to the formation of umbilical granuloma.
- It occurs in any age group, but common in infants and children. Presents as umbilical discharge with tender, red, swelling protruding from the umbilicus which bleeds on touch.
- It has to be differentiated from the anomalies of vitellointestinal duct. It also mimics umbilical adenoma.

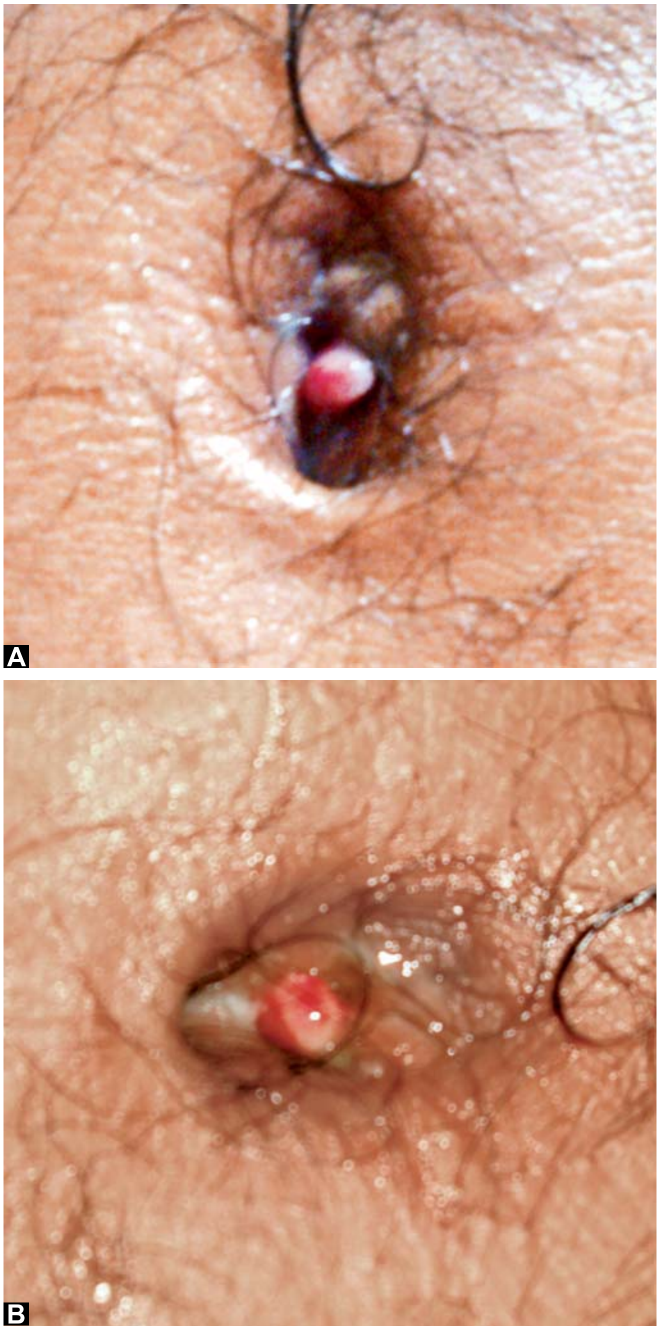

Figs 17.5A and B: Umbilical granuloma due to sepsis.

Treatment

- Antibiotics, application of silver nitrate and dry dressings.
- Occasionally might require excision of granuloma or umbilectomy (rarely).

ANOMALIES OF VITELLOINTESTINAL DUCT

During development of fetus, midgut communicates with yolk sac through vitellointestinal or omphalomesenteric duct. This duct slowly narrows during the development of abdominal wall to lie within the cord, later this communication obliterates to free intestine from yolk sac. Any defect in obliteration causes different anomalies.

Anomalies are:

1. May remain completely patent, forming a *intestinal fistula.*
2. Only a small portion near the umbilicus may remain patent forming discharging *umbilical sinus*. Often the mucosa of this retained portion (epithelial lining) protrudes or evertes to form *umbilical adenoma*.
3. Duct is closed on either side, but the intervening portion may remain as a *intra-abdominal cyst*.
4. Vitellointestinal duct which is obliterated can retain *as a band* which may be a seat for intestinal obstruction, volvulus, internal herniation.
5. Intestinal end may remain patent forming Meckel's diverticulum, which may be attached to umbilicus with a fibrous band. *Meckel's diverticulum* itself can cause diverticulitis, obstruction.

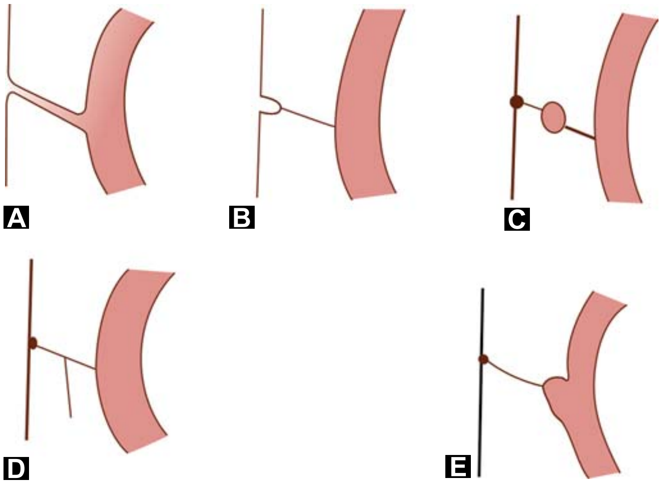

Figs 17.6A to E: Anomalies of vitellointestinal duct. (A) Intestinal fistula, (B) Umbilical sinus, (C) Intra-abdominal cyst, (D) Band, (E) Meckel's diverticulum with band.

Anomalies of vitellointestinal duct

- Intestinal fistula
- Umbilical sinus
- Umbilical adenoma
- Intra-abdominal cyst
- Meckel's diverticulum
- Intra-abdominal band

Investigations

- Fistulogram.
- U/S abdomen.
- Radioisotope study.
- If obstruction is present, plain X-ray abdomen in erect posture is useful to visualise the multiple air-fluid levels.

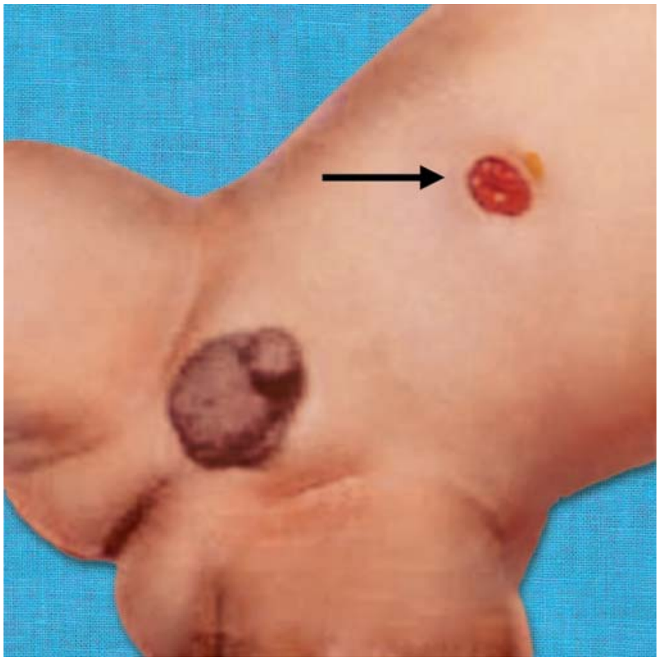

Fig. 17.7: Patent vitellointestinal duct.

Treatment

Excision of vitellointestinal duct is done, along with resection of bowel segment containing Meckel's diverticulum.

UMBILICAL SINUS

Causes

- Persistent vitellointestinal tract partially towards umbilical side.
- Persistent urachus.
- Tuberculosis.
- Umbilical infection.
- *Umbolith*—desquamated epithelium of umbilicus gets collected and inspissated in the umbilical recess causing black/brown coloured stone (umbilical stone). It causes recurrent umbilical infection and sinus.
- Pilonidal sinus of the umbilicus.
- Malignancy in the urachus.

Clinical Features

- Pain, swelling, discharge and tenderness in the umbilicus.
- Features of the specific causes.

Investigations

- Study of the dischargeculture and sensitivity, cytology for malignant cells, AFB study.
- Sinusogram.
- U/S abdomen, CT abdomen.
- Chest X-ray, ESR.

Treatment

- Treat the cause.
- Antibiotics.
- Antitubercular drugs in case of tuberculosis.
- Umbilectomy.

UMBILICAL ADENOMA (Raspberry Tumour)

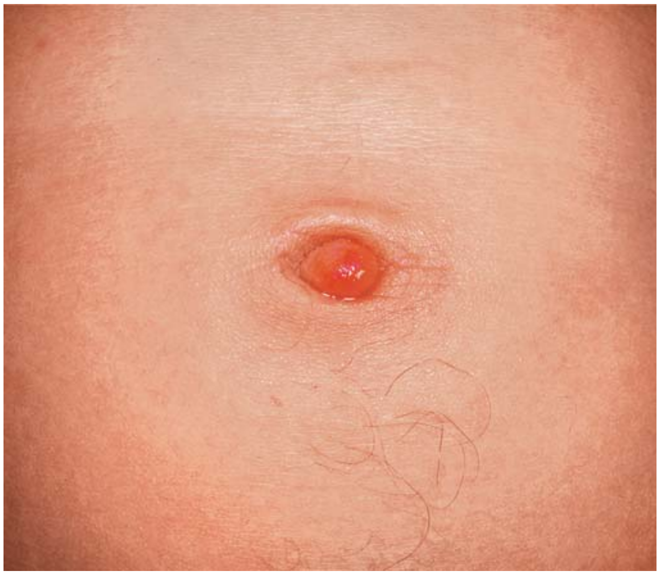

Fig. 17.8: Umbilical adenoma (Raspberry tumour).

- It is commonly seen in infants. It is due to partially obliterated vitellointestinal duct towards umbilical end, causing *prolapse of the mucosa* giving rise to *umbilical adenoma* also called *as Raspberry tumour*.
- It protrudes out as a red swelling, which is moist with mucus and tends to bleed on touch. It often gets infected, discharging pus through the umbilicus.
- Histologically, it consists of columnar epithelium rich in *goblet cells*.

Differential Diagnosis

- Umbilical granuloma.

Treatment

- If the tumour is pedunculated, a firm ligature is tied round it, so that tumour will fall off in few days. If it reappears, umbilectomy is done.
- Sometimes there may be a patent vitellointestinal duct or a Meckel's diverticulum, which may require resection along with the bowel segment.

***Note*:**

Umbilical adenoma *does not disappear after* silver nitrate application.

Tanyol's sign: Umbilicus is equidistant from xiphoid process and symphysis pubis. It is displaced downwards in ascites and upwards in pelvic masses.

UMBILICAL FISTULA

Causes

- Patent vitellointestinal duct.
- Patent urachus.
- Postsurgical.
- Tuberculosis.

Clinical Features

- Faecal discharge or urinary discharge, mucoid discharge.
- Recurrent infection.
- Pain, tenderness and excoriation in and around the umbilicus.

Investigations

- Fistulogram, CT fistulogram.
- Discharge study. C/S, AFB and cytology.
- U/S abdomen.

Treatment

- Fistulectomy, along with resection of bowel segment, patent vitellointestinal tract, and anastomosis of the bowel.
- Fistulectomy and excision of vitelline duct up to the antimesentric surface of the ileum. Opening in the ileum is closed transversely using vicryl or silk sutures.

PATENT URACHUS

Allantoic duct/stalk which is remnant of cranial part of ventral urogenital sinus forms urachus. It gets fibrosed and forms median umbilical ligament.

When urachus is patent it can form:

- *Patent urachus (Urachal fistula)* between umbilicus and dome of the urinary bladder (It is persistent median umbilical ligament).
- *Urachal sinus* if only umbilical side of the urachus is patent.
- *Urachal cyst* if only middle portion of the urachus is patent with lining and fluid content.
- *Urachal diverticulum* when bladder side of the urachus is patent.

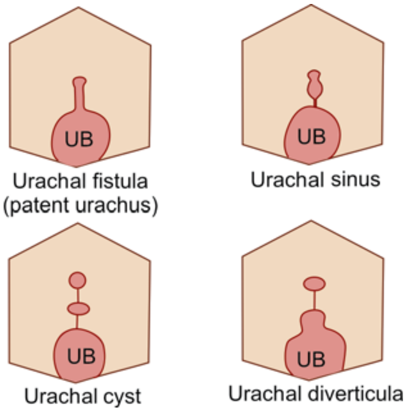

Fig. 17.9: Different types of urachal anomalies.

Features

- Persistent discharge from the umbilicus often stained with urine if it is fistula.
- Recurrent infection and bleeding.
- Pain in the umbilicus and below.
- Recurrent urinary infection.

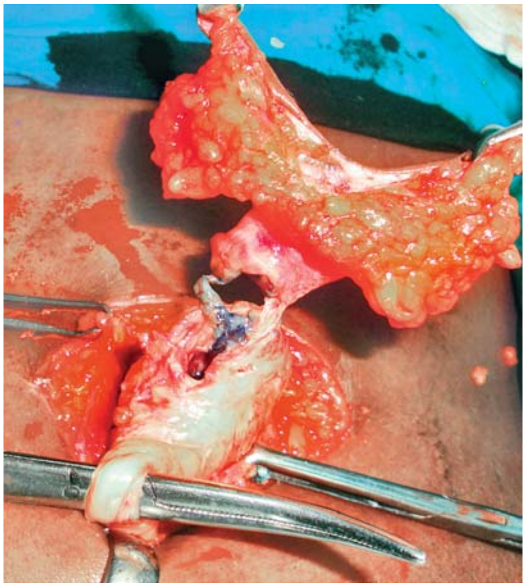

Fig. 17.10: Tuberculosis of urachus which is excised.

Investigations

- Fistulogram to see the extent.
- U/S abdomen.
- Discharge analysis and culture.
- Urine analysis.

Treatment

- Surgical excision of the tract.
- Often umbilical excision may be required.

Note:
Malignancy (adenocarcinoma) or tuberculosis can occur in patent urachus.

BURST ABDOMEN (Abdominal Dehiscence) (ACUTE WOUND FAILURE)

- It is disruption of a laparotomy wound, occurring usually between 5th and 8th postoperative day.
- Usually sutures opposing the deep layers, i.e. peritoneum and rectus sheath tear through, causing burst abdomen—*Acute wound failure*.
- There is postoperative separation of the abdominal musculoaponeurotic layers.
- It is 2% common.
- Other than technical errors, deep wound infection causes localised separation of the wound which in addition to raised intra-abdominal pressure leads to wound dehiscence.

Clinical Features (2%)

- A sudden feeling of giving way from the wound—on 5th to 8th postoperative day often precipitated by bouts of severe cough.

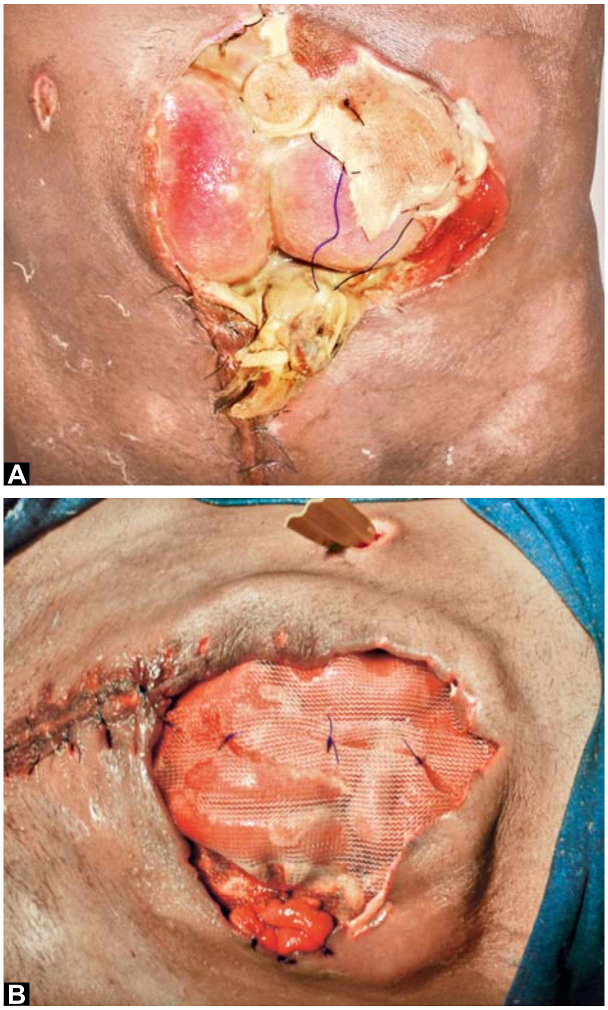

Figs 17.11A and B: Burst abdomen wound exposing intestines outside. It is covered with a mesh later.

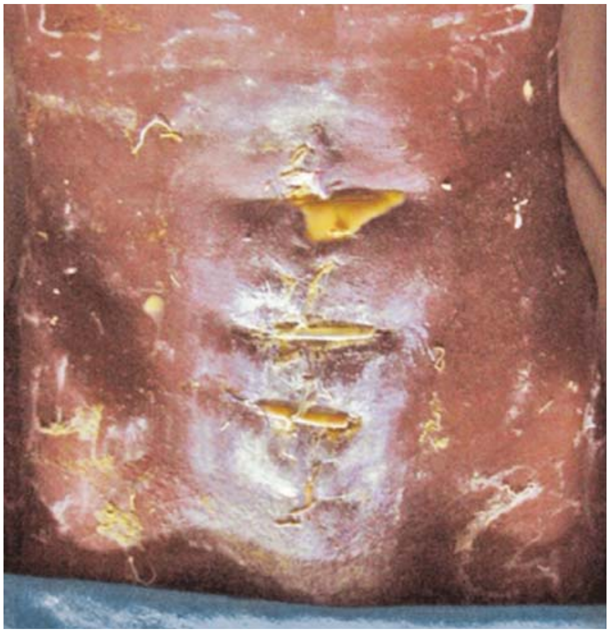

Fig. 17.12: Abdominal faecal fistula. Note the tension sutures and faecal discharge.

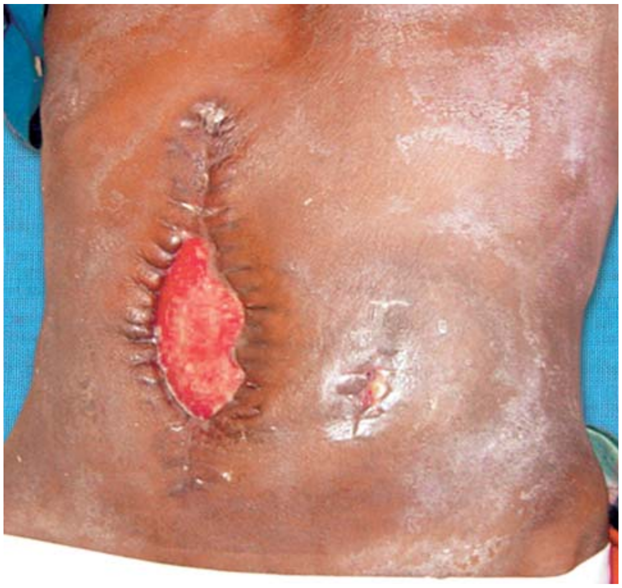

Fig. 17.13: Ulcer over the postoperative wound.

Factors related to burst abdomen

- Choice of suture materials used
- Method of closure; continuous closure is more likely to disrupt than interrupted sutures
- Upper midline and vertical wounds are more likely to disrupt than transverse
- Surgical wounds of peritonitis, acute abdomen, major surgeries like pancreatic, hepatic, gastric, surgeries for malignancies have a high incidence of disruption
- Severe cough, vomiting and distension in postoperative period
- Poor general condition of the patient—anaemia, jaundice, hypoproteinaemia, obesity, uraemia and diabetes mellitus, old age, steroids, radiation, malignancy, chemotherapy.

- *Pinkish serosanguineous discharge* (salmon coloured large quantity of fluid) from the wound.
- Often omentum or coils of intestine are forced out of the wound.
- *Clinically burst abdomen can be diagnosed without fail*.
- Probing of the wound using gloved finger appreciates dehiscence of musculoaponeurotic layer.

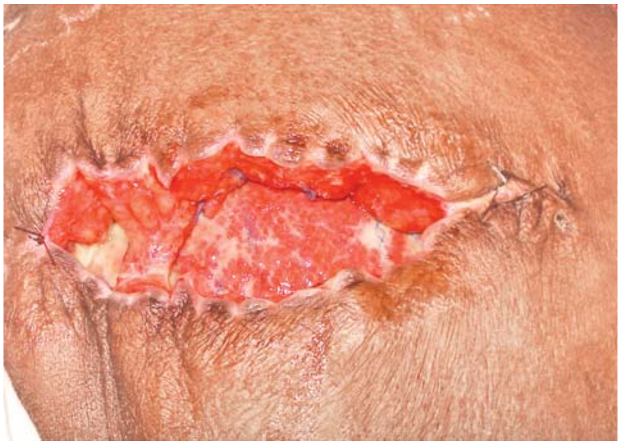

Fig. 17.14: Incisional hernia which was repaired earlier has gaped exposing the mesh.

Truth is what is actually should be. Reality is what is happening. Every reality need not be truthfully right.

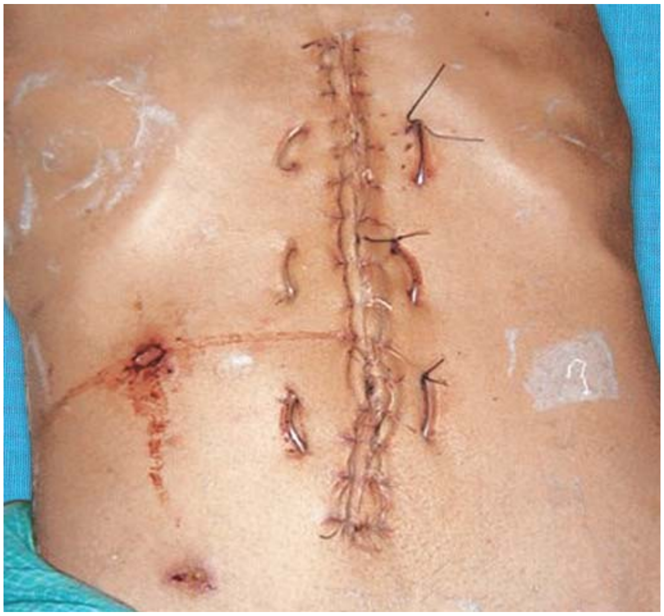

Fig. 17.15: Horizontal tension sutures.

Treatment

- Nasogastric aspiration, IV fluids.
- *Emergency surgery*, i.e. under anaesthesia, wound is opened up properly.
  - Coils of intestines are replaced into the abdominal cavity. Thorough wash is given. Wound is closed by *all layer sutures,* passing a nonabsorbable suture material through the red rubber or plastic collar—*tension sutures* (which is kept for 14 days).
  - *Modified Smead-Jones closure* mainly used to prevent burst abdomen—it is interrupted specialised suture used in the closure of abdomen as single layer excluding the skin. Linea alba is held with Allis' forceps. Number one polyethylene or PDS suture material is used. First bite on one side taken 3 cm away (width) from the margin from outside to inside; it is then passed through the corresponding opposite edge with 3 cm width from inside to outside; later again one small loop of 5 mm width from the edges of each side of the wound from first bite site to second bite site is taken; suture is knotted on the free edge of the first bite side. Full thickness bite holds the suture and maintains the tension in the wound. Smaller loop keeps the linea alba in apposed place. Large curved Ferguson needle is better to place these sutures. Each suture is placed at 2 cm interval. This is the type of suturing used at present in acute abdominal conditions instead of the retention sutures. Here also it is better to place all sutures under proper vision and knotting is done at the end. At least four knots should be placed. Excessive tension should be avoided. In upper abdomen peritoneum need not be included in the bite; but in lower abdomen as linea alba is indistinct, peritoneum is included in this.
- Antibiotics and IV fluids are continued.
- Wound usually heals well without much second dehiscence. Late problem, may be development of incisional hernia.
- Biological dressings, wound vacuum systems are newer modalities used.

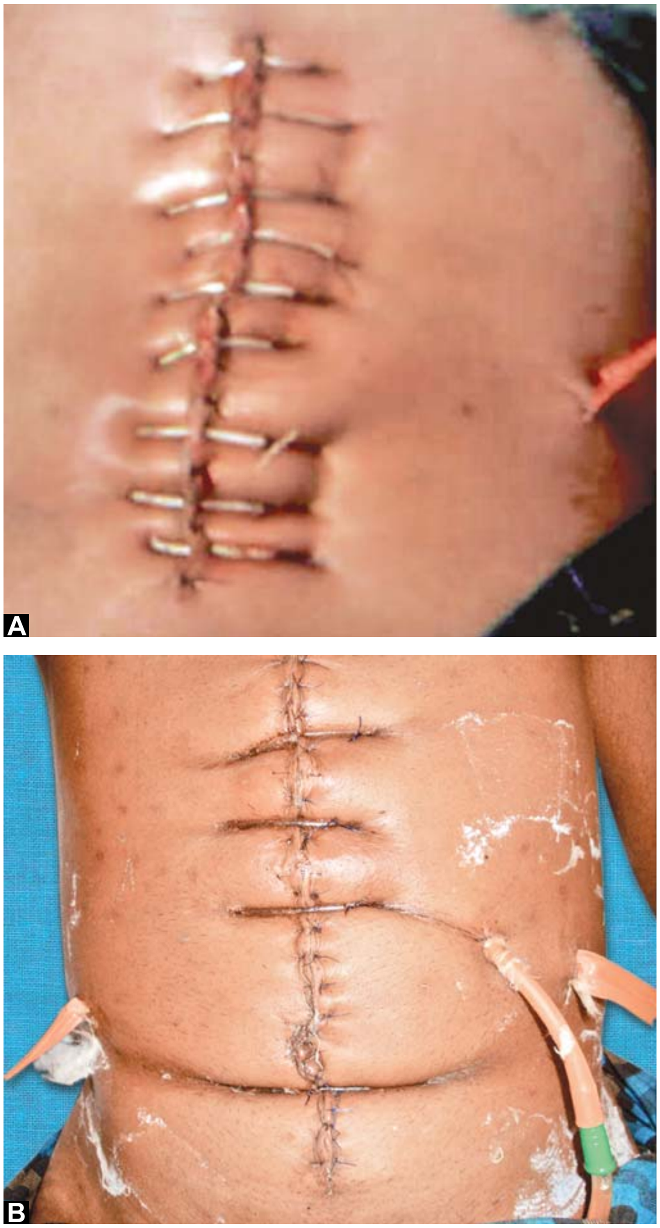

Figs 17.16A and B: (A) Vertical tension sutures. (B) Shows tension sutures with multiple drains.

Newer present methods of managing the burst abdomen

- If the fascial edges are necrotic, *proper debridement* of the wound edges is done. One should not attempt forcible wound closure as it will increase intra-abdominal pressure further and re-dehiscence occurs (When fascia is strong and intact primary closure can be done). Wound and contents are covered with *absorbable mesh or biological prosthesis* like decellularised porcine submucosa and dermis or human cadaveric dermis. They prevent bowel desiccation, bacterial infection.
- *Wound vacuum system* using—open cell foam, semiocclusive drape over the foam and suction apparatus. It provides immediate coverage, minimises heat loss, by negative pressure it clears interstitial fluid, reduces the bowel oedema and contamination, increases wound blood flow and promotes wound healing. Once wound granulates wound is closed with primary closure or with skin graft or flaps depending on the size of the wound.

ABDOMINAL WALL TUMOURS

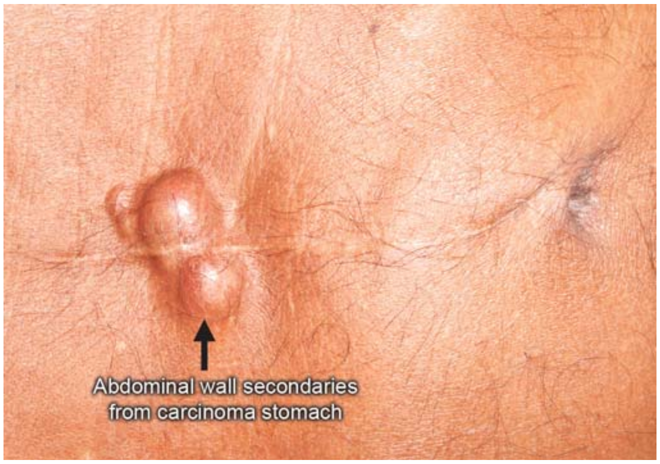

Fig. 17.17: Abdominal wall secondaries from carcinoma stomach.

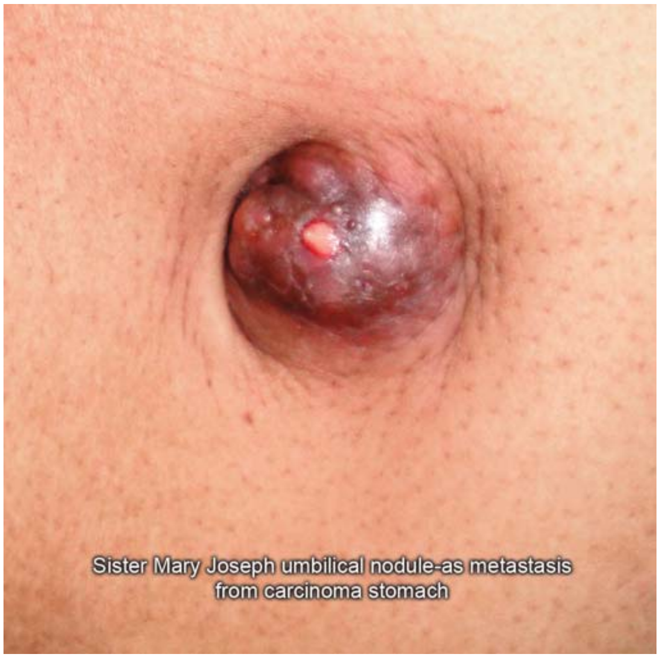

Fig. 17.18: Sister Mary Joseph (Nee Julia Dempsey, Mayo Clinic USA) metastatic umbilical nodule from carcinoma stomach.

- They are not uncommon but often present late as they are usually asymptomatic.
- Common tumours are lipoma, fibroma, neurofibroma, and fibromatosis.
- Malignant tumours occasionally when occurs, are either from skin or soft tissues. They may be desmoid tumour, soft tissue sarcoma like fibrosarcoma, dermatofibrosarcoma, liposarcoma, umbilical secondaries (*Sister Joseph Mary tumour*).
- Presents usually as painless progressive swelling. Often ulceration can occur. Attaining large size is also known. It is dull to percuss. On contracting the abdominal wall muscles, swelling becomes prominent and less mobile.
- Differential diagnoses are—abdominal wall abscess, haematoma, intra-abdominal tumours (adherent to abdominal wall).
- U/S abdomen, CT abdomen is diagnostic. Biopsy is essential.
- *Treatment* is wide excision with adequate clearance with removal of adjacent skin and musculoaponeurotic layer. Defect is covered with a *large mesh*.

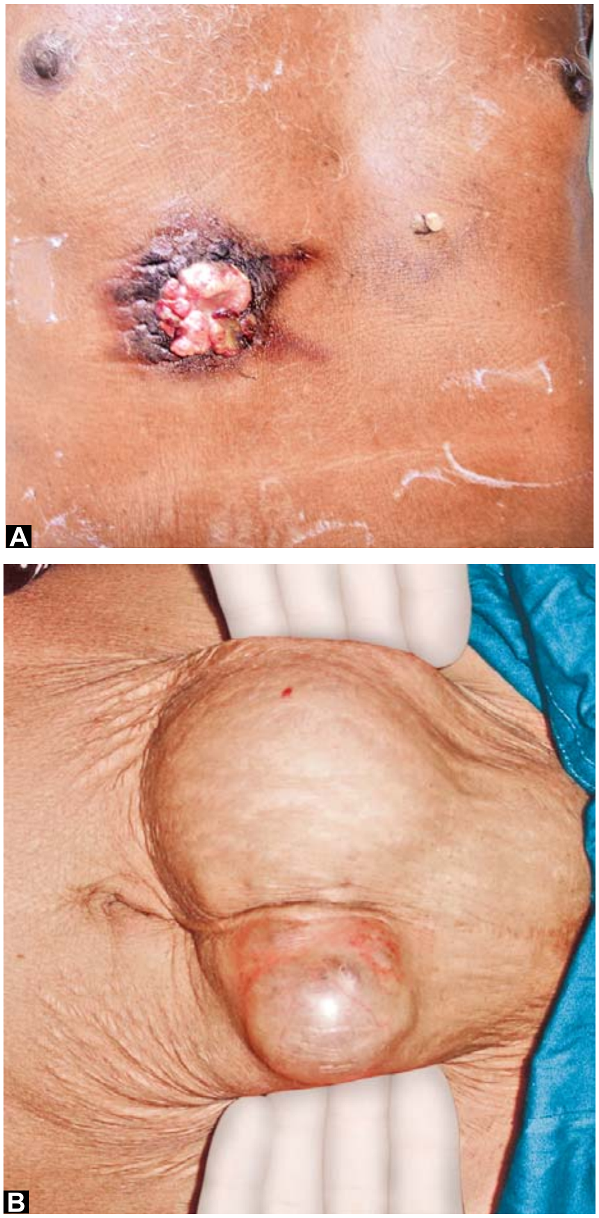

Figs 17.19A and B: Abdominal wall proliferative ulcer (A) and large tumour (B). (B) May be fibromatosis of abdominal wall also.

DESMOID TUMOUR

(*'Desmo' Means Band/Tendon in Greek*)

- It is a tumour arising from the musculoaponeurotic layer of abdomen, *below* the level of the umbilicus.
- It is unencapsulated, hard, fibroma, presently classified under ***aggressive fibromatosis.***

- 80% of cases occur in women, commonly after deliveries. It is common over old abdominal operation scars (lower abdomen), may be due to old haematomas. Oral contraceptive (OCP) use is more commonly associated with desmoid.
- It is often associated with the familial polyposis colon (FAP), osteomas, odontomes epidermal cysts—***Gardner's syndrome.*** Desmoid tumour is 1000 times more common in FAP.
- It is a slow growing tumour involving muscle and soft tissue of the abdominal wall, locally spreading, often undergoes myxomatous changes.
- Unlike fibromas, it never turns into sarcoma.
- It is often classified as superficial and deep. It is also classified as abdominal (common in females); extra-abdominal (common in males, common in back, head, scars, chest wall, neck).

Histologically it contains multinucleated plasmodial giant cells.

Gardner's syndrome

- Familial polyposis colon
- Osteomas
- Odontomes
- Epidermal cyst—20%

Management

- The possible association of Gardner's syndrome is looked for and Barium enema, X-ray, U/S abdomen are done. *MRI is useful.* Biopsy is done often to confirm the diagnosis.
- *Wide excision* of the tumour with a margin of 2.5 cm is done along with *placement of a mesh to* the abdominal defect.
- It is moderately radiosensitive.
- Drugs like sulindac and tamoxifen are also used.
- Chemotherapeutic agents like doxorubicin, actinomycin D, dacarbazine, carboplatin are used but with significant toxicity.
- Recurrence rate is high—20-40%.

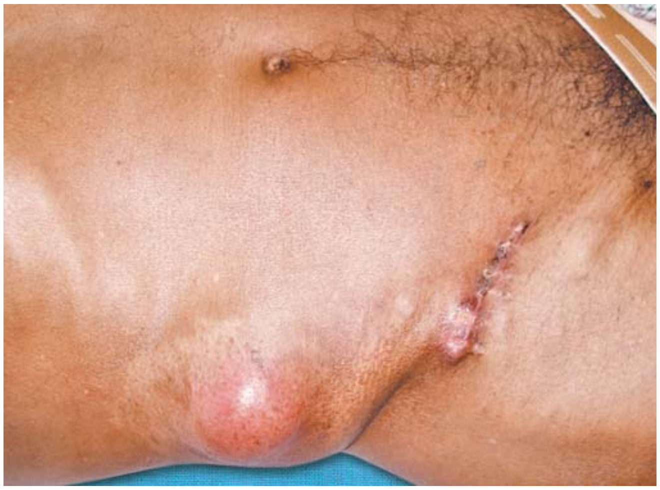

Fig. 17.20: Recurrent abdominal wall tumour. It may be fibromatosis or recurrent dermatofibroma.

EXOMPHALOS (Omphalocele)

- It is the failure of all or a part of the gut to return to the coelomic cavity during early foetal life, as coelomic cavity has not developed properly.

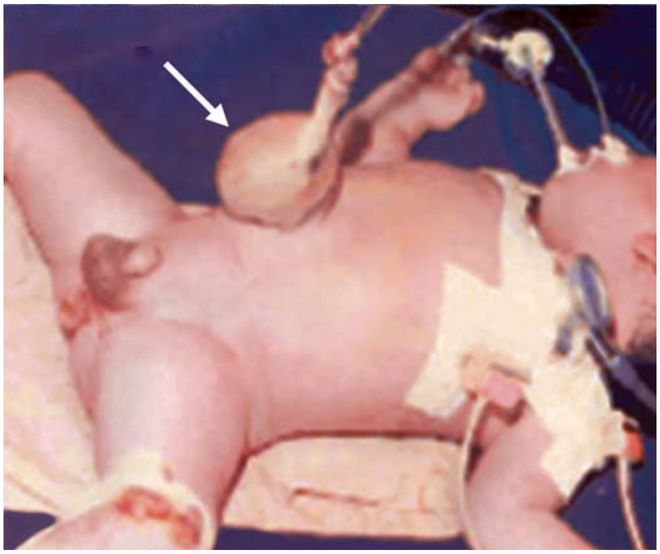

Fig. 17.21: Exomphalos.

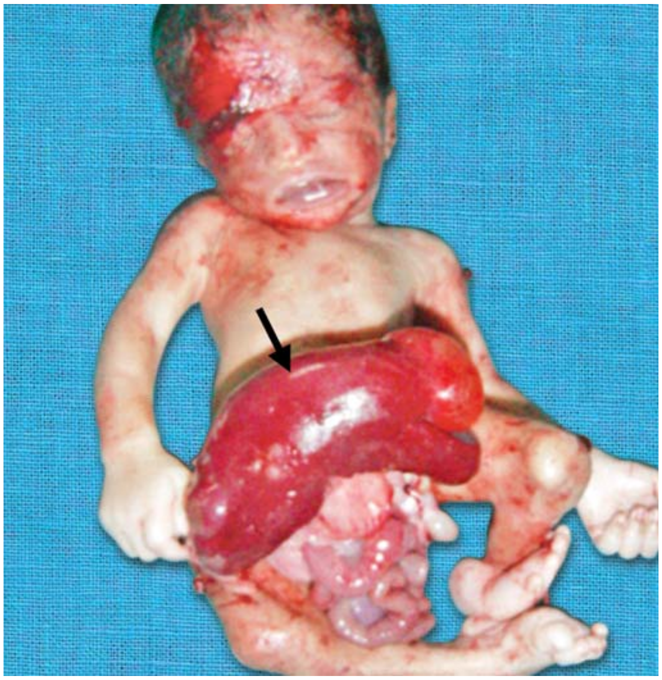

Fig. 17.22: Exomphalos major. Note the liver, small and large bowel.

- Sac covering the content is very thin, consists of three layers—outer amniotic membrane, middle Wharton's jelly and inner peritoneal layer. Sac may get ruptured during birth.
- *Omphalocele is often associated with congenital anomalies of cardiac and genitourinary system—70%.*
- Vitellointestinal anomaly, diaphragmatic hernia, malrotation can co-exist, exomphalos, macroglossia, gigantism (EMG) can co-exist in 2/3 of such babies—is called as *Beckwith-Weidman syndrome.*

- Chromosomal trisomies—13, 15, 18 and 21, bladder extrophy, imperforate anus, sacral vertebral anomaly, meningomyelocele.
- Abdominal wall defect can be confirmed *in utero* by U/S. Amniocentesis and chromosomal analysis can be done. This will allow deciding the possible need for termination of pregnancy, or management plan immediately after delivery.
- Abdominal wall muscle is normally developed but peritoneal layer is hypoplastic.

Types

Two types:

Differences between exomphalos minor and exomphalos major

Exomphalos minor	*Exomphalos major*
Small sac	Large sac
Defect < 5 cm	Defect > 5 cm
Content is small bowel only	Liver, small and large bowel
Primary closure is possible	Primary closure is not possible
Good prognosis	Poor prognosis

Exomphalos minor:

- Here the sac is small and *umbilical cord is attached to the summit*, with small bowel as the content. Treatment for E. minor is relatively easier. The sac has to be twisted so as to reduce the content to the peritoneal cavity through the umbilical defect and the abdomen is strapped. Later the defect is closed (Defect is < 5 cm).

Exomphalos major:

- A large defect (> 5 cm) is present with contents lying completely outside. *Umbilical cord is attached to the inferior aspect of the sac.* Contents being small bowel, large bowel and liver.
- Often the sac ruptures during delivery, which in turn leads to severe infection and high mortality. Here immediate surgery (within hours) is the only hope to save the child.

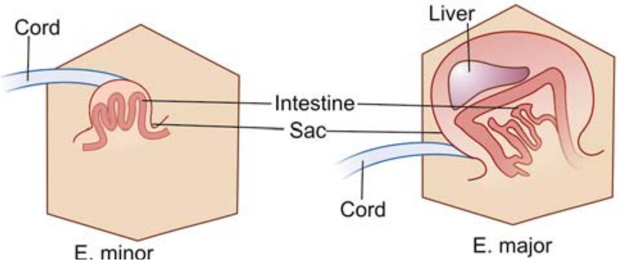

Fig. 17.23: Diagram showing differences between exomphalos minor and exomphalos major.

- As there is hardly any properly developed abdominal cavity, it is not possible to reduce the contents to the peritoneal cavity. So initially, relaxing incisions are placed over the lateral abdominal wall and also the subcutaneous layer has to be undermined so as to accommodate whatever possible contents in the cavity. The contents are not to be forced into the cavity, which in turn may cause intestinal obstruction, respiratory distress, venous engorgement.
- Often a sterile polythene bag *(Silastic silo)* can be wrapped over the content carefully. Gradual twisting of the bag over the summit at regular intervals over two to three weeks will stimulate the peritoneal cavity to increase in capacity and eventually the contents get reduced. Later the defect is closed properly.
- Antibiotics, IV fluids, nutritional support are required during this period.
- *Mortality is higher* in spite of all these—due to sepsis, dehydration, hypothermia and respiratory failure.
- Late ventral hernia should be corrected at later period.

Management of exomphalos

- Vitamin K injection, TPN
- Sepsis control with antibiotics
- Evaluation for other anomalies
- In E. major when sac is intact application of 0.5% mercurochrome with 65% alcohol to promote granulation tissue formation
- Prevention of aspiration
- Prevention of hypothermia
- Wrapping the content with sterile bag/wrapper
- Correction of dehydration
- Definitive surgical procedures like release incisions and closure

GASTROSCHISIS (Belly Cleft)

- It is a defect of the anterior abdominal wall just lateral to the umbilicus. It is common in premature babies.
- It is associated with defect in the involution of 2nd umbilical vein. It is common in mothers *younger than 20 years*, those who take, aspirin, ibuprofen, pseudoephedrine during 1st trimester and who regularly smoke and take alcohol.

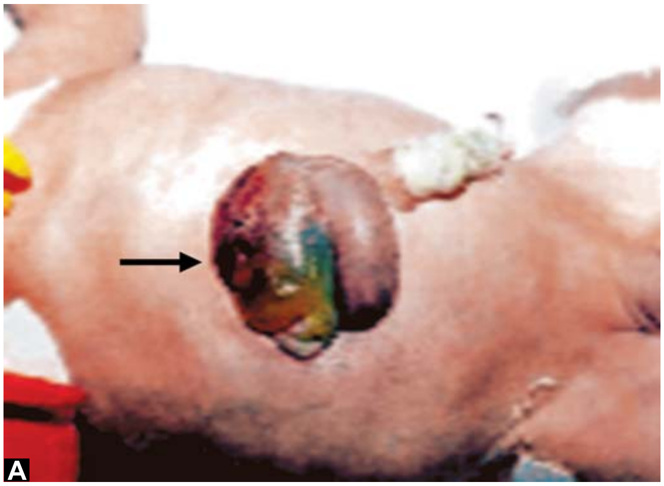

Fig. 17.24A

A deaf husband and a blind wife are always a happy couple.

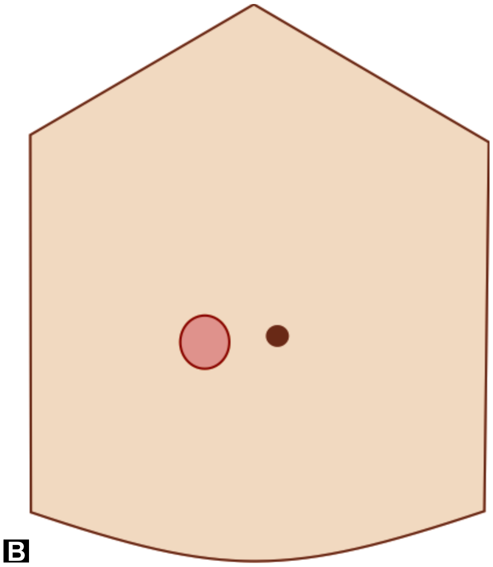

Fig. 17.24B

Figs 17.24A and B: Gastroschisis.

- Umbilicus is normal. The defect is almost always to the right of an intact umbilical cord. Evisceration of the bowel occurs through the defect during intrauterine life. There is no peritoneal sac and the irritating effect of amniotic fluid causes chemical peritonitis with formation of a thick, oedematous membrane.
- *Nonrotation and intestinal atresia are common associations (15%).* Cardiac anomaly is not common as in omphalocele.
- After delivery, these infants are more prone for fluid loss, hypothermia, hypovolaemia, sepsis, metabolic acidosis.
- Necrotising enterocolitis is also common in such infants (20%). They are also more prone for paralytic ileus.

Treatment

General:
- Fluid management.
- TPN.
- Antibiotics.
- R-T aspiration.
- Calorie supplement.

Specific:
- Later under G/A, intestines are pushed into the abdomen through the defect and defect is closed with interrupted nonabsorbable sutures. Often when bowel is not accommodating in the abdominal cavity, bowel is initially placed in sterile silastic silo bag. In later period, it is pushed into the abdomen gradually. Often a part of the bowel may not be viable, then resection and anastomosis has to be done.
- With proper surgery, nutrition, resuscitation, survival rate is 90% which is better than omphalocele.
- *Prolonged postoperative ileus* is the common problem in these patients.

Exomphalos	*Gastroschisis*
1. Defect through umbilicus	1. Defect lateral to umbilicus
2. Covered by sac which has three layers	2. No peritoneal sac
3. Associated with anomalies	3. Not associated (except malrotation)

RECTUS SHEATH HAEMATOMA

- Rectus abdominis muscle is supplied by superior and inferior epigastric arteries. Injury to one of these vessels will cause bleeding and haematoma in rectus sheath.
- Commonly it is *due to bleeding from inferior epigastric artery* in the lower abdomen.

Causes

- Trauma.
- Surgery.
- Spontaneous haematoma, typhoid fever.
- Blood dyscrasias, haemophila, anticoagulant therapy.
- Severe straining and exercises.
- Tetanus and other convulsions.
- Patients on anticoagulants.
- Puerperium.

Features

- Common in females.
- Sudden onset of swelling in lower abdomen, which is tender, warm, firm on one side of the abdomen. Swelling does not cross the midline.
- Bluish discoloration over the swelling.
- U/S and aspiration confirms the diagnosis.
- Should be differentiated from other masses and parietal hernias.
- Coagulation profile is a must.
- CT scan abdomen is often needed.

Treatment

- Usually conservative with analgesics and antibiotics.
- Angiographic *embolisation* of inferior epigastric artery.
- Occasionally requires drainage of haematoma and ligation of inferior epigastric artery.

ABDOMINAL WALL ABSCESS

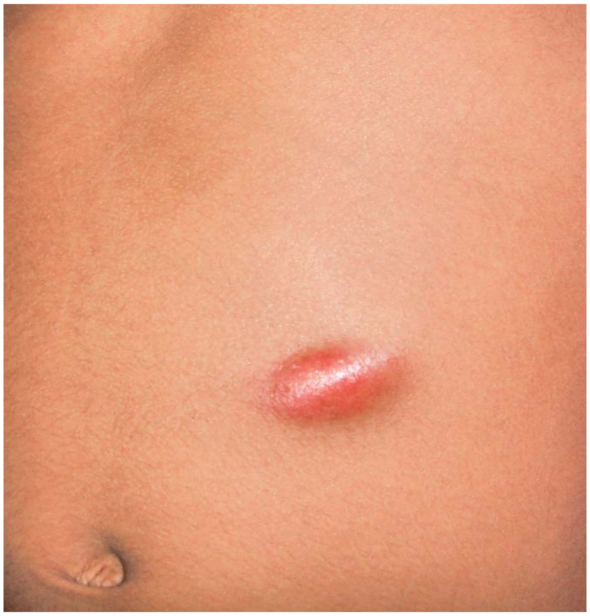

Fig. 17.25: Abdominal wall abscess in a child.

Causes

- Infected haematoma.
- Umbilical sepsis spreading into the abdominal layers causing the abscess.
- Blood spread from distant focus.

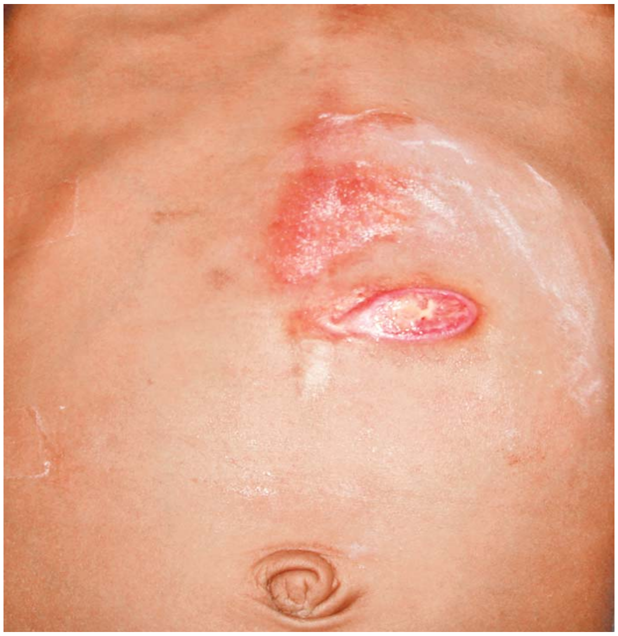

Fig. 17.26: Abdominal wall infection with a wound.

Features

- Tender, soft/firm swelling which is well-localised, adherent to skin and abdominal muscles underneath and not mobile.
- Aspiration will show pus.
- Should be ruled out from intra-abdominal mass, cold abscess, parietal hernia.
- U/S and needle aspiration is confirmative.

Treatment

It is antibiotics and drainage under general anaesthesia.

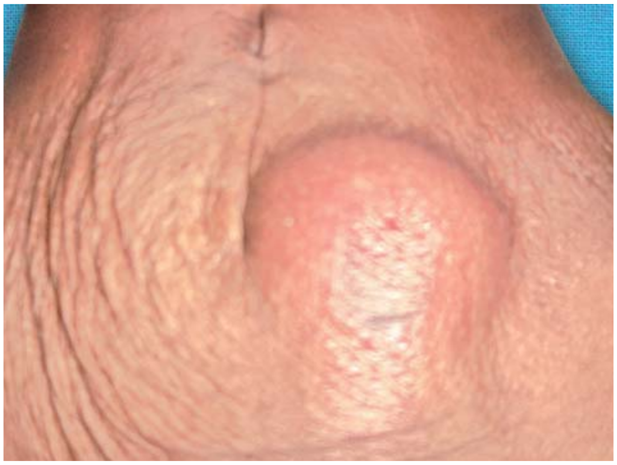

Fig. 17.27: Abdominal wall abscess.

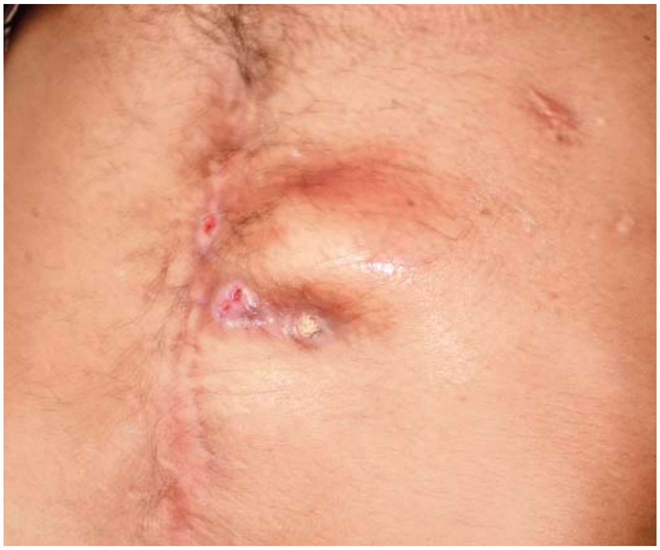

Fig. 17.28: Postoperative abdominal wall abscess.

MELENEY'S PROGRESSIVE SYNERGISTIC BACTERIAL GANGRENE OF ABDOMINAL WALL

- It is due to infection by microaerophilic streptococci, staphylococci and other anaerobes of the postoperative abdominal or thoracic wounds.
- It is common in HIV, diabetic and immunosuppressed people.

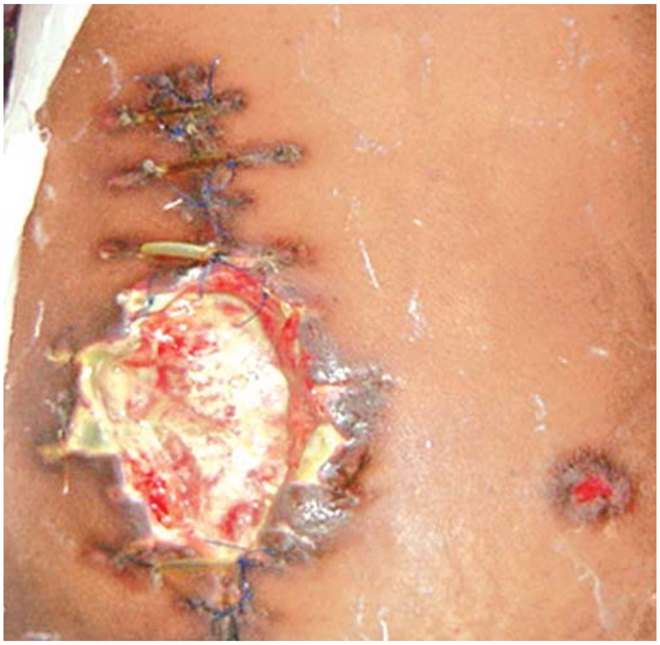

Fig. 17.29: Severe abdominal wall sepsis in a postoperative patient.

- Sudden pain, redness, blackening and gangrene of the skin of the abdomen with abdominal wall necrosis.
- Toxicity, septicaemia, renal failure can occur.

Treatment

- Antibiotics like penicillins and metronidazole.
- Excision of necrotic and gangrenous tissue until it bleeds.
- Blood transfusion, nutrition supplement.
- Maintaining adequate urine output.
- Management of toxaemia, hyperbaric oxygen and critical care.
- Skin grafting and coverage, when it granulates well.

DIVERICATION OF RECTI (DIASTASIS RECTI)

- It is thinning of linea alba in midline in epigastrium.
- Abdominal wall protrudes in midline with prominent divaricated edges of both recti.
- Transversalis fascia remains intact and hence hernia will not be present; so impulse on coughing will be absent.
- Diastasis will become prominent on lifting the head from the bed.
- It does not require any treatment.

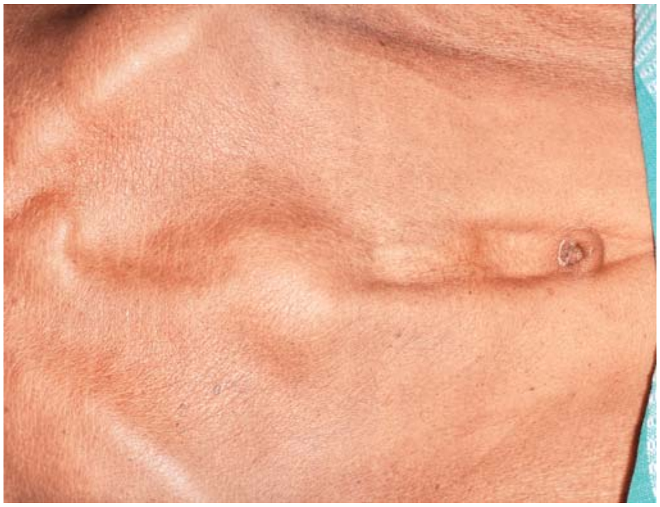

Fig. 17.30: Diverication of the recti with visible mass in epigastrium.

Chapter 18 Hernia

CHAPTER OUTLINE

- Hernia
- Aetiology
- Parts of Hernia
- Classification of Hernia
- Inguinal Hernia
  - Surgical Anatomy of Inguinal Canal
  - Classification of Inguinal Hernia
  - Indirect (Oblique) Inguinal Hernia
  - Newer Classification of Hernia
  - Direct Hernia
  - Recurrent Hernia
  - Hernioplasty
  - Malgaigne Bulging
  - Incarcerated Hernia
- Strangulated Hernia
- Sliding Hernia
- Pantaloon Hernia
- Infantile Hernia
- Femoral Hernia
- Ventral Hernia
- Incisional Hernia
- Umbilical Hernia
- Paraumbilical Hernia
- Epigastric Hernia
- Spigelian Hernia
- Obturator Hernia
- Richter's Hernia
- Lumbar Hernia
- Phantom Hernia
- Sciatic Hernia
- Complications of Hernia Surgery
- Parastomal Hernia
- Different Types of Hernia

HERNIA

Hernia means—*'To bud' or 'to protrude', 'off shoot' (Greek) 'rupture' (Latin).*

- *A hernia is defined* as an area of weakness or disruption of the fibromuscular tissues of the body wall. Often hernia is also defined as an actual anatomical weakness or defect. 75-85% of abdominal wall hernias are groin hernias. 15% of males and 5% of females will develop groin hernia. Presently all hernias in groin are grouped as groin hernias. But in this chapter discussion of the indirect/direct/femoral hernias are given, in detail as it is still practiced and followed in most of the centers and still it is important when surgical technical aspects are considered.
- Hernia is defined as an abnormal protrusion of a viscous or a part of a viscous through an opening, artificial or natural with a sac, covering it.
- *Inguinal hernia is the most common hernia (73%)* because the muscular anatomy in the inguinal region is weak and also due to the presence of natural weakness like deep ring and cord structures.
- Femoral is 17%; umbilical is 8.5%; others are 1.5% (Excluding incisional hernia).
- In general, incisional hernia is next to inguinal hernia in occurrence.

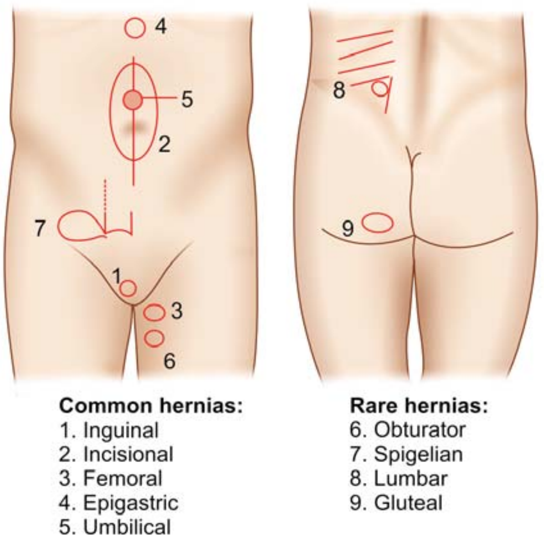

Fig. 18.1: Common and rare sites of hernias.

AETIOLOGY

- Straining.
- Lifting of heavy weight.

In order to achieve a radical cure of inguinal hernia it is absolutely essential to restore those conditions in the area of hernial orifice which exist under normal circumstances.

—Edoardo Bassini

- Chronic cough (tuberculosis, chronic bronchitis, bronchial asthma, emphysema).
- Chronic constipation (habitual, rectal stricture).
- Urinary causes
  - Old age—BPH, carcinoma prostate.
  - Young age—stricture urethra.
  - Very young age—phimosis, meatal stenosis.
- Obesity.
- Pregnancy and pelvic anatomy (especially in femoral hernia in females).
- Smoking.
- Ascites.
- Appendicectomy through McBurney's incision may injure the ilioinguinal nerve causing right sided *direct* inguinal hernia.

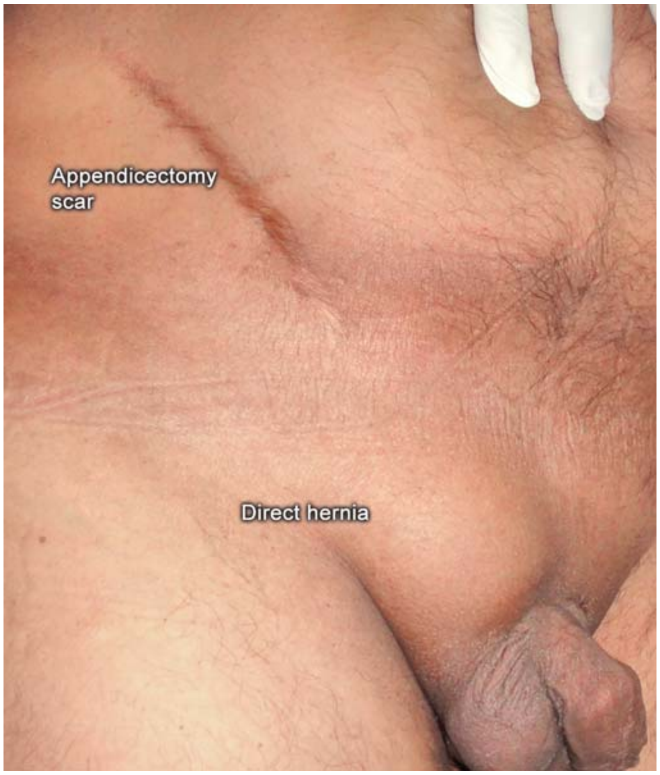

Fig.18.2: Old appendicectomy scar with direct inguinal hernia. It is due to injury to ilioinguinal nerve during appendicectomy.

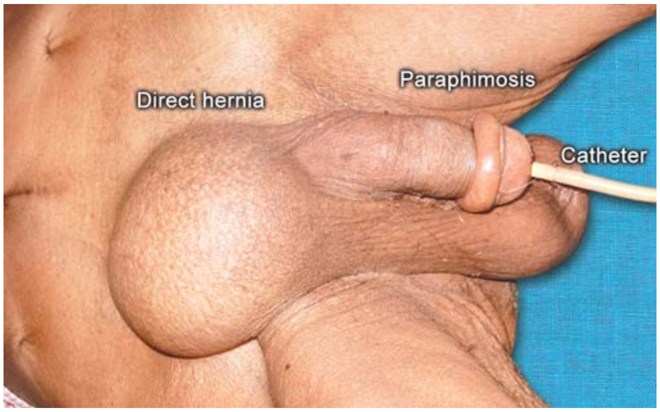

Fig. 18.3: Inguinal hernia in a patient who is having benign prostatic hyperplasia (BPH) with Foley's urinary catheter. He is having paraphimosis probably for not replacing back the retracted prepuce after catheterisation.

- An indirect inguinal hernia occurs in a congenital, preformed sac, i.e. the remains of processus vaginalis. Chances of presence of bilateral preformed sac is 60%.
- Familial collagen disorder—*Prune Belly syndrome*.
- Acquired herniation is also probably due to collagen deficiency called as *metastatic emphysema of Read.*

PARTS OF HERNIA

Hernia comprises of:

- *Covering.*
- *Sac.*
- *Content.*

Sac is a diverticulum of peritoneum with mouth, neck, body and fundus. Neck is *narrow in indirect sac but wide in direct sac*. Body of the sac is thin in infants, children and in indirect sac but is thick in direct and long-standing hernia.

  - Hernia without neck: Those hernias with larger mouth lack neck, e.g. direct hernia, incisional hernia.
  - Hernia without sac: Epigastric hernia—it is protrusion of extraperitoneal pad of fat.

Coverings of the sac are the layers of the abdominal wall through which the sac passes.

Contents of Sac

  - Omentum—*Omentocele* (Epiplocele). Difficult to reduce the sac later, initially it can be reduced easily.
  - Intestine—*Enterocele*— commonly small bowel, but sometimes even large bowel.
  - Difficult to reduce the sac initially.
  - *Richter's hernia*: A portion of circumference of bowel is the content.
  - Urinary bladder may be the content or part of the posterior wall of the sac—*cystocele*.
  - Ovary, often with fallopian tube.
  - Meckel's diverticulum—*Littre's hernia*.

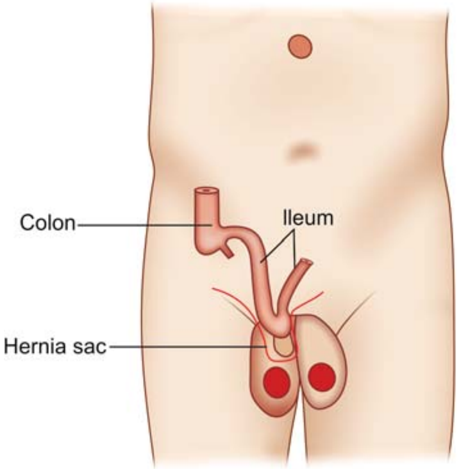

Fig. 18.4: *Littre's hernia* with Meckel's diverticulum as content.

  - *Fluid:* Fluid is secreted from congested bowel or omentum. It may be an infected fluid or ascitic fluid or blood from the strangulated sac.

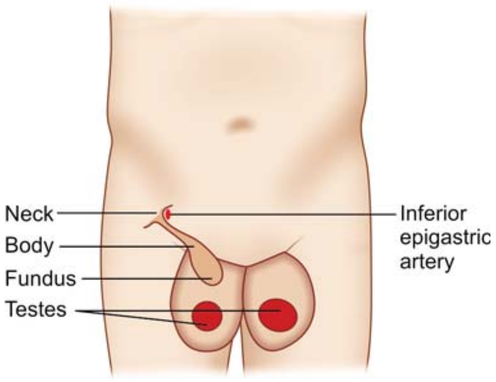

Fig. 18.5: Parts of hernia. Note the fundus, body and neck of the hernia.

CLASSIFICATION OF HERNIA

Classification I (Clinical)

1. Reducible Hernia

Hernia gets reduced on its own or by the patient or by the surgeon. Intestine reduces with gurgling and it is difficult to reduce the first portion. Omentum is doughy, and it is difficult to reduce the last portion. Expansile impulse on coughing present.

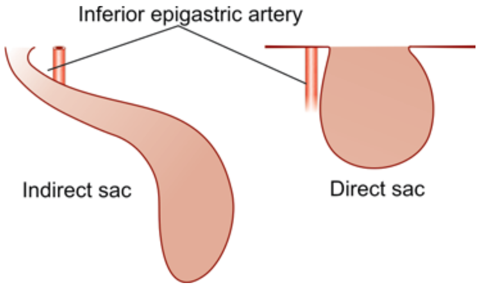

Fig. 18.6: Diagram showing the differences between indirect and direct sacs.

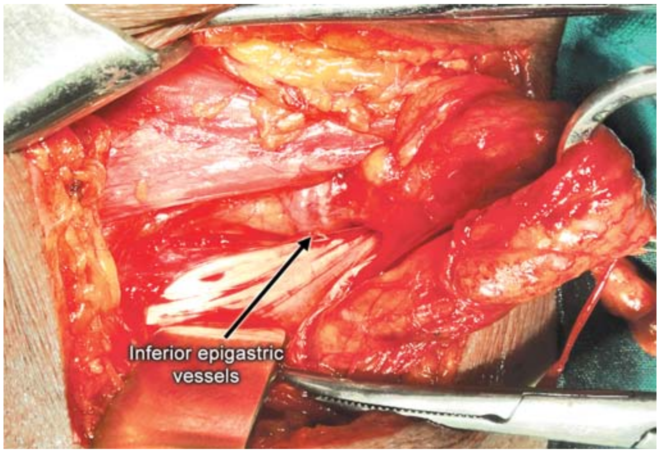

Fig. 18.7: Inferior epigastric artery on table. It is very important landmark during surgery whether open or laparoscopic approach.

2. Irreducible Hernia

Here contents cannot be returned to the abdomen due to narrow neck, adhesions, overcrowding. Irreducibility predisposes to strangulation.

3. Obstructed Hernia

It is an irreducible hernia with obstruction, but blood supply to the bowel is not interfered. It eventually leads to strangulation.

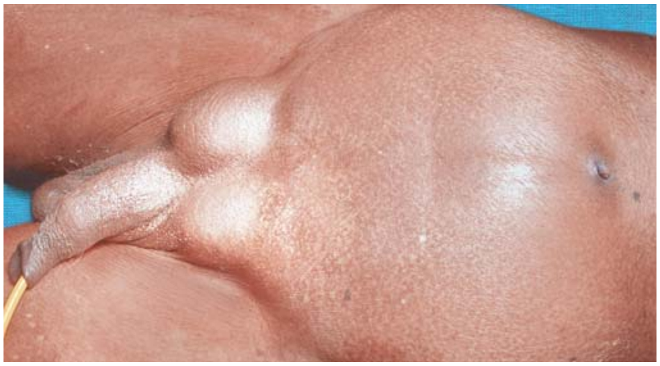

Fig. 18.8: Inguinal hernia presenting with features of intestinal obstruction.

Garrey's stricture: Constriction that occurs due to ischaemic narrowing of small bowel which has reduced from an obstructed hernia.

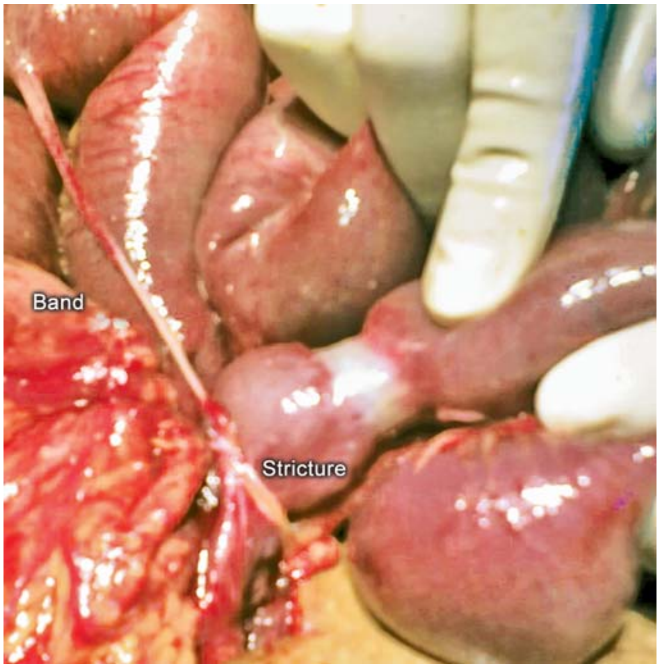

Fig. 18.9: Garrey's stricture on table. Due to ischaemia, a stricture is formed in the bowel at the constriction ring after reduction.

4. Inflamed Hernia

It is due to inflammation of the contents of the sac, e.g. appendicitis, salpingitis. Here hernia is tender but not tense; overlying skin is red and oedematous.

5. *Strangulated Hernia*

It is an irreversible hernia with obstruction to blood flow. The swelling is tense, tender, with no impulse on coughing and with features of intestinal obstruction.

Features of intestinal obstruction may be absent in case of omentocele, Richter's hernia, Littre's hernia.

Classification II

Congenital—Common

It occurs in a preformed sac/defect. Clinically may present at a later period due to any of the precipitating causes like in indirect inguinal hernia.

Acquired

It is secondary to any causes which raise the intra-abdominal pressure leading into weakening of the area like in direct inguinal hernia.

Classification III: According to the Contents

- Omentocele—omentum.
- Enterocele—intestine.
- Cystocele—urinary bladder.
- Littre's hernia—Meckel's diverticulum.
  Note: Littre described Meckel's diverticulum in a hernial sac 81 years before Meckel was born.
- Maydl's hernia.
- Sliding hernia.
- Richter's hernia—part of the bowel wall.

Classification IV: Based on Sites

- Inguinal hernia—occurring in inguinal canal.
- Femoral hernia—occurring in femoral canal.
- Obturator hernia.
- Diaphragmatic hernia.
- Lumbar hernia.
- Spigelian hernia.
- Umbilical hernia.
- Epigastric hernia.

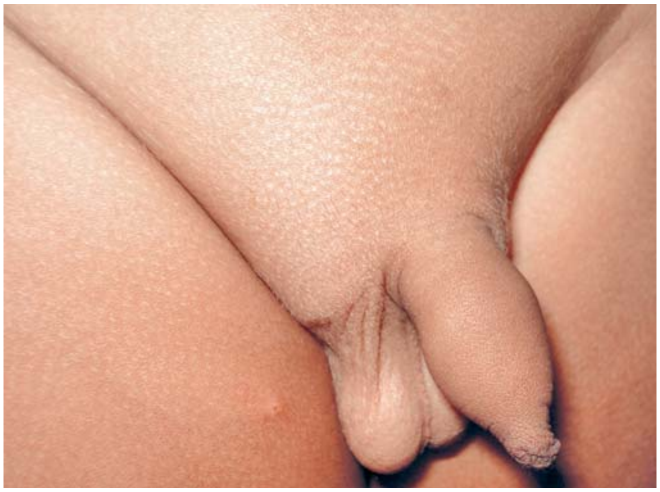

Fig. 18.10: Inguinal hernia on right side in a child. It needs only herniotomy.

Herniography (By Gullmo)

- Injection of contrast material into the peritoneal cavity and taking films in supine and prone positions to diagnose small protrusions of peritoneal sac is called as herniography.
- It was earlier also used for diagnosing undescended testis.

INGUINAL HERNIA

SURGICAL ANATOMY OF INGUINAL CANAL

Superficial inguinal ring is a triangular opening in the external oblique aponeurosis and is 1.25 cm above the pubic tubercle. The ring is bounded by a superomedial and inferolateral crus. Normally the ring does not admit the tip of little finger.

Deep inguinal ring is a U-shaped condensation of the transversalis fascia, lies 1.25 cm above the inguinal ligament midway between the symphysis pubis and the anterosuperior iliac spine.

***Inguinal (Poupart's) ligament*:** It is formed by the lower border of the external oblique aponeurosis which is thickened and folded backwards on itself, extending from anterosuperior iliac spine to pubic tubercle.

Inguinal canal: It is an oblique passage in lower part of abdominal wall, 4 cm long, situated above the medial ½ of inguinal ligament, extending from deep inguinal ring to superficial inguinal ring.

In infants both superficial and deep rings are *superimposed* without any obliquity of the inguinal canal.

Inguinal canal in female is called as **'canal of Nuck.'**

Contents of inguinal canal

- Spermatic cord in males
- Round ligament in females
- Ilioinguinal nerve

Contents of spermatic cord

- Vas deferens
- Artery to vas
- Testicular and cremasteric artery
- Genital branch of genitofemoral nerve
- Pampiniform plexus of veins
- Remains of processus vaginalis
- Sympathetic plexus around the artery to vas

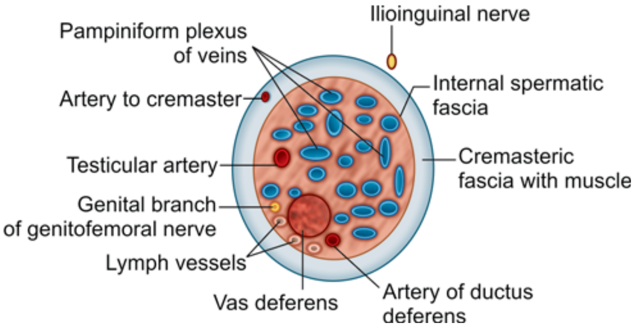
Pampiniform plexus of veins
Ilioinguinal nerve
Artery to cremaster
Internal spermatic fascia
Testicular artery
Cremasteric fascia with muscle
Genital branch of genitofemoral nerve
Lymph vessels
Vas deferens
Artery of ductus deferens

Fig. 18.11: Structures related to cord in the inguinal canal.

Coverings of Spermatic Cord

- Internal spermatic fascia from fascia transversalis
- Cremasteric fascia.
- External spermatic fascia from external oblique aponeurosis is seen below the external ring in the scrotum.

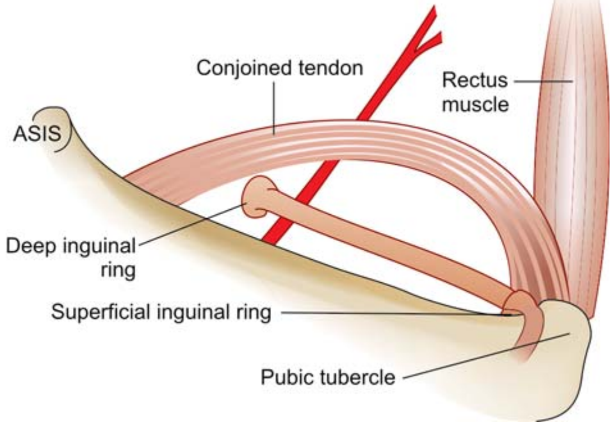
Conjoined tendon
Rectus muscle
ASIS
Deep inguinal ring
Superficial inguinal ring
Pubic tubercle

Fig. 18.12: Surgical anatomy of inguinal canal.

Boundaries

In front: External oblique aponeurosis and conjoined muscle laterally.

Behind: Inferior epigastric artery, fascia transversalis and conjoined tendon medially.

Above: Conjoined muscle (Arched fibres of internal oblique).

Below: Inguinal ligament.

Defence mechanism of inguinal canal

- Obliquity of inguinal canal
- Arching of conjoint tendon
- '*Shutter mechanism'* of internal oblique
- '*Ball valve mechanism'* due to contraction of cremaster muscle which plugs to superficial ring
- When external oblique muscle contracts, intercrural fibres of superficial ring appose causing '*slit valve mechanism'*
- Hormones

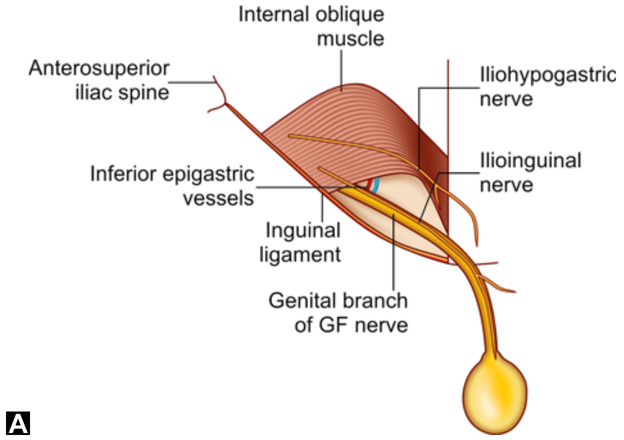
Internal oblique muscle
Anterosuperior iliac spine
Iliohypogastric nerve
Inferior epigastric vessels
Ilioinguinal nerve
Inguinal ligament
Genital branch of GF nerve
A

Fig. 18.13A

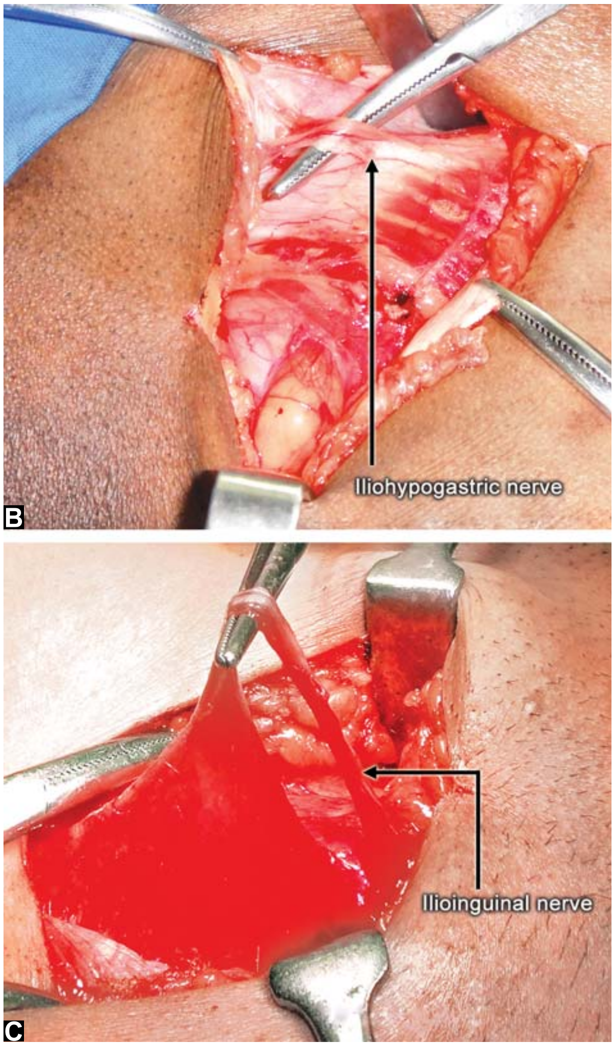
Iliohypogastric nerve
B
Ilioinguinal nerve
C

Figs 18.13B and C

Figs 18.13A to C: Nerves in inguinal canal: *Iliohypogastric nerve* (T_{12}, L_1) runs between transversus abdominis and internal oblique divides into lateral and anterior branches; anterior branch pierces internal oblique 2 cm medial to anterosuperior iliac spine and later pierces the external oblique 3 cm above the superficial inguinal ring to supply abdominal skin above the pubis. *Ilioinguinal nerve* (L_1) pierces the transversus abdominis near anterosuperior iliac spine, pierces the internal oblique just above the internal ring, and enters the inguinal canal within cremasteric fascia outside the cord supplies medial thigh, base of penis and proximal scrotum. *Genital branch of genitofemoral nerve* (L_1, L_2) enters the cord through internal ring supplying the cremaster and scrotum.

Fruchaud's Myopectineal Orifice

It is an osseo-myo-aponeurotic tunnel. *It is through this tunnel all groin hernias occur.*

It is bounded*:*

- Medially by lateral border of rectus sheath.
- Above by the arched fibres of internal oblique and transversus abdominis muscle.
- Laterally by the iliopsoas muscle.
- Below by the pectin pubis and fascia covering it.

> *When spermatic cord is rolled transversely beneath the gentle pressure of index finger. Thickening of the cord denotes presence of a hernia.*
> **— William E Ladd**

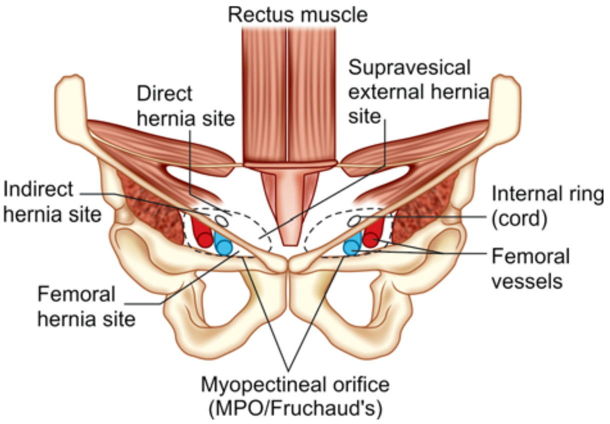
Rectus muscle
Direct hernia site
Supravesical external hernia site
Indirect hernia site
Internal ring (cord)
Femoral vessels
Femoral hernia site
Myopectineal orifice (MPO/Fruchaud's)

Fig. 18.14: *Fruchaud's myopectineal orifice* is bound—by lateral border of rectus, iliopsoas, conjoint tendon, pectin pubis.

CLASSIFICATION OF INGUINAL HERNIA (EARLIER)

Classification I

Anatomical Classification (in Inguinal Hernia)

Indirect hernia

It comes out through internal ring along with the cord. It is lateral to the inferior epigastric artery.

Direct hernia

It occurs through the posterior wall of the inguinal canal through 'Hesselbach's triangle' (bounded medially by lateral border of rectus muscle, laterally by inferior epigastric artery, below by inguinal ligament). Sac is medial to the inferior epigastric artery.

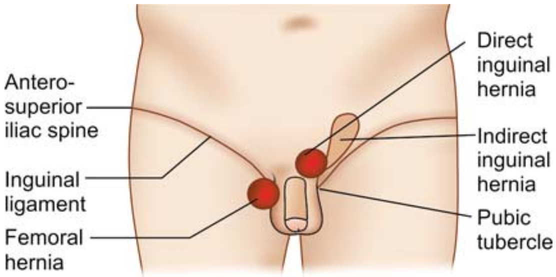
Antero-superior iliac spine
Inguinal ligament
Femoral hernia
Direct inguinal hernia
Indirect inguinal hernia
Pubic tubercle

Fig. 18.15: Location of direct and indirect inguinal hernia and femoral hernia.

Classification II

According to the Extent

- *Incomplete*:
  - **Bubonocele:** Here sac is confined to the inguinal canal.
  - **Funicular:** Here sac crosses the superficial inguinal ring, but does not reach the bottom of the scrotum.
- *Complete:* Here sac descends to the bottom of the scrotum.

Saddle-bag or pantaloon hernial sac has got both medial and lateral component.

Note:
Inguinal hernia is above and medial to the pubic tubercle. Femoral hernia is below and lateral to pubic tubercle.

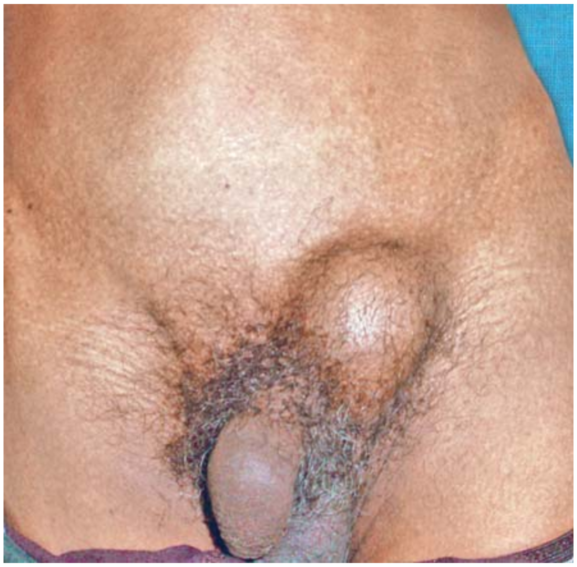

Fig. 18.16: Bubonocele.

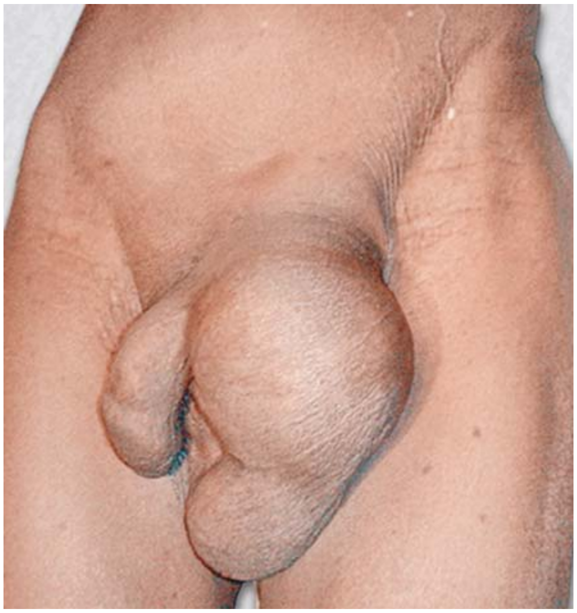

Fig. 18.17: Irreducible left side inguinal hernia. It is above the upper part of the testis (funicular). Taxis is the method used to reduce it.

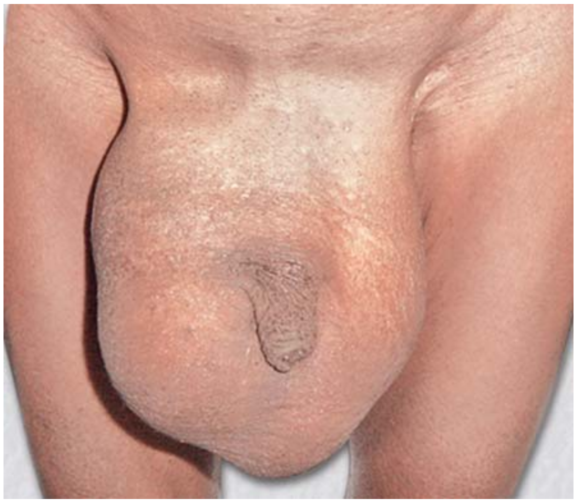

Fig. 18.18: Bilateral complete inguinal hernia. Here hernia descends up to the bottom of the scrotum.

NEWER CLASSIFICATION OF HERNIA

I. Gilbert Classification

***Type I*:** Hernia has got snug internal ring through which a peritoneal sac passes out as indirect sac.

Type II: Hernia has a moderately enlarged internal ring which admits one finger but lesser than two finger breadth. Once reduced it protrude during coughing or straining.

Type III: Hernia has got large internal ring with defect more than two finger breadth. Hernia descends into the scrotum or with sliding hernia. Once reduced it immediately protrudes out without any straining.

Type IV: It is direct hernia with large full blow out of the posterior wall of the inguinal canal. The internal ring is intact.

Type V: It is a direct hernia protruding out through punched out hole/defect in the transversalis fascia. The internal ring is intact.

Type VI: Pantaloon/double hernia.

Type VII: Femoral hernia.

Type VI and VII are Robbin's modifications.

II. NYHUS Classification

Type I: Indirect hernia with normal deep ring.
Type II: Indirect hernia with dilated deep ring.
Type III: Posterior wall defect.
  a. Direct.
  b. Pantaloon hernia.
  c. Femoral hernia.
Type IV: Recurrent hernia.

III. BENDAVID Classification [Type, Staging, Diameter (TSD) Classification]

Type I: Anterolateral defect (indirect).
Type II: Anteromedial (direct).
Type III: Posteromedial (femoral).
Type IV: Posterior prevascular hernia.
Type V: Anteroposterior defect: Inguino-femoral hernia.

IV. Classification—Casten's Staging

Stage 1: An indirect hernia with a normal internal ring.
Stage 2: An indirect hernia with enlarged internal ring.
Stage 3: All direct or femoral hernias.

V. Halverson and McVay Classification

Class 1: Small indirect hernia.
Class 2: Medium indirect sac.
Class 3: Large indirect hernia or direct hernia.
Class 4: Femoral hernia.

VI. Ponka's Classification

1. Indirect inguinal hernia:
    - Uncomplicated indirect hernia.
    - Sliding indirect hernia.
2. Direct hernia:
    - Small defect on the medial aspect of the Hesselbach's triangle.
    - Diverticular hernia in the posterior wall with an otherwise intact inguinal floor.
    - A large diffuse direct inguinal hernia of the entire floor of Hesselbach's triangle.

INDIRECT (OBLIQUE) INGUINAL HERNIA

- This is the most common type of hernia (65%).
- It is more common in younger age group as compared to direct inguinal hernia which is more common in elderly.
- It is more common on right side in 1st decade but in 2nd decade the incidence is equal on both sides.
- Hernia is bilateral in 30% of cases.
- Sac is thin in indirect type. Neck is narrow and lies lateral to inferior epigastric vessels.

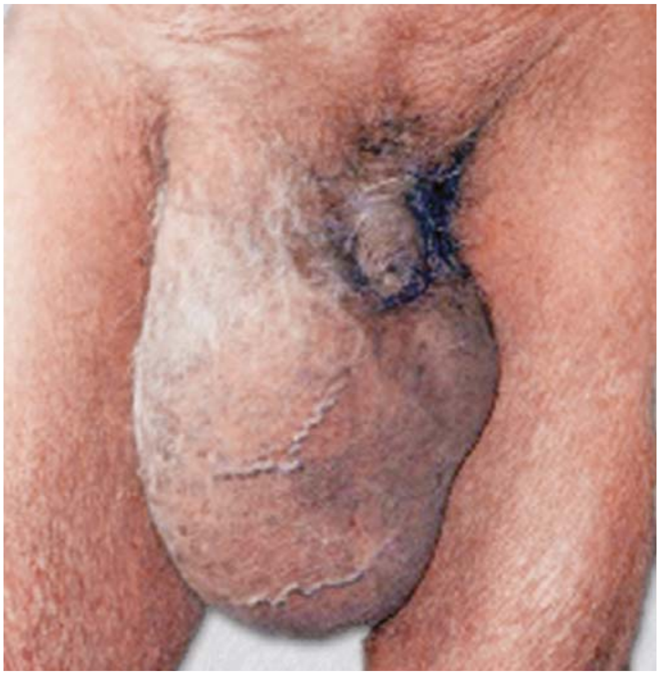

Fig. 18.19: Complete indirect hernia.

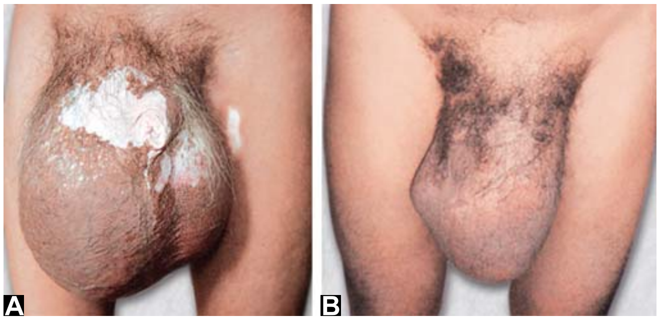

Figs 18.20A and B: Giant inguinal hernias in different sides. (A) Right side and (B) Left side.

Coverings of indirect hernia (from inside out)

- Extraperitoneal tissue
- Internal spermatic fascia
- Cremasteric fascia
- External spermatic fascia
- Skin

Precipitating causes for inguinal hernia

- Smoking
- Obesity
- Respiratory causes like bronchial asthma, tuberculosis, bronchitis
- Ascites
- Previous surgery like appendicectomy which can cause direct inguinal hernia
- Chronic constipation due to anorectal strictures. Rectal stricture may be due to chronic proctitis (amoebic), tuberculosis of anorectum, previous anorectal surgery, rectal carcinoma or stricture due to lymphogranuloma venereum
- Urinary problems like benign prostatic hyperplasia (BPH), urethral stricture
- Straining
- Multiple pregnancies

Types

Three types

1. *Bubonocele*: Where the hernia is limited to inguinal canal.

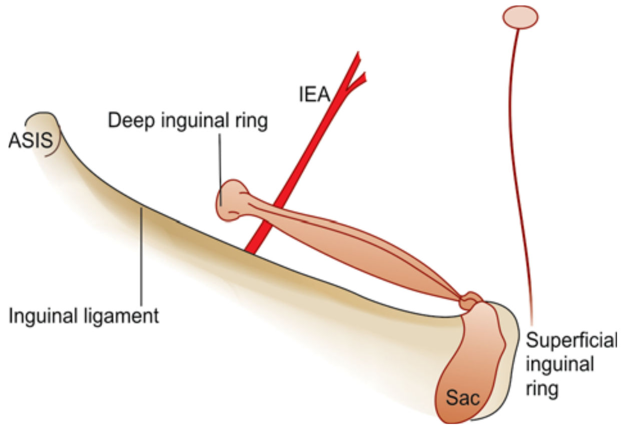

Fig. 18.21: Surgical anatomy of indirect hernia.

2. *Funicular:* Processus vaginalis is closed just above the epididymis. Contents of the sac can be felt separately from testis, which lies below the hernia.
3. *Complete (Scrotal*): Testis appears to lie within the lower part of hernia. It can occur in any age group. It occurs in a congenital preformed sac (processus vaginalis). More commonly contents descend into the pre-existing sac, only when there are precipitating causes which force the content down.

Expansile impulse on coughing

- Hernia
- Laryngocele
- Meningocele
- Empyema necessitans
- Swellings with intracavitary communications like thorax, cranium, and abdomen like dermoid

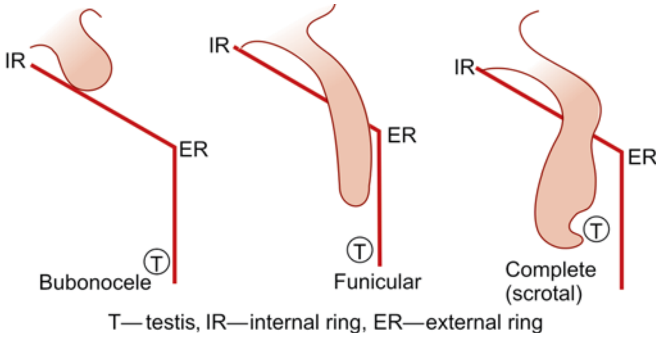

Fig. 18.22: Types of indirect inguinal hernia.

Clinical Features

- Prevalence of inguinal hernia is 25% in males; 2% in females.
- It is more common in males (20 : 1 :: Male : Female).
- Patient presents with dragging pain and swelling in the groin which is better seen while coughing and standing; and felt together with an expansile impulse.
- In *complete* type, the content descends down to the scrotum completely.
- Contents are either small bowel, large bowel, omentum or combination of all these. In females, sometimes ovary and tubes may be the content. In infants, swelling appears when the child cries and is often translucent.
- It is usually reducible, but can go for irreducibility, inflammation, obstruction, strangulation.
- ***Internal ring occlusion test***: Internal ring is located half inch above the mid-inguinal point (center point between anterosuperior iliac spine and pubic symphysis). After reducing the contents, in lying down position, internal ring is occluded using the thumb. Patient is asked to cough. If a swelling appears medial to the thumb, then it is a direct hernia. If swelling does not appear and on releasing the thumb swelling appears during coughing, then it is an indirect hernia confirmed in standing position.
- ***Ring invagination test***: After reduction of hernia, the little finger/index finger of the examiner is invaginated from the bottom of the scrotum, gradually pushed up and rotated to enter the superficial inguinal ring. The impulse on coughing is felt at the tip of the invaginated finger. This test is done only in males.
- ***Zieman's test (Fig. 18.23)***: The examiner places his index finger on the deep inguinal ring and middle finger on the superficial inguinal ring, ring finger over saphenous opening. The patient is asked to cough or to hold the nose and blow. If the impulse is felt on the index finger, it is indirect hernia.
- ***Head*** or ***leg rising test (Fig. 18.24)*** is done to look for abdominal wall muscle tone and Malgaigne bulgings.

  Valsalva manoeuver is also used to check the tone of abdominal wall muscles.
- Abdominal, respiratory, urological examination is done to look for any precipitating factors like chronic bronchitis, ascites, stricture urethra, BPH.
- ***Per rectal*** examination is a must.
- ***Inguinal hernia in females:*** Increased thickness of labium majus on palpation, when compared to contralateral side.

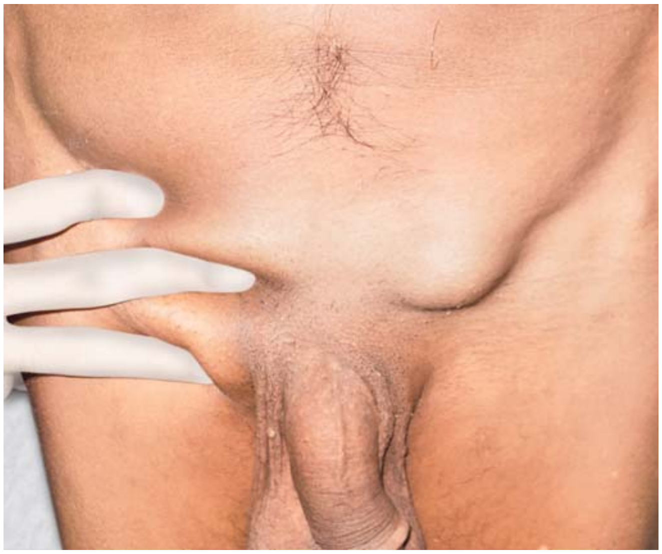

Fig. 18.23: *Zieman's test*. Index finger on deep ring; middle finger on superficial ring and ring finger over saphenous opening—are placed after reducing the content. Patient is asked to cough and impulse is felt in finger corresponding to the existing hernia.

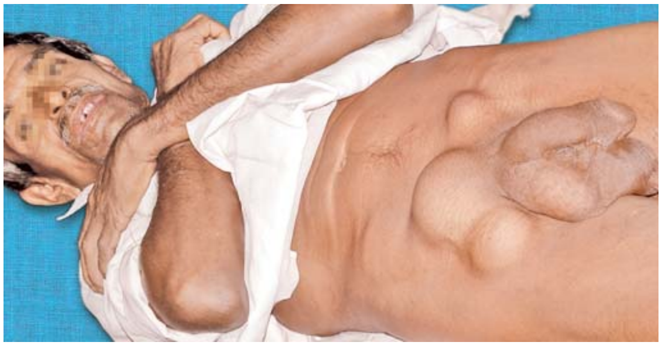

Fig. 18.24: Head or leg rising or Valsalva manoeuvre is important to elicit abdominal muscle tone.

- ***Silk glove sign:*** Index finger is invaginated across scrotum towards the external ring. When patient coughs, inguinal hernia is felt as a slit like sensation.
- Palpation of bulbar urethra for stricture (thickening/crescent like feel/button like depression).

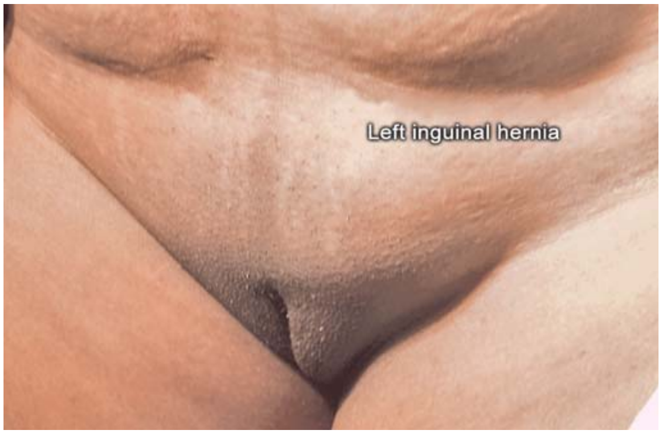
Left inguinal hernia

Fig. 18.25: Left side inguinal hernia in a female.

Use five fingers of the hand to complete all tests for hernia

- Thumb—for deep ring occlusion test
- Index, middle and ring fingers for Zieman's test
- Little finger for superficial ring invagination test

Rules of hernia examination

- Never forget to examine opposite side
- Never forget to do per rectal examination
- Never forget to examine urethra
- Never forget to check abdominal muscle tone

Differential diagnosis

- Hydrocele—infantile/encysted/large vaginal/bilocular
- Undescended testis
- Femoral hernia
- Lipoma of the cord
- Hydrocele of the canal of nuck (in females)
- Inguinal lymph node enlargement
- Groin abscess

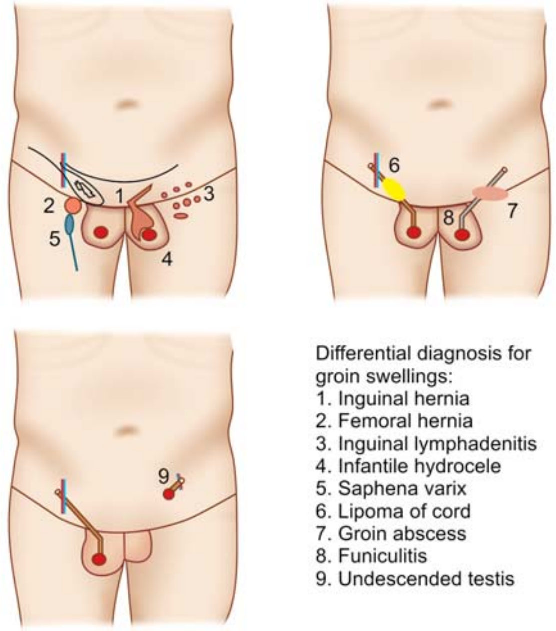
1
2
3
4
5
6
7
8
9
Differential diagnosis for groin swellings:
1. Inguinal hernia
2. Femoral hernia
3. Inguinal lymphadenitis
4. Infantile hydrocele
5. Saphena varix
6. Lipoma of cord
7. Groin abscess
8. Funiculitis
9. Undescended testis

Fig. 18.26: Differential diagnosis for groin swellings.

Indirect inguinal hernia	*Direct inguinal hernia*
Can occur in any age from childhood to adult	Common in elderly
Occurs in a pre-existing sac	Always acquired
Protrusion through the deep ring; herniation occurs later	Herniation through posterior wall of the inguinal canal
Pyriform/oval in shape; descends obliquely and downwards	Globular/round in shape; descends directly forward bulge

Hydrocele of canal of Nuck is the most common differential diagnosis for inguinal hernia in females.

Can become complete by descending down into the scrotum	Descent down into the scrotum is rare
Neck of the sac is narrow and lateral to inferior epigastric artery	Neck of the sac is wide and medial to inferior epigastric artery
Sac is anterolateral to the cord	Sac is posterior to the cord
Ring occlusion test does not show any impulse after occluding the deep ring	Test shows impulse even after occluding the deep ring
Invagination test shows impulse on the tip of the little finger	Impulse is felt over the pulp of the little finger
Zieman's test shows impulse on the index finger	Test shows impulse on the middle finger
Commonly unilateral but can be bilateral	Commonly bilateral
Obstruction/strangulation are common	Rare but can occur
Sac should be opened during surgery	Sac is not necessarily opened unless obstruction is present

In enterocele	***In omentocele (epiploecele)***
First part is difficult to reduce	First part is easier to reduce
but last part is easier. There	but last part is difficult. Has
will be gurgling sound on reduction	a doughy feeling
Resonant on percussion	Dull on percussion
Peristalsis is seen	No peristalsis seen
Bowel sounds may be heard	Bowel sounds not heard

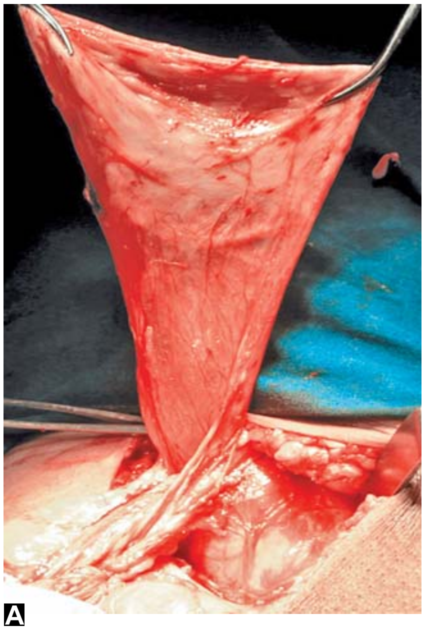

Fig. 18.27A

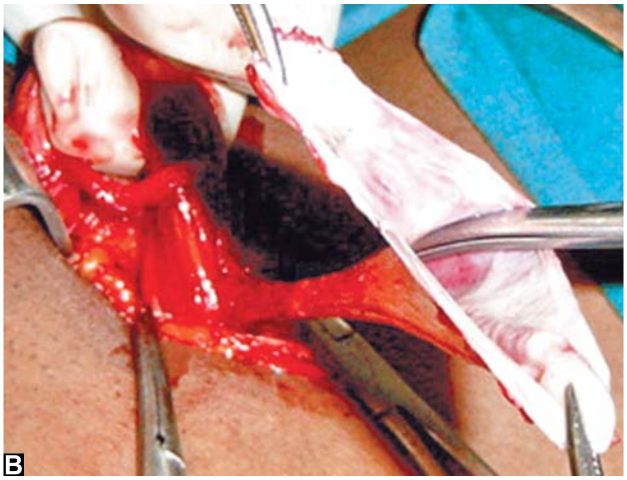

Fig. 18.27B

Figs 18.27A and B: Indirect inguinal hernial sac on table and opening the sac over fundus.

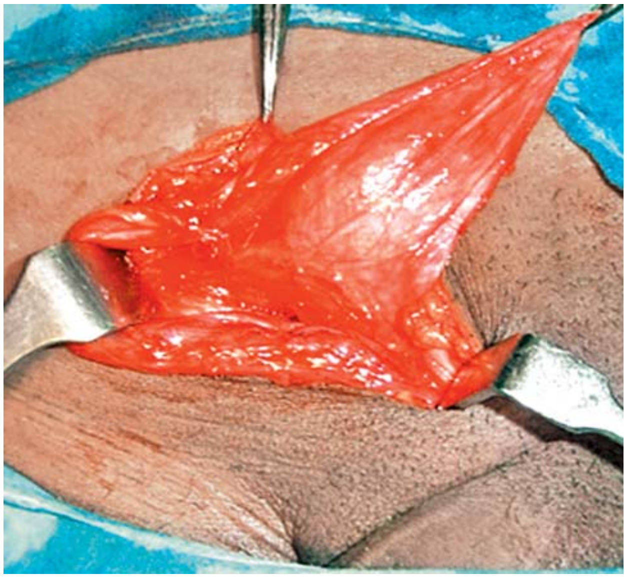

Fig. 18.28: On table direct sac which is medial to inferior epigastric artery. Direct sac is usually not opened. Inverting sutures may be placed if needed.

Investigations

- Chest X-ray to rule out chronic bronchitis.
- Ultrasound of abdomen.
- Tests relevant for precipitating causes.

Treatment

> *This method reconstructs the inguinal canal as it is physiologically, with two rings, one abdominal, the other subcutaneous; and with two walls, one posterior and the other anterior, between which the spermatic cord passes obliquely.*
>
> —**Edoardo Bassini,** 1887

Always Surgery

In infants:

- Whether it is hernia or hydrocele, only *herniotomy* is done through inguinal approach (*Michaelis plank operation*).

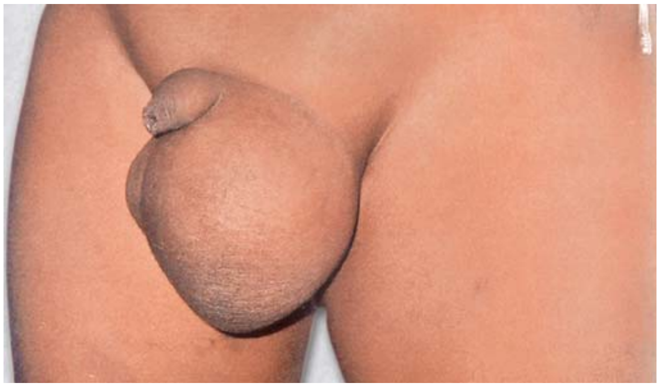

Fig. 18.29: Inguinal hernia in 4-year-old child. Here just herniotomy is required. Repair is not done.

In adults:

- It includes *herniotomy,* i.e. excision of hernial sac and ***herniorrhaphy or hernioplasty (ideal)*** (strengthening of the posterior wall of inguinal canal either by repair or mesh). Precipitating causes should be treated first, like TURP for BPH, dilatation of stricture urethra, treatment of chronic bronchitis. Patient is advised to avoid smoking.
- Hernioplasty is the present choice (ideal) for all inguinal and groin hernias. Mesh is placed either onlay/underlay (over conjoint tendon to inguinal ligament) or inlay (in preperitoneal space). Polypropylene mesh is used. Herniotomy is done prior to mesh placement. TEP (Totally extraperitoneal laparoscopic preperitoneal mesh repair) is preferred method. (For detail refer hernioplasty in later part of this chapter).

Repair may be:

- *Pure tissue repair*: Shouldice, MacVay (Still very useful repairs) and Modified Bassini (not very useful—high recurrence rate as it is repair with tension and nonphysiological)
- *Prosthetic repair*: Lichtenstein, Rives, Gilbert, Stoppa, TEP, TAPP

Repair also can be:

- *Anterior repair*: Through anterior inguinal approach—Bassini's, Shouldice, MacVay, darning, Andrew's, Wilkinson, Copper's, Lichtenstein mesh repair, PHS repair, Rives preperitoneal repair
- *Posterior repair*: Through suprainguinal preperitoneal approach—Nyhus repair, Stoppas, TEP, TAPP, Kugel's repairs

Herniotomy

*Anaesthesia***:** Spinal or G/A or local anaesthesia.

Procedure: After cleaning and draping, skin is incised—1.25 cm above and parallel to the medial two/third of inguinal ligament. Two layers of superficial fascia (outer Camper's fascia and inner Scarpa's fascia) are incised. External oblique aponeurosis is incised. Upper leaf is reflected above and lower leaf is reflected downwards to visualise and expose the inguinal ligament. Ilioinguinal nerve is safeguarded. Cremasteric muscle is opened. Cord structures are dissected. Sac which is anterior and lateral to cord is identified and is pearly white in colour. Dissection is usually started from the fundus and extended towards the neck which is identified by extraperitoneal fat. The neck is narrow and is lateral to inferior epigastric artery. Sac is opened at the fundus. Finger is passed to release any adhesions. Sac is twisted so as to prevent the content from coming back. It is transfixed using absorbable suture material (chromic catgut 2-0 or *vicryl*) and is excised distally.

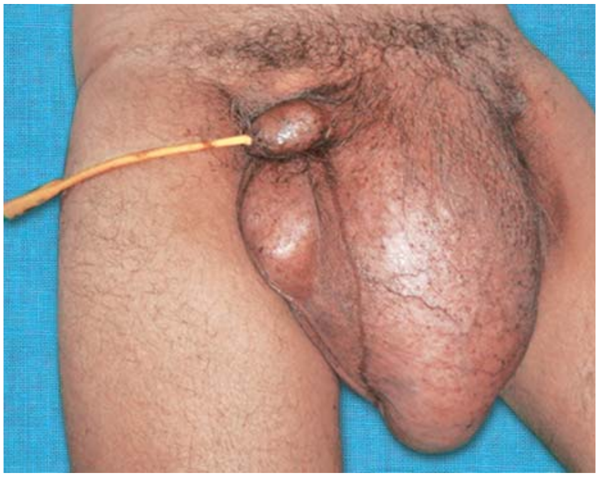

Fig. 18.30: Large hernia with Foley's catheter. Patient had retention of urine due to benign prostatic hyperplasia. He underwent Transurethral resection of prostate [TURP] first and later hernioplasty.

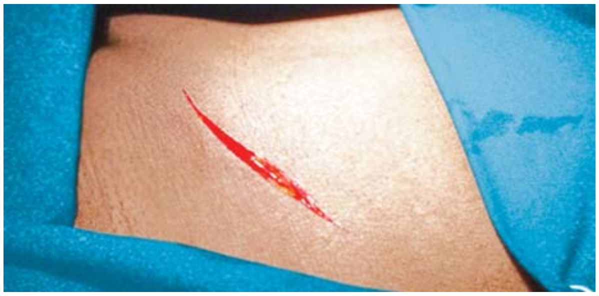

Fig. 18.31: Incision for inguinal hernia. It is 1.25 cm above and parallel to medial third of inguinal ligament.

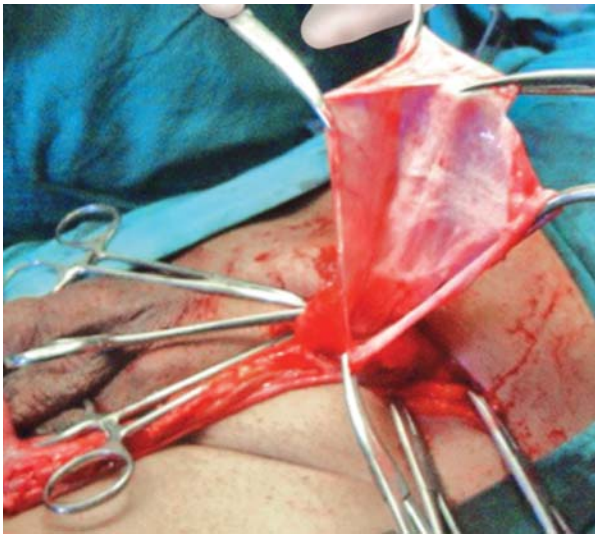

Fig. 18.32: Opened indirect sac in herniotomy.

Usually saphena varix feels softer than a femoral hernia. **— Robert Milnes Walker**

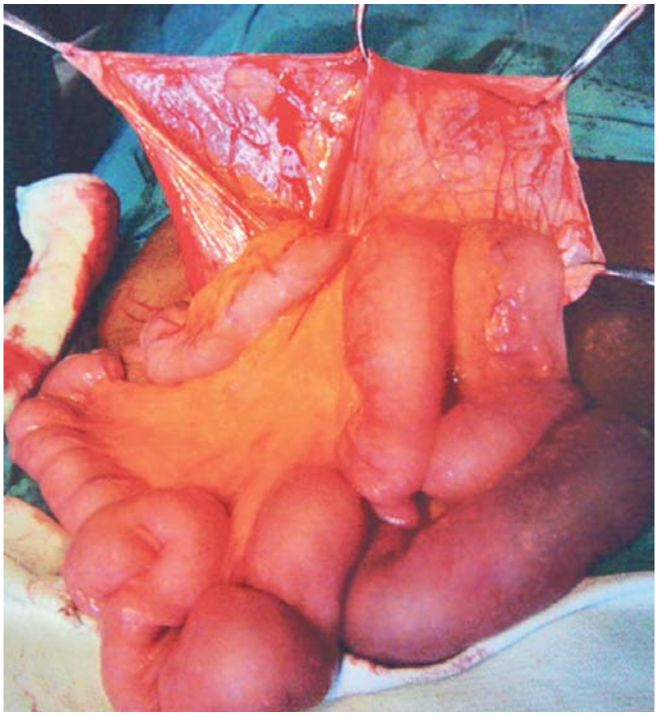

Fig. 18.33: Hernia with enterocele as content.

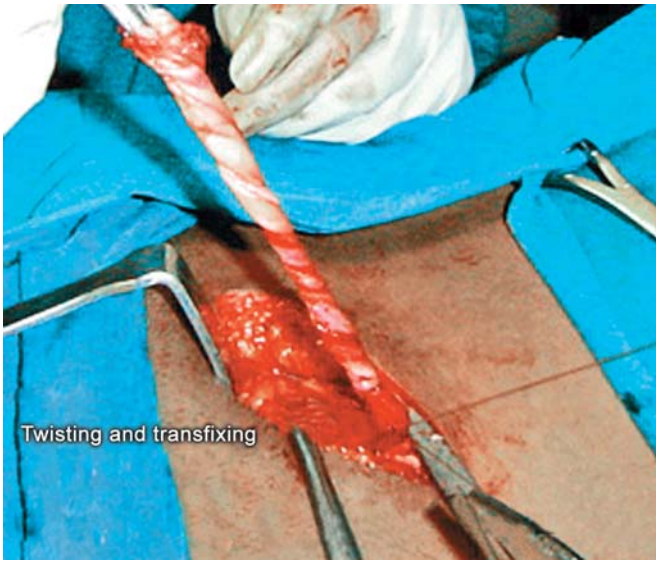

Fig. 18.34: Sac should be twisted before transfixing to prevent contents getting in.

Modified Bassini's Herniorrhaphy

- ♦ Conjoint tendon and inguinal ligament are approximated using interrupted nonabsorbable monofilament sutures {polypropylene (prolene, blue in colour)}. Medial most stitch is taken from the periosteum of pubic tubercle (*called as key or Bassini's stitch*). External oblique is closed and other layers are also closed.
- ♦ It is strengthening of the posterior wall of the inguinal canal by approximation of the conjoint tendon to inguinal ligament using monofilament nonabsorbable suture material. Absorbable suture material like catgut should not be used as 50% of its tensile strength will be lost in 7 days. It takes 6 months to achieve more than 80% of tensile strength in repaired hernial wound; and so nonabsorbable suture material has to be used here to maintain the same adequate tensile strength in this period. Multifilament suture material like silk may precipitate infection because of the crevices in the suture material and tensile strength is not as good as monofilament suture material. Commonly used suture material is either *polypropylene* [*prolene* (blue in colour)] or *polyethylene* [*ethylon* (black in colour)]. Continuous sutures compromises the blood supply and interferes with proper healing; and strength will not be as adequate as interrupted sutures. So always interrupted sutures are used. Earlier, most common surgery done for groin inguinal hernia was modified Bassini's repair. But now hernioplasty is the commonly done procedure for both direct and indirect sac. In direct hernia, sac is usually not opened but in indirect hernia, sac is always opened.

Anaesthesia Used for Inguinal Hernia Repair

General/spinal/epidural or local anaesthesia can be used to do inguinal hernia repair. Local anaesthesia is becoming popular for open approach. General anaesthesia is needed in children and for TAPP and TEP procedures.

Technique of Local Anaesthesia

Around 50-60 ml of xylocaine 0.5% is used. Plain xylocaine 0.5% or xylocaine 0.5% with adrenaline can be used. Plain xylocaine—dose is 2 mg/kg body weight. Xylocaine with adrenaline—dose is 7 mg/kg body weight.

Two methods are used:

1. *Nerve block method (point block)*
   - 10 ml of xylocaine is infiltrated 2 cm above and medial to anterosuperior iliac spine to block the iliohypogastric nerve.
   - Midinguinal point is infiltrated with 10 ml xylocaine.
   - Pubic tubercle point is infiltrated with 10 ml xylocaine.
   - 10 ml of xylocaine is infiltrated just below the inguinal ligament lateral to femoral artery to block the genital branch of genitofemoral artery.
   - Line of skin incision is infiltrated with 10 ml of xylocaine.
   - Later neck of the hernial sac is infiltrated with 10 ml of xylocaine.
2. *Field block method (Shouldice method)*
   - Skin of around 4 cm wide area is infiltrated into the subcutaneous plane as first layer from anterosuperior iliac spine to pubic symphysis. Skin, subcutaneous and two layers of superficial fascia (Camper and Scarpa's) are incised.
   - Area deep to external oblique aponeurosis is infiltrated with 10 ml of xylocaine. External oblique aponeurosis is incised.

- Exposed inguinal canal and hernial sac is infiltrated with 10 ml of xylocaine to continue with the dissection.

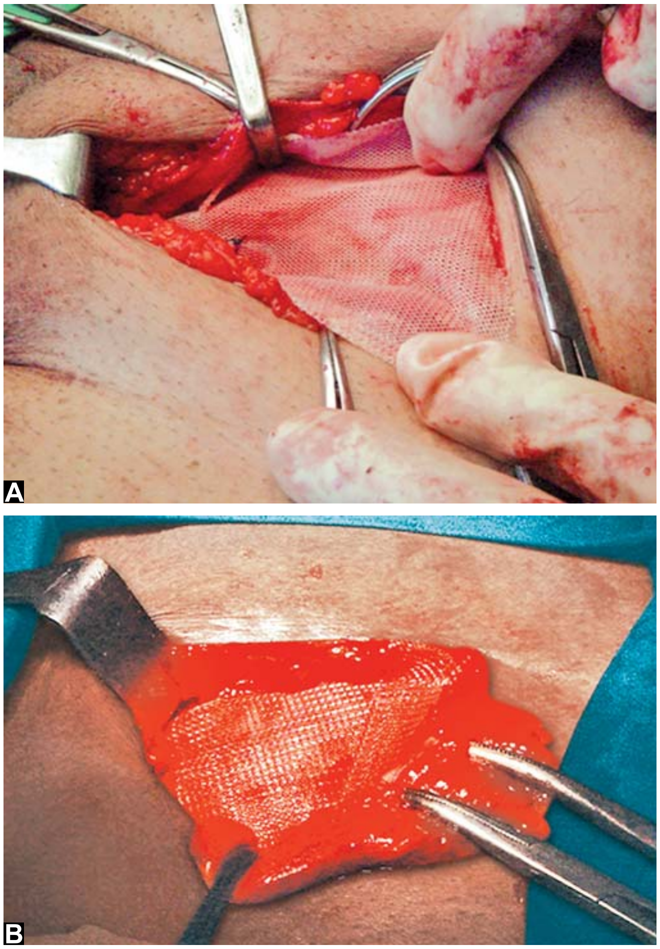

Figs 18.35A and B: Inguinal hernia with mesh placement. Prolene mesh is white in colour. It is sutured to conjoint tendon and inguinal ligament by interrupted sutures using nonabsorbable monofilament suture material.

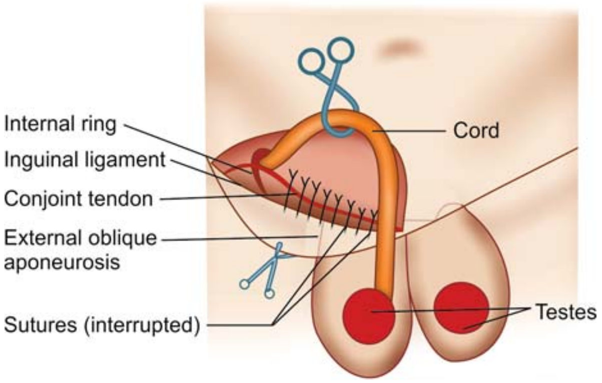
Internal ring
Inguinal ligament
Conjoint tendon
External oblique aponeurosis
Sutures (interrupted)
Cord
Testes

Fig. 18.36: *Modified Bassini's repair*. It is approximation of inguinal ligament to conjoined tendon using interrupted nonabsorbable monofilament sutures.

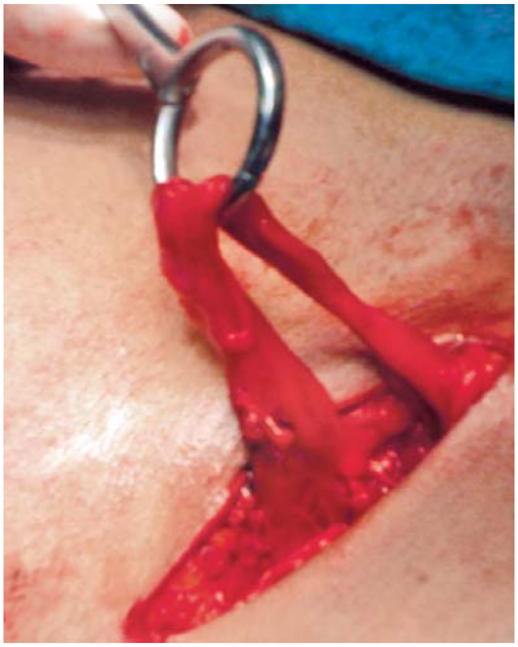

Fig. 18.37: On table inguinal hernia surgery with cord holding forceps.

Complications of herniorrhaphy

- Haemorrhage
- Haematoma, seroma
- Infection—1-5%
- Haematocele
- Post-herniorrhaphy hydrocele, lymphocele
- Hyperaesthesia over the medial side of inguinal canal due to injury to iliohypogastric nerve—neuralgia (15%)
- Recurrence—10-15%
- Osteitis pubis
- Injury to urinary bladder/bowel
- Testicular atrophy, penile oedema rarely can occur

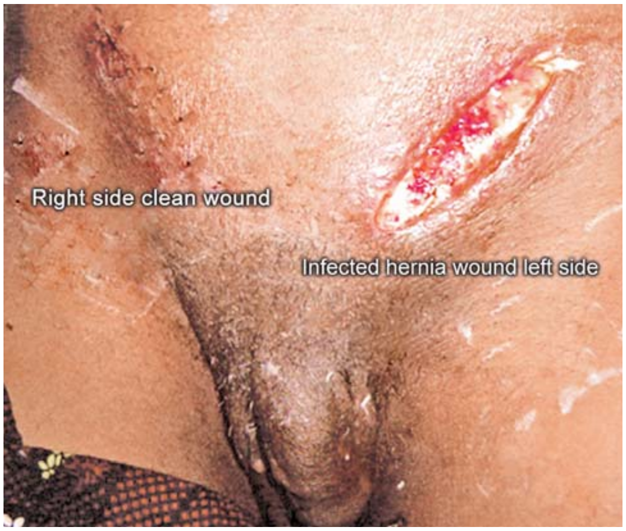
Right side clean wound
Infected hernia wound left side

Fig. 18.38: Left side inguinal hernia wound infected. Right side hernia wound is clean.

In exomphalos, the protruding abdominal contents being covered only by a diphanous membrane, through which viscera are exposed to view, as if exhibited in a show case. **— William E Ladd**

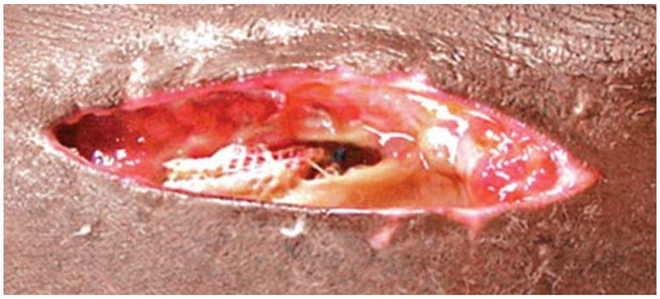

Fig. 18.39: Hernioplasty, mesh is infected. It needs removal of mesh.

Lytle's Repair

Often internal ring is narrowed by placing interrupted sutures over the medial side of the ring to the transversalis fascia using either thread or silk (To narrow the ring and push the cord laterally).

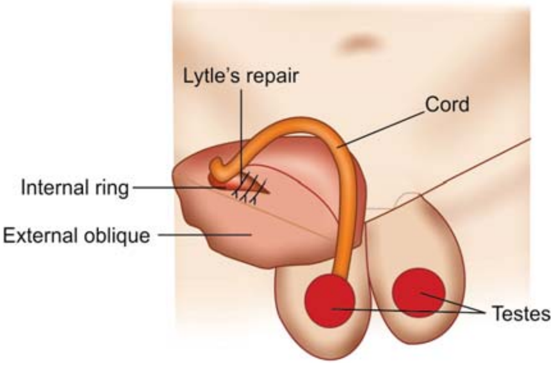

Fig. 18.40: *Lytle's repair*. It is narrowing of the widened internal ring.

Shouldice Repair

Even though transversalis fascia is thin, it is a tough layer and so double breasting of this fascia using continuous sutures (with nonabsorbable material) strengthens the posterior wall of the inguinal wall.

It is a multilayered repair. It was originated at *Shouldice hernia clinic in Toronto* where it was usually done under local anaesthesia. After doing herniotomy as in any other inguinal hernia, transversalis fascia is incised along the line of the wound from deep ring to pubic tubercle. Lower flap of fascia is sutured to posterior part of the upper flap. Upper flap is sutured to the inguinal ligament. It causes double-breasting of the transversalis fascia. Then conjoint tendon and inguinal ligament is further approximated by two layers of continuous sutures. External oblique aponeurosis is sutured in two layers (double-breasting) in front of the cord. Hence the original Shouldice repair is 6 layered procedure. First two layers of transversalis fascia, next two layers of conjoint tendon and last two layers of external oblique aponeurosis. Suture material used here is fine steel wire of 34 gauge (in original Shouldice repair) or polypropylene or polyethylene. Recurrence rate is 1%.

Berliner modified Shouldice repair—It involves double-breasting of the transversalis fascia like in Shouldice repair and single layer closure of the external oblique aponeurosis without any additional two-layered repair of conjoint tendon to inguinal ligament.

Note:
Cremaster is excised in Shouldice repair along with external spermatic fascia to achieve good success rate. As testis loses its support, it causes *hanging/clapper in bell testis*. It also injures genital branch of genitofemoral nerve.

Tanner Slide Operation

To reduce the tension in the repair area, relaxing incision is placed over the lower rectus sheath so that conjoined tendon is allowed to slide downward.

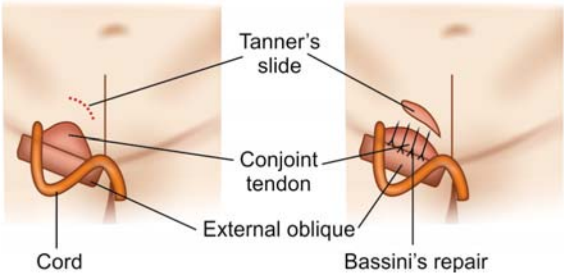

Fig. 18.41: *Tanner slide operation*—relaxing incision placed over the lower medial rectus sheath to reduce the tension after modified Bassini's repair.

Darning (Abrahamson Nylon Darning)

Continuous intervening network of nonabsorbable sutures are placed between conjoint tendon and inguinal ligament to give good support to posterior wall of inguinal hernia.

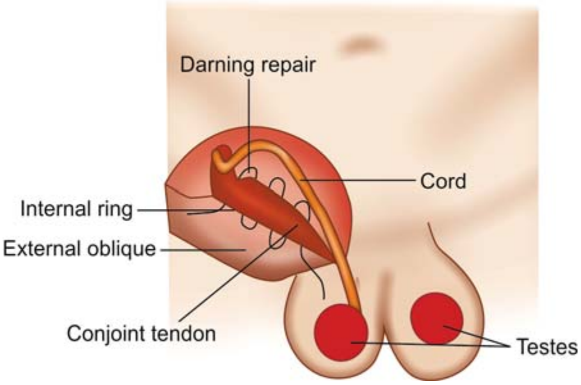

Fig. 18.42: Darning between inguinal ligament and conjoined tendon to act as a support to posterior wall of the inguinal canal.

Koontz Operation (New York, 1963)

In old people after taking consent, orchidectomy is done along with removal of entire cord, testis and total closure of posterior inguinal wall by repair so as to reduce the recurrence.

Removal of Cord at Inguinal Region (Hamilton-Bailey Operation)

Cord is removed from the inguinal canal by ligating both at external and internal ring. But testis is retained (for psychological reason) and closure of inguinal canal by repair is done.

Andrew's Operation

It involves overlapping of the external oblique apo-neurosis.

McVay Operation—1940 (Cooper's Ligament Repair)

It is repair by placing interrupted sutures between transversalis fascia to Copper's ligament (superior pubic ligament) starting from pubic tubercle medially towards femoral sheath and later continued as suture repair between transversalis fascia and iliopubic tract laterally up to the entrance of cord is reached. It is a pure tissue repair. It requires relaxing vertical incision at the lateral border of the anterior rectus sheath, from pubic tubercle point extending superiorly for 4 cm. It covers all three groin defects—indirect, direct and femoral.

Nyhus (Original) Iliopubic Repair

Here with a transverse incision and suprainguinal approach, lateral part of lower end of the rectus is retracted medially after opening anterior rectus sheath. Posterior rectus sheath is opened. Sac is dissected proximally and often ligated. Transaponeurotic arch (transverse abdominis muscle and transversalis fascia) is sutured below to Copper's ligament and iliopubic tract.

Wilkinson Method

Transversus abdominis and internal oblique are sutured to inguinal ligament with continuous monofilament sutures—1st layer; lower free edge of the external oblique is passed above, behind the cord and is sutured to internal oblique surface—2nd layer; upper free edge of the external oblique is brought down in front of the cord and is sutured to lower visible surface of the external oblique—3rd layer. It has less recurrence rate.

Hernioplasty

- Prosthetic mesh repair is used to strengthen the posterior wall of the inguinal canal.
- Onlay like Lichtenstein repair, underlay, PHS repair (Gilbert's), Stoppas, totally extraperitoneal preperitoneal (TEP), transabdominal preperitoneal (TAPP) mesh repair are different prosthetic repairs used. Usually polypropylene mesh is used.
- Recurrence rate has reduced significantly by prosthetic mesh repair. But incidence of mesh *inguinodynia* has increased due to entrapment of ilioinguinal or iliohypogastric nerves in the mesh.
- Even though it is commonly done procedure at present for inguinal hernia, in adolescents whether it should be used or not is a debate. In children it is not used. In strangulated hernia or in presence of sepsis it is not used, only tissue repair is done (Different procedures are discussed at a later part of this chapter).

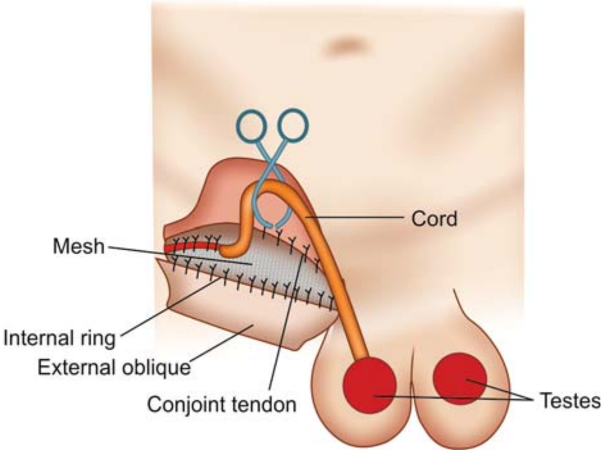

Fig. 18.43: Hernioplasty: Mesh repair—Lichtenstein's method (done under local anaesthesia).

Conservative treatment

- ***Taxis:*** Patient is placed in supine position with hip and knee flexed and hip internally rotated. Contents are pushed with one hand directing with other hand
- ***Use of Truss:*** Rat-tailed sprung truss is used. Measurement is taken from the tip of greater trochanter to third piece of sacrum circumferentially
  - Complications are discomfort, ulceration, strangulation, inflammation, testicular atrophy, ilioinguinal neuritis, femoral neuritis
  - It may be used in elderly people, who are not fit for anaesthesia and surgery
  - Conservative treatment should be avoided in hernia as much as possible
  - Truss is absolutely contraindicated in femoral and sliding hernia
  - Truss increases the surgical bleeding, postoperative oedema, testicular pain

DIRECT HERNIA

- 10-15% of the hernias are direct.
- 50% of direct hernias occur bilateral.
- 35% of inguinal hernias are direct.
- It is uncommon in females and children.
- It is always acquired, due to weakening of posterior wall of inguinal canal.
- Hernia is medial to the inferior epigastric artery with wide neck. Sac is thick and often the medial wall or content may be bladder.
- Direct hernia occurs through *Hesselbach's* triangle which is bounded by inferior epigastric artery laterally, lateral border of rectus medially, inguinal ligament below. It is divided into medial and lateral halves by obliterated umbilical artery (medial umbilical ligament). So direct hernia is classified as *medial or lateral depending on which part of the Hesselbach's triangle, it is arising from.*

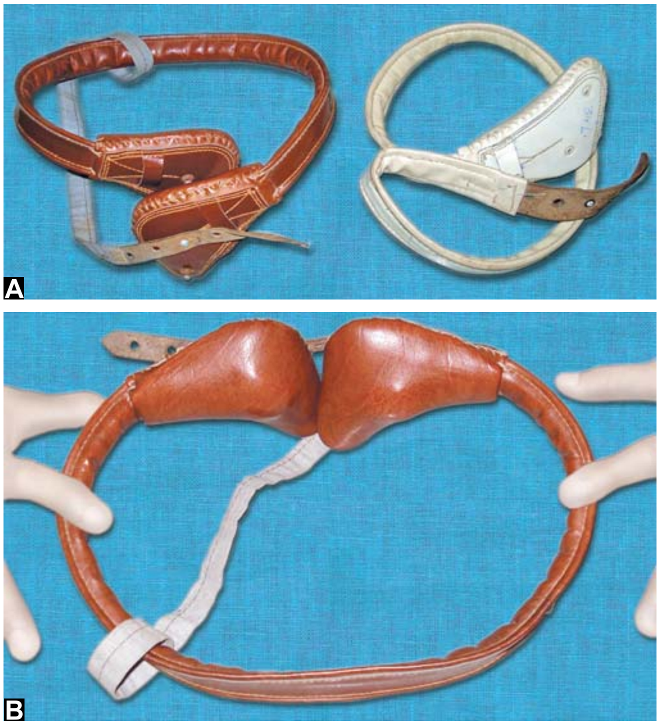

Figs 18.44A and B: *Hernia truss*. It is used only when patient is not fit for surgery. It may precipitate strangulation. Before placing truss, contents of the hernia should be reduced completely.

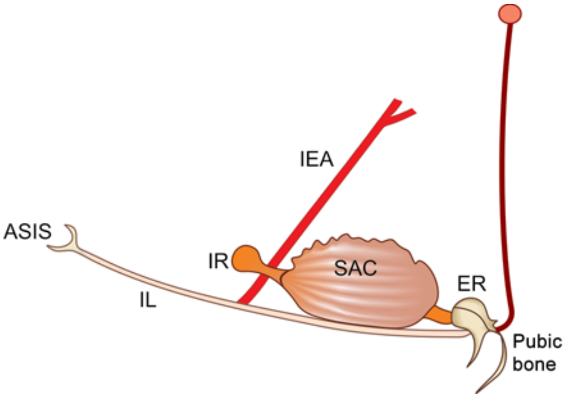

Fig. 18.45: Surgical anatomy of direct inguinal hernia.

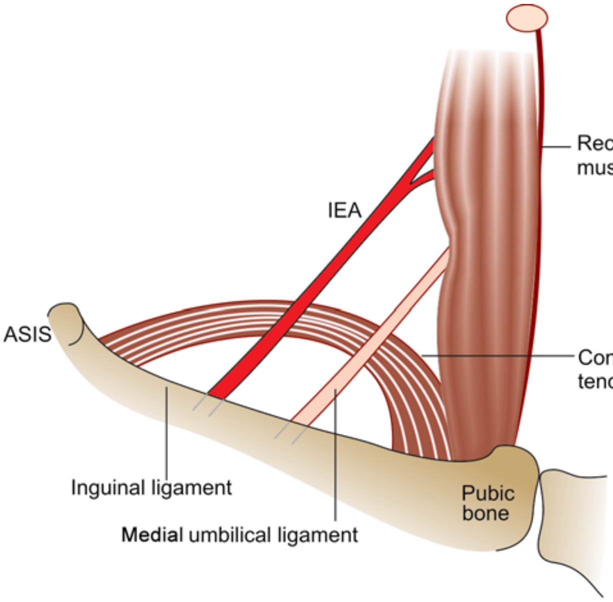

Fig. 18.46: Surgical anatomy of Hesselbach's triangle.

Coverings of direct hernia (from inside out)

- Extraperitoneal tissue
- Fascia transversalis
- Conjoined tendon
- External spermatic fascia
- Skin

Predisposing factors

- Chronic cough, smoking
- Straining
- Constipation
- Heavy work
- Previous appendicectomy

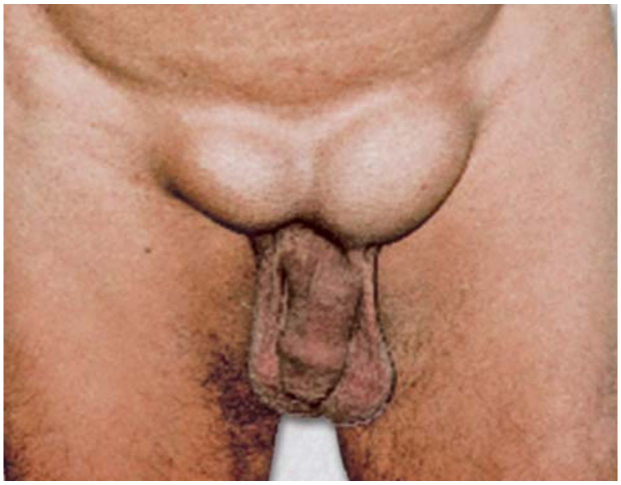

Fig. 18.47: Bilateral direct inguinal hernia.

- ***Malgaigne bulgings*** are often seen in these patients on examination, more often than in indirect hernia. They are protrusion of abdominal wall muscle during leg raising test as weak, soft, supple, swellings which signifies poor abdominal muscle tone.
- Direct hernia rarely descends into the scrotum and strangulation is not as common as in indirect hernia. But in longstanding cases, it can descend down to the scrotum and strangulation can occur.

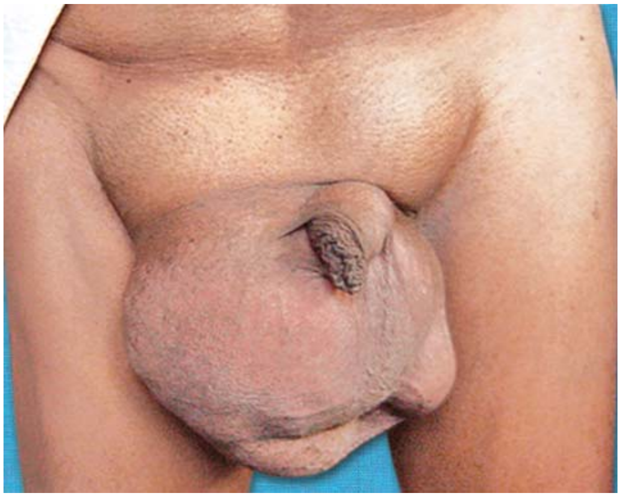

Fig. 18.48: Large bilateral direct hernia. Occasionally direct hernia descends down and becomes complete and then may cause obstruction. Descent is not as common as indirect hernia.

Treatment

Surgery:

- Usually direct sac is not opened.
- Care should be taken at the medial aspect due to the presence of bladder (*bladder should be emptied* before surgery).
- Ideally *hernioplasty (mesh repair*) is done. In case of bilateral hernia, mesh repair can be done on both sides together. Laparoscopic approach (TEP) or suprapubic approach may be better in bilateral cases.

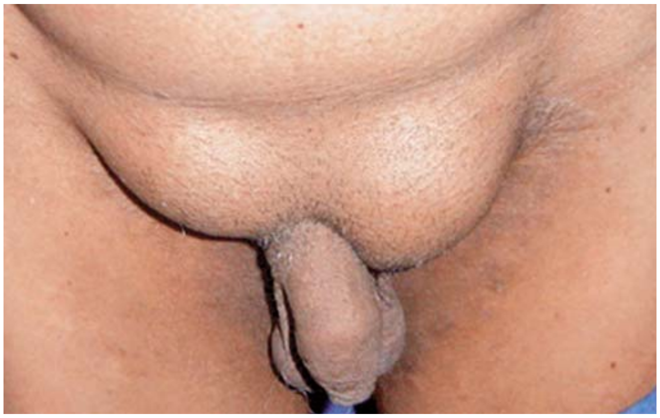

Fig. 18.49: Bilateral direct hernia. Note the medial location of the direct hernias.

Funicular Direct Hernia (Prevesical Hernia)

- It is also called as *Ogilvie's hernia*.
- It is a type of direct hernia which is prone for strangulation.
- It is a narrow necked hernia with prevesical fat and a portion of bladder, and or intestine that herniates through a small defect in the medial part of the conjoined tendon just above the pubic tubercle.
- It occurs in elderly males.

Giant Inguinal Hernia

Hernia reaches below the mid-level of thigh when patient stands.

RECURRENT HERNIA (Inguinal)

- Incidence is 10%.
- If recurrence is within 3 years it is called as *early*; if after 3 years it is *late.*

Predisposing Factors

Preoperative

- Smoking.
- Chronic cough.
- Constipation.
- Old age.
- Anaemia.
- Hypoproteinaemia.
- Straining.
- Increased intra-abdominal pressure of any cause (BPH, carcinoma prostate, stricture urethra).
- Ascites.

Operative

- Tension in the sutures.
- Weak anterior abdominal wall.

Postoperative

- Infection (50%).
- Haematoma formation during earlier surgery.
- Retained sac in pantaloon hernia.
- Straining.

Recurrence Rate

- Bassini's repair—10%.
- Shouldice repair—1%.
- Hernioplasty—1 to 3%.
- Other methods—1 to 5%.

True or false recurrences are the types of recurrence. If hernia occurs in inguinal region after inguinal hernia repair it is called as *true recurrence*. If other groin hernia occurs after inguinal hernia repair like femoral hernia or obturator or other rare types, it is called as *false recurrence*. But presently hernia is classified grossly as groin hernias and so all recurrences are true recurrences.

Clinical Features

- Same as for inguinal hernia. Defect is usually narrow and so more likely to go in for strangulation.
- It can be *medial recurrence* or *lateral recurrence* depending on the location of the sac. Medial recurrence is common as tension in suture line is greatest near the pubic tubercle.

Treatment

- After thorough investigations, the cause of recurrence has to be treated and later *hernioplasty* is done.
- *Laparoscopic (TEP/TAPP) approach* is better for recurrent hernia. In open repair, preperitoneal mesh repair is ideal. In elderly people *Koontz orchidectomy* or cord excision at inguinal canal may be added after proper prior consent.

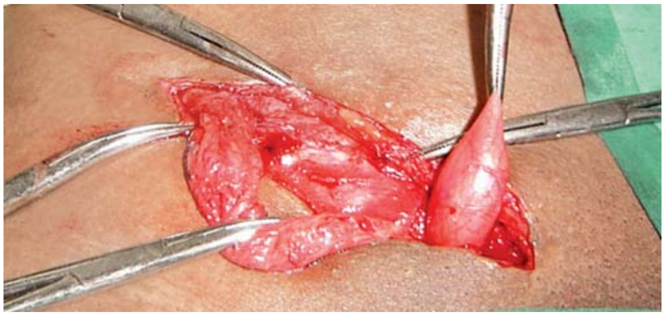

Fig. 18.50: Recurrent inguinal hernia is 10% common. It is commonly medial recurrence. It is treated by mesh repair.

Femoral hernia is never congenital.

Rerecurrent Hernia

1% common.

Causes

Infection, earlier mesh extrusion, failure of treating the precipitating causes.

Treatment

- The cause is treated and *mesh repair is done*.
- GPRVS (Giant prosthesis reinforcement of the visceral sac) or TEP through laparoscopy is ideal.

HERNIOPLASTY

It is strengthening of posterior inguinal wall in case of indirect hernia or in any large hernia with weak abdominal wall using a supportive material. This allows and supports good fibroblast proliferation which in turn strengthens the weak posterior wall of inguinal canal or abdominal wall.

Material Used

- *Synthetic:* Prolene mesh (white in colour) Dacron mesh, Morlex mesh, Mersilene sheath.
- *Biological:* Tensor fascia lata, temporal fascia and skin. (presently biological materials are not well-accepted as infection is common and its efficacy is not proved).

Prolene mesh is commonly used at present. Nonabsorbable interspersed absorbed mesh (Vipro/Ultrapro) are also used nowadays.

Indications

- Direct hernia.
- Recurrent hernia.
- Re-recurrent hernia.
- Incisional hernia.
- Old age.
- Hernia with weak abdominal muscle tone.
- Sliding hernia.

Complications

- Infection.
- Mesh extrusion.
- Foreign body reaction.
- Mesh *inguinodynia* causing hyperaesthesia and pain along the distribution of ilioinguinal or iliohypogastric nerves.
- Mesh erosion into bladder, bowel or vessels can occur occasionally (rare).

Principle

- Size of the mesh should be bigger than the size of the defect.
- Mesh should be fixed above and below to the conjoint tendon and inguinal ligament or abdominal wall using interrupted, nonabsorbable sutures.
- Absolute haemostasis and control (prevention) of infection is important.

Types of mesh repair

- Onlay repair
- Lichtenstein mesh repair—a type of onlay
- Inlay repair—mesh placed at myopectineal level—*bridging repair*
- Underlay repair—mesh placed at deeper plane usually in preperitoneal space
- Gilbert patch and plug repair/Gilbert's PHS repair (onlay + sublay)
- Nyhus preperitoneal mesh repair
- Kugel groin hernia mesh repair
- Modified Rives preperitoneal mesh repair through inguinal approach
- Stoppa GPRVS repair
- TEP mesh repair
- TAPP mesh repair

Hernioplasty is becoming the prime treatment for inguinal hernia.

Types

Different types are:

- *Onlay mesh* repair by placing mesh in front. It is sutured above to conjoint tendon and below to inguinal ligament using monofilament non absorbable suture material.
- *Lichtenstein tension free onlay mesh repair* (1993) where the cord is encircled with mesh which is often done under local anaesthesia. Suturing of mesh is done similar to onlay mesh repair. It has got less recurrence rate.
- *Nyhus preperitoneal mesh repair:* It is done through *suprainguinal*horizontal incision above the pubic symphysis and internal ring. Preperitoneum is approached through lateral border of the lower part of rectus muscle by making an opening in the posterior rectus sheath. Mesh is placed in the preperitoneal space deep to the cord, conjoint tendon, and transversalis fascia. Below, it is folded deep to the iliopectineal ligament of Cooper and sutured to it using two or three interrupted nonabsorbable sutures. It is sutured to transverse abdominis above and transversalis fascia from deep.
- *Modified Rives preperitoneal mesh repair (Read-Rives)* is preperitoneal mesh repair through *transinguinal approach*. Direct sac is inverted with sutures. For indirect sac a high ligation is done. Here mesh is placed in preperitoneal space, folded and sutured below to iliopectineal ligament, above to the transverse abdominis in deeper plane. Often transversalis fascia opened earlier is sutured back using nonabsorbable suture material in front of the placed mesh.
- *Stoppa's giant prosthesis reinforcement of visceral sac (GPRVS):* It is done in large hernias, hernias in elderly, bilateral hernias, recurrent and rerecurrent hernias, hernia with very lax abdomen, and hernia with collagen diseases like Marfan's syndrome. Horizontal length (size) of the mesh is 2 cm less than distance between two anterosuperior iliac spines and vertical length (size) is distance between the umbilicus and pubic symphysis. Large mesh is placed

between peritoneum and lateral, inferior, anterior abdominal wall which stretches in the lower abdomen and pelvis. It is done through lower midline or Pfannensteil incision. Usually such large mesh is placed without any anchorage.

- *Gilbert mesh repair (patch and plug):* After herniotomy, internal ring is plugged by cone-shaped (umbrella plug) piece of prolene mesh. Later onlay/inlay mesh repair of posterior wall of the inguinal canal is done.
- *Gilbert's PHS (Prolene Hernia System) repair (onlay and sublay—sandwich technique)*: It is an open transinguinal approach to place a specially devised mesh in both preperitoneal as inlay and in front as onlay mesh repair. This mesh has got a rounded deeper part which is placed in preperitoneal space and a modified quadrangular part which is placed as onlay. Both parts are connected through a connecting stiff rounded part in between which steadies the mesh in place preventing its displacement.

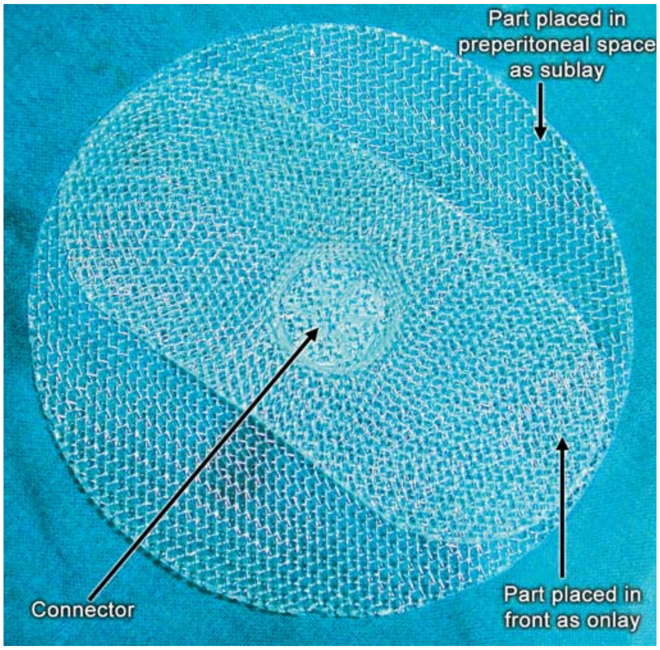
Part placed in preperitoneal space as sublay
Connector
Part placed in front as onlay

Fig. 18.51: Gibert's PHS mesh.

- *Transabdominal preperitoneal laparoscopic mesh repair (TAPP repair):* It is mainly used in irreducible and large hernias.
- *Totally extraperitoneal laparoscopic mesh repair (TEP)*—becoming popular.

Transabdominal Preperitoneal Mesh Repair (TAPP) Using Laparoscope

This is used in large indirect hernia or irreducible inguinal hernia. 10 mm umbilical port is used for laparoscope. 5 mm ports are placed one on each side on pararectal point at the or above the level of umbilicus so as to achieve adequate triangulation. Contents of the hernia are reduced. Hernial sac is dissected in preperitoneal plane after making horizontal incision at the upper part of the sac opening. Vas, gonadal vessels, pubic bone, inferior epigastric vessels are identified. Once sac is dissected and excised, a prolene/vipro/ultrapro mesh of 15 × 10 cm sized or smaller is placed in preperitoneal space. It is fixed to pubic bone using tacks. Peritoneum is closed with continuous prolene sutures.

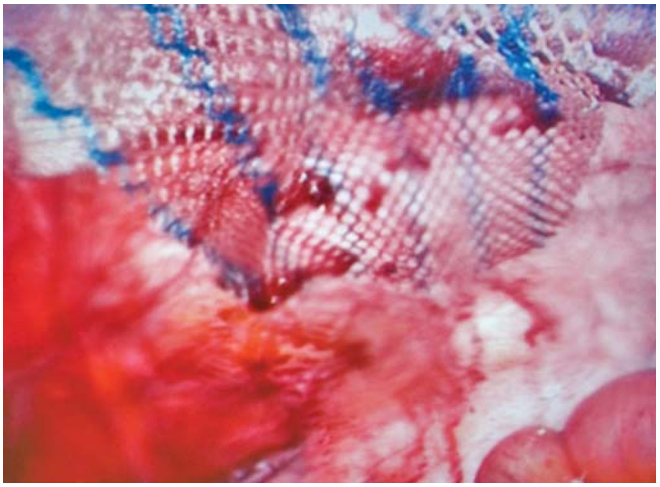

Fig. 18.52: Mesh placement in TAPP procedure.

Totally Extraperitoneal Repair (TEP Repair) Using Laparoscope

- This technique is gaining more popularity than TAPP. Through subumbilical incision (10 mm) extraperitoneal space is reached. After CO_2 insufflation, another 5 mm port is inserted 4 cm below the first port in the midline. Third 5 mm port is inserted in the same line 4 cm below or in the right iliac fossa. Dissection is carried out downwards carefully, then medially up to the pubic tubercle, iliopectineal ligament, laterally to iliac vessels, and inferior epigastric vessels. Once adequate space is dissected 15 × 15 cm mesh is placed and spread. Care should be taken not to have any folding in the mesh. Mesh may be sutured to iliopectineal ligament. Displacement of mesh is not common. Other side also can be done together.

Anatomical Considerations

- Preperitoneal space is a potential space in front of the peritoneum, behind the transversalis fascia and anterior rectus muscle. Below in front of the urinary bladder it is called as *space of Retzius* (medially), laterally it is called as *space of Bogros*. Median umbilical fold is formed by urachus in the midline. Medial umbilical ligament is formed by obliterated umbilical arteries, lateral umbilical fold by inferior epigastric vessels. Three fossae are lying in relation to these folds—supravesical and medial fossae are medial to lateral umbilical fold which are sites of direct hernia whereas

Spigelian line: Runs along the lateral border of rectus sheath and extends from the pubic tubercle to the tip of the 9th costal cartilage.
Spigelian fascia: Is the fascia between the aponeurosis of external and internal oblique, transversus abdominis and recuts sheath.

lateral fossa is lateral to lateral umbilical fold and is site of indirect hernia.

- In 1956 *Fruchaud* described his *myopectineal orifice* bounded medially by the lateral border of rectus abdominis, laterally by iliopsoas, superiorly by conjoint tendon and inferiorly by pectin pubis. This area is the site of groin hernia which should be covered by mesh of adequate size to strengthen the defect and to prevent the recurrence. Iliopubic tract is analogue of the inguinal ligament, extends from Cooper's ligament to anterosuperior iliac spine which divides endoscopic view of preperitoneal space into *superior compartment* (contains inferior epigastric artery, Hesselbach's triangle, cord structures and is site of inguinal hernia) and *inferior compartment* (contains femoral canal, iliac vessels, iliopsoas muscle, genitofemoral nerve, lateral femoral cutaneous nerve). External iliac vessels lie in a triangle formed by gonadal vessels laterally, vas deferens medially and peritoneal reflection inferiorly (*triangle of doom*).

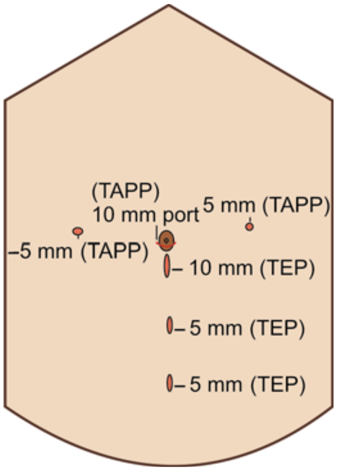
(TAPP)
10 mm port
5 mm (TAPP)
5 mm (TAPP)
10 mm (TEP)
5 mm (TEP)
5 mm (TEP)

Fig. 18.53: Ports used for TEP and for TAPP.

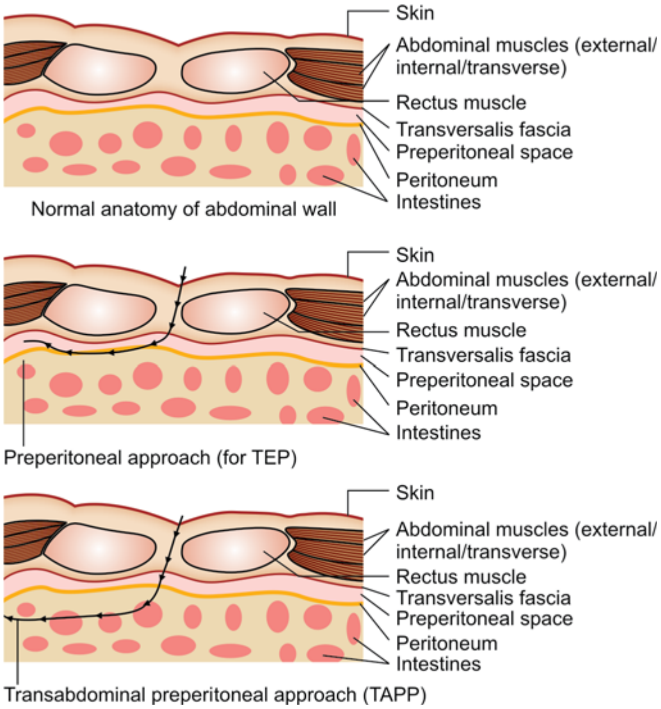
Skin
Abdominal muscles (external/internal/transverse)
Rectus muscle
Transversalis fascia
Preperitoneal space
Peritoneum
Intestines
Normal anatomy of abdominal wall
Skin
Abdominal muscles (external/internal/transverse)
Rectus muscle
Transversalis fascia
Preperitoneal space
Peritoneum
Intestines
Preperitoneal approach (for TEP)
Skin
Abdominal muscles (external/internal/transverse)
Rectus muscle
Transversalis fascia
Preperitoneal space
Peritoneum
Intestines
Transabdominal preperitoneal approach (TAPP)

Fig. 18.54: Diagrammatic representations of TEP and TAPP.

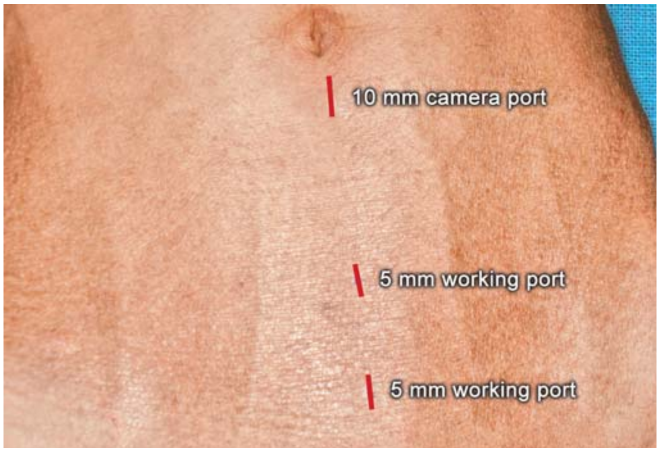
10 mm camera port
5 mm working port
5 mm working port

Fig. 18.55: Port incisions in TEP.

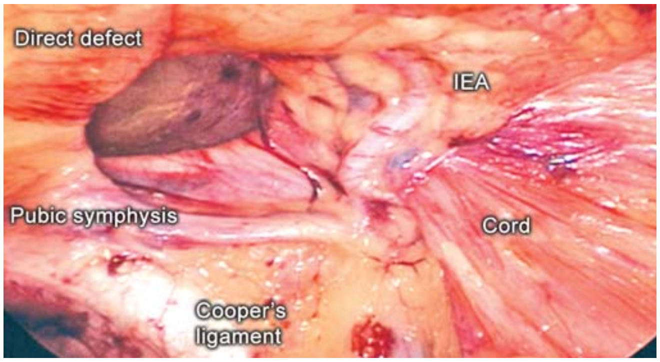
Direct defect
IEA
Pubic symphysis
Cord
Cooper's ligament

Fig. 18.56: TEP after dissection prior to placement of mesh. Direct defect, inferior epigastric artery, pubic symphysis and Cooper's ligament are clearly seen.

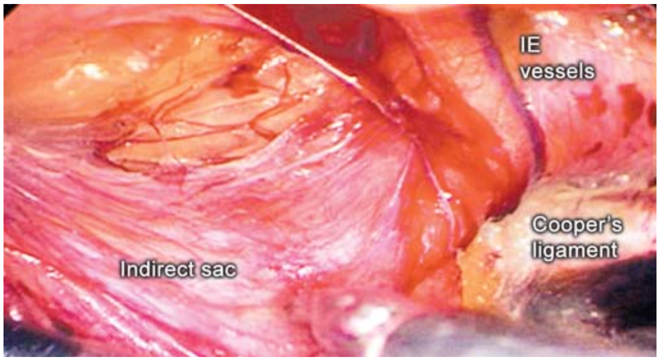
IE vessels
Cooper's ligament
Indirect sac

Fig. 18.57: TEP after dissection showing indirect sac, inferior epigastric vessels, Cooper's ligament.

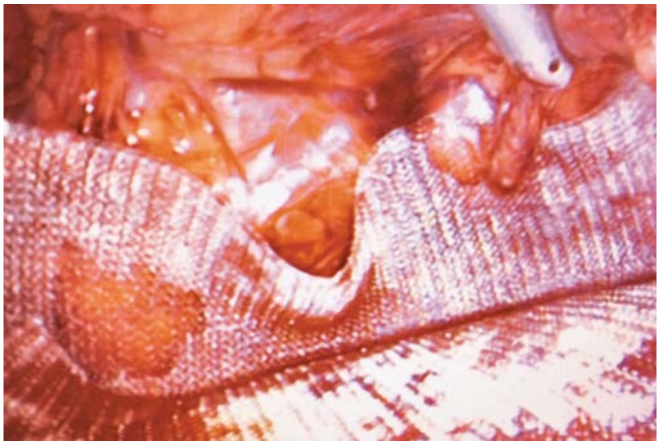

Fig. 18.58: TEP mesh placement in preperitoneal space.

- Aberrant obturator artery is an occasional branch of inferior epigastric artery replacing its pubic branch travels across Cooper's ligament, which during fixation of mesh can cause torrential haemorrhage—*circle of death*. *Triangle of pain* is formed by gonadal vessels medially, iliopubic tract laterally and peritoneal reflection below. Genitofemoral nerve and lateral cutaneous nerve of thigh traverse this triangle. Injury to these nerves either by dissection or by tacks cause postoperative pain. Tacks/staplers should not be placed in this triangle.

Indications for TEP	***Contraindications for TEP***
• Recurrent hernia • Bilateral hernia • Indirect/direct/femoral hernia	• Obstructed/strangulated inguinal hernias • Ascites • Bleeding disorders

Landmarks to be identified in TEP	***Principles in TEP***
• Pubic bone midline • Inferior epigastric artery • Copper's ligament • Iliopubic tract • Cord and vas deferens • Psoas muscle and nerves in relation	• Head-down supine position • Surgeon standing on opposite side of hernia • Camera person placed on opposite side of hernia • Monitor at foot end • Catheterise/empty the bladder properly prior to TEP • Adequate wide space creation • Careful dissection of cord and sac • Ligate indirect sac • Mesh should not be fixed laterally • Size of mesh is 15 × 15 cm • Two point fixation—one at pubic bone other at Cooper's ligament by tacks/staplers

Difficulties and complications in TEP repair

- Difficulty in dissecting indirect sac. Cord/vas injury
- Inadvertent opening of the sac/peritoneum and creation of pneumoperitoneum.
- Injuries to major structures like iliac vessels—0.5-1.0%
- Displacement of mesh or erosion into structures like urinary bladder—rarely may occur
- Nerve injury
- Formation of seroma/haematoma
- Infection
- Recurrence

Advantages of TEP repair

- Approach is totally extraperitoneal
- Small incision
- Proper placement of mesh in right space that is preperitoneal space
- Peritoneal cavity is intact and not opened

MALGAIGNE BULGING

- It is protrusion of abdominal wall muscles during leg rising test as weak, soft, supple, swelling, which signifies poor abdominal muscle tone. It also concludes that particular hernia requires mesh repair ***(hernioplasty)*** as surgical treatment.
- It is common in old age, obese patient.

INCARCERATED HERNIA

- Here the lumen of the portion of colon occupying a hernial sac is blocked with *faeces*. Here scybalous content of the bowel should be capable of being indented with the finger, like putty.
- In incarcerated hernia, sac and contents are densely adherent to each other (contents are fixed to sac). It is always irreducible; often obstructed but may not be strangulated.

STRANGULATED HERNIA

Pathology

It occurs when blood supply of the contents of hernia is seriously impaired leading to formation of gangrene.

Common bacteria in strangulated hernia

- *E. coli*
- Anaerobic streptococci
- Anaerobic bacteria
- *Klebsiella*

Obstruction
↓
Initially venous return is impaired
↓
Congestion of the bowel
↓
Further dilatation of the bowel which becomes purple coloured
↓
Fluid collects in the sac
↓
Eventually arterial blood supply is impaired
↓
Bowel becomes dark, brownish black coloured with flabby and friable wall
↓
Bacteria migrate transerosally and multiply in fluid of the sac
↓
Perforation occurs at the site of constriction ring
↓
Peritonitis occurs.

- Strangulation commonly occurs in the *small bowel* and also in large bowel. Occasionally strangulated omentocele also can occur without any intestinal obstruction.

Smile is the curve which puts everything straight.

- Strangulation can occur in inguinal, femoral, obturator, umbilical or any hernias.
- Indirect inguinal hernia is more prone for strangulation than direct inguinal hernia. It is due to narrow neck, adhesions, narrow external ring in children.

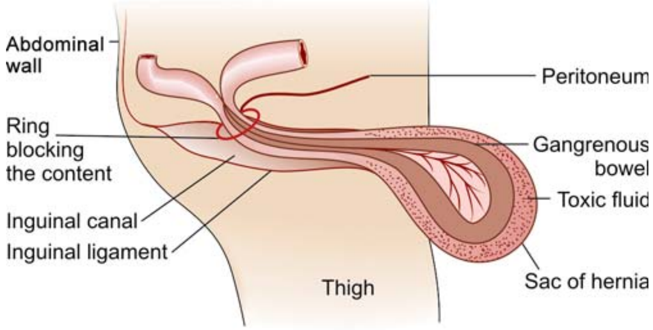

Fig. 18.59: Picture showing strangulated hernia with toxic fluid and site of constriction.

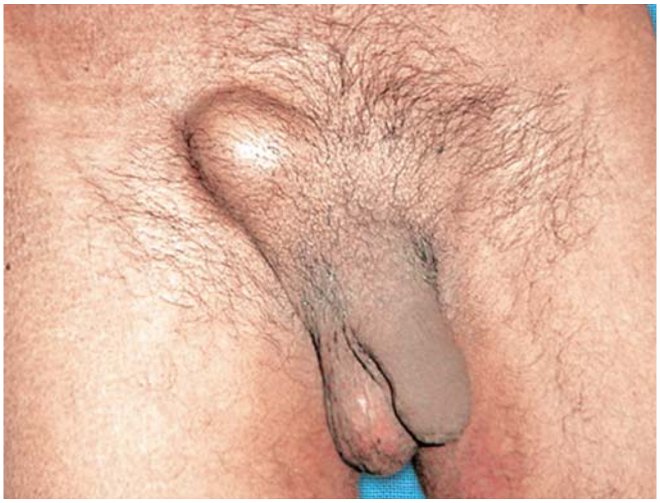

Fig. 18.60: Strangulated inguinal hernia will be tender and irreducible without impulse on coughing.

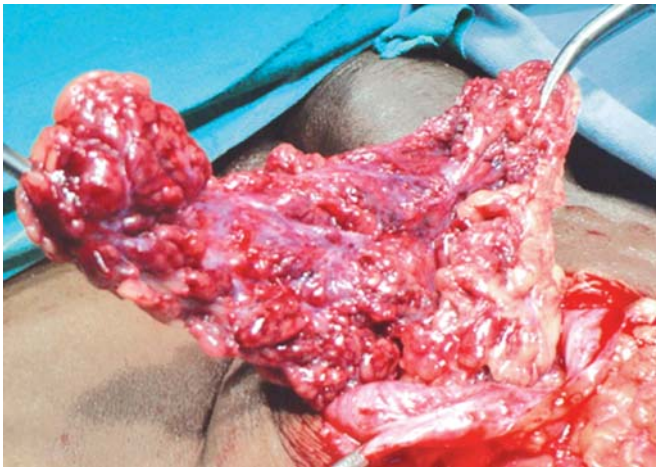

Fig. 18.61: Hernia with gangrenous omentum. Note the colour of the gangrenous omentum.

- Part of circumference of the bowel when strangulated, is called as ***Richter's hernia*** wherein the patient presents with diarrhoea, toxicity mimicking *gastroenteritis.* Richter's hernia is more common with femoral, obturator hernias.

Causes of strangulation

- Narrow neck
- Adhesions
- Irreducibility
- Long-time, large hernia with adhesions

Maydl's hernia (Hernia-in-W): Here a loop of bowel in the form of 'W' lies in the hernial sac and the centre portion of the 'W' loop is strangulated and lies within the abdominal cavity**.** Thus local tenderness over the hernia is not marked and hernia gets reduced with the strangulated loop in the centre of the ***"W"***. Strangulation in this case is often missed during surgery and may lead to peritonitis due to retained gangrenous loop.

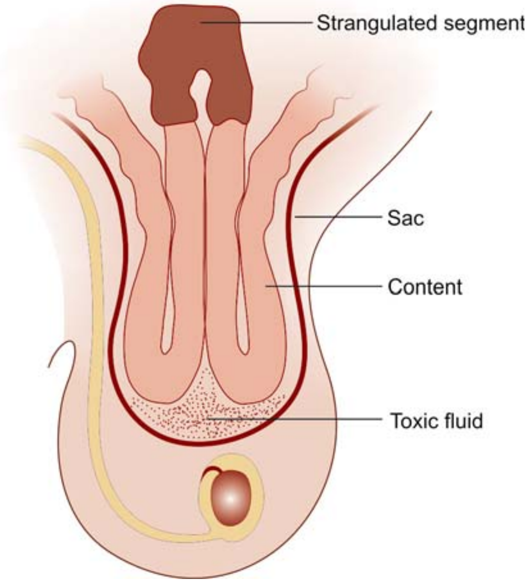

Fig. 18.62: Maydl's hernia (Hernia-in-W).

Clinical Features of Strangulated Hernia

- Sudden severe pain, initially over a pre-existing hernia which later becomes generalised over the abdomen.
- Persistent vomiting, constipation and distension of the abdomen.
- *Hernia is tense, severely tender, irreducible and without any expansile impulse on coughing. Rebound tenderness is diagnostic.*
- Features of toxicity and dehydration.
- Electrolyte imbalance.
- Abdominal distension with guarding and rigidity.
- Oliguria.
- 3% in incidence.

Features of strangulated hernia

- Tense, tender, irreducible
- No impulse on coughing
- Shock, toxicity
- Features of obstruction when there is enterocele
- Abdominal distension, vomiting
- Rebound tenderness

Strangulation during infancy

Incidence is 4%. Female to male ratio is 5:1

In female infants, the content may be ovary with or without fallopian tube

Taxis

- Often in irreducible hernia, *reduction* of hernia is *tried* by elevation, sedation and ***taxis*** (i.e. with flexion and medial rotation of the hip, reduction of hernia is tried).
- *In obstructed hernia, taxis may be dangerous* as during taxis, contusion and rupture of the intestinal wall can occur. Reduction-en-masse may mask the gangrenous bowel existing in the sac. Inner gangrenous loop of Maydl's hernia may be missed. Rupture of the sac extraperitoneally is also a possibility.
- *Taxis has no role in femoral hernia and strangulated hernia.* If tried, contusion, reduction-en-masse and rupture of the sac can occur.

Taxis

- Used in irreducible or partially reducible hernia
- *Trial reduction by flexing and medially rotating the hip*
- It is dangerous in:
  - Obstructed hernia
  - Maydl's hernia

  (As rupture can occur leading to peritonitis)
- No role in femoral and strangulated hernia

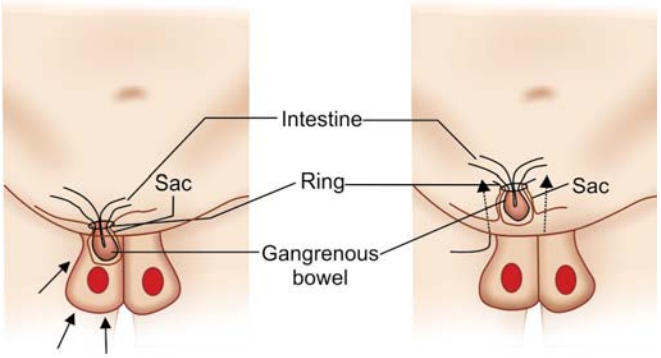

Fig. 18.63: In an obstructed hernia, if reduction is forced it may get reduced with the sac and ring with the obstructed/strangulated bowel—*Reduction-en-masse*.

- ***In strangulated omentum*** features of obstruction are not present (i.e. vomiting, constipation). Omentum becomes congested, oedematous and black in colour which secretes toxic fluid with secondary bacterial infection. But here, initially the sepsis is slower than that of strangulated intestinal obstruction.
- Eventually the infection spreads causing diffuse peritonitis. Downward spread of infection can cause scrotal abscess.

Investigations

- Plain X-ray abdomen in erect posture shows multiple air-fluid levels.
- Serum electrolytes.
- Blood urea and serum creatinine.
- Total count is increased.
- U/S abdomen.

Strangulation without obstruction

- Richter's hernia
- Omentocele

Treatment of Strangulated Hernia

- The patient is admitted.
- Ryle's tube aspiration.
- Intravenous fluids to correct dehydration and electrolyte imbalance.
- Antibiotics.
- Catheterisation to maintain adequate urine output.
- *Emergency surgery*:

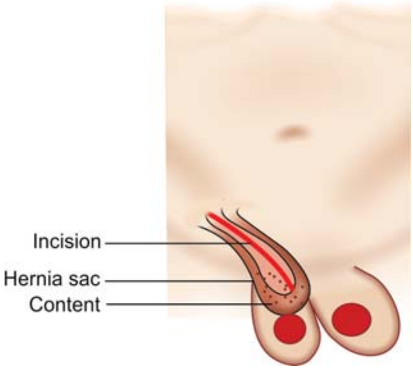

Fig. 18.64: Incision for strangulated hernia is obliquely placed extending to the scrotum.

Groin incision is made with incision extending into the most prominent area of the swelling.

↓

Sac is exposed.

↓

Constriction ring and superficial ring is released (cut).

↓

Sac is opened carefully *without allowing the spillage of fluid* (Usually spillage occurs extraperitoneally)

↓

Fluid is sucked with a suction apparatus.

↓

The bowel is held with fingers so as to prevent it from getting reduced.

↓

The viability of the bowel is checked by *colour, peristalsis, pulsation, bleeding*.

↓

When gangrenous, resection and anastomosis is done and drain is placed.

The most important thing that you wear is the expression on your face.

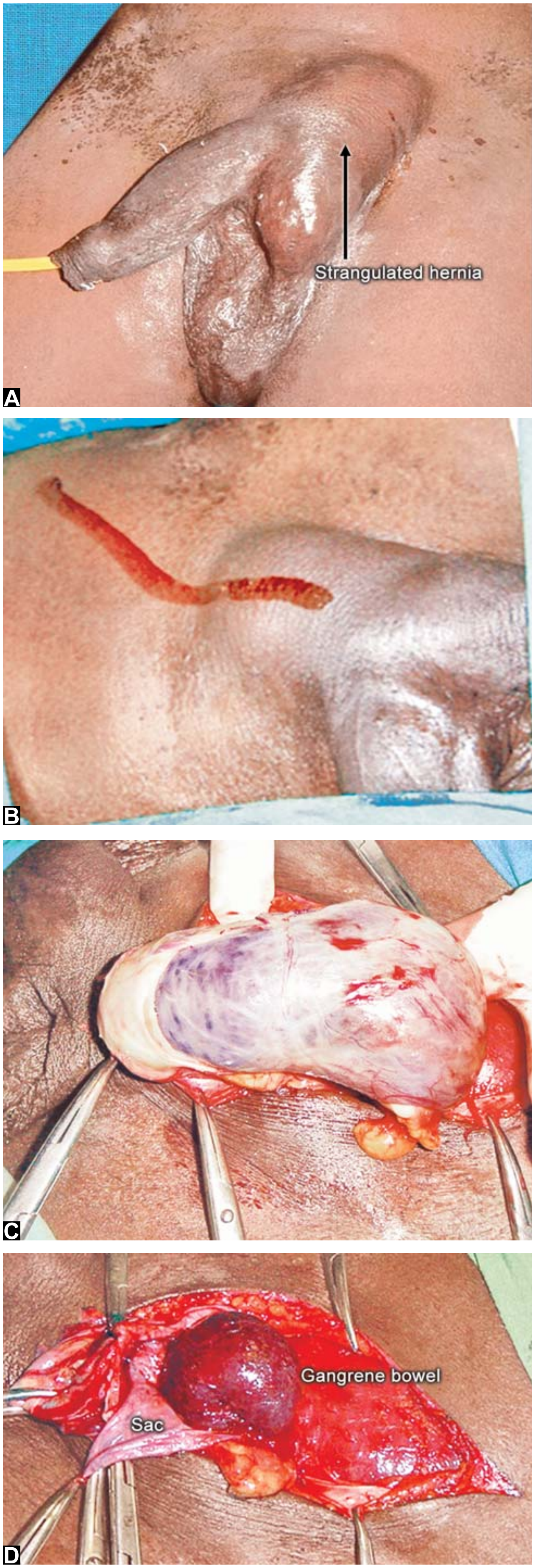

Figs 18.65A to D

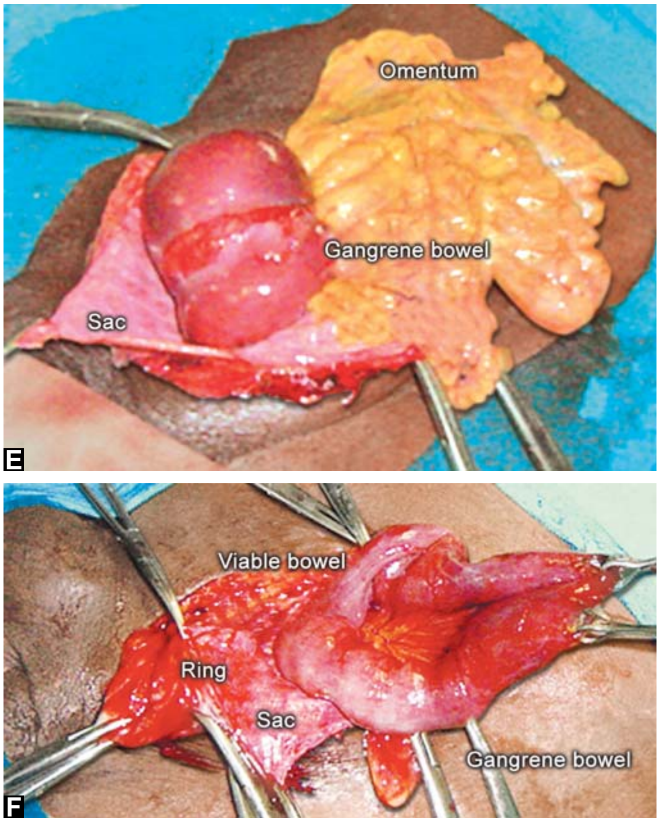

Figs 18.65E and F

Figs 18.65A to F: Strangulated hernia—left side. It is tense, tender, without any impulse on coughing. Note the incision, discoloured sac, content (gangrenous bowel) and proximal viable bowel.

Bassini's repair is done by placing interrupted non-absorbable sutures. Antibiotics, IV fluids are continued. Drain is removed in 4-5 days. Once the bowel movement begins, oral diet is started (in 5 days).

Postoperative Problems

Infection, leak with fistula, septicaemia.

Note:

During surgery for strangulated hernia mesh is usually not used, only repair is done. Biological mesh can be used during the surgery for strangulated hernia.

SLIDING HERNIA (HERNIA-EN-GLISSADE)

- Here posterior wall of the sac is not only formed by the parietal peritoneum, but also by sigmoid colon with its mesentery on left side; caecum on right side and often with portion of the bladder (Both sides) (1:2000).
- Rarely small bowel sliding hernia or sacless sliding hernia can occur (1:8000).
- Content of the sac is usually small bowel or omentum.
- *Sliding hernia occurs exclusively in males. Mainly on the left side* (85%).
- It commonly occurs in *indirect sac* even though femoral and direct sliding hernias are known to occur.
- Bilaterality is extremely rare.
- It is seen commonly in adults and elderly.
- Huge hernia can occur which extends into the scrotum and do not get reduced totally.

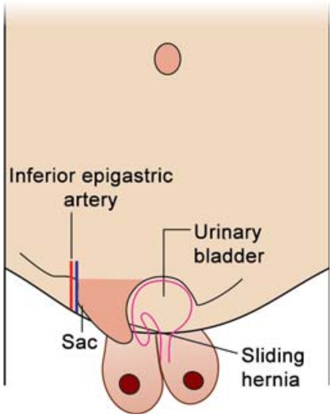

Fig. 18.66: Sliding hernia with urinary bladder as posterior wall of the sac.

Clinical Features (2%)

Large globular swelling in the inguinal region descending into the scrotum, often irreducible.

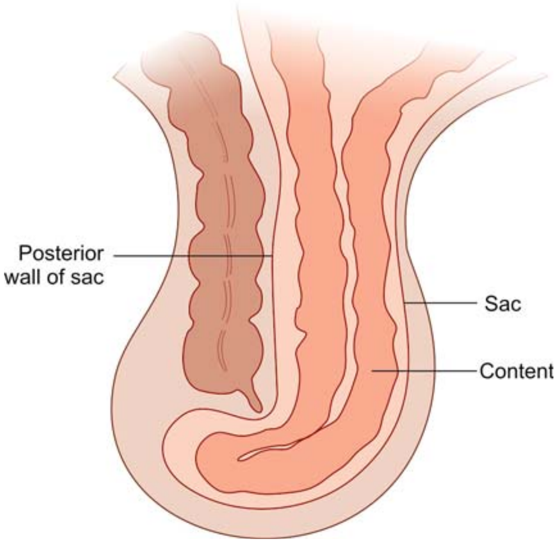

Fig. 18.67: Sliding hernia (*Hernia-en-Glissade*).

Small bowel as the content of sliding hernia can lead to strangulation.

Treatment: *Always Surgery*

- *Posterior wall of the sac should not be separated from large bowel or bladder*. If tried, injury may result to these organs leading to faecal or urinary fistulas.
- Partially excised sac is pushed into the peritoneal cavity with posterior wall and repair is done usually using *prolene mesh* (Hernioplasty).
- In *LaRoque repair* deep ring is incised above and below; reefing of the sigmoid colon mesentery is done which forms the posterior wall of the inguinal canal so that entire sigmoid is replaced into the peritoneal cavity. Posterior wall of the sac should not be separated from large bowel or bladder. If tried it may lead into faecal or urinary fistulas. Often sac can be excised partially; this partially excised sac can be pushed into peritoneal cavity. Right sided sliding hernia will have caecum and appendix in its posterior wall; caecum should not be separated from posterior wall of the sac which may otherwise create faecal fistula. Appendix should not be removed as it may precipitate sepsis. Appendices epiploicae from sigmoid colon should not be removed as there are chances that they may contain small colonic diverticula which may get opened to contaminate the field. Bladder will be present on medial side of the sac and sac should not be separated; if bladder injury occurs it should be sutured in two layers with vicryl. Inside out purse string suture on the opened sac is applied and the sac with its posterior wall is pushed into the abdominal cavity. Urinary catheterisation is a must.
- Truss should not be used in sliding hernia.

PANTALOON HERNIA (DOUBLE HERNIA, SADDLE HERNIA, ROMBERG HERNIA)

- Here both direct and indirect inguinal sacs are present and clinically present *as direct hernia.*
- During surgery, indirect sac may be missed and so leads to *recurrent hernia* through retained (or unidentified) indirect sac.
- Here both medial and lateral sacs *straddle* the inferior epigastric artery.
- It is one of the causes for recurrent hernia.

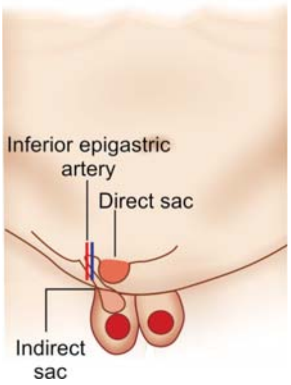

Fig. 18.68: *Saddle/Pantaloon* hernia. It is on either side of the inferior epigastric artery having both direct and indirect sacs.

INFANTILE HERNIA

- Clinically difficult to diagnose.
- Often it is diagnosed on table.
- Processus vaginalis is closed at internal ring and hernial sac either invaginates processus vaginalis as “inverted umbrella’ or comes behind processus vaginalis.

It takes 26 muscles to smile, and 62 muscles to frown.

FEMORAL HERNIA

Surgical Anatomy of Femoral Canal

It is the medial, most compartment of the femoral sheath, which extends from femoral ring above to saphenous opening below. It contains fat, lymphatics, lymph node of Cloquet. It is 1.25 cm long and 1.25 cm wide at the base. Below it is closed by cribriform fascia.

Femoral ring is bounded anteriorly by inguinal ligament; posteriorly by iliopectineal ligament of Cooper, pubic bone and fascia covering the pectineus muscle; medially by concave, sharp lacunar (Gimbernat's) ligament; laterally by a thin septum separating from femoral vein.

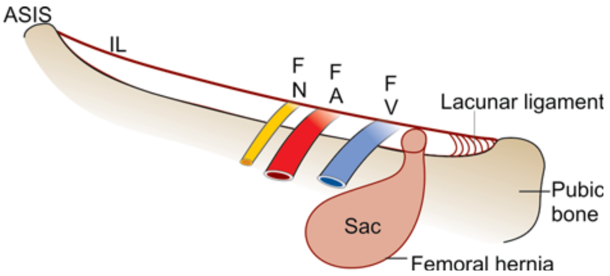

Fig. 18.69: Surgical anatomy of femoral hernia.

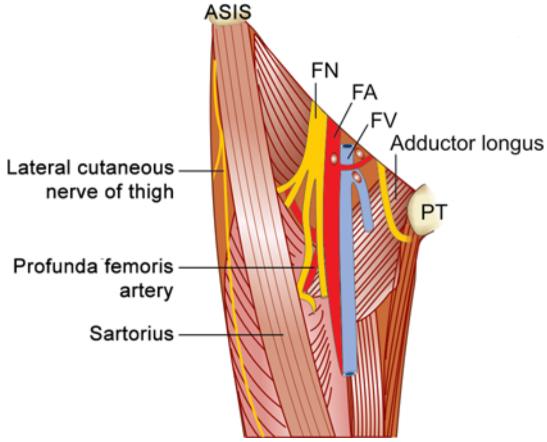

Fig. 18.70: Femoral triangle.

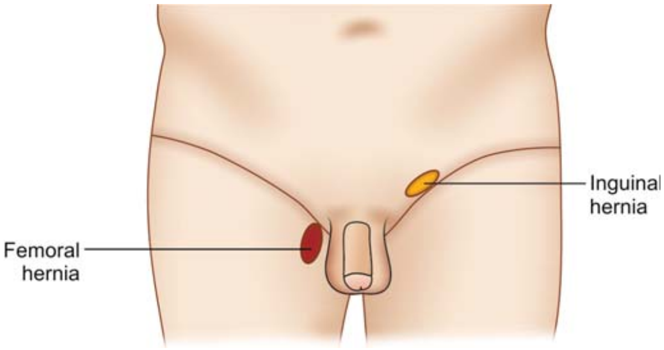

Fig. 18.71: Hernia; inguinal hernia is above and medial to tubercle, and femoral hernia is below and lateral.

Aetiology

- Wide femoral canal.
- Multiple pregnancies.

Pathology in Femoral Hernia

Through femoral canal, hernial sac descends down vertically up to saphenous opening and then escapes out into the loose areolar tissue to expand out like a ***retort.*** Because of its irregular pathway and narrow neck, it is more prone for obstruction and strangulation. During surgery, precaution should be taken about the femoral vein and pubic branch of obturator artery (or accessory obturator artery) which often may get injured leading to torrential haemorrhage.

Clinical Features

- Common in females (2:1 ratio), common in multipara.
- Rare before puberty. 20% occurs bilateral, however, more common on right side.
- Presents as a swelling in the groin *below and lateral* to the pubic tubercle (Inguinal hernia is *above and medial* to the pubic tubercle).
- Swelling, impulse on coughing, reducibility, gurgling sound during reduction, dragging pain, are the usual features.
- When obstruction and strangulation occurs which is more common, presents with features of intestinal obstruction—painful, tender, inflamed, irreducible swelling without any impulse. They also present with abdominal distension, vomiting and features of toxicity.
- Often femoral hernia can be associated with inguinal hernia also.
- 40% of femoral hernias present as emergency hernia with obstruction/strangulation.

Gaur's sign: In femoral hernia, distension of superficial epigastric and/or circumflex iliac veins occurs due to the pressure by the hernial sac.

Differential diagnosis

- Inguinal hernia
- An enlarged Cloquet lymph node of any cause
- Psoas abscess—psoas spasm with flexed hip but difficulty in extension
- Lipoma
- Femoral aneurysm
- Distended psoas bursa (Disappears on hip flexion)
- *Saphena varix*—it is enlarged terminal portion of long saphenous vein. It is soft, disappears on lying down, fluid thrill, impulse on coughing and venous hum on auscultation are present. There is associated varicose veins on leg
- Haematoma in the region

Hydrocele of femoral hernia occurs when adherent omentum which is the content secretes fluid into the sac.

Herniation through a gap in the lacunar ligament (medial) is always strangulated and is called as *Laugier's femoral hernia (L for L).*

In congenital dislocation of hip, femoral hernia occurs *behind* the femoral vessels—*Narath's femoral hernia.*

If sac lies under the pectineal fascia, is called as *Cloquet's hernia.*

Strangulation and Richter's hernia are *common* in femoral hernia.

Often on medial side, a portion of bladder forms the wall of the femoral hernial sac—*sliding—femoral hernia*.

- Femoral hernia—occurs medial to femoral vein
- Narath's hernia—occurs behind femoral artery, in congenital dislocation of hip
- Hesselbach's hernia—occurs lateral to femoral artery
- Serofini's hernia—occurs behind femoral vessels
- Laugier's hernia—through lacunar ligament
- Teale's hernia—in front femoral vessels
- Callison-Cloquet hernia—through pectineal fascia
- Cooper's hernia—femoral hernia with two sacs

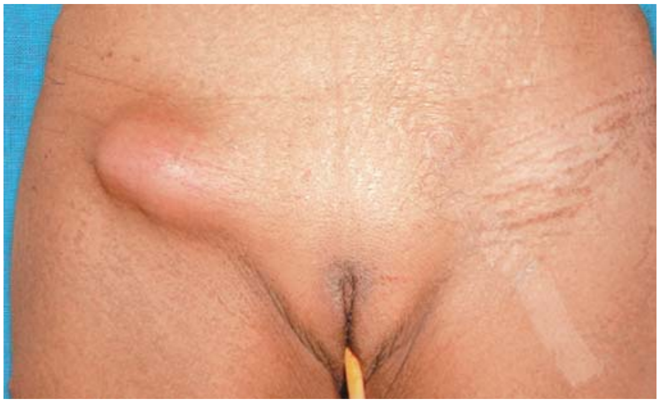

Fig. 18.72: Femoral hernia on right side. Note the location below the inguinal ligament. Femoral hernia is common on right side. It is common in females.

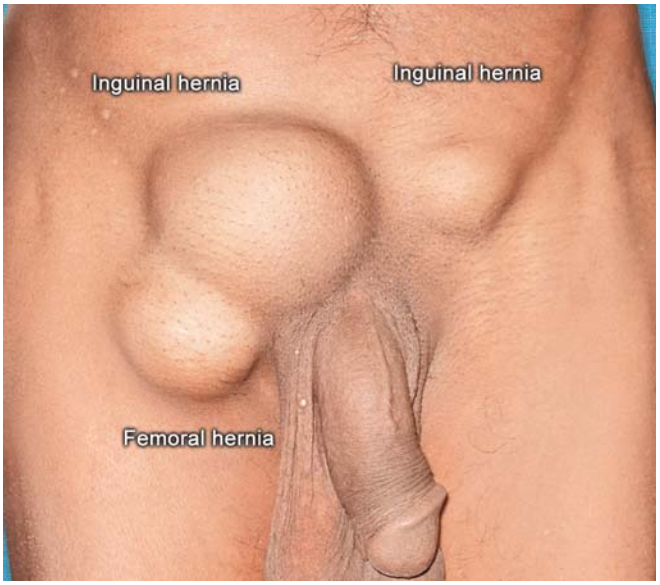
Inguinal hernia
Inguinal hernia
Femoral hernia

Fig. 18.73: Right side femoral hernia in a male patient. Patient is also having bilateral inguinal hernia. Femoral hernia is rare in males.

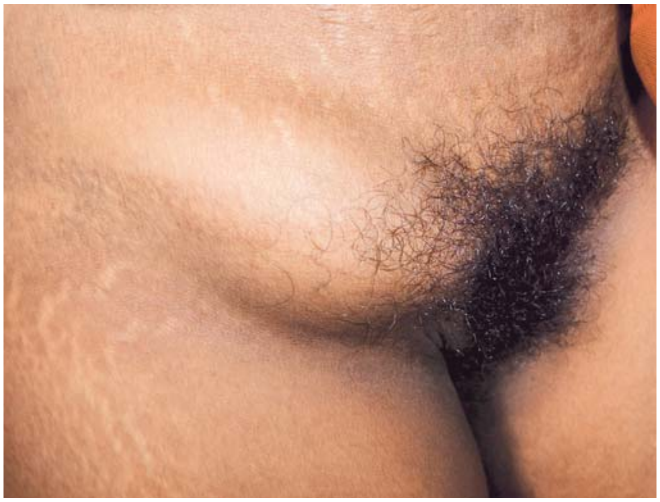

Fig. 18.74: Right side femoral hernia in a female patient. Femoral hernia is common in females. In females the most common hernia is inguinal hernia.

Treatment

Approaches

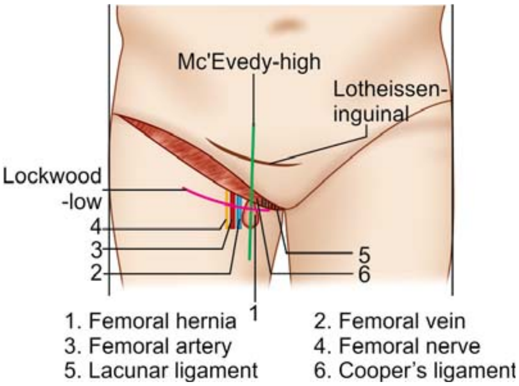
Mc'Evedy-high
Lotheissen-inguinal
Lockwood -low
4
3
2
1
5
6
1. Femoral hernia
2. Femoral vein
3. Femoral artery
4. Femoral nerve
5. Lacunar ligament
6. Cooper's ligament

Fig. 18.75: Different surgical approaches for femoral hernia.

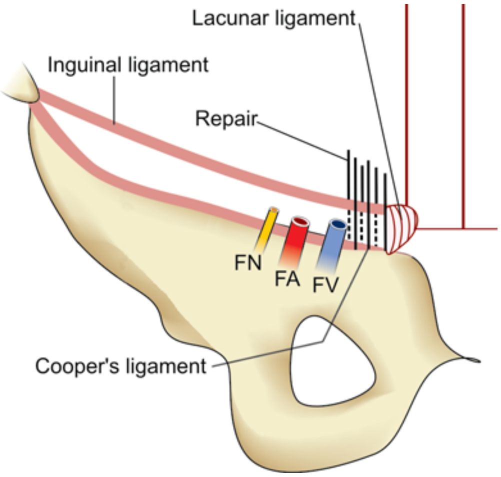
Lacunar ligament
Inguinal ligament
Repair
FN
FA
FV
Cooper's ligament

Fig. 18.76: Femoral hernia repair. Inguinal ligament is approximated to Cooper's (Iliopectineal line) ligament. In Lotheissen's repair conjoined tendon is sutured to iliopectineal line.

1. ***Lockwood-low operation***: Here sac is approached below the inguinal ligament through groin crease incision (or over the swelling) so that fundus of sac is dissected by direct vision and repair is done from below.
   Here inguinal ligament is sutured to Cooper's ligament. standard and ideal (Cooper's ligament repair).
2. ***Mc'Evedy-high operation:*** A incision is made over the femoral canal extending vertically above the inguinal ligament. Sac is dissected from below, neck from above and repair is done from above. It gives a very good exposure of both neck, fundus of sac and repair is also easier. It is done in strangulated femoral hernia.
3. ***Lotheissen's operation***: It is through inguinal canal approach (like for inguinal hernia). Transversalis fascia is opened and neck of the sac is identified in the femoral ring. Sac is dissected from above, neck is ligated and repair is done.
   Lotheissen's repair: After herniotomy, *conjoined tendon is sutured to iliopectineal line* (ligament) by interrupted sutures (2 or 3), using nonabsorbable monofilament sutures. Care should be taken to avoid injury to femoral vein, pubic branch of obturator artery, bladder. It is not as strong as Cooper's ligament repair.
   Complications: Bleeding haematoma, abscess formation.
4. ***AK Henry's approach***: Repair of bilateral femoral hernia through lower abdominal incision.
5. Polypropylene mesh can be buttressed over the femoral canal to close the defect.
6. Laparoscopic mesh repair—TEP/TAPP.

VENTRAL HERNIA

Any protrusion through abdominal wall with the exception of hernia through the inguinofemoral region is defined as *ventral hernia*. Incisional hernia (80%) and primary defects in abdominal fascia which can cause umbilical hernia, epigastric hernia, paraumbilical hernia or Spigelian hernia are grouped under ventral hernia.

Ventral hernia can be:

- Reducible.
- Irreducible.
- Obstructed.
- Strangulated.
- Single.
- Multiple small defects (*Swiss cheese hernia*).

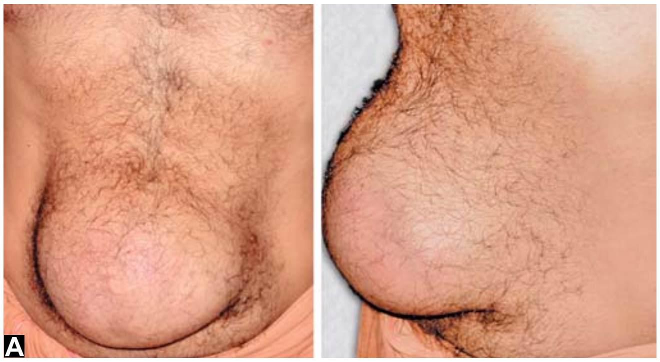

Fig. 18.77A

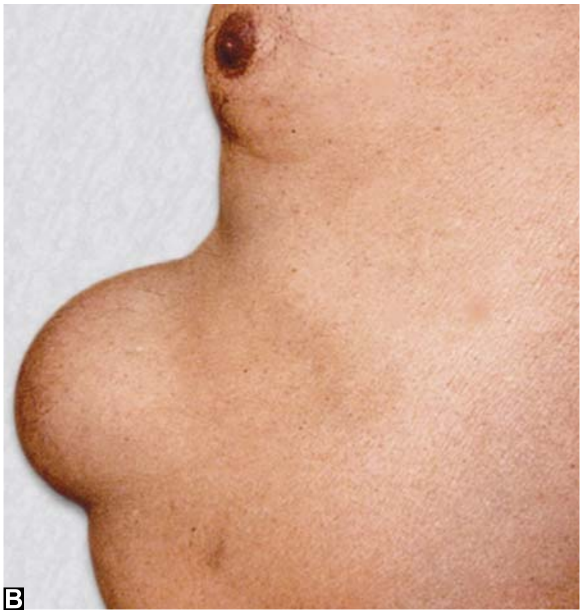

Fig. 18.77B

Figs 18.77A and B: Large ventral hernias. It needs proper mesh repair.

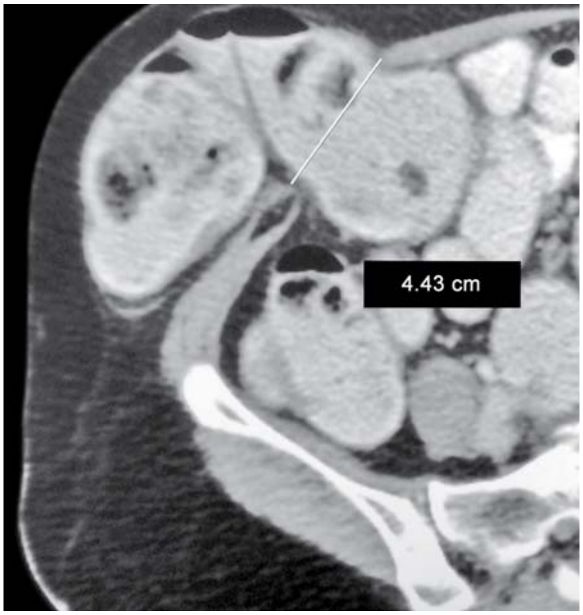

Fig. 18.78: CT scan showing ventral hernia which is irreducible.

Causes
- Congenital defect.
- Obesity.
- Smoking.
- Chronic cough.

Preoperative preparation
- Proper skin hygiene.
- Respiratory care.
- Control of obesity.
- Antibiotics.

Management
- Relevant investigations
- Open mesh repair—inlay; onlay; sublay—retrorectus; underlay.

- Laparoscopic mesh placement—underlay (directly in front of the contents) using dual mesh or four layered mesh.

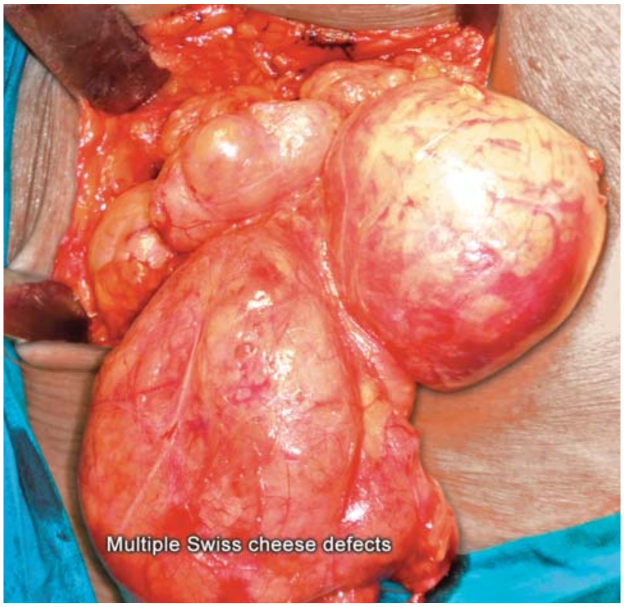

Fig. 18.79: Ventral hernias with Swiss cheese multiple defects.

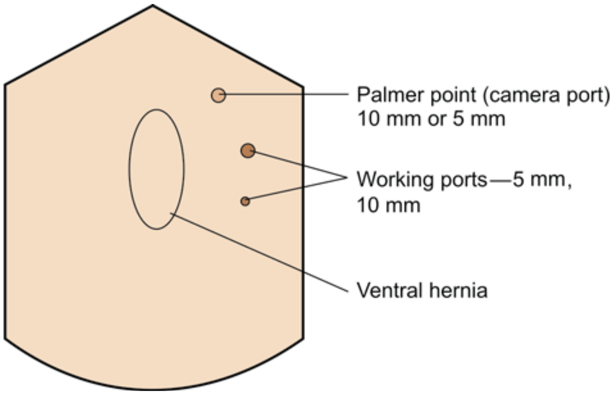

Fig. 18.80: Port placement for laparoscopic ventral hernia repair. Left subcostal point is *Palmer point*.

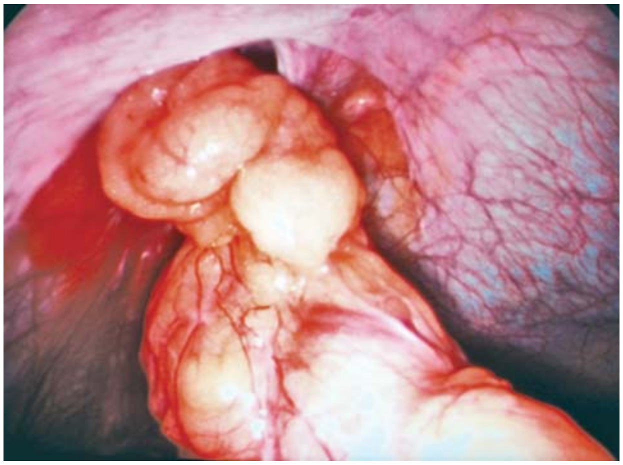

Fig. 18.81: Laparoscopic view of ventral hernia with omentum as irreducible content. Content is released and duval or four-layered mesh is fixed with tackers or sutures.

INCISIONAL HERNIA

- It is herniation through a weak abdominal scar (scar of previous surgery).
- It is common in old age and obese individuals.
- It occurs in 10% of abdominal surgeries; 70% occurs in first 5 years; 30% occurs in 5-10 years.

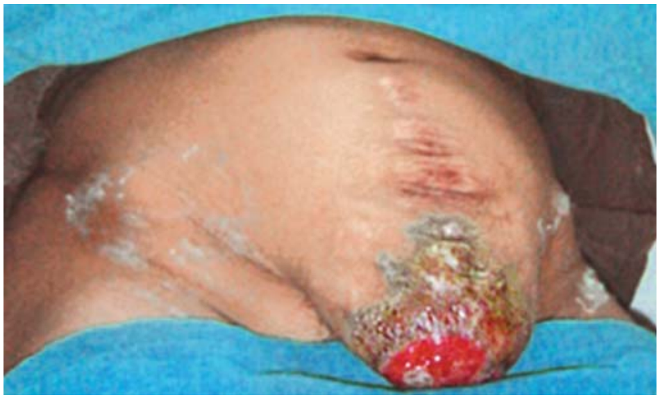

Fig. 18.82: Incisional hernia adherent to skin. Note the ulceration over the summit. Patient underwent resection and anastomosis with removal of the ulcerated skin with mesh placement.

Additional History to be Collected in Incisional Hernia

- Details of surgery that patient has undergone earlier. Duration after how long incisional hernia has occurred is important.
- History of wound infection, wound dehiscence, whether surgery done was an emergency or elective, and tension sutures placed or not.
- History of pain, irreducibility and details of precipitating factors to be asked.
- Other precipitating factors are similar to inguinal hernia like smoking, urinary/respiratory/abdominal symptoms.

Predisposing Factors

- Vertical scar, midline scar, lower abdominal scar—may injure the nerves of the abdominal muscles.
- Scar of major surgeries (biliary, pancreatic).
- Scar of emergency surgeries (peritonitis, acute abdomen).
- Faulty technique of closure.
- Poor nutritional status of the patient.
- Presence of cough, tuberculosis, jaundice, anaemia, hypoproteinaemia.
- Malignancy, immunosuppression.
- Smoking in postoperative period.
- Causes which increases the intra-abdominal pressure (BPH, straining, stricture urethra or rectum, ascites).

Factors responsible for development of incisional hernia

- Vertical incision has got higher chances of incisional hernia than horizontal incision
- Layered closure of the abdomen has got higher chance than single layer

- Continuous closure has got higher chances than interrupted closure
- Using absorbable suture material has got higher chances of hernia than nonabsorbable sutures
- Emergency surgical wound has higher chances than elective surgical wound
- Laparotomy for peritonitis, acute abdomen, and trauma can commonly cause incisional hernia
- Drainage through the main laparotomy wound may precipitate formation of incisional hernia
- Chronic cough, smoking, obstructive uropathy, constipation can precipitate incisional hernia
- Diabetes, old age, malnutrition, malignancy, anaemia, hypoproteinaemia, jaundice, ascites, liver disease, uraemia, steroid therapy, immunosuppressive diseases are other precipitating factors

Clinical Features

- Swelling in the scar region.
- Pain.
- Impulse on coughing.
- Gurgling sound.
- Often bowel peristalsis may be visible under the skin.
- Eventually features of irreducibility, obstruction, strangulation is seen.

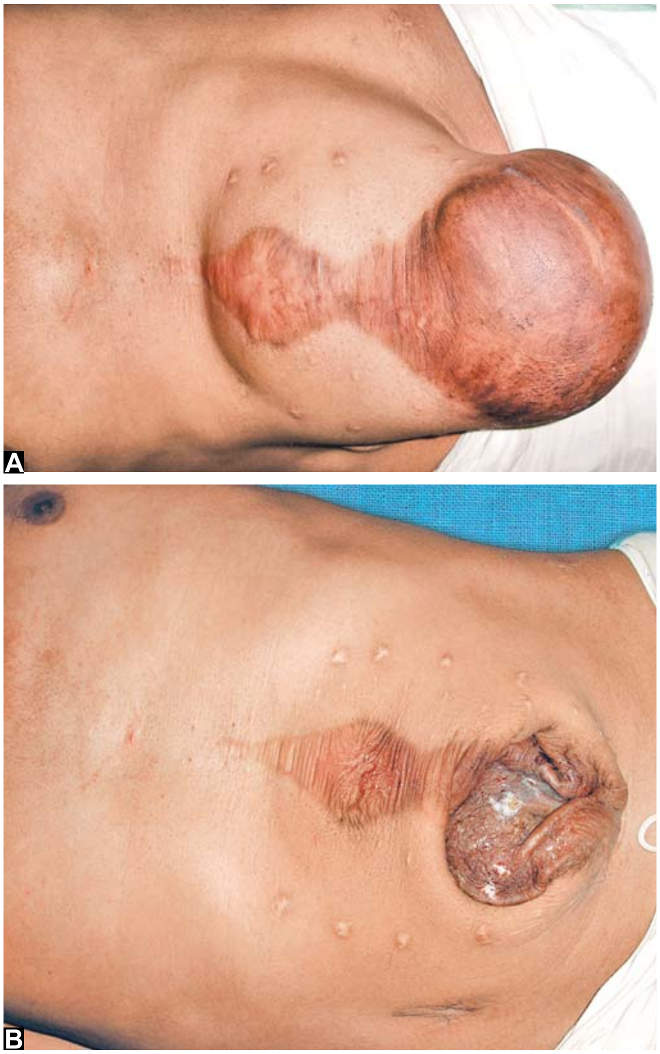

Figs 18.83A and B

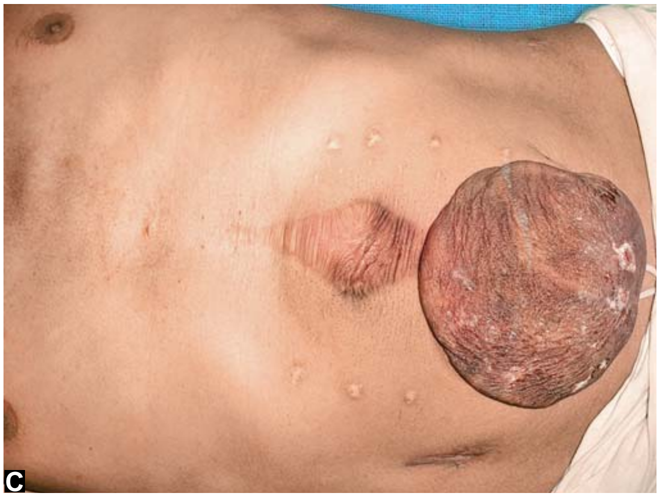

Fig. 18.83C

Figs 18.83A to C: Incisional hernia in a patient during straining and during relaxing.

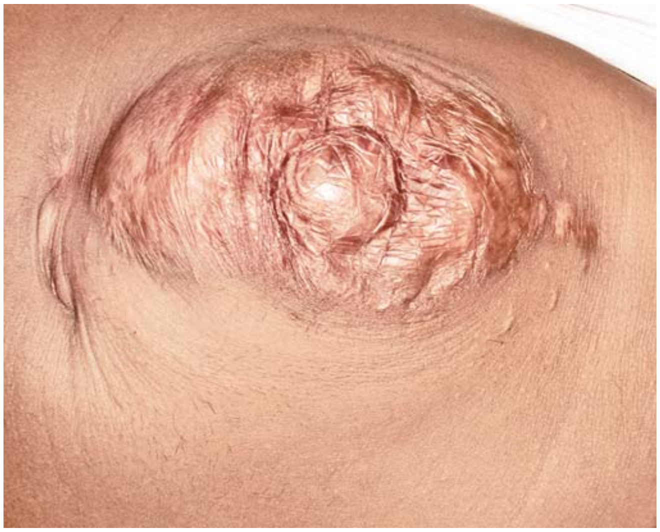

Fig. 18.84: Incisional hernia which is irreducible with visible bowel beneath.

Defect is felt and assessed during head rising, by placing fingers over the scar horizontally
Size of the defect is very important

- Hernia is common in lower abdomen.
- It may be *small or large; huge or massive (diffuse).*
- Scar, its extent and location, whether healed primarily or secondarily, skin over the scar and swelling is noted. Details of the swelling with expansile impulse on coughing and *examination both in lying down and standing are done.*
- *Gap cannot be assessed in an irreducible hernia.*

Type of defects in incisional hernia

- Small defect < 2 cm
- Large and wide defect > 2 cm
- Very large defect
- Massive/diffuse
- Multiple defects—*Swiss cheese pattern*

***Note*:**
- Size of the defect is important to decide the type of surgical closure in incisional hernia.
- Midline hernia expels the content more outwards due to contraction of rectus muscles on both sides.

Investigations: Always the precipitating factors must be looked for:

- Chest X-ray.
- U/S abdomen.
- Tests relevant for causes.

Complications of incisional hernia: Irreducibility, obstruction, strangulation, incarceration

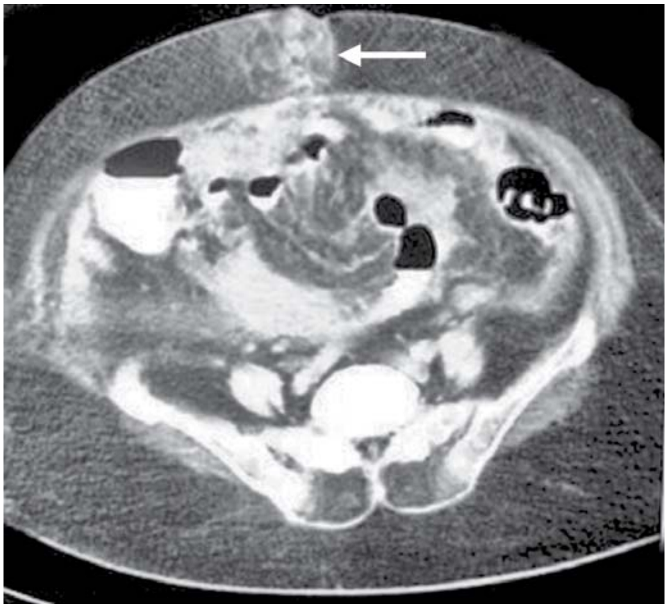

Fig. 18.85: CT scan abdomen showing incisional hernia.

Preoperative Preparations for Incisional Hernia Surgery

- Reduction in weight and control of obesity.
- Nutrition, control of anaemia.
- Treatment for diabetes, hypertension, cardiac diseases, respiratory problems.
- Treating the precipitating causes.
- Chest X-ray, U/ S abdomen to be done.
- Massive incisional hernia after reduction might cause IVC compression, paralytic ileus and diaphragmatic elevation with respiratory embarrassment (*abdominal compartment syndrome*). It is prevented by prior increasing the capacity of peritoneal cavity by creating pneumoperitoneum using CO_2 so as to increase the peritoneal pressure by 12-15 cm of H_2O, daily for 3-6 weeks. Later definitive surgery is done.

Treatment Strategy for Incisional Hernia

- **Mesh repair of the incisional hernia** defect is always better and ideal with less chances of recurrence.
  - Adequate sized mesh is placed either outer to peritoneum *(sublay)*, or outer to musculoaponeurotic abdominal layer *(onlay/overlay)*, or occasionally combined sublay and onlay mesh placement, both deep and outer to musculoaponeurotic layer.
  - *Rive's Stoppa's mesh placement* for incisional hernia is placing mesh between posterior rectus sheath and rectus muscle—**retrorectus mesh placement** is easier for dissection and mesh placement.
  - Commonly polypropylene mesh is used. Other materials used are Dacron, polytetrafluoroethylene (PTFE) mesh, polyglycolic mesh (vicryl mesh) or combined polypropylene and polyglycolic acid mesh (vipro mesh). Drain (suction drain) must be placed after surgery.
  - Mesh repair is ideal approach for incisional hernia. Prevention of infection/haemostasis and drainage are important.
- **Laparoscopic mesh repair is done for incisional hernia** by placing a mesh under the defect laparoscopically in intraperitoneal plane. The only problem of this underlay placement is chances of adhesion and GI fistula formation, but still it is found to be safer. Laparoscopic preperitoneal mesh placement is also done for smaller defects. Now dual mesh (PTFE) or four layered mesh are available. Here mesh is placed under the peritoneum deep to the defect after reducing the contents. Mesh is fixed with sutures and tacks. In four-layered mesh, deepest 1st layer is absorbable cellulose which allows new peritoneum to creep underneath. 2nd layer is PDS/PTFE mesh 3rd layer is polypropylene mesh and the last 4th layer is again PDS/PTFE mesh. It is ideal but costly.
- **Cattell's operation**: When the defect is less than 3 cm, and if the patient is having adequate abdominal muscle tone, *layer by layer anatomical repair* is done using monofilament nonabsorbable suture material like *polypropylene/ polyethylene,* ideally with interrupted sutures. Sac should be dissected, ligated and excised prior to repair. Peritoneum and posterior rectus sheath is apposed as first layer and anterior rectus sheath as second layer.
- **Double breasting of the rectus sheath** using interrupted nonabsorbable sutures using monofilament suture material. It is overlapping the rectus sheath in two layers with two rows of sutures.
- **Keel operation** is done in large defect. Scar is excised and is dissected beyond the margin of the defect. Sac is never opened unless there is obstruction of the content. Sac in inverted using continuous/interrupted inverting nonabsorbable sutures, layer by layer until the defect margins are apposed together which is then again sutured with interrupted sutures. Keel is inverted beam of the ship.

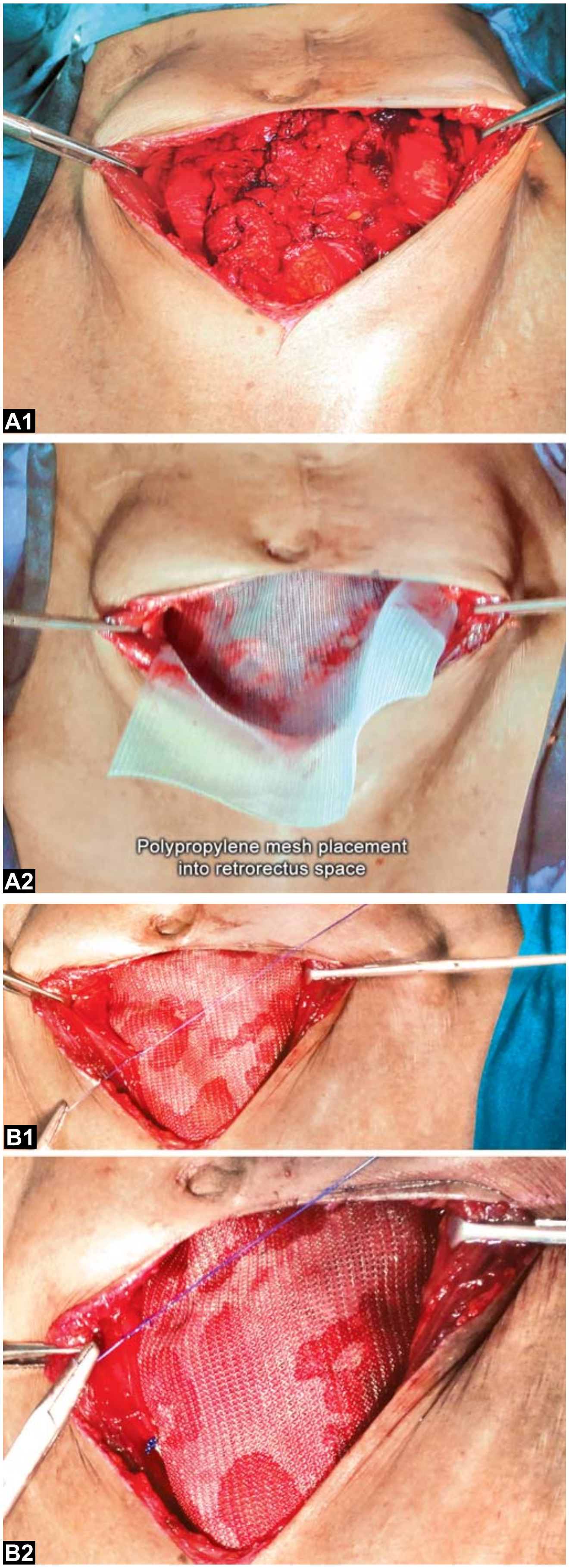

Figs 18.86A and B: On table incisional hernia defect also showing mesh placement in retrorectus plane.

- **Nuttall's operation** is done for lower midline incisional hernia. Recti attachments are detached from the pubic bones and are crossed over to fix to opposite pubic bones so as to create a firm abdominal wall support by crossed recti muscles.

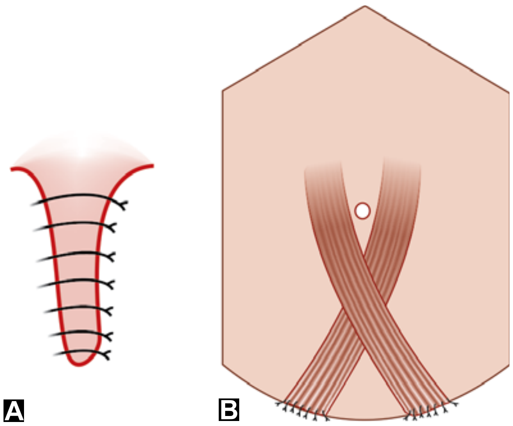

Figs 18.87A and B: (A) Keel operation and (B) Nuttall's operation.

Different Types of Mesh Repair for Incisional Hernia

- *Outer to peritoneum is ideal method (sublay):* Large sized mesh is placed in preperitoneum. It need not be fixed as abdominal pressure keeps it in position (according to Pascal's law).
- *Under the peritoneum, directly over the content (underlay):* Now it is accepted but there are chances of adhesions/fistula formation. It is used in laparoscopic repair. Dual mesh/four-layered mesh is used.
- *Overlay* mesh placed outer to musculoaponeurotic layer. Here mesh is placed under subcutaneous tissue; it carries high recurrence rate (30%). So it is *not recommended.*
- *Combined inlay and overlay* with two layers of mesh.
- *Rive's Stoppa's method* of placing mesh between posterior rectus sheath and rectus muscle widely. Dissection is done at least 10 cm beyond the defect under rectus muscle in front of the posterior rectus sheath; mesh should cover 5 cm beyond the defect all around. Recurrence rate with this is less than 10%.
- *Components separation technique* is better method in large defects. Here skin and subcutaneous fat are dissected of the anterior rectus sheath → external oblique is incised 2 cm lateral to rectus abdominis → external oblique is separated from internal oblique until posterior axillary line → additional relaxing incisions are made over rectus sheath, internal oblique, transverse abdominis → it is repeated on opposite side → 20 cm mobilisation is achieved → also reinforced with a large mesh. Component separation technique causes often lateral bulges on both sides. If it is done without mesh support recurrence rate is high. Advantage is defect up to 20 cm can be easily brought together. Technique is also called as *Autologenous repair by vascularised innervated muscle flaps) (Ramirez, 1990).*

- *Bridging at myoaponeurotic level outer to peritoneum (inlay):* Adequate sized mesh is placed in the defect level bridging the gap to cover preperitoneum. It is better by principle but still shows high recurrence.

Surgeries for incisional hernia

- Mesh repair—commonly done and ideal
- Component separation technique
- Layer by layer closure—Cattell's operation
- Double breasting of the rectus sheath
- Keel operation—not commonly used now
- Nuttall's operation—not commonly used now

Postoperative Care

- Antibiotics.
- Analgesics.
- Nasogastric aspiration.
- Abdominal binder for support.
- Prevention of paralytic ileus.
- Control of obesity and other precipitating factors.
- Stop smoking and treat other associated causes.
- Early ambulation.
- Fluid management, catheterisation.
- Drain should be kept until drainage becomes minimal.
- Abdominal binder is used to support abdominal wall during recovery period.

Complications of incisional hernia surgery are:

- Wound infection, seroma formation
- Paralytic ileus, abdominal compartment syndrome in large hernias
- Wound sinus, enterocutaneous fistula
- Infection of the mesh, recurrence

Additional Problems in Large Incisional Hernia

- While reducing the bulky contents like bowel and omentum, inadequate intra-abdominal capacity leads to increased intra-abdominal pressure causing IVC compression, mesenteric oedema followed by stasis of splanchnic bed, paralytic ileus, diaphragmatic elevation and respiratory distress (*abdominal compartment syndrome*), urinary and bowel disturbances. Abdominal capacity can be raised by creating regular pneumoperitoneum over the period of 3-6 weeks.
- Lordosis and back pain may be presenting feature.
- Sac and contents may get adherent to the thin skin over the summit of the hernia leading to skin ulceration and occasionally fistula formation.
- Often might need resection of the adherent bowel segment.
- Large mesh placement is required.

Note:

- *It is now universally accepted that prosthetic repair is gold standard for all incisional hernia.* Prostheses used are polypropylene mesh, e PTFE, Dacron, dual intraperitoneal mesh, biologic grafts.
- Nonprosthetic repair has got only historical value as recurrence rate is very high (as high as 55%).

UMBILICAL HERNIA

Anatomy of Umbilical Region

Umbilical ring is a complex structure which is related to linea alba, falciform ligament, obliterated urachus (median umbilical ligament) and umbilical fascia (*Richet's fascia)*. It is located at the level of L_4 and L_5 vertebral disc. It is at lower position in infants. It is water shed area for venous and lymphatic drainage—above umbilicus it drains to axillary vein or lymphatics; below to inguinal area. Umbilical skin is supplied by T_{10} spinal cord. It is the meeting point of four folds of embryonic plate and three systems—GI (vitellointestinal), urinary (urachus), and vascular (umbilical vessels). Umbilical hernia develops due to either absence of umbilical fascia or incomplete closure of umbilical defect. Weakest part in the umbilical cicatrix is upper part where hernia begins.

- Umbilical hernia can be **congenital** in newborn and infants (common in males) or **acquired** in adults (common in females). Congenital umbilical hernia is *common* in Africa or in African origin people (8 times). Acquired is like any ventral hernia.
- It is herniation through a weak umbilical scar (cicatrix).
- It is common in infants and children, occurs commonly due to neonatal sepsis.
- Male : female :: 2:1.
- It is seen in 20% of newborn infants.
- Umbilical hernia is common in Down's syndrome, *Beckwith-Weidman syndrome*.

Clinical Features

- Presents with a swelling in the umbilical region within first few months after birth, the size increases during crying. It is hemispherical in shape.
- Defect can be felt with finger during crying.
- Occasionally it can go for irreducibility and obstruction which presents with pain, distension, vomiting.

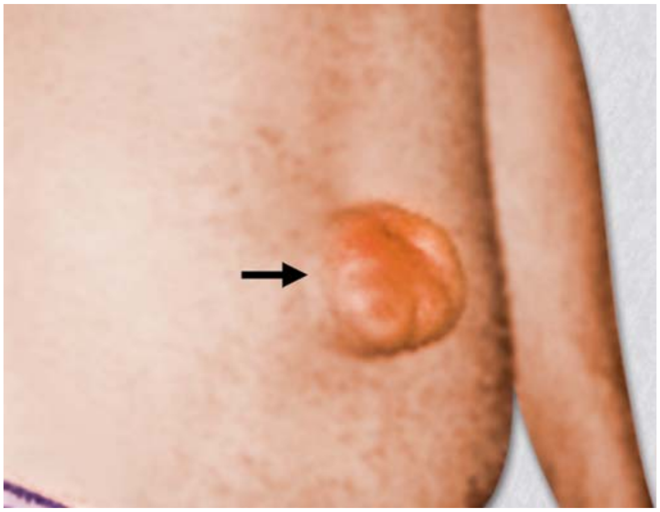

Fig. 18.88: Umbilical hernia.

Those who can't hear the music think the dancer's mad.

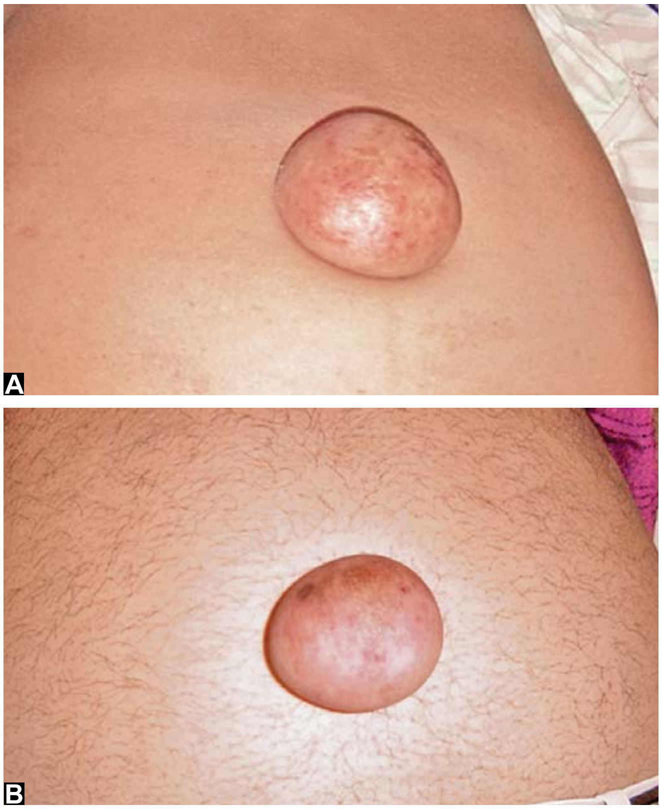

Figs 18.89A and B: Umbilical hernia. Often it can attain large size.

Treatment

Initially conservative. In 93 to 95% of cases, it disappear spontaneously in few months after birth (masterly inactivity). It can be hastened by adhesive strapping across the abdomen.

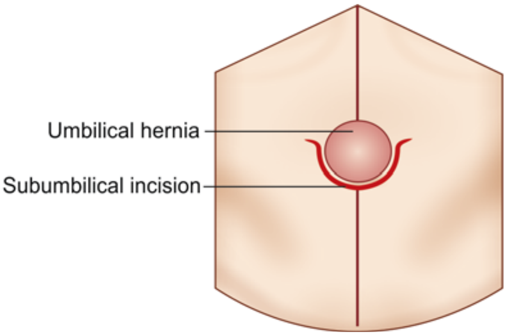

Fig. 18.90: Incision for umbilical hernia.

Indications for Surgery

- If persists even after the age of two years.
- If the defect is more than 2 cm in size.
- Acquired/adult umbilical hernia.

Different surgeries are:

Primary closure of the defect:

An infraumbilical incision is made encircling its lower half. Sac is dissected circumferentially and is released off from the umbilicus and subcutaneous tissue. Sac is opened; contents are reduced; excess part is excised up to the umbilical ring. Defect is closed with interrupted nonabsorbable polypropylene sutures.

Mayo's operation:

Here horizontally opened rectus sheath is approximated as double breasting with upper flap overlapping lower flap in front. It is rarely done today.

Sublay mesh repair:

In a large umbilical hernia (> 3 cm size defect) with a degenerated skin on its surface it is often difficult to retain the umbilicus. When umbilicus is tried to be saved, infraumbilical incision should extend laterally about 2 cm on each side at 3 and 9 o'clock positions. Sac is dissected similarly. Sac is excised after excision of redundant sac. Presently it is standard to use polypropylene mesh as sublay or in retrorectus position and then rectus sheath is closed.

Umbilectomy:

Unhealthy thin skin over the large umbilical hernia is a real problem. It is better to do umbilectomy in these patients (excision of umbilical cicatrix). It is done only in adult with large umbilical hernia with thinning of umbilical skin. Prior consent and eventual creation of umbilicus is needed.

Open dual PTFE and polypropylene mesh placement:

Umbilical hernia is dissected similarly through subumbilical incision. Redundant sac is excised. Peritoneum is not closed. A special composite mesh containing wider PTFE on the inner side with little smaller polypropylene mesh on the outer aspect is used. It has also got two additional straps of the mesh attached to the outer part of the main composite mesh which are used to hold and later to fix into the defect anteriorly. This has got excellent results with low recurrence rate.

Laparoscopic umbilical hernia repair:

It is similar like any ventral hernia repair through laparoscopy which is done under general anaesthesia. It is useful only in large umbilical hernia.

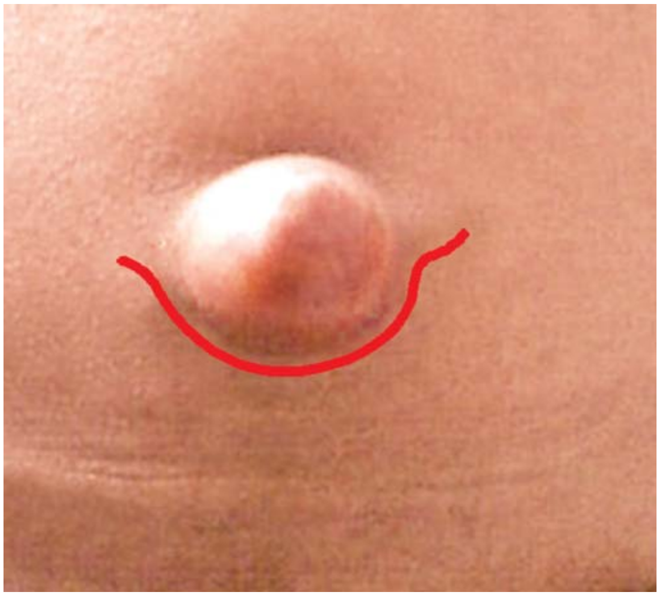

Fig. 18.91: Umbilical hernia. It needs surgical intervention.

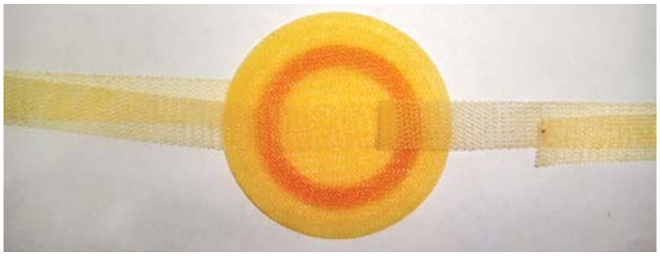

Fig. 18.92: Umbilical hernia dual mesh to place both intraperitoneally and retrorectus plane with straps to fix.

PARAUMBILICAL HERNIA (Supra- and Infraumbilical Hernia)

- It occurs commonly in adults. It is a protrusion or herniation through linea alba, just above or below the umbilicus.
- It enlarges ovally, often attains a large size and sags downwards.

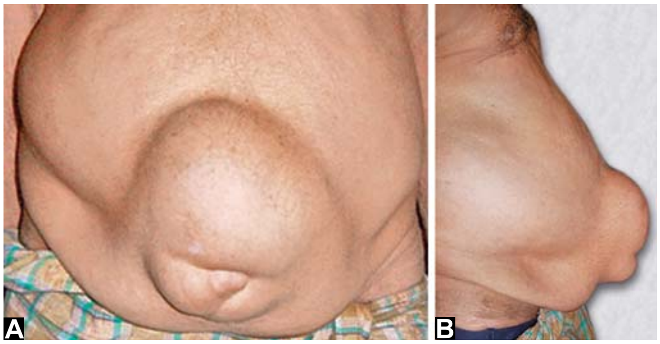

Figs 18.93A and B: Large ventral hernia front and side views. It needs preperitoneal mesh repair.

- Neck of the sac is relatively narrow. Contents are usually omentum, small bowel, sometimes large bowel.
- It has got tendency to go for adhesion, irreducibility and obstruction.

Predisposing factors

- Obesity
- Multiple pregnancies
- Flabby abdominal wall

Clinical Features

- Common in females (5:1 ratio).
- It presents as a swelling which has smooth surface, distinct edges, soft, resonant with dragging pain and impulse on coughing. Large hernias can present with intestinal colic due to subacute intestinal obstruction. Eventually strangulation can occur.

Treatment

Is always surgery

- Dissection of hernial sac and placement of mesh in retrorectus plane and under the umbilicus is the ideal treatment.
- Often *umbilectomy* is required and also *mesh placement is beneficial* (when defect is > 4 cm in size).
- If there is strangulation, resection of bowel segment and anastomosis is done followed by repair of the hernia.
- ***Mayo's operation*:** Through a transverse elliptical incision, sac is identified and dissected. Herniotomy is done. *Double breasting of the defect* in the rectus is done by interrupted nonabsorbable sutures.
- Additional lipectomy (panniculectomy) may be done in case of pendulous abdomen.
- Postoperative weight reduction, use of abdominal binder is required for these patients.

Mayo's operation

- It is done for umbilical and paraumbilical hernia
- Once lower flap or umbilicus is raised above, sac is identified, dissected and opened
- After reducing the contents sac is transfixed using vicryl
- Rectus sheath is repaired with double breasting using nonabsorbable sutures between superior and inferior fascial margins as 'vest on pant' imbrications
- Skin flap is closed with a drain. Infection, recurrences (30%) are known complications
- It is not commonly used now

EPIGASTRIC HERNIA (Fatty Hernia of Linea Alba)

- It is 10% common; common in males.
- 20% of epigastric hernias are multiple—Swiss *cheese like*.
- It occurs usually through a defect in the decussation of the fibres of linea alba, any where between xiphoid process and umbilicus.
- Extraperitoneal fat protrudes through the defect as fatty hernia of the linea alba presenting like a swelling in the upper midline with an impulse on coughing.
- *It is sacless hernia*. Later protrusion enlarges and drags a pouch of peritoneum, presenting as a *true epigastric hernia*.
- If the defect is less than 1.5 cm, lateral margin of the defect is formed by only anterior and posterior lamina of the rectus sheath; if the defect is > 1.5 cm, then lateral margin is also formed by rectus muscle.
- Content of true epigastric hernia is usually omentum, sometimes it may be small bowel.
- Common in muscular men; manual labourers.

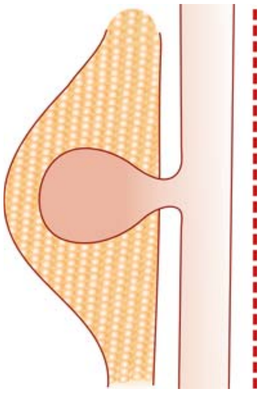

Fig. 18.94: Epigastric hernia.

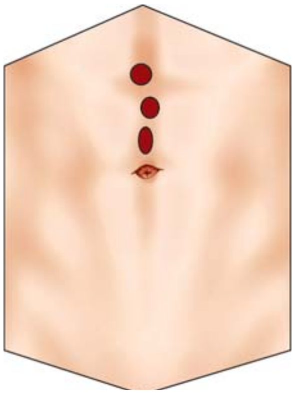

Fig. 18.95: Epigastric hernia with *Swiss cheese defects*.

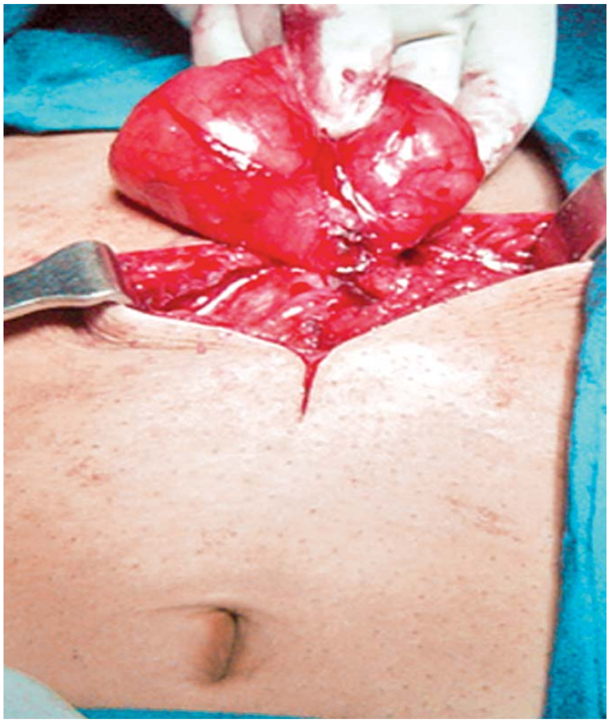

Fig. 18.96: Epigastric hernia. Note the location of epigastric hernia above the umbilicus. Initially it is sacless but later sac with content develops.

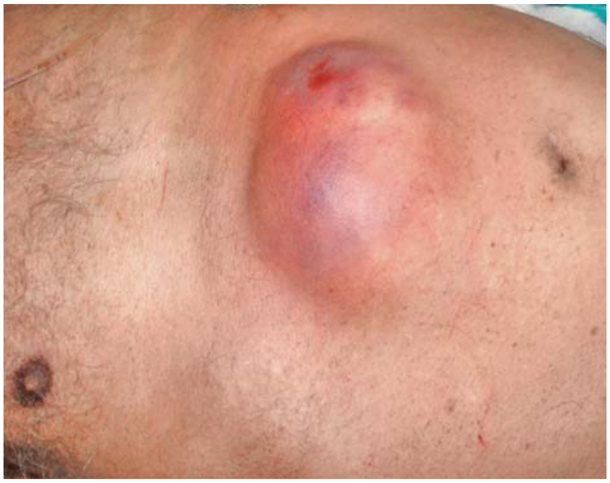

Fig. 18.97: Epigastric hernia showing features of strangulation.

Clinical Features

- *Often asymptomatic.*
- Swelling in the epigastric region which is tender.
- Pain in epigastric region. It is often associated with peptic ulcer and so pain may be due to peptic ulcer. So gastroscopy is done to rule out acid peptic disease.
- Impulse on coughing; defect in the epigastric region are also found.
- Irreducibility, obstruction, strangulation as seen in any other hernia can also occur in epigastric hernia.

Treatment

- Through a vertical incision, sac is dealt with. Defect is closed with nonabsorbable interrupted sutures.
- Large defect is supported with preperitoneal mesh.
- Complete reconstruction of linea alba is needed from xiphisternum to umbilicus especially in *Swiss cheese* type using different methods like—interrupted primary closure using polypropylene sutures; *modified shoelace technique* is used after removing strip of medial margins of the linea alba; double breasting of the linea alba.

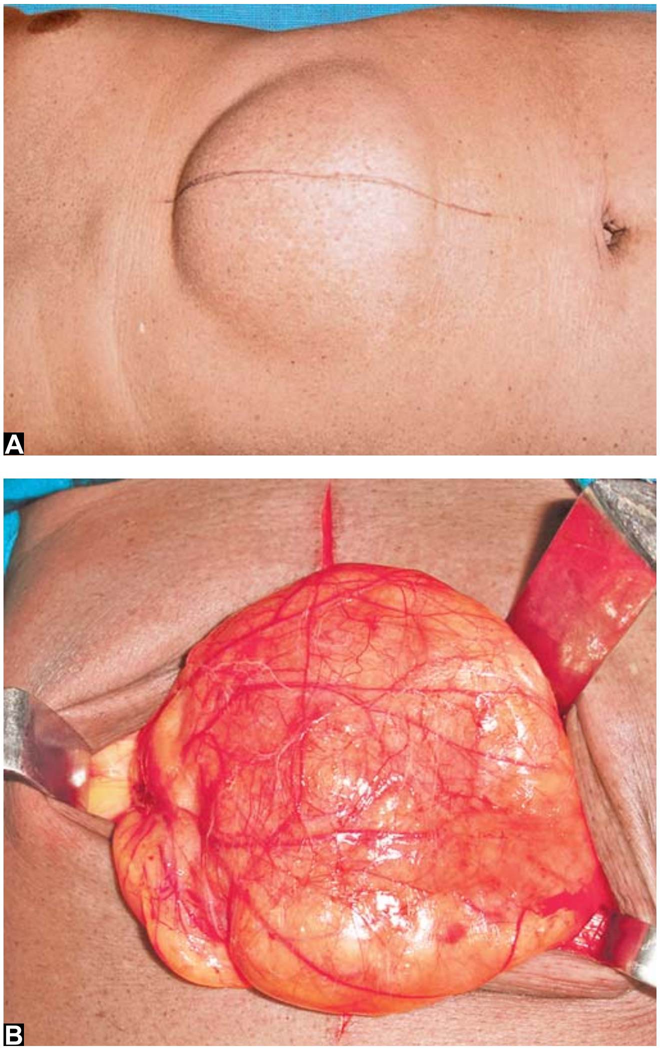

Figs 18.98A and B

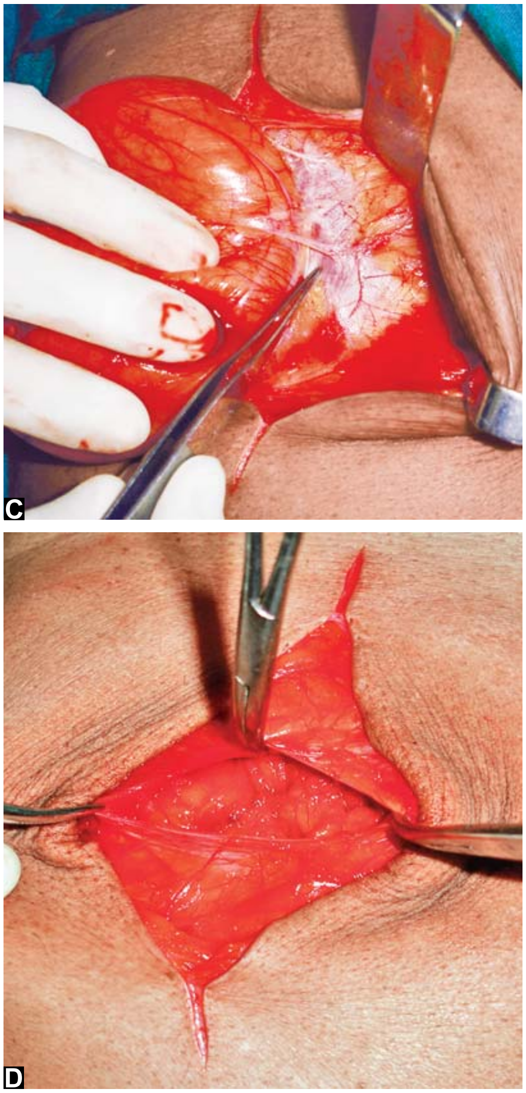

Figs 18.98C and D

Figs 18.98A to D: Epigastric hernia—incision, sac and defect.

Interparietal hernias/interstitial hernias

- Herniation through parietal peritoneum into various layers of the abdominal wall
- Common in Down's syndrome, Prune Belly syndrome
- Often it can attain large size
- May mimic abdominal wall lipoma; haematoma
- As neck of the sac is often narrow, can present with irreducibility or obstruction
- Commonly it is deep to external oblique aponeurosis
- *Types:*
  - Preperitoneal—between peritoneum and transversus abdominis muscle—20%
  - Interparietal/intermuscular—between external oblique and internal oblique; most common—60%. It is commonly associated with inguinal hernia
  - Extraparietal (inguinosuperficial)—herniates through external oblique aponeurosis into subcutaneous plane—20%
- U/S; X-ray abdomen; often CT confirms the diagnosis
- Through a transverse incision, surgical ligation of sac, repair or mesh placement is the treatment

SPIGELIAN HERNIA

- It is a type of interparietal hernia occurring at the level of arcuate line through spigelian point. Hernial sac lies either deep to the internal oblique or between external and internal oblique muscles.
- It is lateral ventral hernia through *Spigelian fascia* at any point along its line. *Semilunar line of Spigel* is a line from pubic tubercle to tip of 9th costal cartilage. It marks the lateral margin of the rectus sheath. Semicircular arcuate line (fold) of Douglas is lower end of posterior lamina of rectus sheath below the umbilicus and above the pubis. Spigelian fascia is area between lateral border of the rectus muscle and external and internal oblique and transversus abdominis muscle.
- Spigelian hernia can occur above (10%) or below (90%) the umbilicus. Below the umbilicus it occurs at the junction of linea semilunaris and linea semicircularis (wider and weaker point). In Spigelian hernia, defect is formed by internal oblique and transversus abdominis muscle. External oblique is outer to the hernial sac.

Clinical Features

- Presents as a soft, reducible mass lateral to the rectus muscle and below the umbilicus, with impulse on coughing. Strangulation is common in spigelian hernia.
- Presence of precipitating factors like obesity, chronic cough, old age, multiple pregnancies.
- Common in females after 50 years of age.

Differential Diagnosis

- Abdominal wall lipoma.
- Soft tissue sarcoma.
- Abdominal wall haematoma.

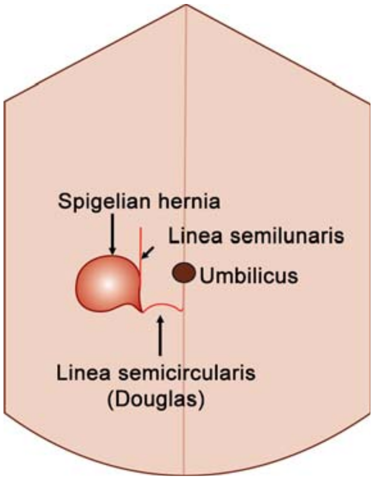

Fig. 18.99: Spigelian hernia.

Investigation

- Ultrasound abdomen.

The wind of anger blows out the lamp of intelligence.

Treatment

- Through a lengthy transverse incision herniotomy and later closure of the defect layer by layer using nonabsorbable interrupted sutures.
- But *ideally* mesh is required to cover the defect properly.
- Laparoscopic dual mesh placement is also useful.

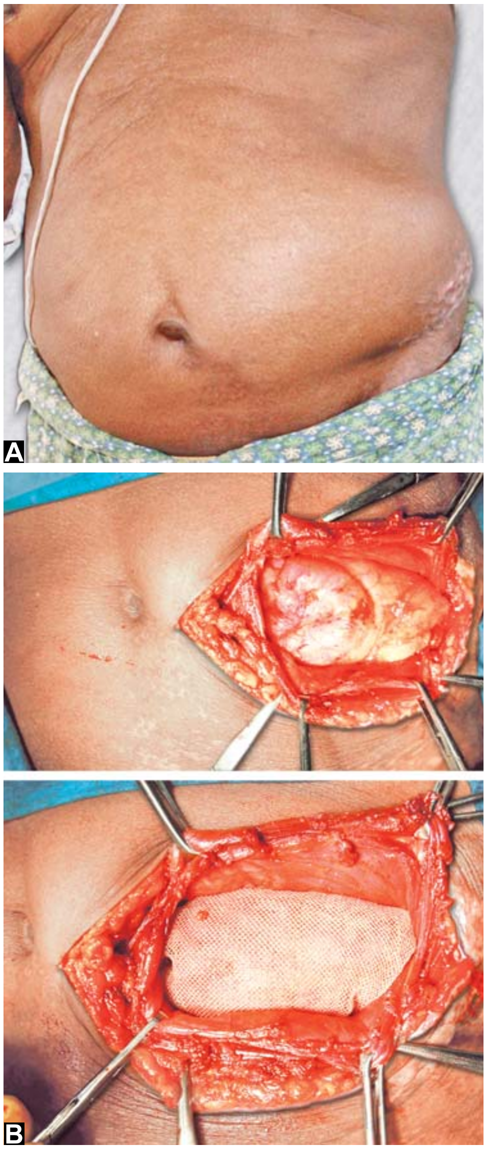

Figs 18.100A and B: Spigelian hernia. On table finding and placement of mesh in same patient.

OBTURATOR HERNIA

- It is hernia occurring through obturator canal between superior ramus of pubis and obturator membrane. It is a rare entity, seen in elderly females (6::1 ratio female to male).

Clinical Features

- Usually presents with features of intestinal obstruction (85%) and more often confirmed only on laparotomy.
- Rarely seen as a swelling in Scarpa's triangle (20%) deep to the pectineus muscle, with limb in flexed and abducted position. Movement of limb is painful.
- Referred pain in knee joint through geniculate branch of obturator nerve signifies not only obturator hernia but also strangulation—*Howship-Romberg sign* (50%).
- On per vaginal examination, tender swelling is felt over the obturator foramen.
- Here strangulation is usually of Richter's type.

Treatment

- Laparotomy is done and the sac is identified. It is dissected and ligated. If strangulation is present (common), resection and anastomosis is done. Broad ligament is stitched over the opening to prevent recurrence.
- Mesh placement is the ideal way of repairing the obturator defect.
- *In nonobstructive type,* if diagnosed clinically and by CT scan imaging, either TEP (laparoscopic) or lower abdominal midline extraperitoneal approach may be the method to manage the obturator hernia.
- In strangulated hernia, rarely additional exposure in thigh often may be required to reduce the content through a vertical medial pectineus muscle split incision.

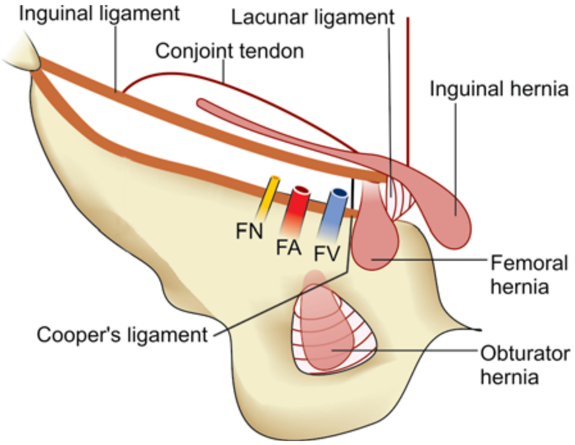

Fig. 18.101: Obturator hernia through obturator canal in obturator foramen. Note also the locations of inguinal and femoral hernias.

RICHTER'S HERNIA

It is a hernia in which the sac contains only a portion of the circumference of the intestine (small bowel). It is usually *seen in femoral and obturator hernia.*

Clinical Features

- It *mimics gastroenteritis* with pain abdomen, diarrhoea, toxicity, vomiting.
- There are no features of intestinal obstruction.
- Constipation does not occur in Richter's hernia.
- Gangrene (strangulation) of a part of bowel occurs, eventually leading to peritonitis.

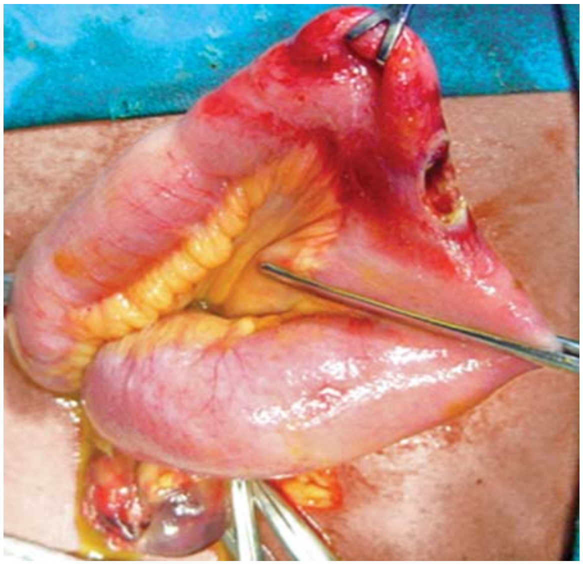

Fig. 18.102: Strangulated femoral hernia, Richter's type with perforation. Patient underwent resection and anastomosis.

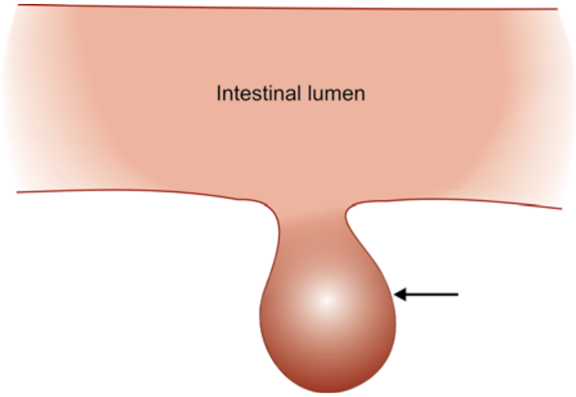

Fig. 18.103: Richter's hernia.

Treatment

- Resection and anastomosis is done.
- The type of hernia is treated.
- Mortality increases with delay in surgical intervention.

LUMBAR HERNIA

- It is herniation either through superior or inferior lumbar triangle.
- Superior lumbar triangle *(Grynfelt's/Lesgaft's triangle)* is bounded by sacrospinalis, 12th rib and posterior border of internal oblique.
- Inferior lumbar triangle is bounded by latissimus dorsi, external oblique and iliac crest *(triangle of Petit).*
- Lumbar hernia is more common through superior lumbar triangle.
- It can be:
  - **Primary.**
  - **Secondary,** which is due to previous renal surgery, more common.

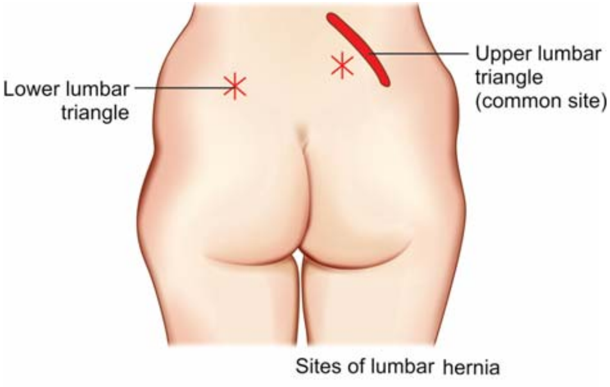

Fig. 18.104: Lumbar triangles.

Differential Diagnosis

- Lipoma.
- Cold abscess.
- Lumbar phantom hernia.

Treatment

Repair using fascial flaps or mesh.

PHANTOM HERNIA

- It is a muscular bulge as a result of local muscular paralysis due to interference with nerve supply of the affected muscles, like poliomyelitis. It is common in lumbar region. It is often seen in lower abdomen.

Phantom hernia is muscular bulge due muscle paralysis in lumbar region
Phantom limb is feeling of pain in amputated toe or limb
Phantom tumour is tumour like lesion in lung like interlobar pleural effusion

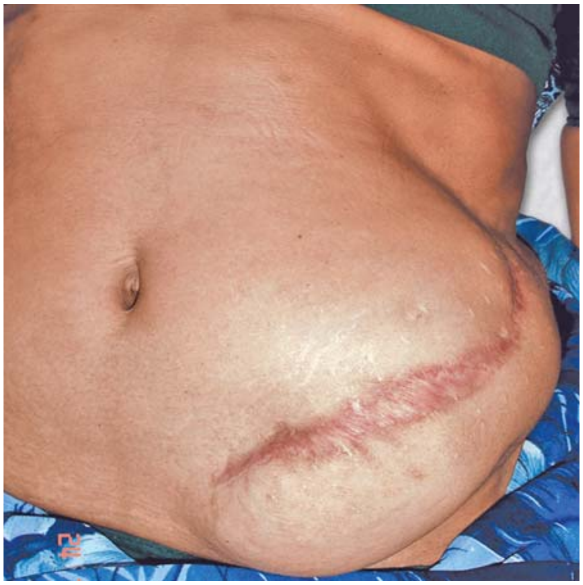

Fig. 18.105: Lumbar incisional hernia—large. Previous lumbar incision scar is visible.

SCIATIC HERNIA

- It is a rare hernia.
- It is the protrusion of the peritoneal sac through the greater or lesser sciatic foramen.

Types

Classified based on their relationship to the pyriformis muscle and ischial spine.

1. Suprapyriformis.
2. Infrapyriformis.
3. Subpyriformis.

Sac lies deep to gluteus maximus. Large hernias protrude below the buttock crease.

Diagnosis

Usually on laparotomy for intestinal obstruction.

Treatment

Defect is covered by fascia mobilised from pyriformis muscle after reducing the sac contents.

COMPLICATIONS OF HERNIA SURGERY

Proper idea of complications of hernia surgery are important as often complications may become more problematic than hernia itself. 20 years back where tissue repair was popular recurrence was the most worried complication. Now since recurrence rates have come down due to standard usage of mesh, mesh related complications have become more common than recurrence. Inguinodynia, mesh extrusion, mesh infection, mesh erosion are the worried problems.

• Complications of open hernia surgery	• Early
• Complications of TEP/TAPP	• Delayed
• Complications of other hernia surgery—ventral/femoral	• Late • Life-threatening

Complications of open hernia surgery	***Complications of TEP/TAPP***
• Infection • Groin pain; ostetitis pubis • Ischaemic orchitis—due thrombosis of pampiniform venous plexus (0.5%) • Injury to vas • Injury to viscera • Recurrence • Hydrocele formation • Seroma, haematoma • *Inguinodynia*—neural • *Dysejaculation*—painful, burning sensation just before/during/after ejaculation due to cremaster dysfunction or vas stricture (0.25%)	• SC emphysema • Pneumothorax, hypercarbia • Vascular • Neural • Visceral • Infection, ileus • Conversion • Recurrence

Immediate	***Late/delayed***
• Vascular—injury to iliac vessels/IE vessels • Visceral injury—bowel/bladder • Vas injury • Anaesthetic complications	• Seroma/haematoma • Neural complications • Intestinal obstruction • Bowel adhesion/ fistula • Testicular atrophy • Mesh related complications • Recurrence

Inguinodynia

- It is chronic inguinal pain seen in post-hernia surgery patients (30%)—whether tissue or mesh repair.
- *Causes are*—traction, cautery, transection, entrapment. Even though it can occur in both tissue and mesh repair, it is more observed in mesh repair especially in onlay mesh repair. It is due to entrapment of nerve in the suture or in the mesh itself or nerve gets adherent to the mesh during fibrosis (perineural fibrosis) (*Mesh inguinodynia*).
- It may be transient or persistent.
- Nerves involved are—iliohypogastric, ilioinguinal, genital branch of genitofemoral nerve, paravasal nerves.
- *Features are*—distressing pain in the groin which often radiates to thigh, scrotum and loin. Arch and twist mobility of pelvis reproduce the pain. Bupivacaine injection relieves the pain. Imaging/nerve conduction studies are of no use.
- Open method has higher (32-38%) incidence of inguinodynia than TAPP/TEP. In open hernia surgery inguinodynia has replaced recurrence as a primary complication. It is distressing discomfort to both patient and surgeon. Incidence is less in posterior or laparoscopic approach.
- It is treated with analgesics/nerve block (Injection of steroid, local anaesthetic agents, phenol, and alcohol)/transcutaneous stimulation/cryotherapy/radiofrequency therapy/ neurectomy.
- Neurectomy is done via groin or suprainguinal or laparoscopic or laparotomy approach. Commonly groin approach is used. Ilioinguinal, iliohypogastric and genitofemoral nerves are carefully dissected. Nerve which is causing the problem is transected at a point where it comes out of the internal oblique muscle and proximal end should be ligated, otherwise end neuroma may form to cause recurrence of the pain. This ligated stump is buried in the internal oblique muscle. Just neurolysis may be sufficient but chances of re-adhesion are higher with recurrence of symptoms. In difficult cases nerve may be transected through laparotomy/ laparoscopy approach.

PARASTOMAL HERNIA

- It is herniation of intestine on the side of bowel stoma—ileostomy/colostomy/other stomas.
- Incidence is 8%.
- *Classification (Devlin's)*: (1) Subcutaneous—intestine herniates along the side of stoma to reach subcutaneous plane—most common. (2) Interstitial—along the side of stoma intestine herniates into intermuscular plane. (3) Intras-

tomal—bowel herniates between emerging and everted parts of the stoma. (4) Perstomal—herniation occurs between the layers of the prolapsed stoma.

- *Features:* Pain, features of obstruction, stomal malfunction, swelling which is tender, toxicity.
- *Diagnosis*: X ray; CT abdomen.
- *Treatment*: Surgery is indicated in recurrent obstruction, narrow neck, strangulation, interference with stoma.
  - Stoma is dissected, hernia content is reduced and defect is repaired. But it shows high recurrence.
  - Subcutaneous mesh repair as onlay placement after reduction of hernia.
  - Extraperitoneal mesh repair around the stoma.
  - Intraperitoneal mesh repair deep to peritoneum surrounding the stoma and defect.

DIFFERENT TYPES OF HERNIA

1. Gibbon's hernia—It is hernia with hydrocoele.
2. Berger's hernia—Hernia in pouch of Douglas.
3. Romberg hernia—Saddle hernia.
4. Obturator hernia—Hernia through obturator foramen (canal).
5. Grynfelt's hernia—Upper lumbar triangle hernia.
6. Petit's hernia—Lower lumbar triangle hernia.
7. Femoral hernia—Hernia medial to femoral vein.
8. Cloquet's hernia—Hernia through pectineal fascia.
9. Narath's hernia—Behind femoral artery, in congenital dislocation of hip.
10. Hesselbach's hernia—Lateral to femoral artery.
11. Serofini's hernia—Behind femoral vessels.
12. Laugier's hernia—Through lacunar ligament.
13. Teale's hernia—In front of femoral vessels.
14. Richter's hernia—Part of circumference of bowel wall is gangrenous.
15. Littre's hernia—Hernia with Meckel's diverticulum as the content.
16. Sliding hernia—Posterior wall of the sac is formed by colon or bladder.
17. Maydl's hernia—'W' hernia.
18. Phantom hernia—Localised muscle bulge following muscular paralysis.
19. Spigelian hernia—Hernia through spigelian fascia.
20. Mery's hernia—Perineal hernia.
21. Sciatic hernia—Hernia through greater or lesser sciatic foramen.
22. Little's hernia—Appendix in hernial sac.
23. Beclard's hernia—Femoral hernia through the saphenous opening.
24. Barth's hernia—Hernia between abdominal wall and persistent vitellointestinal duct.
25. Holthouse's hernia—Inguinal hernia that has turned outwards into the groin.

Chapter 19 Oesophagus

CHAPTER OUTLINE

- Anatomy
- Lower Oesophageal Sphincter
- Dysphagia
- Contrast Study of Oesophagus
- Oesophagoscopy
- Oesophageal Endosonography
- Gastro-oesophageal Reflux Disease
- Hiatus Hernia
- Rolling Hernia
- Reflux Oesophagitis
- Barrett's Oesophagus
- Barrett's Ulcer
- Oesophageal Motility Disorders
- Achalasia Cardia
- Plummer-Vinson Syndrome
- Corrosive Stricture of Oesophagus
- Schatzki's Rings
- Boerhaave's Syndrome
- Mallory-Weiss Syndrome
- Tracheo-oesophageal Fistula
- Oesophageal Diverticulum
- Carcinoma Oesophagus
- Benign Tumours of the Oesophagus
- Oesophageal Perforation
- Foreign Body Oesophagus

ANATOMY

Oesophagus is a hollow muscular tube which begins at the lower edge of the cricoid cartilage (C_6 vertebra) and ends at oesophagogastric junction (T_{12} vertebra). It is 25 cm in length. Upper end is closed by cricopharyngeus muscle which is 18 cm from upper incisors. Lower end is 40 cm from the upper incisors (Upper jaw is fixed and so is used as the landmark to measure, but not the lower jaw which is mobile).

It lies anterior to vertebral column and posterior to the trachea. It lacks serosal layer but is surrounded by a layer of loose fibroareolar adventitia.

Anatomical specialities

- Lacks serosa (other structure without serosa is rectum).
- Contains 2 different types of muscles (striated and smooth at proximal 1/3 and distal 2/3 respectively)
- Contains 2 different types of epithelium.
- Segmental blood supply.
- Only part of GIT which shows *very thinly scattered* Meissner's plexus.
- Longitudinal arrangement of veins and lymphatics.

It is lined by squamous epithelium throughout the length except the last 3 cm (OG junction) which is lined by columnar epithelium. Submucosa of the oesophagus is thick and strongest part.

Parts

1. ***Cervical oesophagus:*** It extends from cricopharyngeus which is the horizontal part of inferior constrictor muscle. Upper oblique part is called as thyropharyngeus. Gap between the two is called as *Killian's dehiscence* which is a *site of occurrence of pharyngeal pouch*. Cervical oesophagus is related to trachea and recurrent laryngeal nerve.
2. ***Thoracic oesophagus:*** Lies initially towards the right side. In lower third it deviates towards the left and continues as abdominal oesophagus. It is related to azygos vein, thoracic duct (which crosses the oesophagus posteriorly from right to left), aorta, pleura and pericardium.
3. ***Abdominal oesophagus*** is 2.5 cm long and grooves behind the left lobe of the liver.

Three areas of anatomic narrowing

- Cervical constriction—occurs at the level of cricopharyngeal sphincter—*narrowest point of GIT*—15 cm from upper incisor—site of F/B impaction
- Bronchoaortic constriction—located at the level of T_4 – 25 cm from upper incisor—site of endoscopic perforation
- Diaphragmatic constriction—occurs where oesophagus traverses the diaphragm (Level of T_{10})—40 cm from upper incisor

Arterial supply of oesophagus

By inferior thyroid artery, oesophageal branches of the aorta, gastric arteries and inferior phrenic arteries.

Venous drainage of oesophagus

By inferior thyroid vein, brachiocephalic vein, left hemiazygos vein, azygos vein, coronary vein, splenic vein and inferior phrenic vein. Veins are longitudinal and they lie in submucosal plane in lower third and in muscular plane above.

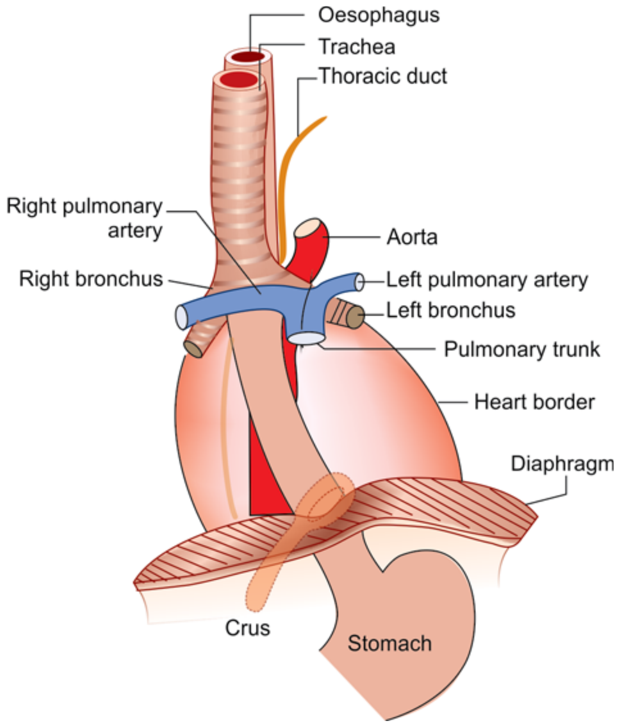
Oesophagus
Trachea
Thoracic duct
Right pulmonary artery
Aorta
Right bronchus
Left pulmonary artery
Left bronchus
Pulmonary trunk
Heart border
Diaphragm
Crus
Stomach

Fig. 19.1: Anatomical relations of the oesophagus.

Lymphatic drainage

Lymphatic arrangement in oesophagus is longitudinal and so spread of carcinoma to distant lymph nodes occurs early. Longitudinal lymphatics are 6 times more than transverse one There are *more lymph vessels* in submucosa than blood vessels. Lymph nodes are:

- *Paraoesophageal groups* located in the wall of the oesophagus and are cervical, thoracic, paraoesophageal and paracardial nodes.
- *Perioesophageal groups* located immediately adjacent to oesophageal wall. They are deep cervical, scalene, paratracheal, mediastinal, diaphragmatic, gastric and coeliac lymph nodes.
- *Lateral oesophageal groups* receive lymph from para and perioesophageal lymph nodes.

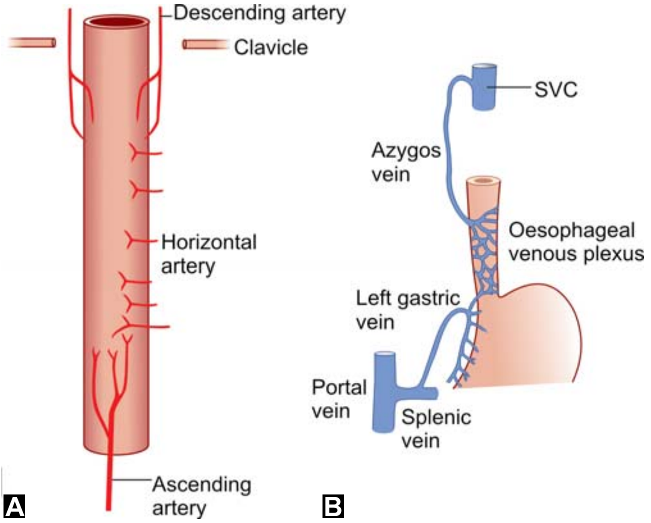
Descending artery
Clavicle
Horizontal artery
Ascending artery
A
SVC
Azygos vein
Oesophageal venous plexus
Left gastric vein
Portal vein
Splenic vein
B

Figs 19.2A and B: (A) Arterial supply and (B) Venous drainage of oesophagus.

- Upper oesophagus drains into supraclavicular nodes.
- Thoracic oesophagus drains into paraoesophageal and tracheobronchial nodes in thorax.
- Abdominal oesophagus drains into coeliac nodes.
- However due to longitudinal arrangement of lymphatic plexus in the wall of oesophagus, any group of (neck/thorax/coeliac) lymph nodes can get involved in diseases at any level.

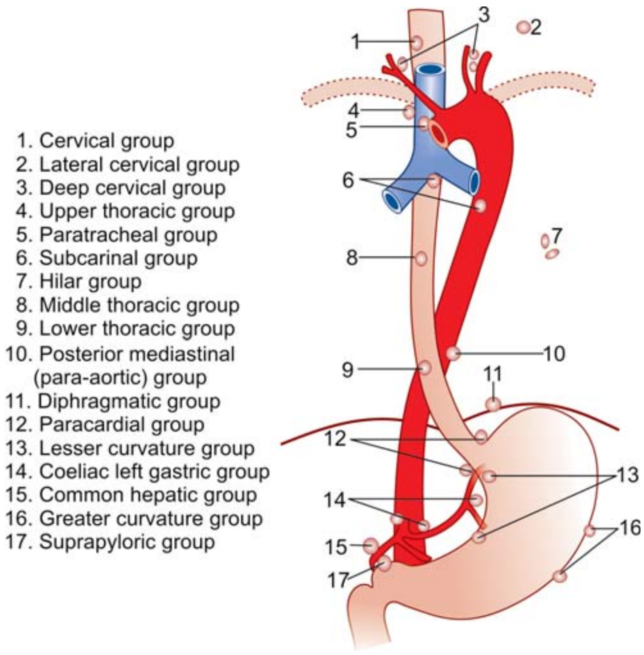
1. Cervical group
2. Lateral cervical group
3. Deep cervical group
4. Upper thoracic group
5. Paratracheal group
6. Subcarinal group
7. Hilar group
8. Middle thoracic group
9. Lower thoracic group
10. Posterior mediastinal (para-aortic) group
11. Diphragmatic group
12. Paracardial group
13. Lesser curvature group
14. Coeliac left gastric group
15. Common hepatic group
16. Greater curvature group
17. Suprapyloric group
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Fig. 19.3: Lymphatic drainage of the oesophagus.

Nerve Supply

- Oesophagus is innervated by vagus.
- It has got both sympathetic and parasympathetic innervation.
- It has got mainly *Auerbach's plexus* between longitudinal and circular muscle layers.

Leugart's pouch is prolapse of the wall of the oesophagus like a diverticulum over a band *(Leugart's band)* between the left bronchus and adjacent vertebra.

Physiology

Three types of contractions in the oesophagus:

1. *Primary:* Progressive, triggered by swallowing.
2. *Secondary:* Progressive, generated by distension or irritation.
3. *Tertiary:* Nonprogressive.

Presbyoesophagus is less efficient oesophageal peristalsis.

LOWER OESOPHAGEAL SPHINCTER (LOS)

- It is a high physiological pressure zone located in the lower end of the oesophagus in terminal 4 cm, with a resting pressure of 10-25 mmHg.
- LOS prevents reflux of gastric and duodenal contents.
- It is influenced by food, gastric distension, gastric pathology, smoking, GI hormones, alcohol.

- Normally there is a *transient relaxation period* wherein reflux (*physiological*) occurs but then immediately gets cleared by oesophageal clearance mechanism.
- So *pathological reflux or GORD* can occur due to decreased LOS tone, altered relaxation time, reduced oesophageal clearance mechanism, or other altered mechanical factors.
- *Oesophageal clearance mechanism* is due to primary oesophageal peristalsis which carries saliva with high bicarbonate content which neutralizes and clears the transient physiological reflux. Manometry with special microtransducers are used to measure the LOS pressure.

DYSPHAGIA

- *Dysphagia* is difficulty in swallowing. Painful swallowing is *odynophagia*.
- It can be *acute*—due to foreign body impaction or acute infection or *chronic* due to causes like stricture or carcinoma, etc.
- Associated hoarseness of voice may be present in advanced pharyngeal or post cricoid carcinomas. At late stage laryngeal carcinoma also can cause dysphagia along with hoarseness of voice.
- Dysphagia can be *oropharyngeal* or *oesophageal* depending on the cause.
- Dysphagia may be due to pathology in *voluntary/pharyngeal phase* of the swallowing wherein patient also develops cough while swallowing. Dysphagia due to problem in *oesophageal involuntary phase* of swallowing is specified by food getting stuck in the pathway. But site of "*food getting stuck*" feeling is not relevant.
- Dysphagia can be *progressive or intermittent*.

Causes of Dysphagia

Common causes:

- Gastro oesophageal reflux diseases (GORD/GERD/Hiatus hernia).
- *Carcinoma oesophagus*: Here dysphagia is of short duration and progressive. 2/3 of the lumen should be blocked by tumour to develop dysphagia.
- *Foreign body in oesophagus*: It may be coin/bone piece/denture. It is common in children. It causes acute dysphagia. It may be often life threatening.
- Carcinoma of pharynx or posterior 1/3rd of the tongue.
- *Corrosive strictures*: It is usually alkali stricture. Squamous mucosa is resistant to acid effect to certain extent.
- *Oesophageal candidial infection*: It is becoming common due to immunosuppression in association with HIV infection; steroid therapy; cancer chemotherapy; post-transplant period, etc. Presentation is dysphagia and odynophagia. Oral candidiasis (thrush) is obvious. Endoscopy shows whitish curd-like plaques in the oesophageal mucosa which cannot be moved (whereas food particles can be moved). Barium swallow shows mucosal ulceration and irregular areas. Biopsy confirms the diagnosis. Treatment is oral antifungal as well as topical antifungal therapy.

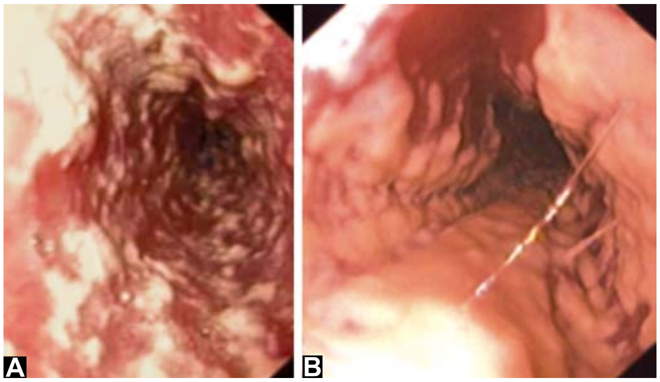

Figs 19.4A and B: Endoscopic view of oesophageal candidiasis infection. Note the whitish curd-like plaques.

- Plummer-Vinson syndrome.
- Mediastinal swellings—primary tumours/nodal mass either lymphoma or secondaries or tuberculosis.

Rare causes:

- *Diffuse oesophageal spasm*: They are incoordinated contractions of oesophagus causing chest pain or dysphagia. It is common in distal 2/3rd of the oesophagus. Hypertrophy of circular muscle fibres with very high persistent pressure of 400-500 mmHg is specific. Treatment is calcium channel

Causes of dysphagia

Extraluminal causes	*Causes in the wall of oesophagus or other area*	*Causes in the lumen*
• Mediastinal nodes—secondaries/lymphoma/tuberculosis • Aortic aneurysm • Rolling hiatus hernia • Thyroid enlargement—malignant • Dysphagia lusoria • Congenital anomalies • Mediastinitis/mass	• Carcinoma oesophagus • Corrosive/tuberculous/inflammatory/congenital stricture oesophagus • GERD • Achalasia cardia • Plummer-Vinson syndrome • Oesophageal diverticulum • Carcinoma posterior 1/3rd of tongue/pharynx • Diffuse oesophageal spasm • Congenital anomalies • Retropharyngeal abscess/peritonsillar abscess (Quinsy)/acute tonsillitis/pharyngitis	• Foreign body in the oesophagus—coin/dentures/fish or meat bone *Other causes* Cranial causes (neurological): • Bulbar palsy/infarction/hemiplegia • Vertebro basilar insufficiency

blockers, vasodilators, endoscopic dilatation and extended oesophageal surgical myotomy up to the aortic arch (very useful especially for dysphagia; not much for chest pain).

- Oesophageal diverticula, Chaga's disease.
- *Dysphagia lusoria (Lusoria means in Latin "sport of nature"):* It is a congenital vascular anomaly of aortic root. *Aortic arch anomalies* are—double arch (40%), right arch and left ligamentum arteriosum (25%), anomalous innominate or common carotid artery or aberrant right subclavian artery (10%). It is due to disappearance of proximal right 4th aortic arch instead of distal portion. All patients having this anomaly (*dysphagia lusoria*) have got an *aberrant right subclavian artery* in a transposed position arising from descending aorta that courses posterior to oesophagus. Often there will be a complete *vascular ring* around trachea and oesophagus. It is categorised based on their specific subclavian anomaly—depends on presence of aneurysm, occlusive disease and compression. Commonly they are asymptomatic. Presentations may be dysphagia, chest pain, stridor, wheeze, recurrent respiratory infection (usually presents after the age of 40).
  Investigations: CT chest, MRI, chest X-ray, barium swallow (T4 level diagonal impression) and endoscopy (Shows pulsating extraluminal compressive mass).
  Treatment is reconstruction or ligation of aberrant right subclavian artery by sternotomy/by neck approach.
- *Thyroid swelling*: It is uncommon to develop dysphagia in a thyroid swelling. There will be always dyspnoea when dysphasia develops. Large malignant thyroid or anaplastic thyroid can cause dysphagia with dyspnoea or stridor.
- *Boerhaave's syndrome*: It is vertical full thickness tear of lower oesophagus due to vomiting with closed glottis. It is often life threatening and emergency.
- Neurological causes like stroke, bulbar palsy, motor neuron disease, Parkinson's disease, etc.
- Congenital anomalies of oesophagus.
- *Drug induced dysphagia*: Drugs like KCl, quinine, NSAID can cause dysphagia.
- Mediastinal fibrosis.

Evaluation of a Patient with Dysphagia

- Proper history.
- Haematocrit.
- Chest X-ray often shows mediastinal mass lesion/ foreign body.
- *Oesophagoscopy*: Once lesion is detected, it is treated accordingly. Biopsy from lesions, endotherapy if needed should be carried out (like F/B removal; stricture dilatation; sclerotherapy).
- Barium swallow may show irregular filling defect or extrinsic compression.
- CT scan chest is very useful method to identify the anatomical location of the cause (nodes/tumour/aorta/cardiac cause/ congenital). Extent, spread, nodal status, size and operability of tumour also well-assessed.
- Oesophageal manometry in achalasia cardia/GERD.
- 24 hours pH monitoring is ideal and most accurate for GERD. Small pH probe (transnasal catheter) is passed into the distal oesophagus 5 cm proximal to upper margin of LOS under manometry guidance. Probe is connected to a digital recorder worn by the patient for 24 hours. Record is analysed using a computer. A pH less than 4 for more than 4% of total 24 hours period (more than near to one hour in toto in 24 hours) is pathological reflux. It is often assessed by scoring system. Radio-telemetry pH probes are used now without any nasal tube. It is passed and placed on the oesophageal wall using endoscope.
- Endosonography is very useful in many conditions causing dysphagia. It can assess site, layers of the oesophagus, nodes, spread, etc. properly. Different layers are seen as alternating hyperechoic and hypoechoic bands.
- Ultrasound abdomen to see abdominal nodes/liver/ascites.
- MRI study.

Treatment for Dysphagia

Depends on cause—modified Heller's myotomy; oesophageal resection; dilatation; F/B removal, etc.

CONTRAST STUDY OF OESOPHAGUS

Types

1. Barium swallow using barium sulphate (thick paste).
2. Using water soluble contrast like 'Gastrografin'.

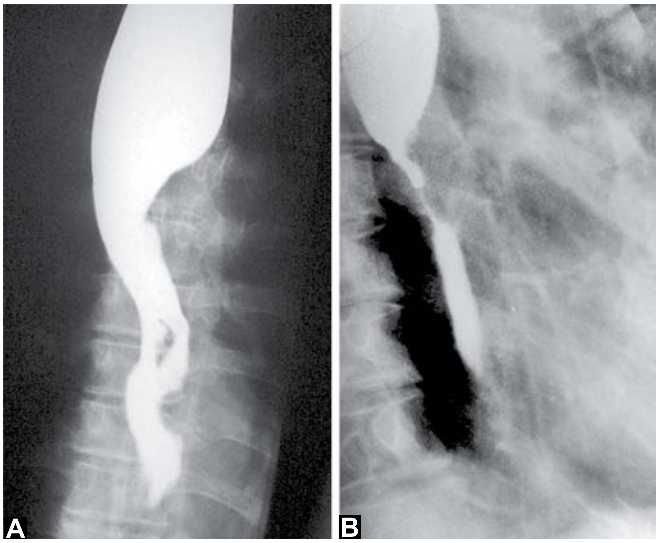

Figs 19.5A and B: Barium swallow X-rays showing narrowing and irregularity in carcinoma oesophagus.

Indications

1. *Barium swallow*
   - Dysphagia due to motility disorder like achalasia cardia, diffuse oesophageal spasm.
   - Dysphagia due to mechanical causes like carcinoma, benign strictures and neoplasms, external compression.
   - Pharyngeal pouch and other diverticula.
   - GORD.

Instrumentation is the most common cause of oesophageal rupture.

2. *Water soluble contrast radiograph*
    - In suspected oesophageal perforation.
    - Leaking oesophageal anastomosis.

Important Findings in Barium Swallow

- Achalasia cardia—*'Bird beak'* appearance, as the oesophagus is grossly dilated above an apparent narrowing at the cardia. In long standing cases—'sigmoid oesophagus'.
- Diffuse oesophageal spasm—*Corkscrew* appearance.
- GORD—shows reflux when done in Trendelenburg's position.
- Oesophageal carcinoma—irregular stenosing lesion with shouldering (*'Rat tail'* is a fluoroscopic finding).
- Pharyngeal pouch—demonstration of the pouch.
- External compression—indentation of barium column by superior or posterior mediastinal mass, enlarged left atria as in mitral stenosis.

OESOPHAGOSCOPY

Indications

Diagnostic

- To identify the lesion and to take biopsy in carcinoma oesophagus.
- For diagnosing other oesophageal conditions.

Therapeutic

- To remove foreign body
- To dilate stricture
- To place endostents for inoperable carcinoma oesophagus
- To inject sclerosants for varices

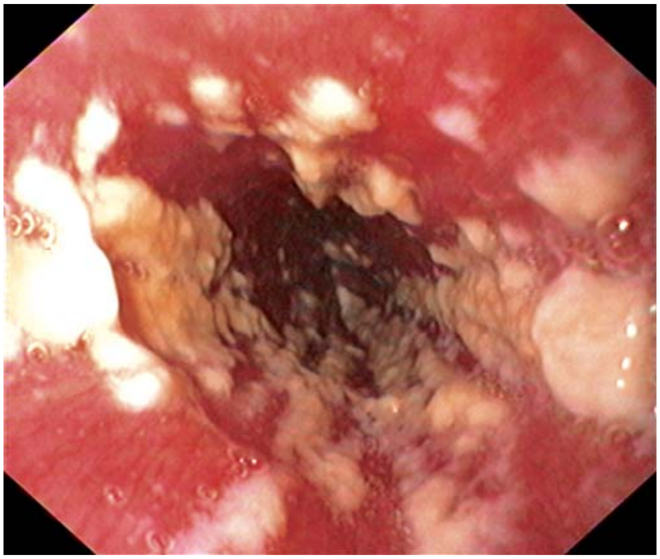

Fig. 19.6: Oesophageal candidiasis—note the multiple curd-like lesions.

Types

1. ***Rigid oesophagoscope (Negus type)***
    - It is done under general anaesthesia. Head is extended and head end of the table is tilted upwards, scope is passed behind the epiglottis and cricoid through the cricopharyngeal opening. This is the most difficult part in oesophagoscopy. After that, negotiating through the oesophagus is easier. The lesion is identified and biopsy is taken if required.
    - *Complications* are perforation (most common site of perforation is at the level of cricopharyngeus), bleeding.
2. ***Fibreoptic flexible oesophagoscope***
    - It can be done under local anaesthesia. Reflux and hiatus are well identified. Stomach also can be visualised. Easy to pass and perforation is unlikely.
    - But tissue taken for biopsy is smaller and removal of foreign body is also difficult.

OESOPHAGEAL ENDOSONOGRAPHY

- It is useful method of finding and assessing involvement or pathology of different layers of oesophagus especially in carcinoma oesophagus. It shows all layers clearly and distinctly and so invasion can be better made out and operability can be decided.
- **Endoscopic oesophageal staining** using labelled iodine is used to identify early carcinoma in oesophagus. Normal mucosal cells contain glycogen which takes up iodine and so stains brown, whereas carcinoma cells will not take up iodine and so mucosa appears pale (not stained).

GASTRO-OESOPHAGEAL REFLUX DISEASE (GORD/GERD)

- It is a pathological reflux from the stomach into the lower oesophagus.
- It accounts for 75% of oesophageal pathology (most common).
- Symptoms may be due to peptic oesophagitis, LOS dysfunction, aspiration due to reflux, due to late effects like stricture and obstruction.
- It is due to varied *anatomical and physiological factors.*

Anatomical Factors

- Obesity.
- Altered length of intra-abdominal oesophagus. Level of acid exposure is inversely related to length of the intraabdominal oesophagus. More the length more the need of intraabdominal pressure to cause the reflux and so less the reflux with more length.
- Alteration of phreno-oesophageal ligament.
- Altered obliquity of O-G junction (alteration in angle of *'His'*).
- Reduced pinching action (*Pinch-Cock effect*) of right crus of diaphragm.
- Alteration in normal mucosal rosette at O-G junction.
- Alteration in *sling mechanism* of gastric musculature.

Physiological Factors

- Reduced LOS pressure. Normal resting LOS pressure is 10-25 mmHg. During swallowing pressure drops to 1-3 mmHg for few seconds (10 seconds). Resting pressure is increased by gastrin, cholinergic, adrenergic, prokinetic

drugs; decreased by secretin, cholecystokinin, glucagon, calcium channel blockers, deriphylline, coffee, fatty meal.
- Altered transient relaxation period in LOS.
- *Reduced oesophageal clearance* mechanism. Change in oesophageal clearance is due to *altered oesophageal body pump* (primary peristalsis with 30-80 mmHg amplitude propagative downwards waves; secondary by food bolus), *altered salivary bicarbonate* neutralizing mechanism of reflux acid, *altered mucosal defense* mechanism (mucous and surface active phospholipid).
- *Delayed gastric emptying* due to diabetes, neuromuscular block, gastroparesis, medications.
- *Increased gastric distension* and gastric acid hypersecretion.

Other Factors

- Alcohol, smoking, stress, lifestyle.
- It is seen in infants and children and also in pregnant women.

Pathogenesis

GERD begins from *distended fundus* due to variety of causes and other anatomical and physiological factors, repetitively unfolding the sphincter to the gastric juice leading to all problems. There is reduced and low LES resistance causing loss of barrier to reflux. It is probably due to gastric distension, delayed gastric emptying, and raised intragastric or intraabdominal pressure. Gastric distension may be due to overeating or aerophagia. Fundic distension → stretching of fundus → LOS squamous epithelium exposed to acid gastric juice → oesophagitis → increased stimulus to swallow saliva to neutralise oesophagitis → further fundal distension → cycle repeats → sphincter is taken into stretched fundus → effects like erosion, ulceration, fibrosis, mucosal metaplasia.

Types

Classification I
- *Symptomatic uncomplicated GORD* (NERD—*Nonerosive Reflux Disease*).
- *Symptomatic, complicated GORD*.

Classification II
- *Primary*: Incompetent LOS.
- *Secondary*: Due to surgery/disease.

Classification III
- *GORD with sliding hernia.*
- *GORD without sliding hernia.*

Clinical Features

- Fatty dyspepsia.
- Chest pain and heart-burn **(pyrosis)**—80%—main symptom. Pain is more in lying down position and at night. It may mimic cardiac pain/angina. It often begins in the epigastrium and becomes substernal.
- Odynophagia (painful swallowing).
- Appearance of symptoms within seconds of ingestion of food is typical.
- Regurgitation—50%. It is return of oesophageal content.
- Laryngeal symptoms.
- Dysphagia will occur once complications begin.
- Symptoms are more with change of position.
- Chronic cough, shortness of breath and hoarseness—Nocturnal reflux—reflux is return of gastric content.
- Haematemesis.
- Postprandial fullness, choking, wheezing, recurrent pneumonia are other features.

Triad

- Heart burn
- Epigastric pain with dysphagia
- Regurgitation

Typical Symptoms

Heart burn, regurgitation, dysphagia mainly; odynophagia, haematemesis often.

Atypical Symptoms

Cough, hoarseness, wheezing, sore throat, noncardiac chest pain and palatal/dental erosions. It will eventually lead into recurrent pneumonia and pulmonary fibrosis. If pH of the cervical oesophagus is below 4, for less than 1% of time, then respiratory symptoms are due to GERD.

De Meester's scoring system

Grade	*Features*
	Heart burn
0	None
1	Minimal—occasional episodes
2	Moderate—medical therapy visits
3	Severe—interferes daily activities
	Regurgitation
0	None
1	Minimal—occasional episodes
2	Moderate—on position and straining
3	Severe—features of aspiration
	Dysphagia
0	None
1	Minimal—occasional episodes
2	Needs fluid to clear
3	Causes food impaction

Complications

- Reflux oesophagitis.
- Sliding hiatus hernia.

Reflux is precipitated by flexion of the trunk—Boot lacing sign.

- Stricture lower end oesophagus.
- Oesophageal shortening.
- Barrett's oesophagus.
- Carcinoma (adeno) oesophagus (10% of GERD).

Oesophageal	*Extraoesophageal*
• Oesophagitis • Stricture • Barrett's oesophagus • Oesophageal shortening • Carcinoma (adeno)	• Laryngitis • Recurrent pneumonia • Progressive pulmonary fibrosis

Investigations

- Barium study in head down position.
- Endoscopy to exclude other disease and to assess any mucosal injury—red inflamed mucosa often with ulceration. Sliding hernia can be identified through endoscopy. When patient retches, gastric mucosa will enter the OG junction and ascends upwards to variable distance.
- Mucosal biopsy to confirm metaplastic transformation.
- Oesophageal manometry to assess the function of LES. Length and pressure of LOS is important. Transient relaxation period which is increased is most important.
- 24 hours oesophageal pH monitoring—*gold standard*. PPI should be stopped for 3 weeks prior to pH monitoring.

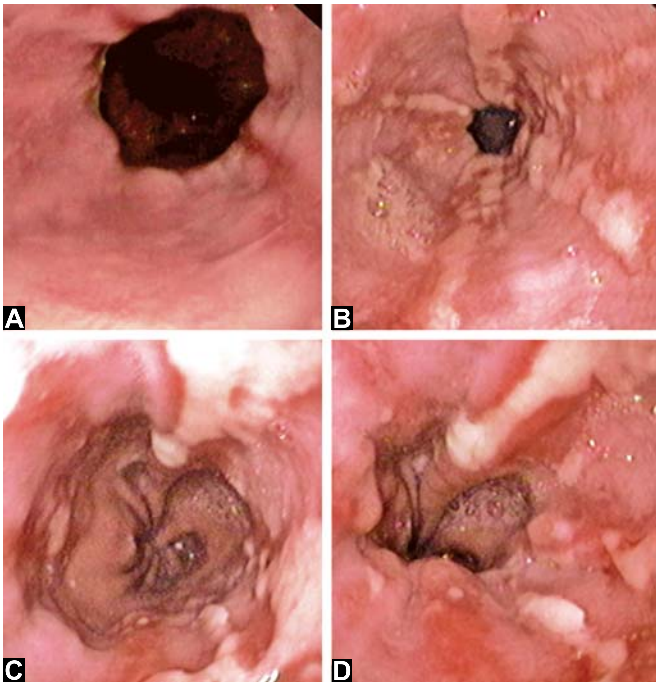

Figs 19.7A to D: Gastro-oesophageal reflux disease (GORD) endoscopic views.

- ***Bernstein test:*** Instillation of 1:10 HCl into stomach of patient with reflux will reproduce the symptoms; and pH in the lower end of the oesophagus decreases. Later pain disappears with saline infusion.
- Dual probe pH monitoring (one in distal oesophagus and one in proximal oesophagus/trachea) is used to confirm the respiratory complications of GORD.
- *DeMeester scoring system* is used to assess the severity of the GORD.
- 99Tc sulphur colloid mixed with saline scan to study OG reflux and 99Tc HIDA scan to study and detect duodeno-gastric reflux are used.

Differential Diagnosis

- Achalasia cardia.
- Carcinoma oesophagus.
- Peptic ulcer.
- Gallstones.
- Pancreatic diseases.
- Gastritis.
- Cardiac angina.

Treatment of Uncomplicated GORD

A. ***General measure—Life-style changes***
  - Control of obesity.
  - Stop smoking and alcohol.
  - Avoid tea, coffee and chocolate.
  - Propped up position.
  - Small frequent meals.

B. ***Drugs***
  - **H_2 antagonists; antacids.**
  - **Proton pump inhibitors (PPIs)—very effective**
    - Omeprazole 20 mg—BD for 3-6 months.
    - Lansoprazole 30 mg.
    - Pantoprazole 40 mg.
    - Esomeprazole 20 mg.
    - Rabeprazole 20 mg.
  - **Prokinetic drugs**
    - Metoclopramide, Domperidone 30 mg.
    - Cisapride, Mosapride.
    - Itopride 50 mg—does not cause cardiac arrhythmias.
    - **Defoaming** semethicone, alginic acid along with antacids.
    - **Anticholinergic drugs**—pirenzipine.
    - **Mucosal protector drugs**—sucralfate, colloid bismuth.

C. ***Endoluminal therapies for GORD***
  - Endoluminal *suturing* (Wilson-Cook)
  - *Plexiglass microspheres* (PMMA). Microsphere suspended in gelatin is injected through endoscopic needle. Gelatin gets absorbed and spheres cause a tissue bulking
  - *Gatekeeper reflux* repair system. Endoscopic delivery of preformed radiopaque hydrogel into submucosa
  - *Stretta catheter*. Flexible, soft, 6 mm sized, 65 cm length tube with balloon/basket and 5.5 mm NITI electrode, irrigation and suction
  - *Enteryx* injection technique (estrinyl vinyl alcohol)
  - *Endocinch* technique
  - Endoscopic full thickness *plication.*

D. ***Surgery***

- *Indications for surgical treatment*
  - Failure of drug treatment
  - Sliding hernia
  - Barrett's ulcer
  - Severe pain
  - Presence of complications like bleeding/stricture/ shortening, respiratory problems

Surgeries

Principles of antireflux surgeries

- Operation should restore the LES pressure twice the resting gastric pressure
- The adequate length of the intraabdominal oesophagus should be maintained
- Apposition of diaphragmatic crurae, reduction of hiatus hernia when present
- Repaired OG junction should relax during swallowing adequately
- Tension free fundoplication should be done

Antireflux surgery is the only effective long-term beneficial therapy ideally available and considered now. *Laparoscopic fundoplication* (most popular), Collis-Nissen vertical gastroplasty and fundoplication, resection of OG junction when metaplasia is present, Belsey Mark 4 operation (technically difficult) are the procedures used now.

Fundoplication

- *Nissen's posterior total fundoplication:* Here after narrowing the crus of diaphragm, mobilised *posterior part of the fundus* of the stomach is wrapped totally 360° around the area of OG junction.
  - *Rudolph Nissen* (1959) first did total fundoplication.
  - Opening the peritoneum over the oesophagus to mobilise the lower end of oesophagus.
  - Dissection of the diaphragmatic crura of the oesophagus.
  - Mobilisation of entire fundus by ligating short gastric vessels.
  - Vagi are preserved.
  - In Nissen's, 60 French bougie wrap is ideal; after complete mobilisation, only *posterior part of the fundus* is wrapped around after crural repair (otherwise fundoplication may displace into thorax causing paraoesophageal hernia); intraabdominal length of 2 cm oesophagus is created (not more); fundus of the stomach is known to relax in concert with oesophageal sphincter so only fundus should be used to buttress the sphincter; wrapping should be done only around sphincter oesophagus not around body of stomach; fundoplication should be kept in abdomen without under tension and without displacing into thorax. Nonabsorbable polypropylene sutures are used.
  - Crural repair is done using interrupted polypropylene sutures.
  - Fundoplication should prevent sphincter shortening/ unfolding during gastric distension, should preserve normal swallowing ability, allow proper belching and allow vomiting whenever needed.
  - Bougie; on table and postoperative manometry to assess the pressure of new LOS; on table and postoperative gastroscopy—are commonly used in fundoplication.
  - Mortality of procedure is very less; other parts of the abdomen can be assessed and addressed like gallstones.
  - *Complications* are—*gas bloat syndrome* (inability to belch); dysphagia; inability to vomit; slippage, proximal migration; paraoesophageal hernia, splenic injury. Floppy Nissen's fundoplication reduces the incidence of gas bloat syndrome.
  - *Laparoscopic Nissen's fundoplication* is ideal and equally successful.
  - *Transthoracic approach* is used for Nissen's fundoplication in patients who had hiatus hernia repair earlier; who has short oesophagus; in patients who is having sliding hiatal hernia that does not reduce below the diaphragm, associated pulmonary disease, in obese, in patients who are having narrow subcostal angle or barrel shaped chest.
- *Toupet's partial 180° posterior fundoplication*: It is similar to Nissen's posterior fundoplication; it controls reflux alike Nissen's but gas bloat is very less. It is commonly done procedure now. Short gastric vessels are divided completely or partially.
- *Rosetti Hell anterior fundoplication*: Here *anterior part of fundus* of stomach is used. Superomedial part of the fundus is brought around oesophagus to suture into anterior part of fundus. Here *short gastric vessels are not ligated* and much less fundus need to be mobilised.
- *Dor anterior fundoplication*: Here right margin of the fundus is sutured to left margin of the oesophagus; front part of fundus is sutured to right margin of the oesophagus; 2nd row is also sutured to right crus.
- *Watson's anterolateral fundoplication*: 5 cm intraabdominal oesophagus is created with blunt transhiatal dissection. 120° anterolateral fundoplication augments the LOS function during stomach distension. Bougie insertion, division of short gastric vessels are not needed here.
- *Lind both anterior and posterior fundoplication* of 300° with 60° anteriorly uncovered area is not routinely practiced now.
- *Partial fundoplication with mesh wrap* around is also used. But complications of mesh like erosion, stricture and adhesions are problems.

Fundoplications

- *Nissen's*—total 360° posterior fundoplication
- *Toupet's*—partial 180° posterior fundal/posterolateral
- *Rosetti Hell*—total anterior fundal
- *Dor*—anterior partial
- *Watson's*—anterolateral 120° partial
- *Lind*—posterior and anterior

Other Procedures

- *Hill's operation*: Intraabdominal fixation of OG junction (cardia) into median arcuate ligament to augment the effect of LOS, to enable the effective oesophageal peristalsis. Procedure is done by transabdominal approach which is assessed by intraoperative manometry. But it is technically difficult, may damage celiac plexus.
- *Belsey Mark IV operation* (1967): It is plication of oesophagus to the diaphragm through many interrupted mattress sutures so as to push the oesophagus downwards to make it intraabdominal with adequate length. It is done through left transthoracic approach. 240° fundoplication, creation of intraabdominal oesophagus, crural sling repair are the techniques. It corrects GERD and also maintains normal eructation and physiological reflux. But long hospital stay and respiratory complications are distressing problems.
- *Placement of Angelchik prosthesis*: Annular silicone gel filled angelchik prosthesis with a tape on either end. After insertion around OG junction ends are tied around.
- *Collis' vertical gastroplasty* is done using fundus of stomach. A vertical cut along the line of oesophagus is made on the fundus which is sutured to create extra length of oesophagus and new lesser curvature. It is done whenever oesophagus cannot be mobilised due to oesophageal shortening by a stricture developed as a complication of GORD. Partial fundoplication is added to this usually.
- *Thal's patch procedure*: It is done for localised nondilatable stricture oesophagus. Stricture is incised full thickness longitudinally. Mobilised fundus of stomach is placed and sutured as a serosal patch technique over the defect.
- *Narbona's ligamentum teres cardiopexy* (commonly done in Spain): Ligamentum teres is mobilised from falciform ligament with its blood supply from left lobe of liver; it is detached from umbilicus and brought around the mobilised lower oesophagus with traction until LOS pressure becomes 15-20 mmHg with manometry; flap end is sutured left of OG junction and anterior wall of stomach. It is more physiological and division of short gastric vessels is not required.
- *Boerema operation*: It is fixing OG junction in front into anterior abdominal wall. It is not practiced now.
- *Oesophagogastrectomy* is required whenever there is *failure of antireflux* procedure/*metaplasia* is present / *nondilatable stricture* is present. Lower end of oesophagus and proximal stomach is resected with oesophagogastric anastomosis is done through thoracoabdominal approach.
- *Transhiatal oesophagectomy* with anastomosis between cervical oesophagus and stomach in the neck is an alternate procedure which can be used in severe extensive disease and stricture.

Note:
Stricture oesophagus due to GERD which is dilatable can be treated by *regular oesophageal dilatation* and long term PPI therapy.

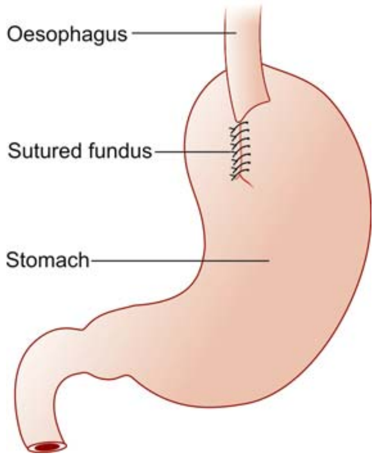

Fig. 19.8: Nissen's fundoplication.

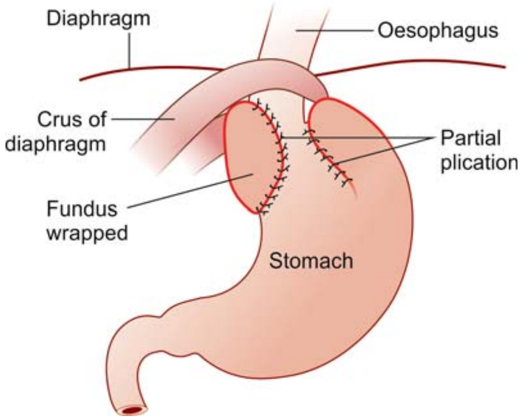

Fig. 19.9: Toupet's partial fundoplication.

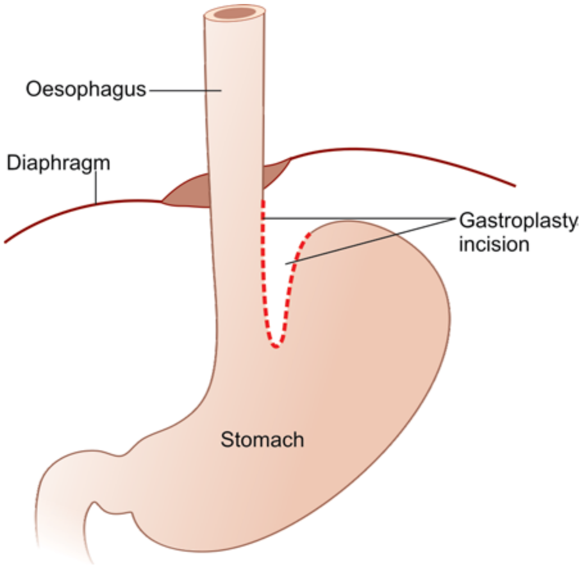

Fig. 19.10: Vertical gastroplasty is done to lengthen the intra-abdominal oesophagus.

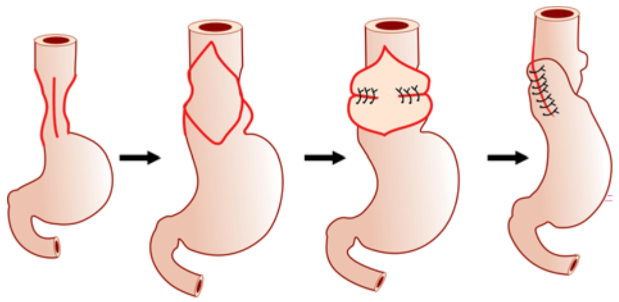

Fig. 19.11: *Thal's patch* done for oesophageal stricture due to reflux oesophagitis.

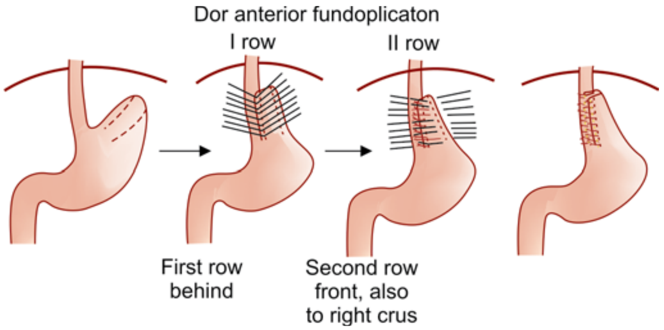
Dor anterior fundoplicaton
I row
II row
First row behind
Second row front, also to right crus

Fig. 19.12: *Dor anterior fundoplication*. Right margin of the fundus is sutured to left margin of the oesophagus. Front aspect of the fundus is sutured to right margin of the oesophagus. Second row is also sutured to right crus.

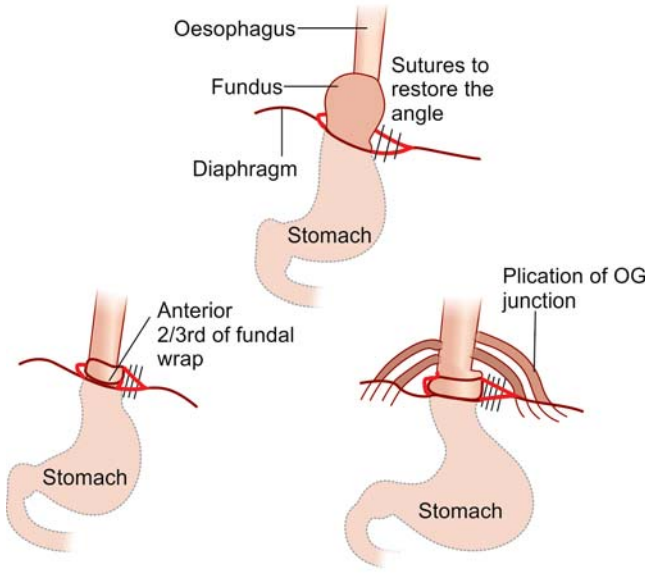
Oesophagus
Fundus
Sutures to restore the angle
Diaphragm
Stomach
Anterior 2/3rd of fundal wrap
Stomach
Plication of OG junction
Stomach

Fig. 19.13: *Belsey mark 4 operation* for sliding hiatus hernia. It is done through thoracic approach. Here angle of oesophagus is restored with sutures and partial anterior fundal wrap is done. Then oesophagus is pushed down by oesophago-diaphragmatic plication sutures.

Complications of Surgery

- Oesophageal/gastric perforation.
- Haemorrhage.
- Pneumothorax/pyothorax.
- Vagus nerve injury.
- Cardiac arrhythmias.
- Sepsis – mediastinitis or septicaemia.
- Disruption/failure of fundoplication.
- Gas bloat syndrome.

HIATUS HERNIA

It is the most common type of diaphragmatic hernia.

Types of hiatus hernia—classification

- *Type I* hiatus hernia: It is the cephalad displacement of the gastrooesophageal junction through the hiatus into the mediastinum. It is usually small, asympto-matic and reducible. It is commonest type
- *Type II* hiatus hernia: It is superior migration of the fundus of the stomach along side the GE junction and oesophagus into the mediastinum with GE junction in normal intraabdominal location. It is rolling hernia
- *Type III* hiatus hernia: It is combination of type I and type II
- *Type IV* hiatus hernia: It is the hernia containing other abdominal viscera as content like transverse colon and omentum

Types

- Sliding hernia (85%).
- Rolling hernia (10-12%).
- Combined.

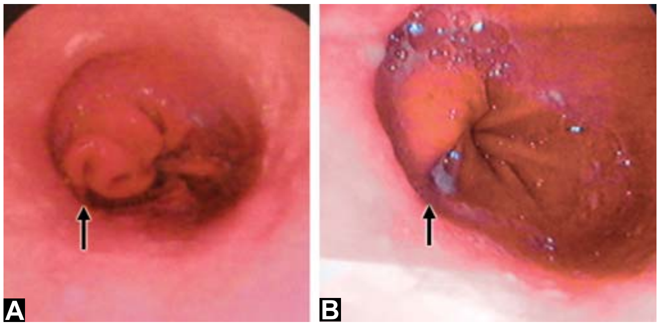

Figs 19.14A and B: Endoscopic view of oesophageal hiatus hernia.

Sliding hernia is commonly associated with GORD (Should be discussed like GORD).

Here the cardia migrates back and forth between the posterior mediastinum and peritoneal cavity.

Saint's triad

- Hiatus hernia
- Diverticulosis
- Gallstones

Kindness is a language which the dumb can speak, the deaf can understand

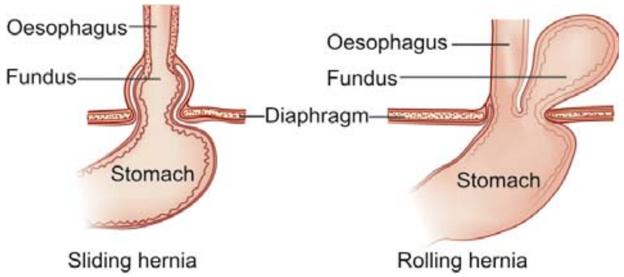

Fig. 19.15: Types of hiatus hernia.

ROLLING HERNIA (PARAOESOPHAGEAL HERNIA)

It is herniation of stomach fundus or rarely other abdominal contents colon/spleen through a hiatus, usually towards left side.

Clinical Features

- Common in elderly.
- Abdominal pain and chest pain.
- Hiccough, early satiety.
- Regurgitation, postprandial bloating.
- Cardiac abnormality (arrhythmia).
- Dysphagia dyspnoea.
- 40% presents as acute features with perforation/gangrene/bleeding.

Complications

- Gangrene of stomach.
- Perforation into the mediastinum.
- Perforation into the peritoneum.
- Gastric volvulus.
- Ischaemic longitudinal ulcer in the herniated stomach in rolling hernia is called as *Cameron ulcer*.

Investigations

- Plain X-ray—lateral and PA erect view showing retrocardiac air-fluid level.
- Barium meal study very useful.
- ECG.
- 3D CT scan is useful.

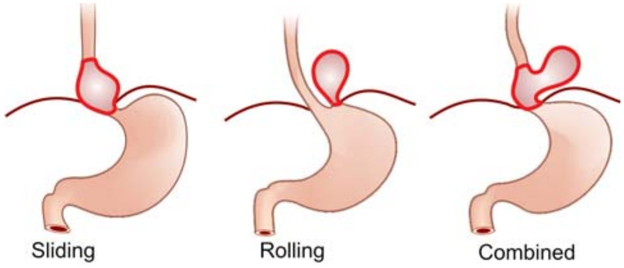

Fig. 19.16: Different types of hiatus hernia.

Treatment

- Treatment is always *surgical*.
  - Excision of sac and repair of the defect.
  - If it is gangrenous, gastrectomy is required.
  - Either abdominal or thoracic or laparoscopic approach can be used in treating rolling hernia surgically.
  - Mesh reinforcement to hiatus to close the defect may be needed.

REFLUX OESOPHAGITIS

Types

1. ***Acute:*** Following *burns, trauma, infection, peptic ulcer.*
2. ***Chronic*****:** Reflux of acid in sliding hernia, after gastric surgery.
   Reflux is quite common in pregnancy.
   Site is always in lower oesophagus.

Pathology

There is bleeding granulation tissue in lower oesophageal mucosa with spasm of longitudinal muscle which pulls the adjacent gastric area upwards into the oesophagus causing sliding hernia.

Grading

1. Mucosal erythema
2. Mucosal erythema + superficial ulceration
3. Mucosal erythema + superficial ulceration + submucosal fibrosis
4. Mucosal erythema + extensive ulceration + paramural fibrosis

Savary-Miller classification of reflux oesophagitis—grading

1. Single or isolated erosive lesion(s), oval or linear but affecting only one longitudinal folds
2. Multiple erosive lesions, noncircumferential affecting more than one longitudinal fold with or without confluence
3. Circumferential erosive lesions
4. Chronic lesion: Ulcers, strictures and or short oesophagus alone or in association with grade 1-3 lesions
5. Columnar epithelium in continuity with the Z line, non-circular, star shaped or circumferential, alone or in association with grade 1-4 lesions

Los-Angeles classification of reflux oesophagitis—grading

A. One or more mucosal break *no longer than 5 mm* that does not extend between the tops of two mucosal folds
B. One or more mucosal break *more than 5 mm* long that does not extend between the tops of two mucosal folds
C. One or more mucosal break that continues between the tops of two or more mucosal folds but which involves *less than 75%* of the circumference
D. One or more mucosal break, which involves *at least 75%* of the oesophageal circumference

Clinical Features

- It is a part of GORD.
- Pain and burning sensation in retrosternal area often referred to shoulder, neck, arm.
- Heart burn is common.

- Dysphagia.
- Anaemia.

Diagnosis

- Barium meal X-ray.
- Gastroscopy and biopsy.

Barrett's ulcer is an ulcer with gastric (columnar) metaplasia in lower oesophagus.

Treatment

- *Antacids*
- *H_2 blockers*: Ranitidine, famotidine.
- ***Proton pump inhibitors****—Main method and more effective.*
  - Omeprazole 20 mg BD one hour before food (Morning) for 6 months
  - Lansoprazole 30 mg
  - Pantoprazole 40 mg
  - Esomeprazole 20 mg
  - Rabeprazole 20 mg (can be given with food).
- ***Prokinetic drugs*** like metochlopramide, domperidone, cisapride, mosapride.
- *Treating GORD and associated causes*. By fundoplication and other surgeries.
- *Resection in severe cases*.

Note:
Erythromycin is a prokinetic drug (motilin agonist), which acts by binding with motilin receptor on GI smooth muscle cells.

BARRETT'S OESOPHAGUS (Norman Barrett, British, 1950)

- It is the metaplastic changes in the mucosa of the oesophagus as the result of GORD.
- Squamous epithelium of lower end of the oesophagus is replaced by diseased columnar epithelium (columnar metaplasia).
- There is macroscopic visible length of columnar mucosa with microscopic features of intestinal metaplasia.
- It affects lower oesophagus commonly often middle oesophagus also.

Types (Based on Length)

a. If the length of metaplasia is more than 3 cm, it is called as *long segment Barrett's oesophagus*—classic Barrett type.
b. If the length is less than 3 cm, it is called as *short segment Barrett's oesophagus*.

Histological Types

a. *Gastric type:* Contains chief and parietal cells.
b. *Intestinal type:* Contains goblet cells.
c. *Junctional type:* Contains mucous glands alike of gastric cardia.

Cardia metaplasia is metaplasia at OG junction without any macroscopic change in gastroscopy.

This diseased columnar epithelium is more prone for malignant transformation, i.e. when there is intestinal metaplasia, risk of malignant transformation increases. More the amount of dysplasia more is the risk of malignant transformation. *Dysplasia may be indefinite; low grade or high grade.*

Clinical Features

- Features of GORD.
- Haematemesis.
- Common in *men*; common in whites.

Complications of Barrett's oesophagus

- Ulcerations and stricture
- Dysphagia
- Bleeding
- Perforation
- Adeno carcinoma of O-G junction (25 times more common)

Management

- Regular endoscopic biopsy and surveillance for low grade dysplasia.
- Ablation of Barrett's oesophagus by laser.
- Photodynamic therapy—through endoscopy.
- Argon beam coagulation.
- Proton pump inhibitors—high dose for 3-6 months.
- Antireflux treatment by surgery.
- Resection—*Always better choice*—for high grade dysplasia. *Transhiatal* oesophagectomy is preferred.
- Endoscopic mucosal resection.

BARRETT'S ULCER

It is an ulcer in columnar epithelium lined Barrett's oesophagus at or just above the squamocolumnar junction.
It is more prone for:

- *Bleeding.*
- *Perforation.*
- *Adenocarcinoma of oesophagus.*

Treatment for Barrett's ulcer is endoscopic biopsy and resection.

OESOPHAGEAL MOTILITY DISORDERS

Primary

- ***Achalasia***, vigorous achalasia.
- ***Diffuse and segmental oesophageal spasm***: It is 5 times less common than achalasia, mainly primary spasm of the oesophagal body, predominantly presenting only as chest pain and less severe dysphagia. There is hypertrophy of oesophageal wall muscle with degeneration of vagal fibers; Simultaneous multipeak waveforms of high amplitude and long duration with 20 or more simultaneous waveforms out of 10 wet swallows is diagnostic. Often it can be segmental or involve distal 2/3rd of oesophagus. LES resting pressure is normal with normal deglutition relaxation time.

Barium swallow shows *corkscrew* pattern. Treated with nitrates, calcium channel blockers, bougie dilatation, long oesophageal myotomy either video assisted or thoracotomy approach.

- ***Nutcracker oesophagus*** (*supersqueeze oesophagus*): Common hypermotility disorder with high amplitude peristalsis, equal in both sexes,. Severe chest pain (noncardiac), dysphagia without regurgitation, odynophagia are the features. Waves of hypertensive amplitude pressure > 180 mmHg often can become very high > 400 mmHg; long duration contraction waves > 6 seconds with normal LES pressure and LES relaxation. It is treated with nifedipine, nitrates, antispasmodics, occasionally long myotomy.
- ***Hypertensive LES:*** LES pressure is above normal (> 26 mm Hg). LES relaxation, amplitude pressure, contraction waves, oesophageal body peristalsis are normal. Botulinum, balloon dilatation, modified Heller's myotomy are the treatment.
- ***Ineffective oesophageal motility disorders***: It is irreversible contraction abnormality of distal oesophagus in association with GORD. It is often due to secondary inflammation of oesophageal body following more exposure to gastric contents and poor oesophageal acid clearance. Reflux, dysphagia, heartburn, chest pain, are the features. Manometry shows—sum of total number of low amplitude contractions of less than 30 mmHg and nontransmitted contractions exceeds 30% of wet swallows.
- ***Nonspecific oesophageal motility disorders***: It shows slow transit incomplete emptying, with normal or hypertensive LES pressure, LES shows incomplete relaxation, decreased amplitude pressure < 35 mmHg, > 20% nontransmitted, prolonged waves of > 6 secs with abnormal peristalsis. Dysphagia, chest pain reflux and regurgitations are the features.

Secondary

- *Neurological*—stroke, bulbar palsy, motor neuron disease, multiple sclerosis, Parkinson's disease, poliomyelitis.
- *Muscular*—myasthenia gravis, muscular dystrophy, dermatomyositis.
- *Autoimmune* disorders—systemic sclerosis, polymyositis, SLE, CREST syndrome, sclreoderma.
- Eosinophilic allergic oesophagitis, alcoholic neuropathy.
- Endocrine and metastatic diseases.

ACHALASIA CARDIA (Cardiospasm)

> *The symptom—complex in cardiospasm (the term by which oesophageal achalasia was formerly known) is as a rule almost pathognomonic. It may be divided into the three stages—first, cardiospasm without food regurgitation; second, cardiospasm with immediate food regurgitation; third, cardiospasm with dilated oesophagus, the retention of food in the dilated portion and its regurgitation at irregular intervals after taking.*
>
> — **Henry Stanley Plummer,** 1908

It is failure of relaxation of cardia (oesophagogastric junction) due to disorganised oesophageal peristalsis, as a result of failure of integration of parasympathetic impulses causing *functional obstruction* (*Achalasia* means failure to relax—Greek). It is first identified by Thomas Willis in 1672.

Aetiology

- There is absence or less numbered ganglions in myenteric plexus.
    - Stress.
    - Vit B_1 deficiency.
    - Chaga's disease. It is caused by *Trypanosoma cruzi,* which is common in South America called as sleeping sickness.
    - Diffuse oesophageal spasm *(Corkscrew oesophagus).*
    - Most commonly it is *idiopathic*. There is *degeneration of Auerbach's myenteric plexus* along the entire length of oesophagus more so in LOS.
- There is pencil shaped narrowing of cardia (O-G junction) with enormous dilatation of proximal oesophagus, which contains foul smelling fluid and is more prone for aspiration pneumonia.
- Achalasia cardia *is a precancerous* condition—seven times chances of getting *squamous cell carcinoma* (8%, after 15 years).

Pathology

- Thickening of circular muscle fibres in distal oesophagus.
- Myenteric inflammation; depletion of ganglion cells; neural fibrosis, reduced nitric oxide and VIP (mediators of LES relaxation).
- Absence of peristalsis; Raised LES pressure; failure of relaxation with functional obstruction of OG junction.
- Dilatation of proximal oesophagus with atony.

Clinical Features

- Common in females between 20 and 40 years age group.
- Incidence is 6 per 1, 00,000 population.
- Chest pain occurs in early stage.
- Achalasia with diffuse oesophageal spasm is called as '*vigorous achalasia'*.
- Presents with progressive dysphagia, which is more for liquid than to solid food.
- Regurgitation and recurrent pneumonia are common (10%).
- Walking while eating, chin thrusting, neck and shoulder extension, Valsalva manoeuvre facilitates emptying of food from the oesophagus.
- Heartburn (50%) is common.
- Malnutrition and general ill health.
- Lung abscess formation.
- Odynophagia and weight loss.

Triad

- Dysphagia
- Regurgitation
- Weight loss

Staging

I: Proximal dilatation < 4 cm
II: Dilatation between 4-6 cm
III: Dilatation > 6 cm
IV: Sigmoid dilatation

Investigations

- *Barium swallow is diagnostic*—shows.

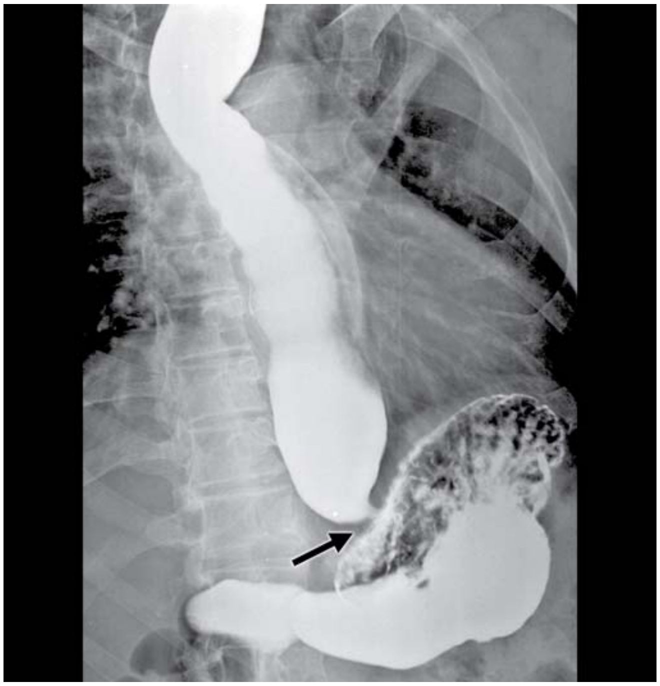

Fig. 19.17: Barium swallow X-ray showing features of achalasia cardia.

- Pencil like smooth narrowing of lower oesophagus—*Bird beak appearance*
- Dilatation of proximal oesophagus
- Absence of fundic gas bubble
- Sigmoid oesophagus or megaoesophagus

- Chest X-ray shows patches of pneumonia. *Double mediastinal strip* of dilated oesophagus is typical with air fluid level in posterior mediastinum on lateral view.
- Oesophageal manometry shows unrelaxed lower oesophageal sphincter with high resting pressure—very useful and gold standard. It shows failure of LES to relax completely during swallowing and complete absence of peristalsis. LES pressure is > 35 mmHg. Base line oesophageal pressure will be high without progressive oesophageal peristalsis with low amplitude muscular tone. Intraoesophageal pressure in relation to intragastric pressure is elevated.
- Oesophagoscopy is done to confirm the diagnosis and to rule out carcinoma oesophagus. It shows totally closed LES with atonic, dilated proximal oesophagus. Biopsy of mucosa at LES should be done.

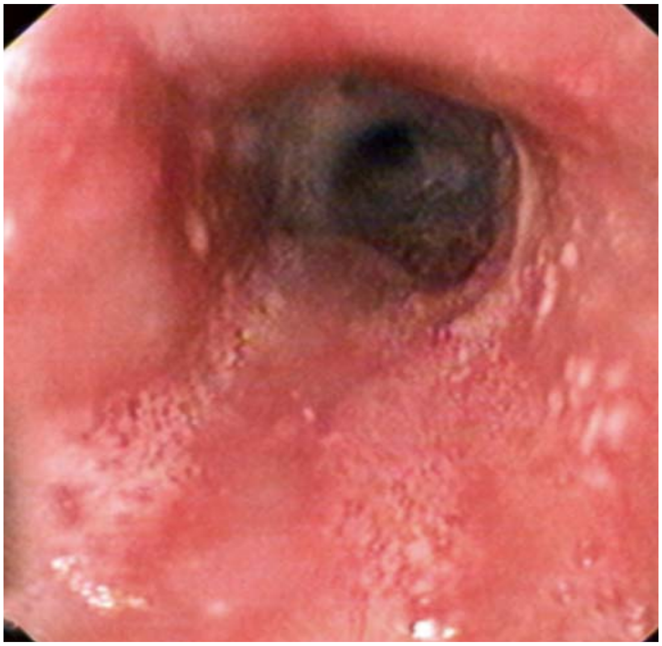

Fig. 19.18: Endoscopic view of achalasia.

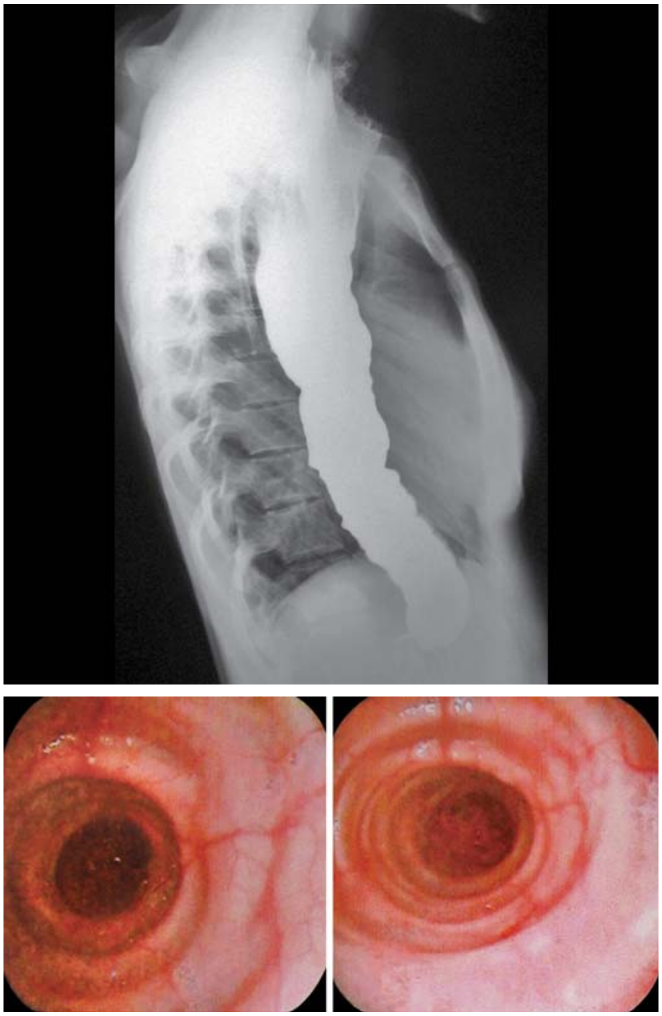

Fig. 19.19: Sigmoid oesophagus in achalasia cardia—Barium X-ray and endoscopic view.

Differential Diagnosis

- Carcinoma oesophagus.
- Stricture oesophagus.
- Scleroderma.

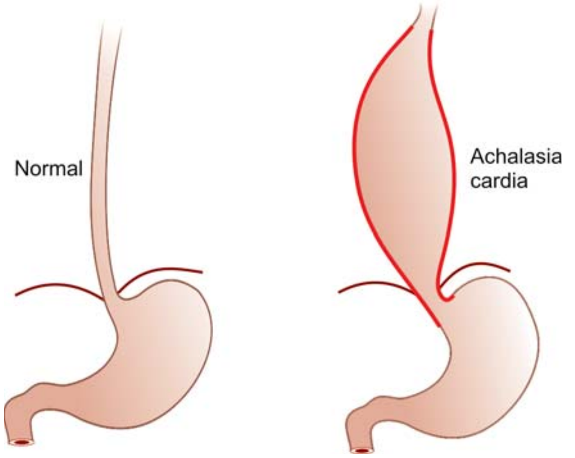

Fig. 19.20: Diagrammatic representation of achalasia cardia.

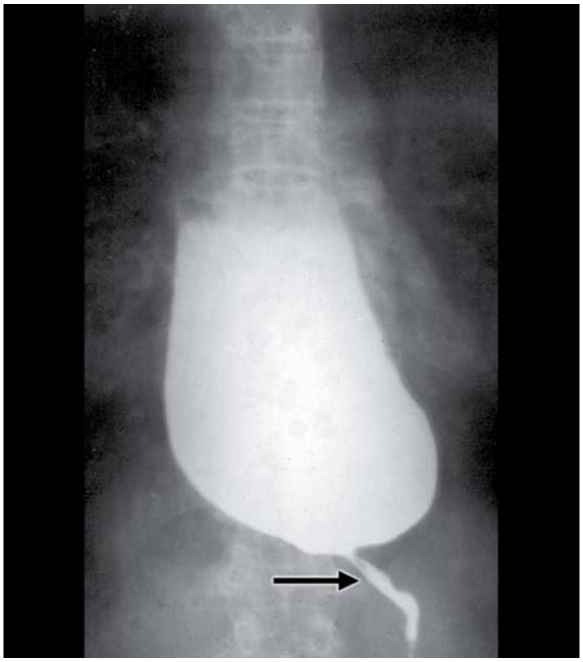

Fig. 19.21: X-ray picture of achalasia cardia.

Treatment

Treatment for achalasia

- *Forcible dilatation*:
  - Plummer's pneumatic dilatation
  - Negus hydrostatic balloon dilatation
- Modified *Heller's* cardiomyotomy
- *Drugs*
  - Botulinum toxin A
  - Nitroglycernie, nifedipine

Surgery

- *Modified Heller's operation* (Heller—German, 1913): Oesophagocardiomyotomy. Success rate is 85%.

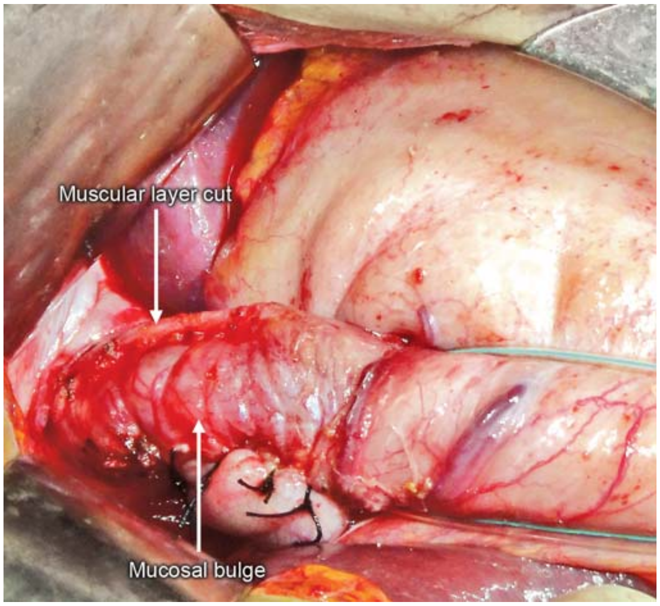

Fig. 19.22: Heller's cardiomyotomy—on table look with partial fundoplication (*Courstesy:* Dr Yogishkumar, Professor of Surgery, KMC, Mangalore).

  - Either through thoracic or through abdominal approach, thickened circular muscle fibres are cut longitudinally for about 8-10 cm, 2 cm proximal to the thickened muscle to 1 cm distal to OG junction. Care should be taken not to open the mucosa (Anterior myotomy is done now. Original Heller's is both anterior and posterior myotomies).

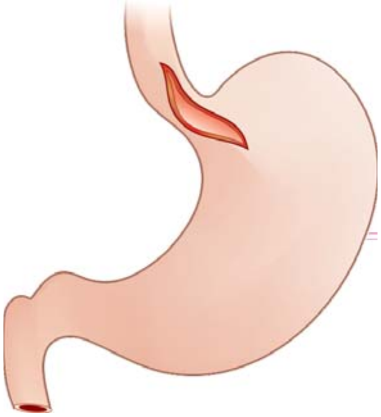

Fig. 19.23: Heller's cardiomyotomy for achalasia cardia. Only circular muscle layer is cut longitudinally in OG junction until mucosa protrudes out without perforating the mucosa.

  - *Nissen's or Toupet's fundoplication* is done along with the above procedure to prevent reflux.
- *Forcible dilatation* stretches spasmodic segment. Gradual repeated dilatations are required. Dilatation up to 54 French bougie can be done. Two types of dilatations are used—*pneumatic* and *hydrostatic*.
  - *Plummer's pneumatic dilatation* is done using balloons of 30-40 mm diameters. It is inserted over a guidewire. Risk of perforation (1%), need for repeated dilatations are the problems.
  - *Negus hydrostatic dilatation* is done to dilate O-G junction. It is not very well accepted method as chances of

perforation is high. Success rate is 65%. 30 mm diameter balloon is inflated for 3 minutes.

- *Laparoscopic/thoracoscopic* cardiomyotomy—ideal.
- Resection is done when failure of myotomy occurs or when megaoesophagus or metaplasia is present. *Transhiatal* total oesophagectomy with gastric pull up and oesophagogastric anastomosis in the neck is a good option in such patients.

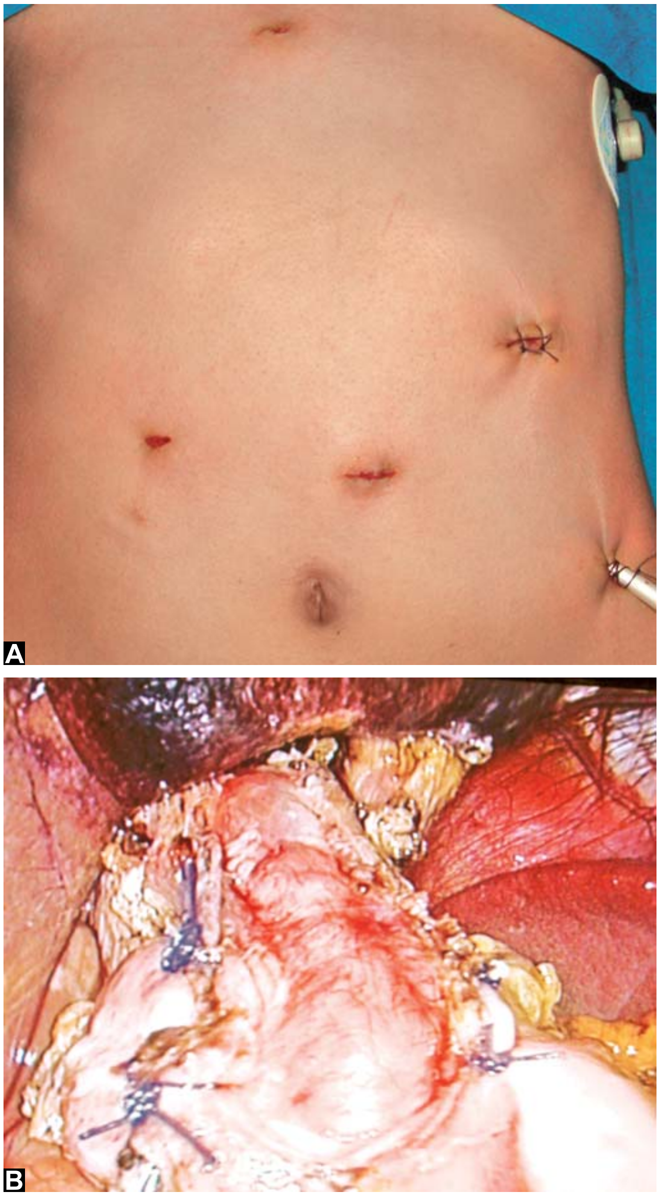

Figs 19.24A and B: Laparoscopic port positions and view of Helller's cardiomyotomy for achalasia cardia (*Courtesy:* Dr Roshan, MS DNB, Laparoscopic Surgeon, City Hospital, Mangalore).

Drugs

a. Endoscopic injection of botulinum toxin to sphincter—high recurrence rate.
b. Calcium channel blocker, nitroglycerin sublingually.

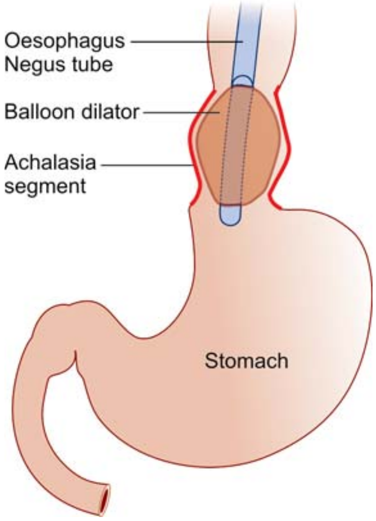

Fig. 19.25: Negus hydrostatic dilator used for achalasia cardia. It has got only 65% success rate.

Note:

- *Pseudoachalasia* shows features like of achalasia cardia with dysphagia and weight loss, seen in an elderly due to carcinoma. Amylnitrate inhalation causes sphincter relaxation in Achalasia cardia but not in pseudoachalasia.
- *Botulinum toxin* is neurotoxic protein derived from Clostridium botulinum is highly toxic poisonous substance. Very small dose is used for therapeutic purpose. Seven types of toxins are found. A (A1 /A2 /A3) type is used for therapy. It blocks the cholinergic nerve ends reducing the cholinergic acetylcholine release causing flaccid paralysis of muscles. It is used in cosmetic facial line, strabismus, focal dystonia, tremor, tics, muscle spasms, achalasia, smooth muscle hyperactivity, Frey syndrome, and hyperhidrosis.

PLUMMER-VINSON SYNDROME (Paterson-Kelly Syndrome)

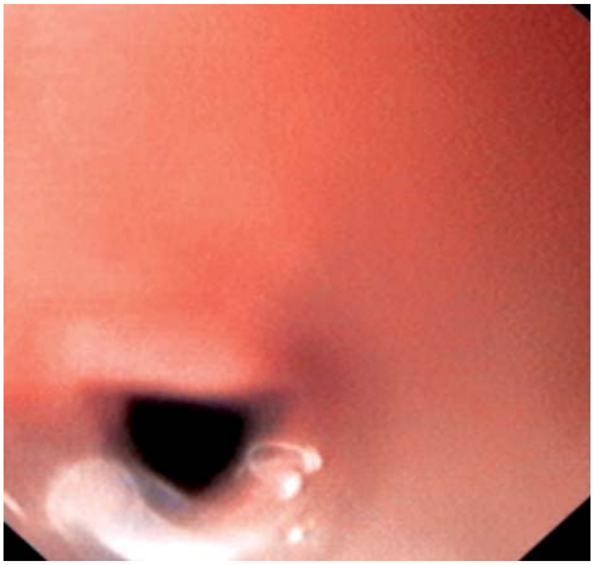

Fig. 19.26: Oesophageal web, endoscopic view.

- Here *oesophageal webs* are seen in uppermost portion of the oesophagus with spasm of circular muscle fibres.
- Common in patients with long standing *iron deficiency anaemia*.
- Common in females.
- *Superficial glossitis*, *cheilitis, koilonychia commonly seen.*
- Splenomegaly may be present.
- In oesophageal webs, mucosa is hyperkeratotic, friable, desquamated.
- It is a *premalignant condition* and presents with severe dysphagia.
- Oesophagoscopy and biopsy is required to rule out malignancy.

Treatment

- Oral iron—ferrous sulphate 300 mg TDS with vitamins.
- Blood transfusion is given when there is severe anaemia (Transfusion of packed cells).
- IV or IM iron therapy.
- Once anaemia comes under control, webs will clear and patient can swallow.
  Follow-up endoscopy is a must.
- Dilatation of web may be required.

CORROSIVE STRICTURE OF OESOPHAGUS

Features of oesophageal corrosive lesion

- *Acute/immediate*
  - Severe pain, shock, laryngeal oedema
  - Mediastinitis, septicaemia, haemorrhage, perforation
- *Late/chronic*
  - Dysphagia
  - Stricture—50%
  - Severe malnutrition
  - Recurrent respiratory infection
  - Oesophageal shortening
  - Malignant changes
  - Tracheo-oesophageal fistula formation

Corrosive strictures can be multiple. Damage is more in lower 1/3rd of oesophagus

- Corrosives are commonest cause of oesophageal stricture.
- Mainly due to ingestion of alkali (*Lye stricture*—Lye is strong alkali sodium hydroxide) sodium hydroxide, occasionally due to acid (sulphuric acid, nitric acid). Acid commonly damages the stomach.
- It causes extensive inflammation of the mucosa with perioesophagitis which, if not treated leads to multiple strictures in oesophagus.
- Sometimes it causes severe life-threatening necrotising lesion which requires immediate surgical intervention.
- Acute phase lasts for 3 weeks.
- Damage is more in lower 1/3rd of oesophagus.
- Alkali is odourless and tasteless and so large volume is ingested. Alkali causes liquefaction, saponification and thrombosis of vessels and later leading to fibrosis and stricture. Acid causes intense pylorospasm, antral pooling of acid, coagulation necrosis and eschar formation.
- Severity depends on type of agent, its concentration and volume.

Phases of tissue injury in corrosive ingestion	***Degrees of burns after corrosive oesophageal and gastric burns***
Phase 1: Acute necrosis—1-4 days	*1st degree*: Mucosal hyperaemia and oedema
Phase 2: Ulceration—granulation—4-12 days	*2nd degree*: Small bleeding, exudates, ulcers, pseudomembrane
Phase 3: Cicatrisation and scarring—3 weeks–6 months	*3rd degree*: Mucosal slough, deep ulcers, massive bleed, complete obstruction, charring, perforation

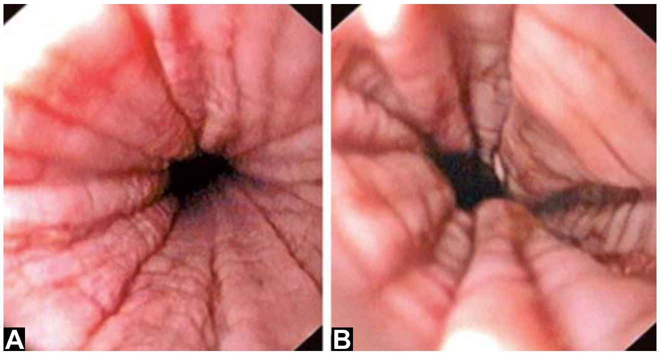

Figs 19.27A and B: Endoscopic view of corrosive injury. Endoscopy should be done very gently in corrosive injury.

Treatment

- *Acute phase management*:
  - *Neutralisation* with vinegar or citrus food if it is alkali ingestion (If pH of the solution is less than 11.5 then damage is less); it is with antacids, milk, egg whites if it is acid ingestion. Early endoscopy is needed to assess the severity and extent.
  - Emetics and $NaHCO_3$ are *avoided* as they may precipitate perforation.
  - *In 1st degree burns*: 48 hours observation; oral feeds are started once patient swallows saliva painlessly. Regular follow-up endoscopy at 1st, 2nd and 8th months. Stricture if formed can be identified by this time.
  - *2nd and 3rd degree burns*: They are treated with fluid therapy, antibiotics, nutrition, resuscitation, PPIs, aerosolised steroids, fiberoptic guided airway intubation if needed / tracheostomy; endoscopic oesophageal stenting, feeding jejunostomy, laparoscopy for evaluation. Unstable patients have high mortality. Laparotomy is done in such patients. If oesophagus and stomach shows full thickness necrosis, resection of these parts is done and end cervical oesophagostomy with jejunostomy is done. When in doubt re-exploration for second look is done after 36-48 hours to assess the stomach.
- Careful early gentle repeated endoscopy is mandatory.
- Though advocated often for 2-3 weeks, use of steroids is controversial and under debate.

- Antibiotics if there are chances of aspiration or perforation.
- *Regular oesophageal dilatation* is done for stricture. Stricture is dilated endoscopically using guidewire. Dilators are solid type with gradual increase in diameters. Often radiologic C-ARM guidance is needed to pass the guide wire into the stomach. Dilatation should be done up to minimum 16 mm diameter. Pneumatic or balloon dilatation is also practiced. Gum elastic dilators, mercury weighted dilators, Eder-Puestow dilators, Savary-Gilliard dilators, balloon dilators are other dilators used. Earlier, blind dilatation using oesophageal bougies of increased diameters was the practice, which is followed even now in many places, but chances of perforation is more.
- Oesophageal resection in corrosive strictures is technically difficult and may be hazardous. Oesophageal bypass is better and easier, and following later by regular endoscopic surveillance for malignant transformation (5%). Colon is used as replacing conduit as stomach itself may be diseased in corrosive pathology.
- In multiple strictures oesophageal resection and colonic transposition may be advocated if risk of malignancy is considered.

Note:

Malignancy can develop in corrosive strictures—5%. It is 1000 fold greater than general population.

Causes of stricture oesophagus

- Peptic stricture (oesophagitis induced)
- Corrosives—most common cause
- Foreign body
- Postsurgical, radiotherapy
- Congenital
- Infection like tuberculosis
- Drugs like tetracycline, vitamin C

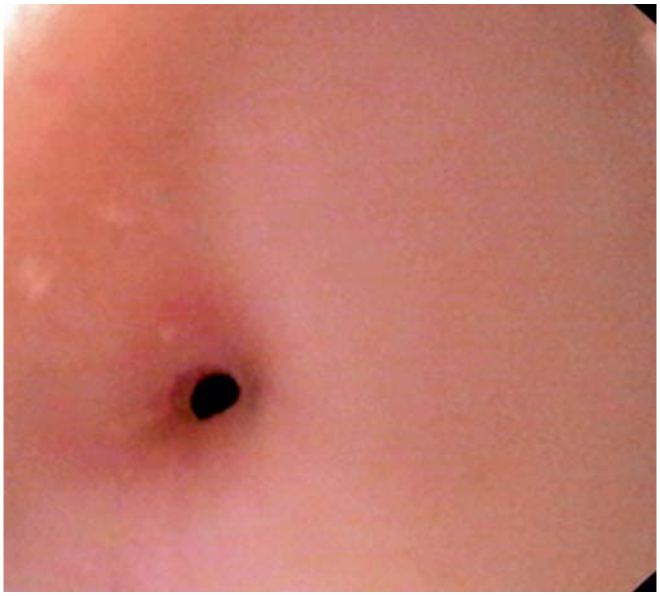

Fig. 19.28: Endoscopic view of oesophageal stricture.

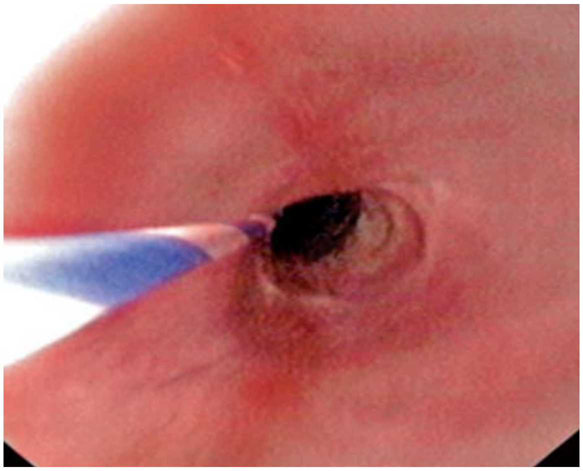

Fig. 19.29: Oesophageal dilatation is being done endoscopically.

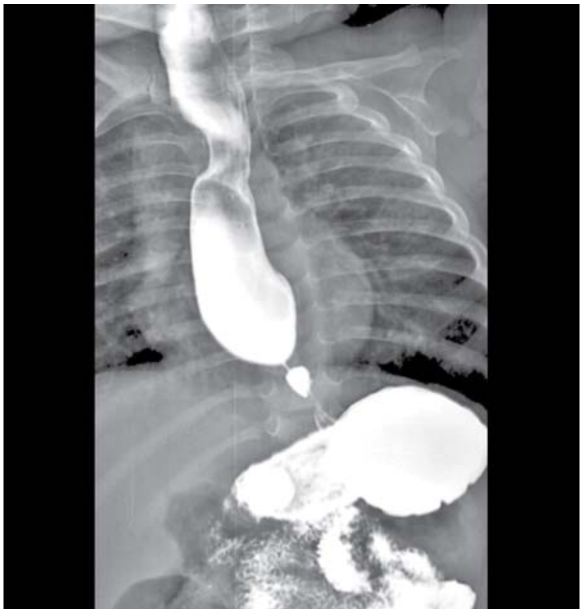

Fig. 19.30: Barium swallow showing stricture at lower oesophagus. It may be due to oesophagitis, congenital, tuberculosis, malignancy, drug induced, etc.

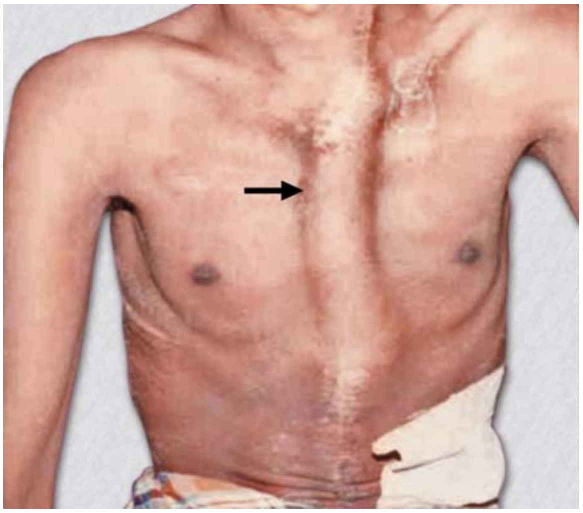

Fig. 19.31: Patient underwent total oesophagectomy with colonic transposition. Colon placed in subcutaneous plane.

All are equal in their ignorance.

SCHATZKI'S RINGS

They are semicircular protrusion of lower oesophageal mucosa located at or just above the oesophagogastric junction (squamocolumnar junction). Its under-surface is lined by columnar gastric epithelium.

- They involve only the mucosa and submucosa of the oesophagus, not the muscle.
- They present with dysphagia and reflux.
- *Episodic aphagia* can occur causing, food bolus or meat bone to get impacted which requires emergency rigid oesophagoscopy to remove the food. 5 ml of 2.5% oral papain every 30 minutes to digest food protein along with 50 mg meperidine IV to dislodge the impacted food bolus can be tried initially.

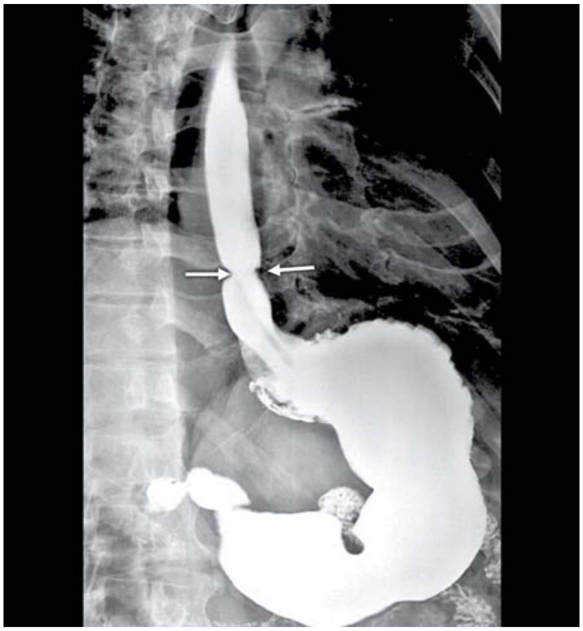

Fig. 19.32: Barium swallow X-ray showing Schatzki ring, also shows features of hiatus hernia.

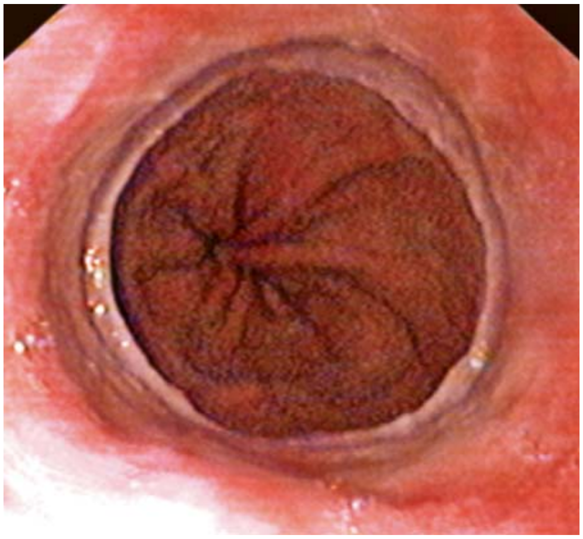

Fig. 19.33: Schatzki's ring—endoscopic view.

Treatment

- Intermittent oesophageal bougienage.
- Antireflux drugs.

Note:

Ring *should not* be excised.

BOERHAAVE'S SYNDROME

It is a tear in the lower third of the oesophagus which occurs when a person vomits *against a closed glottis,* causing leak into the mediastinum, pleural cavity and peritoneum.

Site

2-10 cm of posterolateral part of lower oesophagus.

Investigations

- Chest X-ray—shows pneumomediastinum *('V' sign of Naclerio*).
- MRI/CT chest.
- Total count.

Treatment

- Nil by mouth.
- Antibiotics, IV fluids, TPN.
- Feeding by jejunostomy.
- Most often surgery with resection may be required (thoracotomy and repair).
- *When severe mediastinitis occurs, condition has high mortality.*

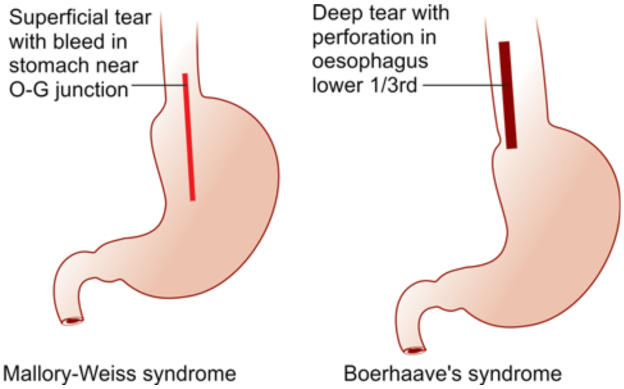

Superficial tear with bleed in stomach near O-G junction
Deep tear with perforation in oesophagus lower 1/3rd
Mallory-Weiss syndrome
Boerhaave's syndrome

Fig. 19.34: Diagrams showing Mallory-Weiss syndrome and Boerhaave's syndrome.

MALLORY-WEISS SYNDROME

- It is seen in adults with severe prolonged vomiting, causing longitudinal tear in the mucosa of stomach at and just below the cardia, *leading to severe haematemesis*.
- Violent vomiting often may be due to migraine or vertigo or following a bout of alcohol.
- Presents with severe *vomiting* and later *haematemesis*, with features of *shock.*
- It is common in one o'clock position.
- Only 10% of cases involve lower oesophageal mucosa.

Features of Boerhaave's syndrome

Presentations	***Differential diagnosis***	***Complications***
Sudden onset of symptoms	Myocardial infarction	Mediastinitis
Severe chest pain	Pancreatitis	Septicaemia
Pain abdomen		Peritonitis
Subcutaneous emphysema Shock		Emphysema

'Crunching effect in the chest' is called as *Hamman's sign*.

Mackler's triad: (1) vomiting, (2) chest pain, (3) subcutaneous emphysema.

Investigations

- Gastroscopy, Hb%, PCV, blood grouping.
- During gastroscopy, if stomach is not inflated properly, 50% cases may be missed.

Differential diagnosis

- Bleeding peptic ulcer
- Oesophageal varices
- Erosive gastritis
- Carcinoma stomach

Treatment

- Conservative, as it is only a mucosal tear.
- Blood transfusion.
- IV fluids.
- Sedation.
- Haemostatic agents like vasopressin.
- Endoscopic injection therapy is used if required.
- Surgery is rarely required.

TRACHEO-OESOPHAGEAL FISTULA

Types

- In **85% cases,** it is a blind upper end with lower end communicating with trachea.
- It may be associated with **VACTER** anomalies.

V — Vertebral defects
A — Anal atresia
C — Cardiac defect (PDA/VSD)
TE — Tracheo-oesophageal fistula
R — Radial hypoplasia and renal agenesis

Clinical Features

- TOF should be recognised within 24 hours of birth.
- Newborn baby regurgitates all feeds and there is continuous pouring of saliva from the mouth which is a *diagnostic feature.*
- *Cough and cyanosis*.
- It is commonly associated with maternal *hydramnios (50%).*

Investigations

- Obstruction is revealed while passing nasogastric tube.
- Contrast study will reveal fistula and obstruction (Dionosil 1 ml).
- Other anomalies are looked for.
- Chest X-ray.
- Echocardiography.

Treatment

Surgery

- Child requires feeding gastrostomy commonly.
- Often the procedure is staged one.
- Through right sided thoracotomy (opposite to the side of aortic arch), fistula is identified and resected. Lower segment is anastomosed to the blind upper segment. Occasionally if the length is inadequate or the atretic segment is long one then, colonic or gastric transposition is required.

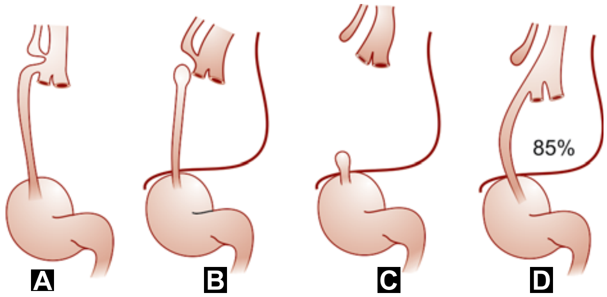

Figs 19.35A to D: Types of tracheo-oesophageal fistula. (A) H type, (B) Lower end blind, upper end connected to trachea, (C) Both ends blind, (D) Upper end blind, lower end connected to trachea (85%).

Complications of surgery

- Pneumonia
- Leak from the anastomotic site
- Reflux
- Dysphagia

OESOPHAGEAL DIVERTICULUM

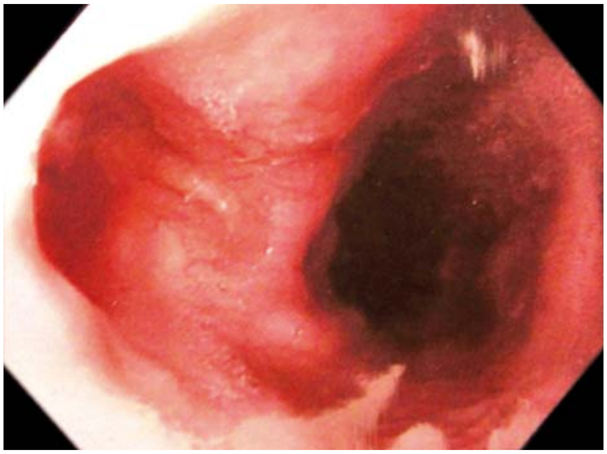

Fig. 19.36: Endoscopic view of oesophageal diverticulum

All things must change to something new and strange.

Types

1. ***Pulsion diverticulum:***

Pulsion diverticula are *false type* containing *mucosa and submucosa* only; is due to high abnormal intraluminal oesophageal pressure developed due to various motility disorders.

   a. Pharyngeal pouch through Zenker's or Killian's dehiscence (*Refer chapter NECK*).

   b. *Epiphrenic pulsion* diverticulum occurs in lower oesophagus, usually towards right side, due to obstruction in the distal oesophagus or due to incoordinated LOS relaxation.

      – *Site* is within 10 cm of OG junction. It is *false* type.
      – It is associated with *nonspecific oesophageal motility* disorders and often with achalasia and diffuse oesophageal spasm. Ehler Danlos syndrome and trauma are other causes.
      – It is common on *right* side with wide mouth.
      – *Features* are of motility disorders like dysphagia, regurgitation, cough, weight loss, chest pain.
      – Barium study, CT chest are *diagnostic*; endoscopic evaluation with EUS, manometry is a must.
      – *Treatment:* Diverticulopexy/diverticulectomy (excision) + oesophageal myotomy (Heller's) + repair of associated hiatus hernia/antireflux procedure.

2. ***Traction diverticulum:***

Occurs in mid-oesophagus or in parabronchial region, is due to mediastinal granulomatous disease like tuberculosis.

- Traction diverticulum is a *true* type containing all layers in its wall and is due to traction by the healing fibrosing mediastinal lymph nodes.
- It is seen commonly towards right side. It has got wide mouth and it rests on the spine.
- *Presentation* is dysphagia, chest pain and regurgitation.
- CT scan (chest), barium study, manometry, endoscopy to assess mucosa with EUS, blood test for tuberculosis (ESR, peripheral smear) are the *investigations*.
- *Treatment:* Treating the *cause* like tuberculosis, histoplasmosis. Diverticulum less than 2 cm is observed; progressive symptoms, size > 2 cm needs surgery. *Surgeries* are—diverticulopexy, long oesophageal myotomy.

CARCINOMA OESOPHAGUS

- Carcinoma oesophagus is common in China, South Africa and Asian countries.
- It is 6th most common cancer in the world.
- It is less than 1% of all cancers. It is 7% of all GI malignancies.
- It is less common in America and European countries.
- In India, it is common in Karnataka and Orissa.
- When patient presents with dysphagia, often it is fairly advanced and inoperable and only palliation is the possibility. But then surgery is the treatment of choice in early growths.

Aetiology

Diet, deficiencies (vit. A, C, Riboflavin)
Mycotoxin
Alcohol and tobacco
Fungal contamination of food

Achalasia cardia
Oesophageal webs
Barrett's oesophagus
Plummer –Vinson's syndrome
Corrosive strictures
Tylosis*
Nitrosamines

5% common
Common after 45 years
Common in men
Common in ***China*** – ***Henan*** province
In India, common in Orissa and Karnataka

* *Tylosis* is an inherited disease with thickening of skin of palm and sole

Tylosis

- Autosomal dominant condition seen from childhood
- Soles and palms are involved called as palmoplantar keratoderma
- Waxy, yellow lesions, which does not itch
- 60% of members of families develop carcinoma oesophagus after the age of 60
- Systemic retinoids are the drugs used for tylosis

Pathology

- Common in:
  - Middle third—50%.
  - Lower third—33%.
  - Upper third—17%.
- Lower 3 cm of oesophagus is lined by columnar epithelium, and so adenocarcinoma is common here. Barrett's columnar metaplasia which occurs in lower third of oesophagus is also more prone for adenocarcinoma.
- Squamous cell carcinoma is commonest type in India and Asian countries.

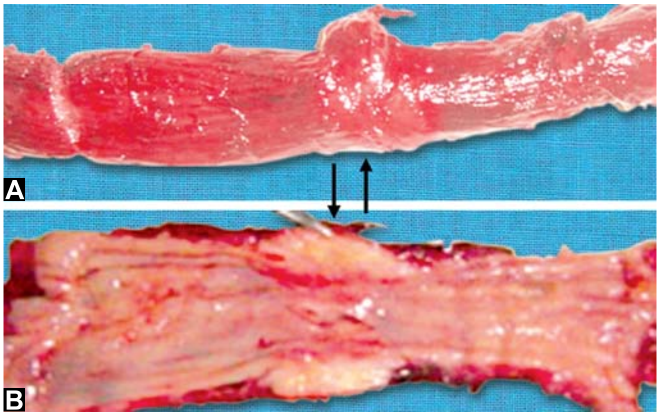

Figs 19.37A and B: Carcinoma oesophagus middle third—pathology specimen, gross and cut section.

Note:

- *In India 90% are squamous cell carcinomas.*
- *In western countries, adenocarcinoma is becoming more common.*

Adenocarcinoma arises from submucosal oesophageal glands/heterotrophic columnar epithelium/Barrett's oesophagus.

Gross Types

- Annular (15%).
- Ulcerative (20%).
- Fungating—cauliflower like (60%).
- Polypoid.
- Varicoid—diffuse submucosal type.

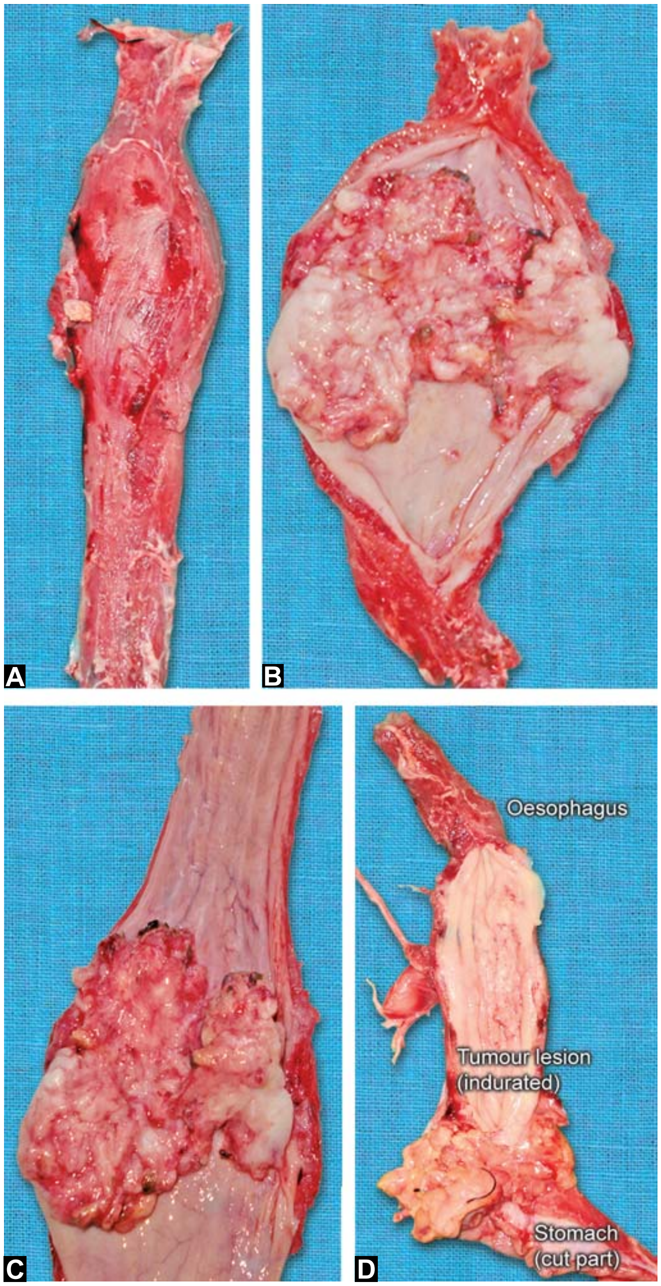

Figs 19.38A to D: Gross outer look and cut section of proliferative and indurated lesions of carcinoma oesophagus.

Spread

- *Direct*
  - Lack of serosal layer in oesophagus favours local extension. In upper third it spreads through muscular layer and get adherent to left main bronchus, trachea, and left recurrent laryngeal nerve (causes hoarseness), aorta or its branches (causes fatal haemorrhage, but rare).
  - It may perforate and cause mediastinitis.
  - It may get adherent to pleura also.
- *Lymphatic spread*
  - It spreads both by lymphatic permeation and lymphatic embolisation.
  - It can cause satellite nodules elsewhere in the oesophagus, away from the main tumour.
  - Above in the neck, it spreads to supraclavicular lymph nodes.
  - In thorax, it spreads to paraoesophageal, tracheobronchial lymph nodes to subdiaphragmatic lymph nodes.
  - In abdomen, it spreads to coeliac lymph nodes.
- *Blood spread occurs* to liver, lungs, brain and bones.

Clinical Features

- *Recent onset of dysphagia* is the commonest feature. *For the dysphagia to develop, two-third of the lumen should be occluded.*
- Regurgitation.
- Anorexia and loss of weight (severe), cachexia.
- Pain-substernal or in the abdomen.
- Liver secondaries, ascites.
- Bronchopneumonia, melaena.
- Features of broncho-oesophageal fistula in carcinoma of upper third oesophagus (30%).
- Left supraclavicular lymph nodes may be palpable.
- Hoarseness of voice due to involvement of recurrent laryngeal nerve.
- Hiccough, due to phrenic nerve involvement.
- Back pain—due to nodal spread (paraoesophageal/*coeliac* nodes).
- Male to female ratio is 3:1. In adenocarcinoma it is 15:1.

Investigations

- *Barium swallow*: Shouldering sign and irregular filling defect.
- *Oesophagoscopy*—to see the lesion, extent and type.
- *Biopsy*—for histological type and confirmation.
- *Chest X-ray*—to look for aspiration pneumonia, *to see vocal cord palsy, to identify fistula.*
- *Bronchoscopy*—to see invasion in upper third growth.
- *Oesophageal endosonography*—to look for the involvement of layers of oesophagus, nodes, cardia and left lobe of the liver. Nodes smaller than 5 mm can be very well visualised by EUS which may be missed in CT scan. EUS guided transmucosal nodal needle aspiration cytology can also be done.
- *CT scan* (95% accuracy)—to look for local extension, nodal status, perioesophageal/diaphragmatic/pericardial (1%)/ vascular infiltration, obliteration of mediastinal fat and status of tracheobronchial tree in case of upper third growth.

- *U/S abdomen*—to look for liver and lymph nodes status in abdomen.
- *Endoscopic oesophageal staining* with labelled iodine results in normal mucosa being stained brown, but remains pale in carcinoma (as mucosa in carcinoma will not take up iodine).
- *Blood tests*: Haematocrit; ESR; Liver function tests.
- *Laparoscopy*: It is useful to see peritoneal spread, liver spread and nodal spread. It is the only reliable method to detect peritoneal seedlings. Biopsy from different places can also be taken. It will prevent unnecessary laparotomy for anticipated surgical resection.
- *PET scan* using 18 F-fluorodeoxyglucose (FDG): 18 FDG is given to the patient. FDG enters highly active cells and gets phosphorylated to FDG 6 phosphate. It stays in the cell as end product and get polarised there. PET with CT scan is used to see response for therapy.
- *Video assisted thoracoscopic approach*—to stage oesophageal carcinoma.
- *Endoscopic mucosal resection (EMR)* is done using double channeled endoscope with tip having a soft plastic cap. Cap is firmly placed over the lesion and suction is created; a snare is brought over the lesion; biopsy specimen is snared off 1.5 cm in size containing mucosa and submucosa. It is basically a diagnostic biopsy tool; but can be therapeutic in early and premalignant lesion. Endoscopic submucosal dissection (ESD, Japan) is now devised using hook cautery and scissors to remove the lesion up to muscularis propria.

Treatment

Gastrostomy should not be done as a palliative procedure.

Principles

- Only 20% of oesophageal cancers present early and becomes *curable*. In such early growths confirmed with absence of nodal spread, curative surgery is the main approach—*radical oesophagectomy*. Proximal extent of resection should be 10 cm above the macroscopic tumour and distal extent of resection is 5 cm from macroscopic distal end of tumour. Proximal stomach has to be removed commonly especially in lower 1/3rd of tumour. Sufficient removal of contiguous structures may be needed in curative resection.
- If nodes are present, then multimodal approach should be used like—curative resection; radiotherapy and chemotherapy. Outcome of surgery depends on location of tumour; number, location and size of nodes; tumour grading.
- Neoadjuvant therapy by chemotherapy and/or radiotherapy prior to surgery may improve the survival.
- Aggressive chemoradiation also may be used as curative therapy in some patients especially upper 1/3rd growths and in patients who are unfit for surgery.
- In remaining patients (80%) *palliation* is the main modality of treatment. Palliation therapy is done if patient is not fit for major surgery; if there is blood spread; if there is spread to adjacent organ; if there is peritoneal/liver spread. It is to relieve pain and dysphagia and also to prevent aspiration and bleeding.

Indications for Curative Treatment

1. Early growth when patient is fit.
2. When there is no involvement of adjacent perioesophageal structures, bronchus, liver or distant organs.

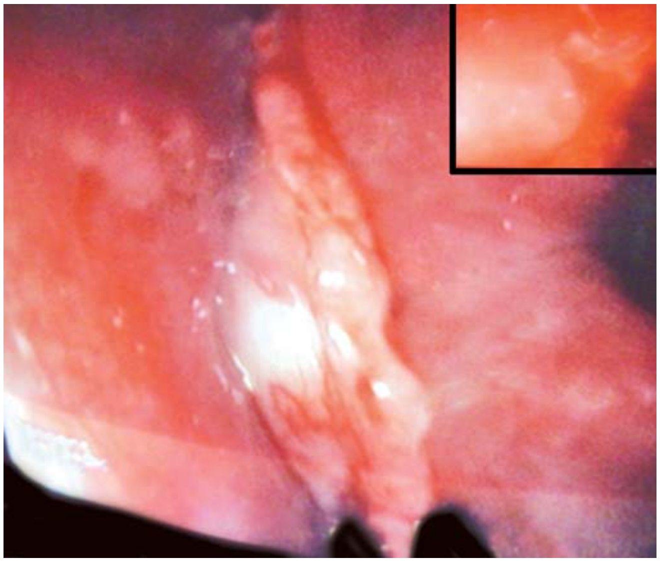

Fig. 19.39: Carcinoma in upper oesophagus (Endoscopic view).

Indications for palliative therapy

- Relieve pain
- Relieve dyspahgia
- Prevent bleeding
- Prevent aspiration

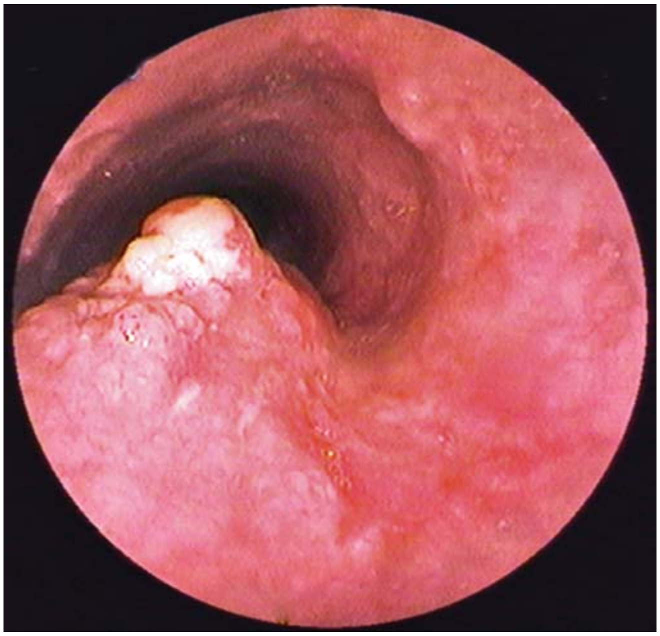

Fig. 19.40: Carcinoma oesophagus (Endoscopic view).

TNM staging	*WNM staging (Wall penetration, Node, Metastases)*
T – Tumour: T_x - Tumour cannot be assessed T_0 – No tumour Tis – High grade dysplasia T_1 – Tumour invades lamina propria, muscularis mucosa or submucosa without breaching submucosa T_2 – Tumour invades into muscularis propria but not beyond T_3 – Tumour invades paraoesophageal tissues without adjacent structure invasion T_4 – Tumour invades adjacent structures **N – Nodal spread:** N_x – Regional nodes cannot be assessed N_0 – No regional node spread N_1 – Regional node spread present **M – Distant metastases:** M_x – Distant spread cannot be assessed M_0 – No distant spread M_{1a} – Upper oesophageal tumour spreads to cervical nodes; middle oesophageal tumour spreads to mediastinal nodes; lower oesophageal tumour spreads to celiac nodes. M_{1b} – Upper oesophageal tumour spreads to mediastinal or celiac nodes; middle oesophageal tumour spreads to cervical or celiac nodes; lower oesophageal tumour spreads to cervical or upper mediastinal nodes. *Staging TNM:* Stage 0 – Tis,N_0,M_0 Stage I – T_1,N_0,M_0 Stage IIA – T_2,N_0,M_0; T_3,N_0,M_0 Stage IIB – T_1,N_1,M_0; T_2,N_1,M_0 Stage III – T_3,N_1,M_0; T_4,anyN,M_0 Stage IVA – AnyT,AnyN,M_{1a} Stage IVB – AnyT,AnyN,M_{1b}	**Wall penetration:** W0 – Intramucosal mucosa penetration W1 – Intramural mucosa penetration W2 – Transmural mucosa penetration **N – regional nodes:** Nx – Regional nodes cannot be assessed N0 – No regional nodes N1 – 1-4 regional nodal spread N2 – More than 4 nodal spread **M – Distant metastases:** Mx – Distant spread cannot be assessed M1 – No distant spread M2 – Distant spread present *Staging WNM:* Stage 0 – W_0,N_0,M_0 Stage I – W_0,N_1,M_0; W_1,N_o,M_0 Stage II – W_1,N_1,M_0; W_2,N_0,M_0 Stage III – W_2,N_1,M_0; W_1,N_2,M_0; W_0,N_2,M_0 Stage IV – AnyW,AnyN,M_1
Variables considered for management: Histology of tumour—SCC / adeno Location of tumour—cervical, mid thoracic, lower Extension of tumour in depth Staging of tumour Distant spread Nutritional status of the patient—good/fair; serum albumin more or less than 3.4 mg% Pulmonary function, cardiac reserve Weight loss less or more than 10% Dysphagia present/absent	*Japanese added to TI of TNM system following:* T_1 – Mucosal epithelial—no node chance T_1 – Lamina propria—5% nodes chance T_1 – Muscularis mucosa—18% nodes chance T_1 – Superficial submucosal—50% node chance T_1 – Middle submucosal T_1 – Deep submucosal—55% node chance

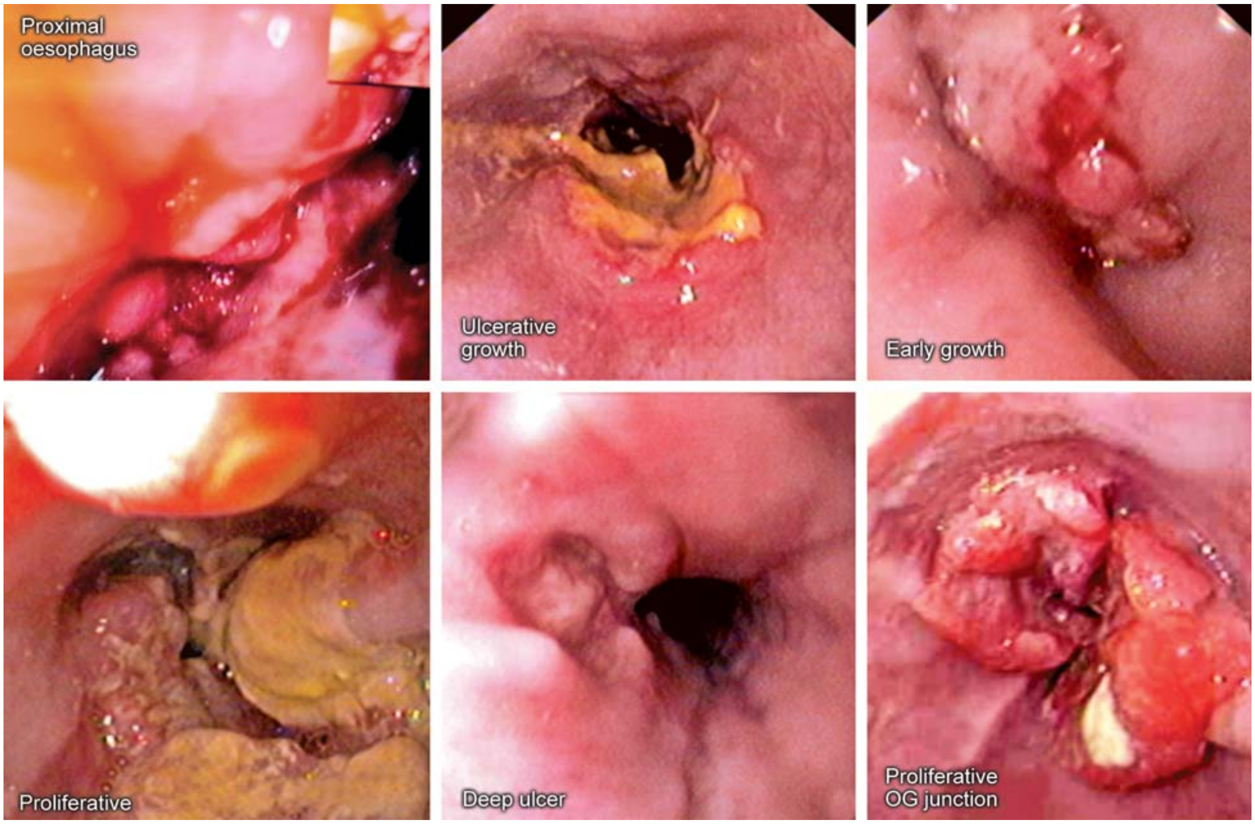
Proximal oesophagus
Ulcerative growth
Early growth
Proliferative
Deep ulcer
Proliferative OG junction

Fig. 19.41: Endoscopic views of oesophageal carcinomas—at different levels and different gross types.

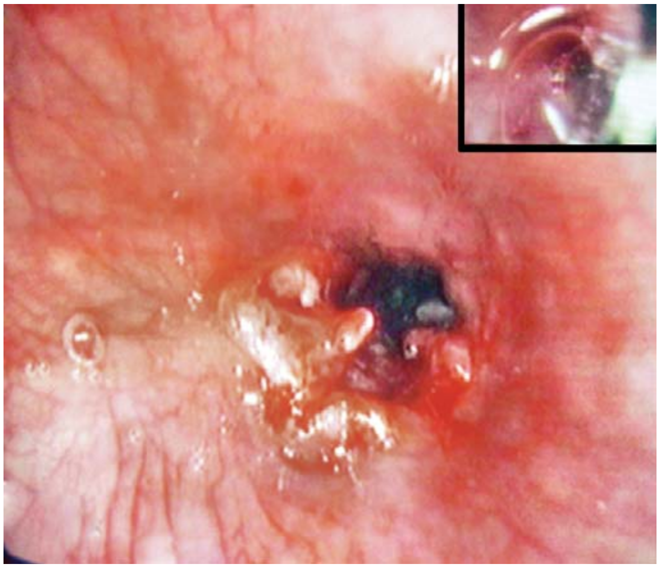

Fig. 19.42: Endoscopic view of carcinoma oesophagus in lower third.

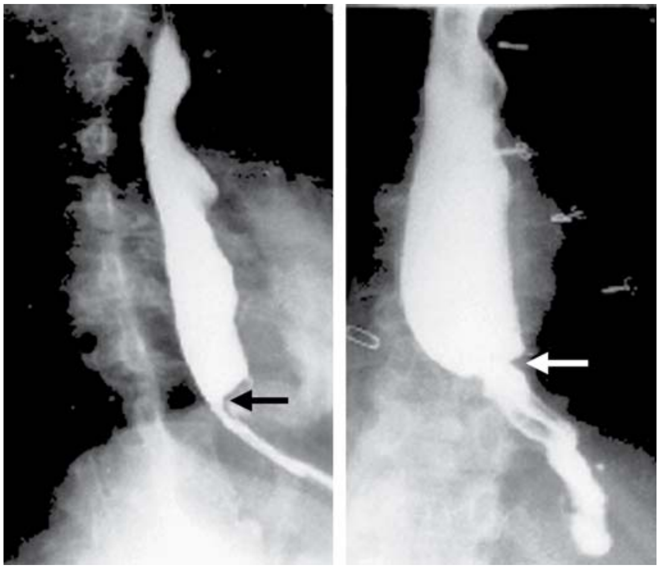

Fig. 19.43: Barium swallow showing irregular filling defect and shouldering sign in middle third oesophagus.

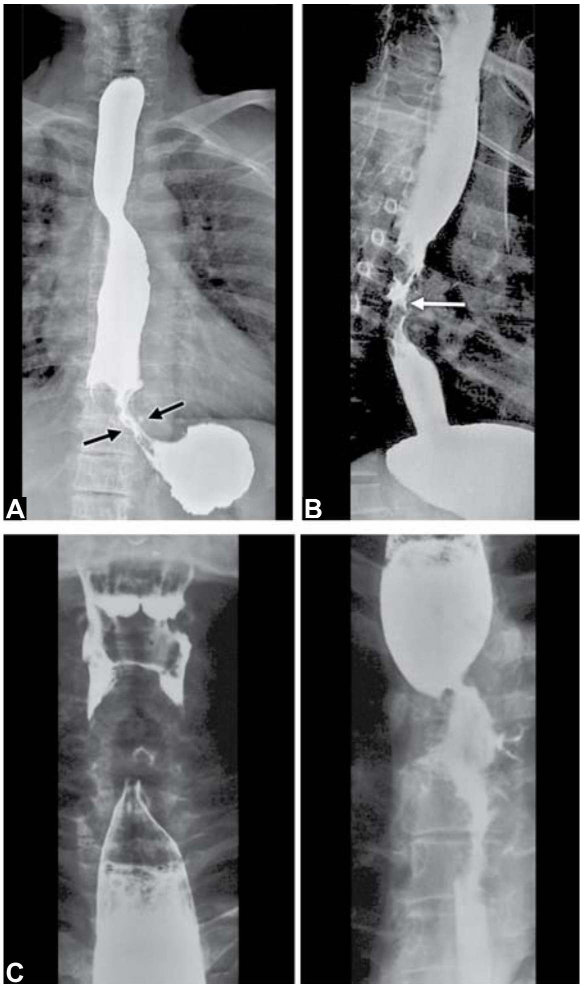

Figs 19.44A to C: Barium swallow X-rays showing irregular filling defect at different levels—feature of carcinoma of oesophagus.

Approaches for Different Level Tumours

Post cricoid tumour (Squamous cell carcinoma):

- Treated mainly by radiotherapy. Radical radiotherapy—5000-6000 rads.
- Often pharyngolaryngectomy is done along with gastric or colonic transposition. But complications are more in this procedure. Free jejunal transfer is the other option.

Upper third growth (Squamous cell carcinoma):

- Treated mainly by radiotherapy.
- Commonly it is advanced with left recurrent laryngeal nerve palsy and bronchial invasion.
- If it is early and operable, *McKeown three phased oesophagectomy* and anastomosis is done in the neck. Initially laparotomy is done to mobilise the stomach. Then thoracotomy through right 5th space is done and oesophagus is mobilised. Through right side neck approach, oesophagus with growth is removed. Anastomosis between pharynx and stomach is done in the neck.
- Split sternum approach oesophagectomy is also practiced.

Middle third growth (Squamous cell carcinoma):

Ivor Lewis operation (Lewis-Tanner two-phased oesophagectomy): After laparotomy stomach is mobilised. Pyloroplasty is done. Through right 5th space thoracotomy is done and growth with tumour is mobilised. Partial oesophagectomy and oesophagogastric anastomosis is done in the thorax. Intercostal tube drainage is placed during closure. Right gastroepiploic vessels should be retained safely (essential). Azygos vein should be ligated securely. Mediastinal nodes should be dissected. Thoracic duct should be ligated if needed. Feeding jejunostomy is better to maintain nutrition.

If the growth is inoperable, palliative radiotherapy is given.

Lower third growth (Squamous cell carcinoma + adenocarcinoma):

- Here through left thoracoabdominal approach, partial oesophagogastrectomy is done with oesophagogastric anastomosis.
  Often jejunal Roux-en-Y loop anastomosis is done.
- *Orringer approach,* i.e. *transhiatal blind total oesophagectomy* with anastomosis in the left side of the neck. Through laparotomy, stomach and lower part of the oesophagus are mobilised. Through left sided neck approach, upper part of the oesophagus is mobilised using finger. Blind dissection is completed by meeting both fingers above and below in the thorax. Later oesophagus is pulled up out above through the neck wound and removed. Continuity is maintained in the neck. It is a *palliative* surgery.

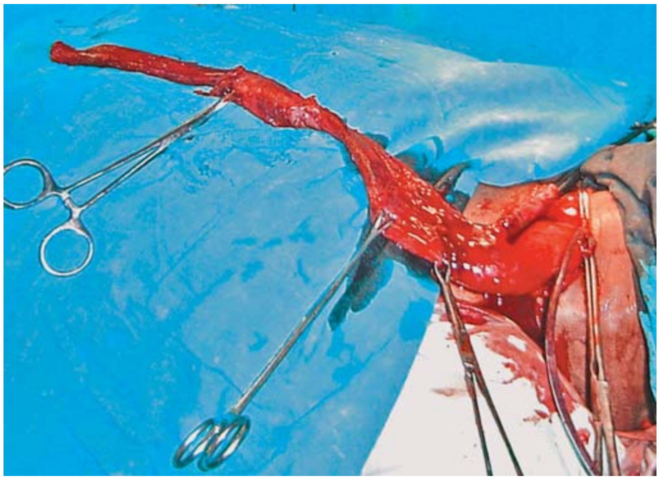

Fig. 19.45: Transhiatal blind oesophagectomy—on table surgical specimen.

Other Approaches

- *Thoracoscopic—laparoscopic oesophagectomy* and lymphadenectomy is becoming popular, safer and effective.
- *Radical oesophagectomy with 3-field clearance* of abdominal/thoracic and cervical nodes is also practiced in many centres. 3-Field clearance (coeliac, thoracic and neck) is done for mainly squamous cell carcinoma as spread in SCC

is upwards. In adenocarcinoma 2-field clearance is sufficient (abdominal—coeliac and thoracic) as spread is downwards.

Note:

When oesophagus is removed totally or subtotally an interposition is required between cervical oesophagus/pharynx and distal stomach. Different parts of GI is used for the same.

Oesophageal substitutes:

- **Stomach**: It is preferred one and is based on right gastric and right gastroepiploic arteries. It needs only one anastomosis and take up is well due to good vascularity. But it can cause postprandial fullness with bile or acid regurgitation (> 50%). It also needs a pyloroplasty for gastric drainage as vagotomy is done. Entire stomach or tubed part or reverse tubed part can be transposed.
- **Colon**: It is better as there is less postprandial problems. But technique is more with the need of three anastomoses. Right colon is used with middle colic artery based as isoperistaltic loop. Left colon is used with ascending branch of left colic artery as antiperistaltic loop. Complications are leak, fistula formation.
- **Jejunum**: Isolated required length of pedicled jejunum is transposed. Jejunal free transfer is also tried; vessels are sutured to internal mammary vessels or to the vessels available the neck. Pedicled jejunum is also *supercharged* with microvascular anastomoses with internal mammary vessels. It is used as a last option only.

Transposition is done through posterior mediastinum (*shortest route*), right pleural space (transpleural), retrosternal or subcutaneous route.

- *Skinner en bloc resection* used to be of practice in olden days in which tumour with oesophagus is resected along with thoracic duct, azygos vein, intercostal vessels crossing the vertebral bodies, pericardium and mediastinal pleura. There is no survival benefit with this procedure.
- *Vagal sparing oesophagectomy* with HSV is tried to preserve pylorus functioning with avoiding pyloroplasty to reduce the chances of postprandial fullness. But it is technically tedious.
- *Lymphadenectomy* in carcinoma oesophagus:
  - *Standard*: Nodes removed are—paratracheal, parabronchial, carinal, paraoesophageal, posterior mediastinal, paracardial, left gastric, along lesser curve of stomach.
  - *Extended/three field/ultraradical*: As like standard above, with bilateral cervical lymphadenectomy, removal of upper mediastinal, coeliac, retroperitoneal, subhepatic nodes.

Surgical approaches

- Ivor-Lewis 2-phase oesophagectomy
- McKeown 3-phase oesophagectomy
- Left abdominothoracic approach for lower oesophagus—*Sweet* approach
- Thoracoscopic—laparoscopic
- Transhiatal blind approach—Orringer's

Postoperative Management

- Fluid and electrolyte management.
- Antibiotics and proper analgesia.
- Respiratory care; ICT care; physiotherapy.
- Prevention of DVT—elevation, exercises, heparin.
- Monitoring for bleeding, sepsis, leak, oxygen saturation.
- TPN only during initial postoperative period and *early jejunostomy feeding* for nutrition.

Palliative Treatment

80% of patients with carcinoma oesophagus, when present first, have fairly advanced tumour and so they are amenable for only palliative treatment. It is to *relieve pain, dysphagia and to prevent aspiration and bleeding*. Palliation therapy is done –

- If patient is not fit for major surgery.
- If there is blood spread.
- If there is adjacent organ spread.
- If there is peritoneal/liver spread.

Different methods are

- ***Palliative external radiotherapy***
  - 3000 Rads. Severe mucositis, stricture and fistula formation are the complications.
- ***Intraluminal RT***
  - Brachytherapy (radiation intraluminally). Loading catheter is placed using endoscope and applicator is fixed to mouth to give 1500 cGy radiation with least systemic effects.

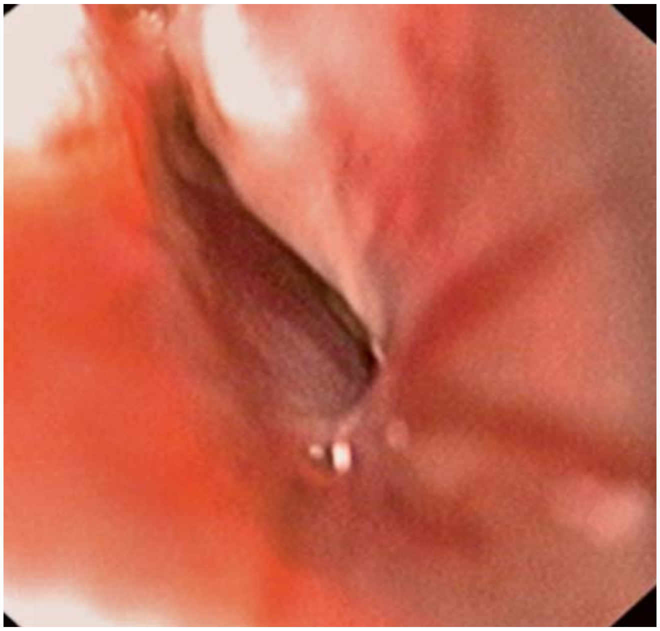

Fig. 19.46: Stricture oesophagus after radiotherapy given for carcinoma oesophagus.

- ***Chemotherapy***
  - Cisplatin; methotrexate; 5 FU; ***palcitaxel***, etoposide, bleomycin. *Platinum based chemotherapy* is beneficial especially in advanced adenocarcinoma of oesophagus.
- ***Intubation***

  Intubation was described by Symmonds in 1887. It is commonly used method. Guidewire is passes across the growth under X-ray screening or C-arm guidance; flexible introducer and prosthetic tube is pushed across the tumour along the guidewire. It carries 90% success rate. Problems are—tube intolerance, poor drainage, airway compression, reflux, aspiration, displacement, food blockage, tumour overgrowth beyond the prosthesis causing its failure.

Intubation is used for *tracheo-oesophageal fistula* or external compression. Prosthesis with a sponge filled balloon is used for fistula closure. Standard tube wrapped with multilayered polyvinyl sponge is other option. It is less expensive, single time, rapid acting.

They can be traction or pulsion tubes. Perforation chance is 10%.

Different tubes used are:

- *Atkinson tube.*
- *Celestin tube* (armoured rubber tube with a long tail)—ideal, commonly used tube. It is wider proximally. It can be passed through endoscopy or laparotomy.
- *Souttar tube (coiled German silver wire)*—block is first dilated with bougies; tube is passed over small sized bougie and pushed ascross the block. It is mainly useful for lower oesophagus.
- *Mousseau-Barbin tube*—cheaper but needs laparotomy. After laparotomy, anterior wall of the stomach is opened. MB tube is passed from mouth with the help of anesthetist by stitching its tip to nasogastric tube and pulled down into the stomach. Tip is cut near the cone part. It is sutured to anterior wall of the stomach. Gastrotomy is closed.

♦ ***Endoscopic therapy:***

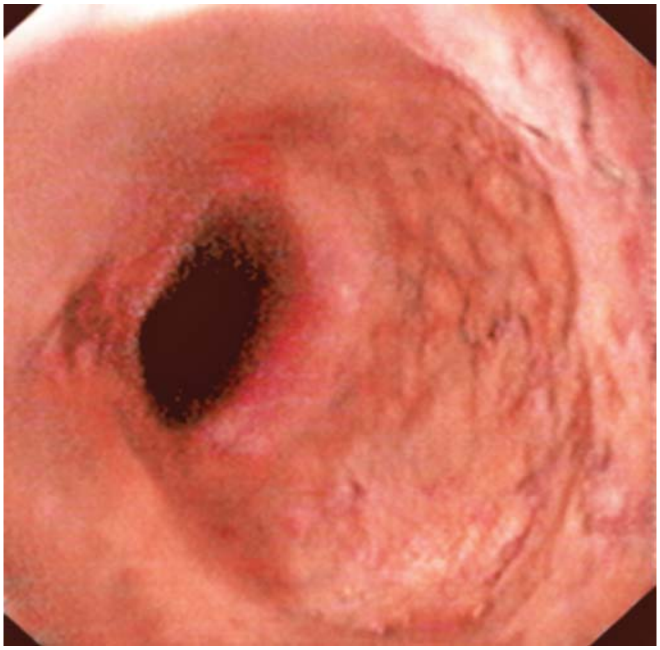

Fig. 19.47: Oesophageal stent *in situ*—a look endoscopically.

- **Self-expanding metal stents (SEMS)** are passed through endoscope under C-arm guidance. It is the *ideal method of palliation*. Stent is collapsed during insertion and released once it is placed in proper position. There is no need to dilate oesophagus more than 8 mm to pass this expanding stent and so chances of perforation is minimal.

  Uncovered SEMS—here tissues project through the mesh to have a better grip with less chances of migration. But stent occlusion is more.

  Plastic covered SEMS—it shows less stent occlusion and less friction. Stent migration is more.

  SEMS may be—stainless wall stent; knitted nitinol Strecker stent; stainless steel covered *Gianturco-Rosch* stent.

  Problems of stents are—aspiration, displacement, erosion, bleeding, tumour growth across or beyond mesh, food bolus obstruction, retrosternal pain, need for reinsertion (40%). Mortality is 1-2%.
- **Endoscopic laser** is used to core a channel through the tumour to improve dysphagia (Nd YAG laser; Diode laser). It causes thermal destruction of tumour. It improves dysphagia but needs repeated laser ablation. It may be used more effectively to remove tumour block in previously placed stents.

  Exophytic tumour less than 6 cm is *suitable* for laser.

  Noncontact high power Nd:YAG 50-100 W laser from distal to proximal end facilitates visualisation of lumen and also reduces the chances of perforation.

  Contact low power Nd:YAG 10-20 W laser is used for fully occluded tumour with less smoke formation and less perforation chance.

  Success rate of palliation is 85%.

  Problems are—fever, chest pain, 3% mortality, perforation (2%) and fistula formation—5%, costly, takes one week to relieve dysphagia
- **Endoscopic bipolar diathermy** using a olive tip; argon beam plasma coagulation; endoscopic alcohol injection into the tumour.
- **Endoscopic photodynamic therapy** (PDT) is used to destruct tumor and to relieve dysphagia. It is often used as a therapy in early cancer. Photosensitive haematoporphyrin agent is injected intravenously 48 hours before endoscopy. It is activated over tumour using laser. Visible infrared light also can be passed to tumour endoscopically to create tumour necrosis by released cytotoxic singlet oxygen through photosensitiser. Sunburn, fever, perforation, pleural effusions are complications. It is effective only to superficial cancers; effects will be seen only after 1 week; need to avoid direct sunlight exposure by the patient for one month is the drawback.

♦ ***Surgery***

- Transhiatal Orringer's blind oesophagectomy is a palliative surgical procedure.
- *Kirschner palliative gastric bypass* done in advanced carcinoma oesophagus wherein mobilised stomach is brought to neck via retrosternal or subcutaneous route and anastomosed to divided cervical oesophagus. Lower cut end of oesophagus is anastomosed to a jejunal loop. Here oesophagus is not addressed (left alone).

Palliative procedures

- External or intraluminal RT (Brachytherapy)
- *Traction tubes* like Celestin or MB tubes through open surgery
- *Pulsion tubes* like self-expandable metal stents through endoscopes using C-arm
- Endoscopic laser
- Chemotherapy
- Transhiatal oesophagectomy—Orringer

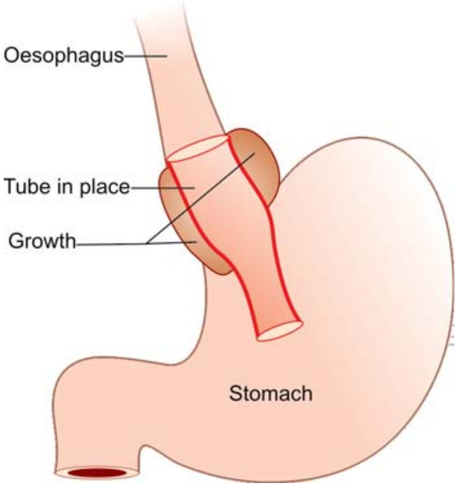

Fig. 19.48: *Mousseau-Barbin* tube placed as a palliative method for carcinoma oesophagus.

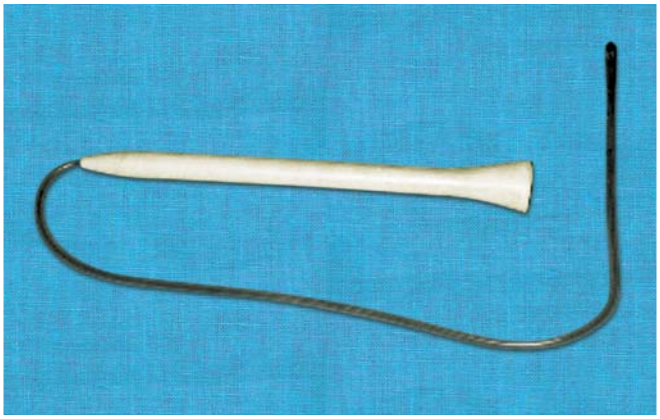

Fig. 19.49: *Mousseau-Barbin* tube.

Complications of oesophagectomy

- 5-10% mortality
- Haemorrhage
- Respiratory infection, often severe
- Septicaemia
- Chylothorax, injury to thoracic duct
- Anastomotic leak—thoracic leak is most dangerous (5-10%)
- Hoarseness due to recurrent laryngeal nerve palsy
- Stricture formation (40%)
- Gastro-oesophageal *reflux* in gastric transfer
- *Conduit necrosis* due to ischaemia either stomach or colon
- *Colonic dysmotility* causing partial obstruction in colon transfer

Terminal Events in Carcinoma Oesophagus

- Cancer cachexia.
- Sepsis, mediastinitis.
- Immunosuppression.
- Malignant tracheo-oesophageal fistula (causes severe respiratory infection and death. Here expansile (self-expandable) *endoluminal stents* are used at the site of fistula to have temporary benefit).
- Erosion into major blood vessel—bleeding.

Prognosis

- Not good because of early spread, longitudinal lymphatics, aggressiveness, difficult approach, late presentation.
- Nodal involvement carries bad prognosis.
- 5-year survival rate is only 10%.

BENIGN TUMOURS OF THE OESOPHAGUS

- *Benign tumours* of the oesophagus are rare (0.5-1% of all oesophageal tumours).
- It grows slowly like a balloon by expansion, compressing surrounding structures. It never infiltrates or spreads.
- It can be solid, cystic, polypoid.
- It is usually in submucosal plane.
- It can cause obstruction/regurgitation/aspiration/ mediastinal compression.
- It can be squamous papilloma/polyp/inflammatory pseudo tumour/*leiomyoma (commonest benign tumour of oesophagus—65%)*/neurofibroma/rhabdomyoma/ lipoma. True adenoma in oesophagus is very rare.
- Features may be asymptomatic (85%—identified incidentally during contrast X-ray/endoscopy); dysphagia/airway obstruction/pneumonia/spluttering during swallowing; stridor/regurgitation.
- ***Leiomyoma*** (commonest—65%) is smooth, sessile, lobulated, firm, with grey-white whorled appearance.
  - Only when leiomyoma reaches 5 cm in size it causes obstruction.
  - Multiple localised leiomyomas can occur which can be enucleated independently.
  - *True diffuse leiomyomas* can occur occasionally in females (4%) in lower oesophagus, as diffuse hyperplasia and thickening of both muscular layers; often as part of the Alport's syndrome which needs total oesophagectomy with gastric pull up, even though benign.
  - Benign leiomyoma of oesophagus rarely turns into leiomyosarcoma.
  - 90% of oesophageal leiomyomas occur in *lower third of the oesophagus.*
  - Leiomyomas are common in men in 5th decade.
  - Leiomyoma which expresses the c-kit oncogene (CD117) is considered as GIST.
- *Investigations*—barium swallow X-ray (smooth circular outline/eccentric filling defect)/oesophagoscopy/endosonography/CT scan.
- *Treatment*—if tumour is more than 5 cm/symptomatic tumour/intraluminal tumour/when diagnosis is doubtful

surgical enucleation is indicated. Enucleation is the therapy of choice. Ideally it should be done through right-sided thoracotomy. Occasionally oesophageal resection is needed if tumour is very large/tumour with mucosal ulceration/if tumour is near OG junction. Thoracoscopic resection can be done. Leak, empyema, sepsis and stricture are the occasional complications.

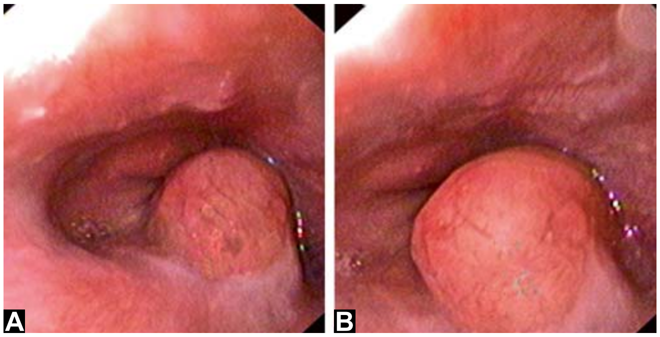

Figs 19.50A and B: Oesophageal submucosal tumour—endoscopic view—leiomyoma.

Note:
Unlike in the stomach and intestine (gastric leiomyoma more than 6 cm/intestinal leiomyoma more than 4 cm are potentially malignant) increased size of the oesophageal leiomyoma does not predispose the malignant transformation.

- ***Oesophageal cysts***
  - It is 2nd commonest benign tumour of oesophagus.
  - It can be congenital or acquired. Congenital is derived from foregut. It contains mucous. It is lined by ciliated columnar epithelium. In infants it is common in upper third of oesophagus; often with a fistula into airway. It can cause obstruction.
  - Acquired cyst is from obstruction of the excretory ducts of oesophageal glands.
  - *Treatment*: Enucleation or resection. If fistula is present it should be ligated and divided.

OESOPHAGEAL PERFORATION

Causes

- Instrumental injuries—commonest cause, 75% commonest site is just above the level of cricopharyngeus.
- Foreign bodies.
- Alkali injuries.
- Carcinoma oesophagus 1%.
- Boerhaave's syndrome 15%.
- Trauma 9%.
- Surgical trauma (Vagotomy, thyroidectomy, Heller's operation, pneumonectomy, spine surgery).

Clinical Features

Chest pain, vomiting, shock, subcutaneous emphysema.

Investigations

- Chest X-ray—shows pneumomediastinum.
- CT scan.

Complications

- Mediastinitis.
- Septicaemia.
- Empyema, ARDS.

Treatment

- *Conservative treatment*: It is advocated in small perforations due to instrument where there is minimal air leak and contamination of mediastinum with less septic load. Crepitus should be absent; pleura should be clear and without any obstruction. Treatment is—antibiotics, nutrition (TPN/enteral through tube), fluid management, proper observation and monitoring the patient by repeated blood counts, and imaging. Biodegradable removable self expanding stents also can be used. It may also act as a bridge therapy for eventual major surgical exploration. Stents which are used for carcinomas cannot be used as they cannot be removed.
- *Thoracotomy,* proper saline wash to pleura, mediastinum and repair with buttressing the area using *pedicled intercostal musculopleural flap is done*. Nasogastric tube for long duration, jejunostomy tube for feeding, ICT for drainage is essential. Often in late cases decortication of lung is needed to achieve full expansion of the lung.
- Repair over T tube so as to create a controlled fistula along with feeding jejunostomy and ICT on both sides.
- Intraluminal stents/mediastinal irrigation and drainage.
- Resection of oesophagus with gastric pull up. As condition is an emergency situation and with sepsis it carries high mortality.
- Oesophageal exclusion with cervical oesophagostomy above and feeding jejunostomy below.
- Diversion surgeries using colon/stomach/jejunum.

FOREIGN BODY OESOPHAGUS

Common Foreign Bodies

- Coins, metals, plastics.
- Dentures.
- Pins, toothpicks, batteries.
- Fish or meat bones—dangerous—40%.
- Food (meat—common/vegetables) impaction—45%.

Sites of Impaction in Oesophagus

- Cervical constriction—C6.
- Broncho-aortic constriction—T4.
- Diaphragmatic constriction—T10.
- Pre-existing malignancy or inflammatory stricture site.

You can become strongest in your weakest place.

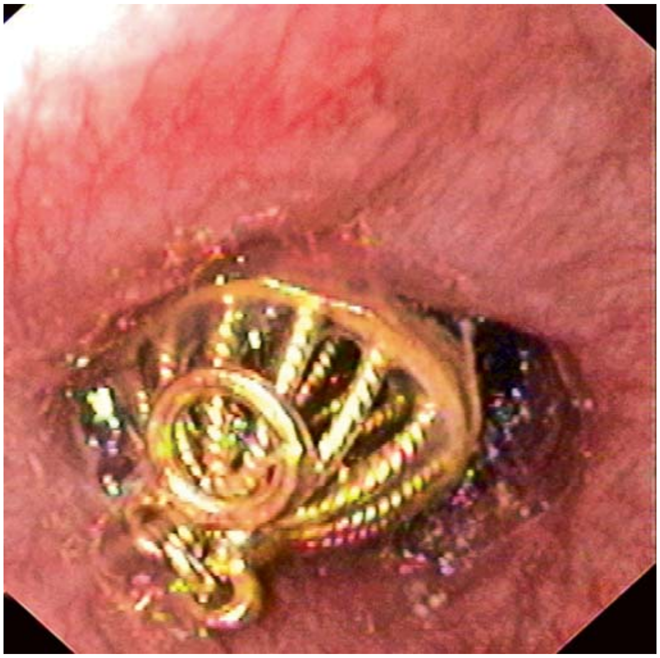

Fig. 19.51: Foreign body in the oesophagus.

Features

- Sudden dysphagia with chest pain and breathlessness.
- Later features of shock, sepsis, mediastinitis, empyema.

Management

- X-ray shows site and level of the F/B.
- Endoscopic removal can be tried.
- Impacted large F/B should be removed by thoracotomy.
- Antibiotics, jejunostomy, TPN, ICT are also required.

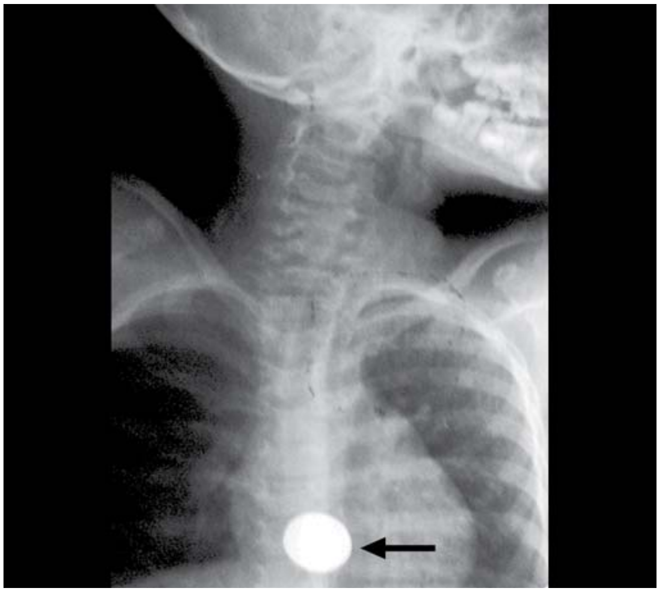

Fig. 19.52: Foreign body (COIN) in the lower oesophagus. Usually it can be removed by endoscope.

Chapter 20 Stomach

CHAPTER OUTLINE

- Anatomy
  - Blood Supply of Stomach
  - Nerve Supply of Stomach
  - Histology
  - Lymphatic Drainage of Stomach
  - Duodenum
- Gastric Physiology
- Gastric Function Tests
- Gastrin
- Barium Meal Study
- Gastroscopy
- Helicobacter Pylori
- Congenital Hypertrophic Pyloric Stenosis
- Gastritis
- Acute Peptic Ulcer
- Gastric Ulcer
- Duodenal Ulcer
- Pyloric Stenosis due to Chronic Duodenal Ulcer
- Perforated Peptic Ulcer
  - Perforated Duodenal Ulcer
  - Perforated Gastric Ulcer
  - Perforated Stomal Ulcer
  - Dry Perforation
- Bleeding Peptic Ulcer
  - Bleeding Duodenal Ulcer
  - Bleeding Gastric Ulcer
- Haematemesis
- Complications of Gastric Surgery
  - Duodenal Blow Out
  - Recurrent Ulcer
  - Dumping Syndrome
  - Roux Stasis Syndrome
  - Gastrojejunocolic Fistula
- Trichobezoar
- Chronic Duodenal Ileus
- Durban's Syndrome
- Acute Gastric Dilatation
- Gastric Volvulus
- Gastric Polyp
- Menetrier's Disease
- Duodenal Diverticula
- Carcinoma Stomach
- Gastric Lymphoma
  - Primary Gastric Lymphoma
  - Secondary Gastric Lymphoma
- Gastric Sarcomas
- Gastrointestinal Stromal Tumours
- Pyloroplasty
- Gastrostomy
- Gastrectomy
- Gastrojejunostomy
- Retrograde Jejunogastric Intussusception
- Vagotomy

ANATOMY

The stomach contains four anatomic regions:

1. Fundus.
2. Cardia.
3. Body.
4. Antrum.

The *duodenum* is 20-30 cm in length. It extends from pyloric sphincter to *ligament of Treitz*. It is divided into four parts.

90% of duodenal ulcer occurs in the 1st part of duodenum (duodenal bulb/cap).

CBD and pancreatic duct merges to form *ampulla of Vater* and enter the 2nd part of duodenum. The 3rd part of duodenum is wedged between aorta and superior mesenteric artery.

Goals and actions must meet.

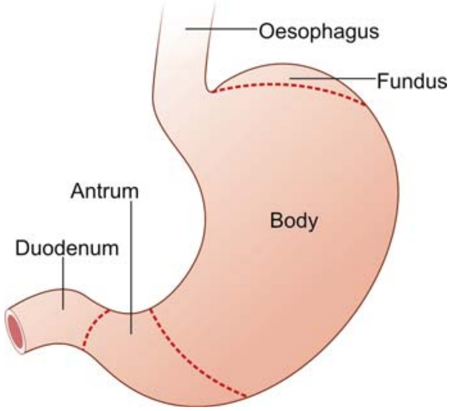

Fig. 20.1: Parts of stomach.

BLOOD SUPPLY OF STOMACH

Arterial Supply

- Left gastric artery, a branch of coeliac artery (Smallest branch of coeliac axis).
- Right gastric artery, a branch of hepatic artery.
- Gastroduodenal artery, a (largest) branch of hepatic artery.
- Right gastroepiploic artery, a branch of gastroduodenal artery.
- Left gastroepiploic artery, a branch of splenic artery.
- Short gastric arteries, branches of splenic artery.

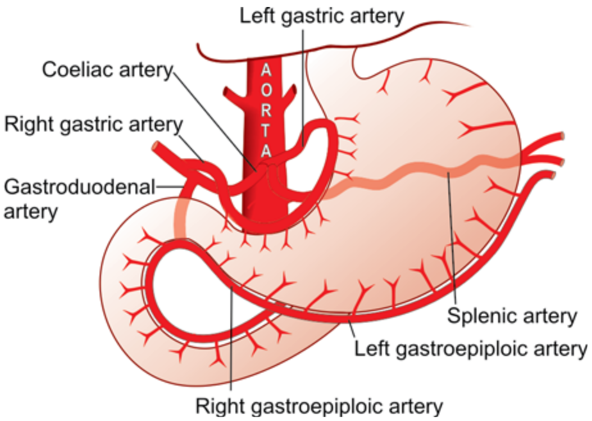

Fig. 20.2: Blood supply of the stomach.

Venous Drainage

- Right and left gastric veins drain into portal vein.
- Right gastroepiploic vein drains into superior mesenteric vein.
- Left gastroepiploic vein and short gastric veins drain into splenic vein.
- Prepyloric vein of *Mayo* distinguishes pyloric canal from the first part of duodenum.

NERVE SUPPLY OF STOMACH

- Intrinsic innervation occurs through myenteric plexus of *Auerbach* and submucous plexus of *Meissner.*
- Right vagus is posterior and left vagus is anterior.
- Posterior vagus gives *criminal nerves of Grassi*, which supply lower oesophagus and fundus of stomach, which, if not cut properly during vagotomy, may lead to recurrent ulcer.
- Vagus also gives splanchnic branches (*hepatic and coeliac branches),* ends as *nerve of Latarjet* which supplies the antrum and maintains the antral pump.
- Parietal branches help in HCl secretion, which is an important concept in vagotomy that is done as a treatment in duodenal ulcer.
- *Truncal vagotomy* with posterior gastrojejunostomy is done for chronic duodenal ulcer with pyloric stenosis.

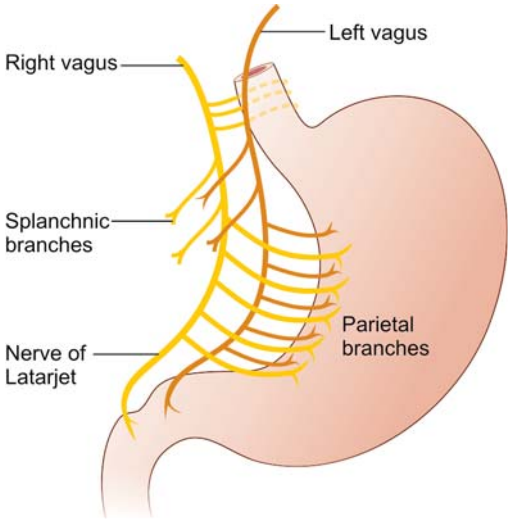

Fig. 20.3: Anatomy of nerve supply of stomach.

- *Highly selective vagotomy* **(***HSV)* is done in case of uncomplicated chronic duodenal ulcer which is not responding to available medical line of treatment. In HSV, nerve of Latarjet is retained so as to retain antral pump and no drainage is required. Here only the fibres entering the stomach are ligated close to the lesser curve to reduce the acid secretion.
- In *selective vagotomy* splanchnic branches are retained but it is presently not done.
- Gastroduodenal pain is sensed via sympathetic fibres (T_5-T_{10}).

HISTOLOGY

- The fundus and body contains parietal and chief cells.
- *Parietal cells* secrete acid and intrinsic factor.
- *Chief cells* produce pepsinogen. There are two types of pepsinogen secreted by chief cells—I and II. Pepsinogen I is produced only in stomach. In gastric atrophy pepsinogen I is decreased.
- In the antrum, *endocrine cells* produce gastrin (G cells) and somatostatin (D cells).
- 12% of epithelial cells of stomach are parietal (oxyntic) cells; 45% chief (zymogenic cells); 40% mucous cells; 3% endocrine cells.
- Pyloric sphincter is a thick circumferential layer of smooth muscle.

- Submucosa is strongest, collagen rich layer of gastric mucosa.
- Gastric mucus is a mucosal barrier containing mucopolysaccharides which maintains the integrity of gastric mucosa.

LYMPHATIC DRAINAGE OF STOMACH

- Lymphatics of proximal half of stomach drain into left gastric, splenic, and superior pancreatic lymph nodes. From antrum, it drains into right gastric, right gastroepiploic, and subpyloric lymph nodes. From pylorus, it drains into right gastric and subpyloric lymph nodes.
- Efferent lymphatics from suprapyloric region drain into para-aortic lymph nodes and so into left supraclavicular lymph nodes. Efferent lymphatics from subpyloric lymph nodes drain into superior mesenteric lymph nodes. Lymphatics near *oesophagogastric* (OG) junction communicate with oesophageal lymphatics.
- In carcinoma stomach if upper lymphatics are blocked, retrograde spread through lower lymphatics can occur.

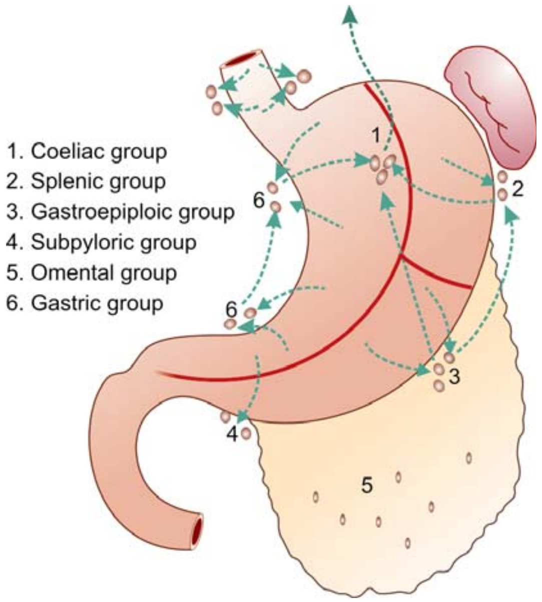

Fig. 20.4: Lymphatic drainage of stomach.

- Presently in carcinoma stomach different resections are classified as R0, R1, R2, R3, or D1, D2 (dissection) based on levels of lymph nodes in the abdomen in relation to the stomach. R0 is *no residual disease*; R1 is microscopic residual disease; R2 is macroscopic residual disease; R3 is inoperable.

DUODENUM

Arterial Supply

It is mainly supplied by superior and inferior pancreaticoduodenal arteries. First part also gets supply from right gastric artery, supraduodenal artery, a branch of hepatic artery.

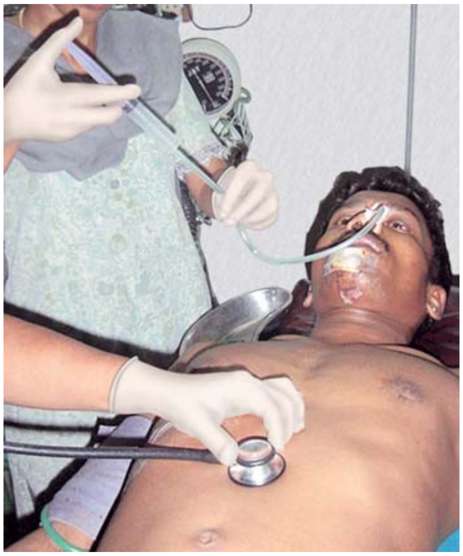

Fig. 20.5: After passing the Ryle's tube to the stomach, air from the syringe should be infused through the tube. When auscultated gush of air into the stomach can be heard which confirms that tube tip is in the stomach.

Venous Drainage

Drains into the splenic, superior mesenteric and portal veins.

Nerve Supply

- Sympathetic from spinal segments T_8 and T_{10}.
- Parasympathetic from vagus.

Lymphatic Drainage

Drains mainly to pancreaticoduodenal nodes present along the inside of the curve of duodenum.

GASTRIC PHYSIOLOGY

- *Gastric function* is regulated by hormonal and neural methods. Hormonal mediators control function through endocrine (through release into blood), paracrine (diffusion across interstitial space) and neurocrine (diffusion across synopsed target cell and receptor binding).
- *Gastric acid secretion* is regulated by acetylcholine, histamine and gastrin. Acetylcholine is the principal mediator of acid release through vagal parasympathetic ganglion cells. Vagus innervates parietal, G and *enterochromaffin-like* (ECL) cells. Basal acid secretion is 10% of maximal acid output (1-5 mmol/hour). It is reduced by 90% after vagotomy or H_2 receptor blockage. ***Phases of acid secretion*** are—(1) *Cephalic phase*—through central meditation (smell/sight/taste → vagus → acetylcholine → muscarinic receptors. (2) *Gastric phase*—food enters the stomach → antral G cells → acid release though gastrin (gastric distension causes direct acid release). Gastric phase lasts until stomach is empty and releases 70% of total acid release. (3) *Intestinal phase*—it is 10% of acid release and is mediated by chyme entering the small bowel—nongastrin related.

- *Receptors of acid secretions* are—(1) Gastrin—CCK receptors: Two types are present in parietal cell; Gastrin CCK type A receptor has high affinity for CCK but less to gastrin; Gastrin CCK type B receptor has got high affinity for both CCK and gastrin. (2) Muscarinic receptor (parietal cell)—its M3 type is mediator for acetylcholine. (3) Histamine receptor (parietal cell)—H_2 subtype binds to histamine to cause effect. (4) Somatostatin receptor—these are 5 subtypes. Its inhibitory action is through parietal cell subtype. (5) Second messengers involved are intracellular cAMP and calcium.
- *Luminal gastric pH* is 2. The pH at surface epithelial cells is 7. These cells secrete bicarbonate continuously into the lining mucus gel to keep the pH of surface mucus at 5. Bicarbonate in mucus barrier reduces once luminal pH reduces below 1.4.
- *Functions of gastric acid*: Conversion of pepsinogen to pepsin which in turn hydrolyses proteins into polypeptides; promotes release of duodenal secretin; prevents bacterial colonisation of upper GIT; formation of food chyme which contains food particles of 1 mm size or less.
- *Gastric juice* contains HCl, mucus, swallowed saliva, reflux content from duodenum. Parietal cells secrete electrolytes of 160 mmol/L. They also secrete *intrinsic factor*, a mucoprotein which is essential for absorption of vitamin B_{12} in ileum; its secretion is independent of acid secretion from parietal cell and is not influenced by PPIs. Pepsinogen is a proteolytic proenzyme (42,000 mol wt). Type I is secreted from chief cells and mucus neck cells of only acid secreting part of the stomach (acid secreting mucosa). Type II is secreted from surface epithelial cells of entire stomach and proximal duodenum. Mucus is secreted by surface mucus cells and mucus neck cells from acid secreting area of stomach and antrum. Gel like viscous mucus contains 85% water and 15% glycoproteins. It contains bicarbonate secreted from surface epithelial cells. Mucus is strong gastric barrier. Mucus secretion is inhibited by anticholinergics and NSAIDs. *H. pylori* break the mucin and so barrier.
- *Gastric motility* begins from *pacemaker cell of Cajal* located at proximal stomach. Special myoelectric migrating complex slow waves with electric spikes maintain gastric motility in three phases. Immediately after food intake resting tone of fundus and proximal stomach decrease causing receptive relaxation and gastric accommodation mediated by vagus. Vagotomy eliminates this causing early fullness, early satiety and rapid gastric emptying. Delayed gastric motility occurs in diabetes, *H. pylori* infection and after vagotomy. Many factors like stress, hormones increase the gastric motility. Prokinetics, erythromycin helps in gastroparesis.
- *Mucosal defense factors* are—mucus production, mucosal HCO_3, mucosal blood flow, growth factors, cell renewal, endogenous prostaglandins, epithelial barrier (hydrophobic phospholipids, restitution, NO, epidermal growth factor and microcirculation). *Aggressive factors* are—*H. pylori*, HCl secretion, pepsin, smoking, alcohol, duodenal bile reflux, ischaemia, NSAIDs.

GASTRIC FUNCTION TESTS

Gastric function tests

- Pentagastrin test/Kay's augmented histamine test
- Hollander's insulin test
- Radioisotope labelled gastric emptying study
- 24 hours intragastric pH monitoring
- Gastrin level estimation

Peak Acid Output Test (Pentagastrin/Augmented Histamine Test)

Initially stomach is emptied completely. Basal acid level of aspirated stomach content (aspiration is done for 1 hour on empty stomach) is analysed. Pentagastrin—6 µgm/kg or histamine 2 µgm/kg is injected IV/SC/IM and 15 minutes samples for subsequent hours (usually 1 hour) are collected and analysed. Peak acid output level is calculated. Test is useful in Zollinger-Ellison (Z-E) syndrome, duodenal ulcer and gastric ulcer. Basal acid output (BAO) is 2-3 mEq/hour. Peak acid output (PAO) is highest rate of secretion obtained in any of the 15 minutes samples following stimulation. Maximal acid output (MAO) is obtained by averaging the output of two final 15 minutes samples. In duodenal ulcer and Z-E syndrome basal and stimulated acid is increased; in pernicious anaemia, gastric atrophy and gastric cancer both are decreased.

	Basal	*Peak*
Normal	5 mmols	25-27
Gastric ulcer	2	15
Duodenal ulcer	6	35-38
Z-E syndrome	Very high	> than 60

Achlorhydria is defined as a condition in which stomach cannot produce secretions with a pH of less than 7.0 even after maximal stimulation. It is seen in 20% of carcinoma stomach.

Note:
Achylia means—no pepsin in the gastric juice.

Insulin Test (*Hollander*)

- This test is useful postoperatively to confirm the completeness of vagotomy. 0.2 units/kg body weight of insulin is injected intravenously to a fasting patient so as to create hypoglycaemia of below 35 mg%, which in turn stimulates the parietal cells through hypothalamus and through vagus to cause increased acid secretion. Patient who had undergone complete vagotomy does not show increase in acid secretion.
- *Early response* signifies incomplete vagotomy wherein there is increase in acid concentration in first hour of > 20 mmols/L.
- *Delayed response* signifies an increase in acid concentration in between first and second hours, and is probably due to vagal gastrin release.

GASTRIN

- Gastrin is secreted by the G-cells from gastric antrum.
- *Types of gastrin* are—*Big gastrin* (G_{34}); *little gastrin* (G_{17})—most common 90%; *mini gastrin* (G_{14}). Pentapeptide at the C terminal end of G_{17} is active part which is identical to CCK. Luminal peptides and amino acids are most potent stimulator of gastrin release. Luminal acid is the most potent inhibitor of gastrin release.
- *Gastrin promotes* the release of acid and also regulates it. It also maintains mucosal defense, has trophic effects on parietal and ECL cells (enterochromaffin like cells).
- Number of G-cells are increased in duodenal ulcer, G-cell hyperplasia, but not in gastric ulcer. Normal plasma value is 50 ng/L of plasma (fasting). It is analysed by radioimmunoassay.
- It increases to many 1000's in gastrinomas.
- *Hypergastrinaemia* can occur in *ulcerogenic conditions* like—antral G cell hyperplasia, retained excluded antrum, Z-E syndrome, short gut syndrome, gastric outlet obstruction; *nonulcerogenic conditions* are—PPIs, pernicious anaemia (–ve feedback), vagotomy, atrophic gastritis (hypochlorhydria), *H. pylori*, chronic renal failure.
- Gastrin is raised very high in *gastrinomas*.

Gastric hormones

- *Gastrin*
- *Somatostatin*—2 types, 14 and 28. In stomach, type 14 is common; from D cells of gastric mucosa of fundus and antrum. Antral acidification is the main stimulus; vagal acetylcholine inhibits its release. It has *paracrine* effect inhibiting acid secretion from parietal cell, gastrin release from G cell and histamine release from ECL cell. *H. pylori* decrease antral D cells and so somatostatin, leading into increased gastrin release
- *Gastrin releasing polypeptide* (GRP, Bombesin)—it stimulates gastrin and somatostatin release by binding to receptors on G and D cells. It is released from sympathetic nerve terminals of the gastric mucosa of body and antrum. Its half life is 1½ minutes
- *Histamine* is stored in ECL and mast cells. Its release is stimulated by gastrin, acetylcholine and epinephrine
- *Gherlin* (28 amino acids) is mainly secreted from endocrine cells of gastric mucosa. It has mainly got growth hormone releasing action; prolactin, ACTH, cortisol, aldosterone release is also promoted. Its action on islet cells reduces insulin release. *Gherlin* stimulates appetite. Gastric bypass, gastrectomy reduces the appetite.

Investigations for gastroduodenal diseases

- Gastroduodenoscopy is ideal and most common investigation used to visualise mucosa. Lesions if any biopsy is taken for *H. pylori* and carcinoma. Also often used for endotherapy
- Endosonography (EUS) is very sensitive method to assess tumours, visualise stomach layers (90% accuracy), lymph nodes (80%), and to detect early liver metastasis which may not be identified by CT especially from left lobe
- CT scan is good imaging method to detect the stage, spread, nodal status, liver secondaries, and status of lungs
- CT and PET scan (together) using FDG is better to assess early spread
- Laparoscopy is very good investigating tool to identify peritoneal secondaries, and to stage the disease. Laparoscopic US is very sensitive to detect the liver secondaries
- Barium meal studies are used to detect hiatus hernia
- Celiac angiography is useful in bleeding ulcers, both for diagnostic and therapeutic purpose (therapeutic embolisation)

BARIUM MEAL STUDY

Indications

1. Gastric ulcer—shows a *niche* which is the ulcer crater, a *notch* which is due to spasm of circular muscle on the greater curvature.
2. Chronic duodenal ulcer—shows absence or deformed duodenal cap (due to spasm of 1st part of duodenum, barium will not stay and so cap will not be formed).
3. Gastric outlet obstruction—the cause may be chronic duodenal ulcer with pyloric stenosis or carcinoma pylorus. Features are:
   - Enormous dilatation of stomach.
   - Greater curvature below the level of iliac crest.
   - Absence of duodenal cap.
   - No filling of dye in 2nd part of duodenum.
   - Mottled appearance of stomach because of retained food particles.
   - Evidence of gastritis.
4. Carcinoma stomach—irregular filling defect.

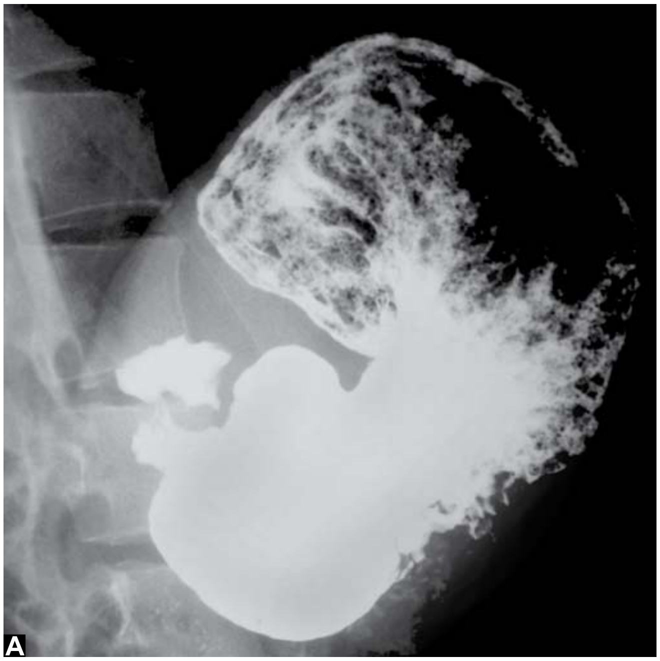

Fig. 20.6A

The stomach you can hear, the stomach you can see, the stomach you can feel.—**Sir James Walton**

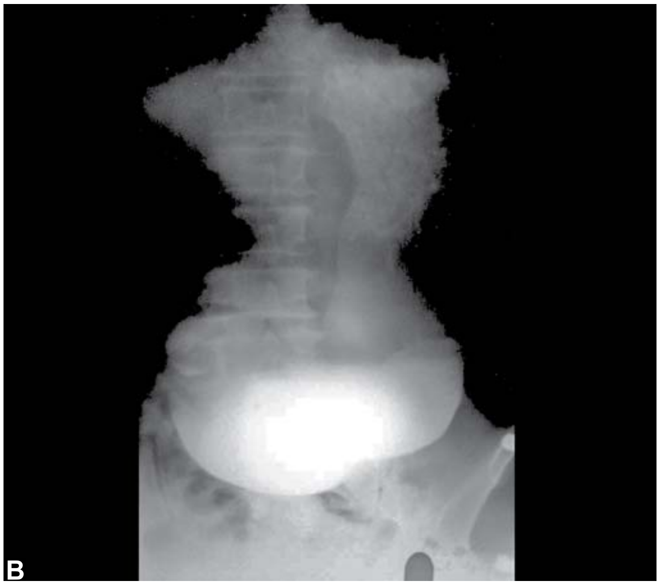

Fig. 20.6B

Figs 20.6A and B: Barium meal X-rays showing normal picture in one and dilated stomach with features of gastric outlet obstruction in another.

5. Pseudocyst of pancreas—widened vertebrogastric angle.
6. Stomal ulcer in previous gastrojejunostomy.
7. In chronic duodenal ileus (*Wilkie's syndrome)*—shows dilatation of stomach, 1st and 2nd part of duodenum, proximal portion of 3rd part of duodenum.
8. Others—gastric volvulus, duodenal diverticula, trichobezoar, gastric fistulas, diaphragmatic hernias when stomach is the content.
9. Carcinoma of head of pancreas shows "*pad sign"*; in periampullary carcinoma—*reverse "3'* sign.
   Frostberg's reverse '3' sign (inverted '3' sign).
   Rose thorn duodenum in carcinoma head of pancreas.
10. Hiatus hernia.

Procedure

- Barium sulphate solution is used [Barium is neurotoxic, but in sulphate media it will not get absorbed and so barium sulphate is used (Barium phosphate is not used)].
- About 300 ml solution is given to the patient to drink and its flow down to the stomach is observed under fluoroscopic guidance. Films are taken as required. Commonly oblique views are taken.
- Microcrystalised barium sulphate (*Microbar solution*) is better as it does not get precipitated.

Complication

It may precipitate intestinal obstruction.

GASTROSCOPY

It is visualisation of interior of stomach, duodenum, oesophagus. It can be done as OP procedure.

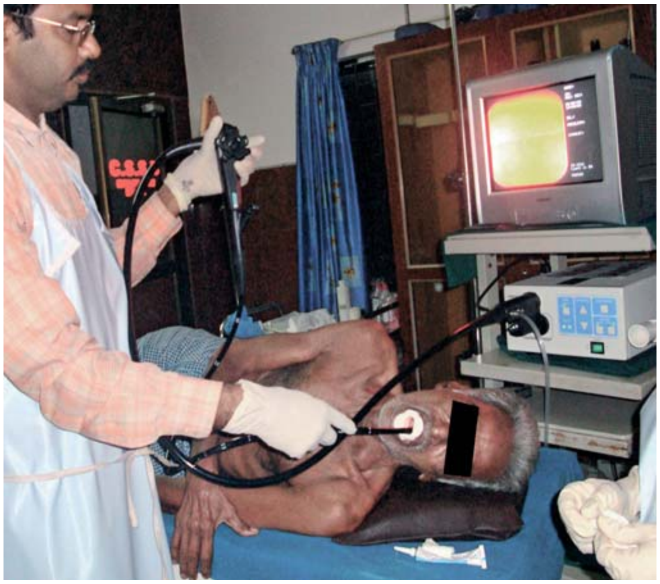

Fig. 20.7: Technique of doing gastroscopy using videoendoscope.

Uses

1. *For diagnosing any pathology,* e.g.
   - Gastric ulcer.
   - Duodenal ulcer.
   - Gastritis.
   - Stomal ulcer.
   - Carcinoma stomach.
   - Oesophagitis.
   - Varices.
   - Biopsies from the suspected cases of malignancy or for *Helicobacter pylori* can be taken.
     - ***Endosonography*** can be done to assess the staging, operability of carcinoma stomach or oesophagus.
     - Presently fibreoptic, flexible, or video gastroduodenoscopy is used.
2. *Therapeutic*
   - ***Videoendoscopy*** is used not only for diagnosis but also mainly for therapeutic procedures.
   - Both end viewing and side viewing gastroscopes are available.
   - For therapeutic procedures and ERCP, side viewing gastroscope is required.

Therapeutic procedures done are:

- Variceal injection or ligation or glueing or banding
- Stenting of pseudocyst of pancreas through gastroscopy
- Polyp removal
- Submucosal resection
- For ERCP diagnostic and therapeutic procedures
- Percutaneous gastrostomy (PEG)

Procedure

Gastroscopy is done following eight hours of fasting. After lignocaine spray into the oral cavity, gastroscope is passed gently down the oesophagus when the patient does the swallowing action. Once the scope is inside the stomach, air is

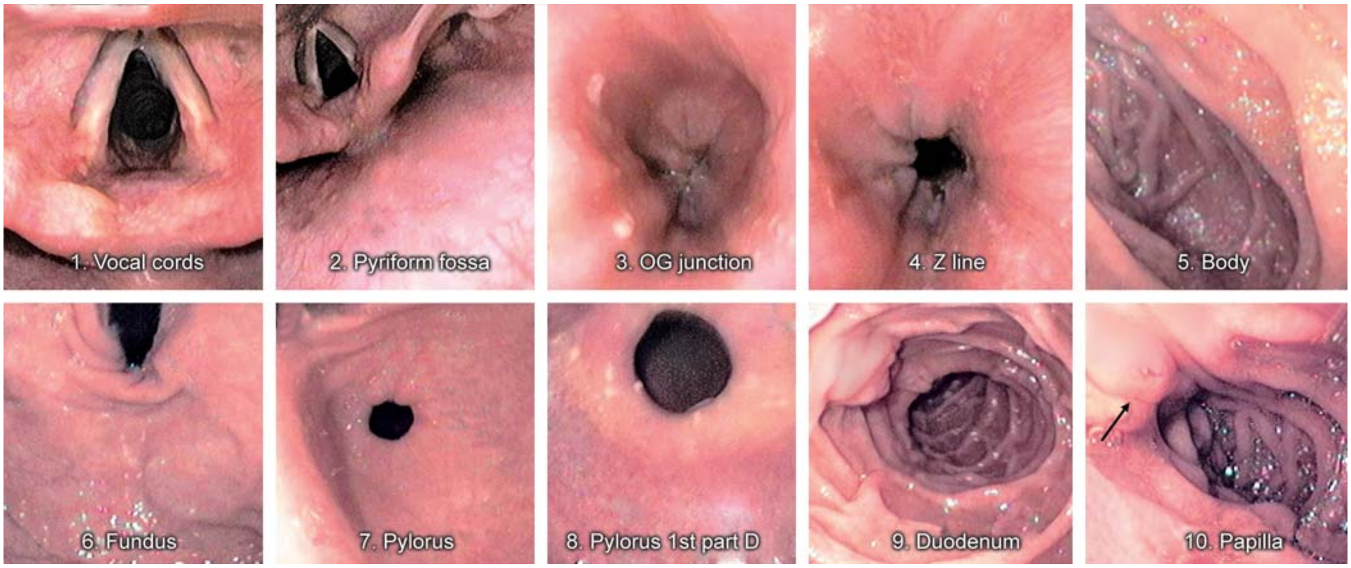

Fig. 20.8: Gastroscopic views of different parts—vocal cords; pyriform fossa; oesophagus; oesophagogastric (OG) junction; body, fundus and pylorus of stomach, duodenum with clearly visible duodenal papilla (*Courtesy:* Dr Tantry and Dr Sandeep Gopal, Gastroenterologists, KMC, Mangalore).

inflated and different parts of the stomach is visualised. Fundus is visualised by retropulsion. Scope is passed through the pylorus to see the 1st and 2nd parts of duodenum and looked for any pathology. If required biopsy is taken.

Often sedation with midazolam is beneficial to have an easy passage.

- *Buscopan is used to relax stomach wall.*
- *Z-line signifies squamocolumnar junction.*
- *Multi-byte biopsy forceps is used.*

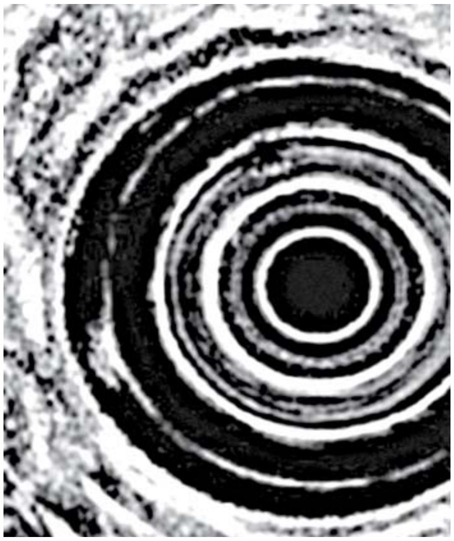

Fig. 20.9: Endosonography of stomach showing different layers with echogenicity.

Complications

- Bleeding.
- Aspiration.
- Perforation (rarely). Perforation occurs mainly in therapeutic procedures like oesophageal dilatation, endoscopic mucosal resection (EMR). It is common in proximal and middle oesophagus but can occur anywhere. Patient presents with hypotension, tachycardia, sudden severe pain in chest, abdomen, neck and surgical emphysema.

HELICOBACTER PYLORI

- It is gram –ve spiral like flagellated organism, first studied by Warren and Marshall (both got Nobel prize), which is commonly present in stomach. It is involved in pathogenesis of duodenal ulcer, gastric ulcer, gastritis (type B) and gastric cancer.

H. pylori

- Duodenal ulcer—95%
- Gastric ulcer—70%
- Gastritis—70-90%
- Gastric cancer—No. 1 carcinogen
- Gastric MALTOMA (mucosa associated lymphoid tissue lymphoma)

Pathogenesis

- It is the most common bacterial infection in the world. Rhesus monkey is the only natural reservoir. Its incidence increases with age.
- It releases enzymes like *urease* that hydrolyses urea so as to release ammonia which through negative feedback mechanism increases the gastrin release from G-cells.
- Infection occurs in stomach, i.e. body, fundus, antrum, disrupts the mucosal barrier, causes chronic inflammatory response leading to gastritis, gastric ulcer.
- Other than urease it also secretes dehydrogenase (converts alcohol to aldehyde which is toxic to mucosa), endopeptidase (disrupts mucosal barrier). Urease creates alkaline environment around it in mucus layer of gastric epithelium.

Control of *H. pylori* Infection

It is treated with antibiotics and other drugs with different combinations

Clarithromycin 500 mg BD	Metronidazole 400 mg BD	Omeprazole 20 mg BD
OR	OR	OR
Amoxycillin 500-750 mg BD	Tinidazole 600 mg BD	Lansoprazole 30 mg BD
OR		OR
Tetracyclines, or Bismuth.		Pantoprazole 40 mg BD

The above regime is given for 7-14 days and then only proton pump inhibitors are continued.

It can survive only in gastric epithelium or gastric metaplasia in duodenum or Barrett's oesophagus or in heterotopic gastric mucosa in Meckel's diverticulum or rectum. It is because receptors for organisms to adhere into mucosa are available only in gastric mucosa.

- *H. pylori* impair mucosal healing, cause degranulation of eosinophils. It releases various protease and lipases that break mucus and so strong protective mucus barrier.
- It also secretes ***cytotoxins*** (cagA and vacA) which may also be involved in inflammatory reaction or malignancy.
- Even though normal duodenum cannot harbour the *Helicobacter*, duodenum with gastric metaplasia can very well get infected by *Helicobacter* which explains why *Helicobacter* is involved in duodenal ulcer.
- *Helicobacter* is normally not found in saliva.
- Infection is transmitted through faeco-oral route, with a infection rate of 80-90% in a population (common and high).
- It is more common in lower socioeconomic group.
- It is considered to be a ***carcinogen for stomach***. It is not linked with carcinoma of OG junction.

Enzymes and toxins released by *H. pylori*

- *Urease*
- Dehydrogenase
- Endopeptidase
- *Vacuolising cytotoxin*
- Haemolysin
- *cagA and vacA*

It is identified by:

a. *Rapid urease test* (cod liver oil test/CLO test)
b. C13-C14 breath tests (labelled urea breath test)
c. Biopsy from the different parts of the stomach and staining to identify the organism.

For *H. pylori* three biopsies are taken one each from antrum, body and fundus.

Tests for *H. pylori*

- Rapid urease test—90% sensitivity, 98% specificity
- C13/ C14 breath tests—95% sensitivity and specificity—'gold standard'
  - C13 requires spectrometry and costly
  - C14 uses radioactivity
- Serology to identify IgG antibody—ELISA test with 90% sensitivity and specificity
- Biopsy and culture—very costly
- *Warthin's starry silver stain*, acridine orange are special stains used
- *Newer methods* – special fluorescent technique, PCR products of *H. pylori*—urease gene, 165 rRNA identified using specialised probes when organisms are in less number

CONGENITAL HYPERTROPHIC PYLORIC STENOSIS

It is hypertrophy of musculature of pyloric antrum, especially the circular muscle fibres, causing primary failure of pylorus to relax. Duodenum is normal.

Clinical Features

- Common in first born *males* (4:1).
- Incidence is 4 in 1000 births.
- It is familial.
- It is seen between 3rd and 6th weeks of age of an infant, the time taken by the hypertrophied muscle to cause complete obstruction.
- Vomiting—forcible, projectile and non-bilious.
- Visible gastric peristalsis (VGP).

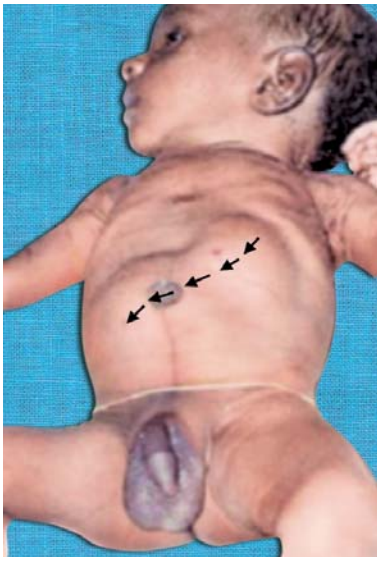

Fig. 20.10: Visible gastric peristalsis.

- Palpable lump of hypertrophied pylorus which is better felt from left side, as a mobile, smooth, firm mass, with all borders well made out, moves with respiration, with impaired resonance on percussion. It is the most important clinical feature (95%).
- Constipation.
- Dehydration and loss of weight.
- Electrolyte imbalance—hypokalaemic metabolic alkalosis.

Clinical features of congenital pyloric stenosis

- Vomiting
- VGP
- Palpable mass
- Constipation and dehydration

In premature infants:

- VGP and mass is better seen and felt.
- Vomiting is regurgitant.
- Anorexia is common.

Diagnosis is established by:

- Clinical examination.
- U/S abdomen (very useful)—*Doughnut sign*.
  - Pyloric muscle 4 mm or more in thickness.
  - Length of pyloric canal > 1.8 cm.
- Barium meal shows obstruction.

Differential Diagnosis

- Duodenal atresia (Bilious vomiting is present).
- High intestinal obstruction (e.g. volvulus neonatorum).
- Intracranial haemorrhage.

Treatment

1. Correction of dehydration and electrolyte imbalance.
2. ***Surgery***: *Ramstedt's operation*—After laparotomy, hypertrophied muscle is cut along the whole length adequately until the mucosa bulges out. Mucosa should not be opened (pyloromyotomy). If mucosa is injured, it should be sutured horizontally using interrupted vicryl or silk sutures.

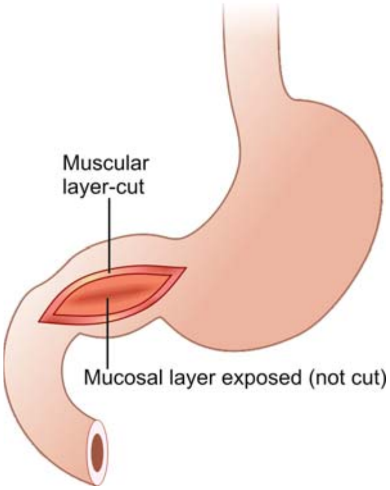

Fig. 20.11: Ramstedt operation for congenital hypertrophic pyloric stenosis. Note that here muscular layer is cut but not mucosal layer.

Complications of surgery

- Postoperative pyrexia (Hyperthermia).
- Gastroenteritis
- Electrolyte imbalance

3. ***Medical treatment***: Not advisable as cure is not guaranteed. Atropine methyl nitrate orally is tried to relax the pylorus muscle.

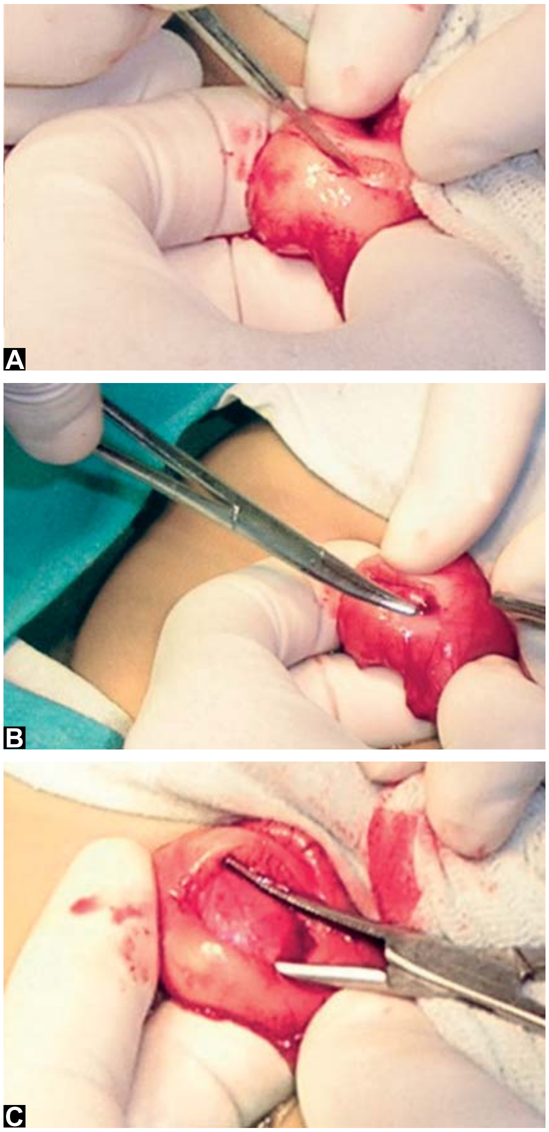

Figs 20.12A to C: *Ramstedt pyloromyotomy* for congenital hypertrophic pyloric stenosis. Only muscular layer is cut to allow mucosa to bulge out.

GASTRITIS

Types

1. Type A gastritis.
2. Type B gastritis.
3. Reflux gastritis.
4. Erosive gastritis.
5. *Others*: Stress gastritis, lymphocytic gastritis, granulomatous gastritis, phlegmonous gastritis.

Troubles waste the stomach like rust waste iron.—**Croatian proverb**

Type A Gastritis

- Autoimmune disease.
- There is formation of antiparietal cell antibodies.
- Parietal cell dysfunction occurs causing achlorhydria and vitamin B_{12} deficiency.
- Antrum is not affected.
- 'G' cell hyperplasia occurs with raised serum gastrin level.
- There is formation of microadenoma of enterochromaffin like cells (ECL cells) with predisposition to gastric carcinoma.

Type B Gastritis

- Occurs due to *Helicobacter pylori* infection.
- Antrum is affected.
- Peptic ulcer is common.
- *Helicobacter* related pangastritis commonly occurs which may turn into gastric cancer.

Reflux Gastritis

- Usually occurs after gastric surgeries.
- Prokinetic drugs are useful—metochlopramide, domperidone, cisapride, mozapride.

Erosive Gastritis

- *Occurs due to disturbed gastric mucosal barrier.*
- Induced by NSAIDs/alcohol.
- Due to inhibition of cyclo-oxygenase type 1 (COX-1) receptor enzyme, resulting in decreased prostaglandin production (Prostaglandin is cytoprotective).
- COX-2 mediated NSAIDs will not cause erosive gastritis.

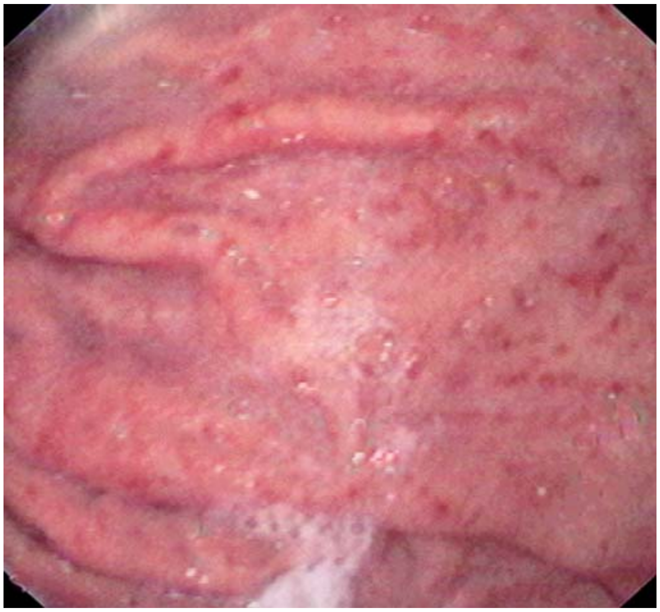

Fig. 20.13: Endoscopic view of erosive gastritis.

Stress Gastritis

Treatment:

- IV ranitidine 50 mg 8th hourly.
- IV omeprazole/pantoprazole.
- Sucralfate orally.

Lymphocytic Gastritis

It is associated with *H. pylori* infection.

Granulomatous Gastritis

It is seen in Crohn's disease and tuberculosis.

Phlegmonous Gastritis

It is due to severe bacterial infection of stomach. It is rare but dangerous.

Nonulcer dyspepsia

- Symptom complex with pain and discomfort in the upper abdomen
- It is intermittent upper abdomen pain in the absence of peptic ulceration
- It occurs in 25% of population—large number
- Anatomical or biochemical abnormalities are not discovered in this condition
- *H. pylori* is not associated with this condition
- Often it lasts for long time decreasing the quality of life
- Differential diagnosis – GERD/acid peptic diseases/gallstones/pancreatitis/carcinoma
- *H. pylori* eradication is not required and there is no surgical role

ACUTE PEPTIC ULCER (DUODENAL OR GASTRIC ULCER)

They are usually multiple erosions due to disruption of the mucosal barrier.

Causes

Stress, drugs like analgesics, steroids, surgeries.

Clinical Features

- Sudden onset of acute pain and tenderness in epigastric region.
- Vomiting with or without haematemesis.
- Often acute peptic ulcers can lead to perforations.
  - Acute ulcers after cerebral trauma or neurosurgeries are called *as Cushing's ulcers*.
  - Acute ulcers after major burns are called *as Curling's ulcers*.

Diagnosis is by gastroscopy.

Treatment

- Intravenous ranitidine 50 mg, 8th hourly.
- IV fluids. IV pantoprazole/rabeprazole/omeprazole.
- Blood transfusions if there is bleeding.

Most of the time surgery is not required for acute ulcers. During follow-up patients are advised to take antiulcer drugs for 4-6 weeks—ranitidine, omeprazole or lansoprazole.

Curling's Ulcers

They are acute ulcers which develop after major burns, presenting as pain in epigastric region, vomiting or haematemesis. Treatment is conservative—IV ranitidine. IV pantoprazole 80 mg in 100 ml DNS—slow, later 40 mg IV maintenance.

Note:
Curling's ulcer occurs when burn injury is more than 35%. It is observed in the body and fundus not in antrum and duodenum.

Cushing's Ulcers

They are acute ulcers which develop after cerebral trauma or after neurosurgical operations. It is commonly single, deeper ulcer more frequently perforates. It can occur in oesophagus and duodenum also. Treatment is conservative by IV ranitidine.

GASTRIC ULCER

Aetiology

It occurs due to imbalance between protective and damaging factors of gastric mucosa.

- Atrophic gastritis, duodenogastric bile reflux, gastric stasis, abnormalities in acid and pepsin secretion. Acid becomes ulcerogenic even to normal gastric mucosa.
- Smoking, alcohol, NSAIDs, steroids.
- *Helicobacter pylori* infection (70%).
- There is either normochlorhydria or hypochlorhydria.
- Altered mucosal barrier mechanism.
- Lower socioeconomic group.

Factors Involved in Gastric Ulcer Formation

- Duodenogastric reflux—reflux containing bile salts and lysolecithin break the mucosal barrier making it more vulnerable for injury, action of drugs and pepsin injury.
- Gastric stasis.
- Ischaemia of the gastric mucosa.
- Type II and III gastric ulcers show acid hypersecretion.

Pathology

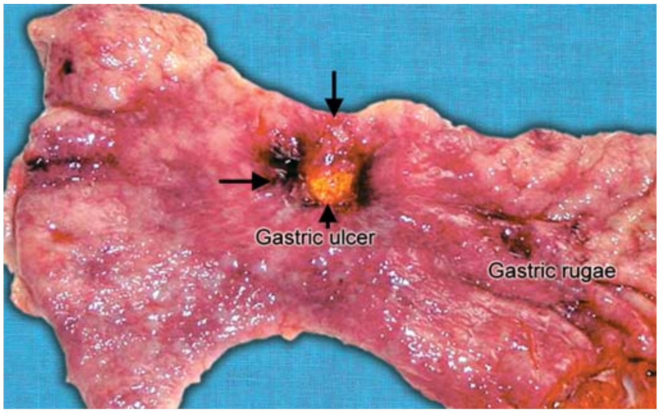

Fig. 20.14: Specimen of stomach (identified by the mucosal pattern and rugae) showing deep ulcer near lesser curvature. Margin of the ulcer is clear, not everted with gastric mucosal folds converging towards the base of the ulcer. 95% of benign gastric ulcer occurs towards lesser curve. Benign gastric ulcer is more common in lesser curvature, as it takes more burden of passage of food and so more of wear and tear. Benign gastric ulcer is rare in greater curvature, fundus and cardia. Histologically it shows destruction of epithelial lining; proliferation of margin; destruction of the part of the muscle layer; granulation tissue in the floor; infiltration with chronic inflammatory cells; endarteritis and fibrosis in the base.

- Gastric ulcer is large in size, usually lies in the lesser curvature, its floor being formed by the muscular layer.
- Posteriorly it may penetrate into the pancreas; it may cause torrential bleeding by eroding left gastric (commonly) vessles or splenic vessels or vessels in the gastric ulcer wall.
- Anteriorly it may perforate or penetrate into the liver. It may lead into *hour glass contracture, or tea-pot deformity*.
- Microscopically, it shows ulcer crater with chronic inflammatory cells and granulation tissue, endarteritis obliterans and epithelial proliferation.
  (Ulcer to the right of the incisura is malignant unless proved otherwise).
- Gastric ulcer > 3 cm is called as giant gastric ulcer. It has got 6-23% chances to turn into malignancy.
- Grossly, margin of the benign gastric ulcer is clear; deep; near lesser curve; edge is not everted with gastric mucosal folds *converging* towards the base of the ulcer.
- 95% of benign gastric ulcer occurs towards lesser curve, as it takes more burden of passage of food and so more of wear and tear. Benign gastric ulcer is rare in greater curvature, fundus and cardia.
- **Acute ulcer**: It is confined to mucosa and submucosa. It is commonly due to NSAIDs.
- **Chronic ulcer:** It penetrates muscularis layer of stomach.

Clinical Features

- Equal in both sexes. It is becoming more common in females.
- Common after the age of 40 years.
- *Pain* in epigastric region after taking food, lasting up to two hours. Pain is uncommon during night. It is relieved by vomiting or by inducing vomiting.

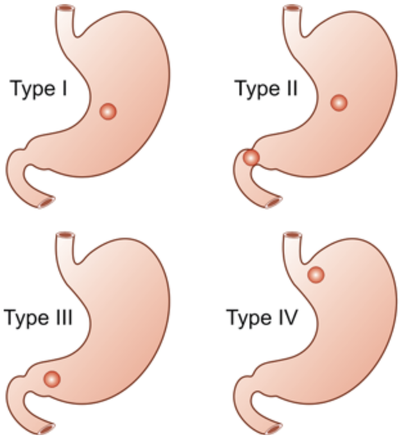

Fig. 20.15: Types of gastric ulcer (*Daintree Johnson*).

- *Periodicity*: Symptom free interval may be 2-6 months. Often with seasonal variation.
- *Vomiting* relieves pain and often it is induced by the patient for relief of pain.
- *Haematemesis and melaena*: Haematemesis is more common.
- Appetite is good but hesitant to eat, because eating induces pain and that results in loss of weight. But once complica-

It is consequences which are more important than rewards or punishments.

Types of gastric ulcer (Daintree Johnson)

Type	*Location*	*Incidence*	*Acid level*
Type I	In the antrum, near the lesser curve	55%	Normal
Type II	Combined gastric ulcer (in the body) with duodenal ulcer	25%	High
Type III	Prepyloric ulcer	15%	High
Type IV	Gastric ulcer in the proximal stomach or cardia	5%	Normal

tions occur, appetite decreases. Aversion to spicy, fried foods occurs.

- On deep palpation, tenderness is felt in epigastric region.

Note:
Often in lesser curve, saddle shaped ulcer can occur.

Differential Diagnosis

- Hiatus hernia.
- Cholecystitis.
- Chronic pancreatitis.
- Chronic gastritis.
- Dyspepsia.
- Carcinoma stomach.

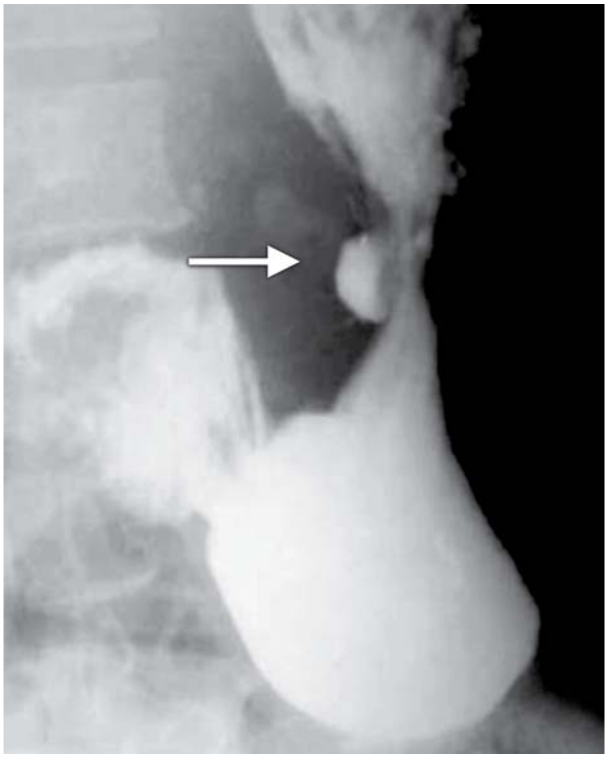

Fig. 20.16: Barium meal showing *Niche* in the lesser curve as benign gastric ulcer.

Investigations

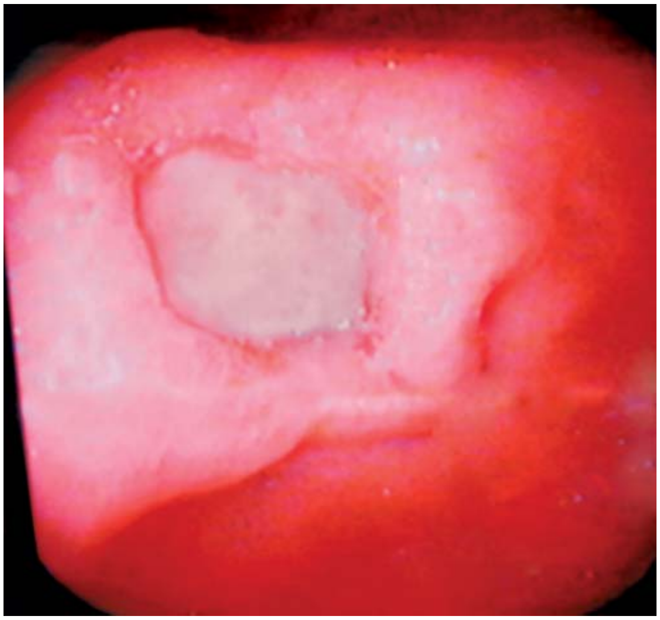

Fig. 20.17: Benign gastric ulcer endoscopic view. Biopsy is a must. Ideally 10 biopsies should be taken from the edge.

- Barium meal X-ray to see *niche and notch.*

Barium meal X-ray features of benign gastric ulcer

- *Niche* on the lesser curve with *notch* on the greater curvature
- Ulcer crater projects beyond the lumen of the ulcer
- Regular/round margin of the ulcer crater—*stomach spoke wheel pattern*
- Overhanging mucosa at the margins of a benign gastric ulcer—projects inwards towards the ulcer—*Hamptom's line*
- Converging mucosal folds towards the base of the ulcer
- Symmetrical normal gastric mucosal folds

- Gastroscopy is done to see the location, type of ulcer and also to take biopsy (10 biopsies).
- US abdomen mainly to rule out other diseases and to confirm associated diseases.

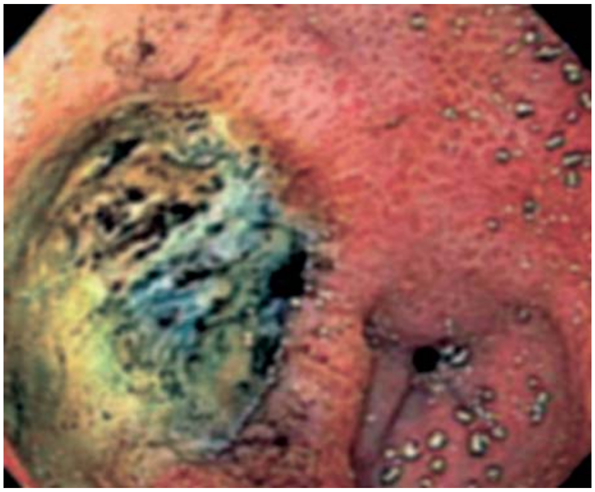

Fig. 20.18: Gastric ulcer in prepyloric region.

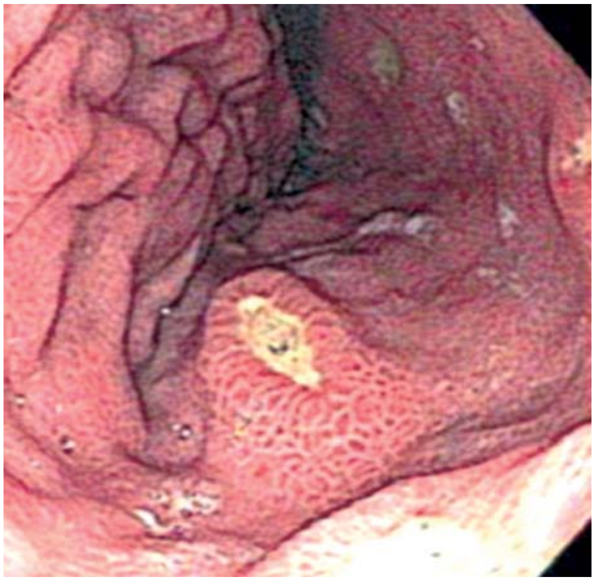

Fig. 20.19: Multiple ulcers visualised on gastroscopy.

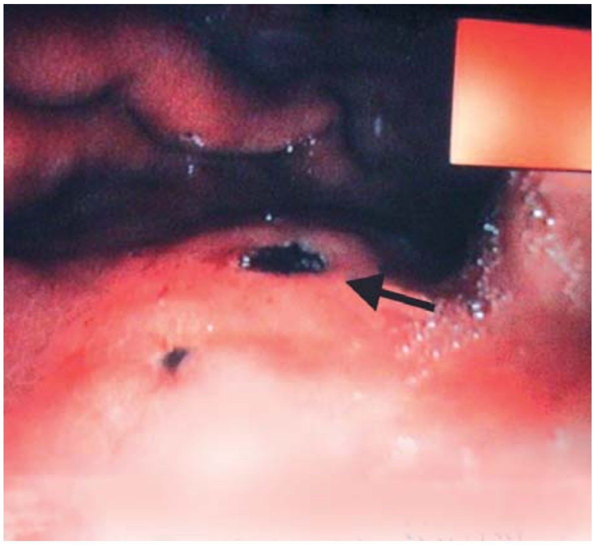

Fig. 20.20: Multiple gastric ulcers on endoscopy.

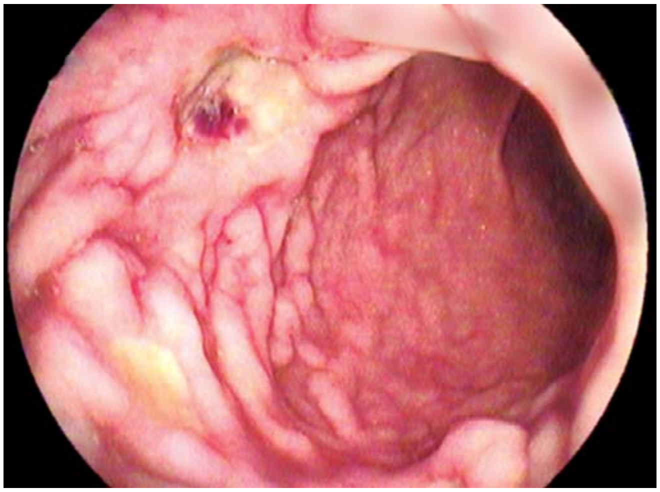

Fig. 20.21: Gastric ulcer in the body of the stomach.

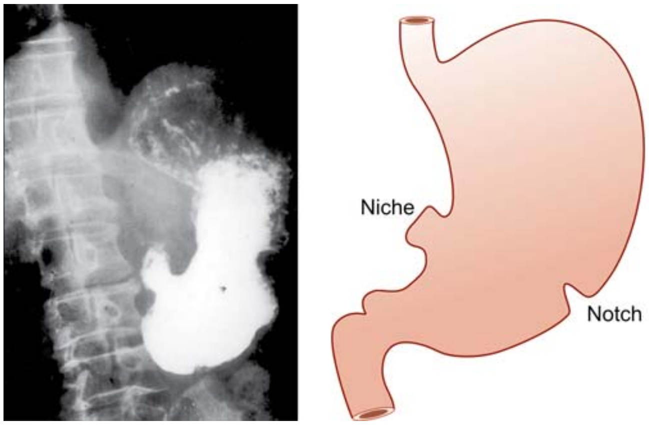

Fig. 20.22: Barium meal study showing *niche and notch*—gastric ulcer.

Treatment

- Drugs like H_2 blockers, proton pump inhibitors, carbenexolone (Biogastrone, Sucralfate, prostaglandins which coats the ulcer and so creates a mucosal barrier) helps in reducing or eliminating the symptoms.
- But asymptomatic ulcer may exist silently and may turn into malignancy.
- ***So surgery is the preferred line of treatment. Partial gastrectomy and Billroth I gastroduodenal anastomosis is done.***
- Type IV proximal gastric ulcer is difficult to manage. It is treated by subtotal gastrectomy. Often distal gastrectomy with selective sleeve like extension cut along the lesser curve to remove the ulcer is done—*Pauchet's procedure.*
- *Other surgical procedures:*
    1. *de Miguel's antrectomy*: Distal antrectomy, pylorectomy with excision of ulcer along with gastroduodenal anastomosis is done. It preserves gastric reservoir function, shows less recurrence rate and less operative morbidity.
    2. *Mak's pylorus preserving gastrectomy*: Hemigastrectomy with excision of pyloric ulcer but retaining 2 cm prepyloric stomach. It is only used in type I gastric ulcer. Even though it has got fewer incidences of postoperative diarrhoea and dumping, it has got high recurrence rate.

3. *HSV with excision of ulcer.*
4. *Kelling Madlener procedure*: It is antrectomy and excision of proximal gastric ulcer Type IV.
5. *Csendes procedure*: It is subtotal gastrectomy with sleeve extended resection along the lesser curve for type IV proximal gastric ulcer.

Complications of Gastric Ulcer

1. ***Hour glass contracture***: It occurs exclusively in ***women***, is due to cicatricial contracture of lesser curve ulcer.
    - Here stomach is divided into two compartments.

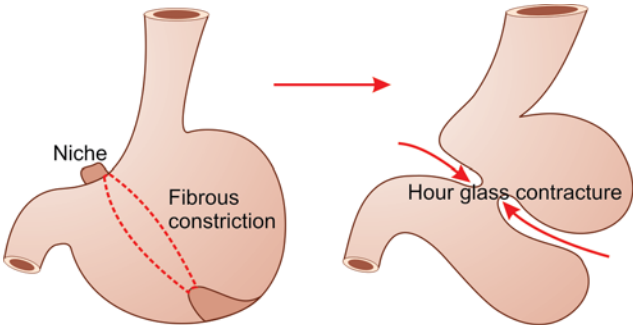

Fig. 20.23: Hour glass contracture.

- *Clinical features*
    - Loss of periodicity.
    - Persistent pain.
    - Vomiting.
    - Loss of appetite and weight.
- *Diagnosis*
    - *Barium meal*: It shows filling only in the proximal stomach or double pouched stomach.
    - Gastroscopy.
- *Treatment*

  Partial gastrectomy wherein gastric ulcer with lower compartment of the stomach is removed and Billroth I anastomosis is done.

2. ***Tea-pot deformity* (Hand-bag stomach):** It is due to cicatrisation and shortening of the lesser curvature.
    - They present with features of pyloric stenosis.
    - Treatment is partial gastrectomy with Billroth I anastomosis.

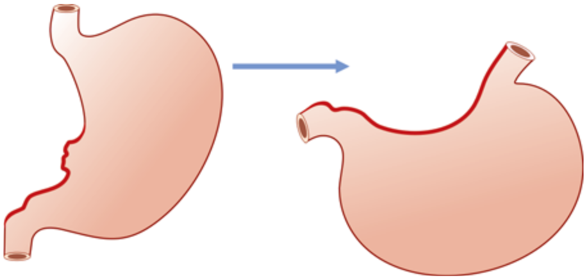

Fig. 20.24: Tea-pot deformity.

3. ***Perforation***—most frequent.
4. ***Bleeding*** *by erosion into the left gastric and rarely splenic vessels or to vessels in the wall of ulcer*—35%. It is common in type II and III gastric ulcers.
5. ***Penetration*** *posteriorly into pancreas, anteriorly into liver.*
6. ***Malignant*** *transformation usually into* ***adenocarcinoma of stomach (2-5%)***.

DUODENAL ULCER

Aetiology

- Common in people with blood group O +ve.
- Stress, anxiety—'*hurry, worry, curry*'.
- *Helicobacter pylori* infection is an important aetiology for duodenal ulcer (90%).
- NSAIDs, steroids.

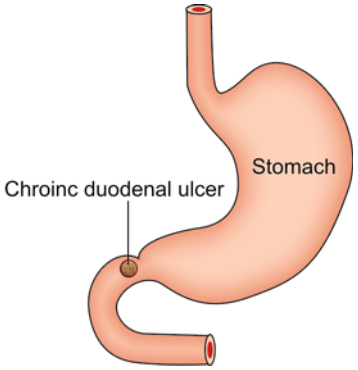

Fig. 20.25: Anatomical location of chronic duodenal ulcer.

- Endocrine causes: Zollinger-Ellison syndrome, MEN syndrome, hyperparathyroidism.
- Other causes: Alcohol, smoking, vitamin deficiency.
- *Dragstedt dictum*: "No acid – No ulcer".

Pathology

- Ulcer occurs in the first part of duodenum, usually with in the first inch,involving the muscular layer.

  Sites:

  a. In the bulb (bulbar)—95%.

  b. Post-bulbar (5%).
- Eventually it shows cicatrisation causing pyloric stenosis. Serosa overlying the site of duodenal ulcer shows petechial haemorrhages with speckled red dots, appearing like sprinkled *cayenne pepper*.

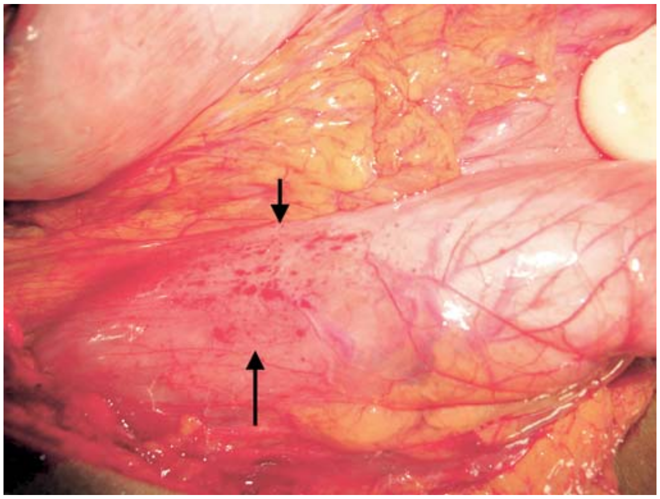

Fig. 20.26: Serosal surface in a chronic duodenal ulcer. It appears like *cayenne pepper*.

	Benign gastric ulcer	*Malignant gastric ulcer*
Mucosal folds	Converging mucosal folds upto the margin	Effacing mucosal folds
Site	95% lesser curve	Greater curvature
Margin	Regular margin	Irregular margin
Floor	Granulation tissue in floor	Necrotic slough in floor
Edge	Not everted; punched or sloping	Everted edge
Surrounding area	Surrounding area and rugae are normal	Surrounding area shows nodules, ulcers and irregularities
Size and extent	Small, deep up to part of muscle layer	Large and deep

- Microscopically, ulcer with chronic inflammation with granulation tissue, gastric metaplasia of duodenal mucosa, endarteritis obliterans are visualised.
- Sometimes two opposing ulcers, i.e. over anterior and posterior surfaces of duodenum are present and are called as *kissing ulcers.*
- An anterior ulcer *perforates* commonly, posterior ulcer *bleeds or penetrates* commonly.

Clinical Features

- In India, ratio of duodenal ulcer to gastric ulcer is 30 : 1. A very high incidence.
- It is common in all socioeconomic group, more with stressed professionals (Type A personality).
- *Pain* is more before food, in early morning and decreases after taking food. It is classically called as ***hunger pain*** as it is relieved by taking food. Night pains are common.
- Common *in males*.
- *Periodicity* is more common than in chronic gastric ulcer with seasonal variation.
- *Water-brash, heart burn, vomiting* may be present.
- *Melaena* is more common, *haematemesis* also can occur.
- Appetite is good and there is gain in weight. It decreases once stenosis develops.
- Eats more frequently without any restriction.
- Chronic duodenal ulcer can ***be uncomplicated or complicated.***

Differences between clinical features of gastric ulcer and duodenal ulcer

Gastric ulcer	*Duodenal ulcer*
Pain after food intake	Pain before food intake
Periodicity less common	Periodicity more common
Haematemesis more common	Malaena more common
Weight loss occurs	Weight gain occurs
Equal in both sexes	Common in males

Complications of Duodenal Ulcer

1. Pyloric stenosis: Due to scarring and cicatrisation of first part of the duodenum.
2. Bleeding (10%).
3. Perforation (5%). Both acute and chronic ulcers can perforate. Anterior ulcers perforate.
4. Residual abscess.
5. Penetration to pancreas.

Note:
- Chronic duodenal ulcer will not turn into malignancy.
- Ulcer which is more than 2 cm is called as *giant* duodenal ulcer.

Investigations

- Barium meal X-ray shows deformed or absence of *duodenal cap* (because of spasm). Appearance of *'trifoliate'* duodenum is due to secondary duodenal diverticula which occurs as a result of scarring of ulcer.

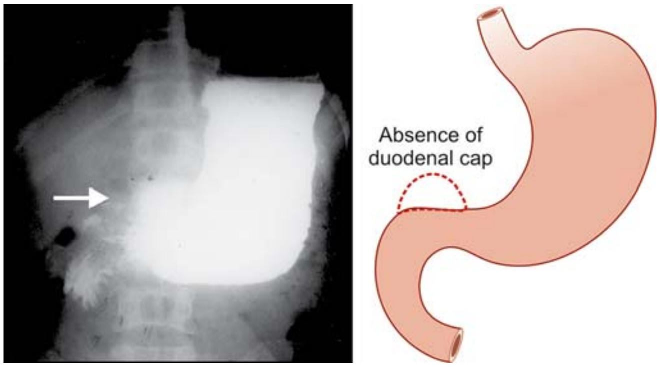

Fig. 20.27: Absence of duodenal cap—chronic duodenal ulcer.

- Gastroscopy reveals the type, location of ulcer, narrowing if any. Biopsy also can be taken to look for the presence of *Helicobacter pylori*. Usually biopsies are taken from duodenum, pylorus, antrum, body, fundus, and confirmed by rapid urease test or C13 or C14 breath tests.
- Estimation of serum gastrin level, serum calcium level.

Differential diagnosis

- Carcinoma stomach (pylorus)
- Dyspepsia due to other causes
  - Hiatus hernia
  - Oesophagitis
  - Cholecystitis
  - Chronic pancreatitis

> *Patients with peptic ulcer were regularly "milked" for acid, which was used for patients suffering from hypochlorhydria.*
> — **Albert E Coates**

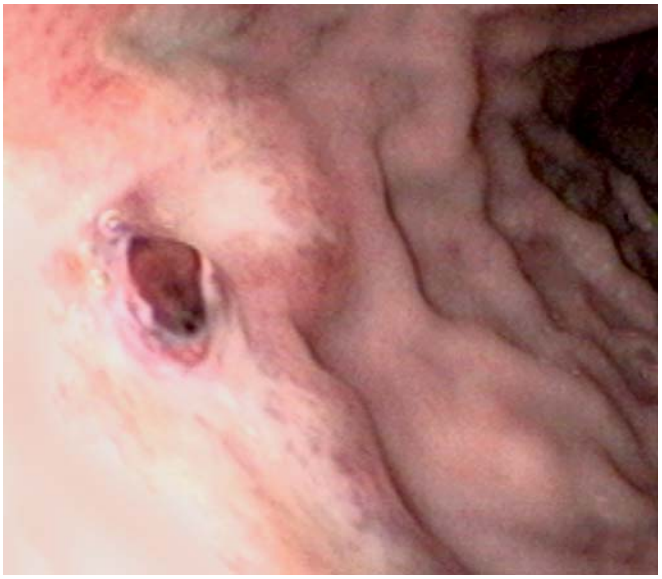

Fig. 20.28: Endoscopic view of duodenal ulcer.

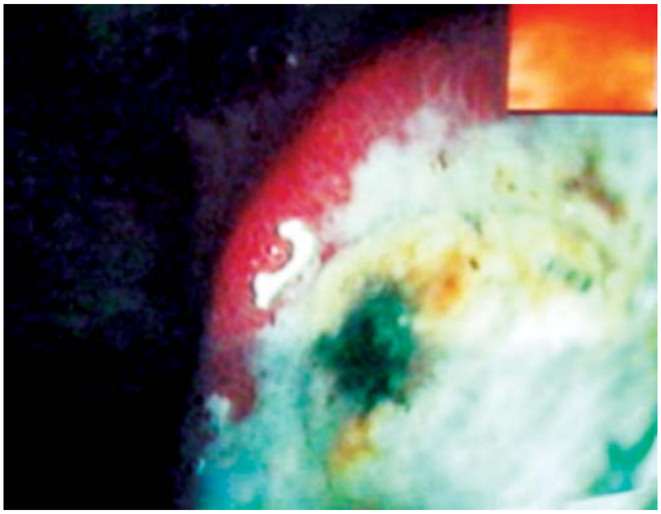

Fig. 20.29: Chronic duodenal ulcer, gastroscopic view.

Treatment

Aim of therapy:

To *relieve* symptoms; to heal ulcer; to prevent recurrence.

I. ***General measures:***
  - Avoid alcohol, NSAIDs, smoking, spicy foods. Have more frequent food.

II. ***Specific measures*:**
  - Intragastric pH should be maintained above 5.

Drugs

1. *H_2 Blockers:*
   - Promotes ulcer healing in 4-8 weeks, by reducing acid secretion.
     - Tab cimetidine.
     - Tab ranitidine (300 mg HS or 150 mg BID), (IV preparation is available).
     - Tab famotidine (IV is available) *Most potent H_2 blocker*. Dose is 20-40 mg/day.
     - Tab roxatidine.
     - Tab nizatidine.
2. *Proton pump inhibitors*:
   - Inhibit parietal cell H^+, K^+ ATPase enzyme responsible for acid secretion. They are used for 6-12 weeks. They stop acid secretion completely.
     - Omeprazole 20 mg OD 1 hour before food—IV preparation is available.
     - Esomeprazole 40 mg.
     - Lansoprazole 30 mg.
     - Pantoprazole 40 mg—IV preparation available.
     - Rabeprazole 20 mg—IV preparation available.
3. *Antacids*:
   - Neutralises the HCl to form water and salt and also inhibits peptic activity.
   - Aluminium hydroxide and magnesium trisilicate are commonly used.
   - Dose is 2 gram 2 hours *after food*.
   - Aluminium hydroxide causes constipation, magnesium trisilicate causes diarrhoea. Osteomalacia, milk alkali syndrome, rebound ulcer due to gastrin release are other complications.
4. *Sucralfate*
   - It is an aluminium salt of sulfated sucrose which provides a protective coat to ulcer crater thereby promotes healing. It inhibits peptic activity.
   - It binds to ulcer bed and stays for 12 hours; prevents back diffusion of hydrogen ion; raises endogenous prostaglandin level in tissues; binds bile acid and pepsin; prevents colonisation of gastric mucosa by bacteria.
   - Dose is 1 g qid for 6 weeks (Before food). It is an effective drug.
5. *Anti-Helicobacter pylori regime*:
   - It is very useful, given for 7-14 days—later the proton pump inhibitors are continued.
   - *Triple or quadruple* (tetracycline, bismuth, tinidazole, pantoprazole) regimes are used.
6. *Colloid bismuth sulphate* is a good drug for ulcer, but it stains the oral cavity and mucosa.
7. Misoprostol (200 mg tid) is the only prostaglandin agonist accepted.
   - PG E1 (mesoprostol) and E2 increase mucus and bicarbonate secretion, improves mucosal blood flow, but reduces acid secretion.

Follow-up gastroscopy is a must, to confirm that ulcer has healed.

***Note*:**

Antacids and H_2 blockers should not be used along with PPI as these drugs will reduce the action of PPIs by creating alkaline media.

Surgery for Uncomplicated DU

Indications for surgical intervention for chronic DU (Uncomplicated DU):

1. Uncomplicated DU, not responding to drug therapy of 8-12 weeks—**intractable duodenal ulcer**
2. Repeated recurrences
   Presently most of the uncomplicated DU does not require surgery

- *Highly selective vagotomy (HSV).*
  - In HSV, only fibres supplying the parietal cells are ligated. Nerve of Latarjet which supplies the antrum

Anti-Helicobacter regime (Triple regime)

Clarithromycin 500 mg BD	Metronidazole 400 mg BD	Omeprazole 20 mg BD
OR	OR	OR
Amoxycillin 500-750 mg BD	Tinidazole 600 mg BD	Lansoprazole 30 mg BD
OR		OR
Tetracyclines, or Bismuth, etc.		Pantoprazole 40 mg BD

  pump is retained and so no drainage procedure is required in HSV. HSV is also called *as parietal cell vagotomy or superselective vagotomy*. Here nerve fibres in last 6 cm of stomach, just proximal to pylorus are preserved (Crow's foot). Vagotomy reduces acid secretion, hence ulcer heals. *No acid, No ulcer*.
  - HSV was first described by *Amdrup and Johnston* in 1969.
  - Distal 6 cm oesophageal nerve fiber clearance is essential. Fibers up to 6 cm proximal to pylorus in stomach are cleared; nerve of Latarjet is preserved; adequate distal greater curve clearance is essential.
  - *Intraoperative test for completion of vagotomy* should be done—*Grassi test or insulin test*.
  - It has got distinct *advantages*—low operative mortality (0.2%) and postoperative morbidity (0.5%); post-vagotomy diarrhoea and dumping syndrome is very low; chances of developing anaemia, weight loss, osteoporosis, tuberculosis, and carcinoma are very less.
  - *Problems*—lesser curve necrosis due to ischaemia; recurrent ulcer 10-15% in 10 years.
- *Selective vagotomy with pyloroplasty* (SV + P).
- *Truncal vagotomy with gastrojejunostomy* (TV + GJ).
- *Posterior truncal vagotomy with anterior seromyotomy—Taylor's operation*. It can be done through laparoscopy.
- *Vagotomy with antrectomy*: Gastrin producing antrum, vagal cholinergic pathway from ulcer bearing area is removed with gastroduodenal anastomosis. Ulcer recurrence is very low but has got morbidity.
- *Posterior truncal vagotomy* with HSV without drainage procedure (Kim's) often through laparoscopy is also done.
- *Linear gastrectomy* with posterior truncal vagotomy through laparoscopy.
- Most of these procedures presently can be done through *laparoscope*.

Note:

Presently, there is *no role of gastrectomy or gastrojejunostomy (Just GJ) for uncomplicated DU.*

PYLORIC STENOSIS DUE TO CHRONIC DUODENAL ULCER

Pathology

Chronic DU after many years undergoes scarring and cicatrisation causing total obstruction of the pylorus, leading to enormous dilatation of stomach.

Clinical Features

- Pain is severe, persistent, in epigastric region, and also with feeling of fullness.
- Vomiting—large quantity, foul smelling and frothy, vomitus contains food consumed on previous day (partially digested or undigested food).
- Loss of periodicity.
- Loss of appetite and weight.
- *Visible gastric peristalsis* (VGP)—may be elicited by asking the patient to drink a cup of water.
- *Positive succussion* splash which is done with 4 hours empty stomach, by placing a stethoscope over the epigastric region and shaking the patient adequately.
- *Auscultopercussion test* shows dilated stomach.

  Test is done by placing a stethoscope over the epigastric region. Skin is scratched from left side downwards, at several points away from the epigastrium (towards left side) using finger and these points are joined. Normally greater curvature of stomach is above the level of umbilicus (midway between the umbilicus and epigastrium). In gastric outlet obstruction it lies below the level of the umbilicus (Stomach we see; stomach we feel; stomach we hear).
- Confused status because of alkalosis and electrolyte changes.
- ***Electrolyte changes***: Because of vomiting, *hypochloraemic, hyponatraemic, hypokalaemic, hypocalcaemic, hypomagnesaemic alkalosis occurs. It causes **paradoxical aciduria**.*
- Mass is never palpable.
- *Goldstein saline load test*—half an hour after installation of 750 ml of saline, if volume remained and if more than 250 ml, suggest obstruction.

Investigations

1. *Barium meal study:*
   - Absence of duodenal cap.
   - Dilated stomach where greater curvature is below the level of iliac crest.
   - Mottled stomach.
   - Barium will not pass into duodenum.
2. Gastroscopy to rule out carcinoma stomach and to visualise the stenosed area.
3. Electrolyte study for correction of electrolyte imbalance.
4. ECG to check for hypokalaemia.

An inn has one entrance and three exist—to hospital, to asylum and to prison. —**Croatian Proverb**

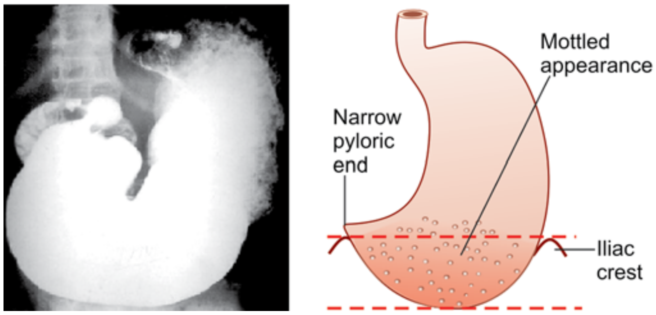

Fig. 20.30: Pyloric stenosis with gastric dilatation.

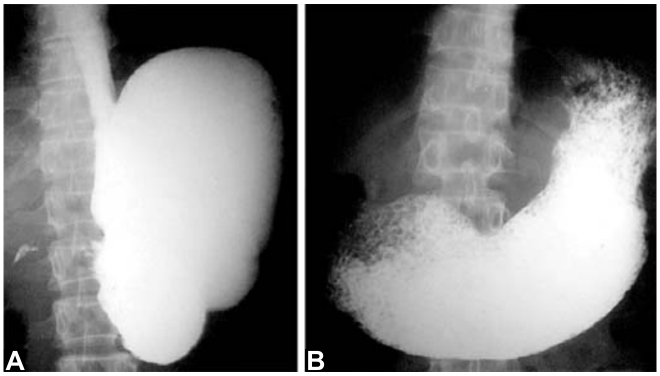

Figs 20.31A and B: Barium meal pictures showing gastric outlet obstruction.

Differential Diagnosis

Carcinoma pylorus—here mass may be palpable.

Pyloric stenosis—causes

- Congenital
- Chronic DU—fibrosed/cicatrised
- Carcinoma pylorus
- Adult pyloric stenosis—it is treated by pyloroplasty (not by pyloromyotomy)
- Pyloric mucosal diaphragm—it should be excised surgically or endoscopically

Treatment

- Correction of dehydration and electrolytes by IV fluids—normal saline or double strength saline, calcium, potassium, magnesium.
- Blood transfusion is given if there is anaemia.
- TPN support.
- Stomach wash to clean the stomach contents (using normal saline) is given using stomach tube like Eswald's. It also reduces the oedema of stomach wall and improves gastric emptying time by increasing the gastric muscle tone.
- Surgery
  - *HSV with gastrojejunostomy* is present recommendation even though it is technically difficult. HSV is better than TV as it maintains the nerve supply of the chronically obstructed antrum and so may eventually reduce the chronic emptying problems.
  - Truncal vagotomy along with gastrojejunostomy of *Mayo (posterior, vertical/oblique, short loop, retrocolic, isoperistalsis) is done—ideal, commonly advocated procedure.*
  - Vagotomy, antrectomy (acid secreting area) with Billorth I anastomosis along with feeding jejunostomy for nutrition is the other option.

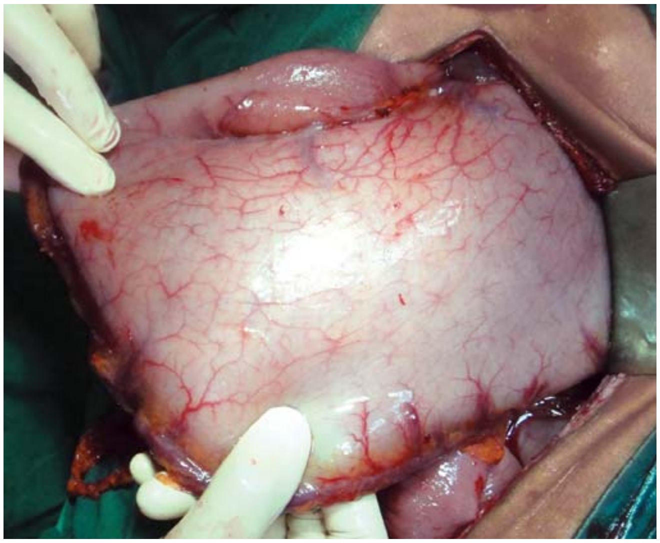

Fig. 20.32: *Gastric outlet obstruction* showing dilated stomach on table.

Note:

- Often gastric emptying may be delayed for 2-4 weeks after surgery in pyloric stenosis. It usually recovers in 7 days.
- After recovery eradication of *H. pylori* infection is routinely done even though infection may not be evident in many of patients with outlet obstruction.
- *No role of pyloroplasty or HSV in a scarred duodenum—as it can cause disruption and bile leak.*
- Gastrectomy and other procedures are usually not necessary in pyloric stenosis.
- Endoscopic balloon dilatation of the stenosed area benefits only temporarily with high recurrence rate.

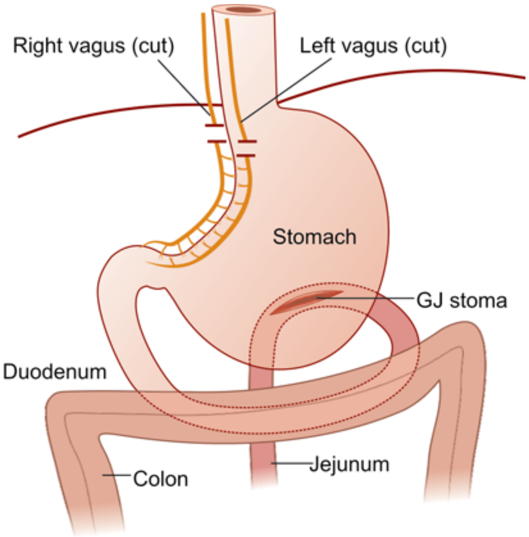

Fig. 20.33: Truncal vagotomy and gastrojejunostomy. *It is posterior, vertical, short loop, retrocolic, isoperistaltic GJ of Mayo.*

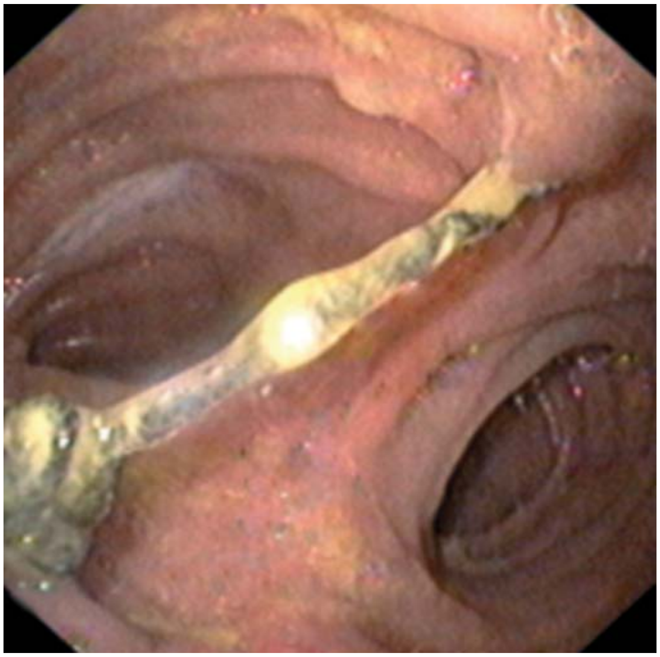

Fig. 20.34: Gastrojejunostomy stoma view on endoscopy.

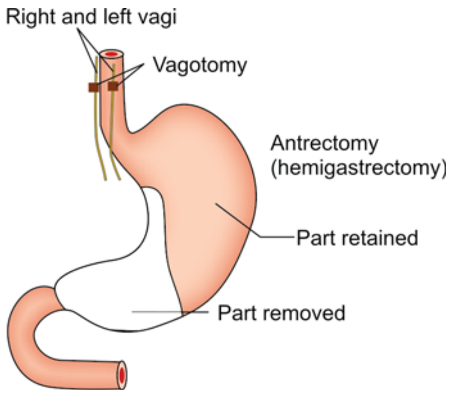

Fig. 20.35: Vagotomy and antrectomy. It is occasionally used for intractable resistant duodenal ulcer or chronic duodenal ulcer with obstruction. HSV is commonly used in intractable cases. GJ with vagotomy is commonly used in pyloric stenosis due to chronic duodenal ulcer.

ELECTROLYTE CHANGES IN PYLORIC STENOSIS

- Hyponatraemia.
- Hypokalaemia.
- Hypomagnesaemia.
- Hypochloraemia.
- Metabolic alkalosis.
- Paradoxical aciduria.

Because of severe vomiting which contains acid as well as undigested retained food, sodium, chloride, potassium, magnesium levels drop, causing metabolic alkalosis. Alkalosis can lead to hypocalcaemia causing tetany. It is called as ***'gastric tetany'.***

To control alkalosis, kidney secretes excess bicarbonate (In alkalosis without hyponatraemia, bicarbonate is secreted along with sodium). Here, due to hyponatraemia, body conserves sodium and so bicarbonate is secreted along with hydrogen ion. So urine becomes *acidic*. It is called as ***paradoxical aciduria.***

Clinical Features

- Irritability, confused status, dehydration.
- Often convulsions can occur.
- Features of alkalosis like rapid breathing—Cheyne-Stokes breathing and tetany.

Investigations

- Serum electrolytes.
- Arterial blood gas analysis.
- Serum calcium level estimation.

Treatment

- Double strength normal saline.
- IV potassium given slowly under ECG monitoring.
- The cause is treated.
- IV magnesium.

Note:

Metabolic changes are not severe in carcinoma pylorus as it is seen in pyloric stenosis due to chronic duodenal ulcer because in carcinoma stomach often there is hypochlorohydria or achlorohydria.

PERFORATED PEPTIC ULCER

- It is the terminology used for *perforation of duodenal ulcer or gastric ulcer or stomal ulcer.* Otherwise all clinical features and management are similar.
- Perforation is common in duodenal ulcer (75% of perforated peptic ulcers). Mortality is more in gastric ulcer perforation and perforation in elderly.

A. PERFORATED DUODENAL ULCER

- It is common in males (8:1) between 35-45 years of age group, but can occur in any age group.
- ***Anterior ulcer*** perforates, commonly
    - In 80% of cases, there is a history of chronic DU.
    - In 20% cases, it is silent perforation.
- Perforation can occur in acute ulcers or in acute presentation of a pre-existing chronic ulcer.
- Perforation may be precipitated by steroids, analgesics (NSAIDs), alcohol, antimalarials.
- Overall incidence is 5%.
- Active ulcers perforate commonly.

Stages of Perforation

- ***Stage of chemical peritonitis:***
    - Once perforation occurs, stomach contents escape into the peritoneal cavity. The acid from the stomach causes chemical peritonitis leading to severe pain in epigastric region, vomiting, tenderness, guarding, rigidity, tachycardia, sweating.

Things turn out best for people who make best of the way things turn out.

- ***Stage of reaction (Stage of illusion):***
  - Peritoneum secretes lot of fluid to neutralise the escaped content and so temporarily the pain reduces, and the patient feels better. This phase lasts for about 6 hours.
- ***Stage of diffuse bacterial peritonitis:***
  - After about six hours, bacteria from GIT (escape) migrate from the site of perforation causing diffuse peritonitis.

Clinical Features

- Presents with *severe persistent pain* in the epigastrium initially, later in the right side abdomen (as the inflammatory fluid spills along the right paracolic gutter) and finally becomes generalised. Pain is of sudden in onset, is due to contact of expelled gastric contents with the parietal peritoneum. Pain often radiates to right scapular region. Pain becomes more on movements.
- Tenderness and rebound tenderness is seen (*Blumberg sign*) all over the abdomen.
- Fever, vomiting, dehydration, oliguria occurs.
- Patient is toxic, with tachycardia, hypotension, tachypnoea.
- Abdominal distension occurs.
- Guarding and rigidity, initially in the epigastrium but later all over the abdomen.
- Dullness over the flank because of fluid.
- *Obliteration of liver dullness*—as a result of collection of escaped gas under the diaphragm.
- Silent abdomen with absence of bowel sounds.
- Tenderness felt on per rectal examination.
- Sometimes fluid from supracolic region slowly trickles down along the right paracolic gutter and collects in right iliac region causing pain and tenderness in RIF mimicking appendicitis.
- Often slow, small perforation presents with subacute features, but diffuse peritonitis eventually sets in 24-48 hours.

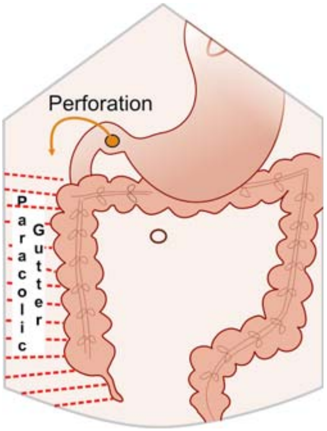

Fig. 20.36: Spread in DU perforation.

Terminal stage:

Patient may have oliguria, septicaemia, shock, *Hippocratic facies* (sunken eyes, cold periphery and shallow rapid breathing, ill look), with MODS (Multiorgan dysfunction syndrome).

Investigations

- *Plain X-ray abdomen (erect posture)*: Shows gas under diaphragm in 70% of cases. In 30% of cases, there is no gas under diaphragm. It may be due to, either the gas leak is less than 1 ml or due to previous surgery causing adhesions between liver and diaphragm, or sealed peptic ulcer.

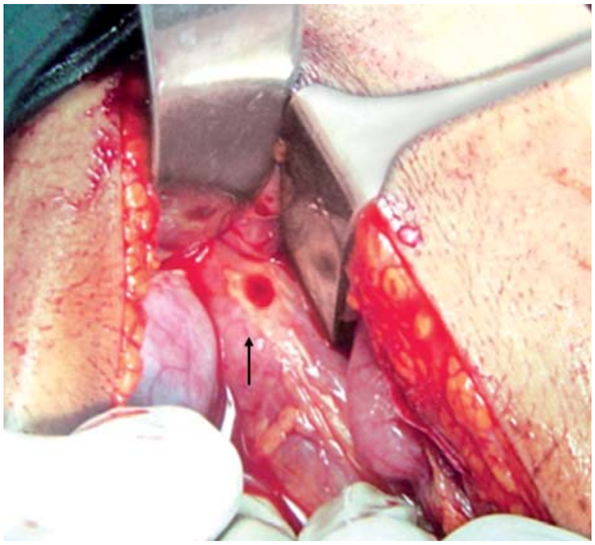

Fig. 20.37: Note the site of the DU perforation, on table.

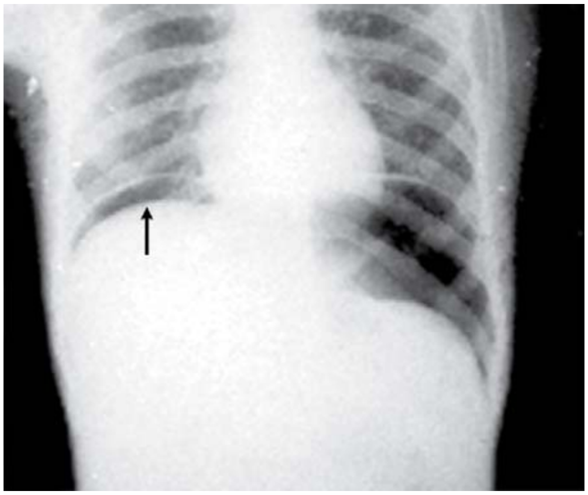

Fig. 20.38: Plain X-ray abdomen in erect position with ground-glass appearance and gas under diaphragm—perforation.

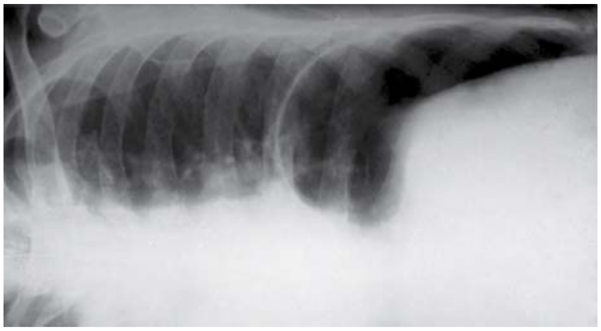

Fig. 20.39: Lateral decubitus X-ray showing air under diaphragm. This is done when patient is critically ill and cannot make the patient stand in erect posture.

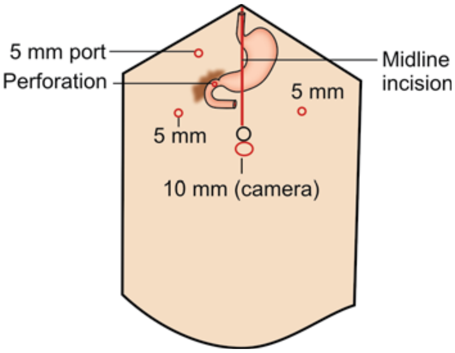

Fig. 20.40: Approaches to manage duodenal ulcer perforation can be either open (upper midline incision) laparotomy incision or laparoscopic.

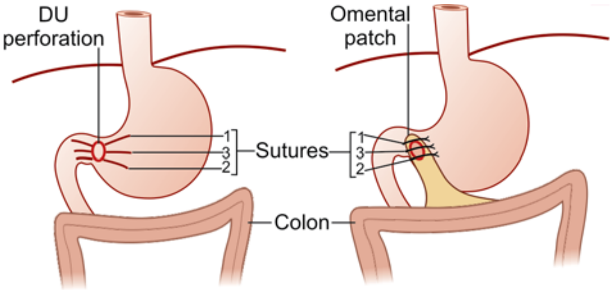

Fig. 20.41: Duodenal ulcer perforation closed horizontally using interrupted silk/vicryl sutures with omental patch over it.

(***Chilaiditi's syndrome*** is the interposition of the colon in front and above the liver. It is common in children and elderly. It may be mistaken for gas under diaphragm in plain X-ray abdomen).

- U/S abdomen shows free fluid and often gas.
- Blood urea, serum creatinine, total count, electrolytes, are helpful.
- CT scan abdomen is very sensitive investigation whenever there is absence of gas under diaphragm. It rules out other conditions like pancreatitis. Gastrograffin upper GI study also confirms the perforation.

Differential diagnosis

- Acute appendicitis
- Acute pancreatitis
- Acute cholecystitis
- Ruptured aortic aneurysm
- Myocardial infarction
- Mesenteric ischaemia
- Pneumonia

Treatment

- Patient is advised admission.
- IV fluids—Ringer lactate, normal saline, dextrose saline.
- Antibiotics—Cefotaxime, metronidazole, amikacin.
- Catheterisation.
- Ryle's tube aspiration.

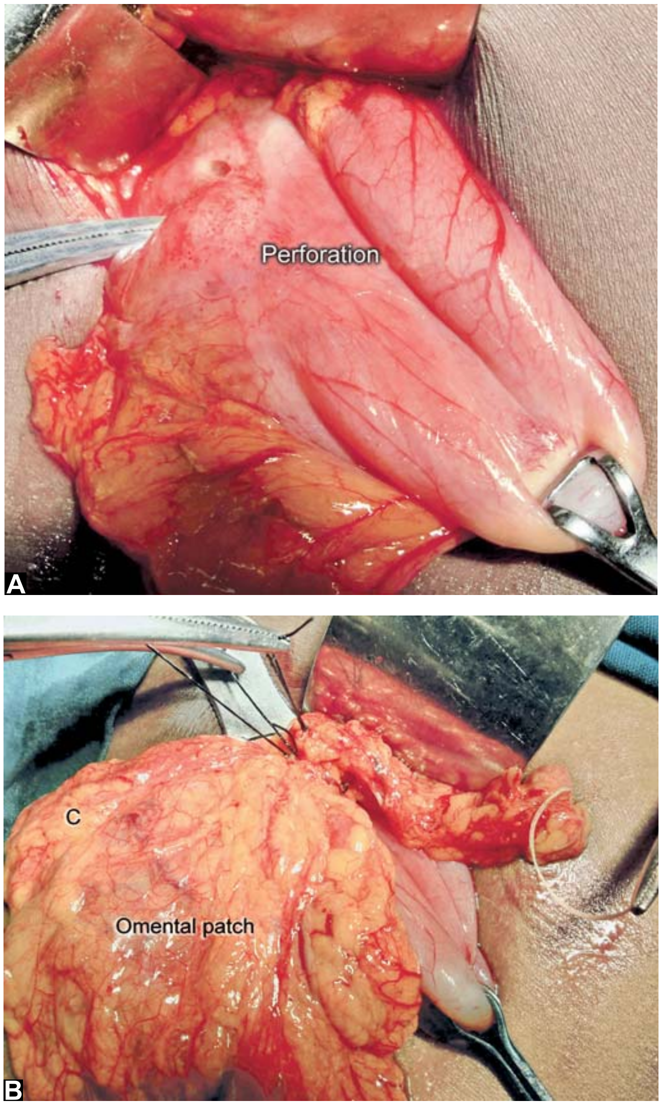

Figs 20.42A and B: Duodenal ulcer perforation and its closure using omental patch.

- *Surgery:*
  - Emergency laparotomy through upper midline incision is done.
  - All infected fluid is sucked out.
  - Perforation is identified and closed with interrupted, horizontal sutures using *either silk or vicryl*. Omental patch is placed *before suturing*—it is called as ***Rosoe-Graham Operation.*** Because of its *adhesion property* it seals perforation; good vascularity and lymphatics promote the healing effectively.
  - Peritoneal wash (toilet) using 5-10 litres of saline is given.
  - Drain is placed and abdomen is closed, often if required with *tension sutures*. Drain is removed in 3-5 days. During discharge, patient is advised to avoid alcohol and to take H_2 blockers or proton pump inhibitors for 6-12 weeks.
  - After 12 weeks *follow-up gastroscopy must be done.*

65% of haematemesis is due to bleeding peptic ulcer. Among peptic ulcer, duodenal ulcer (35%) is the most common cause.

Note:

- Very rarely, in elderly people or in cases of subacute perforation which gets sealed on its own, *conservative treatment is* tried with careful observation, i.e. by giving IV fluids, antibiotics, Ryle's tube aspiration, maintenance of urine output and electrolytes. It is called as *Hermen-Taylor regime*. But it should not be a standard treatment.
- Laparoscopy is beneficial for both diagnoses, as well as for therapy for perforation closure using omental patch. It is proposed and widely used for early duodenal perforation. It can be useful for most of the duodenal ulcer perforations. If it is problematic, conversion options should always be kept in mind.
- Once perforation is closed, either by open or laparoscopic method, after recovery patient should be advised, *anti-Helicobacter pylori* therapy (triple regime for 14 days) with PPI for 3 months. It reduces rate of reperforation and ulcer recurrence in duodenal ulcer.
- In patients with severe peritonitis and critically ill, after perforation closure it is better to insert a nasojejunal tube or feeding jejunostomy for nutrition in postoperative recovery period. Even if patient develops a temporary duodenal leak, this supports nutrition well until the leak stops.
- Occasional large perforated duodenal ulcer with oedema which cannot be closed is managed by serosal patch/duodenal drainage and pyloric exclusion/gastrostomy, duodenostomy and jejunostomy/Roux-en-Y jejunal patch over the perforation.
- *Manheim peritonitis index* or *APACHE II* scoring system is used to assess the patient properly.
- Doing concomitant acid reducing surgical procedure like HSV is under debate and not used routinely (not advised) even though some centers advocate. HSV may be done if there is not much contamination of the peritoneal cavity or in early perforation of a chronic duodenal ulcer.

B. PERFORATED GASTRIC ULCER

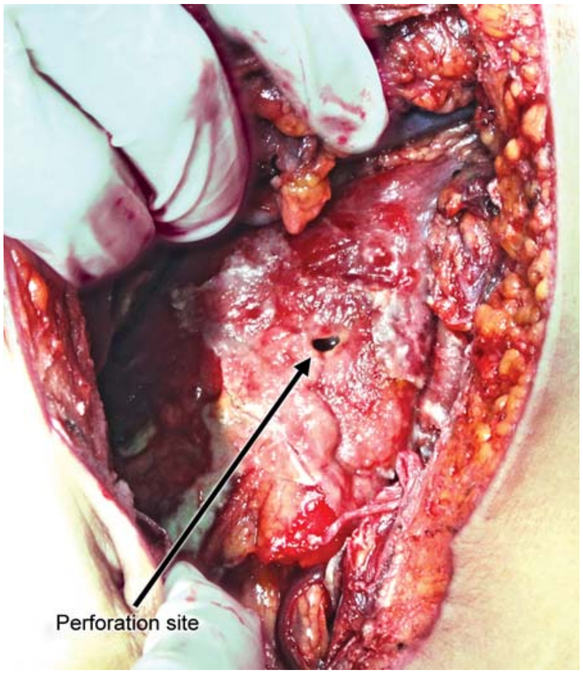

Fig. 20.43: Malignant gastric ulcer showing perforation.

- ♦ Commonly ulcer in the lesser curve near the antrum perforates. Amount of gas escaped is more than the perforated DU. Malignancy should always be suspected and so *biopsy from the edge is a must.*
- ♦ Mortality in gastric ulcer perforation is high (20%).
- ♦ Commonly they are prepyloric in position.
- ♦ Primary closure with an edge biopsy is commonly used. Distal gastrectomy including ulcer area is better option if patient's general condition is favourable.
- ♦ Posterior gastric ulcer perforation is often difficult to diagnose both clinically and radiologically.

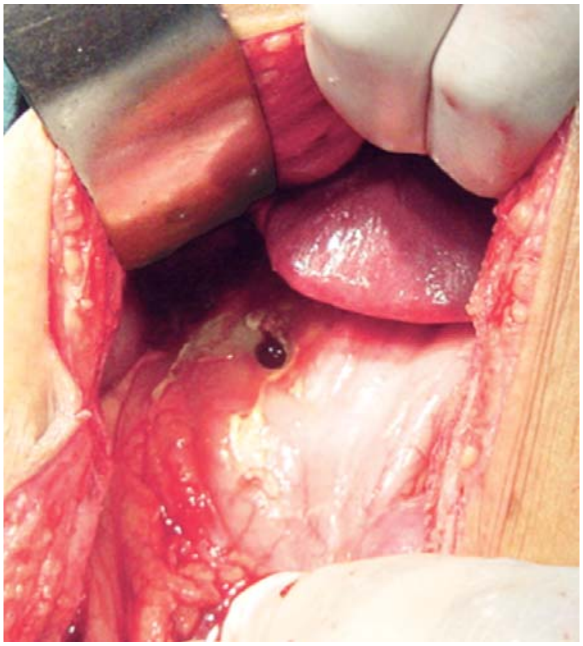

Fig. 20.44: Gastric ulcer perforation on table finding.

C. PERFORATED STOMAL ULCER

In perforation of stomal ulcer*, often undoing of GJ or partial gastrectomy may be required.*

D. DRY PERFORATION
(Perforated Duodenal Ulcer Sealed by Omentum)

- ♦ Patient is ambulatory.
- ♦ Vomiting is absent.
- ♦ Rigidity confined to epigastrium and right hypochondrium.

Different signs in X-ray in perforation

- ❖ *Cupola sign*—crescent shaped radiolucency under the diaphragm
- ❖ *Riglers sign*—visualisation of both aspects of the bowel wall being outlined by gas on either side
- ❖ *Inverted V sign*—gas on either sides of the falciform ligament
- ❖ *Football sign*—collection of gas in the center of the abdomen like a foot ball
- ❖ *Triangle sign*—gas between bowel loops

Conditions which mimic pneumoperitoneum—pseudopneumoperitoneum

- Subpulmonary pneumothorax
- Chilladiti syndrome
- Subphrenic abscess due to infections by gas forming organism like *Clostridium welchii*
- Subdiaphramatic fat or omental fat under the diaphragm may rarely mimic gas under the diaphragm

BLEEDING PEPTIC ULCER

- It is **bleeding either from duodenal ulcer, or gastric ulcer or stomal ulcer**.
- In bleeding from stomal ulcer, partial gastrectomy is required (Fig. 20.45).
- Mortality in bleeding peptic ulcer is high (20-30%). Elderly age, associated systemic diseases increase the mortality.
- NSAID's and *H. pylori* infection, coagulopathy, and anticoagulant drugs are common precipitating factors. Concomitant use of NSAID and steroids increase risk by 10 fold.
- Need of more than 5 units of blood transfusion during hospital stay is called as *massive haemorrhage*.

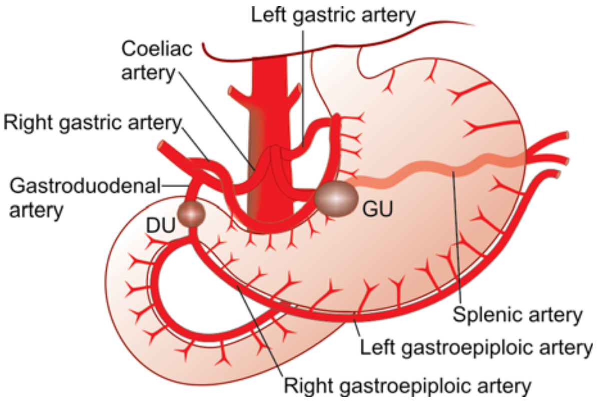

Fig. 20.45: Sites of gastric ulcer and duodenal ulcer bleeding. Note GU can erode into the left gastric artery. DU can erode into the gastroduodenal artery.

A. BLEEDING DUODENAL ULCER

- 10% common.

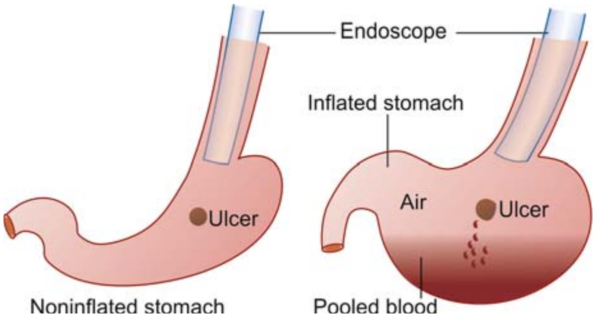

Fig. 20.46: In bleeding ulcer during endoscopy stomach should be inflated with air adequately so that bleeding ulcer is visualised well.

- Risk of bleeding in chronic duodenal ulcer increases to 35% if patient has not taken specific anti-*Helicobacter pylori* therapy and PPI.
- Bleeding from DU is either from the *small vessels in the wall of ulcer* crater or due to erosion into the *gastroduodenal artery*.

Precipitating causes

- Alcohol
- NSAIDs, steroids
- Excessive fibrosis
- Atherosclerotic disease

Forrest classification of bleeding or bled ulcer

Ia	Spurting and bleeding
Ib	Nonspurting but active bleeding
IIa	Visible vessel with red or blue protrusion or pulsatile pseudoaneurysm
IIb	Nonbleeding ulcer with clot overlying
IIc	Ulcer with haematin base
III	Clean ulcer—no clot, no vessel

- Usually *posterior* duodenal ulcer bleeds.
- Bleeding from small vessels in the wall of ulcer is due to sloughing of the ulcer. It is less severe, gradual and most often well-controlled by conservative treatment.
- Bleeding from erosion of gastroduodenal artery is severe, torrential and almost always needs early surgical intervention.

Clinical Features

- *Haematemesis and melaena.*

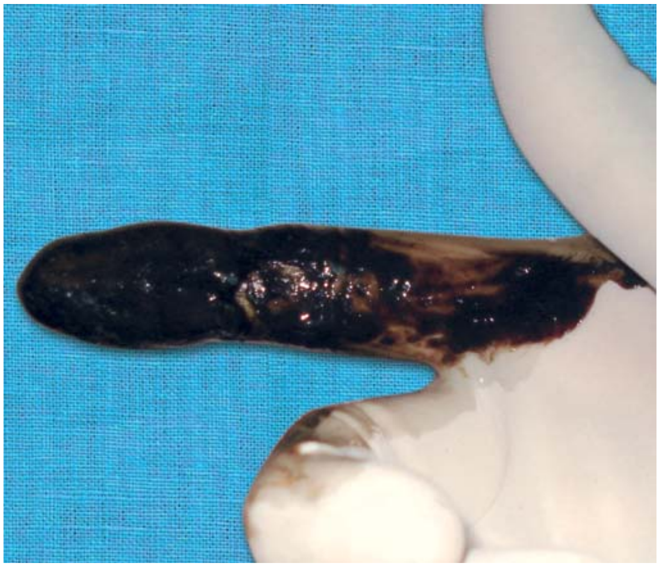

Fig. 20.47: Melaena—typical *black, tarry* coloured *foul* smelling stool. It signifies upper GI bleed. At least 50 ml bleeding should be there to cause melaena.

- Features of shock: Pallor, tachycardia, sweating, hypotension, tachypnoea, dry tongue, cold periphery.
- Past history of chronic DU may be present. But it is not always necessary in every patient, as some may have a silent ulcer which may present as bleeding and haematemesis to begin with.
- History of pain and tenderness in epigastric region which has increased in intensity recently.

Differential diagnosis

- Erosive gastritis
- Oesophageal varices
- Carcinoma stomach
- Bleeding gastric ulcer
- Mallory-Weiss syndrome
- Gastric polyps
- Bleeding disorders

Investigations

To look for in endoscopy, in bleeding ulcer

- Spurter
- Clot
- Visible vessel
- Aneurysmal dilatation of the arteriole in the wall of ulcer
- Ooze

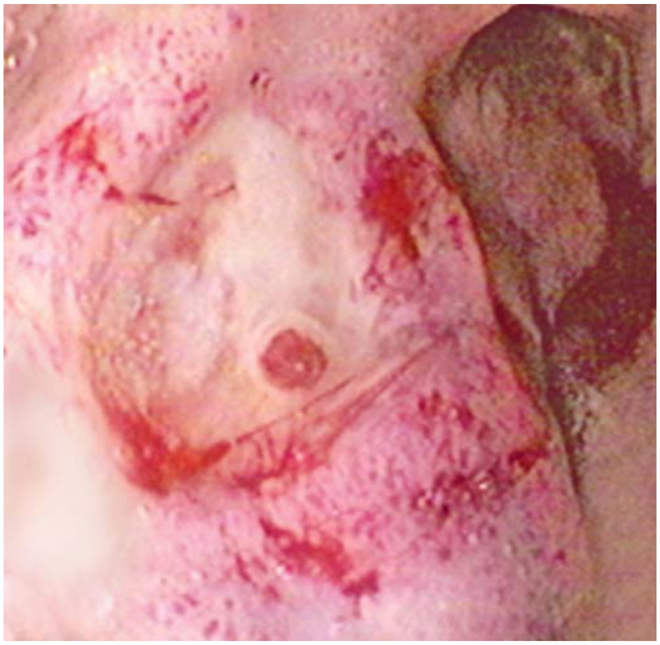

Fig. 20.48: Bleeding duodenal ulcer endoscopic view.

- *Gastroscopy is confirmative*—it is a must. It identifies ulcer bleed in 90% of cases clearly. Possibility of rebleed is also assessed by endoscopy. A flat clear based ulcer is less likely to rebleed. Active ulcer/fresh clot/visible vessel/pseudoaneurysm/large ulcer are more likely to rebleed. *Rockall scoring system* is used to predict rebleed.
- *Coeliac angiogram* to identify the bleeder may be helpful.
- Hb% and PCV—should be repeated at regular intervals (once in 2-3 hours).
- Blood group and cross-matching.
- Estimation of serum electrolytes, blood urea, serum creatinine, platelet count.

Treatment

Seventy percent of bleeding duodenal ulcers are treated conservatively.

- The shock is corrected initially by:
  - Foot end elevation.
  - IV fluids, plasma expanders (haemaccel, dextran, crystalloids). CVP line is better in these patients.
  - Sedation.
  - Catheterisation—to assess urine output.
  - Blood transfusion to replace the lost blood.
- Stomach wash is given—1 : 2,00,000 adrenaline in saline wash is given to the stomach through Ryle's tube.
- IV ranitidine 50 mg 6th or 8th hourly. IV famotidine.
- IV pantoprazole 80 mg in 100 ml dextrose saline is given slow IV as starting dose and later 40 mg in dextrose saline IV OD/BD. Slow continuous infusion of pantoprazole 40 mg in dextrose saline, 500 ml IV can also be given.
- *Endoscopic cauterisation* of small vessel with either gastroscopic bipolar cautery or through Laser or through heater probe or through haemoclips can be tried to stop the bleeding.
- *Sclerotherapy*—ethanolamine oleate, distilled water. *Epinephrine injection* is also used *commonly.* Other agents used are absolute alcohol, polidocanol.
- Observation
  - Patients with bleeding from small vessels in the wall of DU will commonly respond to conservative treatment.
- *Angiographic embolisation of gastroduodenal artery.*

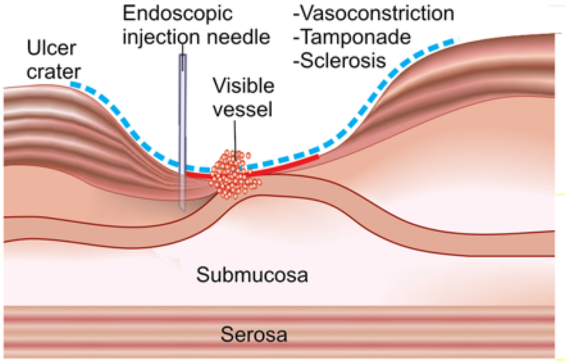

Fig. 20.49: Sclerosants like ethanolamine oleate or distilled water are used through endoscopy to control bleeding ulcer. It causes tamponade, vasoconstriction, sclerosis. It is the best method to control the bleeding.

Indications for surgical treatment

- In spite of conservative treatment condition of patient deteriorates
- Bleeding from gastroduodenal artery
- Recurrent bleeding
- Elderly patient
- If more than 4 units of blood required immediately

- ***Surgery***
  - After laparotomy, pyloric channel and first part of the duodenum (gastroduodenum) is opened longitudinally, bleeder is identified. *Underruning of* the bleeding area with vicryl is done.
  - If the bleed is from gastroduodenal artery, then it has to be *ligated* to stop the bleeding. Opened gastroduodenum is closed by ***Finney's pyloroplasty.***
  - Truncal vagotomy may be done together, if the general condition of the patient is good.

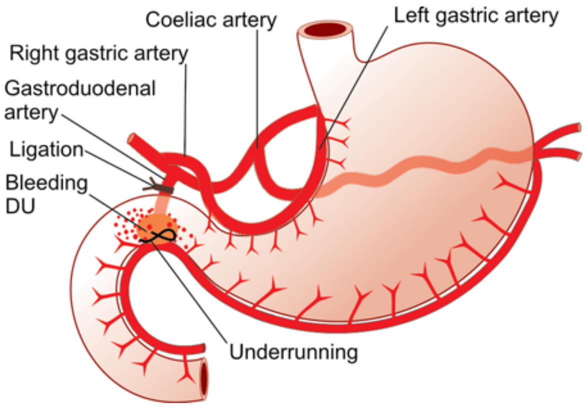

Fig. 20.50: Gastroduodenal artery ligation is required in case of severe massive upper GI duodenal ulcer bleed.

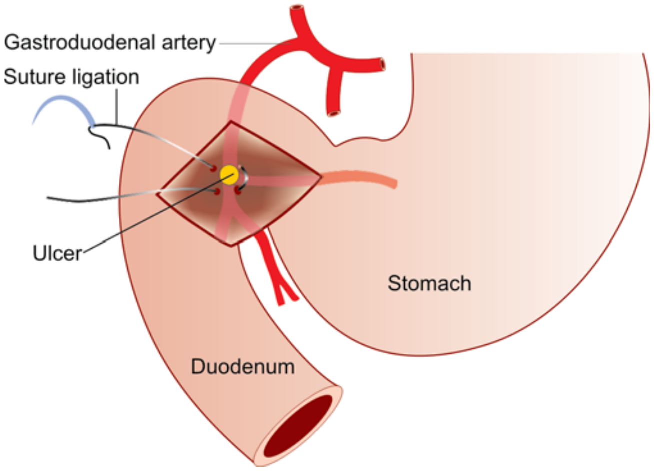

Fig. 20.51: Underrunning of the bleeding duodenal ulcer at ulcer bed is done to control haemorrhage. After that duodenum is closed with Finney's pyloroplasty.

- ***Further treatment***
  - During discharge, patient is advised to take *anti-Helicobacter pylori* (triple) therapy, proton pump inhibitors for 6-12 weeks (Omeprazole, or Lanzoprazole).
  - Healing of ulcer can be confirmed by doing a gastroscopy after 6-12 weeks.

Treatment schedule for bleeding duodenal ulcer

- Treatment of shock
- Blood transfusion
- Stomach wash with—1 : 2,00,000 adrenaline saline
- IV ranitidine, famotidine, pantoprazole, omeprazole
- Endoscopic sclerotherapy using ethanolamine oleate or distilled water
- Endoscopic bipolar cauterisation, laser therapy, heater probe, haemoclips
- Angiographic embolisation of gastroduodenal artery
- Open surgery and underrunning of the bleeding ulcer bed
- Ligation of gastroduodenal artery and then Finney's pyloroplasty
- Proton pump inhibitors during discharge
- Follow-up gastroscopy after 3 months

Sclerotherapy is the most popular endoscopic method used at present

B. BLEEDING GASTRIC ULCER

- It is similar to bleeding DU. Bleeding may be either from the ulcer bed or from the erosion of gastric vessels commonly but occasionally splenic vessels. They commonly present with severe haematemesis and shock.
- Bleeding is much more severe than bleeding DU.
- Surgery is the main treatment:
  After initial resuscitation, blood transfusion and a trial of conservative management, laparotomy is done.

On laparotomy one of the following procedures is done.

Underrunning of the ulcer bed.

OR ligation of splenic vessels with splenectomy with or without partial gastrectomy and gastroduodenostomy.

Postoperative management and follow-up is same like bleeding DU.

Treatment for bleeding gastric ulcer

- Underrunning of the ulcer bed
- Partial gastrectomy with Billroth I anastomosis
- Vagotomy with antrectomy
- Occasionally splenic vessel ligation with splenectomy may be required

HAEMATEMESIS

Causes:

- Chronic peptic ulcer (duodenal + gastric) (65%).
- Acute peptic ulcer.
- Acute erosive gastritis (Steroids, NSAIDs).
- Oesophageal varices.
- Mallory-Weiss syndrome (5-15%)
- Carcinoma stomach (5%)
- Gastric polyps, lymphomas, leiomyomas.
- Portal gastropathy.

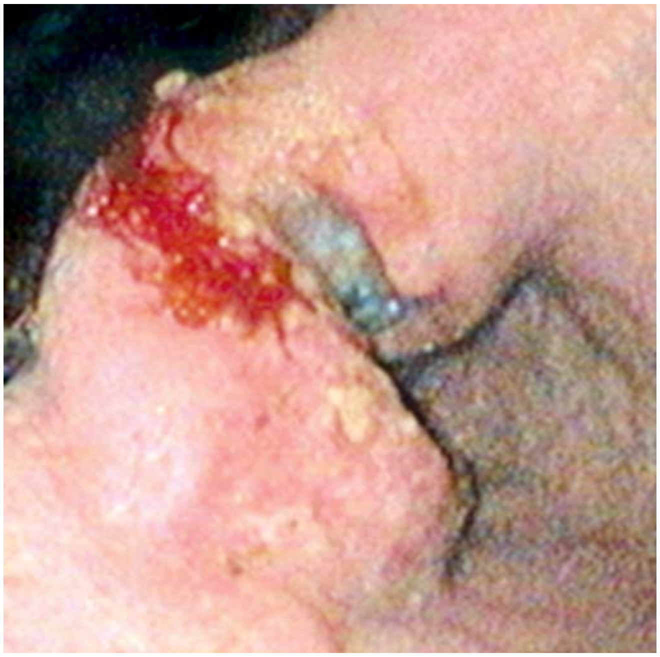

Fig. 20.52: Gastroscopy showing carcinoma stomach which is bleeding. Note the slough and everted edge.

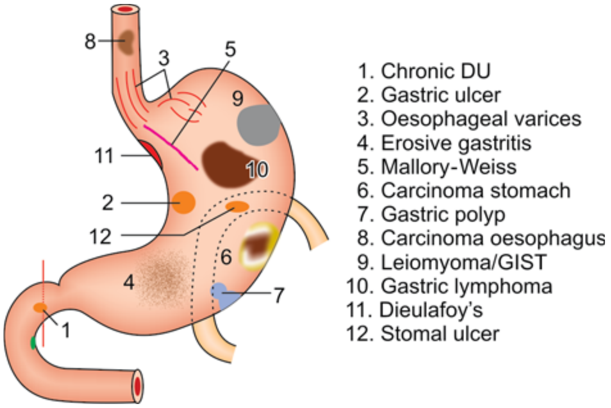

Fig. 20.53: Various causes of haematemesis.

- Bleeding disorders.
- Pernicious anaemia.
- Thrombocytopenia.
- ***Gastric antral vascular ectasia***
  - It is a rare endoscopically confirmed condition which shows segmented dilated vessel meshes in the antral mucosa (*water-melon/tiger stripe stomach*).
  - It is often associated with achlorhydria and hypergastrinaemia.
  - It is *common* in middle aged females; *common* in liver diseases (25%) and autoimmune connective tissue disorders.
  - *Pathologically* it shows mucosal fibromuscular hyperplasia and hyalinisation.
  - After confirmation with *gastroscopy*, often *antrectomy* is needed.
- Osler-Weber Rendu syndrome, aortoduodenal fistula, Crest syndrome are rare causes.
- ***Dieulafoy's disease***
  - A gastric arteriovenous malformation which is covered by apparently normal mucosa.
  - It occurs in proximal stomach near OG junction (within 6 cm) along lesser curve (80% of cases).
  - Bleeding often may be severe and torrential.
  - Vasculitis or atheroma are absent in the vessel.
  - It is 5% of nonvariceal upper GI bleed.
  - A large 1-3 mm tortuous abnormal submucosal artery (AVM) is the cause, which due to its pulsation erodes the mucosa to expose itself to acid which further erodes the artery causing bleeding.
  - Endoscopy and endoscopic therapy or excision of the lesion is required.
  - Angiography can be done to confirm the disease and to do therapeutic embolisation using gel foam.
  - Failure of endoscopic or angiographic therapy needs gastrotomy and excision of the entire lesion—*gastric wedge resection*. It can be done by open/laparoscopic approach. Prior endoscopic tattooing is mandatory to identify the lesion during resection.

Management of Haematemesis

- *Evaluation of patient* by measuring BP, pulse, respiration looking for features of shock, oxygen saturation, investigating for Hb%, blood grouping, blood urea, serum creatinine, LFT, prothrombin time, platelet count, arterial blood gas analysis, gastroscopy.
- *Initial treatment* is central line insertion, fluid and blood replacement; catheterisation; IV PPI; nasogastric tube placement (controversial but now universally accepted); FFP, platelet transfusion if needed; gastroscopic sclerotherapy, banding, laser, haemoclip application; SB tube for varices; pharmacotherapy.
- *Further treatment*—critical care (ICU); antibiotics; treatment of complications like sepsis, DIC, ARDS.
- *Specific treatment*—open surgery for uncontrolled bleeding, ligation of gastroduodenal artery and underrunning of ulcer bed with pyloroplasty, gastrectomy, ligation of varices with devascularisation.
- *Definitive surgery* for underlying cause—shunt surgery; vagotomy and GJ; gastrectomy; TIPSS; splenectomy, etc.

COMPLICATIONS OF GASTRIC SURGERY

- Haemorrhage.
- Stomal obstruction.
- Biliary fistula.
- Injury to CBD.
- Duodenal blow out—on 4th postoperative day.
- Pancreatitis.
- Recurrent ulcer/stomal ulcer.
- Gastrojejunocolic fistula.
- Dumping syndrome (Postcibal syndrome).
- Nutritional disturbances, diarrhoea.
- Pulmonary tuberculosis.
- Carcinoma in gastric remnant (after 10-15 years).
- Gallstone formation.
- Alkaline gastritis.
- Afferent and efferent loop syndrome.
- Afferent loop obstruction—*common.*
- Efferent loop obstruction.

Complications may be classified as:

Intraoperative—bleeding, injury to spleen, pancreas, CBD.

Early postoperative—bleeding either intraluminal (from stomal site) or extraluminal; leak along anastomotic line; delayed opening of stoma; gastroparesis; duodenal stump blow out; lesser curve necrosis; fistula formation—gastric/duodenal/jejunal; omental infarction.

Late postoperative—reflux gastritis, recurrence of ulcer; dumping syndrome; malnutrition; gastrojejunocolic fistula; motility disorders; small stomach syndrome; afferent or efferent loop obstruction; retrograde jejunogastric intussusception; remnant carcinoma; gallstones; pulmonary or GIT tuberculosis; bezoar formation.

Stomal obstruction after gastric surgeries is due to:

- Mucosal oedema—usually subsides eventually
- Retrograde jejunogastric intussusception
- Hypertrophied stomal mucosa causing ball-valve mechanism
- Efferent loop obstruction
- Gastric atony eventhough there is wide patent stoma causing apparent stomal obstruction

Nutritional deficiencies following gastrectomy

❖ Iron deficiency anaemia	40%
❖ Megaloblastic anaemia (after 5 years)	25%
❖ (100 µg B_{12} is given IM weekly, later monthly)	
❖ Vitamin B deficiency	10%
❖ Calcium deficiency	40%
❖ Calcium deficiency with bone changes (after 5 years)	5%

A. DUODENAL BLOW OUT

- It is a very serious complication of Billroth II gastrectomy, occurs usually on 4-5th day after surgery.

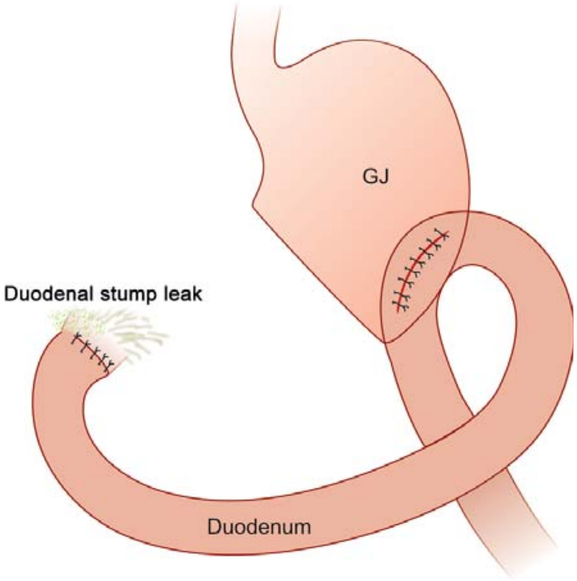

Fig. 20.54: Duodenal blow out after Billroth II anastomosis usually occurs 4 days after the surgery. It is often life-threatening condition.

- Here contents in the afferent loop are not having free flow which in turn increases the pressure in the duodenal 'C' loop leading to giving way of the closed duodenal stump.
- It is due to improper closure of duodenal stump, oedematous inflamed duodenum, afferent loop block, distal obstruction, ischaemia of least vascular duodenum and sepsis.

Features

- Sudden, severe pain abdomen postoperatively with features of shock.
- Forming a duodenal fistula through the drain placed.
- May cause peritonitis.
- Severe electrolyte imbalance.
- Features of biliary peritonitis, septicaemia.
- Skin excoriation and its problems.

Treatment

- *Conservative therapy* with nasogastric aspiration, IV fluids, TPN, antibiotics. Usually, patient will recover in 40-60 days.
- *Surgery is indicated* when there is peritonitis, fistula not responding, when there is distal obstruction.
- *Surgeries are*—afferent and efferent loop connection, relieving the obstruction distally, serosal patch over the duodenal stump after excision of the fistula, creating a controlled fistula after placing a tube into the duodenum.

Duodenal fistula

Causes

- Due to duodenal blow out—ischaemia, afferent loop obstruction, sepsis
- After closure of perforated DU
- Injury to duodenum by trauma, surgeries like hemicolectomy, renal surgeries, pancreatic surgeries, etc

Types

- High > 500 ml/24 hours
- Low < 500 ml

Features

- Pain abdomen
- Bile leak through the drain site or fistula track
- Skin excoriation
- Electrolyte imbalance
- Malnutrition, anaemia
- Recurrent infection

Investigations

- Electrolyte estimation, haematocrit
- CT fistulogram

Treatment

- TPN, feeding through jejunostomy
- Management of anaemia, electrolytes
- Blood transfusion
- Antibiotics whenever sepsis is suspected

What we need is cup of understanding, barrel of love and an ocean of patience.

- Care of skin—zinc oxide cream application, controlled fistula by placing a Foley's catheter into the fistula track
- Usually responds well by conservative therapy in 8-12 weeks
- If not improved then surgery is needed—anastomosis between afferent and efferent loops (jejuno-jejunostomy); serosal patch techniques; surgical closure of fistula, Roux-en-Y patch over fistula

B. RECURRENT ULCER

It is a wide terminology which includes stomal ulcers, recurrence of ulcer in the original ulcer site or different parts of stomach/duodenum or jejunum after a therapeutic surgical procedure.

Stomal ulcer can occur in gastrojejunostomy stoma or in gastroduodenostomy stoma. GJ stomal ulcer *per se* can occur after only gastrojejunostomy or after partial gastrectomy with Billroth II anastomosis. Recurrent ulcer at the original site is commonly referred to *ulcer recurrence* (after HSV commonly).

Recurrent ulcer is 3-7%. It may be after:

- Gastrojejunostomy.
- Billroth I or Billroth II anastomosis.

It may be:

- Gastrojejunostomy stomal ulcer
- Gastroduodenostomy stomal ulcer
- Jejunal ulcer
- Gastric ulcer
- *Ulcer recurrence*—10%

Causes

- Incomplete vagotomy.
- Zollinger-Ellison syndrome, gastrinomas.
- Alcohol, smoking.
- Hyperparathyroidism.

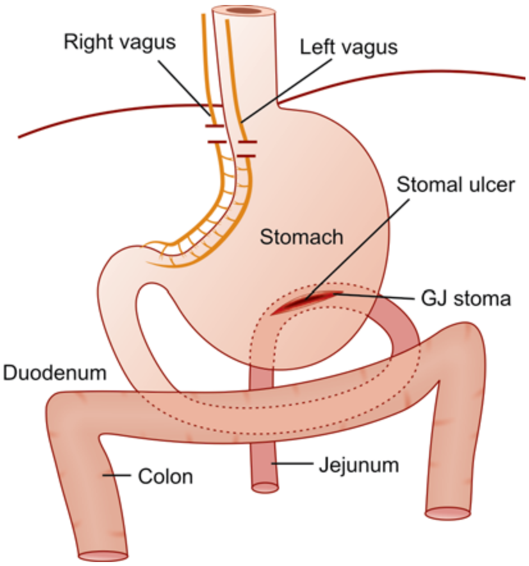

Fig. 20.55: *Stomal ulcer* after GJ may be due to incomplete vagotomy, Zollinger-Ellison syndrome.

- Gastrojejunostomy when done without vagotomy—40%.
- After HSV—10%.

Clinical Features

- Pain in the umbilical and left hypochondrium (severe, persistent).
- Back pain.
- Haematemesis.
- Anaemia, loss of weight.
- Features of obstruction—10%.

Note:

Stomal ulcer may complicate to gastrojejunocolic fistula, perforation, bleeding, obstruction, penetration.

Investigations

- Barium meal.
- Gastroscopy.
- Hollander's insulin test.
- Pentagastrin test.
- Gastrin level estimation.
- Serum calcium to rule hyperparathyroidism.
- *Visick grading* is used to assess the response to therapy to various procedures (Grade 1, 2, 3, 4 and 5).

Treatment

- Complete the vagotomy.
- Undoing of GJ.
- Conversion into another procedure, Roux-en-Y procedure.
- Treat hypergastrinaemia, hyperparathyroidism.
- Eradication of *H. pylori* infection using triple regime for 2 weeks and then continuation of therapy using PPI for 3 months.
- Subtotal gastrectomy in resistant recurrent ulcer cases.

Stomal ulcer

- It is the ulcer in anastomotic site of GJ/GD
- Overall incidence is 10% in 10 years after surgery
- It is common on jejunal side but can occur on gastric side or both sides or at junction

Causes

- After only gastrojejunostomy procedure—40%
- After vagotomy and GJ—5-7%
- After partial gastrectomy and gastroduodenostomy—Billroth I—3%
- After partial gastrectomy and GJ—Billroth II—3%
- Other causes like gastrinomas, hyperparathyroidism, incomplete vagotomy, smoking, alcohol

Features

- Pain and haematemesis/melaena is the main presentations
- Pain is persistent, severe over left upper abdomen often radiating to back
- Pain may be in the chest if it is antecolic anastomosis
- Weight loss and features of obstruction can occur

Investigations

- Acid tests (Hollander's), gastroscopy, tests specific for the cause
- Serum gastrin/calcium assay

Complications
- Bleeding stomal obstruction
- Perforation
- Gastrojejunocolic fistula
- Nutritional deficiency

Treatment
- Completion vagotomy
- Subtotal gastrectomy/revision gastrectomy
- Roux-en-Y anastomosis
- Treat the cause

C. DUMPING SYNDROME

(Post-cibal Syndrome)

It is common in *females*, seen after Billroth II surgery.

Types

1. *EARLY DUMPING SYNDROME:*
- It is common and more severe type
- Incidence is 10%
- Vasomotor symptoms appear immediately after food, lasts for 30-40 minutes, aggravated by bulky food. It is relieved by lying down, aggravated by more food

***Clinical features*:**
- Sweating, tachycardia, colicky pain and diarrhoea
- Hypotension and features of hypovolaemia

***Pathogenesis*:** It is a *primary disorder of carbohydrate metabolism* wherein initial transient hyperglycaemia prevents further absorption of glucose, which in turn draws fluid from the bowel wall by high osmolarity resulting in increased intestinal activity, resulting in diarrhoea and fall in blood volume

Treatment: Small, dry, more frequent food, with avoidance of carbohydrates. Octreotide 100 µg given subcutaneously before meals is found to be beneficial

***Surgical treatment*:**
- Conversion of Billroth II to Billroth I
- Interposition of reversed jejunal loop (*Henley's loop*)

Early dumping syndrome can last for long time (many years)

2. *LATE DUMPING SYNDROME*:
- It is of less severe type
- Incidence is 5%
- It usually occurs 2 hours after meal
- It is relieved by glucose and aggravated by exercise

Pathogenesis: Due to initial hyperglycaemia insulin secretion is stimulated which in turn leads to hypoglycaemia

***Clinical features*:**
- Tremor, fainting, nausea
- Features of hypoglycaemia

***Treatment*:** Symptoms are less severe and so treated conservatively, by giving glucose and food

D. ROUX STASIS SYNDROME

- It is occurrence of *gastric atony* after subtotal or partial gastrectomy if Roux-en gastrojejunostomy is done.
- They present with fullness, vomiting, loss of appetite and weight, early satiety.
- It occurs in 25% of Roux-en-Y GJ. It is common in *females*.
- It is common in Roux-en-Y limb if *more* than 40 cm. It is late complication seen in months or year.
- There may be dysfunction of both gastric remnant and Roux limb. It is due to *ectopic pacemaker* in *Roux limb* which delays the gastric emptying.

Treatment

- Completion total gastrectomy is the choice.
- Isolated jejunal loop interposition of 40 cm length—Henley's loop.
- Intestinal pacing, OR usage of '*uncut Roux*' limb.

E. GASTROJEJUNOCOLIC FISTULA

It is a complication of gastrojejunal ulcer (stomal ulcer), wherein ulcer penetrates and erodes into the transverse colon leading to gastrojejunocolic fistula, a dreaded complication.

Clinical Features

- *Sudden onset of severe diarrhoea* after every meal with past history of GJ done and past history of gastrojejunal ulcer.
- Foul gas eructation.
- Dehydration.
- Rapid loss of weight, anaemia, cachexia, hypoproteinaemia (bacteria from colon enters the jejunum disrupting the absorptive mechanism).
- Steatorrhoea.
- Rarely faecal vomiting.

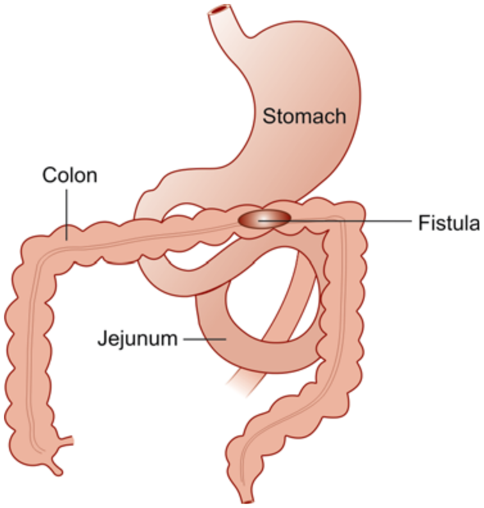

Fig. 20.56: Gastrojejunocolic fistula.

Investigation

Barium enema (not barium meal). CT scan, endoscopy (upper and lower).

Treatment

- IV fluids.
- Total parenteral nutrition (TPN).

Carcinoma stomach is the commonest cause of Krukenberg's tumour.

- Blood transfusions.
- Antibiotics.
- Rehydration, electrolyte management.
- Surgery: Resection of involved *stomach, jejunum, colon* has to be done and continuity has to be maintained (*Triple resection*). Surgery is the only treatment.

It has got high mortality.

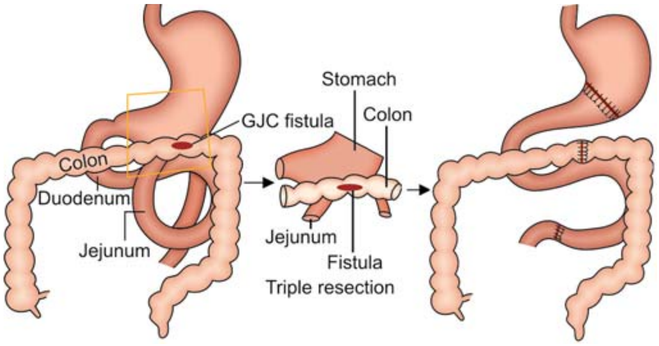

Fig. 20.57: Resection of the gastrojejunocolic fistula (removal of involved parts of the stomach, jejunum and colon—triple resection) with maintaining the continuity.

TRICHOBEZOAR (Rapunzel Syndrome)

- It is a hair-ball commonly seen in stomach of *females with psychiatric illness,* who swallow hair regularly.
- It forms a ball like mass occupying the full stomach.

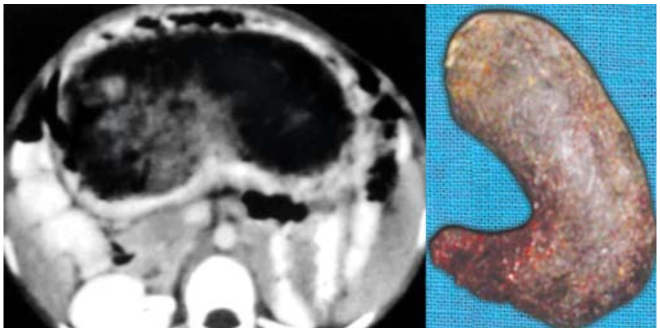

Fig. 20.58: Trichobezoar—CT and hair ball mass (*Courtesy*: Dr (Proffessor) BM Nayak, KMC, Mangalore)

Clinical Features

- Epigastric pain, early satiety.
- Haematemesis.
- Features of gastritis.
- Loss of appetite.
- Perforation.
- Epigastric mass.

Investigation

- Barium meal and endoscopy is confirmative.
- CT scan is very useful.

Treatment

- Gastrotomy and removal of hair ball. Duodenum and small bowel should be examined on table for additional bezoars in these sites.
- Psychiatric counselling.
- Enzymatic digestion and endoscopic breaking of the bezoars is also tried.
- It can also be removed by laparoscopy.

Phytobezoar

  - It is due to solid ball formation in the stomach comprising plant fibres and seeds.
  - Phytobezoar is common in gastric remnant after partial gastrectomy and in diabetics with autonomic neuropathy.
  - *Enzymatic digestion* can be tried using papain (found in Adolph's meat Tenderizer, as one tsp in 300 ml of water taken orally many times a day to fragment bezoar), cellulase substance. If fails, removal by gastrostomy.

CHRONIC DUODENAL ILEUS (WILKIE'S SYNDROME)

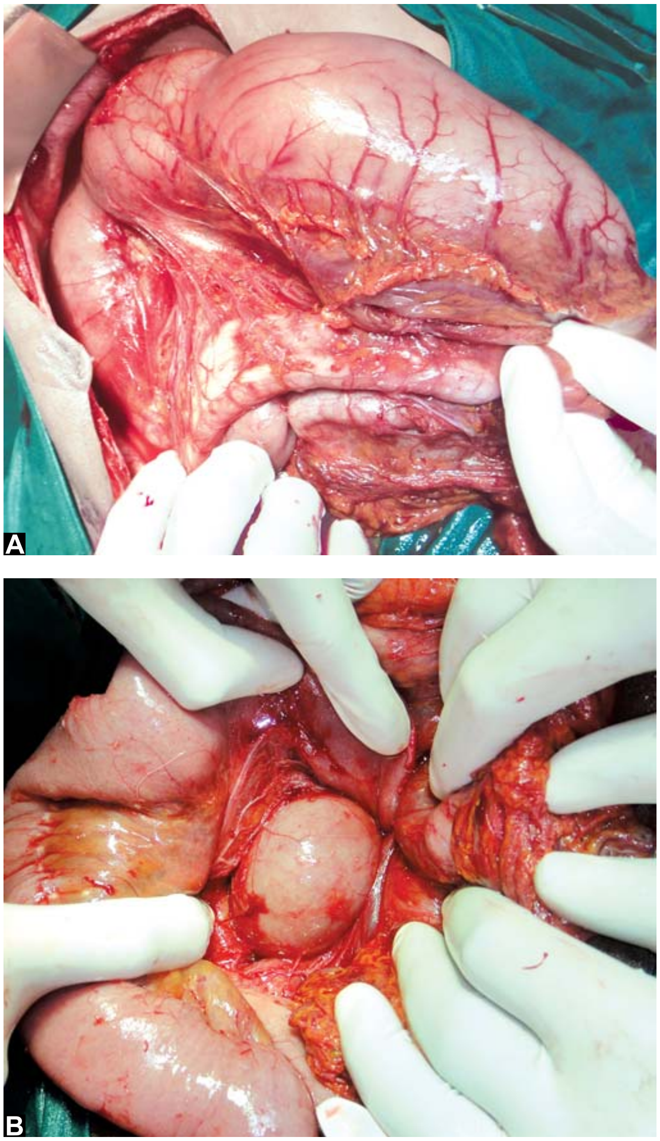

Figs 20.59A and B: Superior mesenteric artery—SMA (*Wilkies'*) syndrome showing site of obstruction and dilated proximal gastroduodenum.

- It is also called as *superior mesenteric artery (SMA) syndrome.*
- It is due to obstruction of the 3rd part of the duodenum due to decreased angle between SMA and aorta.
- It can be congenital or due to low insertion of SMA or high insertion of duodenal end or due to traumatic aneurysm of SMA.
- It can be aggravated by plaster casts, lordosis, pancreatic tumour, enlarged lymph node in the 3rd part of duodenum (*Cast syndrome).*

Clinical Features

- Bilious vomiting.
- Upper abdominal fullness.
- Visible peristalsis.
- Dehydration.

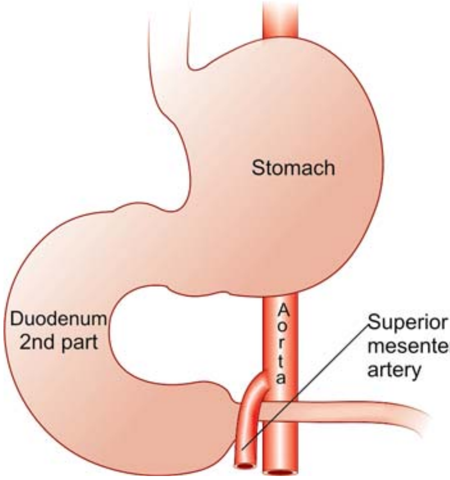

Fig. 20.60: Wilkie's syndrome.

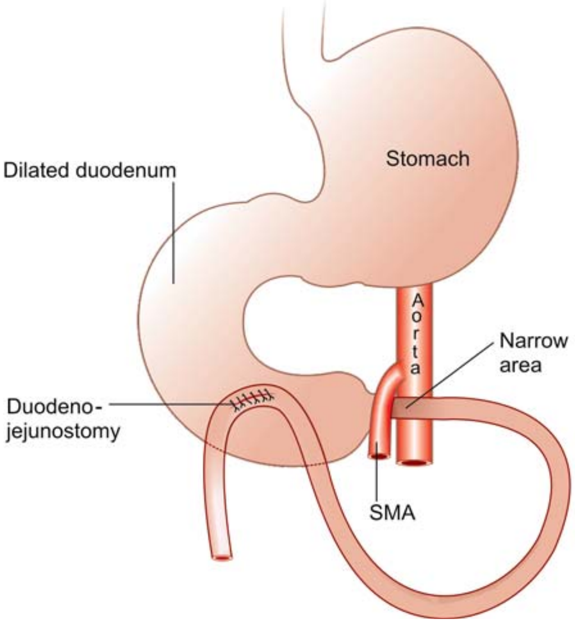

Fig. 20.61: Duodenojejunostomy, in case of chronic duodenal ileus.

Differential Diagnosis

- Pyloric stenosis with gastric outlet obstruction.
- Annular pancreas.

Investigation

- Barium meal is diagnostic.
- CT scan abdomen.

Treatment

- Duodenojejunostomy.

Note:

Gastrojejunostomy—not done.

DURBAN'S SYNDROME

- There is a connecting band between right and left crura which causes compression of celiac artery and SMA.
- It may cause foregut ischaemia due to inadequate collaterals, or midgut ischaemia due to diversion of blood to foregut.
- Presentation—chronic abdominal pain, murmur in epigastric region.
- It is often difficult to diagnosis. CT scan, CT angiogram may show poor perfusion. MRI may be useful.
- Treatment—cutting/excising the fibrous band through laparotomy or laparoscopy.
- Condition is also called as *median arcuate ligament syndrome* or *coeliac artery syndrome* or *coeliac band syndrome*.

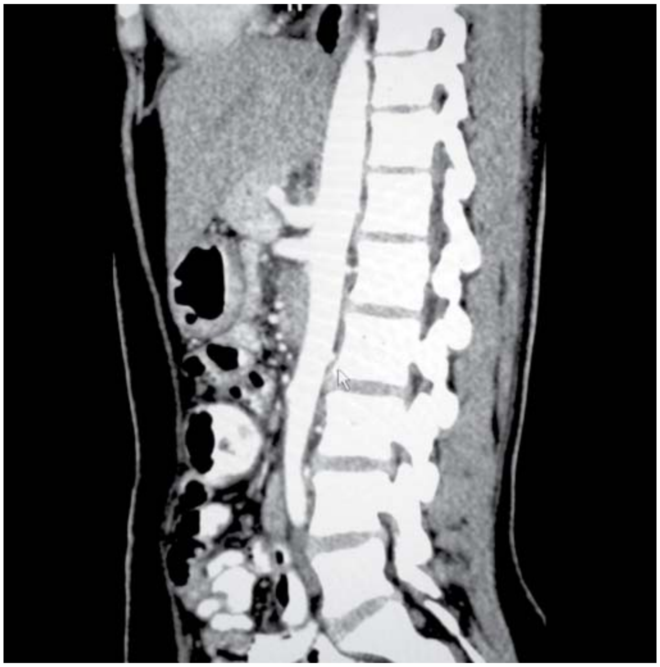

Fig. 20.62: CT scan showing coeliac artery syndrome.

Many very skillful operators are not good surgeons.— **William J Mayo**

ACUTE GASTRIC DILATATION

- It is an enormous acute dilatation of stomach with atonic gastric wall without peristalsis.
- Stomach distends enormously occupying most of the abdomen and pelvis causing sequestration of lot of fluid resulting in hypovolaemia.

Causes

- After major surgery (abdomen, neurosurgery).
- Trauma, burns.
- Retroperitoneal haematoma.
- Electrolyte imbalance.
- Other causes: Anorexia nervosa, bulimia, polyphagia, drug abuse, diabetes, anaesthesia, debilitating diseases, spinal cord diseases, muscle dystrophy.

Clinical Features

- Features of hypovolaemia and shock.

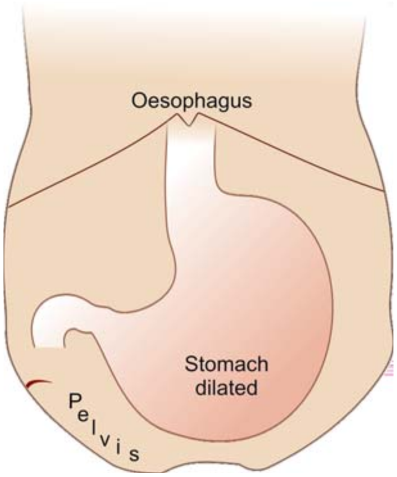

Fig. 20.63: Acute gastric dilatation showing dilated stomach occupying entire abdomen and also pelvis.

- Vomiting, hiccough. Vomits large quantity of brownish black fluid like *"the storm water of a peat-laden stream."* Vomitus when placed in a test tube and held in a strong light, myriads of small particles may be suspended in the fluid.
- Dilated stomach confirmed by ausculto percussion test.
- Positive succussion splash.
- Electrolyte imbalance is seen.

Investigations

Plain X-ray abdomen, serum electrolytes, U/S abdomen.

Treatment

- Conservative treatment is given initially.
  - Large amount of intravenous fluids.
  - Ryle's tube aspiration.
  - Electrolyte management, blood transfusion.
- The cause is treated.
- Rarely surgical decompression is required. Condition has got high mortality.

Complications

- Aspiration pneumonia *(Mendelson syndrome).*
- Severe hypovolaemia and electrolyte imbalance.
- Raised intragastric pressure causes *venous congestion*, venous infarction, necrosis and perforation.

GASTRIC VOLVULUS

It is twist in the axis of the stomach.

- Rotation occurs around the axis made by two fixed points—cardia and pylorus.
- It can be idiopathic or secondary to—hiatus hernia, left sided eventration, adhesions or pyloric obstruction with long-standing gastric dilatation.

Twist may be:

- ***Organo-axial***: Common in elderly—horizontal.
- ***Mesenterico-axial***: Common in children—vertical.

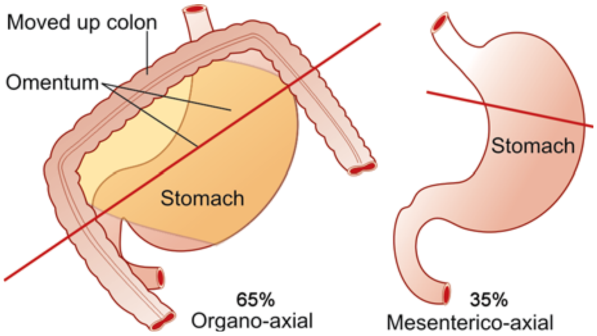

Fig. 20.64: Types of gastric volvulus.

Pathology

- Here stomach twists upwards between oesophagogastric junction and pyloroduodenal junction. Colon along with omentum also moves upwards initially.
- It is often associated with rolling hiatus hernia or diaphragmatic eventration.

Clinical Features

Borchardt's triad

- ❖ Acute epigastric pain
- ❖ Violent ineffective vomiting
- ❖ Inability to pass a nasogastric tube

Types

- Acute.
- Chronic recurrent. It is common type.

Complications

- Perforation.
- Gangrene of the stomach.
- Bleeding.

Investigations

- Plain X-ray abdomen in erect posture.
- Barium meal X-ray.
- CT scan abdomen.

Treatment

- Untwisting of the volvulus and gastropexy by fixing the anterior wall of the stomach to anterior abdominal wall.
- Gastrojejunal fixation (gastrojejunostomy without a stoma).
- Gastrectomy when stomach is gangrenous.
- Treating hiatus hernia or eventration.
- Displacing colon downwards by dividing gastrocolic omentum—*Tanner's operation.*

GASTRIC POLYP

Gastric polyps are observed in 3% of total gastroscopies. 45% of them are found in fundus which does not show any malignant potential. These polyps are often associated with FAP or Gardener's syndrome and colorectal neoplasms; but these gastric polyps perse are non-neoplastic.

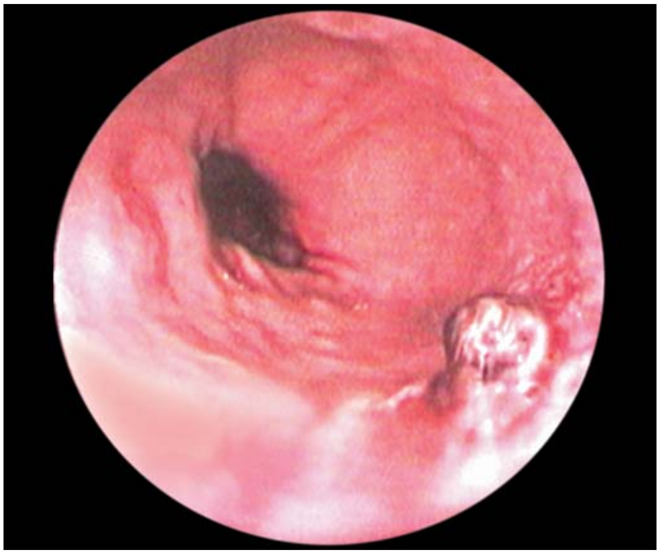

Fig. 20.65: Gastric polyp in prepyloric region.

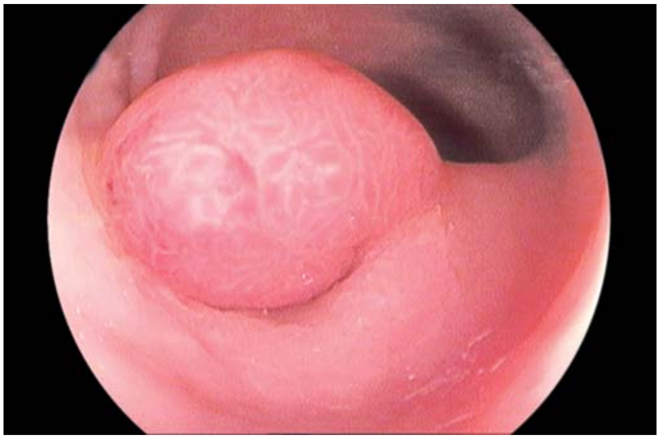

Fig. 20.66: Polyp in the duodenum first part.

Types

1. Hyperplastic polyp—75% common, minimal risk of malignancy (2%).
2. Adenomatous polyp (10%)—neoplastic in origin. Size > 2 cm is potentially malignant (25% chance).

Features

- *Hyperplastic polyp* is randomly distributed throughout the stomach. They are usually less than 2 cm in size, often multiple. Spontaneous regression can occur. Often it is associated with *H. pylori* infection.
- *Adenomatous polyp* is usually single. It is common in antrum. It can be sessile (common) or pedunculated. It is usually 2 cm or more (80%). It has got hyperchromatic nuclei arranged in a picker fence pattern. Size more than 2 cm has got high malignant potential.
- Inflammatory, hamartomatous and heterotrophic are other gastric polyps of negligible malignant potential.
- 50% are asymptomatic.
- Fundal polyps are associated with PPI therapy and familial polyposis.
- Gastric carcinoids arise from enterochromaffin like cells (ECL). They usually present as small polyps with pernicious anaemia.
- It may present with pain, haematemesis and gastric outlet obstruction. Gastroscopy is diagnostic.
- Endoscopy and resection is needed. If not possible open resection is done.
- Small polyp can be under regular surveillance by endoscopy.

MENETRIER'S DISEASE

- It is a condition with giant gastric mucosal folds (hypertrophic gastropathy) in the fundus and body of the stomach-*'cobblestone appearance'*.
- Histologically, it shows hyperplasia, mucosal thickening and gastric gland atrophy.
- Hypoalbuminaemia, anaemia and hypochlorhydria is common.
- Antrum is not involved.
- There is over expression of transforming growth factor alpha (TGFα) peptide which binds to epidermal growth factor (EGF).
- There is foveolar surface mucus cell hyperplasia with absence of *parietal cells* causing excessive mucus production, protein loss and achlor/hypochlorhydria.
- It is associated with CMV infection in children and *H. pylori* infection in adult.

Clinical Features

- Epigastric pain, anaemia and weight loss.

Charm strikes the sight, but merit wins the soul.

Investigations

- Serum total proteins, albumin.
- Gastric function tests.
- Barium meal studies, gastroscopy.

Menetrier's disease

- Giant folds in fundus and body
- Gastric gland atrophy
- Hypochlorhydria
- Hypoalbuminaemia
- It may turn into malignancy
- Antral sparing

Treatment

- *Total gastrectomy* is the treatment if there is massive protein loss or dysplasia or carcinoma.

Menetrier's disease may turn into malignancy.

DUODENAL DIVERTICULA

Types

1. **Primary:**
- Occurs in the 2nd or 3rd part of duodenum.
- Are congenital in origin, with all the three layers of duodenum.
- Are usually asymptomatic.
- Diagnosed by endoscopy.
- Treated if required—excision by endoscope or open method.

2. **Secondary:**
- Most common type.
- Always of acquired type.
- Occurs in the 1st part of duodenum.
- Due to cicatrised chronic duodenal ulcer.
- Barium meal X-ray shows '*trifoliate duodenum*'.
- Treatment of chronic duodenal ulcer with pyloric stenosis—truncal vagotomy with posterior GJ.

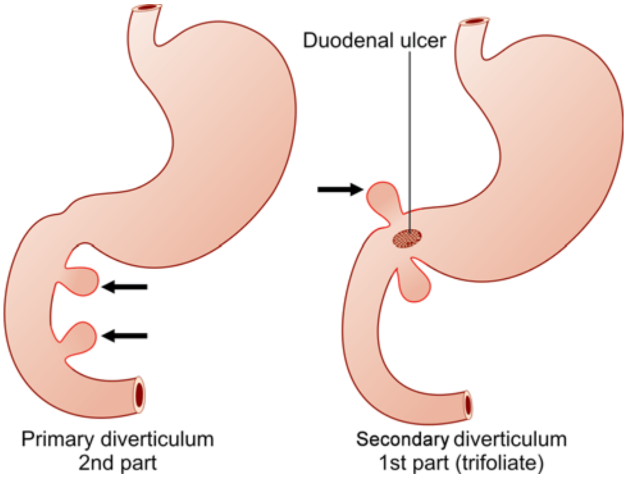

Fig. 20.67: Primary and secondary diverticula of the duodenum.

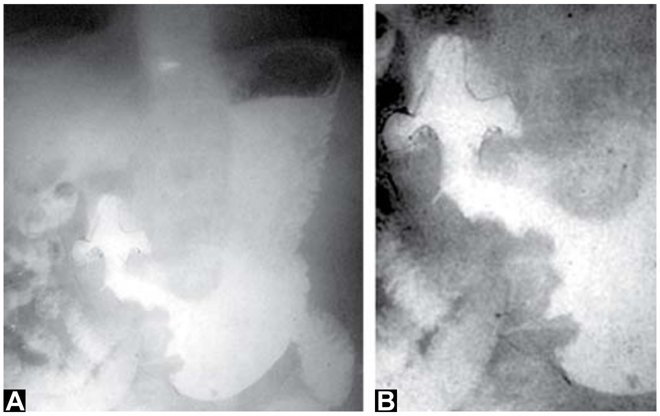

Figs 20.68A and B: Barium meal picture showing trifoliate duodenum as secondary diverticulum due to chronic duodenal ulcer.

CARCINOMA STOMACH

> *My kinsman, Antonio Bruno, retained the food he had eaten for too short a time, and then threw it up undigested.... His body wasted away through lack of nourishment till little more than skin and bone remained. At last he was brought to his death. The body was cut open for reasons of public welfare. It was found that the opening of his stomach had closed up and it had hardened...with the result that nothing could pass through to the organs beyond, and death inevitably followed.*
>
> —**Antonio Benivieni,** 1507

- '*It is the captain of men of death*'.
- It is more common *in Japan*—70 per 1,00,000 population.
- It is more common in males 2:1.
- Decrease incidence in western world (Western Europe and US)—last four decades. But this decrease is confined to distal gastric cancers. Incidence of proximal gastric cancer is increasing. Carcinoma proximal stomach is not associated with *H. pylori* infection unlike cancers of body and distal stomach.

Risk Factors (Aetiology)

- *Familial*—10%. Napolean and many members of his family died of carcinoma stomach. *Familial gastric cancer* in associated with mutation of *e-cadherin gene* (90% risk). It causes hereditary diffuse gastric cancer. Relatives of such family show mutation of this gene. Whether prophylactic total gastrectomy is needed or not in these high-risk individuals is a debate.
- *Inactivation of p53, over expression of growth factors, bcl-2 gene mutations* are other genetic causes.
- *HNPCC, Li-Fraumen syndrome.*
- Gastric mucosa of people with *blood group 'A'* is more susceptible for carcinogens—diffuse type. It is due to different mucopolysaccharide secretion in stomach of blood group A patients who are more susceptible for carcinogens.
- *Gastric polyps, adenomatous polyp* > 2 cm.
- *Pernicious anaemia*—high-risk 6 times.
- *Gastric remnant*—15 years after gastrectomy and GJ.

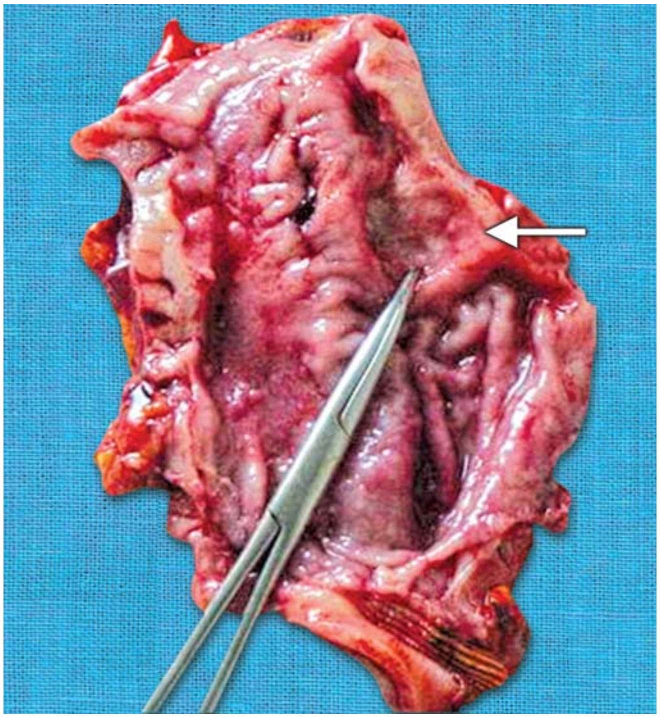

Fig. 20.69: Gastrectomy specimen showing ulcerative gastric carcinoma.

- *Diet*—high salt diet, food with more nitrosamines increases the risk. Smoked salmon fish increases the risk. They release polycyclic hydrocarbons. Ingested nitrates and nitrites from preserved food are converted to nitrosamines by GI bacteria.
- Fruits and vegetables rich in vitamin 'C' protect from carcinoma stomach.
- *Chronic gastritis (atrophic, autoimmune)*, intestinal metaplasia. Type A causes proximal gastric cancer. Type B causes distal gastric cancer.
- *Gastric dysplasia*: Chronic gastritis → gastric atrophy → intestinal metaplasia → dysplasia → carcinoma *in situ* → carcinoma—this cycle occurs in body and distal stomach and is often called as *Correa cycle*.
- Smoking, alcohol.
- *Helicobacter pylori infection*—high-risk (Cag A strain) 6 fold increase in incidence. It causes intestinal type of gastric cancer.
- Agammaglobulinaemia—high-risk (2-5%).
- Chronic benign gastric ulcer. Carcinoma arising from benign gastric ulcer is called as *ulcer cance*r of stomach.
- Giant hyperplasia of gastric mucosal folds (*Menetrier's disease).*
- Carcinoma in proximal stomach is common in young and upper socioeconomic group.
- Carcinoma in distal stomach is common in old and lower socioeconomic groups.
- In western countries, carcinoma stomach is more common in proximal, near OG junction. Obesity, young individual, white people, smoking, alcohol intake, gastro-oesophageal reflux, higher social status, high calorie diet and probably genetic factors are the causes for proximal gastric cancers. It is more aggressive, spreads early due to thin muscularis mucosa. It is often diagnosed late. Signet ring type is common. It carries poor prognosis. It needs oesophageal resection.
- In Asian countries, it is still common in distal stomach.
- In India, it is common in south India (4 times than North India).
- Epstein-Barr virus, radiation exposure.
- Occupational—Rubber workers, coal workers.
- Zinc, lead, talc, asbestos all can cause carcinoma stomach.

Precursor Lesions of Carcinoma Stomach

- *Chronic atrophic gastritis*: It is the most common precursor lesion mainly intestinal subtype. In Japan 95% of atrophic gastritis develop early gastric cancer. Incidence is higher in elderly and in those associated with *H. pylori* infection.
- *Adenomatous gastric polyps.*
- *Intestinal metaplasia*: Risk of carcinoma depends on extent of metaplasia in mucosa. *H. pylori* eradication is important here.

  Based on histological and biochemical nature, two types are found:
  - *Complete*: Glands are completely lined with goblet cells and intestinal absorptive cells indistinguishable histologically and biochemically from their small bowel counterparts.
  - *Incomplete*: It contains columnar cells, goblet cells but without intestinal absorptive cells.

It also can be:
  - *Type I*—Mature; goblet cells secret sialomucin.
  - *Type II*—Cells in different levels of dedifferentiation. Cells secrete sialomucin and an abnormal sialomucin (sulphomucin)—a small quantity.
  - *Type III*—Marked dedifferentiation of cells, secreting mainly sulphomucin.
- *Menetrier's disease.*
- *Benign gastric ulcer*: Here risk is 2-5%. But it is related to the size, extent and duration of the benign gastric ulcer. Giant gastric ulcer has got as high as 6–23% risk. Cancer developing in a pre-existing benign gastric ulcer is called as *ulcer cancer.*
- *Stomach remnant (stump carcinoma)*: It can occur after Billroth II gastrojejunostomy (common) or vagotomy GJ. It takes around 15 years or more to develop cancer. Common site is close to the stoma. Atrophic gastritis—metaplasia—dysplasia and carcinoma develops. Altered acid level, duodenogastric bile refluxes are the pathogenesis which increases the peptic activity causing mucosal metaplasia and dysplasia. Earlier history of gastric surgery, recent history of loss of appetite and decreased weight with often a palpable mass are the features. Liver secondaries, ascites may develop in late cases. Gastroscopy with biopsy and CT scan confirms the diagnosis. Treatment is gastrectomy with nodal clearance.

Aetiology for gastric cancer

Environmental/ occupational/diet/habits

- Smoking/alcohol/ obesity
- Low vegetables, diet with low vitamin A and C
- Consuming red meat, smoked salmon fish, cabbage, diet rich in nitrosamines, lead
- Viral infections like EB virus
- Occupational—rubber/ coal workers
- Lower social status—distal cancers
- Higher social status—proximal cancers

Precancerous lesions:

- *H. pylori* infection, chronic gastritis
- Pernicious anaemia
- Intestinal metaplasia
- Adenomatous polyps more than 2 cm
- Agammaglobulinaemia
- Benign gastric ulcer
- Previous gastric surgery
- Menetrier's disease

Genetic and familial

- *E-cadherin* gene mutation in diffuse cancers
- Mutation in APC gene for β-catenin in intestinal cancers—50%
- Inactivation tumour suppressor gene p53 in 30% cases
- Loss of heterozygosity in the BCL2 gene, an inhibitor of apoptosis (intestinal type)
- Amplification/over expression of different growth factor receptors
- HNPCC—carries 5-10% risk of gastric cancer
- Li Fraumen syndrome
- Blood group A
- First degree relative carries 3-6 folds increased risk of gastric cancer.
- Monozygotic twins carry more risk than dizygotic
- When both parents have gastric cancer, the siblings are at risk of diffuse proximal gastric cancer
- FAP—10 folds more risk of gastric cancer
- Mutation of H-ras oncogene and over expression of c-erb B2 gene

Pathology

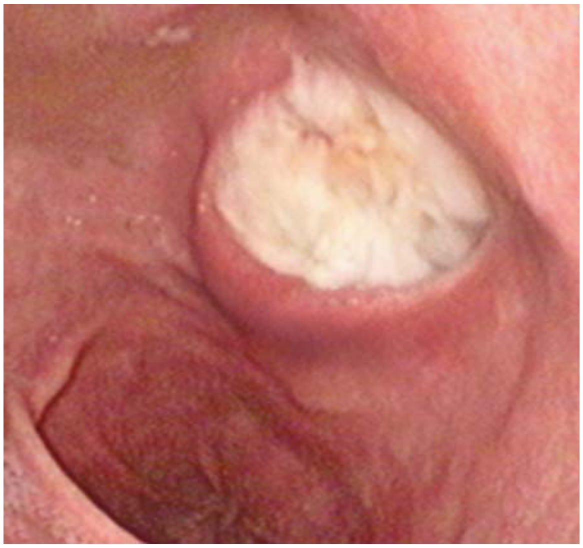

Fig. 20.70: Endoscopic view of carcinoma of stomach.

I. ***Gross types***:

Cauliflower type
Ulcerative type
Leather-bottle (Linitis plastica)

II. ***Lauren's classification*** (DIO classification):

a. *Intestinal type (53%)*:
- Has got favourable prognosis.
- Polypoid and superficial types are intestinal varieties—common in *H. pylori* induced. Gland formation and definite cellular architecture are the features.
- Synchronous/polypoid/superficial types are common.
- Gastric mucosa is replaced with epithelium that resembles small bowel mucosa.
- It is common in men and aged individuals, probably has environmental factors, commonly shows haematogenous spread, *microsatellite instability*, APC gene mutations, p53, p16 inactivation.

b. *Diffuse type (33%)*:
- It has got poor prognosis. Common in blood group A, familial type, young people and females.
- It is usually poorly differentiated/signet type with early gastric wall penetration and both submucosal and subserosal lymphatic spread.
- Linitis plastica, ulcerative growth without glandular formation is common in this type.
- It shows decreased E-*cadherin* with p53, p16 inactivation.

c. *Others*—Unclassified (14%).

III. ***Depending on the depth of invasion***:

a. *Early gastric cancer* is defined as involvement of *mucosa and/or submucosa* only with or without involvement of lymph nodes – T1 + any N.

It is being classified as:

(Japanese's classification)

1. Protruded.
2. Superficial—elevated (a), flat (b), depressed (c).
3. Excavated.

It is:
- More common in Japan—50% of gastric cancers treated are early gastric cancers. In USA—20%.
- 10% of early gastric cancers will have nodal metastases. 3% in only mucosal lesions. 20% in submucosal lesions.
- Nodal spread in early gastric cancer depends on tumour size (> 2 cm) and differentiation.
- About 70% are well-differentiated.
- Endoscopic mucosal resection (EMR) is possible in cancer of only mucosal lesions.
- Overall cure rate with adequate gastric resection and lymphadenectomy—95%.

b. *Advanced gastric cancer* is defined as involvement of *muscularis and/or serosa* with or without involvement of lymph nodes.

It is classified as:

(Borrmann's classification):

I. Single, polypoid carcinoma.
II. Ulcerated carcinoma with clear cut margin.
III. Ulcerated carcinoma without clear cut margin.
IV. Diffuse carcinoma—linitis plastica.
V. Unclassified (See Fig. 20.71).

IV. ***Ming's classification***:

- Expanding—favourable prognosis.
- Infiltrative—unfavourable, poor prognosis.

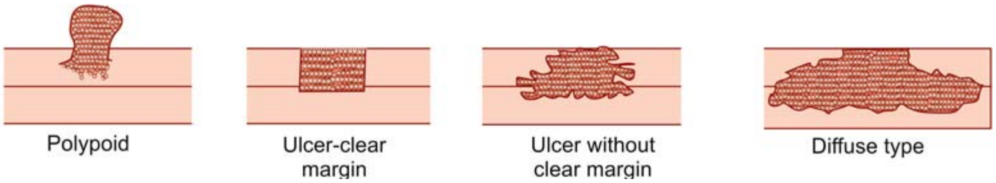
Polypoid
Ulcer-clear margin
Ulcer without clear margin
Diffuse type

Fig. 20.71: Borrmann's classification of the advanced gastric cancer.

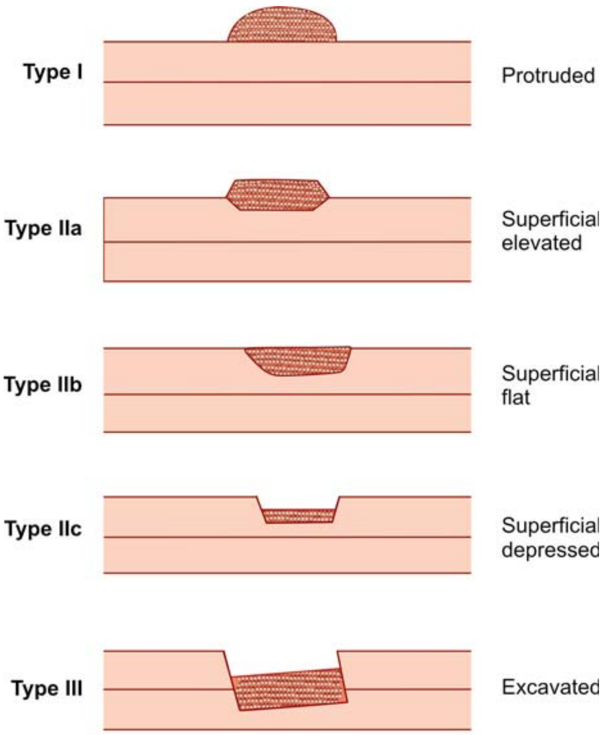
Type I
Protruded
Type IIa
Superficial elevated
Type IIb
Superficial flat
Type IIc
Superficial depressed
Type III
Excavated

Fig. 20.72: Japanese's classification for early gastric cancer.

V. ***WHO histological classification of gastric cancer (microscopic)***

- Adenocarcinoma—most common.
  - Papillary adenocarcinoma.
  - Tubular adenocarcinoma.
  - Mucinous adenocarcinoma.
  - Signet-ring carcinoma
- Adenosquamous carcinoma.
- Squamous cell carcinoma—rare, when occurs it is common near OG junction.
- Undifferentiated carcinoma.
- Others—unclassified carcinoma.

VI. ***Proximal gastric adenocarcinoma (Siewert classification)***

Type 1: Carcinoma in Barrett's oesophagus/true oesophageal carcinoma extending to GE junction. Here total oesophagectomy with gastric pull through to neck is needed.

Type 2: Tumour within 2 cm of squamocolumnar junction. Here total gastrectomy with Roux-en-Y oesophagojejunostomy is needed.

Type 3: Tumour in subcardial region. Here total gastrectomy with Roux-en-Y oesophagojejunal anastomosis is done.

VII. ***Morphovolumetric classification:***

It is based on ratio of invasion into muscle to mucosa in advanced carcinoma as:

- *Funnel type*—mucosa involvement is more compared to muscle with a ratio of < 0.15. It carries better prognosis.
- *Columnar type*—equal involvement in ratio of 0.75-1.25.
- *Mountain type*—muscle invasion is more with ratio < 1.25.

VIII. ***Morson and Dawson classification:***

- Nodular.
- Ulcerated.
- Fungating.
- Linitis plastica.

IX: ***Kajitani classification:***

- Localised.
- Intermediate.
- Infiltrating.

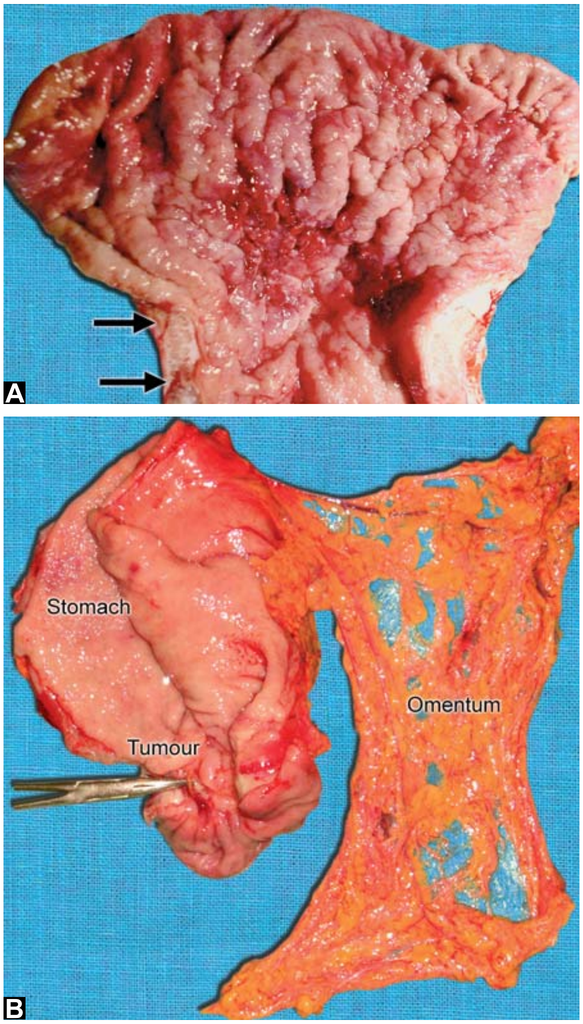
A
Stomach
Omentum
Tumour
B

Figs 20.73A and B: Specimen of stomach showing thickening of wall of part of the stomach with loss of rugosity. Omentum, proximal parts are also seen. Thickening is extending into serosa.

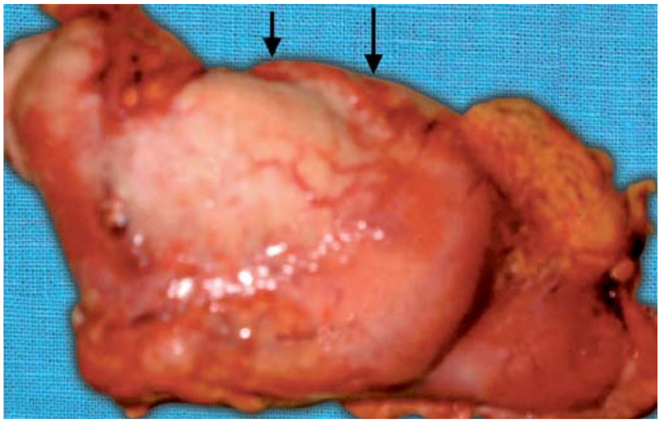

Fig. 20.74: Outer look of carcinoma stomach extensively involving including serosa.

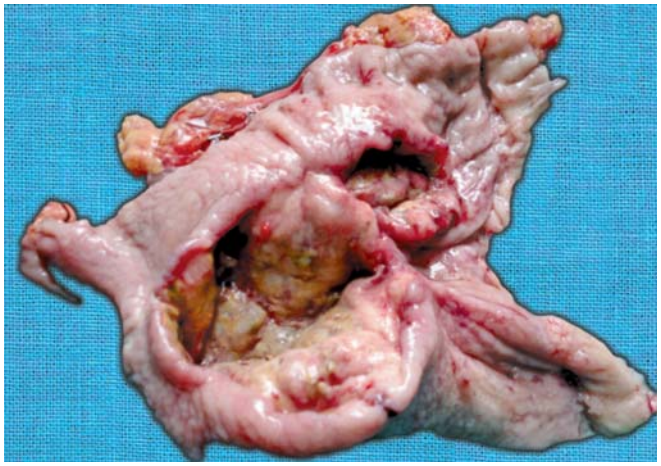

Fig. 20.75: Specimen of stomach showing large ulcerative deep lesion in the stomach with raised and everted edge. Loss of rugosity also observed.

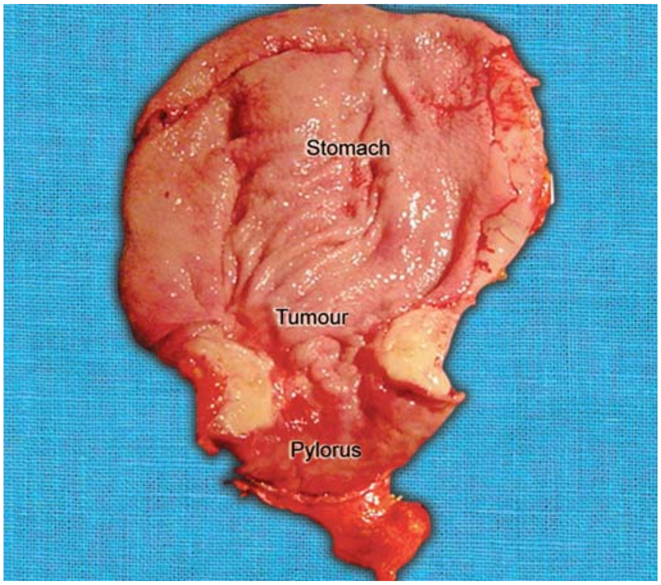

Fig. 20.76: Specimen of stomach showing thickening of pylorus—feature of carcinoma stomach. In Asian countries, pylorus is the most common site. In Western countries proximal stomach is the most common site.

Common Site of Occurrence

- Prepyloric and pyloric region (65%) (most common site)
- Body (25%).
- Fundus, OG junction.

But now the incidence of growth in the upper part, near oesophagogastric (OG) junction is increasing.

Note:

- *Synchronous gastric cancer* (Moertel) is histologically proved malignant conditions separated by normal stomach wall. Local extension and secondaries should be ruled out.
- *Distal gastric cancers* are diet related, environmental, epidemic, intestinal type, have better prognosis, with 30% respectability, treated with distal subtotal gastrectomy.
- *Proximal gastric cancers* are not diet related, not environmental, endemic, diffuse type, carries poor prognosis, with less than 20% respectability, treated with total gastrectomy.

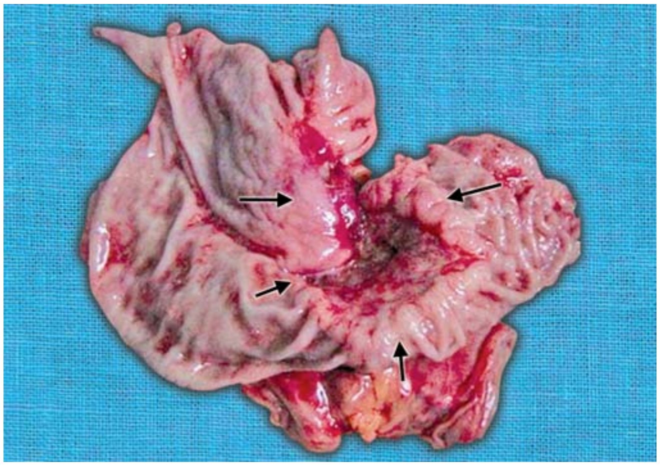

Fig. 20.77: Ulcerative lesion with everted edge in the gastrojejunostomy stoma. It usually occurs only after 15 years of gastrojejunostomy surgery.

Spread (Biological Behaviour)

- ***Direct spread***:
  - *Horizontal submucosal* spread along stomach wall.
  - *Vertical spread by invasion* across to adjacent structures like—pancreas, colon, mesocolon, liver.

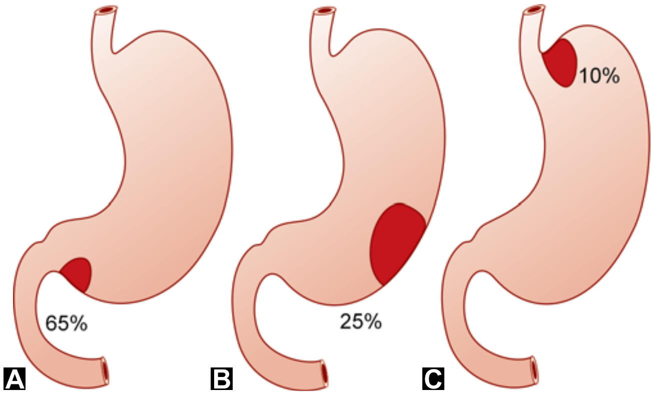

Figs 20.78A to C: Common sites of carcinoma stomach, (A) Pylorus, (B) Body, (C) Fundus.

- ***Lymphatic spread:***
  - Spread occurs by permeation and embolisation through lymphatics to subpyloric, gastric, pancreaticoduodenal,

splenic, coeliac, aortic, and later to left supraclavicular lymph nodes (Virchow's lymph node—*Troisier's sign).*

  - Retrograde spread to mesenteric nodes of small and large bowel can occur which signifies poor prognosis.

Zones of lymphatic drainage in stomach: Four zones.

Zone 1 lies in gastrocolic omentum along the right gastroepiploic vessels, draining pyloric portion of the greater curve to pyloric, coeliac and aortic lymph nodes.

Zone 2 lies in gastrocolic and gastrosplenic omentum along the left gastroepiploic vessels draining from upper half of greater curve to pancreaticosplenic and aortic lymph nodes.

Zone 3 is drainage from proximal two thirds of the stomach and the upper lesser curve along the left gastric artery. Zone 3 drains into perioesophageal lymph nodes.

Zone 4 is from distal portion of lesser curve and pylorus along hepatic artery into para-aortic nodes.

Lymph node groups

- *Group I:* Perigastric nodes
- *Group II:* Along the root of major vessels
- *Group III:* At the root of superior mesenteric artery and hepatoduodenal ligament
- *Group IV:* Distant lymph nodes

- ***Blood spread***
  - It occurs to liver (most common) causing multiple liver secondaries presenting as multiple, hard, nodules with umbilications due to central necrosis.

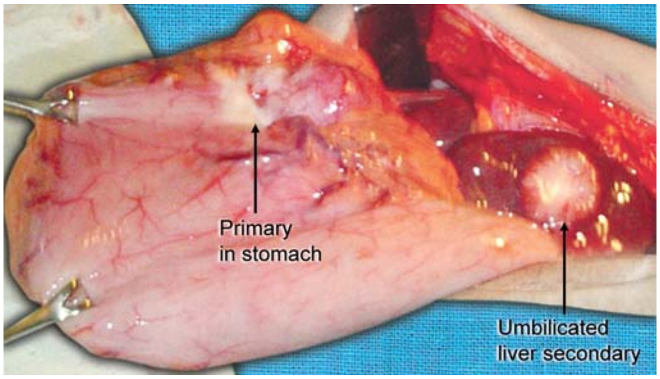

Fig. 20.79: Carcinoma stomach showing umbilicated secondaries in liver.

  - Later lungs and bones can get involved.
- ***Transperitoneal spread***
  - It can cause peritoneal seedlings (leading to ascites) and also can cause *Krukenberg's tumours in ovary* in menstruating age group (Commonly seen in colloid carcinoma of stomach).
  - Krukenberg's tumour can often occur without ascites. This is explained by possible *retrograde lymphatic spread* from stomach to ovary. This is favoured by absence of ascites, no denudation/implantation/adhesions on the ovarian surface. Through lymphatic spread, even early gastric cancer can cause *Krukenberg's tumour*. In Japan, 17% of ovarian tumours is Krukenberg's secondaries from carcinoma of stomach. Occasionally Krukenberg's tumour can arise from carcinoma breast (lobular), other abdominal organs also.
  - Rectal secondaries *(Blumer shelf), Sister Mary Joseph* umbilical secondaries are through transperitoneal spread.
  - Transperitoneal spread is best identified through laparoscopy and confirmed by laparoscopic biopsy.

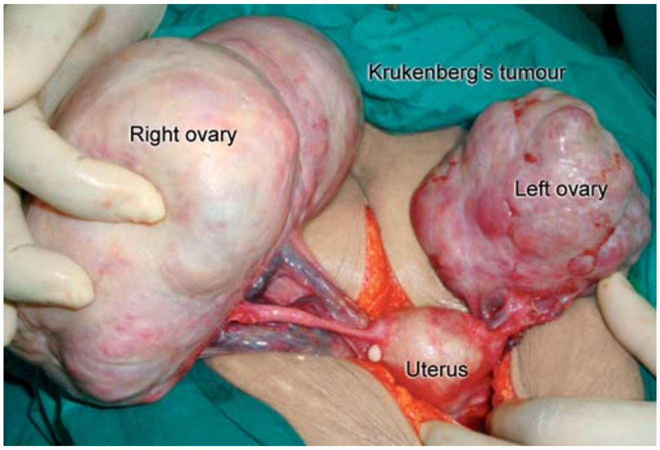

Fig. 20.80: Krukenberg's tumour involving both ovaries from carcinoma of stomach. It is due to transcoelomic spread or through peritoneal lymphatics or blood spread. It usually occurs during ovulation period.

TNM Staging of Carcinoma Stomach

TNM Staging, 7th edition, 2010

Tumour – T

Tx—tumour cannot be assessed
T0—no evidence of tumour
Tis—carcinoma *in situ*, intraepithelial tumour without invasion
T1a—invades lamina propria or muscularis mucosa
T1b—invades submucosa
T2—invades muscularis propria
T3—invades subserosal connective tissue without invasion of visceral peritoneum or adjacent structures
T4a—invades serosa/visceral peritoneum
T4b—invades adjacent structures

Nodes – N

Nx—cannot be assessed
N0—no nodal spread
N1—1-2 regional nodes
N2—3-6 regional nodes
N3a—7-15 regional nodes
N3b—16 or more nodes

Metastases – M

M0—no distant spread
M1—distant spread present

Staging

0—T_{is} N_0 M_0
IA—T_1 N_0 M_0

His heart cannot be pure whose tongue is not clean.

IB—$T_1\ N_0\ M_0$ / $T_2\ N_0\ M_0$
IIA—$T_1\ N_2\ M_0$;$T_2\ N_1\ M_0$;$T_3\ N_0\ M_0$
IIB—$T_1\ N_3\ M_0$;$T_2\ N_2\ M_0$;$T_3\ N_1\ M_0$;$T_{4a}\ N_0\ M_0$
IIIA—$T_2\ N_3\ M_0$;$T_3\ N_2\ M_0$;$T_{4a}\ N_1\ M_0$
IIIB—$T_3\ N_3\ M_0$;$T_{4a}\ N_2\ M_0$;$T_{4b}\ N_{0\text{-}1}\ M_0$
IIIC—$T_{4a}\ N_3\ M_0$;T, $N_{2\text{-}3}\ M_0$
IV—Any T Any N M_1

Stage I, II,II—*locoregional disease*
Stage IV—*systemic disease*

Japanese classification of gastric carcinoma (JCGC)

It includes following *principles*:

- Basic rules for clinical, surgical, pathological, and final features
- Specific rules for histology
- Gastric biopsy specimen group classification
- Chemo RT response

Nodal spread as per JCGC:

N_x—nodes unknown/cannot be assessed
N_0—no nodal spreads
N_1—Group I nodes only involved
N_2—Group II nodes are involved but not group III
N_3—Group III nodes are involved

Birmingham staging

1—Confined to mucosa/muscularis propria
2—Muscularis/serosal involvement
3—Nodal spread
4a—Residual disease; 4b—metastatic disease

Leather-Bottle Stomach (Linitis Plastica) (7-10%)

- It is an aggressive diffuse type of carcinoma stomach wherein there is enormous proliferation of fibrous tissue involving submucosa of stomach which is thickened, (*Mother of pearl appearance*) but mucosa looks and feels normal.
- It is type IV gastric carcinoma.
- It is *poorly differentiated* type *lacking* glandular formation, showing clusters of small uniform cells often with *signet ring* cells.

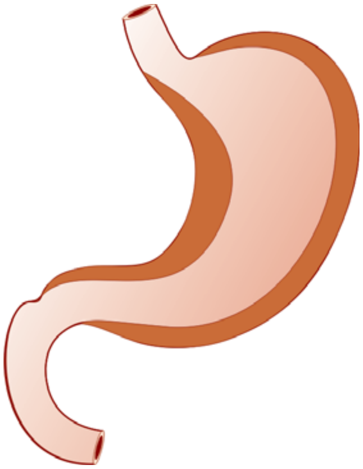

Fig. 20.81: Linitis plastica.

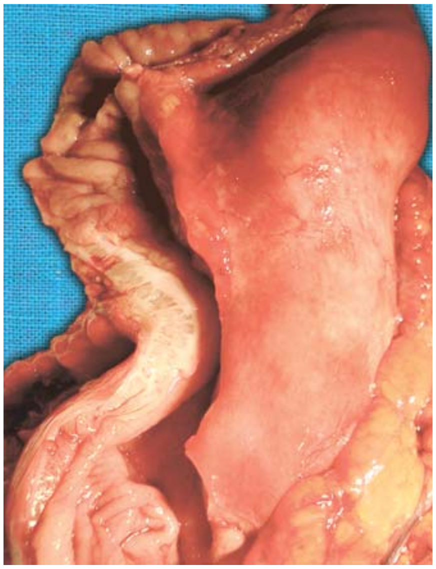

Fig. 20.82: Specimen showing features of *linitis plastica*—a diffuse type of carcinoma stomach mainly involving submucosa and deeper layer (mother of pearl appearance). It carries poor prognosis. It is type IV gastric carcinoma. It may be generalised or localised. It is also called as *leather-bottle stomach*.

- It spreads *submucosally* without much inflammatory cells.
- Transmural, lymphatic, intraperitoneal spread and early metastasis is common.
- *Common* in young, females, blood group A and is familial.
- *Clinical features:*
    - Loss of appetite and weight. Later mass may be palpable from the body of the stomach.
- *Investigation:*
    - Small sized stomach in barium meal study.
    - Endosonography is also diagnostic.
- *Treatment:*
    - Total gastrectomy with oesophagojejunal anastomosis.
    - Prognosis is poor.
    - It may be occasionally localised linitis plastica in pyloric antrum which is better operable.

Colloid carcinoma of stomach

- It is 6% common among all gastric cancers, wherein tumour cells contain *colloid* which is histologically typical
- It also signifies *poorer prognosis*
- Common in women

Presentations of Carcinoma Stomach

- *Asymptomatic* in early gastric cancer or often in cancer of body of stomach.
- *Nonspecific symptoms*—indigestion/vague epigastric discomfort, constant nonradiating pain which is not related to food intake.
- *Specific* symptoms depend on the site of tumour—obstruction, dysphagia, mass, etc.
- *Metastatic disease*—liver secondaries, ascites, secondaries in ovary, rectovesical pouch, umbilicus, supraclavicular nodes, lung and bone secondaries.
- *Unusual presentations*—acanthosis nigricans, Irish nodes in the axilla.

Clinical Features

- Recent onset of loss of appetite and weight, early satiety, fatigue. *Microcytic, hypochromic anaemia* (iron deficiency) is common (40%).
- Upper abdominal pain.
- Vomiting with features of gastric outlet obstruction, i.e. [VGP +ve, +ve ausculto-percussion test, +ve succussion splash (to be checked with 4-6 hours empty stomach)].
- *Mass abdomen:* Mass in pylorus lies above the umbilicus, nodular, hard, with impaired resonance, mobile, moves with respiration, all border well made out.
- Dysphagia with mass in upper epigastrium.
- When it arises from the body of stomach, it may present as only mass abdomen.
- Along with jaundice, liver may be palpable with secondaries which are hard, nodular (50%) with umbilication.
- Ascites.
- +ve Troisier's sign.
- +ve rectovesical secondaries. (Blumer shelf) on per rectal examination.
- *+ve Trousseau sign*—migrating thrombophlebitis, also seen in carcinoma pancreas.
- Anaemia, cachexia.
- Haematemesis (15%), melaena.
- Occasionally carcinoma stomach can present as *perforation* to begin with (4%).
- Rarely as secondaries in the liver with silent primary in stomach.
- Secondaries in umbilicus, as Sister Joseph's nodules (spread through ligamentum teres).
- Cutaneous secondaries.
- Krukenberg tumours.

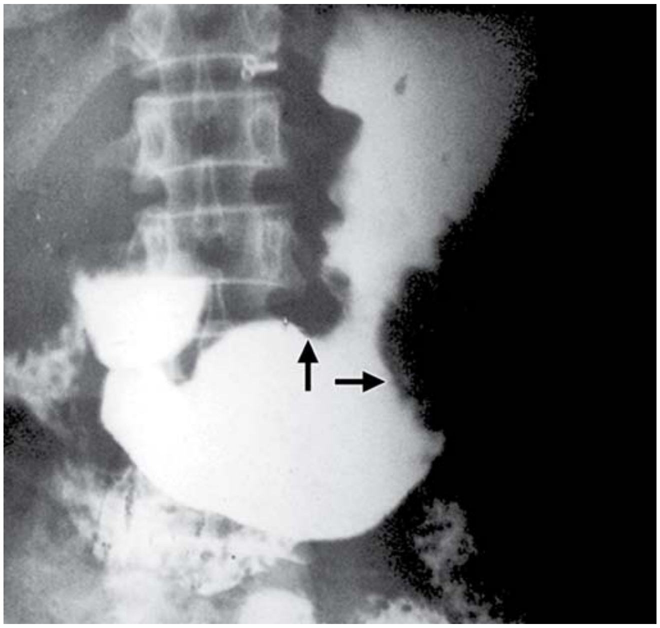

Fig. 20.83: Barium meal showing irregular filling defect in the body of the stomach suggestive of carcinoma stomach.

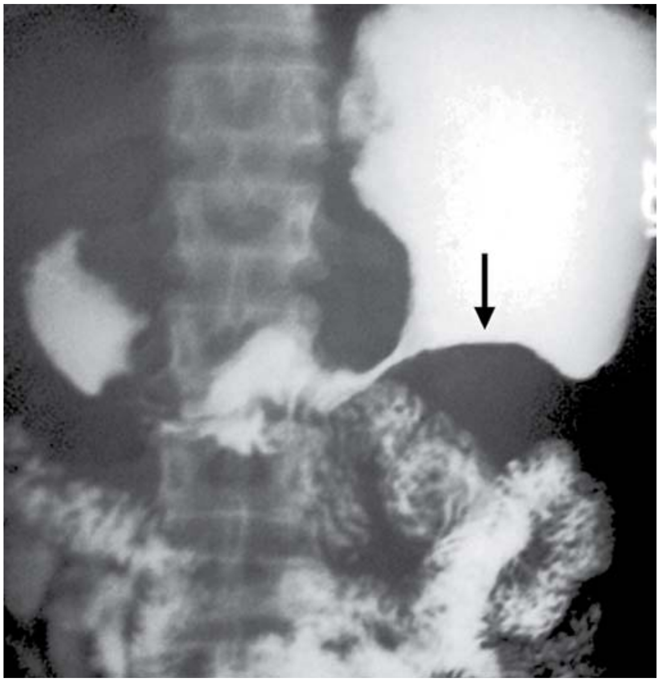

Fig. 20.84: Barium meal showing irregular filling defect in the pylorus suggestive of carcinoma pylorus.

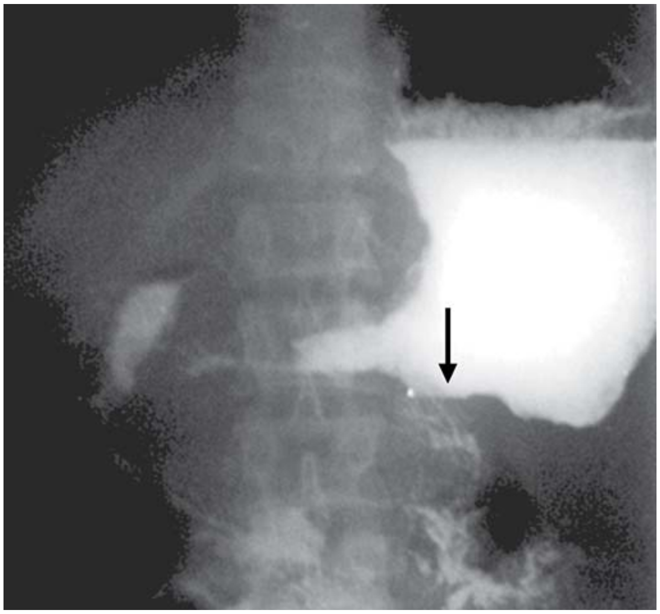

Fig. 20.85: Barium meal showing irregular filling defect in the body of the stomach due to carcinoma.

Differential Diagnosis

- Acid peptic disease; pyloric stenosis with gastric outlet obstruction.
- Gastritis.
- Pancreatic mass—carcinoma.
- Transverse colon mass—carcinoma.
- Advanced fixed stomach mass may mimic retroperitoneal or nodal mass.

Without human being this world would have been heaven. Production of noble human being was the Perfect Tragedy to this world !!

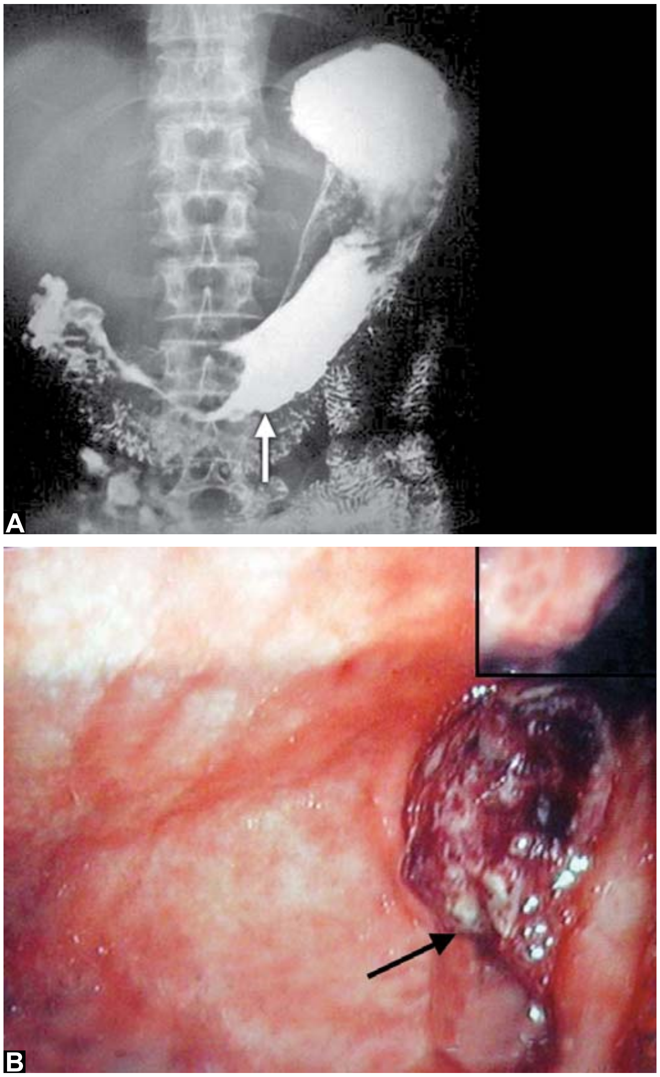

Figs 20.86A and B: Polypoid carcinoma stomach seen in barium meal picture and also through endoscopy.

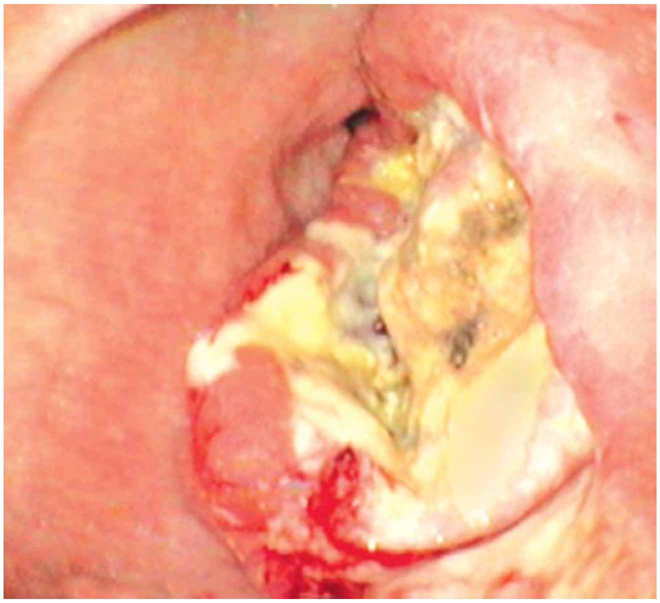

Fig. 20.87: Ulceroproliferative growth in the antrum—an endoscopic view.

Investigations

- Hb%, haematocrit.
- Barium meal (Irregular filling defect).
- Single contrast barium studies—sensitivity is 75%.
- Double contrast barium studies—sensitivity is 90-95% in the detection of gastric cancer, comparable to endoscopy.

Barium meal findings in carcinoma stomach

- Irregular filling defect
- Loss of rugosity
- Delayed emptying
- Dilatation of stomach in carcinoma pylorus
- Decreased stomach capacity in linitis plastica
- Margin of the lesion projects outward from the ulcer/lesion into the gastric lumen—*Carmanns meniscus sign*

- Gastroscopy with biopsy—10 targeted biopsies.
- Endosonography: EUS is useful to detect the involvement of layers of the stomach, nodal status and to define whether tumour is early or advanced. All 5 layers are visualised as alternate hypo- and hyperechogenic areas. It gives 90% accuracy for T staging and 80% of nodal staging.

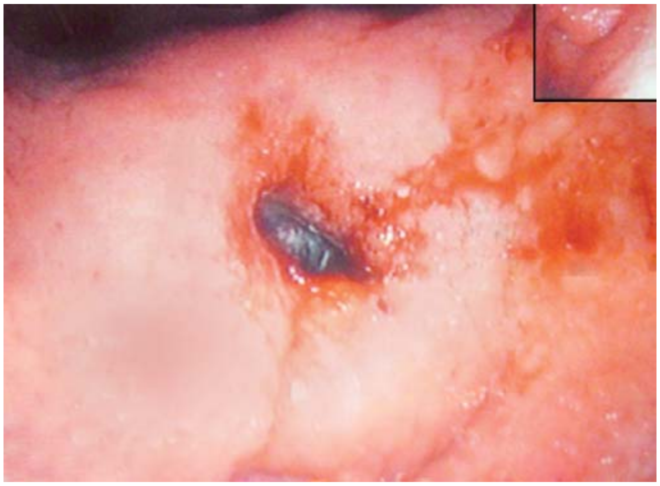

Fig. 20.88: Endoscopic view of the growth in the pylorus.

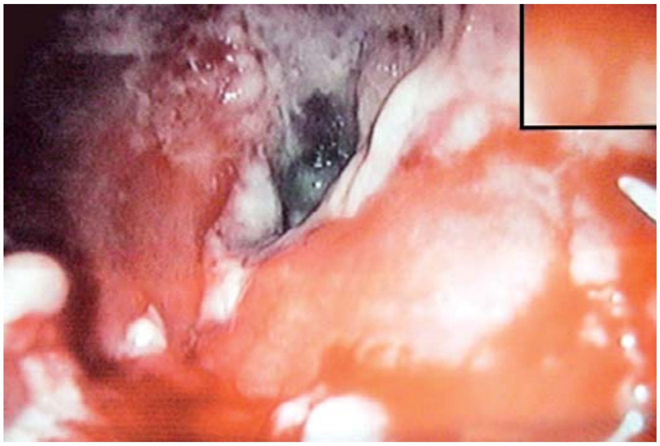

Fig. 20.89: Carcinoma stomach endoscopic view showing irregular tumour.

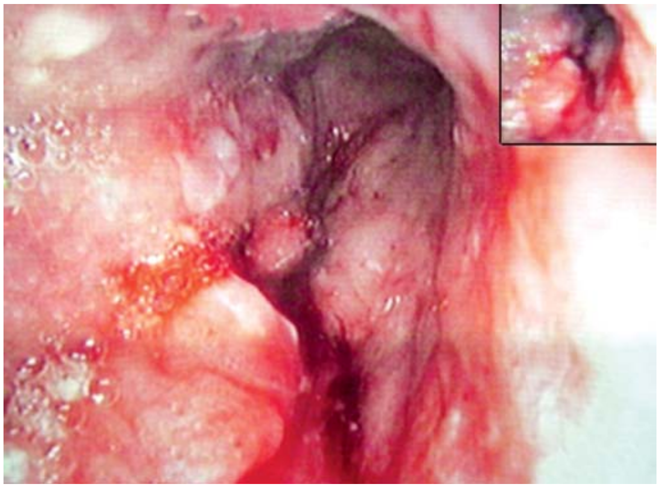

Fig. 20.90: Carcinoma stomach as seen through gastroscopy.

- US abdomen to see liver secondaries, ascites, nodes, ovaries. It is not as sensitive as EUS and CT scan.
- Liver function tests and prothrombin time. Secondaries in liver can cause hepatic dysfunction and liver cell failure. Enlarged hepatic nodes can compress CBD and cause obstructive jaundice also.
- Often FNAC from left supraclavicular lymph node when it is significantly palpable.
- *Laparoscopy* to stage disease. Laparoscopy will pick up peritoneal secondaries; detect occult metastases, organ invasion. Guided biopsy of peritoneum and nodes, peritoneal lavage for cytology and laparoscopic US can be done. It will avoid unnecessary laparotomy. *Signs of inoperability will be better identified* like—peritoneal deposits (CT may miss), fixity, liver secondaries, fixed coeliac nodes, para-aortic nodes, ascitic fluid positivity on analysis. It is more accurate than CT.
- CT scan abdomen and CT thorax in proximal tumours to see size, extent, infiltration, LN status, secondaries and operability.

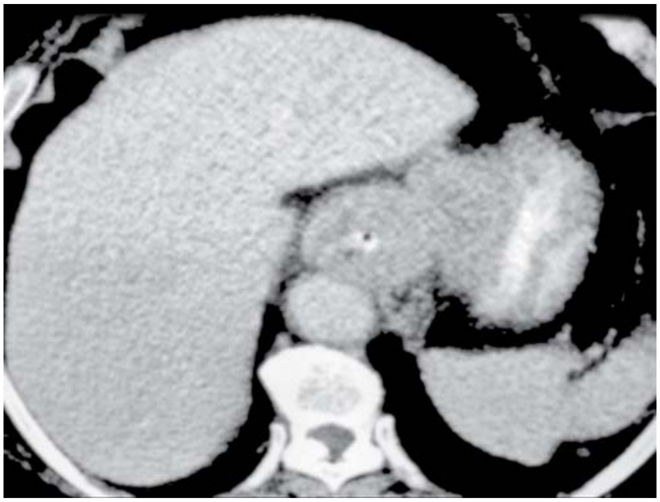

Fig. 20.91: CT scan showing carcinoma pylorus with luminal obstruction.

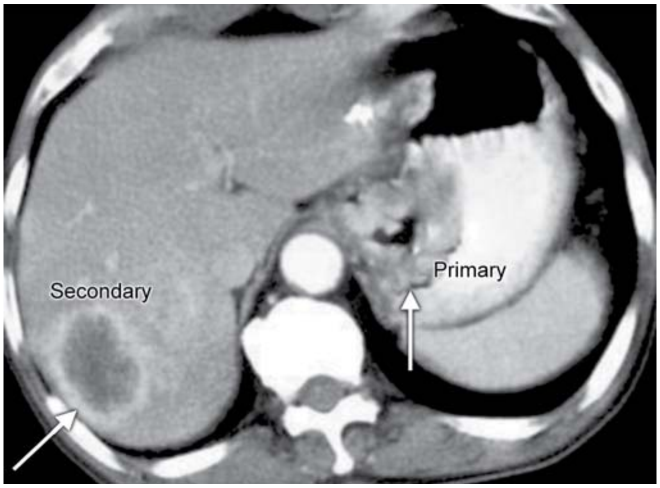

Fig. 20.92: CT scan showing primary tumour in the stomach with secondaries in the right lobe of the liver. Secondaries are commonly multiple; only occasionally it will be solitary.

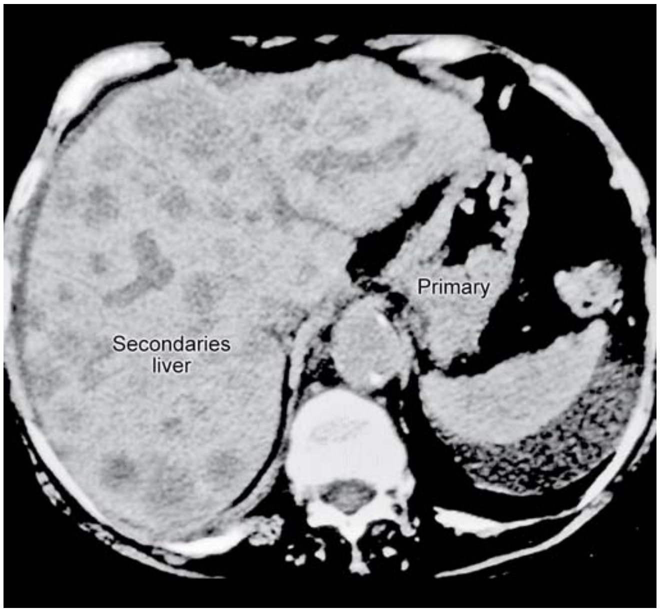

Fig. 20.93: CT scan showing multiple secondaries in liver with primary in the stomach.

- *Tetracycline fluorescence test*—gastric cancer cells take up tetracycline given orally which becomes yellow in colour. Under UV light it shows yellow fluorescence.
- CA 72-4 is important tumour marker to evaluate the relapse. CEA, CA 19-9, CA 12-5 are other markers.
- Combined *PET*—CT scan is useful. It is to evaluate metabolic, physiologic and functional activity. It is mainly useful to identify recurrent gastric cancer.
- *Sentinel node biopsy*—99mTc colloid peritumoral injection during gastrectomy. It has got sensitivity—80%, specificity—100% but not widely used.

Treatment

Surgery is the treatment of choice for carcinoma stomach.

Most intelligent animal is human being, but he happened to be the most cruel animal also.

Preoperative preparation

- ❖ Correction of anaemia, nutrition, fluid and electrolyte
- ❖ Cardiac, respiratory and renal status assessment
- ❖ Stomach wash using normal saline
- ❖ Prophylactic antibiotic as achlorhydria in gastric lumen allows colonisation of *Streptococcus faecalis*, *E. coli*, bacteroides, *Staphylococcus ablus*
- ❖ Blood/FFP may be needed preoperatively and for surgery

1. If it is an early growth, as in case of pylorus, *lower radical gastrectomy* with proximal 5 cm clearance is (*Subtotal gastrectomy*), along with removal of greater omentum, lesser omentum, all lymph nodes, spleen (to clear splenic lymph nodes), tail of pancreas (when required); and later Billroth II anastomosis.
2. In case of growth in the OG junction or upper part of the stomach, *upper radical gastrectomy* is done along with removal of spleen, both omentum, lymph nodes; and later *oesophagogastric* anastomosis. *Total gastrectomy is ideal here.*
3. In case of growth in the body or *linitis plastica, total gastrectomy (radical) with* oesophagojejunal anastomosis is done.

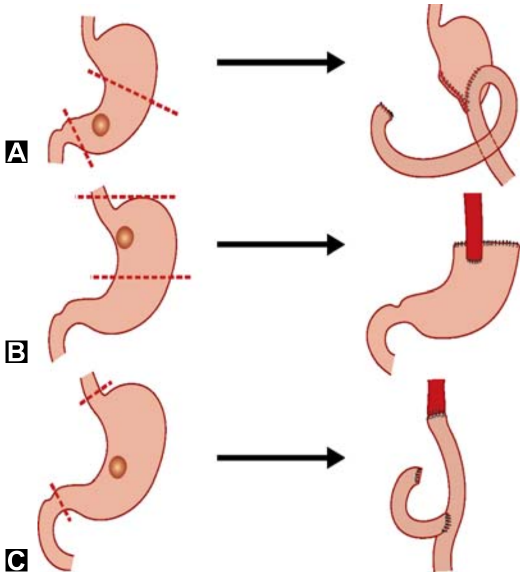

Figs 20.94A to C: Gastrectomies for carcinoma stomach: (A) Lower radical gastrectomy; (B) Upper radical gastrectomy; (C) Total radical gastrectomy.

Note:

- In tumours of antrum or body of the stomach, *subtotal gastrectomy is ideal*. Total gastrectomy is not done.
- *Total gastrectomy* with reconstruction is better option for *proximal growth, aggressive growth from body or linitis plastica*. It reduces heartburn, dumping, improves appetite better with better functional outcome. Roux-en-Y oesophagojejunostomy is the reconstruction done.
- *Distal clearance* towards duodenum is 1 cm from tumour end; proximal clearance is 5 cm.
- *D1 gastrectomy* is clearance of perigastric nodes; *D2 gastrectomy* is clearance of nodes along major vessel trunks—along coeliac and splenic arteries. D2 is always better than D1.
- ***D2 gastrectomy and adjuvant chemoradiation is ideal present concept therapy in operable carcinoma stomach.*** *D2 gastrectomy includes*—removal of stomach with growth, omental bursa, greater and lesser omentum, anterior layer of mesocolon, anterior pancreatic capsule, D2 lymphadenectomy.
- Prophylactic splenectomy once practiced is no longer found useful. It is done only if the tumour is adherent to spleen or its vessels.
- In early well circumscribed lesion, clearance of 2.0 is adequate but infiltrating tumours need 5 cm clearance.
- D2 dissection is better than D1 dissection of lymph nodes.
- *Omentectomy* (both greater and lesser) and *bursectomy* (removal of outer leaf of transverse colon baring colic vessels) should be added.
- Absolute curative resection is done—when there is absence of serosal, peritoneal, hepatic spread; with free adequate resection margin, with adequate nodal clearance (D dissection exceeds one level more the N nodal involvement).
- In operable cases, *neoadjuvant chemotherapy* improves the survival.
- Ideal reconstruction is either directly into the duodenum or in total gastrectomy using jejunal interposition with creation of reservoir. In distal gastrectomy Polya reconstruction is used.
- When the tumour is located upper or middle posterior wall of the stomach with serosal involvement without distant metastases *left upper abdominal evisceration* can be done. This *extended combined resection* includes—total gastrectomy, pancreaticosplenectomy, transverse colectomy always, occasionally may include left hepatectomy, left nephrectomy, and left adrenalectomy with resection of part of the left diaphragm. Its significance is doubtful.
- *Extended resection* means total gastrectomy with distal pancreatectomy with splenectomy.
- Para-aortic dissection is removal of fatty tissue and nodes from aortic hiatus to origin of inferior mesenteric artery often with left adrenalectomy to clear area adjacent to left renal artery.

Presently newer staging for dissection is practiced in Japan—D1, D2, D3, D4 depending on levels of lymph nodes cleared ('D' for dissection). D1 or D2 dissections are commonly done.

- ♦ **D1** is done when nodes are N_0. It is removal of nodes along the lesser and greater curves and pylorus (Group I: stations 3-6).
- ♦ **D2** is done when nodes are N_1. Here nodes like left gastric/common hepatic/splenic/retropancreatic are removed. Splenic nodes can be cleared with or without splenectomy along with removal of tail of pancreas. Here node stations 1-11 are removed. Stations 12-18 are not removed. Average 27 nodes are removed.
- ♦ **D3** is done when nodes are N_2. Nodes removed for clearance are hepatoduodenal, nodes along mesentery, middle colic (1-16 stations). Average 43 nodes are removed.
- ♦ **D4** is *not commonly advocated*. It is removal of stations 1-18.

D1—involvement of group I lymph nodes
D2—involvement of group I and II lymph nodes
D3—involvement of group I, II and III lymph nodes
D4—involvement of group I, II, III and para-aortic nodes

Note:

All lymph nodes are numbered differently as 'stations' or 'echelons'.

Lymph node stations in gastric carcinoma (Japan)—18 stations are there

1. Right cardiac
2. Left cardiac
3. Nodes along the lesser curvature
4. Nodes along the greater curvature
    a. Along short gastric vessels—4sa
    b. Along left gastroepiploic vessels—4sb
    c. Along right gastroepiploic vessels—4sd
5. Suprapyloric nodes
6. Subpyloric nodes
7. Along left gastric artery
8. Along common hepatic artery
9. Along celiac axis
10. At splenic hilum
11. Along splenic artery
12. At hepatoduodenal ligament
13. Retroduodenal lymph nodes
14. At root of mesentery
15. Around middle colic artery
16. Para-aortic nodes
17. Around lower oesophagus
18. Supradiaphragmatic

First tier nodes: Nodes within 3 cm from primary tumour (stations 1-6)—N_1 (old TNM)

Second tier nodes: Nodes in main and intermediate arterial trunk (stations 7-11)—N_2 (old TNM)

Third tier nodes: Nodes at stations 12-18 (para-aortic and above)—N_3 (old TNM)

Note: 15 lymph nodes must be removed for adequate staging of nodes

4. *Endoscopic mucosal resection (EMR)*
    - It is practiced in Japan in early gastric cancer. Tumour less than 2 cm, elevated, well-differentiated tumours without nodal disease is the ideal selection for EMR. Proper regular endoscopic surveillance is essential in these patients after EMR.
    - Lesion is elevated by injecting normal saline submucosally and elevated lesion with adjacent mucosa is resected using cautery up to muscularis propria; later with clips defect is apposed.
    - It is done in protruded early gastric cancer or ulcer-free depressed cancer which is well-demarcated.
5. *Photodynamic therapy*
    - Intravenous injection of haematoporphyrin derivative photosensitiser is given initially; 48 hours later laser light is delivered to tumour through an endoscope to create interaction between two, releasing highly reactive singlet O_2 which causes tumour necrosis.

Adjuvant Therapy

Chemotherapy

- Mitomycin 10 mg IV single dose in monthly cycles.
- 5-fluorouracil, 250 to 500 mg IV in 5% dextrose daily for 5 days.
- Cisplatin and epirubicin, adriamycin, oxaliplatin, capecitabine are other drugs used at present.
  (*FAM regime* is 5-fluorouracil, adriamycin, mitomycin C).
- Chemotherapy is also used as palliation in inoperable cases/metastatic cases whether surgical palliation (if obstruction or bleeding is present) is needed or not.
- Neoadjuvant chemotherapy to downstage the tumour is also used to make it operable later. ECF (Epirubicin, cisplatin, 5 FU), EAP (Epirubicin, adriamycin, cisplatin), FDT (% FU, doxorubicin, triazinate) are used.
- *Bevacizumab* (anti-VEGF) is currently under trial.
- Instillation of mitomycin C impregnated charcoal intraperitoneally is done to control lymphatic disease (Japan). It specifically kills the malignant cells in lymphatics and lymph nodes in the abdomen.

Chemoradiotherapy

- Postoperative radiotherapy (RT) with chemotherapy using 5 FU and leucovorin. Chemoradiation has become the *standard* adjuvant treatment.
- Some benefits are shown but controversial. It is said to be better than adjuvant chemotherapy alone.

Note:

Very less *role for radiotherapy at present.* Many radiosensitive structures in the gastric bed prevents tumour to have effective RT. RT is used as palliative method for painful bone secondaries.

Immunochemotherapy as an Adjuvant

- It is given in *stage III* carcinoma *after* radical gastrectomy. It starts from 5th postoperative day to the end of 2 years.
- It is based on the fact that preoperative T cell counts, percentage, are reduced which can be improved/modulated by this therapy.
- Residual cancer cells or micrometastases can be eradicated by this therapy.
- *Regime contains*—initially *immunotherapy* using weekly intramuscular injections of Picibanil (*Streptococcus pyogenes* derived) immune agent 1.0 clinical unit is given on 5th postoperative day. Later from 10th postoperative day onwards, *chemotherapy* regime consisting of mitomycin C and 5 FU is used as follows—mitomycin C 4 mg/50 kg IV twice weekly for 2 weeks, then weekly for 6 weeks; 5 FU as 500 mg/50 kg IV twice weekly for 2 weeks, then weekly for 6 weeks, then 600 mg/50 kg orally daily for 2 years.

Signs of inoperability

- *Adherent to pancreas or colon or mesocolon*
- *Ascites*
- *Para-aortic lymph nodes*
- *Secondaries in liver*
- ***Palpable mass is incurable but can be resectable surgically***
- Blumer shelf
- Left supraclavicular nodes
- Sister Mary Joseph nodule
- Irish node (Left axillary lymph node secondaries)

Goodness is the permanent investment.

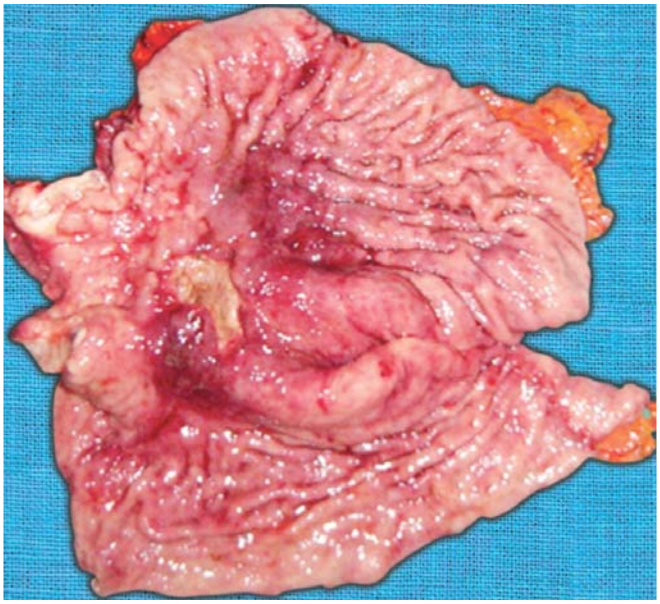

Fig. 20.95: Carcinoma stomach resected specimen.

Management of gastric carcinoma

- *Early growth in pylorus*: Lower radical gastrectomy with removal of tumour, proximal 5 cm clearance, nodal clearance, greater and lesser omentum, distal pancreas and spleen (now not regularly removed; it is removed to clear splenic nodes—one of the node stations) and Billroth II anastomosis or Roux-en-Y anastomosis is done. Postoperatively adjuvant chemotherapy should be given—5-fluorouracil, mitomycin, epirubicin, cisplatin, oxalilatin, capecitabine
- *D1 or D2 nodal dissection* should be done for adequate clearance in curative surgery
- *Growth in body, proximal growth, diffuse carcinoma and generalised linitis plastica* are the indications for total radical gastrectomy with oesophagojejunal anastomosis.
- Neoadjuvant chemotherapy in advanced gastric cancer prior to surgery and later gastrectomy.
- Instillation of mitomycin C impregnated charcoal intraperitoneally to control lymphatic disease (Japan)
- Palliative procedures like *palliative partial gastrectomy*, anterior gastrojejunostomy, Devine's exclusion procedure, luminal stenting in proximal inoperable growths, chemotherapy are used in inoperable cases
- In early carcinoma proper lymph nodal clearance is important

Palliative Treatment

- To palliate pain
- To palliate vomiting
- When there is bleeding
- To improve appetite
- *Partial gastrectomy (palliative) is the best method*

Palliative Procedures

- Palliative anterior gastrojejunostomy with jejuno-jejunostomy.
- *Devine's exclusion procedure* wherein instead of removal of tumour, it is excluded with Billroth II anastomosis.
- For unresectable growth in upper part of the stomach, palliation can be attained by inserting tubes through the oesophagus to the stomach like Mousseau-Barbin tube, Celestin tube (like in carcinoma oesophagus). Endoscopic stenting can also be tried.
- Palliative chemotherapy is used in advanced stage.
- RT and analgesia like morphine is used to relieve pain.
- Surgery / endotherapy is used to relieve obstruction. Palliative partial gastrectomy (limited local resection), anterior GJ, laser, stenting are different options.
- Surgery (partial gastrectomy) or endotherapy is required in bleeding.
- Palliation for ascites is difficult even though repeated tapping and chemotherapy is done.

Palliative procedures for carcinoma stomach

- *Palliative partial gastrectomy is the best palliation*
- Palliative anterior gastrojejunostomy with jejuno-jejunostomy
- Devine's exclusion procedure
- M-B tube/Celestin tube insertion for proximal stomach growths
- Endoscopic stenting/dilatation
- Laser recanalisation
- Palliative chemotherapy—FAM regime

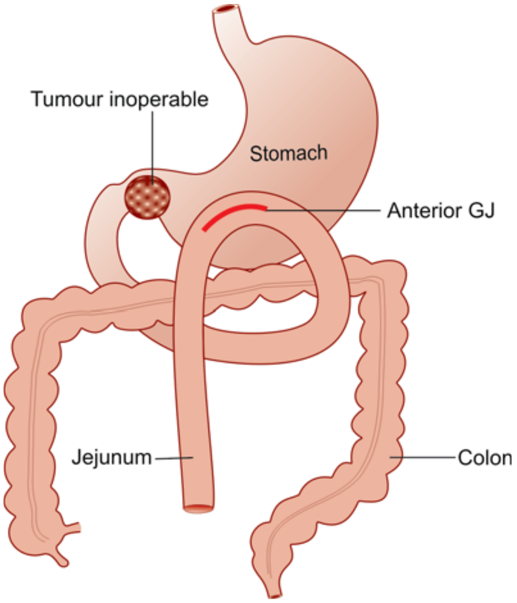

Fig. 20.96: *Anterior gastrojejunostomy* is one of the palliative procedures used in case of inoperable carcinoma stomach. It is anterior, antecolic GJ.

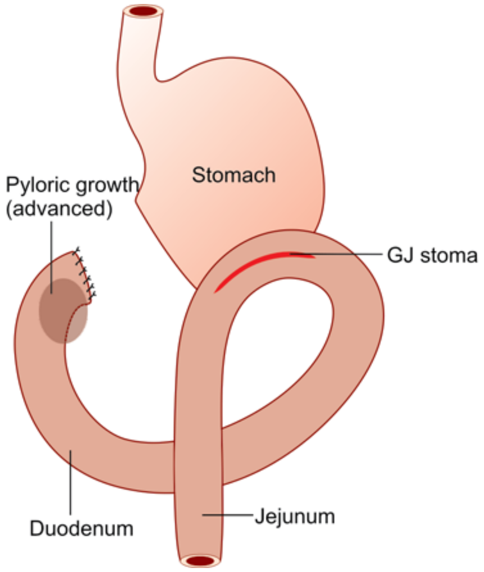

Fig. 20.97: *Devine's exclusion* procedure. Here inoperable pyloric tumour is left and excluded. Proximal stomach is transected and used for Billroth II like gastrojejunostomy.

Prognosis

- In early gastric cancer which has undergone good surgical resection, 5-year survival rate is 70-90% in Japan. In India, 5-year survival rate is 20%.
- In advanced gastric cancer it reduces to 20-25% in Japan. In other countries it is still worser.
- Overall prognosis is worse in carcinoma stomach.
- When serosa is not involved 5-year survival is 50%. When serosa is involved it is 20%.
- Nodal spread is a bad prognostic factor.
- Involvement of more than 4 nodes carry poor prognosis.
- In Japan, prognosis is better due to early diagnosis and better technical and staging approach.
- Early gastric cancer has got better prognosis.
- Intestinal type has better prognosis than diffuse type.
- Size of the tumour (2 cm); tumour depth; and histological type and grading.

Prognostic factors

- Early or advanced
- Histological grading
- Staging
- Gross types
- Lymph node status
- Liver secondaries
- Serosal involvement—an important factor
- Intestinal type—got better prognosis than diffuse type
- Ascites
- Response to treatment

Note:

- Recurrence rate is very high in first 3 years after therapy.
- If primary tumour can be removed surgically, it is called as *resectable tumour*. Resectability can be for palliation or along with radical surgery for cure.
- *Operable tumour* means radical gastrectomy with (amenable for) nodal clearance.
- *Inoperable tumour* means not amenable for radical surgery.
- Inoperable tumour can be resectable for palliation.
- **R_0** is resection with complete clearance—no *residual disease*. **R_1** is resection but with microscopic residual tumour. **R_2** is resection with macroscopic residual tumour, **R_3** is unresectable.
- D1 and D2 dissections are different for proximal and distal tumours.
- In early gastric cancers, laparoscopic endoluminal transgastric mucosal resection of the lesion is under trial now.
- Adjuvant chemoradiation (RT with 5-FU as radiosensitiser) is also under trial for proximal tumour, undifferentiated high grade tumour.
- Antrectomy is also called as *hemigastrectomy*. Removal of 60-75% stomach is called as *partial gastrectomy*. Removal of more than 80% of stomach is called as *subtotal gastrectomy*.
- Perforation in gastric cancer has got high mortality.
- Most common site of recurrence is stomach bed, then lymph node and anastomotic site.
- Will Roger's stage migration phenomenon is thorough staging with stage for stage to improve the outcome.

GASTRIC LYMPHOMA

There are two types of gastric lymphomas.

- Primary.
- Secondary.

Note:

- Gut lymphomas are 10% of all lymphomas. It is common in males (2:1). Coeliac disease, alpha chain disease, ulcerative colitis, HIV, transplantation—are the aetiology in general for gut lymphomas. *H. pylori* is relevant for gastric lymphoma.
- It occurs in stomach (55%), small bowel (30%), colon (10%), ileocaecal region (5%).

A. PRIMARY GASTRIC LYMPHOMA

- Primary gastric lymphoma is most common one among GI primary lymphomas.
- It is second common malignant neoplasm of the stomach (5% of gastric malignancies).
- It is common in elderly men and common in the antrum of the stomach.
- It can be infiltrative, ulcerative, nodular, polypoid or combined.
- Gastric lymphomas are commonly of non-Hodgkin's type (NHL—B cell type 98%).
- It is arising from B cells—derived from mucosal associated lymphoid tissue (MALT). So often called as MALTOMA
- Diffuse mucosal thickening which eventually ulcerates—is the pathology.
- Disease remains in the stomach for long time and later spreads to liver and other lymph nodes.
- Association of primary gastric lymphoma with *H. pylori* is well-established. Normal gastric mucosa does not contain lymphocytes but mucosa with *H. pylori* infection contains lymphocytes.
- Usually disease is localised to stomach without any other lymph nodes in the abdomen or mediastinum with normal liver, spleen, bone marrow and total count.

- It is basically a localised disease of stomach to begin with but later to have systemic spread.
- *Main presentation* is mass abdomen, pain, vomiting and loss of weight. Perforation, bleeding, obstruction, gastrocolic fistula are also not uncommon.
- *Dawson's criteria for primary GI lymphoma* (NHL): No palpable superficial lymph nodes; no thoracic lymph nodes in CT scan; normal WBC count and bone marrow; predominant bowel pathology; liver, spleen are normal.
- *Ann Arbor staging:* Stage IE—confined to GIT; IIE – GIT + regional nodes; IIIE – GIT + nodes away from and beyond regional; IVE – GIT + other intra-abdominal organs or extra-abdominal.
- *It can be*—B cell, T cell, low grade, high grade, pure centroblastic, unclassified.

Clinical Features

- Pain abdomen, melaena and mass abdomen which is smooth and firm.
- Loss of weight, loss of appetite.

Gastric lymphoma is associated with:

- Wiskott Aldrich syndrome
- Klinfelter syndrome
- HIV, *H. pylori* (75%)
- Ataxia telangiectasia
- SLE
- Chronic gastritis
- MALT

Complications

- Obstruction.
- Perforation.
- Bleeding.
- Spread to other lymph nodes and liver—only late event.

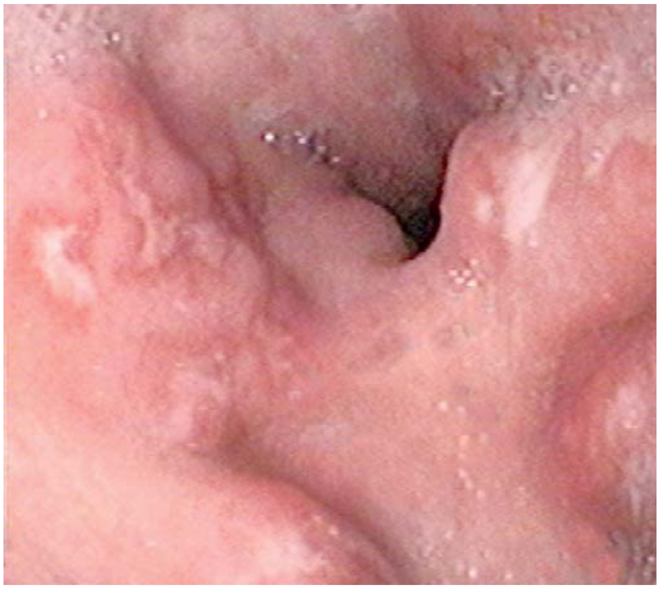

Fig. 20.98: Gastric lymphoma, endoscopic view. It can be primary or secondary. Secondary is common. Primary occurs first in stomach and later other lymphatic system are involved. Gastrectomy is the treatment with chemotherapy.

Investigations

- Gastroscopy and biopsy.
- Endosonography.
- Peripheral smear, bone marrow aspiration.
- CT abdomen and CT chest and mediastinum.

Treatment

- Radical subtotal gastrectomy is done in primary gastric lymphoma at antrum. Radical total gastrectomy is done in primary lymphoma of proximal stomach.
- *Anti-Helicobacter pylori* therapy. Early primary gastric lymphoma may regress by triple regime.
- Chemotherapy as adjuvant. Drugs are same as used for NHL. RT is also useful.

B. SECONDARY GASTRIC LYMPHOMA

- It is the most common type of lymphoma occurring in stomach. It differs from primary in many aspects.
- Here mainly systemic lymph nodes are involved first and the disease begins in lymph nodes at different regions, later to occur in stomach.
- Peripheral smear shows lymphocytosis; bone marrow shows changes; CT abdomen and chest shows multiple nodal enlargement.
- It is basically a systemic disease extending to stomach.
- Perforation and bleeding are the complications.
- Treatment is mainly chemotherapy—CHOP/ABVD/MOPP/BACOP and other regimes.
- Surgery is not done unless any local complications occur.

Gastrointestinal mesenchymal tumours

- *GIST*—CD 117; CD 34; smooth muscle actin
- *Smooth muscle tumours* —CD 34; smooth muscle actin; Desmin
- *Neural tumours*—S-100; CD 34

GASTRIC SARCOMAS

- Stomach is the most common site of GI sarcoma.
- Leiomyosarcoma is the most common type of gastric sarcoma. Other type is malignant leiomyoblastoma.
- Leiomyosarcoma is common in proximal stomach.
- It attains enormous size with rubbery consistency.
- Mainly spreads through the blood to the liver.

Clinical Features

- Pain, vomiting, anorexia.
- Mass in upper abdomen.

Complication

- GI bleeding.

Investigations

- U/S abdomen and CT scan.
- Gastroscopy is useful in less than 50% of cases.

Treatment

- Subtotal gastrectomy.
- Chemotherapy.

Leiomyosarcoma has got better prognosis compared to adenocarcinoma and lymphoma of stomach.

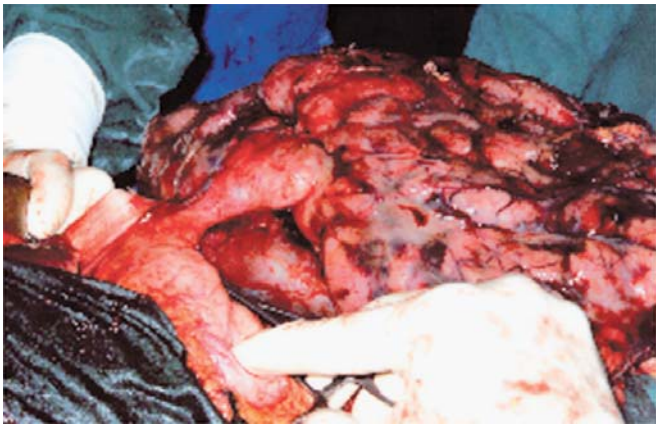

Fig. 20.99: Gastric sarcomas.

GASTROINTESTINAL STROMAL TUMOURS (GIST)

- It is a rare tumour of GI tract—0.2% of all GI tumours.
- But it is the most common nonepithelial tumour of the gastrointestinal tract. *Stomach is the most common site* of all GISTs—50%; 25% of all GISTS are from small bowel; rectum—15%; colon—10%.
- Equal in both sexes and common in 50-70 years age group.
- GIST arises from interstitial cell of Cajal (pacemaker cell which intercalates between smooth muscle cells and intramural neurons). Mutation of tyrosine kinase and platelet derived growth factor alpha (PDGFα) are the newer pathogenetic theories.
- GIST is classified as very low-risk (2 cm); low-risk (2-5 cm); intermediate risk (5-10 cm) and high-risk (> 10 cm) based on *tumour size* and *mitotic activity* of cells.
- 95% of GISTS express c-kit – CD117 mutations – a specific molecular marker.
- *Kit protein* (CD117, stem cell factor receptor) is a transmembrane tyrosine kinase receptor which is detected by immunohistochemistry. It distinguishes from true smooth muscle neoplasms. 80% of GISTs are positive for CD34—haematopoietic progenitor cell antigen.
- Abdominal pain; weight loss; *GI bleed (most* common presentation) and large mass abdomen are typical. Mass is extraluminal as it is of submucosal origin but expands and compresses the mucosa. 50% of GIST can present as metastatic disease to liver and peritoneum (ascites).
- GIST's almost *never* metastasise to regional lymph nodes.
- It can be spindle shaped (70%); epithelioid (25%); combination.
- *Carney's triad*—extra-adrenal paraganglioma; pulmonary chondroma; gastric GIST.

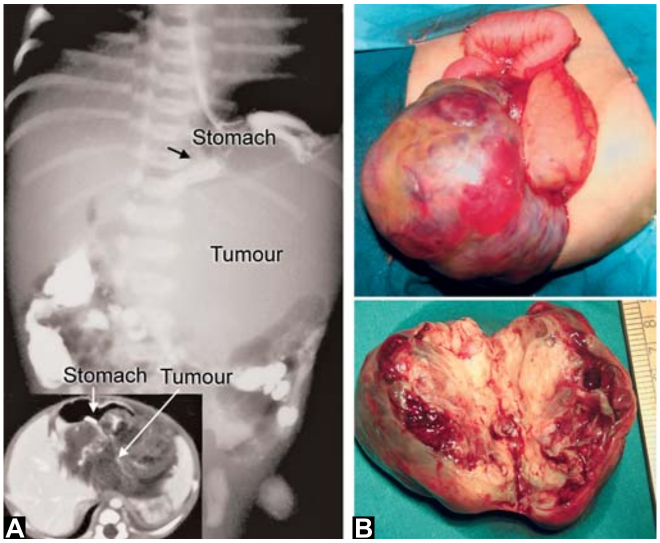

Figs 20.100A and B: Barium meal X-ray showing gastric teratoma presenting as a large mass lesion compressing gastric lumen. Inset shows CT picture also. Operated specimen is also shown.

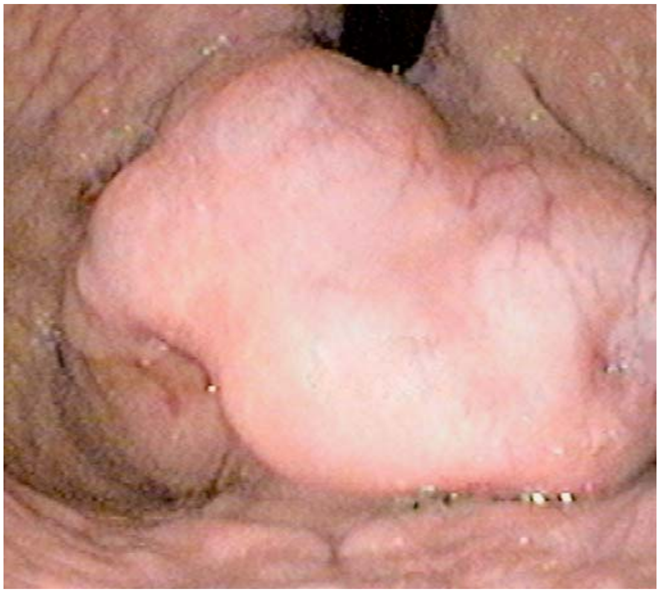

Fig. 20.101: Gastric leiomyosarcoma near the fundus—endoscopic view. It is now classified under GIST.

***Investigation*:**

- Mainly CT scan.
- Tumour/molecular marker to differentiate it from sarcomas (immunohistochemistry).
- Endosonography guided biopsy/guided FNAC are important to get histological confirmation.
- 18 FDG PET scan is very useful adjunct to CT but reserved for difficult/equivocal cases.

***Treatment*:**

- Surgical resection of the tumour with part of the adjacent bowel.
- *Imatinib mesylate*—a specific *oral* drug (year 2000) that inactivates tyrosine kinase kit and so prevents phosphorylation of the receptor and proliferation of tumour is very much

The remedy for injuries is not to remember them.

beneficial in advanced cases. Now it is also used if tumour size is more than 10 cm; in intraperitoneal rupture/spillage; haemorrhage in GIST; multifocal occurrence. Duration of imatinib is usually one year.
- Newer drug—*SU11248* inhibits tyrosine kinase receptor as well as blocks PDGFRA. Another newer derivative—*sunitinib and dasatinib* is used in imatinib refractory cases.
- Doxorubicin ± ifosfamide is tried in metastatic disease but with only 5% response rate.

Prognostic factors:
- Size of GIST more than 5 cm, male sex.
- High mitotic activity (> 15 mitoses per 30 high power field), mixed pathology (spindle + epithelioid).
- Liver spread.
- KIT exon 9 mutation which is more aggressive than KIT exon 11 mutations—carries poor prognosis.

GIST

- GIST is the most controversial topic.
- Primary nonepithelial tumours are classified as: Group I—smooth muscle type expressing smooth muscle actin and desmin; Group II—gastrointestinal autonomic nerve tumours (GANs/myenteric plexus tumour/plexosarcomas) express neuron specific enolase and s-100 and are aggressive; Group III—dual differentiation; Group IV—uncommitted type.
- GIST can be benign, intermediate/borderline, or malignant. Size and mitotic activity are main factors of malignancy; others are—necrosis, haemorrhage, hypercellularity.
- GIST is common in fundus of stomach. It can be submucosal (60%), subserosal (30%), intramural. Fundal GIST carries better prognosis than small bowel GIST.
- Small bowel GIST is common in jejunum (40%), ileum and duodenum. It can be submucosal or subserosal. It presents with GI bleed, abdominal pain, intestinal obstruction, mass abdomen, intussusception, perforation and cachexia. It often produces HCG causing hyperemesis.
- *GISS*—is gastrointestinal stromal sarcoma. It is common in stomach (50%), small bowel. GISS accounts for 2% of soft tissue sarcomas. 70% of them are symptomatic at early stage. Surgical resection is the treatment.

PYLOROPLASTY

Indications

- Bleeding duodenal ulcer.
- As a drainage procedure after vagotomy in uncomplicated-DU which is not responding to medical line of treatment.
- As a part of procedure like Ivor-Lewis operation and Siguira-Futagawa operation.

Contraindications

- Duodenal ulcer with cicatrisation and scarring.

Types

- ***Finney's pyloroplasty.***
- ***Heineke-Mikulicz pyloroplasty***.

Complications

- Duodenal leak—very dangerous one.
- Later reflux.

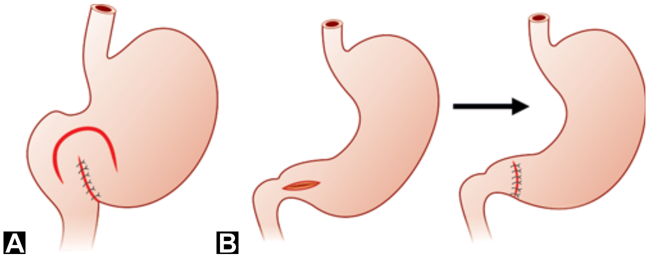

Figs 20.102A and B: Types of pyloroplasty. (A) Finney's pyloroplasty, (B) Heineke-Mikulicz pyloroplasty.

GASTROSTOMY

- It is the procedure wherein a tube is passed into the stomach per abdominally for the purpose of enteral feeding.
- It can be *temporary or permanent.*

Types

A. *Serous lined*—for temporary gastrostomy.
B. *Mucosa lined*—for permanent one.

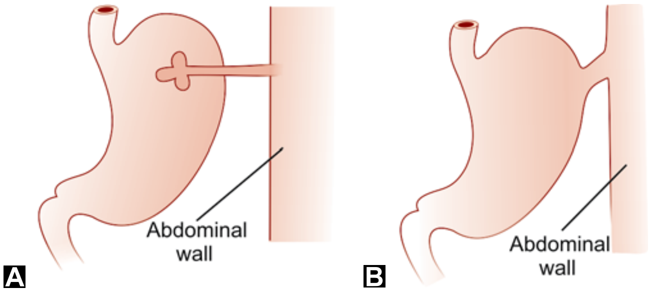

Figs 20.103A and B: Types of gastrostomy. (A) Temporary gastrostomy, and (B) Permanent gastrostomy.

After minilaparotomy, gastrotomy is done. Foley's or Malecot's catheter is placed in the stomach. It is fixed to parietal peritoneum through two circular purse string sutures.

Other Types

- Stamm's temporary gastrostomy.
- Kader-senn temporary gastrostomy.
- Janeway's permanent gastrostomy.
- *Endoscopic percutaneous gastrostomy*—becoming popular (PEG).

Indications

- Oesophageal strictures.
- Any conditions where tube feeding is required for more than 4 weeks (e.g. burns, severe sepsis).
- Major neck surgeries.

Contraindications

- Previous gastrectomy.
- Gastric diseases with impaired gastric emptying.
- Intestinal obstruction.

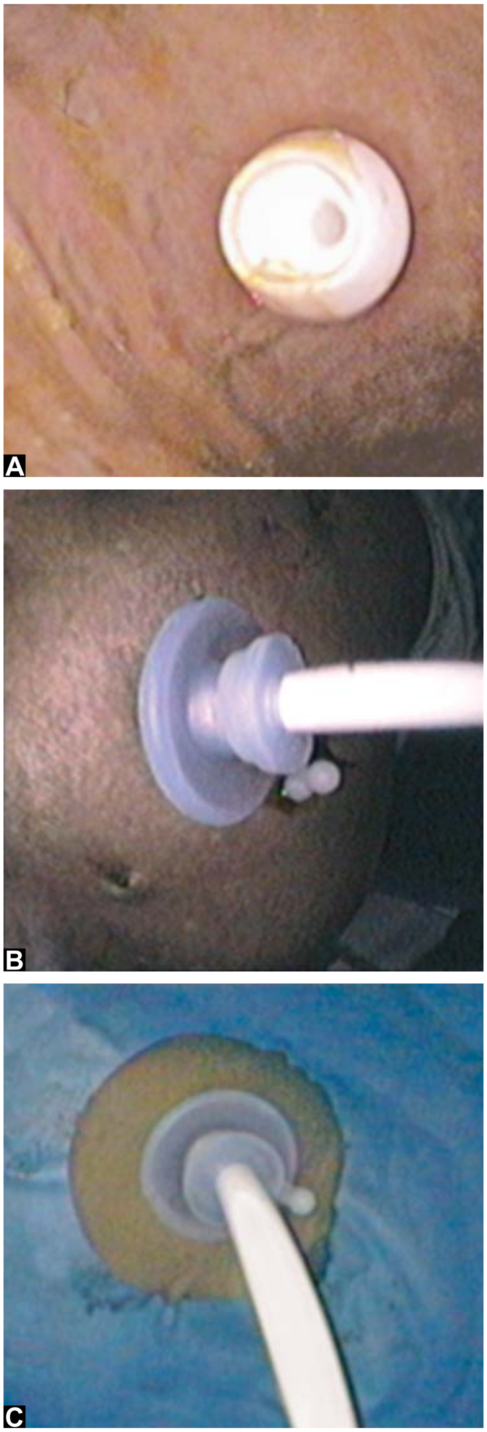

Figs 20.104A to C: Percutaneous endoscopic gastrostomy (PEG). It is easier and acceptable method and less invasive.

Complications

- Infection.
- Trauma to other organ, e.g. colon.
- Leak from gastrostomy site.
- Aspiration pneumonia.
- Blockage.

GASTRECTOMY

Types

1. *Billroth I* is done for benign condition. Here along with partial gastrectomy, gastroduodenostomy is done.
2. *Billroth II* is done for carcinoma stomach. After partial gastrectomy, gastrojejunostomy is done and duodenal stump is closed.
3. *Lower radical gastrectomy* is done in early carcinoma pylorus. Here along with the growth and proximal 5 cm of stomach, omentum, lymph nodes, spleen with tail of pancreas are removed and Billroth II anastomosis is done.
4. In growth of upper part or OG junction, *upper radical gastrectomy* is done along with *oesophagogastric anastomosis*.
5. In some cases like linitis plastica, *total gastrectomy* along with *oesophagojejunal anastomosis is* done.

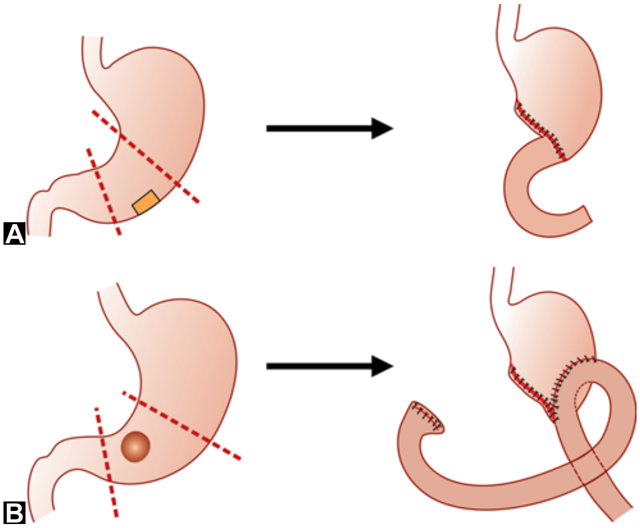

Figs 20.105A and B: Types of gastrectomy. (A) Billroth I anastomosis and (B) Billroth II anastomosis.

Indications

- Chronic benign gastric ulcer.
- Benign tumours of stomach (Leiomyoma).
- Carcinoma stomach.
- Stomal ulcer.
- Bleeding ulcer.
- Leimyosarcoma, gastric lymphoma.
- Menetrier's disease.

Complications

- Bleeding.
- Bile leak.
- Duodenal blow out.
- Gastric fistula.
- Dumping syndrome.
- Anaemia.

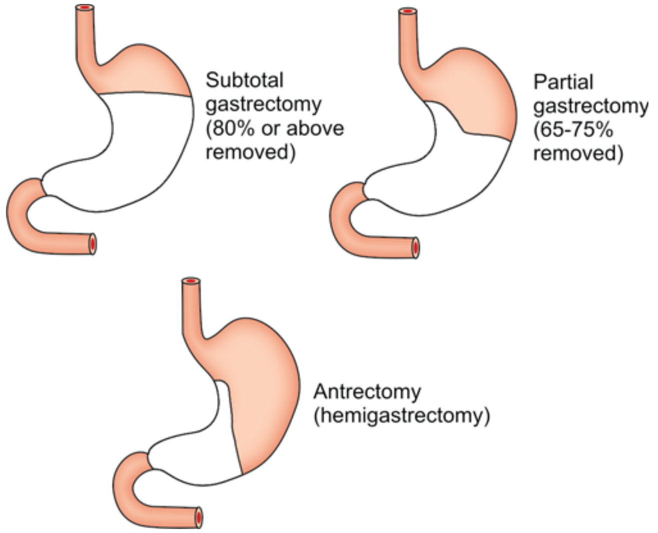
Subtotal gastrectomy (80% or above removed)
Partial gastrectomy (65-75% removed)
Antrectomy (hemigastrectomy)

Fig. 20.106: Different types of gastrectomy. Subtotal—more than 80% of stomach is removed; partial—60-75% removed.

GASTROJEJUNOSTOMY (GJ)

Types

1. Anterior GJ.
2. Posterior GJ.

Anterior gastrojejunostomy is done as a palliative procedure in case of advanced inoperable (adherent posteriorly) carcinoma pylorus to palliate vomiting. It is anastomosis between jejunum and anterior surface of stomach, in front of transverse colon (antecolic).

Posterior GJ is done along with truncal vagotomy, in pyloric stenosis due to chronic duodenal ulcer. *It is posterior, vertical, retrocolic, short loop, isoperistaltic (of Mayo*).

GJ is also done as part of the Billroth II gastrectomy.

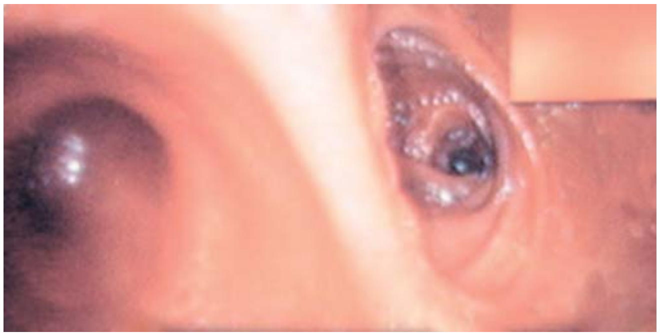

Fig. 20.107: Gastrojejunostomy stoma view on endoscopy.

Problems with GJ

- Stomal obstruction.
- Afferent or efferent loop obstruction.
- Dumping syndrome.
- Duodenal blow out.
- Retrograde intussusception.
- Reflux gastritis.
- It is a rare entity occurs after gastrojejunostomy.
- It can occur immediately, in weeks or in months.

RETROGRADE JEJUNOGASTRIC INTUSSUSCEPTION

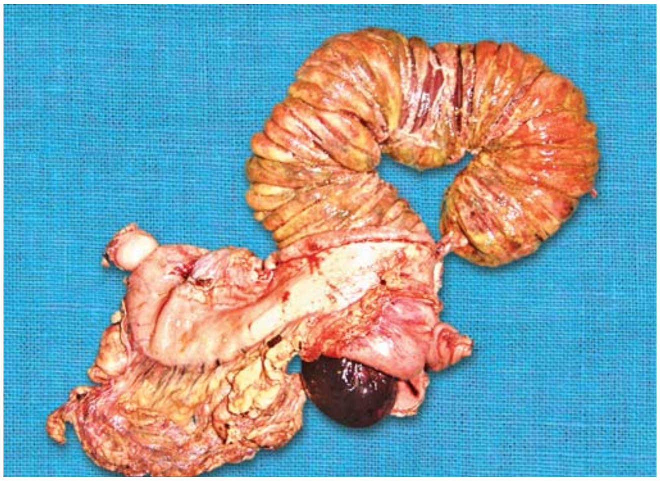

Fig. 20.108: Retrograde jejunogastric intussusception with gangrene—resected specimen.

- Presents with *pain* above and left of the umbilicus, *haematemesis,* firm, tender *mass* above and towards the left side. Patient becomes better in erect posture.
- Condition causes stomal obstruction.
- Barium meal shows *coiled spring look* within the stomach remnant. *Gastroscopy* is diagnostic.
- It is treated by open reduction of the intussusception with anchoring stitch to the bowel or enteroanastomosis. When gangrenous, resection and *Roux-en-Y anastomosis* is done.

VAGOTOMY

> *In an inquiry which I had formerly instituted, respecting the functions of the stomach, I divided these nerves (the vagi) in the neck of a dog, for the purpose of ascertaining the influence which they possess on the secretion of the gastric juice.... We may conclude that the suppression of the secretions...sufficiently demonstrate, that the secretions of the stomach and intestines are very much under the control of the nervous system.*
>
> —**Benjamin Collins Brodie,** 1814

Types

1. ***Truncal vagotomy*** along with GJ as drainage procedure, is done in case of pyloric stenosis due to chronic DU.

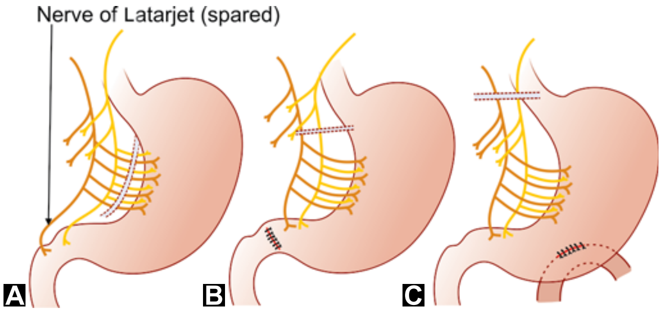

Figs 20.109A to C: Types of vagotomy. (A) Highly selective vagotomy, (B) Selective vagotomy with pyloroplasty, (C) Truncal vagotomy with gastrojejunostomy.

In uncomplicated DU, with failure of medical treatment and when HSV cannot be done, then truncal vagotomy with pyloroplasty is done.

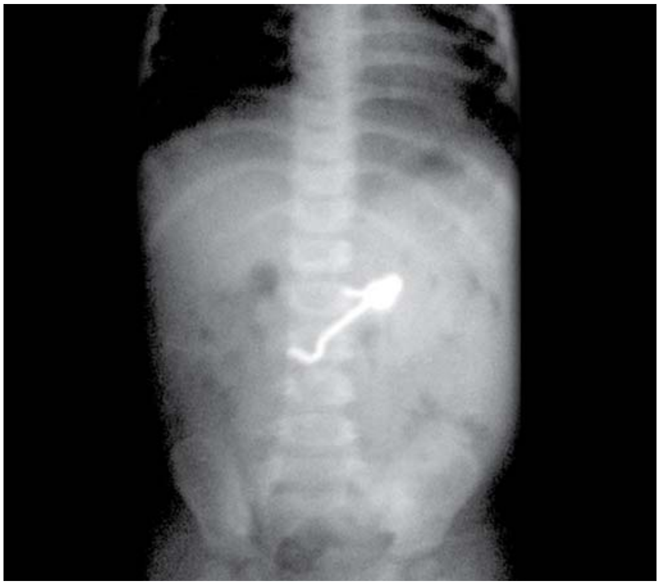

Fig. 20.110: Foreign body in the stomach. It usually does not harm and commonly passes down.

2. ***HSV (Highly selective vagotomy)*** or ***Parietal cell vagotomy or Super selective vagotomy*** is done in case of uncomplicated DU where medical treatment fails. Here only fibres entering the stomach is divided both anteriorly as well as posteriorly. Nerve of Latarjet is retained to supply antrum and so *no drainage procedure is required* (Amdrup). *It has got 10% recurrence rate.*
3. ***Selective vagotomy*** is at present not done often. Here along with a drainage procedure, either GJ or pyloroplasty, vagotomy is done with retaining of coeliac and hepatic branches. It has got 10% recurrence rate.

Complications of vagotomy

Intraoperative	*Late complications*
Injury to spleen/oesophagus/liver/pancreas/thoracic duct stomach Bleeding from phrenic veins/gastric vessels/perioesophageal vessels Pneumothorax Oesophagogastric disconnection *Early postoperative* Gastric atony Lesser curve necrosis in HSV Transient dysphagia	• Vagotomy diarrhoea—20% – Most common complication of vagotomy – Gastric stasis – Hypoacidity with fermentation with bacterial overgrowth causing enteritis – Alteration in intestinal villi and enzyme content, altered small bowel motility due denervation – Altered biliary and pancreatic exocrine function – Cholestyramine 4 gram tid/reverse jejunal loop is the treatment for intractable case • Reflux oesophagitis • Oesophageal stricture • Gallstones—due to bile stasis and gastric atony—15%

The days that make us happy make us wise.

Chapter 21 Small Intestine

CHAPTER OUTLINE

- Anatomy
- Meckel's Diverticulum
- Regional Enteritis
- Surgical Complications of Typhoid
- Surgical Complications of Roundworm
- Pneumatosis Cystoides Intestinalis
- Mesenteric Vessel Ischaemia
- Necrotising Enterocolitis
- Small Bowel Tumours
- Benign Tumours of Small Bowel
- Malignant Tumours of Small Bowel
- Carcinoid Tumour
- Short Bowel Syndrome
- Massive Bowel Resection
- Small Bowel Enema
- Capsule Endoscopy
- Small Bowel Enteroscopy
- Enteric/Gastrointestinal Fistula

ANATOMY

The small intestine is 6 m in length. This includes upper fixed (duodenum, 25 cm), lower mobile part (proximal 2/5th is jejunum and distal 3/5th is ileum).

Differences Between the Features of Jejunum and Ileum

	Jejunum	*Ileum*
1.	Long and few vasa rectae	Short and numerous vasa rectae
2.	Long plicae	Small plicae
3.	Thick wall	Thin wall
4.	Mesentery transparent	Mesentery contains fat
5.	Peyer's patches are scanty	Peyer's patches are abundant, located in the antimesenteric border
6.	Villi, leaf like and more abundant	Villi, finger like and less abundant.
7.	Proximal 40% of small bowel	Distal 60%

Mesentery attaches the small intestine to the posterior abdominal wall, contains blood vessels, lymphatics, fat.

Nerve Supply

Sympathetic supply is from T_9-T_{11} and parasympathetic supply is from vagus. Both pass through coeliac and superior mesenteric plexus.

Bowel wall contains the myenteric plexus of Auerbach which lies between the circular and longitudinal muscle coats; and submucous plexus of Meissner.

Sympathetic nerves are motor to sphincter, inhibitory to peristalsis; parasympathetic nerves are inhibitory to sphincter and stimulates peristalsis.

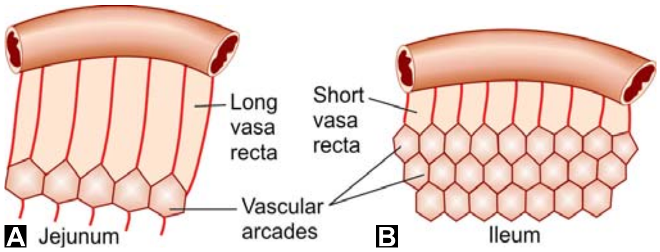

Figs 21.1A and B: Vascular pattern in (A) jejunum, (B) ileum.

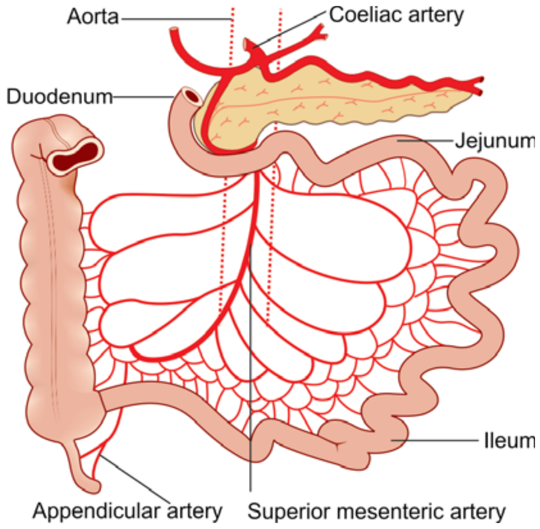

Fig. 21.2: Blood supply of the small bowel.

Normal GIT secretions

Saliva	1000 ml
Gastric	1500 ml
Intestinal	4000 ml
Bile	1000 ml
Pancreas	1500 ml

Cell type	***Function***
Goblet cells	Mucous
Paneth cells	Lysozyme, tumour necrosis factor, cryptidins
Enterocytes	Absorption
Enteroendocrine cells	Different hormones

Investigations for small bowel diseases

- Barium meal follow through
- Small bowel enema *(Enteroclysis)*
- Plain X-ray abdomen to see intestinal obstruction
- Selective splanchnic angiography to see vascular bleed or malformation (DSA)
- Isotope scintigraphy to see Meckel's diverticulum
- Estimation of faecal fat in 24 hours—normal is less than 7 gram
- *Schilling test* to find out the absorption of vitamin B_{12} from terminal ileum
- SeHCAT bile acid absorption test (selenium labelled) to find out the absorption of bile acids from the terminal ileum
- Many different breath tests are used to find out malabsorption, bacterial overgrowth and transit time
- Jejunal biopsy
- CT scans to see fistula, tumour, and spread
- Capsule endoscopy is very useful

MECKEL'S DIVERTICULUM

> *It appears at a specific site in the ileum and the wall contains each of the several layers of the intestinal tract…. The proof that the diverticulum is a residuum of the communication between the intestinal canal and the umbilical stalk rests in the findings which I have observed in three stillborn, full-term fetuses.*
>
> —**Johann Friedrich Meckel,** 1809

It is a congenital diverticulum arising from the terminal ileum and is part of the unobliterated proximal portion of the vitello intestinal duct.

It is:

- 2% common.
- 2 feet from the ileocaecal valve.
- 2 inch in length.
- 2% of Meckel's diverticulum only will be symptomatic.
- 50% of symptomatic are below 2 years of age.
- 2:1 female preponderance is seen.
- It is congenital, results from incomplete closure of vitello-intestinal duct.
- It is the most common congenital anomaly of small intestine.
- Arises from the antimesenteric border of the ileum, containing all three layers of the bowel with independent blood supply.
- In 20% of cases mucosa contains heterotopic epithelium like gastric (commonest—50%), colonic and pancreatic tissues (5%).
- It may be connected to or communicated with the umbilicus through a band or fistula.
- It may be associated with oesophageal atresia, exomphalos, and anorectal malformations.

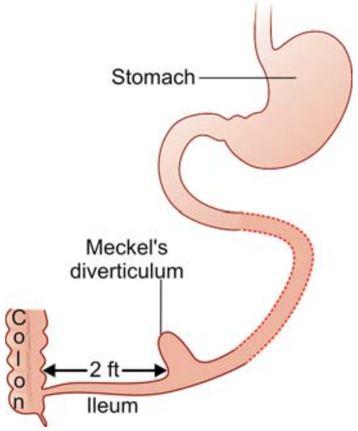

Fig. 21.3: Anatomy of Meckel's diverticulum.

Presentations in Meckel's Diverticulum

- Asymptomatic—in majority cases.
- Severe haemorrhage most common, seen in children aged 2 years or younger (*Maroon coloured blood)*.
- Intestinal obstruction due to bands/adhesions/ intussusception.
- Perforation.
- Intussusception, volvulus of small bowel.
- Peptic ulceration.
- Diverticulitis (20%)—features mimic acute appendicitis.
- *Littre's hernia*—it is presence of Meckel's diverticulum in hernial sac as content. It is observed in inguinal/femoral hernia.
- Silent Meckel's diverticulum found during laparotomy or laparoscopy or by radioisotope study.
- Carcinoid or GIST can occur in Meckel's diverticulum.

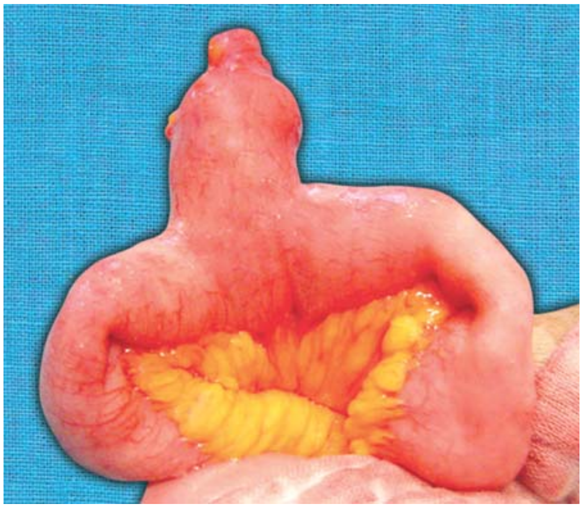

Fig. 21.4: Meckel's diverticulum.

One cannot love what he cannot respect, whether it be himself or another.

Diagnosis

- Technetium (Tc^{99}) radioisotope scan is very useful (90-95% accuracy). Ninety per cent of heterotrophic gastric mucosa can be identified in Meckel's diverticulum by radioisotope study. It can detect Meckel's diverticulum with minimal bleeding also (0.1 ml/minute). So it is very useful investigation in children presenting with bleeding.
- X-ray abdomen to see complications like obstruction, perforation.
- *Laparoscopy is very useful.*
- Enteroclysis/small bowel enema under fluoroscopy may show the Meckel's diverticulum. It is probably the most accurate investigation.

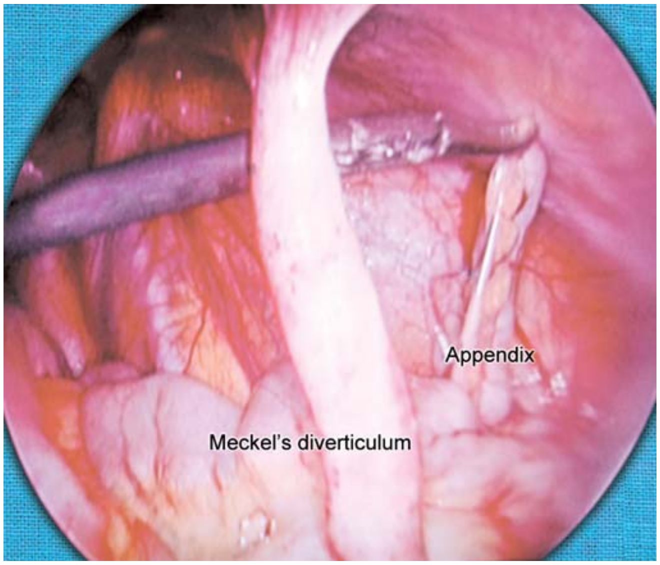

Fig. 21.5: Laparoscopic view of Meckel's diverticulum. Patient presented with features of acute appendicitis. On laparoscopy, Meckel's diverticulum which was adherent to umbilicus by a cord was found. Appendix was identified separately. Appendicectomy and Mecklian diverticulectomy was done.

Treatment

- Asymptomatic Meckel's diverticulum can be left alone when identified during laparotomy.
- *Resection of a short segment of ileum containing Meckel's diverticulum and end-to-end anastomosis is done*.
- *Meckelian diverticulectomy* with closure of enterotomy also can be done, but chances of retaining heterotopic tissues and stenosis are higher.

Indications for Surgery

- Surgery is done whenever the base is narrow, and in lengthy diverticulum.
- Presence of adhesions or band which may precipitate obstruction, intussusception or volvulus.
- Symptomatic patients or presence of complications.
- If it is found in children below 2 years.

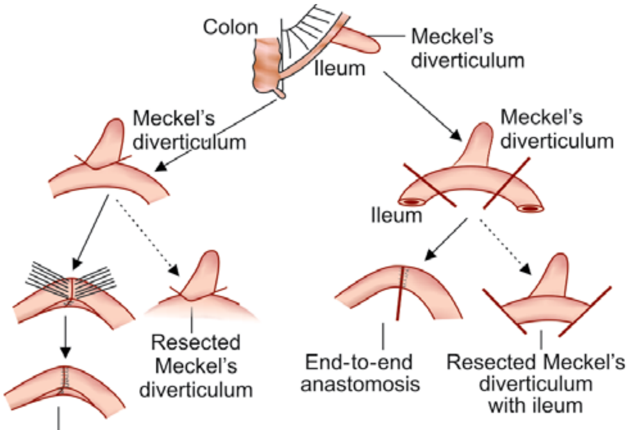
Colon
Meckel's diverticulum
Ileum
Meckel's diverticulum
Meckel's diverticulum
Ileum
Resected Meckel's diverticulum
End-to-end anastomosis
Resected Meckel's diverticulum with ileum
Meckelian diverticulectomy linear stapler can be used

Note:
Care of clearing heterotopic gastric epithelium should be taken. Should avoid stricture formation later.

Fig. 21.6: *Meckel's diverticulum*—surgical treatment. Meckelian diverticulectomy is done by obliquely clamping beyond the base of the Meckel's diverticulum. Care should be taken not to retain heterotopic gastric (or other) epithelium which can be felt like indurated area. Enough precaution should be taken so that stricture will not form later. When heterotopic tissues extend beyond the Meckel's diverticulum into the ileum, then resection of ileum with Meckel's and anastomosis is done. Clamping distal to the base of the Meckel's diverticulum like appendicectomy should not be done. It will cause stricture, leak or retaining heterotopic tissues in the part.

Note:
"Meckel's diverticulum frequently suspected; often sought; seldom found"—Charles Mayo.

- *Duodenal diverticulum* is the most common acquired diverticulum of small bowel.
- *Meckel's* is the most common true congenital diverticulum of small bowel.
- *Duodenal diverticulum*:
  It is common in females; 65% occur within 2 cm of ampulla. Commonly they are asymptomatic; < 5% need surgery. It can be congenital/acquired; true/false; intra or extraluminal. Biliary obstruction, pancreatitis, cholangitis, haemorrhage are *complications*. Diverticulectomy; duodenotomy and invagination of diverticulum when it is very close to ampulla embedded in pancreas; sphincteroplasty are the *treatment*. Perforated diverticulum is treated with serosal jejunal patch/duodenal diversion/duodeno or gastrojejunostomy. Trifoliate diverticula is an acquired one, occurs in chronic duodenal ulcer.
- *Jejunoileal diverticula*:
  It is rare (0.1%). It is false type, common in elderly. They are common in jejunum, often multiple, protrudes from mesenteric border. Chronic pseudo-obstruction, vague pain, malabsorption, haemorrhage, perforation can occur occasionally. Diverticulitis leads into perforation and peritonitis or abscess. Bile salt deconjugation with steatorrhoea, megaloblastic anaemia (vitamin B_{12} deficiency) can occur. *Treatment*—enterotomy with removal of enterolith or resection and anastomosis.

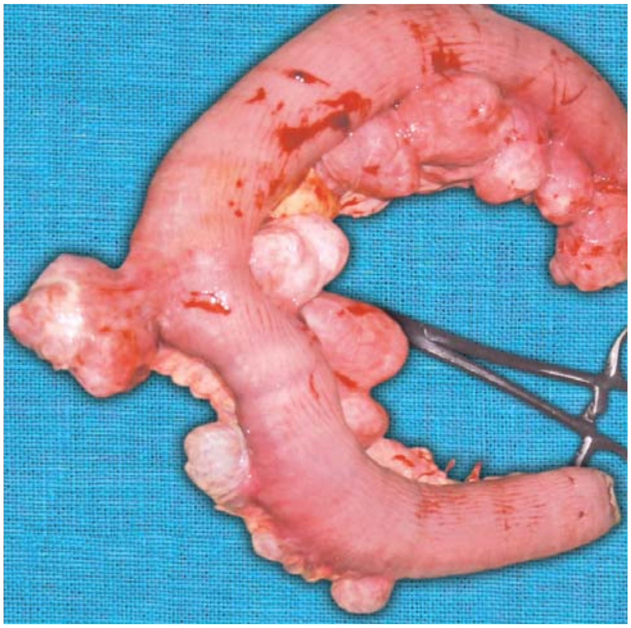

Fig. 21.7: Jejunal diverticula are rare. Often can cause obstruction, haemorrhage. It needs surgical resection.

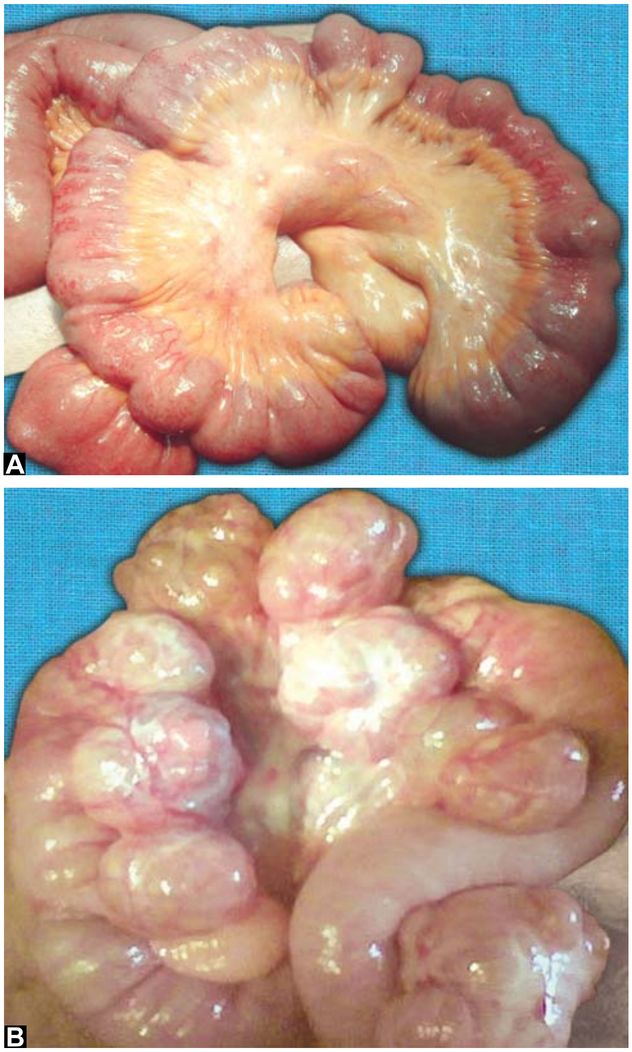

Figs 21.8A and B: Jejunal diverticula—multiple.

REGIONAL ENTERITIS (Crohn's Disease)

> *We propose to describe, in its pathologic and clinical details, a disease of the terminal ileum, affecting mainly young adults, characterised by a subcutate or chronic necrotising and cicatrising inflammation. The ulceration of the mucosa is accomplished by a disproportionate connective tissue reaction…a process which frequently leads to stenosis of the lumen of the intestine, associated with the formation of multiple fistulas.*
>
> —**Burrill Bernard Crohn, Leon Ginzburg,**
> **Gordon D Oppenheimer,** 1932

- It is a *granulomatous, noncaseating inflammatory condition of the ileum commonly and of the colon often.*
- It is independent of age, sex, socioeconomic status and geographic areas.

Aetiology

- Unknown, but a familial and infective nature is thought of.
- Increased autoantibodies.
- Diet and food allergy.
- It is slightly more common in females.
- DNA of Mycobacterium paratuberculosis was found in intestines of 60% of patient's with Crohn's disease but antituberculous drug therapy has not helped them.
- Focal ischaemia as a vasculitis may be the cause.
- 10% of first degree relatives; 50% of monozygotic twins develop Crohn's disease. Genes NOD2/CARD15 in chromosome 16q12 has got strong association with Crohn's disease. CARD15 is expressed in Paneth cells of the ileum.
- Smoking is related to Crohn's disease as aetiology, as for relapse and for exacerbations.

Causes for Crohn's

- *Infectious—Mycobacterium paratuberculosis* and atypical mycobacteria
- *Immunologic*
- *Genetic*—chromosome 16q—IBDI with CARD$_{15}$/NOD$_2$ gene (40 fold risk)
- *Environmental*
- *Jews* are more prone
- Smoking, diet, OCPs (controversial), psychosocial factors

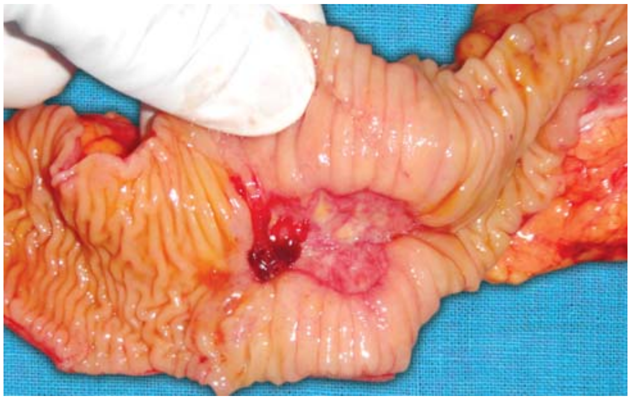

Fig. 21.9: Multiple ulcers in the jejunum—Crohn's or tuberculosis.

The successful man is the average man, but focused.

Pathology

Transmural inflammation
↓
Granuloma formation with linear snake like ulcers
↓
Cicatrisation
↓
Thickening of the bowel wall *(Hose pipe pattern)*
↓
Adhesions
↓
Fistula formation

- There is increased mucous membrane permeability → antigen induced cell mediated inflammatory response → release of cytokines like TNF, interleukin 2 → defect in suppressor T cell → granuloma and other pathology.
- Fibrosis, stricture formation, deep ulcers, oedema of mucosa between ulcer areas which looks like '*cobble stone*', skipped normal areas in between, serosal opacity, mesenteric fat stranding, enlarged mesenteric lymph nodes, abscesses in the mesentery, fistula are the pathology. Small mucosal aphthous ulcers are earliest gross feature.
- Disease may be *inflammatory, stricturing or perforating types*.
- Noncaseating giant cell granuloma with chronic inflammation of all layers; focal arterial blocks in muscularis propria are the microscopic features.
- Extensive *fat wrapping around bowel* which is been thickened, firm, rubbery, incompressible, segmental is typical.

Main features of Crohn's disease

- Ileum—most common site of occurrence—60%
- Rectal sparing is usual and common
- Skip lesion
- Hose-pipe pattern
- Linear ulcers and cobble stone appearance of mucosa
- Transmural

- Mesentery is thickened, oedematous, with enlarged lymph glands which will neither break nor calcify.
- Rarely jejunum, stomach and other parts of GIT like oral cavity, oesophagus are involved.
- In colon (30%), it is commonly observed in caecum and ascending colon.
- Toxic megacolon with acute colitis even though rare, can occur in Crohn's disease.

Note:
Anal fissure is most common anal problem in Crohn's disease. It may lead into perianal abscess and fistula.

Vienna classification of Crohn's disease

- *Age in years*:
  - A_1: < 40; A_2: ≥ 40
- *Behaviour*:
  - B_1: Nonstructuring, nonpenetrating
  - B_2: Stricturing
  - B_3: Penetrating
- *Location:*
  - L_1: Terminal ileum
  - L_2: Colon
  - L_3: Ileocolon
  - L_4: Upper GIT

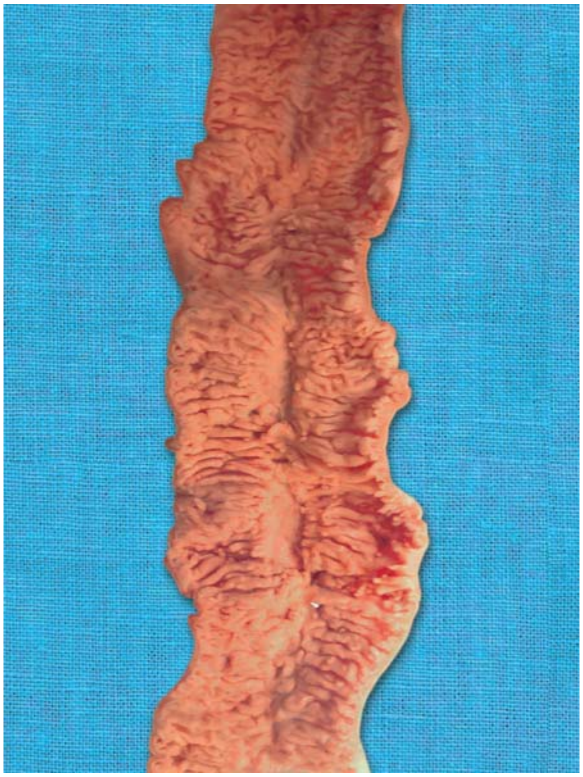

Fig. 21.10: Crohn's disease in small bowel. It is transmural full thickness disease. In 30% cases colon is involved. Skip lesion, stricture, fistula are common. It can involve any part of GIT occasionally.

Clinical Features

- It is common in young age group.
- Abdominal pain and diarrhoea is the initial slow insidious presentation. There is also asymptomatic period in between.
- Diarrhoea is usually less severe without blood, pus or mucous.
- Mild fever, weight loss, lethargy.
- Crohn's disease may present as tender, firm, resonant mass in right iliac fossa
- Obstruction, fistula formation, often perforation.
- Bleeding which is usually chronic but occasionally massive can occur.

- Perianal disease with fissure, fistula, and abscess can occur in 25% of patients with small bowel Crohn's. It can be the only presentation of Crohn's in 5% of cases. 50% of colonic Crohn's will have perianal disease.
- Extraintestinal manifestations occur in 30% of Crohn's disease.

Presentations

a. ***Acute presentations*** *(5%):*
It mimics acute appendicitis with severe diarrhoea. Often there will be localised or diffuse peritonitis.

b. ***Chronic Crohn's:***

First stage—Mild diarrhoea, colicky pain, fever, anaemia, mass in right iliac fossa which is tender, firm, nonmobile along with recurrent perianal abscess.

Second stage is either acute or chronic intestinal obstruction due to cicatrisation with narrowing.

Third stage—Fistula formation—enterocolic, enteroenteric, enterovesical, enterocutaneous, etc.

It is *precancerous condition but not as much as ulcerative colitis*.

Extraintestinal manifestations of Crohn's disease

- Skin: Erythema nodosum, pyoderma gangrenosum—most common
- Eyes: Iritis, uveitis
- Joints: Arthritis, ankylosing spondylitis
- Sclerosing cholangitis
- Nephrotic syndrome
- Pancreatitis
- Amyloidosis
- *Blood:* Anaemia, thrombocytosis, DVT, arterial thrombosis.

Investigations

- Plain X-ray abdomen, ultrasound abdomen.
- *Barium meal follow through or small bowel enema shows*—
  - Straightening of valvulae conniventes.
  - Multiple defects (*cobblestone* appearance).
  - Cicatrisation of ileum (*string sign of Kantor*).
  - *Rose thorn appearance of the bowel wall.*
  - Radiologically Crohn's disease is classified as non-stenosing type or stenosing type.
- CT scan and CT fistulogram is useful method.
- Colonoscopy usually shows normal rectum; with colon showing aphthoid like ulcers and reddened mucosal margin. Deep ulcers, stricture and fistula will be evident in late cases. Colonoscopy also shows segmental, deep, cobblestone look.
- Blood tests for anaemia, protein loss, mineral and trace element loss like magnesium, zinc, and selenium. There will be raised *C reactive protein and orosomucoid* in active disease.
- Capsule endoscopy is useful investigation.
- MRI to diagnose anal disease.
- *Serum markers*: 90% of patients with Crohn's disease show ASCA (anti saccharomyces cerviside antibody) positive and pANCA (perineural antineutrophil cytoplasmic antibody) negative, whereas in 98% of patients with ulcerative colitis, ASCA is negative but pANCA positive.

Complications of Crohn's

- Intestinal obstruction
- Stricture
- Bleeding
- Fistula formation
- Carcinoma small and large bowel
- Perianal abscess
- Peritonitis
- Pericolic abscess

Differential Diagnosis

- Radiation enteritis and *Yersinia enteritis*.
- Ulcerative colitis.
- Intestinal tuberculosis, *Salmonella*, *Shigella*, CMV
- Carcinoma ileum or caecum.
- Differential diagnosis for mass in the right iliac fossa (carcinoma caecum, actinomycosis, appendicular mass, ileocaecal TB, ectopic kidney, mesenteric lymphadenitis).

Treatment

Medical

- *Cessation* of smoking
- Bed rest, protein and vitamin supplementations. Often nasogastric tube nutrition or TPN is required.
- *Steroids* are mainly used to induce remission of the disease in initial phase. It is less useful for maintenance. Dose is 20-40 mg/day of perdnisolone (0.5 mg/kg/day) for 3-6 weeks. Methylprednisolone infusion IV 60 mg/day for 5-7 days can be given initially. *Budesonide*, a newer steroid shows high first phase metabolism in liver and so in high dose achieves targeted delivery into the intestine. It is often combined with mesalamine or used as an alternate first line therapy.
- *Azathioprine* is used for maintenance therapy. It inhibits the cell mediated immunity. 6-mercaptopurine and cyclosporine are also used. *Tacrolimus* (FK-506) inhibits production of IL_2 and is effective in improving fistula.
- *Salazopyrines* are mainly used in acute ileitis and colitis. It is not useful in inducing remission. 5-aminosalicylic acid (5-ASA) and sulfasalazine are also used. 5-ASA inhibits leokotriene, TNF, interleukin production. It acts mainly on colonic and partly on small bowel mucosa. *Mesalamine* which causes slow release of 5ASA is better with a dosage of 4 gram/day. It is released throughout its passage along the small and large bowel. Mesalamine remains the *first line therapy* for Crohn's disease.
- *Metronidazole* is useful in reducing the anal and colonic pathology by suppressing the cell mediated immunity and also as antibacterial. It is not beneficial for small bowel disease.

- Monoclonal antibody like *infliximab* is used in severe refractory cases which act against tumour necrosis factor alpha (TNFα). Single dose is used for induction. Later given after 2nd week, 6th week and then once in 8 weeks at a dose of 5 mg/kg. It is also useful in promoting closure of fistula in 60 - 80% cases. But its long term side effects are not confirmed.
- Anti IL_1, anti IL_{12}, anti IL_{18}, anti interferon γ antibodies are tried. *Natalizumab* is recombinant human monoclonal antibodies against α integrin.
- *Antibiotics* are useful in controlling sepsis in fistula, colitis.

Medical therapy

- To induce remission—steroids
- For maintenance—immunomodulating drugs like azathioprine
- Antibiotics, metronidazole (as immunomodulator)
- Monoclonal antibody—infliximab
- Nutritional support

Note:
Patients with Crohn's disease should avoid NSAIDs

Surgery

Indications

- Failure of medical treatment.
- If patient cannot be weaned off systemic steroid after 6 months.
- Intestinal obstruction—*most common indication.*
- Fistula formation, bleeding, malignant change.
- Perforation, fulminant colitis.
- Perianal problems.
- Crohn's disease children with growth retardation.

Note:
Surgery is not to cure the disease, but to correct complications. Recurrence of complications and relapse of disease can occur even after surgery. Patient should be on postoperative azathioprine/6 MP/5ASA.

Surgeries

- Ileocaecal resection (common procedure done because commonly ileocaecal region is involved).
- Segmental resection—conservative resection is better.
- Total colectomy and ileorectal anastomosis. It is only done in extensive colonic Crohn's. Ileoanal anastomosis with pouch is not done after total proctocolectomy or continent ileal pouch is not done as disease may later recur in the pouch itself.
- Stricturoplasty.
- Temporary ileostomy.
- Right hemicolectomy is done occasionally.
- Emergency colectomy may be needed in 8% of patients with extensive severe colitis but with rectal sparing.
- Laparoscopic resection is good alternative.
- Occasionally if rectum is diseased or anal disease is severe then total proctocolectomy with ileostomy is done.
- Corrective surgery for anal diseases like fissure, abscess and fistula.
- Definitive procedures for internal fistulas like ileovesical, ileocolic with faecal and urinary diversions.
- Bypass and exclusion procedures are not commonly used at present.
- In free perforation and peritonitis ileostomy is needed. But later, it is difficult to decide the timing of closure of ileostomy as disease is extensive; it is wiser to do right ileocolic resection or right hemicolectomy in these patients.
- Enterocutaneous fistula is treated with excision of fistula with resection and anastomosis of that particular segment. Organ fistulas are also treated with resection of adherent bowel and closure of organ like urinary bladder wall.
- Small bowel Crohn's causes chronic bleeding; colonic Crohn's often may cause massive bleed. In massive bleed angiographic control or colonoscopy is needed.

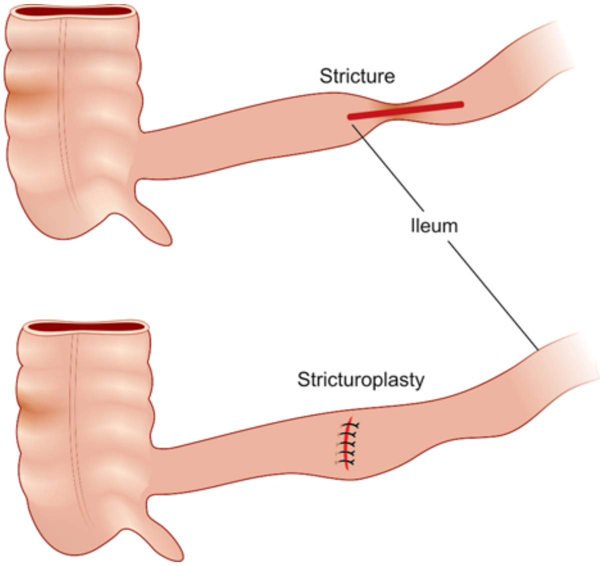

Fig. 21.11: Stricturoplasty can be done in Crohn's disease in case of short strictures in the absence of active inflammation.

Note:
Appendicectomy should not be done in acute phase of appendicitis as it may lead to external faecal fistula; if there is no acute inflammation, *elective appendicectomy can be/should be done* as it eliminates appendix as a source of future abdominal pain.

SURGICAL COMPLICATIONS OF TYPHOID

- Typhoid (Enteric fever) is caused by organism—*Salmonella typhi.*
- Mode of transmission is faecoral route.
- Presents initially with fever, chills, abdominal pain.

Typhoid

- *S typhi, S paratyphi* A, B and C
- Gram-negative flagellated bacilli
- Faecooral route is commonest mode of spread
- Asymptomatic human carriers are most important reservoirs (*Typhoid Mary*)
- Peyer's patches, mesenteric lymph nodes, gallbladder, spleen, liver and bone marrow are commonly involved

Complications of typhoid

- Paralytic ileus is the *most common* complication
- Intestinal haemorrhage—as the Peyer's patches in terminal ileum are enlarged and ulcerated—20%. Usually blood transfusion and antityphoid therapy is sufficient to control bleeding. Rarely in massive bleed resection anastomosis is needed

- Perforation
- Cholecystitis and carrier stage
- Venous thrombosis particularly of common iliac vein is known
- Typhoid cystitis, bacilluria, epididymo-orchitis
- Arthritis
- Typhoid osteomyelitis (granular sequestrum)
- Typhoid perichondritis and laryngitis leading to airway obstruction
- Typhoid abscess in liver, spleen, brain and parotid
- Myositis, myocarditis, meningitis

Typhoid (Enteric) Perforation

Perforation usually occurs in 3rd week of the infection:

- Ulcers are multiple, arranged in parallel and in antimesenteric border of the ileum.
- One or more ulcers might perforate and many ulcers may be on impending perforation.
- Patient is toxic, presents with:

- *Severe diarrhoea*
- *Relative bradycardia*
- *Soft abdomen*
- *Obliterated liver dullness*
- *Abdomen without guarding and rigidity (because of Zenker's degeneration)*
- *Initial history of fever for few days then pain abdomen and tenderness, which is progressive*

Associated features

- Splenomegaly
- Fever, headache
- Rose spots in the skin

Investigations

- Possibility of missing typhoid perforation is very high.
- X-ray in erect posture shows gas under diaphragm.
- Widal test is positive.
- Neutropenia is seen in blood.
- Blood culture (90% + ve) and stool cultures are often required. Marrow culture is useful.
- Identification of specific and sensitive markers (to detect Ig G/IgM). These tests are positive even when blood culture is negative.
- PCR assay for typhoid infection is also useful.

Note:

- Incidence of typhoid perforation is 2% of total cases.
- 75-80% of perforations are single which needs simple closure after taking biopsy from the edge of ulcer.
- 20-25% of perforations are multiple which requires either resection or exteriorisation as ileostomy.
- In paratyphoid B fever colon perforation can occur.
- CT scan abdomen is ideal investigation to detect early pneumoperitoneum and fluid collection.

Treatment

- Antityphoid drugs (Quinolones, chloramphenicol, ceftriaxone sodium) are started.
- Later laparotomy is done to close the perforation.
- Ulcers which are on impending perforation should be sutured with serosa.
- Presently, resection and anastomosis is also accepted as treatment when multiple perforated ulcers are present.
- ZIP technique (laparostomy) to reopen the abdomen whenever required is also beneficial.
- Exteriorisation of ileum is ideal in a critically ill patient. Once patient recovers properly after 6-12 weeks closure and continuity is maintained.

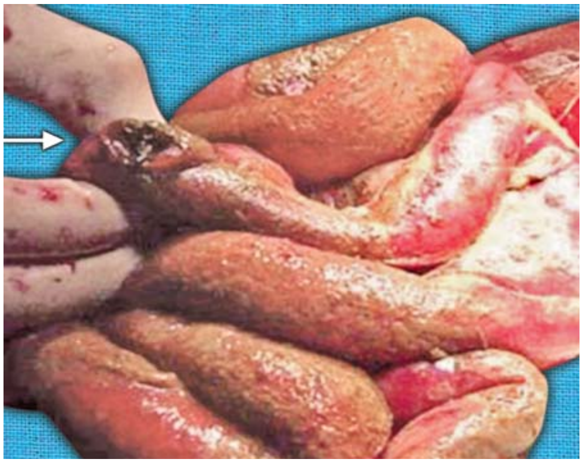

Fig. 21.12: Enteric perforation is often severe causing faecal peritonitis. Resection/exteriorisation may be needed in severe cases.

- Closure of perforation with ileotransverse anastomosis is the other option.
- ICU care, TPN/enteral feeding/monitoring/blood transfusion/ fluid and electrolyte management/ prevention or management of other complications like DVT/DIC/ARDS/septicaemia.

SURGICAL COMPLICATIONS OF ROUNDWORM *(Ascaris Lumbricoides)*

Life Cycle

Eggs → faecooral route → larvae → portal circulation → liver → hepatic veins → IVC, heart, pulmonary artery → lungs → maturation of larvae (8 weeks) → alveoli, trachebronchial tree → coughed out or swallowed to intestine → adult roundworms → male + female worms union → release of eggs → stool → reinfection.

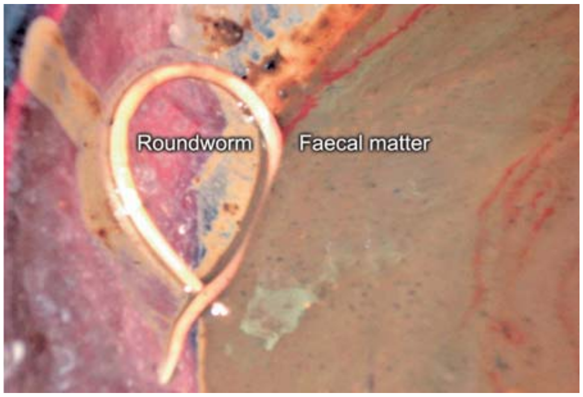

Fig. 21.13: Roundworm removed from peritoneal cavity. Note the faecal contamination.

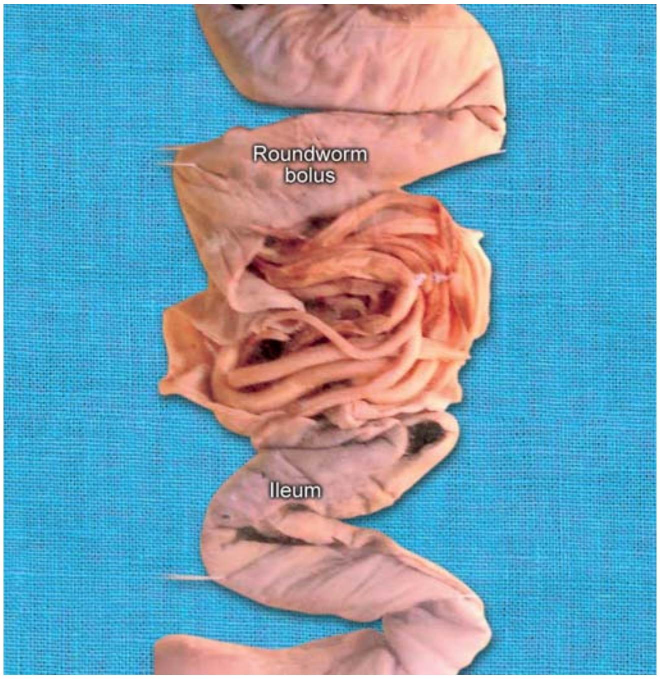

Fig. 21.14: Specimen of ileum showing roundworm *bolus*. Bolus causes intestinal obstruction, commonly near terminal ileum.

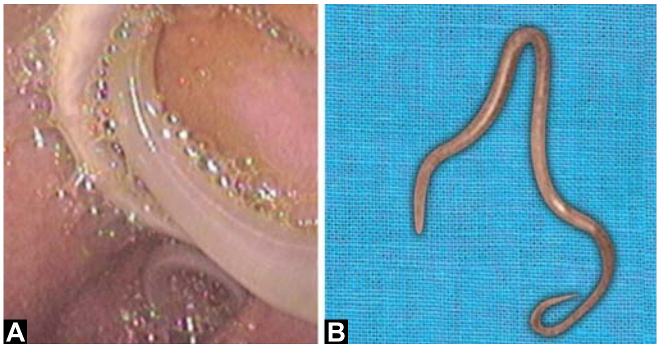

Figs 21.15A and B: Gastroduodenoscopy view of roundworm. It was removed through endoscopy. Roundworm in proximal bowel can cause pancreatitis, cholangitis. Vomiting of roundworms signifies intestinal obstruction. It need not be due to roundworm obstruction. Because of the obstruction and proximal bowel irritation, worms move proximally into stomach and come out through the mouth. (*Courtesy:* Dr Tantry and Dr Sandeep Gopal, Gastroenterologists, KMC, Mangalore).

Features

- Worm colic.
- Toxicity—fever, tachycardia.
- Subacute intestinal obstruction.
- Acute intestinal obstruction with palpable roundworm bolus per abdomen.
- Perforation—common in the ileum.

Note:

Perforation usually occurs at the site of pre-existing disease like nonspecific ileal ulcer, amoebic ulcer, typhoid ulcer, and suture line.

- Intraperitoneal abscess.
- Dyspepsia, malabsorption, iron deficiency anaemia.
- Due to migration of worm into the CBD/pancreatic duct causes ascending cholangitis with fever, jaundice and upper abdominal pain or features of pancreatitis.

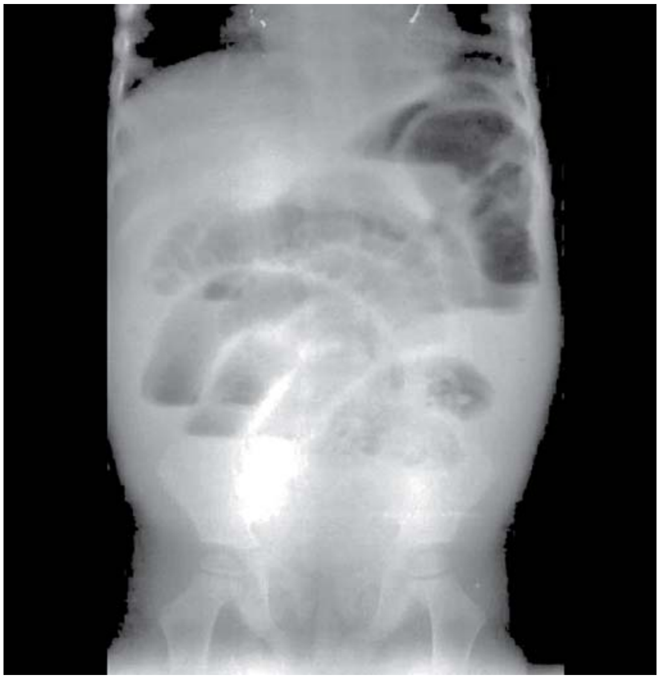

Fig. 21.16: Plain X-ray abdomen in a child showing features of intestinal obstruction due to roundworms in the bowel.

- *Loeffler's syndrome* is due to larvae in the alveoli causing fever, dry cough, chest pain, dyspnoea.

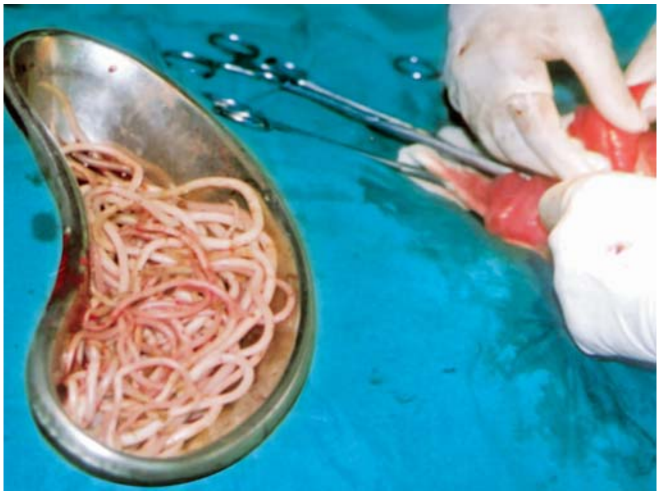

Fig. 21.17: Roundworm obstruction. Worms removed through an enterotomy.

Investigations

- Small bowel enema/barium meal follow through may show roundworms in the ileum.
- US can demonstrate the worms/worm bolus/worm in CBD or pancreatic duct.
- Blood may show eosinophilia, anaemia, hypoalbuminaemia. LFT for worm in CBD.
- Chest X-ray may show bronchitis.
- Sputum or bronchial wash may show larvae or Charcot-Leyden crystals.
- Stool examination may show ova.
- CT/MRI will show worms/obstruction/worm in CBD or pancreatic duct.

Treatment

- Antihelminthic drugs like piperazine citrate 60 ml is given immediately if there are features of obstruction, otherwise at night time. Albendazole, mebendazole are other drugs used. Proper nutrition is important.
- Nasogastric aspiration; IV fluids; observation is sufficient in most of the situations. In 48 hours, obstruction will be relieved and child passes worm bolus per anally.
- Occasionally laparotomy and milking of the worm bolus into the caecum without enterotomy is needed. Once worms are in the colon, hypertonic saline wash per anum makes them to disperse and pass outside.
- Removal of worms by enterotomy is rarely done. Enterotomy should *be securely sutured using silk otherwise reperforation* from sutured site can occur by worm toxins or by worm activity and migration.
- If there is perforation (common site is ileum) or gangrene with peritonitis, peritoneal worms should be essentially removed. Ideally ileum should be exteriorised to prevent suture line breakdown after closure due to worms. Once patient recovers bowel ends are sutured to maintain the continuity and abdomen is closed. Even though it is *ideal method*, it is not commonly performed.
- ERCP, worm extraction and stenting is done for worm in CBD/pancreatic duct.

Treatment of roundworm obstruction

- Drugs—piperazine citrate, mebendazole, albendazole
- Most often by conservative treatment, worms get dispersed and passed per anally. But patient requires nasogastric aspiration, IV fluids, antibiotics, and observation
- If patient is not responding then laparotomy is done. *Worm bolus* in the distal ileum is milked into the caecum. Often enterotomy and removal of worms is required
- Perforation due to worm requires immediate laparotomy, removal of worms and closure of perforation
- Only rarely, resection and exteriorisation is required

PNEUMATOSIS CYSTOIDES INTESTINALIS

- It is transient, thin cysts containing nitrogen in submucosa and subserosa of ileum.
- It is due to increased intraluminal pressure, which forces N_2 through the layers of the bowel due to hyperperistalsis.
- It is associated with chronic duodenal ulcer, chronic pulmonary disease, bowel obstruction.
- In neonates it is associated with necrotising enteritis.
- It is common in subserosal and submucosal planes. Cysts contain nitrogen and hydrogen.
- They can occur anywhere in GIT but commonly seen in jejunum. Later in ileocaecal and colon region.
- Mesentery and peritoneum may also be involved.
- It is equal in both sexes.
- These gas filled cysts in small bowel may rupture into peritoneal cavity to cause *sterile pneumoperitoneum*. Peritonitis is unusual.
- Abdominal pain, distension, bowel disturbances, vomiting are the presentations.
- Intestinal obstruction, haemorrhage, perforations are complications.

Treatment

- *Oxygen therapy* by 70% oxygen for 5 days or hyperbaric oxygen 2.5 atmospheres for 2 hours daily for 3 days.
- Metronidazole therapy.
- Treating the cause; resection only in refractory cases.

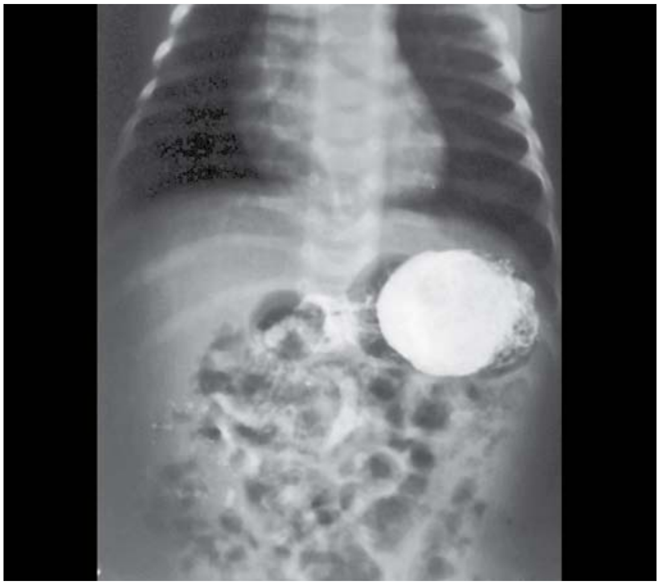

Fig. 21.18: Pneumatosis cystoides intestinalis.

Small bowel strictures

Causes

- Tuberculosis of ileum/jejunum
- Crohn's disease
- Radiation enteritis
- Ischaemic strictures—after bowel ischaemia, necrotising enterocolitis
- Nonspecific causes
- *Garrey's stricture* following long standing irreducible hernia occurs after reduction and surgery (due to constriction band causing ischaemia at the site)

Management

- Small bowel enema/capsule endoscopy/CT scan—investigations
- *Treatment*: Treating the cause; resection anastomosis, and stricturoplasty

MESENTERIC VESSEL ISCHAEMIA

Superior mesenteric artery is commonly involved than inferior mesenteric artery. Often superior mesenteric vein can also get involved.

Causes

- *Embolism* (50%)—*Sources*
    - From left auricle, as seen in atrial fibrillation.

  - A mural infarct.
  - Atheroma from aorta or aneurysm.
  - Endocarditis vegetations.
  - Left atrial myxoma.
- *Thrombosis*
  - It may block the origin of the superior mesenteric artery and can cause ischaemia of full length of small bowel. It is life-threatening.
  - It may be due to atherosclerosis or occasionally TAO.
  - Often all main splanchnic vessels—coeliac, superior and inferior mesenteric arteries may be involved by atherosclerosis.
- *Nonocclusive*
  - It is due to hypotension/hypoperfusion.
  - It is due to *vasospasm* due to shock—*nonocclusive mesenteric ischaemia (NOMI)*.
- *Superior mesenteric vein thrombosis,* occurs in patients with portal hypertension, portal pyaemia, sickle cell disease, women with contraceptive pills (OCP).

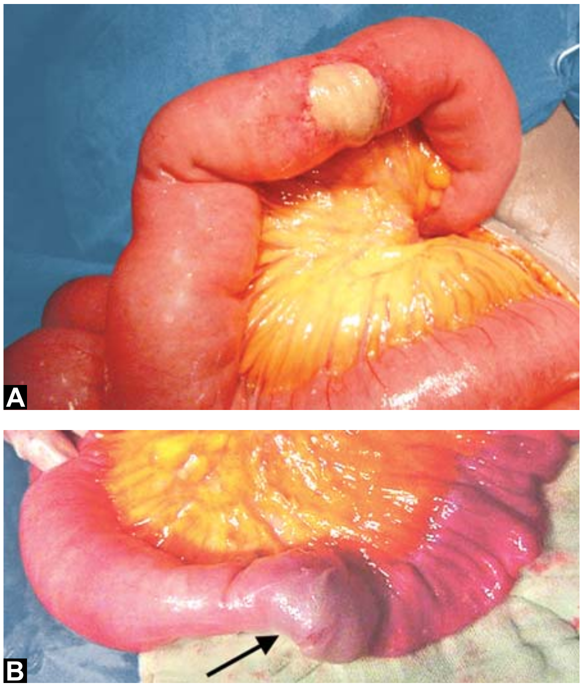

Figs 21.19A and B: Ischaemic bowel necrosis. It will lead to perforation and peritonitis. So always requires surgical resection.

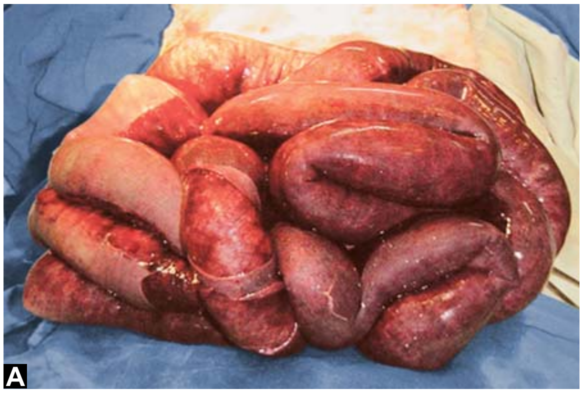

Fig. 21.20A

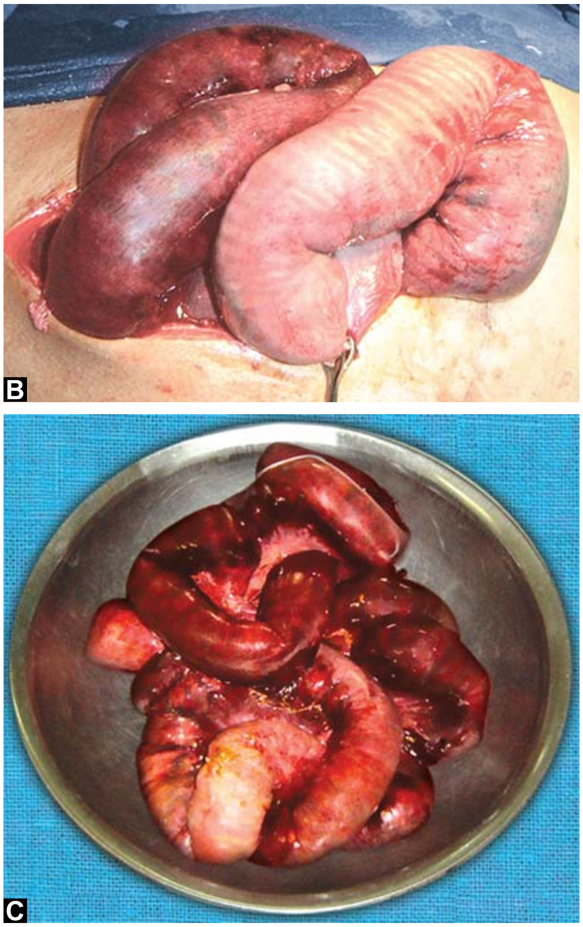

Figs 21.20B and C

Figs 21.20A to C: Acute extensive mesenteric ischaemia causing small bowel gangrene. Patient will develop short gut syndrome.

Pathology

- Bowel and mesentery will be oedematous, friable, discoloured, and collected with fluid and blood.
- Once gangrene sets in, perforation can occur leading to peritonitis.

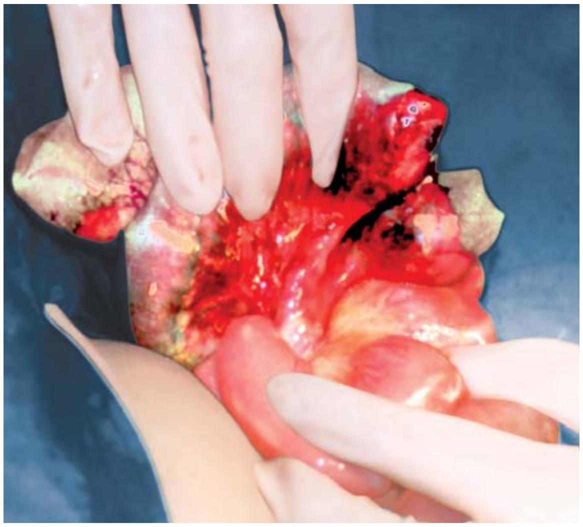

Fig. 21.21: Gangrene of small bowel due to mesenteric vascular ischaemia.

Clinical features

In acute

- *Abdominal pain that is out of proportion* in relation to tenderness; pain is around umbilicus to begin with later diffuse.
- Persistent vomiting, bloody diarrhoea; later shock and toxicity.
- Initially abdomen will be soft but later develops tenderness, rebound tenderness, distension, guarding and rigidity.
- Rectal examination shows bloody stool.

In chronic

- *Post-prandial abdominal pain* is the most important symptom with aversion to food and weight loss.
- *Abdominal angina*—recurrent colicky pain, diffuse in nature occurs with or without food intake.
- *Bloody diarrhoea*—it is a feature of both acute and chronic and is confirmed by rectal examination.
- Later, once severe infarction of bowel occurs; it may present with acute features.

Types

It can be:

- *Acute type,* sloughing of intestinal mucosa occurs in 3 hours; infarction of entire thickness of bowel occurs in 6 hours.
- *Chronic type,* it is slow and gradual process.

It can also be:

- *Extensive* involving large segment of bowel, small or large.
- *Localized:*
  - Colonic localised ischaemia at splenic flexure – commonest.
  - Focal segmental ischaemia often with necrosis in small bowel.
  - Chronic mucosal infarction.

Specific Features of Mesenteric Ischaemia

- It is often considered as intestinal failure.
- It can be *thrombotic; embolic; and nonocclusive* (low flow status—is due to reduced SMA perfusion by underlying causes like cardiogenic shock, mesenteric vasoconstriction or part of systemic disease. There is no documented thrombosis or embolus).
- It is rapidly progressive causing early mortality.
- Clostridium, anaerobic and streptococci infections are evident.
- *Gas bubble in the mesenteric vein* is pathognomonic sign of irreversible mesenteric ischaemia.
- CT/CT angiogram is very useful. An occlusion at the ostium of SMA signifies thrombosis. Embolic occlusion is smooth filling defect, usually distal to origin of middle colic artery sparing proximal small bowel, right side colon and transverse colon. Emboli can be fragments causing multiple patchy necrosis of bowel at different sites.
- Decreased blood volume; poor tissue perfusion; metabolic acidosis; myocardial depression by depressant factor; endotoxaemia; ARDS; renal failure; septicaemia and MODS are the pathological events.
- On laparotomy—*musty smell and dusky look of extensive gangrenous bowel is typical*. When in doubt, ischaemia should be confirmed by oxygen support; warming; on table Doppler.

Investigations

- Plain X-ray in erect posture and US of abdomen.
- CT scan is the investigation of choice.
- Angiogram or CT angiogram is the *most reliable* investigation especially in NOMI. It is also *ideal* for chronic ischaemia.
- Angiogram—on-table angiogram
- Doppler study—on table.
- *Blood tests*: Total count is raised with drop in haemoglobin. Raise in blood urea and serum creatinine is common. Serum phosphate level is increased very early in 4 hours as small bowel is rich in phosphate. Arterial blood gas analysis is often essential. Blood culture may be needed.

Treatment

- *Emergency laparotomy* is done. With the help of Doppler the block is identified and the vessel is opened. The block is removed and the bowel is reperfused.
- If patient has presented after 24-48 hours, gangrene might have already occurred, then resection and anastomosis is done.
- If patient presents within 6 hours, it is possible to prevent gangrene and to salvage the bowel. Emergency SMA angiography is done. Papaverine is injected for vasodilatation (by many studies its benefit is not confirmed). Often heparin (20,000 units loading dose and then maintenance dose of 5000-10000 units 6th hourly) or thrombolytics are injected. Immediate laparotomy is done. SMA is opened (arteriotomy) over the obstruction and thrombus/embolus is removed using Fogarty catheter. Perfusion is maintained. Close monitoring is essential for possibility of formation of bowel gangrene and if it is so relaparotomy should be done for bowel resection.
- In acute condition, *thrombolysis* using streptokinase, urokinase and recombinant tissue plasminogen activator can be tried. It should be done within 8 hours of onset of pain.
- NOMI is often treated with *selective infusion of vasodilator* ***papaverine*** into SMA. It is also used in all types of mesenteric ischaemia. Catheter is placed into SMA under guidance; 45 mg of papaverine given as a bolus initially, later 30 mg/hour continuous infusion with 1 mg/ml concentration. Catheter is usually kept for 2 weeks and repositioning with regular angiographic confirmation is needed. Papaverine being a vasodilator increases the bowel perfusion through end arteries as well as collaterals.
- *Chronic mesenteric arterial ischaemia* is treated with *surgical revascularisation* using aortomesenteric bypass graft and mesenteric endarterectomy. It also can be treated with *PTA with or without stenting*. PTA should not be done if there is bowel infarction. It is done through transfemoral or transaxillary route; balloon should be longer than the lesion and 10% wider than the arterial diameter.

A fault once denied is twice committed.

- *Chronic mesenteric vein thrombosis* is treated with anticoagulation or propranolol or oesophageal variceal banding or portocaval shunt.
- Eventually patient requires *oral anticoagulant therapy* using warfarin for 12 months.
- Repeated *post-therapy angiography* is needed in all patients.
- Steroids are given to reduce lysosomal membrane destruction and effects of endotoxins; to maintain arterial resistance and complement activity. High dose hydrocortisone is used.
- Higher generation antibiotics; critical care; ventilator, fluid, electrolytes and nutritional support (TPN) are essential.

Note:
- Minimum bowel length required to be retained is 1.2 meters, otherwise the patient will have high mortality (due to short bowel syndrome).
- Small bowel transplantation is under trial for the same.
- Mesenteric bowel ischaemia has got very high mortality.
- Development of septicaemia, ARDS, DIC, postoperative leak are common.

NECROTISING ENTEROCOLITIS

- Necrotising enterocolitis is an acquired inflammatory disease commonly seen in infants and newborn but occasionally can occur in children and adults. It is more commonly seen in premature babies. It is more common in formula fed babies than breastfed babies. Reduced gut flora make virulent pathogens to act and cause sepsis.
- Common site is terminal ileum, caecum and ascending colon. Often it can involve entire small bowel.
- Gas in the bowel wall and often in portal vein is typical.

Incidence

- Often associated with low birth weight babies.
- Preterm babies 90%.
- Term babies 10%.

Pathogenesis

Hypoxia
Hypothermia
Hypotension
Hyperviscosity
Acidosis

Umbilical artery cannulation
Exchange transfusion
Hyperosmolar feeds
Packed cell transfusion
Over dosage with calcium antagonists

↘ ↙

Mesenteric ischaemia.
↓
Bacterial invasion of mucosa and bowel wall.
↓
Necrosis—skip lesions are common.
↓
Necrotising enterocolitis.

Sites

- Common in small bowel.
- Terminal ileum and caecum.

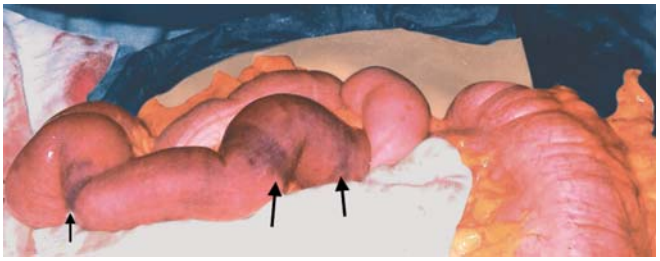

Fig. 21.22: Necrotising enteritis. Note the segmental involvement of the disease.

Clinical Features

Most babies present with:
- Bilious vomiting.
- Abdominal distension, guarding and rigidity.
- Bradycardia.
- Apnoea, lethargy.
- Gross or occult blood in stool (75%).
- Erythema and oedema of anterior abdominal wall—late feature.
- DIC.
- Mortality is 20%.

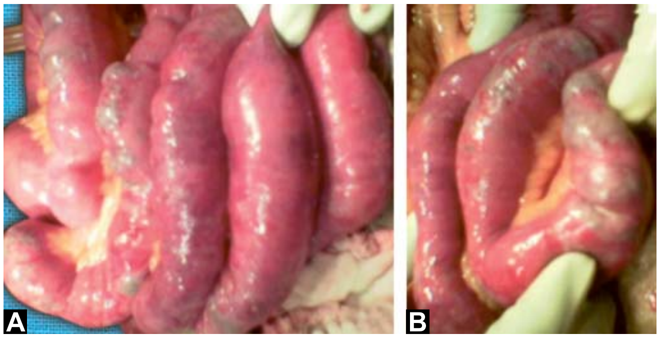

Figs 21.23A and B: Necrotising enterocolitis. Skin necrotic lesions are typical. It is common in infants and children but can be seen in adolescents occasionally (*Courtesy:* Dr Rupen, MS, KMC, Mangalore).

Investigations

a. Plain X-ray abdomen:
   - Linear gas shadow in bowel wall.
   - Free intraperitoneal air implies perforation.

b. Blood examination—leukocytosis, anaemia, hyperkalemia, metabolic acidosis.

Complications

- Stricture—small bowel.
- Perforation—peritonitis.
- Colonic stricture can develop lately after healing of ischaemic area. It is seen in 10% of cases.

Management

I. ***Medical therapy***
   - Nasogastric aspiration.
   - Broad spectrum IV antibiotics.

- TPN for 10-14 days.
- Oxygen supplement is required.

II. ***Surgery***

Indications

1. Pneumoperitoneum.
2. An abdominal mass.
3. Dilated intestinal loops.

Procedures done

- Bowel necrosis and perforation—resection with stoma.
- Localised disease—resection and primary anastomosis.

Complication: Extensive bowel resection—short bowel syndrome.

SMALL BOWEL TUMOURS

- They are rare neoplasms—3% of all GI tumours (Even though 80% total length and 90% of mucosal surface area of the GIT is small bowel).
- Common in elderly men.
- Common in New Zealand and Hawaii.
- It can be benign or malignant.
- Early diagnosis is difficult.
- Presentations are vague initially.

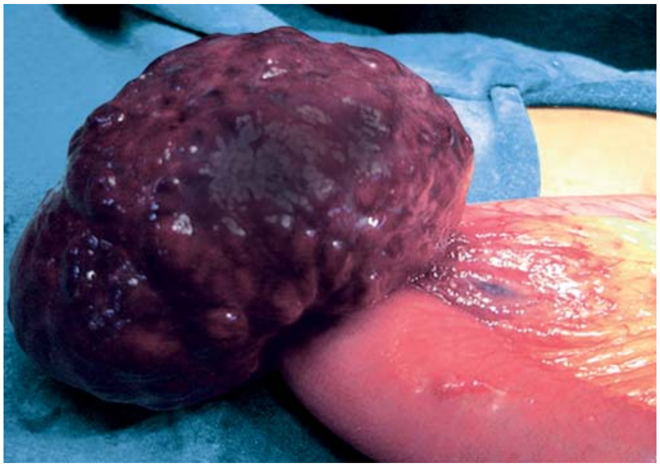

Fig. 21.24: Jejunal haemangioma. Surgical resection cured the condition.

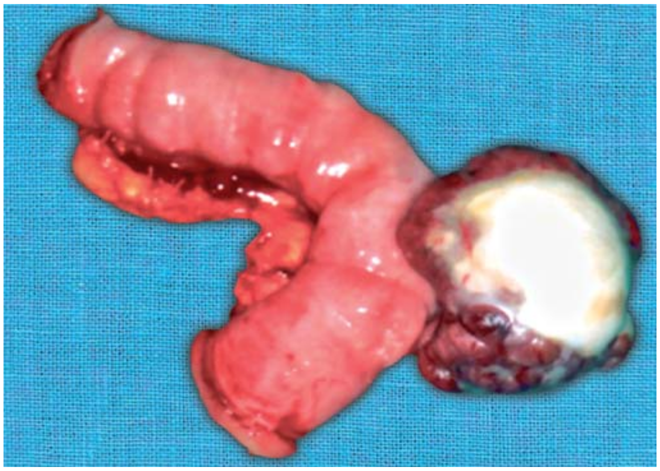

Fig. 21.25: Jejunal tumour.

Reasons why malignancy is uncommon in the small bowel (even though it comprises 75% of length and 90% of GI mucosa):

- Rapid transit time, 30 minute to 2 hours. So exposure of mucosa to toxins and metabolites is less.
- Alkaline mucus rich luminal content is protective.
- Cells of small bowel produce the enzyme benzopyrene hydroxylase which detoxifies the carcinogen—benzopyrone.
- High levels luminal IgA provides immunity.
- Plenty of lymphoid tissue in the wall provides immunity.
- Healthy small bowel has got less bacterial load and so their toxic metabolites (as compared to colon).

Risk factors implicated are:

- Bile acids and their metabolites.
- Post-cholecystectomy status.
- Familial adenomatous polyposis (FAP) especially with duodenal adenomas has got very high chance of developing adenocarcinoma.
- Crohn‘s disease has got 100 times increase in incidence of adenocarcinoma.
- Celiac disease increases the risk of lymphoma.
- Peutz-Jegher’s syndrome has got increased risk of small bowel adenocarcinoma.
- Patients with chronic immunosuppression have very high-risk of developing NHL (PTLD—post-transplant lymphoproliferative disorder, HIV) or sarcoma.
- Patient with von Recklinghausen’s disease of neurofibromatosis is prone to develop GI neurofibroma or neurofibrosarcoma.
- Smoking, red meat, alcohol, salt food.

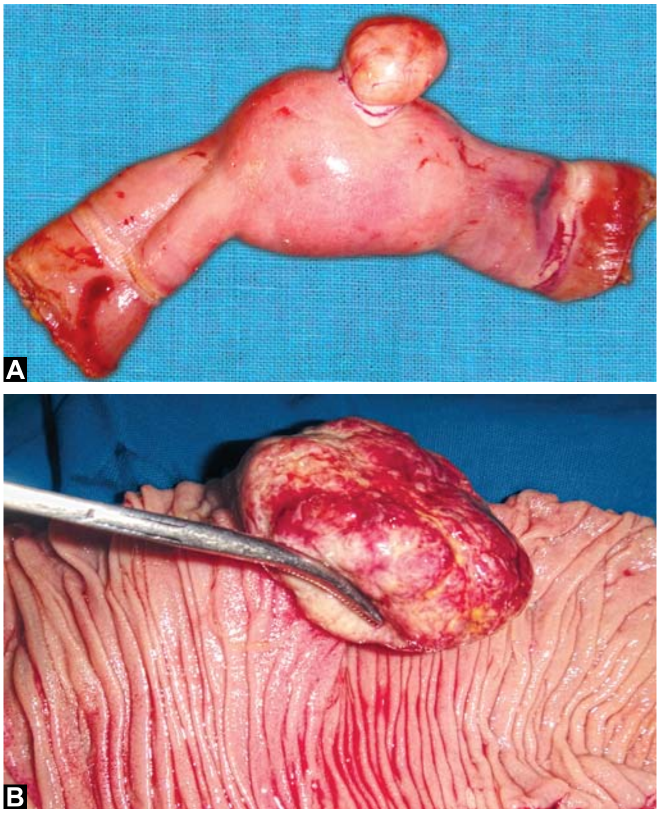

Figs 21.26A and B: Jejunal tumour coming out of serosa.

Presentations

- Asymptomatic initially.
- Features of obstruction/intussusception/bleeding.
- Vague abdominal discomfort.

Investigations

- Often it is very difficult to assess small bowel. It may be on table finding during surgery, while presenting as complication like obstruction.
- Small bowel enteroclysis.
- CT abdomen is better investigation to assess bowel/ nodes/ organs. *CT enteroclysis* shows 90% accuracy.
- Video capsule endoscopy enables to visualise small bowel mucosa properly.
- CT angiography is useful in identifying vascular tumours.
- Enteroscopes (push type or Sonde pull enteroscopes) are technically difficult.
- Intraoperative enteroscopes are useful.

BENIGN TUMOURS OF SMALL BOWEL

- Benign tumors of small bowel are 50% of primary small bowel tumors.
- Adenoma can be potentially malignant.

Types

Leiomyoma or GIST

- It is the most common symptomatic benign tumour of small bowel.
- It arises from interstitial cell of *Cajal*. It can be spindle cell type (70%) or epithelioid type.
- Commonly they are benign in small bowel.
- GIST express CD_{117} (90%), the c kit proto oncogene protein membrane receptor for stem cell growth factor and CD_{34} (80%), a human progenitor cell antigen.

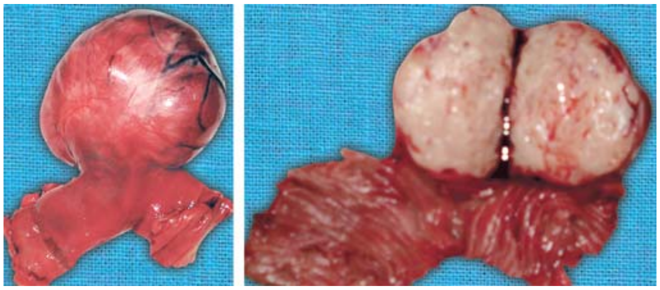

Fig. 21.27: Leiomyoma of jejunum.

Adenomas

- It is 15% of all benign small bowel tumours.
- 50% are in ileum; 30% in jejunum; 20% in duodenum.
- Commonly single but can be multiple when familial or associated with FAP.
- Commonly presents as bleeding and obstruction.

Adenomas can be:

- *Brunner gland adenoma*
  - It occurs as benign hyperplastic tumour in proximal duodenal submucosa which secretes alkaline bicarbonate rich fluid.
  - Bleeding is the usual problem.
  - It is treated by endoscopic resection.
  - It never turns into malignancy.
- *Villous adenoma*
  - It can occur anywhere in small bowel presenting as intestinal obstruction or haemorrhage.
  - But commonly seen in periampullary region presenting as obstructive jaundice, upper abdominal pain.
  - Size more than 3 cm can lead into adenocarcinoma with an adenoma—carcinoma sequence.
  - EUS is ideal tool to assess size and depth.
  - ERCP is done to relieve obstruction, to biopsy and to remove if it is small in size.
  - Malignant potentiality is very high—50%.
  - Transduodenal excision/pancreaticoduodenectomy are the surgical options. Small bowel adenomas are treated by reception.
- *True adenomas*
  - It can be tubular or tubulovillous.
  - It is usually single.
  - Endoscopic excision is the treatment; resection/pancreaticoduodenectomy is done if there is invasive lesion.
- *Familial adenomas*
  - It is associated with FAP with 5% risk for adenocarcinoma in duodenum.
  - It is diffuse throughout the duodenum.
  - *Spigelman classification* is used for assessment of FAP related duodenal adenomas. Villous adenomas larger than 3 cm or suspicious lesions identified on endoscopy are biopsied.
  - *Screening endoscopy* to be done every year.
  - Endoscopic or open polypectomy; argon beam or photodynamic therapy is tried in benign lesions.
  - Pancreaticoduodenectomy is done in high grade, carcinoma in situ, Spigelman IV classification.

Spigelman classification of duodenal adenomatosis

Parameter about polyps	Point 1	Point 2	Point 3
Number	1-4	5-20	More than 20
Size	1-4 mm	5-10 mm	More than 10
Histology	Tubular	Tubulovillous	Villous
Dysplasia	Mild	Moderate	Severe

Stage	**Points**
0	0
1	1-4
2	5-6
3	7-8
4	9-12

Lipoma

- It mainly causes intussusception.
- It is common in *elderly men*.
- It is usually *single intramural submucosal* lesion.
- Lipoma *does not* have malignant potential.

Peutz-Jegher's Syndrome

- Autosomal dominant condition with melanotic pigment patches and GI polyps.
- Brown black pigments of 1-2 mm diameter in circumoral face, cheek, forearms, palms, soles, digits, perianal region.
- Hamartomas are located in entire jejunum and ileum; 50% of colorectum; 25% of stomach.
- Intussusception, bleeding, anaemia—presentations.
- Cancer of small intestine, stomach, pancreas, ovary, lung, uterus and breast can occur.
- Cure is not possible. Segment which is causing complication is resected.

Haemangioma

- It is 4% of all small bowel benign tumours.
- They are multiple in 60% of patients.
- They are usually submucosal abnormal proliferative vessels.
- Jejunum is the commonest site.
- They are often seen in association with *Osler Weber Rendu disease, Turner's syndrome*.
- It presents as small bowel bleed.
- It can be diagnosed by angiography, $^{99}m_{Tc}$ RBC scanning or capsule endoscopy.
- *Treatment*: Endoscopic sclerotherapy or angiographic embolisation or resection of bowel segment.

General Features

- They cause haemorrhage, colicky pain, and intussusception.
- Diagnosis is by suspicion. Small bowel enema, capsule endoscopy, radioisotope study, CT scan or MRI may help in identifying the disease. Laparoscopy and proceed may be useful.
- Commonly it is on table finding during exploration for acute presentations.
- Adenomas are potentially malignant depending on the size (> 3 cm).

Treatment

- It depends on type, number, nature, size of the tumour.
- Resection and anastomosis cures the condition.
- Surgery is not done for Peutz-Jegher's syndrome unless it presents with complications.

MALIGNANT TUMOURS OF SMALL BOWEL

1. Adenocarcinoma

- Its incidence is 40% of small bowel tumours.
- It is most common primary malignant small bowel tumour.

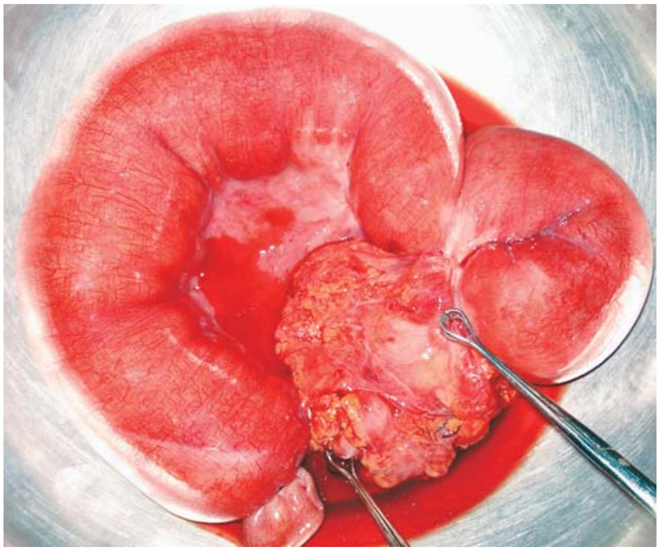

Fig. 21.28: Resected small bowel specimen showing small bowel adenocarcinoma.

- In 80% cases it is in the duodenum and jejunum.
- FAP, adenoma, Crohn's, coeliac diseases are the causes.
- Adenocarcinoma in *Crohn's disease* occurs in younger age group and commonly in ileum (70%).
- Nonspecific features, anorexia, crampy pain abdomen, bleeding and diarrhoea, obstruction or features of metastases (liver).
- CT scan, capsule endoscopy are the investigations.
- TNM system is used for staging.
- Nodal spread is the important prognostic factor.
- Surgical resection is the choice of therapy. *Pancreaticoduodenectomy* is done for duodenal tumour; for ileal/jejunal site tumours *radical resection* with 10 cm margin with adjacent mesenteric clearance is done. Adenocarcinoma of terminal ileum is treated with *right hemicolectomy*.
- RT and chemotherapy are less beneficial.
- It carries a poor prognosis.

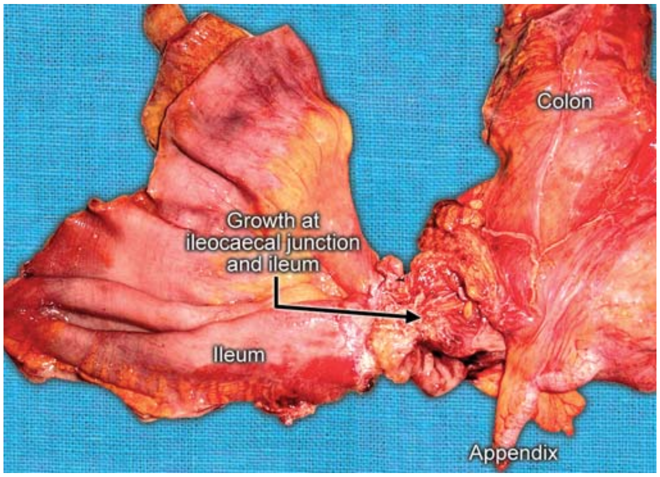

Fig. 21.29: Carcinoma ileum near ileocaecal junction. It (adenocarcinoma of ileum) carries poor prognosis. It is often on table diagnosis.

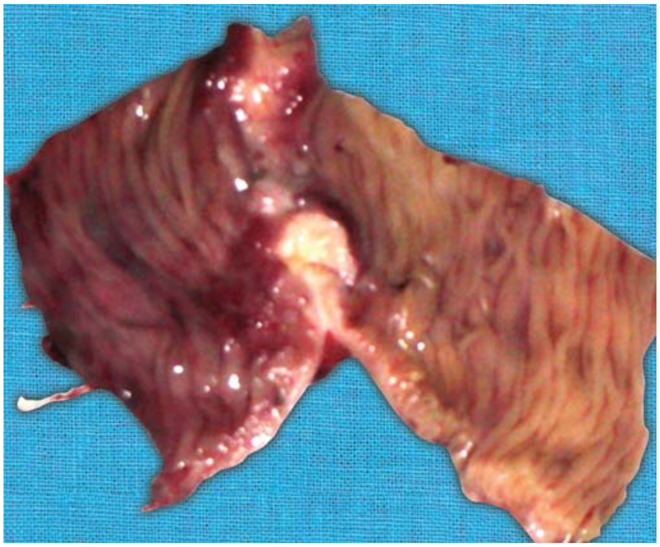

Fig. 21.30: Jejunal adenocarcinoma presenting as stricture with features of proximal small bowel obstruction.

2. Non-Hodgkin's lymphoma (NHL)—25%

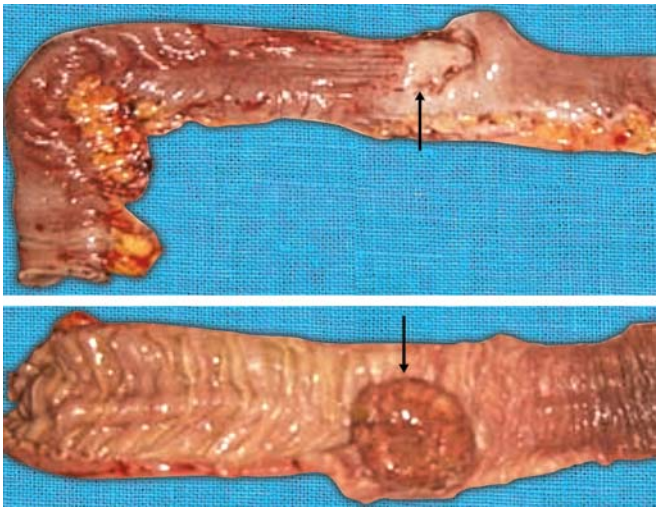

Fig. 21.31: Gross look of small bowel lymphoma.

- GI is the commonest extranodal site of NHL (20%), with small bowel (primary intestinal NHL) as second common site (30%) for extranodal site of NHL (First common site is stomach—60%).
- Lymphadenopathy/mediastinal lymph node enlargement are absent. Normal spleen, liver, blood peripheral smear are observed.
- B cell type is the commonest type 75%. It is common in ileum.
- T cell type (25%) carries poor prognosis.
- Presentations are malabsorption, obstruction, perforation, haemorrhage or palpable mass.
- In children lymphomas are the most common intestinal neoplasm.
- 25% of patients develop perforation.
- Fever when present suggests systemic spread.
- CT scan with CT guided biopsy or laparoscopic biopsy is needed.
- Surgical resection and chemotherapy are the treatment.
- Prognosis is poor.

3. Carcinoid tumour—30%.

4. GIST (gastrointestinal stromal tumour)

- It is rare but most common nonepithelial small bowel tumours.
- 25% of GIST occurs in small bowel (stomach—50 to 60%). It is 0.2% of all GI tumours.
- It is equal in both sexes.
- It arises from interstitial cell of Cajal.
- More than 95% show c-kit (transmembrane receptor tyrosine kinase) mutation.
- Many mesenchymal tumours of small bowel are now classified as GIST. GIST attains massive size when presenting clinically.
- Palpable mass, compression, haemorrhage are the features; has *less affinity* for lymphatics.
- CT is diagnostic. Histochemistry and tumour markers are needed.
- GIST can be low risk or high risk group based on tumour size and mitotic index.
- Surgical wide resection is the treatment.
- Imatinib mesylate a KIT kinase inhibitor is used successfully. SU11248 with similar effect is also used.

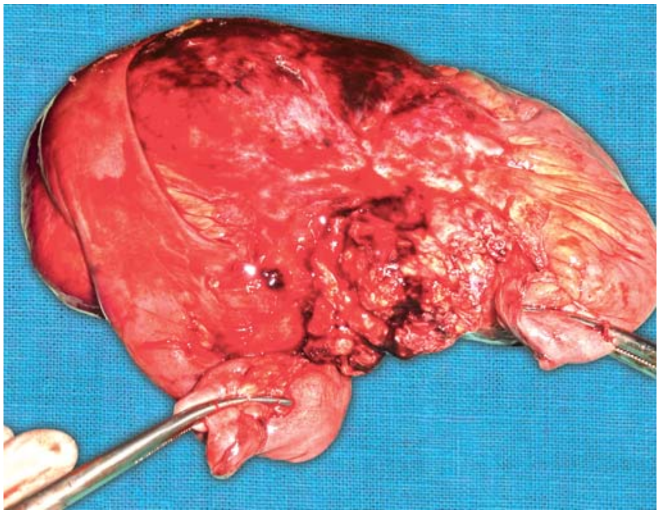

Fig. 21.32: GIST of small bowel (*Courtesy:* Dr Chinivalar, MS, Surgeon).

5. Liposarcoma and myxoliposarcoma.

6. Secondaries in small bowel

They are very rare—if it is present, primary being commonly melanoma.

Malignant tumours of small bowel

- Adenocarcinoma is commonest—40%. Order of occurrence—duodenum, jejunum and ileum
- GIST is new entity comprising smooth muscle tumours also
- Carcinoid and lymphoma has got better prognosis
- Adenocarcinoma and GIST has got poor prognosis
- Presentations are pain, diarrhoea, weight loss, anaemia, melaena, palpable mass, intestinal obstruction and metastatic features
- Bleeding and perforation are common in lymphoma and sarcoma
- Enteroclysis/capsule endoscopy/CT scan/histochemistry are needed to diagnose
- Surgery is the main therapy
- Imatinib is used in GIST with favourable results

CARCINOID TUMOUR

- The term *'Karzinoids'* was coined by Oberndarger in 1907.
- They arise from the enterochromaffin cells (Kulchitsky cells) found in the crypts of Lieberkuhn.
- These cells are capable of APUD (Amine precursor uptake and Decarboxylation) and can secrete vasoactive peptides.
- It commonly occurs in appendix (45%), ileum (25%) and rectum (15%). Other (15%) sites are—other parts of GIT (including pancreas and biliary tract), bronchus and testis. In the ileum it is almost always in the terminal 2 feet.
- Metastasis occurs in 3% of appendiceal carcinoid; 35% of ileum.
- 75% of carcinoids are less than 1 cm and 2% of them will spread; carcinoid of 1-2 cm shows 50% chances of spread; more than 2 cm shows 85% chances to spread.
- 75% of carcinoids are asymptomatic and found incidentally.
- It secretes amines (**5 HT**, 5 HIAA, 5 HTP—85%, histamine, and dopamine), tachykinins (kallikrein, substance P, neuropeptide k), peptides (**chromogranins**—100%, pancreatic polypeptide—40%, neurotensin, HCG α, HCG β, motilin), prostaglandins.
- 10% of cases are associated with MEN syndrome type I.
- Pathologically it is smooth, firm, yellowish submucosal nodule seen in antimesenteric border of bowel with mesenteric nodal mass having desmoplastic reaction.
- Carcinoid in appendix is usually single. But commonly it causes luminal obstruction and so presents with features of appendicitis. Common site is at tip/distal 2/3rd.
- Small bowel carcinoids are multiple in 40% cases. In 50% of cases other primary malignancy is observed like of breast and colon.
- Small bowel carcinoids (jejunoileal) < 1 cm incidence, of nodal and liver spread is 20-30%. If it is 1-2 cm, nodal spread is 60-80% and spread to liver is 20%. If it is > 2 cm, nodal spread is more than 80% and spread to liver is 50%.

Types of Small Bowel Carcinoids

1. *Foregut carcinoids* (Bronchial, thymic, gastroduodenal, pancreatic): Produce low levels of serotonin, but may secrete 5 HPT (5 hydroxy tryptophan) and ACTH.
2. *Midgut carcinoids* (Jejunal, ileal, appendiceal, right colic): Secretes high levels of serotonin.
3. *Hindgut carcinoids (*Distal colon, rectum): Rarely produce serotonin, but produce somatostatin and peptide YY. They present as submucosal nodules without ulcer. Hormonally inactive.

Foregut carcinoids—argyrophillic (cells can be stained with metallic silver only in the presence of reducing agent)
Midgut carcinoids—argentaffin (stained with metallic silver without a reducing agent) and argyrophillic both
Hindgut carcinoids—do not stain with silver

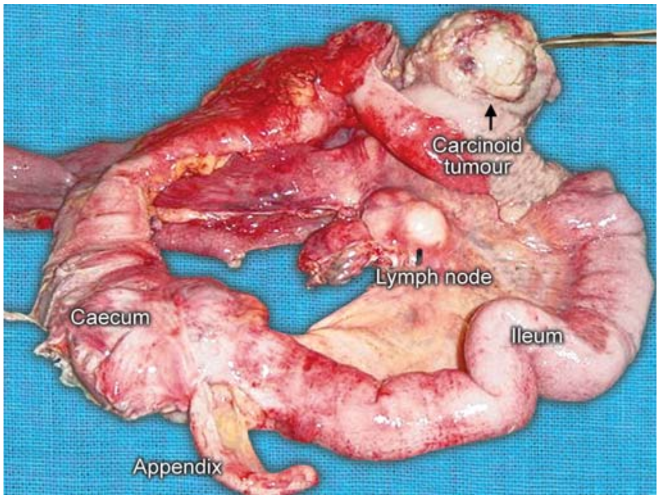

Fig. 21.33: Carcinoid of ileum. It often presents when liver secondaries develop as carcinoid syndrome. Resection and anastomosis is the treatment.

Features of Small Bowel Carcinoids

- Primary tumour is usually small < 1 cm.
- Age: Seen in 50-60 years.
- May be multicentric in small bowel—40%.
- May coexist with synchronous adenocarcinoma of small or large bowel (10%).
- Highest propensity to metastasise to liver to produce carcinoid syndrome.

Clinical Features

- Most often asymptomatic—an incidental finding.
- May present with abdominal pain, features of intestinal obstruction, diarrhoea.
- Hindgut carcinoids present with constipation, bleeding per rectum, rectal tenesmus.
- Once secondaries develop in the liver (which is yellowish) ***Carcinoid syndrome*** develops (10%), which is due to release of 5-HT, kinins, prostaglandins, histamine and indoles causing flushing, diarrhoea, cyanosis, asthmatic attacks, hepatomegaly, cardiac lesion on right side.
- *Attacks can be induced by alcohol.*

Investigations

- *Urine shows increased 5-hydroxyindoleacetic acid (5HIAA) levels. (Normal value—2-8 mg/24 hours).*
- ^{111}In-octreotide scintigraphy to detect the tumour.
- CT scan is very useful for evaluation.
- I_{131} MIBG scan is also done.
- *Plasma level of chromogranin A* will be elevated in 80% of patients.
- Provocative tests using pentagastrin, calcium or epinephrine may be used.
- PET scan using 1C 5 HTP and 18 F L DOPA isotopes when fused with CT scan will give the best image; with urine 5 HIAA and serum chromogranin A, diagnosis will be accurate.

Our senses don't deceive us; our judgement does.

Treatment

- If it is in the tip of the appendix (*tip is common*), appendicectomy and regular follow up with urine 5HIAA is sufficient.
- If it is in the base of appendix or appendicular lesion > 2 cm size or terminal ileum, right hemicolectomy is required.
- In small bowel, if the primary tumour is < 1 cm, with no lymph nodes, then segmental intestinal resection and anastomosis is sufficient. Radical resection of bowel is necessary in case of large, multiple carcinoids, with involvement of lymph nodes.
- In liver secondaries, along with surgical debulking hepatic resection, hepatic artery ligation or embolisation is tried.

Surgical treatment

- *Appendix*
  - Tip/lesion less than 2 cm but not involving base—appendicectomy
  - Lesion more than 2 cm/involving base—right hemicolectomy
- *Gastroduodenum*
  - Less than 1 cm—endoscopic resection
  - More than 1 cm—subtotal gastrectomy/pancreaticoduodenectomy
- *Small bowel lesion*
  - Less than 1 cm—segmental resection
  - More than 1 cm—radical resection with adjacent mesentery
  - Terminal ileum—right hemicolectomy
- *Rectal lesion*
  - Less than 1 cm—endoscopic resection
  - 1 cm invasive—wide excision
  - More than 1 cm—anterior resection

Medical treatment

- Mainly symptomatic
- Long acting somatostatin analogue—octreotide can be given—90% symptom palliation is achieved
- Slow release formulation of octreotide (*Sandostatin LAR*) is also used
- *Pasireotide* is 2nd generation octreotide which has got wide somatostatin receptor inhibition action (40 fold than octreotide) is said to be very effective
- Ketanserin (to control diarrhoea), cyproheptadine, ondansetron are also effective
- Bevacizumab (antiangiogenesis factor), sunitinib (Tyrosine kinase inhibitor), everolimus (m TOR inhibitor) are other agents tried
- Others: Serotonin antagonists, antihistamines, alpha methyl dopa, 5-fluorouracil
- Bromocryptine, streptozocin, doxorubicin, dacarbazine, 5 FU, methylsergide, diphenoxylate hydrochloride, interferon
- Radiolabelled somatostatin analogue
- Indium 111 labelled pentetreotide is also effective

Note:

- Methysergide causes retroperitoneal fibrosis and so not used now.
- Anaesthesia may precipitate carcinoid syndrome during surgical treatment; IV octreotide infusion 50 μg/hour with steroid and antihistamines are needed in such patients.

Carcinoid Syndrome

- It occurs in 10% of carcinoid tumours.
- It is common in small bowel carcinoid, but can occur in other sites also—bronchus, pancreas, testis, ovary.
- It is the (first pass metabolism site) *liver secondaries* which cause the *carcinoid syndrome*. Carcinoids of ovary, testis and retroperitoneum bypass the first pass metabolism in liver and so causes syndrome without liver secondaries. Rarely carcinoid syndrome develops in noncarcinoid malignant tumour and dermatomyositis.
- Syndrome is due to release of 5 HT, 5 HTP, histamine, dopamine, bradykinin, prostaglandin, kallikrein and substance P.
- **Presentations**
  - *Vasomotor:* Flushing (80%) is the *commonest manifestations* which may be short lived, *diffuse, erythematous* involving face, neck and upper chest; longer, often permanent *cyanotic violaceous flush* with watery eyes; *prolonged flush* for 3 days affecting entire body with oedema face and hypotension; *red patchy flush* seen in carcinoid stomach.
  - *GIT:* Diarrhoea (75%) is next common which is episodic, watery, explosive occurs after meals and is due to raised serotonin level. Hepatomegaly (70%) is common.
  - *Cardiac:* Pulmonary stenosis (90%), tricuspid insufficiency (45%), tricuspid stenosis (40%) are features of right valvular fibrosis—due to serotonin. Bronchospasm (asthma—25%) is due to serotonin and bradykinin.
  - *Diversion of dietary tryptophan causes malabsorption and pellagra.*
- Symptoms are precipitated by intake of alcohol, cheese, chocolate and red wine.
- Severe bronchospasm, circulatory collapse, when not diagnosed can lead into '***carcinoid crisis***'.
- *Treatment*
  - Assessment by 5 HIAA, and other methods mentioned above.
  - IV octreotide infusion, management of circulatory failure effectively using critical care.
  - Treatment of hepatic secondaries—embolisation, chemotherapy, TACE (transarterial chemoembolisation), hepatic artery ligation, radiofrequency ablation.

Prognosis

- Prognosis is very good for carcinoid of appendix. 5-year survival is 90% and commonly appendiceal carcinoid is identified incidentally; they are commonly in the tip or distal 2/3rd of the appendix; nodal spread and liver spread is rare; 75% are less than 1 cm. Two histological types of carcinoid *appendix are goblet cell and classical type*. Goblet cell type carries poor prognosis.

- Small bowel carcinoid has got poorer prognosis compared to carcinoid of appendix. Prognosis depends on size of primary tumour; nodal spread; multicentricity; hepatic spread. Carcinoid has better prognosis than adenocarcinoma of small bowel.

SHORT BOWEL SYNDROME (Short Gut Syndrome)

- It is the symptom complex after massive small bowel resection, i.e. *resection more than 70% of the bowel*. Minimum bowel required is 1.2 meters.
- Proximal jejunal resection is better tolerated than distal ileal resection.
- Ileum is more adaptive and has got capacity to increase the absorption capacity more efficiently. So in massive resections, patients with retained ileum will do better. Adaptation is better if ileum and caecum is preserved.

Causes

- Massive small bowel resection as in superior mesenteric artery ischaemia.
- Multiple sequential resections.
- Recurrent Crohn's disease—common cause.
- Necrotising enterocolitis in neonates and children.
- Intestinal atresia.
- Midgut volvulus.
- Radiation enteritis.

Adaptive changes in the retained small bowel

- Villous hypertrophy and hyperplasia
- Increased absorptive surface
- Increased capacity of small bowel
- Lower transit time of bowel content

Factors Deciding the Outcome in Short Gut Syndrome

- Age—infants tolerate massive resection better.
- Retaining ileocaecal valve shows better outcome. Ileocaecal resection interrupts the enterohepatic circulation of bile salts which in colon get metabolised into secondary bile salts. Secondary bile salts block the absorption of water and electrolytes.
- Delayed clearance of gastrin causes hyperacidity and peptic ulcer.
- Adaptability of the small bowel. Adaptation is better if ileum and caecum are preserved.
- Additional colectomy increases the morbidity. In such occasion retained jejunum will function as either of two ways. '*Net absorbers*' have adequate absorption of water and sodium and usually will have more than 100 cm of retained jejunum. '*Net secretors*' have less than 100 cm of jejunum and lot of sodium and water are secreted into the lumen from plasma. These patients should avoid hypotonic solutions like water/ tea; instead should take hypertonic saline with glucose containing 90 mmol sodium/L. WHO cholera fluid is ideal (contains 90 mmol sodium/L).

Clinical features

- Diarrhoea
- Fluid and electrolyte deficiency
- Severe malnutrition
- Recurrent bacterial enteritis

Outcome of Short Bowel Syndrome

- Severe malabsorption.
- Severe dehydration.
- Gallstone formation due to altered bile metabolism.
- Urinary calculi due to increased oxalate level.
- Water and electrolyte imbalance.
- Diarrhoea is common.
- Recurrent bacterial enteritis.
- Fulminant hepatic failure often can occur.
- Osteomalacia, tetany, hypomagnesaemia.
- Bleeding diathesis, peripheral neuropathy.

MASSIVE BOWEL RESECTION

- Resection of *more than 200 cm* of small bowel is called as massive resection.
- *Metabolic sequelae* in massive resection depend on – extent and anatomical site of resection; functional capacity and adaptation of the small or large bowel; cause for the resection. It causes gastrin hypersecretion, increased parietal cell mass in stomach, peptic ulceration; increased endocrine pancreatic secretion; hypoalbuminaemia, oedema.
- *Resection of more than 100 cm of distal ileum* causes diarrhoea and steatorrhoea (steatorrhoea is faecal fat > 6 gram/ day); increases bile lithogenicity; causes deficiency of vitamins A, D, E, K; colonic bacteria converts unabsorbed fatty acids and conjugated bile salts into hydroxy fatty acids and deconjugated bile salts.
- *Ileum* maintains enterohepatic circulation of bile salts, absorption of vitamin B_{12} and vitamin D. Resection of middle part of small bowel is better tolerated.
- *Removal of ileocaecal valve* reduces the intestinal transit time; reduces absorption of vitamin B_{12}, calcium, magnesium, zinc; increases diarrhoea and contamination of shortened small bowel by bacteria.
- In massive resection, *colonic bacteria* degrade fatty acids into lactate and short chain fatty acids; lactate reduces colonic pH which inhibits bacteroides; due to this acid resistant anaerobes will increase in colon which produces large amount of D lactate which is absorbed causing *D lactic acidosis* presenting with confusion, ataxia and nystagmus.
- *TPN* was introduced by Dudrick. It is very useful in short gut syndrome.

Treatment

Early phase

- Total parenteral nutrition.

- Fat and fibre free but protein rich liquid diet with essential fatty acids.
- Diarrhoea is controlled by loperamide/codeine phosphate.
- Oral cholestyramine to bind bile salts is needed in massive resection which includes ileum.
- Parenteral vitamin B_{12} injection regularly.
- H_2 antagonists/PPIs/somatostatin (to reduce secretions from stomach, liver and pancreas).
- Octreotide reduces the secretion and reduces GI motility.
- Fluid and electrolyte management.
- Control of diarrhoea.

Late phase

- Enteral nutrition to stimulate intestinal adaptation.
- TPN supplement—often permanently required.
  Home parenteral nutrition is universally used method in western countries. But in Asian countries it is yet not practicable.
- Hormones and glutamine administration are under trial.
- Surgical techniques to delay the intestinal transit time or small bowel transplantation are under trial but with less success.
  - *Reversal of 10 cm intestinal segment* to delay transit time.
  - *Intestinal lengthening* to delay transit time and increase the absorption surface.
  - *Small bowel transplantation*, ideal but graft rejection and failure is the problem.
  - Antiperistaltic colonic interposition into the small bowel segment.
  - Regeneration of intestinal mucosa over denuded serosa.
  - Longitudinal splitting of the intestine and closing of these as separate tubes which are anastomosed to each other to achieve lengthening—*Binachi's surgery.*
  - Mucosal *stem cell transplantation using enterocytes* without lymphoid tissue.
  - Seeded mucosal autografts in a prosthetic tubes.

SMALL BOWEL ENEMA (Enteroclysis)

It is the contrast study of the small bowel by infusing contrast agent either high density microbar solution with methyl cellulose or water soluble iodine dye through Nolan intestinal tube.

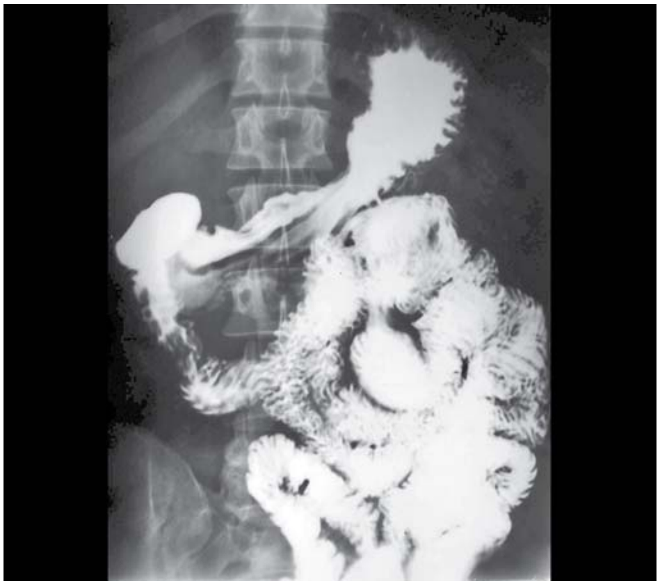

Fig. 21.34: Enteroclysis X-ray film showing entire small bowel. It is still commonly used investigation to assess small bowel pathology.

Indications

- Small bowel tumours.
- Stricture small bowel.
- Congenital anomalies.
- Crohn's disease, intestinal tuberculosis.

Findings to look for:

- Narrowing.
- Filling defects, irregular or regular.
- Mass lesions.

Problems

- Poor patient acceptance.
- Technically difficult.

CAPSULE ENDOSCOPY

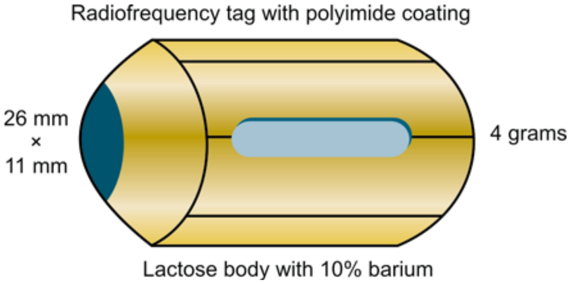
Radiofrequency tag with polyimide coating
26 mm × 11 mm
4 grams
Lactose body with 10% barium

Fig. 21.35: Capsule endoscopy—diagrammatic look.

- It is a type of endoscopy wherein highly sensitive miniature video-camera which after activation is swallowed by the patient. It is 26 mm × 11 mm in size. It weighs 4 grams.
- It contains single use video-camera, with 6 light emitting diodes, a lens, a colour camera chip, two batteries, radiofrequency transmitter with an antenna. Capsule takes 2 pictures per second which is transmitted to a worn recording device through radiofrequency. From the recording device data is downloaded to special computer.
- It is swallowed with 12 hours fasting.
- After activation it functions for 8 hours. It is mainly used to study small bowel diseases like vascular malformations, narrowing, tuberculoses, ulcers and tumours.
- This capsule camera sends signals and endopictures at regular intervals to the receiver (digital recorder) tied over the patient's waist.
- This receiver is later attached to specialised computer software to get different level pictures for study.
- Capsule gets deactivated in 8 hours and is passed out in the stool.
- It is available only at few centers and is costly.
- Capsule retention is the complication in 5% cases.
- It may not give proper evaluation in obstructive pathology and motility disorders of small bowel.

SMALL BOWEL ENTEROSCOPY

It is a difficult technique done to visualise the small bowel.

Indications

- Occult or obscure GI bleed. 5% of GI bleed is not diagnosed by any methods like gastroduodenoscopy, colonoscopy, contrast imaging. Angiodysplasia and ectasias are common causes for such type.
- Small bowel tumours.
- Crohn's disease.
- Celiac disease.
- Refractory sprue.
- HIV related small bowel diseases.
- Intraoperative enteroscopes to assess lumen on table.

Technique

Push Enteroscopy

- It is easier and faster. It reaches 60 cm beyond the ligament of Treitz.
- Flexible enteroscopes or paediatric colonoscopes are used.
- Simethicone is given prior to enteroscopy.
- BT, CT, LFT and prothrombin time should be checked. Aspirin should be stopped 5 days prior if patient is taking it.
- Oesophageal overtube is used to pass the entero-scopes through upper GI.
- Enteroscopes is passed through overtube into stomach, duodenum, and jejunum. Proper evaluation of the mucosa is done during withdrawal. If there is a bleeder point, it is fulgurated.

Sonde Enteroscopy

- It is a 5 mm diameter, 275 cm enteroscope passed through nose by a piggyback technique with a paediatric colonoscope. Its two internal channels are for air inflation and for balloon inflation. Biopsy or working channel is not there.
- Scope is passed through pylorus and further it is moved with subsequent inflation of balloon and distal progression by peristalsis. Distal position is confirmed by fluoroscopy. It takes 6 hours to pass through the entire small bowel.
- Once it is in the terminal part of the small bowel, small bowel is inflated with air and mucosa is inspected.

Intraoperative enteroscopy

- Abdomen is opened. By a small enterotomy scope is passed into the lumen. It is guided manually by compression and release to visualise entire mucosa of small bowel.

Complications and Problems

- Epistaxis due to nasal irritation.
- Perforation.
- Pancreatitis.
- Failure to pass enteroscopes as needed.
- Time consuming and not acceptable by the patient.

ENTERIC/GASTROINTESTINAL FISTULA

Enteric fistula is a challenging problem due to its high mortality of 30%, management of electrolyte imbalance, malnutrition, and sepsis.

Classifications

Anatomical

- *External:* They are **enterocutaneous fistulas** which are discussed in detail here.
  - 75% of enteocutaneous fistulas are of *postoperative* cause due to disruption of the anastomotic site.
  - 25% are *spontaneous* due to malignancy, radiotherapy, inflammatory bowel disease (most common in spontaneous cause), diverticular disease and bowel ischaemia.
  - 70% of external fistula will close spontaneously.
- *Internal*
  - Most of the internal fistula need surgical correction, if they are symptomatic or causing complications.

Aetiological

- Anastomotic disruption in postperative period.
- Trauma to normal bowel—inadvertent enterotomy/injury.
- Disease of the bowel extending into the adjacent structures.
- Disease outside the bowel extending into the normal bowel.

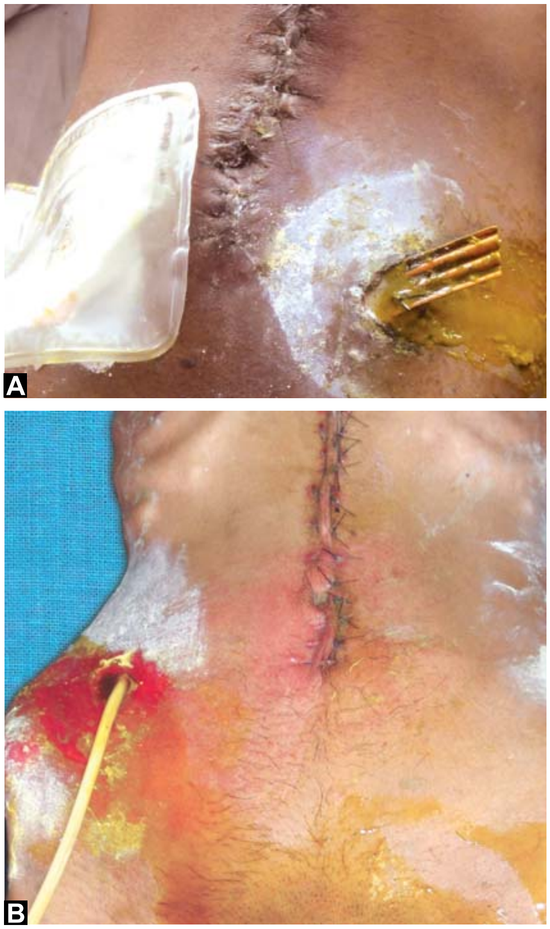

Figs 21.36A and B: Enterocutaneous fistula in two different patients.

Depending on the site

- *Oropharyngeal*
- *Oesopharyngeal*
- *Gastric:* Commonest cause is the site of gastrostomy tube, either with tube *in situ* or after tube removal. Other causes are—post-gastrectomy and anastomotic leak after surgery for carcinoma stomach, surgeries for benign gastric diseases, reflux diseases, and obesity.
- *Duodenal*: It occurs after surgery for pancreas, biliary system, duodenal stump leak, gastroduodenal anastomosis, renal, aortic, colonic surgeries. Spontaneous fistula occurs in Crohn's, ulcers, malignancy and trauma. It can be duodenal stump fistula which often closes spontaneously; lateral duodenal fistula which less commonly closes spontaneously.
- *Small bowel*: It is the commonest site of GI fistula. More than 75% are postoperative. Crohn's disease is the most common cause of spontaneous small bowel fistula in which 50% each are internal and external.
- *Colonic:* It can be due to postappendicectomy, postcolonic surgeries. It is common after emergency surgeries, and surgeries in unprepared bowel. Appendicectomy in acute phase of Crohn's may cause fistula due to adhesion of inflamed terminal ileum to abdominal wall wound, not from appendicular stump. Spontaneous fistula even though rare, occurs in diverticulitis, malignancy, ulcerative colitis, radiation and pancreatitis.

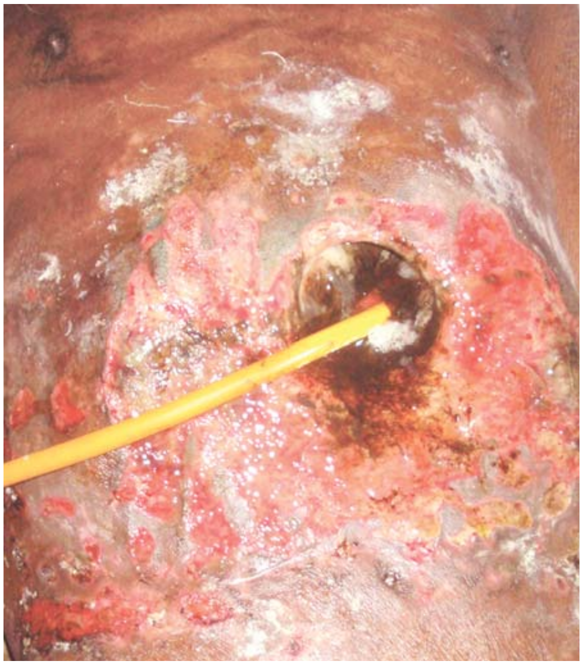

Fig. 21.37: Fistula at gastrostomy site. This is becoming the common cause.

Fistula may be due to:	***Physiological—based on quantity of daily output:***	***Fistula may be:***
• Complete disruption of anastomotic site • Partial disruption • Lateral fistula with distal obstruction • Fistula in stricture bowel	• *High output*—> 500 ml/day—usually small bowel; 50% mortality; less chance of spontaneous closure. • *Moderate output*—200-500 ml/day—colonic and small bowel mixed. • *Low output*—< 200 ml/day—colonic; mortality is 15%; more chance of spontaneous closure.	*Favourable:* • Site—oropharyngeal, oesophsgeal, duodenal stump, pancreaticobiliary, jejunal, colonic. • Cause—postoperative, appendicitis, diverticulitis. • Low output, absence of sepsis, transferrin level more than 200 mg/dl. • Healthy adjacent tissues with intestinal continuity, without distal obstruction. • Fistula track more than 2 cm; bowel wall defect less than 1 cm^2. *Unfavourable:* • Site—gastric, lateral duodenal, ileal. • Cause—malignancy, inflammatory bowel disease. • High output, poor nutrition, sepsis, transferrin < 200 mg/dl. • Diseased adjacent bowel, presence of distal obstruction, large abscess, complete bowel disruption, radiation. • Track < 1 cm; bowel wall defect > 1 cm^2. Track epithelialisation, presence of foreign body.
Fistula may be: • *End fistula* • *Lateral fistula* **Fistula may be:** • Single • Multiple	**Fistula stoma categories:** *Category 1*: A single orifice fistula through an intact abdominal wall or healed scar with normal skin adjacent *Category 2*: Single or multiple fistulas passing through abdominal wall close to bony part or umbilicus or surgical scar *Category 3*: Fistula through a small dehiscence through the main wound *Category 4:* Fistula through a large dehiscence through main wound or dehiscence at the bottom of wound	
Fistula may be: • *Simple* • *Complicated*—if multiple tracks, track with an abscess, communication to more sites or more than one viscus. **Note:** Fistula of duration of more than 3 months is called as *chronic fistula*		

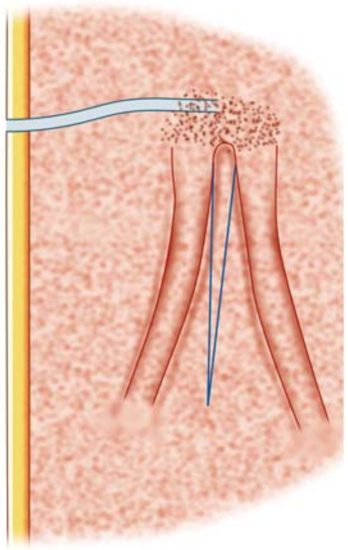

Fig. 21.38: Complete disruption of the anastomotic site causes fistula which is unlikely to close spontaneously.

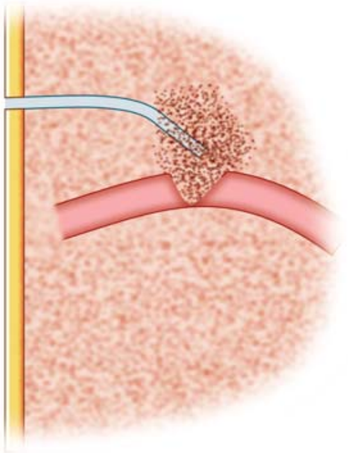

Fig. 21.39: Partial disruption also is equally problematic.

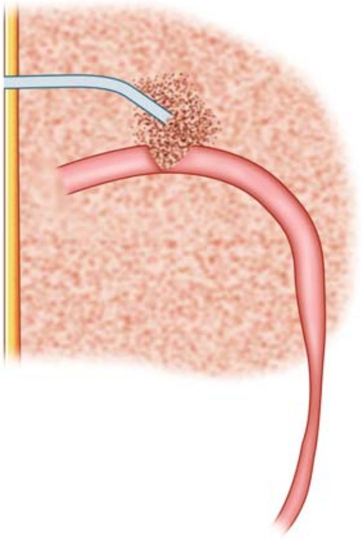

Fig. 21.40: Distal obstruction precipitates fistula formation and also prevents its closure.

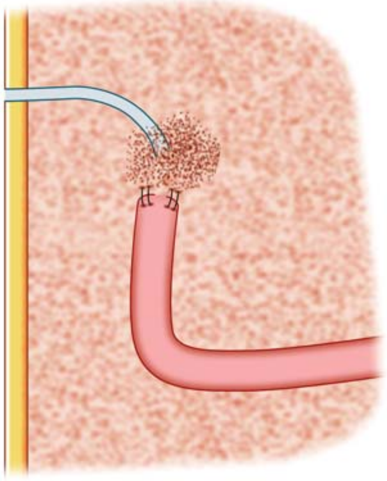

Fig. 21.41: End fistula. It is seen in duodenal stump leak.

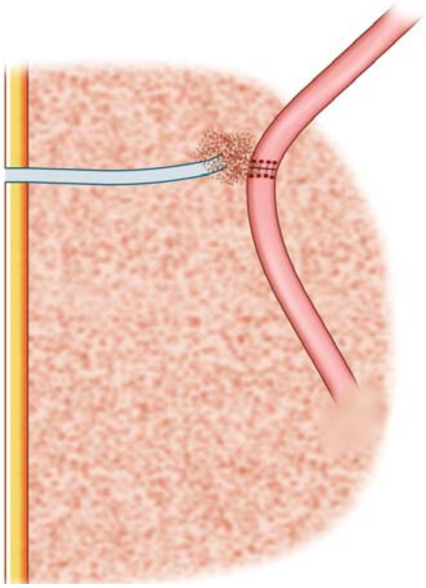

Fig. 21.42: Lateral fistula is seen in anastomotic leak from suture line.

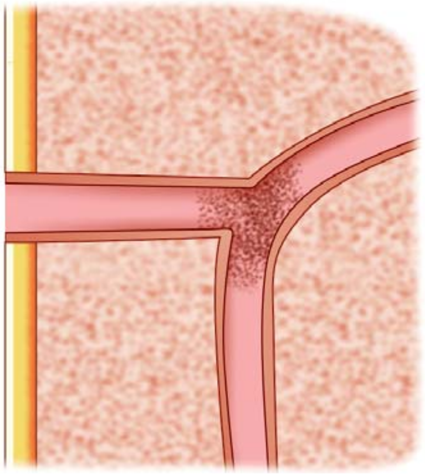

Fig. 21.43: Epithelialisation of the fistula track prevents spontaneous closure.

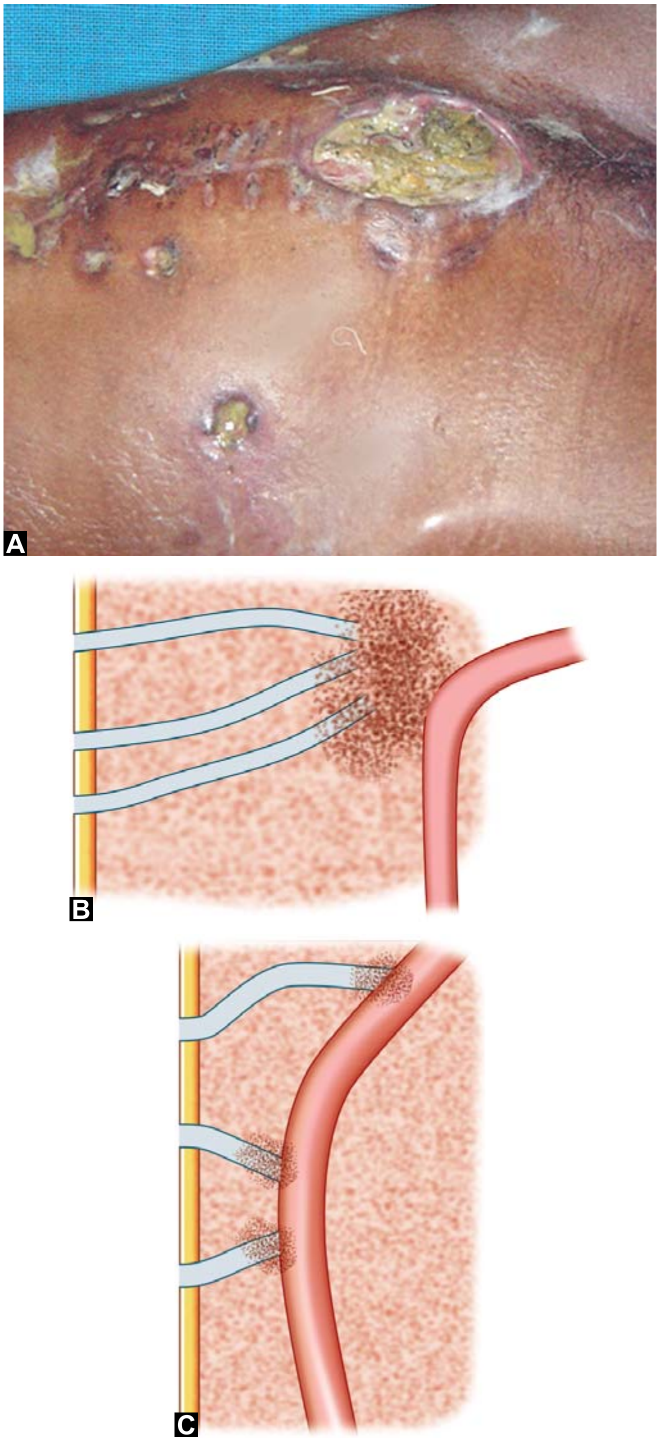

Figs 21.44A to C: Multiple enterocutaneous fistulas can occur from same site or from different sites.

Factors Precipitating the Fistula Formation

- *Patient factors:* Anaemia, malnutrition, sepsis, hypotension, hypothermia, poor oxygen saturation, associated respiratory diseases/failure, emergency surgeries, specific diseases like malignancy, Crohn's, tuberculosis, adhesions.

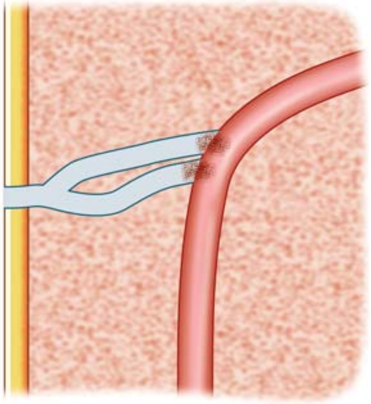

Fig. 21.45: Fistula can have multiple tracks even though it is opening at one point in the skin.

- *Technical factors*: To reduce the chances of fistula formation technical care is essential like—proper mobilisation, meticulous dissection, tension free anastomosis, proper secure anastomosis, most important is proper blood supply at the bowel cut edges, care to avoid cautery injury as bowel is very sensitive to heat/cautery injury, proper prior preparation of patient and bowel.

Problems with enterocutaneous fistulas	***Factors which prevents spontaneous closure of fistula***
• Skin excoriation • Electrolyte imbalance • Severe malnutrition • Recurrent sepsis—locally and systemic like pneumonia, candida, varicella • Pellagra, osteomalacia, zinc deficiency	• *Distal obstruction* • Foreign body in the track • Abscess in the track, infection • Definitive disease like carcinoma, Crohn's, tuberculosis • Epithelialisation of the track • Anaemia, hypoalbuminaemia (2.5 g/dl) • Complete disruption of anastomotic site

Management of Fistula

Phase 1

- *Resuscitation and restoration* of volume with crystalloids and colloids, blood transfusion to achieve haematocrit of 30%, maintenance of albumin level at 3.0 gm/dl with albumin infusion.
- *Sepsis control* with antibiotics, percutaneous drainage of abscess under guidance or open drainage.
- *Skin care* to prevent excoriation using Karya powder, zinc oxide cream/powder, ion exchange resins, stoma adhesive and controlled fistula drainage using sump constructed suction catheter drain system or vacuum assisted closure (VAC) system or silicone barrier or created inverted cone system.

- *Reduction of output of fistula*—proton pump inhibitors, histamine antagonists, sucralfate, octreotide, infliximab (in fistula in Crohn's patients). Long term nasogastric aspiration should be avoided.
- *Nutrition*: Nutritional status should be assessed by clinical (weight, anthropometry), biochemical methods. 30 Kcalories/kg/day; 1.5 grams/kg/day of protein is the basic need. Initially TPN is used. Once patient tolerates oral, enteral feeding should be started ideally. Enteral feeding (oral/gastrostomy/jejunostomy) is contraindicated in presence of distal obstruction. Enteral feeding reduces the sepsis, improves the bowel activity, caliber, thickness and ability to hold sutures. It also avoids TPN related problems.

Phase 2

- *Investigations* are done to assess fistula and its causes. It is done in 7-10 days of fistula formation.
- *Fistulogram* using water soluble contrast, *CT fistulogram* to see the pathological anatomy of fistula—site, number, length, status of bowel, distal obstruction, presence of abscess cavity.
- *Biochemical analysis* (electrolytes, haematocrit and albumin) and renal, hepatic, respiratory, cardiac status should be assessed carefully.

Phase 3

- *Decision* by observation and assessment, whether fistula will close spontaneously or not. Favorable fistulas are likely to close spontaneously but not unfavorable.
- *Definitive procedure* is done for fistula only after 6 weeks. Mortality and recurrence is higher if operated prior to 6 weeks due to obliterative peritonitis.
  - *Definitive surgical procedure* is lengthy, complex and *team work*. Optimum nutrition, proper planning, prophylactic antibiotics are needed.
  - *Reopening* should be done through a *new distant* often transverse incision.
  - *Bowel refunctionalisation* by freeing entire bowel from ligament of Treitz to rectum should be done to clear adhesions and obstructions and all areas of sepsis and abscesses. *Sharp dissection* using scissor should be done to clear adhesions. As much as possible bowel injury should be avoided; if occurs it should be closed transversely using interrupted 3 zero silk sutures.
  - *Resection of the bowel adjacent to fistula* with track is the ideal procedure with end to end meticulous two layered closure using interrupted 3 zero silk sutures.
  - When it is not possible, fistula area bypass, Roux-en-Y drainage, serosal patch technique is used. Duodenal fistula is better managed by bypass using gastrojejunostomy and vagotomy without intervening the fistula.
  - *Proper irrigation of abdominal cavity* with saline and antibiotics during procedure, omental flap around the anastomosis, various solutions to prevent repeat adhesions are also often done. Supportive jejunostomy may be added for enteral feeding.
  - *Abdominal wall closure* is important by primary closure or by using myocutaneous flap. Mesh should not be used for closure as *recurrent fistula* may occur.

Phase 4

- Treatment during recovery and healing time also should be adequate and optimum.
- Supplementing of nutrition, protein, vitamins and essential element are important.
- Physical, psychological therapy is needed.

Happiest people are not only happy in themselves; they are the cause of happiness to others.

Chapter 22 Large Intestine

CHAPTER OUTLINE

- Anatomy
- Hirschsprung's Disease
- Diverticular Disease of the Colon
- Ulcerative Colitis
- Ischaemic Colitis
- Pseudomembranous Colitis
- Surgical Complications of Intestinal Amoebiasis
- Tumours of Colon
  - Benign Tumours/Polyp of the Colon
  - Juvenile Polyps
  - Metaplastic/Hyperplastic Polyp
  - Peutz-Jegher's polyp
  - Adenoma of Colon
  - Familial Adenomatous Polyp
  - Gardner's Syndrome
- Carcinoma Colon
- Angiodysplasia of Colon
- Ogilvie's Syndrome
- Colostomy
- Stoma Care
- Stoma Appliances
- Faecal Fistula
- Preparation of Large Bowel for Surgery
- Surgical Pouches
- Barium Enema

ANATOMY

- The colon (large intestine) is 135 cm long, and is divided into caecum, ascending colon, transverse colon, descending colon and sigmoid colon.
- The wall of the colon is composed of mucosa, submucosa, innercircular muscle layer and outer longitudinal muscle layer which, in turn, is concentrated into 3 separate longitudinal strips—*taeniae coli.*
- Small pockets of fat filled peritoneum—*appendices epiploicae* is scattered all over the colon except appendix, caecum and rectum.
- *Haustra* are sacculations between the taeniae.
- All the 3 above are important features of colon.
- *Ileocecal valve* serves as a sphincter to prevent the back reflux to terminal ileum.

Blood Supply

- Iliocolic, right colic, and middle colic arteries which are branches of superior mesenteric artery supply the colon from caecum to splenic flexure.
- Left colic, sigmoid, superior rectal arteries which are branches of inferior mesenteric artery supply the descending and sigmoid colon.
- The anastamotic arcade formed between the branches of superior and inferior mesenteric arteries is called '*arc of Riolan'*.
- Venous drainage occurs into superior mesenteric vein (which joins the splenic vein to form the portal vein) and inferior mesenteric vein (drains into the splenic vein).

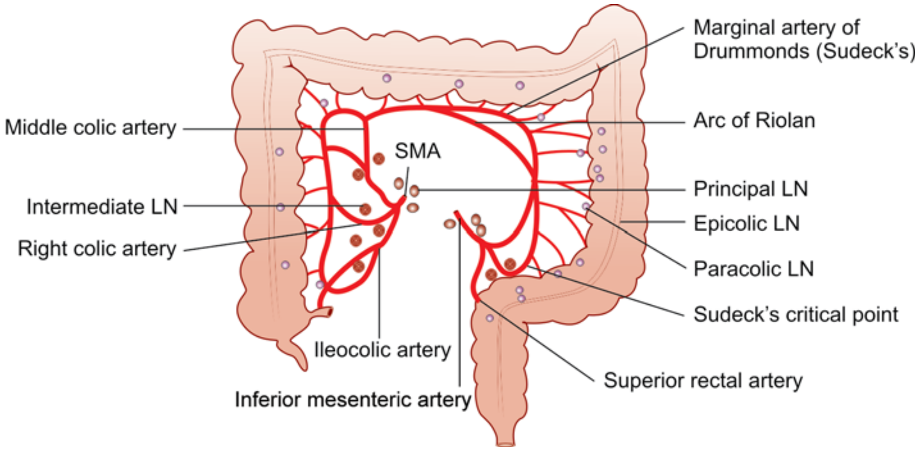
Marginal artery of Drummonds (Sudeck's)
Middle colic artery
Arc of Riolan
SMA
Principal LN
Intermediate LN
Epicolic LN
Right colic artery
Paracolic LN
Sudeck's critical point
Ileocolic artery
Superior rectal artery
Inferior mesenteric artery

Fig. 22.1: Arterial supply and lymphatic drainage of colon.

Lymphatic Drainage

- Mucosa contains no lymph channels, so mucosal cancers rarely metastasize.
- Nodes are *epicolic* (located in the colonic wall), *paracolic* (located along the inner margin), *intermediate* (located near mesenteric vessels), *principal* (located near main mesenteric vessles).

Nerve Supply

- Colonic motility is under control of autonomic nervous system; parasympathetic via vagi and pelvic nerves, sympathetic via superior and inferior mesenteric ganglia.
- The effect of meal on colonic activity is termed as *gastrocolic reflex.*

HIRSCHSPRUNG'S DISEASE (CONGENITAL MEGACOLON)

> *My first specimen is a colon, but a colon of such a size that it will no doubt surprise you to learn that it comes from a child only 11 months old when it died.... Only (the) rectum was not dilated, nor indeed subject to any obstruction.*
>
> —**Harald Hirschsprung,** 1888

- It is a congenital, familial condition, occurring in newborn due to the *absence of ganglion cells—Auerbach's* and *Meissner's plexus in anorectum*, which may extend proximally either a part or full length of the colon.
- It always involves the anus, internal sphincter and rectum (partly or entirely).
- There is narrow, spasmodic, non relaxing pathological segment.
- Transitional zone proximal to it contains only few ganglion cells with formation of cone.
- Still proximal to it, colon is dilated enormously with hyperaemia, multiple ulcers and hypertrophied circular muscle fibres.
- It is one of the causes of neonatal intestinal obstruction.
- Severe enterocolitis can occur which may be fatal. Perforation, peritonitis and septicaemia can occur.
- Often there will be a chronic course of the disease with malnutrition, abdominal distension.

Types

1. *Ultrashort segment HD*—only anal canal and terminal rectum is aganglionic.

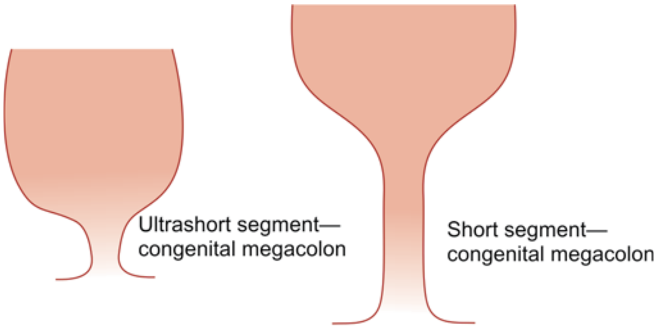

Fig. 22.2: Types of Hirschsprung's disease.

2. *Short-segment HD*—anal canal and rectum is completely involved (80%).
3. *Long-segment HD*—anal canal, rectum and part of the colon is involved (10%).
4. *Total colonic HD*—anal canal, rectum and full length of the colon is involved - 10%.

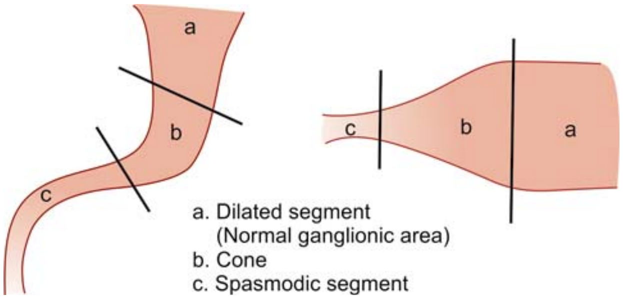

Fig. 22.3: Zones in Hirschsprung's disease.

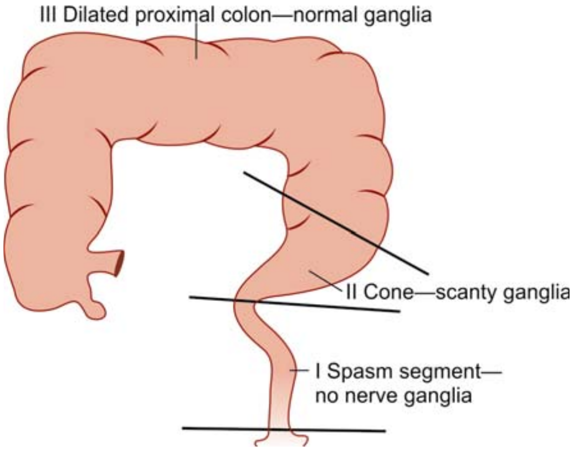

Fig. 22.4: Congenital megacolon showing spasmodic aganglionic segment, coning (transitional zone), and proximal dilated normal ganglionic segment.

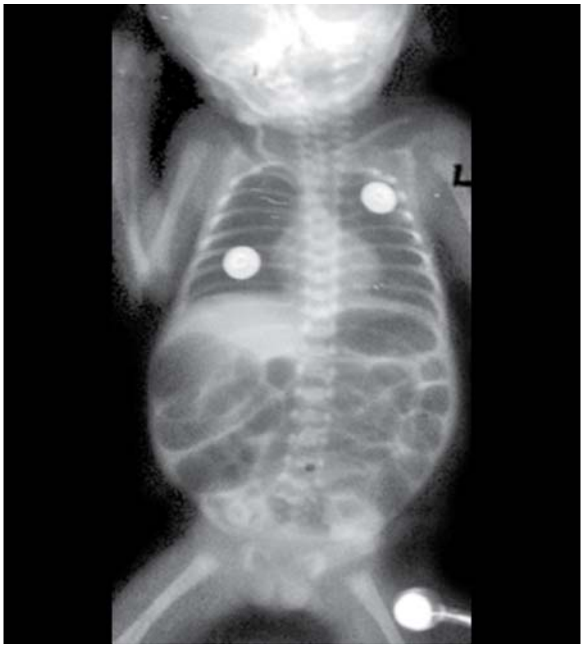

Fig. 22.5: Newborn with Hirschsprung's disease with intestinal obstruction (*Courtesy:* Dr Vivek Prabhu, MCh, Mangalore).

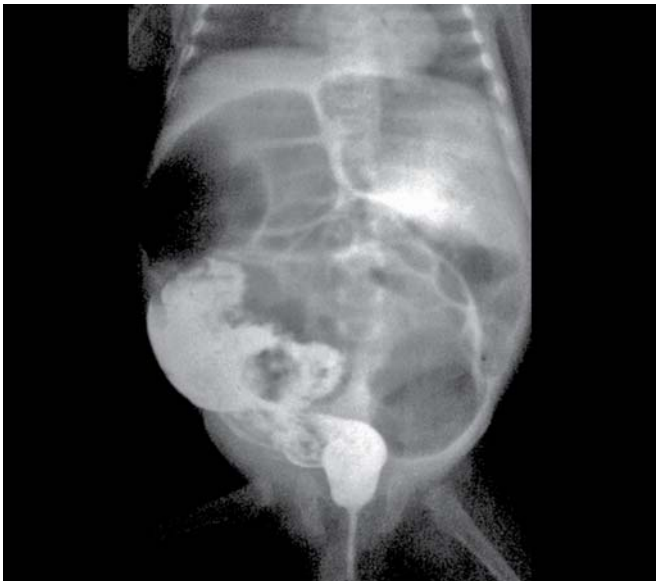

Fig. 22.6: Barium enema X-ray showing parts of congenital megacolon—spasmodic area, cone, proximal dilated segment.

It has got ***three zones***

I. *Distal immobile spastic segment,* i.e. aganglionic zone.
II. Proximal, *middle transitional zone* of about 1–5 cm length with less, sparse number of ganglions (*cone).*
III. A still more proximal, *hypertrophied dilated* segment is actually the normal ganglionic area.

Clinical Features

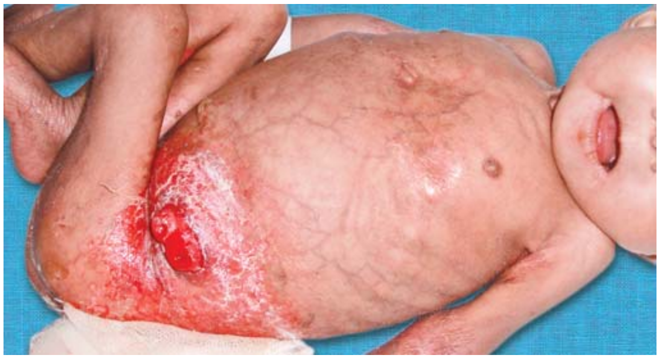

Fig. 22.7: Child suffering from Hirschpurung's disease undergone initial sigmoid colostomy.

Presentations: Acute, recurrent, chronic.

- It is common in males (80%).
- Its incidence is 1 in 5000 live births.
- It is common in infants and children, occasionally it occurs in adults also.
- Often it is associated with Down's syndrome (10%). (Commonest association).
- In 90% of cases, symptoms appear in early neonatal period, i.e within three days of birth. *The child fails to pass meconium.* After introducing finger into the rectum, child passes *toothpaste like stool*, with evidence of straining. Distension of the abdomen with features of intestinal obstruction is seen.
- In children, there is passage of *goat pellet* like stools, malnutrition, abdominal distention—chronic type. Constipation, with history of passing stools once in 3-4 days with straining is seen throughout the childhood and also in adolescent period. Occasionally, condition can cause intestinal obstruction.
- 10% familial.
- Gene mutation can occur in chromosome no. 10 commonly; occasionally in chromosome no. 13.
- ***Rectal examination*** shows tight sphincter with empty rectum. Child passes lot of gas and meconium.

Diagnosis

- History of failure of passing meconium.
- Plain X-ray abdomen—shows intestinal obstruction. Useful in case of perforation.
- Biopsy from all three zones to study the *ganglions and hypertrophic nerve terminals* in spasmodic segment. Starting from 2 cm above the dentate line, a full thickness rectal biopsy is ideal.
- ***Barium enema*** is done to look for the extent of disease and three zones. *Foley's catheter should not be used while doing barium enema in case of Hirschsprung's disease.*
- ***Anorectal manometry***—shows the absence of rectoanal reflex in Hirschsprung's disease, which is diagnostic.
- *Acetylcholine esterase staining shows hypertrophied nerve bundles*.

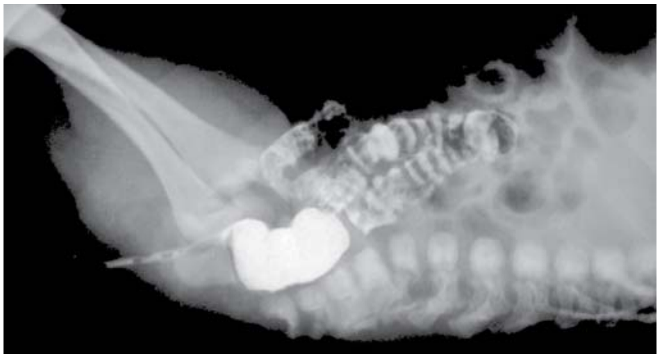

Fig. 22.8: Barium enema X-ray showing features of Hirschsprung's disease in a newborn.

Complications

- Colitis (Intramucosal gas in plain X-ray). Enterocolitis may be fulminant and fatal.
- Intestinal obstruction.
- Growth retardation.
- Constipation.
- Perforation.
- Peritonitis.
- Septicaemia.

Differential Diagnosis

- Total neuronal dysplasia
- Acquired megacolon—rectum is loaded with stool
- Anorectal malformations (ARM)
- Hypothyroidism
- Meconium plug syndrome

Treatment

- Initially, colostomy is done either transverse or transitional, so to have normal bowel function.
- Nutritional supplementation.
- Once the child attains 10 kg of weight, definitive procedure is done, i.e.
  a. Excision of aganglionic segment (spasmodic segment).
  b. Maintenance of continuity by doing coloanal anastomosis.
  c. Closure of colostomy later.

  Common procedures done are:
- *Modified Duhamel Operation*—resection of upper part of the rectum and a part of colon; anastomosis of colon to posterior part of the lower rectum and crushing the spurs to create the rectal pouch. It is *technically easier and a retro-rectal pull through*. New pouch is created by anterior part of the aganglionic rectum and by ganglionic proximal pulled down colon. Biopsy should be taken from proximal pulled down colon to look for evidence of ganglions. Pulled down proximal colon is sutured to full thickness posterior anal canal just above the dentate line. Spur between these two segments is crushed by Kocher's forceps or specialized instrument to create a single pouch.
- Soave's mucosectomy and pull through operation.
- Coloanal anastomosis after proctocolectomy.
- Total proctocolectomy with ileo-anal anastomosis in case of total colonic HD.
- *Swenson's operation*—through abdomino-anal approach, aganglionic segment is resected and colo-anal anastomosis is done.
- *Anorectal myectomy* is found to be very useful for ultrashort segment and short-segment Hirschsprung's disease. Mucosa in incised horizontally 1 cm proximal to mucocutaneous junction. A strip of muscularis with part of internal sphincter is excised with both muscle layers of the rectum for about 6-10 cm length. Mucosa is sutured back. Complications are abscess formation and failure.

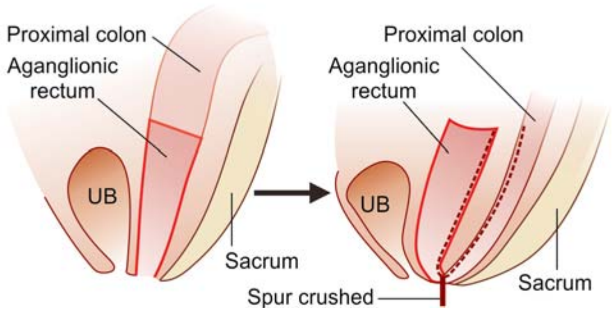

Fig. 22.9: Modified Duhamel operation. Normal colon is brought behind the aganglionic rectum and anastomosed just above the dentate line.

Complications of surgery

- Severe colitis
- Faecal fistula
- Stenosis
- Stunted growth

Acquired megacolon

- No contracted segment in rectum
- Seen in children with faulty toilet training
- Rectum and sigmoid colon are dilated
- Normal ganglions in all levels
- Improper bowel habit causing chronic bowel dilatation
- Repeated enemas, manual evacuation, toilet training, educating the parents are required
- Should be differentiated from Hirschsprung's disease

DIVERTICULAR DISEASE OF THE COLON

- They are acquired herniations of colonic mucosa through circular muscles at the points where blood vessels penetrate (points of least resistance).
- It is more commonly localised to *sigmoid colon (90%)* but occasionally seen in full length of the colon.
  Rectum is not affected.

Saint's triad (5%)

- Diverticulitis
- Hiatus hernia
- Gallstones

- It is rare in Asian and African countries because of the high fibre diet. It is common in western countries.
- Colonic diverticulosis is usually of false type with only mucosal herniation.

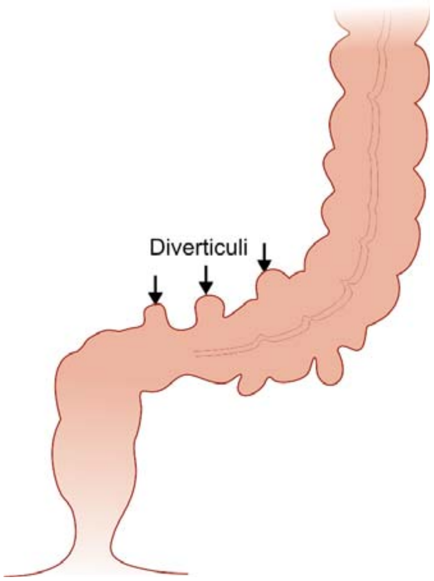

Fig. 22.10: Diverticular disease of colon.

Aetiology

- *Diet*—it is the main factor. Low fibre diet increases the stool transit time, reduces the stool weight, reduces the bulkiness of stool which increases the intraluminal pressure and muscle hypertrophy. High fibre diet prevents this.
- Disease is more common in females. It is more common in aged.
- It is more common in nonvegetarian than in vegetarian.
- NSAID intake by inhibiting prostaglandin synthesis may cause diverticular disease. It is more common in individuals with steroid therapy or immunocompromised people.
- Smoking and alcohol.
- Long standing constipation increases the stool transit time and causes diverticulosis.

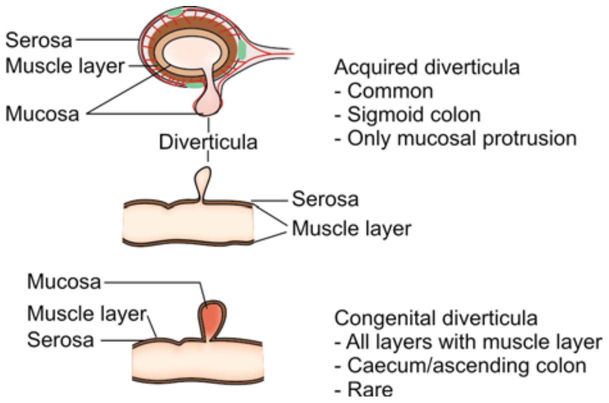

Fig. 22.11: Diagram showing the formation of diverticula. Note the difference between congenital and acquired diverticula.

Types

- ***Diverticulosis*** is the initial primary stage of the disease, wherein there is hypertrophy, muscular incoordination leading to increased segmentation and increased intraluminal pressure. At this stage they are asymptomatic, but often get severe spasmodic pain due to colonic segmentation called as *painful diverticular disease*.
- ***Diverticulitis*** is the second stage due to inflammation of one or more diverticula with pericolitis.It presents with persistent pain in left iliac fossa, fever, loose stool, recurrent constipation, tenderness in left iliac fossa, palpable and thickened sigmoid colon. P/R may reveal a tender mass.

Diverticulitis can be:

- *Uncomplicated diverticulitis* presenting with pain and spasm over left iliac fossa, fever. Recurrent attacks need sigmoid colectomy. It can be done electively through laparoscopy.
- *Complicated diverticulitis* presents with *haemorrhage, abscess, fistula* formation, perforation, peritonitis and obstruction due to stenosis. Vesicocolic fistula presents with *pneumaturia.*
- *Diverticula associated colitis* (DAC) presents with tenesmus, diarrhoea, haematochezia. DAC is a distinct clinical entity with segmental colitis.

Pathology

There is hypertrophy and thickening of the muscle layer with progressive colonic narrowing and segmentation with raised intraluminal pressure causing pulsion diverticula of only mucosa adjacent to taenia in antimesentric region. Pathology is common in sigmoid colon.

Features of Diverticular Disease

- In western countries, 50% risk to develop diverticular disease for an individual is at the age of 60 years. Only 15% of patients with diverticulosis develop diverticulitis. 75% of patients with diverticulitis have uncomplicated course with features of only diverticulitis. 25% of patients with diverticulitis develop complications like abscess, perforation, stenosis and fistula. Abscess can be commonly pericolic and pelvic, rarely in buttock and ischiorectal fossa.
- Features of diverticulosis—fullness of abdomen, bloating, flatulence, vague discomfort.
- Features of diverticulitis—pain in left iliac fossa which is constant radiates to back and groin, tenderness, bloody stool, often massive haemorrhage, fever, rigidity and mass in left iliac fossa. Mass is usually tender, firm, resonant, nonmobile.
- Features of fistula—colovesical is the commonest type of fistula. It causes passage of gas in the urine (*pneumaturia*) commonly and occasionally feces.

Complications of diverticulitis

- Perforation and pericolic abscess or peritonitis
- Progressive stenosis and intestinal obstruction
- Profuse colonic haemorrhage (17-20%)
- Fistula formation (5%)—vesicocolic, vaginocolic, enterocolic, colocutaneous

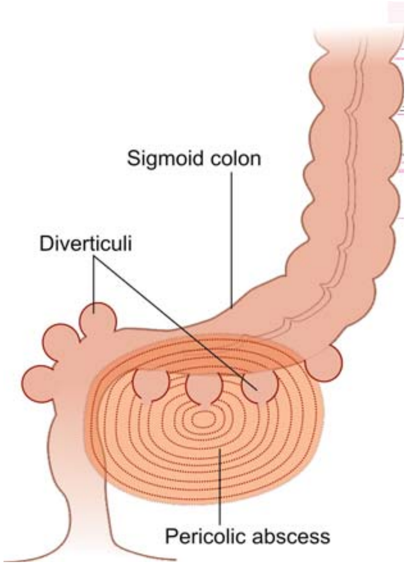

Fig. 22.12: Sigmoid diverticula causing pericolic abscess as a known complication.

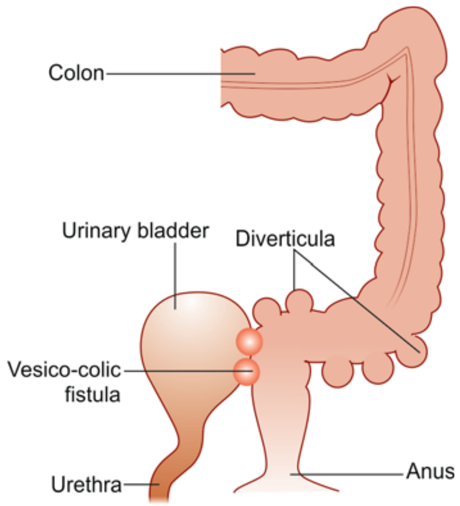

Fig. 22.13: Diverticular disease causing colovesical fistula. Patient will have urine stained with faecal matter and severe urinary infection. Contrast dye study, CT scan is required. Initial colostomy and ureterostomy as diversion is required before definitive procedure.

Note:
Diverticulitis is not a precancerous condition.

It may coexist with malignancy, irritable bowel syndrome, Crohn's disease.

Hinchey's classification of diverticulitis

- Stage I: Pericolic or mesenteric abscess
- Stage II: Walled off pelvic abscess
- Stage III: Generalised purulent peritonitis
- Stage IV: Generalised faecal peritonitis

Investigations

- *Barium enema (best method to diagnose)* shows *'saw-teeth'* appearance. *Champagne glass sign*—partial filling of diverticula by barium with stercolith inside—seen in sigmoid diverticula.
- *Sigmoidoscopy* is useful but should not be done in acute stage. Once acute stage subsides, barium enema, sigmoidoscopy, colonoscopy can be done (To rule out associated malignancy).
- CT scan in acute phase to see pericolic abscess.
  - CT scan shows thickening of muscle layer, abscess, perforation, fistula, involvement of organs like urinary bladder and associated pathology. CT scan is the ideal investigation.
- Cystoscopy and colonoscopy in case of fistula. Ureteric stenting is needed to make eventual surgery easier.

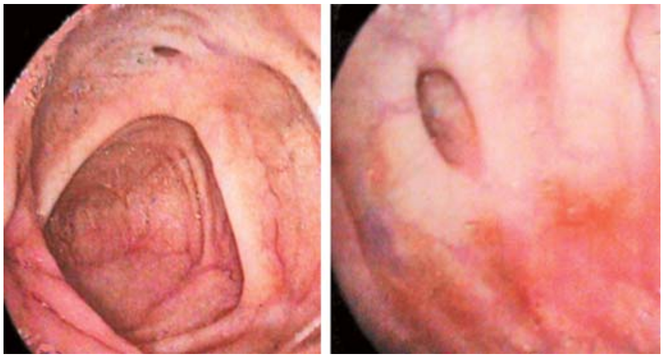

Fig. 22.14: Colonoscopic view of sigmoid diverticula.

Differential Diagnosis

- Carcinoma sigmoid colon.
- Amoebic colitis, ulcerative colitis, ischaemic colitis and Crohn's disease.
- Tuberculosis.

Coexistence of carcinoma and diverticulitis can occur in 12% of cases.

Treatment

Medical treatment

- High fibre diet
- Antibiotics
- Bulk purgatives
- Avoid constipation
- Regular follow-up of the progress of disease and onset of complications
- Abscess can be drained by CT guided aspiration or percutaneous drainage tube

- In acute stages, conservative treatment like bowel rest, antispasmodics, antibiotics are advised.
- *Guided aspiration* of the abscess is sufficient if abscess is small. Proper antibiotics are needed.
- Later, surgery is required**.** *Resection of sigmoid colon and anastomosis (colorectal) is done.*

Indications for surgery

- Recurrent diverticulitis
- Diverticulitis with complications

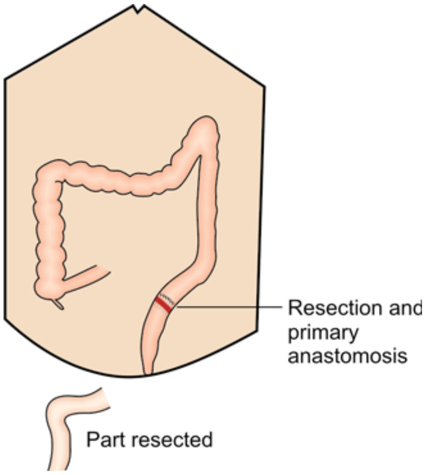

Fig. 22.15: Resection and primary anastomosis can be done in sigmoid diverticula after proper bowel preparation electively.

- If colon is loaded wth faecal matter, initially a transverse colostomy is done. Then resection and anastomosis; later colostomy closure is done as a staged procedure.

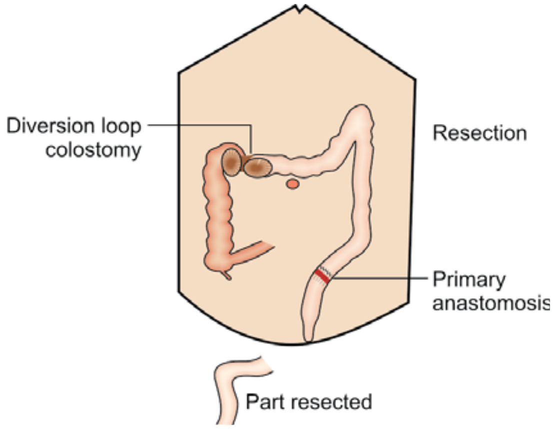

Fig. 22.16: Resection of sigmoid colon for sigmoid diverticula with primary anastomosis and defunctioning diversion colostomy.

- In case of perforation, proximal colostomy and exteriorisation of the affected bowel with later resection and anastomosis is done as it has high mortality.

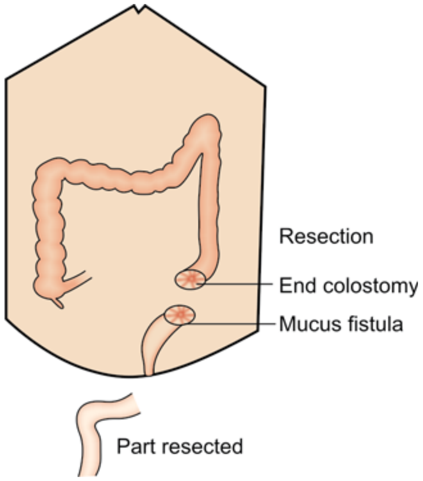

Fig. 22.17: Resection and formation of end colostomy and mucus fistula for acute sigmoid diverticulitis.

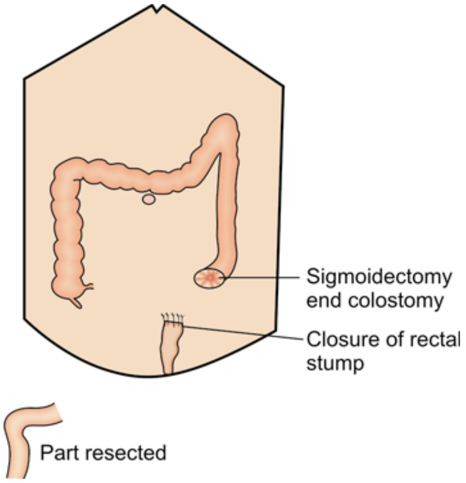

Fig. 22.18: Hartmanns operation for acute diverticulitis with sepsis/perforation/peritonitis. Hereafter resection proximal colon is exteriorised as end colostomy. Distal rectal stump is closed.

- Occasionally, Hartmann's procedure (combination of sigmoidectomy, end colostomy and closure of rectal stump) is better and life saving.
- Fistulas are treated by resection of the diseased bowel and closure of the fistula along with diversion procedures like colostomy, cystostomy, ureterostomy.
- In certain cases of diverticulosis, a longitudinal incision through the taenia and muscular layer without opening the mucosa is sufficient (like Heller's/Ramstedt's myotomy)—*Reilly's sigmoid myotomy*.

Note:

- It is basically **a benign condition**, therefore the prognosis is good. High fibre diet is advised.
- Complications like *perforations, fistulas are dangerous and life threatening*.
- Right sided diverticulum in the caecum and ascending colon is usually solitary and congenital.

Myotomy

- Heller's myotomy—achalasia cardia
- Ramstedt's myotomy—congenital pyloric stenosis
- Reilly's myotomy—diverticulosis

ULCERATIVE COLITIS

- It is an inflammatory condition of rectum and colon of unknown aetiology perhaps related to stress, westernized diet, autoimmune factor, familial tendency, allergic factor.
- Disease commonly starts in the *rectum,* spreads proximally to the colon and often into the ileum as ***back wash ileitis*** (5%).

Aetiological Factors

- Westernized diet, red meat; less common in vegetarians.
- Defective mucin production in the colonic mucosa and mucosal immunological reaction.
- Autoimmune factors—cytotoxic T lymphocytes against colonic epithelial cells and presence of anticolon antibodies. Association with HLA DR2 is observed in ulcerative colitis. DR 1501 is associated with less severe type' DR 1502 is associated with more severe form.
- Appendicectomy and smoking protects ulcerative colitis especially from extraintestinal features and from postoperative complications.
- Familial in nature.
- Allergy to milk (cow milk) and other dietary factors.
- Excess reactive oxidative metabolism in ulcerative colitis.
- Psychological aspects, stress, life style, personality disorders.

Pathology

To begin with, multiple minute ulcers
(*pin point ulcers*) occur with proctitis and colitis
↓
These ulcers extend into the deeper layer
↓
Spasm of the bowel

↓
Stricture of the colon
↓
Permanently contracted colon (*pipe stem colon*)
↓
In between ulcers, epithelial thickening occurs which appears like polyps
↓
Pseudopolyposis.

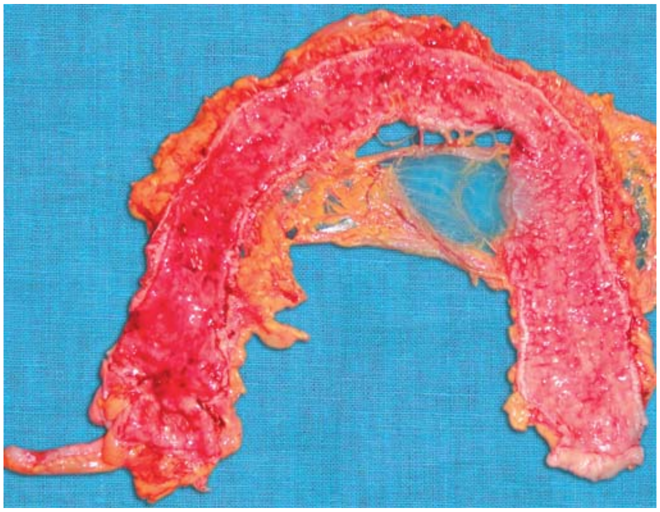

Fig. 22.19: Operated specimen showing colon with features of ulcerative colitis.

- It is a disease confined to mucosa and submucosa.
- There is no bowel wall thickening and no granuloma formation.
- There are no skip lesions. Rectum is always involved. Distal bowel is involved to begin with, then spreads proximally. Distal involvement is more severe. Entire colon including caecum and appendix may be involved.
- Only rectum involvement, as proctitis occurs in 25% cases. Such patients will have 5% risk of developing rectal cancer. In 15% cases, it is left sided ulcerative colitis presenting with severe recurrent diarrhoea. In 25% patients, total proctocolitis is the presentation. Bloody diarrhoea, malnutrition, complications like *toxic megacolon*, perforation (steroid may mask the features) and carcinoma are common here.
- Pseudopolyps are of inflammatory in nature. Absence of normal mucosa between these pseudopolyps is important to differentiate it from neoplastic polyps.
- Intense inflammation in the ***mucosa and submucosa*** with typical feature like:
  - *Multiple crypt abscesses*
  - *Sparing of the deeper layers* of the colonic wall
  - Inflammatory *pseudopolyps*
  - Multiple *pin point ulcers*
  - Increase in *substance p containing nerve fibres*
  - *Lymphoid hyperplasia* in mucosa and submucosa (25%)
  - Presence of *anti neutrophil cytoplasmic antibodies* with a perinuclear staining pattern (86%)
  - Decreased goblet cell mucin
- Only in **toxic megacolon** (1.5-2.5%) there is acute inflammation extending to entire thickness of the colonic wall including the serosa. It is not the colon that is toxic but it is the patient who is toxic, hence the name. It is precipitated by non specific causes, during barium enema study, due to drugs like opiates, antidiarrhoeal drugs and anticholinergics. Toxic megacolon commonly affects the *transverse colon* which will be more than 6 cm in diameter. Left colon or entire colon also may be involved. Caecum when rarely involved; becomes more than 10 cm in diameter. Colon will be like *wet blotting paper*. It is prone for perforation, peritonitis which carries high mortality (25-50%). C-reactive protein will be increased. It needs emergency laparotomy with total colectomy/proctocolectomy with ileostomy. Occasionally toxic megacolon can occur in pseudomembranous colitis, amoebic colitis or typhoid colitis.
- **Carcinoma in ulcerative colitis** is more prevalent than in Crohn‘s disease.

  Factors involved are:
  - *Extent* of involvement (more in total colonic).
  - *Duration* of the disease; *continuous active disease* than intermittent disease.
  - Incidence is 5% when duration is 15 years; 25% in 25 years; 35% in 30 years; 45% in 35 years; 65% in 40 years.
  - Incidence of developing cancer in left sided colitis is 10 years later than universal ulcerative colitis.
  - Incidence of carcinoma developing in ulcerative colitis is *equal* in both sexes.
  - Carcinoma in ulcerative colitis is commonly *aggressive and poorly differentiated, multicentric, synchronous, infiltrative and scirrhous;* half the patients will have colloid carcinoma (signet ring), more advanced at the time of presentation, *dysplasia* developing into cancer is common.
  - Ulcerative colitis with primary sclerosing cholangitis has still increased risk of developing cancer.
  - In ulcerative colitis, dysplasia is very important factor to transform into carcinoma. It may be mild, moderate or marked. It is often called as *dysplasia associated lesion or mass* (DALM).

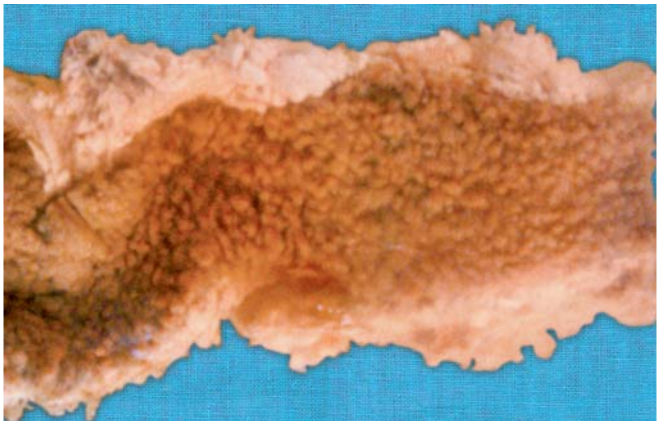

Fig. 22.20: *Multiple pseudopolyposis* involving entire colon in ulcerative colitis. Disease involves only mucosa and submucosa.

Clinical Features

Disease usually begins in rectum as **proctitis** later becomes **left sided colitis** and eventually causes severe **total proctocolitis.**

- More common in females (2:1), begins in 3rd decade.
- Watery diarrhoea, mucus or blood stained discharge per rectum.
- Colicky pain, spasms.
- Decreased appetite and loss of weight.
- Relapses and remissions at regular intervals.

Clinical grading of ulcerative colitis

	Mild	**Moderate**	**Severe**
Stool frequency	< 4	4-6	> 6
Pulse	< 90	90-100	>100
Haematocrit	Normal	30-40	< 30
Weight loss in %	None	1-10%	> 10%
Temperature	Normal	99-100	> 100
ESR	< 20	20-30	> 30
Albumin	Normal	3-3.5	< 3.5

Two types of *presentations:*

a. ***Fulminant type,*** 5% common:
   - It is a severe form, with continuous diarrhoea with passage of blood, mucus and pus.
   - Patient is ill and dehydrated.
   - Mimics fulminant amoebic colitis; severe typhoid and dysentery.
   - Fever, hypokalaemia, acidosis, dehydration and shock.
   - Abdominal distension occurs.
   - *Acute toxic dilatation* (1.5%) of *transverse colon* may occur where the diameter of transverse colon > 6 cm. It has high mortality and requires *emergency surgery,* i.e. either colostomy or resection with ileostomy and later ileo-anal anastomosis. Here colon is like *wet blotting paper*.

b. ***Chronic type*** (95%):
   - Lasts for months to years with diarrhoea, blood loss, anaemia, invalidism, abdominal discomfort and pain.
   - Severe malnutrition and hypoproteinaemia.

Note:

- Prevalence is 40-100/1,00,000.
- It is common before 30 years.
- Smoking increases Crohn's but protects ulcerative colitis.
- Incidence of colonic stricture is 10% in ulcerative colitis and is due to muscular hypertrophy. Strictures are usually benign. But 60% stricture appear after 20 years; 80% of stricture occur proximal to splenic flexure; stricture causing obstruction are likely to be malignant.
- *Perinuclear staining antinuclear cytoplasmic antibodies (pANCA)* are seen in 85% of patients with ulcerative colitis and is diagnostic test to differentiate from Crohn's.
- Arthritis (20%), ankylosing spondylitis (5%), erythema nodosum (15%), pyoderma gangrenosum are extraintestinal features *which resolve after total colectomy.*
- Primary sclerosing cholangitis (PSC, 5%) is common in ulcerative colitis before 40 years of age; common in males; related to HLA B_8 and HLA DR_3 (10 times). Malignant transformation in colon is 5 times more in these patients compared with ulcerative colitis without PSC. It causes pain abdomen, obstructive jaundice, later liver cirrhosis and failure. Total colectomy *will not resolve* PSC.
- *Surveillance colonoscopy* to identify transformation to carcinoma in ulcerative colitis is done as follows – Every year from 8 years after the onset of pancolitis and 15 years after the onset of left sided colitis. 10-30 random biopsies should be done. No dysplasia (less chance of carcinoma)/low grade dysplasia (10% chance)/high grade dysplasia (40% chance)/DALM (50% chance). All dysplasia types need proctocolectomy. Flow cytometry of biopsy specimens to study DNA aneuploidy or polyploidy is needed for further confirmation. 30% of high grade dysplasias show invasive carcinomas.

Investigations

- *Barium enema*—shows loss of haustrations, narrow contracted colon *(hose pipe colon)*, mucosal changes, pseudo-polyps. *It is avoided in fulminant cases.*
- *Sigmoidoscopy and biopsy.*
- *Colonoscopy* is also required.

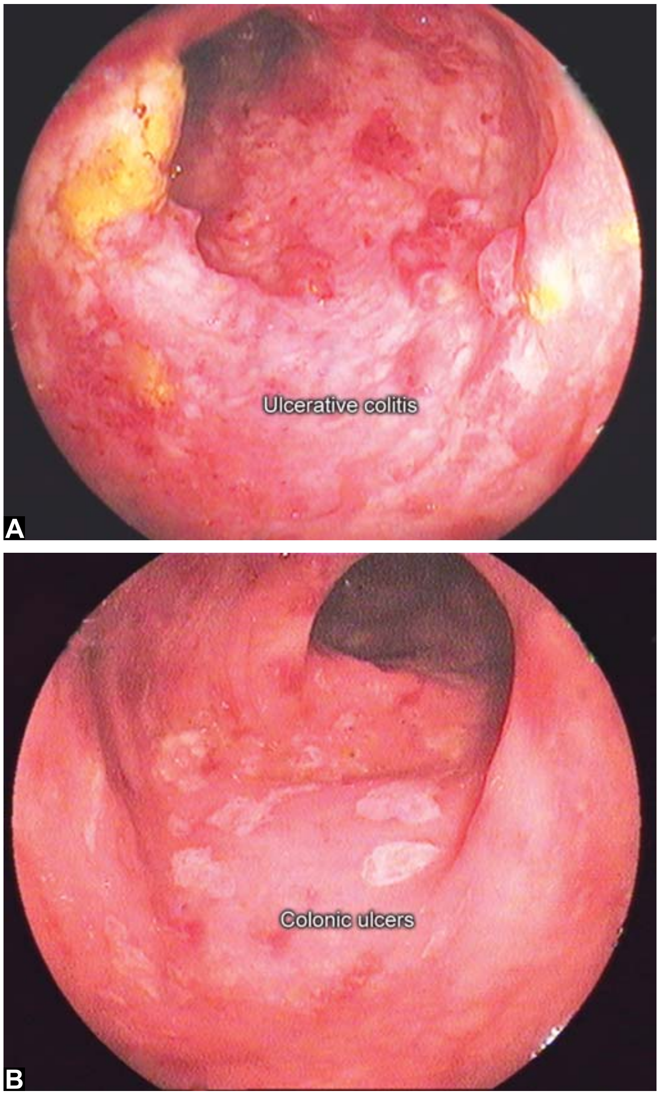

Figs 22.21A and B: Colonoscopic view of ulcerative colitis and multiple ulcers.

Due to very high incidence of malignant transformation in ulcerative colitis (10-20%), multiple biopsies should be taken from suspected areas of the colon. Risk increases with age of the patient and duration of the disease (20%).

Sigmoidoscopic grading of ulcerative colitis

0—Normal mucosa
1—Loss of vascular pattern
2—Granular, non-friable mucosa
3—Friability on rubbing
4—Spontaneous bleeding, ulcerations

- Plain X-ray abdomen is useful in obstruction, toxic megacolon, perforation.
- C reactive protein will be very high in acute phase.

Differential diagnosis

- Crohn's disease
- Ischaemic colitis
- Irritable bowel syndrome
- Amoebic colitis
- Bacillary dysentery
- Carcinoma colon
- Collageous colitis in females
- Infectious colitis by *Clostridium difficile, Campylobacter jejuni*

Differences between Crohn's disease and ulcerative colitis

Crohn's disease	Ulcerative colitis
It affects ileum and often colon but can involve any part of GIT—rectal sparing is common	It affects rectum and colon from distal to proximal
Full thickness—transmural disease	It affects mucosa and submucosa—not deeper
Skip lesions are typical	Skip lesions are not observed
Granulomatous lesion on histology—deep ulcers; pseudopolyps are not found	Not a granulomatous lesion—superficial ulcers, pseudopolyps are present
Stricture and fistula are common	Narrowing can occur but not very common
Anal fissure and perianal abscess are more common	Fissure and perianal disease can occur but not as common as in Crohn's disease
May mimic appendicitis	Will not mimic appendicitis
Bleeding is not common	Bleeding is common
Fever is common	Fever is uncommon
Mass in RIF is common	Mass is not a common feature
Anal pathology very common	Anal pathology is rare
Discontinuous segmental asymmetrical colitis	Continuous mucosal disease—colitis
Toxic megacolon is rare	Toxic megacolon is common
Normal vascular pattern	Distorted vascular pattern
Fistula is very common	Fistula is uncommon
Recurrence is common after resection	Total proctocolectomy cures the disease

Complications

GIT

- Pseudopolyposis
- Turning into malignancy
- Stricture formation, commonly in recto sigmoid and anal canal—10%
- Toxic megacolon in transverse colon
- Massive haemorrhage—1%
- Fistula in ano—20%
- Perforation—10-20%

Extraintestinal

- Severe malnutrition
- Liver cirrhosis (50%)
- Skin lesions—pyoderma, erythema nodosum
- Arthritis, iritis, ankylosing spondylitis—common
- Sclerosing cholangitis, carcinoma of bile duct

Treatment

General

- Correction of anaemia.
- Fluid and electrolyte supplimentation.
- Nutrition (high protein, carbohydrate, vitamin, but low fat diet), TPN.
- Sedatives and tranquillisers.
- Psychological counselling.

Drugs

In active disease, drugs are used to *induce remission*. Later drugs also should be given for *maintenance* of remission and to prevent relapses.

- *Salazopyrine/sulfasalazine* splits into 5 amino-salicylic acid and sulphapyridine in colon. It is used as *first line therapy*. Its dose is 2-4 gm/day. It is mainly used to induce remission. It suppresses PG E1 and PG E2 to reduce the inflammation, inhibits proteolytic enzymes and also causes immunosuppression. 5 ASA is the most active ingredient. Side effects are skin rashes, bone marrow suppression, folic acid deficiency, haemolysis in glucose 6 phosphate dehydrogenase deficiency patients, temporary fertility problems in men.
- *5 ASA (Mesalamine)* is also used in active disease as *first line therapy*. It is used with an azo bond to prevent its absorption in small bowel. Its induction of remission is same as sulfasalazine. Idiosyncrasy and temporary infertility is not present in mesalamine. Its *oral* dose is 2-4 gm/day. It is also used as *retention enema* which is better than steroid enema in left sided ulcerative colitis/proctitis. But mesalamine enema (4 gm in 100 ml saline) is costly. Enema is combined with oral mesalamine or oral steroid in acute cases. Mesalamine 500 mg s*uppository* is also used in BID doses. Daily oral (1.6

gm) and topical 4 gm enema, twice weekly is often used for maintenance therapy. Headache, dyspepsia and myocarditis are the complications.

- *Steroids* are used in cases where salazopyrine fail to induce remission. It is a drug for *refractory cases*. Oral prednisolone 60 mg/day tapering in 4 weeks is the dose. Intravenous hydrocortisone 100 mg 8th hourly is used in acute cases. 100 mg hydrocortisone enema/40 mg methylprednisolone enema (steroid retention enema) are used for 2 weeks in acute active diseases. Most patients with moderate to severe disease need steroid therapy. *Budesonide* is a newer steroid hydrocortisone analogue is also equally effective with lesser side effects and less adrenal suppression.
- *Immunomodulators* are used often to induce remission and for maintenance. *Azathioprine and 6-mercaptopurine* are used. They act at DNA level (purine ribonucleotide) and inhibit lymphocyte function (of T cells). But this action is slow.
- *Cyclosporin* is used in refractory, fulminant severe ulcerative colitis as a reserve drug. Dose is 4 mg/kg/day IV.
- *Mebeverine HCl* and *Tegaserod* (6 mg) are the other drugs used in ulcerative colitis. Other measures are using sucralfate, short chain fatty acids, probiotics, antidiarrhoeal drugs (diphenoxylate, loperamide, codeine), avoiding milk products, fibre, fruits.
- Antitumour necrosis factor alpha—*Infliximab* is also used selectively (in some studies). It is given intravenously at interval of 6 weeks. It shows 70% remission in ulcerative colitis.
- Newer drugs with a targeted delivery into the colon is used nowadays, e.g. Olsalazine (azobond); Balsalazine (4 aminobenzoyl carrier).

Management of severe ulcerative colitis

- It is more than 6 stools per day, with cramping pain, toxicity, fever, raised ESR, anaemia, tachycardia, more than 10% weight loss, hypoalbuminaemia less than 3.5 gm%
- It always needs hospitalisation
- IV corticosteroids [hydrocortisone] 300 mg/day/oral prednisolone/5 ASA orally or as enema
- IV cyclosporine with azathioprine or 6 mercaptopurine
- Fluid and electrolyte management
- TPN/enteral nutrition
- Proper monitoring of the patient for complications
- Surgery

Indications for Surgery—30% Cases

- Intractability—commonest indication
- Toxic dilatation
- Perforation
- Haemorrhage
- Risk of malignant transformation, dysplasia (DALM)
- Onset at early age
- Chronic invalidism
- Progressive disease with stricture, abscess, fistulae
- Steroid dependency, persistent active disease
- Malignancy
- Severe extraintestinal manifestations
- Growth retardation in children

Toxic megacolon

- 1.5% incidence
- Common in transverse colon
- Colon is more than 6 cm in diameter
- Colon is like wet blotting paper
- Commonly seen in ulcerative colitis
- Can occur in bacterial colitis, pseudomembranous colitis, fulminant amoebic colitis
- May be precipitated by antidiarrhoeal drugs
- IV fluids, blood transfusion, antibiotics, steroids are also needed
- Can be life threatening
- Plain X-ray is very useful
- Emergency surgery, colectomy, colostomy/ileostomy may be required
- In olden days *blow hole procedure* was advocated for toxic megacolon, may be still useful in severe cases. Skin level transverse colostomy, sigmoid colostomy and a loop ileostomy is done to decompress entire small and large bowel. After many months (6-8 months) once patient recovers from acute illness, total proctocolectomy with ileal pouch anal anastomosis is done

Surgeries

- *Total proctocolectomy with ileo-anal anastomosis with pouches as reservoir* ("J', 'S', or "W' pouches). It is called as *restorative proctocolectomy with ileal pouch anal anastomosis (IPAA)*. It is *ideal curative procedure* for ulcerative colitis. Anal sphincter complex is preserved. 30 cm ileal pouch is created either hand sewn or with stapler. Created ileal pouch reservoir is anastomosed just above the anal canal using end to end stapler. Diameter of the pouch should be twice the diameter of the ileum. Obstruction (25%), diarrhoea, *pouchitis (25%)*, leak, sepsis, incontinence are the complications.

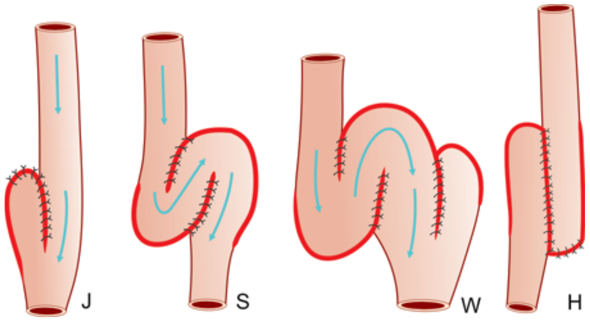

Fig. 22.22: Different pouches are used after restorative proctocolectomy and during ileoanal anastomosis. Pouch acts as reservoir and reduces the frequency of stool. It avoids ileostomy with natural stool passage with continence. Pouchitis can occur with diarrhoea, blood in stool, perianal pain. Occasionally, redoing or conversion may be required. J type is commonly used.

- Total proctocolectomy with ileostomy (permanent, continent Kock's ileostomy, with one way valve is done). It is used at present only *in revision surgery* when needed after restorative proctocolectomy.
- *Total proctocolectomy with end non continent ileostomy* was the earliest operation done for ulcerative colitis. It is now *reserved only* for ulcerative colitis complicated with

carcinoma of rectum. *Ileostomy site* should be marked in standing and sitting position. Stoma should be within the right rectus abdominis muscle at the summit of infraumbilical fat, away from midline incision, bony prominences and umbilicus.

- *Total colectomy* with ileorectal anastomosis. Proper follow-up at regular intervals by regular sigmoidoscopy evaluation should be done as rectum is also diseased and vulnerable for complications. It is *not commonly* practiced now.
- Total colectomy with rectal mucosectomy and anastomosis above the dentate line on posterior aspect is also occasionally used.

Ileostomy

It is usually end ileostomy. But often loop ileostomy is also done

Indications:

- After total proctocolectomy for ulcerative colitis, total colonic Hirschsprung's diseases, Crohn's disease (occasionally), and total colectomy for carcinomas
- Loop ileostomy is done in critically ill patient with acute ileal conditions like multiple ileal perforations, ileal gangrene or distal fistulas or sepsis. It is temporary ileostomy
- Ileostomy is *placed in right iliac fossa* through the right rectus muscle.
- Ileostomy may be temporary (loop) or permanent (end)
- Brooke's classic end non continent ileostomy or Koch's continent ileostomy (with continent intra-abdominal pouch) are used
- Ileum should project as a spout at least 4 cm above the skin surface, which facilitates the effluent to pass directly into the bag
- Ileostomy usually acts in 48 hours
- Ileostomy bag, ileostomy care, nutrition and electrolyte management are important
- **Complications** like prolapse, stenosis, haemorrhage, retraction can occur

Complications of Surgery for Ulcerative Colitis

- Pouchitis (20%) with pain, diarrhoea, fever, bleeding, toxicity, pouch—vaginal fistula, foecal incontinence (5%). Pouchitis disease activity scoring index is at present used which is based on clinical, endoscopic and histological inflammatory features.
- Stenosis, pelvic abscess formation.
- Leak, fistula formation.
- Problems with ileostomy—psychological trauma, skin excoriation, retraction, stenosis (25%), prolapse, bleeding, enteritis, ileal necrosis, ileal volvulus, paraileostomy hernia, paraileostomy abscess.
- Sexual dysfunction following proctocolectomy by nerve injury.
- Persistent perineal sinus after total proctocolectomy.
- Risk of cancer persists if only total colectomy is done.

ISCHAEMIC COLITIS

It occurs in **splenic flexure** where blood supply is precarious. Splenic flexure is the *water shed* area of colon, receiving blood supply from terminal branches of superior and inferior mesenteric arteries.

- Ischaemic colitis is common in females; common in aged.
- It is related to atherosclerosis, emboli, vasculitis, diabetes, chronic renal failure, autoimmune diseases, polycythaemia, haemodialysis, etc.
- *Water shed point*—Griffith's point in artery of Riolan gets poor perfusion by arterial disease/low perfusion pressure/ altered viscosity causing ischaemic colitis.

Types (Marston's Classification)

1. Gangrenous type—*ischaemia of full thickness* colon causing peritonitis.
2. Stricture type— *ischaemia of muscularis* layer causing scarring.
3. Transient type—most vulnerable layer, *mucosal* involvement usually recovers completely.

Clinical Features

- Pain in left iliac fossa and left hypochondrium.
- Vomiting, diarrhoea.
- Passing blood in the stool.

Differential Diagnosis

- Carcinoma colon.
- Ulcerative colitis.
- Crohn's disease.
- Tuberculosis.

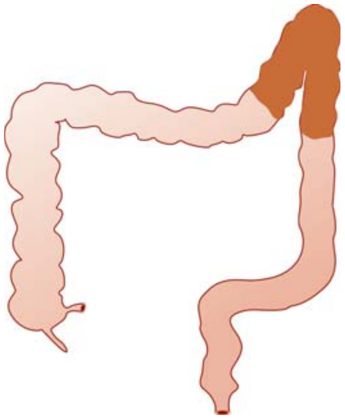

Fig. 22.23: Ischaemic colitis in splenic flexure.

Investigation

- *Plain X-ray* reveals '*thumb printing sign*' due to mucosal oedema and submucosal haemorrhage. Perforation is also diagnosed.
- *CT scan shows* colonic wall thickening with posterior fat shadowing. Angiography is not helpful.
- *In acute stage,* barium enema or contrast study or sigmoidoscopy or colonoscopy is avoided due to risk of perforation.
- *In chronic stage*, colonoscopy and contrast study is a must.

Treatment

- 80% of patients will recover from conservative treatment—bowel rest, fluids, antibiotics, adequate perfusion.
- Surgery is indicated in gangrene and peritonitis, stricture (15%), segmental ischaemia (20%).
- Laparotomy, resection and anastomosis (splenic flexure) is done.
- In acute phase with peritonitis (covering), diversion colostomy is needed.

Colonic ischaemia

- Colon is the *commonest* site of intestinal ischaemia
- Most of colonic ischaemia results from *small vessel occlusion* or low flow ischaemia.
- Vascular diseases, vasculitis, diabetes, hypotension, aortic surgery with ligation of inferior mesenteric artery, aortic atherosclerosis blocking opening of IMA, OCP intake, cocaine abuse, coagulopathies, CMV, *E. coli* infections, long distance running—are the causes
- *Splenic flexure* is the most common site; but any segment like sigmoid colon can be involved
- Rectum is *spared* due to rich collaterals

PSEUDOMEMBRANOUS COLITIS

- It is an acute diarrhoea due to toxins produced by the overgrowth of *Clostridium difficile* after antibiotic therapy (usually after clindamycin).
- *Clostridium difficile* is a Gram positive, anaerobic, spore forming bacillus. It produces toxin A (enterotoxin) and toxin B (cytotoxin, more potent).
- It is also often seen in immunocompromised patients and patients who are on cancer chemotherapy.
- Incidence is 2%. Mortality is 30%.
- Stool cytotoxin assay is highly sensitive and specific. ELISA test for toxins is also useful.
- Colonoscopy is ideal as right side involvement is more common.
- It can occur up to 6 weeks after stopping the drug.
- Diarrhoea, toxaemia, perforation, haemorrhage also can occur.

Treatment

- IV vancomycin 500 mg 8th hourly.
- IV metronidazole 8th hourly.

SURGICAL COMPLICATIONS OF INTESTINAL AMOEBIASIS

Trophozoites of *Entamoeba histolytica* by digesting mucosa, submucosa of *rectosigmoid* (75%) region commonly or *ileocaecal* region of the colon, causes *retort shaped amoebic ulcers* which exudes blood, pus and necrotic material.

- Cyst is the infective agent. It enters through faeco oral route. It forms trophozoite which multiplies and causes inflammation and *flask shaped ulcers* in the rectosigmoid, caecal and often in the terminal ileal region.
- It causes blood with mucus diarrhoea, toxaemia, secondary bacterial infection, fulminant colitis, perforation and peritonitis, rarely toxic megacolon, extraintestinal amoebiasis.
- Amoebic liver abscess is the *commonest form* of extraintestinal amoebiasis. Others are cutaneous amoebiasis, amoebic empyema, and very rarely amoebic brain abscess, amoebic pericarditis.
- *Chronic amoebiasis* causes vague abdominal pain, decreased appetite, intestinal colic and psychological trauma to the patient.

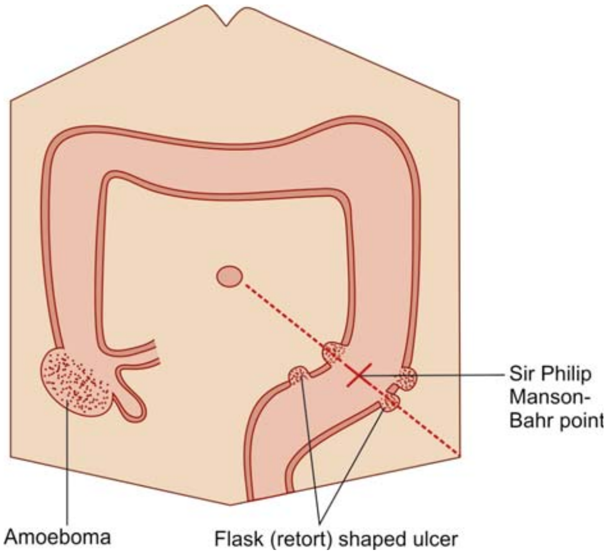

Fig. 22.24: Amoebic ulcer is common in left side (sigmoid colon). Amoeboma can occur in caecal region where it forms granuloma in pericolic area presenting as mass abdomen. Amoebic ulcers are classically flask/retort shaped.

Presentations

1. *Amoebic dysentery* with diarrhoea, colicky pain, tenderness in left iliac fossa (in" *Sir Philip Manson-Bahr amoebic point'*).
2. *Amoebic typhlitis* (inflammation of caecum) presentation with pain and tenderness in right iliac fossa and often also as a *mass in the right iliac fossa mimicking carcinoma caecum* called as ***amoeboma*** (amoebic granuloma) (1.5%).

Differential diagnosis for amoeboma

- Appendicular mass
- Ileocaecal tuberculosis
- Carcinoma colon
- Retroperitoneal tumour
- Lymph node mass

3. *Acute fulminant amoebic colitis* is a severe type with sloughing of colonic and rectal mucosa causing torrential bleeding, toxicity which can be life-threatening.

Investigations for intestinal amoebiasis

- Saline wet mount of fresh stool—trophozoites in 90% cases
- Serology—indirect haemagglutination test
- PCR
- P/R and proctoscopy

Complications of amoebic colitis

- *Amoeboma*—right side common
- Perforation in rectosigmoid/caecal region bleeding, peritonitis
- Stricture rectum and colon
- Intestinal obstruction
- Pericolic, paracolic, ischiorectal abscess and fistula formation
- Amoebic typhlitis, amoebic liver abscess

Treatment

- In acute cases, hospitalization and proper rehydration is done.
- Antispasmodics to alleviate pain is given.
- Metronidazole IV or metronidazole retention enema is given. It is very effective in both intestinal and extraintestinal amoebiasis.
- In less severe disease, tab metronidazole (400-800 mg) 3 times daily is given for 10 days or tab tinidazole 2 gm daily for 3 days.
- Diloxanide furoate is very effective in chronic amoebiasis and cyst passers. Dose is 500 mg, 3 times daily for 10 days.
- Other drugs used are—paromomycin, iodoquinol.

TUMOURS OF COLON

BENIGN TUMOURS/POLYP OF THE COLON

- *Polyp* is a tumour/swelling arising from mucosal surface with a pedicle/stalk. '*Poly*' means many; '*pous*' means foot in Greek.
- *Polyp* is a mass projecting into the bowel lumen beyond the surface epithelium.

Classification

Inflammatory

- Ulcerative colitis.
- Segmental colitis.
- Crohn's disease.
- Diverticulitis.
- Dysenteric colitis.

Hyperplastic (Metaplastic)

- Also called as metaplastic mucosal polyps.

Hamartomatous

- Peutz-Jegher's syndrome.
- Juvenile polyp.
- Cronkhite—Canada syndrome.

Neoplastic

- Tubular—pedunculated.
- Villous—sessile.
- Tubulo—villous.
- FAP—familial adenomatous polyposis.

Others

- Lipoma, haemangioma, leiomyoma.

JUVENILE POLYPS

- *Commonest* polyp of colorectum in *infants and children.*
- Can cause intussusception, prolapse through rectum, bleeding.
- Colonoscopic polypectomy is done.
- Not a premalignant condition.

METAPLASTIC/HYPERPLASTIC POLYP

- Metaplastic—indicates a difference in appearance from normal mucosa.
- Very minute in size—1-2 mm. Multiple.

Polyps

Inflammatory	Hyperplastic	Hamartomatous	Neoplastic
Ulcerative colitis	(Metaplastic)	Peutz-Jegher syndrome	Adenomatous
Segmental colitis		Juvenile polyp	– Tubular
Dysenteric colitis		Cronkhite-Canada syndrome	– Tubulovillous
Crohn's disease			– Villous
Diverticulitis			Malignant (Carcinomatous)

Children, like animals use all their senses to discover the world.

- Common in rectum. Also occurs in other parts of colon. It is most common colonic polyp.
- Contains columnar epithelium, cystic dilatation, goblet cells, and lymphocytes.
- Not a pre-malignant entity.

PEUTZ-JEGHER'S POLYP

- It is common in *small intestine* (jejunum) but can also occur in large intestine.
- ***Features*** are multiple, familial, hamartomatous intestinal polyps.
- Asssociated with *melanosis* of the oral mucosa, lips (lower lip) and occasionally digits (*not in tongue*).
- Microscopically it contains tree like branching filaments of mucosa with smooth muscle wall.
- It can occasionally turn into malignancy.
- It is autosomal dominant disease with germline defect in suppressor serine threonine kinase 11 (STK11).

Complications

Bleeding or intussusception, when occurs requires surgery either resection—anastomosis or colonoscopic removal.

ADENOMA OF COLON

- It can be *tubular, villous, tubulovillous (*Histologically). Tubular is the commonest—70%. 25% tubulovillous; 5% villous.
- It also can be *solitary or multiple*, sessile or pedunculated.
- Tubular is commonly 1-2 mm sized and sessile is 3-4 mm sized.

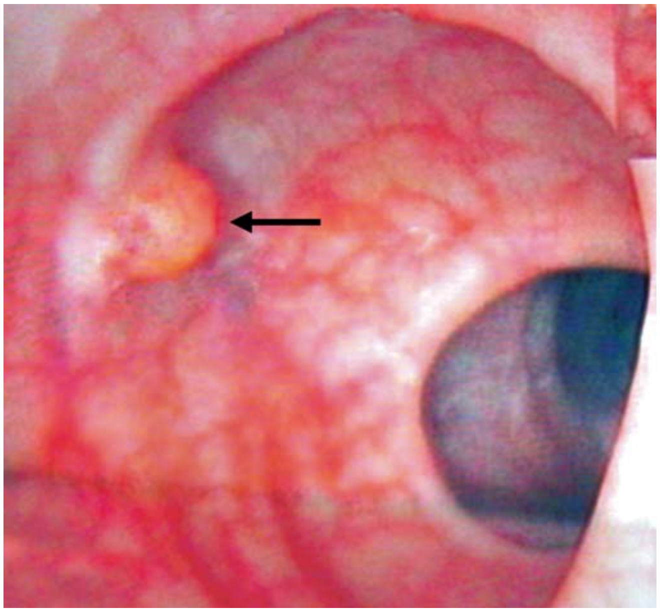

Fig. 22.25: Colonoscopic view of colonic polyp.

- ***It has malignant potential***.
  Potentiality increase with:
  - *Size, is an important factor* in causing carcinoma. If size of adenoma is > 2 cm—30-50% chances of developing carcinoma; 1-2 cm —10% chances of carcinoma; < 1 cm —1-5% chances of carcinoma.
  - Sessile nature.
  - Villous architecture.
  - Dysplasia.

Grading of adenoma

- I: Minimal hyperplasia, no cellular atypia
- II: Mild hyperplasia, cellular atypia
- III: Moderate hyperplasia, cellular atypia
- IV: Severe hyperplasia, cellular atypia
- V: Carcinoma *in situ*

Features of Adenoma

- Can be asymptomatic.
- Bleeding per anum is usually chronic but rarely can be acute.
- Anaemia.
- Prolapse—common in tubular type.
- Diarrhoea common in villous type; mucus discharge.
- Tenesmus, colicky abdominal pain.
- Spurious diarrhoea.
- Poor general health.
- Electrolyte imbalance—*hypokalaemia.*
- Per-rectal examination should be done to feel the adenoma/polyp.
- Carcinomatous changes.

Investigations

- Serum electrolytes.
- *Barium enema* study—shows filling defect—usually multiple.
- *Colonoscopy—biopsy is a must.*
  - Size should be noted.
  - Texture—harder tumour more likely to be malignant.
  - Colour of the lesion—pale is benign, pink, red and active—could be carcinoma.
  - Ulceration on the surface if it has turned into malignancy.

Treatment

- Colonoscopic polypectomy using snare. Any adenoma more than 5 mm in size should be removed colonoscopically.
- Diathermy excision/coagulation with sigmoidoscope.
- Per-anal polypectomy.
- Per-anal excision with clear margin of the rectal sessile adenoma.
- Open-abdominal colotomy and polypectomy in case of huge adenoma.
- Segmental resection of the colon if polyps/adenomas are limited to one segment of the colon confirmed by colonoscopy.

- Total colectomy/proctocolectomy if multiple polyps present all over colorectum and if associated with FAP.
- Blood transfusions, correction of electrolytes, protein supplements.

Problems in therapeutic colonoscopy

- Perforation due to necrosis
- Haemorrhage—secondary
- Intracolonic explosions
- Sepsis

FAMILIAL ADENOMATOUS POLYP (FAP)

- It is inherited as an autosomal dominant neoplastic condition (chromosome no. 5q21).
- It presents in younger age group—15-20 years.
- Incidence is equal in both sex, involving commonly the large intestine but can also occur in stomach, duodenum and small intestine.
- It is familial with a *high potential for malignant transformation*. If there is no adenoma at the age of 30 years, then it is not FAP of colon.
- It can be associated with duodenal or ampullary carcinomas, *Gardner's syndrome* [Desmoid tumour in the abdomen, osteomas (75%) and epidermoid cysts] and also *Turcot"s syndrome* [FAP + brain tumour (medulloblastoma or gliomas)] or sarcoma of bone.
- Usually multiple (over 100).
- Presents with lower abdominal pain, loose stools with blood and mucus, weight loss.

Investigations

- Double contrast barium enema.
- Colonoscopic biopsy.

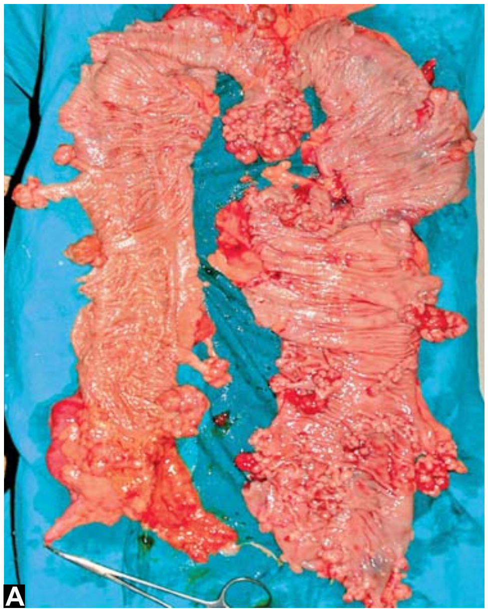

Fig. 22.26A

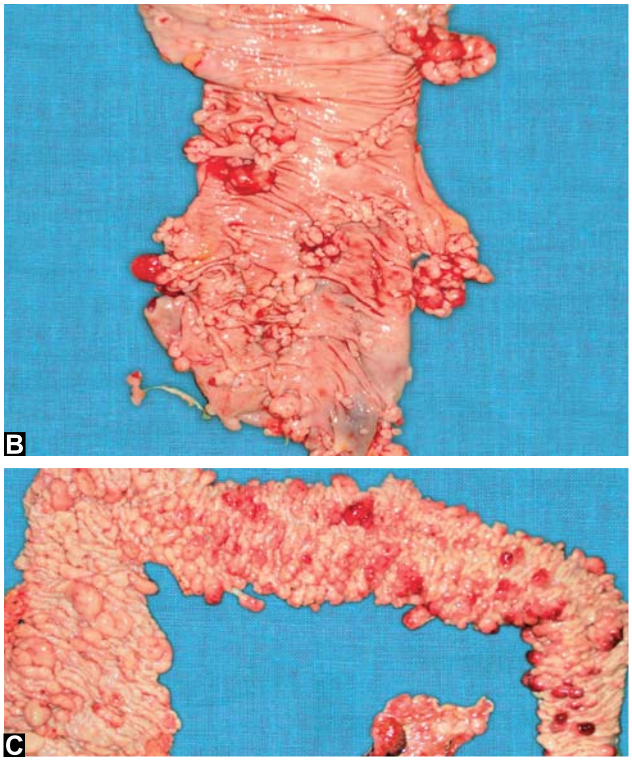

Figs 22.26B and C

Figs 22.26A to C: Multiple polyps extensively involving the colon.

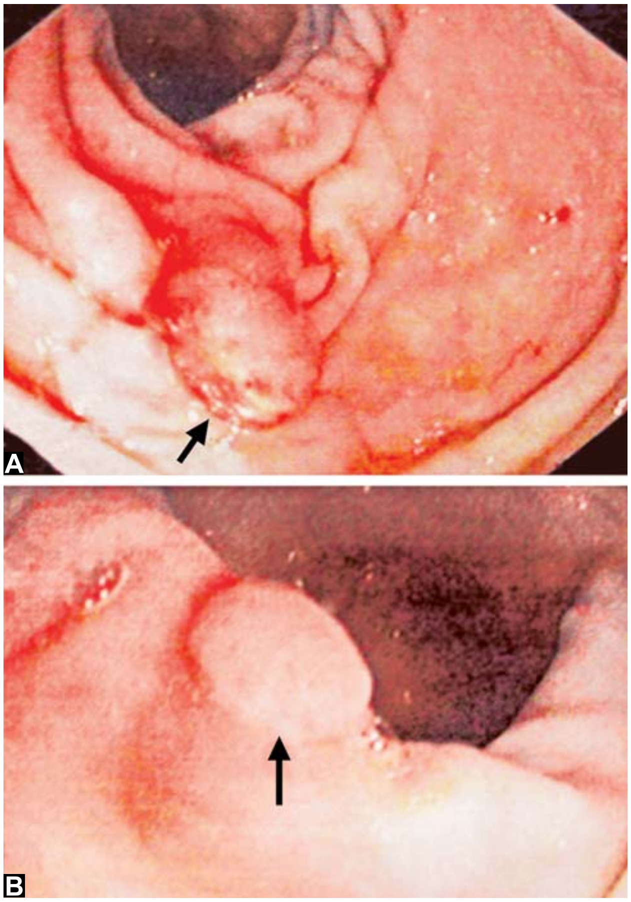

Figs 22.27A and B: Colonoscopic view of polyp colon in two different patients.

Self-respect — that a corner stone of all virtue.

- Screening of *all the members of the family is a must*.
  Pigment spots in retina (Congenital hypertrophy of the retinal pigmented epithelium of the iris, CHRPE / CHIRPES) and DNA tests for FAP are good screening methods, but cost and availability limits its use.
- Patients with FAP will develop colorectal cancer almost 100% in the absence of surgical intervention.

Treatment

1. *Proctocolectomy with ileoanal anastomosis with ileal pouch* is commonly done.
2. Alternatively a conservative *total colectomy* with ileo rectal anastamosis can be done, but with a regular follow-up with sigmoidoscopy for any *rectal polyps.* If present *snaring of polyps* should be done.

Sulindac (NSAID) 300 mg BD/*aspirin* 325 mg OD is given to these patients, causes disappearance of polyps in the rectum.

GARDNER'S SYNDROME

- It is commonly associated with FAP—10%.
- Presents with bone, skin, soft tissue and dental abnormalities. Jaw osteomas are very common. Other features are epidermoid cysts (50%), exostoses, fibromas, lipomas.
- Associated with desmoid tumours seen in the scar, abdomen, intraabdominal region and mesenteric fibromatosis.
- Congenital hypertrophy of pigment layer of retina (seen as pigment spots)—commonly seen.
- Often associated with MEN IIb syndrome.

Turcot's syndrome

- Autosomal recessive disorder
- Colonic polyps + brain tumours like
- medulloblastoma or gliomas
- Often associated with FAP

Cronkite-Canada syndrome

- Polyps in stomach, duodenum and colorectum. No polyps in oesophagus, small intestine
- Intractable diarrhoea, pigmentation, cachexia, alopecia, onychodystrophy
- In females
- 15% will have malignancy

Cowden syndrome
(multiple hamartoma—neoplasia syndrome)

- Autosomal dominant
- Penetrance at age 20 with PTEN suppressor gene mutation
- Benign ectodermal tumour—tricholemmoma—80%
- Macrocephaly—40%
- GI polyposis—35% and no risk of GI malignancy
- 10% risk of thyroid cancer
- 30-50% risk of breast cancer

Bannayan-Riley-Ruvalacaba syndrome
(Ruvalacaba Myhre Smith)

- Autosomal dominant
- GI hamartomatous polyp
- Macrocephaly, mental retardation
- Delayed psychomotor development
- Lipid storage myopathy
- Hashimoto's thyroiditis
- Hyperpigmentation of penile skin
- No risk for GI or extra GI malignancy

CARCINOMA COLON

- It is commonly *adenocarcinoma*.
- Very rarely adenosquamous, squamous carcinoma can occur.

Adenocarcinoma

- Sigmoid colon (21%) is the most common site of malignancy after rectum (38%).
- In caecum it is 12% common.

Aetiology

- ***Diet:***
  - Red meat and saturated fat increases the incidence of colonic cancer.
  - Cholesterol increases the bile acid concentration in the intestinal lumen which acts as cocarcinogen.
  - High fibre diet protects the colon against cancer.
  - Calcium in diet prevents colonic cancer by combining with bile salts and reducing bile salt concentration in the colon. It directly acts on the colonic mucosal cells to reduce their proliferative potential.
  - Diet with lack of fibre increases the risk. Diet with high fat increases the risk.
  - Dietary vitamins A,C, E and zinc reduces the risk.
- ***Genetic:***
  - Carcinoma colon is more common in individuals with *adenoma colon* or with *familial adenomatous polyposis (FAP*), Gardner's syndrome, Turcot's syndrome.
  - Relatives of colonic cancer patient have got 2-4 times increased risk of developing carcinoma of colon.
- Long standing *ulcerative colitis, Crohn's disease* has high risk of colonic cancer. Crohn's disease is a premalignant condition but not as much as ulcerative colitis.
- *Alcohol and cigarette smoking increases the risk.*
- *Hereditary nonpolyposis colonic cancer (HNCC)* has got high incidence (25%) of synchronous and metachronous growth, so total colectomy is needed.
- After cholecystectomy and ileal resection there is increased bile salts and so more prone for carcinoma colon.
- Radiation increases the risk (mucinous type).

- Ureterosigmoidostomy increases the risk by 100-500 times.
- Acromegaly may increase the risk.

Note:

Aspirin, calcium and other NSAIDs protect against colonic cancer.

Pathogenesis

- Adenoma—carcinoma sequence
  - Most of the colonic carcinoma develops from polyp/adenoma pathway.
  - Normal epithelium →initiation by 5q loss APC gene → dysplasia (hyperproliferative) → DNA methylation → early adenoma → 12p activation K ras → intermediate adenoma →18q loss DCC → late adenoma → action by 17p loss p53 →carcinoma → spread .
- 80% of colorectal cancer arises from *loss of heterozygosity (LOH) pathway*. LOH pathway is due to APC gene defects (in FAP), K ras mutation altering the cell cycle [K ras binds to GTP (guanosine triphosphate) hydrolyse to GDP which inactivates G protein normally; K ras mutation blocks GTP hydrolyse leading into permanently active form of G protein causing carcinoma]; loss of DCC tumour suppressor gene; mutation of tumour suppressor gene p53. LOH pathway is microsatellite stable (MSS) and carries poor prognosis compared with MSI.
- 20% of colorectal cancer develops from mutation from *RER (Replication Error Repair) pathway* wherein repair mechanism of DNA replication error is lost. It causes microsatellite regions of genome to have repeated sequences leading into error and is called as *microsatellite instability* (MSI). In colon, it is seen in right side growths and is associated with better prognosis.

Colonic cancer may be:

Nonhereditary colon cancer

- It can be sporadic colon cancer—60%.
- It can be familial colon cancer—30%. Common in Ashkenazi – Jewish population.

Hereditary colon cancer

- FAP.
- HNCC.
- Peutz Jeghers syndrome—2-3% risk of cancer colon.
- *Cronkite—Canada syndrome.*
- *Juvenile polyposis syndrome*—it differs from isolated juvenile polyps discussed earlier. It is an autosomal dominant condition, occurs in children and adolescent. Germ line mutation of SMAD-4 gene is observed. It increases the risk of colonic cancer.

HNCC (hereditary nonpolyposis colonic cancer)

- No polyps. Autosomal dominant
- Three members of the family have colonic cancers
- Two first degree relatives will have same cancer
- Two consecutive generations observed
- One relative with less than 50 years age will have colonic cancer
- *Lynch syndrome* I is site specific—commonly right sided, occurs in early age group, 40% are metachronous
- *Lynch syndrome* II has other malignancy in, stomach, breast, ovary, endometrium and urinary bladder. It is cancer family syndrome
- Microsatellite instability (MSI) at DNA level occurs in HNCC
- Accounts for 3-5% of colonic cancers
- Amsterdam criteria I (1990); Amsterdam criteria II (1999) and revised Bethesda guidelines (2002) are used to diagnose HNCC

Types

Patient can have *de novo multiple primary* carcinomas in different parts of the colon at the same time, i.e. ***synchronous* (5-10%),** or can present with growth in different parts of the colon in different periods, i.e**. *metachronous* (10-20%).**
Gross types: *Annular, tubular, ulcerative, cauliflower like.*

Amsterdam criteria	Bethesda criteria
	Amsterdam criteria or one of the following:
• At least 3 relatives of common cancer	• 2 cases of HNPCC associated cancer in one patient including synchronous or metachronous cancer
• One affected person is 1st degree relative of other two affected persons	• Colon cancer and 1st degree relative with HNPCC associated cancer and or colonic adenoma
• Two successive generations affected	• Colon or endometrial cancer diagnosed before the age of 45
• At least one case of colon cancer diagnosed before the age of 50 years	• Right sided colon cancer that has an undifferentiated pattern or signet cell type before the age of 45
• FAP excluded	• Adenoma diagnosed before 45
Modified Amsterdam criteria	
Same as Amsterdam except cancer must be associated with HNPCC (colon, endometrium, small intestine, renal pelvis and ureter) instead of specifically colon cancer	

Read fully; understand exactly; write precisely; speak essentially.

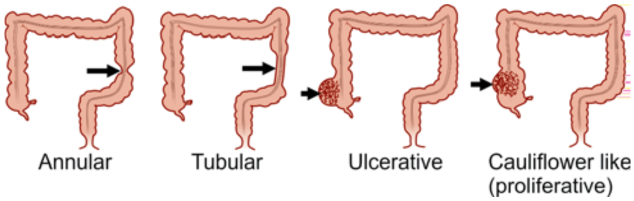
Annular
Tubular
Ulcerative
Cauliflower like (proliferative)

Fig. 22.28: Gross types of carcinoma colon. Annular and tubular types are common in left side, ulcerative and proliferative are common in right side.

Annular (stenosing) type:

- It is more common on left side.
- Here the growth spreads round the internal wall and so it often presents with intestinal obstruction.

Ulcerative type:

- It is common on right side.

Proliferative type:

- Common in right side. It is fleshy, bulky and polypoid. It is less malignant.

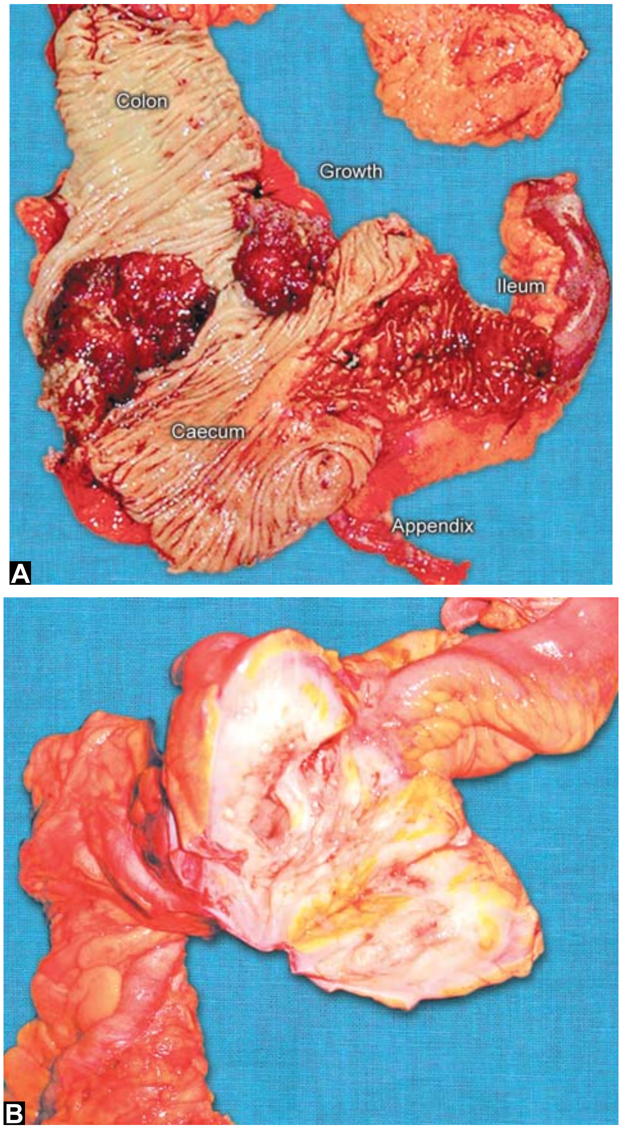
Colon
Growth
Ileum
Caecum
Appendix
A
B

Figs 22.29A and B: (A) Carcinoma caecum showing proliferative lesion. (B) Thickening of the wall is seen.

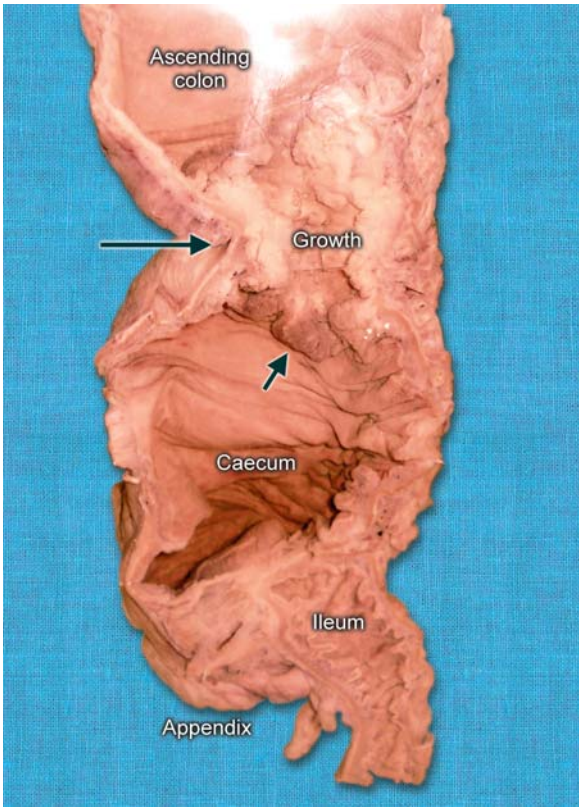
Ascending colon
Growth
Caecum
Ileum
Appendix

Fig. 22.30: Carcinoma ascending colon. Note the growth, caecum and ileum. It is proliferative type.

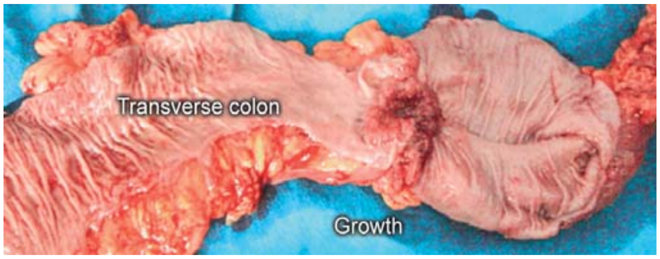
Transverse colon
Growth

Fig. 22.31: Carcinoma transverse colon. It is ulceroproliferative lesion.

Histology (WHO)

- Adenocarcinoma—90%.
- Mucinous adenocarcinoma—5-10%.
- Signet ring cell carcinoma.
- Small cell/oat cell carcinoma—rare—extremely poor prognosis.
- Squamous cell carcinoma.
- Undifferentiated carcinoma.

Duke's histological grading of carcinoma colon (Now modified Morson-Dawson)

- Grade I—low grade.
- Grade II—average grade.
- Grade III—high grade.
- Grade IV—anaplastic.

Carcinoma confined to muscularis mucosa does not metastasize.

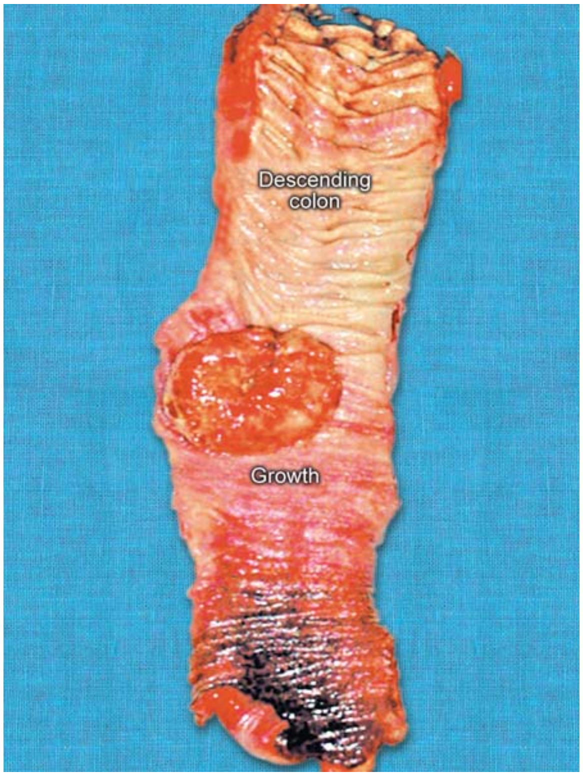

Fig. 22.32: Carcinoma descending colon. Note the proliferative tumour. Commonly it is stricture type/annular type on left side.

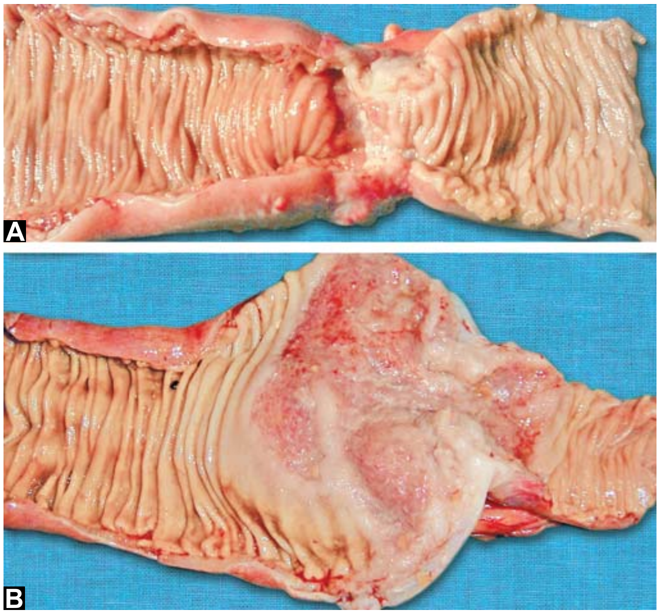

Figs 22.33A and B: Annular type of growth observed in left sided/ sigmoid colon.

Haggitt's Invasion of Malignant Polyp

In pedunculated polyp

Level 0—noninvasive carcinoma over the summit.
Level 1—invasion to head of the pedunculated polyp.
Level 2—invasion to neck of the pedunculated polyp.
Level 3—invasion to stalk of the pedunculated polyp.
Level 4—invasion to base of the pedunculated polyp.
In sessile polyp—all lesions are level 4.

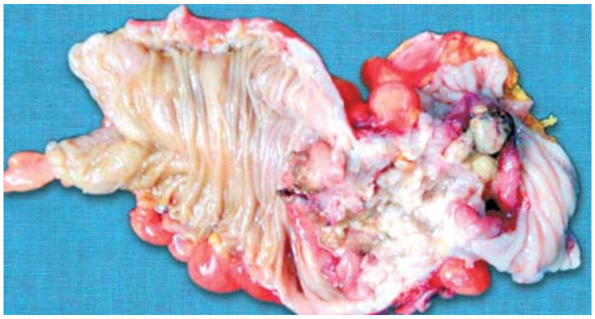

Fig. 22.34: Ulceroproliferative lesion in the colon. Note the everted edge.

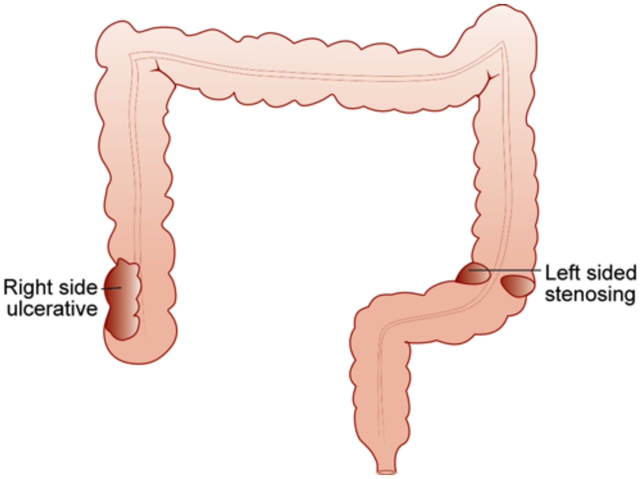

Fig. 22.35: Carcinoma colon—types.

Sessile Malignant Polyp Invasion

Sm 1: Submucosal invasion into upper 1/3rd (superficial/ inner).
Sm 2: Submucosal invasion into middle 1/3rd (inner 2/3rd).
Sm 3: Submucosal invasion lower 1/3rd (deep).

Staging of carcinoma colon

Duke's

A. Confined to bowel wall, mucosa and submucosa
B. Extends across the bowel wall to the muscularis propria with no lymph nodes involved
C. Lymph nodes are involved

Modified Duke's

A. Growth limited to rectal wall
B. Growth extending into extrarectal tissues but no lymph node spread
   B1: Invading muscularis mucosa
   B2: Invading into or through the serosa
C. Lymph node secondaries
D. Distant spread to liver, lungs, bone, brain

Astler-Coller's grading of colorectal/rectal cancer

A. Intramucosal
B1 Involvement up to muscularis propria
B2 Spread through the wall in to peritoneum
C1 B1 + involvement of lymph nodes
C2 B2 + involvement of lymph nodes
D Distant spread

Let us not look back in anger or forward in fear but look around for awareness.

Spread

Direct spread:

- Locally it can invade the bladder, obstruct ureter and so cause hydronephrosis.
- Can perforate and cause peritonitis/pericolic abscess/faecal fistula.
- Growth may get adherent to psoas muscle posteriorly.
- Carcinoma sigmoid colon can infiltrate and cause colovesical or colovaginal fistula. It can infiltrate ureter, ovary, uterus etc. It can cause pericolic abscess or abscess in lateral abdominal wall.

Lymphatic spread:

- Growth through lymphatics spreads to pericolic, epicolic, intermediate and principal group of lymph nodes.

Groups of lymph nodes draining colon

- N1: Nodes immediately adjacent to bowel wall.
- N2: Nodes along ileocolic/right colic/middle colic/ left colic/ sigmoid arteries.
- N3: Nodes near the origin of SMA and IMA.
- Nodal spread in carcinoma colon is sequential from N1 → N2 → N3.

Blood spread:

- 40% of carcinoma colon spreads to *liver via portal veins.*
- *Secondaries* may be either *solitary* or *multiple*, present as liver with hard, umbilicated nodules.
- Rarely it spreads to *bone, lung, skin.*

Cunderson-Sosin staging

A – Lesion limited to mucosa
B_1 – Through mucosa, still within bowel wall
B_2 – Through entire bowel wall
B_3 – Adherent to or invading adjacent organs
C_1 – Limited to bowel wall but node +ve
C_2 – Through entire bowel wall, node +ve
C_3 – Adherent, invasion to adjacent organs with node +ve
D – Distant spread / locally unresectable tumour

TNM staging of colorectal cancer

Tumour—T

T_x – Primary tumour cannot be assessed
T_0 – No evidence of tumour
Tis – Carcinoma *in situ*—intraepithelial/invasion into lamina propria
T_1 – Invasion into submucosa
T_2 – Invasion into muscularis propria
T_3 – Invasion into pericolorectal tissues/fat
T_{4a} – Invasion into surface of the visceral peritoneum
T_{4b} – Direct Invasion or adherent to adjacent structures/organs

Regional nodes—N

N_x – Nodes cannot be assessed
N_0 – No nodal spread
N_1 – Regional nodes 1-3 involved
  - N_{1a} – 1 regional node
  - N_{1b} – 2 to 3 regional nodes
  - N_{1c} – Tumour deposits in serosa/mesentery/nonperitonealised pericolic or perirectal tissues without regional nodes

N_2 – Regional nodes 4 or more involved
  - N_{2a} – 4-6 regional nodes
  - N_{2b} – 7 or more regional nodes

Distant metastases—M

M_0 – No distant spread
M_1 – Distant spread present
  - M_{1a} – Spread confined to one organ or site—liver/lung/ovary/nonregional nodes;
  - M_{1b} – Spread to more than one organ or site/peritoneum.

Histological grade—G

G_x – Grade cannot be assessed
G_1 – Well differentiated
G_2 – Moderately differentiated
G_3 – Poorly differentiated
G_4 – Undifferentiated

Residual tumour—R

R_0 – No residual tumour after resection
R_1 – Microscopic residual tumour after resection
R_2 – Macroscopic residual tumour after resection

Stage Group

0 – Tis N_0 M_0
I – T_1 N_0 M_0; T_2 N_0 M_0
IIA – T_3 N_0 M_0
IIB – T_{4a} N_0 M_0
IIC – T_{4b} N_0 M_0
IIIA – $T_{1\text{-}2}$, $N_{1\text{-}1c}$, M_0; T_1 N_{2a} M_0
IIIB – $T_{3\text{-}4a}$ $N_{1\text{-}1c}$ M_0; $T_{2\text{-}3}$ N_{2a} M_0; $T_{1\text{-}2}$ N_{2b} M_0
IIIC – T_{4a} N_{2a} M_0; $T_{3\text{-}4a}$ N_{2b} M_0; T_{4b} $N_{1\text{-}2}$ M_0
IVA – Any T Any N M_{1a}
IVB – Any T Any N M_{1b}
V_0 – No venous invasion; V_1—Presence of venous invasion
L_0 – No lymphatic vessel invasion; L_1—Presence of lymphatic vessel invasion

Clinical Features

- Occurs usually after 50 years. Familial type can present in younger age group. Common in males (M : F :: 3 : 2).
- Commonly present with loss of appetite and weight, anaemia, abdominal discomfort and mass per abdomen.
- 20% of cases present as an acute intestinal obstruction.
- 20% of colonic/colorectal cancer has stage IV disease at the time of first presentation.
- ***Right sided growth*** commonly presents with *anaemia, palpable mass* in the right iliac fossa, which is not moving with respiration, mobile, nontender, hard, well-localised with impaired resonant note.

Differential diagnosis for mass in the right iliac fossa

- Ileocaecal tuberculosis
- Appendicular mass
- Actinomycosis
- Ectopic kidney
- Mesenteric lymph nodes
- Ovarian tumour in females
- Retroperitoneal tumour
- Amoeboma

Carcinoma caecum occasionally presents like acute appendicitis or intussusception with intestinal obstruction.

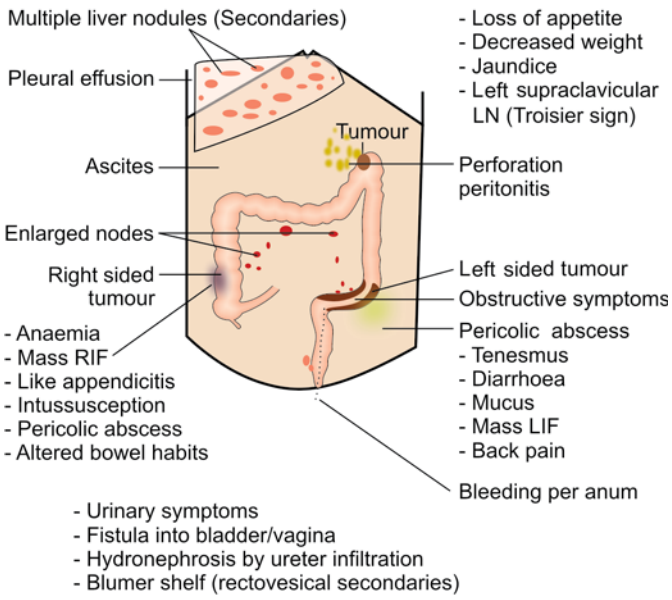
Multiple liver nodules (Secondaries)
Pleural effusion
Ascites
Enlarged nodes
Right sided tumour
- Anaemia
- Mass RIF
- Like appendicitis
- Intussusception
- Pericolic abscess
- Altered bowel habits
Tumour
- Loss of appetite
- Decreased weight
- Jaundice
- Left supraclavicular LN (Troisier sign)
Perforation peritonitis
Left sided tumour
Obstructive symptoms
Pericolic abscess
- Tenesmus
- Diarrhoea
- Mucus
- Mass LIF
- Back pain
Bleeding per anum
- Urinary symptoms
- Fistula into bladder/vagina
- Hydronephrosis by ureter infiltration
- Blumer shelf (rectovesical secondaries)

Fig. 22.36: Clinical features of carcinoma of colon—diagrammatic representation.

- ***Left sided growth*** presents with *colicky pain, altered bowel habits (alternating constipation and diarrhoea*), palpable lump, distension of abdomen due to *subacute/chronic obstruction*. Later may present like complete colonic obstruction. Tenesmus, with passage of blood and mucus, with alternate constipation and diarrhoea, is common. Bladder symptoms may warn colovesical fistula.
- Features of pericolic abscess/obstruction (15%)/perforation/peritonitis may be the first presentation.
- *Closed loop obstruction* can occur in transverse colon growth (stricture type causing block) with competent ileocaecal valve. Enormously dilated right sided colon is prone for *stercoral ulcer,* perforation and faecal peritonitis.
- Enlarged liver with multiple umbilicated hard secondaries, ascites, rectovesical secondaries, palpable left supraclavicular lymph nodes are other presentations.
- Faecal strength of *Streptococcus bovis* bacteria increases many fold in patients with colonic cancer compared to individuals without colonic cancer.

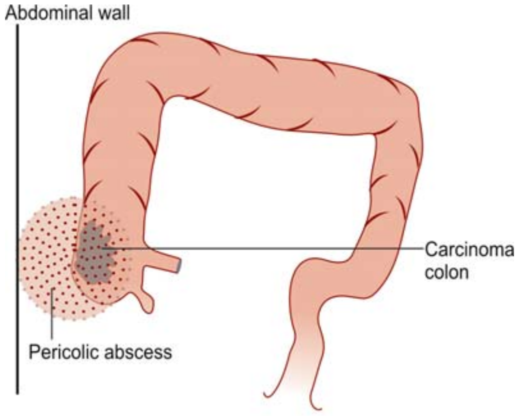
Abdominal wall
Carcinoma colon
Pericolic abscess

Fig. 22.37: Carcinoma with pericolic abscess right sided. Mass here will be tender and localised. Patient may be having fever, tachycardia. Associated obstruction may be there occasionally.

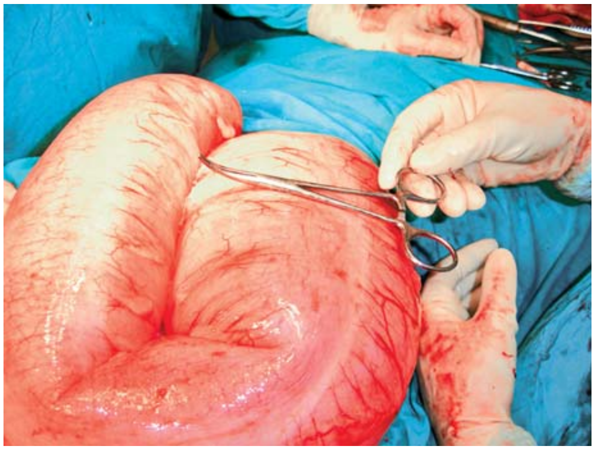

Fig. 22.38: Dilated colon in colonic growth with obstruction.

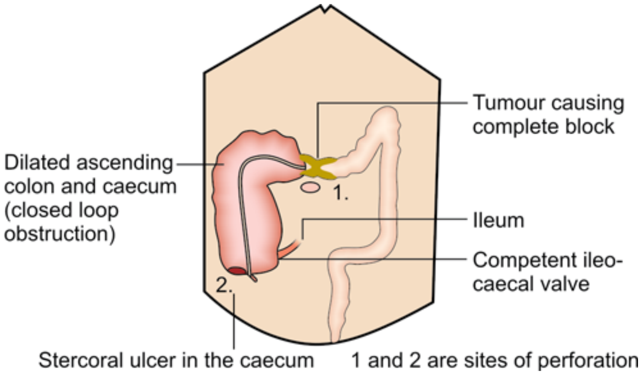
Tumour causing complete block
Dilated ascending colon and caecum (closed loop obstruction)
1.
Ileum
Competent ileo-caecal valve
2.
Stercoral ulcer in the caecum
1 and 2 are sites of perforation

Fig. 22.39: Carcinoma transverse colon or near hepatic flexure can cause complete block, leading into the enormous dilatation of caecum. If ileocaecal valve is competent, then there will be closed loop obstruction of right-sided colon causing *stercoral ulcer* due to pressure ischaemia over the summit of the caecum. Perforation can occur in stercoral ulcer or in growth site itself. It is treated by emergency surgery with resection and temporary ileostomy.

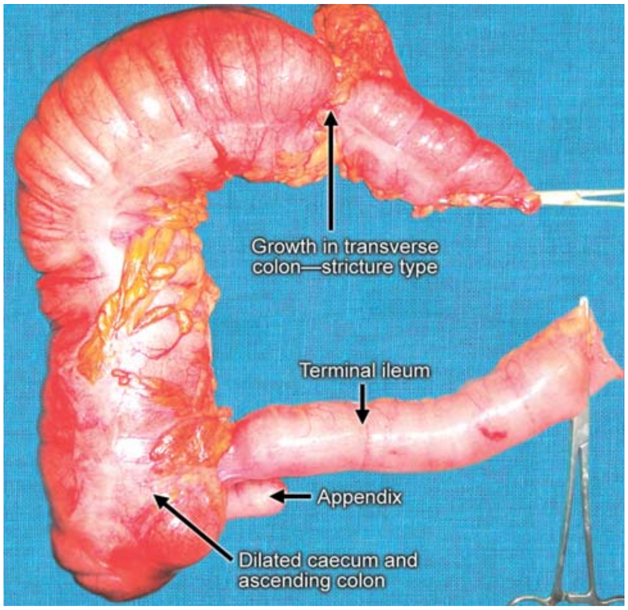
Growth in transverse colon—stricture type
Terminal ileum
Appendix
Dilated caecum and ascending colon

Fig. 22.40: *Closed loop obstruction* specimen showing in a operated patient with growth in mid-transverse colon. Extended right hemicolectomy was done.

Little discipline multiplies rewards.

Local complications of carcinoma colon

- Intestinal obstruction
- Closed loop obstruction
- Perforation and peritonitis
- Vesicocolic fistula
- Invasion of ureter
- Pericolic abscesss

Investigations

Screening and surveillance for colon cancer

- *Faecal occult blood test (FOBT)*—it is nonspecific test for peroxidase contained in haemoglobin. It is simple but with low specificity
- *Flexible sigmoidoscopy*—once in 5 years to identify the adenoma; it is often combined with FOBT
- *Colonoscopy* is the most accurate and most complete method for evaluating the entire colon. It allows identification of small polyps (< 1 cm), allows biopsy, polypectomy, control of bleeding, stricture dilatation if needed. Problem as a screening method is—prior need for mechanical bowel preparation
- *Air contrast barium enema (ACBE)* detects polyps greater than 1 cm. Its accuracy is more in proximal colon than in sigmoid colon as one may misinterpret a polyp for diverticulosis
- *CT colonography (Virtual colonoscopy)*—it is helical CT 3 dimensional intraluminal colon imaging. It needs bowel preparation, air insufflation, CT imaging

- *Barium enema*: Shows irregular filling defect and *'apple core' lesion* (in left sided carcinoma). It also helps in finding colonic polyps (*Air-contrast barium enema*).
- *Colonoscopy and biopsy* confirms the diagnosis.
- Virtual colonoscopy (CT colonography) is also useful to visualize entire colon.

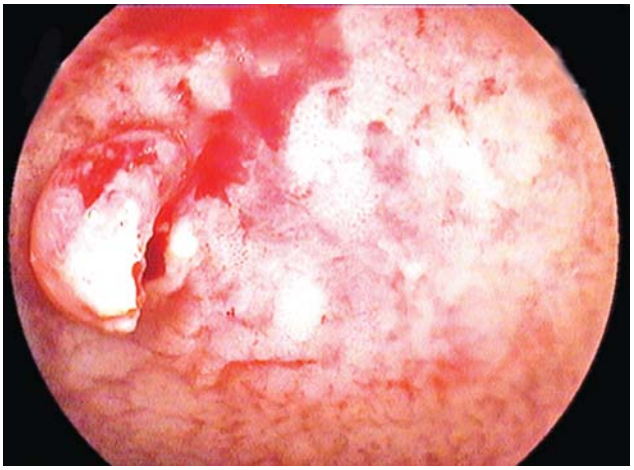

Fig. 22.41: Carcinoma colon—colonoscopic view.

- U/S, to see secondaries in liver, peritoneum, lymph node status, rectovesical secondaries.
- **CEA (*Carcinoembryonic antigen*):**
  - It is a *cell surface glycoprotein* discovered by Gold and Freedman.
  - It is normally produced by colonic epithelium.
  - Its serum ½ life is up to 10 days and is cleared by liver through Kupffer cells. So its ½ life prolongs in cholestasis and hepatocellular dysfunction.
  - Normal level is < 2.5 ng/ml. Level > 5 ng/ml is significant.
  - Even though it is a widely used tumour marker, it has got low sensitivity.

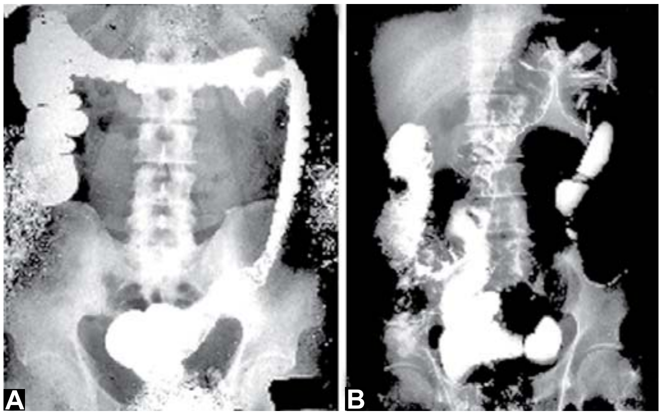

Figs 22.42A and B: Carcinoma colon. (A) Barium enema study, (B) Air-contrast study.

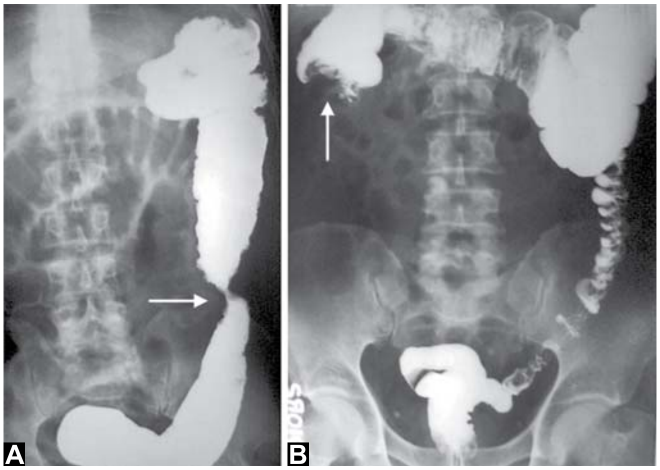

Figs 22.43A and B: X-rays with barium enema showing narrowing in sigmoid region (A) and irregularity (B).

  - CEA is primarily associated with colorectal cancers, but it can also increase significantly in pancreatic, gastric, lung, breast carcinomas. Often its level also increases in nonmalignant conditions like pancreatitis, hepatitis, obstructive jaundice, BPH.
  - *Uses in colorectal cancers are*:
    a. Preoperative levels >7.5 ng/ml signifies poor prognosis.
    b. If postoperative level does not fall, it indicates either incomplete resection, or occult metastasis elsewhere.
    c. Increase CEA during follow-up indicates recurrence or secondaries.

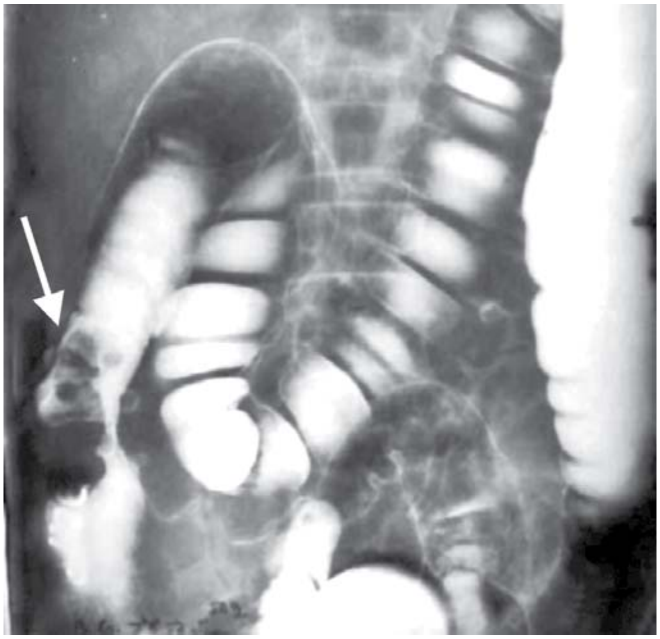

Fig. 22.44: Barium enema X-ray showing irregular filling defect in caecum—feature of carcinoma colon.

  - A slow rise indicates loco regional disease.
  - A rapid rise signifies metastasis.
  - It is not useful in assessing follow-up in poorly differentiated adenocarcinoma as such tumour will not produce CEA.
- Left supraclavicular lymph node if palpable, its FNAC may clinch the histological diagnosis.
- Hb%, PCV, haematocrit, ESR. Look for *occult blood in stool* is the initial test for anaemia.
- CT scan abdomen and pelvis—to see local spread, invasion, size and extent, stage, nodal status and liver secondaries.
- LFT—mainly enzyme studies like alkaline phosphatase, SGPT.

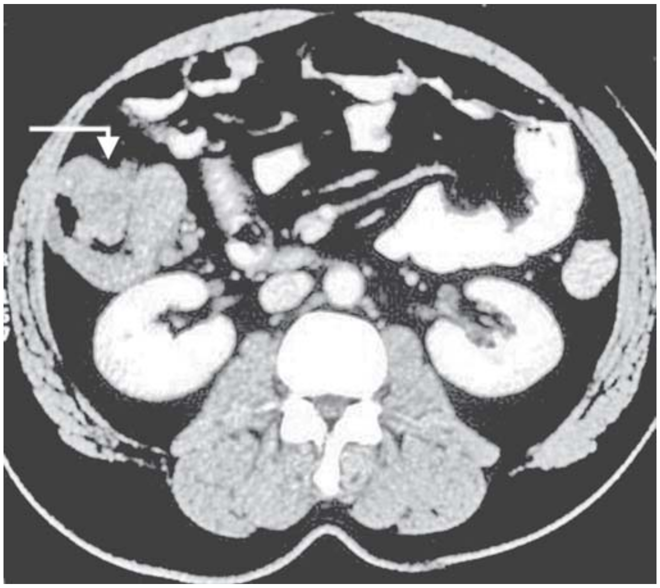

Fig. 22.45: CT scan picture showing growth in the right side colon.

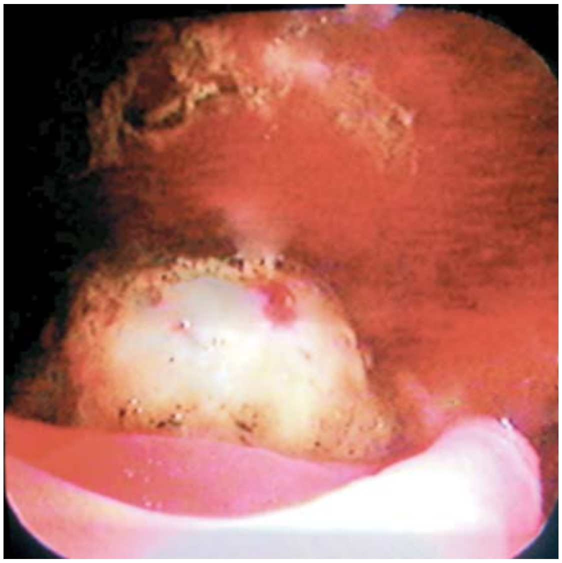

Fig. 22.46: Colonoscopic view of carcinoma colon.

Treatment

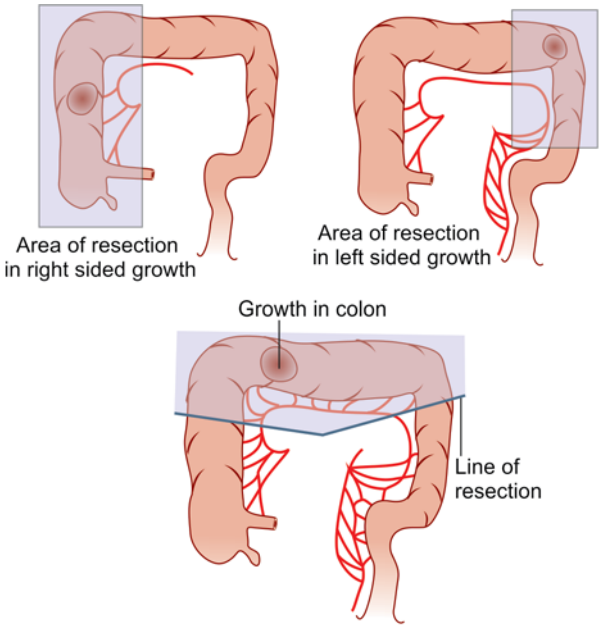
Area of resection in right sided growth
Area of resection in left sided growth
Growth in colon
Line of resection

Fig. 22.47: Levels of resection in growths in different portions of the colon.

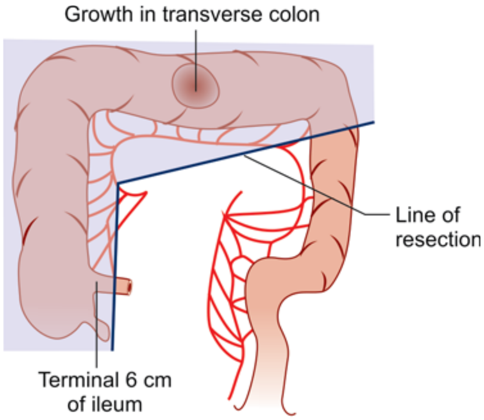
Growth in transverse colon
Line of resection
Terminal 6 cm of ileum

Fig. 22.48: Diagram showing extended right hemicolectomy.

Art is long, life short, judgement difficult, opportunity transient.

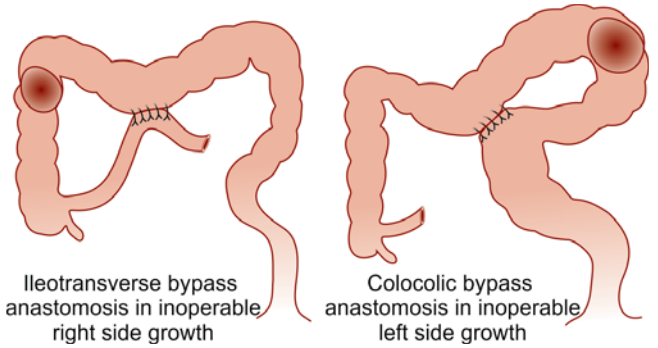
Ileotransverse bypass anastomosis in inoperable right side growth
Colocolic bypass anastomosis in inoperable left side growth

Fig. 22.49: Ileotransverse anastomosis in case of inoperable hepatic flexure growth. Colosigmoid bypass in case of inoperable growth in splenic flexure. .

Mainly Surgical

- *Right sided early growth*:
  - *Right radical hemicolectomy* with ileo transverse anastomosis is done. Structures removed are terminal 6 cm of ileum, caecum and appendix, ascending colon, 1/3 of transverse colon, lymph nodes (epicolic, paracolic, intermediate).

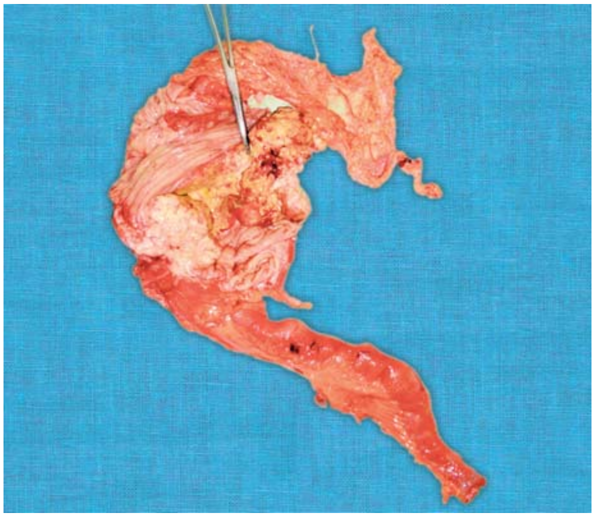

Fig. 22.50: Specimen of hemicolectomy done for carcinoma caecum. Note the different parts—terminal ileum, ascending colon, hepatic flexure and omentum.

  - In inoperable right sided growth, *ileotransverse anastomosis* is done as a by-pass procedure.
- *Transverse colon growth*
  - *An extended right hemicolectomy* is the procedure done for transverse colon growth which includes division of right colic, middle colic arteries at their origin, with removal of terminal 6 cm ileum, ascending and transverse colon; anastomosing terminal ileum and proximal part of the descending colon—*ileocolic*.
  - Alternatively, in mid-transverse colon growth, transverse colon with both flexures can be removed; anastomosing cut ends of ascending and descending colon—*colocolic*.
- *Left sided early growth*:
  - Left radical hemicolectomy is done, where in left ½ of transverse colon and descending colon is removed along with lymph nodes.
  - *Left sided stenosing type of growth* can present with acute intestinal obstruction, in which case initially colostomy is done. Later, after 3-6 weeks, following proper preparation, required formal procedure is done, followed by colostomy closure after 8 weeks (*3 stage operation*).
  - Often growth in the transverse or left sided colon, which is stenosing or obstructive type, can cause *closed loop obstruction* because of the competent ileocaecal valve. As a result, pressure increases in the caecum eventually leading to its perforation. Perforation can also occur occasionally at the site of tumour. They need *emergency intervention*—caecostomy or colostomy, or ileostomy with resection and anastomosis. If severe peritonitis sets in, it may be life-threatening.
  - Left sided colon growth *with obstruction* can be treated with resection of tumour, saline lavage and cleaning the rest of the colon by passing a catheter through the appendix into the caecum followed by primary anastomosis thus avoiding colostomy.
  - In severely *obstructed sigmoid growths* are often alternatively treated as tumour resection with removal of entire proximal colon and continuity is maintained with ileo sigmoid/ileorectal anastomosis. It also avoids diversion.
  - During exploration, presence of any synchronous growths has to be looked for. Now it is found that carcinoma colon can be *associated with tuberculosis or lymphoma of colon*. So this association is also looked for.
  - In acute left sided growths, *trans anal self-expanding metal stents* can be used instead of defunctioning colostomy.
  - *Turnball's 'No touch technique':* Here the vessels are ligated at its origin, at the beginning, in order to prevent the tumour spread—due to tumour handling.
- *Multiple synchronous primaries* in the colon.
  - *Total abdominal colectomy* with ileorectal anastomosis is done if there are multiple primary tumours or in HNPCC. Significant chronic diarrhoea is the problem due to defective water absorption.
- Surgical treatment of *liver secondaries.*
  - *In solitary liver secondary*, segmental hepatic resection is useful.
  - In case of multiple secondaries confined to one lobe of the liver, hemihepatectomy can be tried.
  - *Metastasectomy*—it is done in one secondary/one lobe secondaries/less than 3 metastases in both lobes/without any extrahepatic spread.
- '*A second look operation'* is most often helpful in carcinoma of colon, to resect the residual or recurrent tumours (*Owen Wangensteen's Second look surgery).*
- If there are synchronous growths or growth with other area having multiple potential polyps, then total colectomy with colorectal anastomosis is done.
- Recurrent tumour should be treated by re-exploration and resection with adjuvant chemotherapy of different regime.
- Laparoscopic evaluation and resection is becoming popular. Laparoscopic assisted colonic resection is also done [*Hand assisted laparoscopic surgery (HALS)*].

Note:

- *Caecum and ascending colon*—right radical hemicolectomy (terminal ileum to mid transverse colon); *hepatic flexure and transverse colon*—extended right radical hemicolectomy (terminal ileum to proximal descending colon); *splenic flexure*-extended left radical hemicolectomy (hepatic flexure to rectosigmoid); *descending colon and sigmoid*—left radical hemicolectomy (splenic flexure to rectosigmoid).
- Minimum *5 cm* margin clearance at colonic side/sides is needed; *6 cm terminal ileum* clearance is sufficient in case of right sided growth.
- Duodenum, ureter, IVC, superior mesenteric vein should be taken care of during dissection in right hemicolectomy; spleen, stomach, tail of the pancreas should be taken care of in left hemicolectomy.
- It is *ideal* to mobilize both flexures adequately in all colonic surgeries to reduce the anastomotic tension.
- Bleeding *at anastomotic site edges* should be adequate; in emergency obstructive situation, better to do temporary covering colostomy or ileostomy to prevent anastomotic leak.
- On table and postoperative *hypoxia and hypotension* should be prevented to minimize anastomotic leak.
- *Preoperative mechanical bowel preparation* is important in elective colonic resection using polyethylene glycol / mannitol, oral bowel antiseptics like neomycin and metronidaziole.
- Infiltration into adjacent local structures is *not a contraindication* for surgical resection.
- At least *12 nodes should be examined* during histology after resection.
- *Blood transfusion* before or during surgery for colonic cancer may alter the immunological aspect of the colonic cancer and may increase the recurrence rate and so may alter the prognosis. It is presently not very well-proved. *Reason thought is*—blood transfusion raises the suppressor T cells leading into immunosuppression. So autologous blood transfusion is better when in case of need.
- Proper preoperative bowel preparation by mannitol/ polyethylene glycol/hypertonic saline/bowel wash/ bowel antiseptics (neomycin or gentamycin or streptomycin and metronidazole orally).
- Antibiotic prophylaxis; DVT prophylaxis by crepe bandage or stockings bandage, heparin or low molecular weight heparin.
- Pulmonary function tests and pulmonary exercise preoperatively as well as postoperatively.
- Urinary catheterisation, nasogastric tube placement.
- Postoperative proper care with fluid, electrolyte, drain care, early mobilisation, monitoring.
- Complications of surgery are leak (fecal fistula), bleeding, infection like intraperitoneal/pelvic abscess, respiratory problems, DVT, wound infection, burst abdomen.

Adjuvant Therapy

Chemotherapy

Indications for chemotherapy

- Positive nodes.
- T4 lesions.
- Venous (microscopic) spread.
- Signet cell type.
- Poorly differentiated tumour/aneuploidy.
- Changes in CEA level.

Postoperative chemotherapy is used commonly. Occasionally also given preoperatively.

Regimes are:

- 5 fluorouracil (5 FU) with folinic acid (leucoverin/ LV) is the most commonly used regime for 6 months as monthly cycles. Folinic acid potentiates the action of 5 FU.
- Levamisole 150 mg/day for 3 days given once in 15 days for one year with intravenous 5 FU monthly for one year.
- Irinotecan/5 FU/LV—IFL regime is also used.
- Folinic acid (LV)/5 FU/oxaliplatin—FOLFOX regime is also used. It is becoming treatment of choice.
- Irinotecan/oxaliplatin—IROX regime is used in previously untreated metastatic colonic cancer.
- Capecitabine (xeloda) an oral drug which generates 5 FU at tumour tissue, shows significantly greater response than 5 FU/LV regime.
- Phase II trials are going on for capecitabine/oxaliplatin and capecitabine/irinotecan combination regimes.
- 5 FU infusions into the portal vein during and immediately after surgery have shown benefits in terms of outcome and recurrence.

EGFR and VEGF blockers (EGFR is epithelial growth factor receptor; VEGF is vascular endothelial growth factor):

- They are used as single agent and also in combination with chemotherapy drugs in phase II and III trial.
- Drugs are monoclonal antibody, cetuximab which blocks EGFR, bevacizumab which binds VEGF.

- Dose of 5 FU is 400 mg/m^2/day IV × 5 days once in 3-4 weeks for 6-12 months
- Dose of leucoverin is 20 mg/m^2/day × 5 days once in 3-4 weeks for 6 months
- Dose of levamisole is 150 mg/day × 3 days once in 2 weeks for one year
- Capecitabine is given orally 1250 mg/m^2 twice daily

Radiotherapy

- Usually there is no role for RT as tumour is radioresistant.
- It is often used in locally advanced tumour, infiltrating the psoas major muscle or lateral abdominal wall, left sided colonic growth.
- It is also used in inoperable recurrent tumour.

Follow-up of Carcinoma Colon

- For 3 years at regular intervals, once in 3-6 months.

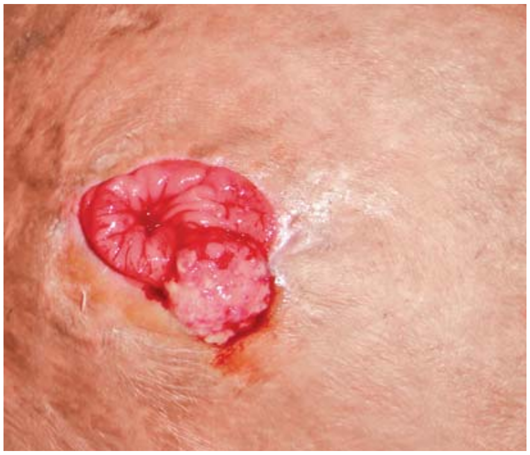

Fig. 22.51: Patient has undergone left hemicolectomy earlier for sigmoid growth with a colostomy; now presenting as recurrence at colostomy site. It could be recurrence of original disease or could be new metachronous disease in the mucosa of the colostomy site. CT scan/CEA/PET scan will help to differentiate. Re-excision can be done if it is metachronous local growth.

This is by:

- Regular CEA analysis
- Ultrasound abdomen
- Barium enema X-ray
- Colonoscopy
  Rise in CEA is a definite indicator of recurrence or secondaries
  In patients with raised CEA, radioisotope antibody study will show the site of recurrence or secondaries
- Serum alkaline phosphatase

Note:

- ***Local recurrence***: Incidence of local recurrence after radical resection is 15%. 85% of recurrence is detected in 2 years. It is more in colocolic and colorectal anastomosis than ileocolic. *Recurrence may be*—local/locoregional/regional/metastatic. *Recurrence may be*—true recurrence at anastomotic site, probably mucosal; adjacent and close to anastomotic site, at the original primary tumour bed; peritoneal recurrence; recurrence in adjacent organ. It is due to implantation of spilled tumour cells during easier surgery, incomplete resection, metachronous new lesion, retrograde lymphatic spread. It is identified by CEA, CT scan, radioisotope/radioimmuno assay, colonoscopy. It is treated by curative surgery with clearance, chemotherapy, RT.
- Colonoscopy after 1 year to see anastomotic site, new / missed polyp/metachronous growth.
- Later colonoscopy is done once in 3-5 years unless there is a family history (here yearly colonoscopy is needed).
- CEA estimation once in 3 months for 2 years. If there is a raise any time, further screening is done for recurrent/metastatic disease by CT/MRI/PET scan.

Prognosis

Depends on:

- *Site*—left sided tumours has got better prognosis as they present early.
- *Type*—colloid carcinoma has got poorer prognosis.
- *Size* of the tumour.
- *Lymph nodes status: Number of lymph nodes* involved decides the prognosis.
- *Liver secondaries has poor prognosis.*
- *Age of the patient.*
- Associated diseases like *HIV.*
- *Stage of* the tumour.
- Presence of complications, perforation, peritonitis.

On the whole, it is a curable malignancy with proper surgery and adjuvant therapy.

5 year survival is:

Stage I – 90%.
Stage II – 75%.
Stage III – 50%.
Stage IV – less than 5%.

ANGIODYSPLASIA OF COLON

- It is a vascular ectasia seen *commonly in right side* colon.
- It is uncommon in healthy individual.
- It is commonly seen in elderly of 70 years age. It is acquired malformation of aging.
- It is not common in left side colon. It is often associated with calcified aortic stenosis; and ectasia bleeding if present, stops once aortic stenosis is corrected.
- Common cause of rectal bleed in adults and elderly.
- *An acquired condition seen in caecum* and ascending colon due to degeneration of the mucosal and submucosal vessels of the colon.
- Diagnosis is by mesenteric angiography. *Angiographic criteria in angiodysplasia*—early and prolonged filling of draining veins; cluster of small arteries; visualisation of vascular tuft.
- Colonoscopic fulguration or resection is the treatment.
- Therapeutic embolisation is useful in 85% of cases.

OGILVIE'S SYNDROME

- *Described by* Sir William Heneage Ogilvie in 1948.
- Also called as *colonic pseudo-obstruction.*
- Severe colonic ileus occurs without any mechanical obstruction.
- *Tympanic, nontender, distended* abdomen and bowel sounds will be normally present—are the clinical features. Features of causative etiology may be present.
- Occurs in critically ill patients.
- Due to sacral parasympathetic nerve dysfunction.
- Commonly seen in *right and transverse colon.*
- Due to sacral parasympathetic malfunctioning, there is atony of descending colon causing functional obstruction. Splenic flexure will be the junction of dilated and collapsed parts of the colon wherein parasympathetic supply of vagus ends and of sacral parasympathetic begins. Increased sympathetic activity causes colonic dilatation.
- It is seen in:
  - Scleroderma, SLE, dermatomyositis.
  - Chaga's disease.
  - Myotonic dystrophy, multiple sclerosis.
  - Neuropathies.
  - Myopathies.
  - Hypothyroidism, diabetes mellitus.
  - CRF, renal transplantation.
  - Poisoning, sepsis, hypoxia.
  - Radiotherapy, orthopedic procedures.
  - Psychiatric disorders, drug abuse.
  - Retroperitoneal irritation by blood, urine, pancreatic enzymes.
  - Lumbar spine and pelvic trauma, shock, stroke.
  - Septicaemia, burns, MI.
  - Can be *idiopathic* also.
- *Investigations:*
  - Plain X-ray abdomen shows dilatation of colon.
  - Barium enema will be normal.
  - Careful gentle colonoscopy is useful as therapeutic also.
  - CT abdomen to rule out mechanical obstruction of large bowel.

Treatment

- Mainly conservative.
- Motility enhancing drugs like neostigmine and erythromycin are used. *Neostigmine* is given 2.5 mg IV for 3 minutes, showing response in 10 minutes in 90% of patients. But bradycardia should be watched for and atropine should be kept ready.

- Prokinetic drugs like cisapride/mosapride are used.
- Ceruletide decapeptide 0.3 μg/kg IM is used to stimulate intestinal motility.
- Ryle's tube aspiration.
- Flatus tube insertion. It is not very useful as disease is common in proximal colon.
- *Epidural anaesthesia* causes sympathetic blockade relieving pseudo-obstruction.
- *Colonoscopic decompression* is also useful.
- Correction of electrolyte imbalance.
- Sodium enema.

Caecal diameter more than 12 cm (critical) on X-ray indicates surgical exploration.

- Tube caecostomy; resection and exteriorisation.

COLOSTOMY

It is an artificial opening made in the colon to the exterior (skin) to divert faeces and flatus.

Types

1. ***Temporary:*** Is done in conditions wherein diversion is required to facilitate healing distally in the rectum or distal colon. And this is closed once the purpose is over.

   Site of temporary colostomy is usually right hypochondrium and left iliac fossa.

   It can be *loop colostomy* or *Devine's double-barrel colostomy* (wherein there is a gap between the two openings of colostomy which prevents spillage into the distal loop).
2. ***Permanent*** colostomy is always *end colostomy* placed in left iliac fossa, 6 cm above and medial to the anterior superior iliac spine.

Indications	
Temporary	**Permanent**
Congenital megacolon	AP-Resection.
Anorectal malformations	Carcinoma anal canal.
Sigmoid volvulus	After Hartmann's operation
Perforation of left sided colon	
Left sided colonic growth	
High anal fistula	
Trauma to left sided growth	

Colostomy can be:

- Loop, end, double barrel.

It can be:

- *Diversion colostomy*: It is done when there is breach in bowel wall, trauma, destruction, sphincter injuries, Crohn's, carcinoma rectosigmoid. It is usually brought out as an end colostomy with a mucus fistula or with a Hartmann's resection. It can be at sigmoid, descending and transverse colon.

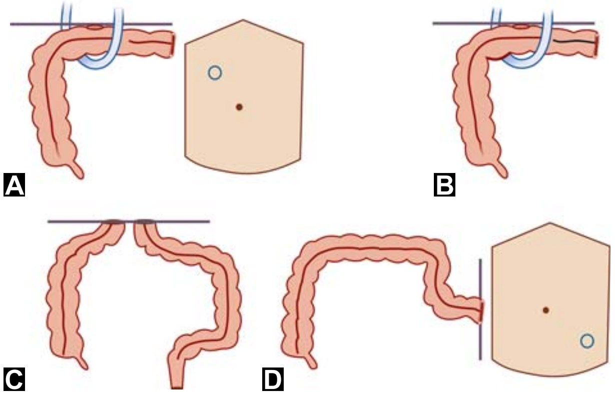

Figs 22.52A to D: Types of colostomy. (A) Temporary colostomy, (B) Loop colostomy, (C) Devine's double barrel colostomy, (D) Permanent colostomy.

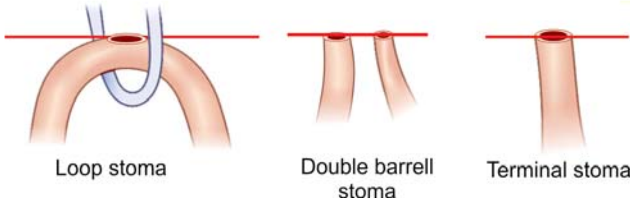

Fig. 22.53: Different types of stomas used in colostomy.

- *Decompression colostomy*: It is done for obstructive lesions in rectosigmoid, toxic megacolon. Types—(1) *Blow hole* procedure is done as a single or multiple small stomas onto the skin to decompress colon in acutely ill patients with massively dilated or impending perforation of colon. Problems are—remaining possible ischaemic parts of the colon is not identified still risking perforation; peristomal skin complications and mucosal prolapse is common. (2) *Tube caecostomy*—It is technically easier without much complications; but tube getting blocked limits its use regularly. (3) *Loop transverse colostomy*—It serves as decompressing and diverting long stoma without getting blocked. It is done only in mobile colon.
- *Irrigation colostomy*: It avoids need for appliance wear; reduces the passage of uncontrolled gas with less leak of stools. But it is time consuming with chance of water intoxication due to excess water absorption during irrigation.

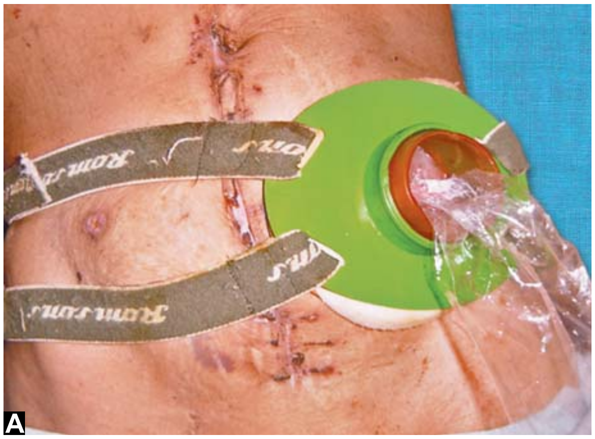

Fig. 22.54A

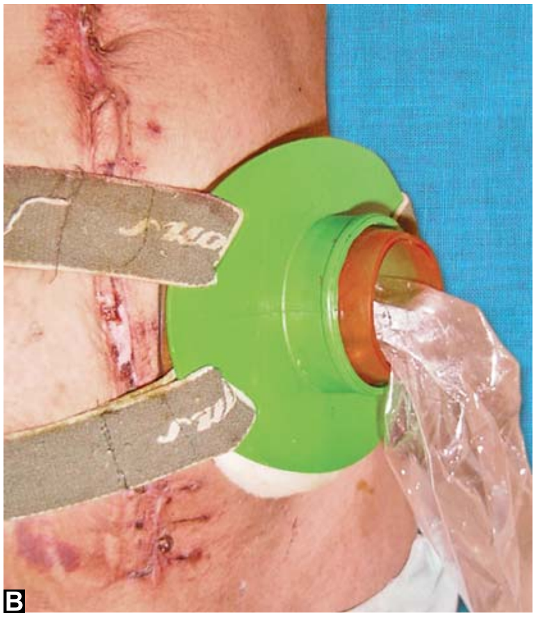

Fig. 22.54B

Figs 22.54A and B: Diagram of colostomy in a patient with colostomy bag.

Complications of colostomy

- Prolapse of mucosa (prolapse of distal loop mucosa is common)—commonest complication
- Retraction
- Necrosis
- Stenosis
- Herniation
- Bleeding
- Diarrhoea
- Enteritis
- Skin excoriation

Educating the patient regarding the proper usage of colostomy bags and proper care of the colostomy is very essential.

Colostomy Care

- Similar to stoma care.
- Initially stoma bag should be transparent as content is liquid stool but later it can be opaque.
- Regular consultation with stoma therapist.
- Care of the skin.
- Training for managing colostomy, its care to prevent leak, odour, and discomfort.

Closure of Colostomy

Criteria for temporary colostomy closure:

- Integrity of distal colon should be normal and adequate.
- Anorectal sphincter should be normal.
- Cause for construction of colostomy is cured completely without any suspicious of recurrence of same disease distally.
- When temporary colostomy is done, it is closed usually after 3 months.
- *Two types of closure are present*—extraperitoneal and intraperitoneal type. *Now intraperitoneal closure is done*. Earlier extraperitoneal closure was done in loop colostomy by placing a spur in between and closing the antimesentric part of the colon. It prevents the peritoneal contamination. But inadequate closure, leak, adhesions are the problems. *Intraperitoneal closure*—commonly advocated technique now, is done by placing a circumferential incision over the margin with skin edge. Incision is deepened to enter the peritoneum and pull out the colostomy stoma. Part adjacent to the skin is resected and anastomosed using silk/vicryl. Sutured bowel is placed into the peritoneal cavity. Drain is placed into the peritoneal cavity. Abdomen is closed in layers.
- Closure of colostomy is done after proper bowel preparation, under general or spinal anaesthesia.
- Proper postoperative care is important.
- Patient should perform anal sphincter exercises to prevent sphincter atrophy and to maintain sphincter tone.

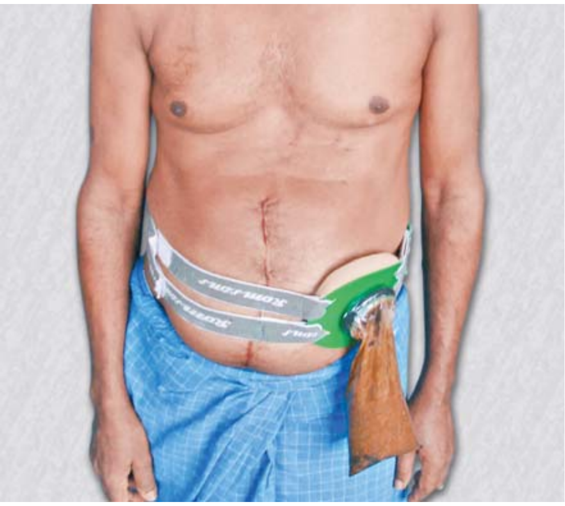

Fig. 22.55: Patient with colostomy bag. He has undergone abdominoperineal resection (APR) for carcinoma rectum.

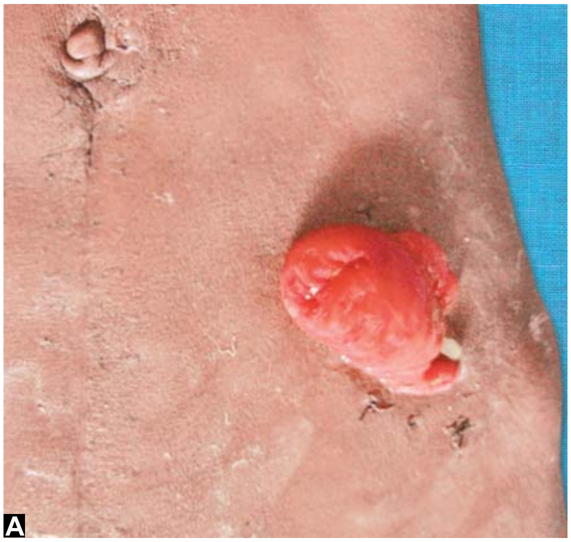

Fig. 22.56A

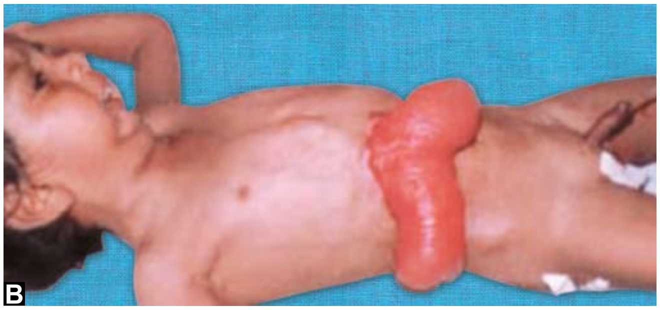

Fig. 22.56B

Figs 22.56A and B: Prolapse of colostomy is a common complication.

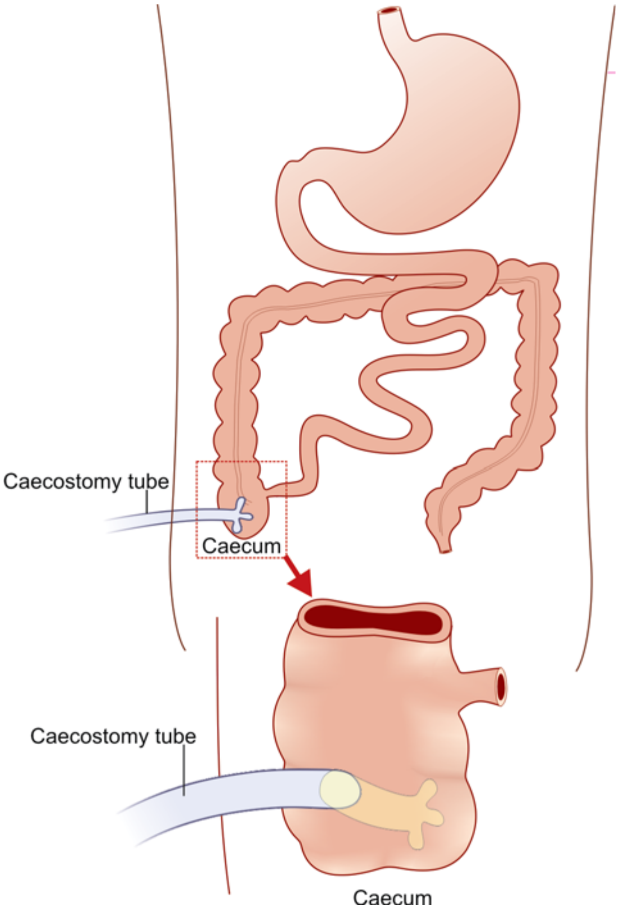

Fig. 22.57: Caecostomy is often done in acute colonic conditions as diversion procedure after surgery distally like perforation closure and resection. But it is not as popular as colostomy because it may not function adequately, may get displaced or blocked. But it is technically easier and usually done by doing appendicectomy and passing Malecot's tube through appendicular stump. Removal is easier without any surgical intervention.

STOMA CARE

Definition of Stoma

Stoma is an artificial opening or 'mouth like' to the exterior, the abdominal wall so as to drain the content from the tubular structures inside, like bowel or ureter. It is done for diversion of urine or faecal matter in case of malignancy, trauma, and sepsis or after surgery.

Types

- ***Ileostomy***—terminal 5 cm ileum is projected out, on to the skin of abdominal wall to drain semi-liquid, faecal matter.
- ***Colostomy***—colon at different levels, can be brought out to the skin as required as colostomy, to divert faecal matter.
- ***Cutaneous ureterostomy***—cut ends of one or both ureters are apposed to the skin of abdominal wall.
- ***Ileal urinary conduit***—segment of isolated ileum can be used to drain urine from the ureter as urinary ileal conduit. Ureters are anastomosed to a closed ileal conduit. Ileal stoma is brought out as stoma. Different types of continent ileostomies are in use to prevent leak, soakage and discomfort.
- ***Vesicostomy***—it is done in children. Here anterior wall of bladder is brought out and bladder mucosa is sutured to the skin of abdominal wall.

Stoma created may be *round* (commonly) or *square* in shape.

Preparation and Counselling of the Patient for Stoma

- To certain extent stoma of any type causes psychological and physical trauma to the patient, as it is nonphysiological, distressing and socially not acceptable.
- Patient should be explained about the procedure and should be convinced and consoled about the stoma.
- Detailed meaning, explanation and after care of the stoma should be discussed.
- Indication for the stoma and consent for the same should be taken.
- Reassurance about the stoma, its care, and its position should be diagrammatically explained to the patient and his close relative.
- In case of obstructive disease, stoma is done as an inevitable procedure to relieve the obstruction, often it may be temporary.
- Proper bowel preparation by bowel wash, gut irrigation is required before surgery.
- *The surgeon selects the site of the stoma.*
    - Stoma is usually sited midway between anterior superior iliac spine and umbilicus.
    - It should be away from the belt line.
    - It should be away from the scar, creases, and bony points.
    - Patient should be assessed for proper size, adequacy for stoma in lying down, sitting, and standing positions.
    - Proper stoma appliances should be decided after thorough check-up and discussion with patient and patient's relative.
    - Stoma site should be marked properly before surgery.
    - Ileostomy is usually sited in the right iliac fossa, colostomy in left iliac fossa.
    - Allergy for the particular appliances should be checked for.
    - The patient should consult stoma therapist.

Postoperative Care for the Stoma

- Stitches are removed in 6-10 days.
- Dressing should be done first over the stoma and after placement of appliance, laparotomy wound is dressed otherwise stoma appliance will not sit properly.
- Patient should be observed for any complications.
- Once wound has healed patient can take bath by removing the appliances. After bath skin is dried up and stoma appliances can be fit again.
- Patient should be taught about the stoma care and its appliances.
- Care and prevention of skin excoriation due to leak is also looked into.
- Psychotherapy is given for the patient.
- Skin should be *absolutely dry* prior to placing the stoma appliances.

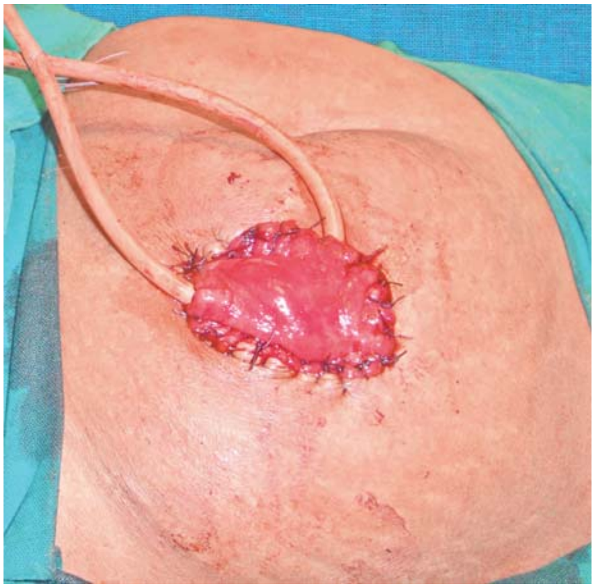

Fig. 22.58: Right sided temporary loop colostomy. Note the tube kept in between.

Complications of Stoma

- Skin excoriation.
- Mucosal prolapse—common complication.

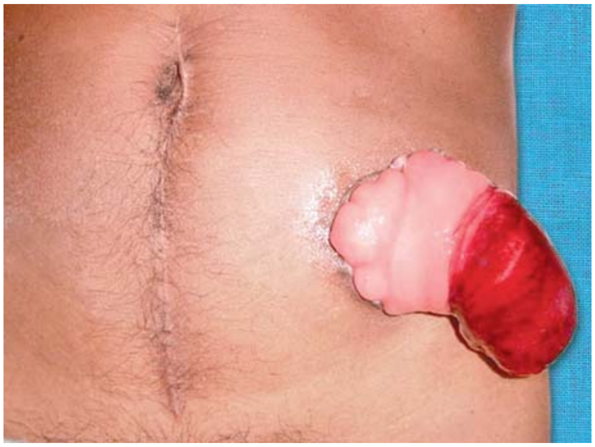

Fig. 22.59: Colostomy prolapse is the common complication.

- Stenosis and block.
- Infection either bacterial or candidial.
- Diarrhoea due to irritation.
- Leak due to improper fitting of the appliances, scar, irregularity of stoma, prolapse.
- Bleeding from the stoma edge.
- Herniation of the abdominal contents adjacent to stoma.

Skin Excoriation

It is a major problem in stoma patients. It is basically due to leak adjacent to appliances.

Causes for Excoriation

- Leak due to improper appliances.
- Wet skin before placing the appliance.
- Inadequate stoma hole.
- Improper and inadequate adhesive sheet usage.
- Allergy.
- Infection like of bacteria and candida.
- Altered weight of the patient.
- Stoma bag is overfilled or kinked or air in the stoma bag.

Treatment of Excoriation

- Control of infection by antibiotics and control of moniliasis.
- Allergy has to be confirmed, and if it is the cause, the agent is found out and treated as required.
- Zinc oxide cream application.
- Change of the type of appliance.
- Refashioning of the stoma.

STOMA APPLIANCES

Stoma appliances are devices, which are used to collect and dispose the effluent materials which come out of the stoma.

Ideal stoma appliance is*:*

- Leak proof
- Should not damage the stoma and surrounding skin
- Should prevent odour
- Should be available
- Easier to use

Types of Appliances

It can be:

- *Closed type* is discarded when full and is used in patients with well-formed stool.
- *Drainable type* is used in patients with loose liquid stool. It can be emptied and retained and reused. Immediately after colostomy, drainable appliance is used. Later, it can be changed over to closed type.

It can also be:

- *One-piece stoma* appliance as a bag with adhesive system attached which adheres to skin around the stoma.

- *Two-piece stoma* appliance has got a flange with adhesive system and a bag over it, which can be removed and replaced with a new one without disturbing the flange underneath.

Bag can be:

- *Transparent,* in which fluid can be visualised. It is used in initial period of the stoma.
- *Opaque,* in which fluid cannot be visualised. It is used eventually later.

General Care and Advice to Patients with Stoma

- Patient can have normal diet. Diet, which regulates the bowel action, is better. Plenty of water is advisable.
- Patient can go for normal work, exercise like sports, swimming, tennis. Stoma appliances suitable for these works are available.
- Antidepressants, anticholinergics might cause constipation. So these drugs should be taken carefully.
- Using irritant solutions near stoma should be avoided. It may lead to dangerous complications.
- Patient can have normal sexual activity.
- Patient should have additional stoma bags in hand so as to use if required urgently.
- Patient should be aware of different appliances available and should be well-versed with its use. He can take the help of the stoma societies.

FAECAL FISTULA

- This is a dangerous entity.
- It commonly occurs after appendicectomy (gangrenous), ileal resection, colonic surgery, malignancy, ileocaecal TB, actinomycosis and Crohn's disease.
- It may be from the ileum or the colon.
- It may be single or multiple openings.
- More proximal the fistula, more is the severity in fluid and electrolyte imbalance.
- It may be from the main incision wound or from the drain wound.
- If there is no distal obstruction, the fistula will heal spontaneously, but may take a longer time.
- *Fistulogram* often delineates the track.
- CT fistulogram with CT scan abdomen is essential.
- Often gastrograffin study is also useful to study the fistulous track in detail.

Factors preventing closure (FRIENDS)

- F : Foreign body in fistula tract
- R : Radiation enteritis
- I : Inflammatory bowel disease
- E : Epithelialisation of tract
- N : Neoplasm
- D : Distal obstruction
- S : Sepsis

Factors preventing spontaneous closure of faecal fistulas

- High output > 500 ml/24 hr
- Distal bowel obstruction
- Sepsis, inflammatory bowel disease
- Tuberculosis
- Cancer
- Radiation enteritis
- Foreign body in fistula tract
- Epithelialisation of tract

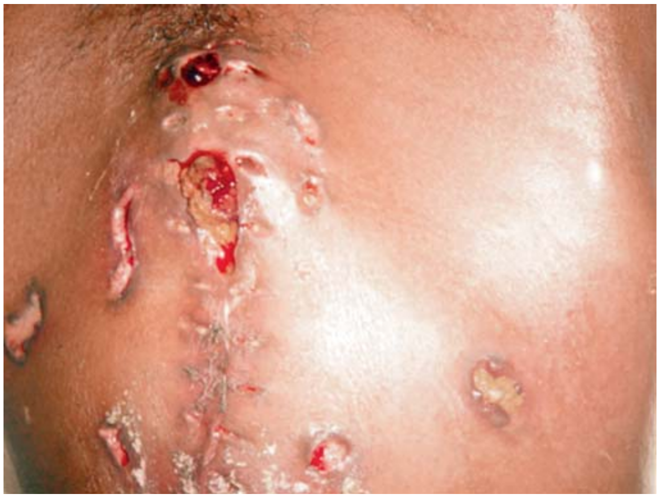

Fig. 22.60: Multiple faecal fistulous openings in the abdominal wall. Patient has undergone laparotomy for acute peritonitis.

Investigations for faecal fistula

- Fistulogram using gastrograffin
- CT scan and CT fistulogram
- Barium enema study

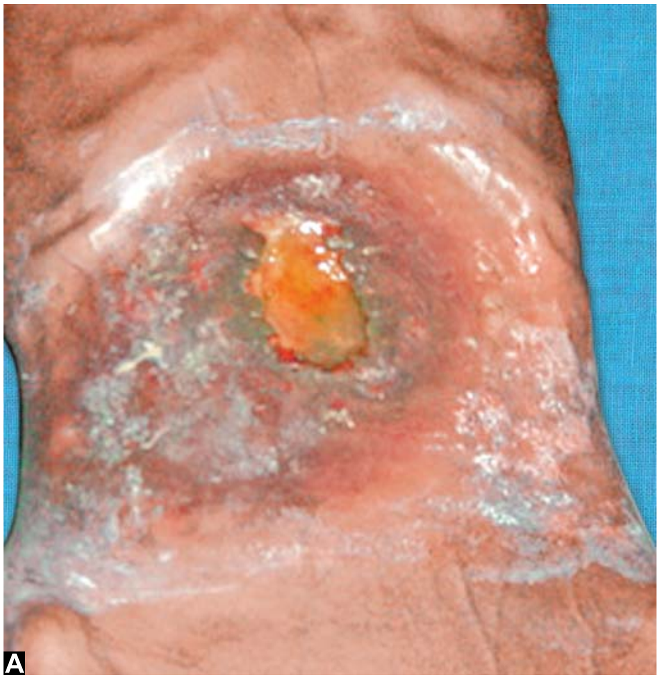

Fig. 22.61A

An essential aspect of creativity is not being afraid to fail.

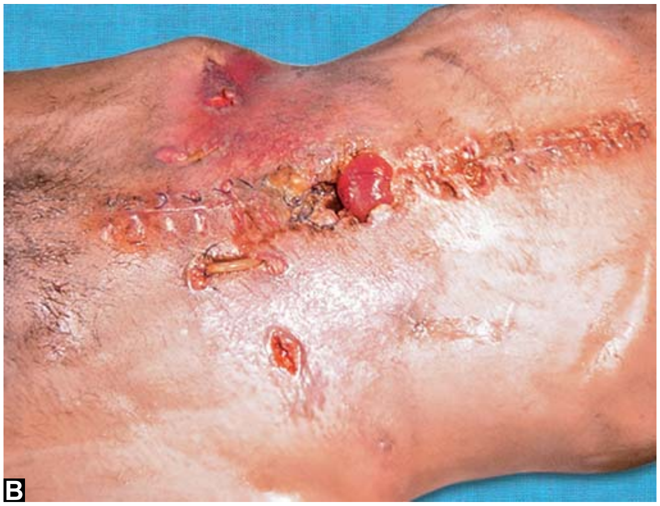

Fig. 22.61B

Figs 22.61A and B: Gastrointestinal faecal fistula after an emergency surgery for ileal perforation with severe peritonitis. Patient recovered after long-term stay in the hospital.

Treatment

- TPN and blood transfusion may be required to improve the nutritional status of the patient.
- Sepsis is controlled with proper antibiotics.
- Management of fluid and electrolyte loss.
- Skin is protected from excoriation by using zinc oxide cream.
- If it persists, later exploration and resection is done.
- The cause is treated.
- Bypass may be required. (*Most of the time, the decision is, taken on table during surgery*).

PREPARATION OF LARGE BOWEL FOR SURGERY

Principle Behind Bowel Preparation

Colon contains large amount of bacteria up to 10^9 / ml of faeces. Most common anaerobe is bacteroides; commonest aerobe is *Escherichia coli; Pseudomonas, Enterococcus, Proteus, Klebsiella, Streptococcus* are other organisms. Bowel preparation is done to clear this bacterial load to reduce postoperative complications.

Methods Used

1. **Mechanical bowel preparation**
    - *Polyethylene glycol (PEG)* is a non absorbed sodium sulphate solution, 2-3 litres of which is asked to drink by the patient along with plenty of additional fluids orally. It cleans the bowel by passing loose stool for 10-15 times in 12 hours. It acts by its hygroscopic action. *Side effects* are—nausea, vomiting, and abdominal cramps. Antiemetics are often needed. It is ideal in renal failure, ascites, cirrhosis, CCF.
    - *Sodium phosphate* is an alternative to PEG as smaller volume is sufficient to take. But it causes electrolyte imbalance. Its efficacy is similar to PEG. But patient compliance is better with sodium phosphate. It can cause hyponatraemia, hypocalaemia, hypocalcaemia and hypophosphataemia in case of renal dysfunction. In such patients it is contraindicated.
    - *Other methods*—(1) *Total gut irrigation* daily using 200 ml of *oral mannitol* or through nasogastric tube for 3 days prior to surgery. (2) *Bowel wash* daily for 3 days prior to surgery using 2 litres of normal saline (not water as it will cause water intoxication); it cleans entire large bowel. (3) Senna, castor oil, bisacodyl. (4) Repeated enemas. *PEG or sodium phosphate has taken over all this methods*.

    Contraindications for mechanical bowel preparation: Complete bowel obstruction and perforation.
2. **Antibiotics—parenteral and as bowel antiseptics**
    - *Oral neomycin* (gentamycin, streptomycin were used in olden days) 1 gram, *erythromycin* 1 gram, is used 3 days prior to surgery. Alternatively *ciprofloxacillin and metronidazole* are used. IV fluids should be given in addition to these patients to maintain adequate hydration.
    - *IV antibiotics 4 hours before* making incision, reduces the incidence of sepsis. Usually cephalosporins are given.

Controversies

- There is doubt about the advantages of preoperative mechanical bowel preparation even though it is practiced universally.
- Oral antibiotics to reduce the bacterial load in the colon is also controversial. However, preoperative single dose parenteral antibiotic improves the result and reduces the sepsis.

Indications for Large Bowel Preparation

- Carcinoma colon (especially left sided).
- Anorectal malformations.
- Megacolon.
- Carcinoma rectum.
- Surgery for ulcerative colitis.
- FAP.
- Diverticulitis.
- High pelvirectal fistulas.
- Before colonoscopy.

SURGICAL POUCHES

Pouches are created as reservoir whenever needed as a replacement for the existing reservoir like stomach, rectum, and urinary bladder.

Types

- *Jejunal pouches:* It is used for stomach. Hunt Lawrence pouch, omega loop, Roux en loop are the examples. It is commonly used after total gastrectomy. Hunt Lawrence pouch is 15 cm in length.
- *Ileal pouch*: It is used as rectal reservoir, urinary reservoir, ileostomy reservoir. Examples are—J, S, W, H pouches

Pathological pouches	Experimental physiological pouches	Natural pouches
• *Zenker's diverticulum* causing pharyngeal pouch • *Physick's pouch* in the rectal wall in between rectal valves • *Hartmann's pouch* is protruded pathological gallbladder infundibulum	• *Heidenhein pouch*: To study gastric physiology done in stomach • *Pavlov pouch:* Gastric pouch created with retaining *vagii* to study gastric physiology	• *Rathke's pouch* in pituitary • *Seessel's pouch* in pharynx • *Morrison's hepatorenal* pouch • Rectouterine pouch of *Douglas*

after proctocolectomy with ileo anal anastomosis; Koch's pouch for ileostomy; Ileal pouch for bladder replacement or cutaneous ureterostomy.

- *Colonic*: Sigmoid reservoir or caecal/ileocaecal reservoir is often used to replace urinary bladder.

Principles

- Blood supply should be preserved.
- Mesentery should extend adequately.
- Bowel is opened on the antimesenteric border and sutured using vicryl posterior and anteriorly.
- Created pouch is anastomosed to the required area from its summit or from open end.

Complications

- Obstruction (25%) is the common complication. It may be due to stomal stricture, internal hernia, volvulus and adhesions.
- Pouch disruption and leak can occur which is usually treated conservatively.
- Pelvic abscess formation in ileoanal pouch is confirmed by clinical features of fever, back pain, raised total count, CT scan. It is treated with antibiotics, drainage per abdomen or reservoir excision and ileostomy.
- Pouchitis: It is the common problem (20%). Tenesmus, bloody diarrhoea, spasm, back pain, fever, cramps, dehydration are the presentations. It is treated conservatively with antibiotics, hydration, *probiotics*.
- Vaginal fistula formation.
- Anal canal stricture formation.
- Faecal incontinence.
- In urinary conduits, sepsis, acidosis, electrolyte imbalance, stone formation, urine leak are the complications.

Note:
Pouch is contraindicated in *Crohn's disease.*

BARIUM ENEMA

- It is the contrast X-ray done to visualise large bowel.
- Therapeutic barium enema is done in intussusception in children.

Barium enema

Indications	Finding
Carcinoma colon	Irregular filling defect
Ileocaecal tuberculosis	Pulled up caecum, obtuse ileocaecal angle
	Filling defect, incompetent ileocaecal valve
Ulcerative colitis	Loss of haustrations, lead pipe appearance
Colonic polyps	Smooth, regular filling defect
Congenital mega colon	Narrow zone, zone of cone, dilated proximal segment
Diaphragmatic hernia	Colonic shadow in the left thoracic cavity
Gastrojejunocolic fistula	Leak into the stomach from colon

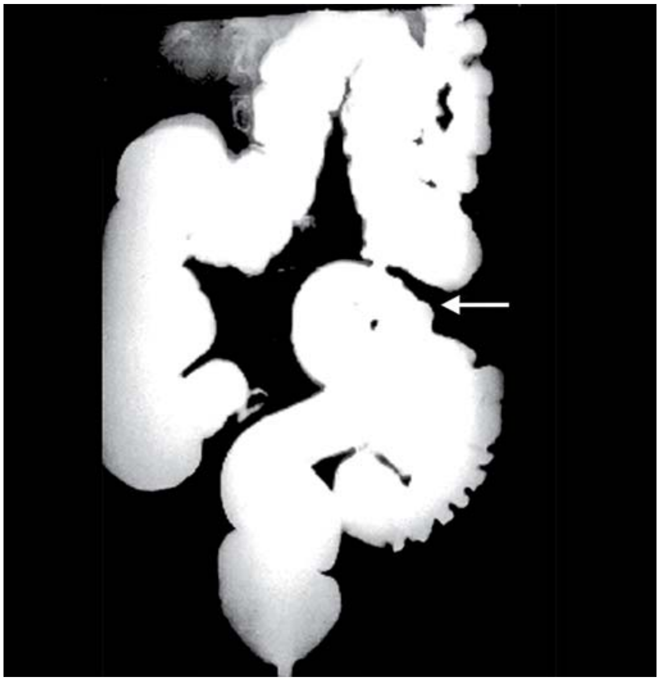

Fig. 22.62: Barium enema showing normal peristaltic movement.

Procedure

Laxative is given previous evening. Enema is given on same day before doing the procedure. About one litre of barium sulphate solution is infused into the colorectum per anally through enema tube. In children 12F Foley's catheter is used. Once complete filling occurs, X-ray is taken. Patient is asked to evacuate the barium and postevacuation film is taken. After that, air is inflated into the colon which gives a better contrast to visualise thin mucosa of the colon (**air-contrast barium enema).**

Contraindications

Acute colonic conditions.

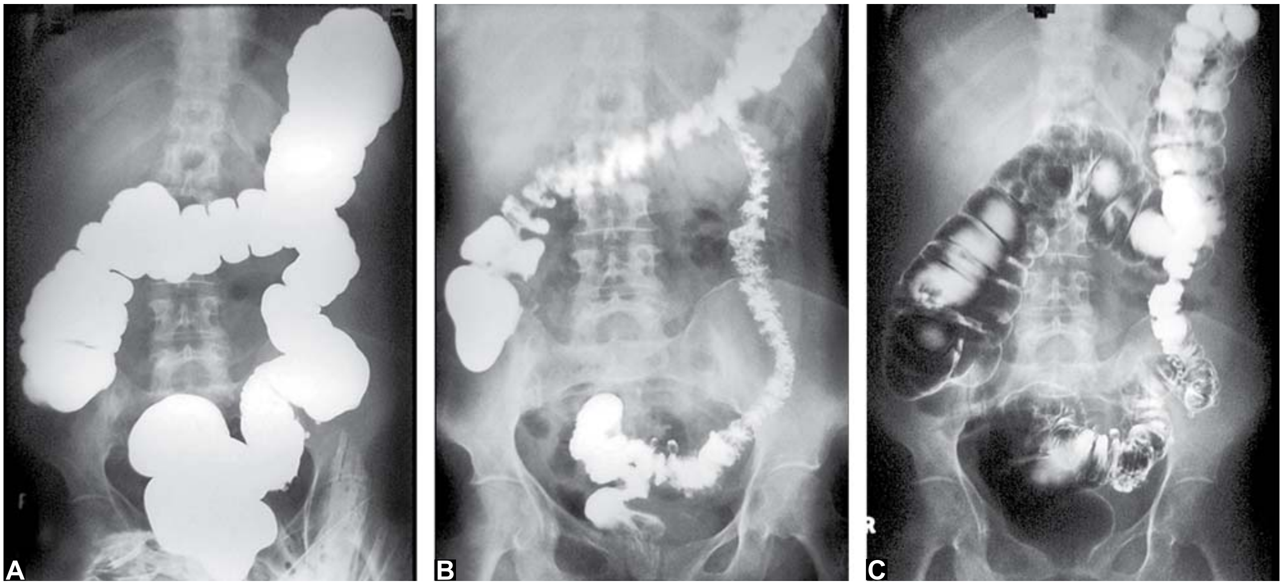

Figs 22.63A to C: Barium enema X-ray in different indications should be taken with complete filling, postevacuation and after air-contrast.

Chapter 23 Intestinal Obstruction

CHAPTER OUTLINE

- Intestinal Obstruction
- Dynamic Obstruction
- Duodenal Atresia
- Small Intestine Atresia
- Malrotation
- Meconium Ileus
- Intussusception
- Volvulus
- Sigmoid Volvulus
- Paralytic Ileus
- Adhesions and Bands
- Internal Hernias

INTESTINAL OBSTRUCTION: TYPES

Classification I: Depending on Aetiopathology

A. Dynamic.
B. Adynamic.

Classification II: Depending on Type of Obstruction

1. Acute: Common in small bowel.
2. Chronic.
3. Acute on chronic: Common in large bowel.
4. Closed loop obstruction.

	Dynamic		***Adynamic***
Outside the wall	*In the wall*	*In the lumen*	
Hernia—25% Adhesions—40% Volvulus Intussusception	Tuberculous stricture Crohn's disease Malignancy	Gallstones Roundworm Inspissated faeces Meconium ileus	Cessation of peristalsis Postoperative period Electrolyte imbalance Spinal injuries Uraemia Diabetes mellitus Retroperitoneal—haematomas and surgeries Renal surgeries Mesenteric ischaemia Pseudo-obstruction

Classification III: Depending on Site of Obstruction

	Proximal small bowel	***Distal small bowel***	***Large bowel***
Site of obstruction	*Duodenum and jejunum*	*Ileum*	*Any where in large intestine*
Causes:	Congenital Lipomas Leiomyomas Malignancy Bands and adhesions	Tuberculosis strictures Malignancy Crohn's Gallstones Hernias—common cause Roundworm Congenital	Malignancy Tuberculosis stricture Anorectal malformation Volvulus Congenital megacolon Bands
Clinical features:	Severe vomiting, dehydration, no or less distension, colicky pain	Central distension, vomiting, dehydration central abdominal pain	Constipation, distension—early; Late vomiting less pain
Special features:			
Plain X-ray	Valvulae conniventes	Characterless Central fluid level	Dilatation and haustration

Today's friction is tomorrow's destruction.

Classification IV

A. Congenital.
B. Acquired.

Congenital	***Acquired***
Anorectal malformations	Hernia (commonest)
Congenital megacolon	Postoperative
Adhesions	Intussusception
Duodenal atresia	Roundworm
Intestinal atresia (ileal)	Gallstones
Bands and adhesions	Tuberculosis
Malrotation	Malignancy
Volvulus neonatorum	Internal hernias

Bolus obstruction

- Worms
- Gallstones
- Trichobezoars
- Phytobezoars
- Meconium pellets
- Foreign body
- Food bolus
- Stercoliths

Intestinal obstruction still is a challenging, commonly acute condition in surgical practice. *Adhesions* are becoming the more common cause of intestinal obstruction than hernia nowadays. Common causes for adhesions are gynaecologic surgeries, appendicectomy, colonic and pelvic surgeries. 20% cases of intestinal obstruction are due to malignancy or its peritoneal *carcinomatosis* spread. More than 50% of intestinal obstruction in western countries are due to adhesions; 20% due to malignancies; 10-15% due to hernias. Crohn's disease is also becoming a common cause of obstruction.

Note:
Adhesions and hernias are most common causes of intestinal obstruction.

DYNAMIC OBSTRUCTION

- It is mechanical blockage of normal propulsion and passage of intestinal contents.
- Obstruction may be external/internal; partial (incomplete or subacute)/complete; acute/acute on chronic/or chronic; simple/closed loop/strangulation; congenital/acquired; proximal/distal.
- Earlier, hernia was the commonest cause of intestinal obstruction. Now adhesions (40%) are the commonest cause especially in developed countries and hernia being 15-25% common. In developing countries both adhesions and hernia are the common causes of intestinal obstruction.
- Adhesions commonly cause small bowel obstruction than large bowel.
- Eighty per cent of intestinal obstruction occurs in small bowel; 20% in colon. 70% of colonic obstruction is due to malignancy. Other 30% is due to volvulus; diverticulitis, inflammatory cause like tuberculosis, etc.
- Mortality is 3% in obstruction without strangulation; 30% in obstruction with strangulation.
- Recurrent obstruction is more common in adhesions.

Pathology and Pathogenesis

- ***Changes proximal to the bowel obstruction***:

Intestinal obstruction
↓
Increased peristalsis
↓
Becomes vigorous
↓
Obstruction not relieved
↓
Peristalsis ceases.
↓
Flaccid, paralysed, dilated bowel

  - Fluid collects just proximal to the obstruction which is derived from saliva, stomach, pancreas and intestine. Because of oedema and inflammation absorption decreases, sequestration of fluid from the circulation into the lumen occurs and bacteria *(E. coli, Klebsiella,* anaerobes, bacteroides and other organisms) multiply, toxins are released—*toxaemia* occurs. This leads to severe dehydration, electrolyte imbalance.

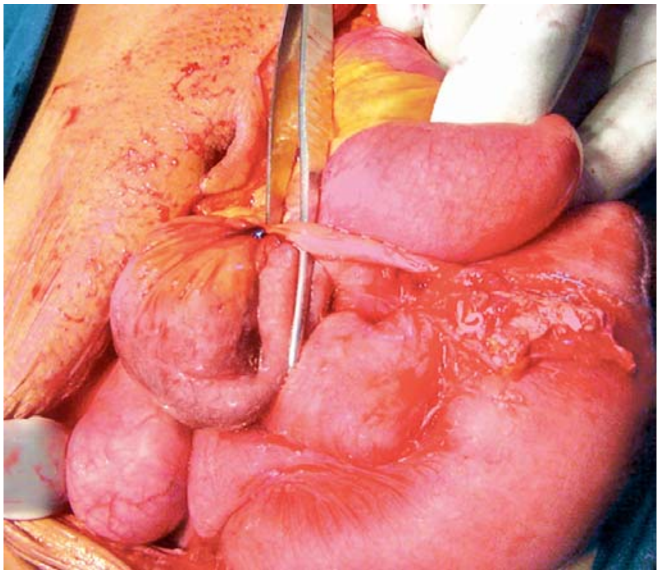

Fig. 23.1: Intestinal obstruction due to band. Note the ischaemic bowel under the band.

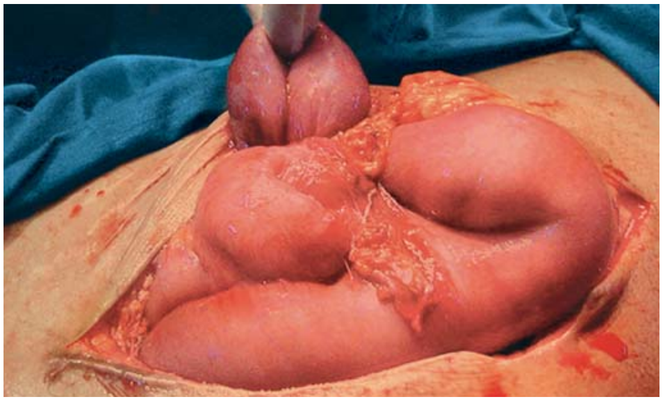

Fig. 23.2: Postoperative internal hernia with obstruction.

- Proximal to the collected fluid, air accumulates (derived from swallowed air (70%), diffusion from blood into the lumen (20%), from digested product and bacterial action (10%)), in which, main component is nitrogen (90%) and also hydrogen sulphide. During vigorous peristalsis air enters the distal fluid, results in c*hurning,* is the reason to cause *multiple air-fluid levels in plain X-ray abdomen.*
- *Defective absorption*, decreased fluid intake, loss of fluid by vomiting, sequestration of fluid into the bowel lumen—leads into severe dehydration, fluid and electrolyte imbalance.
- Inflammatory response in the bowel wall (*intramural inflammation*) causes accumulation of activated neutrophils and macrophages in the muscle wall which release reactive enzymes and cytokines. These substances damage secretory and motor process of muscle leading into dilatation of the bowel. Increased release of *nitric oxide* in muscle wall and production of intramural reactive oxygen metabolites alter gut motility and permeability.
- *Intestinal wall hypoxia* is also the cause for dilatation.
- In first 12 hours of obstruction, there is *only decreased absorption* which causes accumulation of fluid and electrolytes in the lumen.
- After 12 hours, there is also *increased intestinal secretion* causing further accumulation of the fluid.
- Accumulation of bacterial toxins, bile salts, prostaglandins, and mucosa derived free radicals, VIP—all increases the luminal secretion of fluid in obstructed bowel.
- Dilatation of bowel wall increases *intraluminal pressure* which exceeds the bowel wall *venous pressure* causing ischaemia which causes further dilatation and *ischaemic injury*. This leads into eventual blockage of arterial perfusion causing bowel wall necrosis/ gangrene.
- *Increased bacterial colony* in the bowel (Normal flora is less than 10^6 colonies/ml in jejunum and 10^8 colonies /ml in the ileum) due to altered luminal content and environment → multiplication → toxins → further mucosa damage → disrupted mucosal defense/ barrier/ integrity → *translocation of bacteria across mucosa* into submucosa and also absorption of bacterial and other toxins into the circulation → bacteraemia/toxaemia/ septicaemia/SIRS/MODS.

Factors causing systemic problems in intestinal obstruction

- Dilatation of the bowel
- Decreased absorption across mucosa
- Increased secretion into the lumen
- Intramural inflammation and hypoxia
- Increased intraluminal pressure
- Venous congestion and increased venous pressure
- Disrupted mucosal barrier → bacterial translocation

- ***Changes at the site of the obstruction:***

Initially venous return is impaired.

↓

Congestion, oedema of bowel wall occurs which turns purple.

↓

Later this jeopardizes the arterial supply.

↓

Loss of shineness, blackish discolouration, loss of peristalsis.

↓

Gangrene.

↓

Perforation occurs.

↓

Bacteria and toxins migrate into the peritoneum.

↓

Peritonitis.

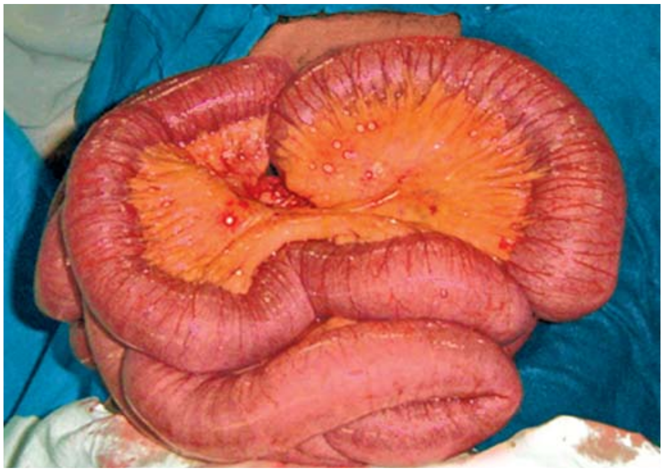

Fig. 23.3: Intestinal obstruction showing dilated bowel loops.

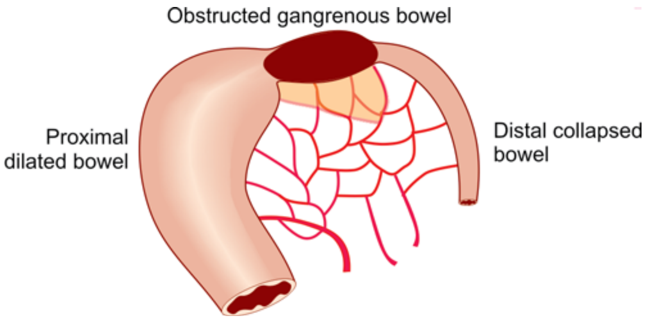

Fig. 23.4: Gangrenous bowel with proximal dilated and distal collapsed bowel segments.

> *The coexistence of intestinal colic and borborygmi, establishes the diagnosis of obstruction of the small intestine in more than 9 out of 10 cases.*
> —**Clarence Dennis**

Note:

- Causes for strangulation are—external like hernia (by constriction ring of the sac), adhesions, bands; compression in the wall causing mechanical block and compromised blood supply like in volvulus, intussusception; increased intraluminal pressure like closed loop obstruction; mesenteric ischaemia.
- Morbidity, poor outcome and mortality are more common in intraperitoneal strangulation than in strangulation of hernia as it is localised with less fluid loss.
- Outcome also depends on age of the patient, extent of the disease and time at which patient reaches for therapy.
- *Massive 3rd space fluid loss*, dehydration, hypovolaemia, hypotension, hypochloraemia, hypokalaemia, metabolic alkalosis, oliguria, azotemia, haemoconcentration (after fluid therapy haematocrit falls indicating the need for blood transfusion).
- *Translocation of bacteria* across oedematous small bowel wall causes septicaemia.

♦ ***Closed loop obstruction*:**

- When there is obstruction in the large bowel, with ileocaecal valve competence (40%), pressure increases in the caecum.

↓

Stercoral ulcer in the caecum.

↓

Gangrene.

↓

Perforation.

↓

Peritonitis (Faecal).

- Perforation also can occur at the site of obstruction due to the malignant growth.

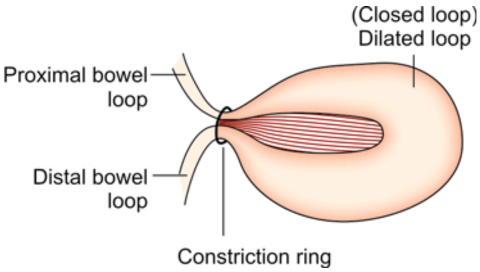

Fig. 23.5: Closed loop obstruction. Here loop of the bowel is obstructed at its point of entry and exit creating closed loop.

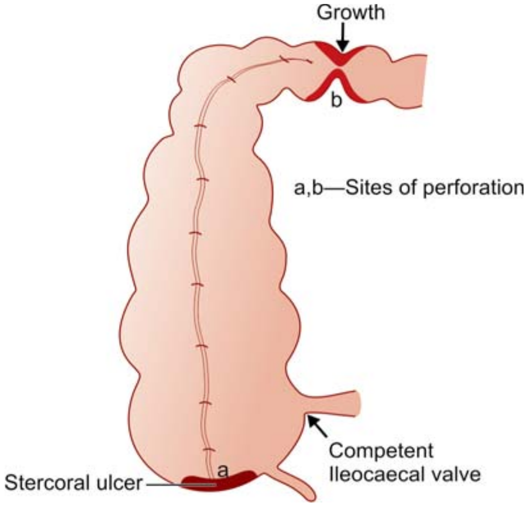

Fig. 23.6: Closed loop obstruction.

- Closed loop obstruction also can occur when bowel get obstructed at both proximal and distal parts of the loop of the bowel. It can occur in external or internal hernias, volvulus, etc. Necrosis and perforation are both common at obstructed site and over the convex summit of the bowel content.

♦ ***Bowel distal to the obstruction*** *is inactive and collapsed.*

Clinical Features

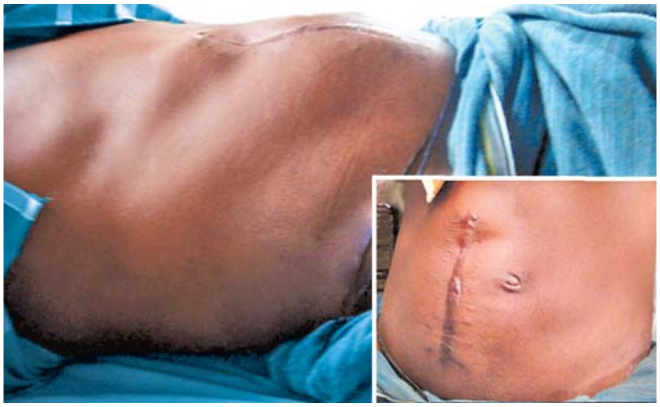

Fig. 23.7: Step ladder peristalsis seen in small bowel obstruction. Note the previous laparotomy scar signifies obstruction is due to adhesions. Laparoscopy and proceed—is the needed management.

♦ ***Abdominal pain*:**

- Initially colicky and intermittent: later continuous and severe.
- Pain is the first symptom to develop which is sudden and severe. Initial colicky pain suggests obstruction and eventual diffuse persistent pain suggests strangulation. Pain begins usually around umbilicus in small bowel obstruction.
- In small bowel obstruction, it is crampy, recurrent paroxysms occurring as short crescendo/decrescendo episodes (of 30 seconds).
- In large bowel obstruction, it is of longer episodes of minutes (In paralytic/adynamic ileus, pain is diffuse and mild).

♦ ***Vomiting*:**

- In *jejunal* obstruction, it is early and persistent.
- In *ileal* obstruction, it is recurrent occurring at an interval; initially bilious later faeculent.
- In *large bowel* obstruction, vomiting is a late feature.

♦ ***Distension*:**

- It is absent or minimal in case of *jejunal* obstruction
- Obvious with visible intestinal peristalsis (VIP) and borborygmi sounds in case of *ileal* obstruction—*Step ladder peristalsis*.
- It is enormous in case of *large bowel* obstruction.

♦ ***Constipation*:**

- It is *absolute,* i.e. neither faeces nor flatus is passed.

Exceptions

- Richter's hernia obstruction
- Gallstone obstruction
- Mesenteric vascular occlusion
- Intestinal obstruction with a pelvic abscess

- ***Dehydration***:
  - Leads to oliguria → renal failure.
- ***Features of toxaemia and septicaemia***:
  - Tachycardia, tachypnoea, fever, sunken eyes, cold periphery.
- ***Abdominal tenderness***:
  - It is initially localised but later becomes diffuse—is a feature of intestinal obstruction. Rebound tenderness and guarding will not be present in simple obstructions which are features of strangulation.
- ***Features of strangulation***:
  - Continuous severe pain, shock, tenderness, rebound tenderness *(Blumberg's sign).*
  - Guarding and rigidity, absence of bowel sounds.
  - In case of strangulated hernia, a swelling which is tense, tender, rigid, irreducible, no expansile impulse on coughing and history of recent increase in size is seen.
- ***Temperature***:
  - Fever signifies inflammation in the bowel wall/ ischaemia/perforation.
  - Hypothermia can occur when septicoemia develops due to lack of pyrogenic response. It suggests poor prognosis.
- ***Bowel sounds***:
  They are increased—high pitched metallic (rushes and groans) sounds followed by metallic tinkling sounds of dilated bowel. Eventually once fatigue occurs or gangrene develops, bowel sounds are not heard—silent abdomen of peritonitis develops (In paralytic ileus, there are only continuous metallic sounds of dilated bowel).
- ***Per-rectal examination***:
  - Shows *empty, dilated rectum*, often with tenderness. If rectal growth is the cause for obstruction, it may be palpable.

Investigations

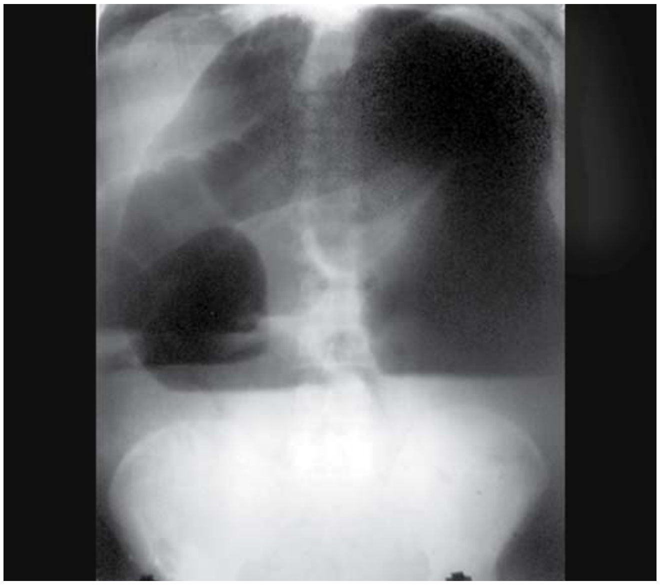

Fig. 23.8: Plain X-ray abdomen in erect posture showing dilated bowel and colon—a feature of intestinal obstruction.

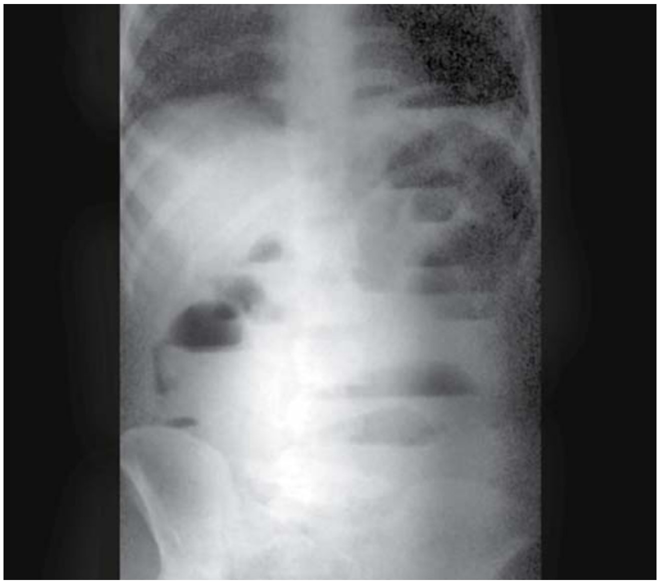

Fig. 23.9: Multiple air fluid levels in intestinal obstruction.

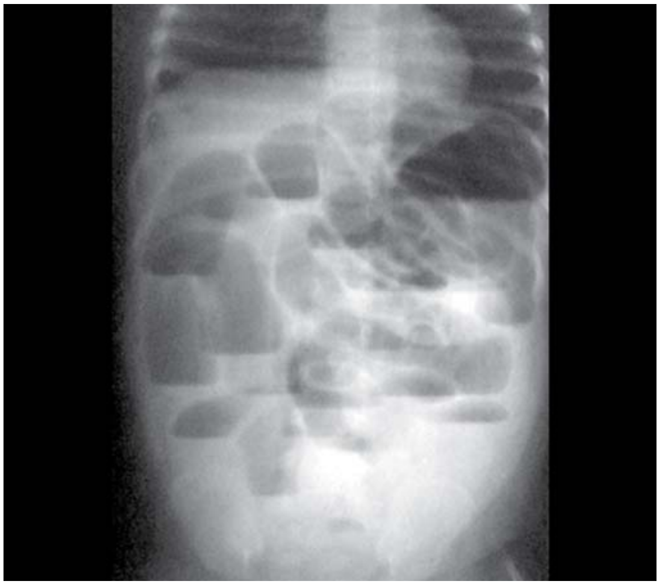

Fig. 23.10: Plain X-ray abdomen showing multiple air fluid levels due to bowel obstruction.

- ***Plain X-ray abdomen***: (initially supine abdominal X-ray is taken; later if needed X-ray in erect posture is taken if perforation is suspected).
  - *Multiple air-fluid levels.*
  - Proximal the obstruction → Lesser the air fluid level.
  - Distal the obstruction → More the air fluid level.

Normally, *three fluid levels can be seen* in plain X-ray film—at fundus of stomach, at duodenum and often at caecum.

  - *Jejunum* shows *concertina effect* due to *valvulae conniventes (Herring bone pattern)—by the valves of Kerckring.*
  - *Ileum is smooth and characterless* (by Wangensteen).
  - *Large* bowel shows *haustration.*
  - *Pneumobilia* (gas in biliary tree) may be due to gallstone ileus.

  - Distended caecum is shown as round gas shadow in the right iliac fossa. Dilated caecum signifies large bowel obstruction.
  - Small bowel > 3 cm diameter; proximal large bowel > 9 cm; transverse colon > 5.5 cm; sigmoid colon > 5 cm are suggestive of intestinal dilatation. But this increased diameter need not suggest intestinal obstruction everytime.

Triad of small bowel obstruction in plain X-ray

1. Dilated small bowel loops > 3 cm
2. Multiple air fluid levels in erect X-ray
3. Paucity of air in the colon

- *Barium (micro bar solution) enema or gastrografin contrast enema X-ray* is useful in intussusception. [*Barium meal is usually contraindicated in acute intestinal obstruction*. However dilute (micro bar) barium meal/gastrografin meal follow through X-ray may be done with caution in suspected subacute/partial intestinal obstruction under fluoroscopy, otherwise it may precipitate complete obstruction or may cause perforation and barium peritonitis which is very dangerous].
- Haematocrit, blood urea and serum creatinine; arterial blood gas analysis (acidosis is common), LFT, platelet count (In severe sepsis there will be altered LFT with thrombocytopenia).
- Serum electrolytes estimation. Hypokalaemia is common.
- Total count is increased. But can be significantly low in severe stage of sepsis.
- Estimation of serum D lactate, CPK – BB isoenzyme, intestinal fatty acid binding protein are different investigations may be useful to predict bowel ischaemia/gangrene.
- *US abdomen* is useful to see dilated bowel and fluid in the peritoneal cavity. It is better than X-ray but not as good as CT scan. It has got 95% sensitivity; 80% specificity; 80% accuracy. Doppler US is useful in detecting strangulation.
- *CT scan* is very reliable investigation for intestinal obstruction. It has got 93% sensitivity; 94% accuracy and 100% specificity. In CT scan small bowel loop > 2.5 cm suggests dilatation. It can show dilated loop, transition zone and collapsed part which are definitive features of intestinal obstruction. It can also give idea of changes in the bowel wall, ischaemia, strangulation, mesenteric oedema and thickening. It also shows *bowel wall gas, portal venous gas* and mass lesion.
- Basic electrical rhythm of small bowel will be changed in ischaemia. It can be determined by noninvasive method using *superconducting quantum interference device (SQUID).*

Complications of intestinal obstruction

- Peritonitis
- Hypovolaemic and septic shock
- Renal failure
- ARDS
- Intra-abdominal abscess formation
- Moribund status

Differential diagnosis

- Paralytic obstruction
- Pseudo obstruction
- Ascites

Treatment

- *Nasogastric aspiration*: To reduce toxic effects, to reduce bowel distension which indirectly improves pulmonary ventilation and to reduce possibility of aspiration pneumonia.
- *Replacement of fluid and electrolytes*.
- *Antibiotics:* Ampicillin, gentamycin, metronidazole, cephalosporins.
- *Blood transfusion*: FFP or platelet transfusions are often needed in critical patient.
- *ICU critical care*: Systemic management of complications like ARDS, DIC, SIRS are important. If there is hypotension, dopamine/dobutamine are also needed.
- *CVP for fluid and monitoring*: PCWP (pulmonary capillary wedge pressure) monitoring are often needed in haemodynamically unstable patient.
- *Surgery***:**
  - ***Immediate laparotomy*** is done and the site (by finding the junction of dilated proximal and collapsed distal bowel) and cause of the obstruction is identified. The obstruction is relieved.

To check for viability of bowel, look for

- Peristalsis
- Pulsations
- Bleeding in mesentery and bowel wall
- Friability—friable, flabby muscle is seen in ischaemia
- Colour (black/pink)—dull and lusterless serosa is seen in ischaemia
- Serosal shining

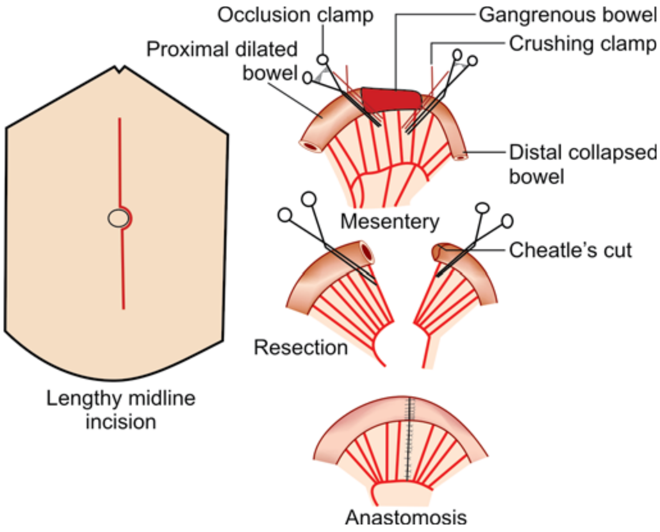

Fig. 23.11: Bowel resection and anastomosis for gangrene in case of intestinal obstruction. Cheatle's cut on the antimesenteric margin of the collapsed distal segment is often needed to avoid discrepancy in luminal width. Single layer interrupted (silk/vicryl) or two layered continuous sutures can be used. Single layer is better in acute conditions.

- *Warm saline soaked mop* is placed over the doubtful area with 100% oxygen inhalation for 20 minutes; if colour becomes normal with peristalsis then bowel is viable.
- *On table Doppler* study may be useful.
- *Fluorescein fluorescence study* may be helpful on table to check the viability.
- If bowel is not viable *resection and anastomosis is done.* A good peritoneal wash is given and the abdominal cavity is drained.
- Abdomen is closed in layers using nonabsorbable sutures (polyethylene, polypropylene, nylon). Often tension sutures are required.
- Small bowel can be decompressed using *Savage's decompressor.*
- In case of *right-sided colonic* obstruction, *right hemicolectomy with ileocolic* anastomosis is done.
- In case of *left-sided colonic* obstruction, *left hemicolectomy* (resection) and *colo-colic* anastomosis is done with a *defunctioning colostomy* (right-sided transverse) which is closed after 6 weeks.
- Obstruction due to *rectosigmoid* growth with patient being severely ill—*Hartmann's operation* can be done to save the life of the patient wherein distal stump after removal of the growth is closed, proximal colon is brought out as *end colostomy.*
- *Second look operation* may be needed in doubtful cases or multiple segment obstructions in 24-48 hours to confirm viability.
- *Laparoscopic approach* may be useful in partial obstruction, proximal obstruction, obstruction due to band. Conversion when needed should be done without hesitation.
- *Acute postoperative obstruction* is difficult to identify and manage. CT is very useful. Initially it is treated conservatively (90%), but suspected ischaemic cases or persistent obstruction becomes an indication for surgery. Resection with exteriorization may be the choice.

Postsurgery Complications

- Pelvic abscess.
- Subphrenic abscess.
- Biliary or faecal fistulas.
- Burst abdomen.
- Bands and adhesions.
- Incisional hernias.

DUODENAL ATRESIA

- It is the commonest site of intestinal atresia.
- It is usually a complete stenosis of the second part of duodenum at the level of ampulla of Vater.
- It is defective fusion of foregut and midgut with failure of recanalisation.
- Incidence is one in 10,000 live births.

Types

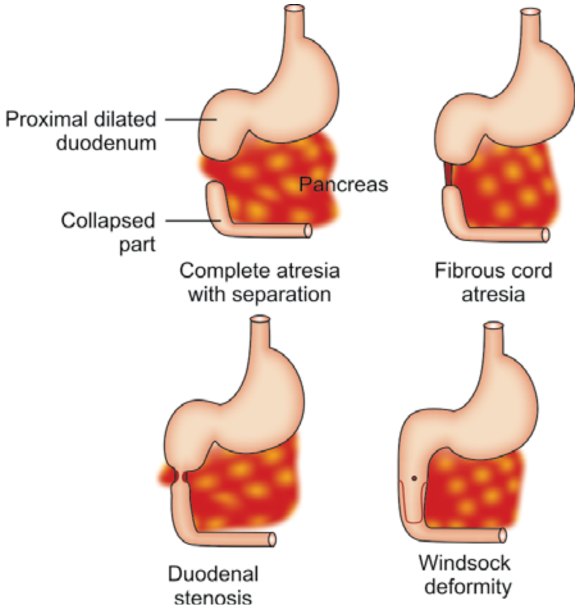

Fig. 23.12: Types of duodenal atresia—complete; fibrous cord; duodenal stenosis and windsock deformity.

- Type 1: Duodenal complete atresia: It is the commonest type (50%). It is usually complete separation with intact wall. In 25% cases complete separation with wall also occurs.
- Type 2: Fibrous cord.
- Type 3: Incomplete or partial obstruction. It can be stenosis or web with an aperture.
- Duodenal diaphragm/web can present as complete; incomplete with a fenestra; incomplete diaphragm with central aperture which causes '*windsock' deformity* due to proximal dilatation. Here actual diaphragm will be proximal to the site of obstruction.
- Duodenal atresia may be *preampullary* (nonbilious vomiting) or *postampullary* (bilious vomiting). Postampullary is common (80-90%).

Associations

- Duodenal stenosis is often associated with annular pancreas.
- It can be isolated duodenal atresia or in association with Down's syndrome (30%)/incomplete rotation of gut (20%)/congenital heart diseases (30%)/trisomy (30%)/anorectal malformations (10%), etc.
- It is commonly associated with maternal polyhydramnios (50%). Antenatal US can confirm it. 50% infants are premature.

Features

- Jaundice.
- Bilious/nonbilious vomiting immediately after birth.

- Features of gastric outlet obstruction.
- Dehydration. Electrolyte changes are common.
- Growth retardation of newborn due to deprived nutrition (by swallowed amniotic fluid in fetus).

Investigations

- Plain X-ray shows classic *double-bubble sign with absence of air in the distal part.*

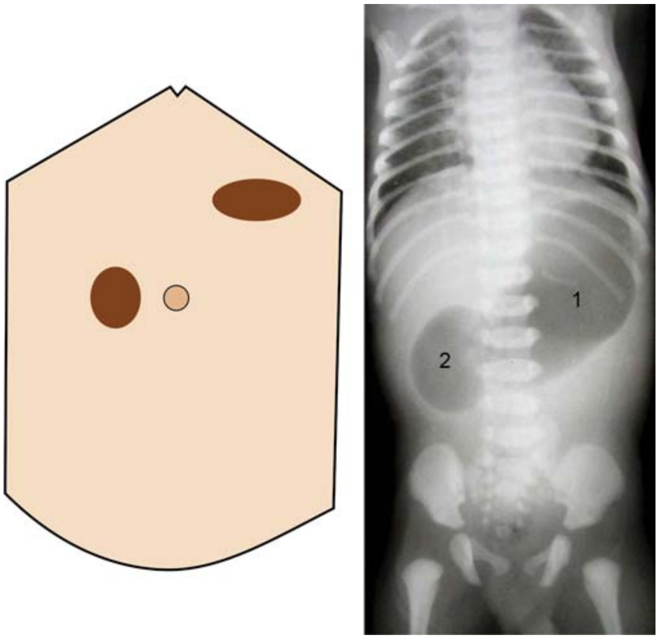

Fig. 23.13: Double-bubble sign in duodenal atresia.

- In partial obstruction, air may be present in distal loop and so contrast study has to be done. Risk of aspiration in newborn infant should be remembered.
- U/S will show distended stomach and proximal duodenum, *rail road track* duodenum and features of associated anomalies. Maternal and fetal ultrasound in pregnancy may identify the pathology and also maternal polyhydramnios.

Treatment

- Proper preoperative preparation like correction of fluid and electrolytes; gastric decompression; TPN; injection vitamin K; evaluation for associated anomalies.
- *Duodenoduodenostomy* is done. Associated malrotation should be corrected in these patients (Ladd's operation). Side-side duodenoduodenostomy may cause dilated duodenum (megaduodenum—30%); anastomotic dysfunction; and delayed transit of the content.
- *Kimura's diamond shaped anastomosis* between transversely opened proximal pouch and longitudinally opened distal pouch reduces the problems of anastomosis. Presence of bile in the duodenum and proximal and distal patency should be confirmed by saline irrigation.
- Transanastomotic nasojejunal or gastrostomy tube for feeding purpose is needed as prolonged postoperative ileus is the usual problem.

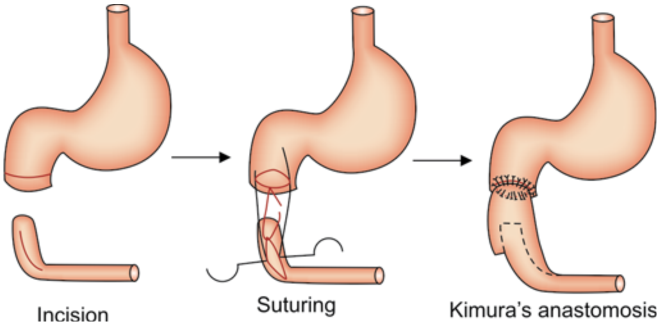

Fig. 23.14: Duodenoduodenostomy is ideal in duodenal atresia. Kimura's modification has got lesser chance of developing side effects like gastroparesis.

- Duodenojejunostomy may cause blind loop problems. Gastrojejunostomy should not be done due to high incidence of marginal ulceration and bleeding.

SMALL INTESTINE ATRESIA (INTESTINAL ATRESIA)

- It is jejunoileal atresia.
- It is due to intrauterine mesenteric vascular accidents (occlusion) of the segments affected. V-shaped mesentery; presence of bile pigments in the distal segments suggesting earlier period of patency—are supportive for the above theory.
- It is often associated with malrotation, gastroschisis, volvulus—20%.
- Common site is proximal jejunum; next common is distal ileum.
- Maternal polyhydramnios occurs in 35% of jejunal atresia.
- Proximal bowel wall is dilated with hypertrophy but villi are normal. Distal bowel is collapsed but with hypertrophied villi. Narrow collapsed large intestine (microcolon) is seen in proximal atresia.

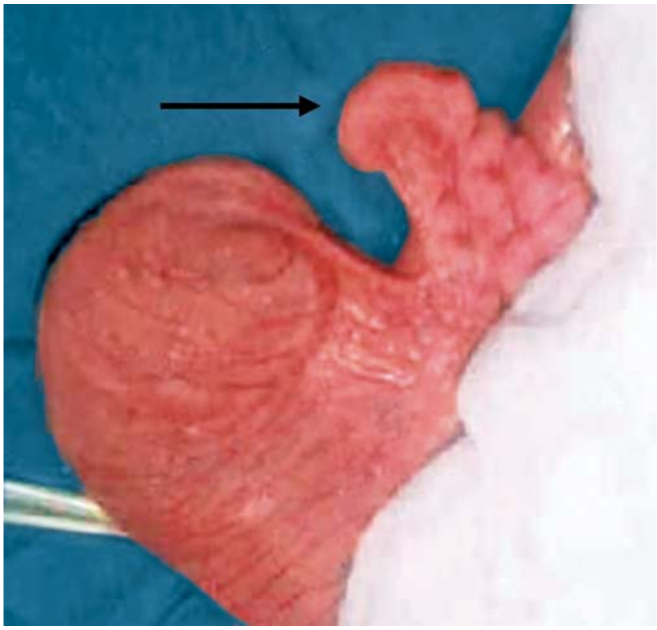

Fig. 23.15: Typical intestinal atresia.

Griesfield Modification of Martin's Classification of Intestinal Atresia

- Type I: Membranous/mucosal with normal mesentery—20%.
- Type II: The lumen is atretic—fibrous cord between proximal and distal parts of the segment involved (only one atretic segment) but mesentery is normal—40%.
- Type III (a): Atresia with complete separation of proximal and distal ends and V-shaped defect of mesentery—35%.
- Type III (b): Atresia with christmas tree-shaped defect in mesentery with distal bowel being supplied by single artery—right colic/ileocolic/superior mesenteric—*apple peel atresia*.
- Type IV: Multiple atresias—5%.

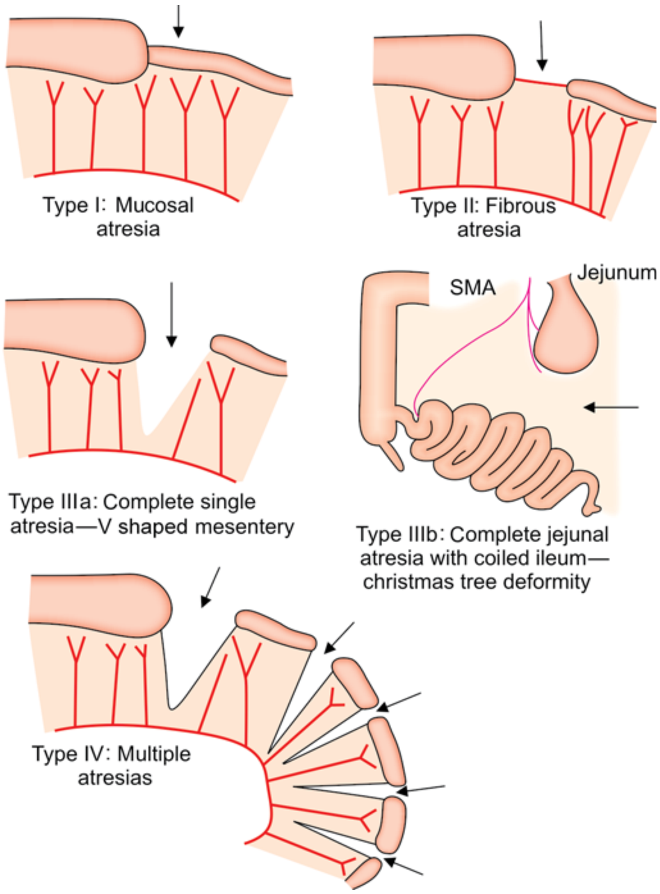
Type I: Mucosal atresia
Type II: Fibrous atresia
SMA
Jejunum
Type IIIa: Complete single atresia—V shaped mesentery
Type IIIb: Complete jejunal atresia with coiled ileum—christmas tree deformity
Type IV: Multiple atresias

Fig. 23.16: Types of intestinal atresia.

Clinical Features

- Equal in both sexes.
- Commonly infant is of low birth weight.
- Bilious vomiting with features of intestinal obstruction like distension.
- Features of associated anomalies.
- Respiratory distress.
- Jaundice.

Investigations

- Plain X-ray abdomen shows—*triple-bubble appearance* in jejunal atresia; multiple air fluid levels in ileal atresia.
- Barium enema may show narrow microcolon. Calcification suggests antenatal fetal bowel perforation.
- US abdomen to confirm associated anomalies.

Differential Diagnosis

- Other causes of neonatal intestinal obstruction like duodenal atresia, malrotation, volvulus of midgut.

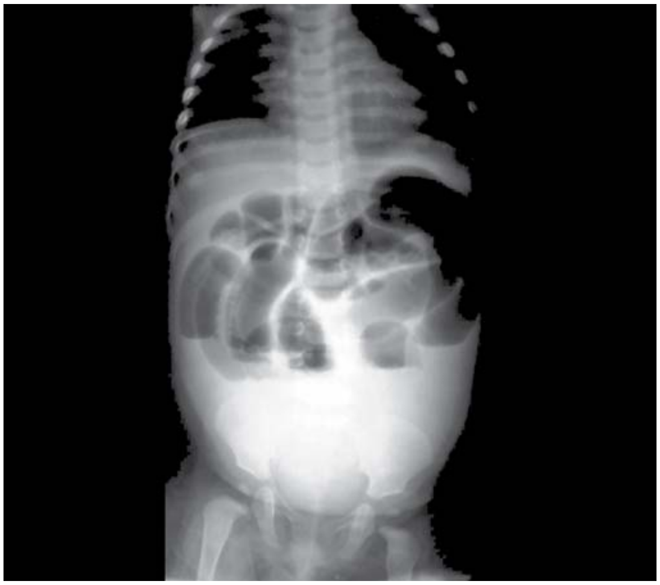

Fig. 23.17: X-ray abdomen of a patient with intestinal obstruction in a newborn as neonatal intestinal obstruction.

Treatment

- Resection and anastomosis is the choice of therapy.
- Treatment of associated anomalies.
- Tapering proximal jejunoplasty is done if extensive length of jejunum is involved.
- Prognosis depends on length of atretic bowel; multiple atresias; associated anomalies.

MALROTATION

It is interference in the process of normal rotation of midgut in fetus and its mesenteric fixation.

Stages of Normal Rotation of Midgut

- *Stage 1*: In 4th-8th week of intrauterine period, midgut supplied by superior mesenteric artery (SMA) grows rapidly. As coelomic cavity cannot accommodate growing midgut during that period, it protrudes into the umbilical cord as *physiological hernia*.

> *If the hand is kept flat upon the abdomen the underlying coil may be felt to harden and soften alternately much like in a pregnant uterus, in a case of intestinal obstruction.*
> — **Arthur H Burgess**

- *Stage 2*: In 10th-12th week, midgut migrates into coelomic cavity. First, small bowel returns towards the left side of the abdomen. Then caecocolic loop returns to left lower abdomen. It rapidly rotates 270° counterclockwise to reach right iliac fossa. Then duodenojejunal segment rotates 270° counterclockwise to reach left of SMA and behind the colon.
- *Stage 3*: Fusion of different parts of mesentery and posterior peritoneum.

Different Errors of Rotation

- *Stage 1*: Exomphalos major/minor or gastroschisis.
- *Stage 2*: Errors of rotation in this stage is important and is usually considered under malrotation.
    - *Nonrotation*: Causing small bowel in right side; colon in left side; caecum in midline as suspension.
    - *Incomplete rotation*: It is the *most common type* of malrotation. Caecum is located in subhepatic right hypochondrium. *Ladd's peritoneal band* connects from caecocolic loop to posterior abdominal wall compressing the 2nd part of the duodenum. Entire midgut is hanging down along with SMA with a narrow based mesentery causing midgut volvulus.

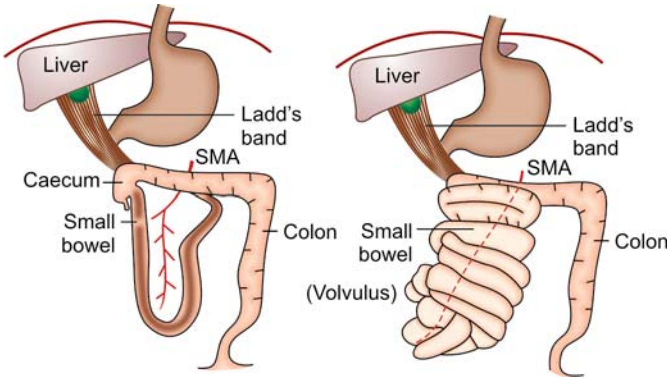

Fig. 23.18: Diagram showing incomplete gut rotation and small bowel volvulus.

    - *Reverse rotation*: Final 180° rotation occur clockwise bringing colon posterior to duodenum and SMA.
    - *Hyperrotation*: Rotation up to 360° or 450° causing caecum on left side of the abdomen.
    - Encapsulated small bowel occurs while fetal midgut returns into coelomic cavity.
- *Stage 3*: Final defect in fixation causes mobile caecum and ascending colon leading into caecal volvulus.

Associated Anomalies (20%)

- Congenital diaphragmatic hernia of Bochdalek.
- Prune belly syndrome.
- Duodenal atresia.
- Oesophageal atresia.

Presentations

- Acute/recurrent/subacute intestinal obstruction.
- Midgut volvulus (30%) (usually clockwise rotation) with features of strangulation, perforation, peritonitis.
- Shock, septicaemia, passage of dark blood per rectum, oedema and erythema of anterior abdominal wall.
- In children—failure to thrive, recurrent abdominal pain, cyclical vomiting, constipation and diarrhoea.

Investigations

- Plain X-ray abdomen shows air fluid levels.
- Barium meal (dilute/microbarium) and follow through X-ray is the investigation of choice.
- US abdomen/CT abdomen, if needed.
- Haematocrit, serum electrolytes estimation.

Treatment

- Resuscitation, antibiotics, fluid and electrolytes, blood transfusion.

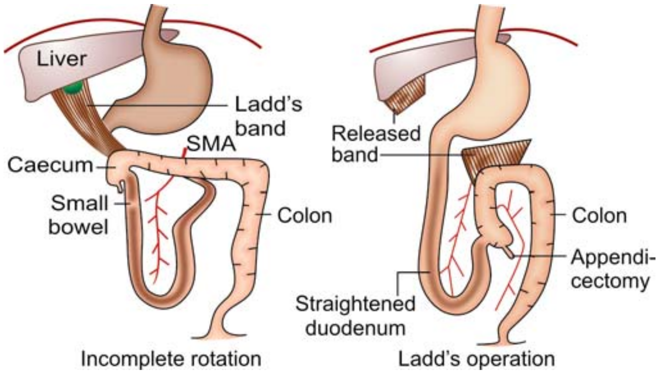

Incomplete rotation

Ladd's operation

Fig. 23.19: Ladd's operation for incomplete gut rotation. Ladd's band is released. Duodenum is straightened. Appendicectomy is done.

- *Laparotomy* through horizontal incision is done. Clockwise rotated midgut which is congested and cyanotic is identified. Untwisting of the midgut in counterclockwise direction is done. Viability of bowel is confirmed (colour, vessels in mesentery, peristalsis, on table Doppler). *Ladd's band* is divided *large bowel is repositioned in left side.* The entire duodenum is *Kocherised* and the ligament of Treitz is divided so that *duodenum becomes straight towards right* iliac fossa. This achieves *wide root of the mesentery* and places the *small bowel in the right* side of the abdomen thus preventing further volvulus. A *complementary appendicectomy* is done—*Ladd's operation.*
- After laparotomy, if bowel is gangrenous, it is resected and remaining parts of the bowel are exteriorised as enterostomies. A second look operation is done to look for viability of remaining bowel and also to maintain the continuity. Often it will be extensive bowel resection leading into poor prognosis.

MECONIUM ILEUS

- It is neonatal manifestation of fibrocystic disease of the pancreas wherein thick meconium, which is viscid and paste like, gets collected in the terminal ileum. Because of inspissation it forms a firm bolus leading to obstruction of the ileum.
- Neonates present with features of ileal obstruction as well as respiratory dysfunction, exocrine pancreatic insuffi-

ciency, and high salt in the sweat (Na^+ and Cl^- more than 90 mmol/L).

- Meconium ileus occurs in 15% of patients with cystic fibrosis. Cystic fibrosis is an autosomal recessive disease involving bronchioles, exocrine pancreas and sweat glands. Condition is common in Caucasians. Exocrine pancreatic insufficiency and malabsorption is seen in 90% of patients with cystic fibrosis.
- There is hypertrophy and dilatation of the proximal ileum containing thick, viscid, tenacious dark green meconium. Distal ileum and colon are narrow and contracted having grayish meconium pellets. Meconium gets calcified very rapidly. Gangrene, perforation, volvulus can occur in 50% of cases.
- Intrauterine perforation causes fetal meconium peritonitis which leads to dense adhesions and calcification in peritoneum. Fetal meconium peritonitis is sterile.

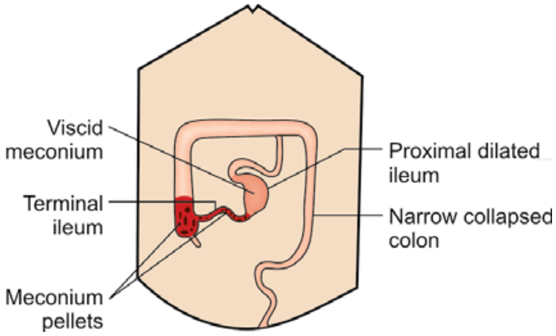

Fig. 23.20: Pathology of meconium ileus. Note the meconium pellets in the distal ileum and colon; thick viscid meconium in proximal ileum.

- Plain X-ray shows calcified meconium pellets with multiple air fluid levels which appear as 'soap-bubbles' (*Neuhauser sign*).
- Vomitus of the patient which does not contain trypsin, when poured on the exposed X-ray film will not digest the gelatin of the film whereas the vomitus of individual with normal pancreas will digest the gelatin of X-ray film—*very useful test.*
- Pilocarpine, a cholinergic drug is injected into skin to stimulate the sweating and collected sweat (100 µg sweat) is analysed for sodium and chloride.
- Elevated albumin level in meconium, sodium level assay in nail clipping, serum immune active trypsin assay are other investigations.

Meconium ileus

- Commonly associated with cystic disease of pancreas but not necessarily always
- Respiratory dysfunction
- Exocrine pancreatic insufficiency
- High salt in the sweat > 90 mmol/L
- Ileal obstruction
- Soap-bubble appearance in X-ray
- Gelatin in X-ray film will not get digested by patient's vomitus
- Bishop-Koop operation, ileostomy, dissolving ileal meconium pellets are required

Complications of meconium ileus

- Intestinal bolus obstruction
- Perforation and peritonitis
- Gangrene, volvulus formation

Treatment

Nonoperative measures

- Dissolution through enema can be tried using *gastrografin* which is hyperosmolar and contains Tween 80 as dissolving agent. *Gastrografin* is diatrizoate meglamine with Tween 80 (polysorbate 80) as dissolving agent. Acetyl choline (2-4%) also can be used for irrigation per-anally.
- Acetyl choline wash through nasogastric tube also can be used—5-10 ml 6th hourly.
- Treatment for cystic fibrosis.

Operative measures

- When patient's condition is critical with obstruction, *Bishop-Koop operation* is done. Proximal dilated segment is resected and resected end is anastomosed to side of the distal collapsed ileum. End of the distal ileum is brought out as ileostomy. Through the ileostomy gastrografin or acetyl cystine wash is given regularly to dissolve meconium pellets. This ileostomy can be kept for long time. Continuity is maintained at later period.
- *Santulli operation* is done where proximal ileum is brought as ileostomy for irrigation using a fine tube and distal ileum is sutured to proximal ileum as end to side anastomosis.
- *Resection and anastomosis* is ideal if patient is fit and if proximal bowel is suitable for anastomosis.

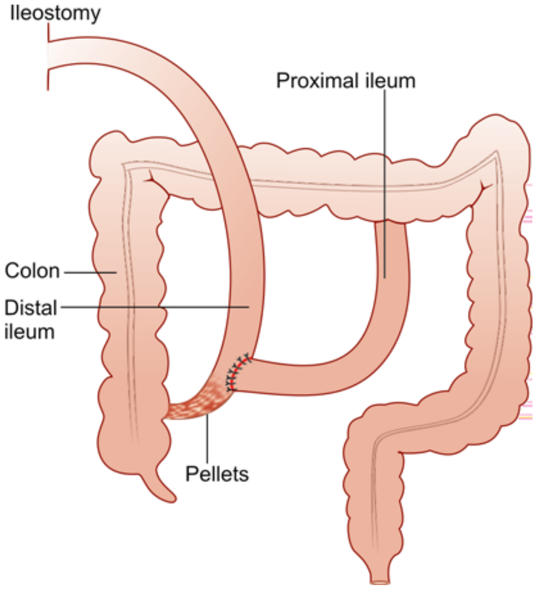

Fig. 23.21: *Bishop-Koop operation*. Here distal ileum is brought as ileostomy and proximal part sutured to distal bowel.

Management of meconium ileus

- *Nonoperative wash per rectally* using acetyl cysteine or gastrografin through Foley's catheter passed per anally. Acetyl cysteine can also be passed for irrigation per orally through a Ryle's tube as 10 ml 6th hourly.
- *Bishop-Koop and Santulli operations* done in very sick neonates.
- *Resection and anastomosis* is ideal when child is adequately fit.

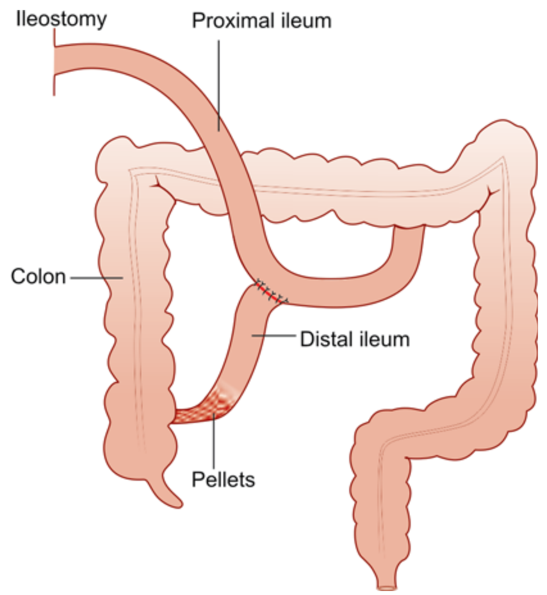

Fig. 23.22: *Santulli operation* done for meconium ileus wherein proximal ileum is brought out as ileostomy. Distal ileum is sutured to proximal.

Causes neonatal intestinal obstruction

- Hirschsprung's disease
- Duodenal atresia
- Intestinal atresia
- Malrotation
- Midgut volvulus/volvulus neonatorum
- Meconium ileus
- Anorectal malformation

INTUSSUSCEPTION (ISS)

Definition

It is telescoping or invagination of one portion (segment) of bowel into the adjacent segment.

Types

1. ***Antegrade***: Most common.
2. ***Retrograde***: Rare (jejunogastric in gastrojejunostomy stoma).

In elderly intussusception:

- Colocolic is most common type
- Apex is formed usually by growth
- No role of hydrostatic reduction

- It can be single or multiple (rare).
- It can be *ileo-colic (most common type, 75%),* colocolic, ileoileocolic, colocolic.
- It is common in weaning period of a child (common in male), between the period of 6-9 months.
- It is the *commonest cause* of intestinal obstruction in children of *6-18 months* age.

Aetiology

- Idiopathic ISS is common in children, occurs in terminal 50 cm of ileum.
- During weaning, change in diet causes inflammation and oedema of Peyer's patches—may stimulate ISS.
- Upper respiratory tract viral infection which causes oedema of Peyer's patches is also thought as an aetiology for intussusception in children.
- Other causes in adolescents and adults are submucous lipoma, leiomyoma, polyps in jejunum (Peutz-Jegher syndrome), other polyps and carcinomas with papillary projections.

Causes

- Change in diet during weaning
- Upper respiratory tract viral infection
- Intestinal polyps
- Submucous lipoma
- Leiomyoma of intestine
- Meckel's diverticulum
- Carcinoma
- Purpuric submucosal haemorrhages

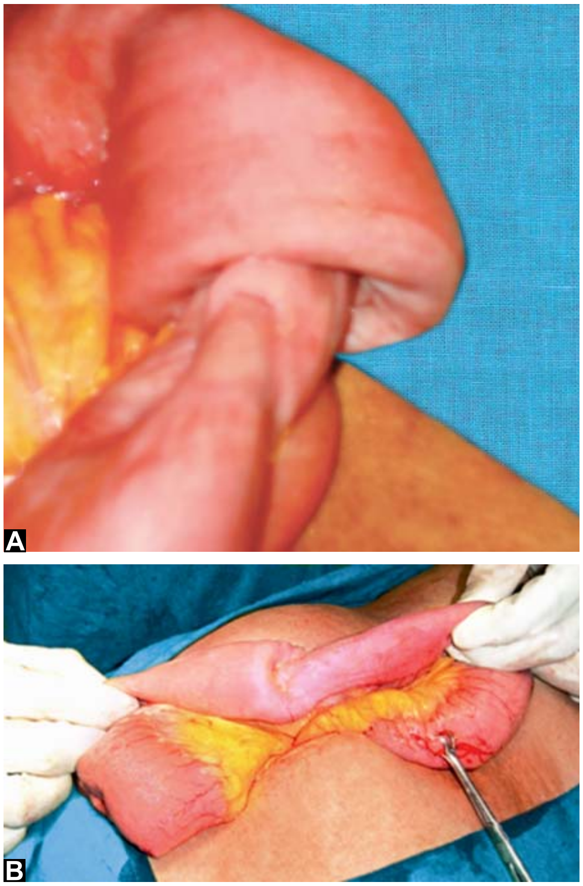

Figs 23.23A and B

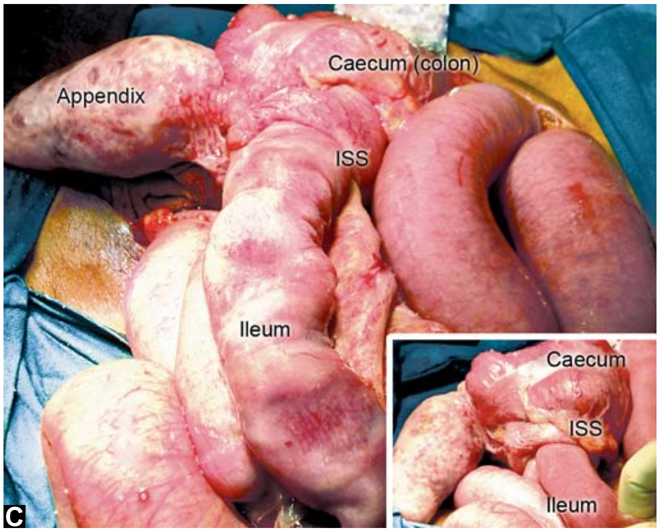

Fig. 23.23C

Figs 23.23A to C: Intussusception in different patients. Different parts and apex point is also clearly seen.

Pathology

- ***Apex*** is the one which advances;
- ***Intussuscipiens*** is the one which receives (outer sheath);
- ***Intussusceptum*** are the tubes which advances (middle and inner sheath).
  - Apex and inner tubes will have compromised blood supply which lead to gangrene.
  - Because of ischaemia, apex sloughs off and bleeds, which mixes with the mucous to produce the classic *red-currant jelly* that is passed per anum.
  - Gangrene which sets in leads to perforation and peritonitis.
  - Red currant jelly is not commonly observed in ISS in adult, but it can occur.

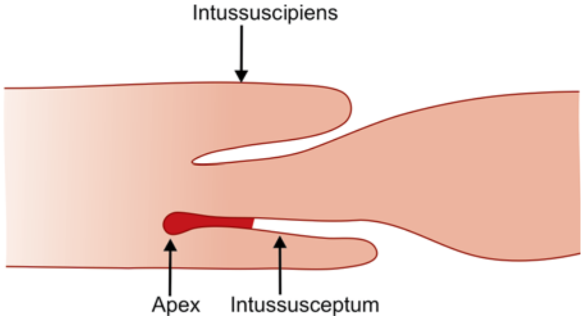

Fig. 23.24: Parts of intussusception.

Clinical Features

- Common in males (3:2).
- Common in 6-9 months. But can also occur at later age grouped children.
- Common in spring and winter, coinciding with the gastroenteritis and respiratory infections in respective periods.
- Commonest cause of intestinal obstruction in infancy.
- Initial colicky abdominal pain (75%) which eventually becomes severe and persistent.
- Sudden onset of pain in a male child, with progressive distension of the abdomen, vomiting, with passage of *"red-currant-jelly" stool.* It is usually not found in adult ISS.
- Often ISS is recurrent, when it gets reduced, child automatically becomes asymptomatic (Mother often complains "Bachha rotha he, Bachha sotha he". It means child cries during an episode and sleeps peacefully once it gets reduced).

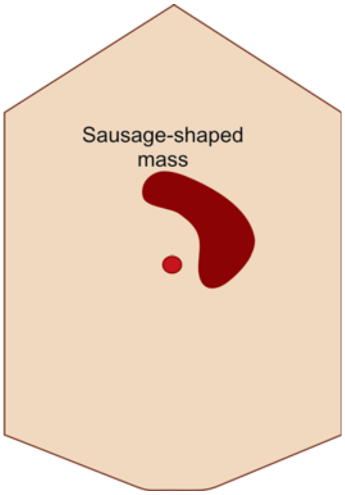

Fig. 23.25: Sausage-shaped mass of intussusception.

- *On examination*, a mass is felt either on the left or right of the umbilicus which is sausage shaped with concavity towards umbilicus, smooth, firm, resonant, not moving with respiration, mobile, *contracts under the palpating fingers.* Often mass appears and disappears.
- Right iliac fossa is empty *(Sign of Dance).*
- After 24-48 hours, abdominal distension appears and increases progressively with features of intestinal obstruction.
- Features of intestinal obstruction with *step-ladder peristalsis*.
- Blood stained stool is often obvious on digital examination of the rectum. Occasionally ISS can be seen per anally and felt with a long mesentery.
- Eventually, gangrene and perforation occurs with features of the peritonitis.

Abdominal palpatory findings in ISS

- Palpable mass (85%)
  - Sausage shaped smooth, firm mass
  - Mass does not move with respiration
  - Mobile in all directions
  - Resonant
  - Mass contracts under the palpating fingers
  - Mass appears and disappears
- Empty right iliac fossa
- Features of intestinal obstruction/peritonitis—later

In the field of observation, chance favours only the prepared mind.

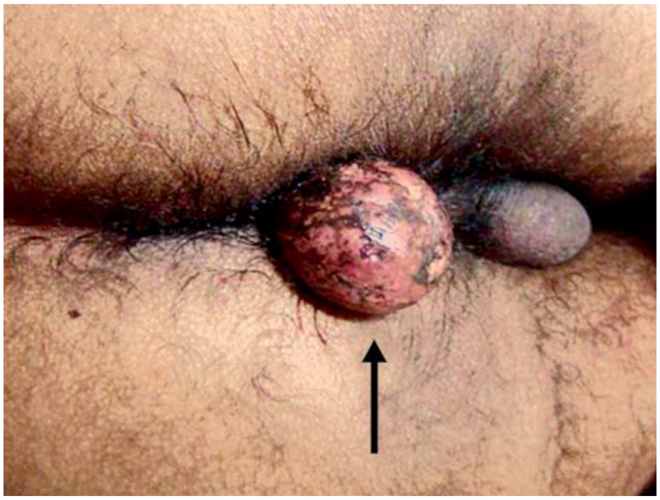

Fig. 23.26: A rare condition where colonic intussusception has occurred through rectum, coming out per anally. Part is already gangrenous.

Complications

- Intestinal obstruction
- Perforation
- Peritonitis

Differential Diagnosis

In children:	Acute gastroenteritis. Purpura with intestinal symptoms.
In adults:	Carcinoma colon. Mesenteric mass.

Investigations

- Barium enema shows typical *claw sign or coiled spring* sign (Pincer end).
- Ultrasound shows *target sign or pseudokidney sign or bull's eye sign, which is diagnostic*.
- *Doppler* may show mass with *doughnut sign* and is useful to check *blood supply* of bowel.
- Plain X-ray abdomen shows *multiple air fluid levels*.

Treatment

Initial management

- Ryle's tube aspiration.
- IV fluids.
- Antibiotics.
- Catheterisation.

Later management

Nonoperative management

- *Reduction by hydrostatic pressure* using either *warm saline or microbarium sulphate solution or air* (popular in ***China).*** Barium or saline is infused into the rectum through a catheter (Foley's catheter). Under fluoroscopy, reduction can be observed. Child will pass large quantity of air and faeces; distension reduces; child shows recovery and stops crying. Air or contrast enters the proximal bowel freely. *Palpable abdominal mass disappears*. Rare complication is perforation of colon. It is done in early stage within 24 hours of presentation. 70% cases of ISS will respond to nonoperative method. It is *contraindicated* in complete obstruction; perforation and peritonitis.

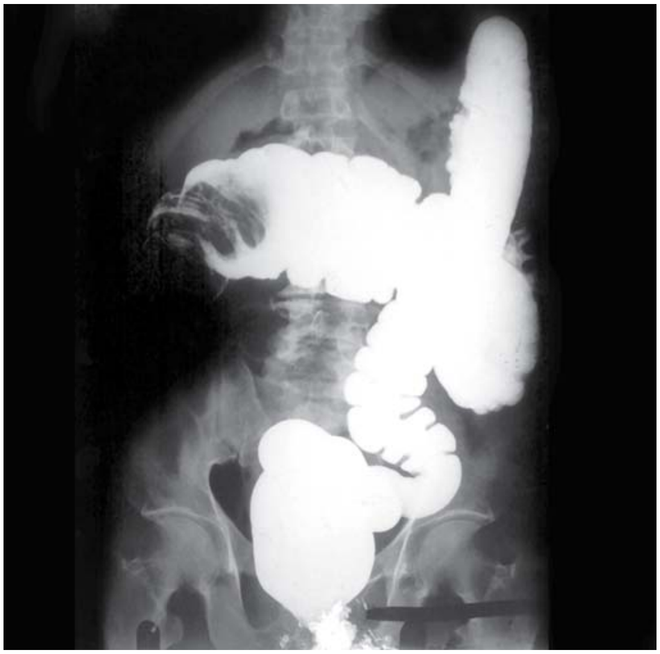

Fig. 23.27: Barium enema showing the typical 'Claw sign' of intussusception.

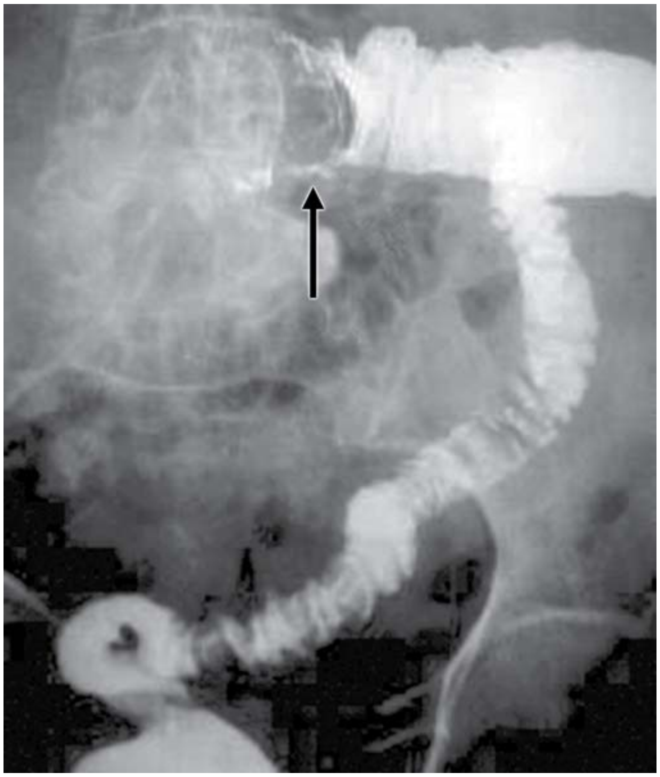

Fig. 23.28: Barium enema X-ray showing claw sign in intussusception.

Indications for surgery in ISS

- ISS more than 48 hours
- Features of perforation, strangulation, peritonitis
- Recurrent ISS
- In adult commonly resection is required

Surgery

- *Cope's method:* If reduction does not occur, *laparotomy* is done under G/A. By gently *milking out* the intussusception with warm packs, it is reduced. After reduction, viability of the bowel is checked carefully. If manual reduction is not possible, it is understood that the bowel is likely to be gangrenous which requires resection and anastomosis. In case of viable bowel, often terminal ileum is anchored to the ascending colon and *Jackson veil band is* cut. Patient also requires nasogastric tube aspiration, IV fluids, antibiotics.
- Laparoscopic approach may be used to reduce the intussusception.
- If intussusception persists for more than 48 hours or intussusception in adult requires resection.
  - Ileocolic resection is sufficient.

Recurrence Rate

- In hydrostatic reduction—10%.
- In open manual reduction—2%.
- In resection—very less < 1%.

VOLVULUS

Definition

It is the *twist (rotation)* in the axis of the loop of the bowel either clockwise or anticlockwise.

- 15% of large bowel obstruction is due to volvulus.
- Sigmoid colon is the commonest site (*anticlockwise*)—65%.
- ***Caecal volvulus***
  - Caecum is the second common site (*clockwise*) (C for C)—30%.
  - It is common in females, present as intestinal obstruction.
  - *Caecal bascule* is the presence of constricting band across the ascending colon (*Bascule*—French—see-saw and balance).
  - Caecum will be markedly distended and found in the centre of the abdomen.
  - It is due to lack of fixation of the caecum—*mobile caecum*.
  - Occasionally it is associated with malrotation.
  - Caecal volvulus is the commonest cause of large bowel obstruction in pregnancy.
  - X-ray shows *round gas shadow* in right iliac region. CT scan is very useful. Barium enema is also helpful.
  - *Resection and anastomosis (surgery)* is the only treatment.
- Volvulus of small intestine (midgut), volvulus neonatorum, gastric volvulus are other volvulus which can occur.

SIGMOID VOLVULUS (Volvulus of Pelvic Colon)

- It is common in Asia, common in India (7% of intestinal obstruction) and especially South India because of high fibre diet.
- It is very common cause of large bowel obstruction in Peru and Bolivia due to high altitude.
- More common in males and old age.
- It is common in patients with chronic constipation with laxative abuse.

It is common in:

- Ogilvie's syndrome
- Mentally retarded individuals
- Chaga's disease
- Hypothyroidism
- Anticholinergic drugs
- Multiple sclerosis
- Scleroderma
- Parkinson's disease

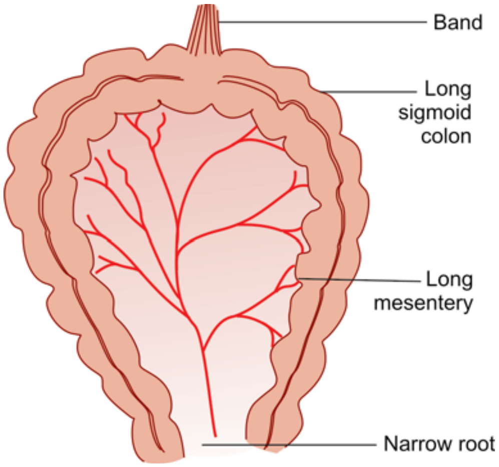

Fig. 23.29: Causes of sigmoid volvulus.

Predisposing factors

- Adhesions
- Peridiverticulitis
- Overloaded redundant pelvic colon
- Long pelvic mesocolon
- Narrow attachment of sigmoid mesocolon

- Always rotation is *anticlockwise.*
- It requires *one and half turn* of rotation to cause vascular obstruction and gangrene which eventually leads into perforation either at the root or at the summit of the sigmoid loop.

- Enormous distension of the colon occurs. Sometimes ileum comes to the root of the sigmoid volvulus and encircles it causing ***compound volvulus*** in which case knotted small bowel also becomes gangrenous—***ileosigmoid knotting.***

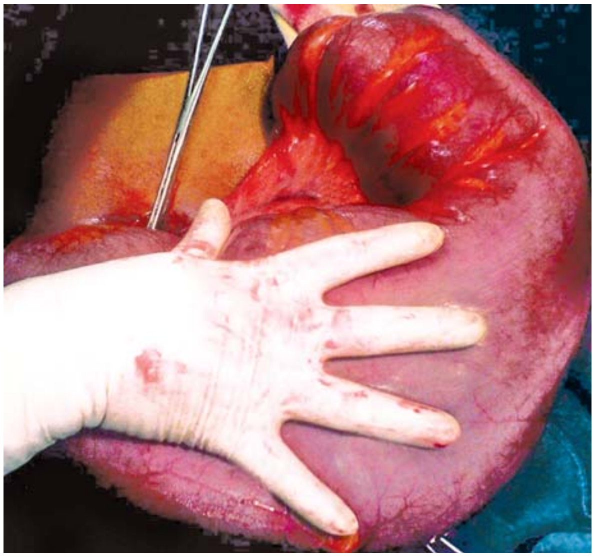

Fig. 23.30: Sigmoid volvulus. Note the dilated sigmoid colon.

Types

1. Acute.
2. Recurrent.

Clinical Features

- Pain in the abdomen—initially left sided, eventually all over.
- Absolute constipation *(obstipation*—no faeces, no flatus).
- Enormous distension of abdomen, starting from left iliac fossa extending to the whole of the abdomen (Tympanic abdomen).
- Late vomiting and eventually dehydration.
- Features of peritonitis.
- Hiccough and retching can occur.
- *Tyre like feel* of the abdomen is diagnostic.

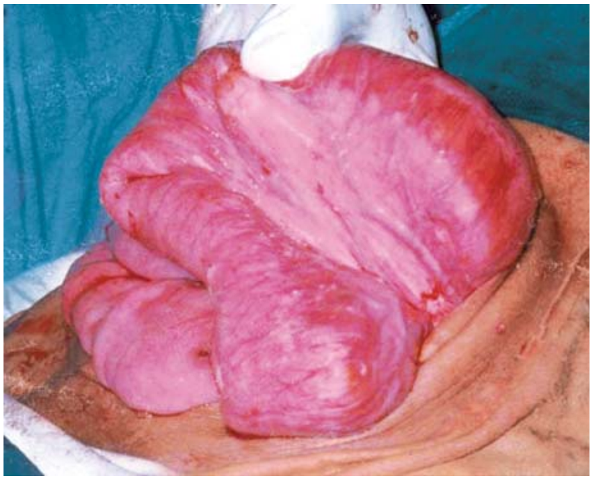

Fig. 23.31: Dilated sigmoid in early volvulus.

Differential Diagnosis

- *Ogilvies's syndrome*: It is acute colonic pseudo obstruction. It is due to malfunctioning sacral parasympathetic nerves. Splenic flexure is the junction of collapsed and dilated large bowel. Descending colon is atonic causing acute functional obstruction. It may be due to trauma, retroperitoneal irritation, antidepressants, uraemia, diabetes, myxoedema, hypokalaemia, etc. Prokinetic drugs, colonoscopic decompression, tube caecostomy are the treatment.
- Faecal impaction.
- Carcinoma rectosigmoid region.
- Idiopathic megacolon.

Investigations

1. ***Plain X-ray***:(diagnostic in 70-80%)
    - Ω *sign (omega sign)*—single, grossly distended loop of colon arising out of the pelvis and extending towards the diaphragm.
    - *Coffee-bean sign* or *Bent-inner tube sign*.
2. ***Contrast enema:*** (dilute barium/water soluble contrast media is used):
   Birds beak sign (ace of spades appearance)—Upper end of barium column tapers into the spirally twisted distal sigmoid colon.
3. ***CT scan*** (for difficult cases):
   Shows characteristic whirl pattern.
4. **Blood**—Haematocrit, renal function tests, serum electrolytes.

Treatment

- RT aspiration.
- IV fluids.
- Catheterisation.
- Antibiotics.
- *A flatus tube or Sigmoidoscope* is passed in operation theatre (proper care is taken otherwise it leads to perforation). If it derotates, patient will pass flatus and faeces; and distension reduces.
- If derotation does not occur, laparotomy through midline incision should be done. Dilated sigmoid colon is derotated manually and checked for viability. If viable, it can be fixed to the lateral wall of abdomen or pelvis—***sigmoidopexy***.
- If sigmoid colon is gangrenous, it is resected and proximal cut end is brought out as colostomy and distal end is brought out as mucous fistula, from the rectum *(Paul-Mikulicz Operation*). Later after 6 weeks, continuity is maintained after proper preparation.
- Resection of the gangrenous sigmoid done; proximal cut is brought out as end colostomy: distal end closed—*Hartmann's operation*. Later in 6-12 weeks colorectal anastomosis is maintained.
- Primary resection and anastomosis is not advisable because of the contamination and anastomotic leak.

Mortality in sigmoid volvulus is due to *perforation, peritonitis and septicaemia*.

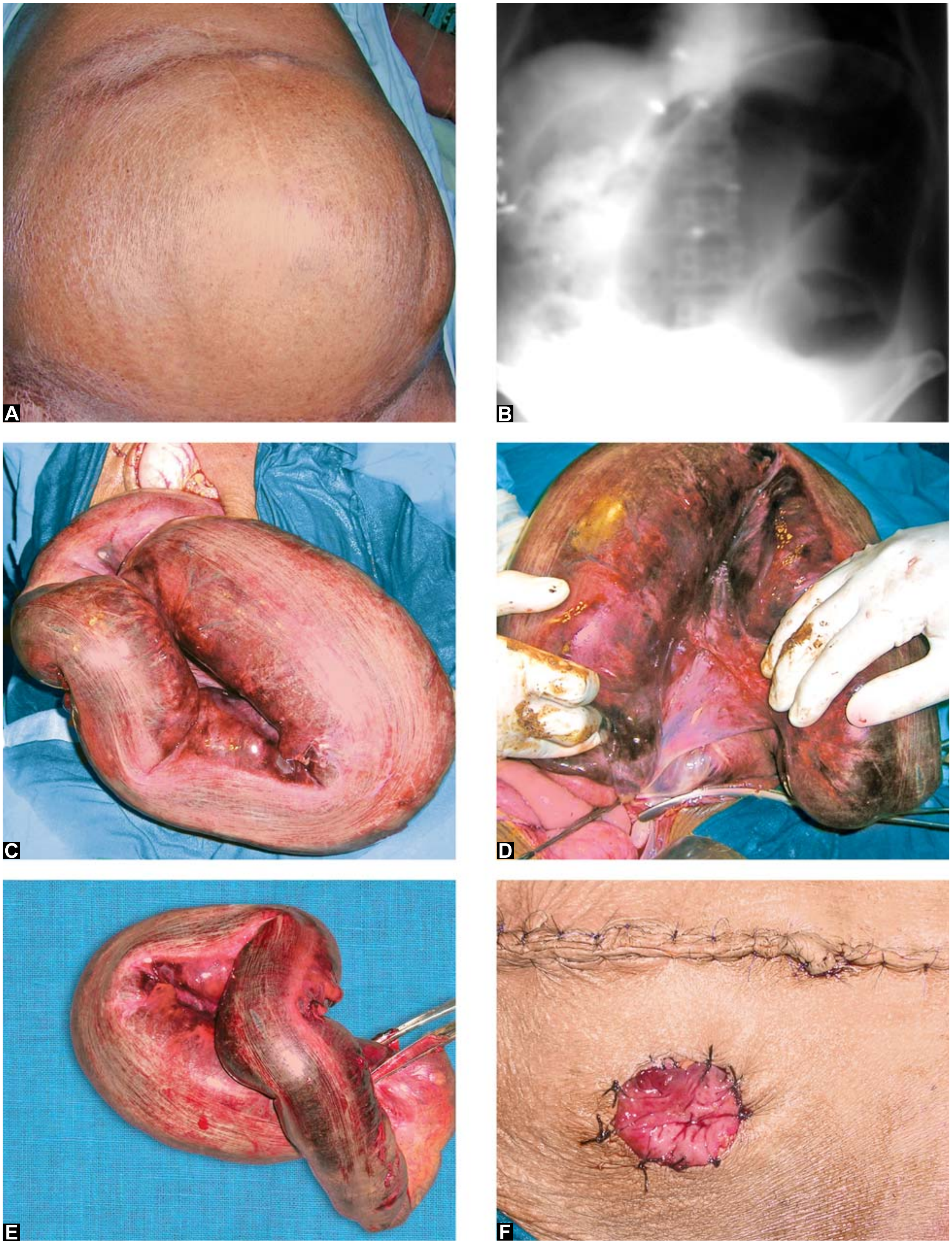

Figs 23.32A to F: Sigmoid volvulus—note distended tympanic abdomen; X-ray showing typical sigmoid volvulus; on table gangrenous sigmoid volvulus; underwent resection with colostomy and mucous fistula (exteriorisation).

A little may save a deal of friction.

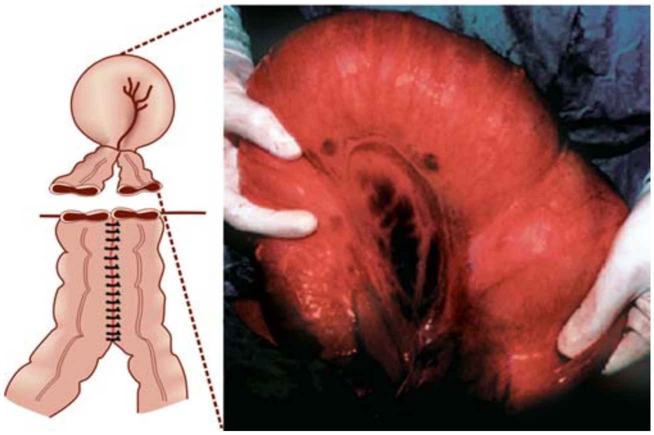

Fig. 23.33: Paul-Mikulicz operation.

PARALYTIC ILEUS (Adynamic Intestinal Obstruction)

- It is a state in which intestines fail to transmit peristalsis due to failure of neuromuscular mechanism, i.e. Auerbach's and Meissner's plexus.
- It may be localised or generalised.

Causes

- Postoperative
- Infective—pus, blood, bile, toxins, enteritis
- Uraemia
- Hypokalaemia
- Spinal injury
- Retroperitoneal haemorrhage
- Spinal surgery
- Plaster jacket

Clinical Features

- No passage of flatus.
- No bowel sounds.
- Marked abdominal distension.
- Vomiting of large volume of fluid.
- Tachycardia.
- Respiratory distress due to pressure over the diaphragm.
- High pitched tinkling note '*like bells at evening pealing*'.
- Dull abdominal pain (not colicky).
- Features of fluid/protein/electrolyte imbalance.

Investigations

- Serum electrolyte estimation: Especially serum potassium.
- ECG.
- X-ray abdomen.
- Ultrasound abdomen to find out the possible cause of ileus, e.g. sepsis.

Treatment

- Nasogastric aspiration.
- The primary cause is treated.
- IV fluids.
- Electrolyte management.
- Catheterisation and urine output measurement.
- *Do not stimulate the peristalsis ("Don't flog a tired horse").*
    - **Measurement of abdominal girth** is necessary to see whether patient is recovering or not.
    - Decompression of the large bowel can be tried by inserting a flatus tube per anally into the rectum and keeping in place for few hours.
    - Most often, patient recovers in 3-6 days by conservative treatment.
    - In prolonged, life threatening paralytic ileus, laparotomy is done and bowel is decompressed with *Savage's decompressor* and closed with tension sutures.
- *Catchpole regime*: Adrenergic blocking agent along with cholinergic stimulant (neostigmine) is used rarely only in resistant cases as medical therapy.

ADHESIONS AND BANDS

Adhesions and bands are the most common causes of intestinal obstruction in Western countries. In India, hernia and then adhesions are the two common causes of intestinal obstruction.

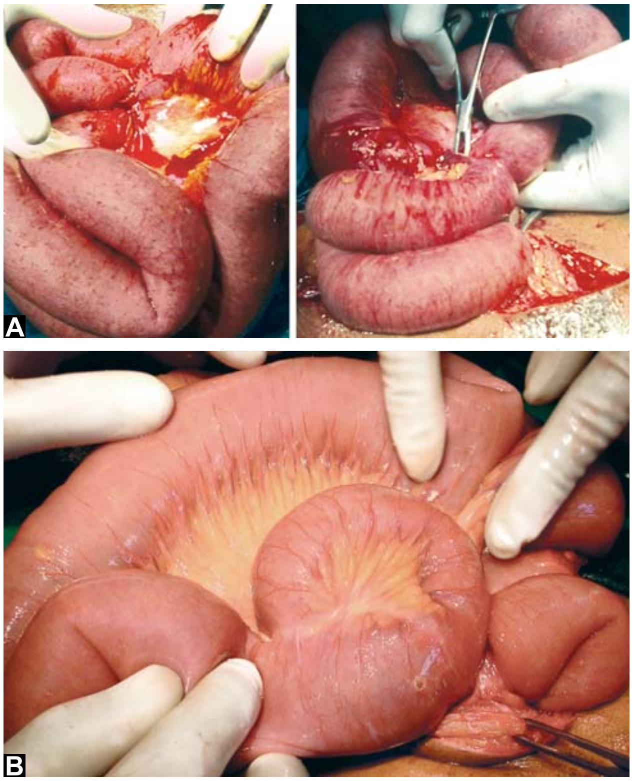

Figs 23.34A and B: Adhesions causing intestinal obstruction.

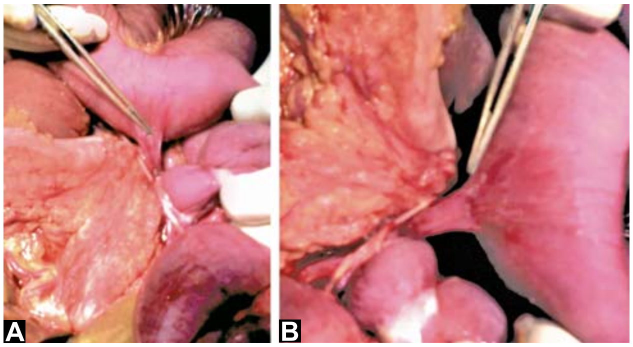

Figs 23.35A and B: Congenital band causing intestinal obstruction. These bands are often vascular.

Causes

- Infection due to peritonitis, appendicitis, postlaparotomy, and other acute infective abdominal conditions.
- Materials used during surgery can cause dense inflammatory reactions—suture materials like silk, thread, and foreign body, mop, and gauze, talc powder, drugs like sulphonamides and penicillins.
- Ischaemia of bowel due to poor blood supply, sepsis.
- Gynaecological conditions, bowel injury, radiation-induced enteritis, Crohn's disease, other inflammatory bowel diseases.
- Specific conditions like tuberculosis, malignancy.

Note:

- It is often called as *adhesive small bowel obstruction* (ASBO). 80% cases showed earlier laparotomies; 35% multiple laparotomy. 75% of surgeries in such patients are emergency one.
- *Causes may be classified as—Congenital* adhesions; Ischaemic—Reduced blood supply of the bowel, *Traumatic*—Patients have undergone laparotomies for intraabdominal injuries; *Irritants*—Talc powder used for gloves, starch, antibiotics, nonabsorbable suture materials, gauze and cotton filaments, barium; *Inflammatory:* Appendicitis, cholecystitis, pelvic infection, Crohn's disease, tuberculosis; *Others*—Serosal denudation during surgery precipitates adhesions.
- Initial cause ischaemia/injury → inflammatory exudates → fibrinous adhesions → fibrous adhesion → fibrinolysis. Fibrinous adhesion is avascular. Often thick vascular adhesions develop. Adhesions can be localized or extensive.

Types

Type I—Fibrinous adhesions occur during 5-10th post-surgical period. It usually gets resolved completely. It is avascular and flimsy.

Type II—Fibrous adhesions. Due to lack/poor blood supply, bowel gets attached to part of peritoneum or omentum or other parts of the bowel with dense vascular adhesions to maintain blood supply. It will persist and precipitate intestinal obstruction, often-subacute and recurrent type.

Adhesions due to tuberculosis are severe, dense and difficult to separate.

Clinical features

- Pain abdomen—colicky type and recurrent and episodic
- Distension, vomiting
- Constipation
- Reduced bowel sounds on auscultation
- Previous surgical scars commonly observed
- Dehydration, tachycardia, hypotension
- Commonest symptom in adhesions is ***pain*** which is sharp, sudden, recurrent, episodic and colicky. Colicky pain is more common in midlevel obstruction. *Gilroy Bevan triad* of adhesive pain is—(1) Pain may get aggravated or relieved on change of posture. (2) Pain in the region of old abdominal scar. (3) Tenderness is elicited by pressure over the scar
- Lower adhesions will have more distension compared to proximal adhesion
- Presents with features of obstruction initially, later features of strangulation (toxicity, fever, tachycardia, guarding, rigidity, rebound tenderness)
- Adhesions may be asymptomatic for many years but later become symptomatic

Investigations

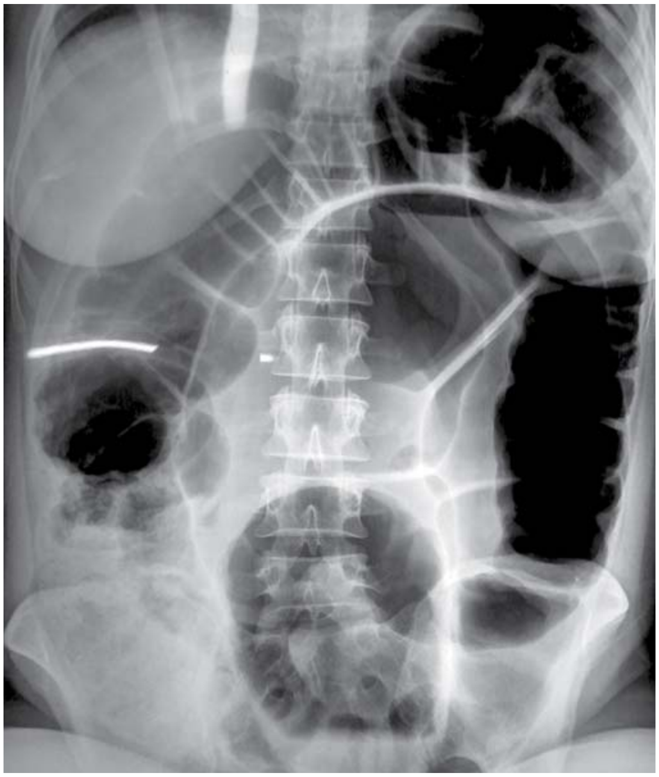

Fig. 23.36: X-ray showing intestinal obstruction with dilated colon—colonic obstruction.

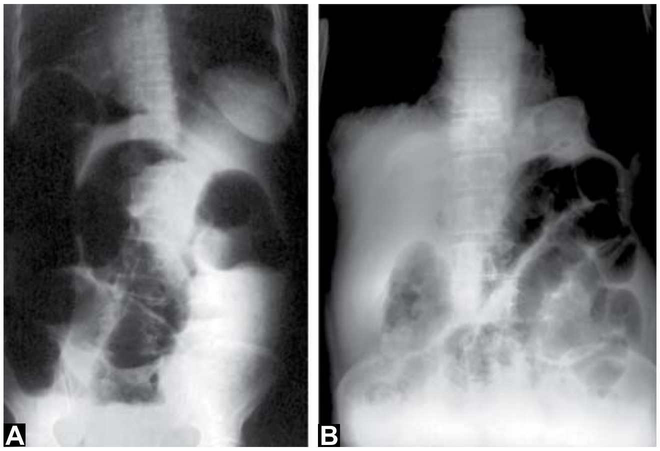

Figs 23.37A and B: X-ray showing small bowel obstruction with multiple air fluid levels.

- *Plain X-ray abdomen shows dilated bowel loops.*
- *Oral contrast with CT scan is very useful.*
- *In partial obstruction only enteroclysis with water-soluble iodine dyes can be done.*

Level of obstruction should be assessed probably by X-ray and CT scan.

- CT may show radiological features of strangulation -
  - Intramesenteric fluid.
  - Mesenteric oedema, congestion.
  - Bowel wall thickening more than 2 mm.

  - Reduced mural enhancement in strangulated bowel compared to adjacent bowel.
  - Small bowel loop may go for volvulus showing *Whirl sign* in CT scan with whirlpool mesentery of that part of the bowel.
- Electrolyte estimation, total count, serum creatinine, blood grouping.

Treatment

Initial Treatment

- IV fluids.
- Nasogastric aspiration.
- Antibiotics.
- Often per anal flatus tube insertion.
- Commonly patient responds in 4-5 days of therapy.

Later Treatment

- Open surgical adhesiolysis using fingers gently.
- *Laparoscopic adhesiolysis* is becoming popular, safer, ideal with less recurrent adhesion rate and gives good results.

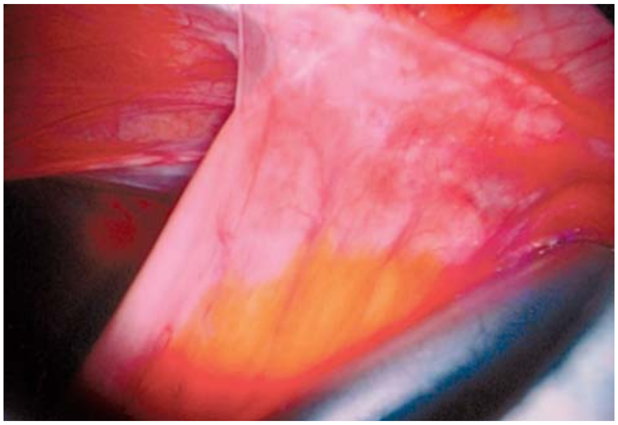

Fig. 23.38: Laparoscopic view of dense adhesions—postoperative cause.

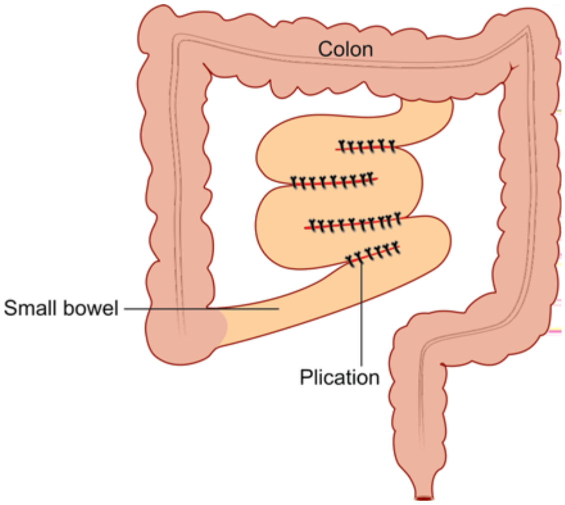

Fig. 23.39: Methods used to prevent recurrent adhesions and so to reduce the chances of intestinal obstruction. *Nobles plication* is plication of intestines.

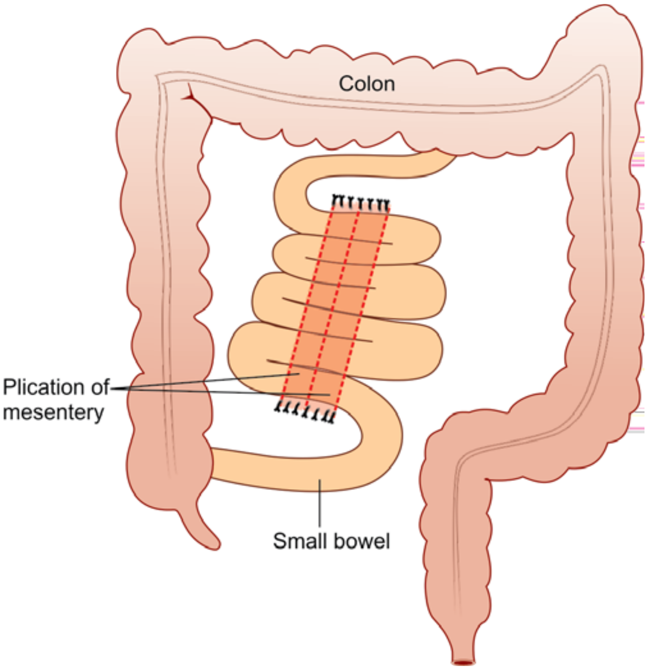

Fig. 23.40: *Childs-Phillips operation* done to prevent adhesive obstruction by plicating mesentery.

- *Noble's plication* of intestines—wherein adjacent coils of small bowel are sutured to prevent in going for further recurrent adhesions.
- *Childs-Phillips mesenteric plication* is similar way of plication of intestinal mesentery, which will prevent crumpling of bowel and adhesion formation.
- Passing long stiff tube from exterior through an opening in the proximal bowel up to the entry into the large bowel, which acts as intraluminal splinting.
  - Intestinal intubation is technically difficult with chances of perforation and fistula formation.
  - *Baker's tube* insertion through *jejunostomy* site; *Stewardson insertion of Baker tube* through nasogastric route; In *Nelson and Nyhus* method, 2nd balloon to Baker tube 25 cm proximal to the first to facilitate easier insertion of the tube—are different tubes and techniques used.

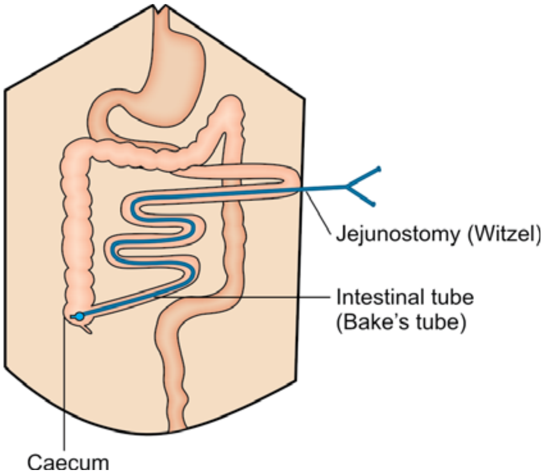

Fig. 23.41: Intestinal intubation via a jejunostomy tube used for adhesions.

- Expandable PTFE membrane barriers or hyaluronic acid and carboxymethyl cellulose membrane barriers are also used.

Prevention of adhesions

- Instilling drugs like hyaluronidase, steroids, dextrans, dextrins, polyvinyls, silicone, fibrinolysins, chondroitins, streptomycins
- Gentle handling of the bowel
- Prevention of spillage of contents like bile, faeces
- Ideal peritoneal wash using 8-10 liters of saline before closing the abdomen
- Careful placing of the drains
- *Laparoscopic procedures has got lesser chance of getting adhesions than open method*

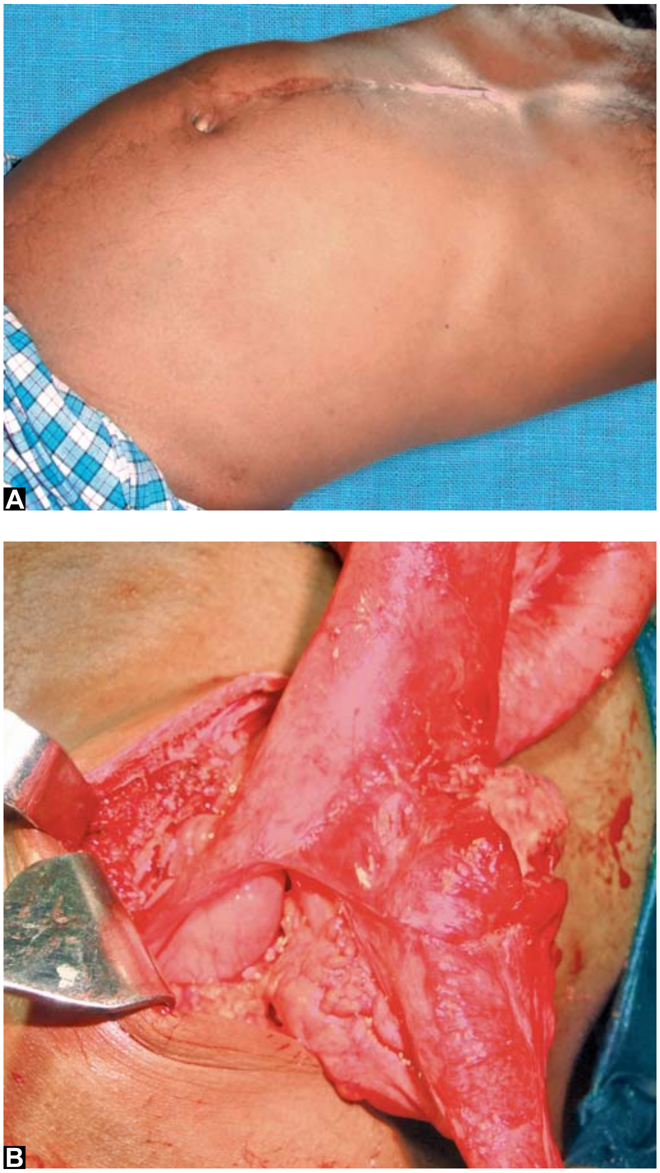

Figs 23.42A and B

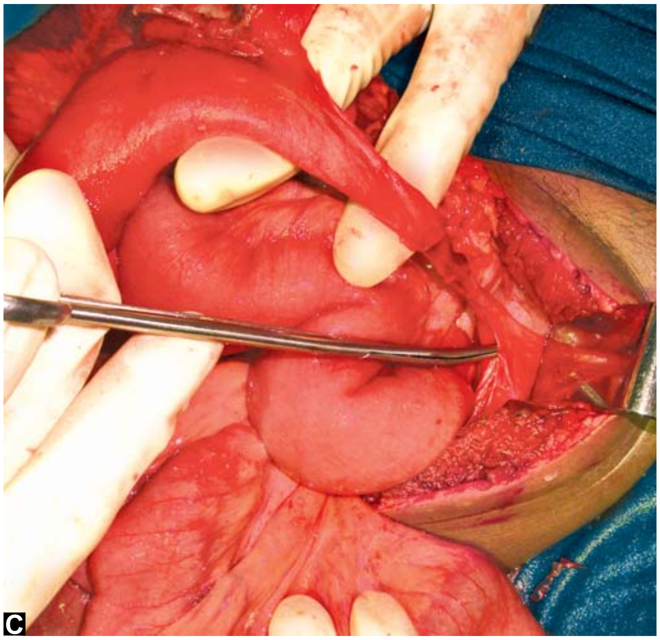

Fig. 23.42C

Figs 23.42A to C: Postoperative adhesions have become commonest cause of intestinal obstruction. Note the abdominal distension in a previous laparotomy patient and on table adhesions.

Obstruction due to Bands

- These are *dense fibrous strings* attached from one portion of the abdomen to another area or bowel causing entrapment of intestines leading into obstruction and often dangerous *strangulation.*
- Common causes are vitellointestinal duct, Ladd's band, omental band, postsurgical fibrous band, tuberculous band.
- Clinical features are like of intestinal obstruction.
- Management is release of the band either through laparoscopy or through laparotomy.

INTERNAL HERNIAS

- It is rare entity.
- A portion of small bowel gets entrapped in one of the retroperitoneal fossae or congenital internal defects (e.g. mesenteric defect).

Sites

- Mesenteric defect; defect in transverse mesocolon.
- Foramen Winslow.
- Diaphragmatic hernia.
- Hernia through duodenal/caecal/appendiceal/ intersigmoid retroperitoneal fossae.
- Hernia through broad ligament.

Mesocolic/paraduodenal hernias: They are rare congenital small bowel herniation behind the mesocolon due to midgut malrotation. It can be right or left types. In right type small bowel is right of the superior mesenteric artery (SMA is medial to neck of the sac) behind right side colonic mesentery with ileocolic, right colic, middle colic vessels in front in the sac. Left type is commonest (75%) with herniation of small bowel between inferior mesenteric vessels and attachment of descending mesocolon. Presentation is acute intestinal obstruction. Contrast CT is diagnostic. Often condition is diagnosed on table during surgery. Incising peritoneum sac adjacent to vessels carefully and adequately is the needed surgical treatment. Malrotation should be surgically corrected.

Mesenteric defect hernias (small bowel mesentery/ mesocolon) even though can occur anywhere, are common near ileocolic junction.

Acquired internal hernias occur through defects developed after surgeries like gastrojejunostomy (*Stammer's* hernia); colostomy; ileostomy; surgeries for morbid obesity.

Features

- Herniation occurs in these sites whenever there are adhesions.
- Features of intestinal obstruction—acute/subacute/recurrent.
- Diagnosis is usually on table. But enteroclysis, contrast CT scan are useful.

Treatment

- Release of adhesion, reduction of hernia, if bowel not viable resection and anastomosis.
- Often major vessels may be present adjacent to the constriction ring which may get injured while releasing causing torrential haemorrhage. Decompression is first done prior to reduction.

Chapter 24 Appendix

CHAPTER OUTLINE

- Surgical Anatomy
- Duplication of Appendix
- Acute Appendicitis
- Incidental Appendicectomy
- Appendicular Mass
- Appendicular Abscess
- Faecal Fistula after Appendicectomy
- Mucocele of Appendix
- Neoplasms of the Appendix
- Laparoscopic Appendicectomy

SURGICAL ANATOMY

It is located at the terminal end of the caecum where three taeniae join, about 2 cm below the ileocaecal orifice. Usually, around 5-10 cm in size but can be varible. Size of its lumen is that of matchstick. Diameter of appendix is 3-8 mm; diameter of lumen is 1-3 mm (matchstick).

Mesoappendix is extension of the mesentery contains appendicular artery, a branch of ileocolic artery. Often an accessory appendicular artery *(of Seshachalam*) may be present. Thrombosis of these vessels leads to gangrenous appendicitis.

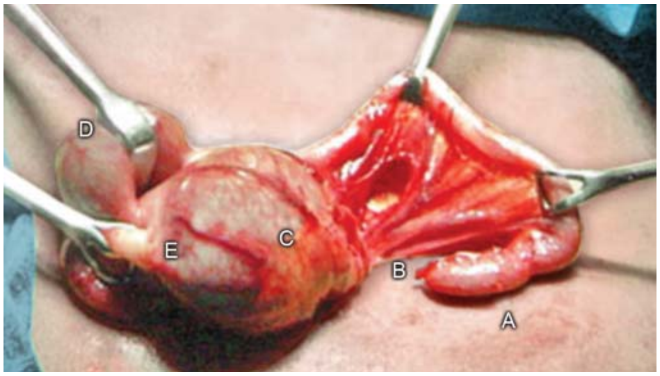

Fig. 24.1: On table demonstration of mesoappendix with appendicular artery. Note the appendix (A); mesoappendix (B); caecum (C); ileum (D); taenia (E).

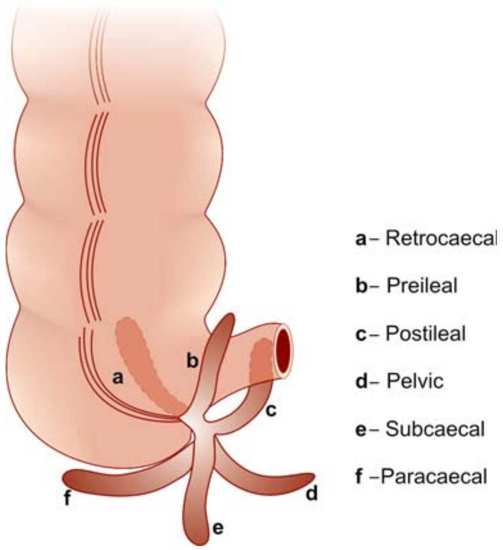

Fig. 24.2: Different anatomical positions of the appendix.

DUPLICATION OF APPENDIX (Wallbridge Classification) (Fig. 24.3)

Type A: Partial duplication in a single caecum.
Type B: Two separate appendices in a single caecum.
Type C: Double caecum with each one having one appendix.

- **Most common position** *is retrocaecal (75%). Next common is pelvic (21%).*

Charm strikes the sight, but merit wins the soul.

Other sites are:
- Preileal—rarest (1%)
- Postileal
- Paracaecal
- Subcaecal
- Subhepatic

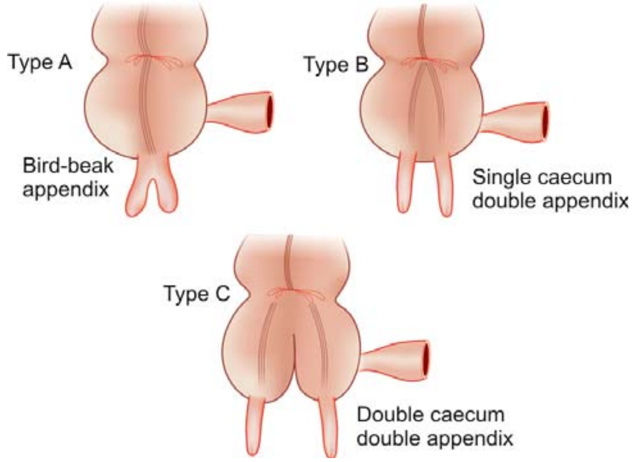

Fig. 24.3: Duplication of appendix.

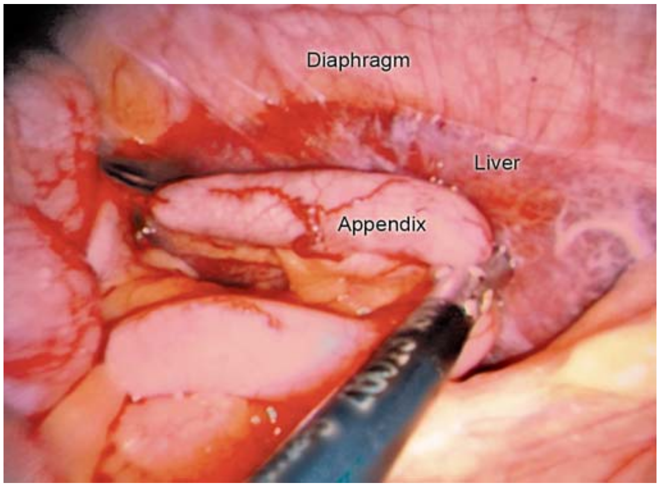

Fig. 24.4: Laparoscopic view of subhepatic appendix.

- Often it may be in subserosal plane leading to difficult dissection during appendicectomy.
- In *situs inversus* appendix is on *the left side.*
- Mucosa of appendix is lined by columnar cells with crypts.
- Submucosa contains numerous lymphatic aggregations (follicles*) ('abdominal Tonsil').*
- Opening of the appendix into caecum is guarded by *'valve of Geralch'*.

ACUTE APPENDICITIS

> *The partial inflammation of the peritoneum, in the liac fossa, is sometimes set up by disease in the Appendix caeci…. The appendix having been perforated by ulcerations, occasioned by the lodgement of the faecal concretions in its cavity, extravasation takes place, and inflammation of a more severe and serious kind is originated…. Nature sometimes succeeds in limiting the inflammation to a part of the right side; but it is at other times diffused over the whole abdomen …and quickly proves fatal.*
>
> —**Thomas Hodgkin,** 1836

Aetiology

- It is common in young males.
- It is common in white races.
- Fibre rich diet prevents appendicitis. Less fibre diet increases chance of appendicitis.
- It is common in May and August—seasonal variation—often called as *epidemic appendicitis*.
- Viral infection may cause mucosal oedema and inflammation which later gets infected by bacteria causing appendicitis.
- Family history may be relevant in 30% of appendicitis in children with appendicitis occurring in first degree relatives.
- Obstruction of the lumen of appendix causing obstructive appendicitis.
  - Blockage occurs due to—faecoliths, stricture, foreign body, round worm or threadworm.
  - Adhesions and kinking—carcinoma caecum near the base, ileocaecal Crohn's disease.
- Distal colonic obstruction.
- Abuse of purgatives.
- Faecolith is the *most common* cause.

Organisms:

E. coli (85%), enterococci, (30%), streptococci, *Anaerobic streptococci, Cl. welchii,* bacteroides.

Pseudoappendicitis is appendicitis due to acute ileitis following Yersinia infection. It is often due to Crohn's disease.

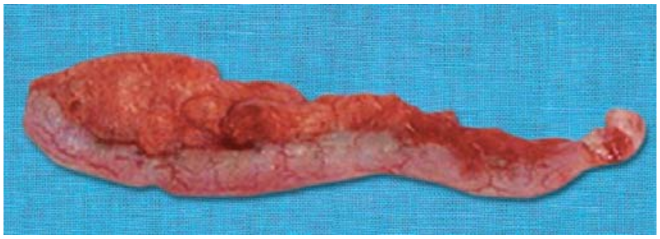

Fig. 24.5: Removed appendix with its mesoappendix.

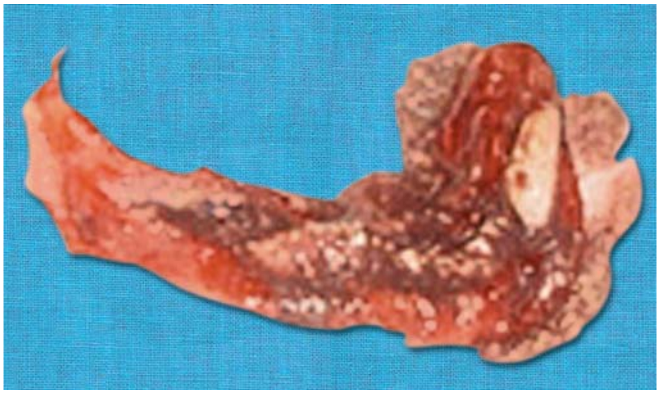

Fig. 24.6: Severely infected appendix with faecolith inside. Faecolith is one of the causes of obstructive appendicitis.

Pathogenesis

- Acute inflammation of the mucus membrane with secondary infection without obstruction causes *acute nonobstructive appendicitis*. It may lead into resolution, fibrosis, recurrent appendicitis or eventual obstructive appendicitis.

- Luminal obstruction by faecolith, lymphoid hyperplasia, pinworm (oxyuris vermicularis), other worms, foreign body, carcinoma/Crohn's disease → mucus and inflammatory fluid collects inside the lumen → increases intraluminal pressure → leads to blockage of lymphatic and venous drainage → resulting in increased oedema of mucosa and wall → causes mucosal ulceration and ischaemia → bacterial translocation → bacterial spread through submucosa and muscularis propria → *acute obstructive appendicitis* → thrombosis of appendicular artery → ischaemic necrosis of full thickness of the wall of the appendix → gangrene of the appendix → perforation at the tip or at the base → peritonitis.

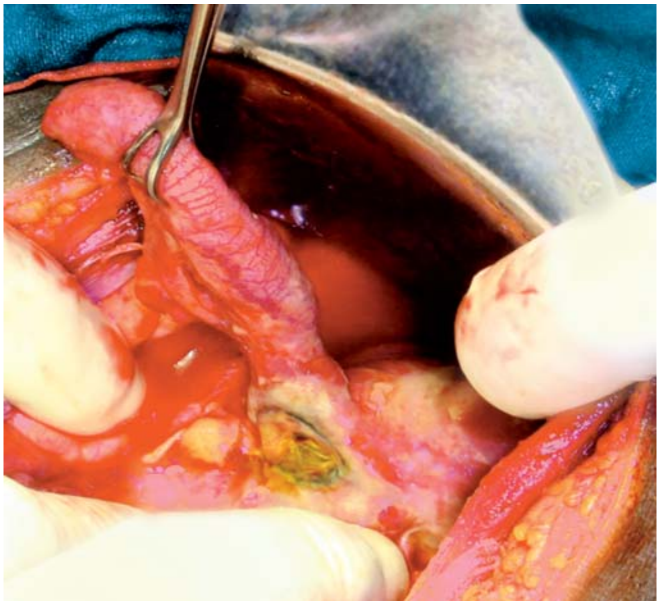

Fig. 24.7: Laparotomy showing perforated appendix with peritonitis.

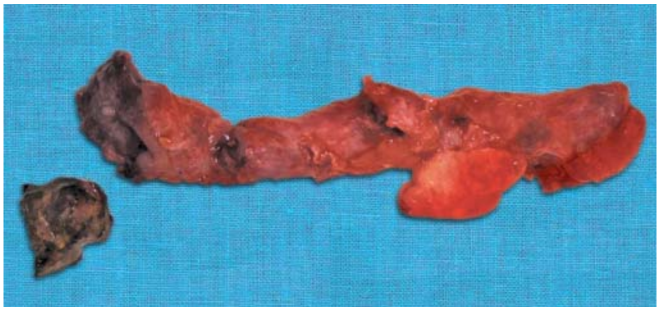

Fig. 24.8: Appendicolith with appendicitis.

- After perforation → localisation by greater omentum and dilated ileum occurs → with suppuration and pus inside forming *appendicular abscess*.
- In severe acute appendicitis → localisation can occur by omentum and dilated ileum without pus inside → forming *appendicular mass*.
- Acute appendicitis with blockage at the opening of the lumen → inflammation rarely subsides → mucus collects inside the lumen of the appendix resulting in its enlargement → *Mucocele of the appendix*.

Types

1. ***Acute nonobstructive appendicitis (catarrhal):***

   Inflammation of mucous membrane occurs with redness, oedema and haemorrhages which may go for following courses:
   - Resolution.
   - Ulceration.
   - Fibrosis.
   - Suppuration.
   - Recurrent appendicitis.
   - Gangrene—rare initially in nonobstructive type but later can occur.
   - Peritonitis.

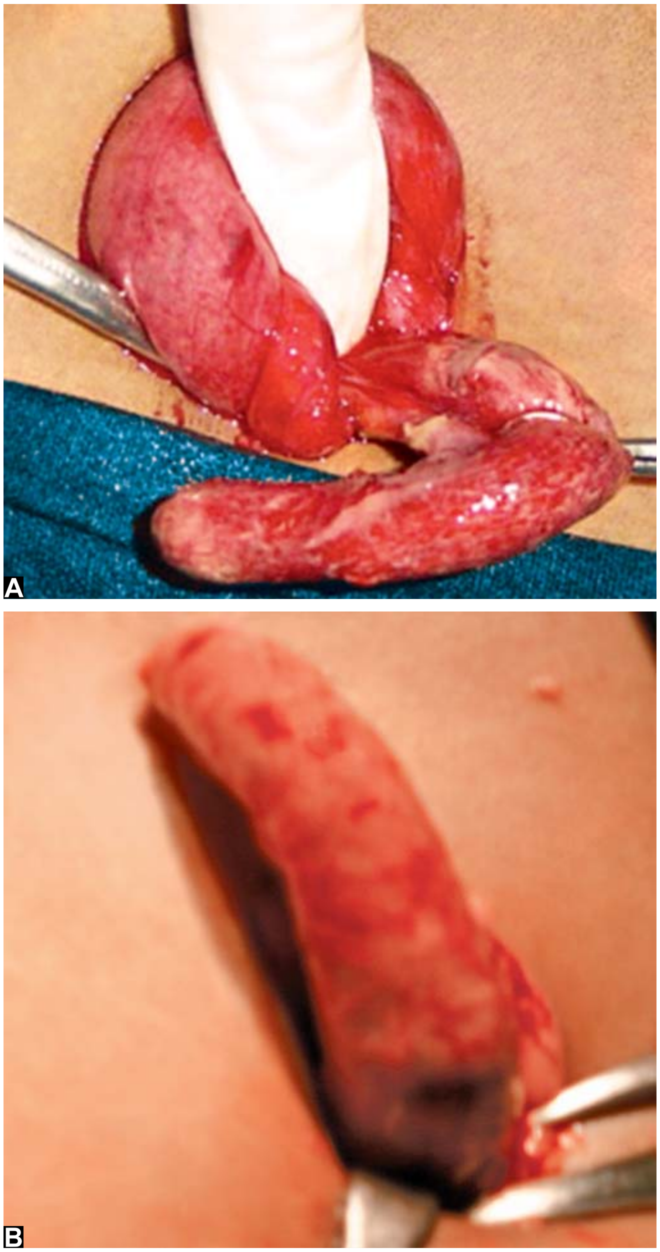

Figs 24.9A and B: Gangrenous obstructive appendicitis—on table.

Appendices epiploicae are absent in rectum, appendix and caecum.

2. ***Acute obstructive appendicitis*:** Here pus collects in the blocked lumen of appendix which is blackish, gangrenous, oedematous and rapidly progresses leading to perforation either at the tip or at the base of appendix. This leads to peritonitis, formation of appendicular abscess or pelvic abscess. Most often, there will be thrombosis of the appendicular artery.
3. ***Recurrent appendicitis:*** Repeated attacks of nonobstructive appendicitis leads to fibrosis, adhesions causing recurrent appendicitis.
4. ***Subacute appendicitis*** is milder form of acute appendicitis.
5. ***Stump appendicitis*** is retained long stump of appendix after commonly laparoscopic appendicitomy.

Clinical Features

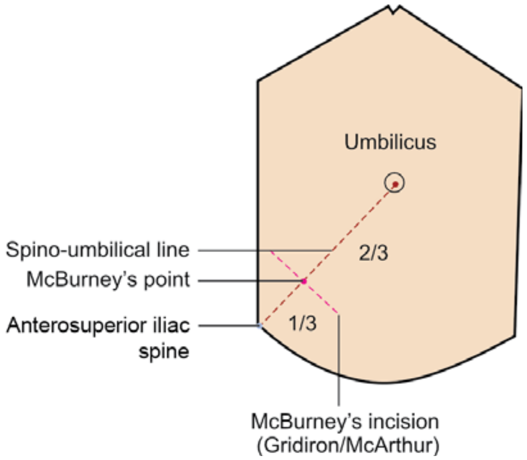
Umbilicus
Spino-umbilical line
McBurney's point
2/3
Anterosuperior iliac spine
1/3
McBurney's incision (Gridiron/McArthur)

Fig. 24.10: *McBurney's point* is junction of lateral 1/3rd and medial 2/3rd of spinoumbilical line. McBurney's/Gridiron's/McArthur's incision is perpendicular to this line in this point.

- It is rare before the age of two, common in children and other age groups.
- *Pain*: It is the earliest symptom. Visceral pain starts around the umbilicus due to distension of appendix, later after few hours, somatic pain occurs in right iliac fossa due to irritation of parietal peritoneum due to inflamed appendix. Pain eventually becomes severe and diffuse which signifies spread of infection into the general peritoneal cavity.
- *Vomiting:* Due to reflex pylorospasm.

Murphy's triad

- Pain
- Vomiting
- Temperature

- *Constipation* is the usual feature *but diarrhoea* can occur if appendix is in postileal or pelvic positions.
- *Fever, tachycardia, foetor oris* are other features.
- *Urinary frequency:* Inflamed appendix may come in contact with bladder and can cause bladder irritation.
- Tenderness and rebound tenderness at *McBurney's point* in right iliac fossa *(release sign—Blumberg's sign)* are typical.
- *Rovsing's sign:* On pressing left iliac fossa, pain occurs in right iliac fossa which is due to shift of bowel loops which irritates the parietal peritoneum.
- Hyperextension (in case of retrocaecal appendix—*Cope's psoas test*) or internal rotation (in case of pelvic appendix—*obturator test*) of right hip causes pain in right iliac fossa due to irritation of psoas muscle and obturator internus muscle respectively.
- *Baldwing's test* is positive in retrocaecal appendix—when legs are lifted off the bed with knee extended, the patient complains of pain while pressing over the flanks.
- P/R examination shows tenderness in right side of the rectum.
- Hyperaesthesia in *'Sherren's triangle'*. This triangle is formed by anterosuperior iliac spine, umbilicus, pubic symphysis.

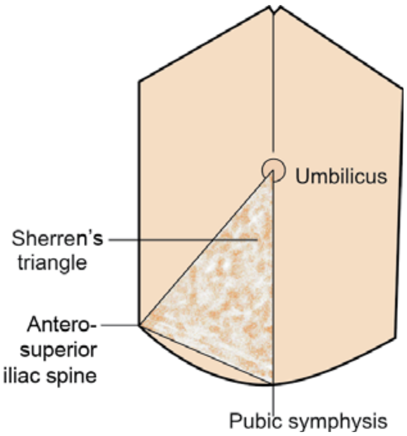

Fig. 24.11: *Sherren's triangle*. It is area of hyperaesthesia found in acute appendicitis between anterosuperior iliac spine; umbilicus and pubic symphysis.

Often infection gets localised by omentum, dilated ileum and parietal peritoneum leading to *appendicular mass.* Often suppuration occurs in the localised area resulting in *appendicular abscess.*

Clinical signs in appendicitis

- Rovsing's sign
- Blumberg's sign (Release sign)
- Cope's psoas test
- Obturator test
- Baldwing's test
- Bastede sign
- *Dumphy's* cough tenderness sign (Refer fascinating signs for detail)
- Bapat bed shaking test
- Heel Drop test

- ***Acute appendicitis in infancy:***
  - Eventhough it is rare, when it occurs, it has got 80% chances of perforation with high mortality (50%).
- ***Acute appendicitis in children:***
  - Here localisation is not present, and so peritonitis occurs early.
  - It *requires early surgery.* Dehydration, septicaemia are common.
- ***In elderly:***
  - Gangrene and perforation are common. Because of lax abdominal wall, localisation is poor and so peritonitis sets in early.

- ***In pregnancy:***
    - *Incidence* is 1 in 2,000 pregnancies. It is more common in 1st and 2nd trimesters.
    - Appendix shifts to upper abdomen. So pain is higher and more lateral.
    - Rebound tenderness and guarding *may not* be evident.
    - TC will be very high with neutrophilia.
    - Risk of *premature labour* is 15%.
    - *Foetal death* in early appendicitis is 5% but becomes 29% once appendix perforates in pregnancy.
    - After 6 months, maternal mortality increases by 10 times than usual and also leads to premature labour.
    - Appendicitis is the most common non-gynaecologic surgical emergency during pregnancy.
    - Incidence of perforation is highest in 3rd trimester.
    - Surgery is the treatment.

Differential Diagnosis for Acute Appendicitis

Many conditions mimic acute appendicitis. It differs in children, adult, elderly and females.

- *Perforated duodenal ulcer*: In duodenal ulcer perforation, fluid trickles down along right paracolic gutter and mimics appendicitis. Upper abdominal pain, obliterated liver dullness, gas under diaphragm in X-ray and CT scan differentiate it from acute appendicitis.
- *Acute cholecystitis*: Pain in right upper abdomen, fever, jaundice, upper abdominal guarding are the features of acute cholecystitis. US; HIDA scan, LFT will differentiate it from acute appendicitis.
- *Acute pancreatitis*: Pain in epigastrium, radiating to back, raised serum amylase and lipase, CT abdomen with a history of alcohol intake often are diagnostic.
- *Right ureteric colic*: Pain is colicky in nature which often refers to genitalia. Haematuria, urinary symptoms are common. It mimics retrocaecal/pelvic acute appendicitis. Often in ureteric stone, abdomen is soft and nontender. CT is the important way to differentiate.
- *Acute typhlitis*: Inflammation of caecum is called as typhlitis. Often it is difficult to differentiate it from acute appendicitis. Intravenous/oral metronidazole completely controls the disease.
- *Acute bacterial enterocolitis*: It presents with pain abdomen, diarrhoea, toxaemia, dehydration. Often it is difficult to differentiate from acute appendicitis.
- *Acute mesenteric lymphadenitis* is difficult to differentiate from acute appendicitis. It is treated conservatively. CT may be helpful to identify it. Laparoscopic evaluation is ideal.
- *Right sided acute pyelonephritis*: Here there will be pain and tenderness in loin. Urine analysis, US are diagnostic. Often DTPA scan may be needed.
- *Crohn's disease* presenting with acute symptoms will have similar features of acute appendicitis.
- *Pelvic inflammatory disease* like salpingo-oophoritis mimics acute appendicitis. Twisted/haemorrhagic/ruptured ovarian cyst/ruptured ectopic gestation/endometriosis/tubo-ovarian abscess mimics acute appendicitis. US, laparoscopy helps to differentiate it from others. *Mittelschmerz* is lower abdominal pain due to rupture of follicular cyst during midcycle. It subsides on its own. There are no systemic features.
- *Meckel's diverticulitis* presents clinically like acute appendicitis. It is not possible to differentiate between two clinically.
- *Intussusception* mimics acute appendicitis in children. ISS is common before the age of 2 years. Acute appendicitis is rare before the age of 2 years. Palpable mass, features of intestinal obstruction, barium enema X-ray, US are useful methods to differentiate.
- *Worm infestation (round worm bolus/ball):* It often presents as pain in right iliac fossa. Features of intestinal obstruction are common here.
- *Right-sided lobar pneumonia* and pleurisy are often not easy to differentiate from acute appendicitis. Pleural rub, change in breath sounds, chest X-ray can identify pneumonia.
- *Testicular torsion/acute severe orchitis* often look like acute appendicitis. Referred pain in iliac fossa, and if scrotum is not palpated clinically these conditions are mistaken for acute appendicitis. These problems are much more obvious if *testis is undescended* one.
- *Sigmoid diverticulitis* in elderly with loop lying towards right side may present as pain in the right iliac fossa.
- *Carcinoma caecum* may present with features of acute appendicitis without any earlier typical features.
- *Ruptured aortic aneurysm*, *acute intestinal obstruction, mesenteric ischaemia* may present like acute appendicitis.
- Rare conditions like preherpetic pain of the right 10th and 11th dorsal nerves may mimic acute appendicitis. Guarding and rigidity will not be present. There will be significant hyperaesthesia.
- Tabetic crisis, tuberculosis of spine, secondaries in spine, multiple myeloma, osteoporotic pain often can mimic acute appendicitis.
- Acute crisis of porphyria and diabetes mellitus mimic acute appendicitis with severe abdominal pain.

Differential diagnosis

- Perforated peptic ulcer
- Ruptured or twisted ovarian cyst
- Acute cholecystitis
- Right ureteric colic
- Enterocolitis
- Right acute pyelonephritis
- Mesenteric lymphadenitis
- Lobar pneumonia
- Crohn's disease
- Acute pancreatitis
- Meckel's diverticulitis
- Acute crisis of porphyria
- Salpingitis
- Diabetic abdomen
- Ectopic gestation—ruptured
- Typhlitis

Perforated appendix rarely causes pneumoperitoneum.

Differential diagnosis in children

- Meckel's diverticulitis
- Acute colitis
- Acute iliac lymphadenitis
- Intussusception
- Roundworm colic
- Lobar pneumonia

Differential diagnosis in females

- Ruptured ectopic gestation
- Mittelschmerz rupture of ovarian follicle during mid-menstrual period
- Ovarian cyst torsion
- Salpingo-oophoritis

Differential diagnosis in elderly

- Acute diverticulitis
- Carcinoma caecum—acute features
- Mesenteric ischaemia
- Intestinal obstruction
- Aortic aneurysm leak
- Crohn's disease

Sequelae of acute appendicitis

- Resorption
- Relapse and recurrent appendicitis
- Appendicular mass
- Appendicular abscess
- Perforation—has got 20% mortality
- Peritonitis, septicaemia
- Portal pyaemia
- *Intestinal obstruction* due to obstructive ileus, inflammatory adhesion, formation of band between appendix and omentum or between appendix and small bowel

Investigations

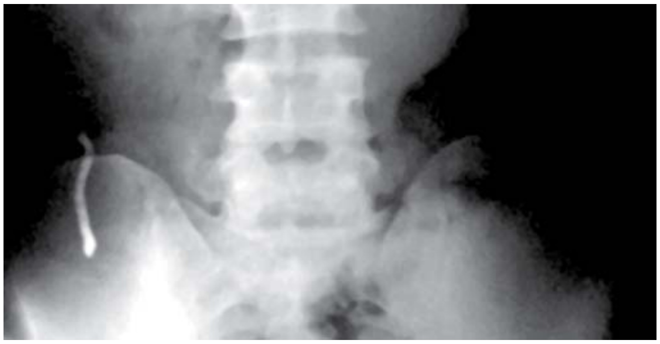

Fig. 24.12: Radio-opaque appendix in a plain X-ray. It could be calcified or have calcified content.

- Total leucocyte count is increased.
- *Ultrasound* is done to rule out other conditions like ureteric stone, pancreatitis, ovarian cyst, ectopic pregnancy and also to confirm appendicular mass or abscess.
- Laparoscopy is the most useful method.

Sonographic criterias for appendicitis (85% Specificity)

- Noncompressible appendix of size > 6 mm AP diameter, hyperechoic thickened appendix wall > 2 mm—*target sign*.
- Appendicolith.
- Interruption of submucosal continuity.
- Periappendicular fluid.

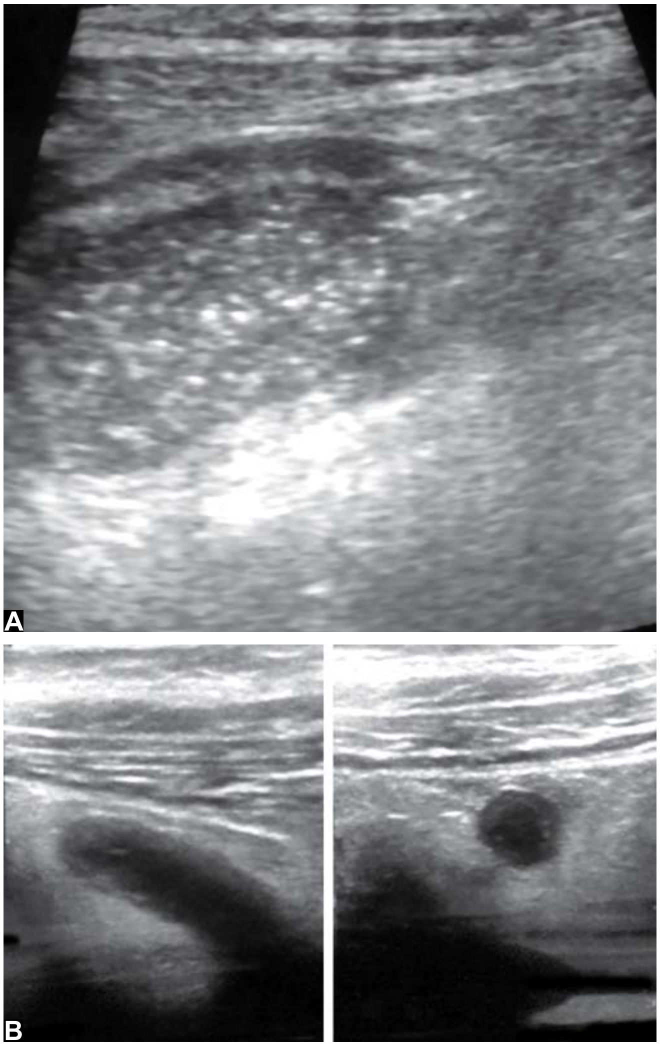

Figs 24.13A and B: US showing dilated noncompressible appendix in two different patients.

- *Contrast CT scan* is very much useful when diagnosis is difficult especially in old people. Dilated appendix; dilated lumen; thickened wall; nonfilling of the lumen by contrast or air; periappendicular fluid collection; presence of mass/abscess/associated pathology like carcinoma can be identified. It has 95% sensitivity and specificity with 95% accuracy. *Dirty fat thickened mesoappendix*, appendicular *phlegmon*, appendicular *faecolith* and thickened caecum funneling contrast into the orifice of the appendix as *arrowhead sign*—are all relevant features in CT scan.
- *C-reactive protein,* even though nonspecific increases in acute phase. *99mTc HMPAO labeled leukocyte* imaging may give guidance in deciding the management.

- *Plain X-ray may show* lumbar scoliosis towards right due to psoas spasm which is not uncommon; faecolith on the right side; obliteration of preperitoneal fat line due to retrocaecal appendicitis; segmental ileus in caecum and terminal ileum; speckled extraluminal gas in right iliac fossa, gas in appendix, pneumoperitoneum (very rare); intestinal obstruction (occasionally only); soft tissue mass in mass or abscess of appendix—all these features are very much nonspecific. X-ray is useful to rule out DU perforation, intestinal obstruction, ureteric stone.

Different Scoring Systems Used

Alvarado scoring for appendicitis (1986):	*Score*
Migrating pain	1
Anorexia	1
Nausea and vomiting	1
Tenderness in right iliac fossa	2
Rebound tenderness	1
Elevated temperature	1
Leucocytosis with count more than 10,000	2
Shift to left with neutrophilia in peripheral smear	1
Total score	**10**
Score less than 5: Not sure.	
Score between 5-6: Compatible.	
Score between 6-9: Probable.	
Score more than 9: Confirmed.	

- *Kalam modified* Alvarado scoring system (1994) where shift to left is removed.
- *Tzanakis scoring system* 2005—lower abdominal tenderness—4; rebound tenderness—3; total count > 12,000/cm—2; USG features—6.
- *RIPASA scoring system* (2010)—with 15 parameters.
- *Anderson scoring system*—8 parameters.

Treatment

Surgery-Appendicectomy:

Approaches

1. ***Gridiron incision***: Incision is placed perpendicular to the right spinoumbilical line at the McBurney's point (i.e. at the junction of lateral one-third and medial two-third of spinoumbilical line). (Gridiron is a frame of cross beams to support a ship during repairs. This incision was first described by *McArthur*).
2. ***Rutherford Morison's muscle*** cutting incision *(Muscles are cut upwards and laterally*).
3. ***Lanz crease incision*** centering at McBurney's point—cosmetically better.
4. ***Right lower paramedian incision/lower midline incision***—when in doubt or when there is diffuse peritonitis.
5. ***Laparoscopic approach:*** Becoming popular and better.
6. ***Fowler-Weir approach*** by cutting muscle medially over the rectus.

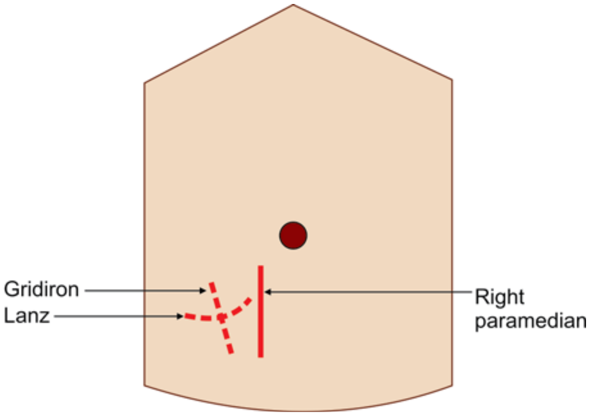

Fig. 24.14: Approaches for appendicectomy.

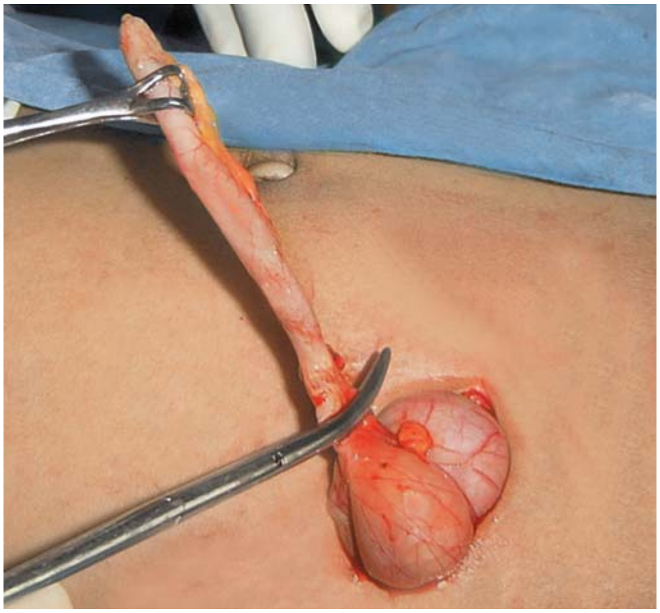

Fig. 24.15: Appendix, on table during appendicectomy.

Procedure

- Under general anaesthesia, skin is incised. Two layers of superficial fascia are cut. External oblique aponeurosis is opened in the line of incision. Internal oblique and transverse muscles are split in the line of fibres. Peritoneum is opened in the line of incision. Caecum is identified by taeniae, and ileocaecal junction. Omentum when adherent is separated. Appendix is held with Babcock's forceps. Mesoappendix with appendicular artery is ligated. Using thread or silk, a purse—string suture is placed around the base of the appendix. Base of the appendix is crushed with artery forceps and transfixed using vicryl (absorbable). Appendix is cut distal to the suture ligature and removed. Stump is cleaned with antiseptics. Purse string suture is tightened so as to bury the stump.

Poverty knows how extremely expensive it is to be poor.

- In difficult cases—*Retrograde appendicectomy* can be done. In presence of pus or burst appendix, the peritoneal cavity is drained.
- Postoperatively, IV fluids, antibiotics are given. Once bowel sounds are heard, oral diet is started.

Complications after appendicectomy

- Paralytic ileus
- Reactionary haemorrhage due to slipping of ligature of the appendicular artery
- Residual abscess (pelvic, paracolic, local, subdiaphragmatic)
- Pylephlebitis (Portal pyaemia)
- Adhesions, kinking and intestinal obstruction
- Right inguinal hernia *(direct)*—due to injury to ilioinguinal nerve
- Wound sepsis 10%
- Faecal fistula
- Respiratory problems and DVT

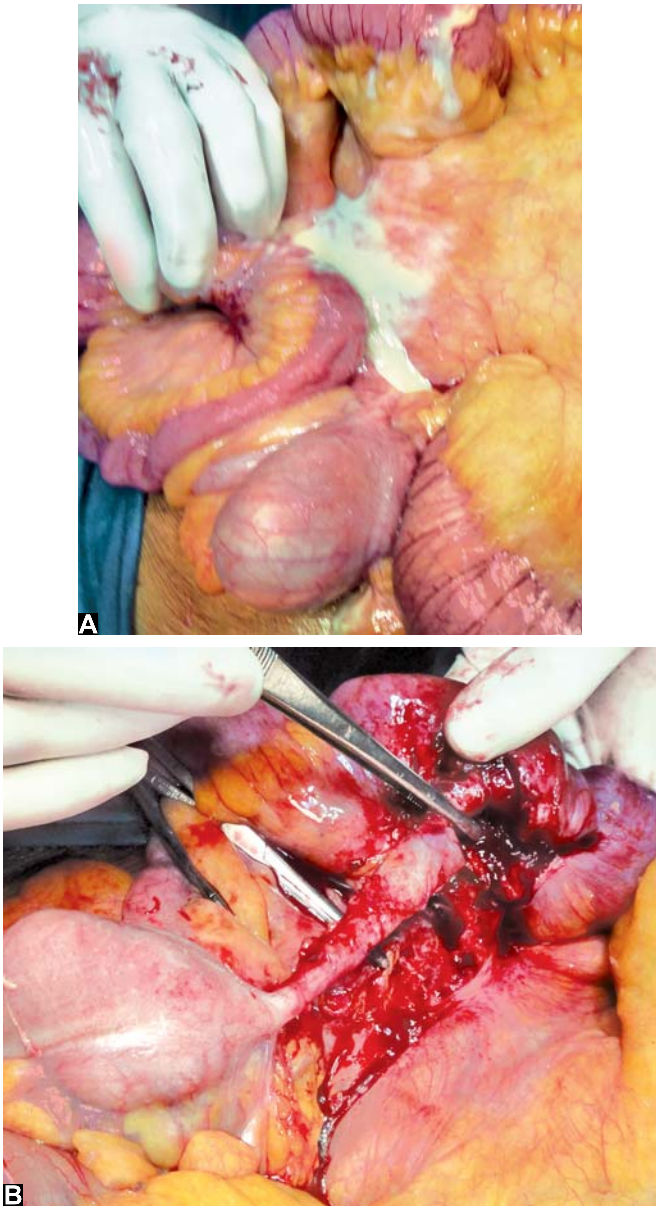

Figs 24.16A and B: Burst appendicitis showing pus. Tip of appendix is burst causing collection of pus adjacent. Usually proximal site of appendix just distal to the site obstruction will burst.

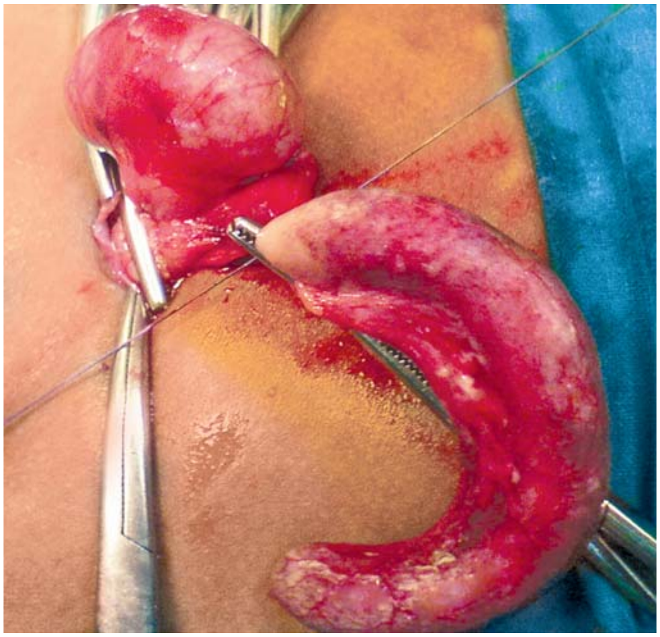

Fig. 24.18: Gangrenous appendix. It is gently ligated at the base using 2 zero vicryl. It is not buried as stump will be friable (Burying is not necessary to any appendicectomy).

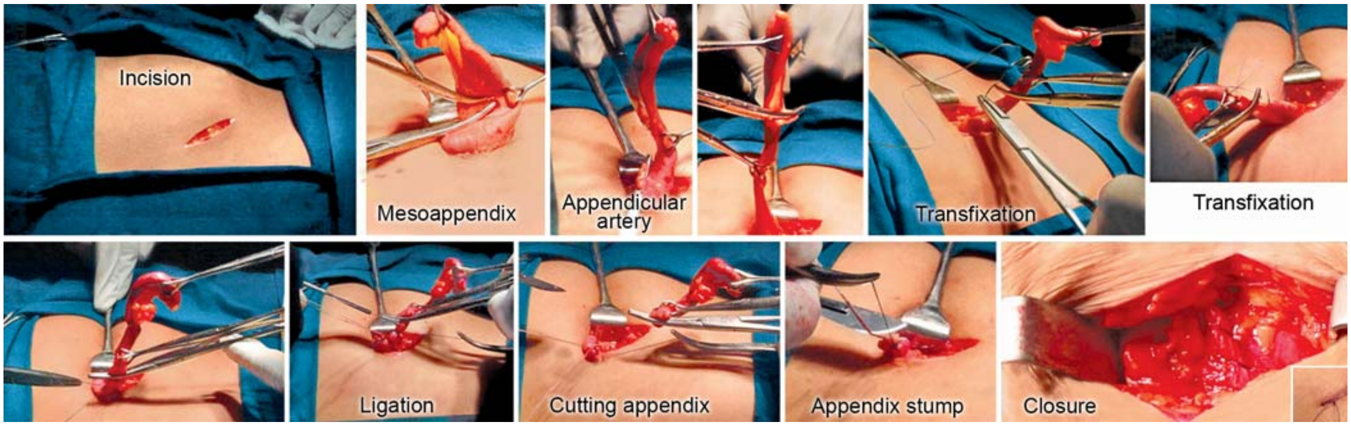

Fig. 24.17: Steps in open appendicectomy. Note the base of appendix, mesoappendix; its ligation; transfixation of the base using 2 zero vicryl; cutting of appendix and closure of the wound. Burying is not essential; it is optional. Burying is done using silk/vicryl using purse string suture.

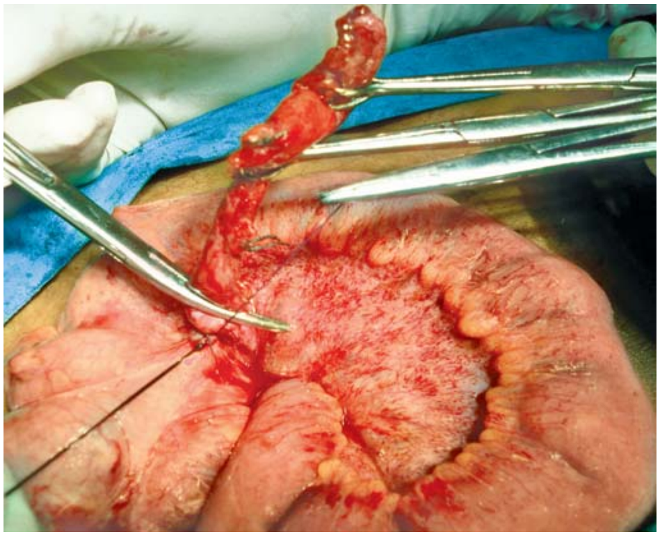

Fig. 24.19: Transfixation of the base of appendix using 2 zero vicryl is an important step after secured ligation of appendicular artery.

Portal pyaemia

- It is rare nowadays
- It is septic portal system thrombosis
- Commonly seen in immunosuppressed individuals
- Infection spreads to liver through portal vein causing rapid multiplication of virulent organisms leading into septicaemia (toxaemia with hypotension, tachycardia), jaundice, tender palpable liver. Patient will be drowsy
- Treatment—antibiotics like cefoperazone, amikacin, metronidazole, meropenem; fluid management; ventilator support
- It carries poor prognosis

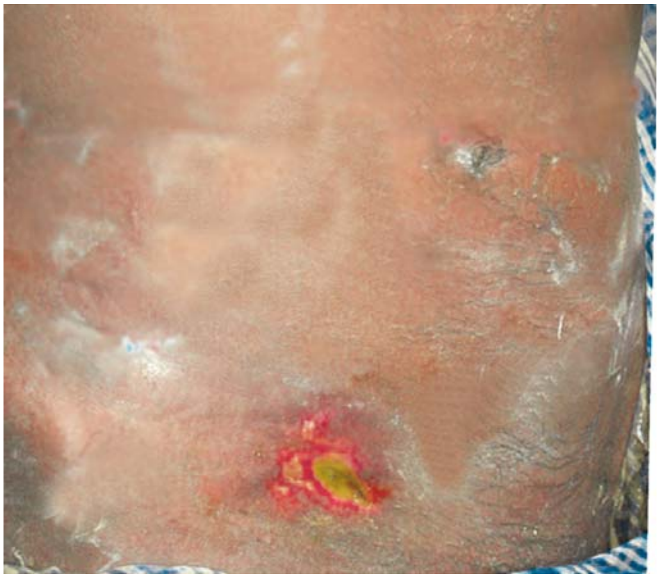

Fig. 24.20: Faecal fistula after appendicectomy. Most of the time faecal fistula subsides by conservative treatment unless there is distal obstruction or specific causes like Crohn's, tuberculosis or malignancy.

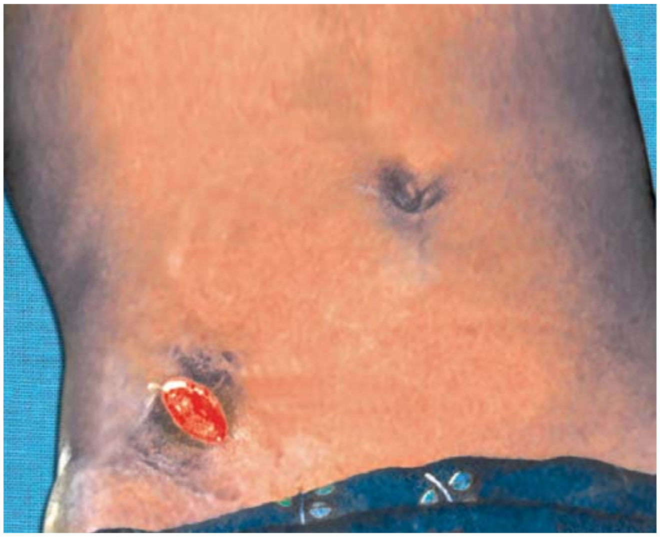

Fig. 24.21: Wound infection after appendicectomy in patient with burst appendicitis.

Troubles in Appendicectomy

- During surgery if appendix is found normal, other cause for symptoms should always be looked for like Meckel's diverticulum, Crohn's disease, ovarian/pelvic causes in females, malignancy, etc.
- Appendicular tumour may be found. If it is in the tip, appendicectomy is sufficient. It could be carcinoid tumour. If it is in the base right hemicolectomy is done.
- Absence of appendix—a rare occasion can occur. Caecum and taeniae should be traced properly before finalising it.
- Appendicular abscess/pelvic abscess formation.
- Malignancy in the caecum is identified on table—right hemicolectomy should be done.
- If Crohn's disease is identified during surgery, appendicectomy can be done with care, if base of the appendix is normal. But in rare occasion where appendix is involved by Crohn's disease, appendicectomy should not be done but treated only with antibiotics and steroids, otherwise fistula can develop.

INCIDENTAL APPENDICECTOMY

- Here removal of normal appendix is done at laparotomy for other conditions, e.g. hysterectomy.
- It is done in vague lower abdominal pain of doubtful severity.
- It is a useful procedure to tackle *'Munchausen syndrome,* i.e. the patient is always worried of pain abdomen and gets relieved after the procedure (psychological benefit).
- It is done along with Ladd's procedure for malrotation.
- It is also done during on table colonic lavage *(Doodleys lavage).*
- It is not done in Crohn's disease (during acute phase of appendicitis), postradiation, immunosuppression, aortoiliac grafts.

Remember

- Appendicitis is common in white races, young males and in those who are on westernised diet
- It most commonly affects individuals of age group 10-20 years. 2/3rd will develop perforation due to rapid progression and poor localisation
- Gangrene, perforation and peritonitis are rare in nonobstructive type; but recurrent appendicitis of nonobstructive origin can cause perforation
- Pneumoperitoneum is not common in appendicular perforation
- Appendicular artery which is an end artery can undergo infective thrombosis and can cause gangrene and perforation
- In retrocaecal appendicitis rigidity is not common; psoas spasm is known to occur
- Pelvic and post-ileal appendicitis can cause diarrhoea. Post ileal appendicitis is difficult to diagnose
- It is difficult to remove subhepatic appendix through McBurney's incision
- Pain will be above and lateral in appendicitis in pregnant women. Appendicitis is the most common acute abdominal condition in pregnancy (1:1500 pregnancies). Incidence of foetal loss is 5% without perforation and it becomes 20% if there is perforation. It is better to do laparotomy to remove the appendix in pregnancy
- In elderly atypical features are more common and so diagnosis is often missed. Gangrene and perforation are common in elderly. Often it mimics subacute obstruction
- In obese patients diagnosis is often difficult
- Appendicitis is rare before 2 years. But when it occurs perforation and peritonitis are common carrying poor prognosis
- *Negative appendicectomy*—incidence is 30%
- *Reginald Fitz of Boston* coined the term appendicitis. *McBurney* described clinical features; *Claudius Amyand* (1736) did first appendicectomy
- Appendix is found on the *left side in situs inversus* patient. Situs inversus may be *both thoracic and abdominal or only abdominal*
- Acute pancreatitis (straw/haemorrhagic chicken broth fluid), DU perforation (bile fluid), perforated Meckel's diverticulum, twisted ovarian cyst/ectopic pregnancy (bloody fluid) are *important life-threatening conditions* which may be missed for appendicitis and patient might undergo appendicectomy as a wrong procedure in these patients
- ***Simple appendicitis*** is one where the symptoms are of less than 48 hours duration with imaging studies showing appendicitis without abscess or phlegmon
- ***Chronic appendicitis'*** earlier this term was not used, but is presently accepted terminology; few attacks of recurrent appendicitis will lead into chronic appendicitis. It presents with episodic often vague discomfort with colicky pain in RIF, anorexia, malaise, pain with movement and is often called as *grumbling appendicitis*. TC, US, CT scan may be normal in these patients
- *Perforation rate* in appendicitis is 25% in general; in children and elderly it becomes 45-50%. *High fever* more than 102°F, *TC* > 18,000/- are suspected features of rupture
- *Mortality rate of appendicectomy* is less than 1%. Morbidity and complications are more after surgery for perforated appendicitis
- *Surgical site infection* is 5% in uncomplicated appendicitis; 20% in perforated appendix after surgery
- *Small bowel obstruction* postoperatively is 1% with simple appendicitis; 3-5% in perforated appendicitis after surgery. More than 50% of obstruction occurs in first year of postoperative period
- In children with appendicitis, there is poor localisation and so peritonitis is common. So conservative therapy should be avoided. Surgery is the only choice of treatment otherwise early peritonitis is the danger.
- Appendicular mass is initially treated with *Ochsner Sherren* regime. After 6 weeks, *interval appendicectomy* is done.
- Children, old age, faecolith, laxative abuse, diabetes mellitus, immunosuppression and pelvic appendix are high-risk factors for perforation in appendicitis
- In pelvic and retrocaecal appendicitis, adjacent ureteral inflammation can occur in which urine on analysis shows blood cells and pus cells
- Incidence of removal of normal appendix is 30%
- *Stump appendicitis* is inflammation and infection in the remaining portion of the appendix in the stump after appendicectomy. It is a rare entity
- On table during surgery, normal appendix if found, it is called as '*Lily white appendix*'. Then other pathology like Meckel's, ileal/mesenteric lymph node/ovarian disease has to be looked for

APPENDICULAR MASS (Periappendicular Phlegmon)

- It is the localisation of infection occurring 3 to 5 days after an attack of acute appendicitis.

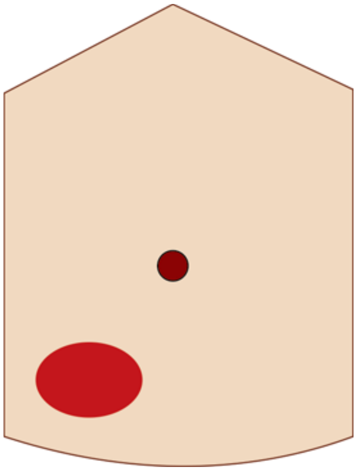

Fig. 24.22: Appendicular mass—a well localised one.

- Inflamed appendix, greater omentum, oedematous caecum, parietal peritoneum and dilated ileum (Ileus) forms a mass in the right iliac fossa.
- This mass is tender, smooth, firm, well localised, not moving with respiration, not mobile, all borders well made

out (well localised) and resonant on percussion. Patient may have fever and features of toxicity.

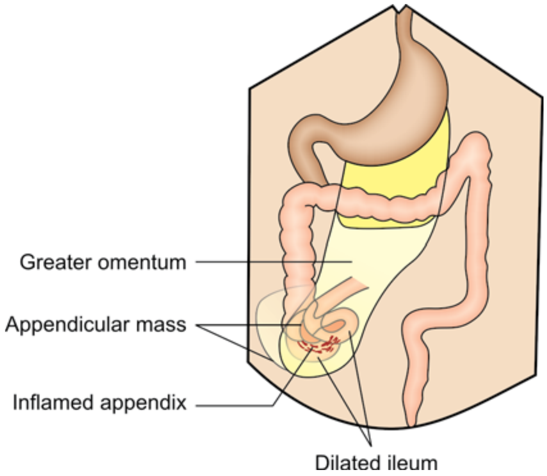

Fig. 24.23: Appendicular mass is formed by dilated ileum; greater omentum; inflamed appendix and caecum. It is resonant, smooth, firm, and tender with well defined borders which does not move with respiration and does not have mobility.

Differential diagnosis

- Carcinoma caecum
- Crohn's disease
- Ovarian disease
- Twisted ovarian cyst
- Actinomycosis
- Mesenteric lymphadenitis
- Ruptured ectopic pregnancy
- Ileocaecal tuberculosis

Investigations

- TC is increased.
- U/S confirms the mass.

Treatment

Conservative ***(Ochsner-Sherren Regimen)***, as nature has already localised the infection, if now disturbed will cause faecal fistula.

Includes observation:

- Temp, BP, pulse chart.
- *Marking the mass to identify the progression/regression.*
- Antibiotics (Ampicillin, metronidazole, gentamicin, or other drugs given depending on severity and requirement).
- IV fluids.
- Analgesics.
- Initial nasogastric aspiration.
  - Patient usually shows response by 48 to 72 hours and mass reduces in size, temperature and pulse becomes normal. Appetite is regained. 90% of patients respond to conservative therapy.
  - Patient is discharged and advised to come *for **interval appendicectomy*** after 6 weeks.

Contraindications for Ochsner-Sherren regimen

1. When diagnosis is in doubt.
2. In acute appendicitis in children and elderly.
3. In burst, gangrenous appendicitis.
4. In patients in whom diffuse peritonitis sets in.

Criteria to discontinue *Ochsner-Sherren* regimen

- Patient becomes more toxic (tachycardia, high fever)
- Persistent vomiting
- Increase or spread of pain abdomen (means onset of diffuse peritonitis)
- Increased size of the mass
- Suppuration (abscess formation) in the mass

In these patients the regimen is discontinued. The patient is taken for immediate surgery, either through laparotomy or through classic approaches

APPENDICULAR ABSCESS

- It occurs due to suppuration in an acute appendicitis or suppuration in an already formed appendicular mass.
- Abscess commonly *occurs in retrocaecal* region but often can occur in subcaecal, preileal lumbar or postileal regions.
- *Pelvic abscess is also common after an attack of acute appendicitis*.

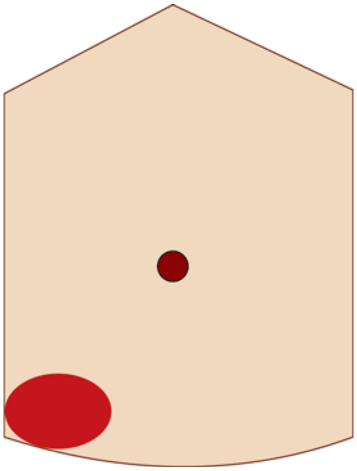

Fig. 24.24: Appendicular abscess. Swelling with ill-defined lower border.

Clinical Features

- High fever, features of toxicity, tender, smooth, dull (to percuss), soft swelling in right iliac fossa which lies towards right lateral and lower side with clear upper margin but indistinct lower margin**.**
- U/S confirms the diagnosis.

Treatment

- Antibiotics are started.
- Under G/A, incision is made in the lower lateral aspect of the swelling above the inguinal ligament. Skin, external oblique muscle is cut. Abscess cavity is opened and pus is drained *extraperitoneally*, which is sent for culture and sensitivity. Wound is closed. A drain is placed through a separate incision. Antibiotics are continued.

An essential aspect of creativity is not being afraid to fail.

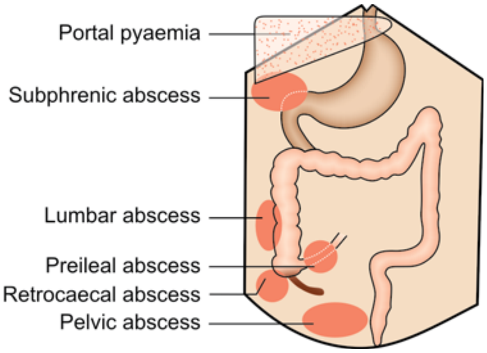

Fig. 24.25: Different sites where abscess can occur after appendicitis retrocaecal, appendicular; pelvic; subphrenic; preileal and lumbar. Portal pyaemia can occur with multiple pyaemic abscesses in the liver.

- *Interval appendicectomy is done after 3 months.*

Pelvic abscess is drained per-rectally or through posterior colpotomy (in females).

FAECAL FISTULA AFTER APPENDICECTOMY

Causes

- It can occur when appendicectomy is done in gangrenous/perforated/friable base appendix.
- It can occur after drainage of appendicular abscess.
- It can occur if appendicectomy is done/attempted in appendicular mass.
- If there is underlying additional pathology like Crohn‘s disease/carcinoma/ileocaecal tuberculosis/actinomycosis during appendicectomy, fistula can occur.

Features

- Faeculent, foul smelling discharge from either main wound or drain site.
- Features of infection.
- Skin excoriation.
- Features suggestive of cause.

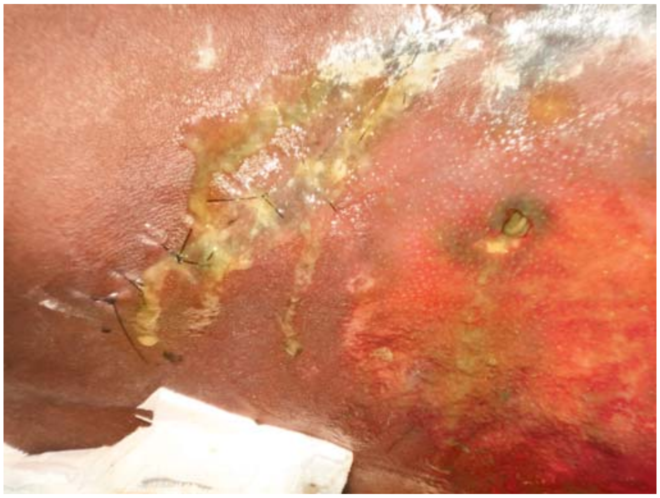

Fig. 24.26: After appendicectomy fistula can occur. Note the faecal discharge.

Investigations

- CT fistulogram to delineate the track.
- CT scan abdomen to find out the other pathology.
- Other relevant investigations, Hb%, albumin level, etc.

Treatment

- Conservative—antibiotics, IV fluids, dressing, zinc oxide cream over the skin, observation.
- Most of the time fistula subsides provided there is no distal obstruction by adhesions or kinking or specific causes like carcinoma or tuberculosis.
- If persists even after 6 weeks, resection of ileocaecal segment and anastomosis is done.

MUCOCELE OF APPENDIX

- It can be neoplastic or non-neoplastic.
- It occurs when proximal end of the lumen of appendix gets slowly and completely occluded, usually by a fibrous stricture causing collection of sterile fluid (mucus) in the lumen. It is a retention cyst.
- Appendix is grossly enlarged with features of sub-acute appendicitis.
- Mucocele can get infected leading to *empyema* of appendix.
- Rupture of mucocele can lead to *pseudomyxoma peritonei*. Neoplastic type causes generalised pseudomyxoma peritonei; non-neoplastic type causes localised pseudomyxoma peritonei.
  (Other cause for pseudomyxoma peritonei is ruptured mucinous carcinoma of ovary).
- Often mucocele of appendix is also caused by a mucus secreting adenocarcinoma and if it is so right hemicolectomy is done.

Clinical Features

- Colicky pain in right iliac fossa.
- Tenderness in the right iliac fossa.

Investigations

- U/S abdomen.

Pseudomyxoma peritonei

- Jelly like mucoid yellowish-brown substance accumulates in peritoneal cavity
- Due to ruptured adenocarcinoma appendix/mucocele or mucinous carcinoma of ovary
- Common in females
- Painless progressive distension of abdomen with intestinal obstruction occurs eventually
- Shifting dullness is absent
- Surgical debulking, oophorectomy, appendicectomy, omentectomy are often done
- Chemotherapy is useful—cisplatin
- Carries poor prognosis

Treatment

- Appendicectomy.

Note:
Pseudomyxoma peritonei is presently considered to be due to neoplastic adenocarcinoma of appendix with gelatinous fluid collection in the peritoneal cavity. It is also seen in cystadenocarcinoma of ovary. Treatment is surgery and chemotherapy.

NEOPLASMS OF THE APPENDIX

- It is rare.
- It is often postappendicectomy histological diagnosis.
- ***Cystic neoplasms of appendix***: Simple cyst (non-neoplastic mucocele); mucinous cystadenoma; mucinous cystadenocarcinoma (most common form of cystic neoplasms); pseudomyxoma peritonei. Simple cyst is non-neoplastic obstruction of the lumen and is less than 2 cm in size which contains mucin. Mucinous cystadenoma attains progressively large size of up to 8 cm with CT showing calcification of the wall. Laparoscopic appendicectomy is not used in mucinous cystadenoma. Hemicolectomy is done in mucinous cystadenocarcinoma and cystadenoma of large size and if base is involved.
- ***Carcinoid tumour*** is the most common type. It is less aggressive. It is often incidentally found. It is arising from Kulchitsky cells in crypts of Lieberkuhn (argentaffin tissue). It is ten times more common than other types (One in 400 appendices). Commonly its location is in the tip. 75% are less than 1 cm; 15% are 1-2 cm; 10% are > 2 cm in size. It stains chromograninB immunohistochemically. Distant and nodal spread occurs if tumour is more than 2 cm. Carcinoid of appendix may be goblet cell type or classical type histologically. Goblet cell has got more mortality than classic type. Treatment is appendicec-tomy. Right hemicolectomy is done if base is involved or size is more than 2 cm or nodes are involved. 5-year survival is 90%.
- ***Primary adenocarcinoma of the appendix*** is rare. It can be mucinous (common) or colonic (less common) type. Acute presentation as appendicitis is common in colonic type. It is staged as Duke's staging A, B, C and D. 5-year survival rate for each is 100%; 65%; 50% and 5% respectively. Mucinous type has got better prognosis. 5-year survival for mucinous type is 70% and colonic type is 40%. Mucinous type can rupture into the peritoneal cavity and can cause pseudomyxoma peritonei.

Tumours of the appendix

- Carcinoid tumour—most common site is appendix
- Primary adenocarcinoma of appendix
- Mucocele of appendix leading into pseudomyxoma peritonei

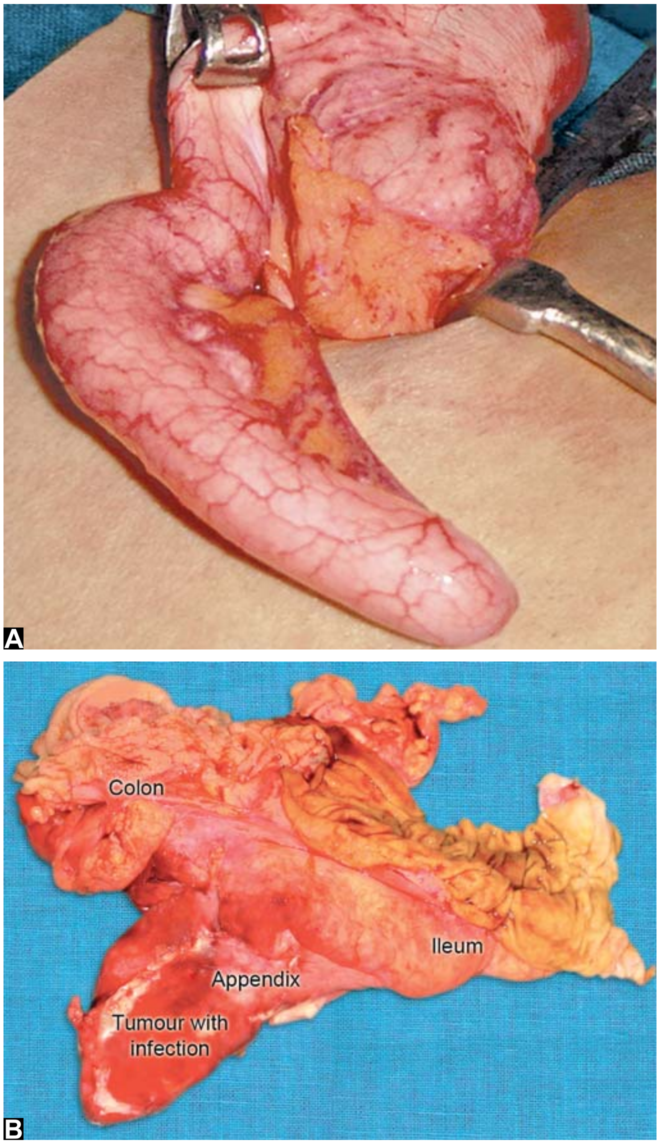

Figs 24.27A and B: Carcinoid tumour in two different patients. If it is in the tip or away from the base, then appendicectomy is sufficient. If it is in the base or extending into the caecum then right hemicolectomy is needed.

LAPAROSCOPIC APPENDICECTOMY

This is newer, popular and ideal method of appendicectomy. *It has become gold standard method of treatment*.

Advantages

- Diagnosis is confirmed.
- Other parts of the abdomen are visualised.
- In females pelvic structures are assessed properly.
- Trauma of access is less.
- Faster recovery.
- Laparoscopic appendicectomy is definitely better whenever there is vague abdominal pain; atypical pain; situs inversus; in women; subhepatic appendix and as interval appendicectomy.

Art is long, life short, judgement difficult, opportunity transient.

Disadvantages

- Technical difficulties especially in burst appendix.
- Cost factor and availability.

Note:

Consent for conversion should be taken.

Technique

- Procedure is done under general anaesthesia. Head down position with right tilt is needed. Surgeon and camera man stands on the left side. Scrub nurse on the right side. Monitor is kept on the foot end right side.
- 10 mm camera port is placed at the umbilicus. Working ports are two 5 mm, one on each side of lower abdomen or one on left side and another on the lower midline. One of the working ports can be 10 mm in difficult appendectomies.

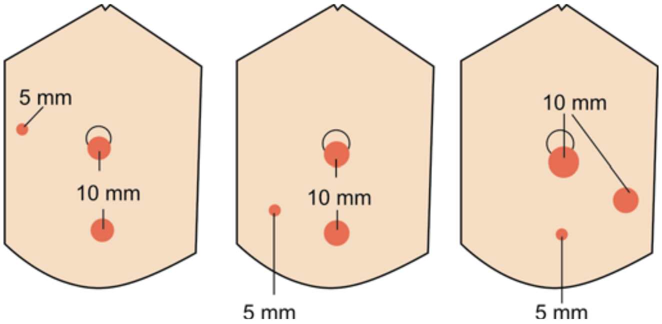

Fig. 24.28: Different possible port placements for appendicectomy.

- Pneumoperitoneum is created using CO_2.
- Appendix is held with grasper or *Babcock's forceps*. Mesoappendix is cauterised by bipolar or unipolar cautery.
- Appendix is dissected up to the base of the appendix.
- Base of the appendix is ligated with loop ligature. Intracorporeal ligature also can be placed using vicryl 2 zero suture material. Appendix is removed through 10 mm working port along with reducer. Often retrieval bag can be used to remove the appendix.
- Umbilical port is closed in two layers. Other ports are closed by skin sutures. If gangrenous or burst appendix drain can be placed through one of the ports.
- Oral food is started in 12 hours.

Laparoscopic assisted appendicectomy can be done if caecum is mobile without much adhesions and mobilised appendix can be delivered through the 10 mm umbilical port. As in regular laparoscopic procedure, appendix is identified and mobilised. Appendix which is not friable and not turgid can be delivered through umbilical port gently. During this procedure gas flow should be stopped so that all gas in the cavity will be empty. Mesoappendix is ligated. Appendix is transfixed like in open method and removed. Stump is pushed back into the place. Telescope is passed again to confirm the position and security of the stump. Advantage is that it is faster.

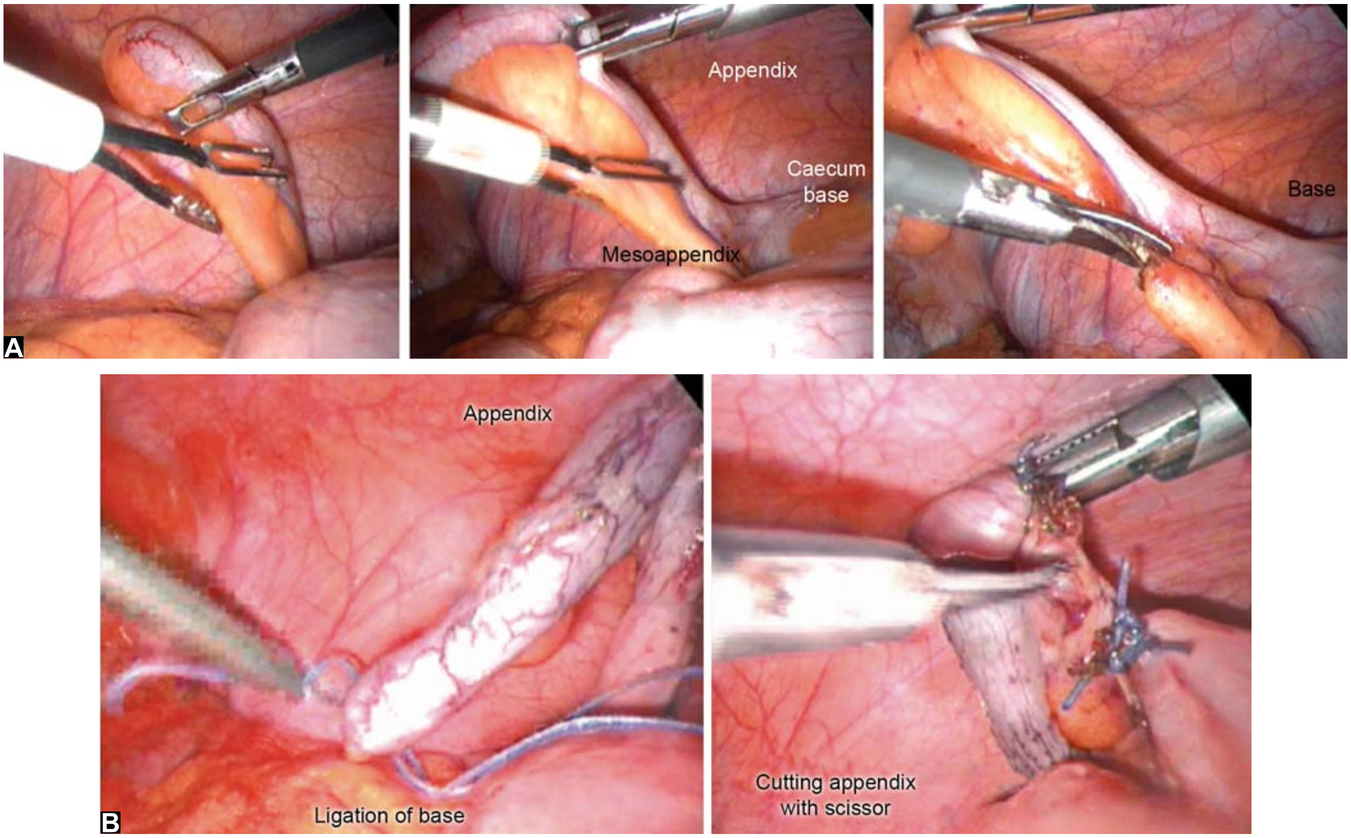

Figs 24.29A and B: Laparoscopic appendicectomy. Note the mesoappendix is being cauterised. Ligation of the appendix is done using vicryl (*Courtesy:* Dr Keshava Prasad, MS, KMC, Mangalore).

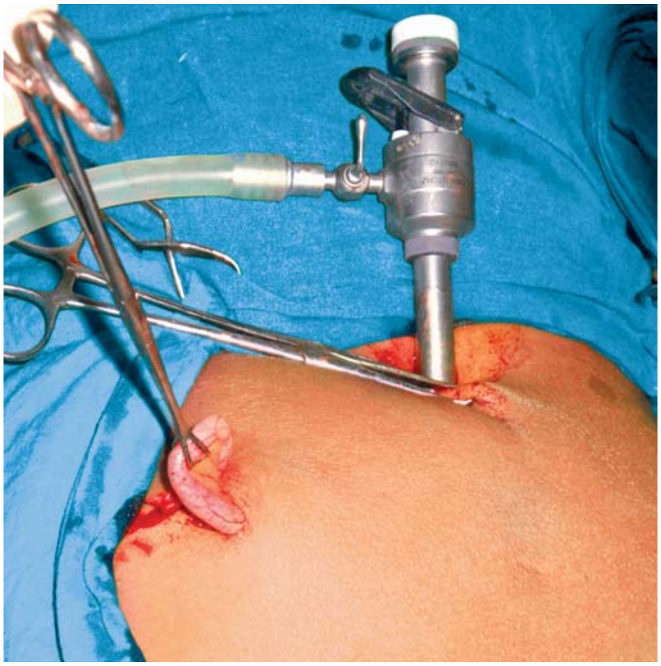

Fig. 24.30: Laparoscopic assisted appendicectomy.

Complications

- Injury to bowel, vessels while passing the ports.
- Complications of pneumoperitoneum.
- Accidental cautery injury to bowel, vessels and other vital structures.
- Bleeding.
- Bowel perforation, peritonitis.
- Ligature slipping, leak, peritonitis, fistula formation.

Chapter 25 Rectum and Anal Canal

CHAPTER OUTLINE

- Anatomy
- Per-rectal Examination
- Proctoscopy (Kelly's)
- Sigmoidoscopy
- Colonoscopy
- Carcinoma Rectum
- Solitary Ulcer Syndrome
- Rectal Prolapse
- Anorectal Malformations
- Pilonidal Sinus
- Piles/Haemorrhoids
  - External Piles
- Anal Fissure
  - Sentinel Pile
- Anorectal Abscess
- Fistula-In-Ano
- Anorectal Strictures
- Condyloma Acuminata
- Anal Intraepithelial Neoplasia
- Malignant Tumours of Anal Area
- Anal Margin Tumours
- Sacrococcygeal Teratoma
- Anal Incontinence
- Descending Perineal Syndrome
- Proctitis
- Proctalgia Fugax
- Hidradenitis Suppurativa of Anal Region
- Pruritus Ani
- Gastrointestinal Haemorrhage
  - Upper GI Bleed
  - Lower GI Bleed
  - Obscure GI Bleed

ANATOMY

Rectum

- It is the distal portion of the large gut, placed between the sigmoid colon above and anal canal below; in front of last three pieces of sacrum and coccyx (From S3).
- The three cardinal features of large intestine (sacculation, appendices epiploicae and taeniae) are absent.
- The upper third of rectum is covered by peritoneum on front and sides, mid third only on the front, lower third is infraperitoneal.
- The rectum is pulled forward by the puborectalis muscle forming the anorectal sling which is primarily responsible for rectal continence.
- It has got three lateral flexions left, right and left from below upwards (*Valves of Houston*).

Rectosigmoid Junction

- Implies a segment of bowel comprising the last seven centimetre of sigmoid colon and upper five centimetre of rectum.
- On sigmoidoscopic examination it is taken as a point 15 cm from the anal verge.

Supports of Rectum

- Pelvic floor.
- *Fascia of Waldeyer*: It is the condensation of pelvic fascia behind rectum, contains superior rectal vessels and lymphatics.
- *Lateral ligaments of rectum*: It is the condensation of pelvic fascia, attaches rectum to the posterolateral wall of lesser pelvis.
- *Denonvillier's fascia:* It is the fascial condensation which separates rectum from prostate in males and vagina in females.
  - Rectum is supplied by rich network of vessels that originates from superior and middle rectal arteries and median sacral artery.
  - Lymphatic drainage from upper half of rectum is to inferior mesenteric nodes; from lower half to internal iliac nodes.

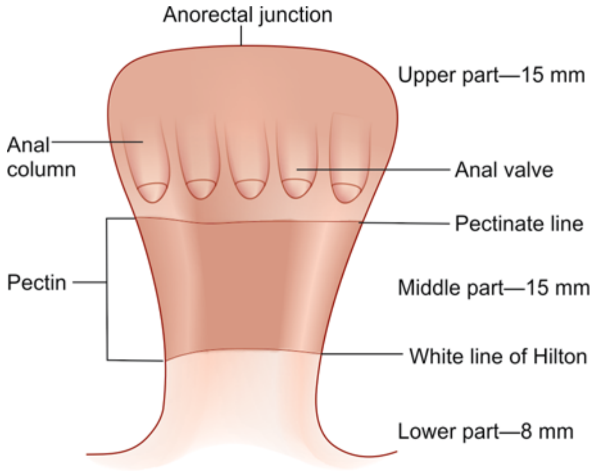
Anorectal junction
Upper part—15 mm
Anal column
Anal valve
Pectinate line
Pectin
Middle part—15 mm
White line of Hilton
Lower part—8 mm

Fig. 25.1: Interior of the anal canal.

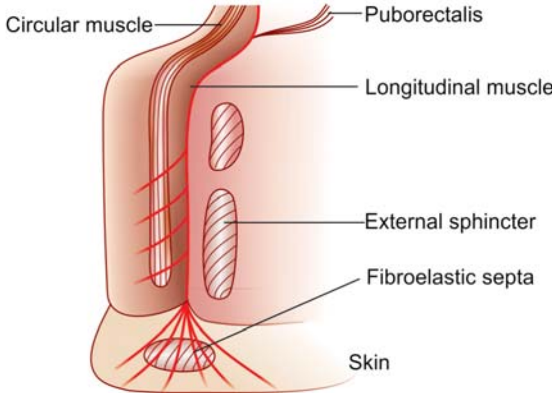
Circular muscle
Puborectalis
Longitudinal muscle
External sphincter
Fibroelastic septa
Skin

Fig. 25.2: Sphincters of anal canal.

It is innervated by autonomic nervous system; sympathetic (L_1, L_2) is motor to sphincter and inhibitory to musculature; parasympathetic (S_{234}) is motor to musculature and inhibitory to sphincter. Sensation of distension is carried through parasympathetic; pain sensation is carried by both.

Anal Canal

- It is 4 cm long, extends from levator ani muscle to anal verge.
- The *dentate line* represents the former site of the embryonic anal membrane.
- The lining of the canal above this line is columnar epithelium and below is skin.
- The mucosa above this line has an autonomic nerve supply, below is by pudendal nerve.
- The venous drainage above this line is by inferior mesenteric and portal circulation, whereas below to systemic venous circulation.
- Internal haemorrhoids develop above this line.

Sphincters of Anal Canal

Internal sphincter: Downward extension of circular muscle of rectum, under control of autonomic nervous system.

External sphincter: Surrounds the internal and continuous with the levator muscle.

Blood supply is from inferior rectal artery.

Venous drainage: Internal rectal venous plexus lies in the submucosa of the anal canal. It drains mainly into the superior rectal vein but communicates freely with external plexus. It is an important site of portasystemic communication. They are situated in anal column at 3,7,11 o'clock. Their saccular dilatation forms *'primary internal piles'*.

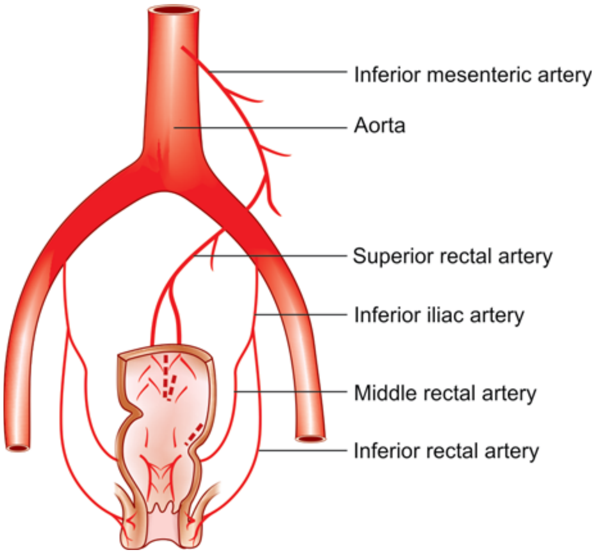
Inferior mesenteric artery
Aorta
Superior rectal artery
Inferior iliac artery
Middle rectal artery
Inferior rectal artery

Fig. 25.3: Blood supply of rectum.

PER-RECTAL EXAMINATION (Digital Examination of the Rectum)

> *We had almost come to the conclusion that the case (of vasovesiculitis) was one of acute appendicitis, but decided to make a rectal examination for the sake of completeness.*
>
> **—Ulysses Grant Dailey, WS Grant,** 1924

No abdominal examination is complete without a per rectal examination.

a. It is done to palpate.
   1. Carcinoma rectum.
   2. Stricture rectum.
   3. Polyps.
   4. Thrombosed piles.
   5. BPH and carcinoma prostate.
   6. Secondaries in the rectovesical pouch (Blumer shelf).
   7. Sphincter tone.
   8. Pelvic abscess (is felt as boggy swelling).
b. To feel the internal opening of anal fistulas.
c. In bimanual palpation of the bladder or pelvic tumours.
d. In acute abdominal conditions—it reveals dilated empty rectum with tenderness.

Positions for Per-rectal Examination

- Right lateral position.
- Left lateral position.

in the case of acute abdomen, it is more important to insert the finger into the lower end than to put the thermometer into the upper end of the alimentary tract.
— sir zachary cope

- Dorsal position in ill-patients.
- Lithotomy position.
- Knee-elbow position.
- Picker position: Patient in standing position leans forward by grasping a chair or stool. This method is used to palpate seminal vesicles which is involved by tuberculous seminal vesiculitis (as craggy feeling) or in trichomonas vaginalis infestation of seminal vesicle.

Per-rectal examination is contraindicated in acute fissure-in-ano.

PROCTOSCOPY (KELLY'S)

Indications

a. *Diagnostic*—piles, fissure in ano, polyps, stricture, etc.
b. *Therapeutic*—injection therapy for partial prolapse or piles, cryotherapy for piles, polypectomy, biopsy for carcinoma rectum or anorectum.

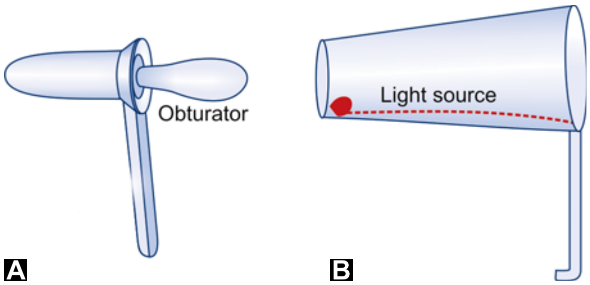

Figs 25.4A and B: Types of proctoscopes: (A) Non-illuminating, (B) Illuminating.

Types

a. Illuminating.
b. Nonilluminating.

Parts (10 cm)

Proctoscope is conical shape, with proximal diameter more than the distal, so as to illuminate the light at the required site properly. *Obturator* is the inner part which allows the easy insertion of the proctoscope.

Positions for Proctoscopy

- Left lateral position (common).
- Right lateral.
- Lithotomy.
- Knee-elbow position.

Technique of Proctoscopy

After doing digital examination, proctoscope with the obturator is introduced inside, through the anal canal in the direction towards the umbilicus.The obturator is removed. Proctoscope is withdrawn and during the course of withdrawal, any pathology has to be looked for.

Acute anal fissure is contraindication for proctoscopy.

SIGMOIDOSCOPY

> *Annual sigmoidoscopy for all, after their fortieth birthday: something to look forward to.*
>
> — **Henry George Miller,** 1968

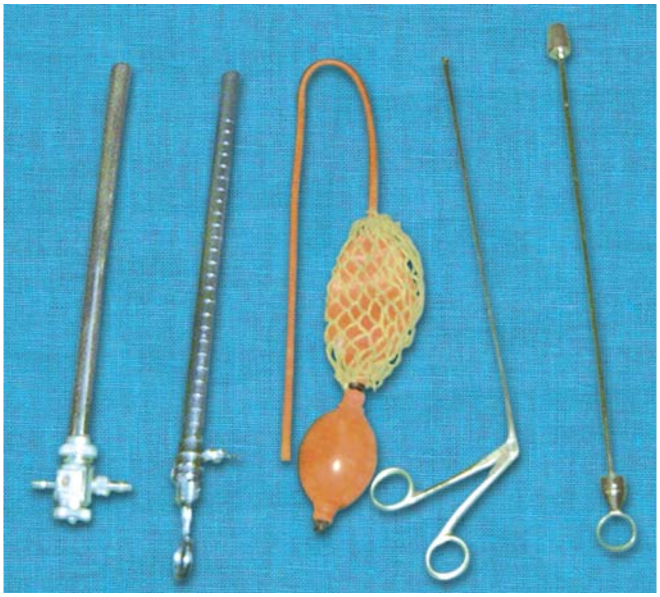

Fig. 25.5: Rigid sigmoidoscope with inflation balloon and biopsy forceps.

- It is used to visualize rectum and sigmoid colon, take biopsies from suspected lesions and do therapeutic procedures (polypectomy, control of bleeding, etc).
- There are *two* types:
  1. ***Rigid***—25 cm long, with illumination.
  2. ***Flexible***—60 cm long.
- In lateral position as in P/R examination or proctoscopy, sigmoidoscope with obturator is passed into the rectum and obturator is removed. Rectosigmoid is inflated with air and scope is negotiated into the sigmoid through *Alpha* (α) *manoeuvre*. Looked for any disease, biopsies are taken and also any required procedure is done.

Precaution: Care should be taken in acutely inflamed sigmoid colon, because chance of perforation is high.

COLONOSCOPY

- It is 160 cm long, flexible.
- Technique is same as sigmoidoscopy, but is passed up to the caecum.

Technique

It is often done under GA using propofol or with laryngeal mask airway (LMA). It can be also done under high sedation. But patient finds difficult to tolerate pain and distension. Passage by *elongation; looping with a manoeuvre; dither-torquing* (clockwise-anticlockwise rotations) methods are used. Difficulty is encountered while passing through sigmoid colon, splenic flexure, and hepatic flexures. Continuous air inflation is important. It is better to visualise the lumen and then pass the colonoscope. Often it can also be negotiated into the terminal ileum. Changing position, abdominal pressure is required for better negotiation of the colonoscope. Technique differs in patients after haemicolectomy or through colostomy.

Indication

- Bleeding per rectum, resistant anaemia.
- To take biopsies from different parts of the bowel.
- To identify synchronous growths, ulcerative colitis.
- To remove polyps.
- When barium enema shows irregularity.
- For therapy—colonoscopic polypectomy, dilatation of stricture colon, fulgaration.

Contraindication

Acute ulcerative colitis.

Advantage

It helps to visualise full length of the colon. GA is not used, except in children.

Disadvantage

Takes a long time and requires expertise to do the same.

Hazards

- Perforation of bowel, splenic flexure is the commonest site.
- Trauma to anorectum.
- Sepsis.
- Haemorrhage.
- Problems of incomplete therapeutic procedures.

CARCINOMA RECTUM

> *Bubo is an apostem breeding within the anus in the rectum with great hardness but little aching. This I say, before it ulcerates, is nothing else than a hidden cancer.... Out of bubo (cancer) goes hard excretions and sometime they may not pass, because of the constriction caused by the bubo, and they are retained firmly within the rectum.... I never saw nor heard of any man that was cured... but I have known many that died of the foresaid sickness.*
>
> —**John of Arderne,** 1414

- It is common in females.
- Usually originates from a pre-existing adenoma or papilloma (tubular polyp).
- In 3% of cases, it occurs in multiple sites (synchronous).

Aetiology

- Red meat and saturated fatty acids increase the risk.
- High fibre diet reduces the risk.
- Alcohol and smoking increases the risk.
- FAP and adenomas are more prone to carcinomas.

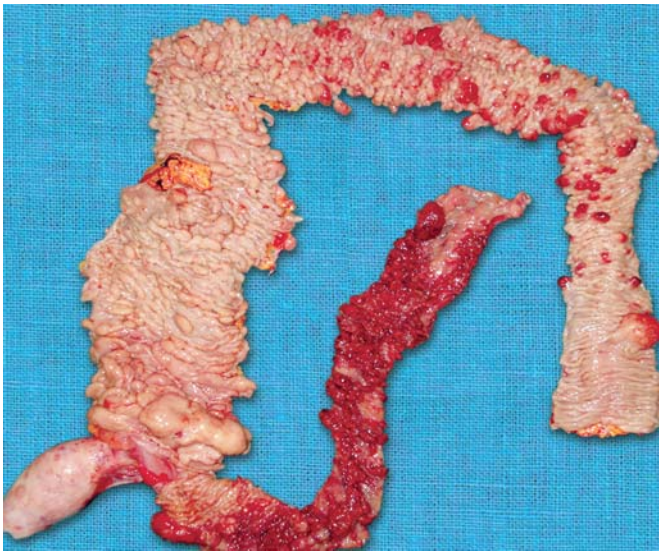

Fig. 25.6: Polyps of colon. It is important premalignant condition for colorectal cancers.

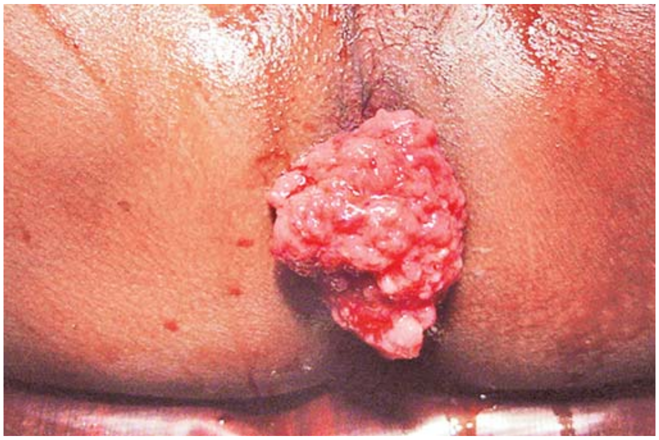

Fig. 25.7: Large rectal polyp which has come out of the anal canal. It is potentially malignant.

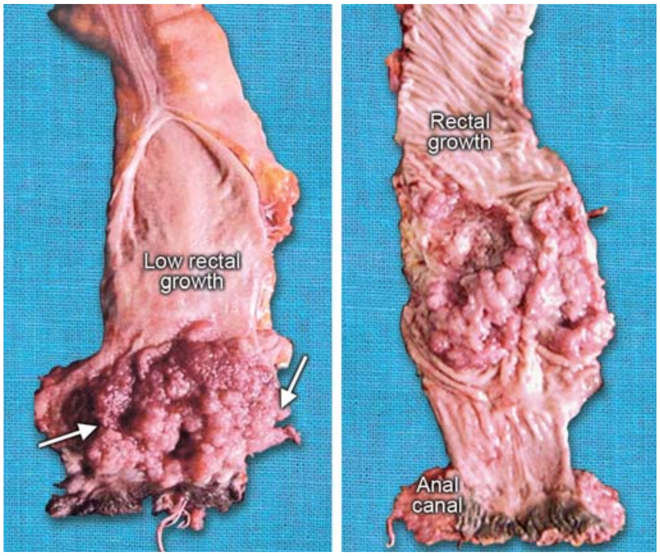

Fig. 25.8: Pathological specimens of anal canal, rectum and sigmoid colon after abdominoperineal resection for low rectal carcinoma and midrectal carcinoma.

Never insult the vagina by examining the rectum first —**An old axiom**

- Villous adenoma has 40% chance of turning malignancy, size more than 2 cm is at high-risk.
- Ulcerative colitis; Crohn's disease; HNPCC carries higher incidence of carcinoma of rectum.
- Family history of rectal cancer—any first degree relative of a person with rectal cancer will show two times increased risk of carcinoma rectum.
- Risk of developing other cancers like of endometrium (40%); stomach (20%); biliary tree (20%); ovary (10%) in the same patient also increases.
- '*Adenoma—carcinoma sequence'* like in carcinoma of colon is known common method of occurrence.

Gross: It can be:

- *Ulcerative.*
- *Papilliferous.*
- *Infiltrative.*
- *Annular*: It is common in rectosigmoid junction.
- *Diffuse type:* Often observed in patients with ulcerative colitis which carries poor prognosis.

Histologically: It is *adenocarcinoma* which may be:

- *Well-differentiated—10%*
- *Moderately differentiated—65%*
- *Undifferentiated—25%*

Spread

- *Local spread*: Initially, it spreads, locally circumferentially (takes 12-18 months to complete the circumference of the bowel). Later spreads out to the muscular coat and perirectal tissue. Then to prostate, bladder, seminal vesicles in males, and uterus and vagina in females. Posteriorly into the sacrum and sacral plexus, laterally into the ureters.
- *Lymphatic spread:* Above the peritoneal reflection, spread occurs upwards along the colonic lymph nodes. In midrectum, into the para rectal and mid-rectal lymph nodes. Downward spread is rare occurs when growth is close to the anal canal into the inguinal lymph nodes. Obturator nodes may be involved in 8% of lower rectal growths.
- *Venous spread* occurs to the *liver 35%, lungs 20%, adrenals 10%* and other areas.
- *Perineural spread* carries poor prognosis.

Haggitt's invasion of malignant polyp
(Similar to carcinoma colon)

In pedunculated polyp
Level 0—noninvasive carcinoma over the summit
Level 1—invasion to head of pedunculated polyp
Level 2—invasion to neck of the pedunculated polyp
Level 3—invasion to stalk of the pedunculated polyp
Level 4—invasion to base of pedunculated polyp
In sessile polyp—all lesions are level 4

Duke's staging of carcinoma rectum

A. Confined to bowel wall, mucosa and submucosa
B. Extends across the bowel wall to the muscularis propria with no lymph nodes involved
C. Lymph nodes are involved

Modified Duke's staging

A. Growth limited to rectal wall (15%)
B. Growth extending into extra rectal tissues but no lymph node spread (35%)
  B1: Invading muscularis mucosa
  B2: Invading in to or through the serosa
C. Lymph node secondaries (50%)
D. Distant spread to liver, lungs, bone, brain

Note:
Astler-Coller's grading (Refer Page No. 967, Chapter 22).

Colloid carcinoma of the rectum

It is 12% common in young people
Types

- Primary and secondary
- *Secondary* colloid carcinoma is common type and is due to mucoid degeneration of adenocarcinoma itself.

Primary is mucus within the cell with displaced nucleus *(signet ring)*

Primary type has got poorer prognosis compared to secondary

TNM staging of rectal cancers

Tx—Primary not assessed
T_0—No primary tumour
Tis—Carcinoma *in situ*
T_1—Invasion to submucosa
T_2—Invasion to muscularis propria
T_3—Invasion of subserosa or non-peritonealized perirectal tissues
T_4—Involvement of visceral peritoneum, other organs or structures
N_0—No nodal spread
N_1—1-3 nodal spread
N_2—4 or more nodal spread
M_0—No distant spread
M_1—Distant spread present

Clinical Features

- Bleeding per rectum/anum (may mimic haemorrhoids)—earliest symptom.
- *Spurious diarrhoea*: It occurs in *early morning* due to overnight mucus accumulation in the rectum causing urgency for defecation, but results in spurious diarrhoea with incomplete evacuation.
- *Tenesmus*: It is painful incomplete defecation with bleeding.
- *Bloody slime*: Mucus with blood in stool.
- Sense of incomplete evacuation, constipation.
- Presenting as piles due to proximal venous congestion by tumour or as fistula in perianal region (which itself is tumour extension into the anal canal).
- Anaemia, malnutrition, loss of appetite and weight.
- Altered bowel habits.
- Urinary symptoms are due to infiltration of bladder or prostate.
- Back pain, due to invasion of sacral plexus.
- Ascites, liver secondaries, urinary symptoms.

90% of rectal growths can be felt by per-rectal examination. Depth of tumour penetration can be assessed through digital examination as superficial tumours are mobile; deep penetrating tumours are not mobile.

Investigations

- Proctoscopy.
- Sigmoidoscopy.

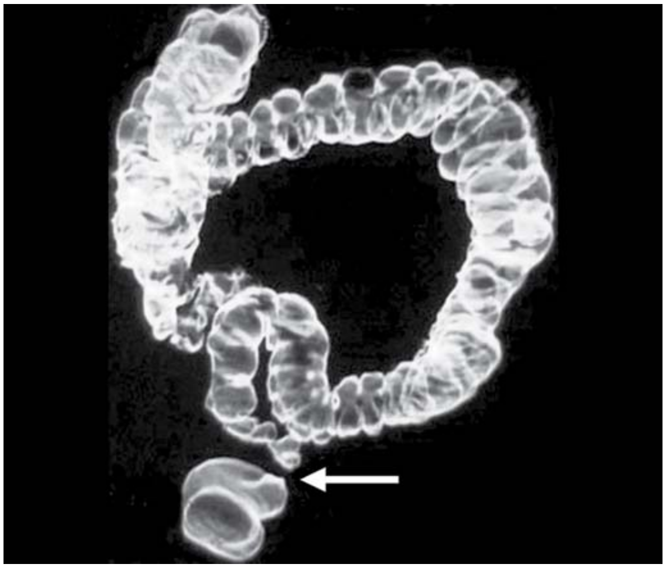

Fig. 25.9: Air-contrast CT of colon and rectum showing significant narrowing in rectosigmoid junction.

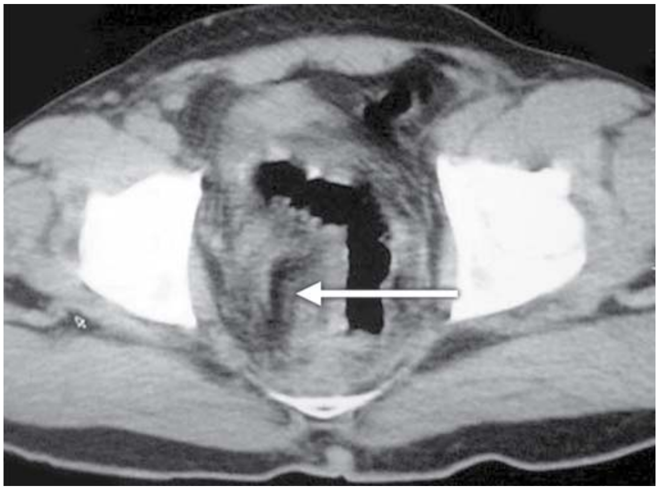

Fig. 25.10: CT picture of carcinoma rectum. CT is useful in evaluating the extent, nodal status and operability.

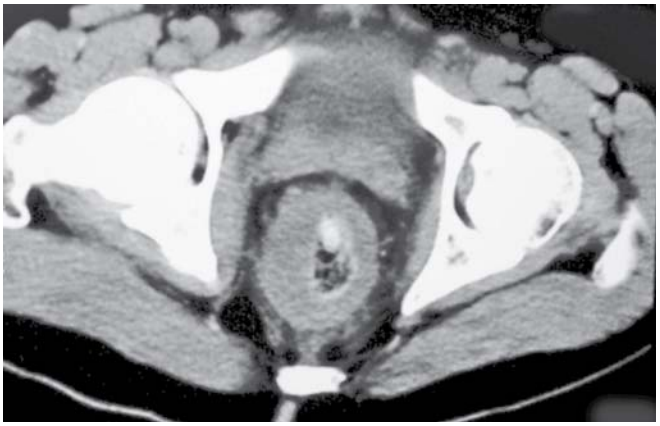

Fig. 25.11: CT scan showing carcinoma of rectum.

- Biopsy using *Yeoman's* forceps.
- Barium enema in case of FAP and synchronous growths.
- Colonoscopy is ideal to rule out presence of any synchronous growths proximally (5%) or polyps.
- Even though *colonoscopy* is done *rigid proctosigmoidoscopy* is a must to identify the precise location of the tumour and to measure the tumour distance from anal sphincter accurately.
- U/S abdomen—to look for secondaries in liver, ascites.
- CT scan to see operability, local extension, size, nodal status, ureteral involvement, presence of perforation or fistula. CT is very useful to assess *nodal status*. Local extension is better assessed by TRUS. Any mesorectal node detected in CT is considered as malignant spread. Liver secondaries are well-identified in CT. Ureteral involvement in CT scan signifies requirement of stenting prior to surgery. Chest CT is essential to look for secondaries.
- *Endorectal ultrasonography*—very useful to assess the local extent of the tumour. *Transrectal ultrasound (TRUS)*/endorectal ultrasound gives more accurate picture of primary tumour, layers, perirectal tissues and nodes. TRUS is superior in T staging of rectal cancers. Its accuracy is 95% compared to MRI (85%); and CT (75%). Endorectal US based T staging and N staging used now depends on layers involved and presence of nodes. TRUS detects nodes more or equal to 5 mm size.
- *Endorectal coil MRI* (*EC MRI)* is very useful as it gives larger field of view compared to TRUS; extent, adjacent organ spread are better assessed by MRI. Recurrent tumour is better assessed by MRI.

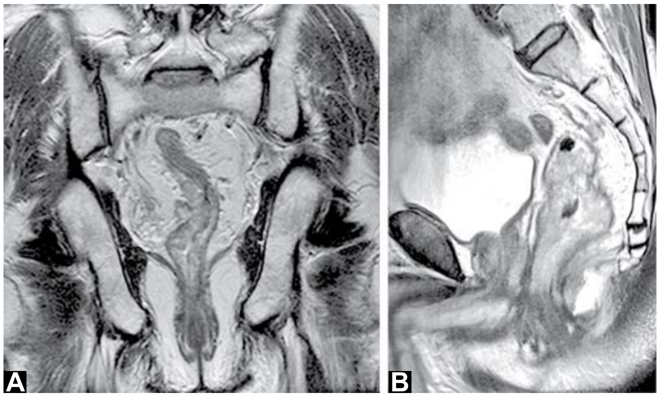

Figs 25.12A and B: MRI showing carcinoma of rectum and adjacent structures. MRI is good method to evaluate the spread and staging.

- *Fluorine—18 fluorodeoxyflucose PET scan* is useful to detect recurrent local tumours; metastatic disease; to detect pathologic response in preoperative chemoradiation. PET is not accurate for nodal spread.
- Blood tests like haematocrit; CEA; blood urea and serum creatinine; serum electrolytes and proteins for management purpose. *CEA estimation*—it is raised in metastatic disease. It is important during follow-up after treatment.

Differential diagnosis

- Inflammatory stricture
- Amoebic granuloma
- Tuberculosis
- Carcinoid
- Solitary ulcer syndrome

Treatment

Surgery

Principles

- Surgery is the main method of treatment. Preoperative chemoradiotherapy is often used if growth is invading into adjacent tissues (T4). Adjuvant chemotherapy and radiotherapy is a must.
- Genetic, morphologic, biologic features of rectal cancers are similar to colonic cancers.
- But *anatomical factors* make it more complex than colonic cancers, like its location deep in the pelvis, relation to important structures like ureters, bladder, genital, autonomic nerves and anal sphincters. So surgical approach is very difficult.
- *Avascular endoplelvic fascial plane* is important during dissection to avoid injury laterally to autonomic nerves (will cause impotence in men and bladder dysfunction in both sexes); more medial dissection leads into incomplete clearance and high local recurrence.
- *Abdomino perineal resection (APR)* is the gold standard. But if tumour is well-differentiated and if there is adequate margin above the anal canal, a sphincter saving *anterior resection (AR)* can be done. *Low anterior resection (LAR)* is possible if EEA stapler is used for anastomosis. But anterior resection should not be done by compromising the adequacy of tumour clearance. Tumour clearance is still the priority in rectal cancer as it decides the eventual outcome.
- *Total mesorectal excision (TME) should be the goal* in all procedures as mesorectum contains nodes and lymphatics, clearance of which gives better result.
    - It is a *sharp dissection* (not blunt) in avascular areolar plane between fascia of rectum which encroach the mesorectum and parietal pelvic wall fascia.
    - Mesorectum should not be breached.
    - Absolute haemostasis, *preservation of autonomic nerves* and dissection under vision are the essential principles.
    - Both layers of membranous anterior Denonvillier's facia should be dissected off the prostate and seminal vesicles in male to have proper clearance.
    - In TME for middle and lower rectum, entire mesorectum should be removed.
    - For upper rectal tumors, TME is done 5-6 cm below the lower margin of the tumour.
    - TME improves the quality of life in relation to *impotence, retrograde ejaculation and urinary incontinence.* These complications are around 50% or more in APR whereas in TME it is less than 20%.
- Recommended *distal rectal margin clearance is 5 cm*, however 2 cm distal margin is an *acceptable* clearance.
- *Circumferential resected margin—CRM* (*radial margin, > 2 mm*) is more important than proximal and distal margin. A 5 cm clearance of mesorectum from the primary tumour is essential as tumour implants can occur only for up to 4 cm from primary tumour margin.
- *Principles to be followed*—adequate lymphatic and vascular clearance; en bloc resection of primary tumour; no / less touch technique; avoiding spillage; adequate radical surgery.
- *Ultra low anterior resection or intersphincteric* resection can be considered in low rectal tumours.
- After resection (on table*) irrigation of rectal bed* with cetrimide or hypertonic solution like distilled water is often practiced as they are tumoricidal.
- Selection of the procedure AR or APR is decided by proper staging using TRUS and MRI.
- Neoadjuvant chemotherapy or neoadjuvant chemoradiotherapy is often used in T3 lesions which also may avoid APR; and AR may be sufficient.
- Local wide excision approaches are often used when tumour is < 4 cm; < 40% circumference involved; T1 N0/ T2 N0 tumour:
    - *Transanal approach* is used in small tumours that is 3-5 cm above the dentate line but within 10 cm from the anal verge:
    - *Transcoccygeal Kraske's approach* for posterior wall rectal tumours (dangerous posterior faecal fistula can occur);
    - *Transanal endoscopic microsurgery* (Buess) using operating microscope and videoscope can also be done. This *endosurgical device* is 4 cm in diameter, specialized sealed proctoscope with ports for CO_2 insufflation, water irrigation, and suction and for monitoring intrarectal pressure. CO_2 insufflation distends the rectum and local excision is done with proper positioning of the device.
- In females, partial vaginectomy with or without hysterectomy and bilateral oophorectomy may be needed in T4 lesions to achieve surgical resection. Removal of uterus and ovaries prevents patient from developing possible associated cancers of these organs. Carcinoma rectum also spreads to ovaries commonly which can be prevented by oophorectomy.
- Often resection of liver secondaries can be undertaken in selected patients when one lobe is involved or solitary secondaries are present.
- *Laparoscopic APR/AR* is becoming popular.

  Features are:
    - Dissection will be more meticulous.
    - Less blood loss, less postoperative pain.
    - Early bowel function.
    - Clearance is same as open method in relation to primary tumour and nodes.
    - Short hospital stay, mortality and morbidity are similar to open method.
    - Port site recurrence chances are 0.5-2% (Earlier it was higher; now it has reduced due to proper technique, careful handling of the specimen, specimen isolation prior to extraction, trocar site irrigation with cytotoxic agents and povidone iodine).

- For carcinoma rectum presenting with obstruction, an initial proximal colostomy is done. Neoadjuvant chemoradiation is given. Patient is reassessed for operability. Then APR is done with permanent colostomy.
- Incidence *of local recurrent rectal cancer* is 30%.
    - 80% of local recurrence occurs within 2 years of surgery.
    - Common site of recurrence is in the pelvic wall. It also can occur at distal anastomotic margin.
    - Intractable pelvic pain, urinary symptoms, sepsis, bleeding, perineal sinus, swelling and induration, bowel disturbances are the features.
    - It is evaluated by CEA, biopsy, CT abdomen, MRI of pelvis and PET scan.
    - It is often difficult to manage.
    - Incidence of recurrence will come down to 5% in proper TME.
    - Palliative chemoradiation, end colostomy, ureteral stenting are the palliation. RT controls pain and bleeding.
    - Extensive radical surgery like removal of tumour with pelvic structures can be undertaken with diversions but with a high mortality rate and failure rate.
- Proper *preoperative bowel preparation* is a must in all rectal cancers which reduces the postoperative problems like sepsis, leak and increases the early chance of recovery.
    - Bowel preparation by polyethylene glycol; electrolytes *orally*; bowel wash *per anally*; bowel antiseptics like neomycin 1 gm tid previous day/erythromycin 1 gm tid/ metronidazole tid *orally.*
- Systemic antibiotics like cefazolin, metronidazole, gentamycin.
- Prophylactic heparin or low molecular weight heparin.
- Foley’s catheterisation and nasogastric tube should be passed.
- Preoperative adequate hydration using IV fluids.
- Blood grouping and crossmatching and required bottles of blood should be reserved if needed to transfuse during the procedure.

Different surgeries for carcinoma rectum are:

1. **Abdominoperineal resection** (A-P resection) (APR) wherein sigmoid, descending colon and upper rectum is mobilised per abdominally. Anal canal with perianal and perirectal tissues are dissected per anally. Retained colon is brought out as *end colostomy* in *left iliac fossa*.

APR—types

- *Miles*—abdomen first, perineum later
- *Gabriel*—perineum first, abdomen later
- *Lioyd-Davis*—synchronised (together), (combined)

- APR is done through lower midline incision in lithotomy position. Left-sided colon and entire rectum is mobilised from above. Rectum is mobilised posteriorly in avascular plane in front of nerve plane (hypogastric nerve) between mesorectum and sacrum. Inferior mesenteric artery is ligated high proximal (as lymphovascular ligation) at its origin or just beyond its first branch. Colon is transected and proximal cut end is fashioned for end colostomy in *left iliac fossa*. Through perineum, a purse string suture is placed around anal margin. Circumferential incision is placed around the anus. Dissection is deepened using scissor and cautery into the perineal body, coccyx, ischial tuberosity, ischiorectal fossa. First posterior and lateral dissections are undertaken until it reaches above. Lastly, anterior dissection is done to reach above and specimen is removed through perineal wound. Perineal wound is closed in layers often with a drain. Abdomen drain is placed. Colostomy is created by suturing skin to mucosa using silk.

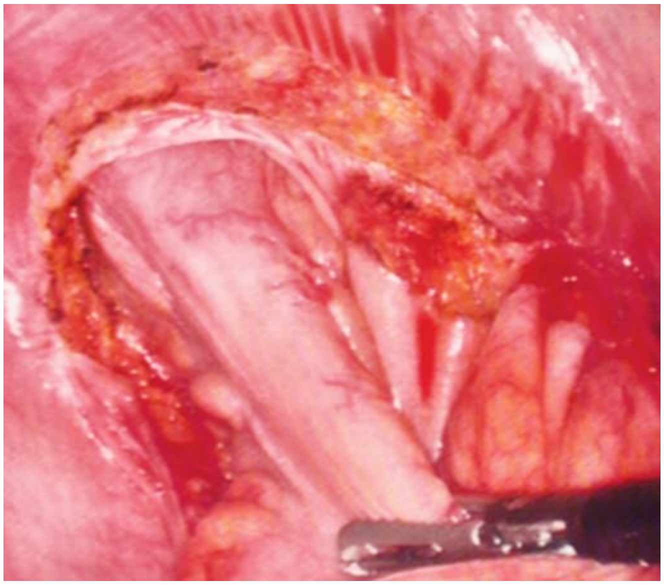

Fig. 25.13: Laparoscopic view of mobilised rectum in APR.

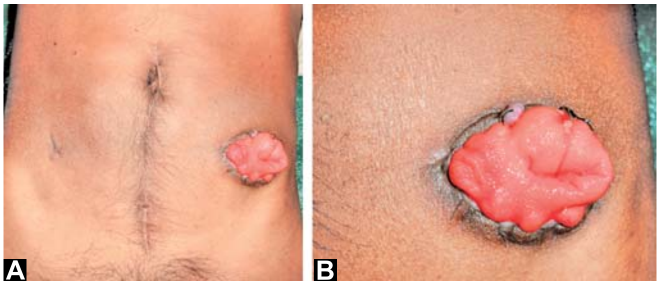

Figs 25.14A and B: Colostomy done in left iliac fossa after APR. Colostomy care is important.

- *Complications of APR* are:
    - Bleeding.
    - Infection of perineal wound.
    - Complications of colostomy like prolapse, stenosis, and infection.
    - Injury to urinary system, ureter, impotence, urinary incontinence.
    - Operative mortality is less than 2%.

APR is the treatment of choice when mesorectum is involved or when it is poorly differentiated tumour or when nodes are involved. It gives adequate clearance.

‘Knowledge’ is proud that she knows so much, ‘wisdom’ is humble that she knows no more.

- *Sphincter saving APR with coloanal anastomosis*: It is done in operable distal rectal tumour in young individual wherein within oncological principle anal sphincter need not be sacrificed but adequate oncological tumour clearance can be achieved. Here a permanent colostomy stoma is avoided. Approach is both abdominal and perineal. Initial dissection of rectal mobilisation is done from above; dissection of rectal mucosa from the anal sphincter at the level of dentate line and complete dissection of distal rectum is done through anal canal. Entire rectosigmoid is removed retaining only the anal sphincter. Colonic J pouch or coloplasty reservoir is created in the mobilised descending colon; coloanal anastomosis is done per anally using hand sutures under direct visualisation.
- *APR with neo sphincter reconstruction* is also occasionally sought; but technically difficult with complications. Perineal colostomy is done with gracilis muscle wrap which is made to produce sphincter like muscle twitch using an implanted pacemaker.

2. **Anterior resection** (Abdominal radical restorative operation) is done in growths located in the mid and upper part of the rectum, which is well-differentiated, small-sized and with a clear adequate length for anastomosis after resection.

- *Anterior resection* is also called as *anterior proctosigmoidectomy* through *abdominal* approach wherein rectum *above* the peritoneal *reflection* is resected with colorectal anastomosis.
- *Low anterior resection (LAR)* is resection of rectum *below the peritoneal reflection* along with the sigmoid colon (as sigmoid should be removed due its precarious blood supply after dissection), with total mesorectal excision (TME) *through abdominal approach* (laparotomy) and *colorectal anastomosis* using *circular* stapler device (EEA stapler). Stoma should be inspected using proctoscope for integrity and when in doubt a covering temporary proximal colostomy should be done. LAR often leads into frequent small bowel movements causing more frequent stools called as *low anterior resection syndrome (LAR syndrome)/clustering*. It can be avoided by creating reservoir either by doing *colonic J pouch* or by doing *coloplasty* 6 cm from proximal divided end of colon (longitudinal colostomy between taeniae of 10 cm which is sutured horizontally).

Criteria for anterior resection—(Low anterior resection, LAR)

- Upper and middle third rectal growth
- Above peritoneal reflection
- Well-differentiated tumour
- < 4 cm size tumour
- In females, growth 7 cm above the anal verge
- T1 N0/T2 N0 tumour
- Tumour without lymphatic or venous spread

Advantages:
- Avoids permanent colostomy.
- Sphincter is retained.
- Patient's acceptance.

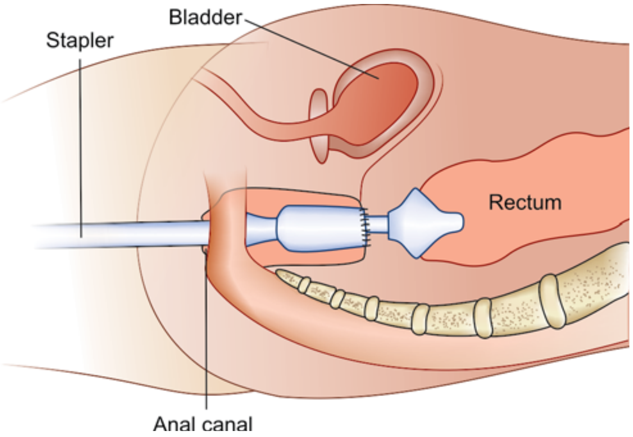

Fig. 25.15: Anterior resection for carcinoma upper rectum using stapler device.

Disadvantages:
- Uncertainity of clearance which is very important in cancer surgeries and so chances of **local recurrence** is high.
- Anastomotic leak, infection; stenosis.

3. **Hartmann's operation** is an excellent palliative procedure—done in elderly people who are not fit for major surgery like AP resection and also in locally advanced tumours. Here rectal growth is resected and upper end of the rectum is closed completely. Proximal colon is brought out as *end colostomy*.
4. **Pelvic evisceration (Brunschwig's operation):** It is removal of rectum with the tumour, all the lymph nodes, urinary bladder, fat, fascia, uterus, vagina, with colostomy and urinary diversion. It is neither favourable nor popular.

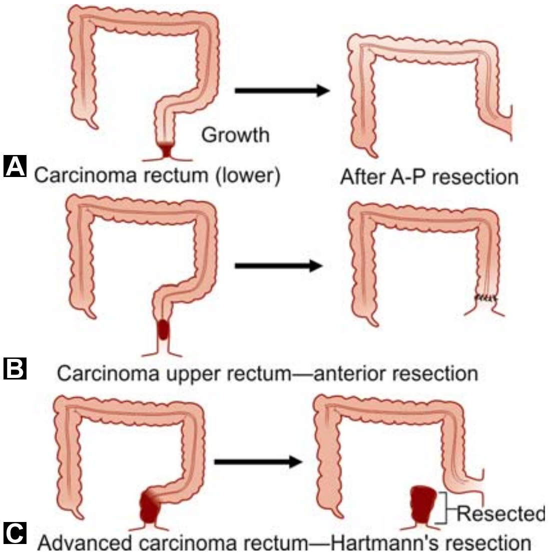
Growth
A
Carcinoma rectum (lower)
After A-P resection
B
Carcinoma upper rectum—anterior resection
Resected
C
Advanced carcinoma rectum—Hartmann's resection

Figs 25.16A to C: Surgeries for Ca rectum (A) A-P resection, (B) Anterior resection, (C) Hartmann's resection.

5. **Palliative colostomy** is done in advanced unresectable growth which presents with intestinal obstruction.

Radiotherapy

It is beneficial in carcinoma rectum showing increased survival rate. It is useful when growth is below the level of peritoneal reflection.

Radiotherapy in carcinoma rectum

- Only rectal adenocarcinoma in GIT responds well for RT
- Preoperative RT can be given to down stage the tumour so as to make it amenable for APR or make it for anterior resection
- Postoperative RT is commonly used
- In small well-differentiated growths *papillon's* intracavitary curative RT can be tried with proper follow-up
- Palliative RT
- IORT (Intraoperative RT) is used in pelvic wall disease. It can cause peripheral neuropathy and ureteral stenosis
- As a component of chemoradiation
- Short course 25 Gy in 5 fractions in 5 days
- Long course 5040 Gy in 28 fractions in 6 weeks
- RT sterilises field; causes down staging of tumour; preserves sphincter

Chemotherapy

It has been tried using endoxan, 5 FU, semustine also with leucoverin (Folinic acid) or levamisole.

- Capecitabine, oxaliplatin are newer drugs used (refer chapter large intestine for detail). Indications are positive nodes; T2 stage; blood spread; for chemoradiation; recurrent tumour; advanced/ metastatic disease.
- Biologic agents.
  - Bevacizumab VEGF receptor antagonist.
  - Cetuximab EGFR antagonist.

Other Methods

- Electrocoagulation and decoring of the tumour, as a palliative procedure; stenting.
- Laser photocoagulation, cryotherapy.
- *Portal vein infusion*; hepatic artery infusion for metastases.
- *Tumour vaccines*: Tumour antigen does not elicit immune response *in situ*; but vaccines are injected to evoke immune response. (1) BCG with irradiated tumour cells (2) Monoclonal antibodies 17-1A [Murine Ig G2A] (3) CEA vaccines.

Chemoradiation in Carcinoma Rectum

- It is very useful *adjuvant therapy.*
- To prevent recurrence after anterior resection.
- Neoadjuvant (preoperative) chemoradiation can be used in tumour like T3 to downstage the disease and make it possible for AR.
- Postoperative chemoradiation.
- Palliative chemoradiation in locally advanced disease or metastatic disease also.
- Recurrent local carcinoma of rectum.
- In carcinoma of rectum presenting with obstruction, chemoradiation is given after loop colostomy.

Prognosis in carcinoma rectum

5-year survival is	***Prognostic factors are***	***Follow-up***
• Stage I 90% • Stage II 75% • Stage III 40% • Stage IV 5%	• Size of the tumour • Differentiation • Mesorectal involvement • Stage of the disease • Nodal status, perineural spread • Distant spread • Circumferential resected margin • Adjuvant therapy used	• Regular colonoscopy • CEA assessment • PET scan • CT/MRI • Colostomy care in APR

SOLITARY ULCER SYNDROME

- It is mainly thickening and disorganisation of muscularis mucosa with superficial ulceration.
- It is usually 4-12 cm from the anal verge in the anterior wall of the rectum. But often can occur in sigmoid colon.
- Attempt to defaecate in the closed pelvic floor causes funneling of the rectum and descent of the anterior rectal wall. Raised intrarectal pressure and hidden intussusception is the cause. It is often seen in sexual abused individuals. Often typical crater like ulcer is seen/felt on the *anterior rectum*. Chronic ischaemia at that point may be the cause.
- In 30% cases, there are multiple ulcers.
- Often there will be inflammation and induration of the area without an ulcer.
- ***Presentations are***—common in young females, constipation, bleeding, mucosal prolapse, chronic pain in the anal canal, incontinence. But sphincter tone on rectal examination is usually normal.
- ***Investigations:*** *Defaecography* shows nonrelaxing persistent puborectalis impression waves. EMG shows decreased electrical activity. Colonoscopy should be done to rule out other conditions like neoplasm, ulcerative colitis. Colonic transit time shows rapid filling of the rectum but delayed clearance of 7 days from the rectum.
- Condition is commonly associated with rectal prolapse.
- ***Differential diagnosis*** are carcinoma, tuberculosis, ulcerative colitis.
- ***Treatment:***
  - High fibre diet.
  - Treatment for rectal prolapse.
  - Avoid surgical excision in solitary ulcer syndrome as much as possible.

RECTAL PROLAPSE

- It is *circumferential descent* of recetum (bowel) through the anal canal.
- It is commonly seen in infants, children and elderly individual**.**
- It is common in females (6 : 1).
- *Faecal incontinence* is very common feature; urinary incontinence occurs in 35% of patients; 15% of patients are associated with vaginal vault prolapse.
- Rectal prolapse can be:
  - *Partial.*
  - *Complete.*
  - *Hidden/concealed*—it is internal intussusception of the sigmoid into the rectum or part of the rectum distally; they do not come out of the anal orifice. Here only mucosa and submucosa separates from muscularis layer and descends.

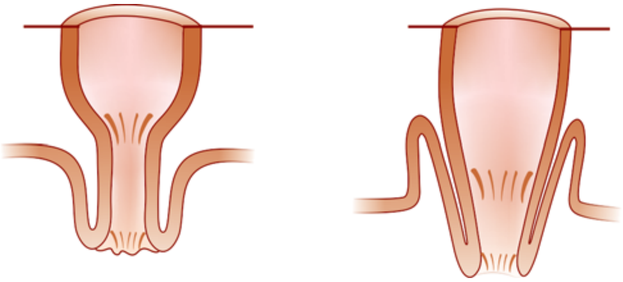

Fig. 25.17: Rectal prolapse.

Aetiology

- *Alexis Moschowitz* put his theory of 'rectal prolapse is due to sliding herniation of the pouch of Douglas through pelvic floor fascia into the anterior aspect of the rectum'.
- *Broden and Snellman* proposed that 'procidentia is a full thickness rectal intussusception starting approximately 7.5 cm above the dentate line which is extending beyond the anal verge'.
- Decreased sacral curvature and decreased anal canal tone are the probable causes in infants.
- *Chronic constipation* with straining is the common cause.
- Diarrhoea, cough, malnutrition are the additional factors in children.
- It may be due to reduced ischiorectal fossa fat, neurological causes, fibrocystic disease of pancreas or poorly developed pelvis.
- There is diastasis of the levator ani, abnormally deep cul de sac, redundant sigmoid colon, patulous sphincter, loss of rectal sacral support, lax and atonic pelvic floor musculature.
- *Pudendal nerve damage* is said to be the cause for pelvic floor and anal sphincter weakness. It may be due to obstetric injury, diabetes, sacral nerve damage.
- In adults, it is common in females, common in multipara—repeated birth injuries to perineum results in damage to the perineal nerve supply.
- Additional factors are due to increased intra-abdominal pressure due to any cause like chronic cough, stricture urethra.

Types

Partial Rectal Prolapse

Here only mucosa and submucosa of the rectum descends, not more than 3.75 cm. There is no descent of the muscular layer. It is the *commonest type of rectal prolapse*.

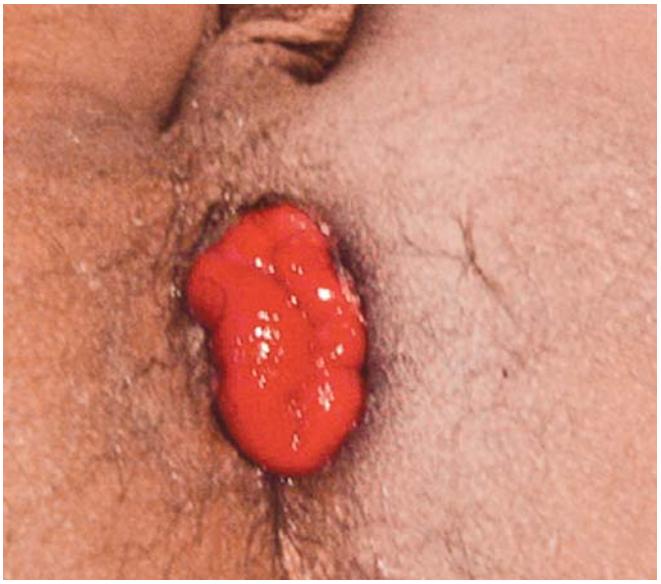

Fig. 25.18: Partial rectal prolapse is less than 3.75 cm and only mucosal.

Clinical Features of Partial Rectal Prolapse

- History of mass per anum, which can be observed when child is allowed to strain in squatting position.
- It is *pink in colour and circumferential*.
- It differs from *piles* (differential diagnosis), the piles are not circumferential and *are plum or blue coloured (not pink)*.

Treatment for Partial Prolapse

1. The nutrition of the child is improved and digital repositioning is tried. *Correction of constipation* is important.
2. Submucosal injections of 10 ml of 5% phenol in almond oil or ethanolamine oleate is given into the apex of the prolapse under. G/A so as to create an aseptic inflammation leading to tethering of mucosa to the underlying muscular coat. Injection can also be given at the base or both at the apex and base.
   - Alternatively 30 ml of tetracycline or oxytetracycline or hypertonic saline injection can also be used. Initial injection is supported by *Thiersch wiring* using (temporary) chromic catgut until adhesion occurs between the mucosa and muscular layer (in 3 weeks).
   - In adults, injection therapy is tried for partial prolapse, results are not as good as in children.
3. Thiersch wiring alone is tried with good success rate in children.
4. *Goodsall's operation* is excision of the prolapsed mucosa at its base, usually in three positions.

5. Stapled transanal rectal resection surgery *(STARR)* is also often used.

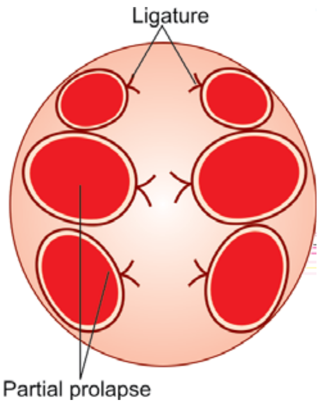

Fig. 25.19: Goodsall's ligature for partial prolapse. Prolapsed mucosa is ligated at different positions using nonabsorbable suture material often with double needled. Usually, it is done for one side in three portions.

Complete Rectal Prolapse

- Also called as *procidentia,* is *less common than partial prolapse.*
- It is common in females (6 : 1 :: female : male).
- It is due to weakened levator ani and supporting pelvic tissues.
- The descent is always more than 3.75 cm, contains all layers of the rectum (i.e. including muscular layer). Often descends down up to 10-15 cm.
- It is often associated with the uterine descent *(uterine prolapse*).
- It is also thought to be as an intussusception of the rectum.
- Once complete prolapse is more than 5 cm, anteriorly it drags peritoneum as pouch which often contains small intestine. On digital pushing it reduces with gurgling.
- Patulous anal sphincter is typical with mucus discharge and faecal incontinence.

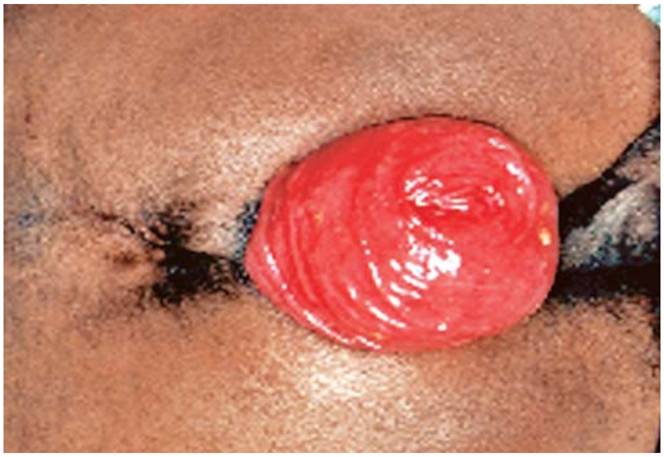

Fig. 25.20: Complete rectal prolapse.

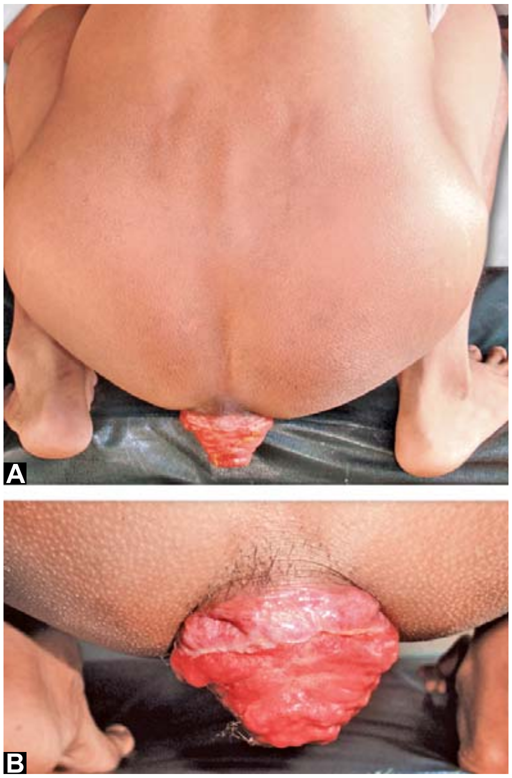

Figs 25.21A and B: Complete rectal prolapse. It should be confirmed by observing the patient during straining in squatting position.

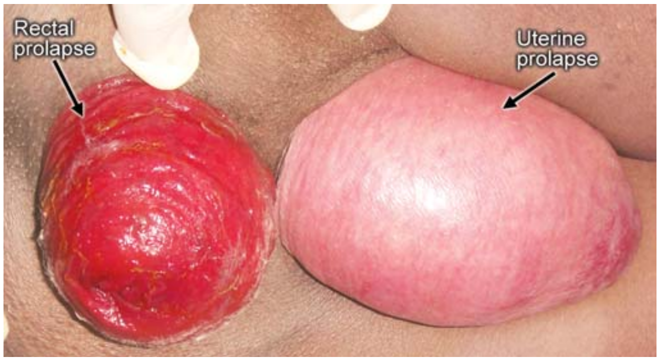

Fig. 25.22: Female patient presented with both complete uterine and rectal prolapse (Procidentia).

- Mucosa of the chronic rectal prolapse is *thickened, ulcerated, bleeds*, and often *incarcerated* below the level of anal verge.

Aetiological factors

- Weak anus, external sphincter and pelvic muscle
- Lax, mobile rectum
- Obliterated ano-rectal angle
- Abnormally mobile rectum with descent

Clinical Features of Complete Rectal Prolapse

- *Complete descent of rectum as mass per anum* circumferentially which is *red* in colour. Mass is usually reducible and painless. Incarcerated or infected rectal prolapse is painful.
- Rectal prolapse may be associated with the *uterine prolapse* (uterine procidentia) in females.

He who is angry is seldom at ease.

- Faecal incontinence (75%) is very common. It is due to disruption of the anal sphincter and prolapsed rectal mucosal discharge.
- Bleeding can occur because of the congestion.
- Sepsis, discharge, fever, anaemia are other features.
- P/R examination shows lax sphincter. Anteriorly, peritoneal sac comes down as a pouch which may contain small bowel.

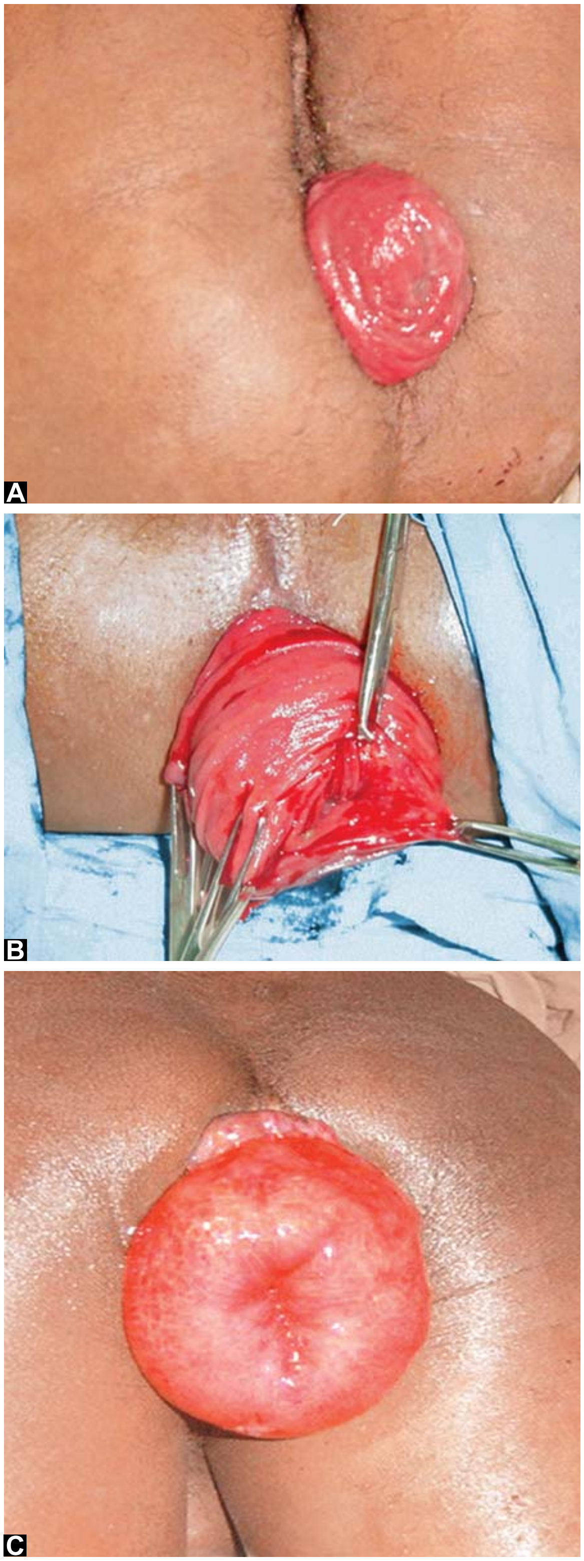

Figs 25.23A to C: Complete rectal prolapse. It is more than 3.75 cm, prolapse of mucosa with muscular layer.

Differential Diagnosis

Rectosigmoid intussusception, third degree piles, large rectal polyp.

Folds in rectal prolapse are *concentric* and *red* in colour; in piles there are *radial nonconcentric* invaginations of haemorrhoidal cushions which are *plum* coloured with a cutaneous component.

Complications of rectal prolapse

- Ulceration, bleeding, anaemia
- Proctitis, sepsis
- Irreducibility, gangrene
- Rupture with evisceration

Investigations

- Defecography reveals increased mobility of the rectum from sacral fixation point with redundant mesorectum and funnel formation. It is fluoroscopic and spot filming in lateral projection after instilling radio-opaque material into the rectum done in sitting posture over a radiolucent commode.
  - Cinedefecography, triple contrast cinedefecography, dynamic MRI defecography, colpocystodefecography are helpful to delineate complex pelvic floor problems.
  - *Defecography detected abnormalities*—megarectum, incontinence, nonrelaxing puborectalis, abnormal perineal descent (2.5 cm), mucosal prolapse, solitary ulcer rectocoele, enterocele.
  - *Preprolapse* (in defecography): Rectum is funnel shaped, lack of fixation to sacrum, excessive rectosigmoid mobility, ring pocket formation, intussusception.
  - *Rectal prolapse* (in defecography)**:** Redundant sigmoid colon, wide deep pouch of Douglas.

Defecographic grading of rectal prolapse

N—normal rectal fixation and sphincter relaxation and rectal emptying
1—nonrelaxed puborectalis
2—mild intussusception
3—moderate intussusception
4—severe intussusception
5—prolapse
R—rectocele

- *Sigmoidoscopy:* It is to detect the tumour in the intussuscepted prolapsed rectum which is an occasional cause.
- *Anal manometry: Resting* (40 mmHg of internal sphincter) and *squeeze* (80 mmHg, exteranal sphincter) pressures at various points in anal canal is measured by placing water filled balloons attached to catheters and transducers placed in the anal canal.
- *Pudendal nerve latency study: Specialized transducer* attached to a glove like device is to be worn on the finger through which digital rectal examination is done. Electrode in the finger is directed over the right and left levator ani complex to measure pudendal nerve terminal motor latency (PNTML) which is normally 1.8-2.2 msec. It is prolonged in pudendal nerve damage.
- *Electromyography study* of the puborectalis muscle tone is also very useful.

Treatment for Complete Prolapse

Aim of Treatment

- To *control* the prolapse; to *restore* continence; to *prevent* constipation.
- *In young males, abdominal repair should be avoided as it injures pelvic nerves leading to sexual impotency*. So perineal approach is better.
- Rectopexy is fixing the rectum to sacrum by sutures or mesh after complete mobilisation of the rectum. Laparoscopic rectopexy using polypropylene mesh and sutures gives good result and has become very popular.
- *Delorme's operation* is better option in young individual with complete prolapse.
- In elderly perineal proctectomy with anterior and posterior reefing of the sphincter muscle is accepted method now. It is similar to Altemeier technique (It is perineal rectosigmoidectomy). Here entire prolapsed rectum and redundant sigmoid is resected through perineum prior to reefing of the sphincter.
- Anal encircling surgeries using synthetic wires/mesh/suture materials are limited to extremely ill patients and elderly who will not withstand perineal proctectomy.
- *Choice of procedure* depends on age, sex, operative risk, pelvic floor defect, degree of incontinence, history of constipation.

Surgery—Types

Abdominal Procedures

Laparoscopic rectopexy

- It is ideal and good approach to fix the rectum to sacrum.
- *Laparoscopic posterior mesh rectopexy (LPMR)* is the procedure done. Prior bowel preparation is needed. Head down, low lithotomy position is needed. Ports are placed as shown in diagram. Sigmoid colon is held with left sided port. Surgeon does dissection from right side. Peritoneum on the right of the rectum is opened from sacral promontory downwards to reach presacral avascular plane. Care should be taken to avoid injury to autonomic nerves, ureters. Dissection is extended down as posterior mobilisation into the pelvis with adequate mobilisation of the rectum. Lateral ligaments are either divided or left alone. *Anterior mobilisation* is also important. Anterior mobilisation along the Denonvillier's fascia is done 5 cm below the peritoneal reflection. 15 × 10 cm polypropylene mesh is placed in the presacral space deep to rectum which is fixed to presacral fascia along the sacrum and sacral promontory. Mesh is sutured to rectal wall also on both sides using interrupted polypropylene sutures. Only partial wrapping of mesh is done. Peritoneum is closed using vicryl.
- Many advocate laparoscopic mobilisation and fixation of mobilised rectum to sacral promontory using polypropylene sutures without mesh.
- *Laparoscopic sigmoid resection and rectopexy* (Laparoscopic resection rectopexy, LRR) is done when there is rectal prolapse with constipation, with excess redundant sigmoid colon with kinking.

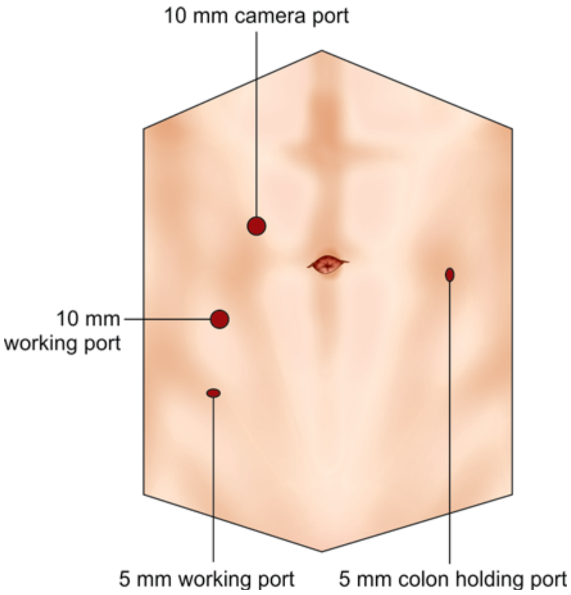

Fig. 25.24: Laparoscopic rectal prolapse—ports.

Well's operation

- *Polyvinyl alcohol sponge* is wrapped around the mobilised rectum and is fixed to sacrum. Infection, fistula formation is high.
- Polypropylene mesh is used as a modification now instead of polyvinyl sponge; wrapping is only partially done to reduce the incidence of constipation.

Ripstein operation

After mobilisation of the rectum, 5 cm width Teflon mesh sling is passed around the rectum to fix it behind to fascia 5 cm below and in front of the sacral promontory. Sling is also fixed in front and laterally to rectum.

Goligher's operation

Rectum is entirely mobilised up to anorectal ring and its posterior muscular layer is fixed to presacral fascia using interrupted polypropylene sutures.

Devadhar rectal plication

Through abdominal approach, junction between thicker lower part and thinner upper part of the intussusception is identified. A purse string suture using silk is placed in front and laterally; further 3-4 interrupted submucosal Lambert sutures are placed to create reverse intussusception.

Lahaut's operation

Rectosigmoid is mobilised fully; mobilised loop of rectosigmoid is passed in front through posterior rectus sheath behind the rectus muscle; extraperitonealisation is done to pull the rectus forward to prevent descent.

Rosoe Graham operation

- After mobilisation of the rectum, levator muscles are exposed and sutured in front of the rectum along with removal of pouch of Douglas.

- *In Dumphy operation* it is done through combined abdominoperineal approach with perineal rectosigmoidectomy. Both procedures control prolapse well but not incontinence.

Muir low anterior resection

In a redundant rectosigmoid with prolapse, rectosigmoid resection is done; excision of redundant pouch of Douglas is done; rectum is fixed to sacrum behind.

Perineal Procedures

Delorme's operation (Mucosal sleeve resection and plication)

After prior bowel preparation, under spinal anaesthesia, in lithotomy position, completely prolapsed rectum is held with Babcock's forceps. 1 in 2,00,000 adrenaline solution is injected into the submucosal plane of the rectum to cause haemostasis. By making longitudinal incision, with sharp scissor and cautery dissection, mucosa is stripped off from the deeper muscular layer from 1 cm below the anal margin to the apex of the prolapsed rectum. Muscular layer of rectum is plicated using absorbable vicryl 2 zero interrupted sutures all around; approximately 12-15 plication stitches are needed. These sutures are tied after finishing the passage of all sutures. Cut end of the mucosa of the apex is sutured to anal margin using interrupted vicryl sutures. It is technically easier. But incontinence may persist and recurrence chance is high.

Perineal posterior fixation of the rectum of Lockhardt-Mummery

Retrorectal space is dissected through perineal approach and posterior rectal wall is sutured to sacrum and coccyx, additionally placing a retrorectal gauze pack to stimulate adhesions.

Wyatt operation

Through perineal post anal approach dissected retrorectal space is placed with a Marlex/mersilene mesh which is sutured high to the sacral promontory and rectal wall laterally.

Mickulicz Miles perineal transanal rectosigmoidectomy/amputation of prolapse

Prolapsed rectum is excised and sigmoid is sutured to the anal margin.

Altemeir's rectosigmoidectomy

After rectosigmoidectomy, colonic anastomosis and pelvic floor is supported by suturing puborectalis muscle in front of the rectum using nonabsorbable sutures.

Anal encircling

- In 1891, Thiersch did anal encircling using silver wire to provoke inflammatory fibrosis as well as to give mechanical support to the anal orifice—*Thiersch wiring*. It is done under local anaesthesia. Two small incisions are made at lateral parts of the anal canal. Silver wire is passed around deeper to these incisions and tied after placing the *index finger* into the anal canal. Wire can be removed after 12 months. Polypropylene, nylon are the other materials that can be used. *Complications*: Pain due to wire erosion, infection, faecal impaction, incarceration, high recurrence of 50% or more. In children with rectal prolapse, *temporary wiring* along with Goodsall's ligature or injection sclerotherapy using thick catgut are often advocated.
- *Supralevator high encirclement of anal orifi*ce (of Notaras) is done above the level pf levator muscles by placing Teflon or nylon ribbons through anterior and posterior incisions. It gives better support and will not cut through. *Thoralsksen* modified this by placing plastic/mersilene tape around the bowel high up.

Procedures for repair of rectal prolapse

Perineal operations	***Abdominal operations: Open/Laparoscopic***
• Mucosal sleeve resection (Delorme)	• Suture rectopexy—Goligher, Lahaut, Devadhar, Rosoe Graham
• Perineal rectosigmoidectomy (Altemeier)	• Prosthetic or mesh rectopexy—anterior/posterior—Well's, Ripstein
• Posterior fixation of the rectum of Lockhardt-Mummery	• Resection rectopexy—Muir's
• Wyatt operation	• Anterior resection
• Mickulicz Miles perineal transanal rectosigmoidectomy/amputation of prolapse	• Lahaut's operation
• Anal encircling (Theirsch)	

Complications of Surgery

- Injury to hypogastric nerve causing impotence.
- Bladder dysfunction.
- Bleeding from sacral venous plexus.
- Injury to rectum and colon causing faceal fistula.
- Constipation after rectopexy is a known complication.
- Recurrence of prolapse.
- Improper correction of continence occurs in 50% cases.
- Infection—proctitis/pelvic abscess, etc.

ANORECTAL MALFORMATIONS (ARM)

- It is due to imperfect fusion of the post-allantoic gut with the proctodaeum.
- Incidence is one in 4500 newborns.

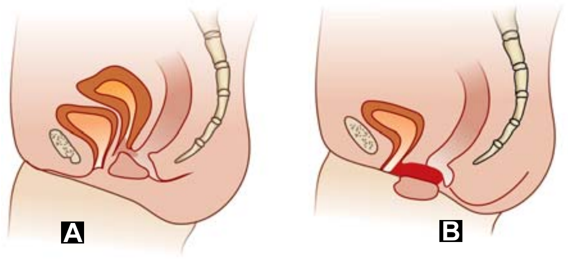

Figs 25.25A and B: Anorectal malformation with fistula in (A) Females and, (B) Males.

Wingspread classification of anorectal malformations
('Wingspread'—name of the place where the conference was held)

Low	*Intermediate*	*High*
It is below the level of pelvic floor, (Puborectalis). Easy to diagnose and treat with good outcome. It may be—	It occurs at the level of puborectalis, with or without fistula	It can be with or without a fistula into the bladder urethra, uterus, vagina It may be:
– Covered anus		– Anorectal agenesis
– Anovestibular fistula		– Rectal atresia
– Anal stenosis		– Cloaca (only in females, with confluence of rectum, vagina, bladder and urogenital sinus)
– Anal membrane.		

Clinical Features

- New born presents with inability to pass meconium, abdominal distension, features of intestinal obstruction, improper anal dimple, sometimes with complaints of passing meconium per urethra.

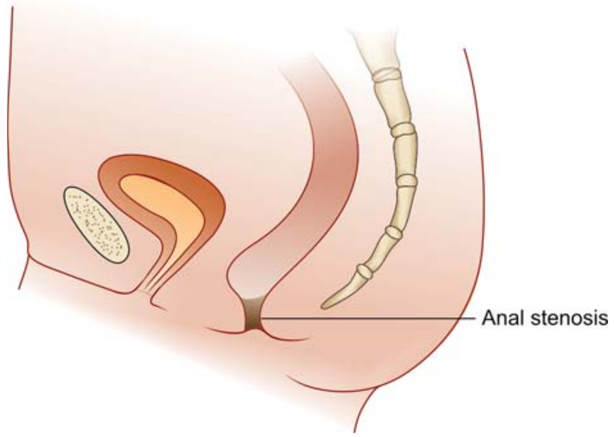

Fig. 25.26: Anal stenosis.

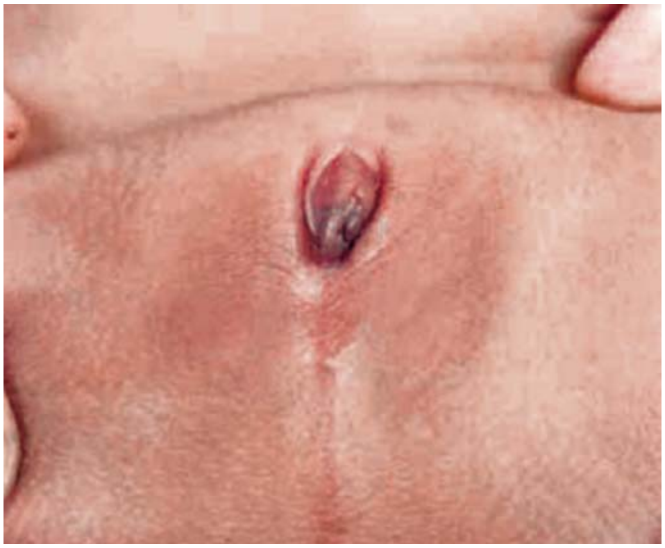

Fig. 25.27: Anorectal malformation, Cloaca type.

- It can be associated with—cardiac anomaly, tracheo-oesophageal fistula, renal anomalies, spinal anomalies. VATER/VACTERL anomaly.
- Most common ARM in boys is *rectobulbar fistula* with fistula beginning distal to puborectalis from distal rectum to bulbar urethra.
- Most common ARM in females is *rectovestibular fistula*.

Investigations

1. *Wangenstein's invertogram:* Usually done 6-12 hours after birth, so as to allow air to reach the rectal pouch. A metal coin (marker) is strapped at the presumed site of anus and X-ray is taken. Length between the rectal pouch and anal dimple marker is more than 2.5 cm in high anal fistula.
    a. In *low fistula*, rectal pouch is distal to the *Stephen's line (Pubococcygeal line).*
    b. In *intermediate,* pouch is at the level of ischial spine *(Kelly's point*).
    c. In *high fistula,* rectal pouch is proximal to the Stephen's line.
2. *Murugassu's technique:* Through visible anal dimple, meconium is aspirated by passing a needle into the rectal pouch in sitting propped up position. Watersoluble iodine dye is injected. Lateral X-ray is taken to study the level through Stephen line and Kelly's point.
3. U/S abdomen.
4. Evaluation of cardiac function is also important.
5. MRI spine.

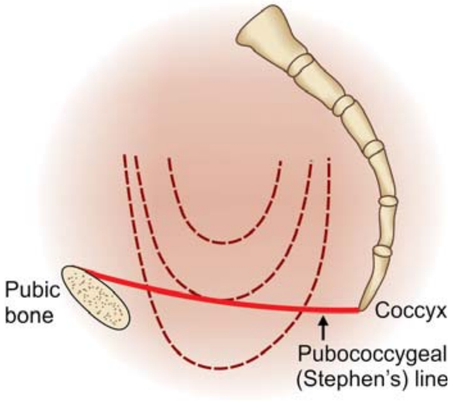

Fig. 25.28: *Stephen's line*.

Treatment

1. In ***low fistula***, single stage reconstruction is done under G/A with very good results.
   i. Anoplasty.
   ii. Anovestibuloplasty.
   iii. Anal dilatation.
   iv. Incision of anal membrane.
2. In ***high fistula***, initial colostomy is done. Later definitive procedure, i.e. *Pull through operation* through puborectalis and anastomosis of rectal pouch to create the anal canal is done. Closure of colostomy is done later.
   - *Posterior sagittal anorectoplasty* is commonly done procedure.

Note:

Level of rectal pouch and normal/abnormal sacrum are the deciding factors for good results.

Complications

- Infection
- Faecal fistula
- Stenosis
- Colitis
- Malnutrition
- Faecal incontinence

PILONIDAL SINUS/DISEASE (Jeep Bottom; Driver's Bottom)

Pilus—hair; Nidus—nest

- It is of infective origin and occurs in sacral region between the buttocks, umbilicus, axilla.
- It is epithelium lined tract, situated short distance behind the anus, containing hairs and unhealthy diseased granulation tissue. It is due to penetration of hairs through the skin into subcutaneous tissue. It forms granuloma/unhealthy granulation tissue in the deeper plane.
- *Types of hair* (H), *force of hair insertion* into subcutaneous tissue (F), *vulnerability of the skin* (V) are the three factors that cause pilonidal sinus. Number of hairs collected, acuteness of root end of hairs, type of hair—tough/silky, shape of hair—straight/curled, scaliness of hair are the deciding features of hair. Cut hairs from above descend into cleft and stay there to get buried deep into pilonidal sinus. Depth, narrowness, friction movements in the natal cleft; soft/macerated skin with erosions, splits, wide skin pores, wounds, presence of moisture and sweat are other factors.
- It is common in hair dressers (seen in interdigital clefts), jeep drivers.
- It is common in 20-30 years of age. It is common in males and mostly affects hairy men.

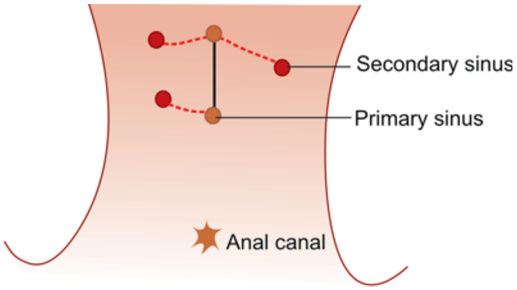

Fig. 25.29: Site and types of pilonidal sinus.

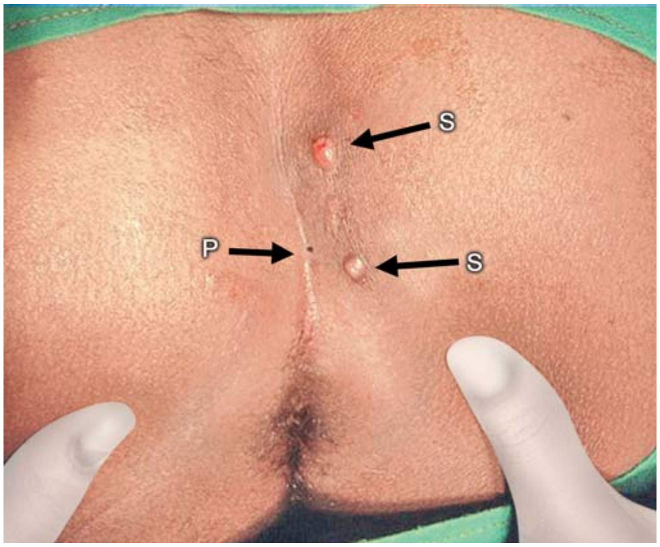

Fig. 25.30: Typical site of pilonidal sinus in the sacral region. Note the primary and secondary sinuses.

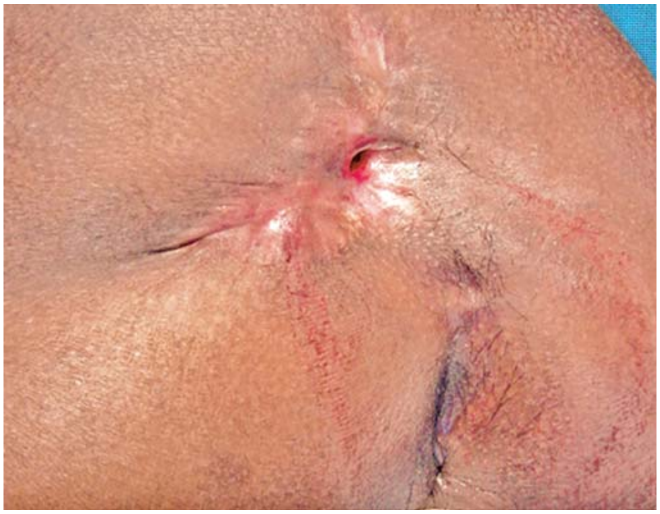

Fig. 25.31: Recurrent pilonidal sinus. 25-40% of pilonidal sinuses can recur.

Commonest site: Interbuttock sacral region.

Pathology

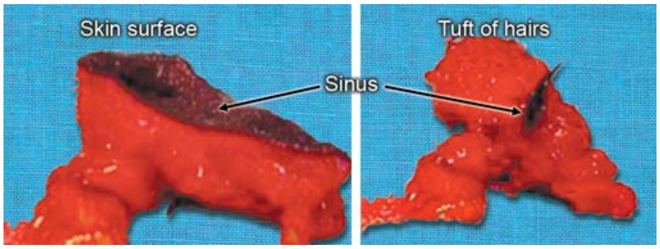

Fig. 25.32: Specimen of typical excised pilonidal sinus. Note the tuft of hair.

Hair penetrates the skin
↓
Dermatitis
↓
Infection
↓
Pustule formation
↓
Sinus formation
↓
Hair gets sucked into the sinus by negative pressure in the area
↓
Further irritation and granulation tissue formation
↓
Pus forms
↓
Multiple discharging sinus

Primary sinus occurs in the midline.

Secondary sinus occurs laterally (paramedian).

Note:

Theories like Preen gland, medullary canal vestige, traction dermoid, inclusion dermoid are no longer accepted. Now it is considered as acquired condition.

Remember about pilonidal sinus

- Congenital theory is no longer considered; it is an acquired entity
- Hair follicles have never been demonstrated in the wall of the sinus (only hairs have found)
- Number, sharpness, nature, shape of hairs; depth and narrowness of the natal cleft; friction movements; nature of the skin, moisture and sweating are the factors predisposing pilonidal sinus
- Hair need not be local, tract always traverses cephalad
- Male preponderance—74%, male sex hormone effect, hairy body, more sweat and maceration
- Occurs in young—20-29 years, who are having active pilosebaceous glands
- Dark haired—stiff hairs, rare in Negroes
- Obese and overweight—deep natal cleft
- Prolonged sitting
- Many procedures for treatment are available with each one having their own advantages and disadvantages

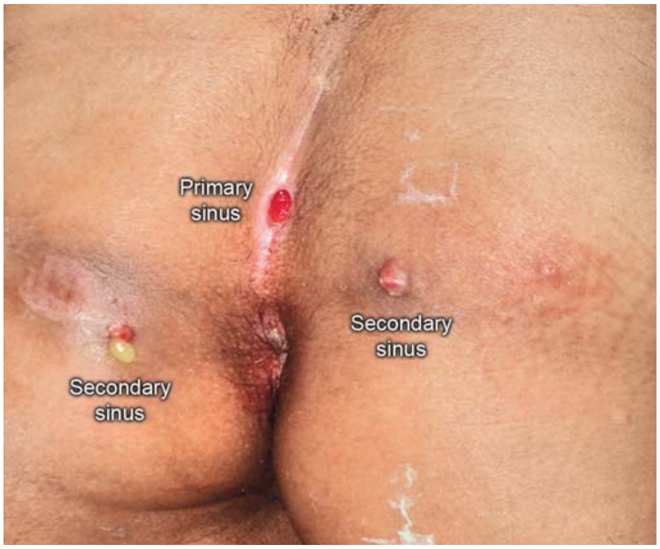

Fig. 25.33: Pilonidal sinus—primary and secondary sinuses are clearly seen.

Clinical Features

- *Discharge*—either sero sanguinous or purulent.
- *Pain*—throbbing and persistent type.
- *A tender swelling* seen just above the coccyx in the midline *(primary sinus*); and on either sides of the midline *(secondary sinus*).
- Tuft of hairs may be seen in the opening of the sinus.
- Presentation may be *as an acute* exacerbation (abscess), or as a *chronic one.*
- It causes recurrent infection, abscess formation which bursts open forming recurrent sinus with pain, discharge and discomfort.

Complications

- Chronic pilonidal sinus can cause occasionally sacral osteomyelitis, necrotising fasciitis and rarely meningitis.
- It is not a life threatening condition but often it can be a morbid disease because of high recurrence rate.

Treatment

In acute phase initially—drainage of the abscess and antibiotics; later definitive treatment is undertaken.

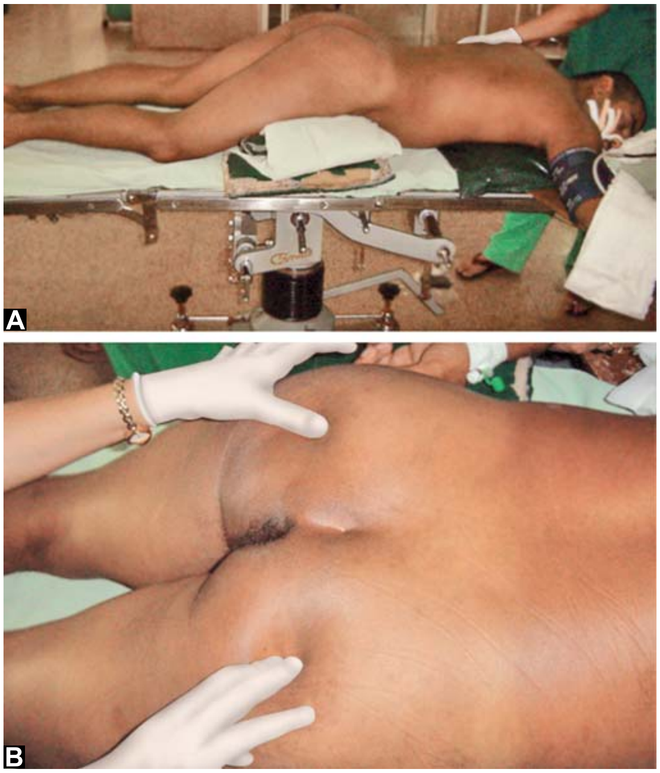

Figs 25.34A and B: Pilonidal sinus—operative jack knife position.

Definitive treatment:

- In prone position (*jack knife* position, i.e. prone with buttocks elevated) excision and primary closure is done under general anaesthesia or local anaesthesia. All sinus tracks, unhealthy granulation tissues with hairs are removed completely. Methylene blue is injected to demonstrate the multiple tracks properly.

Trust your hopes, not your fear.

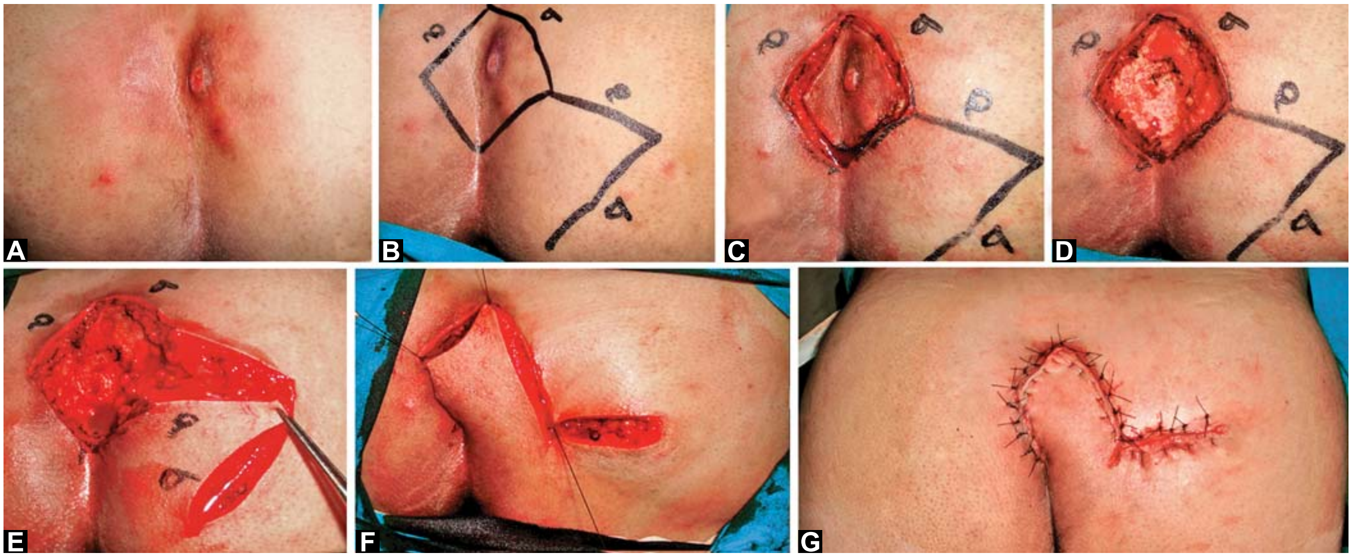

Figs 25.35A to G: Rhomboid-Limberg flap for pilonidal sinus. It gives good result.

- Excision and skin grafting—has got high recurrence rate.
- Excision with Z plasty—good result.
- Excision with multiple Z plasty.
- Karydakis excision through a semilateral incision and lateralised suturing of the wound away from the midline gives good result.
- Excision with closure using Limberg (Rhomboid) buttock flap (single or double rhomboid flaps)—good result.
- V-Y gluteal advancement flap.
- *Bascom technique* of excision through lateral approach is a good method. Through small lateral incision or multiple small lateral incisions 2-4 mm sized sinus is approached and pus is drained; hairs are removed with only minimal excision of sinus done. Cavity walls are not excised. Lateral small wounds are either sutured or left open for spontaneous healing.
- *Lahey and Cattell's relaxing skin incisions* on one buttock to relieve tension on main wound sutures with later closure of secondary wounds by sutures or advancement.
- *Davies and Starr buttock skin flap rotation* into the defect and secondary defect is closed at a later period.
- After excision of entire sinus completely, wound is left open to granulate and heal by epithelialisation with regular dressings.
- *Buie's marsupialisation of the sinus track*—after making incision on the sinus track, edge of the laid opened area is sutured to the skin edge all round using silk or vicryl. This reduces healing time and promotes healing.
- *Lord and Millar's limited excision of primary track* for 0.5 cm depth with removal of tuft of hairs, debris and unhealthy granulation tissue using tiny brush and nylon bristle.
- *Injection of phenol to the track* destroys the epithelium after removal of the tufted hairs. Phenol is allowed to be in contact with epithelium for 3 minutes to create a blanch in the track orifice.

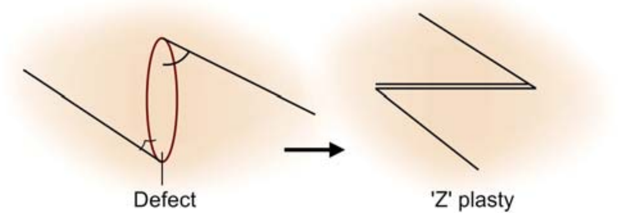

Fig. 25.36: 'Z' plasty is done for pilonidal sinus. It gives good result. Multiple Z plasties are also used.

Nonoperative treatment/prevention of recurrence after surgery:

- Regular shaving of natal cleft to have meticulous hair control. *Laser, depilatory cream, electrolysis* are other methods used.
- Proper perineal hygiene.

Causes for high recurrence rate

- Improper removal
- Overlooking of existing diverticulum
- Entry of new tuft of hairs
- Breakage of scar

Note: Condition has got high recurrence rate (20%)

PILES/HAEMORRHOIDS

> *If bile or phlegm be determined to the veins in the rectum, it heats the blood in the veins: and these veins becoming heated attract blood from the nearest veins, and being gorged the inside of the gut swells outwardly, and the heads of the veins are raised up, and being at the same time bruised by the faeces passing out, and injured by the blood collected in them, they squirt blood, most frequently along with the faeces, but sometimes without faeces.*
>
> —**Hippocrates,** (460-375 BC)

Piles = a ball or mass, Haemorrhoids = blood to ooze, Figs = a fruit (Anjoora).

- The word '*Haemorrhoids*' is derived from Greek word Haima (bleed) + Rhoos (flowering), means bleeding. The *pile* is derived from the Latin word 'Pila' means Ball.
- It is downward sliding of anal cushions abnormally due to straining or other causes.
- *Anal cushions* (Thomson, 1975) are aggregation of blood vessels (arterioles, venules), smooth muscles and elastic connective tissue in the submucosa that normally reside in left lateral, right posterolateral and right anterolateral anal canal.
- Piles can be mucosal or vascular (Graham Stewart, 1963). Vascular type is seen in young; mucosal is seen in old.
- *Present concept* is weakening of Park's ligament which is the lower end of the external sphincter.

Types

Internal—above the dentate line, covered with mucous membrane.
External—below the dentate line, covered with skin.
Interno-external—together occurs.

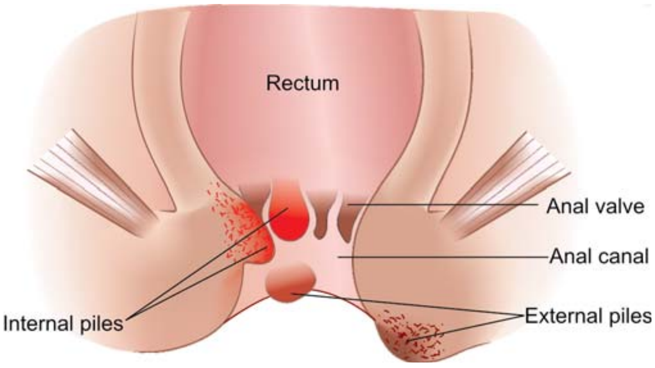

Fig. 25.37: Anatomical locations of internal and external piles.

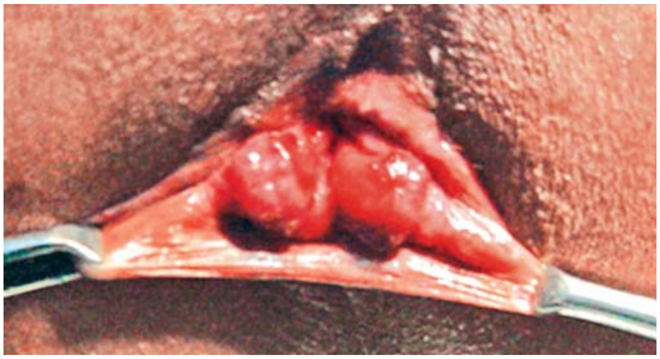

Fig. 25.38: Parts of piles—plum coloured internal part and black cutaneous external component.

Classification I

***Primary haemorrhoids*:** Located at 3, 7, 11 o'clock positions, related to the branches of the superior haemorrhoidal vessel which divides on the right side into two; left side it continues as one.

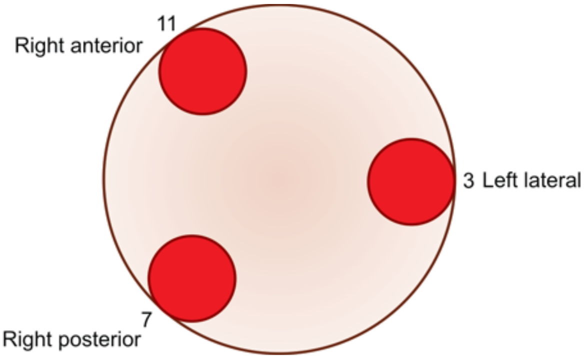

Fig. 25.39: Positions of haemorrhoids.

Secondary haemorrhoids: One which occurs between the primary sites.

Classification II

First degree haemorrhoids
Piles within that may bleed but does not come out

Second degree haemorrhoids
Piles that prolapse during defecation, but returns back spontaneously

Third degree haemorrhoids
Piles prolapsed during defecation, can be replaced back only by manual help

Fourth degree haemorrhoids
Piles that are permanently prolapsed

Piles begin as pedicle and it is located at the origin of the internal pile, i.e. at the level of anorectum.

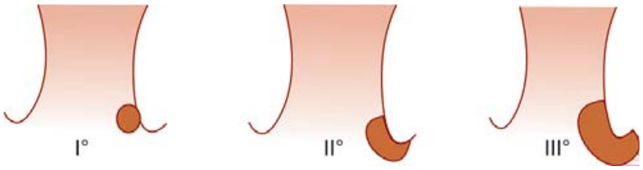

Fig. 25.40: Degrees of haemorrhoids.

Aetiology

- Hereditary.
- Morphological—weight of the blood column without valves causes high pressure. Veins in the lower rectum are in loose submucosal plane, but the veins above enter the muscular layer, which on contraction increases the venous congestion below (more prevalent in patients with constipation). Superior rectal veins have no valves (as they are tributaries of portal vein) and so more congestion.
- Other causes are straining, diarrhoea, constipation, hard stool, low fibre diet, overpurgation, carcinoma rectum, pregnancy, portal hypertension (rare cause). *During pregnancy* factors causing haemorrhoids—raised progesterone relaxes the venous wall and reduces its tone, enlarged uterus compresses the pelvic vein, and constipation is common problem.

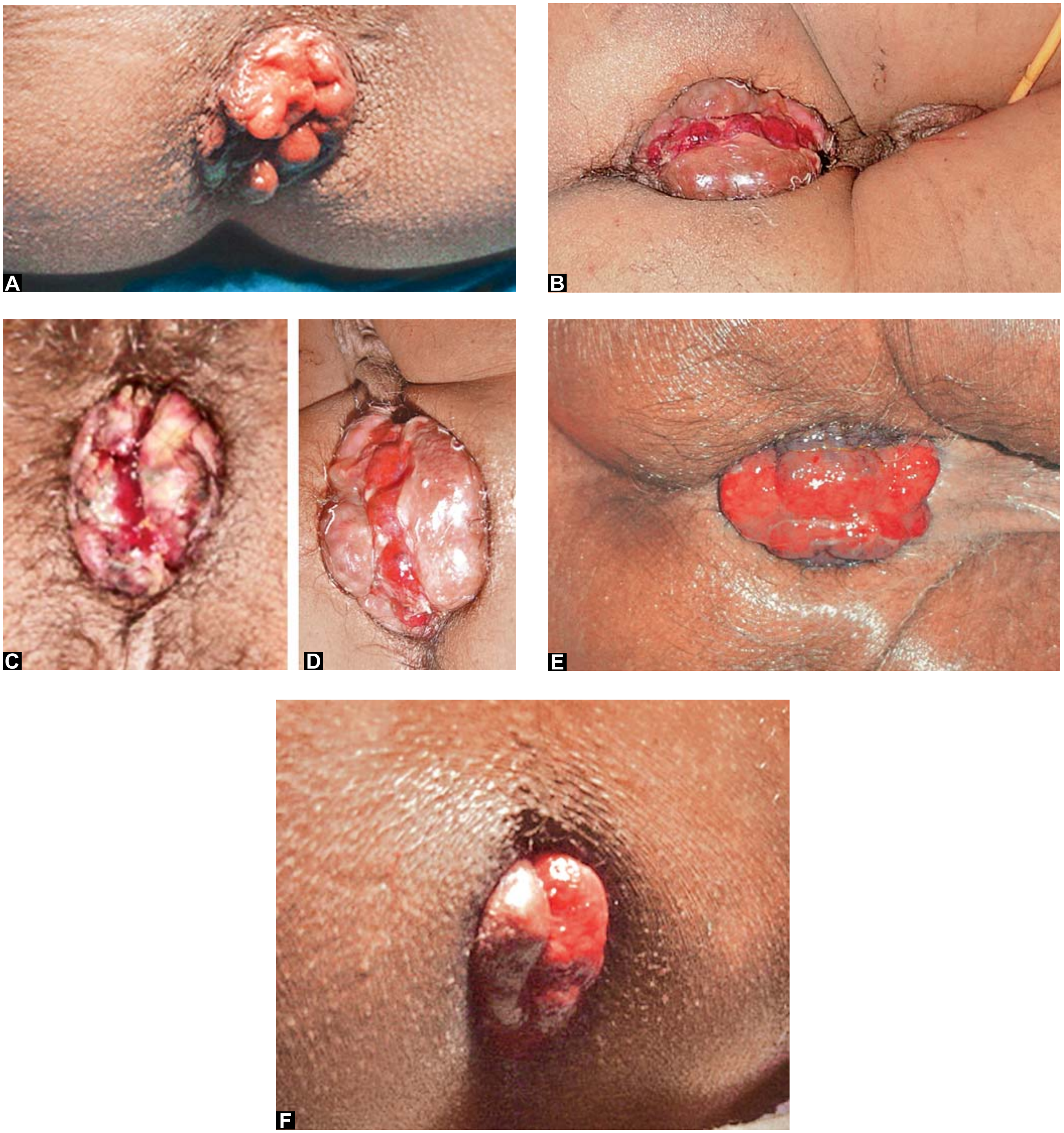

Figs 25.41A to F: Different types of prolapsed piles.

- *Bulging of haemorrhoidal plexus* (*anal corpus cavernosum, by Stelzner*) occurs due to raised luminal pressure and transmission of arterial pressure; pressure in *rectal ampullary pump* (Wannas) during straining raises the portal as well as systemic pressure causing obstruction to venous outflow causing haemorrhoids. *Disruption of suspensory tissues* which hold plexus in position (*sliding lining theory*); raised *basal anal pressure;* unsupported superior haemorrhoidal vein in the loose submucosal connective tissue in the anorectum when passes through the muscular coat gets *constricting effect* leading into congestion of haemorrhoidal plexus—are the *other theories* of haemorrhoid formation.
- *Idiopathic cause:* It is very difficult to pinpoint the cause for production of piles.

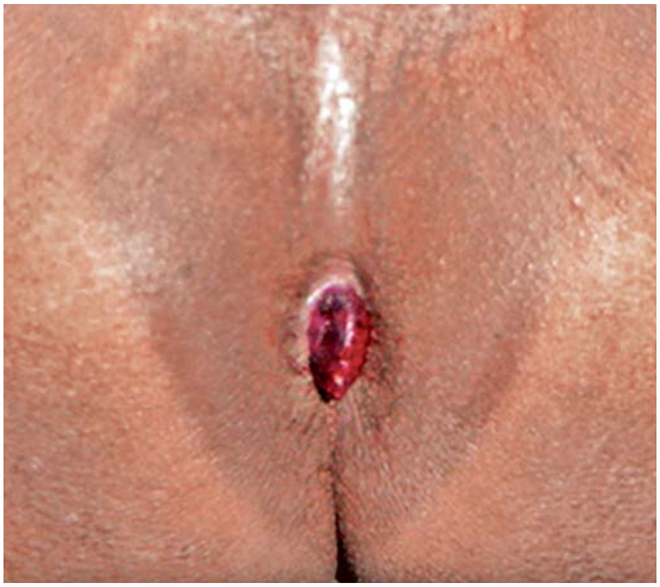

Fig. 25.42: Thrombosed prolapsed pile. Note the colour of the thrombosed pile. It is very painful and tender.

An arterial pile is haemangiomatous condition of superior rectal artery entering the pedicle of internal haemorrhoid which will bleed profusely.

Clinical Features

- The *prevalence rate* of piles is 4.4% in the world, in about 10 million people.
- It may occur at any age but mostly seen in the age between 30 to 65 years.
- Incidence is equal in both the sexes.
- *Bleeding*—1st symptom—'Splash in the pan'—'bright red and fresh'—occurs during defecation.
- Mass per anum.
- Discharge—a mucoid discharge.
- Pruritus.
- Pain—may be due to prolapse, infection or spasm.
- Anaemia—secondary.
- On inspection, prolapsed piles will be visualized.
- On P/R examination, only thrombosed piles can be felt.
- Through *proctoscopy,* exact position can be made out as a bulge into the proctoscope.
- Points to be noted during proctoscopy:
  - The numbers, degrees and size.
  - The surface and appearance of piles.
  - Features chronicity of the prolapse.

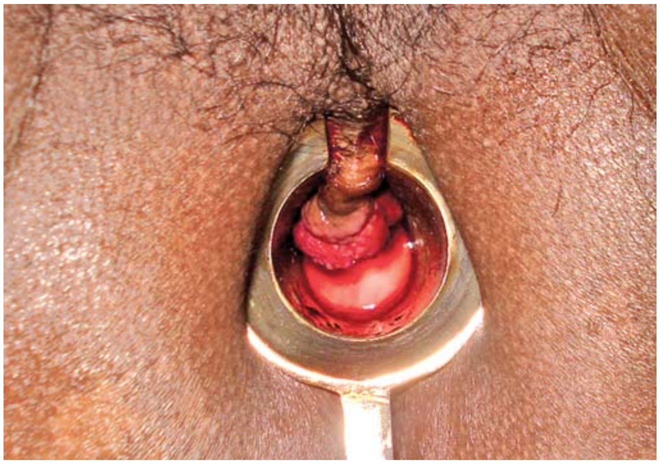

Fig. 25.43: Proctoscopic view of the internal pile (grooved proctoscope).

Causes for bleeding per anum	
Piles	Carcinoma rectum
Fissure-in-ano	Carcinoma colon
Polyps	Diverticulitis
Ulcerative colitis	Intussusception
Amoebic colitis	Vascular anomaly of the colorectum
Fistula-in-ano	Mesenteric ischaemia

  - One should look for other rectal lesion such as external tags, anal papillae and fissure, proctitis.

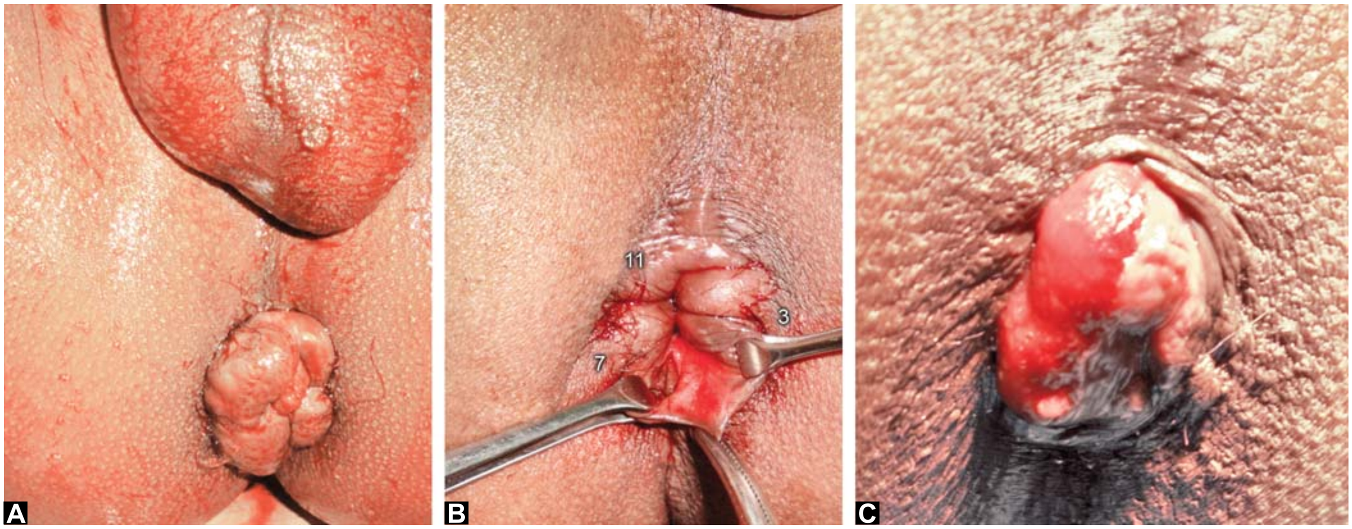

Figs 25.44A to C: Different positions of piles and parts.

We do not see things as they are, we see them as we are.

  - Any gynaecological, genitourinary or abdominal conditions like—carcinoma of rectum, polyps, tumours, features of ulcerative colitis should be identified.
  - Presence of other discharge like blood, pus, mucous.
- Sigmoidoscopy or colonoscopy or barium enema should be done if there is any suspicion of associated malignancy.

Differential Diagnosis

- Carcinoma.
- Rectal prolapse.
- Perianal warts.

Investigations

- Haematocrit.
- Colonoscopy to evaluate proximally for any cause.
- Barium enema X-ray.

Complications

- Profuse haemorrhage which may require blood transfusion.
- Strangulation—piles is being gripped by anal sphincter.
- Thrombosis—piles appear dark purple/black, feels solid and tender.
- Ulceration.
- Gangrene.
- Fibrosis.
- Stenosis.
- Suppuration, leads to perianal or submucosal abscess.
- Pylephlebitis (Portal pyaemia) is rare, but can occur in 3rd degree piles after surgery.

Treatment

Preventive	***Therapeutic/curative***
Diet—more fibre/liquid Laxatives	**Medical**—local applications; sitz bath, diet, laxatives, drugs—analgesics **Parasurgical** • Sclerotherapy • Banding • Cryotherapy • Infrared coagulation (IRC) • Laser therapy • Doppler guided haemorrhoidal artery ligation (DGHAL) **Surgical** • Open haemorrhoidectomy • Closed haemorrhoidectomy • Stapled haemorrhoidopexy • Anal stretching—Recamier, Lord's

1. ***Nonoperative:***

- *Sitz bath*—means the patient has to sit in warm water with the anal region dipped in water for 20 minutes, 2-3 times a day.This reduces the oedema, pain and promotes healing.

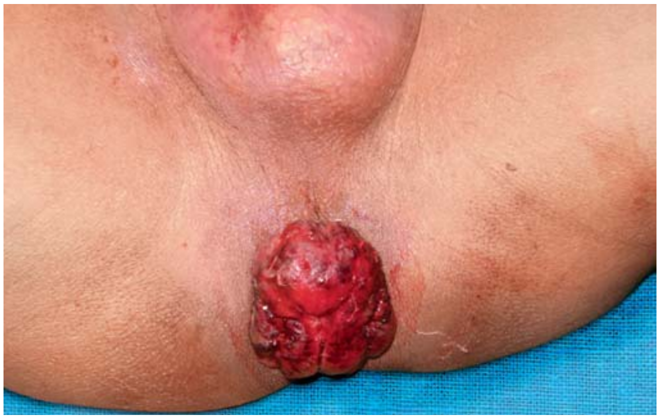

Fig. 25.45: Prolapsed, strangulated piles.

- *Local applications* to reduce pain, itching and oedema can be used.
- Antibiotics, laxatives, anti-inflammatory drugs are beneficial.
- Fibre diet 35 gram/day, plenty of water. *Fibre alternatives* (bulk-forming agents e.g. ispaghula husk, sterculia, methylcellulose) can be used to supplement a high-fibre diet; squatting position may reduce the incidence of piles; plenty of liquid intake.
- *Laxatives* such as lactulose solution which soften bowel motions and relieve the constipation. Bulk laxatives such as psyllium mucilloid, Konsyl or polycarbophil may be necessary.

2. ***In case of inflamed, permanently prolapsed, oedematous piles,*** initially, manual stretching of the anal canal sphincter is tried. This prevents congestion of anal cushions and relaxes the anal sphincter, as a result of which the prolapsed piles gets reduced—*Lord's dilatation (8 fingers)*. Once oedema subsides, in 1-2 weeks, formal procedure is done.

Complications of anal dilatation (3-4 fingers dilatation)

- ❖ Incontinence—rectal—usually temporary
- ❖ Infection
- ❖ Haemorrhage/haematoma
- ❖ Prolapse rectum

3. ***Injection-Sclerosant therapy*:**

- It is done in 1st degree and early 2nd degree piles (internal)—outpatient procedure.
- Using proctoscope and Gabriel syringe, 3-5 ml of 5% phenol in almond oil is injected into the submucosal plane just above the anorectal ring to the pedicle. All three piles can be injected separately—3-5 ml to each site in single sitting. Technique can be repeated after 6 weeks. This technique is not done in the presence of sepsis or prolapse.
- The drug causes fibrosis in the submucosal region (sclerosis leading to mucosal fixation on to deeper planes and occlusion of lakes) and thereby fixation of the anal cushions which do not prolapse, causes strengthening of the vessel wall and obliteration of the vessel lumen.
- It is quick and painless; gives 95% cure rate in 1st degree piles; done on OP basis.

- *Contraindications are*—thrombosed/prolapsed piles, presence of proctitis/fissure/fistula-in-ano, pregnancy and diabetes mellitus.
- Sclerotherapy has not gained popularity, oily solution is difficult to handle and inject.
- *Complications*—recurrence (15%), hypochondriac pain due to entry of drug into the portal system, tenesmus, mucosal sloughing/ulceration, submucosal abscess, anal canal pain, anal stricture.
- Inadvertent deep injection can precipitate pelvic abscess, prostatitis, impotence, rectovaginal fistula.

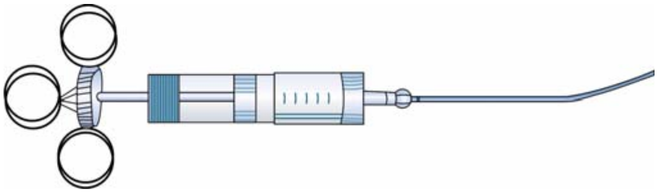

Fig. 25.46: *Gabriel syringe*—It is used for injection sclerotherapy for internal piles using 5% phenol in almond/olive oil (almond/vegetable oil acts as a vehicle which holds phenol for long period of action). It is a stainless steel syringe with two metallic finger brims near the proximal end. One more metallic finger brim is present on the proximal end of the piston to place thumb while injecting.

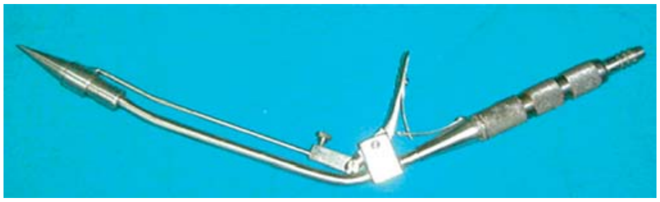

Fig. 25.47: *Barron's band* instrument for 1st and early 2nd degree piles.

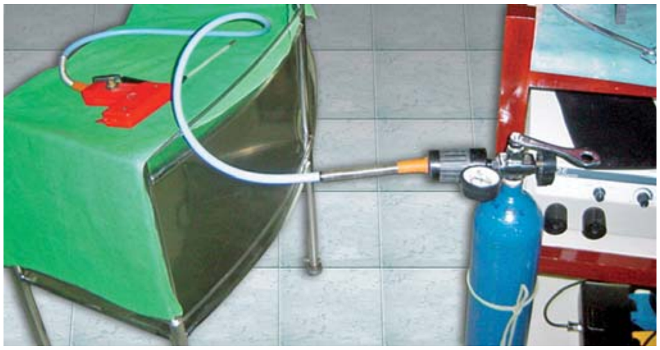

Fig. 25.48: Cryosurgery instrument set up for cryohaemorrhoidectomy.

4. ***Barron's banding:***

- It is done for 2nd degree piles. It causes ischaemic necrosis and piles fall off.
- At one time only two piles can be banded. Repeat banding can be done only after 3 weeks.
- Band should be placed 2 cm above the dentate line.
- Usually 2 bands are used for pile mass to take care of breakage. Tissue sloughs off in 1-2 weeks leaving an ulcer which heals by scarring.
- Equipment is inexpensive, simple to perform; done without anaesthesia on OP basis; results are consistent, stops bleeding and tackles the prolapsing anal cushion.

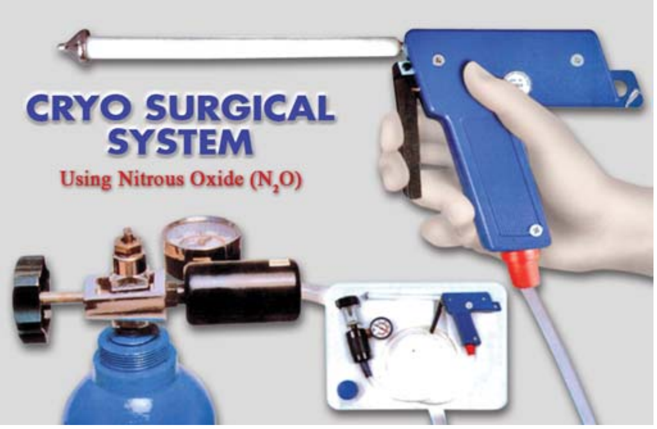

Fig. 25.49: Cryo instrument for cryosurgery.

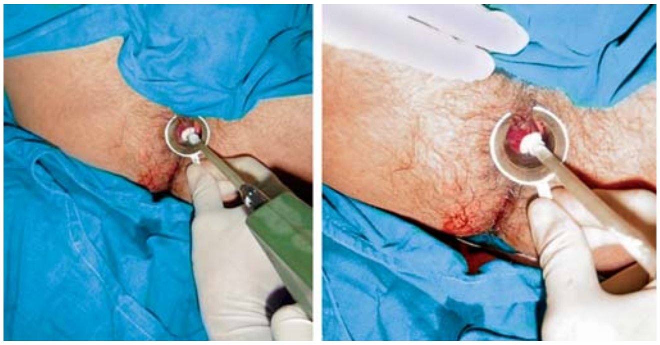

Fig. 25.50: Cryosurgery for piles. Note the probe, grooved proctoscope, freezing the internal pile mass.

- It is contraindicated in fissure/fistula/proctitis.
- *Complications*: if applied low into skin it causes severe pain; discomfort; secondary haemorrhage, ulceration.

Note:

Suction banding is used presently. Suction is used to suck the internal pile into the banding gun.

5. ***Cryosurgery:***

- Using nitrous oxide (–98°) or liquid nitrogen (–196°), extreme cold temperature is used to coagulate and cause necrosis of piles which gets separated and falls off subsequently.
- It is relatively painless and can be done on OP basis.
- All masses can be tackled at one sitting.
- This is carried out with the help of a *cryoprobe*. Nitrous oxide is preferred since it produces adequate freezing. Nitrous oxide delivery system/cryoprobe does not require special rewarming circuit. Liquid nitrogen produces quick destruction and damage of sphincter muscle; also needs special rewarming circuit for its release and is costlier.
- The patient is put in lithotomy position; cryoprobe is applied in the longitudinal axis of internal pile above the dentate line; the pressure must be maintained above 700 lb continuously; the rapid adhesion with freezing (white area formation on piles) occurs; traction and slight rotation in both directions is done to draw entire pile mass

to come in contact with the probe; when entire tissue is frozen in 20-30 seconds, the probe is detached after warming (defrosting/thawing). Procedure is repeated on other pile masses.

- *Advantages*: It is reasonably painless, simple, safe, can be done on OP basis, with less bleeding.
- *Disadvantages:* It causes *profuse watery discharge*, itching, pain if skin is frozen; marked oedema of adjacent skin, incontinence occasionally.

6. ***Infrared coagulation:***

- Here heat is used to burn the piles so as to allow it to fall off.
- Pulses of infrared radiation are applied through a handheld applicator. The specific infrared wavelengths produce chemical changes that cause blood coagulation within the haemorrhoid itself, which causes the haemorrhoid to seal, shrivel, shrink or slough off. Source of infrared rays of 14 volt Wolfram halogen lamp with a gold plated reflector rays are transmitted from fiber optic cable which terminates in a probe or pistol for coagulation.
- In left lateral position, the probe is applied at the base of pedicle above the dentate line and bursts are given in clover leaf fashion. Timer is set at 2 secs giving a depth of 2 mm; total time taken 2-5 minutes. It is done in 1st, 2nd, and 3rd degree piles.
- It produces a discrete area of necrosis (coagulates tissue proteins and evaporates water from the cells) which heals to form a scar; reduces or eliminates blood flow through the haemorrhoid thereby shrinking it and mucosa becomes fixed to the underlying tissue.
- Often 3 or 4 sittings are needed at 1 month intervals.
- It does not cause noncontact coagulation; does not cause interference with electromagnetic devices such as pacemakers. It is *contraindicated* in external pile, proctitis. Long-term results are not good. Equipment is expensive; multiple sessions are needed.

7. ***Laser therapy for piles***—for 3rd degree piles.

- Nd-YAG , diode and carbon dioxide lasers can be used but are expensive and tedious.
- The intense beam of light interacts with tissue and can be used to cut, coagulate or ablate the tissue, sealing off nerves and tiny blood vessels can be done by laser beam. By sealing superficial nerve endings patients have minimum postoperative discomfort. Laser is used for dissecting and excising pile masses.
- It is done for internal haemorrhoids.
- *Advantages*—less operative time; less intraoperative and postoperative bleed and pain; rapid healing; quick recovery; done under LA/SA; less complications; minimal pain, constipation and urinary retention.
- *Disadvantages*—needs skill; sphincter should be taken care of; non contact burning can occur; secondary haemorrhage can occur due to heat tissue destruction; and also injury to sphincter can occur.

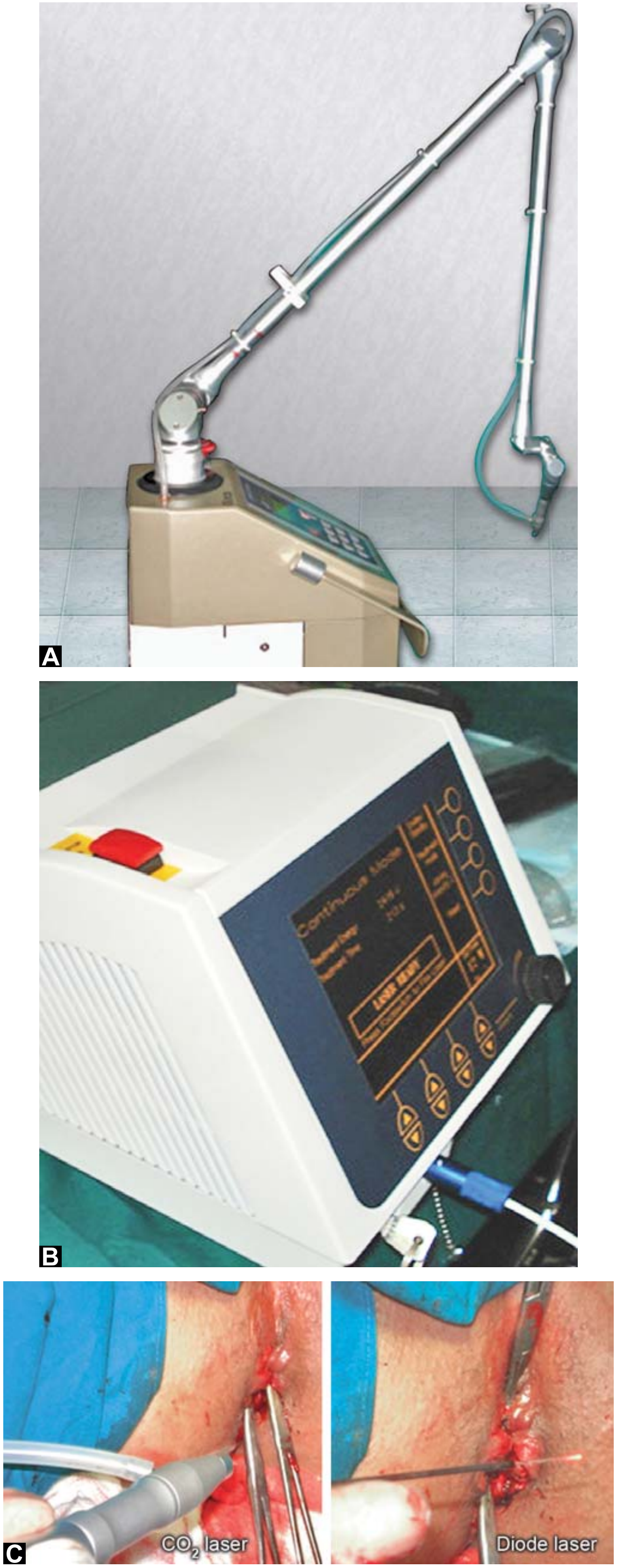

Figs 25.51A to C: Laser surgery for piles—CO_2 laser and diode laser.

8. ***Doppler guided haemorrhoidal artery ligation (DGHAL)***

- DGHAL is an advanced instrument that works under Doppler guided ultrasound. It is painless, 20-minute procedure that cures all degrees of haemorrhoids. It causes choking and blocking of the blood supply of piles. It is done using proctoscope with an incorporated Doppler probe.
- This proctoscope is inserted and used to locate the haemorrhoidal arteries by an audible signal. Once located, a needle holder is inserted into the lumen of the proctoscope and the artery ligated with a 'figure of eight' absorbable suture into the submucosa. The procedure is repeated until no more Doppler signals are identified.
- Advantages: Anaesthesia is not needed; blood loss, pain, residual problems are minimal; done as day care surgery; early return for work; may be safe in diabetic, cardiac, old age patients, and in pregnancy.
- It is under trial and too early to confirm the efficacy.

9. ***Stapled haemorrhoidopexy (Antonio Longo)***

- It is circumferential excision of the mucosa and sub-mucosa 4 cm above the dentate line using circular haemorrhoidal stapler passed per anally (MIPH—minimally invasive procedure for haemorrhoids).
- *Advantages* are—it is less painful; less blood loss; faster recovery; short hospital stay and equally efficacious.
- It is done only for prolapsed piles.
    - This procedure avoids wound in the sensitive peri of anal skin thereby reducing the postoperative pain. Using stapling gun, a unique circular stapler which reduces the degree of prolapsed piles by excising a circumferential strip of mucosa from the proximal anal canal. The strip of mucosa and sub mucosa is excised circumferentially above the dentate line. The veins leading to the haemorrhoids are thus incorporated in this excision. Activation of the gun also simultaneously recovers the cut mucosa and sub mucosa by stapling the edges together.
    - *Disadvantages*—need for experience in advanced surgical skill; costlier; may cause a full-thickness excision of the rectal wall; may injure the anal sphincter. *Improper purse string* can cause incomplete doughnut leading to severe haemorrhage.
    - *Contraindication*: Associated anorectal disease like fissure, fistula in ano.

***Note*:**
Doughnut should be *sent for histology* for muscular layer; proper doughnut should not contain muscular layer.

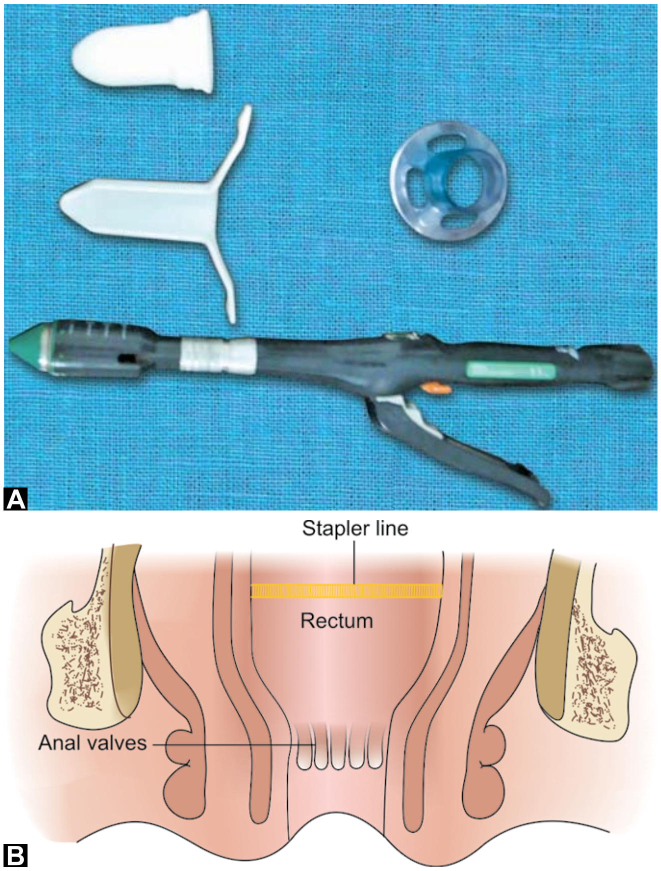

Figs 25.52A and B: Stapler haemorrhoidopexy. It is minimally invasive procedure for haemorrhoids. It is high mucosal rim excision including vascular pedicles using circular (single use) staplers. It is done for prolapsed piles.

10. ***Open operative methods: Still gold standard.***

Indications

- ❖ 3rd degree piles
- ❖ Failure of nonoperative methods
- ❖ Fibrosed piles

Haemorrhoidectomy is the best treatment for haemorrhoids The haemorrhoidectomy is performed using an open or closed technique. The open technique is commonly used in U.K. and is known as *Milligan-Morgan operation*. The *closed technique (Hill Ferguson)* is more popular in U.S.A. Both involve ligation and excision of the haemorrhoid, but in the open technique the anal mucosa and skin are left open to heal by second intention, and in the close technique the wound is sutured. Efficacy of surgery is 95% with 2-5% recurrence rate.

Methods are:

- *Ligation and excision of piles*—commonly done procedures *(Milligan-Morgan) (Open method).*
    - Under anaesthesia, in lithotomy position, initially the sphincter should be dilated to reduce the postoperative pain. Later skin is held with Allis forceps, internal pile is held with artery forceps. Skin is cut in "V' shaped manner and internal sphincter is separated and pushed up. Pedicle is transfixed with vicryl or catgut and distal part is excised. All the three piles can be dealt in a single sitting.
    - Postoperatively**,** *sitz bath*, antibiotics, laxatives, analgesics, local applications are given. Often few finger dilatation of the anal canal is required to prevent stenosis.
- *Submucosal haemorrhoidectomy of 'Parks'*—approach is above the skin through submucosal plane.
- *Hill-Ferguson closed method*: Here patient in prone position, under GA/caudal anaesthesia, retraction is done using Hill-Ferguson retractor. Incision is made around pile mass, pedicle is dissected to its proximal base; it is ligated with transfixation using 2 zero vicryl or silk; mucosa and anal skin is sutured using 3 zero vicryl/dexon after proper haemostasis using cautery.

One cannot love what he cannot respect, whether it be himself or another.

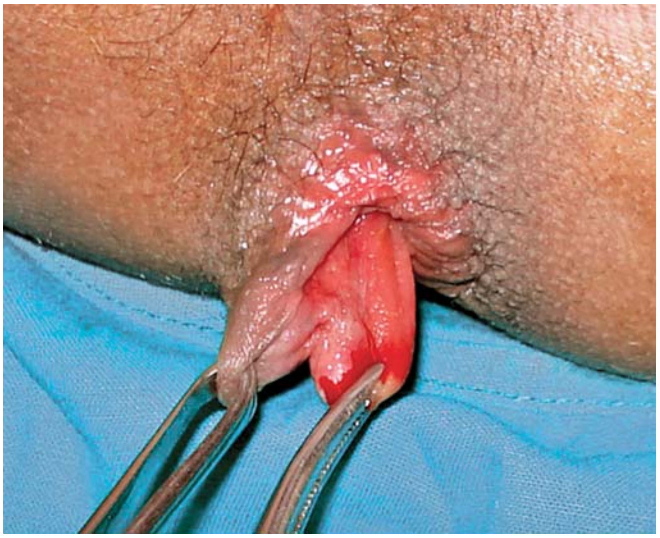

Fig. 25.53: During haemorrhoidectomy, skin part is held with Allis forceps; internal pedicle is held with artery forceps. A 'V' cut is placed over the outer skin up to mucocutaneous junction. Dissection is deepened to visualise the internal sphincter. Once pedicle is dissected, it is transfixed using vicryl suture material. Distal tissue is excised. Technique is repeated on other sites also.

Postoperative complications

- Pain—due to spasm, nerve irritation, muscle injury
- Retention of urine—commonest—50%
- Reactionary or secondary haemorrhage
- Anal stricture
- Anal fissure
- Recurrence
- Anal discharge for sometime
- Incontinence for faeces or gas
- Ectropion (Whitehead deformity)

11. ***Management of strangulated/thrombosed/gangrenous piles***: Here initially conservative treatment is done using warm water sitz bath; antibiotics; elevation; bed rest; saline compression dressing; analgesics. This reduces the oedema and piles shrink. Later in 4-5 days haemorrhoidectomy is done. Doing haemorrhoidectomy immediately may precipitate portal pyaemia and also increases risk of developing anal stricture.
12. ***Newer methods***: Using ultrasound or controlled electric energy (Harmonic scalpel or ligasure), haemorrhoidectomy can be done with less postoperative pain. But tissue charring may precipitate secondary haemorrhage.

EXTERNAL PILES

Causes

- As a part of internal piles.
- Sentinel pile associated with anal fissure.
- Anal skin tags.

Treatment

- The cause is treated.
- Sitz bath.
- Excision.

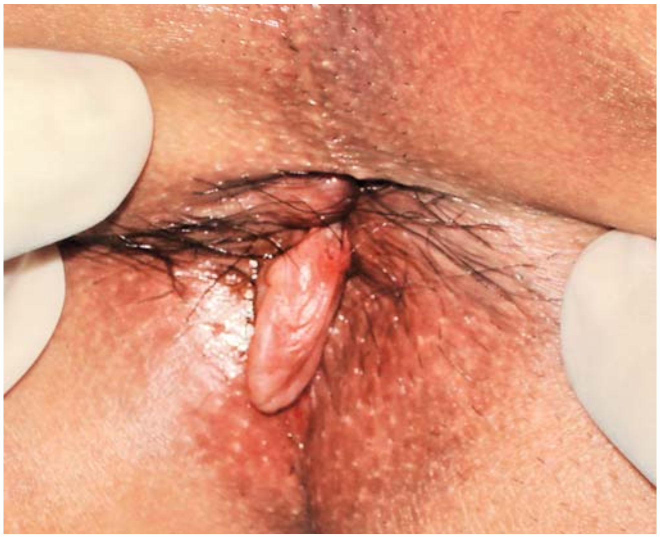

Fig. 25.54: Typical external pile. It causes haematoma, abscess, pain, itching.

Problems

- Pruritus ani.
- Perianal haematoma.
- Perianal abscess formation.

ANAL FISSURE (FISSURE-IN-ANO)

- It is an ulcer in the longitudinal axis of the lower anal canal.
- Commonly it occurs in the *midline, posteriorly (more common in males),* but can also occur in the midline anteriorly (more common in females)**.**
- 95% of anal fissures in men are posterior; 5% are anterior. 80% of anal fissures in females are posterior; 20% are anterior. Anterior anal fissure is common in females.
- It is superficial, small but distressing lesion.
- Fissure ends above at the *dentate line***.**

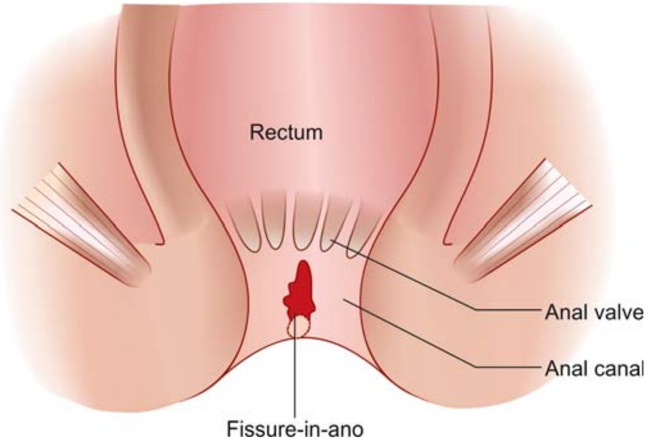

Fig. 25.55: Anatomy of anal canal with fissure-in-ano. It is commonly posterior.

Internal Sphincter

- The internal sphincter is formed from a thickening of the smooth muscle of circular coat of upper end of anal canal.

Circular muscle is the continuation of the inner coat of the rectum. This involuntary muscle commences where the rectum passes through the pelvic diaphragm and ends just within the anal orifice, where its lower border can be felt.

- The internal anal sphincter is 2.5 cm long and 2 to 4 mm thick. The internal sphincter is closed by a sheath of striped muscle.
- Spasm and contracture of this muscle play a major part in fissure and several other anal infections.

External Sphincter

It can be divided into three parts—deep, superficial and subcutaneous portion. It is considered to be one muscle.

- Deep part encircles the upper end of anal canal and has no bony attachment.
- Superficial part is attached posteriorly to the coccyx, anteriorly inserted into the mid-perineal point in the male, in female it fuses with the sphincter vagina.
- Subcutaneous part encircles the lower end of the anal canal and has no bony attachment.

Causes

- Because of the curvature of the sacrum and rectum, hard faecal matter while passing down causes a tear in the anal valve leading to posterior anal fissure.
- Anterior anal fissure is common in females due to lack of support to pelvic floor.
- Hard stool; diarrhoea; *increased sphincter tone; local ischaemia;* trauma; sexually transmitted diseases.
- *Other causes*—haemorrhoidectomy, Crohn's disease, venereal disease, ulcerative colitis, tuberculosis.

Types

- Anal fissure can be *acute or chronic*.

Acute Anal Fissure

- It is a deep tear in the lower anal skin with severe sphincter spasm without oedema or inflammation.
- It presents with *severe pain and constipation*.

Chronic Anal Fissure

- It has got inflamed, indurated margin with scar tissue.
- Ulcer at its inferior margin is having a skin tag which is oedematous, acts like a guard—*sentinel pile.*
- Proximally *hypertrophied anal papilla* is observed.
- It can cause repeated infection—fibrosis—abscess formation—fistula formation.
- Chronic fissure is less painful than acute one.
- Multiple fissures are seen in inflammatory bowel disease, homosexuals and venereal diseases.
- Chronic fissure can cause *complications* like—abscess, fistula formation.

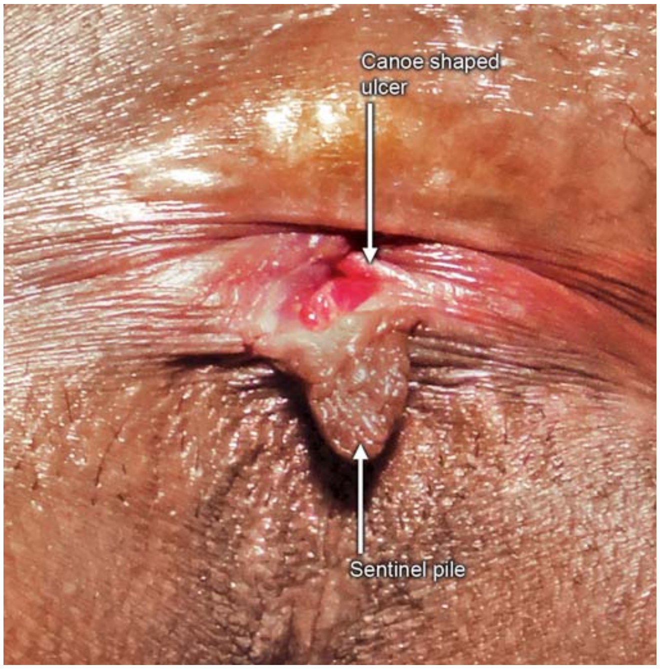

Fig. 25.56: Chronic fissure-in-ano with ulcer and sentinel (pile) skin tag.

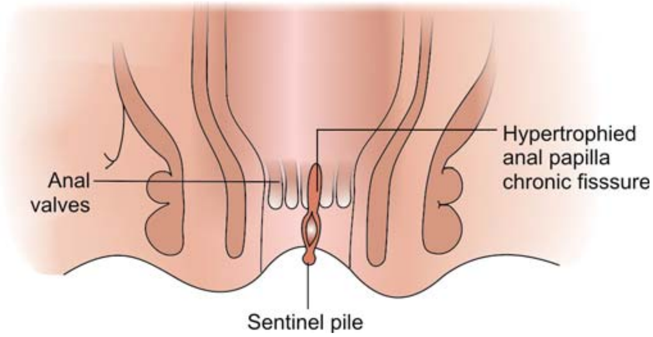

Fig. 25.57: Parts of chronic fissure-in-ano. Ulcer in the lower anal canal; sentinel pile below and hypertrophied papilla above.

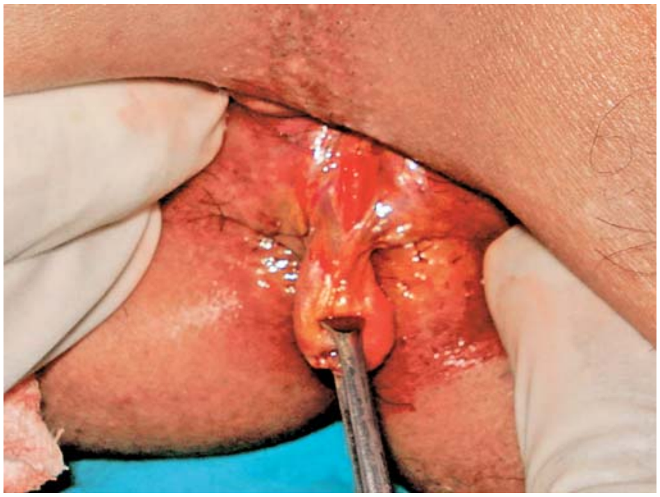

Fig. 25.58: Fissure-in-ano with sentinel pile.

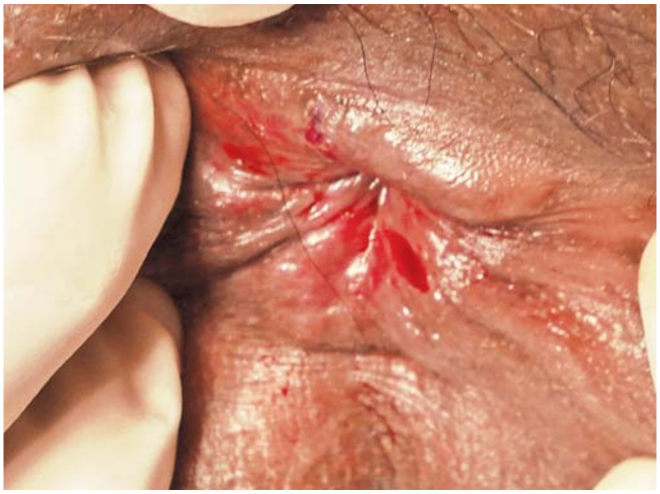

Fig. 25.59: Lateral fissure-in-ano.

Clinical Features

- Common in middle aged women, not in elderly.
- Pain is severe in nature in acute type, whereas less severe in chronic.
- Constipation, bleeding and discharge.
- P/R examination and proctoscopy is not possible in acute fissure-in-ano. General anaesthesia is required for examination.
- In chronic fissure, ulcer is felt with button like depression, induration and often sentinel pile.

Differential Diagnosis

- Carcinoma anal canal.
- Inflammatory bowel disease.
- Venereal diseases.
- Anal chancre (painful).
- Tuberculous ulcer.
- Proctalgia fugax.

Treatment

General measures for anal fissure

- Adequate fluid intake (6-8 glasses of liquids)
- Fiber rich diet (vegetables, fruits, brown rice)
- Bulk forming agents (psyllium husk, bran)
- Stool softeners (lactulose)
- Local anaesthetic agents (lignocaine 5%)
- Sitz bath
- Avoid constipation
- Once recovers, regular anal dilatation

***In an acute* case**

- *Lord's dilatation* is done under G/A to relax the sphincter. It is the *manual dilatation* (Lord, 1969) of the anus under general anaesthesia with relaxation using *four fingers of each hand* (*8 fingers*) to cause vigorous stretching of the anal canal to break the circular constricting band in the wall of the anorectum.
- Later, use of laxatives, xylocaine surface anaesthetic application, and anal dilatation with finger can be carried out for certain period**.**
- Bed rest; 2% nifedepine ointment.
- *Stretching of the anal sphincter* (Recamier, 1829) using *two fingers of each hand* (*4 fingers*) under anaesthesia is also an alternative one. It is better than Lord's dilatation as complications are less.

For chronic fissure

- *Dorsal fissurectomy with sphincterotomy* is done under anaesthesia. Specimen should be sent for biopsy to rule out carcinoma, tuberculosis, etc. Here transverse fibres of internal sphincter is divided in the floor of the fissure.
- *Lateral anal sphincterotomy*
  - Here internal sphincter is divided partially away from the fissure either in right or left lateral positions (also gives a good result).
  - Here *closed or open methods* (Notaras) are used. Sphincterotomy is done below the dentate line. In *closed method* no 11 blade is inserted into the intersphincteric groove to pass upwards. Blade is moved medially to cut lower 1/3 or 1/2 of the internal sphincter. In *open method* skin is incised laterally, external to anal verge. Hypertrophied band of lower part of internal sphincter is dissected and divided. Wound is left open.
  - Haematoma, perianal abscess, bruising, fistula, incontinence are the complications of lateral sphincterotomy.
- Topical nitroglycerine 0.2% is also used to relax the sphincter. It causes severe headache.
- Botulinum toxin 25 units injected into the internal sphincter. It causes temporary denervation of the internal sphincter; reducing its tone, improving the blood supply and control of ischaemia. It causes temporary incontinence for flatus (10%).
- Diltiazem (2%), L arginine are the other agents used.
- Regular anal dilatation is also often important to prevent recurrence.

SENTINEL PILE

('sentinel' means guard)

- It is commonly associated with *Fissure-in-ano* of chronic type wherein, in the lower part of fissure, skin enlarges and appears like guarding the fissure.

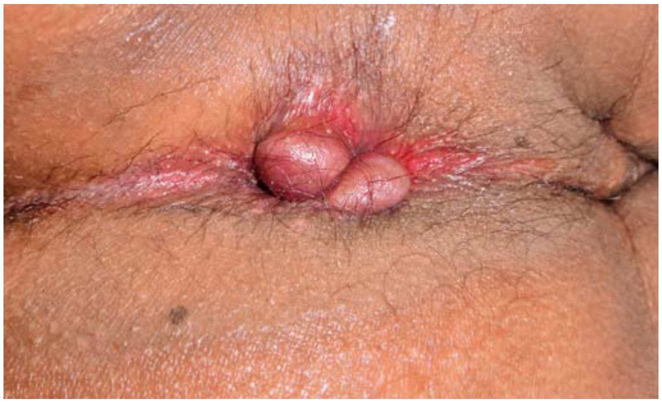

Fig. 25.60: Perianal haematoma—it should be drained to remove clot. Small haematoma may resolve on its own. Larger one if not drained will form an abscess.

- It can cause perianal haematoma, abscess formation, discomfort.
- The chronic fissure is treated along with excision of the sentinel pile.
- There may be low grade *infection* and *lymphatic oedema*. Haematoma/abscess can develop in it.

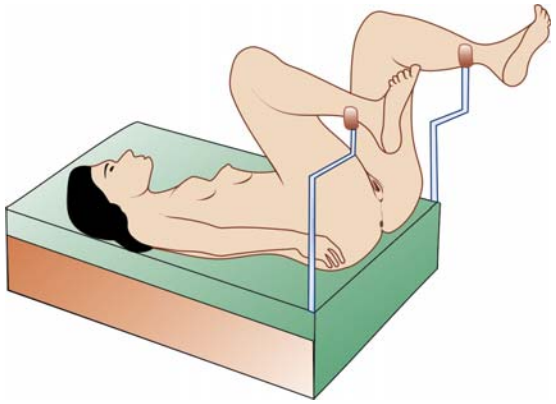

Fig. 25.61: Lithotomy position used for all perineal surgeries like for fissure, piles, fistula, and APR.

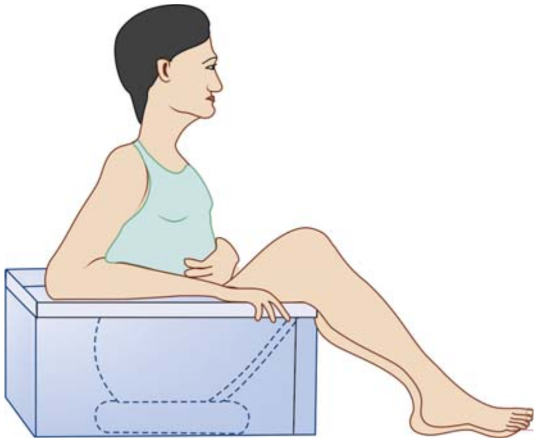

Fig. 25.62: *Sitz bath* is used after perineal surgeries like for piles, fissure, and fistula. Patient sits in warm water bath with perineum dipped in the water. Often small quantity of antiseptics or $KMNO_4$ is added.

ANORECTAL ABSCESS

- Most common causative organism is *E. coli (60%)*. Others are *Staphylococcus, Bacteroides, Streptococcus, B. proteus*. Commonly occurs due to infection of anal gland in perianal region. 95% of anorectal abscesses are due to infection of anal glands in relation to crypts—*cryptoglandular* disease.
- Common in diabetics and immunosuppressed.

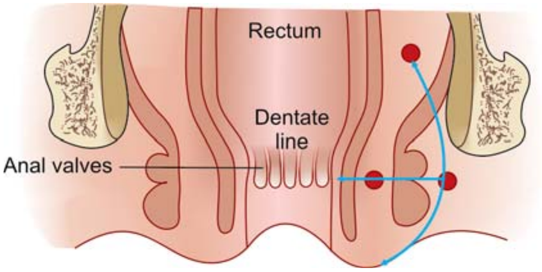

Fig. 25.63: Anatomy of anorectal abscess.

- **Other causes:**
  - Injury to anorectum.
  - Cutaneous infection (e.g. Boil).
  - Blood born infections.
  - Many anorectal abscesses are associated with anal fistulas.
  - Fissure-in-ano.
  - Perianal haematoma.
  - Post anorectal surgery.
  - Crohn's disease.
  - Tuberculosis.

Differential diagnosis of anorectal abscess

- Periurethral abscess
- Bartholin abscess
- Tuberculous abscess

Investigations

- MRI is the *investigation* of *choice* for anorectal abscess.
- *Perineal and anal US* is also very useful.
- Investigations relevant to specific cause may be done.
- Proctosigmoidoscopy is needed to identify secondary cause in anorectum.

Classification

1. Perianal.
2. Ischiorectal.
3. Submucous.
4. Pelvirectal.
5. *Fissure abscess* (in relation to fissure-in-ano).

Perianal Abscess (60%)

- This usually results due to suppuration of anal gland or suppuration of thrombosed external pile or any infected perianal condition.
- It lies in the region of subcutaneous portion of external sphincter.

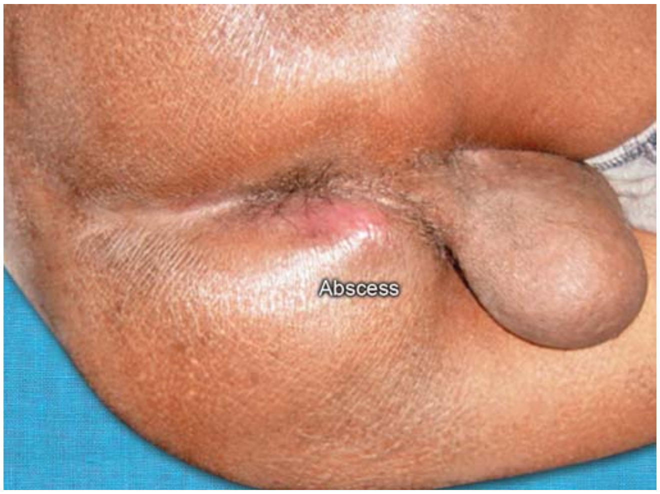

Fig. 25.64: Perianal abscess. Note the swelling and inflammation.

Our senses don't deceive us; our judgement does.

Clinical Features

- Severe pain in perianal region with difficulty to sit.
- Tender, smooth, soft swelling in the region.

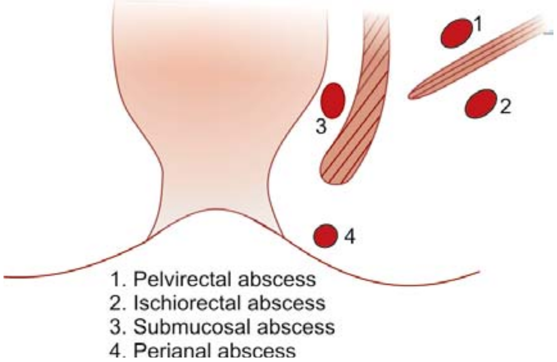

Fig. 25.65: Sites of anorectal abscess.

Treatment

- Sitz bath, antibiotics, analgesics, local application of anaesthetic agents and laxatives.
- Drainage under G/A.

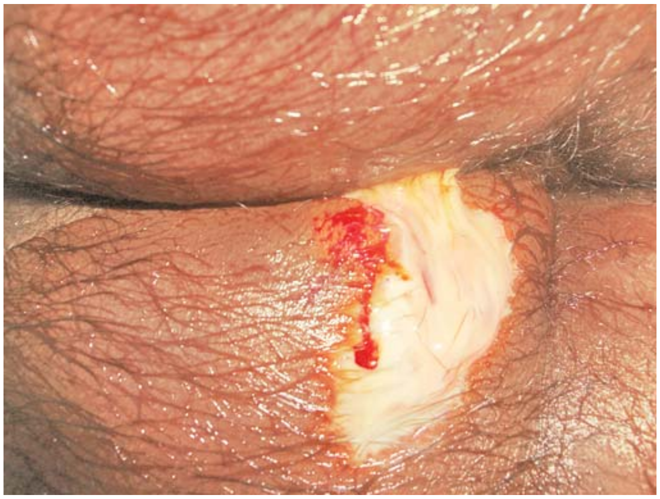

Fig. 25.66: Perianal abscess showing pus oozing out.

Ischiorectal Abscess (30%)

Surgical Anatomy

Ischiorectal fossa (pyramidal shape 5 cm depth and 2 cm width) lies between anal skin and levator ani. Right and left communicates with each other. Laterally, it is related to fascia covering obturator internus; medially to levator ani and external sphincter; posteriorly sacrotubercous ligament and gluteus maximus; anteriorly urogenital diaphragm; below, the floor by skin. Above it is related to lunate fascia and pudendal neurovascular bundle in pudendal canal (Alcock's canal).

- Commonly it is due to extension of low intermuscular anal abscess, laterally through external sphincter.
- But often it can be blood or lymphatic born.
- Fat in the fossa is more prone for infection because it is least vascularised.

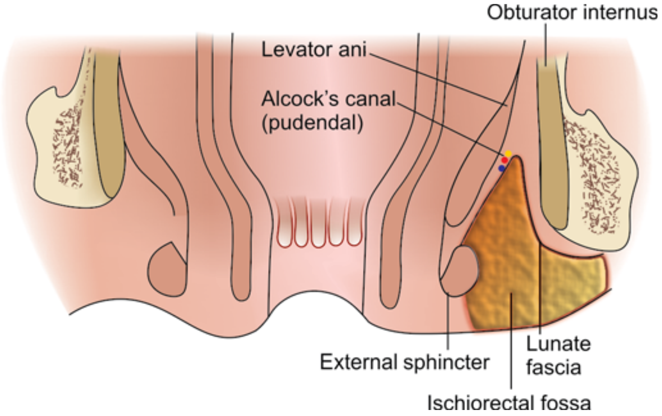

Fig. 25.67: Anatomy of ischiorectal fossa. Note the boundaries and Alcock's canal.

- Fossa communicates with that of opposite side through postsphincteric space and so *horse-shoe like* abscess can occur.
- It presents with tender, indurated, brawny swelling in the skin over the ischiorectal fossa with high fever.
- *Swelling is not well-localised and fluctuation is absent in ischiorectal abscess.*

Treatment

Under G/A in lithotomy position, through a cruciate incision a portion of skin is excised (de roofing) and pus is drained. Pus is sent for C/S and presence of any internal opening to rectum should be looked for (for possibility of an existing fistula).

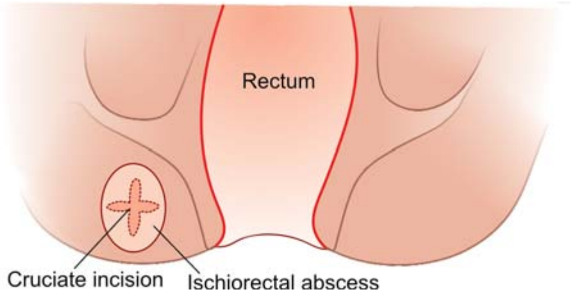

Fig. 25.68: Cruciate incision is used for drainage of ischiorectal abscess.

Submucous Abscess (5%)

- It occurs above the dentate line, which can be drained with sinus forceps, through a proctoscope.
- Aching pain in the anorectum with significant perineal discomfort.
- On digital examination (P/R), tender, soft, smooth swelling in the lower rectum and anal canal.
- It may be missed clinically as there is no obvious swelling externally.
- *Treatment* is proper antibiotics; incision and drainage under general anaesthesia.

Pelvirectal Abscess

- It is situated between the upper surface of levator ani and pelvic peritoneum. It is almost like a pelvic abscess,

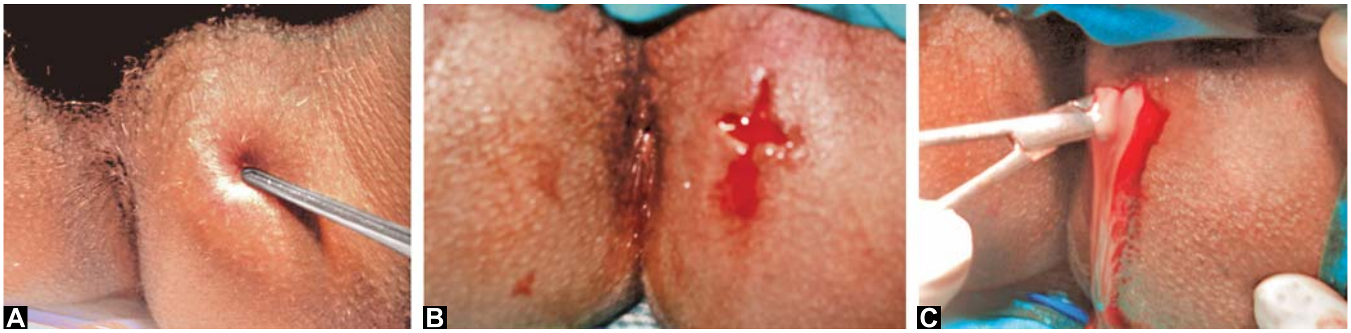

Figs 25.69A to C: Ischiorectal abscess drainage done under general anaesthesia. Note the cruciate incision placed and pus gushing out.

occurs secondary to appendicitis, salpingitis, diverticulitis, Crohn's.

- U/S abdomen is done to rule out the above factors.
- Treated accordingly, after thorough investigations for diabetes, Crohn's and other conditions.

Problems with anorectal abscess

- Recurrent abscess formation
- Fistula formation

FISTULA-IN-ANO

- It is a track lined by granulation tissue which connects perianal skin superficially to anal canal; anorectum or rectum deeply.
- It usually occurs in a pre-existing anorectal abscess which burst spontaneously.

Fistula-in-ano can be:

- Cryptoglandular—90%.
- Non cryptoglandular (other causes)—10%.

Cryptoglandular Hypothesis

- The intersphincteric space is the surgical plane between the internal and external sphincters and is found between the longitudinal muscle and external sphincter, where it exists as a sheet of fat containing loose areolar tissue. The fat-filled ischioanal fossa lies lateral to the sphincter complex and is traversed by a network of fibroelastic connective tissue. Proximal half of the anal canal is characterized by longitudinal mucosal folds, the anal columns of Morgagni. The distal aspect of each column is linked to its neighbour by a small semilunar fold (the anal valves), which in turn forms small pockets (the anal sinuses, or *crypts of Morgagni*). The distal undulating limit of these valves is the dentate (pectinate) line, which also marks the most distal aspect of the anal transitional zone, a histologic junction between anal squamous and rectal columnar epithelium. The dentate line lies 2 cm proximal to the anal verge and is important landmark in fistula-in-ano because the anal glands empty into the crypts that lie proximal to the valves. These glands secrete mucus to lubricate anus, and are the source of infection. These glands present in the subepithelium, internal sphincter, and two-thirds of these glands are located within the intersphincteric space. It is the infection of these intersphincteric glands that initiates the fistula in ano, known as the "*cryptoglandular hypothesis*".
- These glandular infection leads into an intersphincteric abscess due to blockage of the draining duct by infected debris. This abscess may resolve by spontaneous drainage into the anal canal or may progress to an acute anorectal abscess. Treatment of this abscess is incision and drainage; but source of infection in the intersphincteric space persists, leading into development of a fistula in ano. Acute anorectal abscess and fistula in ano are believed to be acute and chronic manifestations, respectively of the same disease.
- While most fistulas start as a *simple single primary* tract, recurrent infection eventullaly causes formation of *extensions (secondary tracts)*. Extensions may be intersphincteric, ischioanal, or supralevator (pararectal). The ischioanal fossa is the commonest site for an extension. Extensions also occur in the horizontal plane known as "horseshoe" if there is ramification on each side of the internal opening.

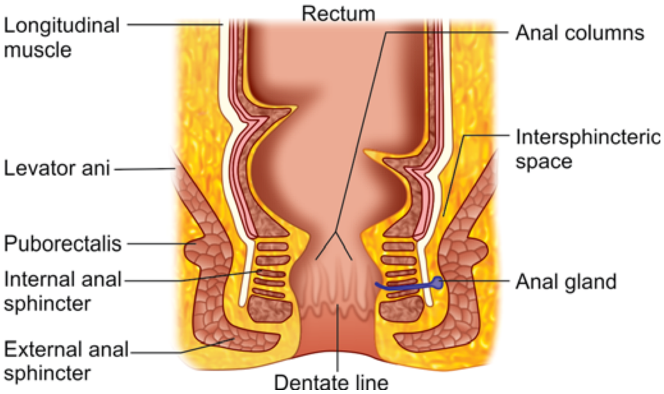

Fig. 25.70: Anatomy of anal glands (Cryptoglands of Morgagni).

Other causes are (Non-cryptoglandular)

- Tuberculosis
- Carcinoma
- Crohn's disease
- Ulcerative colitis
- Lymphogranuloma venereum
- Hydradenitis suppurativa
- Traumatic

Plenty of lymphoid aggregates surround the anal glands, which explain the high incidence of anal fistula in *Crohn's* disease.

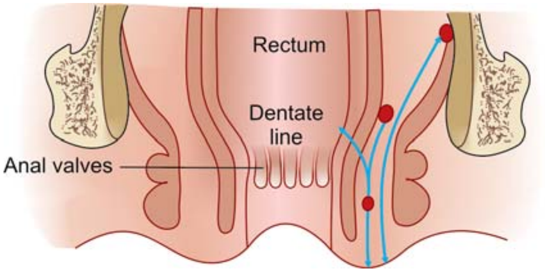
Rectum
Dentate line
Anal valves

Fig. 25.71: Anatomy of fistula-in-ano.

Classifications

Standard (Milligan Morgan, 1934; Goligher 1975)	***Park`s classification (1976)***
• Subcutaneous commonest • Low anal—common • Submucous • High anal • Pelvi rectal	• Intersphincteric – commonest 70% • Transphincteric 25% • Supralevator/ suprasphincteric 4% • Extrasphincteric 1%

It can be:

1. ***Low level fistulas****—these open into the anal canal below the internal ring.*
2. ***High level fistulas****—these open into the anal canal at or above the internal ring.*

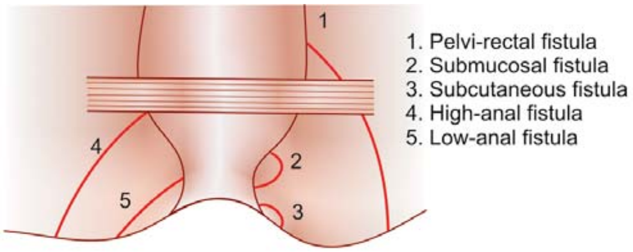
1
2
3
4
5
1. Pelvi-rectal fistula
2. Submucosal fistula
3. Subcutaneous fistula
4. High-anal fistula
5. Low-anal fistula

Fig. 25.72: Classification of fistula-in-ano (*Standard classification).*

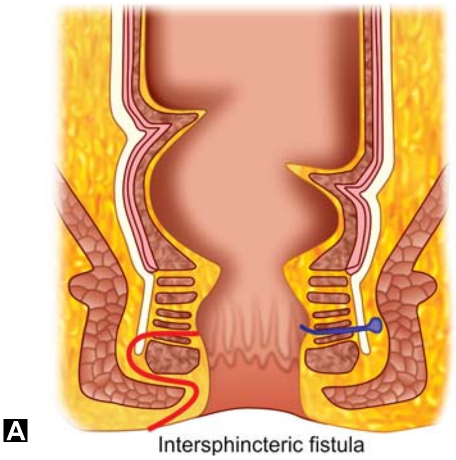
A
Intersphincteric fistula

Fig. 25.73A

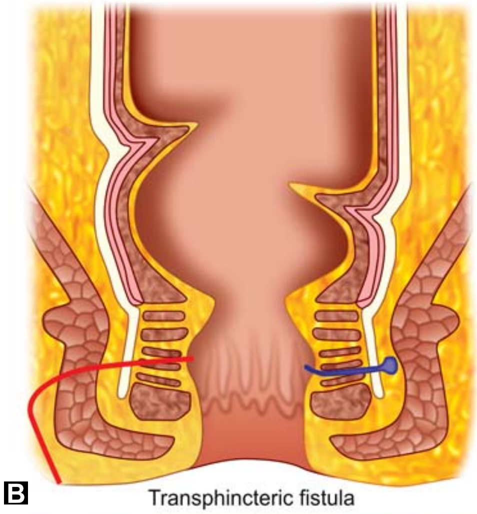
B
Transphincteric fistula

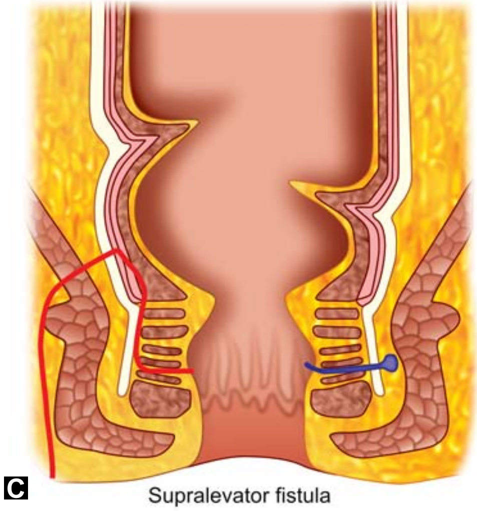
C
Supralevator fistula

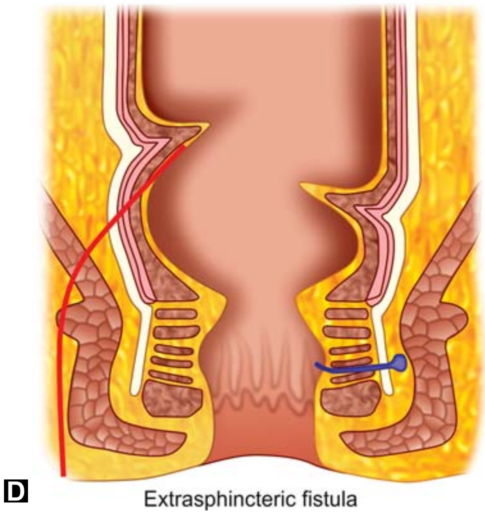
D
Extrasphincteric fistula

Figs 25.73B to D

Figs 25.73A to D: *Park's classification* of fistula-in-ano—(A) Intersphincteric fistula. (B) Transphincteric fistula. (C) Supralevator fistula. (D) Extrasphincteric fistula.

It can be:

- *Simple fistula* without extensions.
- *Complex fistula* with extensions.

It can be with:

- *Single external opening*.
- *Multiple external openings* which are often seen in tuberculosis, ulcerative colitis, Crohn's disease, LGV, hidradenitis suppurativa, actinomycosis.

LOW-LEVEL FISTULAS

Clinical Features

- It has a prevalence of 0.01% and is common in young adult males (2:1, male to female).
- It presents with seropurulent discharge (65%), along with skin irritation and one or more external opening may be present with induration of the surrounding skin.
- Often it may heal superficially but pus may collect beneath forming an abscess which again discharges through same or new opening.
- Ischiorectal fossa on each side, most often communicates with each other behind the anus causing *horseshoe* fistula.

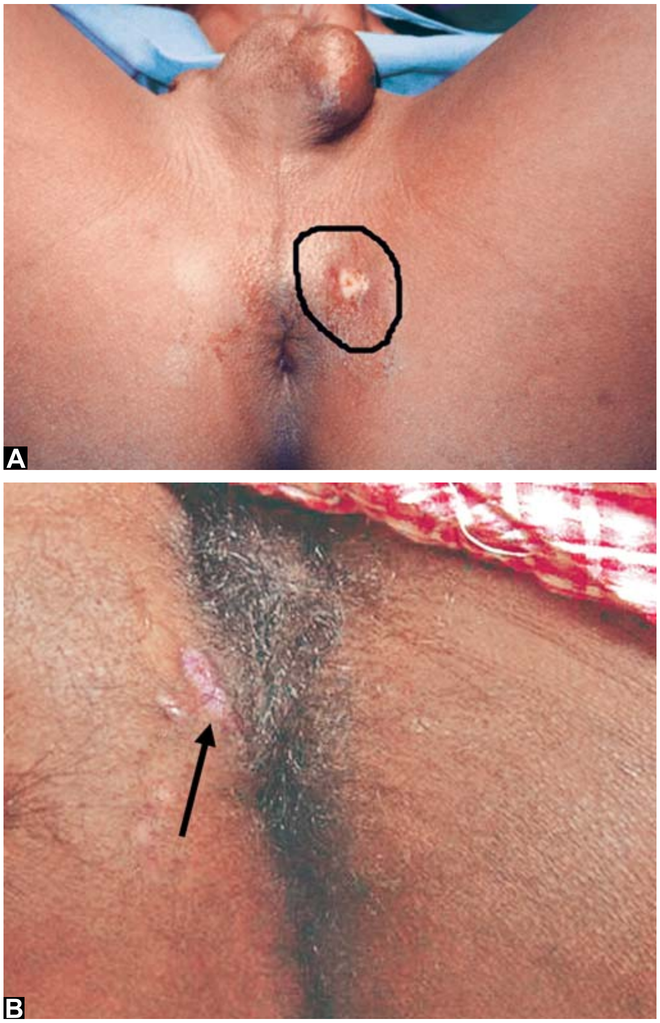

Figs 25.74A and B

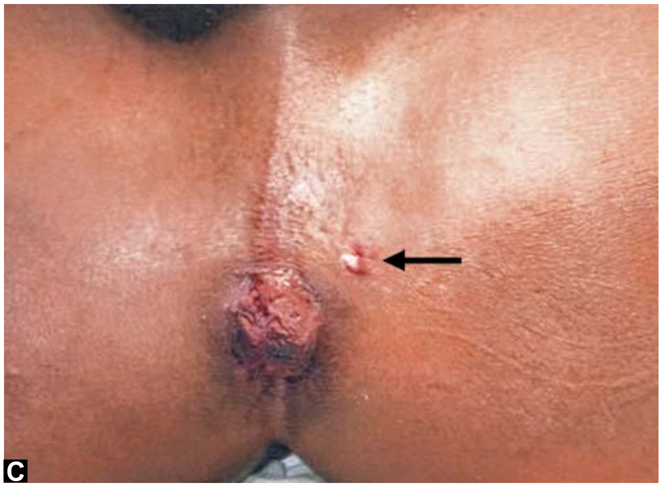

Fig. 25.74C

Figs 25.74A to C: Fistula-in-ano (*simple type*)—typical look and site.

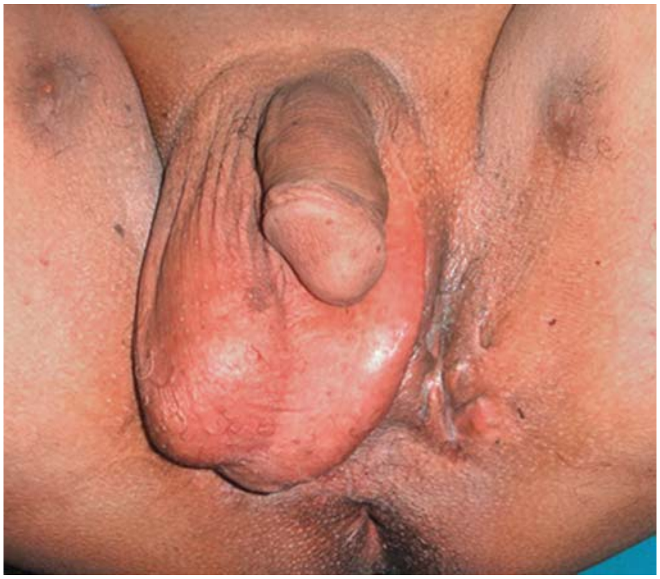

Fig. 25.75: *Complex* fistula-in-ano.

Goodsall's (1900) Rule

- Fistulas with an external opening in relation to the anterior half of the anus is of *direct type.*
- Fistulas with external openings in relation to posterior half of the anus, has a curved track may be of horse-shoe type, opens in the midline posteriorly and may present with multiple external opening all connected to a single internal opening.
- *P/R examination shows indurated internal opening usually in the midline posteriorly***.**
- Most of the fistulas are on posterior half of anus.
- *Probing in the ward and fistulogram* in the ward before surgery using *Lipiodol* is not advisable as it may cause

You never get the second chance to make the first impression.

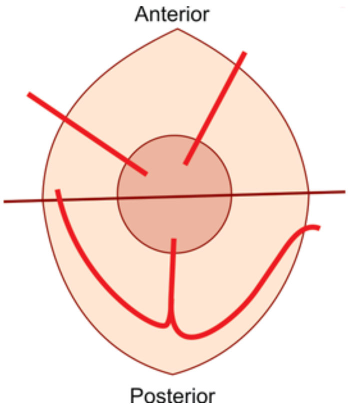

Fig. 25.76: *Goodsall's rule*. Anterior fistulas are having straight track. Posterior fistulas are having curved track with internal opening in the posterior midline.

recrudescence of inflammation. It can be done with adequate precaution. Probing is done under *general anaesthesia* gently with care without creating extensions.

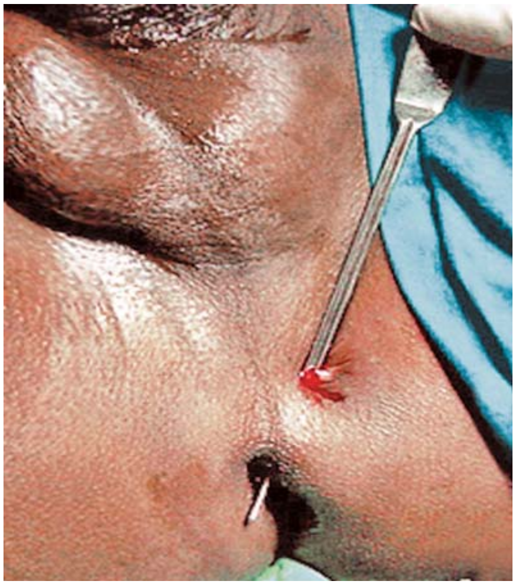

Fig. 25.77: Anterior fistula-in-ano with probe in place. Anterior low fistula has got straight track. Both internal and external openings are seen.

Investigations

- Chest X-ray, ESR and barium enema X-ray.
- *If required fistulogram is done only under anaesthesia.*
- **MRI/MRI fistulogram** ideal.
- Endorectal U/S (US perineum) is useful to assess deeper plane.
- Discharge study, methylene blue dye study, biopsy.
- Colonoscopy often when ulcerative colitis/Crohn's is suspected.
- Specific blood test.

Tuberculous fistulas do not have induration, will have pale granulation tissue with watery discharge and they are most often multiple. Here, the infection occurs in lymphoid tissue over the lower part of anal canal, around anal gland opening.

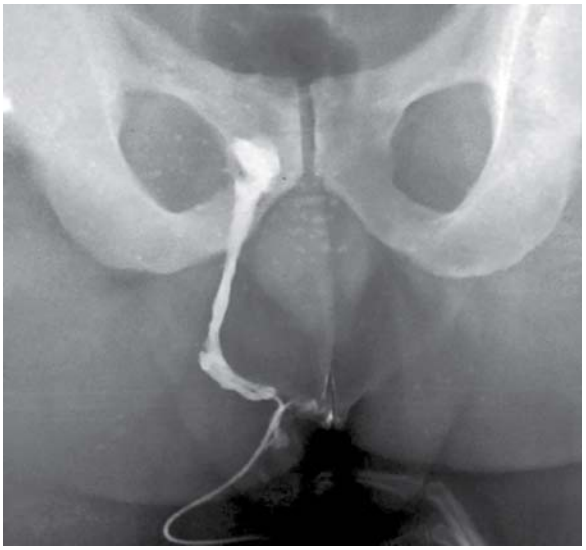

Fig. 25.78: Fistulogram X-ray showing track.

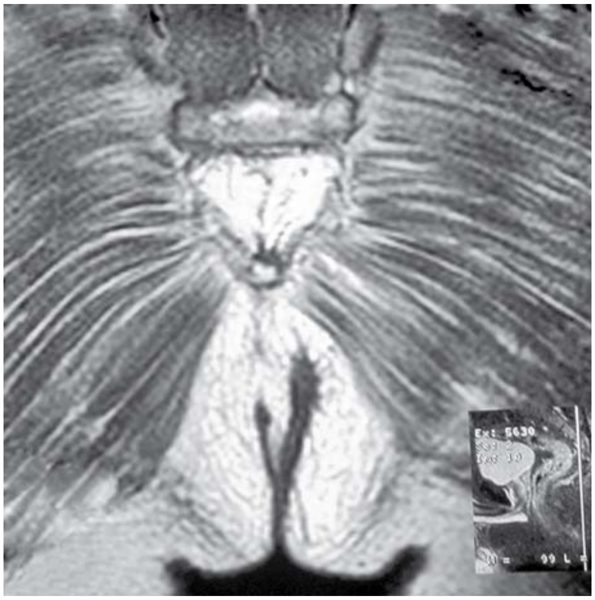

Fig. 25.79: MR fistulogram.

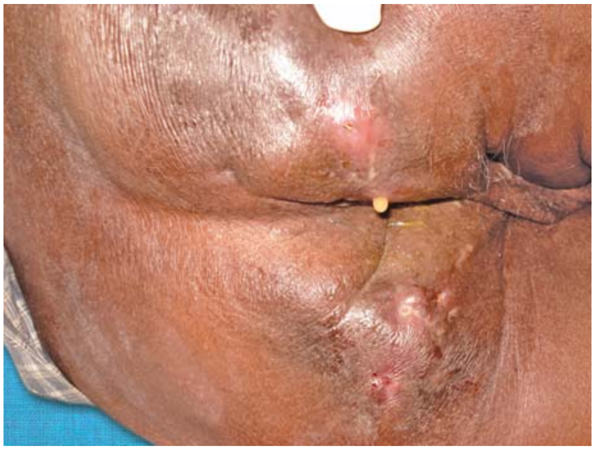

Fig. 25.80: Multiple fistulas-in-ano. It may be Crohn's disease; carcinoma or HIV. Biopsy should be done prior to formal therapy. MRI of perineum is of great help in such patients to anatomically evaluate the fistula-in-ano.

Differential diagnosis for fistula-in-ano

- Urethral fistula in male
- Chronically infected Bartholin's gland
- Pilonidal sinus
- Hidradenitis suppurativa
- Carcinoma
- Crohn's tuberculosis, ulcerative colitis

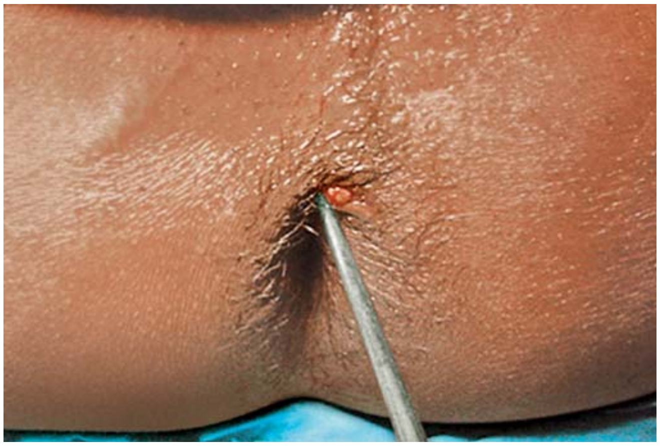

Fig. 25.81: Probing of the fistula track during surgery to find out its inner opening.

Treatment

Principles in management	***Procedures***
• Identify the cause—cryptoglandular or other • Delineate exactly the fistula anatomy —MRI/EUS • Identify relation of fistula to anal sphincter • Drain all sites of infection • Eradicate **track** and secondary **extensions** • Preserve anal incontinence function The *primary objectives* are to eradicate the tract and drain all associated sites of infection while simultaneously preserving anal continence	• Laying open the fistula—fistulotomy with curetting • Fistulectomy • Gluing of fistula—not much useful • Mucosal flap procedure • Fistulectomy with primary repair • Fistulectomy with primary repair with episiorectopexy • Fistulectomy with secondary repair • Anal fistula plug repair (AFP) • LIFT Technique (Ligation of intersphincteric fistula track) • Seton technique (Latin—seta—*a bristle*) • Colostomy—lay open, fistulectomy, later closure of colostomy—for high type • VAAFT procedure (Video assisted anal fistula track ligation)

- *Fistulectomy*
  - Under G/A or spinal anaesthesia, probe is passed through external opening up to the internal opening which is felt as an induration. Fistula is opened along the probe using a knife. Fibrous track along with unhealthy granulation tissue and additional external openings are excised. Specimen is always sent for histopathology.

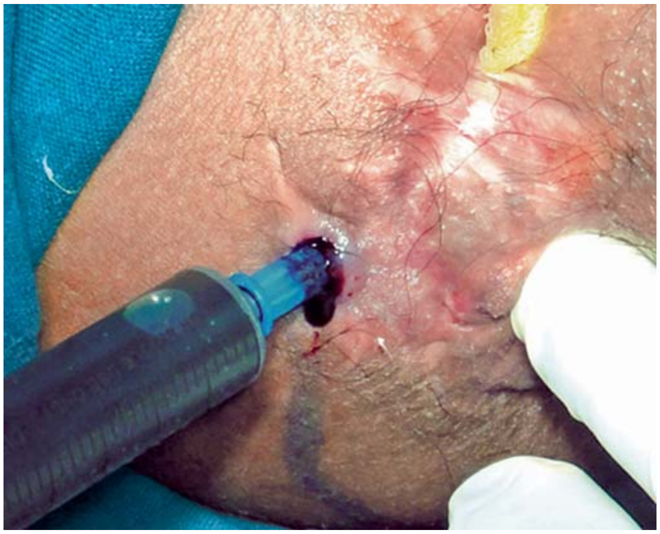

Fig. 25.82: Methylene blue injection to delineate track on table in fistula-in-ano.

  - Postoperatively—sitz bath, antibiotics, analgesics, laxatives are given.
  - *Fistulectomy for low level fistulas do not cause rectal incontinence.*
  - Proper curetting of the infected anal gland area is essential.
- *Fistulotomy*
  It is done in low anal fistula. It is technically easier. After passing the probe through the entire fistulous track, incision is made over the probe to cut and lay open the fistulous track. It is allowed to granulate and heal from the floor/surface. Technique is safer, easier and can be done on outpatient basis.
- *Advancement flaps* are used occasionally to get better result—mucosal flap procedure.
- *Gluing of the fistula* track is tried but success rate is not good. Fibrin glue is a multicomponent system containing mainly human plasma fibrinogen and thrombin. Once prepared it is injected into the fistula track which hardens in few minutes and fills the entire track. Success rate is 70%.
- *Anal fistula plug (AFP) repair*: Surgisis anal fistula plug (*porcine small intestine submucosa, SIS*) is used with 85% success rate in simple fistula. It contains naturally derived extracellular matrix which acts as scaffolding, ingrowth of tissue, remodeling.
- *LIFT technique* (Ligation of intersphincteric fistula track): Under anaesthesia in lithotomy position, intersphincteric space is reached through a transverse incision. Fistula running across is identified and ligated using vicryl on either side. Part is excised; outer part is curetted through external opening.
- *VAAFT procedure* (Video assisted anal fistula track ligation): Fine specialized endoscope is passed through the outer opening into the fistula track; with continuous irrigation fistula track is cleaned and wall is cauterized. Inner opening is ligated through vicryl from luminal side.

HIGH-LEVEL FISTULAS

- Its upper opening is at or above the anorectal ring. It is difficult to treat.

Common causes are:

- Crohn's disease
- Ulcerative colitis
- Trauma
- Carcinomas
- Foreign body

- Incontinence may follow after lay opening of these fistulas.

Investigations

- Barium enema X-ray, colonoscopy, chest X-ray, biopsy.

Treatment

- Requires staged procedure—*initial colostomy* is done followed by definitive procedure. This prevents sepsis and promotes faster healing.
- Later closure of colostomy is done.

Seton Technique

- A *silk or linen ligature* is passed across the fistula and left in place with a tie.
- *Striated muscle* superficial to fistula track is encircled with Seton material and tied securely and left *in situ* to create ischaemic necrosis, dividing the muscle slowly without allowing it to spring apart avoiding gutter deformity. Even though internal sphincter is divided in this, causing adequate laying open of entire fistula track, it will well-preserve the sphincter function and pressure.
- This allows the fistula to granulate and heal from above and to close completely. Usually takes longer duration to heal.
- It is done for intermediate and intersphincteric fistula. It is used prior to definitive procedures like fistulectomy or advancement flap.

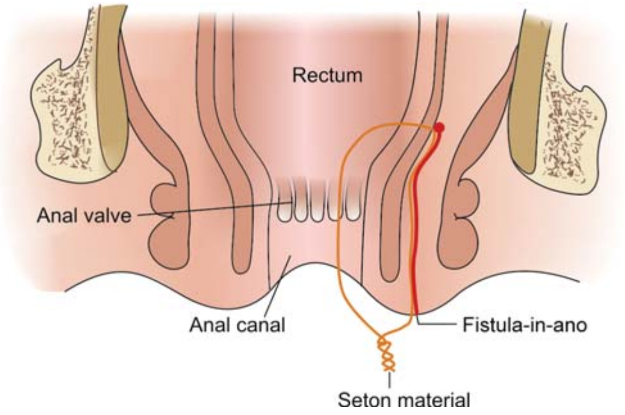

Fig. 25.83: *Seton technique* for intersphincteric fistula-in-ano. It is kept for 3 months. It can be tight or loose seton depending on the indications. It promotes formation of granulation tissue, healing by fibrosis and track recedes downwards.

- Seton can be kept for 3 months. It can be regularly replaced by new silk or any material by rail road technique without anaesthesia.
- *Two types of setons are present*.
  - Loose setons are used mainly to drain for long period in recurrent/postoperative fistulas and due to specific causes like Crohn's. There is no tension in seton.
  - Cutting setons are used when enclosed muscle is needed to cut (cheese wiring through ice effect). It is placed tight.

ANORECTAL STRICTURES

Causes

- Congenital
- Fibrotic anal fissure
- Irradiation
- Senility
- LGV (in females)
- Ulcerative colitis
- Crohn's disease
- Carcinomas
- Postoperative (surgery for piles, coloanal anastomosis)

Clinical Features

- Progressive constipation.
- On P/R examination, stricture can be felt as a tight ring.
- Features relevant of specific cause.

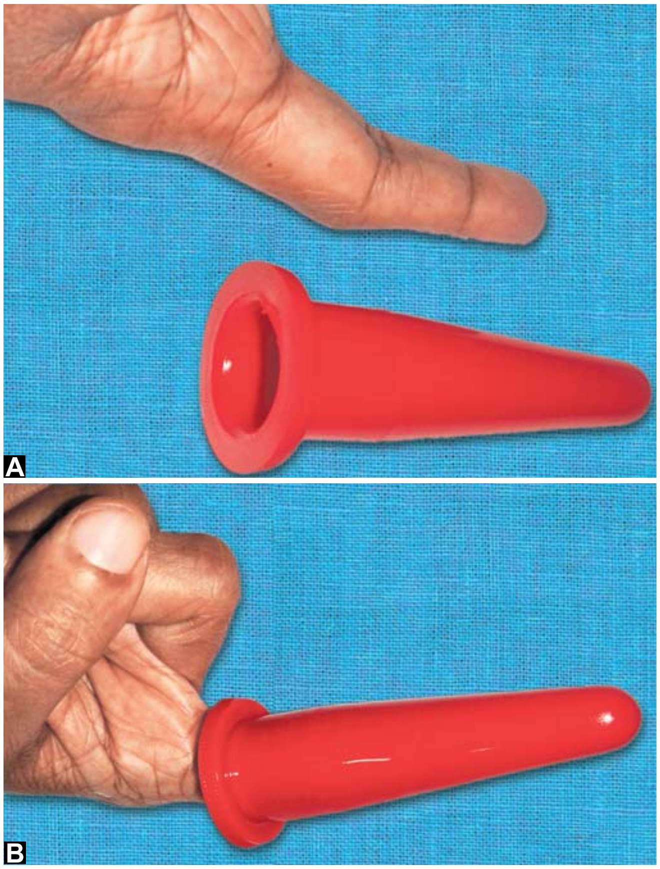

Figs 25.84A and B: *Anal dilator*. Many proctology surgeries require regular anal dilatation postoperatively. Note how to hold the dilator to pass into the anus after applying lubricant.

Investigations

- Barium enema X-ray, biopsy, colonoscopy

Treatment

- The cause is treated.
- Dilatation of the anal canal under general anaesthesia.
- Resection in severe recurrent cases.

CONDYLOMA ACUMINATA

- It is most common sexually transmitted anal disease. It is common in homosexual men.
- Penile warts or female genital warts may be present.
- It is caused by Human Papilloma Virus (HPV).
- Pruritus, discharge, pain and bleeding are the features.
- Pinkish white warts in anal canal, often attaining large size causing *Buschke Lowenstein tumour*.
- Large wart may block the anal canal orifice.
- Whitening occurs on applying acetic acid on it.
- Biopsy confirms the diagnosis.
- Treatment is local application of 25% podophyllin cream; surgical excision of the wart; intralesional injection of interferon α 2b.
- Malignancy should be ruled out by histology.

ANAL INTRAEPITHELIAL NEOPLASIA (AIN)

- It is dysplasia of anal or perianal epidermis.
- It is seen in individuals with HIV infection; HPV infection (16, 18); individual who do anorectal intercourse.

Classification:

AIN I—absence of keratocyte maturation and cellular atypia observed in outer 1/3rd of epithelium—low grade.

AIN II—cellular atypia observed in middle 1/3rd—low grade squamous intraepithelial neoplasia.

AIN III—cellular atypia full thickness—high grade squamous intraepithelial neoplasia.

- 30% of anal warts will show AIN.
- It is raised scaly white/pigmented/cracked lesion.
- Biopsy confirms the disease.
- Treatment: Excision; topical imiquimod, 80% tricholoroacetic acid and oral retinoids.

MALIGNANT TUMOURS OF ANAL AREA

- *Anal malignant tumours* are < 2% of large bowel tumours. It can be below the dentate line (SCC, 80% of anal tumours); above the dentate line (Basaloid/transitional/cloacogenic).
- Causes may be HPV infection; HIV infection; AIN (Anal intraepithelial neoplasia); organ transplant recipients; immunosuppression.

Classification of tumours of anal area

It can be:

- Benign
- Malignant

It also can be:

- Tumour of the anal canal (proximal to dentate line)—SCC, adenocarcinoma, melanoma
- Anal margin tumour (distal to dentate line)—ANI, Bowen's disease, Paget's disease, BCC, anal margin SCC

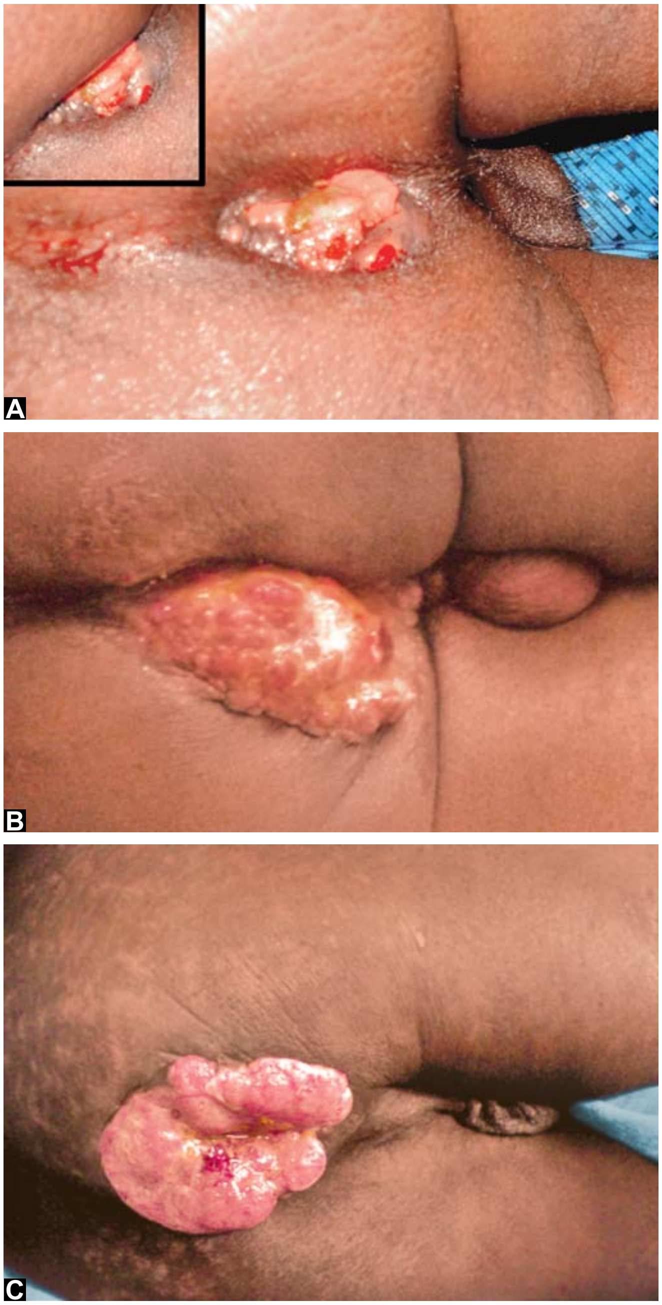

Figs 25.85A to C: Anal canal carcinoma. Squamous cell carcinoma is commonest type—80%.

Types

1. *Squamous cell carcinoma* is the commonest type. Predisposing causes: Papilloma, irradiation, dermatitis, long standing fistula-in-ano.
2. Basaloid carcinoma—it is rare, non-keratinising squamous cell carcinoma. Highly malignant.
3. Muco-epidermoid carcinoma—arises near squamo–columnar junction.
4. Basal cell carcinoma.
5. Melanoma—blue/black in colour mistaken for thrombosed pile—poor prognosis (5 years—10%).
6. Adenocarcinoma from the anal glands in a pre-existing fistula-in-ano.

God gives every bird its food, but he does not throw it into the nest.

Features	*Investigations*
• Ulceration	• Biopsy from anal region
• Bleeding	• FNAC of inguinal node
• Pain, pruritus and discharge	• U/S abdomen
• Irregular indurated mass	• MRI perineum is very useful
• Anovaginal fistula in females	• *P/R is a must to assess upper extent of the growth*
• Faecal incontinence in late cases	*AJCC staging* Tis—carcinoma *in situ*
• Inguinal node—hard, nontender	T1—tumour < 2 cm T2—tumor 2-5 cm T3 —tumor > 5 cm T4—invasion into adjacent organs like vagina/urethra N1—perirectal nodes + N2—unilateral internal iliac/inguinal nodes N3—perirectal + inguinal/ iliac or bilateral nodes M 0—no distant spread M1—distant spread present
• Iliac nodes	
• Later—late constipation—obstruction	

- ***Squamous cell carcinoma of anal canal,*** usually present as a fungating or ulcerative growth, which spreads to inguinal lymph nodes.
  - Biopsy and FNAC of lymph nodes are the essential investigations.
  - *Treatment:* Wide excision of the lesion with 3 to 5 cm clearance and ilioinguinal block dissection for lymph nodes are done. Follow-up radiotherapy is also often given.
  - Nigro regime

Nigro Regime

Nigro Regime for Anal Carcinoma (Norman Nigro, et al. 1974)

- Initial radiotherapy for 3 weeks 3000 rads (30 Gy total) to perineum and pelvis
- Then chemotherapy—5 FU, for 4-5 days; is a radiosensitizer, started on 1st day of RT as 1000 mg/m^2 continuous infusion. Mitomycin C is 15 mg/m^2 as single dose on 1st day of RT
- Later after 3 weeks abdominoperineal resection (APR)

  - ***Chemoradiation*** is becoming popular for carcinoma of anal canal.
  - Drugs used for chemotherapy are 5 FU, bleomycin, vincristine, adriamycin.
  - In advanced growths radiotherapy is the only treatment.
- ***All other tumours**: Abdominoperineal resection with* permanent colostomy is done.

Paget's

- Paget's disease of breast
- Paget's disease of anal margin—of apocrine glands
- Paget's disease of penis
- Paget's test
- Paget's disease of bone

ANAL MARGIN TUMOURS

It is the tumour below the level of dentate line.

1. Bowen's disease.
2. Paget's disease.
3. Basal cell carcinoma.
4. Squamous cell carcinoma.
5. Verrucous carcinoma (Giant condyloma acuminatum or Buschke-Lowenstein tumour).
6. Epidermoid carcinoma (SCC; cloacogenic/transitional/ basaloid carcinoma).

SACROCOCCYGEAL TERATOMA

- It is an uncommon tumour, but most common of the large tumours in first 3 months of life.
- More common in females.
- Retention of large amount of primitive toti-potential cell in this region may be the reason for this tumour.
- It occurs between coccyx and rectum
- It is attached to coccyx, extends commonly downwards as a huge mass, occasionally upwards into the pelvis.

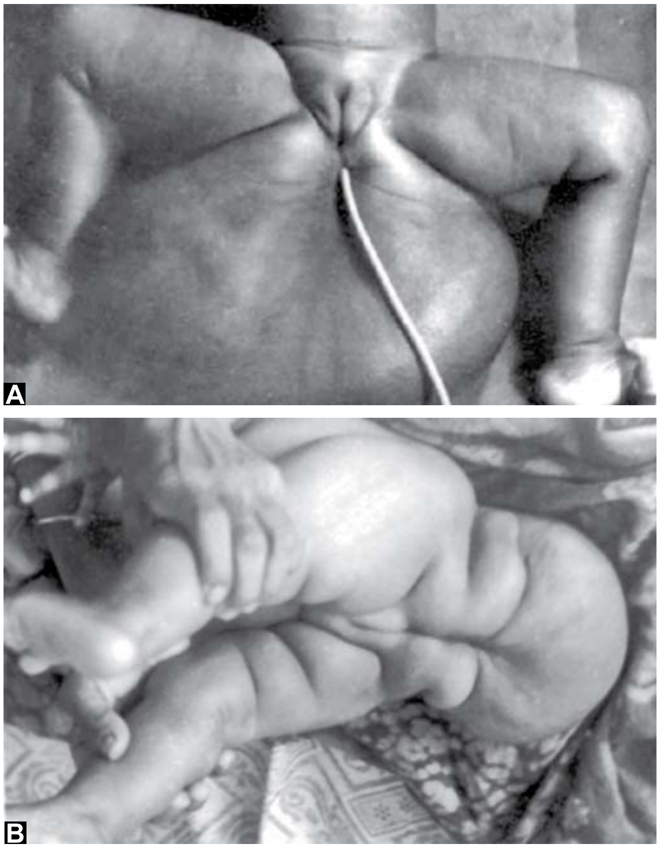

Figs 25.86A and B: Sacrococcygeal teratoma, typical site (*Courtesy:* Professer Suresh Kamath, MS, Mangalore).

- It is a congenital condition arising from totipotent cells.
- X-ray and CT scan are must.

Complications	***Differential diagnosis***
Ulceration	Sacral meningocele
Infection	Sacral chordoma
Urinary obstruction	Postanal dermoid
Malignant changes	

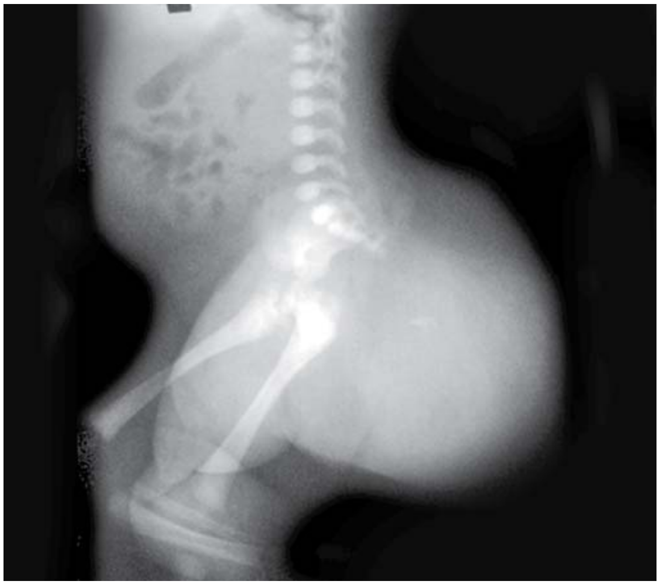

Fig. 25.87: X-ray of a patient with sacrococcygeal tumour (*Courtesy:* Professor Suresh Kamath, MS, Mangalore).

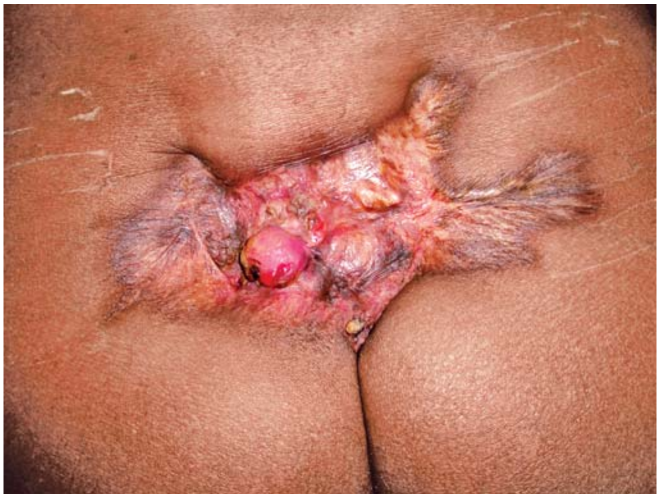

Fig. 25.88: Recurrent chordoma of sacrum.

Treatment

- Excision soon after birth.

ANAL INCONTINENCE

Continence of anal canal is maintained by two factors:

- Normal rectal and colonic pressure and activity.
- Normal pelvic floor function.

Types

- *Urge incontinence*—here rectal and colonic pressure and activity is increased but normal pelvic floor.
- *True incontinence*—here rectal and colonic pressure and activity is normal but defective pelvic floor function.
- *Full incontinence*—here rectal and colonic pressure and activity is reduced and also defective pelvic floor function.
- *Temporary*—treated by reassurance. Often seen after Lord's dilatation.
- *Permanent*—needs definitive therapy.

Causes

Causes of anal incontinence

- Denervation—spinal injury, spina bifida
- Damage—childbirth, wounds, surgeries
- Descent—rectal prolapse, perineal descent
- Debility—old age, diseases
- Destruction—RT, malignancy
- Dementia—senility, psychosis
- Deficiency—congenital anomalies

- Irritable bowel syndrome, severe diarrhoea.
- Prolapsed piles, rectal prolapse.
- Old age, malnutrition, debilitating illness.
- Congenital anomalies.
- Trauma, surgeries, injury during childbirth in females.
- Spina bifida, spinal tumours, spinal injuries and surgeries.
- Malignancy, postirradiation.
- Psychological causes.

Evaluation of the patient

- For specific causes
- Anorectal manometry
- Per-rectal examination
- Sigmoidoscopy
- Electromyography
- Defaecography
- Perineometer to assess level and angle of anorectal junction

Treatment

- Suturing of the torn sphincter.
- Repair of puborectalis muscle and plication of external sphincter.
- Encircling operations around anal canal to give support using gracilis sling or mersiline sutures.
- Electrical stimulation of the puborectalis.
- *Secca therapy*: Temperature controlled radiofrequency energy in the anal canal and distal rectum to create scarring and fibrosis of internal sphincter and adjacent tissues.

DESCENDING PERINEAL SYNDROME

- When a healthy person increases the intra-abdominal pressure and relaxes the pelvic floor muscles, there will not be any changes in the concavity of the perineum.

- In chronic ill-patients, malnourished, and people with preprolapse, perineal descent can occur with obliteration of the normal concavity of the perineum. It is called as *descending perineal syndrome*.
- Levators got injured directly or indirectly causing weakening of pelvic floor.
- Anal canal is situated many centimeters below the pubo coccygeal line. Usually 3-4 cm low during straining.
- *Defecography* is ideal to evaluate such patients. Perineometer is also used.
- Presentations are tenesmus; incomplete evacuation; incontinence.
- *Treatment* is diet, laxatives, avoiding straining, and suppositories.
- Restoration of pelvic floor by various surgical methods may be needed often with rectal resection and suspension.
- Total pelvic marlex mesh repair; transcoccygeal posterior hitching of the rectum (Kraske); correction of cystocele, rectocele and enterocele.
- Results are *not very good* as recurrence or residual problems may persist.

PROCTITIS

It is inflammation of rectal mucosa often with the inflammation of colon and anal canal.

Types

- Acute.
- Chronic.
- Nonspecific—common.
- Ulcerative proctocolitis as part of ulcerative colitis.
- Specific
  - Bacillary dysentery.
  - Amebic proctitis—common.
  - Combined amoebic and bacillary.
  - Gonococcal proctitis.
  - Lymphogranuloma inguinale (LGV).
  - Tuberculous proctitis.
  - Bilharzial proctitis due to schistosoma haematobium.
  - Enema induced proctitis especially of herbal enemas.

Clinical features

- Pain per rectum and anum
- Tenesmus
- Passage of mucus and blood
- Frequent urge to pass stool
- Fever, loss of appetite
- Pain and tenderness in left lower abdomen
- P/R is tender

Investigations

- Sigmoidoscopy is more relevant than just proctoscopy.
- Stool study, stool culture.
- Mucosal biopsy.
- Serological tests.
- Relevant investigations like ESR, blood smear, and chest X-ray.

Treatment

- Antibiotics, antiamoebic drugs like metronidazole.
- In severe cases, retention enema using metronidazole, prednisolone, salazopyrin.
- IV fluids, IV antibiotics and IV metronidazole are often required.
- Treating the specific causes like tuberculosis, gonococcal infection and bilharzial infection.

PROCTALGIA FUGAX

- It is sudden severe recurring pain in the rectum of unknown cause with segmental pubococcygeal spasm.

Features

- It is common in young people may be due to stress, straining.
- Common at night, starts suddenly, lasts for few minutes and then subsides spontaneously.
- Pain is unbearable and severe with often constipation.
- Gradually subsides on its own.
- Occasionally, only cutting of puborectalis muscle is required but with danger of developing incontinence.

HIDRADENITIS SUPPURATIVA OF ANAL REGION

- It is a chronic suppurativa condition of apocrine glands of the skin in axilla/perineum/mons pubis/ thighs/scrotum etc.
- Apocrine gland duct obstruction → bacterial infection → multiple glands involvement → secondary infection (*Staphylococcus aureus*, streptococci) → skin oedema, multiple raised pustules → multiple communicating fistulae formation.
- Disease does not extend above dentate line or into sphincter.
- It is common in young obese females.
- Sinus; scarred areas; discharge; skin changes; pain and tenderness; foul smelling fluid are the presentations.
- *Differential diagnosis* are—Crohn's disease; fistula-in-ano; pilonidal sinus; tuberculosis; actinomycosis; LGV.
- *Treatment*:
  - Weight reduction; proper hygiene.
  - Antibiotics; analgesics.
  - Incision and drainage of abscess.
  - Laying open of all communicating tracks and regular dressing.
  - Radical local excision of entire apocrine bearing perineal skin with reconstruction using flap.
  - Recurrence is known to occur.

PRURITUS ANI

- It is intractable itching in and around anal canal.
- Skin is reddened, hyperkeratotic, cracked and moist.

Causes

- Poor hygiene.
- Anal discharge due to fissure/fistula/piles/warts/ polyp.
- Trichomonas vaginalis infection of vagina in females.
- Parasites.
- Epidermophytosis.
- Allergic cause.
- Dermatitis/psoriasis.
- Intertrigo/erythrasma (Corynebacterium minutissimum).
- Diabetes mellitus; psychological cause.

Treatment

Proper cause should be assessed and treated. Good hygiene; local steroid application; topical xylocaine; strapping of the buttocks are needed.

GASTROINTESTINAL HAEMORRHAGE (GI BLEED)

GI bleed is classified as upper GI and lower GI bleed.

- ***Upper GI bleed*** is bleeding above the level of ligament of Treitz.
- ***Lower GI bleed*** is bleeding below the level of ligament of Treitz.

  Ligament of Treitz is a fibromuscular band, which extends from right crus of diaphragm to duodenojejunal flexure with upper part made up of striped muscle fibres, lower part smooth muscle fibres and middle part with elastic fibres.

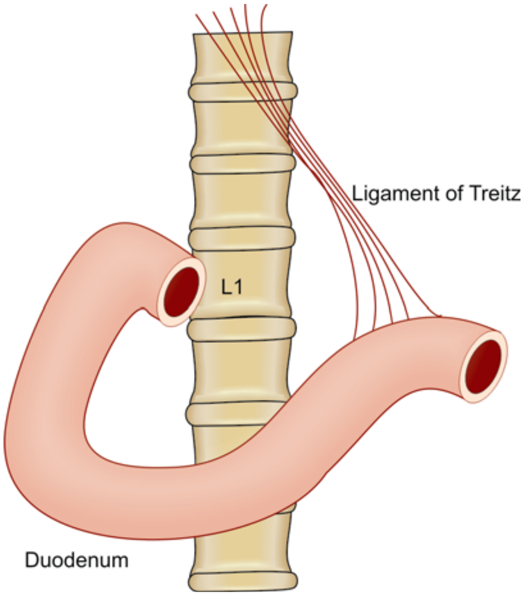

Fig. 25.89: Anatomy of ligament of Treitz.

UPPER GI BLEED

It is considered as:

- Variceal.
- Nonvariceal.

Causes

- Peptic ulcer 55%. Ulcer bleeding is precipitated by NSAIDs, steroids, alcohol. Ulcer bleeding is overall common in men. But NSAID induced ulcer bleeding is common in females. *Duodenal ulcer* (35%) more commonly bleeds than *gastric ulcer* (20%)
- Gastroduodenal erosions.
- Oesophageal varices.
- Oesophagitis and erosions.
- Carcinoma stomach—5%
- Mallory-Weiss syndrome—5-15%.
- Vascular anomalies—Dieulafoy's syndrome (A-V malformation in the fundus of the stomach), Osler-Weber-Rendau syndrome, Ehlers-Danlos syndrome.
- Aortoduodenal fistula.
- Bleeding disorders.
- H/O drug intake—anticoagulants, clopidogrel, ecospirin.

Factors which aggravate the bleeding

- Gastric acid which inhibits the platelet aggregation
- Pepsin, by its proteolytic action causes erosion of the ulcer into the vessel. It also digests the clot, so as to aggravate the bleeding
- Mucosal blood supply pattern
- Gastric motility
- Alcohol, drugs

  Major haemorrhage occurs when erosion of gastroduodenal artery or left gastric artery or splenic artery occurs or when bleeding occurs from varices

Clinical Features

Acute Bleed

- Features of shock.
- Haematemesis.
- Melaena.

Chronic Bleed

- Hypochromic microcytic anaemia, glossitis, koilonychia, congestive cardiac failure.
- Mortality in upper GI bleed is 10%.

Investigations

- *Gastroscopy* to see the spurting vessel, oozing, clot in the ulcer, collected blood in the lumen of the stomach.
- CT angiography of coeliac trunk and SMA.
- Hb%, packed cell volume, CVP measurement, blood grouping and crossmatching. U/S abdomen.
- LFT; prothrombin time; platelet count; blood urea and serum creatinine; serum electrolytes.
- *Modified Forrest classification* (refer Chapter 20—Stomach) and *Rockall scoring system* (better) is used. In ***Rockall scoring system*** parameters used are—*age* (60 [0], 60-79 [1], > 79 [2] Years); *shock* (none [0], pluse > 100 [1],) pluse > 100 with hypotension [2]); *comorbidity* (none [0], IHD [2], renal/liver failure/advanced maligancy [3]); *condition diagnosed* (Mallory Weiss [0], all but mailigancy [1], upper GI maligancy [2]); *endoscopic finding* (none [0], dark spot, blood/clot/visible vessel [2]). *Score*: < 3 low-risk; 3-8 moderate; > 8 high.

Happiness lies in the joy of achievement and the thrill of creative effort.

Treatment

Treatment of upper GI bleed

- *General*
- *Medical*—**PPI**/tranexamic acid/octreotide
- *Endoscopic*
- *CT angiography guided embolisation*
- *Surgical*

- General: IV fluids, catheterisation, Ryle's tube aspiration, blood transfusion.
- Injection ranitidine IV 50 mg 8th hourly, or famotidine IV, omeprazole IV, pantoprazole IV.

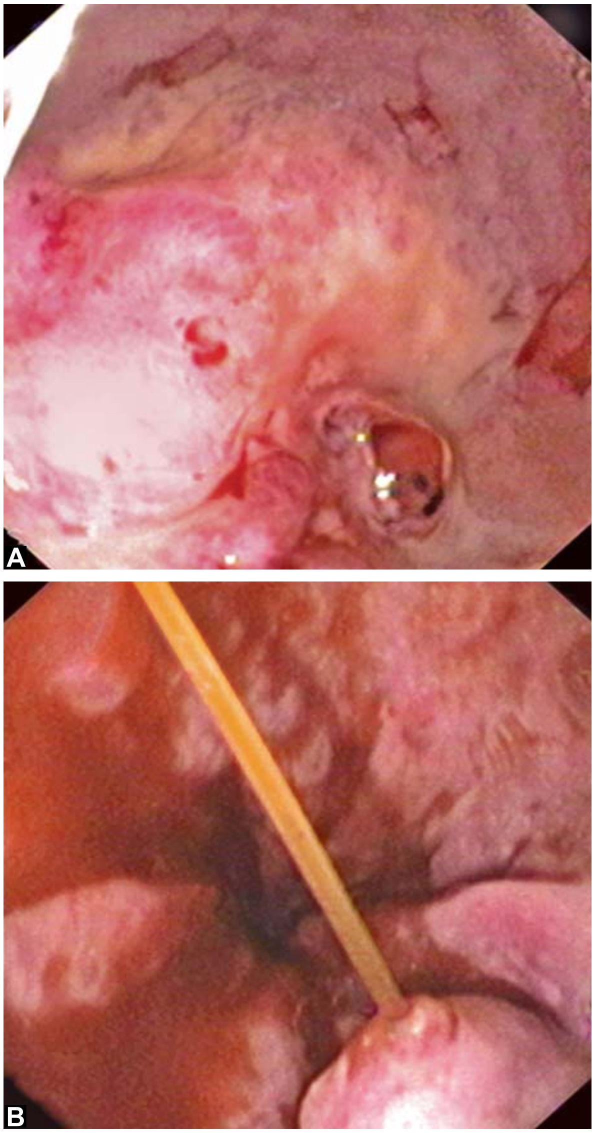

Figs 25.90A and B: Bleeding peptic ulcer (duodenal ulcer) is the commonest cause of upper GI bleed. Oesophageal varix causes dangerous severe life-threatening bleed.

- Antifibrinolytics like tranexamic acid, EACA.
- Somatostatin or octreotide, *PPI infusion.*
- *Endoscopic therapy* (tamponade, laser, haemoclip, banding, sclerotherapy, etc.) is the first line of therapy in all upper GI bleed. *CT angiography guided transcatheter embolisation* of artery (gastroduodenal) is very useful in bleeding duodenal ulcer if endoscopic therapy fails. *Persistent recurrent bleeding* needs surgical intervention.
- For varices, vasopressin, propranolol, isosorbide dinitrate, Sengstaken tube tamponade, sclerotherapy, Boerema-crile operation, Hasaab operation (devascularisation with splenectomy), oesophageal transection, Siguira-Futagawa operation or TIPSS.
- For peptic ulcer, saline wash with 1 : 2,00,000 adrenaline, Nd:YAG laser therapy, cautery coagulation, thrombin injection, Finney's pyloroplasty, partial gastrectomy, ligation of gastroduodenal or left gastriartery, cytoprotectant mesoprostil injection.
- The cause is treated, once acute episode is under control.

Prognosis

- Varices and gastric ulcer bleed has higher mortality.
- Bleeding duodenal ulcer has got better prognosis.

LOWER GI BLEED

- Bleeding in the GIT below the level of the ligament of Treitz.
- Normal faecal blood loss is 1.2 ml/day. A loss more than 10 ml/day is significant.

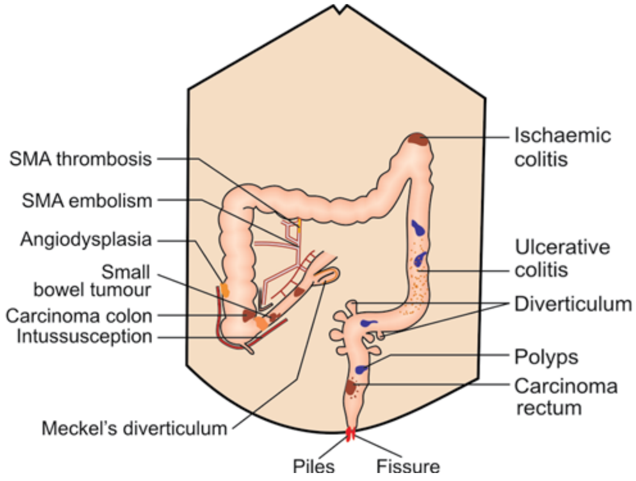

Fig. 25.91: Causes of lower GI bleed.

Causes

- Angiodysplasia.
- Diverticular disease—commonest cause in Western countries.
- Tumours of colon or small bowel.
- Anorectal diseases—haemorrhoids, fissure-in-ano.
- Ulcerative colitis, Crohn's disease.
- Colorectal polyps; rectal carcinomas.
- Intussusception.
- Tumours, either benign or malignant of colon or small bowel.
- Meckel's diverticulum.
- Ischaemic colitis.
- Stercoral ulcer.
- Infectious colitis.
- Mesenteric artery occlusion.

Lower GI bleed can be:

- Occult bleed: > 10 ml/day but not revealed
- Overt bleed: Bleeding which is revealed
  - Overt acute
  - Overt acute massive (bleed > 1.5 litre/day)
  - Overt chronic

Common causes

- Internal piles
- Diverticular disease
- Neoplasia
- Inflammatory bowel diseases
- Angiodysplasia

Classification

I. Bleeding may be:

Small bowel bleed: Polyp', Meckel's diverticulum, mesenteric ischaemia, intussusception; small bowel tumor.

Large bowel bleed: Angiodysplasia, carcinoma, colitis, diverticulitis, carcinoma.

Anorectal diseases: Piles, fissure-in-ano, carcinomas.

II. Bleeding may be:

Congenital: Polyp's, Meckel's diverticulum.

Inflammatory: Ulcerative colitis, infective, amoebic, Crohn's disease.

Neoplastic: Adenomas, carcinomas, polyps.

Vascular: Angiodysplasia, mesenteric artery ischaemia, colitis.

Others: Piles, fissure-in-ano.

Angiodysplasia is common in caecum and ascending colon.

Bleeding more than 1.5 litres per day is called as *acute massive GI bleed.*

Acute bleed occurs in:

- Mesenteric ischaemia
- Angiodysplasia
- Ischaemic colitis
- Meckel's diverticulum
- Intussusception
- Acute episodes of ulcerative colitis

  80% of acute bleed regress spontaneously

  20% will become either massive or recurrent

Presentations

- Acute bleeding presents with features of shock.
- Chronic blood loss occurs in piles, fissures, colitis. Presents with hypochromic, microcytic anaemia.
- Tenesmus, subacute obstruction, loss of appetite, decreased weight, bloody diarrhoea is seen in carcinoma distal, large bowel.
- Per-rectal examination is a must which may reveal polyp, growth, ulcerations.
- Haematochezia.
- Mass palpable per abdomen in left or right iliac fossa or mass of intussusception.

- *Blood with mucus*—colitis, carcinoma
- *Fresh blood as splashes in the pan*—piles
- *Maroon coloured stool*—Meckel's diverticulum
- *Red currant jelly in stool*—intussusception
- *Bright red blood in stool*—polyps

Investigations

- Hb%, packed cell volume, ESR.
- Bleeding time; clotting time; prothrombin time; platelet count; blood urea and serum electrolytes.
- Occult blood in the stool—positive faecal blood test using Guaiac reagent.
- Barium enema.
- Proctoscopy for piles and sigmoidoscopy for rectosigmoid diseases.
- Colonoscopy for colitis, carcinomas, polyps.
- Small bowel enema (enteroclysis).

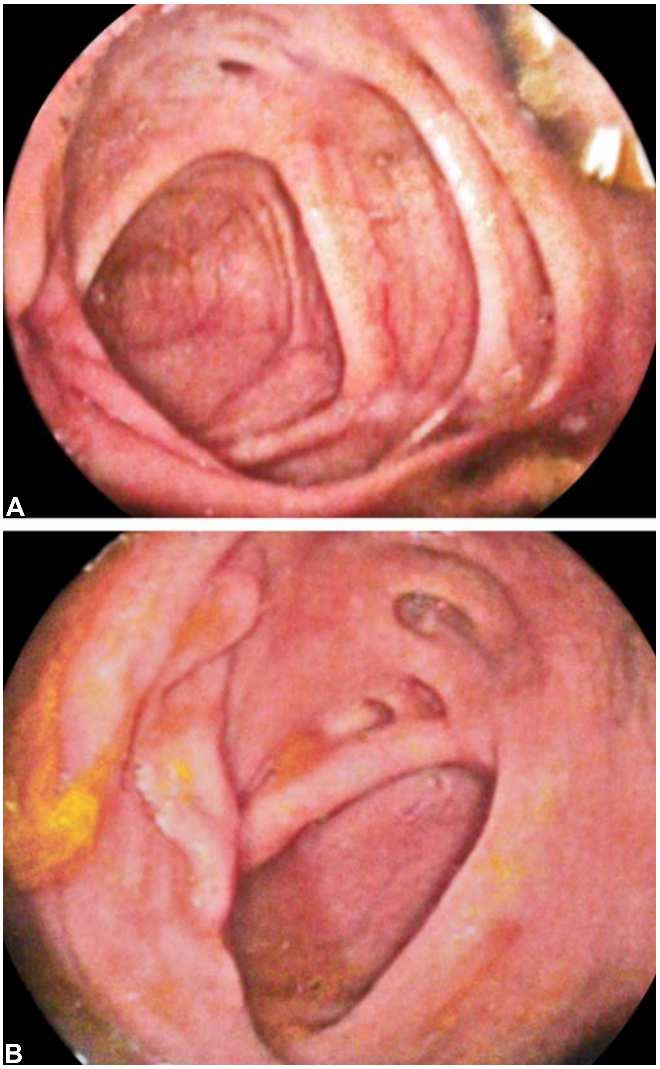

Figs 25.92A and B: Sigmoid diverticula on colonoscopy. It is the common cause of lower GI bleed in Western countries.

- U/S abdomen.
- Mesenteric angiogram is very useful investigation in acute bleed, especially in angiodysplasia.
- Technetium scan for Meckel's diverticulum.
- Capsule endoscopy.

Treatment

- Endoscopic fulguration or therapeutic embolisation or right hemicolectomy (for angiodysplasia).
- Endoscopic polypectomy for polyps.
- Treatment for ulcerative colitis with mesacol enema or drugs or total proctocolectomy with ileo-anal anastomosis.
- Surgical resection of colonic carcinoma.
- Massive resection of small bowel in mesenteric ischaemia.
- Sigmoid colectomy in sigmoid diverticula.
- Cause is treated.
- Proper exploration through a lengthy midline incision is essential.

OBSCURE GI BLEED

- It is intermittent GI bleed for which *no source has been found* endoscopically/radiologically.
- It is 5% common.
- Commonly, it is due to either missed common cause or angiodysplasia.
- If it is angiodysplasia, angiography, nuclear scintigraphy, capsule endoscopy; then angiographic embolisation or resection of the part of the bowel is done.
- Enteroscopy, upper and lower GI scopies are needed in other conditions.
- All other conditions are treated accordingly.

Investigations for GI bleed

Blood

- Haematocrit, LFT, blood urea, serum creatinine
- Serum ferritin, iron, binding capacity
- Coagulation profile—platelet count, BT, CT, PT, APTT

Occult stool blood test (1-2 ml/day)

- Benzidine test
- Guaiac test
- Haemoccult—guaiac impregnated electrophoresis paper
- Fecal test
- Immunological test

Endoscopy

- Gastroduodenoscopy
- Proctosigmoidoscopy
- Colonoscopy—often difficult in acute bleed for clarity visualisation, if bleeding point is seen it can be controlled by cautery, laser, haemoclips, injection sclerotherapy, heater probe etc. But still it becomes the test of choice once bleeding has controlled/stopped temporarily
- Small bowel endoscopy—push type (160 cm); Sonde enteroscope (275 cm, 5 mm diameter with balloon tip) are used. Entire small intestine is visualised while withdrawing but therapy or biopsy is not possible

Angiography

- It identifies when bleeding rate is 0.5 ml / min; useful in active bleed; in therapeutic embolisation, injection vasopressin can be done. Embolisation of small bowel vessel may cause bowel infarction which is dangerous and resection is needed in such situation. Visualisation in angiography can be improved by selective infusion of vasodilators like tolazoline and prostaglandins, using magnification films, using vasoconstrictor drugs
- *Angiogragraphic criteria in angiodysplasia are*—early and prolonged filling of draining veins; cluster of small arteries; visualisation of vascular tuft

Nuclear scintigraphy

- Identifies 0.1 ml/min of bleed ; *Tc sulphur colloid scan* is very sensitive and is completed in 1 hour but increased uptake in spleen and liver obscures bleeding point and needs repeated administration due to rapid clearance; *Tc labeled RBC* recirulates and so effective for 1 day with better localisation
- *Advantages are*—high-sensitivity even with active continued bleed; screening test prior to angiography
- *Problems are*—no specificity; therapy is not possible

Other methods

- *CT* angiography; *MR* angiography
- *Aortography* for aortoenteric fistula
- *Intraoperative localisation*—on table enteroscopy; on table Doppler; on table bowel lavage and colonoscopy

Chapter 26 Urology

A. Kidney

> *It is no exaggeration to say that the composition of the blood is determined not by what the mouth ingests but by what the kidneys keep; they are the master chemists of our internal environment, which, so to speak, they synthesize in reverse.*
>
> **—Homer William Smith,** 1939

CHAPTER OUTLINE

- Anatomy of Kidney and Ureter
- Plain X-ray—Kidney, Ureter and Bladder
- Intravenous Urogram
- Retrograde Pyelography
- Renal Angiogram
- Micturating Cystourethrography
- Ascending Urethrogram
- Isotope Renography
- Cystoscopy
- Catheters
  - Foley's Catheter
  - Malecot's Catheter
  - Red Rubber Catheter
- Nephrostomy
- Suprapubic Cystostomy
- Haematuria
- Horseshoe Kidney
- Cystic Diseases of the Kidney
  - Polycystic Kidney Disease
  - Solitary Renal Cyst
- Duplication of Renal Pelvis and Ureter
- Retrocaval Ureter
- Uretrocele
- Injuries to Kindney
- Renal Tuberculosis
- Hydronephrosis
- Pyonephrosis
- Carbuncle of Kidney
- Perinephric Abscess
- Renal Calculus
- Ureteric Calculi
- Staghom Calculus
- Benign Tumours of Kidney
- Wilm's Tumour
- Renal Cell Carcinoma
- Approaches to Kidney

ANATOMY OF KIDNEY AND URETER

KIDNEY—ANATOMY

- They are a pair of excretory organs situated retroperitoneally, on the posterior abdominal wall; one on each side of vertebral column.
- Vertically it extends from the upper border of T_{12} vertebra to the centre of the body of L_3 vertebra. The right kidney is slightly lower than the left, left kidney is slightly nearer to the median plane.
- Each kidney is about 11 cm long, 6 cm broad, 3 cm thick. Lateral border is convex. Medial border is concave with a hilum. Structures seen in the hilum are (from anterior to posterior)—(1) renal vein, (2) renal artery, (3) renal pelvis.
- Upper pole is related to suprarenal gland; lower pole lies 1 inch above the iliac crest.
- Capacity of renal pelvis is 10 ml.
- The angle between the 12th rib and the outer border of the sacrospinalis is the kidney angle. Kidney pain is usually referred here and pressure over this point elicit pain in kidney lesions.

- Hepatorenal pouch is a peritoneum lined deep recess, related to the upper pole of kidney. It is the lowest site when the body is in horizontal position, excluding the pelvis. Collection of extravasated fluid is likely to occur in this pouch following surgeries to liver and biliary tract.

Capsules of Kidney

1. Proper capsule—fibrous membrane which can easily be stripped off from the organ.
2. Perirenal fat is in the space of Gerota.
3. Renal fascia of Gerota—has got anterior layer *(Fascia of Toldt)* and posterior layer *(Fascia of Zuckerkandl).*
4. Pararenal body of fat.

Structure of Kidney

Gross

Coronal section shows outer brownish cortex, inner pale medulla, renal sinus.

Renal sinus is a space that extends into the kidney from hilus. It contains (1) branches of renal artery, (2) branches of renal vein, (3) renal pelvis.

The pelvis divides into 2-3 major calyces which in turn divides into 7-13 minor calyces.

Histology

Each kidney is composed of 1-3 millions of uriniferous tubules; each of which has a collecting part *(collecting tubules)* and secretory part *(nephron).*

Blood Supply

It is from renal artery, which arises from the aorta at right angles, at the level of intervertebral discs between L_1 and L_2 vertebrae. About 95% of abdominal aortic aneurysm arises below the level of origin of renal artery.

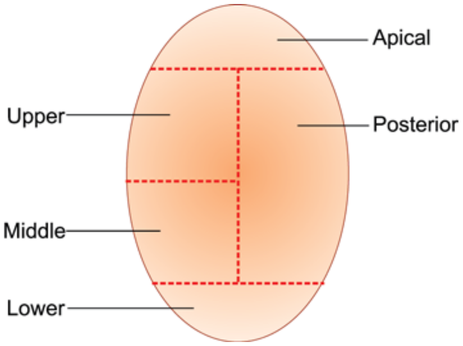

Fig. 26.1: Segments of kidney.

The kidney is divided into anatomic segments based on blood supply. In the hilum, the main artery, divides into anterior and posterior divisions. The anterior division supplies the apical, upper, middle and lower segments; posterior segment is supplied by the posterior division. The knowledge of this is important in the operation of anatrophic nephrolithotomy, where a functionally *avascular plane* (*Brodel*) between the posterior segment and the upper and middle segments exist on the posterior half of the kidney, about two-thirds of the way from the hilum to the lateral margin of the kidney.

Nerve Supply

$T_{10,11,12}$ through the lesser and lower splanchnic nerves.

Lymphatics

They are drained into para-aortic nodes. Perinephric fat is removed during nephrectomy (for malignancies), as intrarenal lymphatics freely communicates with the plexus in the perinephric fat.

URETER—ANATOMY

- Each ureter is 25-30 cm long. It begins within the renal sinus as a funnel shaped dilatation called renal pelvis.
- It enters the bladder wall obliquely to open at the lateral angle of the trigone.
- It lies in the retroperitoneal space not attached to any fixed structures, so can be displaced or obstructed by retroperitoneal masses like tumours or aneurysms.
- Abdominal portion of ureter lies on the medial portion of psoas major muscle near the tips of transverse processes of lumbar vertebra (which is very well seen in IVP).
- The ureter enters the pelvis crossing the end of common iliac or beginning of external iliac artery.
- In females, uterine artery crosses the ureter from lateral to medial side, about 2 cm lateral to the cervix (this has to be remembered while ligating the uterine arteries in hysterectomy).

Normal sites of constrictions: (1) PUJ, (2) At the brim of lesser pelvis, (3) Along its passage through the bladder wall

- ***Arterial supply:*** Upper segment of ureter receives blood supply from branches of renal and adrenal arteries; the middle portion from the branches of arteries of posterior abdominal wall; the pelvic portion from the branches of internal iliac arteries.
  - Vessels reach the abdominal portion of ureter on the medial side; whereas the pelvic portion receives on the lateral side (This has to be remembered while mobilising the ureter). The vessels form an anastomotic plexus in the adventitia of the ureter, which is found to be deficient in 10-15% of individuals leading to necrosis of cut ends of ureter following extensive mobilisation.

PLAIN X-RAY—KIDNEY, URETER AND BLADDER (KUB)

- Preparation of the patient: Enema/bowel wash/laxative is given on the previous day and the patient is asked to fast in order to reduce the bowel gas shadows in X-ray.
- High penetration X-ray is taken in supine position which covers *pubic symphysis and lower two ribs*.
- Interpreting the film:
  a. First bony parts are looked for, i.e. the hip, pelvis, lumbar vertebrae for fractures, scoliosis, spina bifida, secondaries in the spine.
  b. ***Kidney shadow***: Kidney shadows are visualised in plain X-ray KUB due to difference in the density between kidney (high vascularity) and perinephric fat (low vascu-

larity). Findings noted are size, location, calcification and stones. In children, perinephric fat is absent and so kidney shadows are not visualised.

c. ***Psoas shadow***: It is visualised well in normal KUB.

Psoas shadow is obliterated in:

- In enlarged kidney
- In scoliosis due to inflammatory or infiltrative causes
- In malignancy
- Tuberculous spine with cold abscess (psoas abscess)
- Splenic injury—in left sided shadow
- Retroperitoneal tumours

d. ***Ureteric line****:* It is looked for any radio-opaque shadow (ureteric stone). It runs along the tips of the *transverse processes of the lumbar vertebrae*, crosses the *sacroiliac joints* and reaches up to a point medial to the *ischial spine.*

e. Bladder, prostate and urethral areas are visualised for any lesion.

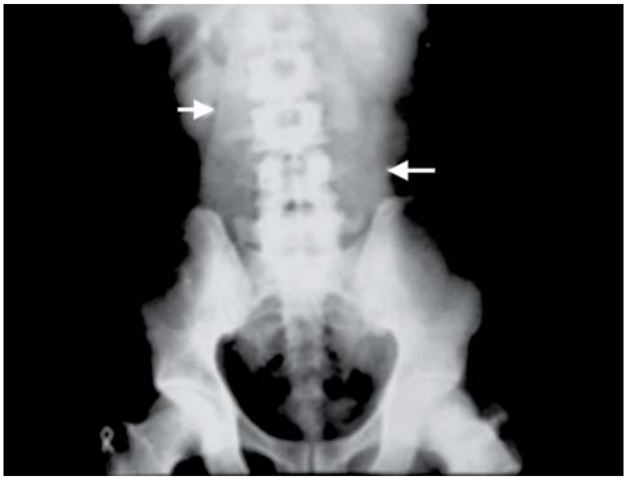

Fig. 26.2: Plain X-ray KUB AP-view. Note the psoas shadow.

INTRAVENOUS UROGRAM (IVU)

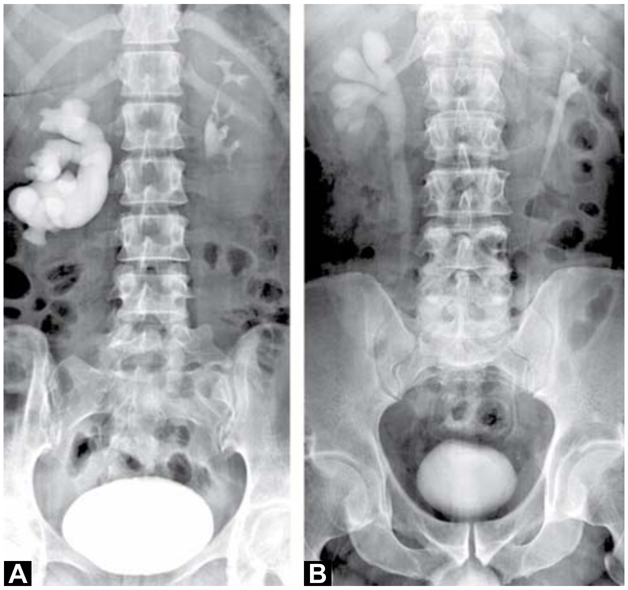

Figs 26.3A and B

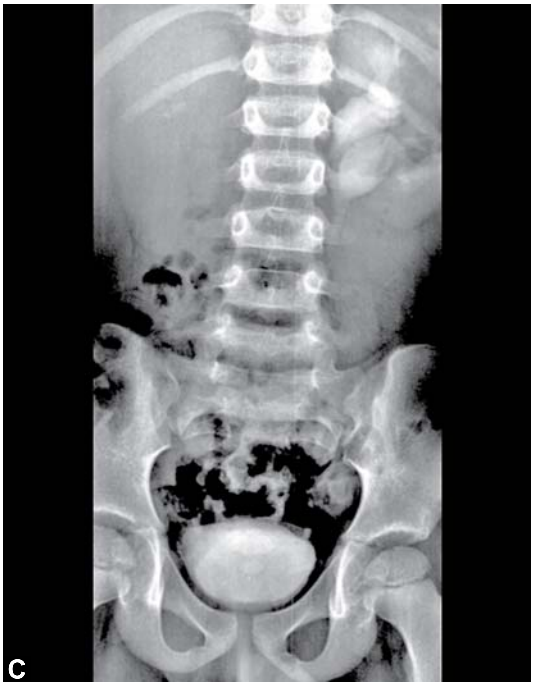

Fig. 26.3C

Figs 26.3A to C: IVU showing hydronephrosis with clubbing of calyces and dilatation. Delayed post-lasix film often should be taken whenever there is poor secretion in initial films. Film can be taken as late as 72 hours. This delayed film shows dilatation (*Courtesy:* Dr Navinchandra Shetty, MD, HOD of Radiology, KMC, Mangalore).

Procedure

Renal function must be normal.

- Overnight fasting for 8 hours is advised. Laxatives are given to reduce bowel shadow and get a good quality film.
- First, a plain X-ray KUB is taken (IVU should not be read without doing KUB).
- Then 1 ml test dose of **sodium diatrizoate** (*Urograffin)* or **meglumine iothalamate** is injected IV and waited for

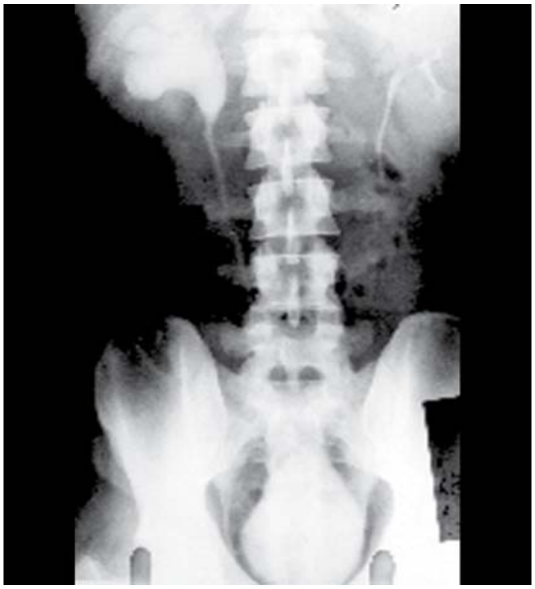

Fig. 26.4: IVU showing hydronephrosis.

IVU (Intravenous Urogram)

Indications	*Findings*
1. Hydronephrosis	Clubbing of calyces
2. Congenital anomalies	
a. Horseshoe kidney	Flower vase appearance
b. Duplex kidney and double ureter	
c. Ureterocele	Adder (cobra) head appearance
d. Polycystic kidney disease	Spider leg appearance
e. Retrocaval ureter	Reverse 'J' sign with hydronephrosis
3. Renal cell carcinoma	Irregular filling defect; spider leg appearance
4. To see the function of the kidneys in bilateral diseases	Bilateral stones, obstructive uropathy
5. After surgery for urinary diseases	To see the function of kidneys and outcome of the surgery
6. Renal injury	To see the function of other kidney (a very specific investigation)

Contraindications

1. Iodine sensitivity—may go for anaphylaxis. Hence, all precautions must be taken and essential drugs should be available while doing IVU
2. Multiple myeloma and hypergammaglobulinaemias (Acute renal failure may be precipitated due to dehydration)
3. Toxic thyroid

5-10 minutes for any reaction. If no adverse reaction occurs, then full dose 1 ml/kg body weight, IV urograffin is given (40-50 ml).

- X-ray is taken in 1-5 minutes, which shows the nephrographic and secretory function of the kidneys.
- Later 15 minutes and then 20-30 minutes films are taken.
- Further films are taken depending on the need.
- Film can be taken as late as 72 hours. Late films show bladder pathology as well as residual urine.
- In case of *renal failure* with high blood urea, dose of dye is increased to 2 ml/kg body weight to get a better film—***Infusion IVU***. Often diuretics are used in these patients to have better secretion.

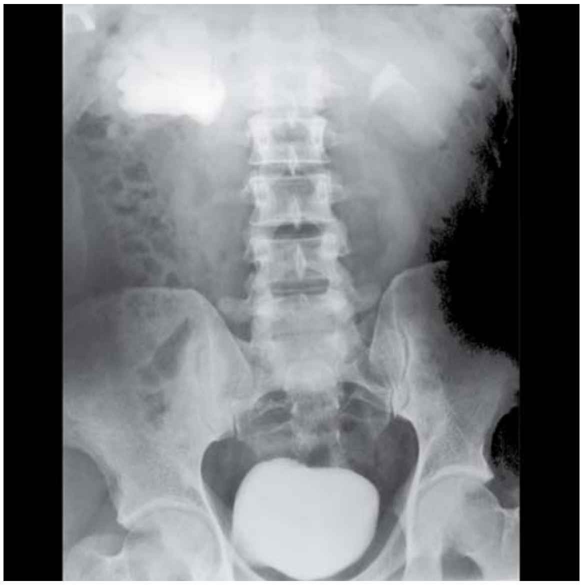

Fig. 26.5: IVU showing hydronephrosis on right side; normal secretion on left side; with bladder filling.

- Lower abdominal compression is done for 10 minutes to have better definition of calyces, but not done in children and patients with abdominal aortic aneurysm.
- ***Minute IVU***: In case of renal artery stenosis, within first minute many films are taken to see nephrographic shadow (where a small, concentrated kidney is seen).
- ***Nonvisualization of kidney:*** No contrast is seen in the film even after 12 hours.

RETROGRADE PYELOGRAPHY (RGP)

Indications

- ❖ Failure of showing any secretions in an IVU as late as 72 hours film
- ❖ Urinary tuberculosis
- ❖ Urothelial tumours from the renal pelvis

Procedure

- Under G/A, cystoscope is passed. Ureteric orifice is visualised.
- Ureteric catheter is passed. Dye, sodium diatrizoate is injected.
- Patient is put in 15° head down position to allow the dye to reach upper urinary system.
- X-ray is taken.

Advantages

a. Prior to dye injection selective urine sample can be taken from each ureter.
b. Brush biopsy from suspected urothelial tumours of upper urinary tract can be taken.
c. Better-delineation of anatomy (due to more concentration of dye).

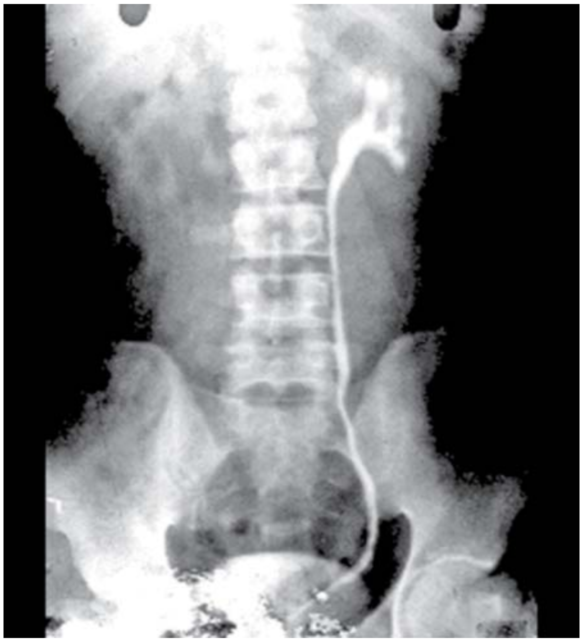

Fig. 26.6: Left RGP—normal study.

Disadvantage

Anaesthesia is required and is laborious.

RENAL ANGIOGRAM

Indications

- ❖ Renal artery stenosis
- ❖ Renal artery atheroma
- ❖ Renal artery aneurysm
- ❖ Occasionally renal cell carcinoma
- ❖ Arterial anomalies

Procedure

Retrograde Seldinger technique:

- ♦ Through femoral artery, using Seldinger's needle selective angiogram is done to visualise tumour vascularity, narrowing, anomalies. Dye used is Hypaque. Dose: 6-7 ml.
- ♦ Therapeutic embolisation, transluminal balloon angioplasty for renal artery stenosis can also be done by this approach.
- ♦ Translumbar approach for angiogram (through aortogram) is also used. Hypaque used is 30 ml.

Complications

- ❖ Paraplegia
- ❖ Embolism
- ❖ Dissecting aneurysm
- ❖ Bleeding
- ❖ Renal tubular necrosis

Renal pharmacoangiogram:

- ♦ Noradrenaline is injected along with the dye. Normal vessels will constrict in response to noradrenaline. But since tumour is autonomous, vessels in renal cell carcinoma do not respond to noradrenaline and so *tumour blush* is seen.

MICTURATING CYSTOURETHROGRAPHY (MCU)

Indications

a. Vesicoureteric reflux.
b. Posterior urethral valve.

Procedure

Catheter is passed into the bladder. Dilute iodine dye is infused. X-ray is taken during micturition. *Free reflux* is looked for. Applying pressure over the suprapubic region, X-ray is taken. *Pressure reflux* is studied.

Vesicoureteric reflux is graded depending on the severity of the reflux—as:

I—Ureters seen
II—Ureters and pelvis are seen
III—Ureters, pelvis, calyces are seen
IV—With grossly distended calyces
V—Tortuous elongated serpentine ureters

It can be unilateral or bilateral. Often it is associated with posterior urethral valve. It is often complicated by infection, pyonephrosis and renal failure

Investigations: MCU, IVU, U/S, blood urea and serum creatinine

Treatment: Tailoring of ureter with reimplantation

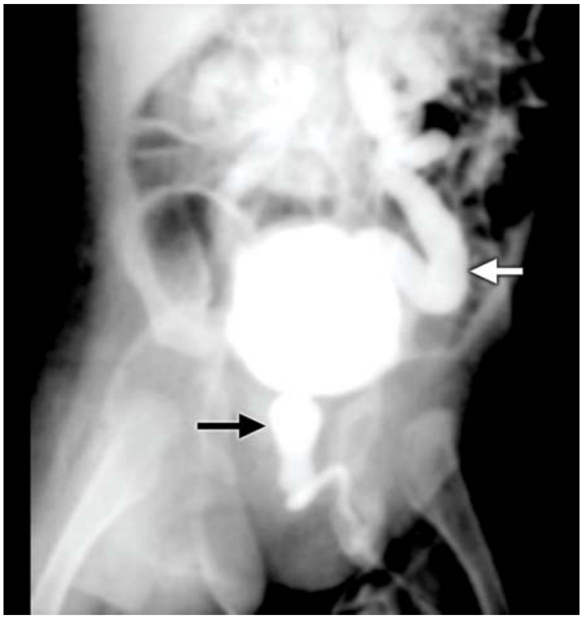

Fig. 26.7: Micturating cystourethrogram showing concomitant existence of posterior urethral valve (causing dilatation of proximal urethra) and vesicoureteric reflux (left side).

ASCENDING URETHROGRAM

- ♦ It is the investigation of choice for *stricture urethra*.
- ♦ Red rubber catheter is passed into the external meatus.

If someone is greedy, bribe him, if someone is foolish, advise him; if someone is wise, listen to him.

- Water soluble iodine dye is injected through the catheter.
- Oblique X-ray films are taken to visualise the urethra.
- Site, size, extent of stricture and extravasation can be found out in urethrogram.

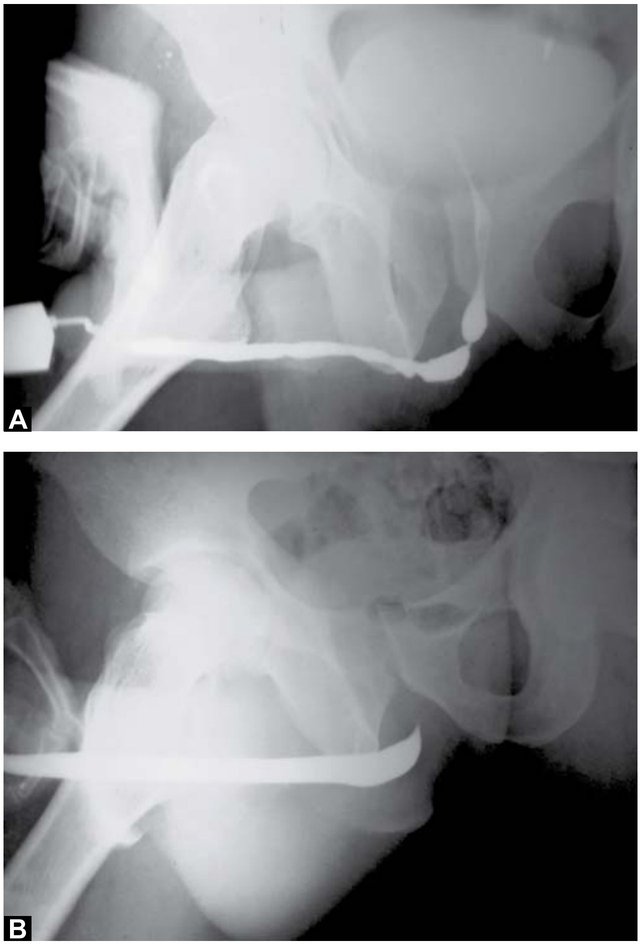

Figs 26.8A and B: Ascending urethrogram showing stricture urethra.

ISOTOPE RENOGRAPHY

- A measure of individual kidney function is obtained by this method using a gamma camera.
- Radiolabelled Technetium 99m **DMSA** (Dimercaptosuccinic acid) **or DTPA** (Diethylenetriamine-penta acetic acid) is given intravenously.

DTPA—renal function.

DMSA—renal parenchymal changes.

It shows:

a. *Early* vascular phase.
b. *Then* secretory phase.
c. *Later* excretory phase.

- This allows the assessment of renal plasma flow to each kidney and the efficiency and effectiveness of pelvicalyceal excretion also.
- Secretion < 20% is precarious.

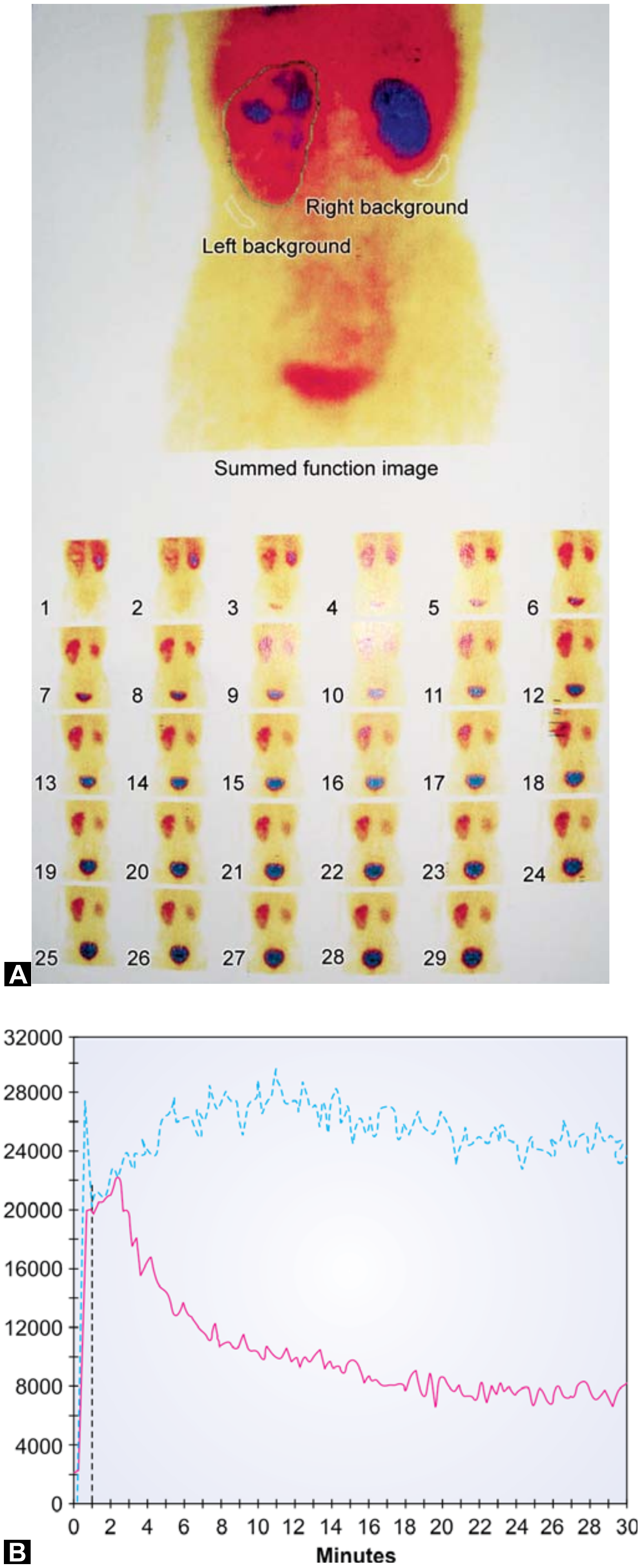
Right background
Left background
Summed function image
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
A
32000
28000
24000
20000
16000
12000
8000
4000
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Minutes
B

Figs 26.9A and B: Isotope renogram showing reduced secretion on right side.

Problems

1. Positioning of counters.
2. Often difficult to differentiate from muscle mass. It is only a supportive investigation.

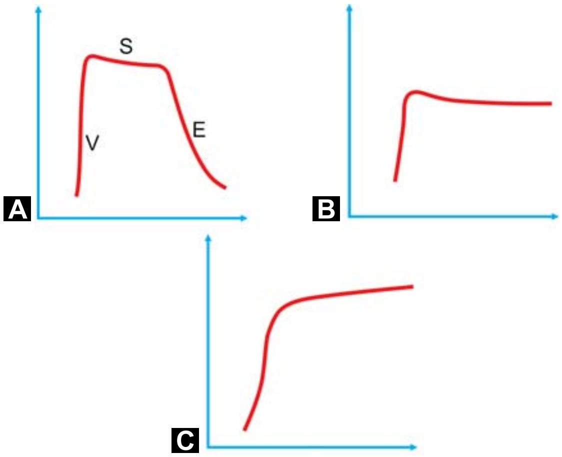

Figs 26.10A to C: Renography showing (A) Normal study, (B) Hydronephrosis and (C) Renal failure.

CYSTOSCOPY

Uses

- To examine urethra, bladder, ureteric orifice for any pathology (tumour, infection)
- To visualise any bladder fistulas
- To treat
  - Urethrotomy (in stricture urethra)
  - TURP for BPH and carcinoma prostate
  - Bladder tumour resection
  - Bladder stone removal (cystolithotripsy, cystolitholapaxy)
  - Ureteric catheterisation
  - Fulguration of posterior urethral valve

Contraindication: Acute cystitis and prostatitis

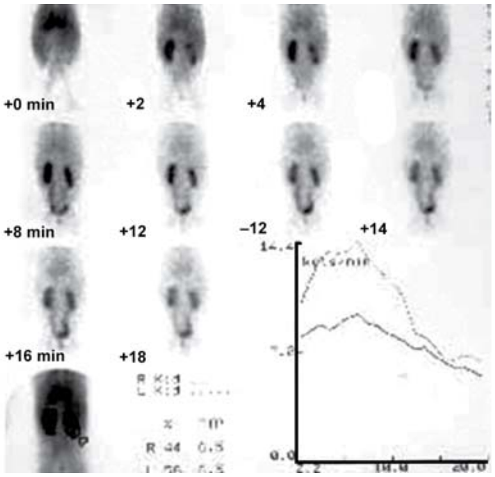

Fig. 26.11: Isotope renogram showing kidney function.

Types

- ***Rigid.***
- ***Flexible.***

Procedure

- Patient is placed in lithotomy position.
- Under G/A or spinal anaesthesia after cleaning and draping, through urethra, cystoscope is passed using continuous *glycine irrigation* (to avoid TURP syndrome).
- The parts of the urethra are visualised while passing the cystoscope.
- Once bladder is reached, it is looked for diverticula, hypertrophy and other pathologies.
- Ureteric orifices are visualised.

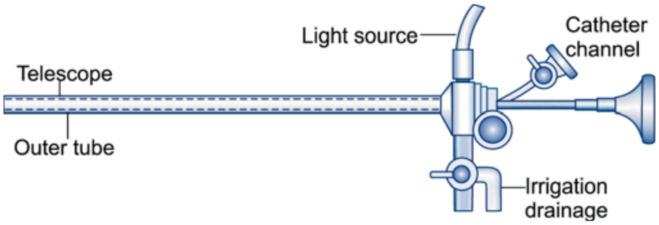

Fig. 26.12: Cystoscope.

Complications

- Urethral injury.
- Bleeding.
- Water intoxication.

Fig. 26.13: Cystoscopy is being done for different procedures like TURP; diagnostic, etc. (*Courtesy:* Dr Ashok Pandit, MCh, Urologist, Mangalore).

CATHETERS

They are hollow tubes used to relieve urinary retention, obtain urine for analysis, irrigate bladder and to instill drugs into bladder.

The skilful doctors know what is wrong by observing alone, the middling doctor by listening and the inferior doctor by feeling the pulse.
—Chang Chung-Ching

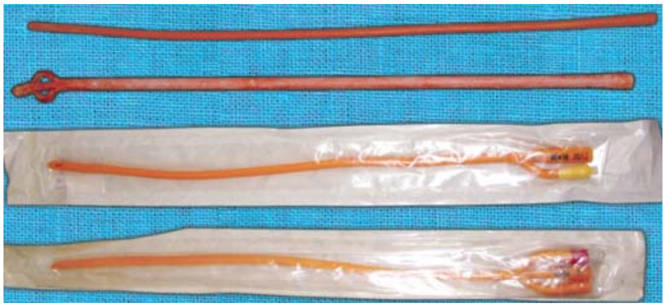

Fig. 26.14: Photo showing red rubber catheter; Malecot's catheter and Foley's catheters.

Types

a. Nonself-retaining catheter: Simple red rubber catheter.

b. Self-retaining catheter: Foley's catheter, Malecot's catheter, Gibbon's catheter, De-Pezzer catheter.

Types of Catheterisation

a. ***Indwelling catheterisation***: When a catheter is left behind in bladder and remains so it is called an indwelling catheter. *It is achieved by:*
   - Ballon tip of catheter—Foley's catheter.
   - Flower tip of catheter—Malecot's catheter, De- Pezzers.
   - Straping catheter externally—Gibbon's catheter.

b. ***Intermittent catheterisation***: A sterile catheter is introduced intermittently by the patient or by others.

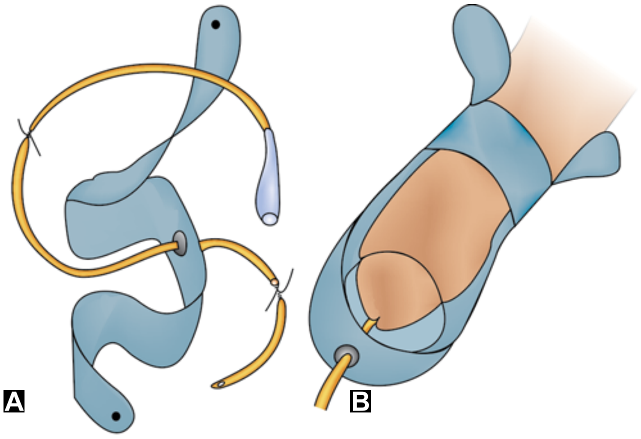

Gibbon's catheter with strap

Catheter in male

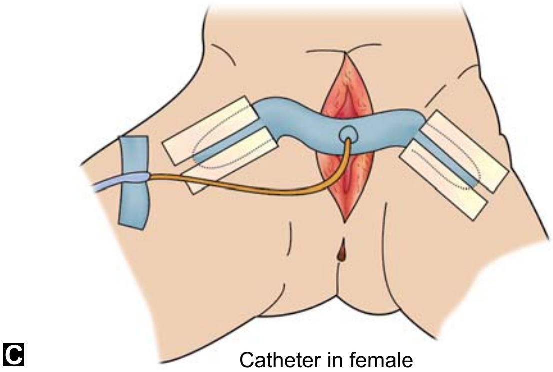

Catheter in female

Figs 26.15A to C: Gibbon's catheter. Note the look and method of fixing in male and female. It is a self-retaining catheter made up of plastic with a stillete in it which makes its passage easier through urethra. Sizes differ for males and females because of the difference in length of the urethra in male and female. Catheter has two ribbons attached which can be used to fix to the genitalia by adhesive plaster. Gibbon's catheter is least irritant to urethra.

FOLEY'S CATHETER (Fredrick Eugene Basil Foley—American Urologist)

It is a *self-retaining* urinary catheter made of *latex*. It has got a balloon near the tip into which distilled water is infused to make it *self-retainable*. Usually, Foley's catheter is kept for 7 days. *Sterilized* by γ-radiation.

Size: Adults—16 F.
Children—8 F or 10 F.

(F—French unit, Charriere unit, where each unit equals 0.33 mm). 16 F means circumference of the catheter is 16 mm. Diameter is one-third of circumference).

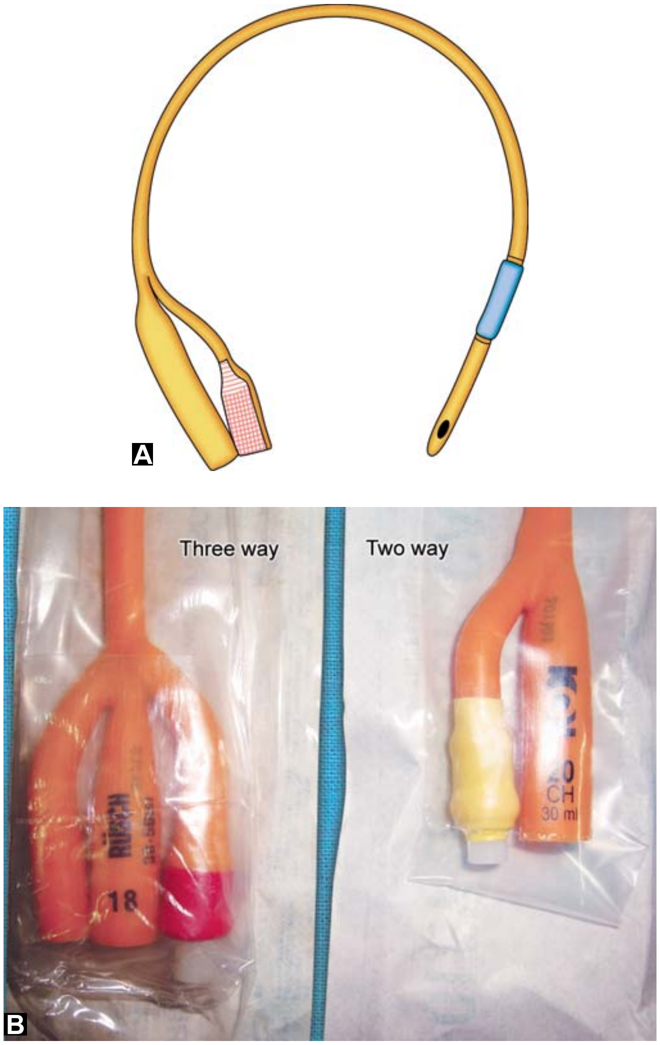

Figs 26.16A and B: Foley's catheter. Two way/three way Foley's catheters are available.

Uses

1. To pass per urethrally in retention of urine of any cause (BPH, stricture, trauma).

2. To measure the urine output in renal failure, postoperative patients, terminally ill patients.
3. Percutaneous cystostomy.
4. Cholecystostomy.
5. To drain fistulas.

Types

1. Two-way Foley's.
2. Three-way Foley's—to give bladder irrigation, e.g. following TURP.
3. Silicon coated Foley's—to reduce reaction and so as to keep for longer period (3 months).

Procedure

- After cleaning under strict asepsis, lignocaine gel is lubricated onto the urethral meatus. Catheter is passed into the urethra.
- Sometimes *Maryfield introducer* is used to pass Foley's catheter.
- Once catheter is in the bladder, urine flows out. It is now connected to a *urosac bag*.
- Balloon is inflated with 20-30 ml (amount is written on the catheter) of distilled water to make it self-retainable.
- During removal of the catheter same amount of water should be removed from the balloon before pulling out the catheter.

Complications

- Infection
- Encrustation
- Bleeding
- Stone formation
- Blockage
- Stricture
- Difficulty in removal of the catheter due to blockage of the balloon channel. Here bulb of Foley's can be punctured either from above under U/S guidance or injecting ether into the balloon so as to burst it but may cause chemical cystitis or passing a stilette into the channel

MALECOT'S CATHETER

- Self-retaining urinary catheter with an umbrella or flower at the tip. It is made of red rubber, contains sulphur and so it is radio-opaque.
- It is never introduced per urethrally.
- It is sterilised by boiling.

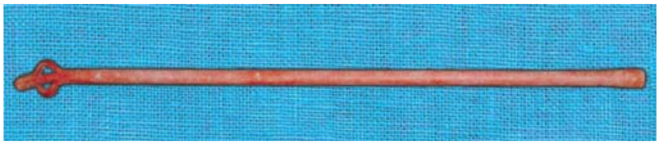

Fig. 26.17: Malecot's catheter—photo.

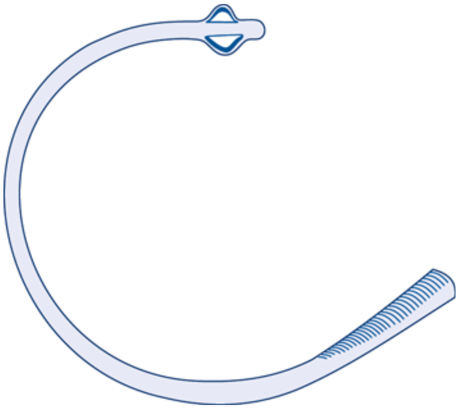

Fig. 26.18: Malecot's catheter—diagram.

Uses

- Suprapubic cystostomy (SPC)
  - In case of urinary retention when Foley's catheterisation fails (after two trials)
  - For diversion of urine following bladder, prostate or urethral surgeries
- Perinephric abscess
- Pyonephrosis
- Cabot's nephrostomy
- Subphrenic abscess
- Cholecystostomy
- Infected pseudocyst of the pancreas
- Amoebic liver abscess
- Gastrostomy

Advantages

- Malecot's catheter can be kept for a longer duration (3 months).
- It drains fluid adequately.
- Less infection rate.
- Removal is easier.

Disadvantage

- Surgery (open method) is required to insert the catheter.

RED RUBBER CATHETER

- It is a non self-retaining urinary catheter.
- Its tip is blunt. It has got only side opening.

Uses

1. Used to drain urine from the bladder temporarily.
2. To find out residual urine. After passing urine, catheter is introduced into the bladder. The amount of retained urine is measured. If it is more than 30-50 ml it signifies obstruction. It often increases up to 200-300 ml in conditions like BPH.

When the patient dies the kidneys may go to the pathologist, but while he lives the urine is ours. It can provide us a serial story of the major events going on within the kidney.
—Thomas Addis

3. Other uses
   - To administer nasal oxygen.

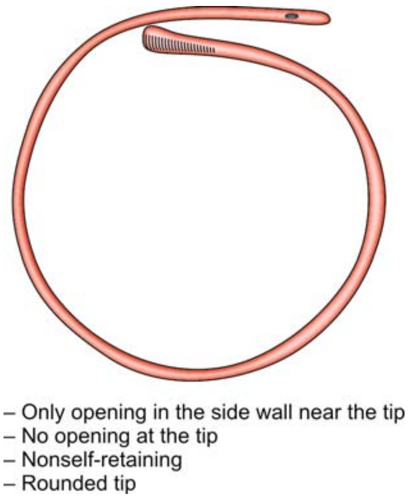

Fig. 26.19: Simple red rubber catheter. It has got only side opening. No opening in the tip (Flatus tube has got opening on both sides as well as at the tip).

   - For throat suction.
   - As a tourniquet for venesection and surgeries of fingers and toes.

NEPHROSTOMY

Nephrostomy is diversion of urine by passing a tube through the kidney into the calyces**.**

Indications

- Pyonephrosis
- Severe hydronephrosis
- Postoperative—following renal surgeries
- Obstruction in ureters

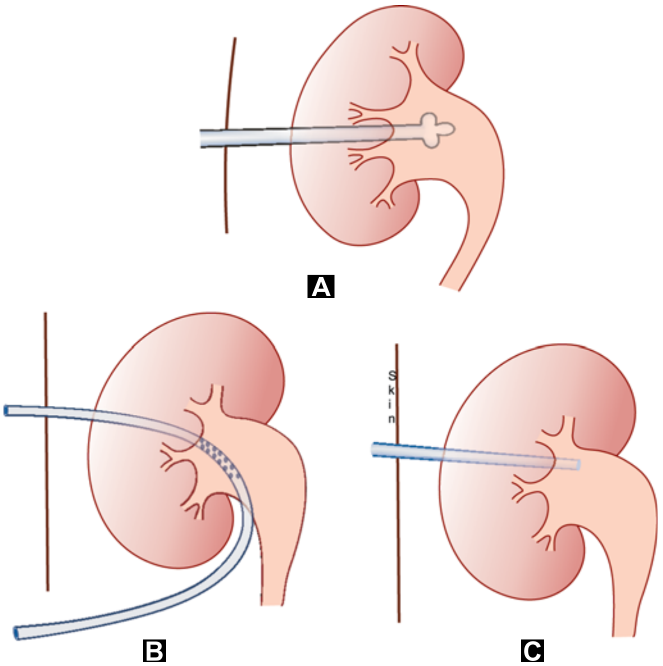

Figs 26.20A to C: Nephrostomy types: (A) Cabot's nephrostomy, (B) Tresidder's nephrostomy and (C) Percutaneous nephrostomy.

Types

1. ***Cabot's nephrostomy***: Malecot's catheter is placed into the calyces.
2. ***Tresidder's through and through loop nephrostomy***: Loop tube is placed with central part in the kidney and ends outside percutaneously to drain urine. This achieves double drainage. New tube can be placed using rail road technique.
3. ***Percutaneous nephrostomy***: Done under U/S guidance using guide wire and dilators.

Complications of nephrostomy

- Bleeding
- Sepsis
- Blockage

SUPRAPUBIC CYSTOSTOMY (SPC)

- It is placing of Malecot's catheter into the bladder above the pubis by open method, or percutaneously under the guidance using Foley's catheter.
- It is a temporary opening through the abdominal wall into the bladder.

Prerequisite

Bladder must be full and is confirmed by dullness below the umbilicus or by U/S.

Indication

Retention of urine, when urethral catheterisation fails.

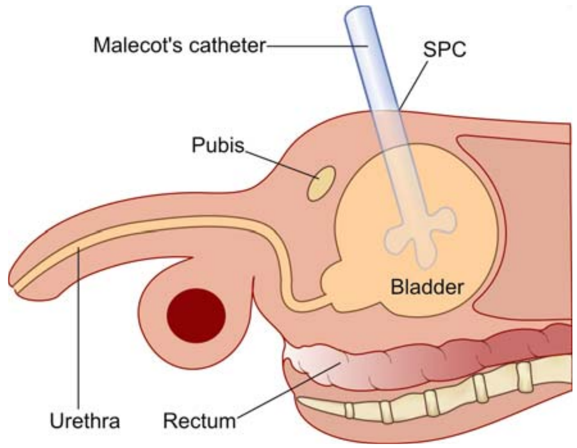

Fig. 26.21: Suprapubic cystostomy (SPC). Malecot's or Foley's catheter can be used.

Procedure

- Under local anaesthesia (Xylocaine 2% above the pubis, in the midline) or G/A, a vertical midline incision of 3 cm in length is placed through linea alba. Skin, fascia, anterior rectus sheath are incised. Recti are retracted. In extraperitoneal space, peritoneum with pad of fat is reflected upward.
- Bladder is identified by the pattern of detrusor muscle and vesical venous plexus and is also confirmed by aspirating urine through a syringe.

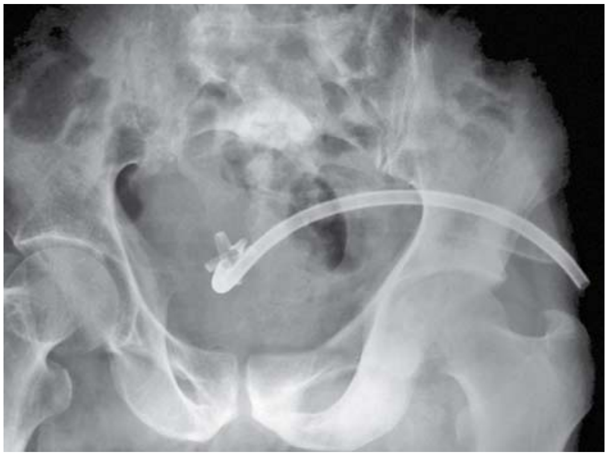

Fig. 26.22: X-ray showing radio-opaque Malecot's catheter.

- The bladder is opened near the fundus. Urine is aspirated. Bladder wall is inspected for any pathology.
- Malecot's catheter is straightened using artery forceps and placed in the bladder. Sutures are placed around the Malecot's catheter. Wound is closed in layers.
- In percutaneous SPC, Foley's catheter is passed into the bladder using trocar and cannula.

Complications

- Injury to bowel, peritoneum.
- Infection.

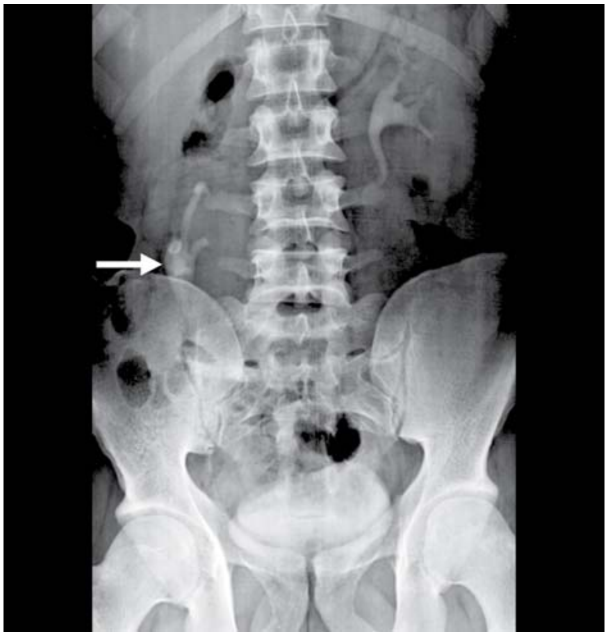

Fig. 26.23: Ectopic kidney right sided IVU picture. It is un-ascended kidney. It often presents as mass in the right iliac fossa. CT scan is diagnostic. Differential diagnoses for masses in right iliac fossa are nodal mass; carcinoma caecum; tuberculosis; psoas abscess; retroperitoneal tumour. Sepsis; stones; tumour; can occur in ectopic kidney. Management of whatever problems arise is done through open method. Bilateral ectopic kidney may lead into eventual renal failure. Isotope scan is diagnostic.

HAEMATURIA

Types

a. *Gross* (visible to unaided eye).
b. *Microscopic* (> 5 RBC's/HPF).

- *Early* (initial) haematuria: Urethral origin, distal to external sphincter
- *Terminal* haematuria: Bladder neck or prostate origin
- *Diffuse* (total) haematuria: Source is in the bladder or upper urinary tract

False haematuria: Discolouration of urine from pigments such as food colouring and myoglobin.

Silent haematuria is due to tumours of kidney or bladder unless proved otherwise.

Causes

- Renal injury
- Urinary stones
- Wilm's tumour
- Tuberculosis
- Renal cell carcinoma
- Cystitis
- Bladder tumour
- Urinary bilharziasis
- BPH, carcinoma prostate
- Renal infarct
- Glomerulonephritis
- Blood dyscrasias

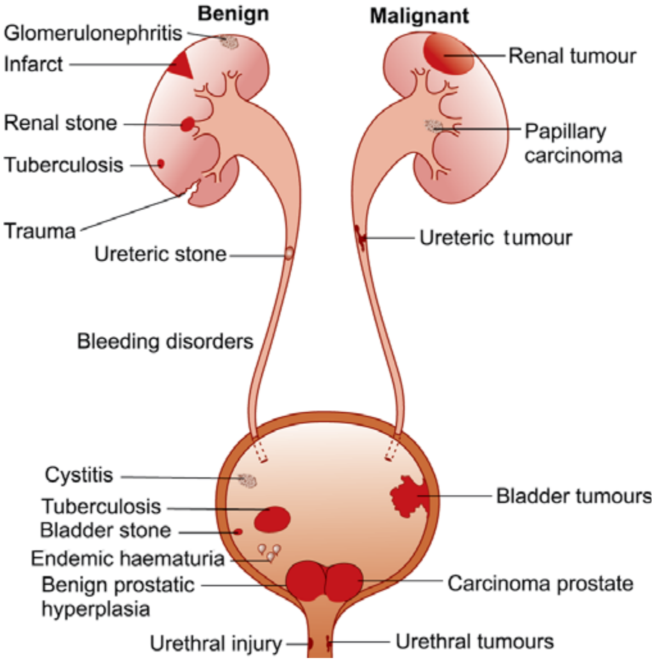
Benign
Malignant
Glomerulonephritis
Infarct
Renal stone
Tuberculosis
Trauma
Ureteric stone
Bleeding disorders
Renal tumour
Papillary carcinoma
Ureteric tumour
Cystitis
Tuberculosis
Bladder stone
Endemic haematuria
Benign prostatic hyperplasia
Urethral injury
Bladder tumours
Carcinoma prostate
Urethral tumours

Fig. 26.24: Causes of haematuria.

Good health is a serious business: like life itself, it has to be worked at and it takes on added meaning with effort.
—Normour Cousins

Investigations

- Urine culture and sensitivity (urine test for haematuria—*Benzidine test).*
- Ultrasound to look for the stone, tumour in the urinary tract.
- Cystourethroscopy to look for bladder or urethral pathology.
- IVU look for function of the kidneys.
- Urinary cytology for diagnosing urothelial malignancy.
- Bleeding time; clotting time; prothrombin time; platelet count.
- CT abdomen.
- Renal function tests—blood urea, serum creatinine.

Management

- Cause should be identified and treated.
- Blood transfusion.
- Antibiotics.
- Nephroureterectomy for RCC; removal of stone from kidney, ureter, urinary bladder.
- Treatment of bladder tumour by cystoscopic resection; intravesical chemotherapy using BCG; radiotherapy; systemic chemotherapy.
- Treatment of medical causes like glomerulonephritis.
- Correction of BPH.
- Correction of bleeding diathesis.

HORSESHOE KIDNEY

- It is a developmental anomaly where there is failure of complete ascent of kidneys with the fusion of lower or upper poles. It is due to fusion of subdivisions of mesonephric duct, when the embryo is as early as 30-40 days old.
- This condition is common in males.
- Fusion of lower pole is common (rarely upper poles).

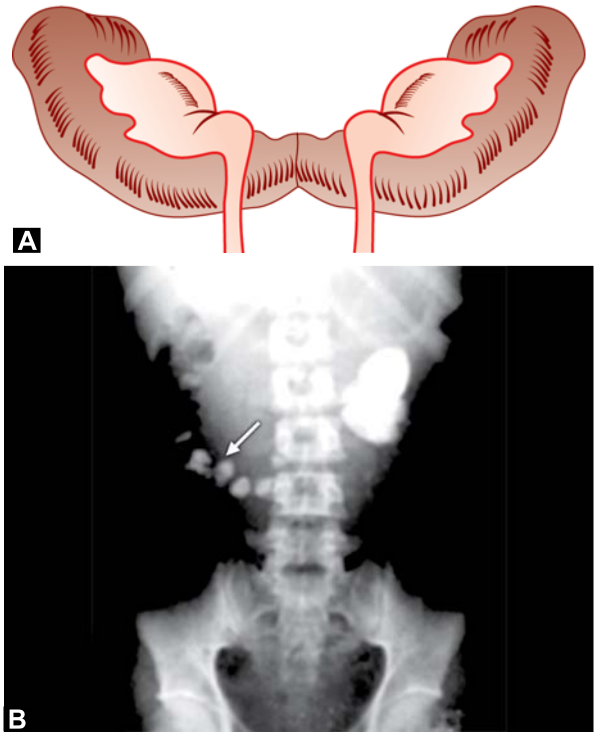

Figs 26.25A and B: (A) Horseshoe kidney and (B) Plain X-ray KUB showing horseshoe kidney with renal stones.

- Most common site is in front of 4th lumbar vertebrae. The part in front of the vertebra is called as *isthmus*. It has blood supply which freely communicates one kidney to other. Isthmus usually lies in front of aorta.

Clinical Features

- Presents as a fixed, nonmobile, firm mass in the midline at the level of 4th lumbar vertebra which is resonant on percussion.
- It is more prone for infection, stone formation, hydronephrosis, tuberculosis.

Diagnosis

- IVU—medialisation of lower calyces and curving of ureter like a ***'flower vase'.***

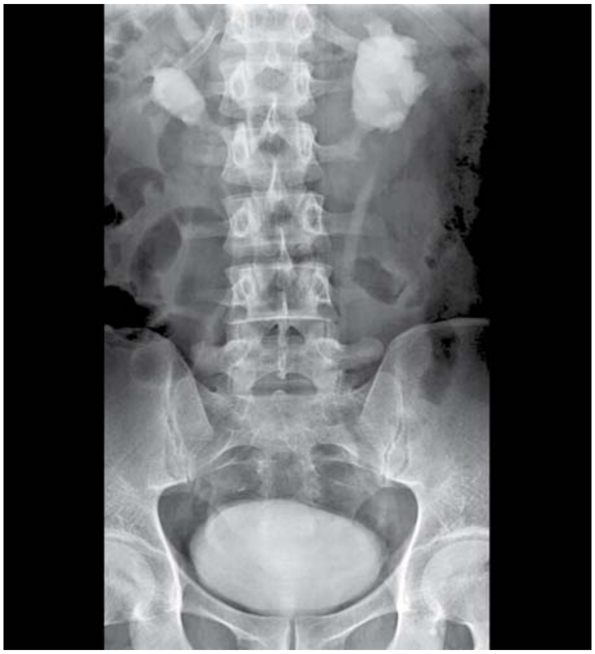

Fig. 26.26: IVU showing flower vase appearance of horseshoe kidney.

- U/S abdomen.
- Urine analysis, blood urea and serum creatinine are supportive investigations.
- CT abdomen.

Treatment

- Whatever the complication occurs in horseshoe kidney, it is treated accordingly.
- *Per se* separation of isthmus is not indicated, unless to approach aorta for aortic diseases.

CYSTIC DISEASES OF THE KIDNEY

Types

a. ***Genetic***
  - Adult polycystic kidney disease (Autosomal dominant).
  - Infantile polycystic kidney disease (Autosomal recessive)—fatal.

b. ***Nongenetic***: Simple cyst, multicystic kidney, medullary sponge kidney.

c. ***Acquired renal cystic disease*** may develop in patient on long-term dialysis.

POLYCYSTIC KIDNEY DISEASE (PCKD)

- Adult PCKD is inherited as autosomal dominant disease. It is common in females.
- It is bilateral and presents in third decade. One side presents little earlier than other side.
- *Associations*
  - Polycystic diseases of liver (18%), pancreas and lungs.
  - Berry aneurysm in the circle of Willis.
- Cyst formation occurs at the junction of the distal tubule and the collecting duct.
- Grossly it contains multiple cysts with a clear or brownish fluid (due to haemorrhage).

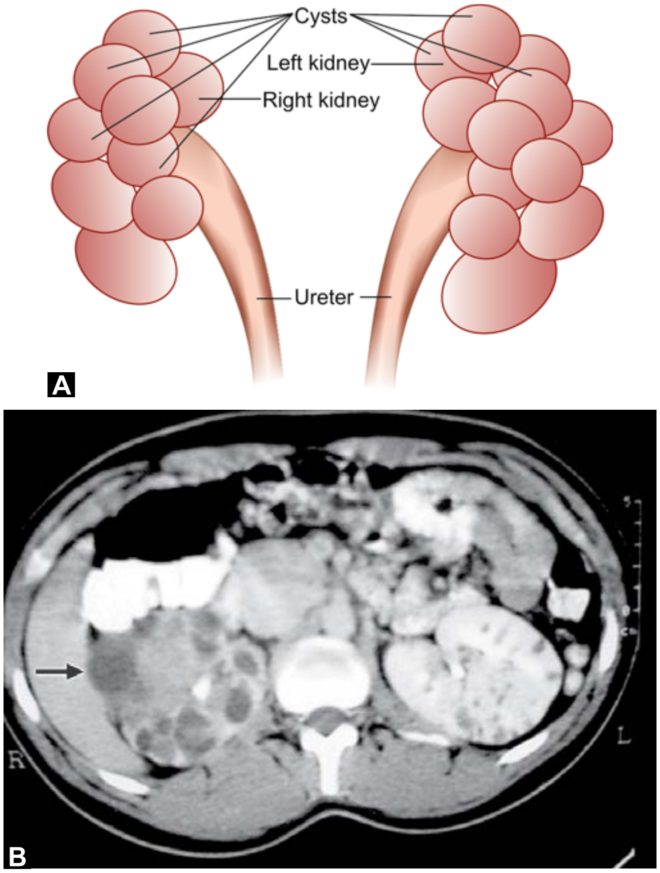

Figs 26.27A and B: Polycystic kidney disease, picture and CT scan. It is bilateral, multiple.

Clinical features

- ❖ Bilateral palpable renal mass
- ❖ Loin pain
- ❖ Haematuria
- ❖ Infection
- ❖ Hypertension
- ❖ Uraemia

- Renal mass, which is lobular, firm, mobile, moves with respiration, ballotable, with dull renal angle and resonant band in front.
- Pain, which is due to stretch of renal capsule or haemorrhage into a cyst.
- Haematuria (25%), due to overdistended cyst rupturing into the renal pelvis.
- Infection is due to stasis.
- Hypertension (75%).
- Uraemia occurs in late stage due to renal failure.

Differential diagnosis

- ❖ Renal cell carcinoma
- ❖ Hydronephrosis
- ❖ Solitary renal cyst

Investigations

- U/S confirms the presence of cysts.
- IVU—*Spider leg pattern* with an elongated compressed renal pelvis, narrowed and stretched calyces.
- Blood urea and serum creatinine.
- Urine shows low specific gravity.

Treatment

- Wait and watch policy.
- If one of the cysts overdistends causing pain, haemorrhage, infection, then surgical intervention is required.
- ***Rovsing operation***: The kidney is exposed. The cyst is opened. The fluid is evacuated. The cut edge is marsupialised.
- Presently U/S guided aspiration is done as a simpler approach.
- Laparoscopic/retroperitoneoscopic aspiration/de-roofing of the renal cyst.
- Once renal failure sets in, then initial haemodialysis followed by bilateral nephrectomy, is done and later renal transplantation should be planned for.

SOLITARY RENAL CYST

- Solitary renal cyst is never congenital.
- It is due to an earlier trauma or infection resulting in blockage of tubule, leading to cyst formation.
- It is usually unilateral, presents as a renal mass which is smooth, often tender if infected or haemorrhagic**.**

Differential Diagnosis

- Renal neoplasm.
- Hydronephrosis.
- Polycystic disease.
- Hydatid cyst.

Investigation

- U/S and IVU confirms the diagnosis.
- CT scan.

The value of experience is not in seeing much but in seeing wisely—**William Osler**

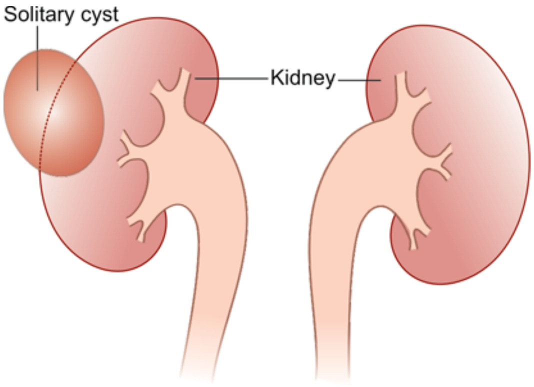

Fig. 26.28: Solitary renal cyst. It is acquired, unilateral.

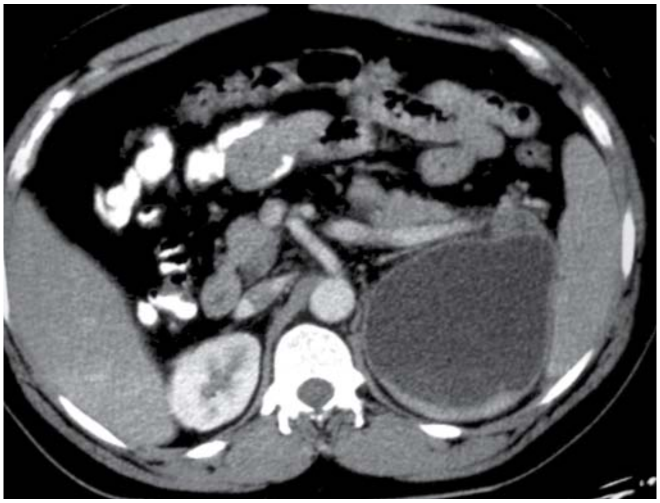

Fig. 26.29: Left side renal cyst—CT picture. It could be solitary renal cyst. It can be aspirated or de-roofed by laparoscopy/retroperitoneoscopy.

Treatment

- Kidney is exposed. The cyst is aspirated and a portion of the cyst wall is removed ***(Kirwin's operation*)** and cavity is filled with perinephric fat.
- Occasionally if the cyst is in one of the pole, partial nephrectomy is done.
- Laparoscopic approach.

DUPLICATION OF RENAL PELVIS AND URETER

- It is most common congenital anomaly of the upper urinary tract (4%).
- Usually unilateral. Common on the left side.
- In 3% of cases it is associated with duplication of ureter.
- Upper renal pelvis is small, drains the upper calyces. Lower renal pelvis is larger, drains the middle and lower calyces.
- Double ureter when associated, may be *partial* where two ureters join in lower third or *complete* where upper ureter opens into the bladder at a lower level and lower ureter opens into the bladder at the upper, normal ureteric orifice. This is called as "*Weigert Meyer Law*".
- In partial duplex, there is reno-renal reflux resulting in infection, stone formation and hydronephrosis.

Investigations

- IVU—diagnostic.
- U/S to look for complications**.**
- Cystoscopy shows **double ureteric orifices** on the same side.
- DTPA scan to see the function.

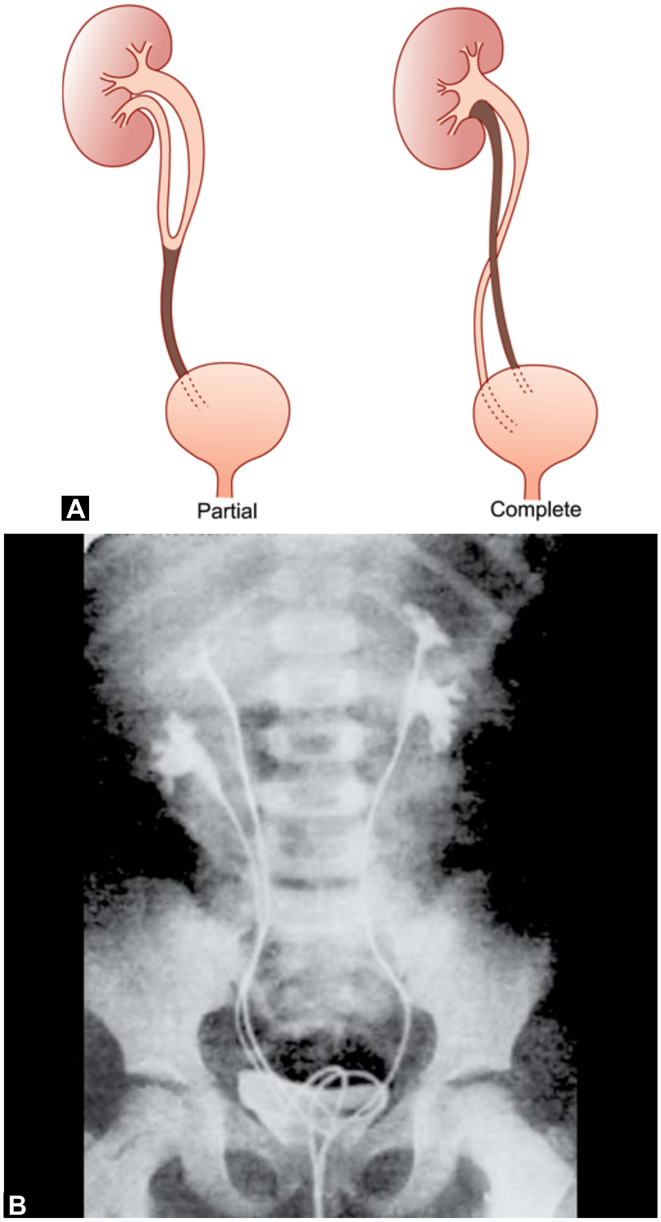

Figs 26.30A and B: (A) Duplex renal system and (B) X-ray showing bilateral duplex system.

Treatment

- Ureteric meatotomy is done if there is narrowing of the orifice.
- The co-existing complications are treated.

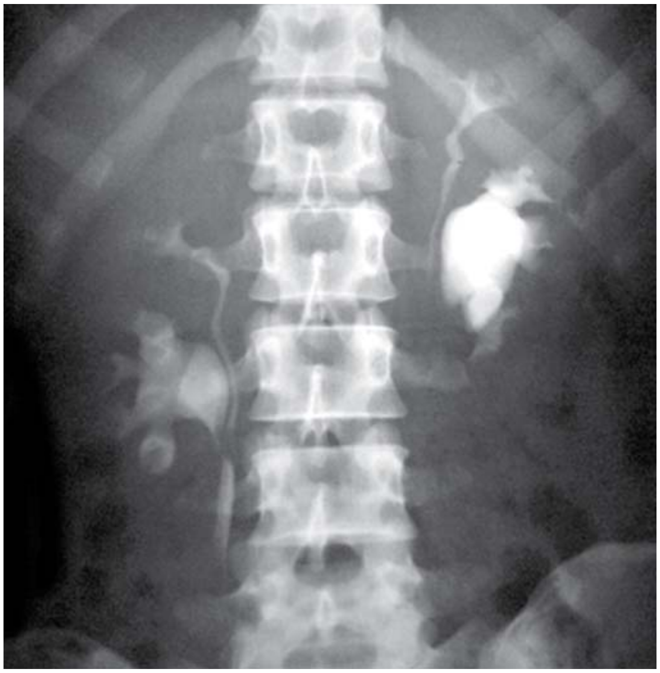

Fig. 26.31: IVU showing bilateral duplex kidney.

- Often heminephrectomy including removal of corresponding ureter may be essential as treatment.
- In females with complete duplication, lower ureteric orifice is ectopic, causing urinary incontinence which needs partial nephrectomy or ureteric reimplantation.

RETROCAVAL URETER

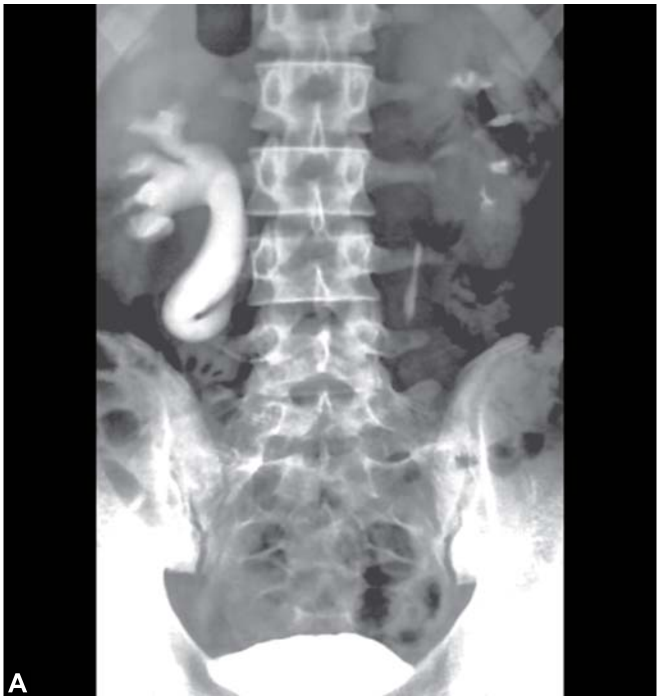

Fig. 26.32A

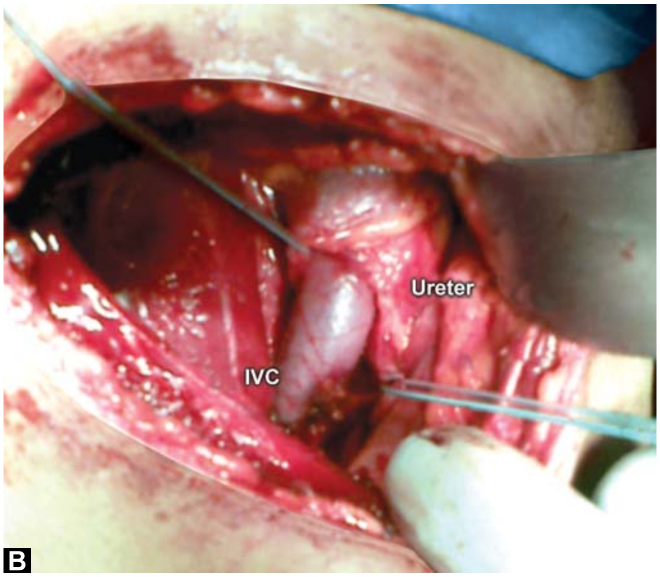

Fig. 26.32B

Figs 26.32A and B: Retrocaval ureter—IVU picture showing reverse J sign. On table finding of retrocaval ureter. It is treated by Anderson's Hynes operation. It causes hydronephrosis. It is due to anomalous development of IVC.

- It is due to developmental defect of IVC, as a result of which right ureter passes behind the IVC, causing right sided hydronephrosis with upper third hydroureter.
- IVU shows hydronephrosis with '*reverse J sign*'.
- *Treatment:* Anderson Hynes' operation.

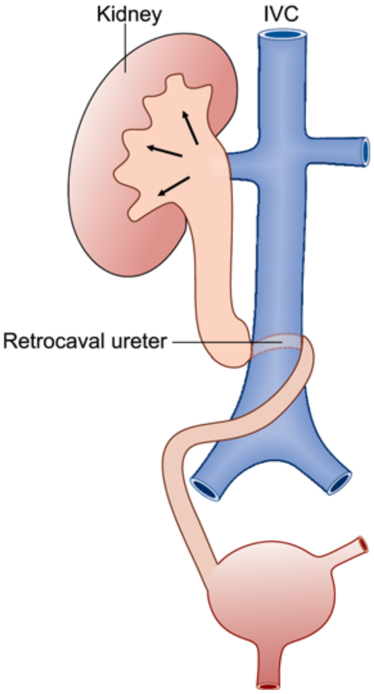

Fig. 26.33: Retrocaval ureter. It is due to developmental problem of IVC.

URETEROCELE

- It is a *cystic enlargement of the intramural portion of ureter* due to congenital atresia of the ureteric orifice. Its wall contains mucous membrane only.
- It is common in females, often bilateral (10%).

Complications

- Stone formation
- Recurrent infection
- Hydronephrosis

Stephen classification:
Stenotic, sphincteric, sphincterostenotic.

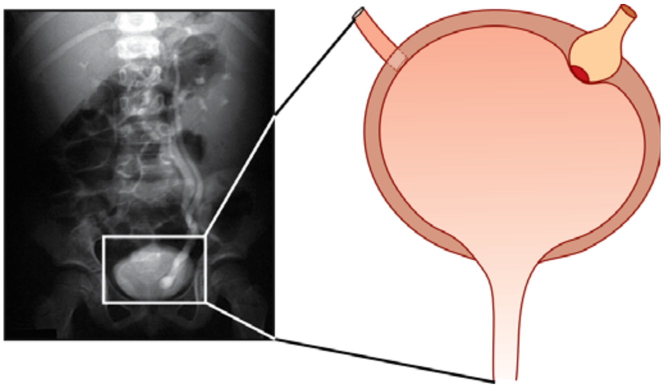

Fig. 26.34: IVU reveals left sided ureterocele with duplex kidney. Note the characteristic Cobra (Adder) head pattern of left ureterocele.

Investigations

- IVU—shows A*dder head* appearance or *cobra head* appearance.
- Cystoscopy—shows translucent cyst which is thin walled surrounding the ureteric orifice.

Treatment

- *Cystoscopic ureteric meatotomy* with the removal of cyst wall.
- In addition to that co-existing complications should be treated.
- Often ureteric reimplantation is needed.

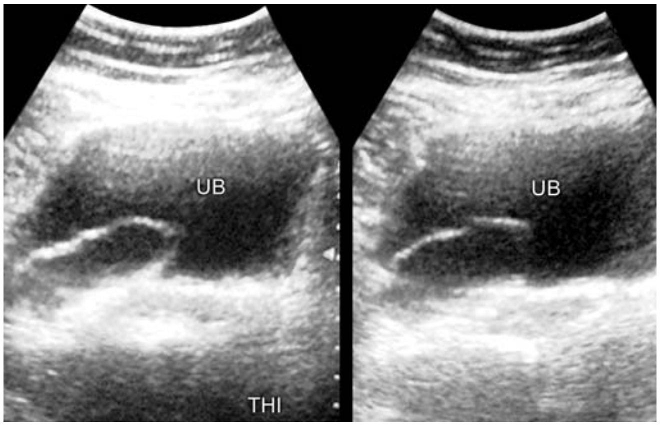

Fig. 26.35: U/S picture of ureterocele—Adder head appearance.

INJURIES TO KIDNEY

- Commonly it is due to a blunt injury.
- Often it is associated with other abdominal injuries—of liver, spleen, bowel, mesentery, etc.
- *Per se renal injury is extraperitoneal.*

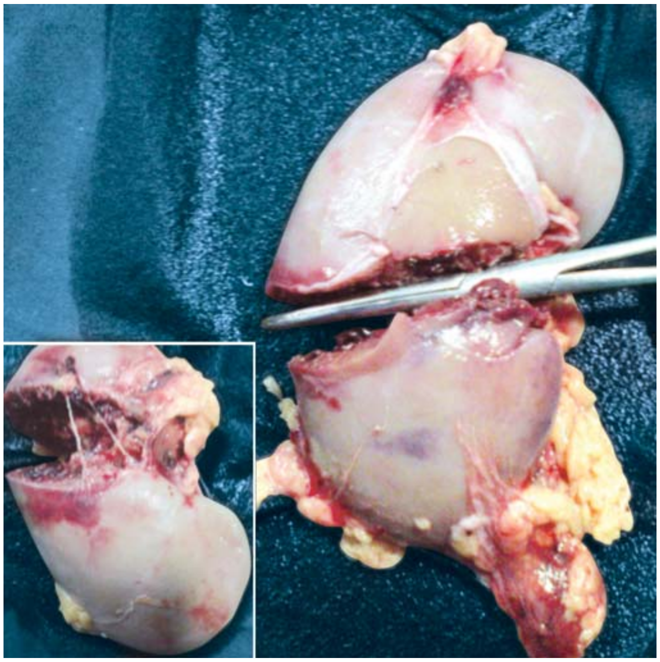

Fig. 26.36: Renal trauma causing extensive laceration. Patient underwent nephrectomy (*Courtesy:* Dr Ashfaque Mohammed, DNB; KMC, Mangalore).

Types

1. Small subcapsular.
2. Large subcapsular.
3. Cortical laceration.
4. Laceration with perinephric haematoma.
5. Medullary laceration with bleeding into the renal pelvis.
6. Corticomedullary complete rupture.
7. Hilar injury (most dangerous).

Grading of renal injury

1. Subcapsular nonexpanding haematoma without parenchymal laceration
2. Cortical laceration < 1 cm of parenchymal depth, no extravasation; perirenal haematoma
3. Cortical laceration > 1 cm depth; no urine extravasation
4. Parenchymal laceration extending through cortex and medulla with collecting system; with extravasation of urine
5. Renal pedicle avulsion; shattered kidney

Clinical Features

- Features of *shock.*
- *Haematuria*—may be mild to profuse depending on the type of injury.
- *Sudden delayed profuse haemorrhage* causing haematuria can occur between 3rd day to 3rd week after trauma.

- *Clot colic*.
- Bruising, swelling and tenderness in the loin.
- Paralytic ileus with abdominal distension occurs due to retroperitoneal haematoma implicating splanchnic nerves.

Investigations

- ***IVU (high dose)***: It is the investigation of choice. Here function of not only the injured kidney but also of the contralateral kidney can be seen. It is observed that often opposite renal artery undergoes a reflex spasm, temporarily ceasing the function of the contralateral kidney.
- ***U/S abdomen***: Done to see the type of injury, amount of haematoma and other associated injuries in the abdomen. U/S is repeated at regular intervals to see the progress (at 12-24 hourly).
- Blood urea and serum creatinine should be repeated at regular intervals.
- Blood grouping and cross-matching for blood transfusion.
- Emergency CT scan is very useful.

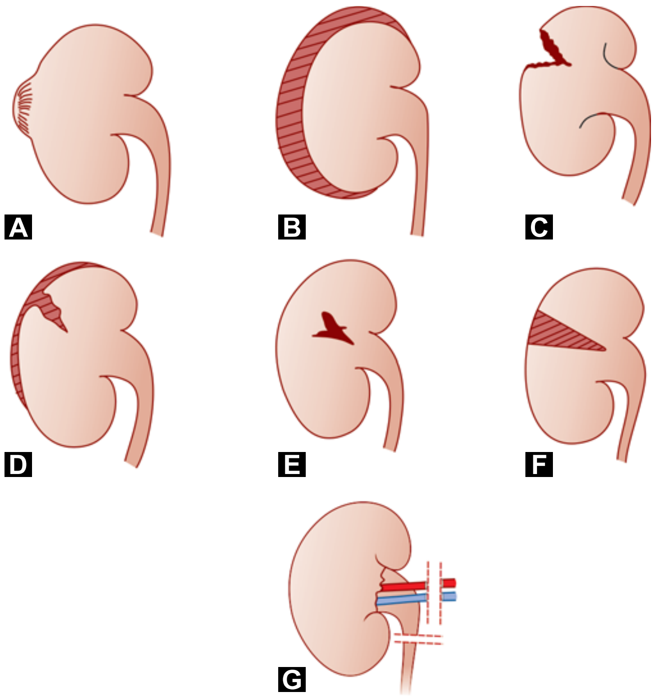

Figs 26.37A to G: (A) Small subcapsular, (B) Large subcapsular, (C) Cortical laceration, (D) Laceration with perinephric haematoma, (E) Medullary laceration with bleeding into the renal pelvis, (F) Corticomedullary complete rupture, (G) Hilar injury (most dangerous).

Treatment

I. *Initially always conservative:*
  - Catheterise and watch the urine colour and output.
  - Blood transfusion.
  - Strict observation with regular monitoring of the pulse, BP, temperature, *U/S follow-up daily*.
  - Sedation, analgesic and antibiotics.
  - 75% of patients respond to conservative management.
  - While treating conservatively, regular monitoring of blood urea and serum creatinine is a must.
  - If the patient goes in for renal failure, haemodialysis should be done for 6-8 weeks. Meanwhile, other kidney starts functioning again and patient recovers without any further problem.

II. *Indications for surgical intervention:*
  - When there are signs of progressive blood loss with the condition of the patient deteriorating.
  - Formation of progressive perinephric haematoma.
  - When there are associated other injuries.
  - Hilar injury.

Surgery (Only in 10-20% of Patients)

Options:

- Gentle suturing of the laceration. Often kidney is friable, this is not possible.
- Then nephrostomy (Cabot's) is done.
- When the injury is in the poles partial nephrectomy is done.
- In *hilar injury and severe laceration, nephrectomy is the only choice*.

	Types	*Treatment*
I	Only bruise/contusion	Conservative
II	Breach in calyceal system rupture of one of the small branches of renal artery	Conservative/ nephrectomy
III	Rupture of pelvi-calyceal system/renal substance	Nephrectomy

Complications

- Clot retention in the bladder and may go for renal failure
- Pararenal pseudohydronephrosis
- Infection
- Perinephric abscess
- Aneurysm of the renal artery
- Renal failure
- Hypertension occurs 3 months later

RENAL TUBERCULOSIS

- Commonly it is *secondary*. Primary may be in the lung.
- Tuberculous bacilluria occurs with an early lesion in the renal cortex, and the disease spreads along the ureter causing tuberculous ureteritis and stricture ureter.
- Tuberculous kidney results in any of the following pathological types.

Pathological Types

Through blood, bacteria reach the glomeruli causing caseating granuloma with Langhan's type of giant cells and epithelioid

> *Variability is the law of life, and so no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease.* **—William Osler**

cells. These granulomas coalesce to form a papillary ulcer and other consecutive different forms.

- *Tuberculous papillary ulcer.*
- *Cavernous form.*
- *Hydronephrosis*.
- *Pyonephrosis* [due to (secondary) superadded infection by *E. coli*, *Klebsiella*].

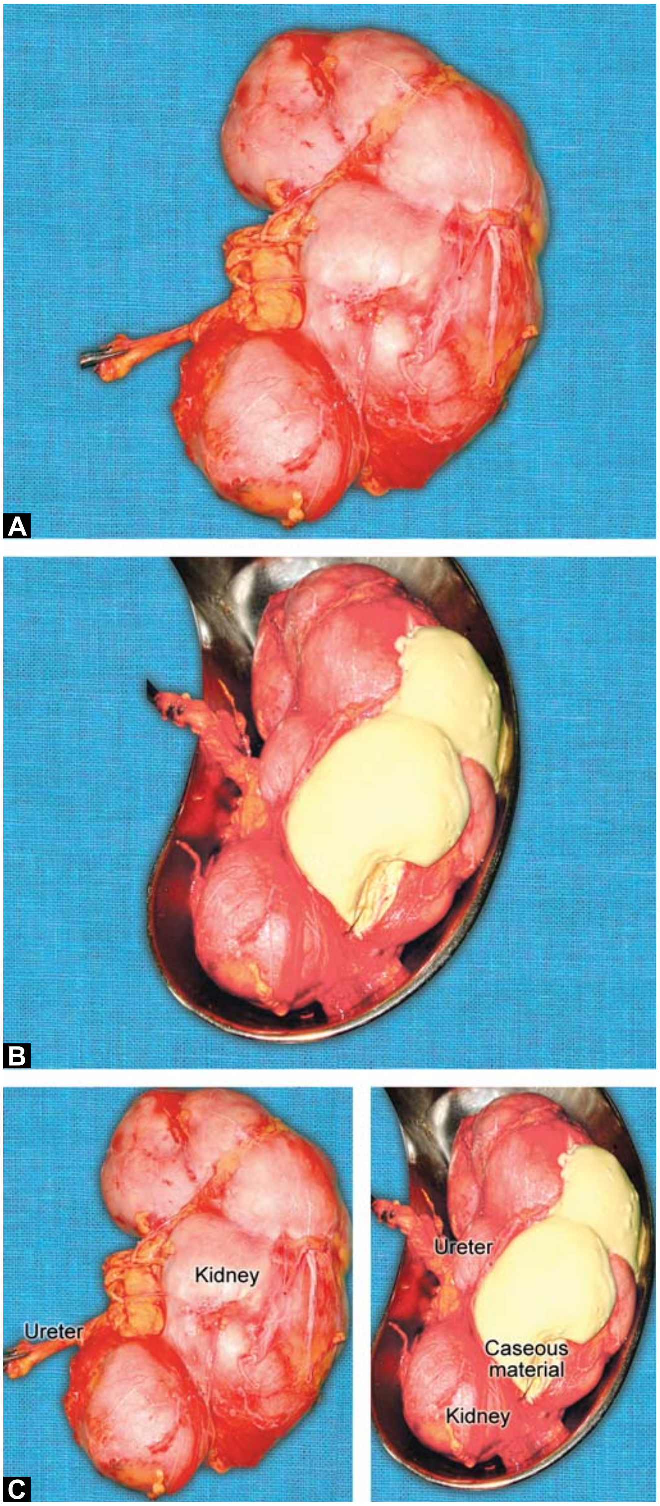

Figs 26.38A to C: Specimen of kidney showing dilatation and caseous material as content. It is tuberculous pyonephrosis. Ureter is visualised in the specimen. Often there may be ureteric stricture due to tuberculosis.

- *Tuberculous perinephric abscess*.
- *Calcified tuberculous area* (mimics calculi, hence called as pseudocalculi).
- *Caseous kidney*—often called as *putty kidney* or *cement kidney* (it goes for autonephrectomy).
- *Miliary tuberculosis*.
- Tuberculous bacilluria occurs from an early stage of the disease which causes tuberculous ureteritis and stricture ureter. Most common site is *ureterovesical junction*; second common site is pelviuretric junction.
  - Tuberculous cystitis eventually results in *golf hole ureter and thimble bladder* (cystoscopic findings). This is due to fibrosis causing rigid withdrawn dilated ureteric orifice looking like *golf hole*. Entire urinary bladder gets fibrosed, stiff and unable to dilate and accommodate urine causing *thimble systolic bladder*.
  - Tuberculous prostatitis, seminal vesiculitis (P/R—palpable seminal vesicle), tuberculous epididymitis and funiculitis are other associations. Thickened epididymis with ulcer on the posterior aspect of the scrotum may be often found. Tuberculous funiculitis with beaded, thickened vas deferens.

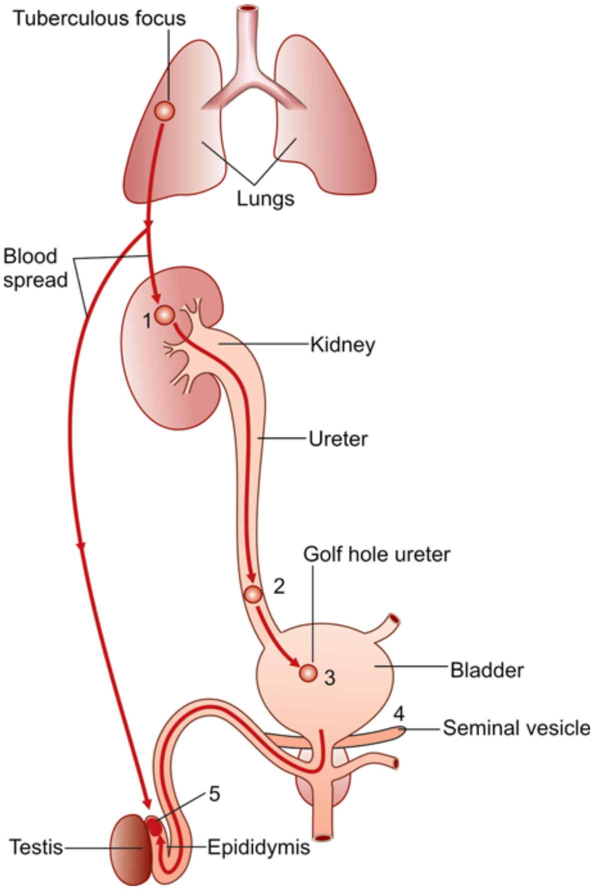

Fig. 26.39: Mode of spread in urinary tuberculosis.

Clinical Features

- Common in males.
- Common on right side.
- Frequency—both day and night.
- Polyuria.

- ***Sterile pyuria***: Urine is pale and opalescent with presence of pus cells without organisms in an acid urine—*abacterial aciduria* (Other causes: Interstitial cystitis, chlamydia).
- ***Painful micturition with often haematuria.*** Haematuria may be overt or microscopic (50%).
- Renal pain and suprapubic pain. Suprapubic pain is more common due to cystitis.
- ***Tuberculous kidney is rarely palpable*** unless there is hydronephrosis or perinephric abscess.
- Enlarged prostate and seminal vesicle, thickened beaded vas, thickened epididymis, impotence, infertility are other features.
- Presentation like acute pyelonephritis.
- Features of urinary stones; recurrent urinary tract infection; renal failure if both kidneys are diseased; hypertension.
- Haemospermia; pelvic pain.
- Dyspareunia; menstrual dysfunction; vaginal discharge; infertility in females.
- Fever and weight loss.
- Often cough with expectoration and haemoptysis may be present.

Investigations

- ↓Hb%.
- ↑ESR. Mantoux skin test is usually positive.
- Chest X-ray.
- U/S abdomen.
- *Three consecutive early morning samples of urine* **(EMSU)** are collected and sent for microscopy (Ziehl-Neelsen staining), culture (L-J media) or guinea pig inoculation.
- Plain X-ray KUB—shows calcification.
- CT scan of abdomen and pelvis to see hydronephrosis, shrunken kidney, stricture, necrosis.
- IVU—hydrocalyx, narrowing of calyx, stricture ureter which are often multiple with dilatations in between.
- Often RGP is very useful, as better definition of ureter, pelvis, calyces and selective sampling of urine are possible.

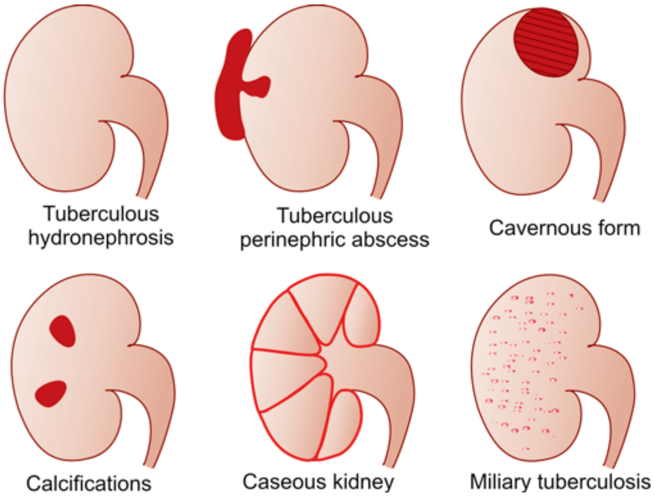

Tuberculous hydronephrosis

Tuberculous perinephric abscess

Cavernous form

Calcifications

Caseous kidney

Miliary tuberculosis

Fig. 26.40: Renal tuberculosis types.

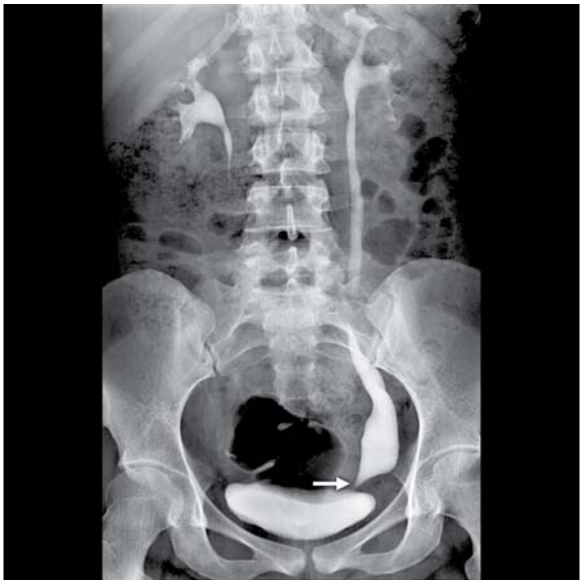

Fig. 26.41: Stricture of left ureter lower part—IVU picture. Common cause is tuberculosis of ureter.

- Cystoscopy reveals multiple tubercles, bladder spasm, oedema of ureteric orifice eventually forming "*golf hole ureter*", scarring, ulceration, bleeding, stone formation.
- Polymerase chain reaction (PCR) for tuberculosis. Radioisometric culture.
- Voiding cystourethrography (MCU) to see ureteric stricture and reflux.

Treatment

- Antitubercular therapy is started. INH, rifampicin, ethambutol and pyrazinamide. Duration of treatment is *one year*.
- After 6-12 weeks of drug therapy, surgical treatment is planned. Kidney is exposed. Pyocalyx is drained. Cut edge of the capsule is sutured—*Hanley's renal cavernostomy.*
- Hydronephrosis—*Anderson Hynes* operation or nephrostomy or stenting ("J" stent) of ureter is done.
- Renal tuberculous abscess not resolving for 2 weeks should be drained.
- Ureteral stricture—stenting/reimplantation of the ureter into the bladder/psoas hitch/Boari's flap/ileal conduit (*Koch's ileal conduit*).
- Thimble bladder—hydraulic dilatation/ileocystoplasty/caecocystoplasty/sigmoid colocystoplasty is done.
- In unilateral lesion, with gross impairment of renal function—*nephroureterectomy* is done.

Indications for nephroureterectomy

- ❖ Nonfunctioning kidney
- ❖ Disease extensively involving the kidney
- ❖ Disease causing hypertension and severe obstruction
- ❖ Tuberculous pyonephrosis
- ❖ Coexisting renal cell carcinoma

Difficulties are like mirror on the wall, that show a person what they are in reality.

HYDRONEPHROSIS (HN)

It is an aseptic dilatation of pelvicalyceal system due to *partial* or *intermittent obstruction* to the outflow of urine.

Aetiology

It can be unilateral or bilateral.

Unilateral

A. *Extramural*:
1. Aberrant renal vessels (vein or artery). It is common on *left side.*
2. Compression by growth (carcinoma cervix, carcinoma rectum).
3. Retroperitoneal fibrosis.
4. Retrocaval ureter.

B. *Intramural:*
1. Congenital PUJ obstruction.
2. Ureterocele.
3. Neoplasm of ureter.
4. Narrow ureteric orifice.
5. Stricture ureter following removal of stone, pelvic surgeries or tuberculosis of ureter.

C. *Intraluminal:*
1. Stone in the renal pelvis or ureter.
2. Sloughed papilla in papillary necrosis.

Bilateral

A. *Congenital:*
- Congenital stricture of external urethral meatus, pin-hole meatus.
- Congenital posterior urethral valve.

B. *Acquired:*
- BPH.
- Carcinoma prostate.
- Postoperative bladder neck scarring.
- Inflammatory/traumatic urethral stricture.
- Phimosis.
- Carcinoma cervix.
- Bladder carcinoma.

- *Congenital PUJ is the most common cause of HN***.**
- Often it is bilateral and presentation on one side is earlier than the other side.
- ***Aberrant renal artery*** or ***vein*** in the lower pole of kidney can compress the PUJ causing HN. Renal angiogram confirms the diagnosis.

Treatment of aberrant renal vessels

If it is a vein it can be ligated safely. But if it is an artery, it exclusively supplies the lower pole of the kidney and so cannot be ligated. So kidney is mobilised; upper and lower poles are approximated together so that artery is made to slip away from the site of compression—*Hamilton Stewart operation***.**

- ***In pregnancy*** dilatation of ureters and both pelvis occur due to atony of ureteric musculature by progesterone. It starts as early as in the first few weeks of pregnancy and lasts until few weeks after delivery. Involution occurs 2-12 weeks after delivery.

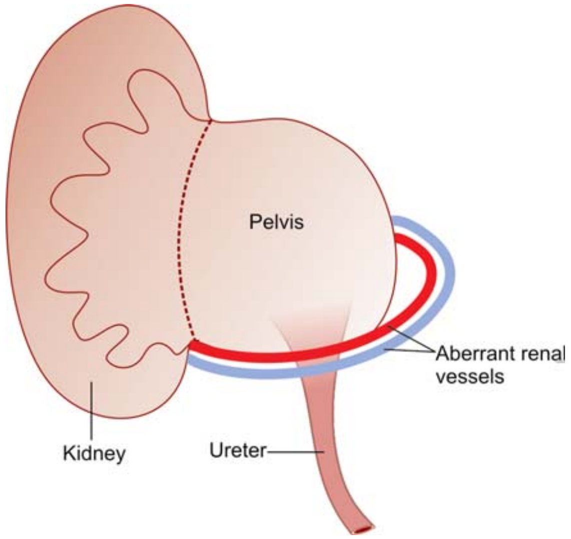

Fig. 26.42: Aberrant renal vessels are one of the known causes of hydronephrosis.

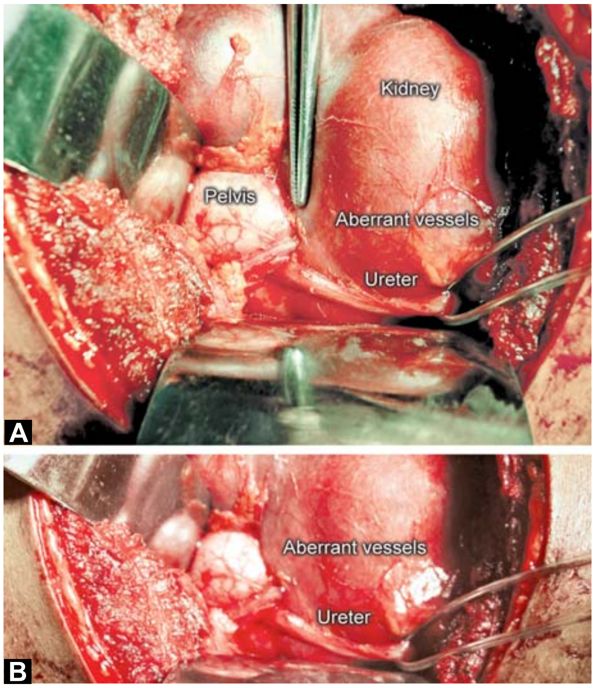

Figs 26.43A and B: Aberrant renal vessels causing obstruction and hydronephrosis. As this often supplies lower pole of kidney exclusively, it need to be retained.

Classifications

Classification I
- Unilateral HN
- Bilateral HN without renal failure
- Bilateral HN with renal failure

Classification II
- *Intermittent HN*: Obstruction occurs, swelling and pain appear in the loin. After sometime patient passes large amount of urine following which swelling and pain disappear—*Dietl's crisis*
- *Persistent HN*: It is due to persistent partial obstruction

Classification III
- HN only
- HN with hydroureter

Classification IV
- Extrarenal pelvic HN (80%)
- Intrarenal pelvic HN (20%). Destruction of kidney is earlier and severe here, as compared to extrarenal pelvic HN

Pathology

- Initially pressure burden is taken up by the pelvis; later calyces and renal parenchyma. Gradually, parenchyma thins out due to destruction and it dilates. Eventually leading to compromised secretory function. Parenchymal thickness of less than 2 mm is unlikely to function. In bilateral cases such patients will go for renal failure.

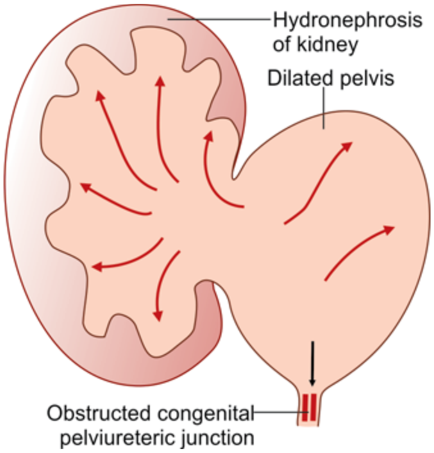
Hydronephrosis of kidney
Dilated pelvis
Obstructed congenital pelviureteric junction

Fig. 26.44: Hydronephrosis due to congenital PUJ obstruction. It is the common cause of hydronephrosis.

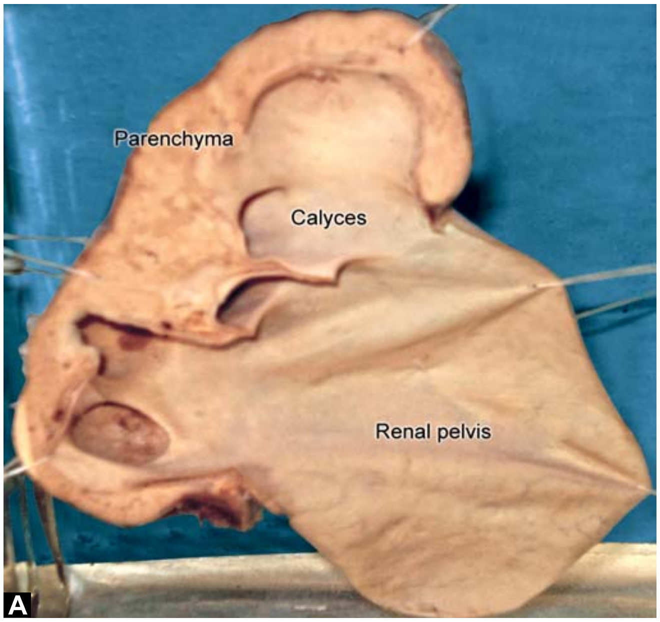
Parenchyma
Calyces
Renal pelvis
A

Fig. 26.45A

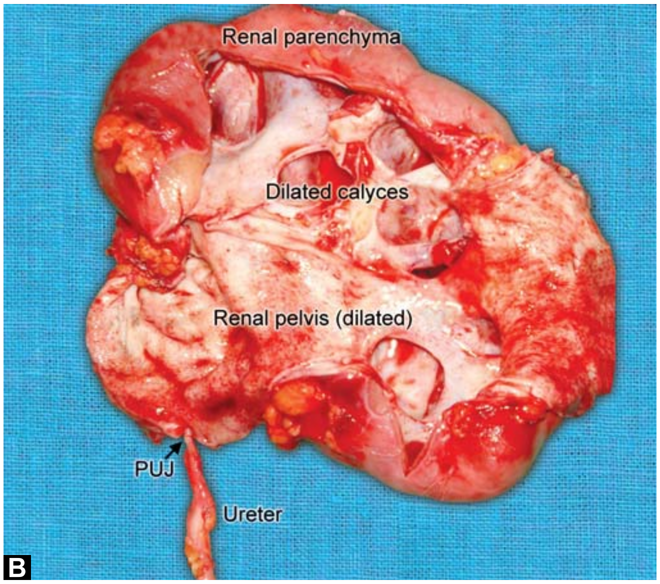
Renal parenchyma
Dilated calyces
Renal pelvis (dilated)
PUJ
Ureter
B

Fig. 26.45B

Figs 26.45A and B: Specimens showing hydronephrosis.

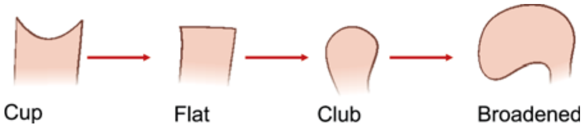
Cup
Flat
Club
Broadened

Fig. 26.46: Stages of HN.

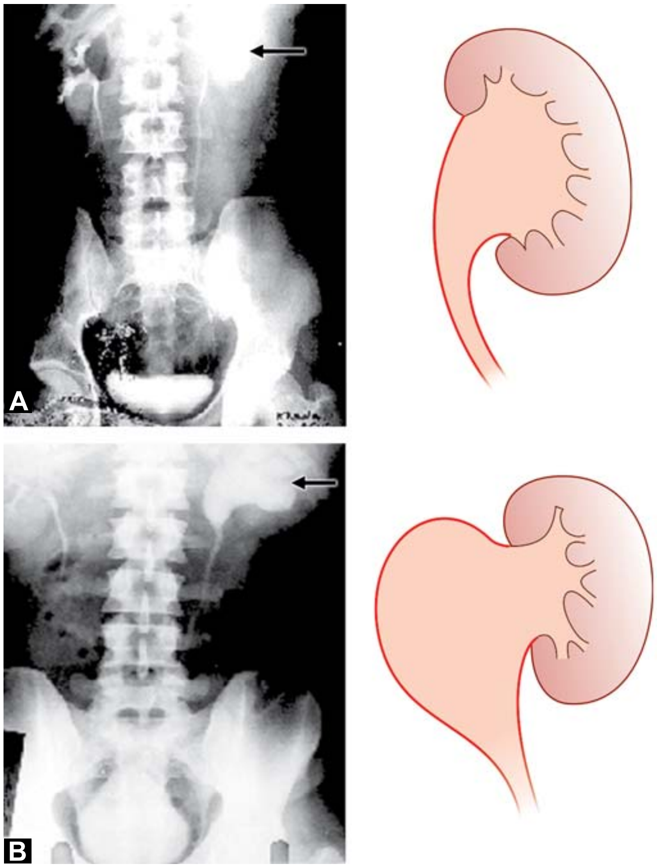

Figs 26.47A and B: Types of renal pelvis (A) Intrarenal and (B) Extrarenal.

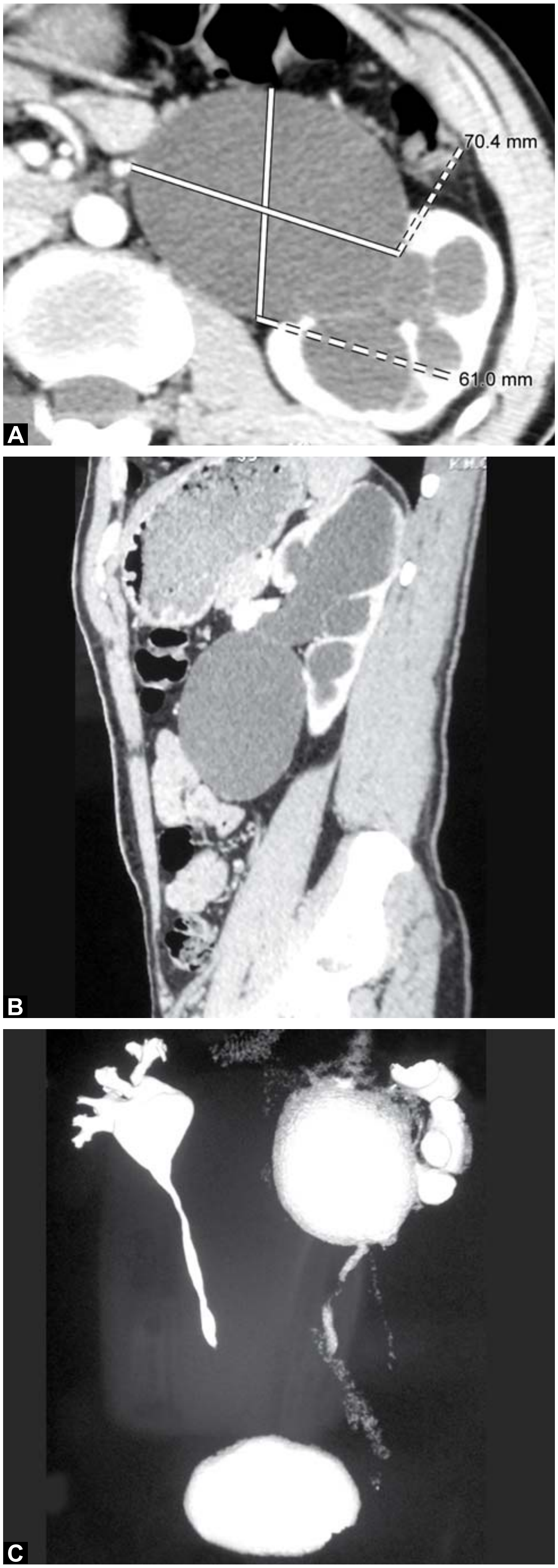

Figs 26.48A to C: CT scan showing hydronephorosis of left kidney with extrarenal pelvis. Patient underwent laparoscopic pyeloplasty (*Courtesy:* Dr Ashok Padit, MCh, Urologist, Mangala Hospital, Mangalore).

Clinical Features

A. *In unilateral cases:*

- Congenital PUJ obstruction and calculus are the most common causes.
- M : F : : 2 : 1.
  - Right side kidney is affected more commonly.
  - Dull aching loin pain with dragging sensation or heaviness.
  - Mass in the loin which is smooth, mobile, ballotable, moves with respiration with dullness in renal angle and a band of colonic resonance in front.
  - Attacks of acute renal colic.
  - Often patient may be having ***Dietl's crisis***—after an acute attack of renal colic, swelling in the loin is seen which disappears after sometime following passage of large volume of urine.
  - Dysuria, haematuria, if infected fever and tenderness in renal angle.
  - Occasionally hypertension.

B. In bilateral cases:

- From lower urinary tract obstruction.

❖ Loin pain
❖ Features of bladder outlet obstruction—frequency, hesitancy, poor stream
❖ Kidneys are often not palpable if renal failure develops early

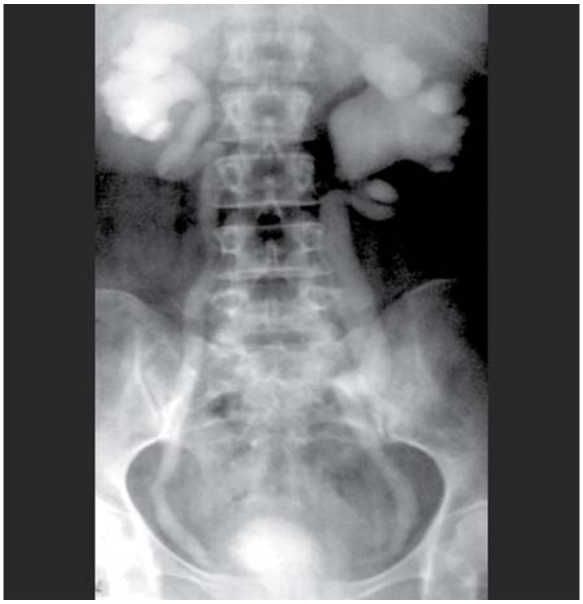

Fig. 26.49: IVU showing bilateral hydronephrosis and hydroureter.

- From bilateral upper urinary tract obstruction.

❖ Loin pain, mass in the loin, attacks of renal colic
❖ In bilateral cases, when it is severe, features of renal failure like oliguria, oedema, hiccough may be present

Complications

1. Pyonephrosis.
2. Perinephric abscess.
3. Renal failure in bilateral cases.

Investigations

- Blood urea and serum creatinine.
- Urine for microscopy.
- U/S abdomen: Investigation of choice.

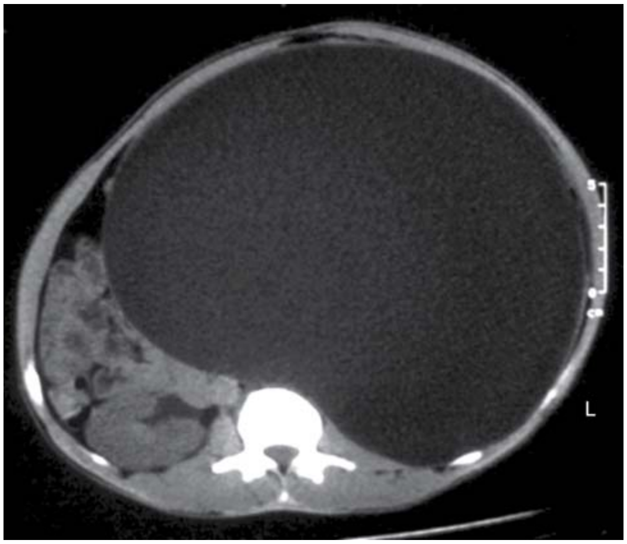

Fig. 26.50: Hydronephrosis as seen in CT scan.

Type of pelvis, thickness of parenchyma, site of obstruction and cause of obstruction, e.g. stones, can be made out (Refer Fig. 26.54).

- *IVU*: To find out the function of diseased as well as opposite kidney. Normal calyx is *cup* shaped. It gets *flattened* and later *club shaped* which eventually becomes *broadened* in hydronephrosis (Refer Fig. 26.47).
- ***Whitaker test***: A fine needle is passed into the renal pelvis through loin. Pelvis is perfused with saline at a rate of 10 ml/minute. Normally, initially the pressure increases and later it will remain constant. Persistent increase in pressure suggests HN.
- CT scan is diagnostic.
- ***Isotope renography*** is also useful to study the function of the kidney before and after the surgical treatment and also to see the efficacy of surgery as far as function is considered—DTPA scan.

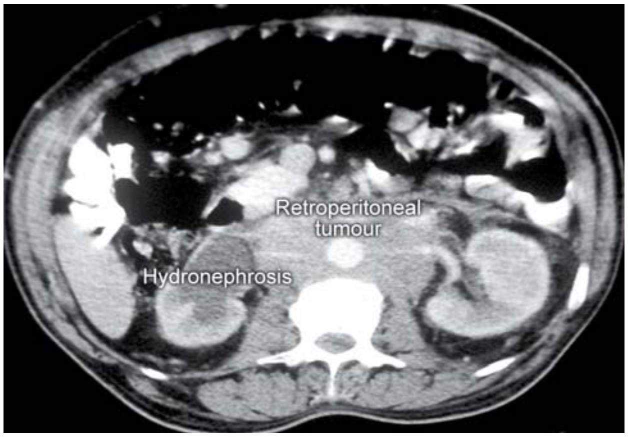

Fig. 26.51: Retroperitoneal tumour causing ureteral obstruction with hydronephrosis.

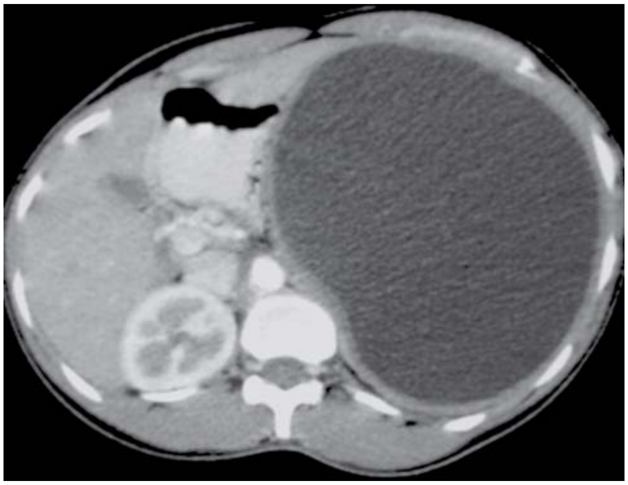

Fig. 26.52: CT scan showing bilateral hydronephrosis. One side large (left side); another side early hydronephrosis.

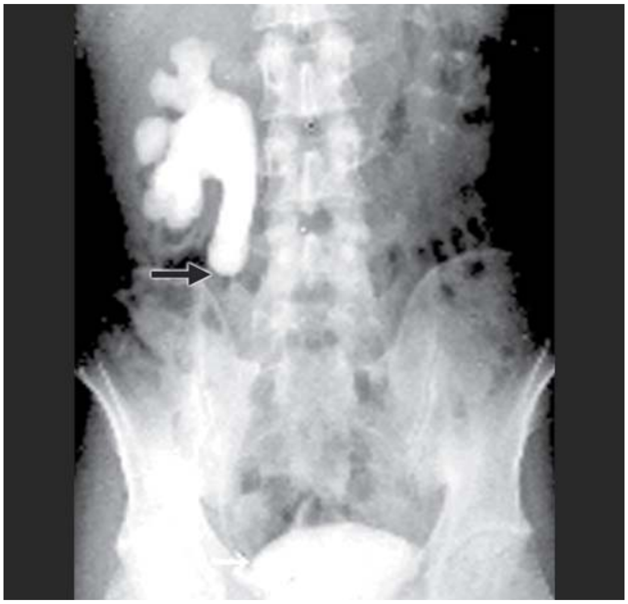

Fig. 26.53: IVP showing right hydronephrosis. Note the dilated right renal pelvis and delay in the excretion of the dye from the affected kidney.

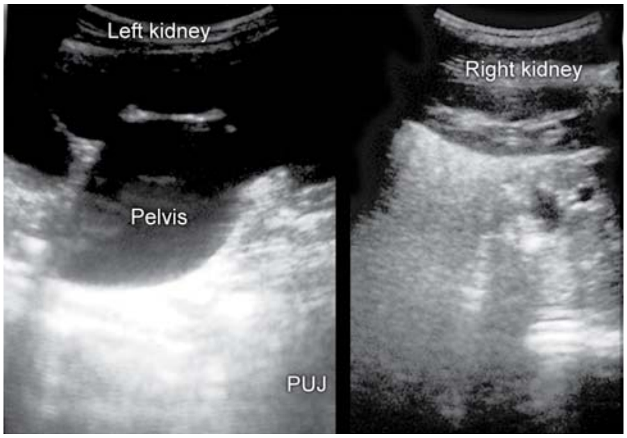

Fig. 26.54: U/S picture of hydronephrosis showing dilated pelvis.

> *In the case of perinephric abscess bending the trunk away from the side of the abscess may produce pain, whereas the patient can bend his body towards the lesion without much discomfort.* **—George S**

Treatment

- Always conservative surgeries which are aimed at conserving the kidneys are done. Nephrectomy is not done unless indicated.

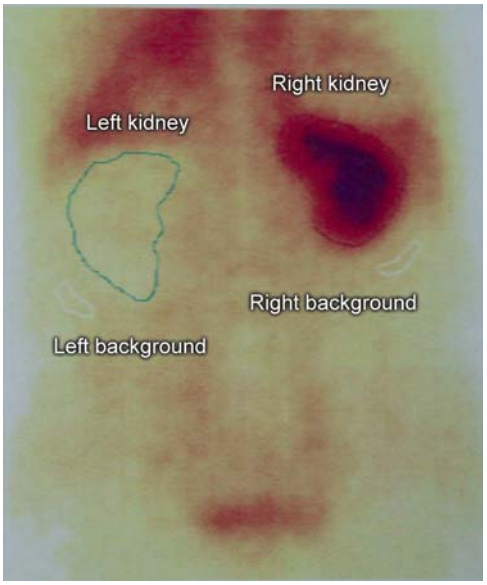

Fig. 26.55: Isotope scan showing hydronephrosis on the left side.

1. The cause is treated: Stone, congenital anomaly, aberrant renal vessels, stricture urethra (dilatation, urethrotomy, urethroplasty); phimosis (circumcision); BPH (TURP); posterior urethral valve (cystoscopic fulguration of valve).
2. ***Anderson-Hyne's operation*** (*Dismembered pyeloplasty*): In congenital PUJ obstruction, the spasmodic segment and redundant pelvis are excised. A new pelvis is created and the cut end of pelvis is anastomosed to the ureter in the *dependent* position.

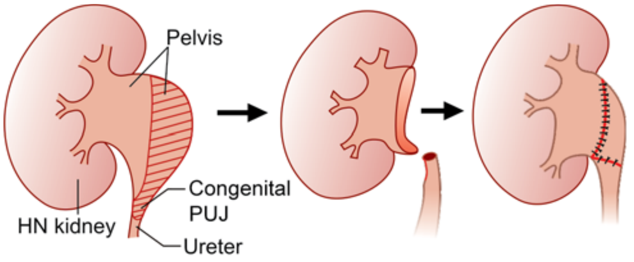

Fig. 26.56: Anderson-Hyne's operation.

3. *Davis T-tube ureterostomy:* Placement of T tube in the ureter by making longitudinal incision.
4. *Non-dismembered pyeloplasties*: Here PUJ is not transected. Reconstruction is done without PUJ transection by different methods, e.g. **Foley's Y-V plasty.**
5. In bilateral HN, without renal failure, *kidney which is functioning better should be operated first.* Three months later, otherside kidney is dealt with.
6. In bilateral HN with renal failure, *bilateral nephrostomy and haemodialysis* support is required initially. After 3-6 weeks IVU is done and the functions of both kidneys are looked for. If they function, then treated accordingly by Anderson-Hyne's operation.

   If kidneys are still not functioning (renal function < 20% on isotope study) then nephrectomy and renal transplantation is the only option left.
7. *Laparoscopic* or *retroperitoneoscopic pyeloplasty* is becoming popular but is expensive and time consuming. It gives good result as well.
8. *Endoscopic pyelolysis*, even though technically easier results are not assured.

PYONEPHROSIS

It is collection of pus in pelvicalyceal system, which is converted into a multiloculated sac. Occurs due to:

a. Infection of pre-existing hydronephrosis.
b. Following acute pyelonephritis.
c. As a complication of renal calculus—either stone in the renal pelvis or staghorn calculus.

Clinical Features

- Usually unilateral.

Triad

- Anaemia
- Fever
- Loin swelling

- Tender mass in the loin which is smooth, soft, not mobile, not moving with respiration.
- Patient may also have cystitis, pyuria, burning micturition.
- Features of toxicity such as fever with chills and rigors.

Investigations

- Plain X-ray, KUB, may show renal calculus.
- IVU shows HN poor secretion.
- *Cystoscopy* reveals cystitis with *efflux of purulent pus* through the ureteric orifice**.**
- U/S shows dilatation.
- DTPA scan later.

Treatment

- After starting antibiotics, pus is immediately drained from the kidney through a loin incision and *nephrostomy* tube (Cabot's nephrostomy) (Malecot's catheter) is placed.
- If kidney is totally destroyed, *subcapsular nephrectomy* is done. This also prevents other kidney from getting infected through perirenal lymphatic connections.
- In bilateral pyonephrosis, *bilateral nephrostomy* is the only choice. *'J' stenting* is done often to keep the ureters patent.

CARBUNCLE OF KIDNEY (RENAL CARBUNCLE)

- A localised inflammatory necrotic mass of tissue involving renal parenchyma, caused by *Staphylococcus aureus* and coliform organisms, source of which is cutaneous infections like boil and carbuncle.

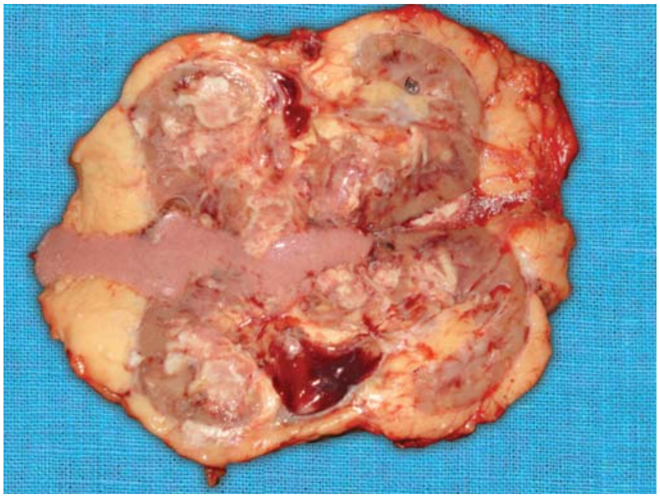

Fig. 26.57: Cut section of specimen of kidney showing pus, necrotic area, and multiple infected abscesses—renal carbuncle with multiple abscesses. Condition is common in diabetics.

- It presents as ill-defined tender swelling in the loin, with pyrexia and leucocytosis.
- Staphylococci can be isolated from the urine.
- IVU shows obliteration of group of calyces, ***mimics renal cell carcinoma***.
- **Treatment**: Antibiotics, drainage of carbuncle, with placement of Malecot's catheter.
- Often, in severe type nephrectomy may be needed.
- Severe septicaemia may occur in renal carbuncle.

Renal abscess/renal carbuncle

- A localised inflammatory necrotic mass of tissue involving renal parenchyma, caused by *Staphylococcus aureus* and coliform organisms, source of which is cutaneous infections like boil and carbuncle. Eventually infection may spread to entire kidney causing multiple abscesses
- It presents as ill defined tender swelling in the loin, with pyrexia and leucocytosis
- Staphylococci can be isolated from the urine
- IVU shows obliteration of group of calyces, *mimics renal cell carcinoma*
- Treatment: Antibiotics, drainage of carbuncle with Malecot catheter placement
- Often, nephrectomy may be needed in severe type
- Life-threatening septicaemia can occur which often needs not only nephrectomy but also higher antibiotics like mereponem, linezolid, and intensive critical care, ventilator support

PERINEPHRIC ABSCESS

Causes

- Infection of a perinephric haematoma.
- Perforation through renal capsule from pyonephrosis or renal carbuncle.
- Tuberculous perinephric abscess.
- Extension of cortical abscess.
- Haematogenous spread.
- Extension of appendicular abscess.
- Periureteral lymphatic spread.

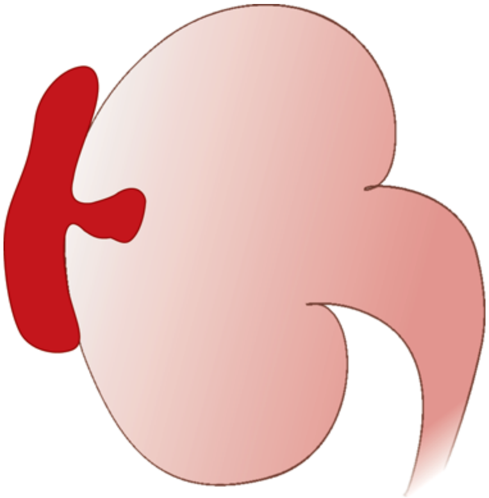

Fig. 26.58: Perinephric abscess.

Clinical features

- High fever
- Fullness in the loin
- Tenderness and rigidity
- Scoliosis with concavity towards the side of abscess

Investigations

- Total count is increased.
- Plain X-ray KUB—obliteration of psoas shadow, scoliosis, elevation of hemidiaphragm.
- IVU—two films taken one in lying down position and another in erect posture.

  Normally in erect posture, downward displacement of the kidney is seen. Downward displacement is not seen in case of perinephric abscess—*Mathe's sign.*
- U/S abdomen.
- CT scan is diagnostic.

Treatment

1. Antibiotics are started.
2. Under G/A through loin approach (lumbar incision), pus is drained adequately. Pus is sent for culture and sensitivity. Collection in the cortex of the kidney should also be drained. A Malecot's catheter is kept in place.

RENAL CALCULUS

> *The Times whereat you often in a day have the Urinary Excretion performed with Ease, are times which invite you very frequently to lift up your Hearts unto God with such an Acknowledgment as This; 'O, My most merciful God, I bless thee, that the grinding Torments of the Stone, are not now grinding of me'*
>
> —**Cotton Mather,** 1724

It is more common in males; 90% are radio-opaque (gallstones are more common in females; 90% are radioluscent).

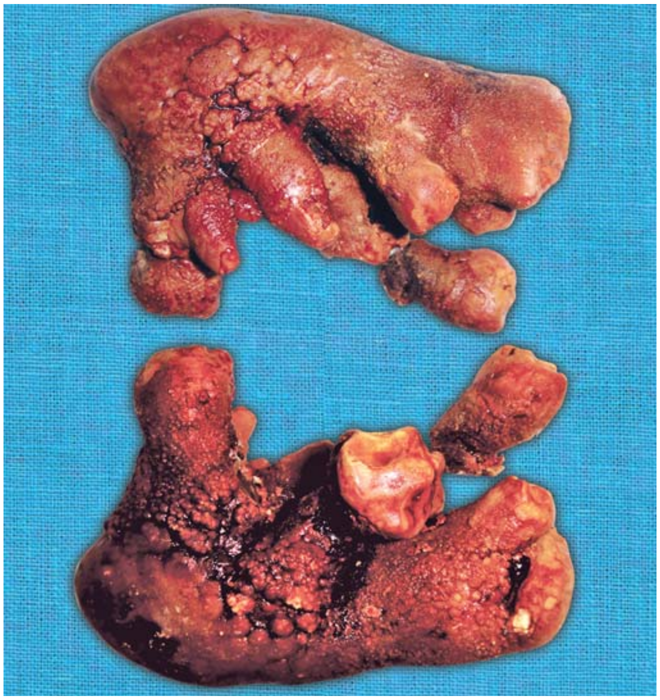

Fig. 26.59: Large staghorn calculus—both surfaces of the stone were shown.

Aetiology

- *Diet:* Vitamin A deficiency—it causes desquamation of epithelium which acts as a nidus for stone formation.
- *Climate*: In hot climate urinary solutes will increase with decrease in colloids, which leads to chelation of solute with calcium forming a nidus for stone.
- *Citrate level* in urine (300-900 mg/24 hours) maintains the calcium phosphate and carbonate in soluble state and any decrease in citrate level in urine causes stone formation.
- *Infection* in kidney: Urea splitting organisms commonly cause stone formation, i.e. *E. coli*, *Staphylococcus, Streptococcus, Proteus.*
- *Prolonged immobilisation* causes decalcification of bones and so hypercalciuria leading to stone formation.
- *Hyperparathyroidism* causes hypercalciuria causing multiple bilateral stones or often bilateral nephrocalcinosis (5%).
- *Hyperoxaluria,* as a result of altered glycine metabolism.
- *Cystinuria* (Autosomal recessive).
- *Stasis* due to obstruction to urine flow.
- *Medullary sponge kidney*.
- *Randall's plaque theory* is erosion and deposition of urinary salts as Randall's plaque at the apex of renal papillae.
- *Carr's postulates* states that minute concretions called as microliths normally develop in the subendothelial part of the tubule which will be carried away as particles by renal lymphatic network vessels. If these lymphatics are blocked, microliths enlarge and act as nidus for stone formation.
- *Others:* Sarcoidosis, myelomatosis, gout, idiopathic hypercalciuria, hypervitaminosis D, neoplasms on treatment, hypomagnesuria (Mg^{++} in urine acts as a complexing agent and prevents nucleation normally).
- *Renal tubular acidosis*: Commonly causes calcium phosphate stone (10%).

Stages of stone formation

I. Supersaturation
II. Nucleus formation
III. Crystallisation
IV. Aggregation
V. Matrix formation
VI. Stone

Epitaxy: Growth of one type of stone on another type.

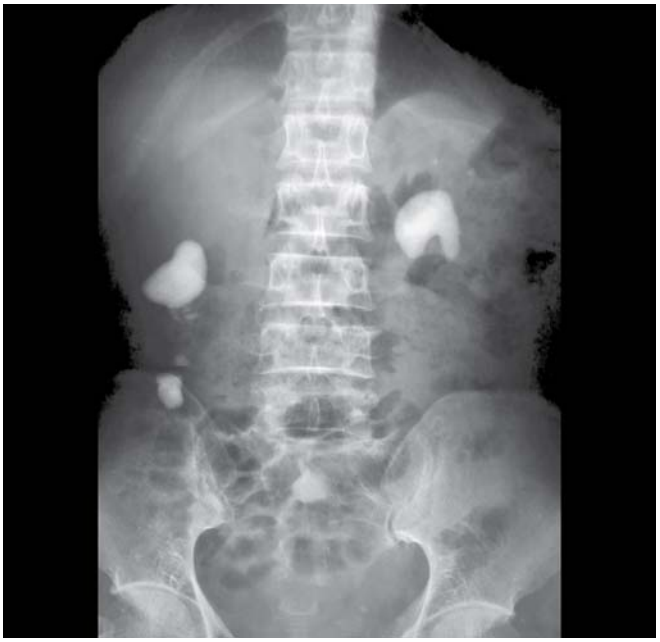

Fig. 26.60: Plain X-ray KUB showing bilateral large renal stones and right ureteral stone.

Types

1. *Oxalate stones (75%):* Also called as *mulberry stone* as it is brown in colour, with sharp projections. It is invariably calcium oxalate stone, shows *envelope crystals in urine.*
2. *Phosphate stones (10-15%):* It is either calcium phosphate or calcium, magnesium, ammonium phosphate stone usually occurring in an infected urine. It is smooth and white in colour. In an alkaline urine it enlarges rapidly, filling renal calyces taking their shape called as *staghorn calculus*. It is radio-opaque and attains a large size.
3. *Uric acid stones (5%)* are smooth, hard, yellowish, multiple and radioluscent. They are seen in gout, hyperuricosuria, altered purine metabolism.
4. *Urate stones*.
5. *Cystine stones (2%)* occur in *cystinuria* where there is defective absorption of cystine from the renal tubules (autosomal recessive condition).

It is seen in young girls, occurs only in *acidic urine.* It is multiple, soft, yellow in colour and the colour changes to *greenish hue* on exposure. It attains large size. It is radio-opaque because it contains sulphur.

6. *Xanthine stones* are very rare, smooth, brick red in colour, due to altered xanthine metabolism.

   Here there is deficiency in xanthine oxidase enzyme.
7. *Indigo stones*: Very rare. Blue in colour.
8. *Struvite stone*: It is compound of magnesium, ammoniumphosphate mixed with carbonate. It occurs in presence of ammonia and urea splitting organisms in urine, e.g. *Proteus; Klebsiella.*

Shapes of Stone Crystals in Urine

	Type of crystal	*Shape of the crystal*
a.	Calcium oxalate monohydrate	Dumbell shaped
b.	Calcium oxalate dihydrate	Envelope shaped
c.	Uric acid	Yellowish of varying size and shape
d.	Cystine	Hexagonal, very soft stones
e.	Triple stone	Coffin lid shaped

Clinical Features

- *Pain—renal pain* is located over renal angle, hypochondrium and lumbar region. Often severe radiating to groin and testis in male, with vomiting due to pylorospasm. Pain worsens on movements.
- *Haematuria* is common.
- Pyuria.
- Fever.
- Tenderness in renal angle, with often a mass in the loin due to hydronephrosis which moves with respiration and is bimanually palpable, ballotable, smooth, soft.
- As urinary tract infection.
- Incidental finding.
- Often hypertension.

Note:
Food rich in oxalates are, spinach, tea, cola, alcohol, citrus fruits.

Investigations

- *Blood*: ESR, serum calcium, phosphate, creatinine, blood urea, uric acid, PTH level.
- *Urine*: Calcium, urate, cystine if suspected only, pH, specific gravity.
- *Plain X-ray, KUB*: To see kidney shadow, stones (90%—*radio-opaque*).
- IVU to see renal functions and HN.
- RGP if required.
- U/S abdomen—can detect even radioluscent stones and gives information about the changes in renal parenchyma.
- Urine analysis and C/S to identify bacteria.
- *CT scan* will identify the small missed stones in ureter.

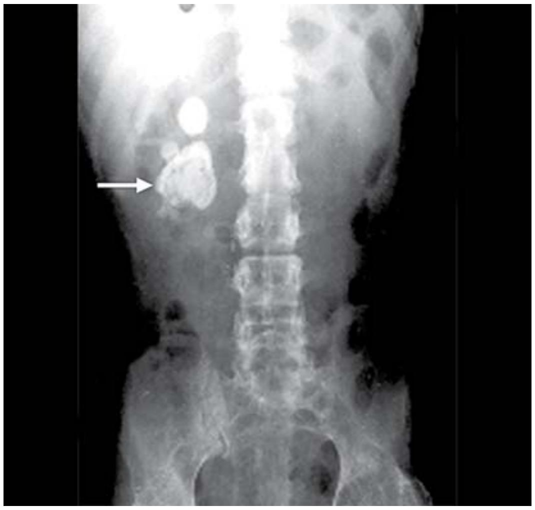

Fig. 26.61: Plain X-ray, KUB, showing multiple stones in the right kidney.

Treatment

1. ***PCNL (Percutaneous Nephrolithotomy).***

Indications

- Stones more than 2.5 cm in size
- Multiple stones
- Stones not responding for ESWL

Procedure: Initially cystoscopy is done and ureteric stent/catheter is placed and renal pelvicalyceal system is identified under C-Arm guidance. Under the guidance of C-Arm or U/S, needle puncture is made in the loin percutaneously. Through kidney, calyx and pelvis are approached. Guidewire is passed. Graduated dilators are passed and so track is widened. Then through that, a nephroscope is passed. After fragmentation, stone is removed using different methods [Laser (Holmium), pneumatic, ultrasonic or electrohydraulic].

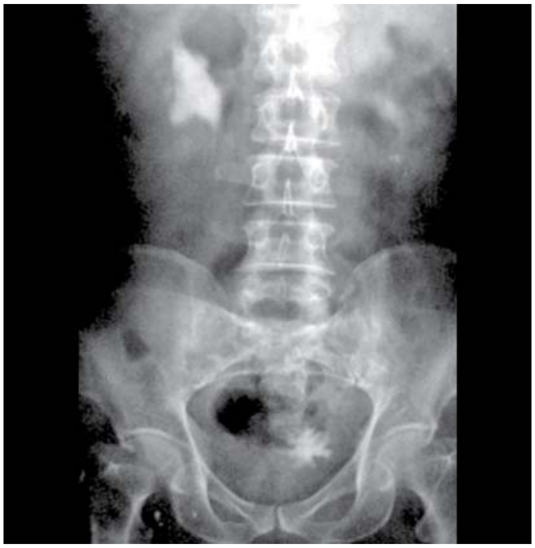

Fig. 26.62: Plain X-ray KUB showing phosphate stone in the kidney and oxalate stone in urinary bladder.

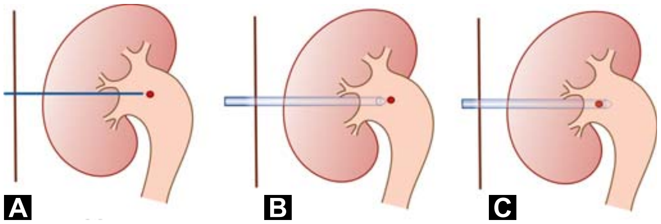

Figs 26.63A to C: Stages in PCNL.

Complications of PCNL

- Haemorrhage
- Perforation of collecting duct causing extravasation of irrigation fluid
- *Injury to colon or pleura* while creating initial track for nephroscope

2. ***ESWL (Extracorporeal Shock Wave Lithotripsy):***

A
MEDILITH
MEDICO
B

Figs 26.64A and B: ESWL table and machine used for renal stones. It needs C-Arm to guide the exact point of stone to be fragmented.

- *Piezo-Ceramic* or *Electromagnetic* shock waves are passed to the stone through water bath or water cushion which acts as a media. Shocks are produced at 2/sec. 1000-4000 shocks are required for each stone.
- *Dornier Lithotripter* is used for fragmenting stones.
- Stone is located and observed through fluoroscope (C-Arm) or ultrasound. Shock waves are triggered to create *compressive waves* over the stone, to *fragment* it. These fragments are flushed out later.

Advantages

- No anaesthesia is required
- Can be done as an OP procedure
- Less than 2.5 cm sized stones are well fragmented
- Hard stones, oxalate stones are better eliminated by ESWL
- ESWL can be done repeatedly in different sittings
- If it is not successful one can switch over to PCNL

Complications

- Renal haematoma
- Severe haematuria
- Injury to adjacent structures
- Fragmented stone retains in the ureter

Contraindications

- Pregnancy
- Bleeding disorders
- Patients with abdominal aneurysms
- Sepsis and renal failure (Serum creatinine more than 3 mg%)

3. ***Conservative treatment:***
    - Flush therapy—mainly *used for lower ureteric stones.*
    - IV fluids.
    - Inj frusemide 60-80 mg IV.
    - Anti-inflammatory and antispasmodic agents are given to relieve the pain.
4. ***Surgery for renal stones*:** Presently most of the renal stones can be removed without open surgery (PCNL, ESWL, URS). But limiting factors are cost and availability.

Surgeries

1. *Pyelolithotomy.*
    - Suitable for stones in extrarenal pelvis.
    - By loin (posterior subcostal) incision, kidney is approached.
    - Renal pelvis is opened, the stone is removed and the pelvis is closed. A drain is placed and wound is closed.
2. *Extended pyelolithotomy (Gil-Vernet*): In case of intrarenal pelvis, incision is done on the hilum between the pelvis and kidney over the renal sinus, dissection is carried out so as to remove the stones from pelvis as well as calyces.
3. *Nephrolithotomy*: By placing incision just behind the most convex surface (*Brodel's line*), stone is removed.
4. *Nephropyelolithotomy*: By making incisions both over the kidney and pelvis, stone/stones are removed. It is often done in *staghorn calculus.*

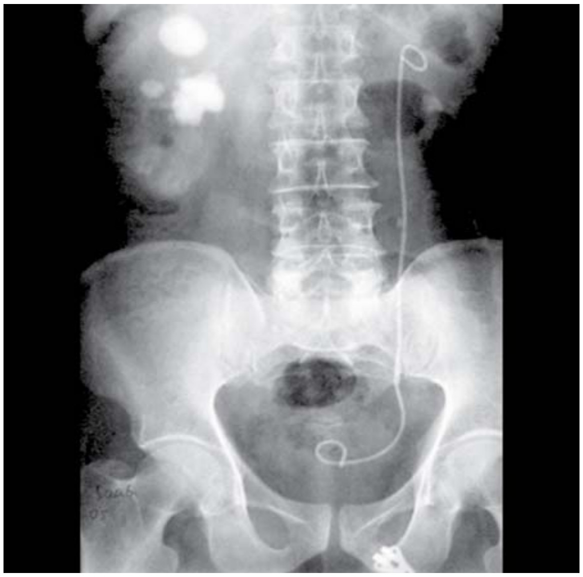

Fig. 26.65: Plain X-ray showing right-sided renal stone with 'J' stent on left side.

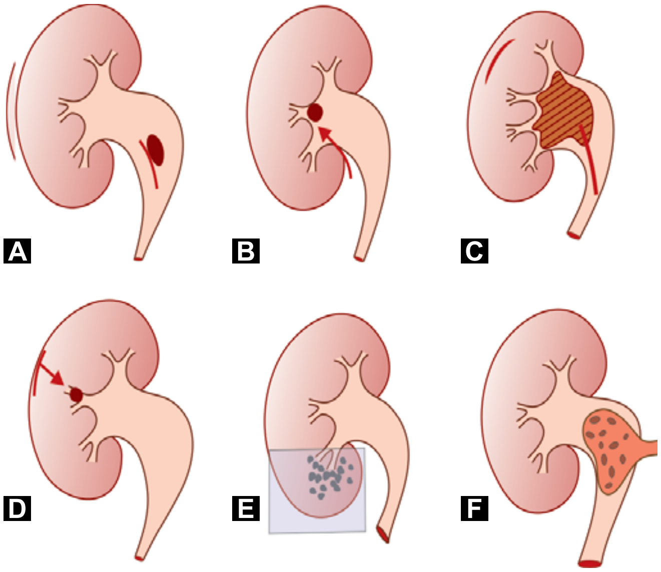

Figs 26.66A to F: Surgeries for renal stones. (A) Pyelolithotomy, (B) Extended pyelolithotomy, (C) Nephropyelolithotomy, (D) Nephrolithotomy, (E) Partial nephrectomy, (F) Coagulum pyelolithotomy.

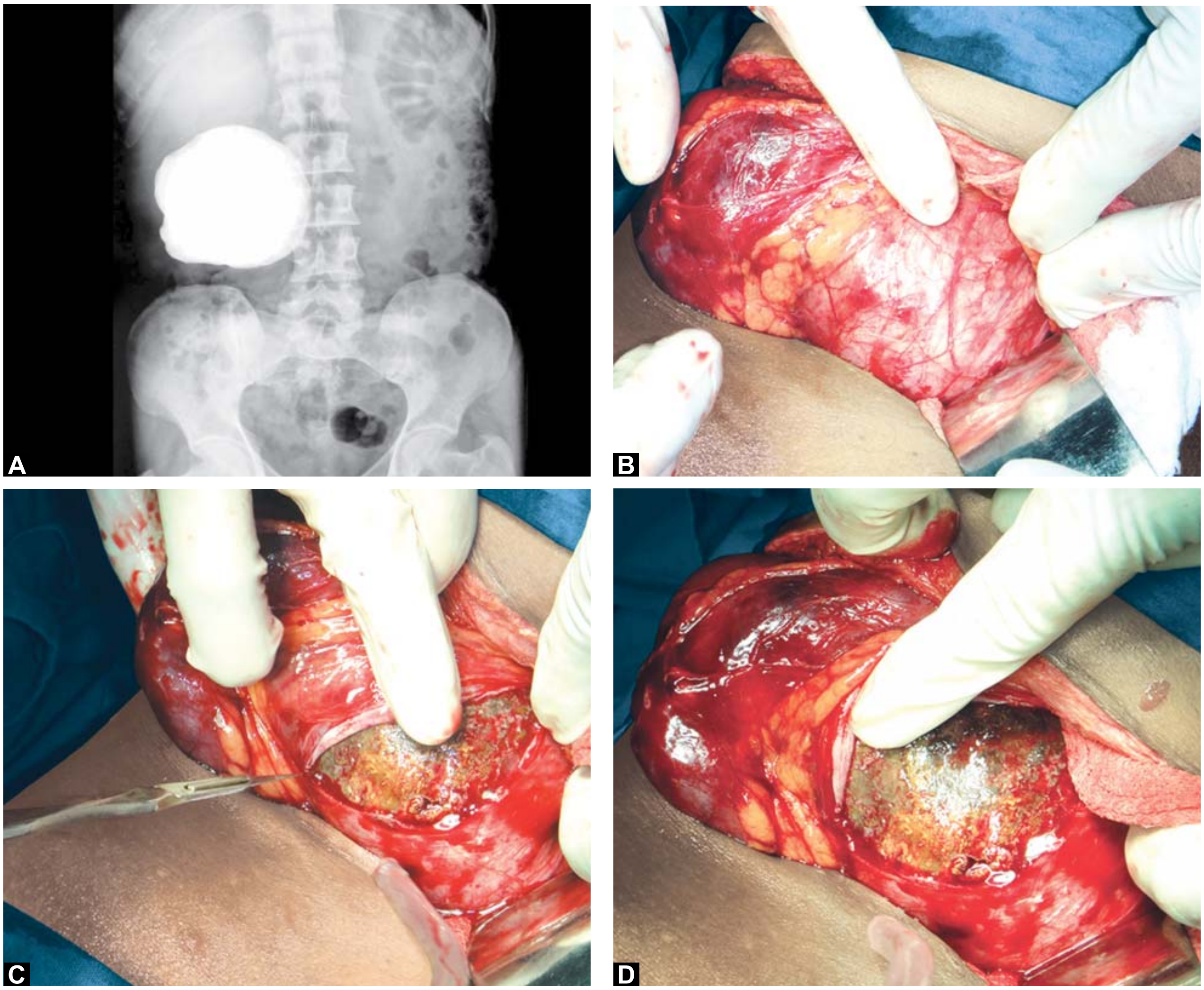

Figs 26.67A to D

You may be disappointed if you fail, but you are doomed if you don't try.

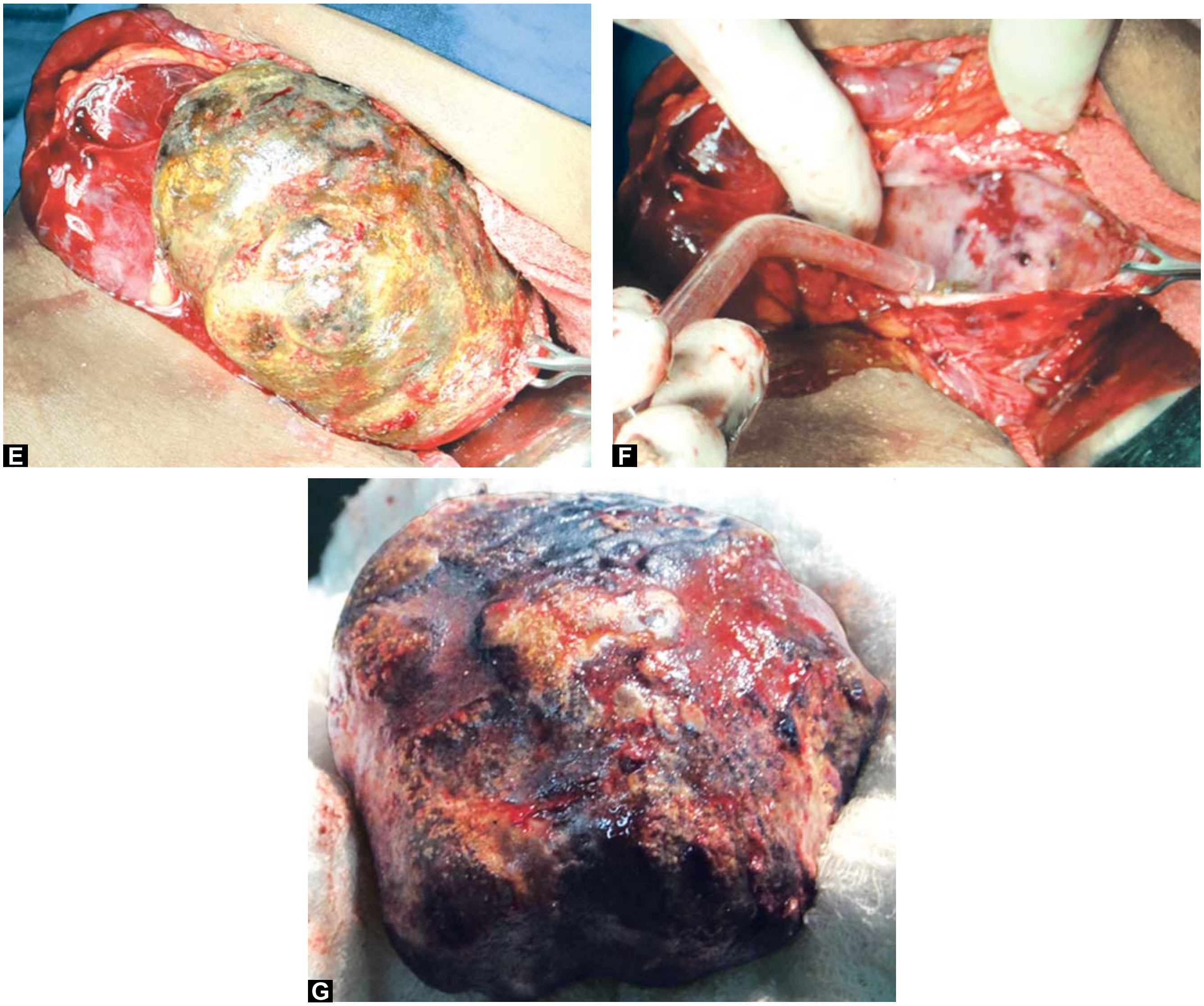

Figs 26.67E to G

Figs 26.67A to G: Very large renal pelvic stone of more than 8 cm in size weighing 610 grams. It is one of the largest renal stone extracted (*Courtesy:* Dr Ashok Padit, MCh, Urologist, Mangala Hospital, Mangalore).

5. *Partial nephrectomy:* Done when there are multiple stones occupying a pole, usually lower pole of the kidney or when there is damage to the calyx, if not removed may encourage further stone formation.
6. *Bench surgery:* Kidney is removed out temporarily, cooled by ice packs or inosine or liquid nitrogen. Stones are searched and removed completely. Later kidney is replaced in right iliac fossa.
7. *Coagulum pyelolithotomy*: Coagulum solution which contains fibrinogen is poured into the renal pelvis. It is activated so that it solidifies, meanwhile entangling the stones in renal pelvis. This entangled mass is removed en masse.
8. *Anatrophic pyelolithotomy*: After exposing the kidney, it is cooled with ice packs for 20 minutes and posterior branch of the renal artery is clamped temporarily using bull-dog clamp. The most avascular plane behind the Brodel's line is thus visualised properly. Kidney is opened through this line and stone/stones are removed (anatrophic means "to prevent atrophy").

URETERIC CALCULI

- Always of renal origin.
- Nature of stones are same as that of renal stones.
- They are commonly of elongated shape.
- They can get impacted at various narrow junctions.

Sites

- PUJ.
- Where ureter crosses the iliac vessels.
- Where ureter crosses vas deferens/broad ligament.
- Where ureter penetrates outer layer of bladder muscle.
- In the intramural portion of ureter near the ureteric orifice. Stones less than 5-8 mm size may pass spontaneously.

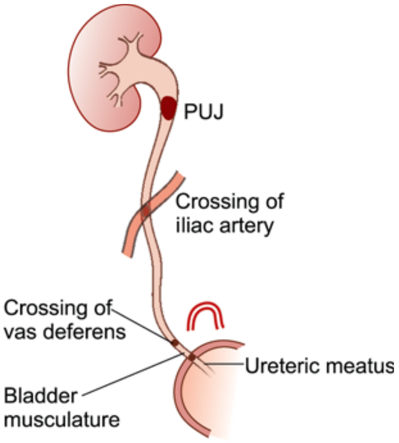

Fig. 26.68: Sites of ureteric narrowing.

Problems with ureteric stones

- Obstruction
- Hydronephrosis
- Infection
- Impaction
- Ureteral stricture

Clinical Features

1. *Pain*—it is of colicky type and radiates from loin to groin often to the tip of the genitalia, testis in males, labia majora in female (referred along the genitofemoral nerve).

   It is severe in intensity, increases with exercise.

   It mimics *appendicitis, cholecystitis, ovarian* or *tubal pathology*.
2. Nausea, vomiting, sweating due to pain and reflex pylorospasm.
3. Haematuria, dysuria, frequency, strangury.
4. Tenderness in iliac fossa and renal angle (no rebound tenderness).

Investigations

- Urine—microscopy, C/S.
- Plain X-ray, KUB—radio-opaque stones are visible in 90% of cases—in the line of ureter (near the tips of transverse processes of lumbar vertebrae, sacroiliac joint and medial to ischial spine).

  Lateral or oblique films are required to differentiate from other opacities which mimic stone.
- IVU shows hydronephrosis and hydroureter. Function may be accurately assessed by isotope renogram.
- Blood urea, serum creatinine, serum calcium, uric acid level.
- U/S is useful.
- CT scan is *diagnostic.*

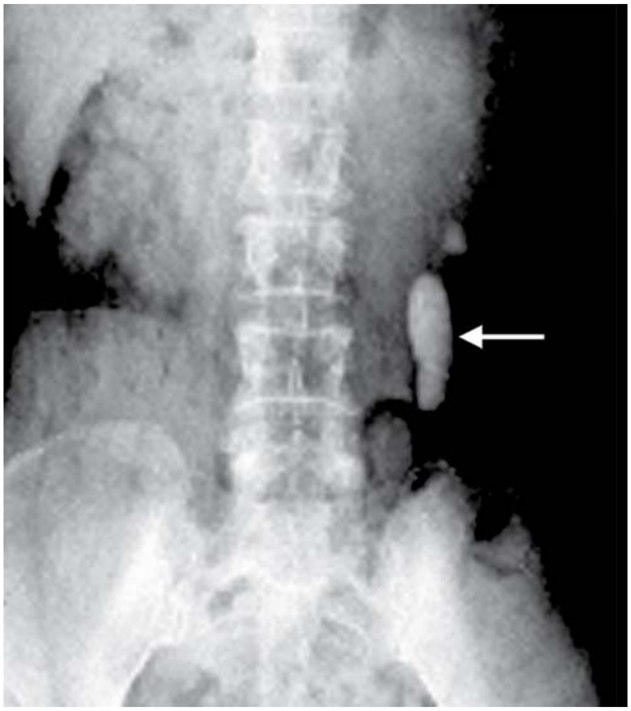

Fig. 26.69: A plain X-ray, KUB, showing left ureteric calculi.

Differential diagnosis

- Appendicitis
- Cholecystitis
- Ovarian cyst
- Tubo-ovarian disease
- Mesenteric adenitis
- Ruptured ectopic gestation

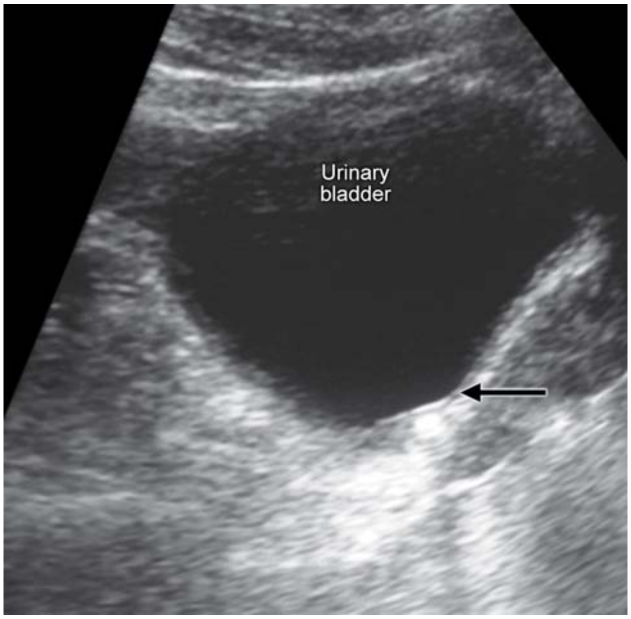

Fig. 26.70: USG showing vesicoureteric junction (VUJ) stone which needs cystoscopic meatotomy and extraction.

Treatment

1. Plenty of water orally.
2. Diuretic—oral frusemide to flush the stone.
3. Suitable antibiotics to control sepsis; antispasmodics to relieve pain.
4. IV fluids—fast infusion of about 1.5 to 2 litres and injection frusemide 60 to 80 mg. Usually given for 3 to 5 days.
5. Surgical intervention for ureteric stones:

Indications

- Size of the stone more than 5 to 8 mm
- IVU showing deterioration of function
- Co-existing infection
- If stone is impacted in the ureter with persistent symptoms

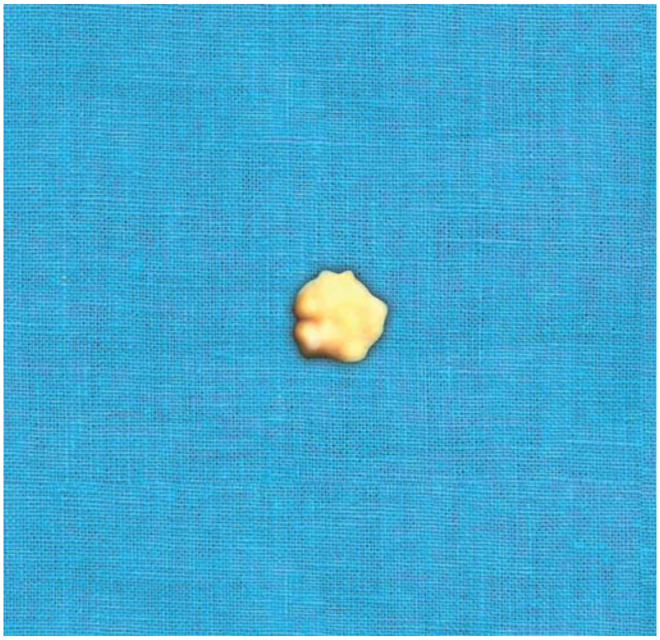

Fig. 26.71: Ureteric stone.

Procedures

Upper third stone:

- ESWL for stone in upper third ureter.
- The stone is pushed into the renal pelvis and then PCNL is done.
- URS—*Ureterorenoscopic stone removal*:
  - Through ureteroscope, stone is visualised and often fragmented using pneumatic bombarder. It is then extracted by ureteroscope.
  - Complications are perforation of ureter and extraperitoneal leakage of urine, bleeding.
- Open *ureterolithotomy* through loin incision.

Stone in middle third ureter:

- URS.
- Open ureterolithotomy.

Stone in lower third ureter:

- URS.
- *Dormia basketing*:

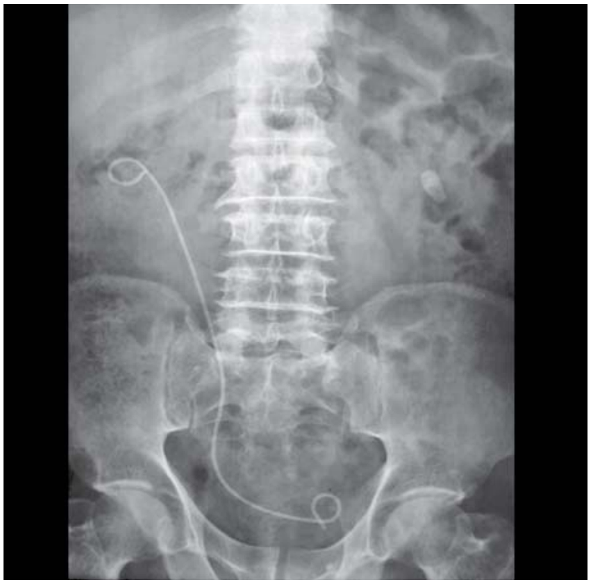

Fig. 26.72: 'J' stent right side—used after URS (ureterorenoscopy) in ureteric stone extraction.

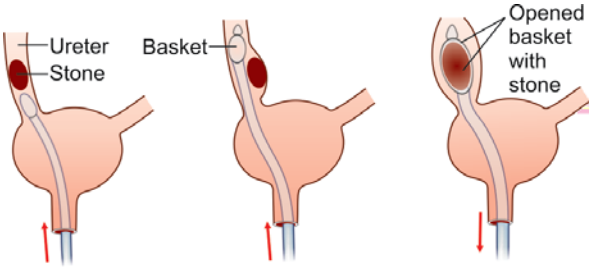

Fig. 26.73: Dormia basket is used to remove the stone in the lower ureter. Single stone less than 1 cm can be removed.

Indications for Dormia basketing

- Stone in lower third ureter
- Stone below pelvic brim
- Stone less than 10 mm size
- Single stone

Basket is passed into the proximal ureter beyond the stone and opened. The stone is then pulled out.

Complications of Dormia basket

- Stone dislodgement
- Urethral injury
- Avulsion of ureter
- Stricture ureter

- Open ureterolithotomy.
- Using cystoscope under general anaesthesia, *ureteric meatotomy* is done for stones impacted at the ureteric orifice. It is released by cutting the orifice at upper and lateral aspects.

Note:

All ureteric stones can be removed through laparoscopy or retroperitoneoscopy.

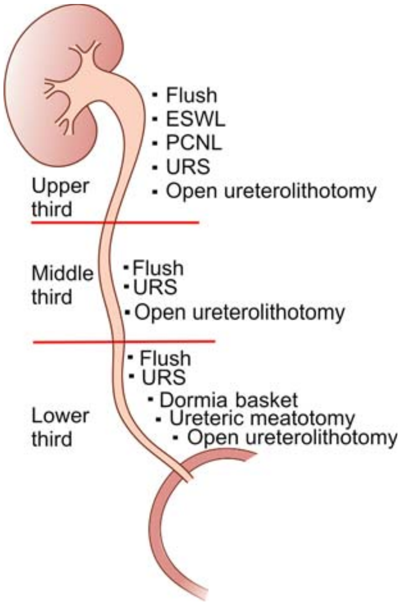

Fig. 26.74: Approaches for removal of ureteric stones.

RECURRENCE OF STONES

Types of recurrence:

a. False recurrence—during intervention, tiny fragments may be left behind.
b. True recurrence—recurred again once all stones are cleared by intervention.

Causes for recurrence:

Existing causes like hyperparathyroidism, hypercalciuria, hyperuricaemia, gout, cystinuria.

Prevention and advice

- Hydration is the main method of prevention. Per day 3 to 4 litres of fluid should be taken.
- Diet:
  - Avoid diets rich in calcium oxalate, sodium (natriuresis causes hypercalciuria) and vitamin C (gets converted into oxalate).
  - Increased intake of dietary fibre—binds with intestinal calcium and decreases the calcium absorption. *Sodium cellulose phosphate is used* for the same.
- *Allopurinol* reduces the uric acid level and so uric acid stone. It also reduces the oxalate level.
- Diet rich in *magnesium* makes calcium oxalate less soluble.
- *Aluminium gel* or *ammonium chloride* prevents the excessive alkalinity and so prevents the recurrence of phosphate stones.
- *Penicillamine* or *alpha-mercaptopropinyl-glycine* may reduce the recurrence rate of cystine stones.
- In idiopathic hypercalciuria, strict low calcium diet along with 5 mg bendrofluzide which reduces loss of calcium in urine or indomethacin 25 mg tds which reduces the calcium secretion or sodium phytate can be given.
- Acetohydroximic acid due to structural similarity to urea is a potent inhibitor of urease and so reduces the bacterial originated stone.

STAGHORN CALCULUS

- It is the stone occupying the renal pelvis and calyces.
- It is usually phosphate or ammonium, magnesium phosphate *(Triple phosphate*) stone.
- It is white in colour, soft, smooth, occurs in pre-existing infection (commonly *E. coli*).
- It can be *unilateral* or *bilateral***.** Patient with *bilateral* stones may go in for *renal failure.*

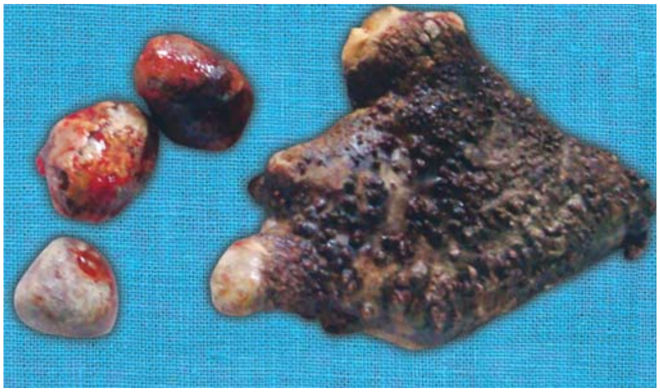

Fig. 26.75: Staghorn calculus.

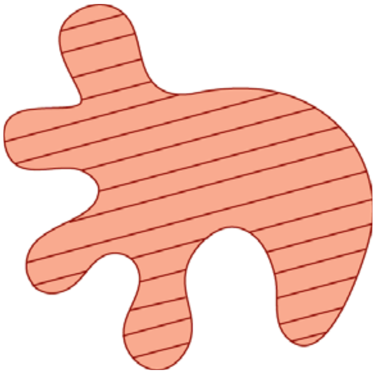

Fig. 26.76: Staghorn calculus (diagrammatic representation).

Presentation

- Pain in loin
- Fever
- Burning micturition
- Haematuria
- In bilateral cases, symptoms are bilateral and often with features of renal failure (hiccough, oedema feet, oliguria, anaemia)

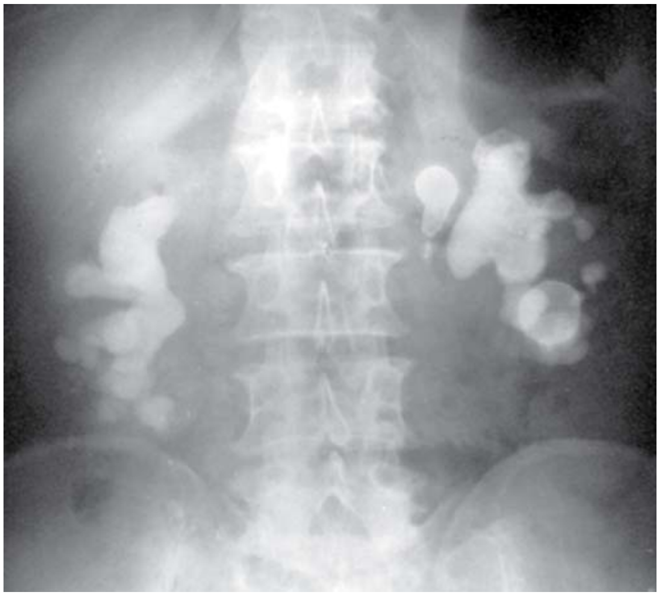

Fig. 26.77: Bilateral staghorn calculi.

Investigations

- U/S, abdomen.
- Plain X-ray, KUB.
- IVU to see the renal function.
- Blood urea and serum creatinine.
- Urine microscopy and urine C/S.
- Isotope renogram—DTPA.

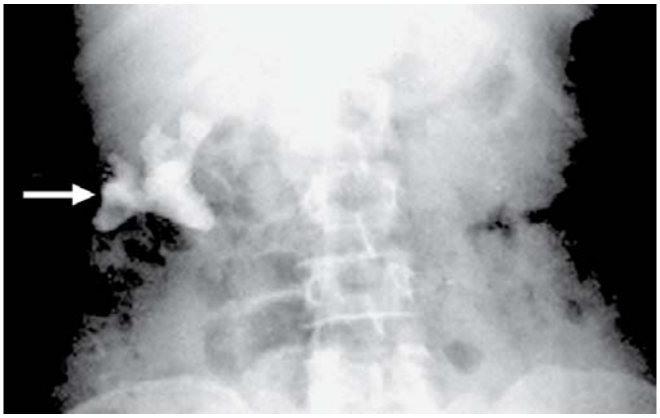

Fig. 26.78: A plain X-ray, KUB, revealing staghorn calculus on right side.

Treatment

Antibiotic is started.

1. *Unilateral stone* is removed by *nephropyelolithotomy*. Multiple incisions are made over the Brodel's line (avascular plane) and assisted by pelvic incision. Stone is removed. Blood transfusion may be necessary as bleeding can occur.

   After that, nephrotomy and pyelotomy incisions are sutured with chromic catgut or vicryl.
2. In *bilateral cases* IVU is very essential. The kidney which is functioning better should be treated first. After 3 months, the other side kidney should be operated upon.
3. If there are *bilateral staghorn calculi with pyonephrosis,* then initially bilateral nephrostomy is done using Malecot's catheter *(Cabot's nephrostomy)*. IVU is done later to see the renal function. If function is present, then both kidneys are operated one after the other in a gap of 3 months. They often need *haemodialysis.*
4. Sometimes in case of severe infection *nephrectomy* is required.

Presently, PCNL is becoming a popular procedure for unilateral or bilateral staghorn calculi.

But when required, *conversion to open nephropyelolithotomy* should be considered.

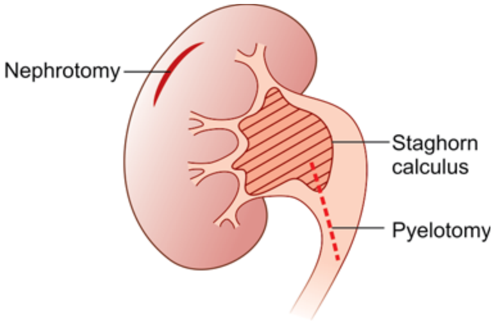

Fig. 26.79: Nephropyelolithotomy. Note the incisions placed.

Complications of staghorn calculus

- Pyelonephritis
- Pyonephrosis
- Perinephric abscess
- Renal failure

Differential diagnosis for radio-opaque shadow which mimics renal stone

- Calcified lumbar or mesenteric lymph node
- Gallstone (10% are radio-opaque)
- Concretion in appendix
- Phleboliths
- Ossified tip of 12th rib
- Chip fracture of transverse process of vertebra
- Calcified renal tuberculosis
- Calcified suprarenal gland
- Drugs or foreign body in the alimentary canal

BENIGN TUMOURS OF KIDNEY

They are rare.

Types

- *Renal papillary adenoma*: Small discrete adenoma arises from tubular epithelium, < 5 mm in size, cortical, yellow circumscribed tumour. It is potentially malignant, may turn into papillary type of RCC, especially if size > 3 cm.
- *Angiomyolipoma*: It contains fat, vessels and smooth muscles. It is commonly associated with tuberous sclerosis (50%). Tumour is susceptible for spontaneous haemorrhage.
- *Oncocytoma:* It is an epithelial tumour arising from intercalated cells of collecting ducts. Its cells show large nucleoli. These large eosinophilic cells contain numerous mitochondria. Tumour is homogenous well encapsulated tan coloured. It can attain large size and can be multicentric.

WILM'S TUMOUR (NEPHROBLASTOMA)

> *The patient was a 3-year-old girl with a kidney tumour which had grown to immense proportions in a short time. The child, anemic and emaciated, was admitted with an enormous mass in the right abdomen and with definite ascites. After nephrectomy, the little child recovered uneventfully. A few months later, however, a recurring abdominal mass was again palpable and shortly afterward the child died.*
>
> —**Max Wilms,** 1899

- It arises from embryonic connective tissue containing epithelial and connective tissue elements.
- It is located in one of the poles of the kidney.
- It is bilateral in 5% cases. It is common in first 4 years of life.

Pathology

Gross

It is smooth, soft, fleshy, pinkish white in colour, often with haemorrhagic areas.

Microscopically

Malignant primitive glomeruli and primitive tubules, with epithelial and connective tissue cells exist side by side, one of the types is usually prominent.

Types

A. *Favourable histology (FH)—Wilms'—means without anaplastic features*
B. *Unfavourable histology (Un FH)—Wilms'—means with anaplastic features*

Histological Types

- Cystic nephroma.
- Mesoblastic nephroma.
- Nephroblastoma.

Spread

Mainly through blood into the lungs, liver and rarely to bones.

Clinical Features

- Incidence is equal in both sexes.

Triad

Mass, fever, haematuria

- *Mass abdomen is commonest presentation*. Mass is smooth, mobile, firm or hard, lobular, located in the loin, moves with respiration, bimanually palpable, ballotable, with dullness in renal angle and with resonant band in front. It does not cross the midline (Differential diagnosis is adrenal neuroblastoma which is knobby and nodular, does not move with respiration and crosses the midline).

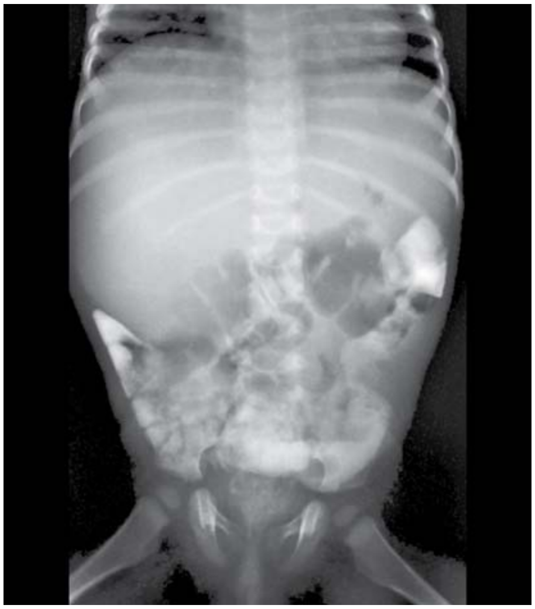

Fig. 26.80: X-ray showing mass lesion in right abdomen. It could be Wilm's tumour.

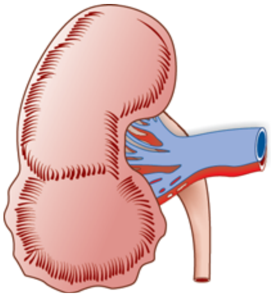

Fig. 26.81: Renal tumour.

- *Fever*—may be due to tumour necrosis.
- *Haematuria* is a grave sign as it signifies rupture of tumour into the renal pelvis.
- *Hypertension*—in 25% cases.
- 12% of cases are associated with congenital anomalies and syndromes (Glaucoma, aniridia, Beckwith syndrome).

Staging of Wilm's

I. Disease confined to kidney and completely resectable
II. Tumour extends beyond kidney, but can be excised completely
III. Residual disease after resection with positive lymph nodes and massive spillage
IV. Blood born metastasis
V. Bilateral disease

Investigations

- U/S abdomen, CT scan abdomen.
- IVU.
- Renal angiography.

- X-ray, abdomen—*egg shell* peripheral calcification is diagnostic.
- MRI.

Differential diagnosis

- Adrenal tumour
- Retroperitoneal tumour
- Renal cyst
- Polycystic kidney disease

Treatment

- *Nephrectomy and postoperative radiotherapy in case of unilateral tumours.*
- In bilateral cases, either bilateral partial nephrectomy or nephrectomy on one side with partial nephrectomy on the other side is done *(Nephron sparing surgery)*.

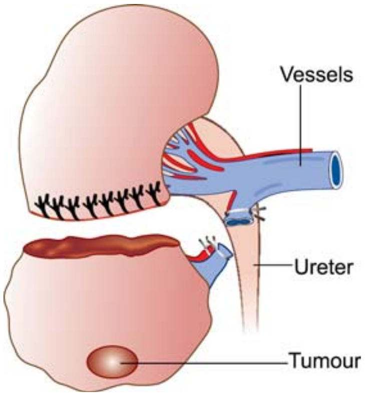

Fig. 26.82: Partial nephrectomy.

- *Chemotherapy:* Actinomycin D, vincristine, doxorubicin.
- Preoperative RT in locally inoperable cases—later surgery and then chemotherapy is given.

Prognosis

Five years survival rate is 80% under the age of one year. Recurrence rate is high within one year.

Prognostic Factors

Age, bilaterality, size of the tumour, presence of haematuria, spread of tumour, histology type.

RENAL CELL CARCINOMA (RCC)

Also known as *Hypernephroma—(it is a misnomer), Grawitz tumour, clear cell carcinoma, Internist tumour.*

- It is an adenocarcinoma arising from renal tubular cells, most common site is proximal renal tubule. More common in males; more common in 5th-6th decade of life.
- It is 3% of all adult malignancies.
- **RCC types**:
  - *Clear cell* (75%)—nonpapillary, associated with loss of sequence in short arm of chromosome no.3. It can be sporadic (common, 95%) or familial.
  - *Papillary* (15%)—multifocal commonly, with papillary pattern.
  - *Chromophobe RCC* (5%)—arising from intercalated cells of collecting duct; shows halo around the nucleus; carries excellent prognosis.
  - *Collecting duct of Bellini carcinoma* (1%)—arises from collecting duct of medulla; tumour with fibrous stroma typically in medullary location.

Aetiology

- It is associated with von Hippel-Lindau disease (cerebellar haemangioblastoma, retinal angiomatosis, tumour or cysts of pancreas). Here RCC is commonly bilateral
- Diet high in animal fat
- Environmental factors like asbestos, lead, cadmium and tobacco
- Cigarette smoking
- Chromosomal aberration, tuberous sclerosis
- Acquired cystic kidney disease after long-term dialysis
- *Birt-Hogg-Dube' syndrome* with hereditary chromophobe RCC and oncocytoma
- Cortical renal adenoma could be RCC by itself
- Analgesic nephropathy

Pathology

Gross

It attains a large size. Commonly located in *upper pole*, sometimes in lower pole but rare in the middle. Cut section is *yellowish* due to lipoid content with areas of haemorrhage and necrosis. This noncapsulated tumour is very vascular.

Microscopy

Malignant cells which are cubical or polyhedral containing lipid, cholesterol and glycogen.

Histological types—clear (75%), granular, spindle, sarcomatoid, papillary (15%)

Spread

1. *Local*: Into the perinephric pad of fat, calyces and renal pelvis.
2. *Blood spread*: RCC enters the renal vein as *proliferating tumour thrombus,* which extends into the IVC and later gets detached causing "*cannon ball secondaries*" in the lung, which are often calcified. Once primary tumour is removed, secondaries may regress due to tumour immunity. Occasionally, secondaries occur in bone, liver and brain. Left testicular vein which drains into left renal vein may gets blocked by proliferating tumour thrombus resulting in *irreducible left sided varicocele. More than 50% spread to lungs; 30% to bones*; then liver, adrenal and brain.
3. *Lymphatic spread*: To hilar lymph nodes, para-aortic lymph nodes.

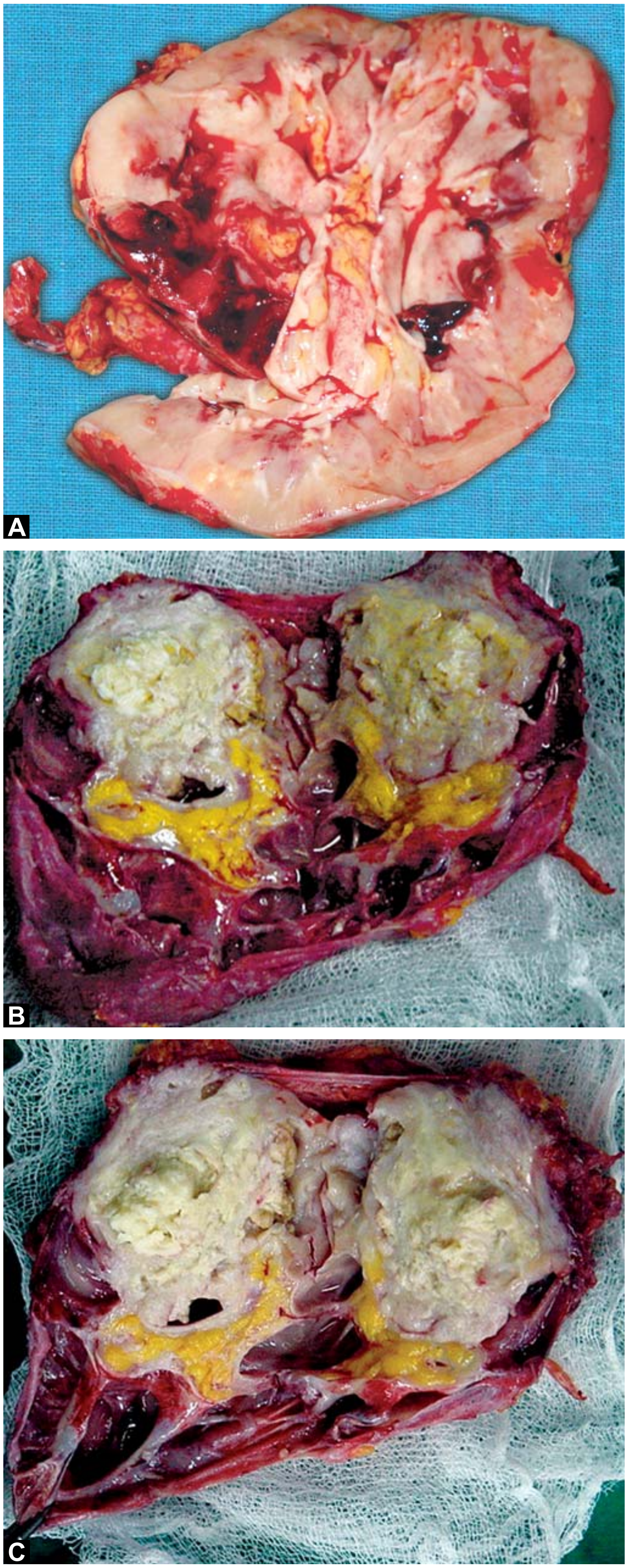

Figs 26.83A to C: Renal cell carcinoma in upper pole of kidney—resected specimen. It is from two different patients.

AJCC (American Joint Committee on Cancer) Staging: TNM Staging

Tx: Primary tumour cannot be assessed
T0: No primary tumour
T1: Tumour less than 7.0 cm size, limited to kidney
T1a is tumour 4 cm or less; T1b is tumour 4-7 cm in size
T2: Tumour more than 7.0 cm size, limited to kidney
T3: Tumour extends into major veins, adrenals, perinephric fat but not into the Gerota's fascia
  T3a : Into adrenal or perinephric tissue
  T3b : Into renal vein or IVC below diaphragm
  T3c : Tumour extends into IVC above the diaphragm
T4: Tumour invades Gerota's fascia and extends beyond
N0: No lymph nodes
N1: Spread to single regional lymph nodes
N2: Spread to more than one regional lymph nodes
M0: No blood spread
M1: Distant spread to—lungs (75%); soft tissues (25%); bones (20%); liver (15%); CNS (8%); skin (8%)

Robson-Flocks and Kadesky staging

Stage 1	Tumour confined to renal parenchyma
Stage 2	Tumour invasion to perinephric fat but confined within Gerota's fascia
Stage 3	a. Tumour invasion to renal vein or IVC b. Invasion to regional lymph nodes c. Both a + b
Stage 4	Invasion to adjacent organs other than adrenal Distant metastasis

Note:
45% present as early disease; 25% as locally advanced disease; 30% as metastatic disease.

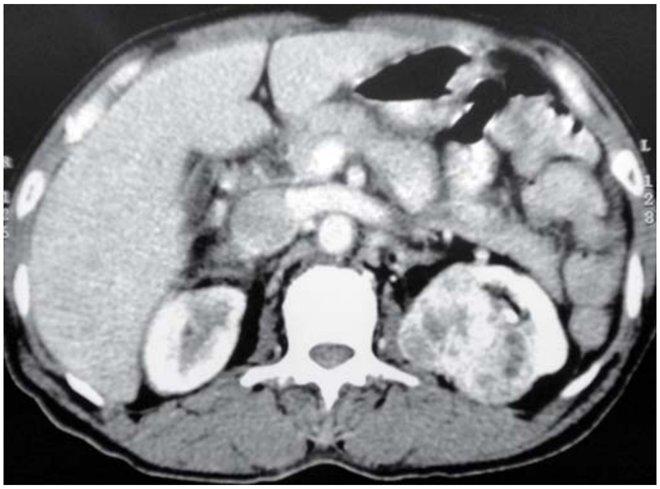

Fig. 26.84: CT scan showing RCC—left sided.

Clinical Features

- M : F :: 2 : 1.
- Haematuria—30%.

The men who make history have no time to write it.

- Clot colic.
- Dragging discomfort in the loin.
- Mass in the loin which moves with respiration, mobile, nodular, hard, with dull renal angle and resonant band in front.
- Left-sided *varicocele* which is *irreducible* in left-sided RCC.

Triad of RCC

- Pain
- Haematuria
- Palpable renal mass

Atypical presentations: 25%

- Due to secondaries:
  1. Pathological fractures.
  2. Persistent cough and haemoptysis.
- *Persistent pyrexia* with no evidence of infection ***(P***yrexia of ***U***nknown ***O***rigin)—20%
- *Constitutional symptoms:* Malaise, lethargy and severe anaemia.
- *Polycythemia:* 4%—due to increased secretion of erythropoietin.
- Hypercalcaemia, hypertension. Hypercalcaemia due to PTH like hormone secretion, hypertension is due to increased secretion of renin from kidney tissue adjacent to tumour.

Surgical renal conditions associated with hypertension

- Polycystic kidney disease
- Renal cell carcinoma
- Renal artery stenosis

- *Nephrotic syndrome:* Very rare.
- *Stauffer's syndrome:* Nonmetastatic reversible liver dysfunction which gets corrected after nephrectomy. It is 7% common. It carries poor prognosis.
- Cushing's syndrome.
- Leukaemoid reaction due to bone marrow stimulation.
- Secondary amyloidosis 5%.

Investigations

- Urine microscopy for RBCs.
- IVU—shows mass lesion and irregular filling defect.
- U/S, abdomen—to know the size, extension, lymph node involvement, spread to the liver, status of renal vein and IVC.
- *CT scan*: It is confirmatory and also helps to know the status of renal vein and IVC. Multi-detector CT and CECT (contrast enhancement CT) are very useful in detecting early lesion/function/spread/venous status. Lymph node status, tumour extension are well made out with CT. Contrast enhancement CT scan helps to find out function of opposite kidney and tumour thrombus in renal vein or IVC.

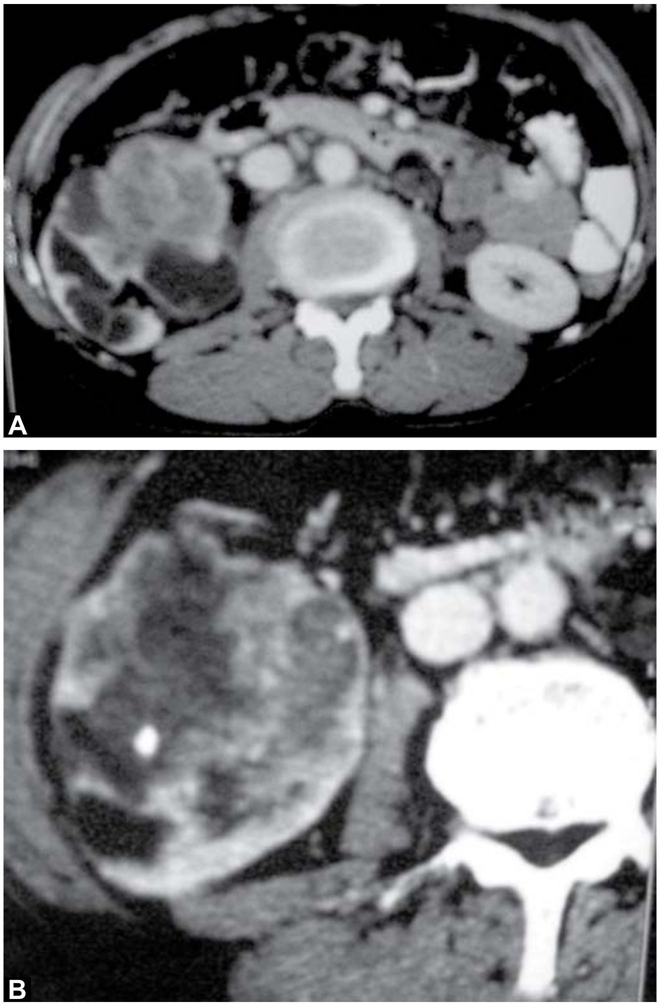

Figs 26.85A and B: CT scan showing features of renal cell carcinoma.

- *Renal angiogram* through Seldinger technique via transfemoral route, to see the vascularity. *Pharmacoangiogram* (Inject noradrenaline along with dye while doing angiogram). As tumour vessels are autonomous they will not constrict whereas adjacent normal vessels will constrict, so *tumour blush* is visualised. Through angiogram therapeutic embolisation of tumour can be done to reduce the vascularity of tumour.
- Chest X-ray shows *cannon ball s*econdaries. Often it is *calcified.*
- *CT chest is ideal and more reliable.*
- Bone scan to see bone secondaries.
- Peripheral smear, serum calcium, haematocrit and ESR.
- *Inferior venacavagram* to assess the tumour extension through renal vein into IVC and extension in the IVC infra- or supradiaphragmatic.
- MRI is diagnostic to assess the tumour thrombus in IVC. MRI/MR angiogram is unique in identifying the spread into IVC especially in the thorax. In such occasion oesophageal endosonography is also useful to visualise thoracic extension of the tumour thrombus.

Differential diagnosis

- Polycystic kidney disease
- Solitary cyst of kidney
- Adrenal tumour
- Retroperitoneal tumour
- Carcinoma colon

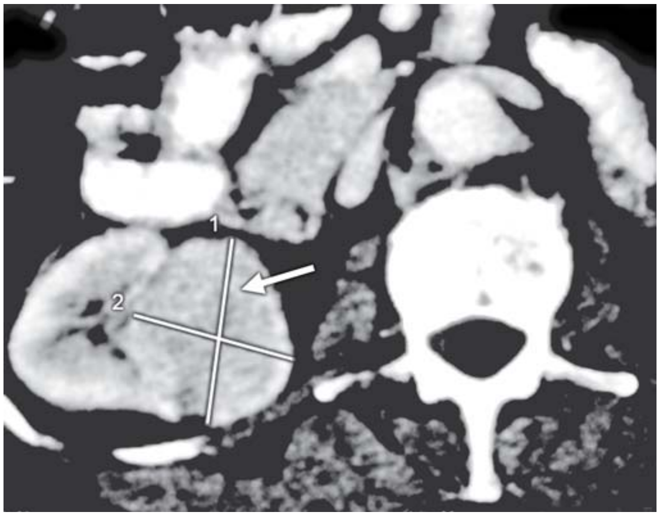

Fig. 26.86: CT scan of carcinoma right kidney.

Treatment

Surgery is the treatment of choice.

- *Radical nephrectomy*:

Structures removed are:

- Entire kidney along with tumour
- Perinephric tissue
- Ipsilateral adrenal gland
- Proximal 2/3rd ureter/as low as possible
- Lymph nodes from crus of diaphragm to aorta bifurcation with renal hilar nodes

    - Transperitoneal approach is often used. Retroperitoneal/Nagamatsu (resection of 11th ribs) approach/thoracoabdominal approach/posterior vertical are other approaches used. Patient will be in lateral position. After laparotomy, colon is mobilised medially. Vessels are identified and dissected and ligated securely (transfixation and three ligatures proximally using nonabsorbable silk/polypropylene sutures).
    - Preoperative renal artery embolisation can be done to decrease vascularity and to facilitate the removal of entire tumour.
    - Always renal vein is ligated early so as to prevent tumour dislodgement (But early ligation of renal vein may cause congestion and torrential bleeding and so renal artery is ligated first commonly in practice). But usually renal artery which is posterior is dissected and clamped or ligated in continuity; then renal vein is ligated and divided; later ligated renal artery is divided. In case of IVC extension, after applying an oblique vascular clamp, IVC is opened, tumour thrombus is removed and IVC is sutured. In supradiaphragmatic vena caval extension of thrombus, cardiopulmonary bypass is necessary. First renal vessels (artery and vein) should be ligated to prevent tumour dissemination and bleeding.

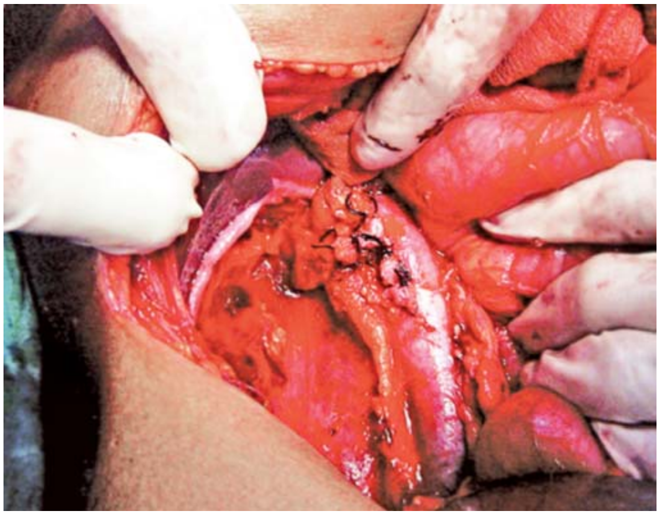

Fig. 26.87: Renal cell carcinoma, on table after nephrectomy. Note the vascular pedicles.

- Even in large fixed tumour, palliative nephrectomy or debulking is advised as it may cause regression of secondaries.
- *Nephron sparing surgeries* is done in bilateral RCC (Bilateral partial nephrectomy). Renal artery is temporarily occluded using vascular clamps and kidney is cooled to have proper control of bleeding and adequate visualisation of the line of resection. Partially resected specimen is assessed by frozen section biopsy. Retained partial capsule is sutured after haemostasis. Renal arterial clamp is released.
- *No role for radiotherapy.*
- *Chemotherapy:* Vinblastine and progesterone can be tried. RCC is a remarkably refractory solid tumour—chemoresistant.
- Interferons and interleukins have shown beneficial effects.
- Preoperative renal artery embolisation can be done to decrease the vascularity of the tumour (Using clot, gelfoam, spheres).
- Antiangiogenic drugs like endostatin and angiostatin are under trial. RCC is highly vascular and there is highly angiogenic environment to give possible benefit by anti-angiogenesis.
- Patient with solitary lung metastasis may get benefited by wide resection of the secondary.
- Humanised monoclonal antibodies like bevacizumab which neutralises VEFG are under trial.

Note:

- Laparoscopic approach is becoming popular. Here renal artery is ligated first.
- ***Sisitinab***, a multitargeted tyrosine kinase inhibitor prevents angiogenesis and tumour proliferation is said to be useful.

Prognosis

- Five-year survival rate is 40%.
- In early localised disease, it is 70-90%. In advanced and metastatic disease it is 10-15%
- 5-year survival is 65% for stage I and II; 40% for stage III; 10% for stage IV

Prognostic factors are:

- Tumour size more than 4 cm carries poor prognosis
- Extension into the renal vein
- Presence of secondaries
- Differentiation
- Local extension
- Hypercalcaemia and Stauffer's syndrome carry poor prognosis

Note:

- Most common type of renal pelvis and ureteric tumours is transitional cell carcinoma (90%).
- Renal pelvic tumours account for 5% of renal tumours. It is common in Taiwan.
- It is more aggressive.
- Treatment is radical nephroureterectomy with removal of cuff of bladder adjacent to ureteric orifice.

APPROACHES TO KIDNEY (SURGICAL)

- Posterior subcostal (oblique lumbar incision).
- Laparotomy approach (anterior approach).
- Thoracoabdominal approach.
- Nagamatsu approach (approach following excision of 11th rib in case of upper pole tumour).
- Posterior vertical lumbotomy approach (Gilvernet's approach).
- Laparoscopic approach.
- Retroperitoneoscopic approach.

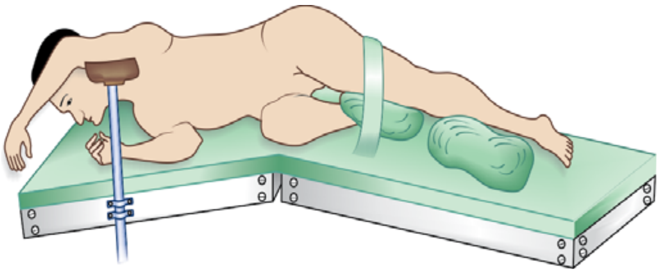

Fig. 26.88: On table position to approach kidney surgically.

B. Urinary Bladder

CHAPTER OUTLINE

- Anatomy
- Ectopia Vesicae
- Urachal Anomalies
- Vesical Calculus
- Cystitis
- Recurrent Cystitis
- Interstitial Cystitis
- Schistosoma Haematobium
- Thimble or Systolic Bladder
- Bladder Tumours
- Transitional Cell Carcinoma
- Ureterosigmoidostomy
- Rupture Bladder
- Residual Urine
- Malakoplakia
- Neurogenic Bladder
- Vesicoureteric Reflux
- Bladder Diverticula
- Urinary Diversion
- Urinary Fistulas

ANATOMY

- It is a muscular reservoir of urine situated in the pelvis, posterior to the pubic bone from which it is separated by the retropubic space.
- The size, shape, position of bladder depends on the amount of urine it contains.
- In females, the peritoneum is reflected from the superior surface of the bladder onto the anterior wall of the uterus close to the junction of the body of the uterus and cervix.
- In males, the peritoneum is reflected from the superior surface of the bladder over the superior surface of the ductus deferens and seminal vesicles.
- The neck of the bladder is the most fixed part, lies 3-4 cm behind the pubic symphysis. In males it lies on the prostate and in females it rests on the anterior vaginal wall which is also supported by the levator ani muscle.

Structure of the Bladder

- The bladder wall consists of interlacing fibres of smooth muscle. The urothelium consists of transitional epithelium which is loosely attached and easily separated from the bladder wall except over the *trigone*.
- *Trigone* is a triangular area over the lower part of the base of the bladder. The ureters opens into the posterolateral angles of the trigone. The internal urethral orifice opens at the inferior angle of trigone.

Sphincters

- *Internal sphincter*: Smooth muscle, situated at the neck of the bladder, micturition follows its relaxation. It is innervated by autonomic nervous system.
- *External sphincter*: It consists of striated muscle, innervated by the perineal branch of pudendal nerve (S_2,S_3).

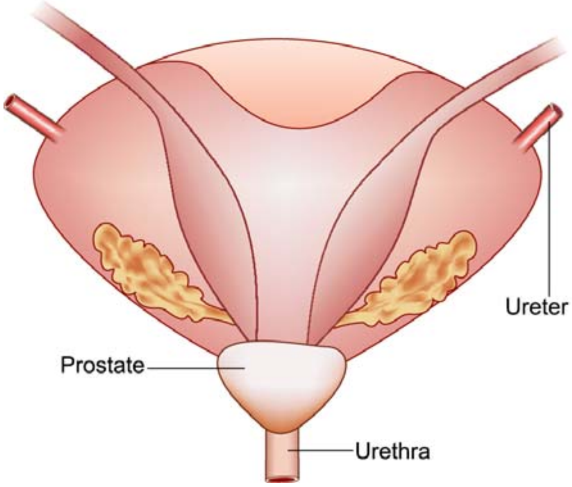

Fig. 26.89: Anatomy of urinary bladder—posterior aspect.

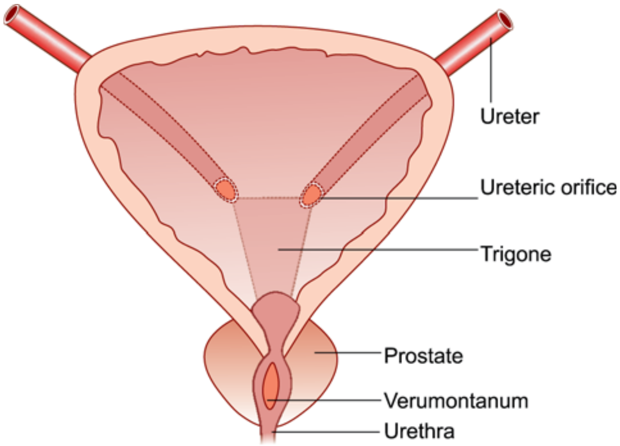

Fig. 26.90: Urinary bladder—inner anatomy.

Retrograde cystogram is the most reliable investigation in case of bladder rupture.

Blood Supply

- Superior vesical artery, a branch of internal iliac artery.
- Inferior vesical artery:
  - Venous drainage is by the vesical plexus of veins, which drains into the internal iliac veins. These veins connect directly to veins of hip bones, the heads of femoral and vertebral bodies. This accounts for the site of occurrence of bony metastasis in carcinoma of the bladder and prostate.

Lymphatic Drainage

Lymphatics drain into internal iliac and external iliac nodes.

ECTOPIA VESICAE (Extrophy of the Bladder)

- It is incomplete development of the infraumbilical part of the anterior abdominal wall and anterior wall of the bladder. It is of embryological origin.
- It is often associated with the spina bifida and other congenital anomalies.
- It is more common in males (4 : 1).
- Red mucus membrane of posterior bladder wall protrudes out with visible urine efflux from ureteric orifice.
- Umbilicus is absent.
- There is separation of pubic bones.
- In males, *epispadias* is commonly present with rudimentary prostate and seminal vesicle.

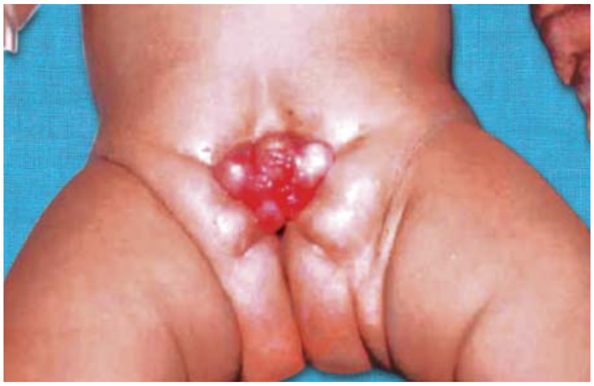

Fig. 26.91: Child showing extrophy bladder.

- Testes are normal.
- Bilateral inguinal hernias are common.
- In females defective external genitalia is common.

Problems

- Repeated soakage
- Ulceration
- Pain
- Recurrent pyelonephritis
- Renal failure
- Metaplastic changes in mucosa can lead into *adenocarcinoma*
- **50%** of patients die of renal failure

Treatment

- Staged procedure.
- Initial diversion of urine to colon/rectum.
- Iliac osteotomy and closure of the abdominal wall.
- Correction of epispadias.
- Sometimes cystectomy and permanent ureterosigmoid diversion is required.
- Condition has got high mortality due to infection and renal failure.

URACHAL ANOMALIES

- *Patent urachus with urine leak*: It signifies obstruction distal to bladder.
- *Urachal sinus with discharge.*
- *Urachal cyst:* Present as immobile, fluctuant swelling in the abdominal wall.

Sepsis, tuberculosis, adenocarcinoma, can occur in patent urachus.

VESICAL CALCULUS

> *The signs of a stone in the bladder are, great and frequent irritations to make water, a stoppage in the middle of making it, and a pain with heat just after it is made; a tenesmus, pain in the extremity of the urethra, incontinence or suppression of urine, together with a quiet pulse, and the health in no bad state.*
>
> —**William Heberden,** 1802

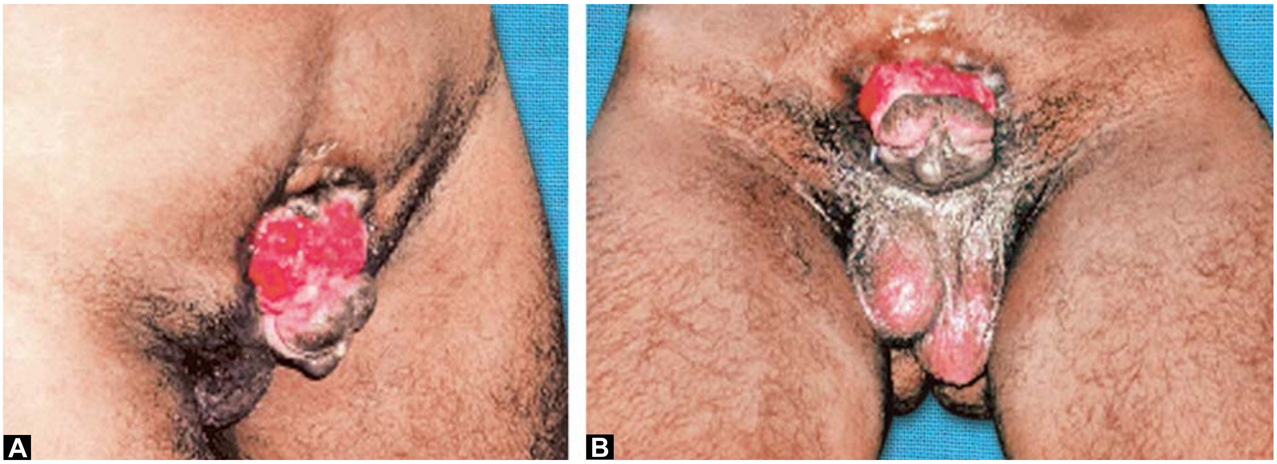

Figs 26.92A and B: Ectopia vesicae in a man of 25 years old.

Types

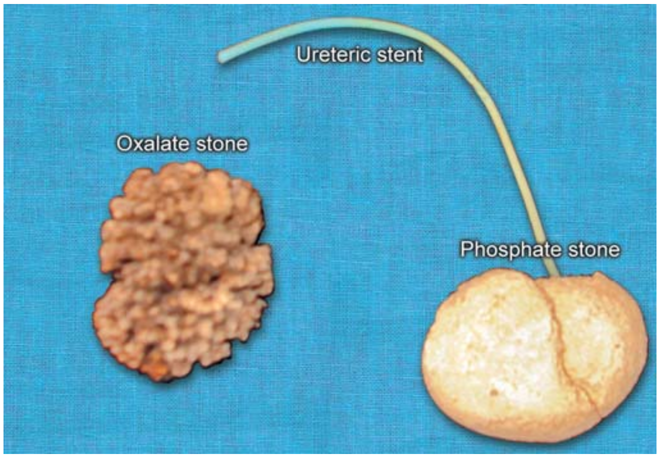

Fig. 26.93: Bladder stones—oxalate and phosphate. Phosphate stone has occurred in an indwelling J ureteric stent.

- ***Primary vesical calculus*:** Occurs in sterile urine.
  - Usually comes down from kidney through ureter into the bladder and it gets enlarged here. It is usually *oxalate* stone (Jack stone).
  - Oxalate stone is usually single, primary stone, brownish black in colour (due to deposited blood pigment over the surface), hard and with spikes over the surface which irritates bladder mucosa causing haematuria *(mulberry stone).*
- ***Secondary vesical calculus***: Occurs in the presence of infection. Most common bladder stone.
  - It is usually *phosphate stone*, occurs in bladder only.
  - Phosphate stone is smooth, soft, ivory-white in colour. It is either calcium phosphate or ammonium, calcium and magnesium phosphate *(Triple phosphate stone).*
  - *E. coli* is the common organism.
- ***Uric acid and urate stones*** are single or multiple, primary, nonradio-opaque, smooth, pale yellow in colour.
- ***Cystine calculus**:* Occurs in cystinuria and is radio-opaque due to high sulphur content.

Aetiology

- Infection
- Hypercalciuria of any cause
- Hyperoxaluria
- Cystinuria
- Bed ridden and paraplegic patients
- Gout and other hyperuricaemic patients
- Diverticula bladder
- Obstruction to urine flow by BPH, urethral stricture, bladder neck obstruction
- Neurogenic bladder
- Schistosomiasis
- Foreign body in bladder

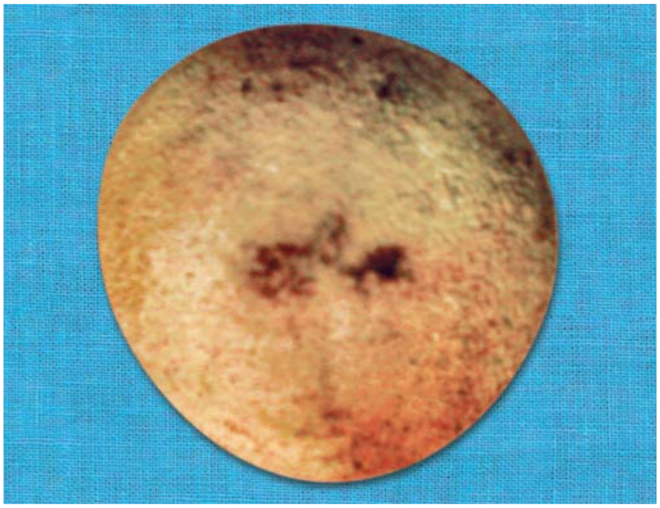

Fig. 26.94: Large bladder stone removed by open suprapubic extraperitoneal cystolithotomy.

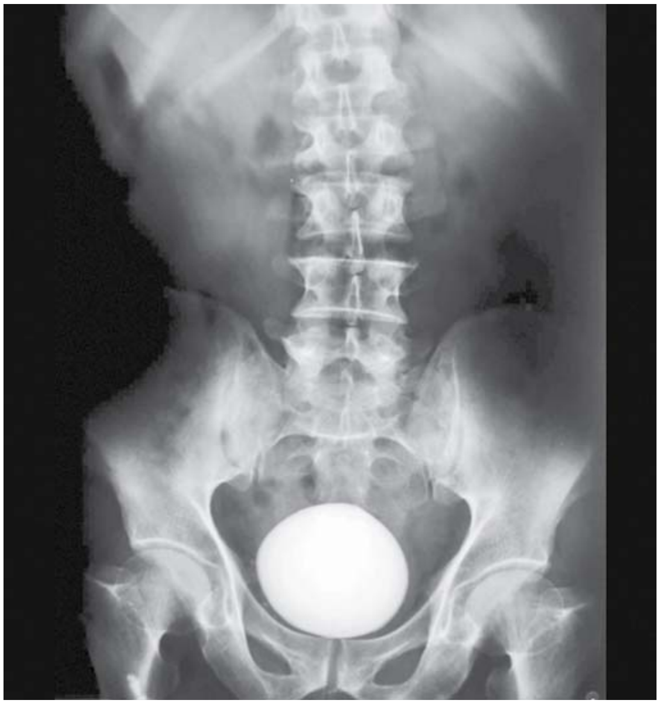

Fig. 26.95: Plain X-ray KUB showing large bladder stone.

Clinical Features

- Common in males. Often occurs in children.
- *Frequency* is more during day than night, because during day, due to ambulation stone comes in contact with the trigone of the bladder and irritates, whereas during night, stone slips towards the fundus, away from the trigone and so less frequency and pain.
- *Pain*: More during day which is referred to the tip of penis or labia. Also increases during jolting movements. Suprapubic pain and tenderness may be present.
- *Haematuria***:** Often terminal.

It is better to have an idea than to have thousand opinion.

- Interruption of urinary stream and often acute urinary retention.
- Features of cystitis: Burning micturition, fever, pain.
- P/R or P/V: Large stone may be palpable.
- Stone may be identified incidentally in plain X-ray, KUB or U/S, abdomen.

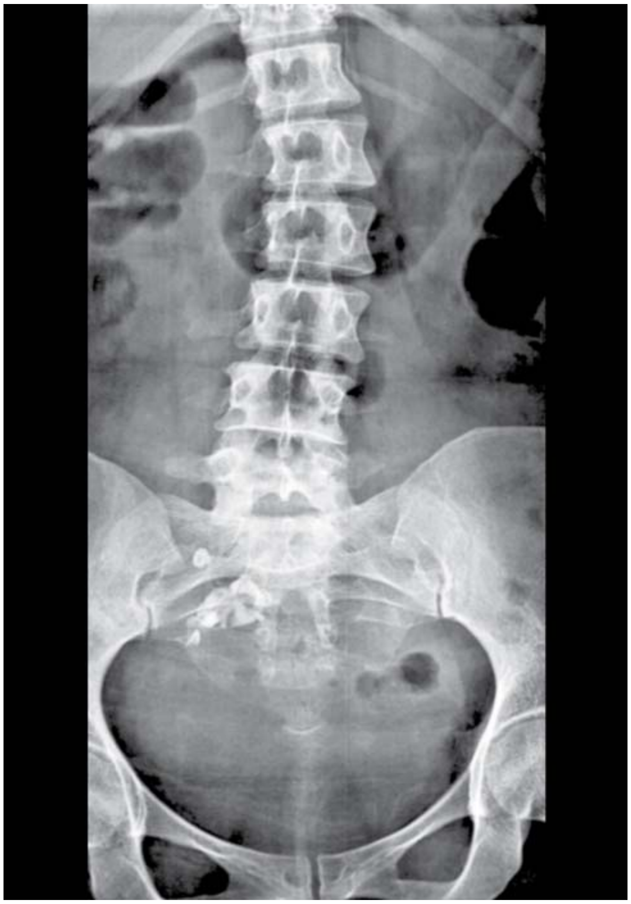

Fig. 26.96: Plain X-ray showing dermoid cyst of ovary with calcification. It is teratoma of ovary.

Investigations

- Urine microscopy
  - Envelope crystals in oxalate stone,
  - Hexagonal type in cystine stone.
- Urine C/S.
- Blood urea, serum creatinine, serum calcium, inorganic phosphate, uric acid.
- Plain X-ray, KUB shows radio-opaque stones—90% are radio-opaque.

Radiological D/Ds for vesical calculi in females:

- Calcified fibroid uterus
- Dermoid cyst ovary

- IVU to see function of the kidney.
- U/S abdomen is diagnostic.
- Cystoscopy to see radioluscent stone.

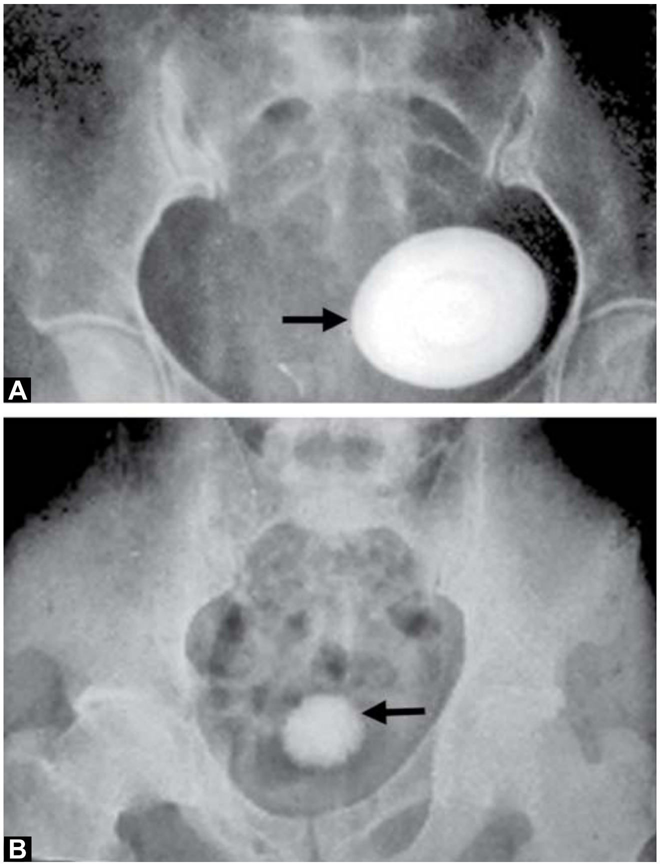

Figs 26.97A and B: A plain X-ray, pelvis showing (A) Phosphate stone. Note the laminations. (B) Oxalate stone. Note the spiculated margins.

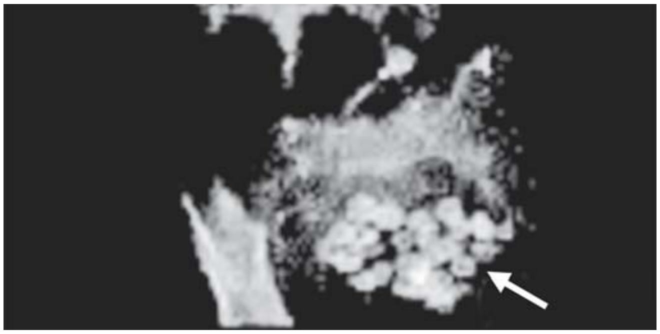

Fig. 26.98: Plain X-ray revealing multiple bladder calculi.

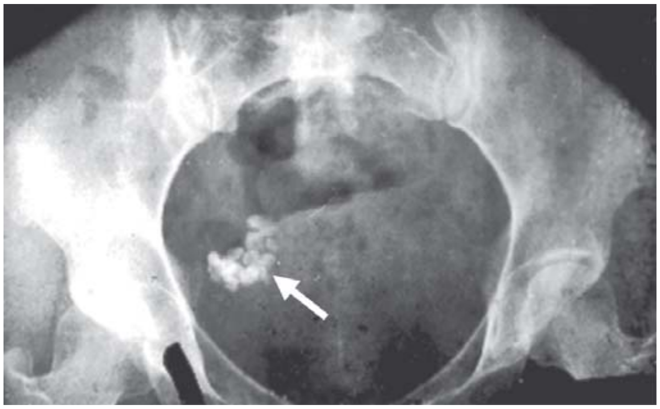

Fig. 26.99: Dermoid cyst ovary. Note the radio-opaque teeth.

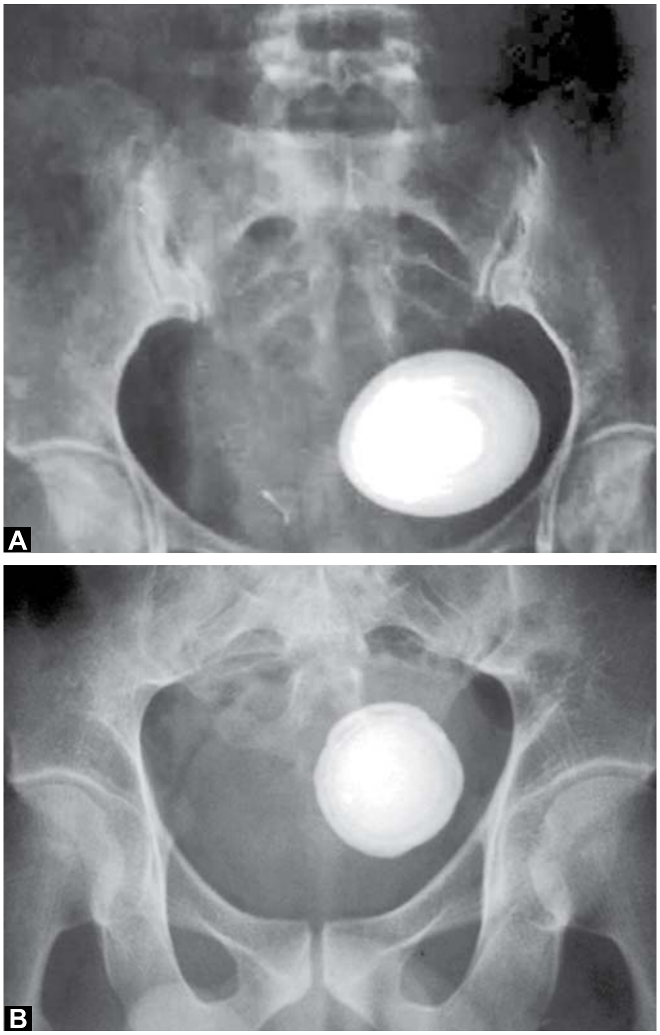

Figs 26.100A and B: X-ray showing vesical calculus. Note the laminated appearance.

Treatment

1. **Cystoscopic litholapaxy**
   Under GA, cystoscope is passed and the stone is visualised. It is fragmented by pneumatic, laser, electromagnetic waves or mechanohydraulic lithotripsy. The bladder is flushed using an irrigator (Freyer's evacuator or irrigator or Ellik's evacuator).

Contraindications

- Too large stone
- Too small stone. Too soft stone. Too many stones
- Stone in bladder diverticula
- Bladder tumour
- Contracted bladder
- Patient's age below 10 years
- When patients general condition is poor, as the procedure takes a longer duration it is avoided and open removal is advised

2. **Suprapubic open cystolithotomy**
   Through pfannenstiel incision, bladder is approached extraperitoneally. Bladder is identified by its detrusor muscle pattern, which is criss-cross and also its venous pattern. Bladder is opened near the fundus and stone is removed. Bladder is closed often with SPC using Malecot's catheter and Foley's catheter is passed per urethera. Wound is closed in layers with a drain.

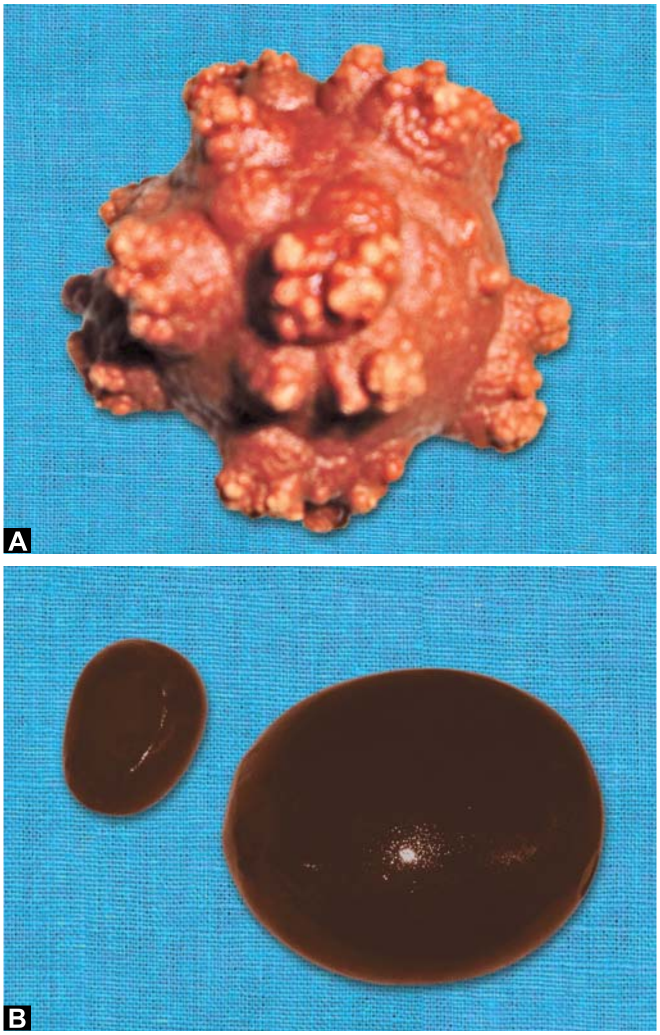

Figs 26.101A and B: Two types of bladder stones—one with surface projections, other with smooth surface.

3. **Suprapubic percutaneous litholapaxy**
   This procedure is becoming popular. When cystoscope cannot be passed per urethra, bladder is approached suprapubically. Through a needle, guidewire and dilators, a track is created through which a nephroscope is passed to remove the stone after fragmenting.
   The cause is treated.

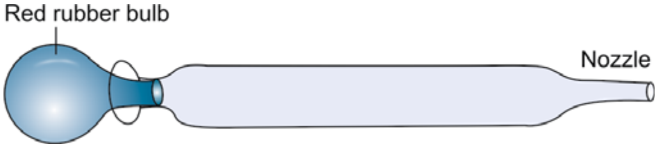

Fig. 26.102: Asepto syringe used for bladder wash in patients with urinary catheter which has been blocked; or when haematuria is noticed in the catheter. It is also used to fill the bladder with saline.

Men take only their needs into consideration, never their abilities.

CYSTITIS

Inflammation of the bladder mucosa due to different causes.

Causes

- Acute bacterial cystitis
- Chronic cystitis due to tuberculosis, syphilis
- Interstitial cystitis
- Radiation cystitis
- Cystitis due to schistosomiasis
- Postmenopausal atrophic cystitis

Predisposing Factors

- Congenital urinary tract anomalies
- Short urethra in females may cause ascending infection and cystitis.
- Initial period of sexual contact in females can cause cystitis—Honeymoon cystitis.
- Catheters, instrumentation.
- Bladder stone.
- BPH, carcinoma prostate.
- Cystocele, bladder diverticulum.
- Stricture urethra, bladder neck obstruction.
- Bladder tumours.
- Pregnancy.
- CNS diseases, spinal injury.

Organisms

- *E.coli, Klebsiella, Pseudomonas, Staph. aureus, Staph. albus, Proteus.*
- *Candida albicans*, fungal infection.

Features

- Painful urination, frequency, strangury, incomplete emptying, with often retention.
- Occasionally haematuria can occur.
- Burning urine, discoloured foul smelling urine.
- Fever, chills, rigors, suprapubic pain and tenderness and often loin pain.
- Septicaemia can develop in severe cystitis.

Investigations

- Total count will be raised.
- Urine will show pus cells and culture will reveal the organisms.
- X-ray may show a stone. IVU, cystogram are also helpful.
- U/S abdomen will show thickening of bladder mucosa, stone, tumour.
- Cystoscopy reveals bladder inflammation, mucosal changes.

Treatment

- Appropriate antibiotics like quinolones, aminoglycosides, cephalosporins.
- Plenty of water intake to flush the bacteria from the bladder.
- Often admission and parenteral antibiotics may be needed.
- Evaluation and therapy for cause should be done.
- Strict aseptic precaution is used while catheterisation, cystoscopy.
- Often long-term antibiotics are needed to prevent recurrent infection.

RECURRENT CYSTITIS

Causes

- Failure of early treatment for acute cystitis
- Reinfection by same or different organisms.

Types

- Chronic cystitis.
- Phosphate deposition in the ulcerated mucosa causing encrusted cystitis.
- Proliferation of mucosa and submucosa causing proliferative cystitis. It may be cystitis follicularis, cystitis polyposa, cystitis glandularis, cystitis cystica.
- Gas filed cysts in the submucosa causing cystitis emphysematosa.
- Necrosis of bladder wall can occur leading to gangrenous cystitis.

Management

- Urine microscopy and culture.
- Antibiotic and urinary antiseptics like hippuric acid, mandelic acid.
- Proper hygiene.
- Plenty of oral fluid intake and frequent emptying of the urine.

INTERSTITIAL CYSTITIS (Hunner's Ulcer, Elusive Ulcer)

- It is common in females
- Common in western countries
- Common in psychic females

Pathology

There is pancystitis with fibrosis of vesical musculature along with linear ulcers in the bladder mucosa. Microscopically, severe inflammation of all layers of bladder with fibrosis is observed. Bladder eventually becomes **thimble (systolic) bladder** with decreased bladder capacity up to 30-60 ml (less than 100 ml).

Clinical Features

- Pain.
- Decreased bladder capacity.
- Pain increases with bladder distension.
- Frequency and often haematuria.

Investigation

Cystography and cystoscopy are diagnostic.

Treatment

- Hydrostatic dilatation.
- Instillation of dimethyl sulphoxide (Rimso 50).
- Ranitidine instillation.
- Ileocystoplasty or caecocystoplasty, to increase the bladder capacity.
- Urinary diversion.

Differential Diagnosis

Other causes for thimble bladder.

SCHISTOSOMA HAEMATOBIUM (Endemic Haematuria, Urinary Bilharziasis) (Swimmer's Itch)

Life Cycle

Fresh water snail—(Intermediate host)
↓
Bifid tailed embryos (cercariae) in infected water
↓
Penetrate the skin
↓
Through circulation enters the liver
↓
Male and female worms formed
↓
Sexual fusion and maturity occurs
↓
Flows through the portal vein in retrograde direction along the inferior mesenteric vein
↓
Vesical venous plexus
↓
Bladder wall and submucosa
↓
Ova are released
↓
Enters the bladder through the mucosa
↓
Ova are released into the urine and to water
↓
Ciliated miracidium
↓
Fresh water snail
↓
Death of snail releases thousands of cercariae.

Pathology in the Bladder

- Bilharzial pseudotubercles—earliest sign.
- Nodules.
- Sandy patches—calcified dead ova with degeneration of overlying epithelium.
- Granulomas.
- Ulceration and papilloma.
- Fibrosis and thimble bladder formation.
- In due course of time, development of *squamous cell carcinoma*.
- Other pathologies: Ureteral and urethral stricture, recurrent UTI, bladder calculi, urinary fistula.

Clinical Features

- Initially cutaneous lesions like utricaria develops lasting for few days.
- Then after a period of 4-8 weeks, fever, along with features of eosinophilia develops.
- Eventually, after many months it causes intermittent, painless, *terminal haematuria*.

Investigations

- Cystoscopy and biopsy.
- IVU.
- Cystography.
- Urine microscopy: Last few ml of early morning urine sample is collected to see for ova and RBCs.

Differential diagnosis

- Tuberculosis
- Recurrent cystitis
- Malignancy

Treatment

- Long-term *Praziquantel* or metrifonate.
- Surgery for thimble bladder—ileo-or caecocystoplasty.
- Cystoscopic diathermy fulgaration of papillomas.
- Radical cystectomy, if it is squamous cell carcinoma of the bladder.

THIMBLE OR SYSTOLIC BLADDER

It is inability of the bladder to relax and distend and so inability to retain the urine as required.

Causes

- Tuberculous cystitis
- Schistosoma haematobium
- Interstitial cystitis
- Radiotherapy
- Malignancy
- Previous bladder surgery

- Bladder is fibrotic and contracted, with difficulty in dilating and accommodating urine as needed.
- Bladder capacity is less than 100 ml (60 ml).

Clinical Features

- Features of decreased bladder capacity
- Frequency of urine
- Features of recurrent cystitis

Investigations

- Cystoscopy.
- Cystography.
- IVU.
- Urine C/S.
- Specific diagnostic tests.

Treatment

- The cause is treated.
- Hydrostatic dilatation.
- Augmentation of bladder by doing ileocystoplasty or *caecocystoplasty* or *sigmoidocystoplasty*.
- Steroids are also tried.

BLADDER TUMOURS

They are urothelial tumours. They are commonly malignant. Benign tumours are rare. Transitional cell carcinoma (TCC) is the most common.

1. Primary

a. ***Epithelial***
  - *Transitional cell carcinoma* (90%).
  - *Adenocarcinoma*, arising from urachal remnant or in exstrophy bladder or from glandular metaplasia (2%).
  - *Squamous cell carcinoma* originates from bilharzial infection (5%) or calculus.

b. ***Connective tissue tumour:***
  - Myoma, angioma, fibromas, sarcomas.
  - Extra-adrenal phaeochromocytoma.

2. Secondary

From adjacent organs like sigmoid colon, rectum, uterus, ovary, prostate.

TRANSITIONAL CELL CARCINOMA (TCC)

- It is the most common type of bladder tumour.
- It is the 4th most common nondermatological cancer in man.

Aetiology

3C's
  - Chemical carcinogens.
  - Cigarette smoking.
  - Cyclophosphamide.

- Chemical carcinogens are the main factor. 2-Naphthylamine, aminobiphenyl, benzidine, chloro-O-toluidine, chloro-aniline, other dyes.
- Occupation wise it is common in textile, dye, cable, tyre, petrol, leather workers, painters, chemical workers, sewage workers.
- Abnormal tryptophan metabolism can cause bladder cancer.
- Schistosoma haematobium; chronic irritation, etc.

Classification I

1. ***Nonmuscle invasive tumour without involving lamina propria*:** Has got excellent prognosis (70%).
2. ***Nonmuscle invasive tumour with involvement of lamina propria.***
3. ***Muscle invasive type*** (25%): Carries poor prognosis.
4. ***Carcinoma in situ*** *(flat noninvasive—5%)*:
   - Contains irregularly arranged cells with large nuclei, with high mitotic index, replacing normal urothelium.
   - This may occur alone—*Primary carcinoma in situ.*
   - It may occur in association with a new tumour—*Concomitant carcinoma in situ.*
   - It can occur in a patient who had a previous tumour—*Secondary carcinoma in situ.*
   - It has got high malignant potential with 50% mortality rate.
   - It was called earlier as 'malignant cystitis' as it causes severe dysuria, suprapubic pain and frequency (terminology not used presently).

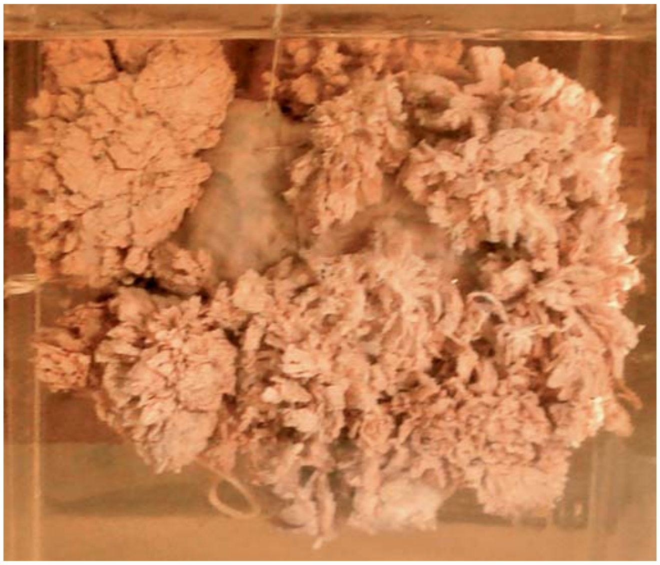

Fig. 26.103: Pathology specimen of bladder (tumour) showing multiple papillary tumours projecting into the lumen.

Classification II

a. ***Superficial bladder tumour***:
  - It may be *papillary*, pedunculated with narrow stalk, which is often multiple.
  - It may be *sessile* with a wide base, can be single or multiple, and has got tendency to invade the muscle earlier.
  - Mucosa in and around the tumour is oedematous, red, with dilated vessels, often with encrustations. It is 70% common. Less than 5% of superficial TCC develop metastatic carcinoma.

b. ***Muscle invasive TCC***:
- Almost always they are solid, sessile, with a broad base and with irregular ulcerated surface.

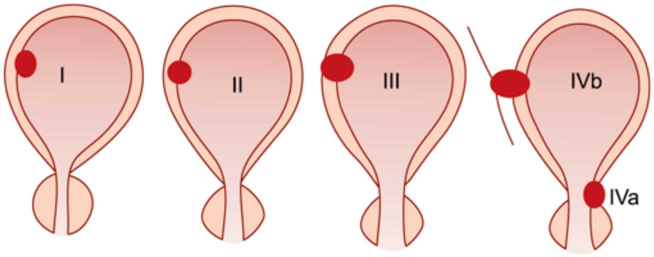

Fig. 26.104: Staging of bladder tumour.

- It may spread through lymphatics to pelvic lymph nodes or through blood to the lung, liver and bones.
- It has got poor prognosis.

c. ***Carcinoma in situ:***
- Bladder tumor can be papillary or villous type. It can be low grade or high grade.

Sites

- *Lateral wall—most common (35%).*
- Trigone—next common (32%).

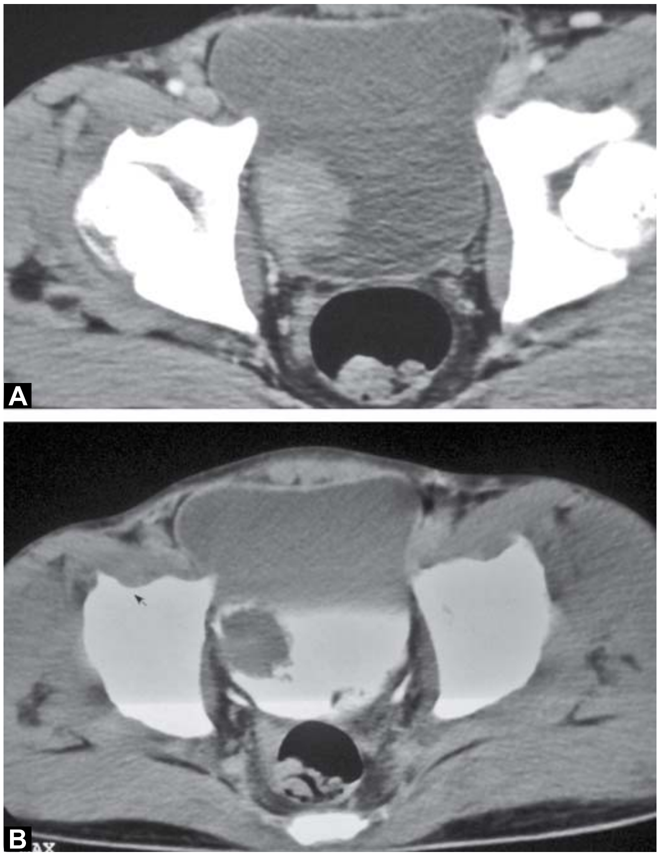

Figs 26.105A and B: CT scan showing bladder tumour.

Staging (Refer Fig. 26.104)

Jewett-Strong-Marshall staging

I Tumour confined to subepithelial connective tissue
II Muscle infiltration superficially
III Full thickness muscle and perivesical tissue infiltrated, but mobile
IV Fixed to adjacent organs
  IVa (prostate)
  IVb (pelvic wall)

Staging is done by bimanual palpation under G/A.

TNM staging (AJCC)

Tis – Carcinoma *in situ*
Ta – Noninvasive papillary tumour
T1 – Invading only lamina propria
T2 – Tumor invades muscularis propria
- p T2a – Invading the inner half of the muscle
- p T2b – Invading the outer half of the muscle

T3 – Invading the perivesical tissues
- p T3a – Microscopic invasion
- p T3b – Macroscopic invasion

T4 – Invasion into prostate/uterus/vagina/pelvic wall/abdominal wall
- T4a – Into prostate/uterus/vagina
- T4b – Into pelvic/abdominal wall

N0 – No nodes
N1 – Single regional nodal spread up to 2 cm in size
N2 – Spread to single node 2-5 cm in size or multiple nodes less than 5 cm
M0 – No distant spread
M1 – Distant spread present

Clinical Features

- Painless haematuria anaemia.
- Features of cystitis, with suprapubic pain, frequency, dysuria.
- Hydronephrosis can occur when tumour obstructs the ureteric orifice.
- Pain in groin, back, perineum, when tumour invades the pelvic wall.
- Common in males—3:1.

Investigations

- Urine microscopy: For RBC's and malignant cells.
- Blood: Hb%, blood urea, serum creatinine.
- IVU: Shows filling defect with distortion and often hydronephrosis.
- *Cystoscopy* and biopsy/resection.
- *Bimanual examination under G/A*—to stage the tumour.
- U/S abdomen to see bladder wall, pelvis, liver, lymph nodes.
- CT scan to evaluate the extension.
- MRI to see invasion and pelvic wall status.

Many receive advice, only wise profit from it.

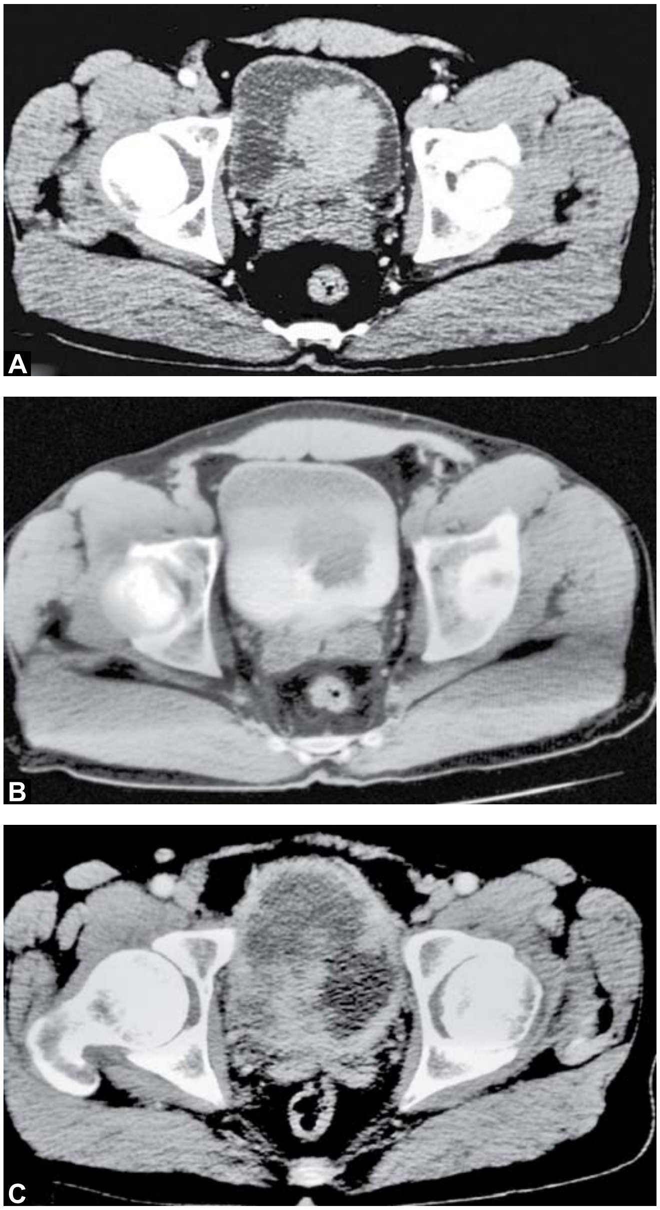

Figs 26.106A to C: CT scan picture showing irregularity in bladder mucosa—carcinoma urinary bladder.

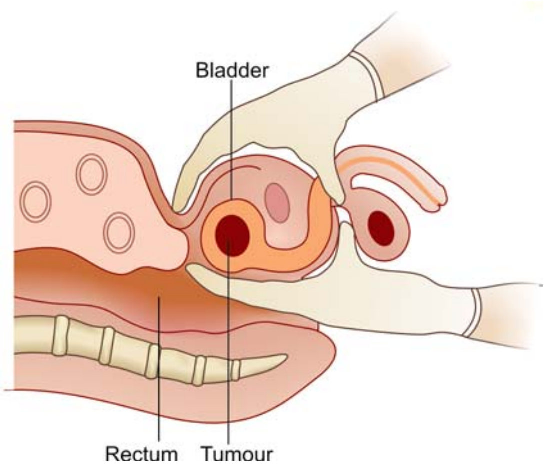

Fig. 26.107: Bimanual palpation is done for bladder tumour under general anaesthesia to evaluate the tumour.

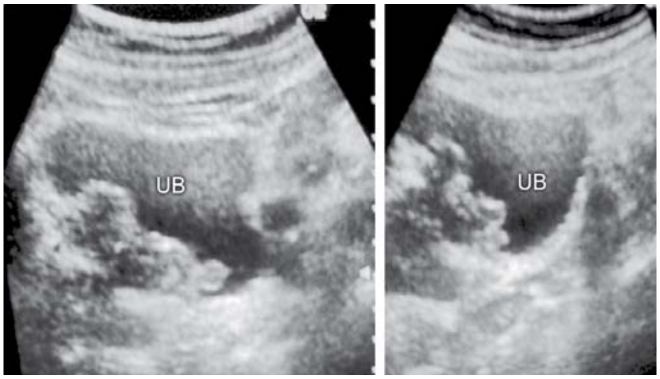

Fig. 26.108: Carcinoma urinary bladder—U/S picture.

Treatment

a. ***Noninvasive tumour***:

- *Endoscopic resection* of tumour.
- *Intravesical chemotherapy* using BCG (Dose: 120 mg in 150 ml of normal saline weekly for six weeks), mitomycin C, epirubicin, adriamycin, thiotepa, methotrexate can be given especially for carcinoma *in situ*.
  - BCG is very useful. Very rarely *BCG provocation* can occur. Fever, joint pain, granulomatous prostatitis, sinus formation, disseminated tuberculosis are the features. Dysuria, frequency and urinary irritation can occur. BCG is *live attenuated vaccine*; its advertent IV entry can be *life threatening* and so if there is haematuria its infusion should be postponed.
  - Mitomycin C causes skin desquamation and rash.
- *Systemic chemotherapy*: Using cisplatin, 5 FU, adriamycin, mitomycin.
- *Helmstein balloon degeneration* for large papillary tumour. Balloon is passed into the bladder and inflated so as to cause pressure necrosis of the summit of the tumour. Later remaining part of the tumour can be resected easily through cystoscopy.

b. ***Invasive bladder tumour*:**

- *Curative interstitial radiotherapy* using implantation of radioactive gold grains (Au 198) (half life is two and half days), or radioactive tantalum wires (Ta 182) (half life is 4 months).
- *Radical deep external beam radiotherapy* (45 Gy) using cobalt 60 is useful, as bladder is retained and so normal act of micturition and potency can be maintained. Complication is that it may eventually lead to formation of thimble bladder.
- ***Surgery***:

Indications for surgery

- Multiple bladder tumours
- Sessile tumours
- Recurrent tumours
- Poorly differentiated tumours
- *In situ* carcinomas
- Squamous cell carcinoma
- Adenocarcinomas

  - *Partial cystectomy* done when tumour is confined to fundus of the bladder and is single, with a margin of clearance of 2.5 cm. External beam RT and chemotherapy should be given.

  - *Radical cystectomy*: CT scan is a must before doing radical cystectomy to see the pelvis and lymph node status. Here urinary bladder, urethra, paravesical tissues, lymph nodes (pelvic) are removed. In females hysterectomy with removal of part of the vagina is done. After surgery urinary diversion is done either by doing continent ileal conduit, or ureterosigmoidostomy or by creating rectal urinary pouch.
- *Intravesical chemotherapy* by BCG, mitomycin C, adriamycin and interferons.
- *Systemic chemotherapy* by cisplatin, adriamycin, mitomycin, vinblastine—adjuvant therapy. Neoadjuvant cisplatin based chemotherapy is beneficial improving survival by 7%. MVAC regime is methotrexate; vinblastine; adriamycin and cisplatin.

Prognosis

- Depends on type, differentiation, location, stage, invasion, number, lymph node status, pelvis involvement and response to treatment.
- TCC alone—85%; TCC with SCC or adenocarcinoma—6%; SCC/adenocarcinoma alone—3%; spindle cell carcinoma—2%.

URETEROSIGMOIDOSTOMY

- It is a procedure done for urinary diversion after radical cystectomy or as a permanent diversion for many other causes, e.g. bladder dystrophy.
- *Left ureter is implanted to the sigmoid colon, right one to the upper rectum or rectosigmoid junction.*

Complications

- Because the pressure in sigmoid colon is more, it causes *pyelonephritis* and recurrent upper urinary tract infection.
- Because of diarrhoea it causes *hypokalaemia.*
- Because of reabsorption of chloride, *hyperchloraemic acidosis* occurs.

Treatment

- IV sodium bicarbonate (8.4%) with slow IV potassium.
- IV antibiotics.
- Patient is advised to take potassium citrate and sodium bicarbonate and to **avoid salt.**
- In many cases conversion to *ileal conduit* may be required.

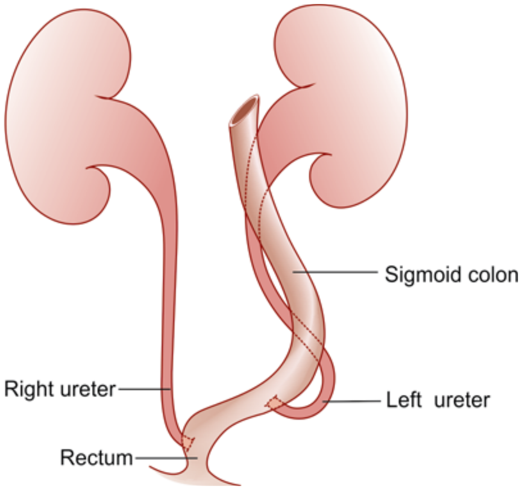

Fig. 26.109: Ureterosigmoidostomy.

RUPTURE BLADDER (BLADDER INJURY)

Causes

- Blow, kick or fall
- Road traffic accidents
- Stabs, gunshot injuries
- Endoscopic trauma
- Diathermy
- Instrumentations

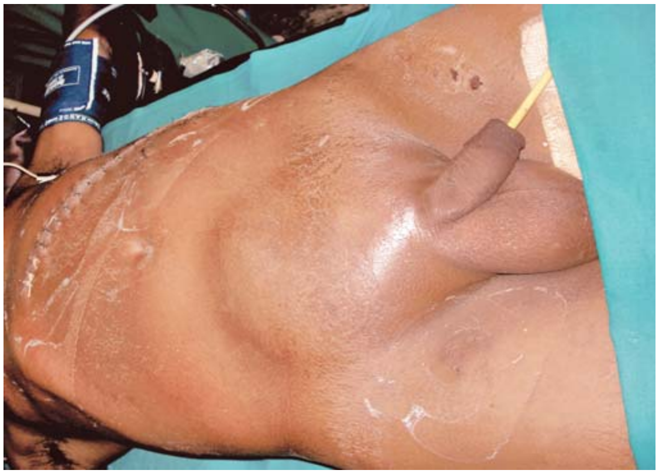

Fig. 26.110: Rupture urinary bladder causing extravasation and lower abdominal swelling. It needed surgical intervention, repair of bladder; suprapubic cystostomy.

Types

I. *Intraperitoneal rupture*—20% common.
Occurs in fully distended bladder due to blow, kick, or fall.

II. *Extraperitoneal rupture*—80% common.
Due to road traffic accidents, golf playing, fall over the manhole.
Its features and management are same as rupture of posterior (membranous) urethra.

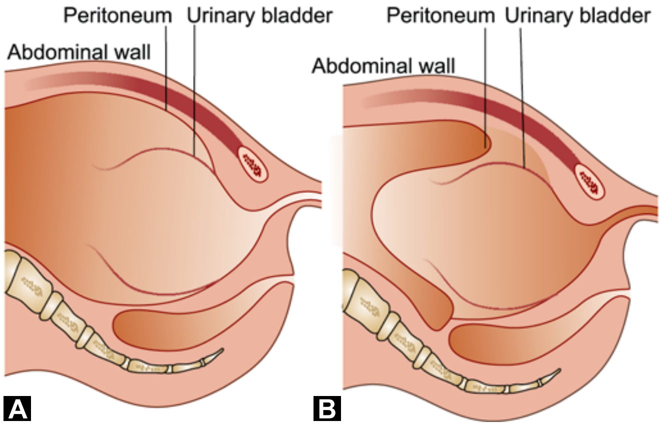

Figs 26.111A and B: (A) Intraperitoneal rupture. (B) Extraperitoneal rupture.

I. Intraperitoneal Rupture

- It is 20% common.
- It occurs in full bladder.

Clinical features

- Sudden pain in suprapubic region.
- Shock and syncope.
- Diffuse abdominal pain.
- Urine leaks into the abdominal cavity causing distension of abdomen.
- Later it causes features of peritonitis, with guarding, rigidity, tenderness and rebound tenderness, dull flank.
- *Patient does not have the desire to micturate.*

Investigations

- Plain X-ray shows ground glass appearance.
- Peritoneal tap is done to confirm urine.
- Cystogram: After passing a small catheter gently per urethra, water soluble iodine dye is passed to visualise the tear in the bladder and entry of the dye into the free peritoneal cavity. This can be done now through C-Arm image intensifier easily.
- U/S abdomen to look for other injuries in the abdomen.
- CT scan abdomen.

Treatment

- *Surgery is the only treatment for intraperitoneal rupture of the bladder*. Emergency laparotomy is done.
- Bladder tear is sutured in two layers using **vicryl**. Peritoneal wash is given.
- Malecot's catheter is placed from above as SPC.
- Prevesical space and peritoneal cavity are drained separately.
- Foley's catheter from below is also passed.
- Adequate specific antibiotics are given to prevent sepsis.

Complications of rupture bladder

- Cystitis and pyelonephritis
- Peritonitis
- Pelvic abscess
- Fistula formation (vesicovaginal or rectovesical)
- Paralytic ileus
- Haemorrhage

Without surgery mortality is 100%

II. Extraperitoneal Rupture of the Bladder

- It is 80% common.
- Occurs in road traffic accidents, in a nondistended bladder.

Clinical features

- There is collection of urine and blood in the extraperitoneal space in front, with fullness, diffuse pain and tenderness in lower abdomen.
- Swelling in the scrotum or labia, and abdominal wall.
- Strangury and inability to pass urine.
- Often blood in the external meatus is noted.
- Features of shock and of other associated injuries may be noted.

Investigations

- Plain X-ray pelvis shows fracture pelvis.
- Cystogram shows leak from the bladder.

During cystoscopy

- Bladder cannot be distended
- Endoscopy light may be shining through the abdominal wall
- Irrigating fluid cannot be retrieved back

Associated urethral injury is looked for

Treatment

- The bladder is exposed extraperitoneally, the tear is identified and sutured. The extraperitoneal space is irrigated with saline. Bladder is closed with a SPC using Malecot's catheter and a drain is placed in prevesical space (cave of Retzius).
- If there is urethral injury it should be treated accordingly.

Note:

- **Bladder injury can also occur** during hystrectomy (both abdominal and vaginal), surgery of colon or rectum, repair of direct inguinal or femoral hernias.

RESIDUAL URINE

- It is the amount of urine retained in the bladder after voiding (at the end of completion of the act of micturition).
- Normal value is 30 ml.
- Amount more than 50 ml is significant. It signifies obstruction in the urethra like BPH.
- Residual urine more than 200 ml in BPH indicates the need for surgical intervention.
- High residual urine is also seen in different types of neurogenic bladder.
- Residual urine precipitates infection because of stasis.

Residual urine is assessed by the following methods:

- The patient is asked to pass urine and then a red rubber catheter is passed to empty the retained urine which is measured to quantify.
- Ultrasound for evaluation of the bladder after voiding urine.
- Intravenous urogram (IVU)—postmicturition film.

MALAKOPLAKIA

- It is usually associated with chronic cystitis of unknown aetiology causing greyish raised patches in the bladder mucosa.
- Microscopically it shows infiltration of submucosa with lymphocytes, plasma cells and large multinucleated malakoplakia giant cells with concretions, called as ***Michaelis Gutmann bodies*** present both inside the giant cell as well as outside.
- It does not turn into malignancy.
- It presents as cystitis.
- Management is urine analysis; culture; antibiotics; urinary antiseptics like pyridium; cystoscopic evaluation and biopsy to rule out other causes.

NEUROGENIC BLADDER

It is the altered bladder function due to defects anywhere in the pathway of micturition reflex.

Causes

- Spinal cord trauma
- Disc prolapse
- Spina bifida
- Spinal cord root lesions
- Tabes dorsalis
- Diabetes mellitus
- Cerebral diseases
- Poliomyelitis

Types

Uninhibited Neurogenic Bladder

- It occurs in CNS diseases due to block in the corticospinal tract.
- There is increased frequency, urgency and incontinence.
- There is early desire to micturate with inability to hold urine.
- Bladder capacity is decreased.
- Awareness of filling and distension, and voiding pressure is normal.
- No residual urine.

Reflex Neurogenic Bladder (Automatic Bladder)

- Lesion here is above the level of micturition centre S_2, S_3 and S_4.
- There is no desire to micturate.
- No awareness of filling or distension.
- Bladder capacity is normal or increased.
- Voiding pressure varies.
- Residual urine present.
- Infection is common.

Autonomous Bladder

- Here micturition centre is destroyed (S_2, S_3, S_4). Both sensory and motor parts are destroyed.
- There is no awareness of filling.
- Bladder is incapable of contraction and so bladder capacity is increased.
- No voiding pressure.
- High residual urine and infection.
- Stress incontinence is present.
- Patient empties the urine by manual pressure.

Sensory Paralytic Bladder

- Here sensation from the bladder to micturition centre is selectively affected.
- It is common in tabes dorsalis and pernicious anaemia.
- There is complete loss of awareness of filling.
- Varying voiding pressure.
- Increased bladder capacity and residual urine.
- Infection is common.

Motor Paralytic Bladder

- It is due to selective destruction of motor pathway from the micturition centre to the bladder.
- It is seen commonly in polio, tumours, trauma.
- There is normal awareness of filling and painful distension.
- Voiding pressure is zero. It is atonic bladder.
- There is residual urine with normal bladder capacity.

Investigations

- Spine X-ray.
- Micturating cystogram.
- Urodynamic studies.
- U/S abdomen to find out residual urine.
- Urine culture and microscopy.

Treatment

- The cause is treated.
- Condom drainage.
- Intermittent catheterisation.
- Drugs like carbachol, distigmine can be given.
- Alpha adrenergic agents.
- Physiotherapy and pelvic exercises.
- Electrical stimulation of perineal muscles.
- Treating the stress incontinence surgically.
- Antibiotics whenever there is infection.

VESICOURETERIC REFLUX

- Normal flow of urine is always from ureter to bladder through ureteric orifice. It is called as efflux.
- If urine enters the ureter from the bladder through ureteric orifice, it is called as reflux. It is always *pathological.*

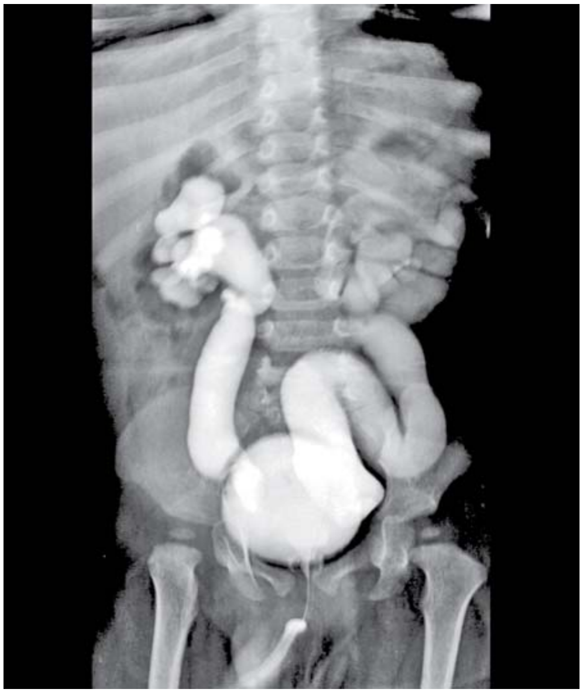

Fig. 26.112: Micturating cystourethrogram (MCU) showing gross vesicoureteric reflux on both sides.

Aging is “immobility, incontinence, instability and intellectual deterioration.

Causes

- *Congenital:* It is commonly associated with posterior urethral valve.
- *Acquired:* Trauma, after surgery or intervention.
  - Reflux has got **5 grades** depending on the amount of reflux (for grades refer back—MCU).
  - Reflux can be unilateral or bilateral.

Clinical Features

- Features of recurrent urinary tract infection.
- Renal failure in bilateral cases.
- Palpable kidney mass in the loin.

Investigations

- Urine microscopy and C/S.
- Blood urea and serum creatinine.
- IVU.
- Micturating cystourethrogram (MCU).

Treatment

- Tailoring of the ureter with ureteric re-implantation.
- Antibiotics.
- In severe cases nephrectomy, renal transplantation are required.

BLADDER DIVERTICULA

Normal intravesical pressure is 30-50 cm of water. Pressure reaches as high as 150 cm of water in obstruction. This causes hypertrophy of bladder mucosa → saccule formation → diverticulum.

Diverticula can be:

- *Acquired*—common
  - Pulsion diverticulum occurs due to increased intravesical pressure
  - Traction diverticula occurs due to traction into a hernial sac in the groin—often as a sliding hernia
- *Congenital*—due to remnant urachus in midline

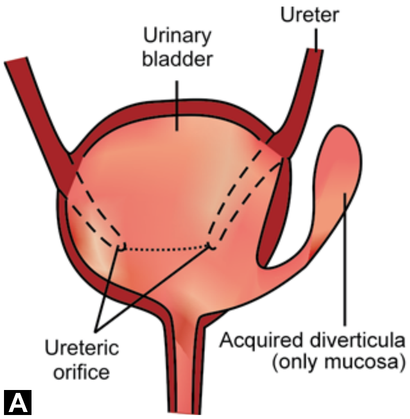

Fig. 26.113A

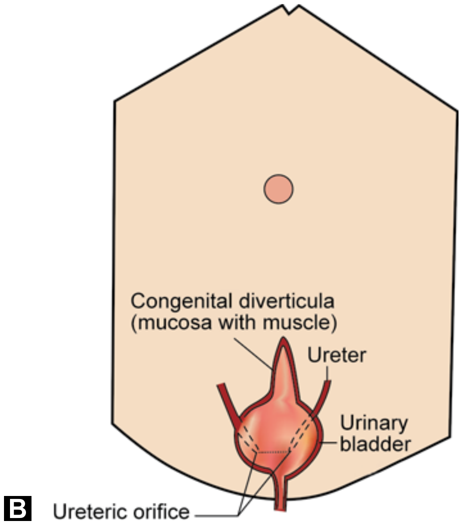

Fig. 26.113B

Figs 26.113A and B: Diverticula bladder. In acquired type only mucosal protrusion is seen in paraureteric orifice region. In congenital type it is full thickness protrusion, commonly over the dome of the bladder.

1. *Congenital diverticula:*
   - It occurs over the dome of the bladder.
   - It is due to unobliterated end of the urachus.
   - As it contains muscle in the wall and it contracts during micturition.
   - Rare type.
2. *Acquired diverticula*:

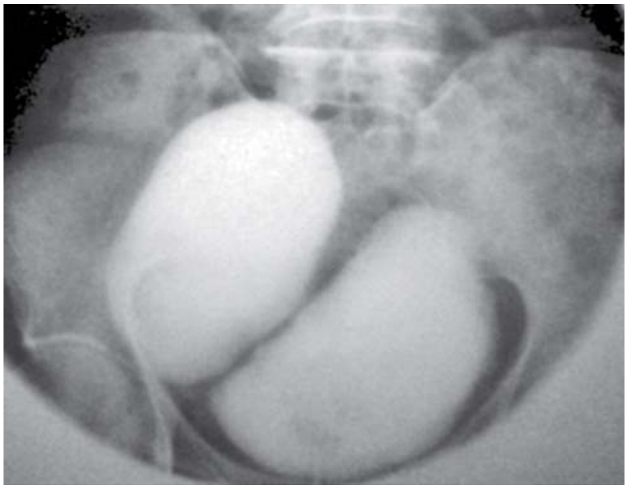

Fig. 26.114: X-ray showing bladder diverticula. It can be pulsion or traction diverticula. It can be congenital or acquired. Acquired occurs in paraureteric region. Congenital occurs in midline over the dome of the urinary bladder.

- They are more common.
- It occurs as a result of chronic urinary obstruction like BPH.
- It occurs adjacent to ureteric orifice and contains only mucosa (above and outer to ureteric orifice).
- It is prone for stasis, sepsis, stone formation, metaplasia and malignancy.
- Common in males (95%). Common in adults of 50 years.
- Haematuria (30%); fever; pyelonephritis.
- Micturition twice in rapid succession to empty the diverticular urine—is often typical.

Complications of bladder diverticula

- Recurrent infection
- Bladder stone formation
- Obstructive uropathy
- Malignant transformation—less than 5%

Investigation

- Urine C/S, cystoscopy, cystogram, IVU, U/S abdomen.
- CT scan abdomen.

Treatment

- Removal of the diverticula through open surgery.
- Approach is combined intravesical and extravesical. Through cystoscopy, ureteric stent is placed. Suprapubic transverse extraperitoneal open approach is used. Anterior wall of bladder is vertically opened in midline. Diverticulum is packed with gauze. Junction between urinary bladder and, diverticulum is dissected from outside; transected and diverticulum is excised. Bladder surface is closed with vicryl 2 zero single layer sutures. Midline vesicotomy is closed with a Malecot's catheter. Retropubic space is drained. Indwelling urinary catheter is kept *in situ* for 7-10 days.

URINARY DIVERSION

- It is diversion of urine temporarily or permanently, proximal to the site of obstruction.
  - Temporary type is done in benign conditions and distal obstructions, to promote healing.
  - Permanent type is done in advanced malignancies which block the urine flow distally, after radical cystectomy, after radiotherapy with frozen pelvis, etc.

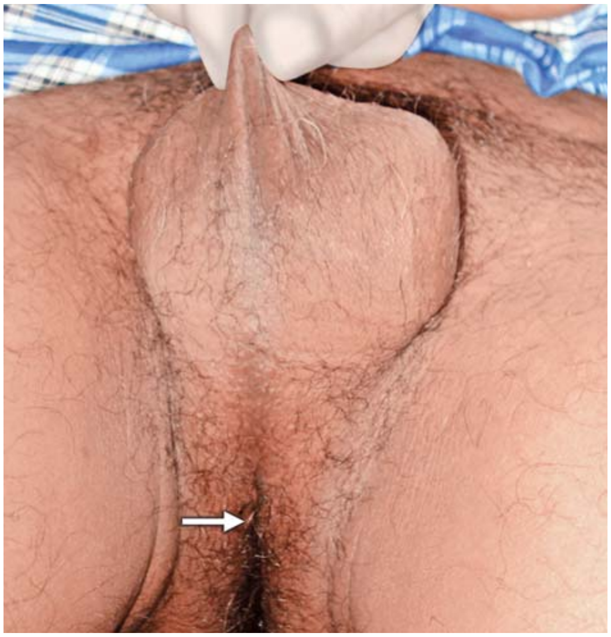

Fig. 26.115: Perineal urethrostomy after total amputation of penis.

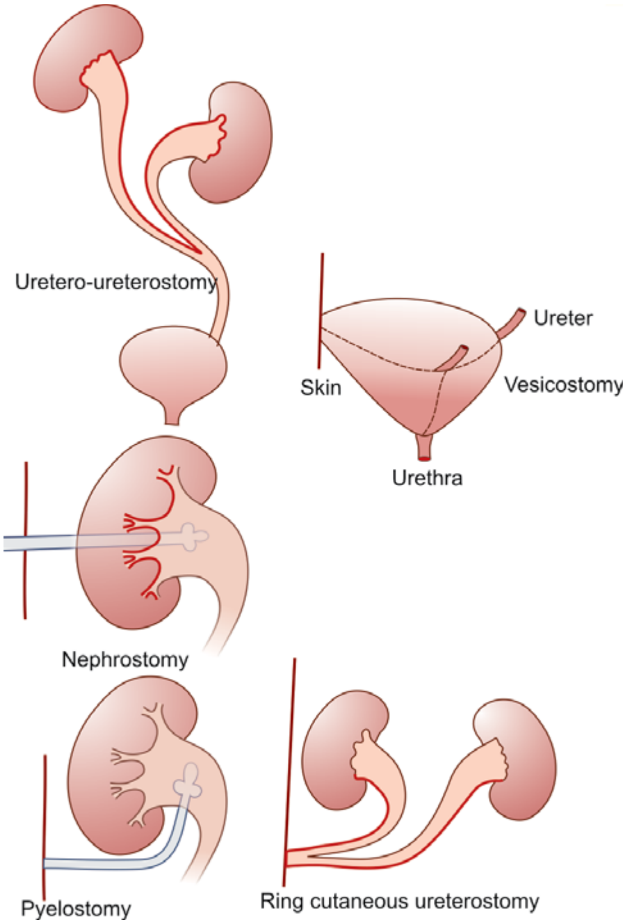
Uretero-ureterostomy
Ureter
Skin
Vesicostomy
Urethra
Nephrostomy
Pyelostomy
Ring cutaneous ureterostomy

Fig. 26.116: Types of urinary diversions.

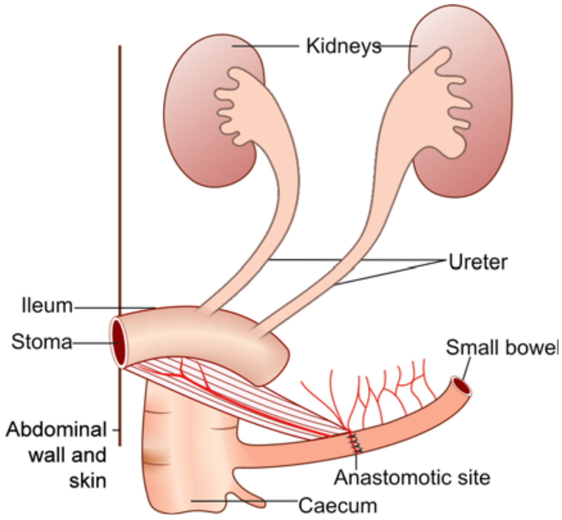
Kidneys
Ureter
Ileum
Stoma
Small bowel
Abdominal wall and skin
Anastomotic site
Caecum

Fig. 26.117: Ileal conduit as urinary diversion.

Temporary Diversions

- Suprapubic cystostomy (SPC).
- Vesicostomy in children in ectopiae vesicae, epispadias, urethral anomalies.
- Ureteric J stent—pigtail stents kept for 3 months.
- Urethrostomies.

A man who does nothing never has time to do any thing.

- Nephrostomies in pyonephrosis and high obstructive uropathies.
- Pyelostomy.

Permanent Diversions

External Diversions

- *Noncontinent ileal conduit*
  Here 20 cm of vascularised distal ileum (30 cm from IC valve) is prepared. Ureters are re-implanted over the ileum. Other end of ileum is brought out as a 3 cm stoma in right iliac fossa. Patient has to wear a bag over it.
- *Continent ileal conduit* is similar conduit with creation of specialised valve (Koch's continent) within, using ileal mucosa or appendix. Patient has to empty the conduit by passing self catheters once in 6 hours.
- Cutaneous ureterostomy/cutaneous ring ureterostomy.

Internal Diversions

- Ureterosigmoidostomy (*Leadbetter-Palitano*).
- Rectal bladder creation.
- Ileocystoplasty.
- Caecocystoplasty/sigmoidocystoplasty.
- Creation of neobladder using ileum/caecum/sigmoid colon which is sutured to urethra.

Types of urinary diversions

- Nephrostomy
- Pyelostomy
- Cutaneous ureterostomy
- Cutaneous ring ureterostomy
- Suprapubic cystostomy
- Vesicostomy in children
- Urethrostomy
- Ureterosigmoidostomy
- Rectal bladder
- Urinary ileal conduit (Koch's conduit)

Complications of Diversions

- Urine leak, faecal leak, intestinal obstruction, sepsis.
- Hyperchloraemic hypokalaemic metabolic acidosis.
- Pyelonephritis.
- Osteomalacia, growth retardation.
- Calculi formation—4%.
- Continent failure, blockage of diversion stoma (stenosis).
- Reflux of urine.
- Vitamin B_{12} and folate deficiency due to absence of ileum.
- Complications of stoma.
- Renal failure.
- Metaplasia and adenocarcinoma especially at uretero sigmoidostomy junction.

URINARY FISTULAS

It is the leak of urine at abnormal sites.

Causes

Congenital

- Patent urachus.
- Ectopiae vesicae.
- In association with anorectal malformation—rectovesical fistula.

Acquired

- Trauma to perineum.
- Pelvic surgery.
- *Vesicovaginal fistula*—due to obstructed labour; after hysterectomy, bladder may get injured while pushing it down; anterior colporrhaphy in elderly after vaginal hysterectomy; radiation induced; infiltrating carcinoma cervix. Usually it takes one week to cause necrosis and present as fistula. Features are—passage of urine per vagina, dribbling of urine in vagina on per speculum examination, *swab test is positive* (Methylene blue is injected through urethra after placing a swab in the vagina. Swab turns blue if there is fistula). Treatment is difficult. Surgical exploration—transvaginal repair in low fistula. Suprapubic repair if fistula is high. Often urinary diversion is needed.
- Specific causes like tuberculosis.
- Staghorn calculi can cause fistula after nephrostomy.
- Fistula can develop after surgery of renal pelvis, ureter, bladder. It will persist if there is distal obstruction like stricture urethra, otherwise it will subside spontaneously.
- Renal pelvis involved by Crohn's disease.

Investigations

- Discharge study for creatinine level.
- Urine C/S; leaking fluid for C/S.
- U/S abdomen; CT scan.
- MR fistulogram.

Treatment

- Correction of cause.
- Urinary diversion in initial period.
- Excision of fistula track and repair.
- Fistula can recur based on cause, nutrition, local blood supply, specific causes.

C. Prostate

CHAPTER OUTLINE

- Anatomy
- Acid Phosphatase
- Prostate Specific Antigen
- Benign Prostatic Hyperplasia
- Prostatitis
- Bladder Outlet Obstruction
- Carcinoma Prostate

ANATOMY

It is an accessory gland of male reproductive system. It is composed of glandular tissue embedded in fibromuscular stroma. It surrounds the first 3 cm of the urethra.

Lobes

- It is composed of 5 lobes—anterior, posterior, 2 lateral, 1 middle/median lobe.
- Primary carcinoma is said to begin in posterior lobe. Middle lobe produces an elevation in the lower part of the trigone of bladder *(uvula vesicae*). Adenoma is more common here due to more amount of glandular tissue.

Capsules

1. *True capsule*: Formed by the condensation of the peripheral part of the gland; continues with the stroma of the gland.
2. *False capsule*: Derived from the pelvic fascia. On each side prostatic venous plexus are embedded in it.

McNeal Divided Prostate into Three Zones

- *Peripheral zone*—prone for carcinoma.
- *Periurethral transition zone* where BPH arises.
- *Central zone*.

Note:
- Prostate carcinoma develops in prostatic gland proper.
- BPH develops in submucosal glands.

ACID PHOSPHATASE

- It is the enzyme that splits organic phosphates.
- It is found in many human tissues, but more concentrated in prostate.
- It is active at pH 5.
- Acid phosphatase secreted by prostate drains into the urethra through prostatic ducts and so blood levels of this enzyme remain low.
- Serum acid phosphatase estimation should be done ***on empty stomach*** because heavy meals alter the level of the acid phosphatase.
- Normal value is *0-5 King Armstrong units per 100 ml of serum*.

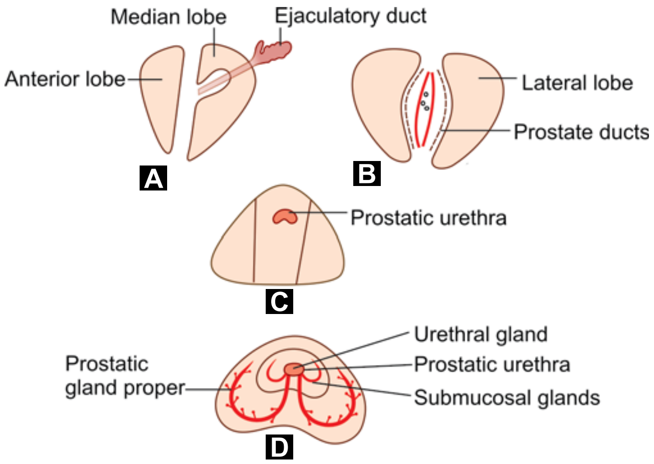

Figs 26.118A to D: Parts of prostste.

- It is raised significantly in carcinoma prostate with metastases.
- *It does not increase in BPH.*
- Slight increase in acid phosphatase level occurs in acute prostatitis, Paget's disease of bone and hepatic cirrhosis.
- Prostatic fraction of acid phosphatase is more relevant in carcinoma prostate.

PROSTATE SPECIFIC ANTIGEN (PSA)

- It is a protease, produced from the prostatic epithelium secreted in the semen to cleave and liquefy the seminal coagulum formed after ejaculation.
- PSA is organ specific. *Normal value is 4 ng/ml of plasma. More than 10 ng/ml is significant*.
- PSA elevation occurs not only in carcinoma but also in prostatic hyperplasia and prostatitis.
- But the increase is much more in carcinoma than in benign conditions.
- PSA is in two forms: *Major,* bound form and *minor,* free form.
- Major, bound form increases in carcinoma. Minor, free form is increased in benign conditions.
- PSA density, i.e. PSA level per gram of prostate tissue is more relevant.
- Serial estimation of PSA is very useful to suspect spread and recurrence after treatment.

The formula to be complete happy is to be very busy.

- 25% of men with PSA 4-10 ng/ml show prostate carcinoma.
- 20% of men with normal PSA (1-4 ng/ml) will show prostate carcinoma.
- PSA more than 10 ng/ml is suggestive of carcinoma prostate.
- PSA more than 35 ng/ml is almost diagnostic of advanced carcinoma of prostate.
- Decrease in PSA after therapy suggests adequate ablation.
- Men aged > 50 years with PSA ≥ 3 ng/ml should undergo prostatic biopsy.

Prostatic calculi:
It is the calcification of corpora amylaceae which contains calcium phosphate, protein and fat.

BENIGN PROSTATIC HYPERPLASIA (BPH)

> *He was very often, both in the Day and the Night, forced to make Water, seldom in any Quantity, because he could not retain it long enough.*
>
> —**Edward Hyde (first Earl of Clarendon),** 1759

It is benign enlargement of prostate which occurs after 50 years, usually between 60 and 70 years.

Aetiology

Theories

- It is involuntary hyperplasia due to disturbance of the ratio and quantity of circulating androgens and estrogens.
- BPH is a benign neoplasm, also called as ***fibromyoadenoma.***
- Hypothalamus → pulsatile release of LHRH → release of luteinising hormone (LH) from anterior pituitary → stimulates Leydig cells of testes → releases testosterone (TS) → reaches prostate → releases 5α reductase type II of prostate → converts TS to DHT (dihydrotestosterone) for its effects. DHT is five times more potent than TS. 90% TS is from testes. Remaining 10% TS is from adrenal cortex.
- With age TS level drops slowly. But fall of oestrogen level is not equal. So prostate enlarges through intermediate peptide growth factor.
- BPH arises from submucosal glands of periurethral transitional zone with stromal proliferation and adenosis. It eventually compresses the peripheral zone and enlarges as lateral lobe.
- BPH arising from subcervical glands of central zone enlarges as middle lobe projecting up into the bladder.

Pathology

- BPH usually involves median and lateral lobes or one of them.
- It involves adenomatous zone of prostate, i.e. submucosal glands.
- Median lobe enlarges into the bladder.
- Lateral lobes narrow the urethra causing obstruction.
- *Urethra above* the verumontanum gets elongated and narrowed.
- *Bladder* initially takes the pressure burden causing *trabeculations, sacculations* and later *diverticula* formation.
- Enlarged prostate compresses the prostatic venous plexus causing congestion, called as *vesical piles leading to haematuria.*

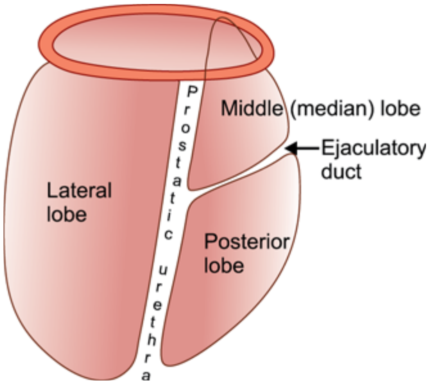

Fig. 26.119: Prostate anatomy.

- Incrimination of BPH as the source of haematuria before excluding other causes is termed as ***"Decoy prostate"***.
- *Kidney and ureter*: Backpressure causes *hydroureter and hydronephrosis*.
- Secondary ascending infection can cause *acute* or *chronic pyelonephritis*.
- Often severe obstruction can lead to obstructive uropathy with *renal failure.*
- BPH causes *impotence.*

Clinical Features

- Frequency occurs due to introversion of sensitive urethral mucosa into the bladder or due to cystitis and urethritis.

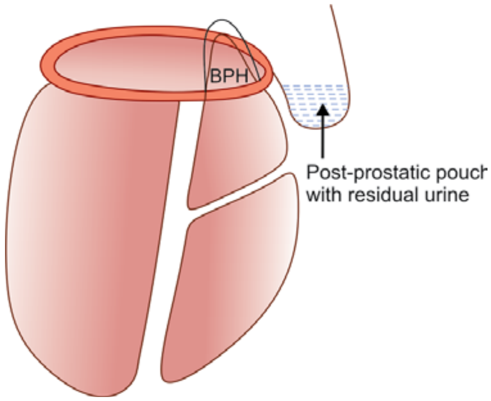

Fig. 26.120: Retention of urine in bladder due to BPH.

- Urgency, hesitancy, nocturia.
- Overflow and terminal dribbling.
- Difficulty in micturition with weak stream and dribble.
- Pain in suprapubic region and in loin due to cystitis and hydronephrosis respectively.
- Acute retention of urine.
- Retention with overflow. High pressure chronic retention with functional obstruction.
- Impaired bladder emptying with its problems like cystitis, urethritis, stone formation and residual urine.
- Haematuria.
- Renal failure.
- *Prostatism* is a combination of symptoms like frequency both at day and night, poor stream, delay in starting and difficulty in micturition.

- Tenderness in suprapubic region, with palpable enlarged bladder due to chronic retention. Hydronephrotic kidney may be palpable.
- **Per rectal examination shows enlarged prostate. It should be done when bladder is empty.**
- Features of urinary infection like fever, chills, burning micturition.
- International prostate symptom score is available now.

Lower urinary tract symptoms (LUTS)

Symptoms of voiding	*Symptoms of storage*
• Hesitancy • Poor flow not improving by straining • Dribbling even after micturition • Intermittent stream—stops and starts • Poor bladder emptying • Episodes of near retention	• Frequency • Nocturia • Urgency • Urge incontinence • Nocturnal incontinence

Differential Diagnosis

- Stricture urethra.
- Bladder tumour, carcinoma prostate.
- Neurological causes of retention of urine like diabetes, tabes, disseminated sclerosis, Parkinson's disease.
- Idiopathic detrusor activity.
- Bladder neck stenosis; bladder neck hypertrophy.

Investigations

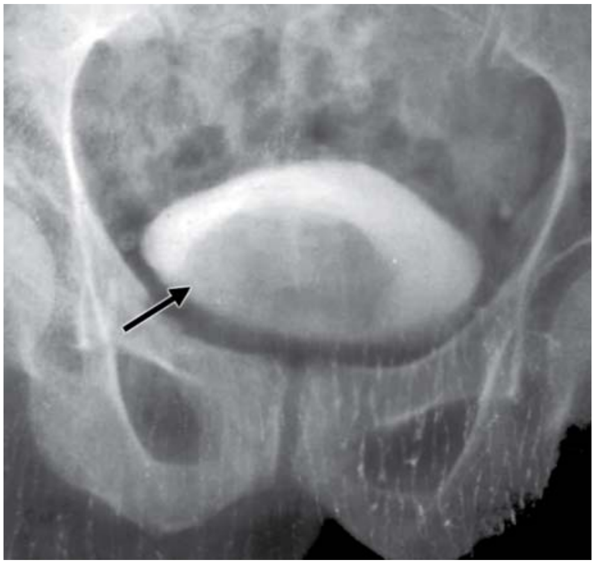

Fig. 26.121: IVU showing bladder with enlarged median lobe of prostate.

- Urine for microscopy and C/S.
- Blood urea and serum creatinine.
- U/S abdomen—look for presence of residual urine.
- Urodynamics.
    - Urine flow rate > 15 ml/sec is normal. 10-15 ml is equivocal; < 10 ml is low.
    - Voiding pressure < 60 cm of water is normal; 60-80 is equivocal; > 80 is high.
- Cystoscopy.
- Transrectal US (TRUS) is useful to find out nodules/possibility of carcinoma prostate. It is not done routinely.
- Acid phosphatase.
- Prostate specific antigen (PSA).
- IVU—to see kidney function.
- Serum electrolytes.

Note:
Normal peak urine flow rate is 20 ml/sec. In obstruction it is less than 10 ml/sec.

Management

- Patient with ***acute retention*** of urine requires *urethral catheterisation.*
- If urethral catheterisation fails, then *suprapubic cystostomy (SPC)* is done.
- If patient presents with ***uraemia***, then urethral catheterisation is a must. That allows the kidney to function adequately and further obstructive damage is prevented.
- Serum electrolytes should be corrected properly in these patients.

Indications for Surgery

- Prostatism (frequency, dysuria, urgency).
- Acute retention of urine.
- Chronic retention of urine with residual urine more than 200 ml.
- Complications like hydroureter, hydronephrosis, stone formation, recurrent infection, bladder changes.
- Haematuria.

Surgeries

1. *Transurethral resection of prostate (TURP)*:

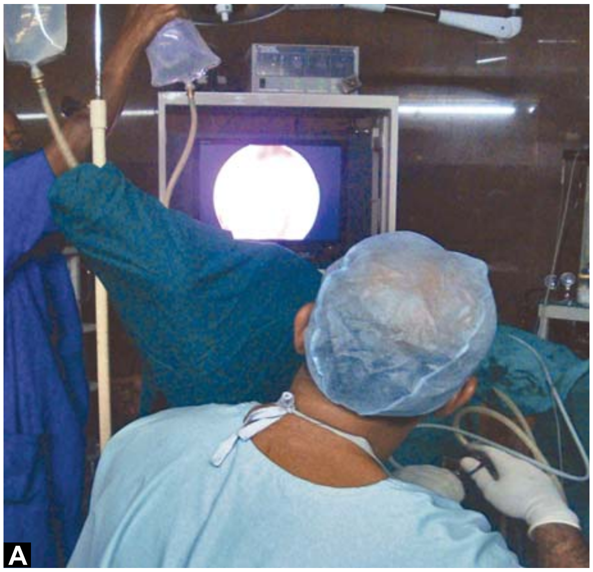

Fig. 26.122A

In cases of complete intrapelvic rupture of the urethra, the tip of the catheter can be felt per rectum protruding from the urethral tear as a longitudinal cord lying beneath the anal canal. **—Wallace**

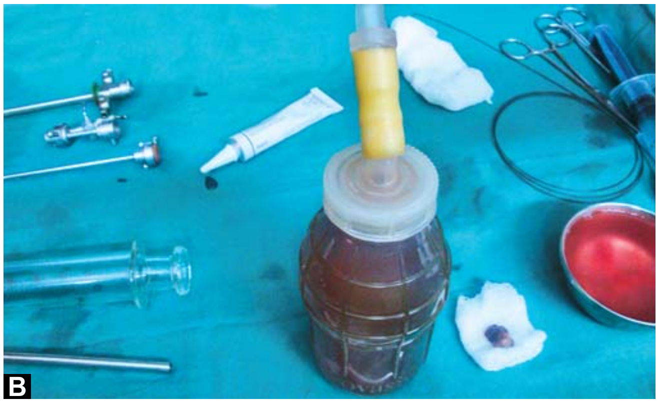

Fig. 26.122B

Figs 26.122A and B: Endotherapy for BPH—TURP is the most common procedure; laser (holmium), stenting—are different procedures. TURP set is shown on the table with evacuator to remove fragmented prostate tissue.

- Using cystoscope with fluid like glycine irrigating continuously, enlarged prostate is identified and resected using a loop with a hand control. Resection is done using high frequency diathermy current, above the level of verumontanum.
- After surgery, continuous bladder irrigation using normal saline is done using three way Foley's catheter. Antibiotics should be given. Catheter is removed within 72 hours.

Postoperative complications are:

- Water intoxication with congestive cardiac failure—TURP syndrome
- Hyponatraemia
- Haemorrhage
- Infection
- Incontinence
- Perforation of the bladder or prostatic capsule
- Stricture urethra
- Retrograde ejaculation and impotence
- Recurrence

TURP

- No suprapubic incision
- Most common and popular method
- Done using resectoscope
- Recurrence may occur in large prostate
- TURP syndrome can develop
- Continuous postoperative irrigation using glycine solution is needed for 72 hours
- Faster recovery and early discharge is possible

2. *Freyer's suprapubic transvesical prostatectomy*. Before TURP, it was a popular and was the procedure of choice for enlarged prostate.

   Complications are haemorrhage, infection, stricture urethra, incontinence, impotence, bladder neck contracture.
3. *Millin's retropubic prostatectomy*. It is done without opening the bladder. (It is not commonly practiced).
4. *Young's perineal prostatectomy* through perineal approach.
5. Microwave hyperthermia treatment with temperature of 45–50 degree.
6. Laser treatment is becoming popular, using holmium laser.
7. High energy electromagnetic treatment.
8. Placement of intraurethral stents at prostatic urethra.
9. Placement of extraurethral stents which are inert.
10. Transurethral balloon dilatation of the prostate.
11. If patient presents with acute retention of urine, initial urethral catheterization is done. If not possible then *suprapubic trocal cystostomy* ideally or formal open SPC is done. Once patient's obstructive uropathy is under control, TURP is done after 7-14 days after evaluation.

Specific Problems after Surgical Intervention of Prostate

- Retrograde ejaculation—65%.
- Erectile dysfunction—5%.
- Failure, recurrence of symptoms and enlargement—10%.
- Need for re TURP/surgery in 10 years—15%.
- Severe sepsis—6%.
- Recurrent late urinary infection—20%.
- Postoperative haematuria—3% needs transfusion.
- Mortality in TURP is 0.5%.

Other methods used for BPH

- Transurethral incision of prostate—TUIP
- Laser ablation
- Transurethral microwave therapy—TUMT
- Transurethral needle ablation using high frequency radiowaves—TUNA
- High intensity US energy
- Water induced thermotherapy—through a balloon placed in prostatic urethra
- Prosthetic stents
- Balloon dilatation

Drugs Used for BPH

- *Alpha 1 adrenergic blocking agents*—which inhibit smooth muscle contraction of prostate. They reduce the bladder neck resistance so as to improve the urine flow.
  - Short acting drugs are *prazosin* and *indoramin.*
  - Long acting drugs are *terazocin* and *doxazosin*.
  - Selective alpha$_{1A}$—adrenoceptor blocking agent: *Tamsulosin* 0.2 to 0.4 mg OD for 12 weeks.
- *5-alpha reductase inhibitor* inhibits conversion of testosterone to dihydrotestosterone.
  - It is effective in palpably enlarged prostate.
  - Drug used is *finasteride* 5 mg daily. Duration of treatment is about 6-8 months.
  - It is contraindicated in obstructive uropathy or carcinoma prostate.

PROSTATITIS

Types

Acute or chronic.

Acute Prostatitis

Causes:

- Due to instrumentation.
- Ascending infection from below.
- Haematogenous.
- Descending infection from above.

Bacteria involved:

- *E. coli, Klebsiella, Proteus.*
- *Staphylococcus.*
- *Streptococcus faecalis.*
- *Gonococcus.*

Clinical features:

- Pain, frequency, fever with chills and rigors.
- Retention of urine.
- **Perineal heaviness, pain on defaecation.**
- Tender prostate on per rectal examination.
- Initial fraction of urine is turbid which is sent for culture and sensitivity.

Investigations:

- Urine C/S.
- U/S abdomen.

Treatment:

- Prolonged rigorous antibiotics—for 2 months.
- Avoidance of alcohol and sexual intercourse for 6 weeks.

Complications

- Formation of single or multiple abscess. It is common in diabetics
- Chronic prostatitis
- Retention of urine

Prostatic abscess has to be drained—through perineal approach under U/S guidance

Chronic Prostatitis

Caused by *E. coli*, *Staphylococcus, Streptococcus, Trichomonas, Chlamydia.*

- There is always associated posterior urethritis.
- Epididymitis.
- Pain in the perineum, rectum, low back pain, leg pain.
- Fever.
- Sexual dysfunction.
- Per rectal examination shows tender prostate.
- Prostatic fluid obtained by prostatic massage shows 15 or more pus cells/HPF.
- In three glass urine test—first glass contains prostatic threads.

Treatment: Antibiotics—Co-trimethoxazole, trimethoprim, doxycycline.

Prostatic Abscess

- It is infection, suppuration and pus formation in the prostate gland.
- Presentation is fever, rigors, perineal pain, urinary disturbances, and tender soft fluctuant swelling in the prostate on rectal examination.
- Often presentation may be retention of urine.
- Total count will be increased.
- Urine will show pus cells.
- US is diagnostic. US is often done over perineum also.
- Treatment is antibiotics; US guided aspiration transperineally in lithotomy position or transperineal incision and drainage.
- Suprapubic cystostomy is better in case of retention of urine.
- After drainage antibiotics are needed for longer period of 6 weeks to prevent recurrent infection.

BLADDER OUTLET OBSTRUCTION (BOO)

- It is low urinary flow rate with the presence of high voiding pressure. It is an urodynamically confirmed entity.
- It is diagnosed by urodynamic pressure flow study.
- Flow rate will be les than 10 ml/second with voiding pressure more than 80 cm of water.
- Eventually detrusor inefficiency occurs causing significant residual urine.
- *Causes of BOO*—BPH; bladder neck hypertrophy or stenosis; carcinoma of prostate; urethral stricture; functional bladder neck obstruction.
- *Effects of BOO*—acute retention of urine; chronic retention of urine; impaired bladder emptying; uraemia; infection; stone formation, haematuria.
- US; renal function tests; IVU; PSA are the investigations.
- *Management* is by treating the cause by cystoscopic bladder neck incision, urethrotomy; TURP, etc.

Evaluation of patient with lower urinary tract infection (LUTI)

- Urine analysis and culture
- Hemoglobin, total count, ESR, blood urea and serum creatinine
- Digital examination of rectum (P/R)
- CNS examination
- PSA, acid phosphatase
- Uroflowmetry, flow rate assessment, residual urine analysis
- US abdomen, IVU to see upper urinary tract, CT abdomen in selected patients only
- Cystoscopy, transrectal US (TRUS)
- Prostatic abscess aspiration/drainage through perineal approach under guidance

Do all the good you can, by all the means you can, in all the ways you can, in all the places you can, at all the times you can, to all the people you can, as long as ever you can—

CARCINOMA PROSTATE

> *During the last six months, (I.B.) had suffered …excruciating pain in the region of the kidneys and bladder, attended with almost constant desire to void urine, which was effected with the greatest difficulty.…An examination per rectum proved that there existed an enlarged…prostate gland, and slight pressure occasioned great pain.…The bladder was found (at autopsy) to contain a tumour as big as a large orange…discovered to derive its origin from the prostate gland.… The fungus …plugged up both ureters.… In the liver there were several tumours (and) several…in the lungs.*
>
> —**George Langstaff,** 1817

- It is the *most common malignant tumour* in men over 65 years.
- Carcinoma prostate occurs in peripheral zone in prostatic gland proper, i.e. *commonly in posterior lobe*. So prostatectomy for BPH does not confer protection against development of carcinoma prostate.
- Incidence of prostate cancer in men over 80 years is 70%.

Types of carcinoma prostate

- Microscopically latent
- Tumours incidentally found either by TURP or by PSA estimation
- Early localised carcinoma
- Advanced local prostatic carcinoma
- Metastatic carcinoma either into the bone commonly or other organs

Histology: It is an *adenocarcinoma,* wherein there is loss of myoepithelial cell layer which normally surrounds the prostatic glands (Gleason). Glands here appear in confluence. *Grading of carcinoma* is based on dedifferentiation as proposed by Gleason.

Staging of carcinoma prostate

Occult—Diagnosed after investigation due to suspicion
Stage I—Tumour confined to prostate/local nodule
Stage II—Tumour involving capsule or diffuse type
Stage III—Tumour involving seminal vesicle
Stage IV—Extension into adjacent tissue

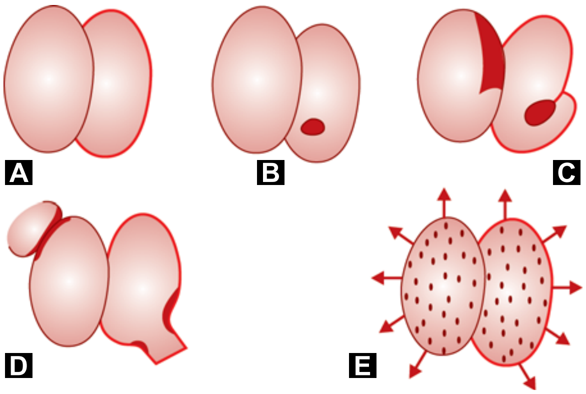

Figs 26.123A to E: Staging of Ca prostate (A) Occult, (B) Stage I, (C) Stage II, (D) Stage III, (E) Stage IV.

TNM staging of carcinoma prostate

T1 – Incidentally confirmed carcinoma after prostatectomy
- T1a – Tumour occupying less than 5% of specimen
- T1b – Tumour occupying more than 5% of specimen
- T1c – Tumour impalpable but suspected by high PSA

T2 – Tumour nodule palpable on rectal examination
- T2a – Single nodule palpable on digital examination within prostate capsule
- T2b – Nodule involving both lobes

T3 – Tumour extends through the capsule
- T3a – Extension through capsule one or both
- T3b – Extension to seminal vesicles
- T4 – Tumour which is fixed or spread to adjacent structures like pelvic wall and rectum other than seminal vesicles

Note:
For T1a progress rate is slow—15-20% in 8 years
For T1b and T2 it is 35%
For T3/T4 M0 it is 50% in 5 years with bone spread
For M1 average survival is 3 years

Note:
Gleason score *more than 7* carries poor prognosis.

Spread

Local spread:

- Upward into seminal vesicles, bladder neck, trigone, later into both ureters causing anuria. Downward into distal sphincter.

Blood spread:

- Into the bones commonly—pelvic bones, lumbar vertebrae, femoral head, ribs, skull—in that order.
  - Pathological fractures can occur in long bones and vertebrae.
  - Paraplegia may occur if spine is involved.
  - Rarely spread to liver and lung can occur.

Lymphatic spread:

  - Into the obturator lymph nodes, then to internal iliac lymph nodes.
  - Through seminal vesicles, into external iliac and retroperitoneal lymph nodes.
  - Eventually mediastinal, left supraclavicular lymph nodes get involved.

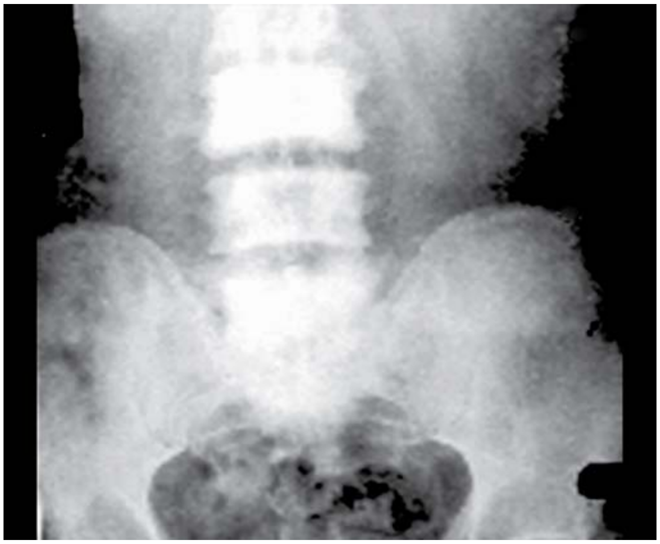

Fig. 26.124: A plain X-ray AP view of lumbar spine in a patient with carcinoma prostate. Note the typical osteosclerotic secondaries.

Clinical features

- ❖ Commonly asymptomatic
- ❖ Bladder outlet obstruction and so retention of urine
- ❖ Haematuria, frequency

- ❖ Pelvic pain, back pain, arthritic pain in sacroiliac joint—features of secondaries
- ❖ On per rectal examination, prostate feels hard, nodular, irregular often with loss of median groove
- ❖ Incidental carcinoma after TURP or after PSA analysis
- ❖ Features of renal failure
- ❖ Anaemia secondary to extensive bone marrow invasion and also due to renal failure

Differential Diagnosis

- ♦ Are other causes of retention of urine and other causes of back pain.

Investigations

- ♦ Hb%, peripheral smear. In metastatic disease, there may be leukoerythroblastic reaction with bone marrow invasion causing anaemia, thrombocytopenia, DIC and increased fibrinogen degradation products (FDP). Anaemia may also be due to renal failure
- ♦ Prostate specific antigen (PSA): More than 10 nmol/ml is suggestive.
- ♦ Prostatic fraction of acid phosphatase is increased.
- ♦ Blood urea, serum creatinine, liver function tests.
- ♦ Transrectal ultrasound (TRUS) is very useful.
- ♦ Transrectal prostatic biopsy. 10 biopsy cores are taken.
- ♦ Plain X-ray, KUB, may show *dense coarse sclerotic* secondaries. Osteolytic or combination of lytic and sclerotic lesions are also often seen.
- ♦ Technetium radioisotope bone scan to see secondaries.

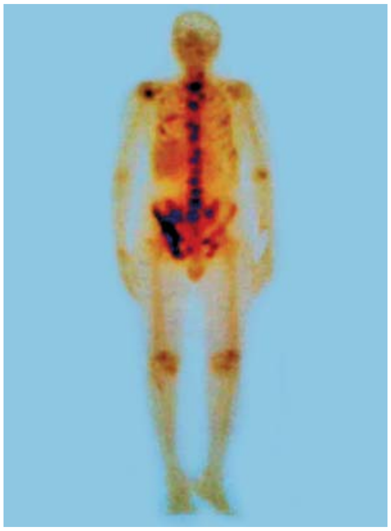

Fig. 26.125: Carcinoma prostate showing skeletal secondaries in bone scan.

- ♦ U/S, abdomen, to see the tumour extension into the bladder and to see kidneys for hydronephrosis.
- ♦ MRI/CT scan is better for staging the disease. MRI is ideal. It can be combined with TRUS to have an accurate staging.

Note:

Osteoblastic secondaries occasionally can be seen in carcinoma breast.

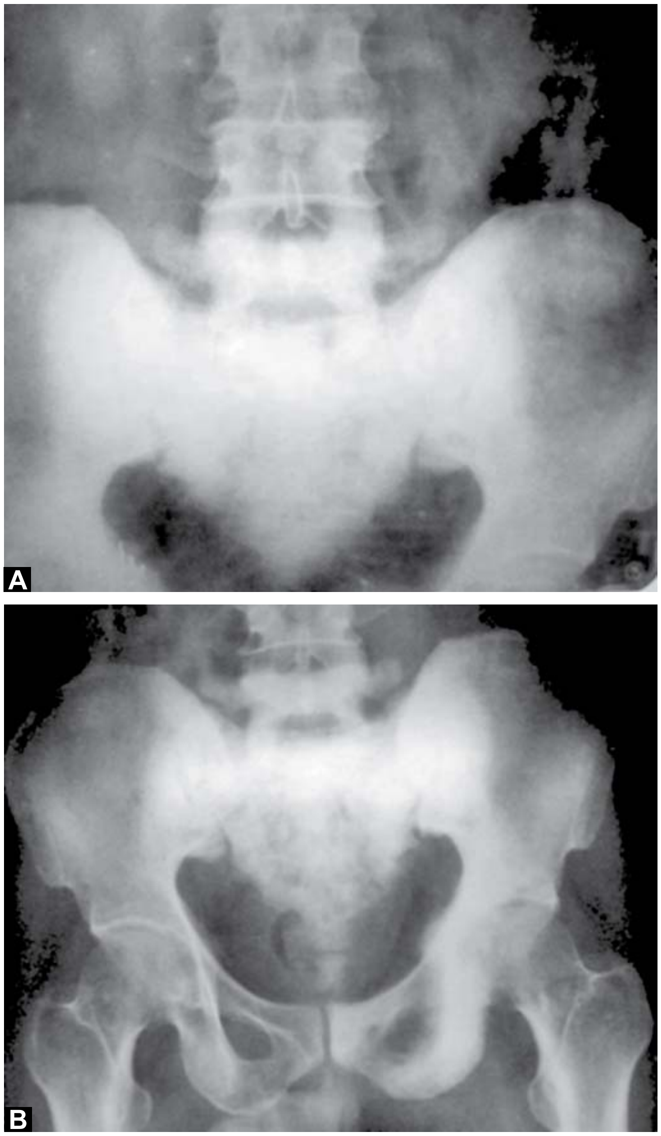

Figs 26.126A and B: Osteoblastic secondaries in pelvic bones and lumbar spine.

Treatment

> *In prostatic cancer with marked elevation of acid phosphatase, castration or injection of large amounts of estrogen caused a sharp reduction of this enzyme to or towards the normal range.... In 3 patients with prostatic cancer, androgen injections caused a sharp rise of serum acid phosphatase.*
> **—Charles Brenton Huggins, Clarence Vernard Hodges,** 1941

- ♦ *Radical prostatectomy* is done in early growth with removal of prostate, seminal vesicle, distal sphincter along with reconstruction of the urethra.

- *Radical radiotherapy* for early carcinoma prostate can be given using both interstitial and external radiation.
- *Pelvic lymph node dissection with I^{125} radiation seeds implantation.*
- *Bilateral subcapsular orchidectomy* is done to reduce the testosterone level. Very useful method. First started by ***Charles Huggins***—Urologist *(Nobel Prize winner)*. (Other condition where orchidectomy is done is carcinoma of male breast.)
- ***TURP*** + ***Bilateral orchidectomy*** + ***External radiotherapy for bone secondaries*** + ***Flutamide*** or ***Honvan***—commonly advocated method.
- External radiotherapy for spine and pelvis.
- Strontium isotope radiotherapy.
- *Drugs:*
  - Phosphorylated diethylstilbestrol ***(Honvan)***—initially given intravenously and later orally. It is very effective. Complications are cardiac congestion, DVT, gynaecomastia.

Phosphorylated diethylstilbestrol

- Initially IV 100 mg/daily later orally 100 mg/daily
- Can cause cardiac congestion, DVT, gynaecomastia
- Prior RT to breasts to prevent gynaecomastia often advocated
- Response starts appearing in 48 hours

  - LHRH agonists (Medical castration)—Leuprolide, Goserelin.
  - Androgen receptor blocking agents *like flutamide*, bicalutamide.
  - Cyproterone acetate (also has got progestogenic effect).
- In elderly people with early carcinoma prostate *inactivity with wait and watch policy* is practised.
- Sometimes permanent suprapubic cystostomy (SPC) is required in these patients.

Prognosis is good

Remember

- Wait and watch policy is ideal in elderly patient with early carcinoma (P/R; PSA) above 70 years
- TURP if there is obstruction with bilateral subcapsular orchidectomy; LHRH agonists; androgen blockage
- Radical prostatectomy/radical RT in young T1a or T1b—early disease. The criteria for radical prostatectomy—life-expectancy should be more than 10 years; PSA < 20 mmol/ml; bone scan should be negative
- Brachytherapy using guided RI seeds of I^{125} and palladium
- Orchidectomy; LHRH agonists (Leuprolide); androgen receptor blockade (flutamide, bicalutamide); TURP if needed—commonly done procedure
- External RT/IV radiopharmaceutical Strontium 89 (bone seeking isotope) for bone secondaries—used

D. Urethra

CHAPTER OUTLINE

- Anatomy
- Urethral Injury
- Stricture Urethra
- Hypospadias
- Epispadias
- Posterior Urethral Valve
- Urethral Calculi
- Urethritis
- Extravasation of Urine
- Retention of Urine

ANATOMY

Posterior urethra:
- Prostatic urethra.
- Membranous urethra.

Anterior urethra:
- Bulbar urethra.
- Penile urethra.

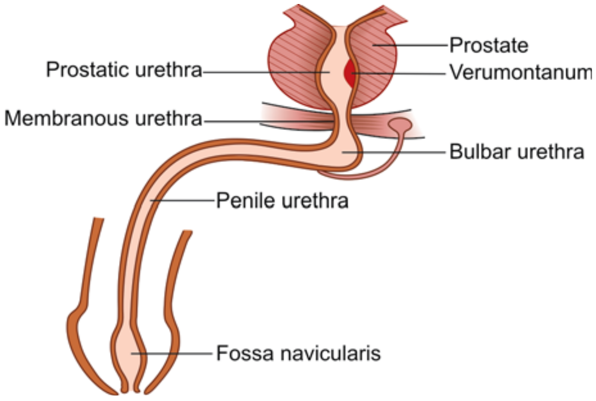

Fig. 26.127: Anatomy of urethra.

URETHRAL INJURY

Classification

I. *Depending on site of rupture:*
  1. Rupture of the membranous urethra.
  2. Rupture of the bulbous urethra.

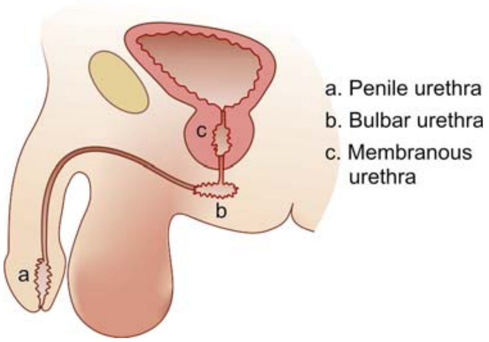

Fig. 26.128: Common sites of urethral injuries.

II. *Depending on circumference of the urethral wall involved:*
  1. Complete.
  2. Incomplete.

III. *Depending on the thickness of the urethra involved:*
  1. Total.
  2. Partial.

RUPTURE OF MEMBRANOUS URETHRA AND/OR PROSTATIC URETHRA (POSTERIOR URETHRA)

Causes

- It is usually associated with pelvic fracture, commonly due to road traffic accidents
- Injury can also occur during instrumentation
- Calculus passage and catheterisation
- In prolonged labour, due to long-standing pressure on the urethra by foetal head

- Prostate is attached to pubis by puboprostatic ligament and disruption of puboprostatic ligament with complete rupture of urethra can lead to *floating prostate—Vermooten's sign.*
- Injury can lead to incomplete rupture of urethra or may be associated with extraperitoneal rupture of bladder.

Based on ascending urethrogram, posterior urethral injury is classified as (McCallum-Colapinto classification).

Type I: Elongation of posterior urethra, but intact

Type II: Prostate "plucked off" membranous urethra with extravasation of urine above sphincter only—*Floating prostate—Vermooten's sign*

Type III: Total disruption of urethra with extravasation of urine both above and below the sphincter

Clinical Features

- Blood in external meatus. Failure or difficulty in passing urine.
- Extravasation of urine to scrotum, perineum and abdominal wall.
- *Shock* with pallor, tachycardia, hypotension.

Silence is not only the absence of sound but also stillness of the mind.

- Features of associated injuries like head injury, thorax and or abdominal organs which take priority in initial phases of management.
- On P/R examination, prostate may be felt high or may not be palpable at all. Signifies *floating* prostate.

Investigations

- X-ray pelvis to see for fracture.
- U/S abdomen to see pelvis and other injuries.
- *Urethrogram* is done to see the site and type of tear (often reserved to do at later stage).

Treatment

The shock and associated injuries are treated.

- In *floating prostate:*
  - *As rupture is complete*, bladder is opened from above. A metal bougie is passed from above through the bladder and one more metal bougie is passed from below through urethra and both are manipulated so as to meet each other.
  - Lower bougie is negotiated along the upper one and so into the bladder. Red rubber catheter is tied to the tip of the lower (urethral) bougie which has already entered into the bladder. When lower bougie is pulled out per urethrally, catheter tied to it will pass through urethra from above, to which Foley's catheter is tied and pulled up, so as to keep it in position.
  - Bladder is closed with a SPC using Malecot's catheter—*Railroad technique*.

In ***incomplete rupture:*** *Two approaches:*

- *First approach proposed by Mitchell*:
  - *Do not pass catheter from below* as it may further damage the urethra and also may damage sphincter mechanism and so may cause incontinence later. Here SPC is done using Malecot's catheter.
  - After three to six weeks, an urethrogram is done. Using endoscope or along with open method Foley's catheter is passed, often after dilatation.
- *Second approach advocated by Blandy*:
  - *Single attempt to pass a small soft catheter per urethrally* gently may lead into the bladder, which will be kept *in situ*, to maintain the continuity.
  - If this fails SPC is done. On second day, in operation theatre (OT), bladder is opened from above and flexible cystoscope is passed from below and using this, catheter is passed from below.
  - Bladder is closed with a SPC.
- If patient with incomplete rupture presents later, then it is managed once a stricture forms, accordingly as stricture urethra, after 3 months. Until then patient may require SPC.
- Other measures: Antibiotics, blood, fluid replacement, treatment of other injuries.

Complications

- Urinary incontinence
- Impotence
- Stricture urethra
- Infection

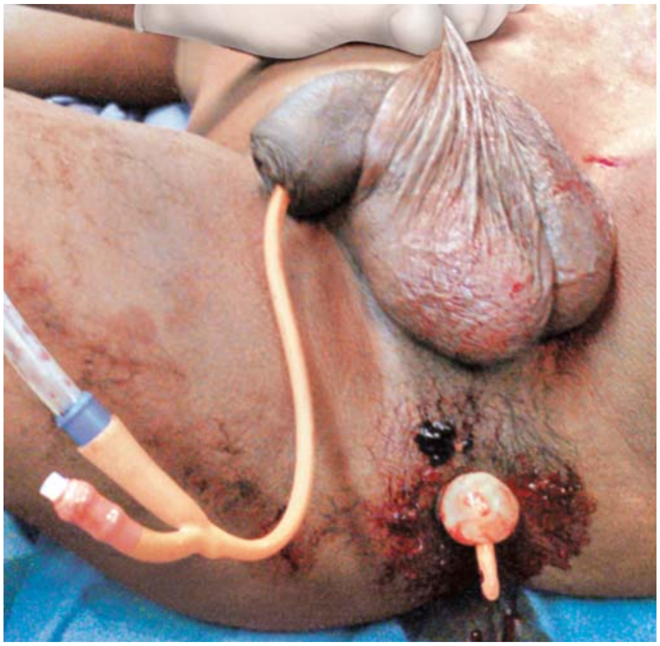

Fig. 26.129: Penetrating injury in the perineum causing rectal and urethral injury which is communicating. Note the catheter passed per urethra has gone into the rectum and come out through the anal canal. It needed colostomy and suprapubic cystostomy diversions. (*Courtesy:* Dr Ashfaque Mohammad, DNB, KMC, Mangalore)

RUPTURE OF BULBOUS URETHRA (Anterior Urethra)

- Usually, due to a fall astride a projecting object, like in sailing ships, cycling, over loose manhole cover, gymnasium.

Clinical features: *Triad*

- Blood in external meatus (Urethral haemorrhage)
- Perineal haematoma
- Retention of urine

- Rupture may be *complete* or *incomplete*.
- *Total* or *partial*.

Investigations

X-ray pelvis, and U/S abdomen. Condition is diagnosed clinically.

Treatment

- *Patient should be told not to try to pass urine*, if passed, then extravasation of urine occurs.
- In operation theatre, *one attempt of urethral catheterisation is tried gently*. If able to pass a catheter, then it is left in place.
- Often perineal haematoma which occurs, has to be drained.
- Antibiotics should be given to prevent sepsis.
- *If catheter fails to pass*, then under general anaesthesia, in lithotomy position, SPC is done. Bulbous urethra is exposed through perineal midline incision and tear is sutured with an indwelling Foley's catheter. Drain is then placed into the perineum.

- If suturing is not possible (sometimes), then *perineal urethrostomy* is done and at later stages continuity is maintained (usually after 3 months).

Complications

- ❖ Infection
- ❖ Extravasation of urine
- ❖ Stricture urethra

STRICTURE URETHRA

Classification I: *Aetiologically*.

1. Congenital.
2. Inflammatory:
   a. Post-gonococcal is most common (70%).
      - Gonococcal stricture occurs one year after infection.
      - Retention develops only 10-15 years later.
      - Common in the bulb of urethra especially in the roof.
      - Here multiple strictures are common, proximal stricture is the narrowest.

   b. Tuberculous.

   c. Other infection (urethritis).
3. Traumatic: Bulbous, membranous.
4. Postinstrumentation: Catheter, dilator, cystoscope.
5. Postoperative: Prostate surgery (4%), urethrostomy.

Classification II:

1. Proximal: Common in *bulbous urethra* (70%).
2. Distal: Congenital (in the external meatus). Often traumatic in children.

Classification III:

1. Permeable: Permits urine to pass.
2. Impermeable.

Classification IV:

1. Passable: Allows catheter to pass.
2. Impassable.

Classification V: It can be single or multiple.

Classification VI: According to the part involved. In the roof (most common) or in the floor.

Clinical features

- ❖ Poor urinary stream
- ❖ Forking and spraying of the stream
- ❖ Incomplete emptying
- ❖ Frequency, dysuria
- ❖ Retention and often with overflow
- ❖ Pain, burning micturition, suprapubic tenderness
- ❖ Thickening and button-like feeling in bulbar urethra (Bulbous urethra is felt clinically by lifting the scrotum in midline in the perineum)

Investigations

- Urine microscopy and culture.
- Blood urea and serum creatinine.
- IVU to see hydronephrosis and function of kidney.
- U/S abdomen.
- X-ray of pelvis to see old fracture with history of trauma.
- *Ascending urethrogram* is an essential investigation to see the site, type, extent and false passage.

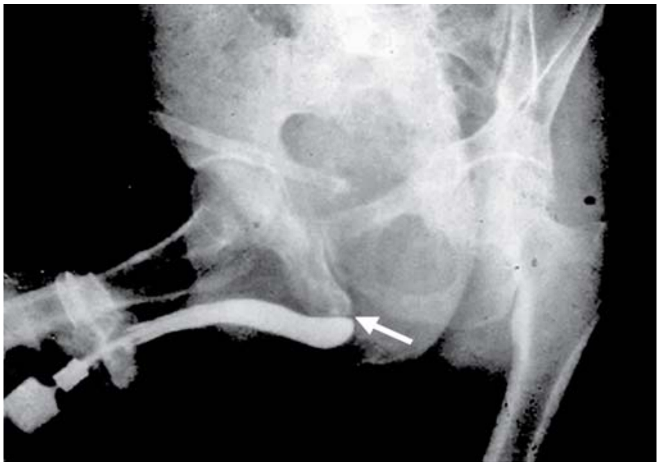

Fig. 26.130: Ascending urethrogram showing failure of dye to pass due to stricture urethra.

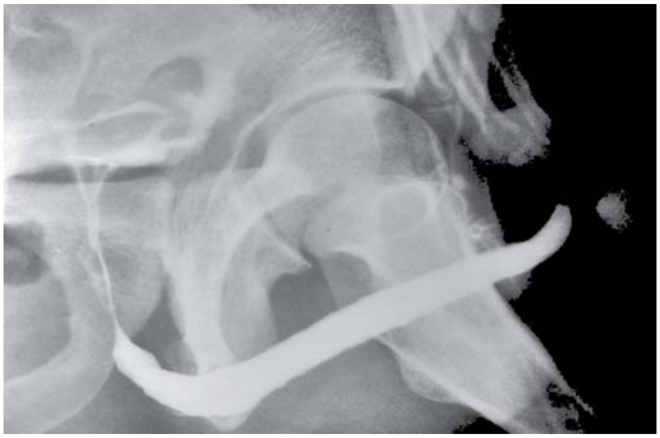

Fig. 26.131: Ascending urethrogram showing stricture urethra.

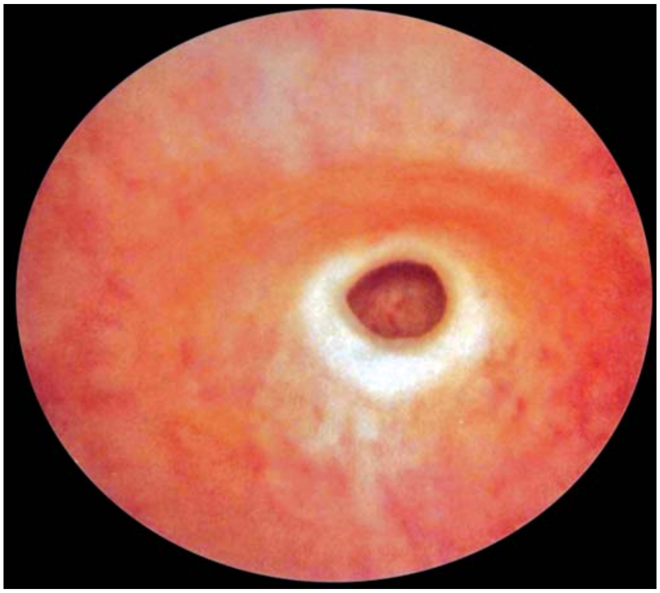

Fig. 26.132: Cystoscopic view of urethral stricture.

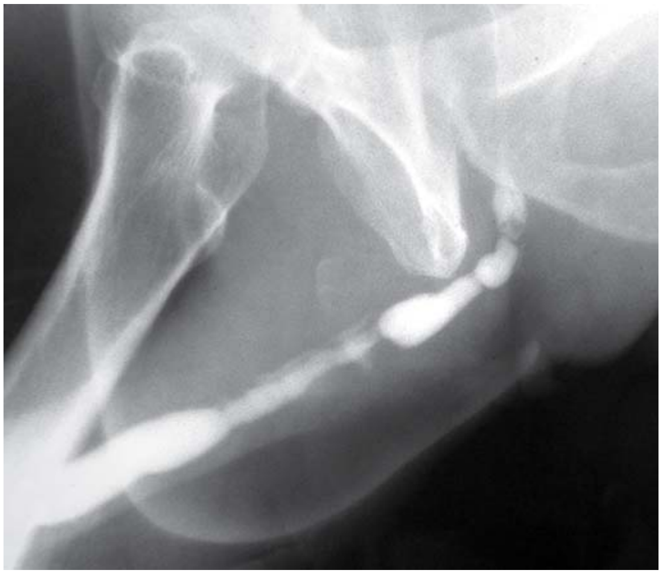

Fig. 26.133: Ascending urethrogram showing multiple urethral stricture.

- The dye is injected through suprapubic needle puncture into the bladder and visualisation is done using C-Arm image intensifier.
- Urodynamic studies.
- Urethroscopy.

Treatment

1. **Intermittent dilatation**:
   Gradual dilatation, initially with thin dilators, later with thicker dilators of increasing size. Dilatation should be done in OT under aseptic precaution.

   Should avoid forcible dilatation or over dilatation.

   Dilatation is done "***Once a week for one month, once a month for one year, and later once a year (on his birthday)"***.

   Dilators used:
   a. Lister's dilator [has got olive tip (blister)].
   b. Clutton's dilator.
   c. Filiform bougies.

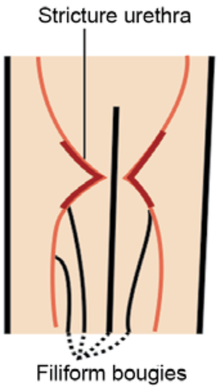

Fig. 26.134: Filiform bougies are used to dilate narrow stricture urethra. One of the bougies will be passed into the urethra.

Complications of dilatation

- Infection and bleeding due to trauma
- False passage
- Fistula formation

2. **Visual internal cystoscopic urethrotomy or stricturotomy**:
   - Here using cystoscope, stricture is visualised and is cut at 12 o'clock position, until it bleeds (fibrous tissue is cut completely).
   - After that Foley's catheter is passed and kept in position for 48 hours.
3. **External urethrotomy** by open method. Presently not commonly done as cystoscopic urethrotomy is more popular. It is presently done as an initial stage surgery for urethroplasty *(Wheelhouse's operation*).
4. **Urethroplasty:** Stricture is excised and urethra is reconstructed using *prepuceal skin* or scrotal skin *(Johanson's urethroplasty).*

Problems in urethroplasty

- Staged procedure and so prolonged hospitalisation
- Infection
- Necrosis of skin flap
- Leak and fistula formation
- Restenosis

Complications of stricture urethra

- Retention of urine
- Urethral fistula
- Infection—urethritis, cystitis, pyelonephritis
- Urethral diverticula
- Periurethral abscess
- Bilateral hydronephrosis
- Stone formation
- Renal failure
- Due to straining—hernia, haemorrhoids, rectal prolapse

HYPOSPADIAS

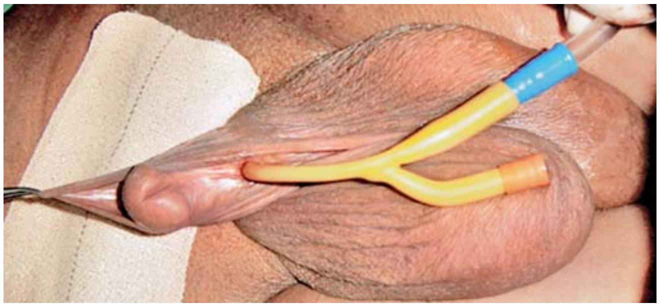

Fig. 26.135: Hypospadias.

It is the most common congenital malformation of urethra wherein external meatus is situated *proximal* than normal, over the *ventral* (under) aspect of the penis.

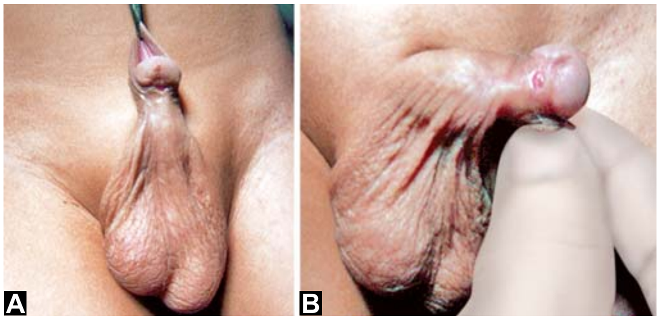

Figs 26.136A and B: Hypospadias in coronal region of glans penis with typical chordee.

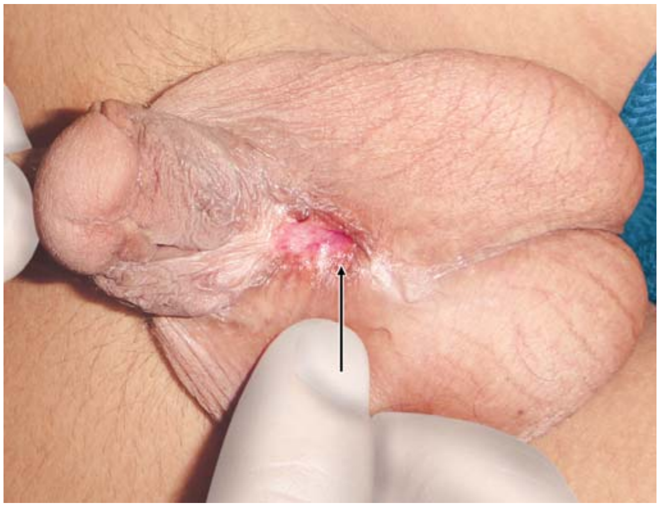

Fig. 26.137: Penoscrotal hypospadias.

Classification

- Glandular: Meatal opening in glans. It is the most common type.
- Coronal.
- Penile.
- Penoscrotal.
- Perineal with split scrotum. This is associated with bilateral undescended testes.

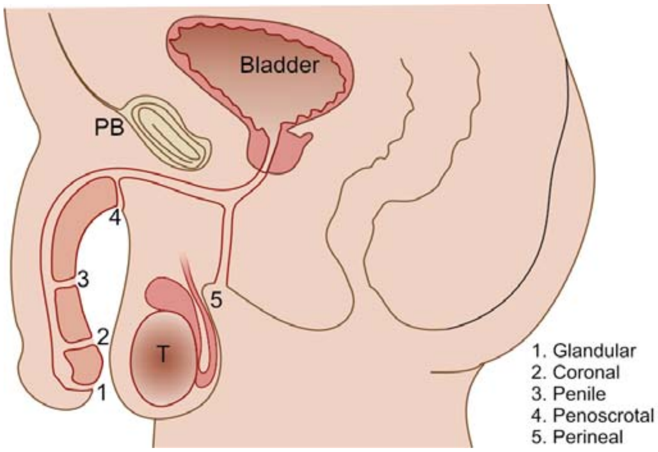

Fig. 26.138: Types of hypospadias.

Features

- Absence of urethra and corpus spongiosum distal to abnormal urethral orifice.
- Bowing or bending of penis distal to abnormal urethral opening *(chordee),* with poorly developed prepuce over inferior aspect.
- Urine soakage over the scrotum with dermatitis and infection.
- Associated congenital anomalies are known to exist.

Treatment: Staged Procedure

I. At the age of one and half years, surgical correction of the chordee is done.

II. At the age of 5-7 years, reconstruction of urethra is done using *prepucial skin* (ideal) or scrotal skin if the patient has been circumcised. Best is prepucial skin.

- *Perineal urethrostomy is done for diversion.*
- Perineal urethrostomy should be there until reconstruction of urethra takes up well *(Denis-Browne procedure).*
- *Magpi repair*: Meatal advancement glandular repair—for glandular hypospadias.

In hypospadias, circumcision is contraindicated as prepucial skin is required for future urethroplasty.

EPISPADIAS

- Here the urethra opens on the dorsum of the penis, proximal to the glans.
- Most common site is at the abdominopenile junction.
- It is associated with a dorsal chordee, ectopia vesicae, urinary incontinence, separated pubic bones.
- It is uncommon in females.

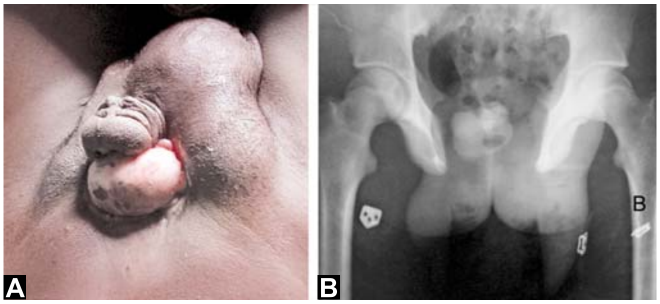

Figs 26.139A and B: Bladder extrophy. X-ray pelvis showing separation of pubic bones. Epispadias is the common association.

Treatment

- Correction of incontinence of urethra.
- Urinary diversion.

POSTERIOR URETHRAL VALVE

- They are congenital symmetrical valves in the posterior urethra, just below the verumontanum.
- It allows the passage of catheter without obstructing its ingress. But it obstructs the outflow of urine.

The greatest service is to share with others the joys of living.

- Proximal urethra is enormously dilated with obstructive pathology in the bladder (sacculations and diverticula formation).
- Bladder wall is thickened and hypertrophied so much so that it is palpable in suprapubic region as a firm swelling *(cricket-ball bladder).*
- There is poor urinary stream, with hydronephrosis, often with infection.
- Child finds difficult to pass urine.
- Often it is associated with vesicoureteral reflux.
- Renal failure and atrophy of kidney occurs due to back pressure.

Investigations

- *Micturating cystourethrography (MCU)* is diagnostic. It shows dilated proximal urethra.
- U/S abdomen.
- Blood urea and serum creatinine.
- IVU.
- Antenatal US shows fetal urinary tract dilatation.

Differential Diagnosis

- ***Marion's disease:*** Bladder neck obstruction due to hypertrophied interureteric bar. MCU differentiates it from posterior urethral valve. Treatment is Y-V plasty ***(Bonin's Y-V Plasty).***
- Neurogenic bladder.

Treatment

Initial SPC and later cystoscopic resection of the posterior valves is done usually after several weeks.

URETHRAL CALCULI

Stone from the bladder is commonly passed out through urethra if it is small, but the stone can get impacted due to a *stricture* or *urethral diverticulum*.

Sites of impaction

- Prostatic urethra
- Bulbous urethra
- Fossa navicularis
- External meatus

Clinical Features

- Painful urination with thin stream and forking of urine.
- Retention of urine.
- Lower urinary tract infection.
- Haematuria.
- Pain in the penis and perineum.
- Stone may be palpable when it is in the bulbar or penile urethra.

Investigations

- Plain X-ray, KUBU (kidney, ureter, bladder urethra).
- U/S abdomen.
- Urine examination.

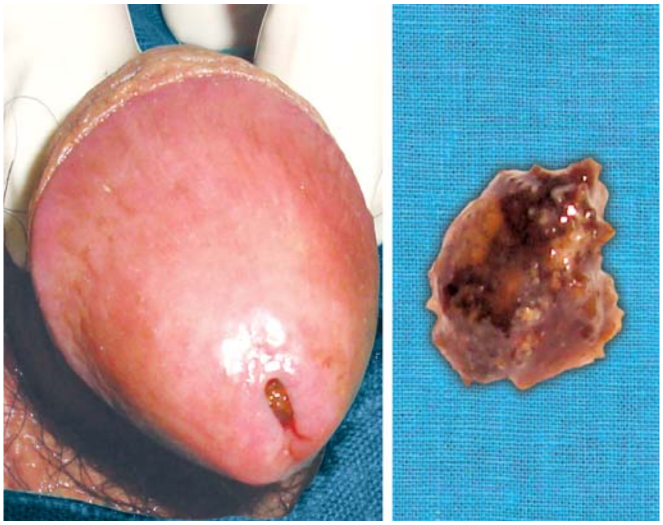

Fig. 26.140: Stone at urethral meatus (external meatus) which is extracted.

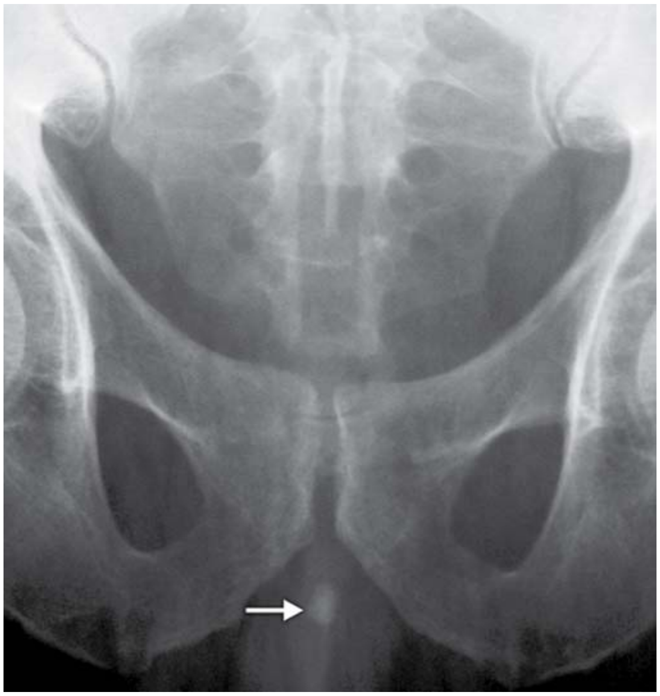

Fig. 26.141: X-ray showing urethral stone.

Treatment

- If it is impacted in the external meatus or fossa navicularis, meatotomy and extraction of the stone is done (Meatotomy is cutting and widening of external meatus so as to have an easy stone extraction).
- If it is impacted in other parts of the urethra, urethroscopy and removal of stone from urethra is done or through a cystoscope the stone is pushed into the bladder and then extracted.
- In case of retention of urine, initial suprapubic cystostomy is required before removal of the stone.
- Diverticulum and stricture urethra should be treated accordingly.

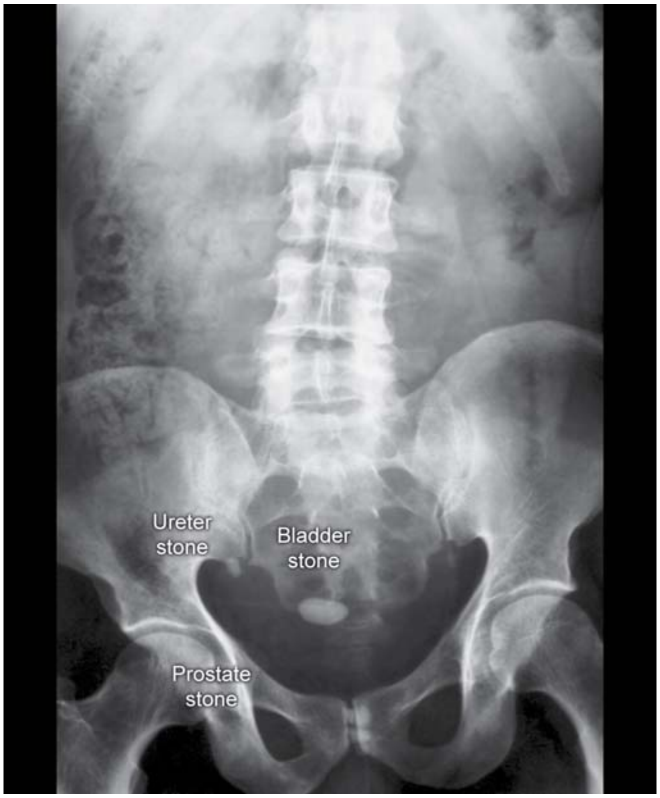

Fig. 26.142: Plain X-ray KUB showing stone in the bladder; stone in the terminal end of ureter; stone in the prostatic urethra.

Complications

Bleeding, stricture urethra and infection.

URETHRITIS

Causes

- Gonococcal urethritis.
- Nongonococcal urethritis:
  - Trichomonal urethritis.
  - Lymphogranuloma venereum (LGV).
  - Mycoplasma.
  - Ureoplasma ureolyticum.
  - *Candida*.
- Trauma, catheters, cystoscopes, stones.
- Chemical urethritis.
- *Reiter's disease*: Arthritis, conjunctivitis, urethritis.
- Allergy.

Reiter's syndrome

- Nongonococcal urethritis
- Conjunctivitis
- Arthritis

Clinical Features

- Urethral discharge, dysuria, burning micturition, haematuria and increased frequency.
- Perineal pain, tenderness over the site.
- Suprapubic pain and tenderness.

Investigations

- Urine microscopy and culture.
- U/S abdomen.

Treatment

- Antibiotics given according to culture and sensitivity.
- The cause is treated.

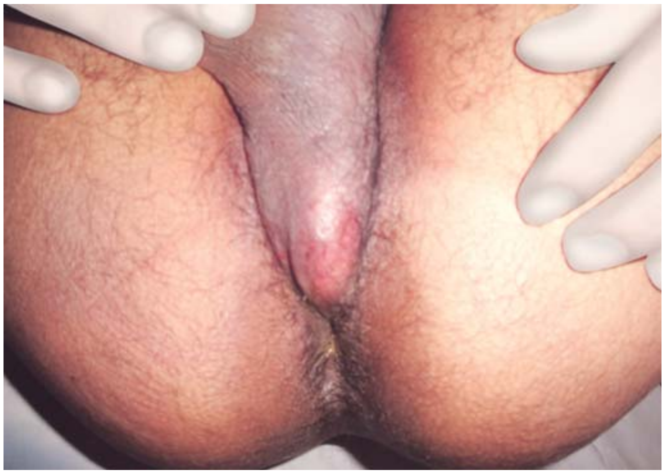

Fig. 26.143: Periurethral abscess.

EXTRAVASATION OF URINE

Urine extravasates from the bladder or urethra and collects in the layers of the perineum or in the prevesical space (cave of Retzius).

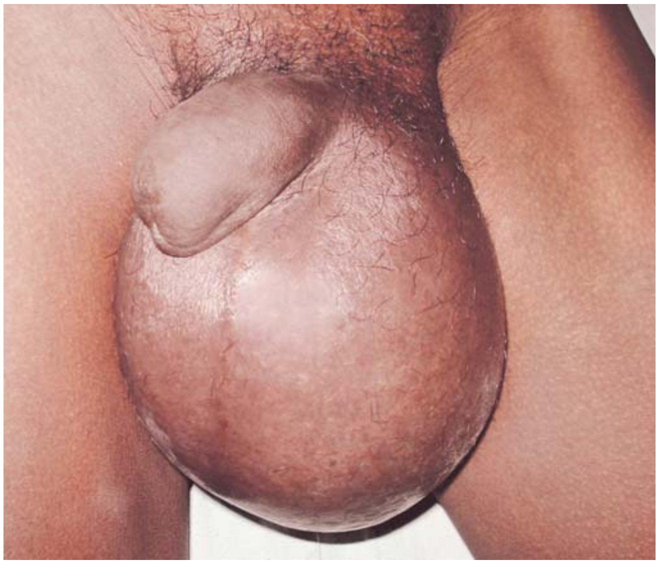

Fig. 26.144: Extravasation of urine showing the collection of urine in the scrotum and perineum after trauma.

Causes

- Trauma to bladder or urethra.
- Instrumentation like dilatation, catheterisation, cystoscopy.
- Infection like gonococcus *(water can perineum)*.

True victory means complete control over sense-organs.

Types

It may be superficial or deep.

1. *Superficial:* It is due to either bulbar urethral injury or due to bursting of periurethral abscess after urethral stricture. Once urine extravasates due to full thickness disruption of the urethra anteriorly, it collects in superficial perineal space. This space is a closed cavity all round except anteriorly where it communicates with scrotal subcutaneous tissue deep to fascia Colles, penis between superficial fascia and deeper Buck's fascia, in the anterior abdominal wall deep to Scarpa'a fascia. It does not spread to thigh and ischiorectal space as Scarpa'a fascia is attached firmly to fascia lata of thigh. Superficial perineal space is closed above by inferior fascia of perineal membrane; below by fascia of Colles; laterally by ischiopubic rami. It is open and communicating only anteriorly. Entire scrotum, penis and often lower abdominal wall are swollen containing urine. It is painful; patient cannot pass urine through urethra; severe pain and shock due to pelvic injury. Often sepsis occurs and skin sloughs of leading into urinary fistulas.
2. *Deep:* Urine spreads upwards into the extraperitoneal space of the pelvis around the bladder and prostate into the anterior abdominal wall causing deep extravasation of the urine. Here rupture of urethra is at membranous part of the urethra much more proximal than superficial type.

Clinical Features

- Pain, swelling in the perineum.
- Retention of urine.
- Suprapubic pain, tenderness and fullness.

Treatment

- The cause is treated:
  - Antibiotics.
  - Suprapubic cystostomy.

Often multiple incisions may be placed in the perineum to evacuate the extravasated urine.

Complications

Abscess formation, sinus/fistula formation.

RETENTION OF URINE

- It is *accumulation of urine* in the urinary bladder. Kidneys excrete urine in normal quantity. But patient passes only small amount of urine or does not pass urine at all causing *distended bladder*. In anuria due to renal failure, patient does not pass urine as kidney does not secrete any urine and bladder is not distended.
- Retention can be *acute; chronic; retention with overflow*. Retention differs in male, female and in infants and children.
- Chronic retention is due to BPH, bladder neck contracture, urethral stricture, etc. and is painless with suprapubic dullness.
- It is one of the common conditions seen in surgical casualty.

Causes

Urethral injury
Stricture urethra
Stone in the urethra
Phimosis with meatal stenosis
Posterior urethral valve

Prostatic hyperplasia (BPH)
Carcinoma prostate
Bladder tumour near bladder neck
Bladder neck hypertrophy
Prostatitis or urethritis

Neurogenic causes like head and spinal injuries
Postoperative period following pelvic and perineal surgeries
Drug induced retention of urine: Anticholinergics, antihistamine drugs, antidepressants

In males

- BPH in old.
- Stricture urethra in young.
- Trauma—urethral/pelvic.
- Postsurgical—perineal surgeries like haemorrhoidectomy; fistulectomy; fissurectomy; pelvic surgeries; surgeries for fracture hip/fracture pelvic bone.
- Bladder outlet obstruction.
- Carcinoma of prostate.

In females

- Uterine prolapse.
- Urethral stricture.
- Uterine/Ovarian surgeries.

In infants and children

- Posterior urethral valve.
- Meatal stenosis.

Clinical Features

- Pain and swelling (fullness) in the suprapubic region.
- Inability to pass urine.
- Smooth, soft, tender swelling in hypogastric region which is dull on percussion.
- Features relevant of specific causes.
  - Per rectal examination to see prostate and other causes. In full bladder often prostate cannot be assessed properly.

Types

- Acute
- Chronic
- Retention with overflow

Acute retention: It is rare. It is sudden inability to pass urine. It is painful distension of the bladder. It is seen in urethral trauma, due to anaesthesia, or surgery (perineal/abdominal). There is increased desire to pass urine.

Chronic retention is gradual collection of urine in the bladder due to ineffective emptying of the bladder completely. Bladder is distended and is painless. It is common in elderly. Frequency, difficulty in urination, overflow incontinence is common. Infection in such chronic retention makes it painful.

Acute on chronic retention: Patient is having chronic obstructive condition like BPH; due to infection and acute inflammation and oedema of mucosa of urethra sudden total blockage sets in causing acute on chronic retention of urine.

Investigations

- U/S abdomen.
- Blood urea, serum creatinine.
- Urine microscopy.

Treatment

- Urinary catheterisation using Foley's catheter.
- If it fails then suprapubic cystostomy (SPC) using Malecot's catheter is done.
- The cause is treated.
- Intermittent catheterisation in case of neurogenic bladder.

Management of retention of urine

- Clinical and sonological assessment of kidney, ureter and bladder
- Evaluation for cause
- Evaluation for obstructive uropathy and renal function
- Urinary catheterisation by Foley's catheter/red rubber catheter/metallic catheter.SPC with Malecot catheter by open method or using Foley's catheter by trocal puncture method
- Intermittent catheterisation
- Vesicostomy in children
- Treating the cause
- Postoperative retention is a common cause. It is treated by reassurance, placing hot water bag over suprapubic region, change of position or standing, only occasionally by catheterisation

The quality of our thoughts determines our own personal degree of happiness.

E. Penis

> *The functioning role of the penis is as well established as that of any other organ in the body. Ironically, there is no organ about which more misinformation has been perpetrated. The penis constantly has been viewed but rarely seen. The organ has been venerated, reviled, and misrepresented with intent in art, literature, and legend through the centuries.*
>
> **—William H Masters, Virginia E Johnson,** 1966

CHAPTER OUTLINE

- Phimosis
- Paraphimosis
- Circumcision
- Balanoposthitis
- Chordee
- Priapism
- Peyronie's Disease
- Ram's Horn Penis
- Carcinoma Penis
- Buschke-Lowenstein Tumour

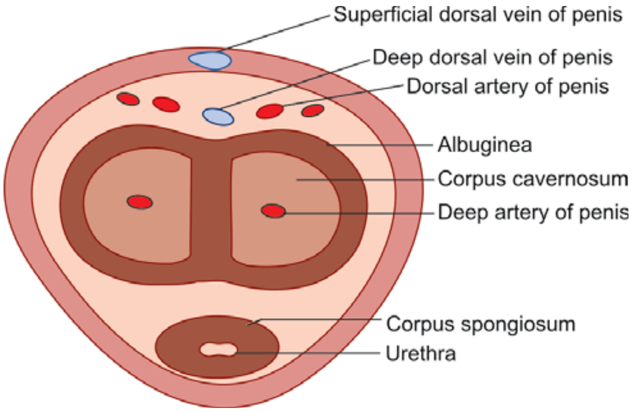
Superficial dorsal vein of penis
Deep dorsal vein of penis
Dorsal artery of penis
Albuginea
Corpus cavernosum
Deep artery of penis
Corpus spongiosum
Urethra

Fig. 26.145: Anatomy of penis.

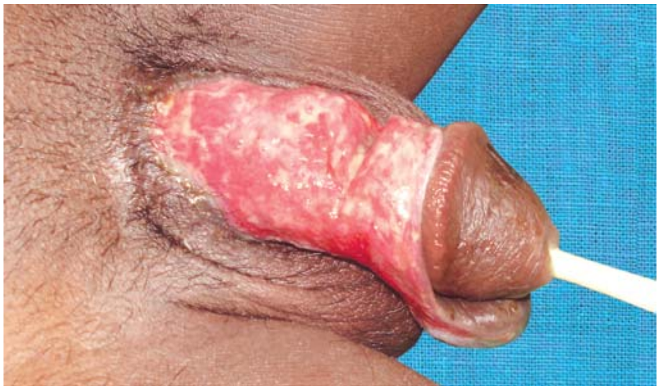

Fig. 26.146: Degloving injury of penis which needed skin grafting later.

PHIMOSIS

It is inability to retract the prepuce over the glans.

Causes

1. Congenital—in which case the child has pinhole meatus and ballooning of prepuce occurs when the child urinates.

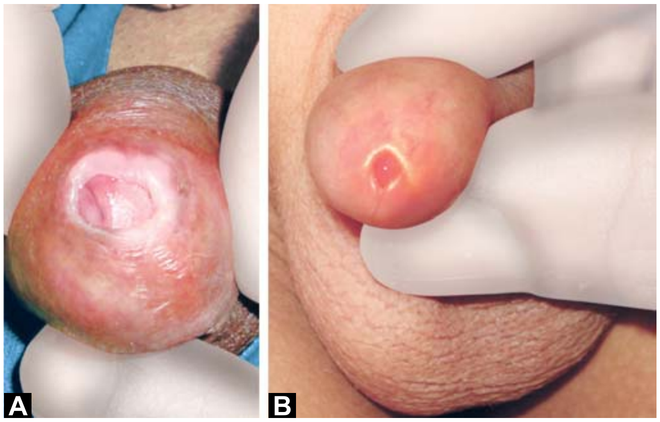

Figs 26.147A and B: Pinhole meatus with phimosis. It is an indication for circumcision. Often clinically there will be ballooning of the prepuce during urination.

2. Balanitis (inflammation of glans) and balanoposthitis (inflammation of glans, prepuce and sac). Common in diabetics.

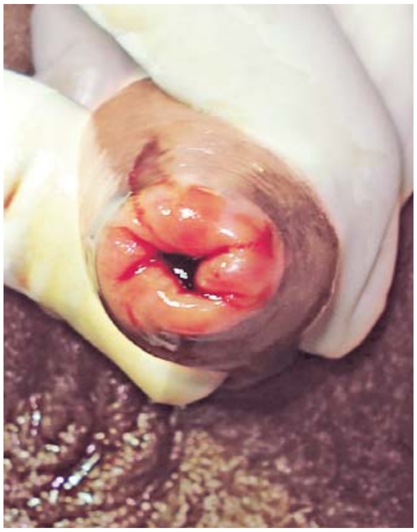

Fig. 26.148: Posthitis and balanoposthitis are common causes of phimosis in adult. It may be common in diabetics. Carcinoma should also be ruled out. Specimen should be sent for histology after circumcision.

Patients with phimosis are more prone for recurrent infection, smegma collection and *carcinoma penis*.

Treatment

Circumcision.

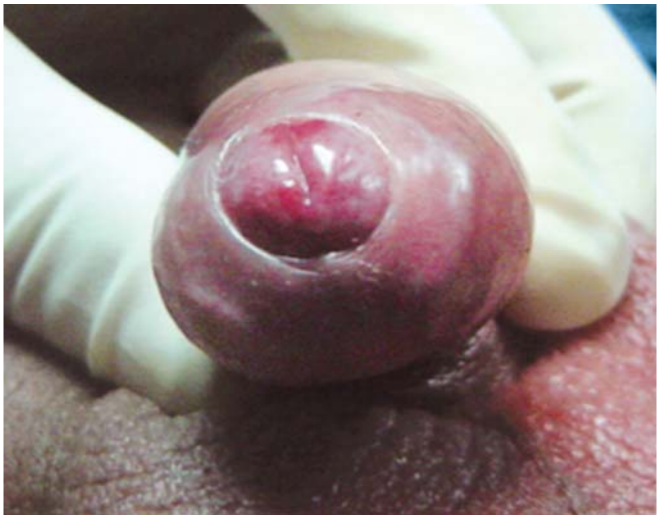

Fig. 26.149: Phimosis. It is inability to retract prepuce over glans.

Problems due to phimosis

- Recurrent balanoposthitis
- Paraphimosis
- Ballooning of prepuceal skin
- Retention of urine
- Prepuceal calculi formation due to smegma collection in prepuceal sac
- Carcinoma of penis later

PARAPHIMOSIS

- Inability to place back (cover) the retracted prepucial skin over the glans is called as paraphimosis.
- It causes ring like constriction proximal to the corona and prepuceal skin.
- As a result the glans will be swollen, oedematous with severe pain and tenderness.
- Often glans will undergo necrosis or becomes gangrenous.

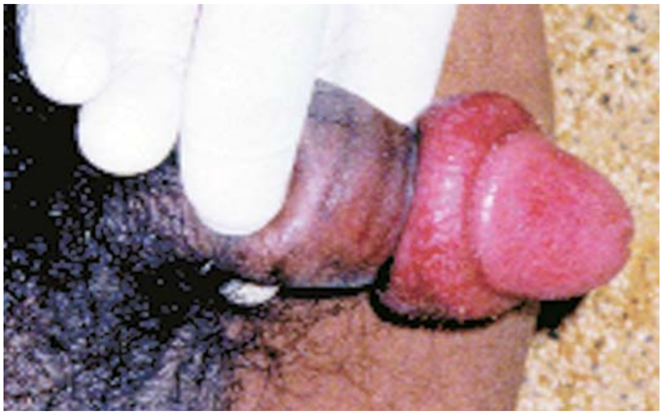

Fig. 26.150: Paraphimosis.

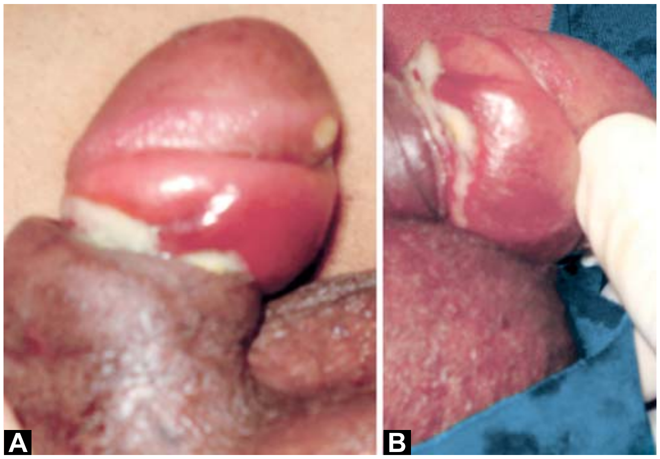

Figs 26.151A and B: Paraphimosis showing constricted ring with infection in the groove.

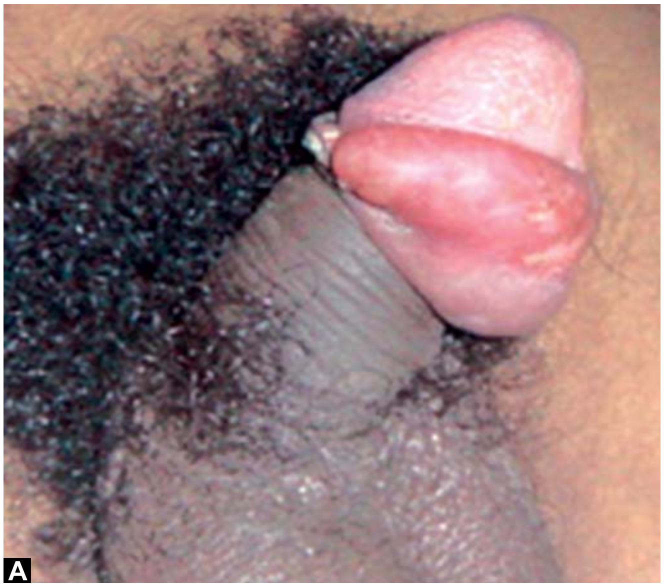

Fig. 26.152A

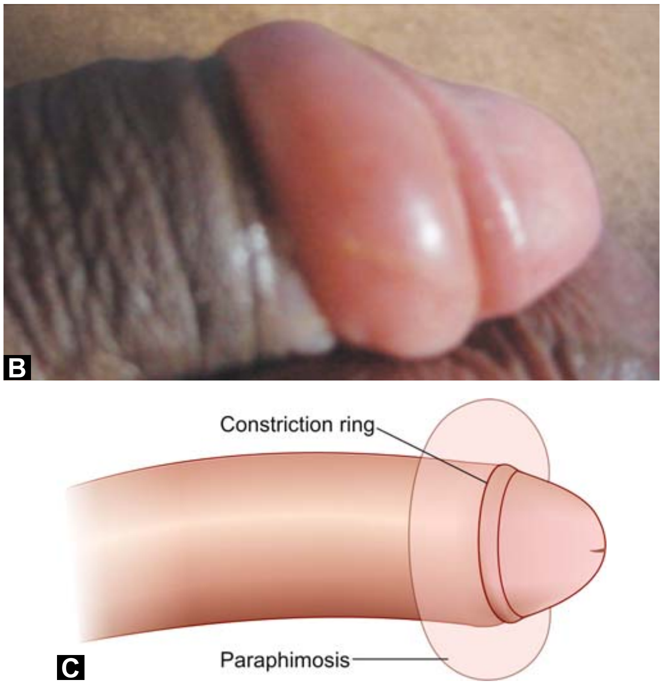

Figs 26.152B and C

Figs 26.152A to C: Paraphimosis is inabiliy to place back the retracted prepuce.

Paraphimosis is precipitated after sexual intercourse or iatrogenically after urethral catheterisation.

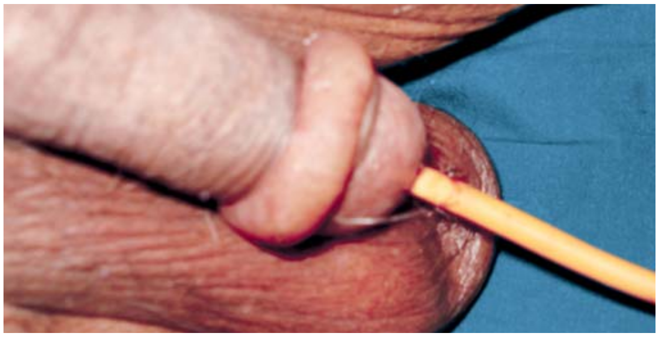

Fig. 26.153: Paraphimosis developed after catheterisation due to failure of replacing back of the prepuce.

Treatment

- Manual reduction of prepuceal skin is to be tried.
- If not possible, *initial dorsal slit* is made to relieve the oedema and compression. Antibiotics and analgesics are given. *Circumcision* is done after 3 weeks.
- Sedation and hyaluronidase injection in 10 ml saline into the constriction ring or multiple needle punctures over the oedematous part, reduces the oedema and makes the paraphimosis to get reduced. Later circumcision should be done.

CIRCUMCISION

> *And God said to Abraham, "As for you, you shall keep my covenant, you and your descendants after you throughout their generations. This is my covenant…. Every male among you shall be circumcised. You shall be circumcised in the flesh of your foreskins, and it shall be a sign of the covenant between me and you. He that is eight days old among you shall be circumcised."***—The Bible,** date unknown

Indications

- Religious
- Phimosis
- Paraphimosis after doing initial dorsal slit
- Balanitis and balanoposthitis (common in diabetics)
- Early carcinoma of prepuce or glans penis—both diagnostic as well as therapeutic purpose
- Certain sexually transmitted diseases, e.g. herpes infection

Procedure (Refer Fig. 26.154)

- In children, it is done under G/A. In adults, it is done under local anaesthesia.
- After cleaning and draping, LA [1% lignocaine *(plain)* injected circumferentially near the root of the penis] is given (ring block).
- Dorsal skin is cut up to the corona and later circumferentially and ventrally.
- The skin is cut with inner layer.
- Care is taken to see that optimum (less) skin is cut ventrally to prevent the occurrence of *chordee*.
- Frenular artery is transfixed and ligated ventrally using chromic catgut (2-0 or 3-0). Small bleeders are also ligated.
- Skin is apposed to the cut edge of corona using interrupted chromic catgut sutures.
- Postoperatively, antibiotics and analgesics are given.

Complications

- Reactionary haemorrhage due to slipping of ligature from frenular artery and dorsal vein
- Infection
- Stricture urethra near the external meatus in *children*
- Chordee due to removal of excess skin on the ventral aspect
- Rarely priapism can occur

- *Hollister Bell cap technique* (Plastibel device): This specially devised plastic cap can be fitted over the glans

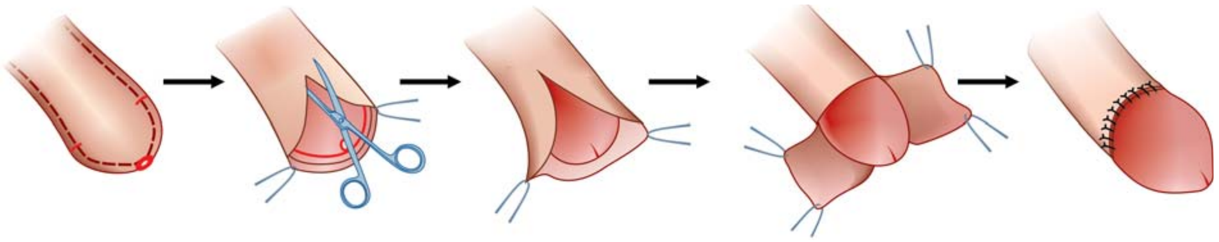

Fig. 26.154: Operative steps of the circumcision.

penis and prepuce is rolled over it. A tight ligature is tied over it near base of the prepuce. In 7 days skin and prepuce sloughs off and sheds with the cap. Bleeding will not occur due to thrombosis of prepuceal vessels. Technique can be used for religious circumcision/balanoposthitis without phimosis. It is contraindicated in phimosis and paraphimosis.

Note:

- Circumcision *by guillotine* is a method done by pulling and stretching the prepuceal skin beyond the glans and cutting the prepuce. It should be condemned and not be done as injury to glans is common. It is practiced in religious circumcision.
- ***Monopolar cautery should not be used*** in circumcision.

BALANOPOSTHITIS

It is inflammation of glans and prepuce—(Inflammation of prepuce is posthitis; of the glans is balanitis).

Causes

- Diabetes mellitus.
- Candidiasis.
- Veneral diseases like syphilis, herpes.

It can cause phimosis, carcinoma penis.

Clinical Features

Pain, swelling, discharge.

Treatment

- Antibiotics.
- Circumcision.
- The diabetes is controlled.

CHORDEE (CORDEE)

It is fixed bending of glans penis, more obvious during erection.

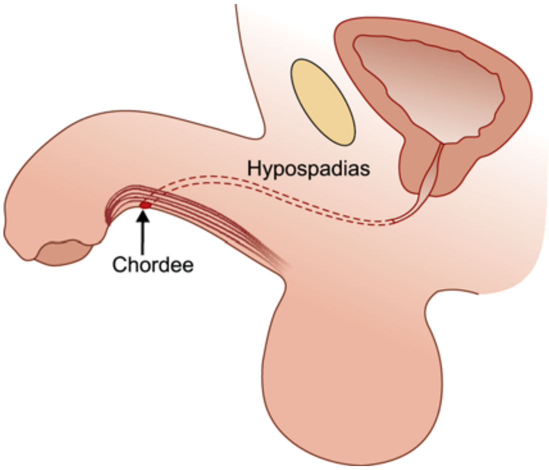

Fig. 26.155: Chordee.

Types

a. Ventral.
b. Dorsal.

Causes

- Hypospadias, where urethra opens more proximally than normal *(ventral cordee).*
- After circumcision, if more skin is cut over the ventral aspect (*ventral cordee*).
- In epispadias *dorsal cordee* occurs.

Treatment

- Chordee due to hypospadias is corrected during staged procedure.
- In chordee following circumcision, initially stilbestrol 6 mg daily is given. Later chordee is corrected surgically by excising fibrous tissue and later doing skin grafting. More often they require surgical intervention.

PRIAPISM

- It is persistent, painful erection of penis.
- Corpora cavernosa are filled with blood due to defective venous drainage. Glans and corpus spongiosum are not involved.

Causes

- Idiopathic *thrombosis of corpora cavernosa*
- Idiopathic thrombosis of prostatic venous plexus
- Sickle cell disease
- Leukaemia
- Secondary deposits in corpora cavernosa
- Spinal injury or diseases

Investigations

Relevant for specific causes.

Treatment

- Anastomosis between corpora cavernosa and saphenous vein.
- Anastomosis between corpora cavernosa and corpus spongiosum.

PEYRONIE'S DISEASE (Induratio-penis Plastica)

- It is development of fibrous tissue plaque on the covering of corpus cavernosum and later involving its full extent resulting in induration of corpus.
- It is a slowly progressive disease of uncertain aetiology, may be due to old trauma, often associated with *Dupuytren's contracture, retroperitoneal fibrosis and plantar fasciitis.*

- Initial active phase has painful erection with changing deformity of penis.
- Later quiescent phase has disappearance of painful erection with development of deformity which is painless.
- Later erectile dysfunction; penile shortening occurs.
- Indurated plaque is noticed in the penis.

Treatment

- Some cases resolve spontaneously.
- Drugs
  - Steroids.
  - Vitamin E.
  - Potassium amino benzoate 12 mg/day.
  - Tamoxifen 20 mg daily.
  - Terfenadine and fexafenadine.
  - Colchicine therapy.
  - Intralesional injection of verapamil 10 mg once in 2 weeks—12 injections.
- Surgery is needed in many cases
  - Excision and plication to opposite side—*Fitzpatric operation*.
  - Multiple incisions over the fibrous plaque and temporal fascia bridging—*Gelhard's operation*.
  - Excision of fibrous plaque and corporotomy is covered with overlay flap like tunica vaginalis flap (*Lockhart's*)/ dermal flap (*Devine and Horton's*).
  - Intracorporeal penile prosthesis placement.

RAM'S HORN PENIS

Filarial involvement of penis where it becomes thick, distorted and resembles horn of a ram.

CARCINOMA PENIS

- It is commonly *squamous cell carcinoma*, but melanoma, adenocarcinoma from Tyson's gland, basal cell carcinoma and secondaries may also occur.

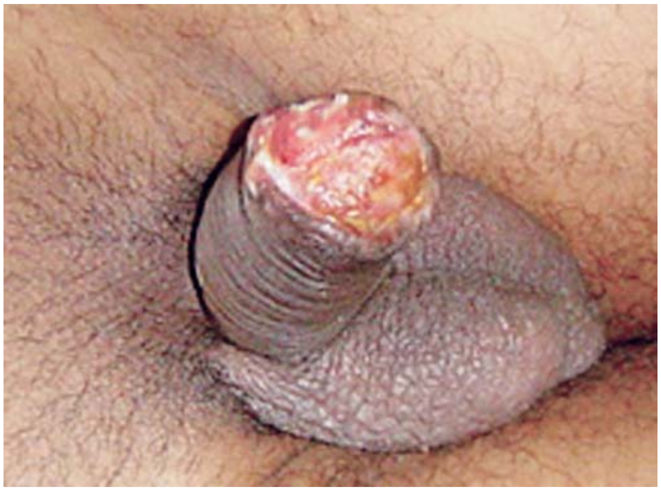

Fig. 26.156: Carcinoma of penis involving glans extensively.

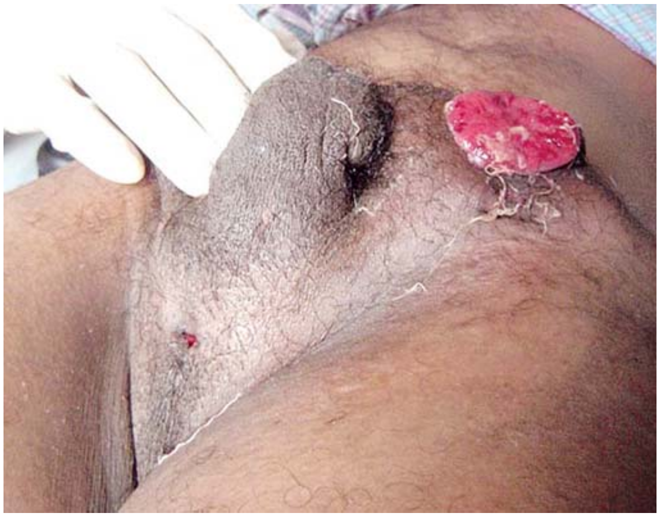

Fig. 26.157: Carcinoma penis earlier operated by total amputation of penis has now developed secondaries in inguinal lymph nodes which has already fungated. Note the perineal urethrostomy done.

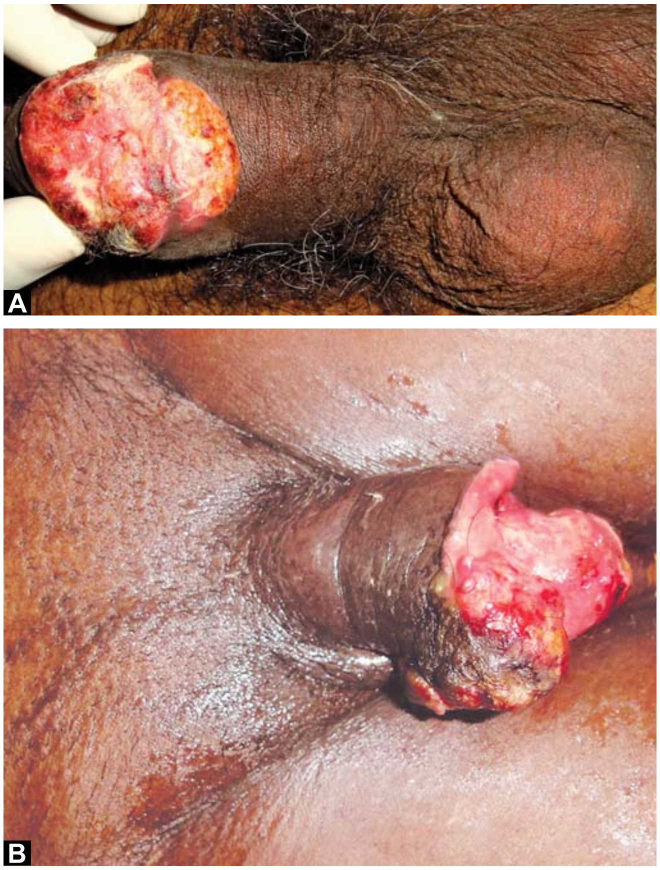

Figs 26.158A and B: Carcinoma penis in two different patients.

Aetiology

- Chronic balanoposthitis, phimosis.
- Sexually transmitted diseases.
- Leukoplakia of glans.
- Long-standing genital warts.
- *Paget's disease* of penis (Erythroplasia of Querat is persistent rawness of glans penis).

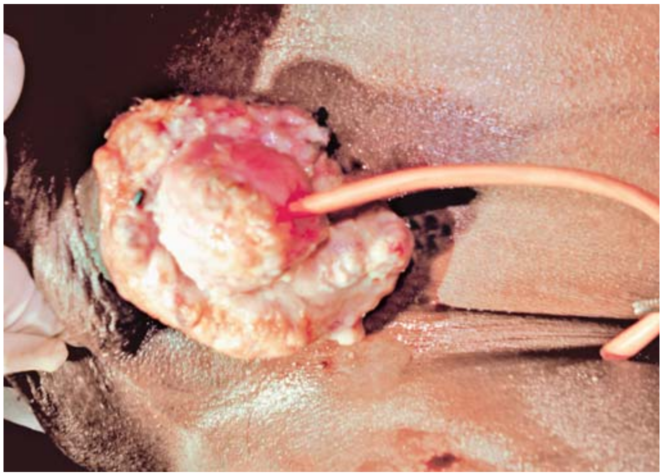

Fig. 26.159: Carcinoma penis—circumferential proliferative exophytic lesion.

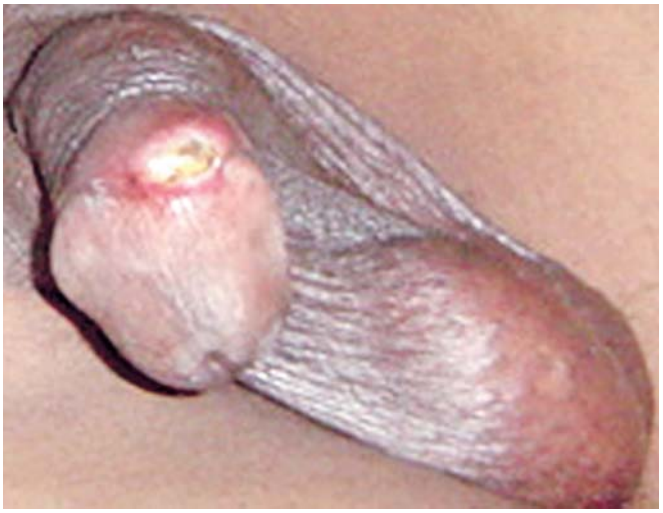

Fig. 26.160: Small ulcerated lesion over glans penis—carcinoma.

- Condyloma acuminata (by human papilloma virus), balanitis xerotica obliterans.
- HIV infection—HPV-16.

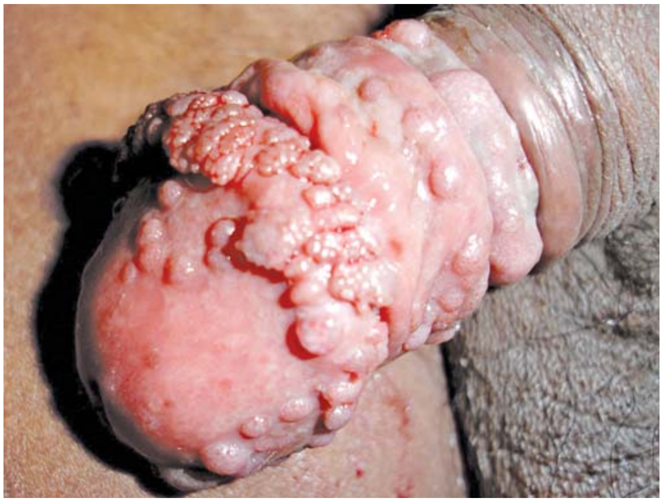

Fig. 26.161: Genital warts—in glans penis. It is a premalignant condition. Biopsy, excision and podophyllin application is needed.

  - *Circumcision during infancy confers total immunity against carcinoma penis*.
  - It is common in Asia and Africa.

Pathology

- *Infiltrating type* occurs in a preexisting leukoplakia. It often presents as indurated area.

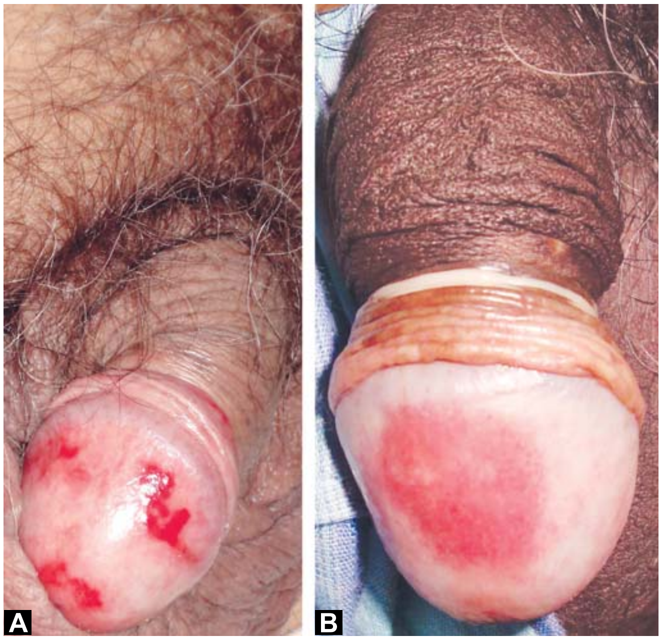

Figs 26.162A and B: Erythroplasia of glans penis. It is a premalignant condition.

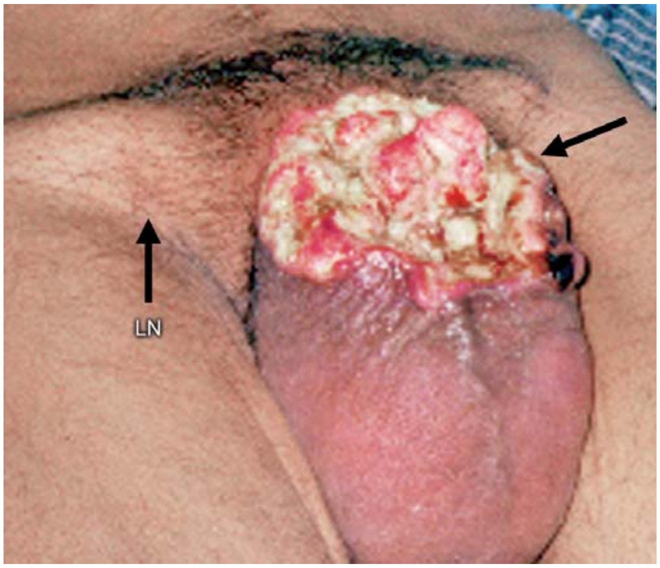

Fig. 26.163: Carcinoma of penis causing destruction of entire penis.

- *Papilliferous type* eventually attains a large size forming a fungating foul smelling lesion which often gets infected.
- *Ulcerative type*.
  - Glans penis is the most common site (coronal sulcus for basal cell carcinoma).
  - 80% are of low grade tumours.

Our senses don't deceive us; our judgement does.

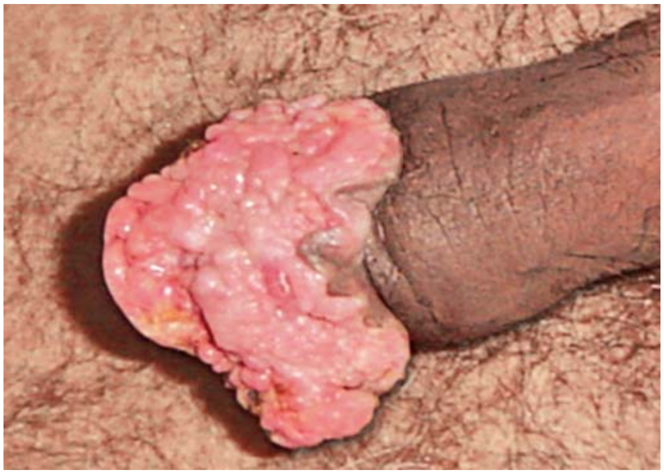

Fig. 26.164: Carcinoma penis with proliferative lesion.

Spread

- Through lymphatics, it spreads to the horizontal group of inguinal lymph nodes which become nodular and hard. Lymph nodes on both sides can get involved. Later, external iliac group are involved (above and on medial aspect of the inguinal ligament).
- Once inguinal lymph nodes are fixed, it causes severe excruciating pain and lymphoedema. *Fixed lymph node status indicates the advancement of the disease.* It may erode into the femoral vessels causing torrential haemorrhage and death. Fungation can occur.
- From glans, it also spreads to Cloquet lymph node which is located in femoral canal.
- Carcinoma from shaft of penis can spread directly to the external iliac lymph nodes.
- It spreads proximally to the body of penis causing induration.
- Urethral meatus may get involved causing alteration in urinary stream. *It is a locoregional malignant disease.*
- *Blood spread is rare*.

Clinical Features

- In an adult, recent onset of phimosis should raise the suspicion of carcinoma penis.

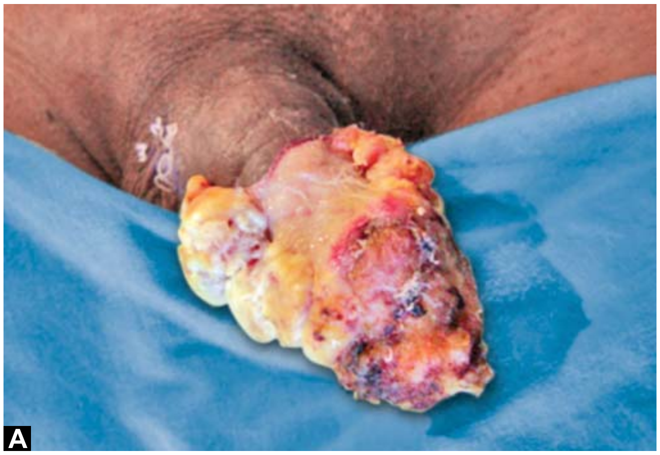

Fig. 26.165A

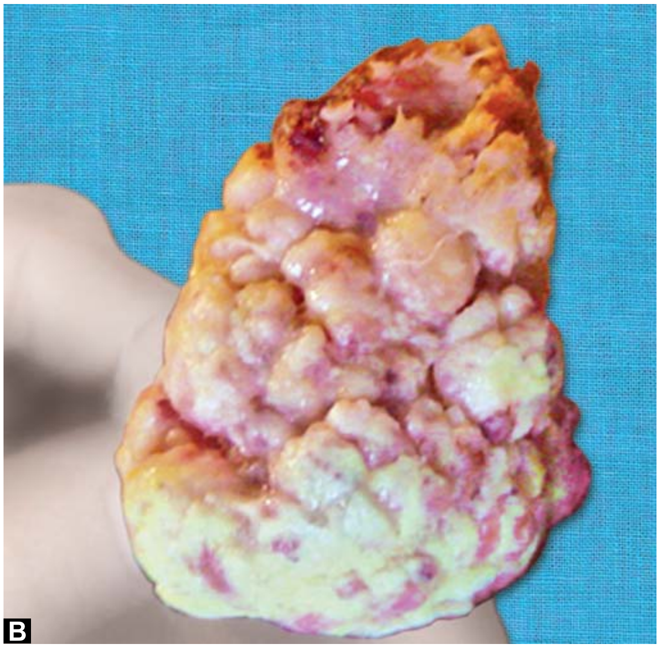

Fig. 26.165B

Figs 26.165A and B: Verrucous carcinoma of penis.

- Lesion is painless initially but later becomes painful due to secondary infection, often accompanied by discharge which is foul smelling, purulent and irritating.
- Altered urinary stream.
- Fungation and induration, everted edge, often extending into the body of penis.
- Palpable hard, nodular inguinal lymph nodes on both sides may be present. External iliac lymph nodes may be palpable.
- Pain, oedema, tenderness, redness develops once infection occurs.

Investigations

1. *Edge biopsy* from the lesion shows squamous cell carcinoma with epithelial pearls.

Broder's grading

- ❖ Very well differentiated (75% epithelial pearls)
- ❖ Well differentiated (50-75%)
- ❖ Moderately differentiated (25-50%)
- ❖ Undifferentiated (< 25%)

2. *FNAC of lymph nodes* (No open biopsy for lymph nodes).
3. U/S abdomen, to look for involvement of external iliac lymph nodes.
4. *SLNB*—**Cabana sentinel node** is located above and medial to the junction of saphenous and femoral vein. It is the first node to get involved in carcinoma penis. So this Sentinel Lymph Node Biopsy (SLNB) after isosulphan blue dye injection in to the primary is done to decide for the necessity for ilioinguinal block dissection.

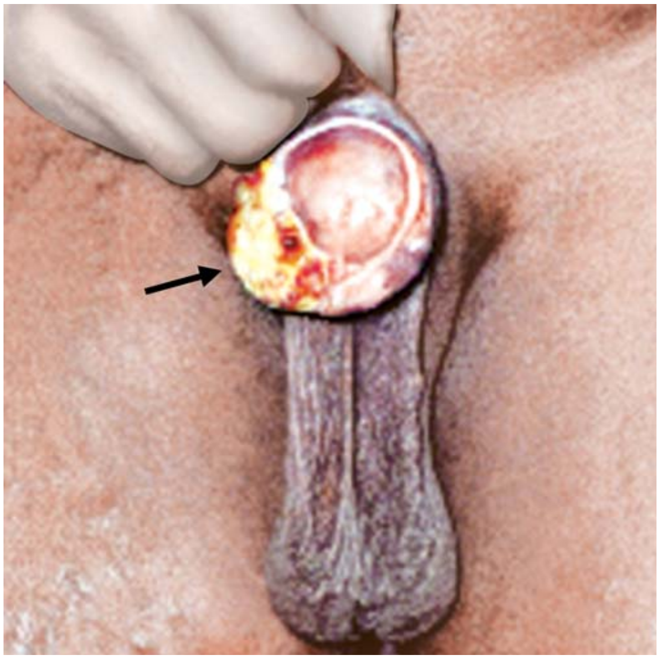

Fig. 26.166: Carcinoma of penis.

Staging of Carcinoma of Penis

Jackson's staging of carcinoma penis

Stage I – Tumour involving only glans/prepuce/both	90% five-year survival
Stage II – Tumour extending into body of penis	70%
Stage III – Tumour having mobile inguinal nodes	50%
Stage IV – Tumour spreading to adjacent structures/fixed nodes	5%

TNM staging

T_0 – No primary tumour
Tis – Carcinoma *in situ*
T_1 – Tumour < 2 cm without deep invasion
T_2 – Tumour between 2-5 cm with minimal deep invasion
T_3 – Tumour > 5 cm with deep invasion/urethral spread
T_4 – Tumour spread to adjacent tissues
N_0 – No nodal spread
N_1 – Mobile regional nodes—unilateral
N_2 – Mobile regional nodes—bilateral
N_3 – Fixed regional nodes
M_0 – No distant spread
M_1 – Distant spread present

Treatment

- If growth involves the glans without extending into the proximal part of shaft of the penis, then *partial amputation of the penis* is done. A length of 2.5 cm stump is retained. Clearance of 2 cm from the proximal extended part of the tumor is needed. Advantage is proper streaming of the urine is possible.
- Partial amputation of penis with bilateral ilioinguinal lymph node block dissection is called as *Young's operation.*

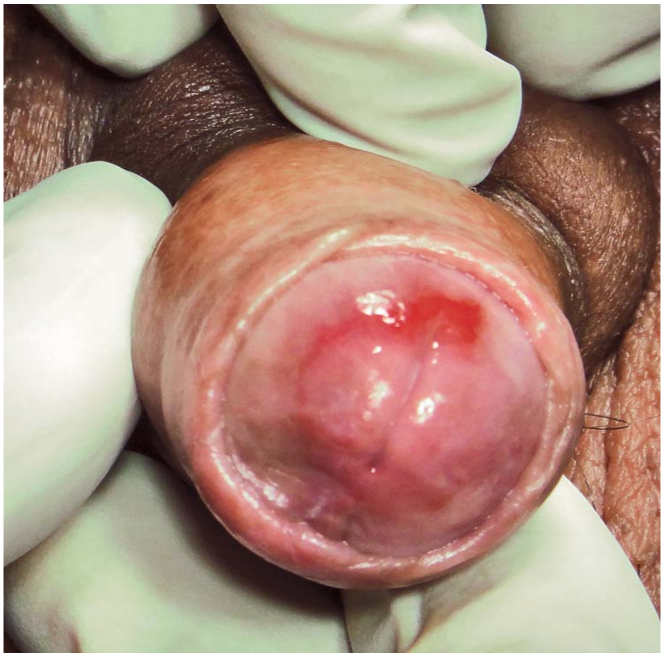

Fig. 26.167: *In situ* carcinoma penis over the glans.

- If tumour involves the proximal part of the body of penis or if it is anaplastic/poorly differentiated tumour *total amputation of penis* is done with *perineal urethrostomy*. Problems with perineal urethrostomy are scrotal ammoniacal dermatitis and stricture at urethrostomy site. Dermatitis is prevented by asking the patient to urinate in sitting position lifting the scrotum upwards. Stricture needs dilatation.

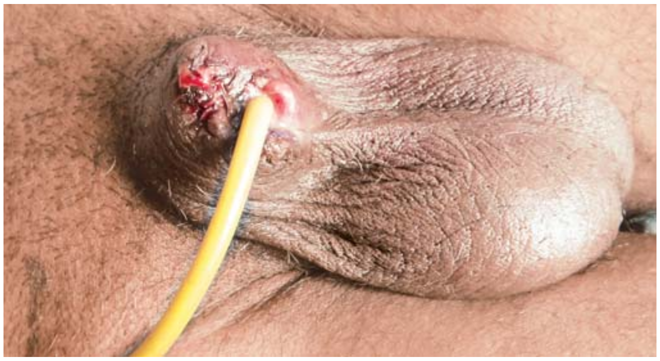

Fig. 26.168: Partial amputation of penis with catheter *in situ*.

- Total scrotectomy with orchidectomy is done along with total amputation of the penis—*Sir Piersey Gold operation*. It prevents frequent dermatitis of the scrotal skin because of the perineal urethrostomy and also reduces the sexual desire.
- In case of carcinoma *in situ*, T_1 lesion of glans penis or well differentiated tumour in young individual, circumcision and curative radiotherapy to the penis can be given using radioactive tantalum wire implantation (6000 cGy in 7 days) or by wearing radium penile mould continuously or intermittently (6000 cGy in 7 days) or by linear accelerator external beam radiotherapy (6000 cGy in 5 weeks). Involve-ment of nodes in these patients is less than 10%.

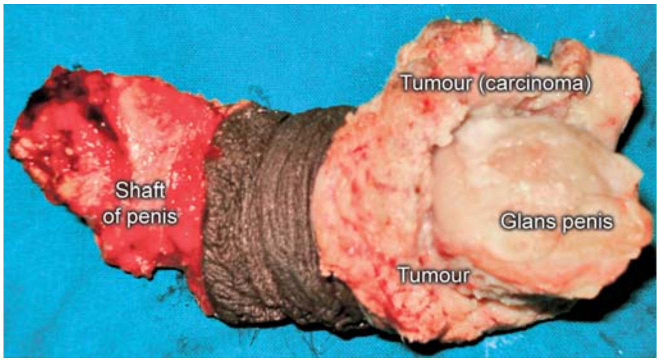

Fig. 26.169: Specimen of total amputation of penis done for carcinoma of penis.

- ♦ When lymph nodes are involved and are mobile, *bilateral ilioinguinal nodal dissection* is done.
  - *Primary inguinal block* is doing block dissection within 4 weeks of surgery for primary tumour.
  - *Secondary inguinal block* is doing block dissection after 4 weeks of surgery for primary disease. Only 50% of palpable inguinal nodes are involved by metastatic spread. So often a course of antibiotic is given and waited for 4-6 weeks. Complications of ilioinguinal block dissection are *flap necrosis*, lymphoedema of lower limb, femoral blow out, infection, lymphorrhoea and haemorrhage.
  - If primary tumour is poorly differentiated, and if tumour is T_2 or above, chances of inguinal nodal spread is more than 50% and so a prophylactic inguinal nodal dissection is done.
  - Often, involvement of inguinal nodes may be due to infection. So a trial of antibiotic therapy is given for 4-6 weeks to reduce the size of the inguinal node.
  - Removal of iliac nodes does not alter the outcome. It is done to confirm the spread, so that further therapy can be planned and prognosis can be predicted.
- ♦ Postoperative radiotherapy to inguinal region is often given.
- ♦ In advanced fixed inguinal nodes palliative external radiotherapy is given. It is to palliate pain, fungation and anticipated erosion into femoral vessels.
- ♦ Topical 5 FU cream or Nd:YAG laser photoirradiation is useful in carcinoma *in situ*.
- ♦ *Chemotherapy:*
  5 FU; methotrexate, bleomucin, cisplatin and vincristine. MBP/VBM combinations are used.
  Bleomycin is a radiosensitiser and so beneficial if RT is planned later.

Note:

- *Dresslers quadrangle*—upper border is formed by line joining anterosuperior iliac spine and pubic tubercle; laterally line joining anterosuperior iliac spine and a point 20 cm below it; medially pubic tubercle and a point 15 cm below it. Nodal block dissection for carcinoma penis should cover this area adequately.
- *Elective prophylactic inguinal block* is done in high-risk group—invasive carcinoma; T_2 and T_3 tumours; with vascular invasion.
- *Therapeutic inguinal block* is done whenever FNAC of node shows positive tumour.
- *Superficial lymph node block* is dissection superficial to fascia lata in N_0 disease. It should undergo frozen section biopsy. If nodes are positive, then formal block dissection should be done.
- *Standard inguinal lymphadenectomy*—is classical block dissection.
- *Modified inguinal block dissection* (Catalona)—small incision, limited dissection, preservation of saphenous vein.

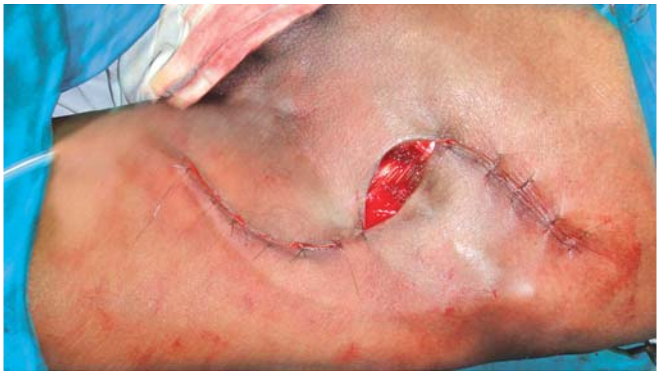

Fig. 26.170: S shaped incicion for ilioinguinal block dissection. Many modification incisions are also used.

Role of radiotherapy in carcinoma penis

- ❖ Carcinoma *in situ*
- ❖ Small lesion less than 2 cm
- ❖ Lesion confined to glans
- ❖ Small lesion in young individual
- ❖ Advanced inoperable disease
- ❖ Palliation to inoperable inguinal nodes in groin
- ❖ Postoperative RT

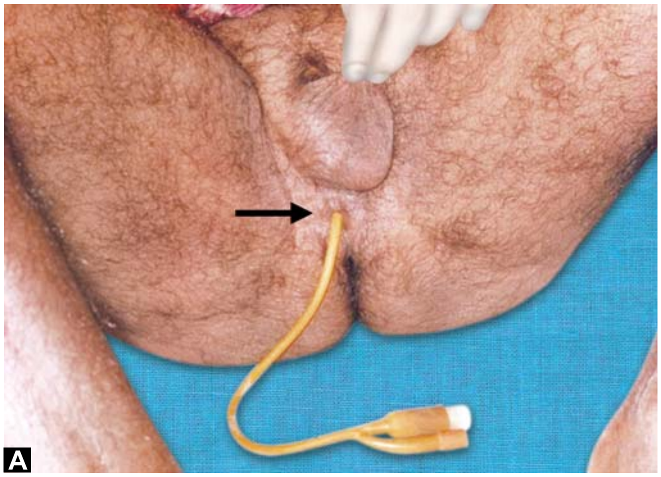

Fig. 26.171A

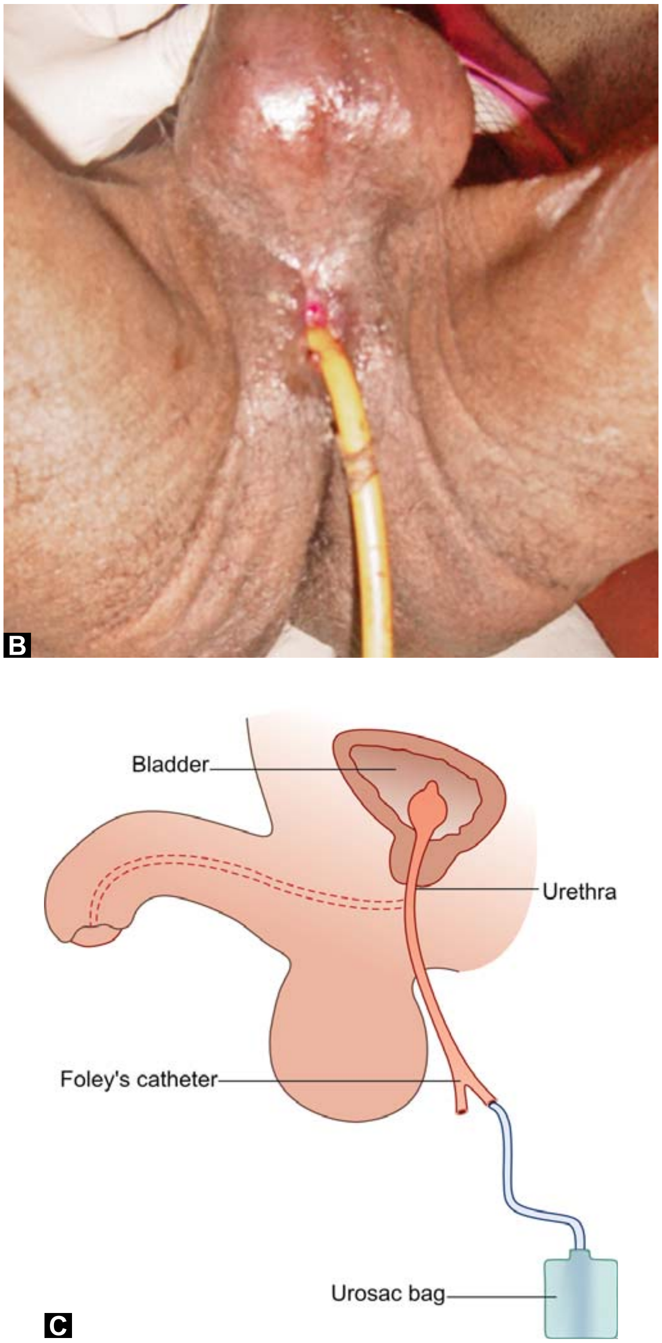
B
Bladder
Urethra
Foley's catheter
Urosac bag
C

Figs 26.171B and C

Figs 26.171A to C: Perineal urethrostomy is done for urethral injury, urethral surgeries like urethroplasties as diversion and after total amputation of penis in carcinoma penis.

Complications of inguinal block dissection

- Haemorrhage
- Lymphorrhoea
- Lymphoedema
- Infection
- Flap necrosis—*common*

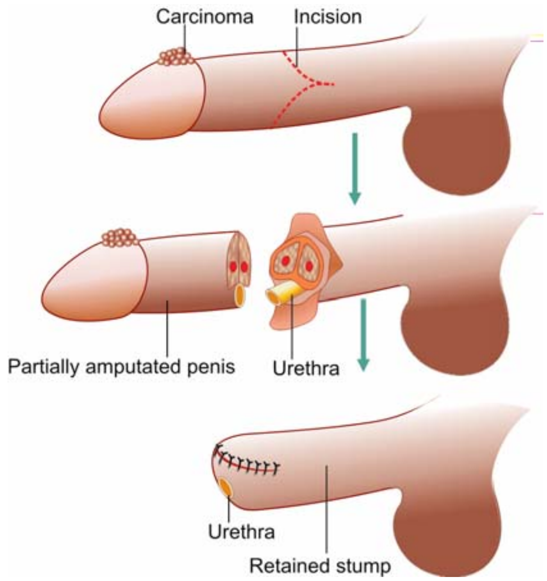
Carcinoma
Incision
Partially amputated penis
Urethra
Urethra
Retained stump

Fig. 26.172: Partial amputation of penis is done in penile carcinoma in the glans not extending into the body of the penis.

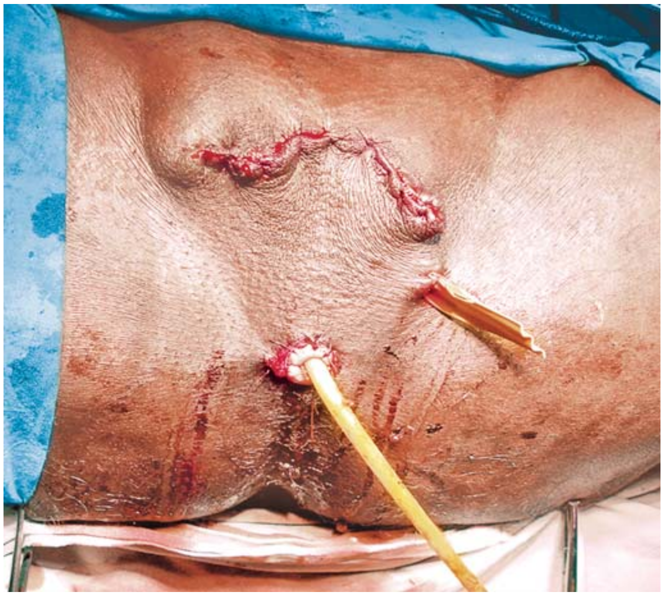

Fig. 26.173: Total penectomy for carcinoma penis. Note the drain placed; sutures; urethrostomy site.

BUSCHKE-LOWENSTEIN'S TUMOUR

It is verrucous carcinoma of penis (5-15% common).

- *It is a curable malignancy*.
- It is locally destructive, locally invasive.
- It is often large, exophytic, dry, verrucae-like growth.
- Neither spreads through lymphatics nor blood.
- After biopsy and confirmation, surgical excision or partial amputation is the treatment of choice.
- Radiotherapy should not be given**.**
- Virus aetiology is proposed—HPV—6 and 11.
- Systemic interferon α therapy and Nd:YAG laser therapy is successful.

F. Scrotum

CHAPTER OUTLINE

- Anatomy
- Fournier's Gangrene
- Hydrocele
  - Primary Vaginal Hydrocele
  - Secondary Hydroscele
- Haematocele
- Pyocele
- Cyst of Epididymis
- Spermatocele
- Varicocele

ANATOMY

It is a cutaneous bag containing the right and left testis, epididymis and lower part of spermatic cord.

Layers (from without inwards)

1. Skin
2. Dartos muscle
3. External spermatic fascia
4. Cremasteric fascia
5. Internal spermatic fascia

Blood Supply

From external pudendal, deep external pudendal and internal pudendal arteries.

Nerve Supply

- Anterior 1/3rd portion by ilioinguinal nerve and genital branch of genitofemoral nerve.
- Posterior 2/3rd portion by posterior scrotal branch of pudendal nerve and perineal branch of posterior cutaneous nerve of thigh.
- Dartos is supplied by sympathetic nerve.

FOURNIER'S GANGRENE

- It is also called as *idiopathic gangrene of the scrotum*.
- It is a vascular gangrene of infective origin, caused by haemolytic streptococci, microaerophilic streptococci, staphylococci, *E. coli*, *Cl.welchii, Bacteroides fragilis*.
- There will be fulminant inflammation of the scrotal skin and subcutaneous tissues resulting in *obliterative arteritis* of the arterioles of the scrotal skin leading into cutaneous gangrene.
- It is common in diabetics, old age, malnourished, immunosuppressed individuals.

Clinical Features

- Condition is common in old age.
- Sudden pain in the scrotum, fever, severe toxicity.
- Very fast spreading cellulitis of scrotal skin, extending to the groin and often to anterior abdominal wall.
- Extensive skin sloughing occurs *leaving normal testis exposed.*
- Sometimes toxicity is so severe that they may go for renal failure and other complications.
- Sometimes the condition may worsen rapidly leading to death (25% mortality).

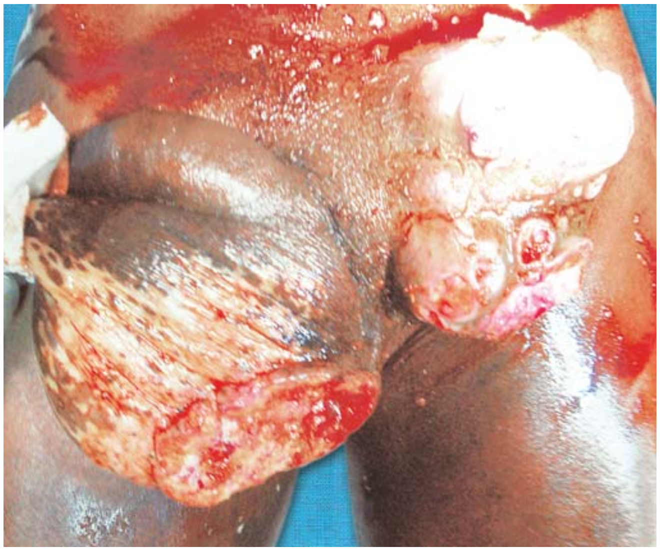

Fig. 26.174: Carcinoma of scrotum with inguinal node secondaries.

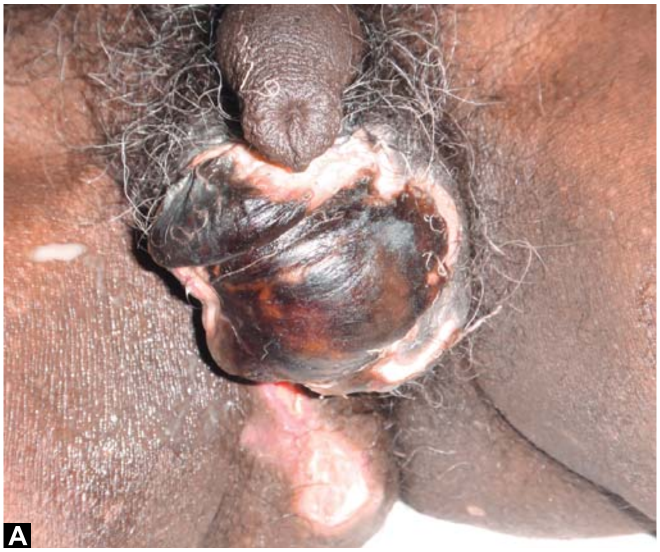

Fig. 26.175A

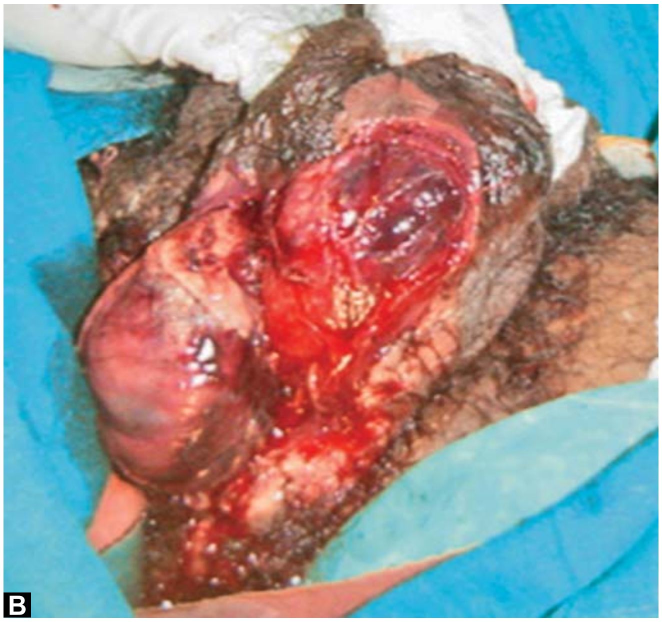

Fig. 26.175B

Figs 26.175A and B: Fournier's gangrene. Testes are normal. It is an infective gangrene of scrotum, rapidly spreading.

Treatment

- Treated as an in-patient always.
- IV fluids and catheterisation—for maintenance of urine output.
- Antibiotics, blood transfusion.
- Nutritional support (TPN, enteral).
- Liberal excision of all slough.
- Once patient recovers and wound granulates well, *skin grafting is done*.
- ***Orchidectomy is not necessary as testis is normal and viable.***
- Testis can be placed in the pouch in medial aspect of the thigh.

HYDROCELE

It is the collection of fluid between the two layers of tunica vaginalis of testis.

Types

- Congenital
- Acquired
  - Primary
  - Secondary

Aetiology

- Defective absorption of fluid by the tunica vaginalis, probably due to damage to the endothelial wall by low-grade infection.
- Excessive production of fluid as in secondary hydrocele.
- Interference with drainage of fluid by lymphatic vessels of the cord.
- Communication with the peritoneal cavity.

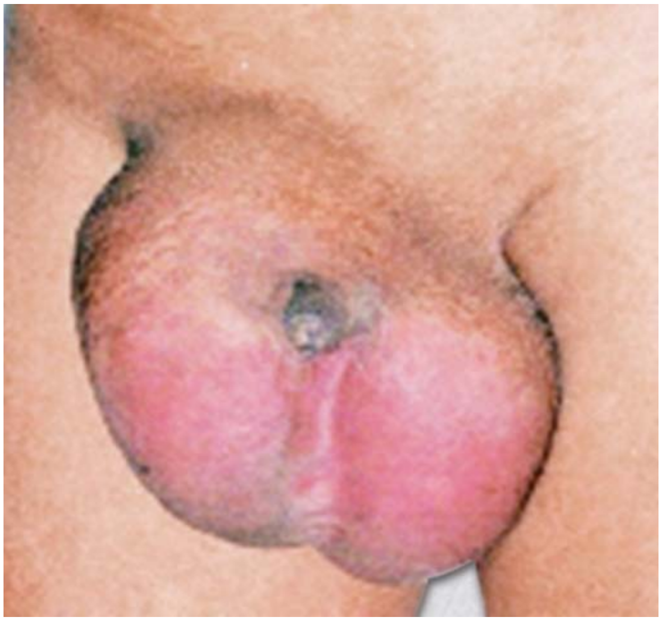

Fig. 26.176: Bilateral hydrocele.

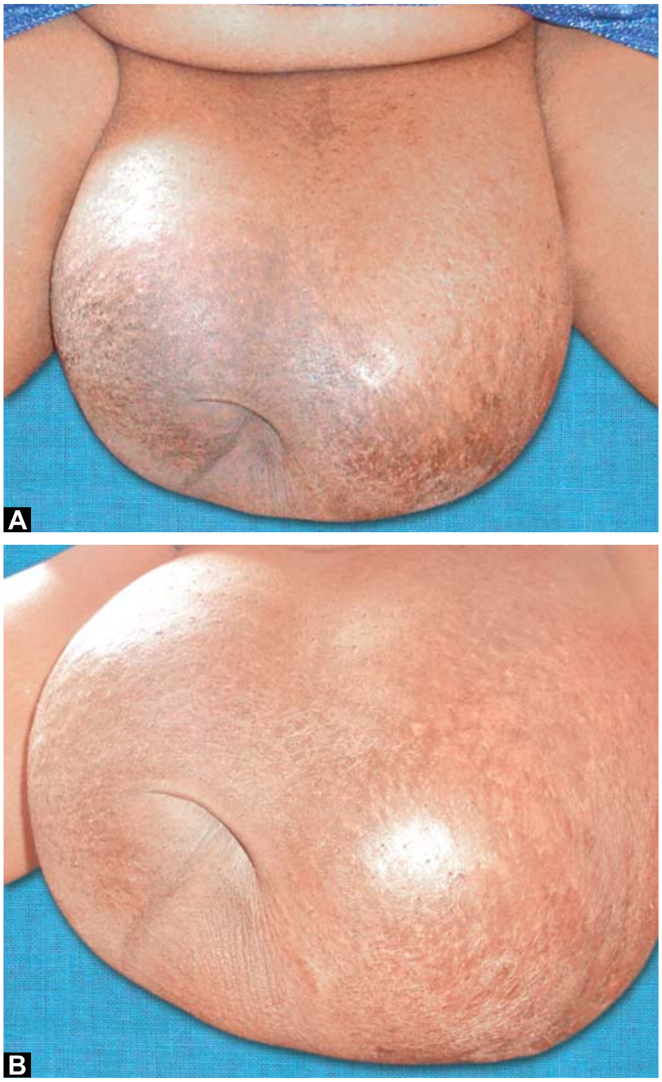

Figs 26.177A and B: Massive hydrocele wherein penis is buried completely within the hydrocele area.

Hydrocele fluid is amber-coloured with specific gravity of 1.022 to 1.024. It contains water, salts, albumin, fibrinogen. Per se, hydrocele fluid does not clot, but if it comes in contact with the blood, fibrinogen gets activated and clots firmly. Often fluid contains cholesterol and tyrosine crystals.

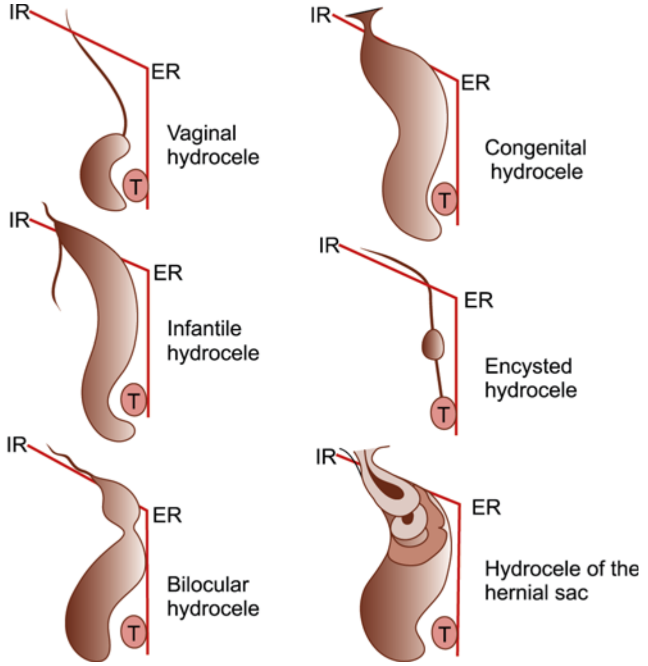

Fig. 26.178: Types of hydrocele.

Cysts which contain cholesterol crystals

- Vaginal hydrocele
- Branchial cysts
- Dentigerous cyst

PRIMARY VAGINAL HYDROCELE

- Occurs in middle-aged, common in tropical countries.
- Testis is not palpable as it usually attains a large size (unlike secondary hydroceles which are small, except in filarial hydrocele).
- *Fluctuant* (elicited by, fixing the hydrocele with hand and feeling for the fluid movement using fingers placed in *two perpendicular directions*).

Note:

A relaxed muscle can demonstrate fluctuation in one direction even though there is *no fluid* in it.

- *Initially transilluminant* (elicited in front of the swelling, side to side), but long-standing hydrocele is nontransilluminant (due to thickened dartos, thickened spermatic fascia, thickened hydrocele sac, infected content, chylous fluid, often filarial hydrocele, haematocele).
- Can *get above the swelling* (you can feel only cord structures and nothing else at the root of the scrotum, unlike in hernia).
- Testicular sensation can be elicited in vaginal hydrocele by transmitting the pressure sensation through the fluid.

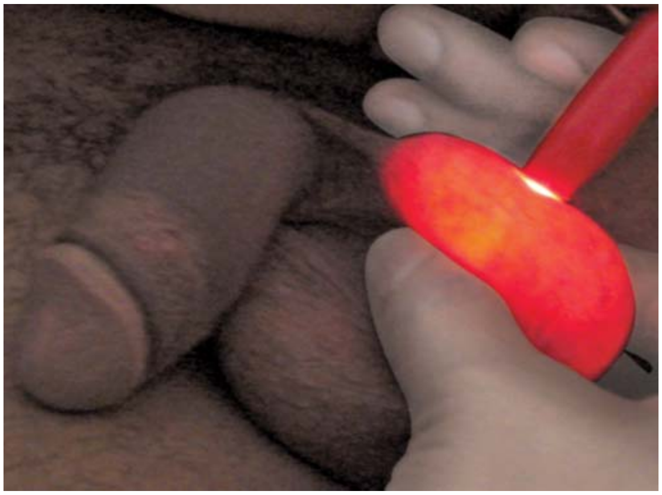

Fig. 26.179: Hydrocele which is brilliantly transilluminant.

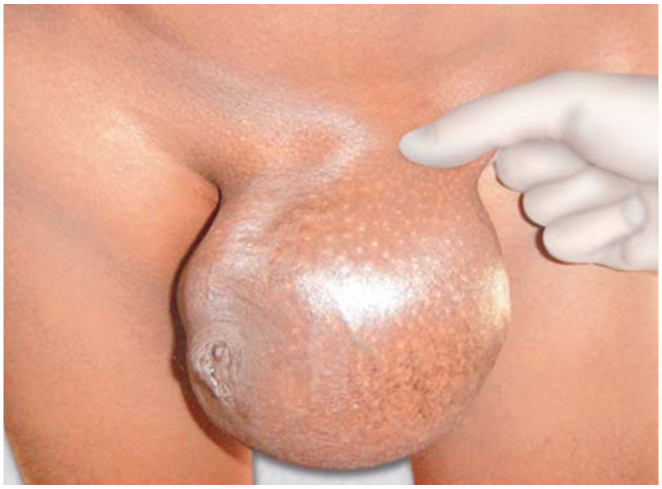

Fig. 26.180: Get above the swelling is possible in case of hydrocele.

Swellings which are brilliantly transilluminant

- Vaginal hydrocele
- Epididymal cyst
- Cystic hygroma
- Ranula
- Meningocele

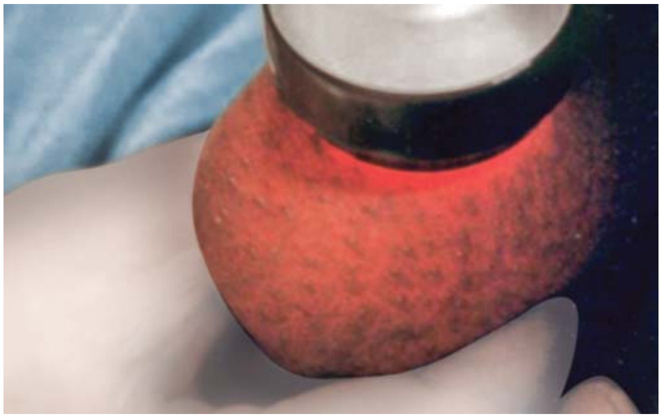

Fig. 26.181: Hydrocele is transilluminant. It may be negative in long-standing hydrocele, thick sac, thick dartos, thick skin and infected fluid or chylocele.

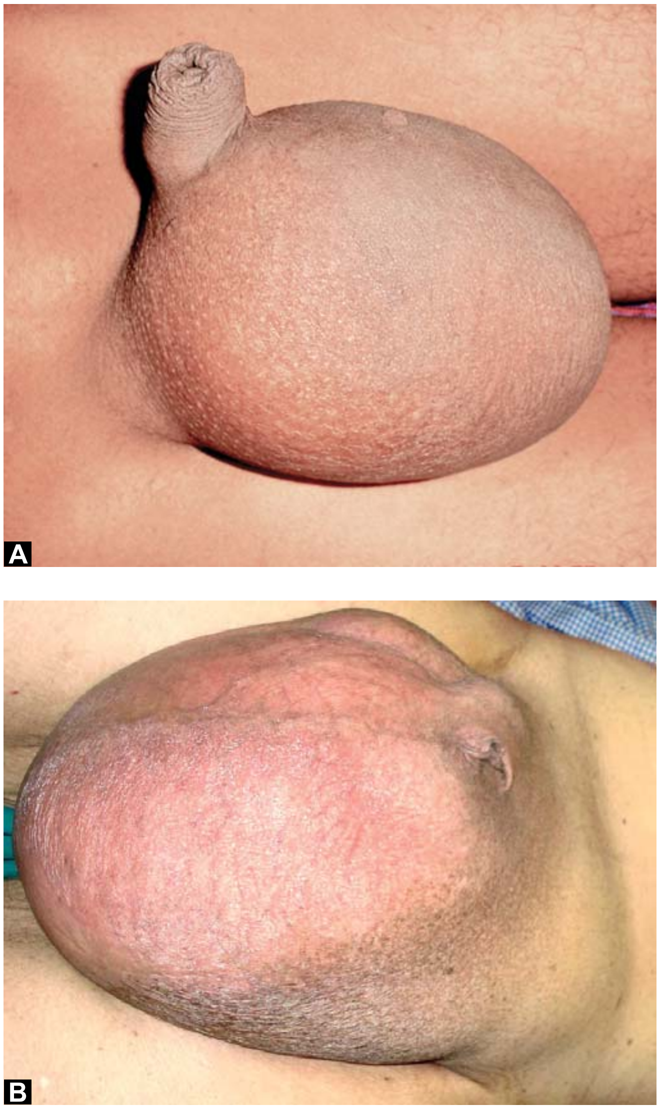

Figs 26.182A and B: Hydrocele in two different patients. One is average sized; another is large and penis is buried into the hydrocele swelling.

Infantile Hydrocele

Here tunica and processus vaginalis (hydrocele) are distended up to internal ring, but sac has no connection with the general peritoneal cavity.

Congenital Hydrocele

- Processus vaginalis communicates with the peritoneal cavity.
- As this communicating orifice is too small, bowel does not descend and so hernia usually will not develop. While lying down, fluid disappears gradually and while standing fluid recollects.
- Hydrocele cannot be emptied by digital pressure as it causes "*inverted ink bottle*" effect. Ascites, tuberculous peritonitis are the aetiologies for the same.

Encysted Hydrocele of the Cord

- It is the *fluid collection* in a portion of patent funicular process part of the tunica vaginalis; but closed above and below; located in inguinal/inguinoscrotal/scrotal part which is fluctuant and transilluminant.
- On gentle traction to the testis, the swelling becomes less mobile (*traction test*).
- Differential diagnosis: Epididymal cyst, inguinal hernia, lipoma of cord, varicocele.
- Treatment is excision under local anaesthesia.

Hydrocele-en-bisac (Bilocular Hydrocele)

Hydrocele has got two intercommunicating sacs, one above and one below the neck of the scrotum. Upper one lies superficial or in the inguinal canal or may insinuate itself between the muscle layers—*cross-fluctuant.*

Other conditions where cross-fluctuation is elicited:

- Plunging ranula
- Compound palmar ganglion
- Psoas abscess

Hydrocele of the Canal of the Nuck

It occurs in females, in relation to the round ligament, always in the inguinal canal.

Hydrocele of the Hernial Sac

It is due to adhesions of the content of hernial sac. Fluid secreted collects in the hernial sac and forms hydrocele of the hernial sac.

SECONDARY HYDROCELE

Causes

Infection:	Filariasis Tuberculosis of epididymis—30% cases have secondary hydrocele Syphilis
Injury:	Trauma, postherniorrhaphy hydrocele
Tumour:	Malignancy

Secondary hydrocele rarely attains large size

It is usually small, lax and testis is usually palpable (unlike primary hydrocele). *Exception is, secondary hydrocele due to filariasis.* It can be very large.

Postherniorrhaphy Hydrocele

It is a secondary hydrocele occurring after the surgery for inguinal hernia. It is due to the damage to lymphatic vessels of the tunica vaginalis and is 0.2% common. It is treated like any hydrocele but usually after about 6 months.

A clear conscience can bear any trouble.

Filarial Hydrocele and Chylocele

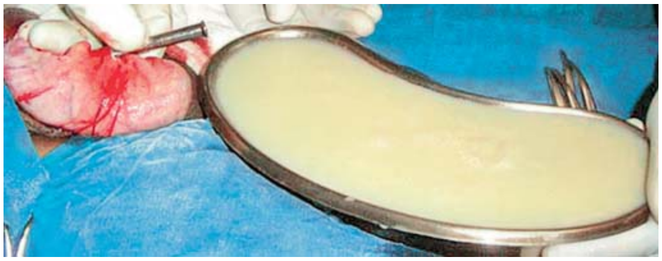

Fig. 26.183: Chylocele.

- Occurs commonly in coastal region, and in and around the equator.
- Usually occurs after repeated attacks of filarial epididymitis.
- Hydrocele is usually of large size and the sac is thickened.
- Fluid contains fat, rich in cholesterol, and is derived from ruptured lymph varix into the tunica.
- It is often difficult to differentiate from primary hydrocele.

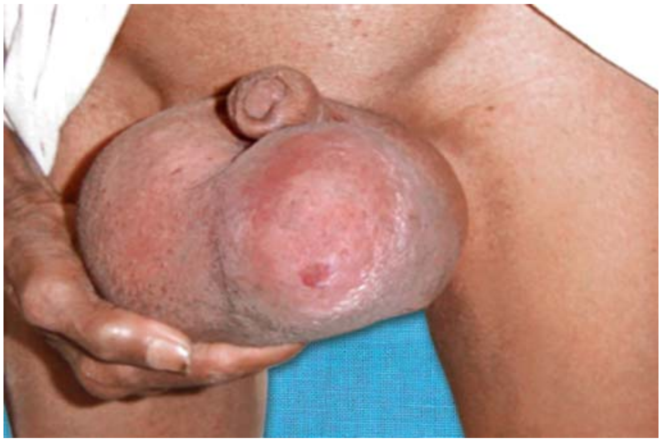

Fig. 26.184: Hernia of hydrocele is a rare complication of hydrocele. Herniation occurs through dartos.

Complications of hydrocele

- Infection
- Pyocele
- Haematocele
- Atrophy of testis, hernia of hydrocele (rare)
- Infertility
- Hernia of hydrocele sac (rare)

Differential diagnosis

- Inguinal hernia
- Epididymal cyst
- Spermatocele
- Testicular tumour
- Scrotal oedema

Treatment for Hydrocele

Surgery

- Sub-total excision
- Partial excision and eversion (Jabouley's operation)
- Evacuation and eversion
- Lord's plication

Procedure

- Under G/A or spinal or L/A, after cleaning and draping, vertical incision of about 6-8 cm in length is made over the scrotum, anteriorly 1 cm lateral to the median raphe.
- Skin, dartos, external spermatic fascia, internal spermatic fascia are incised.

  Bluish hydrocele sac is identified, i.e. parietal layer of the tunica vaginalis of testis.

  Fluid is evacuated using trocar and cannula. Sac is opened.

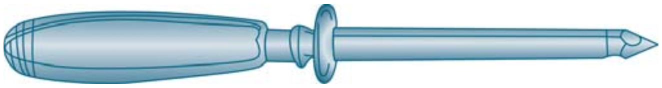

Fig. 26.185: Trocar and cannula used to evacuate the hydrocele fluid.

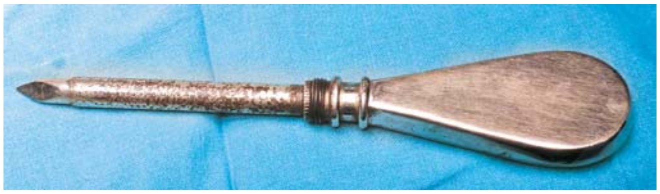

Fig. 26.186: Trocar and cannula used for hydrocele surgery.

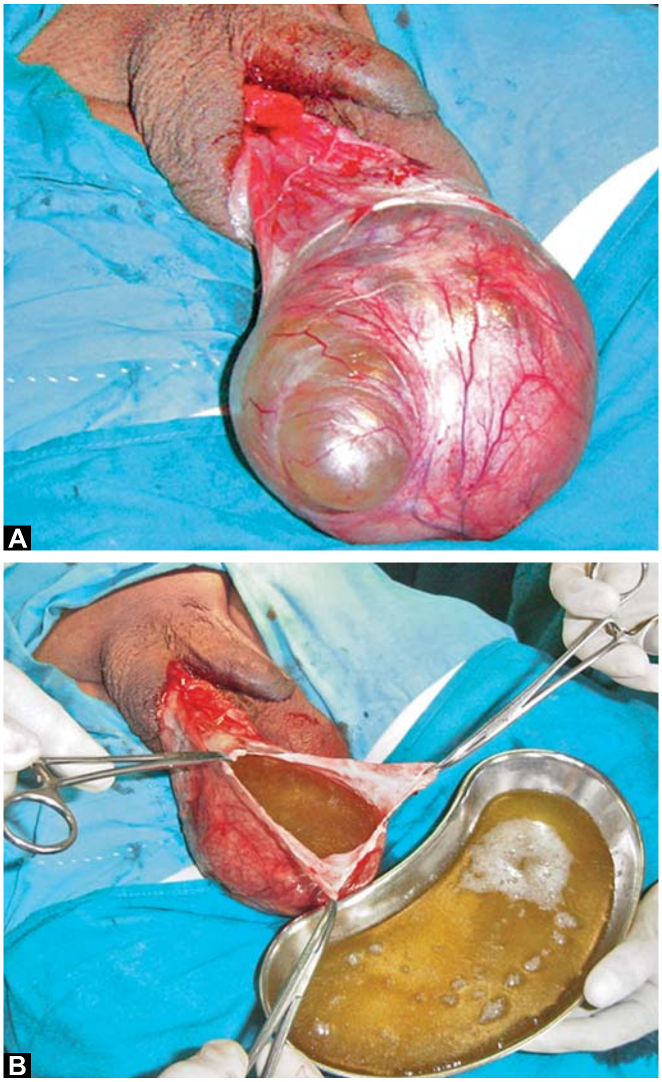

Figs 26.187A and B

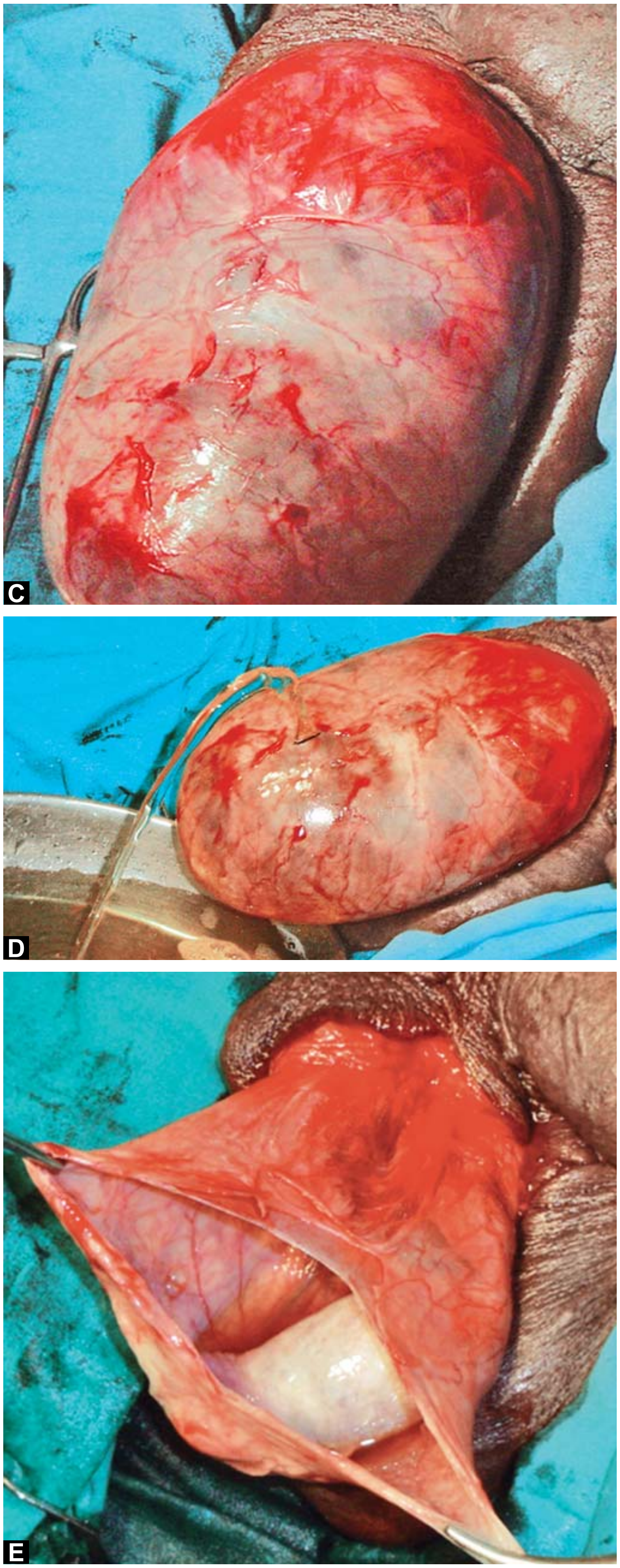

Figs 26.187C to E

Figs 26.187A to E: Bluish hydrocele sac and hydrocele fluid in the sac.

- *If the sac is small, thin and contains clear fluid,* either *Lord's plication*, i.e. tunica is bunched into a "ruff" by placing series of multiple interrupted chromic catgut sutures so as to make the sac to form fibrous tissue (It is relatively avascular and so haematoma will not occur).
- Or *evacuation and eversion* of the sac behind the testis (after eversion, everted sac is sutured with chromic catgut by continuous sutures) is done.
- *If the sac is thick, in large hydrocele and chylocele***,** *subtotal excision of the sac* is done (as tunica vaginalis is reflected on to the cord structures and epididymis posteriorly, total excision leads to orchidectomy with division of cord).
- Often the sac is excised partially and then eversion is done, which is called as *Jabouley's* operation.
- After evacuation, the sac with the testis is placed in a newly created pocket between the fascial layers of the scrotum *(Sharma and Jhawer's technique).*
- *Aspiration must be avoided as much as possible* as it is only a temporary measure (recurrence occurs very early) and chances of haematocele, infection are higher.
- A drain is placed near the root of the scrotum on the lateral aspect because it becomes the most dependent portion, when scrotal support is given. Scrotal support is given to reduce the scrotal oedema.
- Wound is closed in layers.
- Drain is removed in 48 hours.

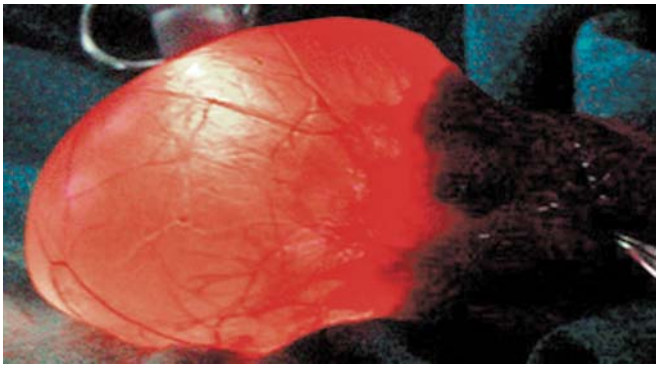

Fig. 26.188: Transillumination of hydrocele sac during surgery just prior to placing the trocar for evacuation after exposing the sac.

Surgeries for hydrocele

- Subtotal excision of the sac
- Jabouley's operation
- Evacuation and eversion
- Lord's plication
- Sharma and Jhawer's technique

Complications of surgery

- Reactionary haemorrhage
- Infection
- Pyocele
- Sinus formation
- Recurrent hydrocele

> *In early testicular tuberculosis there is often a loss of cutaneous elasticity as shown by smoothening out of rugae consequent upon wasting of the cellular tissue beneath the dermis.* **—A Clifford Morson**

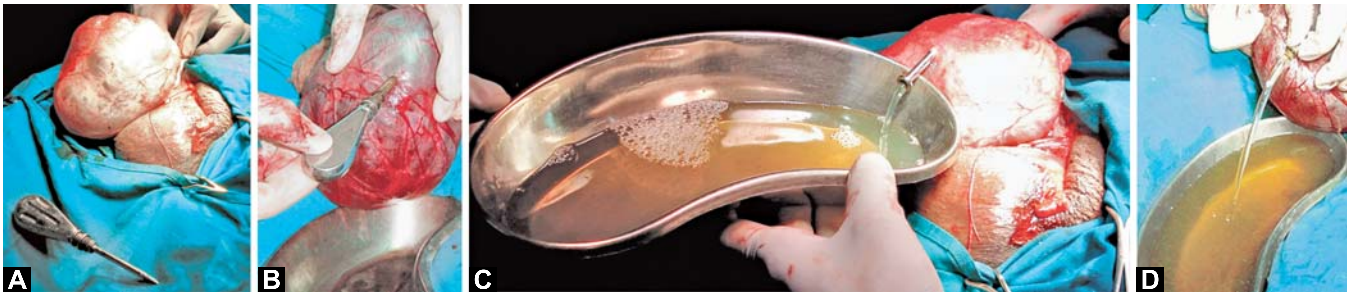

Figs 26.189A to D: Hydrocele surgery showing sac; trocar and cannula; its insertion and evacuation of fluid.

Conditions which cause loss of testicular sensation

i. Testicular tumour
ii. Lepra orchitis
iii. Syphilitic orchitis
iv. Chronic haematocele

Conditions where orchidectomy is done in hydrocele

- Pyocele with testicular destruction
- Clotted haematocele with testicular destruction

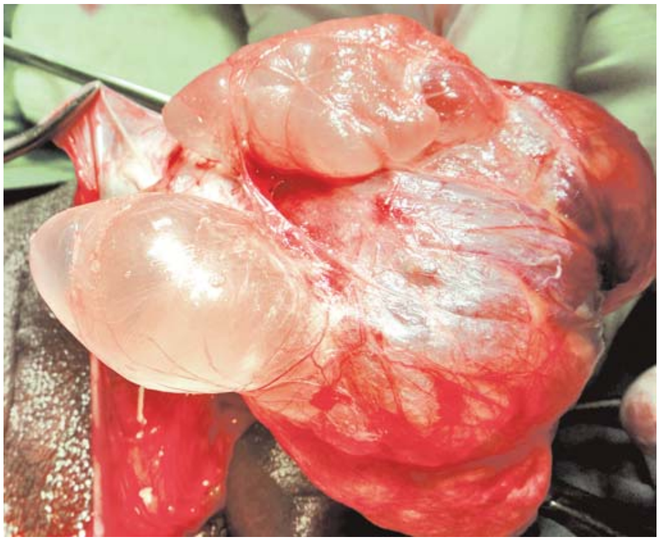

Fig. 26.190: Multiloculated hydrocele—on table look.

HAEMATOCELE

Types

a. Recent Haematocele

- It is due to rupture of one of the vessels in the tunica causing bleeding into the sac.
- Often it may occur following aspiration of a hydrocele.
- It may be precipitated by trauma also.

Clinical features:

- Sudden onset of pain and swelling after an history of trauma.
- It is tender, warm, fluctuant, but nontransilluminant.
- Occasionally aggressive testicular tumour mimics the presentation of acute, recent haematocele.
  - U/S of scrotum is done in such suspected cases to rule out neoplasm and also to find out the viability of testis.

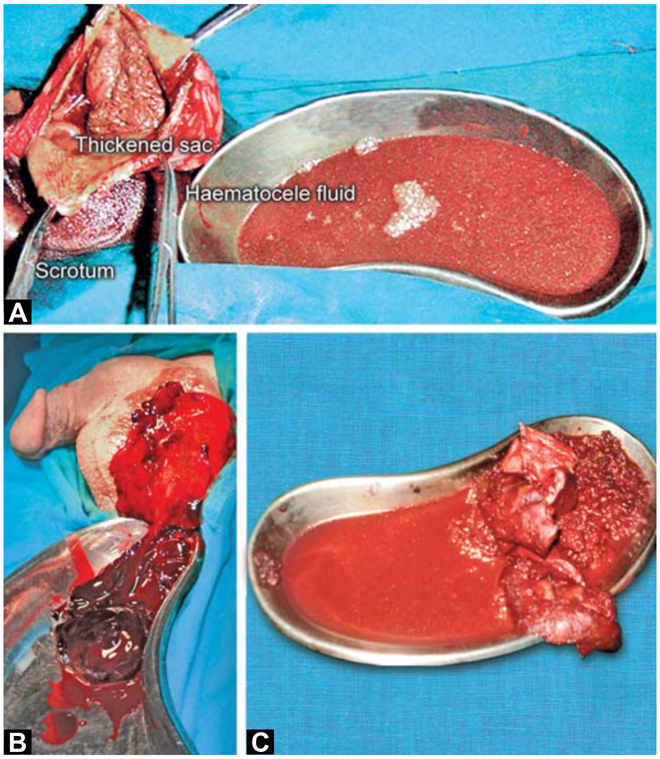

Figs 26.191A to C: Haematocele in different patients. Orchidectomy is often required in these patients if testis is not viable.

Complications

- Chronic (old clotted) haematocele
- Infection
- Pyocele

Treatment: The scrotum is explored under general anaesthesia. The clot is evacuated and the wound is closed with a drain.

b. Chronic or Old Clotted Haematocele

- It is usually due to slow, spontaneous haemorrhage into the tunica vaginalis, without any proper history of trauma.
- It is painless, hard, nontender, nonfluctuant, often calcified swelling, with *loss of testicular sensation*.
- Because of the constant pressure, testicular function and so testicular sensation is lost. It mimics testicular tumour in many aspect.
- U/S of the scrotum is done.
- **Treatment is** *orchidectomy,* as testis is functionless (Low orchidectomy through scrotal approach).

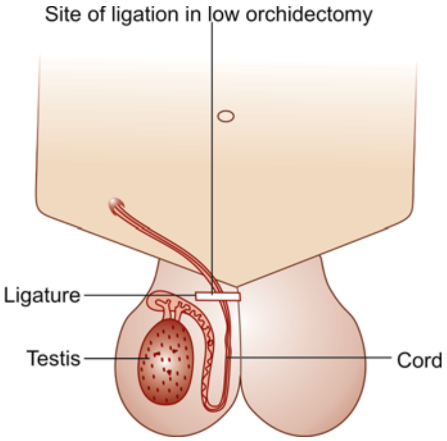

Fig. 26.192: Low orchidectomy is done in case of haematocele, pyocele. Here incision is in the scrotum. Cord is doubly ligated using thick chromic catgut.

PYOCELE

- It is collection of pus in the layers of tunica vaginalis.
- It can occur in a previously normal tunica or in a pre-existing haematocele or hydrocele (which gets infected).

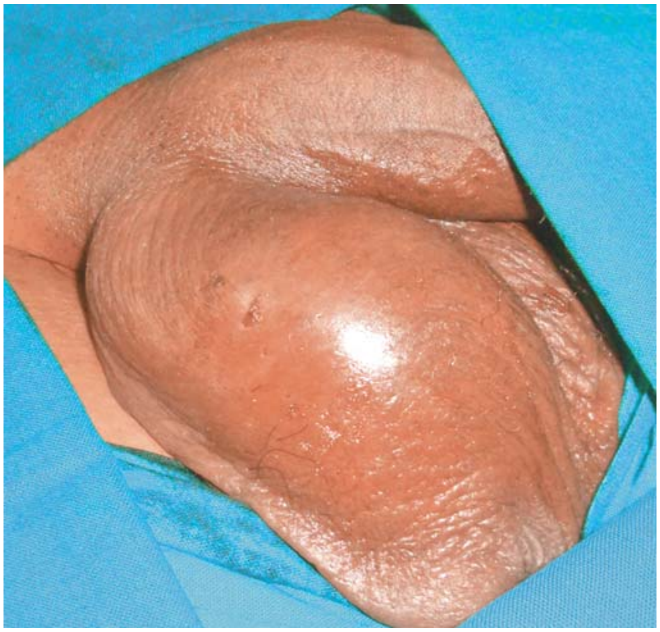

Fig. 26.193: Pyocele—clinical look. US confirms the diagnosis. TC will be raised.

Features

- Fever, toxicity, tender swelling in the scrotum, with scrotal wall oedema.
- Often in young individuals, it may be difficult to differentiate this from the torsion testis.
- Pus under tension eventually causes infective thrombosis of testicular vessels, leading to nonviability of the underlying testis or testicular gangrene.

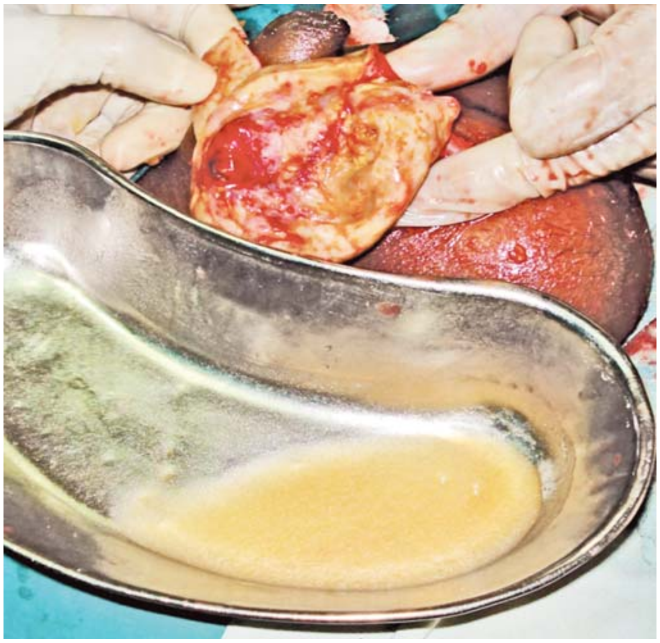

Fig. 26.194: Pyocele of scrotum. It is pus collection in the tunica vaginalis testis. It needs immediate exploration; drainage; often orchidectomy.

Treatment

- Antibiotics are started.
- Scrotum is explored immediately and pus is evacuated.
- Viability of the testis is checked.
- If viable, pus is evacuated and sent for C/S, wound is closed with a drain.
- If the testis is not viable, then *orchidectomy* is done (after taking consent).
- When in doubt, always testis is left in place and observed. Often, testis will show signs of viability. If not, then orchidectomy is done at a later stage.

CYST OF EPIDIDYMIS

It is due to the cystic degeneration of:

- Paradidymis (organ of Geraldes)—is the most common cause
- Appendix of the epididymis
- Appendix of the testis
- The vas aberrans of Haller

- Even though it is of congenital origin, it occurs in middle age.
- It is tensely cystic, contains clear fluid.
- Often bilateral.
- They are aggregation of number of small cysts and so *multiloculated.*
- They feel like '*bunch of tiny grapes*' situated behind the body of the testis.
- It is situated behind the body of testis.

In transposition of the viscera (situs inversus) the right testis hangs lower.—**C Allan Birch**

- Because of numerous septae they are finely tessellated and so are brilliantly transilluminant, appear like ***'chinese lantern pattern'.***

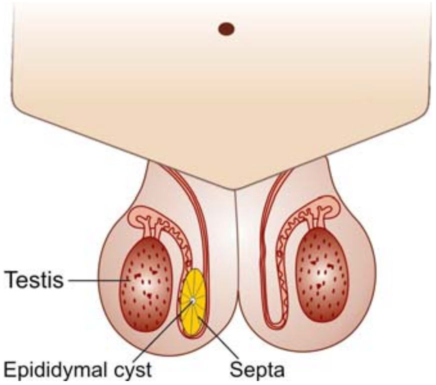

Fig. 26.195: Epididymal cyst showing multiloculated 'Chinese lantern' pattern. Its location is behind the body of testis.

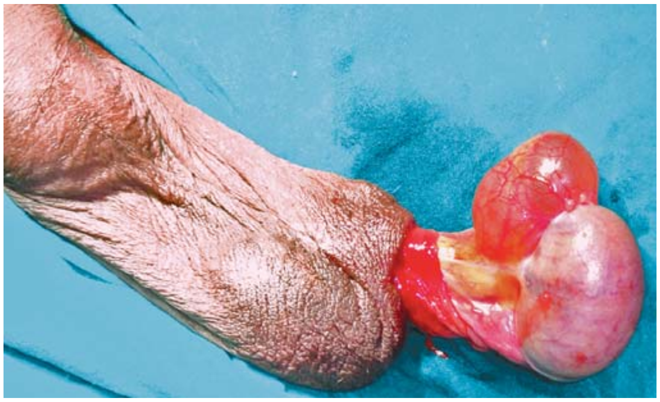

Fig. 26.196: Epididymal cyst.

Differential Diagnosis

- Spermatocele.
- Encysted hydrocele of the cord.

Treatment

- Avoid excision as much as possible as it results in infertility due to blockage.
- In old age, excision can be done.

SPERMATOCELE

It is a *unilocular, acquired retention cyst* derived from blockage of some portion of the sperm conducting mechanism of the epididymis.

- It is situated in the head of the epididymis, above and behind the body of the testis.
- Swelling contains *barley water like fluid,* which contains spermatozoa.
- It is soft, cystic and transilluminant. It is often considered by the patient as having an additional testis.
- Aspiration cytology confirms the diagnosis.

Treatment: It can be left alone. If it is large, excision is done.

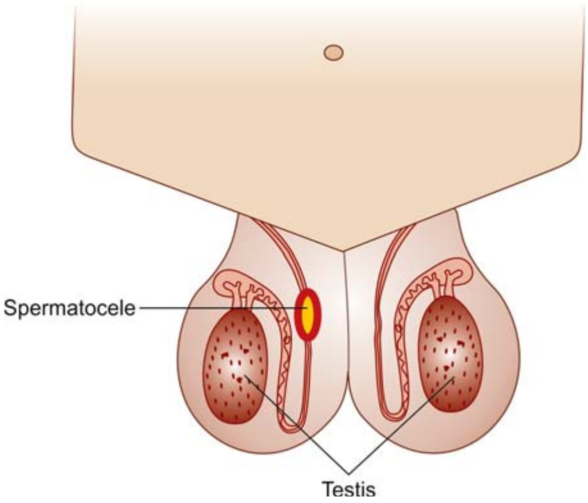

Fig. 26.197: Spermatocele occurs behind and above the testis. It is unilocular containing 'barley' like fluid.

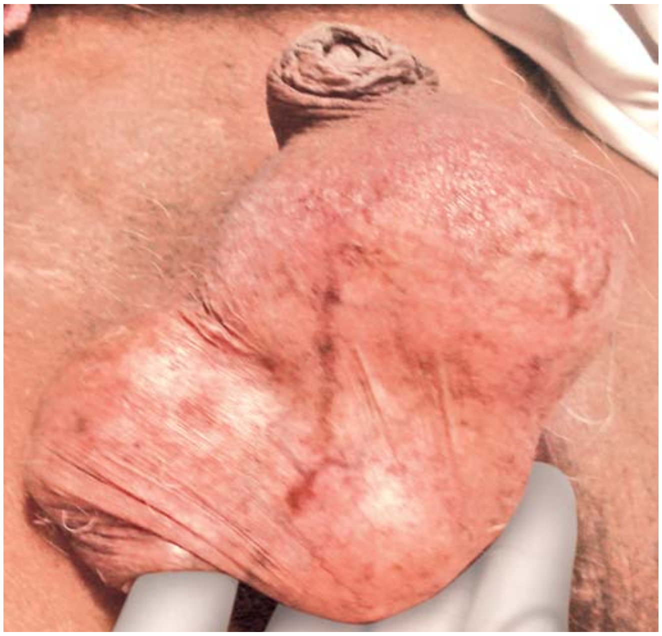

Fig. 26.198: Spermatocele. It was earlier mistakenly called as 3rd testis.

Differential Diagnosis

- Epididymal cyst.
- Hydrocele.

Epididymal cyst	***Spermatocele***
Congenital	Acquired retention cyst
Behind the body of the testis	Behind and above the testis
Multilocular	Unilocular
Bunch of grapes appearance	Looks like 3rd testis
Clear fluid as content	Barley water fluid contains sperms
Brilliantly transilluminant	Transilluminant
Excision should be avoided in young	Can be excised

VARICOCELE

- It is dilatation and tortuosity of the pampiniform plexus of veins and so also the testicular veins. Normally, there will be numerous plexus of veins (pampiniform) in the scrotum, which all join together to form about 4-8 veins in the inguinal canal. Above, in the abdominal cavity, in the posterior abdominal wall all join to form a single testicular vein. On left side, it drains the left renal vein; on the right side it drains the inferior vena cava.
- Varicocele is common in *tall, thin young men*.
- More common on the *left side*, but often can be bilateral.

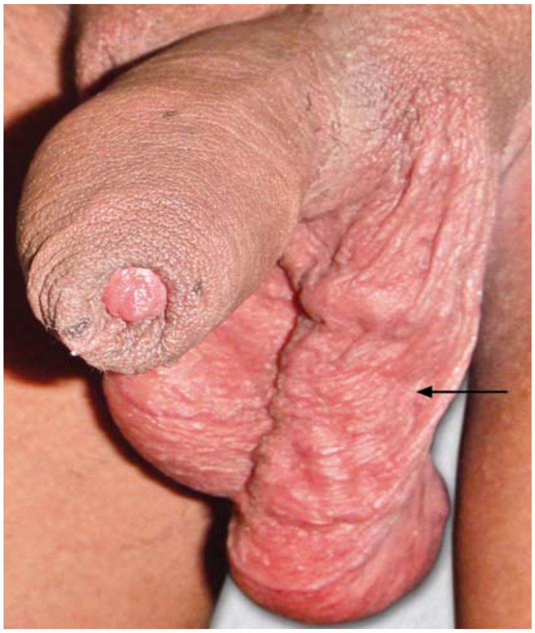

Fig. 26.199: Clinical photo of left sided varicocele—'bag of worms' look.

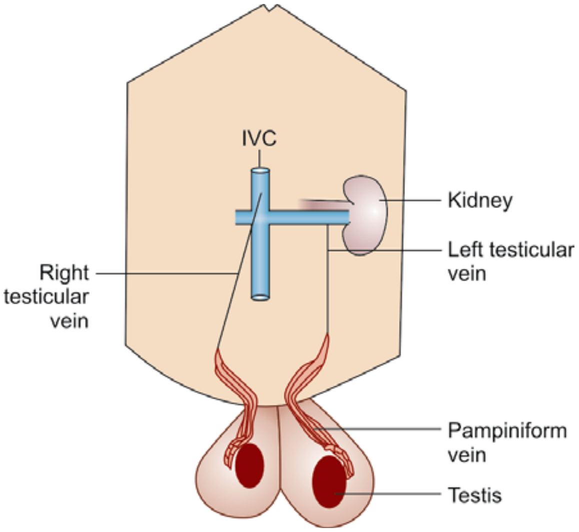

Fig. 26.200: Anatomy of testicular veins. Left testicular vein joins into left renal vein in perpendicular fashion. Right vein joins IVC directly.

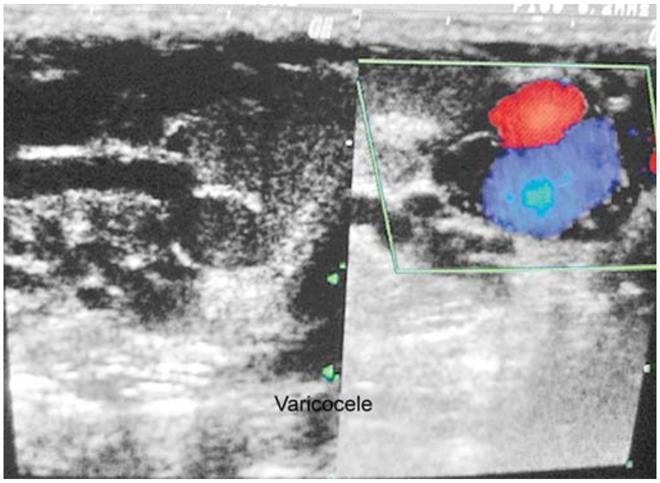

Fig. 26.201: Scrotal U/S showing varicocele.

- Commonly, it is idiopathic in origin, may be due to absence or incompetent valve at the junction of left testicular vein and left renal vein causing inefficient drainage of blood. Other reason is, due to perpendicular (right angle) entry of the left testicular vein into the left renal vein.
- In left sided renal cell carcinoma, tumour proliferates into the left renal vein and blocks the entry of left testicular vein causing varicocele on left side which is *irreducible.*
- Varicocele causes increased temperature in the scrotum which depresses the spermatogenesis and so causes *infertility* (correctable infertility).

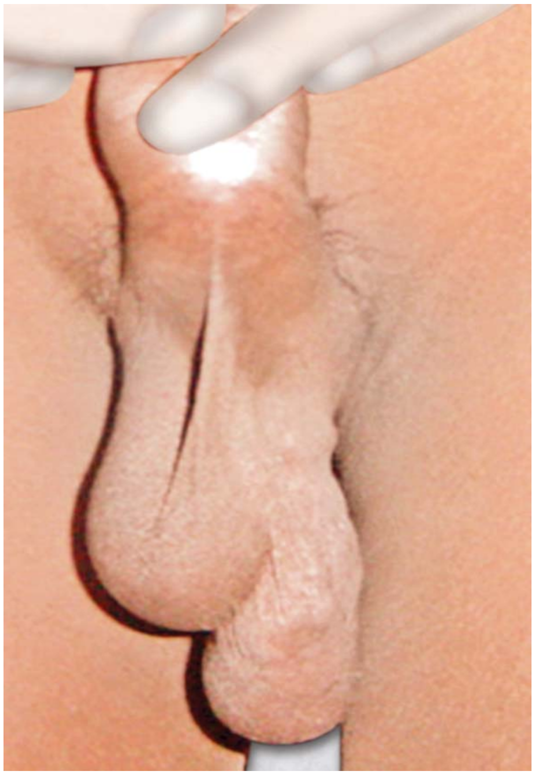

Fig. 26.202: Varicocele it is common on left side.

> *When the scrotal skin of the affected side is less rugose than the normal side, it is probable that, that half of the scrotum never contained a testis.W*
> **—Kenneth I Macrosson**

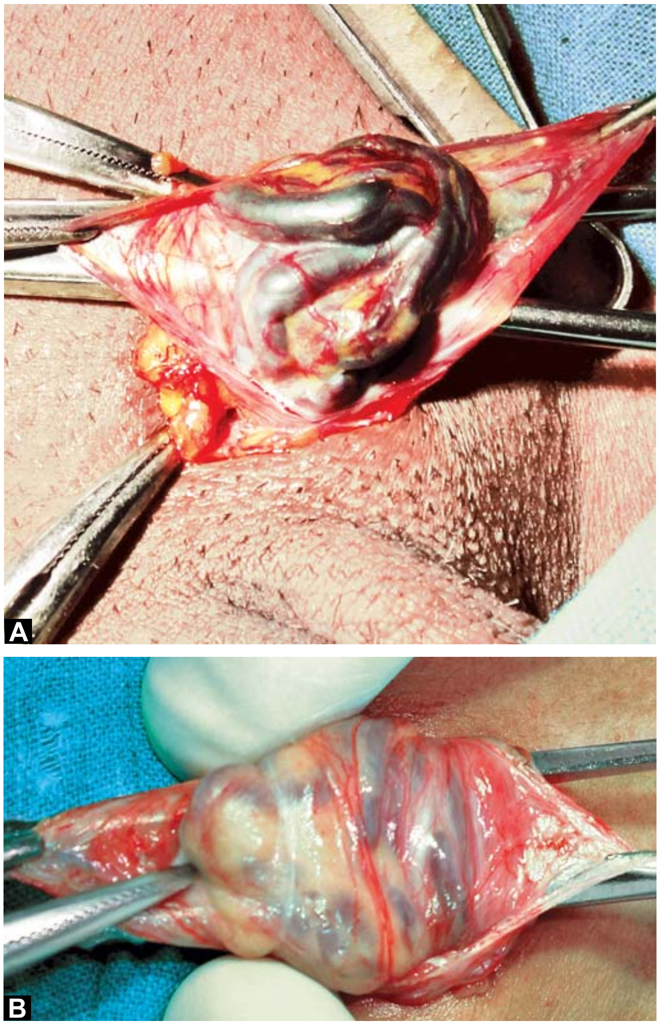

Figs 26.203A and B: Varicocele—on table look of tortuous pampiniform plexus of veins.

Types of Varicocele

- *Primary/idiopathic*—95%: No cause is found. There is incompetence of valves of the testicular vein. It is common on left side because left testicular vein joins left renal vein perpendicularly and left side vein is longer and liable to get compressed by loaded sigmoid colon. Left renal vein is often compressed between aorta and SMA.
- *Secondary*—due to specific cause like left-sided renal cell carcinoma with a tumour thrombus in left renal vein causing obstruction to venous flow of left testicular vein.
- *Subfertility/infertility* are observed even in unilateral varicocele. It is a debate whether it really causes subfertility. Possible causes are:
- Altered heat exchange mechanism of the scrotum due to varicocele → hyperthermia → inhibition of spermatogenesis.
- Increased blood flow → increased temperature in the testes → increases the metabolic activity using glycogen storage → depletion of glycogen → injury of parenchyma of testes → oligospermia.
- Hypoxia of testes.
- Same cause leads into Leydig cell dysfunction, decreasing testosterone levels.
- Final effect is maturation arrest → poor spermatogenesis.

Clinical features

- Swelling in the root of the scrotum
- Dragging pain in the groin and scrotum
- *"Bag of worms"* feeling
- Impulse on coughing
- On lying down it gets reduced (except in renal cell carcinoma)

Bow sign: After holding the varicocele between thumb and fingers, patient is asked to bow. Varicocele gets reduced in size. Bowing reduces the blood flow of testicular vein and pampiniform plexus causing reduction in size.

Grading of varicocele

I. Small
II. Moderate
III. Large
IV. Severely tortuous

Differential Diagnosis

- Hydrocele.
- Inguinal hernia.
- Lymph varix.
- Lipoma of the cord.

Investigations

- Venous Doppler of the scrotum and groin.
- U/S abdomen to look for kidney tumour.
- Semen analysis.

Treatment

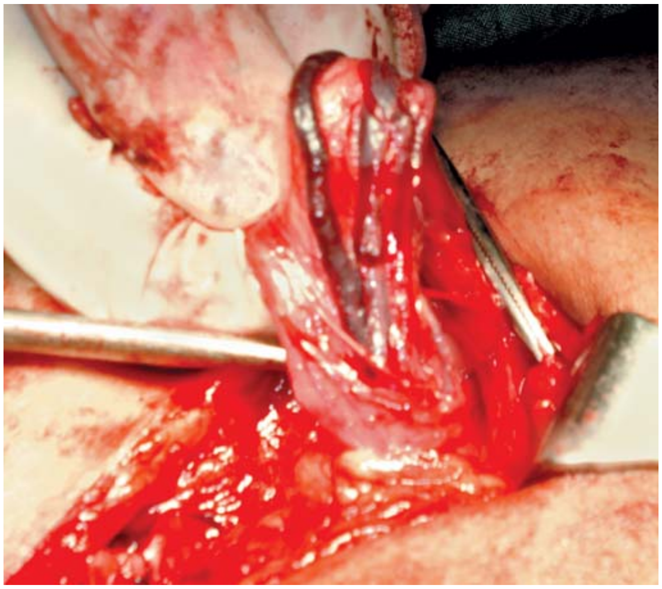

Fig. 26.204: On table varicocele finding.

- ***Palomo's operation***: Suprainguinal extraperitoneal ligation of the testicular vein.
- ***Inguinal approach (Inavissevich approach)***: Easier and safer.
- ***Subinguinal approach*** *(Marc-Goldstein)*: It is subinguinal approach at superficial inguinal ring outside the external oblique aponeurosis without opening the external oblique aponeurosis. Here cord is easily identified through a small incision.
- ***Scrotal approach***: In case of grade IV, veins have to be excised through this approach.
- ***Laparoscopic approach***:Presently accepted, good approach.

Indications for surgery

- Pain
- Oligospermia—usually in 6-12 weeks oligospermia improves very well and also the conception rate

Complications of varicocele surgery

- Haemorrhage and scrotal haematoma
- Infection, pyocele
- Injury to testicular artery
- Injury to ilioinguinal nerve and pain
- Recurrence—5-10%

Chapter 27 Neurosurgery

CHAPTER OUTLINE

- Head Injuries
- Extradural Haematoma
- Subdural Haematoma
- Subarachnoid Haemorrhage
- Depressed Skull Fracture
- Hydrocephalus
- Intracranial Abscess
- Intracranial Aneurysms
- Intracranial Tumours
- Pituitary Tumours
- Craniopharyngiomas
- Spinal Dysraphism
- Meningocele
- Spina Bifida
- Intervertebral Disc Prolapse
- Tuberculosis of Spine
- Spinal Tumours

HEAD INJURIES

> *Now as soon as thou findest that smash which is in his skull like those corrugations which form on molten copper, (and) something therein throbbing and fluttering under thy fingers like the weak place of an infant's crown before it knits together-when it has happened there is no throbbing and fluttering under thy fingers, until the brain of his (the patient's) skull is rent open- (and) he discharges blood from both his nostrils and both his ears, (and) he suffers stiffness in his neck.*
>
> — **(Anonymous),** circa 2500 BC

Mechanism

1. *Distortion of the brain*: Brain is a soft structure, therefore has a 'mobility' and readily distorts. This distortion and mobility is accentuated by CSF and vascular components. Any impact creates shearing forces in the brain causing damage to neurons, supporting tissues and blood vessels. This leads to loss of consciousness, focal neurological deficits. Such distortive damage may be temporary or permanent.
2. *Mobility of the brain in relation to the skull and membranes* causes cerebral damage and bleeding in dural spaces from torn vessels in the dura, commonly the veins. In old age, the brain shrinks, as a result of which 'mobility' of brain increases favouring rupture of veins which cross the subdural space.
3. *Configuration of interior of skull:* Damage is less severe over the smooth area but is more severe over the rough and sharp areas. So the damage is severe over the anterior cranial fossa, over the falx, and over the tentorium.

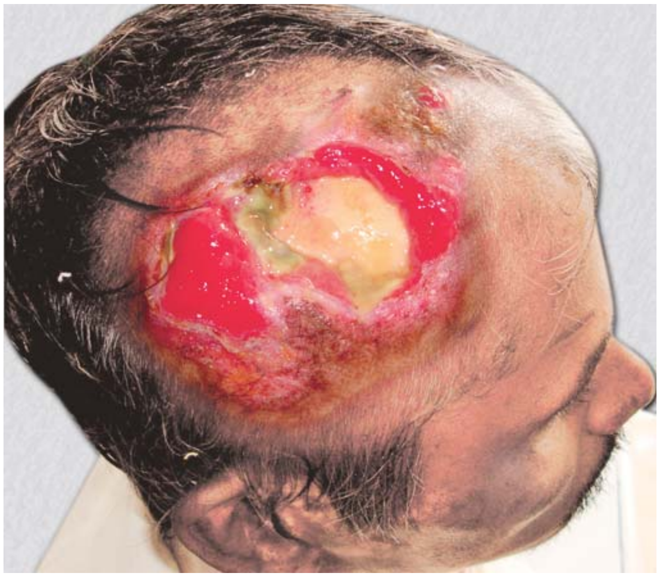

Fig. 27.1: Exposed skull after trauma with sepsis. It needs multiple burr holes to allow granulation tissue formation from depth. Later skin grafting can be done. Otherwise local rotation flap is needed.

4. *Deceleration and acceleration injuries: Deceleration injuries* occur when moving head strikes an immovable object (like in road traffic accidents). *Acceleration injuries* occur when stationary skull is struck by a moving object (like in assault).
5. *Cerebral concussion* is slight distortion causing temporary physiological changes leading to transient loss of consciousness with complete recovery.
6. *Cerebral contusion* is more severe degree of damage with bruising and cerebral oedema leading to diffuse or localized changes.

7. *Cerebral laceration* is tearing of brain surface with collection of blood in different spaces and with displacement of dural parts.

Effects of Brain Injuries

1. *Brain oedema* is accumulation of fluid, both intracellular and extracellular. It is due to congestion and dilatation of blood vessels. It may be diffuse or localized.
2. *Brain necrosis* is of severe variety with destruction and is due to haemorrhagic infarction.
3. *Extradural haematoma* occurs usually in temporoparietal region. It is commonly due to tear in middle meningeal veins and often middle meningeal artery. It causes intracranial hypertension, displacement, Kernohan's effect and often death.
4. *Subdural haematoma* is due to tear of veins between cerebrum and dura due to shearing forces. It is diffuse and commonly associated with cerebral injury.
5. *Intracerebral haematoma* can occur in different parts of cerebrum may be in frontal lobe, temporal lobe.
6. *Intraventricular haemorrhage* is very severe type of haemorrhage.
7. *Brain ischaemia* is due to increased pressure. This in turn leads to alteration in the perfusion of brain which itself aggravates the ischaemia and this forms a vicious cycle, causing progressive diffuse ischaemia of brain.
8. *Coup injury* occurs on the side of the blow to the head. *Contre-coup injury* occurs on the side opposite to the blow on the head.
9. *Coning*: It is due to raise in intracranial pressure causing either:
   i. Herniation of contents of supratentorial compartment through *the tentorial hiatus* or
   ii. Herniation of the contents of infratentorial compartment through the *foramen magnum.*

In supratentorial herniation, there is compression of ipsilateral third cranial nerve and midbrain. Midbrain is displaced away from the mass (haematoma) and midbrain is pressed by the sharp edge of tentorium cerebelli of opposite side leading to dysfunction of corticospinal fibres (which after decussation supplies the opposite side of the body, i.e. same side of the injury).

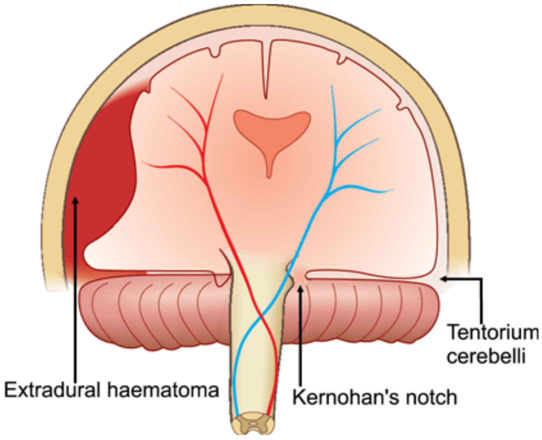

Fig. 27.2: Coning with Kernohan's notch.

This leads to:
   a. Deterioration in the level of consciousness.
   b. Dilatation of pupil on the side of compressing mass (haematoma).
   c. Hemiparesis on the same side of the mass lesion (haematoma) due to compression of the contralateral corticospinal tract. This effect is called as ***"Kernohan's notch"***.

   Herniation of infratentorial contents through the foramen magnum causes obstruction of cerebral aqueduct which further damages the brain function.

10. *Respiratory failure* altering PO_2 and PCO_2 levels.
11. *Raised intracranial pressure* causing bradycardia, hypertension, vomiting. Raised intracranial pressure may precipitate coning and thus aggravates brain ischaemia.
12. *Fluid and electrolyte imbalance.*
13. *Hyperpyrexia.*
14. *Convulsions* due to irritation of grey matter.
15. *CSF rhinorrhoea or CSF otorrhoea.*

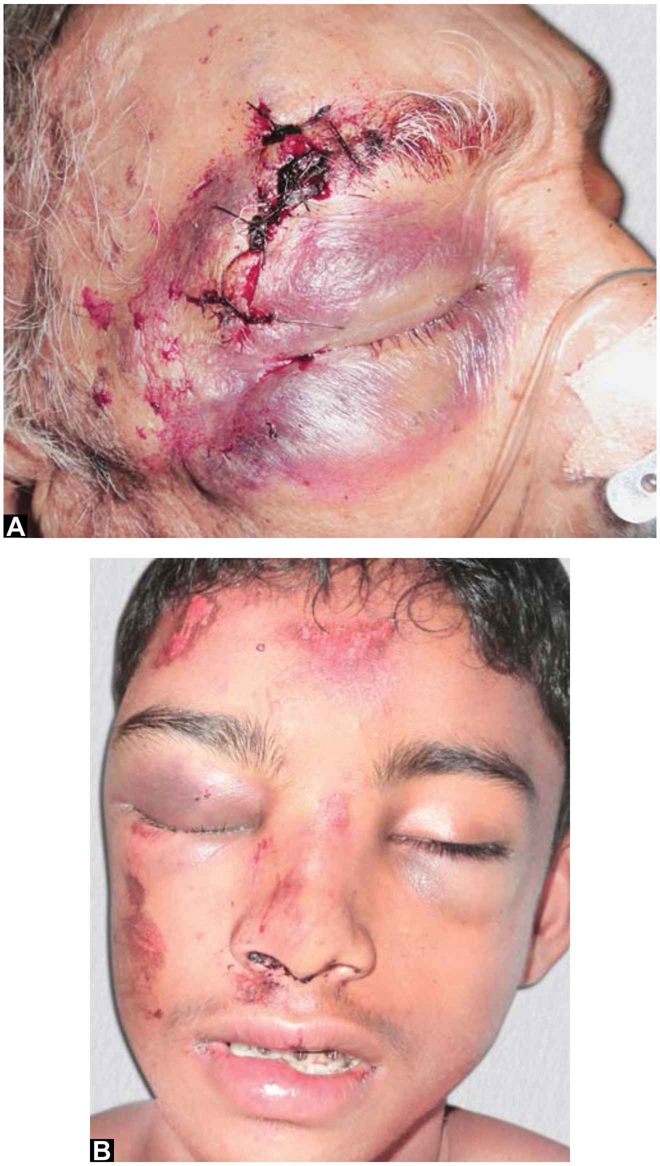

Figs 27.3A and B: Balck eye in two different patients.

Pathology of Head Injury

Primary lesions	*Secondary lesions*
• Diffuse neuronal damage • Shearing lesions • Contusions • Lacerations	Brain swelling • Oedema, venous congestion, hypoxia Intracranial haemorrhage • Extradural • Subdural • Intracerebral Infections • Open head injury – Generalised meningitis – Subdural empyema • Closed head injury – Pott's puffy tumour
Cerebral concussion	***Cause of death in head injury***
• Temporary physiological paralysis of nervous system • Loss of consciousness • Post-traumatic amnesia • Full recovery is expected • Residual complications may develop	• Brain hypoxia • Coning • Diffuse severe irreversible neuronal injury • Death may be due to other injuries like abdo-minal/ thoracic • Metabolic changes • Aspiration in unconscious patient

Other features

- CSF leak or bleeding from nose
- Blood collection in the orbit
- Black eye
- *Battle's sign*—ecchymosis over the mastoid
- Haematoma of scalp
- *Panda sign*—bilateral black eye

Clinical Approach of a Patient with Head Injury

1. Detail history of injury has to be taken and also the process of deterioration—rapid or gradual.
2. History of alcohol intake: Alcohol intake mimics head injury and alcoholism itself may mask the features of head injury.
3. *Neurological assessment*: By

- Level of consciousness
- Glasgow coma scale
- Pupillary reaction to light and size
- Pulse
- Temperature
- Blood pressure
- Respiratory rate
- Reflexes
- Limb movements—normal/mild weakness/ severe weakness/ spastic flexion/extension/no response

4. Status and protection of airway.
5. General assessment and other injuries like fractures, abdominal organ injuries, thoracic injuries are looked for.
6. Presence of any scalp haematoma, fractures of skull bone which may be depressed has to be looked for.
7. Any blood from nose or ear, CSF rhinorrhoea or CSF otorrhoea has to be looked for.

Indications for Hospitalisation

- Any altered level of consciousness
- Skull fracture
- Focal neurological features
- Persistent headache, vomiting, systolic hypertension, bradycardia
- Alcohol intoxication
- Bleeding from ear or nose
- Associated injuries

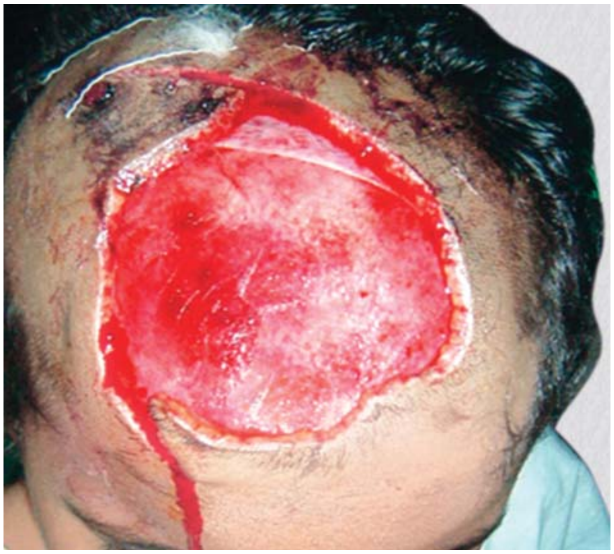

Fig. 27.4: Degloving injury scalp. Note all layers of the scalp stripped off.

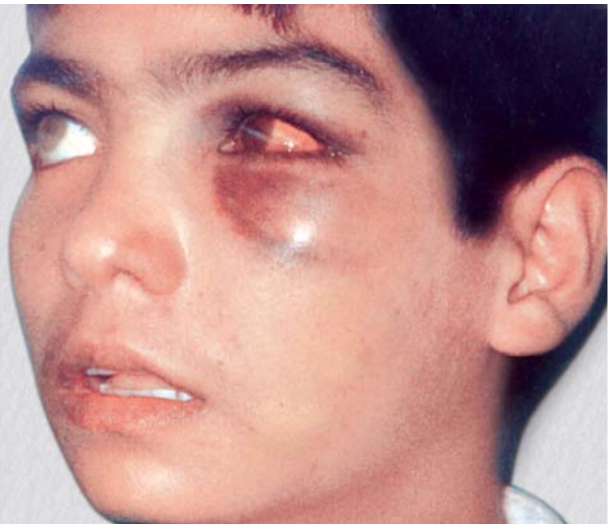

Fig. 27.5: Patient with head injury having black eye.

Glasgow coma scale

Eye opening		*Verbal response*		*Motor response*	
Spontaneous	–4	Oriented	–5	Obeys commands	–6
To speech	–3	Confused	–4	Localises pain	–5
To pain	–2	Inappropriate words	–3	Flexion to pain	–4
None	–1	Incomprehensible words	–2	Abnormal flexion	–3
		None	–1	Extension to pain	–2
				None	–1

Total score-15
Mild head injury: score 13-15
Moderate head injury: 9-12
Severe head injury: less than 8 (3-8)

Adelaide coma scale

It is used in children
Scores for eye opening and motor responses are same as Glasgow coma scale

But *verbal response score differs*—Oriented-5. Words-4. Vocal sounds-3. Cries-2. Nil-1

Orientation cannot be evaluated below 5 years. For first 6 months, the best verbal response is CRY

Investigations

- X-ray skull: To look for fracture, relative position of the calcified pineal gland, presence of intracranial air.
- Serum electrolyte measurement.
- Blood grouping and cross matching.
- CT scan: Plain (not contrast) to look for cerebral oedema, haematomas, midline shift, fractures, ventricles, brainstem injury.
- Carotid arteriography.

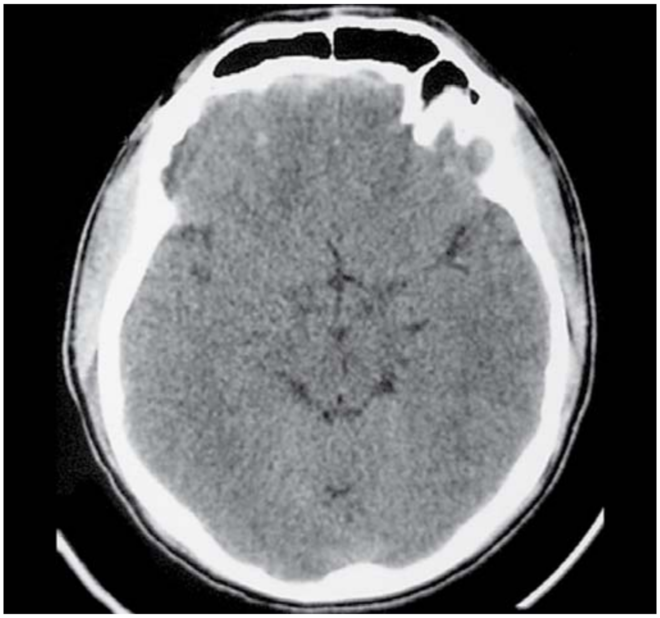

Fig. 27.6: CT scan head showing scalp haematoma on both sides. There is no internal injury.

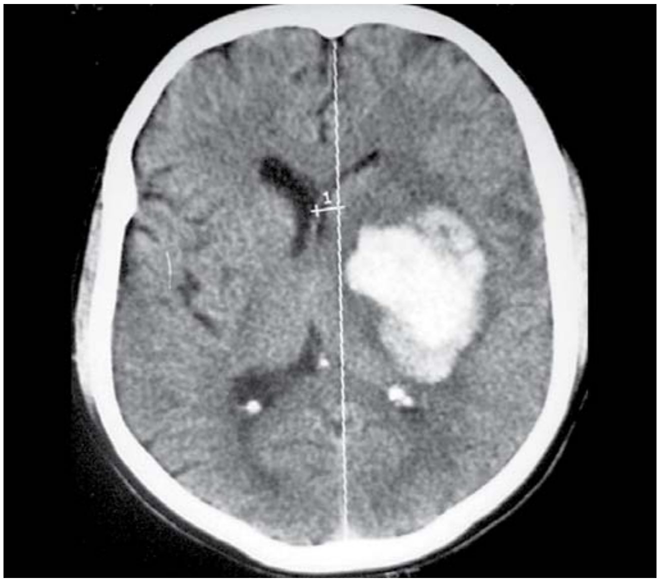

Fig. 27.7: CT scan head showing intracerebral haematoma.

- Investigations for other injuries like ultrasound of abdomen.
- Monitoring of intracranial pressure.

Treatment

General

- Protection of airway using mouth gag, endotracheal intubation or tracheostomy, whenever required.
- Throat suction, bladder and bowel care and good nursing are very essential.
- Nasal oxygen, or often ventilator support.
- IV fluids initially, later Ryle's tube feeding has to be done.
- Electrolyte maintenance.

Drugs

- Sedation is avoided.
- Analgesics and anticonvulsants like phenytoin or phenobarbitone is started.
- Diuretics are given to reduce cerebral oedema—either mannitol 20%, 200 ml IV 8th hourly or frusemide 40 mg IV 8th hourly. It should not be given in case of intracranial haematoma.

In great attempts, it is glorious even to fail.

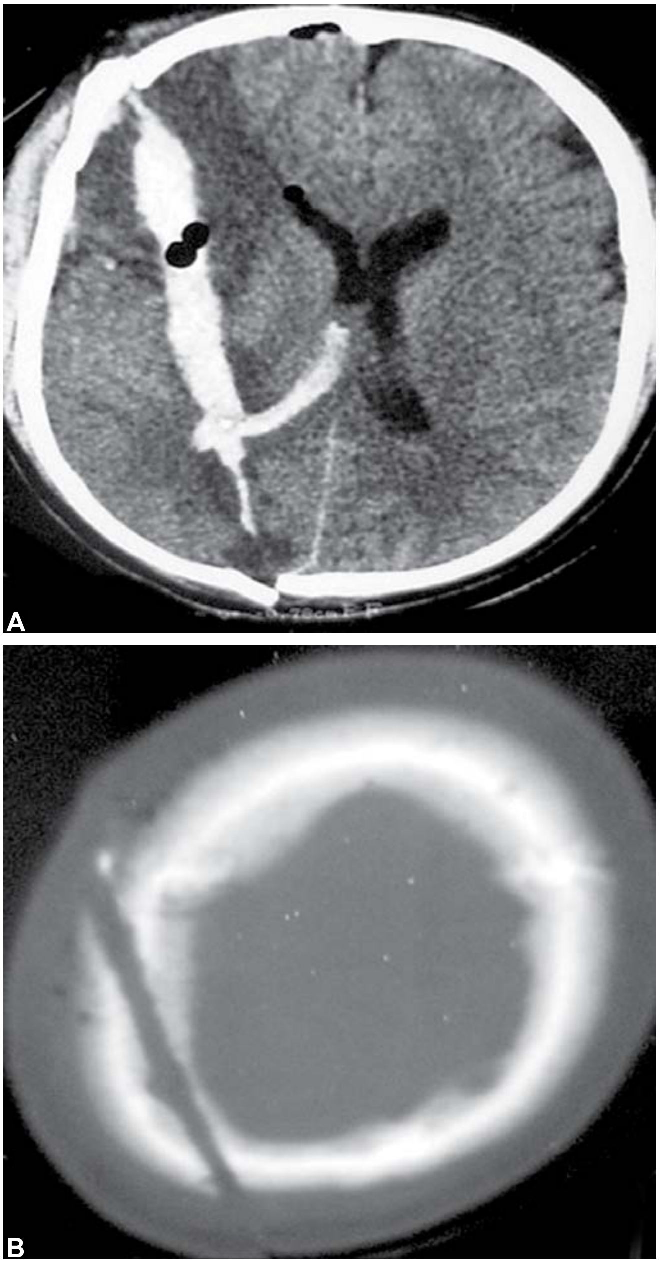

Figs 27.8A and B: CT scan head showing sharp head injury causing through and through cut across skull and brain. Patient survived and went home.

- Antibiotics like penicillins, ampicillins are given to prevent the onset of meningitis.
- Corticosteroids, either dexamethasone or betamethasone is used commonly, but its beneficial effect is not confirmed.
  Indications for surgery
  - Acute extradural haematoma.
  - Acute subdural haematoma.
  - Depressed skull fracture.

Procedure

Craniotomy is done and cranial flap is raised. Clot is evacuated followed by applying *hitch stitches* between dural layer and scalp. Postoperative antibiotics, analgesics, anticonvulsants are given.

Complications of Head Injuries

Early

- Brainstem injury—due to coning.
- Compression over cerebellum and medulla.
- ***CSF rhinorrhoea***: Due to communication between intracranial cavity and the nose. There is a tear in the dura following the fracture involving the sinuses—frontal, ethmoid, sphenoid sinuses. Meningitis is the common complication of CSF rhinorrhoea.
  Treatment: Initial management is only conservative, for 10 days—by antibiotics and observations.

Indications for surgical intervention

- Fracture of middle third face
- CSF rhinorrhoea persisting for more than 10 days
- Fracture of sinuses
- An aerocoele
- An attack of meningitis

  Surgeries:
  i. Reduction of fracture of middle third face.
  ii. Exploration of anterior cranial fossa.
- Meningitis—common.
- Pituitary damage and endocrine failure—requires high dose of hydrocortisone 200 mg, 6th hourly.
- Aerocoele.
- CSF otorrhoea.
- Depressed fractures, often causes injury to dural venous sinuses and may lead to torrential haemorrhage, which may be life threatening. So such depressed fractures should never be elevated.

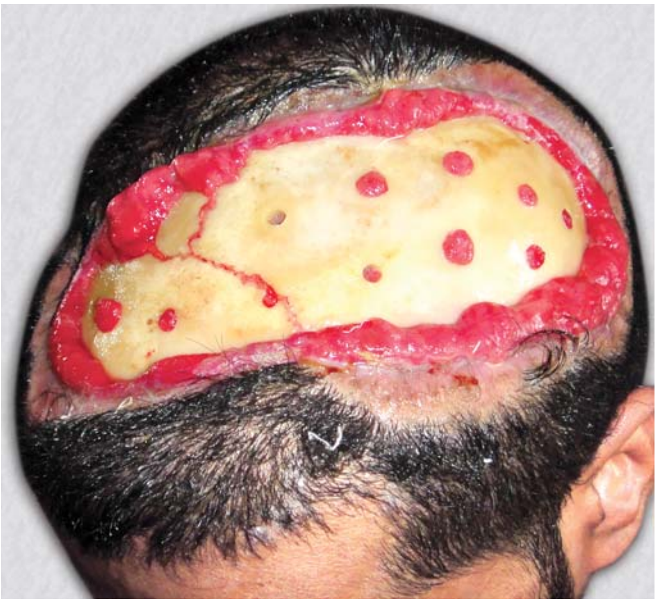

Fig. 27.9: Multiple burr holes done to promote formation of granulation tissue from vascular diploe of the skull. Once granulation tissue covers the entire defect, skin grafting can be placed.

Late

- Chronic subdural haematoma.
- Early post-traumatic epilepsy—they need anticonvulsants for 3 years.

- Late post-traumatic epilepsy is due to scarring and gliosis of cerebrum.
- Post-traumatic amnesia.
- Post-traumatic hydrocephalus.
- Post-traumatic headache.

EXTRADURAL HAEMATOMA

- It is collection of blood in the extradural space between the dura and skull.
- Most common site is *temporoparietal* region. It can *be unilateral or bilateral*

Vessels commonly involved

- Middle meningeal veins
- Anterior branch of middle meningeal artery
- Posterior branch of middle meningeal artery

Usually it is associated with fracture of temporoparietal region.

Pathology

Direct blow, like from cricket ball or road traffic accidents or fall and impact or coup and contre-coup injuries
↓
Fracture of thin temporal bone
↓
Tear of vessels
↓
Bleeding initially outward towards the scalp and under temporalis muscle
↓
Formation of haematoma
↓
Gradual stripping of dura from skull and collection of blood occurs
↓
In 6-12 hours extradural haematoma occurs which raises the intracranial pressure
↓
Coning of supratentorial content (uncus of temporal lobe) through tentorial hiatus
↓
Shift of midbrain towards opposite side which gets injured by sharp edge of the tentorial cerebelli
↓
Corticospinal tract before decussation on opposite side gets injured
↓
So hemiparesis, and pupillary changes occur on the same side of haematoma
↓
This effect is called as ***Kernohan's notch effect***.

- Immediately after injury, there is transient loss of consciousness and the patient soon becomes normal. Later after 6-12 hours, he again falls ill and the condition deteriorates. This is the time taken to develop raised intracranial pressure, coning and its effects. This crucial time gap which is unnoticed and often missed is called as ***"lucid interval".***

Clinical Features

- History of transient loss of consciousness following a H/o blow or fall.
- Patient soon regains consciousness and again after 6-12 hr starts deteriorating ***(Lucid interval)***.
- Later the patient presents with confusion, irritability, drowsiness, hemiparesis on same side of the injury. Initially pupillary constriction and later pupillary dilatation occurs on the same side, finally becomes totally unconscious – *Hutchinsonian pupils*.
- Death can occur if immediate surgical intervention is not done.
- *Features of raised intracranial pressure* like high blood pressure, bradycardia, vomiting is also seen. Occasionally *convulsions* may be present.
- *Wound and haematoma in the temporal region of scalp* may be seen.

Investigations

- X-ray skull may show fracture of temporal bone.
- Electrolyte estimation.
- CT scan head is diagnostic. Extradural haematoma shows *biconvex* lesion.

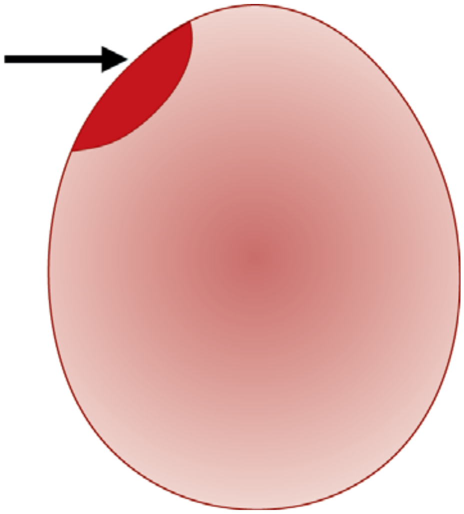

Fig. 27.10: Extradural haematoma. Note the biconvex configuration of the haematoma.

Treatment

Immediate surgical intervention is a must to save the life of the patient.

Fourth nerve when paralysed causes diplopia when going downstairs (that is on looking downwards).
—Sir Benjamin W Ryeroft

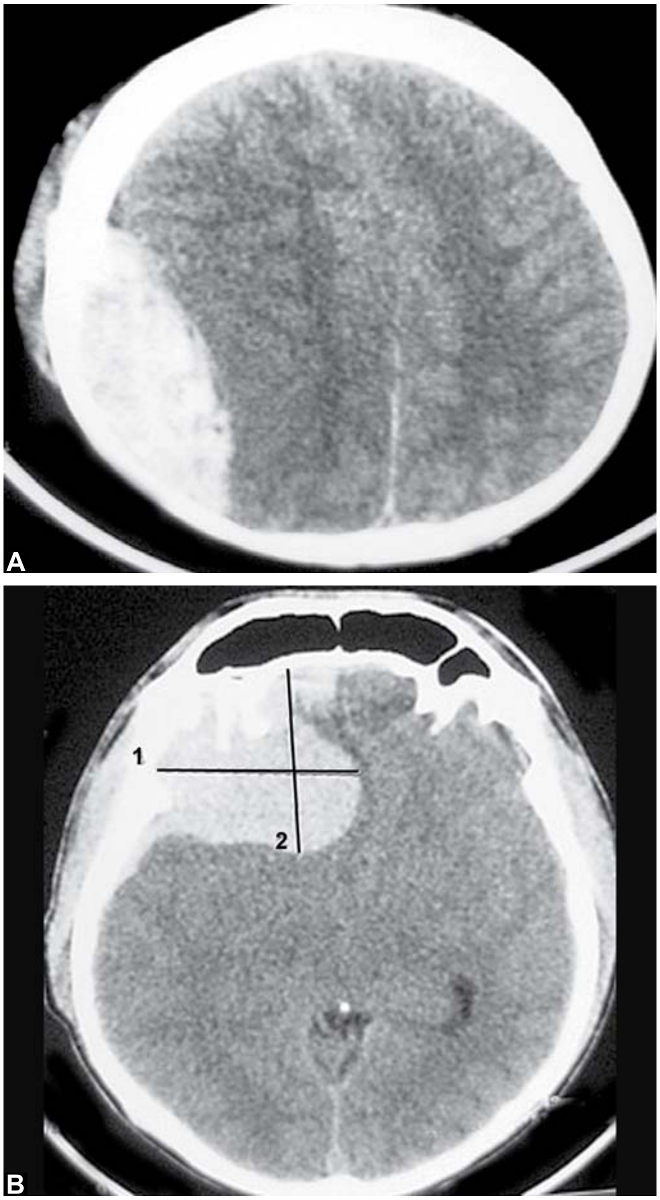

Figs 27.11A and B: CT scan head showing extradural haematoma with biconvex lesion.

- Craniotomy is done and cranial flaps are raised. The dura is opened and the clot is evacuated. The dura is fixed to galea using interrupted sutures—*Hitch stitches*.
- Antibiotics and anticonvulsants are given postoperatively.
- Recovery is good after surgery.

Treatment of extradural haematoma

- Earliest surgery and evacuation is the need
- 5 cm vertical incision in parietal region above the zygoma
- Galea is incised. Skull is opened using perforator and burr
- Meninges are kept aside
- *Black currant jelly clot* is evacuated
- Bleeding vessels are cauterized—bipolar cautery
- Dural hitch stitches are placed
- Opposite side if needed should be evacuated
- Antibiotics; anticonvulsants
- Analgesics
- General measures—catheter; fluid therapy
- Prevention post-traumatic complications

Complications

- Post-traumatic epilepsy
- Meningitis
- Post-traumatic amnesia

SUBDURAL HAEMATOMA

Types

- *Acute*
- *Chronic*

Acute Subdural Haematoma

- It is a collection of blood between the brain and dura. It is due to injury to the cortical veins and often due to laceration of cortex of brain which bleeds and blood gets collected in the subdural space forming a haematoma.
- Here haematoma is extensive and diffuse. There is no lucid interval. There is severe primary brain damage.
- Haematoma may be of coup and contre-coup type.
- Loss of consciousness occurs immediately after trauma and is progressive.
- Convulsion is common.
- Features of raised intracranial pressure is obviously seen—high BP, bradycardia, vomiting.
- Focal neurological deficits or hemiparesis can occur.
- CT scan shows concavo-convex lesion.

Treatment

- Antibiotics, anticonvulsants.
- *Surgical decompression is done by craniotomy*.

Chronic Subdural Haematoma

- It is due to the rupture of veins between dura and brain (cerebral hemispheres), causing gradual collection of blood in subdural space.
- It is commonly seen in elderly people following any minor trauma like fall, slipping (which might have gone unnoticed).
- In elderly people, brain atrophies and even minor injuries can cause shearing and bleeding from these veins.
- Blood collects gradually over 2-6 weeks. Plasma and cellular components get separated. Eventually cellular part gets absorbed leaving only fluid component. It is called as ***chronic subdural hygroma.***
- Usual haematoma collection is 60-120 ml. Often in 50% of cases, it is bilateral.

Clinical Features

- Common in old age, with history of minor trauma.
- Patient presents with confusion, disorientation, gradually with altered level of consciousness and drowsiness.
- Later convulsions, features of intracranial hypertension, features of coning develops.
- Extensor plantar response and pupillary changes develop eventually.

Investigations

- CT scan (shows *concavo-convex* lesion).
- Serum electrolytes.
- Blood grouping and cross matching.

Differential Diagnosis

- Electrolyte imbalance.
- Intracranial space occupying lesion.

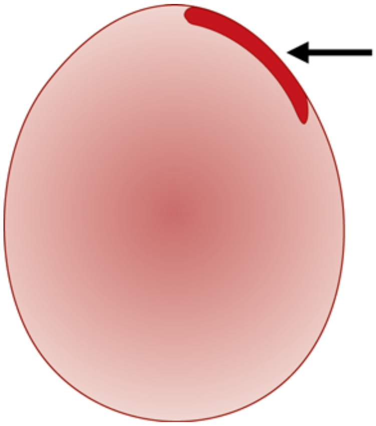

Fig. 27.12: Subdural haematoma. Note the concavo-convex configuration of the lesion.

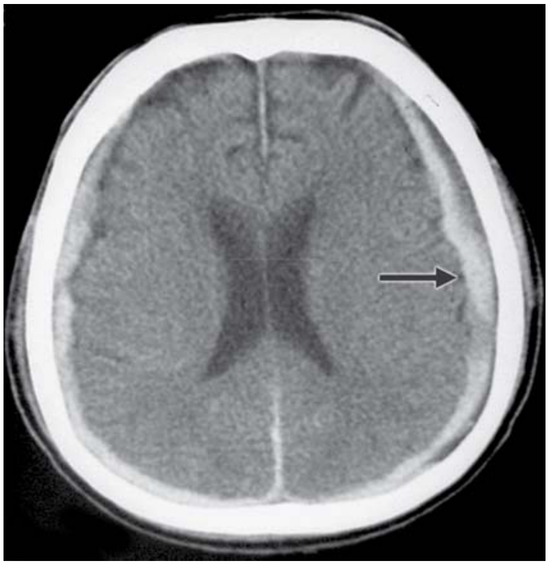

Fig. 27.13: CT scan head showing subdural haematoma with concavo-convex lesion.

Treatment

- Craniotomy and evacuation of clot is done when required on both sides.
- Antibiotics.
- Anticonvulsants for 3 years.

Complications

- Epilepsy
- Meningitis
- Coning
- Neurological deficits

Chronic Subdural Empyema

- It may be primary infection of subdural space from sinusitis focus causing suppuration and pus formation.
- It can be complication of the chronic subdural haematoma.
- It is due to secondary bacterial infection of collected clot/fluid.
- Infection is from sinusitis scalp (common)/through earlier trauma wound/Haematogenous.
- Commonly Gram positive organisms cause empyema like streptococci (viridans/milleri) but other virulent organisms like Gram negative bacteria occasionally can cause.
- There is cortical venous thrombophlebitis and cortical infarction.
- Headache, fever, meningism and convulsions are the features.
- MRI is ideal than CT to diagnose.
- *Treatment*: Antibiotics, craniotomy and drainage; anticonvulsants, ICU care, proper monitoring, regular follow up.
- Condition has got 10% mortality.

SUBARACHNOID HAEMORRHAGE (SAH)

> *The Abbot by chance wished to decide with him the fate of the servants, found (him) prostrate upon the ground, insensible to shouts, to shaking and pinching of the body…senseless….I saw him livid from pallor, deprived of all sensation and animate motion….The first hour after midday of the same day he ceased to live…. By the indulgence of the Most Reverend Lord Abbot… I opened the head….Much blood flowed from the space….The (blood) covered (the brain) all over….The ventricles laid open I found them all filled up with blood.*
>
> —**Johann Jakob Wepfer,** 1658

It is a type of intracranial haemorrhage into the subarachnoid space usually from basal cisterns.

It may be spontaneous or following trauma.

Causes

- Intracranial aneurysms—commonest cause (50%)
- Hypertension
- A-V malformations
- Blood dyscrasias
- Anticoagulant drugs
- Brain tumours (malignant)

Two black eyes following one injury indicate fracture of the base of skull.—**Sir Earnest Finch**

Clinical Features

- Sudden onset of severe headache with vomiting.
- Features of raised intracranial pressure.
- Photophobia.
- Neck stiffness.
- Focal neurological deficits: hemiplegia, dysphasia.
- Eye changes: ptosis, dilated pupil, changes in the eyeball movements.
- Sudden loss of consciousness.
- Features of brain oedema and cerebral ischaemia.

In 40% of recovered patients, rebleeding occurs in 6-8 weeks which is commonly fatal.

Hunt and Hess grading for subarachnoid haemorrhage

Grade 1: Asymptomatic
Grade 2: Severe headache and neck stiffness
Grade 3: Drowsy, confused or mild focal deficit
Grade 4: Stupor, hemiparesis
Grade 5: Decerebrate rigidity, coma

Fischer grading of SAH

I—minimal < 8 mm size
II—moderate 8-15 mm size
III—severe > 15 mm size

Differential Diagnosis

- Meningitis.
- Coning due to any cause.

Investigations

- Lumbar puncture should be done to differentiate from meningitis.
  It has to be done carefully as it may precipitate coning.
  In subarachnoid haemorrhage, blood stained CSF is collected.

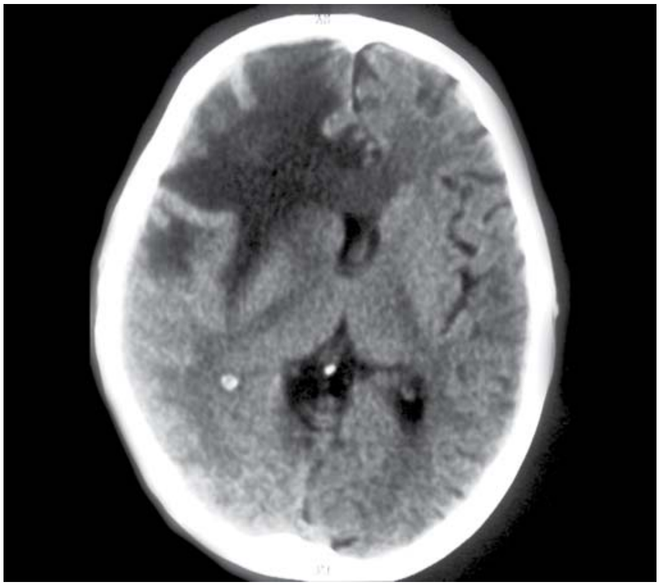

Fig. 27.14: CT scan showing intracerebral brain haemorrhage.

- CT scan.
- Carotid and vertebral angiogram.

Treatment

- Clipping, or wrapping the aneurysm.
- Craniotomy and proceed.
- Ligation of common carotid artery—there is risk of hemiplegia.
- Therapeutic embolisation.
- Excision of vascular malformations.
- Coiling of aneurysm with instruments.

DEPRESSED SKULL FRACTURE

- It is a common neurosurgical problem among the head injuries.
- It means fracture depression is more than the depth of inner table of the skull.

Problems in depressed fracture

- Tear in the dura beneath
- Haematoma in the deeper plane
- Injury to the cerebrum
- Injury to the venous sinuses—may cause life-threatening haemorrhage. *Fracture should not be elevated* in such occasion, as it itself can precipitate bleeding
- Convulsions
- Meningitis

Investigations

- CT scan.

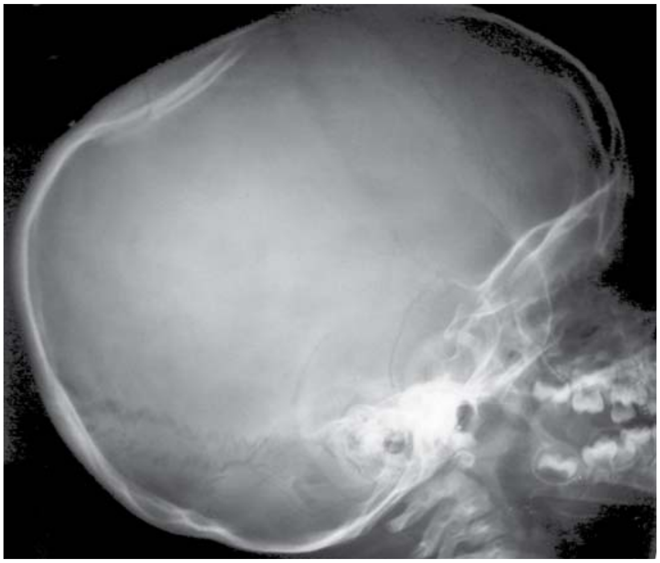

Fig. 27.15: Depressed skull fracture.

Treatment

- Antibiotics, anticonvulsants.
- *Elevation of the depressed fracture*: Burr holes are made in the adjacent normal skull. Fracture is elevated. Bony

fragments and necrotic materials are removed. Dural tear is closed with interrupted sutures.

HYDROCEPHALUS

> *The heads of children sometimes grow enormously large, the sutures give way, and the membranes of the brain are pushed up with the water within, and make a soft tumour rising above the edges of the sutures…. They daily become more and more stupid, with a pulse not above seventy-two.*
>
> —**William Heberden,** 1802

It is dilatation of ventricles due to blockage of cerebrospinal fluid flow (CSF) or due to increased secretion or due to defective absorption of CSF.

Classification I

- **Communicating type:** Ventricles communicate freely into the subarachnoid space.
  Here there is defective absorption of CSF following any inflammation, subarachnoid haemorrhage or trauma.
- **Noncommunicating type**: Obstruction is in the ventricle or its exit, due to any tumours or any inflammatory process.

Clinical Features

Bulging of anterior fontanelle, engorged scalp veins, separation of suture lines, *sun-setting sign* (decreased upward gaze), increasing head circumference, papilloedema, lethargy, ataxia.

Investigations

CT scan, ventriculography, air encephalography, MRI.

Treatment

- Tapping of lateral ventricles.
- Ventriculocysternostomy using polythene catheters—*Torkildsen operation.*
- Ventriculoatrial (VA) shunt.
- Ventriculoperitoneal (VP) shunt.

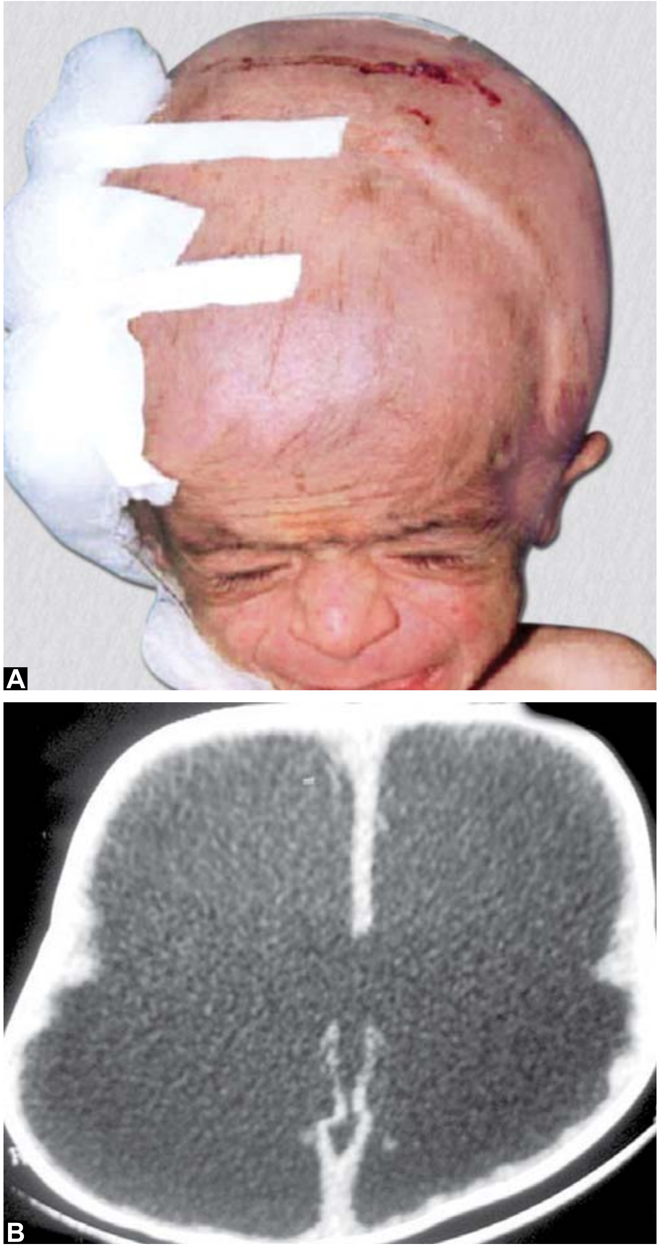

Figs 27.16A and B: Hydrocephalus child. CT scan head picture in hydrocephalus.

Classification II

Congenital	*Acquired*
Associated with spina-bifida and myelomeningocele. It is due to:	May be unilateral or bilateral (midline obstruction), due to:
• Failure of formation of CSF pathways	• Chronic meningitis
• Arnold-Chiari malformation	• Trauma
• Congenital stenosis of aqueduct of Sylvius	• Subarachnoid haemorrhage
• Presents with widening of sutures, tense fontanelles and decreased cortical thickness • Enlargement of head occurs, either prenatal (can cause obstructed labour) or postnatal	• Brain tumours—(pineal/cerebellar/craniopharyngiomas) • Colloid cyst of 3rd ventricle • Arachnoid cysts

The blessing of health is realised on the sick bed.—**Mr Tut-Tut**

INTRACRANIAL ABSCESS

Types

- ***Extradural abscess***: Caused by:
  - Osteomyelitis of skull.
  - Middle ear infection.
  - Frontal sinusitis.

  Pott's puffy tumour is subperiosteal swelling with infection and inflammation of the scalp. There is acute localized headache and tenderness in the skull, localized pitting oedema of the scalp usually in the frontal region.
- ***Subdural abscess***: is caused by septic thrombophlebitis from the frontal sinusitis or other infections. It is often very severe with extension into the venous sinuses.
- ***Intracerebral abscess***: is caused by
  - Extension from middle ear or sinuses.
  - Blood-born infection.
  - After intracranial injuries.

Common sites: Temporal lobe, cerebellum, frontal lobe.

It can be:

a. *Acute*—There is acute septic encephalitis without pus formation. It may cause ventriculitis or localized abscess formation

b. *Subacute*—Occurs in 3 weeks, by the formation of a glial wall, i.e. thickness is more near the cortex and less towards ventricle

c. *Chronic*—Occurs in 6 weeks with thick wall which may persist or may get enlarged behaving like a space occupying lesion

d. *Metastatic*—Abscess in brain occurs either in cerebrum (parietal or temporal lobes) or in ventricles (Ventriculitis is more dangerous and often fatal)

Clinical Features

- Evidence of focus of infections are seen, i.e. middle ear (CSOM), sinusitis.
- Focal neurological features are seen, depending on the location of abscess. In temporal lobe abscess features of dysphasia, contralateral hemiparesis are seen; in cerebellar abscess, all cerebellar symptoms are seen.
- Epilepsy.
- Features of raised intracranial pressure: (a) slow pulse, (b) rising BP, (c) headache and vomiting, (d) papilloedema, (e) deterioration in level of consciousness, (f) visual disturbances.

Differential Diagnosis

- Intracranial tumour.
- Tuberculoma.
- Meningitis.

Investigations

- CT scan.
- MRI.
- Carotid angiogram.
- Ventriculography.
- EEG.
- Isotope brain scan.
- Total count, ESR.
- Investigation specific for focus of infection.

Lumbar puncture should be avoided in acute abscess as coning can occur.

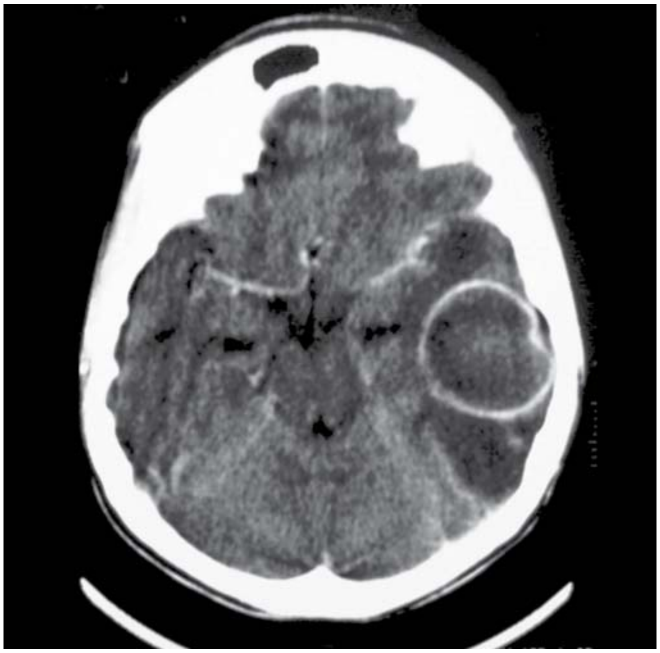

Fig. 27.17: Intracranial abscess. It needs drainage. It could be tuberculosis. It is a space occupying intracranial lesion, mimics malignancy often.

Treatment

- Antibiotics—high dose penicillins, benzyl penicillins are given.
- Burr hole exploration is done and Dandy's brain cannula is placed. Pus is aspirated and sent for culture and cytology.
- In case of chronic abscess, exploration of cranial cavity and excision of brain abscess is done.

Complications

- ❖ Intracranial hypertension
- ❖ Coning
- ❖ Neurological deficits

INTRACRANIAL ANEURYSMS

Types

- ***Subclinoid type*** occurs in the internal carotid artery within the cavernous sinus. It causes ptosis, defective external ocular movements, 5th nerve palsy. It can cause carotid-cavernous fistula.
- ***Supraclinoid type*** is commonest type.
  - *Berry aneurysms: A congenital type* occurs in circle of Willis in relation to internal carotid artery [40%

(most commonly at the origin of posterior communicating artery)], anterior communicating artery, middle cerebral artery, vertebrobasilar artery. It occurs due to weakness in the media of major arteries.
- *Acquired aneurysms* due to atheromas, hypertension.
- *Mycotic aneurysms* occurs due to infection in the wall of cerebral vessels, as a result of any bacteraemia.

Common sites are peripheral branches of middle cerebral artery.

Presentations

- Subarachnoid haemorrhage
- Pressure effects
- Convulsions
- Eye and pupillary signs

Investigations

- CT scan.
- Carotid and vertebral angiogram.

Treatment

- Clipping or wrapping of aneurysms.
- Therapeutic embolisation.
- Open neurosurgical approaches.

INTRACRANIAL TUMOURS

> *There was discovered on (the corpus callosum) of the brain a remarkable round fleshy tumour like an acorn. It was hard and full of holes and was as large as a medium-sized apple. It was covered with its own membrane and was entwined with veins.... We perceived that this ball by compressing the brain and its ducts with its mass and by flooding them, had been the occasion of the lethargy and listlessness and finally of death.*
>
> —**Felix Platter,** 1614

Secondaries are the commonest malignant tumour in the brain. *Metastasis occurs usually from lung (commonest), nasopharynx or from any other organ in the body.*

Secondaries in brain

- Commonest malignant brain tumour
- Primary are—lungs (40%), breast (30%), melanoma (15%), others (15%)
- Usually multiple
- Headache, focal lesions, convulsions, hypertension, altered level of consciousness are the features
- CT scan, CT chest or evaluation of primary—needed
- Radiotherapy and chemotherapy of particular primary is treatment
- Solitary metastasis can be removed surgically
- Steroids reduce the cerebral oedema and so symptoms

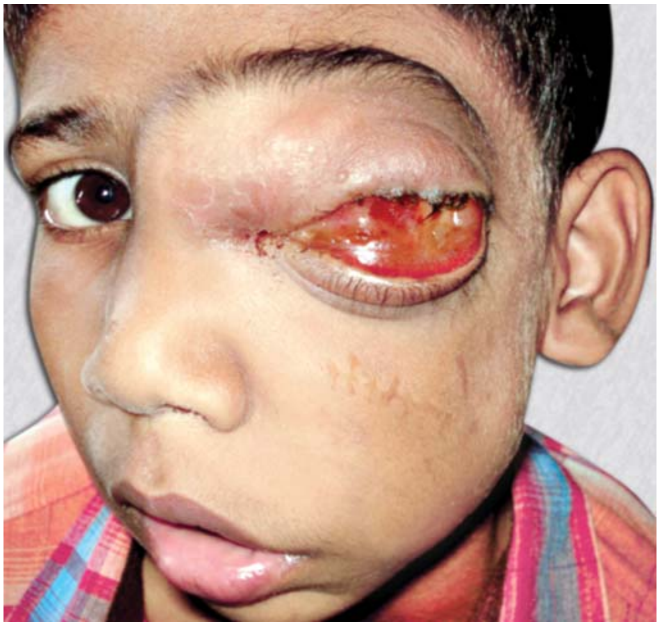

Fig. 27.18: Orbital secondaries from adrenal neuroblastoma.

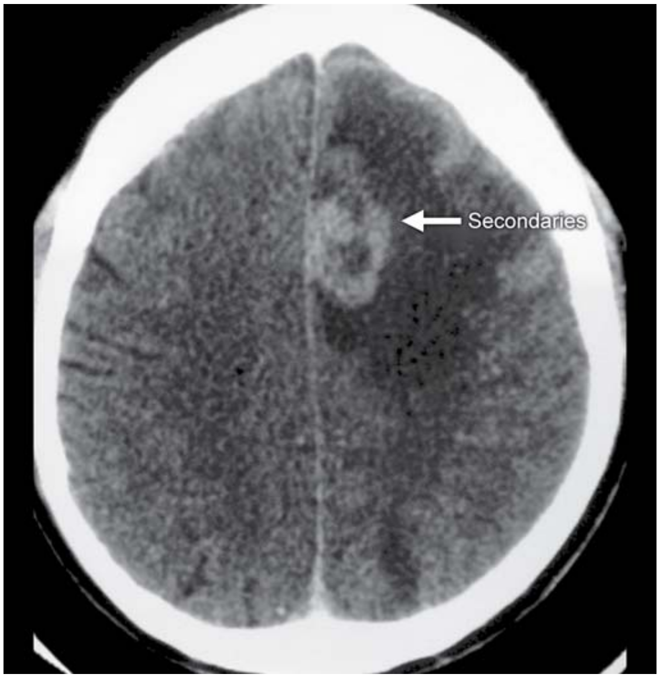

Fig. 27.19: CT scan head showing brain secondaries.

Primary Brain Tumours

1. ***Gliomas (43%).***
   a. *Astrocytomas* are the commonest type. They are usually malignant. They can occur anywhere in the cerebral hemispheres, medulla, brainstem. Peak incidence is in 4th decade.

   They can be *diffuse, solid or cystic*. They contain star-shaped cells resembling adult neuroglial cells. Astrocytic

Disease is no absolute physical entity but a complete intellectual construction, an amalgam of biological state and social definitions.
—Charles E Rosenberg

gliomas are graded as *Grades I, II, III, IV based on the quantity of adult and primitive cells*.

Grade I – Cystic
Grade II – Diffuse
Grade III –Anaplastic
Grade IV – **Glioblastoma multiforme** – It is high grade aggressive type of astrocytoma. It is treated by surgical removal / debulking; high dose radiotherapy, chemotherapy with carmustine inserted into the surgical cavity and oral temozolomide. Median survival is 12 months; 2-year survival rate is 25% or less.

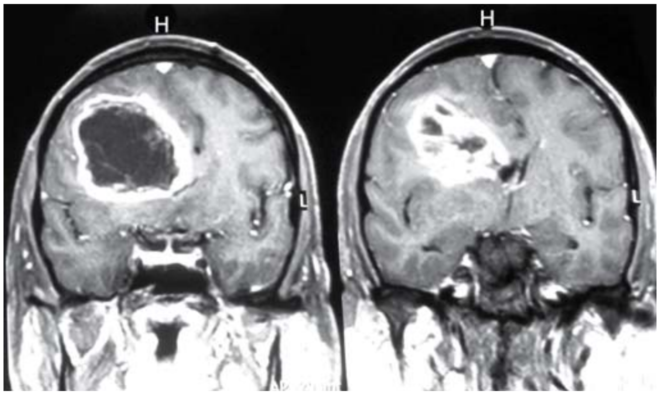

Fig. 27.20: CT scan showing astrocytoma – aggressive variety—grade IV—gliobalstoma multiforme.

b. *Oligodendrogliomas*.

c. *Spongioblastoma polare* arises from the primitive spongioblasts, affects optic chiasma, third ventricle, hypothalamus. They are irremovable but are radiosensitive.

d. *Medulloblastoma* occurs in children, affecting vermis of the cerebellum which grows rapidly with seedling elsewhere in the brain.

e. *Ependymomas*: Cells here resemble ependymal cells. It can occur through out the cerebral hemespheres.

2. ***Meningiomas (18%)*****:** They are usually globular, arising from the arachnoids. Tumour gets attached to the dura. It gets blood supply from dural arteries and veins, from emissary veins and veins of diploe and scalp. Along these veins tumour cells invade the bone, causing bone destruction and reactive hyperostosis.

   Meningiomas are *classified as fibroblastic, endothelial and angioblastic*.

Sites

- Parasagittal
- Frontobasal
- Posterior fossa
- Choroid plexus

Microscopic: It contains whorls of spindle cells, with central hyaline material, with *psammoma bodies*.

Meningioma

- 18% common
- Arising from arachnoids
- Gets blood supply from dural vessels
- Invades the skull bone through emissary and diploe veins.
- Destruction and reactive hyperostosis of bone is common
- *Psammoma bodies* are often seen histologically
- CT/MRI diagnostic
- Surgery is the choice therapy
- It has got good prognosis

3. ***Schwannoma (8%)*****:** Common in auditory nerve, also called as *acoustic neuroma*.

   Occurs in the internal auditory meatus which projects into the cerebellopontine angle (C- P angle), compressing 5, 6, 7, 8th nerves. It presents with compressive features like unilateral deafness, trigeminal neuralgia, squint, cerebellar compression syndrome.

4. ***Pituitary tumours (12%)***.
5. ***Craniopharyngiomas (5%)***.
6. ***Blood vessel tumours (2%)***.

Clinical Features

- Initial period of silent growth.
- Focal syndromes with epilepsy.
- Raised intracranial pressure with headache, vomiting, deterioration of level of consciousness, altered vision, slow pulse, high BP, papilloedema.
- Brain displacement and stage of coning.

Note:
First sign in acoustic neuroma is loss of corneal reflex.

Specific Features

- Frontal lobe tumours: Personality and emotional changes, epilepsy of generalised type, contralateral facial weakness.
- Parietal lobe tumours: Jacksonian epilepsy, progressive hemiparesis, astereognosis, acalculia.
- Occipital lobe tumours: Aura of flashing of light in contralateral field, homonymous hemianopia.
- Temporal lobe tumours: Progressive aphasia, visual, auditory, smell and taste hallucinations, hemiparesis, superior quandrantic hemianopia.
- Midline tumours: Produces bilateral hydrocephalus.
- Tumours of the third ventricle (colloid cyst is common): Causes bilateral hydrocephalus, progressive cerebral atrophy, dementia, sexual precocity, endocrine disturbances.
- Pineal tumours: Causes precocious puberty.
- Cerebellar vermis tumours: Usually medulloblastomas, occur in young children, presents with progressive hydrocephalus and features of herniation of cerebellar tonsils through foramen magnum.
- Cerebellar hemisphere tumours: Commonly are astrocytomas, produce cerebellar syndromes, nystagmus.

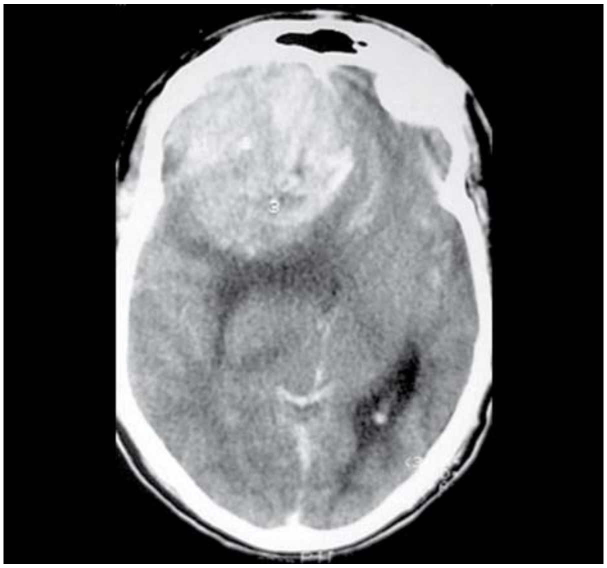

Fig. 27.21: CT scan head showing large meningioma frontal region.

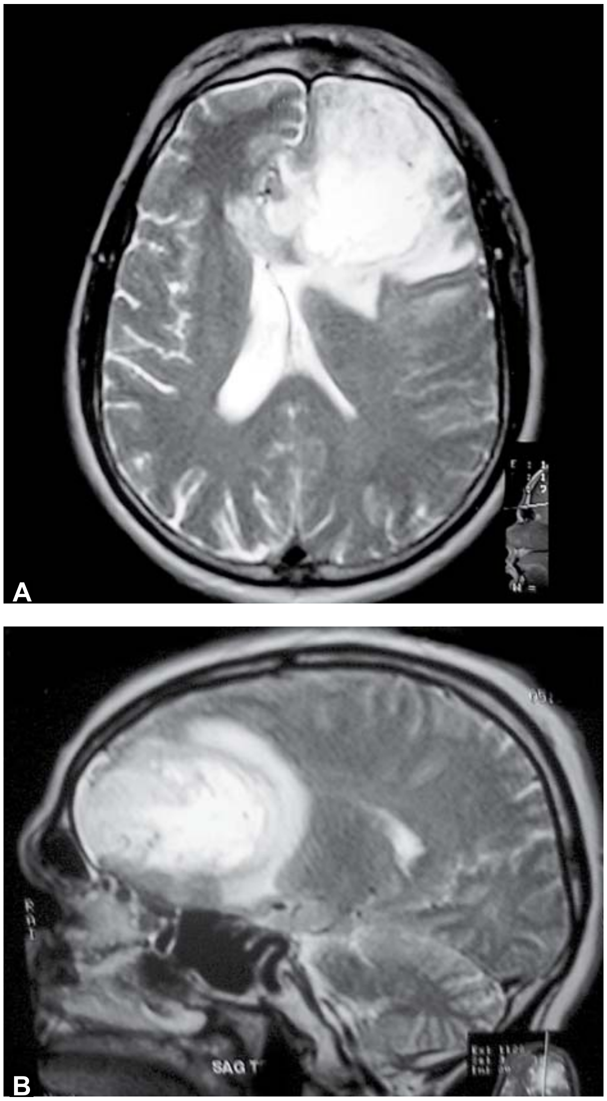

Figs 27.22A and B: MRI brain showing glioma brain.

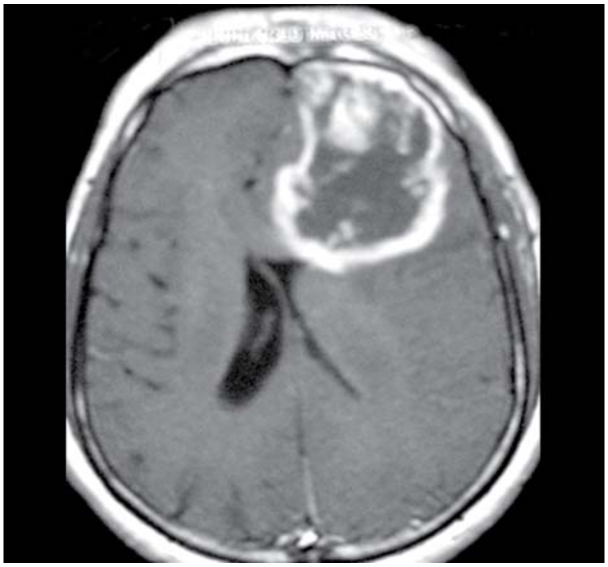

Fig. 27.23: MRI brain showing glioblastoma multiforme. It is very aggressive malignant tumour.

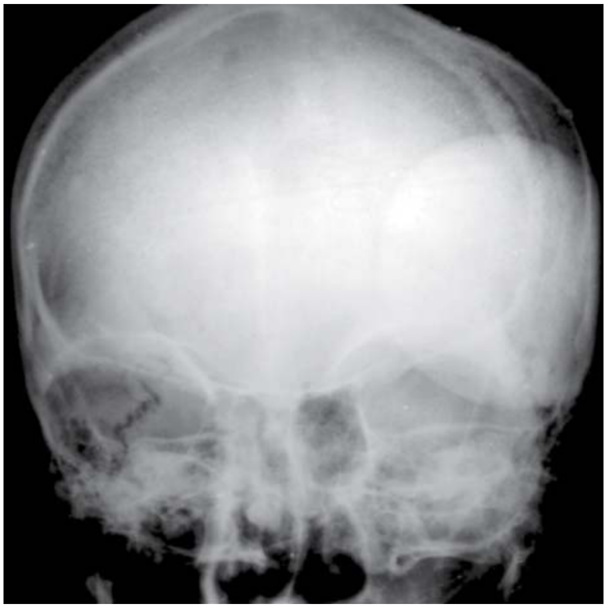

Fig. 27.24: Skull X-ray showing mass calcified lesion. It could be meningioma, dermoid extending intracranially.

Investigations

- X-ray skull.
  - Calcifications like in meningiomas, craniopharyngiomas.
  - Separation of sutures.
  - A beaten silver appearance.
  - Lateral displacement of pineal body.
  - Hyperostosis, expansion, destruction in skull bones.
- Isotope scan.
- CT scan.
- MRI.

A very bold surgeon is the one who realise that, his patient takes all the risks.

- Positron Emission Tomography (PET).
- Carotid angiogram (Introduced by Egas Moniz).
- Ventriculography.
- EEG.

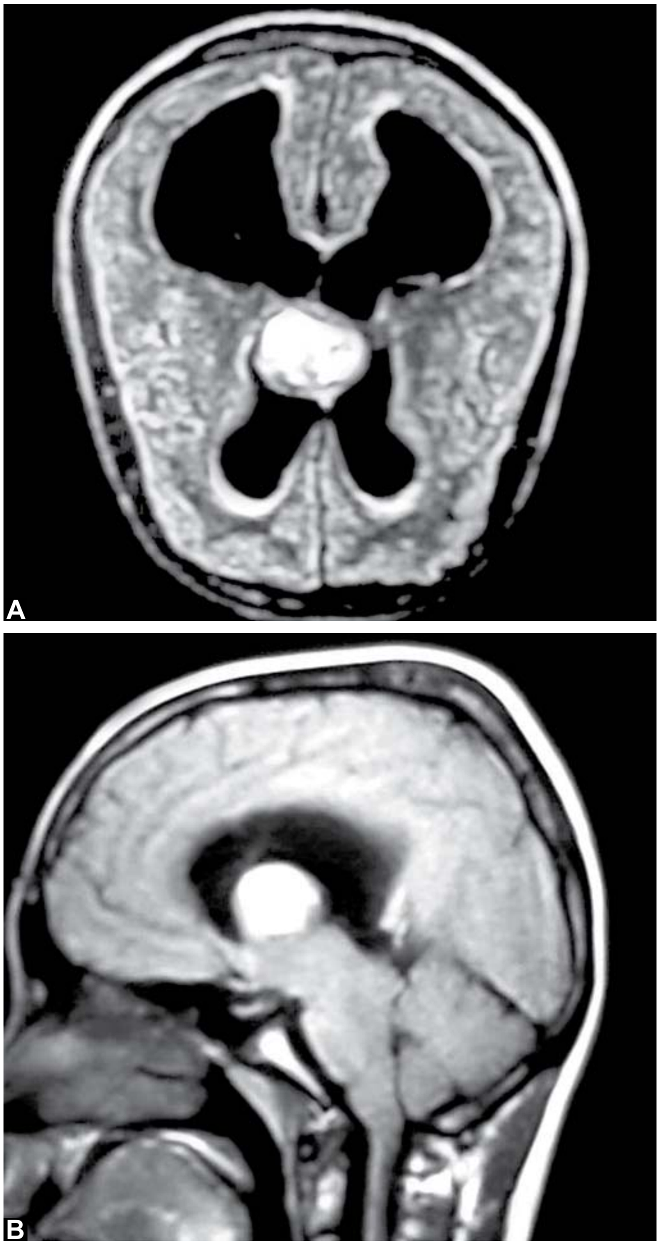

Figs 27.25A and B: Craniopharyngioma from sellar region.

Treatment

- Relief of raised intracranial pressure:
  a. Ventricular tap and drainage through a posterior parietal burr hole.
  b. Tapping of cystic tumours and abscesses.
  c. Administration of mannitol.
  d. Emergency decompression by partial removal of tumour.
  e. Steroid therapy—dexamethasone.
- Establishment of pathological diagnosis:
  a. Burr-hole and biopsy.
  b. Craniotomy and biopsy using brain cannula.
  c. Frozen section biopsy.
  d. CT guided stereotactic biopsy.
- Removal of benign tumours—by different craniotomy approaches.
- Decompressive surgeries for malignant tumours.
- Shunt surgeries to drain CSF—ventriculoperitoneal shunt or ventriculoatrial shunt.
- Radiotherapy—external radiotherapy is used as primary treatment or as an adjuvant therapy after surgery.
- Chemotherapy is occasionally used.

Prognosis

Tumour which is benign and surgically accessible has better prognosis.

PITUITARY TUMOURS

> *A young student…suddenly (sleepy and lethargic) died in convulsions within two weeks. In his brain the ventricles were found swollen with blood… which had gone down to the base of the skull…. The olfactory bulbs were not swollen at all. The pituitary gland…was completely blocked by a…viscous and gelatinous mass about the size of a small bean… Hence a dropsy of the brain ensued.*
>
> —**Richard Lower,** 1672

Classifications I

1. ***Eosinophil (Acidophil) adenomas***: Tumour is usually small. Rarely it causes compressive features.

   It secretes excess growth hormone causing acromegaly in adults and gigantism in children.
2. ***Chromophobe adenomas*** are common in females and in the age group—20-50 years.

   Initially it is *intrasellar* and after sometime becomes *suprasellar*. Later, it extends *intracranially* often massively, causing features of intracranial space occupying lesion. It presents with myxoedema, amenorrhoea, infertility, headache, visual disturbances, bitemporal hemianopia, blindness, intracranial hypertension, epilepsy.

   Differential diagnosis: Meningiomas, aneurysms.

   CT scan, angiogram, X-ray skull are diagnostic.

   Treatment is surgical decompression by craniotomy through subfrontal approach or transsphenoidal approach. Deep external radiotherapy and steroids are also used.
3. ***Basophil adenomas*** are usually small. They secrete ACTH and presents as Cushing's disease with all its features.
4. ***Prolactin-secreting adenomas*** causes infertility, amenorrhoea and galactorrhoea.

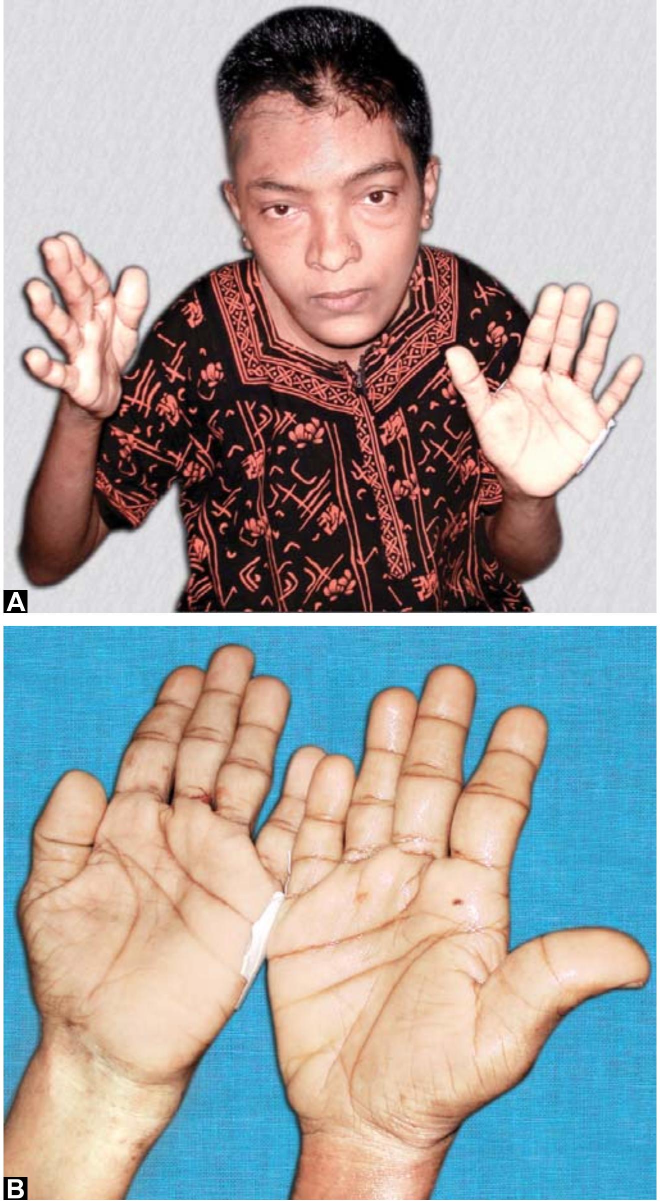

Figs 27.26A and B: Acromegaly due to pituitary tumours. Note the operated scar in the frontal region.

Classification II

- Hypersecreting
- Hyposecreting by compression and atrophy

Classification III

- Micronodular: Tumour size less than 10 mm
- Macronodular: Tumour size more than 10 mm

Stages

- Stage of intrasellar development
- Stage of suprasellar extension
- Stage of massive intracranial extension

Investigations

- X-ray skull—shows calcifications, destruction of sella turcica, mass lesion, enlarged pituitary fossa.
- CT scan.
- MRI.
- Hormone assay—like serum prolactin, growth hormone, ACTH, steroids, sex hormones, etc.

Treatment

- Surgery: By subfrontal craniotomy approach or transphenoidal approach.
  Care should be taken not to injure optic chiasma, arteries, cavernous sinus.
- External radiotherapy.

CRANIOPHARYNGIOMAS

They are large masses with cystic cavities, lined by ciliated epithelium containing cholesterol crystals. Areas of calcifications may be present and coral-like masses may be formed. They are adherent to the basal arteries and adjacent nerves. They are *irremovable.* They are tumours of sellar region.

Clinical Features

- *Intrasellar craniopharyngiomas* inhibits sexual maturation causing obese, impotent dwarf with bitemporal hemianopia (due to compression of optic chiasma)—***Frolich's syndrome.***
- *Suprasellar craniopharyngiomas* produces Frolich's syndrome; pressure on hypothalamus which controls sleep and water metabolism (causes somnolence and diabetes insipidus).
- *Massive intracranial extension* causes intracranial hypertension and also hydrocephalus by obstructing CSF flow.

Investigations

- Skull X-ray shows calcification.
- CT scan is diagnostic.

Treatment

- Through craniotomy or through transsphenoidal approach cystic tumours are evacuated.
- Ventriculoatrial shunt has to be done to drain CSF in case of hydrocephalus.
- Ventriculocisternostomy—*Torkildsen's operation is done in cases of large masses blocking the 3rd ventricle obstructing the CSF outflow.*
- Radiotherapy.

SPINAL DYSRAPHISM

It is posterior midline congenital deformity of spine due to incomplete fusion of ectodermal, mesodermal or neuroectodermal elements either singly or in combination. It is due to arrest of closure of neural tube.

Truth always exists; only lies are invented.

Types

Myelocele

There is failure of closure of neural as well as vertebral arch. There is raw oval defect uncovered over central canal which is incompatible with life and newborn infant dies soon. Loss of lower limb power, incontinence of urine, talipes, CSF discharge, and meningitis is common. Very occasional infant who survives develop a skin coverage from periphery but neurological problems are irreversible. It is common but least survival chance.

Syringomyelocele

It is rarest type. Central canal of spinal cord is dilated with bulging out of the canal under the skin of the dilated centrtal canal as thin cystic sac containing cord with meninges and nerve roots. Gross paralysis is common. It is identified on table as it is difficult to differentiate from myelomeningocele clinically.

Myelomeningocele

It occurs 2 in 1000 live births. Here part of the cord or cauda equina or both present as median cystic swelling at back. It is not transilluminant. It is common in lumbosacral region. Adhesion of spinal cord or cauda equina to sac causes neurological deficit. Kinking, compromised blood supply causes neurological deficit (30%). Monoplegia, paraparesis, paraplegia, talipes and incontinence are the features. It is almost always associated with Arnold Chiari syndrome or Dandy Walker syndrome (atresia of foramina Lushka and Magendie).

Meningocele

Protrusion of spinal meninges through a congenital defect in the spinal laminar defect.

Spina Bifida

Neural arches failure to unite over a limited area and externally there is no protrusion but there is obvious defect.

MENINGOCELE

It is the herniation of meninges through a weak point of spine where bony fusion has not taken place effectively.

Swelling is covered with pia mater and arachnoid mater without dural covering and contains CSF.

Clinical Features

- Present since birth.
- Soft, cystic, fluctuant with transillumination.
- Signs of compressibility.
- Expansile impulse when asked to cough or when child cries.

Location

- Lumbosacral—commonest.
- Occipitocervical—2nd most common.

Investigations

- CT scan head to look for hydrocephalus.
- MRI spine.

Treatment

Excision as early as possible. Transverse elliptical incision. Closure of defect by plication. Approximation of muscles.

Early closure prevents infection.

Complications

- Ulceration.
- Haemorrhage.
- Closure may cause hydrocephalus which needs shunting of CSF.

Meningocele	***Meningomyelocele***
• Membranes content • Soft, cystic • Brilliantly transilluminant • Longitudinal furrow absent • No neurological defect • Good prognosis after repair	• Membranes with nerve roots • Soft to firm • Nontransilluminant • Longitudinal furrow is seen due to adherence of nerve roots • Trophic ulcer, bowel/urine incontinence, motor problems • Neurological deficit present • Not good results

SPINA BIFIDA

- It is also classified under *spinal dysraphism*.
- It is failure of enfolding of nerve elements within the spinal canal during developmental period.
- It is usually seen in lumbosacral region. There is failure of fusion of one or more posterior vertebral arches.
- It is often associated with other anomalies.

Sites

- Lumbosacral
- Thoracolumbar

Types

a. ***Spina bifida occulta***—commonest type
  - There is dimpling of skin with dermoids, lipomas in the site. Impulse on coughing can be seen.

- Initially there is no neurological deficit but later due to tethering, traction on dura, infection, can lead neurological deficits.

b. ***Spina bifida aperta***

- Here neurological deficit is present. It may be *myelomeningocele* wherein spinal cord and nerve roots are in the sac. It may be *meningocele* wherein sac consists of meninges and fluid only.
- Meningocoele is brilliantly transilluminant. Myelomeningocoele is not transilluminant.

Clinical features

- Motor paralysis
- Sensory paralysis
- Visceral paralysis with incontinence of urine and faeces
- Swelling in the spine at the site of the lesion, may be lipoma or dermoid, with impulse on coughing
- Bony defect at the site
- Hydrocephalus

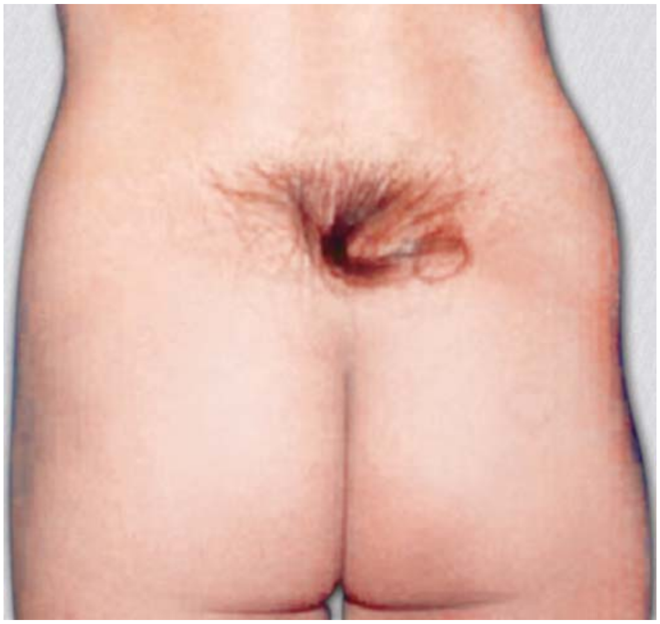

Fig. 27.27: Spina bifida occulta. Note the dimple and tuft of hair.

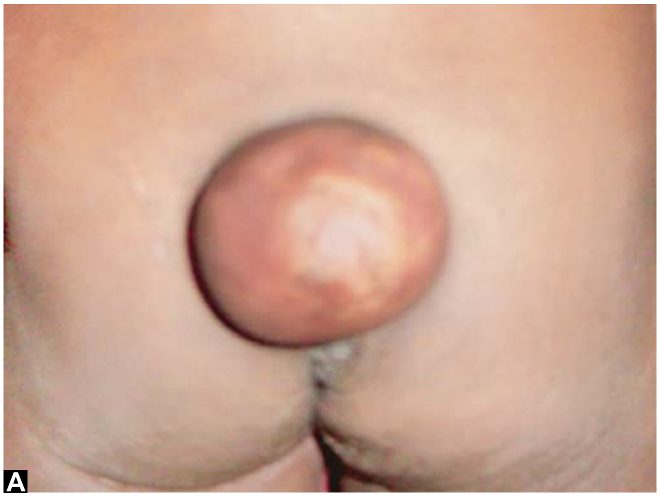

Fig. 27.28A

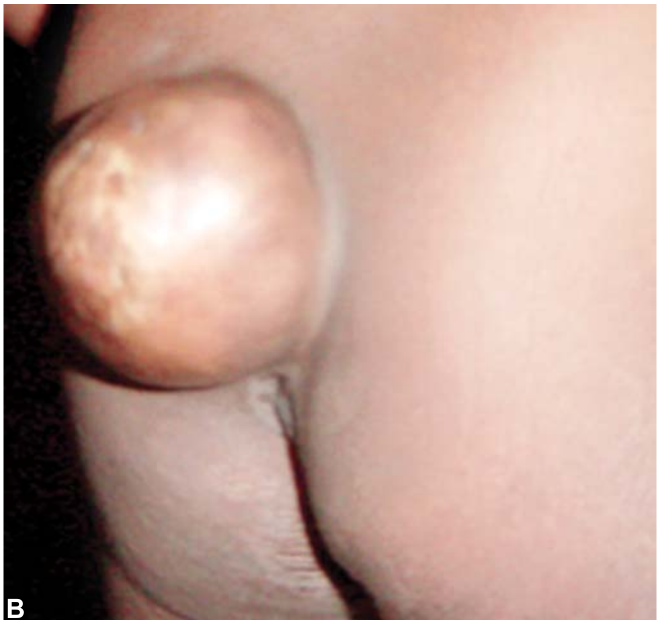

Fig. 27.28B

Figs 27.28A and B: Spina bifida with protruding dermoid through the defect.

Investigations

- Plain X-ray of the spine.
- CT scan/MRI of spine and head.

Treatment

- Correction of deformity.
- Maintaining the visceral function.
- Development of limb function.
- Ventriculoperitoneal shunt or ventriculoatrial shunt surgery for hydrocephalus.

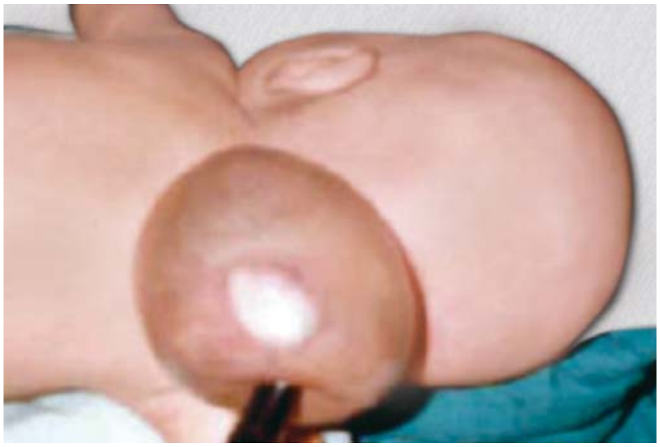

Fig. 27.29: Infant with encephalocele. It is non-transilluminant whereas meningocele is brilliantly transilluminant.

INTERVERTEBRAL DISC PROLAPSE (IVDP)

- Herniation of part of the gelatinous nucleus pulposus through a rent in the annulus fibrosus, commonly at posterolateral part which is a weak point.

The first sign of Pott's disease is deepened midline spinal sulcus at about T10 level due to increased muscular tonicity of sacrospinalis muscle. It is a NATURE'S PLASTER CAST. **—Francis E Jardine**

- ***In cervical spine*** discs between C_5 and C_6; C_6 and C_7 are commonly affected impinging 6th and 7th nerves respectively.
- ***In lumbosacral region,*** discs between L_5 and S_1; L_4 and L_5 are commonly affected impinging first sacral nerve and 5th lumbar nerve respectively.
- Initially protrusion bulges through the pain sensitive posterior longitudinal ligament causing back pain.
- Later it herniates through the posterior ligament compressing the nerve causing typical root pain.

Types

- Posterolateral is common.
- Central-posterior is not common. It is due to trauma. It is severe type, even though rare.

Secondary changes: Osteoarthritis of joints.

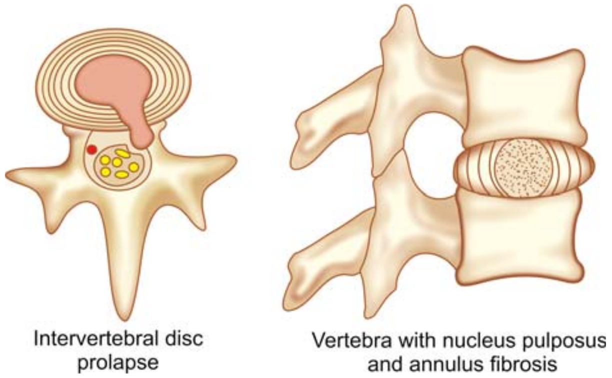
Intervertebral disc prolapse
Vertebra with nucleus pulposus and annulus fibrosis

Fig. 27.30: Intervertebral disc prolapse.

Clinical Features

- Pain in the distribution of the root which gets aggravated by straining, coughing, twisting, stooping.
- Pain radiates along the distribution of the nerve with tingling and numbness.
- Lumbar scoliosis is significant.
- Restricted forward flexion but free lateral flexion.
- Positive straight leg raising on the affected side with positive Lasegue's manoeuvre.
- Wasting in the muscles with blunting of sensation, with absence knee and ankle jerks.
- Loss of bladder sensibility and retention of urine.

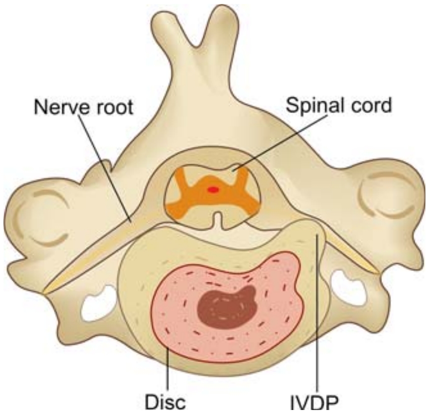

Fig. 27.31: Intervertebral disc prolapse.

Differential Diagnosis

- Spinal cord tumours.
- Tuberculosis of the spine.
- Osteoarthritis.
- Ankylosing spondylitis.
- Spondylolisthesis.
- Vascular problems.
- Retroperitoneal and pelvic tumours.

Investigations

- X-ray spine both AP and lateral. It is not useful in acute prolapse.
- Myelography or Radiculography.
- CT scan.
- MRI is very useful method.

Treatment

- Rest in bed.
- Spinal jacket.
- Continuous or intermittent traction.
- Analgesics and relaxants.
- Intradiscal injection of the chymopapain enzyme which dissolves the fibrocartilaginous tissue and nucleus pulposus.
- Surgical treatment:

  Indications:
  - Persistent pain for 12 weeks which is not relieved by drugs.
  - Severe neurological disturbances like involvement of bladder, bowel or sexual functions.

  Surgery: Discectomy.

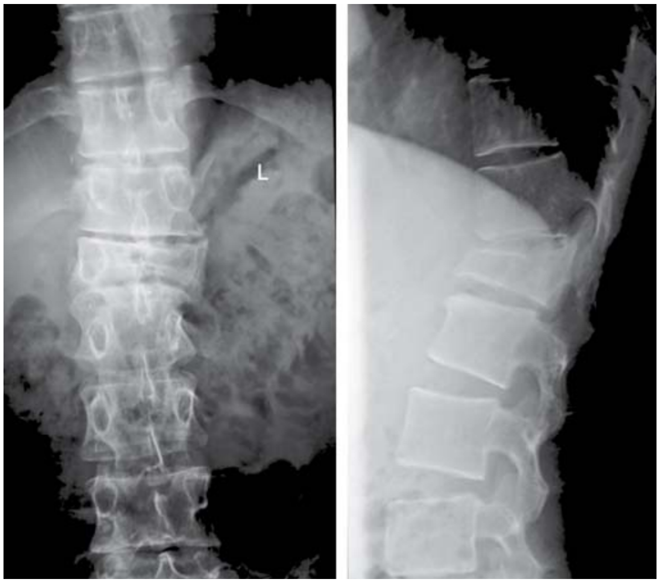

Fig. 27.32: Compression fracture spine—traumatic, pathological.

TUBERCULOSIS OF SPINE (CARIES SPINE)

- It is commonly of *secondary* type.
- Primary focus is in the lungs or lymph nodes.

- Commonly involved vertebra is dorsilumbar vertebra T_{10}.

Mode of infection: Mainly haematogenous.

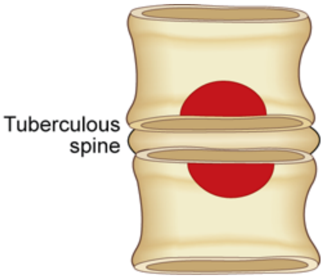

Fig. 27.33: Tuberculosis of spine.

A. Tuberculosis of Cervical Vertebra

Common in C_6 and C_7 vertebra.

Clinical Features

- Pain in the neck, often referred to occipital region through suboccipital or posterior auricular nerves.
- Pain, when referred through anterior cervical primary rami causes *brachial neuralgia.*
- Rigidity in the neck.
- Patient supports the chin on the palm of his hand (*Rust sign*).
- Tenderness and paraspinal spasm with rigidity of the skull muscles.
- All passive and active movements of spine are painful and restricted.
- Cold abscess may be present in the neck, posterior to sternomastoid muscle or in the retropharyngeal region or in the axilla.
- Chronic tuberculous retropharyngeal abscess in midline, situated behind the prevertebral fascia. It is painless, soft, often causing dysphagia and dyspnoea.

Differential diagnosis

- Cervical spondylosis
- Congenital torticollis
- Secondaries
- Spinal cord tumours

Investigations

- X-ray cervical spine, chest X-ray.
- MRI is ideal.

Treatment

- Antitubercular drugs.
- Anterolateral decompression of vertebra.
- Drainage of cold abscess.

B. Tuberculosis of Thoracolumbar Vertebra (Pott's Disease, Caries Dorsilumbar Spine)

Commonest site is dorsilumbar region —T_{10}.

Reasons

- It is the junction of fixed (thoracic) and mobile (lumbar) segments of spine
- It bears the maximum stress
- It has got large amount of cancellous bone
- Bulky nucleus pulposus
- Relative avascularity of the body of dorsilumbar vertebra
- Close proximity to the thoracic duct through which infection can occur
- Adjacent prevertebral veins

Types

- ***Tuberculosis of the body of vertebra***
  - *Tuberculous metaphysitis*—Here contiguous surfaces of the vertebra are involved. It is the most common type in children.
  - *Tuberculous osteomyelitis* (central type—through central artery of the vertebra)—It occurs in the centre of the body of vertebra.
  - *Tuberculous periostitis* (peripheral or periosteal) — It is seen in adults. It is situated deep to the anterior longitudinal ligament.

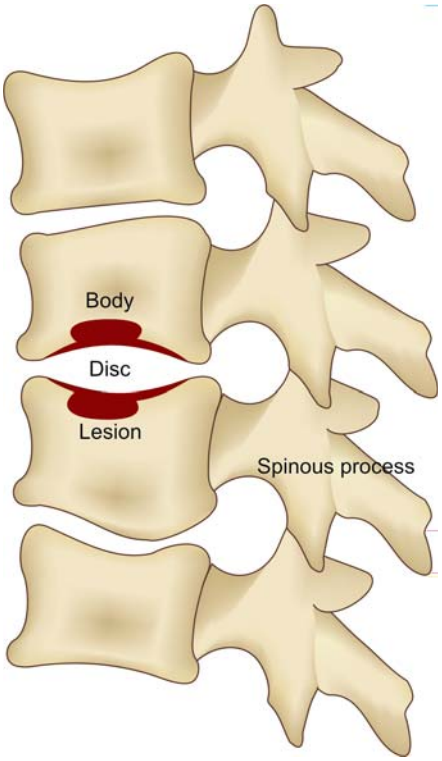

Fig. 27.34: Tuberculosis of spine. Note the area of involvement.

- ***Tuberculosis of the appendages***
  - Spinous process.
  - Pedicle.
  - Transverse process.
  - Lamina.

Features

- Contiguous surfaces of paired vertebra are involved.
- Lower half of one vertebra and upper half of adjacent vertebra below, with intervertebral joints are involved as they share a common blood supply.
- Destruction of the vertebra and intervertebral disc; collection of caseous material behind the anterior longitudinal ligament.

Pathology

- Tuberculous endarteritis
- Decreased blood supply
- Formation of tubercles with caseation
- Destruction of bone
- Failure of subperiosteal new bone formation

Effects

1. Deformity—*Kyphosis (Gibbus*) is excessive posterior curvature of the spine. It is due to complete anterior collapse of the vertebra.
2. *Cold abscess* formation:
   a. Close to the midline, when caseating material passes through the medial dorsal cutaneous nerves.
   b. Away from the midline, when caseating material passes through the lateral cutaneous nerve.
   c. Cold abscess may form in relation to psoas muscle (in dorsilumbar or lumbar).

   It passes through psoas sheath. It causes psoas spasm causing flexion of hip with inability to extend. If abscess extends below the inguinal ligament, then it is cross-fluctuant.
3. Spinal cord involvement—causes *paraplegia* (10%). This is common in thoracic spine, as the spine is narrow.
   a. *Early paraplegia ('Paraplegia in flexion', 'paraplegia with active disease'*).
   It is due to:
      - Pressure by caseating material, granulation tissue and sequestrum.
      - Tuberculous meningitis.
      - Spinal cord oedema.
      - Endarteritis of segmental spinal artery.
      - Pressure from distended disc.

   b. *Late paraplegia* ('*Paraplegia with a healed disease', 'paraplegia with extension').*
   It is due to:
      - 'Gibbus' causing stretching of the spinal cord.
      - Longitudinal shrinkage of the spinal cord due to gliosis.

Clinical Features

- Pain in the back, often with tingling sensation along the distribution of nerve roots.
- In upper thoracic tuberculosis, typical '*military attitude'* with raised shoulder which is drawn backwards is seen.
- Kyphosis (*gibbus)* is the commonest deformity. It can be angular but also can be rounded.
- Tenderness in the spine with paraspinal spasm.
- All active and passive movements of spine are restricted.
- '*Coin test'*—inability to pick up a coin is positive.
- Weakness and decreased power in the limb muscles, with altered sensation.

Investigations

- X-ray spine shows:
  - Narrowing of the disc space.
  - Wedging of the vertebra.
  - Rarefaction of the adjacent vertebra.
  - Soft tissue shadow of cold abscess.
- Chest X-ray PA view.
- ESR.
- Radioisotope bone scan.
- CT scan or MRI of spine.

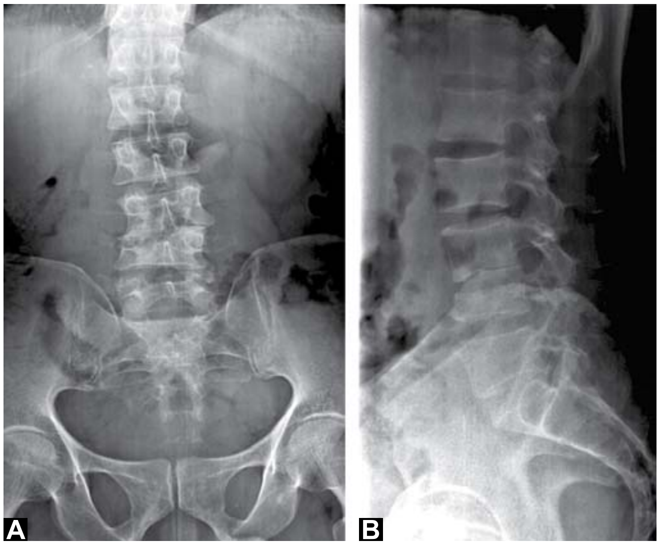

Figs 27.35A and B: Typical tuberculosis of spine involving L4 and L5 spine X-ray features.

Differential Diagnosis

- Secondaries in the spine.
- Spinal tumours.
- Scheuermann's disease (osteochondritis).
- Old compression fracture of the vertebra.
- Ankylosing spondylitis.

Complications of Pott's spine

- Kyphosis
- Cold abscess
- Paraplegia
- Sinus formation
- Dissemination

Treatment

- Antituberculous drugs for one year.
- Rest with plaster jacket (SPICA).

Patient gains weight, X-ray shows re-calcification and healing. ESR becomes normal. Pain and spasm disappears.

Indication for surgery

- When the disease is progressive
- When there is neurological manifestation
- Cold abscess formation

Surgeries

Note:

No posterior approach, no laminectomy for TB spine.

- Anterolateral approach and costotransversectomy with removal of all caseating material with bone grafting.
- Drainage of cold abscess.
- Posterior spinal fusion using bone graft. Bone graft is taken from the iliac crest.

SPINAL TUMOURS

Classification

A. ***Extradural tumours***
- Commonest is secondaries.
  - They lie between the dura, bone and ligamentum flavum.
  - Primary sites are bronchus, breast, prostate, kidney. Rare extradural tumours are sarcomas, osteoclastoma, chondroma, lipoma.

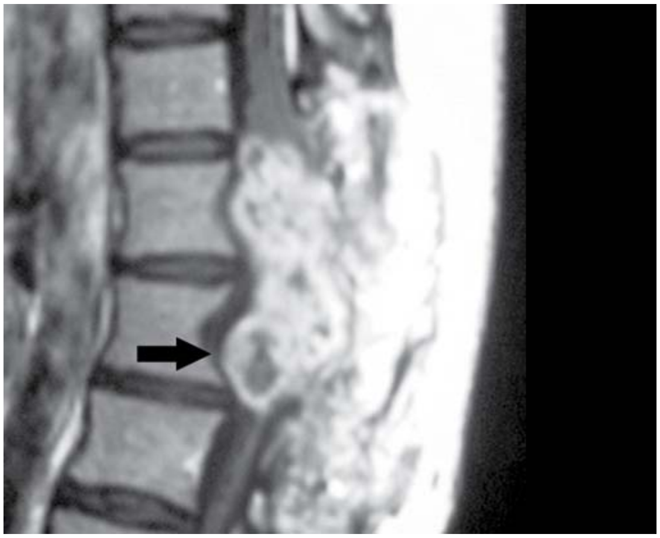

Fig. 27.36: MRI spinal cord showing spinal cord tumour—ependymoma.

B. ***Intradural tumours***
1. *Extramedullary* (75% of intradural)
   - Neurofibromas are the commonest, more common in males.
     - Arises from the posterior nerve root. It can be *'dumb-bell'* tumours.
   - Spinal meningiomas are seen exclusively in females.
2. *Intramedullary* (25% of intradural)
   - Commonest is diffuse gliomas (50%).Others are ependymomas, vascular malformations.
   - Common in cervical cord.

Clinical Features

- Weakness of limbs, often with paraplegia.
- Tingling and numbness.
- Disturbance in micturition.
- Changes in tendon reflexes.
- Pain in the back.

Spinal cord or cauda equina compression is a surgical emergency.

Investigations

- X-ray shows widening of space and destruction.
- Myelography.
- Lumbar puncture *per se* has no role, as it is dangerous.
- CSF below the block may be yellow and proteinaceous and is called as '*Frolin's syndrome'*.
- Protein levels in lumbar CSF is usually raised in the presence of lumbar tumours, commonly 'Schwannomas'.
- CT scan.
- MRI.

Surgery

- Surgery is the main treatment—*Decompression of spinal cord and removal of tumour* by doing laminectomy.
- Adjuvant therapy: Radiotherapy and chemotherapy.
- Intrathecal methotrexate is also beneficial.

No head injury is so slight that it should be neglected, or so severe that life should be despaired of.

—Hippocrates, Father of Medicine

Chapter 28 Thorax

CHAPTER OUTLINE

- Chest Injuries
- Fracture Ribs
- Flail Chest and Stove in Chest
- Pneumothorax
- Tension Pneumothorax
- Haemothorax
- Pleural Tap
- Bronchoscopy
- Empyema Thoracis
- Empyema Necessitans
- Lung Abscess
- Intercostal Tube Drainage
- Shock Lung
- Pulmonary Embolism
- Surgical Emphysema
- Lung Cysts
- Mediastinal Tumours
- Thymomas
- Pancoast Tumours
- Chest Wall Tumours
- Pericarditis
    - Acute Pericarditis
    - Chronic Constrictive Pericarditis
- Pericardial Tap
- Cardiac Tamponade
- Diaphragmatic Hernia
- Pulmonary Complications during Postoperative Period
- Surgical Management of Pulmonary Tuberculosis
- Video Assisted Thoracoscopic Surgery

CHEST INJURIES

Types

- Crush injuries involving lung, pleura, ribs.
- Single rib fracture.
- Two or more rib fractures.
- *Steering wheel injury*—causes multiple rib fractures, bilaterally often with flail chest, with fracture dislocation of upper end of sternum.
- Stove in chest or flail chest.
- Traumatic pneumothorax.
- Haemothorax, haemopneumothorax, with fracture ribs.
- Tension pneumothorax.
- Pericardial, cardiac injuries and rupture of bronchus.
- Associated injuries in liver, spleen, diaphragm, major vessels.

Causes

- Road traffic accidents.
- Industrial accidents.
- Blast injuries.
- Crush injuries.
- Stab injuries.

- In children, ribs are malleable and so fracture ribs are rare.
- In elderly because of rigid ribs fracture is common.
- First and second ribs are protected by clavicle and so their fracture is uncommon.
- 11th and 12th ribs are floating ribs and so their fracture is rare.

Classification of Chest Injuries

As per the *American College of Surgeons Committee on Trauma* chest injuries are classified as:

1. Immediately life-threatening injuries.
2. Potentially life-threatening injuries.

Immediately life-threatening injuries	***Potentially life-threatening injuries***
• Airway obstruction	• Tracheobronchial disruption
• Tension pneumothorax	• Aortic disruption
• Open pneumothorax	• Diaphragmatic disruption
• Massive pneumothorax	• Esophageal disruption
• Cardiac tamponade	• Cardiac contusion
• Flial chest	• Pulmonary contusion

Pathological Effects of Chest Injuries

Immediate	*Late*
Hypoxia	Empyema, fibrothorax
Hypercarbia	Lung abscess
Acidosis	Mediastinitis
Hypovolemic shock	Cardiac arrhythmias
Bronchospasm	

Clinical Features of Thoracic Injuries

- History of trauma, painful breathing, cough, haemoptysis, pain in the chest wall, sometimes external wound may be present (in communicating wounds).
- Features of shock when major vessels are involved, i.e. tachycardia, hypotension, cold periphery.
- Respiratory distress—tachypnoea, cyanosis, respiratory difficulties.
- Tenderness over the fracture site.
- Dullness on percussion with decreased breath sounds signifies haemothorax. Resonant with decreased breath sound confirms pneumothorax.
- Surgical emphysema with palpable crepitus may be present.

Investigations

- Chest X-ray shows haemothorax, pneumothorax, fracture ribs.
- Hb%, PCV to assess blood loss.
- Blood grouping and cross matching.
- Blood gas analysis, i.e. PO_2 and PCO_2.
- U/S abdomen to look for associated abdominal injuries. FAST (Focused abdominal sonar trauma).
- CT chest and CT abdomen.

Treatment

Initial First Aid

- *Airway:* Prevention of aspiration, plastic airway, intubation, tracheostomy.
- *Breathing*: ICT placement, supportive measures.
- *Circulation:* Fluid therapy, CVP line, blood transfusion.
- Look for disability.
- Expose the patient properly for proper breathing and assisting.
- Assess the patient properly.
- Examine the patient thoroughly.
- Evaluate the patient for associated injuries like of head, abdomen, fracture limbs, spine.

Further Treatment

- Fracture rib without complication is treated with analgesics and rest.
- Haemothorax, pneumothorax should be treated with ***intercostal tube drainage*** (ICT) with underwater seal.
- Flail segment should be treated accordingly.
- Blood transfusion is done whenever required.
- Antibiotics like penicillin, cefotaxime, etc. to prevent infection.
- Nasal oxygen, throat suction.
- Ventilator support with IPPV (Intermittent positive pressure ventilation), with regular monitoring of blood gas.
- Tracheostomy/endotracheal intubation is done whenever required.
- Antibiotics, bronchodilators, often steroids are other supportive measures required.

Indications for thoracotomy

- ❖ Haemothorax more than 1500 ml found when ICT is placed or hourly collection in ICT is 200-300 ml
- ❖ ICT placed shows persistent drainage of blood
- ❖ Diaphragmatic injury
- ❖ When associated with liver and spleen injuries
- ❖ Bronchus and major vessel injuries
- ❖ Haemopericardium
- ❖ Oesophageal and thoracic duct injuries

Principles of Management of Chest Injuries

- Pulmonary physiotherapy.
- Aspiration of secretions—trachea, nasotracheal, oral, pharyngeal.
- Pain relief—oral narcotics, intercostal nerve block, epidural anaesthesia.
- Respiratory supports—encourage coughing, chest percussion, deep inspiration efforts, humidification, mobilisation.
- ICT placement for haemo/pneumothorax.
- Management of shock.
- Focused abdominal sonography on trauma (FAST).
- Surgery when indicated—thoracotomy and proceed.
- Management of complications—DVT and embolism, tracheostomy problems, ICT problems, sepsis, ARDS, empyema treatment, bronchopleural fistula, bronchial stenosis, chylothorax, clotted haemothorax, atelectasis.

Complications

- Infections—empyema, lung abscess, pneumonia, septicaemia.
- Respiratory failure.
- Traumatic asphyxia.
- Traumatic shock lung.
- Disseminated intravascular coagulation (DIC).
- ARDS (Adult respiratory distress syndrome).

FRACTURE RIBS

- Rib fractures are rare in children as ribs are malleable.
- Fracture ribs are common in elderly as ribs are rigid and nonmalleable.
- 1st and 2nd ribs are covered by clavicle and so rarely fractures.
- Floating 11th and 12th ribs rarely get fractured.

There are no disease of aged but diseases among the aged.—**Leonard Larson**

Causes for Fracture Ribs

- Road accidents, chest injuries.
- Direct trauma.
- Blast injuries.

Features of Rib Fracture

- Difficulty and pain during breathing.
- Tenderness over fracture site.
- Spring test is positive.
- Propped up position is more comfortable than lying down.
- Features of haemothorax, lung injuries.
- Features of abdominal injuries like of liver and spleen injuries.

Problems in Fracture Ribs

- Surgical emphysema.
- Haemothorax, pneumothorax.
- Cardiac and or major vessel injury.
- Diaphragmatic injury.
- Flail chest.
- Asphyxia, cyanosis, respiratory failure.
- Pneumonia.

Investigations

- Chest X-ray is diagnostic.
- CT chest or abdomen to check deep or solid organ injury.
- Arterial blood gas analysis.

Treatment

- Control of pain.
- Prevention of infection.
- Respiratory physiotherapy.
- Treatment for definitive conditions like flail chest, ICT placing for haemo/pneumothorax.

Non-traumatic rib fracture

- Stress fracture
- Metastatic fracture
- Metabolic cause like hyperparathyroidism
- Osteogenesis imperfecta
- Child abuse
- Old age—osteoporosis

FLAIL CHEST AND STOVE IN CHEST

- It is fracture *of two or more consecutive ribs,* with each rib having *two or more* fracture sites. Such segment is called as ***flail segment.***
- **Stove in chest** is depression of a portion of chest wall due to severe chest injury, otherwise features and management are like flail chest.

Pathophysiology

Flail segment moves independent of the adjacent thoracic cage. During inspiration flail segment moves inwards (unlike normal thoracic cage which moves outward), and during expiration the segment moves outwards (unlike normal cage which moves inward) causing pathophysiological derangements.

- This *paradoxical respiration* causes reduction in ventilatory lung surface and so respiratory dysfunction.
- *Mediastinal flutter*: Movement of mediastinum during different phases of respiration occurs, often causing kinking of great vessels and sudden cardiac arrest.
- *Pendular movement of air* from one lung to other occurs, and thus preventing atmospheric air to get into both injured and otherside normal lung leading to respiratory failure.
- All these derangements gets aggravated by *haemothorax, pneumothorax* and other associated injuries.
- There is *hypoventilation:* carbon dioxide retention and respiratory failure.

Flail chest

Derangements	***Types***	***Problems***
• Paradoxical respiration • Mediastinal flutter • Pendular movement of air • Pulmonary contusion	• Anterior—near costochondral junction • Lateral—in rib shafts • Posterior—safer, due to support of scapula	• Cardiac arrest • Respiratory failure • Haemothorax • Infection • ARDS

First Aid

The paradoxical movement of flail segment is prevented by applying pressure over it using hand or clothes (It is often life saving.) It makes other lung function adequately.

Investigations

- Chest X-ray—to identify the fracture segment.
- Blood grouping.
- Arterial PO_2 and PCO_2 and serum electrolytes.

Treatment

- Intercostal tube drainage.
- Applying clips to fracture ribs and fixing above and below to normal ribs.
- Antibiotics like penicillins, cefotaxime.
- Blood transfusion, IV fluids.
- Bronchodilators, steroids.
- Ventilator support with IPPV—IPPV is treatment of choice. Assisted ventilation is required for several days until the chest wall stabilises. If ventilator support is required for more than 10 days, then tracheostomy is done to prevent laryngeal stenosis which can occur due to prolonged endotracheal intubation.
- Thoracotomy—when required only.

PNEUMOTHORAX

> *The means we possess of reducing (a diseased lung) to a state of collapse, or of divesting it for a time of its peculiar functions, are equally simple and safe. In those cases in which the disease is placed in one of the lungs only, the remedy (i.e. induced pneumothorax) would appear to be simple, safe, and complete*
>
> —**James Carson,** 1820

It is the presence of air between the layers of pleura.

Classification I

a. Pneumothorax.
b. Hydropneumothorax.
c. Pyopneumothorax.
d. Haemopneumothorax.
e. Artificial pneumothorax.
f. Tension pneumothorax.

Classification II

a. Open pneumothorax
b. Closed pneumothorax
  - Simple
  - Tension

Causes

- Traumatic.
- Spontaneous.
  1. Tuberculous.
  2. Non-tuberculous:
     a. Rupture of emphysematous bullae.
     b. Rupture of solitary lung cyst.
     c. Honeycomb or cystic lung.
     d. Idiopathic.

Spontaneous pneumothorax can be acute, chronic, recurrent.

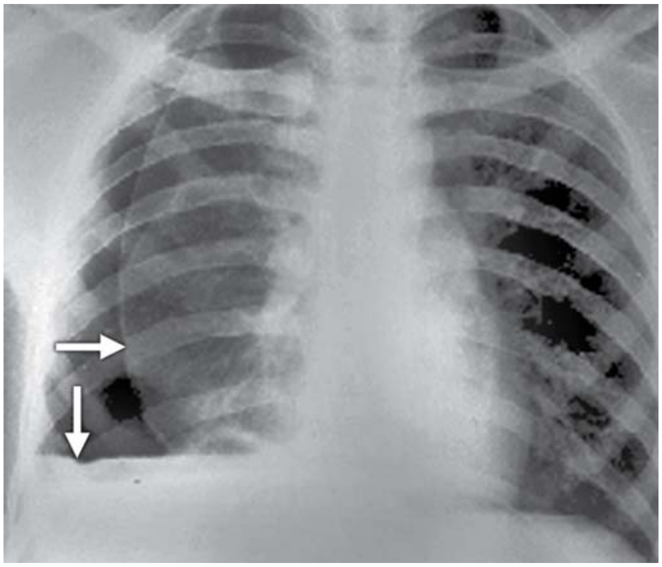

Fig. 28.1: Plain X-ray chest PA-view showing air-fluid level and collapsed lung margin (Hydropneumothorax)

Clinical Features

Hyperresonant, absence of breath sounds, tracheal deviation.

Chest X-ray reveals

- Radiolucency on the affected side
- Absence of lung markings
- Collapsed lung margin

Treatment

- In tension pneumothorax, emergency needle aspiration is done followed by ICT placement.
- ICT placement.
- The cause is treated.
- Often thoracotomy is required, if there is persistent broncho-pleural fistula or ruptured cyst or bullae.

TENSION PNEUMOTHORAX

- During inspiration, air is pumped into the pleural cavity through a valvular opening in the visceral pleura and underlying injured lung.
- Lung collapses first, and as air continuously collects in the pleural cavity, mediastinum shifts towards the opposite side, further decreasing the volume of the functioning lung.
- Further increase in the pleural pressure, reduces the venous return, atrial filling, and ventricular filling and so cardiac output and cardiac function.

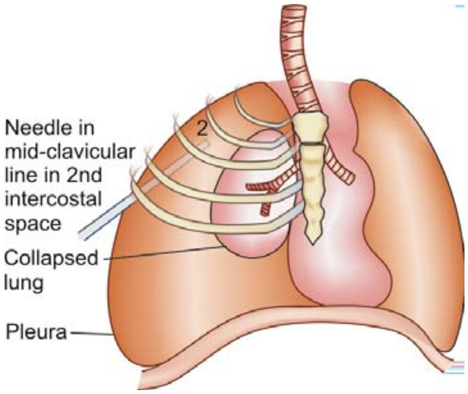

Fig. 28.2: Emergency decompression through 2nd intercostal space.

It causes sudden death and hence emergency treatment is required.

Clinical features

- Tachypnoea and tachycardia
- Decreased/absent breath sounds
- Resonant on percussion with severe mediastinal/tracheal shift
- Cyanosis and hypotension
- Chest pain

Experience teaches slowly at the cost of mistake.

Management

- *Once clinically diagnosed, a wide bore needle is immediately placed in the second intercostal space in midclavicular line*, and a sterile glove is kept on the hub (blunt) end of the needle to create a valve so as to prevent inward sucking of air from outside.
- Nasal oxygen is used.
- Once patient is better, chest X-ray is done.
- Later an intercostal tube is passed.
- Antibiotic, analgesics are given.
- In severe cases ventilator support with IPPV is required.

HAEMOTHORAX

It is blood in pleural cavity. It causes pain, shock, as it is very irritant to pleural cavity.

It is a good culture media for bacteria and so infection is quite common.

Causes

- Trauma.
- Postoperative: pulmonary, cardiac, oesophageal surgeries, cervical sympathectomy, leak from CVP monitor line.
- Tumours of lung, mediastinum, pleura.
- Leaking aneurysms.
- Spontaneous.

There may be rib fractures in traumatic haemothorax.

Clinical features

- Pain in the chest, tenderness
- Difficulty in breathing, dullness on percussion diminished breath sounds
- Features of shock

Investigations

- Chest X-ray.
- Aspiration (pleural tap).
- Chest CT scan.

Treatment

1. ICT placement in the mid-axillary line in the 6th intercostal space.
2. Antibiotics, bronchodilators.
3. Thoracotomy:

Indications of thoracotomy

- Initial chest tube output of 1500 ml of blood or persistent drainage of 200-300 ml/hr.
- Clotted haemothorax is difficult to manage. It requires thoracotomy, evacuation and decortication of lung. Initially liquefaction of the clot is tried by infusing of streptokinase and trypsin into the pleural cavity

Complications

- Infection and empyema.
- ARDS: respiratory failure.

PLEURAL TAP

Indications

- Pleural effusion—both for diagnostic as well as therapeutic purpose. The fluid is sent for culture, cytology, microscopy, specific gravity, biochemical analysis of proteins for diagnosis of tuberculosis, malignancy.
- In empyema thoracis, for diagnostic purpose before placing an ICT.

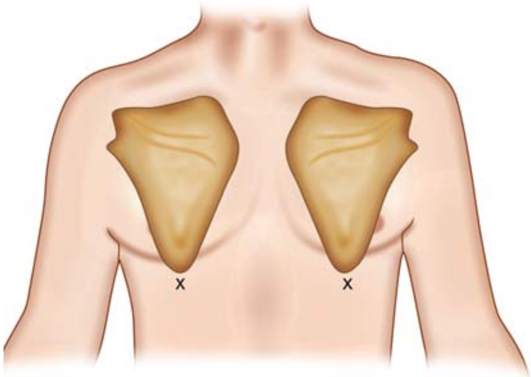

Fig. 28.3: Sites of pleural tap.

Position

In sitting position, leaning forward over a wooden support.

Site

Tip of scapula at 7th intercostal space (posteriorly).
Under local anaesthesia wide bore needle (Abraham needle) is passed to tap the fluid.

Complications

- Infection
- Dry tap or bloody tap
- Sudden vagal shock
- Pain and respiratory distress

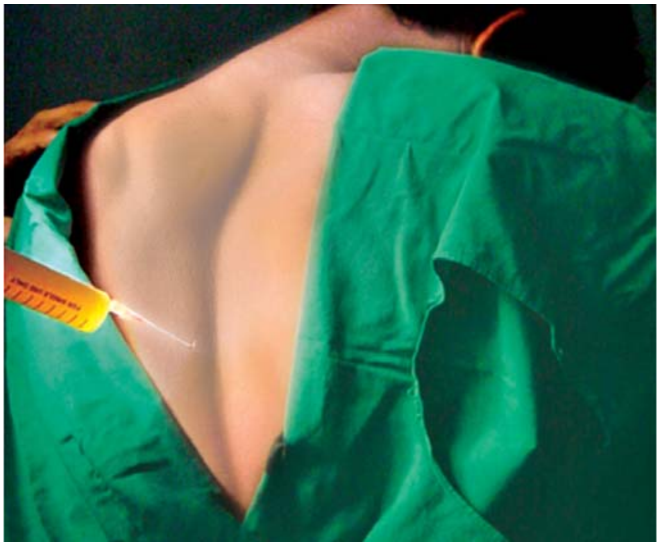

Fig. 28.4: Position for pleural tap is sitting and leaning forward over a support. Site is below the scapula, posteriorly through the seventh intercostal space.

BRONCHOSCOPY

Indications

Diagnostic: To take biopsy in conditions like carcinoma lung, lung abscess, pulmonary tuberculosis.

Therapeutic: To remove foreign body, to suck out the bronchial secretions.

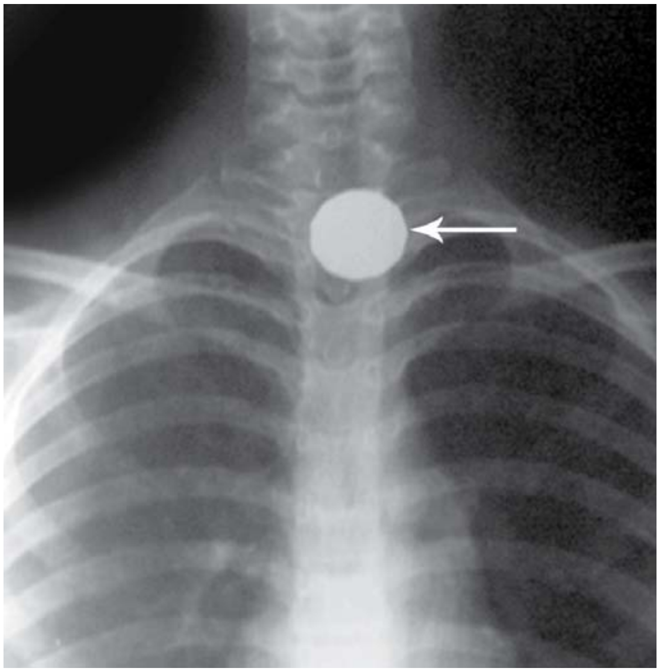

Fig. 28.5: Foreign body in trachea. This can be removed using a bronchoscope. Patient may develop collapse of lung, pneumonia, respiratory failure and ARDS.

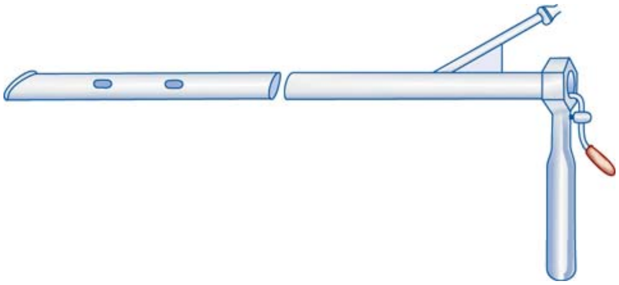

Fig. 28.6: Bronchoscope.

Types

- *Rigid bronchoscopy*—it is used for removal of foreign body and bronchial wash. It reaches up to third generation bronchioles. It is used to take biopsy from carcinoma of proximal divisions but not from carcinoma of peripheral lung. Rigid scope has got multiple holes to allow ventilation during the procedure (Oesophagoscope does not have side holes). It is done under general anaesthesia.
- *Flexible bronchoscopy*—it reaches up to 5th generation bronchioles. It can be done under local anaesthesia. It is mainly used for diagnosis and biopsy.

Complications

- Bleeding
- Infection
- Perforation
- Bronchospasm

EMPYEMA THORACIS

> *The pleurisie is an inflammation of the membrane, investing the ribs, caused by subtile and cholerick blood.... If it tend to suppuration, it commonly infers a pricking pain, a fever and difficulty of breathing.... If nature being too weak...the disease is turned into an empyema, wherefore the Chirurgeon must then be called, who... may make a vent between the third and fourth true and legitimate ribs.... The pus or matter must be evacuated by little and little at several times; and the capacity of the chest cleansed from the purulent matter.*
>
> —**Ambroise Paré,** 1575

It is collection of pus in pleural cavity.

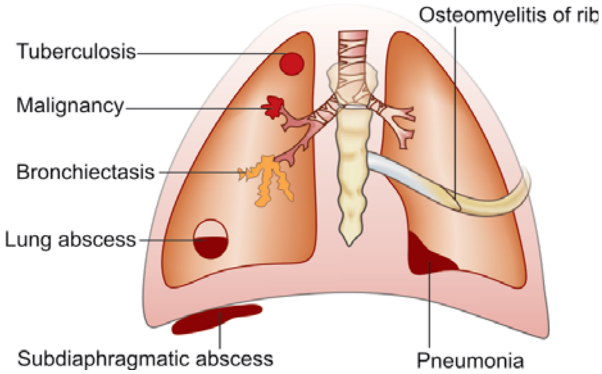

Fig. 28.7: Causes of empyema.

Causes

An empyema is never primary.
Secondary causes are:

- From chest wall: Wounds, osteomyelitis of ribs.
- From lung: Pneumonia, abscess, bronchiectasis, tuberculosis, growth.
- Postoperative: After thoracotomy.
- From oesophagus: Perforations, carcinoma.
- From below diaphragm: Subphrenic abscess.

Tuberculosis and pneumonia are common causes in developing countries.

Pathology

Stages of empyema

- Acute empyema
- Subacute empyema
- Chronic empyema

Initially serous fluid collects, which eventually becomes purulent. Intrapleural clotting of pus occurs with thickening of

For all the happiness mankind can gain, is not in pleasure but in rest from pain.—**Dryden** pleura and later fibrinous adhesion forms resulting in matured empyema.

Chest is withdrawn inwards and is immobile, mediastinum is drawn inwards, diaphragm gets elevated.

It leads to rigid contracted immobile chest with functionless lung underneath—*frozen chest.* Often pus perforates through intercostal space and forms *empyema necessitans*.

Organisms: Initially staphylococci, streptococci, pneumococci, and later *Entamoeba coli*, *Pseudomonas,* drug-resistant staphylococci.

Clinical types

- Acute empyema
- Acute fulminant toxic empyema
- Subacute empyema
- Chronic empyema
- Latent empyema
- Persistent empyema
- Empyema necessitans
- Chronic empyema with sinus
- Interlobar empyema

Pathological types

- Exudative
- Fibrinopurulent
- Organising

Anatomical types (Depending on the anatomical location)

- Apical
- Mediastinal
- Interlobar
- Diaphragmatic
- Lateral

Clinical Features

- Pain in the chest, tenderness, fever.
- Difficulty in breathing.
- Features of toxicity in acute type of empyema.
- Dullness on percussion, absence of breath sounds.
- Decreased chest wall movement.

Complications

- Frozen chest (functionless lung)
- Empyema necessitans
- Osteomyelitis/chondritis of ribs or vertebra
- Pericarditis
- Mediastinitis
- Bronchopleural fistula
- Dissemination of infection

Investigations

- Chest X-ray, ESR.
- Peripheral smear.
- Diagnostic aspiration.
- Pus C/S, AFB.
- Bronchoscopy.
- CT scan and MRI for carcinoma bronchus.

Treatment

Stage 1

- Antibiotics.
- Repeated aspirations.
- Intercostal tube drainage.
- Antituberculous drugs.

Stage 2

- ICT drainage.
- Rib resection *(Eloiser's method)*.
- Antibiotics, antituberculous drugs.
- Respiratory physiotherapy.

Stage 3

- ***Decortication*** *is very* useful and favourable method. Here, thickened pleural sac is stripped off from the lung through thoracotomy approach.
- Often ***lobectomy*** may be required, rarely pneumonectomy is done.
- Physiotherapy, antibiotics, ATD'S are also essential.

EMPYEMA NECESSITANS

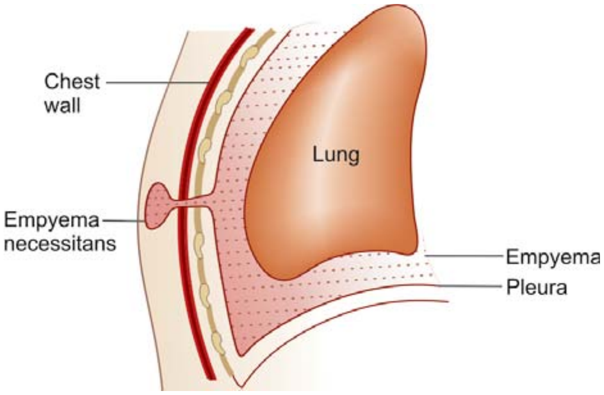

Fig. 28.8: Empyema necessitans.

It is a complication of empyema thoracis, *wherein empyema* which is not drained perforates through the chest wall presenting as subcutaneous collection of pus communicating directly or often with a tortuous route with the pleural cavity.

Causes

- Neglected empyema either of tuberculous or nontuberculous aetiology.
- Following incomplete aspiration of pus of empyema through a thin needle—needle track itself allows pus to form a track and leads to *empyema necessitans*.

Clinical Features

A diffuse, tender, smooth, soft, fluctuant, swelling in the intercostal space having *impulse on coughing.*

Features of empyema necessitans

- Bulge in the intercostal space
- Restricted movement of chest wall
- Tenderness
- Dullness on percussion without breath sounds
- Impulse on coughing

Investigations

- Chest X-ray.
- Total count is increased.
- ESR.

Treatment

- Antibiotics.
- Empyema is drained by placing an ICT. Pus is sent for C/S and AFB.
- If it is tuberculous, antitubercular drugs are started.
- Empyema necessitans subsides on its own when empyema is treated.
- But often may require separate incision and drainage when track is tortuous.

LUNG ABSCESS

- It is localised suppuration in the lung with tissue necrosis.
- It is end-stage of suppurative pneumonitis with thrombosis of associated artery.

Aetiology

- Pneumonias due to *Streptococcus, Pneumococcus, Haemophilus, Staphylococcus,* anaerobic and other bacteria.
- Bronchial obstruction due to tumours or foreign body.
- Chronic upper respiratory infection due to sinusitis, tonsillitis and dental infection (anaerobic infection).
- Septicaemia.
- Aspiration.

As pus accumulates, tension increases inside the abscess cavity causing spread into other areas of the lung or may rupture into the bronchus.

Note:
Lung abscess secondary to aspiration is commonly seen in apical lower lobe.

Clinical features

- Acute onset of fever, but often it is recurrent in nature
- Cough with expectoration
- Haemoptysis with foul smelling sputum.
- Chronic illness with debilitation
- Pleuritic pain

Complications

- Spread into other areas of the lung.
- *Metastatic cerebral abscess*—Lung abscess is the *commonest* focus for metastatic cerebral abscess. Occurs as a result of pyaemic emboli through paravertebral veins.
- Haemorrage, may be torrential due to erosion of the vessel in the abscess wall.
- Empyema thoracis.

Differential Diagnosis

- Pulmonary tuberculosis.
- Carcinoma lung.
- Fungal infections like aspergillosis.
- Pneumoconiosis.
- Lung cysts.

Investigations

- Chest X-ray shows localised opacity with smooth margin and fluid level.
- Bronchoscopy and biopsy is done to rule out carcinoma.
- Sputum for culture, AFB and cytology.
- CT scan.

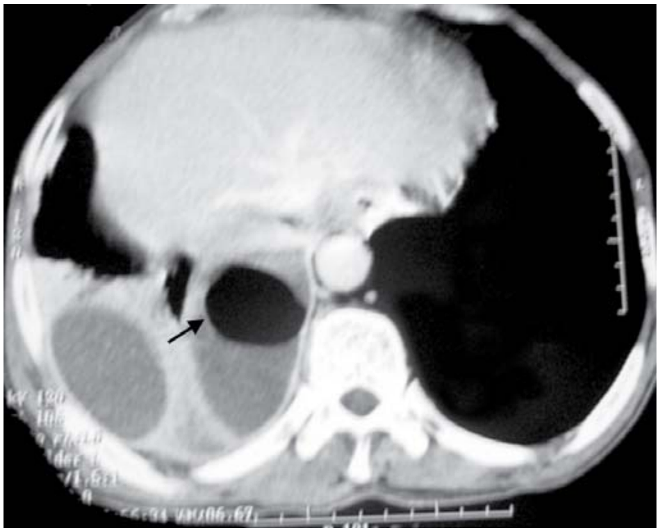

Fig. 28.9: CT scan of lung showing right-sided lung abscess with fluid level.

Treatment

Medical

1. Appropriate *high dose antibiotics* are started depending on sputum culture like penicillins, third generation cephalosporins like cefoperazone, ceftriaxone *for 3-6 weeks.*
2. *Postural drainage* for 2-3 hours, 3 times daily.
3. *Periodic culture* of sputum is done to study the response of treatment.
4. 80-90% of acute abscesses resolve by medical therapy.

Surgical

1. External surgical drainage with rib resection is done in severe acute lung abscess only. Not commonly used procedure.
2. *Lobectomy* is done when medical treatment fails.

> *Clubbing is frequently present in those with a chronic discharging empyema sinus.*
> **—Hippocrates, Father of Medicine**

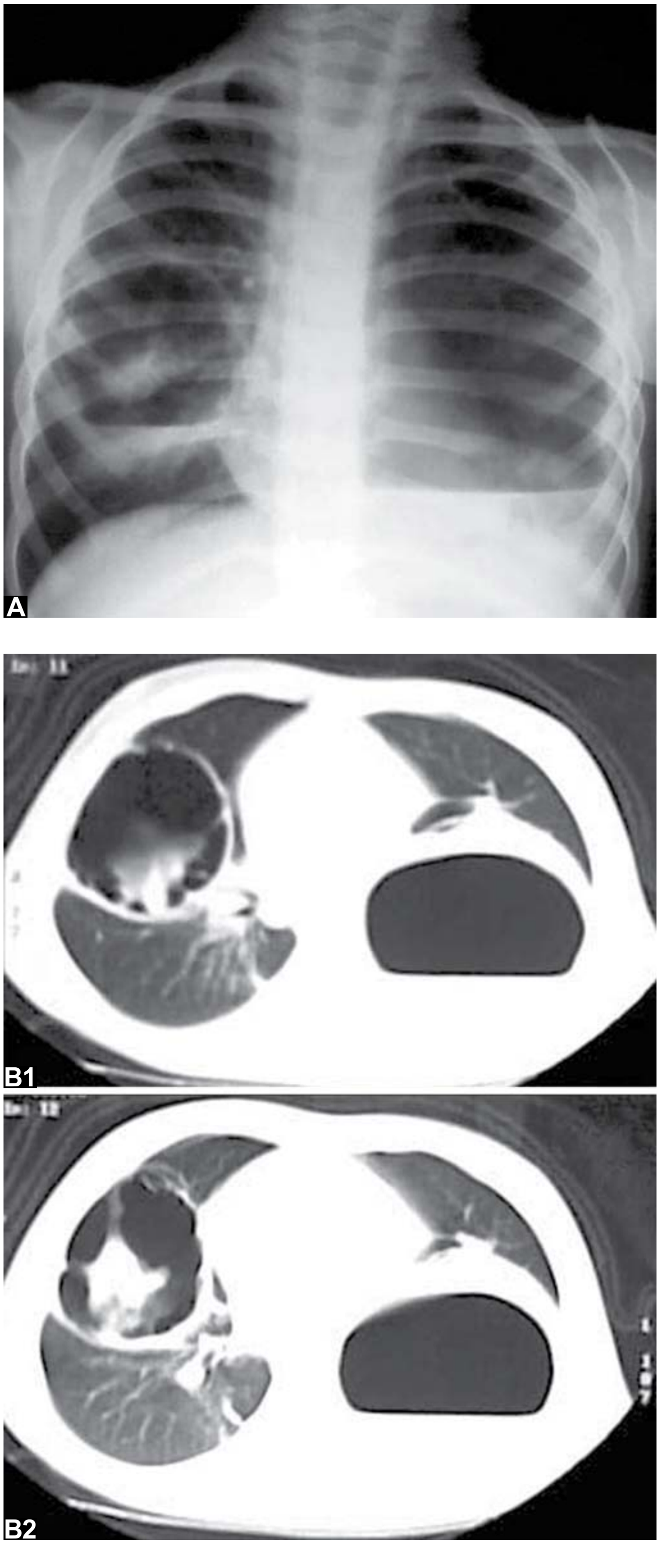

Figs 28.10A and B: Lung abscess X-ray picture and CT of the same patient.

INTERCOSTAL TUBE DRAINAGE

It is the method of draining fluid collected in the pleural cavity safely, so as to allow the underlying lung to expand.

Indications

- Haemothorax
- Pneumothorax
- Haemopneumothorax
- Empyema thoracis
- Traumatic lung contusion
- After thoracotomy, to drain pleural cavity

Procedure

- **Sites:** An ICT is placed in 6th intercostal space in midaxillary line for haemothorax and pyothorax; in 3rd or 4th space in midaxillary line for pneumothorax.

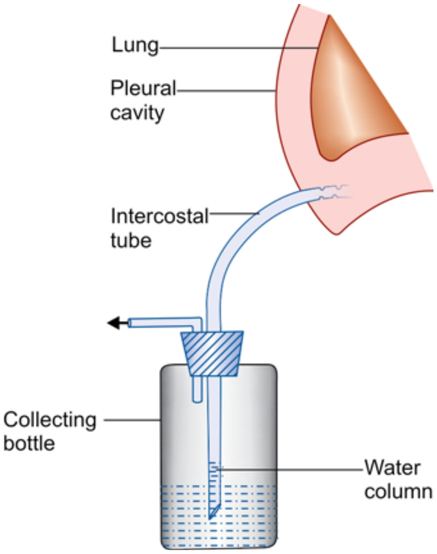
Lung
Pleural cavity
Intercostal tube
Collecting bottle
Water column

Fig. 28.11: Intercostal tube drainage underwater seal.

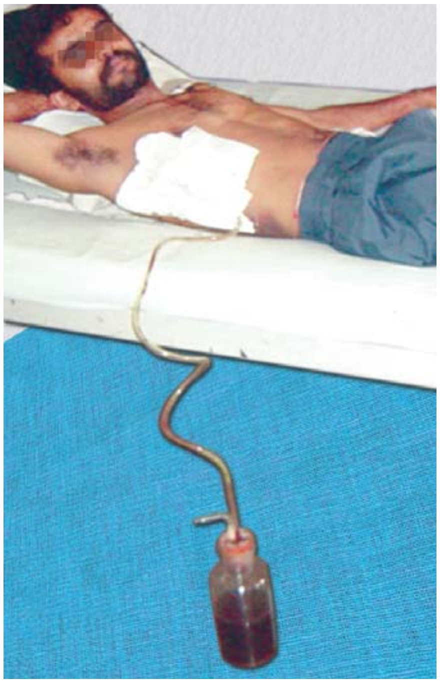

Fig. 28.12: Intercostal tube drainage showing haemothorax.

- Under local anaesthesia, a small incision is made in midaxillary line (as the muscle bulk is less here and so passage of ICT is easier), parallel to intercostal space (above the rib, i.e. in lower part of intercostal space so to avoid injury to neurovascular bundle which are located in the groove in the lower part of rib). Tube with side openings is pushed into the pleural cavity.
- Other end is connected *to under water seal. Air-water column moving* with respiration can be observed. Tube is fixed with skin sutures.
- Usually for pneumothorax, ICT is kept for 2-3 days (till lung expands—confirmed by—repeat chest X-ray). For haemothorax and pyothorax it is kept for 4-6 days or until it stops draining and lung expansion is confirmed by repeat chest X-ray.
- To have proper expansion of lungs, patient is asked to blow football bladder (balloon) and do breathing exercise. If there is bronchopleural fistula, ICT should be placed for a longer time, until the fistula heals.

Complications

- Infection.
- Displacement and inadequate functioning.
- Injury to intercostal vessels and bleeding.
- Pain at the site of ICT placement.

SHOCK LUNG (Stiff lung)

Causes

- Major chest trauma.
- Septicaemia.
- Massive blood transfusions.
- DIC.

Pathogenesis

Development of microthromboembolism in small lung vessels following extensive intravascular coagulation, leading to pulmonary consolidation, which reduces the lung compliance markedly, causing severe depression of gas exchange in the lung—***a stiff lung***.

It has got high mortality as lung cannot expand at all. Outcome is fatal if emergency treatment is not given.

Treatment

- Endotracheal intubation.
- Ventilator support with IPPV.
- Antibiotics.
- High dose of steroids.
- Monitoring the patient with PCWP and cardiac monitor.
- Bronchodilators.
- The cause is treated.

PULMONARY EMBOLISM (PE)

> *From my Stanford days… I had known that excessive bed rest gave rise to thromboembolic complications….The death rate from thromboembolism was always much less at the County Hospital than it was at Stanford Hospital…. When (the County Hospital patients) got up to go to the bathroom, (they) dislodged only tiny clots from their veins and these did not harm them when they got to the lungs and were dissolved, while the wealthier patients (at Stanford) who remained in bed and formed large clots in their legs and pelvises suffered the major consequences of large pulmonary emboli.*
>
> —**William Dock,** 1984

- It is due to deep venous thrombosis (DVT) which gets detached to cause pulmonary embolism.
- It may be from femoropopliteal or ileo-femoral region.

The thrombi most commonly develop in the leg veins due to stasis and hypercoagulable state. They subsequently enlarge and propagate proximally, dislodge to form emboli.

Types

- **Small emboli**: Causes pulmonary hypertension of features of bronchopneumonia.
- **Medium emboli** lodges in branches of pulmonary artery causing chest pain, haemoptysis, dyspnoea.
- **Large (massive) emboli** causes block at bifurcation of pulmonary artery trunk or right/left pulmonary artery leading to sudden chest pain, severe dyspnoea, shock, raised venous pressure and ***sudden death.***

Effects of pulmonary embolism

- Decreased cardiac output
- Pulmonary vasospasm and pulmonary hypertention
- Bronchospasm
- Defective oxygenation of blood

Risk Factors

- Postoperative and trauma patients who are bed ridden.
- Pregnant women.
- Old age.
- Obesity and heart disease.
- Varicose veins.
- All aetiologies which cause DVT.
- Carcinoma.

Clinical features

- Dyspnoea, chest pain and haemoptysis
- Tachycardia, tachypnoea and cyanosis
- Pleural rub and cardiac gallop

The patient dyspnoeic due to pulmonary embolism prefers to lie flat.—**Ronald Gibson**

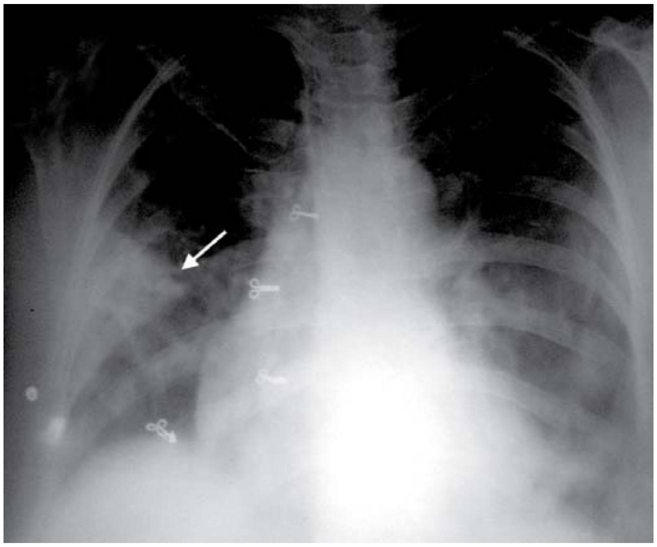

Fig. 28.13: Chest X-ray showing wedge shaped pulmonary infarct.

Investigations

- Chest X-ray PA view—hyperlucency in an area of oligaemia—*Westermark sign.*
- CT scan and MRI—can detect PE.
- Pulmonary angiography: It is diagnostic (100%).
- Arterial blood gas analysis.
- Isotope radionuclide ventilation—perfusion (V/Q) lung scanning: While normal scan rules out PE, evidence of V/Q mismatch is highly suggestive of pulmonary embolism.
- Doppler study and venogaphy—To rule out DVT.

Management

- Thrombolytics: Streptokinase, 6 lakhs units to begin with and later one lakh units hourly.
- Pulmonary embolectomy with or without bypass by immediate thoracotomy.
- Venous thrombectomy using Fogarty catheter.
- Ventilator support.
- To prevent further embolisation of embolus into the lungs, IVC filters can be placed.
- Treatment of DVT.

Note:
Heparin should not be combined with thrombolytics.

SURGICAL EMPHYSEMA

It is collection of gas/air in the subcutaneous or and fascial planes.

Causes

- Lung injury.
- Tracheal injury.
- Chest wall injury.
- After laparoscopic procedure.

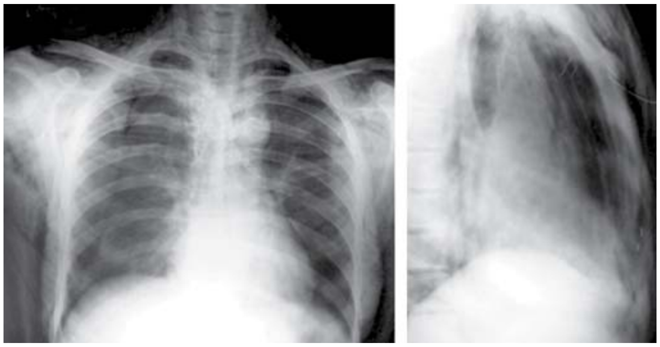

Fig. 28.14: Chest X-ray PA and lateral views showing surgical emphysema.

Types

- Localised—observed at the site of trauma.
- Generalised—extensive in the neck, face, eyelids, chest wall, etc.—in the injury of bronchus, oesophagus and mediastitum.

Features

- Pain, diffuse swelling of the subcutaneous region.
- Palpable *crepitus* is diagnostic.
- X-ray chest and neck is confirmatory.

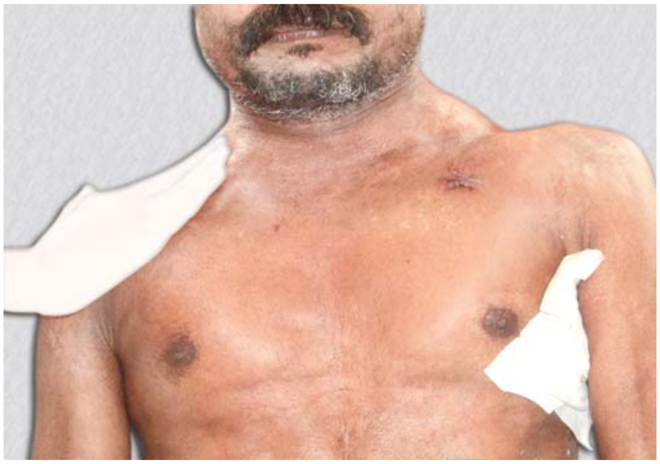

Fig. 28.15: Surgical emphysema. Note the demonstration of crepitus.

Treatment

- Conservative treatment—treat the cause.
- In severe cases—ICT one side or both sides placement.

LUNG CYSTS

Types

- Epithelial cyst
- Emphysematous cyst
- Parasitic cyst (Hydatid cyst)
- Pseudocyst of the lung

Pseudocyst of the lung: It occurs in a cavity due to tuberculosis, lung abscess or *Staphylococcus* pneumonia.

Epithelial Cyst

- It is of d*evelopmental* origin.
- It can be large single cyst or small, multiple cysts.
- It is lined by *respiratory epithelium.*
- It is often associated with cervical rib or cardiac anomalies.
- It is common in *infants and children.*
  - Usual *clinical features* are dyspnoea and chest pain. When infected, presents with fever, cough and haemoptysis.
  - *Treatment* is excision of the cyst. Antibiotics are given when infected.
  - *Here spontaneous pneumothorax is uncommon. Infection and haemorrhage are common.*

Emphysematous Cyst

- *It is a progressive disease of lung with rupture of alveolar wall and distension with air.*
- It is an *acquired* condition.
- Cyst *does not have epithelial lining.*
- Compression of adjacent lung tissue with poor gaseous exchange is common.

Usual *clinical features* are dyspnoea and severe, persistent cough.

Complications

- Spontaneous pneumothorax.
- Severe chronic bronchitis.

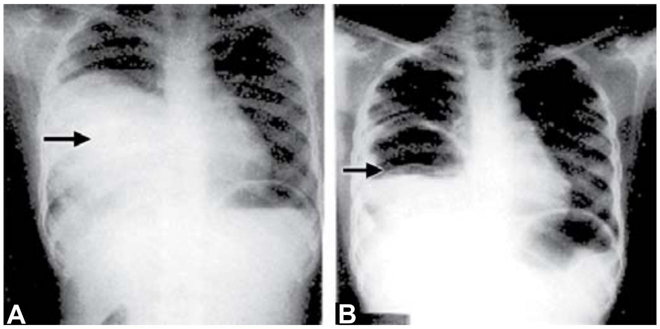

Figs 28.16A and B: (A) Hydatid cyst in lung (B) Ruptured hydatid cyst with water-lily sign.

Treatment

- Localised cyst is excised.
- Generalised cysts are treated conservatively.
- Pleurodesis is often required.

Here spontaneous pneumothorax is very common. Infection and haemorrhage are uncommon.

Hydatid Cyst of the Lung

- It is caused by the parasite, *Echinococcus granulosus.*
- It is often associated with hydatid cysts of the liver.
- Incidence is 15%.
- It is usually single. But multiple hydatids can occur.

Presentations

- Dyspnoea, chest pain and haemoptysis.
- Rupture into bronchial tree causes anaphylaxis.
- Rupture also causes expectoration of fluid and grape skins (vesicles).
- Fever, cough and expectoration due to secondary infection.
- Asymptomatic hydatid, identified during chest X-ray.

Investigations

- Chest X-ray:
  1. Dense homogenous opacity.
  2. Collapsed laminated membrane produces an irregular projections in a fluid level due to rupture of the cyst—*Water-lily* appearance.
  3. *Crescentic cap of air,* when it communicates into bronchial tree.
- Positive Casoni's test.
- Blood shows eosinophilia.
- Complement fixation test.
- Indirect haemagglutination test.
- CT scan—diagnostic.

Complications

- Rupture of the cyst
- Anaphylaxis
- Secondary infection
- Collapse of lung
- Pleural effusion
- Secondary pleural hydatid formation
- Hepatobronchial fistula formation (produces bilious sputum)

Differential diagnosis

- Lung abscess
- Epithelial lung cysts
- Aspergillosis

Treatment

- Thoracotomy and *enucleation of the cyst:* It is achieved by taking a good anaesthetist's help by creating positive pressure ventilation. Cyst extrudes intact. It should not be held with forceps to avoid rupture.
- *Lobectomy* is done in cases with difficulty and complications.
- Drugs like *praziquantel* or *albendazole* is given for a long period.

MEDIASTINAL TUMOURS

- It occurs at any age group, in both sexes.
- It is often identified on a routine chest X-ray, as about 50% are symptomless.

Endurance is not just the ability to bear hard things but to turn it into glory.

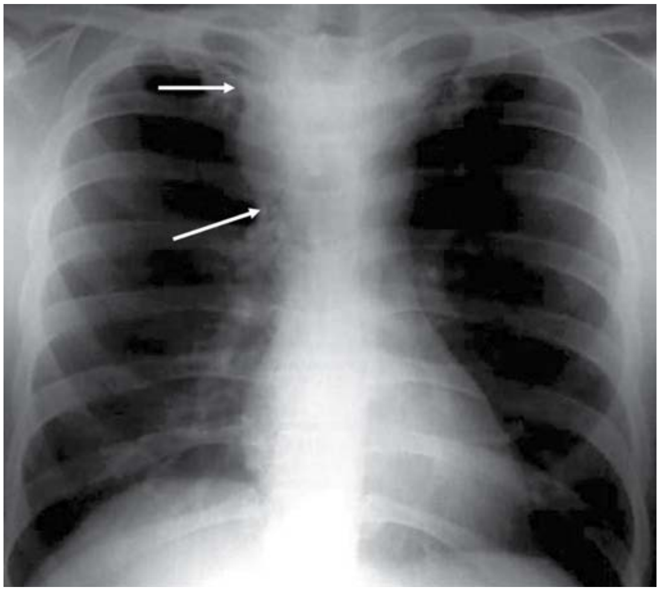

Fig. 28.17: Superior mediastinal tumour—could be thymoma, nodal mass.

Presentations

- Chest pain and back pain.
- Respiratory distress.
- Venous congestion (SVC syndrome).
- Hoarseness of voice (due to compression over recurrent laryngeal nerve).
- Dysphagia, due to oesophageal compression.
- Horner's syndrome, due to compression over sympathetic chain.
- Scabbard trachea.
- Later diaphragmatic paralysis may occur.
- Pleural effusion.
- Haemorrhage, due to erosion of major vessels by malignant tumour.

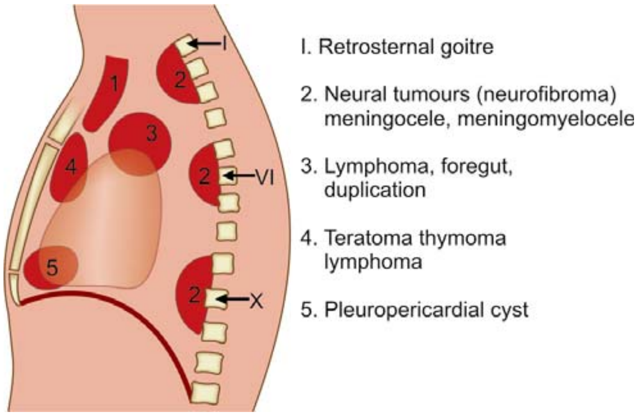
I
1
2
3
4
2
VI
5
2
X
I. Retrosternal goitre
2. Neural tumours (neurofibroma) meningocele, meningomyelocele
3. Lymphoma, foregut, duplication
4. Teratoma thymoma lymphoma
5. Pleuropericardial cyst

Fig. 28.18: Sites of mediastinal tumours.

Classification

Superior mediastinal tumour: Retrosternal goitre.

Anterior mediastinal tumours

- Retrosternal goitre, aortic arch aneurysm lymph node enlargement.
- Thymic tumours—thymomas are commonly associated with myasthenia gravis.
- Teratomas and dermoids.
- Pleuro-pericardial cyst (spring *water cyst*).

Midmediastinal tumours

- Lymphadenopathies of all causes—secondaries, lymphomas, tuberculosis—common.
- Foregut duplication cysts.
- Lipomas.

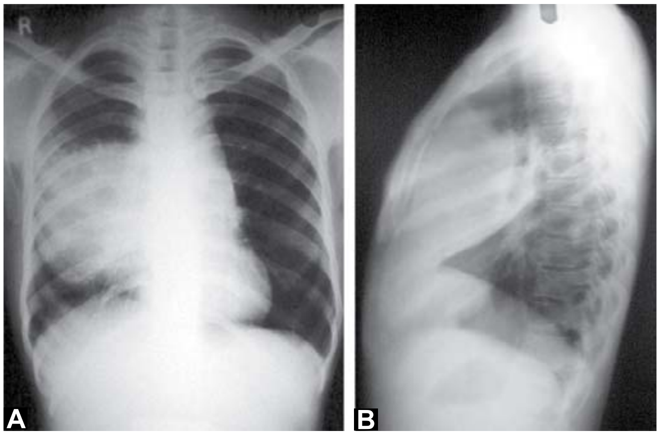

Figs 28.19A and B: Chest X-ray showing mediastinal tumour.

Posterior mediastinal tumours

- Neurofibromas—commonest, 75%.
- Ganglioneuromas.
- Aneurysm of descending aorta.
- Lymph node mass.

Investigations

- Chest X-ray, both PA view and lateral view.
- CT scan.
- MRI.
- Mediastinoscopy and biopsy.

Treatment

- Thoracotomy and removal of tumour.
- If malignant, adjuvant therapy like radiotherapy and chemotherapy are given.

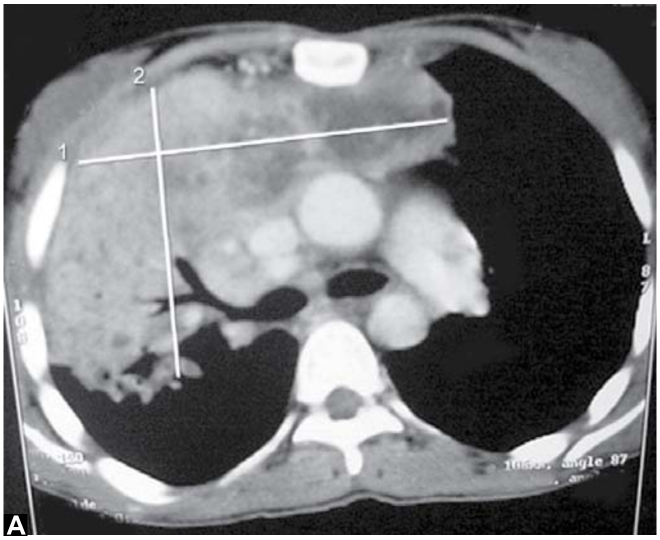

Fig. 28.20A

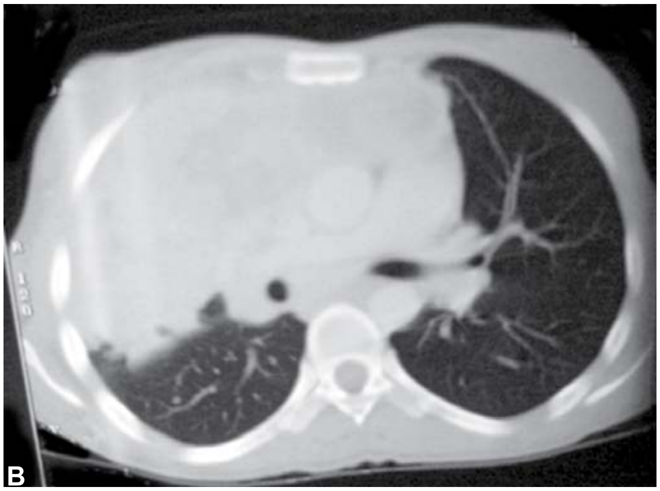

Fig. 28.20B

Figs 28.20A and B: CT scan showing mediastinal tumour.

- Sternotomy for superior and anterior mediastinal tumours.
- Thoracoscopic and mediastinoscopic approaches.

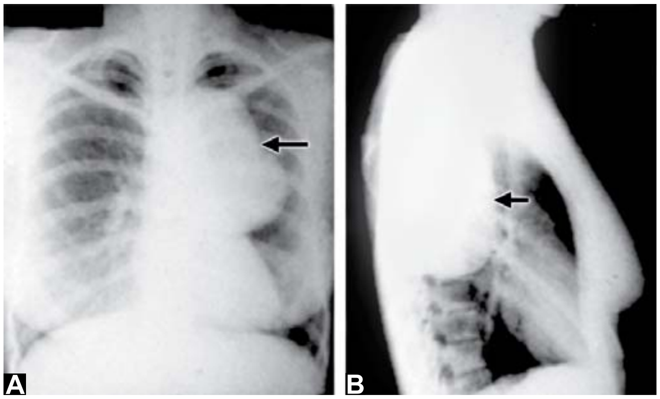

Figs 28.21A and B: Plain X-ray chest showing posterior mediastinal mass.

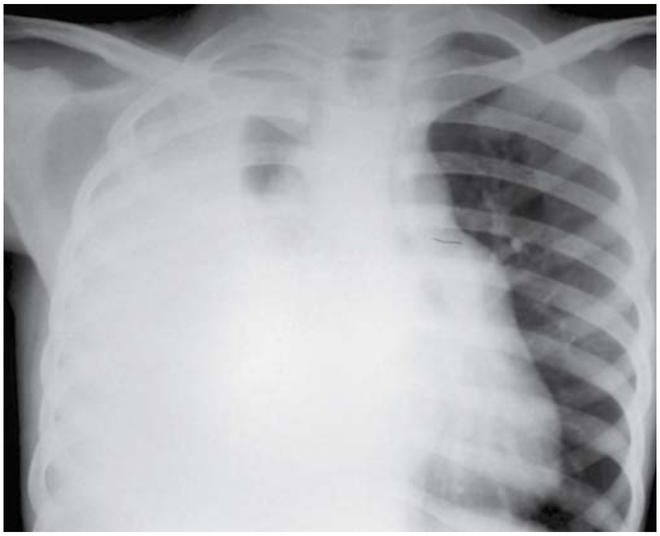

Fig. 28.22: Malignant effusion—right-sided.

THYMOMAS

- Thymomas are the most common tumours of the anterosuperior mediastinum in the adult.
- They are most common in the fifth and sixth decades of life. Both sexes are equally affected.

Histological types

- Epithelial cell—has poor prognosis
- Lymphocytic
- Mixed
- Spindle—has better prognosis

50% of thymomas are malignant

Staging

I — Confined to the capsule
II — Tumour spread to periglandular fat
III — Tumour spread to adjacent tissue

Clinical Features

- Asymptomatic—50%.
- 30-40% of thymomas have associated myasthenia gravis.
- Chest pain.
- Dysphagia and dyspnoea.
- Superior vena caval obstruction.

Investigations

- *Tensilan diagnostic test:* By injecting 10 mg IV to symptomatic myasthenia will be relieved in one minute temporarily.
- Chest X-ray, lateral view shows opacity in the mediastinum.
- CT scan.

Treatment

- *For myasthenia*—tab. neostigmine bromide 15 mg tds, daily.
- *Thymectomy* is very useful.
  - When the disease is less than 5 years.
  - Myasthenia gravis without thymoma.
  - In young, females.

Features of myasthenia gravis

- An autoimmune disease
- Presence of antibodies to acetylcholine receptors in NMJ resulting in reduced muscular contraction
- Weakness of muscle on exertion
- Ptosis, diplopia, dysarthria, dysphagia
- Drooping of jaw
- Paralysis of respiratory muscle
- Periodic remission

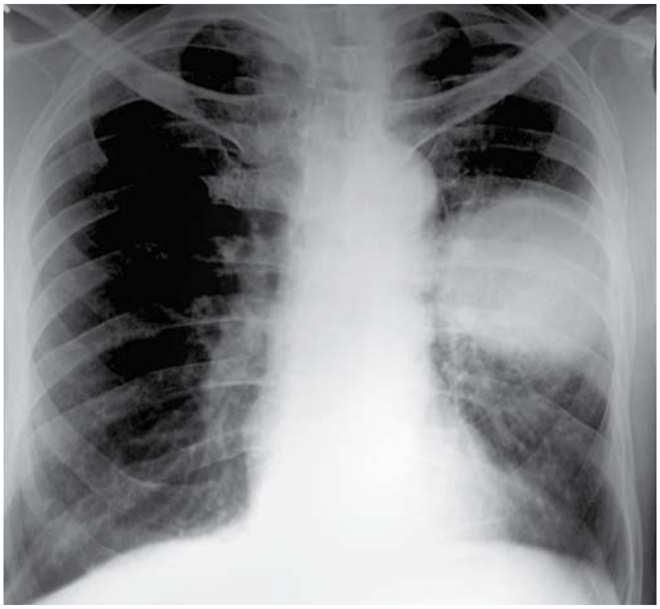

Fig. 28.23: Bronchogenic carcinoma—X-ray picture.

PANCOAST TUMOURS (Superior sulcus tumour)

> *When the tumour extends its feet from all sides of its body into the veins, the sickness produces the picture of a crab.*
>
> —**Galen,** (130-200)

- It is a type of peripheral lung carcinoma arising from the apex of the lung (5% of lung cancers).

Features

- It invades brachial plexus, sympathetic chain, upper ribs and vertebrae.

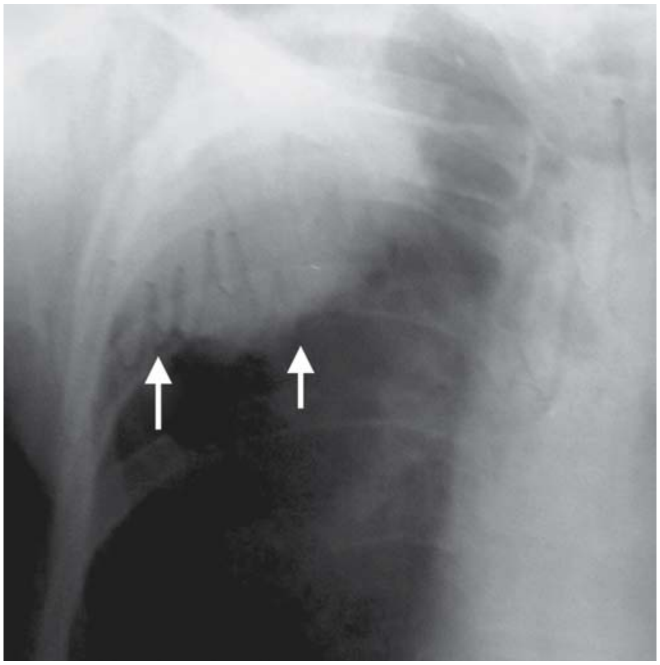

Fig. 28.24: X-ray showing right sided Pancoast's tumour.

Presentations

Pancoast's syndrome

- Lower brachial plexus palsy
- Horner's syndrome
- Rib erosion
- Apical shadow

Horner's syndrome

- Miosis
- Enophthalmos
- Anhydrosis
- Ptosis
- Loss of spino-ciliary reflex

- Intractable pain in upper chest, arm and also weakness in the arm.

Investigations

- Chest X-ray.
- CT scan.
- Bronchoscopy and biopsy.
- CT guided biopsy.
- MRI is better than CT in Pancoast tumour as brachial plexus and sympathetic chain is involved which better visualised in MRI.

Treatment

- Lobectomy/pneumonectomy.
- Radiotherapy—external tele cobalt.
- Chemotherapy—methotrexate.

Prognosis: Poor.

CHEST WALL TUMOURS

- Tumours arising from the chest wall components like muscles or ribs. They can be benign or malignant.
- Commonest benign tumour is chondroma arising from ribs.

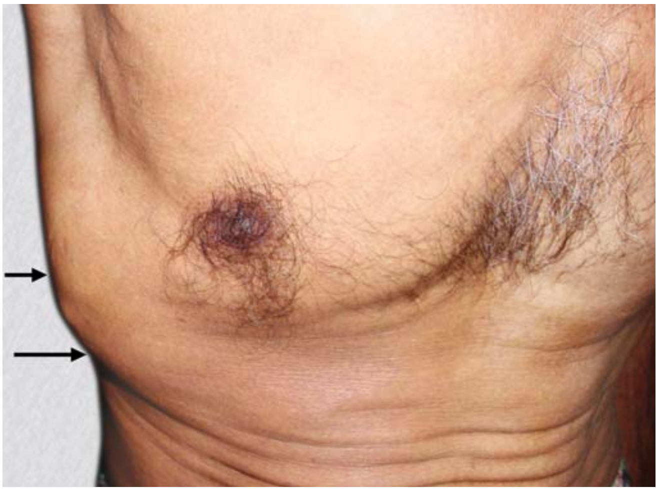

Fig. 28.25: Chest wall tumour. Benign tumours are usually chondromas.

- Malignant tumours are *secondaries (commonest)*, chondrosarcoma arising from ribs (common among primary malignant tumours), rhabdomyosarcoma from muscles, fibrosarcoma from ribs/muscles/other soft tissues, Ewing's sarcoma and invasion from other tumours like from pleura or lungs or breast.

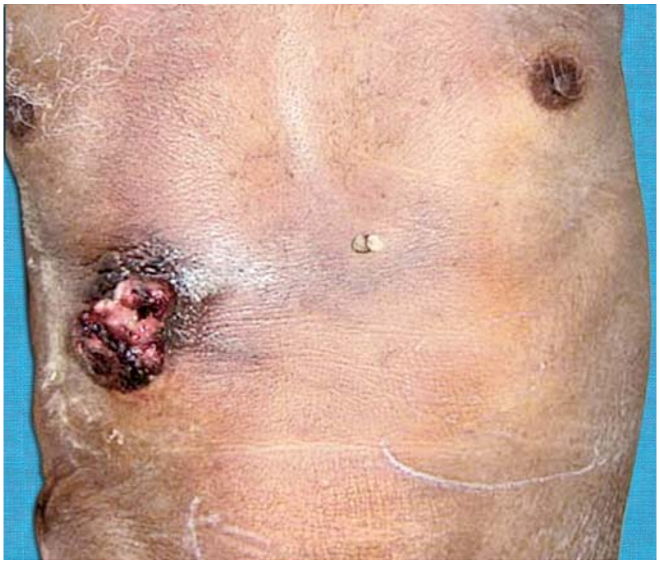

Fig. 28.26: Fungating tumour invading chest wall.

Benign Tumours

- *They are slow growing, nonmobile, painless and usually from the rib cartilage, near costochondral junction.*
- X-ray is diagnostic, shows rib expansion with intact cortex.
- One or more ribs can be involved. Treatment is rib resection.

Primary Malignant Tumour

- *It has got all features of sarcoma—progressive rapid enlargement, attaining large size, warm, vascular, nonmobile, often extends into the thoracic cavity or with skin ulceration.*
- *Secondaries in lung/brain/liver can occur.*
- Chest X-ray, CT chest, US abdomen should be done to see secondaries.
- CT scan can also give idea about the tumour extension and operability.
- Open incision/trucut biopsy is essential for histological confirmation.
- Treatment *is wide excision with chest wall reconstruction using different osteomyocutaneous flaps, rib grafts, mesh or acrylic plates.*
- Postoperative adjuvant chemotherapy is always needed to prevent relapse. In advanced cases radiotherapy to chest wall is advised.

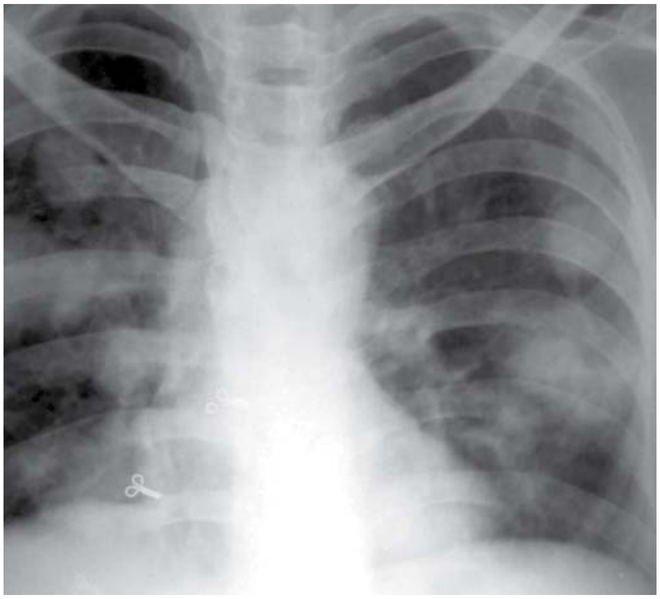

Fig. 28.27: Multiple cannon ball secondaries in lung. Primary sites are—breast, kidney, sarcomas, testis, prostate, thyroid.

PERICARDITIS

> *Water contained in the sac of the heart, likewise if the sac contain within either a fluid putrid and smelling badly, or stones, it causes throbbing….Adhesions of the pericardium to the body of the heart itself…was observed in two dropsical patients.*
>
> —**Guillaume de Baillou (Ballonius),** 1735

Types

1. Acute pericarditis.
2. Chronic pericarditis.
3. Chronic constrictive pericarditis.

ACUTE PERICARDITIS

Usually by bacteria like *Staph. aureus, H. influenzae, Streptococci, Neisseria.* It is uncommon at present, because of availability of good antibiotics.

Other causes

- Viral infection
- Uraemia
- Trauma
- Malignancy
- Connective tissue disorders

- Here pus collects in the pericardial space causing decreased cardiac function and toxicity.

Treatment

- Antibiotics.
- Pericardial aspiration.
- Drainage of purulent fluid by open pericardiotomy.

Don't let the skin stand between you and the diagnosis.

CHRONIC CONSTRICTIVE PERICARDITIS (Pick's Disease)

- Here pericardium is thickened, fibrosed and calcified. Heart is encased in a rigid cavity which decreases the cardiac function as well as the venous return.

Causes

- *Tuberculous pericarditis* is the commonest cause.
- Trauma.
- Viral pericarditis.
- After cardiac surgery.

Clinical Features

- Decreased cardiac output and tachycardia.
- Dyspnoea on exertion, easy fatigability.
- Raised jugular veins, hepatomegaly, ascites and oedema feet.

Investigations

- ECG.
- Echocardiography.
- Chest X-ray.
- CT scan.

Treatment

- Pericardiotomy with biopsy.
- Pericardiectomy.
- Antitubercular drugs is started when tuberculosis is the cause.

PERICARDIAL TAP

Indications

- Pericardial effusion due to any cause—viral, tubercular
- Haemopericardium
- Purulent pericardium

Procedure

A 16 or 18-gauge needle is passed into the pericardium just below the Xiphoid process, directing upwards and backwards, towards left side with an angle of 45° to the surface.

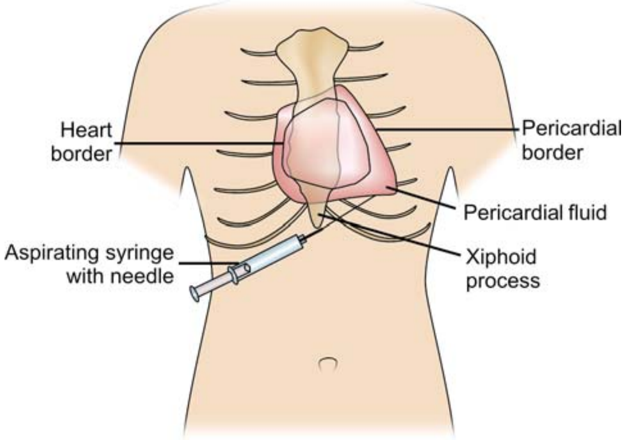

Fig. 28.28: Pericardial tap for pericardial effusion/tamponade.

This site is used because:

- Most dependent aspiration
- Unlikely to traumatise heart
- Pleura is not punctured
- Coronary vessels are not interfered

Presently U/S guided aspiration is commonly done.

Complications

- Injury to heart.
- Infection.

CARDIAC TAMPONADE

Rapid accumulation of fluid or blood in the pericardial space causing increase in the ***intrapericardial pressure*** is called as cardiac tamponade.

This results in compression of cardiac chambers.

↓

↓ venous return

↓

↓ cardiac output

Causes

- Trauma
- Progressive pericardial effusion due to tuberculosis, viral, bacterial infections
- Often, uraemia can cause significant pericardial effusion

Clinical Features

- Hypotension.
- Widened cardiac dullness.
- Muffled or decreased heart sounds.
- Increased venous pressure with raised jugular veins.
- Pulsus paradoxus (pulse becomes weaker on inspiration than expiration).
- In severe cases, heart is unable to expand causing *shock and often sudden death*.

Beck's triad

- Hypotension
- Muffled heart sounds
- Raised jugular venous pressure

Investigation

Chest X-ray and U/S confirms the diagnosis.

Treatment

- ***Pericardial tap,*** as early as possible to allow heart to expand adequately.
- Occasionally, ***open pericardiotomy*** is required.

DIAPHRAGMATIC HERNIA

It is herniation of abdominal content through diaphragm into the chest.

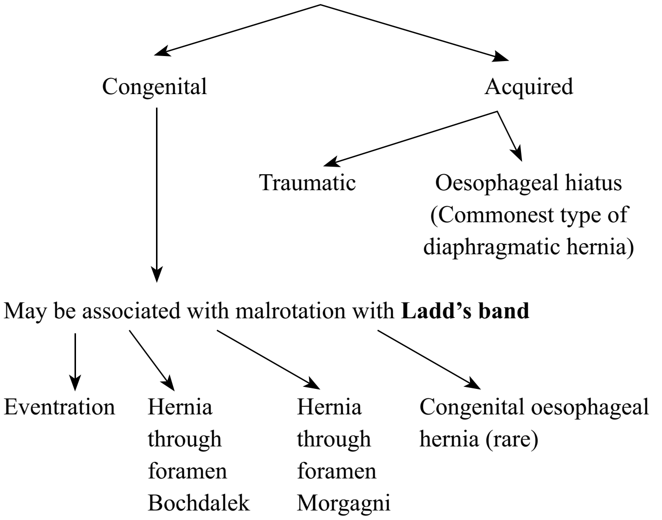
Congenital
Acquired
Traumatic
Oesophageal hiatus (Commonest type of diaphragmatic hernia)
May be associated with malrotation with **Ladd's band**
Eventration
Hernia through foramen Bochdalek
Hernia through foramen Morgagni
Congenital oesophageal hernia (rare)

Etiology of congenital diaphragmatic hernia

- Genetic
- Decreased lung fluid pressure
- Failure of the closure of pleuroperitoneal canal (foramen of Bochdalek)
- Early return of bowel into peritoneal cavity from physiological hernia

Eventration

- It is weakening of diaphragm due to atrophy and loss of muscle of a part or all of one leaf of the diaphragm, with thin fibrous tissue formation, covered with pleura and peritoneum on either side.

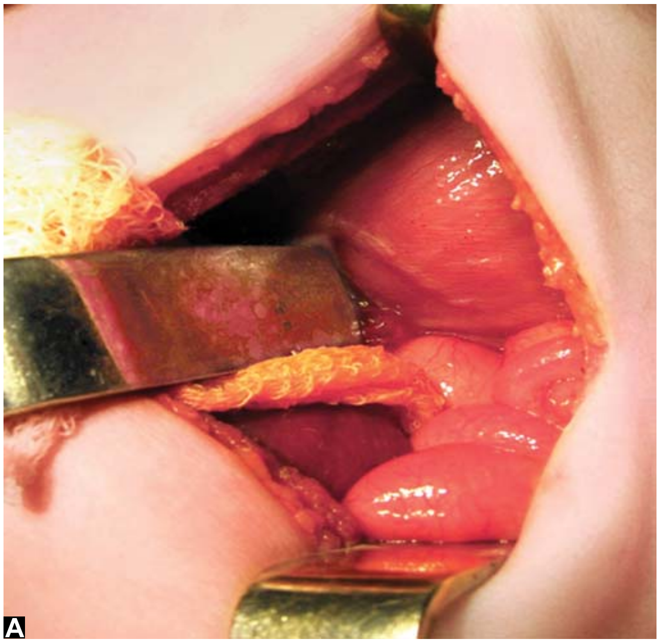

Fig. 28.29A

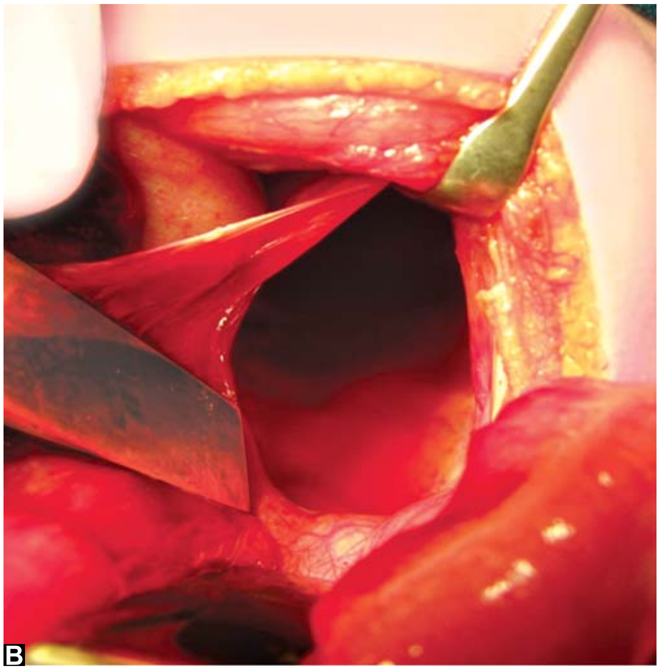

Fig. 28.29B

Figs 28.29A and B: Eventration of diaphragm on table look (*Courtesy:* Dr Sadashiva Rao, MCh, Paediatric Surgeon, KMC, Mangalore).

- Diaphragm is *attenuated and inactive*.
- *Classification of eventration*:
  1. *Congenital*—marked decrease in the muscle fibres in the diaphragm, clinically it may mimic CDH with sac.
  2. *Acquired* or secondary—(a) Phrenic nerve palsy due to trauma (with Erb's palsy); viral—usually in adults (Polio, Herpes zoster, Influenza, Diphtheria). (b) Neoplasia. (c) Autoimmune neuropathy involving diaphragm or phrenic nerve. (d) Iatrogenic.
- Most of them present in the infancy and childhood. A few of them escape to adulthood.
- This thin diaphragm is raised higher and immobile. It is actually not a true herniation. But features mimic hernia.
- *Presentation*: Asymptomatic, wheezing, recurrent LRTI, exercise intolerance, extreme respiratory distress.
- Symptomatic patients have functional deficit in ventilation perfusion (V/P) ratio because lung growth is affected.
- Larger the defect more the hypoplasia. More the hypoplasia lesser the perfusion and severe the symptoms.
- Paradoxical movement compromises the gas exchange.
- *Sniff test*—fluoroscopic evaluation for paradoxical movement.
- Chest X-ray (CT scan/MRI) shows the abnormality
- Differential diagnosis are diaphragmatic hernia through foramen Bochdalek.
- Condition causes respiratory embarrassment.

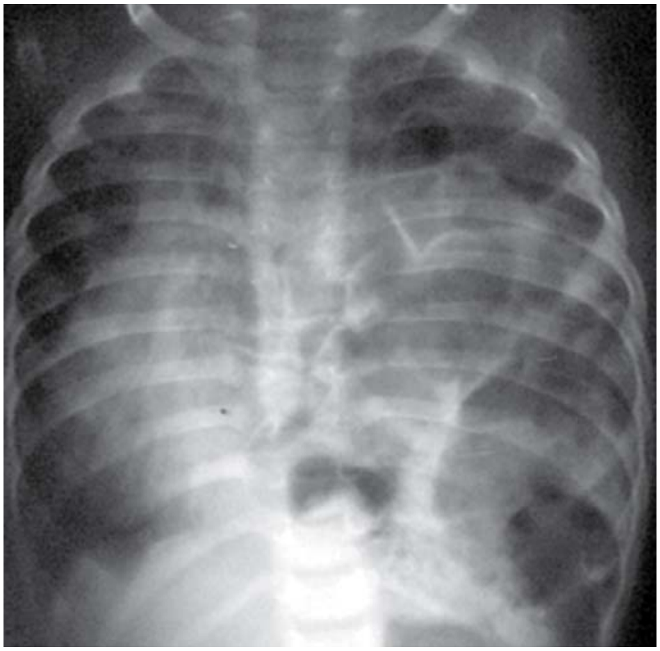

Fig. 28.30: Chest X-ray showing left sided diaphragmatic hernia. Note the bowel shadows in the left side thorax.

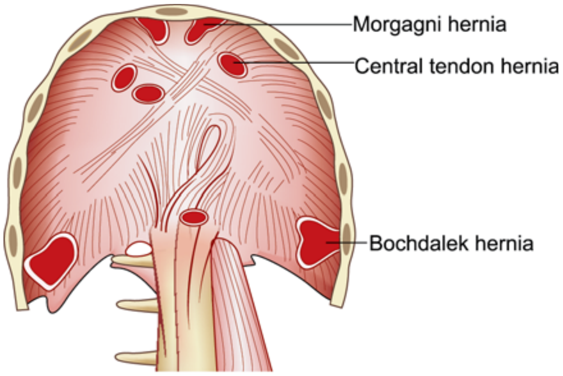

Fig. 28.31: Anatomy of diaphragm showing different sites where hernia can occur.

Types

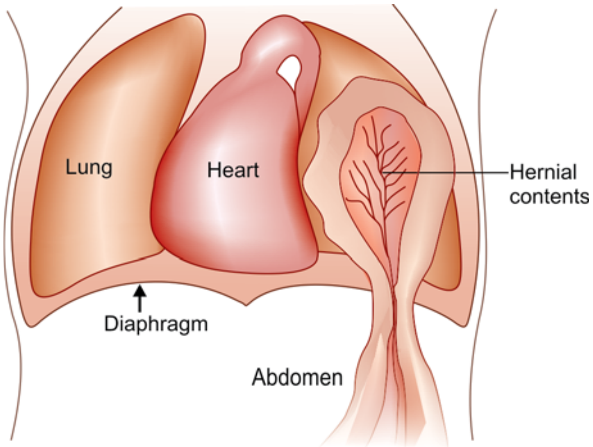

Fig. 28.32: Diaphragmatic hernia.

Treatment is plication of diaphragm through laparotomy.

Indications for surgery

- Symptomatic patients
- Decreased ventilation perfusion ratio
- Prolonged (> 3 to 6 months) exertional dyspnoea
- Impairing daily activity in idiopathic / viral
- Phrenic nerve palsy

Hernia Through Foramen Bochdalek (Through Pleuroperitoneal Canal) (95% Left Sided, Only 5% Right Sided)

- It is commonest congenital diaphragmatic hernia.
- This is a developmental defective condition, due to failure of fusion of pleuroperitoneal canal leaving a direct communication between pleura and peritoneum on left side. This allows herniation of contents of abdomen into the left side thorax.
- Common content is *colon*. Occasionally small bowel, stomach are the contents.
- 80% cases do not have hernial sac. Only 20% cases has got sac.

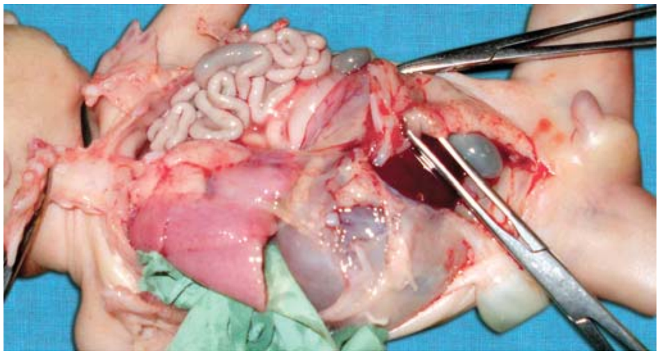

Fig. 28.33: Postmortem look of a newborn baby with classical diaphragmatic hernia (*Courtesy*: Dr Subramanya Bhat K, Sonologist Kanhangad). Note the entire bowel occupying the left side thorax.

Pathogenesis

Pulmonary hypoplasia → *Acidosis* → Pulmonary vascular bed spasm
↗ ↓
Persistent pulmonary hypertension
↙
Persistent fetal circulation

- Hypoplasia not only on left side, it may be bilateral.
- Generations of bronchial branches decreased.
- Extension of tunica media beyond respiratory bronchioles.
- Decreased total lung mass
- Incidence: 1 in 3000 live births, M:F—2:1.
- 90% present in newborn period.
- 90% of CDH are Bochdalek hernias
- Left—80%, right—15%, bilateral—5%.
- Defects are large in babies and small in adults.
- Associated anomalies—10 to 50%, commonest—nonfixation/midgut malrotation. Other anomalies—CNS, cardiac are known to occur and commoner with large/bilateral CDH.

- Very small percentage of patients escape and pass on to adulthood. Incidence—0.17% to 6%
- Incidental detection while evaluating for some other problem. So usually they are asymptomatic.

Poor prognostic criterias in Bochdalek hernia

- Early antenatal detection (before 5th month)
- Small left ventricle
- Large diaphragmatic defect / hemiagenesis
- Stomach in chest
- Polyhydramnios

Clinical features

- Respiratory embarrassment
- Scaphoid abdomen
- Bowel sounds in left side of chest
- Mediastinal shift towards right side
- Occasionally features of intestinal obstruction

Investigations: Chest X-ray, *barium enema (common)* or barium meal, *arterial blood gas analysis (ABG).*

Treatment

- Laparotomy and dissection of the sac (when present) and closure of the defect in the diaphragm (using nonabsorbable suture material).

Note:
Persistent fetal circulation is the major problem in these newborns. It aggravates the postsurgical mortality.

Management

Antenatal	*Postnatal*
• Fetal interventions • Discontinue the pregnancy (poor prognosis) • Continue pregnancy and observation • Shift the mother to higher centre • Mode of delivery—LSCS/ normal	• Stabilise the neonate • Do not resuscitate with Ambu bag • Tracheal intubation and ventilation • Pass nasogastric tube *Surgical correction of CDH* after stabilisation

Triad of congenital diaphragmatic hernia (CDH)

- Respiratory distress
- Scaphoid abdomen
- Mediastinal shift (Pseudodextrocardia)

Association of CDH

- Pulmonary hypoplasia with persistent fetal circulation—is the actual cause for hernia
- Respiratory acidosis

Hernia Through Foramen of Morgagni

- The defect lies between the sternal and costal attachments of diaphragm and is situated in front, towards right. Colon is commonest content. Usually, it is symptom free.

Oesophageal Hiatus Hernia

- Oesophageal hiatus hernia can be congenital or acquired (common).

Traumatic Diaphragmatic Hernia

- Either on right or left side, resulting from road traffic accidents, crush injuries, penetrating injuries, or blunt injuries.
- Patient is under shock.
- On right side along with the liver, the intestines may also get herniated.
- Patient is pale, presents with respiratory distress, guarding and rigidity over the abdomen.

Investigations

- X-ray chest and abdomen.
- U/S abdomen.
- Often CT scan of abdomen and chest is done.

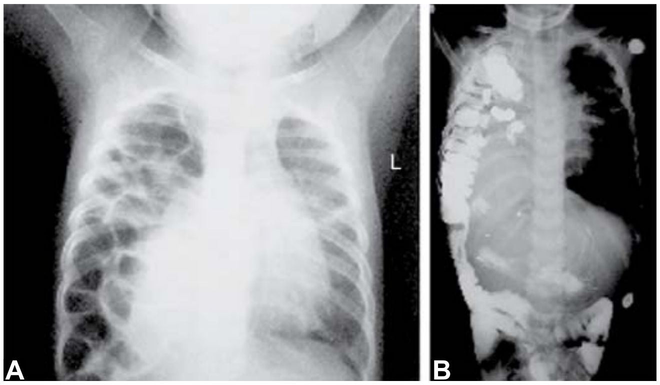

Figs 28.34A and B: (A) Plain X-ray chest, (B) Barium enema, showing bowel loops within thoracic cavity on the right side (diaphragmatic hernia).

Treatment

- Immediate laparotomy and exploration is done. Tear in the diaphragm is sutured.
- Associated injuries in liver or spleen or bowel are treated accordingly.
- Adequate blood transfusion and antibiotics is required.
- Ventilator support is often necessary.

PULMONARY COMPLICATIONS DURING POSTOPERATIVE PERIOD

Precipitating Factors

- Age: Common in infants and elderly.
- Sex: Common in males.
- Common in *smokers.*

An error gracefully acknowledged is a victory.

- Common in people with chronic bronchitis, asthma, tuberculosis or COPD.
- Postoperative pain.
- Deep venous thrombosis and pulmonary embolism.
- Type of surgery: Common in thoracic or upper abdomen surgeries.
- Septicaemia.
- Paralytic ileus.
- Anaesthetic complications.
- Aspiration and its problems.
- Obesity.

Complications

- Bronchopneumonia.
- Lung collapse.
- Bronchitis.
- Lung abscess.
- *Adult respiratory distress syndrome (ARDS).*
- *Respiratory failure.*
- Alkalosis.
- Pleural effusion or empyema.

Investigations

- Chest X-ray.
- Arterial blood gas analysis.

Management

- Suction—aspiration of tracheobronchial tree.
- Respiratory physiotherapy.
- Analgesics to control pain.
- Ventilator support with endotracheal intubation.
- Tracheostomy.
- Control of sepsis by proper antibiotics.

SURGICAL MANAGEMENT OF PULMONARY TUBERCULOSIS

Indications for Surgery

- Failure of medical therapy.
- Tuberculoma.
- Tuberculous bronchiectasis.
- Destruction of lobe.
- Persistent foci, persistent cavity.
- Empyema thoracis.
- Presence of neoplasm with tuberculosis.

Surgeries Done

- Pleural tap.
- Intercostal tube drainage for empyema.
- Rib resection and Eloisier’s drainage.
- Lung resection is undertaken when medical treatment fails or when there is tuberculoma. It is commonly used method.
- Lobectomy or pneumonectomy.
- Decortication.
- Thoracoplasty collapse therapy—it causes elastic relaxation of lung around the lesion leading to closure of cavity.
- Creation of artificial pneumothorax (nitrogen is used).
- Temporary phrenic nerve palsy.
- Extrapleural pneumothorax.
- ***Plombage:*** It is extrapleural pneumolysis using lucite spheres.

VIDEO ASSISTED THORACOSCOPIC SURGERY (VATS)

It is visualisation and doing surgical procedures through video thoracoscopy.

It can be:

- Diagnostic—lung cyst, pleural pathology, malignancy, tuberculosis, lung abscess, emphysematous bullae.
- Therapeutic—cervical sympathectomy, oesophageal mobilisation for carcinoma of oesophagus, lung resection, excision of lung cysts, etc.

Requirements

- Expert skilled personal.
- Double lumen endotracheal tube for one lung anaesthesia.
- Instruments.

Advantages

- Open thoracotomy and its complications are avoided.
- Less painful, fast recovery.
- Magnification and precise dissection.

Disadvantages

- Injury to major structures.
- Longer duration of surgery.
- Learning curve.

Cardiac Surgery

CHAPTER OUTLINE

- Anatomy
- Preoperative Assessment and Preparation of the Cardiac Patient
- Cardiopulmonary Bypass
- Congenital Heart Diseases
- Patent Ductus Arteriosus
- Coarctation of Aorta
- Atrial Septal Defect
- Ventricular Septal Defect
- Pulmonary Stenosis
- Transposition of Great Vessels
- Tetralogy of Fallot
- Acquired Heart Disease
- Mitral Regurgitation
- Aortic Stenosis
- Aortic Regurgitation
- Valve Replacement Surgery
- Ischaemic Heart Disease
- Cardiac Pacemakers
- Postoperative Care

ANATOMY

Heart consists of right and left atrium, right and left ventricle with mitral (bicuspid) valve between left atrium and left ventricle and tricuspid valve between right atrium and right ventricle. Aorta begins from the left ventricle to have systemic circulation. Superior vena cava (SVC) and inferior vena cava (IVC) enters the right atrium to return venous blood. From right ventricle pulmonary artery begins which transport the unoxygenated venous blood into the lungs. From lungs four pulmonary veins collect the oxygenated blood to reach right atrium and to complete the cycle.

Pressure in the right ventricle is low. Pressure in the left ventricle is high.

Chordae tendinae act as guy ropes which connect valves (mitral and tricuspid) to heart through papillary muscles.

Aortic and pulmonary valves are semilunar valves.

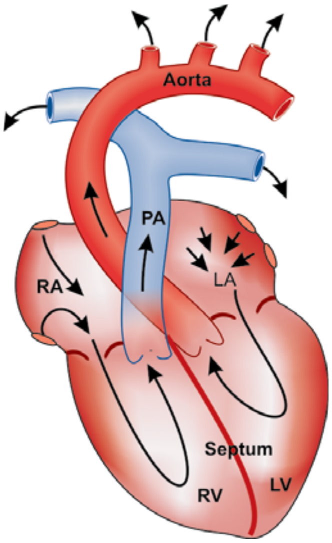

Fig. 28.35: Normal heart anatomy.

First heart sound is produced by closure of atrio ventricular valves. ***Second heart sound*** is produced by closure of semilunar valves. Valvular heart disease may be congenital or acquired. Rheumatic heart disease is the commonest acquired cause for valvular disease. Mitral stenosis is the commonest type.

Right and left ventricles are separated by interventricular septum. Right ventricle is thinner than left ventricle. SVC and IVC opens into right atrium.

Heart derives its blood supply from two ***coronary arteries—right and left***. Left coronary artery once origins at ostium runs for 2 cm, divides into ***anterior descending and circumflex artery***. Anterior descending artery runs in the anterior interventricular groove to reach apex and ascends short distance in posterior interventricular groove. ***The circumflex arter***y runs around the base of left ventricle. Right coronary artery through atrioventricular groove reaches the posterior interventricular groove and runs along it.

Right coronary artery supplies right atrium, entire right ventricle except anterior part, posterior part of the interventricular septum, conducting system except left branch.

Left coronary artery supplies left atrium, entire left ventricle except posterior part, anterior part of the interventricular septum.

There are collaterals within the heart and extracardiac but is not sufficient enough to maintain the adequate perfusion in case of coronary block to the dynamic heart leading into myocardial infarction.

Venous drainage of heart is through superficial and deep veins. Superficial veins (great cardiac, middle cardiac, small cardiac, oblique and posterior veins) form ***coronary sinus*** ending at right atrium. Deep veins directly enter the heart called as ***Thebesian veins.***

First blood passes to the elastic vessels then resistant vessels and capillaries later to veins and then back to heart. In capillary because of great cross-sectional area, rapid transfer of substances across tissues occur.

Have a heart that never hardens and a temper that never tires and a touch that never hurts.

Flow of blood from left side heart through aorta and its branches to capillaries and return to right heart is called as ***systemic circulation.***

Blood from right heart through pulmonary artery to lungs and from lungs through pulmonary veins to left heart is called as ***pulmonary circulation***.

Conduction system of heart consists of ***SA node*** (Sinu atrial node) which is pacemaker of the heart, ***AV node*** (Atrio ventricular node) located in lower part of the atrial septum, ***AV bundle*** (Atrio ventricular bundle) with right and left branches and terminal ***Purkinje fibres*** in subendocardial region.

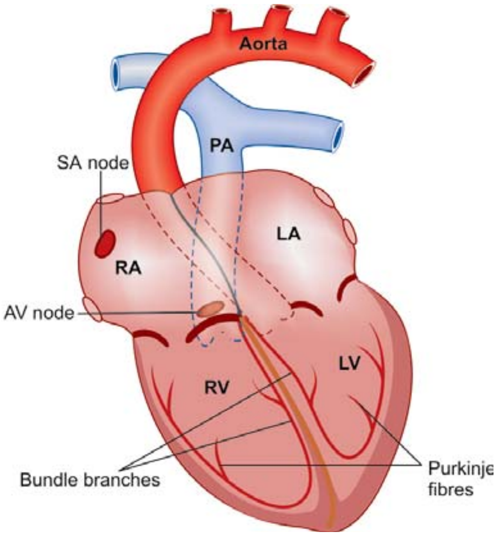

Fig. 28.36: Conduction system of heart.

Surface marking of the heart: Upper border is between 2nd and 3rd costal cartilage at the level of sternum. Lower border is from 2 cm to the right of sternal margin at 6th costal cartilage to the apex of heart at 5th intercostal space 9 cm from the midline.

PREOPERATIVE ASSESSMENT AND PREPARATION OF THE CARDIAC PATIENT

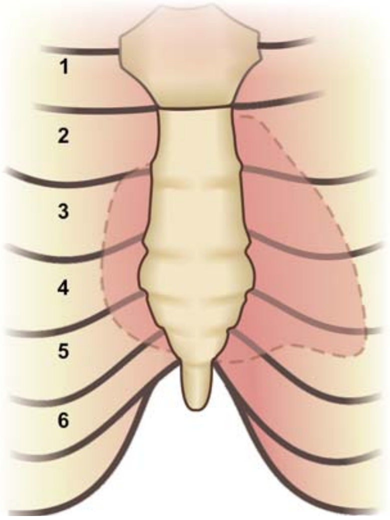

Fig. 28.37: Surface marking of the heart.

- Since cardiac surgery is a major surgery, proper evaluation is essential. Consent, explanation, discussion in initial phases is absolute need.
- Patient will be anxious. He needs proper consolation, moral support.
- Proper history related to smoking, alcohol intake, diabetes, previous cardiac surgery, drug intake is essential.
- ECG, chest X-ray, haemoglobin, haematocrit, blood urea, serum creatinine, blood sugar, lipid profile, blood grouping are essential basic investigations.
- ***Echocardiography*** is ultrasound examination of heart. Heart chambers, pericardium, pumping capacity, valves and their structure, ejection fraction are found through this. Usually cardiologists do this test and give opinion about it.
- **Treadmill stress test** *to find out the response of heart to work.*
- ***Coronary angiogram*** is done to find out the amount of coronary function or its block. Special coronary artery catheter is passed through the brachial or femoral artery like in seldinger angiogram. This catheter is negotiated into the coronary arteries both right and left. Radio-opaque water soluble dye is injected into the catheter and coronaries. X-ray film will confirm the status of the coronaries. Procedure is done under guidance using C-arm. Amount of block is assessed and categorised.
- ***Cardiography*** is the procedure wherein dye is injected into the heart chambers using specialised catheters and size, shape, presence of clot in the chambers, valvular incompetence are studied.
- ***Cardiac catheterisation*** is done to get cardiac ***oximetry and pressures*** in different chambers of the heart. Cardiac catheter is flexible 100 cm long tube passed to the heart under guidance. Right and left sides are assessed separately. Catheter is infused with heparinised solution to prevent intravascular clotting during procedure.
- ***Catheter to the right side*** of the heart is passed through the cubital vein, subclavian vein, SVC. Pressures at systole and diastole are studied in right atrium, right ventricle, pulmonary artery. Blood samples are collected from these three levels to study ***oxygen saturation (oximetry).***
- ***Left side of the heart*** is approached by two approaches—one is through the right atrium penetrating the atrial septum, catheter reaches the left atrium, and so to left ventricle and aorta. Pressure is recorded in each chambers. Blood samples are collected for oxygen saturation.
- ***Pressure*** will alter in valvular diseases like stenosis or regurgitation.
- ***Oxygen saturation*** is altered in shunts between right and left sided. Shunt occurs between systemic and pulmonary circulation.
- *Cardiac isotope scanning.*

- Pressure in atrium is low 0-10 mmHg. Right atrium is little lower than left
- Pressure in the right ventricle is 15-30 mmHg systolic and 0-8 diastolic
- Pressure in pulmonary artery is 15-30 mmHg systolic and 3-12 diastolic
- Pressure in left ventricle is 100-140 systolic and 3-12 mmHg diastolic
- Pressure in aorta is 100-140 systolic and 10-80 mmHg diastolic

- Pulmonary function study is done by chest X-ray, spirometry, blood gas analysis.
- Focus of infection should be looked for like dental infection, throat infection, skin infection.

CARDIOPULMONARY BYPASS

Here functions of the heart and lungs are temporarily replaced by heart—lung machine which acts as a pump and oxygenator to the body.

Heart is exposed through a midline sternotomy incision.

Venous blood from the SVC and IVC is diverted to the heart lung machine by placing tubes to them.

From the heart lung machine oxygenated blood is pumped through a tube which is passed to the ascending aorta so as to bypass the heart and lungs with retaining the function so tissues will have adequate required perfusion. Blood should be heparinised to prevent clotting. Often venous tube is placed in the right atrium.

Temperature is lowered over the heart by poring cold saline. Ventilation is discontinued.

Once procedure in the heart is over, air in the chambers of the heart is cleared. Coronary artery perfusion is restored and heart is made to beat spontaneously or using DC shock. Acidosis and hypocalaemia is corrected. Patient is warmed.

Heart is allowed to fill gradually and slowly heart lung machine blood flow is reduced. Cannulas are removed from the veins (SVC and IVC) and also from the aorta. Protamine sulphate is given to reverse the effect of heparin.

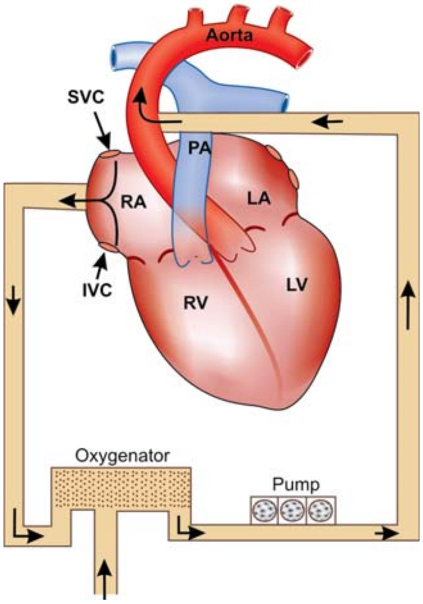

Fig. 28.38: Figure showing cardiopulmonary bypass.

Myocardial protection is achieved by:

- Infusion of solution with 20 mmol/litre of potassium into the coronaries
- Cooling cardioplegic infusate to 6°C
- Surface of the heart is sprayed with 4°C Hartmann's solution
- Body temperature is reduced to 20-27°C
  Heart can be kept arrested for 2 hours.

CONGENITAL HEART DISEASES

Left to right shunt—it leads into overloading of the pulmonary circulation. Usually it is acyanotic.

PDA—patent ductus arteriosus.
ASD—atrial septal defect.
VSD—ventricular septal defect.

Right to left shunt—Blood from the systemic veins enter the systemic arteries bypassing the lungs allowing severe hypoxic blood to enter the systemic circulation. It will cause cyanosis. Examples are tetralogy of Fallot and transposition of great vessels.

Conditions without shunts—coarctation of aorta, aortic stenosis, pulmonary stenosis.

Also classified as ***cyanotic and acyanotic heart diseases.***

PATENT DUCTUS ARTERIOSUS (PDA)—10%

Ductus arteriosus is essential in fetal life wherein blood from pulmonary artery reaches the aorta through this ductus arteriosus by passing the lung. Once child is born, this ductus closes because of the beginning of the lung function. Prostaglandin prevents closure. Indomethacin promotes closure. If it persists in child it is called as PDA.

In PDA, part of the blood from aorta flows back to pulmonary artery because of high pressure in aorta (left to right shunt). This causes pulmonary hypertension.

Occasionally pressure is so severe that flow may reverse and deoxygenated blood from pulmonary artery may flow towards aorta causing reversal of left to right shunt converting acyanotic shunt to a cyanotic one. It is called as ***Eisenmenger's syndrome***. It can occur in other left to right shunt also (ASD, VSD).

Child presents with dyspnoea, chest pain, machinery murmur.

Aortogram and echocardiography are diagnostic.

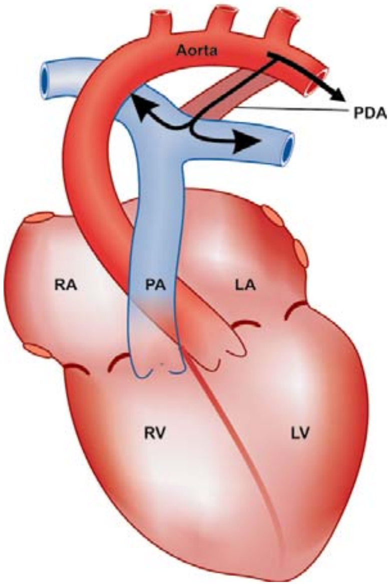

Fig. 28.39: Patent ductus arteriosus.

Treatment—closure of the PDA either through interventional cardiology or more conveniently through open surgery.

Complications of PDA—congestive cardiac failure, bacterial endocarditis.

COARCTATION OF AORTA—5%

Narrowing of the aorta just distal to the origin of left subclavian artery. It is often associated with PDA, VSD.

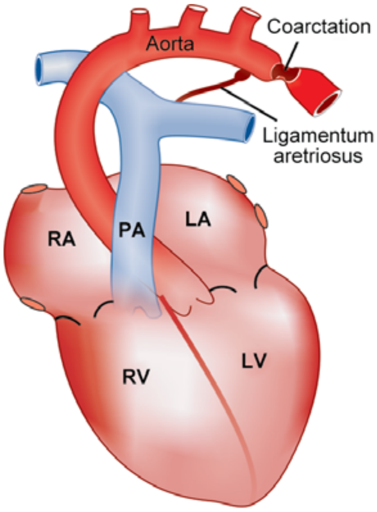

Fig. 28.40: Coarctation of aorta.

There is increased perfusion of upper limb, cranium, face. But less blood supply occurs to lower limb, kidneys.

Types

- It can be preductal or postductal depending on relation to the ligamentum/ductus arteriosus.

Features

- It causes left ventricular hypertrophy, differential blood pressure (pressure upper limb is higher but pressure in lower limb is lower). Often there is differential cyanosis.

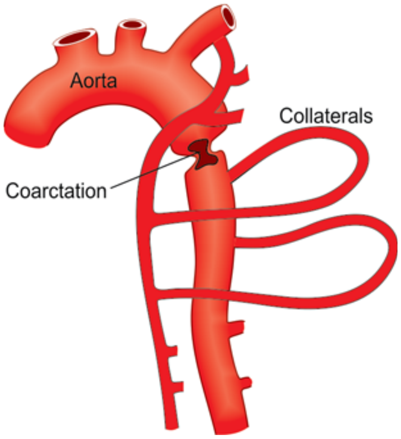

Fig. 28.41: Collaterals development in intercostal vessels in coarctation of aorta.

- Visible dilated intercostal vessels through collateral perfusion is more obvious when patient leans forwards.
- X-ray will show dilated proximal aorta, notching of ribs due to erosion of the intercostals vessels.
- Aortogram is diagnostic.

Treatment *is widening of the narrowed segment of aorta with Dacron patch graft, or removal of stenosed segment and arterial graft.*

Complications

- Congestive cardiac failure, bacterial endocarditis, aortic rupture.

ATRIAL SEPTAL DEFECT (ASD)—7%

- An ASD is the hole in the atrial septum thus causing communication between left and right atrium.

Types

- ***Secundum type*** is the commonest one. It is elliptical, in the middle of the atrial septum and is due to defect in the site of foramen ovale
- ***Primum type*** is lower part. It is rare
- ***Sinus venosus type*** is near SVC opening

Features

- There is shunting of blood from left atrium to right atrium. It causes left ventricular hypertrophy, pulmonary hypertension, eventually occasionally reversal of shunt as right to left leading into cyanosis.
- Dyspnoea, fixed split of second heart sound, systolic murmur due to increased blood flow through the pulmonary valve are the clinical features.
- It is diagnosed by echocardiography, cardiac catheterisation and angiocardiogram.

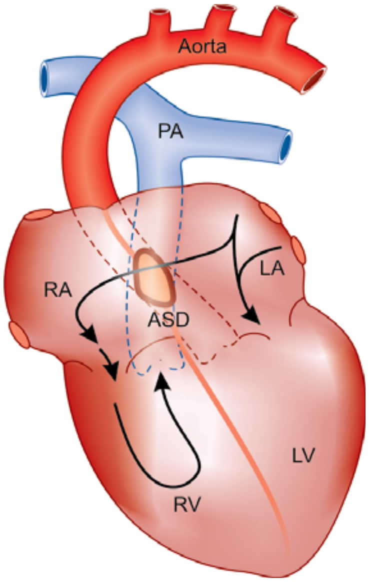

Fig. 28.42: Atrial septal defect.

Treatment

- Closure of the defect by direct suturing or using patch either pericardium or Dacron.

VENTRICULAR SEPTAL DEFECT (VSD)—15%

- VSD is defect in the ventricular septum causing left to right shunt. There is pulmonary hypertension, ventricular hypertrophy.

Types

- Four types of VSD are present depending on location of the defect. Defect can be small or large.
- It may be associated with tetralogy of Fallot.
- *Severe cases of shunt become reversal causing* **Eisenmenger's** *syndrome.*

Clinical Features

- Pansystolic murmur, palpable thrill, split second sound with pulmonary accentuation, recurrent respiratory infections.

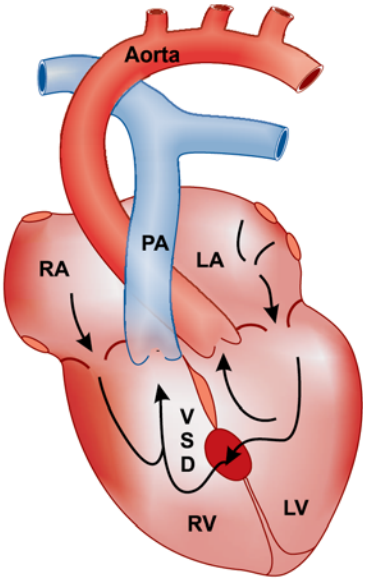

Fig. 28.43: Ventricular septal defect.

- 40-50% of defects especially when they are small, will close spontaneously. Spontaneous closure will not occur after the age of 6.

Investigations

Echocardiography, cardiac catheterisation.

Treatment

Closure of the defect using pericardial patch or Dacron. While suturing care should be taken not to inure atrioventricular conduction system.

Complications of VSD

Recurrent respiratory infection, pulmonary hypertension, cardiac failure.

PULMONARY STENOSIS

It is narrowing of the pulmonary valves due to fusion of cusps. Narrowing may be as less as 2-4 mm. Patient develops right ventricular hypertrophy and later right ventricular failure. Pressure in the right ventricle is higher, in pulmonary artery is lower. Treatment is pulmonary valvotomy.

TRANSPOSITION OF GREAT VESSELS

Here aorta begins from right ventricle and pulmonary artery begins from left ventricle. So venous blood enters the aorta and causes hypoxia and cyanosis. It is usually not compatible with life but often through bronchial collaterals communication may develops and become compatible. When it is often associated with ASD or VSD or PDA mixing occurs and is compatible.

It is cyanotic heart disease.

It is diagnosed by an emergency echocardiography and aortocardiogram.

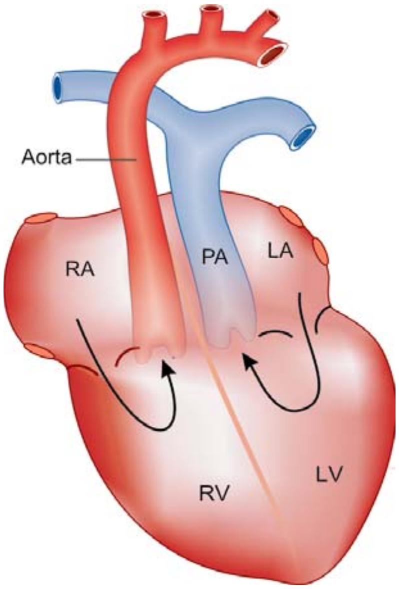

Fig. 28.44: Transposition of great vessels.

Treatment

- Immediate palliative method is ***Rashkind balloon septostomy***. A balloon is passed into the right atrium and through the septum into the left atrium. Balloon is inflated and rapidly pulled out through the atrial septum causing widening of the defect in the septum. It allows flow of oxygenated blood from pulmonary veins through the septum and so to the right ventricle and aorta.
- ***Albert-Mustard atrial redirection of flow of blood:*** Atrial septum is excised and using either pericardial or Dacron

patch left atrial blood is directed to the right ventricle and right atrial blood is directed to the left ventricle.

- Later definitive correction is done either at atrial level or at great arterial level.

TETRALOGY OF FALLOT

It is a cyanotic congenital heart disease.

Components

- Ventricular septal defect
- Infundibular and pulmonary stenosis causing right ventricular outflow obstruction
- Right ventricular hypertrophy
- Dextroposition of aorta

Because of right ventricular outflow obstruction venous blood gets diverted through the existing VSD into the left ventricle and so to the aorta. So systemic circulation is provided with unoxygenated venous blood causing cyanosis.

Features

- Cyanosis with dyspnoea. Ejection systolic murmur which diminishes during cyanotic attack.

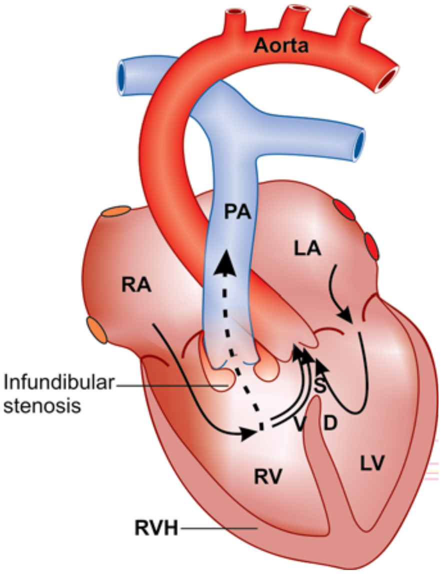

Fig. 28.45: Tetralogy of Fallot.

- Echocardiography, cardiac catheterisation are the essential investigations.

Treatment

- ***Blalock-Taussig operation***—it is done as initial treatment in newborn baby with Fallot's tetralogy. Left subclavian artery and pulmonary artery is connected using a graft which allows the pulmonary arterial perfusion.
- ***Second stage definitive correction*** by relieving the right ventricular outflow obstruction completely and closure of VSD using Dacron patch.

ACQUIRED HEART DISEASE

Mitral Stenosis

Commonest acquired heart disease. Valve here cannot open completely during left atrial contraction.

It is scarring and thickening of the mitral valves due to inflammation of valves with valve area less than 2.5 cm^2 (normal is 4-6 cm^2). It will later lead into fusion of chordae tendinae, calcification of the valve cusps. It is commonly due to rheumatic fever with streptococcal pharyngitis.

Features

Mitral stenosis
↓
Hypertrophy of the left atrium
↓
Sluggish blood flow, left atrial hypertrophy
↓
Atrial fibrillation → atrial thrombus → emboli
↓
Pressure in pulmonary vein increases
↓
Lung congestion, oedema
↓
Pulmonary artery hypertension
↓
Right ventricular failure and tricuspid regurgitation (CCF).

Mitral stenosis—clinical features and investigations

- Dyspnoea, orthopnoea
- Cough, haemoptysis
- Mid-diastolic murmur, with loud heart sound
- Features of atrial fibrillation
- Chest X-ray shows straightening of the left border of the heart
- ECG and echocardiography will demonstrate mitral stenosis, clot
- Cardiac catheterisation shows increased right atrial pressure
- Pulmonary capillary wedge pressure (PCWP) shows increased left atrial pressure

Indications for surgery

- Severe stenosis
- Calcification of cusps
- Clot in the left atrium
- Right ventricular hypertrophy

Surgeries for mitral stenosis

- Closed mitral valvotomy—through left thoracotomy incision, heart is exposed. Left atrium is opened with purse string suture so as to allow the finger to pass inside. Left ventricular tip is incised with a purse string suture and Tubb's valve dilator is passed. Under finger guidance through left atrial purse string wound, valve and cusps are separated using ***Tubb's dilator.***

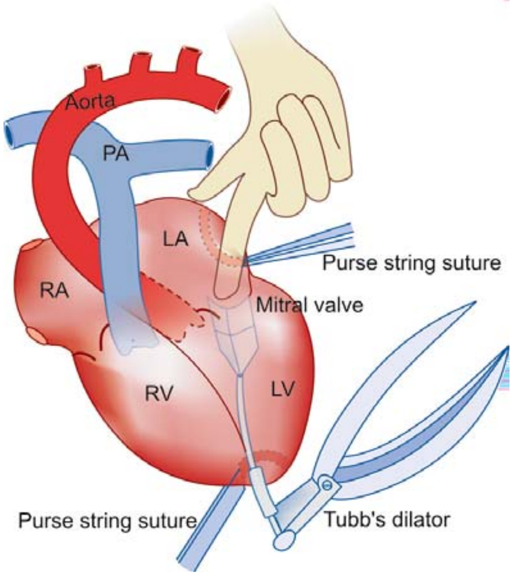

Fig. 28.46: Mitral valvotomy using Tubb's dilator.

- *Open mitral commissurotomy*—it is done with cardiopulmonary bypass, left atrium is opened, valve and commissures are separated, valve is dilated using finger, clot in the left atrium is removed.
- Open valvuloplasty by doing repair of the valve.
- Percutaneous balloon valvotomy.
- Valve replacement surgery.

MITRAL REGURGITATION

Mitral valve here cannot close completely during left ventricular contraction. It can be associated with mitral stenosis also.

Causes

- Rheumatic heart disease causing fibrosis, rigidity and shortening of chordae tendinae leading to inability to close the mitral valve.
- Myxomatous degeneration of the valve.
- Marfan's syndrome.
- Connective tissue disease.
- Sarcoidosis.

Mitral regurgitation
↓
Left ventricular hypertrophy but reduced cardiac output
↓
Left atrial enlargement with increased pressure
↓
Pulmonary venous pressure increases
↓
Lung congestion, oedema (CCF)

Features

- Cough, haemoptysis, chest pain, dyspnoea
- Pansystolic murmur
- Features of failure
- Often associated features of mitral stenosis
- Features of associated diseases
- Echocardiography, cardiac catheterization and PCWP are the investigations

Treatment

Valve replacement surgery.

AORTIC STENOSIS

Narrowing of the aperture of aorta is due to rheumatic heart disease, bicuspid stenosis, calcification. Valve is thickened, fibrosed, calcified and narrowed.

Features

- There is left ventricular hypertrophy with reduced left ventricular capacity.
- Because of the narrowing of the valve, cardiac output decreases.
- Systolic murmur is better heard over aortic area.
- Later may lead into failure.
- Echocardiography is diagnostic.

Treatment

Valve replacement surgery.

AORTIC REGURGITATION

There is incompetence of aortic valve.

Causes

- Rheumatic disease.
- Connective tissue disorder.
- Marfan's syndrome.
- Tertiary syphilis.
- Endocarditis.
- Aortic dissection.

Features

- There is left ventricular hypertrophy, and cardiac failure later
- Early diastolic murmur is characteristic
- Collapsing pulse
- Dyspnoea, chest pain
- Echocardiography is diagnostic
- It may be associated with ischaemic heart disease

Treatment

Valve replacement surgery.

VALVE REPLACEMENT SURGERY

Indications

- Severe valvular disease with fixity, stenosis
- Valve disease with thrombosis like in mitral stenosis
- Multivalvular disease

Types

1. ***Biological valves***—Capentier's glutaraldehyde preserved porcine valve with a sewing ring and frame is used. It is not thrombogenic. So, it is used in child bearing age and in whom warfarin is contraindicated. But its life span is less. So it is useful in elderly.
2. ***Mechanical valves***
    - Starr-Edwards ball—valve prosthesis
    - Bjork-Shiley tilting disc—valve prosthesis.

Mechanical valves are long lasting but they are thrombogenic. Embolic phenomenon can occur. Patient requires oral anticoagulant to keep prothrombin time twice normal. If anticoagulants cannot be used, then aspirin, dipyridamole can be used.

Problems with mechanical valves

- Thrombogenic
- Require warfarin anticoagulant therapy
- Prone for severe epistaxis, intracranial haemorrhage while on warfarin therapy
- Prone for infections by bacteria, fungi, yeasts
- Mechanical haemolysis and anaemia
- Valve malfunction may occur

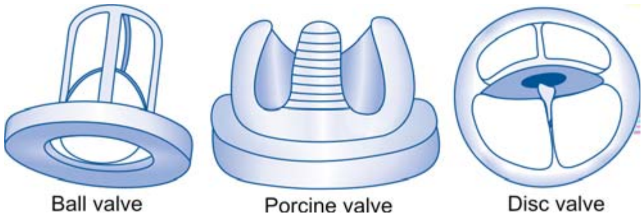

Fig. 28.47: Different types of valve prosthesis.

3. ***Free homograft valve***.

ISCHAEMIC HEART DISEASE (IHD)

Main major arteries of the heart are right coronary, left coronary and circumflex artery. Blockage can occur in one or two or all three vessels. Involvement of all three vessels is called as triple vessel disease.

IHD presents as ***asymptomatic*** with sudden massive cardiac ischaemia with sudden death.

Or ***symptomatic angina,*** is progressive if not treated and eventually may lead in to severe MI.

Atherosclerosis is the commonest cause.

Risk factors are:

- Hypercholesterolaemia
- Cigarette
- Diabetes mellitus
- Hypertension
- Family history and age

Atheroma plaques deposits in the coronary artery and narrows it. Block above 70% is called as critical block. Angina is initially treated with calcium antagonists, beta blockers, nitrates.

Investigations

- ECG.
- Coronary angiography is diagnostic.
- Echocardiography.
- Lipid profile (done on empty stomach).

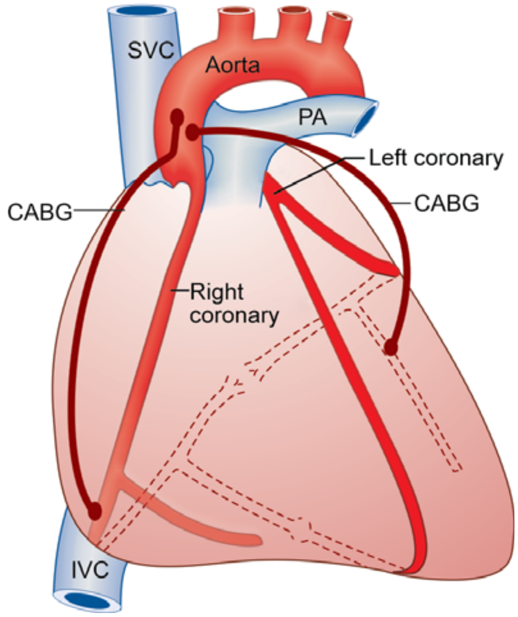

Fig. 28.48: Coronary artery: Anatomy and sites of CABG.

Treatment

1. *Coronary balloon angioplasty* is done in case of angina, only one or two vessels are involved. It is done by interventional radiologists in presence of cardiologists. Balloon tip catheter is passed through femoral or radial artery to reach proximal aorta and entered into the coronary artery opening. Coronary angiogram is done after injecting the dye. Balloon tip of the catheter is passed across the site of the block. Balloon is inflated to crush the atheroma.

Problems with angioplasty

- Restenosis
- May precipitate block

2. Coronary artery stenting after balloon angioplasty.
3. Coronary artery bypass graft (CABG).

Indications

- Triple vessel stenosis
- Left main stem stenosis
- Chronic stable angina
- Angina with persistent chest pain
- Block more than 70%

Grafts Used

- Patient's own long saphenous vein—*commonly used graft*
- Patient's own internal mammary artery—*choice graft* as it is reliable, less occlusion, long-term patency.

Usually 3-4 grafts are used.

Technique—It is done under cardiopulmonary bypass. Aortic cross clamps are used to help proper graft anastomosis.

Complications of CABG

- Graft thrombosis
- Narrowing
- Atherosclerosis in the graft
- Respiratory infection

CARDIAC PACEMAKERS

Cardiac rate is made regular using specialised device which stimulates the heart and acts as pacemaker.

Temporary Pacing

It is done in myocardial infarction, arrhythmias, and cardiac surgery.

Permanent Pacing

It is done in congenital and acquired atrioventricular block, sick sinus syndrome (SA node dysfunction), atrial fibrillation, certain tachycardia.

Routes of Pacing

Transvenous route is commonly used. A needle is passed into the subclavian vein and guide wire is passed through the needle into SVC. Dilators are passed through the guide wire. Then sheath is passed. Electrode is passed into the atrium or ventricle. It is endocardial pacing.

Transthoracic pacing—it is not commonly used. It requires open thoracotomy. One lead is placed to the atrium and two other to ventricle.

Pacemaker may be ***external*** usually used as temporary.

Internal pacemaker is placed in the upper chest wall below the middle of the clavicle in subfacial plane under local anaesthesia (pacemaker pocket).

POSTOPERATIVE CARE

- Proper monitoring and ICU care is essential.
- ECG, temperature, BP, CVP, cardiac monitor, urine output has to be assessed regular intervals.
- Serum electrolytes, blood gas analysis, assessment of isoenzymes, haematocrit, blood urea, creatinine.
- Acidosis and hypercalcaemia to be corrected early.

Complications of pacing

- Arrhythmias
- Infection
- Phlebitis, haematoma, thrombosis, bleeding
- Failure of battery, electrode
- Electrode displacement

Types of electrode activity

- Fixed rate—ventricular pace—no sensing
- Demand release—ventricular pace—sensing-triggered mode
- Inhibited rate—ventricular pace—sensing-inhibited mode

- *Pacemakers are sensitive to electromagnetic interference.*
- Drains from pericardium, mediastinum should be assessed carefully.
- Hypovolaemia, cardiac tamponade, myocardial infarction, hypoxia, acid-base imbalance has to be corrected.
- Dopamine, dobutamine, isoproterenol are the drugs used when there is low cardiac output.
- Digoxin, pacing, cadioversion may be required if there is cardiac arrhythmias.
- Management of respiratory system—when the patient is alert and having adequate cough reflux, endotracheal tube is removed. Breathing exercises, physiotherapy, observation for respiratory distress, prevention of pneumonia to set in are important care required.
- Care of the endotracheal tube, suctioning, checking the parameters in intermittent positive pressure ventilation (IPPV).
- Neurological system should be observed—alertness, speech, level of consciousness, any deficits are looked for.
- Strict asepsis is undertaken in cardiac postoperative period by avoiding unnecessary visitors, separate dress in postoperative ward, wearing gloves, cap, masks while attending the patient.

Chapter 29 Adjuvant Therapy

CHAPTER OUTLINE

- Radiotherapy
- Chemotherapy
- Cell Cycle
- Antimalignancy Drugs
- Hormone Therapy in Cancer
- Immunosuppression
- Immunotherapy
- Hybridoma
- Gene Therapy

RADIOTHERAPY

> *The effect of radiations on living cells is the more intense: (1) the greater their reproductive activity, (2) the longer their mitotic phase lasts, and (3) the less their morphology and function are differentiated.*
>
> —**Jean Alban Bergonié, Louis Mathieu Frédéric Adrien Tribondeau,** 1906

- It is use of ionizing radiation as therapy, mainly in malignant conditions.
- Megavoltage X-rays or gamma rays which generates energy greater than 1×10^6 V are used.

Principles and Factors in Radiotherapy

- Penetration of the beam into the deep seated tumour.
- Building of the radiation dose under the skin over the tumour tissue so as to minimize the skin reactions.
- Precise targeting of radiation towards the tumour.
- Radiosensitivity, tumour volume and tumour size are the factors which determine the radiation efficacy.

Physicochemical Effects of the Radiation

- Direct action on the target tissues.
- Indirect action, by releasing free radicals which cause cell destruction.
- It acts on the different phases of cell cycle to destroy the cell.

Changes in the Tumour Cells after Irradiation

1. Repair of the damaged but retained cells.
2. Repopulation is by proliferation and reproduction of the retained cells.
3. Redistribution of the cells. Cells in the 'S' phase of the cell cycle (where DNA is being synthesised) are radioresistant. These cells after radiotherapy will go to 'G_2' or 'M' phase.
4. Reoxygenation of the cells which were hypoxic at the time of radiotherapy. Hypoxic tumour cells are radioresistant. About 10% of tumour cells are in hypoxic state.

Thus fractionation of the radiotherapy allow retained cells to go for active and sensitised phase making radiotherapy more effective.

Oxygen is a good radiosensitiser. So hyperbaric oxygen is used many times as radiosensitiser.

Radiation Unit

It is called as grey unit. It is defined as the absorption of 1 J of radiation energy by 1 kg of tissue.

Types of Radiotherapy (Therapeutic)

1. ***Curative (radical) radiotherapy:*** It is given in early malignancies as a curative or primary method to tumours which are radiosensitive.
2. ***Palliative radiotherapy:*** It is used to alleviate the symptoms when the tumour is beyond cure.

Symptoms are:

- ***Pain*** in secondaries in the bone
  - Myeloma.
  - Carcinoma breast.
  - Advanced lymph node metastasis.
- ***Bleeding***
  - Carcinoma oral cavity, cervix.
  - Carcinoma bladder, rectum.
  - Advanced secondaries in the lymph nodes.

Common tumours

- Seminoma testis
- Hodgkin's lymphoma
- Squamous cell carcinoma
- Ewing's sarcoma
- Carcinoma lung
- Gliomas
- Bladder tumours
- Carcinoma cervix, vagina, nasopharynx,
- head and neck, oesophagus
- Basal cell carcinoma

Dose is 60 Gy in 30 fractions over 6 weeks.

- ***Fungation***
  - Breast carcinoma
  - Rectal carcinoma.
- ***Obstruction***
  - Bronchogenic carcinoma obstructing inferior vena cava
  - Carcinoma cervix causing obstruction to the ureter.
- ***Pathological fracture*** due to secondaries.
- ***Spinal cord compression*** due to secondaries.

Dose is 10-25 Gy in 1 to 5 fractions.

Principle

Radiation is used to kill actively proliferating malignant cells at their mitotic level.

Common malignancies which are radiosensitive

- Squamous cell carcinoma
- Basal cell carcinoma
- Bladder tumours
- Carcinoma cervix
- Seminoma testis
- Hodgkin's lymphoma

Other areas where radiotherapy is used commonly:

- In carcinoma breast during postoperative period, for secondaries in bone, sometimes as preoperative radiotherapy, as part of *QUART Regime.*
- In follicular carcinoma of thyroid: For secondaries radio active I^{131} 5 m curies is given orally. External radiotherapy is also given for bone secondaries along with internal fixation if there is pathological fracture.
- Malignant brain tumours like astrocytomas.
- Many sarcomas.
- Carcinoma prostate.
- Carcinoma oesophagus.
- Carcinoma lung.
- Fixed secondaries in neck as palliation to palliate pain, fungation and erosion into major vessels.
- Multiple myelomas.

Radioresistant Tumours Include

- All GI malignancies.
- Melanoma.
- Medullary carcinoma of the thyroid.

Source of Radiotherapy

- Cobalt 60 machine—most commonly used.
- Kilovoltage machines.
- Linear accelerator.
- Betatron/microton.
- Radioactive materials like Cesium 137 (pellets), Iridium 192 (wire), Gold 198 (seeds), Iodine 125 (seeds).

Measurement of Ionising Radiation

Roentgen unit is a measure of ionisations produced per unit volume of air by X-rays and gamma rays but it is not used for photon energies above 3MeV. SI unit for exposure is Coulomb per kg.

Radiation Absorbed Dose (RAD)

Absorbed dose is De/dm. It is mean energy imparted by the ionizing radiation to material of mass dm.

Old unit is *rad* which is equivalent to 100 ergs of energy per gram.

Newer International Unit is Presently Used

One Grey *(Gy)* is one joule of energy deposited per kilogram of material. One Gy = 100 rads = 100 cGy. 1 cGy = 1 rad.

Radiotherapy Plan

Type and doses of radiotherapy is decided by following factors:

- Site of tumour.
- Lymphatic field.
- Tumour size and extent.
- Histological type and grading.

Types of radiotherapy (Technical)

- Superficial X-ray therapy (100 KV)
- Orthovoltage X-ray therapy (200/300 KV)
- Caesium (^{137}Cs) or cobalt (^{60}Co) teletherapy
- Intracavitary or intralesional radiotherapy (Brachytherapy)
- Radioactive I^{131} therapy for follicular carcinoma thyroid

Brachytherapy

- It is radiation given with source close to the tumour.
- It is given using iridium^{192} caesium^{137}.
- It is curative radiotherapy.
- It is used in carcinoma oral cavity, penis, breast, cervix and bladder.
- Radiation material placed in the cavity is called *intracavitory* RT.
- Radiation material is inserted into the tissues—*interstitial* RT.
- Implants can be kept permanently or temporarily.
- Radioactive material is placed into the cavity/tissue through applicators under general anaesthesia.
- Intraoperative radiotherapy is also becoming popular.
- It has only localised effect with adjacent tissue being spared.

"Cancer" word itself is lethal many times than cancer.

Radiotherapy sources	*Classification*	*Efficacy depends on:*
• X-rays, electrons-from high energy X-ray machine, i.e. Linear accelerator • Gamma and beta rays from radioactive isotopes like Cobalt-60, Caesium 137, Iridium 197 • Protons, neutrons and heavy ion nuclei from cyclotrons	• *Low LET* (Low Energy Transfer) radiation—X-rays, gamma rays, electron • *High ET* radiation—Neutrons, protons, alpha particles and negative ions. It has heavier mass with dense localisation and hence biologically more effective	• Cell cycle phase • Oxygen content of cells *Methods of delivery* • External beam radiation (Teletherapy) • Brachytherapy • Systemic irradiation using radioisotopes like I^{131}
Preoperative radiotherapy *Advantages* • Down-stage the tumour and reduces the tumour bulk • No change in oxygenation of tissues • Blockage of lymphatics by RT prevents tumour spread during surgical dissection • Reduces the chances of microscopic spread *Disadvantages* • Delayed healing due to reduced vascularity • Flap necrosis, fistula formation • Carotid blow-out	*Postoperative radiotherapy* • More effective and extent of RT is well defined • Flap necrosis, fistula complications are less *Indications* • When resected margin is positive for tumour • When bone/cartilage are involved • Extracapsular nodal spread • Multiple neck nodes/node more than 3 cm in case of neck disease • Poorly differentiated tumour	*Radiotherapy and chemotherapy combined* • Commonly used before or during RT • Chemotherapy after RT is not commonly followed • Chemotherapy before RT is called *as induction chemotherapy* which reduces the bulk of tumour without altering the vascularity • Chemotherapy with RT is called as *concomitant RT* which improves the effect of RT. Methotrexate and bleomycin are radio sensitisers

External beam radiation:

Given using cobalt 60 teletherapy or X-ray source.

Advantages—Deep penetration, skin sparing, better dose distribution.

Internal radiation:

Used for diagnostic and therapeutic purpose, e.g. in carcinoma thyroid for treatment of secondaries.

Complications of Radiotherapy

- Oral mucosal oedema, ulceration, dysphagia, dyspnoea, loss of taste, oral thrush.
- Bone marrow suppression.
- Oophoritis in women and oligospermia in men.
- Effecting lens, eyelashes and lacrymal glands causing dryness, cataract.
- Mucosa in GIT is very sensitive causing nausea, vomiting and diarrhoea.
- Radiation dermatitis, pigmentation in skin.
- Radiation myelitis of the spinal cord—causing hemiplegia, paraplegia.
- Radiation pneumonitis and pulmonary fibrosis.
- Radiation nephritis and renal failure.
- Radiation osteomyelitis.
- Radiation-induced malignancies.
- Radiation-induced thimble bladder.
- Radiation-induced frozen pelvis.
- Infection—bacterial, viral, candidial *(oral thrush).*

Brachytherapy

Type	*Advantages*	*Disadvantages*
• *Intracavitary* like in uterus, urinary bladder, maxillary antrum, bronchial or oesophageal tree • *Interstitial* wherein radioactive needles/wires/ribbons/seeds are inserted into the tumour area like in bladder or oral cavity • *Surface brachytherapy* using moulds like in tumours of skin/eye/breast • Radionuclides used are Caesium 137 (Cs^{137}), Iridium 192 (Ir^{192}), Gold 198 (Au^{198}) and Iodine 125 (I^{125}) • Brachytherapy is often combined with external beam radiotherapy	• High, localised, single continuous dose of RT • Deeper and adjacent tissues are spared • High dose rate with short time • Less side effects • Curative and effective in early cancers • After loading devices are available which reduces personal exposure • Surgery is avoided and part is retained	• Technically difficult • Availability of the facility • Local complications like displacement/erosion

CHEMOTHERAPY

- Chemotherapy, today is an important modality of treatment in managing a case of cancer.
- It is the main modality of treatment in most of the advanced malignancies.
- It is used preoperatively to downstage the tumour so as to increase the possibilities of surgical resection.
- When used postoperatively it reduces the rate of recurrence.
- It is the primary treatment for NHL, leukaemia.
- These drugs mainly act by blocking the mitotic activity in the nucleus in different phases of the cell cycle.
- It is given either oral, intravenous (*systemic*), intra-arterial, intrathecal and intravesical *(regional*) or as isolated limb perfusion.
- Intra-arterial chemotherapy is used in head and neck cancers and hepatic cancers.
- *Isolated limb perfusion* is used in melanoma.

CELL CYCLE

The cell cycle is composed of:

1. ‘G_0’—resting phase or nonproliferative phase.
2. Presynthetic phase (G_1)—resting cells that are not preparing for cell division are said to be in this subphase G.
3. ‘S’ phase—DNA synthesis takes place during this phase.
4. ‘M’ phase—formation of mitotic spindle, i.e. period of mitosis.
5. Postsynthetic phase (G_2)—premitotic interval, which follows the termination of DNA synthesis.

Most potent cytotoxic agents act by damaging DNA. Their toxicity is high during the DNA synthetic phase, ‘S’ phase of the cell. Some agents block the formation of mitotic spindle in M phase (vinka alkaloids and taxanes). So the agents have activity only against cells that are in the process of division. Hence, tumour with high fraction of actively dividing cells are more susceptible to chemotherapeutic agents.

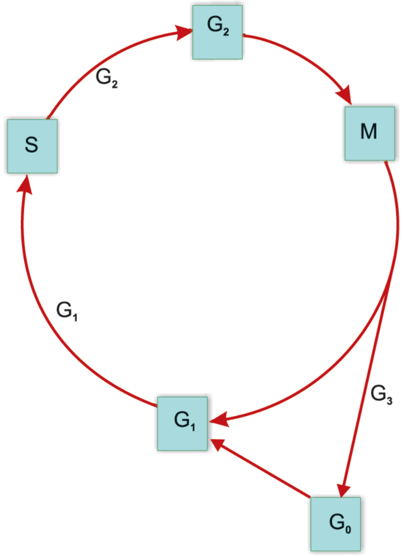

Fig. 29.1: Cell cycle.

When chemotherapy is given with interferons and interleukins, it is called as **biochemotherapy**.

ANTIMALIGNANCY DRUGS

I. Classification

1. *Alkylating agents:*
   a. Nitrogen mustards: Mechlorethamine (mustine hydrochloride), cyclophosphamide, chlorambucil, melphalan.
   b. Ethylenimines.
   c. Alkylsulfonates—busulfan.
2. *Antimetabolites*:
   a. Folic acid antagonists—methotrexate.
   b. Purine antagonists—azathioprine, 6-mercaptopurine.
   c. Pyrimidine antagonists—5-fluorouracil,
3. *Antibiotics:* Actinomycin-D, mitomycin, doxorubicin, bleomycin.
4. *Vinca alkaloids:* Vincristine, vinblastin.
5. *Miscellaneous:* Cisplatin, procarbazine.

II. Classification of antimalignancy drugs based on the action on cell cycle

S-phase specific agents: Cytosine arabinoside, hydroxyurea, 6-mercaptopurine.

M-phase specific agents: Vincristine, vinblastine, palcitaxel.

Phase nonspecific agents: Alkylating agents, nitrosoureas, antitumour drugs, procarbazine, cisplatin, dacarbazine.

Mode of Action

These drugs damage the active cells by affecting the process of cell division. Hence, they also affect haemopoeisis, cellular activity, epithelial tissues and gonads. They also suppress the immune system.

Adverse Effects in General

Bone marrow suppression, alopecia, hepatotoxicity, nephrotoxicity, damage to gonads, damage to GI mucosa and ulceration.

HORMONE THERAPY IN CANCER

Ablative Procedures

- Oophorectomy and adrenalectomy in carcinoma breast.
- Orchidectomy in carcinoma prostate.

Added Hormone

- Prednisolone, progestogens, estrogens and androgens.

Hormone Antagonists

- Tamoxifen is estrogen receptor antagonist. Estrogen receptor level is assayed. If it is more than 10 units/gm of tissue, it is called as ER positive and if less than 10 units, it is called as ER negative.
- Cyproterone acetate, competes with testosterone for binding receptors.

The miserable have no other medicine but only hope.—**William Shakespheare**

	Drugs	Used in	Adverse effects
1.	Mustine hydrochloride	Hodgkin's lymphoma	Alopecia, bone marrow suppression
2.	Cyclophosphamide: Dose: 3 mg/kg IV daily for 5 days in dextrose, as monthly cycle for 6 months Orally as 50 mg tablets.	Ovarian carcinoma, lymphomas, colonic and bronchogenic carcinoma	Alopecia, bone marrow suppression, haemorrhagic cystitis
3.	Melphalan Dose:10 mg daily × 3 weeks. Available as 2 mg tablets	Multiple myeloma and melanoma	
4.	Busulfan: 10 mg for 3 weeks	Chronic myeloid leukaemia	Pulmonary fibrosis
5.	Chlorambucil: 10 mg for 3 weeks	Chronic lymphatic leukaemia	
6.	Methotrexate: 2.5-5 mg/d orally;10 mg intrathecally 1-2 times a week	Choriocarcinoma, ALL, soft tissue sarcoma, breast cancer	Oral and GIT ulceration, bone marrow suppression, hepatic damage
7.	6-Mercaptopurine Dose: 2.5 mg/kg/d orally	Leukaemias and choriocarcinoma	Hyperuricaemia
8.	5-Fluorouracil Dose:15 mg/kg/d IV	Adenocarcinomas of GIT, breast cancers and cancer cervix	Neurotoxicity, stomatitis, bone marrow suppression
9.	Vinca Alkaloids. Vincristine: Dose:1.5 mg/sq.m Vinblastine: Dose: 0.1 mg/kg body weight	Leukaemia Hodgkin's lymphoma	Neurotoxic Bone marrow toxicity
10.	Rubidomycin: Dose: 40 mg/m/ day IV	Acute myeloblastic leukaemia	Myocardial depressant
11.	Adriamycin: 60 mg/m^2 IV	NHL, hepatoma, medullary carcinoma thyroid, osteosarcoma and soft-tissue sarcomas	Cardiotoxic
12.	0'-p DDD(Mitotane): Dose: 10 gm orally for 8 weeks	Carcinoma adrenal cortex	
13.	Bleomycin: 20 units IV or IM	Squamous cell carcinoma of skin and other regions and lymphomas	Pulmonary fibrosis
14.	Cytosine arabinoside Dose: 4 mg/kg IV	Leukaemias and lymphomas	Leukopenia and thrombocytopenia
15.	Lomustine (CCNU) and Carmustine (BCNU)	Brain tumours and lymphomas	
16.	Procarbazine	Hodgkin's lymphoma and Oat cell carcinoma lung	Bone marrow and CNS depression
17.	Cisplatin: Dose: 20 mg/m^2/d	Testicular and ovarian tumours	Nephrotoxicity
18.	Etoposide: Dose: 100 mg/sq.m	Testicular tumour, bladder tumour, lymphomas	
19.	Hydroxyureas	Myeloma, leukaemias	

- LHRH analogue—Goserelin (Zoladex) acts as an antagonist to carcinoma breast.
- Phosphorylated diethylstilbestrol (Honavan) is used in carcinoma prostate.

Drugs that Interfere with Hormone Synthesis or Release

- Aminoglutethimide is an adrenal inhibitor. It is also an aromatase inhibitor inhibiting the peripheral conversion of androstenedione to estrogen.
- *Papillary carcinoma of thyroid is (TSH) hormone dependent.* TSH can be very well-suppressed by giving L-thyroxine, suppressive dose, daily 0.3 mg for life long.

Steroids (corticosteroids) are used as component of chemotherapy regime. They also improve the anorexia and also control hypercalcaemia seen in patients with malignancies on treatment. They reduce cerebral oedema in intracranial malignant neoplasms.

IMMUNOSUPPRESSION

- It is mainly used in transplantation to prevent graft rejection, e.g. in transplantation of kidney, liver, heart, small bowel, pancreas.

Drugs used are:

- *Azathioprine* (Imuran, Purine analogue)—Inhibits purine synthesis.
- *Cyclosporine:* It is a very good immunosuppressant, derived from fungus. It causes inhibition of lymphocytic activity, delayed hypersensitivity, interleukins, memory cells, etc. It binds to cyclophilin, decreases IL-2 and other cytokine release from T cells. It is initially given intravenously and later given orally. It is given usually for long duration, for 12-24 months. It is *nephrotoxic* as well as *bone marrow suppressant*. So, constant monitoring by blood urea, serum creatinine, blood count, Hb% at regular intervals is required. Dose should be adjusted depending on these parameters. Cyclosporine is very effective immunosuppressant in transplantation.
- Antilymphocytic globulin—Antibodies against T-cell receptors (CD_2, CD_3, CD_4, CD_8), B-cells and macrophages.
- Steroids (methylprednisolone)—Alters transcription and translation, affects T-cells and macrophages.
- Cytosine arabinoside—Anti-inflammatory and immunosuppressive effect.
- Tacrolimus (FK506)—Bind to FK binding proteins, effects are similar to cyclosporine.
- OKT3—Acts against CD3 and T-cells.
- Mycophenolate mofetil—Inhibits inosine monosphosphate, affects lymphocytes.

Types of Immunosuppression

1. *Induction regimens*—Mainly aims to avoid rejection and to establish a good graft function during the immediate post-transplant weeks (2 weeks). Antilymphocytic sera, along with either cyclosporine or azathioprine or steroids are used here.
2. *Antirejection regimens*—It aims at reversing the acute rejections. High dose steroids and antilymphocytic sera are used here.
3. *Maintenance therapy*—Using cyclosporine or FK506 along with steroids and azathioprine.

Complications

- Infection—commonest
- Malignancy—squamous cell carcinoma, lymphoma, Kaposi's sarcoma
- Nephrotoxicity
- Bone marrow toxicity
- GIT toxicity
- Neurotoxicity

IMMUNOTHERAPY

It is enhancing the host response of patient whenever required, commonly against malignant cells. Often it is used in other conditions also, where there is severe immunosuppression due to any reason like bone marrow suppression or in specific conditions like severe hepatitis.

Immunotherapy can be:

- Active
- Passive
- Restorative
- Adaptive

Immunotherapy agents

- Monoclonal antibodies.
- Bone marrow transplantation.
- Cell transfer.
- Lymphokines.
- Thymic hormones.
- Levamisole and BCG (as immunomodulators).
- Prostaglandins.
- Interferons, interleukins.
- Immunoglobulins.
- Antibody derived specific to certain tumour like melanoma.

Tumours where immunotherapy is used

- Melanoma
- Bladder tumour
- Carcinoma colon
- Renal cell carcinoma
- In many tumours it is under trial

Disadvantages

- Very costly.
- Nonavailability.
- Effect is not 100%.

HYBRIDOMA

- It is a biotechnological process wherein *multiplication* property of *Myeloma cells* is combined with *synthetic* property of some other required cells, *to achieve rapid* and large quantity manufacturing of required chemical.
- Myeloma cells with only retained multiplication activity is fused with human cells or lymphocytes or other cells (of nature to produce the required product) by ***hybridisation***. The resulting cell has got the capacity to multiply rapidly and to produce required product in large quantity. It is called as ***hybridoma***.
- It is used to generate monoclonal antibodies, insulin and many other antibodies, immunoglobulins, etc.

Life so short, the craft so long to learn.—**Hippocrates**

Monoclonal antibodies are used for
- Immunodiagnosis—for radioimmunoassay, radionuclide scan
- Antibody for detection of tumour antigen
- For cancer therapy
- For serotherapy
- As conjugates (with drugs, toxins, isotopes)
- For production of other chemicals and as a *research tool*

GENE THERAPY

The ability to alter specific genes of interest is, nowadays an exciting and powerful tool in the potential management of a wide range of diseases. Instead of giving a patient a drug to treat or control the symptoms of the genetic disorder, physicians may be capable of treating the basic problem by altering the genetic makeup of the patient's cell.

Typically two methods have been considered: Germ line and somatic cell gene therapy.

- *Germ cell therapy* involves insertion of a gene into fertilised egg for the correction of a genetic disease. Because these genes are dispersed throughout the tissues of the egg, they end up in the germ cells of the foetus, and hence are passed on to the future generations.
- *Somatic cell therapy* involves the insertion of genes or otherwise manipulating the gene machinery of a cell to treat a disease. In this case the cells are restricted to the population that has been treated and any genetic change remains restricted to these cells and is not passed onto the germ cell line.

The goals of human somatic therapy are usually one of the following:

- To repair or compensate for a defective gene.
- To enhance the immune response directed at a tumour or pathogen.
- To protect vulnerable cell populations against treatments such as chemotherapy.
- To kill tumour cells directly.

Several single gene disorders are candidates for gene therapy and in addition, current thinking has expanded to include treatment of AIDS and atherosclersis using gene therapy techniques.

Vectors used for gene therapy fall into ***two main classes***—*Viral and nonviral.*

- **Viral:** Initially retroviruses were used as vectors. Other potential vectors include Adenovirus, Herpes virus and Vaccinia virus.
- **Non-viral systems**
    - Liposome mediated DNA transfer.
    - DNA protein conjugates.

However exciting and appealing the prospects of gene therapy may appear, this technique is still in the experimental stages.

Chapter 30 Anaesthesia

> *It has long been an important problem in medical science to devise some method of mitigating the pain of surgical operations. An efficient agent for this purpose has at length been discovered. A patient has been rendered completely insensible during an amputation of the thigh, regaining consciousness after a short interval.*
>
> —**Henry Jacob Bigelow,** 1846

CHAPTER OUTLINE

- Preoperative Assessment
- General Anaesthesia
  - Anaesthetic Agents
  - Oxygen
  - Muscle Relaxants
  - Reversal Agents
  - Instruments in Anaesthesia
  - Complications of General Anaesthesia
  - Postoperative Care
  - Monitoring the Postoperative Patient
- Regional Anaesthesia
  - Topical Anaesthesia
  - Infiltration Block
  - Field Block
  - Nerve Block
  - Intravenous Regional Anaesthesia
- Spinal Anaesthesia
  - Saddle Block
- Epidural Anaesthesia
  - Caudal Anaesthesia

PREOPERATIVE ASSESSMENT

History

1. Chronic cough, smoking, alcohol, drug intake, drug allergy.
2. Any previous diseases like hypertension, diabetes mellitus, epilepsy, bronchial asthma, tuberculosis, hepatitis, cardiac diseases.
3. Drug therapy: Steroids, antihypertensives, sedatives, antibiotics, antiepileptics.

Examination

General: Posture, teeth, mouth opening, dilated veins, neck movements, tremor, airway.
Anaemia, oedema, jaundice, cyanosis.

Respiratory system: To look for asthma, tuberculosis, emphysema, COPD.

Airway: Mouth opening, Mallampati scoring, thyromental distance, temporomandibular joint assessment.

Thyromental distance: It is the distance between mentum and thyroid cartilage, measured externally. If it is more than 6.5 cm (i.e. more than 4 fingers breadth) intubation is easier, if it is less than 6.5 cm intubation is difficult.

Cardiovascular system: Hypertension, ischaemic heart disease, arrhythmias, cardiac failure, valvular diseases.

Spine: Curvature, intervertebral space, skin over the area for any infection.

Other systems: Abdomen, skeletal system.

Scoring to assess intubation

Mallampati scoring:

Class I	— Faucial pillar, soft palate, uvula are seen
Class II	— Faucial pillar, soft palate are seen
Class III	— Only soft palate is seen

Samsons Young modification:

Class IV	— Only hard palate seen

Class I, II: Easy intubation
Class III, IV: Difficult intubation

Preoperative Investigations

Haematocrit, blood sugar, blood urea, serum creatinine, electrolytes, chest-X ray, ECG, blood grouping, blood-gas analysis, cardiac assessment.

> *When giving mouth-to-mouth resuscitation, the important thing to remember is to remove your denture first.*
>
> **—WU McClenahan**

Preoperative Treatment

- Control of respiratory and cardiac diseases.
- Improvement of Hb% status, if anaemia is present.
- Preoperative antibiotics are given.
- Blood should be kept ready for major cases.
- *Starvation for 4 hours for liquids and six hours for solids.*
- Bladder and bowel should be emptied to prevent soiling on the operation table. Urinary catheter may be passed and enema may be given.
- Dentures, contact lenses, jewellery must be removed.
- Surgical area should be cleaned and properly prepared.

GENERAL ANAESTHESIA

It means abolition of all sensations, i.e. touch, pain, posture and temperature with a state of reversible loss of consciousness.

It has got three components:

(*1) Analgesia. (2) Hypnosis. (3) Muscle relaxation.*

ANAESTHETIC AGENTS

- ***Volatile anaesthetics*****:** They vaporise in room air.
  - Agents used are: Ether, trichloroethylene, halothane, enflurane, isoflurane, sevoflurane.
  - *Ether* which is irritant, unpleasant, flammable, is commonly used agent in developing countries.
  - Enflurane and isoflurane are non-inflammable, non-explosive, non-irritant and stable. Here anaesthesia is rapid with faster recovery.
- ***Gaseous anaesthetics*****:**
  - *Nitrous oxide*: It is non-inflammable, non-irritant, good analgesic but weak anaesthetic agent. It is given along with 30-50% oxygen for balanced anaesthesia (blue coloured cylinder in India).
  - Cyclopropane is highly flammable.
- ***Intravenous anaesthetics*****:**
  - *Thiopentone*: It is ultrashort acting barbiturate which causes hypnosis during induction of anaesthesia. It does not have analgesic effect. It causes hypotension, respiratory depression, laryngeal and bronchospasm. Recovery is rapid. Extravasation of drug can cause skin ulceration. Intra-arterial injection causes vasospasm and gangrene. Dose: 4-7 mg/kg.
    Methohexitone sodium.
    Propanidid. 4-7 mg/kg. It can cause anaphylaxis.
  - *Ketamine:* Dose: 2 mg/kg IV. It is a good analgesic. It causes dissociative anaesthesia. It can lead to hypertension, apnoea, laryngospasm. In children it can be given IM-5 mg/kg. It does not require intubation for small procedures.
  - *Propofol***:** It is widely used induction agent which has got predictable onset and recovery. It has got least side effects on CVS and respiratory system. It is also used for total IV anaesthesia. Dose: 1-2.5 mg/kg.
  - Fentanyl is neuroleptanalgesic. It causes sedation, catatonia, dissociation, hypotension and preferred in asthmatics.

OXYGEN

- Oxygen is given through Boyles apparatus (33.3%).
- Oxygen in high concentration is respiratory depressant and also affects eyes.
- A 5% CO_2 mixture in oxygen is called as carbogen.
- Oxygen is available in black and white coloured cylinder.

MUSCLE RELAXANTS

Depolarising Muscle Relaxants

They act at the level of acetylcholine receptors which widens the refractory period after depolarisation causing paralysis. It is short acting muscle relaxant.

a. *Suxamethonium chloride (scoline*):
  - It lasts for 2-4 minutes.
  - It causes muscle twitching—fasciculations—paralysis.
  - It is metabolised by plasma pseudocholinesterase. Atypical or deficiency of this enzyme prolongs the action of the scoline.
  - *Side effects* are hyperkalaemia, myotonia, apnoea and cardiac arrest.

b. *Suxthonium bromide.*

Non-Depolarising Muscle Relaxants

They block the channels of entry of acetylcholine. They are long acting relaxants.

1. *Tubocurarine*: It lasts for 45 minutes. 30 mg is the dose.
2. *Gallamine. (Flaxedil*): Dose is 1-2 mg/kg. It is cheaper. It is contraindicated in renal diseases.
3. *Pancuronium bromide (Pavulon*): It is synthetic steroid muscle relaxant. It’s action lasts for 45 minutes. Dose is 0.08-0.1 mg/kg.
4. *Vecuronium bromide*: It is a steroid muscle relaxant, given at a dose of 0.05-0.1 mg/kg.
5. *Rocuronium is* short acting steroid muscle relaxant. It starts its action in one minute.
6. *Atracurium*: It lasts for 20-30 minutes. Dose is 0.6 mg/kg.
7. *Mivacurium*: Dose is 0.15-0.25 mg/kg.

REVERSAL AGENTS

They are anticholinesterase drugs which increase the acetylcholine and thus act as antagonising agents for non-depolarising muscle relaxants. They cause bradycardia.

Neostigmine (2.5 mg) is used commonly along with atropine (1.2 mg).

Edrophonium (short acting) and pyridostigmine (long acting) are other drugs.

INSTRUMENTS IN ANAESTHESIA

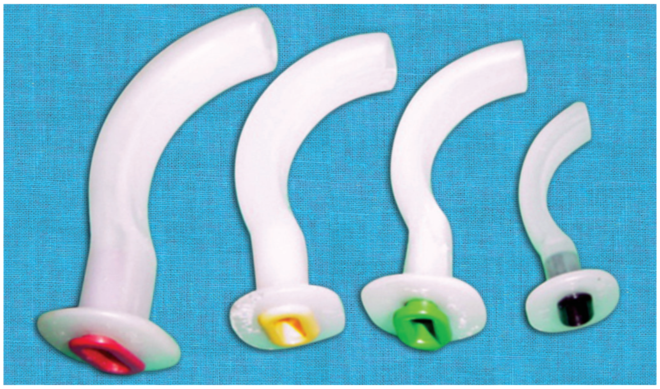

Fig. 30.1: Different sized airways have different colours.

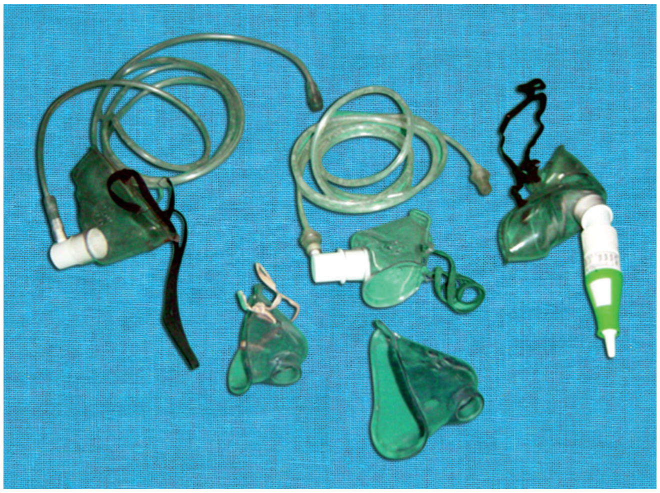

Fig. 30.2: Different sized masks used during anaesthesia.

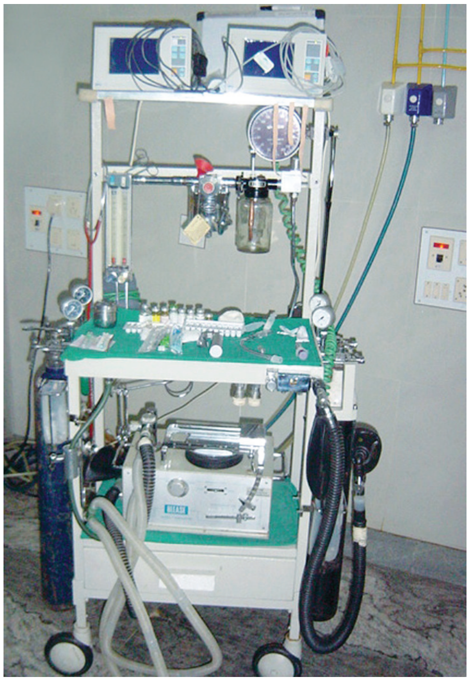

Fig. 30.3: Boyle's apparatus used for anaesthesia.

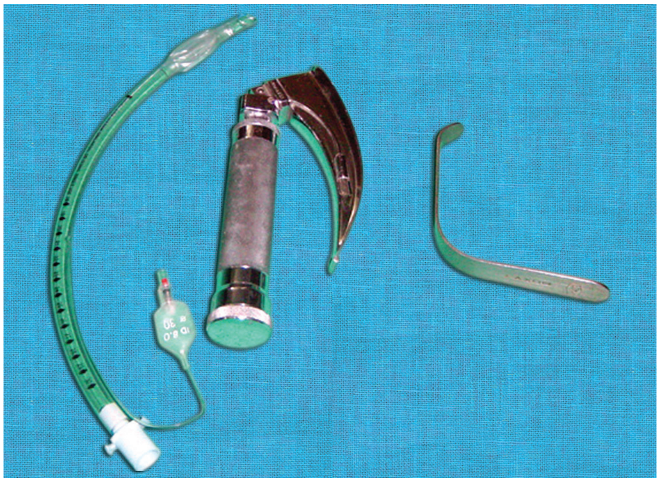

Fig. 30.4: Direct laryngoscope with endotracheal tube and tongue depressor.

1. ***Boyle's apparatus:***
   It consists of:
   a. Cylinders for N_2O and O_2.
   b. Pressure gauge—to know the amount of gas remaining.
   c. Pressure regulator—to regulate the pressure of gas used.
   d. Rotameter—to know the flow of gas.
   e Vaporiser.

Note:

Pin index code: It is safety mechanism to fix gas cylinder onto the Boyle's machine. Different gases have different pin index codes, so that interchanging of cylinders does not take place.
Pin index code for N_2O is 3, 5.
Pin index code for O_2 is 2, 5.
Pin index code for air is 1, 5.

2. ***Endotracheal tube*****:** These are tubes inserted into the trachea and is used to conduct gases and vapours to and from the lungs. Depending on the diameter, it is available in various sizes. It has a cuff at one end which, when inflated stabilises the tube in position and also prevents regurgitation. Non-cuffed tubes are also available. The other end near the mouth is connected to the breathing circuit through which anaesthetic gases are delivered. The tube is inserted using a direct laryngoscope. The proper placement in the airway is confirmed by auscultating for the breath sounds over the chest when the gases are delivered.

Complication

- Postoperative sore throat
- Hoarseness after intubation
- Upper airway oedema

3. Magill's forceps.
4. Mouth gag.

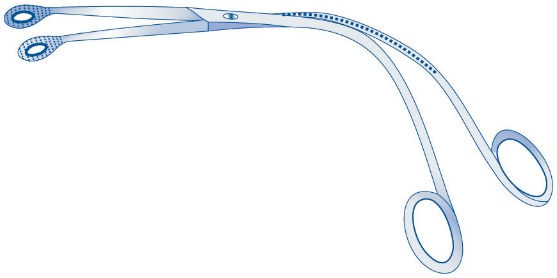

Fig. 30.5: Magill's forceps.

5. Laryngoscope.
6. Connectors.
7. ***Laryngeal mask airway (LMA):***
   - Laryngeal mask consists of a wide bore tube whose proximal end connects to breathing circuit and distal end is attached to an elliptical cuff which can be inflated.
   - These come in various sizes for various age groups. It is made up of silicone rubber.
   - The deflated cuff is lubricated and inserted into the hypopharynx, so that once inflated, the cuff forms a low pressure seal around the entrance into the larynx.

The future belongs to those who believe in the beauty of their dreams.

Advantages

- Does not require laryngoscope for insertion
- Does not cause irritation of airway
- Less incidence of laryngospasm
- In case of inability to intubate can save the life of the patient

Contraindications

- Pharyngeal pathology
- Full stomach as it can cause aspiration
- Bronchospasm

Components of general anaesthesia

- Premedication
- Induction
- Maintenance
- Recovery

Premedication

It is given one hour before surgery:

- *For sedation and relief of anxiety*. Pethidine 50 mg/morphine 10 mg/diazepam 10 mg, midazolam 1-2.5 mg.
- To suppress vagal activity. Atropine 0.6 mg IM.
- *To reduce vomiting*. Promethazine (phenargan) 12.5 mg.

Induction: Patient is preoxygenated with 100% oxygen for 3 minutes then induced with IV thiopentone, given 4-5 mg/kg. Patient *loses consciousness*. Induction is maintained by 67% nitrous oxide and 33% oxygen.

Scoline is given IV to relax muscles so as to facilitate *endotracheal intubation*.

Once intubated, *ventilation* can be either *controlled* using muscle relaxants or *spontaneous* using a volatile anaesthetic agent.

Reversal is done using neostigmine and atropine or glycopyrrolate.

COMPLICATIONS OF GENERAL ANAESTHESIA

- Intra-arterial injection of the drug.
- Myocardial depression and cardiac arrest.
- Hypertension.
- Laryngeal and bronchial spasm.
- Cardiac arrhythmias.
- Respiratory failure.
- ARDS.
- ***Mendelson's syndrome***: It is due to regurgitation of the acid from the stomach causing aspiration of acid leading into bronchospasm, pulmonary oedema and circulatory failure. This is treated with oxygen, suction, hydrocortisone, aminophylline, antibiotics, Ryle's tube aspiration and ventilator support.
- Hypoxia.
- Pneumothorax.
- Anaphylaxis.
- ***Malignant hyperthermia***: It is an inherited myopathic disorder occurs under anaesthesia due to drugs like halothane, scoline. There is marked increase in metabolic rate, with rise of temperature. There is high levels of CPK enzyme. Condition will cause metabolic acidosis and hyperkalaemia. It has got high mortality. Treatment is IV dantrolene, cooling, oxygen and cold IV fluids.
- Hypothermia.

POSTOPERATIVE CARE

Immediate postoperative period is important and critical, because patient may not be fully conscious. Patient should be kept in recovery room until he/she recovers from anaesthesia.

1. *Care of respiratory system*: Adequate breathing is important, otherwise hypoxia sets in, which gradually leads to cardiac arrest.

Respiratory problems may be:

- Laryngeal spasm
- Falling of tongue backwards blocking the airway
- Aspiration
- Bronchospasm
- ARDS
- Respiratory failure

Oxygen supplement through mask, observation, proper positioning are the treatment.

2. Hypercarbia.
3. Circulatory problems:

- Hypotension
- Arrhythmias
- Hypertension
- Cardiac arrest

4. GIT:

- Vomiting
- Regurgitation
- Mendelson's syndrome

5. *Renal problems:* Oliguria, i.e urine output is less than 30 ml/hour. It may be due hypovolaemia, hypotension, acidosis, sepsis, transfusion reaction, toxins.

   The ratio of urine/plasma osmolality of 2:1 signifies prerenal failure. Ratio of 1.7:1 indicates renal failure.

   Blood urea and serum creatinine is done at regular intervals.

   Fluid and electrolyte imbalance, if any is corrected.

   100 ml 20 % mannitol or frusemide 40-80 mg are often required.
6. Other problems:

   Restlessness, shivering, pain.

MONITORING THE POSTOPERATIVE PATIENT

- Pulse, temperature, BP chart.
- Breathing type.
- Level of consciousness.
- Urine output.
- Oxygen saturation and heart rate using pulse oximeter.
- Checking and encouraging limb movements.
- Skin colour, tongue colour for adequacy of oxygenation.
- Tongue for hydration.

- Cardiac monitor.
- Blood gas analysis in case of patient on ventilator.
- Serum electrolytes assessment.

REGIONAL ANAESTHESIA

Carl Koller, an ophthalmologist introduced cocaine as local anaesthetic in ophthalmic practice.
Mode of action: It causes temporary conduction block of the nerve, thus preventing the propagation of nerve impulse.

Advantages of local anaesthetic agent:

- Technically simpler.
- General anaesthesia is avoided.
- Consciousness is retained.
- Patient can have food earlier after surgery.

Drugs used:

- Cocaine, procaine, cinchocaine—amino esters.
- Lignocaine, prilocaine, bupivacaine, ropivacaine—amino amides.

Lignocaine/lidocaine/xylocaine: It is the commonest local anaesthetic agent used. It is available as 0.25-5% concentrations.

It is metabolised in the liver and excreted in the kidney as xylidines. It is also an antiarrhythmic drug and so commonly used in cardiology and cardiac surgery.
Side effects: Giddiness, headache, postural hypotension, tinnitus, circumoral anaesthesia.
Dose: 4 mg/kg effect lasts for 90 minutes.

Uses

- Topical-4%.
- Infiltration block: 0.25%.
- Field block 0.5%.
- Nerve block 1.0 %.
- Epidural 1.5, 2.0%.
- Spinal 5%.

It can be used with or without adrenaline.

Xylocaine with adrenaline has got longer duration of action. It creates relatively bloodless field.

But it should not be used in places where end arteries are present like glans penis, ear lobule, tip of the nose, lip, fingers and toes.

Bupivacaine (Marcaine): It has got prolonged action. It is a vasodilator also.

Dose: 3 mg/kg.
Epidural block: 0.5%
Spinal 0.5% 3 ml.

TOPICAL ANAESTHESIA

- It is used for minor surgeries of eye, laryngoscopy, bronchoscopy, cystoscopy, gastroscopy.
- It is available as instillation, spray, viscous, ointment, gel, EMLA (Eutectic mixture of local anaesthetic).

INFILTRATION BLOCK

Direct injection of local anaesthetic under the skin for small procedures.

FIELD BLOCK

It is achieved by blocking the entire field of excision where lesion is located.

NERVE BLOCK

- Block of inferior dental nerve and lingual nerves in the region of the mental foramen for extraction of teeth.
- Finger block of digital nerves. Here plain xylocaine is used (without adrenaline).
- Intercostal block.
- Ankle block.
- Median and ulnar nerve block.
- Brachial plexus block *(Winnie's block).*
  It can be given through:
  - Interscalene,
  - Axillary,
  - Supraclavicular approaches.

Supraclavicular approach is commonly used. 1 cm above the mid-point of the clavicle, needle is passed downwards, backwards and medially towards first rib. Once needle hits the first rib, 15-20 ml of 1.5% xylocaine is injected (with walking or stepping over the first rib). Complications are pneumothorax and injury to the great vessels.

Other blocks:

Cervical plexus block.
Sciatic nerve block.
Femoral nerve block.

INTRAVENOUS REGIONAL ANAESTHESIA (BIER'S BLOCK)

Limb is exsanguinated and occluded with tourniquet. Pressure in the tourniquet must be 30 mmHg more than the systolic pressure of the patient. Needle is placed in the selected vein. 40 ml of 0.5% xylocaine for upper limb and 80 ml of 0.25% of xylocaine for lower limb is injected into the vein. *Xylocaine with adrenaline should not be used.* It gives very good analgesia for 2 hours.
Side effects: Sudden release of drug into the circulation can cause hypotension, convulsions and often death.

Bupivacaine should not be used.

Indications

For upper and lower limb surgeries, it can be used without G/A or S/A.

SPINAL ANAESTHESIA

It is the injection of local anaesthetic into the subarachnoid space causing loss of sympathetic tone, sensation and motor function. The sympathetic block is 3 segments higher than sensory block, motor block is 3 segments lower than sensory block.

Position: Lateral decubitus position with head, hips and knees being fully flexed so as to open the inter-laminar spaces. Highest point of iliac crest corresponds to 4th lumbar vertebra.

Drugs used:

Lignocaine 5% in 6% dextrose, 2 ml.
Bupivacaine 0.5% in 5% dextrose, 3 ml.
Cinchocaine 0.5% in 6% dextrose, 2 ml.

Technique: 24-26 gauge needle with stillette is used. Needle is passed through the interspinous space and ligamentum flavum to reach the subarachnoid space to get clear fluid (0.5 ml/ sec). Needle is rotated 360 degrees and drug is injected slowly. Patient is repositioned to supine. Drug takes 15 minutes to act.

Types

1. Caudal (up to L_5)
2. Low spinal (up to L_1)
3. Mid-spinal (up to T_{10})
4. High spinal (up to T_6)
5. Unilateral spinal

Advantages

- Economical
- Hypotension reduces the bleeding
- Adequate relaxation is achieved
- Respiratory complications are less

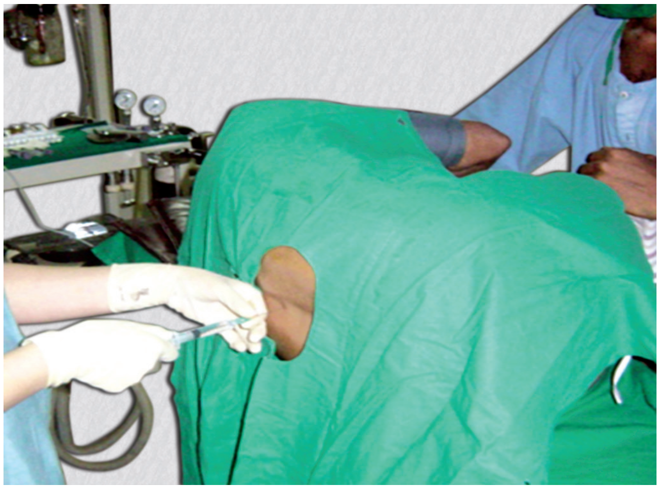

Fig. 30.6: Spinal anaesthesia being given. Note the position of the patient.

Disadvantages and Complications

- CSF leak and aseptic inflammation of meninges causing headache.
- Meningism.
- Infection.
- Paraplegia. It is very rare.
- Occasionally it can become total spinal which requires intubation and ventilator support.

Contraindications

- Cardiac patient
- Allergy
- Increased intracranial pressure. It may precipitate coning
- Sepsis
- Spinal tumours
- Back pain and spinal diseases
- Neurological conditions like syringomyelia
- Kyphosis, scoliosis

SADDLE BLOCK

- It is used for surgeries in perineal and anorectal region.
- It is spinal anaesthesia using xylocaine or bupivacaine given in sitting position.

EPIDURAL ANAESTHESIA

- It is a potential space between dura anteriorly and ligamentum flavum posteriorly which has got negative pressure inside. It extends from foramen magnum to sacral hiatus.
- ***Touhy needle*** is used for epidural anaesthesia. Once the needle is in the space there will be sudden indrawing of air or saline drop.
- An epidural catheter is placed in the space and fixed. 2% xylocaine with adrenaline or 0.5% bupivacaine is injected into the space to achieve anaesthesia up to the desired level.

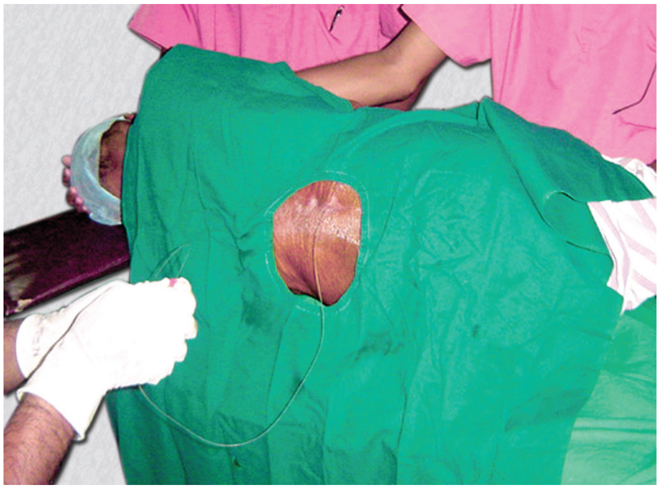

Fig. 30.7: Epidural anaesthesia with epidural catheter placed.

Advantages

- It can be used for continuous repeated prolonged anaesthesia.
- It can be used for postoperative analgesia.
- It can be kept for several days.

CAUDAL ANAESTHESIA

Caudal space is the sacral component of epidural space and access is through the sacral hiatus.

Indications

- Haemorrhoidal surgery
- Circumcision
- Small procedures in the perineum like cystoscopy

Procedure

It is given in lateral position. Needle is inserted through the sacral hiatus to enter the caudal epidural space. Drug is then injected into the space.

Complications

- Trauma to anal canal
- Intravascular injection
- Failure of caudal block

We should have a good heart to pity; a good hand to bless.

Chapter 31 Advanced Imaging Methods

> *What would it be like in a radiologist's shoes? To spend most of my day dealing with images of people: plain black-and-white X-ray images, or speckled images caused by sound waves bouncing off organs, or images caused by dyes outlining arteries and veins, or contrast medium filling loops of bowel, or images reconstructed by computers into cross sections of the body-all without speaking to a patient.*
>
> —**Abraham Verghese,** 1994

CHAPTER OUTLINE

- Ultrasound
- Doppler
- CT Scan
- MRI
- Radionuclide Imaging
- PET Scan

ULTRASOUND

Ultrasound contains waves with a frequency of more than 20,000 cycles/second which the human ears cannot hear.

In medical sonography, frequencies used are commonly 2-10 MHz. The transducer or the probe works as both transmitter of ***sound waves*** and receiver of ***echoes.*** The piezoelectric crystal is the producer of ultrasound waves. Received signals from the patient are fed into the computer which forms the image.

There are three types of ultrasound image display.

1. **A-mode**: Only one dimensional static display as spikes are obtained. It is used only in eye scan.
2. **B-mode**: Two dimensional real time images in the form of grains. It is most widely used type. Using this mode *Transverse, Longitudinal or Oblique* sections can be taken.
3. **M-mode**: Here images are recorded as dots. It is mainly used in moving parts like echocardiography. M-mode is also called as TM mode, i.e. *Time Motion Mode*.

Uses

1. Used in all abdominal and pelvic conditions, often in thoracic conditions.
2. ***Ultrasound of thyroid*** is very useful method to differentiate between solid and cystic lesions.
3. Ultrasound is used in testicular tumours, epididymoorchitis, trauma to testis, erectile dysfunction, etc.
4. Ultrasound breast is used to differentiate solid from cystic tumours.
5. Soft tissue and musculoskeletal ultrasound.
6. Ocular ultrasound is ideal method to image eye and intraocular structures.

Advantages

- No radiation
- Noninvasive
- Effective with efficiency
- Painless
- Low cost
- Available even as portable machines
- Stones are well visualised with *acoustic shadow*

Disadvantages

- Interpretation can be inadequate
- Bowel shadow may prevent proper visualisation
- In obese patient image will be inadequate
- **Interpretation** is based on echogenicity either *hyperechogenic* or *hypoechogenic*

Advanced ultrasound techniques

- Endosonography (EUS) is used in visualisation of walls of oesophagus or stomach through gastroscopy
- Transvaginal ultrasound
- Transrectal ultrasound to see prostate
- Doppler ultrasound to study arterial and venous diseases

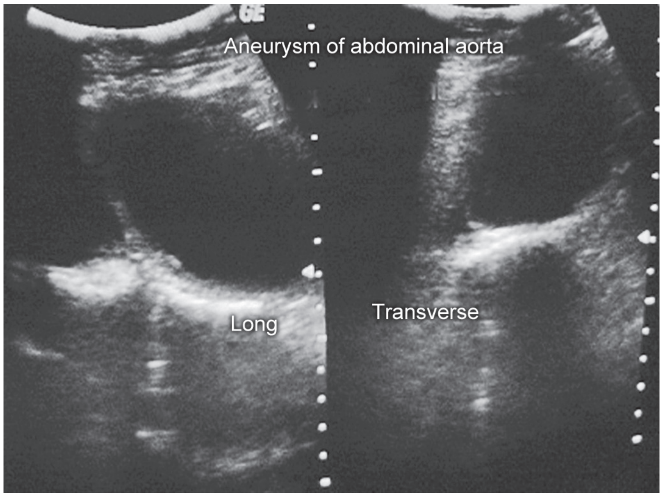
GE
Aneurysm of abdominal aorta
Long
Transverse

Fig. 31.1: US showing abdominal aortic aneurysm.

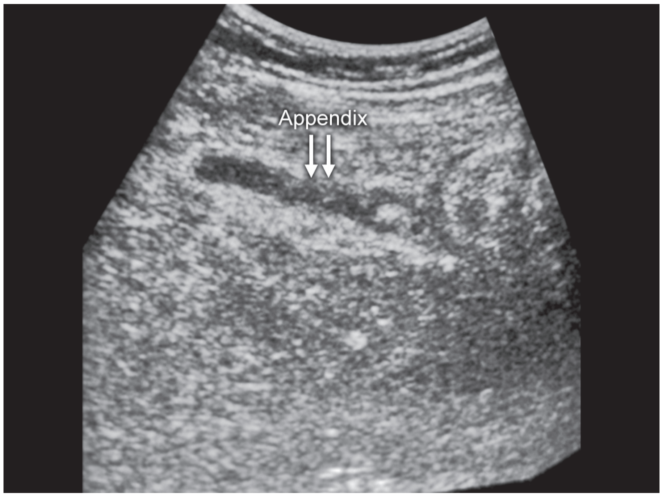

Fig. 31.2: US showing appendix—appendicitis.

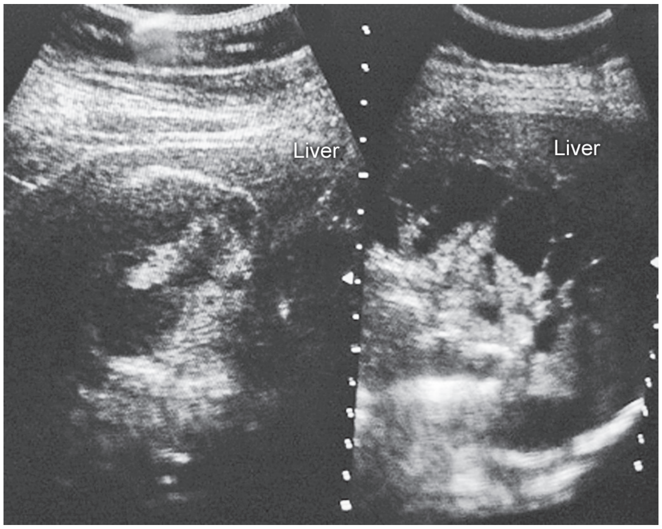
Liver
Liver

Fig. 31.3: US showing liver mass.

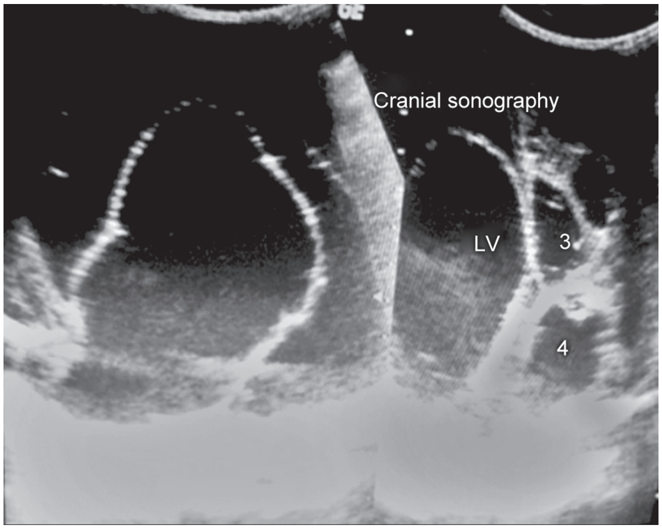
GE
Cranial sonography
LV
3
4

Fig. 31.4: Cranial sonography is done to see intracranial problems through suture lines—usually done in infants and children.

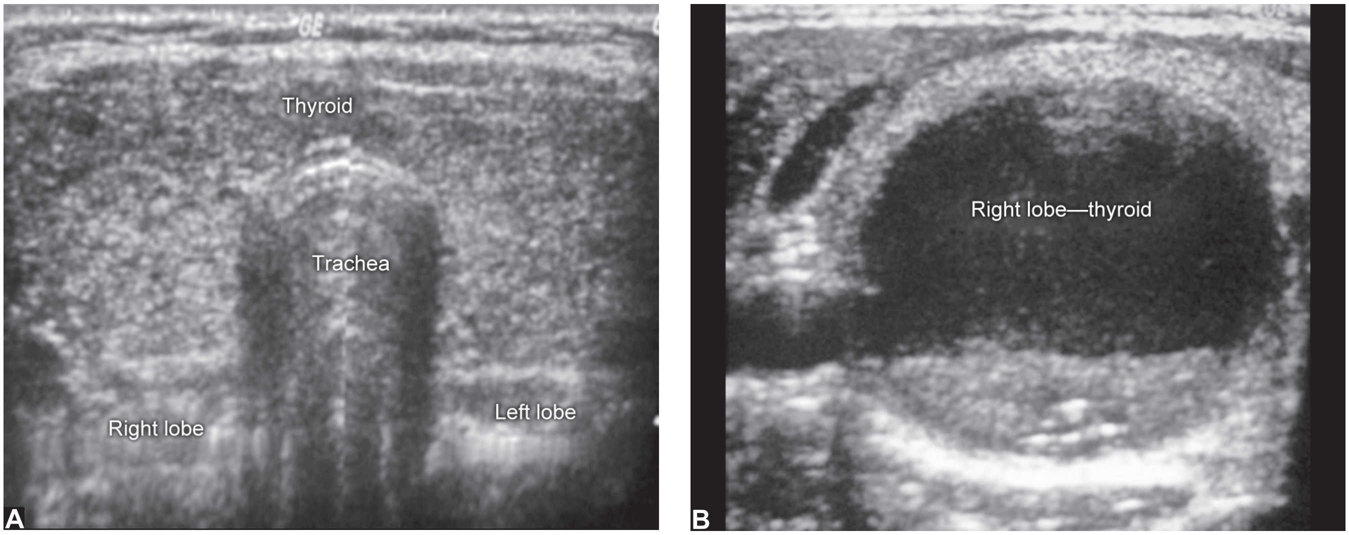
GE
Thyroid
Trachea
Right lobe
Left lobe
A
Right lobe—thyroid
B

Figs 31.5A and B: US thyroid showing thyroid both lobes in one; solid and cystic components in another.

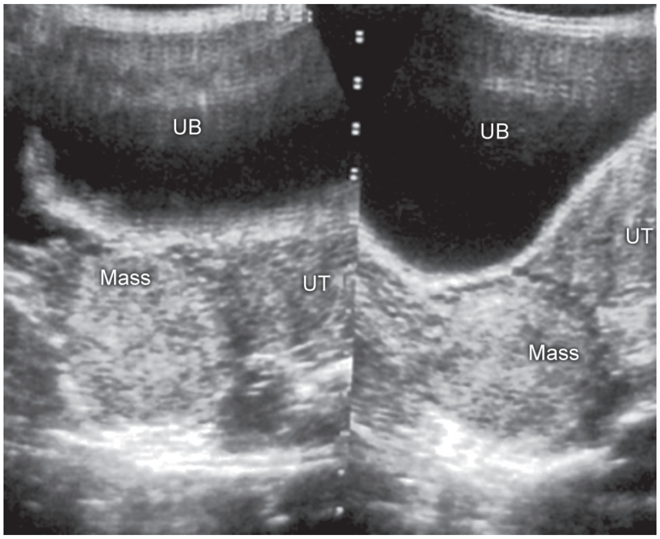

Fig. 31.6: US picture of a pelvic mass.

Ultrasound as therapeutic use

- To guide aspiration of amoebic liver abscess, pericardial tap
- On table ultrasound can be done to assess the operability of tumour. (During laparotomy to assess the extent of tumour, lymph node status, etc.)

DOPPLER

Doppler effect is a change in the perceived frequency of sound emitted by a ***moving source***. So it measures blood flow. Spectral Doppler wave form and ultrasound image are combined in **Duplex scanning**.

Types

1. Continuous waves.
2. Pulsed waves.

Doppler will provide both audio and video signals. ***Colour Doppler imaging displays*** flowing blood as ***red*** when direction of flow is towards the transducer. Image will be ***blue*** if flow is away from transducer.

Uses

- To study cardiovascular system.
- To study vascularity of tumours.
- To study blood flow and velocity in arterial diseases so as to assess stenosis (its extent, cause, etc.) like in atherosclerosis, TAO, cervical rib, aneurysm, A-V fistulas.
- To find out deep venous thrombosis (DVT), varicose veins, perforator incompetence.
- To study grade of varicocele in males.

Advantages

- It has replaced venogram and angiogram in many places as a diagnostic tool.
- It is reliable and non-invasive.

CT SCAN

Computerised tomography scan was invented by ***Godfrey Hounsfield*** in 1963. He was a physicist. He received **Nobel prize** (1972) for the same. The first CAT scan is in the London museum.

Narrow X-ray beams are passed from rotating X-ray generator through **the gantry** where patient is placed. When X-rays pass through the tissues, some of the X-rays get absorbed and some pass through, depending on the tissue density. The different grades of absorption in different tissues are detected through sensitive detectors which are translated to a Gray scale image by a computer.

- Density of tissues is numbered as **Hounsfield Number (HN)**
- Water—Zero HN
- Air—Minus 1,000
- Bone—Plus 1,000

The density of other tissues come in between air and bone with different HNs

Presently spiral CT scan has become popular. They are faster and in a single breath holding time, whole CT scan can be taken.

Both plain and contrast CTs are done whenever required.

Contrast Agents

- **Ionic**: Water soluble iodide dyes like *Sodium diatrizoate, Meglumine iothalamate (Conray, Urograffin, Angiograffin).* They are cheaper but often toxic and cause anaphylaxis.
- **Non-ionic are** safer but expensive, like *Iohexol (Omnipaque), Iopamiro.*

In abdominal CT, contrast agents can be given orally to delineate bowel properly.

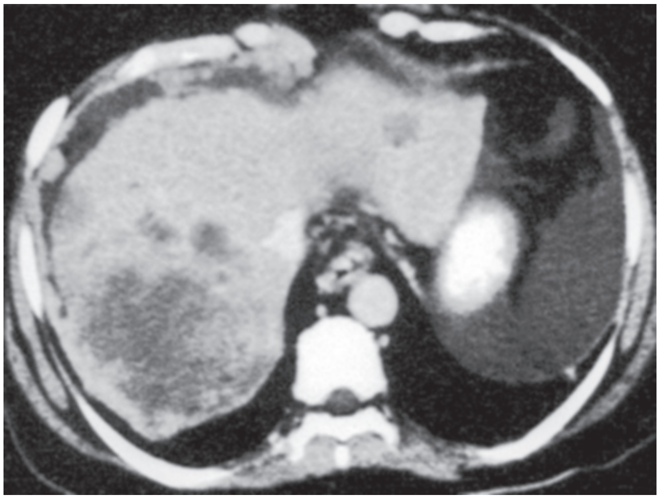

Fig. 31.7: CT scan picture showing liver secondaries.

Indications

- **Trauma** like head injury, chest injury, abdomen trauma. In trauma *only plain CT scan* is taken.
- **Neoplasms:** To see the exact location, size, vascularity, extent and operability.

For example, brain, abdominal, retroperitoneal, thoracic and spinal tumours.

- **Inflammatory conditions,** in various sites.
  For example, psoas abscess, pseudocyst of pancreas.

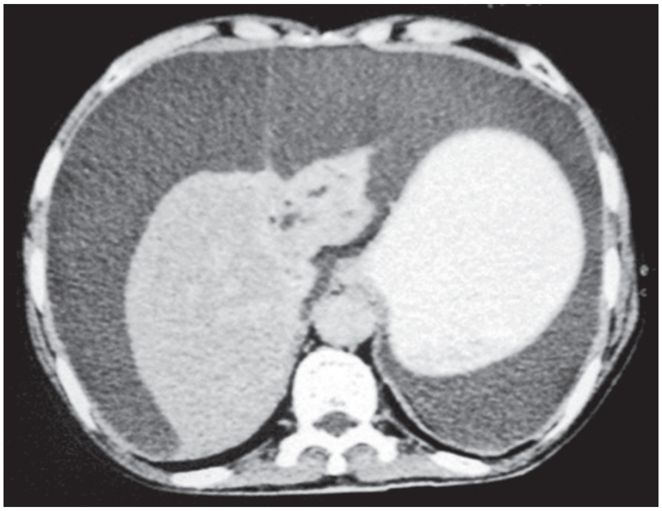

Fig. 31.8: CT scan showing ascites (gross).

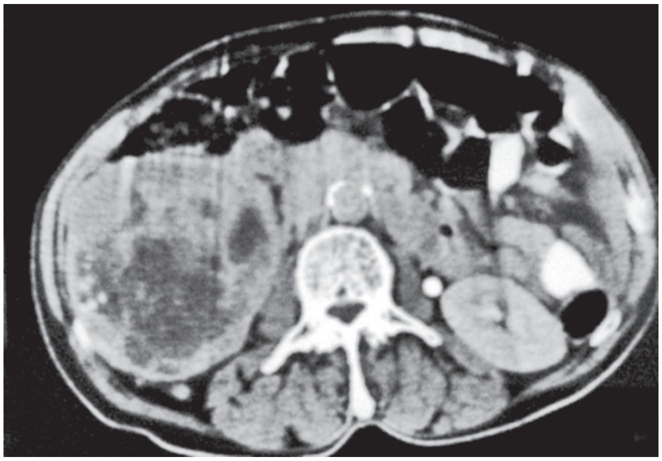

Fig. 31.9: CT scan abdomen showing right renal mass (RCC).

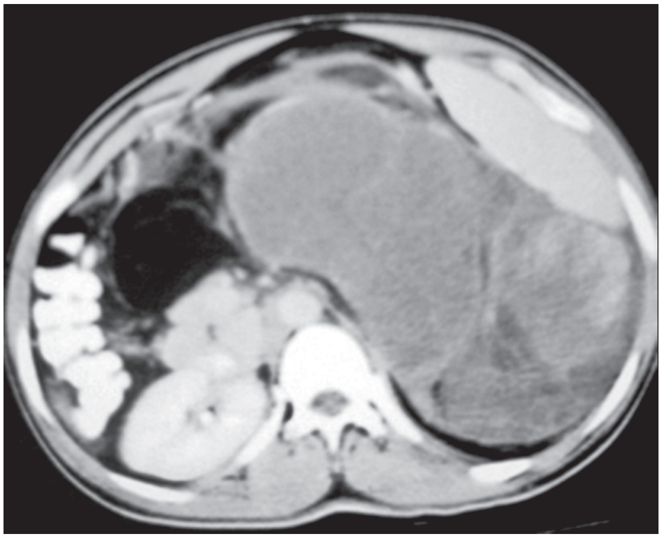

Fig. 31.10: CT scan showing retroperitoneal tumour.

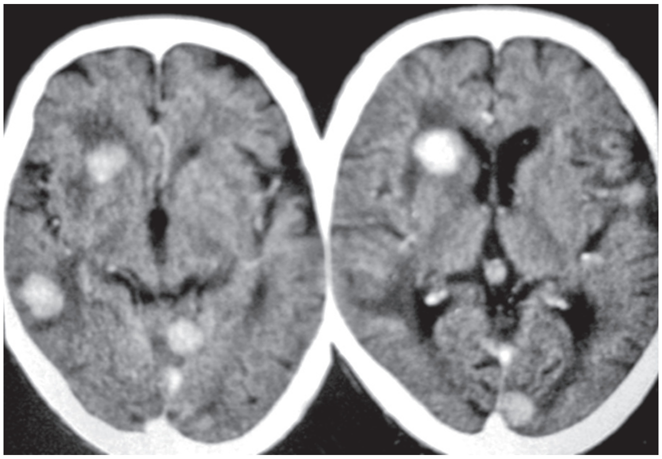

Fig. 31.11: CT scan showing brain secondaries. Secondaries are the commonest malignant tumours of brain. Breast is the common site of primary.

Advantages of CT scan

- One to 2 mm sized sections are possible
- Amount of exposure to radiation is less
- More accurate, sensitive, and specific
- Small lesions are also detected
- CT guided biopsies are done at present, safely

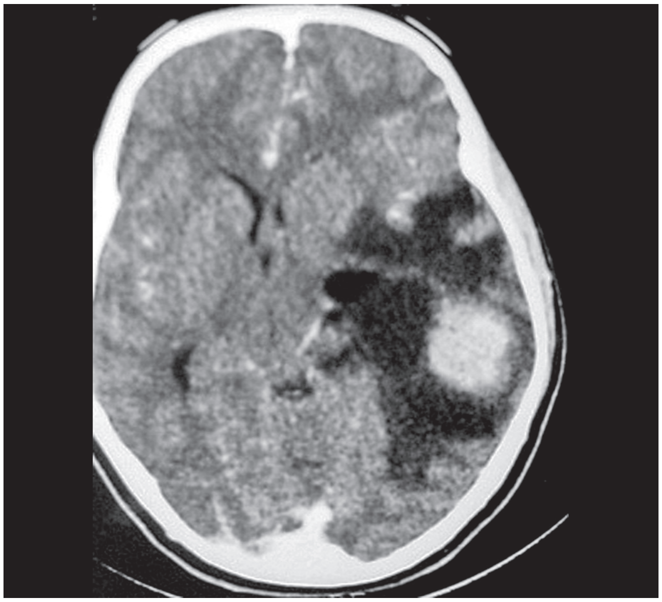

Fig. 31.12: CT picture showing astrocytoma.

Disadvantages

- **Interpretation** by an experienced radiologist is important
- **Artefacts** can be present
- **Cost factor** and **availability**

Findings

- Extradural haematoma—*Biconvex lesion.*
- Subdural haematoma—*Concavoconvex lesion.*

Ultrasound and CT scan are conveyors of morphological information.

- Smooth margin in benign condition.
- Irregular margin in malignant condition.

Advantages of spiral CT scan

- Reduced scan time. Useful in children and critically ill-patients
- Imaging in both arterial and venous phases is possible
- Improved lesion detection. Missing a lesion is uncommon
- Multiplanar and 3-dimensional analysis like CT angiography, complex joint imaging, facial bone imaging is possible

High resolution CT *(HRCT)* is a CT technique used in chest scan where thin sections are taken to have better quality images.

MAGNETIC RESONANCE IMAGING (MRI)

Earlier, named as *Nuclear magnetic imaging*, the term is not used now.

It can be **Plain MRI** or **Contrast MRI**. Contrast agent is *Gadolinium,* given intravenously**.**

Principle

When patient is placed in an *external high magnetic field*, protons of hydrogen atoms rotate in phase with each other and gradually return to their original position releasing small amounts of energy which is detected by sensitive coils. Proton density and relaxation time are assessed by radiofrequency pulse and the computer generates a Gray scale image from this data.

T1 relaxation time is the time taken to return to original axis. *T1 images* are used to find out *normal anatomical details.* It has got high soft tissue discrimination. Here fluid (CSF) looks black.

T2 relaxation time is the time taken by the proton to diphase. It is used to assess *pathological processes. In T2 images* fluid looks white.

In proton density images fluid looks in between black and white.

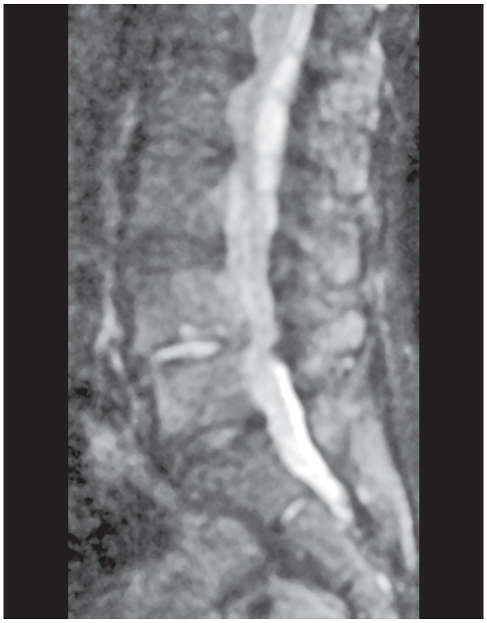

Fig. 31.13: MRI showing spine tuberculosis. Note the lesion and vertebral collapse.

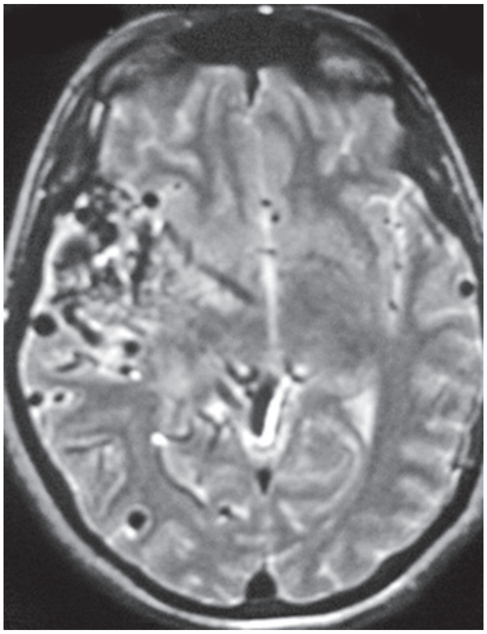

Fig. 31.14: MRI showing AV malformation.

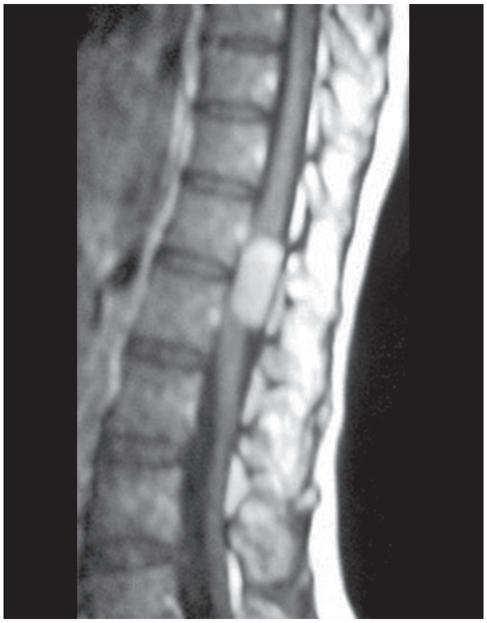

Fig. 31.15: MRI picture, which shows extradural schwannoma.

Uses of MRI

- It is very useful in intracranial, spinal and musculoskeletal lesions including *joint pathologies.*
- It gives direct anatomical sections of the area, with lesions at a high resolution.
- **MR angiogram** is done without injecting IV contrast agents.
- **Cardiac MRI** is very useful.
- **Breast MRI** is used in multifocal recurrent cancers.
- **Magnetic resonance cholangiopancreatography** *(MRCP)* is a very useful noncontrast diagnostic tool which may replace diagnostic ERCP.
- **MR spectroscopy** is chemical analysis of elements in a tissue to differentiate between tumour, inflammation, and degeneration.

Advantages

- Artefacts are not common.
- More sensitive and specific than CT scan.

Contraindications

Patients with ***prosthesis in the body, metallic foreign bodies, pacemakers, cochlear implants, cranial aneurysm clips should never*** undergo MRI

Precaution

Before entering the MRI room, the patient and other personnel should remove all magnetically attractive materials

Disadvantages

- ***Availability*** and ***cost factor.***
- It is time consuming.
- Patient compliance is poor.
- It is not feasible in patients suffering from ***Claustrophobia***.
- It is not ideal in emergencies and critically ill patients.
- It is not useful in lung pathology and subarachnoid haemorrhage.

RADIONUCLIDE IMAGING

- It represents function of an organ than morphology.

Rays:

- α-particles are emitted by the natural radionuclides like radium, which are no longer used in medicine.
- β-particles are useful for therapy but not for diagnosis.
- γ-rays pass out of the body and so used for the diagnostic purpose.
- Mapping is done using sophisticated gamma camera

Technetium 99^m

- It is commonest radionuclide used (99- is mass number; m- metastable.)
- It is administered IV.
- Pure γ-rays emitter.
- Short half life.
- Widely used gamma ray detectors are specified to Tc99^m.

Uses

Tc99^m	labelled serum albumin is used to detect pulmonary emboli
Tc99^m	labelled phosphate is used to image bone
Tc99^m	labelled sulphur colloid is used to detect the functions of liver, spleen, bone marrow
Tc99^m	labelled HIDA (Hippuric immuno diacetic acid) or PIPIDA is used to study the functions of hepatocytes and biliary tract
Tc99^m	labelled DMSA (Dimercapto succinic acid) which is taken up by the renal cortical cells, is used in renal function tests
Tc99^m	labelled DTPA (Diphenyl triamine penta acetic acid) measures G.F.R
Tc99^m	labelled HMPAO (Hydroxy methyl propylamine oxime) is used in Alzheimer's disease and schizophrenia as it crosses the blood brain barrier

Other radionuclides used:

- Thallium 201 chloride for cardiac imaging
- Gallium 67 nitrate to detect tumours and inflammation
- I^{123}

Advantages	***Disadvantages***
Safer	Availability
Easier	Cost factor
No side effects	Not specific
	Fast half life

POSITRON-EMISSION TOMOGRAPHY (PET Scan)

It is a non-invasive diagnostic method to assess the biochemical and physiological status of a tissue.

It is used in complimentary with CT scan and MRI.

Two protons are used, they are positive electrons (*positrons*).

Most clinically used positron emitting radionuclides is fluoro-deoxyglucose (FDG), others are ^{82}Rb, ^{15}O, ^{13}N.

Detectors used are Bismuth germanate (BGO) crystals or sodium iodide crystals.

Principle of *electronic collimation* is used to produce images from the radiation emitted from positron emitting tracers.

Uses

- To assess myocardial perfusion (^{82}Rb) and viability (FDG) study.
- Epilepsy—To localise temporal lobe epilepsy (FDG)
- Cancer imaging—Lung cancer (detection and staging).
- Colorectal cancer.
- Melanoma.
- Head and neck cancer and breast cancer.
- Musculoskeletal tumours.
- Thyroid cancer (I^{131}).

Advantage

Very specific.

Disadvantage

Very expensive and limited availability.

Note:

Most of the U/S pictures, CT and MRI pictures in this book are from Balmatta Scan Center, Mangalore, I am thankful to consultants Dr Raghavendra Bhat and Dr Ravichandra there.

Radioisotope studies are indicators of physiological processes

Chapter 32 Operative Surgery

A. Asepsis and Sterilisation

> *Bearing in mind that it is from the vitality of the atmospheric particles that all the mischief arises, it appears that all that is requisite is to dress the wound with some material capable of killing these septic germs, provided that any substance can be found reliable for this purpose, yet not too potent as a caustic.*
>
> **—Joseph Lister,** 1867

CHAPTER OUTLINE

- Sterilisation
- Disinfection
- Antisepsis
- Asepsis
- Different Methods of Disinfection/Sterilisation

STERILISATION

- It is freeing an article by removing or killing all bacteria, spores, fungi and viruses.

DISINFECTION

- It is killing of all bacteria, fungi and viruses but not spores.

ANTISEPSIS

- It is inhibition of growth of microorganisms.

ASEPSIS

- Asepsis means—organisms are prevented to access the patient or individual.

DIFFERENT METHODS OF DISINFECTION/ STERILISATION

Physical Agents

- ***Burning*** or ***incineration*** is used to disinfect contaminated articles like dressings.
- ***Hot-air oven****:* Here temperature used is 160 to 180 degree for one hour.
- ***Boiling:*** It kills bacteria but not spores and viruses. Temperature is between 90 to 99 degree. It is used to disinfect syringes, utensils. It is not useful for gloves, rubber materials.
- ***Autoclave***: It is steam under pressure. Temperature attained is between 120-135 degree. It is sterilised for 20 minutes with 15 pounds/sq. inch pressure. It kills all organisms including spores. Completeness of sterilisation is confirmed by using specific gelatin protein which precipitates only in steam underpressure for 20 minutes. Green coloured strip turns to black if autoclave is complete *(signaloc).* Surgical gloves, linen, cotton, dressings, surgical instruments are sterilised by this method. Sharp and plastic instruments cannot be sterilised by this method. Bacillus thermophilus spores are used to assess the completeness of the sterilisation in mass scale. Double autoclaving is done for instruments of orthopaedic or ophthalmic surgeries.
- *The Bowie-Dick* method is also used to check the completeness of sterilisation.
- ***Radiation:*** *Ionising type of radiation:* Atomic gamma radiation is used as commercial method to sterilise suture materials, disposable materials in packets. It is viable, safe and cheaper.
- *Non-ionizing radiation* either infrared radiation or ultraviolet radiation is used to reduce the bacteria in air, water. Bacteria and virus are vulnerable to ultraviolet rays below 3000Å. Exposure to eyes and skin can cause burn injury.

Materials	*Method of sterilisation*
All theatre appliances	Autoclave
Sharp instruments (scissors, needles, blades)	Glutaraldehyde 2%, lysol
plastic materials	
Endoscopes	Glutaraldehyde
Rubber equipments	Glutaraldehyde
Syringes	Autoclave, hot air oven, gamma radiation
Heart-lung machine	Ethylene oxide
Disposable articles	Gamma radiation
Operation theatre and rooms	Ideally by U-V radiation or by formaldehyde
Sera and biological materials	Filtration
Lab glassware	Hot-air oven
Ward, sick room, furniture	Formaldehyde, iodophor spray, glutaraldehyde
Clothes, bed sheets especially for burns patients	Autoclaving
Soiled dressings, materials, animal carcasses	Incineration, lysol, iodophors
Excreta	Lysol, iodophors
Cleaning of skin before surgery	Iodophors 2%, savlon, spirit
For cleaning infected wounds	Iodophors, acriflavine, savlon, H_2O_2
To remove slough from the wounds	EUSOL, H_2O_2
Before injection	Spirit is used to clean the skin
Cleaning the ward	Phenol, cresol, lysol
Hand wash	Chloroxylenol, savlon, spirit, iodophors
Bladder wash	0.1% potassium permanganate solution (Condy'slotion), solution of acetic acid and silver nitrate
Water	Chlorination, potassium permanganate
Fruits, vegetables	Potassium permanganate

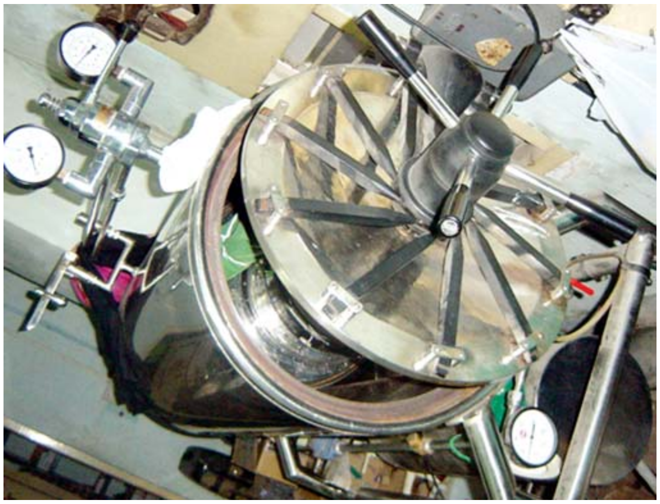

Fig. 32.1: Autoclave machine for sterilisation.

Chemical Agents

- ***Phenol:*** It is used as standard to compare the efficacy of other agents.
- ***Cresol*** is more powerful and nontoxic. 5% solution is used.
- ***Lysol*** is emulsified cresol with soap. 2% solution is effective.
- ***Chlorhexidine (hibitane***) is useful antiseptic.
- ***Hexachlorophane***: It is not used in infants and children because it can get absorbed through intact skin in this age group causing severe neurotoxicity.
- ***Dettol (chloroxylenol***) 5% solution is used.
- ***Cetrimide*** is cationic surfactant (cetavlon) 2% solution is used.
- ***Savlon*** is combination of cetrimide and hibitane. It is very commonly used antiseptic in operation theatres, wards.

Halogens

- Bleaching powder.
- Sodium hypochlorite.
- ***EUSOL: Edinburg University Solution*** contains sodium hypochlorite, boric acid and calcium hydroxide.
  EUSOL bath is dipping the ulcer bearing part in dilute EUSOL solution for 30 minutes 2-3 times a day.
- Iodine
- ***Iodophors:*** These are antiseptics and also sporicidals. They are non-irritant and do not stain skin. ***Povidone-iodine*** is a good example which is commonly used.
- ***Alcohols***: Ethyl or isopropyl alcohols are used.

Treat the patient as a whole, "Half a sheep is mutton".

- ***Formaldehyde:*** It is useful to disinfect the rooms like operation theatre. It is effective at a high temperature and humidity of 80-90%. It is commonly used to fumigate the room. 500 ml of formalin with one litre of water is boiled to get formaldehyde vapour. Formaldehyde vapour can be created by adding potassium permanganate to the same solution. Room is kept closed for 12 hours.
- ***Glutaraldehyde (cidex 2%)*:** It is used to sterilise sharp instruments. Instrument should be dipped for 10 hours to achieve complete sterilisation. It is potent bactericide, sporicide, fungicide and viricide.
- ***Hydrogen peroxide (H_2O_2)*:** It is used as topical oxygen therapy. Because of its effervescence and release of nascent oxygen it removes the tissue debris. It is used to clean wounds, cavities, ulcers, as mouth wash and as ear drops to clear earwax.
- ***Acriflavine and proflavine*** are orange-red coloured dyes used as antiseptics. It is effective against gram-positive and few gram-negative organisms. It retains its activity in pus and body fluids.

Figs 32.2A and B: Operation theatre mop rack to keep 'used mops' during surgery after use.

B. Instruments

CHAPTER OUTLINE

- Cheatle's Forceps
- Sponge Holding Forceps
- Mayo's Towel Clip
- Artery Forceps
- Right Angle Forceps
- Kocher's Forceps
- Allis' Tissue Holding Forceps
- Babcock's Forceps
- Lane's Tissue Holding Forceps
- Morant-Baker's Appendix Holding Forceps
- Volkmann's Retractor
- Langenbeck's Retractor
- Czerny's Retractor
- Morris' Retractor
- Deaver's Retractor
- Doyen's Retractor
- Self-retaining Retractor
- Single Hook Retractor
- Plain Non-toothed Dissecting Forceps
- Toothed Dissecting Forceps
- Surgical Needles
- Needle Holder
- Joll's Thyroid Retractor
- Moynihan's Occlusion Clamp
- Payr's Crushing Clamp
- Desjardin's Choledocholithotomy Forceps
- Bake's Dilator
- Sinus Forceps
- Scissors
- Volkmann's Scoop
- Tracheostomy Tube
- Drains
- Foley's Catheter
- Malecot's Catheter
- Simple Red Rubber Catheter
- Lister's Urethral Dilator
- Ryle's Tube
- Infant Feeding Tube
- Kehr's 'T' Tube
- Proctoscope
- Flatus Tube

CHEATLE'S FORCEPS

- It is used to pick sterilised articles like instruments and drapes so that touching of the instruments is avoided while transferring them. It is kept dipped in antiseptic solutions. It does not have lock.

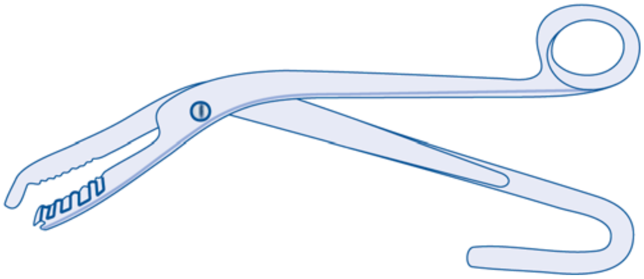

Fig. 32.3: Cheatle's forceps.

SPONGE HOLDING FORCEPS (RAMPLEY'S)

- It has got fenestrated, serrated, flat distal end. It is used to clean the operative field, to swab the cavities, to mop the oozing area, to hold gallbladder and cervix during surgeries, for blunt dissections, as ovum forceps.

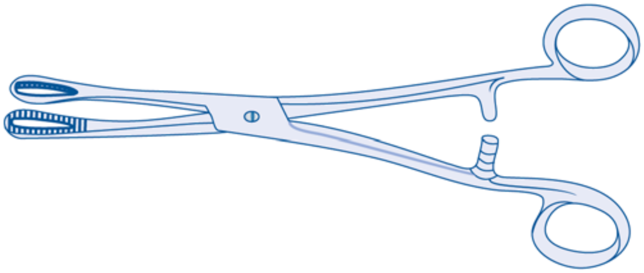

Fig. 32.4: Sponge holding forceps.

MAYO'S TOWEL CLIP (FIG. 32.5)

- It is used to fix drapes in operative field.
- It is used to fix suction tubes, diathermy wires, laparoscopic cables in operative table.
- It is used to fix ribs in flail chest.

ARTERY FORCEPS (HAEMOSTAT)

Types

Based on size:

a. Small or mosquito artery forceps.
b. Medium sized artery forceps.

Anger is a short madness which often creates life-long tragedy.

c. Large artery forceps.

Based on shape:

a. Straight artery forceps.

b. Curved artery forceps.

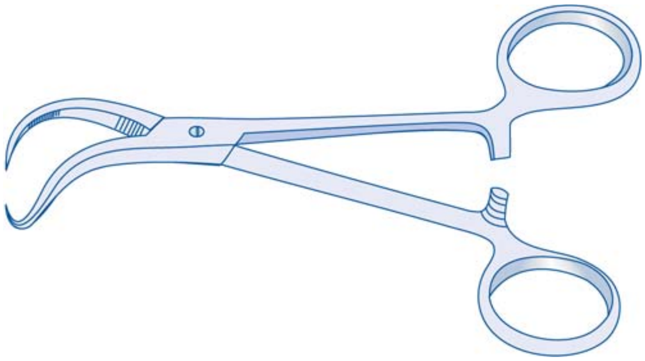

Fig. 32.5: Mayo‘s towel clip

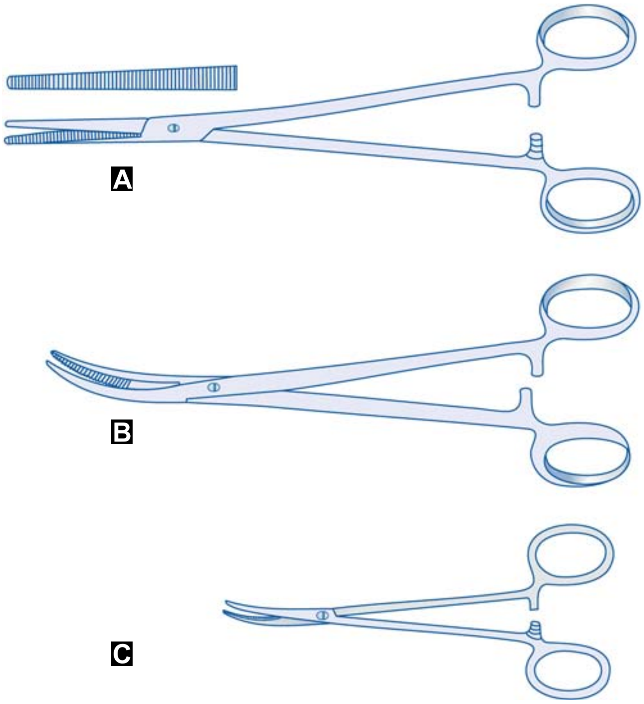

Figs 32.6A to C: Artery forceps: (A) Straight, (B) Curved, (C) Mosquito.

Features of Artery Forceps

Distal blades have transverse serrations which are well-apposed. Lock in the proximal part.

Uses

- To catch bleeding points
- To open the fascial planes in different surgeries
- To pass a ligature
- To hold fascia, peritoneum, aponeurosis
- To hold sutures
- To drain an abscess like a sinus forceps
- To hold gauze as peanut

RIGHT ANGLE FORCEPS

- It is used to dissect pedicles and to pass ligatures.

KOCHER’S FORCEPS

- It has got serrations in the distal blades and apposing tooth in the tip.
- It is used to hold pedicles, tough structures, cut ends of the muscles.
- It is used to hold gauze for blunt dissection, to hold resected bowel, to hold ribs during rib resection.

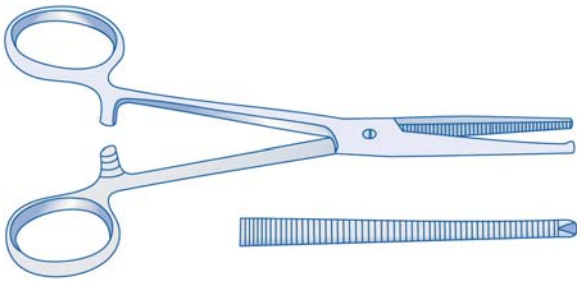

Fig. 32.7: Kocher’s forceps.

ALLIS’ TISSUE HOLDING FORCEPS

- Here distal blades are not apposing each other.
- Tip has got teeth in each blade which are apposing.
- It has got a lock on the proximal part.
- It is used to hold skin flaps, fasciae, aponeurosis, bladder wall.

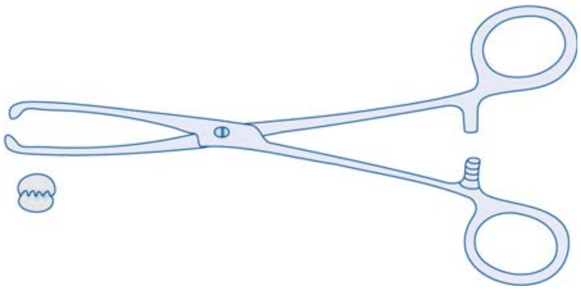

Fig. 32.8: Allis’ tissue holding forceps.

BABCOCK’S FORCEPS

- Its distal part of distal blades are curved with a triangular fenestra in it which allow soft tissues to bulge out. Tip is non-traumatic with transverse serrations on it. It has got a lock in the proximal part.

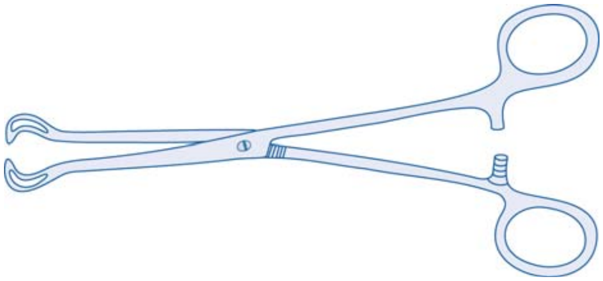

Fig. 32.9: Babcock’s forceps.

- It is used to hold any part of the bowel, fallopian tubes, appendix, ureter, cord, etc.

LANE'S TISSUE HOLDING FORCEPS

- It has got thick, stout distal blades with oval fenestra in each blade.
- It has got apposing tooth in the tip.
- It has got a lock in the proximal part.

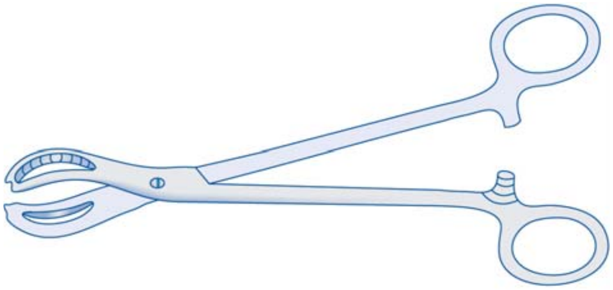

Fig. 32.10: Lane's tissue holding forceps.

- It is used to hold bulky and tough structures, to hold lymph nodes.
- It is also used as towel clip, as sponge holding forceps.

MORANT-BAKER'S APPENDIX HOLDING FORCEPS

- It is like Lane's forceps but with apposing serrations proximal to the tooth. These serrations give a good grip in mesoappendix while holding appendix in appendicectomy. Its use is replaced by Babcock's forceps.

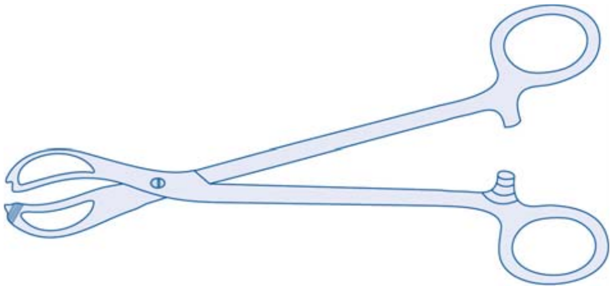

Fig. 32.11: Morant-Baker's appendix holding forceps.

VOLKMANN'S RETRACTOR

- It is used to retract fasciae in soles and palms.

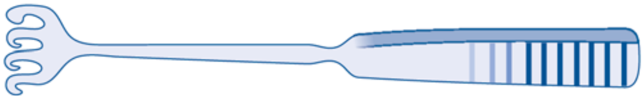

Fig. 32.12: Volkmann's retractor.

Retractors

- Superficial retractors
- Deep retractors
- Self-retaining retractors

LANGENBECK'S RETRACTOR

- It has got a long handle and a small solid blade. It is used in hernia surgery or any superficial surgeries to retract skin, fasciae and aponeurosis, etc.

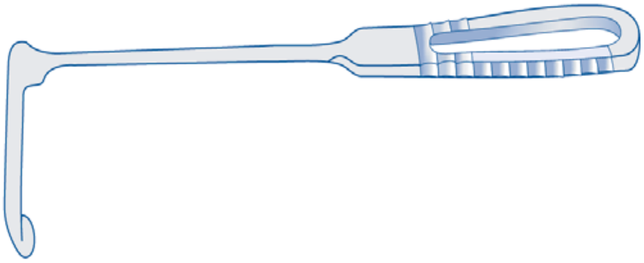

Fig. 32.13: Langenbeck's retractor.

CZERNY'S RETRACTOR (HERNIA RETRACTOR)

- This retractor has got thick, small blade on one side and biflanged hook on the other side in opposite directions. It is used in surgeries like hernia, laparotomy especially during closure.

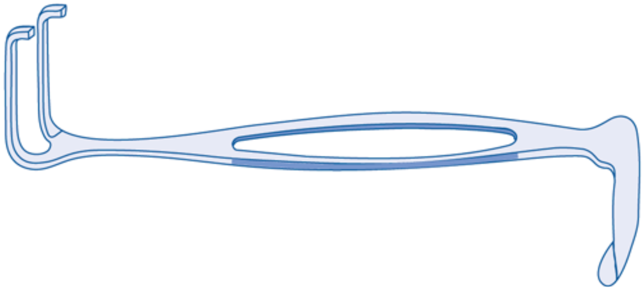

Fig. 32.14: Czerny's retractor.

MORRIS' RETRACTOR

- It may be single blade type or double blade type.
- It is used to retract abdominal wall.

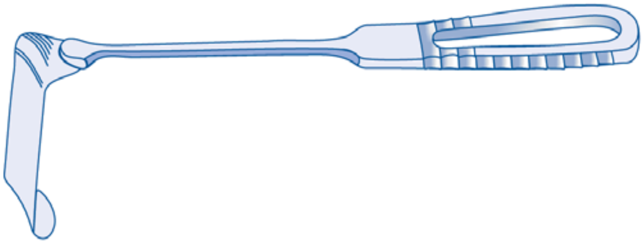

Fig. 32.15: Morris retractor.

DEAVER'S RETRACTOR

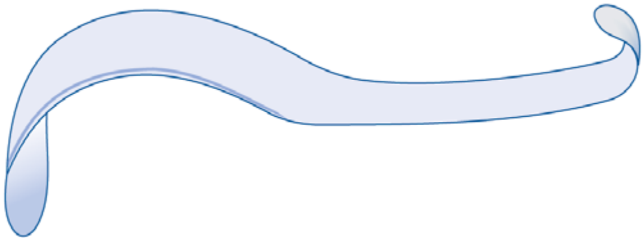

Fig. 32.16: Deaver's retractor.

- It is a retractor with a broad, gently curved blade.
- It is used to retract liver, spleen and other abdominal viscera.

- It is atraumatic and gives adequate exposure of the surgical field.

DOYEN'S RETRACTOR

- It is used in pelvic surgeries.

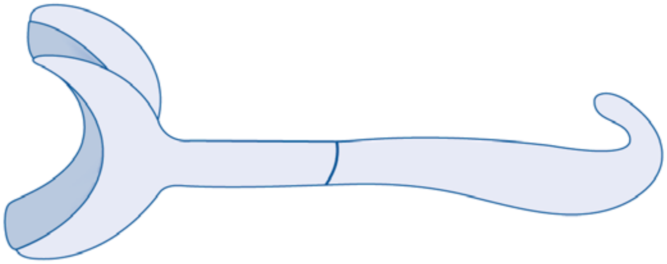

Fig. 32.17: Doyen's retractor.

SELF-RETAINING RETRACTOR

- It has got different adjustable blades so as to retract abdominal wall and tissues during surgery (Balfour's retractor).

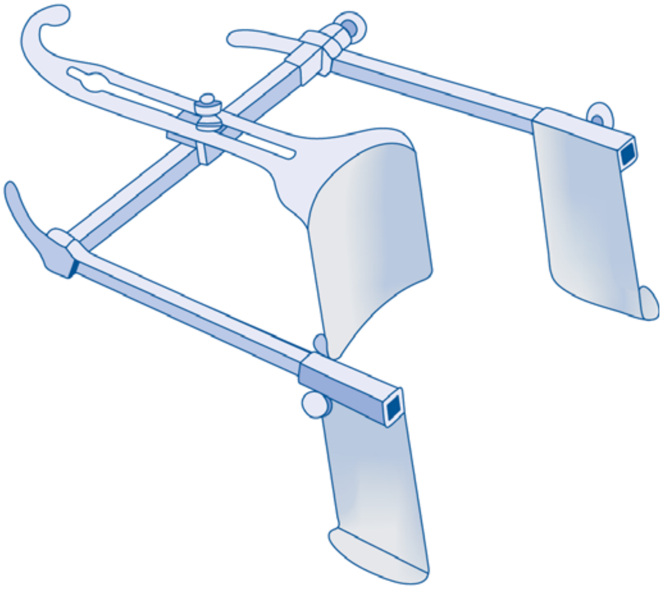

Fig. 32.18: Self-retaining retractor.

SINGLE HOOK RETRACTOR

- It is used to retract skin.

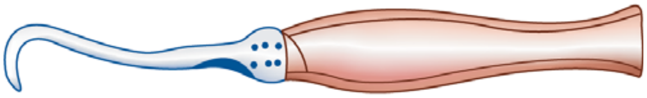

Fig. 32.19: Single hook retractor.

PLAIN NON-TOOTHED DISSECTING FORCEPS

- It is used to hold delicate structures like peritoneum, vessels, bowel, nerves, tendons (Fig. 32.20).

TOOTHED DISSECTING FORCEPS

- It is used to hold skin and tough structures.

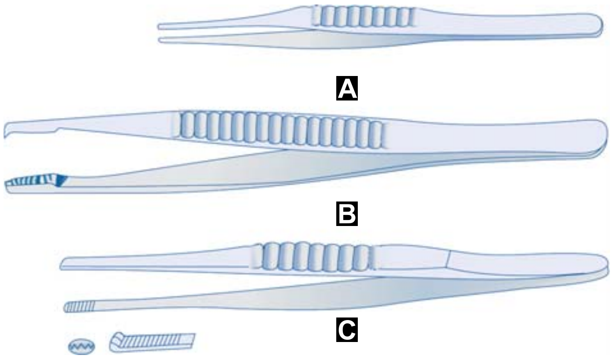

Figs 32.20A to C: Dissecting forceps: (A) Non-toothed, (B and C) Toothed.

SURGICAL NEEDLES

Types

Based on the edge

- Round body needle
- Cutting needle
- Reverse cutting needle
- Taper cut needle
- Side-to-side flat—Hagedron needle

Based on curvature

- Straight needle
- Curved needle. Half circle; 5/8 circle, etc.

Based on Existence of the Eye

- ***Atraumatic needle*** is eyeless. Here suture material is attached to the needle by swaging. Size of the suture material and that of needle is same and so tissue trauma is less. Needle once used is disposed of (not reusable).
- ***Traumatic needle:*** It is eyed needle. Needle in the eye area is wider than the body of the needle and so tissue trauma is more. These needles are reusable.

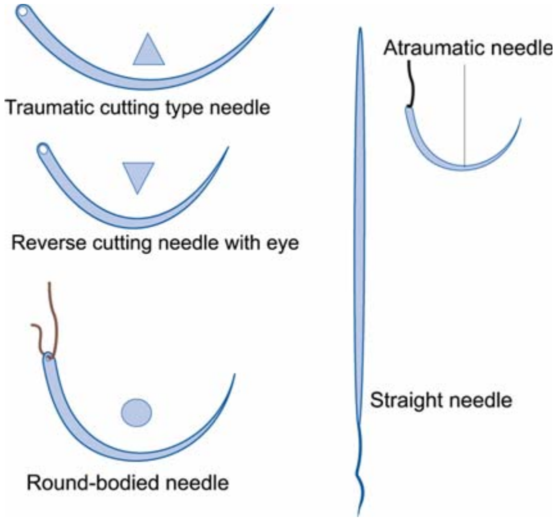
Traumatic cutting type needle
Atraumatic needle
Reverse cutting needle with eye
Straight needle
Round-bodied needle

Fig. 32.21: Needles.

- *Round body needles* are used in soft structures like peritoneum, muscle, vessel, nerves, tendons, bowel, soft tissues.
- *Cutting needles* are used to suture skin, aponeurosis and tough structures.
- Reverse cutting needle is used to suture mucoperiosteum.

NEEDLE HOLDER

- Smaller distal blades with criss-cross serrations often with a groove in the middle are the features of a needle holder.

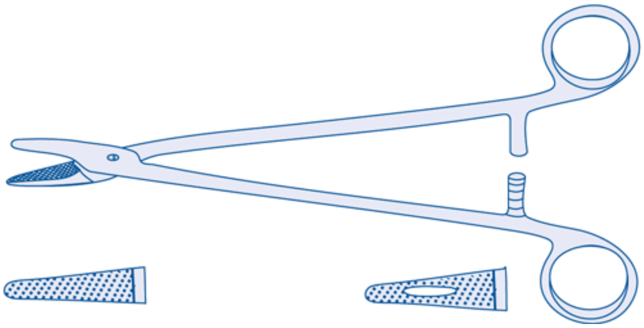

Fig. 32.22: Needle holder.

- It may be straight or curved. It may be available with different sizes. While holding a needle in a needle holder one should get a good control and good grip. This is achieved by placing the needle at the junction of proximal 2/3rd and distal 1/3rd. Needle holder should be held between thumb and ring finger.

JOLL'S THYROID RETRACTOR

- It is a self-retaining retractor specifically used for thyroid surgeries.

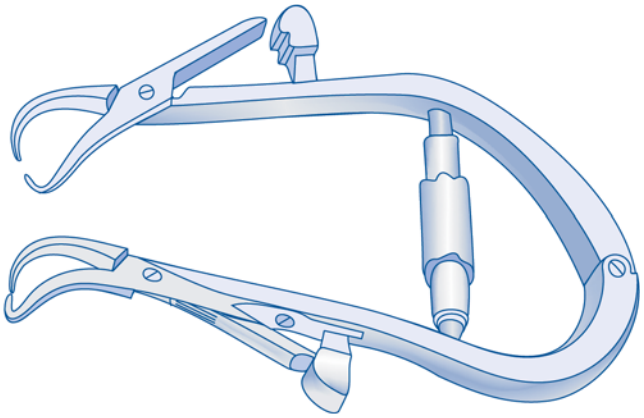

Fig. 32.23: Joll's thyroid retractor.

MOYNIHAN'S OCCLUSION CLAMP

- It has got long distal blades with longitudinal serrations.
- It may be straight or curved.
- It is non-traumatic, non-crushing type.
- It occludes lumen of the bowel/stomach and so prevents spillage of the content of the bowel.
- It also occludes the vessels in the wall of the bowel and so prevents bleeding during surgery.
- It is used during anastomosis of the stomach and other parts of the bowel.

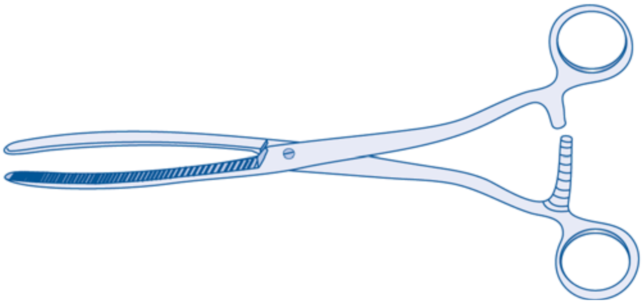

Fig. 32.24: Moynihan's occlusion clamp.

PAYR'S CRUSHING CLAMP (GASTRIC)

- It is stout and heavy instrument with double lever in the handle.

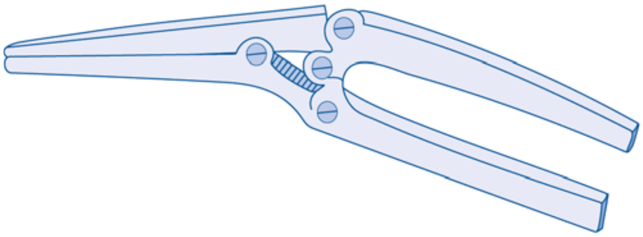

Fig. 32.25: Payr's crushing clamp.

- It crushes the bowel once applied. So before applying it, line of resection of stomach/bowel should be assessed properly. It is applied to the part which is removed. Viability of the bowel is lost once it is applied.
- It is used in gastrectomy and resection and anastomosis of the bowel.

DESJARDIN'S CHOLEDOCHOLITHOTOMY FORCEPS

- It has got long distal blades with smooth serrations and fenestra in the tip. It does not have lock and so accidental damage of CBD mucosa or crushing of the CBD stone are avoided.

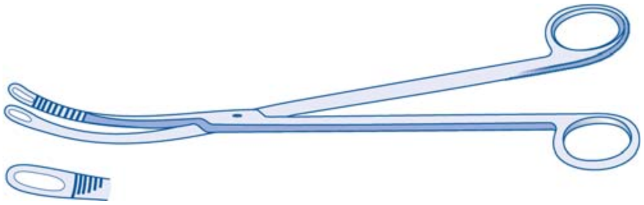

Fig. 32.26: Desjardin's choledocholithotomy forceps.

- It is used for choledocholithotomy (removal of CBD stones).

BAKE'S DILATOR

- It is long malleable metallic instrument with club at the terminal end.

Surgery is always second best; if you can do something else, its' better.—**John Kirklin**

- It is used to assess the CBD, duodenal papilla for patency or block.

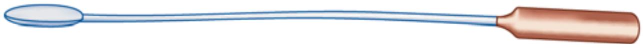

Fig. 32.27: Bake's dilator.

SINUS FORCEPS (LISTER's)

- It has got straight, long blades with serrations in the tip. It does not have a lock.
- It is used to drain pus from abscess cavity (Hilton's method). It is called as sinus forceps because it was initially originated to pack the sinus cavities. It is less traumatic.
- Sinus forceps has no lock; no serrations; broad tip; blunt.

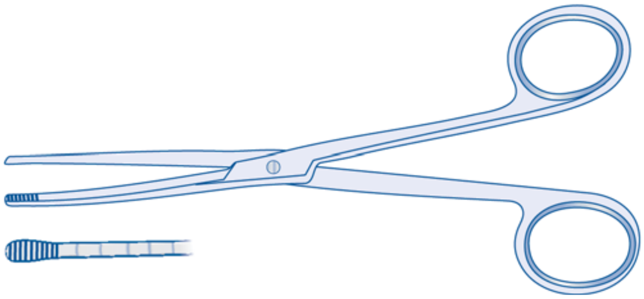

Fig. 32.28: Lister's sinus forceps.

SCISSORS

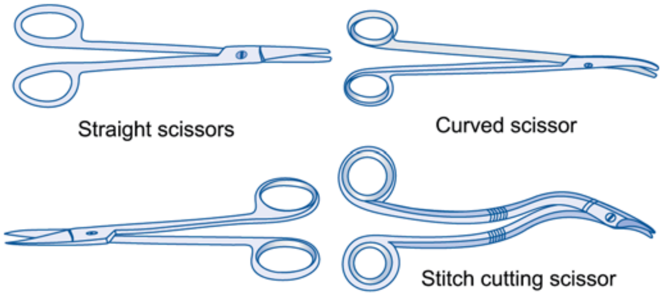

Fig. 32.29: Scissors.

VOLKMANN'S SCOOP

To scoop cavities, ulcer bed, granulation tissues.

On either side different sized scoops are present.

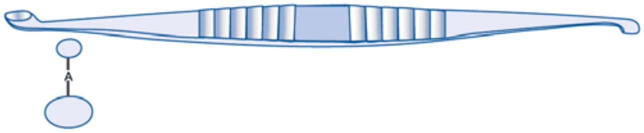

Fig. 32.30: Volkmann's scoop.

TRACHEOSTOMY TUBE

- ***Fuller's bivalved tracheostomy tube***: It has got outer tube and inner tube. Outer tube is biflanged and so insertion is easier. Inner tube is longer with an opening on its posterior aspect. Inner tube can be removed and reinserted easily whenever required.
- ***Jackson's tracheostomy tube***: It has got outer tube, inner tube and an obturator.
- Red-rubber tracheostomy tube.
- PVC tracheostomy tube.

Modern tracheostomy tubes are made of plastic. They are soft, least irritant and disposable. They have inflatable cuff which makes it easier to give assisted ventilation. Cuff should be deflated at regular intervals to prevent tracheal pressure necrosis (For assisted ventilation, endotracheal, tube can be kept for 7 days. Beyond that period, patient needs tracheostomy for further ventilation).

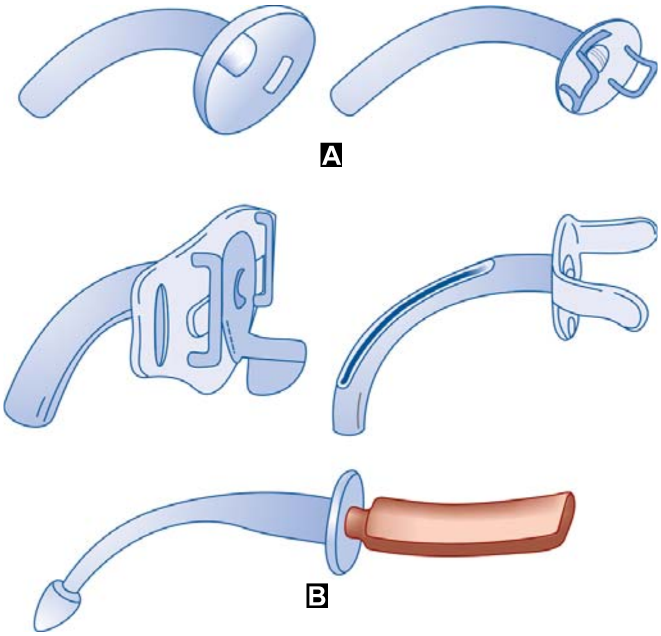

Figs 32.31A and B: Tracheostomy tube: (A) Fuller's, (B) Jackson's tracheostomy tube.

Indications for tracheostomy

- In head, neck and facial injuries
- Tetanus
- Tracheomalacia after thyroidectomy
- Laryngeal oedema/spasm
- Major head and neck surgeries like commando's operation, block dissection, etc.

DRAINS

A drain is a created channel which allows any fluid collected, to come out after closure of the main wound.

Types

- *Corrugated rubber drain:* It drains by capillary action and gravity. It is cheaper and technically easier. But it allows soakage of dressings and causes discomfort to the patient.
- *Tube drains*
  - Malecot catheter can be used as a tube drain.
  - Penrose soft latex rubber tube.
  - Multiple perforated tubes.

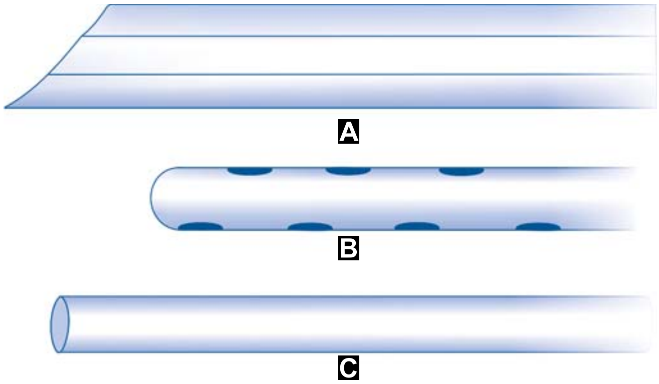

Figs 32.32A to C: Drains: (A) Corrugated rubber drains, (B) Multiple perforated drains, (C) Tube drain.

Advantages of tube drains

- Quantity of fluid like bile, pus can be measured
- It can be kept for longer time
- Skin excoriation will not occur
- Patient remains more comfortable
- Infection rate is less
- Removal is easier
- Dye can be injected and cavity or communication can be assessed using 'C-ARM'

- *Closed suction tube drain system.*
- *Glove drain.*
- *Wick drain* is a gauze drain to drain pus, discharge, etc.
- *Sump drain:* This is a type of drain where parallel air vent prevents the adjacent soft tissues from being sucked into the drain when negative pressure is applied.

Advantages of sump drain

- No drain blockade
- Resists collapse of the structure when suction is applied.

Uses of sump drain

- Collection of irritant discharges (enterocutaneous fistula)
- Collection of secretions having activated enzymes (high small bowel, pancreatic fistula)
- Draining proximal stump in TEF with esophageal atresia to prevent aspiration

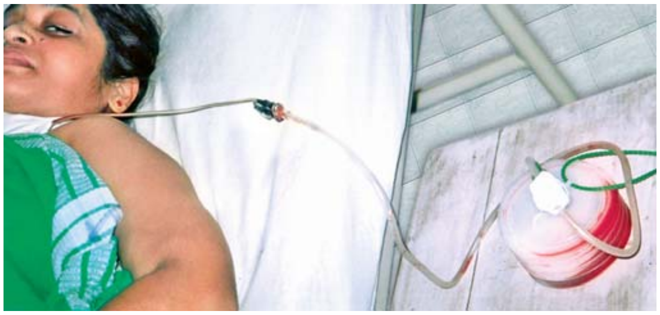

Fig. 32.33: Suction drain used in thyroid surgery.

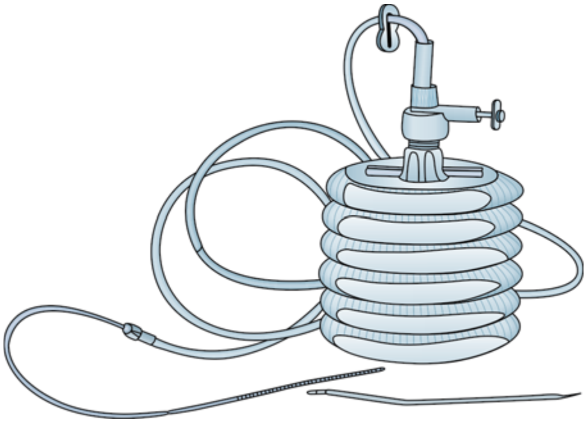

Fig. 32.34: Romovac suction drain. Here suction is created by pressing the suction corrugation. There is a sharp metallic introducer to pass the tube into the required area after puncturing the skin. It is used for thyroidectomy, mastectomy, radical dissection, wide excisions, flap surgeries, etc.

Classification of Drain Systems

- ***Open (static) drain:*** For example, corrugated drain, penrose drain. Infection rate is higher.
- ***Closed siphon drain:*** Here drain is connected to a sterile bag with or without one-way valve. It reduces the infection.
- ***Closed suction drain:*** Here negative pressure of—100 to 500 mmHg is used to create vacuum to drain the secretions.
- ***Sump suction drain:*** Here negative suction with a parallel air-vent is used to prevent the adjacent soft tissues being sucked into the lumen of the drain.
- ***Under water seal drain*** to drain pleural space.

Indications for drains

- In drainage of an abscess
- In bleeding surgical conditions like trauma, peroperative bleed
- In haemo, pyo or pneumothorax
- In acute abdominal conditions like peritonitis, haemoperitoneum
- In major abdominal surgeries like of pancreas, biliary tree, stomach, etc.
- In thyroid surgery
- In hydrocele surgery

Problems in Drains

- Infection can occur through the drain.
- Displacement.
- It may not drain adequately and can give a false information.
- It may interfere with healing process inside.

Presently keeping a drain itself is a questioned debate and controversy all over.

Older dictum was 'when in doubt keep a drain and the surgeon can sleep happily'—is questioned at present.

DRAINS IF NOT USED PROPERLY MAY BE COUNTER-PRODUCTIVE.

> *The postoperative treatment is as essential as the operation and the surgeon is as much responsible for the postoperative treatment as for the operation.*
> **— Roscoe C Giles**

FOLEY'S CATHETER

Refer Urology

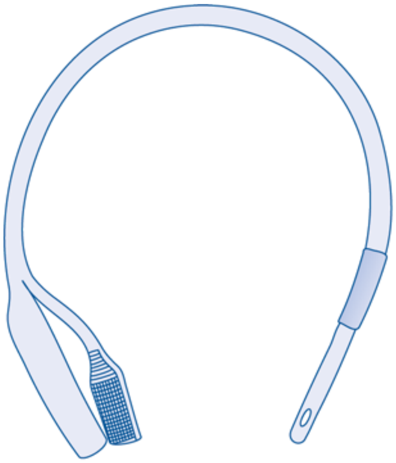

Fig. 32.35: Foley's catheter.

MALECOT'S CATHETER

Refer Urology

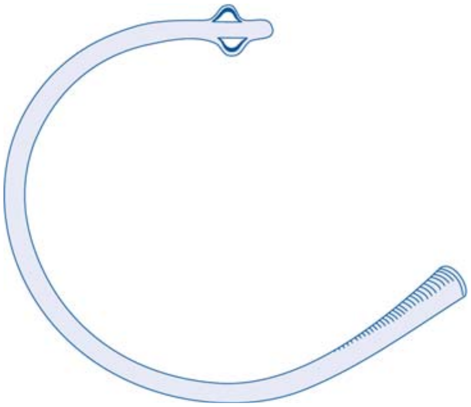

Fig. 32.36: Malecot's catheter.

SIMPLE RED RUBBER CATHETER

Refer Urology

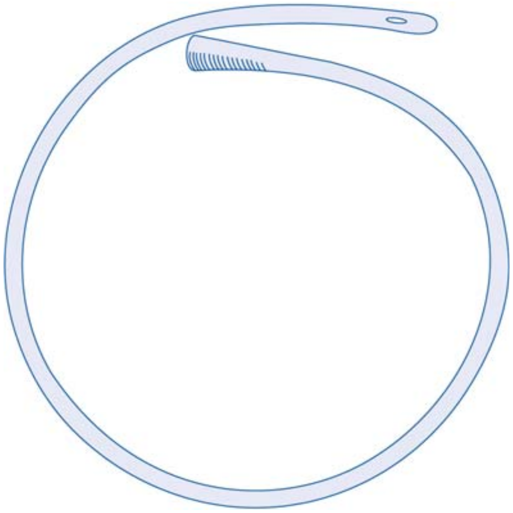

Fig. 32.37: Red rubber catheter.

LISTER'S URETHRAL DILATOR

It has got olive tip and it is used to dilate stricture urethra. Other dilators are Clutton's dilator and filiform bougies.

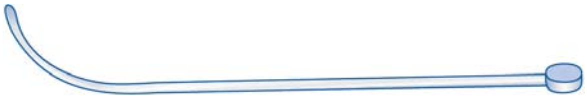

Fig. 32.38: Lister's urethral dilator.

RYLE'S TUBE

It is one meter long and is made of red rubber or plastic.

It has got *three lead shots* in the tip which makes it radiopaque. It also facilitates easy passage of the tube through the oesophagus.

It has got markings at different levels:

- At 40 cm distance, at the level of gastro-oesophageal junction.
- At 50 cm distance, at the level of body of the stomach.
- At 60 cm distance, at the level of the pylorus.
- At 65 cm distance, at the level of the duodenum.

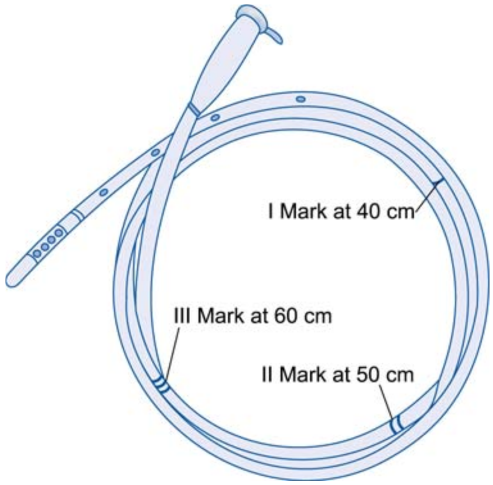

Fig. 32.39: Ryle's tube.

Indications

- Diagnostic:
- For gastric function tests—to assess free acid and total acid
- Hollander's test for completion of vagotomy
- To diagnose tracheo-oesophageal fistula
- Baid test for pseudocyst of the pancreas
- Therapeutic:
- In acute abdominal conditions like peritonitis/ obstruction
- In abdominal trauma
- After abdominal surgeries
- In pyloric stenosis
- In upper GIT bleeding
- In paralytic ileus
- For feeding purpose in conditions like comatose patients, faciomaxillary injuries, major head and neck surgeries

INFANT FEEDING TUBE

- There are *no lead shots* and markings on the tube.
- It is used for feeding purpose in infants who is under coma, with faciomaxillary injuries and anorexia.

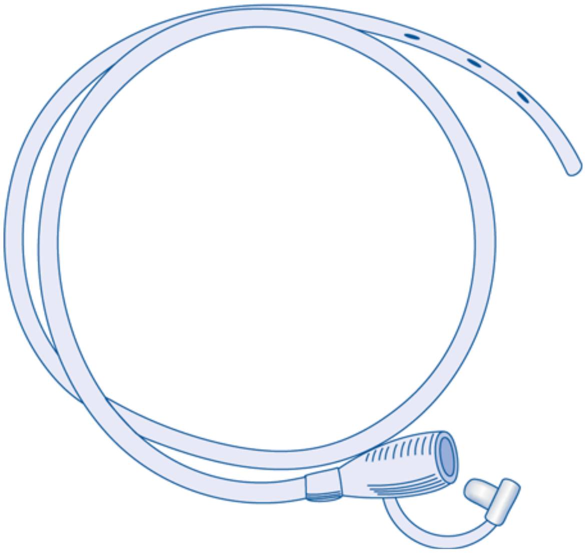

Fig. 32.40: Infant feeding tube.

KEHR'S 'T' TUBE

- It is used after opening CBD (choledochotomy). CBD is closed with 'T" tube placed in the CBD.
- It is made up of latex or red rubber.

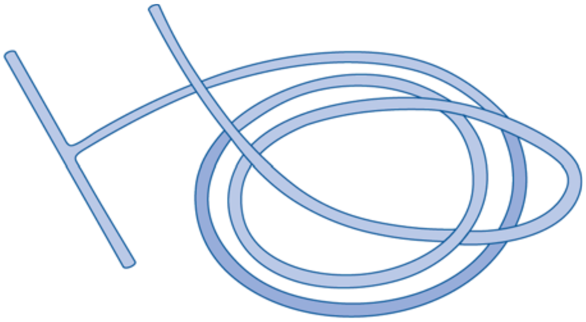

Fig. 32.41: Kehr's 'T' tube.

- 'T' tube has got horizontal part which is placed in the CBD and vertical part which is allowed to come out, to drain bile. Amount of bile draining daily is measured.

Before removal of the "T" tube, patency of CBD should be confirmed.

It is done by following methods:

- The vertical limb is clamped (done in 12-14 days) and the patient is observed for development of pain, fever and jaundice in 24 hours. If normal, then one can presume that there is no obstruction in the CBD.
- Water soluble iodine dye is injected through the tube to visualize biliary tree and free flow of dye into the duodenum. (postoperative 'T' tube cholangiogram). It is done in 14 days which is the time required to develop fibrous track. Once there is free flow, tube is removed and track gets closed on its own.

PROCTOSCOPE

(Refer Rectum and Anal Canal)

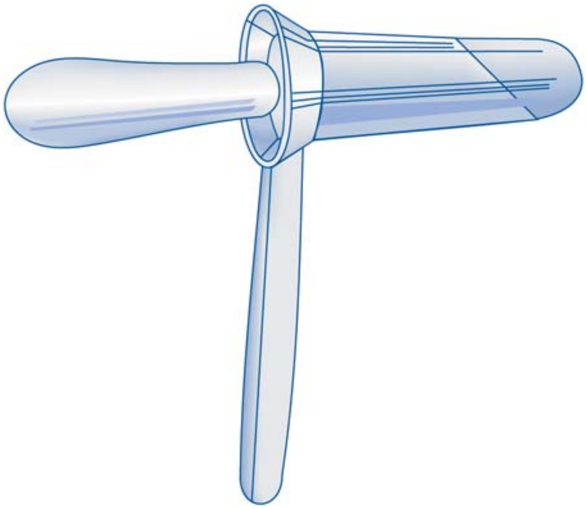

Fig. 32.42: Proctoscope.

FLATUS TUBE

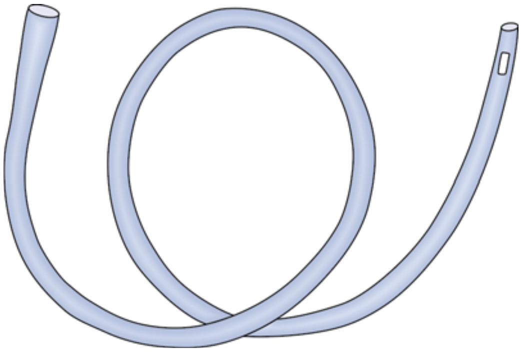

Fig. 32.43: Flatus tube

It is made up of India rubber, 45 cm in length. There is one opening in the tip and another on the side proximal to the tip. (Urinary catheter like red rubber catheter has no opening in the tip, only side opening is present). It is used in sigmoid volvulus to decompress and derotate; in paralytic ileus; in subacute intestinal obstruction. It is passed per anal into the recto-sigmoid area. Proximal end is connected to water container to observe the quantity of air bubble which signifies the amount of gas getting deflated.

C. Suture Materials

CHAPTER OUTLINE

- Classification I
- Classification II
- Classification III
- Classification IV
- Classification V

Features of ideal suture material

- Adequate tensile strength
- Good knot holding property
- Should be least reactive
- Easy handling property
- Should have *less memory*
- Should be easily available and cost effective

CLASSIFICATION I

Absorbable Suture Materials

- **Plain catgut** is derived from submucosa of jejunum of sheep.
    - It is yellowish white in colour.
    - It is absorbed by inflammatory reaction and phagocytosis—absorption time is 7 days.
    - It is used for subcutaneous tissue, muscle, circumcision in children.
- **Chromic catgut** is catgut with chromic acid salt.
    - It is brown in colour.
    - Its absorption time is 21 days.
    - It is used for suturing muscle, fascia, external oblique aponeurosis, ligating pedicles, etc.
- **Vicryl** (Polyglactic acid):
    - It is synthetic absorbable suture material.
    - It gets absorbed in 90 days.
    - Absorption is by hydrolysis.
    - It is violet in colour (braided).
    - It is multifilament and braided.
    - It is very good suture material for bowel anastomosis, suturing muscles, closure of peritoneum.
- **Dexon** (Polyglycolic acid) is synthetic absorbable suture material like vicryl. It is creamy yellow in colour (braided).
- **Maxon** (Polyglyconate) monofilament.
- **PDS (P**oly **D**ioxanone **S**uture material) is absorbable suture material. It is creamy in colour with properties like vicryl. It is costly but better suture material than vicryl.
- **Monocryl** (Polyglecaprone) monofilament.
- **Biosyn** (Glycomer) monofilament.

Uses of absorbable suture materials

- In bowel anastomosis like gastrojejunostomy, resection and anastomosis. Vicryl is used
- In cholecystojejunostomy (CCJ), choledocho-jejunostomy (CDJ), pancreaticojejunostomy. Vicryl is used
- In suturing muscle, fascia, peritoneum, subcutaneous tissue, mucosa
- In ligating pedicles. 1-zero chromic catgut or vicryl are used, e.g. ligation of pedicles during hysterectomy
- In circumcision, usually 3-zero plain or chromic catgut are used

Absorbable suture materials should not be used for suturing tendon, nerves, vessels (vascular anastomosis).

Nonabsorbable Suture Materials

- ***Silk*** is natural, multifilament, braided, non-absorbable suture material derived from cocoon of silkworm larva. It is black in colour. It is coated suture material to reduce capillary action.
- ***Polypropylene*** (Prolene) is synthetic, monofilament suture material. It is blue in colour. It has got high memory. (*Memory of suture material* is recoiling tendency after removal from the packet. Ideally suture material should have low memory.) (Prolene mesh used for hernioplasty is white in colour).
- ***Polyethylene*** (Ethylene) is synthetic monofilament nonabsorbable suture material. It is black in colour.
- ***Cotton*** is twisted multifilament natural nonabsorbable suture material. It is white in colour.
- ***Linen*** is derived from bark of cotton tree.
- ***Steel, polyester, polyamide, nylon*** are other nonabsorbable suture materials.

Uses of nonabsorbable suture materials

- In herniorrhaphy for repair
- For closure of abdomen after laparotomy
- For vascular anastomosis (6-zero), nerve suturing, tendon suturing
- For tension suturing in the abdomen
- For suturing the skin

CLASSIFICATION II

Natural

- Catgut.
- Silk.
- Cotton.
- Linen.

Synthetic

- Vicryl, dexon, PDS, maxon.
- Polypropylene, polyethylene, polyester, polyamide.

CLASSIFICATION III

- **Braided:** Polyester, polyamide, vicryl, dexon, silk.
- **Twisted:** Cotton, linen.

CLASSIFICATION IV

- **Monofilament**: Polypropylene, polyethylene, PDS, catgut, steel.
- **Multifilament**: Polyester, polyamide, vicryl, dexon, silk, cotton.

CLASSIFICATION V

- Coated.
- Uncoated.

Numbering of Suture Material

2-Thick. For pedicle ligation.
1-
0-zero.
1-zero.
2-zero. For bowel suturing.
3-zero.
4-zero.
5-zero. For vascular anastomosis.
6-zero.
7-zero.
8-zero.
9-zero. For ophthalmic surgery. Requires operating microscope.

Fig. 32.44: Photo showing different types of suture materials (with pack).

Types of suturing

- Continuous suturing
- Interrupted simple suturing
- Interrupted mattress suturing
- Subcuticular suturing
- Horizontal tension suturing
- Vertical tension suturing

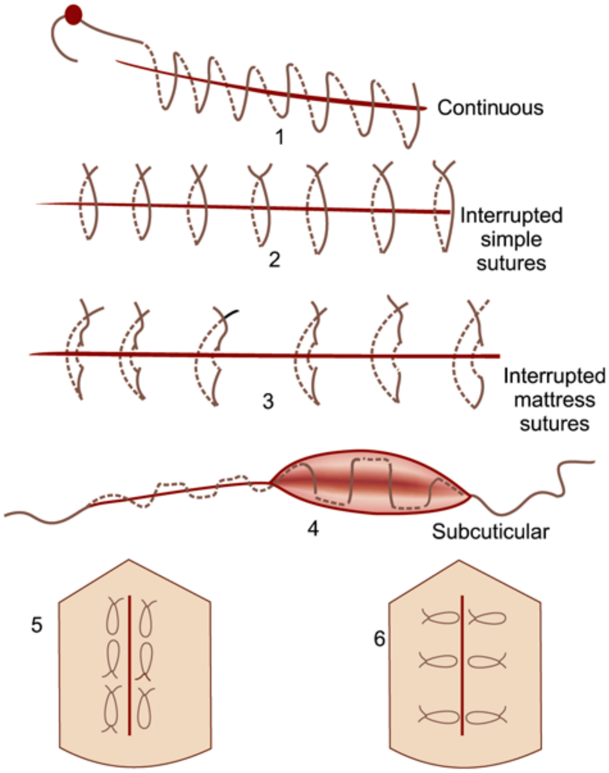

Fig. 32.45: Types of suturing.

Types of knots

- Reef knot
- Granny knot
- Surgeon's knot

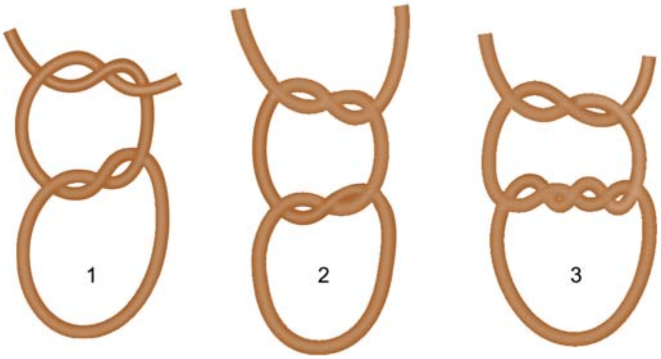

Fig. 32.46: Types of knots.

Sadness is a from of fatigue.

D. Diathermy (Electrocautery)

It is the method to control bleeding or to cut the tissues during surgery.

TYPES

Based on type of current used:

- Unipolar cautery.
- Bipolar cautery. It is safer because its effect is seen only in between electrode points. Adjacent tissues will never get damaged.

Based on type of action:

- ***Coagulation cautery*** which causes haemostasis by tissue coagulation. Here temperature is 100 degree (blue switch).
- ***Cutting cautery***: Here temperature is 1,000 degree which disintegrate the tissues. It is not haemostatic (yellow switch).
- ***Blended current*** is combination of both coagulation and cutting.

Differences Between Unipolar and Bipolar Cautery

Unipolar cautery	*Bipolar cautery*
• Can be used for both coagulation and cutting	• Only for coagulation
• Conducting plate should be kept	• No need
• Cannot be used in patient with artificial valves	• Can be used
• Should be careful about adjacent tissues	• Adjacent tissues will never get damaged

Uses

- For coagulation of bleeders during surgery to achieve haemostasis.
- To cut muscles, fascia, etc.
- It is essential for laparoscopic surgical procedures. Bipolar is commonly used.
- It is used to remove small cutaneous lesions, to control bleeding duodenal ulcer.

Disadvantages

- Infection.
- Cauterisation of normal tissues.
- Problem of explosion.
- Diathermy burn to the patient at the site where diathermy plate is kept.
- Burn injury or electrical shock to surgeon and assisting personnel.

Precautions

- Proper earthing.
- Avoid loose contact of electrodes.
- It should be kept off when not in use during procedure.

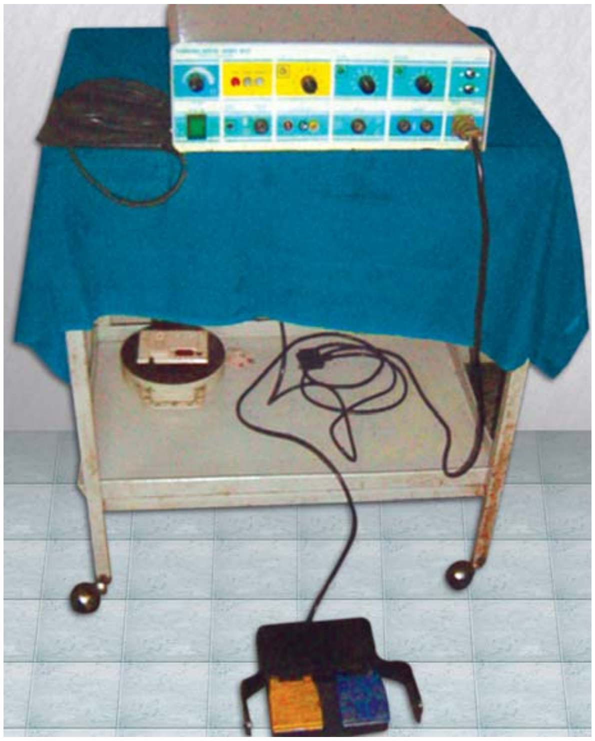

Fig. 32.47: Diathermy machine with plate, foot switch for use.

E. Operative Procedure

CHAPTER OUTLINE

- Abdominal Incisions
- Vasectomy
- Circumcision
- Hydrocele
- Inguinal Hernia
- Appendicectomy
- Thyroidectomy
- Tracheostomy
- Cryosurgery
- Lasers in Surgery
- Staplers in Surgery
- Nasojejunal Tube Feeding
- Gossypiboma

ABDOMINAL INCISIONS

Principles

- Incision should be long enough for a good exposure.
- Splitting the muscle is better than cutting, except rectus muscle.
- Avoid cutting nerves and vessels in the abdominal wall.
- Retract muscle, abdominal organs towards the neurovascular supply.
- Insert a drainage tube through a separate incision.
- Transverse incisions are better than vertical incisions.
- Close the wound layer by layer.

Requirements

- Accessibility
- Extensibility
- Security

Factors affecting the strength of the scar

- Type of surgery (acute abdomen, surgery for malignancy, major surgery)
- Obesity
- Pregnancy
- Straining
- Cough
- Ascites
- Nutrition
- Diabetes
- Immunosuppression
- Type of incision

Complications of Abdominal Incision

- Wound infection.
- Burst abdomen.
- Fistula formation.
- Wound pain.
- Incisional hernia.
- Adhesion and its complication.

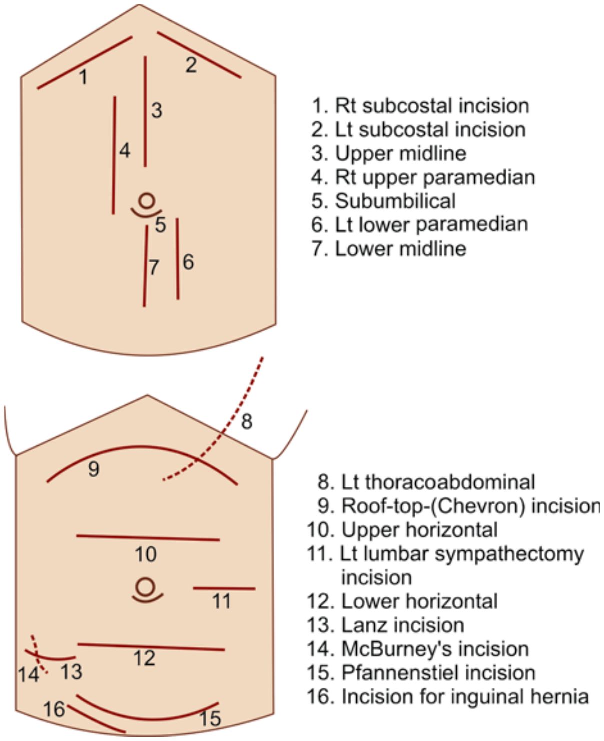
1
2
3
4
5
6
7
1. Rt subcostal incision
2. Lt subcostal incision
3. Upper midline
4. Rt upper paramedian
5. Subumbilical
6. Lt lower paramedian
7. Lower midline
8
9
10
11
12
13
14
15
16
8. Lt thoracoabdominal
9. Roof-top-(Chevron) incision
10. Upper horizontal
11. Lt lumbar sympathectomy incision
12. Lower horizontal
13. Lanz incision
14. McBurney's incision
15. Pfannenstiel incision
16. Incision for inguinal hernia

Fig. 32.48: Different incisions in the abdomen.

Different abdominal incisions are:

- Upper midline.
- Upper right paramedian.
- Upper left paramedian.
- Kocher's incision (right subcostal).
- Left subcostal.
- Bucket handle.
- Upper horizontal.
- Thoracoabdominal.
- Subumbilical.
- Incision for lumbar sympathectomy.
- Lower midline.
- Lower right or left paramedian.
- Incisions for appendicectomy—McBurney's, Rutherfold Morrison's, Lanz, laparoscopic.

Surgery is an irreversible repair but often it can be irreversible damage also !!!

- Pfannenstiel incision.
- Lower horizontal.
  - Upper incisions are always better.
  - Horizontal incisions are better.
  - Paramedian is better than midline.

VASECTOMY

Indications

- Family planning (parents should have two healthy children, consent should be obtained).
- After prostatectomy, vasectomy is done to prevent retrograde infection of testes.

Procedure

After cleaning and draping, 2 ml of 1% injection xylocaine (local anaesthetic) is injected to the root of the scrotum on the antero-lateral aspect.

Skin, external and internal spermatic fasciae are incised, cord is held with Allis forceps. Thick, firm, vas is dissected, two artery forceps are applied. A small piece of vas is excised and cut ends are ligated with silk or thread (nonabsorbable). Skin is sutured. Procedure is repeated on the other side.

Advice

To avoid sexual contact or to use contraception for 3 months.

Complications

- Haematoma
- Haematocele
- Infection
- Pyocele
- Sperm granuloma
- Recanalisation occurs rarely but dangerous

- When there is hernia or hydrocele, vasectomy is done along with specific procedures for hernia or hydrocele.
- ***No scalpel vasectomy,*** using specialised instruments is becoming popular. Procedure does not require any suturing.

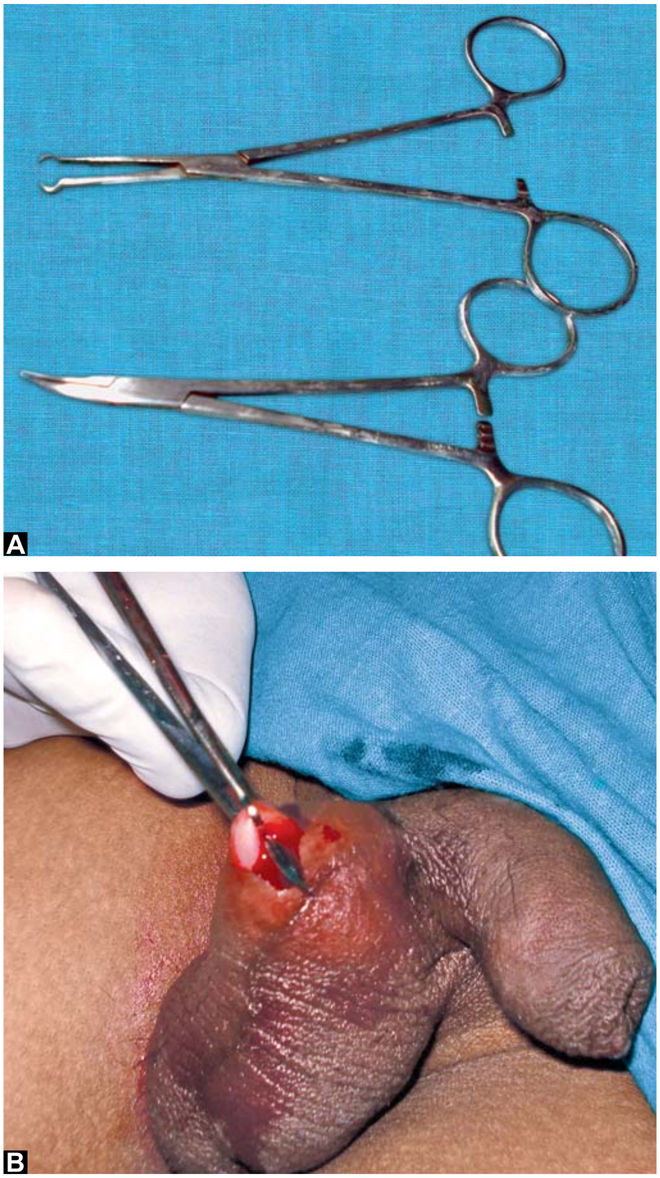

Figs 32.49A and B: No scalpel vasectomy instruments and technique.

CIRCUMCISION

Indications

- Religious.
- Phimosis.
- Paraphimosis after doing initial dorsal slit.
- Balanitis and balanoposthitis (common in diabetics).
- Early carcinoma of prepuce or glans penis—both diagnostic as well as therapeutic purpose.
- Certain sexually transmitted diseases, e.g. herpes infection.

Procedure

In children, it is done under G/A. In adults, it is done under local anaesthesia.

After cleaning and draping, LA (1% lignocaine *(plain)* is injected circumferentially near the root of the penis) is given (ring block). Dorsal skin is cut up to the corona and later circumferentially and ventrally. The skin is cut with inner layer. Care is taken to see that optimum (less) skin is cut ventrally to prevent the occurrence of *chordee*. Frenular artery as well as dorsal vein is transfixed and ligated ventrally using chromic catgut (2-0 or 3-0). Small bleeders are also ligated. Skin is apposed to the cut edge of corona using interrupted chromic catgut sutures.

Postoperatively, antibiotics and analgesics are given.

Complications

- Reactionary haemorrhage due to slipping of ligature from frenular artery and dorsal vein.
- Infection.
- Stricture urethra near the external meatus in *children.*
- Chordee due to removal of excess skin on the ventral aspect.
- Rarely priapism can occur.

HYDROCELE

Types of Surgery

- Subtotal excision.
- Partial excision and eversion (Jabouley's operation).
- Evacuation and eversion.
- Lord's plication.
- Sharma and Jhawers technique.

Procedure

- Under G/A or spinal or L/A, after cleaning and draping, vertical incision of about 6-8 cm in length is made over the scrotum, anteriorly 1 cm lateral to the median raphe.
- Skin, dartos, external spermatic fascia, internal spermatic fascia are incised.

  Bluish hydrocele sac is identified, i.e. parietal layer of the tunica vaginalis of testis.

  Fluid is evacuated using trocar and cannula. Sac is opened.
- *If the sac is small, thin and contains clear fluid,* either *Lord's plication,* i.e. tunica is bunched into a "ruff" by placing series of multiple interrupted chromic catgut sutures so as to make the sac form a fibrous tissue which is relatively avasular and so haematoma will not occur, or *evacuation and eversion* of the sac behind the testis (after eversion, everted sac is sutured with chromic catgut by continuous sutures) is done.
- *If the sac is thick, in large hydrocele and chylocele*, *subtotal excision of the sac* is done (as tunica vaginalis is reflected on to the cord structures and epididymis posteriorly, total excision of the sac leads to orchidectomy with division of cord).
- Often the sac is excised partially and eversion is done, which is called as *Jabouley's* operation.

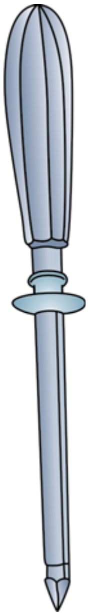

Fig. 32.50: Trocar and cannula used for tapping hydrocele. It prevents spillage of fluid.

- After evacuation, the sac with the testis is placed in a newly created pocket between the fascial layers of the scrotum *(Sharma and Jhawer's technique).*
- *Aspiration must be avoided as much as possible* as it is only a temporary measure (recurrence occurs very early) and chances of haematocele, infection are higher.
- A drain is placed near the root of the scrotum on the lateral aspect because, it becomes the most dependent portion once scrotal support is given. Scrotal support is given to reduce the scrotal oedema.
- Wound is closed in layers.
- Drain is removed in 48 hours.

Surgeries for hydrocele

- Subtotal excision of the sac
- Jabouley's operation
- Evacuation and eversion
- Lord's plication
- Sharma and Jhawer's technique

Complications of surgery

- Reactionary haemorrhage
- Infection
- Pyocele
- Sinus formation
- Recurrent hydrocele

Conditions where orchidectomy is done in hydrocele

- Pyocele with testicular destruction
- Clotted haematocele with testicular destruction

INGUINAL HERNIA

Surgery is the treatment of choice for inguinal hernia.

In infants, whether it is hernia or hydrocele, only *herniotomy* is done through inguinal approach *(Michaelis plank operation*).

In adults: It includes *herniotomy,* i.e. excision of hernial sac and *herniorrhaphy* (strengthening of the posterior wall of inguinal canal either by repair or mesh).

1. Herniotomy

*Anaesthesia***:** Spinal or G/A

Procedure: After cleaning and draping, skin is incised 1.25 cm above and parallel to the medial two-third of inguinal ligament. Two layers of superficial fascia (outer Camper's fascia and inner Scarpa's fascia) are incised. External oblique aponeurosis is incised. Upper leaf is reflected above and lower leaf is reflected downwards to visualise and expose the inguinal ligament. Ilioinguinal nerve is safeguarded. Cremasteric muscle is opened. Cord structures are dissected. Sac which is anterior and lateral to cord is identified and is pearly white in colour. Dissection is usually started from the fundus and extended towards the neck which is identified by extraperitoneal fat. The neck is narrow and is lateral to inferior epigastric artery. Sac is opened at the fundus. Finger is passed to release any adhesions.

Sac is twisted so as to prevent the contents from coming back. It is transfixed using absorbable suture material (chromic catgut 2-0) and is excised distally.

2. Modified Bassini's Herniorrhaphy

Conjoined tendon and inguinal ligament are approximated using interrupted nonabsorbable monofilament sutures [polypropylene (prolene, blue in colour)]; medial most stitch is taken from the periosteum of pubic tubercle *(called as key or Bassini's stitch*); external oblique is closed and other layers are closed.

Complications of herniorrhaphy

- Haemorrhage
- Haematoma
- Infection
- Haematocele
- Postherniorrhaphy hydrocele
- Hyperaesthesia over the medial side of inguinal canal due to injury to ilioinguinal nerve
- Recurrence
- Osteitis pubis

Hernioplasty

Prolene mesh is used for hernioplasty (white in colour). Other materials which can be used are Dacron, tensor fascia lata, skin.

Relation of sac to the cord

- Indirect inguinal hernia—anterolateral
- Direct inguinal hernia—posteromedial

APPENDICECTOMY

Approaches:

- ***Grid-iron incision*** *(McBurney's incision*): Incision is placed perpendicular to the right spino-umbilical line at the McBurney's point, i.e. at the junction of lateral one-third and medial two-third of spino-umbilical line.
- ***Rutherford Morison muscle*** cutting incision *(Cut upwards and laterally*).
- ***Lanz crease incision*** centering at McBurney's point-cosmetically better.
- ***Right lower paramedian*** incision—when in doubt or when there is diffuse peritonitis.
- ***Laparoscopic approach:*** Becoming popular.
- ***Fowler-Weir approach:*** Medial muscle cutting incision.

Procedure

Under general anaesthesia, skin is incised. Two layers of superficial fascia are cut. External oblique aponeurosis is opened in the line of incision. Internal oblique and transverse muscles are split in the line of fibres. Peritoneum is opened in the line of incision. Caecum is identified by taeniae, and ileo-caecal junction. Omentum when adherent is separated. Appendix is held with Babcock's forceps. Mesoappendix with appendicular artery is ligated. Using thread or silk, a purse-string suture is placed around the base of the appendix. Base of the appendix is crushed with artery forceps and transfixed using vicryl (absorbable). Appendix is cut distal to the suture ligature and removed. Stump is cleaned with antiseptics. Purse-string suture is tightened so as to bury the stump.

In difficult cases *Retrograde appendicectomy* can be done. In presence of pus or burst appendix, the peritoneal cavity is drained.

Postoperatively, IV fluids, antibiotics are given. Once bowel sounds are heard, oral diet is started.

Complications after Appendicectomy

- Paralytic ileus.
- Reactionary haemorrhage due to slipping of ligature of the appendicular artery.
- Residual abscess (pelvic, paracolic, local, subdiaphragmatic).
- Pylephlebitis (Portal pyaemia).
- Adhesions, kinking and intestinal obstruction.
- Right inguinal hernia *(direct)* due to injury to ilio inguinal nerve.
- Wound sepsis.
- Faecal fistula.

THYROIDECTOMY

Types

- ***Hemithyroidectomy*****:** Along with removal of one lobe, entire isthmus is removed. It is done in benign diseases of only one lobe.
- ***Subtotal thyroidectomy:*** Commonly done in toxic thyroid either primary or secondary and also often for nontoxic multinodular goitre. Here about 8 grams or a tissue, size of pulp of the finger is retained at the lower pole, one or both sides and rest of the thyroid gland is removed.
- ***Partial thyroidectomy*** is removal of the gland in front of trachea after mobilisation. It is commonly done in nontoxic multinodular goitre.
- ***Near total thyroidectomy*****:** Here both lobes, except the lower pole which is very close to recurrent laryngeal nerve and parathyroid is removed. It is done in case of papillary carcinoma of thyroid.
- ***Total thyroidectomy*****:** Entire gland is removed. It is done in case of follicular carcinoma of thyroid, medullary carcinoma of thyroid.

Procedure

Position: Under general anaesthesia patient is put in supine position, with neck hyperextended by placing a sand bag under shoulder—with table tilt of 15 degree head up to reduce venous congestion.
Incision: Horizontal crease incision is done, two finger breadth above the sternal notch, extending from sternomastoid of one side to the other.

Skin and platysma are incised—upper flap raised up to thyroid cartilage, lower flap up to sternoclavicular joint. Deep fascia is opened vertically in the midline. Strap muscles are retracted or cut in between two Kocher's forceps. Pretracheal fascia is opened to mobilise the thyroid. First, short stout middle thyroid vein is ligated, then superior thyroid pedicle is ligated close to the gland so as to avoid injury to external laryngeal nerve. Inferior thyroid artery is ligated away from the gland so as to avoid injury to recurrent laryngeal nerve. Mobilised gland is removed. Bed is sutured with catgut so as to prevent bleeding. Drain is placed. The wound is close in layers.

Thyroid steal: Patient with thyrotoxicosis is taken to operation theatre daily for few days before doing surgery so as to reduce the anxiety of the patient.

Complications of Thyroidectomy

1. ***Haemorrhage***: May be due to slipping of ligatures, either of superior thyroid artery or other pedicles. It causes tachycardia, hypotension, breathlessness and compression over the trachea which inturn may cause severe stridor, respiratory obstruction. As a first aid, immediate release of sutures including that of deep fascia has to be done and pressure over the trachea is released. Then patient is shifted to operation theatre and under general anaesthesia exploration is done and bleeders are ligated. Blood transfusion may be required.
2. ***Respiratory obstruction***: It may be due to haematoma (if it is so, the haematoma has to be evacuated), or due to laryngeal oedema. For laryngeal oedema, *immediate emergency endotracheal intubation* is done along with steroid injections. Often emergency tracheostomy may be required as a life-saving procedure.
3. ***Recurrent laryngeal nerve palsy***: It can be transient or permanent. Transient is 3% common. They usually recover in 3 weeks to 3 months. Often they require steroid supplement and speech therapy. Permanent paralysis is rare.
4. ***Hypoparathyroidism*** is rare (0.5%). Mostly it is temporary due to vascular spasm of parathyroid glands, occurs in 2nd-5th postoperative day. Presents with weakness, +ve Chvostek's sign, carpopedal spasm, convulsions. Serum calcium estimation is done and then 10 ml of 10% calcium gluconate—is given IV 8th hourly. Later supplemented by oral calcium 500 mg 8th hourly. After 3-6 weeks, patient is admitted, drug is stopped and serum calcium level is repeated.
5. ***Thyrotoxic crisis* (Thyroid storm):** Occurs in a thyrotoxic patient inadequately prepared for thyroidectomy and rarely a thyrotoxic patient presents in a crisis following an unrelated operation or stress. They present in 12-24 hours with severe dehydration, circulatory collapse, hypotension, hyperpyrexia and often cardiac failure.

   Treatment is injection hydrocortisone, oral antithyroid drugs, tepid sponging of whole body, beta blocker injection, oral iodides, large amount of IV fluids for rehydration, digitoxin. Cardiac monitor, often ventilator support with close observation is necessary. It has got high rate of mortality with critical period of 72 hours. Correction of fluid and electrolyte imbalance and cardiac monitoring are the important aspects of management.
6. ***Injury to external laryngeal nerve*** causes weakness of cricothyroid muscle leading to alteration in pitch of voice.
7. ***Hypothyroidism:*** Revealed clinically after 6 months.
8. Wound infection, stitch granuloma formation.
9. Keloid formation.

TRACHEOSTOMY

Types

- Emergency tracheostomy.
- Elective tracheostomy.

Tracheostomy Tube

- ***Fuller's bivalved tracheostomy tube***: It has got outer tube and inner tube. Outer tube is biflanged and so insertion is easier. Inner tube is longer with an opening on its posterior aspect. Inner tube can be removed and re-inserted easily whenever required.
- ***Jackson's tracheostomy tube***: It has got outer tube, inner tube and an obturator.
- Red rubber tracheostomy tube.
- PVC tracheostomy tube.

Modern tracheostomy tubes are made of plastic. They are soft, least irritant and disposable. They have inflatable cuff which makes it easier to give assisted ventilation. Cuff should be deflated at regular intervals to prevent tracheal pressure necrosis. (For assisted ventilation endotracheal tube can be kept for 7 days. Beyond that period patient needs tracheostomy for further ventilation.)

Indications for tracheostomy

- In head, neck and facial injuries
- Tetanus
- Tracheomalacia after thyroidectomy
- Laryngeal oedema/spasm/surgeries
- Major head and neck surgeries like commando's operation, block dissection, etc.
- ICU ventilation after 7 days

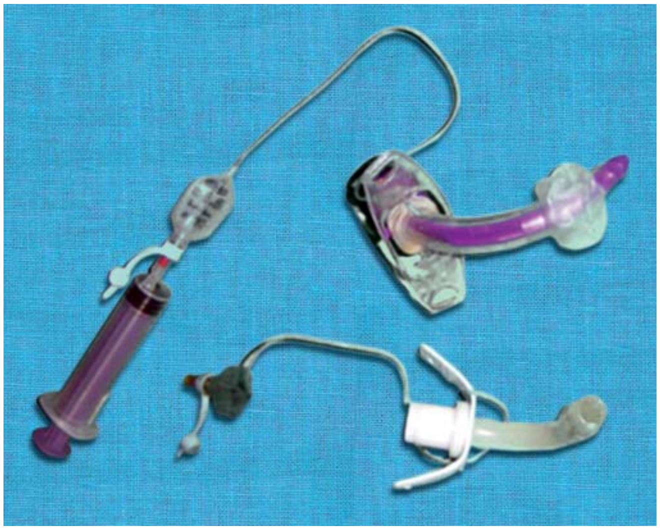

Fig. 32.51: Tracheostomy tube with inflation part and syringe (Inflated with air)

Surgery is not just cutting, but it is an art; it is not only an art but also a merciful art.

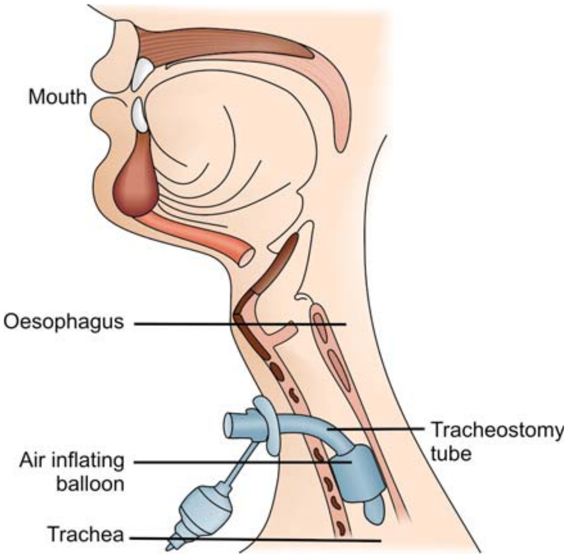

Fig. 32.52: Figure showing the position of tracheostomy tube.

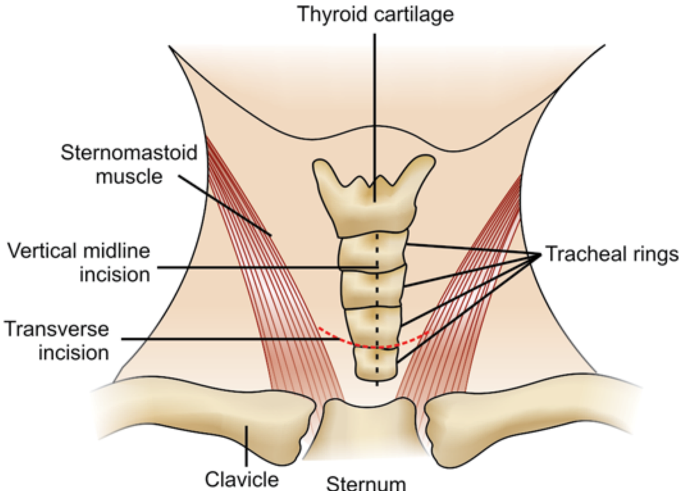

Fig. 32.53: Vertical midline or transverse incisions are used for tracheostomy. ***Vertical midline*** extends from cricoid cartilage to sternal notch. It is used both in emergency and elective tracheostomy and commonly used incision. It gives rapid access with less dissection but leads into poor scar. ***Transverse incision*** can be used in elective tracheostomy. It is placed two finger breadths above the sternal notch with a length of about 5 cm transversely. It has got a better cosmetic scar.

Technique of Tracheostomy

Neck of the patient is hyperextended by placing sand bags under the shoulder. Vertical (midline) or horizontal incision is made. Deep fascia is opened. Strap muscles are retracted laterally. Isthmus is divided or retracted below. 2nd and 3rd tracheal rings are opened and circular opening is made. Tracheostomy tube is placed. It is tied around the neck.

Note:
Endotracheal tube can be kept *in situ* only for 7 days.

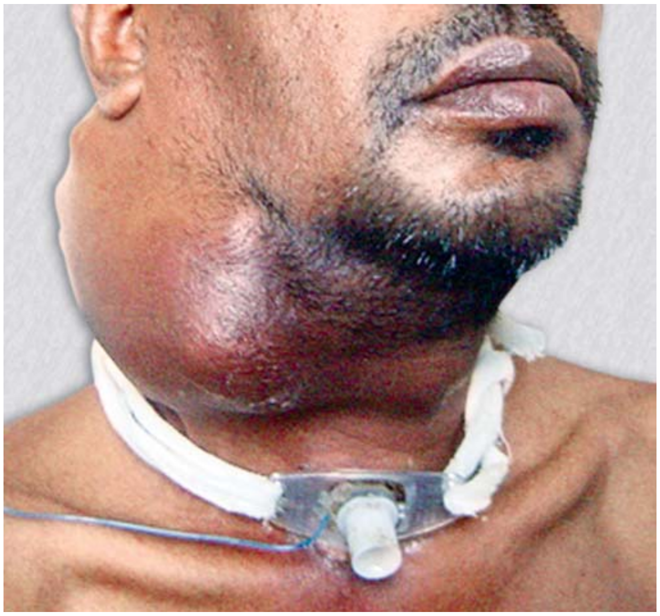

Fig. 32.54: Advanced secondaries in neck with tracheostomy tube to control respiratory stridor.

Tracheostomy Care

- Regular suctioning of the tube.
- Cleaning of tracheostomy tube.
- Humidification of the inspired air.

Complications of tracheostomy

- Tracheal stenosis
- Bleeding
- Aspiration
- Pneumothorax
- Surgical emphysema in the neck
- Mediastinal emphysema
- Tracheostomy dependency

CRYOSURGERY

- It is the destruction of tissues by ***controlled cooling.***
- System contains an automatic defrosting device with a cryoprobe.

Gases used are:

- Nitrous oxide—minus 98°C temperature.
- CO_2—minus 60°C.
- Liquid N_2—minus 180°C.
- Freon—minus 190°C.

Commonly nitrous oxide is used as it is easily available, cheaper and achieves optimum temperature required for different procedures.

Mode of Action

- It produces intracellular crystallisation, dehydration and denaturation of proteins and cell death.
- It causes the obliteration of microcirculation and so cell death.

Indications

- To remove warts and lesions in the skin.
- Cryotherapy for piles.
- For chronic cervicitis.

Advantages

- Relatively bloodless and painless.
- Adequate control of extent and depth in freezing.
- Equally effective.

Disadvantages

- Infection.
- Discharge from the site.

LASERS IN SURGERY (Light Amplification Stimulated Emission of Radiation)

Molecules are placed in a compact area and power is passed through this so as to activate the molecules. Molecules get activated at different periods and move in different directions, which they hit to each other releasing energy. This energy is allowed to act through optical system to the area wherever required.

It is named depending on the molecules used as:

- Argon Laser.
- **Neod**ymium **Y**ttrium **A**luminum **G**arnet Laser (Nd: YAG Laser).
- CO_2 Laser.
- Neon Laser.
- Holmium Laser.
- Erbium Laser.

Uses of Laser

- In *cranial surgery* in children.
- In *ENT* it is used to treat vocal cord lesions, laryngeal lesions.
- In *ophthalmology,* it is very useful in retinal surgery:
  - Detachment.
  - Iridotomy.
  - Dacryocystitis.
  - Capsulotomy.
  - To liquefy human lens.
  - Glaucomas.
- In *general surgery*:
  - In bleeding duodenal ulcer.
  - For palliative decoring of tumours in carcinoma oesophagus.
  - In Ca rectum.
  - In treatment of haemorrhoides (1st and 2nd degree).
  - In resection of bladder tumour.
  - In cervical cancer.
  - To achieve bloodless field.
  - Often in making incisions in abdomen and other places.

Advantages of Laser

- Bloodless field
- Faster
- Small lesions can be removed easily and completely

Precautions

- All reflecting instruments should be avoided otherwise laser gets reflected and injure normal tissues or the working team in the OT.
- All should wear protective spectacles to their eyes.

Disadvantage

Availability and cost factors.

STAPLERS IN SURGERY

Staplers are used for apposition of tissues. Used in skin, bowel, lungs, etc.

Types

- **Cutaneous staplers** give clean apposition. It is faster and technically easier. Problem is removal requires specific instrument and costlier than sutures.
- **Linear staplers** are used to close the bowel either completely or partially.
- **Circular staplers** also called as EEA stapler—**E**nd to **E**nd **A**nastomosis. It is commonly used for colorectal anastomosis in anterior resection for carcinoma rectum, oesophagogastric anastomosis after oesophagogastric resection in case of carcinoma at O-G junction.

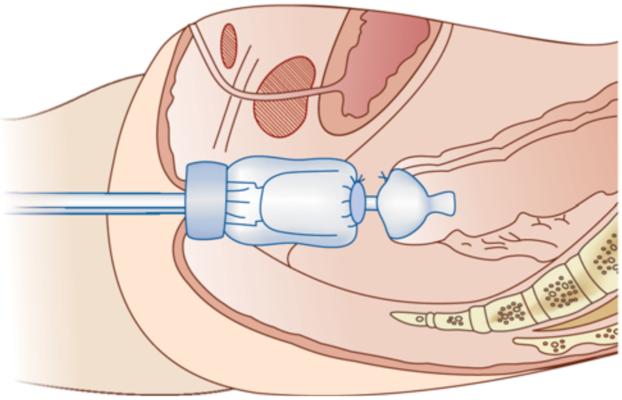

Fig. 32.55: Circular stapler for colorectal anastomosis.

> *Remember, the most important person in an operation theatre is the patient.*
>
> **—Berkehy George Andrew Moynihan**

Parts: *Stapler gun,* and *cartridge* with two rows of stapler pins for apposition. Loaded cartridge is detachable. Cut ends of bowel are placed over *gun* and *cartridge*. Once gun is shot, cartridge moves to the gun and creates anastomosis.

- **GIA stapler** (Gastrointestinal anastomosis stapler) for *side to side* anastomosis, like small bowel or ileo-colic anastomosis.
- **Stapler for lung apposition.**
- **Endostaplers:** These are used during laparoscopic surgeries. It is commonly used for bowel anostomosis. ***Endovascular staplers*** are used to ligate vascular pedicles like renal pedicles during laparoscopic nephrectomy.
- Stapled haemorrhoidopexy—costly.

Disposable staplers are available but are costly.

Advantages: Technically easier and faster.

Disadvantages: Cost factor, availability.

Problems with staplers:

- It is not completely haemostatic and so bleeding can occur.
- Leak from anastomosis, improper apposition.
- Intestinal obstruction.

NASOJEJUNAL TUBE FEEDING

- It is one of the methods of *enteral nutrition*. It is commonly used in acute pancreatitis. It is also useful in other enteral nutrition needs.

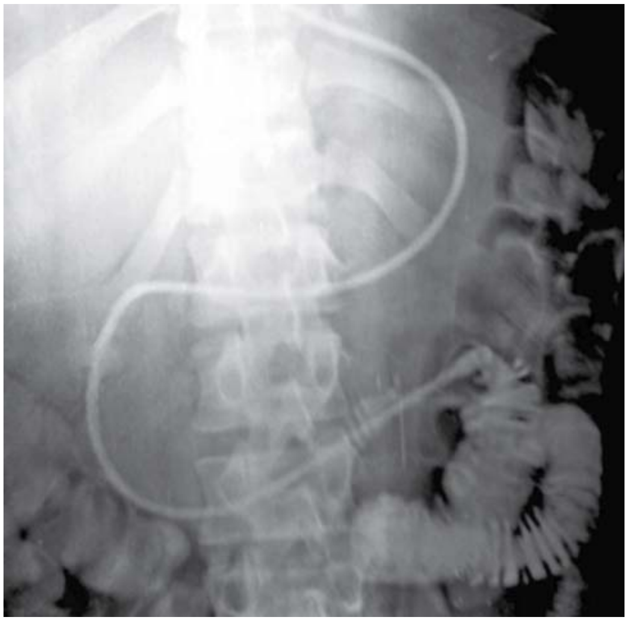

Fig. 32.56: X-ray showing nasojejunal tube for enteral nutrition purpose. It is useful and effective method of enteral nutrition. Its passage needs expertise, C-ARM guidance. It can be kept for long time for 3 months. Its position often to be confirmed by X-ray.

- It is passed under C-ARM guidance per nasally. It is passed up to the first loop of the jejunum across the duodenal C loop. It should be fixed properly to the nostril. Its position should be confirmed by X-ray. One should take care not to displace the tube.
- *Advantages* are—it is safer, easier and can be kept for 3 months. Complications of long-term TPN are not there.
- Disadvantages—irritation by tube, displacement, aspiration.

GOSSYPIBOMA (Gossypium—Cotton Based in Latin)

- By definition, it is the presence of *cotton based foreign body* that is in place of concealment following surgery.
- Forgotten foreign bodies (mop, gauze, etc.) intraoperatively especially in *abdominal cavity* (can occur in any cavities (thorax, pelvis) cause adhesions, provoke sepsis, often get encapsulated.
- It may cause an inflammatory mass, bowel erosion/perforation/peritonitis, intra-abdominal abscess, septicaemia, fistula formation.
- *Presentations* may be as—asymptomatic, pseudotumour, abscess, septicaemia, fistula.
- Often foreign body erodes and enters the bowel lumen and with peristalsis reaches ileocaecal valve causing intestinal obstruction. There are incidences patient has passed the foreign body like mop per anally few months after surgery (in 6% gossipyboms).
- Ultrasound, *CT scan* and MRI identifies the foreign body (gossipyboma). Plain X-ray is of less value.
- Incidende of gossipyboma is 1 in 3,000 surgeries. Migration commonly occurs into intestine either small or large but can occur into urinary bladder or stomach.
- 70% of retained foreign bodies are sponges/mops; 30% are instruments.
- *Commonest site* is abdominal cavity 55%, vagina (20%), thorax (10%) and rest on other cavities.
- Commonly gossipyboma present 3-12 weeks; rarely it can present as later as 5-7 years.
- It is more commonly observed in emergency surgery, truma, and surgery for malignancies.
- *Treatment:* Surgical exploration and extirpation of the foreign body, antibiotics.
- *Legal problem:* Gossypiboma amounts for criminal negligence. Surgeon or team or hospital can be sewed for negligence either in consumer court or criminal court.

F. Laparoscopic Surgery

CHAPTER OUTLINE

- Advantages of Laparoscopic Surgery
- Laparoscopic Cholecystectomy
- Laparoscopic Appendicectomy
- Advanced Laparoscopic Surgeries
- Diagnostic Laparoscopy
- Retroperitoneoscopy
- Natural Orifice Transluminal Endoscopic Surgery

History

First laparoscopic cholecystectomy was done by Muhe of Germany in 1985 and by Mouret in Lyon in 1987.

McKeran and Saye performed the first laparoscopic cholecystectomy in USA in 1988.

First laparoscopic appendicectomy was done by Semm as prophylaxis.

First laparoscopic appendicectomy for acute appendicitis was done by Schreiber in 1987.

Semm changed 75% open gynaecological surgeries into laparoscopic surgeries.

Prof TE Udwadia, Mumbai did first laparoscopic cholecystectomy in India.

ADVANTAGES OF LAPAROSCOPIC SURGERY

- Relatively less painful compared to open surgery. Trauma of access is very less.
- Shorter hospital stay and early return to work.
- Faster postoperative recovery.
- Better visualisation of the anatomy, i.e. better approach for dissection and visualisation of other parts of abdomen for any other pathology.
- Instrumental access to different abdominal locations is many times better compared to open method.
- Minimal scar on the abdomen.

Instruments Used

- Zero degree laparoscope is commonly used. Side viewing scopes are also used to have better visualisation 30°.
- Cold light source either halogen lamp or xenon lamp is used. Halogen lamp is used commonly and is cheaper. Xenon lamp gives high visualisation.
- Camera: 3 chip camera is commonly used with high resolution.
- Video-monitor to display images.
- CO_2 insufflator.
- Long fine dissectors like in open surgical techniques.

Figs 32.57A and B: Laparoscopic set showing trolley, telescope and monitor instruments

- Hooks and spatulas are used along with cautery for dissection.
- Clip applicators.
- Needle holders.
- Endostaplers.
- Veress needle.
- Suction-irrigation apparatus.
- Trocars of different sizes—10 mm, 5 mm.
- Reducers to negotiate smaller instruments through larger ports.

Preparation

Always general anaesthesia. Other preparations are same as for open method.

Technique

Pressure bandages are applied to both legs to improve the venous return and to decrease the stasis.

Head end of the table is lowered to have easier insertion of veress needle and scope.

Ryle's tube and Foley's catheter are essential before insertion of the trocars.

Pneumoperitoneum is created using veress needle through umbilical incision. Access can be achieved by open method through an umbilical incision.

CO_2 is commonly used to create pneumoperitoneum as:

- It is readily available
- It is cheaper
- It suppresses the combustion
- It is easily absorbed by tissues
- It has a high diffusion coefficient
- It is quickly released via respiration

Other gases used are: Air, nitrous oxide, helium, argon.

Pneumoperitoneum is created upto a pressure of 15 mmHg which distends the abdominal cavity adequately to have proper visualisation of the abdominal contents.

Laparoscope is inserted through the umbilical port (10 mm). Abdomen is evaluated for any pathology. Liver, gallbladder, pelvic organs are visualised.

Additional ports (3-4) through trocars are placed depending on the procedure to be done. It may be either 5 mm port or 10 mm port. These ports are placed in such a way to have a proper triangulation of instruments for dissection.

To use clip applicator 10 mm port is required.

Physiologic Changes due to Pneumoperitoneum

- CO_2 causes hypercarbia, acidosis and hypoxia.
- Pneumoperitoneum exerts pressure on the IVC, decreases the venous return and so the cardiac output.
- It increases the arterial pressure also.
- It compromises the respiratory function by compressing over the diaphragm impairing the pulmonary compliance.

Complications

- CO_2 narcosis and hypoxia.
- Sepsis—subphrenic abscess, pelvic abscess, septicaemia.
- IVC compression.
- Bleeding.
- Leak from the site, e.g. bile leak.
- Organ injury during insertion of ports, e.g. major vessels, bowel, mesentery, liver.
- Subcutaneous emphysema and pneumomediastinum.
- Gas emboli, though is rare but fatal.
- Postoperative shoulder pain due to irritation of diaphragm.
- Cardiac dysfunction due to decreased venous return.
- Injury to the abdominal wall vessels and nerves.
- Cautery burn to abdominal structures.
- Abdominal wall hernias.
- Wound infection.
- Mortality—0.5%.

Relative Contraindications

- Patients with compromised cardiac status.
- Peritonitis.
- Previous abdominal surgeries.
- Bleeding disorders.
- Morbid obesity.
- Third trimester pregnancy.
- Portal hypertension.

Basic Laparoscopic Surgeries

- Laparoscopic cholecystectomy.
- Laparoscopic appendicectomy.

LAPAROSCOPIC CHOLECYSTECTOMY

It is becoming popular method of treatment.

Indications

- Gallstones—symptomatic
- Cholecystitis
- Biliary colic

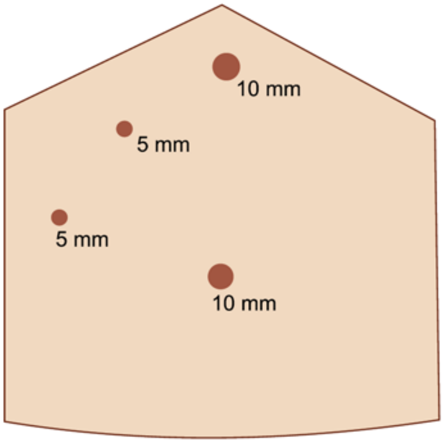

Fig. 32.58: Portals for laparoscopic cholecystectomy.

Relative Contraindications

- End stage cirrhosis, ascites or portal hypertension.
- Cholangitis: Cholecystectomy should be done after the control of cholangitis.
- CBD stones: Here, initially ERCP and stone extraction is done from CBD, then laparoscopic cholecystectomy is done.

Technique

After pneumoperitoneum, patient is placed in head up and slight left tilt position so as to make bowels to fall below and towards the left side. One 10 mm trocar is placed at umbilicus and through this umbilical port, laparoscope is passed. One 10 mm port in the epigastric region and two 5 mm ports in the right subcostal line are placed for grasping the gallbladder and for dissection. Initially, through the working channel gallbladder is held and Calot's triangle is dissected. Cystic duct and cystic artery are clipped.

An intraoperative cholangiogram is done with C-arm. Through the epigastric port, clips or ligatures are applied to the cystic duct and cystic artery, close to the gallbladder. Care should be taken to avoid bleeding and not to injure or clip the CBD or hepatic ducts. Gallbladder is separated from its bed using cautery and spatula and removed through the epigastric port. Abdomen may be drained. Patient is discharged after 48-72 hours.

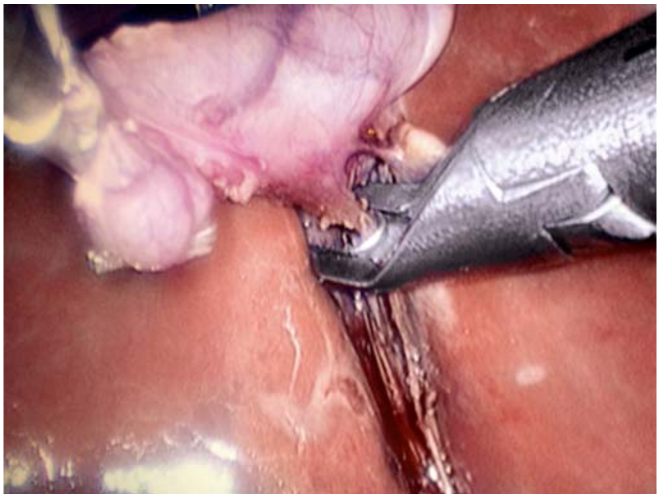

Fig. 32.59: Laparoscopic cholecystectomy. Applying clip to cystic duct is shown.

Complications

- ❖ CBD injury
- ❖ Bile leak
- ❖ Haemorrhage
- ❖ Postoperative jaundice
- ❖ Subphrenic and other intra-abdominal abscess
- ❖ Septicaemia

When problem arises, one should be ready to convert into open cholecystectomy. Conversion rate to open cholecystectomy is 2-10%. It is indicated when there is uncontrolled bleeding, dense adhesions, suspected CBD injury, when anatomy is indistinct.

When required one should not be hesitant to do conversion.

LAPAROSCOPIC APPENDICECTOMY

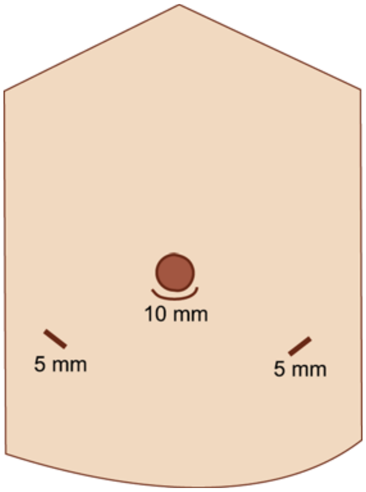

Fig. 32.60: Ports for laparoscopic appendicectomy.

Indications

Acute appendicitis. Here main advantage is confirmation of the diagnosis. Other parts of the abdomen are also visualised.

Relative Contraindications

Appendicular mass and abscess.

Technique: Laparoscope is passed through the umbilical port. Two additional ports are placed, one in lower midline (5 mm), another at right lumbar region. Mesoappendix is clipped or cauterized, using bipolar cautery. Appendix base is clipped or ligated using Roeder knot and ligature.

Complications

- Appendicular stump leak.
- Pelvic abscess.
- Bleeding.
- Injury to caecum, ileum.

ADVANCED LAPAROSCOPIC SURGERIES

- Presently most of the abdominal surgeries can be done through laparoscopy.
- It requires advanced technology and skill. Surgeon should be expert in doing intracorporeal and extracorporeal knotting.

Procedures

- Laparoscopic hernia repair.
- Laparoscopic splenectomy.
- Laparoscopic fundoplication.
- Laparoscopic vagotomy and gastrojejunostomy.
- Laparoscopic Nissen's fundoplication.
- Laparoscopic colectomy.
- Laparoscopic hysterectomy. It is becoming very popular.
- Laparoscopic urologic surgeries.
- Laparoscopic paediatric surgeries.

Laparoscopic Inguinal Hernia Repair

It is becoming popular method, lately. It is a skilled laparoscopic surgery.

1. *Transabdominal preperitoneal repair (TAPP repair)*: Through abdomen, using laparoscope Hesselbach's triangle is exposed and mesh is placed in the preperitoneal space. Peritoneum is sutured back or stapled.
2. *Totally extraperitoneal repair (TEP repair)*: Through subumbilical incision, preperitoneal space is created with the help of a balloon. Laparoscope is passed to this space. Inguinal canal is dissected and mesh is placed.

Triangle of 'Doom': It is bounded by ductus deference medially, testicular vessels laterally, with apex at internal ring. This is dangerous area in laparoscopic hernia repair as dissection may injure iliac vessels and cause torrential haemorrhage.

When light pressure is exerted over area of surgical emphysema, a sensation similar to that of likewise pulpating a horse hair mattress is experienced. **—Denis Dooley**

Complications

- Subcutaneous emphysema.
- Abdominal wall haematoma.
- Injury to major vessels.
- Recurrence.

DIAGNOSTIC LAPAROSCOPY

Indications

- Acute pelvic conditions.
- Tubal pregnancy.
- Ovarian diseases.
- Infertility.
- Staging of the malignancy.
- Biopsy from the tumours.
- In chronic pain abdomen where ultrasound, endoscopies, barium studies are negative, then diagnostic laparoscopy is useful.

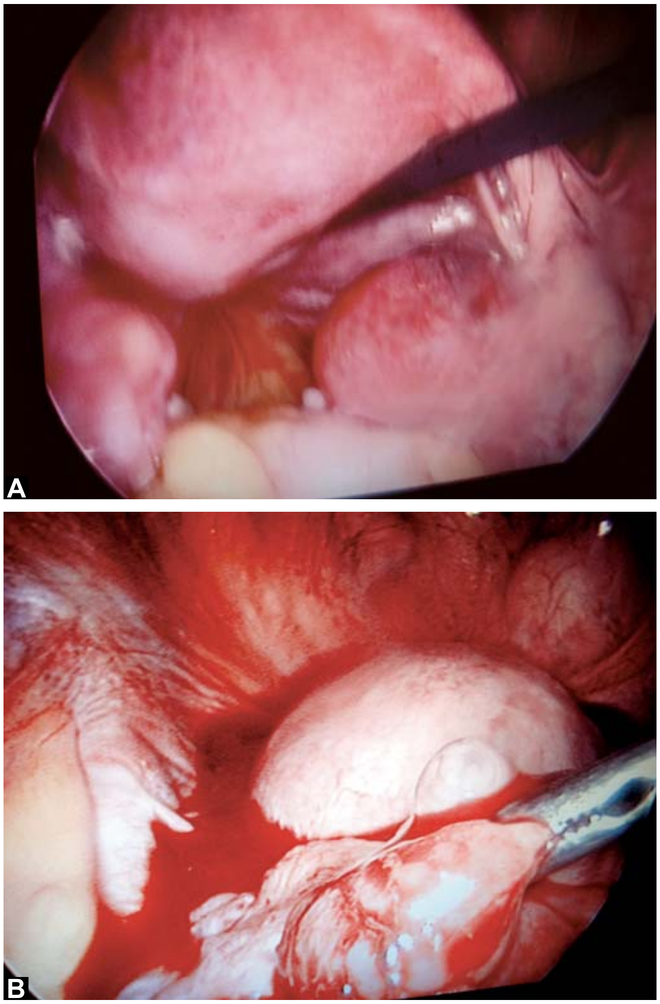

Figs 32.61A and B: Diagnostic laparoscopy showing ectopic pregnancy in right fallopian tube. It is removed by salpingectomy through laparoscopy.

Needle laparoscopy of 2 mm sized becoming popular (especially for diagnostic purpose).

Advantages

- Laparotomy is avoided.
- Once diagnosis is made, therapeutic procedure also can be carried out in the same sitting.

RETROPERITONEOSCOPY

- It is becoming popular in urology to assess kidney, ureter, adrenals for various urologic procedures.
- Through a small loin approach, retroperitoneum is expanded by inflating balloon in the space. Once space is created, different ports are placed to do dissections.

Procedures done through retroperitoneoscopy are:

- Nephrectomy
- Pyeloplasty
- Adrenalectomy
- Pyelolithotomy
- Uretero-lithotomy
- Retroperitoneal lymph node dissection (RPLND)

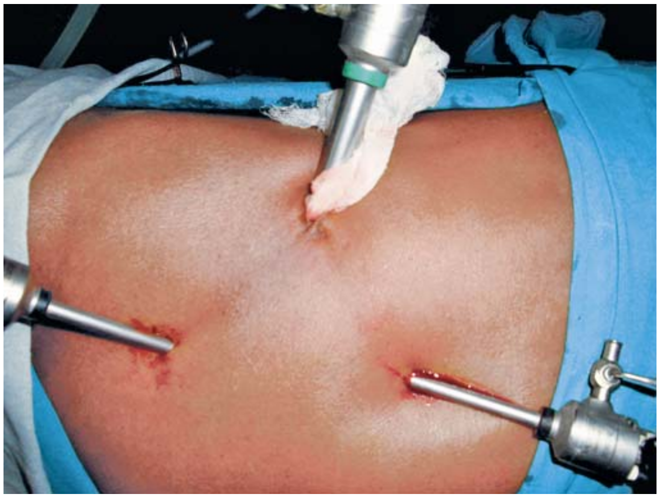

Fig. 32.62: Port positions for retroperitoneoscopy.

Complications

- Injury to vessels.
- Paralytic ileus.
- Bowel (colon) injury.

Advantage

- Complications of pneumoperitoneum is not present and so respiratory reserve is well-maintained.

NATURAL ORIFICE TRANSLUMINAL ENDOSCOPIC SURGERY (NOTES)

It is an experimental surgical technique whereby "*scar less" abdominal operations* can be performed with an *endoscope passed through a natural orifice* (mouth, urethra, anus, etc.) then through an internal incision in the stomach, vagina, bladder or colorectum, thus avoiding any external incisions or scars.

- This technique has been used for diagnostic and therapeutic procedures in animal models, including transgastric (through the stomach) organ removal. The transvesical and the transcolonic approaches are also used. Transgastric and transvesical combined approach is also used to increase the feasibility of moderately complex procedures such as cholecystectomy.
- NOTES was originally described in animals by researchers at Johns Hopkins University (Dr Anthony Kalloo, et al.), and was recently used for transgastric appendectomy in humans in India (by Drs GV Rao and N Reddy).
- On June 25 2007 Swanstrom and colleagues reported the first human transgastric cholecystectomy. In late 2008 surgeons from Johns Hopkins School of Medicine removed a healthy kidney from a woman donor using NOTES. The surgery was called transvaginal donor kidney extraction.
- The ***transvaginal access*** to NOTES seems to be the *safest and feasible*. In 2007, the NOTES Research Group in Rio de Janeiro, Brazil, lead by Dr Ricardo Zorron, performed the first series of transvaginal NOTES cholecystectomy in four patients. With fewer potential complications, the procedure has a disadvantage of being possible only in women.
- Proponents and researchers in this field recognize the potential of this technique to revolutionize the field of minimally invasive surgery by eliminating abdominal incisions. NOTES could be the next major paradigm shift in surgery, just as laparoscopy was the major paradigm shift during the 1980s and 1990s.
- **Advantages are**—lower anaesthesia requirements; faster recovery and shorter hospital stays; avoidance of the potential complications of transabdominal wound infections (e.g. hernias); less immunosuppression; better postoperative pulmonary and diaphragmatic function; and the potential for “scar less” abdominal surgery.
- **Disadvantages are**—it is single port surgery and difficulty in visualisation of the area in need from all directions which is essential for proper surgical dissection. Poor manoeuverability is the problem. Sepsis through this potentially infected area into sterile peritoneal cavity is a real risk. After procedure non closure of the port site opening in these approach sites or if closing their inadequacy are the real risk in NOTES.
- The general impression is that NOTES will be accepted as the newest frontier in minimally invasive surgery. As of today non-Bariatric minimally invasive surgery fellowships offer the best opportunity to train in this new approach.

G. Dressings and Bandages

CHAPTER OUTLINE

- Dressings
- Bandages

DRESSINGS

They are the materials used to cover wounds, ulcers to provide support and to encourage healing.

Advantages

- It covers the wound and so prevents further contamination.
- It gives comfort to the patient.

Disadvantages

- It may get soaked.
- It may delay the epithelial layer formation.

Types

- *Dry dressings:* It is used in clean, sutured operated wound. It is not changed at regular intervals.
- *Wet dressings:* It is used in ulcers and wounds. Dressings are made wet by using jelly or sofra tulle sheets.

Components of Dressing

- *Inner contact layer.* It is non-absorbent and only allows secretion to pass into the absorbent layer. It does not allow penetration of granulation tissue. It is usually kept wet. Either mesh gauze or sofra tulle is used.
- *Intermediate absorbent layer* made up of cotton which absorbs the secretions.
- *Outer layer as supportive* is made up of gauze.

Dressings are fixed to the place by:

- Bandages
- Plasters
- Dynoplast
- Crepe bandages.

Depending on the condition and amount of discharge from the wound/ulcer it is changed as required—twice a day/once a day/once in two days.

Small dressings are done without anaesthesia.

Large areas like burn wounds or dressing in children require general anaesthesia.

BANDAGES

Technique of bandaging is called as ***Dysmergia***.

Indications

- To reduce the swelling like in lymphoedema
- To keep dressings in position
- To support splints
- To stop bleeding/oozing

Types of Bandages

a. *Roller bandages*
  - It is a continuous roll of material, which is rolled over the part to cover the area.
  - It is used in limbs.
  - It is available in different lengths and widths—1 inch, 2 inches, 4 inches or 6 inches.
  - It is used in different ways.
    - *Circular turns:* Continuous rolls placed over the same place.
    - *Spiral turns:* After the initial turn of the bandage, it ascends proximally overlapping the distal 2/3 of the previous turn.
    - *Reverse spiral turn:* Here each spiral turn is reversed in opposite direction so as to attain uniform pressure. It is used in limbs and areas which end as cones.
    - *Figure eight turn:* It is used in knee, elbow, wrist, ankle and for clavicle.
    - *Recurrent turn:* It is used in head, amputation stump. Here initially circular rolls are made and over that half turns are made to cover other parts of the area required.
    - *Spica bandage:* It has got ascending and descending turns, with each turn overlap and cross each other. It is used in hip, groin, shoulder, breast or thumb. *Spica means eye of a bean.*

b. *'T' bandages:* It is used in perineum and groin.

c. *Tailed bandages*: It may be four tailed bandages or many tailed bandages. It is used to support dressings on a wide area like in burns dressing, over abdomen or chest wall.

d. *Tubular bandages:* These are stockings which are unrolled over the limb to give pressure effects. It is used in lymphoedema, varicose veins and in the postoperative period following surgeries of the limb (Tubifix, Tubipress).

e. *Triangular bandage:* These are used for supporting the elbow or forearm. Here a wide gauze is used to cover the arm, forearm and elbow, which again winds around the neck.

f. *Cravat bandages:* It is a folded type of triangular bandage, which is used as sling around the neck, when elbow requires to be rested.

Principles of Bandaging

- Bandage is applied to the part from distal to proximal end.
- Proper positioning of the limb is a must before bandaging.
- After initial few circular turns, the required type of bandaging is then done.
- Bandage is unrolled outwards.
- During bandaging, latter turn should overlap 2/3rd of earlier turn.
- Firm, adequate pressure should be used during bandaging.
- After completing the procedure, the knot should not lie over the area or over the bony points or over the back.
- It should not cause venous or arterial compression.
- Digits should be left open and circulation in the digits should be observed for:

Anatomical place	Types of bandage
Finger/toe	One inch
Arm	Two and half inches
Leg	Four inches
Thigh	Six inches
Trunk	Six inches
Head	Four inches

H. Day Care Surgery

CHAPTER OUTLINE

- Day Care Surgery
- Surgical Audit
- Surgeon and Law

DAY CARE SURGERY

- *Day care surgery* is discharge within 23 hours (USA); surgery done without night stay (UK).
- ***Day care surgery** means* patient is fit to return home in 23 hours usually with overnight stay. *Ambulatory surgery means* patient recovers after surgery and returns home on the same evening. *Office surgery means* patient recovers from surgery and returns home in few hours. *Outpatient surgery* is different from day care surgery in that, patient is not previously fully assessed in outpatient surgery. Only minor procedures are done in this. Patient is not admitted in outpatient surgery. In day care and ambulatory surgery patient is admitted in the hospital.
- ***Day care surgery*** has been defined by *the Royal College of Surgeons* as when the surgical day case patient is admitted for investigation or operation on a planned non-resident basis and who nonetheless requires facilities for recovery. This definition excludes upper and lower GI endoscopies, outpatient procedures such as flexible cystoscopy, and minor superficial surgery under local anaesthetic, none of which require full day case facilities for recovery.
- *Day care surgery* is an upcoming field in surgical practice. It is an unique method wherein general practitioner, nurse at day care ward and theatre, surgeon, anaesthetist work in hand so that hospital stay and so the cost is reduced.
- Patient comes to hospital at morning for surgery and leaves the hospital on same day evening.

Procedures done in day care surgery

- Excision of cysts, lipomas, bursae, neurofibromas.
- Lymph node biopsy, haemorrhoid surgery, endoscopies, circumcision, hydrocele
- Wound suturing, toe amputation

Advantages

- Minimal hospital stay
- Becomes cheaper
- Patient acceptance

- *Contraindications* for day care surgery are—age > 70 years; high risk cardiac and respiratory patients; patients with bleeding disorders.

Basic Requirements are:

- In house anaesthesiologist; recovery room; theatre and recovery room/ward nurse; all essential surgical set up including monitor, ventilator.
- *Alertness; ambulation; analgesia and alimentation—basis for* day care of surgery.

Assessment Done Prior to Surgery are by:

- *Pre anesthetic clinics* with system evaluations; *health questionnaire* by surgeon and physician; *telephonic interviews*.
- Patient selection, patient information, patient acceptance are important parts in day care surgery.

Selection Criteria are

American Society for Anaesthesiologist (ASA) category I and II patients can be taken up for day care surgery. ASA III/or beyond are contraindicated for day care surgery.

ASA (American Association of Anesthesiologists) grading of the patient for surgery

- Normal individual
- Mild-moderate systemic disease—diabetes and hypertension under control
- Severe systemic disease—uncontrolled diabetes and hypertension
- Incpacitating systemic disease
- Moribund status

Class E—emergency surgery

Exclusion criteria for day care surgery are:

- ASA grade beyond III or more
- Obesity (BMT > 35). Hypertension—not controlled
- Surgery requiring more than one hour
- Surgery with anticipation of major fluid/blood loss or needs post operative critical care
- Preterm babies and infants less than 3 month's age
- Patient living in far and not easily reachable or able transport easily
- Unstable psychiatric illness
- If proper caregiver is not available
- Uncontrolled diabetes, alcohol abuse, COPD, severe asthma, epilepsy
- Pregnancy

Problems

Postoperative nausea and vomiting (50%); postoperative pain and postoperative drowsiness / dizziness (50%) are the *common complications* in day care anaesthesia.

Precautions

- Patient should be assessed properly before sending to day care surgery.
- The nurse should give proper instruction to the patient as patient stays in the hospital for a short period.
- Patient should be warned about possible problems like bleeding, vomiting, pain, discomfort, and sedation.
- Before discharging, patient should be seen by the doctor for the fitness.
- All records should be carefully documented.
- Patient should be advised to rush to hospital if any problems arise or to communicate immediately.
- Now hernia; small gynaecology procedures; ENT, cataract surgeries are done as day care procedures.

Nurses hold an important role in day care surgery.

SURGICAL AUDIT

'Clinical audit' is a process used by clinicians who intend to improve the patient care. The process involves comparing various aspects of patient care that includes *structure, process and outcome* against *explicit criteria*.

Aspects of Patient Care

- Structure—includes what is there in that place—the people in place, their training and knowledge, the equipments and facilities provided, the organization, management and their payment, etc.
- Process—includes what procedure is followed in that place in managing referred patients, what antibiotics used, what diagnostic tests done, use of ICU facilities, use of postoperative rehabilitation care, what procedure used for discharge of patients, etc.
- Outcome—includes the overall results that include the morbidity, mortality, readmission, improvement /deterioration of the patient's condition.

Explicit Criteria

Proposal for changes can be made in the care of the patient if it falls short of the criteria chosen which can be undertaken at one or more levels:

- Individual level—more training can be given to the doctors.
- Infrastructure—upgrading of the newer diagnostic tools.
- Team level—nurses getting more trained in handling the procedures along with the doctors.
- Institution—change in the treatment strategy, or antibiotic policy.
- Regional level—providing a good referral centre with all facilities and trained personnel.
- National level—introduction of screening programmes and health campaigns.

Surgical audit is a systematic, critical analysis of the quality of surgical care that is reviewed by peers against explicit criteria or recognised standards, and then used to further inform and improve surgical practice with the ultimate goal of improving the quality of care for patients.

In surgical practice, there will be definitely variations in the results of the surgery done by a trainee or an experienced surgeon, variations in outcome of the operation done in peripheral setup and in referral institutions, usage of modern equipments and technique used.

Step 1
Determine scope

It should be clearly defined, otherwise results in ineffective/ inappropriate data collection.

It should also be relevant, easily measurable.

Common areas in scope of an audit include—duration of hospital stay/unplanned admissions/readmissions/operative specific complications/30 days mortality /morbidity/investigations done/management strategy/patient satisfaction.

Step 2
Select standards

Standards for the selected topic/practice area is decided based on relevant information obtained from:

- Evidence based research and guidelines.
- Local guidelines for local relevance.
- New guidelines developed based on references from a library.

The standard which was already existing or developed must be clearly described, measurable, specific, and realistic.

Step 3
Collect data

It is important aspect of the audit which has to be informative for the audit to be successful. The best quality data collected depends on by whom it is being collected; when—retrospective/prospective collected; how—on form/PDA/computer; at/

Surgical Audit Cycle

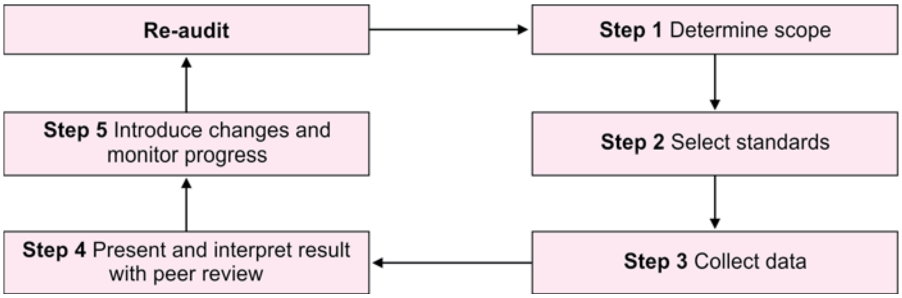
Re-audit
Step 1 Determine scope
Step 5 Introduce changes and monitor progress
Step 2 Select standards
Step 4 Present and interpret result with peer review
Step 3 Collect data

fter the time of surgery; follow-up data when collected; patient identification in a prospective/retrospective study.

Collected data must be relevant to the objectives of the surgical audit. Sometimes, the standards may need to be expanded or reduced/or the data collection methods may need to be modified.

Step 4

Present and interpret result with peer review

Audit aims in continuous improvement by experience and by making changes which is ultimately rewarding.

The outcome of the audit should be presented and discussed in a clinical meeting. It should undergo peer review. It involves viewing and analyzing one's outcome by one's own peers who are none other than other experienced trained surgeons. It should be conducted in an atmosphere of confidentiality, trust and teamwork, should not be an opportunity to blame or brag but exchange of frank, non-confrontational discussions between the colleagues. Mortality/morbidity meetings, grand rounds are one form of peer review.

Step 5

Introduce changes and monitor progress

Based on the conclusion of the audit and meetings certain changes to be made in respect to the patient care are decided and all the personnel involved in the process are informed or educated. The outcome due to changes made are monitored by follow-ups either by reauditing the whole process/ or only the part that has been changed.

SURGEON AND LAW

- It is important to a surgeon to know legal aspects in relation to his profession. Consumer Protection Act and criminal negligence are the two things surgeons are regularly worried about and face often problems.
- It is better to have a fair idea about Consumer Protection Act in relation to patient treatment.
- Surgeon should keep all documents regarding the patient with him or in the hospital.
- Case sheet should be written in detail. Daily follow-up should be written with date and time of visit with progress about the patient.
- It is better to take detailed consent after proper explanation about the disease and treatment protocol to patient and his close attender/relative. It is better to get signature about discussion given from them with date and time. In many centers, it is practiced to record the explanation part to keep it as document.
- Surgical method, its problems, risks due to anaesthesia, high-risk if any, risk of bleeding, complications, duration of hospital stay should be discussed.
- One should make sure that anaesthetist will do preanaesthetic check up prior to surgery; he should also write his preoperative/operative/postoperative anaesthetic notes.
- Daily information sheet should be used wherein patient or party should be informed about the condition of the patient. Timing of this and signature of surgeon and party should be taken.
- After surgery detailed surgical procedure technique should be written in case sheet and should be informed to patient. Specimens should be shown to patient party and should be sent for histology.
- Approximate cost of the procedure and entire bill in the hospital should be informed. One should also inform that it may change depends on complications, number of days in ICU, critical care, need for higher antibiotics, etc.
- Negligence about retaining mops/instruments are legally not acceptable; it is better to take care of enough precautions about that.
- Surgeon has got vicarious liability about the mistakes done by ward boys, nurses, theatre nurses, etc. So it is better to train them for proper care in OT, postoperative wards and ICU.
- It is ideal to show all reports to patient party and discuss/ brief with them about the condition especially when patient is in ICU.
- If patient or party become arrogant or aggressive it is better to make a note of it in case sheet and inform police people about the same.
- It is better to make a professional indemnity insurance policy always to cover these problems in case if needed.
- It is again ideal to have an advocate to discuss these matters whenever needed.
- It is care which surgeon gives not cure always.

Chapter 33 Miscellaneous

A. Fascinating Signs in Surgery

> *"If it is a question of doubt in diagnosis, you may often observe that one man solves the doubt when the others could not, and the way in which one man happened to solve it is this: he applied to the diagnosis of the case some method of examination which others had not applied."*
>
> **—Charles Barrett Lockwood,** 1856-1914

SIGN: Sign is an indication of existence of an objective evidence of a disease, i.e. such evidence as is perceptible to the examining physician, as opposed to the subjective sensation (symptoms) of the patient.

PATHOGNOMONIC SIGN (patho = disease, gnoma = signature, pathognomonic = signature of the disease): Specially distinctive or characteristic sign of a disease or pathological condition on which a diagnosis can be made.

ACCESSORY SIGN (Assident sign): Any nonpathognomonic sign of disease, which adds on to the surety of the diagnosis when present.

ANTECEDENT SIGN: Any precursory indication of an attack of disease. These signs are to be identified at the earliest.

1. **Aaron's sign:** A sensation of pain and/or distress in the epigastric or precordial region on pressure over McBurney's point in appendicitis.
2. **Abadie's sign:**
   Charles Abadie—Spasm of levator palpebrae superioris muscle—A sign of Graves' disease.
   Jean Abadie—Insensibility of Achilles tendon to pressure, seen in tabes dorsalis.
3. **Air cushion sign** (Syn: Klemm's sign): In the radiograph of chronic appendicitis, there is often an indication of tympanitis in the right lower quadrant.
4. **Alder's sign** of shifting tenderness: This sign is useful to diagnose acute appendicitis in pregnancy.
   Locate the most tender spot and mark it on the skin. Now request the patient to turn on the left side and wait for a full minute. If the tenderness is of uterine origin it will shift with the uterus while the position remain constant in case of appendicular origin.
5. **Angell's sign:** Helpful in diagnosing *"Torsion testis"* due to developmental anomaly—The presence of mesentery between the testis and the epididymis is invariably bilateral.

   The sign is usually obscured on the affected side and can be made out by examining the patient in the standing position wherein the opposite testis will be found to lie horizontally instead of in the normal vertical position.
6. **Anghelescus' sign:** Seen in Pott's disease of spine (TB spine) wherein the victim is unable to bend the spine while lying on the back so as to rest on the head and heel alone.
7. **Argyll Robertson pupil sign:** Typically described for neurosyphilis wherein light reflex is lost while accommodation reflex is retained. Such an eye responds poorly to mydriatics. This is due to destruction of fibres between pretectal nucleus and Edinger-Westphal nucleus. Other conditions where this sign may be seen—encephalitis, vascular and traumatic lesions, cerebral tumours, diabetes mellitus and chronic alcoholism.
8. **Auenbrugger's sign:** Bulging of the epigastrium due to massive pericardial effusion. This sign highlights the importance of examining thorax in patients with abdominal symptoms.
9. **Babinski's sign** (Not syn with Babinski's reflex):
   a. Loss or weakening of the Achilles tendon reflex in sciatica. This sign helps distinguish true sciatica from hysterical sciatica.
   b. In hemiplegia the contraction of platysma muscle in the healthy side is more pronounced than on the affected side. It can be elicited by asking the patient to open the mouth, whistling, blowing, etc.

Everything is funny as long as it is happening to somebody else.

   c. When a hemiplegic patient is lying with arms crossed upon the chest and makes an effort to sit up, the thigh on the paralyzed side is flexed upon the pelvis and the heel lifted from the floor while limb on the healthier side does not move.
   d. When the paralyzed forearm is placed in supination it turns over to pronation. It is seen in organic paralysis and is also called 'Pronation sign'.

10. **Baid sign:** Described for pseudocyst of pancreas and is well appreciated in thin individuals. When Ryle's tube is passed into the stomach, it may be palpated over the swelling because the stomach is displaced anteriorly by the pseudocyst pancreas.
11. **Ballances' sign:** Seen in about 25% of ruptured spleen. There is a dull note in both the flanks due to haemoperitoneum. The dullness on the right side can be made to shift, but that on the left side remains constant/fixed because the blood in the vicinity of the ruptured spleen gets coagulated soon.
12. **Ballet's sign:** This sign is helpful in Graves' disease and hysteria wherein there will be persistence of involuntary pupillary and reflex eye movements with loss of all voluntary eye movements (external ophthalmoplegia).
13. **Bamberger's sign:** A sign described in pericardial effusion—Presence of signs of consolidation at the angle of scapula which disappears when the patient leans forwards.
14. **Banana sign:** An ultrasonographic sign described for Arnold-Chiari deformity—a cause for congenital hydrocephalus. Sonography of the fetal skull reveals flattened and curved (banana like) shape of the cerebellar hemisphere.
15. **Bastede's sign:** A sign described in a case of appendicitis—Mentioned to be condemned—when colon is inflated with air through a rectal tube, pain and tenderness can be elicited in the right iliac fossa in a suspected case of appendicitis. Such manoeuvre carries risk of perforation and hence should not be entertained.
16. **Battle's sign:** It is relevant in middle cranial fossa fracture—bruising/ecchymosis over the mastoid process in the line of posterior auricular artery is a tell-tale sign of underlying middle cranial fossa fracture (temporal bone fracture).
17. **Bergman's sign:** A urologic radiographic sign:
    a. The ureter is dilated immediately below an obstructing neoplasm rather than collapsed as seen in cases of obstructing stones.
    b. The ureteral catheter passed in such cases tends to coil in this dilated portion of the ureter.
18. **Berry's sign:** Indicated by the absence of carotid artery pulsation in a patient presenting with goitrous swelling, is an ominous sign of thyroid malignancy (due to carotid sheath infiltration by the malignant tissue).
19. **Bezold's sign:** Described in a case of mastoiditis—An inflammatory swelling seen below the apex of the mastoid process.
20. **Biederman's sign:** A dark colour instead of normal pink colour of the anterior pillar of the throat is seen in some patients with syphilis.
21. **Bird's sign:** Described for hydatid disease of lung wherein a definite zone of dullness with absence of the respiratory sounds may be appreciated.
22. **Biernacki's sign:** Analgesia of ulnar nerve in general paresis and tabes dorsalis.
23. **Blatin's sign** (Syn: Hydatid thrill): A sign elicited in cases of hydatid cystic disease. It is due to displacement of daughter cysts in the fluid of the mother cyst.
24. **Boas sign:** An area of hyperaesthesia, posteriorly extending 2.5 cm lateral to the spinous process of vertebrae to the posterior axillary line and vertically from the level of the 11th dorsal to the 1st lumbar spine—A definitive sign of the presence of cholecystitis.
25. **Blumberg sign** (Syn: Rebound tenderness; Release sign): It is a sign of peritonitis due to presence of an inflamed organ underneath it.
26. **Bonnet's sign:** Pain on thigh adduction in sciatica.
27. **Boston's sign** (Syn: von Graefe's sign): It constitutes the lid lag elicited in cases of thyrotoxicosis.
28. **Bowler's hat sign** (Syn: Double ring sign): A radiological sign which describes the appearance of a gastric polyp seen on end-on position in a double contrast barium meal study. There will be a central lucency with two rims of barium around. May also be seen in cases of sessile intestinal polyp or a diverticulum.
29. **Boyce's sign:** A gurgling sound heard on pressure by the hand on the side of neck, in cases of oesophageal diverticulum.
30. **Bozzolo's sign:** A visible pulsation of the arteries in the nostrils. A sign believed to indicate the presence of aneurysm of the thoracic aorta.
31. **Branham sign** (Syn: Nicoladoni sign): This sign is elicited when arteriovenous fistula is suspected. A pressure on the artery proximal to the fistula will cause:
    a. Reduction in size of the swelling.
    b. Disappearance of bruit.
    c. Fall in pulse rate.
    d. Pulse pressure returns to normal.
32. **Brodie's sign:** A black spot on the glans penis—a sign of gangrene due to urinary extravasation into the corpus spongiosum.
33. **Bald fundus sign:** A radiological sign described for atrophic gastritis. The fundus of the stomach looks like a small dome with absence of mucosal pattern indicated by a very thin smooth appearing gastric wall.

34. **Bent inner tube sign:** A radiological sign described for sigmoid volvulus—an important diagnostic sign. A plain X-ray abdomen taken in the supine position reveals a massively distended ahaustral sigmoid colon arising from the pelvic loop.
35. **Bird of prey sign:** A radiological (Barium enema) sign which helps in confirming the diagnosis in doubtful cases of sigmoid volvulus. Barium enema reveals a smooth tapered narrowing at the point of torsion of the colon. Mucosal folds show a screw pattern around the point of twist.
36. **Border sign:** Describes the appearance of ventral hernia in barium study follow through done in suspected cases, in the postoperative period. Lateral and inferior border of the hernia are sharply outlined while the medial and upper border blends with the abdominal shadows.
37. **Cardarelli's sign:** Transverse pulsations in the laryngotracheal tube in suspected cases of aneurysms and dilatation of the aortic arch.
38. **Chvostek's sign** (Syn: Chvostek Weiss sign; Schultze's sign): A clinical sign typically described for hypocalcaemic tetany. The sign is elicited by tapping over the muscles and/or superficial nerves to induce the muscle spasm. It may be:
    a. **Facial sign:** A light tap over the facial nerve branches in front of the ear lobe causes muscular twitching over the whole of that side of the face.
    b. **Peroneal sign:** Tapping the peroneal nerve near the fibular neck will cause dorsiflexion and abduction of the foot.
39. **Cobra head sign** (Adder head appearance): A radiological sign seen in cases of ureterocele—congenital dilation of lower end of ureter. Urography done in suspected cases produces a characteristic appearance which resembles a cobra head and hence the name.
40. **Crescent sign:** A radiological sign described in relation to two different conditions affecting the lungs and the kidneys:
    a. In plain chest X-ray taken in a patient suspected of hydatid disease or aspergilloma (fungus ball) of the lung reveals a crescent of air in the shadow of partially ruptured hydatid cyst or in the cavity containing fungus ball.
    b. On IVP, the nephrogram reveals a crescent sign in cases of congenital hydronephrosis.
41. **Cullen's sign** (Syn: Cullen Hofstatter sign): A clinical sign which was typically and initially described for ruptured ectopic pregnancy wherein there is discolouration (ecchymosis) of the umbilicus and the surrounding skin (aptly referred to as umbilical black eye). It is due to haemoperitoneum and may be seen in conditions like ruptured ectopic pregnancy (a bluish tinge), acute haemorrhagic pancreatitis (a yellowish tinge).
42. **Carman's sign** (Syn : Carman-Kirklin meniscus sign): A radiographic sign helpful in the analysis of gastric ulcers in barium meal study. It is a reliable indicator of malignancy. Nonprojecting ulcers lying intraluminal in all projections surrounded by an elevated rim of tumour produces a curval interface resembling a meniscus and is called Carman's meniscus sign. Whether the meniscus is convex or concave towards the lumen is mainly dependent upon the site of ulcer in relation to incisura angularis. Meniscus is concave towards the lumen in cases of ulcer being proximal to the incisura while it is convex when the ulcer is distal to the incisura.
43. **Carnett's sign:** A clinical sign which helps differentiate the plane of abdominal swelling. The abdominal wall muscle is made tense by asking the patient to raise both the legs with knee extended. If the lump is intraperitoneal it disappears or becomes less prominent, while it becomes prominent or persists when the lump arises from the abdominal wall.
44. **Chilaiditi's sign** (Syn: Chilaiditi's syndrome; Hepatoptosis): A radiological sign seen in plain X-ray abdomen which is helpful in the diagnosis of this syndrome wherein there is interposition of colon between liver and diaphragm. It needs to be differentiated from the conditions causing gas under the diaphragm. The presence of haustrations in the gas shadow in a plain X-ray abdomen tilts the favour in the diagnosis of Chilaiditi's syndrome.
45. **Coiled spring sign:** A radiological sign (Barium enema) classically described for intussusception. The passage of barium beyond the apex of the intussusception into the intussuscipien gives a coiled spring appearance. The sign may also be seen in cases of:
    a. Post-traumatic haematoma of duodenum.
    b. Acute appendicitis.
    c. Mucocele of appendix.
    d. Endometriosis of appendix.
    e. Intestinal carcinomas.
46. **Coles' sign:** A radiological sign. Barium meal follow through study reveals deformity of the duodenal contour in the presence of duodenal ulcer.
47. **Colon cut off sign:** A radiographic sign of appendicular perforation or colonic spasm. Absence of gas and feces in the right lower quadrant, reflex dilatation of transverse colon and sharp cut-off of gas at the hepatic flexure. This sign is also seen in acute pancreatitis
48. **Coopernail's sign:** Ecchymosis on the perineum and scrotum/labia. A sign of fracture pelvis.
49. **Cope's sign** (Syn: Psoas sign): Clinical signs which are relevant in cases of acute appendicitis. This has two tests:
    a. **Cope's psoas test:** In acute appendicitis there is psoas muscle spasm secondary to the inflamed organ and hence the patient keeps the thigh in a flexed position. This pain can be aggravated by passively hyper-extending the hip joint which stretches the psoas muscle.

   b. **Cope's obturator test:** Principles are same as above except the muscle involved is obturator internus which is passively stretched by internally rotating the right leg which is flexed at the hip and knee.

50. **Cupola sign:** A radiological sign helpful in the diagnosis of pneumoperitoneum. A plain X-ray abdomen (erect position) reveals gas under the diaphragm when there is relatively large amount of air in the peritoneal cavity—Cupola sign.
51. **Crow foot sign** (Syn: Mercedes Benz sign; Seagull's sign): A radiological sign described in relation to cholelithiasis (Gallstones). Nearly 80-90% of the gallstones are radiolucent with only 10-20% being radio-opaque. However rarely a non-opaque gallstone can be diagnosed in plain radiography by the presence of gas containing crevices within the stone. These radiolucent crevices give the appearance of the crow foot; Hence the name crow foot sign, Mercedes Benz sign or the Seagull's sign.
52. **Courvoisier's sign** (Syn; Courvoiser's law): In a patient with obstructive jaundice, if the gallbladder is palpable it is not due to gallstones.
53. **Cowen's sign:** A clinical sign elicited in cases of Graves' disease. In response to a light shone on one eye there is jerky constriction of the contralateral pupil.
54. **Crowe's sign:** Refers to axillary freckling seen in neurofibromatosis.
55. **Cruveilhier's sign** of the saphena varix: A clinical sign elicited in cases of varicose viens (Saphena varix). A thrill is felt over the saphena varix when the patient is asked to perform a Valsalva manoeuvre/cough in an erect position and is due to a jet of blood entering and filling the pouch.
56. **Coleman's sign:** Helps in the clinical evaluation of fracture mandible. There is obvious swelling and bruising over the bony injury and a haematoma in the floor of the mouth if the body of the mandible is fractured.
57. **(Meniscus) Claw sign:** A radiological sign (Barium study) diagnostic of ileo-colic intussusception. The barium in the intussuscipien is seen as a claw around the negative shadow of the intussusceptum.
58. **Coffee bud sign:** It is a radiological sign seen in volvulus of sigmoid colon.
59. **Dalrymple's sign:** It is one of the manifestations of Graves' ophthalmopathy. It consists of retraction of the upper eyelid so that the palpebral opening is abnormally wide and upper sclera is visible.
60. **Delbet's sign:** A prognostic indicator in cases of aneurysm of the main artery of the limb. If the nutrition of the part distal to the aneurysm is maintained then the collateral circulation is said to be sufficient even though the distal pulse is not felt.
61. **de Musset's sign:** Rhythmical jerking movement of the head with each heart beat and is seen in cases of aortic insufficiency and aortic aneurysm.
62. **Dance sign** (Syn: Signe de Dance): A feeling of emptiness in the right iliac fossa—A sign of intussusception.
63. **Demarouay's sign:** Fixation or lowering of larynx during phonation and deglutition. A sign of syphilis of trachea.
64. **Dew's sign:** A clinical sign described in relation to the diaphragmatic hydatid abscess beneath the right cupola of the diaphragm. The area of resonance moves caudally with the patient in knee-elbow position.
65. **Dixon Mann's sign** (Syn: Mann's sign): In case of Graves' orbitopathy the two eyes appear not to be on the same level.
66. **Dorendorf's sign:** The sign identifies the fullness in the supraclavicular groove on one side in aneurysm of aortic arch.
67. **Double bubble sign:** A radiological sign described in cases of duodenal obstruction. A plain X-ray abdomen reveals two foci of gas, one in the stomach and the other in the duodenum—a sign of duodenal atresia. A similar feature is observed in the foetus in ultrasonography done in the antenatal period.
68. **Drummond's sign:** A whiff sound heard over the open mouth during respiration in cases of aortic aneurysm.
69. **DTP sign** (Syn: Distal tingling on percussion; Tinel's sign; Formication sign): A prognostic indicator which is helpful in the evaluation of nerve recovery following nerve injury. If percussion over the site of nerve injury causes tingling sensation in the distal end of the limb it suggests that the nerve injury was a partial one or it heralds the recovery from the nerve injury—a good prognostic sign.
70. **Dubois sign:** Shortness of the little finger in congenital syphilis.
71. **Duchenne's sign:** This clinical sign identifies the sinking of the epigastrium during inspiration (a paradox). It is classically seen in cases of paralysis of the diaphragm and in certain cases of hydropericardium (Pericardial effusion).
72. **Dupuytren's sign:** It is described in two conditions:
    a. A crackling sensation on pressure over a sarcomatous bone.
    b. In congenital hip dislocation, it refers to free up and down movement of the head of the femur.
73. **Dott's sign:** It is helpful in differentiating pain due to acute appendicitis and basal pneumonia with pleuritis. Compression of lower thorax from side to side elicits obvious distress when the lesion is above the diaphragm whereas in appendicitis it has no effect.
74. **Dock sign:** A radiological sign described in relation to coarctation of aorta. A chest X-ray PA view reveals rib notching on the inferior margins of 3rd-9th ribs while sparing the first two ribs and is indicative of collateral circulation developed in coarctation of aorta.
75. **De Weese sign:** A clinical sign helpful in cases presenting with history suggestive of intermittent claudications with palpable peripheral pulses. The patient is asked to exercise by walking or running sufficiently to bring on the pain. If prompt re-examination reveals absence of peripheral pulses it suggests that the pain was truly due to intermittent claudication.
76. **Echo sign:** A percussion sound resembling an echo which is heard over a hydatid cyst.
77. **Elliot's sign:** Refers to presence of indurated edge of a syphilitic skin lesion.

78. **Enroth's sign:** Identifies the abnormal fullness of the eyelids—a manifestation of Graves' orbitopathy.
79. **Escherich's sign** (Syn: Escherich's reflex): Described in relation to tetany, where percussion of the inner surface of the lips or tongue produces contraction of lips, tongue and masseter muscles.
80. **Ewart's sign:** It is said to be positive in cases of pericardial effusion if there is bronchial breathing and dullness on percussion at the lower angle of the left scapula and is due to pressure and collapse of the lingular lobe due to enlarged pericardial sac.
81. **E- sign** (Syn: Reverse 3 sign): A radiological sign seen on barium swallow radiograph in patient suspected of coarctation of aorta and is due to indentation produced on the barium filled oesophagus by the aorta.
82. **Figure of three sign:** A radiological sign seen in the plain chest X-ray PA view in patient with coarctation of aorta. A pair of bulges is seen in the wall of the aortic arch one above and one below the aortic knuckle and is due to pre- and post-stenotic dilatation of the aorta.
83. **Frostberg's sign:** A radiological sign seen in barium study in carcinoma of the head of pancreas involving the duodenum. In carcinoma of the head of the pancreas, there is widening of the 'C' loop of the duodenum and with the involvement of the ampulla of Vater, the expanded loop assumes a reversed 3 configuration—Frostberg's sign. Once the carcinoma of the head affects the duodenum the surgical cure is remote.
84. **Fox sign:** Discolouration near the inguinal ligament—seen in few cases of haemorrhagic pancreatitis.
85. **Fuchsig's sign** (Syn: Crossed leg test): This sign is more relevant when popliteal artery is not clinically palpable along with absent distal pulses. The patient is made to sit with his legs crossed. Normally, oscillatory movements of the foot occur synchronously with the pulse if the popliteal artery is patent while a negative result is more suggestive of popliteal artery block in a patient with absent distal pulses.
86. **Federici's sign:** An interesting sign described in relation to pneumoperitoneum (due to intestinal perforation) wherein on auscultation of the abdomen, the cardiac sounds can be heard.
87. **Flush tank sign:** A clinical sign described in relation to hydronephrosis. The patients classically present with features suggestive of Dietle's crisis, i.e. passage of a large amount of urine with consequent disappearance of a lumbar swelling—Flush tank sign.
88. **Flare sign:** The precursor of varicose ulcer is a splay of fine venules that courses from the medial (sometime the lateral) malleolus, and spreads out to be lost beneath the thick skin of the heel and is known as the "Flare sign".
89. **Gifford's sign:** A clinical sign which helps in differentiating unilateral exophthalmos and proptosis due to intraorbital tumours. If the upper eyelids cannot be easily everted on the affected side, the exophthalmos is more likely to be due to thyrotoxicosis, whereas if the lid is easily everted it is more likely due to an intraorbital mass.
90. **Grey Turner sign:** Skin discolouration (bruising) in the left flank (left costovertebal angle) in cases of acute haemorrhagic pancreatitis.
91. **Gilbert's sign:** Opsiuria indicative of hepatic cirrhosis (opsiuria means excretion of urine more rapidly during fasting than after a meal).
92. **Glasgow's sign:** A clinical sign—presence of systolic murmur over the brachial artery is found in latent aortic aneurysm.
93. **Griffith sign:** It is one of the clinical signs of Graves' orbitopathy wherein there is lower lid lag on upward gaze.
94. **Guyon's sign:** Ballottement and palpation of a floating kidney.
95. **Gaur sign:** A clinical sign seen in femoral hernia. Distension of superficial epigastric and/or circumflex iliac veins on the affected side due to pressure on these vessels by the hernial sac.
96. **Guerin's sign:** Haematoma at greater palatine foramen seen in maxillary fracture.
97. **Homan's sign:** Pain on sudden dorsiflexion of the foot—a sign of deep vein thrombosis of calf veins.
98. **Hamilton Bailey sign:** A clinical sign described for intussusception. A sausage shaped mass with concavity towards the umbilicus which is felt to harden as a wave of peristalsis commences.
99. **Hall's sign:** This sign identifies a tracheal diastolic shock felt in the aneurysm of aorta.
100. **Hatchcock's sign:** Refers to the tenderness towards the angle of the jaw in mumps.
101. **Haudek's sign** (Haudek's niche): A radiological sign identified in barium meal X-ray done in patient suffering from gastric ulcer. It describes a projecting shadow (niche) in radiographs due to settlement of barium in pathological niches of stomach wall.
102. **Hitzelberger's sign:** A clinical sign described in case of acoustic neuroma wherein there is anaesthesia of medial, posterior or superior areas of the external auditory canal because of the tumour tissue compressing on the facial nerve.
103. **Hook sign:** This clinical sign identifies flexion of fingers in a case of acute suppurative tenosynovitis.
104. **Horn's sign:** This sign identifies the pain produced by traction on the right spermatic cord in acute appendicitis.
105. **Howship Romberg sign:** A sign described in the patients suffering from obturator hernia. The patient complains of pain passing down the inner side of the knee due to pressure on the obturator nerve by the hernial sac.
106. **Harvey's sign:** Two index fingers are placed by the side on a vein. The fingers are now pressed firmly and the finger near the heart is moved proximally keeping the steady pressure on the vein so as to empty the short length of vein between

the two fingers. The distal finger is now released. This will allow venous refilling to be observed which will be poor in ischemic limb and increased in arteriovenous fistula.

107. **Hamman's sign** (Syn: Hamman's mediastinal crunch or murmur): A clinical sign—precordial crunching, clicking or knocking sound synchronous with each heart-beat heard on auscultation in conditions such as:
    - Acute mediastinitis
    - Pneumomediastinitis
    - Pneumothorax.
108. **Inflammatory signs** (Syn: Cardinal signs in inflammation):
    - Rubor—Redness
    - Calor—Temperature
    - Dolor—Pain
    - Tumour—Oedema/swelling
    - Functio lesa—Loss of function
109. **Jendrassik's sign:** Paralysis of extraocular muscles—A manifestation of Graves' orbitopathy.
110. **Jugular sign** (Syn: Quekenstedt's sign; Tobey Ayer test): On spinal canal block (e.g. spinal tumour) when pressure is applied over jugular veins, in the manometer connected to LP needle, either pressure will not raise or if raises falls slowly.
111. **Joffroy's sign:** Absence of wrinkling of the forehead when the head is bent down and the patient is asked to look upwards—A sign of Graves' ophthalmopathy.
112. **Kernig's sign:** A sign which is positive in meningitis. With the hip flexed, the knee is extended, normally it can be done upto 175°. In meningitis it is restricted due to spasm of the hamstrings.
113. **Kaposi Stemmer sign:** Failure to pick up or to pinch a fold of skin at the base of the second toe. It is characteristic of lymph oedema.
114. **Kocher sign** (Syn: Mean's sign): A sign of Graves' orbitopathy. The examiner places one hand on a level with the patient's eye and then lifts it higher. The upper eyelid springs up more quickly than the eyeball.
115. **Kanavel sign:** This sign is useful in the diagnosis of ulnar bursitis. In case of ulnar bursitis the site of maximum tenderness is over that part of the ulnar bursa lying between the transverse palmar creases.
116. **Kantor's sign** (String sign of Kantor): A radiological sign described in barium enema follow through in patients suffering from Crohn's disease. A string like configuration of contrast material through a filling defect is seen in the radiography.
117. **Kehr sign:** This sign identifies the pain elicited in the left shoulder in patients with suspected splenic rupture. The pain (referred pain) experienced by the patient is due to blood in the peritoneal cavity irritating the diaphragm.
118. **Klemm's sign:** It is the radiological sign found in chronic appendicitis where there is often an indication of tympanitis in the right lower quadrant.
119. **Knie's sign:** Unequal dilatation of the pupils—a sign of Graves' orbitopathy.
120. **Krisowski's sign:** Cicatricial lines radiating from the mouth in congenital syphilis.
121. **Kenawy's sign:** A clinical sign usually associated with bilharzial fibrosis of liver (Egyptian splenomegaly) but may be present in any type of portal hypertension. Auscultation over the xiphoid process reveals a venous hum (splenic vein engorgement) which becomes prominent on inspiration.
122. **Klein's sign** (of shifting tenderness): A clinical sign useful in acute nonspecific mesenteric lymphadenitis. When patient is shifted to left lateral position, point of maximum tenderness also gets shifted to the left side in contrast to acute appendicitis. It may also be positive in Meckel's diverticulitis.
123. **Lemon sign:** A radiological sign of Arnold-Chiari deformity. There is scalloping of the frontal bones giving the skull a lemon shaped configuration in sonography of the fetal skull during second trimester of pregnancy.
124. **Lennhoff's sign:** A furrow appearing on deep inspiration below the lower rib and above an echinococcal cyst of liver.
125. **Sign of Leser Trelat:** Sudden appearance and rapid increase in size and number of seborrhoeic keratoses, a sign of internal malignancy of the GIT.
126. **Lloyd's sign:** A clinical sign elicited in a patient with renal calculus. Pain elicited in the loin on deep percussion over the kidney even when there is no pain on applying deep pressure.
127. **London's sign:** It is useful in ruptured intestine suspected in an accident case. The presence of pattern of bruising of the skin (i.e. an imprint of the clothing is noted on the skin) indicates that a crushing force has been applied sufficient to rupture the bowel against the vertebral column. This sign is a strong indication to carry a laparotomy.
128. **Mallet Guy's sign:** A clinical sign identified in chronic relapsing pancreatitis. It is elicited by placing the patient in the right lateral position with the patient's hip and knee flexed and deeply palpating the abdomen in the epigastric and the left subcostal region. This will evoke tenderness in chronic relapsing pancreatitis.
129. **McBurney's sign:** Finger tip pressure is made over the McBurney point elicits severe tenderness in patients with appendicitis.
130. **Murphy's sign** (Moynihan's method): This clinical sign is classically described in patients suffering from cholecystitis. It is elicited by asking the patient to breath deeply while exerting moderate pressure with the left hand such that thumb lies over the fundus of the gallbladder. The patient catches his breath as the inflamed gallbladder which is pushed down by the diaphragm gets imposed against the thumb.

131. **McEwen's sign:** A clinical sign elicited in children with hydrocephalus. On percussion of skull behind the junction of frontal, parietal and temporal bones and auscultation over opposite mastoid bone there is a more resonant note than normal, seen in internal hydrocephalus and cerebral abscess in children (syn: cracked pot sign, cranial cracked pot sound).
132. **Meniscus sign:** Refer No. 57.
133. **Moebius sign:** Inability to keep the eyeballs converged due to insufficiency of medial rectus muscle— A clinical sign of Graves' ophthalmopathy.
134. **Milian's ear sign:** A clinical sign which is useful to differentiate facial erysipelas from cellulitis. Erysipelas being a cuticular lymphangitis spreads from the face to the pinna while the cellulitis which is spreading inflammation of the subcutaneous tissue stops short of pinna because of close adherence of the skin to cartilage.
135. **Mahler's sign:** A steady increase of pulse rate without corresponding increase of temperature seen in thrombosis.
136. **Mann's sign** (Syn: Dixon Mann's sign): In Graves' disease the two eye appears not to be on the same level.
137. **Marie's sign:** Tremors of body extremities in Graves' disease and other types of hyperthyroidism.
138. **Mean's sign** (Syn: Kocher's sign): Refer No. 114.
139. **Medusa Lock sign:** A radiological sign described in patients with intestinal obstruction due to roundworm infestation. A plain X-ray erect film of the abdomen reveals trapped intestinal gas within the worm mass giving a characteristic appearance of coiled locks of hair—Medusa Lock sign.
140. **Meitzer's sign:** Loss of normal second heart sound on auscultation of the heart after swallowing. Seen in occlusion or contraction of lower part of oesophagus.
141. **Mercedes Benz sign** (Syn: Crow foot sign): Refer No. 51.
142. **Mexican hat sign:** A radiological sign described in barium enema (filling defect) done in patients with pedunculated polyp of the inferior wall of the colon.
143. **Maulage sign:** Waxy cast appearance of bowel segments—A radiographic sign of coeliac disease.
144. **Mose's sign:** A clinical sign suggestive of deep vein thrombosis. It is elicited by squeezing the relaxed calf muscles from side to side which is painful in case of deep vein thrombosis.
145. **Niche sign** (Syn: Haudek's sign): Refer No. 101.
146. **Nicoladoni's sign** (Syn: Branham's sign): Refer No 31.
147. **Oliver's sign** (Syn: Porter's sign; Tracheal tug): Tracheal tug is seen in:
    - Aneurysm of the aorta.
    - Neoplasm which fixes bronchus to aorta.
148. **Omega sign:** A radiological sign observed in patient suspected of sigmoid volvulus on the X-ray. A plain X-ray abdomen taken in the supine position shows a massively distended sigmoid colon with haustrations arising from the pelvis resembling Greek letter 'omega' (ω).
149. **Payr's sign:** Pressure over the sole of the foot elicits severe pain in cases of thrombophlebitis.
150. **Peroneal sign:** Tapping the peroneal nerve near the fibular neck causes dorsiflexion and abduction of foot in hypocalcaemic tetany.
151. **Perez's sign:** A friction sound heard over the sternum when the patient raises and drops his arms. A sign of mediastinal tumour or aneurysm of arch of aorta.
152. **Parrot's sign** (Syn: Parrot's nodes, hot cross bun skull, Natiform skull): Bony nodes on the outer table of the skull of infants with congenital syphilis, so that it has a hot cross bun or buttock shape.
153. **Pfuhl sign:** A clinical sign which helps in differentiating between subphrenic abscess and pyopneumothorax. Inspiration increases the force of flow in paracentesis in subphrenic abscess, but lessens in case of pyopneumothorax. This distinction is however lost in cases where diaphragm is paralyzed.
154. **Pitre's sign:** Hypoaesthesia of scrotum and testes in tabes dorsalis.
155. **Plummer's sign** (Syn: Quadriceps sign): Inability to step up onto the chair or to walk up steps seen in Graves' disease and other forms of hyperthyroidism.
156. **Pool Schlesinger's sign:** In tetany if patients leg is held at the knee joint and flexed strongly at the hip joint, within a short time there will be an extensor spasm at the knee joint with extreme supination of foot.
157. **Porter's sign:** Tracheal tugging in aneurysm of aortic arch and neoplasms which fix the left bronchus to aorta.
158. **Potain's sign:** Extension of percussion dullness over the arch of the aorta from the manubrium to the third costal cartilage on the right side—seen in cases of dilatation of the aorta.
159. **Prehn's sign:** Elevation and support of the scrotum will relieve the pain in epididymo-orchitis but not in torsion testis.
160. **Psoas sign** (Syn: Cope's test): Refer No. 49.
161. **Puddle sign:** Describes a clinical method to detect small amount of ascites (150-500 ml). Patient lies prone for 5 minutes and then goes for knee-elbow position. In this position dullness is elicited in the umbilicus in case of minimal ascites.
162. **Pemberton's sign:** This sign refers to symptoms of faintness with evidence of facial congestion and external jugular vein distension when the arms are raised above the head touching the ears. This manoeuvre reduces the thoracic inlet thereby hampering venous drainage of the face in the presence of retrosternal thyroid.

Absolute diagnosis are unsafe and are made at the expense of the conscience.—**William Osler**

163. **Pointing sign:** Ask the patient to point to the site of maximum pain. If this proves to be the site of localized tenderness it is also certainly the site of diseased organ, e.g. appendicitis.
164. **Patel's sign:** It is important in paralytic ileus. Apply the stethoscope firmly to the skin just below and right of the umbilicus for full 3 minutes. In paralytic ileus there will be omnious silence, broken by patient's own heart sounds believed to be transmitted via the over distended coils of the intestine with added succussion splashes if the patient moves and very occasionally by faint tinkles.
165. **Pad sign:** A radiological sign identified in patients with carcinoma of the head of pancreas wherein the 'C' loop of barium filled duodenum may be widened.
166. **Panda sign** (Raccoon sign, Spectacle haematoma, Black eye): Haemorrhages in soft tissues around the eye and in the eyelids is known as Panda sign as it resembles panda eyes. It is caused by:
    1. Direct trauma, such as punch to the eye.
    2. Blunt impact to the forehead, the blood gravitating downwards over the supraorbital ridge.
    3. Fracture of the floor of the anterior fossa of the skull.
167. **Queckenstedt's sign** (Jugular sign; Tobey Ayer test): Refer No. 110.
168. **Quenu Muret sign:** A prognostic indicator. In aneurysm, when the main artery of the limb is compressed and a puncture is made at the periphery, if the blood flows, then the collateral circulation is probably established.
169. **Rat-tail sign:** A radiological sign described in relation to oesophageal carcinoma.
170. **Ridge sign:** A clinical sign of dehydration. In considerable dehydration if the skin is picked up in between the thumb and fingers and then released, instead of it springing back with normal elasticity, a ridge is formed that subsides slowly.
171. **Rovsing sign:** In acute appendicitis, when left iliac fossa is pressed; pain is felt in right iliac fossa.
172. **Release sign** (Syn: Rebound tenderness; Blumberg's sign): Refer No. 25.
173. **Reversed 3-sign:** Refer No. 81.
174. **Reisman sign** (Syn: Snellen's sign): The bruit heard over the closed eye in Graves' disease.
175. **Rocher's sign:** A clinical sign which helps in the differentiation of epididymitis from torsion of the testes. In torsion testes, the epididymis cannot be distinguished from the body the testes, whereas in epididymitis the body of the testes can be felt in the enlarged crescent of epididymis.
176. **Rommelaere's sign:** An abnormally small proportion of normal phosphates and of sodium chloride in urine in cancerous cachexia.
177. **Rotch sign:** A clinical sign identified in case of pericardial effusion. Dullness is felt on percussion on the right 5th intercostal space.
178. **Rovighi's sign:** A fremitus felt on percussion and palpation of the superficial hepatic hydatid cyst.
179. **Racoon sign** (Syn: Panda sign, Spectacle Haematoma, Black Eye): Refer No. 166.
180. **Sign of splashing** (Syn: Succussion splash): Seen in gastric outlet obstruction.
181. **Saegesser's sign:** A clinical sign identified in patients with splenic rupture. An excruciating tenderness is elicited in the Saegesser's point or splenic point (seen in the lower part of the posterior triangle of the neck between the left sternomastoid and scalenus medius muscle above the clavicle).
182. **Sternomastoid sign** (Syn: Trail's sign): The sternal head of the sternomastoid muscle will become more prominent on the side to which trachea is deviated.
183. **Saenger's sign:** Refers to the light reflex of the pupil that has ceased, returns after a short stay in the dark. Observed in cerebral syphilis but not in tabes dorsalis.
184. **Sansom's sign:** Described in two separate context:
    - Marked increase in the area of dullness in the second and third intercostal space – Due to pericardial effusion.
    - A rhythmical murmur heard with a stethoscope applied to the lips in aneurysms of the thoracic aorta.
185. **Schlesinger's sign** (Syn: Pool's phenomenon): Refer No. 156.
186. **Schultze sign** (Syn: Schultze Chvostek sign): Refer No. 38.
187. **Silex's sign:** Furrows radiating from the mouth in congenital syphilis.
188. **Sisto's sign:** Constant crying in infancy—a sign of congenital syphilis.
189. **Setting sun sign:** A clinical sign described in patients suffering from raised intracranial tension. There is downwards deviation of the eyes so that each iris appears to set beneath the lower eyelid with white sclera exposed between it and the upper lid. This sign is observed in cases:
    - Hydrocephalus
    - Intracranial haemorrhage
    - Brain tumours
190. **Snellen's sign** (Syn: Reisman sign): Refer No. 174.
191. **Sumner's sign:** This sign is said to be positive when there is increase in the tone of the abdominal muscle on gentle palpation of the right iliac fossa. It may indicate:
    - Appendicitis
    - Right ureteric or renal stones
    - Torsion ovarian cyst.

192. **Simon's sign:** Identifies absence of usual co-relation between the movements of diaphragm and thorax, seen in early cases of meningitis.
193. **Stierlin's sign:** A radiological sign observed in barium enema study of the colon where there is absence of normal shadow due to an indurating or ulcerative process such as tuberculosis of caecum or colon.
194. **String sign** (Syn: Kantor's sign): Refer No. 116.
195. **Suker's sign:** Deficient complementary fixation in lateral eye rotation—a manifestation of Graves' orbitopathy.
196. **String of beads sign:** A series of round shadows resembling a string of beads or pearls, seen on a radiograph of small intestine, indicating of trapped gas surrounded by the fluid of obstructed and distended bowel.
197. **Stellwag's sign:** Identifies the widening of palpebral fissures (staring look) due to retraction of upper eyelids, an early sign of Graves disease.
198. **Stemmer sign** (Kaposi Stemmer sign): Refer No. 113.
199. **Slip sign:** A clinical sign which helps in differentiating a solid swelling, e.g. lipoma, from a cystic swelling. Here, when the edge of the swelling is palpated, the margin of the solid swelling does not yield but slips away from it unlike a cystic swelling which yields to the pressure of the palpating finger and does not slip away.
200. **Solius sign:** A radiological finding observed in a chest X-ray lateral view in patient's with enlarged thymus. An enlarged thymus being a firm swelling does not get flattened against sternum by the pressure of heart and great vessels.
201. **Shrinkage sign:** A radiological sign of thymus enlargement. In a chest X-ray there is a paradoxical alteration of the shape of the chest with respiration (Decrease in transverse diameter with deep inspiratory film than the expiratory film).
202. **Seagull sign:** Refer No. 51.
203. **Suzmann's sign:** A clinical sign described in patient with coarctation of aorta. The collaterals which develop display visible and palpable pulsations together with thrills and murmurs which are most obvious in the inter-scapular and infra-scapular regions of the back. The Dock sign is the radiological counterpart of this sign.
204. **Trail's sign** (Syn: Sternomastoid sign): Refer No. 182.
205. **Tap sign** (Syn: The percussion sign; Chevrier's sign): If the valves are incompetent an impulse will be felt by the fingers overlying the long saphenous vein when the varicosities are percussed below.
206. **Thornton's sign:** Refers to the severe pain complained by the patient in the region of the flanks in nephrolithiasis.
207. **Tinel's sign** (Syn: DTP sign; Formication sign): Refer No. 69.
208. **Tresilian sign:** This sign identifies reddish appearance (congestion) in opening of Stensen's duct in cases of mumps.
209. **Trimadeau's sign:** A radiological sign identified in barium swallow X-ray done in patient's with dysphagia. If the dilatation above an oesophageal stricture is conical, the stricture is fibrous; while it is cup shaped (shouldering) it is likely to be malignant.
210. **Tanyol's sign:** In ascites umbilicus shifts downwards and in mass arising from pelvis it shifts upwards.
211. **Troisier's sign:** Identifies enlargement of left supraclavicular lymph node (Virchow's node). Seen in:
    - Ca stomach
    - Ca testes
    - Ca bronchus
    - Malignancy of any other abdominal organ.
212. **Trousseau's sign:** This sign is described under two different context:
    - The blood pressure cuff is applied to the arm and inflated to pressure above systolic pressure for 3-5 minutes. This will elicit typical carpopedal spasm (obstetrician's hand) in cases of hypoparathyroidism and other conditions associated with hypocalcaemia.
    - Migrating superficial thrombophlebitis—a sign of visceral carcinomas especially of pancreas or the stomach.
213. **Tracheal fluctuation sign:** A unique sign elicited in patients suffering from achalasia cardia.
214. **von Graefe's sign** (Syn: Graefe's sign): Persistent lagging of upper lid behind the corneoscleral limbus when patient is asked to follow the finger moved up and down several times. Seen in Graves' disease.
215. **Vermooten's sign:** A clinical sign helpful in the intrapelvic rupture of urethra. On per rectal examination, the prostate cannot be felt but in its position a doughy swelling (blood and urine) is felt. If prostate is felt, it is displaced upwards.
216. **Vas sign:** This sign is helpful in differentiating testicular neoplasm and an inflammatory lesion of the testes. Inflammatory lesions causes vas deferens to become considerably thickened which remains normal in cases of neoplasm.
217. **Vein sign:** A bluish cord along the mid-axillary line formed by the swollen junction of the thoracic and superficial epigastric vein. Seen in:
    - Tuberculosis involving the bronchial glands
    - Superior vena cava obstruction.
218. **Wegner's sign:** A postmortem finding—A broadened discoloured appearance of the epiphyseal line in infants who have died from congenital syphilis.
219. **Wilder's sign:** An early clinical sign of Graves' disease consisting of slight twitch of the eyeball when it changes its movement from adduction to abduction or vice versa.

Observe, record, tabulate, communicate. Use your five senses.—**William Osler**

220. **Wimberger's sign:** Symmetrical erosions of the proximal tibia seen radiographically in infants with congenital syphilis.
221. **Water lily sign** (Syn: Lily pad sign): A radiological sign identified in cases of hydatid cyst of the lung. When the hydatid cyst ruptures, the daughter cyst floating within the cavity appear like a water lily hence the name water lily sign.
222. **Smith's sign:** Murmur heard in cases of enlarged bronchial lymph nodes on auscultation over the manubrium with the patients head thrown backwards.
223. **'H' Bomb sign:** A radiological sign seen in cases of atrophic gastritis. The gastric folds within fundus and the body of stomach are very thin and the thin walled fundus becomes distended with air (in erect posture) or with barium contrast.

B. Triads in Surgery

1. **Saint's triad:**
   Diverticulosis of colon
   Gallstones
   Hiatus hernia.
2. **Whipples' triad:** Seen in insulinoma.
   Features of hypoglycaemia
   Blood sugar less than 45 mg%
   Symptoms are relieved by glucose.
3. **Charcot's triad:** Seen in ascending cholangitis.
   Intermittent fever
   Intermittent pain
   Intermittent jaundice.
4. **Virchow's triad:**
   Change in the vessel wall
   Diminished rate of blood flow
   Increased blood coagulability.
5. **Murphy's triad:** Seen in acute appendicitis.
   Pain in right iliac fossa
   Vomiting
   Temperature.
6. **Hutchinson's triad:** Seen in late congenital syphilis
   Interstitial keratitis
   8th nerve deafness
   Hutchinson's teeth.
7. **Trotter's triad:** Seen in nasopharyngeal carcinoma.
   Conductive deafness
   Elevation and immobility of same side soft palate
   Pain in the side of the head.
8. **Tillaux's triad:** Seen in mesenteric cyst.
   Soft fluctuant swelling in the umbilical region
   Freely mobile in the direction perpendicular to mesentery
   Zone of resonance all around.
9. **Triad of portal hypertension:**
   Varices
   Splenomegaly
   Ascites.
10. **Cushing's triad:** In intracranial hypertension.
    Increased blood pressure
    Decreased pulse rate
    Decreased respiratory rate.
11. **Triad of renal cell carcinoma:**
    Anaemia
    Haematuria
    Mass in the loin.
12. **Borchardt's triad:** Seen in gastric volvulus
    Acute epigastric pain
    Violent vomiting
    Inability to pass nasogastric tube.
13. **Beck's triad:** Seen in cardiac tamponade.
    Muffled heart sounds
    Distended neck veins
    Hypotension.
14. **Pancoast's triad:** Seen in Pancoast tumour.
    Excruciating pain in the arm
    Horner's syndrome
    Erosion of ribs.
15. **Prune-Belly triad:** Seen in Prune-Belly syndrome.
    Cryptorchidism
    Abdominal wall defects
    Genitourinary defects.
16. **Mackler's triad:** Seen in Boerhaave's syndrome.
    Vomiting
    Chest pain
    Subcutaneous emphysema.
17. **Triad of Sandblom:** Seen in haemobilia
    Jaundice
    Pain
    Melaena.
18. **Galezia triad:**
    Dupuytren's contracture
    Retroperitoneal fibrosis
    Peyronie's disease of penis.
19. **Dieulafoy's triad:** Seen in appendicitis
    Hypersensitiveness of skin
    Reflex muscular contraction
    Mac Burney's tenderness.
20. **Triad of congenital diaphragmatic hernia**
    Respiratory distress
    Apparent dextrocardia
    Scaphoid abdomen.
21. **Carney's triad**
    Functioning adrenal para ganglioma—nonfamilial
    Gastric leiomyosarcoma—GIST
    Pulmonary chondroma.
22. **Triad of Ohashi in IPMN in ERCP**
    A bulging ampulla of Vater
    Mucin secretion
    Dilated main pancreatic duct.
23. **Haimovici triad of revascularisation of an acutely ischaemic limb**
    Muscle infarction
    Myoglobinuria
    Acute renal failure.
24. **Triad of small bowel obstruction in plain X-ray**
    Dilated small bowel loops > 3 cm
    Multiple air fluid levels in erect X-ray
    Paucity of air in the colon.
25. **Gilroy Bevan triad of adhesive pain is**
    Pain may get aggravated or relieved on change of posture
    Pain in the region of old abdominal scar
    Tenderness is elicited by pressure over the scar.

C. Misnomers in Surgery

1. *White bile*: It is neither white nor bile. It is opalescent. It contains mucous. It signifies severe obstructive jaundice due to which secretion of bile from liver is stopped. Mucous is derived from biliary tree lining.
2. *Mycotic aneurysm*: It is not due to fungal infection. It is due to bacterial infection.
3. *Lateral aberrant thyroid*: It is not an aberrant thyroid. It is secondaries in neck lymph node from occult primary in the thyroid, i.e. papillary carcinoma thyroid.
4. *Adenolymphoma of parotid gland*: It is not lymphoma. It is a benign tumour of the parotid. It never turns into malignancy.
5. *Pretibial myxoedema*: It is not seen in myxoedema. It is seen in thyrotoxicosis.
6. *Sternomastoid tumour:* It is not a tumour. It is due to birth trauma which causes organised haematoma in sternomastoid muscle.
7. *Dissecting aneurysm:* It is aortic dissection, not dissecting aneurysm.
8. *Malignant hydatid:* It is not malignant. It is due to *Echinococcus alveolaris*. It behaves like a malignant condition.
9. *Malignant exophthalmos:* It is due to primary thyrotoxicosis. It is not a malignant condition.
10. *Spina ventosa*: It is tuberculous dactylitis. It is not related to spine.
11. *Mycosis fungoides:* It is cutaneous "T" cell lymphoma. It is not due to fungal infection.
12. *Hypernephroma*: It is renal cell carcinoma. It is not above the kidney.
13. *Hepatoma:* It is hepatocellular carcinoma. It is not benign tumour.
14. *Melanoma:* It is melanocarcinoma. It is not benign.
15. *Surgical ganglion:* It is arising from the synovial sheath. It is not from nerve ganglion.
16. *Brain fungus*: It is not a fungal infection. It is seen in cranial injury wherein due to injury brain protrudes out of the wound as fungus.
17. *Cock's peculiar tumour*: It is not a tumour. It is ulcerated sebaceous cyst. It mimics SCC of skin.
18. *Pott's puffy tumour:* It is not a tumour. It is osteomyelitis of the frontal bone with cellulitis of the frontal region of the scalp. It may spread intracranially through emissary veins which endangers the life of the patient.
19. *Tumour alba*: It is tuberculosis of synovial sheath of knee joint.
20. *Umbilical adenoma*: It is not a tumour. It is prolapse of mucosa of cutaneous end of vitello-intestinal duct
21. *Madelung's deformity*: It is not associated with any of the conditions in the lung. It is a congenital subluxation or dislocation of the lower end of the ulna, from malformation of the bones.
22. *Ray fungus:* Actinomycoses, a bacteria infection.
23. *Malignant pustule:* Anthrax infection.
24. *Malignant oedema:* Gas gangrene.
25. *Juvenile melanoma:* It is not a melanoma. It is benign SPITZ naevus.
26. *Hydrocele of the neck*: It is cystic hygroma
27. *Kuttner tumour*: It is chronic sclerosing sialadenitis of submandibular salivary gland.
28. *Aneurysmal bone cyst:* It is not related to artery. It is expanding lesion in the bone containing blood mixed fluid of unknown aetiology.
29. *Hairy leukoplakia*: It is hairy leukoplakia with white confluent patches of fluffy / hairy hyperkeratotic thickening almost always situated in the lateral border of the tongue. It is not hair in tongue; not related hair diseases; not structurally hair component.

D. Triangles in Surgery

1. *Hesselbach's triangle:* The medial border of the triangle is formed by the rectus sheath, superolateral border by the inferior epigastric artery, inferior border by the inguinal ligament.

   Importance: Helps in differentiating direct from indirect hernias.
2. *Bryant's triangle:* Patient lies in supine position, three lines are drawn, one from the anterior superior iliac spine vertically down to the bed; another from the tip of the greater tronchanter to join the first line at the right angles; third line from anterior superior iliac spine to the tip of the greater tronchanter—this forms a triangle.

   Importance: Diminution in the length of the second line indicates upward displacement of the greater trochanter. Third (oblique) helps determine the anteroposterior displacement of the greater tronchanter.
3. *Triangle of Doom:* Bounded by the ductus deferens medially, spermatic vessels laterally in the male and the apex of the triangle is at the level of internal inguinal ring.

   Importance: External iliac artery and vein, femoral nerve are located in the triangle and so stapling is avoided in this triangle while doing laparoscopic preperitoneal repair for hernia.
4. *Calot's triangle:* Formed medially by the common hepatic artery, laterally by the cystic duct and the apex is formed by the junction of the cystic and the hepatic ducts. Base is formed by the cystic artery.

   Importance: Cystic lymph node of 'Lund' is present in the fork created by the junction of the cystic and common hepatic artery.
5. *Anal triangle and urogenital triangle:* A transverse line joining the anterior parts of the ischial tuberosities and passing immediately anterior to the anus divides the perineum into two triangles, anal and urogenital triangle. Anal triangle contains the termination of the anal canal in the median plane and an ischiorectal fossa on each side.

   Importance: Both the perineal and ischiorectal spaces in the anal triangle are common sites of abscess.
6. *Femoral triangle:* It is triangular depression on the front of the upper one-third of the thigh below the inguinal ligament, bounded medially by the medial border of adductor longus, laterally by the medial border of sartorius, apex is formed by the meeting point of the medial and lateral boundaries.

   Importance: Femoral vein in this site is most suitable for intravenous injection in the infants; femoral artery pulsations are an important aid for clinicians (in diagnosing peripheral vascular disease).
7. Lumbar triangle:
   a. *Inferior lumbar triangle of Petit*—bounded by the crest of ilium below, external oblique laterally and medially by the latissimus dorsi.

      Importance: Most primary lumbar hernia occur through this triangle.

   b. *Superior lumbar triangle of Grynfelt*—bounded by the 12th rib above, medially by the sacrospinalis and laterally by the posterior border of the inferior oblique.

   Importance: Grynfelt hernia appears through superior lumbar triangle.
8. *Triangle of auscultation*: This is a small triangular interval bounded medially by the lateral border of the trapezius, laterally by the medial border of the scapula and inferiorly by the upper border of latissimus dorsi. Floor of the triangle is formed by the 7th rib, 6th and 7th intercostal spaces and *rhomboids minor*.

   Importance: This is the only part of the back which is not covered with muscles. Respiratory sounds are better heard with a stethoscope here than that heard elsewhere in the back.

   On the left side, the cardiac end of the stomach lies deep to this triangle and in days before X-rays were discovered, sounds of swallowed liquids were auscultated over this triangle.
9. *Simon's triangle:* Bounded anteriorly by the recurrent laryngeal nerve, posteriorly by the common carotid artery and base is formed by the inferior thyroid artery.

   Importance: Aids in identification of the recurrent laryngeal nerve.
10. *Lumbosacral triangle of Marcille:* It is a triangular interval on each side of the body of the 5th lumbar vertebra, bounded medially by the body of the 5th lumbar vertebra, laterally by the medial border of the psoas muscle, apex by the junction of the psoas major muscle and the body of the 5th lumbar vertebra. Base is formed by the upper surface of the ala of the sacrum and floor by the transverse process of the 5th lumbar vertebra and the iliolumbar ligament.

    Importance: Ureter crosses the common iliac vessels at the lateral angle of the triangle.

11. *Retromolar trigone:* Base overlies the ascending ramus of the mandible from the last molar, apex terminates at the maxillary tuberosity, laterally continuous with the buccal mucosa and medially blends with the anterior tonsillar pillar.
    Importance: Common site for oral malignancy.
12. *Sherren's triangle:* Bounded by the umbilicus, symphysis pubis and the anterior superior iliac spine.
    Importance: Indicates the area of hyperaesthesia in an acute episode of appendicitis.
13. *Scalene triangle:* Bounded by scalenus anticus anteriorly, scalenus medius posteriorly and first rib inferiorly.
    Importance: Subclavian artery and trunks of the brachial plexus pass through the scalene triangle where they may be compressed causing thoracic outlet syndrome.

E. Drugs at a Glance

> *Remember how much you do not know. Do not pour strange medicines into your patients.*
> **—William Osler,** 1903

ANTIBACTERIALS

> *It seems likely that in the next few years a combination of antibiotics with different antibacterial spectra will furnish a "cribrum therapeuticum" from which fewer and fewer infecting bacteria will escape.*
> **—Alexander Fleming,** 1946

SULFONAMIDES

They act by inhibiting folic acid synthetase which converts PABA to folic acid.

Side effects: Intolerance, crystalluria, agranulocytosis, goitre, neuritis, jaundice.

Drugs

Sulfadiazine: 1 gm 6th hourly.
Sulfamethizole: 200 mg 6th hourly.
Sulfadimidine: 1 gm 6th hourly.
Sulphamethoxazole:
Sulfaguanidine: 3 gm 6th hourly.
Sulfamethoxale: 1 gm 6th hourly.
Used in meningitis, UTI, chancroid, trachoma, ulcerative colitis, toxoplasma, Bacillary dysentery.
Silver sulphadiazine cream 1% for burns.
Sulphacetamide as ophthalmic solution.

Cotrimoxazole: Combination of trimethoprim and sulfamethoxazole in a ratio of 1 : 5 in different strengths

Trimethoprim inhibits dihydrofolate reductase enzyme which converts dihydrofolate into tetrahydrofolate. It has got synergistic action with sulphamethoxazole.

It is used as BID dose in typhoid, plague, UTI, prostatitis.

Side effects: Severe skin reactions, megaloblastic anaemia.

QUINOLONES

They act by interfering with the synthesis of DNA.

They are useful against gram-negative organisms, enteric fever, respiratory infections, gastroenteritis, urinary infections, tuberculosis.

Side effects: Allergic reactions, CNS manifestations, hallucinations, nephritis, arrhythmias.

Drugs

Nalidixic acid 1 gm 6th hourly.
Norfloxacin 400 mg BD for 10 days.
Ciprofloxacin 500 mg BD for 10-14 days.
Pefloxacin 400 mg BD.
Ofloxacin 200 mg OD or BD for 10-14 days.
Lomefloxacin 400 mg OD.
Sparfloxacin 200 mg BD.
Levofloxacin 500 mg OD for 7-14 days.
Gatifloxacin 200 mg OD.

He is the best physician who knows the worthlessness of many medicines.—**Benjamin Frunklin**

PENICILLIN

It acts by interfering with the cell wall synthesis of bacteria.

Used in *Streptococcus, Pneumococcus, Meningococcus* and gonococcal infections, syphilis, tetanus, gas gangrene, actinomycosis, plague. It is also used as prophylaxis in rheumatic fever. It is used in pelvic infections, lymphoedema.

Side-effects: Anaphylaxis, serum sickness, Jerisch-Herxheimer reaction.

Drugs:

Benzyl penicillin. 4-10 lakhs 4th or 6th hourly intravenously after test dose.

Procaine-penicillin 4-10 lakhs IM after test dose OD for 7-14 days.

Benzathine penicillin 12-24 lakhs deep IM into the buttocks after test dose once in 3 weeks.

AMPICILLIN/AMOXYCILLIN/TALAMPICILLIN/PIVAMPICILLIN

Used in respiratory infections, meningitis, endocarditis, cutaneous infections.

It is not useful in penicillin-resistant staphylococci infections.

Dose: 500 mg 6th hourly.

METHICILLIN

It is a penicillin group of antibiotic which is used in penicillin resistant staphylococcal infections.

Dose: 1-2 gm. IM 6th hourly.
1 gm in 5-10 ml normal saline can be used as intrapleural or intra-articular therapy.

CLOXACILLIN/DICLOXACILLIN/FLUCLOXACILLIN

These drugs are used in penicillin resistant staphylococcal infections.

Dose: 500 mg 6th hourly.

CARBENICILLIN/TICARCILLIN

It is used in septicaemias, urinary tract infections mainly due to *Pseudomonas* and *Proteus* infections.

Dose: 1 gm 6th hourly IV.

PIPERACILLIN/AZOCILLIN/MEZLOCILLIN

It is active against *Pseudomonas.*

Dose: 2 gm 4th–6th hourly.

CLAVULANIC ACID

It is a beta lactamase inhibitor used against beta lactamase producing bacteria.

It is used along with amoxycillin/ticarcillin/ampicillin.

SULBACTUM

It is a beta lactamase inhibitor.

It is combined with ampicillin.

TAZOBACTUM

It is a beta lactamase inhibitor combined with piperacillin.

CARBAPENEM/IMIPENEM/MEROPENEM

These are bactericidal beta lactam antibiotic.

They are used in septicaemia, *Pseudomonas, Klebsiella* and *Proteus* infections.

LINCOMYCIN/CLINDAMYCIN

They act on bacterial ribosomal RNA.

Side effect: Pseudomembranous colitis.

Dose: 500 mg tds.

VANCOMYCIN/TEICOPLANIN

It inhibits cell wall synthesis.
Side effects: Nephritis, ototoxicity.
Used in pseudomembranous colitis.
Dose: 500 mg 6th hourly IV.

CEPHALOSPORINS

They are bactericidal. They inhibit bacterial cell wall synthesis.
Useful in most of the bacterial infections.
Side effects: Anaphylaxis, Hepatotoxicity, Nephrotoxicity.

Ist Generation	*IInd Generation*
Cephalexin 500 mg TDS oral. Cephadroxyl 500 mg BD oral. Cephadrine 250 mg QID oral. Cephazolin 1 gm BD IV. Cephalothin 2 gm 6th hourly. Cepharin 2 gm 6th hourly.	Cefuroxime 500 mg BD oral. Cefaclor 500 mg 8th hourly. Cefamandone 2 gm 6th hourly. Cefoxitin 2 gm 8th hourly. Cefonicid 2 gm 6th hourly. Ceforanide 2 gm 12th hourly.
IIIrd Generation	*IVth Generation*
Cefetoxime 1 gm 6th or 8th hourly. Ceftriaxone 1 gm 12th hourly. Ceftizoxime 2 gm 8th hourly. Ceftazidime 2 gm 8th hourly. Cefaperazone 2 gm 12th hourly. Cefixime 200 mg BD orally.	Cefepime 200 mg BD orally. Cefpodoxime 200 mg BD orally.

MACROLIDES

They act by inhibiting protein synthesis and also by blocking ribosomal activity.
Side effects: Allergic reactions, gastric irritation, cholestasis.
Uses: Respiratory infections, skin infections.
Drugs:
Erythromycin: 500 mg QID.
Roxithromycin: 150 mg BD.
Azhithromycin 500 mg OD before food for 3 days.
Clarithromycin 500 mg BD. It is also used as anti-*Helicobacter pylori* in treating duodenal ulcer.
Spiramycin 1-5 millions BD.

AMINOGLYCOSIDES

They block the RNA-ribosome combination and also inhibits the enzymes involved in Krebs' cycle.
Side effects: Ototoxicity (vestibular), nephrotoxicity.

Aminoglycoside will not get absorbed through the GIT because of high pH. So they are given either IM or IV. As bowel antiseptic they can be given orally.

Drugs:
1. Streptomycin: 0.75 gm IM OD for 3 months. Used in tuberculosis, plague, chancroid, granuloma inguinale, brucellosis, respiratory infections, gut sterilisation during large bowel preparation.
2. Kanamycin: 1.5 gm/day. It is more cochlear toxic.
3. Gentamicin: 80 mg BD IM or IV. 1 mg/kg body weight. Antipseudomonal drug.
4. Tobramycin 5 mg/kg body weight. More effective against *Pseudomonas* than gentamicin.
5. Amikacin 250 mg BD IM/IV. 15 mg/kg body weight.
6. Netilmycin: 3-6 mg/kg body weight.

7. Neomycin 1 gm 6th hourly. Mainly used in hepatic failure and in gut sterilisation.
8. Framycetin as 0.5% ointment.
9. Paromomycin: 2 gm QID. Used in amoebic dysentery.

TETRACYCLINES

They are bacteriostatic.

They act by inhibiting enzyme system, protein synthesis and ribosomal activity.

Side effects: Hepatic dysfunction, gastric irritation, permanent yellow staining of teeth, suppresses the bone growth.

Uses: Plague, cholera, sexually transmitted diseases like syphilis, gonorrhoea, chancroid, actinomycosis.

It is also used to identify malignant cells as brilliant yellow fluorescence under UV light after giving tetracycline for 5 days.

Dose: 500 mg QID.

Doxycycline: 100 mg OD/BD for 10-14 days.

CHLORAMPHENICOL / THIAMPHENICOL

It interferes with protein synthesis.

Side effects: Bone marrow suppression, Grey-baby syndrome, liver damage.

Uses: Typhoid fever, meningitis, plague.

Dose: 2 gm/day.

DRUGS FOR TUBERCULOSIS

RIFAMPICIN:

It inhibits DNA-dependent RNA polymerase. It is bactericidal.

Side effects: Hepatotoxicity, orange-red coloured urine, flu-like syndrome.

Uses: Tuberculosis.

Other uses: Leprosy, Meningococcal carrier, brucellosis, mycetoma, Q-fever, *Legionella, Chlamydia.*

Dose: 450-600 mg/day (OD) before food.

RIFABUTIN: 150 mg/day.

ISONICOTINIC ACID HYDRAZINE (INH):

It inhibits the synthesis of phospholipid synthesis of cell wall of bacteria. It is bactericidal.

Side effects: **I**ntolerance, **N**euritis, **H**epatitis.

It crosses the blood-brain barrier and placenta.

Dose: 300 mg orally OD.

ETHAMBUTOL: It is bacteriostatic.

Side effects: Gastric intolerance, retrobulbar neuritis (green colour vision is defective).

It crosses the blood-brain barrier and concentrates in CSF.

Dose: 25 mg/kg body weight. 800 mg OD after food.

PYRAZINAMIDE: It is bactericidal.

It acts on dormant bacteria, bacteria inside the macrophages and caseating material.

Side effects: Hepatotoxicity, hyperuricaemia, photosensitivity.

Dose: 1.5 gm OD after food or 750 mg BD after food for 2 months.

MORPHAZINAMIDE: 3 gm/day.

STREPTOMYCIN.

CYCLOSERINE: 2 gm daily. It is a reserve drug.

VIOMYCIN: 1 gm twice a week. IM. It is used in multi-drug-resistant tuberculosis.

AMIKACIN/KANAMYCIN.

CAPREOMYCIN.

PAS: Para-amino salicylic acid: It interferes with PABA metabolism of the bacteria.

Side effects: GIT intolerance, hepatotoxicity, blood abnormalities.

ETHIONAMIDE and PROTHIONAMIDE.

CLARITHROMYCIN.

AZHITHROMYCIN.

ANTIAMOEBIC DRUGS

DEHYDROEMETINE:
It acts on the trophozoite. But not on the cyst.
It is used for extra-intestinal amoebiasis.
Side effects: Myocarditis, myalgia.
Dose: 60 mg deep IM OD for 10 days given under monitor.

DIIDOHYDROXYQUINOLINE: 2 gm/day. It is more useful against trophozoites.

IDOCHLOROHYDROXYQUINOLINE: 0.75 gm/day. It is more useful in cyst passers.

CHLOROQUINE: Used for only extra-intestinal amoebiasis.
500 mg BD for 2 days.
250 mg BD up to 3 weeks.

METRONIDAZOLE:
400-800 mg tds for 10 days. Or 500 mg (1 mg/ml) tid IV.
It is used for intestinal and extra-intestinal amoebiasis.

TINIDAZOLE:
300-600 mg tid for 3-5 days.

SECNIDAZOLE:
2 gm single dose after food for intestinal amoebiasis.
2 gm OD for 5 days.

ORNIDAZOLE:

DILOXANIDE FUROATE:
500 mg tid for 10 days.
Used in cyst passers and chronic carriers.

TETRACYCLINES:
2 gm/day for 10 days.
It is used only as an adjuvant drug.

IMPORTANT ANTIHELMINTHICS

PIPERAZINE CITRATE:
Dose: 5 gm single dose. 1 ml = 150 mg.
Used for Ascariasis.
Side effects: Cerebellar ataxia, vertigo, convulsion.

TETRAMISOLE: 150 mg for adult, 50 mg for children. Used in ascariasis and ankylostomiasis.

MEBENDAZOLE: 100 mg BD for 3 days. 600 mg tid for 21 days.
Used in Ascariasis, Ankylostomiasis, Trichuris trichura, hydatid cyst.

ALBENDAZOLE: 400 mg single dose.

PRAZIQUANTEL: 500 mg tid for 15 days.

THIABENDAZOLE: 25 mg/kg for 3 days.

PYRANTEL PALMOATE: 11 mg/kg. 15 ml. (1 ml = 50 mg).

LUCANTHONE 1 gm tid for 3 days. **HYCANTHONE** 4 mg/kg orally for 4 days.
Used in S. haematobium, S. mansoni.

METRIFONATE: 7.5 mg/kg for 3 days.

NIRIDAZOLE: 25 mg/kg. Used in schistosomiasis.

ANTIMONY COMPOUNDS: Given intramuscularly.
Used in S. haematobium and leishmaniasis.

DIETHYL CARBAMAZINE CITRATE (DEC): 100 mg tid for 21 days.
Used in Filariasis, tropical eosinophilia, larva migrans.

LEVAMISOLE: 150 mg OD.
Used in ascariasis, hookworms, strongyloidosis, as immunomodulator in cancer.

> *It seems likely that in the next few years a combination of antibiotics with different antibacterial spectra will fernish a "cribrum therapeuticum" from which fewer and fewer infecting bacteria will escape.* **—Alexander Fleming**

ANTICOAGULANTS

HEPARIN: 10,000 units IV and later 5,000 subcutaneously 8th hourly.
Uses: DVT, pulmonary embolism, DVT prophylaxis.
Heparin therapy should be monitored by partial thromboplastin time (PTT).
Side effects: Haemorrhage, alopecia.
Protamine sulphate neutralizes the action of heparin. It is given IV slowly. Dose should not exceed 50 mg.

LOW MOLECULAR WEIGHT HEPARINS:
It has got longer duration of action. Monitoring is not required. Given as once a day regime.
Dalteparin 2500-10, 000 IV- SC/IV.
Enoxaparin: 40 mg SC.
Parnaparin.
Fraxiparin.

WARFARIN SODIUM:
It is an oral anticoagulant. It is monitored using prothrombin time.
Available as 5 mg tablets.
Dose: 10 mg OD.
It causes haemorrhage and drug interactions.

OTHER DRUGS

- **VITAMIN K:** 10 mg OD IM for 3-5 days. It is used in bleeding disorders, Vitamin K deficiencies in children, liver disorders.
- **ERYTHROPOIETIN**: 500 IU/kg thrice a week. It is used in chronic renal failure.
- **BROMOCRIPTINE**: *Dose:* 20 mg/day.
  It is used to suppress lactation, galactorrhoea, hyperprolactinaemia, parkinsonism, acromegaly.
- **CLOMIPHENE CITRATE**: Used in infertility.
  Dose: 50 mg for 5 days from 5th day of menstruation.
- **OCTREOTIDE:** 0.1 mg BD. Used after pancreatic surgeries in haematemesis, gut endocrine tumours.
- **VASOPRESSIN:** 20 IU IV/SC. Used in bleeding oesophageal varices.
  DESMOPRESSIN 4 microgram/day-SC-IV-NASAL SPRAY.
  TERLIPRESSIN: 2 mg IV.
- **L-THYROXINE**: 25, 50, 100, 200 microgram tablets.
  Uses: Myxoedema, cretinism for maintenance therapy after total thyroidectomy, as suppressive therapy in papillary carcinoma thyroid (300 μgm) in physiological goitre, in dyshormonogenesis.
- **ANTITHYROID DRUGS:** *Side effects:* Agranulocytosis and alopecia.
  CARBIMAZOLE: 5-20 mg daily exactly 8th hourly.
  METHIMAZOLE: 5-20 mg daily.
  PROPYLTHIOURACIL: 600 mg tid. Used in pregnancy and children.
- **DANAZOL**: It is used in fibrocystic disease of breast, gynaecomastia, mastalgia, infertility, precocious puberty.
  Dose: 800 mg OD.
- **ANTACIDS**: Aluminium hydroxide, magnesium trisilicate, sodium bicarbonate.
  Dose is variable.
- **CARBENOXOLONE:** 100 mg tid. It is used in peptic ulcer.
  It causes water and sodium retention, hypokalaemia.
- **H_2 RECEPTOR BLOCKERS**: Used in peptic ulcer.
  Cimetidine 1000 mg/day.
  Ranitidine 150 mg BD or 300 mg at bed time. Injection ranitidine 50 mg IV 6th-8th hourly.
  Famotidine: 40 mg OD.
  Roxatidine 75 mg OD.
  Nizatidine.
- **PROTON PUMP INHIBITORS:** Used in peptic ulcer.
  Omeprazole: 20 mg OD/BD 1 hour before food.
  Lanzoprazole: 30 mg OD before food.
  Esomeprazole 40 mg OD.
  Pantoprazole 40 mg OD before food.
- **PIRENZEPINE**: 50 mg BD. Used in peptic ulcer.

- **BISMUTH COLLOIDS**: Used in peptic ulcer. It is also anti-*Helicobacter pylori.*
  It stains the oral mucosa.
- **SUCRALFATE:** Used in peptic ulcer.
  Dose: 1 gm tid.
- **ONDANSETRON**: 8 mg BD. Orally or IV. It is used to prevent vomiting before starting chemotherapy.
- **LAXATIVES:**
  Castor oil: 16 ml HS.
  Phenolphthalein 300 mg HS.
  Bisacodyl 5 mg orally. HS or 100 mg as rectal suppositories.
  Isabgol as bulk laxative. 15 mg HS.
  Liquid paraffin: 30 ml HS.
- **ANTIDIARRHOEAL:**
  Bismuth kaolin.
  Diphenoxylate, atropine.
  Loperamide 2-4 mg 8th hourly or as required up to maximum of 16 mg/day dose.
- **PROKINETIC DRUGS:**
  Metoclopramide: Used in vomiting, dyspepsia, hiccough. It causes extra-pyramidal reactions.
  Domperidone: *Dose:* 10 mg tid before food. It does not cause extra-pyramidal reaction as it does not cross the blood-brain barrier.
  Cisapride: 10-40 mg/day before food. It can cause cardiac arrhythmias.
  Mosapride.
- **POTASSIUM:** It should be given only in IV drip slowly. IV *Bolus should not be given*. 2 mEq/ml.
  Dose: 20-40 m equivalents. IV should be given under ECG monitoring.
- **SODIUM BICARBONATE**: 7.5% w/v in correction of acidosis.

Appendix

Important Laboratory Values

URINE

Specific gravity:
Normal: 1.010 to 1.025. Low: less than 1.010. High: more than 1.025.
Fixed: 1.010 to 1.014.
Reaction: Acidic with a pH of 6.0.
Colour: Clear and amber coloured.

Parameter	**Values**
Aldosterone	2-10 µg/day
Nitrogen	0.4-1.0 gm/day
Amylase	30-250 Somogyi units/hour
Calcium	< 3.8 mmol/day
Catecholamines	< 100 U/day
Copper	< 25 µg/day
Creatinine	1.0-1.6 gm/day
Creatinine clearance	140-150 ml/min in males 105-130 ml/min in females
Estrogens	4-25 µg/day in males 5-100 µg/day in females. Very high in pregnancy
17-hydroxy corticosteroids	2-10 mg/day
5-HIAA	2-9 mg/day
17-ketosteroids	7-25 mg/day in men 4-15 mg/day in women
Magnesium	6.0-8.5 mEq/24 hours
Metanephrines	1.3 mg/day
Urine osmolarity	38-1400 mOsm/kg water
Phosphorus	0.9-1.3/day
Porphyrins	
Coprophyrin	50-250 µg/day
Uroporphyrin	10-30 µg/day
Potassium	25-100 mmol/day
Protein	< 150 mg/day
Sodium	100-250 mEq/day
Urobilinogen	1-3.5 mg/day
VMA	< 8 mg/day

BLOOD

Acetoacetic acid	< 0.3 mmol/litre
Acid phosphatase	1.0-5.0 King-Armstrong units
Alkaline phosphatase	20-90 IU/litre

Humor is hazardous to your illness; but not for you.

Amino nitrogen	3.5-5.5 mg/dl
Amylase	60-180 Somogyi units
Ascorbic acid	0.4-1.0 mg/dl
Bicarbonate	23-29 mmol/litre
Bilirubin	Total 0.3-1.0 mg/dl. Direct 0.1-0.3 mg/dl Indirect 0.2-0.7 mg/dl
Calcium	9-11 mg/dl
Ionised	4.5-5.6 mg/dl
CO_2 in plasma	20–30 mmol/L. (50-70 volume %)
CO_2 tension in artery	35-45 mmHg
Ceruloplasmin	27-37 mg/dl
Chlorides	98-106 mmol/L
Cholesterol—total	150-250 mg%
Copper	115 + or –15 µg/dl
Cortisol	5-20 µg/dl
Creatine phosphokinase (CPK).	25-90 units/ml in males. 10 m-70 units/ml in females
Creatinine	< 1.5 mg/dl
Free fatty acids	0.7 mmol/litre
Gastrin	40-200 mg/dl
Glucose	70-110 mg/dl
17-OH corticosteroids	2-10 mg/day
IgG	800-1500 mg/dl
IgM	40-150 mg/dl
IgA	90-320 mg/dl
Insulin	6-26 µU/ml
17-keto steroids	7-25 mg/day in males. 4-15 mg/day in females
Lipase	1.5 units
Magnesium	0.8-1.3 mmol/litre
5' nucleotidase	0.3-2.6 Bodansky units/dl
Osmolality	280-300 mOsm/kg of water
Oxygen.	17-21 volume % in arterial blood 10-16 volume % in venous blood
Oxygen saturation	97% in arterial blood. 60-85% in venous blood
pH of blood	7.36-7.44
Phosphorus, inorganic	1-1.4 mmol/litre
Potassium	3.5-5.0 mmol/litre
Protein—total	5.5-8.0 gm/dl
Albumin	3.5-5.5 gm/dl
Globulin	2.0-3.5 gm/dl
Sodium	136-145 mmol/litre
Sulphate inorganic	0.8-1.2 mg/litre
Testosterone	< 100 ng/dl in females. 300–1000 ng/dl in males
TSH	0-5 IU/ml plasma
Thyroxine (T4)	5-12 µg/dl
Triiodo thyronine (T3)	80-200 ng/dl
SGOT/AST	6-18 units/litre
SGPT/ALT	3-26 units/litre
Uric acid	2.5-8.0 mg/dl in males 1.5-6.9 mg/dl in females
Urea nitrogen (BUN)	10-20 mg/dl

RBC count	4.6-6.2 millions/mm^3 in males 4.2-5.4 millions/mm^3 in females 4.5-5.1 millions/mm^3 in children
Reticulocyte count	25,000-75,000/mm^3
WBC count	4,300-10,000/mm^3
Platelet count	1,50,000-4,40,000/cu mm
Bleeding time	1-4 minutes

Haematocrit	40-54 ml/100 ml
Clotting time	2-15 minutes
Clot retraction time	Apparent in 60 minutes, complete in 24 hours Usually < 6 hours
Plasma fibrinogen	160-400 mg/dl
Partial thromboplastin time	68-82 seconds Activated is 32-46 seconds
Prothrombin time	11-15 seconds
Haemoglobin	14-18 gm/dl males 12-16 gm/dl females 11-16 gm/dl children 16-19 mg/dl newborn
Fetal haemoglobin	< 2%
HaemoglobinA2	1.5-3.5%
Osmotic fragility	Begins in 0.45-0.39% NaCl and completes in 0.33-0.30%
Sedimentation rate (ESR)	< 10 mm in one hour in males < 20 mm in one hour in females

STOOL EXAMINATION

Bulk	100-200 gm
Water	75%
Osmolarity	250 mOsm/L
Colour	Brown
pH	7.0-7.5
Fat	< 7 gm/day
Stercobilinogen	50-280 mg/day
Urobilinogen	30-200 mg/100 gm
Nitrogen	< 2.5 gm/day
Calcium	0.6 gm/24 hours
Trypsin	20-90 units/gram

Further Reading

1. Bailey and Love’s Short Practice of Surgery, 25th edition.
2. Oxford Textbook of Surgery, 2nd edition.
3. Maingot’s Abdominal Operations, 11th edition.
4. Fischer’s Mastery of Surgery, 6th edition.
5. Sabiston Textbook of Surgery, 19th edition.
6. The Washington Manual of Surgery, 5th edition.
7. Farquharson’s Textbook of Operative General Surgery, 9th edition.
8. Surgery of Anus, Rectum and Colon by John Goligher, 5th edition.

Index

A

Abdomen quadrants 773
Abdomen, blunt trauma 150
Abdominal cocoon 621
Abdominal compartment syndrome 827, 829
Abdominal dehiscence 788
Abdominal diagnostic paracentesis 151
Abdominal fluid 663
Abdominal incisions 1259
Abdominal tonsil 1006
Abdominal trauma 148
Abdominal trauma, management 151
Abdominal wall abscess 794
Abdominal wall, Meleney's bacterial gangrene 795
Abdominoperineal resection 1026, 1027
Aberrant renal artery 1084
Aberrant right subclavian artery 841
Abscess 40
Abscess drainage, Hilton's method 43
Abscess, alveolar 386
Abscess, amoebic liver 774
Abscess, anorectal 1049
Abscess, appendicular 781
Abscess, breast 556
Abscess, cold 45, 472, 473, 474
Abscess, collar stud 162, 472
Abscess, complications 42
Abscess, dental 386
Abscess, extradural 1182
Abscess, intracerebral 1182
Abscess, intracranial 1182
Abscess, ischiorectal 1050
Abscess, lung 1201
Abscess, metastatic 46
Abscess, metastatic cerebral 1201
Abscess, parapharyngeal 466
Abscess, parotid 438
Abscess, pelvic 607
Abscess, pelvirectal 1050
Abscess, perianal 1049
Abscess, perinephric 1089
Abscess, psoas 781
Abscess, pyaemic 45, 46
Abscess, pyogenic 40
Abscess, sites 41, 42
Abscess, splenic 723
Abscess, subdural 1182
Abscess, submucous 1050
Abscess, subphrenic 608
Abscess, treatment 43
Abscess, types 40
Acanthosis 392
Accessory appendicular artery of Seshachalam 1005
Accessory parotid tumour 454
Achalasia cardia 850
Achalasia cardia, barium swallow 851
Achalasia cardia, staging 851
Achalasia cardia, treatment 852
Achalasia cardia, triad 850
Achalasia, laparoscopic cardiomyotomy 853
Achalasia, Negus hydrostatic dilatation 852
Achalasia, Plummer's pneumatic dilatation 852
Achlorhydria 872
Achrochordon 84
Achylia 872
Acid phosphatase 1121
Acid-base balance 98
Acidosis, metabolic 99
Acidosis, respiratory 100
Acriflavine 1246
Acrocyanosis 211
Actinomyces israelii 56, 642
Actinomycosis 56
Actinomycosis jaw 386
Acute haemolytic reaction 125
Acute pyomyositis 67
Acute suppurative tenosynovitis 164
Adamantinoma 383
Adelaide coma scale 1175
Adenoma colon 962
Adenoma colon grading 962
Adenoma colon, therapeutic colonoscopy 963
Adenoma sebaceum 295
Adenomas, basophil 1186
Adenomas, chromophobe 1186
Adenomas, eosinophil 1186
Adenomas, prolactin-secreting 1186
Adhesiolysis, laparoscopic 1002
Adhesion, Gilroy Bevan triad 1001
Adhesion, prevention 1003
Adhesions and bands 1000
Adhesions, Child's-Phillips plication 1002
Adhesions, CT scan 1001
Adhesions, Noble's placation 1002
Adiposis dolorosa 71
Admiron's triangular hypothesis 680
Adrenal gland, secondaries 539
Adrenal incidentalomas 538
Adrenal mass 779
Adrenalectomy 542
Adrenalectomy, laparoscopic 543
Adrenocortical carcinoma 539
Adrenocortical carcinoma, Mc Falane staging 539
Adrenocortical tumours 538
Advanced trauma life support 143
Adynamic obstruction 983
Aerophobia 59
α-fetoprotein 645
Aflatoxin B1 643
AGES scoring 510, 512
Aggressive fibromatosis 84, 326, 791
Ainhum 215
Alcohol effects on pancreas 728
Alemtuzumab 356
Alkalosis, metabolic 99
Alkalosis, respiratory 99
Allograft 340
Allopurinol 1097
Alopecia, moth eaten 28
Alpha fetoprotein 628
Altemeir's rectosigmoidectomy 1034
Ambulatory venous pressure 242
Ameloblastoma 383
AMES scoring 510, 512
Aminoglutethimide 585, 1228
Amoebiasis cutis 632
Amoebic colitis, acute fulminant 961
Amoebic colitis, complications 961
Amoebic liver abscess, aspiration 634
Amoebic liver abscess, chronic 633
Amoebic liver abscess, percutaneous drainage 634
Amoebic liver abscess, sequelae 632
Amoebic liver abscess, treatment 634
Amoebic typhlitis 960
Amoeboma 781, 960
Ampulla of Vater 670
Amputation of penis, partial 1145

Amputation of penis, total 1145
Amputation stump, conical bearing 331
Amputation, 'Peg-leg' 335
Amputation, above-knee 213, 335
Amputation, below-knee 213, 334
Amputation, Boyd's 333
Amputation, Buerger's 213
Amputation, Burgess 334
Amputation, Chopart's 333
Amputation, complications 337
Amputation, forequarter 336
Amputation, Gritti-Stokes 335
Amputation, Guillotine 336
Amputation, hind quarter 335
Amputation, ideal stump 330
Amputation, incisions 331
Amputation, interscapulothoracic 336
Amputation, Krukenberg 336
Amputation, Lisfranc's 332
Amputation, modified Syme's 333
Amputation, patient evaluation 331
Amputation, Pirogoff's 334
Amputation, principles 331
Amputation, prosthesis 339
Amputation, Ray's 332
Amputation, Sir-Gordon Taylor's 335
Amputation, Syme's 333
Amputation, transmetatarsal (Gillies') 332
Amputation, Wagner's 333
Amputations 330
Amputations, types 330
Amsterdam criteria 965
Amsterdam criteria, modified 965
Amylase creatinine clearance ratio 730
Anaemia, Cooley's 718
Anaemia, erythroblastic target cell 718
Anaemia, mediterranean 718
Anaesthesia, caudal 1237
Anaesthesia, epidural 1236
Anaesthesia, general 1232
Anaesthesia, instruments 1101
Anaesthesia, intravenous regional 1235
Anaesthesia, postoperative care 1234
Anaesthesia, preoperative assessment 1231
Anaesthesia, regional 1235
Anaesthesia, spinal 1235
Anaesthesia, topical 1235
Anaesthetic agents 1232
Anaesthetic, gaseous 1232
Anaesthetics, intravenous 1232
Anaesthetics, volatile 1232
Anal canal, anatomy 1021
Anal canal, malignant tumours 1057
Anal canal, malignant tumours, AJCC staging 1058
Anal canal, sphincters 1021
Anal carcinoma, NIGRO regime 1058
Anal circling 1034
Anal cushions 1039
Anal dilatation 1042
Anal fissure, acute 1047
Anal fissure, chronic 1047
Anal fissure, chronic, dorsal fissurectomy 1048
Anal fissure, chronic, lateral anal sphincterotomy 1048
Anal fistula plug repair 1055
Anal incontinence 1059
Anal incontinence, secca therapy 1059
Anal intraepithelial neoplasia (AIN) 1057
Anal manometry 1032
Anal margin tumours 1057, 1058
Anal sphincter, external 1021, 1047
Anal sphincter, internal 1021, 1046
Anaplasia 282
Anastomosis, Kimura's diamond shaped 990
Anchovy sauce 631, 632
Anchovy sauce pus 33
Andy gump deformity 399
Anel's procedure 203
Aneurysm 44
Aneurysm, Berry 1182
Aneurysm, carotid artery 209
Aneurysm, cirsoid 225
Aneurysm, dissecting 209
Aneurysm, intracranial 1182
Aneurysm, mycotic 1183
Aneurysm, popliteal 208
Aneurysm, ruptured aortic 207
Aneurysm, splenic artery 722
Aneurysms 202
Aneurysms, abdominal 203
Aneurysms, abdominal, aortic 204, 205
Aneurysms, berry 202
Aneurysms, dissecting 202
Aneurysms, endoluminal stent 207
Aneurysms, false 202
Aneurysms, mycotic 203
Aneurysms, peripheral 208
Aneurysms, true 202
Angelchik prosthesis 846
Angiodysplasia, colon 974
Angiogram, aortic 180
Angiogram, coronary 1216
Angiogram, renal 1069
Angiogram, transbrachial 187
Angiography 180
Angiography, digital substraction, DSA 180
Angiography, indication 180
Angiography, retrograde transfemoral Seldinger 180
Angiolipoma 72
Angiomyolipoma 1098
Angiomyoneuroma 82
Angioplasty, balloon 216
Angioplasty, coronary balloon 1222
Angioplasty, percutaneous transluminal balloon 190
Angioscope 192
Angioscopy 216
Angle of 'His' 842
Anion gap 100
Anorchism 1161
Anorectal malformation, 'Wingspread' classification 1035
Anorectal malformation, treatment 1036
Anorectal malformations 1034
Anorectal myectomy 951
Anorectal strictures 1056
Antacids 884
Anterior resection 1028
Anthrax 61
Antibioma 42, 556, 558
Anticoagulants 252
Anticoagulants, *in vitro* 253
Anticoagulants, oral 253
Anti-*Helicobacter* regime 884
Anti-HIV antibody 63
Antilymphocytic globulin 355, 1229
Antimalignancy drugs 1227, 1228
Antiplatelet antibodies 718
Antiplatelet drugs 253
Antirabies serum 60
Antirabies vaccines, indication 59
Antireflux surgeries 845
Antisepsis 1244
Antitetanus globulin 51
Antitetanus serum 51
Antithyroid drugs 506
Antitubercular drugs 474
Antrectomy, de Miguel's 881
Antylus procedure 203
Aortic aneurysm 778
Aortic clamping 152
Aortic regurgitation 1221
Aortic stenosis 1221
Aortocaval fistula 205
Aortoenteric fistula 205
Aortofemoral bypass graft 184
Aortoiliac endarterectomy 183
Aortoiliac occlusive disease 183
Aortoiliac occlusive disease, types 183
APACHE II scoring system 890
Ape thumb deformity 276
Appendicectomy 1011, 1262
Appendicectomy, complications 1012
Appendicectomy, Fowler-Weir approach 1011
Appendicectomy, Fowler-Weir incision 1262
Appendicectomy, grid-iron incision 1262
Appendicectomy, incidental 1013
Appendicectomy, interval 1014, 1015
Appendicectomy, Lanz crease incision 1262
Appendicectomy, laparoscopic 1017
Appendicectomy, laparoscopic assisted 1018
Appendicectomy, negative 1014
Appendicectomy, retrograde 1012, 1262
Appendicectomy, Rutherford Morison incision 1262
Appendicectomy, troubles in 1013
Appendicitis sequelae 1010
Appendicitis, acute nonobstructive 1006, 1007

Appendicitis, acute obstructive 1007, 1008
Appendicitis, Alvarado scoring 1011
Appendicitis, Alvarado scoring, Kalam modified 1011
Appendicitis, children 1008
Appendicitis, chronic 1014
Appendicitis, CT scan 1010
Appendicitis, differential diagnosis 1009
Appendicitis, elderly 1008
Appendicitis, epidemic 1006
Appendicitis, grumbling 1014
Appendicitis, infancy 1008
Appendicitis, pregnancy 1009
Appendicitis, recurrent 1008
Appendicitis, simple 1014
Appendicitis, sonographic criterias 1010
Appendicitis, stump 1008, 1014
Appendicitis, subacute 1008
Appendicular abscess 1007, 1015
Appendicular mass 781, 1007, 1014
Appendix, carcinoid tumour 1017
Appendix, cystic neoplasms 1017
Appendix, duplication (Wallbridge classification) 1005
Appendix, duplication 1005
Appendix, mucocele 1007, 1016
Appendix, neoplasms 1017
Appendix, primary adenocarcinoma 1017
Appendix, surgical anatomy 1005
Apple-jelly nodule 26
Apudomas 536, 755
Arc of Riolan 948
Arimidex 585
Arm-foot venous pressure 242
Arterial diseases 175
Arterial diseases, treatment 190
Arterial graft 192
Arterial occlusion, acute 196
Arterial occlusion, traumatic 197
Arterial pile 1041
Arterial substitutes 220
Arteries, lower limb 175
Arteries, upper limb 175
Arteriovenous fistula 225
Arteriovenous fistula, acquired 227
Arteriovenous fistula, congenital 225
Artery, inferior thyroid 485, 523
Artery, superior thyroid 485
Artificial blood 127
ASCA 927
Ascending urethrogram 1069
Ascites 661
Ascites in portal hypertension 663
Ascites, encysted 622
Ascitic tap 662, 663
Ascoli's thermoprecipitation test 61
Asepsis 1244
Asepsis, wound scoring 36
Asepto syringe 1109
Askanazy cells 519
Astrocytomas 1183
Astrup formula 100
Atherectomy 191
Atherosclerosis 182
Atherosclerosis, management 183
Atherosclerosis, pathogenesis 182
Athlete's foot 172
Athyreosis 486
Atkinson tube 865
Atresia, apple peel 991
Atresia, duodenal 989
Atresia, intestinal (small intestine) 990
Atresia, intestinal, Griesfield modification of Martin's classification 991
Atrial septal defect, primum type 1218
Atrial septal defect, secundum type 1218
Atrial septal defect, sinus venosus type 1218
Atrophic blanche 238
Audit, clinical 1275
Audit, surgical 1275
Auerbach's plexus 839, 870
Autoclave 1244
Autograft 340
AV node 1216
Axillary nerve sampling 547
Axillary nodes management 589
Axillary nodes, Berg's levels 547
Axillary tail of Spence 545
Axonotmesis 273
Azathioprine 356, 927, 1229

B

Bacillus anthracis 61
Back wash ileitis 954
Bactec MGIT broth culture 619
Bacteraemia 45
Baker's tube 1002
Balanoposthitis 1141
Ball valve mechanism 801
Bandage, spica 1272
Bandages 1272
Bandages, cravet 1273
Bandages, roller 1272
Bandages, 'T' 1272
Bandages, tailed 1272
Bandages, triangular 1272
Bandages, tubular 1272
Bandaging, principles 1273
Bands 1003
Bariatric surgery 107
Barium enema 981
Barium enema, air contrast 982
Barium meal study 873
Barium swallow 841
Barnard's aphorism 610
Barrett's ulcer 849
Basal cell adenoma 445
Basal cell carcinoma, treatment 306
Basal cell naevus syndrome 303
Basal cell papilloma 298
Bascom technique 1038
Base ball finger 170
Bassini's repair 820
Bassini's stitch 808
Bat rabies 58
Bazin's disease 27
Beck's triad 1210
Bednar's tumour 296
Bedsore 21, 214
Beger's procedure 744
Beger's lavage 732
Belatacept 356
Belly cleft 793
Bench surgery 1094
Benign prostatic hyperplasia 1122
Benign prostatic hyperplasia, drugs in 1124
Benign prostatic hyperplasia, surgical management 1123
Benzidine test 1076
Berry picking 512
Berry's ligament 484, 492, 520
Bethesda criteria 965
Bevacizumab 585
Bile 670
Bile duct carcinoma 702
Bile duct injuries 708
Bile duct injuries, Hannover classification 709
Bile duct injuries, Stewart-Way classification 709
Bile duct injuries, Strasberg classification 708
Bilhemia 705
Biliary atresia 677
Biliary colic 683
Biliary dyskinesia 710
Biliary fistulas 703
Biliary obstruction, Benjamin's classification 696
Biliary papilloma, benign 699
Biliary sphincter dyskinesia 710
Biliary strictures 697
Biligram 671
Biliopancreatic diversion 110
Billing gate hump 86
Billroth I gastroduodenal anastomosis 881, 886
Biloma 709
Binachi's surgery 942
Biobrane 136
Biochemotherapy 1227
Biological valves 1222
Biopsy 286
Biopsy, excision 289
Biopsy, frozen section 290
Biopsy, incision 288
Biopsy, trucut 289
Biosyn 1256
Bisgaard method 250
Bite, bee 363
Bite, black widow spider 362
Bite, brown spider 363
Bite, mammalian 363
Bite, snake 362
Bite, spider 362

Bladder carcinoma, transitional cell 1112
Bladder carcinoma, treatment 1114
Bladder diverticula 1118
Bladder diverticula, acquired 1118
Bladder diverticula, congenital 1118
Bladder mass 781
Bladder outlet obstruction 1125
Bladder systolic 1082
Bladder tumour, carcinoma *in situ* 1113
Bladder tumour, Helmstein balloon degeneration 1114
Bladder tumour, intravesical chemotherapy 1114
Bladder tumour, muscle invasive 1113
Bladder tumour, radiotherapy 1114
Bladder tumour, superficial 1112
Bladder tumours 1112
Bladder tumours, Jewett-Strong-Marshall staging 1113
Bladder tumours, TNM staging (AJCC) 1113
Bladder, autonomous 1117
Bladder, cricket ball 1134
Bladder, extraperitoneal rupture 1115, 1116
Bladder, intraperitoneal rupture 1115, 1116
Bladder, motor paralytic 1117
Bladder, neurogenic 1117
Bladder, reflux neurogenic 1117
Bladder, sensory paralytic 1117
Bladder, systolic 1110, 1111
Bladder, thimble 1082, 1110, 1111
Bladder, uninhibited neurogenic 1117
Blair's incision 438
Bleeding duodenal ulcer, endoscopy 892
Bleeding duodenal ulcer, sclerotherapy 892
Bleeding gastric ulcer 893
Bleomycin 482
Block dissection, complication 481
Block, Bier's 1235
Block, field 1235
Block, infiltration 1235
Block, nerve 1235
Block, saddle 1236
Block, Winnie's 1235
Blood coagulation, mechanism 129
Blood fractions 124
Blood loss, measurement 122
Blood substitutes 126
Blood transfusion 124
Blood transfusion, autologus 127
Blood transfusion, blood collection 124
Blood transfusion, complication 125, 126
Blood transfusion, donor criteria 124
Blood transfusion, massive 126
Blood vessels, palpation 179
'Blow-out' fracture, orbit 376
Blue toe syndrome 207
Bluedome cyst of Bloodgood 551
Blumer's shelf 749, 907, 909
Blunt abdominal trauma 150
BM Achauer classification of burns contracture 137
Body mass index 106
Boil 46
Boiling 1244
Bolus obstruction 984
Bonin's Y-V plasty 1134
Borchardt's triad 900
Bottle nose 297
Botulinum toxin 853
Bouin's fluid 1163
Bourneville's disease 295
Bowen's disease 298
Boyle's apparatus 1233
Brachytherapy 1225
Brachytherapy, interstitial 1225, 1226
Brachytherapy, intracavitary 1225, 1226
Brachytherapy, surface 1226
Bradley's repeated laparotomy 732
Brain tumours, primary 1183
Brain death, criteria 353
Brain ischaemia 1173
Brain necrosis 1173
Brain oedema 1173
Brain, secondaries 1183
Branchia fistula, acquired 461
Branchia fistula, congenital 460
Branchial cyst 459
Branchial cyst, treatment 460
Brasdor's ligation 203
Breast cancer, Foote Stewart classification 565
Breast carcinoma 563
Breast carcinoma, tumour receptors 577
Breast carcinoma, advanced 590
Breast carcinoma, axillary clearance 589
Breast carcinoma, axillary sampling 579, 589
Breast carcinoma, bilateral 573
Breast carcinoma, biological behaviour 569
Breast carcinoma, bone secondaries 592
Breast carcinoma, BRCA genes 563
Breast carcinoma, chemotherapy 586
Breast carcinoma, chemotherapy regimes 586
Breast carcinoma, classification 565
Breast carcinoma, Columbia classification 575
Breast carcinoma, conservative surgeries 582
Breast carcinoma, core biopsy 576
Breast carcinoma, cutaneous manifestation 569
Breast carcinoma, discharge study 577
Breast carcinoma, ductal endoscopy 579
Breast carcinoma, ductography 579
Breast carcinoma, early, management 586
Breast carcinoma, edge biopsy 577
Breast carcinoma, Elston-Ellis modified Bloom-Richardson grading 569
Breast carcinoma, excision biopsy 576
Breast carcinoma, Fischer concept of spread 564
Breast carcinoma, FNAC 575
Breast carcinoma, FNAC scoring 576
Breast carcinoma, grading of the tumour 568
Breast carcinoma, haematogenous spread 571
Breast carcinoma, Halsted concept of spread 564
Breast carcinoma, hormone therapy 584, 585
Breast carcinoma, image guided biopsy 576
Breast carcinoma, locally advanced 590
Breast carcinoma, lymphatic spread 570
Breast carcinoma, male 593
Breast carcinoma, Manchester staging 573
Breast carcinoma, metastatic 591
Breast carcinoma, metastatic, chemotherapy 591
Breast carcinoma, metastatic, hormone therapy 591
Breast carcinoma, metastatic, radiotherapy 591
Breast carcinoma, metastatic, treatment 591
Breast carcinoma, modified Gail risk assessment model 564
Breast carcinoma, MRI breast 577
Breast carcinoma, neoadjuvant chemotherapy 586
Breast carcinoma, newer investigation modalities 579
Breast carcinoma, oestrogen receptor study 576
Breast carcinoma, Paget's cells 567
Breast carcinoma, pleural effusion 592
Breast carcinoma, pregnancy 594
Breast carcinoma, progesterone receptor status 577
Breast carcinoma, prognostic factors 592
Breast carcinoma, radiotherapy 583
Breast carcinoma, sentinel node biopsy 578, 589
Breast carcinoma, thermography 579
Breast carcinoma, TNM staging 573
Breast carcinoma, triple assessment 579
Breast carcinoma, trucut biopsy 576
Breast carcinoma, ultrasound 575
Breast carcinoma, wide local excision 582, 587
Breast conservative surgery 587
Breast cyst 559
Breast implants 596
Breast reconstruction 594
Breast self-examination 573
Breast, anatomy 544
Breast, ANDI 548, 551
Breast, atrophic scirrhous 567
Breast, blood supply 546
Breast, cystic swelling 560
Breast, ductal carcinoma 564
Breast, ductal carcinoma *in situ* 564, 565
Breast, fibrocystic disease 551

Breast, frozen section biopsy 576
Breast, hard swellings 597
Breast, inflammatory carcinoma 566, 591
Breast, lactating carcinoma 566
Breast, lactational abscess 556
Breast, lobular carcinoma 564
Breast, lobular carcinoma *in situ* 564, 567
Breast, lymphatic drainage 546
Breast, massive enlargement 597
Breast, medullary carcinoma 566
Breast, milk secretion 546
Breast, nerve supply 546
Breast, nerves-related 546
Breast, nipple discharge 597
Breast, nipple inversion 597
Breast, nipple retraction 597
Breast, nonlactational abscess 557
Breast, scirrhous carcinoma 566
Breast, secretory apparatus 544
Breast, topography 544
Breast, transcoelomic spread 572
Breast, tuberculosis 559
Brittle diabetes 744
Bronchial cast 141
Bronchoscopy 1199
Bronchoscopy, flexible 1199
Bronchoscopy, rigid 1199
Brown's vasomotor index 181
Brunner gland adenoma 936
Bubo 28
Buerger's angle of vascular insufficiency 179
Buerger's disease 184
Buerger's, exercise 188
Buerger's, position 188
Buffalo hump 260
Bunion 86, 173
Bupivacaine 1235
Burhenne technique 693
Burn contracture classification, BM Achauer 137
Burn contracture classification, Ifeanyichukwu 137
Burns 130
Burns contracture 137
Burns contracture, classification 137
Burns contracture, complication 137
Burns contracture, treatment 137
Burns, acid 141
Burns, alkali 141
Burns, assessment 132
Burns, assessment of airway injury 133
Burns, biologic dressings 136
Burns, chemical 141
Burns, classification 130
Burns, effects 134
Burns, first aid 134
Burns, first degree 130
Burns, fluid resuscitation 135
Burns, fourth degree 131
Burns, full thickness 132
Burns, inhalation 141
Burns, local management 135
Burns, management 134
Burns, Marjolin's ulcer 137
Burns, medicolegal aspects 142
Burns, metabolic changes 133
Burns, partial thickness 132
Burns, pathophysiology 132
Burns, second degree 131
Burns, sepsis 133
Burns, third degree 131
Burns, topical agents 136
Burns, toxic shock syndrome 134
Burns, types 130
Burns, wound coverage 136
Bursa anserina 86
Bursa bicipito-radial 85
Bursa, adventitious 86
Bursa, anatomical 85
Bursa, Brodie's 86
Bursa, complication 85
Bursa, infrapatellar 85
Bursa, olecranon 85
Bursa, porter's 86
Bursa, prepatellar 85
Bursa, psoas 85
Bursa, radial 164
Bursa, retrocalcaneum 86
Bursa, semimembranosus 86
Bursa, subacromial 85
Bursa, subhyoid 85
Bursa, tailor's 86
Bursa, ulnar 164
Bursa, Weaver's 86
Bursae 85
Bursae, types 85
Bursitis 85
Bursitis, retro-Achilis 86
Bursitis, subcalcaneal 86
Burst abdomen 788
Burst abdomen, modified Jones closure 790
Buschke Lowenstein tumour 1057

C

Cadaver donor 353
Caecal bascule 997
Café-au-lait spots 91, 92
Calcifying epithelioma of Malherbe 295
Calcinosis cutis 89
Calcitonin 518, 519
Calcitonin gene related peptide 1161
Calcium 531
Calculus, staghorn 1090, 1097
Calculus, urethral 1134
Calculus, vesical 1106
Calculus, uerteric 1094
Calf musculovenous pump 231
Callosity 170
Calot's triangle 670
Campbell de Morgan spots 229
Canal of Nuck 800
Canalicular adenoma 445
Cancer, Chimney sweep 299
Cancer, Kang 299
Cancer, Khangri 299
Cancer, management strategy 291
Cancer, tear 303
Cancer-en-cuirasse 569
Cancrum oris 391
Cantlie's line 666
Capitonnage 641
Capsule endoscopy 942
Carbimazole 506
Carbuncle 47
Carbuncle, renal 48
Carcinoembryonic antigen 970
Carcinogens, chemical 284
Carcinogens, microbial 284
Carcinogens, radiation 284
Carcinoid crisis 940
Carcinoid syndrome 939, 940
Carcinoid tumour 939
Carcinoma 282
Carcinoma alveolus 423
Carcinoma cheek 396
Carcinoma cheek, advanced 401
Carcinoma cheek, approaches 406
Carcinoma cheek, biological behaviour 398
Carcinoma cheek, Broder's classification 401
Carcinoma cheek, chemotherapy 407
Carcinoma cheek, clinical features 399
Carcinoma cheek, halitosis 399
Carcinoma cheek, premalignant lesions 396
Carcinoma cheek, radiotherapy 407
Carcinoma cheek, reconstruction after surgery 405
Carcinoma cheek, staging 399
Carcinoma cheek, TNM staging 399
Carcinoma cheek, treatment 401
Carcinoma cheek, verrucous 398
Carcinoma colon, adjuvant therapy 973
Carcinoma colon, air contrast barium enema 970
Carcinoma colon, Aster-Coller's staging 967
Carcinoma colon, barium enema 970
Carcinoma colon, chemotherapy 973
Carcinoma colon, colonoscopy 970
Carcinoma colon, Cunderson-Sosin staging 968
Carcinoma colon, Duke's histological grading 966
Carcinoma colon, Duke's staging 967
Carcinoma colon, local complications 970
Carcinoma colon, local recurrence 974
Carcinoma colon, modified Duke's staging 967
Carcinoma colon, Owen Wangensteen's second look operation 972
Carcinoma colon, radiotherapy 973
Carcinoma colon, screening 970
Carcinoma colon, TNM staging 968
Carcinoma colon, treatment 971
Carcinoma floor of the mouth 423

Carcinoma gallbladder 699
Carcinoma gallbladder, Nevin's staging 700
Carcinoma gallbladder, TNM staging 700
Carcinoma gallbladder, treatment 701
Carcinoma hard palate 427
Carcinoma *in situ* 284
Carcinoma lip 410
Carcinoma lip, treatment 412
Carcinoma oesophagus 840, 858
Carcinoma oesophagus, endoscopic mucosal resection 860
Carcinoma oesophagus, palliative treatment 864
Carcinoma oesophagus, TNM staging 861
Carcinoma oesophagus, treatment 860
Carcinoma oesophagus, types 859
Carcinoma of gingivobuccal complex 395
Carcinoma of penis, Broder's grading 1144
Carcinoma pancreas 747
Carcinoma pancreas, barium meal X-ray 752
Carcinoma pancreas, biopsy 751
Carcinoma pancreas, CA19-9 751
Carcinoma pancreas, endosonography 751, 752
Carcinoma pancreas, ERCP 751, 752
Carcinoma pancreas, laparoscopic US 751
Carcinoma pancreas, periampullary 748
Carcinoma pancreas, preoperative preparartion 752
Carcinoma pancreas, PTC 751
Carcinoma parathyroid 532
Carcinoma penis 1142
Carcinoma penis, Jackson's staging 1145
Carcinoma penis, radiotherapy 1146
Carcinoma penis, TNM staging 1145
Carcinoma penis, treatment 1145
Carcinoma peritonei 614
Carcinoma prostate 1126
Carcinoma prostate, Gleason grading 1126
Carcinoma prostate, TNM staging 1126
Carcinoma rectum 1023
Carcinoma rectum, tumour vaccine 1029
Carcinoma rectum, chemoradiation 1029
Carcinoma rectum, chemotherapy 1029
Carcinoma rectum, CT scan 1025
Carcinoma rectum, Duke's staging 1024
Carcinoma rectum, endorectal ultrasonography 1025
Carcinoma rectum, Haggitt's invasion of malignant polyp 1024
Carcinoma rectum, local recurrence 1027
Carcinoma rectum, lower anterior resection 1026, 1028
Carcinoma rectum, modified Duke's staging 1024
Carcinoma rectum, radiotherapy 1029
Carcinoma rectum, TNM staging 1024
Carcinoma rectum, total mesorectal excision 1026
Carcinoma stomach 902
Carcinoma stomach, adjuvant therapy 913
Carcinoma stomach, aetiology 902
Carcinoma stomach, barium meal study 910
Carcinoma stomach, biological behaviour 906
Carcinoma stomach, Birmingham staging 908
Carcinoma stomach, Borrmann's classification 904
Carcinoma stomach, chemoradiotherapy 913
Carcinoma stomach, chemotherapy 913
Carcinoma stomach, colloid 908
Carcinoma stomach, endomucosal resection 913
Carcinoma stomach, endosonography 910
Carcinoma stomach, immunochemotherapy 913
Carcinoma stomach, Japanese classification 904
Carcinoma stomach, laparoscopy 911
Carcinoma stomach, Lauren's classification 904
Carcinoma stomach, lymph node station (Japan) 913
Carcinoma stomach, lymphatic spread 906
Carcinoma stomach, Ming's classification 904
Carcinoma stomach, palliative treatment 914
Carcinoma stomach, precancerous lesion 904
Carcinoma stomach, precursor lesions 903
Carcinoma stomach, presentations 908
Carcinoma stomach, sentinel node biopsy 911
Carcinoma stomach, Siewert classification 905
Carcinoma stomach, TNM staging 907
Carcinoma stomach, transperitoneal spread 907
Carcinoma stomach, WHO histological classification 905
Carcinoma tongue 417
Carcinoma tongue, chemotherapy 422
Carcinoma tongue, clinical features 418
Carcinoma tongue, Commando operation 421
Carcinoma tongue, posterior one-third/ base 422
Carcinoma tongue, Price-Hill regimen 422
Carcinoma tongue, radiotherapy 422
Carcinoma tongue, spread 420
Carcinoma tongue, terminal events 422
Carcinoma tongue, treatment 421
Carcinoma, basal cell 303
Carcinoma, basisquamous 303
Carcinoma, branchiogenic 476, 478
Carcinoma, caecum 781
Carcinoma, cuniculatum 302
Carcinoma, cylindromatous 446
Carcinoma, Merkel cell 295
Carcinoma, nasopharyngeal 423
Carcinoma, squamous cell 24
Carcinoma, verrucous 24, 300, 301
Card test 458
Cardiac arrest 118
Cardiac catheterisation 1216
Cardiac massage, external compression 119
Cardiac massage, internal open 119
Cardiac pacing, postoperative management 1223
Cardiac pacemakers 1223
Cardiac patient, preoperative assessment 1216
Cardiac tamponade 1210
Cardiography 1216
Cardiopulmonary bypass 1217
Cardiospasm 850
Caries spine 1190
Carney's triad 917
Carotid blow out 482
Carotid body 468
Carotid body tumour, Shamblin's classification 468
Carotid body, syncope 468
Carpopedal spasm 538
Carr's postulates 1090
Catchpole regime 1000
Catgut, chromic 1256
Catgut, plain 1256
Catheter, Fogarty 198
Catheter, Foley's 1072
Catheter, Malecot's 1073
Catheter, red rubber 1073
Catheterisation, indwelling 1072
Catheterisation, intermittent 1072
Catheters 1071
Causalgia 275
Cautery, bipolar 1258
Cautery, blended 1258
Cautery, coagulation 1258
Cautery, cutting 1258
Cautery, unipolar 1258
Cavernous sinus thrombosis 37
CBD stones, primary, treatment 693
CBD stones, retained/residual, management 693
CBD strictures 697
CBD strictures, Bismuth's classification 697
Celestin tube 865
Celiac angiography 645
Cellcycle 1227
Cellulitis 37
Central node compartment neck dissection 515
Central venous pressure 116

Cerebral concussion 1172
Cerebral contusion 1172
Cerebral laceration 1173
Cervical rib 457
Cervical rib, treatment 459
Cetrimide 1245
CFTR 739
Chancroid 28
Charcot's triad 691
Charles Huggins 1128
Chassaigne tubercle 179
Cheek, anatomy 396
Cheilitis 408
Cheilitis, actinic 409
Cheilosis 391
Chemodectoma 468
Chemotherapy 1227
Chemotherapy, head and neck cancers 482
Chernobyl nuclear disaster 509
Cherry angiomas 229
Cherubism 388
Chest injuries, classification 1194
Chest injuries, management 1195
Chest injuries, thoracotomy indications 1195
Chest wall tumours, benign 1209
Chest wall tumours, primary malignant 1209
Chiba needle 672
Chick embryo cell culture vaccine 59
Chief cell 870
Child turcote pugh (CTP) scoring system 357
Child's grading 659
Child's grading, Pugh's modification 659
Chiropady 188
Chlorhexidine 1245
Cholangiocarcinoma 702
Cholangiocarcinoma, perihilar, Bismuth classification 702
Cholangiocarcinoma, risk factors 702
Cholangiocarcinoma, TNM staging 702
Cholangiogram IV 671
Cholangiogram, peroperative 673
Cholangiogram, postoperative T-tube 673
Cholangitis, sclerosing 699
Cholangitis, sclerosing, primary 699
Cholangitis, sclerosing, secondary 699
Cholecystectomy 705
Cholecystectomy, complications 706
Cholecystectomy, extended 701
Cholecystectomy, laparoscopic 705, 706
Cholecystectomy, laparoscopic, single incision 708
Cholecystectomy, open 705
Cholecystectomy, prophylactic 705
Cholecystitis acute, complications 684
Cholecystitis acute, treatment 686
Cholecystitis glandularis proliferans 689
Cholecystitis, acute 684
Cholecystitis, acute acalculous 684, 686, 694
Cholecystitis, chronic 688
Cholecystitis, emphysematous 685
Cholecystitis, *Salmonella* 694
Cholecystitis, xanthogranulomatous 694
Cholecystolithiasis 110
Cholecystoses 689
Cholecystostomy 686, 752
Choledochal cyst 675
Choledochal cyst, Todani modification of Alonso-Lej classification 676
Choledochal cyst, treatment 678
Choledochoduodenostomy 693
Choledochojejunostomy 693, 745
Choledocholithiasis 690
Choledocholithiasis, complications 691
Choledocholithiasis, primary stones 690
Choledocholithiasis, recurrent stones 691
Choledocholithiasis, retained stones 691
Choledocholithiasis, secondary stones 691
Choledochotomy 692, 693
Cholesterol crystals 460
Cholesterol polyposis 689
Cholesterosis 689
Chondrosarcoma 327
Chordee 1133
Chordee 1140, 1141
Chordee, dorsal 1141
Chordee, ventral 1141
Chordoma 93
Choudary Millis criteria 574
Christmas disease 129
Chronic duodenal ileus 898
Chronic venous insufficiency 237
Chylocele 1152
Chylothorax 271
Chylous ascites 271
Chyluria 271
Cidex 1246
Cilostazol 218
Circle of death 817
Circulatory failure 122
Circumcision 1139, 1140
Circumcision 1260
Cisplatin 482
Cisterna chyli 255
Clamp, Moynihan's occlusion 1251
Clamp, Payr's crushing 1251
Classification, deBakey's 210
Classification, Stanford 210
Claudication, arterial 176
Claudication, neurogenic 176
Claudication, venous 176
Claw hand 278
Claw hand, median 279
Claw hand, ulnar 278
Cleft disorder, LAHS classification 368
Cleft disorder, problems 368
Cleft lip 368
Cleft lip, Millard criteria 369
Cleft palate 368, 370
Cleft palate, treatment 372
Cleft surgery, Delaire timing 370
Clergyman's knee 85
Climatic bubo 29
Closed loop obstruction 969, 986, 972
Clostridium difficile 960
Clostridium tetani 49
Clostridium tetani, exotoxin 50, 53
Clostridium welchii 52
Clostridium welchii, exotoxins 53
Clotting factors 129
Clutton's joint 28
Clutton's joint 28
Coagulation cascade system 129
Coagulation, extrinsic 129
Coagulation, intrinsic 129
Coarctation of aorta 1218
Coffin corner 395
Coin test 1192
Cold abscess 474, 1192
Cold abscess, sequelae 475
Cold recovery time 189
Colitis, ischaemic 959
Colitis, pseudomembranous 960
Colitis, ulcerative 954
Collecting duct of Bellini carcinoma 1100
Collis' vertical gastroplasty 846
Colloid bismuth sulphate 884
Colloid carcinoma rectum 1024
Colocalisation hypothesis 729
Colon, benign tumours/polyp 961
Colon, diverticular disease 951
Colon, hamartomatous polyp 961
Colon, juvenile polyps 961
Colon, metaplastic/hyperplastic polyp 961
Colonic ischaemia 960
Colonic mass 777
Colonic mass, left side 778
Colonoscopy 1022
Colostomy 975
Colostomy, care 976
Colostomy, closure 976
Colostomy, complications 976
Colostomy, decompresseion 975
Colostomy, Devine's double-barrel 975
Colostomy, diversion 975
Colostomy, irrigation 975
Colostomy, loop transverse 975
Colostomy, permanent 975
Colostomy, temporary 975
Combined recurrent/superior laryngeal nerve palsy 529
Compartment excision 94
Compartment syndrome 8
Compartment syndrome, problems 9
Complete rupture urethra, treatment, railroad technique 1130
Compound palmar ganglion 165
Compression bandages, grading 251
Condyloma acuminata 1057
Condyloma lata 28
Congenital heart diseases 1217
Congenital hypertrophy of the retinal pigmented epithelium of iris (CHIRPES) 964

Congenital PUJ obstruction 1084, 1086
Coning 1173, 1177
Conley's point 452
Contracture, Dupuytren's 168
Contracture, hour-glass 882
Contracture, treatment 137
Contracture, Volkmann's ischaemic 168
Contre-coup injury 1173
Cooper's ligament repair 811
Corn 171
Corona phlebectatica 235
Coronary arteries 1215
Coronary balloon angioplasty 1222
Coronary sinus 1215
Correa cycle 903
Costoclavicular compression manoeuvre 458
Couinaud segments 666
Countryman's lip 411
Coup injury1173
Courvoisier's law 694, 749
Craniopharyngiomas 1187
Cresol 1245
Criles operation 480
Criminal nerve of Grassi 870
Crohn's disease 925
Crohn's disease, barium meal follow through 927
Crohn's disease, Vienna classification 926
Crossed leg test 178
Crurum puellarum frigidum 211
Crush syndrome 10
Crush syndrome, effects 10
Crush syndrome, treatment 10
Crutch palsy 280
Cryoprecipitate 125
Cryosurgery 1264
Cryptorchidism 1161, 1166
Csendes procedure 882
CSF rhinorrhoea 1176
CT scan 1240
CT scan, contrast agents 1240
CT scan, high resolution 1242
CTscan, spiral 1240, 1242
Culling 712
Cultured skin 136
Cushing's syndrome 539
Cutaneous neurofibromatosis 91
Cutaneous ureterostomy 977
CVP 116
Cyclical mastalgia 551
Cyclophosphamide 482
Cyclosporine 1229
Cyclosporine A 356
Cyst, acquired 74
Cyst, apoplectic 74
Cyst, chylolymphatic 611
Cyst, classification 74
Cyst, clinical features 74
Cyst, congenital 74
Cyst, degenerative 74
Cyst, dental 385
Cyst, dentigerous 384
Cyst, distention 74
Cyst, effects 75
Cyst, enterogenous 611
Cyst, epidermoid 79
Cyst, epididymis 1155
Cyst, false 74
Cyst, lymph 88
Cyst, mesenteric 780
Cyst, Morrant Baker's 87
Cyst, omental 614, 780
Cyst, parasitic 74
Cyst, periapical 385
Cyst, pseudopseudopancreatic 737
Cyst, radicular 385
Cyst, retention 74
Cyst, retroperitoneal 780
Cyst, sebaceous 79
Cyst, solitary renal 1077
Cyst, splenic, nonparasitic 724
Cyst, splenic, parasitic 724
Cyst, splenic, pseudo 724
Cyst, spring water 1206
Cyst, thyroglossal 487
Cyst, traumatic 74
Cyst, true 74
Cystectomy, partial 1114
Cystectomy, radical 1115
Cystic fibrosis 758
Cystic hygroma 89, 464
Cystic lesion, abdomen 780
Cystic swelling, neck 469
Cystitis 1110
Cystitis, interstitial 1110
Cystitis, recurrent 1110
Cystocele 800
Cystogastrostomy 736
Cystogram 1116
Cystosarcoma phylloides 552
Cystoscopic litholapaxy 1109
Cystoscopy 1071
Cysts 74
Cysts, lung 1204
Cysts, mesenteric 611

D

Dacron patch graft 1218
Danaparoid 252
Danazol 552
Dandy's brain cannula 1182
Darning, Abrahamson nylon 810
Davies and Starr buttock skin flap rotation 1038
Davis T-tube ureterostomy 1088
Day care surgery 1274
Day care surgery, ASA grading of patient 1274
Day care surgery, exclusion criteria 1274
DCIS 565
DCIS, management 566
DCIS, Nottingham prognostic index 566
DCIS, Van Nuy's prognostic index 566
D-dimer test 233
Deceleration injuries 1172
Decoy prostate 1122
Deep veins 230
Defaecography 1029, 1032
Defibrillation 120
Degloving injuries 10
Delayed capillary filling 178
Delayed venous refilling 178
Delphian nodes 485
Demodex folliculorum 79
Denis-Browne procedure 1133
Denonviller's fascia 1020
Dentate line 1021
Dermal flares 235
Dermal histiocytoma 296
Dermatofibroma 296
Dermatofibrosarcoma 326
Dermatofibrosarcoma protuberans 296
Dermoid, external angular 75
Dermoid, implantation 78
Dermoid, internal angular 75
Dermoid, sublingual 390
Dermoid, teratomatous 78
Dermoid, tubulo 78
Dermoids, sequestration 75
Dermoids, sequestration, complication 77
Dermoids, submental 78
Dermoids, types 75
Desmoid tumour 84
Dettol 1245
Devine and Horton's dermal flap 1142
Devine's exclusion procedure 914
Dexon 1256
Dextrans 126
Diabetic foot 213
Diagnostic peritoneal lavage (DPL) 149, 151
Diagnostic peritoneal lavage 714
Dialysis 359
Dialysis, peritoneal 360
Diaphragmatic defence 600
Diaphragmatic hernia, congenital 1211
Diaphragmatic hernia, traumatic 1213
Diastasis recti 796
Diathermy, types 1258
Diethylstilbesterol 585
Dietl's crisis 1084, 1086
Dieulafoy's disease 894
Dilator, Bake's 1251
Dilator, Clutton's 1132
Dilator, filiform 1132
Dilator, Lister 1132
Dilator, Lister's urethral 1254
Dionosil 857
Diosmin 242
Disease, Basedow's 499
Disease, Caisson's 202
Disease, Caroli's 676, 677
Disease, decompression 202
Disease, Dercum's 71
Disease, Goetsch's 499
Disease, Grave's 490, 499
Disease, hide porter's 61
Disease, Marion's 1134
Disease, Menetrier's 901, 903

Disease, Mondor's 546
Disease, Nonne-Milroy's 257
Disease, Ormond's 764
Disease, Parry-Romberg 229
Disease, Peyronie's 1141
Disease, Pick's 1210
Disease, Plummer's 490, 499
Disease, Reclus 568
Disease, Reiter's 1135
Disease, Schimmelbusch's 551
Disease, Sipple 518
Disease, Sipple's 536
Disease, Von Ricklinghausen 532
Disease, Wool sorter's 61
Disinfection 1244
Disinfection, methods 1244
Dislocation, mandible 380
Disobliteration 191
Disseminated intravascular coagulation 128
Distal run off 180, 186
Diverication of recti 796
Diverticular disease, colon 951
Diverticulitis 952
Diverticulitis, Hinchey's classification 953
Diverticulosis 952
D-lactic acidosis 941
DMSA 1070
Dohlman's operation 463
Donor organ retrieval 353
Doodleys lavage 1013
Doppler 179, 1240
Doppler test 240
Dormia basket 692, 693
Drain system, classification 1253
Drain, closed siphon 1253
Drain, closed suction 1253
Drain, corrugated rubber 1252
Drain, indications 1253
Drain, sump 1253
Drain, sump suction 1253
Drain, tube 1252
Drain, under water seal 1253
Drain, wick 1253
Drains 1252
Drains, problems 1253
Dressings 1272
Dressings, components 1272
Dressings, dry 1272
Dressings, wet 1272
Dresslers quadrangle 1146
Driver's bottom 1036
Dry perforation 890
DSA 180
DTPA 1070
Duct ectasia 558
Duct of Luschka 673
Duct papilloma 562
Duct, Bartholin 434
Duct, Santorini 725, 759
Duct, Stensen's 431
Duct, Wharton's 433
Duct, Wirsung 725, 759
Duodenal adenomatosis, Spigelmn classification 936
Duodenal atersia, 'Wind sock' deformity 989
Duodenal blow out 895
Duodenal diverticula 902
Duodenal diverticulum 924
Duodenal injury grading 153
Duodenal ulcer, barium meal X-ray 883
Duodenal ulcer, bleeding 891
Duodenal ulcer, bleeding, Forrest classification 891
Duodenal ulcer, bleeding, treatment 892, 893
Duodenal ulcer, complications 883
Duodenal ulcer, perforated 887
Duodenal ulcer, surgeries 884
Duodenal ulcer, treatment 884
Duodenoduodenostomy 990
Duodenum, anatomy 871
Duoderm 136
Duplex scan 179, 233, 241
Duplex scanning 1240
Duval procedure 745
DVT, prevention 235
Dynamic obstruction 983
Dysejaculation 836
Dyskeratosis 392
Dysmergia 1272
Dysphagia 840
Dysphagia lusoria 841
Dysphagia, causes 840
Dysplasia 283

E

EAST 179
E-cadherin gene 904
ECF excess 96
ECF loss 96
Echinococcus granulosus 637
Echinococcus granulosus, life cycle 637
Echocardiography 1216
Echosclerotherapy 243
Ectopia vesicae 1106
EGFR and VEGF blockers 973
Electric burns 139
Electric burns, complications 140
Electric burns, types 140
Electron microscopy 290
Elephant boot 333, 339
Elephantiasis, endemic 260
Elephantiatic neurofibromatosis 91
Eloiser's method 1200
Embolectomy 198, 200
Emboli lodging sites 197
Emboli, paradoxical 197
Embolisation, therapeutic 201
Embolism 197
Embolism, air 201
Embolism, arterial 197
Embolism, fat 201
Embolism, features 198
Embolism, investigations 198
Embolism, pulmonary 254, 1203
Embolus, saddle 200
Emil Theodor Kocher 526
Emphysema, surgical 1204
Emphysematous cyst, lung 1205
Emprosthotonus 51
Empyema necessitans 1200
Empyema thoracis 1199
Empyema thoracis, decortication 1200
Empyema thoracis, types 1200
Empyema, chronic subdural 1179
EMSU (early morning samples of urine) 1083
Endarterectomy 191
Endoscopic mucosal resection 860
Endoscopic necrosectomy 732
Endoscopic oesophageal staining 842, 860
Endoscopic pyelolysis 1088
Endoscopic retrograde cholangiopancreatography (ERCP) 671
Endoscopic surgery, natural orifice transluminal 1270
Endoscopic ultrasound 696
Endotracheal tube 1233
Endovascular aneurysm repair (EVAR) 206
Endovascular stapler 1266
Enlarged spleen 778
Entamoeba histolytica 960
Entamoeba histolytica, life cycle 631
Enteral feeding 102
Enteral feeding, advantages 103
Enteral feeding, complication 102
Enteral feeding, contraindication 103
Enteric perforation 929
Enteritis, regional 925
Enterocele 800, 806
Enteroclysis 942
Enteroscopy, push 943
Enteroscopy, Sonde 943
Enucleation 94
Ependymomas 1184
Epididymitis 171
Epididymitis, tuberculous 1171
Epididymo-orchitis, acute 1165
Epignathus 93
Epineurorrhaphy 274
Epi-perineurorrhaphy 274
Epiplocele 806
Episodic aphagia 856
Epispadias 1106, 1133
Epitaxy 1090
Epithelioid cells 473
Epithelioma 24, 299
Epstein pearls 367
Epstien-Barr virus 270
Epulis 382
Epulis, carcinomatous 383
Epulis, congenital 382
Epulis, fibrous 382
Epulis, giant cell 383
Epulis, granulomatous 383

Epulis, myelomatous 382
Epulis, pregnancy 382
ERCP, therapeutic 692
Erysipelas 38
Erysipeloid disease 39
Erythema induratum 27
Erythrocyanosis frigida 27
Erythromelalgia/erythralgia 210
Erythroplakia 393
Erythropoietin 127
Eschar 131, 136
Eschar, circumferential 137
Escharotomy 136
Eschmann blade 340
Eskimoma 441
Essen regimen 60
Esthiomene 29
Ether 1232
Ethrombosis 232
Ethylene 1256
Ettapan's lavage 604
Euro-Collins solution 354
Eusol 1245
EUSOL 251
Eusol bath 1245
Evan's formula 135
EVAR 206
Eve's disease 383
Eventration 1211
Exocrine pancreatic disease 763
Exomphalos 792
Exomphalos, major 793
Exomphalos, minor 793
Exophthalmometer 503
Exophthalmos 503
Exophthalmos, grading 504
Exophthalmos, malignant 503
Exophthalmos, pulsating 505
Exophthalmos, severe 503
Exophthalmos, treatment 503
Expansile impulse on coughing 804
External laryngeal nerve 527
External laryngeal nerve, injury 1263
External piles 1046
Extracorporeal shock wave lithotripsy (ESWL) 1092
Extrahepatic biliary tree 670

F

Facial nerve 432
Facial nerve injury 453
Facial nerve palsy, surgeries 454
Facial nerve, identification 452
Faciomaxillary diseases 367
Faciovenous plane of Patey 432
Factor VI and IX concentrate 125
Faecal fistula after appendicectomy 1016
Familial adenoma 936
Familial adenomatosis polyp 963
Familial hypocalciuric hypercalcaemia 533
Fascia of Waldeyer 1020
Fasciotomy 7, 9, 199
Fatty hernia of linea alba 831
Felon 161
Femoral canal, anatomy 822
Femoral canal, pathology 822
Femoral hernia 822
Femoral hernia, AK Henry's approach 824
Femoral hernia, hydrocele 822
Femoral hernia, Laugier's 822
Femoral hernia, Lotheissen's repair 824
Femoral hernia, Narath's 822
Femoral hernia, sliding 823
Femoropopliteal bypass graft 184
FFP 124
Fibrin cuff theory 236
Fibrinogen 125
Fibrinolysins 199, 234
Fibroadenoma 548
Fibroadenoma, complex 549
Fibroadenoma, giant 548
Fibroadenoma, juveline 549
Fibrocystadenosis 548, 551
Fibroepithelial polyp 416
Fibrolipoma 71
Fibroma 84
Fibroma, classification 84
Fibroma, hard 84
Fibroma, soft 84
Fibromyoadenoma 1122
Fibrosarcoma 336
Fibrous dysplasia of bone/jaw 386
Filariasis 259
Filariasis, extrinsic incubation period 259
Filarisis, biological incubation period 259
Finasteride 1124
Fine needle aspiration cytology (FNAC) 289
Fine needle non-aspirating cytology (FNNAC) 290
First heart sound 1215
Fish handler's disease 39
Fissure-in-ano 1046
Fistula 30
Fistula, aortoenteric 205
Fistula, aortocaval 205
Fistula, cimino 227, 360
Fistula, duodenal 895
Fistula, ECK 659
Fistula, enteric/gastrointestinal 943
Fistula, external 30
Fistula, faecal 979
Fistula, gastrojejunocolic 897
Fistula, internal 30
Fistula, parotid 438
Fistula, rectobulbar 1035
Fistula, rectovestibular 1035
Fistula, thyroglossal 488
Fistula, tracheo-oesophageal 857
Fistula, types 30
Fistula, umbilical 788
Fistula, urinary 1120
Fistula, vesicovaginal 1120
Fistula-in-ano 1051
Fistula-in-ano, cryptoglandular hypothesis 1051
Fistula-in-ano, gluing 1055
Fistula-in-ano, high level 1052, 1055
Fistula-in-ano, low level 1052, 1053
Fistula-in-ano, Park's classification 1052
Fistula-in-ano, standard classification 1052
Fistulectomy 1055
Fistulogram 33
Fistulotomy 1055
5-fluorouracil 482
Flap, Abbe 412
Flap, Bakamjian 348
Flap, cross leg 350
Flap, cutaneous 345
Flap, deltopectoral 348
Flap, Estlander's rotating 412
Flap, fasciocutaneous 345
Flap, forehead 348
Flap, free 347
Flap, Fries modified Bernard 412
Flap, gastrocnemius muscle 350
Flap, Gillies fan 412
Flap, groin 349
Flap, island 347
Flap, Karapandzic 412
Flap, latissimus dorsi muscle/myocutaneous 349
Flap, LD 594, 596
Flap, Limberg 350
Flap, local rotation 346
Flap, muscle 346
Flap, myocutaneous 346
Flap, Narayanan's 348
Flap, nasolabial 412
Flap, osteomyocutaneous 346
Flap, pectoralis major myocutaneous 349
Flap, radial forearm 350
Flap, Ruben's 595
Flap, saltatory 347
Flap, TRAM 594, 596
Flap, transposition 346
Flap, transverse rectus abdominis muscle (TRAM flap) 350
Flaps 344
Flaps, axial pattern 345
Flaps, random pattern 344
Flariasis, clinical incubation period 259
Flatulent dyspepsia 683
Flexor retinaculum 156
Flial chest 1196
Flial segment 1196
Fluid therapy 100
Fluids, colloids 101
Fluids, crystalloid 101
Fluorine-18 fluorodeoxyflucose PET scan 1025
Fluosoleda 127
Fluoxymestrone 585
Flutamide 1128
Foaming liver 53
Focused abdominal sonar trauma (FAST) 151
Fogarty catheter 198, 693
Foley's Y-V plasty 1088

Follicular odontome 384
Foot drop 280
Foramen of Langer 545, 546
Forceps, Allis' tissue holding 1248
Forceps, artery 1247
Forceps, Babcock's 1248
Forceps, Cheatle's 1247
Forceps, Desjardin's choledocholithotomy 1251
Forceps, Kocher's 1248
Forceps, Lane's tissue holding 1249
Forceps, Lister's sinus 1252
Forceps, Morant-Baker's appendix holding 1249
Forceps, non-toothed dissecting 1250
Forceps, Rampley's 1247
Forceps, right angle 1248
Forceps, sinus (Lister's) 1252
Forceps, sponge holding 1247
Forceps, toothed dissecting 1250
Fordyce's disease 81
Formaldehyde 1246
Fracture middle third area 375
Fracture nose 373
Fracture ribs 1195
Fracture ribs, non-traumatic 1196
Fracture, depressed skull 1180
Fracture, Guardsman 378
Fracture, Guerin 374
Fracture, mandible 378
Fracture, pyramidal 375
Fracture, zygomatic bone and arch 376
Fracture, zygomatic complex 376
Fractures, Le Fort 374
Fragility test 716
Freckle 308
French disease 56
Frequency 1107
Frey procedure 743, 744
Frog hand 162
Frostbite 23, 215
Frozen chest 1200
Frozen tongue 418
Fruchaud's myopectineal orifice 801
Fuezalida vaccine 59
Functional neck node dissection 513
Fundoplication, Dor anterior 845
Fundoplication, laparoscopic 845
Fundoplication, Lind anterior and posterior 845
Fundoplication, Nissen's 845, 852
Fundoplication, partial with mesh wrap 845
Fundoplication, Rosetti Hell anterior 845
Fundoplication, Toupet's 852
Fundoplication, Toupet's partial 180° posterior 845
Fundoplication, Watson's anterolateral 845
Fundoplications 845
Furuncle 46

G

Gabriel syringe 1043
Gaiter's zone 28, 249
Galactocele 554
Galactorrhoea 560
Galactorrhoea, primary 560
Galactorrhoea, secondary 560
Galezia triad 168
Gallbladder dyskinesia 710
Gallbladder polyp 699
Gallbladder, congenital anomalies 673
Gallbladder, diverticulosis 689
Gallbladder, double/triple 673
Gallbladder, empyema 687
Gallbladder, hydrops 688
Gallbladder, mobile/floating 673
Gallbladder, mucocele 688
Gallbladder, palpable mass 775
Gallbladder, polyp 689
Gallbladder, porcelain 694
Gallbladder, Strawberry 689
Gallbladder, surgical anatomy 669
Gallstone colic 683
Gallstone ileus 703
Gallstones 679
Gallstones, asymptomatic 694
Gallstones, dissolution therapy 690
Gallstones, effects 682
Gallstones, management 684
Gallstones, mixed 679
Gallstones, pigment 679
Gallstones, stones 679
Galveston regime 135
Ganglion 93
Ganglioneuroma 89
Gangrene 177, 211
Gangrene, diabetic 213
Gangrene, dry 177, 212
Gangrene, Fournier's 1148
Gangrene, wet 177, 212
Garrey's stricture 799, 931
Garrod's pads 168
Gas gangrene 52
Gas gangrene, complications 53
Gas gangrene, organism 52
Gas gangrene, polyvalent antiserum 54
Gas gangrene, prevention 54
Gas gangrene, treatment 54
Gas gangrene, types 53
Gastic pH 872
Gastrectomy 919
Gastrectomy, Billroth I 919
Gastrectomy, Billroth II 919
Gastrectomy, D1 912
Gastrectomy, D2 912
Gastrectomy, linear 885
Gastrectomy, lower radical 912, 919
Gastrectomy, Mak's pylorus preserving 881
Gastrectomy, nutritional deficiency 895
Gastrectomy, partial 881
Gastrectomy, total 912
Gastrectomy, upper radical 912, 919
Gastric acid, secretion 871
Gastric acid, function 872
Gastric adenocarcinoma, proximal, Siewart classification 905
Gastric antral vascular ectasia 894
Gastric cancer, advanced 904
Gastric cancer, distal 906
Gastric cancer, early 904
Gastric cancer, familial 902
Gastric cancer, Kajitani classification 905
Gastric cancer, morphovolumetric classification 905
Gastric cancer, Morson and Dawson classification 905
Gastric cancer, proximal 906
Gastric cancer, synchronous 906
Gastric cancer, WHO histological classification 905
Gastric carcinoma 776
Gastric carcinoma, Japanese classification 908
Gastric dilatation, acute 900
Gastric dysplasia 903
Gastric function test 872
Gastric hormones 873
Gastric lymphoma 915
Gastric lymphoma, Ann Arbor staging 916
Gastric lymphoma, Dawson's criteria 916
Gastric lymphoma, primary 915
Gastric lymphoma, secondary 916
Gastric motility 872
Gastric physiology 871
Gastric polyp 901
Gastric sarcomas 916
Gastric surgery, complications 894
Gastric tetany 887
Gastric ulcer 879
Gastric ulcer, bleeding 893
Gastric ulcer, complications 882
Gastric ulcer, pathology 879
Gastric ulcer, perforated 890
Gastric ulcer, types (Daintree Johnson) 880
Gastric varices 654
Gastrin 873
Gastrin releasing polypeptide 873
Gastrin-CCK receptors 872
Gastrinoma triangle 757
Gastrinoma, gastrin assay 757
Gastrinomas 757
Gastritis 877
Gastritis, chronic atrophic 903
Gastritis, erosive 878
Gastritis, granulomatous 878
Gastritis, lymphocytic 878
Gastritis, phlegmonous 878
Gastritis, reflux 878
Gastritis, stress 878
Gastritis, type A 878
Gastritis, type B 878
Gastrocolic reflux 949
Gastroduodenal diseases, investigation 873

Gastrograffin wash 993
Gastrointestinal mesenchymal tumours 916
Gastrointestinal stromal tumours 917
Gastrointestinal stromal sarcoma 918
Gastrojejunostomy of Mayo 886, 920
Gastro-oesophageal reflux disease (GORD/GERD) 840, 842
Gastro-oesophageal reflux disease, endoluminal therapies 844
Gastroschisis 793
Gastroscopy 874
Gastroscopy, therapeutic procedures 874
Gastrostomy 103, 918
Gastrostomy, Janeway's 104
Gastrostomy, Kader-Senn 104
Gastrostomy, percutaneous endoscopic 104
Gastrostomy, Stamm temporary 104
Gate control theory 364
Gelatin 126
Gene therapy 1230
General anaesthesia, complications 1234
General anaesthesia, premedication 1234
Geographic tongue 417
Germ cell therapy 1230
Gherlin 873
GI bleed 1061, 1064
GI bleed, angiography 1064
GI bleed, endoscopy 1064
GI bleed, modified Forrest classification 1061
GI bleed, nuclear scintigraphy 1064
GI bleed, occult 1063
GI bleed, Rockall scoring system 1061
Giant cell arteritis 190
Giant cell reparative granuloma 384
Gianturco-Rosch stent 865
Gibbus 1192
Gieger Muller's gamma ray counter 508
Gimbernat's ligament 822
GIST 917, 918
Glands of Blandin 389
Glasgow coma scale 1175
Glioblastoma multiforme 1184
Gliomas 1183
Glisson's capsule 627
Glomangioma 82
Glossitis 416
Glossitis, chronic superficial 417
Glossitis, Hunter's 417
Glossitis, median rhomboid 416
Glossitis, migrans 417
Glossitis, pellagra 417
Glucagonoma 758
Glutaraldehyde 1246
Glypressin 656
Godwin's tumour 441
Goiter, colloid 491
Goiter, diffuse hyperplastic 490, 491
Goiter, Hokkaido 499
Goiter, multinodular 490, 491
Goiter, plunging 497
Goiter, retrosternal 497
Goitre, classification 490
Goitre, WHO grading 491
Goodsall's operation 1030
Goodsall's rule 1053
Goose's foot 86
GORD, De Meester's scoring system 843
GORD, endoluminal therapy 844
GORD, treatment 844
Goserelin 1128
Gossypiboma 1266
Grading of duodenal injury 153
Graduated elastic compression garments 262
Graft 340
Graft contracture 343
Graft intake, stages 342
Graft rejection 359
Graft, endovascular 216
Graft, full thickness 344
Graft, partial thickness 340
Graft, skin 340
Graft, split thickness 340
Graft, Thiersch 340
Graft, Wolfe 344
Grafts, arterial 192
Granular cell myoblastoma 416
Granulation tissue 17
Granulation tissue, exuberant 19
Granulation tissue, healthy 18
Granulation tissue, unhealthy 19
Granuloma pyogenicum 49
Grawitz tumour 1100
Gray 1225
Great pox 56
Green IVC filter 254
Grynfelt's triangle 835
Gummatous ulcer 28
Gunthur-Tulip filter 254
Gustatory, sweating 452
Guyon's canal 278
Gynaecomastia 560
Gynaecomastia, neonatal 561
Gynaecomastia, pubertal 561
Gynaecomastia, senescent 561
Gynaecomastia, Simon's classification 561
Gynaecomastia, treatment 561

H

H_2 blockers 881, 884
Haemaccel 126
Haemangioma 221, 937
Haemangioma, capillary 222
Haemangioma, cavernous 223
Haemangioma, strawberry 222
Haemangiosarcoma 328
Haematemesis, causes 893
Haematemesis, management 894
Haematocele 1154
Haematoma, acute subdural 1178
Haematoma, chronic subdural 1178
Haematoma, extradural 1173, 1177
Haematoma, extradural, treatment 1178
Haematoma, rectus sheath 794
Haematoma, subdural 1173, 1178
Haematuria 1075
Haematuria, diffuse 1075
Haematuria, early 1075
Haematuria, endemic 1111
Haematuria, false 1075
Haematuria, microscopic 1075
Haematuria, silent 1075
Haematuria, terminal 1075, 1111
Haemobilia 631
Haemodialysis 360
Haemolytic anaemia, alloimmune 717
Haemolytic anaemia, autoimmune 718
Haemolytic anaemia, drug-induced immune 717
Haemolytic anaemia, immune 717
Haemophilia 129
Haemophilia, B 129
Haemophilia, classic 129
Haemorrhage 121
Haemorrhage, acute 121
Haemorrhage, acute on chronic 122
Haemorrhage, arterial 121
Haemorrhage, capillary 121
Haemorrhage, chronic 121
Haemorrhage, classification 121
Haemorrhage, concealed 121
Haemorrhage, gastrointestinal 1061
Haemorrhage, nonsurgical 122
Haemorrhage, pathophysiology 122
Haemorrhage, primary 121
Haemorrhage, reactionary 121
Haemorrhage, revealed 121
Haemorrhage, secondary 121
Haemorrhage, subarachnoid 1179
Haemorrhage, subarachnoid, Fischer grading 1180
Haemorrhage, subarachnoid, Hunt and Hess grading 1180
Haemorrhage, surgical 122
Haemorrhage, treatment 123
Haemorrhage, venous 121
Haemorrhagic shock, classification 122
Haemorrhoidectomy 1045
Haemorrhoidectomy, Hill-Ferguson method 1045
Haemorrhoidectomy, submucosal (Parks) 1045
Haemorrhoids 1038
Haemorrhoids, Barron's banding, 1043
Haemorrhoids, classification 1039
Haemorrhoids, cryo surgery, 1043
Haemorrhoids, Doppler guided artery ligation 1045
Haemorrhoids, infrared coagulation 1044
Haemorrhoids, laser therapy 1044
Haemorrhoids, primary 1039
Haemorrhoids, sclerosant therapy 1042
Haemorrhoids, secondary 1039
Haemorrhoids, treatment 1042
Haemostasis 129
Haemostat 1247

Haemothorax 1198
Haemothorax, thoracotomy indications 1198
Haggitt's invasion of malignant polyp (Colon) 967
Haimovici triad 199
Hairy tongue 417
Halitosis 420
Hallux valgus 173
Halogens 1245
Halsted nodes 547
Hamartoma 229
Hamartomata 229
Hand-bag stomach 882
Hand infections 157
Hand infections, complications 160
Hand infections, general features 157
Hand infections, management 159
Hand infections, types 157
Hand, blood supply 156
Hand, modified Verdan zones 156
Hand, muscles 156
Hand, nerve supply 156
Hand, surgical anatomy 156
Hard chancre 28
Hard corn 171
Hare lip 368
Harris Benedict equation 104
Hartmann's pouch 669
Hashitoxicosis 500
Hashitoxicosis 519
Hassab procedure 658
Heart, anatomy 1215
Heart, conducting system 1216
Heart, surface marking 1216
Heberdon's nodes 170
Heinz bodies 712
Helicobacter pylori 874, 875
Helicobacter pylori, enzymes and toxins 876
Helicobacter pylori, tests 876
Hemicolectomy, left sided 972, 973
Hemicolectomy, right 972, 973
Hemihepatectomy 645, 649, 666
Hemipelvectomy 336
Hemobilia 704
Henderson equation 99
Henderson-Hasselbalch equation 99
Hepar lobatum 28, 649
Heparin 252
Heparin, low molecular weight 252
Hepatectomy, total 645
Hepatic adenoma 642
Hepatic cystadenocarcinoma 651
Hepatic encephalopathy 665
Hepatic encephalopathy, grading 665
Hepatic failure 664
Hepatic failure, fulminant 664
Hepatic resection 665
Hepatic resection, techniques 666
Hepatobiliary cystadenoma 650
Hepatoblastoma 643
Hepatocellular carcinoma 643
Hepatoma 643, 774
Hepatoma Okuda staging system 644
Hepatoma, AJCC staging 644
Hepatoma, CLIP scoring 644
Hepatoma, CUPI scoring 644
Hepatoma, fibrolamellar varient 644
Hepatoma, intra-arterial chemotherapy 646
Hepatoma, neoadjuvant therapy 646
Hepatoma, Okuda classification 644
Hepatoma, palliative treatment 646
Hepatoma, percutaneous ethanol injection 646
Hepatoma, radiofrequency ablation 646, 649
Hepatoma, spread 644
Hepatoma, treatment 645
Hepatorenal syndrome 665
Her2/Neu receptor 577
Hereditary elliptocytosis 717
Hereditary nonpolyposis colonic cancer (HNCC) 964
Hereditary spherocytosis 716
Hermen-Taylor regime 890
Hernia 798
Hernia examination, rules 805
Hernia surgery, complications 836
Hernia through foramen Bochdalek 1212
Hernia through foramen of Morgagni 1213
Hernia, acquired 800
Hernia, Callison-Cloquet 823
Hernia, classification 799
Hernia, Cloquet's 822
Hernia, congenital 800
Hernia, contents 798
Hernia, Cooper's 823
Hernia, coverings 798
Hernia, diaphragmatic 1211
Hernia, different types 837
Hernia, double 821
Hernia, epigastric 831
Hernia, funicular direct 813
Hernia, giant inguinal 813
Hernia, Hesselbach's 823
Hernia, hiatus 847
Hernia, hiatus, classification 847
Hernia, incarcerated 817
Hernia, incisional 825
Hernia, infantile 821
Hernia, inflamed 799
Hernia, inguinal 800
Hernia, irreducible 799
Hernia, Littre's 800, 923
Hernia, lumbar 835
Hernia, Maydl's 818
Hernia, mesenteric defect 1004
Hernia, mesh repair, material used 814
Hernia, mesh repair, types 814
Hernia, mesocolic 1003
Hernia, obstructed 799
Hernia, obturator 834
Hernia, pantaloon 821
Hernia, paraduodenal 1003
Hernia, paraoesophageal 848
Hernia, parastomal 836
Hernia, parastomal, Devlin's classification 836
Hernia, paraumbilical 831
Hernia, parts 798
Hernia, prevesical 813
Hernia, recurrent 813
Hernia, reducible 799
Hernia, re-recurrent 814
Hernia, Richter's 800, 818, 834
Hernia, rolling 848
Hernia, Romberg 821
Hernia, sac 798
Hernia, saddle 821
Hernia, sciatic 836
Hernia, Serofini's 823
Hernia, sliding 820, 847
Hernia, Spigelian 833
Hernia, strangulated 800, 817
Hernia, strangulated omentum 819
Hernia, strangulated, in infancy 819
Hernia, strangulated, treatment 819
Hernia, Swiss cheese 824, 831, 832
Hernia, Teale's 823
Hernia, truss 812
Hernia, umbilical 829
Hernia, umbilical, laparoscopic repair 830
Hernia, umbilical, open dual PTFE and polypropylene mesh placement 830
Hernia, umbilical, sublay mesh repair 830
Hernia, ventral 824
Herniaaetiology 797
Hernia-en-glissade 820
Hernia-in-W 818
Hernias interparietal/interstitial 833
Herniography 800
Hernioplasty 811, 814, 1262
Herniorrhaphy, complications 809
Herniorrhaphy, modified Bassini's 807, 808, 1262
Herniotomy 807, 1261
Herpes labialis 409
Hesselbach's triangle 802, 811
Hexachlorophane 1245
Hiatus hernia, oesophageal 1213
Hibernoma 71
Hibitane 1245
HIDA scan 673
Hidradenitis suppurativa 46
Hidradenitis suppurativa, anal region 1060
Hilar injury, spleen 713
Hip disarticulation 335
Hippocrates facies 605
Hippocratic facies 888
Hirschsprung's disease 949
Histamine 873
Histamine receptors 872
Hitch stitches 1176
HIV infection, clinical classification 63
HIV, CD4 count 63
HIV, GIT problems 64
HIV, hospital and surgeon 65
HIV, modes of transmission 63

HIV, neurological problems 64
HIV, pulmonary problems 64
HIV, tests 63
HIV, treatment 64
HIV, tumours 64
HIV, universal precautions 65
HNPCC 902
HO's triangle 423
Hogarth Pringles's manoeuvre 697
Hollister bell cap technique 1140
Home parenteral nutrition 942
Honey comb liver 57
Honvan 1128
Hormone antagonist 1227
Hormone therapy, cancer 1227
Horse antirabies serum 60
Hose pipe colon 956
Hot-air oven 1244
Hounsfield number 1240
Housemaid's knee 85
Howell-Jolly bodies 712
Human albumin 4.5% 126
Human diploid cell vaccine 59
Human leucocyte antigen 353
Human rabies immunoglobulin 60
Humburg classification 221
Humby's knife 340
Hunger pain 883
Hunter's ligation 203
Hunterian chancre 28, 56
Hutchinson's contagious condyloma 392
Hutchinson's freckle 307
Hutchinson's melanotic freckle 311
Hutchinson's teeth 56
Hutchinson's triad 56
Hutchinsonian pupil 1177
5-hydroxyindoleacetic acid (5HIAA) 939
Hybridoma 1229
Hydatid cyst, complication 1205
Hydatid cyst, liver 637
Hydatid cyst, liver, Hassen Gharbi's ultrasound based classification 638
Hydatid cyst, liver, treatment 639
Hydatid cyst, liver, WHO classification 638
Hydatid cyst, lung 1205
Hydatid cyst, Water-lily appearance 1205
Hydatid disease, malignant 641
Hydatid fluid 638
Hydatid thrill 638
Hydradenoma 306
Hydrocele 1149
Hydrocele fluid 1150
Hydrocele of the neck 464
Hydrocele, bilocular 1151
Hydrocele, canal of Nuck 1151
Hydrocele, complications 1152
Hydrocele, congenital 1151
Hydrocele, encysted 1151
Hydrocele, filarial 1152
Hydrocele, hernia sac 1151
Hydrocele, infantile 1151
Hydrocele, postherniorrhaphy 1151
Hydrocele, primary vaginal 1150
Hydrocele, secondary 1151
Hydrocele, surgeries 1153
Hydrocele, tapping 1261
Hydrocele, treatment 1152
Hydrocele, types of surgery 1261
Hydrocele-en-bisac 1151
Hydrocephalus 1181
Hydrocephalus, classification 1181
Hydrogen peroxide 1246
Hydrohepatosis 695, 702, 749, 773
Hydronephrosis 1083
Hydronephrosis, intermittent 1084
Hydronephrosis, persistent 1084
Hydrophobia 58
Hydrotherapy 141
Hydroxyethystarch 126
Hygroma, chronic subdural 1178
Hypaque 434, 1069
Hypen veins 235
Hyperabduction manoeuvre 458
Hyperbaric oxygen 118
Hyperkalaemia 98
Hypermagnesaemia 98
Hypernatraemia 97
Hypernatraemia, types 97
Hypernephroma 1100
Hyperparathyroidism 532
Hyperparathyroidism, acute 532
Hyperparathyroidism, acute, treatment 533
Hyperparathyroidism, primary 532, 534
Hyperparathyroidism, secondary 532
Hyperparathyroidism, tertiary 532
Hyperparathyroidism, X-ray 533
Hypersplenism 723
Hyperthyroidism 499
Hyperthyroidism, apathetic 500
Hyperthyroidism, postpartum 500
Hyperthyroidism, signs 502
Hyperthyroidism, subclinical 500
Hypertrophic scar 12
Hypocalcemia, spurious 537
Hypocalcemia, temporary 537
Hypokalaemia 97
Hypokalaemic alkalosis 98
Hypomagnesaemia 98
Hyponatraemia 96
Hyponatraemia, hypervolemic 96
Hyponatraemia, normovolemic 96
Hypoparathyroidism 535, 537, 1263
Hypoparathyroidism, permanent 537
Hypoparathyroidism, temporary 537
Hypospadias 1132
Hypothyroidism 526
Hypovolaemia, types 113

I

I^{123} 508
I^{131} 508
Idiopathic gangrene of the scrotum 1148
Ifeanyichukwu classification of burns contracture 137
Ileal conduit 1120
Ileal interposition with sleeve gastrectomy 110
Ileal urinary conduit 977
Ileostomy 959, 977
Ilzarov method 188
Imatinib mesylate 917
Immunohistochemistry 290
Immunosuppression 1229
Immunosuppression, risks related 356
Immunosuppression, types 1229
Immunosuppressive agents 355
Immunotherapy 755, 1229
Immunotherapy, tumours used in 1229
Impetigo 49
In situ hybridisation, fluorescence (FISH) 290
In situ hybridisation 290
Incineration 1244
Incision, Galliard Thomas 550, 552
Incision, Grid-iron 1011, 1262
Incision, Lanz crease 1011
Incision, Mc'Burney 1014, 1262
Incision, Rutherford-Morrison 1011
Incision, Webster's 550
Incisional hernia, laparoscopic mesh repair 827
Incisional hernia, mesh repair 827
Incisional hernia, mesh repair, overlay 827
Incisional hernia, mesh repair, sublay 827
Incisional hernia, mesh repair, underlay 827
Incisional hernia, mesh replacement, Rive's Stoppa's 827, 828
Incomplete rupture urethra, treatment, Blandy approach 1130
Incomplete rupture urethra, treatment, Mitchell approach 1130
Index, ankle-brachial pressure 180
Index, pack year 185
Index, smoking 185
Induratio-penis plastica 1141
Infection, apical subungual 161
Infection, deep palmar space 162
Infection, HIV 62
Infection, hospital acquired 61
Infection, nosocomial 61
Infection, opportunistic 61, 62
Infection, terminal pulp space 161
Infection, web space 162
Infliximab 928, 958
Infrainguinal arterial occlusive disease 184
Infusion IVU 1068
Ingrowing toe nail 172
Inguinal block dissection, bilateral 1146
Inguinal block dissection, modified (Catalona) 1146
Inguinal block dissection, prophylactic 1146
Inguinal block dissection, superficial 1146
Inguinal block dissection, therapeutic 1146
Inguinal canal, anatomy 800
Inguinal canal, contents 800

Inguinal canal, defence mechanism 801
Inguinal hernia repair, anaesthesia 808
Inguinal hernia repair, local anaesthesia 808
Inguinal hernia, Bendavid classification 803
Inguinal hernia, classification 802
Inguinal hernia, direct 803, 805, 811
Inguinal hernia, direct, coverings 812
Inguinal hernia, Gilbert classification 803
Inguinal hernia, Gilbert classification, Robbin's modification 803
Inguinal hernia, indirect 802, 803, 805
Inguinal hernia, indirect, bubonocele 802, 804
Inguinal hernia, indirect, Casten's staging 803
Inguinal hernia, indirect, complete 802, 804
Inguinal hernia, indirect, coverings 804
Inguinal hernia, indirect, funicular 802, 804
Inguinal hernia, indirect, Halverson and McVay classification 803
Inguinal hernia, indirect, Ponka's classification 803
Inguinal hernia, indirect, scrotal 804
Inguinal hernia, NYHUS classification 803
Inguinal hernia, precipitating factors 804
Inguinal ligament 800
Inguinal ring, deep 800
Inguinal ring, superficial 800
Inguinodynia 811, 814, 836
Injuries, blast 147
Injuries, bullet 146
Injuries, chest 1194
Injuries, hand 166
Injuries, hand, classification 166
Injuries, hand, treatment 166
Injuries, head 1172
Injuries, kidney 1080
Injuries, kidney, grading 1080
Injuries, maxillary 374
Injuries, maxillofacial 372
Injuries, neck 146
Injuries, penetrating 148
Injury, axillary nerve 281
Injury, bladder 1115
Injury, brachial plexus 275
Injury, colon 154
Injury, common peroneal nerve 280
Injury, duodenal 152
Injury, duodenal, grading 153
Injury, lacrimal apparatus 373
Injury, liver 154
Injury, long thoracic nerve 281
Injury, medial poplitial nerve 281
Injury, median nerve 276
Injury, pancreatic 153
Injury, parotid duct 373
Injury, radial nerve 279
Injury, renal 154
Injury, reperfusion 199
Injury, seatbelt 155
Injury, small bowel 153
Injury, spinal 145
Injury, splenic 154, 712
Injury, steering wheel 1194
Injury, ulnar nerve 278
Injury, urethral 1129
Injury, urinary bladder 154
Insulinomas 755, 756
Insulinomas, radioimmunoassay 756
Insulinomas, treatment 757
'In transit nodules' 284, 312
Intercostal tube drainage 1202
Interleukin 2 receptor 356
Intermittent claudication 175
Intermittent claudication, Boyd's classification 176
Internal hernias 1003
Internal mammary nodes 547
Internal piles, primary 1021
Internist tumour 1100
Intervertebral disc prolapse 1189
Intestinal amoebiasis, surgical complications 960
Intestinal atresia, plain X-ray 991
Intestinal metaplasia 903
Intestinal obstruction, adynamic 1000
Intestinal obstruction, barium enema (microbar solution) 988
Intestinal obstruction, complications 988
Intestinal obstruction, CT scan 988
Intestinal obstruction, gastrograffin contrast enema 988
Intestinal obstruction, neonatal 994
Intestinal obstruction, pathogenesis 984
Intestinal obstruction, plain X-ray abdomen 985, 987
Intestinal obstruction, treatment 988
Intestinal obstruction, types 983
Intestinal obstruction, ultrasound 988
Intestinal tuberculosis, primary 615
Intestinal tuberculosis, secondary 616
Intra-abdominal pressure grading (Busch) 155
Intracranial abscess, acute 1182
Intracranial abscess, chronic 1182
Intracranial abscess, metastatic 1182
Intracranial abscess, subacute 1182
Intraductal ultrasound 696
Intravenous urogram 1067
Intrinsic factor 872
Intrinsic minus deformity 228
Intrinsic plus deformity 228
Intussusception 780, 994
Intussusception, Cope's method 997
Intussusception, retrograde jejunogastric 920
Intussusceptum 995
Intussuscipiens 995
'Inverted ink bottle' effect 1151
Iodophors 1245
Iohexol 1240
Iopamiro 1240
Iopanoic acid 670
Ischaemia, features 178
Ischaemia, upper limb 216
Ischaemia, upper limb, bypass surgeries 218
Ischaemia, upper limb, investigations 217
Ischaemia, upper limb, management 217
Ischaemic colitis, Marston's classification 959
Ischaemic heart disease 1222
Isograft 340
Isotope lymphoscintigraphy 257
Isotope renography 1070, 1087
IVC filter 254
IVU, Adder head appearance 1068
IVU, flower vase appearance 1068
IVU, horseshoe kidney 1076
IVU, hydronephrosis 1087
IVU, spider leg appearance 1068
IVU, ureterocele 1080
IVU-high dose 1081
IVU-polycystic kidney 1077

J

Jack stone 1107
Jack-knife position 1037
Jaffe's tumour 384
Jarisch-Herxheimer reaction 56
Jaundice 749
Jaundice, intermittent 749
Jaundice, obstructive 695, 749, 752
Jaundice, obstructive, classification 696
Jaundice, obstructive, pruritus 697
Jaundice, surgical 695
Jeep bottom 1036
Jejunoileal bypass 108
Jejunoileal diverticula 924
Jejunostomy 104
Jejunostomy, needle 104
Jejunostomy, Witzel 104
Johansen 'stepladder' procedure 412

K

K ras gene 748
Kaposis's sarcoma 328
Kaposis's sarcoma, African 328
Kaposis's sarcoma, AIDS associated 328
Kaposis's sarcoma, Europian 328
Kaposis's sarcoma, transplant associated 328
Kasai porto enterostomy 678
Keetly-Torrek procedure 1163
Keller's operation 173
Kelling Madlener procedure 882
Kelly's point 1035
Keloid 11
Keratin pearls 300, 401
Keratoacanthoma 296
Keratoconjunctivitis sicca 439
Kernohan's notch 1173
Kernohan's notch effect 1177
Ketamine 1232

Khaini chewers 410
Kidney shadow 1066
Kidney, anatomy 1065
Kidney, approaches 1104
Kidney, avascular plane (Brodel) 1066
Kidney, benign tumours 1098
Kidney, capsule 1066
Kidney, caseous/putty 1082
Kidney, cystic diseases 1076
Kidney, horseshoe 1076
Kidney, palpable mass 779
Kidney, polycystic disease 1077
Kidney, structure 1066
Killian's dehiscence 462, 838
Kirschner palliative gastric bypass 865
Kiss cancer 284
Kistner's valvuloplasty 250
Kit protein 917
Klatskin tumour 702
Koch's ileal conduit 1083
Kocher's vein 485
Koontz orchidectomy 813
Kous Netzoff aluminium needle 631
Krukenberg secondaries 572
Krukenberg's tumour 285, 907
Kuntz nerve 194
Kussmaul's breathing 100
Kuttner tumour 436
Kyphosis 1192

L

Ladd's band 992, 1211
Ladd's operation 990, 992
Ladd and Gross procedure 1163
Lahey and Cattell's relaxing skin incision 1038
Langerhan's cell 293
Laparoscopic adjustable gastric banding 108
Laparoscopic APR 1026
Laparoscopic cholecystectomy 1268
Laparoscopic posterior mesh rectopexy (LPMR) 1033
Laparoscopic sigmoid resection and resctopexy 1033
Laparoscopic surgery, advantages 1267
Laparoscopic surgery, complications 1268
Laparoscopic surgery, history 1267
Laparoscopic surgery, instruments 1267
Laparoscopic surgery, pneumoperitoneum 1268
Laparoscopic TAPP repair 1269
Laparoscopic, advanced surgeries 1269
Laparoscopic, appendicectomy 1269
Laparoscopic, TEP repair 1269
Laparoscopy, diagnostic 1270
Laparostomy 604
Large bowel, preparation for surgery 980
Large intestine, anatomy 948
Large intestine, lymphatic drainage 949
Laryngeal tumours, glottic 428
Laryngeal tumours, subglottic 428
Laryngeal tumours, supraglottic 428
Laryngeal cancer, radiotherapy 428, 429
Laryngeal mask airway 1233
Laryngeal nerve, recurrent 485, 523, 528
Laryngeal nerve, superior 485
Laryngeal surgery, conservative 430
Laryngocele 463
Laryngocele, external 464
Laryngocele, internal 464
Larynx, muscles 528
Lasers in surgery 1265
Latent period of Bandet 713
Leflunomide 356
Left to right shunt 1217
Leiomyosarcoma 327
Leprosy 55
Leprosy, multibacillary 55
Leprosy, paucibacillary 55
Leprosy, surgical complications 55
Leprosy, types 55
Lesgaft's triangle 835
Letrozole 585
Leugart's band 839
Leugart's pouch 839
Leukoplakia 392
Leukoplakia, hairy 392
Leuprolide 1128
Lid lag 502, 503, 504
Lid retraction 502, 503, 504
Lid spasm 503
Ligament of Cooper 546
Ligament of Treitz 1061
Ligation and excision of piles (Milligan-Morgan) 1045
Ligation of intersphincteric fistula track (LIFT) 1055
Ligation, Hunter's, 229
Ligation, quadruple 228, 227
Lignocaine 1235
Lily white appendix 1014
Limb ischaemia 176
Limb ischaemia, critical 177, 185
Limb ischaemia, Fontaine classification 176
Limb ischaemia, functional 177
Limb ischaemia, Rutherford classification 176
Limb ischemia, acute, classification 198
Limey gallbladder 681
Line of demarcation 177, 178
Lingual colic 436
Linitis plastica 908
Lintons tube 657
Lip, mucous cyst 408
Lip, verrucous carcinoma 410
Lipiodol 33, 434
Lipodermatosclerosis 236, 238, 250
Lipoedema 260
Lipoma 71
Lipoma arborigens 71
Liposarcoma 74, 325, 766
Lithogenic bile 680, 694
Littoral cell angioma 724
Livedo reticularis 210
Liver 627
Liver abscess, pyogenic 636
Liver abscess, pyogenic, open drainage 637
Liver abscess, pyogenic, percutaneous drainage 637
Liver biopsy 629, 645
Liver cyst 649
Liver cyst, congenital 649
Liver cyst, neoplastic 650
Liver cyst, simple 649
Liver cyst, traumatic 651
Liver function tests 628
Liver injury 629
Liver injury, CT grading 630
Liver injury, management 630
Liver transplantation, orthotopic 645, 661
Liver tumours 642
Liver tumours, benign 642
Liver, actinomycosis 642
Liver, amoebic abscess 631
Liver, blood supply 627
Liver, focal nodular hyperplasia 643
Liver, functions 628
Liver, haemangioma 642
Liver, honeycomb 642
Liver, massive enlargement 649
Liver, palpable left lobe 776
Liver, palpable mass 773
Liver, primary malignant tumours 643
Liver, secondaries 647
Liver, secondaries, classification 647
Liver, secondaries, treatment 649
Liver, segmental anatomy 628
Liver, space occupying lesions 668
Liver, surgical anatomy 627
Local haemostatic agents 123
Local recurrent rectal cancer 1027
Lockhart's tunica vaginalis flap 1142
Lord's anal dilatation 1042, 1048
Lord's plication 1153
Lowenstein-Jensen media 474, 619, 626
Lower GI bleed 1061
Lower GI bleed, classification 1063
Lower oesophageal sphincter 839
Lucid interval 1177
Ludwig's angina 38, 465
Lugol's iodine 506
Lung cyst 1204
Lung cyst, epithelial 1205
Lupus vulgaris 26
Lye stricture 854
Lympedim 262
Lymph node of '*Lund*' 670
Lymph node, microanatomy 256
Lymph nodes, deep cervical 455
Lymph nodes, facial 455
Lymph nodes, submandibular 455
Lymph nodes, submental 455
Lymph nodes, superficial cervical 455
Lymphangiography 257
Lymphangioma 88

Lymphangioma ab igne 464
Lymphangioma circumscripta 464
Lymphangioma diffusum 88, 464
Lymphangioma, capillary 88
Lymphangioma, cavernous 89, 464
Lymphangioma, circumscriptum 88
Lymphangiosarcoma 262, 328, 583
Lymphangitis 39
Lymphatic watersheds of skin 255
Lymphatics 255
Lymphatics of head and neck, anatomy 455
Lymphatics, function 256
Lymphoedema ab igne 88
Lymphoedema, Browse classification 257
Lymphoedema, Brunner's grading 260
Lymphoedema, complex decongestive therapy 262
Lymphoedema, lymphangiographic classification 257
Lymphoedema 257
Lymphoedema, filarial 259
Lymphoedema, Kinmonth classification 257
Lymphoedema, pathophysiology 258
Lymphoedema, primary 257
Lymphoedema, secondary 257, 259
Lymphoma, Ann-Arbor staging 267
Lymphoma, Burkitt's 270
Lymphoma, cutaneous T cell 270
Lymphoma, Hodgkin's 265
Lymphoma, Hodgkin's, treatment 268
Lymphoma, malignant of Africa 270
Lymphoma, MALT 270
Lymphoma, non-Hodgkin's 269
Lymphoma, non-Hodgkin's, Rappaport classification 269
Lymphoma, non-Hodgkin's, treatment 269
Lymphoma, non-Hodgkin's, working classification 269
Lymphoma, REAL classification 265
Lymphoma, Rye's classification 265
Lymphoma, staging laparotomy 267
Lymphomas 264
Lysol 1245
Lytle's repair 810

M

MACIS scoring system 512
Mackler's triad 857
Macroamylasaemia 726
Macrocheilia 408
Macroglossia 416
Madura foot 57
Madura hand 58
Madurella mycetomi 57
Mafenide acetate 136
Magnetic resonance cholangiopancreatography (MRCP) 673, 727, 1242
Magnetic resonance imaging 1242
Magpi repair 1133
Major histocompatibility complex 353
Major papilla of Vater 759
Malabsorptive bariatric surgery 107
Malaena 891
Malakoplakia 1116
Malgaigne bulging 812, 817
Malignant fibrous histiocytoma 326
Malignant histiocytoma 326
Malignant hyperthermia 1234
Malignant laryngeal tumours 427
Malignant laryngeal tumours, staging 428
Malignant mesothelioma 613
Malignant oedema 52
Malignant peripheral nerve sheath tumour 328
Malignant pustule 61
Mallampati scoring 1231
Mallet finger 170
Malrotation 991
MALTOMA 915
Mammary dysplasia 551
Mammary fistula of Atkins 563
Mammography 547, 574
Mammography, bilateral 590
Mammography, diagnostic 548
Mammography, screening 548
Mandibular resections 406
Mandibulectomy, marginal 406
Mannheim peritonitis index 606, 890
Manoeuvre, costoclavicular compression 179
Manoeuvre, hyperadduction 179
Manoeuvre, scalene 179
Marjolin's ulcer 300, 302
Maryfield introducer 1073
Mass abdomen, differential diagnosis 771
Mass abdomen, investigations 782
Mass, epigastrium 775
Mass, hypogastrium 781
Mass, left hypochondrium 778
Mass, left iliac fossa 781
Mass, lumbar region 778
Mass, right hypochondrium 773
Mass, right iliac fossa 780
Mass, umbilical region 780
Massive bowel resection 941
MAST 116
Mastalgia 553
Mastalgia, cyclical 551, 553
Mastalgia, noncyclical 553
Mastectomy, Auchincloss modified radical 581
Mastectomy, complications 583
Mastectomy, extended radical 583
Mastectomy, Halsted radical 582
Mastectomy, prophylactic 593
Mastectomy, simple 480
Mastectomy, skin sparing 583, 588
Mastectomy, subcutaneous 552
Mastectomy, toilet 583
Mastectomy, total 580
Mastectomy, total with axillary clearance 580
Mastitis 555
Mastitis carcinomatosis 566
Mastitis, intramammary 556
Mastitis, periductal 558
Mastitis, plasma cell 558
Mastitis, retromammary 557
Mastitis, subareolar 555
Matas endoaneurysmorrhaphy 203
Maxillary tumours 424
Maxillary tumours, behaviour 424
Maxillary tumours, Lederman's classification 425
Maxillary tumours, Ohngren's classification 424
Maxillary tumours, TNM staging 425
Maxillofacial injuries, classification 372
Maxillofacial injuries, Le Fort's classification 374
Maxillofacial injuries, primary care 374
Maxon 1256
Mayo's procedure 173
McFarlane staging 539
McBurney's point 1008
McBride procedure 173
McCune-Albright syndrome 92
MCR foot wear 212
Mebeverine HCl 958
Mechanical valves 1222
Meckel's diverticulectomy 924
Meckel's diverticulum 786, 923
Meconium ileus 992
Meconium ileus, complication 993
Median nerve palsy 276
Mediastinal flutter 1196
Mediastinal tumours 1205
Medusa head 61
Megacolon, acquired 951
Megacolon, congenital 949
Megaloglossia 416
Meggitt's classification 213
Meglumine ioglycamate 671
Meglumine iothalamate 1067, 1240
Meissner's plexus 870
Melanoblast 308
Melanocyte 308
Melanoma 308
Melanoma, acral lentiginous 312
Melanoma, AJCC classification 309
Melanoma, amelanotic 312
Melanoma, biochemotherapy 316
Melanoma, Breslow's classification 309
Melanoma, Clark's concept 312
Melanoma, Clark's level 309, 310
Melanoma, clinical types 310
Melanoma, desmoplastic 312
Melanoma, endolymphatic therapy 316
Melanoma, Glasgow criteria 312
Melanoma, immunotherapy 316
Melanoma, isolated limb infusion 316
Melanoma, isolated limb perfusion 316
Melanoma, lentigo maligna 311
Melanoma, newer TNM staging 313
Melanoma, nodular 310
Melanoma, occult 313
Melanoma, ocular 312

Melanoma, risk factors 309
Melanoma, satellite nodules 312
Melanoma, sentinal lymph node biopsy 315
Melanoma, staging 313
Melanoma, subungual 312
Melanoma, superficial spreading 310
Melanoma, treatment 314
Melanoma, tumour markers 313
Melanophages 308
Melanophores 308
Melanosis 962
Melanuria 313
MELD score 357
Meleney's infection 65
Melon seed bodies 165
Melphalan 316
MEN syndrome 536
MEN syndrome type 71, 91
Menghini needle 629
Meningiomas 1184
Meningocoele 1188, 1189
Meningomyelocoele 1188, 1189
Meralgia paraesthetica 281
Mercurochrome 343
Mesalamine 957
Mesenteric desmoid tumour 612
Mesenteric lymphoma 612
Mesenteric malignancy 612
Mesenteric tear 152
Mesenteric trauma 613
Mesenteric vessel ischaemia 931
Mesenteric vessel ischaemia, treatment 933
Mesh repair, Gilbert 815
Mesh repair, Gilbert's PHS repair 815
Mesh repair, inlay 814
Mesh repair, Lichtenstein tension free 814
Mesh repair, modified Rives preperitoneal 814
Mesh repair, Nyhus preperitoneal 814
Mesh repair, onlay 814
Mesh repair, totally extraperitoneal laparoscopic (TEP) 815, 817
Mesh repair, transabdominal preperitoneal laparoscopic (TAPP) 813, 815
Mesher 343
Metastatic emphysema of Read 798
Methimazole 506
Methotrexate 482
Michaelis Gutmann bodies 1116
Mickulicz Miles perineal transanal rectosigmoidectomy 1034
Microbar solution 874
Microdochectomy 562
Microfilaria 259
Micturating cystourethrography 1069, 1134
Mikulicz disease 440
Military antishock trouser 116
Military attitude 1192
Milker' nodes 166
Miller procedure 264
Mill-Wheel murmur 201
Milnes-Walker thoracic oesophageal transection 658
Milwaukee's classification 710
Miner's beat knee 85
Miner's elbow 85
Minimal incision aortic surgery (MIAS) 206
Minnesota tube 657
Minute IVU 1068
Mitotane 586
Mitral commissurotomy, open 1221
Mitral regurgitation 1221
Mitral stenosis 1220
Mitral volvotomy, closed 1220
Modified Brooke formula 135
Modified radical neck dissection 512
MODS 45, 46, 112, 114, 117, 132, 888
MOHS 306
Mole, hairy 307
Mole, nonhairy 307
Molluscum sebaceum 296
Mongolian's spot 308
Monilial thrush 391
Monks localising zones 153
Monoclonal antibodies 1230
Monoclonal antibody 356
Monocryl 1256
Montgomery tubercles 545
Morison's aphorism 37
Moth eaten alopecia 28
Mount Vernon box frame 376
Mousseau-Barbin tube 865
Moynihan's aphorism 680, 694
Moynihan's caterpillar hump 673
Mucocele of frontal sinus 75
Mucus glands of Nuhn 389
Muir and Burclay regime 135
Muir low anterior resection 1034
Mulberry stone 1107
Muller's muscle 502
Multiple cylindroma 306
Murphy's triad 876
Murugassu's technique 1035
Muscarinic receptors 872
Muscle relaxant, depolarising 1232
Muscle relaxant, nondepolarising 1232
Myasthenia gravis 1207
Mycetoma pedis 57
Mycobacterium tuberculosis 54
Mycophenolate mofetil 356, 1229
Mycotic aneurysm 203
Myelinic neuroma 89
Myelocele 1188
Myotomy 954
Myotomy, cricopharyngeal 463
Myotomy, Heller's 954
Myotomy, Ramstedt's 954
Myotomy, Reilly's 954
Myxoedema coma 527
Myxoedema, pretibial 504

N

Naevi 306
Naevolipoma 71
Naevus flammeus 222
Naevus of Ota 307
Naevus sebaceous of Jadassohn 294
Naevus spilus 307
Naevus, blue 307
Naevus, compound 307
Naevus, dysplastic 307
Naevus, giant 308
Naevus, halo 307
Naevus, intradermal 307
Naevus, Ito 307
Naevus, junctional 307
Naevus, juvenile 307
Naevus, spindle cell 307
Naevus, Spitz 307
Nagler reaction 54
Nail, floating 160
Nail, hang 160
Napkin ring stricture 617
Narbona's ligamentum teres cardiopexy 846
Narcotic analgesics 366
Nasojejunal feeding 1266
Nasopharyngeal carcinoma 423
Nasopharyngeal carcinoma, treatment 423
Neck dissection, bilateral 481
Neck dissection, classic radical 480
Neck dissection, extended radical 481
Neck dissection, modified radical 480
Neck dissection, supraomohyoid block 481
Neck dissection, types 480, 482
Neck lymph nodes, memorial Sloan-Kettering cancer centre levels 475
Neck lymph nodes, secondaries 475
Neck lymph nodes, secondaries with occult primary 477, 479
Neck lymph nodes, secondaries, blind biopsy 477
Neck lymph nodes, secondaries, features 476
Neck lymph nodes, secondaries, FNAC 478
Neck lymph nodes, secondaries, staging 478
Neck lymph nodes, secondaries, treatment 479
Neck lymph nodes, secondaries, types 477
Necrosis 177
Necrotising enterocolitis 934
Necrotising fasciitis 65, 67
Needle holder 1251
Needle, Hagedorn 1250
Needles, atraumatic 1250
Needles, cutting 1251
Needles, round body 1251
Needles, surgical 1250
Needles, traumatic 1250

Negri bodies 58
Neoadjuvant chemoradiotherapy 755
Neoplasia, etiological factors 284
Neoplasm 282
Neoplasm, benign 282
Neoplasm, malignant 282
Neoplasm, malignant, locally 282
Neoplasm, malignant, locoregionally 282
Neoplasm, malignant, systemic 282
Neostigmine 1232
Nephrectomy 1100
Nephrectomy, partial 1094
Nephrectomy, radical 1103
Nephrectomy, subcapsular 1088
Nephroblastoma 1099
Nephrolithotomy 1092
Nephron sparing surgery 1103
Nephropyelolithotomy 1092, 1098
Nephrostomy 1074
Nephrostomy, bilateral 1088
Nephrostomy, Cabot's 1074, 1088, 1098
Nephrostomy, percutaneous 1074
Nephrostomy, Tresidder's loop 1074
Nephroureterectomy 1076, 1083
Nerve of Latarjet 870
Nerve suturing 274
Neurilemmoma 92
Neuroblastoma 89, 540
Neuroblastoma, Hutchinson's type 540
Neuroblastoma, Pepper type 540
Neuroblastoma, *Shimada* histological classification 541
Neuroblastoma, staging 541
Neurofibroma 90
Neurofibroma, complications 92
Neurofibroma, nodular 91
Neurofibroma, plexiform 91, 92
Neurofibroma, vagus nerve 468
Neurolipoma 71
Neuroma 274
Neuroma, end 89
Neuroma, false 89
Neuroma, lateral 89
Neuroma, true 89
Neuroma, types 89
Neuropraxia 273
Neurotmesis 273
Niche 873, 880
Nitrous oxide 1232
NNIS system 68
'No re-flow' phenomenon 199
No scalpel vasectomy 1260
Nocardia asteroids 57
Nocardia madurae 57
Noma 391
No-man's land 157
Nonchromaffin paraganglioma 468
Non-heart beating donors 354
Non-narcotic drugs 366
Non-nucleoside reverse transcriptase inhibitor (NNRTI) 64
Nonocclusive mesenteric ischaemia (NOMI) 932
Nonulcer dyspepsia 878
Normal rotation of midgut, errors 992
Normal rotation of midgut, stages 991
Notch 873, 880
Nucleoside reverse transcriptase inhibitor (NRTI) 64
Nutrition 101
Nutritional support 102
Nyhus iliopubic repair 811

■ O

Obesity 106
Obesity, causes 106
Obesity, medical management 107
Obesity, morbid 106
Obesity, morbid, problems 106
Obstipation 998
Obstruction, proximal small bowel 983
Obturator 1022
Ochsner-Sherren regimen 1015
Ochsner-Sherren regimen, contraindication 1015
Ochsner-Sherren regimen, criteria 1015
Octreotide 656, 731
Odontomes 382
Odynophagia 840
Oesophageal carcinoma, endoscopic bipolar diathermy 865
Oesophageal carcinoma, endoscopic laser 865
Oesophageal carcinoma, endoscopic photodynamic therapy 865
Oesophageal carcinoma, endoscopic therapy 865
Oesophageal clearance mechanism 840
Oesophageal cyst 867
Oesophageal dilatation 855
Oesophageal diverticulum 857
Oesophageal diverticulum, pulsion type 858
Oesophageal diverticulum, traction type 858
Oesophageal endosonography 842, 859
Oesophageal manometry 844
Oesophageal motility disorders 849
Oesophageal perforation 867
Oesophageal substitutes 864
Oesophageal varices 654
Oesophageal varices, balloon tamponade 657
Oesophageal varices, definitive management 659
Oesophageal varices, emergency management 656
Oesophageal varices, endoscopic gluing 657
Oesophageal varices, endoscopic therapy 656
Oesophageal varices, endoscopic variceal banding 657
Oesophageal varices, endoscopic variceal sclerotherapy 657
Oesophageal varices, endoscopic variceal thrombin injection 657
Oesophageal varices, grading 654
Oesophageal varices, pharmacotherapy 656
Oesophageal varices, treatment 654
Oesophagectomy, complications 866
Oesophagectomy, McKeown three phased 863
Oesophagectomy, Orringer approach 863
Oesophagectomy, radical 860, 863
Oesophagectomy, vagal sparing 864
Oesophagogastrectomy 846
Oesophagoscope, fibreoptic flexible 842
Oesophagoscope, rigid 842
Oesophagoscopy 841, 842
Oesophagus, 'bird beak' appearance 842
Oesophagus, 'corkscrew' appearance 842
Oesophagus, abdominal 838
Oesophagus, anatomical narrowing 838
Oesophagus, anatomical specialities 838
Oesophagus, anatomy 838
Oesophagus, arterial supply 838
Oesophagus, Barrett's 849
Oesophagus, benign tumours 866
Oesophagus, candidiasis 840
Oesophagus, cervical 838
Oesophagus, contrast study 841
Oesophagus, corkscrew 850
Oesophagus, corrosive stricture 854
Oesophagus, corrosive stricture, degree of burns 854
Oesophagus, corrosive stricture, treatment 854
Oesophagus, diffuse spasm 840, 849
Oesophagus, foreign body 840, 867
Oesophagus, leiomyoma 866
Oesophagus, lymphatic drainage 839
Oesophagus, nutcracker 850
Oesophagus, sigmoid 842
Oesophagus, thoracic 838
Oesophagus, venous drainage 838
Oil of evening primrose 552
OK 432 465
Okuda classification 774
Okuda needle 672
Omental torsion 614
Omental tumour 614
Omentocele 800, 806
Omentoplasty 188, 195
Omentoplasty, complications 196
Omphalitis 784
Omphalocele 792
Oncocytoma 445, 1098
Onghren's line 303
Onychocryptosis 172
Onychogryphosis 172
Onychomycosis 172
Operation Koontz' 810
Operation, Anderson- Hyne's 1088
Operation, Andrew's 811
Operation, Belsey Mark IV 846
Operation, Bishop-Koop 759, 993
Operation, Blalock-Taussig 1220
Operation, BOCCA 480

Operation, Boerema 846
Operation, Boerema-Crile 658
Operation, Brunschwig's 1028
Operation, Cattell's 827
Operation, Charle's 263, 264
Operation, child's 743
Operation, Commando 481
Operation, Delorme's 1033, 1034
Operation, Fitzpatric 1142
Operation, Fowler's 172, 279
Operation, Gelhard's 1142
Operation, Goligher's 1033
Operation, Hadfield-Adair 558
Operation, Hamilton-Bailey 810
Operation, Hamilton-Stewart 1084
Operation, Hartley-Dunhill 493, 521
Operation, Hartmann's 989, 998, 1028
Operation, Hill's 846
Operation, Homan's 263
Operation, in Dumphy 1034
Operation, Ivor Lewis 863
Operation, Jabouley's 1153, 1261
Operation, Jurasz 736
Operation, Keel 827
Operation, Kirwin's 1078
Operation, Kondolean's 264
Operation, Ladd's 991
Operation, Lahaut's 1033
Operation, Lily's 677
Operation, Lockwood low 824
Operation, Lotheissen's 824
Operation, Macey's 264
Operation, Max-Page 169
Operation, May-Husni 234
Operation, Mayo's 830, 831
Operation, McEvedy-high 824
Operation, McVay 811
Operation, Michaelis plank 806
Operation, modified Duhamel 951
Operation, modified Heller's 852
Operation, Newman-Seabrock's 439
Operation, Nielubowitz 263
Operation, Nuttall's 828
Operation, O'Brien's 263
Operation, Ombredanne's 1163
Operation, Palma 234
Operation, Palomo's 1159
Operation, Partington-Rochelle 743, 744
Operation, Patey's 581
Operation, Paul Brand's 279
Operation, Paul-Mickulicz 998
Operation, Puestow's 743
Operation, Ramstedt's 877
Operation, Riordan 279
Operation, Ripstein 1033
Operation, Rose Graham 1033
Operation, Roscoe-Graham 889
Operation, Rovsing 1077
Operation, Santulli 993
Operation, Scanlon's 581
Operation, Sever's 275
Operation, Sir Piersey Gold 1145
Operation, Sistrunk 263, 488, 489
Operation, Smith 736
Operation, Stye-Bunnell's 279
Operation, Sugiura-Futagawa 658
Operation, Tanner slide 810
Operation, Tanner's 901
Operation, Taylor's 885
Operation, Thompson's 263
Operation, Torkildsen 1181, 1187
Operation, Trendelenburg 244
Operation, Wardill-Kilner push back 372
Operation, Well's 1033
Operation, Wheelhouse's 1132
Operation, Wyatt 1034
Operation, Young's 1145
Operation, Zadik's 172
Operative surgery 1244
Opisthotonus 51
OPSI 722
OPSI, prevention 722
OPSI, treatment 722
Opsite 136
Opsomyoclonus 541
Optease filter 254
Oral and upper aerodigestive cancers 394
Oral carcinoma, prognostic factors 414
Oral cavity, premalignant condition 394
Oral cavity, syphilitic lesions 392
Oral cholecystogram 670
Oral submucosal fibrosis 393
Orbital cellulitis 37
Orchidectomy 1163
Orchidopexy 1163
Orchitis 1170
Orchitis, bilateral interstitial 1170
Orchitis, syphilitic 1170
Orf 166
Organ of Geraldes 1155
Orphan Annie eye nucleus 511
Orthopantomogram 367, 401
Orthotonus 51
Orthoses 339
Osler Weber Rendu disease 937
Osteitis fibrosa cystica 532
Ovarian mass 781
Oximetry 1216
Oxygen 1232
Oxygen therapy 118
Oxyphil adenoma 445

P

PABA excretion in urine 763
Pacemaker cell of Cajal 872
Pachydermatocele 91
Pachyglossia 416
Pack year index 185
Packed cells 124
Paget's disease of nipple 567
Paget's disease penis 1142
Paget's, recurrent fibroid of 84
Pain 364
Pain pathway 364
Pain, clinical assessment 364
Pain, deep 365
Pain, grading 365
Pain, periodicity 365
Pain, psychogenic 365
Pain, radiation 365
Pain, referred 365
Pain, segmental 365
Pain, shifting/migration 365
Pain, superficial 365
PAIR (Puncture-Aspiration-Injection-Reaspiration) 638, 640
Palate, diseases 367
Palatoplasty, principles 372
Palcitaxel 586
Palmar aponeurosis 156
Panarteritis 185
Pancoast tumours 1208
Pancreas pseudocyst 732, 734
Pancreas pseudocyst, complication 737
Pancreas pseudocyst, D'Egidio classification 735
Pancreas, accessory 759
Pancreas, acute pseudocyst 732
Pancreas, anatomy 725
Pancreas, annular 759
Pancreas, blood supply 726
Pancreas, cystadenocarcinoma 749, 777
Pancreas, cystic lesions 763
Pancreas, ectopic 759
Pancreas, function 726
Pancreas, intraductal mucinous neoplasm (IPMN) 747
Pancreas, mucinous cystic neoplasm 746
Pancreas, nerve supply 726
Pancreas, periampullary carcinoma 748
Pancreas, pseudocyst 777
Pancreas, serous cystic neoplasm 747
Pancreatectomy, distal 627, 636, 743
Pancreatectomy, Fortner's regional 753
Pancreatectomy, total 743, 744, 754
Pancreatic abscess 761
Pancreatic ascites 760
Pancreatic calculus 760
Pancreatic carcinoma, adjuvant therapy 755
Pancreatic carcinoma, pain control 755
Pancreatic carcinoma, prognostic factors 755
Pancreatic carcinoma, TNM staging 752
Pancreatic cholera 758
Pancreatic divisum 759
Pancreatic duct decompression 743
Pancreatic enzyme replacement therapy (PERT) 742
Pancreatic fistula 733, 760
Pancreatic fistulae, complications 761
Pancreatic juice 726
Pancreatic necrosis 732, 761
Pancreatic polypeptide tumour 758
Pancreatic pseudoaneurysm 733
Pancreatic resection 743
Pancreatic sphincter dyskinesia 710
Pancreatic trauma 762
Pancreatic trauma, classification 762

Pancreatic trauma, grading 762
Pancreatic tumours 746
Pancreatic tumours, acinar cell carcinoma 746
Pancreatic tumours, cystadenocarcinoma 749
Pancreatic tumours, endocrine 755
Pancreatic tumours, exocrine 746
Pancreatic tumours, infiltrating ductal adenocarcinoma 746
Pancreatic, abscess 733
Pancreaticoduodenectomy 743
Pancreaticoduodenectomy, Traverso-Longmire pylorus preserving 753
Pancreaticojejunostomy 743
Pancreatitis 727
Pancreatitis, acute 727
Pancreatitis, acute interstitial 729
Pancreatitis, acute physiology and chronic health evaluation 734
Pancreatitis, acute, APACHI II scoring 734
Pancreatitis, acute, Atlanta classification 729
Pancreatitis, acute, Balthazar CT scoring system 730
Pancreatitis, acute, complications 732
Pancreatitis, acute, conservative treatment 731
Pancreatitis, acute, metabolic/biochemical changes 729
Pancreatitis, acute, plain X-ray 730
Pancreatitis, acute, spiral CT 730
Pancreatitis, acute, surgical treatment 730
Pancreatitis, autoimmune 729
Pancreatitis, chronic 738
Pancreatitis, chronic, classification 739
Pancreatitis, chronic, conservative treatment 742
Pancreatitis, chronic, CT abdomen 741
Pancreatitis, chronic, endocrine dysfunction 740
Pancreatitis, chronic, endoscopic therapy 742
Pancreatitis, chronic, endosonography 741
Pancreatitis, chronic, ERCP 740, 741
Pancreatitis, chronic, ERCP Cambrige grading 741
Pancreatitis, chronic, exocrine dysfunction 740
Pancreatitis, chronic, pathogenesis 738
Pancreatitis, chronic, surgery 742
Pancreatitis, chronic, TIGAR-O risk factor classification 738
Pancreatitis, early-onset idiopathic 739
Pancreatitis, emphysematous 734
Pancreatitis, fulminant 728
Pancreatitis, gallstone, Ranson's prognostic criteria 734
Pancreatitis, Glasgow-Imrie prognostic criteria 734
Pancreatitis, haemosuccus 733
Pancreatitis, hereditary 728
Pancreatitis, idiopathic 729
Pancreatitis, Marseille's classification 727
Pancreatitis, necrotising 729
Pancreatitis, non-gallstone, Ranson's prognostic criteria 734
Pancreatitis, Trapnell's classification 727
Pancreatitis, tropical 739
Pancreatitis, Ulm classification 729
Pancreatoblastoma 746
Pancreolauryl test 763
Panendoscopy 477
Papilloma 83
Papilloma, complications 84
Papilloma, duct 84
Papilloma, fibroepithelial 84
Papilloma, infective 83
Papilloma, true 83
Papillomatosis 563
Pappenheimer bodies 712
Para-aortic lymph node mass 777
Paradidymis 1155
Paradoxical aciduria 887
Paradoxical respiration 1196
Parafollicular 'C' cell 518
Paraganglioma 765
Parakeratosis 392
Paralysis, Erb-Duchenne 275
Paralysis, Klumpke's 275
Paralytic ileus 1000
Paraneoplastic syndromes 286
Paraoesophageal lymph nodes 839
Paraphimosis 1139
Paraplegia in extension 1192
Paraplegia in flexion 1192
Paratesticular tumours 1070
Parathormone 531
Parathyroid anatomy 531
Parathyroid, autotransplantation 526, 536
Parathyroidectomy 534, 536
Parathyroidectomy, complications 534
Parathyroidectomy, endoscopic 535
Parathyroidectomy, indications 534
Parathyroidectomy, minimal invasive 535
Parathyroidectomy, remedial 536
Parathyroidectomy, subtotal 536
Parathyroidectomy, video assisted 535
Parathyromatosis 532
Parenteral nutrition, home 105
Parenteral nutrition, peripheral 104
Parenteral nutrition, total 104
Parietal cell 870
Parkland regime 135
Paronychia, acute 160
Paronychia, chronic 161
Parotid gland, accessory 431
Parotid gland, anatomy 431
Parotid lymphoma 451
Parotid, lymphatic drainage 432
Parotidectomy 448
Parotidectomy, chemotherapy 449
Parotidectomy, complications 449
Parotidectomy, radical 449
Parotidectomy, radiotherapy 449
Parotidectomy, superficial 451
Parotidectomy, suprafacial 452
Parotidectomy, total conservative 449
Parotitis, acute 438
Parotitis, suppurative 438
Parotitis, suppurative, complications 438
Passaro's triangle 757
Patent ductus arteriosus 1217
Patey's stylomandibular tunnel 442
Patient controlled analgesia 366
Patterson operation 402
Pavulon 1232
PCWP 116
PDS 1256
Peau d'orange 569
Pel-Ebstein fever 266
Pelvic evisceration 1028
Penis, Ram's horn 1142
Peptic ulcer, acute 878
Peptic ulcer, bleeding 891
Peptic ulcer, perforated 887
Peptic ulcer, perforation, plain X-ray abdomen 888
Percarditis, chronic constrictive 1210
Percussion 771
Percutaneous nephrolithotomy (PCNL) 1091
Percutaneous transbiliary drainage (PTBD) 752
Percutaneous transhepatic biliary drainage 672
Percutaneous transhepatic cholangiography (PTC) 672
Percutaneous transluminal angioplasty 184
Percutaneous transluminal balloon angioplasty 190
Perdomo procedure 641
Perforator veins 231
Periappendicular phlegmon 1014
Pericardial tap 1210
Pericardiotomy, open 1210
Pericarditis 1209
Pericarditis, acute 1209
Pericystectomy 641
Pericystectomy, laparoscopic 641
Pericystojejunostomy 641
Perineal urethrostomy 1131
Perioesophageal lymph nodes 839
Periodicity 879, 883
Peripheral nerve injuries, causes 275
Peripheral nerve injuries, Seddon's classification 273
Peripheral nerve injuries, Sunderland's classification 274
Peripheral nerves 273
Peripheral nerves, classification 273
Peritoneal carcinomatosis 614
Peritoneal cavity 599
Peritoneal cavity, spaces 599
Peritoneal malignancy 613
Peritoneal malignancy, primary 613
Peritoneal tuberculosis 615
Peritoneum, anatomy 599

Peritoneum, parts 599
Peritoneum, physiology 599
Peritonitis, acute 600
Peritonitis, biliary 606
Peritonitis, diagnostic laparoscopy 603
Peritonitis, diffuse 602
Peritonitis, emergency laparotomy 604
Peritonitis, familial periodic 607
Peritonitis, laparotomy 603
Peritonitis, localised 602
Peritonitis, parturition/abortion 607
Peritonitis, postoperative 606
Peritonitis, primary 600
Peritonitis, sclerosing 606
Peritonitis, secondary 600
Peritonitis, spontaneous bacterial 605
Peritonitis, tertiary 600
Perleche 391
Perniosis 23
Per-rectal examination 1021
Pes anserinus 432
Peutz-Jegher's polyp 962
Peutz-Jegher's syndrome 937
Phagedena 27, 392
Phantom hernia 835
Phantom limb 338, 835
Phantom tumour 835
Pharyngeal pouch 462
Pharyngeal pouch, problem 462
Phenol 1245
Pheochromocytoma 542
Pheochromocytoma, malignant 543
Phimosis 1138
Phlebectomy, ambulatory 248
Phlebectomy, electrodesiccation 248
Phlebectomy, transilluminated 248
Phlebography 241
Phlegmasia alba dolens 232
Phlegmasia cerulea dolens 232
Phosphorylated diethylstilbestrol 1128
Phrygian cap 673
Phytobezoar 898
Picibanil 465
Picker position 1022
Piezo-ceramic shock waves 1092
Piles 1038
Pilomatrixoma 295
Pin index code 1233
PIPIDA scan 673
Pitting 712
Pituitary tumours, classification 1186, 1187
Plain- X-ray (KUB) 1066
Plantar fasciitis 172
Plasma 124
Platelet concentrate 125
Pleomorphic adenoma, recurrence after excision 444
Plethysmography 241
Plethysmography, air 242
Plethysmography, photo 241
Pleural tap 1198
Pleurothotonus 51
Plombage 1214
Pneumothorax, tension 1197
Pneumatosis cystoides intestinalis 931
Pneumaturia 952
Pneumobilia 987
Pneumococcal vaccine 722
Pneumothorax 1197
Pneumothorax, classification 1197
Podoconiosis 260
Policeman's heel 172
Polyarteritis nodosa 210
Polycystic liver disease 650
Polyethylene glycol (PEG) 980
Polymerase chain reaction (PCR) 290
Polyvalent antisnake venom 362
Porta hepatic injury 631
Portal biliopathy 668
Portal hypertension 651
Portal hypertension, management 653
Portal hypertension, segmental 659
Portal hypertension, sinistral 651
Portal hypertensive gastropathy 661
Portal pyaemia 46, 637, 1013
Portosystemic collateralisation 652
Port-wine stain 222
Positron-emission tomography 1243
Post-cricoid tumour 863
Posterior urethral valve 1133
Postoperative obstruction, acute 989
Postoperative patient, monitoring 1234
Postoperative synergestic gangrene 26
Postphlebitic limb 250
Post-thrombotic syndrome 237
Potato nose 297
Pott's puffy tumour 1182
Pott's puffy tumour 48
Pott's spine 1191
Pott's spine, complication 1192
Pouch of Douglas 981
Pouch, coonic 981
Pouch, Hartmann's 981
Pouch, Heidenhain 981
Pouch, ileal 980
Pouch, jejunal 980
Pouch, Morrison's hepatorenal 981
Pouch, Pavlov 981
Pouch, pharyngeal 981
Pouch, Physick's 981
Pouch, Rathke's 981
Pouch, Seessel's 981
Pouches, surgical 980
Pouchet's procedure 881
Poupart's ligament 800
PPV 23 722
Pregangrene 177
Pre-hospital trauma life support 143
Prepyloric vein of Mayo 870
Presbyoesophagus 839
Pressure sore 21
Pressure sore, factors causing 21
Pressure sore, staging 22
Pressure sores 214
Pretracheal fascia 484
Priapism 1141
Pringle manoeuvre 154, 629
Proctalgia fugax 1060
Proctitis 1060
Proctoscope 1022, 1255
Proctoscopy 1022
Profundaplasty 184, 188, 191
Prolene 1256
Propofol 1232
Propranolol 506
Propylthiouracil 506
Prostate specific antigen 1121
Prostate, anatomy 1121
Prostate, capsule 1121
Prostate, floating 1129, 1130
Prostate, McNeal's three zones 1121
Prostatectomy, Freyer's suprapubic transvesical 1124
Prostatectomy, Millin's retropubic 1124
Prostatectomy, radical 1127
Prostatectomy, Young's perineal 1124
Prostatic abscess 1125
Prostatic calculi 1122
Prostatism 1122
Prostatitis 1125
Prostatitis, acute 1125
Prostatitis, chronic 1125
Prosthesis, Canadian Syme's 339
Prosthesis, computer-assisted designing 339
Prosthesis, internal 339
Prosthesis, lower limb 339
Prosthesis, nonsuction type 339
Prosthesis, patellar tendon bearing (PTB) 339
Prosthesis, SACH 339
Prosthesis, suction type 339
Prosthesis, tilting table 339
Prosthesis, types 339
Prosthesis, upper limb 339
Protamine sulphate 252
Prothrombin complex concentrate 125
Prothrombin time 628
Proton pump inhibitors 884
Proud flesh 19
Pruritus ani 1060
Psammoma bodies 512, 1184
Pseudoachalasia 853
Pseudoappendicitis 1006
Pseudocyst, lung 1204
Pseudoeschar 137
Pseudogynaecomastia 561
Pseudohyponatraemia 96
Pseudolipoma 71
Pseudomesenteric cyst 625
Pseudomyxoma peritonei 614, 1016
Pseudopneumoperitoneum 891
Psoas abscess 769
Psoas shadow 1067
PTA 190
PTFE mesh 827
Pudendal nerve latency study 1032
Pulmonary capillary wedge pressure 116
Pulmonary circulation 1216
Pulmonary complications, postoperative period 1214

Pulmonary stenosis 1219
Pulseless disease 188
Pumping veins 230
Purified chick embryo cell culture vaccine 59
Purified vero cell rabies vaccine 59
Purpura, idiopathic thrombocytopenic (ITP) 718
Purpura, thrombotic thrombocytopenic (TTP) 720
Pyaemia 45
Pyelolithotomy 1092
Pyelolithotomy, anatrophic 1094
Pyelolithotomy, coagulum 1094
Pyelolithotomy, extended 1092
Pyeloplasty, dismembered 1088
Pyeloplasty, laparoscopic 1088
Pyeloplasty, non-dismembered 1088
Pylephlebitis 637
Pyloric stenosis, barium meal study 885
Pyloric stenosis, chronic duodenal ulcer 885
Pyloric stenosis, congenital hypertrophic 876
Pyloric stenosis, electrolyte changes 885, 887
Pyloric stenosis, treatment 886
Pyloroplasty 918
Pyloroplasty, Finney's 893
Pyloroplasty, Finney's 918
Pyloroplasty, Heineke-Mikulicz 918
Pyocele 1155
Pyogenic granuloma 19, 49
Pyonephrosis 1088
Pyrosis 843

Q

Quadrantectomy 582
QUART therapy 582, 588

R

Rabies 58
Rabies, avian vaccine 59
Rabies, BPL inactivated vaccine 59
Rabies, cell culture vaccine 59
Rabies, dogs 60
Rabies, fixed virus 58
Rabies, immunisation in animals 60
Rabies, nervous tissue vaccine 59
Rabies, passive immunity 60
Rabies, pathogenesis 58
Rabies, post-exposure prophylaxis 60
Rabies, pre-exposure prophylaxis 60
Rabies, street virus 58
Rabies, sylvatic 58
Rabies, urban 58
Rabies, vaccines 59
Rabies, wild life 58
Rabies, wound classification 59
Rabies, wound treatment 59
Radiation 1244
Radiation absorbed dose 1225
Radiation unit 1224
Radical excision 94
Radio-remnant ablation 514
Radioactive I^{123} scan 489
Radioactive I^{125} fibrinogen study 233
Radioactive iodine 508
Radionuclide imaging 1243
Radioresistant malignancies 1225
Radiosensitive malignancies 1225
Radiotherapy 1224
Radiotherapy, complications 1226
Radiotherapy, concomitant 1226
Radiotherapy, curative 1224
Radiotherapy, induction 1226
Radiotherapy, interstitial 1225
Radiotherapy, intracavitory 1225
Radiotherapy, palliative 1224
Radiotherapy, physicochemical effects 1224
Radiotherapy, postoperative 1226
Radiotherapy, preoperative 1226
Radiotherapy, sources 1225, 1226
Radiotherapy, types 1224, 1225
Rains factor 123
Ram's horn nail 172
Ram's horn penis 260
Randall's plaque theory 1090
Ranula 389
Ranula, marsupialisation 390
Ranula, plunging 389
Rashkind balloon septostomy 1219
Ray fungus 56
Raynaud's disease 189, 216
Raynaud's phenomenon 188, 216
Raynaud's phenomenon, treatment 219
Raynaud's syndrome 188, 216
Raynaud's syndrome, Coffman criteria 189
Reactive hyperaemia time 189
Reboring 191
Reconstruction 340
Rectal plication, Devadhar 1033
Rectal prolapse 1030
Rectal prolapse, aetiology 1030
Rectal prolapse, defecographic grading 1032
Rectal prolapse, hidden/concealed 1030
Rectal prolapse, partial 1030
Rectal prolapse, treatment 1030
Rectal proplapse, complete 1031
Rectopexy 1033
Rectosigmoid junction 1020
Rectum, anatomy 1020
Rectum, perineal posterior fixation of Lockhardt-Mummery 1034
Rectum, supports 1020
Recurrent childhood parotitis 438
Recurrent fibroid of Paget's 84
Recurrent laryngeal nerve palsy 523, 525, 527, 1263
Recurrent laryngeal nerve palsy, bilateral 529
Recurrent laryngeal nerve palsy, unilateral 529
Recycled blood 127
Red currant jelly 780
Red currant jelly stool 995
Reed-Sternberg cells 265
Refeeding syndrome 102, 106
Reflux oesophagitis, Los Angeles classification 848
Reflux oesophagitis, Savary-Miller classification 848
Renal abscess 1089
Renal angiogram 1102
Renal calculus 1089
Renal calculus, aetiology 1090
Renal calculus, indigo 1091
Renal calculus, struvite 1091
Renal calculus, urate 1090, 1107
Renal calculus, uric acid 1090, 1107
Renal calculus, xanthine 1091
Renal carbuncle 1088
Renal cavernostomy, Hanley's 1083
Renal cell carcinoma 1100
Renal cell carcinoma, AJCC staging 1101
Renal cell carcinoma, chemotherapy 1103
Renal cell carcinoma, CT scan 1102
Renal cell carcinoma, MRI 1102
Renal cell carcinoma, Robson-Flocks and Kadeskey staging 1101
Renal cell carcinoma, triad 1102
Renal cell carcinoma, types 1100
Renal injury, grading 1080
Renal pain 1091
Renal papillary adenoma 1098
Renal pelvic tumour 1104
Renal pelvis, duplication 1078
Renal pharmacoangiogram 1069
Renal stones, conservative management 1092
Renal stones, cystine 1090, 1107
Renal stones, oxalate 1090, 1107
Renal stones, phosphate 1090, 1107
Renal stones, recurrence 1097
Renal stones, struvite 1091
Reperfusion injury 199
Residual urine 1116
Resistive index 496
Rest pain 176
Resting salivary flow 433
Restrictive bariatric surgery 107
Retention of urine, acute 1136
Retention of urine, acute on chronic 1137
Retention of urine, chronic 1136
Retractor, Balfour's 1250
Retractor, Czerny's (hernia) 1249
Retractor, Deaver's 1249
Retractor, Doyen's 1250
Retractor, Joll's thyroid 1251
Retractor, Langenbeck's 1249
Retractor, Morris' 1249
Retractor, self-retaining 1250
Retractor, single hook 1250
Retractor, Volkmann's 1249
Retractors 1249

Retrograde pyelography 1068
Retrograde Seldinger technique 1069
Retromammary space 545
Retromolar trigone 396
Retroperitoneal cyst 765
Retroperitoneal fibrosis 764
Retroperitoneal fibrosis, triad 764
Retroperitoneal mass 765
Retroperitoneal radical lymph node dissection (RPLND) 1169
Retroperitoneal radical lymph node dissection, modified (RPLND) 1169
Retroperitoneal sarcoma 769
Retroperitoneal sarcoma, Dutch/memorial Sloan-Kettering cancer centre classification 769
Retroperitoneal sarcoma, features 768
Retroperitoneal space, anatomy 764
Retroperitoneal swellings 765
Retroperitoneal tumours 765, 766
Retroperitoneal tumours, AJCC grading 767
Retroperitoneal tumours, TNM staging 767
Retroperitoneal tumours, treatment 768
Retroperitoneoscopy 1270
Retropharyngeal abscess 466
Retropharyngeal abscess, acute 466
Retropharyngeal abscess, chronic 467
Reversal agents 1232
Reynold's pentad 691
Rhabdomyosarcoma 327
Rhinophyma 297
Richet's fascia 829
Riddle's triangle 523
Right to left shunt 1217
Ringer's lactate 101
RIPASA scoring system 1011
Risus sardonicus 50
Rituximab 356
River struma 519
Robertson's cooked meat media 53
Rockall scoring system 892
Roentgen unit 1225
Rokitansky-Aschhoff's sinuses 688
Rose Bengal I^{131} study 678
Rotter's nodes 547, 570
Roundworm obstruction 931
Roux- en-Y procedure 754
Roux-en-Y gastric bypass 108
Roux-en-Y gastric bypass, laparoscopic 109
Rupture bladder 1115
Rupture bladder, complications 1116
Rupture, anterior urethra 1130
Rupture, bulbous urethra 1130
Rupture, bulbous urethra, treatment 1130
Rupture, membranous urethra 1129
Rupture, membranous urethra, MaCallum-Colapinto classification 1129
Rupture, posterior urethra 1129
Rutherford Morison's kidney pouch 609

S

SA node 1216
SAAG 662
Sabre tibia 28
Sacrococcygeal teratoma 1058
Saddle embolus 200
Saegesser's tender point 713
SAG-M blood 125
Saint's triad 680, 694, 847, 951
Salazopyrine 927, 958
Saliva 434
Salivary tumour, acinic cell 446
Salivary tumour, adenocarcinoma 446
Salivary tumour, adenoidcystic carcinoma 446
Salivary tumour, adenolymphoma 444
Salivary tumour, AJCC, TNM staging 447
Salivary tumour, carcinoma expleomorphic adenoma 446
Salivary tumour, carcinosarcoma 446
Salivary tumour, facial nerve sacrifice 449
Salivary tumour, FNAC 448
Salivary tumour, malignant mixed 446
Salivary tumour, malignant, general features 447
Salivary tumour, management 447
Salivary tumour, mucoepidermoid 445
Salivary tumour, open biopsy 443
Salivary tumour, parotid, surgery 448
Salivary tumour, radiotherapy 449
Salivary tumour, squamous cell carcinoma 447
Salivary tumour, submandibular 447
Salivary calculus 435
Salivary colic 435
Salivary gland excision, submandibular 437
Salivary gland tumour, minor 450
Salivary gland, ectopic 434
Salivary gland, minor 434
Salivary gland, sublingual 434
Salivary neoplasm 440
Salivary neoplasm, WHO classification 440
Salivary nucleus, inferior 454
Salivary nucleus, superior 454
Salivary tumour, mixed 441
Salivary tumour, papillary cystadenolymphomatosum 444
Salivary tumour, pleomorphic adenoma 441
Salmon patch 222
Sandblom triad 704
Santulli operation 993
Saphena varix 238, 822
Saphenous neuralgia 244
Saphenous vein graft, *in situ* 192
Saphenous vein graft, reverse 192
Saphenous vein, long 230
Saphenous vein, short 230
Sarcoidosis 272
Sarcoma 282
Sarcoma, Enneking classification of surgical resection 323
Sarcoma, FNAC 320
Sarcoma, incision biopsy 320, 321
Sarcoma, isolated limb perfusion 325
Sarcoma, trucut biopsy 320
Sarcomas 318
Sarcomas, chemotherapy 325
Sarcomas, features 318
Sarcomas, lymph node spread 329
Sarcomas, radiotherapy 324, 325
Saturday night palsy 280
Savage's decompressor 989
Savlon 1245
Scabbard trachea 526
Scalene manoeuvre 179
Scalene triangle 174, 456
Scalvo's serum 61
Schatzki's rings 856
Schiller-Duval bodies 1166
Schistosoma haematobium 1111
Schwannoma 92, 1184
Scissors 1252
Scleroderma/systemic sclerosis 211
Sclerosing adenosis 552
Sclerosing angioma 296
Sclerotherapy, advantages 244
Sclerotherapy, catheter directed 243
Sclerotherapy, disadvantages 244
Sclerotherapy, foam 243
Sclerotherapy, micro 243
Sclerotherapy, transillumination 243
Scolicidal agents 641
Scoline 1232
Scoliosis capitis 470
Scoop, Volkmann's 1252
Scrofuloderma 473
Scrotum, anatomy 1148
Scrum pox 49
Sebaceous horn 81
Seborrhoeic keratosis 298
Seborrhoeic wart 298
Second heart sound 1215
Segmental pressure measurement 179
Segmentectomy 666
Self-expanding metal stents (SEMS) 865
Semilunar line of Spigel 833
Seminoma testis 1166
Semon's law 529
Semple vaccine 59
Sengstaken-Blakemore tube 657
Senile keratosis 298
Sentinal lymph node biopsy 566, 1144
Sentinel node biopsy 317
Sentinel node, Cabana 1144
Sentinel pile 1047
SEPS 247
Septal defect, atrial 1218
Septal defect, ventricular 1219
Septicaemia 45
Septicaemia, complications 45
Septicaemia, stages 45

Septicaemia, types 45
Sequestrum 34, 177
Serocystic disease of Brodie 552
Serum amylase 726
Serum ascites albumin gradiant 662
Serum lactescence 730
Serum lipase 730
Seton technique 1056
Sharma and Jhawer's technique 1153, 1261
Shepherd crook deformity 387
Sherren's triangle 1008
Shianoya's criteria 185
Shock 111
Shock index 115
Shock lung 1203
Shock, anaphylactic 111, 115
Shock, cardiac compression 114
Shock, cardiogenic 113
Shock, cardiogenic, management 113
Shock, causes 111
Shock, clinical features 115
Shock, distributive 117
Shock, effects 112
Shock, hypovolaemic 111
Shock, neurogenic 111, 113
Shock, pathophysiology 112
Shock, septic 114
Shock, septic, pathophysiology 114
Shock, septic, stages 114
Shock, septic, treatment 114
Shock, stages 112
Shock, treatment 115
Shock, types 112, 113
Shock, vasovagal 113
Shouldice repair 810
Shouldice repair, Berliner modified 810
Shunt, Denver 664
Shunt, distal splenorenal 659
Shunt, Inokuchi 659
Shunt, LeVeen 664
Shunt, Linton's 659
Shunt, porta-caval 659
Shunt, proximal splenorenal 659
Shunt, Warren's 659
Shunts, nonselective 659
Shunts, selective 659
Shutter mechanism 801
Sialadenectomy 437
Sialadenitis 435
Sialadenitis, acute 435
Sialadenitis, chronic 435
Sialectasis 437
Sialography 434
Sialolithiasis 436
Sialorrhoea 433
Sialosis 437
Sick cell syndrome 113, 122
Sickle cell disease 718
Sideropenia 394
Sideropenic dysphagia 394
Sigmoidopexy 998
Sigmoidoscopy 1022
Sign 'V' 162
Sign of 'V' of Naclerio 856
Sign of Dance 995
Sign of vas 1168
Sign, 'C' loop-pad 750
Sign, Angell's 1165
Sign, Ballance's 713
Sign, Battle's 1174
Sign, bent-inner tube 998
Sign, Berry's 512, 513, 516, 517, 520
Sign, bird's beak 998
Sign, Blumberg's 602, 888, 987, 1008
Sign, Boas's 685
Sign, Bow 1158
Sign, Branham's 228
Sign, bull's eye 996
Sign, Camellotte 638
Sign, Champagne bottle 238, 249
Sign, Champagne glass 953
Sign, Chvostek's sign 538
Sign, claw 996
Sign, coffee-bean 998
Sign, coiled spring 996
Sign, Coleman's 378
Sign, colon cut off 730
Sign, Cowen's 503
Sign, Cullen's 728
Sign, Cupola 890
Sign, Curtain 442
Sign, Dalrymple's 502
Sign, Deming's 1165
Sign, double duct 752
Sign, double-bubble 990
Sign, Doughnut 877, 996
Sign, Enroth 503
Sign, fat ring 612
Sign, Fleischner 618
Sign, football 890
Sign, Fox 728
Sign, Froment's 278, 458
Sign, Frostberg's reverse '3' 874
Sign, Gaur's 822
Sign, Giffrod's 503
Sign, Grey Turner's 728
Sign, Guerin's 375
Sign, Hamman 857
Sign, Harvey's 178
Sign, Homan's 233
Sign, Hood 488
Sign, Hook 164, 778
Sign, Hoover's 610
Sign, Howship-Romberg 834
Sign, Hutchinson 312
Sign, inverted beer bottle 249
Sign, inverted umbrella 618
Sign, inverted V 890
Sign, Jellinek's 503
Sign, Joffroy's 502
Sign, Kanavel 164
Sign, Kehr's 713
Sign, Knie's 503
Sign, Kocher's 503
Sign, Loewi's 503
Sign, London's 153
Sign, Lyre 468
Sign, Mallet-Guys 740
Sign, Mathe's 1089
Sign, Mercedes Benz 680, 694
Sign, Milian's ear 39
Sign, Moebius 502
Sign, Mose's 233
Sign, Moynihan's 689
Sign, Murphy's 689
Sign, Nafziger's 503
Sign, Neuhauser's 993
Sign, Neuhof's 233
Sign, Nicoladoni's 228
Sign, Omega 998
Sign, pad 874
Sign, Panda 373, 376, 1174
Sign, Pemberton's 266, 497
Sign, Prehns 1165
Sign, pseudokidney 996
Sign, puddle 662
Sign, Racoon's eye 541
Sign, reverse '3' 750
Sign, reverse 'J' 1068
Sign, Riglers 890
Sign, Rosenbach's 503
Sign, Rovsing's 1008
Sign, Rust's 474, 1191
Sign, Seagull 680, 694
Sign, Silk glove 805
Sign, slip 72
Sign, Srinivasan costal 782
Sign, Stellwag's 502
Sign, Stemmer's 260
Sign, Stierlien's 618
Sign, string 618
Sign, sun-setting 1181
Sign, Tanyol 662
Sign, target 996
Sign, thumb printing 959
Sign, Tinel's 275
Sign, trial 495
Sign, triangle 890
Sign, Troisier's 907, 909
Sign, Trousseau's 99, 252, 538, 749, 909
Sign, Vermooten's 1129
Sign, Von Graefe's 502
Sign, Westmark 1204
Sign, Whirl 1002
Silastic silo 793
Silk 1256
Simmond's procedure 173
Single photon emission computerised tomography (SPECT) 533
Sinus 30
Sinus forceps 43
Sinus, discharge 33
Sinus, median mental 33
Sinus, pilonidal 1036
Sinus, pilonidal, Buie's marsupialisation 1038
Sinus, pilonidal, Lord and Millar's excision 1038
Sinus, pilonidal, treatment 1037
Sinus, preauricular 35, 380

Sinusogram 33
Sir Philip Manson-Bahr amoebic point 960
Sirolimus 356
SIRS 112, 117, 132
Sister Joseph's nodules 784, 791
Sister Mary Joseph nodules 907, 909
Sister-Joseph nodule 294
Sitz bath 1042
Siver suphadiazine 136
Skin adnexal tumours 284
Skin cancer, TNM staging 300
Skin tag 84
Skin tumours, classification 293, 294
Skin, anatomy 293
Skin, eccrine gland tumours 294
Skin, hair tumour 294
Skin, melanotic cancer 294
Skin, non-melanotic cancer 294
Skin, premalignant conditions 298
Skin, squamous cell carcinoma 298
Skin, squamous cell carcinoma, prognostic factors 301
Skin, squamous cell carcinoma, treatment 301
Skinner en bloc resection 864
Slit valve mechanism 801
Sloan Kettering scoring 510
Slough 177
Small bowel carcinoid 939
Small bowel enema 942
Small bowel enteroscopy 943
Small bowel strictures 931, 934
Small bowel swellings 780
Small bowel tumours 935
Small bowel, adenocarcinoma 937
Small bowel, adenoma 936
Small bowel, benign tumours 936
Small bowel, carcinoid tumour 938, 939
Small bowel, GIST 938
Small bowel, leiomyoma 936
Small bowel, lipoma 937
Small bowel, malignant tumours 937, 938
Small bowel, malignant tumours, treatment 940
Small bowel, non-Hodgkin's lymphoma 938
Small intestine, anatomy 922
Small intestine, diseases investigation 923
Smoking index 185
Snail track ulcer 392
Socia parotidis 431
Sodium diatrizoate 1067, 1240
Sodium tetradecyl sulphate 3% (STDS) 242
Sodium trizoate 434
Soft chancre 28
Soft corn 171
Soft sore 28
Soft tissue sarcoma 318
Soft tissue sarcoma, features 320
Soft tissue sarcomas, amputation in 322, 324
Soft tissue sarcomas, compartment resection 324
Soft tissue sarcomas, grading 322
Soft tissue sarcomas, GTNM staging 322
Solar keratosis 298
Solitary ulcer syndrome 1029
Somatic cell therapy 1230
Somatostatin 656, 873
Somatostatin receptors 872
Somatostatinoma 758
Somatostatinoma, triad 758
Southampton wound grading system 36
Souttar tube 865
Space of Bogros 815
Space of Retzius 815
Space, midpalmar 162
Space, of parona, infection 164
Space, thenar 162, 163
Spermatic cord, contents 800
Spermatic cord, coverings 801
Spermatocele 1156
Sphincter of Oddi 670
Sphincter of Oddi dysfunction 710
Sphincterotomy 692, 693
Sphincterotomy, complications 692
Spider veins 235
Spigelian fascia 833
Spina bifida 1188
Spina bifida aperta 1189
Spina bifida occulta 1188
Spina ventosa 170
Spinal dysraphism 1187
SPINKI 739
Splash in the pan 1041
Spleen, accessory 712
Spleen, blood supply 711
Spleen, function 711
Spleen, ligaments 711
Spleen, red pulp 711
Spleen, rupture 712
Spleen, salutary effect 715
Spleen, surgical anatomy 711
Spleen, wandering 724
Spleen, white pulp 711
Splenectomy 659, 720
Splenectomy, complication 154, 715, 721
Splenectomy, emergency 715
Splenectomy, indications in ITP 719
Splenectomy, laparoscopic 721
Splenectomy, partial 715
Splenectomy, portal hypertension 659
Splenic artery 711
Splenic organ injury scale 714
Splenic rupture, signs 713
Splenic vein 711
Splenic vein cavernoma 659
Splenomegaly 716
Splenoportography 653
Splenorrhaphy 715
Splenosis 713
Spongioblastoma polare 1184
Spread of malignant tumours 284
Spurious diarrhoea 1024
Squamous cell carcinoma, Broder's classification 300
Squamous cell carcinoma, Broder's grading 285
SQUID 988
SSI risk index 68
Stafne bone cyst 434
Staghorn calculus 1097, 1107
Stapler haemorrhoidectomy 1045
Staplers in surgery 1265
Staplers, circular (EEA) 1265
Staplers, cutaneous 1265
Staplers, endo 1266
Staplers, GIA 1266
Staplers, linear 1265
Stenting, intravascular 216
Step ladder paristalsis 986, 995
Stephen's line 1035
Stephen-Fowler's technique 1163
Stercoral ulcer 969
Sterile pyuria 1083
Sterilisaion, methods 1245
Sterilisation 1244
Stiff lung 1203
Stings and bites 362
Stoma appliances 978
Stoma care 977
Stoma, complications 978
Stoma, general care and advice 979
Stoma, patient counselling 977
Stoma, postoperative care 978
Stoma, skin, excoriation 978
Stoma, types 977
Stomach, anatomy 869
Stomach, blood supply 870
Stomach, leather-bottle 908
Stomach, lymphatic drainage 871
Stomach, lymphatic drainage zones 907
Stomach, nerve supply 870
Stomach, palpable mass 776
Stomach, tiger stripe 894
Stomach, water melon 894
Stomal obstruction 895
Stomal ulcer, perforated 890
Stomatitis 390
Stomatitis, aphthous 391
Stomatitis, *Candida* 391
Stomatitis, nutritional 391
Stomatitis, traumatic 391
Stomatitis, Vincent's ulcerative 391
Stool occult blood test 1064
Stoppa's giant prosthesis reinforcement of visceral sac 814
Stork bite 222
Stove in chest 1196
Strawberry scrotum 82
Streptokinase 199
Stretta catheter 844
Stricture urethra 1131
Stricture urethra, classification 1131
Stricture urethra, complications 1132
Stricture urethra, intermittent dilatation 1132

Stridor 497
String sign of Kantor 927
Stripper, Codman's 244
Stripper, Myer's 244
Stripper, Oesch pin 246
Stripping, complications 244
Struma lymphomatosa 519
Struma ovarii 500
Student's elbow 85
Stump carcinoma 903
Stump neuroma 338
Subepithelial benign nodular fibrosis 296
Subhyoid bursitis 467
Submandibular salivary gland, anatomy 432
Subphrenic abscess, pathology 609
Subphrenic abscess, treatment 610
Subphrenic spaces 608
Succussion splash 885
Sucquet-Hoyer canal 82
Sucralfate 884
Sulfamylon 136
Superior mesenteric vein thrombosis 932
Superobesity 106
Supralevator encirclement of anal orifice of Notaras 1034
Suprapubic cystostomy 1074
Suprapubic open cystolithotomy 1109
Suprapubic percutaneous litholapaxy 1109
Surgeon and law 1276
Surgery, endovascular 215
Surgical complications, pulmonary tuberculosis 1214
Surgical complications, roundworm 929
Surgical complications, typhoid 928
Surgical infection 36
Surgical site infection (SSI) 67
Surgical site infection, classification 68, 69
Surgical site infection, management 70
Surgical site infection, prevention 69
Suture aterial, ideal 1256
Suture knots 1257
Suture materials, absorbable 1256
Suture materials, classification 1256
Suture materials, memory 1256
Suture materials, nonabsorbable 1256
Suture materials, numbering 1257
Suturing types 1257
Suturing, delayed primary 7, 8
Suturing, primary 7, 8
Suturing, secondary 7, 8
Suxamethonium chloride 1232
Swelling, cross-fluctuant 94
Swellings, brilliantly transilluminant 1150
Swellings, cross fluctuant 770
Swellings, cross-fluctuation 1151
Swenson's operation 951
Swimmer's itch 1111
Sylvatic rabies 58
Sympathectomy, cervical 190, 194
Sympathectomy, cervicothoracic preganglionic 193
Sympathectomy, chemical 188, 195
Sympathectomy, lumbar 188, 194
Sympathectomy, thoracoscopic 194
Syndactyly 169
Syndrome Cronkite-Canada 964
Syndrome, abdominal compartment (ACS) 155
Syndrome, adrenogenital 540
Syndrome, auroculotemporal 452
Syndrome, Bannayan-Riley-Ruvalacaba (Ruvalacaba Myhre Smith) 964
Syndrome, Beckwith-Weidman 792, 829
Syndrome, Birt-Hogg-Dube 1100
Syndrome, blue toe 207
Syndrome, Boerhaave's 841, 856
Syndrome, Bouveret's 704
Syndrome, Brown-Sequard 146
Syndrome, Budd-Chiari 664
Syndrome, carpal tunnel 276
Syndrome, cast 899
Syndrome, central cord 146
Syndrome, cervical rib 457
Syndrome, Chilaiditi's 889
Syndrome, coeliac artery 899
Syndrome, coeliac band 899
Syndrome, cold agglutinin 718
Syndrome, compartment 196, 197, 199
Syndrome, Conn's 540
Syndrome, Costello 765
Syndrome, Cowden (multiple hamartoma-neoplasia syndrome) 964
Syndrome, Cowden's 509, 563
Syndrome, CREST 189
Syndrome, Cruvielhier-Baumgarten 654
Syndrome, Curtis-Fitz-Hugh 607
Syndrome, dancing eye 541
Syndrome, descending perineal 1059
Syndrome, diGeorge's 538
Syndrome, disappearing pulse 178
Syndrome, dumping 897
Syndrome, Durban's 899
Syndrome, Eisenmenger's 1217, 1219
Syndrome, familial paraganglioma 542
Syndrome, Felty's 720
Syndrome, Ferguson-Smith 300
Syndrome, Frey's 452
Syndrome, Frey's, treatment 453
Syndrome, Frolich's 1187
Syndrome, Frolin's 1193
Syndrome, Gardner's 792, 963
Syndrome, Gorlin 303
Syndrome, Heerfordt's 454
Syndrome, Horner's 193, 476, 477, 1208
Syndrome, Hungry bone 526, 535
Syndrome, juvenile polyposis 965
Syndrome, Kasabach-Merritt 223, 642
Syndrome, Klinefelter's 1166
Syndrome, Klippel-Trenauny 252
Syndrome, Klippel-Trenauny-Weber 223
Syndrome, Koenig's 615
Syndrome, Leriche's 175, 178, 200
Syndrome, Letessier-Meige's 257
Syndrome, Li-Fraumen's 563, 902
Syndrome, Loeffler's 930
Syndrome, Lynch I 965
Syndrome, Lynch II 965
Syndrome, Maffucci 223
Syndrome, Mallory-Weiss 856
Syndrome, McCune Albright's 387
Syndrome, MEA 536
Syndrome, median arcuate ligament 899
Syndrome, MEN 71, 536
Syndrome, MEN II 518
Syndrome, MEN IIb 964
Syndrome, Mendelson's 900, 1234
Syndrome, Mirizzi 683
Syndrome, Munchausen 1013
Syndrome, nurses' 612, 626
Syndrome, Ogilvies's 998
Syndrome, Osler-Rendu-Weber 223, 894
Syndrome, Paget-Schroetter 232, 457
Syndrome, pancoast 1208
Syndrome, Parkes-Weber 223
Syndrome, Paterson-Kelly 853
Syndrome, pectoralis minor 458
Syndrome, Pendred's 489
Syndrome, Pierre-Robin 368, 388
Syndrome, Plummer-Vinson 853
Syndrome, Polands 546
Syndrome, post-cholecystectomy 709
Syndrome, post-cibal 897
Syndrome, primary extraglandular sicca 440
Syndrome, *Proteus* 223
Syndrome, Prune Belly 798, 1161
Syndrome, Rapunzel 898
Syndrome, Reiter's 1135
Syndrome, Roux stasis 897
Syndrome, scalene 457
Syndrome, Sezzary 270
Syndrome, short bowel 941
Syndrome, short gut 941
Syndrome, Shprintzen's 368
Syndrome, Sjögren's 439
Syndrome, Sjögren's, primary 440
Syndrome, Sjögren's, secondary 440
Syndrome, Stauffer's 1102
Syndrome, Stewart-Treves 262, 320, 328, 583
Syndrome, Stickler's 368
Syndrome, Sturge-Weber 223
Syndrome, subclavian steals 196
Syndrome, sump 694
Syndrome, superior mesenteric artery 899
Syndrome, thoracic inlet 457
Syndrome, thoracic outlet 174, 456
Syndrome, thoracic outlet, causes 174
Syndrome, Tietze's 554
Syndrome, Torre-Muir 294
Syndrome, Treacher-Collins 381
Syndrome, Turcot's 963, 964
Syndrome, Turner's 937
Syndrome, Verner-Morrison 758
Syndrome, virilising 540
Syndrome, von Hippel Lindau 542, 747, 763, 765

Syndrome, Waltman-Walters 607
Syndrome, weak-tea 758
Syndrome, Werner' s 536
Syndrome, Wilkie's 874, 898
Syndrome, Zollinger-Ellison 758
Synovial sarcoma 328
Syphilis 56
Syphilis, congenital, 56
Syphilis, late 56
Syphilis, latent 28, 56
Syphilis, primary 28, 56
Syphilis, secondary 28, 56
Syphilis, tertiary 28, 56
Syphilis, types 56
Syphilitic chronic glossitis 392
Syrigomylocele 1188
Systemic circulation 1216
Szilagyi classification 221

T

T3 489
T3 toxicosis 500, 505, 508
T4 489
Tabes mesenterica 625
Tacrolimus 356, 927, 1229
Taeniae coli 948
Takayasu's arteritis 188
Tamoxifen 552, 584
Tamsulosin 1124
Tangential excision 135
Tanner's resection 658
TAO 184
TAO, classification 185
TAO, pathogenesis 185
Tapeworm, dog 637
Taste sensation 454
Taxis 811, 819
^{99m}Tc scan 495
^{99m}Tc sestamibi scan 515
Tea-pot deformity 879, 882
Technetium 99 pertechnetate 508
Technetium 99m 1243
Technetium 99m labeled sestamibi
  scan 533
Tegaserod 958
Telepaque 670
Teletherapy 1226
Temporal arteritis 190
Tendon 351
Tendon graft 352
Tendon healing 351
Tendon healing, stages 351
Tendon repair 351
Tendon suturing 351
Tendon suturing, Goldner method 351
Tendon suturing, Kessler method 351
Tendon transfer 352
Tenesmus 1024
Teratoma testis 1166
Tessari technique 243
Test, Adson's 179, 458
Test, Allen's 179, 458
Test, augmented histamine 872
Test, auscultopercussion 885
Test, Baid 735, 777
Test, Baldwing's 1008
Test, Bernstein 844
Test, Blaxland ruler 662, 781
Test, Brodie-Trendelenburg 238
Test, Buerger's postural 179
Test, C13/C14 breath 876
Test, card 278
Test, Carnett's 781
Test, Casoni's 638
Test, Cope's psoas 1008
Test, EAST-Elevated Arm Stress 458
Test, Ehrlich's 696
Test, Falconer 179
Test, Fegan's 239
Test, Fouchet's 696
Test, Fuchsig's 178
Test, Goldstein saline load 885
Test, Graham cole 670
Test, Halstead 216, 458
Test, Hay's 696
Test, head rising 805
Test, Hess 719
Test, Hollander 872
Test, Ian-Arid 239
Test, insulin (Hollander) 872
Test, internal ring occlusion 804
Test, Javid 196
Test, Kocker's 492, 526
Test, leg rising 804
Test, Linton's 233
Test, Lundh 763
Test, Mantoux skin 54
Test, military attitude 458
Test, modified Adson's 458
Test, modified Perthes' 239
Test, modified Roos 179
Test, Morrissey's cough impulse 239
Test, obturator 1008
Test, Orr chair 1164
Test, pancreolauryl 740, 742
Test, peak acid output 872
Test, pen 276
Test, pentagastrin 872
Test, Perthes' 239
Test, Phalen's 277
Test, Pratt's 239
Test, rapid urease 876
Test, ring invagination 804
Test, Roos 458
Test, Schwartz 239
Test, secretin cholecystokinin 741
Test, secretin cholecystokinin
  stimulation 763
Test, secretin stimulation 757
Test, Simian thumb 276
Test, sniff 1211
Test, Starch iodine 453
Test, tensilan diagnostic 1207
Test, three fingers 638, 773
Test, three tourniquet 239
Test, traction 1151
Test, Triolin breath 763
Test, Whitaker 1087
Test, Zieman's 804
Testicular descent, embryology 1161
Testicular sensation, loss 1154
Testicular tumours 1165
Testicular tumours, chemotherapy 1169,
  1170
Testicular tumours, classification 1166
Testicular tumours, germ cell 1166
Testicular tumours, histological
  classification 1166
Testicular tumours, hurricane type 1168
Testicular tumours, interstitial cell 1167
Testicular tumours, Leydig cell 1167
Testicular tumours, nongermcell 1166
Testicular tumours, Sertoli cell 1167
Testicular tumours, sex cord
  stromal 1166
Testicular tumours, staging 1168
Testicular tumours, TNM staging 1168
Testicular tumours, treatment 1169
Testicular tumours, tumour markers 1168
Testis, 'clapper in bell' 1164
Testis, anatomy 1160
Testis, appendage, torsion 1165
Testis, billiard 1170
Testis, ectopic 1163
Testis, ectopic, Lockwood theory 1163
Testis, gumma 1170
Testis, hanging/clapper in bell 810
Testis, maldescended 1161
Testis, retractile 1164
Testis, torsion 1164
Testis, torsion, treatment 1165
Testis, undescended 1161
Testis, undescended, complications 1162
Testis, undescended, hormone
  therapy 1163
Testis, undescended, laparoscopic
  approach 1163
Testis, vanishing 1161, 1162
Tetanus 49
Tetanus, ascending 50
Tetanus, bulbar 51
Tetanus, cephalic 50
Tetanus, complications 51
Tetanus, descending 50
Tetanus, early 50
Tetanus, incubation period 50
Tetanus, late 50
Tetanus, latent 50
Tetanus, localised 50
Tetanus, neonatorum 51
Tetanus, organism 50
Tetanus, pathogenesis 50
Tetanus, postoperative 51
Tetanus, postures in 51
Tetanus, predisposing factors 49
Tetanus, prophylaxis 52
Tetanus, staging 51
Tetanus, toxoid 52
Tetanus, treatment 51
Tetanus, types 50
Tetanus, urban 51

Tetany 99, 537
Tetracycline fluorescence test 911
Tetralogy of Fallot 1219
Thal's patch procedure 846
Thalassaemia 718
Thebesian veins 1215
Theirsch wiring 1030, 1034
Thiopentone 1232
Thoracic outlet, surgical anatomy 174
Thread veins 235
Throid, incidentaloma 520
Thrombectomy 191
Thrombectomy, venous 234
Thrombin inhibitors 253
Thromboangiitis obliterans, TAO 184
Thrombolysis 199
Thrombolytic agents 253
Thrombophlebitis 251
Thrombophlebitis migrans 252
Thrombosis, axillary vein 232
Thrombosis, deep vein 231
Thrombosis, spontaneous 232
Thrombosis, traveller's 231
Thrombus, fixed 233
Thrombus, free 234
Thymectomy 1207
Thymomas 1207
Thymomas, histological types 1207
Thymomas, staging 1207
Thyrocardiac 508
Thyroglobin estimation 514
Thyroglossal cyst, tugging sensation 488
Thyroglossal duct 484
Thyroid acropachy 504
Thyroid antibodies 489
Thyroid cancer, Degroot's staging 519
Thyroid carcinoma, anaplastic, stage grouping 517
Thyroid carcinoma, blood spread 517
Thyroid carcinoma, differentiated 509
Thyroid carcinoma, differentiated, American thyroid association 510
Thyroid carcinoma, differentiated, postoperative AJCC staging 510
Thyroid carcinoma, differentiated, TNM staging 510
Thyroid carcinoma, features of infiltration 517
Thyroid carcinoma, lymph node spread 517
Thyroid carcinoma, medullary, stage grouping 517
Thyroid carcinoma, papillary/follicular, stage grouping 517
Thyroid constipation 506
Thyroid cyst 497
Thyroid dermopathy 504
Thyroid diseases, stridor 520
Thyroid function tests 489, 505
Thyroid gland, anatomy 484
Thyroid gland, blood supply 485
Thyroid gland, congenital anomalies 486
Thyroid gland, development 484
Thyroid gland, ectopic 486
Thyroid gland, physiology 486
Thyroid hormone synthesis 486
Thyroid myopathy 504
Thyroid neoplasms 509
Thyroid neoplasms, Dunhill classification 509
Thyroid ophthalmopathy, Werner' abridged classification 504
Thyroid opthalmopathy, Werner's abridged classification 504
Thyroid paradox 497, 512
Thyroid steal 523
Thyroid stimulating antibodies 499
Thyroid storm 525
Thyroid, adenoma 494
Thyroid, agenesis 489
Thyroid, anaplastic carcinoma 515
Thyroid, black 520
Thyroid, cold nodule 494
Thyroid, discordant nodule 505
Thyroid, discrete nodule 494
Thyroid, dominant nodule 494, 520
Thyroid, Doppler 496
Thyroid, dyshormonogenesis 489
Thyroid, FNAC 489, 512
Thyroid, follicular carcinoma 512
Thyroid, follicular carcinoma, FNAC 513
Thyroid, hot nodule 494
Thyroid, Hurthle cell carcinoma 515
Thyroid, Kocher's incision 521
Thyroid, lateral aberrant 489
Thyroid, lingual 486
Thyroid, malignant lymphoma 519
Thyroid, medullary carcinoma 518
Thyroid, papillary carcinoma 510
Thyroid, papillary carcinoma, AGES scoring 512
Thyroid, papillary carcinoma, AMES scoring 512
Thyroid, papillary carcinoma, Woolner classification 510
Thyroid, pyramidal lobe 484
Thyroid, radio-remnant ablation 514
Thyroid, radioisotope study 505
Thyroid, solitary nodule 494, 520
Thyroid, Tracheal deviation 495
Thyroid, USG 490, 512
Thyroid, vocal cord position 529
Thyroid, warm nodule 494
Thyroidectomy 521
Thyroidectomy, completion 497, 515
Thyroidectomy, complications 524, 1263
Thyroidectomy, hemi 496, 507, 521, 1262
Thyroidectomy, minimally invasive video-assisted (MIVAI) 523
Thyroidectomy, near total 496, 512, 521, 1262
Thyroidectomy, partial 521, 1262
Thyroidectomy, prophylactic total 518
Thyroidectomy, subtotal 493, 507, 521, 1262
Thyroidectomy, total 493, 507, 512, 513, 518, 521, 1262
Thyroidectomy, total, prophylactic 519
Thyroiditis, autoimmune 499
Thyroiditis, de-Quervain's 499, 519
Thyroiditis, Hashimoto's 519
Thyroiditis, Ligneus 520
Thyroiditis, Reidel's 520
Thyroiditis, Woody's 520
Thyrotoxic crisis 525, 1263
Thyrotoxicosis 499
Thyrotoxicosis factitia 499, 500
Thyrotoxicosis, autoimmune 499
Thyrotoxicosis, cardiac manifestation, Crile's grading 504
Thyrotoxicosis, Jod Basedow 499, 500
Thyrotoxicosis, neonatal 499
Thyrotoxicosis, primary 501
Thyrotoxicosis, secondary 501
Thyrotoxicosis, T3 508
Thyrotoxicosis, trophoblastic 500
Thyrotoxicosis, Wayne's diagnostic indices 502
Tillaux's triad 611
Tinea pedis 172
Tongue fissure 416
Tongue muscles 414
Tongue tie 417
Tongue ulcers 415
Tongue, anatomy 414
Tongue, aphthous ulcer 415
Tongue, benign tumour 416
Tongue, dental ulcer 415
Tongue, development and nerve supply 415
Tongue, herpetic lingual ulcer 415
Tongue, lymphatic drainage 414
Tongue, movements 414
Tongue, papillae 414
Tongue, smoker's ulcer 415
Tongue, syphilitic ulcer 415
Tongue, tuberculous ulcer 415
Topical O_2 therapy 118
Torticollis 469
Torus palatinus 367
Total abdominal colectomy 972
Total laryngectomy 430
Total laryngectomy, after care 430
Total proctocolectomy 958
Touhy needle 1236
Tourniquet palsy 280
Tourniquets 127
Tourniquets, complications 128
Tourniquets, types 128
Towel clip, Mayo's 1247
Toxerutin 242
Toxic adenoma 499
Toxic megacolon 955, 958
Toxic thyroid, children 508
Toxic thyroid, pregnancy 508
Toxic thyroid, radioactive isotopes 509
Toxic thyroid, radioiodine therapy 507
Toxic thyroid, surgery 507
TPN, complications 105
TPN, components 105
Tracheostomy tube 1252

Tracheostomy tube, Fuller's bivalved 1252
Tracheostomy tube, Jackson's 1252
Tracheostomy, care 1264
Tracheostomy, complications 1264
Tracheostomy, indications 1263
Tracheostomy, technique 1264
Transcutaneous oximetry 181
Transcyte 136
Transfusion related graft versus host disease 126
Transfusion related lung injury 126
Transhiatal oesophagectomy 846
Transillumination, 'Chinese lantern pattern' 1156
Transitional cell carcinoma, classification 1112
Transjugular intrahepatic portosystemic shunt (TIPSS) 658, 659
Transplantation 353
Transplantation, bone marrow 359
Transplantation, donor criteria 353
Transplantation, heterotopic liver 357
Transplantation, isolated pancreatic islet 359
Transplantation, liver 357
Transplantation, liver, complications 358
Transplantation, liver, donor criteria 357
Transplantation, organ procurement 354
Transplantation, orthotopic liver 357
Transplantation, pancreatic 359
Transplantation, preoperative evaluation 353
Transplantation, renal 354
Transplantation, small bowel 359
Transposition of great vessels 1219
Transpyloric line 773
Transrectal ultrasound 1025
Transtubercular line 773
Transtuzumab (Herceptin) 585, 591
Transurethral resection of prostate (TURP) 1123
Trauma 143
Trauma, management 143
Trauma, mechanism 145
Traumatic fat necrosis 554
Treponema pallidum 56
TRH stimulation test 489
Triad of CDH 1213
Triad of Ohashi 747
Triad, Tillaux 780
Triage 143
Triage algorithm 143
Triage system, types 143
Triamcinolone, intrakeloidal 11
Triangle of Doom 816, 1269
Triangle of pain 817
Triangle of Petit 835
Trichobezoar 898
Trichoepithelioma 294
Trifoliate duodenum 902
Trigone 1105
Triple anastomosis 745
Triple endoscopy 479
Triple phosphate stones 1097, 1107
Trismus 430
Trivex 248
Trotter's triad 423
Truss 811
T-tube removal methods 693
Tubb's dilator 1220
Tube caecostomy 975
Tube, flatus 1255
Tube, infant feeding 1255
Tube, Kehr's T' 1255
Tube, Ryle's 1254
Tubercle 54
Tuberculosis 54
Tuberculosis spine 1190
Tuberculosis, abdomen 615
Tuberculosis, abdomen, ascitic fluid 619
Tuberculosis, abdomen, barium X-ray 618
Tuberculosis, abdomen, CT scan 619
Tuberculosis, abdomen, treatment 620
Tuberculosis, abdomen, ultrasound 619
Tuberculosis, cervical vertebra 1191
Tuberculosis, ileal 621
Tuberculosis, ileocaecal 616, 663, 781
Tuberculosis, intestinal, acute 615
Tuberculosis, intestinal, chronic 615
Tuberculosis, liver 616
Tuberculosis, pancreas 616
Tuberculosis, peritoneal 615, 621
Tuberculosis, peritoneal, acute 622
Tuberculosis, peritoneal, chronic 622
Tuberculosis, renal 1081
Tuberculosis, spleen 616
Tuberculosis, thoracolumbar vertebra 1191
Tuberculous colitis, diffuse 615
Tuberculous dactylitis 170
Tuberculous lymphadenitis 470
Tuberculous lymphadenitis, clinical types 473
Tuberculous lymphadenitis, stages 472
Tuberculous lymphadenitis, treatment 474
Tuberculous lymphadenitis, types 472
Tuberculous, ano-recto-sigmoidal 626
Tuberculous, drugs, WHO recommendation 626
Tuberculous, mesenteric lymphadenitis 624
Tuberculous, omentum 626
Tuberculous, role of steroids 626
Tumour markers 290, 291
Tumour suppressor gene p53 284
Tumour, Buschke-Lowenstein 1147
Tumour, carotid body 468
Tumour, Cock's peculiar 81
Tumour, desmoid 84, 791
Tumour, dumbbell 441, 468, 1193
Tumour, glomus 82
Tumour, Klatskin 695
Tumour, Lindsay 511
Tumour, phylliodes 552
Tumour, potato 468
Tumour, Raspberry 787
Tumour, staging 285
Tumour, sternomastoid 469
Tumour, sternomastoid, treatment 470
Tumour, superior sulcus 1208
Tumour, turban 306
Tumour, ubiquitous 71
Tumour, universal 71
Tumour, Warthin's 444
Tumours, abdominal wall 791
Tumours, cerebellar vermis 1184
Tumours, chest wall 1208
Tumours, colon 961
Tumours, frontal lobe 1184
Tumours, intracranial 1183
Tumours, jaw 381
Tumours, maxillary 426
Tumours, occipital lobe 1184
Tumours, odontogenic 383
Tumours, parietal lobe 1184
Tumours, pituitary 1186
Tumours, retroperitoneal 779
Tumours, spinal 1193
Tumours, temporal lobe 1184
Tumours, third ventricle 1184
Tumours, tonsil 426
Tunica albuginea 1160
Tunica vaginalis 1160
Turnball's '*no-touch*' technique 972
Tween 80 993
Tylosis 858
Tylosis 858
Typhoid Mary 694
Typhoid perforation 929
Tzanakis scoring system 1011

U

Uereter-golf hole 1082
Uerteric calculus 1094
Uerteric calculus, Dormia basket 1096
Ulcer 14
Ulcer cancer 903
Ulcer, 'girdle' 617
Ulcer, arterial 23
Ulcer, assessment 19
Ulcer, Bairnsdale 23
Ulcer, base 14
Ulcer, Buruli 23
Ulcer, callous 16
Ulcer, Cameron 848
Ulcer, carcinomatous 24
Ulcer, chilblains 23
Ulcer, chronic 20
Ulcer, classification, clinical 15
Ulcer, classification, pathological 16
Ulcer, cleaning 19
Ulcer, clinical examination 29
Ulcer, cortisol 16
Ulcer, Curling's 134, 878
Ulcer, Cushing's 878, 879
Ulcer, debridement 20
Ulcer, decubitus 21, 214
Ulcer, definition 14
Ulcer, diabetic 25
Ulcer, diabetic, problems 26
Ulcer, discharge 18

Ulcer, Douglas 338
Ulcer, dressing 20
Ulcer, Ducrey's 28
Ulcer, duodenal 882
Ulcer, edge 14
Ulcer, edge biopsy 19
Ulcer, edge, everted 14
Ulcer, edge, punched out 14
Ulcer, edge, raised beaded 14
Ulcer, edge, sloping 14
Ulcer, edge, undermined 14
Ulcer, elusive 1110
Ulcer, floor 14
Ulcer, footballer's 21
Ulcer, frostbite 23
Ulcer, gravitational 28, 249
Ulcer, gummatous 28
Ulcer, healing 16
Ulcer, Hunner's 1110
Ulcer, induration 15
Ulcer, investigation 19
Ulcer, ischaemic 23
Ulcer, kissing 883
Ulcer, malignant 16
Ulcer, management 19
Ulcer, margin 14
Ulcer, Marjolin's 13, 24, 28
Ulcer, Martorell's 23
Ulcer, melanotic 24
Ulcer, Meleney's 26
Ulcer, neurogenic 21
Ulcer, neuropathic 21
Ulcer, nonspecific 16
Ulcer, parts 14
Ulcer, penetrating 21
Ulcer, perforating 21
Ulcer, recurrent 896
Ulcer, rodent 24, 303
Ulcer, specific 16
Ulcer, spreading 15
Ulcer, syphilitic 28
Ulcer, traumatic 20
Ulcer, trophic 21, 214
Ulcer, tropical 16
Ulcer, tuberculous 27
Ulcer, venous 28, 249
Ulcer, Wagener's grading 17
Ulcerative colitis, barium enema 956
Ulcerative colitis, sigmoidoscopy 957
Ulcerative colitis, surgeries 958
Ulcerative colitis, treatment 957
Ulcers, in gastrinoma 757
Ulnar paradox 278, 279
Ultrasound 1238
Ultrasound, A-mode 1238
Ultrasound, B-mode 1238
Ultrasound, M-mode 1238
Umbilectomy 830
Umbilical adenoma 786
Umbilical granuloma 786
Umbilical sinus 786, 787
Umbilicus, diseases 784
Umbolith 787
University of Wisconsin solution 353
Unna boots 242, 250, 251
Upper GI bleed 1061, 1062
Urachal anomalies 1106
Urachal cyst 788, 1106
Urachal diverticulum 788
Urachal fistula 788
Urachal sinus 788, 1106
Urachus, patent 788, 1106
Urban rabies 58
Ureter, anatomy 1066
Ureter, golf hole 1082
Ureter, retrocaval 1079
Ureter, sites of constrictions 1066
Ureteric line 1067
Ureteric meatotomy 1078, 1096
Ureterocele 1080
Ureterocele, Stephen's classification 1080
Ureterorenoscoic stone (URS) removal 1096
Ureterosigmoidostomy 1115
Ureterostomy, Davis-T-tube 1088
Urethra, anatomy 1129
Urethral calculus 1134
Urethral dilatation, complications 1132
Urethritis 1135
Urethroplasty, Johanson's 1132
Urethroplasty, problems in 1132
Urethrostomy, perineal 1145
Urethrotomy, external 1132
Urethrotomy, visual internal cystoscopic 1132
Urinary bilharziasis 1111
Urinary bladder, anatomy 1105
Urinary bladder, sphincter 1105
Urinary diversion 1119
Urinary diversion, complication 1120
Urinary diversion, permanent 1120
Urinary diversion, temporary 1119
Urinary diversion, types 1120
Urine, extravasation 1135
Urine, retention 1136
Urine, retention, management 1136
Urograffin 1067
Urokinase 199
Uterine mass 781

V

VAAFT procedure 1055
VACTER anomalies 857
Vacuum assisted closure 22
Vagotomy, complications 921
Vagotomy, highly selective 870, 884, 921
Vagotomy, parietal selectve 921
Vagotomy, posterior truncal 885
Vagotomy, selective 870, 885, 921
Vagotomy, super selective 921
Vagotomy, truncal 870, 885, 886, 920
Vagotomy, with antrectomy 885
Valve of Geralch 1006
Valve prosthesis, Bjork-Shiley tilting disc 1222
Valve prosthesis, Starr-Edwards ball 1222
Valve replacement surgery 1222
Valves of Houston 1020
van den Bergh's test 628
Varices, reticular 235
Varicocele 1157
Varicocele surgery, Inavissevich approach 1159
Varicocele surgery, laparoscopic approach 1159
Varicocele surgery, Marc-Goldstein approach 1159
Varicocele, grading 1158
Varicocele, infertility 1157
Varicocele, irreducible 1157
Varicocele, primary 1158
Varicocele, secondary 1158
Varicography 242
Varicose vein, stripping 244
Varicose vein, subfascial endoscopic perforator ligation 247
Varicose vein, subfascial ligation of Cockett and Dodd 246
Varicose veins 235
Varicose veins surgery, problems 249
Varicose veins, aetiology 237
Varicose veins, CEAP classification 236
Varicose veins, classifications 235
Varicose veins, complications 240
Varicose veins, compression bandages 251
Varicose veins, compression therapy 251
Varicose veins, drugs 242
Varicose veins, endovenous laser ablation 248
Varicose veins, familial 238
Varicose veins, ligation of short saphenous vein 247
Varicose veins, Linton's approach 247
Varicose veins, pathogenesis 236
Varicose veins, radiofrequency ablation 247
Varicose veins, sclerotherapy 242
Vas aberrans of Haller 1155
Vascular anomalies 221
Vascular malformation 221, 225
Vasectomy 1260
Vasopressin 656
VATER anomaly 1035
'Veins of conduit' 230
Veins of lower limb, anatomy 230
Venogram 233, 241
Venography, ascending 241
Venography, descending 241
Venous blood flow, lower limb, physiology 231
Venous disabilty scoring system 240
Venous diseases 230
Venous Doppler 240, 241
Venous flares 235
Venous haemodynamic mapping 241
Venous refilling time 242
Venous segmental disease score 240
Venous ulcer, complications 251
Verocay bodies 92
Veronesi QUART therapy 582
Verruca necrogenica 84
Verruca plantaris 84
Verrucous carcinoma, cheek 398

Vertical banded gastroplasty 107
Vesical calculus, females 1108
Vesical piles 1122
Vesicostomy 977
Vesicoureteric reflux 1069, 1117
Vicryl 1256
Video assisted thoracoscopic surgery (VATS) 1214
Vigorous achalasia 850
Villous adenoma 936
Vim-Silverman needle 629
Vin rose patch 225
Vincent's angina 391
Vipoma 536, 758
Virchow's triad 231
Visible gastric peristalsis 885
Visick grading 896
Visor approach for oral malignancy 406
Vitellointestinal duct, anomalies 786
VNUS closure 247
Voice change 529
Volkmann's sign 169
Volvulus 997
Volvulus, caecal 997
Volvulus, compound 998
Volvulus, gastric 900
Volvulus, pelvic colon 997
Volvulus, sigmoid 997
von Eber's gland 454
Von Meyenburg's complexes 643
Von Recklinghausen's disease 91, 92, 542
von Willebrand's disease 129
Von-Hippel-Lindau disease 1100
Voval cord position 529

W

'W' flap plasty 412
Wagner-Grossman hypothesis 529
Waldeyer's lymphatic ring 455
Wallace's rule of nine 130
Waltzing 347
Wangenstein's invertogram 1035
Wardrop's procedure 203
Warfarin 253
Warthin's starry silver stain 876
Warts 84
Warts, Butcher's 84
Warts, plantar 84
Warts, senile 84
Water can perineum 1135
Water excess 95
Water intoxication 95
Water loss 95
Water shed point 959
Wayne's diagnostic indices 502
Weber-Fergusson incision 427
Weigert Meyer law 1078
Weir-Mitchell's skin 276
Wen 79
Werner's T_3 suppression test 489
Whipple's procedure 743, 744, 753
Whipple's triad 756
White bile 705, 755
White cell trapping theory 236
Wide excision 94
Wild life rabies 58
Wiley's eversion endarterectomy 191
Wilm's tumour 1099
Wilm's tumour, staging 1099
Winging of scapula 281
Witch's milk 558, 560
Wlikinson method 811
Wolf-Chaikoff effect 499
Wound classification 5
Wound failure, acute 788
Wound healing 5
Wound healing, factors affecting 7
Wound healing, phases 6
Wound healing, primary 5
Wound healing, problems 13
Wound healing, secondary 5
Wound healing, stages 6
Wound healing, tertiary 6
Wound healing, types 5
Wound suturing, principles 8
Wound, closed 5
Wound, debridement 7, 8
Wound, excision 8
Wound, open 5
Wound, radical excision 8
Wound, toilet 8
Wounds, management 7
Wrist drop 280
Wry neck 469
Wuchereria bancrofti 39
Wuchereria bancrofti, life cycle 259

X

Xenograft 340
Xenon 133 isotope test 182
Xeroderma pigmentosa 298
Xeromammography 548
Xerostomia 434
Xylocaine 1235

Y

Yaws 29

Z

'Z' plasty 346
Zenker's diverticulum 981
Ziehl-Neelsen stain 619, 626
Zip technique 604
Zoladex 585
Zuckerkandl tubercle 523
Zuska-Atkins disease 563

Latha Ganti
Editor

Atlas of Emergency Medicine Procedures

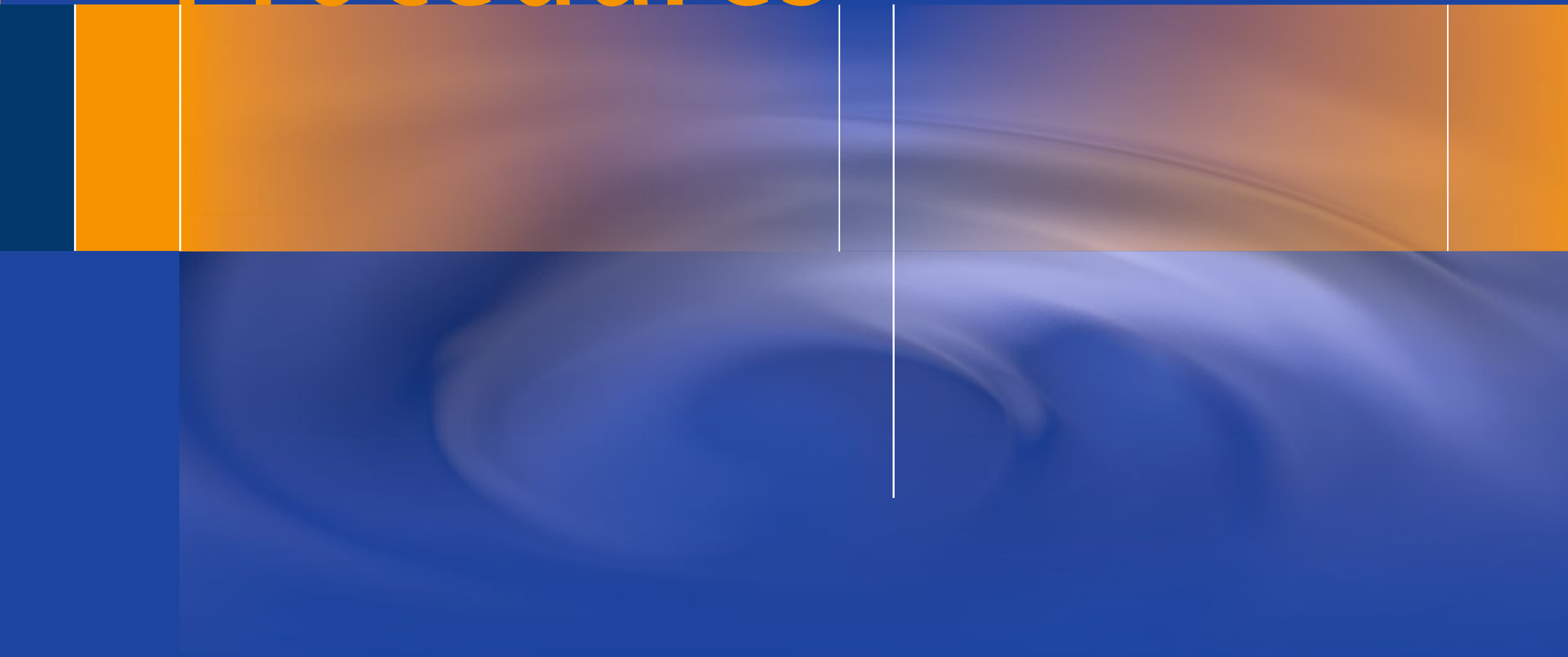

Springer

Atlas of Emergency Medicine Procedures

Latha Ganti
Editor

Atlas of Emergency Medicine Procedures

Springer

Editor
Latha Ganti
Professor of Emergency Medicine
University of Central Florida
Director, SE Specialty Care Centers of Innovation
Orlando Veterans Affairs Medical Center
Orlando, FL
USA

ISBN 978-1-4939-2506-3 ISBN 978-1-4939-2507-0 (eBook)
DOI 10.1007/978-1-4939-2507-0

Library of Congress Control Number: 2016930028

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

To my father, my role model and career coach –

Dr. Ganti L. Rao

Dad, I wouldn't have had the career I do were it
not for you. Your unconditional belief in me, your confidence in
my success, your never ending patience and generosity – you
lead a life of great achievements with utmost humility.
It is my privilege to be your daughter.

Preface

The *Atlas of Emergency Medicine Procedures* is presented in the spirit of “see one, do one, teach one” and “a picture is worth a thousand words.” It can be used at the bedside, both by novice practitioners and seasoned clinicians as a teaching aid. For ease of reference, the most common procedures are grouped by organ systems. Each procedure follows a standardized format, beginning with keywords, a definition if appropriate, indications, and contraindications. These are followed by materials and medications, often accompanied by a photograph of the specific equipment or setup. The procedure itself is numbered rather than bulleted, highlighting the sequence of steps. Photographs are placed where the relevant information is encountered in the text rather than at the end. Every procedure also lists complications. Finally, there is a section on pearls and pitfalls, gleaned from the collective experience of the contributors in addition to traditional teachings.

Emergency medicine, by nature, is a field best suited to a visually appealing, concise text able to deliver the information required. Much effort therefore has been dedicated to images. These are either actual photographs taken of the procedure steps or specially commissioned drawings by Springer’s professional illustration team. A work of this magnitude could not have been accomplished without the expertise of my dedicated Springer team. A huge thank you to Shelly Reinhardt, Megan Ruzomberka, and Lee Klein. It is hoped that readers will find this atlas useful and practical. Feedback and suggestions for future editions are welcomed and can be sent to AtlasEMprocedures@gmail.com.

Orlando, FL, USA Latha Ganti

Contents

Part I Vascular Procedures 1

1 **Arterial Cannulation (Radial and Femoral)** 3
Jeffrey Kile, Katrina John, and Amish Aghera
1.1 Indications 3
1.2 Contraindications 3
1.3 Materials and Medications 3
1.4 Complications 10
1.5 Pearls and Pitfalls 10
Selected Reading 11

2 **Ultrasound-Guided Peripheral Intravenous Access** 13
Coben Thorn and L. Connor Nickels
2.1 Indications 13
2.2 Contraindications 13
2.3 Materials and Medications 13
2.4 Procedure 14
2.5 Complications 15
2.6 Pearls and Pitfalls 15
Selected Reading 15

3 **Central Venous Line Placement: Internal Jugular Vein, Subclavian Vein, and Femoral Vein** 17
Kevin D. Ergle, Zachary B. Kramer, Jason Jones, and Rohit Pravin Patel
3.1 Indications 17
3.2 Contraindications 17
3.3 Materials and Medications 17
3.4 Procedures 18
3.5 Complications 25
3.6 Ultrasound Guided Cannulation: Tips for Each Approach 26
3.7 Removing a Central Line 26
Selected Reading 26

4 **Pulmonary Artery Catheter** 29
Rohit Pravin Patel and Marie-Carmelle Elie
4.1 Indications 29
4.2 Contraindications 29
4.3 Materials 29
4.4 Procedure 30
4.5 Complications 33
4.6 Pearls and Pitfalls 33
Selected Reading 33

5 Noninvasive Cardiac Monitoring: The Edwards Vigileo System 35
Dawood G. Dalaly and Rohit Pravin Patel
5.1 Indications 35
5.2 Contraindications 37
5.3 Definitions and Values 37
5.4 Materials 38
5.5 Procedure 38
5.6 Pearls and Pitfalls 38
Selected Reading 38

6 Peripheral Venous Cutdown 39
Jeffrey Kile, Katrina John, and Amish Aghera
6.1 Indications 39
6.2 Contraindications 39
6.3 Materials and Medications 39
6.4 Choice of Vessel for Cutdown 41
6.5 Procedure 41
6.6 Complications 44
6.7 Pearls and Pitfalls 44
Selected Reading 44

Part II Airway and Vascular Procedures 45

7 Bag-Valve-Mask Ventilation 47
Braden Hexom and Tatiana Havryliuk
7.1 Indications 47
7.2 Contraindications 47
7.3 Materials 47
7.4 Procedure 48
7.5 Complications 50
7.6 Pearls and Pitfalls 50
Selected Reading 51

8 Awake Orotracheal Intubation 53
Benjamin M. Mahon, Justin Bennett, and Lars K. Beattie
8.1 Indications 53
8.2 Contraindications 53
8.3 Materials and Medications 53
8.4 Procedure 53
8.5 Pearls and Pitfalls 56
Selected Reading 56

9 Rapid-Sequence Intubation 57
Ram A. Parekh
9.1 Indications 57
9.2 Contraindications 57
9.3 Materials and Medications 57
9.4 Procedure 60
9.5 Complications 65
9.6 Pearls and Pitfalls 65
Selected Reading 65

10 Direct Laryngoscopy 67
Bharath Chakravarthy and Weston Seipp
10.1 Indications 67
10.2 Contraindications 67

10.3 Materials and Medications 68
10.4 Procedure 69
10.5 Complications 74
10.6 Pearls and Pitfalls 74
References 74

11 Laryngeal Mask Airway 75
Sohan Parekh
11.1 Indications 75
11.2 Contraindications 75
11.3 Types 76
11.4 Materials and Medications 77
11.5 Procedure 77
11.6 Complications 83
11.7 Pearls and Pitfalls 83
Selected Reading 83

12 Combitube 85
Clint Masterson
12.1 Indications 85
12.2 Contraindications 85
12.3 Materials 85
12.4 Procedure 86
12.5 Pearls and Pitfalls 87
Selected Reading 87

13 Assessment of the Difficult Airway 89
Melinda W. Fernandez and Lars K. Beattie
13.1 Indications 89
13.2 Materials and Medications 89
13.3 Procedure 89
13.4 Pearls and Pitfalls 92
References 92
Selected Reading 92

14 GlideScope 93
Sohan Parekh
14.1 Indications 93
14.2 Contraindications 93
14.3 Materials and Medications 93
14.4 Procedure 95
14.5 Complications 96
14.6 Pearls and Pitfalls 96
Selected Reading 97

15 Endotracheal Tube Introducer (Bougie) 99
Joseph Rabinovich
15.1 Indications 99
15.2 Contraindications 99
15.3 Materials and Medications 99
15.4 Procedure 100
15.5 Complications 101
15.6 Pearls and Pitfalls 101
References 101

16 Lighted Stylet Intubation 103
Benjamin M. Mahon and Lars K. Beattie
16.1 Indications 103
16.2 Contraindications 103
16.3 Materials and Medications 103
16.4 Procedure 104
16.5 Pearls and Pitfalls 107
References 107
Selected Reading 107

17 Fiber-Optic Stylet Intubation (Rigid and Semirigid) 109
Joseph Rabinovich
17.1 Indications 109
17.2 Contraindications 109
17.3 Relative Contraindications 109
17.4 Materials and Medications 109
17.5 Procedure 110
17.6 Pearls and Pitfalls 113
References 113

18 Storz Video Laryngoscope 115
Joseph Rabinovich
18.1 Indications 115
18.2 Contraindications 115
18.3 Materials and Medications 116
18.4 Procedure 116
18.5 Complications (As with any Direct Laryngoscopy) 118
18.6 Pearls and Pitfalls 118
References 118
Selected Reading 118

19 Cricothyroidotomy 119
Henry Young II, Shannon Toohey, Bharath Chakravarthy, and Lars K. Beattie
19.1 Indications 119
19.2 Contraindications 119
19.3 Techniques 119
19.4 Complications 126
19.5 Pearls 126
References 126
Selected Reading 126

20 Tracheostomy Tube Malfunction 127
Deena Bengiamin and Bharath Chakravarthy
20.1 Indications 127
20.2 Materials and Medications 127
20.3 Procedure 127
20.4 Pearls and Pitfalls 131
Selected Reading 132

21 Percutaneous Transtracheal Jet Ventilation 133
Clint Masterson
21.1 Indications 133
21.2 Contraindications 133
21.3 Materials and Medications 133

21.4 Procedure . . . 135
21.5 Complications . . . 139
21.6 Pearls and Pitfalls . . . 139
Selected Reading . . . 139

Part III Thoracic Procedures . . . 141

22 Needle Thoracostomy . . . 143
Lucas McArthur and Christian Fromm
22.1 Indications . . . 143
22.2 Contraindications . . . 143
22.3 Materials . . . 143
22.4 Procedure . . . 144
22.5 Complications . . . 147
22.6 Pearls and Pitfalls . . . 147
Selected Reading . . . 147

23 Chest Tube Thoracostomy . . . 149
Brandon R. Allen and Latha Ganti
23.1 Indications . . . 149
23.2 Contraindications . . . 149
23.3 Materials and Medications . . . 149
23.4 Procedure . . . 151
23.5 Complications . . . 153
23.6 Pearls and Pitfalls . . . 153
Selected Reading . . . 153

24 Thoracentesis . . . 155
Lee Richard Donner and Michael Anana
24.1 Indications . . . 155
24.2 Contraindications . . . 155
24.3 Materials and Medications . . . 155
24.4 Procedure . . . 156
24.5 Complications . . . 158
24.6 Considerations . . . 158
24.7 Pearls and Pitfalls . . . 158
Selected Reading . . . 158

25 Open Chest Wounds and Flail Chest . . . 159
Jacob J. Glaser and Carlos J. Rodriguez
25.1 Background . . . 159
25.2 Initial Evaluation of Thoracic Trauma . . . 161
25.3 Open Pneumothorax (“Sucking Chest Wound”) . . . 161
25.4 Flail Chest . . . 162
References . . . 162

26 Emergent Resuscitative Thoracotomy, Open Cardiac Massage, and Aortic Occlusion . . . 163
Kevin M. Jones and Jay Menaker
26.1 Indications . . . 163
26.2 Contraindications . . . 163
26.3 Materials and Medications . . . 163
26.4 Procedure . . . 165
26.5 Complications . . . 169
Selected Reading . . . 169

27 Lung Ultrasonography 171
Ali H. Dabaja, L. Connor Nickels, and Rohit Pravin Patel
27.1 Indications 171
27.2 Contraindications 171
27.3 Materials 171
27.4 Preparations 171
27.5 Procedure 172
Selected Reading 176

Part IV Cardiac Procedures 177

28 Repair of Cardiac Injuries 179
Ronald Tesoriero
28.1 Indications 179
28.2 Contraindications 179
28.3 Materials and Medications 179
28.4 Procedure 181
28.5 Complications 183
28.6 Pearls and Pitfalls 183
Selected Reading 183

29 Synchronized Electrical Cardioversion 185
Jason Jones, Ann Tsung, and Marie-Carmelle Elie
29.1 Indications 185
29.2 Contraindications 185
29.3 Materials and Medications 185
29.4 Procedure 185
29.5 Complications 188
29.6 Pearls and Pitfalls 188
Selected Reading 188

30 Unsynchronized Cardioversion (Defibrillation) 189
Matthew R. Tice, Zachary B. Kramer, and Marie-Carmelle Elie
30.1 Indications 189
30.2 Contraindications 189
30.3 Materials and Medications 189
30.4 Procedure 190
30.5 Complications 192
30.6 Pearls and Pitfalls 192
Selected Reading 192

31 Transcutaneous Pacing 193
Nour Rifai and Christian Coletti
31.1 Indications 193
31.2 Contraindications 193
31.3 Materials and Medications 193
31.4 Procedure 194
31.5 Pearls and Pitfalls 196
Selected Reading 196

32 Transvenous Cardiac Pacing 197
Katrina John, Jeffrey Kile, and Amish Aghera
32.1 Indications 197
32.2 Contraindications 197
32.3 Materials and Medications 198
32.4 Procedure 199

32.5 Complications . . . 202
32.6 Pearls and Pitfalls . . . 202
Selected Reading . . . 202

33 Pericardiocentesis (Optional: Ultrasound Guidance) . . . 203
Katrina Skoog Nguyen, L. Connor Nickels, and Rohit Pravin Patel
33.1 Indications . . . 203
33.2 Contraindications . . . 203
33.3 Materials and Medications . . . 203
33.4 Procedure . . . 203
33.5 Complications . . . 205
33.6 Pearls and Pitfalls . . . 205
References . . . 205

34 Ultrasound Evaluation of Pulmonary Embolism and Heart Strain . . . 207
Giuliano De Portu, L. Connor Nickels, and Marie-Carmelle Elie
34.1 Indications . . . 207
34.2 Contraindications . . . 207
34.3 Materials and Medications . . . 207
34.4 Procedure . . . 207
34.5 Findings . . . 209
34.6 Complications . . . 209
34.7 Pearls and Pitfalls . . . 209
Selected Reading . . . 209

35 Pacemaker Evaluation in the Emergency Department . . . 211
Joseph D. Romano and Christian Coletti
35.1 Pacemaker Function . . . 211
35.2 Common Pacemaker Codes . . . 213
35.3 Pacemaker Malfunction . . . 213
35.4 Initial Evaluation of the Patient with a Pacemaker . . . 214
35.5 Interpreting the Type of Malfunction Based on ECG . . . 214
35.6 Management of Pacemaker-Mediated Tachycardia . . . 214
References . . . 215

Part V Spine Evaluation . . . 217

36 Cervical Collar Placement . . . 219
Justin Bennett and Lars K. Beattie
36.1 Indications . . . 219
36.2 Contraindications . . . 219
36.3 Materials and Medications . . . 219
36.4 Procedure (Aspen Collar) . . . 220
36.5 Pearls and Pitfalls . . . 222
Selected Reading . . . 222

37 Cervical Spine Clearance . . . 223
Braden Hexom and Tatiana Havryliuk
37.1 Indications . . . 223
37.2 Contraindications . . . 223
37.3 National Emergency X-Radiography Utilization Study (NEXUS) and Canadian C-Spine Rules (CCR) to Assess the Need for Imaging . . . 223
37.4 Type of Imaging . . . 224
37.5 Materials . . . 224
37.6 Procedure . . . 225

37.7 Complications 227
37.8 Pearls and Pitfalls. 227
Reference 227
Selected Reading. 227

38 Log Roll 229
Justin Bennett and Lars K. Beattie
38.1 Indications 229
38.2 Contraindications 229
38.3 Materials and Medications. 229
38.4 Procedure 229
38.5 Pearls and Pitfalls. 232
Selected Reading. 232

Part VI Neurologic and Neurosurgical Procedures 233

39 Burr Hole Craniotomy 235
Latha Ganti
39.1 Indications 235
39.2 Contraindications 235
39.3 Materials and Medications. 235
39.4 Procedure 236
39.5 Complications 240
39.6 Pearls and Pitfalls. 240
References. 240
Selected Reading. 240

40 External Ventricular Drain Placement 241
Latha Ganti
40.1 Indications 241
40.2 Contraindications 241
40.3 Materials and Medications. 241
40.4 Procedure 241
40.5 Complications 244
40.6 Pearls and Pitfalls. 244
References. 245

41 Lumbar Puncture in Adults. 247
Kevin Tench, L. Connor Nickels, and Rohit Pravin Patel
41.1 Indications 247
41.2 Contraindications 247
41.3 Materials and Medications. 247
41.4 Procedure 248
41.5 Complications 250
41.6 Pearls and Pitfalls. 250
Selected Reading. 250

42 Reflex Eye Movements (Doll’s Eyes and Caloric Testing). 251
Thomas T. Nguyen, Tina Dulani, and Saadia Akhtar
42.1 Doll’s Eyes (Oculocephalic Reflex Testing) 251
42.2 Caloric Reflex Testing (Vestibuloocular Reflex) 253
Selected Reading. 255

43 Dix-Hallpike Maneuver 257
Rui Domingues and Muhammad Waseem
43.1 Indications 257
43.2 Contraindications 257
43.3 Materials 257

43.4 Procedure 257
43.5 Complications 258
43.6 Pearls and Pitfalls. 258
Selected Reading. 258

44 Epley Maneuver for Vertigo (Particle Repositioning or Canalith Repositioning Procedure) 259
Rui Domingues and Muhammad Waseem
44.1 Indications 259
44.2 Contraindications. 259
44.3 Materials 259
44.4 Procedure 259
44.5 Complications 260
44.6 Pearls and Pitfalls. 260
Selected Reading. 260

45 Clinical Brain Death Examination in Adults 261
Eric S. Papierniak, Hassan M. Alnuaimat, Tracy Timmons, and Deborah M. Stein
45.1 Indications 261
45.2 Contraindications. 261
45.3 Materials and Medications. 261
45.4 Procedure 261
45.5 Pearls and Pitfalls. 267
45.6 Controversies 267
Selected Reading. 268

Part VII Ophthalmic Procedures 269

46 Slit Lamp Examination. 271
Bobby K. Desai
46.1 Indications 271
46.2 Contraindications. 272
46.3 Materials 273
46.4 Procedure 274
46.5 Complications 276
46.6 Pitfalls 276
46.7 Fluorescein Testing 276
Selected Reading. 277

47 Eye Irrigation 279
Bobby K. Desai
47.1 Indications 279
47.2 Contraindications. 279
47.3 Materials and Medications. 279
47.4 Procedure 281
47.5 Complications 282
47.6 Pearls and Pitfalls. 282
Selected Reading. 282

48 Corneal Foreign Body Removal. 283
Bobby K. Desai
48.1 Indications 283
48.2 Contraindications. 283
48.3 Materials and Medications. 284
48.4 Procedure 285
48.5 Complications 285
Selected Reading. 285

49 Ultrasound Evaluation of Retinal Detachment 287
Shalu S. Patel, L. Connor Nickels, and Rohit Pravin Patel
49.1 Indications 287
49.2 Contraindications 287
49.3 Materials and Medications 287
49.4 Procedure 287
49.5 Complications 289
49.6 Pearls and Pitfalls 289
Selected Reading 289

50 Ultrasonography in the Evaluation of Intraocular Pathology 291
Benjamin M. Mahon, Marie-Carmelle Elie, L. Connor Nickels,
and Rohit Pravin Patel
50.1 Indications 291
50.2 Contraindications 291
50.3 Materials 291
50.4 Procedure 291
50.5 Complications 291
50.6 Pearls and Pitfalls 291
50.7 Specific Disease Processes/Pathology 292
References 295
Selected Reading 295

51 Tonometry 297
Bobby K. Desai
51.1 Indications 297
51.2 Contraindications 297
51.3 Overview 297
51.4 Schiøtz Tonometry 299
51.5 Tono-Pen Tonometry 300
51.6 Complications 303
Selected Reading 303

52 Lateral Canthotomy 305
Benjamin M. Mahon and Bobby K. Desai
52.1 Indications 305
52.2 Contraindications 306
52.3 Materials and Medications 306
52.4 Procedure 306
52.5 Complications 308
52.6 Pearls and Pitfalls 308
Selected Reading 308

Part VIII Otorhinolaryngologic Procedures 309

53 Epistaxis Control 311
Benjamin M. Mahon and Bobby K. Desai
53.1 Etiology 311
53.2 Indications 313
53.3 Contraindications 313
53.4 Materials and Medications 313
53.5 Procedure: Anterior Epistaxis 313
53.6 Procedure: Posterior Epistaxis 317
Selected Reading 319

54 Treatment of Septal Hematoma . . . 321
Bobby K. Desai
54.1 Indications . . . 321
54.2 Contraindications . . . 322
54.3 Materials and Medications . . . 322
54.4 Procedure . . . 322
54.5 Pitfalls and Complications . . . 324
Selected Reading . . . 324

55 Nasal Foreign Body Removal . . . 325
Bobby K. Desai
55.1 Indications . . . 325
55.2 Contraindications: Need Urgent Referral . . . 325
55.3 Materials and Medications . . . 326
55.4 Procedures (Based on the Method Chosen) . . . 328
Selected Reading . . . 329

56 Cerumen Removal . . . 331
Bobby K. Desai
56.1 Indications . . . 331
56.2 Contraindications . . . 331
56.3 Materials and Medications . . . 331
56.4 Procedure . . . 334
56.5 Pitfalls and Complications . . . 334
Selected Reading . . . 334

57 Ear Foreign Body Removal . . . 335
Bobby K. Desai
57.1 Indications . . . 335
57.2 Contraindications: Need Urgent Referral . . . 335
57.3 Materials and Medications . . . 335
57.4 Procedures (Based on the Method Chosen) . . . 337
Selected Reading . . . 338

58 Treatment of Auricular Hematoma . . . 339
Bobby K. Desai
58.1 Indications . . . 339
58.2 Contraindications . . . 340
58.3 Materials and Medications . . . 340
58.4 Procedure (Needle Aspiration) . . . 340
58.5 Pitfalls and Complications . . . 346
Selected Reading . . . 346

59 Incision and Drainage of Peritonsillar Abscess . . . 347
Melinda W. Fernandez and Bobby K. Desai
59.1 Indications . . . 347
59.2 Contraindications . . . 347
59.3 Materials and Medications . . . 347
59.4 Procedure: Aspiration . . . 347
59.5 Procedure: Incision and Drainage . . . 349
59.6 Complications . . . 350
59.7 Pearls and Pitfalls . . . 350
Selected Reading . . . 350

60 Incision and Drainage of Sublingual Abscess 351
Melinda W. Fernandez and Bobby K. Desai
60.1 Indications 351
60.2 Contraindications 352
60.3 Materials and Medications 352
60.4 Procedure: Incision and Drainage 353
60.5 Complications (Generally Minimal) 354
60.6 Pearls and Pitfalls 354
Selected Reading 354

61 Incision and Drainage of Parotid Duct Abscess 355
Melinda W. Fernandez and Bobby K. Desai
61.1 Indications 355
61.2 Contraindications 355
61.3 Materials and Medications 355
61.4 Procedure 357
61.5 Complications 358
61.6 Pearls and Pitfalls 358
Selected Reading 358

Part IX Dental Procedures 359

62 Techniques of Mandibular Anesthesia 361
Susana Perry, Joshua Perry, and Rosalia Rey
62.1 Inferior Alveolar Nerve Block 361
62.2 Buccal Nerve Block 365
62.3 Mental Nerve Block 366
Selected Reading 368

63 Reduction of Dislocated Temporomandibular Joint 369
Christopher J. Spencer and Geraldine Weinstein
63.1 Indications 369
63.2 Contraindications for Closed Reduction 369
63.3 Materials and Medications 369
63.4 Procedure 370
63.5 Complications 372
63.6 Pearls 372
Selected Reading 372

64 Dry Socket (Alveolar Osteitis, Fibrinolytic Osteitis) 373
Michael A. Abraham, Amir Azari, Jennifer Westcott, and Franci Stavropoulos
64.1 Indications 373
64.2 Contraindications 374
64.3 Materials and Medications 374
64.4 Procedure 374
64.5 Complications 374
64.6 Pearls and Pitfalls 374
Selected Reading 374

65 Postextraction Hemorrhage 375
Michael A. Abraham, Amir Azari, Jennifer Westcott, and Franci Stavropoulos
65.1 Indications 375
65.2 Contraindications 375
65.3 Materials and Medications 375
65.4 Procedure 375
65.5 Complications 376

65.6 Pearls and Pitfalls 376
Selected Reading 376

66 Fractured Tooth 377
Geraldine Weinstein
66.1 Indications 377
66.2 Methods of Sustaining Dental Fracture 377
66.3 Four Types of Fractured Tooth 378
66.4 Contraindications 379
66.5 Materials and Medications 379
66.6 Procedure 379
66.7 Complications 380
66.8 Pearls 380

67 Dental Avulsion Management 381
Laura Tucker and Abimbola O. Adewumi
67.1 Indications 381
67.2 Contraindications 382
67.3 Materials and Medications 382
67.4 Procedure 383
67.5 Complications 384
67.6 Pearls and Pitfalls 384
Selected Reading 385

Part X Gastrointestinal Procedures 387

68 Excision of Thrombosed External Hemorrhoid 389
Latha Ganti
68.1 Indications 389
68.2 Contraindications 389
68.3 Materials and Medications 389
68.4 Procedure 390
68.5 Complications 392
68.6 Pearls and Pitfalls 392
References 392
Selected Reading 392

69 Diagnostic Peritoneal Lavage 393
Latha Ganti
69.1 Indications 393
69.2 Contraindications 394
69.3 Materials and Medications 394
69.4 Procedure 394
69.5 Complications 397
69.6 Pearls and Pitfalls 397
References 397
Selected Reading 397

70 Manual Reduction of Abdominal Hernia 399
Latha Ganti
70.1 Indications 401
70.2 Contraindications 401
70.3 Materials and Medications 401
70.4 Procedure 401
70.5 Complications 402
70.6 Pearls and Pitfalls 402
Selected Reading 402

71 Extended Focused Assessment with Sonography for Trauma 403
Coben Thorn and L. Connor Nickels
71.1 Indications 403
71.2 Contraindications 403
71.3 Materials and Medications 403
71.4 Procedure 404
71.5 Complications 409
71.6 Pearls and Pitfalls 409
Selected Reading 410

72 Nasogastric Tube Placement 411
David P. Nguyen, L. Connor Nickels, and Giuliano De Portu
72.1 Indications 411
72.2 Contraindications 411
72.3 Materials and Medications 411
72.4 Procedure 412
72.5 Complications 413
72.6 Pearls 413
Selected Reading 413

73 Esophageal Foreign Body Removal 415
David P. Nguyen, L. Connor Nickels, and Rohit Pravin Patel
73.1 Indications 415
73.2 Contraindications 415
73.3 Materials 415
73.4 Procedure 416
73.5 Complications 416
73.6 Pearls 416
Selected Reading 417

74 Activated Charcoal 419
Deylin I. Negron Smida and Judith K. Lucas
74.1 Indications 419
74.2 Contraindications 420
74.3 Materials and Medications 420
74.4 Procedure 421
74.5 Complications 421
74.6 Pearls and Pitfalls 421
Selected Reading 422

75 Gastric Lavage 423
Deylin I. Negron Smida and Judith K. Lucas
75.1 Indications 423
75.2 Contraindications 423
75.3 Materials and Medications 423
75.4 Procedure 424
75.5 Complications 425
75.6 Pearls and Pitfalls 425
Selected Reading 426

76 Whole-Bowel Irrigation 427
Judith K. Lucas
76.1 Indications 427
76.2 Contraindications 427
76.3 Materials and Medications 427

76.4 Procedure 429
76.5 Complications 429
76.6 Pearls 429
Selected Reading 430

77 Sengstaken-Blakemore Tube 431
Thomas T. Nguyen, Etan Eitches, and Stephanie Wetmore-Nguyen
77.1 Indications 431
77.2 Contraindications 431
77.3 Materials and Medications 431
77.4 Procedure 434
77.5 Complications 437
77.6 Pearls and Pitfalls 437
Selected Reading 437

78 Gastrostomy Tube Placement 439
Nathaniel Lisenbee and Latha Ganti
78.1 Indications 443
78.2 Contraindications 443
78.3 Materials and Medications 443
78.4 Procedure (PEG Tube Replacement) 444
78.5 Pearls and Pitfalls 446
78.6 Complications 446
Selected Reading 446

79 Paracentesis 447
Shalu S. Patel and Bobby K. Desai
79.1 Indications 447
79.2 Contraindications 447
79.3 Materials and Medications 447
79.4 Procedure 447
79.5 Complications 449
79.6 Pearls and Pitfalls 449
Selected Reading 449

80 Anal Fissure Management 451
David P. Nguyen, L. Connor Nickels, and Giuliano De Portu
80.1 Indications 451
80.2 Contraindications 453
80.3 Materials and Medications 453
80.4 Procedure 453
80.5 Complications 453
80.6 Pearls and Pitfalls 453
Selected Reading 453

Part XI Genitourinary Procedures 455

81 Bladder Catheterization 457
Maritza A. Plaza-Verduin and Judith K. Lucas
81.1 Indications 457
81.2 Contraindications 457
81.3 Materials and Medications 457
81.4 Procedure 458
81.5 Complications 461
81.6 Pearls and Pitfalls 461
Selected Reading 461

82 Pelvic Examination and Wet Preparation 463
Nauman W. Rashid, Elaine B. Josephson, and Muhammad Waseem
82.1 Indications 463
82.2 Contraindications 463
82.3 Materials and Medications 463
82.4 Procedure 463
82.5 Complications 466
82.6 Pearls and Pitfalls 466
Selected Reading 467

83 Bartholin Gland Abscess/Cysts Drainage 469
Holly H. Charleton, Marylin Otero, Diane F. Giorgi, and Joseph A. Tyndall
83.1 Indications 469
83.2 Contraindications 469
83.3 Procedure Types 469
83.4 Materials 470
83.5 Procedures 470
83.6 Iodoform Packing 472
83.7 Aftercare 474
Selected Reading 474

84 Sexual Assault Forensic Examination 475
Rajnish Jaiswal, Mary T. Ryan, and Muhammad Waseem
84.1 Indications 475
84.2 Contraindications 475
84.3 Materials and Medications 475
84.4 Procedure “Prepare the Patient, Prepare the Room” 476
84.5 Pearls and Pitfalls 481
References 481

85 Treatment of Priapism 483
Jeffrey Kile, Katrina John, and Amish Aghera
85.1 Indications 483
85.2 Contraindications 483
85.3 Materials and Medications 483
85.4 Noninvasive Therapy Preprocedure 485
85.5 Procedures 485
85.6 Complications 491
85.7 Pearls and Pitfalls 491
85.8 Considerations 492
Selected Reading 493

86 Reduction of Phimosis/Paraphimosis 495
Justin Chen and Muhammad Waseem
86.1 Indications 495
86.2 Contraindications 495
86.3 Materials and Medications 495
86.4 Procedure: Manual Reduction for Paraphimosis 496
86.5 Complications 499
86.6 Pearls and Pitfalls 499
Selected Reading 499

87 Manual Testicular Detorsion 501
Brandon R. Allen and L. Connor Nickels
87.1 Indications 501

87.2 Contraindications 503
87.3 Materials and Medications 503
87.4 Procedure 503
87.5 Pearls and Pitfalls 504
References 504

Part XII Skin and Soft Tissue Procedures 505

88 Local Anesthesia 507
Derek Ailes and Muhammad Waseem
88.1 Indications 507
88.2 Contraindications 507
88.3 Materials and Medications 508
88.4 Procedure 509
88.5 Complications 510
88.6 Pearls 510
References 510

89 Regional Anesthesia (Nerve Blocks) 511
Derek Ailes and Muhammad Waseem
89.1 Indications 511
89.2 Contraindications 511
89.3 Materials and Medications 512
89.4 Procedure: General Block 513
89.5 Complications 513
89.6 Pearls and Pitfalls 513
89.7 Selected Specific Blocks 514
References 523

90 Wound Management and Hemostasis 525
Rich Teitell and Muhammad Waseem
90.1 Indications for Primary Wound Repair 525
90.2 Contraindications for Primary Wound Repair 525
90.3 Procedure for Wound Closure and Hemostasis 525
90.4 Complications 528
90.5 Pearls and Pitfalls 528
Selected Reading 528

91 Burn Care 529
Thomas Parry and Jeffrey Pepin
91.1 Burn Description 529
91.2 Indications 530
91.3 Materials and Medications 530
91.4 Sizing 530
91.5 Fluid Resuscitation 532
91.6 Procedure 532
91.7 Complications 532
91.8 Pearls and Pitfalls 532
91.9 Admission Criteria 532
Selected Reading 533

92 Wound Closure 535
Oliver Michael Berrett, Jeffrey Joseph Harroch, Karlene Hosford, and Muhammad Waseem
92.1 Indications 535
92.2 Contraindications 535

92.3 Methods 535
92.4 Preparation 535
92.5 Suture Repair 536
92.6 Alternative Methods of Wound Closure. 541
92.7 Pearls and Pitfalls. 542
Selected Reading. 542

93 Wound Closure with Tissue Adhesive 543
Pratik S. Patel and Latha Ganti
93.1 Indications 543
93.2 Contraindications 543
93.3 Materials and Medications 543
93.4 Optional Materials 543
93.5 Procedure 544
93.6 Complications 545
93.7 Pearls and Pitfalls. 545
Selected Reading. 545

94 Fishhook Removal 547
Judith K. Lucas
94.1 Indications 547
94.2 Contraindications 547
94.3 Materials and Medications 547
94.4 Procedures 548
94.5 Postremoval Wound Care 551
94.6 Complications 551
94.7 Pearls and Pitfalls. 551
Selected Reading. 551

95 Tick Removal 553
David N. Smith and Judith K. Lucas
95.1 Indications 553
95.2 Contraindications 553
95.3 Materials and Medications 553
95.4 Procedure 553
95.5 Complications 554
95.6 Pearls and Pitfalls. 554
Selected Reading. 555

96 Subungal Hematoma Drainage 557
Pratik S. Patel and Latha Ganti
96.1 Causes 557
96.2 Indications 557
96.3 Contraindications 557
96.4 Materials and Medications 557
96.5 Procedure 558
96.6 Complications (Rare) 559
96.7 Pearls and Pitfalls. 559
Selected Reading. 559

97 Incision and Drainage of Abscess 561
Nicholas D. Caputo, Karlene Hosford, and Muhammad Waseem
97.1 Indications 561
97.2 Contraindications 561
97.3 Materials and Medications 561
97.4 Procedure 563

97.5 Complications 565
97.6 Pearls 565
Selected Reading. 565

Part XIII Orthopedic Procedures. 567

98 Splinting. 569
Christopher H. Stahmer and Muhammad Waseem
98.1 Indications 569
98.2 Contraindications. 569
98.3 Materials and Medications. 569
98.4 Procedure 570
98.5 Complications 572
Selected Reading. 572

99 Ulnar Gutter Splint. 573
Jeffrey Kile, Katrina John, and Amish Aghera
99.1 Indications 573
99.2 Contraindications. 573
99.3 Materials 573
99.4 Procedure 575
99.5 Complications 579
99.6 Pearls and Pitfalls. 579
Selected Reading. 579

100 Shoulder Dislocation Reduction Techniques 581
Katrina Skoog Nguyen, L. Connor Nickels, and Rohit Pravin Patel
100.1 Indications 581
100.2 Contraindications. 583
100.3 Materials and Medications. 583
100.4 Procedure 583
100.5 Complications 588
100.6 Pearls 588
Selected Reading. 588

101 Elbow Dislocation Reduction. 589
Katrina John, Jeffrey Kile, and Amish Aghera
101.1 Indications 589
101.2 Contraindications. 589
101.3 Materials and Medications. 590
101.4 Procedure for Posterior Dislocations 590
101.5 Procedure for Anterior Dislocations 592
101.6 Procedure for Radial Head Subluxations. 593
101.7 Complications 595
101.8 Pearls and Pitfalls. 595
Selected Reading. 595

102 Distal Interphalangeal Joint Reduction 597
Justin Chen and Muhammad Waseem
102.1 Indications 597
102.2 Contraindications. 597
102.3 Materials and Medications. 597
102.4 Procedure 598
102.5 Complications 599
102.6 Pearls and Pitfalls. 599
Selected Reading. 599

103 Hip Dislocation Reduction 601
Katrina John, Jeffrey Kile, and Amish Aghera
103.1 Indications 601
103.2 Contraindications 601
103.3 Materials and Medications 601
103.4 Procedure 601
103.5 Complications 606
103.6 Pearls and Pitfalls 606
Selected Reading 606

104 Knee Dislocation Reduction 607
Katrina John, Jeffrey Kile, and Amish Aghera
104.1 Indications 607
104.2 Contraindications 607
104.3 Materials and Medications 607
104.4 Procedure 607
104.5 Complications 612
104.6 Pearls and Pitfalls 612
Selected Reading 613

105 Ankle Dislocation Reduction 615
Katrina John, Jeffrey Kile, and Amish Aghera
105.1 Indications 615
105.2 Contraindications 615
105.3 Materials and Medications 615
105.4 Procedure 615
105.5 Complications 618
105.6 Pearls and Pitfalls 618
Selected Reading 618

106 Arthrocentesis 619
Shalu S. Patel and Bobby K. Desai
106.1 Indications 619
106.2 Contraindications 619
106.3 Materials and Medications 619
106.4 Procedure 619
106.5 Complications 621
106.6 Pearls and Pitfalls 621
Selected Reading 621

107 Intra-articular Injection 623
Bharat Kothakota and Muhammad Waseem
107.1 Indications 623
107.2 Contraindications 623
107.3 Materials and Medications 623
107.4 Procedure 624
107.5 Complications 628
107.6 Pearls 628
Selected Reading 629

108 Sugar-Tong Splint 631
Katrina John, Jeffrey Kile, and Amish Aghera
108.1 Indications 631
108.2 Contraindications 631
108.3 Materials 631
108.4 Procedure 633

108.5 Complications 635
108.6 Pearls and Pitfalls. 635
Selected Reading. 635

Part XIV Obstetric Procedures. 637

109 Fetal Heart Rate Monitoring 639
Nathaniel Lisenbee and Joseph A. Tyndall
109.1 Indications 639
109.2 Contraindications 639
109.3 Methods 639
109.4 Equipment and Procedures 640
109.5 Complications 642
References. 642
Selected Reading. 642

110 Ultrasonography for Ectopic Pregnancy 643
L. Connor Nickels
110.1 Indications 644
110.2 Contraindications 646
110.3 Materials and Medications. 646
110.4 Procedure 646
110.5 Complications 646
110.6 Pearls and Pitfalls. 646
Reference 647
Selected Reading. 647

111 Ultrasonography for Hydatidiform Mole 649
L. Connor Nickels
111.1 Indications 649
111.2 Contraindications 649
111.3 Materials and Medications. 649
111.4 Procedure 649
111.5 Complications 651
111.6 Pearls and Pitfalls. 651
Selected Reading. 651

112 Ultrasonography for Blighted Ovum (Anembryonic Gestation) 653
Katrina Skoog Nguyen and L. Connor Nickels
112.1 Indications 653
112.2 Contraindications 654
112.3 Materials and Medications. 654
112.4 Procedure 654
112.5 Complications 654
112.6 Pearls and Pitfalls. 654
Selected Reading. 654

113 Ultrasonography for Threatened, Incomplete, or Compete Abortion 655
L. Connor Nickels and Giuliano De Portu
113.1 Indications 655
113.2 Contraindications 655
113.3 Materials and Medications. 655
113.4 Procedure 655
113.5 Complications 657
113.6 Pearls and Pitfalls. 657
Selected Reading. 657

114 Ultrasonography for Placenta Previa 659
L. Connor Nickels and Giuliano De Portu
114.1 Indications 660
114.2 Contraindications 660
114.3 Materials and Medications 660
114.4 Procedure 660
114.5 Complications 662
114.6 Pearls and Pitfalls 662
Reference 662
Selected Reading 662

115 Vaginal Delivery 663
Umarfarook Javed Mirza, Christopher Shields, and Muhammad Waseem
115.1 Indications 663
115.2 Contraindications 663
115.3 Materials and Medications 663
115.4 Procedure 665
115.5 Complications 666
115.6 Pearls and Pitfalls 666
Selected Reading 666

116 Shoulder Dystocia Management 667
Irina Fox Brennan and Joseph A. Tyndall
116.1 Indications 667
116.2 Contraindications 667
116.3 Materials and Medications 667
116.4 Procedure 668
116.5 Complications 673
116.6 Pearls and Pitfalls 673
Selected Reading 673

117 Breech Delivery in the Emergency Department 675
Kristin Stegeman, Sapnalaxmi Amin, Anton A. Wray, and Joseph A. Tyndall
117.1 Indications 676
117.2 Contraindications 676
117.3 Materials and Medications 676
117.4 Procedure 676
117.5 Complications 680
117.6 Pearls and Pitfalls 680
Selected Reading 680

118 Management of Primary Postpartum Hemorrhage 681
Megan Kwasniak, Anton A. Wray, and Joseph A. Tyndall
118.1 Indications 681
118.2 Contraindications 681
118.3 Materials and Medications 681
118.4 Procedure 681
118.5 Complications 683
118.6 Pearls and Pitfalls 683
Selected Reading 684

119 Perimortem Cesarean Section 685
Jordana J. Haber, Elaine B. Josephson, and Muhammad Waseem
119.1 Indications 685
119.2 Contraindications 685
119.3 Materials and Medications 685

119.4 Procedure . . . 686
119.5 Complications . . . 688
119.6 Pearls and Pitfalls . . . 688
Selected Reading . . . 688

Part XV Common Pediatric Procedures . . . 689

120 Peripheral Venous Catheterization . . . 691
David N. Smith and Judith K. Lucas
120.1 Indications . . . 691
120.2 Contraindications . . . 691
120.3 Materials and Medications . . . 691
120.4 Procedure . . . 692
120.5 Pearls and Pitfalls . . . 697
Selected Reading . . . 697

121 Umbilical Venous Catheters (Insertion and Removal) . . . 699
Judith K. Lucas
121.1 Indications . . . 699
121.2 Contraindications . . . 699
121.3 Materials and Medications . . . 699
121.4 Procedure (Insertion) . . . 701
121.5 Procedure (Removal) . . . 703
121.6 Complications . . . 703
121.7 Pearls and Pitfalls . . . 704
Selected Reading . . . 704

122 Intraosseous Access . . . 705
Judith K. Lucas
122.1 Indications . . . 705
122.2 Contraindications . . . 705
122.3 Materials and Medications . . . 705
122.4 Procedure . . . 706
122.5 Complications . . . 710
122.6 Pearls and Pitfalls . . . 710
Selected Reading . . . 710

123 Lumbar Puncture in Pediatrics . . . 711
Maritza A. Plaza-Verduin and Judith K. Lucas
123.1 Indications . . . 711
123.2 Contraindications . . . 711
123.3 Materials and Medications . . . 711
123.4 Procedure . . . 713
123.5 Complications . . . 715
123.6 Pearls and Pitfalls . . . 715
Selected Reading . . . 716

124 Suprapubic Bladder Aspiration . . . 717
Maritza A. Plaza-Verduin and Judith K. Lucas
124.1 Indications . . . 717
124.2 Contraindications . . . 717
124.3 Materials and Medications . . . 717
124.4 Procedure . . . 718
124.5 Complications . . . 720
124.6 Pearls and Pitfalls . . . 720
Selected Reading . . . 720

125 Removal of Hair/Thread Tourniquet 721
Judith K. Lucas
125.1 Indications 721
125.2 Contraindications 722
125.3 Materials and Medications 722
125.4 Procedures 722
125.5 Complications 726
125.6 Pearls and Pitfalls 726
Selected Reading 726

126 Use of the Broselow Tape 727
Judith K. Lucas
126.1 Indications 727
126.2 Contraindications 727
126.3 Materials 727
126.4 Procedure 728
126.5 Pearls and Pitfalls 729
Selected Reading 729

127 Nursemaid's Elbow 731
Judith K. Lucas
127.1 Definition 731
127.2 Clinical Diagnosis 732
127.3 Indications 732
127.4 Contraindications 732
127.5 Materials and Medications 732
127.6 Procedure 732
127.7 Evaluation After Procedure 735
127.8 Pearls and Pitfalls 735
Selected Reading 735

Index 737

Associate Editors

Lars K. Beattie, MD
Bobby K. Desai, MD
Marie-Carmelle Elie, MD
Judith K. Lucas, MD
L. Connor Nickels, MD
Department of Emergency Medicine
University of Florida
Gainesville, FL, USA

Sapna Amin, MD
Central Florida Healthcare
Mulberry, FL, USA

Ad-Hoc Reviewers

Linda Sheehan-Foster, PA-C
Jason Williams, RN
Lake City Veterans Affairs Medical Center
Lake City, Florida

Tej Geoffrey Stead
Buccholz High School
Gainesville, Florida

Models

Lars K. Beattie
Eike Flach
Tatiana Havryliuk
Katy Howard
Ben M. Mahon
Pratik S. Patel
Joseph Rabinovich
Thor Shiva Stead
Karthik Dax Stead
Clarence R. Tucker

Contributors

Michael A. Abraham, DMD United States Air Force, Dental Corps, Minot, ND, USA

Abimbola O. Adewumi, BDS, FDSR Department of Pediatric Dentistry, University of Florida College of Dentistry, Gainesville, FL, USA

Amish Aghera, MD Department of Emergency Medicine, Maimonides Medical Center, New York, NY, USA

Derek Ailes, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Saadia Akhtar, MD Department of Emergency Medicine, Mount Sinai Beth Israel, New York, NY, USA

Brandon R. Allen, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Hassan M. Alnuaimat, MD Department of Medicine, University of Florida, University of Florida Health Shands Hospital, Gainesville, FL, USA

Sapnalaxmi Amin, MD Department of Family Medicine/Urgent Care, Bayside Urgent Care Center, Clearwater, FL, USA

Michael Anana, MD Emergency Department, University Hospital, Rutgers New Jersey Medical School, Newark, NJ, USA

Amir Azari, DMD Department of Oral and Maxillofacial Surgery, Oregon Health and Science University, Portland, OR, USA

Lars K. Beattie, MD, MS Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Deena Bengiamin, MD Department of Emergency Medicine, University of San Francisco Fresno, Fresno, CA, USA

Justin Bennett, MD Department of Emergency Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA

Oliver Michael Berrett, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Irina Fox Brennan, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Nicholas D. Caputo, MD, MSc Emergency Department Critical Care, Lincoln Medical and Mental Health Center, New York, NY, USA

Bharath Chakravarthy, MD, MPH Department of Emergency Medicine, University of California Irvine, Orange, CA, USA

Holly H. Charleton, MD Emergency Department, The Brooklyn Hospital Center, Brooklyn, NY, USA

Justin Chen, MD, MSc Department of Emergency Medicine, North Shore University Hospital, Manhasset, NY, USA

Christian Coletti, MD Department of Emergency Medicine and Internal Medicine, Christiana Care Health System, Newark, DE, USA

Ali H. Dabaja, DO Division of Critical Care, Department of Anesthesia, University of Florida, Gainesville, FL, USA

Dawood G. Dalaly, DO Department of Surgery, University of Florida, Gainesville, FL, USA

Giuliano De Portu, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Bobby K. Desai, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Rui Domingues, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Lee Richard Donner, MD Emergency Medicine Department, Lincoln Medical and Mental Health Center, New York, NY, USA

Tina Dulani, MD Department of Emergency Medicine, New York Methodist Hospital, Brooklyn, NY, USA

Marie-Carmelle Elie, MD, RDMS Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Etan Eitches, MD Department of Emergency Medicine, Beth Israel Medical Center, New York, NY, USA

Kevin D. Ergle, MD Department of Internal Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA

Melinda W. Fernandez, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Christian Fromm, MD Department of Emergency Medicine, Maimonides Medical Center, SUNY Downstate College of Medicine, New York, NY, USA

Latha Ganti, MD, MS, MBA Department of Emergency Medicine, Orlando Veterans Affairs Medical Center and University of Central Florida, Orlando, FL, USA

Diane F. Giorgi, MD Department of Emergency Medicine, Mount Sinai Queens, New York, NY, USA

Jacob J. Glaser, MD Combat Casualty Care Directorate, Naval Medical Research Unit, San Antonio, TX, USA

Jordana J. Haber, MD Department of Emergency Medicine, Maimonides Medical Center, New York, NY, USA

Jeffrey Joseph Harroch, MD Department of Emergency Medicine, University of Miami Miller School of Medicine, University of Miami Hospital, Miami, FL, USA

Tatiana Havryliuk, MD Department of Emergency Medicine, University of Colorado Denver, Denver, CO, USA

Braden Hexom, MD Department of Emergency Medicine, Mount Sinai Hospital, New York, NY, USA

Karlene Hosford, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Rajnish Jaiswal, MD Department of Emergency Medicine, New York Medical College, Metropolitan Hospital Center, New York, NY, USA

Katrina John, MBBS Department of Emergency Medicine, Eisenhower Medical Center, Rancho Mirage, CA, USA

Jason Jones, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Kevin M. Jones, MD Trauma and Surgical Critical Care, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA

Elaine B. Josephson, MD Department of Emergency Medicine, Weill Cornell Medical College of Cornell University, Lincoln Medical and Mental Health Center, New York, NY, USA

Jeffrey Kile, MBBS, PhD, MPH Department of Emergency Medicine, Eisenhower Medical Center, Rancho Mirage, CA, USA

Bharat Kothakota, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Zachary B. Kramer, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Megan Kwasniak, MD Emergency Department, The Brooklyn Hospital Center, New York, NY, USA

Nathaniel Lisenbee, MD U. S. Air Force Medical Center Keesler, Biloxi, MI, USA

Judith K. Lucas, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Benjamin M. Mahon, MD Ponciana Medical Center, Kissimmee, FL, USA

Clint Masterson, MD Department of Emergency Medicine, Mayo Clinic Health System in Fairmont, Fairmont, MN, USA

Lucas McArthur, MD Department of Emergency Medicine, Maimonides Medical Center, New York, NY, USA

Jay Menaker, MD Departments of Surgery and Emergency Medicine, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA

Umarfarook Javed Mirza, DO Department of Emergency Medicine, Baylor University Medical Center, Dallas, TX, USA

David P. Nguyen, DO Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA

Katrina Skoog Nguyen, DO Northwest Community Hospital, Arlington Heights, IL, USA

Thomas T. Nguyen, MD Department of Emergency Medicine, Mount Sinai Beth Israel, New York, NY, USA

L. Connor Nickels, MD, RDMS Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA

Marylin Otero, MD Department of Emergency Medicine, Franklin Hospital, Valley Stream, NY, USA

Eric S. Papierniak, DO Division of Pulmonary, Critical Care, and Sleep Medicine, University of Florida and Malcom Randall VA Medical Center, Gainesville, FL, USA

Ram A. Parekh, MD Department of Emergency Medicine, Icahn School of Medicine at Mount Sinai, Elmhurst Hospital Center, Elmhurst, NY, USA

Sohan Parekh, MD Department of Emergency Medicine, University of Texas at Austin Dell Medical School, Austin, TX, USA

Department of Emergency Medicine, University Medical Center Brackenridge, Austin, TX, USA

Thomas Parry, MD Department of Emergency Medicine, Lincoln Hospital and Mental Health Center, New York, NY, USA

Pratik S. Patel, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Rohit Pravin Patel, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Shalu S. Patel, MD Department of Emergency Medicine, Florida Hospital Tampa, Florida Hospital Carrollwood, Tampa, FL, USA

Jeffrey Pepin, MD Department of Emergency Medicine, University of Minnesota Medical Center Fairview, Minneapolis, MN, USA

Joshua Perry, DMD Department of Prosthodontics, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA

Susana Perry, DMD Department of Pediatric Dentistry, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA

Maritza A. Plaza-Verduin, MD Pediatric Division, Department of Emergency Medicine, Arnold Palmer Hospital for Children, Orlando, FL, USA

Joseph Rabinovich, MD Department of Emergency Medicine, Mount Sinai School of Medicine, Elmhurst Hospital Center, Elmhurst, NY, USA

Nauman W. Rashid, MD Department of Emergency Medicine, WellStar Kennestone Hospital, Marietta, GA, USA

Rosalia Rey, DDS Department of Restorative Dental Sciences, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA

Nour Rifai, MBCHB Department of Emergency Medicine, Christiana Care Health System, Newark, DE, USA

Carlos J. Rodriguez, DO Division of Trauma Surgery, Surgical Critical Care, Walter Reed National Military Medical Center, Bethesda, MD, USA

Joseph D. Romano, MD Departments of Emergency Medicine and Internal Medicine, Christiana Care Health System, Newark, DE, USA

Mary T. Ryan, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Weston Seipp, MD Department of Emergency Medicine, University of California Irvine Medical Center, Orange, CA, USA

Christopher Shields, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Deylin I. Negron Smida, MD Department of Emergency Medicine, University of Pittsburgh Medical Center, Saint Margaret Hospital, Pittsburgh, PA, USA

David N. Smith, MD Department of Pediatrics, University of Alabama at Birmingham, Children's of Alabama, Birmingham, AL, USA

Christopher J. Spencer, DDS Department of Restorative Dental Sciences, University of Florida College of Dentistry, Gainesville, FL, USA

Christopher H. Stahmer, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Franci Stavropoulos, DDS Department of Dental Specialties - Oral and Maxillofacial Surgery, Gundersen Health System, La Crosse, WI, USA

Kristin Stegeman, MD Department of Emergency Medicine, The Brooklyn Hospital Center, New York, NY, USA

Deborah M. Stein, MD, MPH Department of Surgery, University of Maryland School of Medicine, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA

Rich Teitell, MD Department of Emergency Medicine, Waterbury Hospital, Waterbury, CT, USA

Kevin Tench, MD Department of Emergency Medicine, Banner Boswell Medical Center, Sun City, AZ, USA

Ronald Tesoriero, MD Department of Surgical Critical Care, University of Maryland School of Medicine, R Adams Cowley Shock Trauma Center, Baltimore, MD, USA

Coben Thorn, MD Department of Emergency Medicine, Bon Secours St. Francis Health System, Greenville, SC, USA

Matthew R. Tice, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Tracy Timmons, MD Department of Trauma/Surgical Critical Care, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA

Shannon Toohey, MD Department of Emergency Medicine, University of California Irvine Medical Center, Orange, CA, USA

Ann Tsung, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Laura Tucker, DDS Department of Pediatric Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA

Joseph A. Tyndall, MD, MPH Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Muhammad Waseem, MD Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA

Geraldine Weinstein, DDS Department of Restorative Dental Sciences, University of Florida College of Dentistry, Gainesville, FL, USA

Jennifer Westcott, DMD Private Practice, Palm Beach Gardens, FL, USA

Stephanie Wetmore-Nguyen, MD Department of Emergency Medicine, New York-Presbyterian Hospital/Columbia University Medical Center, New York, NY, USA

Anton A. Wray, MD Department of Emergency Medicine, The Brooklyn Hospital Center, New York, NY, USA

Henry Young II, MD Department of Emergency Medicine, University of Florida, Gainesville, FL, USA

Part I

Vascular Procedures

1

Arterial Cannulation (Radial and Femoral)

Jeffrey Kile, Katrina John, and Amish Aghera

Arterial cannulation is frequently performed in the care of critically ill patients for purposes of both serial arterial blood gas sampling and continuous intra-arterial blood pressure monitoring. It also provides arterial access for less common procedures, including thrombolysis, embolization, angiography, and infusion of vasoactive drugs. This chapter discusses cannulation of the radial and femoral arteries—the two most common sites for indwelling arterial catheter placement.

1.1 Indications

- Continuous monitoring of blood pressure in acute illness or major surgery
- Serial sampling of arterial blood during resuscitation
- Inability to use noninvasive blood pressure monitoring (*e.g.*, burns, morbid obesity)
- Continuous infusion of vasoactive inotropes (*e.g.*, phentolamine for reversal of local anesthesia)
- Angiography
- Embolization

1.2 Contraindications

- Absolute
  - Circulatory compromise in the extremity
  - Third-degree burns of the extremity
  - Raynaud's syndrome
  - Thromboangiitis obliterans (Buerger's disease)
- Relative
  - Recent surgery in the extremity
  - Local skin infection
  - Abnormal coagulation
  - Insufficient collateral circulation
  - First- or second-degree burns of the extremity
  - Arteriosclerosis

1.3 Materials and Medications

- Radial artery cannulation: standard over-the-needle catheter assembly (Fig. 1.1)
  - Antimicrobial solution and swabs
  - Sterile gloves
  - Local anesthetic (1–2 % lidocaine without epinephrine)
  - Blunt needle
  - 25- or 27-gauge needle
  - Two 5 mL syringes
  - 4″×4″ gauze sponges
  - Standard over-the-needle catheter assembly
- Additional materials required for radial artery cannulation: over-the-needle catheter assembly with integrated guide wire
  - Over-the-needle catheter assembly with integrated guide wire
- Additional materials required for femoral artery cannulation: The Seldinger Technique
  - Introducer needle
  - Guide wire
  - Scalpel
  - Dilator
  - Arterial catheter

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_1

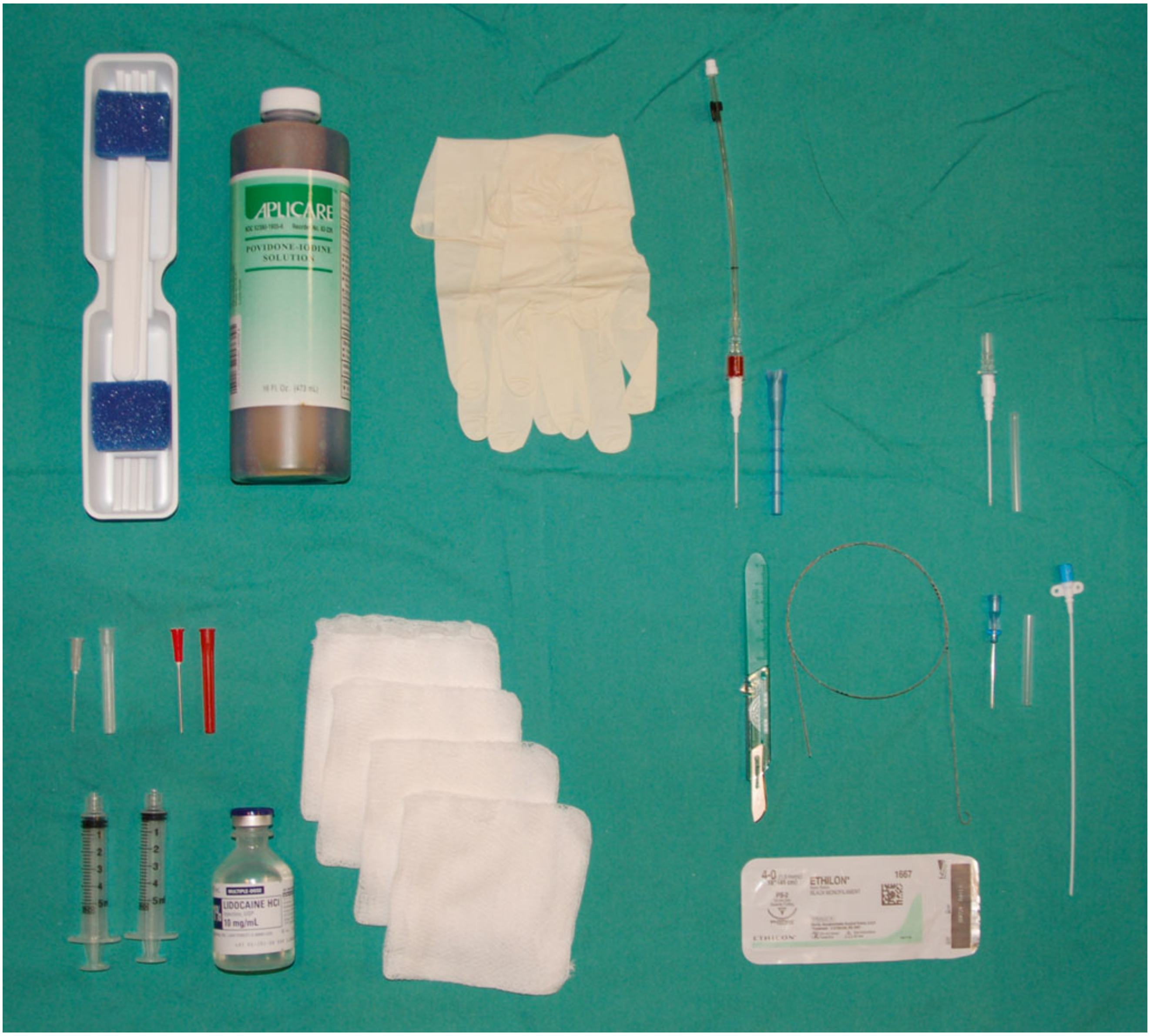

Fig. 1.1 Materials and medications

1.3.1 Procedure: Radial Artery Cannulation—Standard Over-the-Needle Catheter Assembly

1. Ensure adequate collateral flow in the selected extremity using the Allen test (see details below).
2. Immobilize the extremity by dorsiflexing the wrist to approximately 60° over a small towel roll and taping the base of the fingers to an arm board or other flat, fixed surface (Fig. 1.2).
    - Rotation of the wrist may shift the artery from its normal anatomical position, complicating cannulation.
3. Locate vessel by palpation of arterial pulse using the second and third fingers of the gloved nondominant hand.
4. Sterilize overlying skin with antimicrobial solution.
5. Inject local anesthetic to raise a small (0.5 cm) wheal using 25- or 27-gauge needle, and direct needle through wheal to infiltrate skin superficial to the artery with additional local anesthetic.
    - Infiltration of the subcutaneous tissue with local anesthetic may also reduce vessel spasm during arterial puncture.
    - Injection of local anesthetic into the vessel may precipitate arrhythmia, so draw back on the plunger prior to infiltration to ensure the tip of the needle is not inside the vessel.
    - Injection of excessive anesthetic when raising a wheal may obscure palpation of the pulse.
6. Ensure proper function of needle-cannula assembly by checking that cannula advances smoothly over needle.
7. Connect a 5 mL syringe with the plunger removed to the over-the-needle catheter assembly.
    - Attachment of syringe improves control during cannulation.
8. Hold syringe connected to needle-cannula assembly like a pen with needle bevel facing upward.
9. Directing the needle at a 30° angle to the skin, puncture the skin through the anesthetic wheal immediately overlying the palpated artery, and advance needle slowly until tip enters arterial lumen, which is confirmed by visible arterial blood flow ("flashback") into the needle hub and syringe.
    - Avoid self-puncture by maintaining adequate distance between needle tip and index finger (Fig. 1.3).
10. Reduce the angle between needle and skin (by lowering the needle) and advance an additional 2 mm to ensure catheter tip (which sits approximately 2 mm behind the needle tip) has entered the lumen.
    - Advancing the needle too far (or failing to reduce angle between needle and skin) once initial flashback is visualized may result in piercing the back side (or "double puncture") of the artery wall, in which case visible blood flow will cease; if this occurs, slowly withdraw needle several millimeters until pulsatile blood flow reappears.
11. Stabilize position of introducer needle and advance catheter alone into artery over needle until hub of catheter is in contact with the skin; blood flow from catheter hub at this point indicates successful cannulation of the artery.
    - If difficulty is encountered at this step, catheter hub may be rotated slightly to facilitate advancement.
12. Remove the needle without dislodging catheter from artery.
13. Manually apply pressure to proximal aspect of artery to occlude blood flow from the catheter.
14. Attach desired extension tubing, injection cap, and stopcock to the catheter hub.
15. Secure the catheter hub to the skin using silk (2.0) or nylon (4.0) sutures as follows. Take a 0.5 cm bite of skin under the catheter hub with the suture needle, tie several knots in the suture without pinching the skin, then tie a second set of knots around the hub of the catheter firmly. If the catheter assembly contains an integrated suture wing for fixation, take a 0.5 cm bite of skin under suture wing with the suture needle, thread suture through wing perforation, and secure wing against the skin with several knots. If suture wing has two perforations, repeat this process to secure other half of wing to skin (Fig. 1.4).
16. Cover the catheter with an appropriate self-adhesive sterile dressing.
    - A small bead of antibiotic ointment applied to the puncture site prior to dressing reduces the likelihood of cutaneous wound infection.
17. Secure the tubing connected to the catheter with gauze and adhesive tape or other sterile dressing.
18. Ensure all connections extending from catheter are tight and well secured, as accidental disconnection may result in rapid exsanguination.

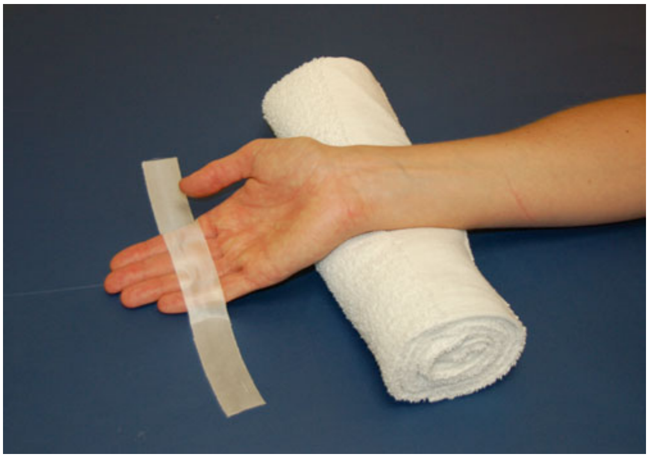

Fig. 1.2 Correct position of wrist prior to cannulation

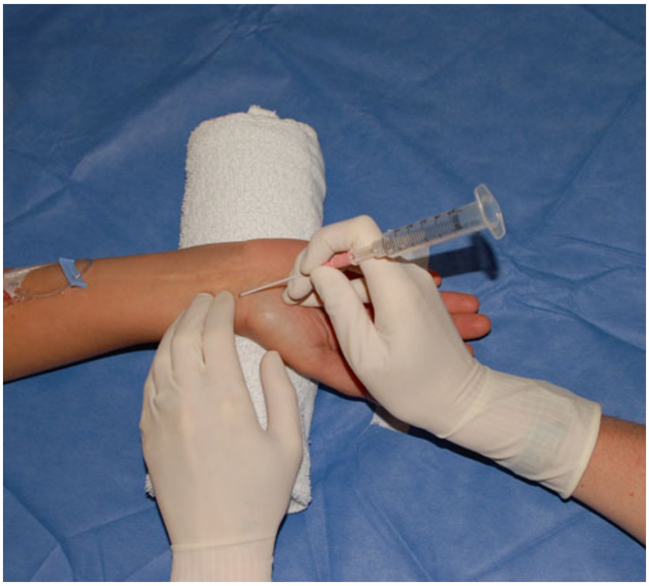

Fig. 1.3 Puncture of radial artery with standard over-the-needle catheter assembly

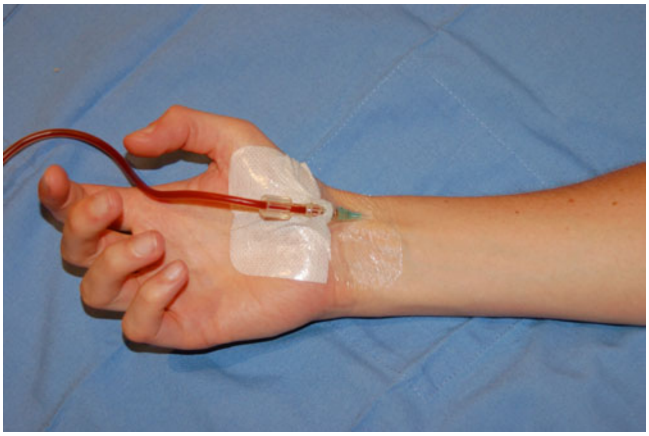

Fig. 1.4 Radial arterial catheter secured to wrist

1.3.2 Procedure: Radial Artery Cannulation—Over-the-Needle Catheter Assembly with Integrated Guide Wire (Arrow and Other Brands)

Perform steps 1–6 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below:

1. Remove protective cap from needle-cannula assembly and ensure proper function by sliding actuation lever along extension tubing to advance and retract guide wire through needle.
2. Retract guide wire as far back as possible using actuation lever to maximize visibility arterial blood flow ("flashback") within introducer hub.

Perform steps 8–10 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below (Fig. 1.5):

1. Hold needle stationery and slowly slide actuating lever forward to feed guide wire as far as possible into artery.
    - If resistance is met while feeding the guide wire, discontinue sliding actuating lever and withdraw entire unit from artery to prevent damage to guide wire or vessel wall
2. Advance entire assembly 1–2 mm further into vessel to ensure catheter tip (which sits approximately 2 mm behind the needle tip) has entered the lumen.
3. Stabilize clear introducer hub in position and advance catheter forward into artery over guide wire until hub of catheter is in contact with the skin.
    - If difficulty is encountered at this step, catheter hub may be rotated slightly to facilitate advancement.
4. Stabilize catheter in position and withdraw introducer needle, guide wire, and feed tube as a single unit; blood flow from catheter hub at this point indicates successful cannulation of the artery.

Perform steps 13–18 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above.

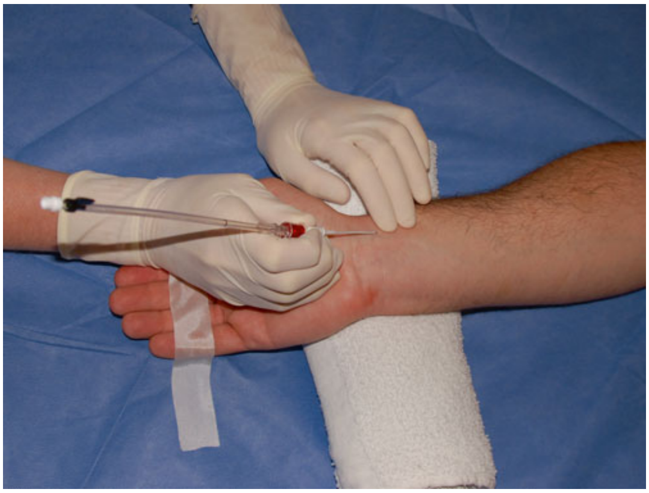

Fig. 1.5 Puncture of radial artery using over-the-needle catheter assembly with integrated guide wire

1.3.3 Procedure: Radial Artery Cannulation—The Allen Test

1. Occlude both radial and ulnar arteries of one extremity with digital pressure at wrist.
2. Instruct patient to repeatedly clench the fist tightly to exsanguinate the hand while occlusion of the arteries is maintained.
3. Without releasing digital pressure on arteries, instruct patient to extend fingers and observe palmar surface to confirm blanching of skin.
4. Release pressure on ulnar artery only and observe palmar surface for reperfusion (Fig. 1.6).
    - If reperfusion of the hand does not occur within 5–10 s, ulnar arterial blood flow may be compromised and radial artery cannulation should not be attempted. If reperfusion is brisk, repeat the test releasing pressure on radial artery only and observing palmar surface for reperfusion. If the return of rubor takes longer than 5–10 s, radial artery puncture should not be performed.

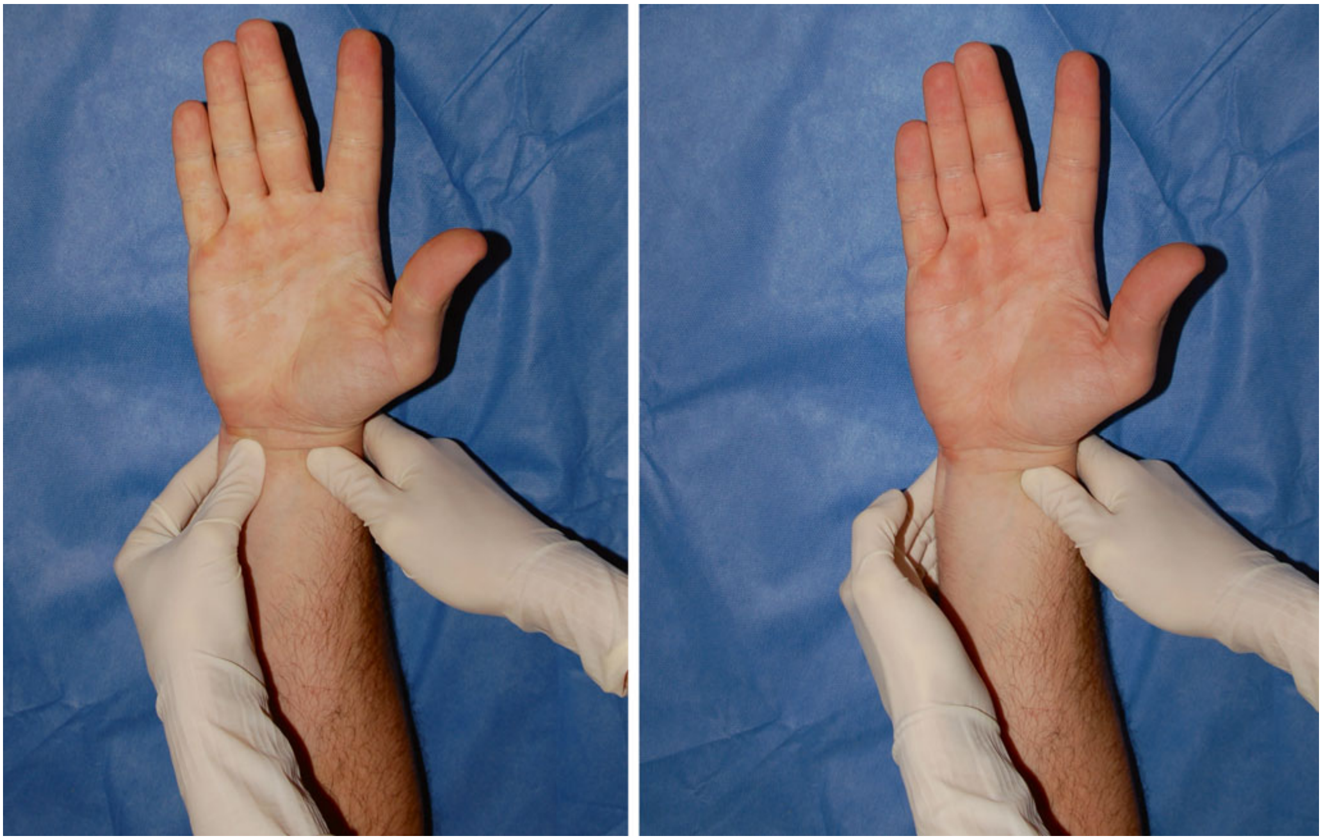

Fig. 1.6 The Allen Test

1.3.4 Procedure: Femoral Artery Cannulation—The Seldinger Technique

1. Place the patient in the supine position with the inguinal region adequately exposed.
2. Palpate the femoral pulse, located at midpoint between pubic symphysis and anterior superior iliac spine, using the second and third fingers of the gloved nondominant hand.

Perform steps 4 and 5 from the "Radial Artery Cannulation: Standard Over-the-Needle Catheter Assembly," above, and then proceed with the steps below (Fig. 1.7):

1. Attach 5 mL syringe to an introducing needle of bore sufficient to accommodate guide wire.
2. Hold syringe connected to introducing needle like a pen with needle bevel facing upward.
3. Directing the needle at a 45° angle to the skin in a cephalic direction, puncture the skin through the anesthetic wheal immediately overlying the palpated artery just distal to the inguinal ligament and advance needle slowly toward palpated artery until tip enters arterial lumen, which is confirmed by visible arterial blood flow ("flashback") into the needle hub and syringe.
    - Avoid self-puncture by maintaining adequate distance between needle tip and index finger.
    - Care must be taken to avoid trauma to the femoral nerve and vein bordering the femoral artery.
4. Hold needle stationery and remove syringe, taking care not to displace the intraluminal position of the needle tip.
    - Advancing the needle too far once initial flashback is visualized may result in piercing the back side (or "double puncture") of the artery wall, in which case visible blood flow will cease; if this occurs, slowly withdraw needle several millimeters until pulsatile blood flow reappears.
5. Occlude needle hub temporarily with gloved finger to prevent unnecessary blood loss and air embolism.
6. Thread blunt end of flexible guide wire smoothly into needle and gently into artery until at least one-quarter of guide wire is intravascular (Fig. 1.8).
    - If resistance is met while threading guide wire, remove wire from needle, reattach syringe, and aspirate blood to confirm continued intraluminal needle tip placement; if resistance is met while *removing* guide wire from needle, remove guide wire and needle from artery *as a single unit* to prevent shearing the guide wire off inside the vessel.
7. Holding the wire securely in place, remove the introducing needle.
8. Using a scalpel, make a small incision (approximately the width of the catheter to be used) through the dermis at the insertion site of the guide wire.
    - Avoid severing the guide wire by facing the sharp edge of the scalpel *away* from the guide wire.
9. While stabilizing the guide wire at its insertion site, thread the dilator over the free end of the guide wire until it is approximately one inch from the skin.
10. Grasp the free end of the guide wire protruding from the tail end of the dilator.
    - If it does not protrude from the tail end of the dilator, the guide wire must be removed sufficiently from the artery to be securely grasped; it must protrude visibly from the tail end of the dilator throughout the subsequent process of threading the dilator into the artery.
11. Holding the dilator firmly near its tip, thread the dilator over the wire into the skin with a back-and-forth twisting motion until it reaches the artery.
    - Only the skin tract should be dilated; dilation of the artery may result in excessive arterial injury and/or hemorrhage.
12. Holding the wire securely in place, remove the dilator.
13. While stabilizing the guide wire at its insertion site, thread the catheter over the free end of the guide wire until it nears the skin.
14. Grasping the guide wire where it protrudes from the tail end of the catheter, thread the catheter into the skin to its appropriate insertion length.
15. While stabilizing the catheter at its insertion site, slowly remove the guide wire.
    - If resistance is met while removing guide wire, remove guide wire and catheter from artery as a single unit to prevent shearing the guide wire off inside the vessel.
16. Secure the catheter to the skin using silk (2.0) or nylon (4.0) sutures. Take a 0.5 cm bite of skin with the suture needle. If the catheter assembly contains integrated "wings" for fixation, thread suture through the perforated wings and secure catheter against the skin with several knots. If no fixation device is included, tie several knots in the suture without pinching the skin, leaving both ends of the suture long. Using the loose ends of the suture, tie a second set of knots around the hub of the catheter, firmly, but without constricting its lumen.

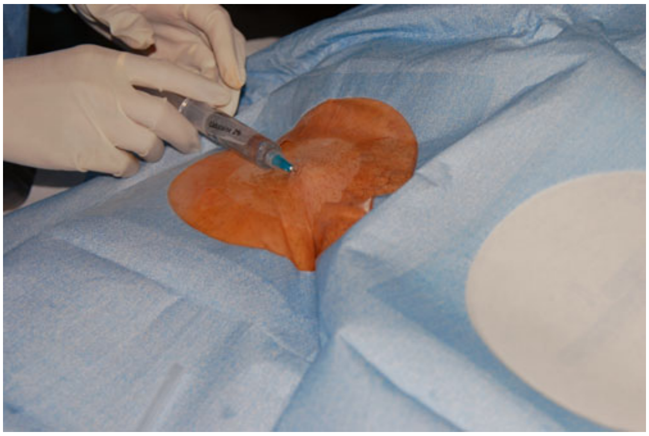

Fig. 1.7 Anesthetic injection over the femoral artery

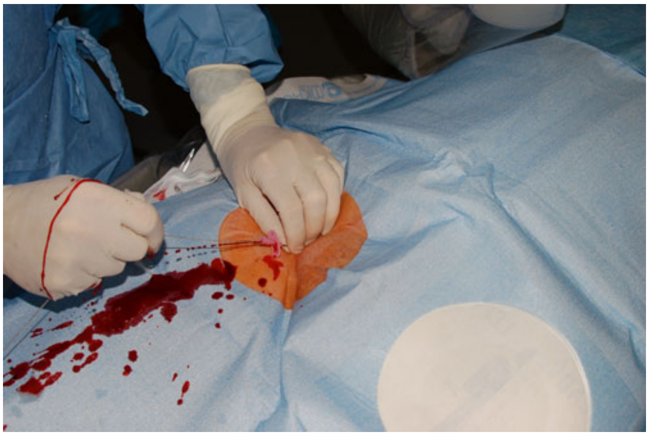

Fig. 1.8 Insertion of guide wire into femoral artery

1.4 Complications

- Hemorrhage
- Hematoma (at puncture site)
- Infection (at insertion site or systemic)
- Thrombosis
- Arteriovenous fistula
- Pseudoaneurysm formation
- Exsanguination (secondary to dislodgement of catheter)
- Cerebrovascular accident (CVA; secondary to air embolism)

1.5 Pearls and Pitfalls

- The shorter and stiffer the plastic tubing connected to the arterial cannula for blood pressure monitoring, the higher its frequency response and the accuracy of measurements.
- Use of an ultrasound probe can facilitate artery location and vessel cannulation.
- Puncture of the femoral artery proximal to the inguinal ligament, or distal to its bifurcation into superficial femoral and deep femoral arteries, may cause massive hemorrhage due to poor vessel compressibility in these regions; the artery should therefore be cannulated just distal to the inguinal ligament, where it is easily compressible against the femoral head if necessary.
- If difficulty is encountered when advancing an over-the-needle catheter into the artery, attach a 10 mL syringe containing 5 mL of sterile normal saline to the catheter hub, aspirate 1 or 2 mL of blood to confirm catheter tip placement within the vessel lumen, and then advance the catheter while gently injecting the saline-blood mixture; the jet of fluid momentarily dilates the lumen, aiding advancement of the catheter.
- An alternative approach to the over-the-needle catheter that will not fully advance is the use of a guide wire. After intraluminal placement of the cannula tip is confirmed by blood return, a guide wire is gently inserted through the catheter into the artery. The cannula is then passed along the guide wire until fully advanced. The guide wire employed must have a blunt, flexible tip to minimize the possibility of vessel wall trauma.
- Two potential consequences of arterial cannulation are vessel obstruction secondary to intravascular thrombosis and hemorrhage (the latter being the most common complication). Choice of puncture site is therefore essential. Due in part to their generous collateral blood flow, as well as ease of compressibility, the radial and femoral are the two most commonly cannulated arteries.
- Repeated puncture following unsuccessful cannulation increases the risk of arterial obstruction secondary to vessel wall damage and thrombosis.

- Double puncture of the cannulated artery by inadvertent over-insertion of the needle has not been shown to increase complications despite the additional trauma to the vessel walls.

Selected Reading

Anderson JS. Arterial cannulation: how to do it. Br J Hosp Med. 1997;57:497.

Gilchrist IC. Reducing collateral damage of the radial artery from catheterization. Catheter Cardiovasc Interv. 2010;76:677–8.

Lemaster CH, Agrawal AT, Hou P, Schuur JD. Systematic review of emergency department central venous and arterial catheter infection. Int J Emerg Med. 2010;3:409–23.

Mitchell JD, Welsby IJ. Techniques of arterial access. Surgery. 2004;22:3–4.

Wilson SR, Grunstein I, Hirvela ER, Price DD. Ultrasound-guided radial artery catheterization and the modified Allen's test. J Emerg Med. 2010;38:354–8.

Ultrasound-Guided Peripheral Intravenous Access

2

Coben Thorn and L. Connor Nickels

2.1 Indications

- Difficulty placing peripheral intravenous (PIV) using traditional methods of direct visualization and palpation
- To reduce needle sticks in a hypercoagulable patient

2.2 Contraindications

- Patient needs emergent central venous access

2.3 Materials and Medications

- Ultrasound machine
- High-frequency (5–8 MHZ) linear probe (Fig. 2.1)
- Ultrasound gel
- Minimum 1.5-in. needle length
- IV setup
- Skilled operator

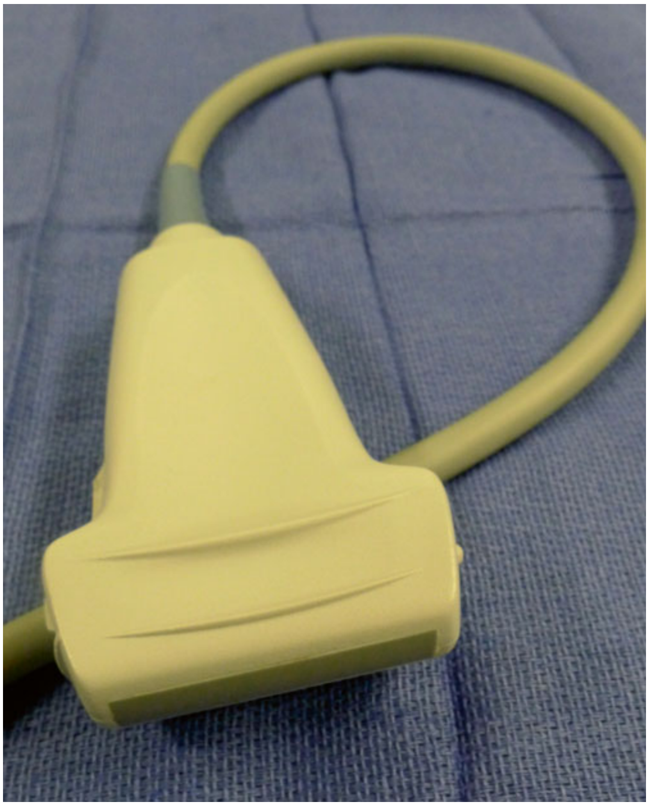

Fig. 2.1 High-frequency (5–8 MHZ) linear probe

C. Thorn, MD
Department of Emergency Medicine, Bon Secours St. Francis Health System, Greenville, SC, USA
e-mail: cobenthorn@gmail.com

L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_2

2.4 Procedure

1. Scan the selected area to identify a target vessel for PIV cannulation.
    - Basilic (runs on the medial side of upper arm) (Fig. 2.2) and cephalic (runs on the lateral side of the upper arm) veins are good superficial veins that are generally not seen without ultrasound. The deep brachial vein is also an option; however, this is a deep vein that runs with the brachial artery, and there is higher chance for complications (Fig. 2.3).
2. Now the site should be prepared for IV insertion. It should be cleaned. An appropriate gauge long needle should be selected, and IV setup should be conveniently located for when access is obtained.
3. The ultrasound probe should be placed in the transverse plane (Fig. 2.4) to best visualize surrounding structures and vein. Alternatively, the probe can be placed longitudinally (Fig. 2.5) for better visualization of needle depth and slope.
    - Most practitioners seem to prefer the transverse approach.
4. The concept of the Pythagorean Theorem is used for accuracy. The needle should be inserted at a 45° angle to the skin and at an equal distance back from the probe as the approximate depth of the vessel vertically. The depth is given on the screen, usually at the bottom in centimeters. As soon as the needle has penetrated the skin the needle tip should be located by fanning the probe toward the needle until it is identified. The needle should then be advanced slowly always keeping the needle tip in view. Once directly on top of the vein, it should tent with pressure and then the needle should be inserted into the vein. Follow the needle in the vein as far as possible while keeping the tip of the needle in the center of the vein to make sure the catheter is securely in the vein and does not infiltrate.

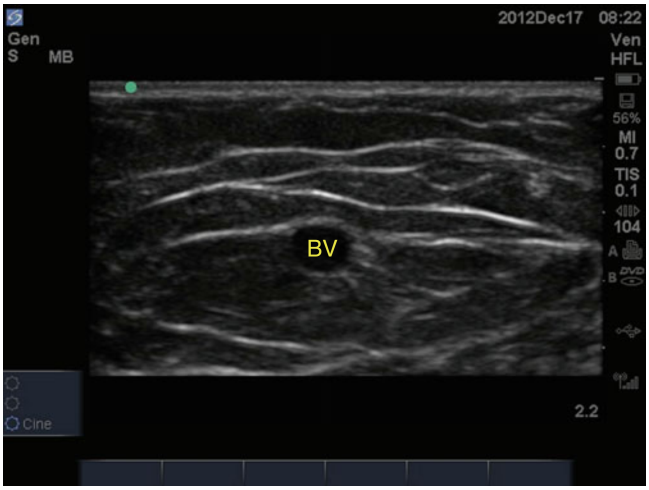
2012Dec17 08:22
Gen
S MB
Ven
HFL
56%
MI 0.7
TIS 0.1
104
BV
Cine
2.2

Fig. 2.2 Basilic vein (BV) located medially when scanning proximally from the antecubital fossa

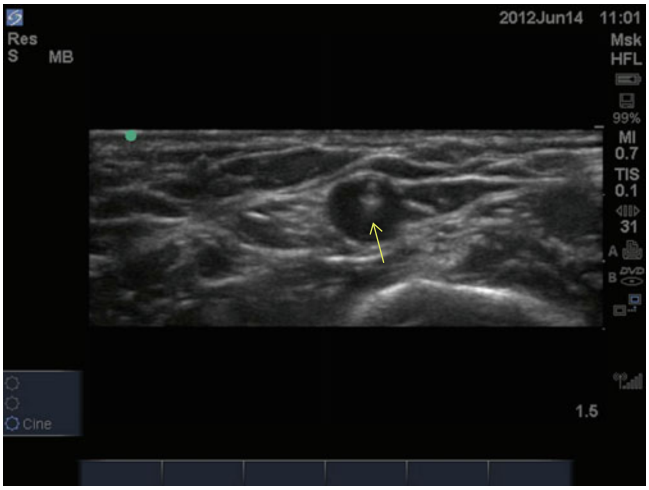
2012Jun14 11:01
Res
S MB
Msk
HFL
99%
MI 0.7
TIS 0.1
31
Cine
1.5

Fig. 2.4 Needle tip in vein seen in a transverse orientation

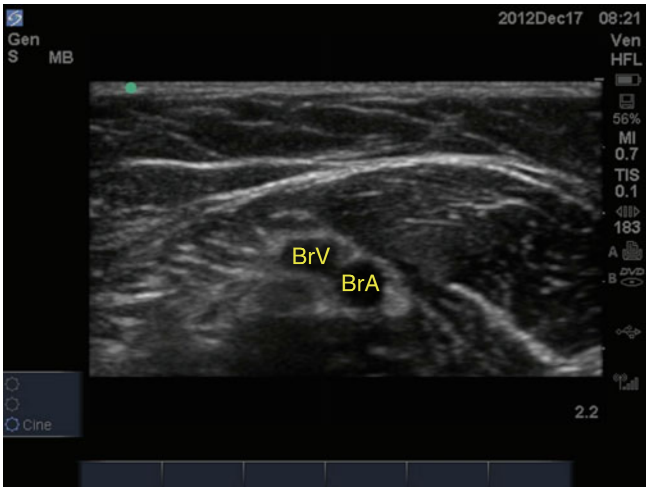
2012Dec17 08:21
Gen
S MB
Ven
HFL
56%
MI 0.7
TIS 0.1
183
BrV
BrA
Cine
2.2

Fig. 2.3 Brachial artery (BrA) and vein (BrV). The less round, slightly compressed, anechoic structure on the left is the vein. The very circular, not compressed, anechoic structure to the right is the artery

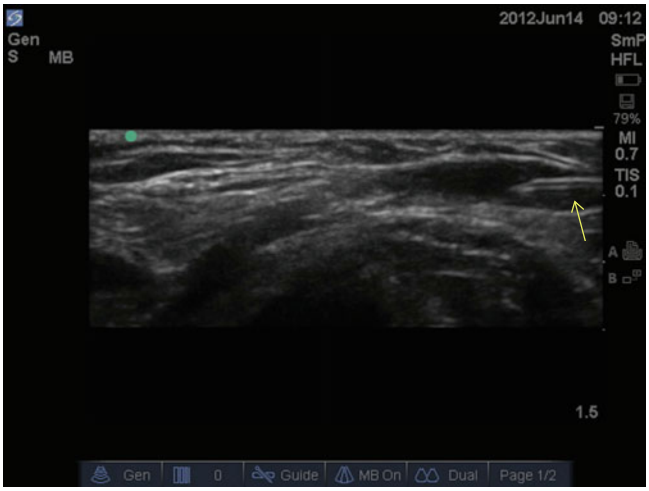
2012Jun14 09:12
Gen
S MB
SmP
HFL
79%
MI 0.7
TIS 0.1
1.5
Gen
0
Guide
MB On
Dual
Page 1/2

Fig. 2.5 Needle tip in vein seen in a longitudinal orientation

2.5 Complications

Inadvertent puncture of artery. Veins should be thin walled, compressible, and have no pulsations.

2.6 Pearls and Pitfalls

- Mistaking the midshaft of the needle for the needle tip. If this occurs, the needle tip is actually deeper than expected. The ultrasound machine will plot a hyperechoic "dot" on the screen for the needle tip, as long as it crosses the ultrasound beam at any point. This same "dot" will appear whether the tip is directly centered under the beam or *any* segment of the needle shaft is intersecting the beam. This can be visually deceiving and makes this procedure difficult to grasp.
- Very slow movements of the needle and the probe are important for keeping the needle tip in view. Once the needle tip is identified, the probe should be fanned forward (away from operator) just slightly and then the needle advanced until the needle tip comes into view again. This is repeated until the needle is securely moved further into the vein.

Selected Reading

Brannam L, Blaivas M, Lyon M, et al. Emergency nurses' utilization of ultrasound guidance for placement of peripheral intravenous lines in difficult-access patients. Acad Emerg Med. 2004;11:1361–3.
Constantino TG, Parikh AK, Satz WA, et al. Ultrasonography-guided peripheral intravenous access versus traditional approaches in patients with difficult intravenous access. Ann Emerg Med. 2005;46(5):456–61.
Dargin JM, Rebholz CM, Lowenstein RA, et al. Ultrasonography-guided peripheral intravenous catheter survival in ED patients with difficult access. Am J Emerg Med. 2010;28(1):1–7.
Ma JO, Mateer JR, Blaivas M. Emerg ultrasound. 2nd ed. New York: McGraw-Hill Companies, Inc.; 2008.
Saul T, Rivera M, Lewiss R. Ultrasound image quality. ACEP News. 2011;4:24–5.

Central Venous Line Placement: Internal Jugular Vein, Subclavian Vein, and Femoral Vein

3

Kevin D. Ergle, Zachary B. Kramer, Jason Jones,
and Rohit Pravin Patel

3.1 Indications

- Volume replacement
- Emergent venous access
- Administration of caustic medications: vasopressors, calcium chloride, hypertonic saline, high dose of potassium
- Dialysis catheter placement (hemodialysis)
- Nutritional support (total parenteral nutrition)
- Long-term antibiotics
- Chemotherapy
- Plasmapheresis
- Frequent or persistent blood draws or intravenous therapy when unable to establish peripheral access due to edema or other causes
- Jugular and subclavian: Central venous pressure monitoring, transvenous pacing wire introduction, pulmonary artery catheterization

3.2 Contraindications

- Absolute
  - Infection at site of insertion
  - Distorted anatomy/landmarks (prior surgery, radiation, or history of thrombus in the specified vein)
  - Subclavian only: Trauma to the ipsilateral clavicle, anterior proximal rib, subclavian or superior vena cava vessels
- Relative
  - Morbid obesity
  - COPD
  - Children less than 2 years (higher complication rates)
    - Coagulopathy (although ultrasound-guided internal jugular can be done in this situation)
  - Agitated or moving patient
  - Jugular only: Trauma to the ipsilateral clavicle, anterior proximal rib, subclavian or superior vena cava vessels
  - Jugular and subclavian: Inability to tolerate potential pneumothorax of the ipsilateral thoracic cage
    - Pneumothorax or hemothorax of the contralateral thorax
    - Patients receiving ventilatory support with high-end expiratory pressures (temporarily reduce the pressures)
  - Femoral only: Intra-abdominal (or retroperitoneal) hemorrhage

3.3 Materials and Medications

- Central venous catheter tray or bundle: single/double/triple/quadruple lumen, dialysis catheter, large-bore introducer (for transvenous pacing or pulmonary artery catheter kit)
- Sterile gloves
- Sterile drapes or towels
- Sterile gown
- Hat/hair cap and mask with eye protection
- Antiseptic solution with skin swabs (e.g., chlorhexidine)
- Sterile saline flushes (one 30 mL syringe or three 10 mL syringes)
- Lidocaine 1 %

K.D. Ergle, MD
Department of Internal Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: kevin.ergle@medicine.ufl.edu

Z.B. Kramer, MD
Emergency Medicine Resident, University of Florida College of Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: zkramer@ufl.edu

J. Jones, MD • R.P. Patel, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: jasonjones@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_3

- Sterile gauze
- No. 11 blade scalpel
- Dressing (sterile waterproof transparent dressing or sterile 4×4 gauze with tape)
- Sterile biopatch
- Suture material with needle driver if needed
- Transducing line (optional)
- Sterile probe cover (if using ultrasound guidance)

3.4 Procedures

3.4.1 Internal Jugular Vein Access Procedure

1. Obtain informed consent if not emergently indicated procedure.
2. Obtain supplies and prepare the room, ensuring that all supplies are within operator reach prior to placing gown and commencing the procedure. Include a sterile ultrasound sheath on the sterile field if ultrasound is being used.
3. Raise bed to a comfortable height for the operator.
4. Place patient with head facing away from side of central line site (if using ultrasound, other positions may be preferred). Place patient in 15–20° Trendelenburg position to help fill the upper central veins and reduce the risk of air embolism.
5. Identify the anatomy. Palpate triangle made by the clavicle and sternal and clavicular heads of the sternocleidomastoid (SCM) muscle to identify the location of the internal jugular vein (Fig. 3.1). If using ultrasound guidance, identify optimal anatomical arrangement.
6. Wash your hands and wear sterile attire using aseptic technique, including cap, mask, gown, and sterile gloves.
7. Prepare the site from the clavicle to the ear and across the trachea with antiseptic solution. Allow the antiseptic (chlorhexidine or iodine) to fully dry.
8. Drape the site and patient with sterile towels and drapes included in most CVL bundles. Make sure to cover the whole area and bed.
9. Cover the ultrasound probe with a sterile sheath. This can be done solo or by holding the sterile ultrasound sheath and having an unsterile assistant hold the probe so that the probe can be covered by the sheath.
10. Prepare the kit by checking the guide wire and flushing the tubing and lines with saline included in the kit.
11. With a 25-gauge needle, use 1 % lidocaine to anesthetize the skin at the apex of the triangle made by the SCM and clavicle. Aspirate to make sure the operator is not in a vessel and make a superficial wheel for the insertion site.
12. Preferred method is with ultrasound guidance (see steps 13–17 and Sect. 3.6 for description of ultrasound guidance). If performing without ultrasound, palpate the carotid artery and insert needle lateral to the artery at the apex of the triangle formed by the SCM, aiming toward the ipsilateral nipple at an angle 30–45° above the horizontal plane (Fig. 3.1). Once blood returned, go to step 18.
13. Place sterile ultrasound gel over the insertion site. Use the ultrasound to identify vessel anatomy including internal jugular vein and carotid artery. Use the ultrasound probe to compress the vein, which is compressible as opposed to the carotid artery, which is not compressible (Fig. 3.2).
14. Prepare the insertion needle and syringe (if long and short needles are available, a short needle may be used to reduce posterior vein perforation) and prime the syringe by pulling back on the plunger prior to making the puncture.
15. Use the ultrasound probe to re-identify the patient's anatomy.
16. Ultrasound can be used in short axis or long axis (Fig. 3.3). Short axis is easier for novice operators due to increased ability to see the artery and vein but has a higher risk of posterior perforation if the needle tip is not visualized well. Once short axis of the vein is found, turning the probe 90° clockwise allows the operator to see the vein in long axis. The needle is better visualized in this view but technically more difficult and has less chance to penetrate the posterior wall. In patients with short necks, it may be difficult to obtain long-axis view and needle insertion in the limited space.
17. Insert the needle using the ultrasound guidance with dynamic approach preferred (see Sect. 3.6 for specifics). Make sure to aspirate while inserting the needle to identify when the venous access is obtained. The needle tip should be visualized through the whole process.
    - If using the static approach (see Sect. 3.6), insert needle lateral to carotid pulsation as this is where the vein anatomically is located. Standard method is to insert the needle as far back as the depth the vessel is visualized (e.g., if the vein is visualized 2 cm below skin surface, the needle should be inserted 2 cm behind the probe at a 45° angle).
    - If inserting the needle ~3 cm does not achieve access, gently withdraw the needle toward the surface of the skin while aspirating. Avoid withdrawing the needle completely from the skin. If needed, redirect the needle and advance until blood is aspirated. Cannulation of the vein often takes place while withdrawing the needle.
18. Hold the needle steady with your nondominant hand and remove the syringe, careful not to advance or withdraw the needle. You can place the base of your hand on the patient's chest to make your hand more stable during

this part of the procedure. Occlude the hub of the needle to prevent air embolus.

19. You may verify that you are in the vein by transducing pressure with a fluid column. The fluid should flow easily into the vein.
    - If the aspirated blood is pulsatile and moves up the column, withdraw the needle completely and apply pressure for 10–20 min while taking the patient out of Trendelenburg position (if nonemergent procedure).
20. Once it is verified that you are in the vein, insert the J-tip of the guide wire into the needle hub and advance into the vein. The J-tip can be straightened with a pinching motion (Fig. 3.4). Always keep one hand on the guide wire until it is removed from the patient. Monitor for arrhythmias as the guide wire is advanced toward the right atrium.
    - If the guide wire does not flow easily, remove the guide wire and reattach the syringe, checking for blood flow.
    - If arrhythmia occurs, slowly withdraw the guide wire until the patient's native rhythm returns.
    - Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.
21. Remove the needle over the guide wire, making sure to always keep control of the guide wire.
22. Make an incision contiguous with the guide wire using a straight (No. 11) blade with the scalpel blade facing upward (away from the wire).
23. Advance the dilator over the guide wire in a twisting motion, keeping control of the guide wire.
    - The dilator only needs to go slightly beyond the anticipated depth of the patient's jugular vein. Do not advance the entire length of the dilator.
24. Withdraw the dilator and hold pressure over the wound site.
25. Advance the catheter over the guide wire while keeping control of the guide wire.
26. With the catheter inserted 10–12 in. from the skin insertion site, retract the guide wire until it comes out of the distal port. Maintain control of the guide wire and advance the catheter to the appropriate length. Usually catheters are inserted 15–16 cm from the right side and 18–20 cm from the left side (Fig. 3.6).
27. Flush each port of the catheter and check aspiration. If difficulty with aspiration or flushing, concern is raised for catheter malposition. The operator can change the depth slightly or twist the catheter and recheck.
28. At this time, an antibiotic ointment or biopatch may be applied to skin around the intersection with the lumen of the catheter. This step is based on local institutional guidelines.
29. Suture the line in place.
30. Enclose CVL site with sterile waterproof transparent dressing.
31. Confirm placement using chest X-ray (CXR). The tip of the catheter should be in the lower third of the superior vena cava (SVC) at the insertion of the SVC into the right atrium.

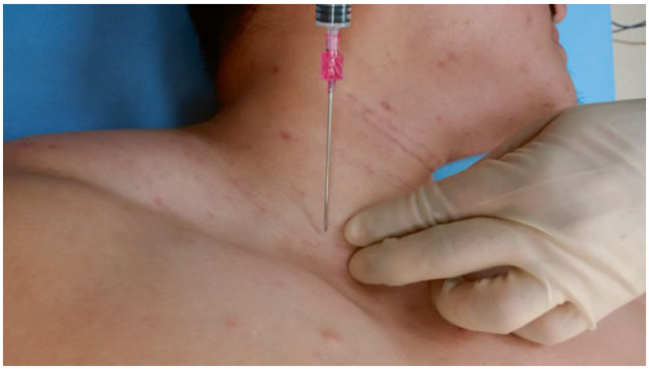

Fig. 3.1 Internal jugular blind approach; this would be the same location for probe placement if doing ultrasound guided

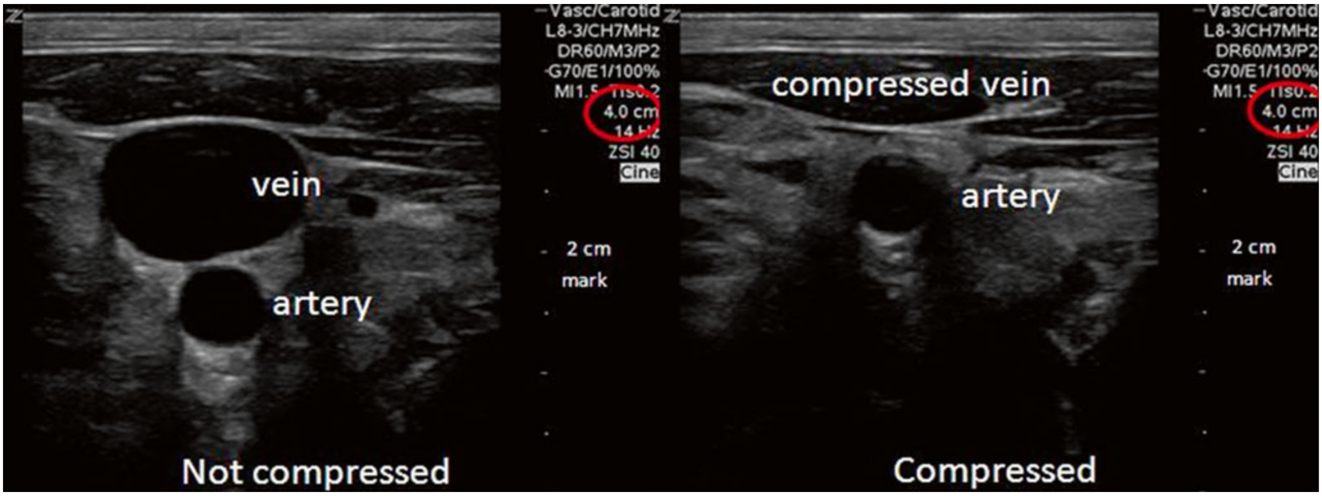
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G70/E1/100%
MI1.5 TIs0.2
4.0 cm
14 Hz
ZSI 40
Cine
vein
artery
2 cm mark
Not compressed
compressed vein
artery
Compressed

Fig. 3.2 Ultrasound showing internal jugular vein and artery with and without compression

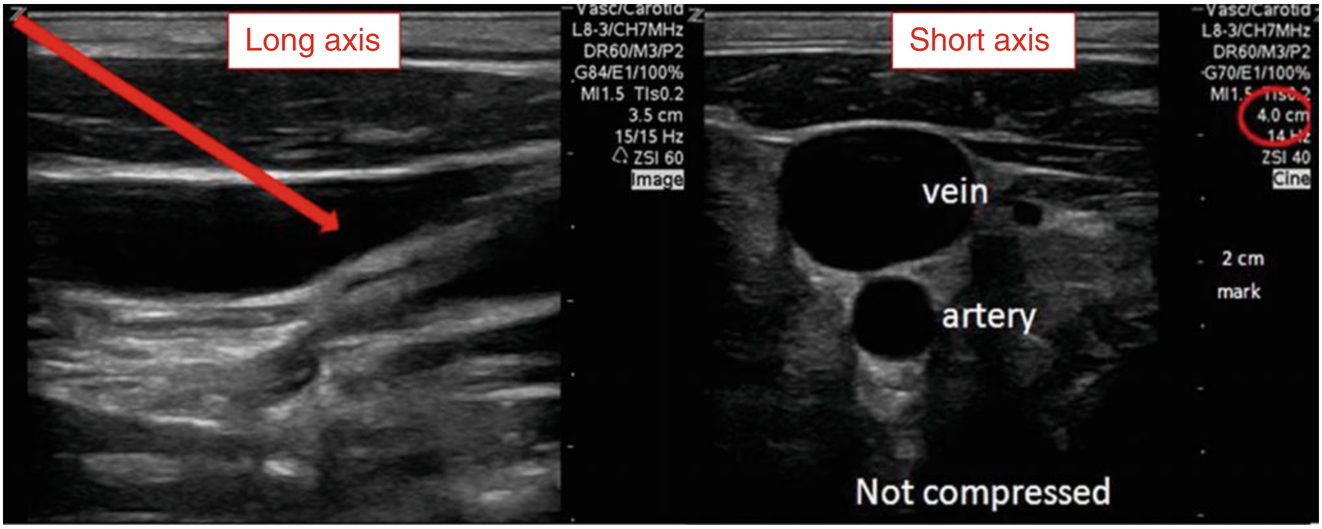
Long axis
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G84/E1/100%
MI1.5 TIs0.2
3.5 cm
15/15 Hz
ZSI 60
Image
Short axis
Vasc/Carotid
L8-3/CH7MHz
DR60/M3/P2
G70/E1/100%
MI1.5 TIs0.2
4.0 cm
14 Hz
ZSI 40
Cine
vein
artery
2 cm mark
Not compressed

Fig. 3.3 Long- and short-axis views of the internal jugular vein

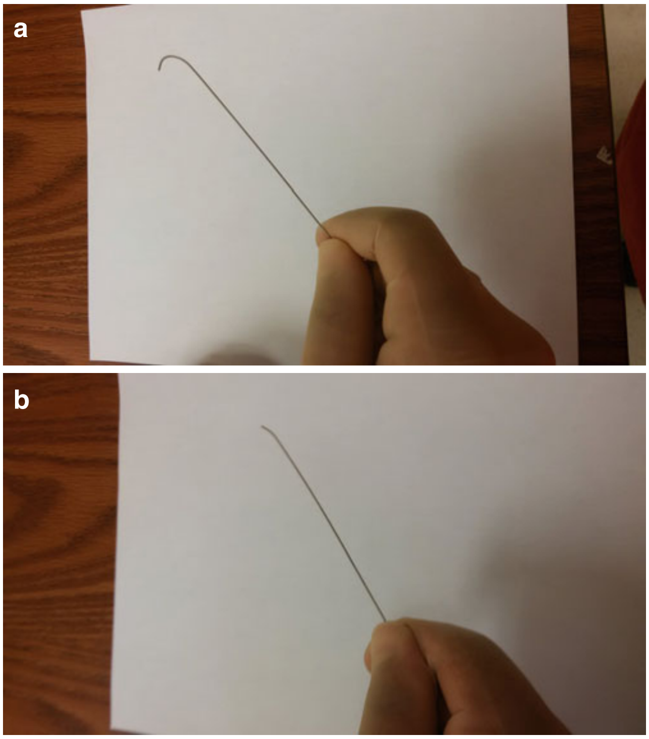

Fig. 3.4 (**a**, **b**) J-tip straightening using pinch/stretch method

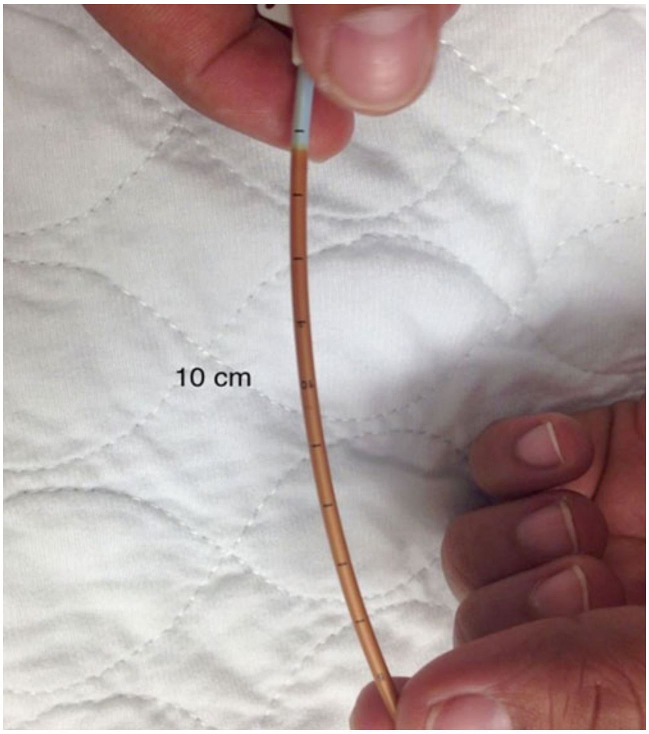

Fig. 3.6 Length: marking seen on typical central venous catheters; number indicates distance in centimeters from distal tip

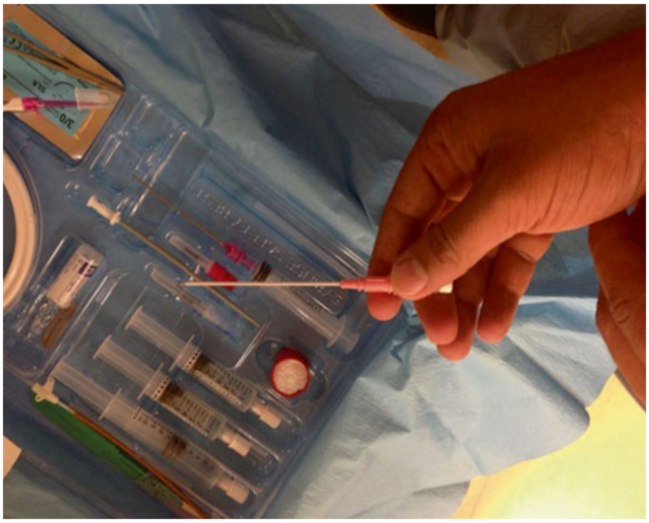

Fig. 3.5 Angiocath that can be used in difficult to cannulate/wire patients

3.4.2 Subclavian Vein Access Procedure

1. Obtain informed consent if not emergently indicated.
2. Raise bed to comfortable height for the operator.
3. Place patient in supine position and position so patient's head is at the top of the bed.
4. Place patient in 15–20° Trendelenburg position (if tolerated) to reduce risk of air embolism. Studies show this will also increase the size of the subclavian vein. Do NOT place towel between shoulder blades (arch shoulder back) as this has been shown to decrease vein diameter and affect reliability of accessibility. Keep shoulders at anatomical location (forward).
5. Prep area chosen from the anterior neck, clavicle, and upper chest (above nipple line) with chlorhexidine prep or iodine.
6. Open kit and place close to operator's dominant hand to allow for easy access. Diameter of catheter/kit used based on clinical situation:
    - Introducer or large bore if requiring large volumes of resuscitation
    - Triple lumen catheters for vasopressors
    - Introducers if anticipating pulmonary artery catheter or venous pacer
7. Operator should prepare with all aseptic techniques (e.g., hand washing) and maximal barrier precautions (e.g., sterile gowns, sterile gloves, caps, masks covering both mouth and nose, eye protection, and full-body patient drapes).
8. Once sterile and able to touch the inside of the CVL kit, the operator may want to retract the curved J-tip wire into the plastic loop sheath for easy directing into the introducer needle. The operator should also uncap all distal lumens and flush all ports with 3–5 cc of the sterile NS syringes to ensure no defects in the lumen of the catheter. Close all ports except the distal tip port (usually marked with words "distal tip") with the slide clamp.
9. Prep area chosen (right or left side) from the anterior neck, clavicle, and upper chest (above nipple line) with sterile chlorhexidine prep (this is the second cleaning).
10. Place full-body drape over patient with opening over selected side where needle will be inserted.
11. Needle insertion site options (Fig. 3.7):
    - One centimeter inferior to the junction of the middle and medial third of the clavicle
    - Just lateral to the midclavicular line, with the needle perpendicular along the inferior lateral clavicle
    - One fingerbreadth lateral to the angle of the clavicle
12. Anesthetize needle insertion site with 5–10 mL of 1 % lidocaine superficially (make sure to pull back on needle syringe to ensure operator is not in the vein or artery).
    - Never place equipment on a patient.
13. Prepare the needle and syringe by placing the long needle on the syringe. Make sure to break seal of syringe by pulling back on the plunger of the syringe prior to making incision with needle.
14. Turn patient's head to opposite of CVL placement and retract ipsilateral shoulder down to improve clavicle-vein relationship. The retraction of the arm can be done a few steps earlier and can be held in position using a person or tape/restraints.
15. Direct the insertion needle toward sternal notch in the coronal plane at an angle no greater than 10–15° while gently withdrawing the plunger of the syringe. Keep bevel of the needle facing up and in line with the numbers on the syringe until operator enters skin, then face bevel caudally to facilitate smooth progression of the guide wire down the vein toward the right atrium.
16. It helps to place nondominant hand (not holding the needle) on the sternal notch so operator can feel where sternal notch is and direct needle in that direction (Fig. 3.7).
17. NEVER increase the angle of the needle greater than 15° as pneumothorax may ensue.
18. Advance the needle under and along the inferior border of the clavicle, making sure the needle is virtually horizontal to the chest wall. Aim medially in the direction of the suprasternal notch, attempting to first aim for the clavicle then "walk" the needle below the clavicle.
19. Once under the clavicle, continue to advance the needle in a plane almost parallel to the skin approximately 2–3 cm until venous blood is freely aspirated into the syringe.
20. When venous blood is freely aspirated, disconnect the syringe from the needle, and immediately occlude the lumen to prevent air embolism and insert the guide wire. If the vein is difficult to locate, remove the introducer needle, flush it clean of clots, and try again. Change insertion sites after three unsuccessful passes with the introducer needle.
21. At this point, the hand holding the needle should be "set in stone." Use the patient's chest wall as a base to keep needle completely still as to not inadvertently advance or retract needle out of the vein.
22. Insert the guide wire through the needle into the vein with the J-tip directed caudally to improve successful placement into the subclavian vein.
    - Beware a return of red pulsatile blood. If this occurs, the wire is in an artery.
    - Beware aspirating air bubbles through the probing introducer needle. This indicates a pneumothorax.
23. Advance the wire until it is mostly in the vein or until arrythmia is seen on the cardiac monitor. Then, retract the wire 3–4 cm.
24. If the wire does not pass easily, remove the wire, reattach the syringe, and confirm that the needle is still in the

lumen of the vein before reattempting. The J-tip can be straightened with a pinching motion (Fig. 3.4).

- Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.

25. Use the tip of the scalpel to make a small incision just against the needle to enlarge the catheter entry site for the dilator and catheter.
26. Holding the wire in place, withdraw the introducer needle and place in needle holder.
27. Thread the dilator over the wire and into the vein with a firm and gentle twisting motion while maintaining constant control of the wire. If a large-bore introducer is placed, the dilator/introducer goes in one step, after the introducer is inserted, hold the wire in place and remove the dilator.
28. If operator is having difficulty threading the dilator, the skin incision with the scalpel may have been too superficial or small. It may help to enlarge this incision to avoid having the dilator get caught on superficial skin or connective tissue.

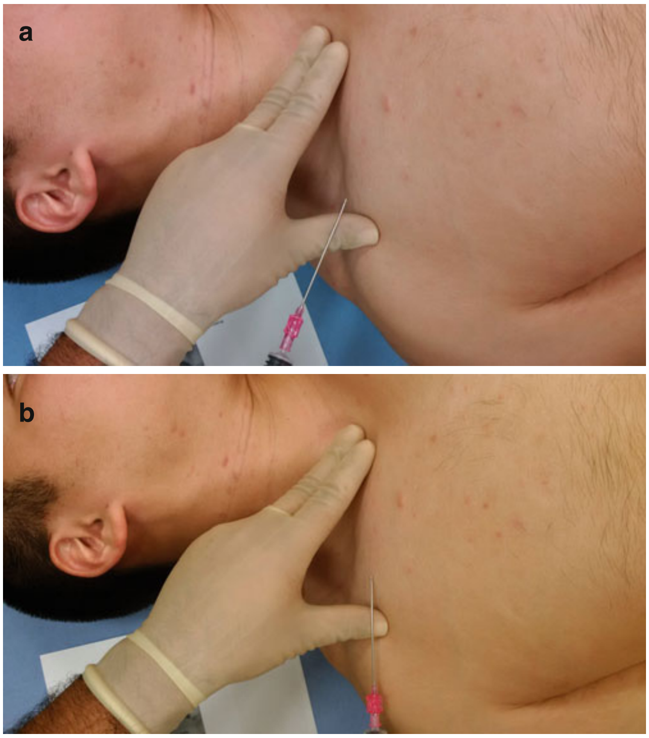

Fig. 3.7 Subclavian vein approach: wrong (**a**) and correct (**b**) angles to take when making skin puncture

29. It is helpful to have sterile gauze handy to apply pressure with the hand not holding the wire as the vein will now bleed profusely from around the wire secondary to dilation.
30. Thread the catheter until it is close to the skin insertion site. Then pull back on the guide wire until it shows outside of the distal port. Grasp the wire outside of the distal port and thread the catheter while holding onto the guide wire. Usually catheters are inserted 15–16 cm from the right side and 18–20 cm from the left side (Fig. 3.6)
31. Hold the catheter in place and remove the wire. After the wire is removed, occlude the open lumen.
32. Attach sterile saline syringe to the hub and aspirate blood. Take needed samples and then flush the line with saline and recap. Repeat this step with all lumens.
33. Place biopatch on skin around the intersection with the lumen of the catheter.
34. Suture the line in place.
35. Enclose CVL site with sterile waterproof transparent dressing.
36. Confirm placement using CXR. Tip of catheter should be in the lower third of the SVC at the insertion of SVC into right atrium (tip at right bronchiotracheal angle or up to 2.5 cm below bronchiotracheal angle).
    - Alternatively, ultrasound can be used for subclavian line access but only a few limited studies have confirmed this as to date so will not describe in detail (see below for typical ultrasound technique used). See references for more information.

3.4.2.1 Subclavian Vein Pearls and Pitfalls

- Inadequate landmark identification: operator should always palpate for landmarks and check anatomy prior to starting the procedure.
- Improper insertion position.
- Insertion of needle through periosteum.
  - Operator should NOT increase angle of needle to avoid the clavicle bone (this can cause a pneumothorax).
  - Operator should press on needle with downward pressure on chest wall to allow needle to maneuver under the clavicle without changing the angle of insertions of the needle.
- Taking too shallow a trajectory with needle.
- Aiming the needle too cephalad (aim for sternoclavicular junction).
- Failure to keep needle in place for wire passage: hand holding the needle should be planted on patient's chest for stabilization.

3.4.3 Femoral Vein Access Procedure

1. Palpate the patient's femoral artery below the inguinal ligament. This is usually found halfway between the anterior superior iliac spine (ASIS) and the midline of the symphysis pubis.
2. Trim overlying hair as necessary.
3. If ultrasound-guided approach is desired, use the linear probe (same as internal jugular) to detect the femoral vein at this location. The femoral vein will be easily compressible, while the femoral artery will be less compressible and pulsatile (Fig. 3.2).
4. Wash hands and use sterile technique to apply iodine or chlorhexidine solution (various forms available).
5. Open your femoral CVL kit and don cap, mask, sterile gown, and sterile gloves. Nonsterile assistants should wear a cap, mask, sterile gown, and sterile gloves. Flush all ports of your CVL kit with saline flushes and check for leaks or malfunction of catheter.
6. Under sterile technique, apply the drape over the area of insertion, and have an assistant extend the drape the length of the bed. Reapply sterile iodine or chlorhexidine at the site.
7. Anesthetize the skin overlying the femoral vein with lidocaine.
8. If ultrasound-guided approach is desired, have an assistant hold up the vascular ultrasound probe. Place your gloved hand through a sterile ultrasound sleeve and grasp the top of the ultrasound probe. Without breaking sterile technique, pull and invert the sterile ultrasound sleeve over the probe and cable. See Sect. 3.6 for detailed ultrasound-guided cannulation technique.
9. Insert the needle at a 45° angle, bevel down, directed superiorly, 1 cm medial to the palpable femoral artery pulse. Once the needle has broken the skin, aspirate by applying a small amount of continuous traction on the plunger of the attached syringe.
10. Advance smoothly and slowly until blood appears in the syringe. Stop once blood is aspirated. If femoral vein is not cannulated, withdraw your needle until just beneath the skin and redirect.
11. Hold the hub of the needle with thumb and forefinger to immobilize in place. Remove the syringe carefully. If blood appears arterial or pulsatile, remove the needle and hold pressure for 5–10 min.
    - Use the palm of your hand on the thigh to stabilize your hand. Not having your hand stable is a common mistake leading to needle movement out of the vein.
12. If blood appears venous (dark color, emerges as a continuous trickle, or transduced), cannulate the needle with the guide wire. Maintain a two-finger grip on the guide wire at all times. Advance the guide wire until approximately 15 cm remains.
    - Must keep handle of guide wire at ALL times and can be done through proper technique.
13. If guide wire does not advance easily, remove guide wire and reposition needle until blood aspirates easily. The J-tip can be straightened using a pinching motion if needed (Fig. 3.4)
    - Alternatively, the catheter/syringe found in most kits can be used as a bridge to guide wire placement. For the author, this has improved success when there is difficulty in wire placement. Use the same steps above with the catheter (Fig. 3.5) and when you have return of blood, advance the angiocath into the vein followed by insertion of the guide wire through the angiocath. This is especially useful in moving/agitated patients, patients who have collapsible veins due to hypovolemia, and patients who have abnormal anatomy and may have veins that take an abnormal angle shortly past the needle tip.
14. Using your scalpel, make a single 1/2 cm stab incision at the site of needle insertion to assist with dilator placement.
    - You can do the stab incision with or without the needle in place, but from experience, most novices have difficulty locating the correct stab location due to small amount of blood accumulation when the needle is taken out.
15. Remove your needle, carefully leaving guide wire in place. Apply dilator over guide wire and advance into the body with gentle pressure and a twisting motion in the same plane that you used to direct the needle.
16. Holding pressure at the insertion site with sterile gauze 4×4 pads, remove the dilator from the guide wire,

leaving the guide wire in place. Insert the central venous catheter over the guide wire until it fits snugly against the skin. Use the markings on the catheter to determine proper length placed (usually with femoral site you can "hub" the catheter) (Fig. 3.6).

17. Using a sterile saline flush, ensure that each lumen of the CVL draws blood easily and flushes easily. Carefully note any that do not and consider repositioning as needed. You can attempt moving catheter in or out a few centimeters or rotating the catheter and recheck. Apply caps to each open lumen of the CVL.
    - Remember to cover the introducer port if not using immediately. If you do not, it poses an infection and air embolism risk to the patient.
18. Suture the line in place.
19. Place an antibiotic biopatch or similar antimicrobial dressing.
20. Enclose CVL site with sterile waterproof transparent dressing.

3.4.3.1 Femoral Vein Pearls and Pitfalls

- Femoral central venous lines cannot accurately transduce central venous pressures.
- Asking the patient to perform a Valsalva maneuver has been shown to increase the width of the femoral vein by 1/3.
- The mnemonic NAVEL (Nerve Artery Vein Empty space Lymphatics) assists in remembering the order of femoral structures from lateral to medial.
- Traditionally, femoral venous lines were thought to have higher rates of infection than subclavian or internal jugular lines, but more recent analyses are challenging this belief.
  - Obesity is a more important risk factor for infection in femoral sites.

3.5 Complications

3.5.1 Jugular and Subclavian Complications

- Pneumothroax/Hemothorax
  - Prevention: Remove patient from ventilator before advancing the needle, choose the right side rather than left, and avoid multiple attempts when possible.
  - Management: Check postprocedure x-ray; if pneumothorax, arrange for thoracostomy depending on the size of the hemo-/pneumothorax.
- Catheter embolization
  - Prevention: Never withdraw a catheter past a needle bevel as this might shear off the catheter.
  - Management: X-ray the patient and contact specialist who can remove the embolized catheter.
- Arterial puncture: Hold compression if this occurs.
- Hematoma: Usually requires monitoring only.
- Thrombosis: This complication may lead to pulmonary embolism.
- Local site or systemic infection: Using maximal sterile precautions has been shown to greatly decrease rate of infection.
- Air embolism
  - May be caused by negative intrathoracic pressure when inspiration by the patient drawing air into an open line hub
  - Prevention: Be sure the line hubs are always occluded; placing the patient in the Trendelenburg position lowers the risk.
  - Management: The patient should be placed in Trendelenburg position with a left lateral decubitus tilt, which may prevent the movement of air into the right ventricle and onward into the left side of the heart. One hundred percent oxygen should be administered to speed the resorption of the air. If a catheter is located in the heart, aspiration of the air should be attempted.
- Dysrhythmias: Due to cardiac irritation by the wire or catheter tip. This can usually be terminated by simply withdrawing the line into the superior vena cava. One should always place a central venous catheter with cardiac monitoring.
- Lost guide wire: If the operator is not careful about maintaining control of the guide wire, it may be lost into the vein. This requires retrieval by interventional radiology or surgery and is an emergency.
- Catheter tip too deep: Check the postprocedure chest radiograph and pull the line back if the tip disappears into the cardiac silhouette.
- Catheter in the wrong vessel: Check the postprocedure chest radiograph for this complication; remove catheter and try again.
- Arterial puncture (subclavian only): The subclavian artery cannot be compressed; so, the subclavian approach should be avoided in anticoagulated patients.

3.5.2 Femoral Complications

- Arterial puncture: The femoral artery site can be compressed, so if punctured hold pressure.
- Hematoma: Usually requires monitoring only.
- Thrombosis: This complication may lead to pulmonary embolism.
- Catheter embolization.
  - Prevention: Never withdraw a catheter past a needle bevel which might shear off the catheter.
  - Management: X-ray the patient and contact specialist who can remove the embolized catheter.

- Lost guide wire: If the operator is not careful about maintaining control of the guide wire, it may be lost into the vein. This requires retrieval by interventional radiology or surgery and is an emergency.
- Local site or systemic infection: Using maximal sterile precautions has been shown to greatly decrease rate of infection.

3.6 Ultrasound Guided Cannulation: Tips for Each Approach

- Venous anatomy is best visualized using high-frequency (5–10 MHz) linear probe. Higher frequencies generate less penetration but better resolution.
- You can use the ultrasound to identify the location of the vessel prior to the procedure and utilize external landmarks during the procedure itself (static technique), or you can use the ultrasound to visualize cannulation of the vessel during the procedure (dynamic technique).
- Static view is advantageous in that the ultrasound transducer is not needed during the sterile portion of the procedure, but it does not allow for direct visualization of cannulation and guidance during the procedure.
- Dynamic view (preferred) allows for direct visualization during the procedure but requires more technique and requires use of transducer during the sterile portion of the procedure.
- The dynamic technique can be used in either a short-axis view, where a cross-sectional view of the vessel and needle is used, or a long-axis view, where a longitudinal view of vessel and needle is used (Fig. 3.3).
- The long-axis view allows for full visualization of the needle throughout the procedure and allows for better visualization and adjustment of needle depth. It is more difficult for lateral changes in positioning and tends to be more difficult technically.
    - – Key in this view is that once a good section of vein is obtained, do not move probe to visualize the needle; move the needle into the ultrasound view by slightly adjusting trajectory.
- The short-axis view allows for lateral changes in position but is not as good at visualizing depth throughout the procedure, as visualization of the needle is in cross-sectional imaging. Perforation of the posterior wall is more common in this view.
- When using the short-axis view, remember to position the ultrasound probe such that the field of the ultrasound intersects the vessel (internal jugular, subclavian, femoral) at the anticipated site of insertion of the needle into the vein. Remember that the needle is only visualized as it intersects the plane of the ultrasound.
- When using the long-axis view, make sure to visualize the vessel with the ultrasound such that you can see the greatest diameter of the vessel along the entire length of the ultrasound probe. Keep the ultrasound steady during the procedure, and insert the needle at an angle at the lateral edge of the ultrasound probe. Using this technique, one can visualize the entire length of the needle.

3.7 Removing a Central Line

1. Place patient in supine or Trendelenburg position (for femoral removal can help decrease bleeding).
2. Remove suturing and dressing.
3. Jugular and subclavian: Have patient exhale and pull the line during the exhalation.
    - Exhalation increases intrathoracic pressure as compared to atmospheric pressure, thereby reducing the risk of air thromboembolism.
4. Hold pressure for approximately 1 min to stop bleeding.
5. Dress with a sterile dressing.
6. If central line-related infection is suspected, cut off the tip with sterile scissors and send for culture.

Selected Reading

Internal Jugular Vein Access

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003a;348(12):1123–33.

Mimoz O, et al. Chlorhexidine-based antiseptic solution vs alcohol-based providone-iodine for central venous catheter care. Arch Intern Med. 2007;167(19):2066–72.

Noble V, et al. Manual of emergency and critical care ultrasound. Cambridge: Cambridge University Press; 2007. p. 196–204.

Parry G. Trendelenburg position, head elevation and a midline position optimize right internal jugular vein diameter. Can J Anaesth. 2004;51(4):379.

Vesely T. Central venous catheter tip position: a continuing controversy. J Vasc Interv Radiol. 2003;14(5):527.

Subclavian Vein Access

Elliott TS, Faroqui MH, Armstrong RF, Hanson GC. Guidelines for good practice in central venous catheterization. Hospital Infection Society and the Research Unit of the Royal College of Physicians. J Hosp Infect. 1994;28(3):163–76.

Fortune JB, Feustel P. Effect of patient position on size and location of the subclavian vein for percutaneous puncture. Arch Surg. 2003;138(9):996–1000; discussion 1001.

Fragou M, et al. Real time ultrasound-guided subclavian vein cannulation versus the landmark method in critical care patients: a prospective randomized study. Crit Care Med. 2011;39(7):1–6.

Kilbourne MJ, Bochicchio GV, Scalea T, Xiao Y. Avoiding common technical errors in subclavian central venous catheter placement. J Am Coll Surg. 2009;208(1):104–9.

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003b;348(12):1123–33.

Femoral Vein Access

Dailey RH. Femoral vein cannulation: a review. J Emerg Med. 1985;2:367–72.

Lim T, Ryu H-G, et al. Effect of the bevel direction of puncture needle on success rate and complications during internal jugular vein catheterization. Crit Care Med. 2012;40(2):491–4.

McGee DC, Gould MK. Preventing complications of central venous catheterization. N Engl J Med. 2003c;348:1123–33.

Marik PE, Flemmer M, Harrison W. The risk of catheter-related bloodstream infection with femoral venous catheters as compared to subclavian and internal jugular venous catheters: a systemic review of the literature and meta-analysis. Crit Care Med. 2012;40:2479–85.

Swanson RS, Uhlig PN, Gross PL, et al. Emergency intravenous access through the femoral vein. Ann Emerg Med. 1984;13:244–7.

Pulmonary Artery Catheter

4

Rohit Pravin Patel and Marie-Carmelle Elie

4.1 Indications

- Prevention or treatment of multiorgan failure in high-risk patients
- Preoperative and postoperative management in high-risk patients with cardiac, pulmonary, or renal dysfunction
- Patients with anticipated large fluid shifts (sepsis, bleeding, burns, cirrhosis)
- Oliguria or hypotension not relieved by fluids
- Suspected cardiac event leading to shock
- For continuous SVO2 (central venous oxygenation) monitoring in shock
- To differentiate shock states
- For monitoring cardiac output in patients requiring high-positive end-expiratory pressure (>14 cm H20)
- Monitoring and management of complicated myocardial dysfunction or cardiogenic shock
- Congestive heart failure with poor response to afterload reduction and diuretic therapy
- Suspected tamponade or contusion from blunt chest injury
- Pulmonary hypertension with myocardial dysfunction
- Diagnosis of primary pulmonary hypertension
- Aspiration of air emboli
- Direct pulmonary artery administration of thrombolytic therapy

4.2 Contraindications

- Tricuspid or pulmonary valve mechanical prosthesis
- Right heart mass (thrombus or tumor)
- Tricuspid or pulmonary valve endocarditis
- Recurrent sepsis (catheter could serve as nidus for infection)
- Hypercoagulopathy (catheter could serve as site for thrombus formation)
- Patient known sensitivity to heparin (catheters with heparin coating)
- Electrocardiographic (ECG) monitoring encouraged in conditions of complete left bundle branch block (risk of complete heart block increased), Wolfe–Parkinson–White syndrome, and Ebstein's malformation (risk of tachyarrhythmias)

4.3 Materials

- Pulmonary artery or Swan–Ganz catheter
- Percutaneous sheath introducer and contamination shield
- Compatible cardiac output computer for measuring cardiac output by the bolus thermodilution method
- Injectate temperature sensing probe (bolus thermodilution method)
- Connecting cables
- Sterile flush system and pressure transducers
- Bedside ECG and pressure monitor system
- Appropriate ECG "slave" cables

4.3.1 Catheter Preparation

- Avoid forceful wiping or stretching of catheter to avoid injury to the thermistor wire circuitry; wiping the heparin coat may cause removal of the coating.
- In vivo calibration is required if in vitro calibration is not done; refer to the monitor operator's manual for detailed calibration instructions.
- Connect catheter's injectate and pressure-monitoring lumens to the flush system and pressure transducers; ensure all lines are free of air.
- Connect the thermistor to the monitor and confirm no fault messages appear.

R.P. Patel, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: rohitpatel@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_4

4.4 Procedure

1. Maintain sterile precautions including sterile cap, mask, gown, gloves.
2. Place a central venous line introducer and verify placement of introducer with chest radiograph; it is also acceptable to obtain the chest radiograph after the insertion of the pulmonary artery catheter if no complication was suspected with the central venous line introducer.
3. Cleanse the skin and introducer thoroughly with chlorhexidine.
4. Have assistant open pulmonary artery catheter kit in sterile fashion.
5. Remove pulmonary artery catheter from kit and have assistant hook up all the ports to the transducers and make sure readings are accurate as the catheter is being manipulated.
6. Gently lift the distal portion of the catheter up from the silicone gripper; do not pull the balloon through the gripper to avoid damage.
7. Have assistant check the proximal and distal ports for patency by flushing with sterile saline. Also have the assistant check the patency of the balloon with the syringe provided in the kit (Fig. 4.1). Check for major asymmetry and for leaks (optional) by submerging in sterile saline or water. Deflate balloon prior to insertion. Carefully wave the distal catheter segment up and down to confirm electrical continuity by observing a pressure tracing on monitor. Ensure proper readings, no information is sometimes better than wrong information. Make sure each port transduces appropriately *prior to* insertion.
8. Familiarize yourself with the catheter line markings. Each thin line indicates 10 cm from the tip and thick line indicates 50 cm from the tip. These are used in combination to indicate length from tip (Fig. 4.2).
9. Place the sterile plastic sleeve (lock side toward patient, Fig. 4.3) over the catheter after flushing all ports to further protect the catheter during manipulation.
10. The distal end of the catheter is inserted into the introducer hub of the central venous line and threaded to the superior vena cava. The catheter must be placed at least 30 cm into the introducer for the balloon to clear the distal end of the introducer prior to inflation. At *no point* should the catheter be withdrawn with the balloon inflated; ensure the assistant has deflated the balloon prior to withdrawal. The balloon assists in directing the catheter through the vascular system using the directional blood flow.
11. At 20 cm, the balloon should be inflated and catheter advanced through right atrium, past the tricuspid valve into the right ventricle, then past the pulmonary valve to the pulmonary artery. The waveform and pressure readings can guide you through the various locations (Fig. 4.4).
12. Once in the pulmonary artery, the catheter should be carefully and slowly advanced to wedge position. The balloon can be deflated and pulmonary artery tracings should reappear. If a wedge is obtained with less than the maximum recommended volume, the catheter should be withdrawn to a position where full inflation volume produces a wedge tracing. Avoid prolonged times when obtaining wedge pressure (2 respiratory cycles or 10–15 s), especially in patients with pulmonary hypertension.
13. General guidelines for distance necessary at various points include: right atrium 20–25 cm, right ventricle 30–35 cm, and pulmonary artery 40–45 cm; catheter usually wedges at 50–55 cm. These are dependent on the starting location you are using to advance the catheter (subclavian, internal jugular, femoral).
14. Once the catheter is in correct position, it should be locked into place with the plastic sleeve tip onto the hub of the introducer.
15. Correct placement is confirmed with chest radiograph (Fig. 4.5).

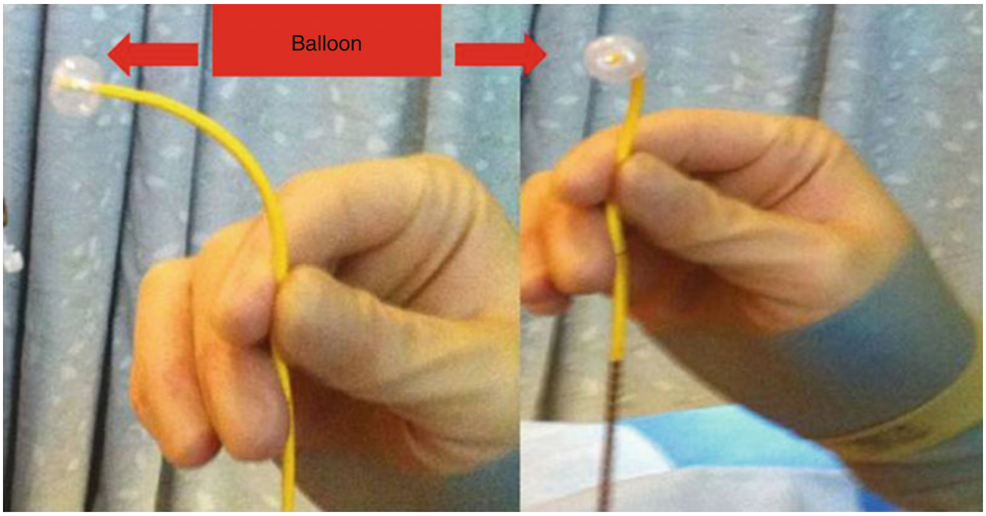

Fig. 4.1 Balloon inflation prior to insertion for evaluation of patency or leaks

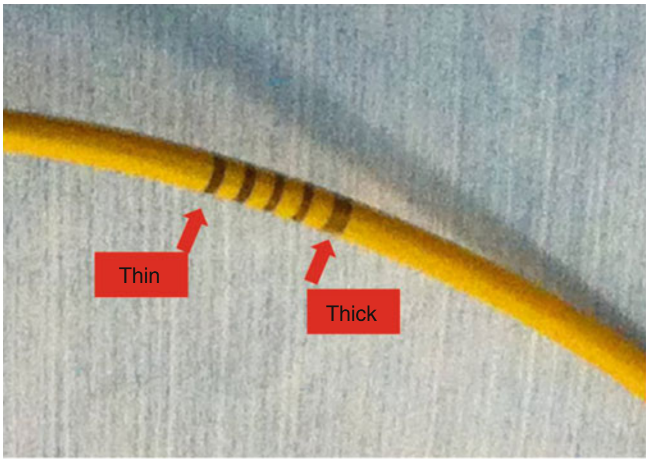

Fig. 4.2 Thick and thin markings found on the catheter representing length from distal tip

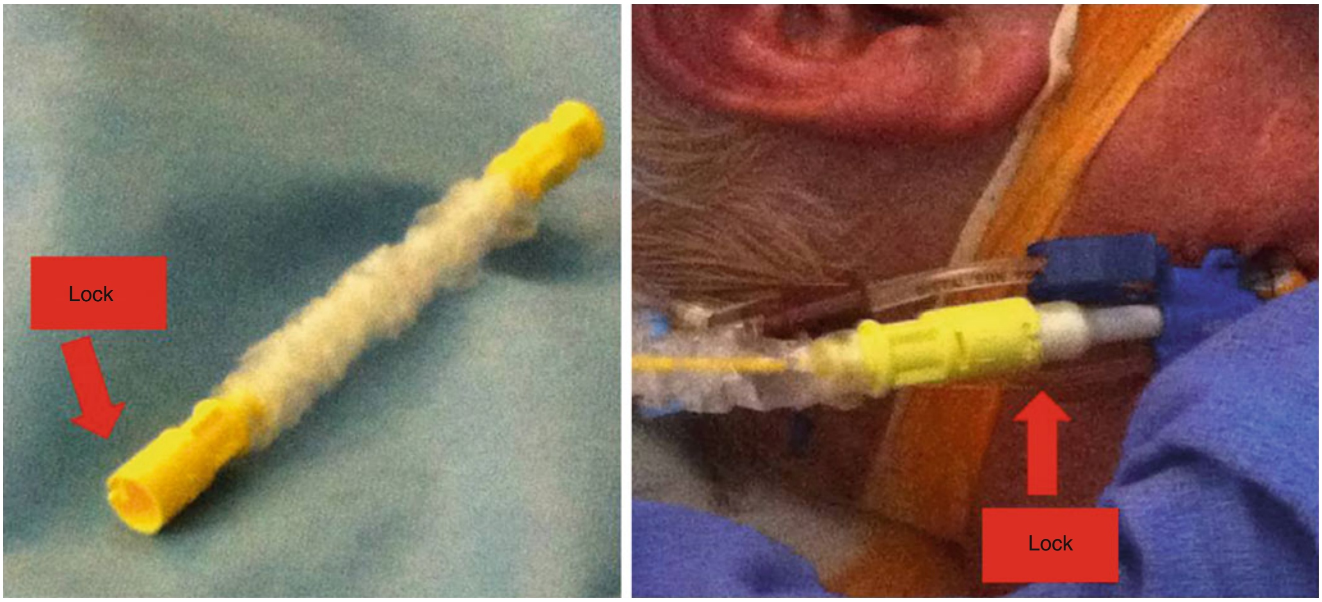

Fig. 4.3 Lock position on catheter for stabilization of catheter

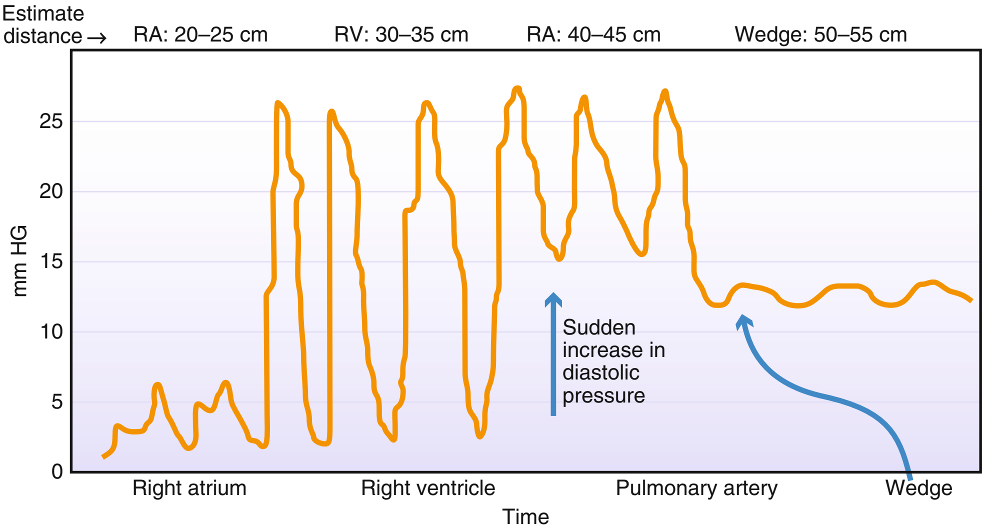

Fig. 4.4 Typical waveform seen at specific locations in the heart with associated estimated distances from catheter tip

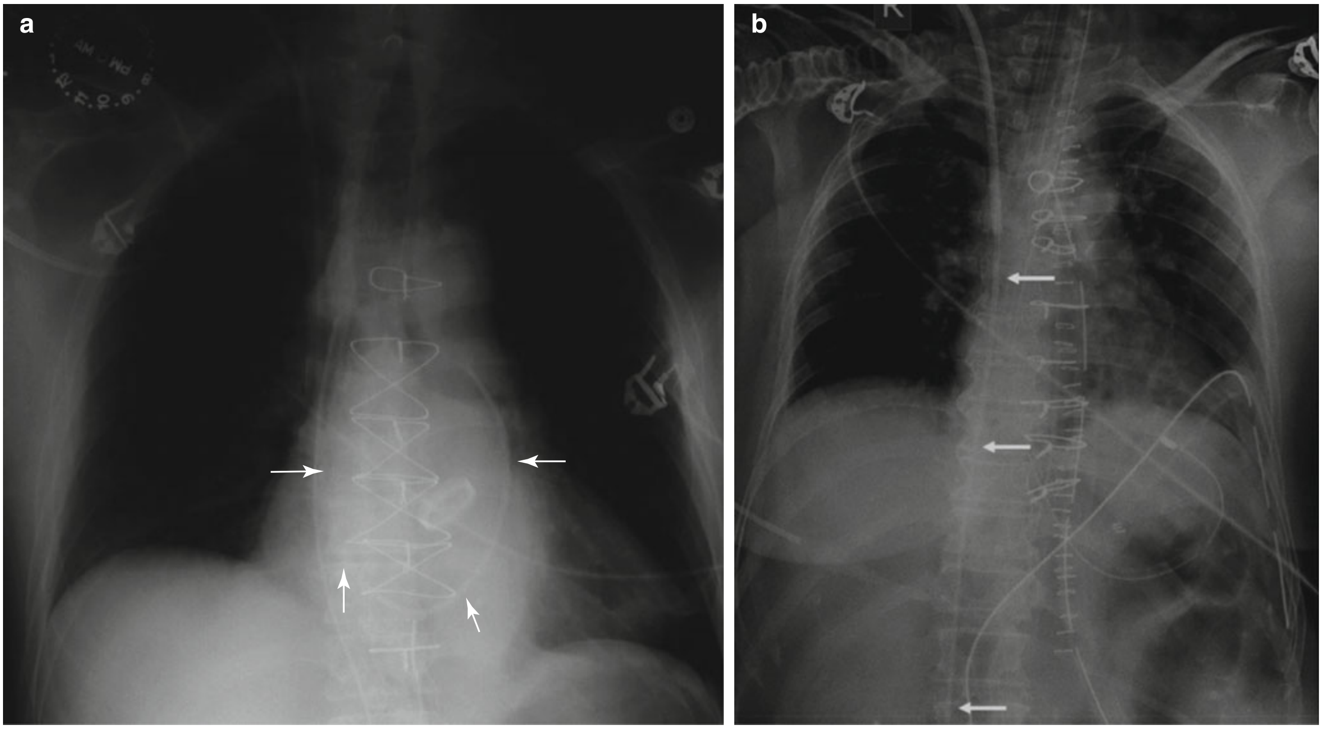

Fig. 4.5 Chest radiograph demonstrating correct (**a**) and incorrect (**b**) positioning of pulmonary artery catheter (*arrows*) (**a** Reproduced with permission from McGraw-Hill: Stead LG, et al. *First Aid for the Radiology Clerkship*. New York: McGraw-Hill; 2009; **b** Reproduced with permission from Wolters Kluwer: Jain SN. A pictorial essay: radiology of lines and tubes in the intensive care unit. *Indian J Radiol Imaging*. 2011;21(3):182–190)

4.5 Complications

- Arrhythmias: most are premature ventricular contractions that are self-limiting and resolve with advancement into pulmonary artery or withdrawal into atrium.
- Right bundle branch block: usually transient after positioning catheter into the pulmonary artery; if has already left bundle branch block may lead to complete heart block; should have temporary pacing equipment on standby.
- Knotting in the right ventricle (RV): risk increased in those with dilated cardiac chambers; a persistent RV tracing (15 cm beyond the point where initial RV tracing was observed) should alert you to this possibility.
- Pulmonary artery rupture: age >60 year, anticoagulation therapy, and presence of pulmonary hypertension increase risk of rupture; hemoptysis shortly after placement is indicative and management includes lateral decubitus positioning (bleeding side down), intubation with double lumen tube, and increasing positive end-expiratory pressure (PEEP).
- Infection.
- Pulmonary infarction: due to unintentional migration of distal tip.

4.6 Pearls and Pitfalls

- If the catheter requires stiffening during insertion, slowly perfuse the catheter with 5–10 mL of cold sterile solution as the catheter is advanced through a peripheral vessel.
- The incidence of complications increases significantly with periods of use longer than 72 h, so assess the need for the catheter on daily basis.
- Anticipate spontaneous catheter tip migration toward periphery of pulmonary bed; if a wedge tracing is observed when balloon is deflated, pull the catheter back.

Selected Reading

1. Edward Lifesciences. Pulmonary artery catheter Instruction manual. 2009. http://ht.edwards.com/resourcegallery/products/swanganz/pdfs/invasivehdmphysprincbook.pdf.
2. Edward Lifesciences. Pulmonary artery catheter Instruction manual. 2009. http://www.edwards.com/products/pacatheters/Pages/ThermodilutionCatheter.aspx.
3. Leatherman JW, Marini JJ. Clinical use of the pulmonary artery catheter. In: Hall JB, Schmidt GA, Wook LDH, editors. Principles of critical care. 3rd ed. New York: McGraw-Hill; 2005. p. 146–50.
4. Moran SE, Pei KY, Yu M. Hemodynamic monitoring: arterial and pulmonary artery catheters. In: Gabrielli A, Layon AJ, Yu M, editors. Civetta, Taylor, and Kirby's critical care. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2009.

Noninvasive Cardiac Monitoring: The Edwards Vigileo System

5

Dawood G. Dalaly and Rohit Pravin Patel

5.1 Indications

Cardiac output monitoring is indicated when trying to determine fluid responsiveness in patients. It assists in directing and assessing results of resuscitative efforts to ensure appropriate tissue perfusion. Although most catheters are systemically invasive, tools like the Vigileo (Edwards Lifesciences; Irvine, CA) are excellent noninvasive devices for determining values such as the stroke volume, stroke volume variation, stroke volume index, cardiac output, and cardiac index (Fig. 5.1).

D.G. Dalaly, DO
Department of Surgery,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: ddalaly@ufl.edu

R.P. Patel, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_5

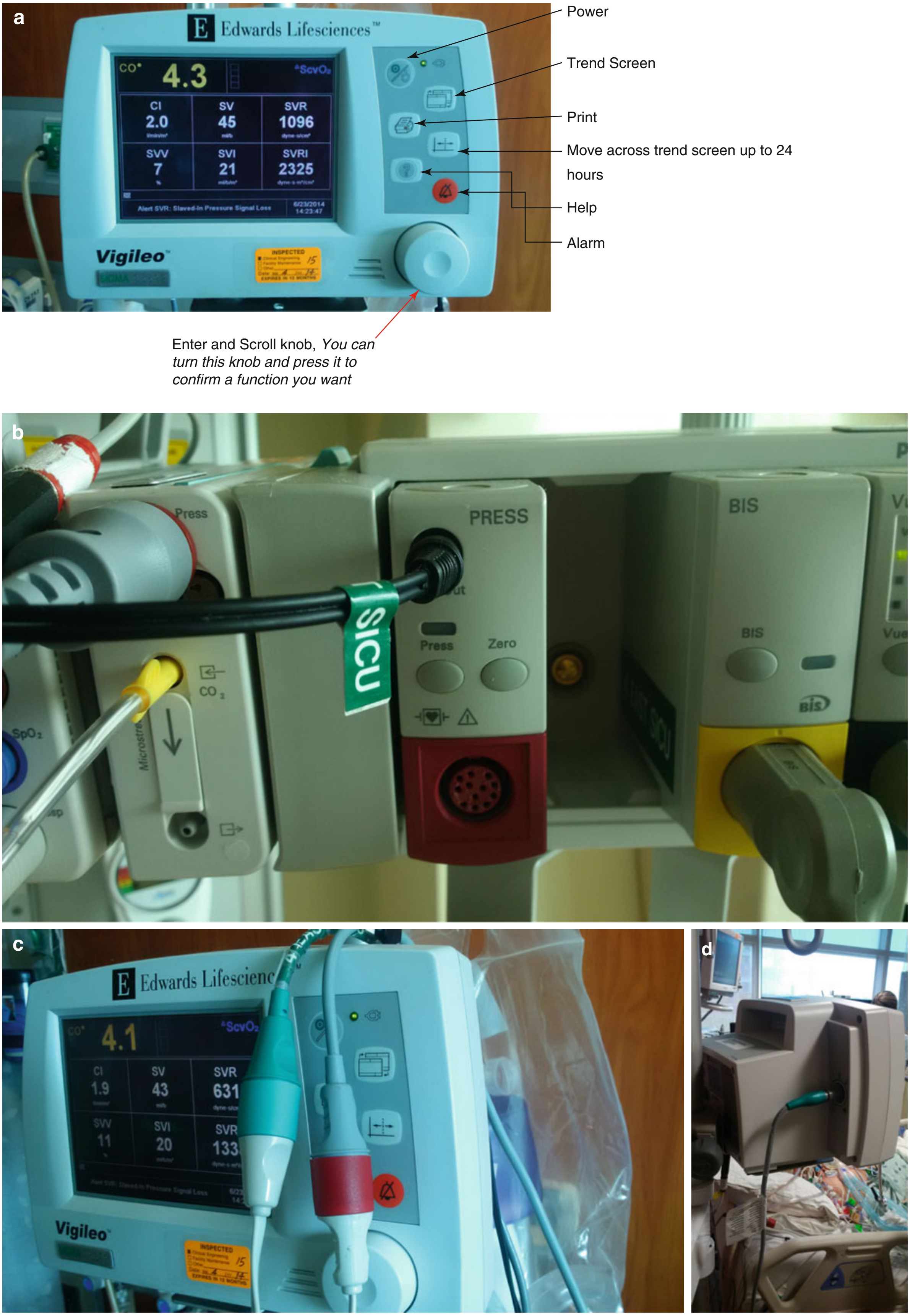
a
Power
Trend Screen
Print
Move across trend screen up to 24 hours
Help
Alarm
Enter and Scroll knob, *You can turn this knob and press it to confirm a function you want*
Edwards Lifesciences
Vigileo
CO
4.3
ScvO2
CI
2.0
SV
45
SVR
1096
SVV
7
SVI
21
SVRI
2325
Alert SVR: Slaved-In Pressure Signal Loss
b
Press
PRESS
Zero
SICU
CO2
BIS
SpO2
c
Edwards Lifescienc
Vigileo
4.1
d

Fig. 5.1 Typical monitoring devices and connecting devices from monitor to patient: (**a**) FloTrac monitor, (**b**) FloTrac red port connects to the heart monitor, (**c**) the FloTrac system connects to the arterial line on one end and the pressure bag/monitors on the other; monitor connections are color coded, (**d**) the green port connects to the green port from the FloTrac

5.2 Contraindications

There are no contraindications to noninvasive monitoring of the heart, although most systems will need to be connected to an arterial line catheter, therefore contraindications to arterial line placement must be determined. Current literature supports the use of noninvasive monitoring to those who are 100 % supported ventilation with tidal volumes ≥8 mL/kg. There is no support for use in those with spontaneous breaths or arrhythmia. Patients with significant variation in respiratory pattern may have results that are unreliable.

5.3 Definitions and Values (Table 5.1)

Table 5.1 Cardiac output monitoring definitions

Term	Reference range	Definition
Cardiac output	4–8 L/min	Volume of blood being pumped by the heart in 1 min
Cardiac index	2.5–4 L/min/m^2	Compares the amount of fluid being pumped by the heart with an individual's body surface area
Stroke volume	40–80 mL/beat	Volume of blood being pumped by the ventricle per beat
Stroke volume variation	10–15 %	Represents percentage of change between minimum and maximum stroke volumes and predictor of fluid responsiveness
Stroke volume index	33–47 mL/m^2/beat	Quantity of blood ejected from the heart per beat
Mixed venous saturation (SvO$_2$)	60–80 %	Percentage of oxygen bound to hemoglobin in blood returning to the right side of the heart; represents oxygen delivery and consumption at the tissue level. Usually obtained from pulmonary artery catheter
Central venous oxygen saturation (ScvO$_2$)	70 %	Surrogate marker for SvO$_{2,}$ usually obtained from internal jugular or subclavian catheters

5.4 Materials

- FloTrac (Edwards Lifesciences, Irvine, CA, USA)
- Vigileo monitor
- Pressure bag

5.5 Procedure

1. Connect FloTrac to arterial line and distal FloTrac port to pressure bag. Inflate bag to 300 mmHg.
2. Connect green FloTrac cord to green Vigileo cord and red FloTrac cord to arterial line port on heart monitor.
3. Turn the port on the FloTrac off to the patient and zero on your heart monitor as you would with an arterial line. At the same time, you should have pressed the "Enter" knob on the Vigileo system, scrolled it to "zero arterial pressure," and pressed the knob again for that function.
4. You should now have an arterial waveform on your heart monitor and your cardiac values on the Vigileo system.

5.6 Pearls and Pitfalls

- Some components of the values obtained are not reliable on spontaneously breathing patients and it is critical to check the ventilator waveforms for these breaths.
- Patients without adequate tidal volumes (at least 8 mL/kg) will also have unreliable values.
- Stroke volume variation usually is more reliable when greater than 13 % and indicates fluid responsiveness rather than when it is less than 13 % (similar to a low central venous pressure (CVP) being more informative than normal or high CVP levels).

Selected Reading

1. Alarcon LH, Fink MP. Chapter 13. Physiologic monitoring of the surgical patient. In: Brunicardi FC, Andersen DK, Billiar TR, et al., editors. Schwartz's principles of surgery. 9th ed. New York: McGraw-Hill; 2010.
2. Holcroft JW, Anderson JT, Sena MJ. Chapter 12. Shock and acute pulmonary failure in surgical patients. In: Doherty GM, editor. CURRENT diagnosis & treatment: surgery. 13th ed. New York: McGraw-Hill; 2009. Available at http://www.accesssurgery.com/content.aspx?aID=5212482. Accessed 22 Aug 2012.
3. Edwards Critical Care Education. Available at http://www.edwards.com/education/Pages/cceducationmap.aspx. Accessed 11 Dec 2012.

Peripheral Venous Cutdown

6

Jeffrey Kile, Katrina John, and Amish Aghera

6.1 Indications

- Distorted anatomy of peripheral venous access sites
- Unavailability of cannulable veins (e.g., in hypovolemia, burn victim, traumatic anatomy, sclerosed veins, etc.)
- Emergency venous access for infusion/transfusion
- Unavailability of central venous access or less invasive means peripherally

6.2 Contraindications

- Absolute
  - Availability of less invasive or less time-consuming means of vascular access
  - Overlying infection, traumatic tissue, burn, etc., at cutdown site
  - Traumatic injury proximal to cutdown site
- Relative
  - Coagulation disorders

6.3 Materials and Medications (Fig. 6.1)

- Sterile gloves
- Antimicrobial solution and swabs
- 4″×4″ gauze sponges
- Local anesthetic (1 % lidocaine 5 mL)
- 5-mL syringe
- Blunt needle
- 25- or 27-gauge needle
- Scalpel
- Vein dilator/lifter
- Peripheral intravenous catheter
- Curved hemostat
- 0-0 silk sutures or 4.0 nylon sutures
- Iris scissors
- Intravenous infusion tubing
- Adhesive tape

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_6

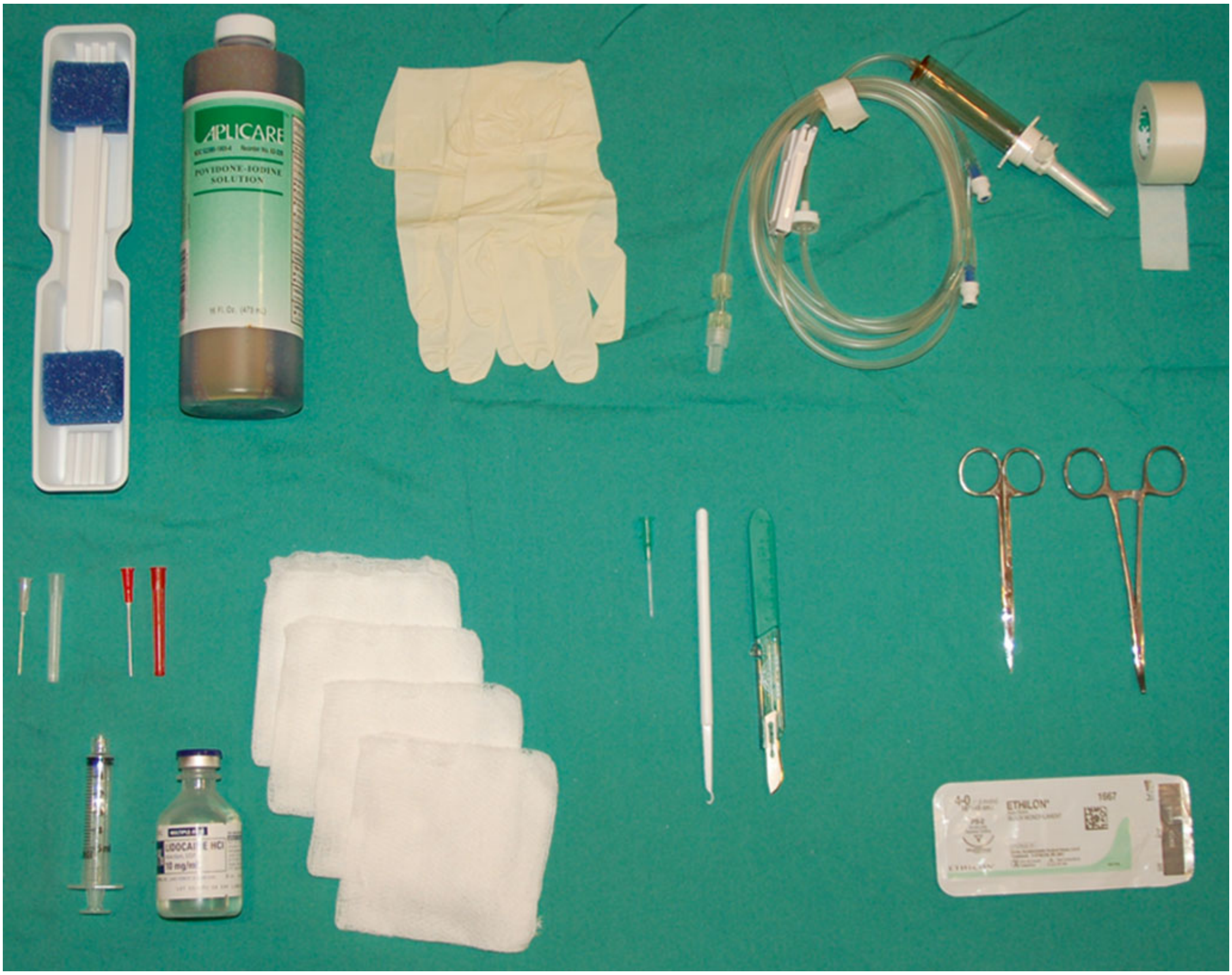

Fig. 6.1 Materials and medications

6.4 Choice of Vessel for Cutdown

- Greater saphenous vein: this vessel is the longest vein in the body, is predominantly subcutaneous, and is exposed with minimal blunt dissection just anterior to the medial malleolus at the ankle.
- Basilic vein: this vessel is reliably located 1–2 cm lateral to the medial epicondyle on the anterior aspect of the humeral region, typically catheterized just superior to the antecubital fossa approximately, and its diameter permits its localization relatively easily even in the hypotensive patient.
- Cephalic vein: this vessel runs anteromedially from the radial aspect of the wrist to the antecubital fossa, is superficial and large in diameter, and is most easily cannulated at the distal flexor crease in the antecubital fossa.

6.5 Procedure

6.5.1 Standard Venous Cutdown Technique

1. Apply antimicrobial solution liberally to the skin surrounding the incisional area.
2. Establish a sterile field by placing drapes around the incisional area.
3. Apply a tourniquet proximal to the planned cutdown site to maximize visualization of vein to be cannulated.
4. Inject local anesthetic to raise a small (0.5 cm) wheal using 25- or 27-gauge needle and then insert the tip of the needle through wheal to infiltrate skin superficial to the artery with approximately 4 mL of anesthetic.
   - Injection of local anesthetic into the vessel may precipitate arrhythmia, so draw back on the plunger prior to infiltration to ensure the tip of the needle is not inside the vessel.
5. Incise the skin with scalpel perpendicular to the course of the vein through all cutaneous layers until subcutaneous fat is visualized (Fig. 6.2).
   - Some practitioners prefer using a longitudinal incision to reduce the risk of transecting neurovascular structures, but this may not produce sufficient exposure of vein.
6. Using a curved hemostat or gloved finger, bluntly dissect the subcutaneous tissue to isolate and mobilize approximately 2–3 cm of the vein (Fig. 6.3).
   - A small self-retaining retractor or tissue spreader can be used in this step to improve visualization of vein if desired
7. Pass suture under the vein distal to the planned venous puncture site using hemostat to stabilize the vein and tie the suture over the vein (Fig. 6.4).
8. Pass a second suture under the vein proximal to the planned venous puncture site using hemostat (Fig. 6.5).
   - This step enables increased visualization, vessel control, and hemostasis during incision.
   - Leave the ends of both sutures long to facilitate maneuvering the vein.
9. Incise one-half to one-third of the diameter of the vein using a scalpel or iris scissors held at a 45° angle to the vessel (Fig. 6.6).
10. Grasping the proximal edge of the incision with a hemostat to apply counter traction (in a distal direction), insert the tip of the catheter into the venous incision (Fig. 6.7).
    - Do not force the catheter if it does not easily advance.
    - Catheter can be introduced directly through the skin incision or via skin puncture adjacent to the skin incision.
    - If the catheter lacks a tapered tip, cut the distal end of the cannula at a 45° angle to fashion a beveled tip.
11. Thread catheter into vein (Fig. 6.8).
12. Aspirate any air which may have entered the cannula during insertion.
13. Connect hub of catheter to intravenous tubing.
14. Tie the proximal suture around the vein just proximal to the venous incision, encircling both the vein and the intraluminal cannula with the suture.
15. Remove tourniquet.
16. Secure the catheter hub to the skin using nylon (4.0) sutures as follows. Take a 0.5 cm bite of skin under the catheter hub with the suture needle, tie several knots in the suture without pinching the skin, then tie a second set of knots around the hub of the catheter firmly.
17. Close the incision using nylon (4.0) sutures.
18. Dress the wound with appropriate self-adhesive sterile dressing or sterile gauze pads and adhesive tape.

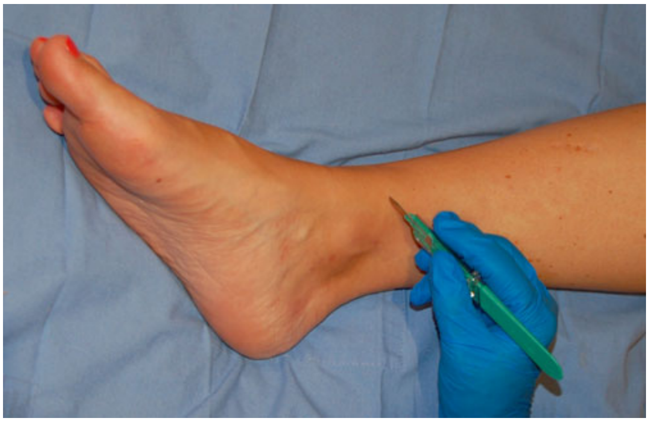

Fig. 6.2 Incision of skin

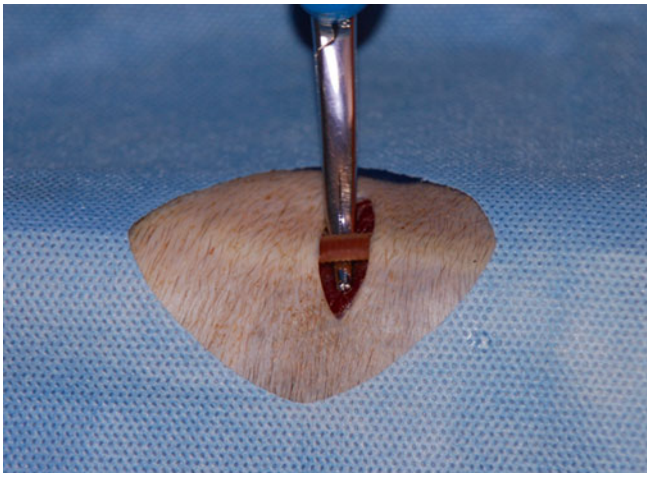

Fig. 6.3 Mobilization of vein

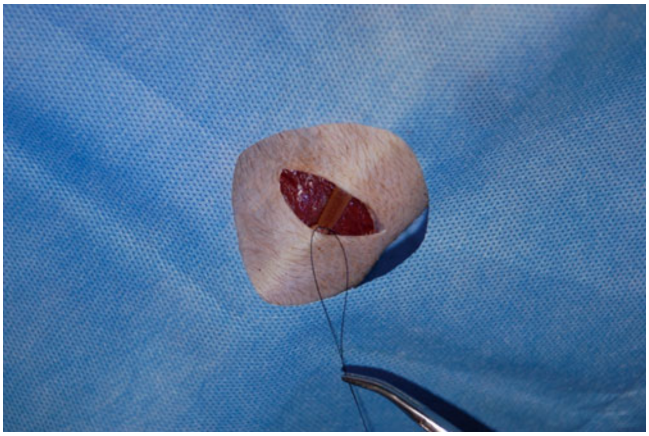

Fig. 6.4 Distal ligature tied around vein

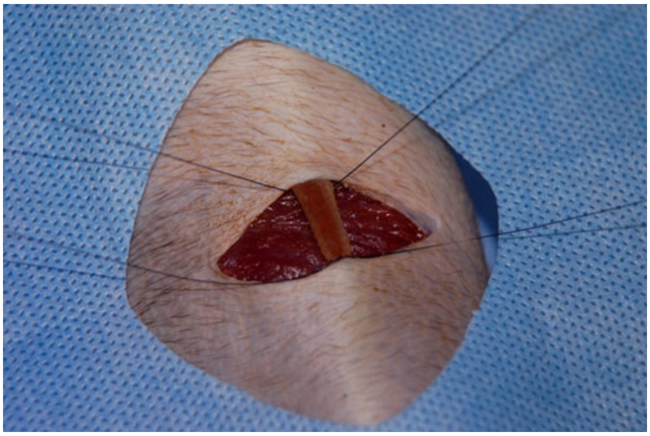

Fig. 6.5 Distal and proximal ligatures in place

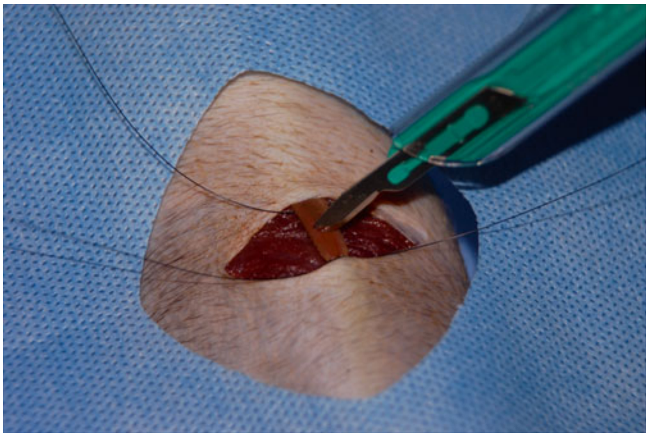

Fig. 6.6 Incision of vein

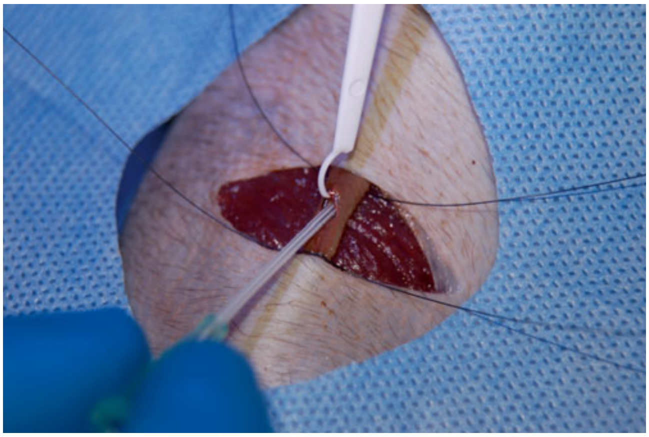

Fig. 6.7 Catheterization of vein

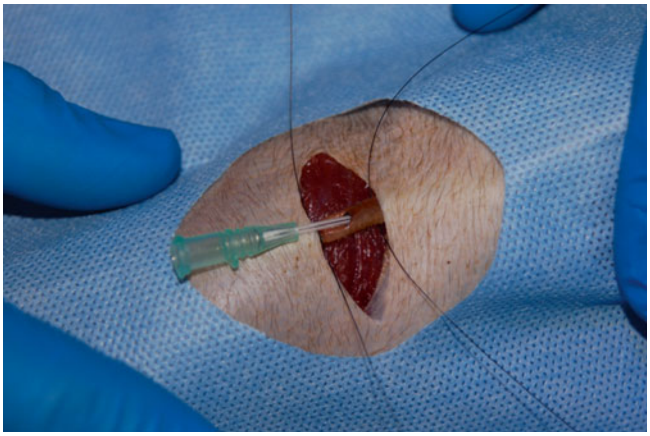

Fig. 6.8 Catheter threaded into vein

6.5.2 "Mini-Cutdown" Technique

(Perform steps 1–6 from the "Standard Venous Cutdown Technique," above, and then proceed with the steps below.)

1. Puncture the vein using a standard over-the-needle venous catheter.
    - Catheter can be introduced directly through the skin incision or via skin puncture adjacent to the skin incision.
2. Thread catheter into the vein over the needle.
3. Remove and discard the needle.
4. Aspirate any air which may have entered the catheter during insertion.
5. Connect catheter to intravenous tubing.

(Continue with steps 15–18 from the "Standard Venous Cutdown Technique," above.)

6.5.3 Modified/Guide Wire Technique

(Perform steps 1–6 from the "Standard Venous Cutdown Technique," above, and then proceed with the steps below.)

1. Insert the blunt end of the guide wire into the incised vein.
2. While stabilizing the guide wire at its insertion site, thread the dilator and sheath assembly over the free end of the guide wire until it is approximately one inch from the skin.
3. Grasp the free end of the guide wire protruding from the tail end of the assembly.
    - If it does not protrude from the tail end of the assembly, the guide wire must be removed sufficiently from the artery to be securely grasped. It must protrude visibly from the tail end of the dilator throughout the subsequent process of threading the dilator into the vein.
    - Never let go of the guide wire during this step, as insertion of the dilator and sheath assembly can otherwise push the guide wire completely into the vein.
4. Holding it firmly near its tip, thread the assembly over the wire into vessel with a gentle back-and-forth twisting motion.
5. Holding the sheath securely in the vein, remove and discard the dilator and guide wire.
6. Aspirate any air which may have entered the sheath during insertion.
7. Connect sheath to intravenous tubing.
8. Remove tourniquet.
9. Secure the sheath to the skin.
10. Close the incision using nylon (4.0) sutures.
11. Dress the wound with appropriate self-adhesive sterile dressing or sterile gauze pads and adhesive tape.

6.6 Complications

- Hematoma
- Infection
- Sepsis
- Phlebitis
- Embolization
- Wound dehiscence

6.7 Pearls and Pitfalls

- Fluids are infused most quickly via short, large-bore catheters.
- If the line is inserted for slow infusion of intravenous drugs, catheter lumen size is relatively insignificant.
- In larger children and adults, intravenous plastic tubing, small-bore pediatric feeding tubes, and Silastic catheters may be used as infusion catheters.
- Threading a 10-gauge intravenous catheter or intravenous tubing directly into the incised vein achieves excellent flow rates.
- If difficulty is encountered while threading the catheter into the incised vein, ensure an appropriately sized catheter has been used and that the vessel lumen has been correctly identified and that no false passage has been created in the adventitia.
- As compared to the standard venous cutdown technique, the mini-cutdown technique is easier and also preserves the vein, permitting repeated catheterization if necessary.
- As compared to the standard venous cutdown technique, the modified/guide wire technique reduces procedure time and increases the likelihood of vein salvage in the event of vessel transection.

Selected Reading

1. Chappell S, Vilke GM, Chan TC, et al. Peripheral venous cutdown. J Emerg Med. 2006;31:411–6.
2. Klofas E. A quicker saphenous vein cutdown and a better way to teach it. J Trauma. 1997;43:985–7.
3. McIntosh BB, Dulchavsky SA. Peripheral vascular cutdown. Crit Care Clin. 1992;8:807–18.
4. Nocito A, Wildi S, Rufibach K, et al. Randomized clinical trial comparing venous cutdown with the Seldinger technique for placement of implantable venous access ports. Br J Surg. 2009;96:1129–34.
5. Shockley LW, Butzier DJ. A modified wire-guided technique for venous cutdown access. Ann Emerg Med. 1990;19:393–5.

Part II

Airway and Vascular Procedures

7 Bag-Valve-Mask Ventilation

Braden Hexom and Tatiana Havryliuk

7.1 Indications

- Hypoxia
- Hypoventilation/apnea
- Rescue maneuver if failed intubation

7.2 Contraindications

- Absolute
  - Inability to ventilate due to lack of seal (thick beard, deforming facial trauma)
  - Inability to ventilate secondary to complete upper airway obstruction
  - Active, adequate spontaneous ventilation
- Relative
  - Full stomach (aspiration risk)
  - After induction and paralysis during rapid sequence intubation (aspiration risk)

7.3 Materials (Fig. 7.1)

- Bag valve mask (BVM) with reservoir
- Oxygen connector tubing
- Nasal pharyngeal airway/oral pharyngeal airway
- Lubricant jelly

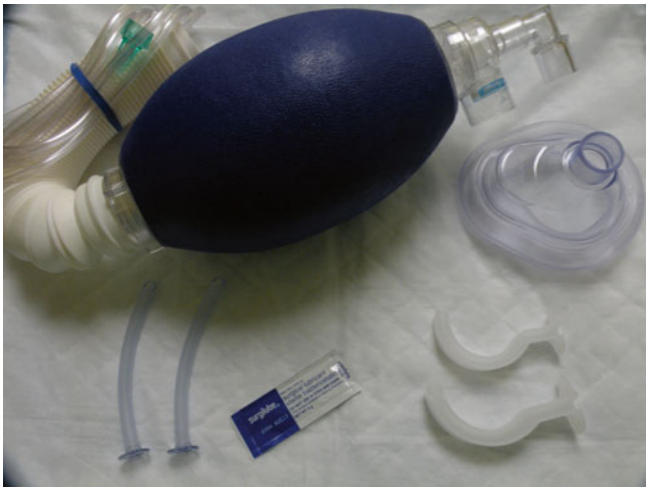

Fig. 7.1 BVM supplies: bag, mask, oral airways, nasopharyngeal airways, lubricant

B. Hexom, MD (✉)
Department of Emergency Medicine, Mount Sinai Hospital,
New York, NY, USA
e-mail: braden.hexom@mssm.edu

T. Havryliuk, MD
Department of Emergency Medicine, University of Colorado
Denver, Denver, CO, USA
e-mail: tatiana.havryliuk@ucdenver.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_7

7.4 Procedure

1. Position patient in "sniffing" position.
2. Open the airway with chin-lift/head-tilt or jaw thrust maneuvers.
3. Place airway adjuncts to maintain airway patency. Use oral airway (Fig. 7.2) in unconscious patients. Use nasal airway (Fig. 7.3) in semiresponsive patients.
4. Attach oxygen tubing to high-flow oxygen (15 L/min).
5. Place appropriately sized mask on patient's face covering the nose and mouth.
    - For one-handed technique (Fig. 7.4), use nondominant hand to make a "C" with index finger and thumb on top of the mask and form an "E" with the rest of the fingers using them to pull up on the mandible (E–C technique). Use the dominant hand to provide bag ventilations.
    - For two-handed (Fig. 7.5), two-person technique (preferred), make two semicircles with index fingers and thumbs of both hands on top of the mask and use the rest of the fingers to pull up on the mandible.
6. Consider the Sellick maneuver (cricoid pressure) to compress the esophagus against the cervical vertebrae, preventing gastric insufflation.
7. Ventilate patient providing reduced tidal volume breaths (500 mL) at a rate of 10–12 breaths per minute.
8. Give each breath gently over 1–1.5 s to avoid high peak pressures, avoiding gastric insufflation.
9. Prepare for definitive airway as dictated by the clinical scenario.

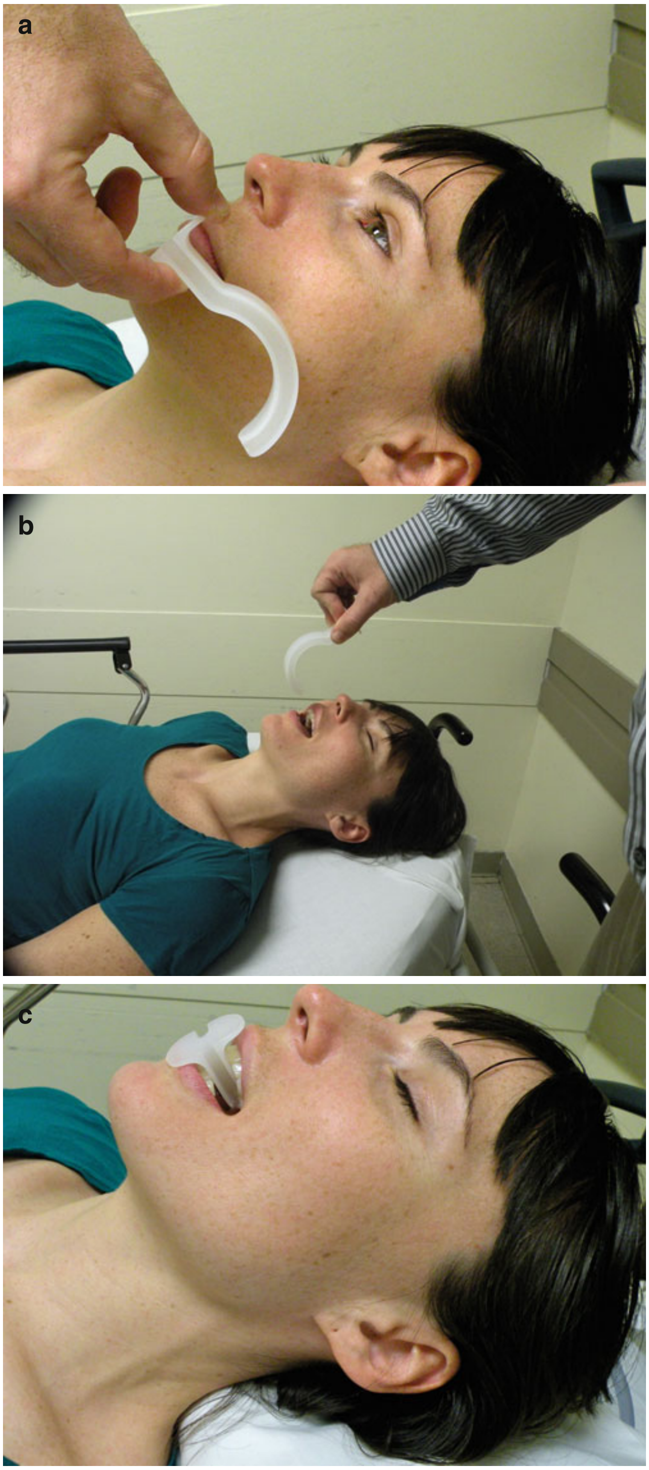

Fig. 7.2 (**a**–**c**) Oral airway insertion

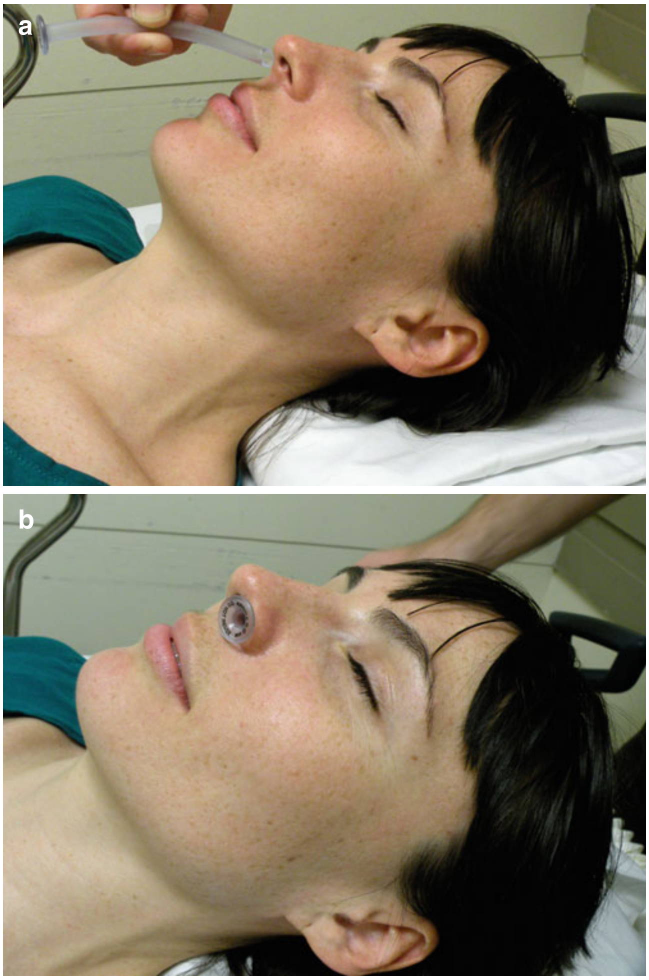

Fig. 7.3 (**a**, **b**) Nasal airway insertion

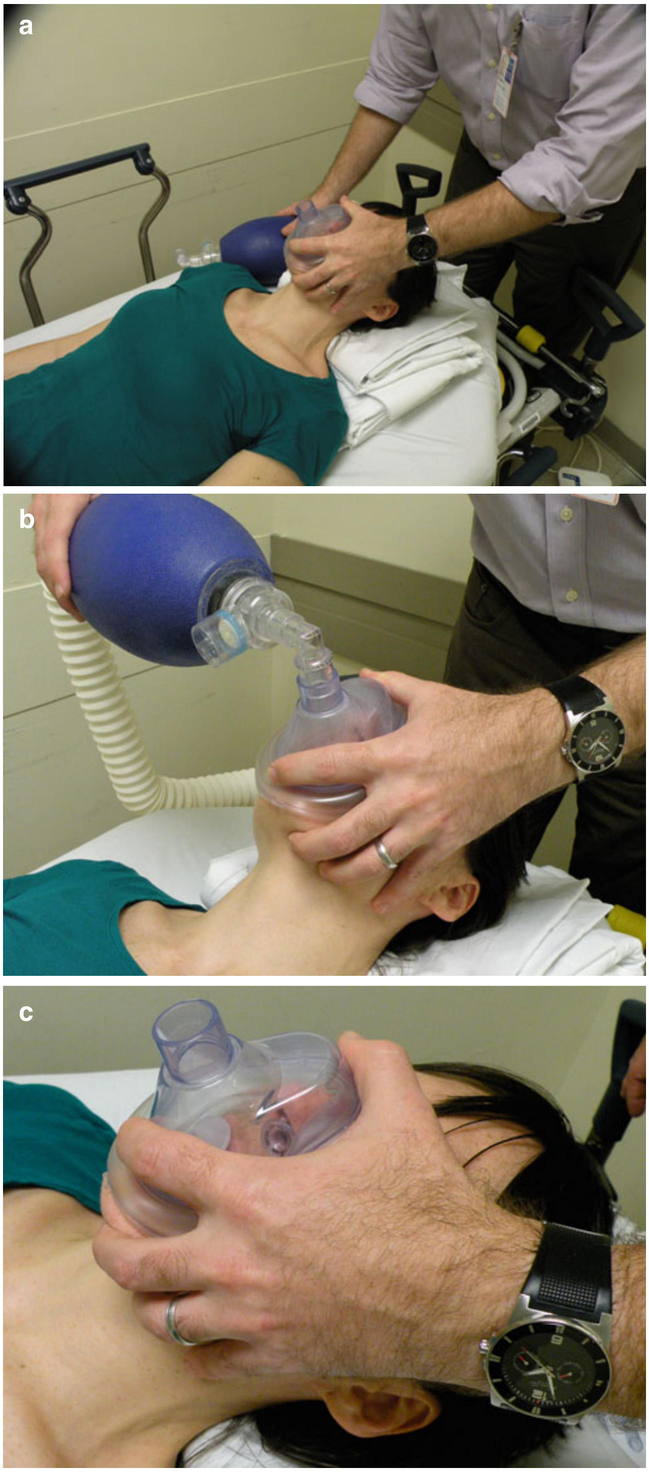

Fig. 7.4 (**a**–**c**) One-handed seal technique

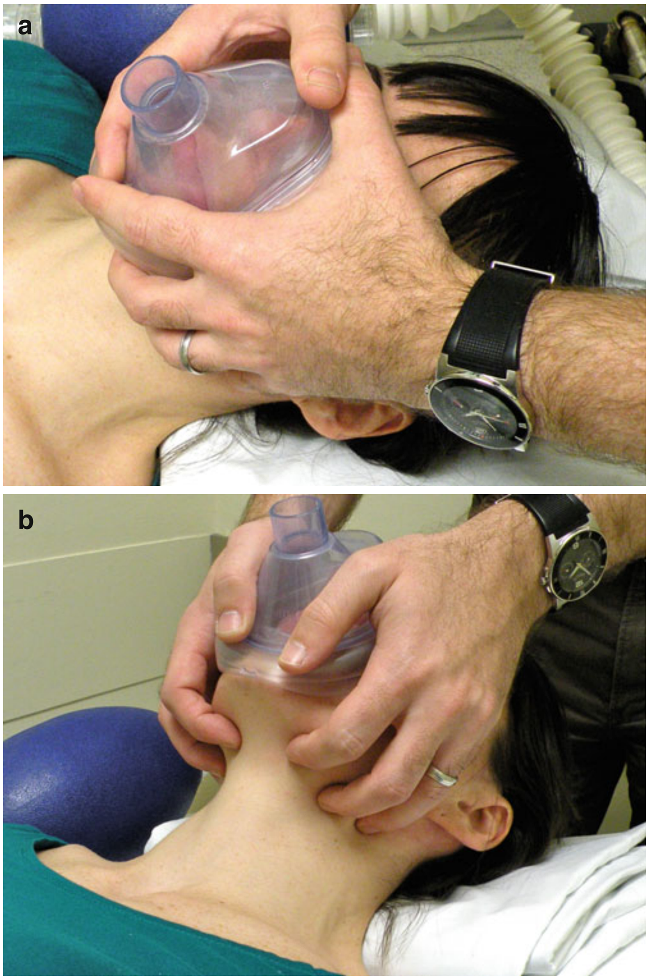

Fig. 7.5 Two-handed seal technique: (**a**) two semicircles, (**b**) alternative thumbs method

7.5 Complications

- Stomach inflation may lead to vomiting and aspiration.
- Increased positive thoracic pressure may cause decreased preload, worsening cardiac output, and/or hypotension.
- Hypoventilation (inadequate O_2 tidal volume, airway patency, or mask seal).

7.6 Pearls and Pitfalls

- Pearls
  - Use jaw thrust to open mouth for patients with possible cervical spine injury.
  - Use airway adjuncts whenever available, especially if prolonged BVM ventilation is anticipated.
  - Use lubricant jelly to insert nasal airway; do not insert in patients with severe facial trauma.
  - Mask should be placed on patient's face prior to attaching to bag.
  - Apply pressure to the bony part of the mandible only to avoid soft-tissue injury.
  - Provide just enough tidal volume to see a chest rise and deliver each breath gently over 1–1.5 s to prevent gastric insufflation.
  - Ensure good seal:
    - Select appropriate mask size.
    - Choose two-handed technique over one-handed, if possible.
    - Keep well-fitting dentures in place if present (and remove before intubation).
    - Lift the mandible toward the mask (as opposed to pushing the mask onto the face).
    - Rock the mask on face until no leak is present.
    - Apply K–Y jelly to beard to improve the seal.
- Pitfalls
  - Overcompression during the Sellick maneuver can compress the cricoid ring, preventing ventilation.
  - Cricoid pressure (Sellick maneuver) is not recommended during cardiac arrest resuscitation.
  - Difficult BVM ventilation: MOANS
    - *M*ask seal
    - *O*besity/obstruction
    - *A*ge
    - *N*o teeth
    - *S*tiff

Selected Reading

ECC Committee, Subcommittees and Task Forces of the American Heart Association. American heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. Circulation. 2010;122:S685–705.

Joffe AM, Hetzel S, Liew EC. A two-handed jaw-thrust technique is superior to the one-handed "EC-clamp" technique for mask ventilation in the apneic unconscious person. Anesthesiology. 2010;113:873–9.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. Philadelphia: Saunders Elsevier; 2010.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

8

Awake Orotracheal Intubation

Benjamin M. Mahon, Justin Bennett, and Lars K. Beattie

8.1 Indications

- Urgent but not emergent endotracheal intubation is required in a patient who:
  - Is awake
  - Is currently protecting his airway
  - Is not a candidate for a supraglottic airway (LMA)
  - The patient
    - Is predicted to have a difficult airway
    - Has structural abnormalities of the airway
    - Will not tolerate a period of apnea
    - May lose his airway (anaphylaxis, angioedema, traumatic airway)
- Patients requiring urgent but not emergent intubation in whom paralytics are contraindicated (i.e., allergies, myasthenic crisis)

8.2 Contraindications

- Absolute
  - Surgical airway indicated
  - Emergent crash airway needed
  - Obtunded patient
  - Allergies to medications needed (lidocaine, glycopyrrolate)
- Relative
  - Inability to maintain airway or tolerate secretions

8.3 Materials and Medications

- Suctioning equipment
- Intravenous access equipment, cardiac monitor, pulse oximetry, blood pressure cuff
- 4 % lidocaine solution
- 2 % viscous lidocaine jelly
- Nebulizer
- Mucosal atomization device
- 4×4 gauze
- Tongue depressor
- Glycopyrrolate/atropine
- Sedation: ketamine, propofol, Versed, and/or fentanyl
- Intubation equipment
- Backup emergency airway adjuncts
- Bag valve mask
- Laryngoscope, fiber optics, oral airway, etc.

8.4 Procedure

1. Preparation
   (a) Establish IV access.
   (b) Place the patient on a cardiac monitor with continuous pulse oximetry.
   (c) Keep backup RSI emergency airway medication and equipment at the bedside.
2. Administer 0.2–0.4 mg of intravenous glycopyrrolate (or a small dose (0.5–1 mg) of atropine to decrease secretions) 15 min prior to procedure.
3. Nebulize 2 mL of 4 % lidocaine with oxygen at 5 l O2 per minute to anesthetize the pharynx (Fig. 8.1).

B.M. Mahon, MD
Ponciana Medical Center,
Kissimmee, FL, USA
e-mail: drfaust2k9@aol.com

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University
School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_8

4. Use Yankauer suction (with the patient's assistance) to dry out the mouth as much as possible. Dabbing the tongue with gauze can assist in this step.
5. Continue preoxygenation.
6. Immediately after the nebulized solution is applied, give the patient a "lidocaine lollipop" (Fig. 8.2).
   (a) A 2 ml dollop of 2 % viscous lidocaine is to be placed on the end of a tongue depressor and is given to the patient to place in his mouth (like a lollipop).
   (b) Have the patient copiously gargle, then swallow the viscous lidocaine.
7. Using a mucosal atomizer, spray 2 ml of 4 % lidocaine in the posterior oropharynx and as far down toward the glottis as possible (Figs. 8.3, 8.4, and 8.5).
8. Sedation
   (a) It is feasible to proceed with the awake intubation in an un-sedated, wide awake but cooperative patient.
   (b) Sedation can be initiated using institutional preferences, but some options include:
      (i) Midazolam 2 mg IV
      (ii) Ketamine 1 mg/kg IV
      (iii) Propofol 1 mg/kg IV
      (iv) Ketofol (ketamine and propofol both at concentrations of 10 mg/ml, 5 ml of each mixed in a 10 cc syringe) titrated at 1–3 ml aliquots
   (c) More atomized lidocaine can be provided prior to endotracheal tube (ETT) passage, but one must be aware of the upper lidocaine dose for your patient.
   (d) Adequate anesthesia is confirmed by the absence of a gag reflex upon direct palpation (Fig. 8.6).
9. Intubation, induction, and gentle direct laryngoscopy can be performed at this point to place the ETT (Fig. 8.7).
   (a) Induction (if no prior sedation) and paralytic agents should be available to immediately administer after ETT placement.
   (b) Thorough discussions on intubation techniques can be found in other chapters in the atlas.

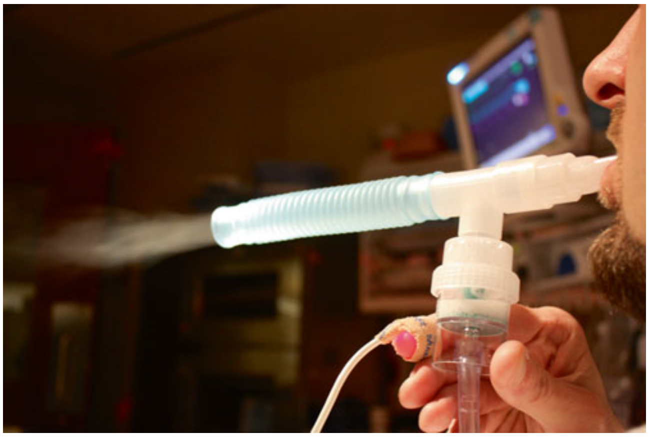

Fig. 8.1 Nebulization of 4 % lidocaine

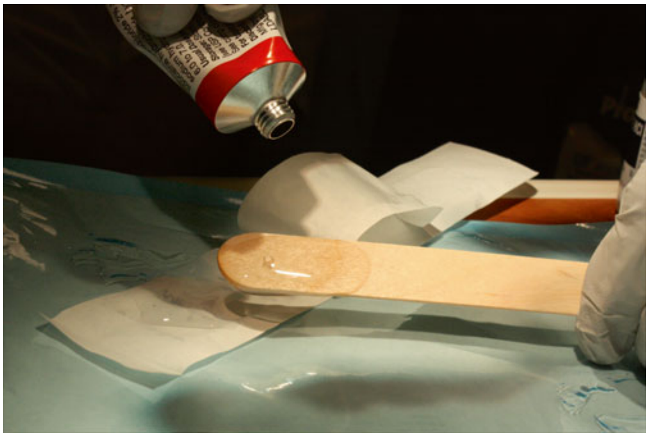

Fig. 8.2 Lidocaine lollipop

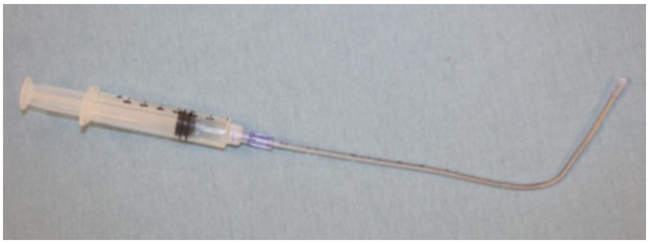

Fig. 8.3 Mucosal atomizer attached to syringe

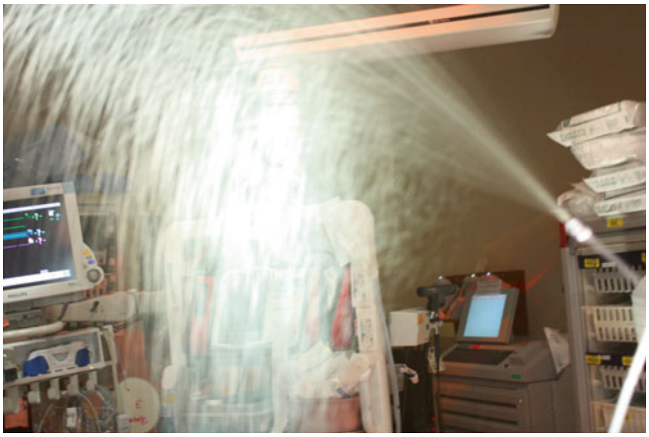

Fig. 8.4 Pushing syringe plunger atomizes lidocaine

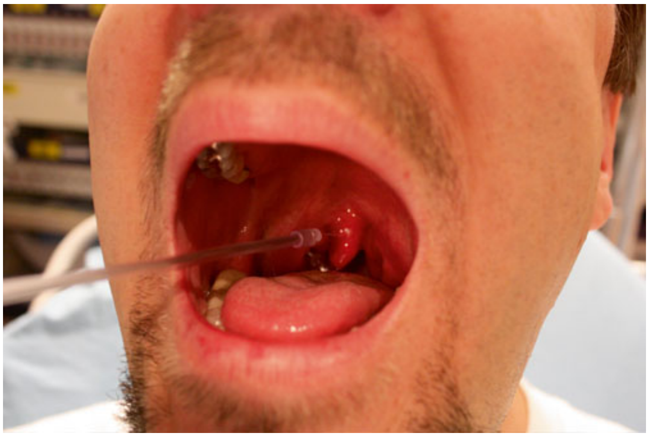

Fig. 8.5 Atomized lidocaine being administered to the posterior pharynx

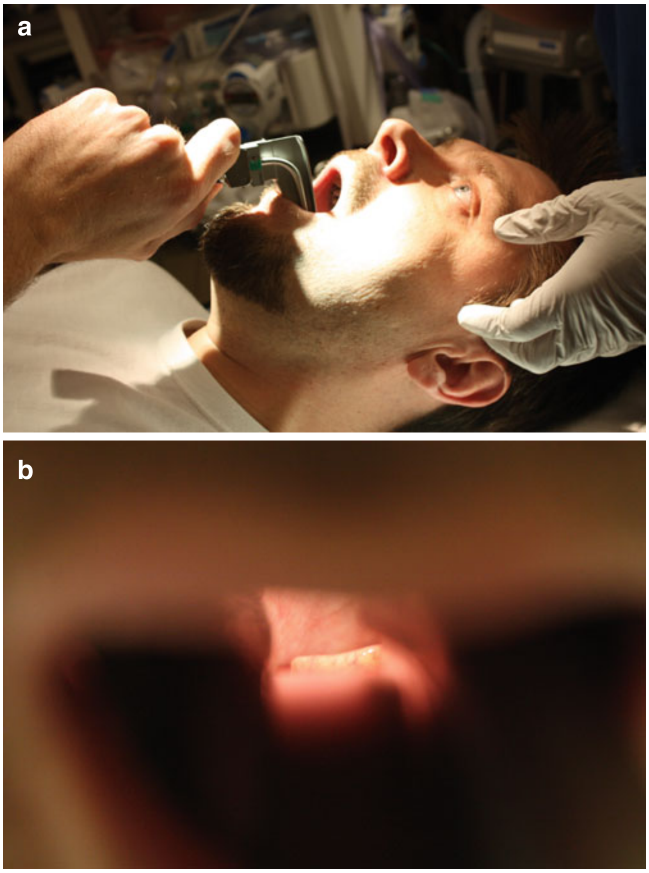

Fig. 8.6 (**a**, **b**) Adequately anesthetized awake patient with laryngoscopic view of epiglottis

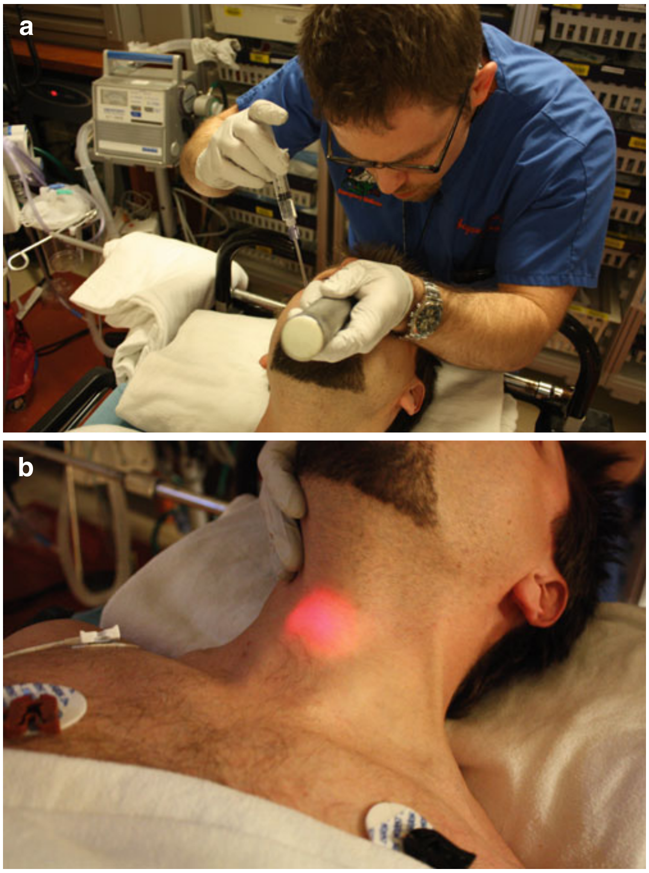

Fig. 8.7 (**a**) Final lidocaine atomization of deep structures and trachea, (**b**) intubation using a lighted stylet

8.5 Pearls and Pitfalls

- Pearls
  - Steps 3–7 should be done successively and as quickly as possible, to capitalize on the short half-life of lidocaine.
  - Simultaneous preoxygenation and anesthetization can be achieved by nebulizing the 4 % lidocaine through a face mask.
  - If the patient has been sedated, soft restraints may help prevent the patient from inadvertently grabbing the tube or your intubating equipment.
  - A nasotracheal intubation can be performed by simply anesthetizing the nares through which the ETT will be placed using lidocaine jelly and/or atomized lidocaine.
- Pitfalls
  - Failure to prepare all equipment beforehand may extend the procedure beyond the lidocaine half-life.
  - The toxic dose of lidocaine is 300 mg or 3–5 mg/kg. The dosages listed are intended for a 70 kg patient and yields a total dose of 280 mg (4 mg/kg). This dose may need to be decreased in smaller individuals.

Selected Reading

Agro F, Hung OR, Cataldo R, Carassiti M, Gherardi S. Lightwand intubation using the Trachlight: a brief review of current knowledge. Can J Anaesth. 2001;48(6):592–9.

Rhee KY, Lee JR, Kim J, Park S, Kwon WK, Han S. A comparison of lighted stylet (Surch-Lite) and direct laryngoscopic intubation in patients with high Mallampati scores. Anesth Analg. 2009;108(4):1215–92.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams and Wilkins, a Wolters Kluwer Business; 2008. Chap 11.

9

Rapid-Sequence Intubation

Ram A. Parekh

9.1 Indications

- Failure to oxygenate
- Failure to ventilate
- Unable to protect airway patency or reflexes
- Projected clinical course deterioration

9.2 Contraindications

- Absolute
  - Complete upper airway obstruction
  - Significant facial and airway trauma with loss of landmarks for orotracheal intubation
- Relative
  - Anticipated difficult intubation
    - Not an absolute contraindication.
    - Patient scenario requires a careful preintubation assessment and plan.
    - Consider an "awake" intubation.
    - Consider alternative airway adjuncts (e.g., extraglottic devices, video laryngoscopy, laryngeal mask airway [LMA]).
  - Induction or paralytic agent-specific contraindications given clinical circumstances
    - Caution: induction agents that lower blood pressure in hypotensive patients
    - Caution: succinylcholine in potentially hyperkalemic patients
  - Crash airway
    - Apneic, arrest, and periarrest situation

9.3 Materials and Medications

- Laryngoscope with appropriate blade (choice based on proceduralist's preference and patient anatomy) (Fig. 9.1)
- Intubating stylet
- Endotracheal tubes (ETTs)
- Syringe, 10 mL (to inflate ETT cuff)
- Surgilube
- Suction catheter
- Oral and nasal airways (Fig. 9.2)
- Ambu bag and mask attached to oxygen source
- Induction, pretreatment, and paralytic agents
- ETT confirmation device—EZ capnometry, quantitative end-tidal carbon dioxide concentration (EtCO$_2$) detection

R.A. Parekh, MD
Department of Emergency Medicine,
Icahn School of Medicine at Mount Sinai,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: ram.parekh@mssm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_9

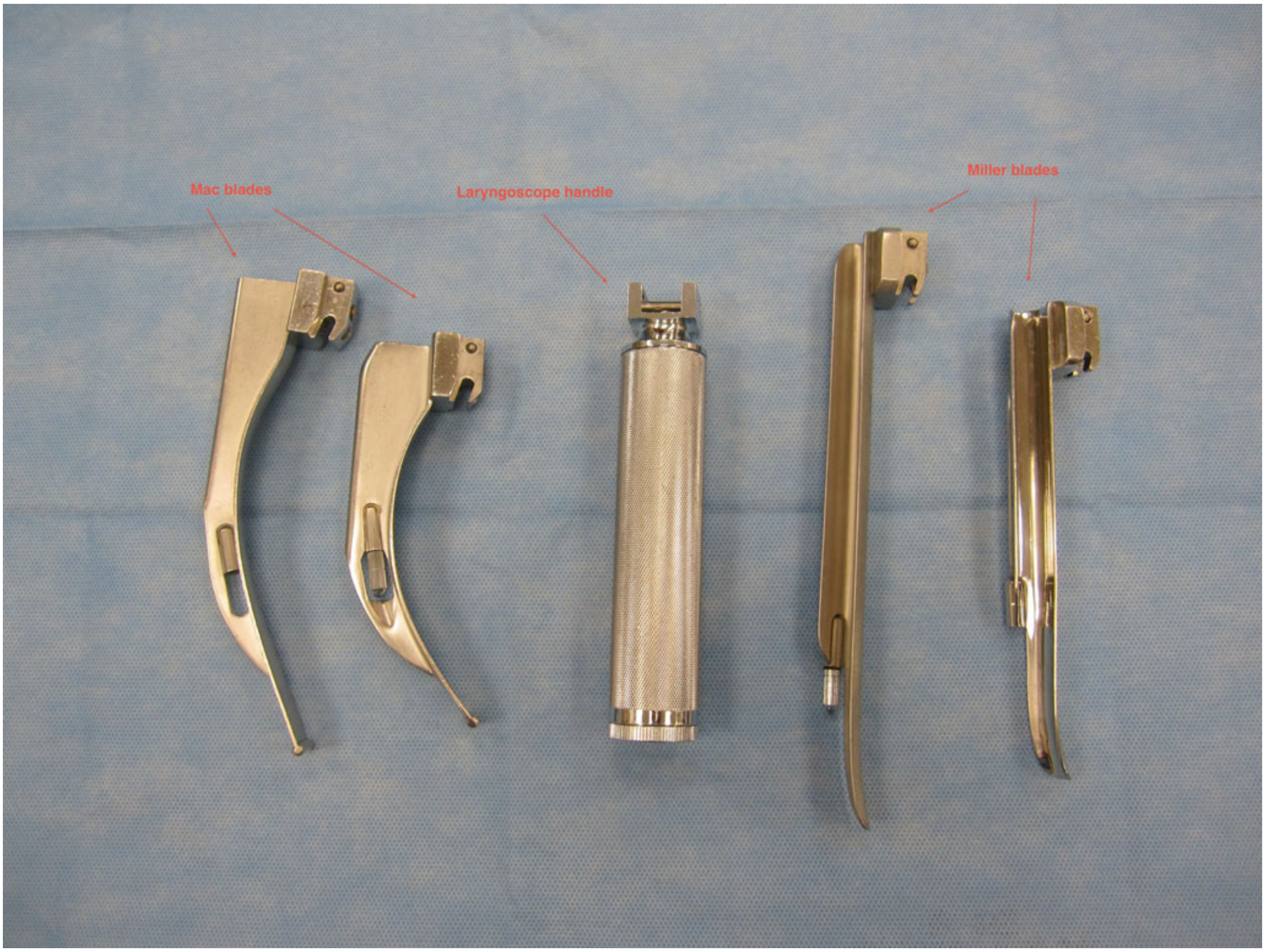
Mac blades
Laryngoscope handle
Miller blades

Fig. 9.1 Laryngoscope and blades

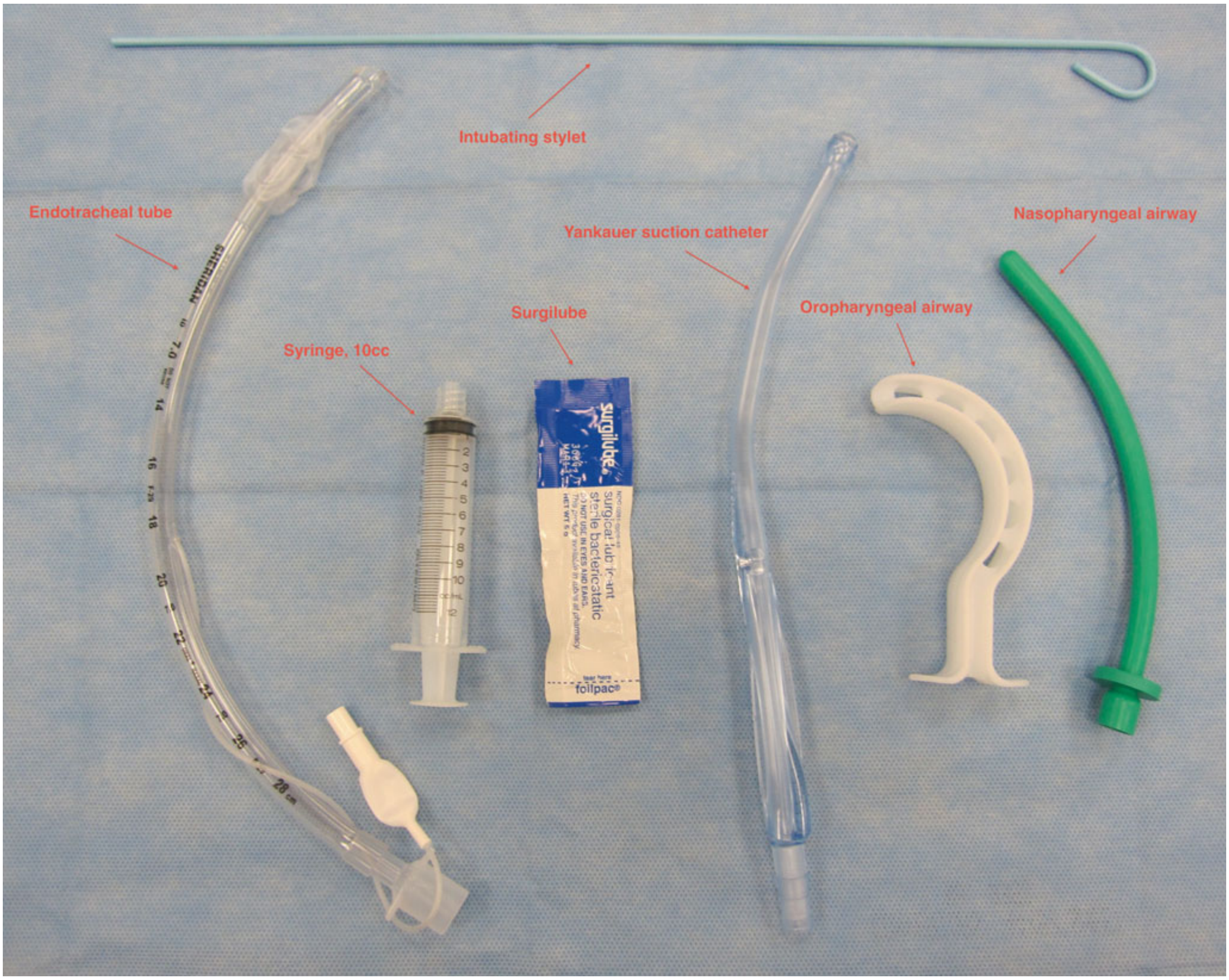
Intubating stylet
Endotracheal tube
Yankauer suction catheter
Nasopharyngeal airway
Surgilube
Oropharyngeal airway
Syringe, 10cc

Fig. 9.2 Oral and nasal airways

9.4 Procedure

1. Preparation
   - Thoroughly assess patient for difficulty of intubation (Fig. 9.3).
   - Develop fallback plans for failed intubation attempt.
   - Establish at least one, but preferably two, secure intravenous (IV) lines.
   - Place on cardiac monitor with pulse oximetry, blood pressure monitoring, and continuous capnography.
   - Yankauer suction device attached to suction, suction on.
   - Pharmacological agents, drawn and labeled.
   - Laryngoscope and blades tested to ensure proper functioning of light source.
   - Desired ETT size, prepared for intubation:
     - Intubating stylet in position (tip at eye of ETT)
     - 10-mL syringe attached to ETT
     - ETT configured per proceduralist's preference (e.g., hockey stick, curved)
     - Cuff tested for air leak
2. Preoxygenation
   - Administer high-flow oxygen for 3–5 min (Fig. 9.4).
     - Nitrogen is exchanged for O_2 in the functional residual capacity of the lungs.
     - Establishes oxygen reservoir within lungs (primarily), blood, and body tissue.
     - Also known as nitrogen "washout."
   - This can be done using:
     - Non-rebreather masks—delivers 65–70 %
       - Difficult intubation not anticipated
     - Well-fitting bag-valve-masks (without positive-pressure ventilation)—delivers greater than 90 % oxygen (Fig. 9.5)
     - Noninvasive positive-pressure ventilation (NIPPV)—delivers 100 % oxygen
       - Consider NIPPV in high-risk patients with moderate to severe shunt physiology.
3. Pretreatment
   - Administer pharmacological agents to mitigate adverse physiological effects of intubation, induction, and paralysis, which may be undesirable in certain clinical circumstances (Table 9.1).
4. Induction and paralysis
   - Administer a rapidly acting induction agent to produce rapid loss of consciousness via IV push (Table 9.2).
   - Immediately follow induction agent with a neuromuscular-blocking agent via IV push (Table 9.3).
5. Positioning
   - If no cervical spine injury suspected, place the patient in the "sniff" position (Fig. 9.6):
     - Flex neck
     - Extend head
   - Ideally, the patient's pinna will be at the level of the sternum.
6. Direct laryngoscopy (see Chap. 10)
7. Proof of placement
   - Visualize ETT passing vocal cords
   - Confirm tube placement via $EtCO_2$:
     - Qualitative detection device—EZ Cap
     - Quantitative continuous $EtCO_2$ waveform on monitor (preferred) (Fig. 9.7)
   - Auscultation of breath sounds:
     - Lung fields bilaterally
     - Epigastric region (ensuring no breath sounds in the stomach)
8. Postintubation management
   - Secure ETT (Fig. 9.8)
   - Initiate mechanical ventilation.
   - Postintubation sedation and analgesia.
   - Postintubation chest x-ray.

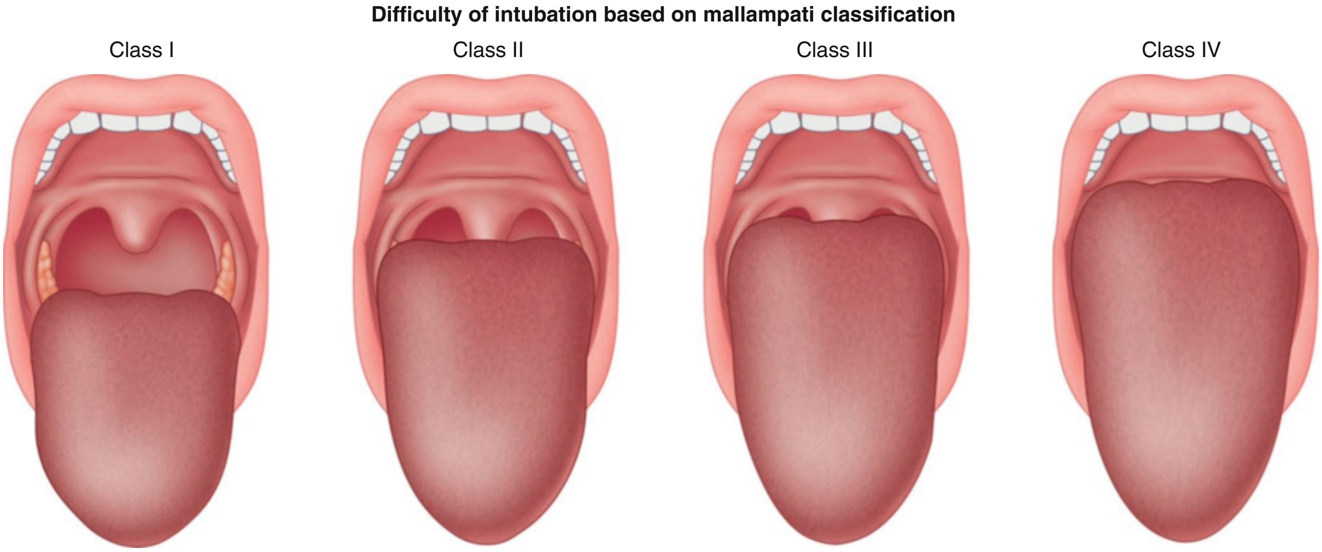
Difficulty of intubation based on mallampati classification
Class I
Class II
Class III
Class IV

Fig. 9.3 Assess patient for difficulty of intubation

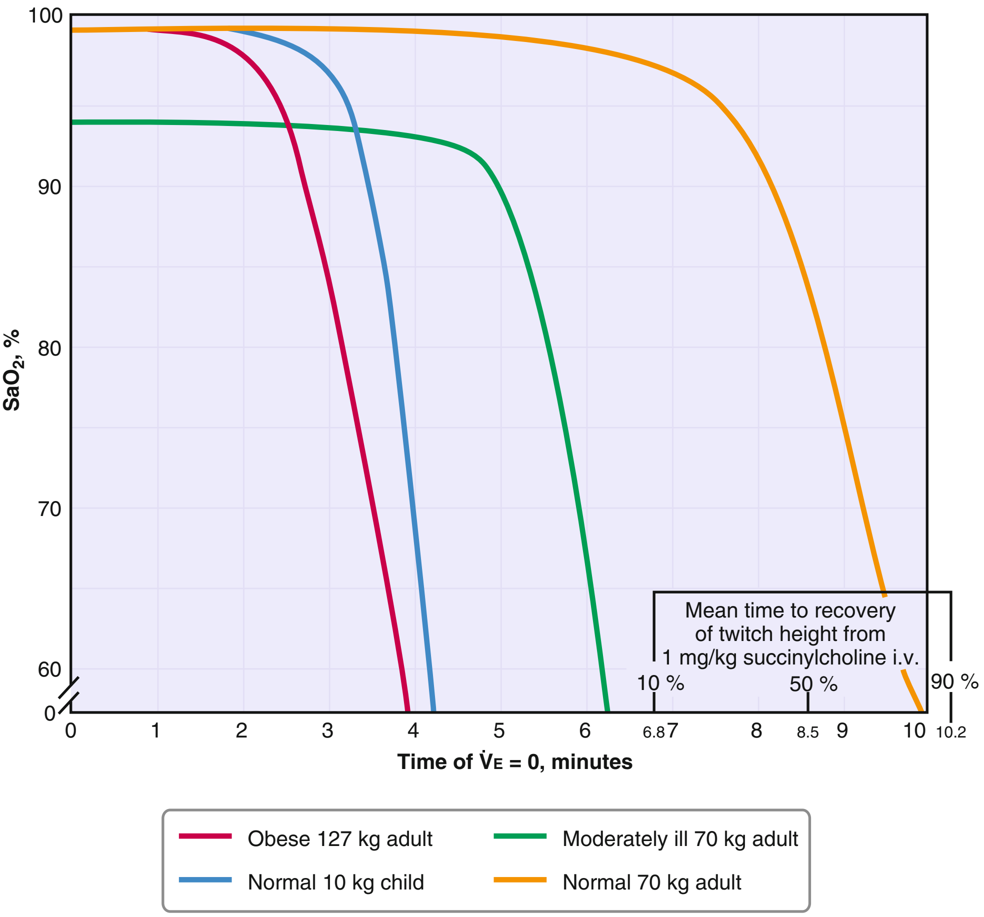

Fig. 9.4 *FaO2* fractional concentration of alveolar oxygen, *SaO2* arterial oxygen saturation, *Ve* expired volume per minute

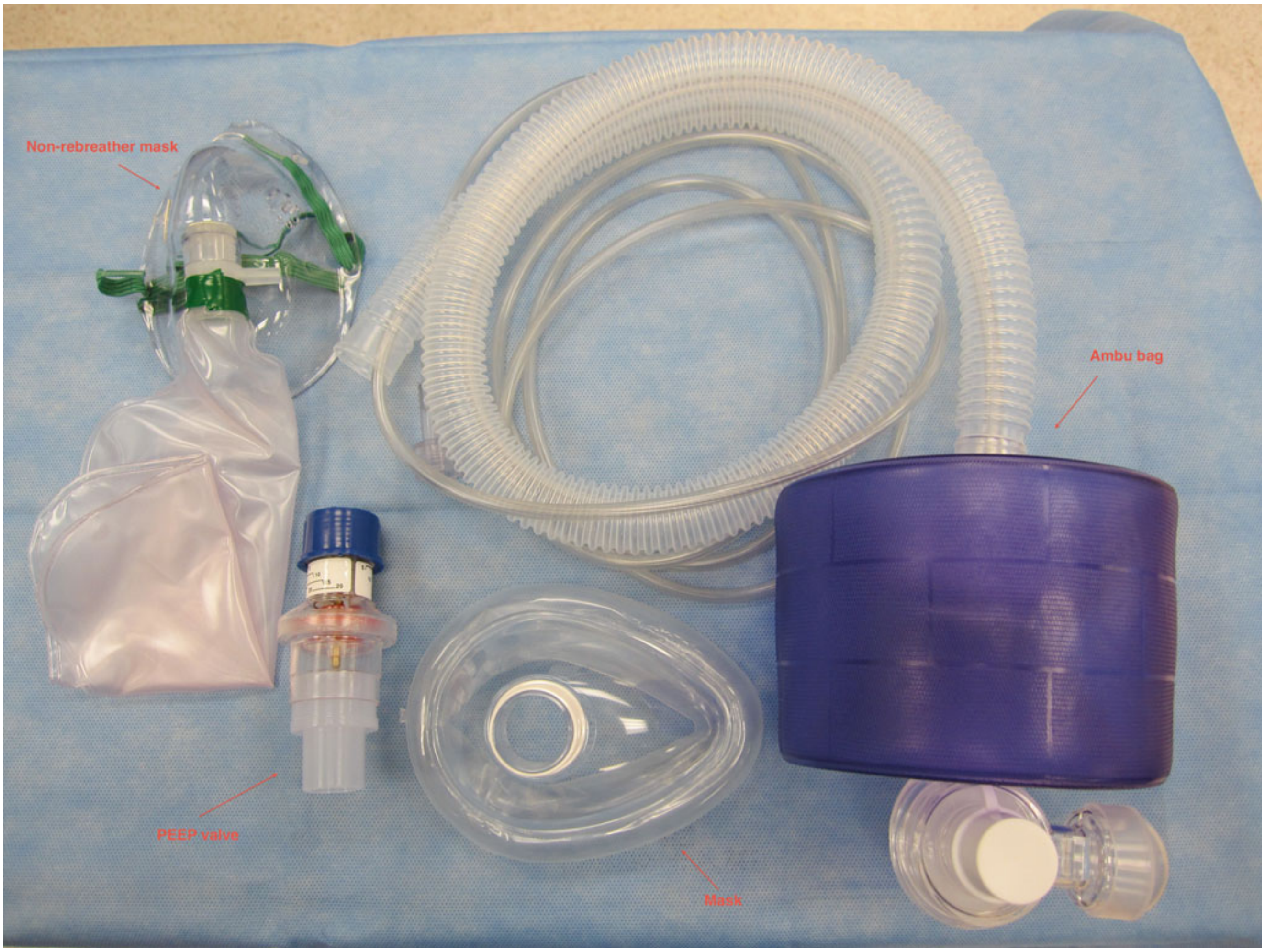

Fig. 9.5 Preoxygenation materials

Table 9.1 Pretreatment: pharmacological agents used to mitigate adverse physiological effects of intubation, induction, and paralysis

Agent	Dose (IV)	Indication
Lidocaine	1.5 mg/kg Rapid push	Use in tight brain to attenuate ICP increase from laryngoscopy/intubation; use in tight lungs to blunt bronchospastic response
Fentanyl	1–3 mcg/kg Slow push	Use in tight brain, tight heart, and tight vessels to blunt reflex sympathetic response to laryngoscopy

ICP intracranial pressure

Table 9.2 Induction: rapidly acting induction agents used to produce rapid loss of consciousness

Agent	Dose (IV) (mg/kg)	Onset (sec)	Duration (min)
Midazolam	0.2–0.3	60–90	15–30
Etomidate	0.3	15–45	3–12
Thiopental	3	<30	5–10
Ketamine	1.5–2.0	45–60	10–20
Propofol	1.5	15–45	5–10

Table 9.3 Paralysis: neuromuscular-blocking agents administered immediately after induction agent

Agent	Dose (IV)	Onset (sec)	Duration (min)
Succinylcholine	1.5 mg/kg	45	6–10
Rocuronium	1.0 mg/kg	60–75	40–60
Vecuronium	0.01 to prime, then 0.15 mg/kg	75–90	60–75

Fig. 9.6 Patient in the "sniff" position. *OA* oral axis, *LA* laryngeal axis, *PA* pharyngeal axis

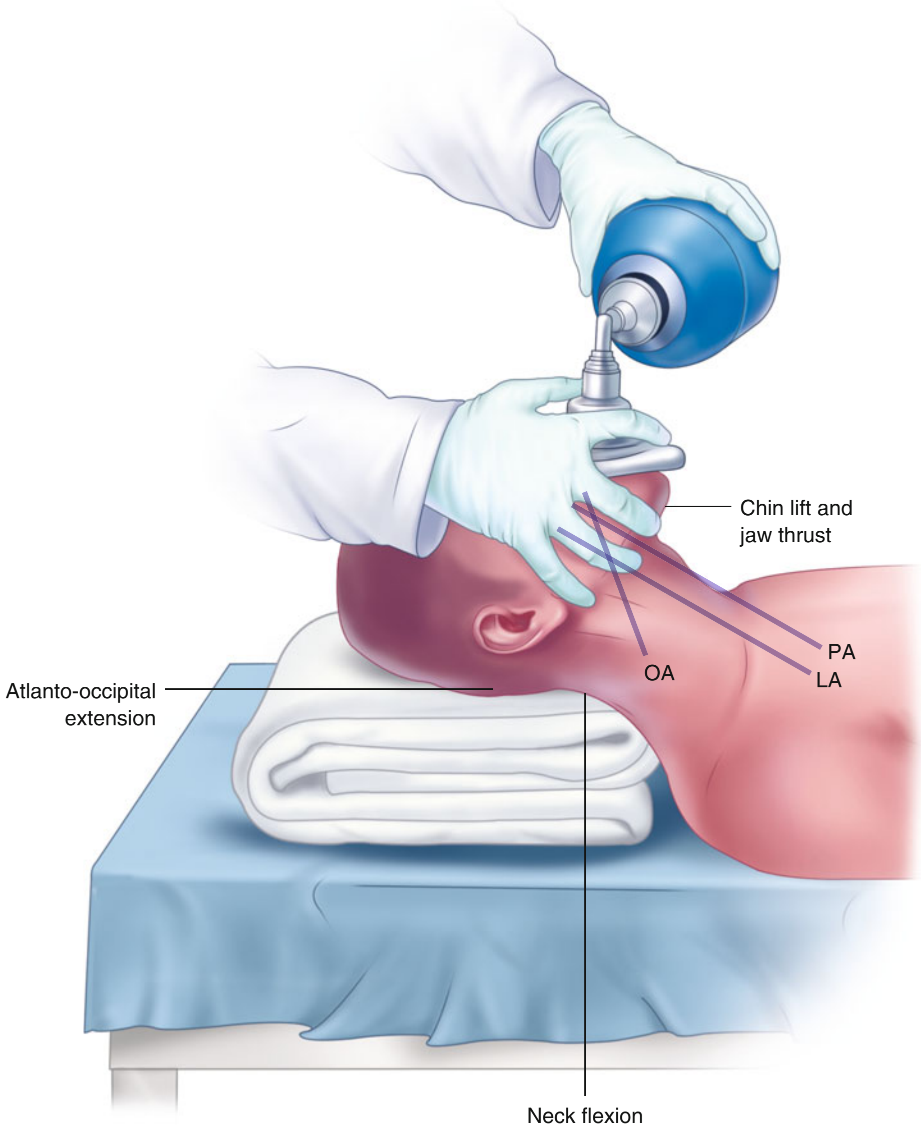
Chin lift and jaw thrust
PA
OA
LA
Atlanto-occipital extension
Neck flexion

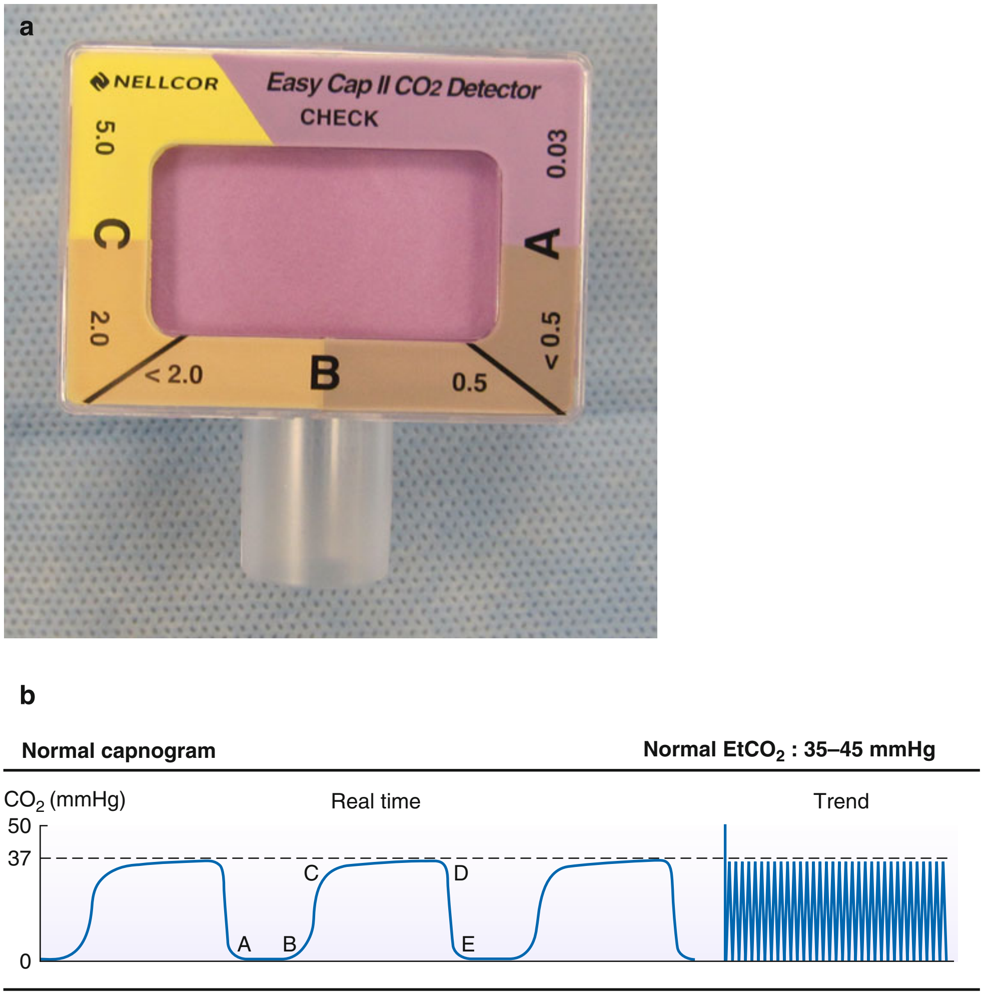
a
NELLCOR
Easy Cap II CO2 Detector
CHECK
5.0
C
2.0
< 2.0
B
0.5
< 0.5
A
0.03
b
Normal capnogram
Normal EtCO2 : 35–45 mmHg
CO2 (mmHg)
Real time
Trend
50
37
0
A
B
C
D
E

Fig. 9.7 $EtCO_2$, end-tidal carbon dioxide concentration: *a* Qualitative detection device—EZ Cap, *b* quantitative continuous $EtCO_2$ waveform on monitor

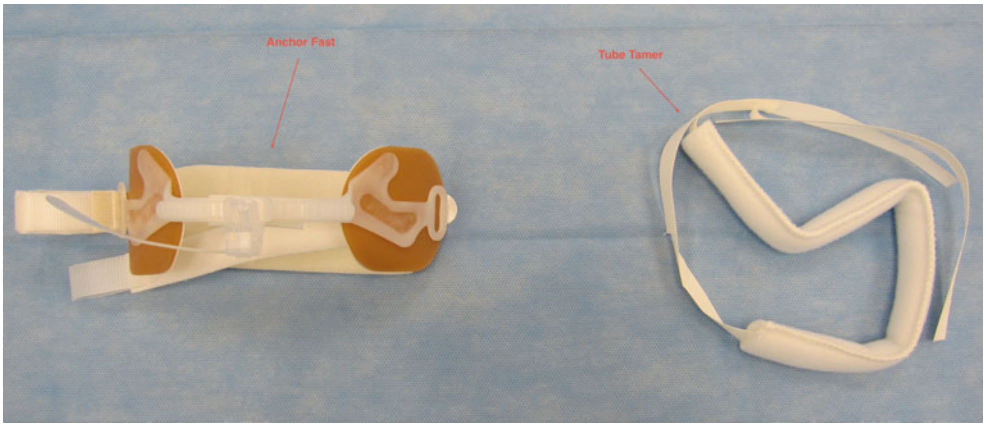
Anchor Fast
Tube Tamer

Fig. 9.8 Equipment used to secure ETT

9.5 Complications

- Esophageal intubation
- Right mainstem intubation
- Pneumothorax from laryngeal trauma
- Aspiration
- Dental trauma
- Vocal cord injury
- Hypotension
  - Induction agent
  - Decreased venous return from positive pressure
  - Pneumothorax
- Hyperkalemia (succinylcholine used in mildly hyperkalemic patient)
- Iatrogenically obstructed airway
- Failure to intubate

9.6 Pearls and Pitfalls

- Utilization of oral and nasal airways will greatly increase the ease of preoxygenation and reoxygenation.
- Inadequate preoxygenation will cause premature desaturation, decreasing time for intubation.
- Suboptimal patient positioning can prevent vocal cord visualization during direct laryngoscopy.
- Consider alternative airway devices before intubation to have a plan in place if a difficult airway is encountered.
- Underdosing induction or paralytic agent will prevent adequate time to perform procedure or create patient discomfort.
- Inappropriately sized McIntosh laryngoscopic blades.
  - Too small—more difficulty in moving tongue and epiglottis out of way for vocal cord visualization
  - Too large—easier to overshoot and go past vocal cords into the esophagus
- Excessive cricoid pressure (Sellick maneuver) may lead to disrupted laryngoscopic view and difficulty passing the ETT.
  - Disrupted view: readjust larynx using dominant hand to allow cord visualization.
  - Difficulty passing ETT: ask for release of some cricoid pressure to allow for ETT passage.
- Inadequate postintubation sedation and analgesia, especially when long-acting paralytics are used.
- Acidic gastric contents can cause CO_2 qualitative capnometry to change to yellow, falsely indicating tracheal placement of the ETT.

Selected Reading

Walls R, Murphy M. Manual of emergency airway management. Philadelphia: Lippincott Williams & Wilkins; 2008.

Weingart SD. Preoxygenation, reoxygenation, and delayed sequence intubation. J Emerg Med. 2011;40:661–7.

Direct Laryngoscopy

10

Bharath Chakravarthy and Weston Seipp

10.1 Indications

- Orotracheal intubation
  - Maintenance of oxygenation/ventilation
  - Airway protection
- Visualization of laryngeal anatomy
- Foreign body retrieval

10.2 Contraindications

- Absolute
  - None
- Relative
  - Presumed difficult airway
    - Anatomical limitations
      - Small oral opening (less than three of the patient's fingers)
      - Small mandible (hyomental distance less than three fingers)
      - Hyoid-thyroid distance (less than two fingers)
    - Clinical limitations
      - Patient with unstable cervical spine
      - Patient with multiple facial or neck trauma
      - Patient with history of tracheal stenosis, irradiation, or history of tracheal mass or surgery

---

B. Chakravarthy, MD, MPH (✉)
Department of Emergency Medicine,
University of California Irvine,
Orange, CA, USA
e-mail: bchakrav@uci.edu

W. Seipp, MD
Department of Emergency Medicine,
University of California Irvine Medical Center,
Orange, CA, USA

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_10

10.3 Materials and Medications (Fig. 10.1)

- Laryngoscope handle
- Laryngoscope blade with light
  - Macintosh blade ("Mac" or "curved blade")
  - Miller blade ("straight blade")
- Bag valve mask attached to 100 % O_2 source
- Endotracheal tube (ETT)
- 10-mL syringe
- Yankauer suction
- End-tidal CO_2 ($EtCO_2$) monitor (colorimetric or quantitative)
- McGill forceps (for foreign body retrieval)
- Postintubation chest radiograph

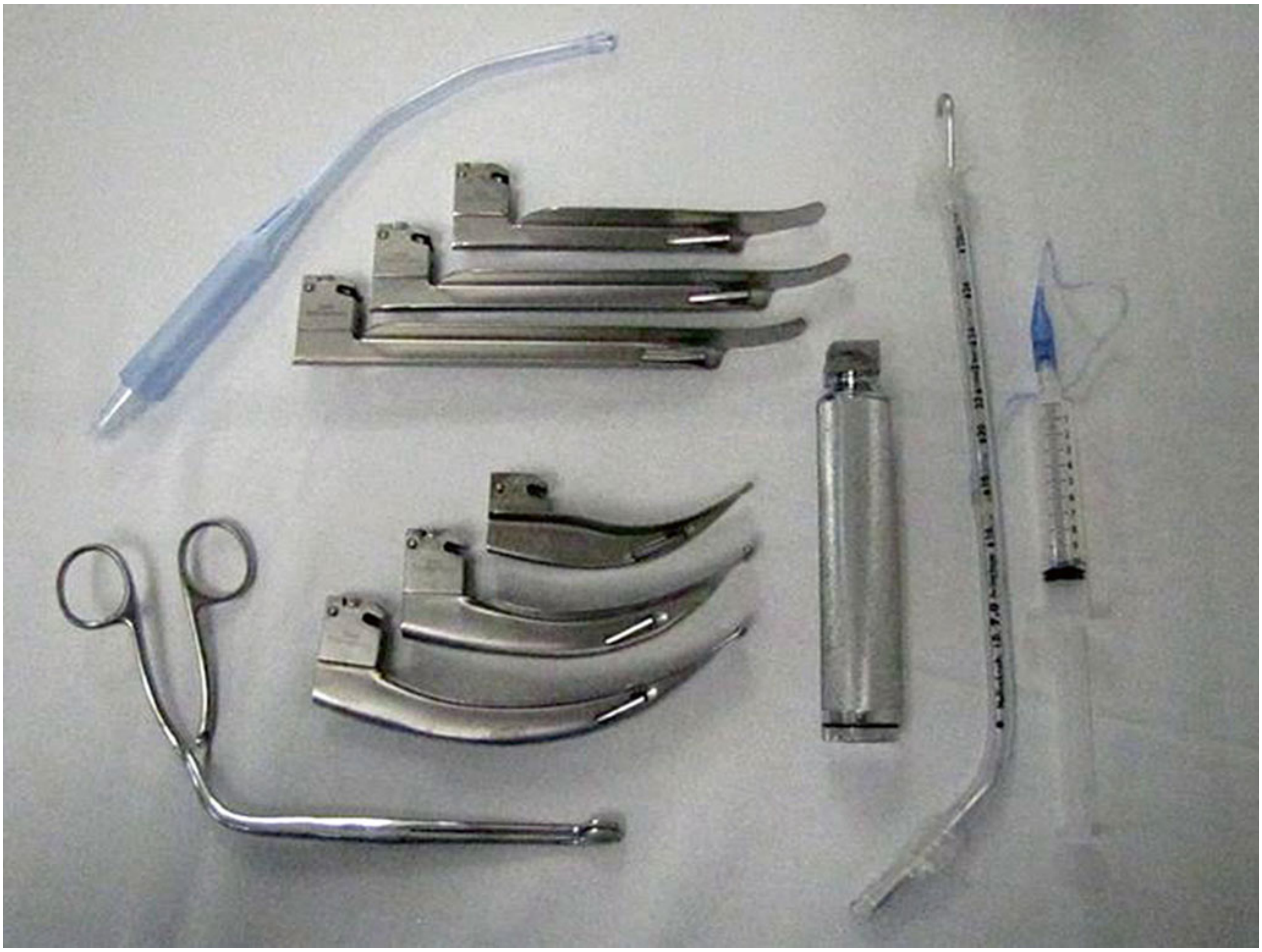

Fig. 10.1 Basic laryngoscopy supplies. *Clockwise from top left*: Yankauer suction, Miller Blades, endotracheal tube, 10-mL syringe, laryngoscope handle, Macintosh blades, and McGill forceps

10.4 Procedure

1. Check the laryngoscope handle and blades to ensure that the light is functioning.
2. Choose the appropriate blade based on patient size.
   - (a) The Macintosh or Miller 3 size is appropriate for the majority of adults, and a 4 can be used for a larger body habitus.
   - (b) The Macintosh blade is generally preferred in adults owing to increased space for ETT passage [1].
   - (c) The Miller blade may be preferable in cases in which the patient has limited mouth opening (owing to its smaller vertical height), or in which the airway is particularly floppy (such as in infants and children) [1].
3. Position the patient (Fig. 10.2).
   - (a) Raise the bed so that the patient's oral opening is at the level of the xiphoid process of the laryngoscopist.
   - (b) The optimal laryngeal view is obtained in the neck flexion/head extension or "sniffing" position. To achieve this, place towels under the patient's occiput to raise it approximately 6–9 cm [2].
   - (c) If patient is in cervical spine precautions, then an assistant must hold the cervical spine in midline immobilization throughout the laryngoscopy; elevation of the occiput is therefore contraindicated.
4. Provide 100 % O_2 via face mask to preoxygenate the patient before laryngoscopy.
5. After ensuring adequate anesthesia and neuromuscular blockade (if performing laryngoscopy for intubation), perform the scissor technique to open the patient's mouth and to lift the tongue base from the glottic opening.
6. Macintosh blade insertion (Fig. 10.3).
   - (a) Insert the laryngoscope into the patient's mouth, starting from the right side, and slowly advance into the oropharynx, using the blade's vertical flange to "sweep" the tongue to the left and away from the glottic opening.
   - (b) Advance the blade along the tongue toward the tongue base, until the epiglottis and posterior arytenoids are viewed.
   - (c) In order to expose the cords, insert the Macintosh blade into the vallecula, which is the potential space anterior to the epiglottis and posterior to the tongue base. This will act as a fulcrum and raise the epiglottis, exposing the vocal cords.
   - (d) To further expose the cords and/or expose the vallecula, exert force outward at a 45° angle to the patient. Do not "rock" the laryngoscope because this may cause injury to the teeth.
7. Miller blade insertion (Fig. 10.4).
   - (a) Insert the blade into the right side of the mouth and slowly advance along tongue toward the tongue base. The Miller blade does not have a flange for isolating the tongue, and thus the Macintosh may be preferable in patients with large tongues.
   - (b) Advance the blade along the right side of the tongue until the epiglottis and posterior arytenoids are visible.
   - (c) In contrast to the Macintosh blade, the Miller blade is used to directly isolate the epiglottis and expose the vocal cords. Using the tip of the Miller, move the epiglottis anteriorly to expose the vocal cords.
   - (d) As with the Macintosh, exert force outward at a 45° angle to the patient to increase the view of vocal cords. Do not "rock" the laryngoscope because this may cause injury to the teeth.
8. Assessing the glottic view (Fig. 10.5).
   - (a) With the epiglottis either directly or indirectly lifted from the glottic opening, assess the Cormack-Lehane laryngeal view grade.
     - (i) Grade I—view of entire laryngeal opening, including cords
     - (ii) Grade II—view of posterior laryngeal cartilages
     - (iii) Grade III—visualization of epiglottis only
     - (iv) Grade IV—no structures visualized
   - (b) A lower grade (higher quality, better) view is predictive of intubation success [3].
   - (c) In the event of a higher grade view, the operator may request airway adjuncts, such as a bougie.
9. Improving the glottic view [4].
   - (a) With the laryngoscope in the desired position, it is possible to improve the glottic view by exerting backward pressure on the thyroid cartilage either with the operator's right hand (bimanual laryngoscopy) or with an assistant applying BURP (backward-upward-rightward pressure) (Fig. 10.5a).
   - (b) Backward pressure increases the vertical distance between the epiglottis and the posterior cartilages, thereby increasing the likelihood of vocal cord visualization.
10. Passing the ETT (Fig. 10.6).
    - (a) With the optimal view of the cords obtained, pass the ETT from the right corner of the mouth through the vocal cords, to a depth of 21 cm at the incisors in females and 23 cm in males.
    - (b) Inflate the ETT cuff with approximately 5 cc of air until the pilot balloon is firm to touch.
11. After completing intubation or after completion of laryngoscopy, slowly remove the blade from the mouth, taking care to avoid dental or lip trauma.
12. Attach the capnography device to the ETT tube to ensure $EtCO_2$ return.
    - (a) Colorimetric devices will turn from purple to yellow in the presence of $EtCO_2$.
    - (b) Quantitative devices will return a CO_2 waveform.
13. Attach the ETT to ventilator or bag-valve-mask connected to an oxygen source.

14. Auscultate breath sounds in both lung fields and ensure absence of breath sounds over the epigastrium (which could signify esophageal intubation).
15. Obtain a postintubation chest x-ray to ensure no right mainstem intubation or pneumothorax.

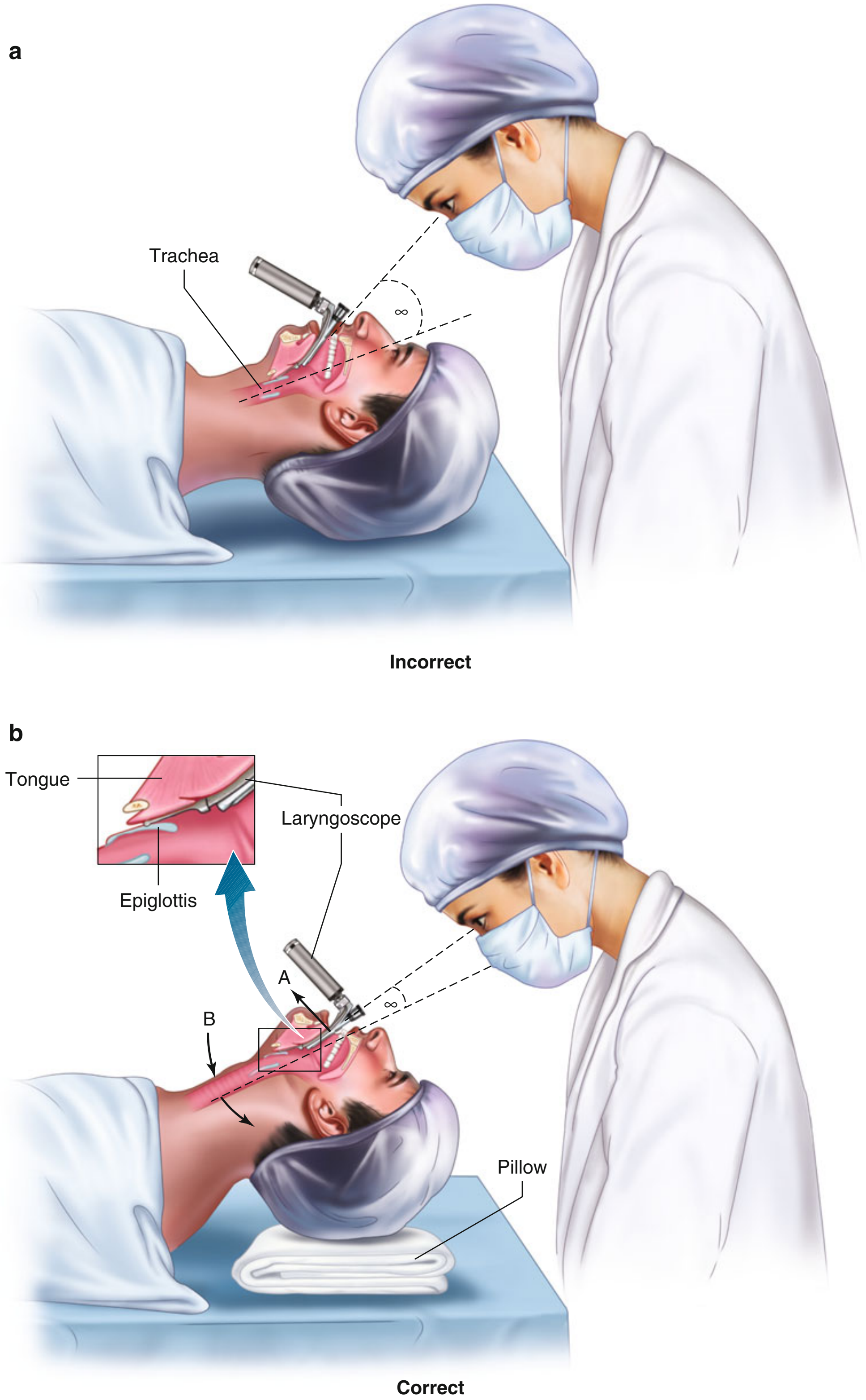
a
Trachea
∞
Incorrect
b
Tongue
Laryngoscope
Epiglottis
A
B
∞
Pillow
Correct

Fig. 10.2 Visualization axis and sniffing position: (**a**) The patient's occiput is not elevated and the neck is not in flexion, thereby creating a steep visual axis, (**b**) the occiput is correctly elevated 6–9 cm, placing the patient in sniffing position and allowing the visual axis to align with the airway axis

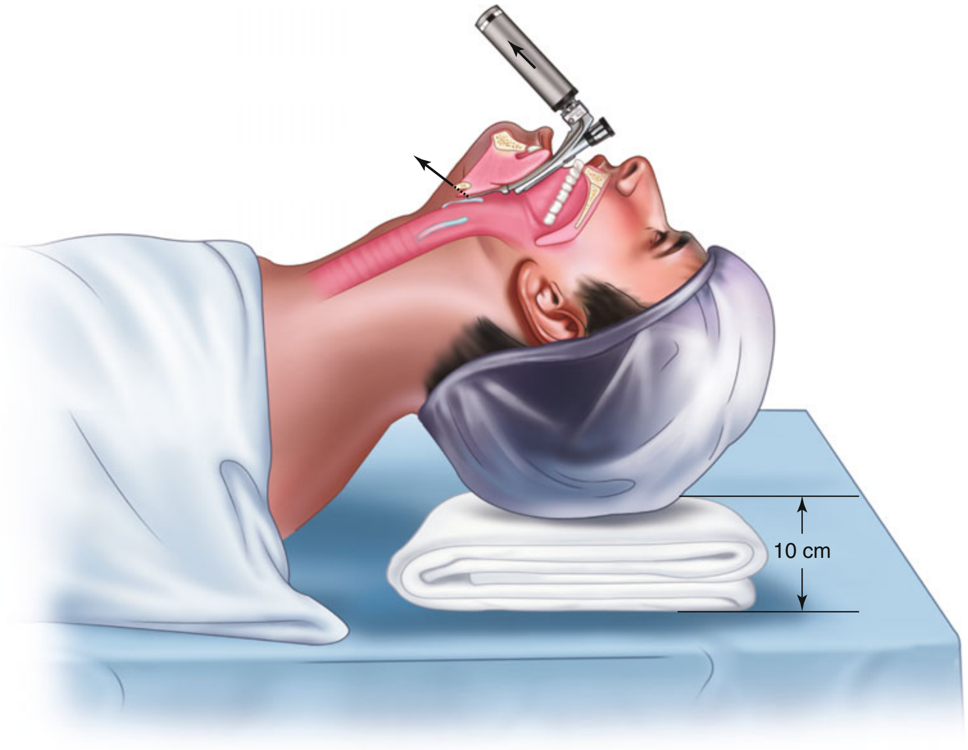

Fig. 10.3 Macintosh blade insertion. The blade is inserted into the vallecula, which raises the epiglottis and exposes the glottic opening

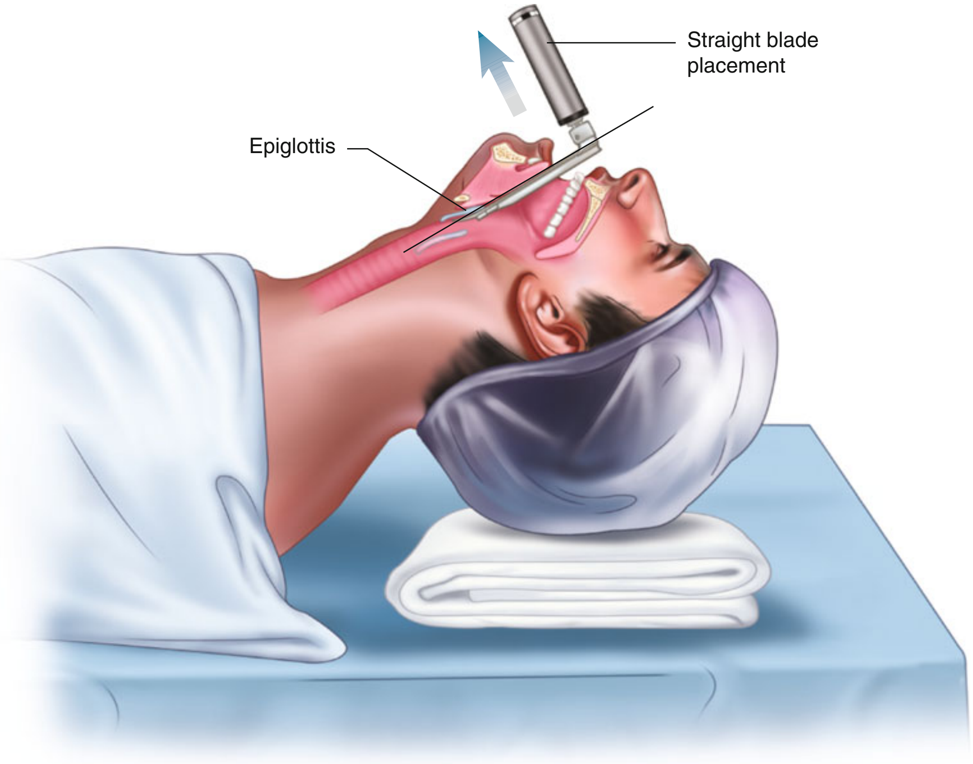
Straight blade placement
Epiglottis

Fig. 10.4 Miller blade insertion. The blade is used to elevate the epiglottis directly, exposing the glottic opening

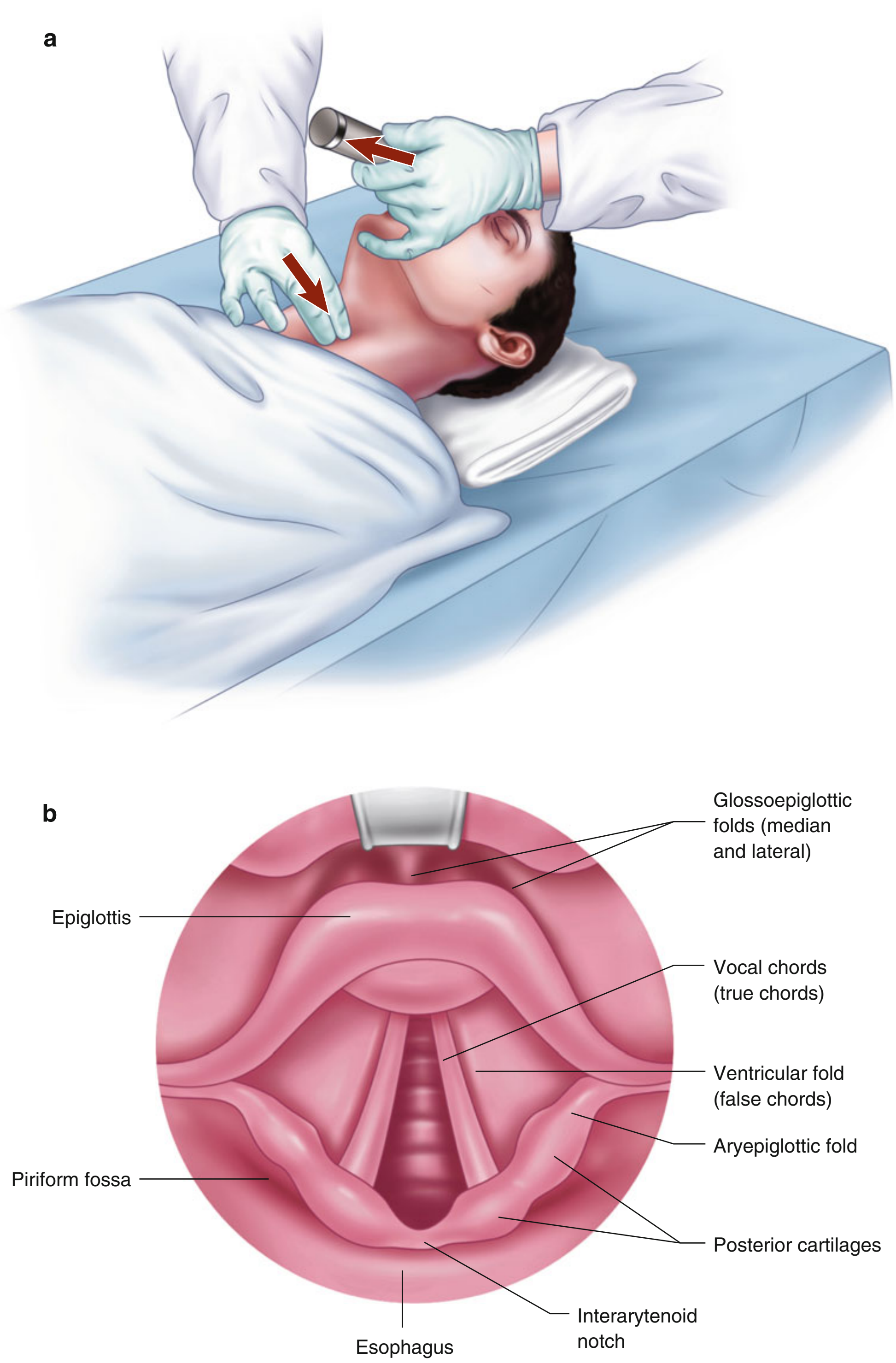
a
b
Glossoepiglottic folds (median and lateral)
Epiglottis
Vocal chords (true chords)
Ventricular fold (false chords)
Aryepiglottic fold
Piriform fossa
Posterior cartilages
Interarytenoid notch
Esophagus

Fig. 10.5 (**a**) Bimanual laryngoscopy. The force on the neck is opposite the direction of lift by the laryngoscope, (**b**) laryngoscopy view, (**c**) Cormack and Lehane

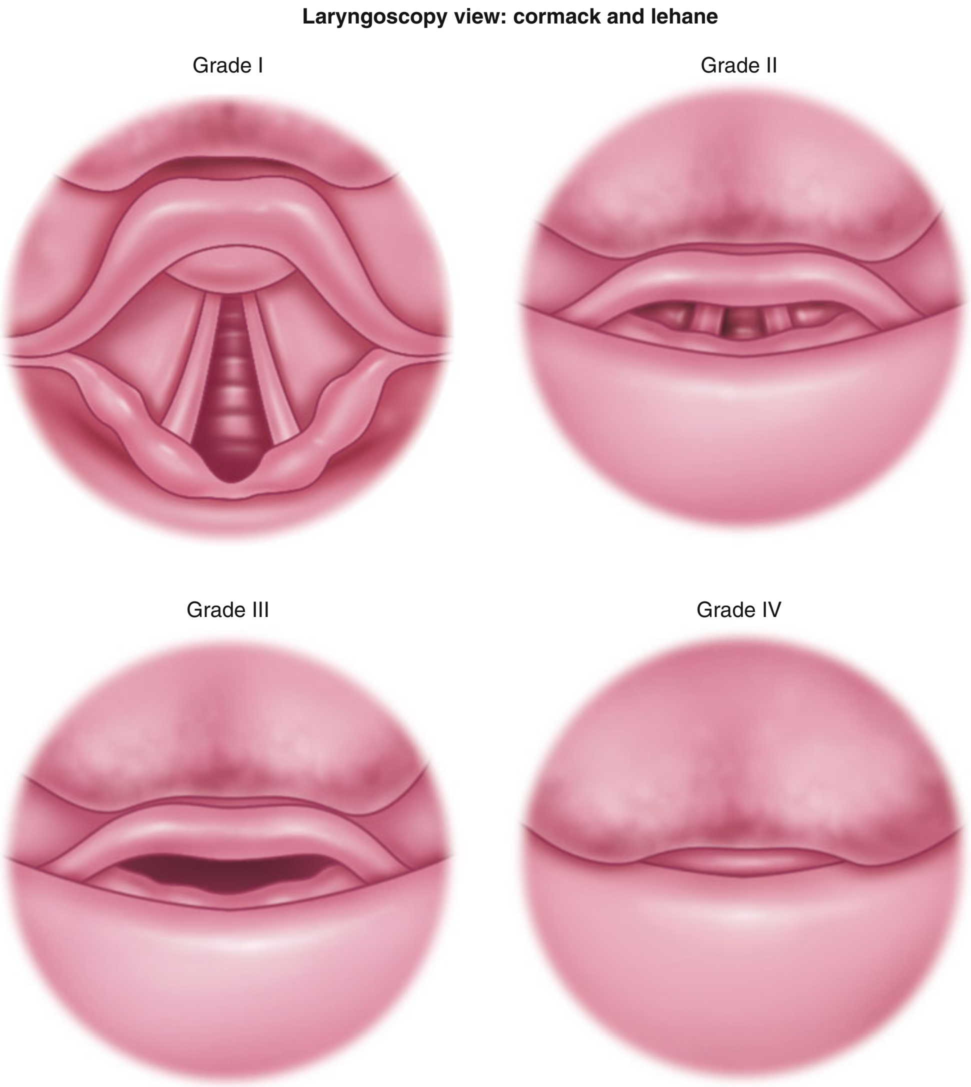
Laryngoscopy view: cormack and lehane
Grade I
Grade II
Grade III
Grade IV

Fig. 10.5 (continued)

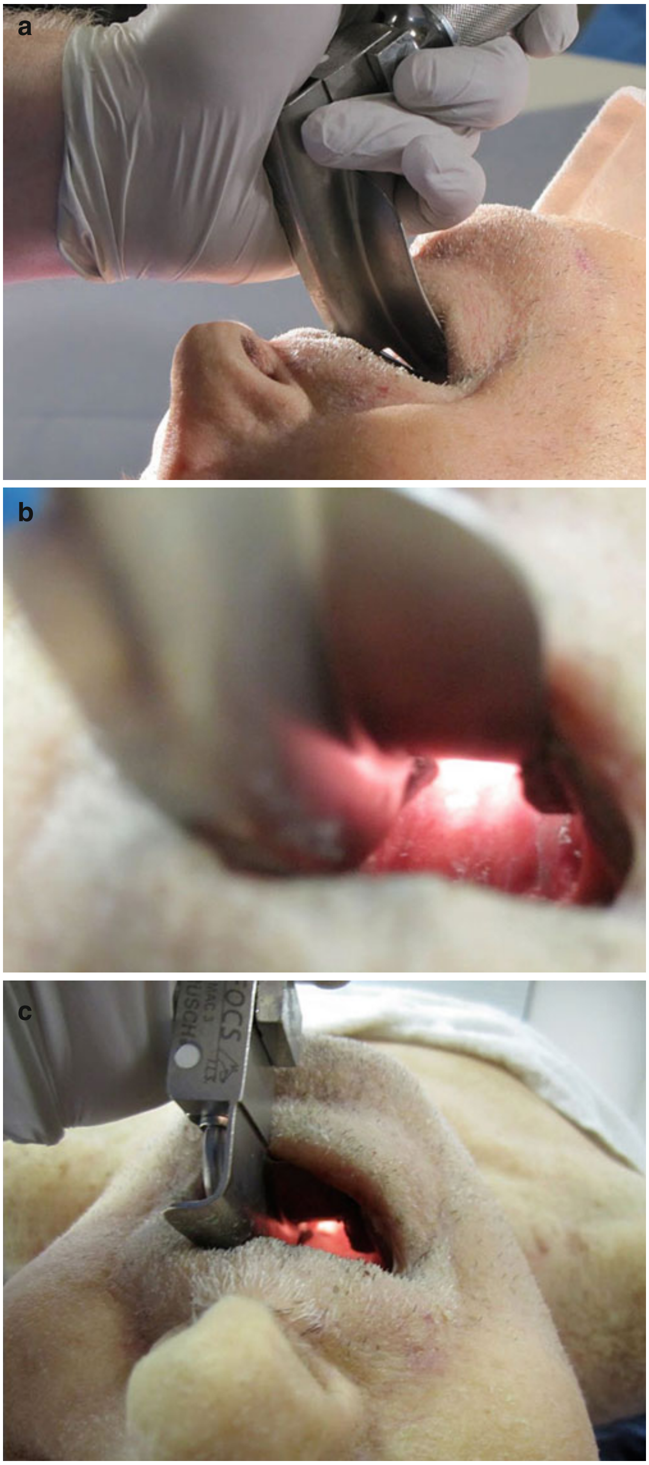

Fig. 10.6 (**a**) Insert the laryngoscope into the mouth, (**b**) sweep the tongue to the left, (**c**) advance the laryngoscope until the epiglottis is visible, and insert the blade into vallecula to expose the cords. When using the Miller blade, insert the blade until the epiglottis is seen. Slide the blade under the epiglottis and lift to expose the cords

10.5 Complications [3, 5]

- Common (1–4 %)
  - Esophageal intubation—*can be fatal if unrecognized*
  - Mainstem bronchus intubation
  - Tachycardia
  - Aspiration
  - Hypotension
- Uncommon (<1 %)
  - Dental/oral trauma
  - Oropharyngeal edema or bleeding
  - Laryngospasm
  - Dysrhythmia
  - Pneumothorax
  - Cardiac arrest

10.6 Pearls and Pitfalls

- Pearls
  - Positioning is of key importance—all patients with a stable cervical spine should be placed in the “sniffing” position to maximize view.
  - Consider the “ramping” position in obese patients with stable cervical spines—elevation of the head and shoulders allows redundant tissue to fall and gives an improved glottic view.
  - Always have suction readily available to remove blood, vomitus, or edema.
  - If structures are not readily visible, withdraw the blade gradually because it is common to insert the blade too deep.
- Pitfalls
  - “Rocking” the laryngoscope instead of lifting outward
  - Failure to recognize esophageal intubation
  - Failure to evaluate postintubation chest x-ray

References

1. Hagberg CA. Benumof’s airway management. Maryland Heights, MO: Mosby; 2007. p. 363–5.
2. Park SH, Park HP, Jeon YT, Hwang JW, Kim JH, Bahk JH. A comparison of direct laryngoscopic views depending on pillow height. J Anesth. 2010;24:526–30.
3. Martin LD, Mhyre JM, Shanks AM, Tremper KK, Kheterpal S. 3,423 emergency tracheal intubations at a university hospital: airway outcomes and complications. Anesthesiology. 2011;114:42–8.
4. Levitan RM, Kinkle WC, Levin WJ, Everett WW. Laryngeal view during laryngoscopy: a randomized trial comparing cricoid pressure, backward-upward-rightward pressure, and bimanual laryngoscopy. Ann Emerg Med. 2006;47:548–55.
5. Walls RM, Brown CA 3rd, Bair AE, Pallin DJ, NEAR II Investigators. Emergency airway management: a multi-center report of 8937 emergency department intubations. J Emerg Med. 2011;41: 347–54.

Laryngeal Mask Airway

11

Sohan Parekh

11.1 Indications

- Rescue device in a failed intubation
- Initial device in a predictably difficult airway
- Temporizing airway prior to definitive endotracheal intubation or surgical airway

11.2 Contraindications

- Absolute
  - Inadequate mouth opening
- Relative
  - Neck trauma/injury/radiation
  - High risk of aspiration

S. Parekh, MD
Department of Emergency Medicine,
University of Texas at Austin Dell Medical School,
Austin, TX, USA

Department of Emergency Medicine,
University Medical Center Brackenridge,
Austin, TX, USA
e-mail: sohan.parekh@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_11

11.3 Types (Fig. 11.1)

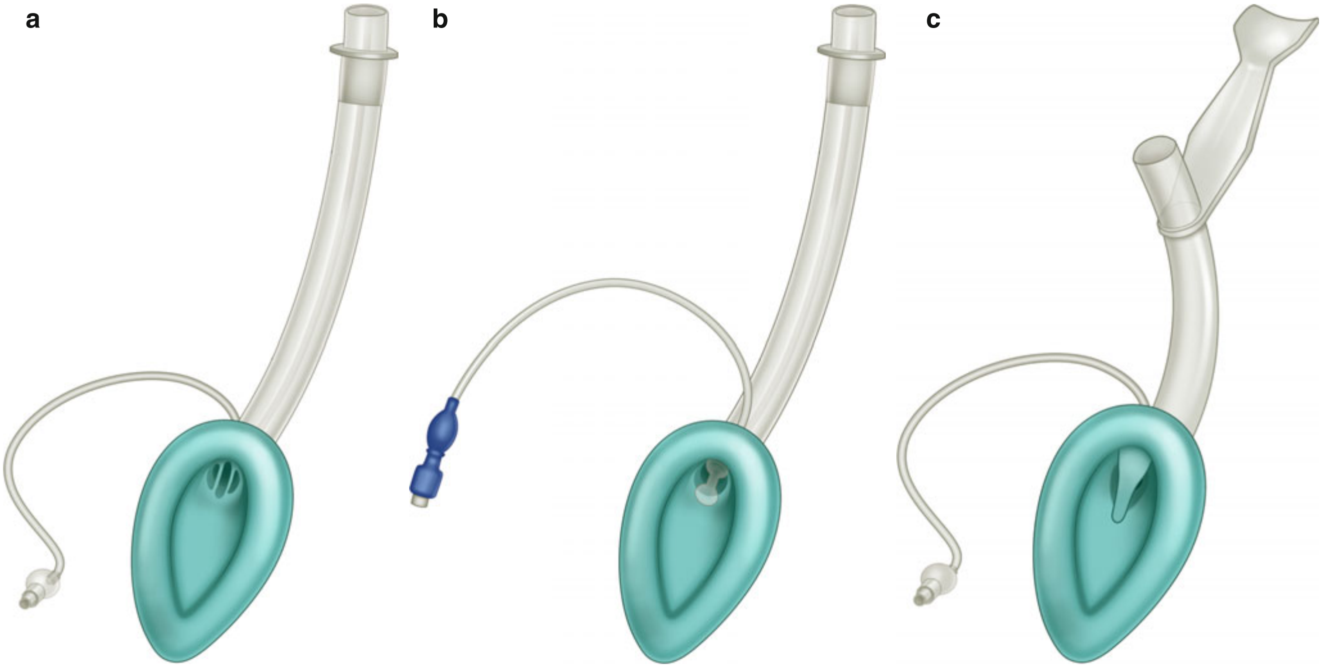

Fig. 11.1 Types of laryngeal mask airway: (**a**) LMA Unique, (**b**) LMA Classic Excel, (**c**) LMA Fastrach

11.4 Materials and Medications

- Appropriately sized laryngeal mask airway (LMA; LMA Unique/LMA Classic Excel/LMA Fastrach) and corresponding syringe (Table 11.1)
- Surgilube
- Bag valve mask
- Continuous end-tidal CO_2 ($EtCO_2$) or colorimetric $EtCO_2$ detector
- 8-mm or smaller endotracheal tube (ETT) (for Fastrach intubating LMA only)

Table 11.1 Laryngeal mask airway sizing

Size	Patient weight (kg)	Maximum cuff inflation volume (mL)	LMA product availability
1	<5	4	Unique
1½	5–10	7	Unique
2	10–20	10	Unique
2½	20–30	14	Unique
3	30–50	20	Unique, Classic Excel, Fastrach
4	50–70	30	Unique, Classic Excel, Fastrach
5	70–100	40	Unique, Classic Excel, Fastrach
6	>100	50	Unique

LMA laryngeal mask airway

11.5 Procedure

11.5.1 LMA Unique or Classic Excel

1. If using a reusable LMA Classic Excel, ensure that it is sterile and inspect it for any damage or wear.
2. Tightly deflate the cuff using a syringe such that it forms a spoon shape (Fig. 11.2).
3. Lubricate the posterior surface of the LMA with sterile lubricating jelly.
4. Stand behind the patient at the head of the bed as in direct laryngoscopy.
5. Place the patient's head in the sniffing position and ensure proper induction and paralysis.
6. Hold the LMA with the index finger of the dominant hand positioned at the juncture of the tube and cuff (Fig. 11.3).
7. Widely open the mouth with the nondominant hand and insert the LMA with the flattened tip flush with the palate.
   - Ensure that the tip of the device does not fold over during insertion.
8. Using the index finger, push the LMA along the curvature of the hard and soft palate (Fig. 11.4).
9. Continue to insert the LMA into the hypopharynx until resistance is felt. (At this point the tip of the LMA is in the esophagus.)
10. Stabilize the tube with the nondominant hand and remove index finger of the dominant hand from the LMA.
11. Inflate the cuff of the LMA to at least half of the maximum value using a syringe.
    - The LMA might move slightly outward during cuff inflation as the LMA positions itself in the hypopharynx.
12. Confirm placement and adequate gas exchange with continuous $EtCO_2$ capnography or colorimetry.

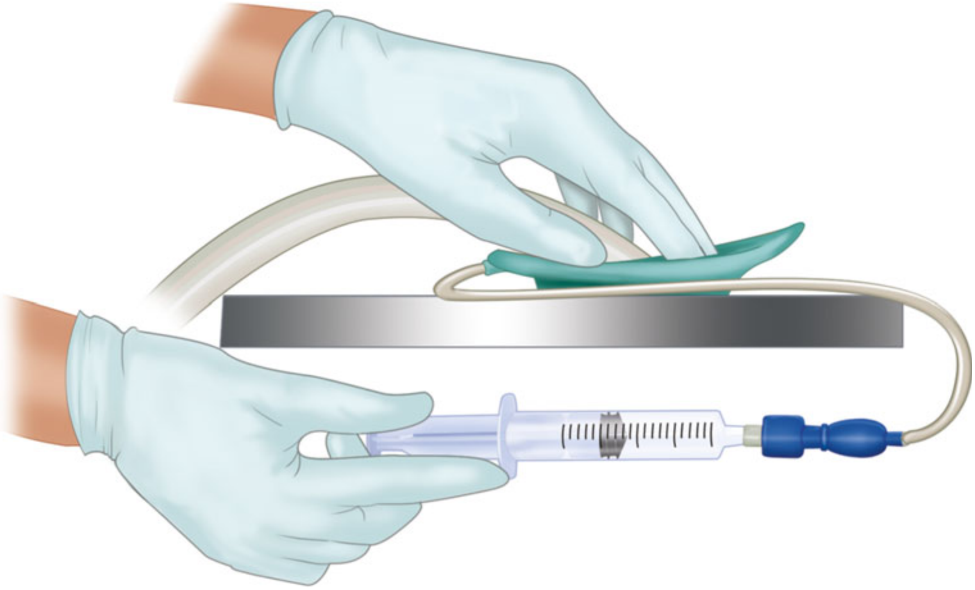

Fig. 11.2 Deflate the laryngeal mask airway (LMA) cuff using a syringe

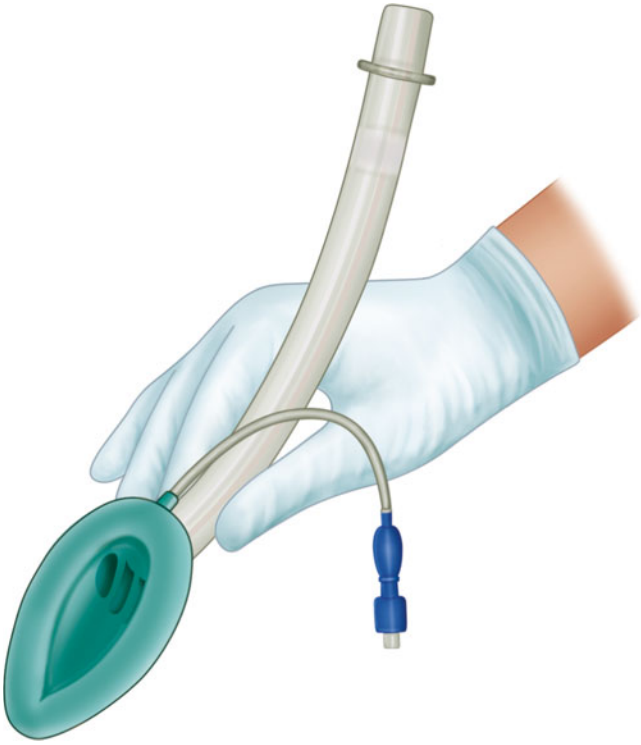

Fig. 11.3 Hold the LMA with the index finger positioned at the juncture of the tube and the cuff

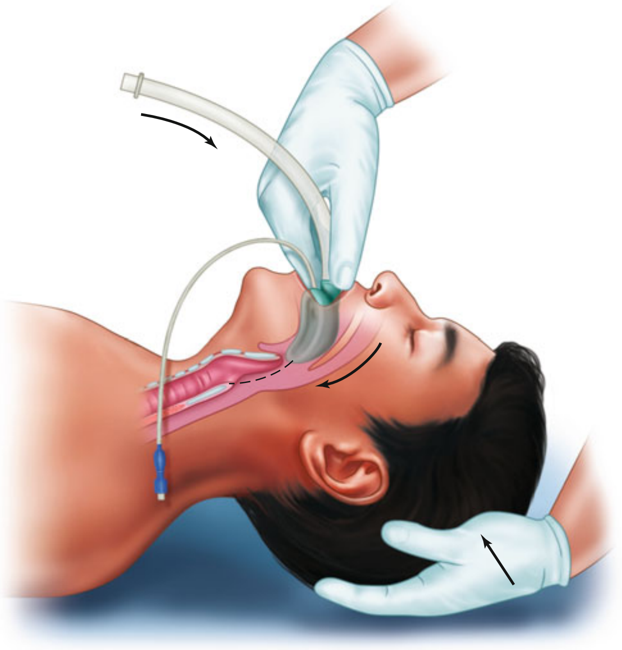

Fig. 11.4 Use the index finger to guide the LMA along the hard and soft palate

11.5.2 LMA Fastrach

1. If using a reusable LMA Fastrach, ensure that it is sterile and inspect it for any damage or wear.
2. Tightly deflate the cuff using a syringe such that it forms a spoon shape.
3. Lubricate the posterior surface of the LMA with sterile lubricating jelly.
4. The LMA Fastrach may be inserted from any position with respect to the patient's head.
5. Position the patient's head in the neutral position. Do not extend the head.
6. Widely open the mouth with the nondominant hand.
7. Holding the handle of the LMA Fastrach, insert the device into the mouth, placing the deflated cuff flush with the superior palate.
    - Distribute the lubricant over the superior palate using a side-to-side motion to allow for easier insertion.
    - Ensure that the tip of the device does not fold over during insertion.
8. Using the handle, gently advance the LMA Fastrach directly into the oropharynx until the curved portion of tube comes into the contact with the patient's chin (Fig. 11.5).
9. At this point use the handle to rotationally advance the device further into the oropharynx following the natural curvature of the palate and posterior pharynx (Fig. 11.6).
    - Do not initiate any rotation until the tube is in contact with the patient's chin.
10. Once resistance is felt, inflate the cuff of the device to at least half of the maximum value using a syringe.
    - Note that the tube is directed slightly caudally when properly inserted.
    - Confirm placement and adequate gas exchange with $EtCO_2$ capnography or colorimetry.

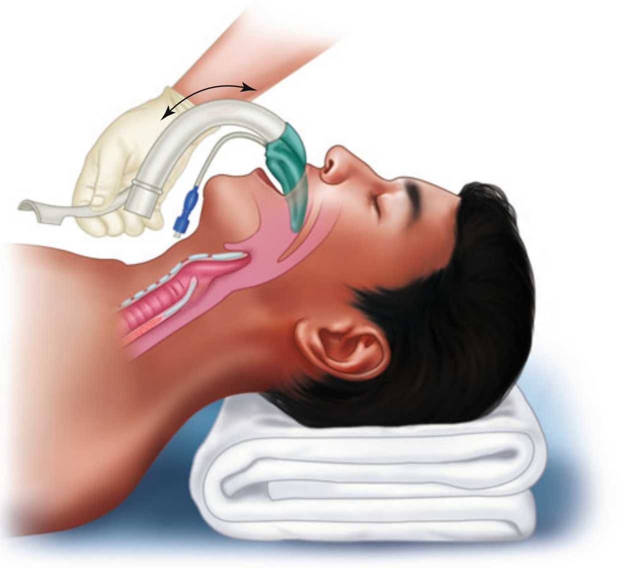

Fig. 11.5 Using the handle, insert the LMA Fastrach such that the posterior surface is in contact with the superior palate

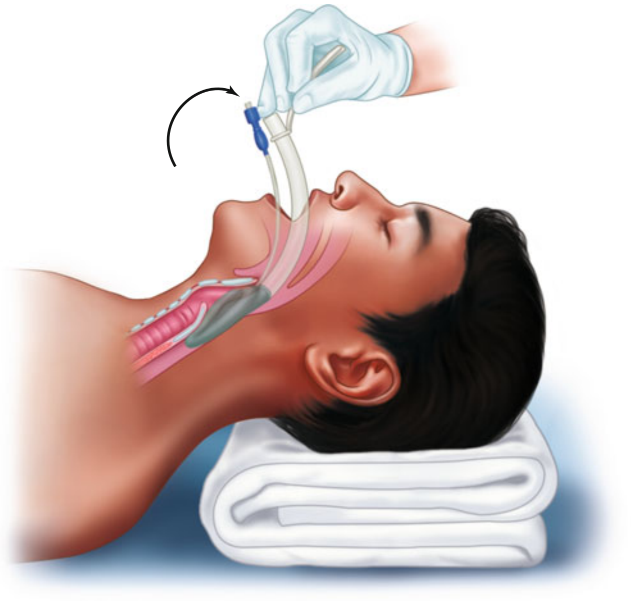

Fig. 11.6 Once the tube is in contact with the chin, use the handle to rotationally advance the device further into the oropharynx

11.5.3 Endotracheal Intubation through the LMA Fastrach

1. Ensure that the ETT will pass freely in the LMA.
2. Lubricate the cuff of the ETT.
3. Firmly hold the handle of the LMA Fastrach with the nondominant hand and insert the ETT to a depth of 15 cm (Fig. 11.7), which places the ETT tip at the point of emergency from LMA Fastrach.
    - Ensure that the tube does not pass beyond 15 cm at this point.
4. Using the handle of the LMA Fastrach, draw the device outward in order to displace the larynx slightly to accommodate insertion of the ETT (Fig. 11.8).
    - Use a lifting rather than a levering motion.
5. Carefully advance the ETT slightly further. If no resistance is felt, continue with insertion of the ETT (Fig. 11.9).
6. Confirm placement and adequate gas exchange with $EtCO_2$ capnography or colorimetry.
7. Once successful confirmation of intubation is established, deflate the cuff pressure on the LMA Fastrach.

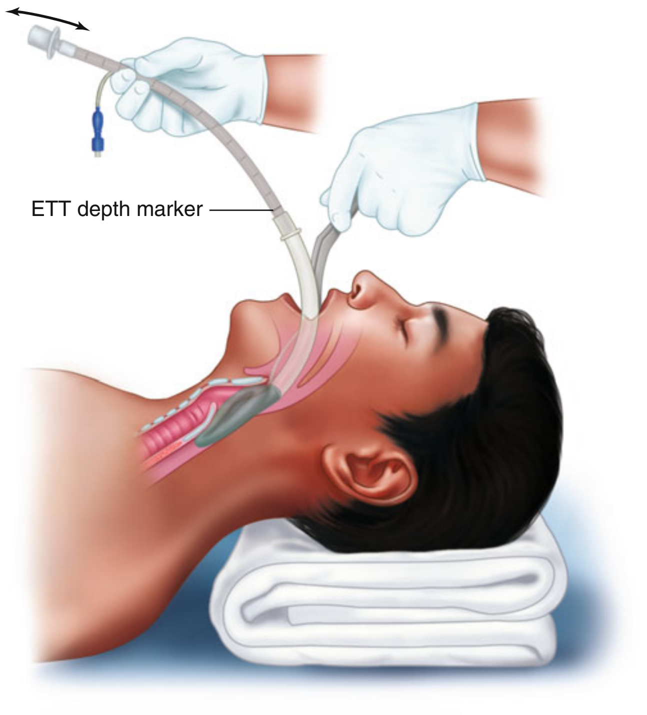

Fig. 11.7 While holding the handle of the LMA, insert the endotracheal tube (*ETT*) to the 15-cm mark

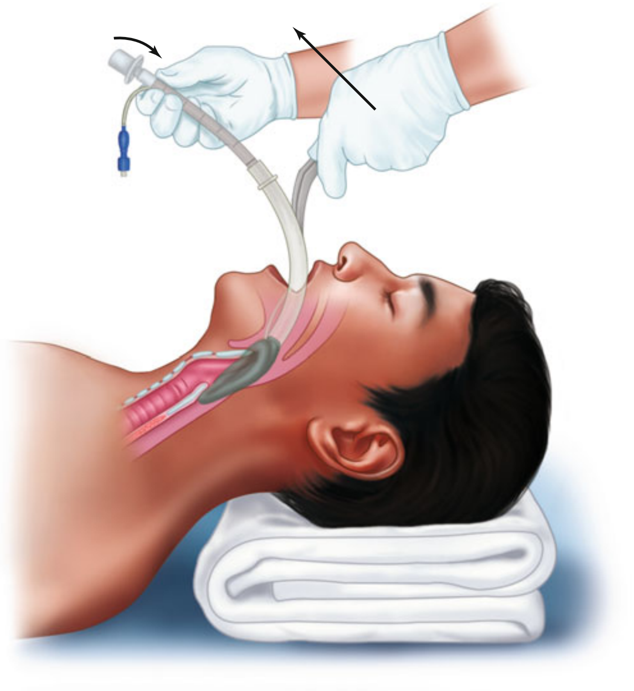

Fig. 11.8 Lift the handle outward to open the glottis for the ETT

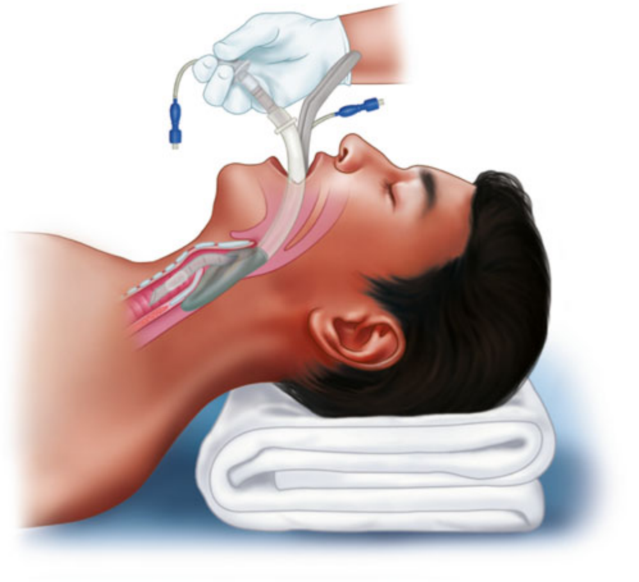

Fig. 11.9 If no resistance is felt during further insertion of the ETT, the ETT may be fully advanced

11.5.4 Removing the LMA Fastrach after Successful Intubation

1. The LMA Fastrach need not be removed immediately, but if this is desired, first adequately oxygenate the patient and then disconnect the patient from the circuit.
2. Remove the airway connector from the proximal end of the ETT.
3. Ensure that the cuff of the LMA Fastrach is entirely deflated.
4. Stabilize the ETT with the nondominant hand, and using the dominant hand, gently ease the LMA Fastrach out by rotating the handle caudally (Fig. 11.10).
5. Once the tube of the LMA Fastrach reaches the proximal end of the ETT, use the stabilizer rod to maintain the position of the ETT while continuing to remove the LMA Fastrach using the handle (Fig. 11.11).
6. After the cuff of the LMA Fastrach has been fully removed from the oral cavity, release the stabilizer rod and ensure stability of the ETT by grasping it distally at the mouth with the nondominant hand (Fig. 11.12).
7. Continue to ease the LMA Fastrach out from around the ETT, ensuring that the pilot balloon and inflation line of the ETT cuff pass through the device (Fig. 11.13).
    - Take care not to rupture the pilot balloon or tear the inflation line of the ETT.
8. Replace the airway connector on the proximal end of the ETT and reconnect the patient to the circuit.

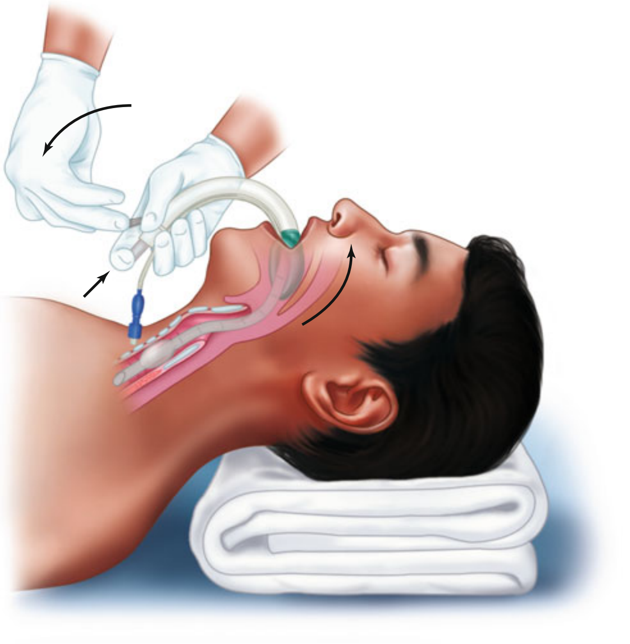

Fig. 11.10 Rotate the handle caudally to gently ease the LMA Fastrach out of the pharynx

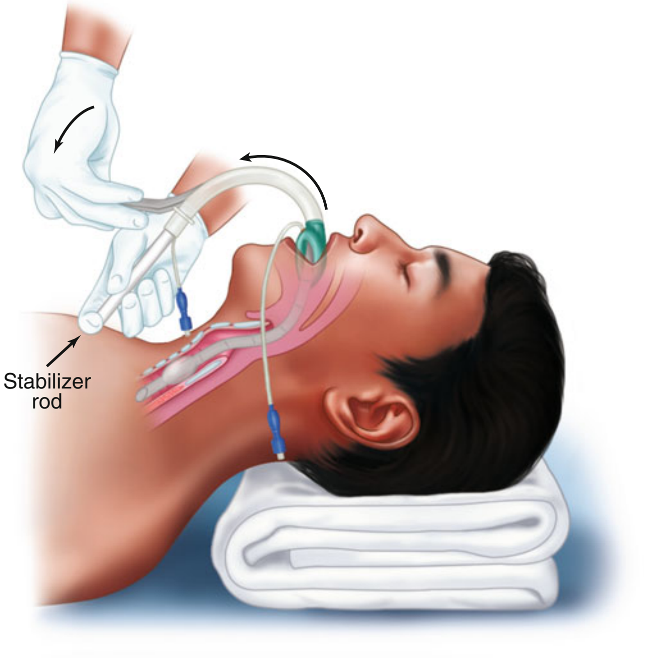

Fig. 11.11 Use the stabilizer rod to allow for further removal of the LMA Fastrach

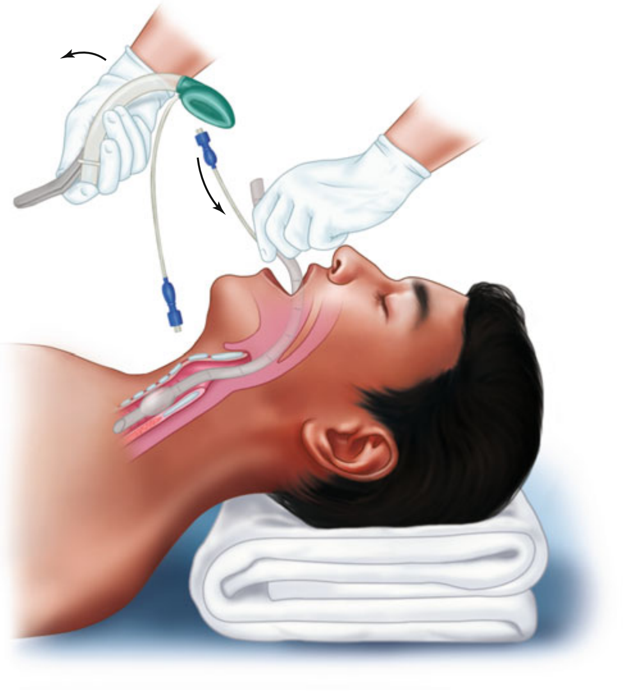

Fig. 11.13 Carefully pass the pilot balloon and inflation line of the ETT cuff through the tube of LMA Fastrach as it is removed

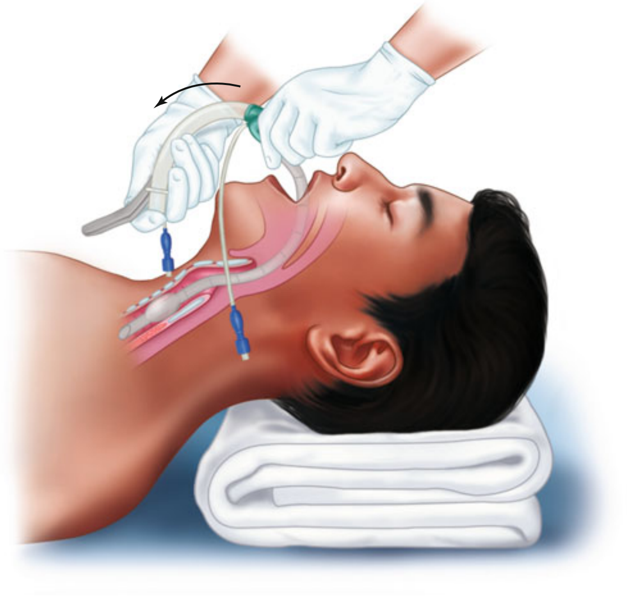

Fig. 11.12 Once the cuff of the LMA Fastrach is out of the mouth, grasp the ETT distally and remove the LMA entirely

11.6 Complications

- Aspiration with resulting pneumonitis
- Ineffective seal resulting in insufficient ventilation
- Coughing, bucking, or breath holding
  - Ensure that the patient is adequately sedated.

11.7 Pearls and Pitfalls

- Cricoid pressure can push the tip of the LMA out of the esophagus and prevent optimal placement.

Selected Reading

Barata I. The laryngeal mask airway: prehospital and emergency department use. Emerg Med Clin North Am. 2008;24:1069–83.

LMA classic LMA flexible LMA classic single use LMA flexible single use instruction manual. Victoria: The Laryngeal Mask Company Limited; 2010.

LMA fastrach LMA fastrach single use instruction manual. Victoria: The Laryngeal Mask Company Limited; 2010.

Pollack CV. The laryngeal mask airway: a comprehensive review for the emergency physician. J Emerg Med. 2001;20:53–66.

Walls RW, Murphy MF, editors. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Combitube

12

Clint Masterson

12.1 Indications

- Need for ventilation and oxygenation in an unconscious, unresponsive, or paralyzed patient
- Rescue airway needed after failed intubation

12.2 Contraindications

- Absolute
  - Awake, responsive patient
  - Intact gag reflex
  - Known esophageal disease
  - Ingestion of caustic substances
  - Child (no Combitubes are made for children)
- Relative
  - D50 or naloxone about to be given
  - Facial trauma

12.3 Materials

- Combitube sized based upon height (Fig. 12.1)
  - >5 ft—size 41 French (cuff inflation 15 and 100 mL)
  - >4 ft to<5.5 ft—size 37 French (cuff inflation 12 and 85 mL)

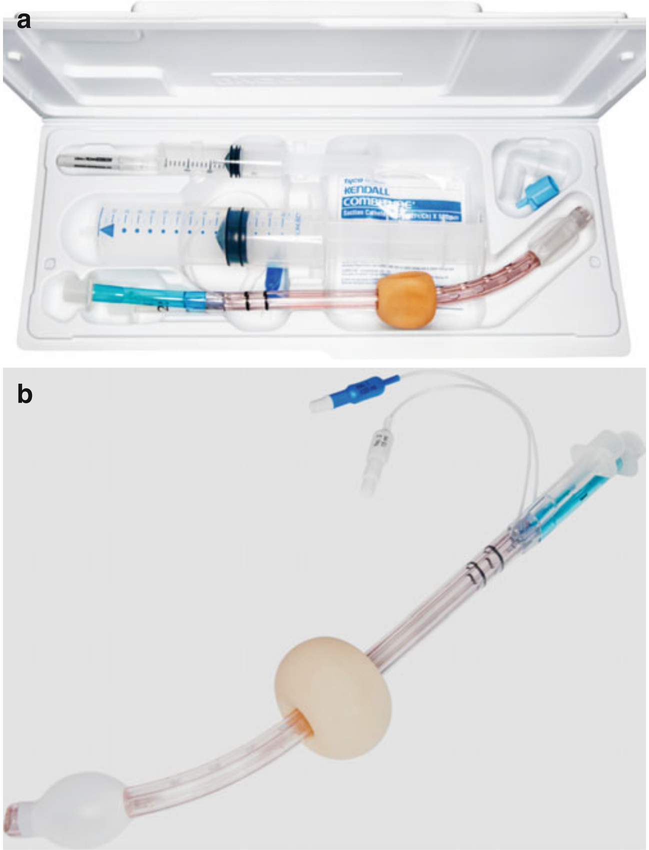

Fig. 12.1 (**a**, **b**) Combitube equipment

C. Masterson, MD
Department of Emergency Medicine,
Mayo Clinic Health System in Fairmont,
Fairmont, MN, USA
e-mail: clint@clintm.us

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_12

12.4 Procedure

1. Test both balloons and cuffs for leaks as one would an endotracheal tube (ETT).
2. Open up the airway.
   - (a) Use a laryngoscope to move the tongue and open the oropharynx.
     OR
   - (b) Use the left hand to elevate the chin, elevating the tongue and pharyngeal tissue.
3. Insert Combitube blindly into the oropharynx until the teeth lie between the two black bands on the proximal Combitube (Fig. 12.2).
4. Inflate the proximal blue cuff until air pressure is produced or the manufacturer-recommended pressure is reached.
   - (a) 85 mL for 37 French Combitube
   - (b) 100 mL for 41 French Combitube
5. Identify placement and attach to oxygen.
   - (a) Ventilate through tube #1 (blue).
   - (b) Auscultate the stomach and lungs.
     - (i) If breath sounds are heard, the Combitube is in its more common esophageal location.
     - (ii) Attach tube #1 to bag valve mask and O_2.
   - (c) ONLY IF gurgling is present over the stomach when tube #1 is ventilated:
     - (i) Ventilate through tube #2.
     - (ii) If breath sounds are heard, the Combitube is in the less common tracheal location.
     - (iii) Attach tube #2 to bag valve mask and O_2.
6. If no breath sounds are heard in either location:
   - (a) Consider obstruction—Combitube may be obstructing the glottis or collapsing the trachea owing to deep proximal cuff inflation.
     - (i) Deflate the cuffs.
     - (ii) Withdraw 3 cm.
     - (iii) Reinflate and start from step 4.
   - (b) Consider equipment failure.
     - (i) Check that balloons are maintaining pressure and intact.
   - (c) Consider reinsertion.
7. Confirm placement with capnogram and pulse oximetry.
8. Secure the Combitube in position (Fig. 12.3).

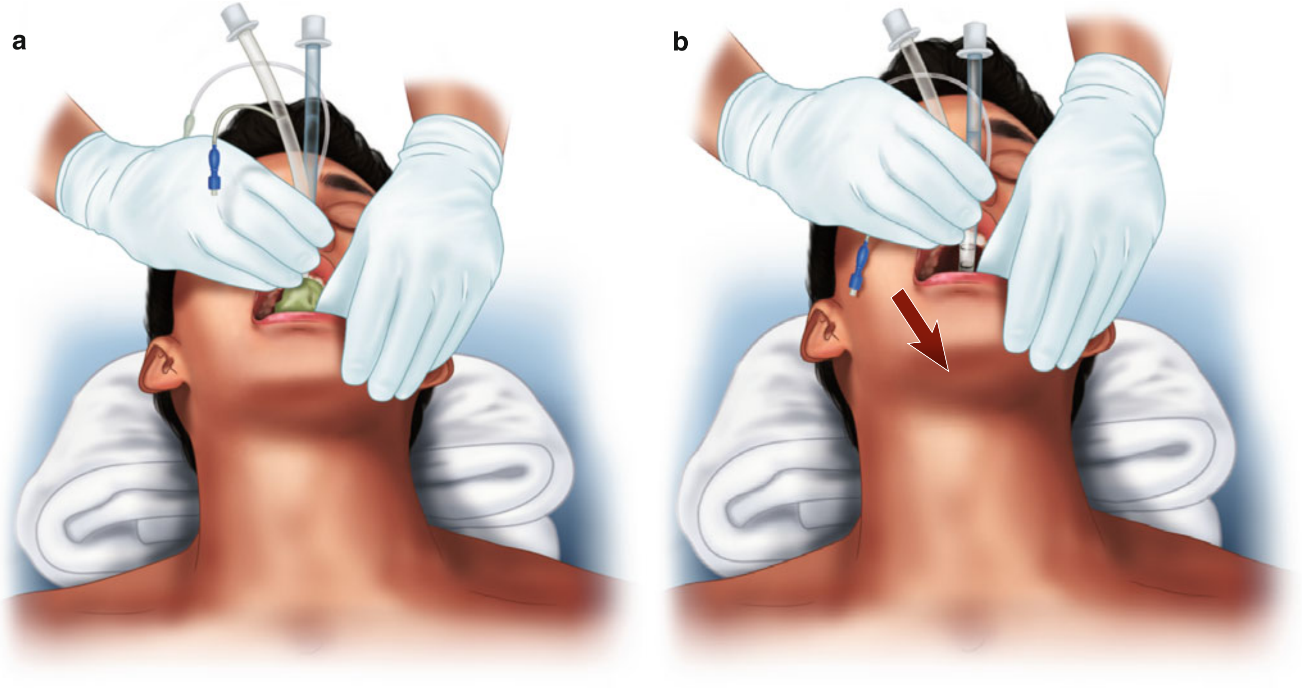

Fig. 12.2 (**a**) Insertion of Combitube. (**b**) Teeth should lie between the two black bands on the proximal Combitube

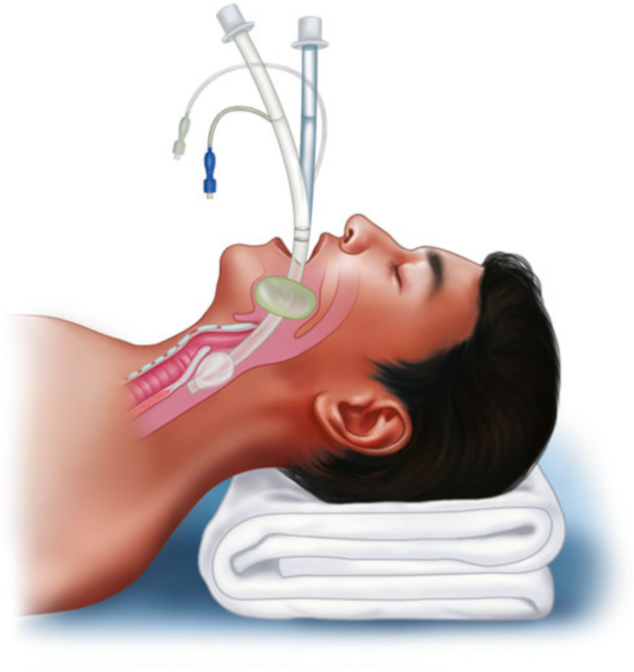

Fig. 12.3 Secure the Combitube in position

12.5 Pearls and Pitfalls

- Pearls
  - In an esophageal intubation situation, a suction tube may be threaded down using tube #2 to decompress the stomach.
  - The Combitube comes with an L-shaped piece that may also be attached to the end of tube #2 to deflect gastric contents away from practitioners.
- Pitfalls
  - After Combitube placement, a definitive airway should be placed when possible.
    - Gastric contents may aspirate despite placement of the Combitube.
    - Combitube should be considered a bridging airway device.
  - Combitubes are associated with a more pronounced hemodynamic stress response than ETTs or laryngeal mask airways (LMAs).
  - Balloon overinflation can lead to esophageal rupture (albeit rare).
  - Combitubes are associated with an increased incidence of sore throat, dysphagia, and upper airway hematomas than ETTs and LMAs.
  - Piriform sinus perforation.

Selected Reading

Agro F, Frass M, Beunmof JL, Krafft P. Current status of the Combitube: a review of the literature. J Clin Anesth. 2002;14:307–14.

Laurin E, Bair A. Devices for difficult airway management in adults. www.uptodate.com. Accessed 14 Mar 2014.

Liem EB. Combitube insertion. University of Florida Department of Anesthesiology, Center for Simulation, Advanced Learning and Technology, Virtual Anesthesia Machine Web site; 2006. http://vam.anest.ufl.edu/airwaydevice/combitube/index.html. Accessed 14 Mar 2014.

Walls R, Murphy M. Manual of emergency airway management. Philadelphia: Lippincott Williams & Wilkins; 2008.

Assessment of the Difficult Airway

13

Melinda W. Fernandez and Lars K. Beattie

13.1 Indications

- Should be performed on all patients that require airway management, conditions permitting
- Respiratory distress
- Airway protection

13.2 Materials and Medications

- None required

13.3 Procedure

1. Anticipating a difficult airway in emergency department patients is the first step to avoiding an unexpected surgical airway.
2. Two mnemonics can be applied quickly and easily to aid in an airway assessment: MOANS and LEMON.

13.3.1 Predicting Bag-Valve-Mask Difficulty

1. Use the mnemonic *MOANS* to assess for possible bag-valve-mask (BVM) difficulty.
   M—mask seal. Will you be able to get a good seal on the face? Predictors of difficulty include facial hair such as a beard, elderly, or emaciated owing to loss of muscle tone in the face.
   O—obesity. Body mass index (BMI) >30.
   A—age (>55 years). Loss of facial muscle tone can make bagging difficult.
   N—no teeth. Although being edentulous makes for an easier intubation, it makes bagging more difficult.
   S—stiff lungs. Acute or chronic lung disease can make a person difficult to bag. In the setting of Trauma, pulmonary contusion(s) and/or other direct lung injuries may increase BVM difficulty.

13.3.2 Predicting Difficult Laryngoscopy

1. Attempts should be made, if at all possible, to assess for a potentially difficult airway. This does not mean you cannot perform direct laryngoscopy if you are anticipating a difficult airway. It does, however, force you to consider all options and to have a solid backup plan in place with backup equipment readily available in the room.
2. Use the mnemonic *LEMON* to predict difficult direct laryngoscopy.
   L—look. A quick look at the patient will tell you a lot. Are there facial injuries; facial anomalies; obesity; short, thick neck; and small mouth or mandible?
   E—evaluate. Use the 3-3-2 rule to quickly assess for the strongest predictors of difficult laryngoscopy.
   - *3:* Open the patient's mouth and three vertically aligned fingers should fit between the incisors.
   - *3:* Three finger widths should fit along the length of the mandible from the mentum to the hyoid bone. Shorter or longer distances may make for a difficult intubation.
   - *2:* Thyromental distance should ideally be two fingers. Measure this from the hyoid to the thyroid.

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida, Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_13

M—Mallampati classification (Fig. 13.1). If patient's condition and situation allow, have the patient open the mouth wide, stick out the tongue, and say "Ahh." Evaluate for visible structures.

- *Class I:* Tonsillar pillars and the entire uvula are visible.
- *Class II:* More than the base of the uvula is visible but no pillars are visible.
- *Class III:* Only the base of the uvula is visible.
- *Class IV:* No uvula or soft palate is visible. Only the hard palate is visible.
- These classifications correlate with the Cormack-Lehane grading system for laryngoscopic views. A Mallampati class I will correlate with a grade 1 view about 99 % of the time, whereas a Mallampati IV will be a grade 3 or 4 view all of the time and a rescue plan with backup equipment immediately available should always be in place [1, 2].

O—obstruction. Observe for anything that can get in the way (e.g., the tongue, dentures, blood, vomit, foreign body, edema, redundant tissue).

N—neck mobility. If patient's condition and situation allow, have the patient flex and extend the neck to evaluate mobility. Many patients in the emergency department have limited neck mobility. Examples include the trauma patient who arrives in cervical collar immobilization or a patient with degenerative or rheumatoid arthritis.

3. The *"6-D"* method is another assessment tool that can be used to predict difficult laryngoscopy and intubation. This method can be remembered by the fact that the word "difficult" begins with the letter "D":

D̲isproportion

- Increased tongue size in relation to pharyngeal size
- Airway swelling or trauma

D̲istortion

- Neck mass, hematoma, abscess, previous surgical airway, arthritic neck changes

D̲ecreased thyromental distance

- Anterior larynx and decreased mandibular space.
- Look for a receding chin or greater than three fingerbreadths from the mentum to the hyoid bone.

D̲ecreased inter-incisor gap

- Reduced mouth opening.
- Look for less than two to three fingerbreadths placed vertically in the patient's open mouth.

D̲ecreased range of motion in any joints of the airway

- Limited head extension
- Previous neck radiation and/or surgery
- Neck contractures

D̲ental overbite

- Oversized, angled teeth disrupt the alignment of airway axes.
- Can decrease the interincisor gap.

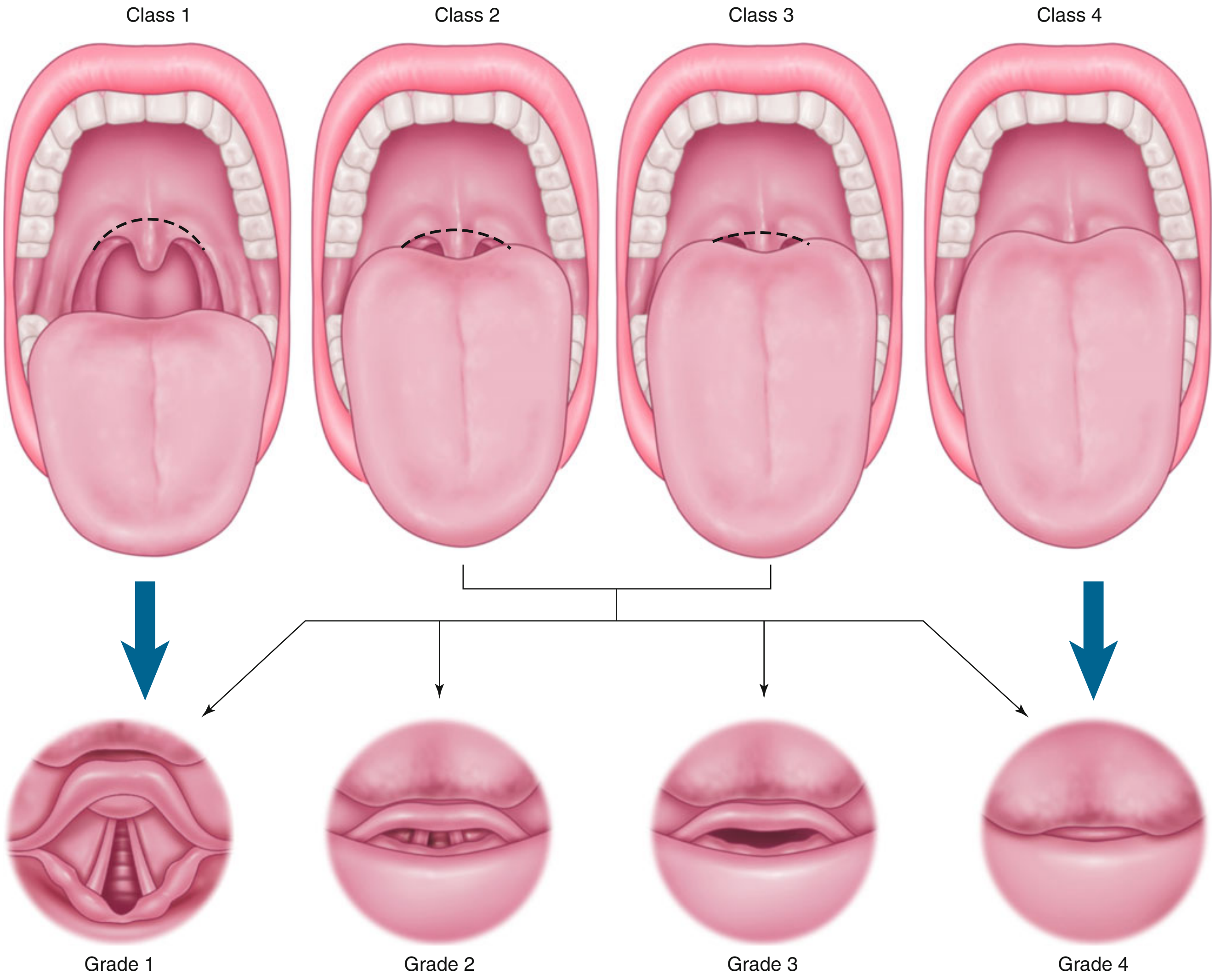
Class 1
Class 2
Class 3
Class 4
Grade 1
Grade 2
Grade 3
Grade 4

Fig. 13.1 Mallampati classification

13.3.3 Optimizing Laryngoscopy in the Obese Patient

- It is essential that emergency medicine physicians are able to successfully intubate the obese patient.
- Proper assessment and positioning will increase the success rate:
  - The goal is to ensure alignment of the oropharyngeal-pharyngeal-laryngeal (OA-PA-LA) airways by placing the patient in the head-elevated laryngoscopy position (Fig. 13.2a).
  - Align the external auditory meatus with the sternal notch along a horizontal line by positioning the patient on a "ramp."
  - The ramp can be created by stacking blankets/towels under the lower back ramping up to the neck and head (Fig. 13.2b).
- While the patient is in position on the ramp, the support is adjusted to minimize head flexion and allow for positioning in the sniffing position.
  - Because of the increased height, a step stool may be required to adequately visualize the airway from the head of the bed.

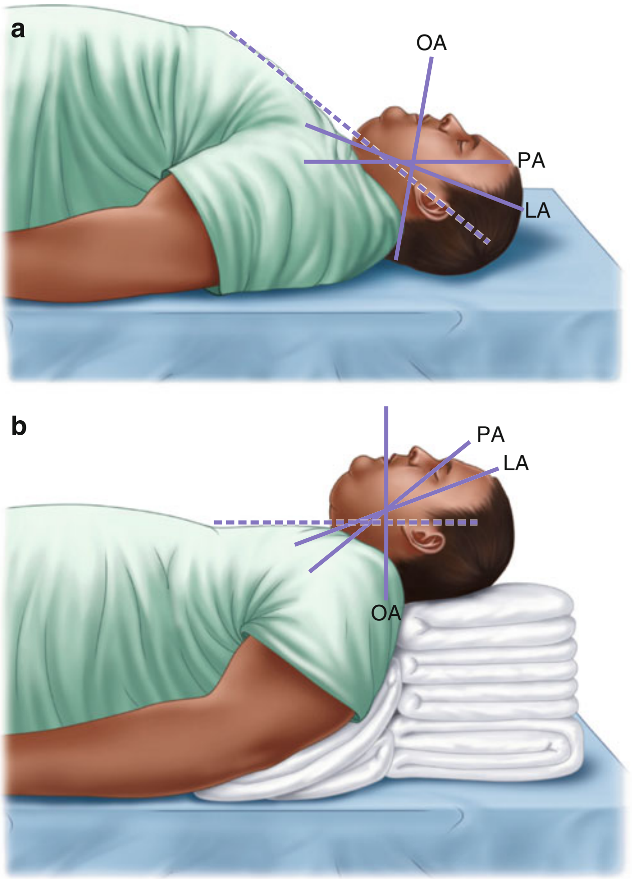

Fig. 13.2 (**a**, **b**) Ramping the obese patient will align the axes of the airway and allow easier direct laryngoscopy when viewed from the head of bed. (*LA* laryngeal airway, *OA* oropharyngeal airway, *PA* pharyngeal airway)

13.4 Pearls and Pitfalls

- Owing to time-sensitive patient care situations, emergency physicians are often not able to perform a thorough airway evaluation on every patient.
- With every airway that you manage and before pushing any drugs, always ask yourself:
  - Will I be able to ventilate this patient?
  - Will I be able to intubate this patient?
  - What is my difficult airway plan if I encounter trouble?
  - Will I be able to perform a surgical airway, if necessary?
- Be sure you have a solid backup plan A, B, and C before pushing any drugs.

References

1. Lee A, Fan LT, Gin T, Karmakar MK, Ngan Kee WD. A systematic review (meta-analysis) of the accuracy of the Mallampati tests to predict the difficult airway. Anesth Analg. 2006;102:1867–78.
2. Boschert S. Think L-E-M-O-N when assessing a difficult airway. ACEP News. Nov 2007.

Selected Reading

Murphy M. Bringing the larynx into view: a piece of the puzzle. Ann Emerg Med. 2003;41:338–41.

Rick J. Recognition and management of the difficult airway with special emphasis on the intubating LMA-Fastrach/whistle technique: a brief review with case reports. BUMC. 2005;18:220–7.

Roberts J, Hedges J. Clinical procedures in emergency medicine. 5th ed. Philadelphia: WB Saunders; 2009. p. 60–2.

Wilson W. Difficult intubation. In: Atlee J, editor. Complications in anesthesia. Philadelphia: WB Saunders; 1999. p. 138–47.

GlideScope

14

Sohan Parekh

14.1 Indications

- Initial device in a predictably difficult airway
- Rescue device in a failed intubation

14.2 Contraindications

- Absolute
  - Inadequate mouth opening
- Relative
  - Blood, vomit, or other secretions that can coat and obstruct the camera lens

14.3 Materials and Medications

- GlideScope video monitor with video cable (GlideScope Video Laryngoscope [GVL] system) (Fig. 14.1) or appropriate-size video baton (Cobalt System) (Fig. 14.2)
- Appropriate-size reusable video laryngoscope (GVL) or single-use laryngoscope blade (GVL Stat) (Table 14.1).
- Endotracheal tube (ETT)
- Malleable stylet or GlideRite rigid stylet
- 10 mL syringe
- End-tidal CO_2 ($EtCO_2$) capnography or colorimetry

S. Parekh, MD
Department of Emergency Medicine,
University of Texas at Austin Dell Medical School,
Austin, TX, USA

Department of Emergency Medicine,
University Medical Center Brackenridge,
Austin, TX, USA
e-mail: sohan.parekh@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_14

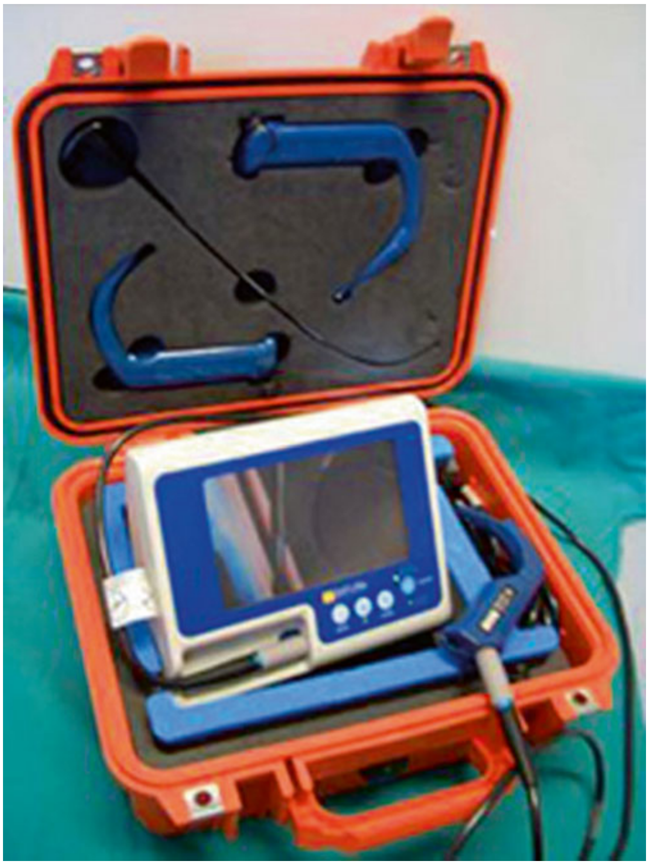

Fig. 14.1 GlideScope GVL system (With kind permission from Springer Science+Business Media: Noppens RR, Werner C, Piepho T. Indirekte Laryngoskopie. *Der Anaesthesist*. 2010;59(2):149–61)

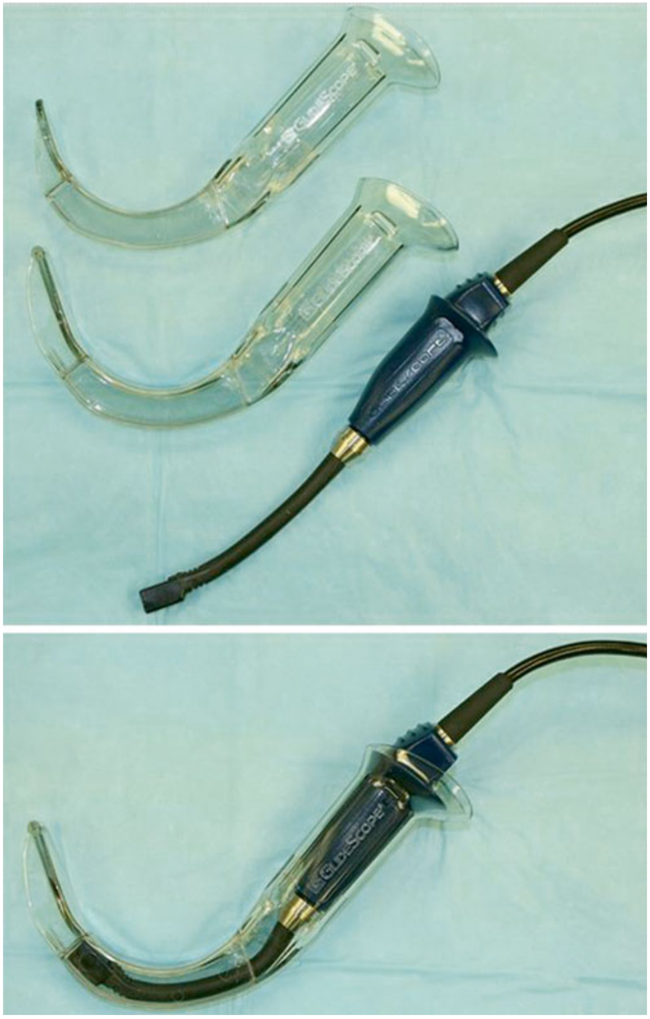

Fig. 14.2 GlideScope Cobalt system (With kind permission from Springer Science+Business Media: Jones PM, Turkstra TP, Armstrong KP, et al. Comparison of a single-use GlideScope® Cobalt videolaryngoscope with a conventional GlideScope® for orotracheal intubation. *Can J Anesthe/Journal canadien d'anesthésie*. 2010;57(1))

Table 14.1 GlideScope sizing

				Video Baton 1–2			Video Baton 3–4	
GVL 1	GVL 2	GVL 3	GVL 4	Stat 0	Stat 1	Stat 2	Stat 3	Stat 4
1.8–10 kg	10 kg—adult	40 kg—morbidly obese	40 kg—morbidly obese	<1.5 kg	1.5–3.6 kg	1.8–10 kg	10 kg—adult	40 kg—morbidly obese

14.4 Procedure

1. Insert the video cable (GVL system) or the video baton (Cobalt system) into the GlideScope video monitor (Fig. 14.3).
2. If using the GVL system, insert the distal end of the video cable into the port on the handle of the reusable video laryngoscope (GVL) (Fig. 14.4).
3. If using the Cobalt system, insert the video baton into the GVL Stat (Fig. 14.5).
    (a) Align the logo on the side of the video baton with the logo on the side of the single-use laryngoscope blade (GVL stat).
    (b) The video baton should slide smoothly and click into the place.
4. Turn on the GlideScope at for at least 30–120 s before use to fully activate the antifog mechanism.
5. Insert a stylet into the ETT. If using a malleable stylet, shape the curvature of the distal end of the tube to conform to the 60° curvature of the laryngoscope blade.
6. Firmly hold the laryngoscope handle in the left hand and ensure that an image can be clearly seen on the video monitor.
7. After ensuring adequate sedation and paralysis, open the mouth wide and insert the laryngoscope blade in the midline underneath the tongue (Fig. 14.6).

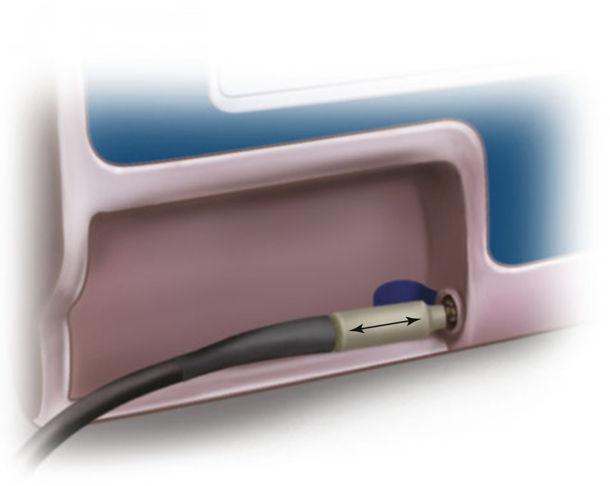

Fig. 14.3 Cable insertion into the video monitor

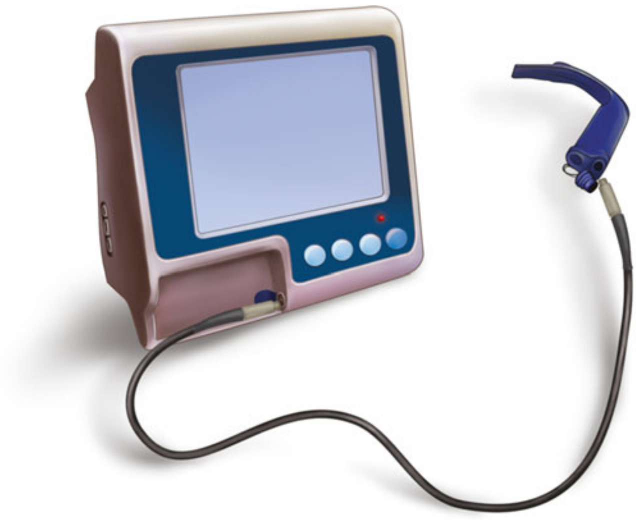

Fig. 14.4 Connect the distal end of the video cable the port on the handle of the GVL (GVL System)

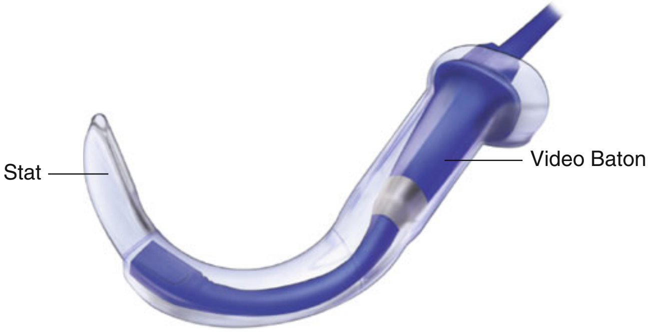

Fig. 14.5 Slide the video baton into the GVL Stat, Cobalt System

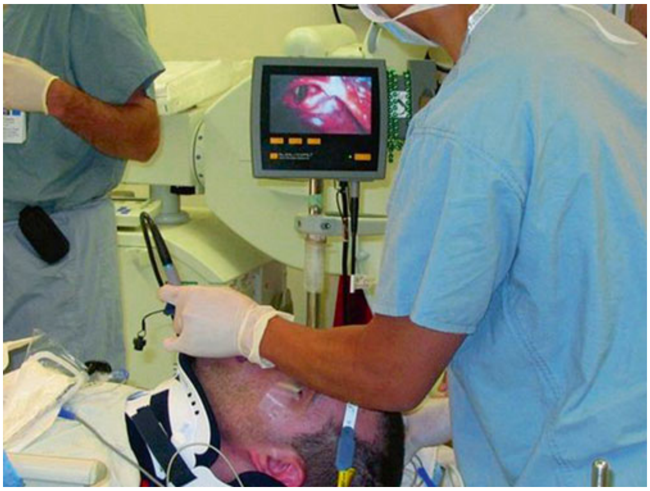

Fig. 14.6 Insert the laryngoscope blade in the midline beneath the tongue (with kind permission from Springer Science+Business Media: Osborn IP, Kleinberger AJ, Gurudutt VV. Chapter 8: Airway emergencies and the difficult airway. In: Levine AI, Govindaraj S, DeMaria S, editors. *Anesthesiology and otolaryngology*. 2013)

8. Looking at the video monitor, advance the laryngoscope blade further into the oropharynx in order to obtain a view of the epiglottis.
    (a) Do not look directly into the oropharynx.
    (b) Movements and adjustments should be guided by the image on the video monitor.
9. Place the laryngoscope blade in the vallecula (analogous to a Macintosh blade) and apply a gentle backward tilt to expose the glottis.
10. In the event that a satisfactory glottic view cannot be obtained, the laryngoscope blade may be advanced and used like a Miller blade to lift the epiglottis out of the way.
11. Directing attention back toward the patient, insert the ETT into the mouth adjacent to the laryngoscope blade.
12. Guide the ETT toward the tip of the laryngoscope such that the end of the ETT emerges on the video monitor.
13. Looking at the video monitor, advance the ETT toward the glottis, and maneuver the tip of the tube between the vocal cords by rotating and altering the angle of the ETT.
    (a) If the ETT tip is posterior to the arytenoids:
        (i) Pull the ETT superiorly, rotate it over the left arytenoid, and gently twist the tube over the epiglottic aperture.
        (ii) Apply external laryngeal manipulation.
        (iii) Withdraw the blade to reduce tilting of the laryngeal axis and lessen the angle of introduction.
    (b) If the ETT abuts the false vocal cords, turn the ETT in the clockwise direction while withdrawing the stylet (Fig. 14.7).
14. Using the thumb, partially withdraw the stylet a few centimeters from the ETT.
    (a) The distal end of the tube should be free of the stylet.
    (b) An assistant can perform this task to allow for greater control and stability of the ETT.
15. Insert the ETT to the desired depth.
16. Fully remove the stylet, inflate the cuff of the ETT using a syringe, and confirm placement with $EtCO_2$, capnography, or colorimetry.

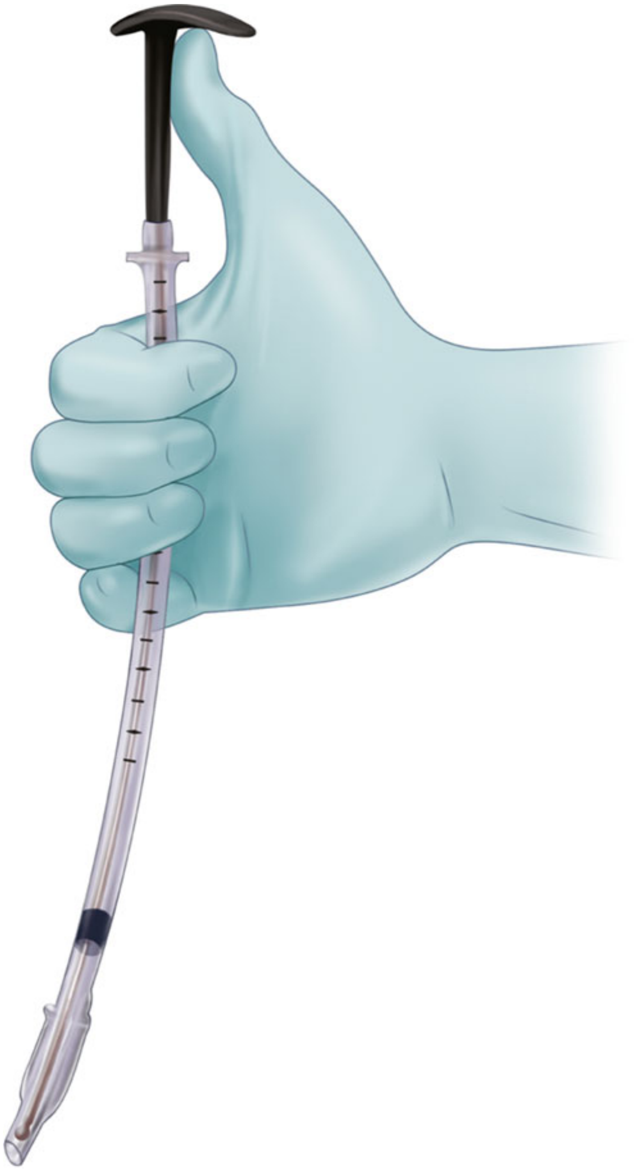

Fig. 14.7 Partially withdraw the stylet from the ETT to allow for passage through the vocal cords

14.5 Complications

- Dental injury
- Airway trauma

14.6 Pearls and Pitfalls

- Unlike conventional laryngoscopy, there is no need to displace the tongue.
- The greatest challenge when intubating with the GlideScope is maneuvering the ETT into the glottis aperture. Alternatives to the conventional technique are:
    – Make a 90° bend in the ETT just proximal to the cuff and insert it in the horizontal direction with the tip toward the right cheek. Once the tube is advanced past the flange of the laryngoscope, rotate it counterclockwise, at which point it should be pointed at the glottis. The tube can then be gently rotated into the glottis.
    – Consider inserting the laryngoscope slightly to the left of the midline upon initial insertion to allow greater space for advancement and maneuvering of the ETT.
- Do not overly lever the laryngoscope or use excessive lifting force after insertion into the vallecula. An adequate view of the glottis is generally easily obtained with minimal effort.
- Display settings can be adjusted using the menu button on the video monitor.

Selected Reading

Cho JE, Kil HK. A maneuver to facilitate endotracheal intubation using the GlideScope. Can J Anaesth. 2008;55:56–7.

GlideScope GVL and Cobalt user's manual & quick reference guide. Bothell: Verathon Inc; 2009–2011.

Kramer DC, Osborn IP. More maneuvers to facilitate tracheal intubation with the GlideScope. Can J Anaesth. 2006;53:737.

Lim HC. Utilization of a GlideScope videolaryngoscope for orotracheal intubations in different emergency airway management settings. Eur J Emerg Med. 2009;16:68–73.

Walls RW, Murphy MF, editors. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Endotracheal Tube Introducer (Bougie)

15

Joseph Rabinovich

15.1 Indications

- During orotracheal intubation, when only epiglottic visualization or partial glottic view is obtained during laryngoscopy.
- Particularly useful when neck mobility is limited, leading to inadequate visualization of the glottis (as in the case with cervical spine immobilization).
- When the glottic opening is narrowed either from pathological causes (burns, trauma, tumor, or other anatomical variation).
- When the direct view of the airway is very narrow, as with limited mouth opening or large tongue. In these scenarios, the endotracheal tube (ETT) can obstruct one's view of the cords during placement.

15.2 Contraindications

- When a failed airway occurs (three unsuccessful attempts at endotracheal intubation and inability to adequately oxygenate)
- When surgical airway is indicated (i.e., upper airway obstruction that prevents passage of the ETT via the orotracheal route)

15.3 Materials and Medications

- ETT introducer (bougie) (Fig. 15.1)
- Water-based lubricant
- Lubricated ETT 6 mm or larger *without* stylet (pediatric bougies are available that accommodate smaller ETTs)
- Standard orotracheal direct laryngoscopy (Miller or Macintosh blade) or video laryngoscopy setup
- Assistant

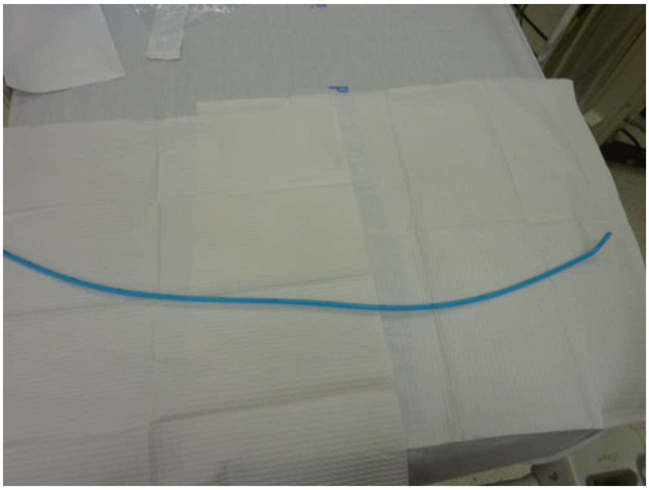

Fig. 15.1 Endotracheal tube introducer (bougie)

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_15

15.4 Procedure

1. The operator first optimizes airway visualization (Fig. 15.2). (Bougie use should not be a substitute for poor technique.)
2. Without losing sight of the airway, the operator asks the assistant to hand him or her the bougie with the coudé tip facing up.
3. The operator directs the bougie tip underneath the epiglottis (Fig. 15.3).
4. Confirmation of placement can be done visually or by tactile sensation:
    (a) A ratchet-like sensation may be felt as the bougie tip is advanced into the airway and slides over the tracheal rings.
    (b) As the bougie is further advanced, the operator may feel the bougie rotate as it enters the bronchus and/or will get a "hold up," the most reliable sign that the bougie is in the trachea [1]. (The "hold-up" sign occurs when the Bougie encounters a terminal bronchus [typically at around 35 cm] and stops advancing.) [2]
5. Once placement is confirmed, the bougie needs to be partially withdrawn to about 25 cm at the lip line.
    (a) Some brands will have a thick black indicator line.
    (b) A sufficient amount of the bougie needs to extend out beyond the proximal end of the ETT.
6. While the operator holds the bougie in place, the assistant threads the ETT over the bougie (Fig. 15.4).
7. The operator now grasps the ETT in her or his right hand and advances it over the bougie.
8. Simultaneously, the assistant holds and stabilizes the proximal end of the bougie.
9. The ETT should be advanced to approximately 23 cm in males and 21 cm in females. The assistant removes the bougie as the operator holds the ETT in place (Fig. 15.5).
10. As the operator holds the ETT firmly in position, the assistant inflates the ETT balloon and withdraws the bougie.
11. Confirmation of proper ETT placement is achieved through traditional means (end-tidal CO_2 detection, auscultation of breath sounds).

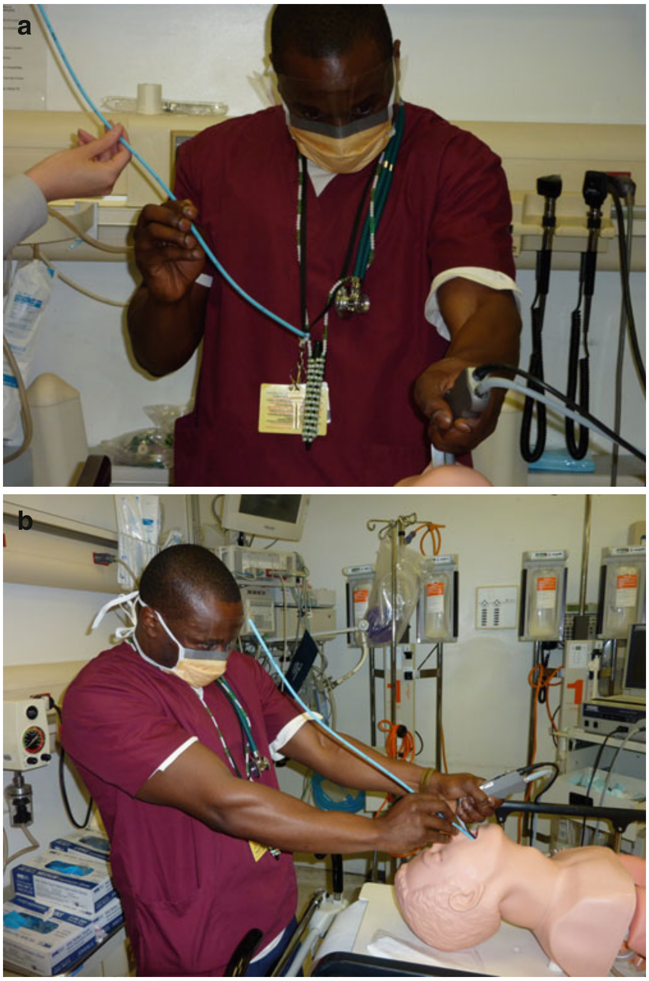

Fig. 15.2 (**a**) Assistant hands operator bougie with coudé tip directed upward, while operator maintains his focus on the target. (**b**) Bougie being placed, parallel with line of sight, underneath epiglottis

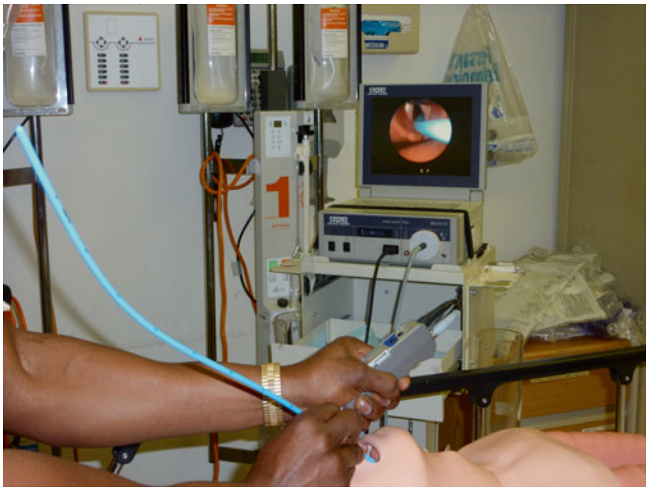

Fig. 15.3 Bougie can be placed into the glottis using direct laryngoscopy or video assistance

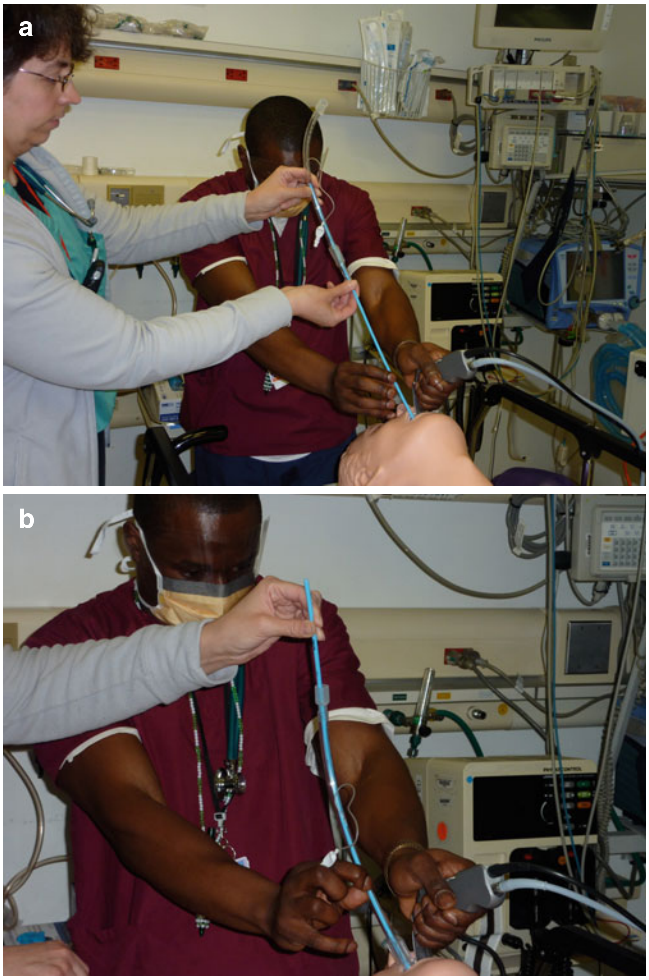

Fig. 15.4 (**a**) Assistant places ETT over bougie and operator withdraws bougie until it protrudes out the top of the ETT, (**b**) while assistant stabilizes the protruding portion of the bougie, the operator railroads the ETT into the airway. The operator continues to support the soft tissues with the laryngoscope blade to facilitate placement

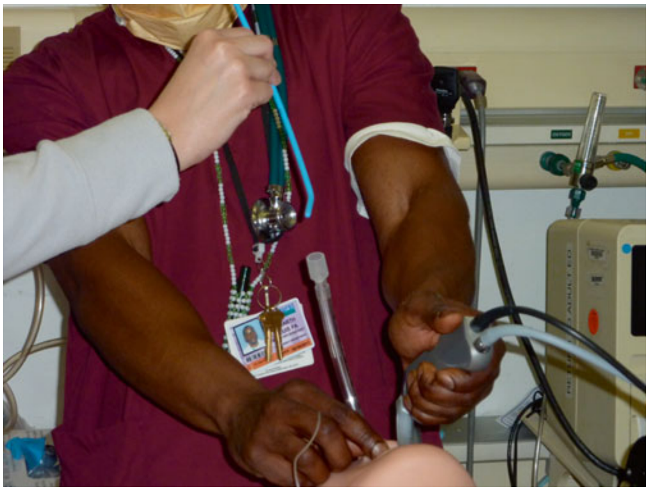

Fig. 15.5 The assistant removes the bougie while the operator stabilizes the endotracheal tube

15.5 Complications

- Trauma to the esophagus, larynx, trachea, or bronchus, including perforation [3, 4].
- In general, complications are rare.

15.6 Pearls and Pitfalls

- The line of sight should be as parallel as possible to the axis of the bougie as it is being passed, allowing better eye-hand coordination. This augments more accurate placement of the bougie tip.
- Maintain the view of the airway as the assistant hands the operator the bougie.
- Keep the laryngoscope in place to support the soft tissues, as the endotracheal tube is slid over the bougie, to facilitate placement.
- If resistance is met during passage, withdraw the ETT slightly (~2 cm) and rotate the ETT counterclockwise one-quarter turn (90°) and reattempt passage.
  - This changes the position of the leading edge of the ETT, which may catch on the posterior laryngeal inlet [2].
  - By rotating the ETT, the leading edge now is anterior facing and is less likely to catch on the arytenoid cartilage along with other laryngeal inlet structures.
- If encountering resistance to ETT placement, consider releasing cricoid pressure (if used).
- Measurement markings on the bougie are aligned with the coudé tip. If not sure of orientation and the loose site of the tip, use the markings to properly orient the tip.

References

1. Kidd JF, Dyson A, Latto IP. Successful difficult intubation. Use of the gum elastic bougie. Anaesthesia. 1988;43:437–8.
2. Murphy MF, Hung OR, Law JA. Tracheal intubation: tricks of the trade. Emerg Med Clin North Am. 2008;26:1001–14.
3. Kadry M, Popat M. Pharyngeal wall perforation – an unusual complication of blind intubation with a gum-elastic bougie. Anaesthesia. 1999;54:404–5.
4. Smith BL. Haemopneumothorax following bougie-assisted tracheal intubation. Anaesthesia. 1994;49:91.

Lighted Stylet Intubation

16

Benjamin M. Mahon and Lars K. Beattie

When lighted stylet intubation is done correctly, the procedure can be very safe, with very little difference in outcome from that of primary laryngoscopy. (*Note*: *several lighted stylet devices*, *such as the TrachlightTM and Light WandTM*, *are no longer being manufactured*, *but these devices are still in use*.)

16.1 Indications

- Difficult/impossible direct laryngoscopy [1, 2]
  - Congenital abnormalities of airway
  - High Mallampati grade [3]
  - Dental appliances
- Failed direct laryngoscopy

16.2 Contraindications

- Absolute
  - Morbid obesity
  - Airway foreign body
  - Expanding neck mass
- Relative
  - Abnormal airway anatomy
  - Airway lesions (e.g., abscess, mass, epiglottitis) that change oropharyngeal anatomy
  - Acute care where concomitant resuscitation requires a well-lit room
  - Lack of familiarity or experience with procedure
  - “Can’t oxygenate, can’t ventilate” situation

16.3 Materials and Medications

- Intravenous (IV) access, O_2, and monitor
- Ambu bag with supplemental oxygen
- Suction (Yankauer and tubing)
- Lighted stylet (LS)
- Endotracheal tube (ETT) 2.5-mm larger than LS with 10-cc syringe
- Surgilube
- Intubation medications (this procedure may be performed as an awake or a rapid-sequence intubation)

B.M. Mahon, MD
Ponciana Medical Center,
Kissimmee, FL, USA
e-mail: drfaust2k9@aol.com

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_16

16.4 Procedure

1. Preoxygenate.
2. Positioning.
   - (a) Sniffing position, pinna at the level of the sternal notch (Fig. 16.1).
   - (b) Skip sniffing position if cervical spine injury is suspected.
3. LS-ETT unit preparation.
   - (a) Insert the wire stylet into the device.
   - (b) Check the LS light.
   - (c) Lubricate the LS with K-Y Jelly.
   - (d) Position the LS just distal to the Murphy eye.
   - (e) Curve the LS to user preference at the line labeled "Bend Here."
4. Administer intubation medications.
5. Have an assistant to apply cricoid pressure.
6. Grasp and elevate the patient's jaw near the corner of the mouth with the operator's thumb, index, and middle fingers, elevating the tongue and epiglottis along with it.
7. Using the free hand, insert the LS-ETT unit into the oropharynx and advance (Fig. 16.2).
8. Use the midline glow in the neck to guide insertion of the LS-ETT (Fig. 16.3).
9. Bright light *below* the thyroid prominence indicates correct placement of the ETT tip.
10. Dim or blurred light or light at the thyroid prominence suggests incorrect positioning (Fig. 16.4).
11. If the transilluminated light is dim, off center, or not seen, esophageal positioning must be considered.
    - (a) Withdraw the LS-ETT unit approximately 2–5 cm.
    - (b) Reposition the patient's head and neck.
    - (c) Reattempt according to steps 5–8.
12. Placement of the ETT (Fig. 16.5).
    - (a) Hold the LS-ETT unit steady with one hand.
    - (b) Check the depth of the ETT and adjust accordingly.
    - (c) Release the LS latch that holds the ETT to the LS.
    - (d) While holding the ETT in position, gently slide the LS out from the ETT.
    - (e) Inflate the ETT balloon.
13. Confirm ETT placement (continuous end-tidal CO_2 [$EtCO_2$], colorimetric capnometry).
14. Secure the ETT.

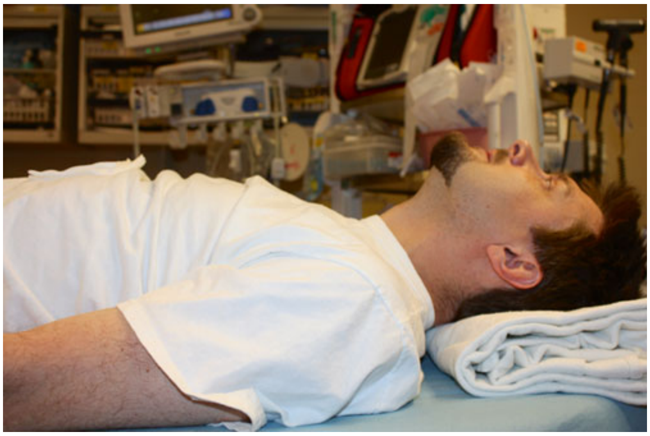

Fig. 16.1 Sniffing position, pinna at the level of the sternal notch

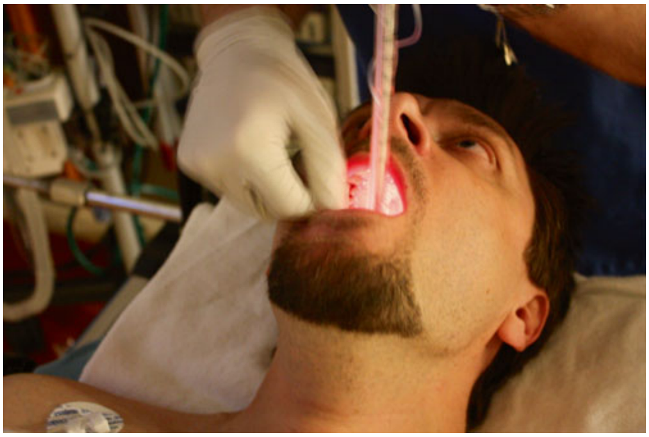

Fig. 16.2 Grasp and elevate the patient's jaw near the corner of the mouth with the operator's thumb, index, and middle fingers, elevating the tongue and epiglottis along with it. Using the free hand, insert the LS-ETT unit into the oropharynx and advance

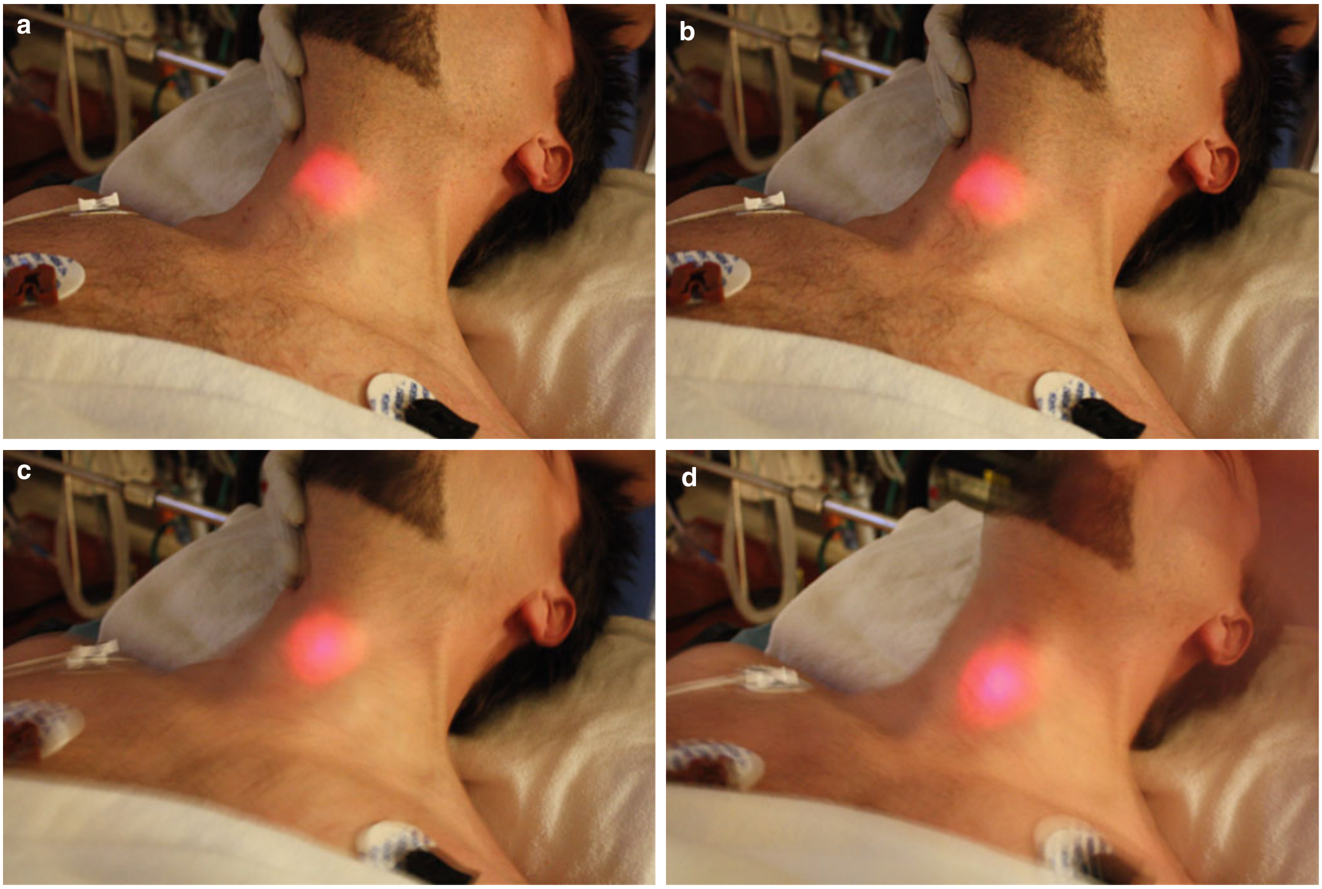

Fig. 16.3 (**a**–**c**) Use the midline glow in the neck to guide insertion of the LS-ETT. (**d**) Bright light *below* the thyroid prominence indicates correct placement of the ETT tip

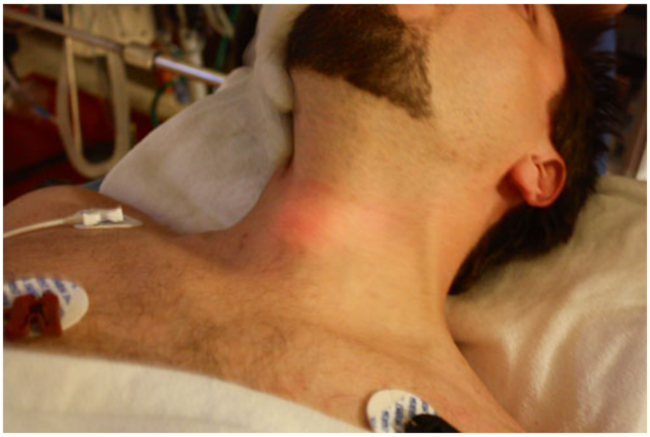

Fig. 16.4 Dim or blurred light or light at the thyroid prominence suggests incorrect positioning

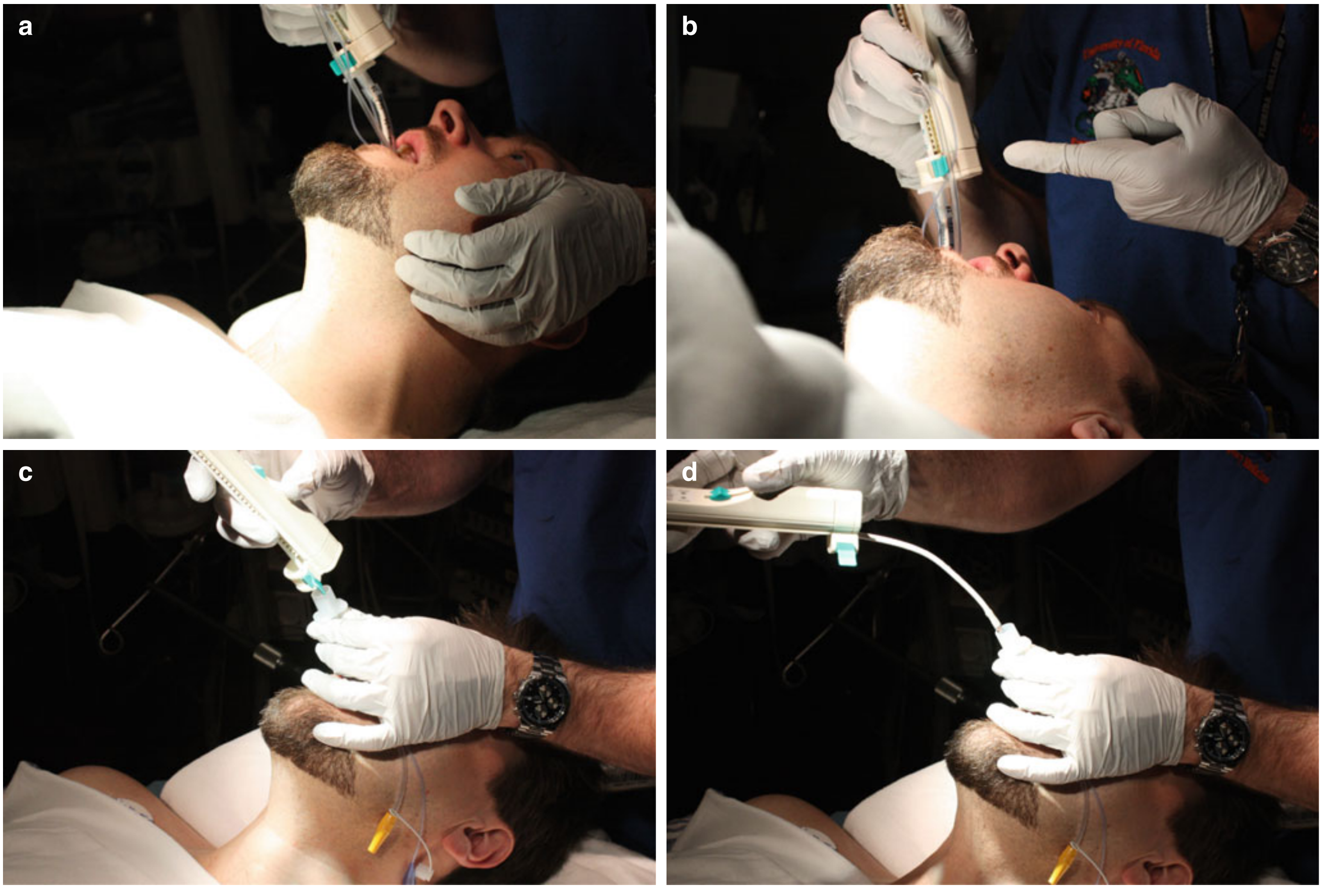

Fig. 16.5 (**a**) Hold the LS-ETT unit steady with one hand. (**b**) Check the depth of the ETT and adjust accordingly. (**c**) Release the LS latch that holds the ETT to the LS. (**d**) While holding the ETT in position, gently slide the LS out from the ETT

16.5 Pearls and Pitfalls

- Pearls
  - – LS-ETT complex—Typically the classic "hockey-stick" shape with the 90° curve just proximal to the cuff is recommended [2].
  - – Dimming the room lights will enhance transillumination.
  - – Pulling the wire stylet out from the LS-ETT unit will make it more pliable and may facilitate its placement in the trachea and removal of the LS.
  - – Some LS devices start to blink after 30 s to prevent bulb overheating.
  - – The LS may be used with nasotracheal intubation, intubation through a laryngeal mask airway (LMA), or conventional laryngoscopy to enhance success.
- Pitfalls
  - – LS intubation should not be used as an emergency airway alternative by a proceduralist unfamiliar with the technique:
    - It is technically complicated and more challenging than many other airway adjuncts in the standard difficult airway algorithm.
    - One study compared the use of four rescue airway devices in the difficult airway algorithm. A success rate of only 20 % was achieved with the Trachlight™ on the first attempt when in the hands of the novice physician when used as a rescue device in their difficult airway algorithm [4].
  - – In very thin patients, transillumination may be visualized quite well even when the LS-ETT unit is in the esophagus:
    - When the unit is in the esophagus, typically the light it will be more diffuse.
    - When the unit is in the trachea, the transilluminated area will be well circumscribed.
  - – In obese patients or patients with significant neck tissue, the transilluminated light from the LS-ETT unit may be dim despite correct positioning in the trachea.

References

1. Agro F, Hung OR, Cataldo R, Carassiti M, Gherardi S. Lightwand intubation using the Trachlight: a brief review of current knowledge. Can J Anaesth. 2001;48:592–9.
2. Davis L, Cook-Sather SD, Schreiner MS. Lighted stylet tracheal intubation: a review. Anesth Analg. 2000;90:745–56.
3. Rhee KY, Lee JR, Kim J, Park S, Kwon WK, Han S. A comparison of lighted stylet (Surch-Lite) and direct laryngoscopic intubation in patients with high Mallampati scores. Anesth Analg. 2009;108:1215–9.
4. Aikins NL, Ganesh R, Springmann KE, Lunn JJ, Solis-Keus J. Difficult airway management and the novice physician. J Emerg Trauma Shock. 2010;3:9–12.

Selected Reading

Langeron O, Birenbaum A, Amour J. Airway management in trauma. Minerva Anestesiol. 2009;75:307–11.

Walls RM, Murphy MF. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. Chap. 11.

Fiber-Optic Stylet Intubation (Rigid and Semirigid)

17

Joseph Rabinovich

17.1 Indications

- For use in routine and predicted difficult oral intubations.
- Similar to a flexible fiber-optic scope with the specific advantages of:
  - Less setup time
  - Less time to perform the procedure
  - Appropriate for routine intubations (and easier to accumulate experience)
  - Rigid enough to lift up the epiglottis
  - Easier to navigate through tissue
  - Less susceptible to being obscured by blood and secretions
  - More durable, more portable, easier to clean, and less expensive
- Particularly useful when neck mobility or mouth opening is restricted.
- Advantageous in awake intubations because it can minimize tissue contact, resulting in less stimulation to the patient's airway and better tolerance.
- Certain stylets can be used to intubate through supraglottic airways such as laryngeal mask airways (LMAs).

17.2 Contraindications

- Complete upper airway obstruction where surgical airway is indicated
- Oral pharyngeal swelling requiring a nasotracheal or surgical approach
- Failed airway and unable to adequately maintain oxygenation

17.3 Relative Contraindications

- Large amounts of blood and secretions may obscure visualization of the airway and cords.
- Very distorted airways. Compared with flexible endoscopy, this device is less maneuverable.

17.4 Materials and Medications

- Endotracheal tube (ETT) 5.5 mm or greater (Fig. 17.1) (Pediatric stylets are also available.)
- Water-soluble lubricant
- Defogging agent
- Optional: Swivel adaptor and meconium aspirator

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine, Elmhurst Hospital Center,
Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_17

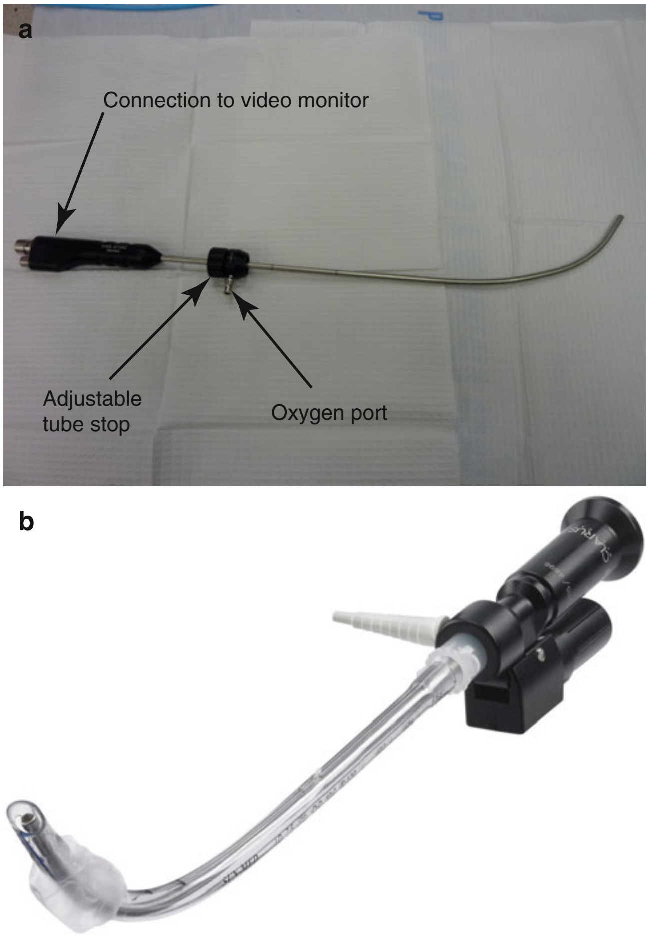

Fig. 17.1 (**a**) Bonfils rigid fiber-optic stylet, Karl Storz endoscopy, (**b**) Levitan FPS malleable fiber-optic stylet

17.5 Procedure

1. Use standard preparation for rapid sequence intubation (RSI) or for awake intubation.
2. Place the ETT over the stylet.
   (a) The ETT should extend slightly beyond the stylet tip.
   (b) If using the malleable stylet:
      (i) Without a laryngoscope: Bend tip to about 70° [1].
      (ii) With a laryngoscope: Bend tip to 35° [1].
   (c) Lubricate the tip of the ETT.
3. Depending on specific scope capability, connect oxygen tubing to the port on the scope.
   (a) This keeps secretions away from the tip while providing an oxygen source.
   (b) Keep flow less than 6 L/min [2].
4. Lens fogging prevention.
   (a) Warm the tip of the scope with the hand or immerse the tip in warm saline.
   (b) Apply the defogging agent.
   (c) Alternatively, chlorhexidine is an effective defogger.
5. Scope insertion.
   (a) Use the nondominant hand to pull the jaw forward while holding the tongue.
      (i) In an awake patient, have the patient protrude the tongue and the operator grasps it with 4×4 gauze. (Alternatively, a Macintosh laryngoscope blade can be used.)
      (ii) The goal is to move the base of the tongue off the posterior pharyngeal wall.
   (b) Initially position the scope horizontally and to the right of the patient's mouth.
   (c) Once the tip is in the oropharynx, reposition the scope vertically (Fig. 17.2a).
   (d) The scope tip should be in the midline or in the retromolar position (per scope design).
   (e) Position the scope and the tip of the ETT in front of the uvula.
   (f) Refer to the eyepiece or video screen to see if there is a clear image of the uvula (Fig. 17.2b).
   (g) Advance the scope very slowly to maintain a view of landmarks, avoiding tissue contact (Fig. 17.3).
6. Once the epiglottis is visualized:
   (a) Continue to advance slowly.
   (b) To get underneath the epiglottis, the tip of the scope may need to be moved posteriorly (by tilting the operator's hand slightly forward) (Fig. 17.4).

7. Once underneath the epiglottis, tilt the scope back, to advance into the more anterior directed airway.
    (a) Make sure the glottic opening is well centered on the screen to facilitate placement (Fig. 17.5).
    (b) If resistance is felt, operator may need to rotate the scope clockwise, or tip scope slightly forward, to prevent the ETT from abutting the anterior aspect of the trachea.
    (c) The ETT may need to be advanced off the rigid stylet to allow further advancement.
8. To remove scope:
    (a) Twist the proximal end of the ETT clockwise.
    (b) Stabilize the tube with the nondominant hand.
    (c) Use the dominant hand to pull the scope forward, following the curvature of the stylet (Fig. 17.6).
    (d) An assistant may be of use during this step.

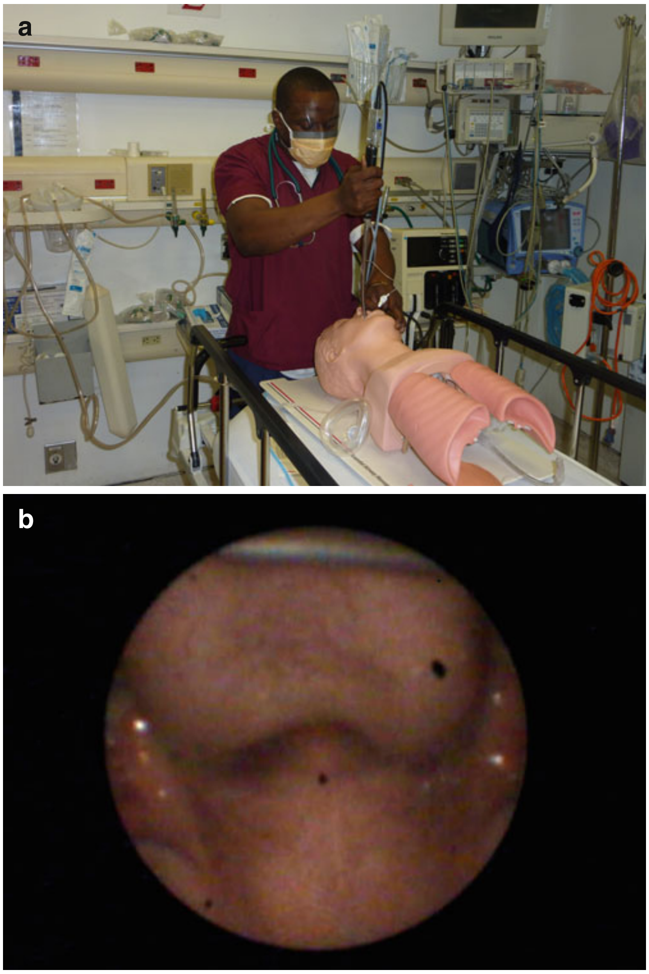

Fig. 17.2 (**a**) Initial placement of scope, under direct visualization; (**b**) tip is positioned in from the uvula

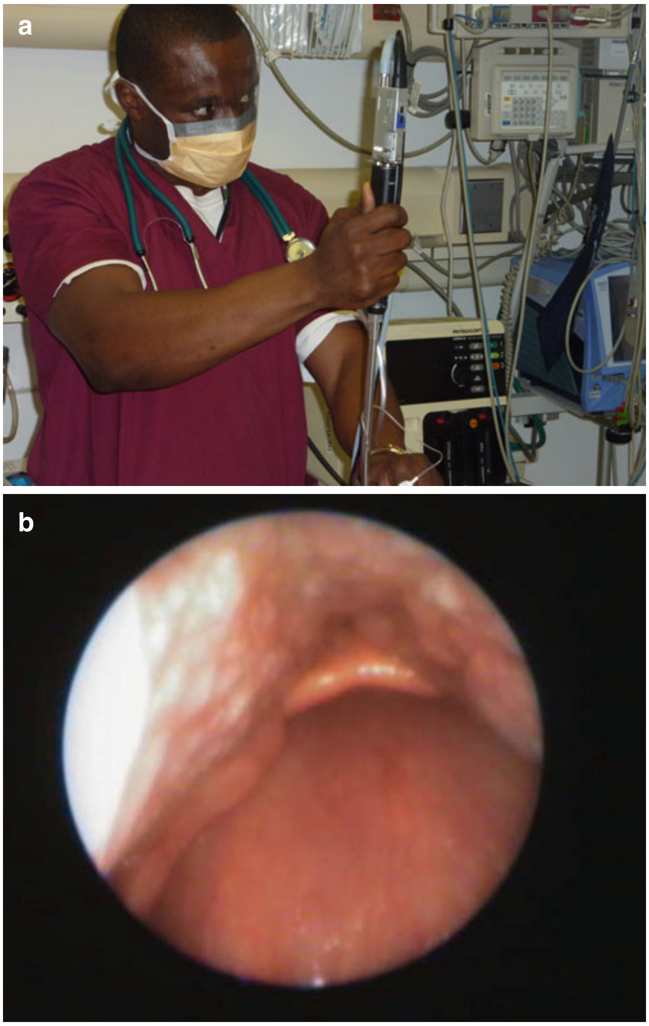

Fig. 17.3 (**a**) Using the video monitor, or through an eyepiece, the operator advances to the next landmark, (**b**) the epiglottis

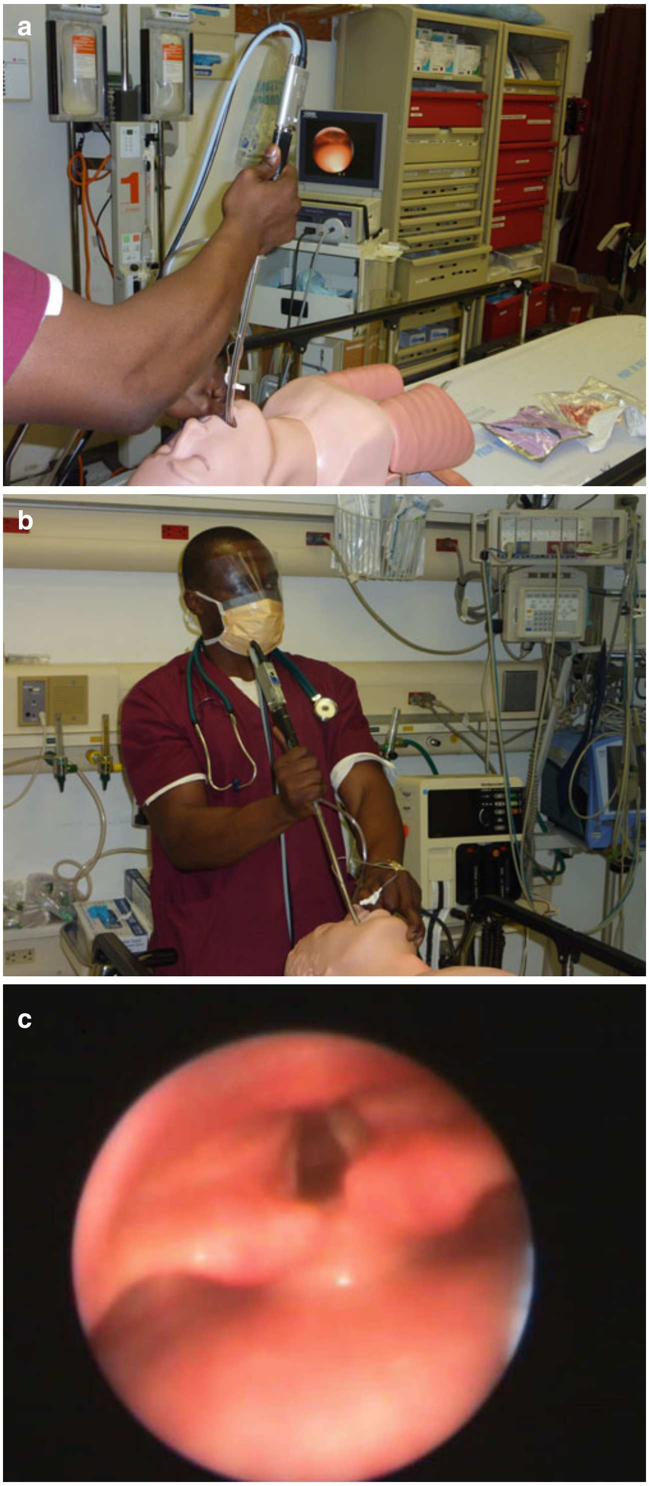

Fig. 17.4 (**a**) Operator tips scope forward to get underneath the epiglottis while advancing; (**b**) once under the epiglottis, the scope may need to be tipped back to advance to the glottic opening; (**c**) operator should try to keep the glottic opening in the center of the screen

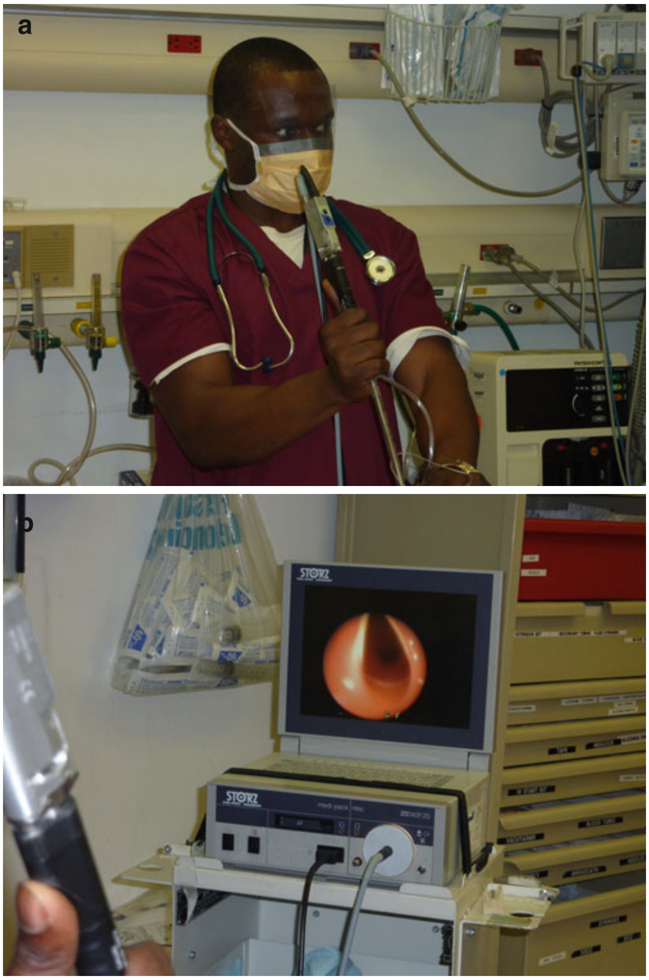

Fig. 17.5 (**a**) Operator is advancing the scope, (**b**) through the glottic opening keeping the image centered

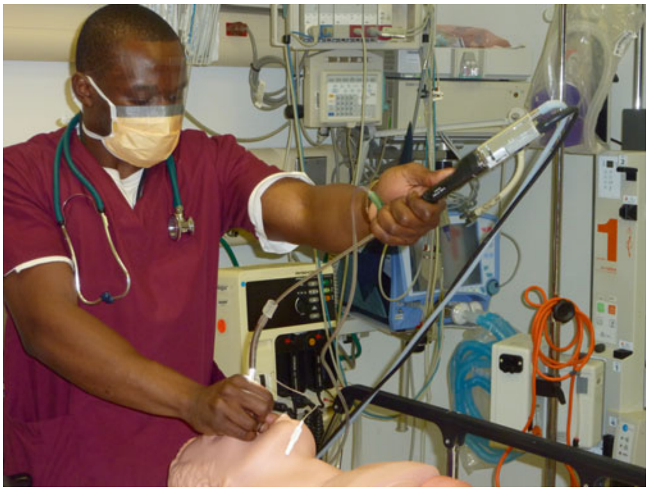

Fig. 17.6 To remove scope, operator must pull it forward while stabilizing the endotracheal tube

17.6 Pearls and Pitfalls

- Pearls
  - When the operator loses perspective or a clear view, withdraw the scope back to the point where identifiable structures are visualized and proceed.
  - The operator can suction through the scope by attaching a swivel adaptor and a meconium aspirator (Fig. 17.7) [3].
- Pitfalls
  - If the scope is advanced too quickly, orientation can be lost.
  - Structures that are too close to the scope will become blurred and unidentifiable.
  - If the scope tip abuts pharyngeal tissue, visualization can become blurred.
  - Flow greater than 6 L/min connected to the oxygen port may result in subcutaneous emphysema (single case report) [2].

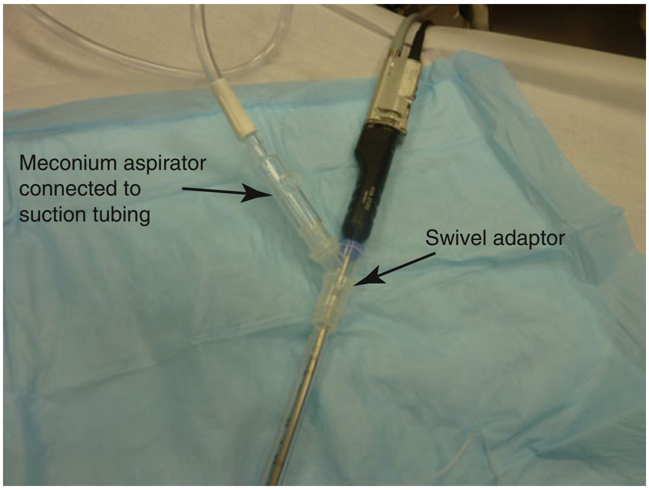

Fig. 17.7 By placing the stylet through the rubber valve of a swivel adaptor, which is then connected to suction via a neonatal meconium aspirator, the operator can now suction through the endotracheal tube

References

1. Levitan RM. Design rationale and intended use of a short optical stylet for routine fiberoptic augmentation of emergency laryngoscopy. Am J Emerg Med. 2006;24:490–5.
2. Hemmerling TM, Bracco D. Subcutaneous cervical and facial emphysema with the use of the Bonfils fiberscope and high-flow oxygen insufflation. Anesth Analg. 2008;106:260–2.
3. Weingart SD, Bhagwan SD. A novel set-up to allow suctioning during direct endotracheal and fiberscope intubation. J Clin Anesth. 2011;23:518–9.

18 Storz Video Laryngoscope

Joseph Rabinovich

18.1 Indications

- Orotracheal intubation for both routine and predicted difficult airways.
- Teaching traditional direct laryngoscopy to novice intubators.
- Ideal for unanticipated difficult airway with the option of intubating indirectly if an adequate direct view is unobtainable.
- An excellent tool when cervical spine precautions need to be taken: Because the video view of the airway is generated by a camera at the tip of laryngoscope blade, less manipulation is required for optimal glottic views.

18.2 Contraindications

- Absolute
  - When orotracheal intubation is contraindicated, e.g., for massive facial trauma, complete upper airway obstruction precluding orotracheal access to the airway
  - In a failed airway (three unsuccessful attempts with inability to maintain adequate oxygenation)
- Relative
  - Blood or copious secretions may prevent indirect viewing of the airway but does not always preclude the use of this device.

J. Rabinovich, MD
Department of Emergency Medicine,
Mount Sinai School of Medicine,
Elmhurst Hospital Center, Elmhurst, NY, USA
e-mail: joerab@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_18

18.3 Materials and Medications (Fig. 18.1)

- Standard materials and medications for endotracheal intubation. Operator should have a backup laryngoscope in case of equipment failure.
- Endotracheal tube (ETT) with or without stylet.
- Water-based lubricant.
- Antifogging agent (not required for C-Mac).

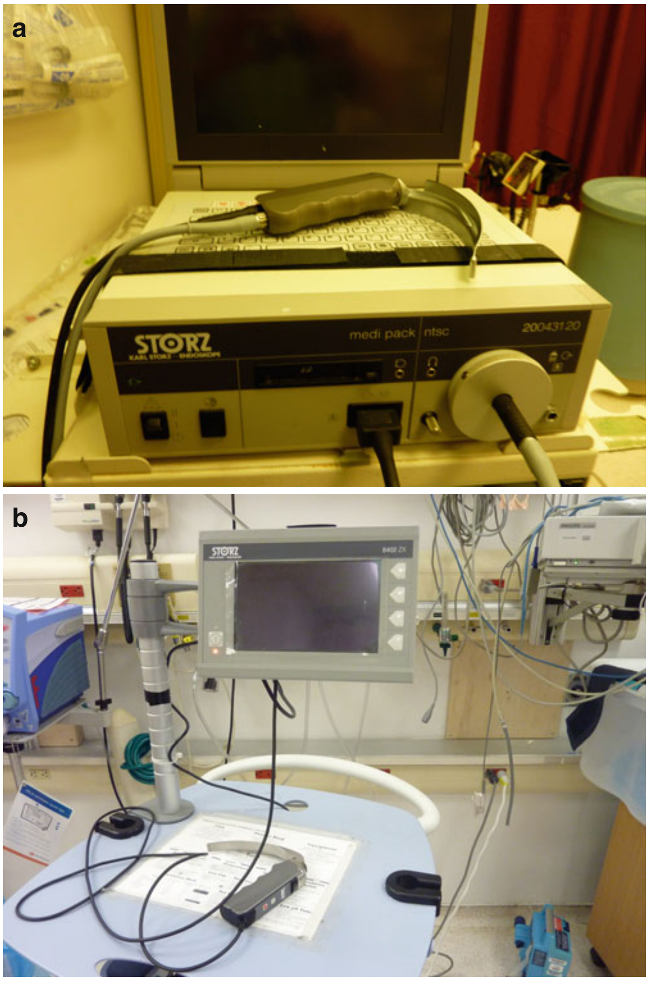

Fig. 18.1 (**a**) Storz video laryngoscope (older version), (**b**) Storz C-Mac (newer version)

18.4 Procedure

1. Standard preparation for orotracheal intubation. If there are no cervical spine precautions, then align the external auditory meatus with the sternal notch [1].
2. Apply antifogging drops to lens at tip of the blade, and/or hold the hand over the blade tip to warm it up to body temperature (older V-Mac model) (Fig. 18.2a).
3. Because blade geometry is the same as in standard laryngoscopes, the insertion technique is identical to that of standard laryngoscopy with a Macintosh blade (Fig. 18.2b).
4. Obtain the best direct view possible.
5. Consider the addition of the backward-upward-rightward pressure (BURP) maneuver [2].
6. Airway maneuvers may be performed by the operator or by the assistant using the video screen as a guide along with operator feedback (Fig. 18.3).
7. The operator has the option of intubating directly with adequate view or indirectly if visualization is improved.
8. Consider using an ETT introducer (bougie—see Chap. 15) if the view is inadequate.
9. Place the ETT, with or without a stylet, into the airway under direct or indirect visualization. If using a stylet, bend the distal end of the tube to approximately 35° as for a standard intubation (Fig. 18.4) [3].
10. Remove the stylet, inflate the cuff of the ETT using a syringe, and confirm placement with end-tidal CO_2, capnography, or colorimetry.

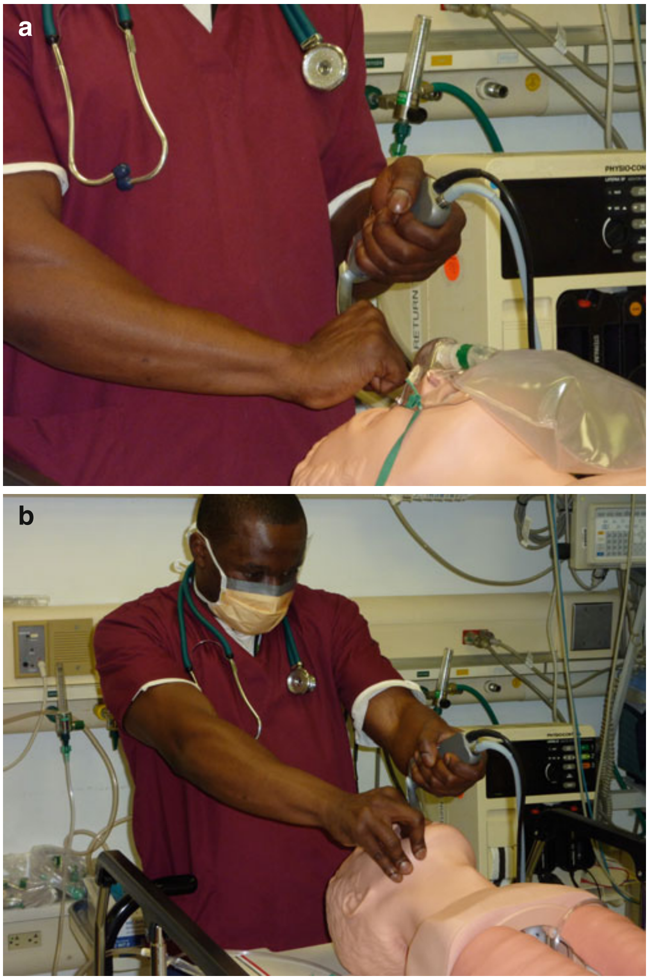

Fig. 18.2 (**a**) Operator warming the blade to prevent fogging, (**b**) laryngoscope blade insertion using the standard direct technique. Operator visually places the blade and optimizes the glottic view

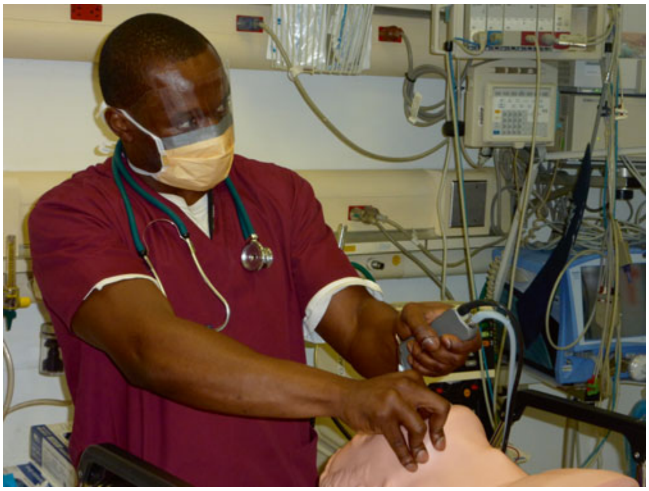

Fig. 18.3 Operator has the choice of using direct visualization (preferred when a novice intubator is learning laryngoscopy) or the indirect video view of the glottis

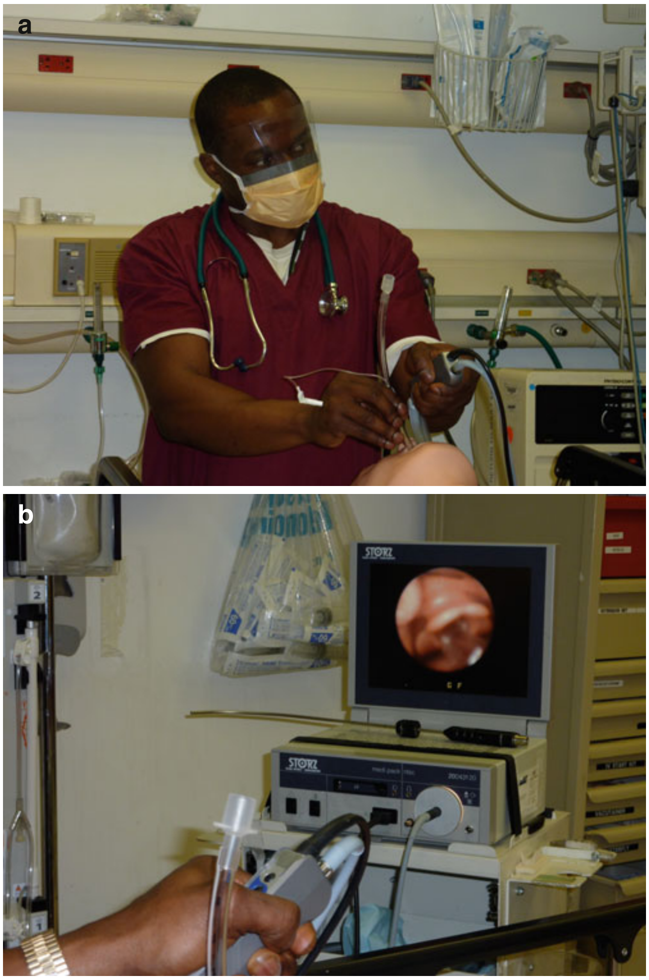

Fig. 18.4 (**a**) Operator initially places the endotracheal tube (ETT) into the oropharynx using direct visualization to avoid injury; (**b**) the tip of the ETT passing through the cords can be confirmed by direct visualization or by watching the video image

18.5 Complications (As with any Direct Laryngoscopy)

- Dental trauma
- Oropharyngeal trauma
- Vocal cord injury

18.6 Pearls and Pitfalls

- The initial placement of the laryngoscope blade and ETT should be done under direct visualization to avoid trauma to the oropharyngeal structures.
- As with direct laryngoscopy, the ETT should not be placed blindly, and the ETT must be seen to pass through the vocal cords to avoid placement in the esophagus.
- View can be obscured by secretions or fogging. If this occurs, the operator may need to remove the blade, wipe it down, and reinsert. The newer model, the C-Mac, is less likely to fog owing to design improvements.
- Observing the ETT pass through the vocal cords can sometimes be easier using the video image because the camera is angled to provide the most optimal view.
- Placement of the blade into the oropharynx can be awkward because the handle of laryngoscope is bulkier than a standard laryngoscope. Its handle is larger and has cables attached to its base. In patients with large anteroposterior diameter chests, the operator may need to rotate the laryngoscope handle toward the right corner of the mouth in order to introduce the blade into the oral cavity and then rotate it back to the proper position.
- These devices are ideal for teaching laryngoscopy. With same blade geometry, the technique is the same as with standard laryngoscopy. As the operator attempts intubation directly, the instructor can observe on the video screen and guide the student. The instructor will also be able to visually confirm that the ETT is entering the trachea.

References

1. Greenland KB, Edwards MJ, Hutton NJ, Challis VJ, Irwin MG, Sleigh JW. Changes in airway configuration with different head and neck positions using magnetic resonance imaging of normal airways: a new concept with possible clinical applications. Br J Anaesth. 2010;105:683–90.
2. Knill RL. Difficult laryngoscopy made easy with a "BURP.". Can J Anaesth. 1993;40:798–9.
3. Levitan RM, Heitz JW, Sweeney M, Cooper RM. The complexities of tracheal intubation with direct laryngoscopy and alternative intubation devices. Ann Emerg Med. 2011;57:240–7.

Selected Reading

Brown 3rd CA, Bair AE, Pallin DJ, et al. Improved glottic exposure with the video Macintosh laryngoscope in adult emergency department tracheal intubations. Ann Emerg Med. 2010;56:83–8.

Niforopoulou P, Pantazopoulos I, Demestiha T, Koudouna E, Xanthos T. Video-laryngoscopes in the adult airway management: a topical review of the literature. Acta Anaesthesiol Scand. 2010;54:1050–61.

19

Cricothyroidotomy

Henry Young II, Shannon Toohey, Bharath Chakravarthy,
and Lars K. Beattie

Up to seven intubation attempts in 1000 end up in a "can't intubate/can't ventilate" situation in the emergency department. These are considered failed airways that may require a surgical airway to maintain ventilation and oxygenation.

19.1 Indications

- Endotracheal tube (ETT) placement attempts unsuccessful
- Failed bag valve mask, laryngeal mask airway, or Combitube ventilation
- Severe facial trauma affecting the upper airway
- Severe oropharyngeal hemorrhage or profound emesis
- Obstruction (foreign body, mass, mass effect)

19.2 Contraindications

- Airway protection achievable using a less invasive strategy
- Tracheal transaction
- Pediatric patients younger than 8 years

19.3 Techniques

- Scalpel-bougie – minimalist
- Scalpel-Trousseau – standard

19.3.1 Scalpel-Bougie

19.3.1.1 Materials and Medications (Fig. 19.1)

- Betadine or chlorhexidine
- Scalpel #11 blade
- ETT (≥6 cm)
- Bougie
- Surgilube
- Bag valve mask

H. Young II, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: hyoungii@ufl.edu

S. Toohey, MD
Department of Emergency Medicine, University of California
Irvine Medical Center, Orange, CA, USA
e-mail: stoohey@uci.edu

B. Chakravarthy, MD, MPH
Department of Emergency Medicine,
University of California Irvine, Orange, CA, USA
e-mail: bchakrav@uci.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_19

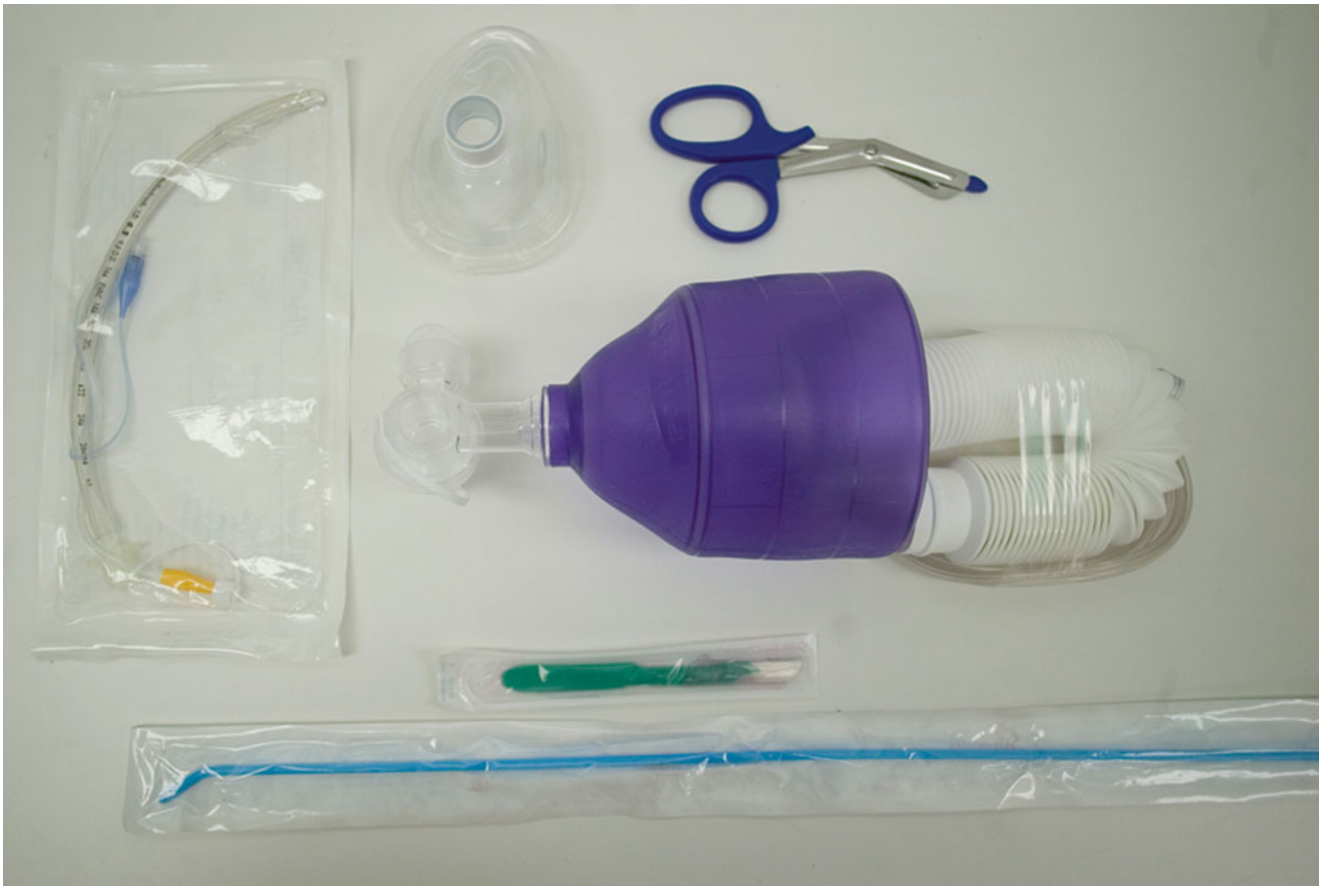

Fig. 19.1 *Right to left*, *top to bottom*: ETT (≥6 cm); bag valve mask; trauma shears; scalpel #11 blade; bougie

19.3.1.2 Procedure

1. Apply topical antiseptic.
2. Remove the 15-mm ETT ventilator connector from the ETT end.
3. Place copious Surgilube on the bougie and railroad over the end of the bougie.
4. Palpate the thyroid notch, cricothyroid membrane, and hyoid bone for orientation (Fig. 19.2a).
5. Stabilize the thyroid cartilage between the thumb and the middle finger of the nondominant hand.
6. Make a vertical skin incision (2–3 cm) over the cricothyroid membrane (Fig. 19.2b).
7. Use the index finger to palpate the cricothyroid membrane.
8. Turn the scalpel 90° and make a 1.5-cm horizontal incision through the lower half of the cricothyroid membrane (Fig. 19.3a).
9. With the scalpel still in the incision, turn it 90°, and insert the bougie into the incision, using the blade as a guide (Fig. 19.3b).
10. Advance the bougie caudally 5–6 cm. Stop if resistance is encountered.
11. Slide the ETT over the bougie into the incision (Fig. 19.4a).
12. Inflate the ETT cuff and ventilate the patient (Fig. 19.4b).
13. Verify the position of the ETT via auscultation, end-tidal CO_2 ($EtCO_2$), and chest radiograph.
14. Secure the ETT.

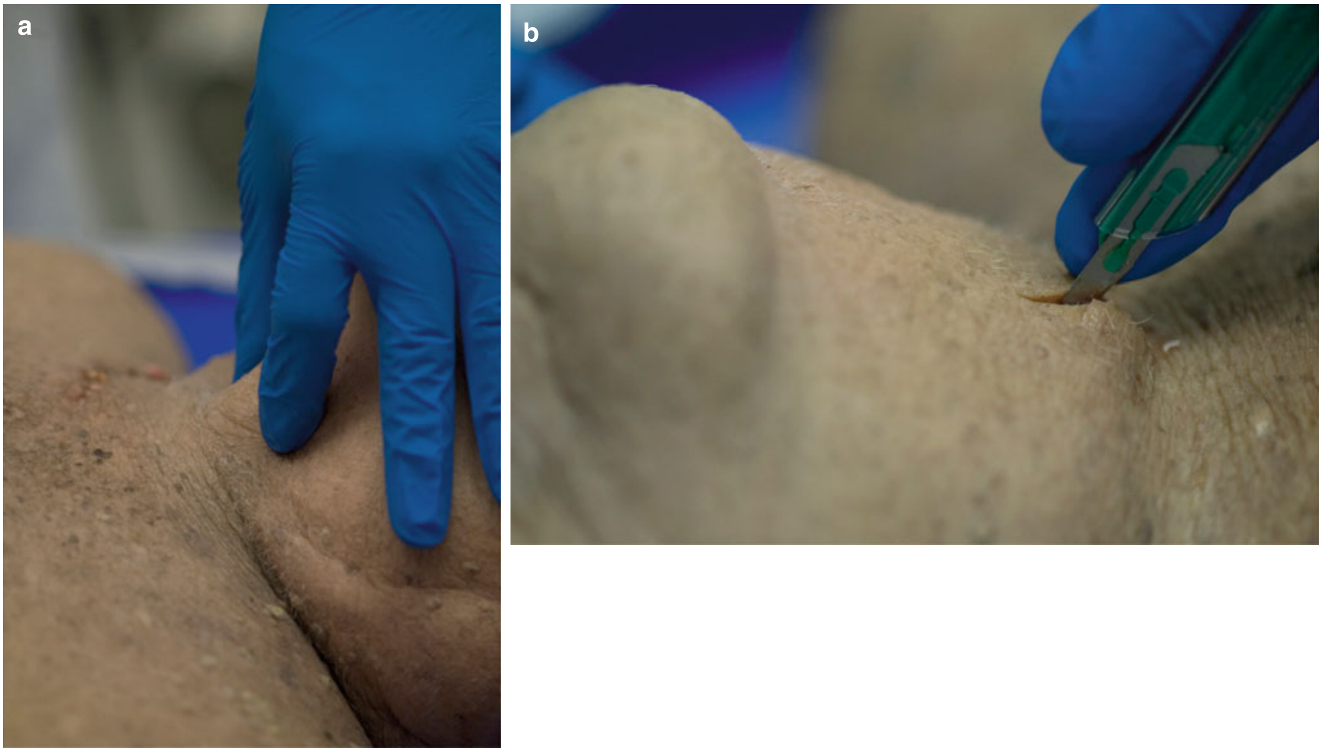

Fig. 19.2 (**a**) Palpate and stabilize the thyroid cartilage, (**b**) make a vertical 2–3 cm incision over the cricothyroid membrane

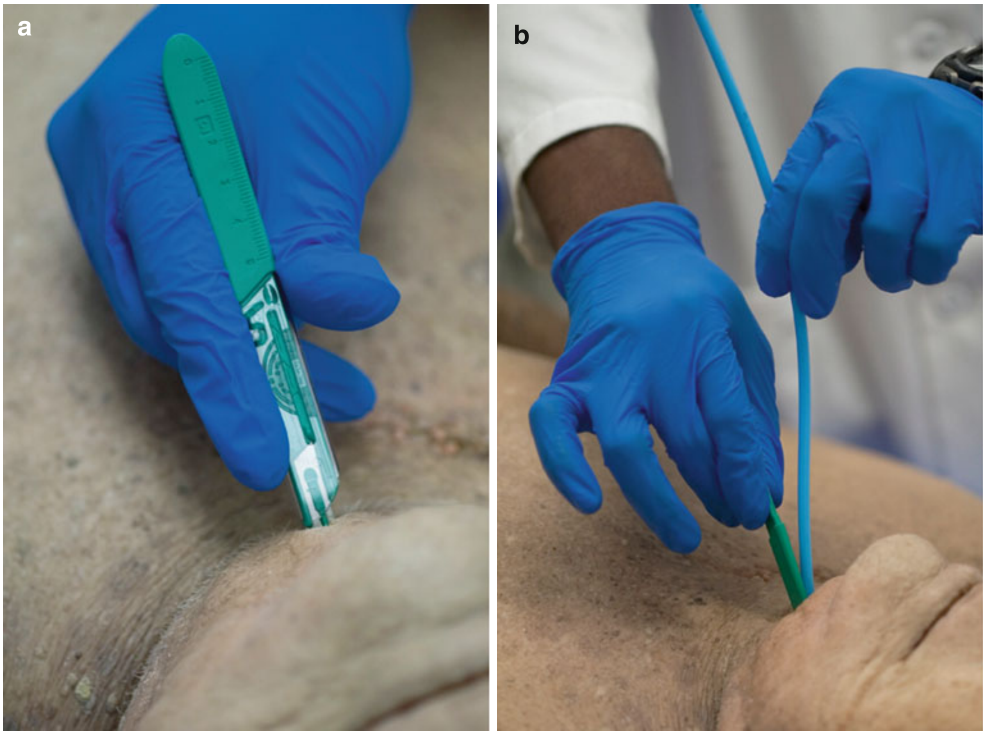

Fig. 19.3 (**a**) Make a horizontal incision into the cricothyroid membrane, (**b**) insert the bougie into the incision made in the cricothyroid membrane

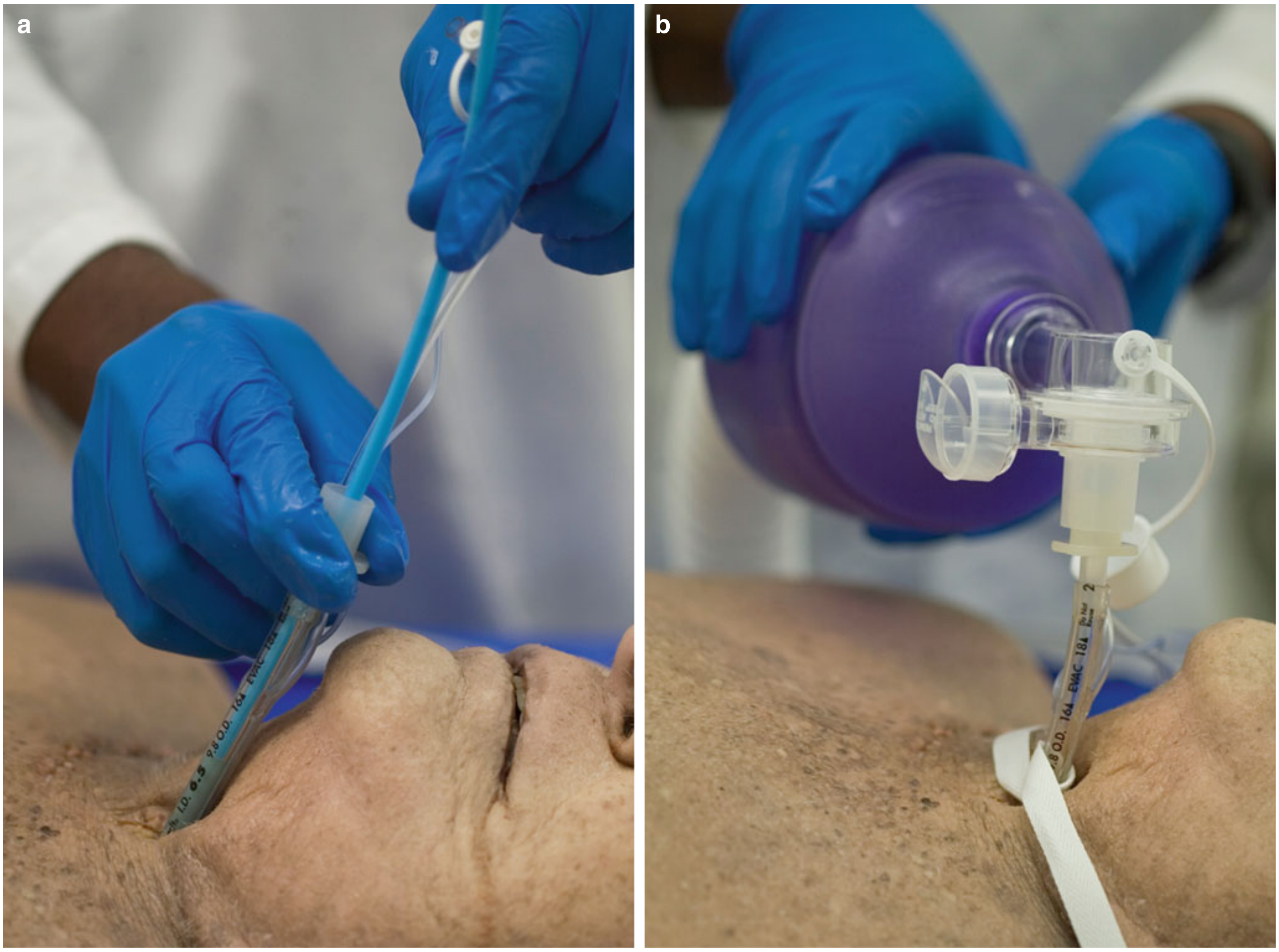

Fig. 19.4 (**a**) Slide the endotracheal tube (ETT) over the bougie into the trachea, (**b**) ventilate the patient

19.3.2 Scalpel-Trousseau

19.3.2.1 Materials and Medications (Fig. 19.5)

- Scalpel with #11 blade
- Tracheal hook
- Trousseau dilator
- Cuffed tracheostomy tube (TT) (6.5 or 7.0) or ETT (5.0, 5.5, or 6.0)
- Antiseptic preparation

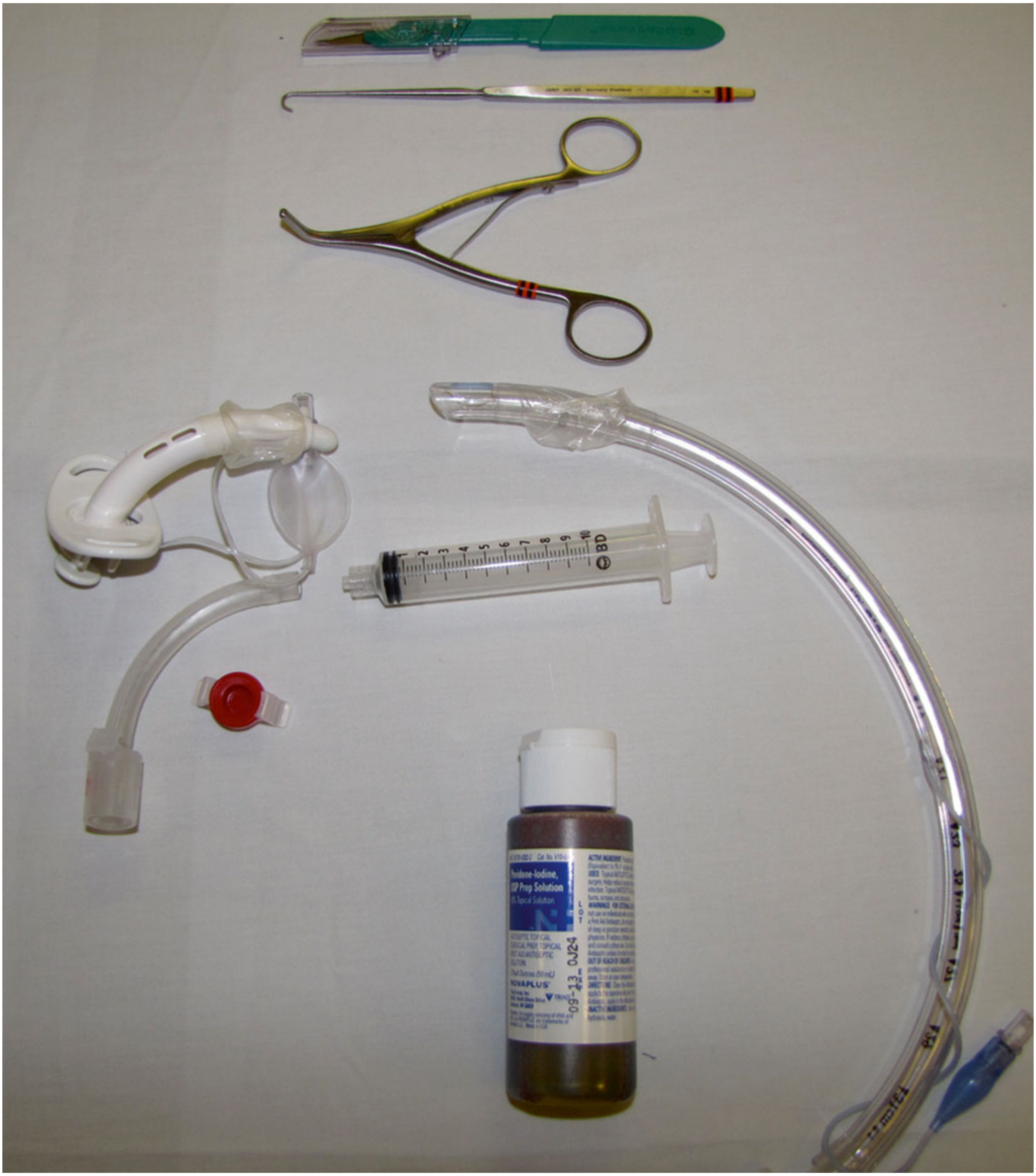
1
2
3
4
5
6
7
8
9
10
BD
Povidone-Iodine, USP Prep Solution

Fig. 19.5 *Top to bottom*, *right to left*: Scalpel with #11 blade; tracheal hook; Trousseau dilator; cuffed tracheostomy tube (TT) (6.5 or 7.0) or ETT (5.0, 5.5, or 6.0); antiseptic preparation

19.3.2.2 Procedure

1. Apply topical antiseptic.
2. Palpate the thyroid notch, cricothyroid membrane, and hyoid bone for orientation.
3. Stabilize the thyroid between the thumb and the middle finger of the nondominant hand (Fig. 19.6a).
4. Make a vertical skin incision (2–3 cm) over the cricothyroid membrane (Fig. 19.6b).
5. Palpate with the index finger to verify the cricothyroid membrane location.
6. Use stabilization of the thyroid and palpation to maintain orientation of the anatomy.
7. Make a 1.5-cm horizontal incision through the lower half of the membrane (Fig. 19.7a).
8. Insert a tracheal hook into the incision, then rotate such that hook faces superiorly (Fig. 19.7b).
9. Withdraw at a 45° angle in a cephalad direction, applying gentle traction to the thyroid cartilage.
10. Place the Trousseau dilator into the incision transversely and open the membrane incision vertically (Fig. 19.8a).
11. Insert a cuffed ETT (5.0–6.0) or TT (6.5–7.0) into the incision between the prongs of the dilator in the horizontal access (Fig. 19.8b).
12. Rotate both the dilator and the ETT toward the head of the patient and then direct the tube downward into the trachea while removing the dilator.
13. Inflate the ETT cuff and ventilate the patient.
14. Verify the position of the ETT via auscultation, $EtCO_2$, and chest x-ray.
15. Once placement of the tube has been verified, the tracheal hook can be removed.
16. Secure the ETT.

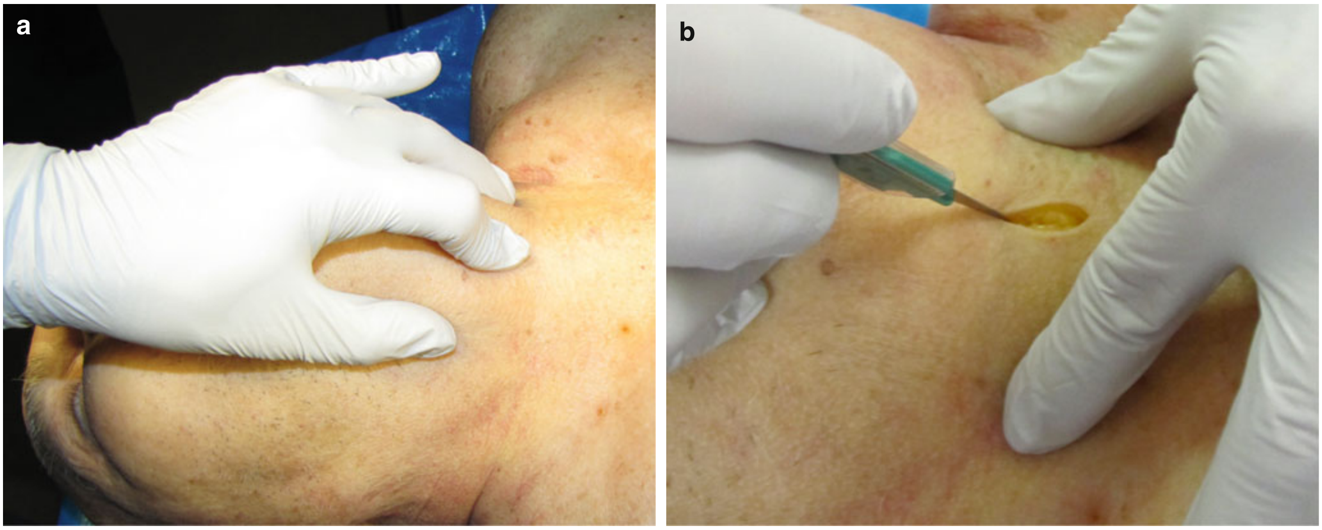

Fig. 19.6 (**a**) Stabilize the thyroid cartilage between the thumb and the middle finger of the nondominant hand; (**b**) make a vertical incision over the cricothyroid membrane

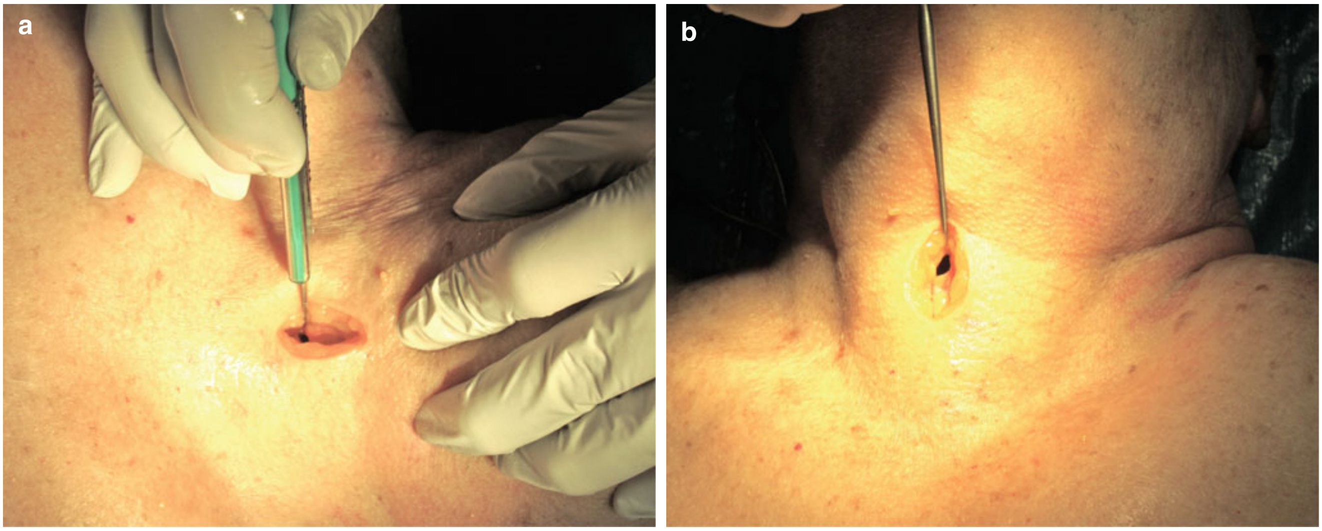

Fig. 19.7 (**a**) Make a horizontal incision in the lower half of the cricothyroid membrane. (**b**) Insert a tracheal hook and apply gentle traction superiorly at a 45° angle

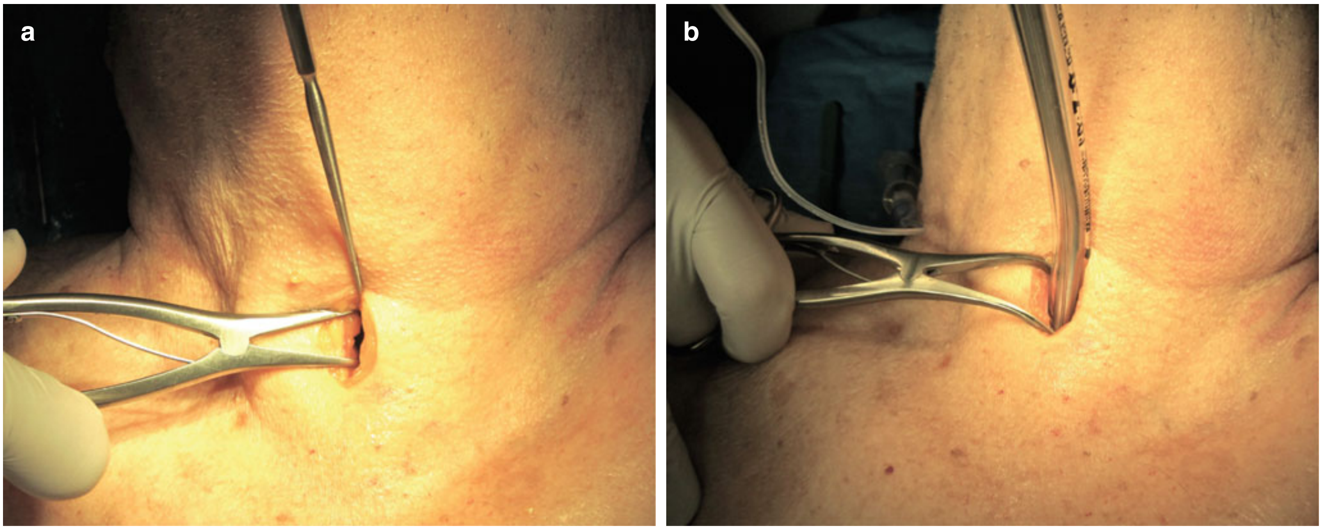

Fig. 19.8 (**a**) Insert the Trousseau dilator into the incision, and open the path for placement of the tracheostomy tube (TT) or the ETT. (**b**) Insert the TT or the ETT into the expanded incision between the Trousseau dilator prongs

19.4 Complications

- Bleeding
- ETT misplacement (false passage, through the thyrohyoid membrane, unintentional tracheostomy)
- Hoarseness, dysphonia, or vocal cord paralysis
- Subglottic or laryngeal stenosis
- Damage to thyroid cartilage, cricoid cartilage, or tracheal rings
- Perforated esophagus
- Infection
- Aspiration

19.5 Pearls

- Predictors of difficult cricothyrotomy: "SHORT" [1]
  - Prior *S*urgery or *S*car tissue
  - *H*ematoma
  - *O*bese
  - Prior *R*adiation
  - *T*umor/abscess
- The incision should cut through the skin and subcutaneous tissue down to the cricothyroid membrane and cartilages.
- Vertical incisions allow for extension in either direction if the cricothyroid membrane is above or below initial incision.
- Blood in the field may hinder visualization of the membrane, but the airway should be established before attempts to control any bleeding.
- Cricothyroid arteries are located cephalad to the cricothyroid membrane.
- Either a TT or an ETT can be used for the procedure.
  - ETTs are more ubiquitous.
  - TTs are easier to secure.
- If an ETT is used, a stylet can help direct placement.
- Cricothyroidotomy is preferred in the emergency setting over tracheostomy owing to the increased risks of bleeding, the mobility of the trachea, and the risk of lacerating the underlying thyroid gland [2].
- Ultrasound can be used to visualize landmarks in patients in whom landmarks are difficult to identify.

References

1. Walls RM, Murphy MF, editors. Manual of emergency airway management. 4th ed. Philadelphia: Wolters Kluwer; 2012.
2. Boon JM, Abrahams PH, Meiring JH, et al. Cricothyroidotomy: a clinical anatomy review. Clin Anat. 2004;17:478–86.

Selected Reading

DiGiacomo C, Neshat KK, Angus LD, et al. Emergency cricothyrotomy. Mil Med. 2003;168:541–4.

Hamilton PH, Kang JJ. Emergency airway management. Mt Sinai J Med. 1997;64:292–301.

Helm M, Gries A, Mutzbauer T. Surgical approach in difficult airway management. Best Pract Res Clin Anaesthesiol. 2005;19:623–40.

Sagarin MJ, Barton ED, Chng YM, Walls RM. Airway management by US and Canadian emergency medicine residents: a multicenter analysis of more than 6,000 endotracheal intubation attempts. Ann Emerg Med. 2005;46:328–36.

Walls RM. Cricothyroidotomy. Emerg Med Clin North Am. 1988;6:725–36.

Tracheostomy Tube Malfunction

20

Deena Bengiamin and Bharath Chakravarthy

Tracheostomy tube (TT) malfunction is the source of airway compromise in patients requiring these airway devices. TT malfunction may create an airway emergency, and the timely replacement of TTs is a challenging procedure in the most experienced hands.

20.1 Indications

- Cuff rupture
  - Can lead to dislodgment
- Dislodgment
  - Most common emergency department TT complication
  - Can lead to air passage obstruction
- Obstruction
  - Caused by blood or thick, dry secretions (formed in the absence of nasopharyngeal air humidification).
  - Dried secretions or blood can act as a one-way valve, allowing air in but restricting outward flow.

20.2 Materials and Medications

- Airway suction catheter
- Oxygen humidifier
- Bougie/nasogastric (NG) tube (12 French)
- Saline
- N-Acetyl-cysteine (NAC)
- Appropriately sized endotracheal tube (ETT) or TT

20.3 Procedure

1. Supply high-flow humidified oxygen through a bag valve mask (BVM), or a non-rebreather face mask.
2. Assess clinical indicators to determine TT problem.
   (a) Indicators of cuff rupture: Air leak with BVM and loose TT.
   (b) Indicators of dislodged TT (if TT is still in the stomal opening): Subcutaneous emphysema, crepitus, and diminished or absent breath sounds.
   (c) Indicators of an obstructed TT: ±stridor and diminished or absent breath sounds.
   (d) If time allows, chest radiograph, continuous capnography, and oxygen saturation can be helpful.
3. Obstruction
   (a) Remove the inner cannula and inspect for obstruction—clean if necessary.
   (b) If this fails, 5–10 mL of saline or NAC can be squirted directly down the TT to loosen secretions.
   (c) Suction thoroughly with a suction catheter (Fig. 20.1).
   (d) In refractory cases, the TT will need replacement.
4. Replacement
   (a) Ideally, the replacement tube should be of the same type and size as the original TT.
   (b) A smaller size TT or ETT can be helpful in settings of airway compromise.
   (c) A 6–7.5-cm ETT tube may be used if a TT is unavailable.
   (d) Remove the existing TT.
   (e) Hyperextend the patient's head and neck to maximize visualization of the stoma.

D. Bengiamin, MD
Department of Emergency Medicine, University of San Francisco Fresno, Fresno, CA, USA
e-mail: dbengiamin@fresno.ucsf.edu

B. Chakravarthy, MD, MPH (✉)
Department of Emergency Medicine, University of California Irvine, Orange, CA, USA
e-mail: bchakrav@uci.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_20

(f) Note: Careful inspection of the area is paramount because the thyroid isthmus may obscure visualization of the tracheal stoma.

(g) Techniques
  - (i) Direct insertion
    1. As soon as possible, insert the new TT or ETT into the stoma to prevent stomal narrowing (Fig. 20.2).
    2. Inflate the new TT/ETT cuff.
  - (ii) Bougie or NG tube
    1. Lubricate the TT or ETT tube.
    2. Lubricate a bougie or 12-French NG tube.
    3. Insert the lubricated bougie or NG tube into the TT or ETT.
    4. Insert the bougie or NG tube into stoma and advance into the trachea (Fig. 20.3a).
    5. Direct the NG tube or bougie caudad toward the lower tracheobronchial tree.
    6. Do not advance more than 7 cm.
    7. If resistance is noted:
      - (a) Either the operator has reached a terminal bronchiole or is in a false passage.
      - (b) Do not force bougie/NG tube farther at this point.
      - (c) Use clinical judgment (bougie/ETT depth, palpation) to determine likely placement.
    8. Advance TT or NG tube over bougie or NG tube is in trachea (Fig. 20.3b).
    9. After the TT is in place, remove the bougie/NG tube.
  - (iii) The fingertip technique
    1. Insert a gloved forefinger into tracheal stoma (Fig. 20.4a).
    2. Formulate a mental plan as to the direction and path of the stoma.
    3. Then place a TT or an ETT into the stoma as the finger is withdrawn (Fig. 20.4b).

5. If placement of a TT or ETT is not possible through the stoma, consider endotracheal intubation.

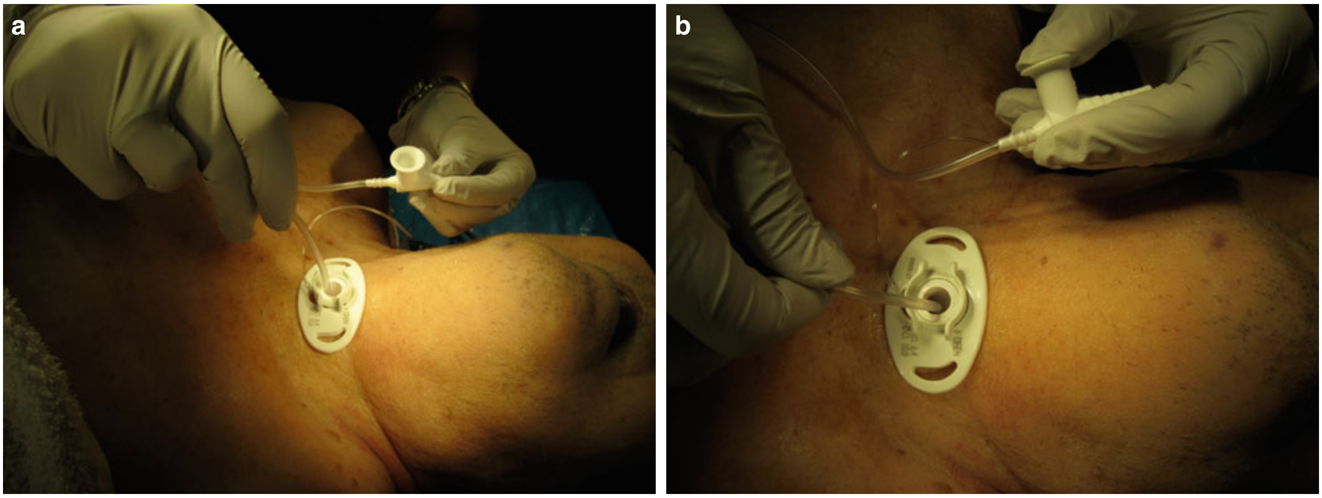

Fig. 20.1 (**a**) Insert the suction catheter to the appropriate depth, keeping the suction port open. (**b**) Slowly withdraw the catheter in a circular motion, keeping the suction port closed

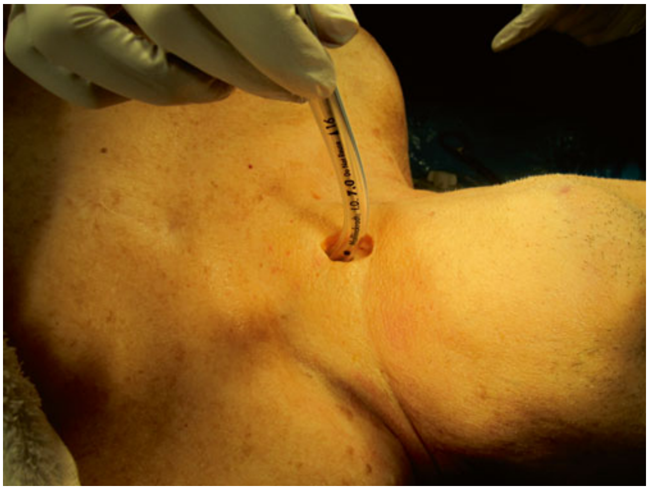

Fig. 20.2 Insertion of the endotracheal tube (ETT) directly into the stoma

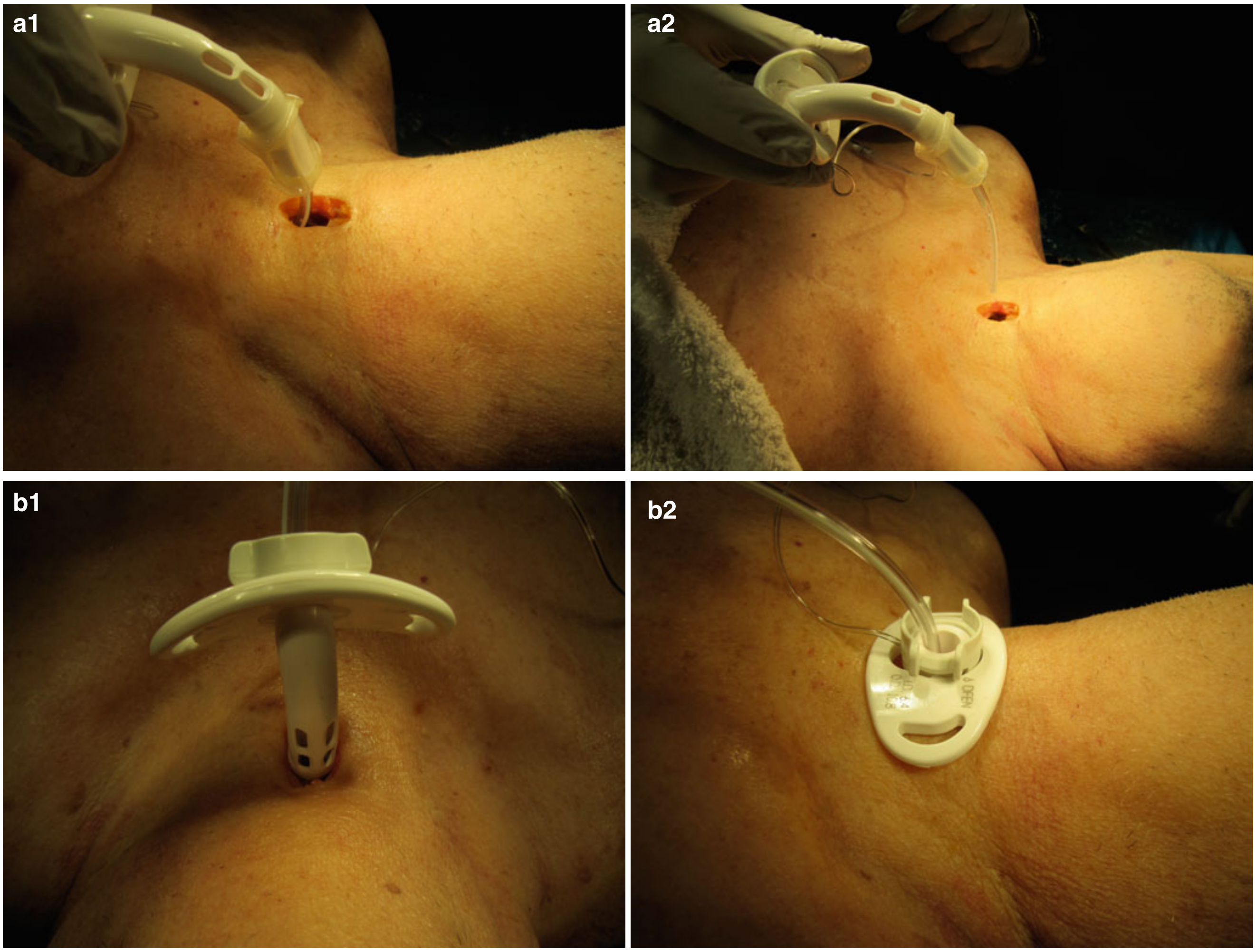

Fig. 20.3 (**a1, a2**) Insert the tracheostomy tube (TT) over the nasogastric (NG) tube or bougie, which is first inserted into the trachea. (**b1, b2**) Advance the TT over the NG tube or bougie, which serves as a guidewire

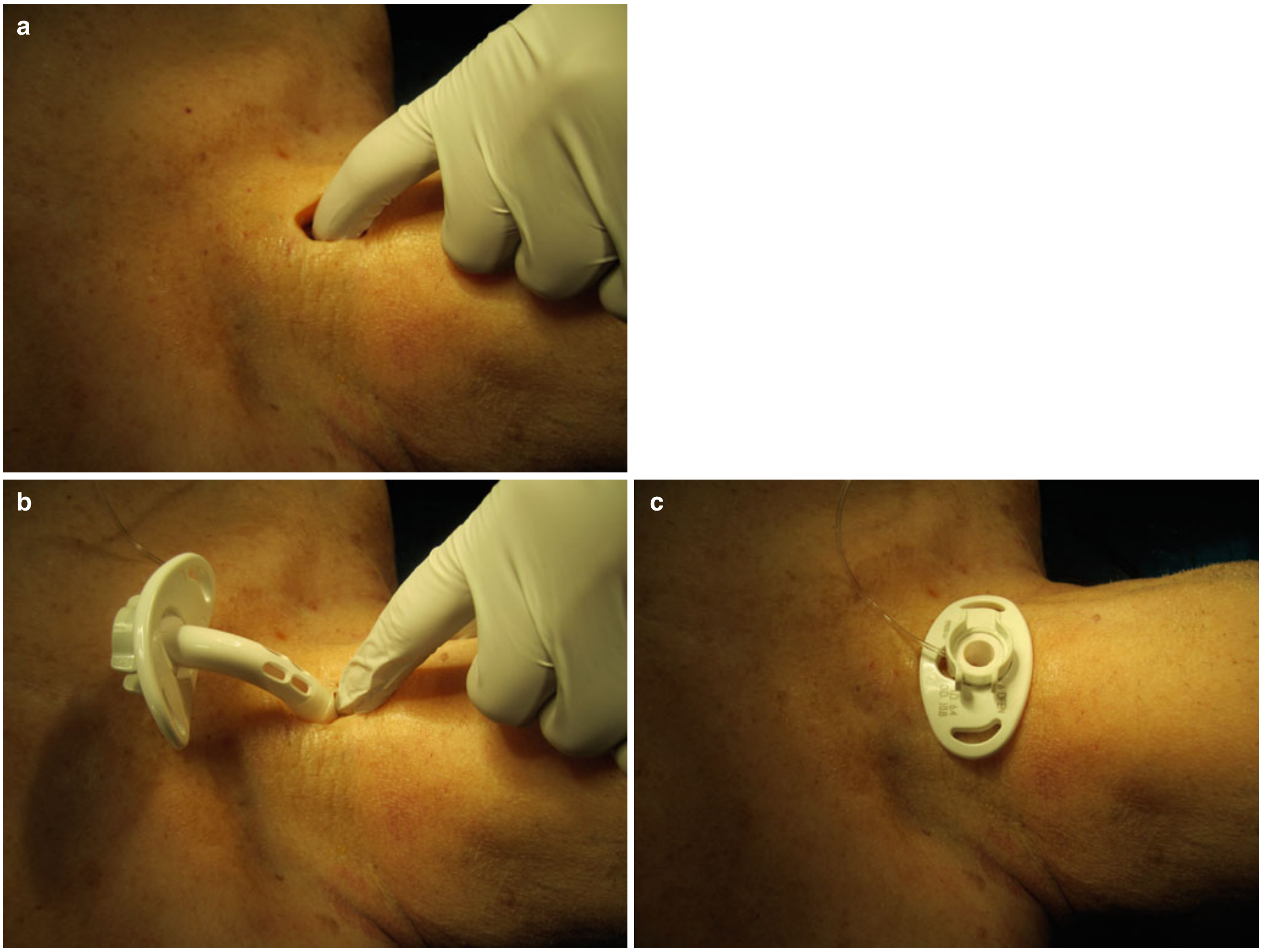

Fig. 20.4 (**a**) Insert a gloved finger into the stoma. (**b**) Insert the TT along the path while withdrawing the gloved finger. (**c**) TT in place

20.4 Pearls and Pitfalls

- Stomal closing
  - – Stomal constriction begins as soon as the TT is removed or displaced.
  - – Forceful attempts to replace a large TT may result in false passages and trauma.
- Tracheal stenosis
  - – Constant TT cuff pressure may cause necrosis, ulceration, and granulation tissue formation, leading to tracheal narrowing.
  - – Complicates TT replacement in the setting of dislodgment or obstruction.
  - – Bougie use can be helpful to place a small(er) ETT until surgical dilation and/or resection can be achieved.
- Creation of a false lumen during recannulation
  - – Subcutaneous emphysema will be an early indicator of this.
  - – Early confirmation of correct placement of TT by confirming:
    - Continuous end-tidal CO_2 monitoring
    - Equal chest rise
    - Bilateral breath sounds
- Unrecognized trachea–innominate artery fistula (Fig. 20.5)
  - – Usually occurs within 3–4 weeks of placement
  - – Presentation: Bleeding around the tracheostomy tube (>10 mL) or massive hemoptysis
  - – Requires
    - ETT cuff overinflation to compress the fistula.
    - Digital pressure on stoma may be helpful to tamponade the bleeding.
    - Place stomal ETT deep to bleeding fistula to protect airway.
    - Definitive surgical intervention in operating room.
  - – Associated with high mortality
- Unrecognized tracheoesophageal fistula
  - – Usually iatrogenic injury from TT placement or NG tube erosion
  - – Presentation: Dyspnea, copious TT secretions, recurrent food aspiration, and gastric distention
  - – Requires
    - Bronchoscopy or swallowing studies to confirm diagnosis
    - Surgical repair or stenting

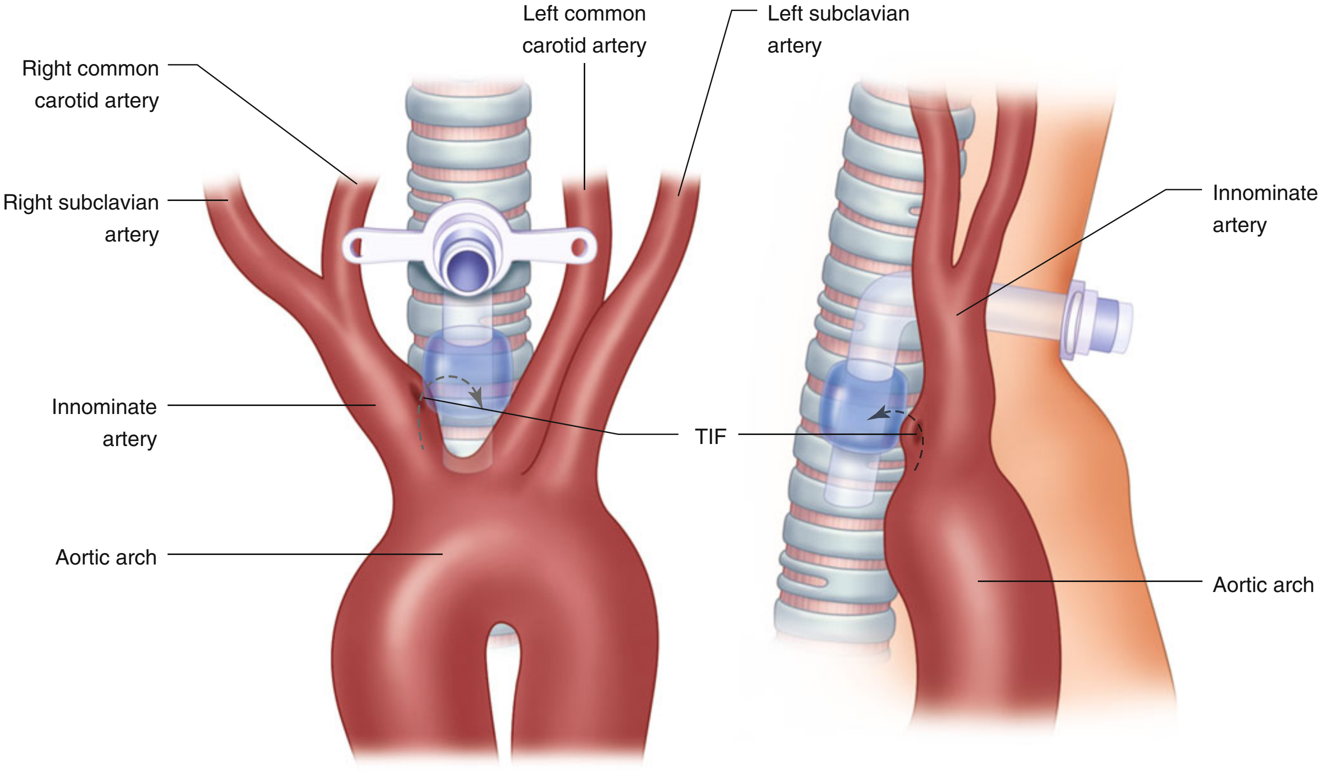

Fig. 20.5 Anatomical relationship between the trachea and the innominate artery. *TIF* trachea–innominate artery fistula

Selected Reading

De Leyn P, Bedert L, Delcroix M, et al. Tracheotomy: clinical review and guidelines. Eur J Cardiothorac Surg. 2007;32:412–21.

Dobiesx VA, Miller SA, Pitzele MJ. Complications of tracheostomies. In: Wolfson AB, Hendey GW, Ling LJ, Rosen CL, Scheider JJ, Sharieff GQ, editors. Harwood-Nuss' clinical practice of emergency medicine. 5th ed. Philadelphia: Lippincott Williams & Wilkins; 2009.

Epstein SK. Late complications of tracheostomy. Respir Care. 2005;40:542–9.

Friedman M, Ibrahim H. The dislodged tracheostomy tube: "fingertip" technique. Oper Technol Otolaryngol. 2002;13:217–8.

Young JS, Brady WJ, Kesser B, Mullins D. A novel method for replacement of the dislodged tracheostomy tube: the nasogastric tube "guidewire" technique. J Emerg Med. 1996;14:205–8.

Percutaneous Transtracheal Jet Ventilation

21

Clint Masterson

21.1 Indications

- Failure to control the airway by other means
- As a temporary measure while preparing for definitive airway control
- Securing the airway in crash airways in infants and small children

21.2 Contraindications

- Absolute
  - Transection of the trachea below the cricothyroid membrane
- Relative
  - Inability to identify the cricothyroid landmarks
  - Anatomical distortion to the cricothyroid membrane
  - Supraglottic obstruction (preventing gas exhalation)

21.3 Materials and Medications (Fig. 21.1)

- Betadine, chlorhexidine, or similar skin sterilization solution
- 12- to 16-gauge angiocatheter or transtracheal jet ventilation (TTJV) purpose-specific catheter
- 10-mL syringe filled with 4 mL of normal saline, 2 % lidocaine, or viscous lidocaine
- Hand-operated regulator valve
- Attach oxygen supply.
  - Connect kit tubing to wall oxygen OR.
  - Connect 7–0 endotracheal connector to bag valve mask (BVM) attached to oxygen.

C. Masterson, MD
Department of Emergency Medicine, Mayo Clinic Health System
in Fairmont, Fairmont, MN, USA
e-mail: clint@clintm.us

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_21

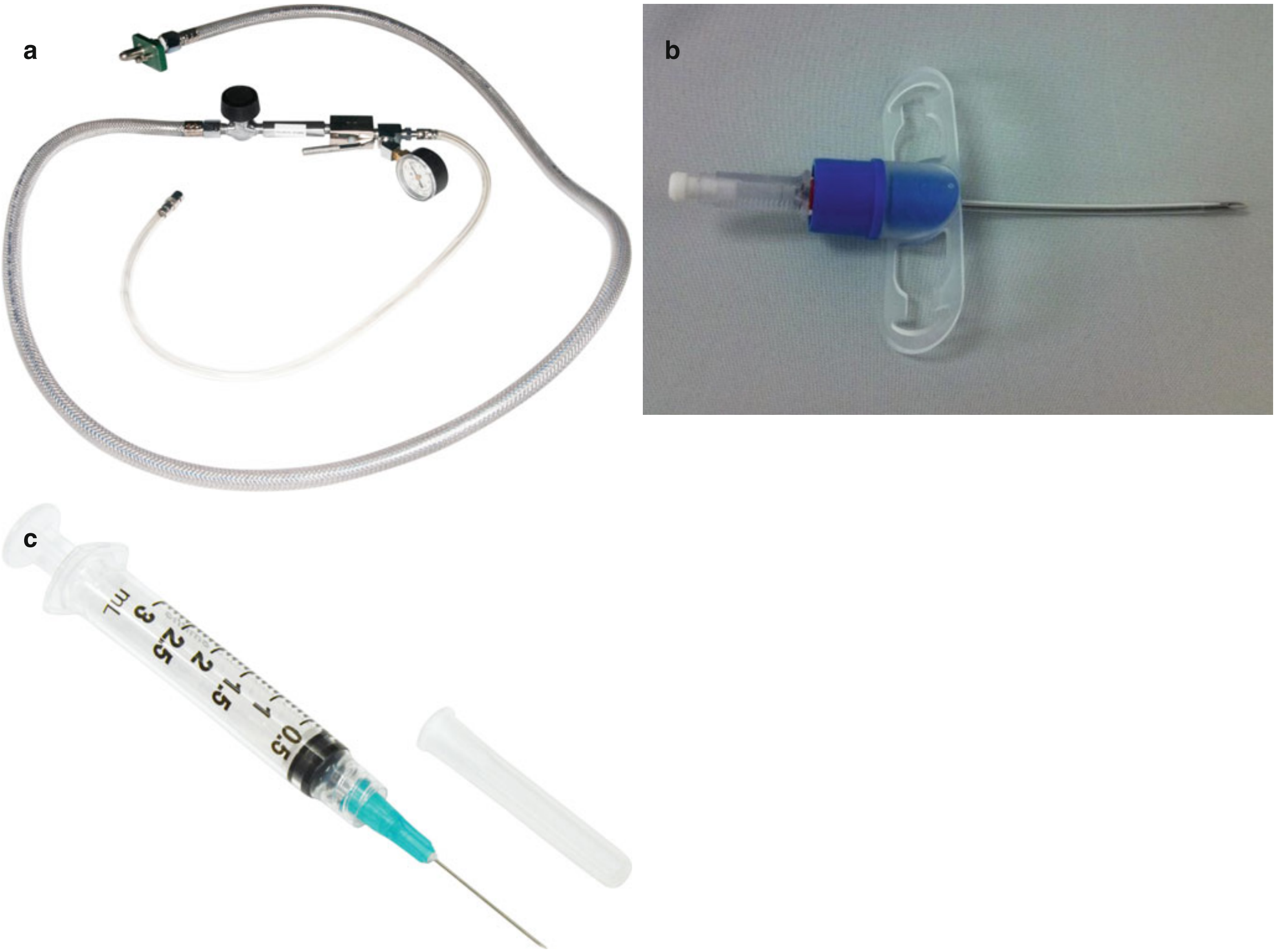
a
b
c
mL
3
2.5
2
1.5
1
0.5

Fig. 21.1 (**a**) Tubing and regulator valve, (**b**) commercially available catheter, (**c**) 3-mL or 10-mL syringe

21.4 Procedure

1. Attach the tubing and the hand-operated regulator valve to wall oxygen (Fig. 21.2a), and place the distal end of the tubing near the patient in preparation for ventilation.
2. Adjust regulator to maximum pressure, 50 psi if possible (Fig. 21.2b).
3. Palpate the cricothyroid membrane just distal to the thyroid prominence (Fig. 21.3).
   (a) Sterilize the area with a suitable cleansing agent.
   (b) Use the thumb and index finger of the nondominant hand to stabilize the trachea for the procedure.
4. Attach the TTJV catheter (or angiocatheter) to the syringe (Fig. 21.4).
5. Advance the catheter through the cricothyroid membrane at a 30–45° caudal direction while aspirating with the syringe (Fig. 21.5).
6. Return of air confirms entry into the trachea.
7. If lidocaine is utilized, it can then be injected to prevent spasm during the procedure.
8. Fully advance the angiocatheter and secure it while the needle and syringe are withdrawn.
9. Remove the needle, secure it to the skin, and connect it to the regulator hose.
10. Secure the distal end of the oxygen tubing (distal to the hand-operated valve) to the catheter (Fig. 21.6).
11. If a BVM is used as the oxygen source:
    (a) Attach a 3-mL syringe to the angiocatheter.
    (b) Attach the BVM with the 7–0 endotracheal tube (ETT) connector to the end of the plungerless 3-mL syringe (Fig. 21.7).
12. Operate the valve 12–20 times a minute with long periods to allow gas exhalation and exchange (Fig. 21.8).
13. Preparations should be made for a definitive airway as soon as possible—preferably within 15 min.

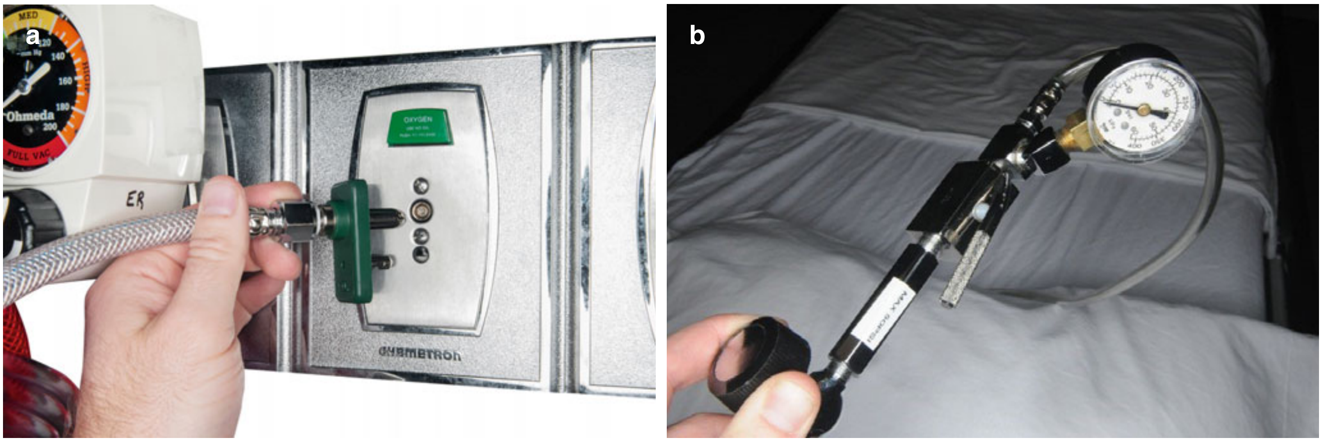

Fig. 21.2 (**a**) Attach tubing and the hand-operated regulator valve to wall oxygen, and (**b**) adjust regulator to maximum pressure (50 psi if possible)

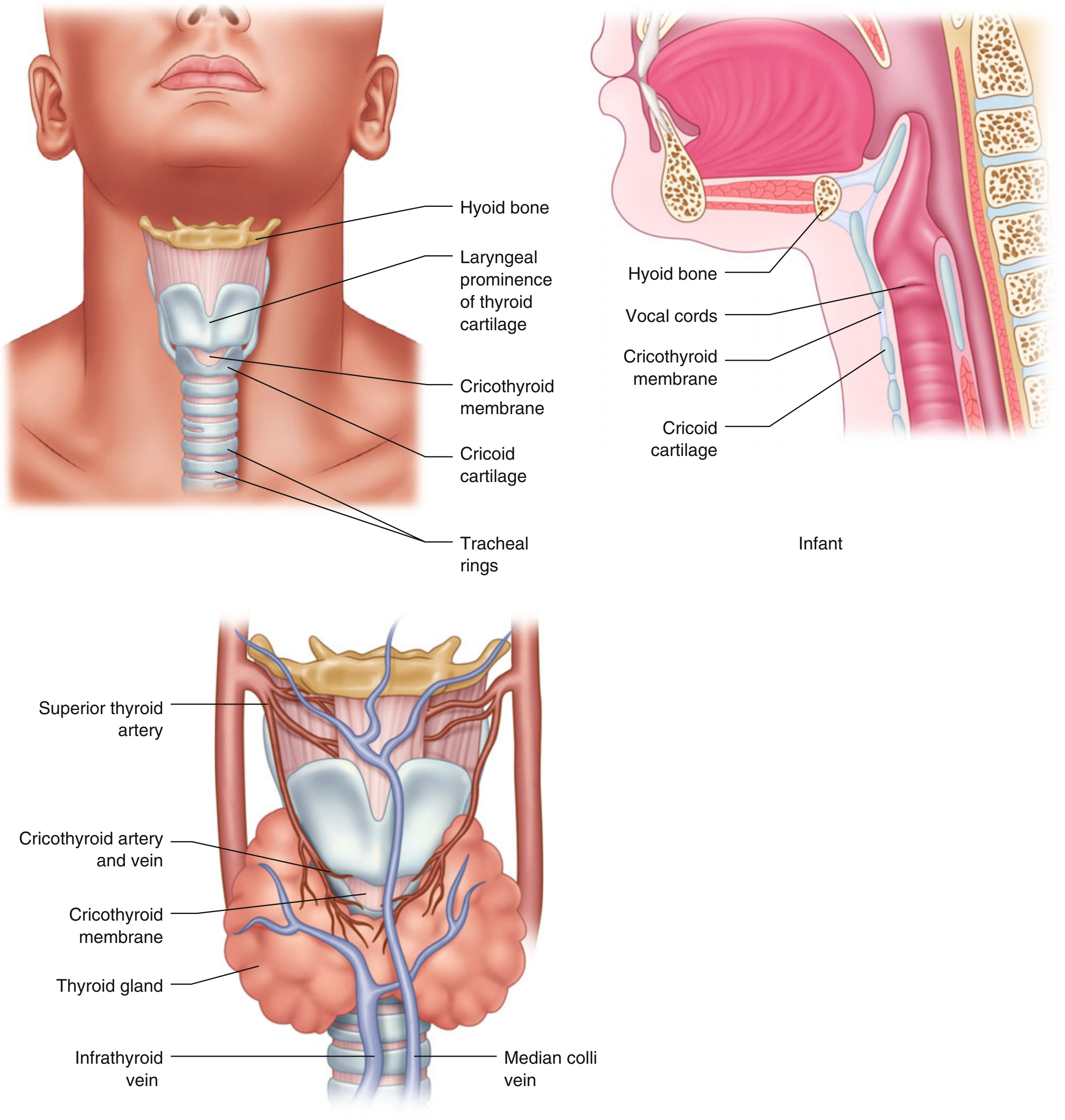
Hyoid bone
Laryngeal prominence of thyroid cartilage
Cricothyroid membrane
Cricoid cartilage
Tracheal rings
Hyoid bone
Vocal cords
Cricothyroid membrane
Cricoid cartilage
Infant
Superior thyroid artery
Cricothyroid artery and vein
Cricothyroid membrane
Thyroid gland
Infrathyroid vein
Median colli vein

Fig. 21.3 Airway anatomy

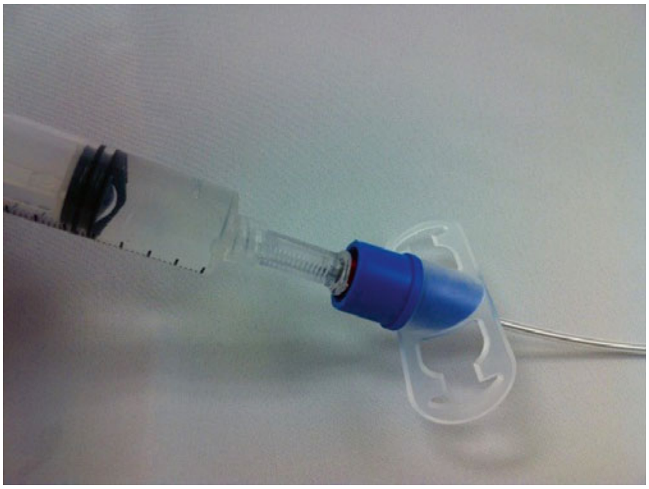

Fig. 21.4 Attach the TTJV catheter to the syringe

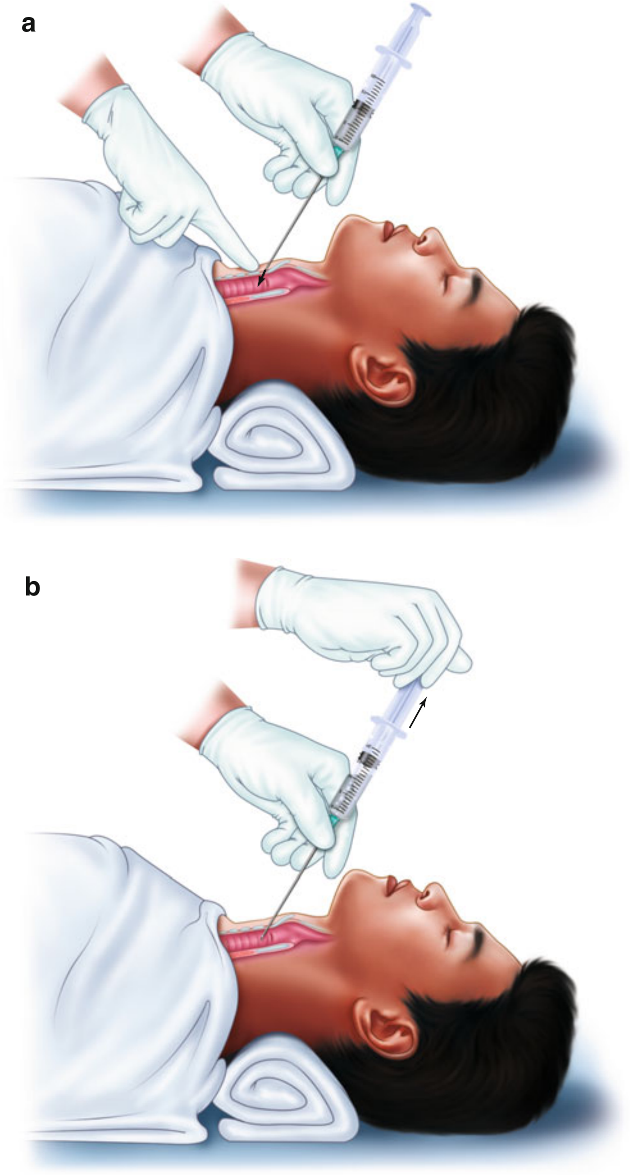

Fig. 21.5 (**a**) Advance the catheter through the cricothyroid membrane at a 30–45° caudal direction (**b**) while aspirating with the syringe

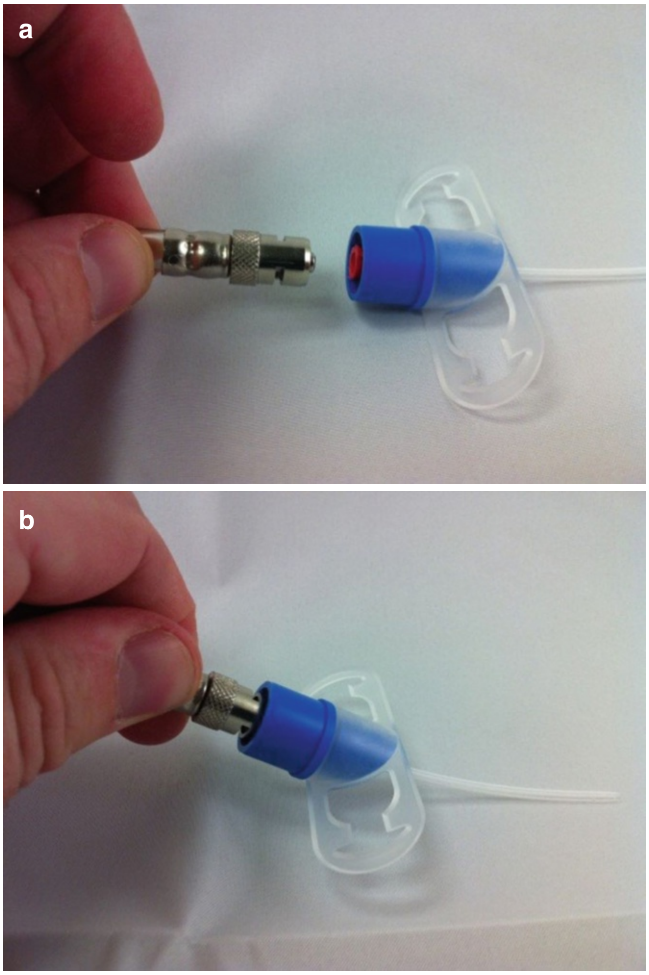

Fig. 21.6 Secure the distal end of the oxygen tubing to the catheter

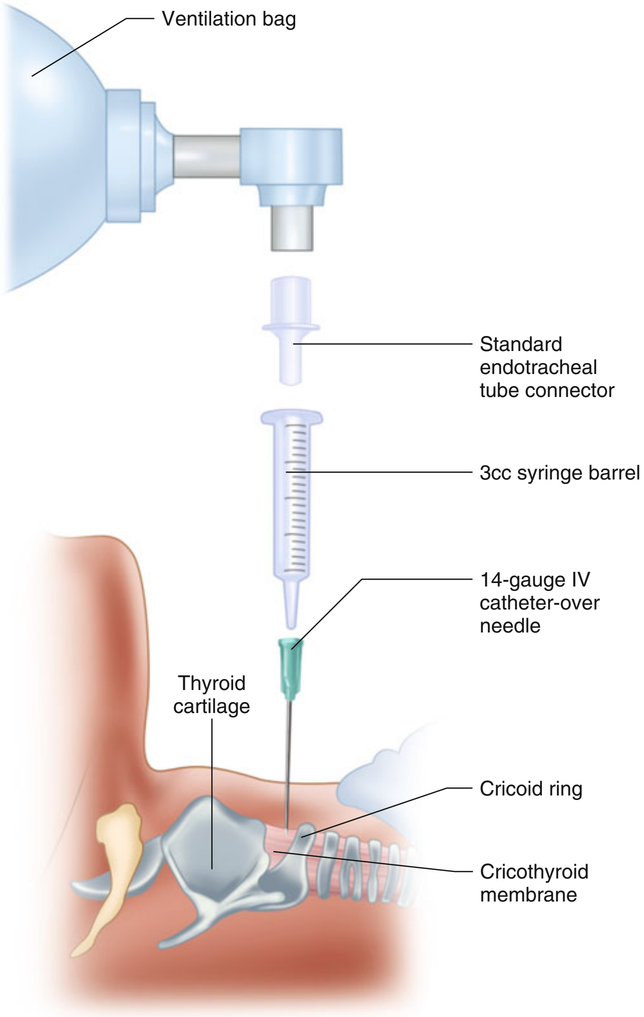
Ventilation bag
Standard endotracheal tube connector
3cc syringe barrel
14-gauge IV catheter-over needle
Thyroid cartilage
Cricoid ring
Cricothyroid membrane

Fig. 21.7 Attach the BVM with the 7–0 endotracheal tube (ETT) connector to the end of the plungerless 3-mL syringe

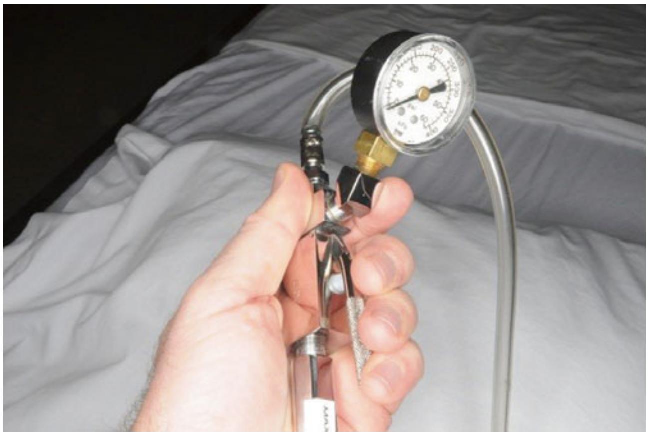

Fig. 21.8 Operate the valve 12–20 times a minute with long periods to allow gas exhalation and exchange

21.5 Complications

- Pneumothorax
- Pneumomediastinum
- Subcutaneous emphysema
- Catheter kink or misplacement
- Hypercarbia and respiratory acidosis
  - Use of TTJV for prolonged periods of time without adequate ventilation will elevate CO_2.
- Barotrauma
- Coughing in conscious patients
- Aspiration
- Persistent stoma

21.6 Pearls and Pitfalls

- If the wall connector does not have a pressure regulator, it can still be used although the risk of barotrauma is greater. Use the endpoint of chest rise to determine the end of each ventilation burst in this case.
- Higher pressures and lack of supraglottic air exchange are risk factors for pneumothorax. If the supraglottic area is obstructed, a Y catheter can be attached to allow gas to escape before the next insufflation.
- TTJV may or may not allow sufficient gas exchange to prevent hypercarbia. Preparations should be made to obtain a definitive airway as soon as possible.
- Endotracheal intubation may be facilitated by the high pressures insufflated in the trachea, and a repeat attempt may be performed after the transtracheal ventilation is achieved.

Selected Reading

Patel R. Percutaneous transtracheal jet ventilation. A safe, quick and temporary way to provide oxygenation and ventilation when conventional methods are unsuccessful. Chest. 1999;116:1689–94.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. Philadelphia: Saunders Elsevier; 2010.

Tintinalli J. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2010.

Walls R. Manual of emergency airway management. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Part III

Thoracic Procedures

22 Needle Thoracostomy

Lucas McArthur and Christian Fromm

22.1 Indications

Needle decompression thoracostomy is a procedure used in the emergent treatment of a tension pneumothorax. Tension pneumothorax is a clinical diagnosis. Decompression treatment should not be delayed in order to obtain radiographic confirmation. The following scenarios illustrate some of the clinical signs that *may* be present in such patients:

- *Awake patient* with suspected or confirmed tension pneumothorax
  - Chest pain
  - Respiratory distress
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema
  - Trachea deviated away from the side of the pneumothorax
  - Tachycardia
  - Falling pulse oximetry (SpO_2)
  - Shock
- *Ventilated patient* with suspected or confirmed pneumothorax (often insidious)
  - Increased resistance to ventilation
  - Hypotension
  - Elevated central venous pressure
  - Tachycardia
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema
  - Trachea deviated away from the side of the pneumothorax
  - Falling SpO_2
  - Shock
- *Injured patient* (especially with penetrating chest trauma) with suspected or confirmed tension pneumothorax
  - In arrest
  - Unexplained hypotension
  - Apnea
  - Decreased breath sounds with hyperresonance and/or subcutaneous emphysema

22.1.1 Absolute Indications

- Patient in acute respiratory distress with rapid decompensation secondary to suspected or confirmed tension pneumothorax
- Injured patient in extremis with apnea, unexplained hypotension, or arrest

22.2 Contraindications

- No absolute contraindications.

22.3 Materials

- Large-bore needle/angiocatheter (minimum of 16 gauge)
- 10-mL syringe (optional)
- One-way valve (optional)
- Betadine (povidone-iodine) swab/chlorhexidine scrub
- Tape

L. McArthur, MD (✉)
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: LMcarthur@jfkhealth.org

C. Fromm, MD
Department of Emergency Medicine, Maimonides Medical Center,
SUNY Downstate College of Medicine, New York, NY, USA
e-mail: cfromm@maimonidesmed.org

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_22

22.4 Procedure

1. Expose the anterior chest at the level of the second intercostal space on the affected side (Fig. 22.1). Alternatively, expose the chest wall at the level of the anterior axillary line in the fourth or fifth intercostal space on the affected side.
2. Cleanse the area with a Betadine swab or chlorhexidine scrub (Fig. 22.2).
3. Using a gloved hand, locate the second intercostal space at the midclavicular line.
    (a) The first rib is normally not felt.
    (b) The second rib is felt just below the clavicle.
    (c) The second intercostal space is the area between the second and the third ribs.

    Note: Alternatively, this procedure may also be performed on the midaxillary line in the fourth intercostal space of the affected side. The same general steps listed later are employed in this approach and care is taken to avoid the neurovascular bundles inferior to the fourth rib.
4. Insert the needle/angiocatheter perpendicular to the chest wall into the second intercostal space just above the superior edge of the third rib to avoid the intercostal neurovascular bundle (Fig. 22.3).
    (a) This step may be done with or without a syringe attached.
    (b) Local anesthesia is usually unnecessary but may be used if the patient is not in extremis.
5. Carefully walk the needle over the third rib and advance until the pleural space is entered.
    (a) Entry into the pleural space is accompanied by a "popping" sound or a sensation of "giving way."
6. If you are able to withdraw air with the syringe or hear a "hiss" of air escaping through the angiocatheter during expiration and inspiration, then placement is considered successful.
7. After removing the needle, secure the angiocatheter in place with tape (Fig. 22.4).

    Caution: Do not reinsert needle into the angiocatheter owing to the danger of sheering the angiocatheter.
8. Assess the patient and evaluate the effectiveness of the procedure.
    (a) The patient should exhibit immediate and obvious improvement in respiratory status including improved lung sounds and vital signs.
    (b) The procedure may be repeated if the patient is not improving.
    (c) Excess pleural air may be aspirated through the angiocatheter with a syringe.
9. Obtain a chest radiograph to confirm success.
    (a) Repeat in 6 h.
10. Because needle decompression is only a temporizing measure, tube thoracostomy (see Chap. 23) must be performed for definitive management of the pneumothorax.

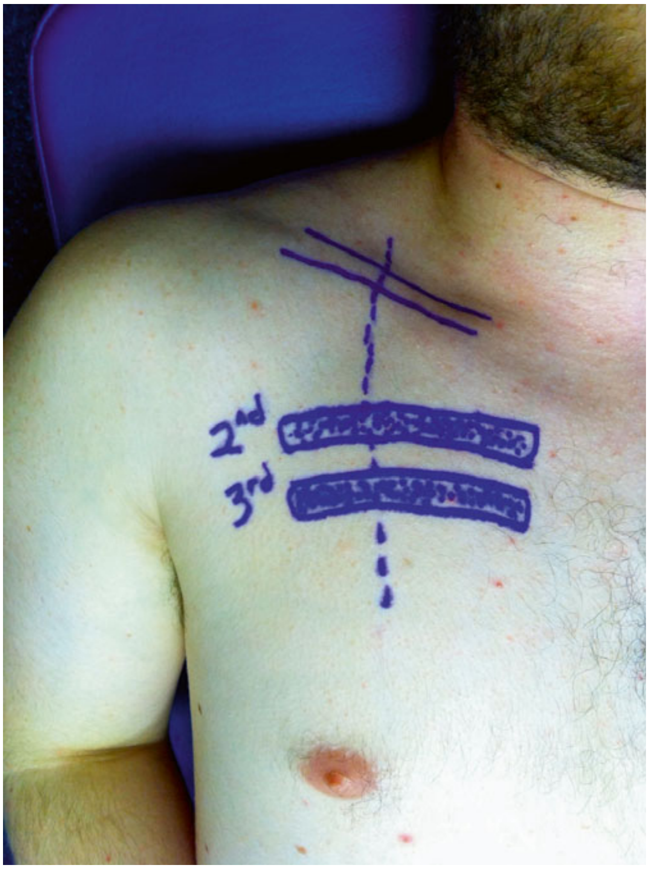

Fig. 22.1 The preferred site for needle thoracostomy is the second intercostal space in the midclavicular line

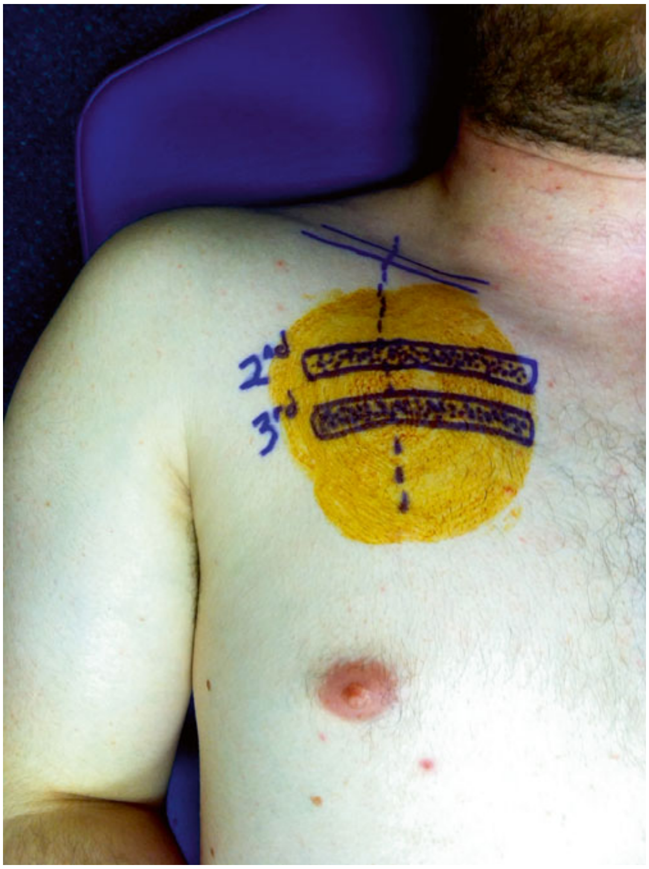

Fig. 22.2 Prepare the skin with povidone-iodine or chlorhexidine

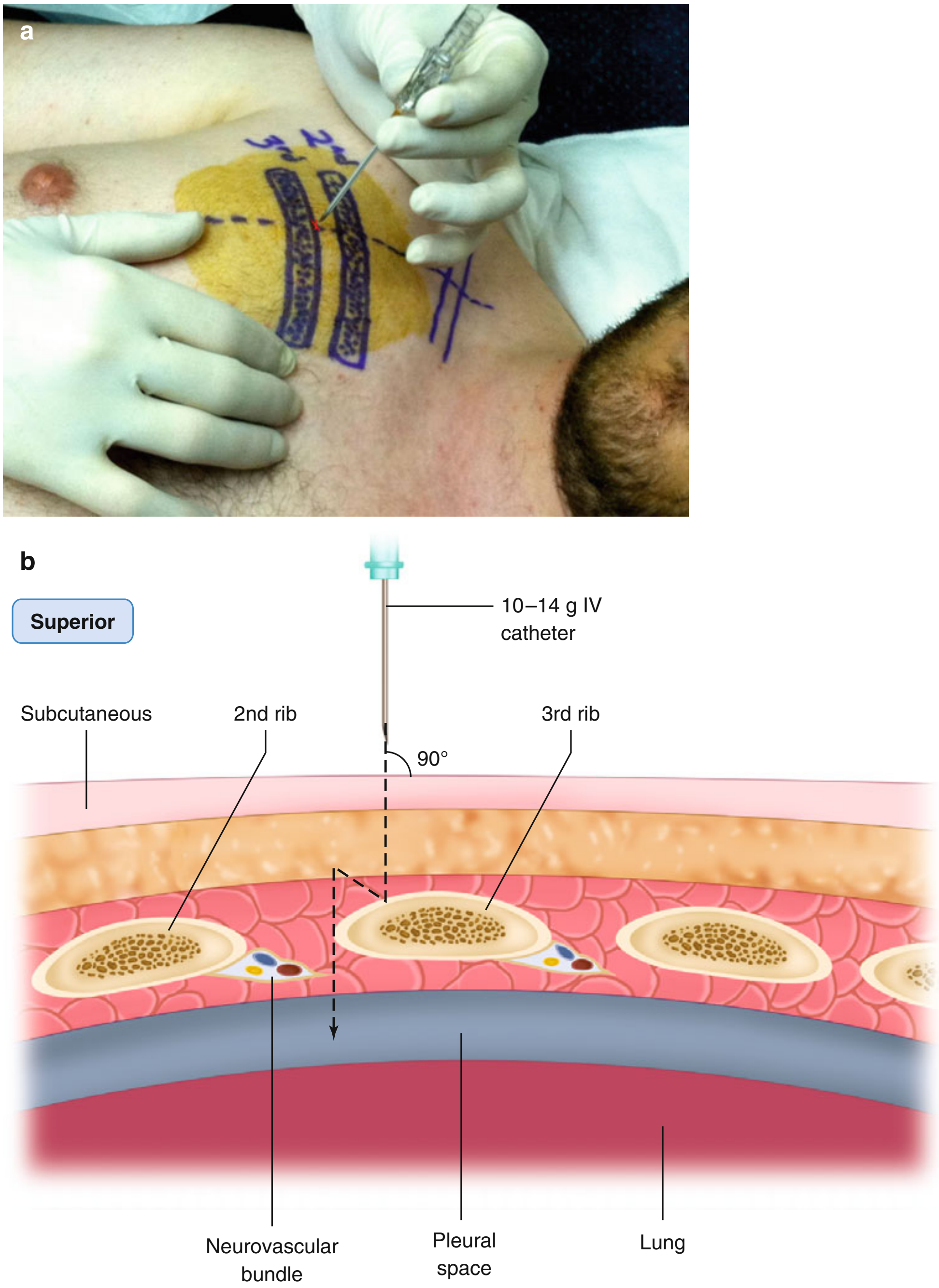
a
2nd
3rd
b
Superior
10–14 g IV catheter
Subcutaneous
2nd rib
3rd rib
90°
Neurovascular bundle
Pleural space
Lung

Fig. 22.3 (**a**) Insert the needle into the second intercostal space just above the superior edge of the third rib and (**b**) avoid the neurovascular bundle by approaching the skin with the needle perpendicular to the chest wall just above the superior edge of the third rib

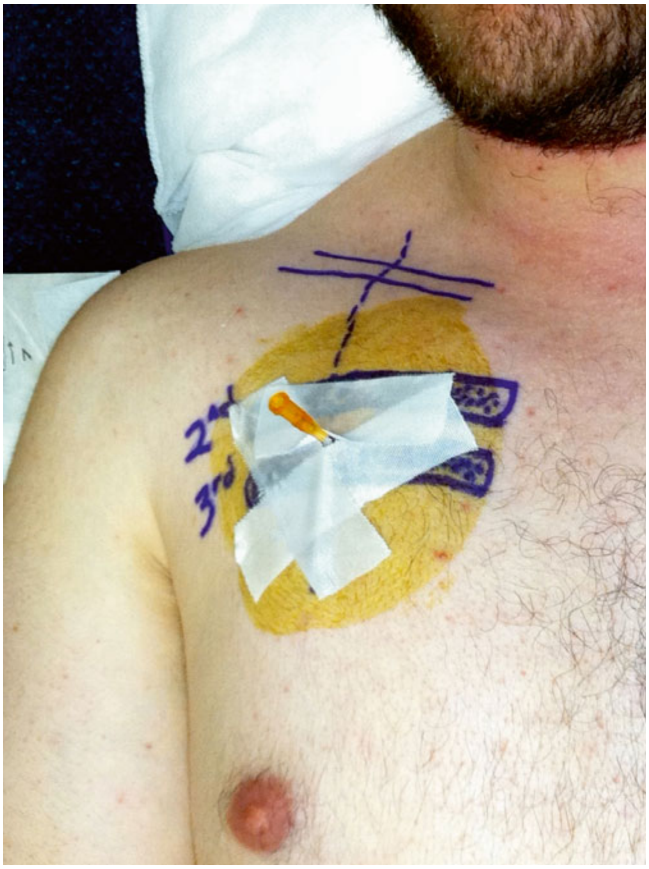

Fig. 22.4 After removing the needle, secure the angiocatheter in place with tape

22.5 Complications

- Failure to resolve the tension pneumothorax.
  - Obese or muscular patients may require a longer needle and catheter to reach the pleural space or, alternatively, may require proceeding immediately to tube thoracostomy.
- Iatrogenic pneumothorax.
- Laceration of intercostal artery or nerve.
- Rapid re-expansion may result in the development of pulmonary edema.
- Infection.

22.6 Pearls and Pitfalls

- Use the sternum as a landmark to more easily locate the second and third ribs.
- The same procedure may also be done on the midaxillary line in the fourth intercostal space, which is usually landmarked with the nipple.
- Primary pneumothorax is unusual in those older than 40 years. Consider the presence of underlying disease in this population.

Acknowledgments We would like to thank Antonios Likourezos, MA, MPH, and Abraham Lederman for assisting with the photographs.

Selected Reading

Britten S, Palmer SH. Chest wall thickness may limit adequate drainage of tension pneumothorax by needle thoracocentesis. Emerg Med J. 1996;13:426–7.

Custalow CB. Color atlas of emergency department procedures. Philadelphia: Saunders; 2005.

Leigh-Smith S, Harris T. Tension pneumothorax—time for a re-think? Emerg Med J. 2005;22:8–16.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 3rd ed. Philadelphia: Saunders; 1998.

Chest Tube Thoracostomy

23

Brandon R. Allen and Latha Ganti

23.1 Indications

- Spontaneous pneumothorax (large and/or symptomatic)
- Tension pneumothorax (or suspected)
- Iatrogenic pneumothorax
- Penetrating chest injuries
- Hemopneumothorax in acute trauma
- Patient in extremis with evidence of thoracic trauma
- Complicated parapneumonic effusions (empyema)
- Chylothorax/hemothorax
- Post-thoracic surgery
- Bronchopleural fistula

23.2 Contraindications

- Absolute
  - Emergent thoracotomy
- Relative
  - Coagulopathy
  - Pulmonary bullae
  - Pulmonary, pleural, or thoracic adhesions
  - Loculated pleural effusion or empyema
  - Skin infection over the chest tube insertion site

23.3 Materials and Medications

- Tube thoracostomy tray
  - #10 scalpel; 18-, 22-, and 25-gauge needles; 10-mL syringes; forceps; clamps; scissors; drape; abdominal pads; 0 or 1–0 silk suture; needle driver; curved clamp (Fig. 23.1a)
- Betadine (povidone-iodine) or other skin antiseptic preparation solution
- Lidocaine (1 % or 2 % with epinephrine)
- Appropriate chest tube size (approximate)
  - Adult male: 28–36 French
  - Adult female: 28 French
  - Child: 12–24 French
  - Infant: 12–16 French
  - Neonate: 10–12 French

Vaseline gauze
Chest drainage system (Fig. 23.1b)

B.R. Allen, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: brandonrallen@ufl.edu

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_23

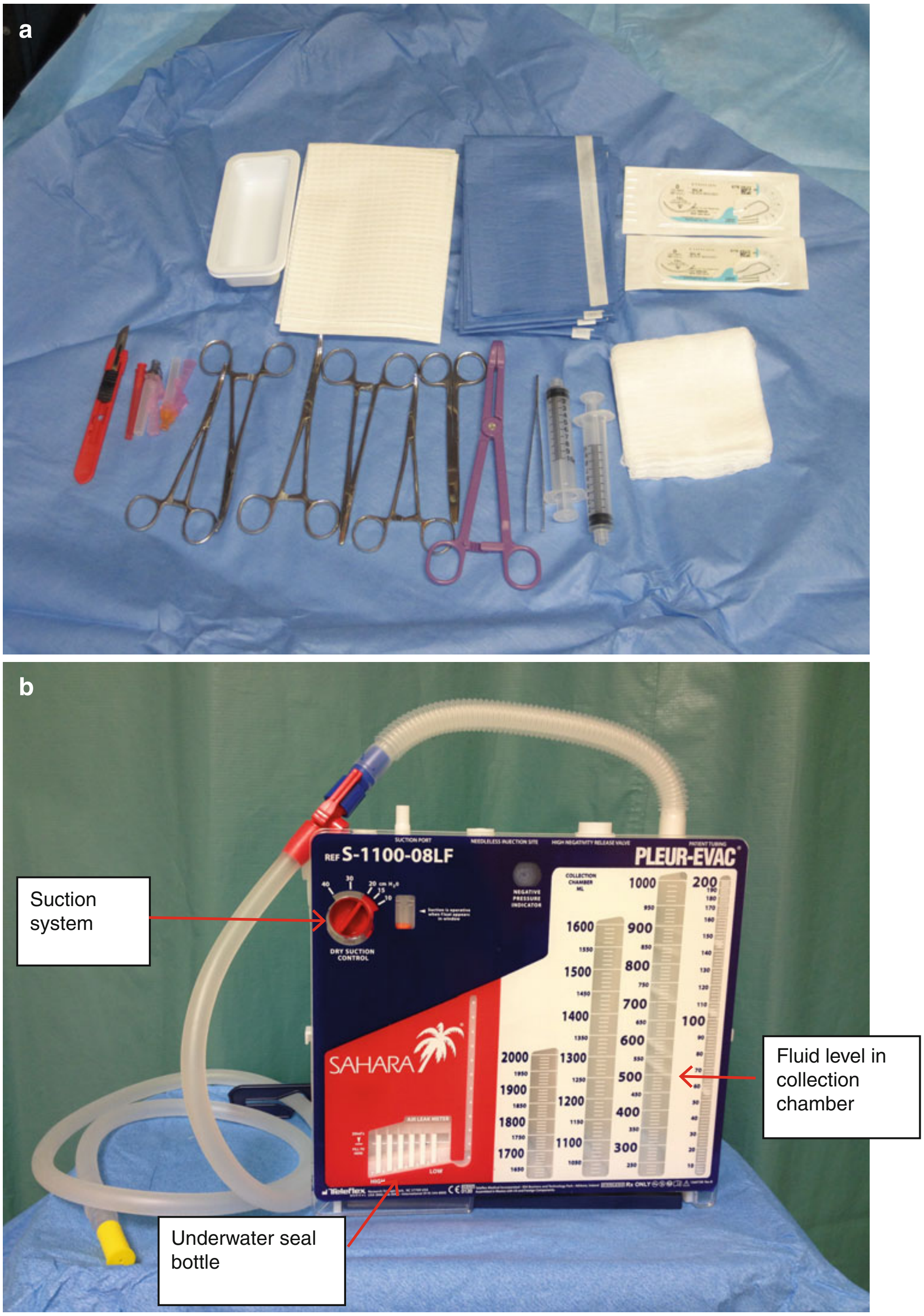
a
b
Suction system
Fluid level in collection chamber
Underwater seal bottle
REF S-1100-08LF
PLEUR-EVAC
SAHARA
DRY SUCTION CONTROL

Fig. 23.1 (**a**) Tube thoracostomy tray, (**b**) chest drainage system

23.4 Procedure (Fig. 23.2)

1. Sterile skin preparation with sterile drape.
2. Anesthetize the appropriate area subcutaneously up to and including the rib periosteum with 5 mL of 1 % lidocaine with epinephrine (Fig. 23.2a).
3. Using a #10 or #11 blade, make an approximately 4-cm skin incision over the desired intercostal level of entry (most often the fourth or fifth intercostal space in the midaxillary line) (Fig. 23.2b, c).

   If the incision is placed *below* the fifth intercostal space, the risk of subdiaphragmatic placement into the abdominal space is increased.
4. Bluntly dissect with a hemostat or Kelly clamp through the subcutaneous tissue to the level of the intercostal muscles with intermittent opening of the dissection instrument during advancement (Fig. 23.2d, e).
5. Digitally palpate the selected intercostal space and the superior margin of the inferior rib (pay careful attention to avoid the neurovascular bundle lying inferiorly) (Fig. 23.2f).
   - If time permits, additional analgesia is recommended at this point of the procedure.
6. Guiding the closed Kelly clamp over the upper margin of the rib, enter the chest wall into the pleural cavity. (This will require some controlled force and a twisting motion.) Once the pleural space is entered, a rush of air or fluid should occur (Fig. 23.2g).

   Uncontrolled force and a lunging motion can result in penetration to the lung, heart, liver, or spleen.
7. Open the Kelly clamp while still inside the pleural space and then withdraw while the clamp is still open to enlarge the dissected tract of entry and allow easier passage of the thoracostomy tube (TT).
8. Explore the dissected tract with a sterile finger to appreciate lung tissue and possible adhesions.
9. To estimate the length the TT is to be inserted, measure the distance between the skin incision and the apex of the lung. If preferred, place a clamp over the tube at the estimated length (Fig. 23.2h).
10. Grasp the proximal end of the TT with the large Kelly clamp and pass the tube through the thoracic cavity along the previously dissected tract.
11. Release the Kelly clamp and continue to advance the tube posteriorly and superiorly.

    Make sure all of the fenestrated holes of the TT are within the thoracic cavity to prevent unnecessary manipulation and/or replacement of the TT.
12. Once the TT is in the desired position, connect the tube to the drainage device (Fig. 42.1b). Once connected, release the cross clamp on the distal end of the TT.
13. Secure the TT to the skin with 0 or 1–0 silk or nylon suture. A simple, interrupted suture above and below the TT with each stitch wrapped tightly around the TT is recommended.

    Incomplete security of the TT leads to dislodging of the tube with routine patient movements.
14. Apply petrolatum gauze over the skin closure surrounding the TT and then apply a support dressing with 4×4 gauze and adhesive tape (4 in.).
15. Obtain a chest radiograph to confirm placement of the TT.

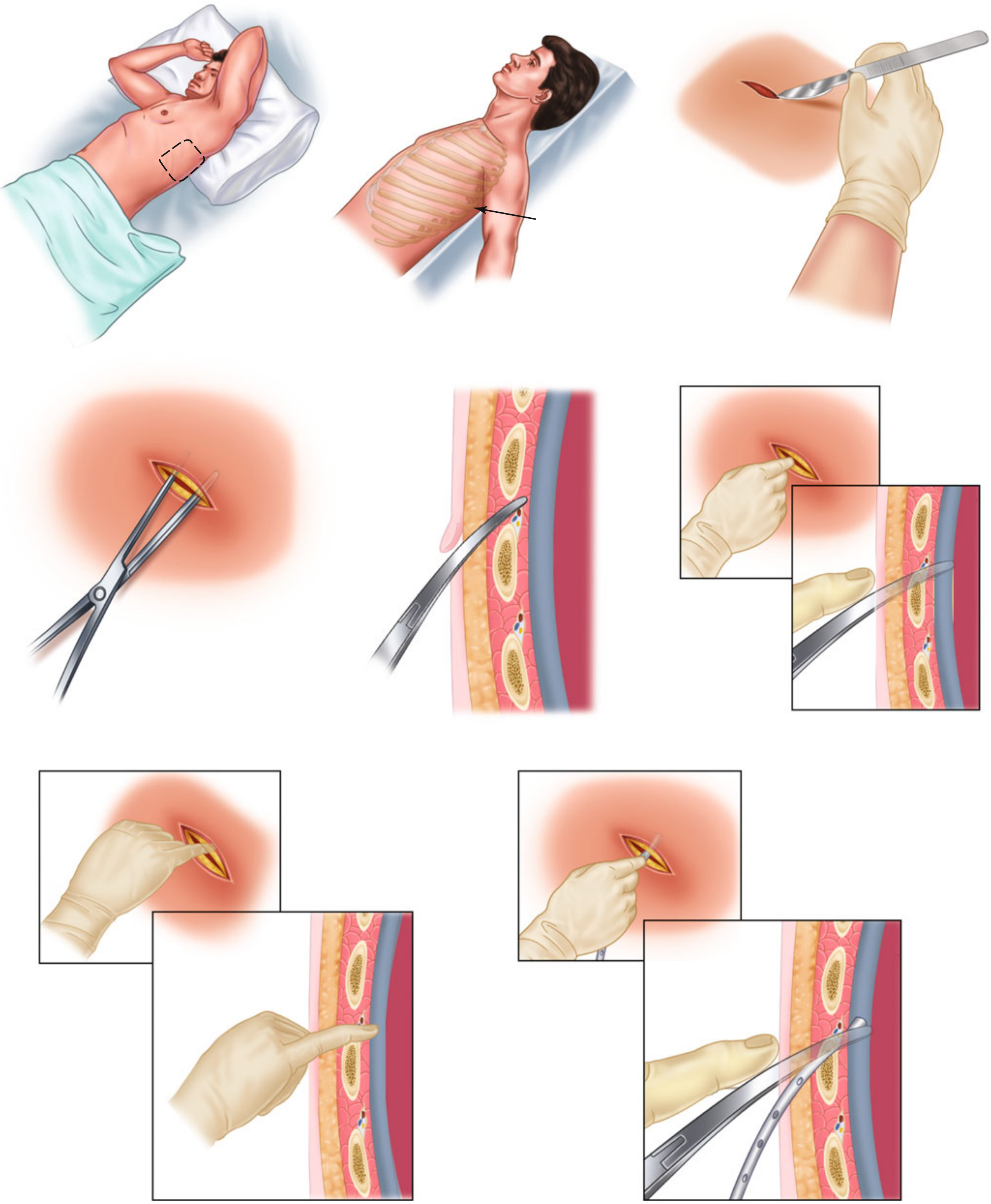

Fig. 23.2 Chest tube thoracostomy procedure

23.5 Complications

- Improper placement for pneumothorax
  - Reposition if:
    - Horizontal (over diaphragm)
    - Subcutaneous
    - Placed too far into the chest (against apical pleura)
  - Remove if:
    - Placed into the abdominal space
- Bleeding (local vs. hemothorax)
- Hemoperitoneum (liver or spleen injury)
- Tube dislodgment
- Empyema (TT introduces bacteria into the pleural space)
- Retained pneumothorax (may require second TT)
- Re-expansion pulmonary edema
- Subcutaneous emphysema

23.6 Pearls and Pitfalls

- Water seal acts as a one-way valve; if the system bubbles, there is an air leak.
- In the Pleur-evac® systems, there is an orange floater which, when static, means the desired suction pressure (usually 20 cmH_2O) has been reached.
- The negative pressure in the chest cavity equals the amount of water in water seal plus amount of suction.
- A chest tube can be removed when there is no air loss or blood for 24 h.
- When removing the tube, have the patient exhale and remove as quickly as possible.
- Leave petrolatum gauze in place for 48 h before changing it (allows wound to heal better).

Selected Reading

Ball CG, Lord J, Laupland KB, et al. Chest tube complications: how well are we training our residents? Can J Surg. 2007;50:450.

Collop NA, Kim S, Sahn SA. Analysis of tube thoracostomy performed by pulmonologists at a teaching hospital. Chest. 1997;112:709.

Dalbec DL, Krome RL. Thoracostomy. Emerg Med Clin North Am. 1986;4:441.

Miller KS, Sahn SA. Chest tubes. Indications, technique, management and complications. Chest. 1987;91:258–64.

Millikan JS, Moore EE, Steiner E, et al. Complications of tube thoracostomy for acute trauma. Am J Surg. 1980;140:738.

Thoracentesis

24

Lee Richard Donner and Michael Anana

24.1 Indications

- *Therapeutic thoracentesis* is performed to relieve dyspnea, hypoxia, or otherwise compromised respiratory function due to a large pleural effusion.
- *Diagnostic thoracentesis* is performed to aid in the diagnosis and workup of:
  - Pleural effusions of unknown cause
  - Unilateral pleural effusions
  - Pleural effusions originally determined to be due to heart failure but persisting after 3 days of diuresis

24.2 Contraindications

- Absolute
  - None
- Relative:
  - Coagulopathy, thrombocytopenia.
  - Small or loculated pleural effusion. These will increase the risk of missing the effusion and causing lung injury.
  - Positive-pressure ventilation.
  - Skin infection over the needle insertion site.

24.3 Materials and Medications

- Thoracentesis tray (commercially available kits generally include the items in the following list) (Fig. 24.1)
  - (1) Fenestrated drape
  - (1) 25-gauge × 1-in. needle
  - (1) 21-gauge × 1.5-in. needle
  - (1) 8-French catheter over 18-gauge needle
  - (1) Small plastic syringe, 5 mL
  - (1) Small plastic syringe, 10 mL
  - (1) Large plastic syringe, 50–60 mL
  - (1) Three-way stopcock
  - Specimen vials and caps
  - (1) Collection bag, 1500 mL, or vacuum container
  - (1) Tubing set
  - (1) Hemostat
  - Betadine (povidone-iodine) or other skin antiseptic preparing solution
  - 10-mL lidocaine 1 % without epinephrine

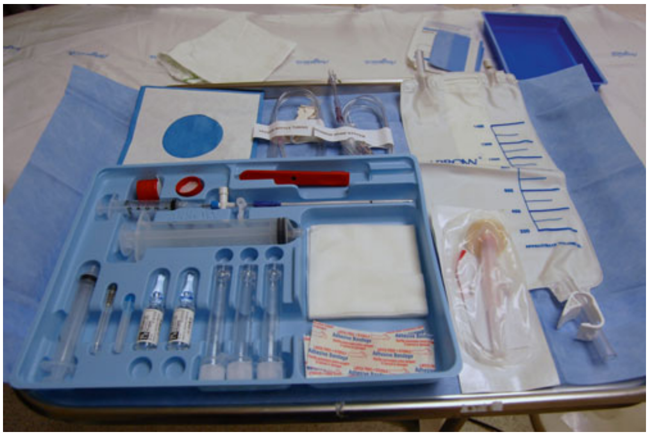

Fig. 24.1 Typical commercial thoracentesis tray

L.R. Donner, MD (✉)
Emergency Medicine Department, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: LeeDonnerMD@gmail.com

M. Anana, MD
Emergency Department, University Hospital, Rutgers New Jersey Medical School, Newark, NJ, USA
e-mail: ananami@njms.rutgers.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_24

24.4 Procedure

1. Place patient in seated position with arms at rest on a bedside table (Fig. 24.2).
2. The location and height of the pleural effusion are confirmed by physical examination. Auscultation of decreased or absent breath sounds, dullness to percussion, and decreased tactile fremitus are physical findings to confirm the location and height of the effusion (Fig. 24.3).
3. Determine and mark the site of needle insertion. This will be at the midscapular line one or two intercostal spaces below the maximum height of the effusion as determined by a combination of imaging and physical examination (Fig. 24.4).
    - *Do not* attempt thoracentesis inferior to the eighth intercostal space because respiratory cycle and anatomical variation place the diaphragm and intra-abdominal organs at risk.
4. Prepare the skin with the sterile skin preparation and sterile drape.
5. Anesthetize the appropriate area subcutaneously using a 25-gauge needle on a 10-mL syringe. Create a wheal in the skin and then infiltrate up to and including the periosteum of the rib inferior to the landmarked space with 5 mL of 1 % lidocaine.
    - Remember that the neurovascular bundle is found at the inferior border of each rib. Keep this in mind as you approach the rib and throughout the rest of the procedure to avoid injury to these structures (Fig. 24.5).
6. Using a 22-gauge needle, slowly walk the needle up and over the superior border of the rib. Continue to advance the needle along the superior border of the rib with the syringe withdrawn and infiltrating lidocaine intermittently along the way.
7. Pleural fluid will be aspirated once the pleural space is reached. Do not advance the needle further. Deposit 1–2 mL more lidocaine. A hemostat can be placed on the needle at the level of the skin to mark the depth of the pleural space and the needle can be removed.
8. Some commercial kits may come with an adjustable depth guard to be positioned at the determined depth. On an 18-gauge catheter-over-needle device, position the depth guard to the appropriate depth determined from the prior step. If a depth guard is not available, use the index finger and thumb on the catheter at the appropriate depth. With the 5-mL syringe attached, advance the device over the superior border of the rib while aspirating, expecting pleural fluid to return again at the determined depth (Fig. 24.6).
9. When the pleural space is reached, do not advance the needle further. Advance the catheter over the needle until the hub reaches the skin. Remove the needle during expiration and immediately cover the open hub with a gloved finger to prevent development of pneumothorax. Some kits provide catheters with one-way valves to prevent air entry.
10. Attach the 50–60-mL syringe to the catheter via the three-way stopcock. Pleural fluid can be drained and transferred to appropriate collection vials for diagnostic thoracentesis. A collection bag may be attached with tubing to the third port of a three-way stopcock for larger volume evacuation in the case of a therapeutic thoracentesis. Employ a syringe pump method to drain 50–60 mL of fluid at a time to the collection bag. Fill the syringe by withdrawing the plunger while the stopcock is closed to the bag. Then, close the stopcock to the patient and pump the contents of the syringe to the bag. Next, close the syringe closed to the bag, and repeat the cycle until the desired volume is drained. A vacuum container is an alternative that simply attaches via tubing to the stopcock (Fig. 24.7).
    - If using a three-way stopcock and a device that does not have a one-way valve on the catheter, be sure to always keep the stopcock closed to the patient unless withdrawing fluid in order to decrease the risk of pneumothorax.
11. When the desired amount of pleural fluid is obtained, remove the catheter during expiration and apply an occlusive dressing.

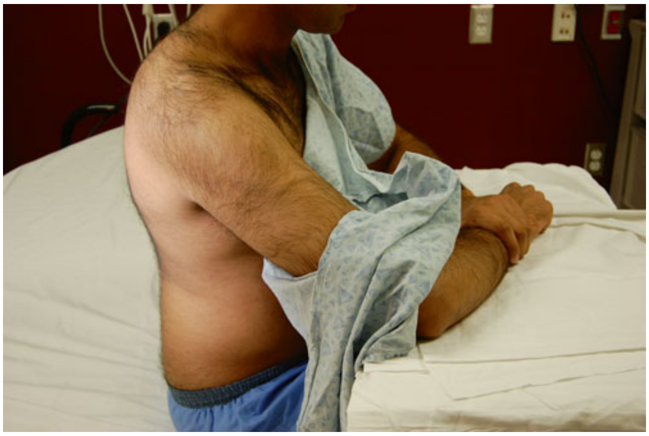

Fig. 24.2 Patient in upright, seated position

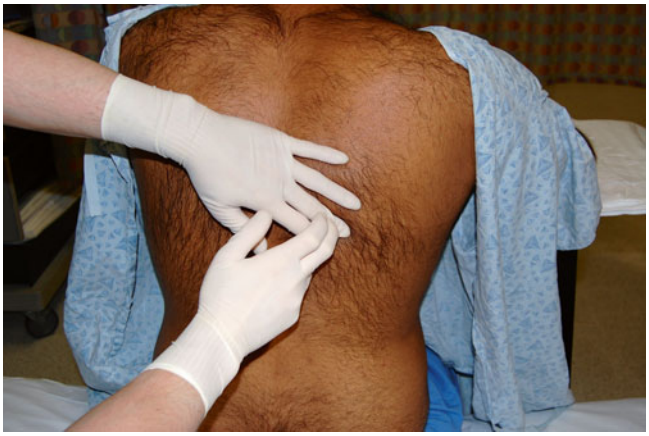

Fig. 24.3 Determining the location and height of the pleural effusion

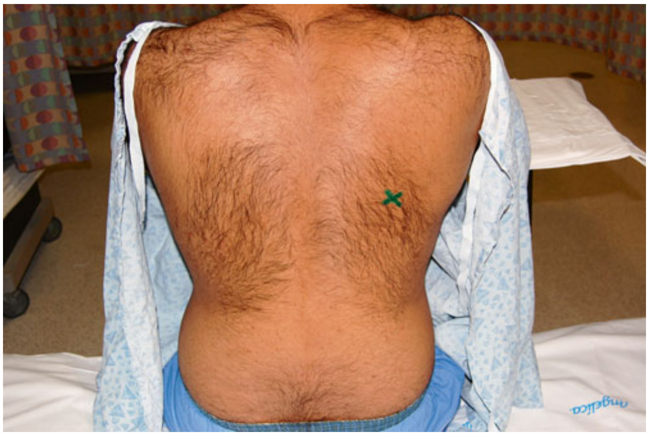

Fig. 24.4 Marked site for needle insertion

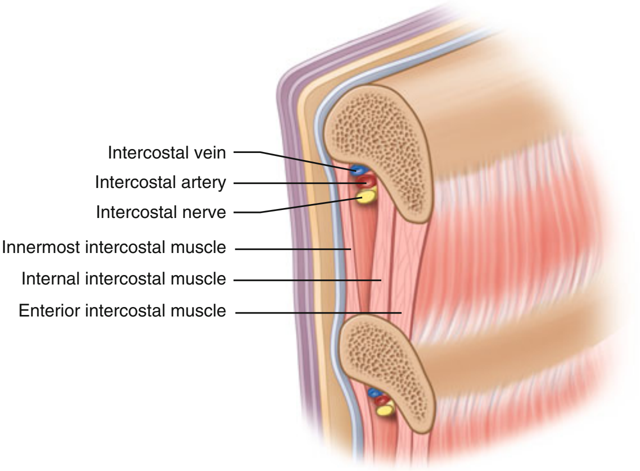
Intercostal vein
Intercostal artery
Intercostal nerve
Innermost intercostal muscle
Internal intercostal muscle
Enterior intercostal muscle

Fig. 24.5 The intercostal neurovascular bundle

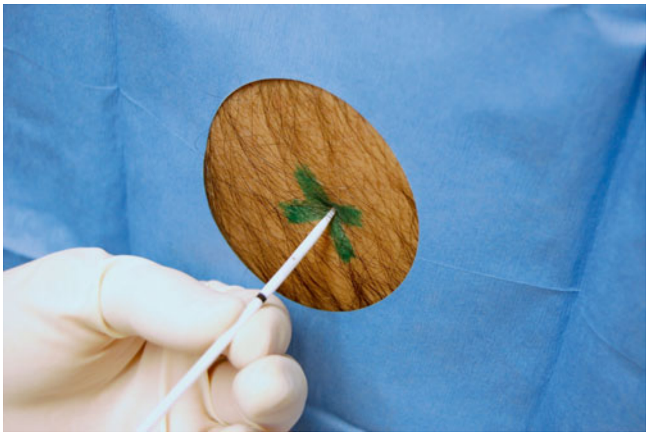

Fig. 24.6 Needle and catheter insertion, thumb and index finger at desired depth

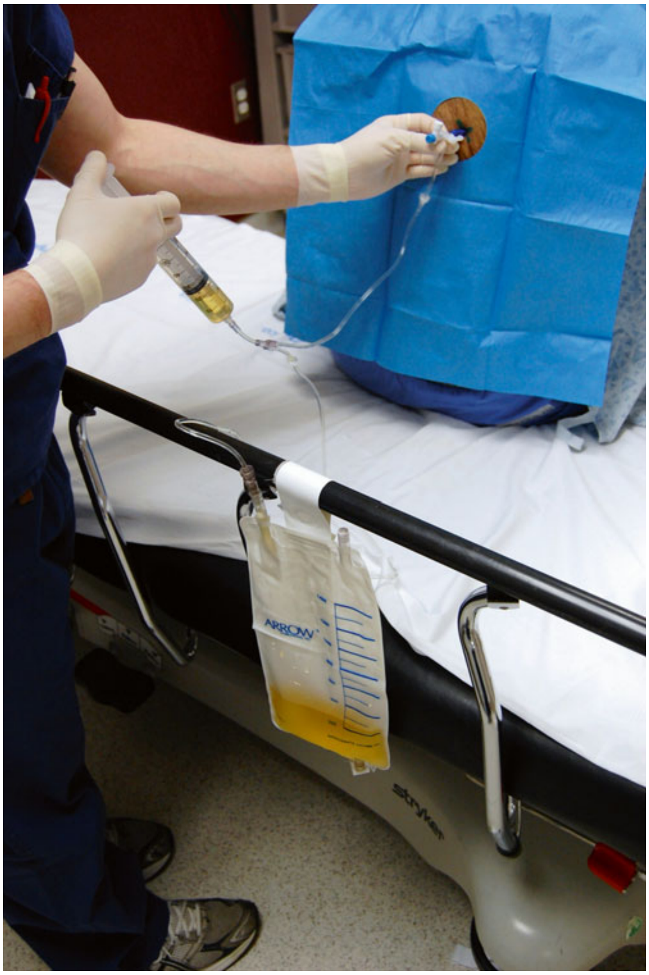

Fig. 24.7 Pleural fluid collection via syringe pump method

24.5 Complications

- Pneumothorax
- Re-expansion pulmonary edema
- Hemothorax, hematoma
- Intra-abdominal organ injury
- Air embolism
- Empyema

24.6 Considerations

- If available, the use of bedside ultrasound is highly recommended because ultrasound guidance has been shown to substantially reduce the risk of pneumothorax. Before the procedure, the height, width, and depth of the effusion can be appreciated by scanning the chest and viewing the effusion through the intercostal spaces. The use of ultrasound aids in selecting the needle insertion site by:
  - Visualizing the distance the needle must pass to reach the parietal pleura
  - Confirming the thickness of the effusion in the site selected is at least a minimum of 1.5 cm
  - Providing the clinician with a view of the effusion and surrounding structures through the complete respiratory cycle

  With these items in mind, the needle insertion site can be selected with confidence and marked before beginning the procedure.
- In addition, the use of bedside ultrasound in real time will allow the clinician to visualize the needle as it passes toward and enters the pleural space. This use requires sterile probe covers.
- Re-expansion pulmonary edema is a rare but feared complication of thoracentesis. The cause is not fully understood. Historically it was thought that re-expansion pulmonary edema was caused by removing too large a volume of fluid from the pleural space (>1–1.5 L). Another theory is that re-expansion pulmonary edema is caused when great negative intrapleural pressures (<20 cmH_2O) are generated during the procedure. The low incidence of this complication has yielded inconclusive evidence.
  - In light of this, it is prudent to continue to limit the volume of pleural fluid removed to no more than 1–1.5 L. Pleural manometry is not widely available for use in the emergency department, but should also be considered if available to maintain intrapleural pressures from reaching more negative values.

24.7 Pearls and Pitfalls

- Bedside ultrasound reduces the risk of complications.
- Never attempt thoracentesis below the eighth intercostal space.
- Thoracentesis should not be performed on pleural effusions demonstrated to be less than 1–1.5 cm thick by ultrasound or on a lateral decubitus film.
- Positive-pressure ventilation mandates extreme care while performing thoracentesis because the lungs can be punctured during inflation. In addition, ventilated patients will not be able to sit upright and will need to have the procedure performed in lateral decubitus position with the effusion side down and the posterior axillary line as the needle insertion site. Alternatively, the patient can be placed supine with the head of the bed elevated to 45° and the midaxillary line as the needle insertion site.
  - Routine post-procedure radiographs are not necessary to exclude pneumothorax. Indications for post-procedure imaging include onset of chest pain during the procedure, persisting cough or chest discomfort after the procedure, air aspiration along any step of the procedure, or positive-pressure ventilation.
  - Pneumothorax rates are higher for inexperienced providers. Although an effusion compromising respiratory function is considered a clear indication for therapeutic thoracentesis, the performance of nonurgent diagnostic thoracentesis might best be delayed for those more practiced in the procedure.

Selected Reading

Dewitz A, Jones R, Goldstein J. Additional ultrasound-guided procedures. In: Ma OJ, Mateer JR, Blavias M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill; 2008. p. 546–50.

Feller-Kopman D, Berkowitz D, Boiselle P, Ernst A. Large-volume thoracentesis and the risk of reexpansion pulmonary edema. Ann Thorac Surg. 2007;84:1656–61.

Gordon CE, Feller-Kopman D, Balk EM, Smetana GW. Pneumothorax following thoracentesis: a systematic review and meta-analysis. Arch Intern Med. 2010;170:332–9.

Light RW. Pleural effusion. N Engl J Med. 2002;346:1971–7.

Thomsen TW, DeLaPena J, Setnik G. Videos in clinical medicine. Thoracentesis. N Engl J Med. 2006;355, e16.

Open Chest Wounds and Flail Chest

25

Jacob J. Glaser and Carlos J. Rodriguez

25.1 Background

Thoracic injuries are commonly associated with penetrating and blunt abdominal trauma and are implicated in 50–70 % of trauma deaths [1]. Cardiac tamponade, tension pneumothorax, massive hemothorax, airway obstruction, flail chest, and open pneumothorax represent the six immediately life-threatening injuries attributed to chest trauma [2]. Accordingly, they must be accurately identified and dealt with urgently.

Open pneumothorax (“sucking” chest wound) is seen in penetrating chest injuries. If the associated chest wound is greater than 2/3 the diameter of the trachea (generally anything greater than 1.5–2 cm), air can preferentially enter the intrapleural space, via the trachea, with each inspiration [3] (Fig. 25.1). This allows equilibration of pressure between the pleural space and the atmosphere, causing the lung to collapse and leading to profound hypoventilation and hypoxia.

Flail chest results from high-energy blunt, crushing chest trauma causing two or more fractures in two or more contiguous ribs. Classically, the fractures are lateral or sternal. Posterior rib fractures rarely cause flail physiology (Fig. 25.2). Flail chest has been reported to have mortality as high as 16 % [4]. This injury pattern is associated with a high incidence of underlying pneumothorax, hemothorax, pulmonary contusion, and chest wall instability. Mortality from flail chest is thought to be correlated with the degree of underlying pulmonary contusion and attendant hypoxia [2].

Both open pneumothorax and flail chest are immediately life threatening and require early appropriate management.

J.J. Glaser, MD (✉)
Combat Casualty Care Directorate, Naval Medical Research Unit,
San Antonio, TX, USA
e-mail: Jacob.glaser1@gmail.com

C.J. Rodriguez, DO
Division of Trauma Surgery, Surgical Critical Care, Walter Reed
National Military Medical Center, Bethesda, MD, USA
e-mail: Carlos.J.Rodriguez.mil@health.mil

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_25

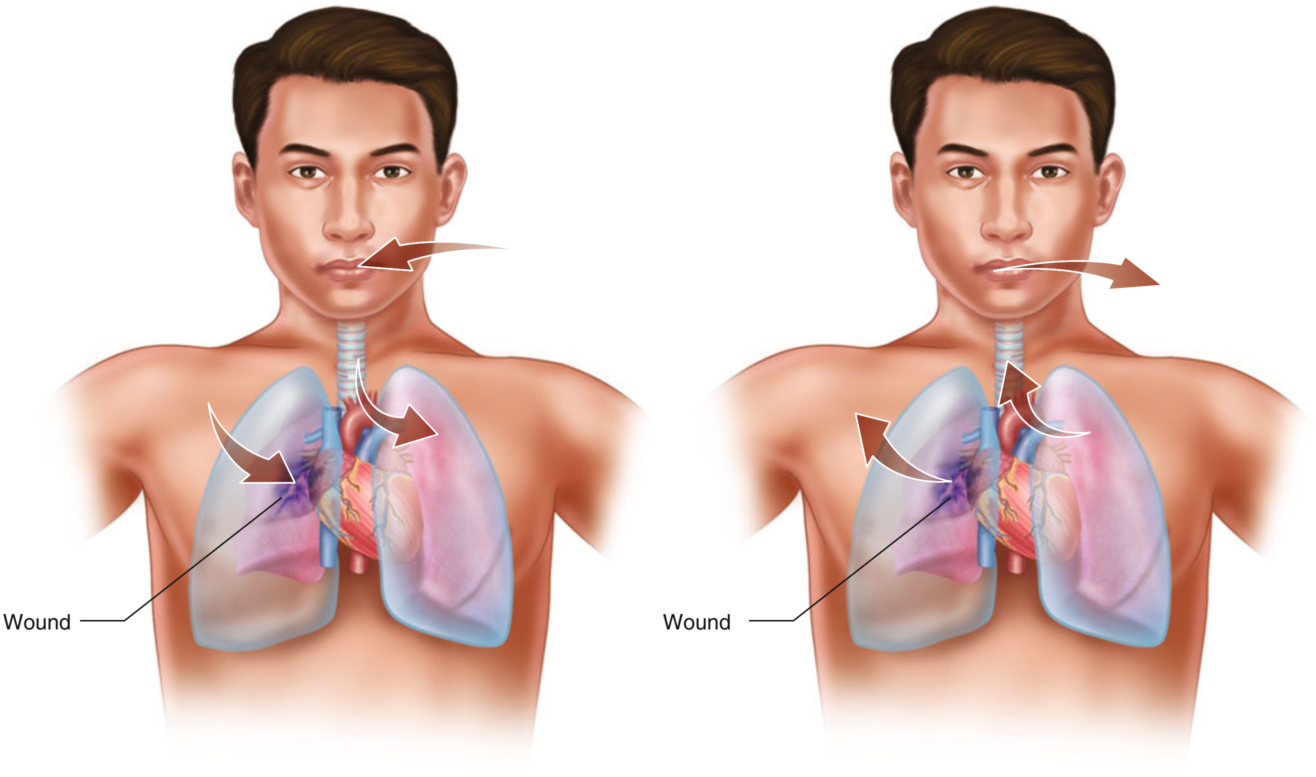
Wound
Wound

Fig. 25.1 Air preferentially will enter the chest via the wound, collapsing the lung on the affected side

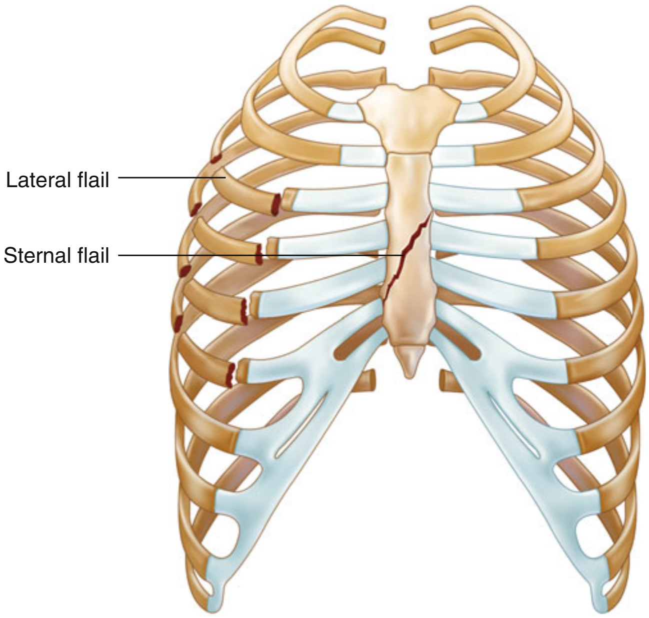
Lateral flail
Sternal flail

Fig. 25.2 Flail chest: two or more rib fractures in two or more segments. Lateral flail (most common) and sternal flail segments represented below

25.2 Initial Evaluation of Thoracic Trauma

- To the best extent possible, obtain good patient handover from the prehospital caregivers.
- Initial physical evaluation.
  - Appropriate attention should be given to the ABCs (airway, breathing, circulation) of ATLS (Advanced Trauma Life Support) management.
  - Evaluation and resuscitation are to be *concurrent* with diagnostic procedures and immediate interventions.
  - Maintain a high degree of suspicion for open chest wound in impalement injury and destructive penetrating trauma (blast injury or shotgun).
  - Maintain a high degree of suspicion for flail chest in high-energy direct impact trauma (motor vehicle crash, fall, crush injury).
- Administer high-flow O_2 with a non-rebreather mask.
  - If patient is in respiratory distress, is unstable, or has an obvious chest wall defect, consider early intubation to secure the airway.
  - Inspect the chest wall for occlusive dressings.
- Decompress the chest if tension physiology is present or suspected.
  - Immediate decompression of a suspected tension pneumothorax can be readily accomplished by removing any existing occlusive dressings *or*
  - Place a large-bore cannula over the rib, second intercostal space, or midclavicular line.
- Specific immediate management appropriate to open chest wound and flail chest (see later).
- Monitor continuous pulse oximetry and electrocardiogram.
- Initiate crystalloid resuscitation via large-bore intravenous (IV) access.
  - Early mobilization of blood products if ongoing hemorrhage or expectation of excessive blood loss.
  - Placement of resuscitative lines concurrent with management of respiratory parameters.
- Early surgical consultation for management of intrathoracic injuries and management of chest wall defect.

25.3 Open Pneumothorax ("Sucking Chest Wound")

- Immediate management requires attention to airway and respirations.
  - If in respiratory distress, *intubate*.
- Close the chest wall defect with an occlusive three-sided dressing.
  - This includes a valve mechanism that allows trapped air to escape, preventing tension (Fig. 25.3).
  - An IV bag cut to fit the wound and then taped on three sides can be useful in an emergent situation.
  - Commercial products are available and appropriate for smaller wounds, including the Asherman Chest Seal and HyFin Vent.
- A completely occlusive dressing may quickly convert an open chest wound into a tension pneumothorax [2, 3, 5] and therefore should *never* be done.
  - The patient and dressing must be serially checked to ensure that trapped air is allowed to escape.
  - If there is any doubt, immediately remove the dressing and replace it with an appropriate dressing.
- These maneuvers are a bridge to definitive care.
- When the timing is appropriate (i.e., time and resources are available), perform a formal tube thoracostomy and convert to a completely occlusive dressing over the wound.
  - Avoid placing the tube through the open wound.
- Once tube thoracostomy, placement of occlusive dressing, and the airway are secured, the pathophysiology of the open pneumothorax becomes inconsequential.
- Immediate consultation with surgery for definitive care of associated intrathoracic injuries is required.
  - The patient may need urgent thoracotomy to treat associated injuries.
  - Irrigation and debridement should take place in the operating room.
  - Depending on injury severity, the patient may need chest wall reconstruction.

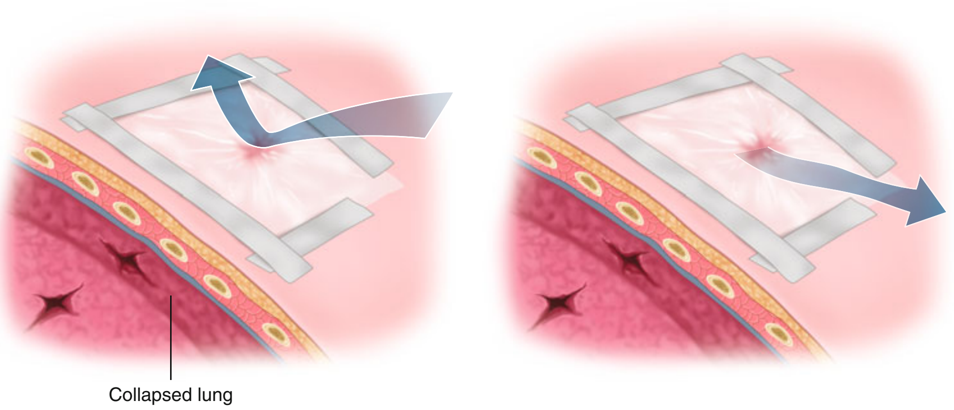

Fig. 25.3 Three-sided dressing, allowing a valve to decompress tension in the chest

25.4 Flail Chest

- Diagnosis is made from mechanism and examination, not radiographically.
  - With inspiration, the affected chest will move inward with negative pressure.
  - With expiration, the affected chest will move outward.
  - Patients who are intubated on positive pressure (and not spontaneously breathing) will often not show this paradoxical chest movement.
- Attention *must* be paid to presumed underlying blunt lung injury and contusion.
  - The degree of underlying contusion (not the flail segment itself) is directly related to the degree of hypoxia and associated morbidity and mortality [2, 3].
  - There should be a low threshold for intubation to manage respiratory distress, hypoxemia, or hemodynamic instability.
  - The patient is at high risk for hemothorax and pneumothorax requiring tube thoracostomy.
- Abdominal injuries may be present in up to 15 % of patients [2].
- After initial stabilization, treatment is supportive.
  - Intensive care unit admission for management of underlying pulmonary contusion.
  - Pain control in the form of epidural or regional block for excellent pain control.
  - Close attention to pulmonary toilet and lung re-expansion.
  - Patients require observation and treatment in a monitored setting until ensured that respiratory parameters and oxygenation are improving.
- Surgical stabilization of the chest wall is rarely performed.
  - Early surgical consultation for chest wall fixation in questionable cases is warranted.

References

1. LoCicero J, Mattox KL. Epidemiology of chest trauma. Surg Clin North Am. 1989;69:15–9.
2. Pietzman AB, Schwab CW, Yealy DM, editors. The trauma manual. 2nd ed. Philadelphia: Wolters Kluwer Health; 2000.
3. Weinberg JA, Croce MA. Chapter 33: Chest wall injury. In: Flint L, Meredith JW, Schwab CW, editors. Trauma: contemporary principles and therapy. 1st ed. Philadelphia: Lippincott Williams & Williams; 2007.
4. Clark GC, Schecter WP, Trunkey DD. Variables affecting outcome in blunt chest trauma: flail chest vs. pulmonary contusion. J Trauma. 1988;28:298–304.
5. Borden Institute Walter Reed Army Medical Center. Emergency war surgery. 3rd ed. Washington: Office of the Surgeon General U.S. Army, Borden Institute; 2004.

Emergent Resuscitative Thoracotomy, Open Cardiac Massage, and Aortic Occlusion

26

Kevin M. Jones and Jay Menaker

26.1 Indications

- Penetrating chest trauma with recent or immanent loss of vital signs
- Consider in blunt trauma with pericardial tamponade or exsanguination where aortic occlusion may provide proximal control

26.2 Contraindications

Absolute
  - Prolonged cessation of vital signs
  - Injury profile obviously incompatible with life
  - Absence of surgical services to whom care can be transferred

Relative
  - None

26.3 Materials and Medications (Fig. 26.1)

- Betadine (povidone-iodine) for rapid skin preparation
- #10 Scalpel
- Mayo or long Metzenbaum scissors
- Finochietto retractor (rib spreader)
- Long DeBakey or other tissue forceps (2)
- Satinsky vascular clamp and/or straight vascular clamp
- Long needle holders (2)
- Lebsche knife or sternal osteotome with hammer
- Lap sponges or gauze pads

K.M. Jones, MD (✉)
Trauma and Surgical Critical Care, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA
e-mail: KJones12@umm.edu

J. Menaker, MD
Departments of Surgery and Emergency Medicine, R Adams Cowley Shock Trauma Center, University of Maryland Medical Center, Baltimore, MD, USA
e-mail: jmenaker@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_26

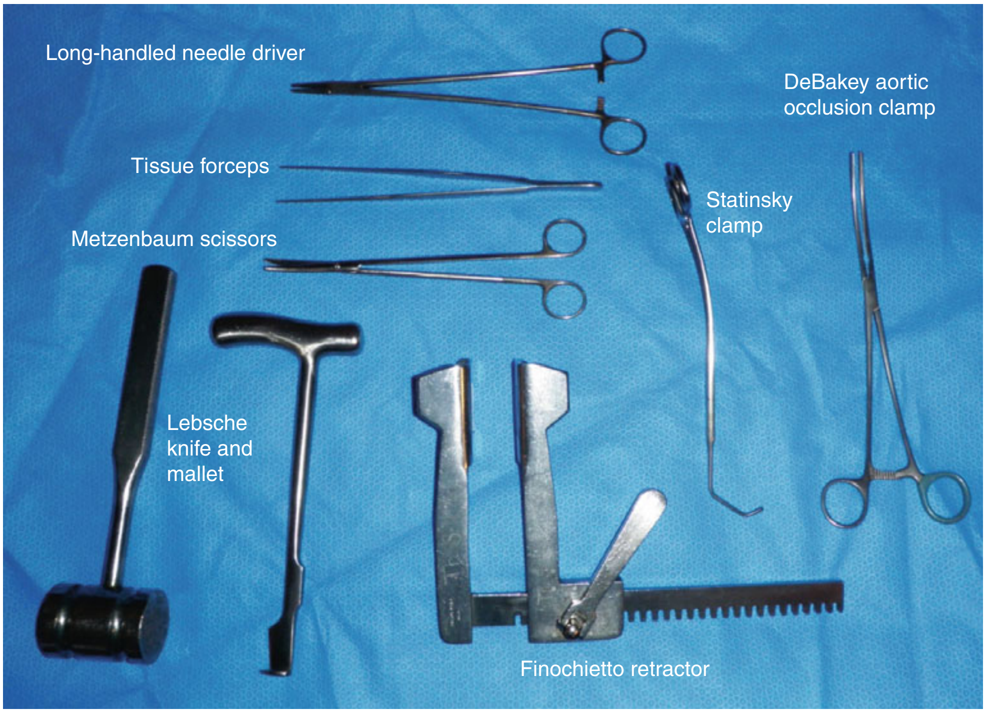
Long-handled needle driver
DeBakey aortic occlusion clamp
Tissue forceps
Statinsky clamp
Metzenbaum scissors
Lebsche knife and mallet
Finochietto retractor

Fig. 26.1 Basic components of an emergency thoracotomy tray: Lebsche knife and mallet for crossing sternum (bone-cutting forceps or sternal osteotome would also suffice), Finochietto retractor, atraumatic vascular clamps (a Satinsky clamp and a DeBakey aortic occlusion clamp), long-handled needle driver, tissue forceps, and Metzenbaum scissors. *Not illustrated* Scalpel with #10 or #20 blade, Mayo scissors

26.4 Procedure

26.4.1 Resuscitative Thoracotomy and Open Cardiac Massage

1. Rapidly prepare the entire anterior and bilateral chest with Betadine.
2. Using a scalpel, incise the skin and subcutaneous tissue from just right of the sternum to the anterior margin of the left latissimus dorsi, following the curvature of the inframammary crease or the fourth or fifth intercostal space (Fig. 26.2).
    - This incision is often made too low on the patient's chest. It should extend across the sternum, not at the level of the xiphoid. Upward retraction of the breast may help provide access to the fourth or fifth intercostal space, where the incision should be located.
3. Bluntly enter the right pleural space through the fourth or fifth intercostal space.
4. Using scissors, cut the intercostal muscles, dividing between the fourth and the fifth ribs from the sternum to the posterior axillary line.
5. If better exposure to the heart is desired, some practitioners advocate extending the incision across the sternum using a Lebsche knife, sternal osteotome, or bone-cutting forceps at this time (Fig. 26.3).
    - If the thoracotomy incision extends to or through the sternum, tie off the internal mammary arteries before closing the chest should the patient be successfully resuscitated, because these will have been divided with the incision.
6. Insert a Finochietto retractor and retract the ribs in order to gain access to the left chest and expose the pericardium (Figs. 26.4 and 26.5).
    - Insert a Finochietto retractor with the rack and pinion bar down and lateral, as in Fig. 26.4, so as not to interfere with extension of the thoracotomy across the sternum into a clamshell maneuver if needed.
7. If massive left pleural hemorrhage is encountered, investigate and control the source at this time.
8. Using tissue forceps, raise a portion of the pericardium anterior to the phrenic nerve, and enter the pericardium using scissors (Fig. 26.6).
9. Widely open the pericardium with scissors, cutting in a cranial-caudal direction anterior to the phrenic nerve and deliver the heart (Fig. 26.7).
10. If hemopericardium is encountered, investigate for and initiate appropriate repair of identified cardiac injuries.
11. Initiate open cardiac massage by cupping the heart between the flattened palmar aspect of the fingers of both hands and rhythmically compressing the heart from apex to base, relaxing completely between compressions to allow filling (Fig. 26.8).

26.4.2 Aortic Occlusion

1. Expose the posterior aspect of the left mediastinum by having an assistant retract the left lung superomedially, dividing the inferior pulmonary ligament if necessary (Fig. 26.9).
2. Bluntly dissect the pleura separating the pleural and mediastinal space just anterior to the vertebral bodies, exposing the aorta (Fig. 26.9).
3. Completely encircle the aorta with the finger of the nondominant hand (Fig. 26.10).
    - Differentiating the aorta from the esophagus when the patient is in a state of profound shock is very difficult. Having an assistant pass an orogastric tube may help distinguish the two. The aorta should be the most posterior structure, lying immediately on the anterior aspect of the vertebral bodies.
4. With the aorta completely encircled, place a vascular clamp across the aorta and verify by sight and feel that the complete vessel is occluded within the clamp (Fig. 26.11).

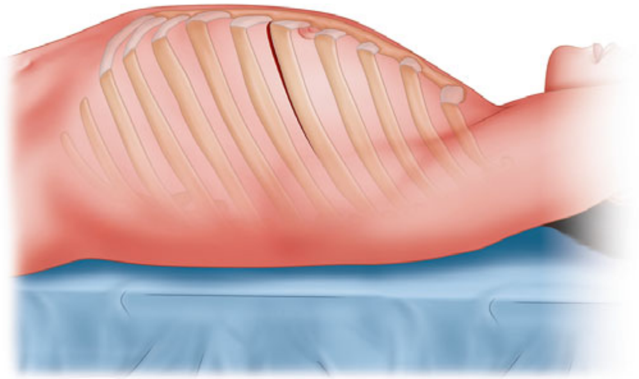

Fig. 26.2 Raise the left arm above the head. Make an incision along the left fourth intercostal space, just below the nipple in a male or at the inframamillary crease in a female

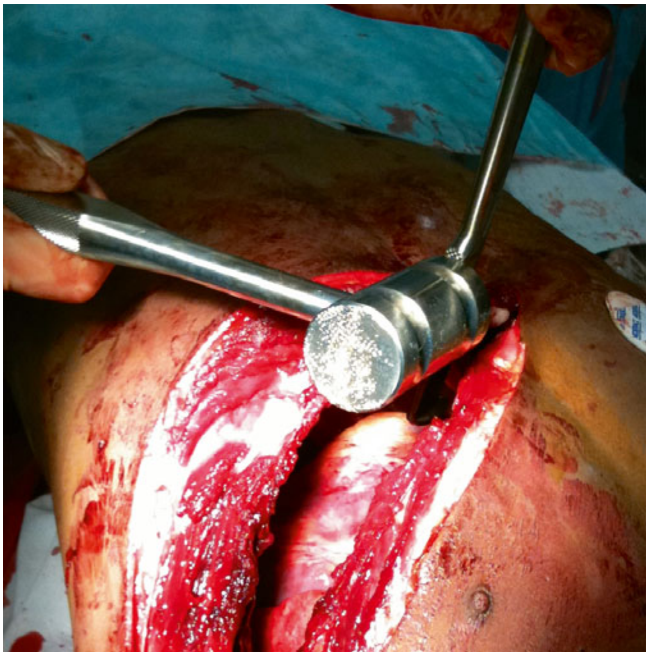

Fig. 26.3 Use a Lebsche knife to extend the incision across the sternum to improve exposure to the heart

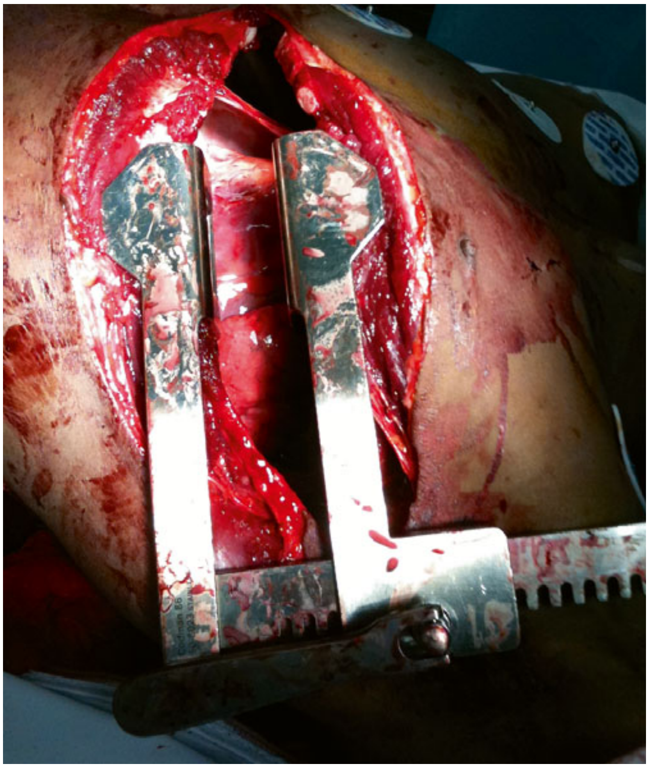

Fig. 26.4 Finochietto retractor placed through the fourth intercostal anterolateral incision. Note the rack and pinion bar placed posterolaterally, where it will not impede access to the midline

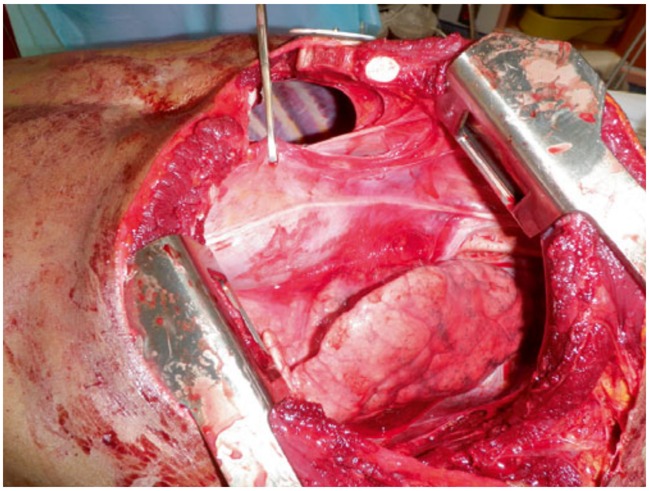

Fig. 26.5 Finochietto retractor extended, exposing the left pleural space and the pericardium. In this case the sternum has been divided as above, and the resultant window into the right pleural space is seen anterior to the pericardium. The lung is deflated in this postmortem picture but would be far more an obstacle in the actively ventilated patient. The clamp seen in the upper portion of the picture is reapproximating the pericardium, which has previously been divided, for the sake of illustration

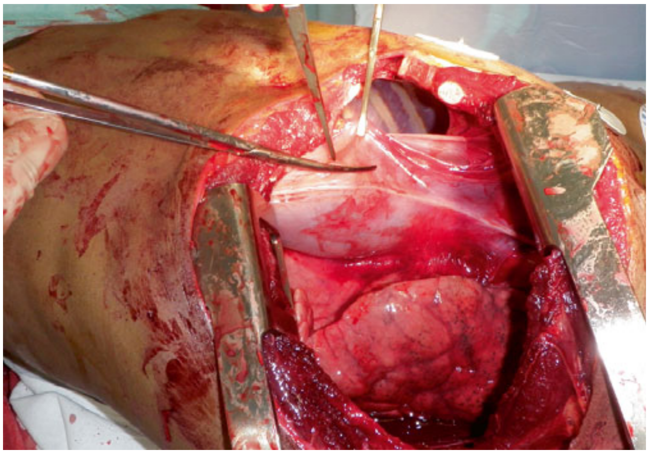

Fig. 26.6 The pericardium is lifted with tissue forceps and opened with a nick using scissors. The phrenic nerve is easily visualized running cranial-caudal just below the scissors in this picture

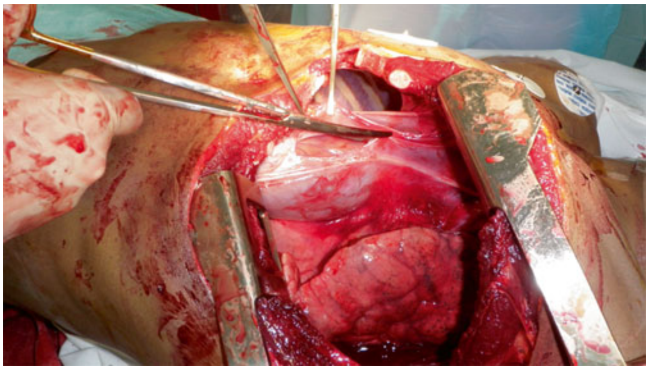

Fig. 26.7 Open the pericardium widely in the cephalad-caudad plane anterior to the phrenic nerve, taking care not to damage the phrenic nerve

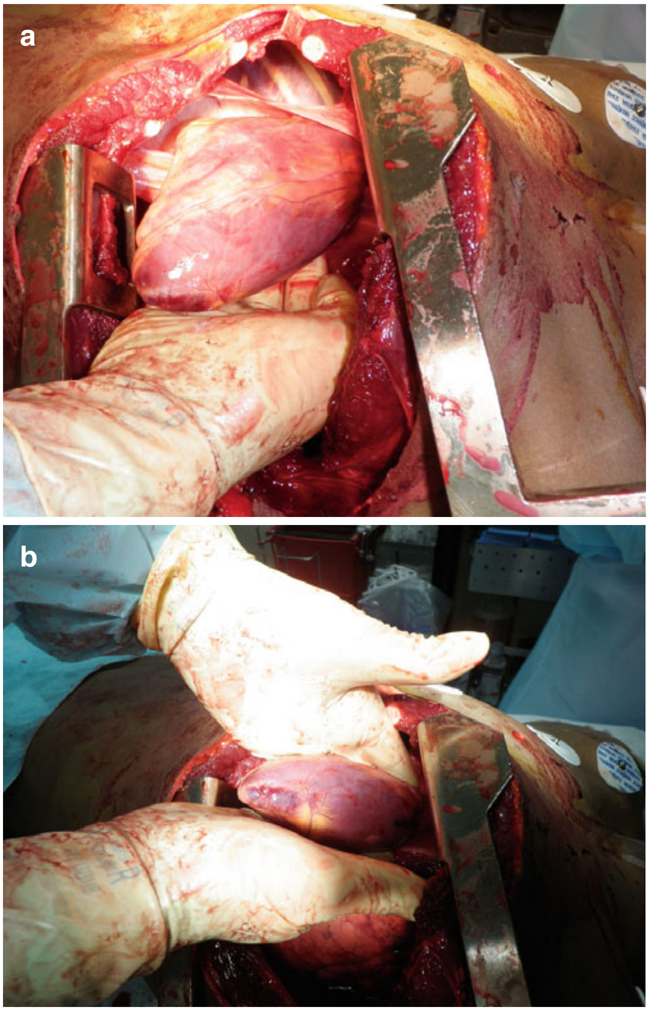

Fig. 26.8 (**a**) Deliver the heart from the pericardium, rapidly assess for cardiac injury requiring damage-control repair; (**b**) initiate open cardiac massage

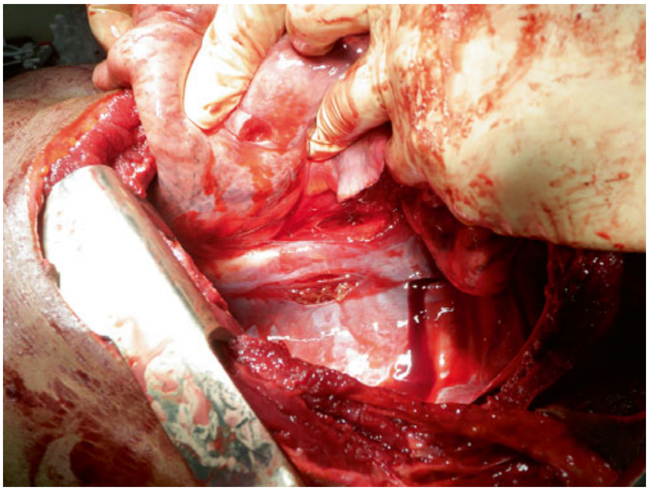

Fig. 26.9 The heart and lung are retracted superomedially, allowing visualization of the left posterior mediastinal pleura; the aorta lies just anterior to the vertebral bodies and has been previously isolated through the pleural interruptions seen here. The heart and lung are assertively retracted here for the benefit of illustration of the posterior mediastinum. Such retraction would completely occlude venous return

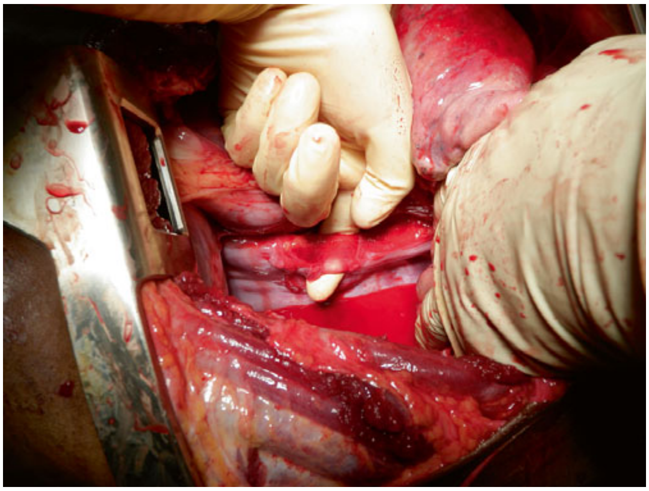

Fig. 26.10 After bluntly dissecting the mediastinal pleura, the aorta is looped using a finger of the nondominant hand

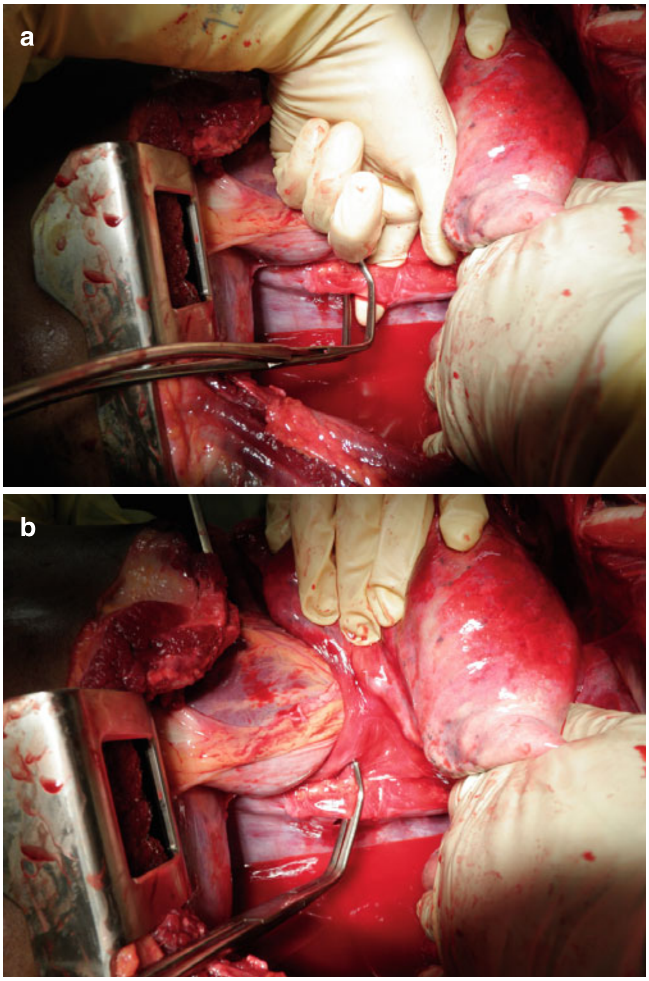

Fig. 26.11 (**a**) A vascular clamp (a Satinsky clamp is used here, although any large atraumatic vascular clamp can be used) is applied across the descending aorta, (**b**) followed by visual and tactile confirmation that the aorta is completely occluded

26.5 Complications

- Injury to care providers, by means of scalpel, needlestick, or sharp foreign body, is the principal concern.
- Post-emergency department thoracotomy infections are rare, even given the less than optimal sterile conditions.
- Damage to lung parenchyma during the initial incision is common and often leads to air leak in survivors.
- Neglect of the mammary arteries, often divided during emergent thoracotomy and not briskly bleeding in the shock state, will result in intrathoracic hemorrhage if not tied off.

Selected Reading

Bartlett RL. Resuscitative thoracotomy. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 3rd ed. Philadelphia: WB Saunders; 1998.

Feliciano DV, Mattox KL. Indications, technique, and pitfalls of emergency center thoracotomy. Surg Rounds. 1981;4:32.

Siemans R, Polk Jr MC. Indications for thoracotomy following penetrating thoracic injury. J Trauma. 1977;17:493.

Wall Jr MJ, Huh J, Mattox KL. Indications and techniques of thoracotomy. In: Feliciano DV, Mattox KL, Moore EE, editors. Trauma. 6th ed. New York: McGraw Hill; 2008.

Lung Ultrasonography

27

Ali H. Dabaja, L. Connor Nickels, and Rohit Pravin Patel

27.1 Indications

- Evaluation of respiratory failure and insufficiency due to pneumothorax, pleural effusion, pulmonary edema, acute respiratory distress syndrome (ARDS), and alveolar consolidation (atelectasis, pneumonia, aspiration)
- Monitoring progress of diseases such as pulmonary edema and pneumothorax
- Procedural guidance during pleural fluid removal or pneumothorax treatment
- Procedural guidance for chest tube placement for complex pleural effusions, hemothorax, pneumothorax, and other pleural diseases

27.2 Contraindications

- None

27.3 Materials

- Ultrasound equipment/machine, ultrasonography gel
- Microconvex or phased-array cardiac probe for sufficient evaluation of lung artifact (use abdominal preset). Linear-array probes, or vascular probes, can be used for more detailed evaluation of the pleura, although this is insufficient for penetration to deeper structures.
- Sterile materials and equipment where appropriate

27.4 Preparations

- Clean ultrasound equipment and use sterile operating procedures and ultrasound conductive gel whenever adequate.
- Ultrasound positioning. Place ultrasound device where it is easily accessible and in adequate view.
- Proper environment. Minimize light interference. Close shades and lights where appropriate.
- Position the patient. This helps optimize and expedite the examination. Supine position should be adequate for critically ill patients. Upright positioning is best for all others, with arms abducted above the head or spaced away from the chest to allow space for the probe. Patients should be moved to the edge of the bed. Uncover appropriate areas, including anterior and lateral chest walls.
- Equipment use. Turn on equipment, enter patient data, and select proper probe (cardiac probe or abdominal probe with abdomen presets). Depth and gain should be adjusted accordingly, where depths in the range of 4–10 cm give proper evaluation of more superficial and deeper structures, respectively.

A.H. Dabaja, DO
Division of Critical Care, Department of Anesthesia, University of Florida Health, Gainesville, FL, USA
e-mail: ali.dabaja@gmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_27

27.5 Procedure

1. Probe position (Fig. 27.1). Start examination at the midclavicular level at the space between the second and the third ribs. Probe positioning should be perpendicular to the ribs (longitudinal positioning) with the ultrasound marker pointed cephalad. This should place the most superficial structures at the top, with deeper structures at the bottom of the monitor. Upon completion, probe positioning can be mapped to evaluate three or four additional areas, typically between the anterior and the posterior axillary lines. Lateral views with the probe should be most posterior, typically along the posterior axillary line, tracking caudad toward the diaphragm. PLAPS (posterolateral alveolar and/or pleural syndrome) pointed posteriorly should also be evaluated, specifically in supine patients. PLAPS is lateral to the scapula and typically requires that the patient be lifted off the bed from one side.
2. Identify the “bat sign” (Fig. 27.2). The initial view observed should be a window of lung flanked by two rib shadows. This view, termed the “bat sign,” should now allow evaluation of the parietal and visceral pleura, seen most superficial as echogenic line (approximately 0.5 cm below start of rib shadows), subsequent lung sliding, and other findings such as “A” and “B” lines, as well as abnormal lung tissue.
3. Identify “A lines” (Fig. 27.3). A lines indicate air. These are multiple echogenic lines appearing horizontally in sequence deep to the pleural line. This artifact represents reverberations of the pleura and can be found in aerated lungs, which can be normal or abnormal (e.g., pulmonary embolism, chronic obstructive pulmonary disease [COPD]). The first true A line, denoted “A1,” is found equidistant from the chest wall to the pleural line. Many other A lines might be seen and are denoted “A′” lines. Subsequent equidistant A lines are “A2,” “A3,” and so on.
4. Identify “B lines” (Fig. 27.4). These artifacts appear in well-aerated lung and are vertical echogenic lines (ray, flashlight, lung rockets) transmitted from the pleura to the deeper parts of the lung on the ultrasound monitor field. They are due to thickened interlobular septa and represent alveolar fluid surrounded by air. True B lines arise from the pleural line and shoot all the way down to the far lung fields, whereas “comet tails” are seen only close to the pleural and are sometimes referred to as “shimmering” or “glimmering” during movement of the pleural line. When multiple B lines are seen in a patient, it is sometimes referred to as “lung rockets” or “flashlights” because many rays are shooting from the pleura. Even though most of the time B lines represent pulmonary edema, they can be seen in other conditions such as aspiration, pulmonary fibrosis, acute respiratory distress syndrome (ARDS), and pneumonia.
5. Identify lung sliding (Fig. 27.5). Lung sliding identifies movement of a normal parietal-visceral interface. Patient breathing causes a rhythmic movement of parietal against the visceral pleura appearing as movement of the hyperechoic line. M-mode can be used to show a timed clip of this through a still image and should only be used as a method of reporting or saving for documentation purposes. A “lung point” is an area adjacent to lung sliding (parietal pleura) that is devoid of movement. This is highly specific for pneumothorax. If one suspects a pneumothorax, quantification of the size can be made by evaluation at the anterior lung points and movement toward the lateral wall of the chest. The more rib spaces found to have absent lung sliding, the larger the pneumothorax. Lung sliding (Fig. 27.5) can be evaluated with M-mode, which can help identify a normal parietal-visceral interface at that level. Obtain an adequate two-dimensional view (“bat sign”) and press the “M-mode” option on the equipment. A normal interface appears as multiple hyperechoic lines, the pleura (termed “seashore”; Fig. 27.5a), followed by a sandlike pattern, the lung tissue. This pattern together is termed a “seashore sign.” Air that disrupts the parietal-visceral interface, as found within a pneumothorax, is identified as horizontal repeating echogenic lines, similar to a barcode (Fig. 27.5b), and is termed “stratosphere sign.” M-mode can be discontinued.
   - Lung sliding can be absent in conditions other than pneumothorax: apnea, right or left bronchial intubation, lung collapse (blebs), pneumonia, and pulmonary fibrosis.
6. Identify lung pulse. This appears as a shimmering of the pleural line due to cardiac activity. This is most apparent on the left side of the chest, closest to the heart. This helps to exclude pneumothorax as well.
7. Move posteriorly. Move the ultrasound probe laterally and posteriorly to the PLAPS point. The transducer can be directed toward the center of the patient’s body in supine patients. Pleural effusions and consolidations are found in the dependent areas of the lung.
8. Move caudally. With the marker still pointing cephalad, move along the posterior axillary line in two or three additional rib spaces. Identification of pleural disease and other pathology requires multiple views and will aid in evaluating the extent of the disease. This will also allow for identification of boundaries of the lung, such as the diaphragm. Identification of the diaphragm is most critical to determine location of fluid.

9. Identify the diaphragm and liver and/or spleen (Fig. 27.6). Along the posterior axillary line, or the posterior chest wall, move the probe caudad to identify the diaphragm. This appears as an echogenic curvilinear structure, with the liver or spleen being subdiaphragmatic and typically of different echogenicity than the lung. Many times the diaphragm is very high in the supine critically ill patient. Massive edema and obesity may also degrade image quality in this location.
    - Always identify the diaphragm. Hypoechoic fluid surrounding the liver or spleen can appear as a pleural effusion and must not be mistaken as such. In addition, lung tissue may mimic hepatic tissue in certain diseases such as dense consolidations termed "hepatization" of the lung. Proper probe positioning, clear identification of the diaphragm, subdiaphragmatic structures, and lung are crucial. This is a common error in novice operators owing to the confusion of the hepatorenal or splenorenal recess for the diaphragm.
    - Identifying the diaphragm can be technically difficult depending on patient position, size, and clinical condition. It maybe useful to start below the diaphragm, first identifying the hepatorenal recess (liver and kidney view interface) and then moving cephalad until the lung and diaphragm are visualized. In addition as ribs change their orientation anatomically, the probe may need to be adjusted while still in the longitudinal axis. Moving the probe clockwise and counterclockwise may be of benefit to bring into view the lung, the diaphragm, and the subdiaphragmatic structures.
10. Identify pleural effusions (Fig. 27.6). Confirming the presence of pleural effusions requires identifying anechoic material between the pleura and the lung. This can be seen as lung movement in an undulating pattern, which typically is facilitated by cardiac activity and respirations. This is termed "jellyfish sign," where the lung flaps as it freely floats in the effusion. Floating debris can also confirm effusion, termed "plankton sign." It is also important to identify the depth of the chest wall to the pleural fluid in order to determine the best location/depth of needle insertion when attempting thoracentesis or chest tube insertion. The challenge is to find a safe path for needle insertion. The key when using ultrasound as guidance is that the angle of the needle/syringe assembly must duplicate the angle of the probe. The time between scan and needle insertion must also be minimized. Real-time guidance is not required. In patients who are obese or edematous, skin indentation during probe placement can result in underestimation of depth of needle insertion required and must be taken into account. Safety margin of thoracentesis is thought to be 10 mm of visceral-parietal distance. Dry taps may be due to loculations the blocking needle, needle plugs, patient movement from scan to tap, and poor angle selection.
    - Exudates, empyemas, and hemothoraces may appear more echogenic, unlike, for example, a transudative effusion that could be anechoic. Complex effusions can also appear as heterogeneous and echogenic. The consistency of the effusion can make identification technically difficult because this can limit lung motion. Sometimes the operator may think there is no effusion when there is an echo-dense effusion.
11. Identify consolidations (Fig. 27.7). Compressed lung appears with alveolar consolidation pattern (tissue-like sign). Alveolar consolidations are devoid typically of air and appear as tissue density; these can be atelectasis, pneumonia, aspiration, or other diseased lung. "Hepatization" of the lung is typical, where the images mimic liver tissue. Images may also have hyperechoic foci representing air bronchograms, which would indicate pneumonia. Probe location should be correlated with an anatomical lobular or segmental area.
    - Lung may slide into the effusion during the respiratory cycle and can be problematic during needle insertion, causing pneumothorax or abnormal wire placement during the performance of pigtail chest tube catheters. This is called a "curtain sign."
12. Sinusoid sign (Fig. 27.8). M-mode is placed in the center of the visible lung when a large amount of pleural fluid is seen. A sinusoid sign strengthens the operator's determination that pleural fluid is present and that the pleural fluid is not necessarily compromising lung dynamics. If the sinusoid sign is absent, it may indicate a "trapped" lung dynamic.
13. Assessment and clinical decision making. Upon completion of ultrasonography of bilateral lung fields, clinical decision-making tools may be of benefit, especially in undifferentiated respiratory failure. A protocol has been developed to organize the exam of a respiratory failure patient (on noninvasive or invasive ventilation only). The BLUE protocol assesses patients based on findings (e.g., A lines, B lines, lung sliding) of both lungs and incorporates them into an algorithm. With acceptable sensitivities and specificities, practitioners can diagnose pulmonary edema, pneumonia, pneumothorax, and COPD/asthma with the BLUE protocol.

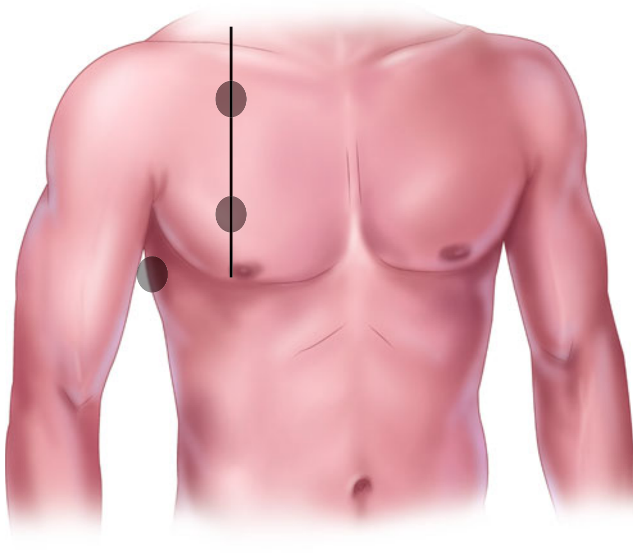

Fig. 27.1 *Black circles* indicate possible ultrasound view points. These are just guides, and one may view a few interspaces left/right and cephalad/caudad from each point to obtain better views

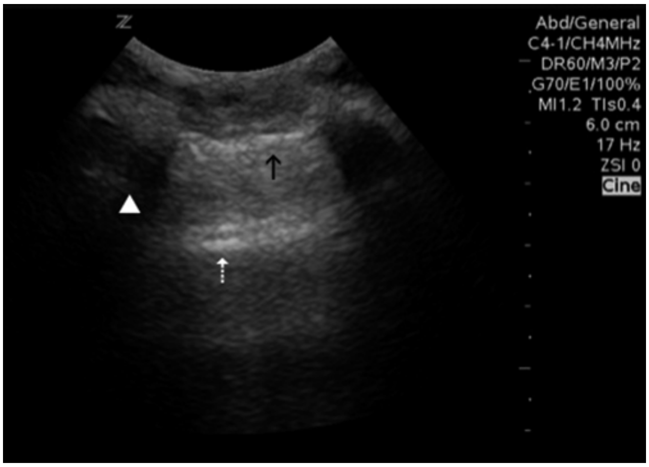
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G70/E1/100%
MI1.2 TIs0.4
6.0 cm
17 Hz
ZSI 0
Cine

Fig. 27.2 Typical bat sign with rib shadows (*arrowhead*) that surround the pleura (*black arrow*) and lung tissue. Incidentally, A lines (*white arrow*) are seen

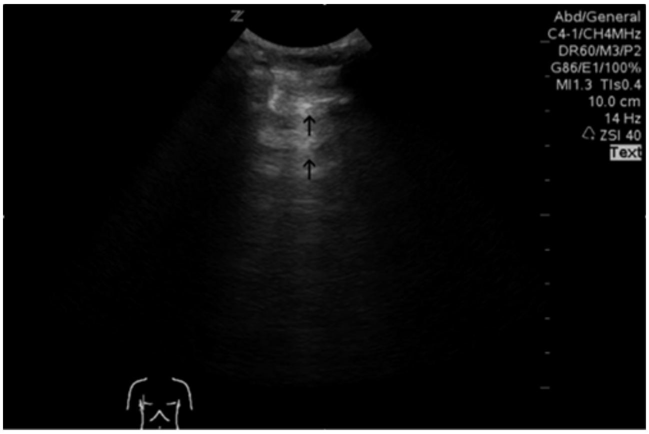
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G86/E1/100%
MI1.3 TIs0.4
10.0 cm
14 Hz
ZSI 40
Text

Fig. 27.3 Rib shadows can be seen, with multiple A lines (*arrows*)

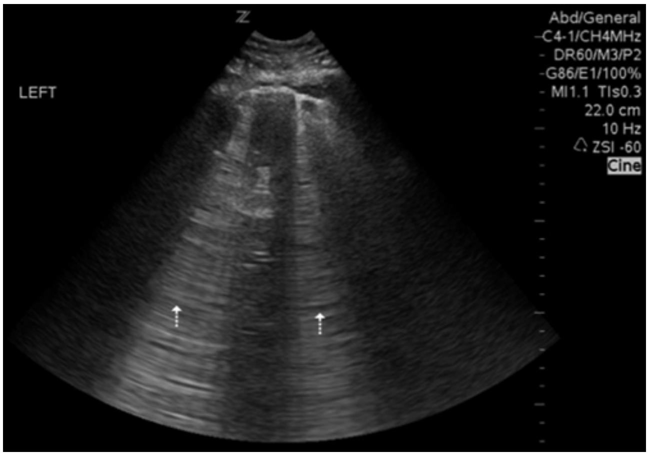
LEFT
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G86/E1/100%
MI1.1 TIs0.3
22.0 cm
10 Hz
ZSI -60
Cine

Fig. 27.4 Two B lines can be appreciated (*arrows*)

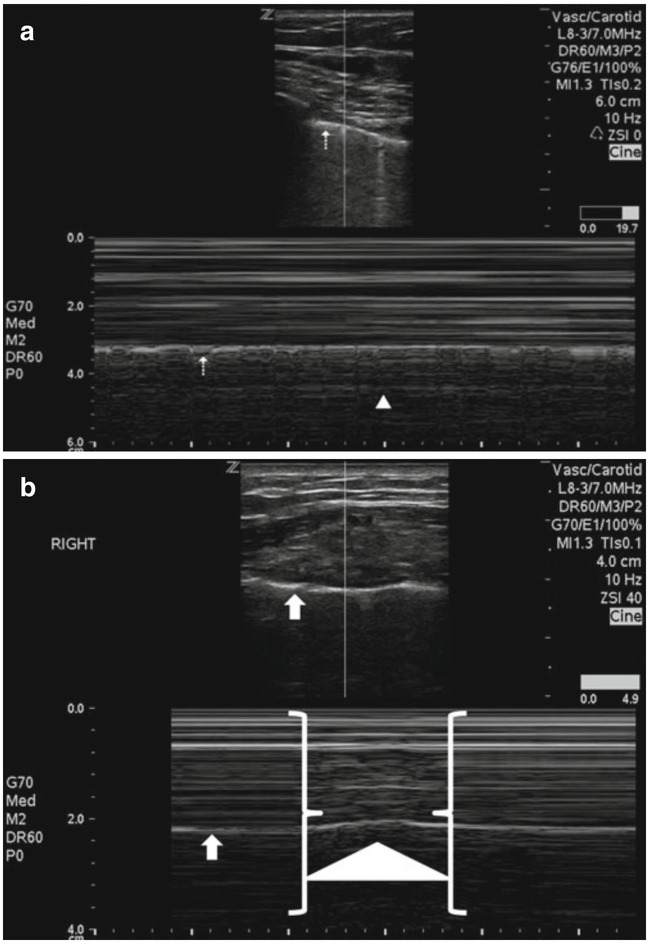
a
Vasc/Carotid
L8-3/7.0MHz
DR60/M3/P2
G76/E1/100%
MI1.3 TIs0.2
6.0 cm
10 Hz
ZSI 0
Cine
0.0
19.7
0.0
2.0
4.0
6.0
cm
G70
Med
M2
DR60
P0
b
RIGHT
Vasc/Carotid
L8-3/7.0MHz
DR60/M3/P2
G70/E1/100%
MI1.3 TIs0.1
4.0 cm
10 Hz
ZSI 40
Cine
0.0
4.9
0.0
2.0
4.0
cm
G70
Med
M2
DR60
P0

Fig. 27.5 (**a**) M-mode identifies the seashore sign, with the pleura (*arrows*) and the lung parenchyma (*arrowhead*) together producing an image similar to a seashore. (**b**) M-mode shows the stratosphere or barcode sign. *White arrows* indicate the pleura, and the *large triangle* shows an area of artifact from movement. Artifact from movement can be distinguished by a similar pattern change above and below the pleural line. On either side of the bracket is a continued horizontal line pattern typical of absent lung sliding

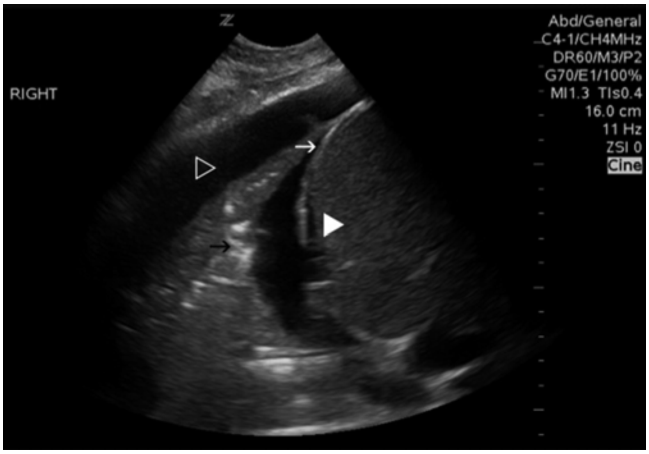
RIGHT
Abd/General
C4-1/CH4MHz
DR60/M3/P2
G70/E1/100%
MI1.3 TIs0.4
16.0 cm
11 Hz
ZSI 0
Cine

Fig. 27.6 Diaphragm (*white arrow*), liver (*white arrowhead*), lung (*black arrow*), and anechoic fluid, likely an effusion (*open arrowhead*). Incidentally, hyperechoic foci in the lung tissue, likely representing air bronchograms, can be seen

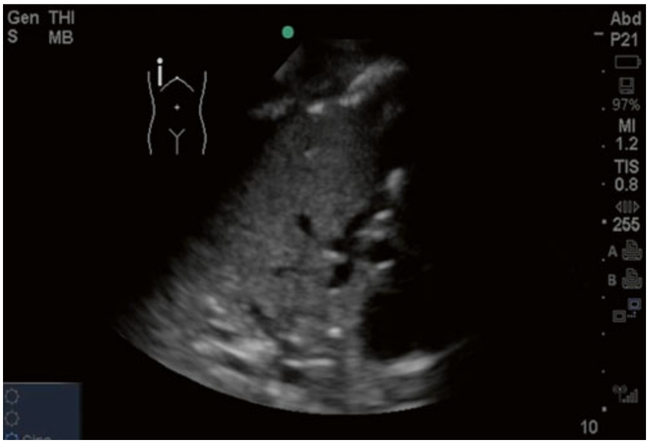
Gen THI
S MB
Abd
P21
97%
MI
1.2
TIS
0.8
255
A
B
10

Fig. 27.7 Lung here appears as tissue density, likely representing the alveolar consolidation pattern seen in pneumonia, and is termed "hepatization" and "tissue-like sign"

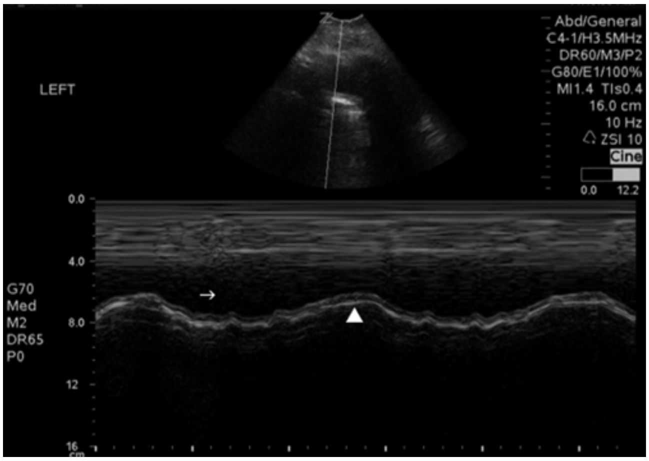
LEFT
Abd/General
C4-1/H3.5MHz
DR60/M3/P2
G80/E1/100%
MI1.4 TIs0.4
16.0 cm
10 Hz
ZSI 10
Cine
0.0 12.2
0.0
4.0
8.0
12
16
G70
Med
M2
DR65
P0

Fig. 27.8 In M-mode, identification of pleural effusion becomes evident with the sinusoidal sign (*arrowhead*) and identification of pleural fluid (*arrow*)

Selected Reading

Levitov A, Mayo PH, Slonim AD. Ultrasound evaluation of the lung. In: Critical care ultrasonography. New York: McGraw-Hill; 2009.

Lichtenstein DA, Menu Y. A bedside ultrasound sign ruling out pneumothorax in the critically ill. Lung sliding. Chest. 1995;108: 1345–8.

Lichtenstein DA, Meziere GA. A lung ultrasound sign allowing bedside distinction between pulmonary edema and COPD: the comet-tail artifact. Intensive Care Med. 1998;24:1331–4.

Lichtenstein DA, Meziere GA. Relevance of the lung ultrasound in the diagnosis of acute respiratory failure: the BLUE protocol. Chest. 2008;134:117–25.

Mayo PH, Doelken P. Pleural ultrasonography. Clin Chest Med. 2006;27:215–27.

Mayo PH, Goltz HR, Tafreshi M, Doelken P. Safety of ultrasound-guided thoracentesis in patients receiving mechanical ventilation. Chest. 2004;125:1059–62.

Part IV

Cardiac Procedures

Repair of Cardiac Injuries

28

Ronald Tesoriero

Whether performed by an emergency physician or a surgeon, the majority of cardiac repairs performed at emergency department thoracotomy will be temporary in nature and require further revision in the operating room. This chapter's main focus is the temporary control of cardiac injuries.

28.1 Indications

- Wounds to the heart in patients presenting with pulseless electrical activity (PEA) or asystole with evidence of cardiac tamponade
  - Penetrating wounds: <15 min of prehospital cardiopulmonary resuscitation (CPR)
  - Blunt wounds: <10 min of prehospital cardiopulmonary resuscitation (CPR)

28.2 Contraindications

- Absolute
  - Presenting rhythm of asystole and no evidence of pericardial tamponade on Focused Assessment with Sonography for Trauma (FAST)
- Relative
  - None

28.3 Materials and Medications

- Diagnostic ultrasound
- Emergency department thoracotomy tray (Fig. 28.1)
  - Sterile drapes, #10 scalpel, curved Mayo scissors, Finochietto retractor, Lebsche sternal knife and mallet, forceps, curved Metzenbaum scissors, surgical skin stapler, Foley catheter, clamps, needle driver, 2–0 and 3–0 polypropylene suture on MH or SH (noncutting) needle, Satinsky vascular clamps
- Betadine (povidone-iodine) or other skin antiseptic preparing solution
- Defibrillator and internal cardiac panels
- Epinephrine

R. Tesoriero, MD
Department of Surgical Critical Care,
University of Maryland School of Medicine,
R Adams Cowley Shock Trauma Center,
Baltimore, MD, USA
e-mail: rtesoriero@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_28

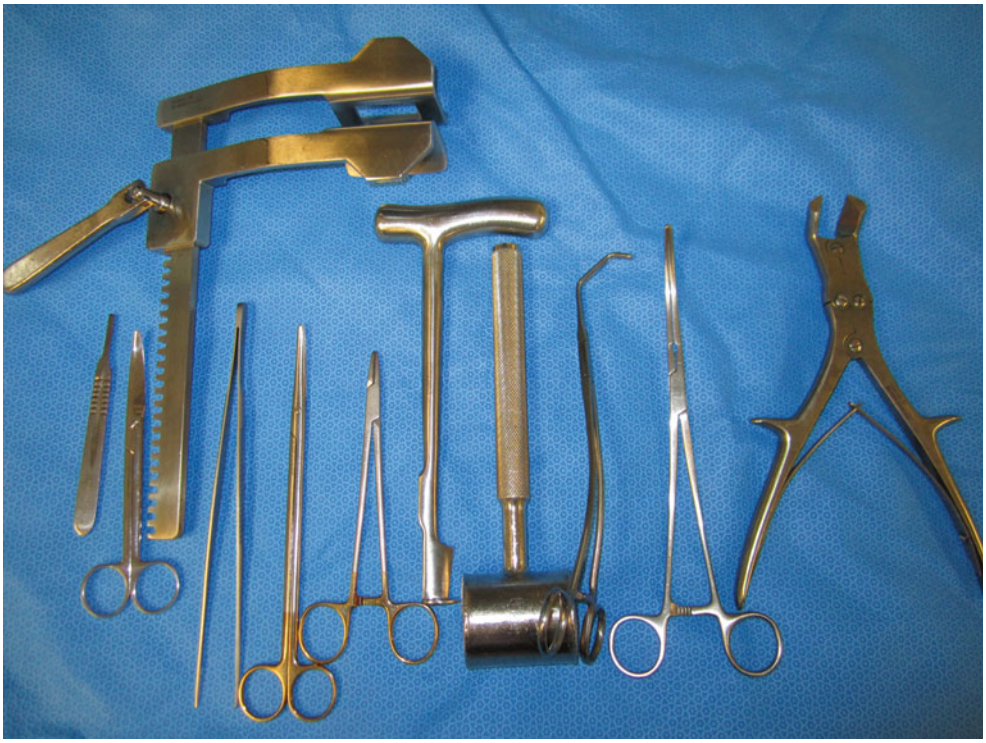

Fig. 28.1 Emergency department thoracotomy tray. *Left to right*: Scalpel, curved Mayo scissors, Finochietto rib spreader, DeBakey forceps, Metzenbaum scissors, needle driver, Lebsche knife and mallet, Satinsky vascular clamp, aortic clamp, bone cutter

28.4 Procedure

1. Identify the pericardium anterior to the phrenic nerve, pinch it between the fingers, and enter it by making a nick with the Metzenbaum scissors. Then open the pericardium longitudinally anterior to the phrenic nerve.
2. If the heart is not contracting effectively (PEA, asystole, fibrillation), immediately begin internal cardiac massage.
    - Unless the injury is so large that it cannot be controlled, taking time out at this point to repair the injury may lead to significant acidosis and inability to reestablish a perfusing rhythm.
3. While continuing internal cardiac massage, identify the area of injury and attempt to control it with manual pressure; if a larger wound, use a Foley catheter to prevent continued bleeding.
4. If necessary owing to the location of injury, extend the incision to a bilateral anterolateral thoracotomy and transect the sternum with a Lebsche knife and mallet or heavy shears.
5. If, after several minutes of internal massage and appropriate red blood cell and plasma transfusion, the patient remains in PEA or asystole, irrigate the heart with warmed saline and administer intravenous (IV) or intracardiac epinephrine.
6. If the heart enters ventricular fibrillation or ventricular tachycardia, cardiovert with internal paddles applied directly to the heart with an energy between 10 and 30 J.
7. Once a perfusing rhythm is reestablished, or if on initial evaluation the wound is so large that it cannot be controlled and will require immediate repair before internal massage can be effective, proceed to cardiac repair.
8. Choose the simplest method that will allow control of the injury until definitive repair can be performed in the operating room.
    (a) Injuries to the atrium: The pliable nature of the atria will often allow placement of a Satinsky clamp for control followed by repair with a running 3–0 polypropylene suture (Fig. 28.2).
    (b) Small injuries to the ventricles: Control the injury with direct manual pressure and close with either interrupted 3–0 polypropylene suture or a surgical skin stapler (Figs. 28.3 and 28.4).
    (c) Medium to large injuries to the ventricle: Attempt to control the wound by placing a Foley catheter through it, then blow up the balloon and apply gentle traction. Either staple or suture the wound closed before deflating the balloon and removing the catheter (Fig. 28.5).
        - Place an occluding clamp on the open end of the catheter or blood loss will continue through the catheter
        - Avoid excessive traction on the Foley because it will pull through the ventricle and make the hole larger
    (d) Extensive or inaccessible injury: Perform temporary inflow occlusion to the heart by manually compressing the right atrium against the heart so that it cannot fill. The heart will likely immediately enter PEA, fibrillation, or asystole giving the physician a couple of minutes to gain control of the injury before the patient becomes unrecoverable.
    (e) Injuries in proximity to coronary vessels: To avoid compression of the vessel, perform a horizontal mattress suture that passes beneath the artery. Teflon pledgets may assist in the repair (Fig. 28.6).

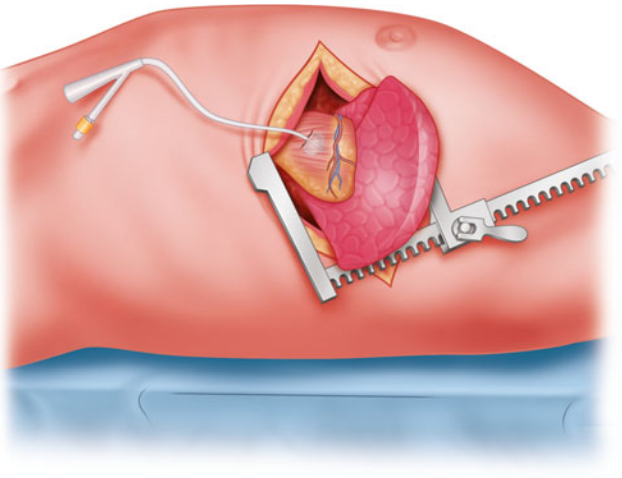

Fig. 28.2 Position the Finochietto retractor with the closed end toward the axilla. A Foley catheter may be used for initial control of large cardiac lacerations

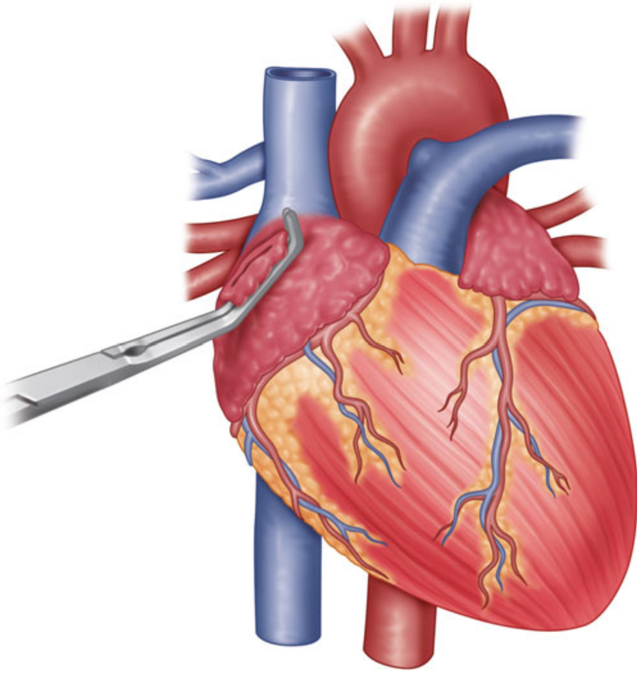

Fig. 28.3 Atrial injuries may be quickly controlled with a Satinsky clamp followed by repair with a 3–0 polypropylene suture

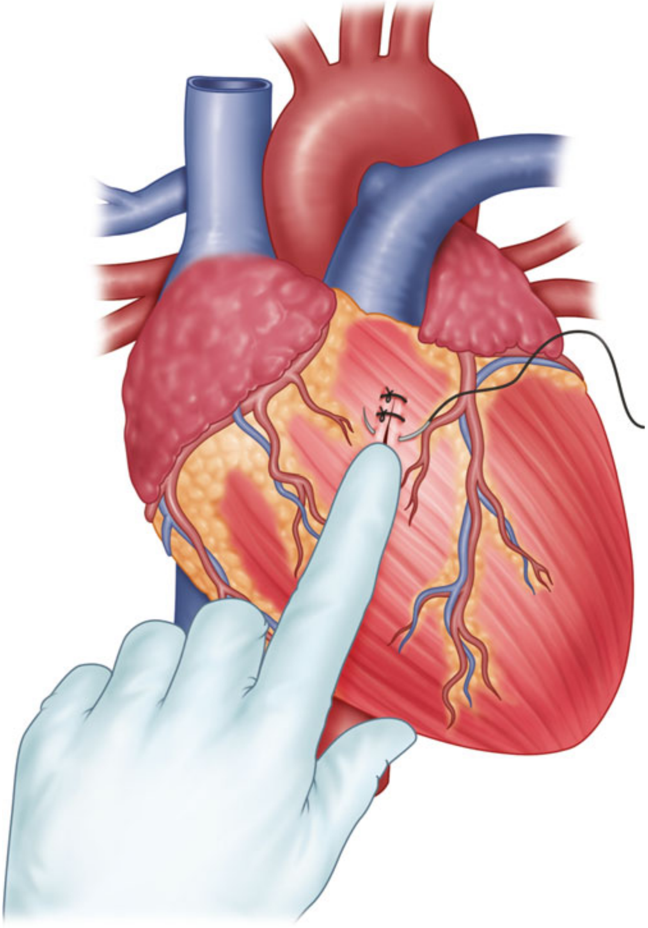

Fig. 28.4 Injuries may be controlled with direct manual pressure and closed with interrupted 3–0 polypropylene suture

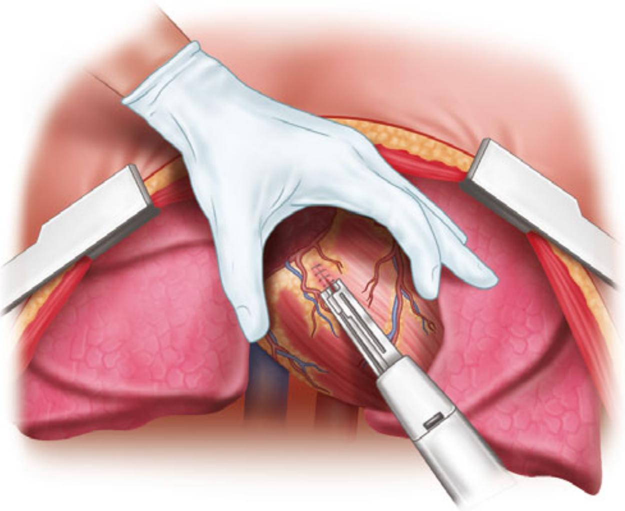

Fig. 28.5 A surgical skin stapler is a quick way to gain initial control of simple cardiac lacerations with minimal risk to the provider

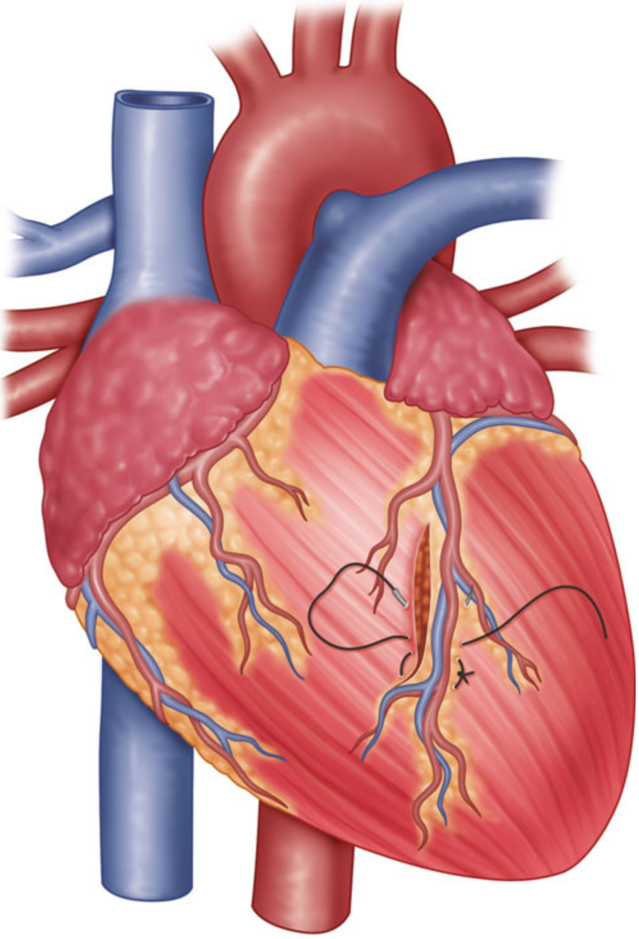

Fig. 28.6 Lacerations in close proximity to coronary vessels can be controlled with horizontal mattress sutures passed beneath the artery to avoid coronary arterial compression

28.5 Complications

- Cardiac and/or pulmonary laceration during entry into the chest
- Bleeding
- Delayed hemorrhage (failure to control internal mammary artery, disruption of sutures or staples)
- Infections: empyema and sternal infection
- Missed intracardiac valvular or septal injury (echocardiography should be performed after repair is complete)
- Air embolus

28.6 Pearls and Pitfalls

- Pearls
  - There may be more than one wound to the heart (especially with gunshot wounds). Look for them. However, if the wound is posterior and not bleeding with the heart in its natural position, it will be more prudent to leave the injury alone until the patient can be transported to the operating room. Elevating the heart can cause both inflow and outflow obstruction, leading to dysrhythmia that may be difficult to recover from.
  - To avoid sutures pulling through in patients with thin, edematous, or friable myocardium, consider horizontal mattress rather than simple sutures. These may be buttressed with Teflon pledgets for added security.
- Pitfalls
  - The myocardium tears easily. When tying sutures, take care to not tighten them too forcefully.

Selected Reading

Asensio JA, Trunkey DD. Current therapy of trauma and surgical critical care. Philadelphia: Mosby/Elsevier; 2008.

Feliciano DV, Mattox KL, Moore EE. Trauma. 6th ed. New York: McGraw-Hill Medical; 2008.

Hirshberg A, Mattox KL. Top knife: the art & craft of trauma surgery. Castle Hill Barns: TFM; 2005.

Moore EE, Knudson MM, Clay CB, et al. Defining limits of resuscitative emergency department thoracotomy: a contemporary Western Trauma Association perspective. J Trauma. 2011;70:334–9.

Wall MJ, Mattox KL, Chen CD, et al. Acute management of complex cardiac injuries. J Trauma. 1997;42:905–12.

Synchronized Electrical Cardioversion

29

Jason Jones, Ann Tsung, and Marie-Carmelle Elie

Tachycardia is defined as >100 beats per minute

29.1 Indications

- Non-sinus-rhythm tachycardias with a pulse including:
  - Atrial fibrillation
  - Atrial flutter
  - Monomorphic ventricular tachycardia (VT)
  - Refractory or unstable supraventricular tachycardia (SVT)
- Unstable signs and symptoms including acute coronary syndrome, decreased level of consciousness, chest pain, dyspnea, pulmonary edema, and hypotension

29.2 Contraindications

- Absolute
  - Ventricular fibrillation and pulseless or polymorphic (irregular) VT require unsynchronized electrical cardioversion (defibrillation), not synchronized cardioversion.
  - Known atrial thrombus.
  - Sinus tachycardia.
- Relative
  - Digitalis toxicity-related tachycardia
  - Atrial fibrillation of greater than 48 h duration without anticoagulation
  - Multifocal atrial tachycardia
  - Electrolyte abnormalities
  - Left atrial diameter greater than 4.5 cm
  - Patients with low probability of maintaining sinus rhythm and readily return to atrial fibrillation
  - Patients with sick sinus syndrome or sinoatrial blockage who will require a pacemaker for maintenance of stable rhythm

29.3 Materials and Medications

- Airway management equipment (laryngoscopes, endotracheal tubes)
- Cardiac monitoring, pulse oximetry, end-tidal CO_2 monitoring
- Cardioverter/defibrillator
- Sedation and analgesic medications

29.4 Procedure

1. Obtain a 12-lead electrocardiogram (ECG) and intravenous (IV) access.
2. If possible, correct underlying electrolyte abnormalities that may cause or contribute to the patient's arrhythmia.
3. Discuss risks, benefits, and alternatives (including pharmacological cardioversion) with the patient and obtain consent.
4. Prepare airway equipment and Advanced Cardiac Life Support (ACLS) code drugs.
5. Consider IV sedation (e.g., propofol, midazolam).
6. Provide IV analgesia (e.g., fentanyl, morphine).
7. Place defibrillator adhesive pads (8- to 12-cm diameter in adults) or paddles on the patient. Pediatric-sized pads/paddles should be used if the patient is less than 10 kg.

J. Jones, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: jasonjones@ufl.edu; elie@ufl.edu

A. Tsung, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: TSUNGANN@UFL.EDU

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_29

8. The first paddle/pad is placed to the right of the sternum at the second/third intercostal space. The second paddle/pad can be placed in one of two equally efficacious positions:
    (a) Anterolateral position—left fourth/fifth intercostal space in the midaxillary line (Fig. 29.1)
    (b) Anteroposterior position—between the spine and the edge of the left scapula (Fig. 29.2)
9. Turn the defibrillator/cardioverter into synchronized mode—marker above R-waves will be present (Fig. 29.3).
10. Select the energy level to be delivered based on the underlying rhythm
    (a) Regular VT (with pulses)—Adults: 100 J (monophasic or biphasic), 200 J for subsequent shocks
    (b) Atrial fibrillation—120–200 J (biphasic), 200 J (monophasic), 360 J for subsequent shocks
    (c) Atrial flutter and paroxysmal SVT—50–100 J (biphasic), 100 J for subsequent shocks
    (d) Pediatric dosage (regular and pulsed VT or SVT)—0.5–1 J/kg, up to 2 J/kg for subsequent shocks (Fig. 29.4)
11. Announce that you are going to deliver the shock on the count of three, and ensure that everyone is clear of the patient.
12. Deliver the shock by pressing button marked "SHOCK."
    - If using paddles, apply firm pressure and keep paddles in place until shock is delivered.
13. Reassess the patient's pulse and cardiac rhythm.
14. Repeat with escalating energy in a stepwise fashion if cardioversion is unsuccessful.

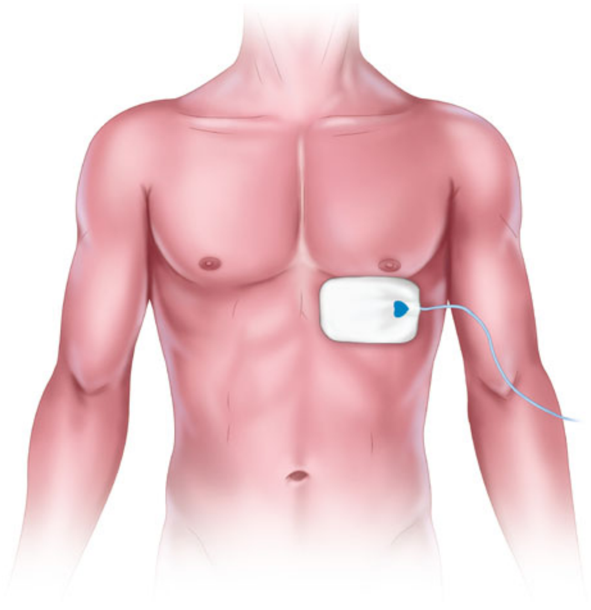

Fig. 29.1 Anterolateral pad placement

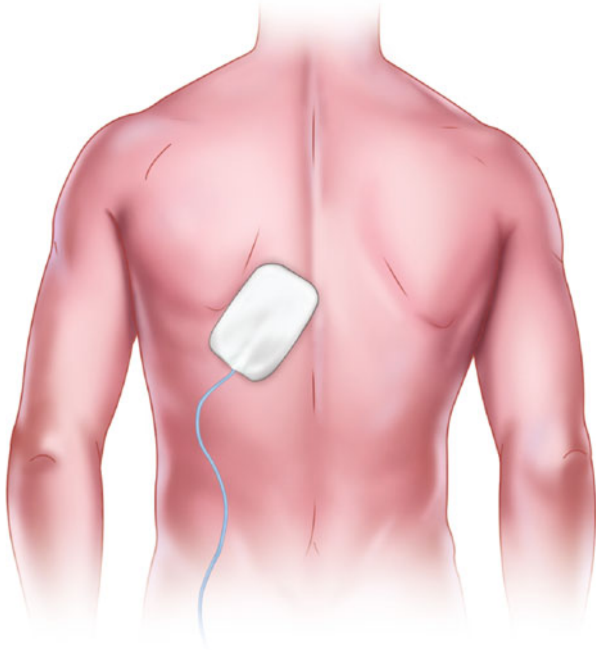

Fig. 29.2 Anteroposterior pad placement

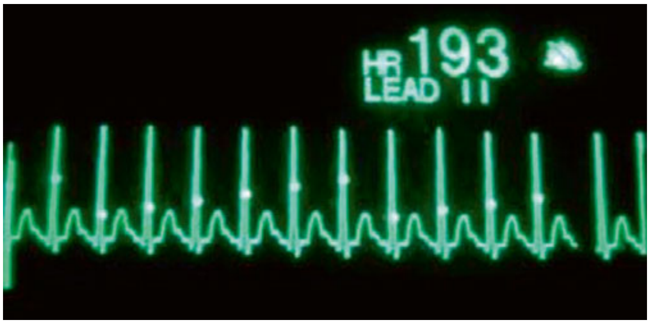

Fig. 29.3 Synchronized cardioversion—mark on R wave

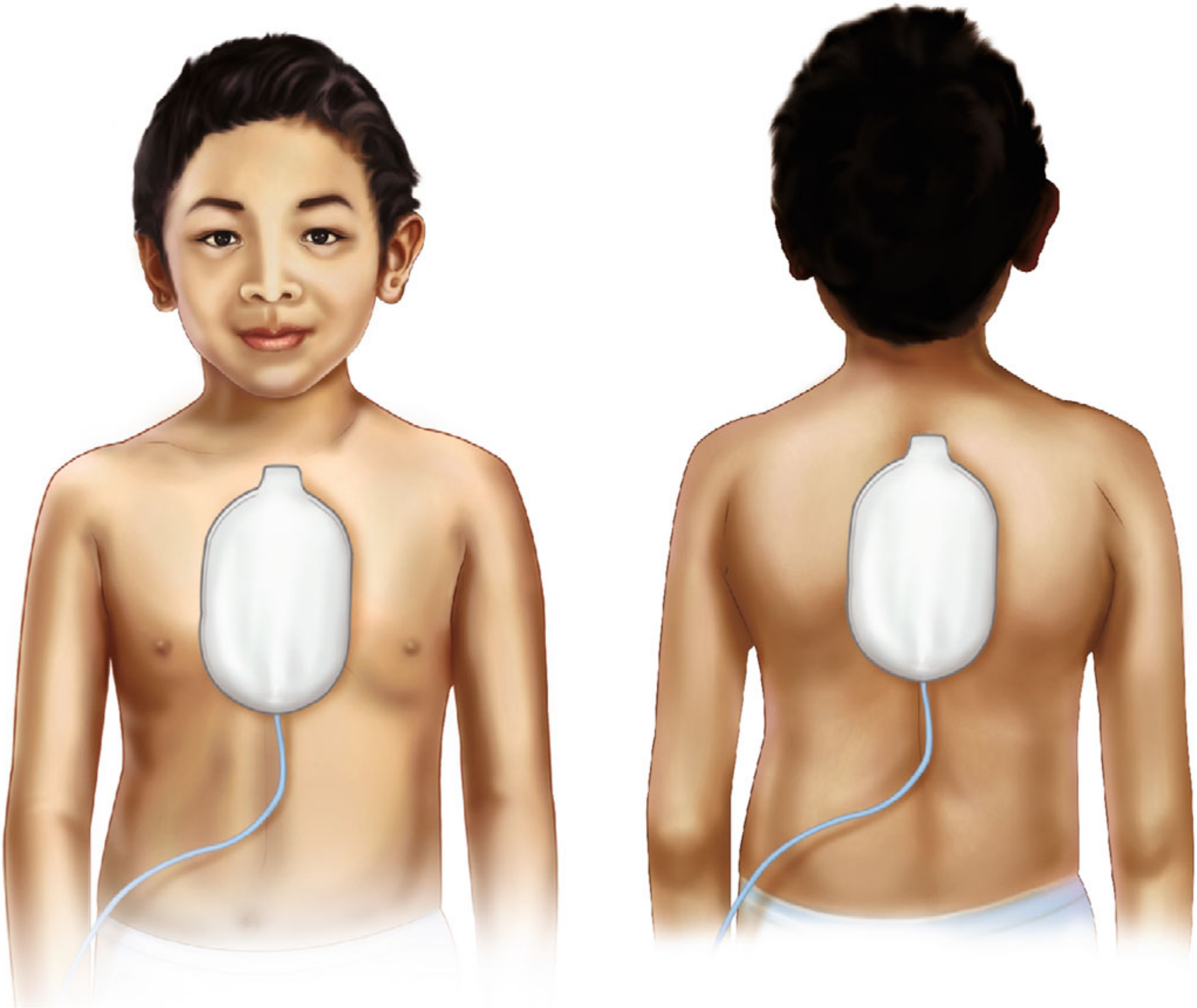

Fig. 29.4 Pediatric pad placement

29.5 Complications

- Superficial burns if there is inadequate gel.
- Induced arrhythmias (bradycardia in patients with previous inferior myocardial infarction, atrioventricular block, VT, ventricular fibrillation, asystole).
- Improperly synched cardioversion may rarely induce ventricular fibrillation.
  - Ectopy of the atria or ventricle in first 30 min after cardioversion
  - Atrial clot embolization in patients without adequate anticoagulation
- Apnea, hypoxia, hypercarbia, or hypotension may occur from sedation/analgesia.
- Medical professionals who incidentally touch the patient during shock delivery may be shocked or burned.
- Rarely, fire has occurred as a consequence of poor pad placement and a hyperoxygenated environment.

29.6 Pearls and Pitfalls

- Hirsute males should be shaved at the pad/paddle placement sites.
- Placing the cardioverter in "synchronized" mode avoids delivering a shock during the relative refractory segment, which could induce ventricular fibrillation.
- Cardioversion of pregnant patients is performed as in the general adult population.
- Keep pacemakers and implantable cardioverter-defibrillators at least 10 cm away from contact with paddles.
- If patient has implanted pacemaker, position pads so that they are not directly over the device.

Selected Reading

Gowda SA, Shah A, Steinberg JS. Cardioversion of atrial fibrillation. Progr Cardiovasc Dis. 2005;48:88–107.

Kleinman ME, Chameides L, Schexnayder SM, et al. Part 14: pediatric advanced life support. 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122 Suppl 3:S876–908.

Link MS, Atkins DL, Passman RS, et al. Part 6: electrical therapies: automated external defibrillators, defibrillation, cardioversion, and pacing. 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122 Suppl 3:S706–19.

Mayeaux EJ. The essentials guide to primary care procedures. Philadelphia: Lippincott Williams & Wilkins; 2012. p. 88–92.

Sirna SJ, Ferguson DW, Charbonnier F, et al. Factors affecting transthoracic impedance during electrical cardioversion. Am J Cardiol. 1988;62:1048–52.

Trohman RG, Parrillo JE. Direct current cardioversion: indications, techniques, and recent advances. Crit Care Med. 2000;28:170–3.

Unsynchronized Cardioversion (Defibrillation)

30

Matthew R. Tice, Zachary B. Kramer,
and Marie-Carmelle Elie

Unsynchronized cardioversion or defibrillation is the delivery of a *high-energy* shock as soon as the button is pushed on defibrillator. This means it can be delivered anywhere in the cardiac cycle. By contrast, synchronized cardioversion (see Chap. 29) delivers a low-energy shock at the peak of the R wave in the cardiac (QRS) cycle.

30.1 Indications

- Ventricular fibrillation (VF)
- Pulseless ventricular tachycardia (VT)
- Cardiac arrest due to or resulting from VF

30.2 Contraindications

- Absolute
  - Conscious patient
  - Presence of a pulse
  - Pulseless electrical activity (PEA)
  - Asystole
  - Multifocal atrial tachycardia
  - Defibrillation without knowing the rhythm
  - A second defibrillation before 2 min (or five cycles) of CPR
  - Advanced Directive, Physician Order for Life-Sustaining Treatment (POLST) indicating no cardiopulmonary resuscitation (CPR) or do not resuscitate (DNR)
- Relative
  - Potential electrical catastrophe (explosive environment [i.e., operating rooms])
  - Dysrhythmias due to enhanced automaticity such as in digitalis toxicity and catecholamine-induced arrhythmia (because mechanism of tachycardia remains after the shock)
- Factors that are not contraindications
  - Pregnancy.
  - Chest trauma.
  - Automatic implantable cardiac defibrillators (AICDs).
  - The patient is on a wet or moist surface.
  - Piercings on the chest.

30.3 Materials and Medications

- Electrocardiogram (ECG) monitor/defibrillator.
- Self-adhesive defibrillation pads or defibrillation paddles (paddles may be more successful than self-adhesive pads, but they have more complications and pose more danger to operators).
- Conductive gel for defibrillation paddles (not ultra sound gel).
- ECG electrodes.
- Supplemental oxygen.
- Intubation equipment as needed.

M.R. Tice, MD • Z.B. Kramer, MD • M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida, Gainesville, FL, USA
e-mail: mtice@ufl.edu; kramer@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_30

30.4 Procedure

Defibrillation is an emergent maneuver and, when necessary, should be promptly performed in conjunction with or before administration of induction or sedative agents to facilitate intubation.

1. Assess the ABCs (airway, breathing, circulation).
    - Open the airway with a head tilt/chin lift (or jaw thrust in a suspected traumatic patient). If the patient is apneic, provide breaths with a bag-valve-mask (BVM) and observe chest rise.
    - Check for pulses. If absent, start CPR.
    - CPR should be initiated before any shock while getting all equipment ready for at least 2 min to provide adequate circulation to the brain and heart.
2. Wipe off the patient's chest if moist or wet.
3. Remove transdermal patches, jewelry, and piercings if possible.
4. Attach ECG electrodes to the patient.
    - The self-adhesive defibrillation pads or defibrillation paddles can be used as ECG electrodes to access the rhythm.
5. Paddles: With conductive gel applied to the metal surface, place one paddle on the patient's right chest, just below the clavicle, near the sternal border. The other should be on the left chest, midaxillary line above the fifth or sixth intercostal space (Fig. 30.1).
    - The long axis of the paddles should be perpendicular to the ribs to allow for better transduction of current through the chest.
    - Pads: Same placement as paddles except that pads can be placed in any orientation as long as they are in full contact with the chest.
    - If a lot of breast tissue is present, push the tissue to one side or lift it away and place the paddles or pads underneath.
    - An error in pad or paddle placement can distort the rhythm into looking like a rhythm that does not require defibrillation.
6. Place the ECG monitor/defibrillator into a mode to acquire a rhythm from the pads or paddles.
    - Steps 1–3 can be done simultaneously.
7. Stop CPR and assess the rhythm and pulse for no longer than 10 s.
8. If VF or pulseless VT is observed, then switch the defibrillator to charge mode.
9. Charge to 200 J.
    - Continue CPR while the defibrillator is being charged.
10. When the defibrillator indicates that it is charged, clearly order that everyone stop touching the patient and observe that there is no physical contact before defibrillating the patient.
    - If using the paddles, apply extra force to the chest through the paddles to deflate the lungs to allow for better defibrillation.
    - The operator should observe a muscle twitch during defibrillation.
11. Restart CPR for 2 min or five cycles.
    - Another operator may charge (but not fire) the defibrillator while CPR is being performed to expedite the time between pulse/rhythm check and the initiation of a shock.
12. After 2 min or five cycles of CPR, assess the rhythm and pulse and repeat steps 6–8 and give appropriate Advanced Cardiac Life Support (ACLS) medications.
13. If successful return of spontaneous circulation (ROSC) occurs, initiate the hypothermia protocol per hospital guidelines.

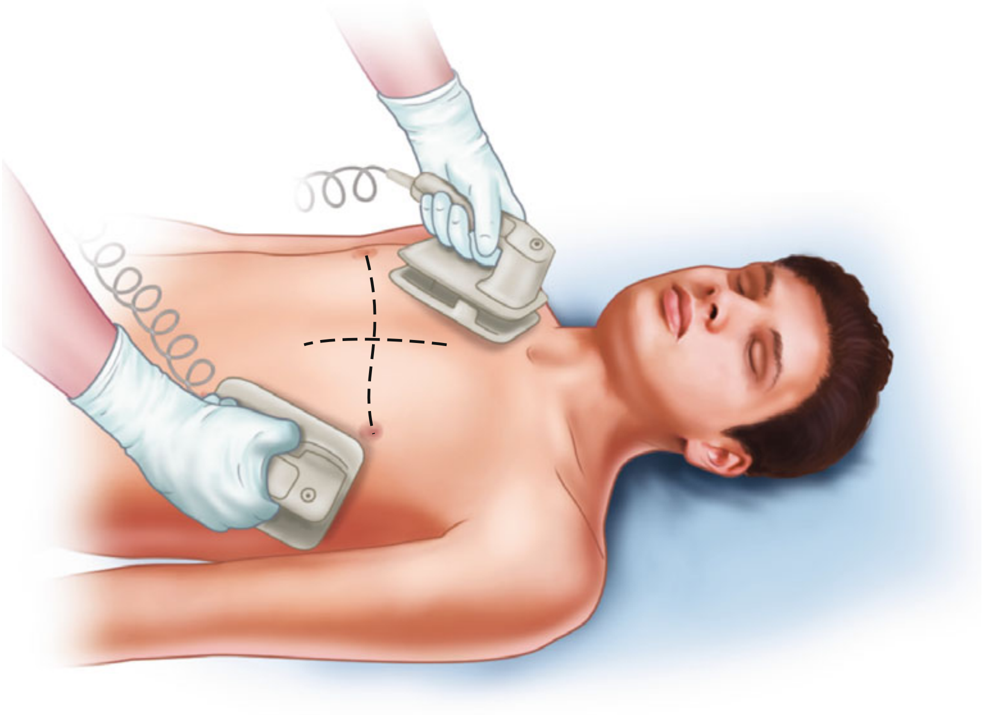

Fig. 30.1 Proper placement of defibrillation paddles

30.5 Complications

- Skin burns (most common and likely due to improper technique).
- Injury to cardiac tissue (myocardial necrosis secondary to burn).
  - ST segment elevation that lasts longer than 2 min usually indicates myocardial injury unrelated to the shock.
- Abnormal heart rhythms (usually benign like atrial, ventricular, and junctional premature beats).

30.6 Pearls and Pitfalls

- Pearls
  - Post-defibrillation cardiac dysrhythmias are more common following prolonged VF and higher-energy-level countershocks. Early defibrillation at the recommended energy levels minimizes this complication. Follow ACLS protocols to manage the resulting dysrhythmias.
  - The presence of liquid (body fluids, medications, or intravenous fluids) may cause electrical arcing thermal burns to the skin and soft tissue and produce ineffective defibrillation by allowing the current to pass across the trunk rather than transthoracically. To minimize this potential complication, ensure that any body fluids or liquids are wiped away from the skin before defibrillation attempts.
- Pitfalls
  - Myocardial and epicardial injury (not direct thermal injury) may result from the electrical current applied in defibrillation. Use the minimal recommended energy levels to minimize injury.
  - Electrical injuries to health-care providers can result if participants remain in contact with the patient during delivery of a countershock because they can serve as a ground for the current discharged.
    - This can be minimized by wearing gloves and using biphasic machines and electrode pads.
  - Fires that can result from sparks in the presence of nitroglycerin patches or ointment, flammable gases, or an oxygen-rich environment can also be a source of injury to the patient or health-care providers. Ensure "all clear" from the patient before delivery of shock to avoid these complications.

Selected Reading

Advance cardiovascular life support: provider manual. Dallas: American Heart Association; 2011.

Deakin CD, Sado DM, Petley GW, Clewlow F. Is the orientation of the apical defibrillation paddle of importance during manual external defibrillation? Resuscitation. 2003;56:15–8.

Pagan-Carlo LA, Spencer KT, Robertson CE, Dengler A, Birkitt C, Kerber RE. Transthoracic defibrillation: importance of avoiding electrode placement directly on the female breast. J Am Coll Cardiol. 1996;27:449–52.

Walker C, Callies F, Langenfeld H. Adverse effects of direct current cardioversion on cardiac pacemakers and electrodes: is external cardioversion contraindicated in patients with permanent pacing systems? Eur Soc Cardiol. 2004;6:165–8.

Transcutaneous Pacing

31

Nour Rifai and Christian Coletti

31.1 Indications

- Hemodynamically unstable (i.e., hypotension, pulmonary edema, chest pain, shortness of breath, or evidence of decreased cerebral perfusion) bradyarrhythmias refractory to medical therapies
- As a bridge to a transvenous or permanent pacemaker
- As an overdrive pacer in tachyarrhythmias
- Controversially, within the first 10 min of a witnessed asystolic cardiac arrest
- In children only with bradycardia associated with a known congenital cardiac defect or after cardiac surgery

31.2 Contraindications

- Absolute
  - None
- Relative
  - Bradyarrhythmia associated with hypothermia (ventricles are more prone to defibrillation-resistant fibrillation)
  - Prolonged cardiac arrest (>20 min)
  - Bradyarrhythmia in children (usually secondary to hypoxia or a respiratory issue)
  - Patient is unable to tolerate the procedure despite sedation and analgesia

31.3 Materials and Medications

- Pacemaker device (modern units offer combined pacer and defibrillator functions) (Fig. 31.1)
- One set of standard electrocardiogram (ECG) electrodes
- One set of pacer pads
- Code cart and airway equipment (prophylactically)
- Sedation and analgesia (typically a short-acting benzodiazepine and an opioid)
  - Midazolam: 0.2–0.10 mg/kg intravenous (IV) push and may repeat with 25 % of initial dose after 3–5 min. Do not exceed 2.5 mg/dose or a cumulative dose of 5 mg.
  - Fentanyl: 1–2 mcg/kg IV slow push over 1–2 min, may repeat dose in 30 min. (Fentanyl is the opioid of choice because it is less likely to exacerbate any hypotension.)

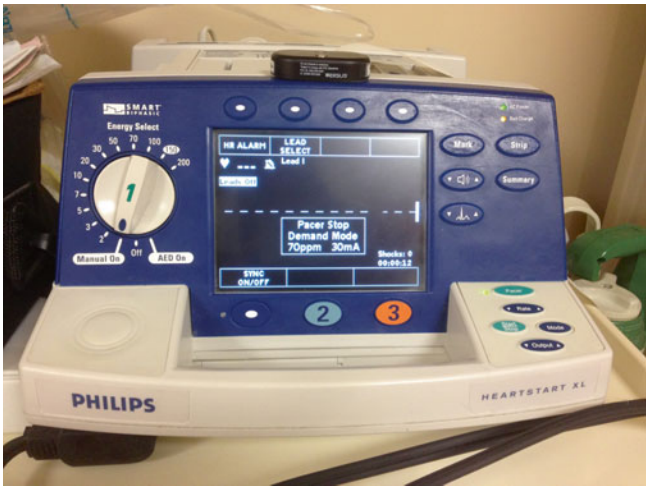

Fig. 31.1 Pacemaker device

N. Rifai, MBCHB • C. Coletti, MD (✉)
Department of Emergency Medicine,
Christiana Care Health System,
Newark, DE, USA
e-mail: nourifai@gmail.com; chris.coletti@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_31

31.4 Procedure

1. Time permitting, clean and dry the skin, and shave any excess hair off the chest.
2. Administer any appropriate sedation and analgesia.
3. Attach the ECG electrodes to both the input port of the pacemaker unit and the patient. On the patient, the white lead is placed just above the right clavicle, the black lead is just above the left clavicle, and the red lead is around the left midaxillary line.
4. Attach the pacer pads either in the anteroposterior or anterolateral positions as pictured (avoid placement over an implanted pacemaker or defibrillator; Fig. 31.2).
5. Turn the machine on and switch it to synchronous (or on-demand) mode.
   - Asynchronous (or fixed) mode fires impulses with no regard to the intrinsic cardiac cycle, increasing the likelihood of an R on T phenomenon, which could result in ventricular tachycardia or fibrillation.
   - Synchronous (or on-demand) mode will not fire an electrical impulse when a QRS complex is sensed; this is the preferred mode for transcutaneous pacing.
6. Set the desired heart rate: typically 60–80 beats/min to achieve adequate perfusion.
7. Select a lead on the pacemaker unit and then press Start.
8. Slowly increase the output current until electrical capture is denoted by a visible pacemaker spike, which will precede every QRS complex on the ECG monitor (Fig. 31.3).
   - Electrical capture is usually achieved between 50 and 100 mA.
   - If a patient is unconscious or truly deteriorating quickly or in cardiac arrest, it may be prudent to set the initial currents at maximum to ensure rapid capture and then decrease the current to just above that at which electrical capture was achieved.
9. After electrical capture is appreciated on the monitor, assess for mechanical capture by palpation of a pulse at a rate that corresponds to that which the machine is set at. An improved blood pressure or resolution of chest pain, shortness of breath, or altered mental status also suggests that the heart rate has improved and perfusion has been restored.
10. When pacing in overdrive for tachyarrhythmias, the pacer rate is set 20–60 beats/min faster than the detected tachycardic rate.
    - Bear in mind that rhythm acceleration or the induction of ventricular fibrillation is a possibility with pacing, hence the recommendation of having a code cart and airway equipment in the room at all times.

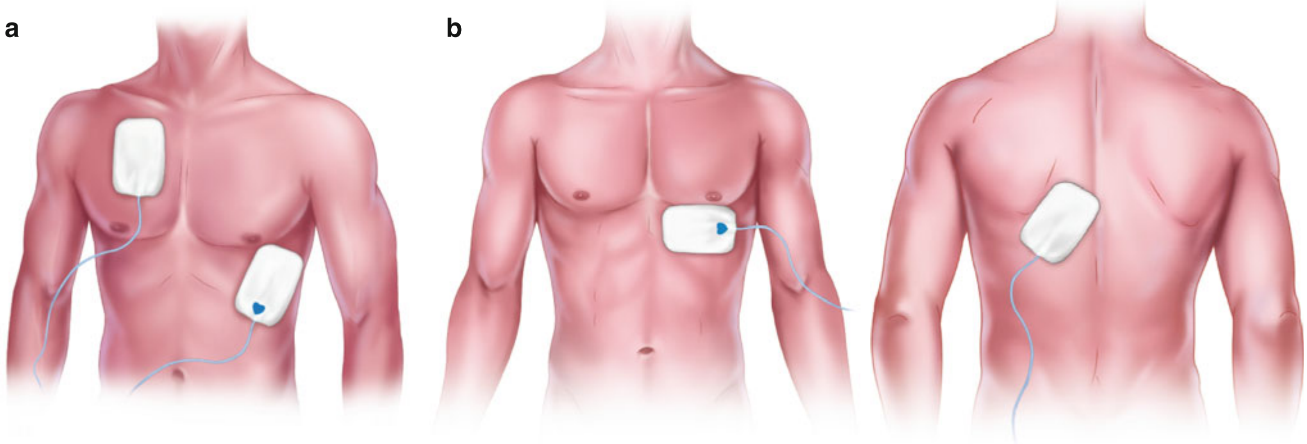

Fig. 31.2 Pacer pads attached in the anterolateral (**a**) or anteroposterior positions (**b**)

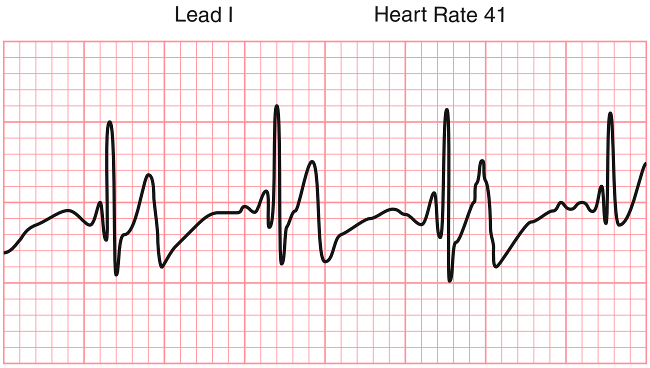
Lead I
Heart Rate 41

Bradycardia prepacing attempt

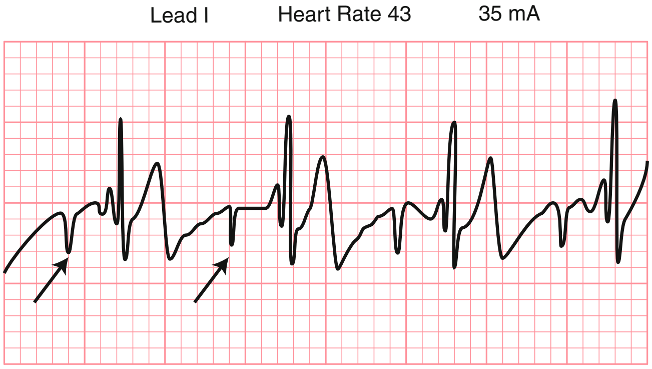
Lead I
Heart Rate 43
35 mA

Pacing attempted, note pacing stimulus indicator (arrow) which is below threshold, no capture

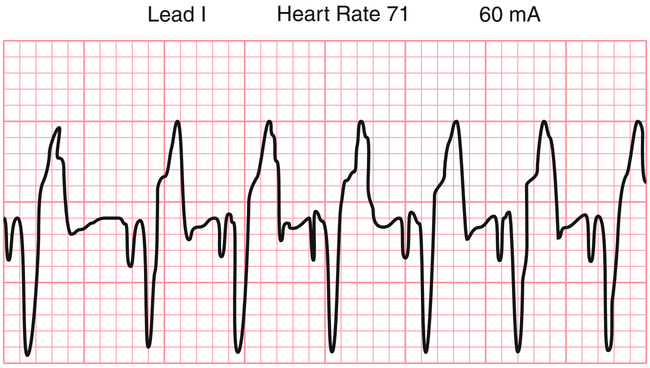
Lead I
Heart Rate 71
60 mA

Pacing above threshold (60mA), with capture (QRS complex broad and ventricular, T wave opposite QRS)

Fig. 31.3 A visible pacemaker spike will precede every QRS complex on the ECG monitor

31.5 Pearls and Pitfalls

- Pearls
  - Pacer pads can be prophylactically placed on bradycardic but stable patients in preparation for potential decompensation.
  - It may be prudent to discuss with a stable but bradycardic patient the option of trying the pacemaker not only to ensure that the machine and pads are working but also to gauge an idea of the output current that will be needed to electrically capture for quicker, more efficient, and potentially lifesaving pacing that may prove necessary.
  - When palpating for mechanical capture, the femoral pulse may be easier because carotid palpation may prove difficult with muscular contractions induced by the pacer impulses.
  - If impulses are not capturing, try slightly readjusting the pacer pads, making sure the leads are still on, the settings were not changed, the machine's battery did not die, or that the machine is plugged in.
  - If electrical capture is achieved but mechanical is not, increase the rate until adequate pulses are palpated.
  - Be careful to not miss an underlying ventricular fibrillation if the monitor is not blanked or dampened by an ECG screen.
  - Cardiopulmonary resuscitation (CPR) can be continued even while transcutaneous pacing is taking place because the pacer pads are insulated and the power delivered in each impulse is minimal, making the risk of injury to health-care workers very low.
- Pitfalls
  - Certain factors including a large body habitus, the presence of large pericardial effusions, scarring secondary to intrathoracic surgeries, or large amounts of intrathoracic air associated with obstructive pulmonary diseases may not only increase the threshold for capture but also may even cause a failure to capture, in which case transvenous pacing should immediately be considered.
  - Be aware that long-term pacing and pacing in children increases the likelihood for cutaneous and soft tissue damage.

Selected Reading

Bonow JS, Mann DL, Zipes DF, editors. Braunwald's heart disease: a textbook of cardiovascular medicine. 9th ed. Philadelphia: Saunders Elsevier; 2008.

Pfenninger JL, Fowler GC, editors. Pfenninger and Fowler's procedures for primary care. Philadelphia: Saunders Elsevier; 2010.

Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2010.

Tintinalli JE, Stapczynski JS, Cline DM, Ma OJ, Cydulka RK, Meckler GD, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw-Hill Medical; 2011.

Transvenous Cardiac Pacing

32

Katrina John, Jeffrey Kile, and Amish Aghera

32.1 Indications

Hemodynamic compromise in the presence of:

- Sinus node dysfunction
- Second- and third-degree heart block
- Atrial fibrillation with slow ventricular response
- New left bundle branch block (LBBB)
- Right bundle branch block (RBBB)
- Bifascicular block
- Alternating bundle branch block (BBB)
- Implanted pacemaker malfunction

32.2 Contraindications

- Absolute
  - Prosthetic tricuspid valve
- Relative
  - Bradycardia in the presence of severe hypothermia

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_32

32.3 Materials and Medications (Fig. 32.1)

- Sterile gloves
- Sterile gown and drapes
- Face mask and surgical cap
- Two 10-mL syringes
- One 3-mL syringe
- Local anesthetic (1 or 2 % lidocaine without epinephrine)
- 22-gauge needle
- Povidone-iodine or chlorhexidine/isopropyl alcohol
- Several 4 × 4-in. gauze sponges
- Sterile dressing pack
- 4-0 nylon or silk sutures
- Scalpel (#11 blade)
- Needle holder
- Scissors
- Introducer needle
- Guidewire
- Dilator
- Introducer sheath
- Collapsed sterile extension sheath
- Balloon-tipped pacing catheter
- Pacing generator with working battery
- Insulated cable with alligator clamps or other suitable connectors
- Electrocardiogram (ECG)/cardiac monitor

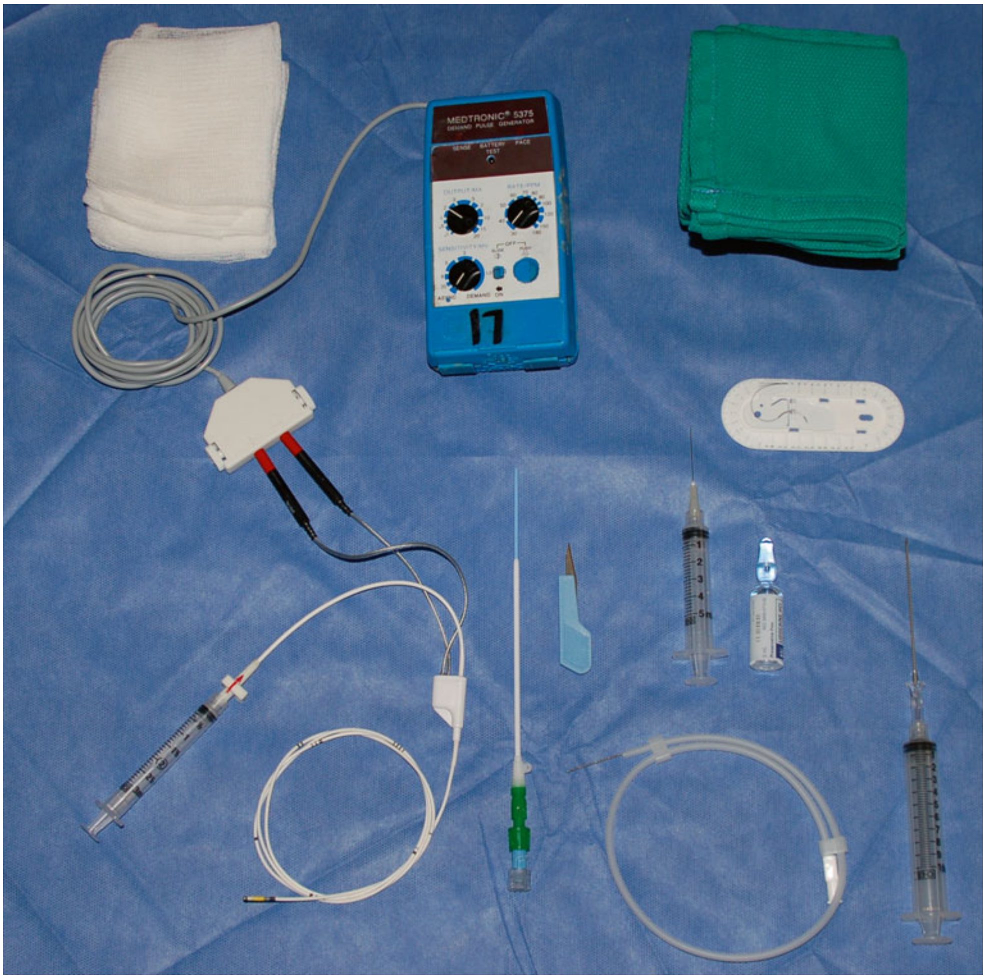

Fig. 32.1 Materials and medications

32.4 Procedure

1. Select appropriate catheterization site.
    - Four sites provide superior access to right atrium: brachial, femoral, internal jugular, and subclavian veins. The right internal jugular and left subclavian veins, respectively, are the most direct routes and, thus, the most commonly used in the emergency department setting.
2. If clinical situation permits, explain the procedure fully to the patient and obtain consent.
3. Sterilize the skin and apply sterile drapes.
4. Anesthetize the intended insertion site with lidocaine.
5. Prepare the pacing wire by inflating and deflating the balloon with 1–1.5 mL of air using a 3-mL syringe, and confirm that locking lever functions correctly to maintain balloon inflation (Fig. 32.2).
6. Using ultrasound guidance by means of a probe covered in a sterile sheath, insert the introducer needle into the vein while applying negative pressure to the attached 10-mL syringe (as with standard central line placement) (Fig. 32.3).
7. When flashback of blood is visualized in the syringe, remove the syringe, stabilizing the introducer needle firmly in place.
8. Pass the guidewire through the needle to a depth of 10–15 cm.
9. Holding the guidewire securely in place, remove the introducing needle.
10. Using a scalpel, make a small incision (approximately the width of the catheter to be used) through the dermis at the insertion site of the guidewire.
    - Avoid severing the guidewire by facing the sharp edge of the scalpel *away* from the guidewire.
11. Pass a dilator over the guidewire to make a tract in the skin and then remove the dilator.
12. Pass the introducer catheter over the guidewire until the hub is in contact with the skin and then remove the guidewire (Fig. 32.4).
13. Attach the collapsible sterile sheath to the hub of the introducer catheter.
14. Attach the positive and negative connectors of the pacing wire to their respective terminals on the cable(s) connected to the generator.
15. Set generator rate at 80 beats/min, output to 5 mA, and sensitivity to 3 mV.
16. Insert the free end of the pacing wire into introducer catheter with the balloon deflated and slowly advance the wire.
17. When the tip of the wire is within the superior vena cava, inflate the balloon, lock the valve to ensure that the balloon remains inflated, and continue to advance the pacing wire.
18. Closely watch the cardiac monitor for signs of capture (i.e., a wide QRS preceded by a pacing spike).
    - Markings on the pacing wire indicate its depth of insertion. At 20 cm, the wire should be in the right ventricle. If there is no capture observed by 25 cm, then the wire should be slowly withdrawn and advanced again.
19. Once capture is achieved, deflate the balloon.
20. Advance the wire a further 1–2 cm.
21. Coil the wire that remains outside the extended sterile sheath and suture it loosely but securely to the skin.
22. Suture the introducer hub to the skin.
23. Reduce generator output to zero and then increase it slowly to determine the minimum pacing threshold (i.e., the minimum voltage at which capture is achieved) and then set the output to twice this value.
24. Set the generator rate.
25. Stabilize the generator near the catheterization site.
    - For example, the generator can be attached to an intravenous (IV) fluid stand to the right of the patient's head if the pacing wire has been placed via the patient's right internal jugular vein.
26. Obtain a chest x-ray to confirm pacing wire position and absence of potential complications (e.g., pneumothorax) (Fig. 32.5).
27. Obtain a 12-lead ECG (Fig. 32.6).

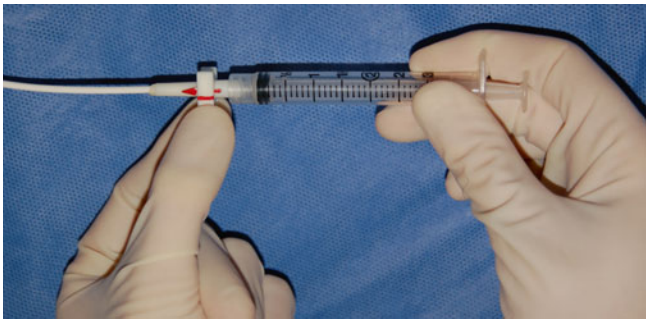

Fig. 32.2 Pacing wire balloon locking lever

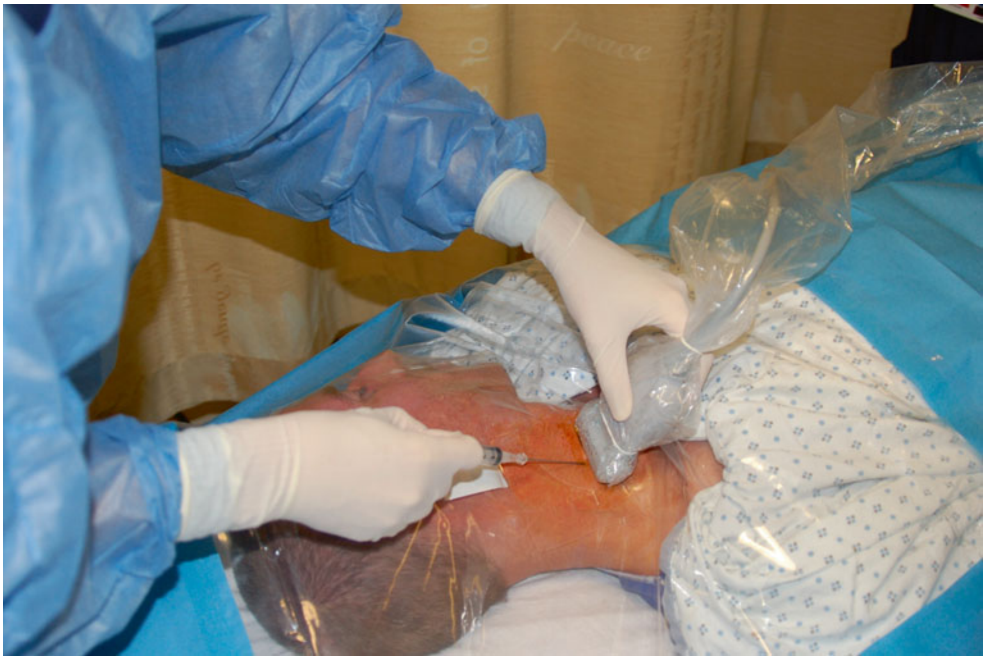

Fig. 32.3 Insertion of introducer needle into internal jugular vein with ultrasound guidance

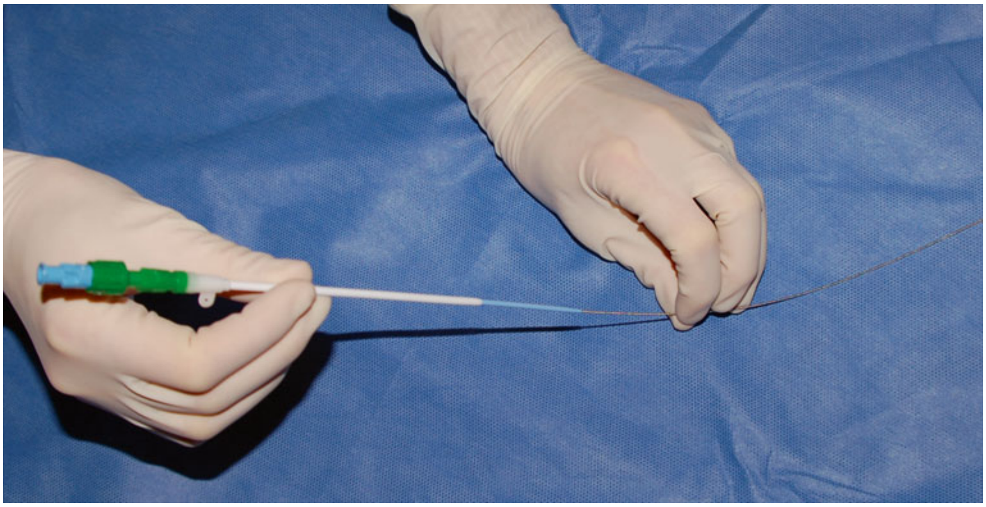

Fig. 32.4 Placement of introducer catheter over guidewire

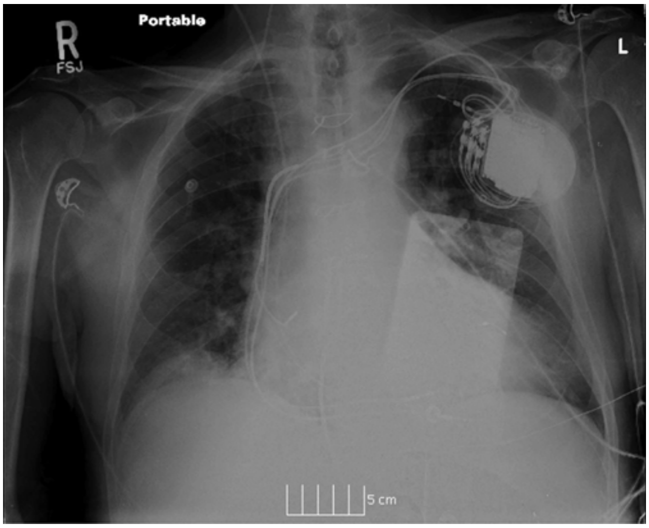
Portable
R
FSJ
L
5 cm

Fig. 32.5 X-ray displaying correct position of pacing wire in the right ventricle

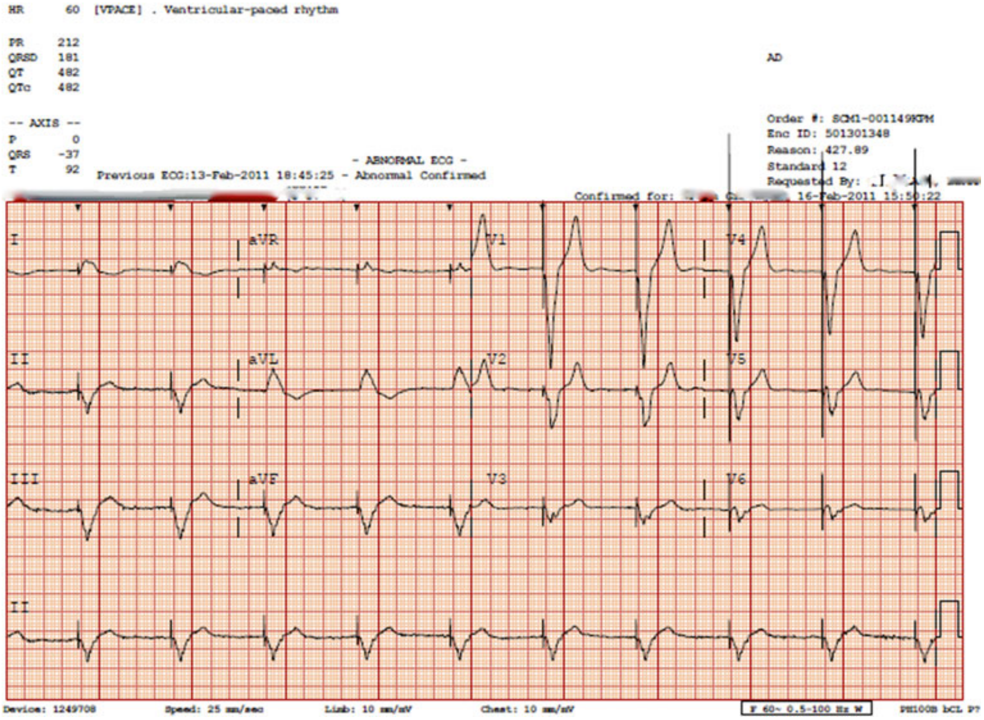
HR 60 [VPACE] . Ventricular-paced rhythm
PR 212
QRSD 181
QT 482
QTc 482
-- AXIS --
P 0
QRS -37
T 92
- ABNORMAL ECG -
Previous ECG:13-Feb-2011 18:45:25 - Abnormal Confirmed
AD
Order #: SCM1-001149KPM
Enc ID: 501301348
Reason: 427.89
Standard 12
Requested By: [redacted]
Confirmed for: [redacted] 16-Feb-2011 15:50:22
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
Device: 1249708
Speed: 25 mm/sec
Limb: 10 mm/mV
Chest: 10 mm/mV
F 60~ 0.5-100 Hz W
PH100B bCL P

Fig. 32.6 Cardiac monitor displaying wide-complex QRS and pacing spike

32.5 Complications

- Arterial puncture
- Venous thrombosis
- Thrombophlebitis
- Pulmonary embolus
- Pneumothorax
- Fracture of guidewire with embolization
- Hemothorax
- Thoracic duct laceration
- Indwelling line infection (local or systemic)
- Dysrhythmias/premature ventricular contractions (PVCs)
- Insertion of pacing wire into coronary sinus/pulmonary artery
- Left (rather than right) ventricular pacing through atrial septal defect/ventricular septal defect (ASD/VSD)
- Septal perforation
- Ventricular perforation
- Entrapment/twisting of preexisting permanent pacing wires by the temporary pacing wire
- Balloon rupture
- Chordae tendineae rupture
- Pacing wire fractures
- Loss of capture owing to wire displacement or fracture
- Generator failure

32.6 Pearls and Pitfalls

- Confirm all necessary equipment is present and functional before beginning the procedure. If a dedicated transvenous pacing set is unavailable, individual components required (e.g., pacing catheter, connector cables, introducer catheter, guidewire) may need to be "cannibalized" from several different procedure kits in an emergency department.
- Become familiar with the locking valve for the pacing wire balloon before beginning the procedure. This small part is essential for inflating and deflating the balloon but is frequently difficult to adjust owing to stiffness (and all the more so if covered in blood).
- The ideal position for the pacing wire tip is against the diaphragmatic aspect of the right ventricle between its midpoint and its apex. This position is confirmed by development of an LBBB pattern with left axis deviation on the cardiac monitor. If difficulty is encountered achieving this position, the tip of the pacing wire can be moved to the right ventricular outflow tract, where the wire is less stable, but adequate pacing can still be achieved. In this latter position, the threshold will need to be kept relatively high and the monitor should display an LBBB pattern with an inferior axis.
- The use of balloon-tipped catheters enables the catheter tip to be directed with blood flow and its position (and successful capture) to be confirmed with a cardiac monitor. Nonballoon-tipped catheters, however, have typically required insertion with guidance of either constant ECG monitoring or fluoroscopy.

Selected Reading

Aguilera PA, Durham BA, Riley DA. Emergency transvenous cardiac pacing placement using ultrasound guidance. Ann Emerg Med. 2000;36(3):224–7.

Birkhahn RH, Gaeta TJ, Tloczkowski J, Mundy T, Sharma M, Bove J, Briggs WM. Emergency medicine-trained physicians are proficient in the insertion of transvenous pacemakers. Ann Emerg Med. 2004;43(4):469–74.

Dalsey WC, Syverud SA, Hedges JR. Emergency department use of transcutaneous pacing for cardiac arrests. Crit Care Med. 1985;13(5):399–401.

Harrigan RA, Chan TC, Moonblatt S, Vilke GM, Ufberg JW. Temporary transvenous pacemaker placement in the Emergency Department. J Emerg Med. 2007;32(1):105–11.

Silver MD, Goldschlager N. Temporary transvenous cardiac pacing in the critical care setting. Chest. 1988;93(3):607–13.

Pericardiocentesis (Optional: Ultrasound Guidance)

33

Katrina Skoog Nguyen, L. Connor Nickels,
and Rohit Pravin Patel

33.1 Indications

- Treatment of hemodynamic compromise from cardiac tamponade
- To diagnose the cause or presence of a pericardial effusion

33.2 Contraindications

- Absolute
  - Aortic dissection
  - Need for immediate surgery for trauma patients
- Relative
  - Coagulopathy
  - Anticoagulant therapy
  - Thrombocytopenia

33.3 Materials and Medications

- Antiseptic (e.g., Chloraprep)
- 1 % lidocaine
- 25-gauge needle, 5/8-in. long
- 18-gauge catheter-type needle, 1½-nch long
- Syringes (10, 20, and 60 mL)
- Ultrasound (US) machine and cardiac/phased array probe
- Sterile US probe cover
- Cardiac monitor

33.4 Procedure

1. Identify the point of maximal effusion with US. Evaluate for hypoechoic or anechoic (dark) effusion around the heart, between the pericardial sac and the myocardium. A patient with hemodynamic compromise from a pericardial effusion or tamponade will have right ventricular collapse, septal bulging, and dilation of the inferior vena cava. Diastolic collapse of the right ventricular free wall can be absent in elevated right ventricular pressure and right ventricular hypertrophy or in right ventricular infarction.
2. Measure the distance from the skin surface to the effusion border to assess the expected needle depth.
3. Choose the needle trajectory based on the point of maximal effusion in the path with the fewest intervening structures. The most commonly used approaches are left parasternal, apical, and subxiphoid. For complex loculated posterior pericardial effusions, optional techniques such as transatrial and transbronchial may be performed by specialists. These types of loculated effusions can occur in autoimmune diseases, infective pericarditis, after cardiac surgery, and after radiotherapy.
4. Sterile preparation: prepare the skin of the entire lower xiphoid and epigastric area with antiseptic. Prepare the US transducer with a sterile sleeve.
5. Local anesthetic: if the patient is awake, anesthetize the skin and planned route of the needle.
6. Pericardial needle insertion: depends on approach used.

K.S. Nguyen, DO
Northwest Community Hospital, Arlington Heights, IL, USA
e-mail: katrinaskoog@hotmail.com;

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida
Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_33

33.4.1 **Subxiphoid Approach** (Fig. 33.1)

- The US transducer is placed just inferior to the xiphoid process and left costal margin.
- Insert the needle between the xiphoid process and the left costal margin at a 30–45° angle to the skin.
- Aim for the left shoulder.

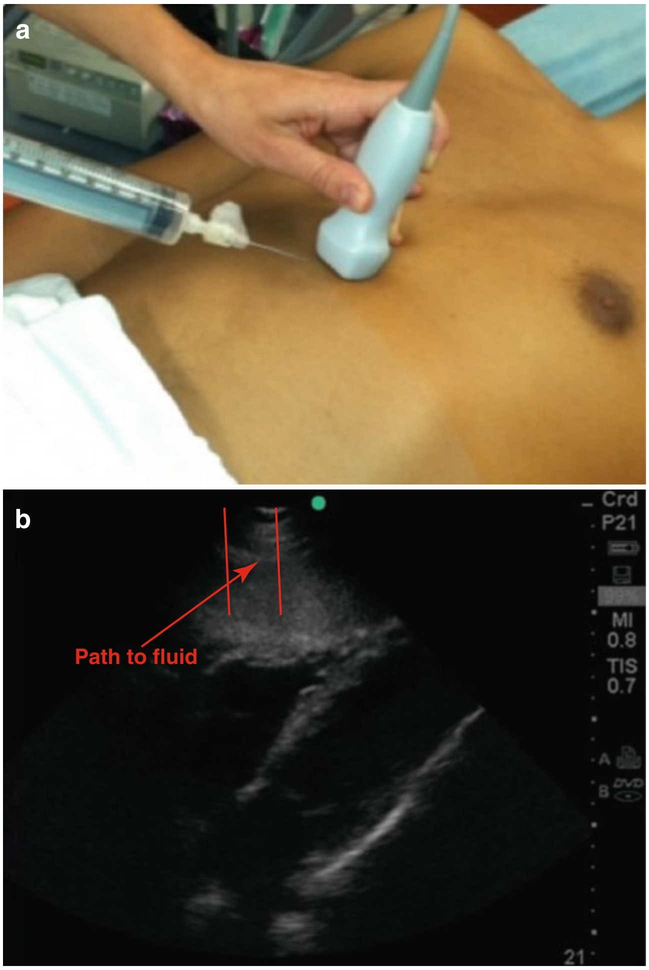
a
b
Path to fluid
Crd
P21
MI
0.8
TIS
0.7
21

Fig. 33.1 Subcostal approach (**a**) and corresponding ultrasound image (**b**). *Red line* through liver to obtain fluid (no fluid present on ultrasound image)

33.4.2 **Apical Approach** (Fig. 33.2)

- The US transducer is placed at the patient's point of maximal impulse and aimed at the patient's right shoulder for a four-chamber view of the heart.
- Insert the needle in the fifth intercostal space 1 cm lateral to and below the apical beat, within the area of cardiac dullness.
- Aim for the right shoulder.
- A Mayo Clinic review showed, in 80 % of total effusions, that the distance to the effusion was least and the size was maximal in the apical approach [1].

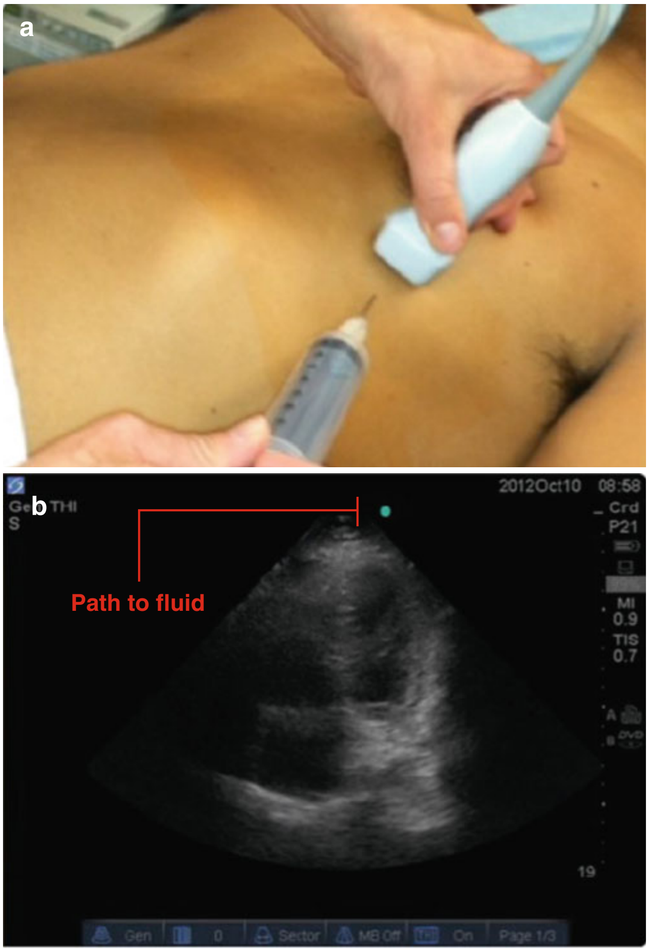
a
b
THI
2012Oct10 08:58
Path to fluid
Crd
P21
MI
0.9
TIS
0.7
19
Gen
Sector
MB Off
On
Page 1/3

Fig. 33.2 Apical approach (**a**) and corresponding ultrasound image (**b**). *Red line* with minimal distance to fluid if present (no fluid present on ultrasound image)

33.4.3 Parasternal Long-Axis Approach

- The US transducer is placed obliquely on the left sternal border between the fourth and fifth ribs with the transducer indicator aimed at the right shoulder.
- Insert the needle perpendicular to the skin in the fifth intercostal space medial to the border of cardiac dullness.
  - Visualize and feel a giving way as the needle penetrates the pericardium. Removal of fluid confirms successful entry. Remove fluid with the goal of restoring hemodynamic stability. Aspiration of fluid should result in improvement in blood pressure and cardiac output.
  - Remove the catheter and apply a dressing. Optional: Place a pigtail catheter using the Seldinger technique for continued drainage.

33.5 Complications

- Blind techniques are associated with 20 % morbidity and 6 % mortality. The complication rate with US-guided approaches is less than 5 % [2].
- Any vital structure within reach of the pericardial needle has the potential for injury: pneumothorax; hemothorax; coronary vessel laceration; hemopericardium; heart chamber lacerations; intercostal vessel injury; dysrhythmias; ventricular tachycardia; puncture of the liver, diaphragm, or gastrointestinal tract; bacteremia; purulent pericarditis; air embolisms; and pleuropericardial fistulas.

33.6 Pearls and Pitfalls

- Pearls
  - A scoring index is available that can be obtained at initial presentation in patients without hemodynamic compromise that identifies those who require pericardial effusion drainage later in the course of treatment. The scoring index uses echocardiographic findings, etiological information, and the size of the effusion at end-diastole [3].
  - Take care to avoid the left internal mammary artery, which travels in a cephalad-to-caudad direction 3–5 cm lateral to the left sternal border [4].
- Pitfalls
  - Tamponade should always be considered in the differential diagnosis of shock or cardiac arrest owing to pulseless electrical activity (PEA). Complications of acute coronary syndrome, aortic dissection, and decompensations in chronic advanced constriction may also need treatment [5].
  - When cardiac compensation mechanisms are exhausted, small increases in pericardial volume can lead to an increase in ventricular diastolic pressure, systemic and pulmonary congestion, and decrease preload and cardiac output.
  - Use US-guidance rather than a blind or an electrocardiogram (ECG)-alone-guided approach for pericardiocentesis to significantly decrease the risk of injury to vital structures [4].
  - Do not delay patient transport to the operating room to perform pericardiocentesis on a traumatically induced effusion unless the patient is hemodynamically unstable and on the verge of cardiac arrest.

References

1. Tsang TS, Enriquez-Serano M, Freeman WK, Barnes ME, Sinak LJ, Gersh BJ, et al. Consecutive 1127 therapeutic echocardiography guided pericardiocentesis: clinical profile, practice patterns, and outcomes spanning 21 years. Mayo Clin Proc. 2002;77:429–36.
2. Guo K, Ding ZP, Tan J. Trans-pleural pericardiocentesis: revisiting an old technique. Catheter Cardiovasc Interv. 2011;78:815–8.
3. Halpern DG, Argulian E, Briasoulis A, Chaudhry F, Aziz EF, Herzog E. A novel pericardial effusion scoring index to guide decision for drainage. Crit Pathw Cardiol. 2012;11:85–9.
4. Seferović PM, Ristić AD, Imazio M, et al. Management strategies in pericardial emergencies. Herz. 2006;31:891–901.
5. Sagristà-Sauleda J, Mercé AS, Soler-Soler J. Diagnosis and management of pericardial effusion. World J Cardiol. 2011;3:135–43.

Ultrasound Evaluation of Pulmonary Embolism and Heart Strain

34

Giuliano De Portu, L. Connor Nickels,
and Marie-Carmelle Elie

34.1 Indications

- Undifferentiated respiratory distress
- Unexplained hypotension
- Evaluation for cardiac tamponade, pericardial effusion

34.2 Contraindications

- Relative
  - Morbidly obese patients
  - Patients with chest wall deformities
  - Patient with subcutaneous emphysema, pneumopericardium
  - Combative or altered patients

34.3 Materials and Medications

- Portable ultrasound machine with appropriate probes (phased array)
- Ultrasonic gel for the probe
- Drapes for the patient (if the condition allows)

34.4 Procedure

1. Begin with the parasternal long-axis view. If possible have the patient turn to the left decubitus side (that helps "move" the heart closer the chest wall).
2. Find the phased array probe, select the "cardiac examination" on the machine, and make sure the orientation marker is pointed toward the right shoulder of the patient.
3. Place the probe along the left side of the sternum over the fourth to sixth intercostal space. This should produce the image shown in Fig. 34.1.
4. Measure the diameter of the ventricle during the end of diastole (normal values, 21 mm ± 1 mm; any measurement >25–30 mm is abnormal).
5. Using the aorta as a landmark, evaluate the structures starting with the pericardium (bright white line around the heart), making sure there is no fluid around it. (Fluid above the aorta indicates a pericardial effusion. Fluid below the aorta indicates a pleural effusion.) If a hypoechoic or anechoic stripe appears in the anterior side of the heart, it is most likely a fat pad. Fluid seen "all around" categorizes it as an effusion.
6. While keeping the probe in the same place, rotate the probe marker 90° clockwise toward the left shoulder to obtain the parasternal short-axis view (Fig. 34.2).

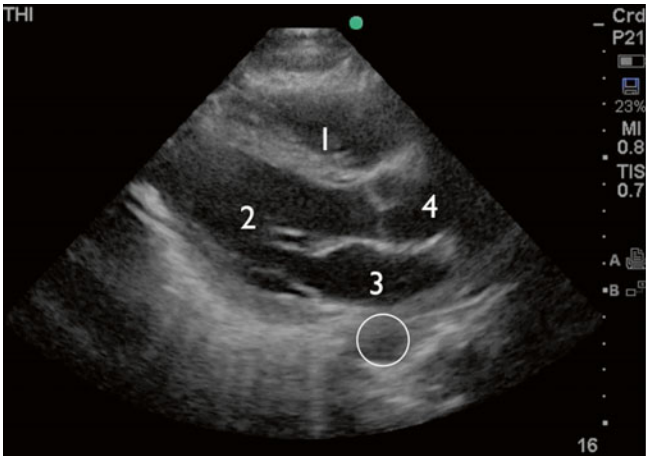

Fig. 34.1 Parasternal long-axis view (PSLA). *1* right ventricle, *2* left ventricle, *3* left atrium, *4* aortic outflow track; *circle* descending aorta (right atrium is not visualized in the PSLA)

G. De Portu, MD • L.C. Nickels, MD, RDMS
M.-C. Elie, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: gdeportu@ufl.edu; cnickels@ufl.edu; elie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_34

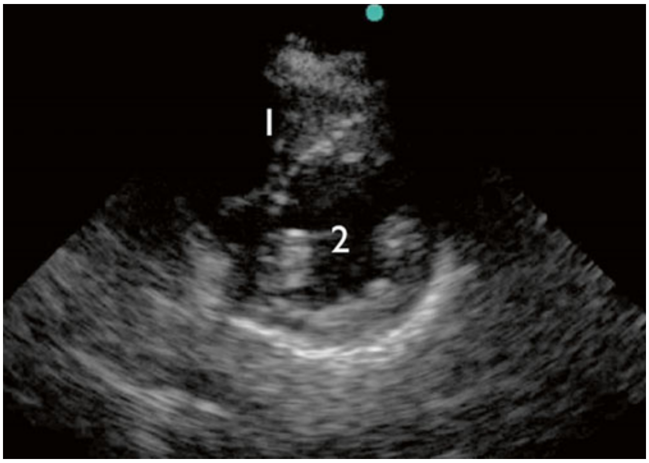

Fig. 34.2 Parasternal short-axis view (PSSA). *1* right ventricle, *2* left ventricle

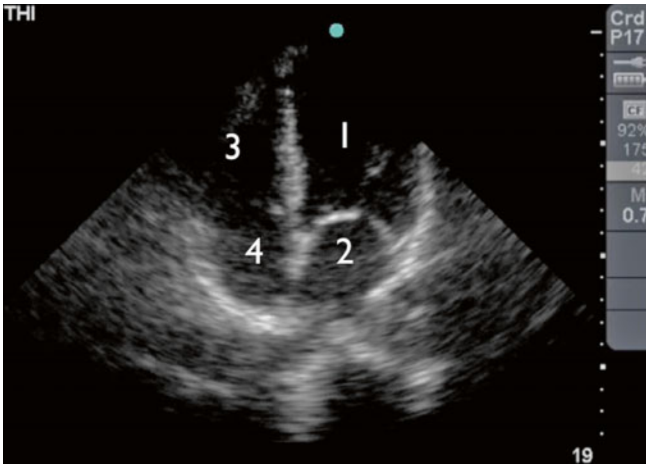

Fig. 34.3 Four-chamber view (4C). *1* left ventricle, *2* left atrium, *3* right ventricle, *4* right atrium

7. The parasternal short-axis view will give information on the contractility of the heart. The right ventricle should be anterior and to the left and the left ventricle to the right. The normal position of the septum bows slightly toward the right ventricle.
8. Obtain a four-chamber view. Place the orientation marker to the patient's left.
9. Palpate for the point of maximum impulse (PMI) and place the transducer. All four chambers should appear in one view (Fig. 34.3).
10. Compare the sizes of the ventricles and note any difference. Notice also the interventricular septum. The normal right-to-left ventricular ratio is less than 0.5. (In an unstable patient, this is probably the most useful because both ventricles can be quickly visualized and the ratio compared.)
11. A subxiphoid approach is also possible. Make sure the orientation marker is toward the right side of the patient (Fig. 34.4).

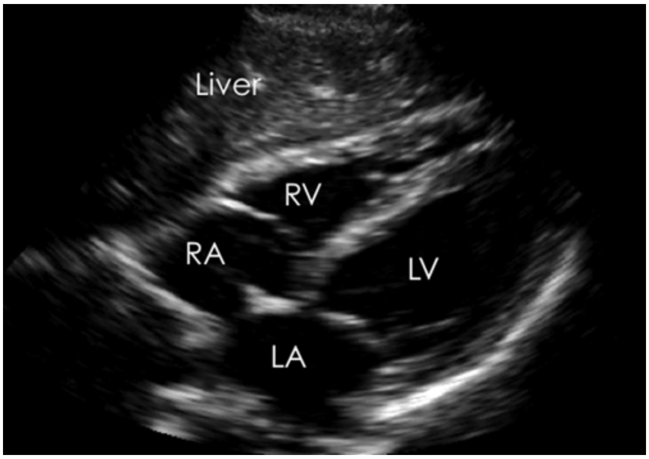

Fig. 34.4 Subcostal/subxiphoid view. *RV* right ventricle, *RA* right atrium, *LV* left ventricle, *LA* left atrium

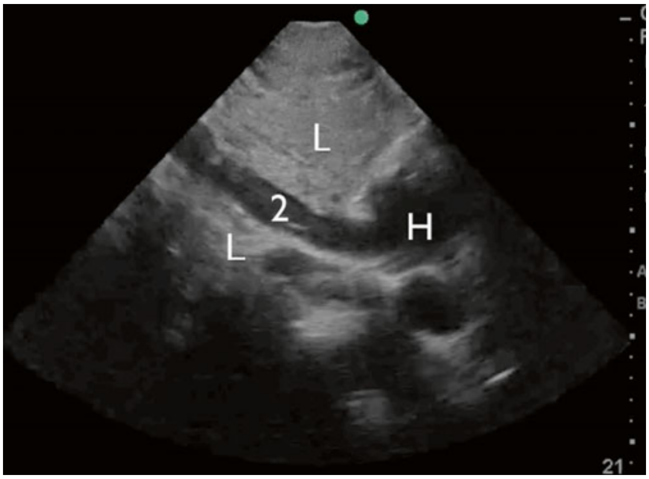

Fig. 34.5 Inferior vena cava. *L* liver, *2* inferior vena cava, *H* heart

12. Use the liver for orientation. The right side of the heart will be nearest to the liver (think that the liver is on the right of the body); again note for any differences in the size ratio.
13. Next take a look at the inferior vena cava (IVC) by placing the curved array transducer just inferior to the xiphoid in a longitudinal fashion. (Switch from the subxiphoid view to the IVC by rotating the probe counterclockwise until the IVC is seen.) Evaluate the IVC. During normal physiological inspiration, the drop in intrathoracic pressure "pulls" blood into the heart, thus decreasing the relative IVC size. If something is preventing venous return, such as a massive PE, collapse will not be as evident and the suspicion for PE increases (fluid overload and increased central venous pressure [CVP] will also account for this finding). Normal IVC diameter is 1.2–2.3 cm, and total collapse and greater than 50 % collapse are normally visualized. An increase in IVC size and less than 50 % or no change has been correlated with increased right atrial pressures (11 to >20 cm Hg) (Fig. 34.5).

34.5 Findings

A right heart that is "strained" or pumping against a higher resistance owing to a PE will show some or all of these changes:

1. Right ventricular dilation (Fig. 34.6).
2. Right ventricular hypokinesis (especially of the middle segment), McConnell's sign but normal motion of the apex.
3. Tricuspid regurgitation.
4. Abnormal septal motion: deviated toward the left ventricle (normally it relaxes during diastole toward the right ventricle); as pressure increases, the right ventricle will not empty properly and septal flattening can be seen.
5. Dilated IVC with little or complete loss of changes in diameter with respiration (variability); the IVC collapses less than 50 % during inspiration.

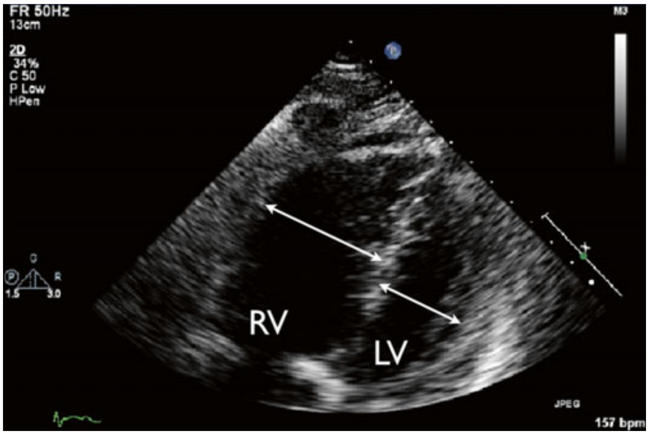

Fig. 34.6 Four-chamber view that shows right ventricular dilatation due to pulmonary embolism. Normally, the LV is greater than the RV during diastole, but in the case of increased pressure, the RV will be enlarged. *RV* right ventricle, *LV* left ventricle

34.6 Complications

- Ultrasound per se has been shown to cause no direct complications with proper use unless it is interfering with advanced airway or life-saving procedures.

34.7 Pearls and Pitfalls

- Pearls
  - Proper interpretation of the images is necessary because some normal anatomy could be confused with a positive finding. Make sure to properly identify the structures because a common mistake is to identify the aorta thinking it is the IVC.
  - Depth: increase to be sure to visualize all structures.
- Pitfalls
  - Ultrasound is user dependent, and the acquisition of images will vary with each user, creating the possibility of false-negatives if images not properly acquired.
  - Reversal of the orientation owing to transducer misplacement could "reverse" the anatomy and allow mistakenly identification of structures.
  - An enlarged right ventricular wall (>5 mm) is considered hypertrophied and would indicate a more chronic etiology of a right ventricular dilation.
  - Gain problems: adjust so the heart chambers are anechoic and the walls are echogenic.

Selected Reading

Kline JA. Thromboembolism In: Tintinalli J, Stapczynski J, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Tintinalli's emergency medicine. 7th ed. New York: McGraw Hill; 2012.

McConnell MV, Solomon SD, Rayan ME, Come PC, Goldhaber SZ, Lee RT. Regional right ventricular dysfunction detected by echocardiography in acute pulmonary embolism. Am J Cardiol. 1996;78:469–73.

Reardon RF, Joing SA. Cardiac. In: Ma OJ, Mateer J, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007. p. 109–44.

Vanden Hoek TL, Morrison LJ, Shuster M, et al. Part 12: cardiac arrest in special situations: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010;122(18 Suppl 3):S829–61.

Pacemaker Evaluation in the Emergency Department

35

Joseph D. Romano and Christian Coletti

35.1 Pacemaker Function

There are approximately 500,000 implanted cardiac pacemakers in the USA and another 100,000 are implanted each year [1]. In the emergency department, a physician may be faced with a pacemaker that is not functioning appropriately. To understand the ways in which a pacemaker can malfunction or lead to medical complications, first it is important to understand how pacemakers work when they do so appropriately.

A common pacemaker system is composed of a pulse generator and insulated wire leads that originate in the pulse generator and end within the myocardium. The pulse generator is implanted in the pectoral region. It contains circuitry and the battery and creates the electrical impulses that depolarize the myocardium. The leads leave the pulse generator and are directed to the heart by following the venous system. Leads may terminate in the right atrium, the interventricular septum of the right ventricle, or biventricular pacing; a third lead navigates the coronary sinus to the left ventricular wall (Figs. 35.1 and 35.2).

Early pacemakers produced impulses only at a set rate. Today, however, virtually all pacemakers have sensing leads that detect intrinsic activity and react with electrical impulses when the intrinsic intervals fall outside of a set time threshold. Some pacemakers are programmed to allow for rate changes based on physical activity. To simplify the classification of different pacemaker types, the North American Society of Pacing and Electrophysiology and British Pacing and Electrophysiology Group (NASPE/BPEG) developed a five-letter code to describe each pacemaker (Table 35.1) [2].

J.D. Romano, MD (✉) • C. Coletti, MD
Departments of Emergency Medicine and Internal Medicine,
Christiana Care Health System,
Newark, DE, USA
e-mail: jromano@christianacare.org; chris.coletti@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_35

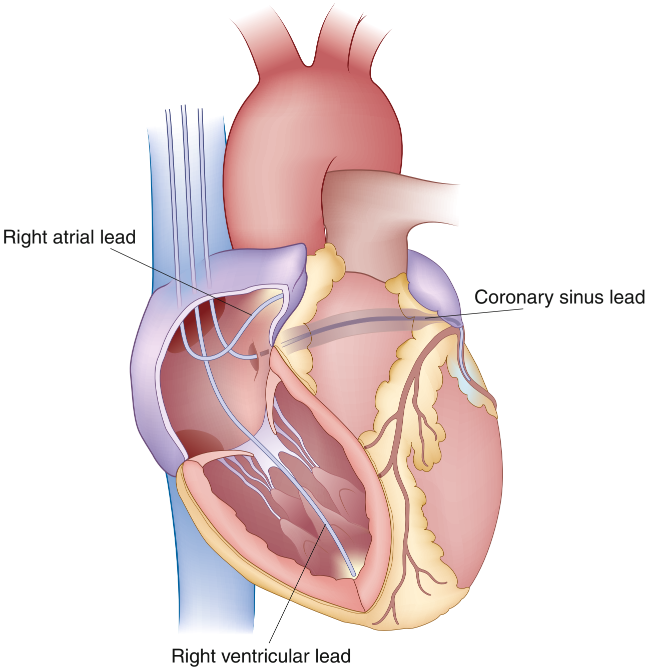
Right atrial lead
Coronary sinus lead
Right ventricular lead

Fig. 35.1 Common pacemaker lead insertion sites

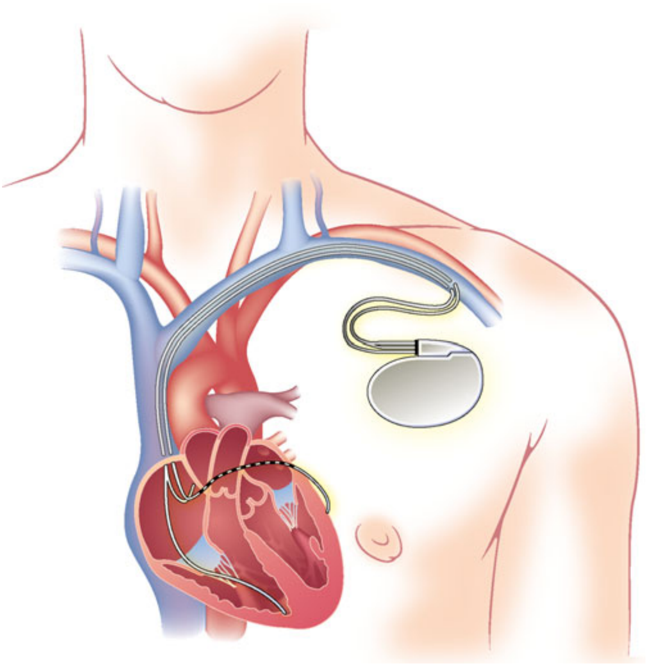

Fig. 35.2 Common generator location and path of pacemaker leads

Table 35.1 NASPE/BPEG five-letter pacemaker code

Paced chamber	Sensed chamber	Response to sensing	Rate modulation	Multisite pacing
A = atrium	A = atrium	T = triggered	R = rate modulation	A = atrium
V = ventricle	V = ventricle	I = inhibited	O = none	V = ventricle
D = dual	D = dual	D = dual		D = dual
O = none	O = none	O = none		O = none

Reproduced with permission from Bernstein et al. [2]

35.2 Common Pacemaker Codes [3]

- VVI: ventricular paced and sensed, inhibited by normal intrinsic pacing; used for patients with a need for pacing if bradycardia occurs
- VVIR: ventricular paced and sensed, inhibited by normal intrinsic pacing; adaptive to changes in intrinsic rate for physically active patients
- DDD: dual-chamber paced and sensed, inhibited by normal intrinsic atrial and ventricular electrical activity; used in third-degree atrioventricular (AV) block to allow for more physiological conduction
- DDDR: dual-chamber paced and sensed, inhibited by normal intrinsic atrial and ventricular electrical activity; adaptive to changes in intrinsic rate; used primarily in sinoatrial (SA) node dysfunction to closely mimic normal, adaptable heart conduction

35.3 Pacemaker Malfunction

Malfunctioning pacemakers can have complicated programming that is best altered after thorough "interrogation" by a trained electrophysiologist. This type of evaluation is beyond the scope of this chapter and, in many clinical settings, subspecialty support is not readily available. In the emergency department, it is vital to stabilize the patient and identify common pacemaker system issues. Pacemaker system malfunctions can be identified as a failure to sense, failure to pace, failure to capture, or pacing at an inappropriate rate [4]. Another common abnormality associated with pacing is the "pacemaker syndrome" [3].

- Failure to sense:
  - Oversensing: The sensor interprets external stimuli as a normal ventricular rate leading to inappropriate inhibition. This may be due to lead fracture, fibrosis of the lead tip, or lead dislodgment. Cross talk is present when the atrial stimulus is sensed on the ventricular lead causing inappropriate inhibition. Oversensing is rarely due to failure of the generator.
  - Undersensing: Present when there is constant pacing despite cardiac activity that has not exceeded threshold. This can be due to low-amplitude intrinsic cardiac activity, lead dislodgment, battery depletion, or metabolic abnormalities.
- Failure to pace:
  - Lack of pacing despite appropriate sensing of the intrinsic electrical activity or activity below the threshold rate.
  - Typically caused by lead fracture, battery depletion, or fibrosis of the lead tip.
  - It is rare for any specific part of the device to fail, but suspicion should be high if the patient had recent radiation therapy, electrocautery, defibrillation, electroshock therapy, magnetic resonance imaging (MRI), lithotripsy, or transcutaneous electrical nerve stimulation (TENS) treatments.
  - Rarely due to battery depletion.
- Failure to capture:
  - A lack of cardiac response despite appropriate sensing and subsequent pacer stimuli.
  - Commonly due to lead dislodgement, but can be due to myocardial perforation, lead fracture, fibrosis of the lead tip, poor lead placement, battery depletion, and antiarrhythmic medications.
  - Functional failure to capture occurs when pacer stimuli fall within the refractory period of previous depolarization.
- Pacing at an inappropriate rate:
  - Likely due to an endless-loop reentry tachycardia known as pacemaker-mediated tachycardia.
    - More common in DDD pacemakers.
    - Initiated by a premature ventricular stimulus, which is carried retrograde through the intrinsic conduction system to the atrioventricular node (AVN) and then the atria. This conduction is sensed by the atrial lead and causes triggering of pacing back in the ventricle. The ventricular depolarization is then sensed again in the atria, forming a loop. The intrinsic conduction system acts as a retrograde limb and the pacer circuit acts as the anterograde limb of a reentry tachycardia.
- The "pacemaker syndrome":
  - Constellation of symptoms found in 20 % of patients with pacemakers.
  - Symptoms include syncope, near-syncope, dizziness, fatigue, weakness, pain, shortness of breath, and cough.
  - Ventricular rates are poorly timed with atrial activity such that atrial contraction occurs against closed mitral and tricuspid valves.
  - Ventricles lose the benefit of the atrial kick, the atria enlarge, and signs and symptoms similar to congestive heart failure ensue.

35.4 Initial Evaluation of the Patient with a Pacemaker [4]

- History:
  - Symptoms of syncope, near-syncope, chest pain, palpitations, and irritation at the generator pocket
  - Brand and NASPE/BPEG code
  - Date of implantation
  - Location of generator pocket and any previous locations
  - Most recent electrophysiological interrogation
  - Medications that interfere with function, such as flecainide or lidocaine
  - Recent procedures, such as MRI, lithotripsy, or defibrillation
- Physical examination:
  - Check vital signs, listen for heart sounds (paradoxically split S2 is normal).
  - Inspect generator pocket; turning of the leads is associated with Twiddler's syndrome, in which a patient compulsively touches the skin around the device.
  - Look for jugular venous distension.
- Laboratory studies and electrocardiogram (ECG):
  - Obtain a chest x-ray to determine effusions, infiltrates, generator placement, and lead placement or fracture.
  - Electrolyte imbalances must be detected and corrected.
  - ECG: The following are examples of appropriate ECG patterns in patients based on type of pacing [3]:
    - VOO (asynchronous pacing): Regular pacer spikes lead to immediate QRS waveforms. Left bundle branch block (LBBB) is a normal finding in a right ventricular lead location.
    - VVI: Pacer spikes and an LBBB waveform should be seen if the intrinsic rate is below the threshold rate.
    - DDD: Various patterns are possible. If the intrinsic rhythm and intervals are normal, then no pacer spikes will be seen. If the atrial rate is slow and AV delay is normal, then an atrial spike will cause a P wave and a normal QRS. If AV delay is prolonged, then two spikes may be observed: a P wave and a QRS with an LBBB waveform.
- If malfunction is suspected:
  - Obtain intravenous (IV) access, place the patient on a heart monitor, and, if possible, consult cardiology for interrogation and reprogramming.
  - Obtain a ring magnet:
    - Positioning over the generator causes the pacing stimuli to revert to an asynchronous ventricular pacing mode at a set rate [5].

35.5 Interpreting the Type of Malfunction Based on ECG

The first step in evaluating the ECG of a malfunctioning pacemaker is to determine the presence of pacing spikes [6]:

- If pacing spikes are noted on the ECG, determine whether capture is present by ensuring that an appropriate waveform follows each pacer spike and that there is an associated pulse.
  - If capture is present, check the rate:
    - If the rate is appropriate, the pacer is functioning normally.
      - If the rate is slow, suspect oversensing.
      - If the rate is rapid, suspect undersensing or pacemaker-mediated tachycardia.
  - If capture is not present, consider metabolic effects or component failure.
- If pacing spikes are not noted on the ECG, determine whether the patient is in an intrinsic rhythm:
  - If the patient is in an intrinsic rhythm and the rate is appropriate, application of a ring magnet will cause pacer spikes to show up at a set rate. This is normal functioning.
    - If application of the magnet does not cause pacing, suspect mechanical failure.
  - If the patient is not in an intrinsic rhythm, place a magnet over the generator.
    - If magnet application causes pacing, consider oversensing.
    - If it does not, consider mechanical failure.
- If a patient is hemodynamically unstable and application of a magnet leads to stability, keep the magnet in place until the patient is able to have the pacemaker interrogated.

35.6 Management of Pacemaker-Mediated Tachycardia

- After ensuring adequate IV access, placing a heart monitor, and interpreting a baseline ECG, put a ring magnet over the generator [5].
  - If a normal rhythm results, keep the magnet on the chest.
  - If it does not change the rate, attempt isometric pectoral exercises by having the patient press the left hand against the right shoulder.
    - This is an attempt to overstimulate the pacemaker sensor and precipitate inhibition of pacer output.
    - If this is unsuccessful, consider transcutaneous pacing.
    - If transcutaneous pacing is unsuccessful, then the leads may require surgical adjustment or removal.

References

1. Ford-Martin PA, Spiwak AJ. Pacemakers. In: Gale encyclopedia of surgery: a guide for patients and caregivers. 2004. Encyclopedia.com: http://www.encyclopedia.com/doc/1G2-3406200337.html. Accessed 01 Apr 2011.
2. Bernstein A, Daubert J, Fletcher R, et al. The revised NASPE/BPEG generic code for antibradycardia, adaptive-rate, and multisite pacing. North American Society of Pacing and Electrophysiology/British Pacing and Electrophysiology Group. Pacing Clin Electrophysiol. 2002;25:260–4.
3. Marx JA, Hockberger RS, Walls RM, Adams J, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2009.
4. Bonow RO, Mann DL, Zipes DP, Libby P, editors. Braunwald's heart disease: a textbook of cardiovascular medicine. 9th ed. Philadelphia: Saunders; 2011.
5. Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004.
6. Kaszala K, Huizer JF, Ellenbogen KA. Contemporary pacemakers: what the primary care physician needs to know [review]. Mayo Clin Proc. 2008;83:1170–86.

Part V

Spine Evaluation

36

Cervical Collar Placement

Justin Bennett and Lars K. Beattie

The cervical spine (C-spine) accounts for the majority of all spinal injuries. In the US prehospital setting, patients are often transported using cervical collars (C-collars) and rigid backboards in trauma and when spinal cord injuries are suspected. After airway, breathing, and circulation, every effort must be made to secure the C-spine.

There are two main subcategories of C-collars: the one-piece and the two-piece. One-piece C-collars include the Stifneck and the Ambu Perfit. Two-piece C-collars have posterior and anterior pieces, with the anterior piece usually the larger of the two. Examples of two-piece collars include the Aspen collar, the Philadelphia collar, and the Miami J collar. The basic features of a C-collar include:

- Adjustable circumference with fasteners (usually Velcro straps)
- Adjustable height with a locking device of different sizes
- Hooks for a nasal cannula
- Exposure of the anterior neck for pulse checks and advanced airway procedures
- Posterior access for cervical palpation
- Padding to protect the soft tissues of the neck

36.1 Indications

- Prehospital suspicion for spine trauma:
  - Emergency department patients or trauma patients who fail clinical rule-out criteria such as NEXUS (National Emergency X-Radiography Utilization Study) and Canadian C-spine rules (see Chap. 37)

36.2 Contraindications

- Absolute:
  - Cervical dislocation with fixed angulation
  - Impaled foreign object in the neck
  - Massive soft tissue swelling in the neck
- Relative:
  - Unsecured airway
  - Surgical airway
  - Vomiting
  - Mandible or soft tissue injuries with potential for airway compromise
  - Preexisting anatomical abnormalities

36.3 Materials and Medications

- Properly fitting C-collar
- Consider:
  - Head blocks, if needed for lateral stabilization
  - Towels or backboard pads for custom support
    - Under shoulders – pediatric patients
    - Under occiput – adults with poor C-spine mobility

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University
School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_36

36.4 Procedure (Aspen Collar)

1. Address airway, breathing, and circulation while maintaining in-line immobilization, before placing a C-collar.
2. Gather personnel:
   - (a) One person is needed to apply the collar to an awake patient.
   - (b) Two or more people may be required when a patient has an altered level of consciousness:
     - (i) One to maintain in-line immobilization in the neutral position
     - (ii) One to place the C-collar
3. While the C-spine is being held in neutral position, assess the airway before placing the C-collar:
   - (a) Anticipate and prepare for airway compromise early to avoid a crash intubation.
   - (b) Place airway if necessary.
4. Palpate and inspect the C-spine, head, and shoulders for evidence of trauma *before* placing the C-collar.
5. While maintaining the neutral position of the C-spine, place the C-collar:
   - (a) Remove loose clothing, jewelry, and earrings that may cause soft tissue pressure wounds.
   - (b) Begin with the piece of C-collar that fits under the occiput (Fig. 36.1).
   - (c) Fold the Velcro straps behind the C-collar.
   - (d) Gently hold back the hair (Fig. 36.2).
   - (e) Slide the occipital section or piece of the C-collar behind the occiput (Fig. 36.3). (Use in-line C-spine stabilization in patients with an altered level of consciousness.)
   - (f) Wrap (one-piece) or place (two-piece) the anterior section of the C-collar around the circumference of the patient's neck and snugly under the chin (Fig. 36.4).
   - (g) Once the Velcro fasteners are in place, ensure that the height is properly adjusted on the C-collar to minimize C-spine mobility.
     - (i) Most C-collars have height adjustments that utilize a locking device that requires releasing a locking mechanism by pulling out.

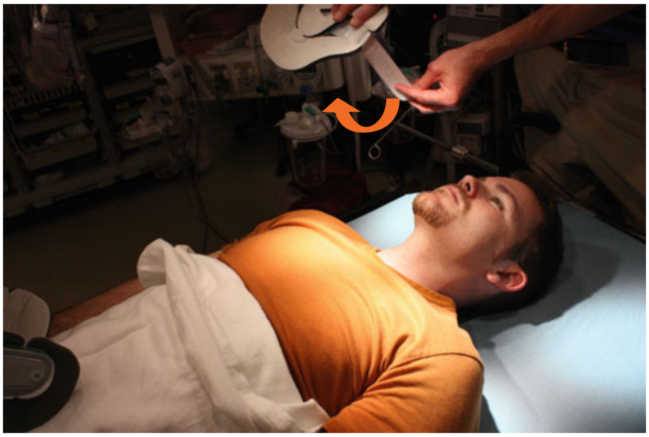

Fig. 36.1 Fold Velcro strap of C-collar posteriorly

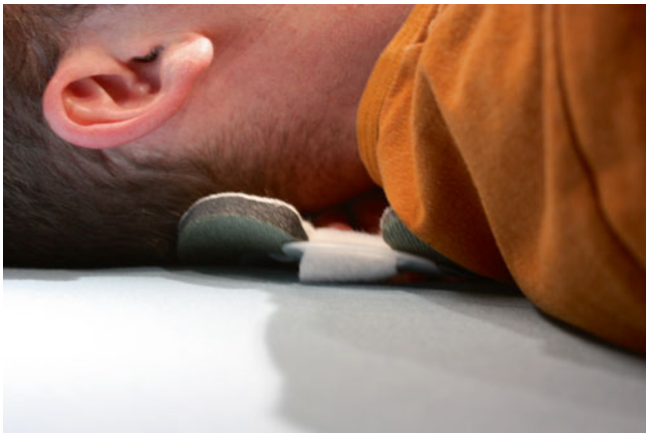

Fig. 36.2 Slide occipital section of collar behind occiput flat against stretcher

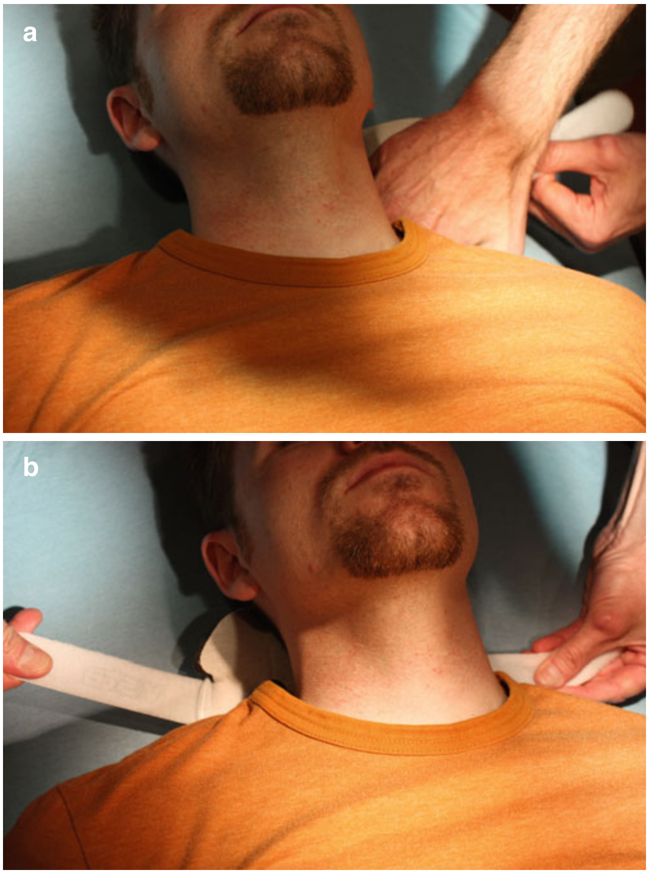

Fig. 36.3 (**a**) Use of hand to guide C-collar behind occiput; (**b**) unfold Velcro strap and position collar behind cervical spine

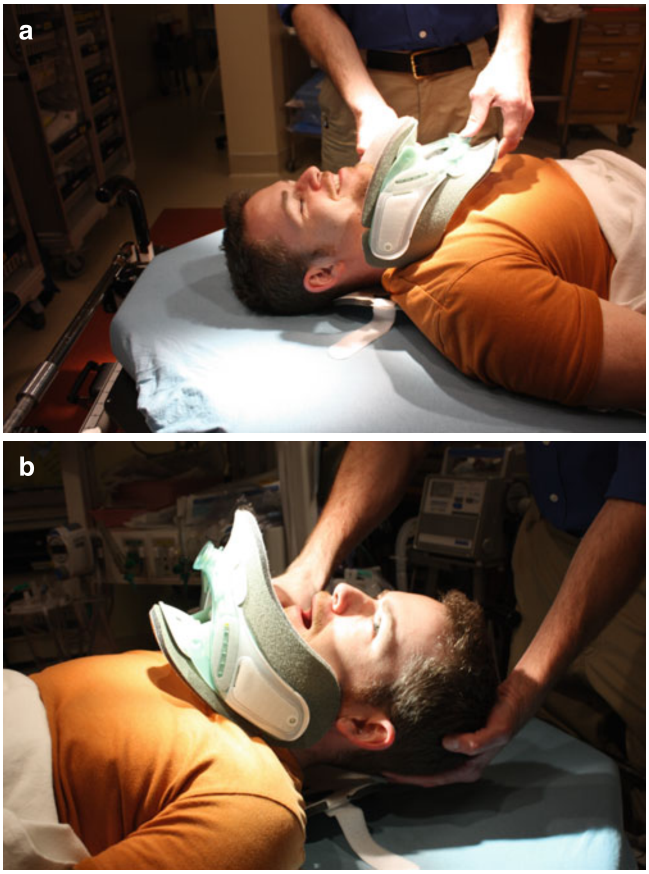

Fig. 36.4 (**a**) Correct orientation of anterior portion of C-collar; (**b**) incorrect orientation of anterior portion of C-collar may cause injury

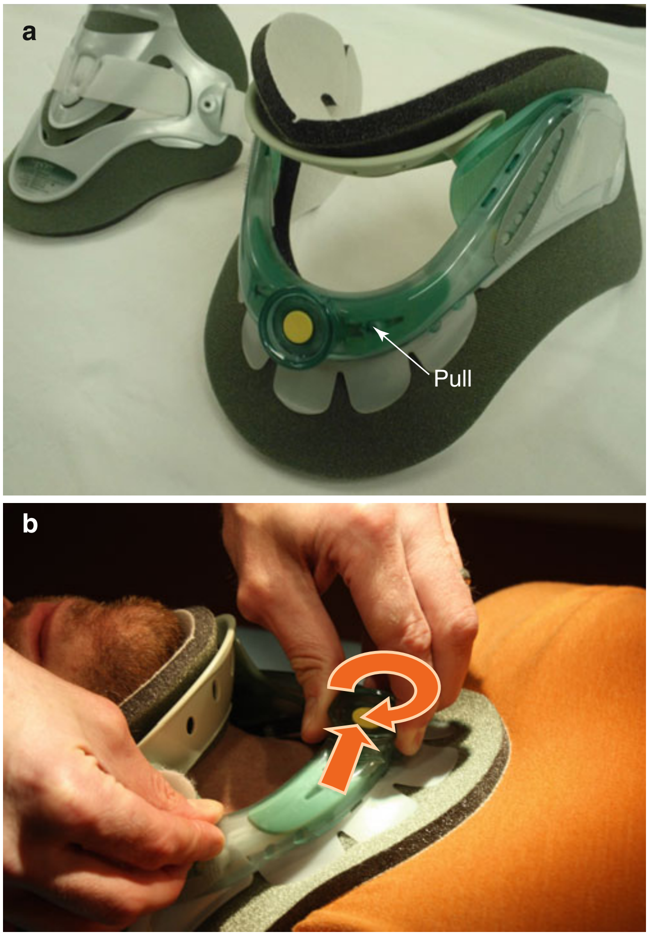

Fig. 36.5 Aspen collar is adjusted with a combination of (**a**) pulling out while (**b**) twisting a round knob at the sternal notch

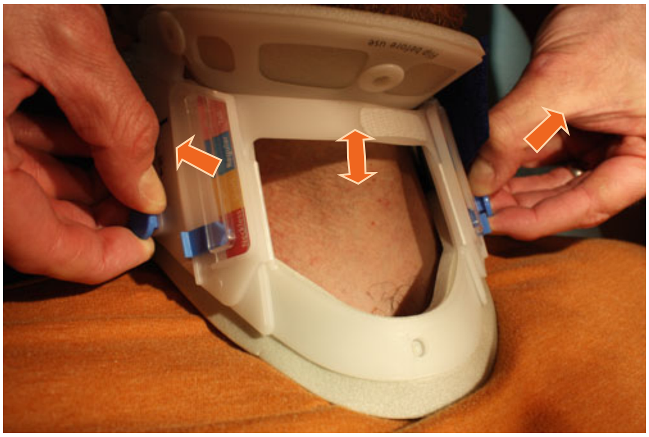

Fig. 36.6 Adjustment the height of the Ambu Perfit one-piece collar by simultaneously pulling (away from the patient) the two locking pins out and adjusting the height of the collar before pushing the pins back in (toward the patient) to lock the collar at the desired height

   - (ii) The Aspen collar is adjusted with a combination of pulling out while twisting a round knob at the sternal notch (Fig. 36.5).
   - (iii) Pull out (away from patient) and twist for height adjustment of the Aspen two-piece collar.
   - (iv) Adjust the height of the Ambu Perfit one-piece collar by simultaneously pulling (away from the patient) the two locking pins out and adjusting the height of the collar before pushing the pins back in (toward the patient) to lock the collar at the desired height (Fig. 36.6).

6. Special cases:
   - (a) Depending on the age of the patient, it may be necessary to place towels under the shoulder blades to keep the neck in a neutral position.
   - (b) Children:
     - (i) Head-to-body ratios are relatively larger than that of adults.
     - (ii) Placement on a backboard may cause significant neck flexion because the occiput rests on a flat board.
     - (iii) Towels can be placed under a child's shoulders to minimize flexion.

   (c) Adults with excess soft tissue or degenerative changes that prevent C-spine straightening:
      (i) A towel or pad is placed behind the occiput to prevent hyperextension.

36.5 Pearls and Pitfalls

- Pearls
  - Airway, breathing, and circulation should be assessed before placement of the C-collar.
  - A high index of suspicion of C-spine injury is needed in intoxicated or comatose patients and the elderly.
  - Always remember to adjust the height and circumference of a C-collar for a snug fit.
- Pitfalls
  - Spinal immobilization increases the risk of aspiration in vomiting patients.
  - Failure to palpate and inspect the C-spine, head, and shoulders before C-collar placement may delay recognition of an impending airway emergency and conceal signs of critical injury needing rapid assessment and care.
  - Lateral C-spine motion is unsecured unless tape, rubber blocks, or towels are used for support.
  - Overlooking C-spine stabilization during C-collar placement leaves the C-spine at risk of further injury in patients with an altered level of consciousness.
  - C-collars that are too tight may decrease venous return from the head and may increase intracranial pressure.

Selected Reading

American College of Surgeons Committee on Trauma. Advanced trauma life support for doctors. 8th ed. Chicago: American College of Surgeons; 2004.

Hankins DG, Boggust A. Prehospital equipment and adjuncts. In: Tintinalli JE, Stapczynski JS, Cline DM, Ma OJ, Cydulka RK, Meckler GD, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2012.

Hoffman JR, Wolfson AB, Todd K, Mower WR. Selective cervical spine radiography in blunt trauma: methodology of the National Emergency X-Radiography Utilization Study (NEXUS). Ann Emerg Med. 1998;32:461–9.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2009.

Stiell IG, Clement CM, McKnight RD, et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med. 2003;349:2510–8.

Cervical Spine Clearance

37

Braden Hexom and Tatiana Havryliuk

37.1 Indications

- Assessment of the need for radiological imaging in trauma

37.2 Contraindications

- Known unstable cervical spine (C-spine) fracture
- Known unstable ligamentous injury
- Intoxication/altered mental status
- Presence of distracting injury

37.3 National Emergency X-Radiography Utilization Study (NEXUS) and Canadian C-Spine Rules (CCR) to Assess the Need for Imaging

- *The NEXUS clinical criteria*
  1. Tenderness at the posterior midline of the C-spine
  2. Focal neurological deficit
  3. Decreased level of alertness
  4. Evidence of intoxication
  5. Clinically apparent pain that might distract the patient from the pain of a C-spine injury
- The presence of any one of the above findings is considered to be clinical evidence that a patient is at increased risk for C-spine injury and requires radiographic evaluation.
- *CCR*
  For alert (Glasgow Coma Scale [GCS] 15) and stable trauma patients
  1. Any high-risk factor that mandates radiography? *YES → Radiography*
     - Age older than 65 years
     - Dangerous mechanism (Table 37.1)
     - Paresthesias in extremities

     ↓ *NO*
  2. Any low-risk factor that allows safe assessment of range of motion? *NO → Radiography*
     - Simple rear-end motor vehicle crash (MVC)
     - Sitting position in emergency department
     - Ambulatory at any time
     - Delayed onset of neck pain
     - Absence of midline C-spine tenderness

     ↓*YES*
  3. Able to actively rotate neck? *NO → Radiography*
     - 45° left and right

     ↓*ABLE*

     No radiography
- CCR found to have higher sensitivity (99.4 % vs. 90.7 %) and specificity (45.1 % vs. 36.8 %) and a lower rate of imaging (55.9 % vs. 66.6 %) than NEXUS criteria for C-spine clearance of low-risk alert trauma patients.

Table 37.1 Dangerous Mechanism

Fall from elevation ≥3 ft/5 stairs
Axial load to head (e.g., diving)
Motor vehicle crash at high speed (>100 km/h), rollover, ejection
Motorized recreational vehicles
Bicycle crash

B. Hexom, MD (✉)
Department of Emergency Medicine, Mount Sinai Hospital,
New York, NY, USA
e-mail: braden.hexom@mssm.edu

T. Havryliuk, MD
Department of Emergency Medicine, University of Colorado
Denver, Denver, CO, USA
e-mail: tatiana.havryliuk@ucdenver.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_37

37.4 Type of Imaging

- Bones
  - C-spine computed tomography (CT)—the new gold standard.
  - Plain X-rays—less sensitive than CT; in one study X-rays missed 45 % of injuries that were picked up by CT [1].
- Ligaments
  - C-spine magnetic resonance imaging (MRI)—consider for possible ligamentous injuries and for further evaluation of obtunded patients.
  - Flexion-extension X-rays—less sensitive than MRI for detection of ligamentous injuries in the acute phase. Delayed flexion-extension films are more sensitive for ligamentous injury than those done the day of injury, but less sensitive than MRI.

37.5 Materials

- Cervical collar (C-collar) (Fig. 37.1)

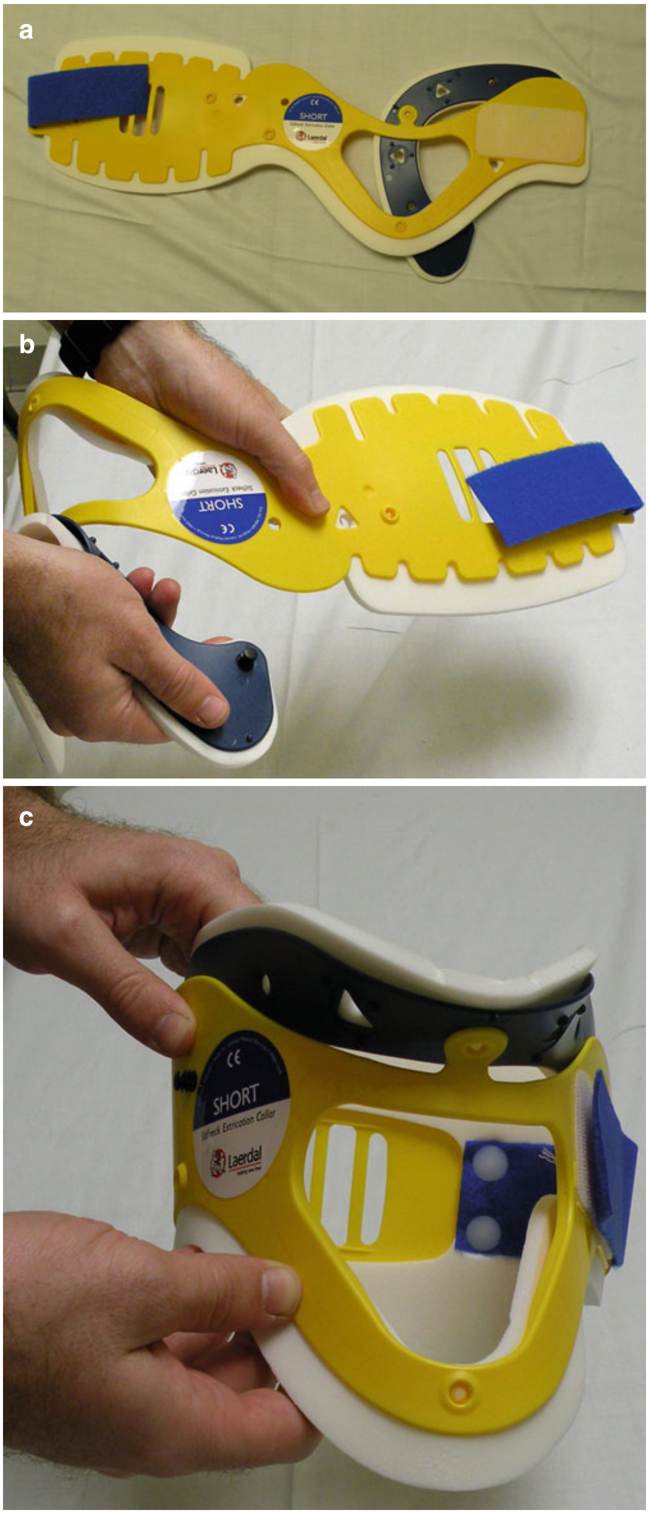

Fig. 37.1 Assembly of C-collar: Select appropriate size (**a**). Invert chinstrap and snap into place (**b**). Assembled C-collar (**c**)

37.6 Procedure (Fig. 37.2)

1. Apply C-collar to stabilize spine. Select appropriate size.
2. Perform a brief neurological examination and assess the patient's mental status; proceed only if both are normal.
3. Loosen the collar to palpate the midline while the patient holds his or her head still. If spinal tenderness exists replace the C-collar and proceed to imaging (Fig. 37.3).
4. Instruct the patient to rotate the neck 45° to each side and flex the neck. If the patient is pain free and with no neurological comprise, the C-collar may be removed and no imaging of the C-spine is required (Fig. 37.4).

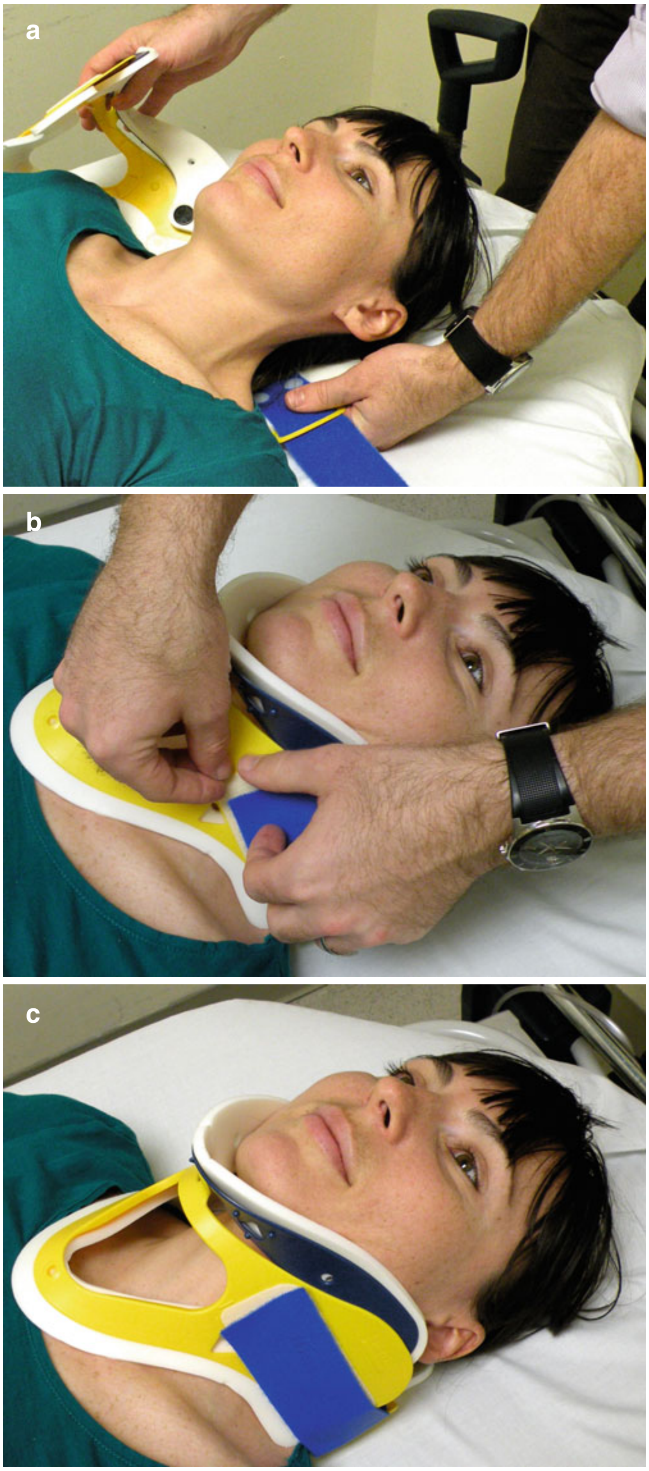

Fig. 37.2 Proper placement of C-collar: Slide C-collar under the neck while keeping neck immobilized (**a**). Secure the collar (**b**). Correct C-collar placement (**c**)

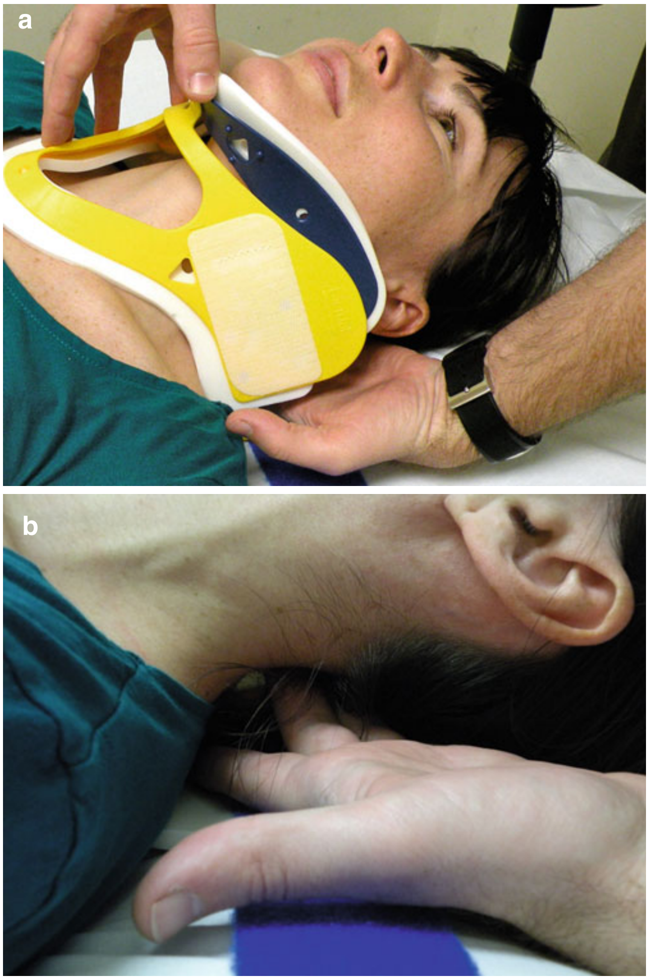

Fig. 37.3 Assessment of C spine tenderness: Loosen C-collar while keeping neck immobilized (**a**). Palpate midline of C spine (**b**)

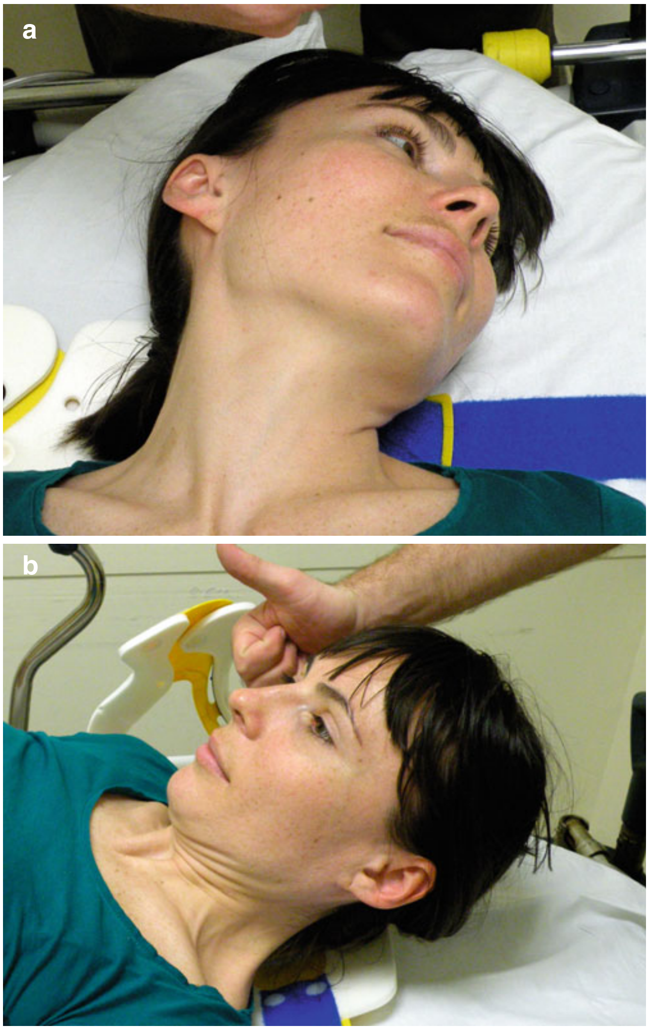

Fig. 37.4 C-spine clearance: Instruct patient to rotate neck 45° each way (**a**). Instruct patient to flex neck (**b**)

37.7 Complications

- Missing clinically important C-spine injury
- Pressure ulcer from prolonged use of C-collar

37.8 Pearls and Pitfalls

- *Pearls*
  - Select appropriate size of C-collar.
  - Ensure adequate mental status because further imaging might be required in obtunded/intoxicated patients.
- *Pitfalls*
  - Avoid prolonged use of C-collar because this can lead to a pressure ulcer.

Reference

1. Schenarts PJ, Diaz J, Kaiser C, Carrillo Y, Eddy V, Morris Jr JA. Prospective comparison of admission computed tomographic scan and plain films of the upper cervical spine in trauma patients with altered mental status. J Trauma. 2001;51:663–8; discussion 668–9.

Selected Reading

Hoffman JR, Schriger DL, Mower W, Luo JS, Zucker M. Low-risk criteria for cervical-spine radiography in blunt trauma: a prospective study. Ann Emerg Med. 1992;21:1454–60.

Platzer P, Jaindl M, Thalhammer G, et al. Clearing the cervical spine in critically injured patients: a comprehensive C-spine protocol to avoid unnecessary delays in diagnosis. Eur Spine J. 2006;15:1801–10.

Stiell IG, Clement CM, McKnight RG, et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med. 2003;349:2510–8.

Stiell IG, Wells GA, Vandemheen K, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. JAMA. 2001; 286:1841–8.

Log Roll

38

Justin Bennett and Lars K. Beattie

38.1 Indications

- Any patient arriving on a rigid backboard in the emergency department
- Assessment of posterior traumatic injuries
- Performed as early as possible after arrival to prevent skin breakdown from pressure ulcers

38.2 Contraindications

- Improperly fitted cervical collar (C-collar) or unsecured cervical spine (C-spine) before log roll
- Unsecured endotracheal tube before log-rolling intubated patients

38.3 Materials and Medications

- Personnel: three or four people
  - One to stabilize the C-spine
  - One or two to roll the patient
  - One to palpate the length of the spine
- Properly fitting hard C-collar
- Trauma shears for removing transport straps

38.4 Procedure

1. Ensure that the airway, breathing, and circulation (ABCs) are established before attempting to remove the patient from the backboard.
2. Gather personnel, at least three, but preferably four people. Larger patients will often require additional assistance.
3. For intubated patients an additional person will be needed to secure the endotracheal tube during the log roll.
4. Ensure the height is properly adjusted on the C-collar to minimize C-spine mobility (Fig. 38.1).
5. Position the stretcher at an ergonomic lifting position for the person responsible for stabilizing the C-spine.
6. Clothing and transport straps should be removed before the log roll to improve visualization of injuries.
7. The patient is asked to cross his or her hands over the chest.
8. C-spine stabilization (at head of bed)
   - Grasp the patient's trapezoids at the midclavicular line.
   - Secure the sides of the patient's head between the operator's forearms.
   - Stabilize the patient's head in neutral anatomical position relative to the body using the operator's forearms (Fig. 38.2).

J. Bennett, MD
Department of Emergency Medicine, Wake Forest University
School of Medicine, Winston-Salem, NC, USA
e-mail: jubennet@wakehealth.edu

L.K. Beattie, MD, MS (✉)
Department of Emergency Medicine, University of Florida,
Gainesville, FL, USA
e-mail: lars.beattie@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_38

9. Thoracic and lumbar spine stabilization
    - One or two (preferably two) people should stand next to the patient on the side to which the patient will be rolled.
        - One person: Place hands over the patient's shoulder and hip.
        - Two people
            - First person places hands on the patient's shoulder and hip.
            - Second person places hands on the patient's hip and knees.
    - The decision to roll to the left or right side is determined by injury sites, to minimize injury exacerbation and pain, and to minimize risk of endotracheal tube dislodgment.

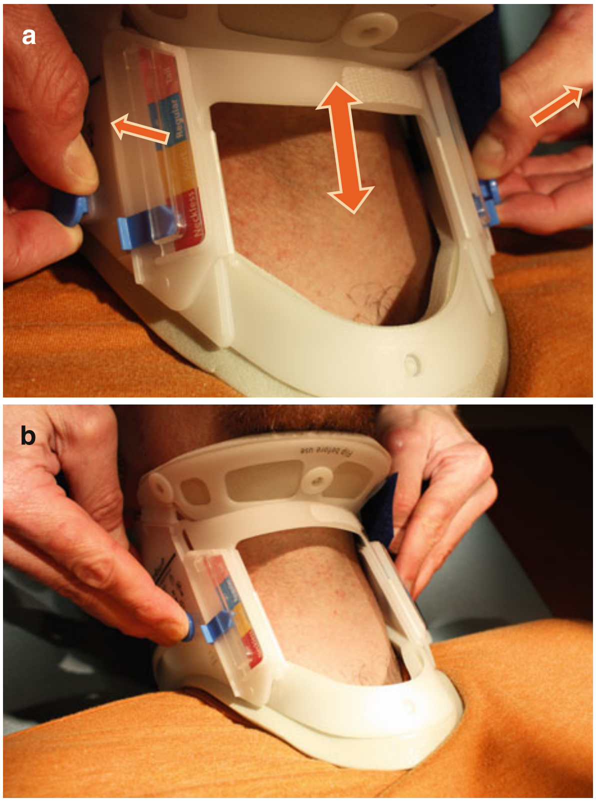

Fig. 38.1 (**a**) The height of the Ambu Perfit one-piece collar is adjusted by simultaneously pulling the two locking pins out (away from the patient) and adjusting the height of the collar, (**b**) then pushing the pins back in (toward the patient) to lock the collar at the desired height

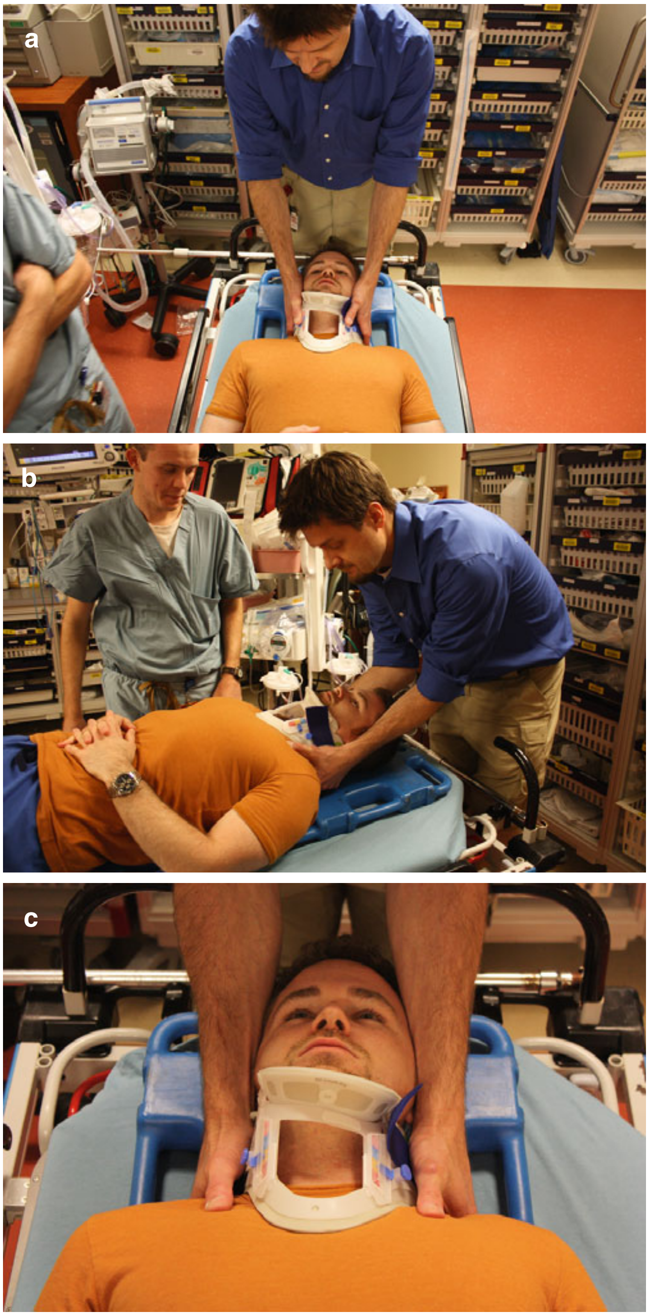

Fig. 38.2 (**a**) Incorrect: Single-point stabilization does not keep head and C-spine in line with torso (**b**, **c**). Correct: The two-point stabilization technique keeps the head and C-spine in line with the torso during the log roll

10. The person at the head of the bed signals the initiation of a synchronized log roll when everyone is in position.
11. Attention should be directed at maintaining a neutral axis along the anatomical line of symmetry during the roll (Fig. 38.3).
12. While the patient is on her or his side, the rigid backboard should be removed and secured to prevent injury of caretakers.
13. The patient's entire posterior should be exposed and examined, taking care to note lacerations and obvious deformities.
14. The entire length of the spine is examined by inspection and then by palpation to assessed for tenderness, step-offs, and deformities (Fig. 38.4).
15. Before returning the patient back onto the stretcher, ensure that any debris, glass, lumps of clothing, or blankets are removed.

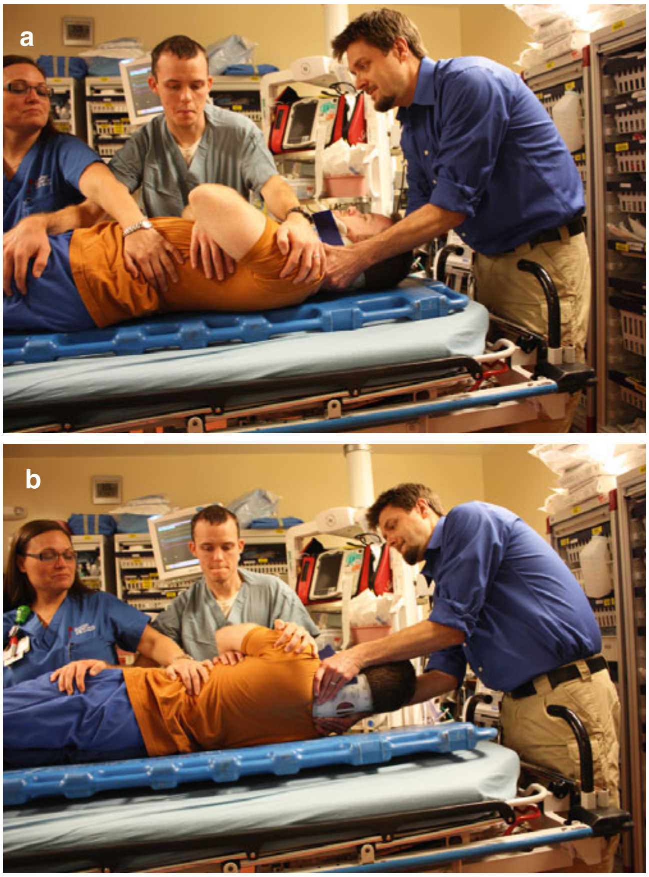

Fig. 38.3 (**a**) The interlocking hands of operators performing lateral log roll help maintain spine immobilization and minimize twisting. (**b**) Two-point cervical spine stabilization is maintained in neutral axis relative to the anatomical line of symmetry during the roll. Appropriate bed height adjustment will increase the ease of the procedure

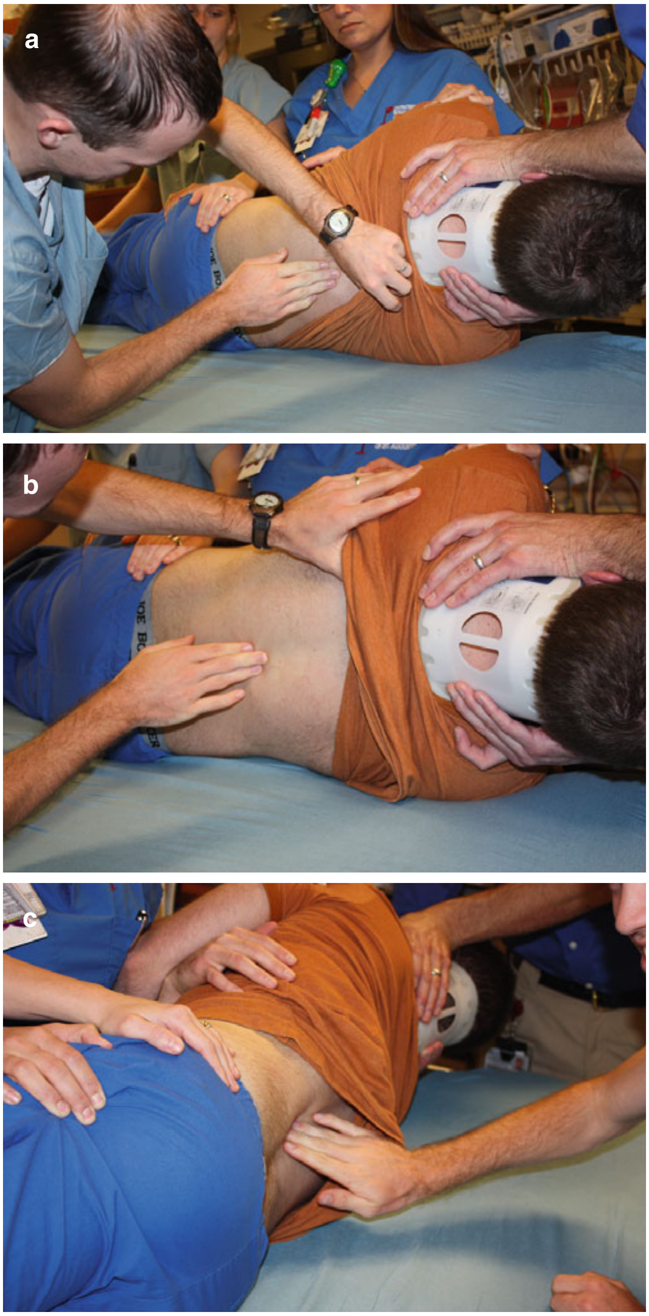

Fig. 38.4 (**a**–**c**) After backboard has been removed, the entire length of the of the spine can be appropriately palpated and assessed for injury to the thoracic and lumbar spine

38.5 Pearls and Pitfalls

- Pearls
  - The log roll should be performed in unison to avoid segmental rotation.
  - The ABCs should be established before initiating the log roll, which is part of the secondary survey in the trauma evaluation.
  - A proper log roll begins with stabilization of the C-spine because it is the most freely mobile part of the spine and, therefore, most frequently injured.
  - Take the patient off the backboard as soon as possible to prevent formation of pressure sores.
- Pitfalls
  - Failure to establish a two-point stabilization between the patient's body and the head leaves the C-spine at risk of further injury during the log roll.
  - Patients should not be left on the backboard for spinal precautions.
  - Failure to ensure that the ABCs are established before attempting a log roll.

Selected Reading

American College of Surgeons Committee on Trauma. Advanced trauma life support for doctors. 8th ed. Chicago: American College of Surgeons; 2008. ISBN 1880696312.

Roberts JR, Hedges JR. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders Elsevier; 2009.

Part VI

Neurologic and Neurosurgical Procedures

Burr Hole Craniotomy

39

Latha Ganti

39.1 Indications

- Extradural hematoma (EDH) or subdural hematoma (SDH) with Glasgow Coma Scale (GCS) <8 and unavailability of timely neurosurgical intervention in the face of increased intracranial pressure
- Signs of increased intracranial pressure [1]
  - Deteriorating level of consciousness
  - Slowing of the pulse
  - Dilating pupils
  - Focal seizures
  - Hemiparesis
  - Extensor posturing of the limbs

39.2 Contraindications

- Absolute
  - GCS >8
  - Immediate availability of neurosurgeon
- Relative
  - Lack of imaging (in this case, decision is guided by neurologic findings and signs of increased intracranial pressure or impending herniation)

39.3 Materials and Medications

- Razor to shave area
- 2 % lidocaine with epinephrine to numb scalp
- 10 % povidone-iodine or chlorhexidine prep
- Light, suction, cautery, dressing tray
- Gelfoam
- Penrose drain
- 3-0 silk
- #10 scalpel blade and #3 handle
- Small self-retractors or rakes
- Drill and drill bits: can be manual (Fig. 39.1) or automatic stopping variety (Fig. 39.2)
- Bone wax or electric cautery apparatus
- Suction apparatus
- Saline irrigation (IV tubing connected to a saline bag with clamp set so flow is low, or a saline syringe flush)

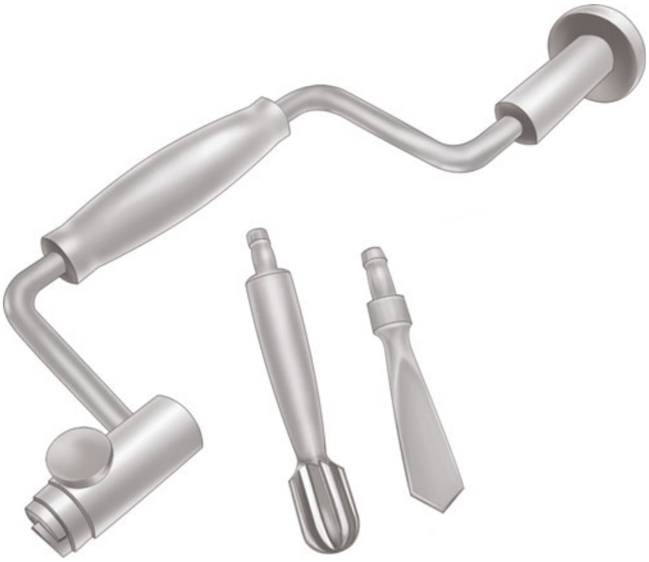

Fig. 39.1 Hudson brace, burr, and perforator

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_39

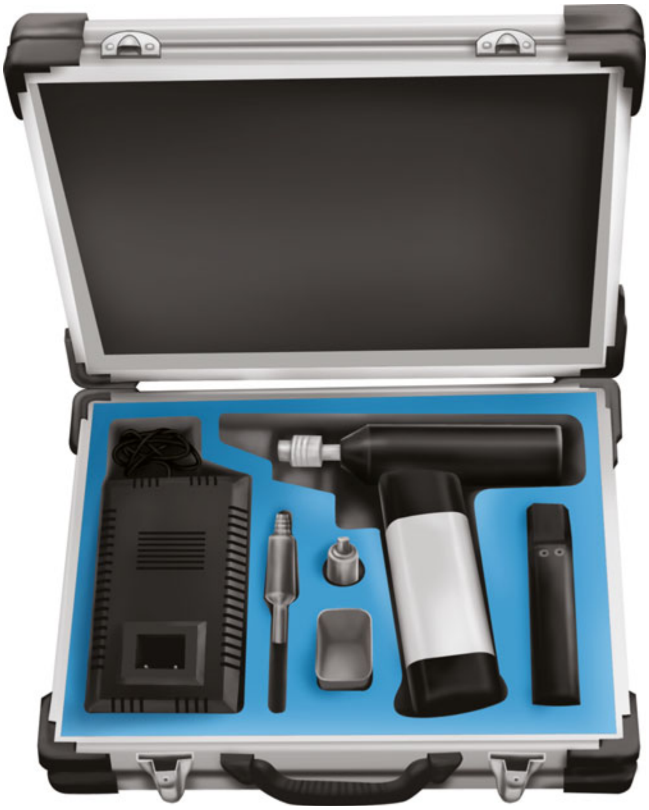

Fig. 39.2 Automatic stopping craniotomy drill set

39.4 Procedure

Preparation

1. Patient should be supine, intubated, with appropriate C-spine precautions
   - Hypertonic saline and/or mannitol can be considered for medical management of increased ICP if/as directed by neurosurgery
2. Determine location and depth for burr hole placement (Figs. 39.3 and 39.4):
   - Have CT images immediately available for viewing
   - Most common location for EDH or SDH is temporal
     - Temporal burr hole placement: 2 fingerbreadths above the ear, 2 fingerbreadths forward (Fig. 39.5)
     - Parietal burr hole placement: 2 fingerbreadths above the ear, 3 fingerbreadths backward (behind the ear)
     - Frontal burr hole placement: 3 fingerbreadths from midline, 3 fingerbreadths above hairline
   - Estimate depth of hematoma by counting the number of slices the hematoma is as seen on CT scan and multiplying by the slice thickness [2]
3. Shave hair around area of hematoma.
4. Prep skin with betadine and chlorhexidine.
5. Anesthetize scalp skin with lidocaine and epinephrine.

   Accessing hematoma
6. Make a vertical incision approximately 4 cm long down to bone.s
7. Use periosteal elevator or end of scalpel blade to scrape muscle and periosteum away from bone.
8. Place self-retaining retractors (or rakes) to keep field open (Fig. 39.6).
9. Using a drill with a clutch mechanism [2], drill through outer table (resistance), diploic space (no resistance), then inner table.
   - Burr hole needs to be placed over the *center* of the hematoma (Fig. 39.7).
10. Control bone bleeding with bone wax; control bleeding from veins and/or muscle with gelfoam sponge or cautery (diathermy).
    - Wound edges may also be cauterized or tamponaded with manual pressure +/– epinephrine gauze.
11. Once in the inner table, separate dura from bone
12. Enlarge opening by switching to a conical or cylindrical burr or use a rongeur

    Evacuating hematoma
13. EDH blood will be visible at this point and should be gently suctioned out.
14. SDH blood will be seen as a tenting of the dura and may be clotted: lift the dura with a hook or make an incision with a fresh scalpel through it to expose the clot and drain.

15. Irrigate with saline (this can be via a hand syringe or via IV tubing connected to a saline bag at low sped flow).
16. Repeat gentle suction.

Closure

17. For temporal burr holes, ligate middle meningeal artery (if visible) or cauterize.
18. Place Penrose drain (Fig. 39.8) and close the dura first with a 3-0 absorbable suture, ensuring there is no dural leak (will lead to infection if present).
    - A watertight seal of the duraplasty is essential to minimize cerebrospinal fluid leakage.
19. Loosely suture scalp using 3-0 silk.

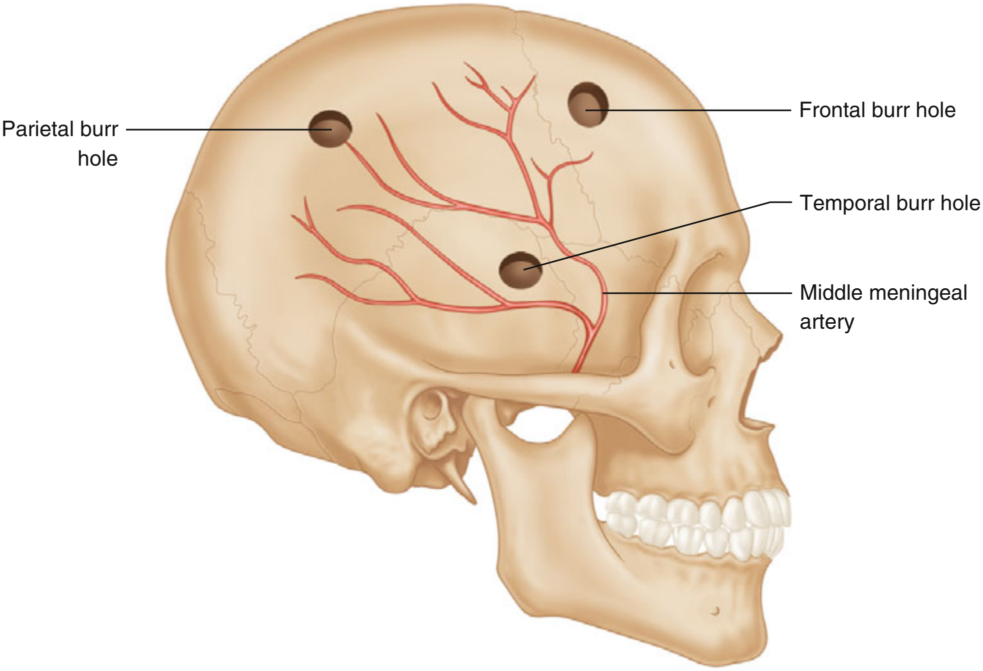
Parietal burr hole
Frontal burr hole
Temporal burr hole
Middle meningeal artery

Fig. 39.3 Positions of burr hole placement

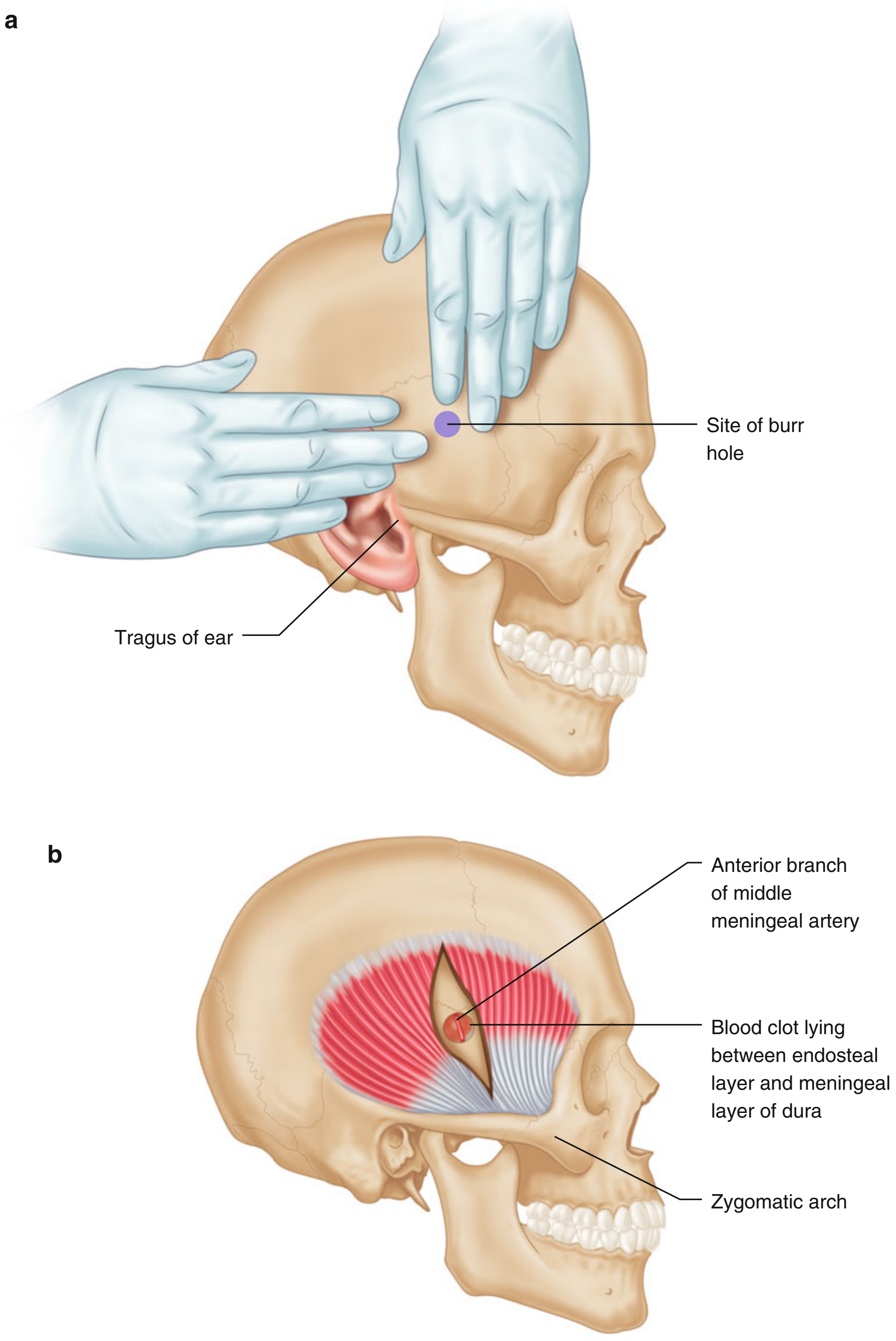
a
Site of burr hole
Tragus of ear
b
Anterior branch of middle meningeal artery
Blood clot lying between endosteal layer and meningeal layer of dura
Zygomatic arch

Fig. 39.4 (**a**, **b**) Anatomy for temporal burr hole placement (most common location)

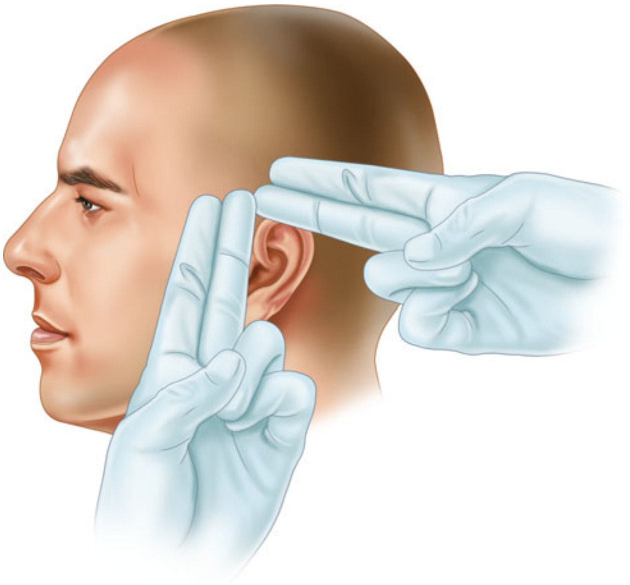

Fig. 39.5 Landmark for temporal burr hole above zygoma, make 5 cm incision over site

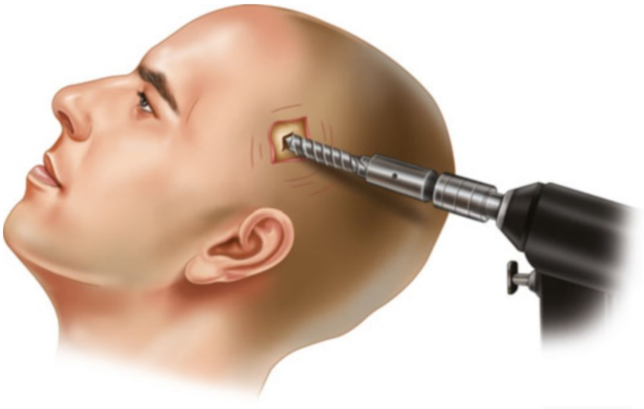

Fig. 39.7 Use penetrator drill to get through outer table. Follow up with a burr to get through the rest of the layers. Can also use an automatic stop craniotomy drill

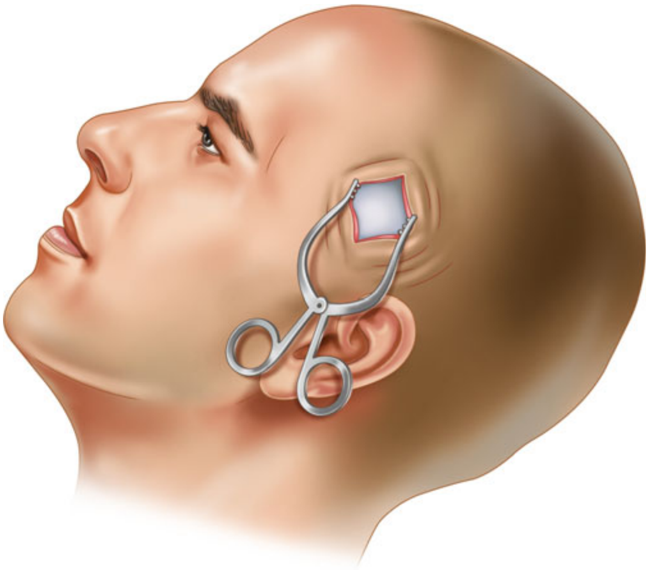

Fig. 39.6 After incision is made, use retractors (preferably self-retaining) to visualize field

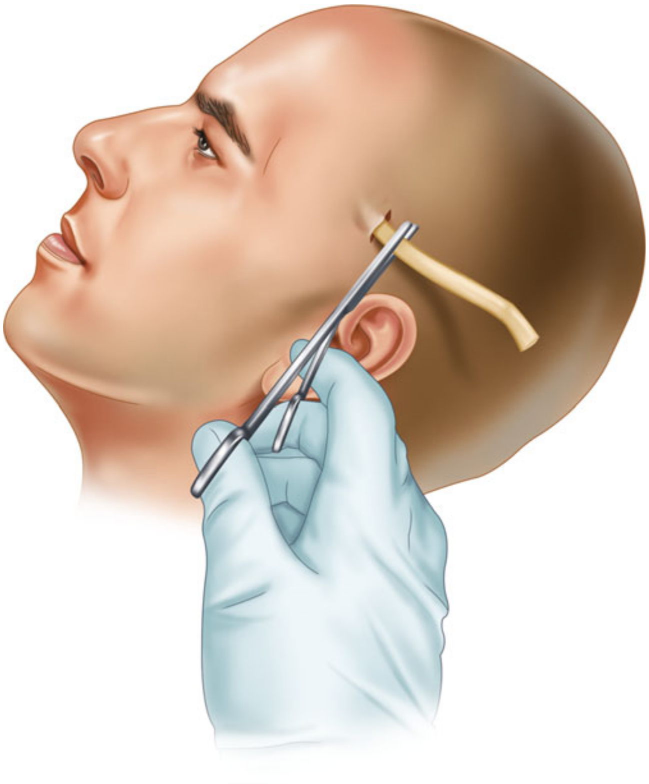

Fig. 39.8 Suture in Penrose drain and close the wound

39.5 Complications

- Brain laceration/perforation
- Temporal artery laceration
- Wrong location (minimized when viewing images during procedure)
- Facial artery laceration
- Meningitis
- Brain abscess

39.6 Pearls and Pitfalls

- Pearls
  - If brain herniates through the burr hole, there is likely a hematoma at another location.
  - Remember this is only a lifesaving procedure that is to be done when timely neurosurgical intervention is not possible. *Do not delay transfer* of the patient
- Pitfalls
  - Bone in temporal area is quite thin; if not using an automatic stop drill, be very careful not to go too far and perforate the brain.

References

1. Wilkinson DA, Skinner MF. The primary trauma care manual for trauma management in district and remote locations. World Health Organization (WHO). http://www.steinergraphics.com/surgical/006_17.6.html. Accessed 22 May 2014.
2. Wilson MH, Wise D, Davies G, Lockey D. Emergency burr holes: "how to do it.". Scand J Trauma Resusc Emerg Med. 2012;20:24. doi:10.1186/1757-7241-20-24.

Selected Reading

http://www.viewmedica.com/vm/pages/library/L_df8516c9#vm_A_ac54d3a1. Accessed 29 June 2014.

External Ventricular Drain Placement

40

Latha Ganti

An external ventricular drain (EVD; also known as a ventriculostomy tube) is placed via a burr hole in one of the ventricles of the brain (Fig. 40.1) to drain excess cerebrospinal fluid (CSF) that causes elevated intracranial pressure.

40.1 Indications

- Emergent need for intracranial pressure (ICP) monitoring and or management
- Enlarged ventricles on neuroimaging with Glasgow coma scale <12
- Subarachnoid hemorrhage with Hunt-Hess grade ≥3 (Table 40.1)
- Coma
- Obstructive hydrocephalus
- Intraventricular hemorrhage
- Signs of increased intracranial pressure

40.2 Contraindications

- Absolute
  - Immediate availability of a neurosurgeon who can do the procedure
- Relative
  - Coagulopathy
  - Scalp infection

40.3 Materials and Medications

- Intubation equipment and medications for sedation as needed
- Sterile gloves, gown, mask
- Ruler
- Surgical marking pen
- 1–2 % lidocaine with epinephrine to numb scalp
- 5 cc syringe and needles to give anesthetic
- 10 % povidone-iodine solution and swabs
- Razor to shave area
- Fenestrated clear drape
- Scalpel #11 blade (for scalp) and #15 blade (for periosteum) with #3 handle (Fig. 40.2a)
- 4×4 sterile gauze
- Adson forceps
- Mosquito forceps
- Self-retaining eyebrow retractor (Fig. 40.2b)
- 3-0 nylon suture and needle holder and/or skin stapler
- Scissors
- A hand drill with variable chuck
- One or more drill bits with depth guards in 5/32″ (3.97 mm), 13/64″ (5.31 mm), and 1/4″ (6.35 mm) sizes
- A hex wrench for depth guard adjustment
- Ventricular catheter
- Primed ventricular drainage collection system

40.4 Procedure

Patient preparation

1. Patient is intubated and placed supine in neutral position with head of the bed elevated 30–45°.
2. Administer one dose of intravenous antibiotics that covers skin flora.
3. Make precise measurement of where the hole and incision will be made. Most commonly, EVDs are placed in

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_40

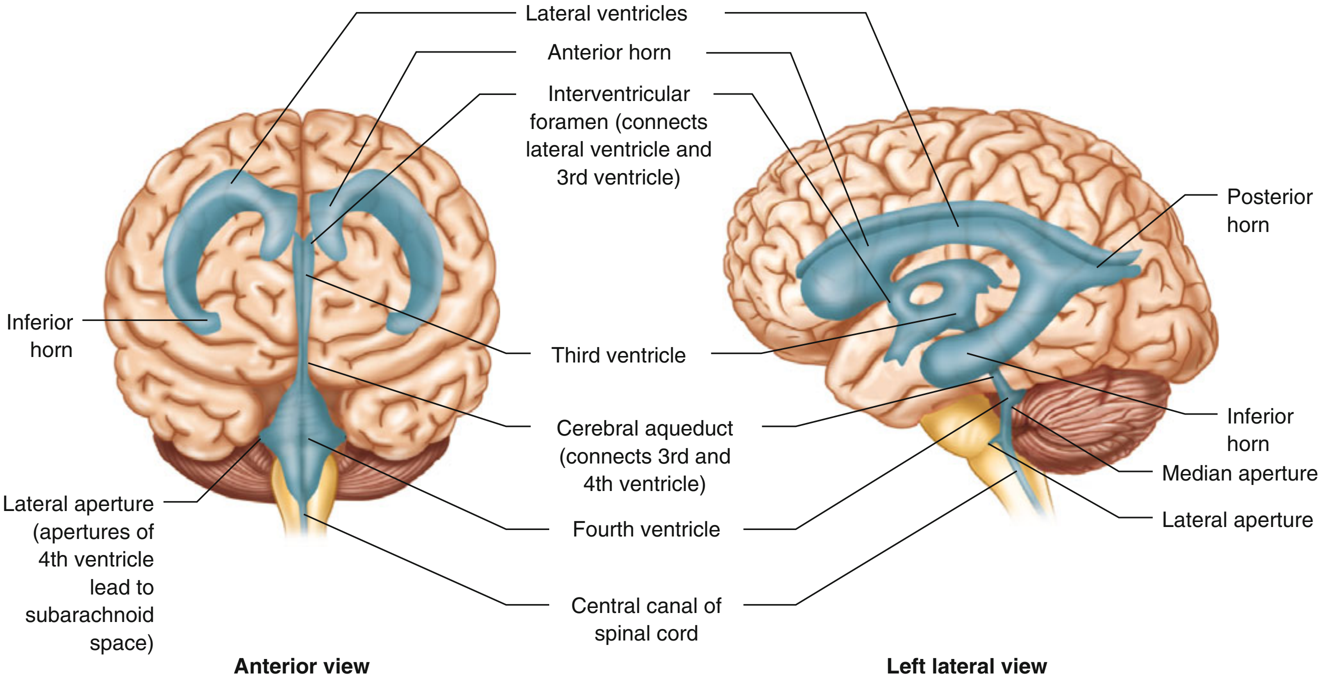
Lateral ventricles
Anterior horn
Interventricular foramen (connects lateral ventricle and 3rd ventricle)
Inferior horn
Third ventricle
Cerebral aqueduct (connects 3rd and 4th ventricle)
Fourth ventricle
Central canal of spinal cord
Lateral aperture (apertures of 4th ventricle lead to subarachnoid space)
Posterior horn
Inferior horn
Median aperture
Lateral aperture
Anterior view
Left lateral view

Fig. 40.1 Ventricles of the brain that contain CSF

Table 40.1 Hunt and Hess scale for subarachnoid hemorrhage [1]

Grade	Signs and symptoms
1	Alert and oriented, mild headache, slight or no nuchal rigidity
2	Alert and oriented, moderate to severe headache, nuchal rigidity, no neurologic deficit
3	Drowsiness, confusion, mild focal neurologic deficit
4	Stupor, moderate-severe hemiparesis
5	Coma, decerebrate posturing

the right frontal scalp, as this is the nondominant hemisphere in >95 % of the population.

4. Using a sterile skin/surgical marker, draw a line 11 cm back from nasion and then a point 3 cm to the right of that, which corresponds to the mid-pupillary line. This is called Kocher's point (Fig. 40.3).
5. Once Kocher's point is marked, shave skin and remove loose hairs so field is wide enough for subcutaneous catheter placement and tunneling.
6. Prep area with povidone-iodine.
7. Place sterile clear drape over field.
8. Prep area of incision once more.
9. Infiltrate scalp with 1–2 % lidocaine with epinephrine (the epinephrine acts as a hemostatic agent, keeping surgical field clean) (Fig. 40.4).
10. Make a 2 cm linear stab incision and extend incision to skull (Fig. 40.5).
11. Use eyebrow self-retaining retractors to hold skin edges back.
12. Drill burr hole with hand drill (Fig. 40.6).
13. Irrigate burr hole with sterile saline.
14. Put on a new pair of sterile gloves (prior to handling ventriculostomy catheter).
15. Remove ventricular tubing from sterile package. It has markings on it that are 1 cm apart.
16. Insert ventriculostomy catheter perpendicular to the skull at the point of insertion. Catheter is aimed at the ipsilateral medial canthus of the eye (anteroposterior plane) and tragus of ear (lateral plane).
17. Advance ventriculostomy tube 5–6 cm with stylet from outer skull table so it sits in the anterior horn of the lateral ventricle (Fig. 40.7).
18. Ensure ventriculostomy catheter is draining CSF.
19. Attach metal trocar to tip of ventriculostomy catheter and tunnel the trocar and catheter under the galea approximately 3–5 cm to the right of the original incision. Bring trocar out through separate stab incision in scalp.
20. Remove trocar, make sure ventriculostomy catheter is still draining CSF.
21. Place temporary cap on ventriculostomy catheter to prevent overdrainage of CSF.

22. Close the original incision with either sutures or staples.
23. Secure catheter to the scalp using staples.

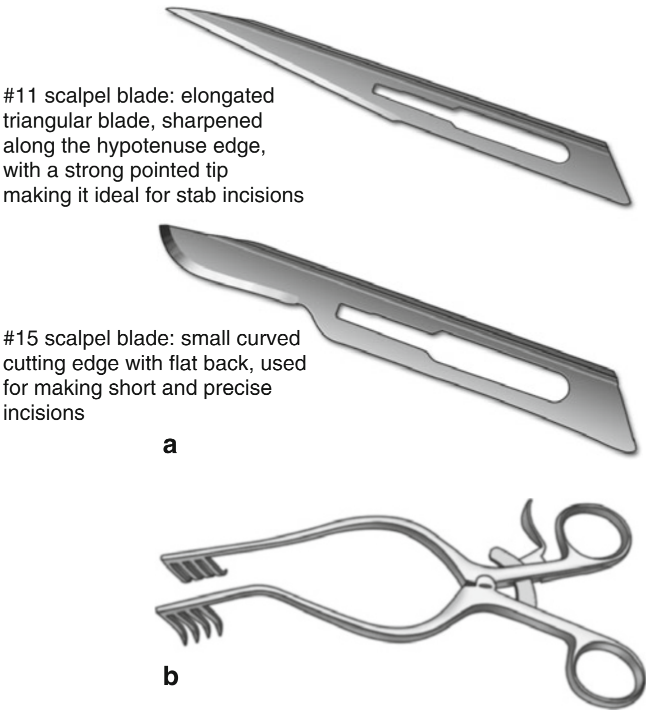
#11 scalpel blade: elongated triangular blade, sharpened along the hypotenuse edge, with a strong pointed tip making it ideal for stab incisions
#15 scalpel blade: small curved cutting edge with flat back, used for making short and precise incisions
a
b

Fig. 40.2 (**a**) Surgical scalpel blades. (**b**) Self-retaining eyebrow retractors

24. Cover incision with sterile transparent dressing.
25. Suture stopcock to ventricular catheter to ensure it does not come off (take care not to make suture too tight or it will occlude ventriculostomy catheter).
26. Remove blind end of stopcock, attach ventricular catheter to ventricular drain system once it is primed.

Priming, setup, and maintenance of ventricular drainage system

27. Prepare drainage system by priming system with sterile, preservative-free saline.
28. Place zero point of drainage system at midbrain (level of patient's ear tragus) or above midbrain at 15–20 cm of H_20 (target ICP set by neurointensivist and will change depending on patient's situation) (Fig. 40.8).
29. Attach drainage tubing.
30. Patients who require an EVD should be closely monitored by nurses trained and competent in assessment and management of the drain and in recognizing signs of increased ICP in the patient.
31. Assessment of the drainage system should be done a minimum of every 4 h, which includes inspecting the EVD from the insertion site along the entire drainage system, checking for cracks in the system or fluid leaking from the insertion site [2].

Post-procedure

32. Obtain CT scan of brain to verify placement (Fig. 40.9).

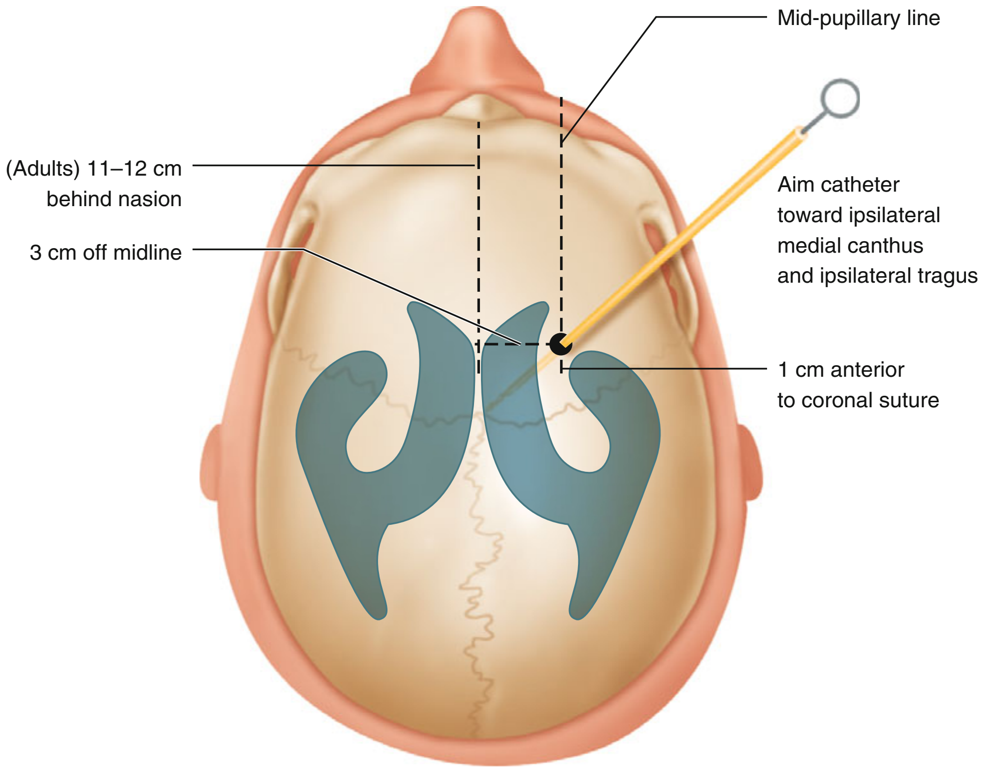
Mid-pupillary line
(Adults) 11–12 cm behind nasion
3 cm off midline
Aim catheter toward ipsilateral medial canthus and ipsilateral tragus
1 cm anterior to coronal suture

Fig. 40.3 Kocher's point for frontal EVD placement

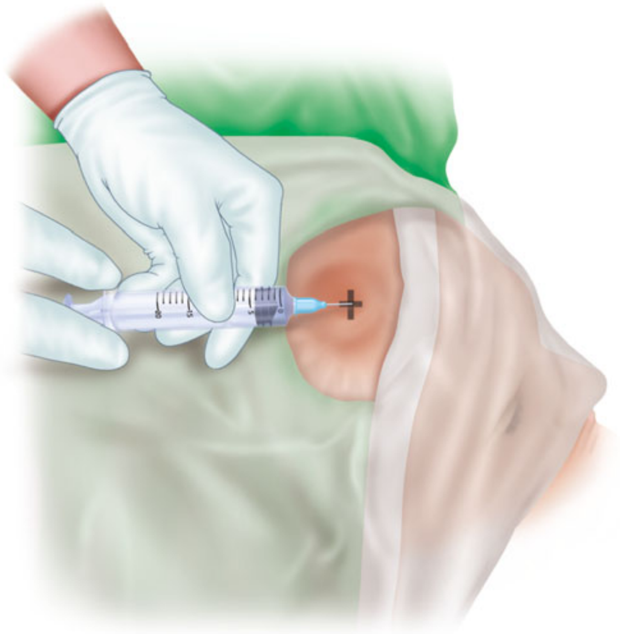

Fig. 40.4 Anesthetize skin

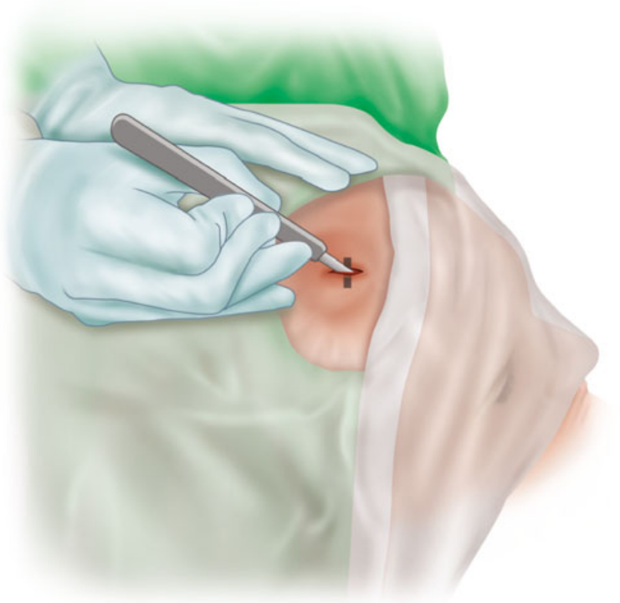

Fig. 40.5 Make stab incision

40.5 Complications

- Hemorrhage
- Infection: meningitis, ventriculitis
- Tract hematoma
- Migration/dislodgement

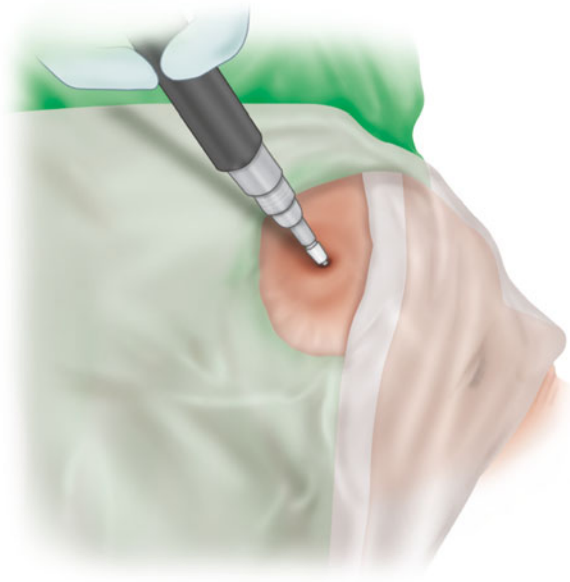

Fig. 40.6 Make burr hole

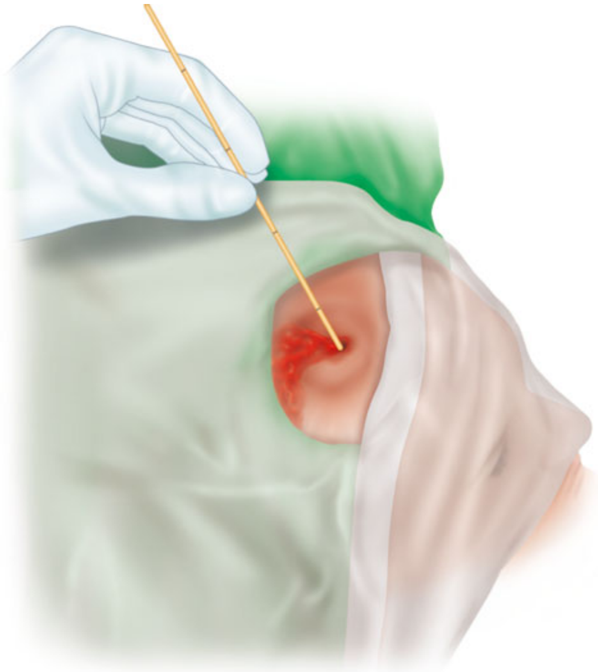

Fig. 40.7 Advance ventriculostomy catheter with stylet

40.6 Pearls and Pitfalls

- Pearls
  - EVD needs to be re-leveled every time the patient moves. Family and all visitors must be informed that any patient movement or change in elevation of head

Fig. 40.8 Zero point of drainage system placed at midbrain (level of patient's ear tragus)

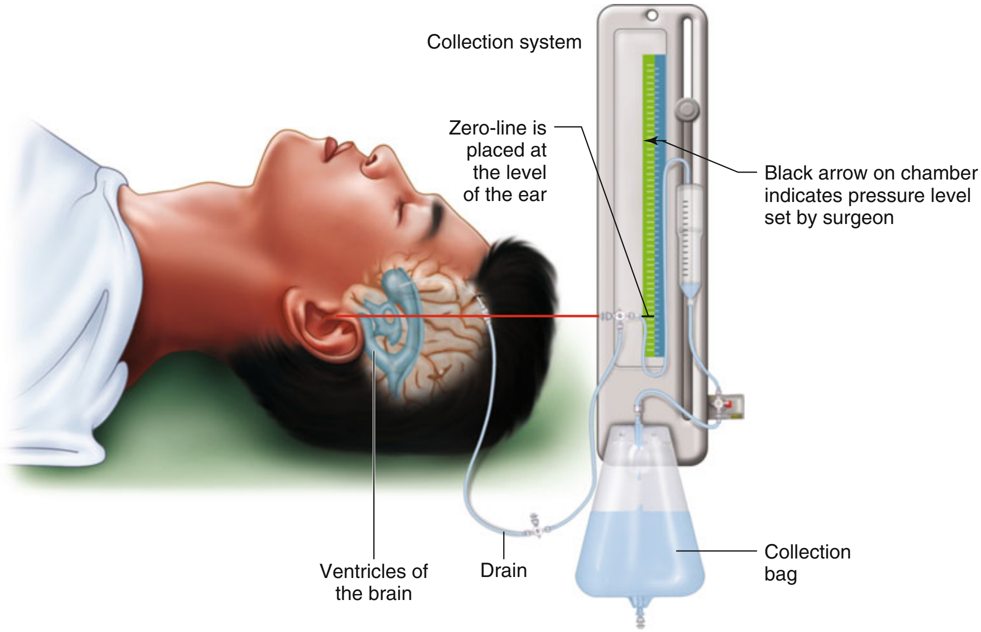
Collection system
Zero-line is placed at the level of the ear
Black arrow on chamber indicates pressure level set by surgeon
Ventricles of the brain
Drain
Collection bag

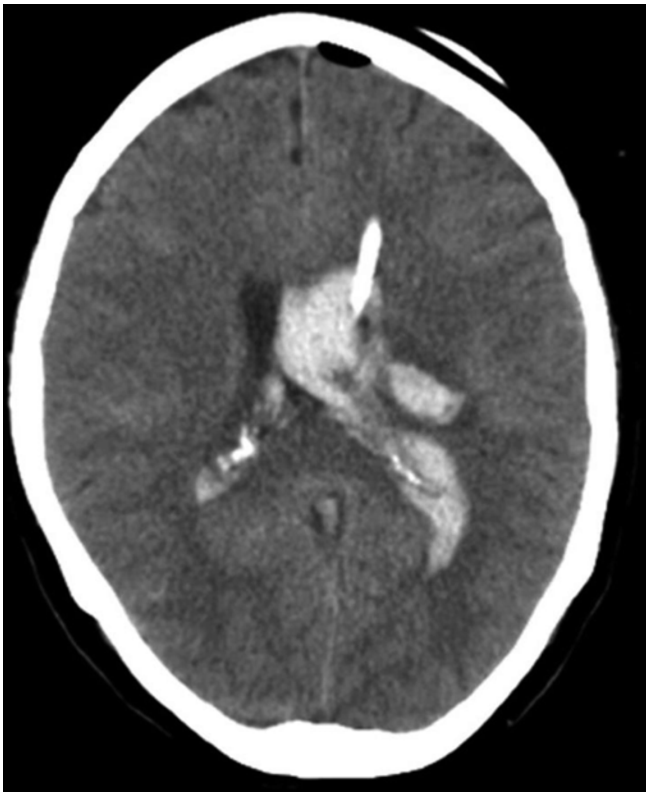

Fig. 40.9 CT scan demonstrating typical placement of ventriculostomy catheter in ipsilateral anterior horn of lateral ventricle [3] (Reproduced with permission from National Institutes of Health: Jaffe et al. [3])

of bed will require re-leveling of EVD in order to ensure it is a the appropriate level for CSF drainage.
  - CSF collection chamber should remain upright to prevent reflux/leakage.
  - EVD placement in the parieto-occipital scalp or frontotemporal or trans-sylvian locations is generally not done at the bedside due to higher rate of complications. This is performed in the operating room.
- Pitfalls
  - If the drain is leveled too low: CSF will drain too easily, which can lead to re-rupture of an unrepaired ruptured cerebral aneurysm or cause a subdural hematoma due to shearing of bridging veins.
  - If the drain is leveled too high, CSF will not drain or will not drain at desired rate, leading to hydrocephalus. The ICP will need to be higher in order for the CSF to drain.

References

1. Hunt WE, Hess RM. Surgical risk as related to time of intervention in the repair of intracranial aneurysms. J Neurosurg. 1968;28(1):14–20.
2. Care of the patient undergoing intracranial pressure monitoring/external ventricular drainage or lumbar drainage. AANN clinical practice guideline series. 2011. http://www.aann.org/uploads/AANN11_ICPEVDnew.pdf. Accessed 27 Sept 2014.
3. Jaffe J, Melnychuk E, Muschelli J, et al. Ventricular catheter location and the clearance of intraventricular hemorrhage. Neurosurgery. 2012;70(5):1258–64.

Lumbar Puncture in Adults

41

Kevin Tench, L. Connor Nickels, and Rohit Pravin Patel

41.1 Indications

- Diagnostic
  - Evaluation for the possibility of a central nervous system (CNS) infection: viral, bacterial, and fungal meningitis and encephalitis
  - Evaluation for inflammatory processes: multiple sclerosis, Guillain-Barré syndrome
  - Evaluation for spontaneous subarachnoid hemorrhage (SAH)
  - Suspicion of CNS diseases: oncological and metabolic processes
- Therapeutic
  - Therapeutic reduction of cerebrospinal fluid (CSF) pressure
  - Procedures requiring lower body analgesia or anesthesia
  - Intrathecal antibiotic administration for some types of meningitis
  - Chemotherapy and methotrexate for some forms of leukemia and lymphomas

41.2 Contraindications

- Presence of infection in tissues at or around puncture site.
- Increased intracranial pressure (ICP) from a space-occupying lesion; patients with signs of cerebral herniation or with potential of increased ICP and focal neurological signs.
- Bleeding diathesis (thrombocytopenia, anticoagulant therapy, hemophilia); may increase risk of spinal hematoma, but level of coagulopathy that increases risk is unclear.
- Patients with cardiorespiratory compromise may worsen owing to position.

Patients with prior history of lumbar surgery, osteoarthritis, ankylosing spondylitis, kyphoscoliosis, or degenerative disk disease might have more success if lumbar puncture is performed by an interventional radiologist using imaging techniques and should be considered.

41.3 Materials and Medications (See Fig. 123.1)

- Spinal needle(s) with stylet
  - Adults: 3.5-in. 20-gauge needle; obese may require 5.0-in. 22- to 24-gauge needle
  - Children: 2.5-in. 22-gauge needle
  - Infants: 1.5-in. 22-gauge needle
- Three-way stopcock (optional: drainage catheter)
- Manometer (optional: extension tube for higher opening pressures)
- Specimen tubes (# may vary, but in general labeled 1–4, important to obtain from 1, 2, 3, 4 owing to cell count obtained from tubes 1 and 3)
- Local anesthetic (lidocaine 1 or 2 %), 5- to 10-mL syringe and needle (25-gauge) for local anesthesia
- Sterile drapes and gauze
- Mask, sterile gown, sterile gloves
- Antiseptic solution for skin preparation (Chloroprep or iodine)

K. Tench, MD
Department of Emergency Medicine, Banner Boswell Medical Center, Sun City, AZ, USA

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_41

41.4 Procedure

1. Positioning
    - Determined by practitioner preference or patient capability.
    - Options: lateral recumbent position, upright sitting position (Fig. 41.1).
    - Lateral recumbent position is preferred to obtain accurate opening pressure and to reduce the risk of post-puncture headache.
    - Both positions require the patient to arch the lower back toward the practitioner in order to open up the intervertebral spaces (obtain the "fetal position" or arch "like a cat").
    - Shoulders and hips should remain aligned during process.
2. Landmarks
    - Determined by palpation.
    - Draw a visual line between the superior aspects of the iliac crests that intersects the midline at the L4 interspace. The L3–4 and L4–5 spaces are preferred because these points are below the termination of the spinal cord.
    - Palpate the posterosuperior iliac crests with the midpoint of a visual line that connects the two crests representing the L4 spinous process.
    - Palpate the space between the L3–4 or the L4–5 spinous processes and mark where the needle will be placed.
3. Ultrasound guidance (optional)
    - Helpful in obese patients, patients with previous surgical scarring, or anyone in whom palpation of the spinous processes is not easily done.
    - Sonographic measurement of the dura mater strongly correlates with needle depth needed to obtain CSF.
    - Identify the spinal process in the short and long axis to determine the midline and the interspinous space.
    - Identify the interspinous ligament for estimation of the depth of needle insertion.

      Commonly only the spinous processes are well visualized, and the interspinous ligament, ligamentum flavum, and subarachnoid space are less clearly seen.
    - High-frequency (5–10 MHz) linear probe to best evaluate anatomy.
    - A marking pen can be used to create a cross-hair-type figure (Fig. 41.1).
    - After placing the patient as described, locate the midline at the lumbar spine in transverse and longitudinal orientations.
    - Bright echogenic structures with shadowing posteriorly identify the spinous processes.
    - Transverse probe positioning to identify midline (Fig. 41.2), and then longitudinal probe positioning to identify interspinous space (Fig. 41.3).
4. Sterile preparation
    - After positioning and palpating the appropriate landmarks, the practitioner should then dress in the appropriate protective gear: mask, gown, and sterile gloves.
    - After dressing, the practitioner can then sterilely prepare the patient.
    - – Make sure the patient's back is completely exposed.
    - – Clean the patient's back with an antiseptic solution (Chloroprep should be scrubbed in an up/down and side/side fashion; iodine in a circular motion starting from the center of the anticipated insertion point).
    - – Apply sterile drapes with the puncture site exposed.

      This is an optimal time to make sure all equipment in a standard lumbar puncture tray connects properly and that the stopcock for opening pressure measurement is assembled. Make sure the stopcock is closed away from the patient so that CSF can flow from the patient to the manometer. If the assembly is done, it will decrease the amount of CSF lost after the puncture.
- Local anesthesia
    - – 1 % Lidocaine or anesthetic cream topically before preparing skin
    - – For injection, form a skin wheal over the insertion site.
    - – Inject into the deep tissues below the wheal in all directions while only breaking the skin once.
    - – Systemic sedatives and analgesics may also be used.

5. Needle insertion
    - Needle should be inserted in the midline between the L3–4 or the L4–5 spinous process, and the stylet should be firmly in place.
    - Initially parallel to the bed, but once into the subcutaneous tissue, the needle should be angled toward the umbilicus (slightly cephalad, 15°) with the bevel facing upward (Fig. 41.4). This sagittal plane orientation spreads rather than cuts the fibers of the dural sac, which run parallel to the spinal axis.
    - If properly positioned, the needle passes through the skin; subcutaneous tissue; supraspinous ligament; interspinous ligament between the spinous processes; ligamentum flavum; epidural space including the internal vertebral venous plexus, dura, and arachnoid; into the subarachnoid space and between the nerve roots of the cauda equina.
    - In most cases, a "pop" will be felt when the needle penetrates the ligamentum flavum, entering into the subarachnoid space; then intermittent withdrawal should be done in 2-mm intervals to assess for CSF flow.
    - If bone is encountered during insertion, the needle should be withdrawn partially without exiting the skin and readjusted to a different angle more cephalad.

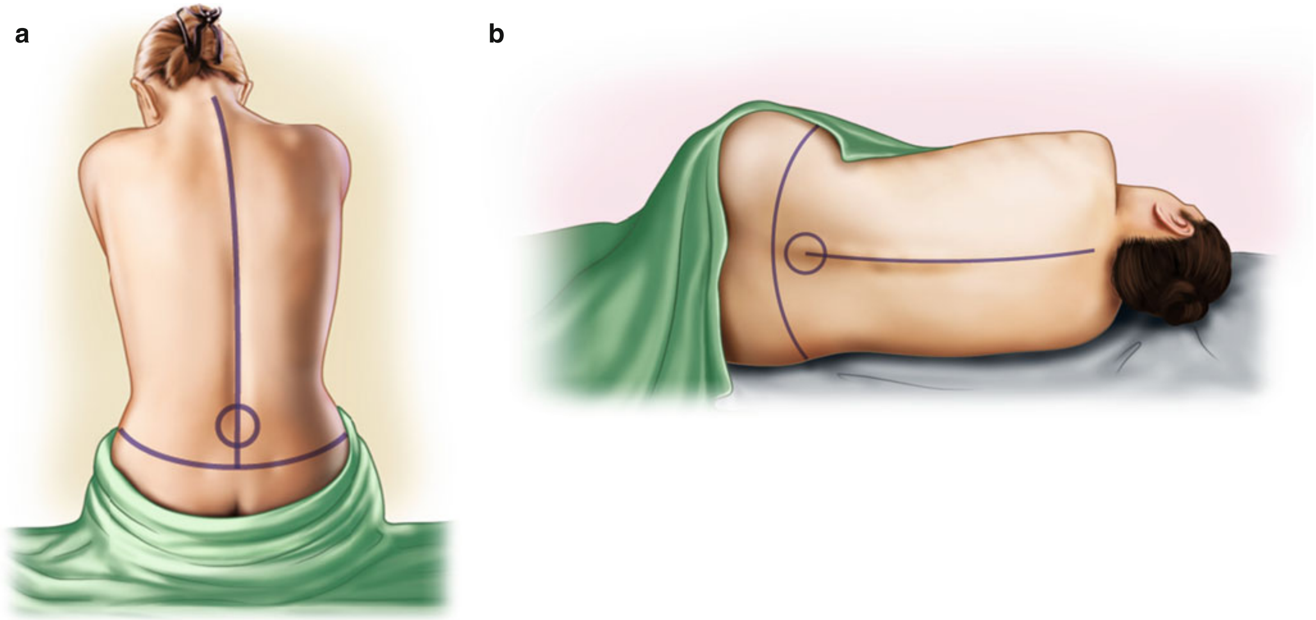

Fig. 41.1 Sitting position (**a**) left lateral decubitus position, (**b**) with general areas of insertion of needle

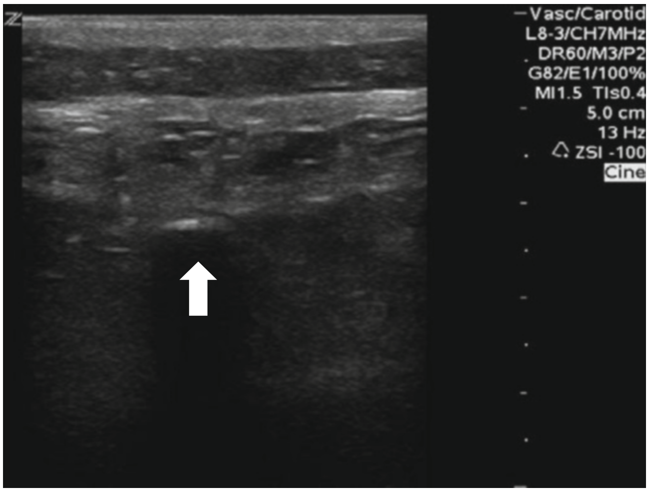

Fig. 41.2 Transverse view, *white arrow* indicates spinous process. Place in middle of ultrasound view to locate appropriate midline access point

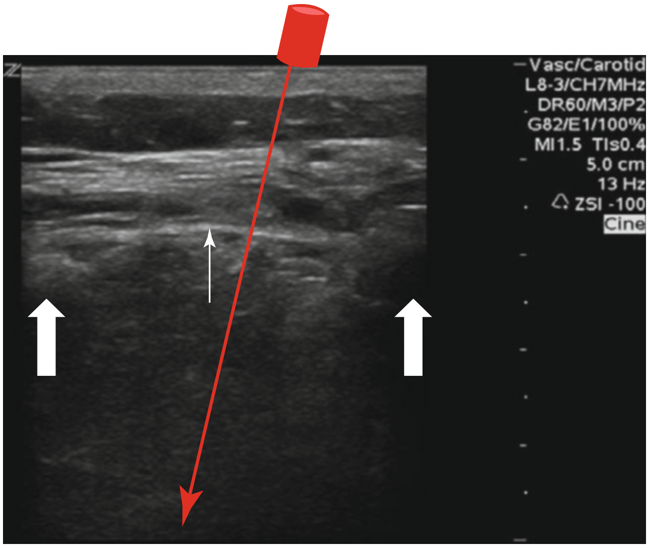

Fig. 41.3 Longitudinal view of lumbar spine. *Thick white arrows* indicate vertebral shadows; *thin white arrow* indicates supraspinous ligament. *Red arrow* indicates trajectory of needle

- If the tap is traumatic, CSF may be blood tinged but should clear as more is collected. If it does not clear, it may indicate intracranial hemorrhage or subarachnoid blood. Also in traumatic patients, clotting will be seen in the tubes; clotting does not occur in SAH owing to defibrinated blood being present in the CSF. Blood-tinged CSF can also be seen in herpes simplex virus (HSV) encephalitis.
- A dry tap is usually due to incorrect positioning and misdirection of needle, often due to a superior direction of the needle with obstruction by the lamina or spinous process of the superior or inferior vertebra. If the needle is too lateral, an inferior or superior articular process may be hit.

  If flow slows down, rotate the needle 90° because a nerve root may be obstructing the opening.

6. Opening pressure measurement
   - Must be performed in the lateral recumbent position. Although there are some conversion formulas from the sitting position, these are not standard of care.

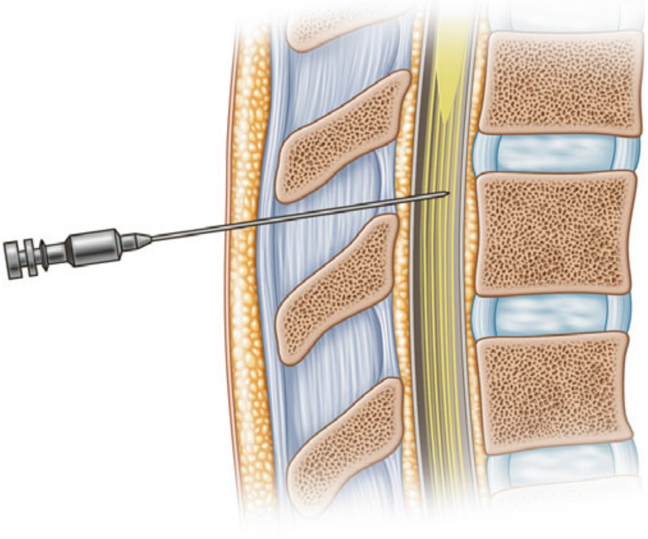

Fig. 41.4 Angle of insertion of needle, cephalad with bevel up

- Once the needle is in the subarachnoid space and CSF is flowing from the needle, the three-way stopcock should be attached to the needle and the manometer should be attached to the stopcock to take a measurement. Use the flexible tube to connect the manometer to the hub of the needle.
- Note the height of fluid in the manometer after it stops rising (normal opening pressure, <20 cm Hg); it may be possible to see pulsations from cardiac or respiratory motion.
  - Elevated CSF pressure is seen with meningeal inflammation, hydrocephalus, pseudotumor cerebri, SAH, and CHF.
  - Decreased CSF pressure is seen in leakage of CSF and severe dehydration.

7. Collecting CSF fluid
   - Collect at least 1–2 mL of CSF fluid in each tube, going from 1 to 4 and never aspirate because this can cause hemorrhage.
   - After collecting the fluid, replace the stylet and remove the needle, clean the skin, and place a bandage over the puncture site.
   - General recommendations
     - Tube 1: glucose, protein, protein electrophoresis
     - Tube 2: Gram stain, bacterial and viral cultures
     - Tube 3: cell count and differential
   - When ruling out SAH, cell count should be performed in tubes 1 and 3 or 1 and 4 to differentiate between SAH and traumatic tap.
     - Tube 4: Any special tests: myelin basic proteins, lactate, pyruvate, and smear on cell concentrates all depend on suspicion.

41.5 Complications

- Implantation of epidermoid tumors: from introducing skin plug into the subarachnoid space and can be avoided by using stylet when advancing.
- Postlumbar puncture headache: most common, occurring in 36.5 % of patients within 48 h
- CSF leak: causes headache when CSF leak through puncture site exceeds rate of production
- Bleeding: most common in patients with bleeding diathesis; may result in spinal cord compression
- Epidural hematoma
- Infection: local cellulitis, abscess (local or epidural), or meningitis
- Herniation syndromes: high risk can be identified by computed tomography but may not completely identify all patients with increased ICP
- Backache: local or referred pain
- Cardiorespiratory compromise

41.6 Pearls and Pitfalls

- Pearls
  - Positioning the patient is key to a successful procedure.
  - In adults the spinal cord may terminate higher than previously thought and it may be okay to go one interspace higher than recommended; but in infants owing to the differential in longitudinal growth of the spinal canal and cord, the spinal cord usually ends in L3. So in children the tap must go L4–5 or L5–S1.
  - Always keep the stylet in place until after the skin barrier is penetrated because this will avoid introduction of epidermoid tissue.
- Pitfalls
  - Postspinal headaches can be avoided with smaller needles and intravenous (IV) fluids.
    - Having the patient lie on the back for 1 h after the procedure has no change in incidence of headache.
    - Treatment consists of IV fluids initially, then caffeine, and, ultimately, if the headache persists, a blood patch.

Selected Reading

Boon JM, Abrahams PH, Meiring JH, Welch T. Lumbar puncture: anatomical review of a clinical skill. Clin Anat. 2004;17:544–53.

Ellenby MS, Tegtmeyer K, Lai S, Braner DA. Videos in clinical medicine. Lumbar puncture. N Engl J Med. 2006;355:e12.

Ferre RM, Sweeney TW, Strout TD. Ultrasound identification of landmarks preceding lumbar puncture: a pilot study. Emerg Med J. 2009;26:276–7.

Peterson MA, Abele J. Bedside ultrasound for difficult lumbar puncture. J Emerg Med. 2005;28:197–200.

Reflex Eye Movements (Doll's Eyes and Caloric Testing)

42

Thomas T. Nguyen, Tina Dulani, and Saadia Akhtar

42.1 Doll's Eyes (Oculocephalic Reflex Testing)

42.1.1 Indications: Doll's Eyes

- To assess brain stem function of a comatose patient
- To assess cerebral function in a comatose patient if brainstem function is intact

42.1.2 Contraindications: Doll's Eyes

- Absolute
  - Occult cervical spine injury; rule out radiographically and clinically
  - Basilar skull fracture
- Relative
  - Rheumatoid arthritis; increased risk of atlantoaxial subluxation resulting in spinal cord compression
  - Osteoporosis; increased risk of cervical spine injury
  - Cervical spine ankylosis; increased risk of cervical spine injury

42.1.3 Procedure: Doll's Eyes (Fig. 42.1)

1. Stand at the head of bed and grasp the patient's head with both hands.
2. Use the thumbs of both of hands to open the patient's eyelids.
3. Rapidly move the patient's head to one side and hold.
4. Simultaneously observe for the presence or absence of horizontal movements.
   (a) Head movement to one side should result in conjugate eye movement to the opposite side and then in spontaneous return of the eyes to the midline (normal test).
   (b) Normal oculocephalic reflex is the observation of conjugate eye movements to the opposite side of head turning. This indicates a functionally intact brainstem in a comatose patient.
   (c) Abnormal oculocephalic reflex: incomplete or absent horizontal eye movements. The eyes remain in the midline. This indicates impairment of the brainstem; caloric testing should be done if not contraindicated.
   (d) A partially abnormal oculocephalic reflex: conjugate eye movement opposite to head turning but does not return to the midline means the brainstem is intact but cerebrum function is not.
5. Repeat by rotating the head to the opposite side.
6. Vertical oculocephalic response can be tested by moving the patient's head up and down. A compensatory vertical eye movement should be observed.
   (a) This test is useful only if the horizontal oculocephalic reflex is negative. An intact vertical oculocephalic reflex with a negative horizontal oculocephalic reflex suggests a pontine lesion.
7. Document the observations.

42.1.4 Pearls and Pitfalls: Doll's Eyes

- Oculocephalic reflex may not be present in the first 10 days of life and is unreliable until 2 years of age.
- Do not attempt the doll's eye maneuver in patients with cervical spinal injuries.

T.T. Nguyen, MD (✉) • S. Akhtar, MD
Department of Emergency Medicine, Mount Sinai Beth Israel, New York, NY, USA
e-mail: tnguyen@chpnet.org

T. Dulani, MD
Department of Emergency Medicine, New York Methodist Hospital, Brooklyn, NY, USA
e-mail: tdulani@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_42

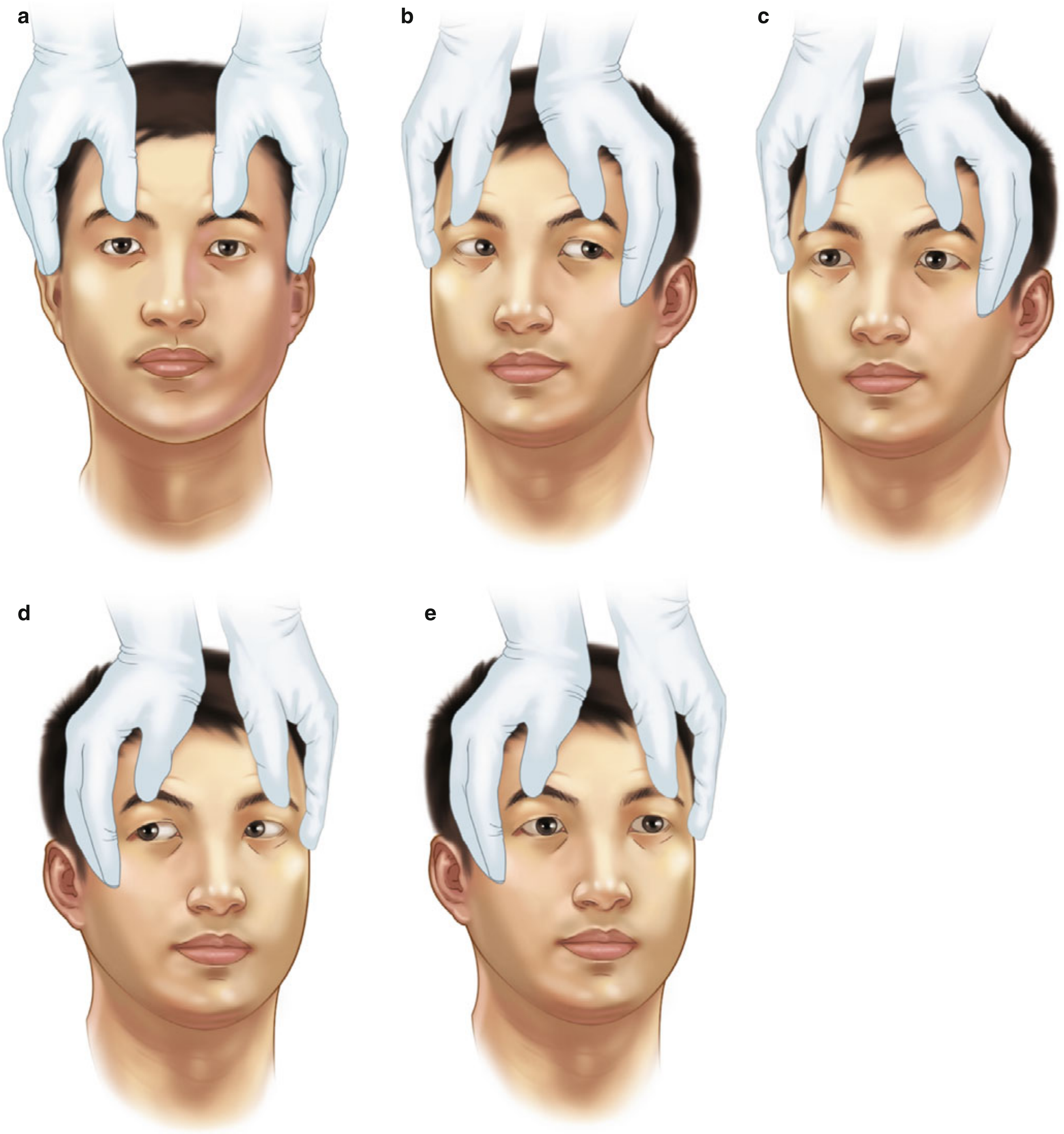

Fig. 42.1 The oculocephalic (doll's eyes) reflex in a patient with an intact brainstem. (**a**) Start with the head facing upright and grasp as depicted using both thumbs to keep the eyes open. (**b**) Rotate the head 90° to the right and the eyes deviate to the left (opposite side). (**c**) The eyes will spontaneously return to the midline. (**d**) Then rotate the head 180° to the left and the eyes should deviate to the right (opposite side). (**e**) The eyes will spontaneously return to the midline

- The doll's eye reflex can be absent or partial in patients with ocular muscle nerve palsy (e.g., cranial nerve [CN] 6).
- Make sure the patient does not have a neuromuscular blockade agent or other toxins present.
- A conscious person can suppress the doll's eye or oculocephalic reflex.

42.2 Caloric Reflex Testing (Vestibuloocular Reflex)

42.2.1 Indications: Caloric Reflex Testing

- In any comatose patient with abnormal doll's eye or if it cannot be performed
- To assess brainstem function of a comatose patient
- To assess asymmetrical function in the peripheral vestibular system

42.2.2 Contraindications: Caloric Reflex Testing

- Absolute
  - Perforated tympanic membrane
  - Presence of tympanostomy tubes
  - Basilar skull fracture, petrous bone fracture
  - Suspicion of cerebrospinal fluid (CSF) otorrhea

42.2.3 Materials and Medications: Caloric Reflex Testing

- Otoscope
- 60-mL syringe
- 16- to 18-gauge angiocatheter
- Thermometer
- Ice-cold water (30–33 °C)
- Warm water (44 °C)
- Emesis basin (to collect water)
- Towels or Chux

42.2.4 Procedure: Caloric Reflex Testing (Fig. 42.2)

1. Place the patient supine and elevate the head to 30° to bring the lateral semicircular canal into the upright position.
   (a) Careful otoscopic examination should be done before caloric testing to make sure there are no contraindications.
   (b) Remove any cerumen from the external auditory canal (EAC); the irrigation fluid should be able to reach the tympanic membrane.
2. Get a 60-mL syringe with an 18-gauge angiocatheter. Remove the plastic angiocatheter from the needle and discard the needle.
3. Place the angiocatheter on the 60-mL syringe and fill it with ice water (30 °C).
4. Place the angiocatheter in the EAC and squirt the water in over a period of 30–40 s. The water should freely enter and exit the auditory canal.
   (a) The stimulus depends upon the temperature of the water and not on water pressure.
   (b) Reflex horizontal movements may be delayed for up to 1 min after irrigation of the EAC.
5. Observe: Have an assistant hold the eyelids open with the head still and facing forward.
   (a) Normal test: Cold water induces slight deviation of the eyes to the side being irrigated followed by a latent period of approximately 20 s and then nystagmus to the opposite side (direction of the fast phase).
   (b) Abnormal test: Eyes do not deviate; this implies brainstem problems.
   (c) The quick phase may return in patients in a persistent vegetative state. Search for the slow, full eye deviation in response to caloric stimulation and not nystagmus when assessing a comatose patient.
6. Pause at least 5 min so the auditory canal can warm up.
   (a) The same EAC can be irrigated with warm water if the contralateral side cannot be tested for any reason.
7. Repeat steps 3–5 in the opposite ear.
   (a) Warm water testing can be used if there is no response to cold water testing. Use warm water (44 °C) and repeat testing, starting with one ear and then, after 5 min, with the other ear.
   (b) Warm water induces nystagmus to the same side (direction of the fast phase).
   (c) COWS: Cold opposite, warm same (referring to the compensatory fast phase of eye movement, the nystagmus).
8. Dry the patient off and reexamine the tympanic membranes to assess for testing-related injury.

42.2.5 Complications: Caloric Reflex Testing

- If contraindications are excluded, no significant complications are expected.
- Tympanic membrane injury or EAC injury can occur from the angiocatheter, forceful irrigation, or injury during cerumen removal.
- Potential complications from caloric testing include meningitis, otitis media, and vomiting.

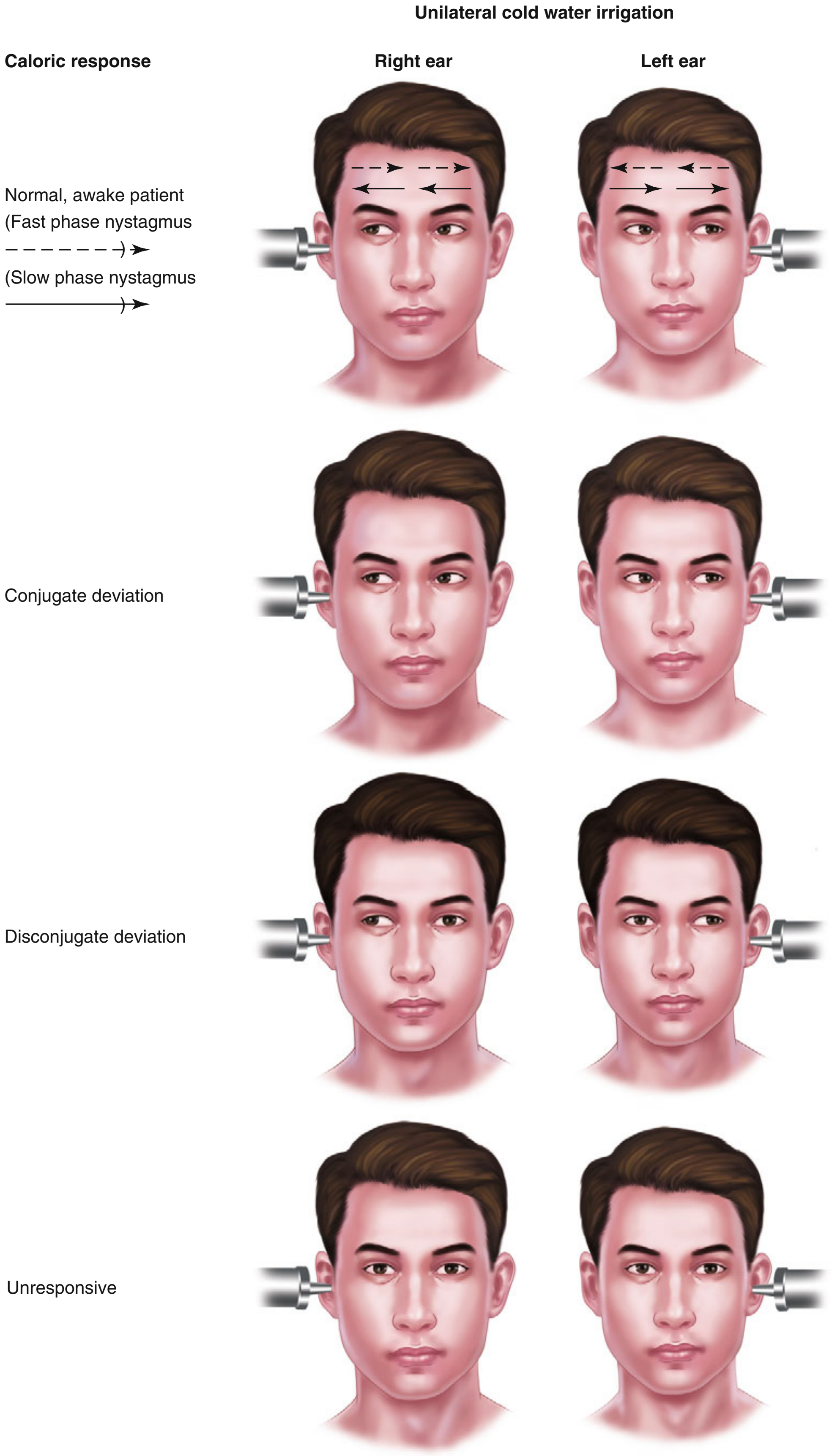
Unilateral cold water irrigation
Caloric response
Right ear
Left ear
Normal, awake patient
(Fast phase nystagmus
(Slow phase nystagmus
Conjugate deviation
Disconjugate deviation
Unresponsive

Fig. 42.2 Depiction of the vestibuloocular reflexes with unilateral cold-water irrigation

42.2.6 Pearls and Pitfalls: Caloric Reflex Testing

- A positive response to caloric testing indicates intact brainstem function.
- In an awake patient, excessive reflex vagal activity may occur (i.e., nausea, vomiting, dizziness).
- Absence of horizontal eye movement means comatose brainstem injury.
- Disconjugate or impaired reflex horizontal eye movements indicate impaired brainstem function at or below the level of the oculomotor nucleus.
- A comatose patient with intact full-reflex horizontal eye movements indicates that the lesion causing coma is in the cerebral hemisphere.

Selected Reading

Gomella LG. Central nervous system. In: Clinician's pocket reference. 11th ed. New York: McGraw-Hill; 2007.

McCann J, et al. Rapid assessment: a flowchart guide to signs and symptoms. Philadelphia: Lippincott William & Wilkins; 2004.

Reeve A, Swensen R. Auditory and vestibular function. In: Disorders of the nervous system: a primer. Hanover, NH: Dartmouth Medical School; On Line; 2009.

Reichman EF, Simon RR. Chapter 103. Reflex eye movements (caloric testing and doll's eyes). In: Emergency medicine procedures. New York: McGraw Hill Education; 2004.

Ropper AH, Samuels MA. Chapter 15. Deafness, dizziness, and disorders of equilibrium. Chapter 17. Coma and related disorders of consciousness. In: Adams and Victor's principles of neurology. 9th ed. New York: McGraw Hill; 2009.

Simon RP, Greenberg DA, Aminoff MJ. Chapter 3. Disorders of equilibrium. In: Clinical neurology. 7th ed. New York: McGraw Hill; 2009. http://www.accessmedicine.com.elibrary.einstein.yu.edu/content.aspx?aID=5146162.

Dix-Hallpike Maneuver

43

Rui Domingues and Muhammad Waseem

The Dix-Hallpike maneuver, also termed the "head-hanging positioning maneuver," is helpful in confirming the clinical suspicion of benign paroxysmal positional vertigo (BPPV). This maneuver provokes abnormal nystagmus, which is a characteristic feature of BPPV.

43.1 Indications

- BPPV is one of the most common types of vertigo.
  - The pathophysiology of BPPV, in brief, is believed to be due to free-moving densities (canaliths/otoliths) in the posterior semicircular canal; with head movement, the particles would alter the flow of the endolymph and cause the stimulation of the ampulla. The particles in the canal cause slow or even reversal of the movement of the cupula and create signals that do not correlate with the actual head movements, therefore causing the sensation of nystagmus.
  - This maneuver locates the cause of vertigo as either the inner ear or the brain; if the problem is in the ear, this maneuver helps localize which ear is affected.
- This maneuver is indicated for patients presenting with vertigo, which is evoked by a change in position and has no symptoms at rest.
- This maneuver is inexpensive, easily done, and part of the physical examination when a patient presents with the complaint of dizziness or vertigo.

43.2 Contraindications

- Severe cervical spine disease.
- Unstable spinal injury.
- High-grade carotid stenosis.
- Unstable heart disease.
- Elderly patients may not tolerate this maneuver.
- There is no need to perform this test in the presence of nystagmus at rest.

43.3 Materials

- Examination table
- Flat cushion
- Frenzel goggles: These high-powered (+20 diopters) magnifying glasses can be placed on a patient during the performance of the maneuver and have shown to increase the sensitivity of the Dix-Hallpike maneuver by preventing the patient from visually fixating on an object, thereby preventing suppression of nystagmus. They are not required to perform the maneuver; they usually are used by specialists.

43.4 Procedure (Fig. 43.1)

1. Have the patient sit at the edge of a bed. The patient is instructed to maintain eye contact with the physician throughout the maneuver.
2. With the patient seated, the examiner will extend the neck, approximately 20°, and turn the head to one side, approximately 30–45°.
3. The examiner then assists the patient by lowering the patient quickly into a supine position, so that the head hangs over the edge of the bed or table, with the neck in a hyperextended position. A flat cushion can be placed

---

R. Domingues, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: rbdomingues@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_43

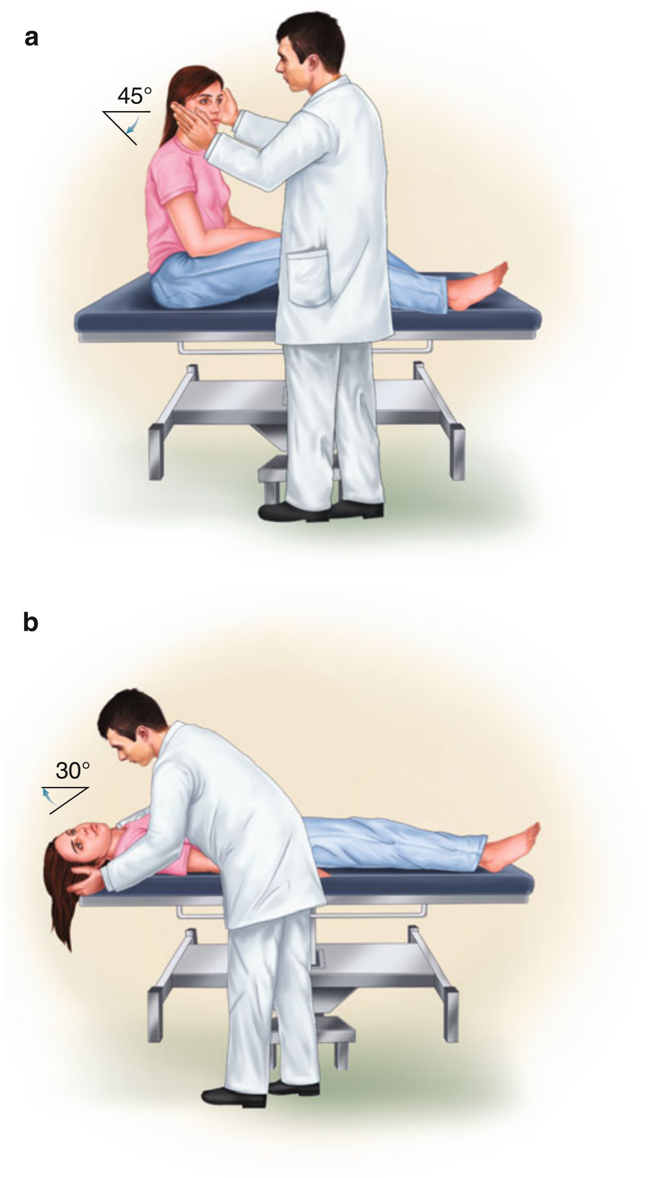

Fig. 43.1 (**a**) With the patient seated, the doctor will extend the neck, approximately 20°, and turn the head to one side, approximately 30–45°. (**b**) The physician then assists the patient by lowering the patient quickly into a supine position, so that the head hangs over the edge of the bed or table, with the neck in a hyperextended position

beneath the person's back in the shoulder blade area to assist with obtaining head extension.

4. This position is held, and the examiner observes for nystagmus for up to 60 s.
    (a) Nystagmus is a rapid, rhythmic movement of the eyes and usually appears after a brief period and lasts less than 30 s.
    (b) The direction of the nystagmus is usually up and twisted; therefore, the eyes will beat toward the ground.
    (c) If the patient becomes dizzy or the doctor sees nystagmus, the test is positive for the ear that is pointed to the floor.
5. The patient is then returned to the upright position and is again observed for nystagmus for 30 s.
6. If nystagmus is not provoked, the maneuver is repeated with the head turned to the opposite direction.
7. If nystagmus is provoked, the patient should have the maneuver repeated to the same side.
    (a) With each repetition, the intensity and duration of nystagmus should decrease.

43.5 Complications

- Vertigo
- Nausea

43.6 Pearls and Pitfalls

- The maneuver can be uncomfortable to the patient because it can cause vertigo and nausea.

Selected Reading

Brandt T, Steddin S, Daroff RB. Therapy for benign paroxysmal positioning vertigo, revisited. Neurology. 1994;44:796–800.

Buckingham RA. Anatomical and theoretical observations on otolith repositioning for benign paroxysmal positional vertigo. Laryngoscope. 1999;109:717–22.

Fife TD, Iverson DJ, Lempert T, et al. Practice parameter: therapies for benign paroxysmal positional vertigo (evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2008;70:69–74.

Halker RB, Barrs DM, Wellik KE, Wingerchuck DM, Demaerschalk BM. Establishing a diagnosis of benign paroxysmal positional vertigo through the Dix-Hallpike and side-lying maneuvers: a critically appraised topic. Neurologist. 2008;14:201–4.

Herdman SJ. Treatment of benign paroxysmal vertigo. Phys Ther. 1990;70:381–8.

Lanska DJ, Remler B. Benign paroxysmal positioning vertigo: classic descriptions, origins of the provocative positioning technique and conceptual developments. Neurology. 1997;48:1167–77.

Li JC, Meyers AD. Benign paroxysmal positional vertigo. http://emedicine.medscape.com/article/884261-overview. Accessed December 20, 2015.

Von Brevern M, Radtke A, Lezius F, et al. Epidemiology of benign paroxysmal positional vertigo: a population based study. J Neurol Neurosurg Psychiatry. 2007;78:710–5.

Epley Maneuver for Vertigo (Particle Repositioning or Canalith Repositioning Procedure)

44

Rui Domingues and Muhammad Waseem

During the Dix-Hallpike test (see Chap. 43), the particles move in the canal and trigger a burst of upbeat-torsional nystagmus. The Epley maneuver causes resolution of positional nystagmus. This maneuver is effective in about 80 % of patients with benign paroxysmal positional vertigo (BPPV).

44.1 Indications

- Performed to alleviate the symptoms of posterior canal BPPV

44.2 Contraindications

- Back or spine injuries or other problems
- Presence of detached retina

44.3 Materials

- Pillow or pad
- Padded table or bed

44.4 **Procedure** (Fig. 44.1)

1. The patient is positioned on the bed with the head slightly extended and turned 45° to the affected side; hold this position for 60 s.
    - The patient may hold the physician's arm for support.
2. The patient is then assisted by the physician to a supine position, with the head in extension, with a pillow or pad placed at the shoulder level allowing for extension; this position is held for a 60 s. This position will likely provoke transient dizziness and vertigo.
3. The head is then turned 90° to the opposite side; this position is held for 60 s. This position will likely provoke transient dizziness and vertigo.
4. The head is then turned through a further 90° while the patient rolls onto the unaffected side; this position is held for 60 s.
5. The patient then sits up slowly, by rolling up from lying on the unaffected side; this position is held for 60 s.
6. The patient should wait for 10–15 min before discharge.
7. Discharge instructions are then provided. Home instructions:
    - Instruct the patient to sleep in a semi-recumbent position for the next two nights; the patient should sleep at a 45° angle, which is most easily done by sleeping in a recliner or with several pillows arranged on a couch.
    - Sleep on the nonaffected side.
    - During the day, try to keep the head in a vertical position; no sudden head movements to the right, left, up, or down.
    - When men shave under their chins, they should bend their bodies forward in order to keep their heads vertical.
    - Do not go to the hairdresser or dentist because these require head movements.
    - Care should be taken when putting in eye drops because it requires head extension.
    - No vacuuming or mopping the floor.
    - Try not to wear clothing that needs to be pulled over the head. Try to avoid bending down to tie shoes.

R. Domingues, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: rbdomingues@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_44

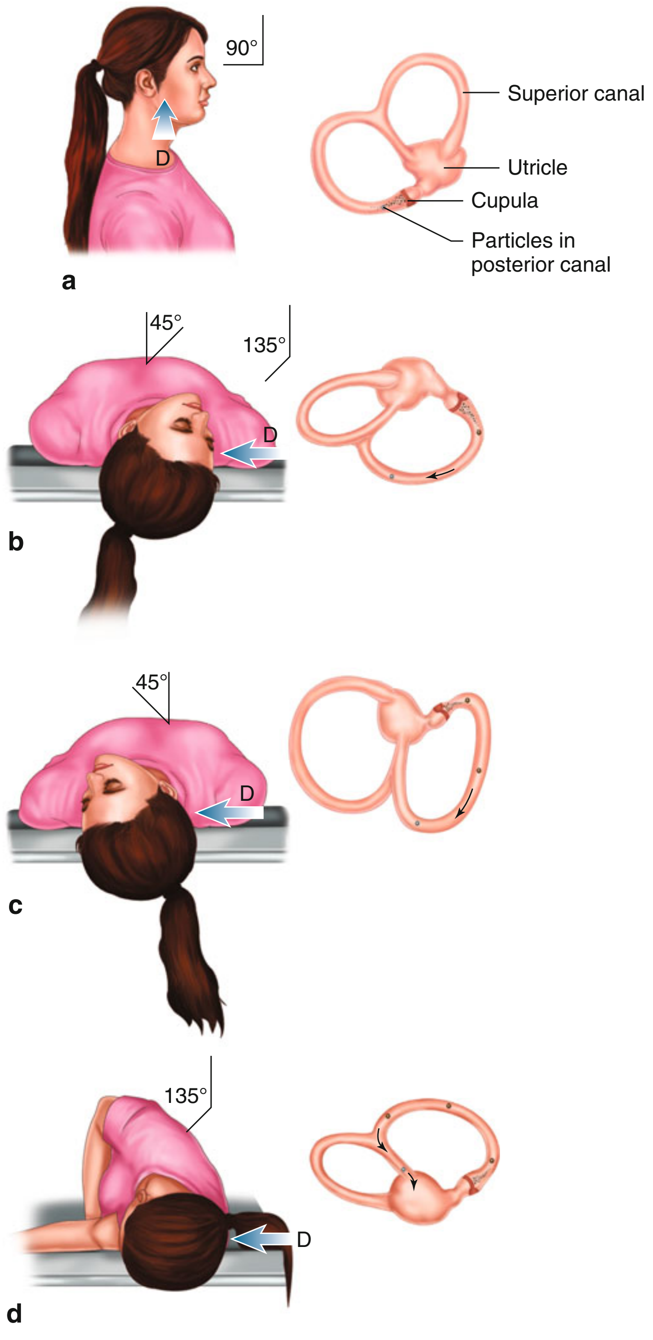

Fig. 44.1 (**a**) Patient seated on the table. (**b**) Patient in supine position, with the head turned 45° to the affected side. (**c**) Head then turned to the unaffected side (45°). (**d**) Head turned additional 90° on the unaffected side. *D* direction of view of labyrinth

44.5 Complications

When performing the previous maneuver, caution is advised should neurological symptoms occur. Occasionally such symptoms are caused by compression of the vertebral arteries; if it persists for a long period of time, a stroke can occur.

44.6 Pearls and Pitfalls

- Pearls
  - This maneuver should initially be performed by a trained therapist or medical physician.
  - It is best to perform the Epley maneuver before going to bed at night.
- Pitfalls
  - The patient should not drive herself or himself home after the procedure.
  - Avoid rapid changes in head position that might provoke BPPV.

Selected Reading

Harvey SA, Hain TC, Adamiec LC. Modified liberator maneuver: effective treatment for benign paroxysmal positional vertigo. Laryngoscope. 1994;104:1206–12.

Herdman SJ. Treatment of benign paroxysmal vertigo. Phys Ther. 1990;70:381–8.

Massoud EA, Ireland DJ. Post-treatment instructions in the nonsurgical management of benign paroxysmal positional vertigo. J Otolaryngol. 1996;25:121–5.

Parnes LS, Price-Jones RG. Particle repositioning maneuver for benign paroxysmal positional vertigo. Ann Otol Rhinol Laryngol. 1993;102:325–31.

Radtke A, Neuhausen H, von Brevern M, Lempert T. A modified Epley's procedure for self-treatment of benign paroxysmal positional vertigo. Neurology. 1999;53:1358–60.

Smouha EE. Time course of recovery after Epley maneuvers for benign paroxysmal positional vertigo. Laryngoscope. 1997;107:187–91.

Welling DB, Barnes DE. Particle repositioning maneuver for benign paroxysmal positional vertigo. Laryngoscope. 1994;104:946–9.

Clinical Brain Death Examination in Adults

45

Eric S. Papierniak, Hassan M. Alnuaimat, Tracy Timmons,
and Deborah M. Stein

45.1 Indications

- Deep coma without any identifiable reversible causes
- Suspected brain death in a mechanically ventilated patient

45.2 Contraindications

- Metabolic derangements
- Acid-base disorders
- Electrolyte abnormalities sufficient to cause the coma
- Profound hypotension
- Hypothermia (core body temperature <36°C)
- Drug overdose or toxic exposure
- Locked-in syndrome

45.3 Materials and Medications

- A bright light
- Intravenous (IV) catheter
- 30-mL syringe
- 50-mL of ice water
- Endotracheal suction catheter
- Long swab or tongue depressor
- Gauze 4 × 4 or saline eye drops

45.4 Procedure

1. Evaluate the motor response to pain.
    - Apply pressure to the supraorbital nerve and nail beds. No motor response is consistent with brain death (Figs. 45.1 and 45.2).
2. Test for the absence of brainstem function.
    - Check pupillary response to bright light.
    - Check the oculocephalic reflex. With the eyelids held open, quickly turn the head to the side. The normal response is for the eyes to move in the opposite direction, maintaining the focus forward. Failure of the eyes to deviate during head rotation is consistent with brain death (Fig. 45.3).
    - Check the oculovestibular reflex with cold caloric testing. Elevate the head of the bed to 30°. Attach an IV catheter to the tip of a 30-mL syringe. Inject 50 mL of cold water into the external ear canal. Observe the pupils for 1 min for deviation toward the ear with the cold stimulus. Wait 5 min before testing the other side. No deviation of the pupils is consistent with brain death (Figs. 45.4 and 45.5).
    - Check for bulbar paralysis.
        - Check for a cough response to bronchial suctioning.

E.S. Papierniak (✉)
Division of Pulmonary, Critical Care, and Sleep Medicine,
Department of Medicine, University of Florida Health,
Gainesville, FL, USA

Malcom Randall VA Medical Center, Gainesville, FL, USA
e-mail: eric.papierniak@medicine.ufl.edu

H.M. Alnuaimat, MD
Department of Medicine, University of Florida, University of
Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: Hassan.Alnuaimat@medicine.ufl.edu

T. Timmons, MD
Department of Trauma/Surgical Critical Care, R Adams Cowley
Shock Trauma Center, University of Maryland Medical Center,
Baltimore, MD, USA
e-mail: ttimmons1@umm.edu

D.M. Stein, MD, MPH
Department of Surgery, University of Maryland School of
Medicine, R Adams Cowley Shock Trauma Center, University of
Maryland Medical Center, Baltimore, MD, USA
e-mail: dstein@umm.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_45

  - Stimulate the posterior pharynx with a long swab or tongue depressor and observe for a gag.
  - Check the corneal reflex by lightly touching the cornea with a cotton swab, corner of gauze, or drops of sterile saline and observe for blinking of the eyelids (Fig. 45.6).
3. Check for respiratory effort by performing an apnea test. Before starting the apnea test, the following conditions must be met:
   (a) The patient must not be hypothermic. The core temperature must be greater than 36°C.
   (b) The patient must be hemodynamically stable with a systolic blood pressure greater than 90 mmHg.
   (c) The arterial blood gas (ABG) must demonstrate a normal arterial partial pressure of carbon dioxide ($PaCO_2$). The arterial partial pressure of oxygen (PaO_2) may be elevated with preoxygenation to minimize the risk of hypoxemia prematurely ending the examination.
      - Monitor the patient with a pulse oximeter.
      - Disconnect the ventilator and deliver oxygen at 6 L/min by cannula into the endotracheal tube.
      - Alternately the patient can remain connected to the ventilator with no applied support, which allows for an in-line negative pressure monitor to be attached to the circuit. Note: Most ventilators are too sensitive to be used as the sole method of sensing respiratory effort as they can produce false-negative results.
      - Watch the chest and abdomen for respiratory motion, or monitor for negative inspiratory force on the pressure gauge. Check an ABG every 10 min and when the test ends.
      - If the $PaCO_2$ increases by 20 mmHg or the $PaCO_2$ is greater than 60 mmHg and there was no respiratory effort, the test is consistent with brain death.
      - The test must be aborted and the patient reconnected to mechanical ventilation if the patient becomes hypotensive, desaturates, or develops cardiac arrhythmias.
      - If there is respiratory effort, the patient must be reconnected to mechanical ventilation (Figs. 45.7 and 45.8).

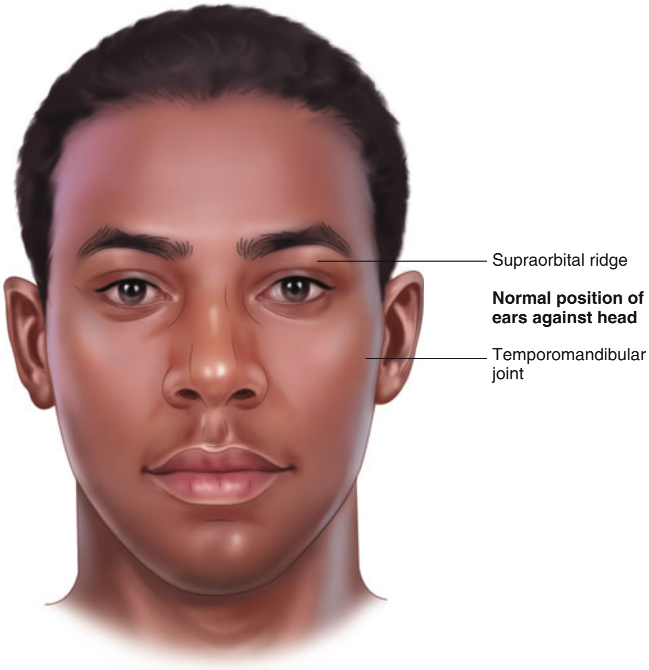

Fig. 45.1 Motor response to pain 1. Temporomandibular joint and supraorbital ridge test supraorbital nerve function

Fig. 45.2 Motor response to pain 2: Firm pressure can be applied to the nail beds with a penlight or similar instrument

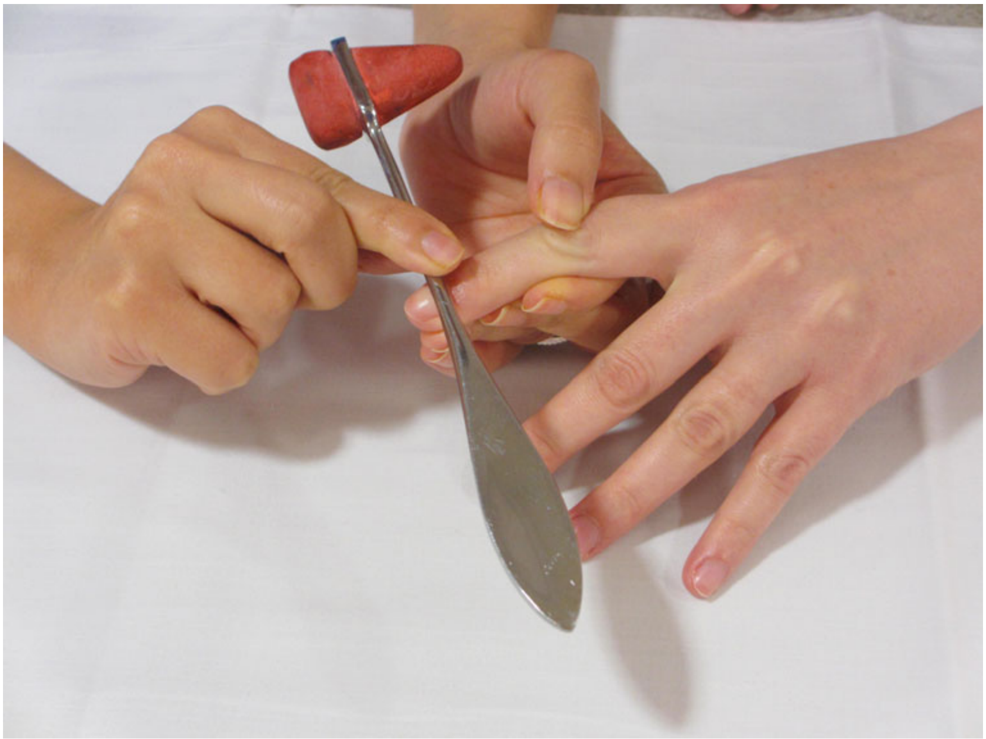

a Normal reaction: Eyes move side to side when head is turned

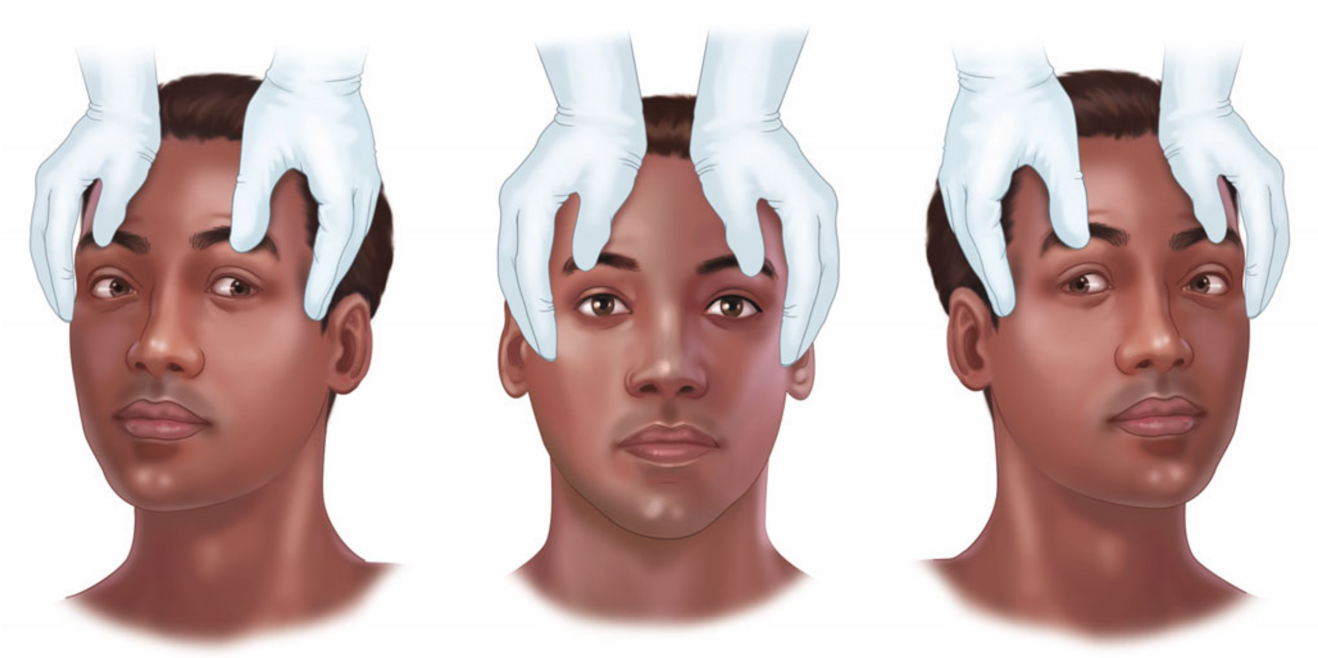

b Abnormal reaction: Eyes remain in fixed position in skull when head is turned

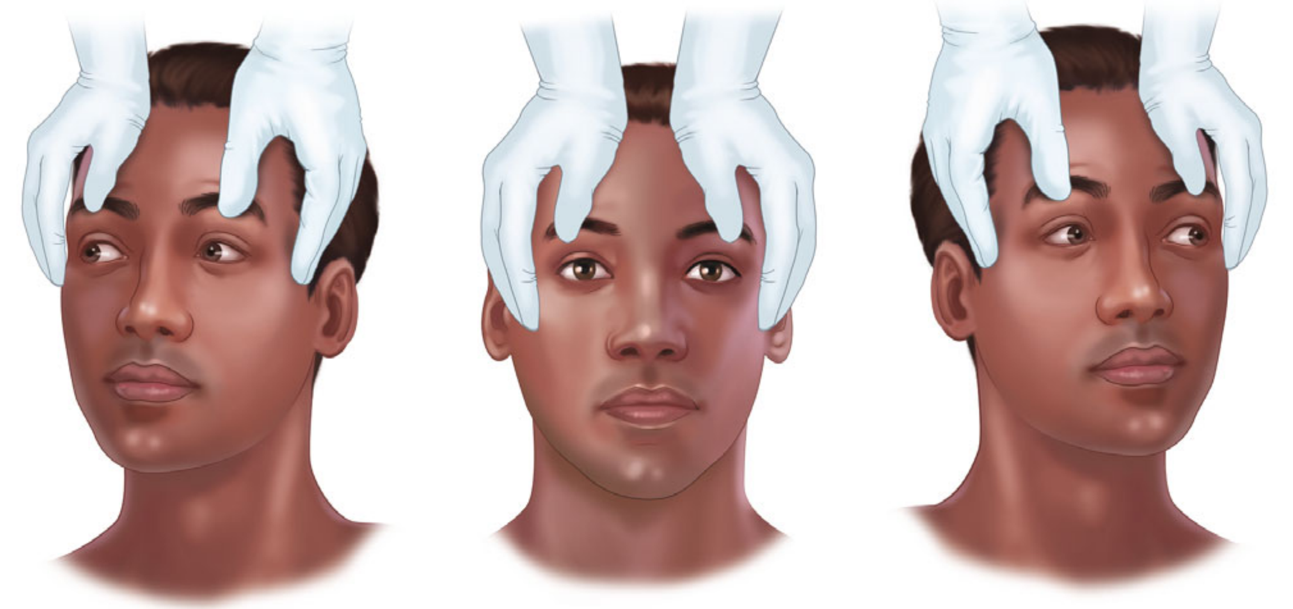

Fig. 45.3 (**a**, **b**) Oculocephalic reflex or "Doll's eyes." Movement of the eyes away from the direction of head turning (i.e., to keep the gaze forward) indicates intact functioning

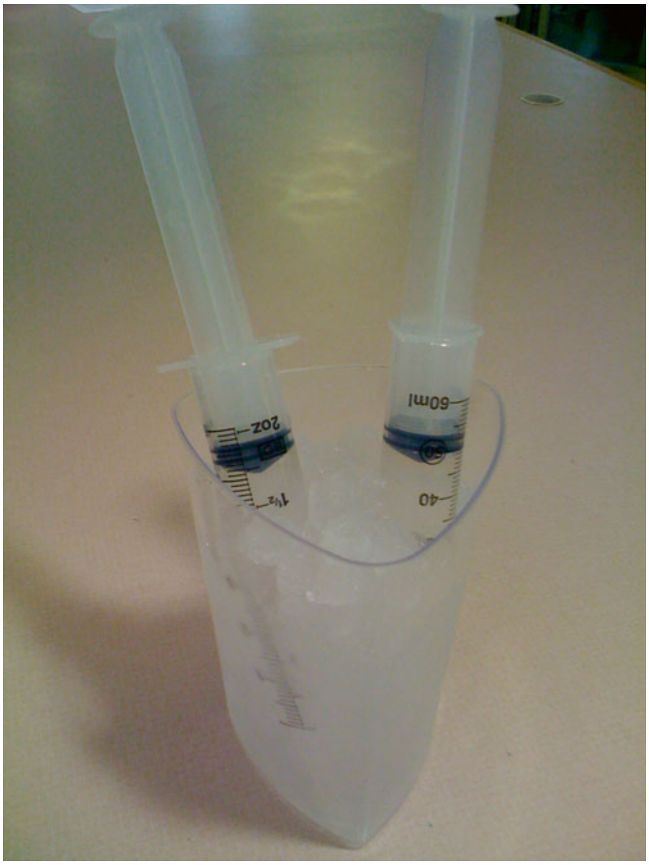

2oz
1½
60ml
40

Fig. 45.4 Preparation of cold saline for testing of the oculovestibular reflex or "cold calorics"

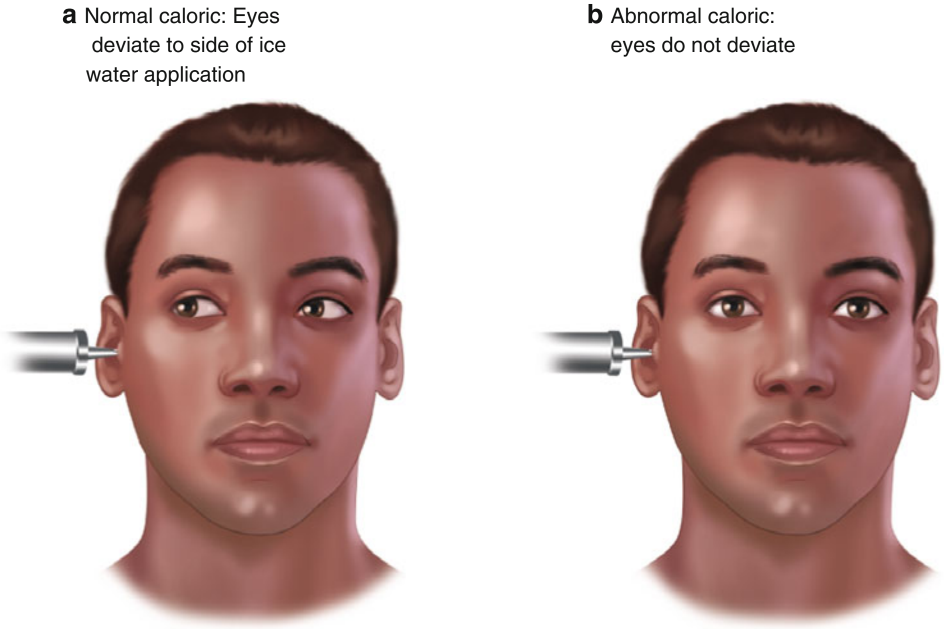

a Normal caloric: Eyes deviate to side of ice water application
b Abnormal caloric: eyes do not deviate

Fig. 45.5 (**a**, **b**) Oculovestibular reflex. Intact functioning is demonstrated by deviation of the gaze toward the side being tested

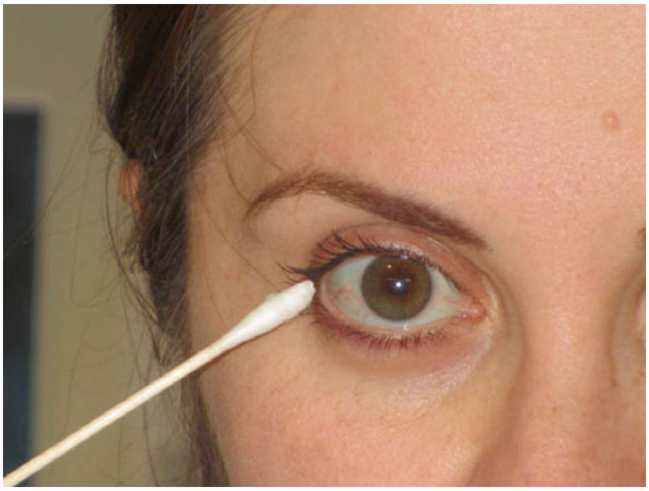

Fig. 45.6 Use of a cotton ball or unwound tip of a cotton swab for testing of the corneal reflex. A piece of sterile gauze (“4×4”) is also commonly used

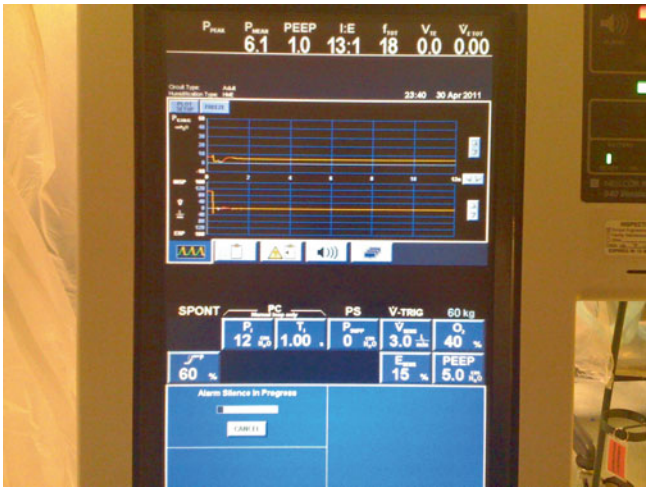
PPEAK
PMEAN 6.1
PEEP 1.0
I:E 13:1
fTOT 18
VTE 0.0
VE TOT 0.00
23:40 30 Apr 2011
SPONT
PC
PS
V-TRIG
60 kg
PI 12
TI 1.00
PSUPP 0
VSENS 3.0
O2 40 %
60 %
ESENS 15 %
PEEP 5.0
Alarm Silence in Progress
CANCEL

Fig. 45.8 Ventilator waveform in a patient without spontaneous respirations during the apnea test

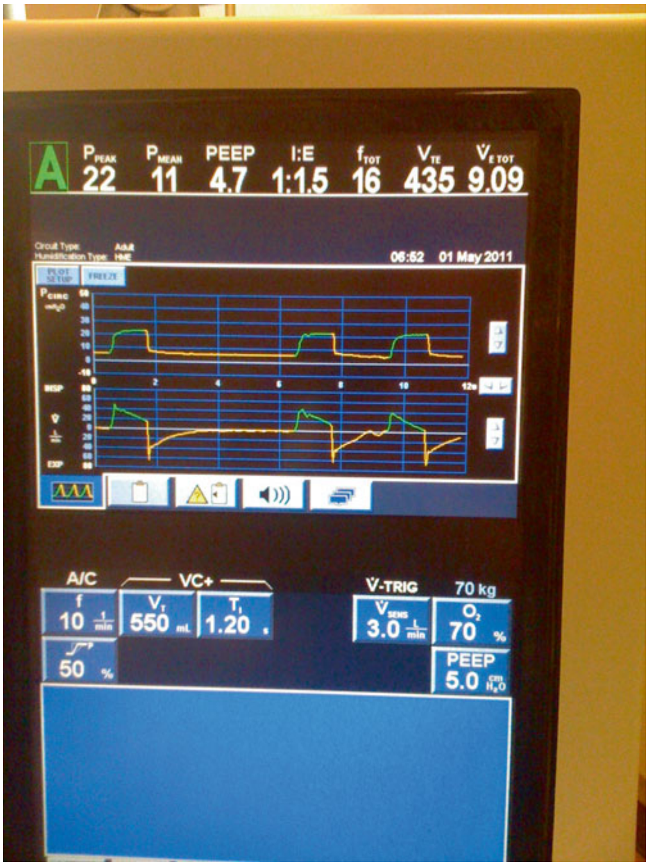
A
PPEAK 22
PMEAN 11
PEEP 4.7
I:E 1:1.5
fTOT 16
VTE 435
VE TOT 9.09
Circuit Type Adult
Humidification Type HME
06:52 01 May 2011
PLOT SETUP
FREEZE
PCIRC
A/C
VC+
V-TRIG
70 kg
f 10 1/min
VT 550 mL
TI 1.20 s
VSENS 3.0 L/min
O2 70 %
50 %
PEEP 5.0 cmH2O

Fig. 45.7 Ventilator waveform before apnea test

45.5 Pearls and Pitfalls

- Pearls
  - The clinical examination should be performed by someone with experience or training in brain death examinations. Local regulations as well as hospital policies should be reviewed in order to determine the requirements for brain death. Some locales require a repeat examination by a different practitioner.
  - A patient with a cervical spine injury causing quadriplegia may not feel nail bed pressure and may be unable to respond to pain with more than facial movements.
- Pitfalls
  - Pupillary constriction less than 3 mm is not consistent with brain death.
  - Do not test oculocephalic reflex (doll's eyes) on patients who may have a cervical spinal cord injury from trauma.
  - Do not test the oculovestibular reflex with cold caloric testing in patients with a ruptured tympanic membrane.
  - A large number of patients may demonstrate spinal reflex movements during brain death. These reflexes may be triggered by touch, noxious stimuli, or removal of the ventilator. These movements can include plantar flexion, upper extremity posturing, eye opening, and the "Lazarus sign" (the arms raise off the bed and cross). Although these movements can be disconcerting to the health care team and family, they do not preclude a diagnosis of brain death.

45.6 Controversies

- Repeat examinations
  - The American Academy of Neurology guidelines recommend repeating the brain death examination at 6 h. The guidelines acknowledge that this is an arbitrary interval. Recent evidence suggests that the second examination may be unnecessary.
- Confirmatory Testing
  - Confirmatory testing is an option if specific elements of the clinical examination cannot be performed. For example, a trauma patient with suspected cervical spinal cord injuries precluding the oculocephalic reflex test would be a candidate for confirmatory testing.
  - Conventional cerebral angiography will demonstrate no filling beyond the carotid bifurcation or within the circle of Willis. Electroencephalography demonstrates no electrical activity. Transcranial Doppler ultrasonography will demonstrate vascular resistance associated with elevated intracranial pressure. Technetium-99m hexamethylpropylene-amine oxime brain scans demonstrate the "hollow skull" sign or no uptake of isotope in the brain (Fig. 45.9).
  - Newer modalities of confirmatory testing include computed tomography (CT) angiography and magnetic resonance imaging/magnetic resonance angiography (MRI/MRA). These tests are being used in some hospitals; however, the recent review by the American Academy of Neurology finds evidence insufficient to recommend using these newer modalities to confirm brain death.

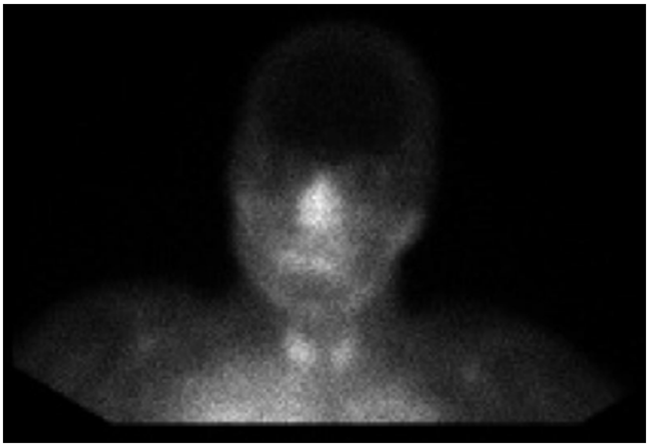

Fig. 45.9 "Empty skull sign" on nuclear medicine flow study confirms brain death

Selected Reading

A definition of irreversible coma: report of the Ad Hoc Committee of the Harvard Medical School to Examine the Definition of Brain Death. JAMA. 1968;205:337–40.

Lustbader D, O'Hara D, Wijdicks EF, et al. Second brain death examination may negatively affect organ donation. Neurology. 2011;76:119–24.

Practice parameters for determining brain death in adults (summary statement). The quality standards subcommittee of the American Academy of Neurology. Neurology. 1995;45:1012–4.

Saposnik G, Maurino J, Saizar R, Bueri JA. Spontaneous and reflex movements in 107 patients with brain death. Am J Med. 2005;118:311–4.

Wijdicks EF. Current concepts: the diagnosis of brain death. N Engl J Med. 2001;344:1215–21.

Wijdicks EF, Varelas PN, Gronseth GS, Greer DM, American Academy of Neurology. Evidence-based guideline update: determining brain death in adults: report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2010;74:1911–8.

Part VII

Ophthalmic Procedures

Slit Lamp Examination

46

Bobby K. Desai

46.1 Indications

- A slit lamp magnifies structures of the eye (Fig. 46.1).
- Gives the operator a three-dimensional view of the area visualized.
- Used to delineate abnormalities that cannot be visualized by other means.
- Helpful in foreign body removal.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_46

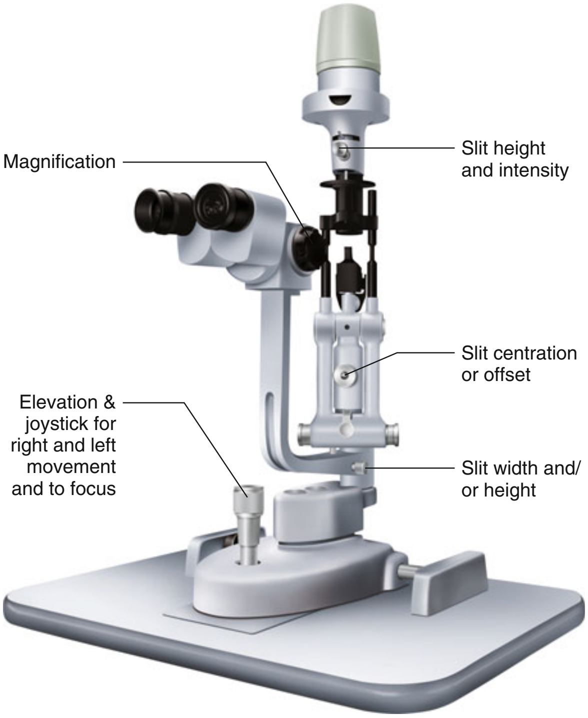
Magnification
Slit height and intensity
Slit centration or offset
Elevation & joystick for right and left movement and to focus
Slit width and/ or height

Fig. 46.1 A basic slit lamp

46.2 Contraindications

- Not to be used in patients who cannot tolerate an upright posture (e.g., trauma patients).
  - In these cases if a slit lamp examination is required, a portable slit lamp may be beneficial (Fig. 46.2).

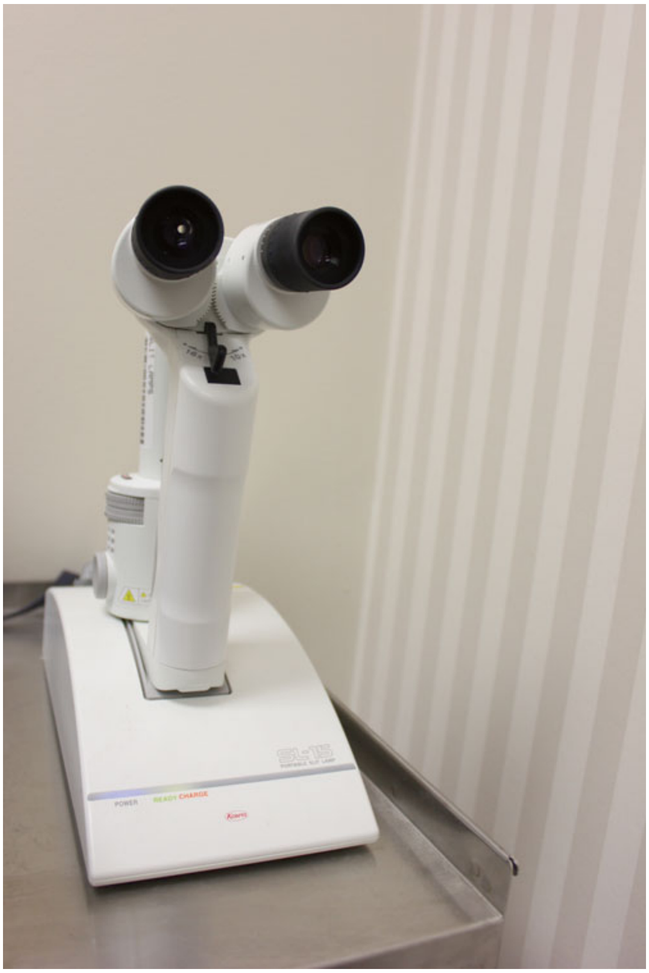

Fig. 46.2 A portable slit lamp

46.3 Materials

- The slit lamp is an eye-specific binocular stereoscope specifically designed to examine the eye and its structures and gives the practitioner a three-dimensional assessment of the eye.
- The operator has full movement of the microscope in all directions and the apparatus can be locked in place if required.
- Its light source can be manipulated to change the characteristics of the beam from its intensity to the angle at which it projects (Fig. 46.3).
  - A low-power setting is typically used for normal examination.
  - A higher-intensity beam is used when evaluating the anterior chamber with a narrow slit beam.
- It has colored filters typically built within the slit lamp.
  - Cobalt blue: Used with fluorescein dye to evaluate for corneal abrasions (Fig. 46.4) and avulsions; the dye will collect where the corneal epithelium is absent. It results in a yellow glow or hue visible through the microscope.
  - Green filter: Used to increase contrast of blood vessels. They appear black and the filter is useful for the assessment of hemorrhage.
- The operator is able to adjust the magnification of the microscopic typically through dial controls.
  - Low magnifications are most helpful for general examination.
  - Higher magnifications are used for examination of a particular area in fine detail.

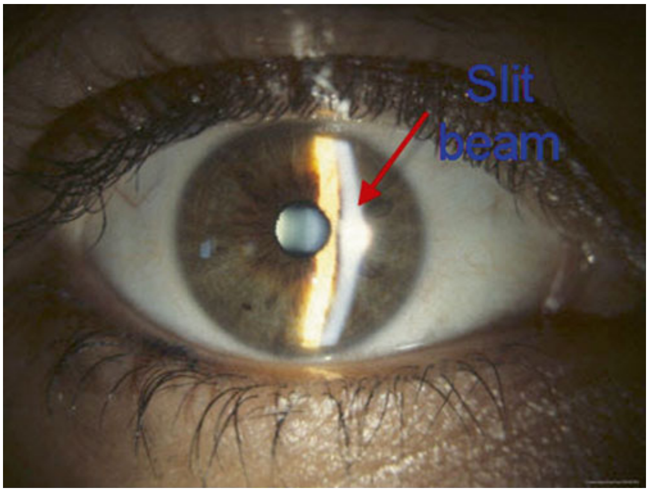

Fig. 46.3 The slit lamp beam and the reflection of the beam

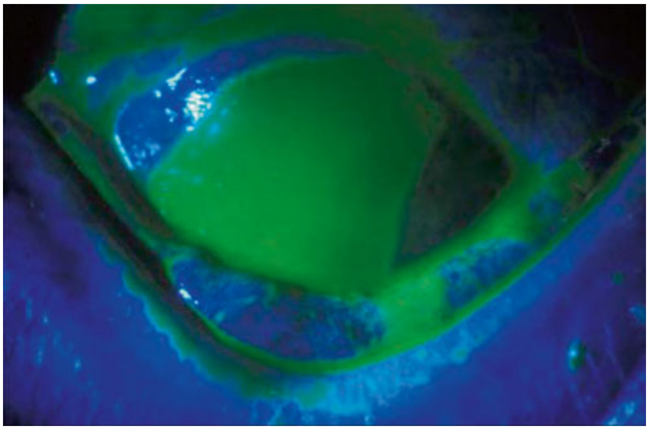

Fig. 46.4 Example of a corneal abrasion seen with a cobalt blue filter (With kind permission from Springer Science + Business Media: Das S, Chohan A, Snibson GR, Taylor HR. Capsicum spray injury of the eye. *International Ophthalmology*. 2005;26(4–5):171–3)

46.4 Procedure

1. As part of a complete eye examination, informed consent is not generally required.
2. An explanation of the procedure and reassurance are helpful.
3. Lock the slit lamp before positioning the patient because unintended movement may inadvertently damage the equipment or cause injury to the operator or patient.
4. The apparatus has a chin and a head rest. Ask the patient to place the chin and forehead in the respective areas (Fig. 46.5).
5. Adjust the height of the apparatus and the patient's chair for optimum comfort (Figs. 46.6 and 46.7). It may be necessary for children to stand during the examination.
6. The patient may place his or her hands on the sides of the table that the lamp is mounted on; this ensures stability of both the lamp and the patient.
7. Adjust the eyepieces of the microscope to fit the operator.
8. Turn on the slit lamp at its lowest setting to avoid a sudden power surge that could potentially damage the bulb.
9. Move the stage forward and narrow the beam and angle it at 45° to the patient. Aim the beam laterally so as not to cause the patient discomfort.
10. Focus the beam by manipulating the joystick to move the apparatus forward and backward so that the beam is clearly visible and its lines are sharp.
11. For each area of the eye to be examined, inspect the area thoroughly using the joystick to slowly manipulate the slit lamp across the eye in all directions, using the height adjustment of the joystick to slightly raise and lower the slit lamp as needed.
    - The operator may find that the slit lamp may move too freely; in that case she or he may find slightly tightening the locking nut of the C-arm may provide better and more precise control.

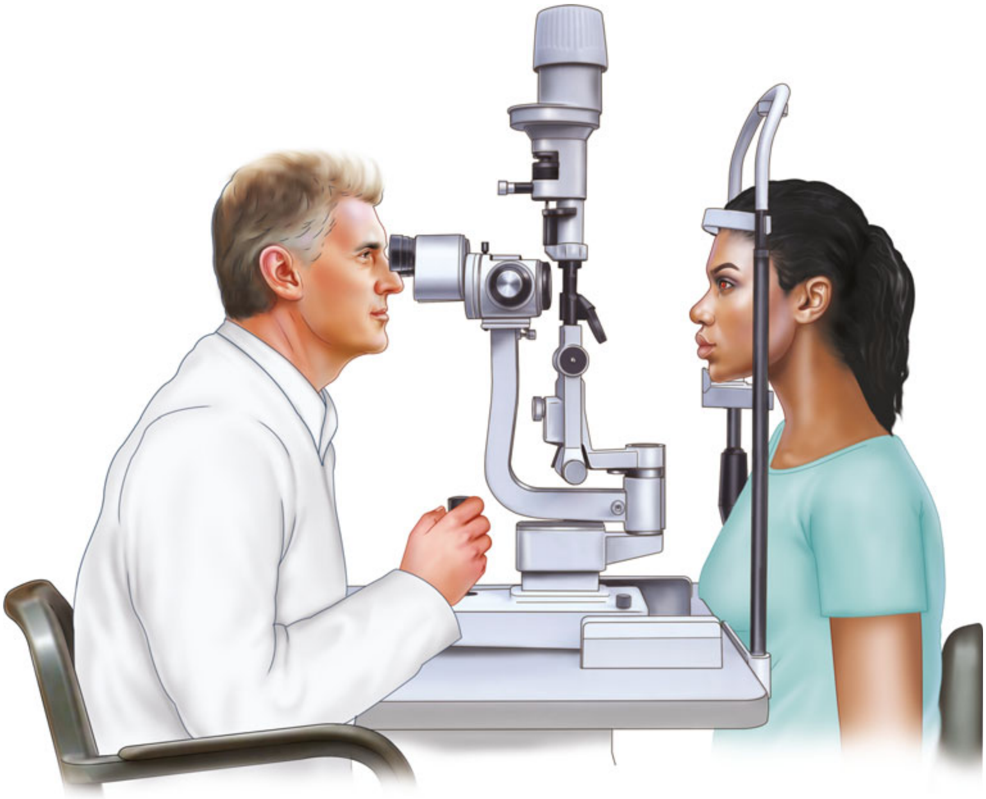

Fig. 46.5 Appropriate positioning for a slit lamp examination

Fig. 46.6 The slit lamp is too high

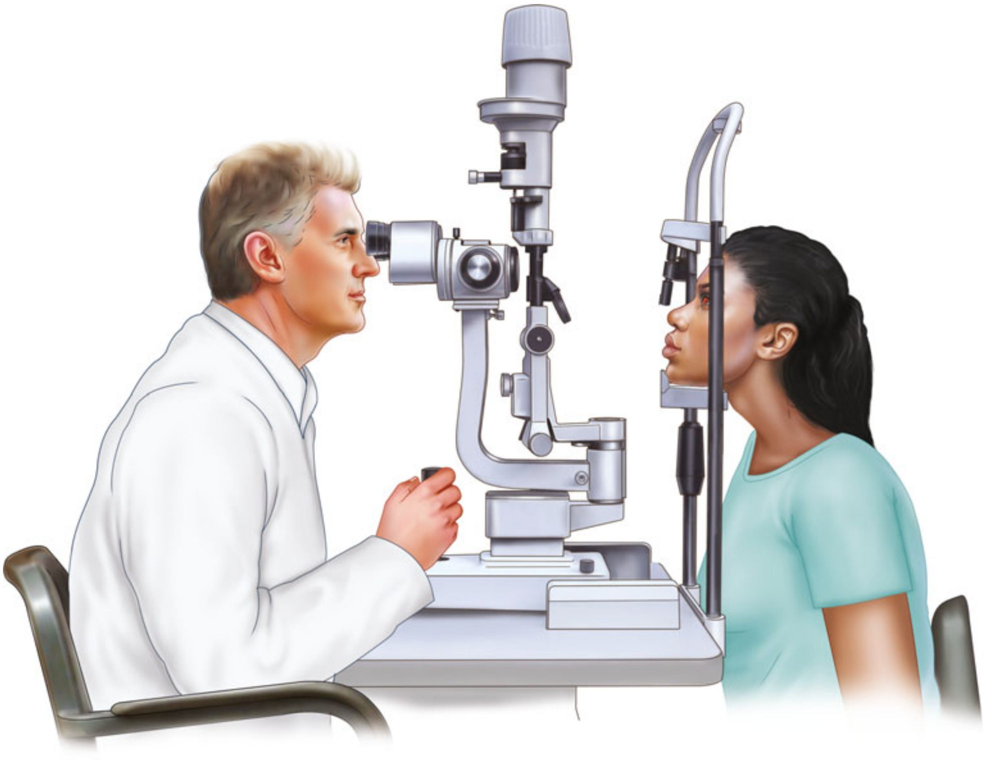

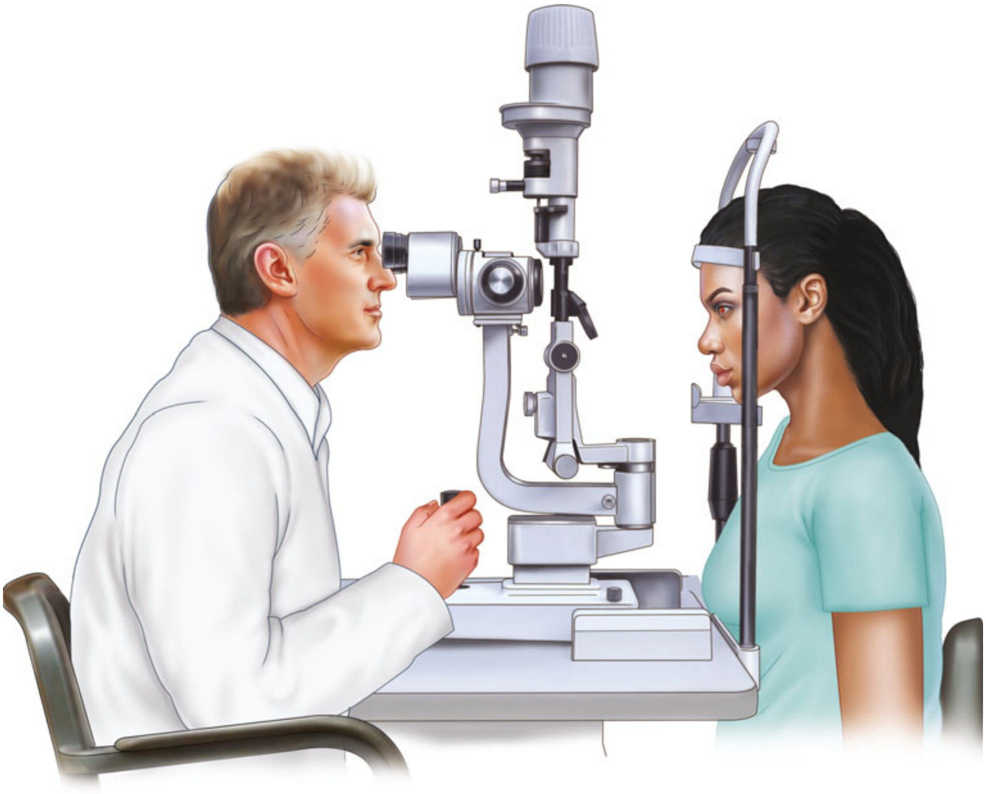

Fig. 46.7 The slit lamp is too low

46.5 Complications

There are no complications following routine slit lamp examination.

46.6 Pitfalls

- Ensure a working bulb.
- Appropriately set the microscope's oculars for the operator's pupillary distance.
- Ensure all dials and knobs are firmly set.
- Inadequate focusing
- Patient noncompliance

46.7 Fluorescein Testing

46.7.1 Indications and Uses

- Examination of the cornea under a cobalt blue filter on a slit lamp or Wood's lamp: The chemical fluoresces under blue light and appears green under these circumstances.
- Useful for detecting corneal abrasions, corneal foreign bodies, and infections of the eye.

46.7.2 Procedure

- Fluorescein is typically used after installation of a topical anesthetic, which provides patient comfort especially for those with significant pathology.
- The fluorescein strip comes packaged in single-use wrappers (Fig. 46.8).
- Hold the strip by the white end and wet the orange end *lightly* because heavily moistening the strip may cause a significant amount of dye to be present, obscuring the examination.
  - A dry strip may be used, but it may irritate the patient's eye especially if already sensitive.
- If this occurs, the clinician may use tissue paper to gently blot the excess solution away.
  - The exception to this is the performance of the *Seidel test*, which is used to assess the eye for potential perforation (Fig. 46.9).
  - The clinician will instill a large amount of dye into the eye by wetting the orange strip copiously.
  - The clinician will next examine the eye for a stream of fluid leaking from the ruptured globe.
  - This stream will fluoresce green or blue in distinction to the rest of the globe, which appears orange.
- The choice of solution to wet fluorescein strip is up to the clinician because saline, tap water, or the recently used anesthetic solution may be used safely.
- Place the now-wet orange end on the lower lid of the patient's eye.
- Ask the patient to blink several times to allow the solution to spread evenly.
- The clinician may use a Wood's lamp, penlight, or the cobalt blue filter on the slit lamp to examine the now-stained eye.
  - The slit lamp is preferable owing to the potential for missing small abrasions.

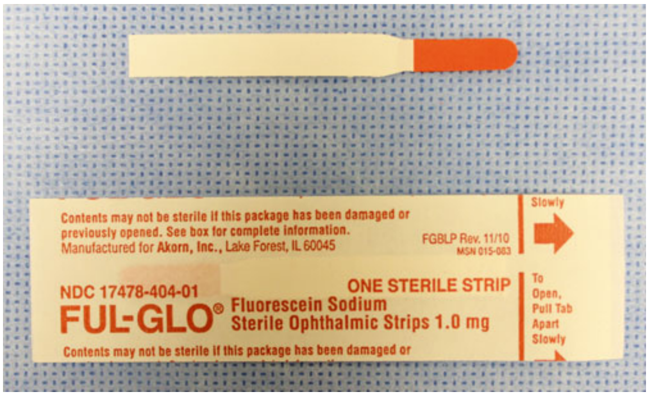
Contents may not be sterile if this package has been damaged or
previously opened. See box for complete information.
Manufactured for Akorn, Inc., Lake Forest, IL 60045
FGBLP Rev. 11/10
MSN 015-083
Slowly
NDC 17478-404-01
ONE STERILE STRIP
FUL-GLO®
Fluorescein Sodium
Sterile Ophthalmic Strips 1.0 mg
Contents may not be sterile if this package has been damaged or
To Open, Pull Tab Apart Slowly

Fig. 46.8 A typical fluorescein strip package

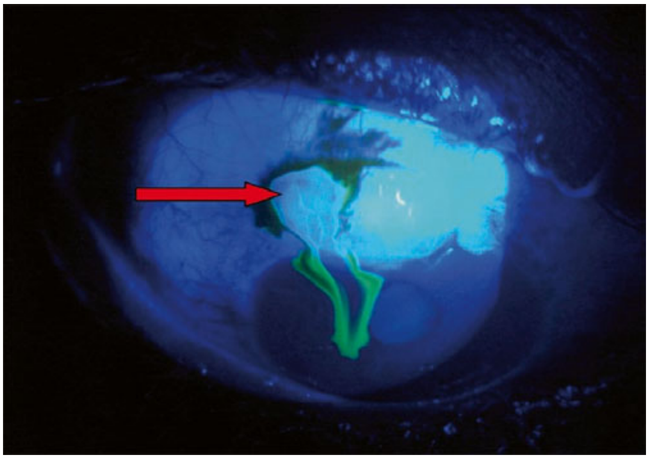

Fig. 46.9 A positive Seidel test (Reproduced with permission from: Lingam V, Panday M, George R, Shantha B. Management of complications in glaucoma surgery. *Indian J Ophthalmol*. 2011;59(Suppl1): S131–40)

46.7.3 Complications

- There is a theoretical risk of the development of superficial punctate keratitis from instillation of topical anesthetic before fluorescein testing.
- Discoloration of soft contact lenses.
- Potential for infection using premixed fluorescein solution.

46.7.4 Pitfalls

- Contact lens wearers should remove their lenses because the fluorescein will permanently stain the lens.
  - The wearer should not put the contacts back in for several hours.

Selected Reading

DuBois L. The slit lamp examination. In: DuBois L, Ledford JK, Daniels K, Campbell R, editors. Clinical skills for the ophthalmic examination: basic procedures. 2nd ed. Thorofare: Slack; 2006. p. 61–70.

Galor A, Jeng BH. Red eye for the internist: when to treat, when to refer. Cleve Clin J Med. 2008;75:137–44.

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Ledford JK, Sanders VN. The slit lamp primer. 2nd ed. Thorofare: Slack; 2006.

47 Eye Irrigation

Bobby K. Desai

47.1 Indications

- Chemical burns to the eye
- Removal of superficial foreign bodies

47.2 Contraindications

- For suspected globe perforation, extreme care must be taken to not exacerbate the injury.

47.3 Materials and Medications

- Irrigating device—the Morgan lens (Fig. 47.1)
- Saline solution or lactated Ringer's (preferably warmed)
- Topical anesthetic drops (Fig. 47.2)
- Basin to secure the extruded solution
- Intravenous (IV) tubing to attach the IV bag to the Morgan lens
- pH paper (Fig. 47.3)

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_47

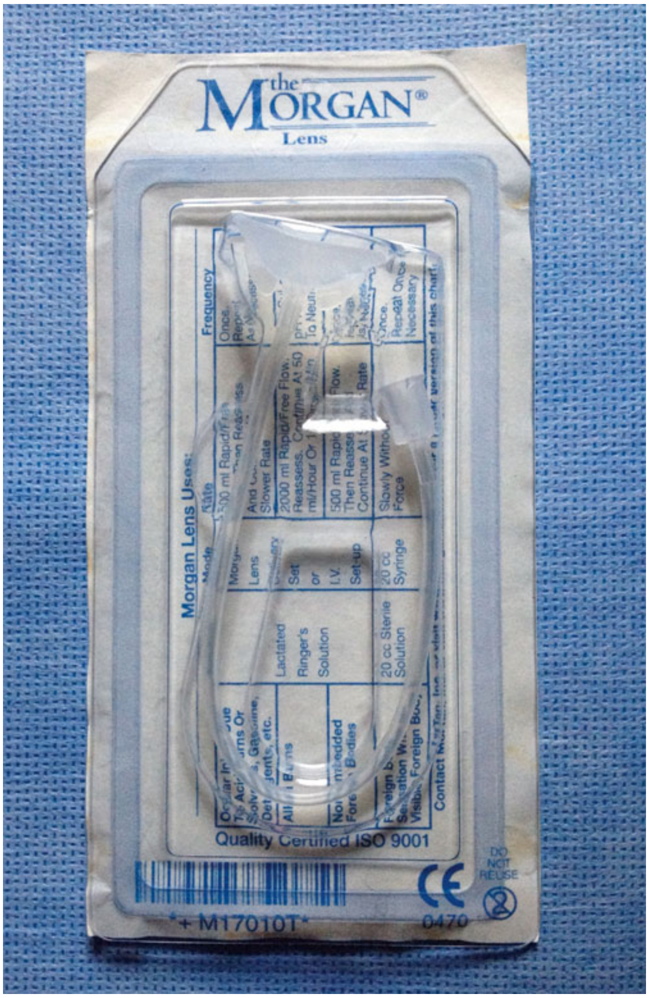
the MORGAN® Lens
Morgan Lens Uses:
Quality Certified ISO 9001
DO NOT REUSE
CE 0470
+M17010T

Fig. 47.1 The Morgan lens and packaging

Alcon®
Alcon Laboratories, Inc.
Ft. Worth, TX 76134 USA

Fig. 47.2 Examples of topical anesthetics

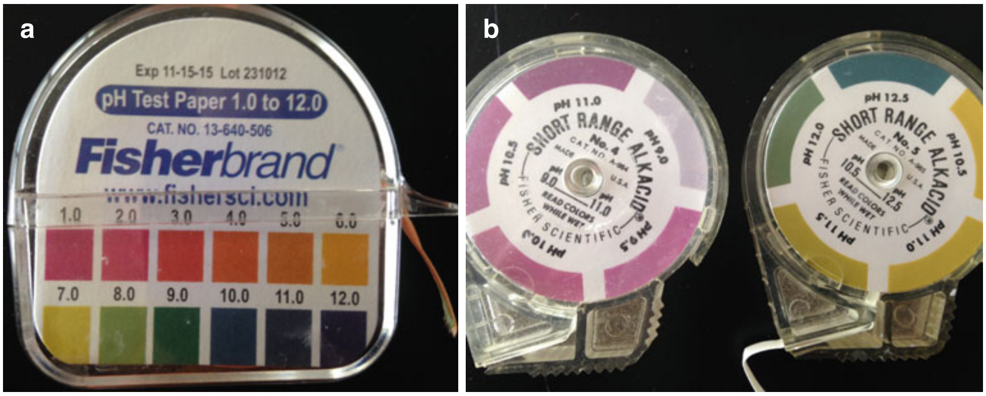
a
Exp 11-15-15 Lot 231012
pH Test Paper 1.0 to 12.0
CAT. NO. 13-640-506
Fisherbrand
www.fishersci.com
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
b
SHORT RANGE ALKACID
No. 4
No. 5
FISHER SCIENTIFIC
READ COLORS WHILE WET

Fig. 47.3 (**a**, **b**) Examples of pH paper

47.4 Procedure

1. Informed consent is generally not required, although a thorough explanation to the patient is warranted.
2. Obtain pH of the eye *before* instilling anesthetic drops.
3. Anesthetize the eye(s) with topical anesthetic of choice by instilling drops within the lower lid and then asking the patient to blink several times in order to evenly distribute the solution.
4. Ensure there are no foreign bodies on the inside of the upper lids by inverting the upper lid.
   - Particulate foreign bodies may be removed with moistened cotton tip applicators.
5. After adequate anesthesia is ensured, place one end of the Morgan lens within the fornix of the upper lid (Fig. 47.4).
6. Next, gently retracting the lower lid will ensure smooth placement of the remaining portion of the Morgan lens (Fig. 47.5).
7. Using the end of the Morgan lens, screw in the prepared IV tubing (Fig. 47.6).
8. Attach the end of the IV tubing to the saline bag and place at height to allow for gravity to ensure a smooth flow of solution.
   - Continue to irrigate the eye until desired pH is obtained.
9. To remove the Morgan lens, use the opposite technique for insertion.

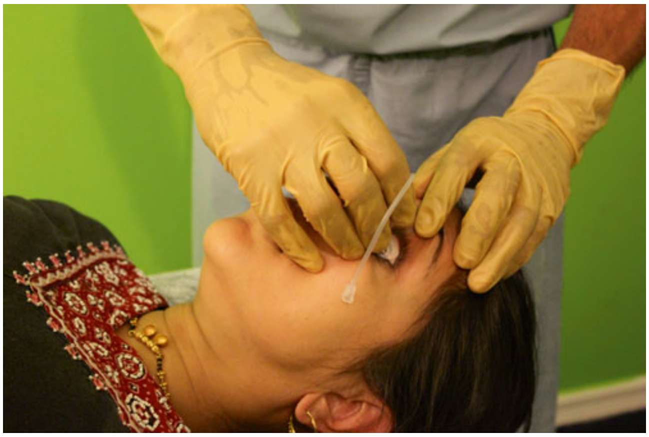

Fig. 47.4 Inserting the Morgan lens under the upper lid

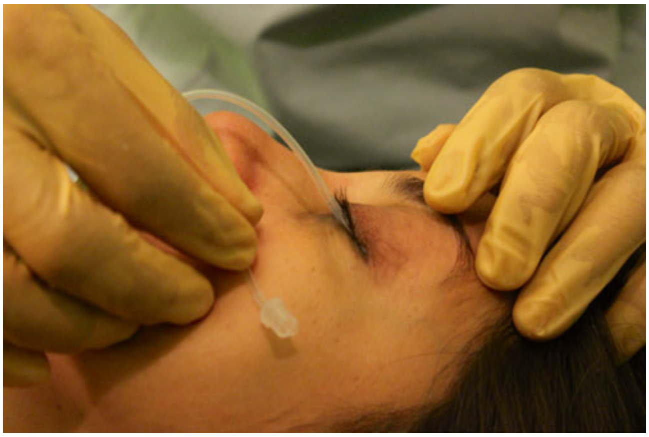

Fig. 47.6 The Morgan lens in place, ready to be attached to IV tubing for saline irrigation

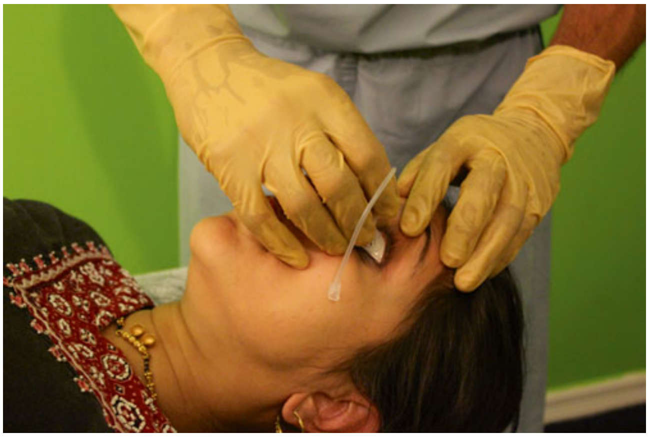

Fig. 47.5 Inserting the Morgan lens under lower lid

47.5 Complications

- Corneal abrasions may be caused by the Morgan lens and are treated in the usual fashion.
- Deep corneal injury may occur with inadequate irrigation.

47.6 Pearls and Pitfalls

- Note that alkali burns will require significant irrigation and more topical anesthesia may be required.
- Ophthalmological consultation may be required, especially for alkaline and hydrofluoric acid burns.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.
Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills Eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

Corneal Foreign Body Removal

48

Bobby K. Desai

48.1 Indications

- Presence of a corneal foreign body (Fig. 48.1)

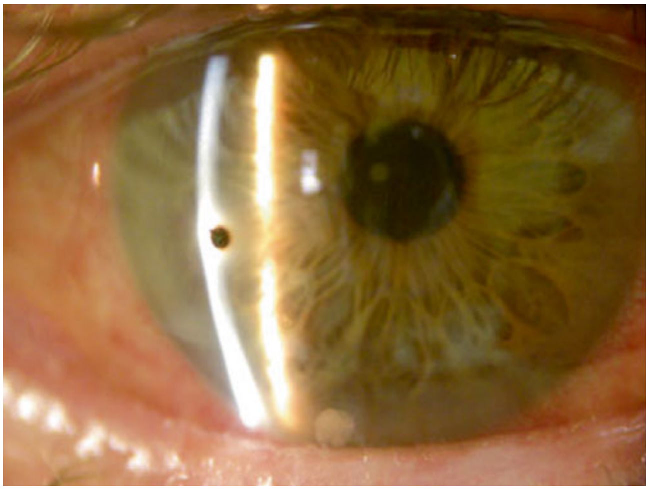

Fig. 48.1 Corneal foreign body (With kind permission from Springer Science + Business Media: Zuberbuhler B, Tuft S, Gartry D, Spokes D. Ocular Surface and Reconstructive Surgery. In: *Corneal Surgery*. 2013. 29–48)

48.2 Contraindications

- An uncooperative patient may require ophthalmological consultation as well as sedation.
- Suspected foreign bodies from high-velocity injuries must be referred to an ophthalmologist subsequent to initial evaluation.
- Any signs of globe penetration (e.g., hyphema) require emergent ophthalmological consultation.
- Overt globe rupture.
- Evidence of an inflammatory process such as iritis requires emergent ophthalmological consultation.
- Foreign bodies associated with corneal burns secondary to chemical exposure such as alkaline corneal burns will require emergent ophthalmological consultation.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_48

48.3 Materials and Medications

- Eye chart for visual assessment
- Cotton tip applicator
- Slit lamp or other source of magnification
- 27-Gauge needle or ophthalmic foreign body needle (Fig. 48.2)
- Topical ophthalmic anesthetic
- Ophthalmic burr (Fig. 48.3)

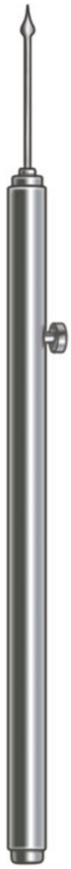

Fig. 48.2 Foreign body needle

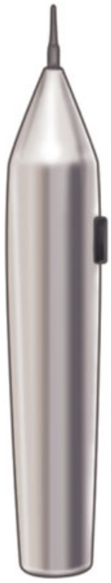

Fig. 48.3 Ophthalmic burr drill

48.4 Procedure

1. Informed consent may be required.
2. Visual acuity and a formal assessment of the eye should be done and documented.
3. Consider intraocular foreign body.
    - Computed tomography (CT) scan may be used to assess for the presence of an intraocular foreign body.
4. Eversion of the upper lid should be performed to evaluate for retained foreign body under the lid.
    - These usually can be removed by:
        - A moistened cotton tip applicator
        - Irrigation
5. Magnification is preferable, but some foreign bodies may be large enough to see with the naked eye.
6. Anesthetize the eye with topical anesthetic.
7. For multiple loose foreign bodies, removal may be attempted with irrigation after appropriate anesthesia.
8. Removal of the foreign body may be attempted with a moistened cotton swab.
    - Metallic foreign bodies recently embedded within the cornea may be difficult to remove in this fashion.
    - However, during the healing process, some metallic foreign bodies may be pushed closer to the surface during reepithelialization of the cornea. These potentially can be removed with this technique.
        - Care must be taken to avoid leaving a rust ring that could permanently stain the cornea.
9. A 27-gauge needle bent at a 90° angle may be used to remove foreign bodies by gently prying it out.
    - An advantage of using a needle is that any rust ring as well as any metallic foreign body can potentially be removed.
10. Technique: A slit lamp may be used to magnify the area.
    - See slit lamp chapter (Chap. 46) for details.
    - The patient should be informed about the procedure; formal informed consent may not be required.
    - Proper positioning is critical for success.
    - The patient's head should be fully forward and firmly placed against the head rest.
    - The patient's hands may be placed on the sides of the slit lamp stage for stabilization.
    - Apply topical anesthesia to the cornea.
11. Using the patient's cheek as a bolster potentially avoids significant movement of the needle if the patient unexpectedly moves.
    - Other modalities for stabilization include supporting the elbow by placing on a box or using towels.
12. Have the patient gaze at one point in the far distance.
    - Using the needle or burr as a scoop, gently manipulate the foreign body out.
        - Using a burr can allow for the total removal of any rust ring.
13. Once the foreign body is removed, patching is not required.
14. Consider application of topical antibiotics.
15. Tetanus prophylaxis should be given as for open other wounds.
16. Arrange follow-up with a primary care physician or preferably an ophthalmologist.

48.5 Complications

- Forceful attempts to dislodge a foreign body may result in corneal perforation.
- Incomplete removal of a foreign body.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

Thomas SH, White BA. Foreign bodies. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010. p. 715–32.

Ultrasound Evaluation of Retinal Detachment

49

Shalu S. Patel, L. Connor Nickels, and Rohit Pravin Patel

49.1 Indications

- To aid in the evaluation of vision loss or change in vision

49.2 Contraindications

- Relative: Suspicion of increased intraocular pressure
- Relative: Suspicion of globe rupture (see Chap. 50)

49.3 Materials and Medications

- Bedside ultrasound machine with high-frequency (7.5- to 10-MHz) linear transducer
- Transparent adhesive such as Tegaderm (optional)
- Sterile ultrasound gel
- Sterile gauze

49.4 Procedure

1. Position the patient supine with his or her eyes closed. The eyelid may be taped closed with a transparent adhesive if desired (Fig. 49.1).
2. Place a liberal amount of ultrasound gel over the eyelid.
3. Place the ultrasound probe over the eye in a transverse position.
4. Adjust the depth of the ultrasound so that the whole eye fills the screen.
5. Scan through the eye fully in the transverse and sagittal planes.
6. When viewing the ultrasound image, the normal eye is a circular hypoechoic structure (Fig. 49.2). The structures should be evaluated from anterior to posterior.
7. Identify the cornea. This is the first thin hypoechoic line in the anterior eye.
8. Continuing posteriorly, identify the anterior chamber, which is an anechoic area bordered by the cornea, iris, and lens. The iris is an echogenic linear structure, and the normal lens is anechoic.
9. Identify the vitreous chamber, the large anechoic region posterior to the lens.
10. Carefully evaluate the posterior globe. The normal retina cannot be distinguished from the other choroidal layers on ultrasound.
11. A detached retina will appear as a hyperechoic linear floating membrane in the posterior vitreous chamber (Fig. 49.3).
12. In the retrobulbar region, the optic nerve can be identified as a hypoechoic linear structure perpendicular to the globe (Fig. 49.3).

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA
e-mail: shalu314@gmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_49

Fig. 49.1 Probe positioning with Tegaderm applied to orbit: (**a**) sagittal view, (**b**) axial view

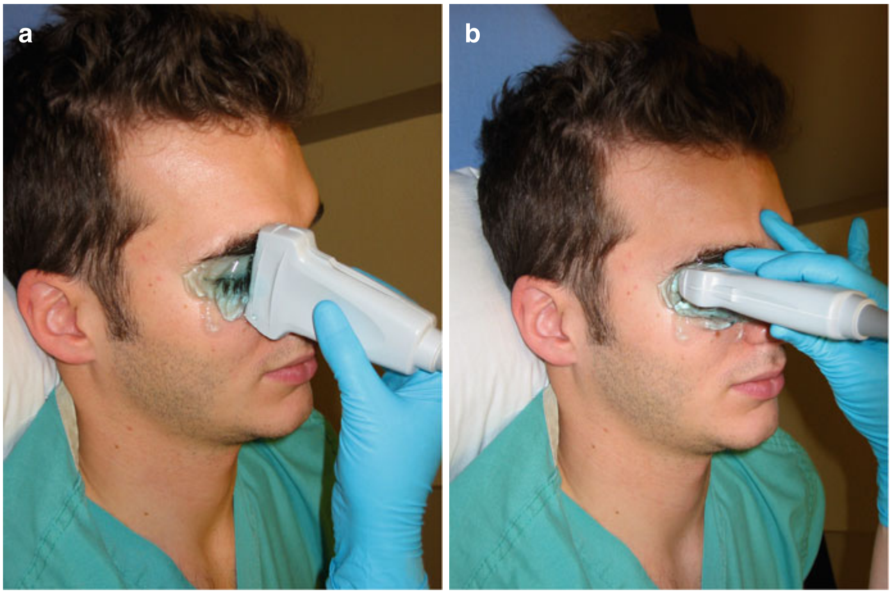

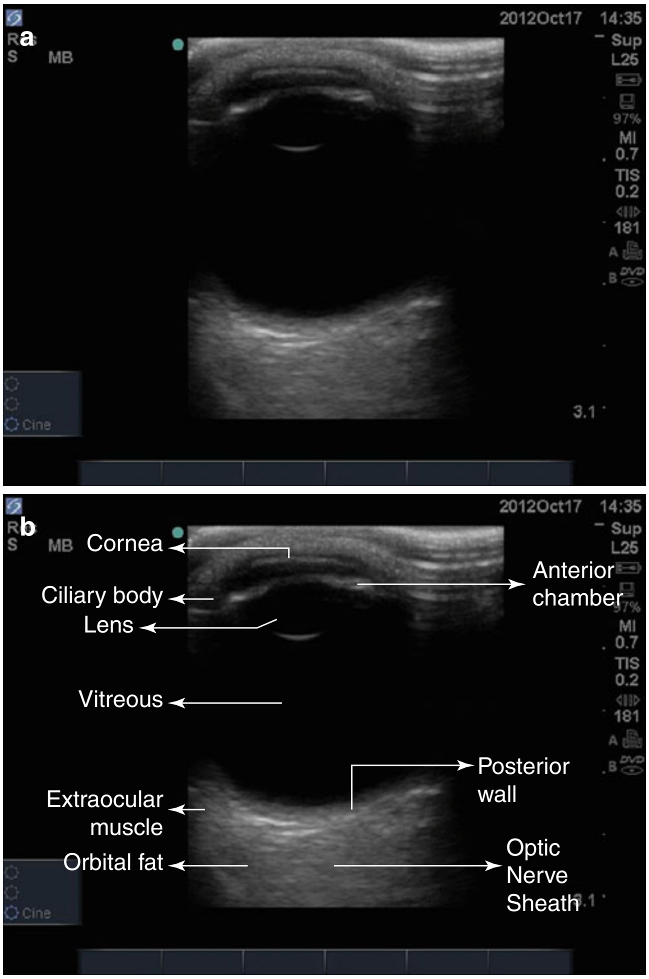
a
2012Oct17 14:35
MB
Sup
L25
97%
MI
0.7
TIS
0.2
181
3.1
Cine
b
2012Oct17 14:35
MB
Cornea
Ciliary body
Lens
Anterior chamber
Vitreous
Posterior wall
Extraocular muscle
Orbital fat
Optic Nerve Sheath
Cine

Fig. 49.2 Ocular ultrasound with normal anatomy findings ((**a**) without labels, (**b**) with anatomy labeled)

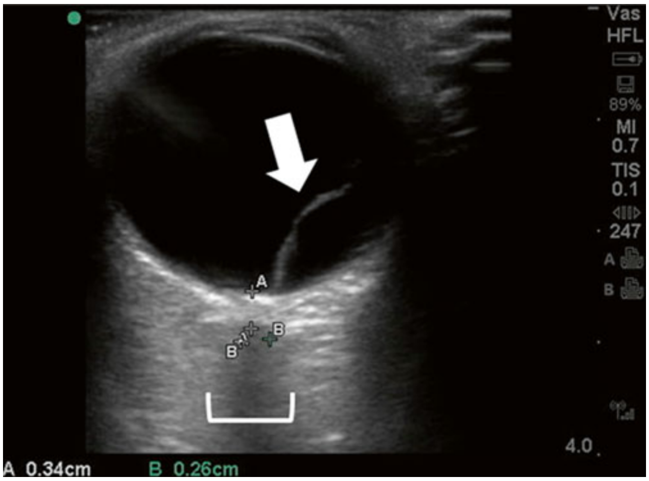
Vas
HFL
89%
MI
0.7
TIS
0.1
247
A
B
4.0
A 0.34cm
B 0.26cm

Fig. 49.3 Retinal detachment appears as a hyperechoic linear structure (*white arrow*). *White* bracket shows the optic nerve sheath shadow (see Chap. 50 for more information)

49.5 Complications

- Conjunctival infection
- Increased intraocular pressure (if too much pressure is applied)

49.6 Pearls and Pitfalls

- Pearls
  - Use a transparent adhesive such as Tegaderm to decrease the risk of conjunctival infection.
  - Use a liberal amount of gel to obtain the best images.
  - Use minimal pressure to obtain the best images.
  - False positives on ocular ultrasound may occur with disk edema or vitreous hemorrhage.
- Pitfalls
  - Subacute retinal detachments may be missed on ocular ultrasound.

Selected Reading

Blaivas M. Bedside emergency department ultrasonography in the evaluation of ocular pathology. Acad Emerg Med. 2000;7:947–50.

Blaivas M, Theodoro D, Sierzenski PR. Elevated intracranial pressure detected by bedside emergency ultrasonography of the optic nerve sheath. Acad Emerg Med. 2003;4:376–81.

Shinar Z, Chan L, Orlinksy M. Use of ocular ultrasound for the evaluation of retinal detachment. J Emerg Med. 2011;4:53–7.

Whitcomb MB. How to diagnose ocular abnormalities with ultrasound. AAEP Proc. 2002;48:272–5.

Ultrasonography in the Evaluation of Intraocular Pathology

50

Benjamin M. Mahon, Marie-Carmelle Elie,
L. Connor Nickels, and Rohit Pravin Patel

50.1 Indications

- Eye trauma
- Loss of vision or decreased vision
- Head injury
- Ocular pain
- Suspected foreign body

50.2 Contraindications

- None

50.3 Materials

- Linear probe (7.5–10 MHz)
- Tegaderm or other protective eye covering
- Copious amounts of water-soluble transmission gel

50.4 Procedure

1. Begin by asking the patient to close her or his eye, and then apply the Tegaderm or other suitable eye cover over both eyes.
2. Apply a copious amount of transmission jelly over both eyes.
3. Gently apply the ultrasound probe to the outer eyelid. If the operator has used a sufficient enough amount of jelly, the probe itself should not need to be in actual contact with the eyelid. This becomes of critical importance in suspected globe trauma or intraocular foreign body, when excess manipulation of the intraocular contents could precipitate further damage or worsening of present damage. Ensure the pressure is gentle enough to prevent this, but sufficient to obtain good imaging (see Chap. 49).
4. Image the normal eye first, to become familiar with the patient's anatomy (see Chap. 49), before proceeding to the suspected abnormal eye.
5. Start with low gain.
6. Ask the patient to move the eye slowly from the left to the right. This is called kinetic echography and will provide information elaborated upon later. Have the patient keep her or his eye straight ahead, as the operator phases up and down and side to side, between the transverse and the sagittal planes.
7. Increase gain slowly while the patient is moving the eye. Shifting from low gain to higher gain during the examination will help to prevent overlooking certain disease processes. As a fluid-filled structure, the eye provides its own acoustic window, and imaging is generally easy.

50.5 Complications

- Through excessive pressure applied to the eye, as described previously, further disruption of the intraocular contents and worsening of preexisting injury can occur.

50.6 Pearls and Pitfalls

- Pitfalls
  - Applying too much pressure to the globe and disrupting already damaged intraocular contents
  - Failing to image with full spectrum of gain
  - Using an insufficient amount of gel

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA
e-mail: faust2k9@ufl.edu

M.-C. Elie, MD, RDMS (✉) • L.C. Nickels, MD, RDMS
R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: elie@ufl.edu; cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_50

50.7 Specific Disease Processes/ Pathology [1]

50.7.1 Vitreous Hemorrhage (Fig. 50.1)

- Best seen in high gain.
- Clinical: Very common pathology. Seen as "floaters" on the visual field. Can cause blindness if large.
- Ultrasound: Appearance may vary. Early on in the evolution of a hemorrhage, these can be seen as small dots or mobile linear opacities that float freely in the back of the posterior chamber with eye movement. Has been described as a "snow-storm appearance." As hemorrhage ages, the blood organizes and membranes form, sometimes layering inferiorly owing to gravity.

50.7.2 Vitreous Detachment

- Best seen with higher gain.
- Clinical: Common in older patients. Presents as "flashers." Vitreous gel loses its attachment from the internal membrane lining.
- Ultrasound: Seen as "swaying seaweed." There are layering opacifications. Also seen as "snow-globe appearance." Concurrent vitreous hemorrhage is also common.
- Note: A small percentage of patients also have a retinal tear, so ophthalmology follow-up is necessary.

50.7.3 Retinal Detachment (Figs. 50.2 and 50.3)

- Best seen with lower gain.
- Clinical: Acute loss of vision, usually painless, often preceded by "flashers" and/or "floaters" as the neurosensory component of the retina pulls away from the retinal pigment epithelium. *An ophthalmological emergency.* Can be caused by ocular trauma or no known precipitant at all.
- Ultrasound: Seen as a taut, hyperechoic, linear opacity that moves with the eye, attached to the optic disc, in the posterior region of the globe.

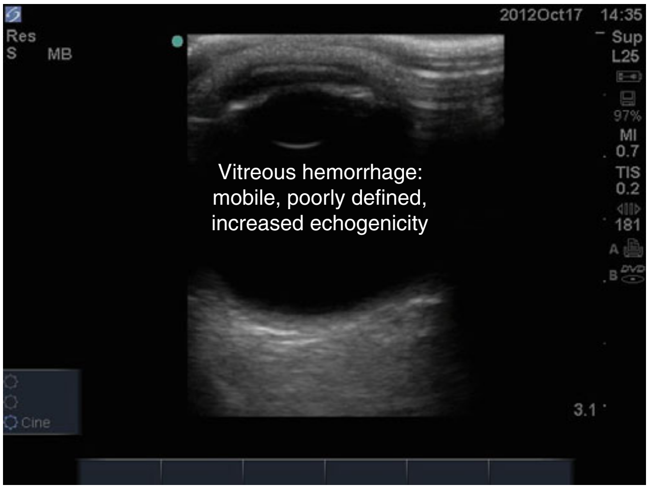

Fig. 50.1 Vitreous hemorrhage

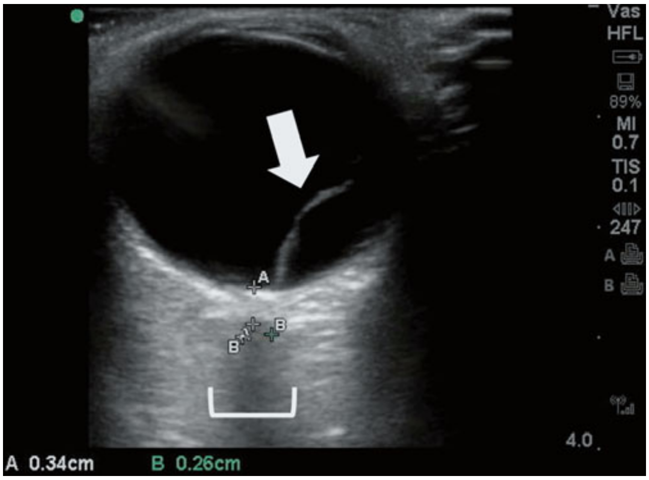

Fig. 50.2 Retinal detachment ultrasound image

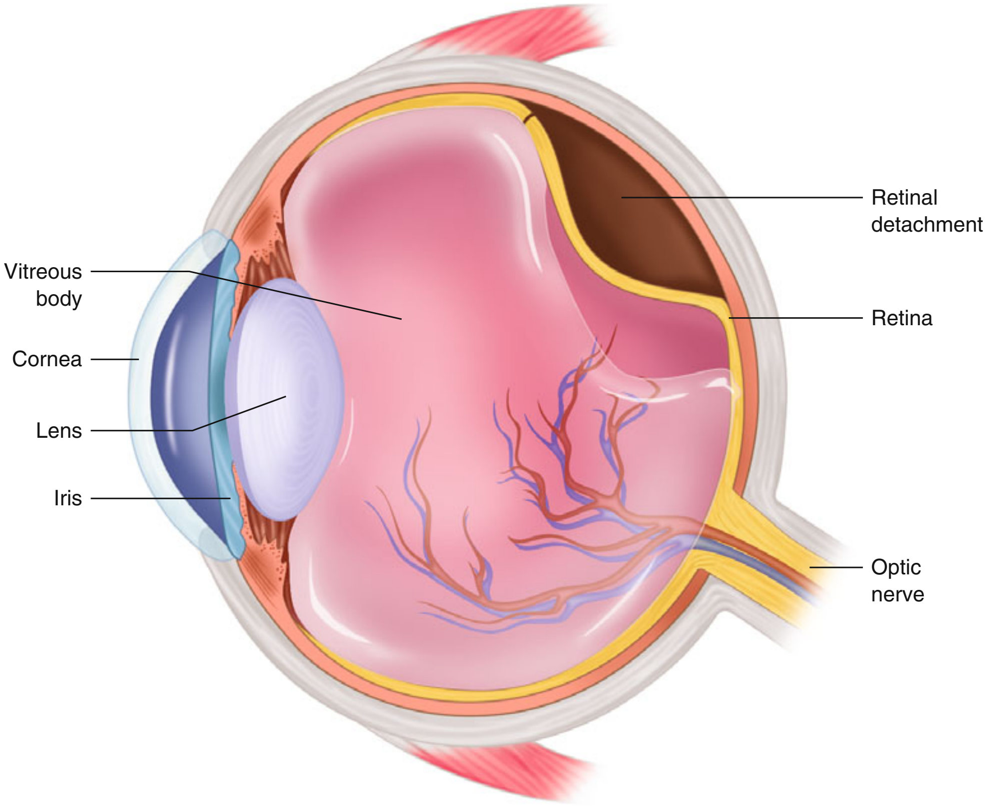
Retinal detachment
Vitreous body
Retina
Cornea
Lens
Iris
Optic nerve

Fig. 50.3 Retinal detachment graphical representation

50.7.4 Intraocular Foreign Body (Fig. 50.4)

- Clinical: Usually a preceding episode of trauma, often from metalworking or landscaping.
- Ultrasound: Foreign bodies are seen as hyperechoic substances in the eye often with distal shadowing. Certain shadow patterns distal to the foreign body, as well as comet tails, can help to differentiate the type of foreign body material.

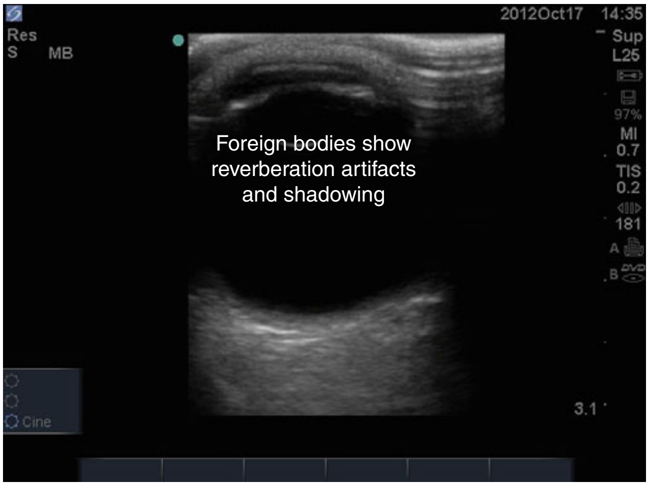
2012Oct17 14:35
Res
S MB
Sup
L25
97%
MI 0.7
TIS 0.2
181
A
B
Cine
3.1
Foreign bodies show reverberation artifacts and shadowing

Fig. 50.4 Foreign bodies

50.7.5 Globe Rupture (Fig. 50.5)

- Clinical: Usually the result of trauma. Frequently, the mechanism causing the injury has subsequently made the eye examination itself impossible, if edema, blepharospasm, hyphema, or distorted anatomy makes visualization of the posterior chamber too difficult. In these circumstances ultrasound can be immeasurably helpful.
- Ultrasound: Findings include decreased size of the anterior chamber, posterior chamber, or both, often with buckled sclera. Globe rupture should be suspected when the injury mechanism is present and any major distortion of the ocular anatomy is appreciated.

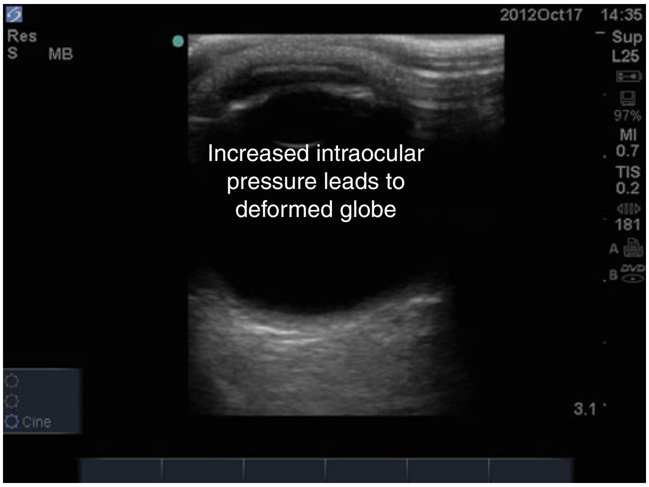
2012Oct17 14:35
Res
S MB
Sup
L25
97%
MI 0.7
TIS 0.2
181
A
B
Cine
3.1
Increased intraocular pressure leads to deformed globe

Fig. 50.5 Globe rupture

50.7.6 Increased ICP

- Clinical: Increased ICP is encountered in a number of pathological processes, such as closed head injury from trauma, altered mental status, pseudotumor cerebri, space-occupying lesion, or any other process causing intracranial injury. Physical examination may be suggestive of an intracranial process, but often key examination findings such as papilledema are not appreciated, owing to both limitations from the patient and their mental status, as well as limitations in the skill of the ED doctor, or the ability to dilate the eyes sufficiently, to appreciate the finding.
- Ultrasound: The size of the *optic nerve sheath diameter* is an indirect measurement of ICP. It is visualized as a hypoechoic linear strip radiating posteriorly from the edge of the back of the globe. A normal diameter is 5.0 mm or less. This measurement is taken at 3 mm back from the posterior edge of the globe (Fig. 50.6). Take measurements bilaterally to compare. Two or more measurements should be taken and then averaged on each side.
- A diameter greater than 5.0 mm, at 3.0 mm distal from the posterior globe border, is strongly suggestive of increased ICP and warrants further evaluation and imaging [2].

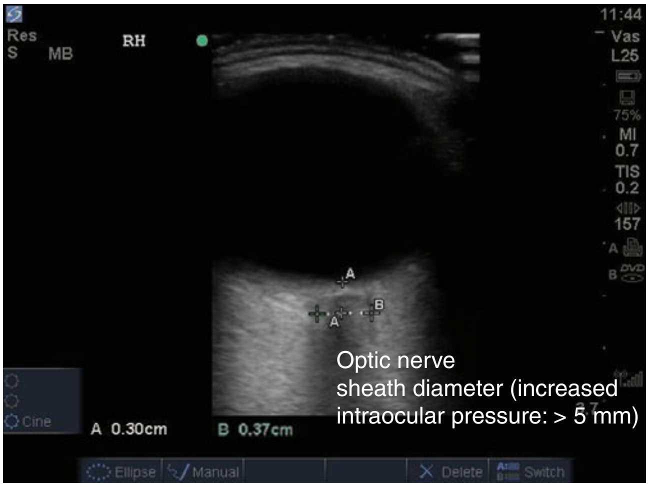

Optic nerve sheath diameter (increased intraocular pressure: > 5 mm)

Fig. 50.6 Optic nerve sheath measurement

References

1. Blaivas M, Theodoro D, Sierzenski P. A study of bedside ocular ultrasonography in the emergency department. Acad Emerg Med. 2002;9:791–9.
2. Blaivas M, Theodoro D, Sierzenski P. Elevated intracranial pressure detected by bedside emergency ultrasonography of the optic nerve sheath. Acad Emerg Med. 2003;10:376–81.

Selected Reading

Mustafa M, Montgomery J, Atta H. A novel educational tool for teaching ocular ultrasound. Clin Ophthalmol. 2011;5:857–60.

Qureshi MA, Laghari K. Role of B-scan ultrasonography in preoperative cataract patients. Int J Health Sci (Qassim). 2010;4:31–7.

Shinar Z, Chan IL, Orlinsky M. Use of ocular ultrasound for the evaluation of retinal detachment. J Emerg Med. 2011;40:53–7.

Vodapalli H, Murthy SI, Jalali S, Ali MJ, Rani PK. Comparison of immersion ultrasonography, ultrasound biomicroscopy and anterior segment optical coherence tomography in the evaluation of traumatic phacoceles. Indian J Ophthalmol. 2012;60:63–5.

Whitcomb MB. How to diagnose ocular abnormalities with ultrasound. AAEP Proc. 2002;48:272–5.

Yoonessi R, Hussain A, Jang TB. Bedside ocular ultrasound for the detection of retinal detachment in the emergency department. Acad Emerg Med. 2010;17:913–7.

Tonometry

51

Bobby K. Desai

51.1 Indications

- To measure intraocular pressure (IOP)

51.2 Contraindications

- A relative contraindication is the presence of an active or indolent infection about the cornea or conjunctiva, unless a one-time-use device is used (e.g., a Tono-Pen XL).
- Recent trauma to the eye.
- Uncooperative patients because improper technique may cause damage to the eye.

51.3 Overview

- Tonometry is the measurement of IOP obtained by evaluating the resistance of the eye to indentation by a force applied to it.
- Can be obtained by several methods.
- Two methods are routinely used in an emergency department setting:
  - *Impression tonometry*: measures the indentation made by a plunger carrying a known amount of weight.
  - More weight can be added to the apparatus.
  - The more weight needed to indent the cornea results in a higher IOP reading.
    - The Schiøtz tonometer is the most commonly used apparatus to utilize this method (Fig. 51.1).
- *Electronic indentation tonometry*: does not exert pressure on the eye (Fig. 51.2).

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_51

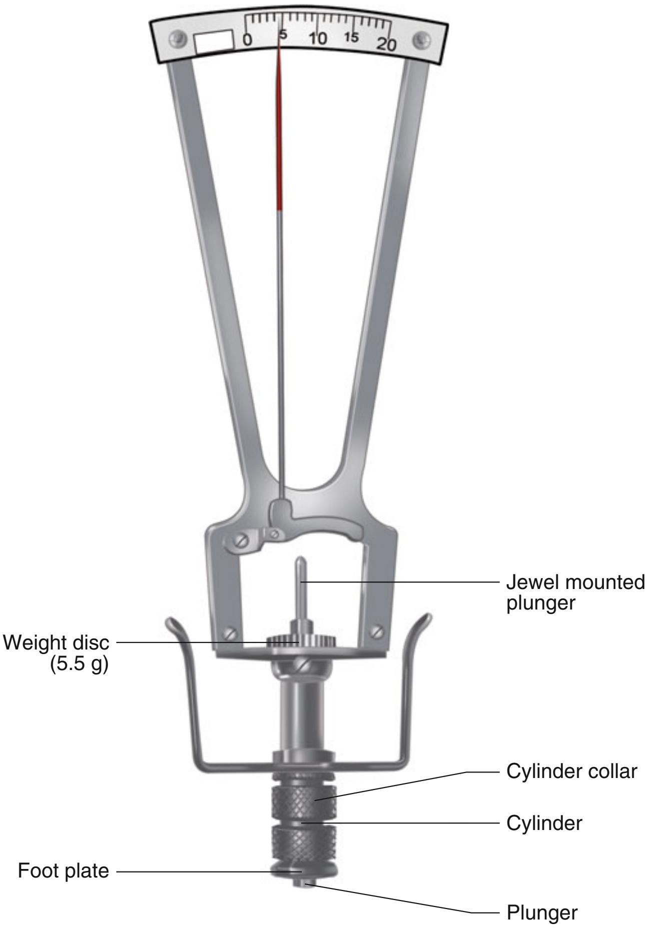
0
5
10
15
20
Jewel mounted plunger
Weight disc (5.5 g)
Cylinder collar
Cylinder
Foot plate
Plunger

Fig. 51.1 Schiøtz tonometer

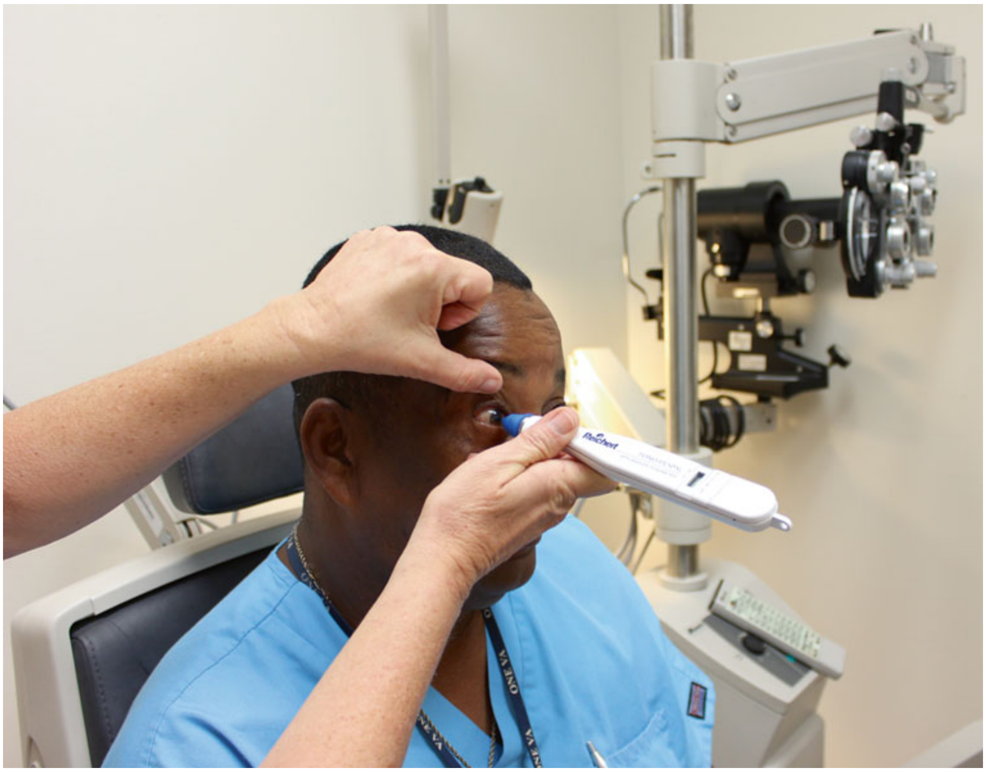

Fig. 51.2 Tono-Pen

51.4 Schiøtz Tonometry

51.4.1 Procedure: Schiøtz Tonometry (Fig. 51.3)

1. Carefully clean the apparatus between each patient.
    - Rubbing alcohol may be used.
    - Because it is metal, it can be autoclaved.
2. Calibrate the Schiøtz tonometer with the metal test block provided.
    - Test by placing the apparatus directly on the metal block.
    - The needle should be at “0” at the far end of the scale.
        - If not, loosen the screw at the base of the needle to rezero it.
        - Ensure that the needle is completely straight because any bend will produce an erroneous reading.
3. An explanation of the procedure is helpful because patient cooperation is critical for accurate results.
4. Anesthetize the eye with topical ophthalmic anesthetic of choice.
5. The patient should be in a recumbent position.
    - Have the patient focus on an area of the ceiling.
6. Hold the instrument with the aid of the curved arms at the side of the tonometer.
    - The operator can rest her or his hand on the patient’s cheek or forehead to maintain stability.
7. Gently rest the tonometer on the patient’s eye such that the instrument is centered on the eye and the instrument is completely vertical; no pressure should be exerted on the eye.
8. Note the scale reading.
9. Lift the Schiøtz directly off the cornea to avoid injury.
10. Using the table provided with the Schiøtz, the operator may convert the scale reading into the IOP.
    - The scale is inversely proportional to the actual IOP.
    - If the scale is low (i.e., high IOP), the additional weights provided with the instrument may be used and the patient retested.

51.4.2 Pitfalls

- Ensure the plunger is clean because it can transmit infection.
- False readings may be obtained without proper calibration.
- Placing pressure on the instrument will cause false readings.

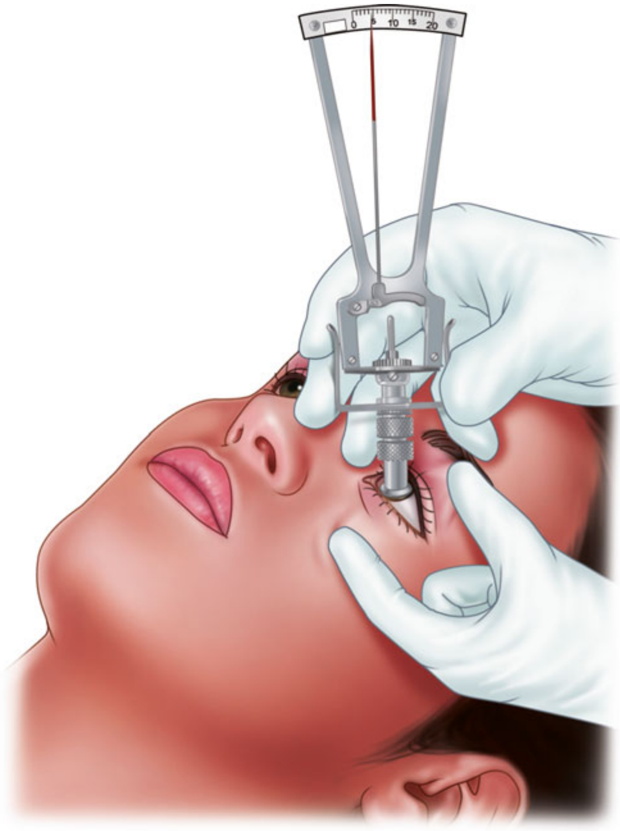

Fig. 51.3 Correct use of Schiøtz tonometer

51.5 Tono-Pen Tonometry

51.5.1 Overview

- Electronically measures IOP
  - Combines applanation and indentation tonometry
- Uses pressure-sensitive electronics to average four successive readings and displays the reading and a reliability factor digitally

51.5.2 Calibration of the Tono-Pen

- Should be performed once daily
- Hold the instrument with the tip down (Fig. 51.4).
- Press the black button twice.
  - The letters "CAL" will appear in the display.
- Press the black button again.
  - After a few seconds the word "UP" will appear.
- Rotate the instrument so that the tip points upward.
  - If "GOOD" appears, the Tono-Pen is ready for use.
  - If "BAD" appears, repeat the process until "GOOD" appears.
    - If "GOOD" does not appear, the device cannot be used.

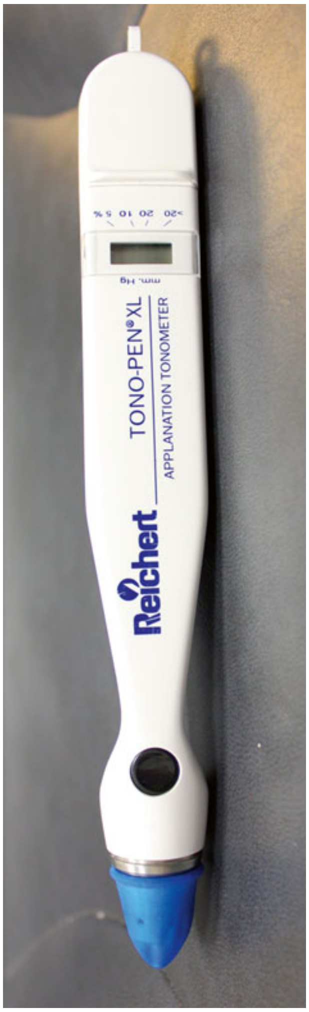

Fig. 51.4 Tono-Pen calibration position

51.5.3 Procedure: Tono-Pen Tonometry

1. An explanation of the procedure is helpful because patient cooperation is critical for accurate results.
2. Anesthetize the eye with the topical ophthalmic anesthetic of choice.
3. Place a probe cover on the unit.
4. The patient should be in a comfortable position because the unit can be used in any position.
    - Have the patient focus on a specific area.
5. The best way to hold the instrument is similar to that of a pen.
    - Ensure that the digital readout is visible.
6. The operator can rest his or her hand on the patient's cheek or forehead to maintain stability.
7. Hold the unit perpendicular to the patient's cornea.
8. Press the black button only once.
    - If "ICALI" is seen, followed immediately by a single row of dashes [– – – –], it indicates that the Tono-Pen requires calibration before it will measure.
    - If "====" is seen and a "beep" is heard, the unit is ready.
    - Proceed with measurement.
9. Gently touch the Tono-Pen to the cornea and withdraw (Fig. 51.5). Repeat several times.
    - Indentation is not required.
10. The unit will "chirp" and a digital reading will be displayed if a valid reading is obtained.
11. After four valid readings are obtained, the average of these measurements as well as a single bar that signifies statistical reliability will appear on the readout.
    - After the final beep if the liquid crystal display (LCD) readout shows "----,“ not enough valid readings were obtained.
    - In this case, the measurement must be repeated.

51.5.4 Pitfalls

- Unsuccessful calibration mandates a repeat attempt at calibration.
- Loosen the Ocu-Film tip cover and repeat for multiple failed attempts.
- Press the reset button and reattempt calibration.
- Replace the battery if necessary.
- If all else fails, use another device or Schiøtz tonometer.

Fig. 51.5 Proper use of Tono-Pen

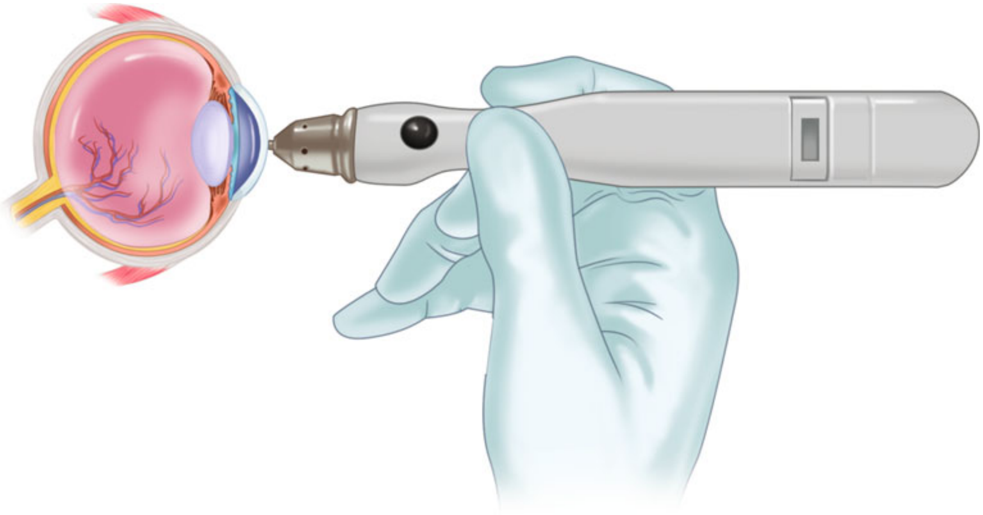

Correct

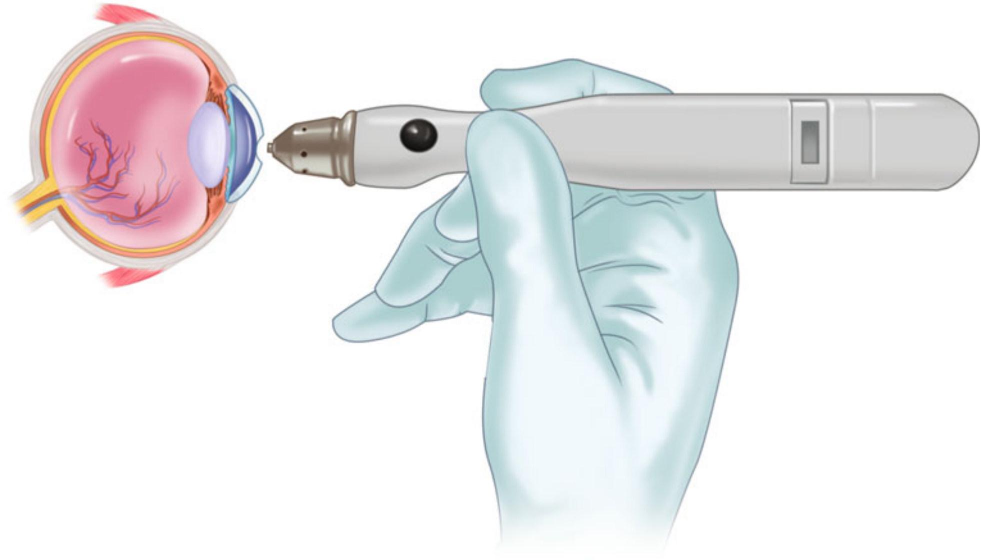

Incorrect

51.6 Complications

- Unusual if proper technique is used.
- If too much force is applied, a corneal abrasion may result.
- Infection when using an improperly sterilized Schiøtz tonometer.

Selected Reading

Lang GK. Ophthalmology: a short textbook. Stuttgart: Thieme; 2000.

Rhee DJ, Pyfer MF, Rhee DM, editors. The Wills eye manual: office and emergency room diagnosis and treatment of eye disease. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 1999.

Lateral Canthotomy

52

Benjamin M. Mahon and Bobby K. Desai

52.1 Indications

- Absolute indications: In the presence of presumed or confirmed retrobulbar hemorrhage
  - – Acute visual loss
  - – Intraocular pressure (IOP) greater than 40 mmHg (normal IOP is 10–20 mmHg)
  - – Proptosis not amenable to retropulsion
- Relative indications: In the context of presumed or confirmed retrobulbar hemorrhage (Fig. 52.1)
  - – Ophthalmoplegia
  - – Cherry red macula
  - – Profound eye pain
  - – Afferent pupillary defect (Marcus Gunn pupil)
    - This defect is seen with the *swinging flashlight test.*
      - – Shine a light into both eyes; a normal response is equal constriction of both pupils.
      - – In those patients with an afferent papillary defect, when light is swung from an unaffected pupil to an affected pupil, the pupil will seem to paradoxically dilate, rather than constrict. The pupil on the nonaffected side will similarly dilate as light is shown into the affected eye. This results from injury to the afferent fibers of cranial nerve (CN) II on the affected side, while the efferent fibers, innervated by CN III, remain intact.
  - – This procedure is most effective if performed as soon as possible because irreversible vision loss can occur secondary to ischemia in as little as 90 min.

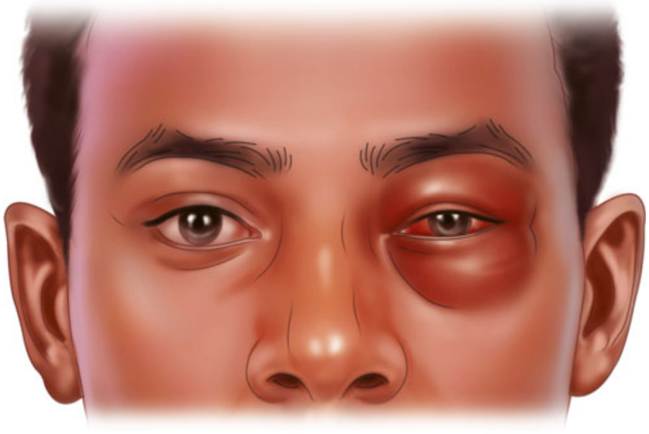

Fig. 52.1 Retrobulbar hemorrhage

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_52

52.2 Contraindications

- Absolute
  - Globe rupture
- Relative
  - Suspected globe rupture: Heralded by a number of signs including hyphema, irregularly shaped pupil, exposed uveal tissue, or profound restriction of extraocular movement

52.3 Materials and Medications

- Straight hemostat
- Suture or iris scissors
- Forceps
- 1–2 % lidocaine with epinephrine
- Sterile gloves
- Sterile towels/drapes
- 4×4 gauze
- Face shield/mask
- Large-bore needle to withdraw anesthetic, 25-gauge needle to infiltrate
- 10-cc syringe
- Normal saline

52.4 Procedure

1. Position the patient
   - While waiting for the procedure to begin, the patient should initially be upright to produce any decrease in IOP that can be provided.
   - At the start of the procedure, lay the patient supine.
   - A cooperative patient is absolutely necessary because even slight movements can cause devastating iatrogenic injury.
   - If the patient is anxious or conscious, sedation may be required.
   - In extreme cases of altered mental status or combative trauma patients, endotracheal intubation and mechanical ventilation may be required.
2. Gently irrigate the affected eye to remove any debris.
3. Inject approximately 1–2 mL of 1–2 % lidocaine with epinephrine into the lateral canthus of the affected eye (Fig. 52.2).
4. Using a curved hemostat, gently crimp the skin over the lateral corner of the patient's eye down to the orbital rim for 3 min to establish hemostasis and set the boundaries for the incision.
   - Using forceps, use the hemostats to pick up the skin just crushed.
   - Raise the skin with forceps and then use scissors to make a 1- to 2-cm incision from the lateral corner of the eye extending laterally outward (Fig. 52.3).
     - This incision can often sufficiently decrease IOP.
     - It is feasible to remeasure the IOP at this point, and if it is still greater than 40 mmHg, proceed to the next step.
   - Visualize the lateral canthal tendon by retracting the inferior orbital lid inferiorly.
   - Direct the scissors along the lateral side of the orbital rim away from the globe, and cut the lateral canthal tendon inferiorly (Fig. 52.4).
   - At this point, again measure IOP, and if greater than 40 mmHg, proceed to the next step.
   - Direct the scissors along the lateral side of the orbital rim away from the globe, this time directed superiorly, and cut the superior crux of the lateral canthal tendon.

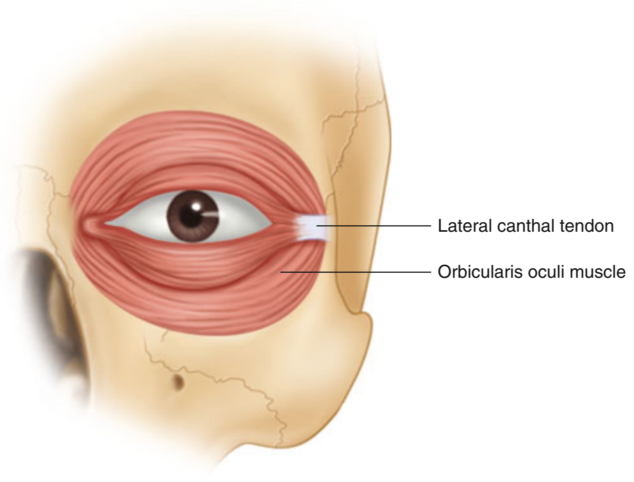
Lateral canthal tendon
Orbicularis oculi muscle

Fig. 52.2 Lateral canthal tendon

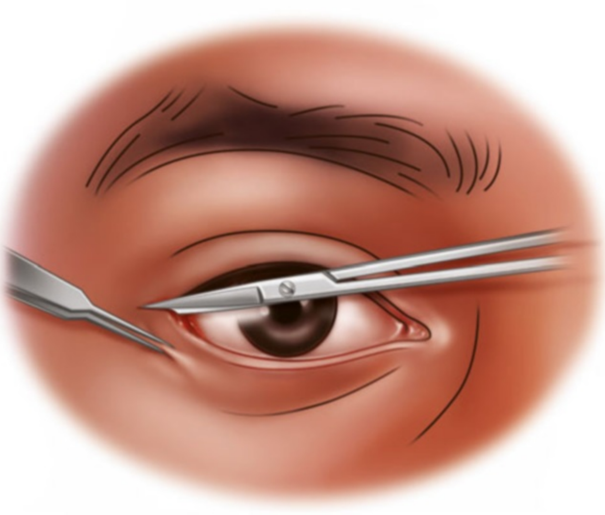

Fig. 52.4 Further cut to reduce intraocular pressure

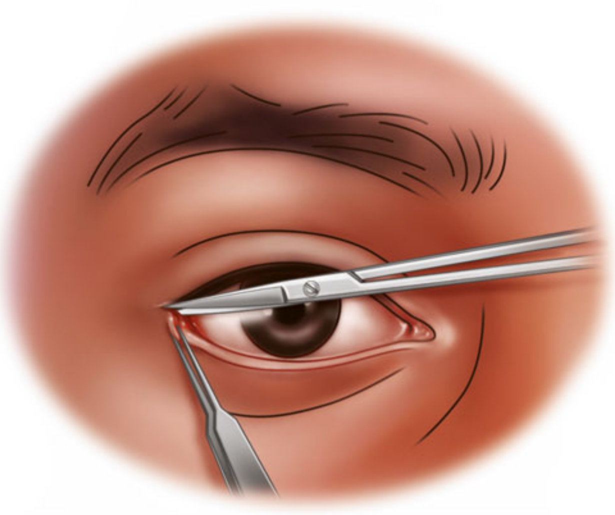

Fig. 52.3 Initial cut

52.5 Complications

- If one is making a superior incision, caution must be taken during this procedure to avoid injury to the lacrimal gland or artery.
- An additional complication of the procedure is ptosis due to iatrogenic injury to the levator aponeurosis, either partially or completely.
- The most obvious complication involves direct injury to the globe itself due to either operator error or poor patient control and immobility.
- Hemorrhage from inadequate hemostasis.
- Infection.

52.6 Pearls and Pitfalls

- Abrupt decrease in pain, decrease in IOP, and resolution of afferent pupillary defect will herald a successful procedure, assuming it was performed within an early enough time frame.
- Having an assistant present is beneficial.
  - In this case, the role of the assistant is to provide lateral retraction of the tissues to decrease likelihood of globe rupture.
- After cantholysis of the inferior component of the lateral canthal tendon, the lower lid will become lax as its attachment to the lateral wall is separated. This signifies a successful incision.
- Incisions made during this procedure generally heal without the need for suturing.
- Emergent ophthalmological consultation should ideally be sought before beginning this procedure.
- Emergent ophthalmological consultation is mandatory subsequent to the procedure.

Selected Reading

Goodall KL, Brahma A, Bates A, Leatherbarrow B. Lateral canthotomy and inferior cantholysis: an effective method of urgent orbital decompression for sight threatening acute retrobulbar haemorrhage. Injury. 1999;30:485–90.

McInnes G, Howes DW. Lateral canthotomy and cantholysis: a simple vision-saving procedure. CJEM. 2002;4:49–52.

Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: WB Saunders; 2009: Chap. 63.

Vassallo S, Hartstein M, Howard D, Stetz J. Traumatic retrobulbar hemorrhage: emergent decompression by lateral canthotomy and cantholysis. J Emerg Med. 2002;22:251–6.

Part VIII

Otorhinolaryngologic Procedures

Epistaxis Control

53

Benjamin M. Mahon and Bobby K. Desai

53.1 Etiology

- Anterior source: approximately 90 % of nosebleeds, usually Kiesselbach's plexus (Fig. 53.1)
  - Nose picking ("epistaxis digitorum")
  - Trauma
  - Infection
  - Nasal foreign body
  - Dry air
  - Atmospheric pressure alterations (e.g., increased altitude, lower arterial partial pressure of oxygen [PaO_2])
  - Allergies
  - Blood dyscrasias
  - Malignancy (e.g., leukemia, lymphoma)
- Posterior source: approximately 10 % of nosebleeds, usually sphenopalatine artery (Fig. 53.2)
  - Iatrogenic coagulopathy (e.g., warfarin, heparin, high-dose aspirin)
  - Blood dyscrasia
  - Liver failure
  - Renal failure
  - Malignancy
  - Older age

B.M. Mahon, MD
Ponciana Medical Center, Kissimmee, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_53

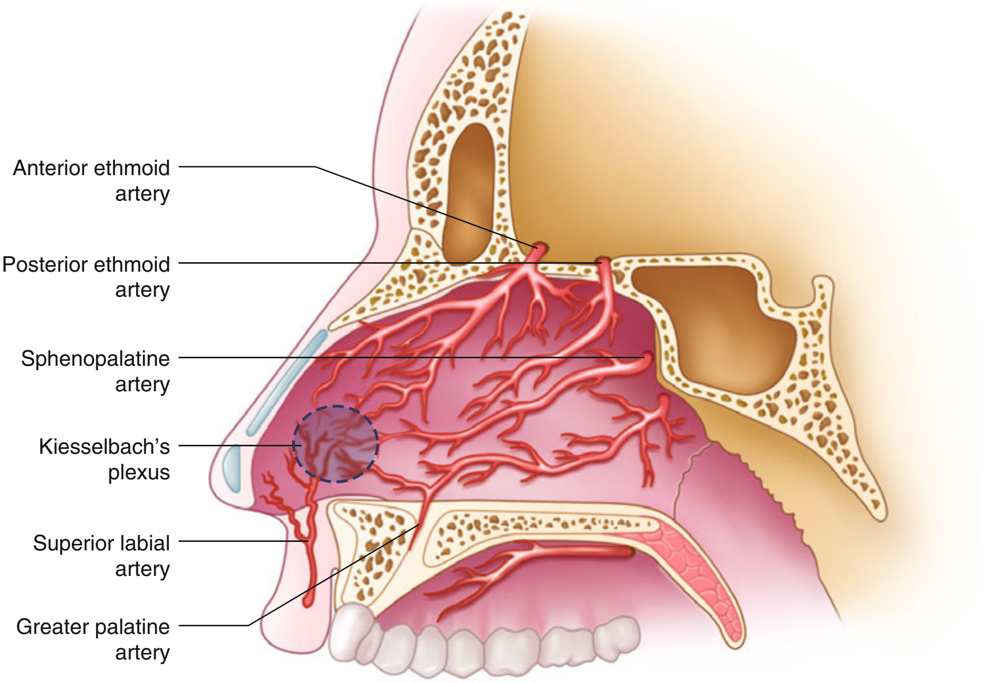
Anterior ethmoid artery
Posterior ethmoid artery
Sphenopalatine artery
Kiesselbach's plexus
Superior labial artery
Greater palatine artery

Fig. 53.1 Vascular supply to the nasal septum

Fig. 53.2 Vascular supply of the lateral wall of the nose

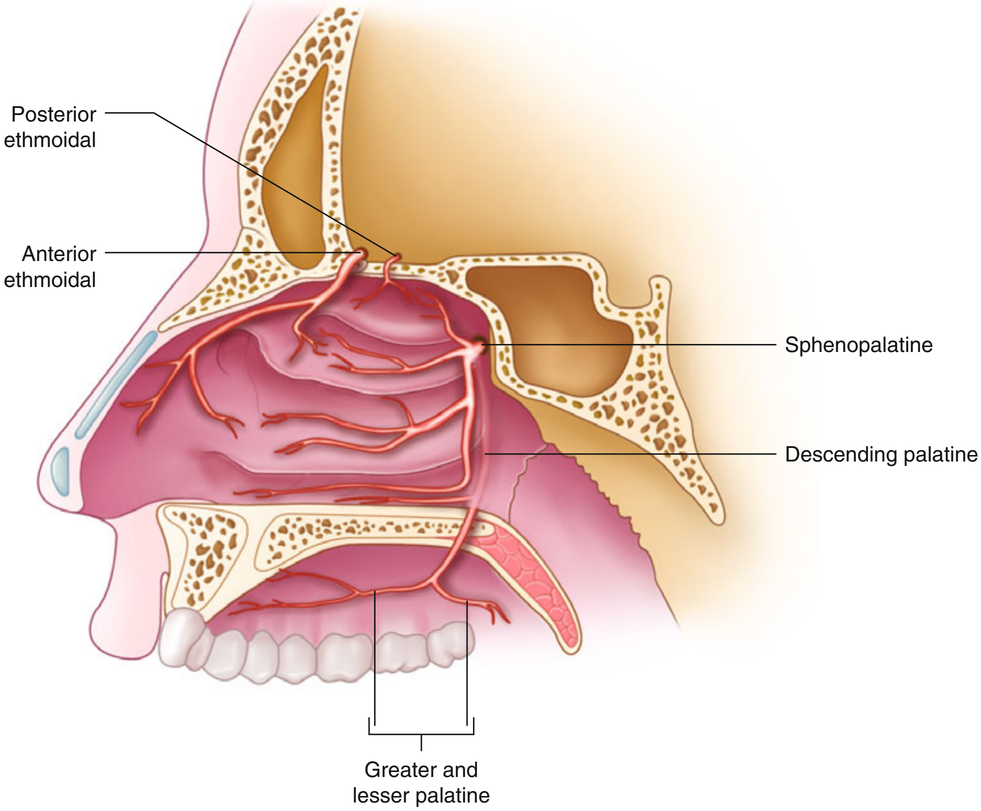
Posterior ethmoidal
Anterior ethmoidal
Sphenopalatine
Descending palatine
Greater and lesser palatine

53.2 Indications

- Acute/recurrent epistaxis

53.3 Contraindications

- Resolution of epistaxis episode before arrival
- Massive facial trauma

53.4 Materials and Medications

- Headlight with focused beam
- Gown
- Gloves
- Mask with visor
- Full-body drape or sterile towels
- Nasal packing material (any or all of the following, as needed [see later]): several cotton pledgets, 3–5 feet of ½-inch ribbon gauze (preferably enriched with petroleum jelly and bacitracin), anterior epistaxis balloon nasal pack
- Topical vasoconstrictor (e.g., 1 % pseudoephedrine or 1:1000 epinephrine)
- Topical anesthetic (e.g., 4 % lidocaine solution or 2 % tetracaine)
- Nasal speculum
- Bayonet forceps
- Anterior epistaxis balloon (e.g., Rhino Rocket)
- Any of the following commercial products, as needed (see later): Gelfoam, Surgicel, Crosseal, and FloSeal
- Foley catheter

53.5 Procedure: Anterior Epistaxis

- Initial preparation
  1. Assemble necessary equipment at the bedside.
  2. Have the patient seated upright, with head and neck in the sniffing position.
  3. Universal precautions are mandatory: gown, glove, and mask.
  4. Drape the patient with towels or a large body drape.
  5. If initial evaluation using the headlamp and nasal speculum with gentle spreading in a vertical fashion or a simple visual examination reveals the source (anterior vs posterior), proceed with the appropriate management pathway as determined later.
     – If clot obstructs visualization, it is recommended for the patient to gently blow her or his nose once to remove any easily friable clots, and then proceed with immediate inspection. If anterior nose bleeds, go to the next step.
- Management of anterior epistaxis:
  1. Initial step in the management of anterior nosebleeds involves pinching the patient's nostrils firmly for at least 10–15 min.
     – This can be done by the patient.
     – Simple pressure is very often enough to provide appropriate hemostasis in anterior nosebleeds. If it fails, proceed to the next step.
  2. A topical vasoconstrictor of choice (e.g., cocaine, oxymetazoline, 1 % pseudoephedrine) can be applied to the septum and lateral walls of the nose either topically with cotton-tipped applicators or by using soaked pledgets.
  3. Continue to apply firm pressure to the nares for 10–15 min, and recheck the nose for evidence of bleeding.
  4. If the bleeding site can be easily visualized, chemical cautery may be attempted with silver nitrate (Fig. 53.3).
     – This can be accomplished with the help of a nasal speculum and a headlight.
     – It is advised to cauterize both the site itself and 0.5 cm around the site, by holding the silver nitrate stick to the site for at least 20 s until the bleeding stops.
     – If bleeding then stops, cover the site with an absorbable gelatin foam or an oxidized cellulose such as Gelfoam or Surgicel or simply apply topical antibiotic to the site.
  5. If there is continued bleeding, the clinician may apply either two elongated cotton pledgets or a commercially available substitute (e.g., a Rhino Rocket or similar device [Fig. 53.4]) presoaked in a few milliliters or a 1:1 mixture of a topical vasoconstrictor and an anesthetic. Combinations can include the previously cited two agents, plus 2 % tetracaine or 4 % lidocaine.
  6. Universal instructions for the insertion of a commercially available anterior nasal pack include (Fig. 53.5):
     - Soak the fabric via the manufacturer's recommended solution if required.
     - Insert antibiotic ointment in the nares to facilitate insertion.
     - Insert the entire length of the balloon along the inferior surface of the nasal cavity.
     - The fluid within the nasal cavity should cause it to expand spontaneously while in the nares, after which the remaining exposed string can be taped to the cheek.
     - For those devices that require air instillation, read the manufacturer's instructions on how much air to instill within the device.

7. Another option is to provide an anterior nasal packing using a few feet of ½-inch ribbon gauze (Fig. 53.6).
8. It is preferable if this gauze is impregnated with petroleum jelly and is enriched with bacitracin ointment.
   - If it is not, bacitracin can be applied to the strip before its insertion into the nasal cavity.
9. To facilitate the insertion, bayonet forceps can be used to gently lay each successive strip on top of the one before it.
10. Begin inferiorly, laying each strip on top of the other, about three or four layers at a time, with the end points protruding from the opening of the nasal cavity.
11. Pack the strands firmly.
12. It may take several feet to pack the entire nasal cavity.
13. This may be both uncomfortable for the patient and time consuming for the provider, so ensure adequate time and copious topical anesthetic.
14. Parenteral analgesic or anxiolytics may be required for any of these techniques.
15. If none of these resources is available, another technique requires the use of two cotton pledgets gently inserted into the nostrils after vasoconstrictor/anesthetic topical solution has been sprayed into the nose either manually or with the addition of bayonet forceps, where they will remain for 20 min, supplemented by firm pressure to the nares as described previously.
    - If the bleeding has then stopped, they can be removed at this time. If it does not, it is advised to repeat the previous step a second time, this time with a longer period of pressure, before proceeding to the next step.
16. If all of these techniques fail to control bleeding, the clinician may apply a sealant spray or foam enriched with thrombin to enhance clotting.
    - Commercially available options include Crosseal or FloSeal.
17. Finally, if none of these techniques manage to halt what one is sure is only an anterior, albeit persistent, nosebleed, consult the otolaryngologist for further guidance.
    - Another consideration at this point may be that the uncontrollable anterior nosebleed is really a posterior nosebleed from an undetermined location.
    - It may be prudent to proceed to the algorithm for treatment of posterior epistaxis, even if a posterior bleed is not specifically confirmed.

53.5.1 Complications

- Using an inadequate amount of packing when inserting the ribbon gauze, the entire product may serve as a plug with potential aspiration risk, rather than as a tool for hemostasis.
- Otitis, sinusitis, and toxic shock syndrome.

53.5.2 Pearls and Pitfalls

- Pearls
  - Be sure to perform a reevaluation of both the posterior oropharynx and the posterior nasopharynx after insertion of any of the previously discussed devices, and inspect for continued bleeding/oozing to ensure hemostasis has been accomplished.
  - For those patients with nasal packs in place, consider the use of oral antibiotics to avoid subsequent sinusitis, otitis, or toxic shock syndrome.
    - Appropriate antibiotic choices include cephalexin, clindamycin, or amoxicillin–clavulanic acid.
    - The use of antibiotics for short-term packing has not been proven.
  - Consider admission for elderly patients or for those patients with potential airway complications such as those patients with chronic obstructive pulmonary disease (COPD) and for those at risk for aspiration.
  - Have the patient follow up with otolaryngology within 2–3 days for reevaluation and packing removal.
  - Provide adequate discharge instructions warning the patient against nose blowing, sneezing with the mouth closed, or any movements or actions that cause a Valsalva maneuver.
  - The patient should try to keep his or her head elevated for the next 24–48 h even while sleeping to prevent potential aspiration.
- Pitfalls
  - Discharging a patient without an adequate period of observation.
  - Failing to ensure appropriate follow-up.
  - Providing inadequate discharge instructions.
  - Mistaking an anterior nosebleed for a posterior nosebleed.
  - Leaving silver nitrite on exposed nasal mucosa can lead to iatrogenic nasal septal perforation.
  - Not informing the patient that minimal oozing of blood from the nose can be expected.

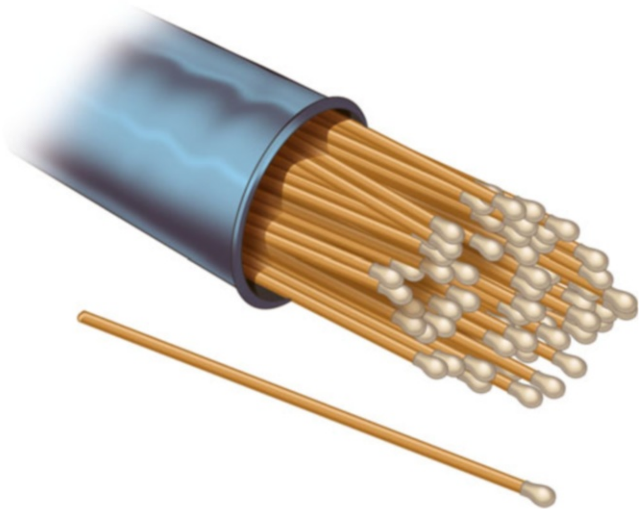

Fig. 53.3 Example of silver nitrate stick and packaging

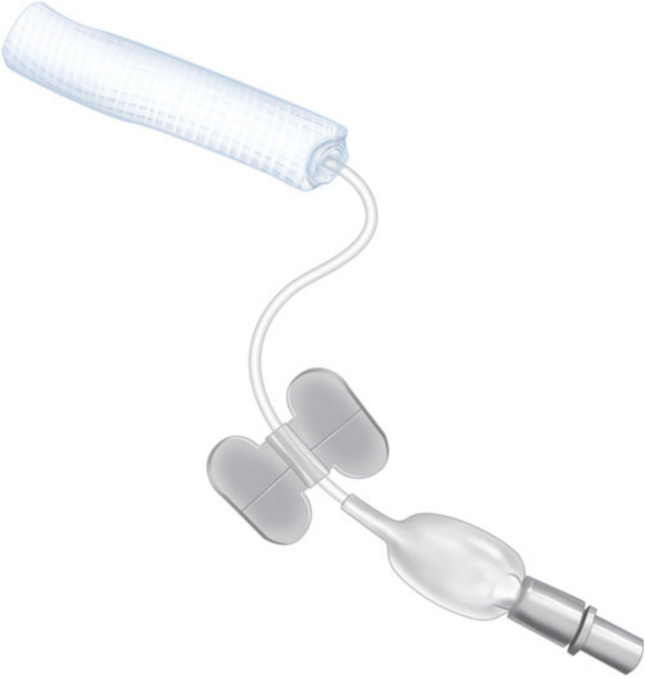

Fig. 53.4 Example of commercial nasal pack

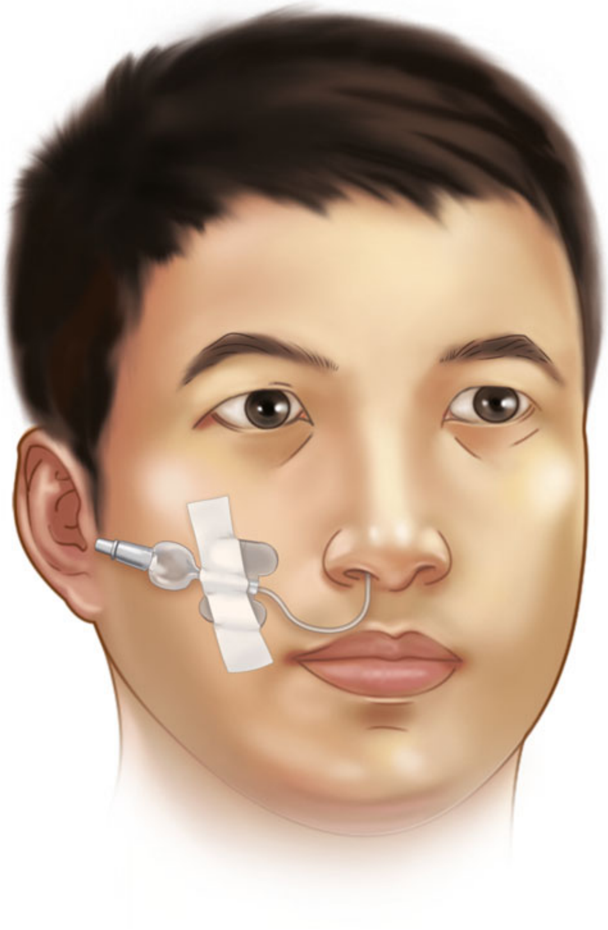

Fig. 53.5 Proper taping of a commercial anterior nasal pack

Fig. 53.6 Gauze method

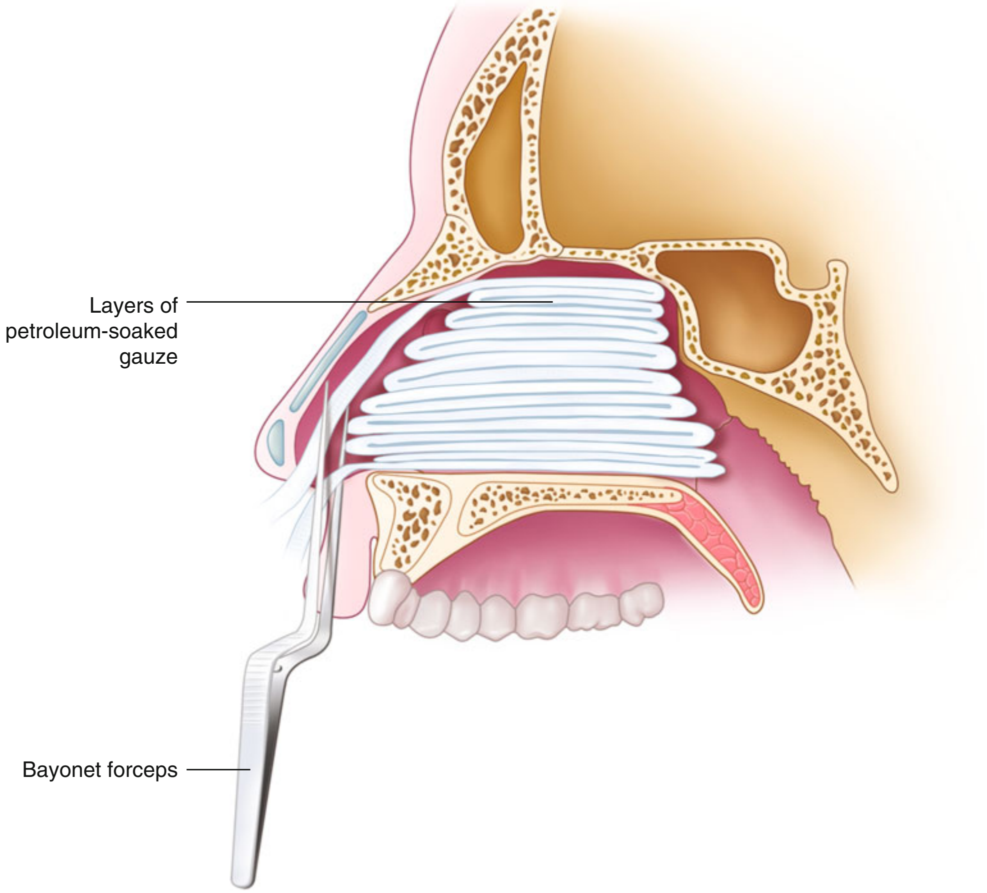
Layers of petroleum-soaked gauze
Bayonet forceps

53.6 Procedure: Posterior Epistaxis

- Preliminary Steps
  1. Owing to the possibility of extreme exsanguination from a posterior source of epistaxis or airway obstruction from a large clot dislodgment or large volume of blood aspiration, the stability of the patient's airway, breathing, and circulation must be ensured first.
     - Appropriate interventions entirely depend on the presentation of the patient and therefore are not described in this chapter.
     - Patients with posterior epistaxis can lose a large volume of blood quickly, so this needs to be an initial consideration.
  2. Consider establishing intravenous access and placing the patient on a cardiac monitor.
  3. Because patients with posterior epistaxis tend to be either older, on various blood thinners, or to potentially have some type of coagulopathy, consider obtaining a complete blood count, prothrombin time, and partial thromboplastin time as part of the preliminary workup.
  4. Consider a blood type and screen.
- Management of posterior epistaxis:
  1. Inspect the nares as described in the previous section on anterior nosebleeds.
  2. Several of the techniques discussed may be attempted but may not be successful.
  3. Assemble necessary equipment at the bedside.
  4. Have the patient seated upright, with the head and neck in the sniffing position.
  5. Universal precautions including a gown, glove, and mask should be employed.
  6. Drape the patient with towels or a large body drape.
  7. Apply a combination of a topical vasoconstrictor and anesthetic to the nose.
     - An appropriate choice (e.g., 1 % pseudoephedrine or 1:1000 epinephrine in a 1:1 ratio with 4 % lidocaine or 2 % tetracaine)
     - Be aware, however, that there are situations in which the bleeding may be too brisk, and this step may need to be avoided owing to time constraints and the volume of bleeding. So it is not unheard of to be required to intervene in posterior epistaxis without this initial step.
  8. Otolaryngology should be consulted as early as possible in the workup of this condition, both for admission and in case the following techniques fail.
  9. An advised initial step for the treatment of posterior epistaxis is to proceed directly to insertion of a unilateral or, preferably, bilateral posterior nasal pack or balloon.
     - An example of these is the elongated version of the Rhino Rocket as described previously.
     - Insertion and securing of the product proceed essentially the same way, except that it is inserted deeper into the posterior oropharynx.
  10. Consider the use of a dedicated posterior balloon, such as a Nasostat or Epistat (Fig. 53.7).
      - These devices are double-balloon systems that are lubricated preferably with bacitracin (sterile jelly is also appropriate) and inserted into the posterior nasopharynx.
      - The posterior balloon is then inflated with the product-specific quantity of air.
      - The product is then withdrawn slightly to ensure that the posterior balloon is situated firmly in its desired location of the posterior nasal cavity.
      - The anterior balloon is then inflated with the device-specific quantity of normal saline, and the device is secured.
- If the previously discussed resources are not available, a clinician can insert an adult-caliber Foley catheter into the posterior nasopharynx, with inflation of the balloon, and gentle traction anteriorly to ensure the bulb is situated similarly to the Nasostat or Epistat, firmly in the posterior nasal cavity to tamponade off the bleeding (Fig. 53.8).
- If all of these techniques still fail to control bleeding, an otolaryngologist will likely need to intervene, performing techniques outside the scope of practice of the emergency physician, including electrocautery, submucosal lidocaine/epinephrine injections, or other surgical interventions.
- The final step in the management of this patient includes admission to a monitored bed.

53.6.1 Pearls and Pitfalls

- Pearls
  - Consult otolaryngology early in the workup as soon as a posterior bleed is identified.
  - Patients receiving packing will still need to be placed on antibiotics, such as a first-generation cephalosporin, clindamycin, or amoxicillin–clavulanic acid.
  - Perform a good physical examination concomitantly with the management of the bleed specifically geared toward assessing the patient's volume status (orthostatics, capillary refill, heart rate), because the patient may have significant volume loss depending on the duration and quantity of the bleed.
  - Take a good history including prior bleed, prior admission requirements owing to epistaxis, any anticoagulant use, blood dyscrasias, and so on.

- Pitfalls
  - Misidentifying a posterior bleed as an anterior bleed. This can be catastrophic if the patient is discharged. Reference the "Etiology" section earlier.
  - Discharging the patient.
  - Consulting otolaryngology too late.
  - Underestimating the volume of blood loss. It may be helpful to think of these patients as trauma patients in their initial management and to proceed down the standard primary survey/intervention/secondary survey/intervention approach, ensuring that the volume needs are identified early.
  - Infection.
  - Pressure necrosis of the septum.
  - Hypoxia.
  - Aspiration.
  - Arrhythmias.
  - Dysphagia.
  - Dislodgment of the pack.

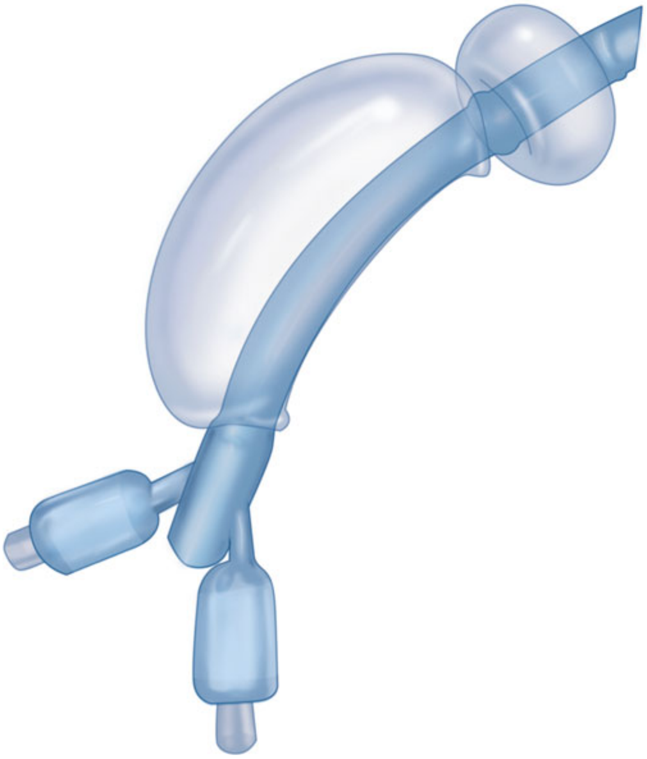

Fig. 53.7 Example of posterior balloon

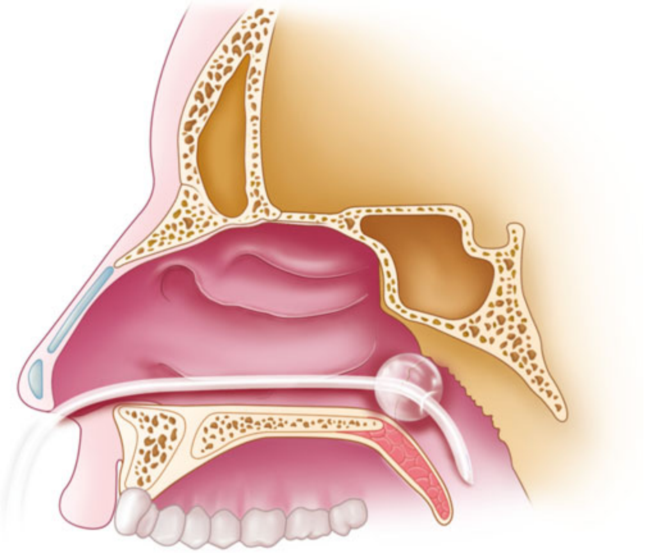

Fig. 53.8 Foley catheter in the posterior nasopharynx

Selected Reading

Buttaravoli P. Minor emergencies, splinters to fractures. 2nd ed. Philadelphia: Mosby; 2007.

Kucik CJ, Clenney T. Management of epistaxis. Am Fam Physician. 2005;71(2):305–11.

Marx JA, Hockberger RS, Walls RM, editors. Rosen's emergency medicine: concepts and clinical practice. 6th ed. Philadelphia: Mosby; 2006.

Pope LE, Hobbs CG. Epistaxis: an update on current management. Postgrad Med J. 2005;81(955):309–14.

Singer AJ, Blanda M, Cronin K, et al. Comparison of nasal tampons for the treatment of epistaxis in the emergency department: a randomized controlled trial. Ann Emerg Med. 2005;45:134–9.

Summers SM, Bey T. Chapter 239. Epistaxis, Nasal Fractures, and Rhinosinusitis. In: Tintinalli's Emergency Medicine, 7th ed. New York, NY: McGraw Hill; 2011.

Treatment of Septal Hematoma

54

Bobby K. Desai

54.1 Indications

- Septal hematomas occur after force applied to the nasal cartilage results in leakage of blood from the perichondrium (Figs. 54.1 and 54.2). They may be unilateral or bilateral.
- Untreated, this hematoma can expand and mechanically obstruct the blood vessels that supply the nasal cartilage.

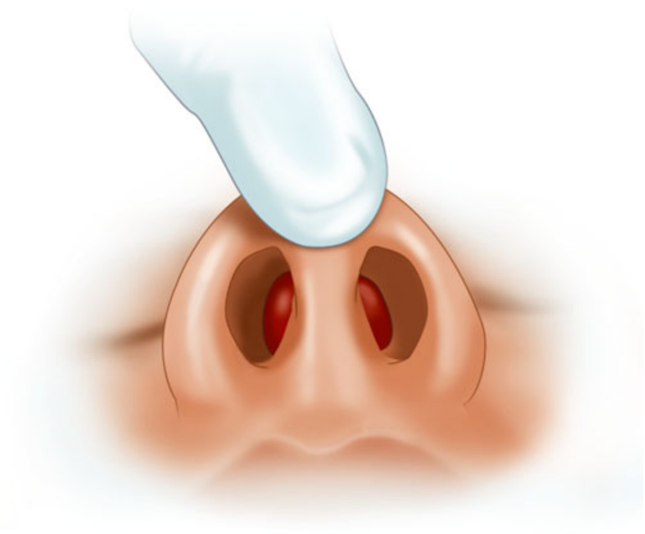

Fig. 54.1 Example of a septal hematoma

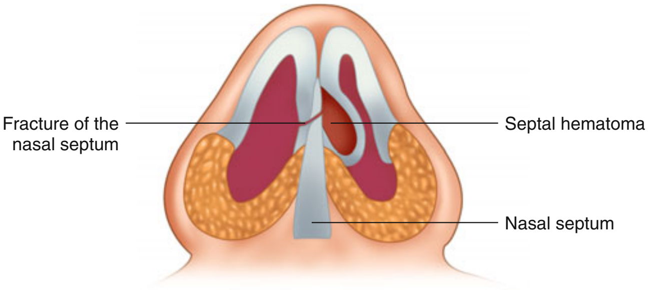

Fig. 54.2 Diagram of a septal hematoma

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_54

54.2 Contraindications

- In the setting of massive facial trauma, airway protection and consultation with otolaryngology or oral surgery may be required.

54.3 Materials and Medications

- 5-cc syringe
- 20-gauge needle (may use larger-gauge needle, but smaller ones may not allow for adequate suctioning).
- Topical antiseptic
- Sterile drapes, gown, eye protection, and gloves
- Light source—preferably a head lamp
- If using procedural sedation, specific medications of choice are needed as well as monitoring and resuscitation equipment
- Nasal speculum
- 4 % liquid cocaine or 4 % liquid lidocaine for anesthesia
- Topical vasoconstrictor of choice (oxymetazoline or phenylephrine)
- Scalpel (#15 blade)
- Suction setup
- Saline for irrigation
- Sterile gauze
- Iodoform packing or sterile rubber band
- Anterior nasal pack (Rhino Rocket)
- Intravenous antibiotics to cover *Staphylococcus aureus*, group A β-hemolytic streptococci, *Streptococcus pneumonia,* and *Haemophilus influenzae*

54.4 Procedure

- Preprocedure
  - Explain the procedure to the patient.
  - Informed consent may be required.
- Procedure—unilateral (Fig. 54.3):

1. Place the patient in a seated position if the procedure is to be done without procedural sedation.
    - If performing the procedure under procedural sedation, the patient may be placed in the supine position with appropriate monitoring and resuscitation equipment on standby.
2. Provide local anesthesia by applying the topical anesthetic of choice to cotton pledgets, wringing out the excess before insertion.
    - Remove after 10 min.
    - If using lidocaine, a topical vasoconstrictor may be required that may be mixed with the lidocaine or applied directly on the nasal mucosa before the procedure.
3. Using a nasal speculum, open the affected nares vertically as wide as possible. Horizontal use may obscure the hematoma.
4. Using a #15 blade, make a vertical incision at the hematoma extending the incision posteriorly at its base.
    - Take care not to make too deep an incision because septal perforation may occur.
5. Suction and gauze should be at the ready to remove as much blood as possible.
6. Pack the incision site with iodoform gauze or a sterile rubber band in order to keep the incision open to prevent accumulation of blood.
7. This packing can be held in place with an anterior nasal pack.
    - The pack should be placed in both nares to prevent deformity of the nasal septum.
8. Consider administration of a dose of intravenous antibiotics.
9. Send the patient home with a prescription for oral antibiotics.
10. Arrange follow-up for 24 h.

- Procedure—bilateral (Fig. 54.4):

1. If bilateral hematomas are present, otolaryngology consultation should be sought.
2. One method involves the incision of one side only through the nasal septum so that both may drain through one side.
    - Care must be taken NOT to penetrate the overlying mucosa of the other side.
3. The second option is to repeat the unilateral drainage procedure on the contralateral side.
4. Care must be taken to not have the incision exactly in the same location as the contralateral side owing to the risk of permanent septal perforation.

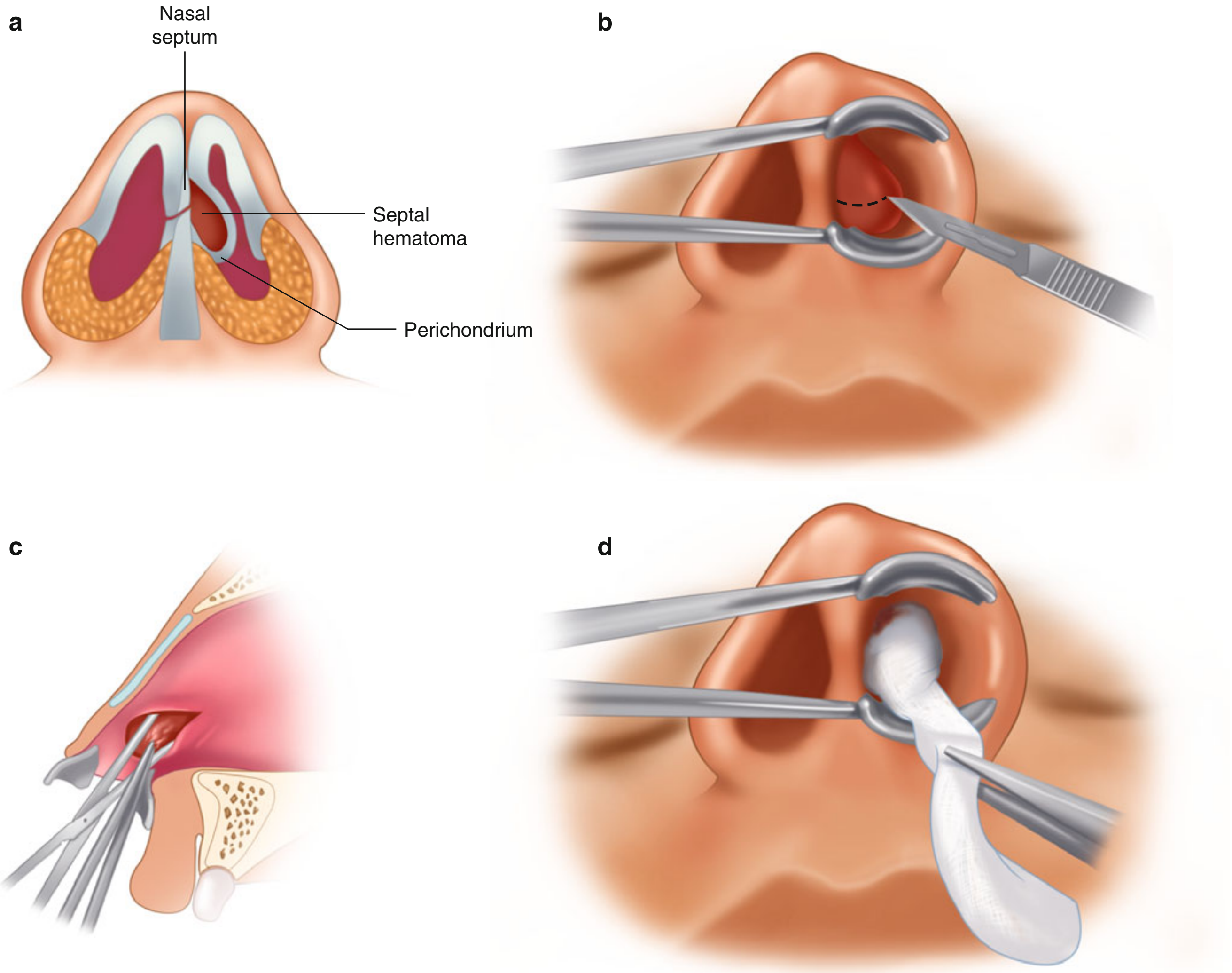
a
Nasal septum
Septal hematoma
Perichondrium
b
c
d

Fig. 54.3 (**a**–**d**) Technique for drainage of a unilateral septal hematoma

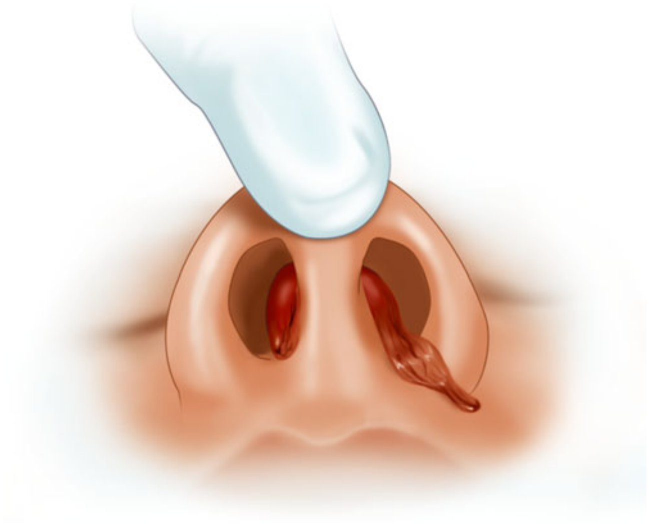

Fig. 54.4 Technique for drainage of a bilateral septal hematoma

54.5 Pitfalls and Complications

- Bleeding
- Infection: toxic shock syndrome
- Deformity of the nose
- Permanent septal perforation
- Inadequate drainage
- Reaccumulation of blood
- Development of septal abscess

Selected Reading

Ginsburg CG. Consultation with the specialist: nasal septal hematoma. Pediatr Rev. 1998;19:142–3.

Savage RR, Valvich C. Hematoma of the nasal septum. Pediatr Rev. 2006;27:478–9.

Nasal Foreign Body Removal

55

Bobby K. Desai

55.1 Indications

- Visualized foreign body (FB) on inspection
- Continued unilateral nasal discharge of unknown etiology
- Recurrent epistaxis

55.2 Contraindications: Need Urgent Referral

- Those cases in which the emergency physician is not confident of success.
- Combative patients who are not candidates for conscious sedation.
- Embedded FBs.
- Penetrating injuries with an FB.
- Button batteries that cannot be removed.
- Patients with bleeding diathesis.
- Respiratory distress.
- FBs superior and medial to the middle turbinate owing to risk of puncture to the cribriform plate.
- Chronic FBs may be difficult to visualize and these should be referred.

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_55

55.3 Materials and Medications

- 1 % lidocaine without epinephrine (solution) for anesthesia
- 2 % lidocaine jelly
- 0.5 % phenylephrine to vasoconstrict vessels
- Nasal speculum (Fig. 55.1)
- Headlamp or other direct lighting instrument
- Ambu bag (Fig. 55.2)
- Alligator forceps (Fig. 55.3)
- Curved hook (Fig. 55.4)
- Foley catheter
- Suction: schuknecht catheter (Fig. 55.5)
- Eye and biohazard protection for the clinician

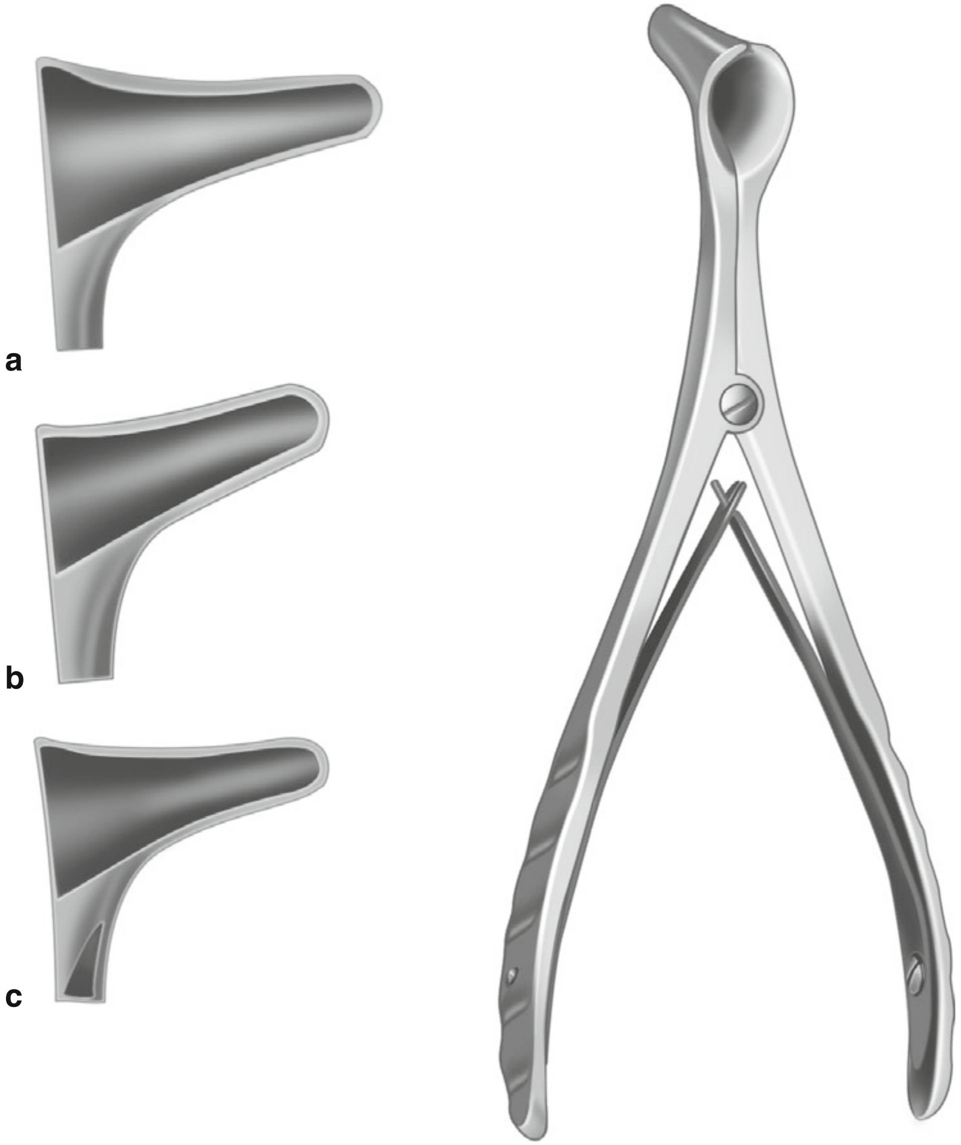

Fig. 55.1 (**a**–**c**) Example of nasal speculum

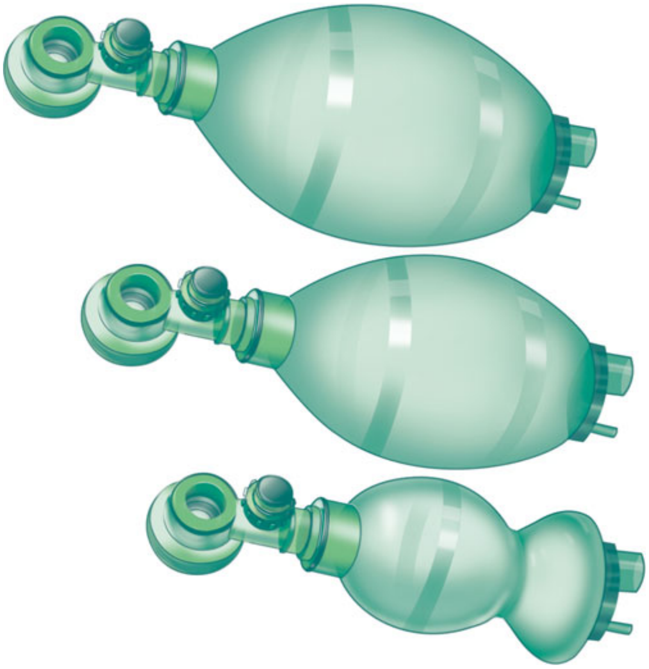

Fig. 55.2 Example of Ambu bags

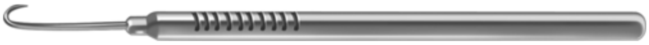

Fig. 55.4 Example of a curved hook

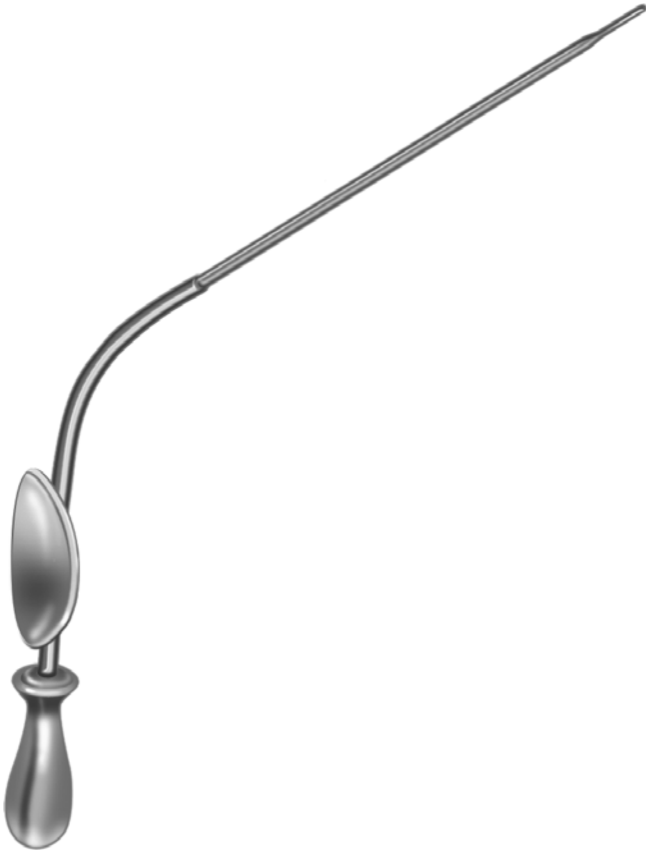

Fig. 55.5 Example of a schuknecht suction tube

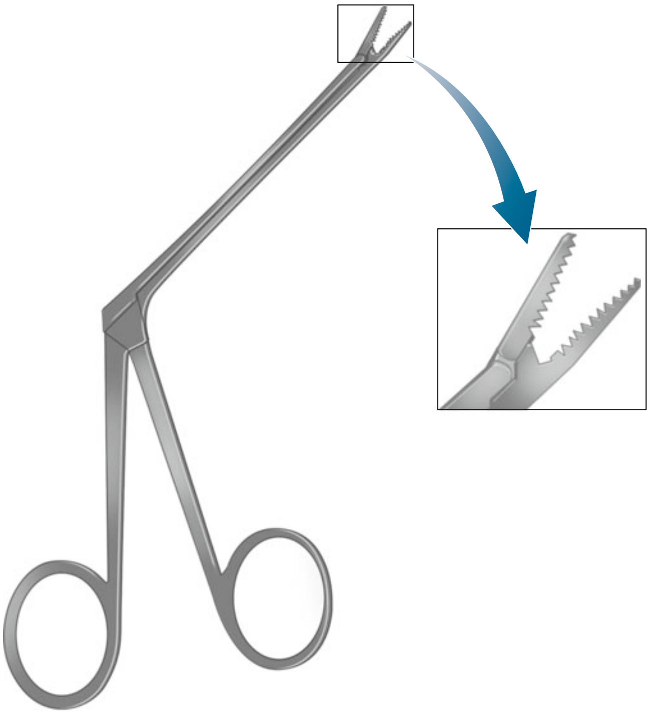

Fig. 55.3 Example of alligator forceps

55.4 Procedures (Based on the Method Chosen)

55.4.1 Alligator Forceps

- Perhaps the easiest method to use for easily visualized FBs.
- Care must be taken for those FBs that are easily broken apart (e.g., peas) to ensure that all of the matter has been removed.
- Using a nasal speculum, spread the nares as much as possible to maximize visualization.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Attempt to grasp the FB with the forceps.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

55.4.1.1 Pitfalls and Complications

- Pushing the object deeper into the nares.
- Inability to totally remove the FB.
- Attempted removal of an embedded FB may cause significant bleeding.

55.4.2 Curved Hook

- Used for nongraspable objects especially in the anterior nares.
- Using a nasal speculum, spread the nares as much as possible to maximize visualization.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pass the hook behind the object, or if the object has holes, the hook may be placed inside the hole to facilitate removal.
- Slowly withdraw the hook, thereby removing the FB.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

55.4.2.1 Pitfalls and Complications

- This technique cannot be used if the object cannot be directly visualized because significant trauma may arise owing to the misplacement of the hook.
- Attempted removal of an embedded FB may cause significant bleeding.

55.4.3 Foley Catheter Removal

- Use a 5 or 6 French catheter.
- Anesthetize the nares with lidocaine with or without epinephrine and vasoconstrict the nasal vessels with phenylephrine.
- Ensure that the Foley balloon has no leaks.
- Place the patient in a supine position.
  - Consider procedural sedation for children and/or a papoose board.
- Lubricate the Foley tip and balloon with lidocaine jelly.
- Advance the tip past the object.
- Blow up the balloon with 2–3 mL of air and gently withdraw the catheter.
  - The amount of air may need to be adjusted depending on the size of the object to be withdrawn.
  - Pulling too quickly may cause a soft FB to break apart.

55.4.3.1 Pitfalls and Complications

- This technique may be used if the FB is not visible.
- In order for this technique to work, the catheter must be able to slide posterior to the object in order for the balloon to pull out the object.

55.4.4 Suction

- A small metal suction catheter—the schuknecht catheter—may be used.
- The suction is placed against the FB and slowly withdrawn.
- This technique works best for round smooth objects visualized in the anterior nares.

55.4.4.1 Pitfalls and Complications

- May not work for posteriorly displaced FBs.
- May not work for objects too tightly entrenched in the nares.

55.4.5 Nasal Positive Pressure

- This technique works best for round FBs occluding the nares.
- Place the patient in a supine position.
  - Consider a papoose board to secure the patient.
- Apply pressure to the contralateral nares to occlude it.
- Connect an Ambu bag to high-flow oxygen.
  - Use a facemask that covers only the mouth.
- Place the mask on the patient's mouth and allow it to expand by covering the thumbhole.
- If this pressure is insufficient to expel the object, the Ambu bag may be compressed to add pressure to the attempt.
- This may allow the object to be seen and grasped in the anterior nares.

55.4.5.1 Pitfalls and Complications

- Too much pressure could theoretically rupture the tympanic membranes.

Selected Reading

Backlin SA. Positive-pressure technique for nasal foreign body removal in children. Ann Emerg Med. 1995;25:554–5.

Chan TC, Ufberg J, Harrigan RA, et al. Nasal foreign body removal. J Emerg Med. 2004;26:441–5.

Kadish H. Ear and nose foreign bodies: it is all about the tools. Clin Pediatr. 2005;44:665–70.

Cerumen Removal

56

Bobby K. Desai

56.1 Indications

- Hearing loss
- Profound dizziness
- Pain

56.2 Contraindications

- Uncooperative patient
- Potential for foreign bodies
- Prior surgery to ear or mastoid
- History of middle or outer ear disease

56.3 Materials and Medications

- Ear speculum (Fig. 56.1).
- Eye and biohazard protection for the clinician
- Headlamp or other direct lighting instrument
- Cerumen-softening solutions
  - Carbamide peroxide
  - Mineral oil or other oil-based agents including almond or olive oil
  - Liquid docusate sodium
  - Acetic acid
  - Sodium bicarbonate
- Alligator forceps (Fig. 56.2)
- Curved hook or cerumen spoon (Figs. 56.3 and 56.4)
- Suction—Schuknecht catheter
- Irrigation setup
  - Ear syringe (Figs. 56.5 and 56.6)
- 20- to 60-mL syringe with an 18-gauge angiocath attached
  - Commercial ear irrigation setup
  - Water (tap water may be used)
  - Basin or other collection device
  - Towels

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_56

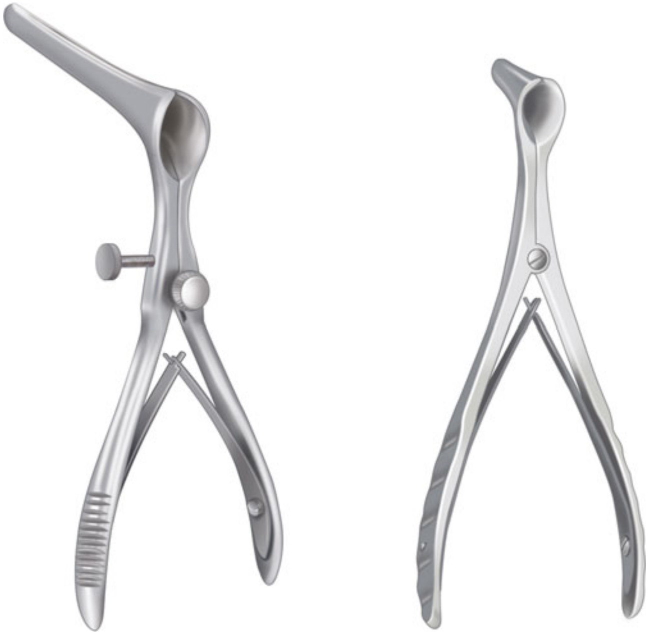

Fig. 56.1 Example of an ear speculum

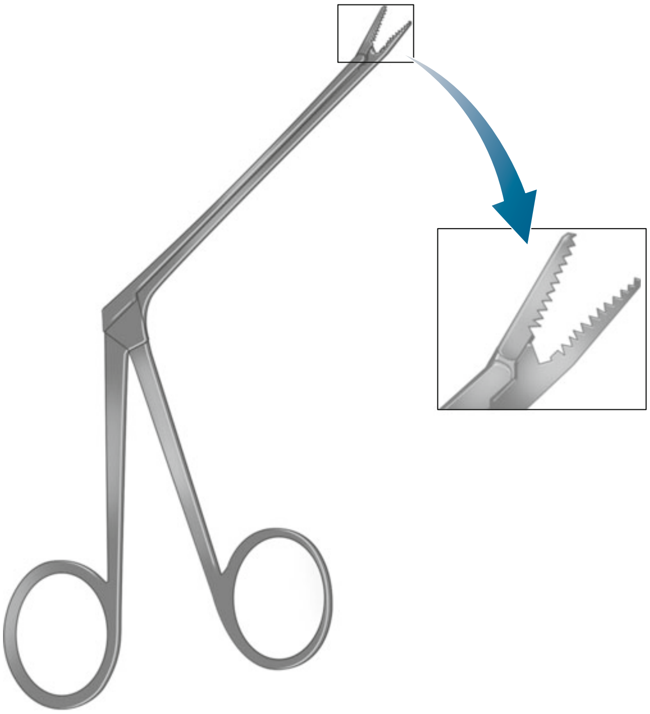

Fig. 56.2 Example of alligator forceps

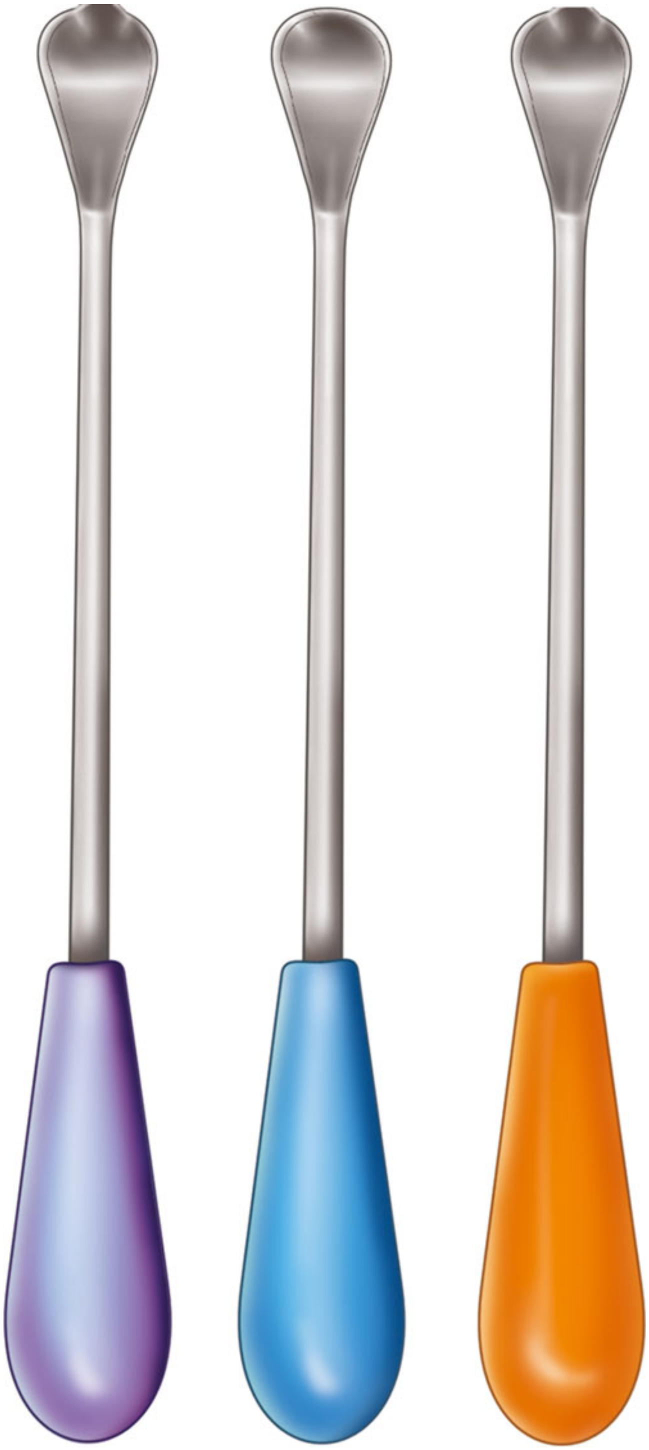

Fig. 56.3 Examples of cerumen spoons

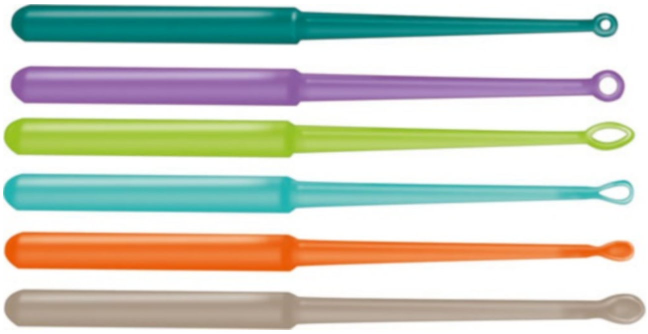

Fig. 56.4 Examples of ear curettes

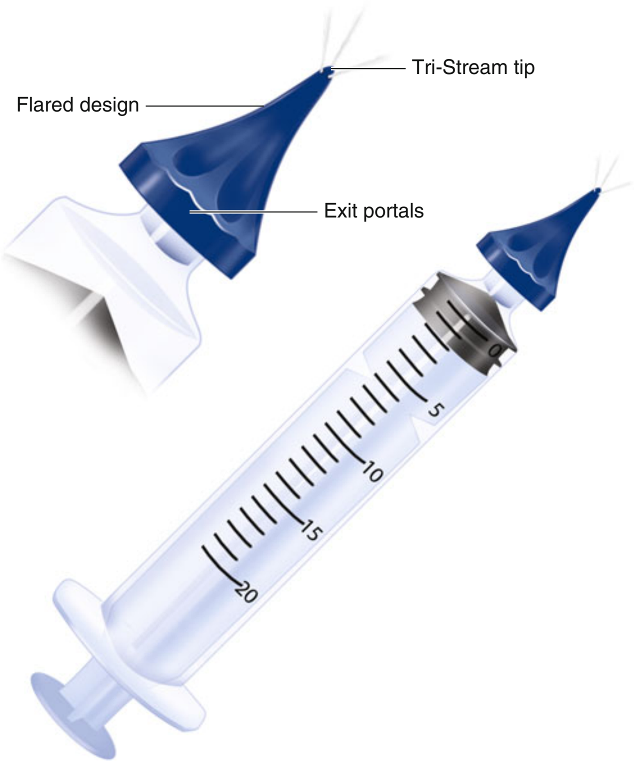
Tri-Stream tip
Flared design
Exit portals
0
5
10
15
20

Fig. 56.6 Example of a specialized ear syringe

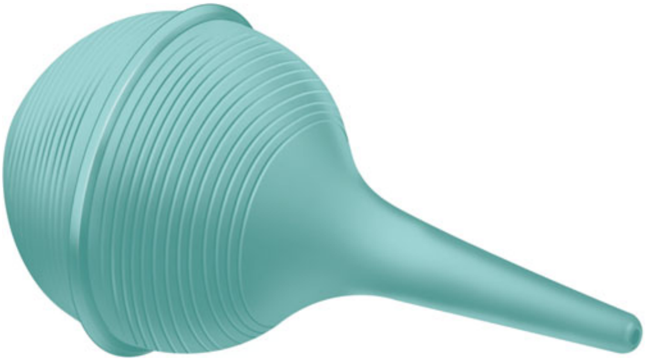

Fig. 56.5 Example of a typical bulb syringe used in irrigation

56.4 Procedure

1. Explain the procedure to the patient to facilitate cooperation.
   - Some discomfort may be involved, including a sensation of pressure and the cold, wet feeling from the water used.
   - Nausea, vomiting, or a sensation of vertigo may also be experienced.
     - Consider pretreatment of vertigo with agent of choice.
2. The patient may be seated or supine with the head slightly turned to the affected side to facilitate the collection of fluid.
3. Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
4. Pull the ear up and out to straighten the ear canal.
5. Before irrigation, if the cerumen is hard, alligator forceps or a cerumen spoon may be used under direct visualization to remove as much cerumen as possible.
6. Use a cerumen-softening agent before irrigation by instilling and leaving in for 30 min to maximize softening.
7. Place a collection device next to the patient's ear, or alternatively, have the patient hold a basin near the ear to collect any liquid.
8. Attach an angiocath to a 20- or 60-mL syringe, place the tip of the catheter within the ear canal, and aim toward the tympanic membrane.
9. Slowly inject the solution into the ear canal.
   - Care must be taken to limit the force used owing to the potential for tympanic membrane rupture.
   - The fluid can be injected in short bursts or all at once.
10. Direct the stream as much as possible toward the rear of the cerumen to be removed.
11. Continually reassess the ear canal to ensure that all cerumen has been removed.
    - If one large piece remains, use a cerumen hook to facilitate removal.
12. If the patient complains of severe sharp pain, stop the procedure and assess the tympanic membrane for rupture.
    - Consider referral or otolaryngological consultation at this point.
13. After successful removal, dry the ear canal with cotton-tipped applicators to prevent otitis externa.

56.5 Pitfalls and Complications

- Owing to instillation of room temperature water, a caloric response may occur including vertigo, nausea, and vomiting.
- Rupture of the tympanic membrane.
- Unsuccessful irrigation secondary to inadequate time given to softening agents before attempting removal.
- Bleeding of the external auditory canal.
- Otitis externa.
- Hearing loss.

Selected Reading

Blake P, Matthews R, Hornibrook J. When not to syringe an ear. N Z Med J. 1998;111:422.

McCarter D, Courtney AU, Pollart SM. Cerumen impaction. Am Fam Physician. 2007;75:1523–8.

Roland PS, Smith TL, Schwartz SR, et al. Clinical practice guideline: cerumen impaction. Otolaryngol Head Neck Surg. 2008;139:S1–21.

Ear Foreign Body Removal

57

Bobby K. Desai

57.1 Indications

- Visualized foreign body (FB) on inspection.
- Continued unilateral otorrhea of unknown etiology mandates a search.
- Recurrent bleeding.
- History of FB placement by patient.

57.2 Contraindications: Need Urgent Referral

- Those cases in which the emergency physician is not confident of success.
- Combative patients are not candidates for conscious sedation.
- Embedded FBs.
- Penetrating injuries with an FB.
- FBs close to the tympanic membrane.
- Consider referral of spherical FBs because these may be difficult to grasp.
- Chronic FBs may be difficult to visualize; these should be referred.
- Consider referral for those patients who have attempts made at other institutions.

57.3 Materials and Medications

- Ear speculum (Fig. 56.1)
- Eye and biohazard protection for the clinician
- Headlamp or other direct lighting instrument
- Alligator forceps (Fig. 56.2)
- Curved hook (Fig. 57.1) or cerumen spoon (Fig. 56.3)
- Suction—Schuknecht catheter (Fig. 57.2)
- Irrigation setup
  - Ear syringe (Fig. 57.3)
  - 20- to 60-mL syringe with an 18-gauge angiocath attached
  - Water (tap water may be used)
  - Basin or other collection device
  - Towels
- Cyanoacrylate (Fig. 57.4)

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_57

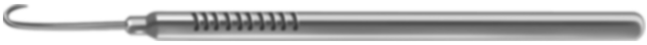

Fig. 57.1 Example of a curved hook

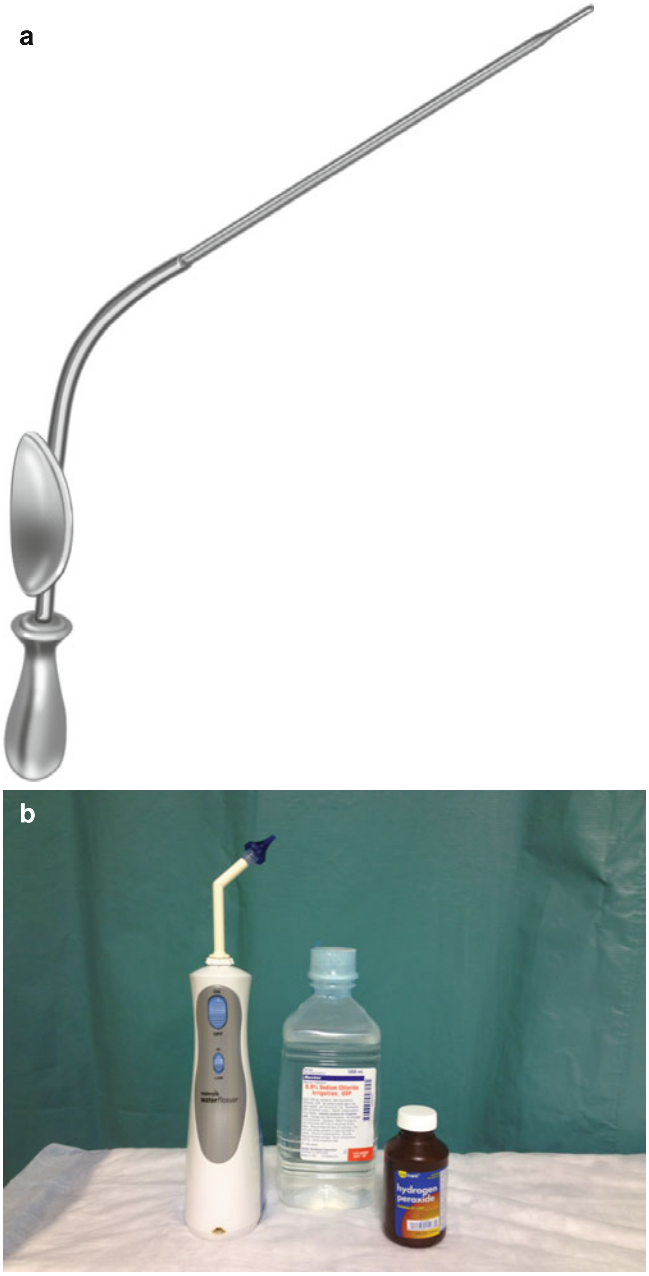
a
b
hydrogen peroxide

Fig. 57.2 (**a**) Example of a Schuknecht suction tube, (**b**) example of ear irrigation setup using a Waterpik system

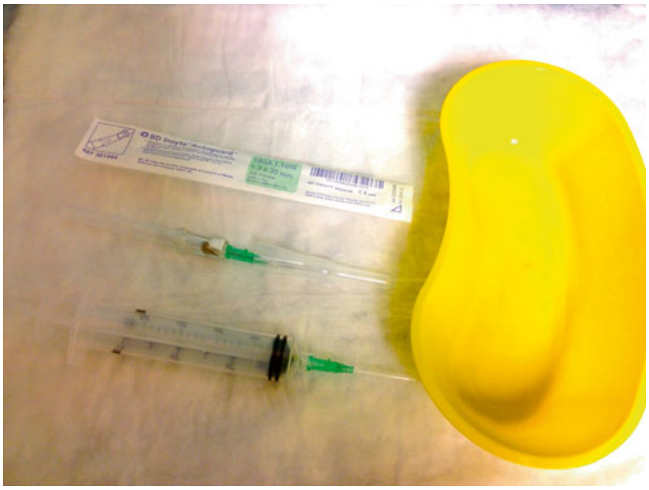

Fig. 57.3 Example of a typical syringe used in irrigation

ETHICON
DERMABOND
ADVANCED

Fig. 57.4 Example of cyanoacrylate

57.4 Procedures (Based on the Method Chosen)

57.4.1 Alligator Forceps

- Ideal for easily visualized FBs, especially graspable FBs in the lateral third of the ear canal.
  - May have limited success with rounded objects.
- Care must be taken for those FBs that are easily broken apart (e.g., peas) to ensure that all of the matter has been removed.
  - May need irrigation subsequent to the attempt.
- Explain the procedure to the patient to facilitate cooperation.
- Pull the ear up and out to straighten the ear canal.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Attempt to grasp the FB with the forceps.
  - If using an ear speculum, attempt to bring the object as close as possible to the speculum in order to stabilize it and slowly remove both.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

57.4.1.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal.
- Inability to totally remove the FB.
- Attempted removal of an embedded FB may cause significant bleeding.

57.4.2 Curved Hook

- Ideal for spherical objects in the lateral third of the auditory canal.
- Can be used for non-graspable objects especially in the lateral third of the auditory canal.
- Direct visualization is necessary to avoid trauma to the ear canal.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Pass the hook behind the object, or if the object has holes, place the hook inside the hole to facilitate removal.
  - In order for this technique to succeed, there must be sufficient area near the object to facilitate the hook going behind it.
- Slowly withdraw the hook, thereby removing the FB.
- After removal, ensure the complete removal of the object and ensure that no other FBs are present.

57.4.2.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal.
- This technique cannot be used if the object cannot be directly visualized because significant trauma may arise owing to the misplacement of the hook.
- Attempted removal of an embedded FB may cause significant bleeding.
- Do not use this technique if the object is close to the tympanic membrane owing to risk of rupture.

57.4.3 Suction

- A small metal suction catheter—the Schuknecht catheter—may be used.
- Ideal for hard rounded objects in the lateral third of the ear canal.
  - Does not work as well for those objects closer to the tympanic membrane owing to lack of suction and the narrow ear canal
- Direct visualization of the object is mandatory.
  - The catheter must be placed directly on the object for this technique to succeed.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place the suction against the FB and slowly withdraw.

57.4.3.1 Pitfalls and Complications

- Pushing the object deeper into the auditory canal
- May not work for objects too tightly entrenched in the ear canal

57.4.4 Irrigation

- Useful for small particulate matter unable to be grasped by forceps or too small for a hook to pull out.
  - Cannot be used for button batteries or vegetable matter.
    - Irrigation may cause an alkaline necrosis with the button battery.
    - Vegetable matter may expand, increasing the difficulty of removal.
- Explain the procedure to the patient to facilitate cooperation.
  - Some discomfort may be involved, including a sensation of pressure and a feeling cold wetness from the water used.
- The patient may be seated or supine with the head slightly turned to the affected side to facilitate collection of fluid.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place a collection device next to the patient's ear, or alternatively, have the patient hold a basin.
- Attach an angiocath to a 20- or 60-mL syringe and place the tip of the catheter within the ear canal, or use a Waterpik system.
- Slowly inject the solution into the ear canal.
  - Care must be taken to limit the force used owing to the potential for tympanic membrane rupture.
- Direct the stream as much as possible toward the rear of the object(s) to be removed.
- Continually assess the ear canal to ensure that all foreign matter has been removed.
- If the patient complains of severe sharp pain, stop the procedure and assess the tympanic membrane for rupture.
  - Consider referral or otolaryngological consultation at this point.

57.4.4.1 Pitfalls and Complications

- Owing to instillation of room temperature water, a caloric response may occur including vertigo, nausea, and vomiting.
- Rupture of the tympanic membrane.

57.4.5 Cyanoacrylate

- Useful for objects that can be directly visualized.
- Ideal for hard rounded objects in the lateral third of the ear canal.
  - Does not work as well for those objects closer to the tympanic membrane owing to lack of visualization
- Direct visualization of the object is mandatory.
  - The stick with the cyanoacrylate must be placed directly on the object for this technique to succeed.
- Consider using a specialized otoscope with magnification (an operating otoscope) to maximize visualization.
- Explain the procedure to the patient to facilitate cooperation.
- The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
- Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
- Pull the ear up and out to straighten the ear canal.
- Place a small amount of cyanoacrylate on the end of a cerumen spoon or wooden end of a cotton-tipped swab.
- Using direct visualization, place the stick in the ear canal and touch the FB.
- Leave the stick on the object for approximately 30–60 s to facilitate drying of the glue.
- Remove the stick and the FB.

57.4.5.1 Pitfalls and Complications

- This technique cannot be used if the object cannot be seen.
- The object must be relatively smooth to facilitate placement of the glue.

Selected Reading

Hanson RM, Stephens M. Cyanoacrylate-assisted foreign body removal from the ear and nose in children. J Paediatr Child Health. 1994;30:77–8.

Kadish H. Ear and nose foreign bodies: it is all about the tools. Clin Pediatr. 2005;44:665–70.

Treatment of Auricular Hematoma

58

Bobby K. Desai

58.1 Indications

- Presence of auricular hematoma
  - These occur after significant force is applied to the ear.
  - This type of hematoma will separate the perichondrium from the cartilage.
    - Untreated, this hematoma prevents the development of new cartilage, subsequently deforming the auricle, causing *cauliflower ear* (Figs. 58.1 and 58.2).

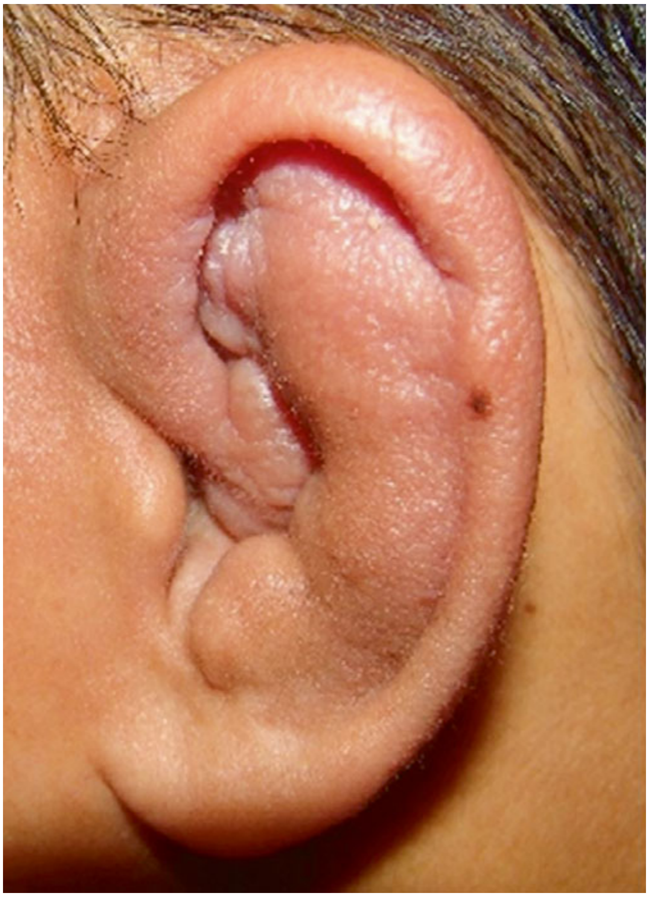

Fig. 58.1 Example of a cauliflower ear

B.K. Desai, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_58

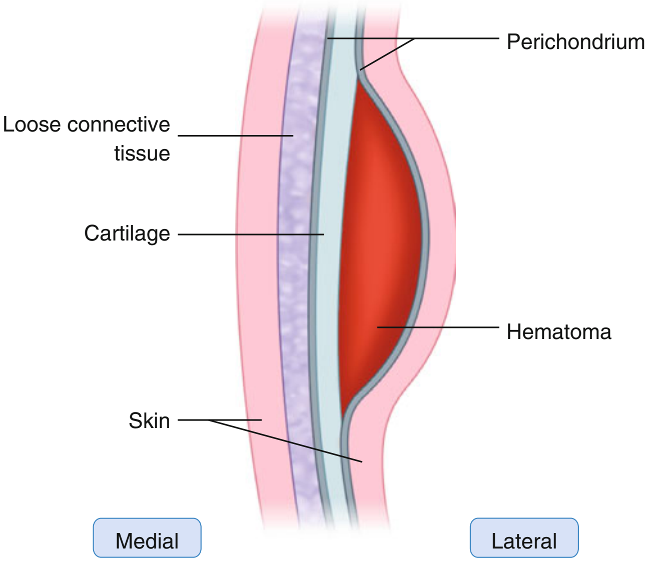

Fig. 58.2 Schematic of auricular hematoma

58.2 Contraindications

- Uncooperative patient.
- Evidence of infection mandates plastic surgery or otolaryngology consultation.
- Significant trauma to area around the ear.

58.3 Materials and Medications

- 5-cc syringe
- 20-gauge needle (may use larger-gauge needle, but smaller ones may not allow for adequate suctioning)
- Topical antiseptic
- Sterile drapes, gown, eye protection, and gloves
- 1 % lidocaine for anesthesia
- Scalpel (#11 or #15 blade)
- Suction setup
- Saline for irrigation
- Hemostat and forceps
- Penrose drain or plain gauze packing
- Xeroform gauze
- Laceration tray setup
  - Needle driver
  - 4–0 Nonabsorbable suture
- Antibiotic ointment
- Gauze bandages

58.4 Procedure (Needle Aspiration)

1. Explain the procedure to the patient.
2. Informed consent may be required.
3. The patient may be seated or supine with the unaffected side turned to the gurney or examining chair.
4. Stabilize the head with the nondominant hand, or for children, consider a papoose board to secure the patient.
   - Consider procedural sedation for children.
5. Carefully clean the area with an antiseptic solution of choice (povidone-iodine [Betadine] or chlorhexidine solution) and allow to dry.
6. Anesthetize the area where the hematoma is of greatest diameter with 1 % lidocaine.
7. Inject the hematoma itself to promote increased anesthesia.
8. The clinician should prepare for the procedure by practicing universal precautions.
9. Drape the patient's ear.
10. Attach a 20-gauge needle to a syringe.
    - A 5-cc syringe should be sufficient.
11. Attempt to aspirate the hematoma contents with the prepared syringe (Fig. 58.3).

12. "Milking" the hematoma may be required to fully remove the clot.
    - If the hematoma has been present for sufficient time for the blood to completely clot, needle aspiration will fail, and an alternative technique, described later, may be necessary to completely evacuate the hematoma.
13. Place antibiotic ointment around the injection site.
14. Loosely dress the area.
15. Refer to an otolaryngologist or a plastic surgeon for follow-up care.

If needle aspiration fails, a larger incision may be required using the following procedure:

1. Perform the previously discussed steps.
2. Consider a local field block of the greater auricular nerve.
3. Make an incision over the hematoma, following the anatomy of the helix (Fig. 58.4).
    - The incision should be 5–6 mm in length and curvilinear, following the concavity of the helix.
    - More than one incision may be required.
4. Remove the hematoma with hemostats or forceps.
    - Suctioning may be used as well.
5. Copiously irrigate the cavity with saline to make certain of complete removal of the hematoma.
6. Insert the Penrose drain or plain packing into the incision.
7. Antibiotic ointment may be applied to the incision site.
8. Place the Xeroform gauze over the incision site.
9. Apply a compression dressing to the ear (Fig. 58.5).
    - Place dry cotton into the ear canal to protect from possible drainage and subsequent infection (Fig. 58.6).
    - Mold Xeroform gauze in and around the contours of the helix (Fig. 58.7).
    - With the aid of an assistant, place gauze behind the pinna and over the Xeroform gauze already molded to the anterior portion of the ear (Fig. 58.8).
    - Place more fluffed gauze over the ear (Fig. 58.9).
    - While the assistant holds the anterior and posterior gauze in place, use a web roll (Kerlix) around the head to keep the gauze in place. Then wrap the head with an ACE wrap or stretchy bandage (Figs. 58.10 and 58.11).
    - These steps are critical to maintain equal compression of the ear, which will help prevent future deformity.
    - Alternatively, dental rolls can be secured with suture to the anterior and posterior site of the drained hematoma (Fig. 58.12).
10. Place the patient on prophylactic antibiotics.
11. Refer to an otolaryngologist or a plastic surgeon for follow-up care.

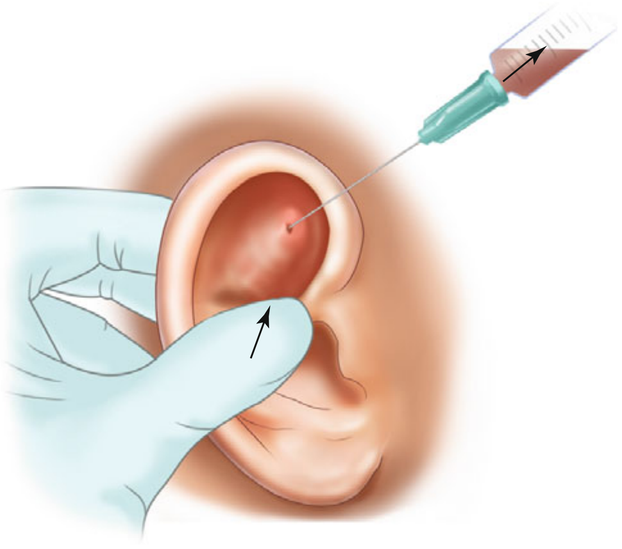

Fig. 58.3 Aspiration of a hematoma

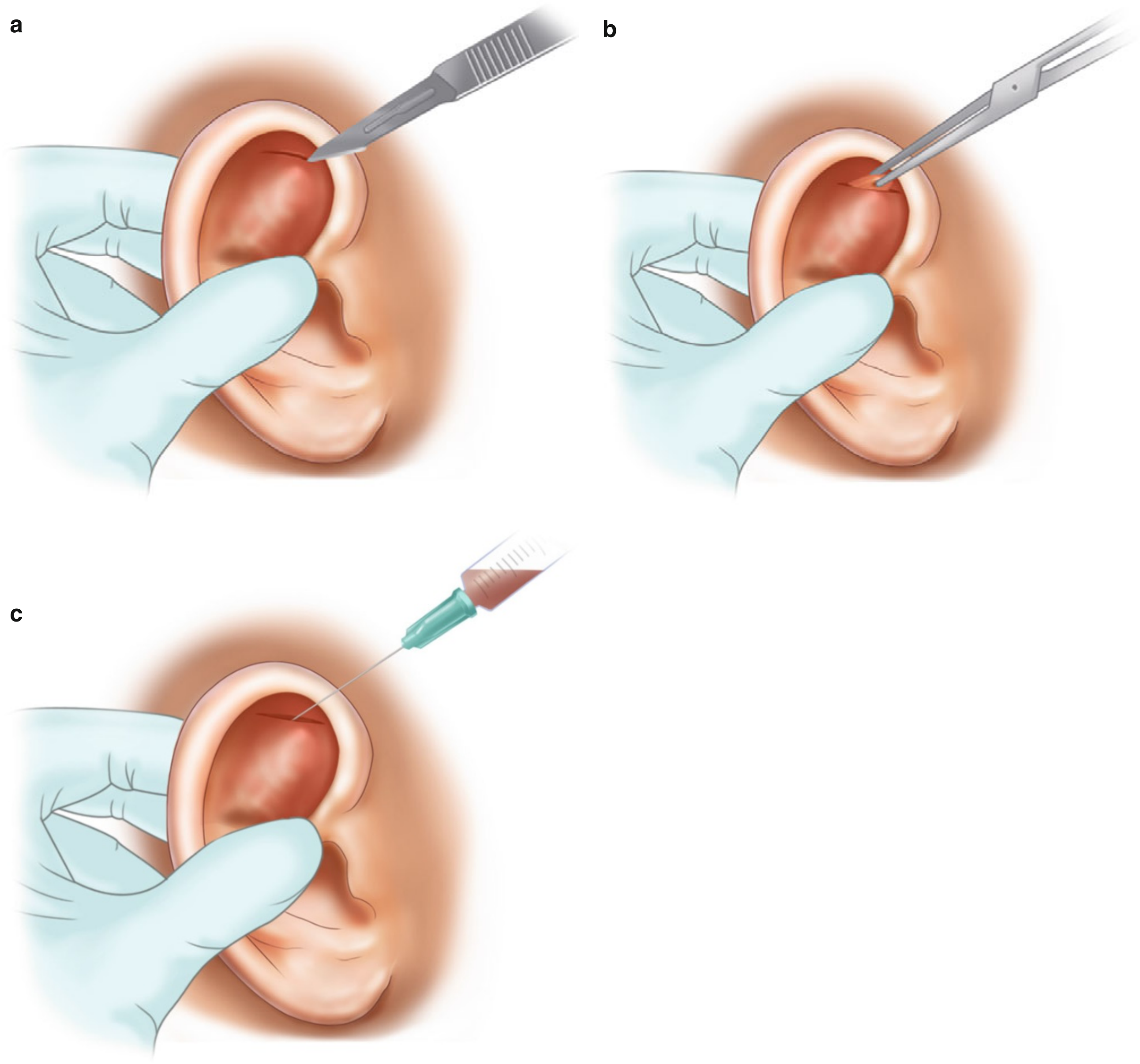

Fig. 58.4 Incision of an auricular hematoma. (**a**) Incision; (**b**) removal of clot with forceps; (**c**) aspiration of blood with syringe

Fig. 58.5 Materials needed for compression dressing

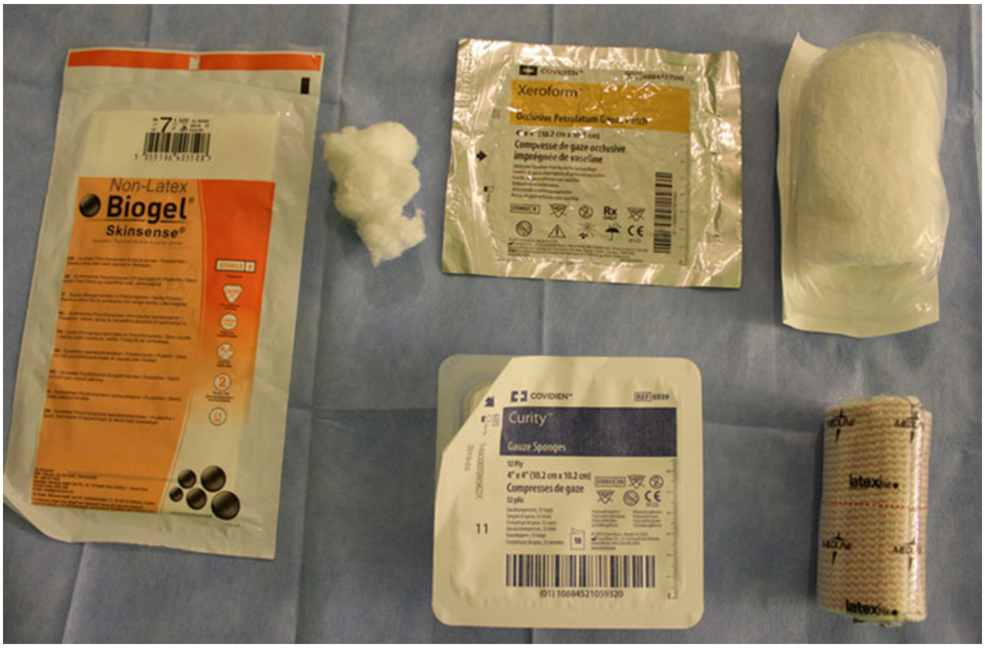
Non-Latex
Biogel
Skinsense
Xeroform
Curity
Gauze Sponges

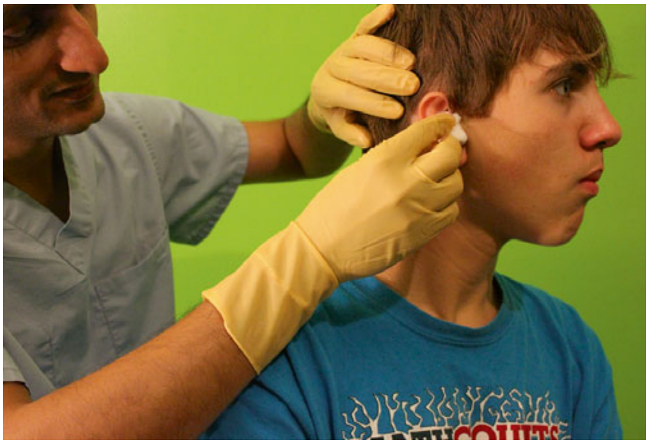

Fig. 58.6 Place cotton in the ear to prevent drainage into ear

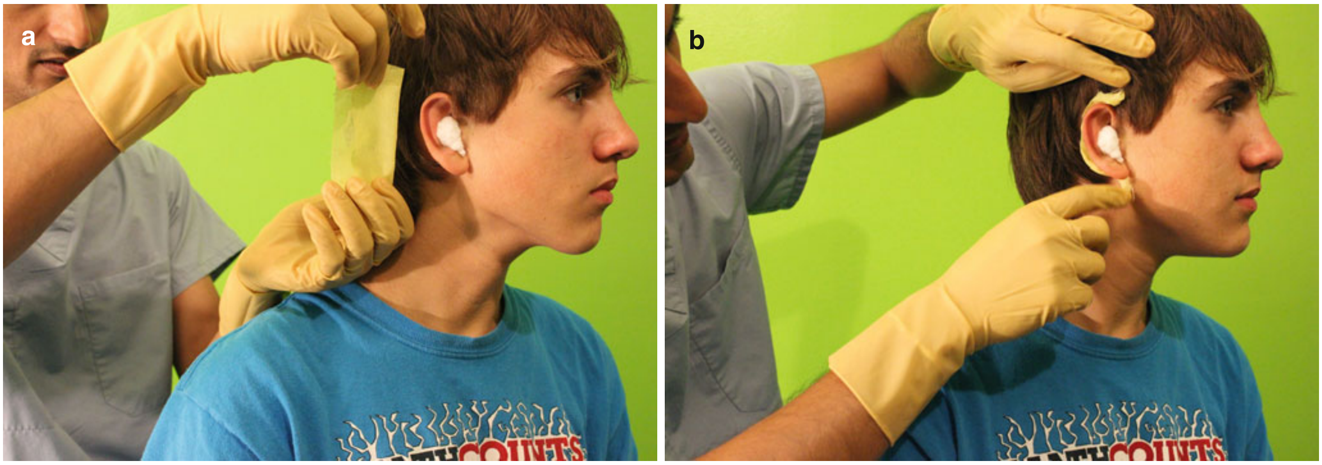

Fig. 58.7 (**a**, **b**) Mold Xeroform gauze around ear helix

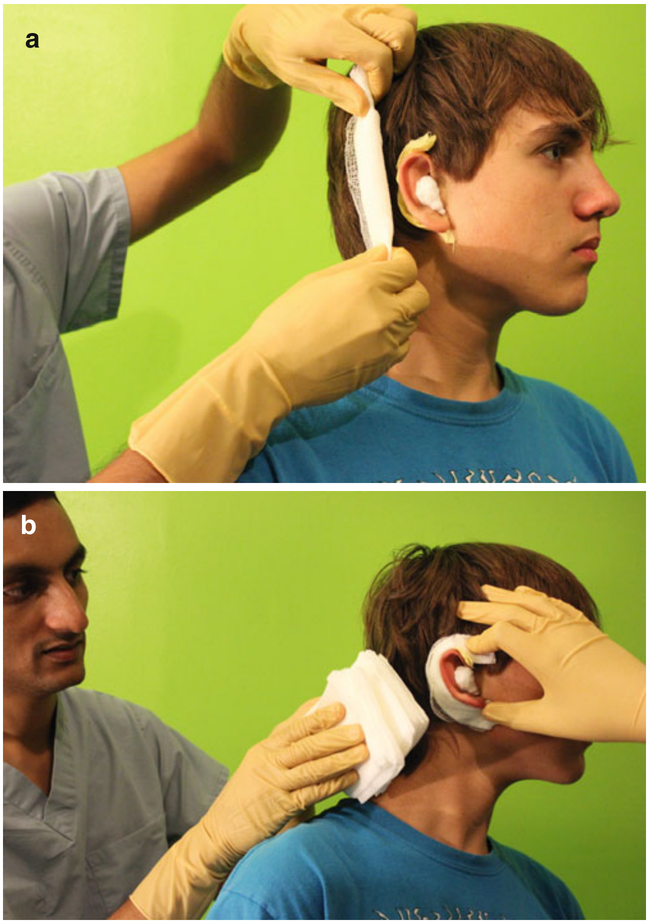

Fig. 58.8 (**a**, **b**) Mold sterile gauze on top of Xeroform gauze around ear helix

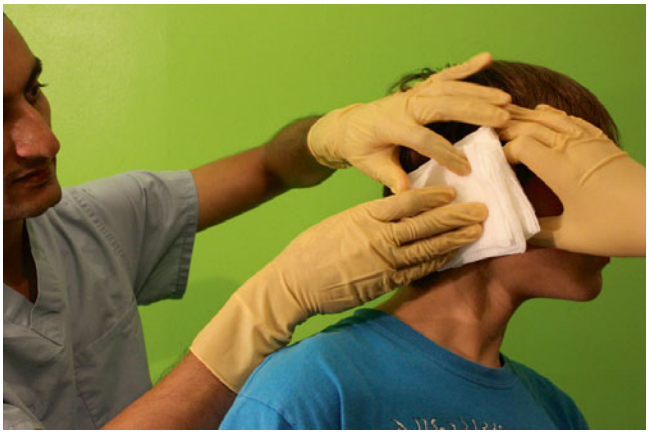

Fig. 58.9 Pad outside of ear with 4×4 gauze pads

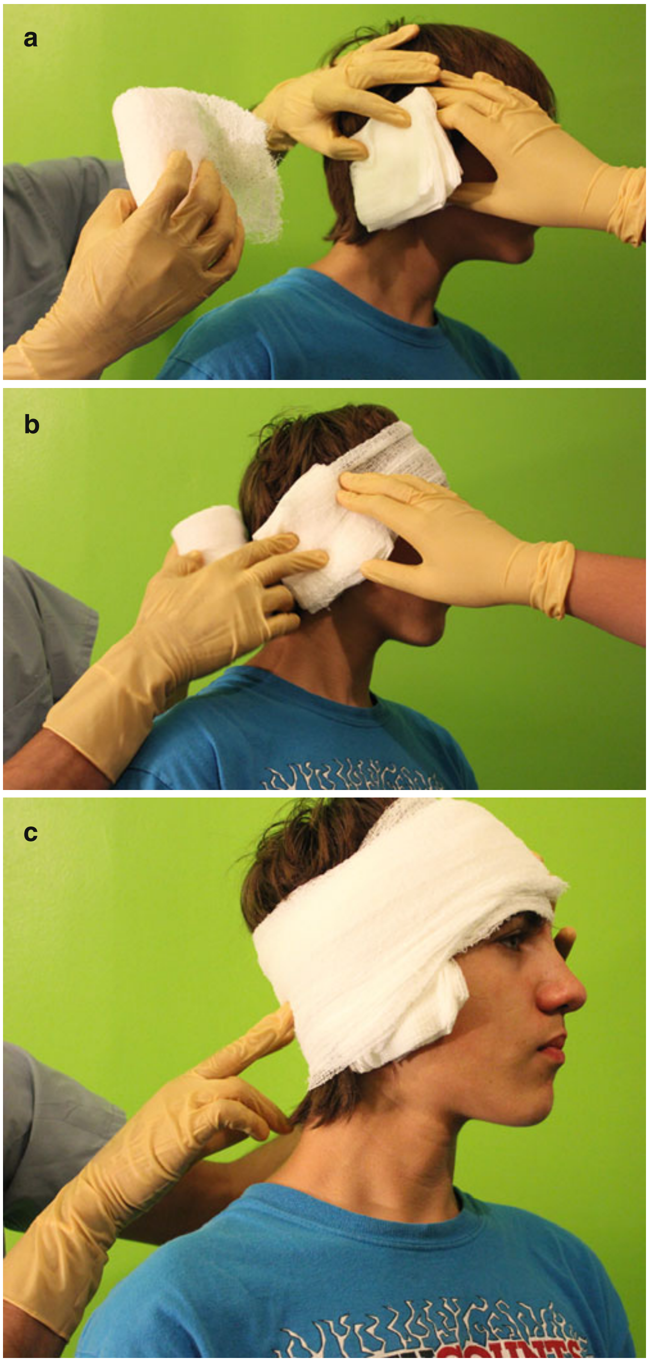

Fig. 58.10 (**a**–**c**) Wrap head with Kerlix (sterile gauze roll)

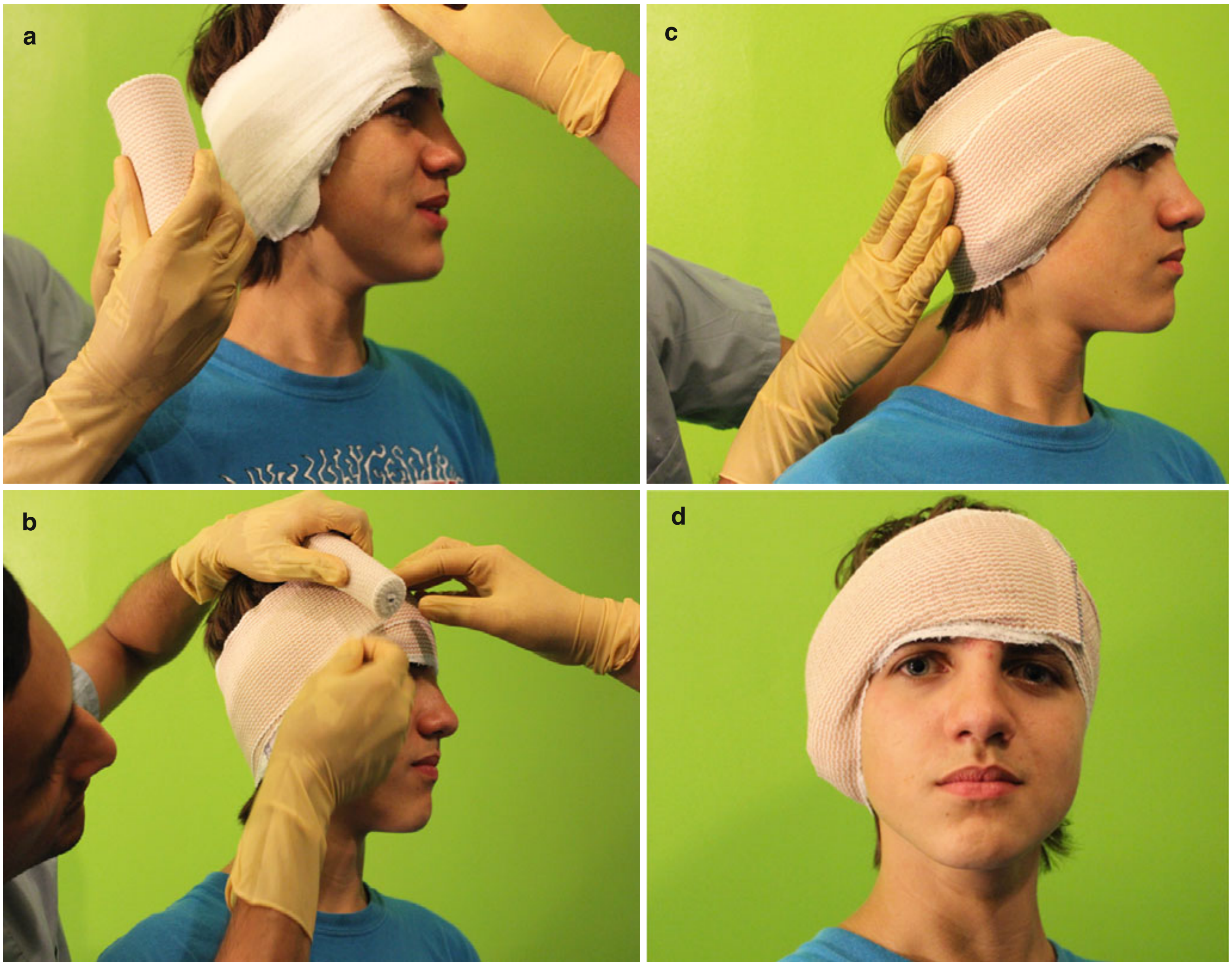

Fig. 58.11 (**a**–**d**) Wrap head with elastic bandage

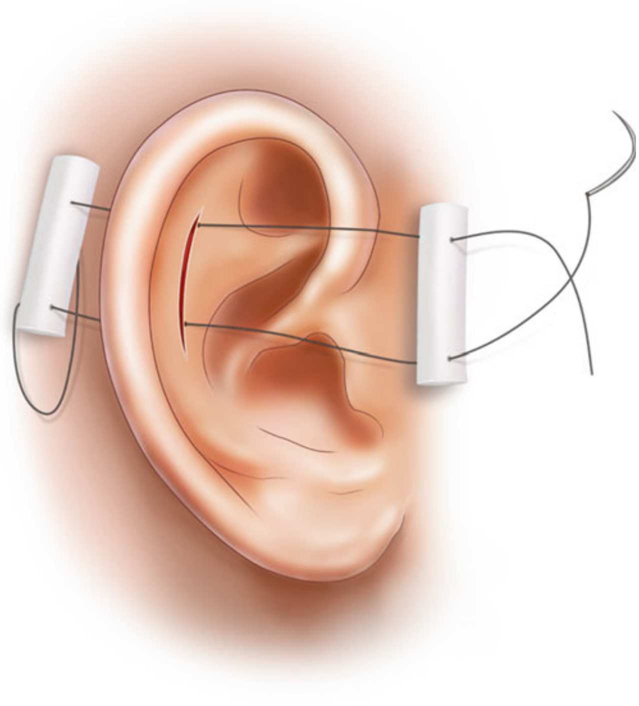

Fig. 58.12 Dental rolls secured to the ear

58.5 Pitfalls and Complications

- Bleeding
- Infection—perichondritis
- Deformity of the auricle
- Inadequate clot removal
- Hematomas present for several days warrant a plastic surgery or an otolaryngology consultation

Acknowledgement The authors would like to thank Thor Shiva Stead for serving as the subject in many of the photographs in this chapter.

Selected Reading

O'Donnell BP, Eliezri YD. The surgical treatment of traumatic hematoma of the auricle. Dermatol Surg. 1999;25:803–5.
Starck WJ, Kaltman AI. Current concepts in the surgical management of traumatic auricular hematoma. J Oral Maxillofac Surg. 1992;50:800–2.

Incision and Drainage of Peritonsillar Abscess

59

Melinda W. Fernandez and Bobby K. Desai

59.1 Indications

- Peritonsillar abscess

59.2 Contraindications

- Absolute
  - Malignancy
  - Vascular malformations
- Relative
  - Pediatric patient
  - Severe trismus
  - Uncooperative patient

59.3 Materials and Medications

- #11 or #15 scalpel
- 27-gauge 1.5-inch needle, 18- to 20-gauge 1.5-inch or longer needle
- 5-mL syringe, 10- to 20-mL syringe
- Tape
- Trauma shears
- Viscous lidocaine
- Lidocaine 1 % with epinephrine
- Laryngoscope with MAC 3 or 4 *or* tongue blade(s) and headlamp or other light source
- Suction setup with Frazier or tonsil suction tip

59.4 Procedure: Aspiration

1. Informed consent may be required.
2. Raise head of bed to at least 60° and place a pillow or other support behind the patient's head.
3. Prepare an 18- to 20-gauge 1.5-inch needle on a 10- to 20-mL syringe with a needle guard.
   - This can be accomplished by using trauma shears to cut 1–1.5 cm off the distal plastic needle cover and replacing the cover over the needle.
   - This now creates a guard to prevent deep penetration into vascular structures (Fig. 59.1).
4. Topically anesthetize area with viscous lidocaine (alternatively, can use Cetacaine [benzocaine, tetracaine hydrochloride, and butamben] spray)
5. Using lidocaine 1 % with epinephrine and a 27-gauge 1.5-inch needle on a 5-mL syringe, infiltrate 1–2 mL into the area. Blanching should be apparent.
6. Palpate the oropharynx with the gloved finger to evaluate for fluctuance.
7. Assemble a MAC 3 or 4 intubation blade on a laryngoscope handle and open into the light-on position. Insert the blade into the patient's mouth, and advance the blade as far posteriorly as possible without inducing gagging. Have the patient hold the laryngoscope handle. Alternatively, use a tongue blade and head lamp or other light source, but the laryngoscope technique allows for an unobstructed view of the area and the weight of the handle helps hold the patient's mouth open.
8. Using the previously prepared needle guard on the long 18- to 20-gauge needle on a 12-mL syringe, insert the needle into the most fluctuant area (as determined from the previous examination), aspirating as the needle advances (Fig. 59.2). The most fluctuant area will usually be located in the superior pole of the tonsil.

- It is very important to be careful not to angle the needle laterally toward the carotid artery. Also remember, this is a peritonsillar abscess so do not aspirate the tonsil itself.

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_59

9. After aspiration, if there is a pus return, remove as much purulent drainage as possible. Typical drainage is between 2 and 6 mL. If a large amount of pus is obtained (>6 mL), incision and drainage (I&D) may be indicated. (See later for details on I&D.)
10. If the aspirate did not return pus, attempt aspiration again by moving the insertion site 1 cm inferiorly to the middle pole of the peritonsillar space. If there is still no pus aspirated, make a final attempt by moving again 1 cm inferiorly to the lower pole of the peritonsillar space.
11. Suction should be set up, readily available, and turned on. Use suction to prevent the patient from aspirating or swallowing any purulent drainage.
12. Expect a small amount of bleeding when the procedure is complete and the needle is removed.

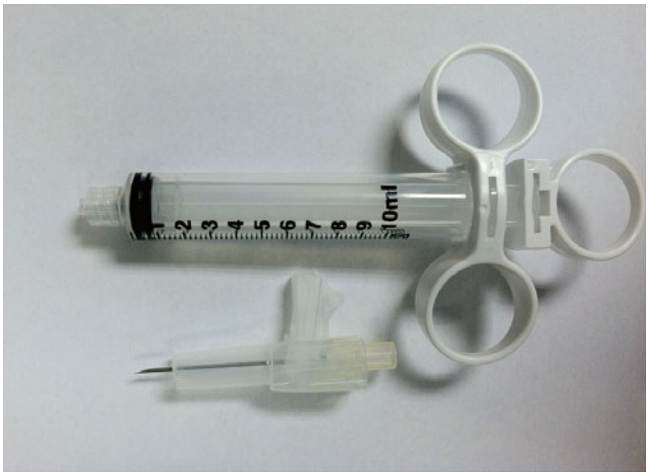

Fig. 59.1 Guard to prevent deep penetration of vascular structures

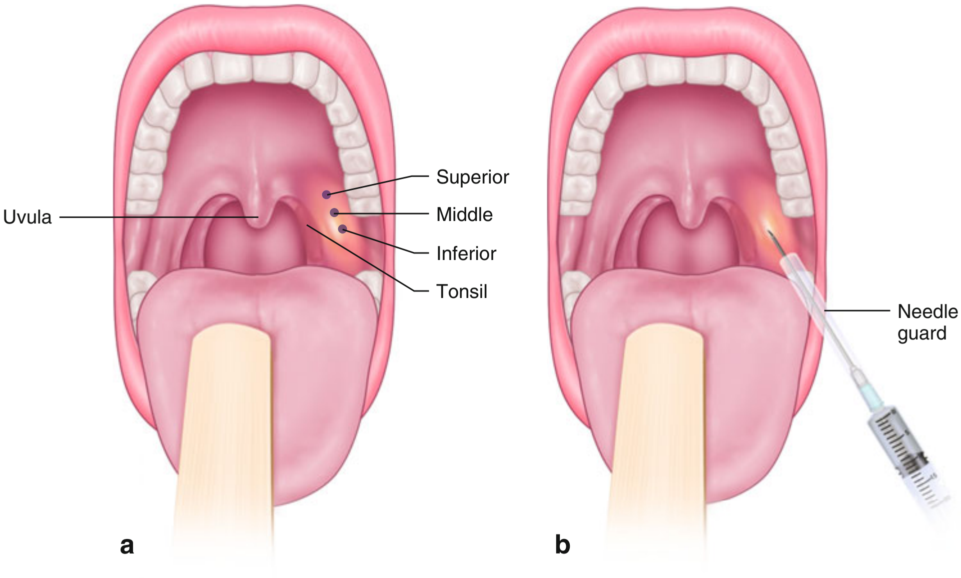

Fig. 59.2 Aspiration of a peritonsillar abscess: (**a**) Aspirate superior pole first. If no pus aspirated, move 1 cm inferior to the middle pole. If still no pus, make final attempt by moving 1 more cm inferiorly to the inferior pole. (**b**) Demonstrates use of needle guard

59.5 Procedure: Incision and Drainage

1. Raise the head of the bed and anesthetize the area as described previously.
2. Have suction set up with a Fraser or tonsil tip, available, and turned on.
3. Have the patient hold the laryngoscope handle after the blade has been inserted orally as described previously.
4. Fashion a blade guard on a #11 or #15 blade scalpel by taping over the blade with only 0.5 cm of the blade exposed at the tip. Incise the area that was previously aspirated with a small stab incision. Suction the area.
5. Insert a curved Kelly clamp into the incision and open gently to enter the abscess cavity and break up loculations (Fig. 59.3). Suction as necessary.
6. Have the patient gargle with a saline solution and expectorate.
7. Do not pack the abscess cavity.
8. Observe for 2–4 h for bleeding.

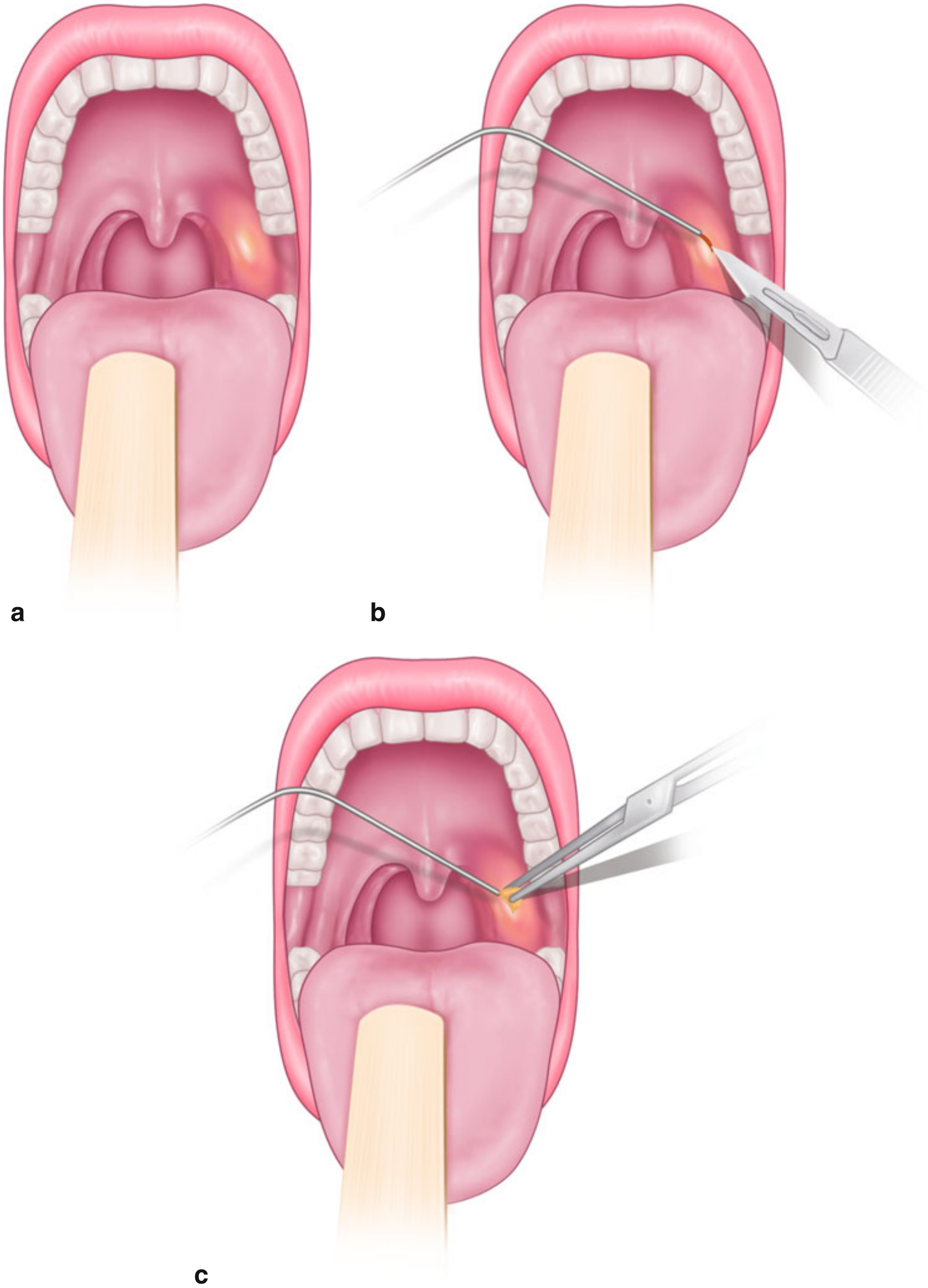

Fig. 59.3 Incision and drainage of peritonsillar abscess: (**a**) Gain access with appropriate lighting. (**b**) Fashion a blade guard on a #11 or #15 blade scalpel by taping over the blade with only 0.5 cm of the blade exposed at the tip. Incise the area that was previously aspirated with a small stab incision. (**c**) Insert a curved Kelly clamp into the incision and open gently to enter the abscess cavity and break up loculations. Do not pack!

59.6 Complications

- Aspiration or incision of the carotid artery
- Excessive bleeding
- Aspiration of purulent material
- Pain

59.7 Pearls and Pitfalls

- Pearls
  - Airway protection should always be considered. Intubation may be necessary for very large abscesses with airway compromise.
  - Bedsides ultrasound is a valuable tool in confirming the diagnosis of peritonsillar abscess. Perform using the intracavitary probe covered with a sterile glove or other appropriate probe cover and insert into the oropharynx. The abscess will appear as any other abscess: an encapsulated, hypoechoic structure. Doppler flow can also be used to locate the carotid artery relative to the abscess.
  - Empirical oral antibiotics after aspiration or I&D are considered usual care and should cover for group A streptococcus and oral anaerobes.
    - Amoxicillin/clavulanate or clindamycin is the most commonly used.
  - A single dose of high-dose steroids may be helpful in relieving symptoms.
  - Patients who cannot tolerate oral fluids, cannot take oral antibiotics, or who appear to have a toxic response should be admitted. Others may be discharged with 24-h follow-up.
- Pitfalls
  - The carotid artery is located approximately 2.5 cm posterolaterally to the tonsils. Take care that the needle is not inserted too far laterally or the risk of aspirating the carotid artery is increased.
  - Incision into the tonsil itself may cause excessive bleeding. Aspiration or incision into the tonsil will likely miss the abscess altogether and may result in misdiagnosis.
  - There is a 1–15 % failure and recurrence rate.

Selected Reading

Afarian H, Lin M. Tricks of the trade—say “ah!”—needle aspiration of peritonsillar abscess. ACEP News. 2008;27(5):38.

Braude DA, Shalit M. A novel approach to enhance visualization during drainage of peritonsillar abscess. J Emerg Med. 2008;35:297–8.

Galioto NJ. Peritonsillar abscess. Am Fam Physician. 2008;77:199–202.

Ozbek C, Aygenc E, Tuna EU, Selcuk A, Ozdem C. Use of steroids in the treatment of peritonsillar abscess. J Laryngol Otol. 2004;118:439–42.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Vieira F, Allen SM, Stocks RM, Thompson JW. Deep neck infection. Otolaryngol Clin North Am. 2008;41:459–83.

Incision and Drainage of Sublingual Abscess

60

Melinda W. Fernandez and Bobby K. Desai

60.1 Indications

- Sublingual abscess (Fig. 60.1)

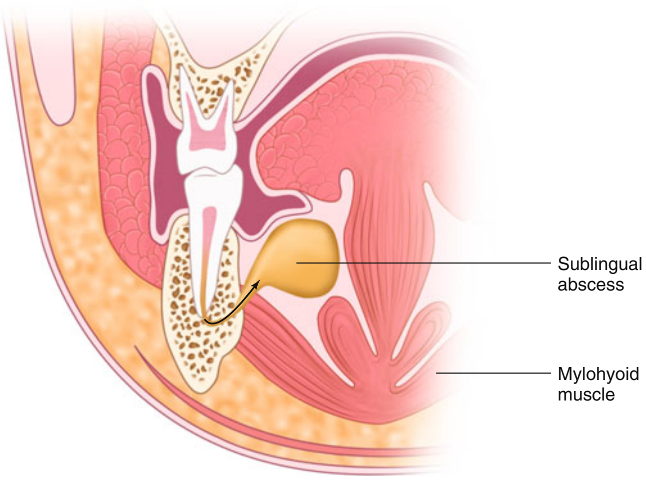

Fig. 60.1 Sublingual abscess

M.W. Fernandez, MD
Department of Emergency Medicine, University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_60

60.2 Contraindications

- Relative
  - Pediatric patient
  - Severe trismus—may need sedation or drainage in the operating room
  - Uncooperative patient—may need sedation or drainage in the operating room
  - Coagulopathy, patients taking anticoagulants, or patients with a known bleeding disorder

60.3 Materials and Medications (Fig. 60.2)

- #11 scalpel
- Hemostats
- Penrose drain
- 4-0 silk suture
- 25- to 27-gauge 1.5-inch needle for injecting anesthetic
- 18- to 20-gauge needle for drawing up anesthetic
- 5- to 10-mL syringe or control syringe
- Viscous lidocaine (or other topical anesthetic)
- Lidocaine 1 % with epinephrine
- Light source
- Suction setup with tonsil suction tip
- Culture swab

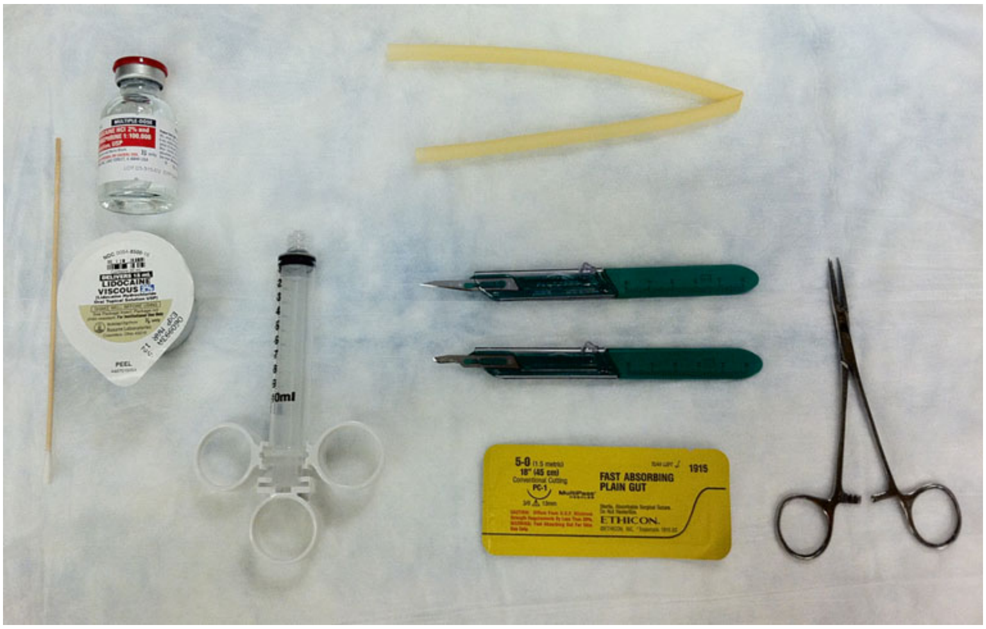

Fig. 60.2 Materials needed: #11 scalpel, hemostats, Penrose drain, 4-0 silk suture, 25–27-gauge 1.5-inch needle for injecting anesthetic, 18- to 20-gauge needle for drawing up anesthetic, 5- to 10-mL syringe or control syringe, viscous lidocaine (or other topical anesthetic), lidocaine 1 % with epinephrine, light source, suction setup with tonsil suction tip, culture swab (optional)

60.4 Procedure: Incision and Drainage

1. After explaining the procedure, risks, and benefits to the patient, put the head of the bed in the most comfortable working position for the clinician and the patient. Alternatively, the clinician may also elect to position the patient in an upright procedure chair that has multiple positions.
2. Apply viscous lidocaine topically with a cotton-tipped swab to the area to be injected with anesthetic. Leave in place for a minute or 2. Alternatively, spray the area with a topical anesthetic spray.
3. Have suction set up with a tonsil tip, available, and turned on.
4. Draw up the appropriate amount of lidocaine with epinephrine into a 5-mL syringe (control syringe if available).
5. Direct light source to area to be worked on. *Adequate lighting is essential.* Options include a headlamp or an overhead adjustable light.
6. Change needle to the 27-gauge needle and inject 1–2 mL of lidocaine with epinephrine using an inferior alveolar block. Alternatively, anesthetize the floor of the mouth around the most fluctuant area. Avoid injecting through infected tissue to avoid possible spread into deeper spaces.
7. Using the scalpel, make an intraoral stab incision superficially at the lowest point of the pus accumulation (Fig. 60.3). This will facilitate evacuation of pus under gravity. Have the suction catheter in the mouth and allow the purulent material to drain into the suction.
8. Obtain a specimen of the purulent fluid for culture and sensitivity.
9. Insert a hemostat into the incision to facilitate drainage, but do not open it up to avoid injury to neurovascular structures. Gently massage the soft tissue surrounding the abscess to assist drainage. Suction as necessary to avoid swallowing or aspirating the pus.
10. Once adequate drainage has been achieved, place a small Penrose drain (or other rubber-type drain) into the cavity, and stabilize on one side with a silk suture that goes through the drain and the mucosa (Fig. 60.4).
11. Have the patient rinse and spit with a half-strength peroxide solution followed by either water or saline.
12. Watch for signs of bleeding or upper airway symptoms.
13. Ensure that the patient can tolerate oral fluids before discharge.

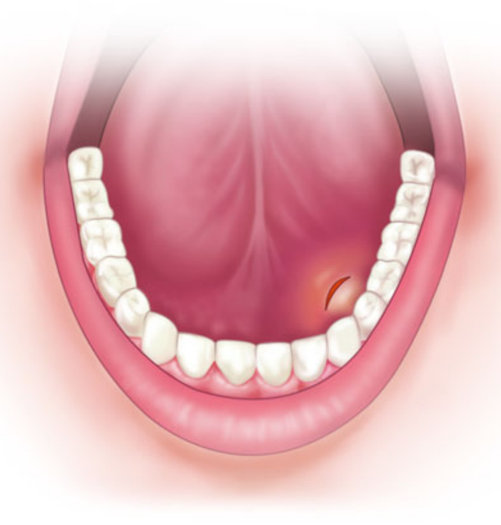

Fig. 60.3 Incise lowest portion of abscess

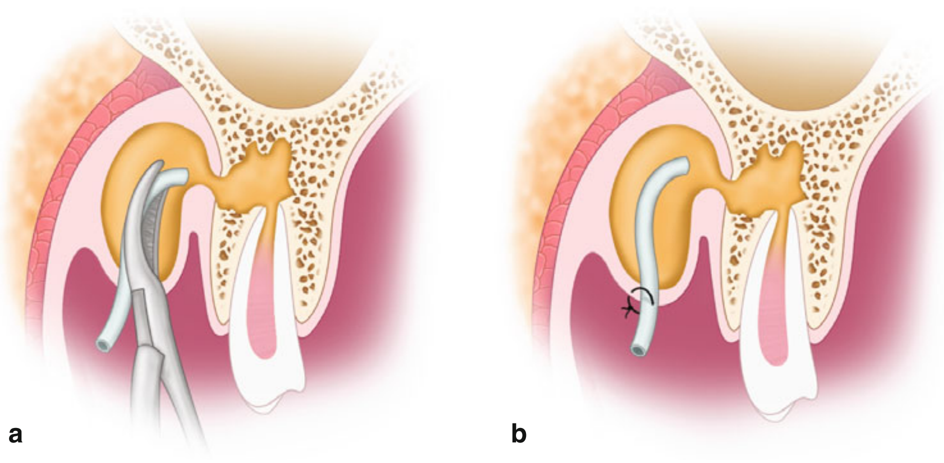

Fig. 60.4 (**a**) Using hemostats, place small Penrose drain into the abscess cavity. (**b**) Suture Penrose in place with a silk suture that goes through the drain and the mucosa

60.5 Complications (Generally Minimal)

- Excessive bleeding (apply pressure—to avoid potential nerve injury, do not cauterize or ligate unless absolutely necessary)
- Aspiration of purulent material (have the patient sit upright and use suction as the abscess is incised)
- Pain

60.6 Pearls and Pitfalls

- Pearls
  - Airway protection is paramount. Infections in this space can quickly spread to the submandibular region and can compromise the airway owing to swelling (Ludwig's angina).
    - Assess for elevation of the floor of the mouth and tongue, ability to lie supine, drooling, stridor, and restlessness.
    - If these signs are present, emergent surgical airway will be necessary before incision and drainage (I&D).
  - The sublingual space is bounded superiorly by the oral mucosa and inferiorly by the mylohyoid muscle.
  - Infected premolars and first molars drain into this space because the apices of their roots are located superior to the mylohyoid muscle.
  - The source molar should be extracted as soon as possible after I&D.
  - Empirical oral antibiotics after I&D should cover for group A streptococcus and oral anaerobes. Penicillin remains the drug of choice, but clindamycin or amoxicillin–clavulanic acid can be substituted.
- Pitfalls
  - The lingual artery, vein, and nerve are contained in the posterolateral area of the floor of the mouth, and the hypoglossal nerve is nearby. These must be avoided when performing I&D for sublingual abscesses.

Selected Reading

Flynn TR, Shanti RM, Levi MH, Adamo AK, Kraut RA, Trieger N. Severe odontogenic infections, part 1: prospective report. J Oral Maxillofac Surg. 2006;64:1093–103.

Reichman E, Simon R, editors. Emergency medicine procedures. New York: McGraw-Hill Education; 2003. p. 1342–5.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Vieira F, Allen SM, Stocks RM, Thompson JW. Deep neck infection. Otolaryngol Clin North Am. 2008;41:459–83.

Incision and Drainage of Parotid Duct Abscess

61

Melinda W. Fernandez and Bobby K. Desai

61.1 Indications

- Parotid duct abscess

61.2 Contraindications

- Absolute
  - None
- Relative
  - Pediatric patient
  - Severe trismus—may need sedation or drainage in the operating room
  - Uncooperative patient—may need sedation or drainage in the operating room
  - Coagulopathy, patients taking anticoagulants, or patients with a known bleeding disorder

61.3 Materials and Medications (Fig. 61.1)

- #11 scalpel
- 4×4 gauze
- Hemostat
- Penrose drain or ¼-inch packing gauze
- Light source (headlamp or overhead light)
- Culture swab
- 4-0 silk suture
- 25- to 27-gauge needle, 1.5–2 inches long
- 18-gauge needle to withdraw anesthetic from vial
- 5-mL syringe
- Viscous lidocaine or other topical anesthetic
- Lidocaine with epinephrine
- Suction setup with Frazier or tonsil suction tip

M.W. Fernandez, MD
Department of Emergency Medicine,
University of Florida Health, Gainesville, FL, USA
e-mail: mindyfernandez@ufl.edu

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_61

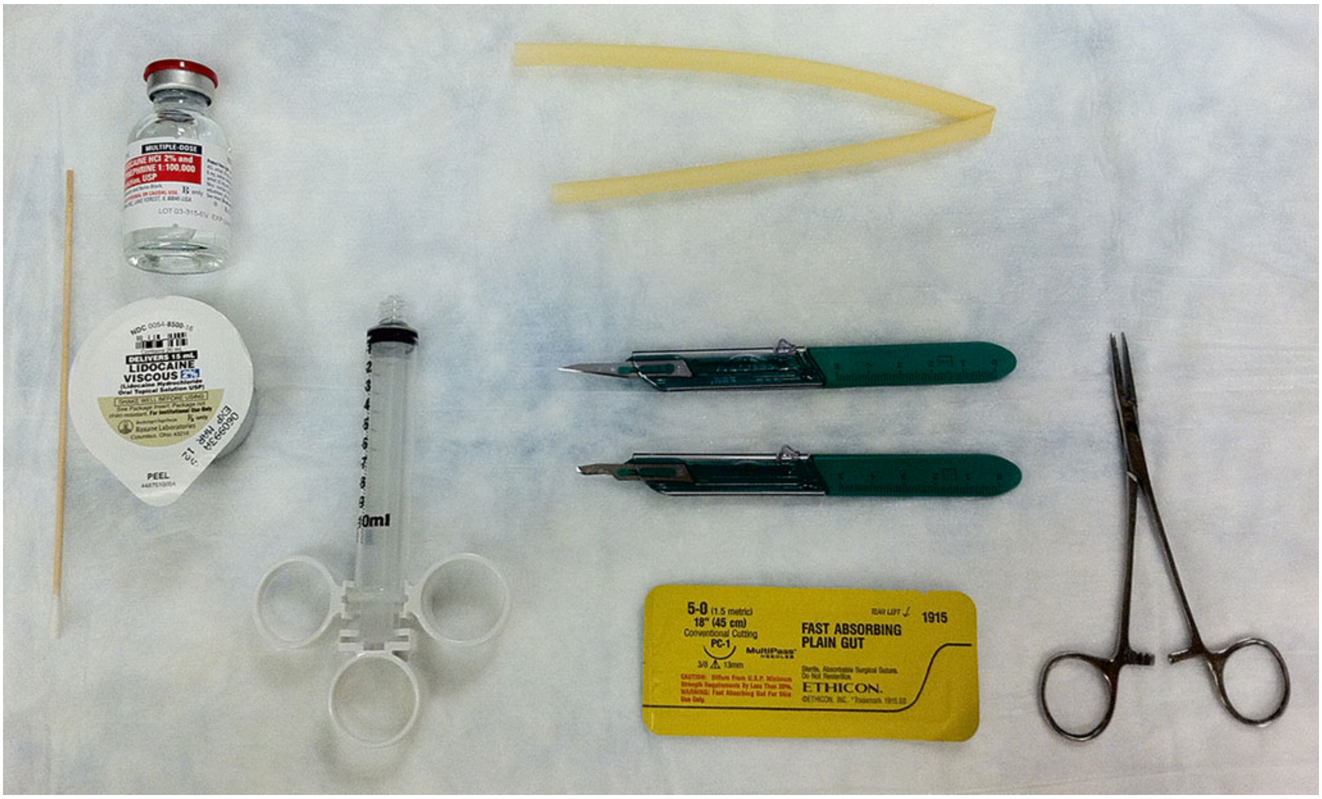

Fig. 61.1 Materials needed: #11 scalpel, hemostats, Penrose drain, 4-0 silk suture, 25- to 27-gauge 1.5-inch needle for injecting anesthetic, 18- to 20-gauge needle for drawing up anesthetic, 5- to 10-mL syringe or control syringe, viscous lidocaine (or other topical anesthetic), lidocaine 1 % with epinephrine, light source, suction setup with tonsil suction tip, culture swab (optional)

61.4 Procedure (Figs. 61.2 and 61.3)

1. After explaining the procedure, risks, and benefits to the patient, put the head of the bed in the most comfortable working position for the operator and the patient. Alternatively, position the patient in an upright procedure chair that has multiple positions.
2. Find the most fluctuant area. Apply viscous lidocaine topically with a cotton-tipped swab to the area to be injected with anesthetic. Leave in place for a minute or 2. Alternatively, spray the area with a topical anesthetic spray.
3. Have suction setup with a tonsil tip, available, and turned on.
4. Draw up the appropriate amount of lidocaine with epinephrine into a 5-mL syringe (control syringe if available).
5. Direct light source to area to be worked on. *Adequate lighting is essential.* Options include a headlamp or an overhead adjustable light.
6. Change needle to a 25- or 27-gauge needle. Inject 1–2 mL of lidocaine with epinephrine just beneath the mucosal surface. Avoid injecting through infected tissue to avoid possible spread into deeper spaces.
7. Using the scalpel, make an intraoral stab incision superficially into the area of greatest fluctuance (Fig. 61.4). Have the suction catheter in the patient's mouth and allow the purulent material to drain into the suction.
8. Obtain a specimen of the purulent fluid for culture and sensitivity.
9. Insert a hemostat into the incision to facilitate drainage, but do not open it up in order to avoid injury to neurovascular structures. Gently massage the soft tissue surrounding the abscess to assist drainage. Suction as necessary to avoid swallowing or aspirating the pus.
10. Have the patient rinse and spit with a half-strength peroxide solution.
11. Place a small Penrose drain (or other rubber-type drain) into the cavity, and stabilize on one side with a silk suture that goes through the drain and the mucosa. Alternatively, cut a strip of ¼- inch packing gauze and insert into the abscess cavity.
12. Have the patient rinse and spit with a half-strength peroxide solution followed by either water or saline.
13. Watch for signs of bleeding or upper airway symptoms.
14. Ensure that the patient can tolerate oral fluids before discharge.

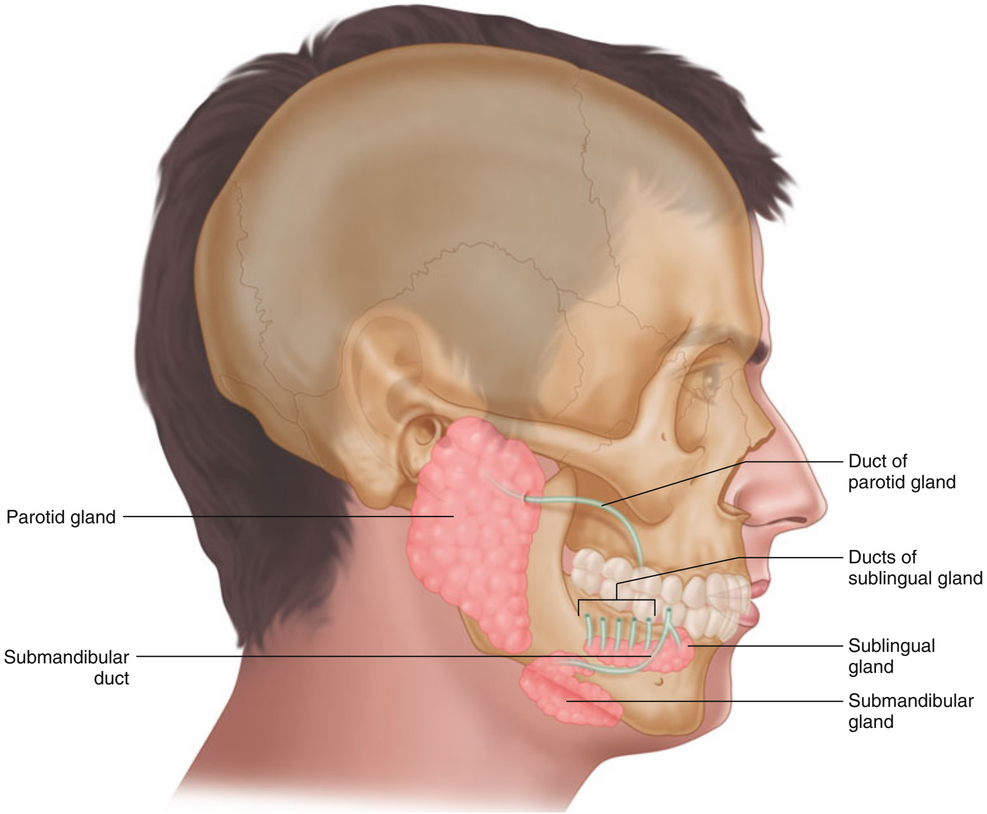

Fig. 61.2 Anatomy of the salivary glands

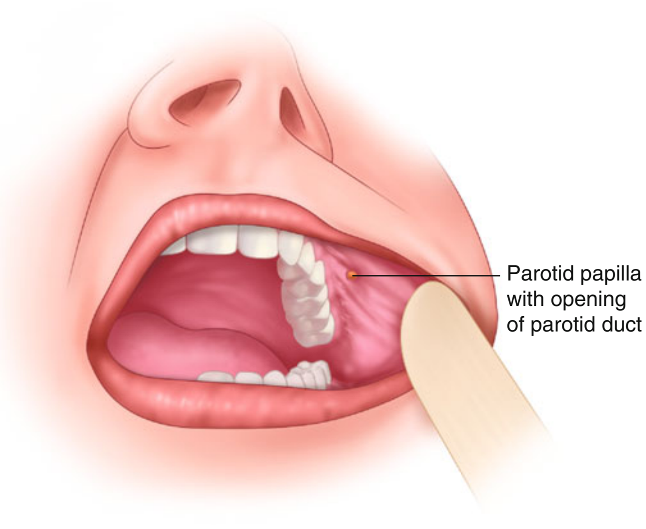

Fig. 61.3 Parotid papilla: opening of the parotid gland located adjacent to third upper molar

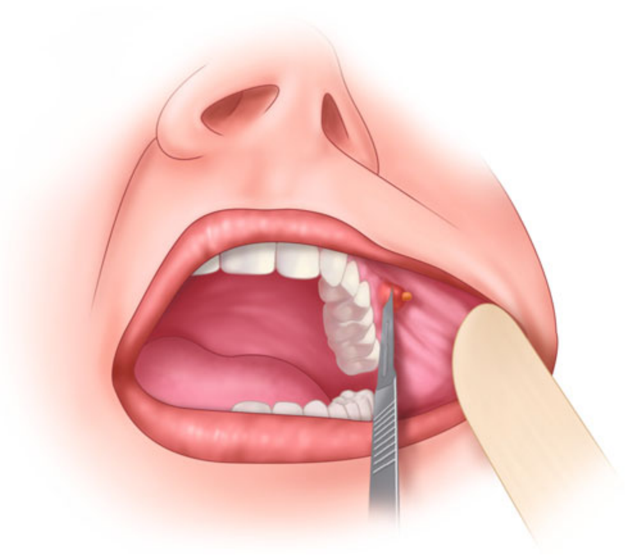

Fig. 61.4 Parotid abscess; make a superficial stab into most fluctuant area

61.5 Complications

- Complications are generally minimal but may include:
  - Excessive bleeding (apply pressure; to avoid potential nerve injury, do not cauterize or ligate unless absolutely necessary)
  - Aspiration of purulent material (have the patient sit upright and use suction as the abscess is incised)
  - Pain

61.6 Pearls and Pitfalls

- Airway protection should always be considered.
- Empirical oral antibiotics after incision and drainage are considered usual care and should cover for group A streptococcal and oral anaerobes. Penicillin remains the drug of choice, but clindamycin or amoxicillin–clavulanic acid can be substituted.
- Patients who cannot tolerate oral medications, cannot take oral antibiotics, or who appear to have a toxic response should be admitted. Pediatric patients should be admitted. Others may be discharged with 24-h follow-up.

Selected Reading

Reichman E, Simon R, editors. Emergency medicine procedures. New York: McGraw-Hill Education; 2003. p. 1346–9.

Roberts J, Hedges J, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. p. 1184–9.

Part IX

Dental Procedures

Techniques of Mandibular Anesthesia

62

Susana Perry, Joshua Perry, and Rosalia Rey

62.1 Inferior Alveolar Nerve Block

62.1.1 Nerves Anesthetized

- Inferior alveolar, branch of the posterior division of the mandibular nerve (V3, branch of the trigeminal nerve)
- Incisive
- Mental
- Lingual (usually)

62.1.2 Areas Anesthetized (Fig. 62.1)

- Mandibular teeth to midline
- Body of the mandible
- Buccal mucoperiosteum, mucous membrane anterior to the mandibular first molar
- Anterior two thirds of the tongue and floor of the mouth (via the lingual nerve)
- Lingual soft tissues and periosteum (via the lingual nerve)

S. Perry, DMD (✉)
Department of Pediatric Dentistry, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: sperry@dental.ufl.edu

J. Perry, DMD
Department of Prosthodontics, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: jperry@dental.ufl.edu

R. Rey, DDS
Department of Restorative Dental Sciences, University of Florida College of Dentistry, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: RREY@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_62

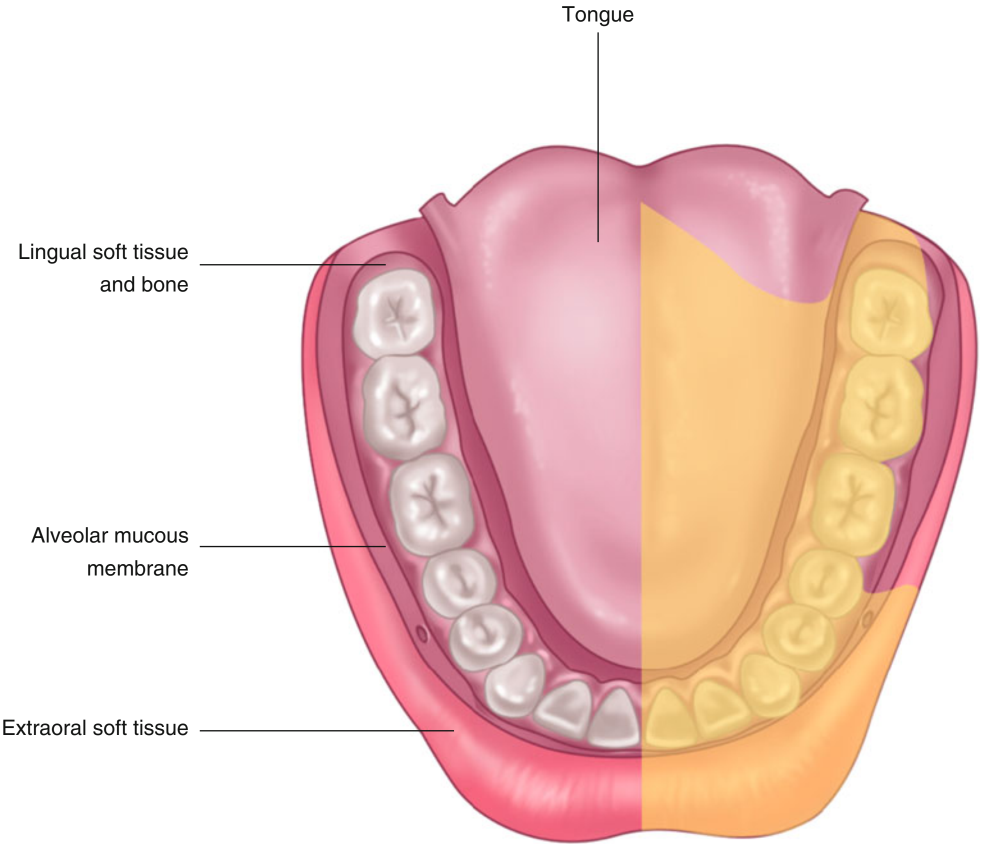

Fig. 62.1 Areas anesthetized with inferior mandibular nerve block

62.1.3 Indications

- When buccal soft tissue anesthesia is necessary
- When lingual soft tissue anesthesia is needed
- When performing procedures on multiple mandibular teeth in one quadrant

62.1.4 Contraindications

- Infection or acute inflammation in the area of injection
- Patients who might bite their lip or tongue (e.g., very young child or physically or mentally handicapped adult or child)
- In relation to local anesthetic use

Absolute

- Local anesthetic allergy
  - Avoid all local anesthetics in the same chemical class (e.g., esters).
- Bisulfite allergy
  - Avoid vasoconstrictor-containing local anesthetics.

Relative

- Atypical plasma cholinesterase
- Methemoglobinemia (idiopathic or congenital)
- Significant liver dysfunction (American Society of Anesthesiologists [ASA] III–IV)
- Significant kidney dysfunction (ASA III–IV)
- Significant cardiovascular disease (ASA III–IV)
  - Avoid high concentrations of vasoconstrictors.
  - Use local anesthetics with epinephrine concentrations of 1:200,000 or 1:100,000 or 3 % mepivacaine or 4 % prilocaine.
- Clinical hyperthyroidism (ASA III–IV)
  - Avoid high concentrations of vasoconstrictors.
  - Use local anesthetics with epinephrine concentrations of 1:200,000 or 1:100,000 or 3 % mepivacaine or 4 % prilocaine.

62.1.5 Materials and Medications

- Local anesthetic carpule (1.7–1.8 mL)
  - Mepivacaine 3 % (+epinephrine 1:100,000)
  - Articaine HCI 4 % (+epinephrine 1:100,000 or 1:200,000)
  - Lidocaine HCI 2 % (+ epinephrine 1:50,000 or 1:100,000) (Fig. 62.2)
  - Bupivacaine HCI 0.5 %+epinephrine 1:200,000
- Aspirating syringe (Fig. 62.3)
- Needle (Fig. 62.4)
  - Gauge refers to the lumen of the needle: The smaller the number, the greater the diameter of the lumen.
  - Needles are color coded by gauge: red=25 gauge, yellow=27 gauge, and blue=30 gauge.
  - *Recommendations: For inferior alveolar nerve (IAN) block, it is best to use a 25-gauge long needle.*
- Mouth props
- Retractors

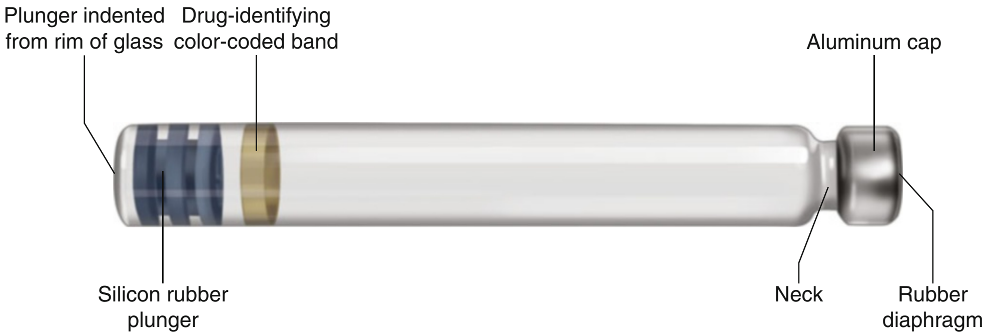
Plunger indented from rim of glass
Drug-identifying color-coded band
Aluminum cap
Silicon rubber plunger
Neck
Rubber diaphragm

Fig. 62.2 Local anesthetic carpule (1.7–1.8 mL)

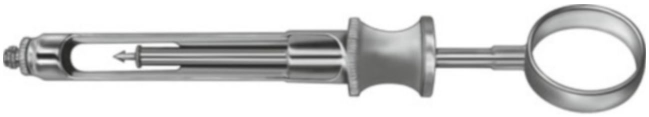

Fig. 62.3 Aspirating syringe

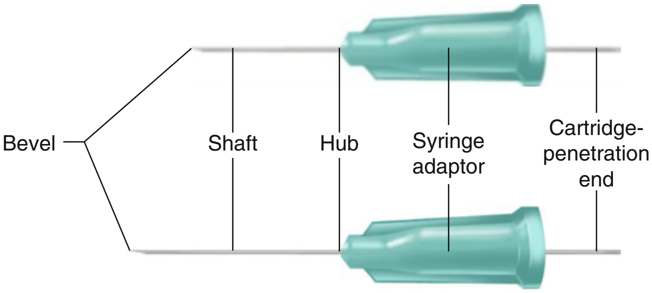
Bevel
Shaft
Hub
Syringe adaptor
Cartridge-penetration end

Fig. 62.4 Needle

62.1.6 Procedure

1. Target area: IAN as it passes downward toward the mandibular foramen.
2. Landmarks.
   - (a) Coronoid notch
   - (b) Pterygomandibular raphe
   - (c) Occlusal plane of mandibular posterior teeth
3. Procedure.
   - (a) Correct position for operator.
     - (i) For a right IAN block, a right-handed administrator should sit at the 8 o'clock position facing the patient.
     - (ii) For a left IAN block, a right-handed administrator should sit at the 10 o'clock position facing in the same direction as the patient.
   - (b) Recommended to position the patient supine and with the mouth wide open.
   - (c) Place thumb on the coronoid notch and index finger extraorally on the posterior border of the ramus in order to estimate the distance between these two points (Fig. 62.5).
     - (i) The needle insertion should be three fourths of the anteroposterior distance from the coronoid notch to the deepest part of the pterygomandibular raphe.
   - (d) Place the barrel of the syringe in the corner of the mouth on the contralateral side, usually corresponding to the premolars.
   - (e) Slowly advance the needle until bony resistance is met.
     - (i) For anxious or sensitive patients, a small volume of anesthetic may be deposited as the soft tissue is penetrated.
     - (ii) Average depth of penetration to bony contact will be 20–25 mm, approximately two thirds to three fourths the length of a long needle.
     - (iii) If the bone is contacted too soon (less than half the length of a long needle), the needle tip is usually located too far anteriorly (laterally) on the ramus. To correct:
       - Withdraw it slightly from the tissues and bring the syringe barrel anteriorly toward the lateral incisor or canine; reinsert to the proper depth.
     - (iv) If the bone is not contacted, the needle tip is usually located too far posterior (medial). To correct:
       - Withdraw it slightly in tissue (leaving approximately one fourth its length in tissue), and reposition the syringe barrel more posteriorly (over the mandibular molars).
       - Continue the insertion until contact with the bone is made at an appropriate depth (20–25 mm).
   - (f) Aspirate. If negative, slowly deposit 1.5 mL of anesthesia over 60 s.
   - (g) Wait 3–5 min before commencing the dental procedure.
4. Precaution: *Do not deposit anesthesia if the bone is not contacted*. The needle tip may be resting within the parotid gland near the facial nerve (cranial nerve VII), and a transient paralysis of the facial never may occur if solution is deposited.

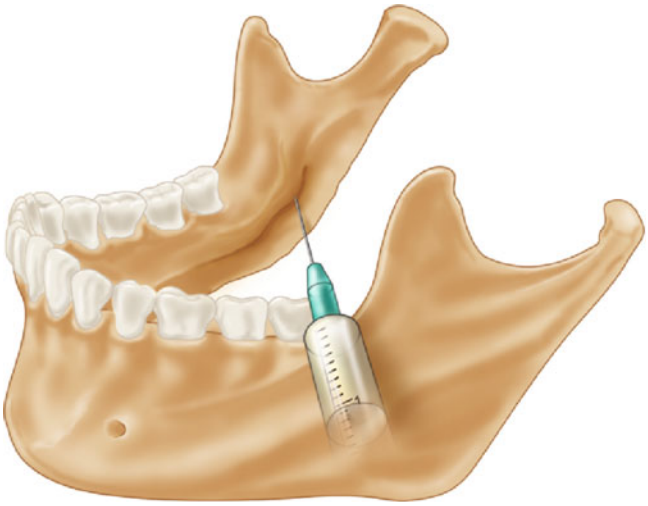

Fig. 62.5 Needle orientation for inferior alveolar nerve block

62.2 Buccal Nerve Block

62.2.1 Nerve Anesthetized

- Buccal nerve, a branch of the anterior division of the mandibular nerve

62.2.2 Area Anesthetized (Fig. 62.6)

- Soft tissues and periosteum buccal to the mandibular molars

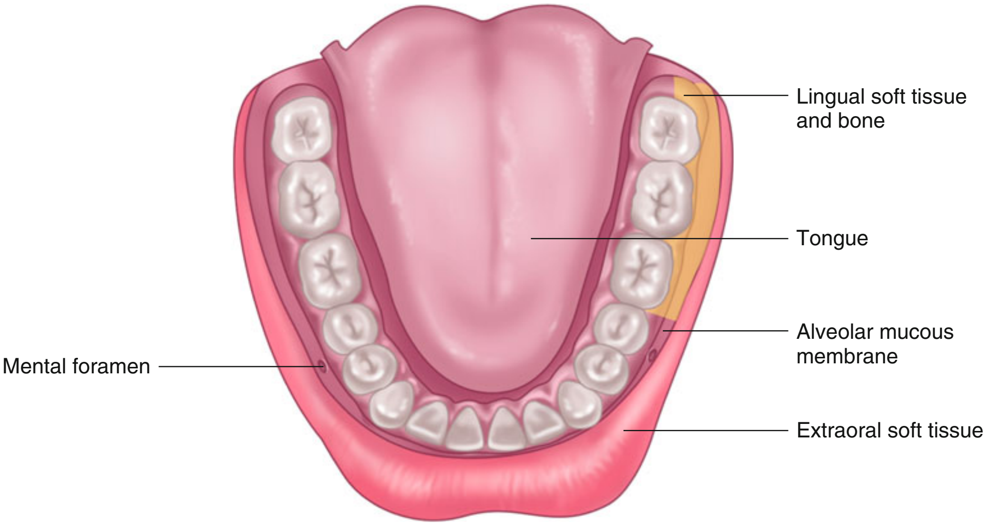
Lingual soft tissue and bone
Tongue
Alveolar mucous membrane
Mental foramen
Extraoral soft tissue

Fig. 62.6 Areas anesthetized with buccal nerve block

62.2.3 Procedure

1. A 25-gauge long needle is recommended.
2. Landmarks: mandibular molars and mucobuccal fold.
3. Orient the bevel of the needle *toward* the bone during injection.
4. Correct positioning.
   (a) For a right buccal nerve block, a right-handed administrator should sit at the 8 o'clock position directly facing the patient.
   (b) For a left buccal nerve block, a right-handed administrator should sit at the 10 o'clock facing in the same direction as the patient.
5. Procedure.
   (a) With the index finger, pull the buccal soft tissues in the area of injection laterally to allow for better visualization.
   (b) Align the syringe parallel to the occlusal plane of the teeth.
   (c) Penetrate the mucous membrane at the injection site, distal and buccal to the last molar (Fig. 62.7).
6. If tissue at the injection site becomes swollen, stop depositing solution.

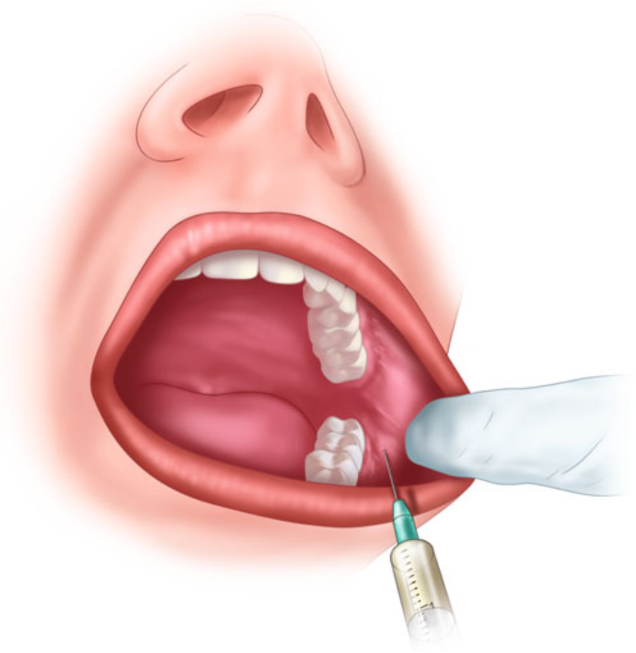

Fig. 62.7 Needle orientation for buccal nerve block

62.3 Mental Nerve Block

62.3.1 Nerve Anesthetized

- Mental nerve, a terminal branch of the inferior alveolar nerve

62.3.2 Area Anesthetized (Fig. 62.8)

- Buccal mucous membranes anterior to the foramen (around the second premolar) to the midline and skin of the lower lip

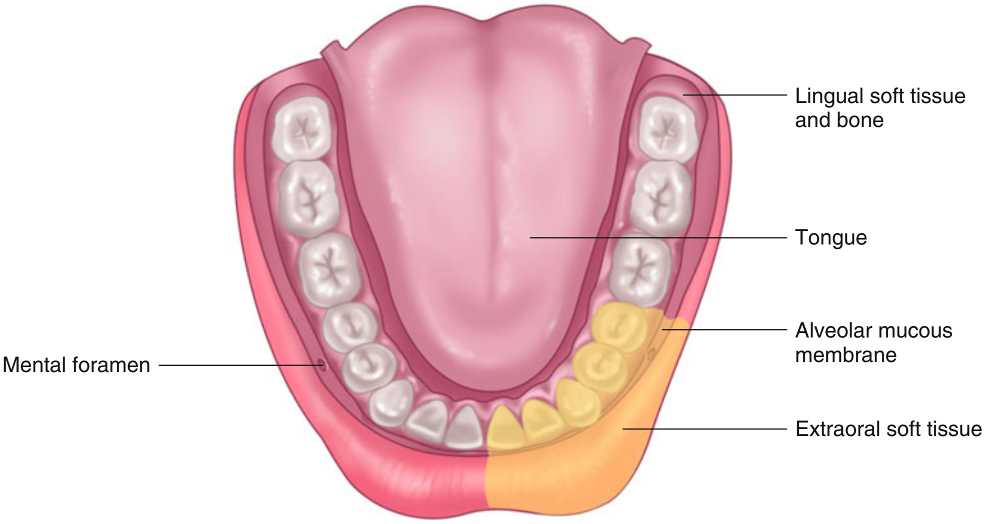
Lingual soft tissue and bone
Tongue
Alveolar mucous membrane
Extraoral soft tissue
Mental foramen

Fig. 62.8 Areas anesthetized with mental nerve block

62.3.3 Indications

- For buccal soft tissue anesthesia in procedures such as:
  - Soft tissue biopsies
  - Suturing of soft tissues

62.3.4 Procedure

1. Area of insertion: mucobuccal fold at or just anterior to the mental foramen.
2. Orientation of bevel should be *toward* the bone during injection.
3. Operator should sit in front of the patient so that the syringe is below the patient's line of sight.
4. Locate the mental foramen.
   (a) Place the index finger in the mucobuccal fold and press against the body of the mandible in the area of the first molar.
   (b) Move the finger anteriorly until the bone beneath the finger feels somewhat concave.
   (c) Mental foramen is usually found around the apex of the second premolar.
   (d) Orient the needle with the bevel directed toward the bone.
   (e) Penetrate the mucous membrane and advance needle slowly; penetration depth is usually 5–6 mm.
   (f) If aspiration is negative, deposit approximately one third of the cartridge over 20 s.
      (i) If the site balloons, stop the deposition of anesthetic and remove the syringe.

Selected Reading

Bennett CR. Monheim's local anesthesia and pain control in dental practice. 6th ed. St. Louis: Mosby; 1978.

Gow-Gates GAE. Mandibular conduction anesthesia: a new technique using extraoral landmarks. Oral Surg. 1973;36:321–8.

Jastak JT, Yagiela JA, Donaldson D. Local anesthesia of the oral cavity. Philadelphia: WB Saunders; 1995.

Malamed SF. The Gow-Gates mandibular block: evaluation after 4275 cases. Oral Surg. 1981;51:463.

Malamed SF. Handbook of local anesthesia. 5th ed. St. Louis: Mosby; 2004.

Reduction of Dislocated Temporomandibular Joint

63

Christopher J. Spencer and Geraldine Weinstein

63.1 Indications

- Open lock: associated with yawning, vomiting, or opening the mouth wide
- Open lock: associated with a dental procedure
- Open lock: associated with endoscopy
- Open lock: associated with oral intubation
- Time duration: acute to 3 weeks or less duration

63.2 Contraindications for Closed Reduction

- Absolute
  - Head trauma with fracture of the skull, maxilla, mandible, or mandibular
    Condyles
- Relative
  - Dislocation of 30 days or longer (will likely be unable to accomplish reduction without general anesthesia and/or open surgical approach)

63.3 Materials and Medications

- Local anesthetic syringe.
- Lidocaine 2 % 1–2 mL.
- 25- to 27-gauge needle (long or approximately 2 inches long).
- Betadine (povidone-iodine) or other skin antiseptic preparation.
- Gauze padding for thumbs.
- Consider a muscle relaxant.
- Consider conscious sedation.

C.J. Spencer, DDS • G. Weinstein, DDS (✉)
Department of Restorative Dental Sciences, University of Florida
College of Dentistry, Gainesville, FL, USA
e-mail: cspencer@dental.ufl.edu; gweinstein@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_63

63.4 Procedure

63.4.1 Manual Closed Reduction *without* Local Anesthesia (Figs. 63.1)

1. Position the patient in an upright posture with the mandible at the physician's flexed elbow height (physician's comfortable position).
2. Place the thumbs on the mandibular molars with wrapping around the thumbs to protect from possible biting force once the mandible reduces.
3. Apply bilateral firm force in an inferior direction.
4. The mandible will move rapidly in an inferior and then a posterior direction as the condyles slide back over the height of their respective articular eminences.

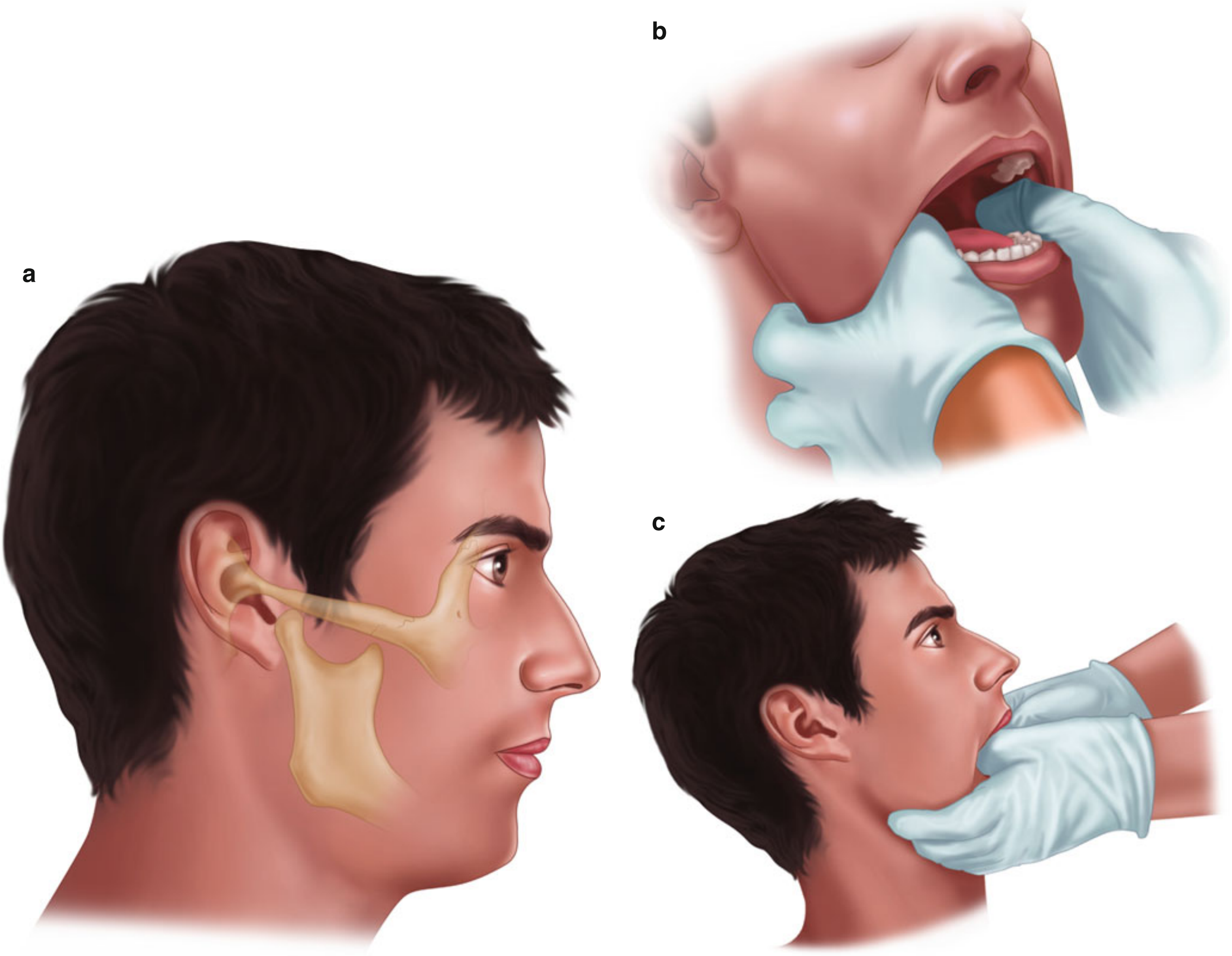

Fig. 63.1 Reduction in progress: (**a**) In a mandibular dislocation, the condyle will be anterior and superior to the articular eminence. (**b**) Position the thumbs on mandibular molars, and apply firm pressure in an inferior direction to distract the TM joint condyles so that they can reposition themselves into the glenoid fossa so that the TM joint can be reduced. (**c**) Lateral view of the distraction force with direction in an inferior direction to distract the condyle of the TM joint

63.4.2 Manual Closed Reduction *with* Local Anesthesia

- If the mandible will not respond to closed reduction with just thumb pressure, likely the masticatory muscles are contracting with sufficient force to prohibit the condyles from being sufficiently distracted owing to pain.
  1. Reduction of pain in the temporomandibular joint (TMJ) with local anesthesia.
  2. *Auriculotemporal block of V3:*
     - The auriculotemporal nerve that innervates the TMJ may be anesthetized inferior to the TMJ capsule. It can be accessed through the skin just anterior to the tragus.
     - With the patient's mouth wide open (it already is in this case), a triangular-shaped hollow will be evident inferior and posterior to the mandibular condyle. Insert the needle at a 20-degree anterior inclination, in the horizontal plane, at the level of the inferior border of the tragus of the ear (Fig. 63.2). The bevel of the needle should be anterior.
     - The needle should be inserted behind the (posterior) ramus and approximately 2 cm deep (aiming for the medial aspect of the posterior border of the ramus). If the posterior border of the ramus is contacted, the needle will need to be directed in a more posterior direction. Then deposit 1–2 mL of lidocaine 2 %.
  3. Then, as before, place the thumbs in a bilateral position on the patient's mandibular molars, and depress the mandible to distract the condyles in an inferior direction.
  4. Conscious sedation may be utilized if the reduction procedure has been arduous and stressful for the patient.

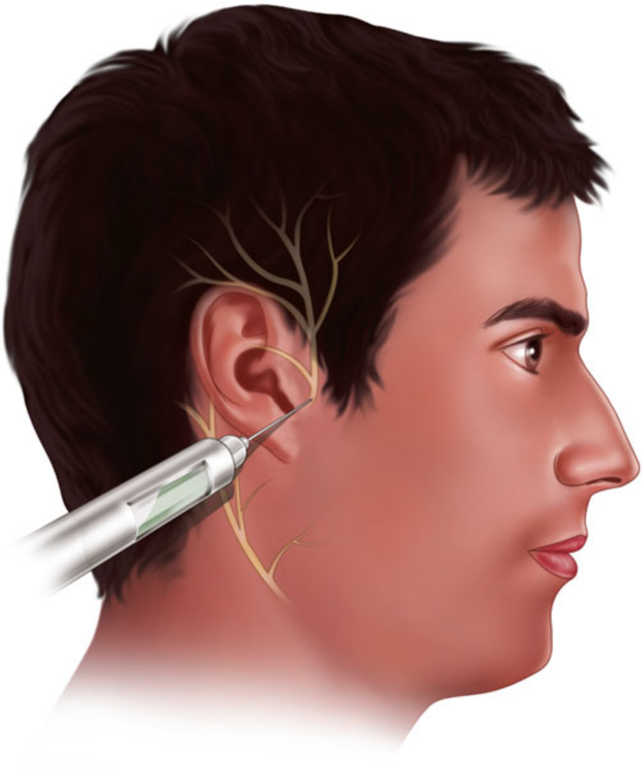

Fig. 63.2 Infiltration of cranial nerve V, the auriculotemporal branch (V3)

63.5 Complications

- Inability to reduce the condyles manually which may lead to more invasive procedures.
- If the condition is acute (≤24 h) and not associated with trauma, there are few if any significant complications or risks for this procedure.

63.6 Pearls

- The manual pressure required on the mandibular molars needs to be sustained and very firm.
- If both condyles are dislocated, it is likely beneficial to attempt one side at a time.
- The clinician needs to protect the thumbs from the impact of the patient's molars during the sudden successful reduction because the biting forces are significant in the molar region.

Selected Reading

Chan TC, Harrigan RA, Ufberg J, Vilke GM. Mandibular reduction. J Emerg Med. 2008;34:435.

Donlon WC, Truta MP, Eversole LR. A modified auriculotemporal nerve block for regional anesthesia of the temporomandibular joint. J Oral Maxillofac Surg. 1984;42:544.

Huang IY, Chen CM, Kao YH, Chen CM, Wu CW. Management of long-standing mandibular dislocation. Int J Oral Maxillofac Surg. 2011;40:810–4.

Prabhakar V, Singla S. Bilateral antersuperior dislocation of the intact mandibular condyles in the temporal fossa. Int J Oral Maxillofac Surg. 2011;40:640–3.

Thagarajah T, Mcculloch N, Thangarajah S, Stocker J. Bilateral temporomandibular joint dislocation in a 29-year-old man: a case report. J Med Case Rep. 2010;4:263.

Dry Socket (Alveolar Osteitis, Fibrinolytic Osteitis)

64

Michael A. Abraham, Amir Azari, Jennifer Westcott, and Franci Stavropoulos

64.1 Indications (Fig. 64.1)

- Definition: severe pain occurring 2–3 days after tooth extraction
- Recent tooth extraction, especially of a mandibular tooth or an impacted third molar
- Partially or completely visible bone socket
- Intense radiating pain (often to the ear)
- Fetid odor without suppuration
- Absence of swelling, lymphadenitis, or bacteremia
- Foreign bodies present in the extraction socket

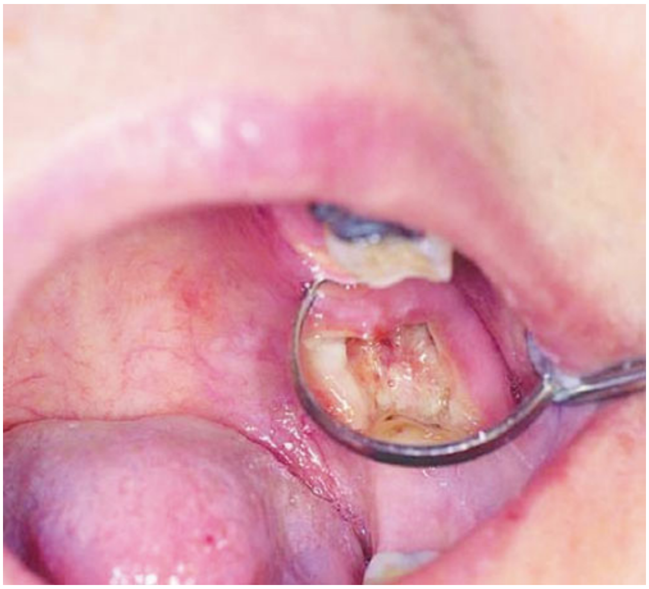

Fig. 64.1 Clinical photograph suggestive of a dry socket. Clinical correlation is necessary

M.A. Abraham, DMD (✉)
United States Air Force, Dental Corps, Minot, ND, USA
e-mail: mikeabrahamdmd@gmail.com

A. Azari, DMD
Department of Oral and Maxillofacial Surgery, Oregon Health and Science University, Portland, OR, USA
e-mail: Azari@OHSU.edu

J. Westcott, DMD
Private Practice, Palm Beach Gardens, FL, USA
e-mail: Jwestcott561@gmail.com

F. Stavropoulos, DDS
Department of Dental Specialties – Oral and Maxillofacial Surgery, Gundersen Health System, La Crosse, WI, USA
e-mail: Mfstavro@gundersenhealth.org

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_64

64.2 Contraindications

- Absolute
  - Osteomyelitis
  - Jaw fracture
- Relative
  - Retained roots
  - Infection

64.3 Materials and Medications

- Warm saline or 0.12 % aqueous chlorhexidine solution
- 12-mL syringe with curved tip (Monoject® syringe)
- 25-gauge needle, syringe
- Local anesthetic, 2 % lidocaine 1:100,000 epinephrine
- Suction with small tip or gauze
- Socket dressing
  - Gelfoam or small gauze strips
- Socket medicament
  - Sultan dry socket paste® (guaiacol, balsam of Peru, eugenol, 1.6 % chlorobutanol), iodoform, or eugenol® (zinc oxide eugenol dental cement)
- Curved forceps

64.4 Procedure

1. Administer local anesthesia as necessary.
2. Remove any sutures closing the extraction site.
3. Irrigate the wound gently with warm saline or 0.12 % aqueous chlorhexidine.
4. Carefully suction or gently dry any excess saline; the socket area should be isolated from saliva by using gauze or cotton rolls.
5. Gently place iodoform-soaked gauze, Gelfoam soaked in eugenol, or Sultan dry socket paste in the extraction socket with forceps/Monoject syringe.
6. Rinse with saline and replace the dressing as needed for the first 2–3 days and every 2–3 days thereafter.
7. Remove the dressing, if it does not dissolve, without replacement once the pain has resolved.
8. Pain medication (nonsteroidal anti-inflammatory drugs [NSAIDS] or narcotics) should be prescribed if necessary.
9. Follow-up with dentist.

64.5 Complications

- Delayed healing
- Wound dehiscence

64.6 Pearls and Pitfalls

- *Pearls*
  - Wound irrigation may be so painful at the first visit that administration of a local anesthetic without a vasoconstrictor should be considered.
  - The patient should experience profound pain relief within minutes of placement of the soaked medicated dressing.
  - If a medicated dressing is necessary for more than 2 weeks, reevaluate for development of osteomyelitis.
  - "Dry socket" is not a progressive disease but may persist for 10–14 days whether treated or not; therapy is palliative.
  - Instruct the patient to avoid the following, which can cause changes of pressure in the mouth:
    - Smoking
    - Using a straw
    - Spitting
    - Drinking carbonated beverages (e.g., soda, seltzer water, beer)
- *Pitfalls*
  - Avoid over manipulating the socket because this will increase the amount of exposed bone and pain.

Selected Reading

Bloomquist D, Hooley J, Whitacre R. A self-instructional guide: surgical complications. 3rd ed. Seattle: Stroma; 1983.

Matocha DL. Postsurgical complications. Emerg Med Clin North Am. 2000;18:549–64.

Roberts G, Scully C, Shotts R. Dental emergencies. West J Med. 2001;175:51.

Postextraction Hemorrhage

65

Michael A. Abraham, Amir Azari, Jennifer Westcott,
and Franci Stavropoulos

65.1 Indications

- Recent tooth extraction site, presenting with more than a slight oozing of blood
- Full evaluation indicating amount of blood loss, present physical condition, and reason for hemorrhage including coagulopathy or medication use

65.2 Contraindications

- Absolute
  - None
- Relative
  - None

65.3 Materials and Medications

- 2×2 gauze pad
- Saline
- 25-gauge needle, syringe
- Local anesthetic without vasoconstrictor—2 % lidocaine plain
- Gelfoam® (absorbable gelatin-compressed sponge) or oxidized cellulose
- Topical thrombin
- Suture kit with 3-0 chromic gut suture or 3-0 Vicryl® suture (synthetic absorbable sterile surgical suture composed of a copolymer made from 90 % glycolide and 10 % L-lactide)
- Hemostat

65.4 Procedure

1. Use suction and saline irrigation to gently rinse the affected area. If a "liver" clot is present, irrigate and remove it with suction.
2. Determine the source of hemorrhage without local anesthesia, if possible, because the use of local anesthetic with an added vasoconstrictor may obscure bleeding sites.
3. Moisten a folded 2×2 gauze pad with saline and place it directly onto the extraction site.
4. Instruct the patient to apply firm biting pressure, and observe for 1 h, changing gauze as necessary.
5. If bleeding persists, an intraoral nerve block should be performed.
   - Blocks are preferred to infiltrations; anesthetic with epinephrine infiltrated near the bleeding site will produce only temporary local hemostasis from vasoconstriction.
6. Gently curette the tooth extraction socket and remove areas of old blood clot or granulation tissue.
7. Check soft tissue for associated arterial bleeding.
   - If hemorrhage is localized to soft tissue, use pressure or tie off vessels.
8. Fold Gelfoam® into a small cylinder to fit into the extraction socket.
9. Place Gelfoam® with topical thrombin or Surgicel® (absorbable hemostat) into the socket and hold in

---

M.A. Abraham, DMD (✉)
United States Air Force, Dental Corps, Minot, ND, USA
e-mail: mikeabrahamdmd@gmail.com

A. Azari, DMD
Department of Oral and Maxillofacial Surgery, Oregon Health and Science University, Portland, OR, USA
e-mail: Azari@OHSU.edu

J. Westcott, DMD
Private Practice, Palm Beach Gardens, FL, USA
e-mail: Jwestcott561@gmail.com

F. Stavropoulos, DDS
Department of Dental Specialties – Oral and Maxillofacial Surgery, Gundersen Health System, La Crosse, WI, USA
e-mail: Mfstavro@gundersenhealth.org

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_65

position with a figure-of-eight stitch using 3-0 chromic gut suture or 3-0 Vicryl suture (Figs. 65.1 and 65.2).

10. Fold 2×2 gauze, moisten it with saline, and place it over the suture.
11. Instruct the patient to bite down with firm pressure for 30 min; repeat as necessary.
12. Follow-up with dentist.

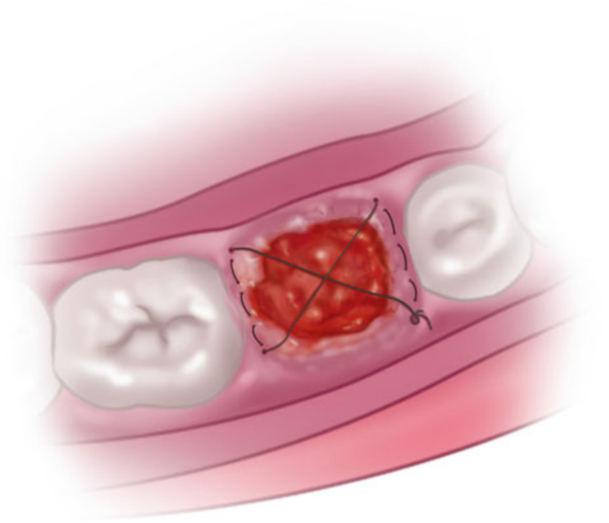

Fig. 65.1 A resorbable figure-of-eight suture placed over an extraction socket

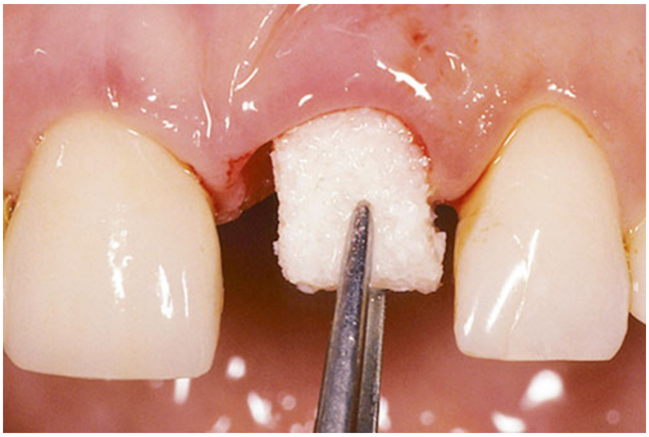

Fig. 65.2 Gelfoam® (absorbable gelatin-compressed sponge) being placed into the extraction socket (Photograph courtesy of Dr. Michael Abraham)

65.5 Complications

- Continued bleeding due to lack of patient compliance.
- Continued bleeding due to coagulopathy or medication use.
- If bleeding persists and coagulopathy is identified, the administration of intravenous blood replacement products may be necessary.

65.6 Pearls and Pitfalls

- Pearls
  - Minor bleeding concerns may be addressed at home by instructing the patient to bite on the affected area with a tea bag for 30 min (tannic acid in tea is a vasoconstrictor).
  - It is normal for an extraction socket to ooze slight amounts of blood for 12–24 h; it is normal for patients to see some blood on their pillow after waking.
  - The patient should be instructed to avoid the following, which can cause changes of pressure in the mouth:
    - Smoking
    - Using a straw
    - Spitting
    - Drinking carbonated beverages (e.g., soda, seltzer water, beer)
- Pitfalls
  - Small amounts of blood mixed with saliva may deceptively appear as large amounts of blood.

Selected Reading

Bloomquist D, Hooley J, Whitacre R. A self-instructional guide: surgical complications. Seattle: Stroma; 1983. pp. 50–5.

Hupp JR, Ellis III E, Tucker MR, editors. Contemporary oral and maxillofacial surgery. 5th ed. St. Louis: Mosby Elsevier; 2008. p. 195–7.

Fractured Tooth

66

Geraldine Weinstein

66.1 Indications (Fig. 66.1)

- Temporary repair of an acute dental fracture until follow-up by a dentist can be secured.

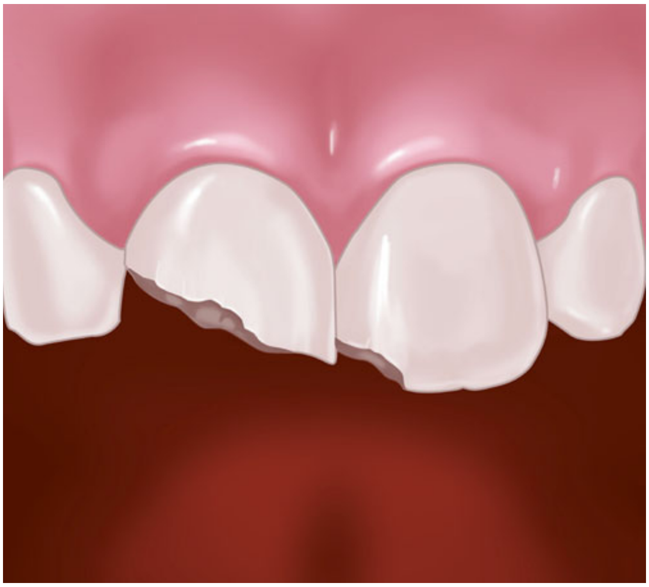

Fig. 66.1 Examples of fractured teeth

66.2 Methods of Sustaining Dental Fracture

- Traumatic injury to the head/facial area
- Falling down
- Extensive tooth decay that has undermined the integrity of the tooth structure
- Biting down on something hard

G. Weinstein, DDS
Department of Restorative Dental Sciences,
University of Florida College of Dentistry,
Gainesville, FL, USA
e-mail: gweinstein@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_66

66.3 Four Types of Fractured Tooth (Fig. 66.2)

- *Type 1*: contained to the enamel of the tooth, asymptomatic, and can be treated easily with a composite filling by a dentist.
- *Type 2*: involves a fracture through the dentin layer of the tooth. The patient may experience some sensitivity to temperature changes and chewing. Depending on the severity, treatment may include a root canal and a restoration by a dentist.
- *Type 3*: involves the pulp of the tooth and will require endodontic treatment by a dentist.
- *Type 4*: a root fracture in the tooth that makes it nonrestorable and requiring extraction. It is diagnosed by means of a periapical radiograph taken in a dental office.

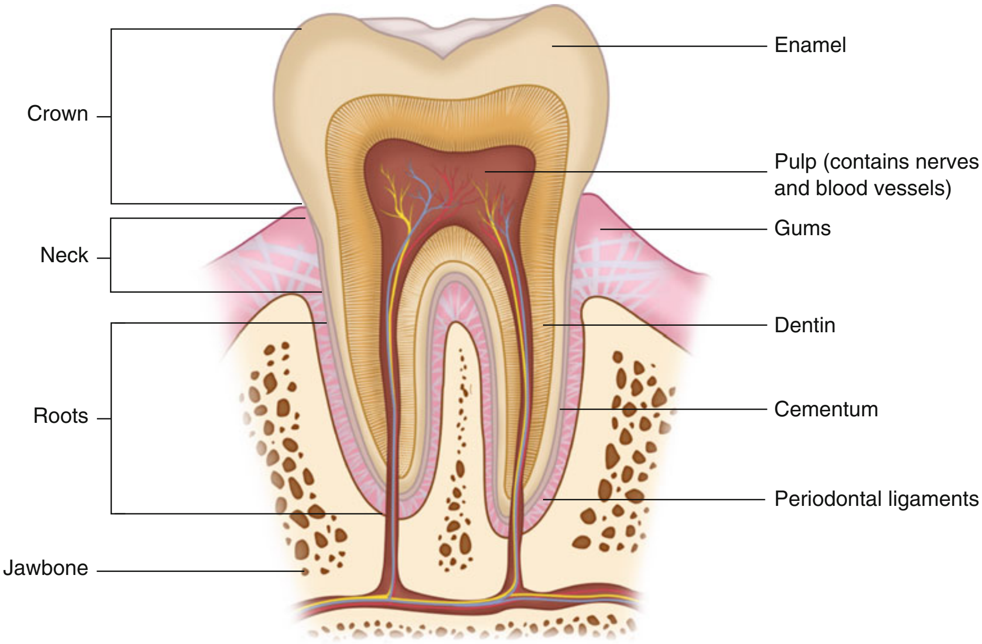
Crown
Neck
Roots
Jawbone
Enamel
Pulp (contains nerves and blood vessels)
Gums
Dentin
Cementum
Periodontal ligaments

Fig. 66.2 Anatomy of a tooth

66.4 Contraindications

- *Relative*
  - Patients at high risk of aspiration owing to intoxication and altered mental status

66.5 Materials and Medications

- Warm saline or 0.12 % aqueous chlorhexidine solution for irrigation of tissues and tooth
- Irrigating syringe
- Suction with a small tip
- Gauze to control hemorrhage
- Resorbable sutures and local anesthetic as needed for soft tissue lacerations
- Temporary tooth restoration material like intermediate restorative material (IRM) and glass ionomer (like Fuji)

66.6 Procedure

1. Have the patient rinse the mouth out with warm water to clean out any debris. Apply an ice pack to the affected cheek to reduce swelling.
2. Anesthetize the area, either locally at the tooth apex or with a nerve block (mental or inferior alveolar).
3. Irrigate the area; assess the fracture intraorally and check for soft tissue lacerations. Use the gauze with finger pressure to control bleeding in the soft tissue or the tooth.
4. Check if the tooth or bony segment is mobile. If so, a referral to a dentist or oral surgeon is necessary as soon as possible for proper assessment.
5. Type II fractures (fracture limited to dentin layer): cover the exposed surface with a temporary dental cement. In a pinch, 2-octyl cyanoacrylate (Dermabond) is an acceptable secondary alternative (Fig. 66.3).
6. Type III (pulp involved)
   - Provide immediate dental follow-up and analgesics.
   - Initiate antibiotic coverage with penicillin or clindamycin.

50% MORE
Than Other Brands
12+
REPAIRS
net wt. 2.64 g (0.09 oz)
Repair
Lost Fillings & Loose Caps
DenTek
Temparin Max
One Step
Formula
MAXIMUM
STRENGTH
Dentist
Recommended
Instant Pain Relief
Easy To Apply
NEW!
Applicator
Tip
Spade Tool
DenTek

Fig. 66.3 Example of temporary dental cement (Reproduced with permission from DenTek Oral Care, Inc.)

66.7 Complications

- Loss of a tooth
- Infection or abscess
- Aspiration of a segment or a whole tooth
- Cosmetic deformity

66.8 Pearls

- Be certain to perform a thorough intraoral examination, looking for tooth fragments or lacerations that may be hiding fragments.
- Dental blocks are very useful for pain control.
- If a tooth is not mobile and the *pulp is exposed*, immediate referral (within a few hours) to a dentist is necessary for extraction or endodontic (root canal) treatment of the tooth. Placement of a temporary-type restoration on this tooth is *not* recommended at this time because it may exacerbate symptoms. Prescribe pain medication and possibly antibiotics when the tooth's pulp is exposed and the patient is unable to see the dentist within 24 h.
- If the tooth is not mobile and *the pulp is not exposed*, a temporary restoration can be placed on the tooth and the patient referred to a dentist for treatment. The fractured part of the tooth should be saved in the event that it can be used. If temporary tooth restoration is unavailable in the emergency department, advise the patient that it is readily available at local pharmacies.
- ALL DENTAL FRACTURES, EXCEPT TYPE I, REQUIRE DENTAL FOLLOW-UP WITHIN 24 h.

67

Dental Avulsion Management

Laura Tucker and Abimbola O. Adewumi

67.1 Indications

- The tooth is completely displaced *out* of its socket, leading to severance of the neurovascular pulp supply and separation of the periodontal ligament (Fig. 67.1).
- Diagnosis
  - Clinically, the socket is found empty or filled with coagulum.
  - Imaging (occlusal, periapical, and lateral views of the affected tooth and surrounding area) (Fig. 67.2):
- Confirm vacuous socket.
- Ensure that the missing tooth is not intruded.
- Diagnose root fracture or alveolar fracture.

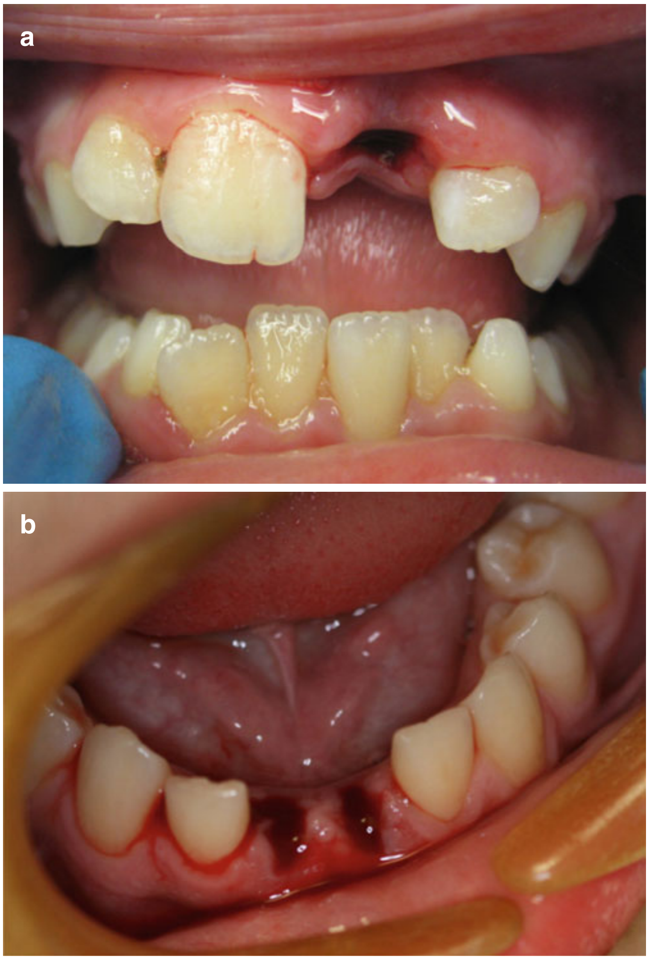

Fig. 67.1 (**a**, **b**) Empty socket following traumatic dental avulsion

---

L. Tucker, DDS • A.O. Adewumi, BDS, FDSR (✉)
Department of Pediatric Dentistry, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: lauragracesullivan@gmail.com; aadewumi@dental.ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_67

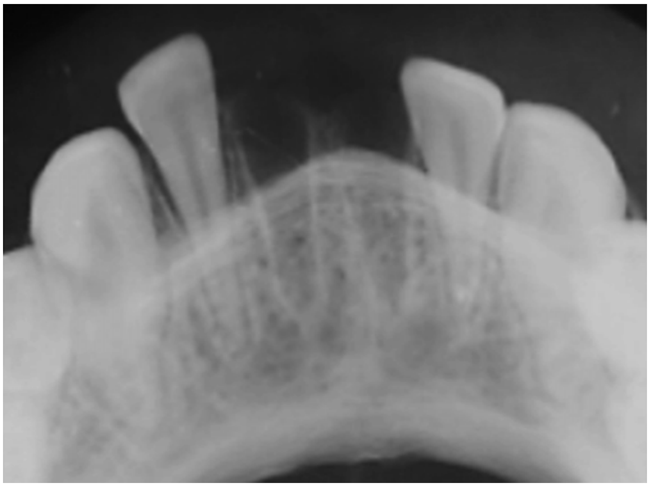

Fig. 67.2 Lower occlusal radiograph shows a complete avulsion of the mandibular right and left central incisors

67.2 Contraindications

- Absolute
  - – Avulsed tooth is a primary tooth.
    - DO NOT REPLANT OR REPOSITION AVULSED PRIMARY TEETH.
    - Replantation of primary teeth increases the potential for damage to the developing permanent tooth owing to the increased frequency of pulpal necrosis.
- Relative
  - – Fractured root (further intervention required before replantation)
  - – Alveolar fracture (further intervention required before replantation)
  - – Prolonged extraoral dry time and out of compatible solution (>1 h)
  - – Immunocompromised host or congenital heart disease
  - – Severe seizure disorder that may put tooth at risk for coming out while the airway is compromised
  - – Patient with potential to lose airway reflexes

67.3 Materials and Medications

- Hank's Balanced Salt (Save-A-Tooth) solution or milk in which to preserve tooth until it can be replanted
  - – Normal saline if neither of those is available
- Normal saline rinse
- 20- or 18-gauge cannula for gentle irrigation of the socket
- Absorbable suture for gingival lacerations, if present
- Flexible splint materials
  - – Round dental wire
  - – A flat pliable metal long enough to cover the affected tooth and the two teeth on other side (e.g., the metal nasal bridge from a respirator mask)
- Fixative
  - – Dental adhesive
  - – Dermabond or some other brand of cyanoacrylate for adhesive skin closures

67.4 Procedure

1. Be certain the tooth is a permanent one and not a primary tooth.
2. If not done by the patient, gently wash the tooth under water for approximately 10 s.
    - Be certain to hold the tooth by the crown, not the root (Fig. 67.3)
3. If the tooth cannot be replanted immediately, place the tooth in Hank's solution or milk.
4. If no such media is available, instruct the patient to hold the tooth inside his or her mouth between the cheek and the gums.
5. Gently replant the tooth, using digital pressure only into as anatomical a position as possible (Fig. 67.4).
    - Assess clinically for alignment.
    - Radiograph for confirmation.
6. Suture any gingival lacerations if present.
7. Apply a flexible splint, securing the affected tooth to the teeth on either side.
    - Consider using skin adhesive both to secure the tooth to its neighbors and, perhaps, to apply a makeshift splint until the patient can be seen by her or his dentist.
        – Towel dry the teeth as best as possible.
        – Apply skin adhesive (using the standard applicator) to the lateral edges of the avulsed tooth where it will make contact with its adjacent teeth.
        – If dental wire is available, apply adhesive to the buccal surfaces of the three teeth (the avulsed tooth central to the other two) and apply length of metal to the Fixodent (Fig. 67.5).
8. Systemic antibiotics with anaerobic coverage is empirical.
9. Ascertain tetanus status for the patient; update if uncertain.

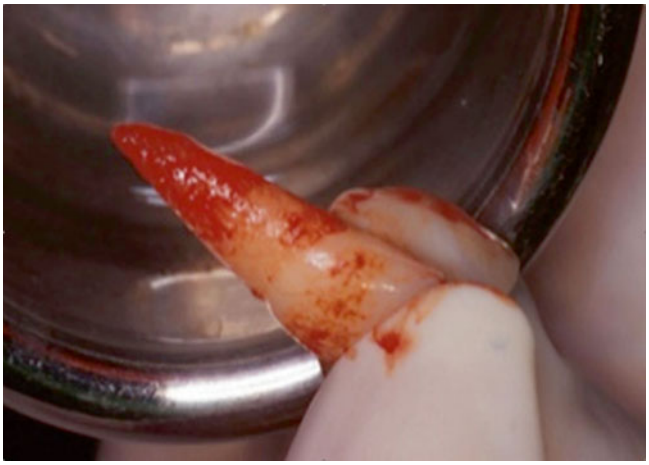

Fig. 67.3 Proper way to hold an avulsed tooth

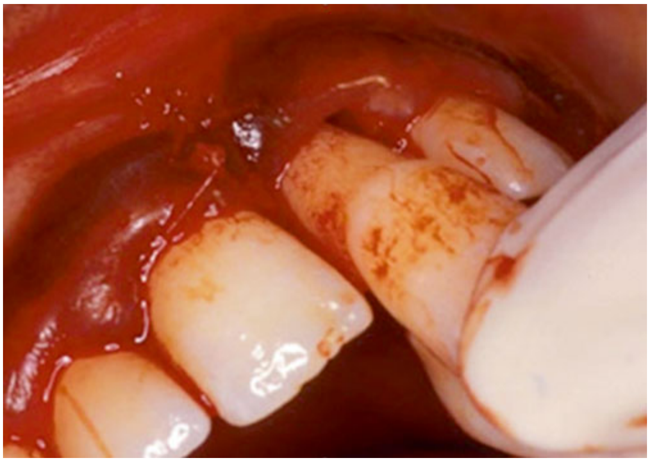

Fig. 67.4 Gentle replantation using digital pressure

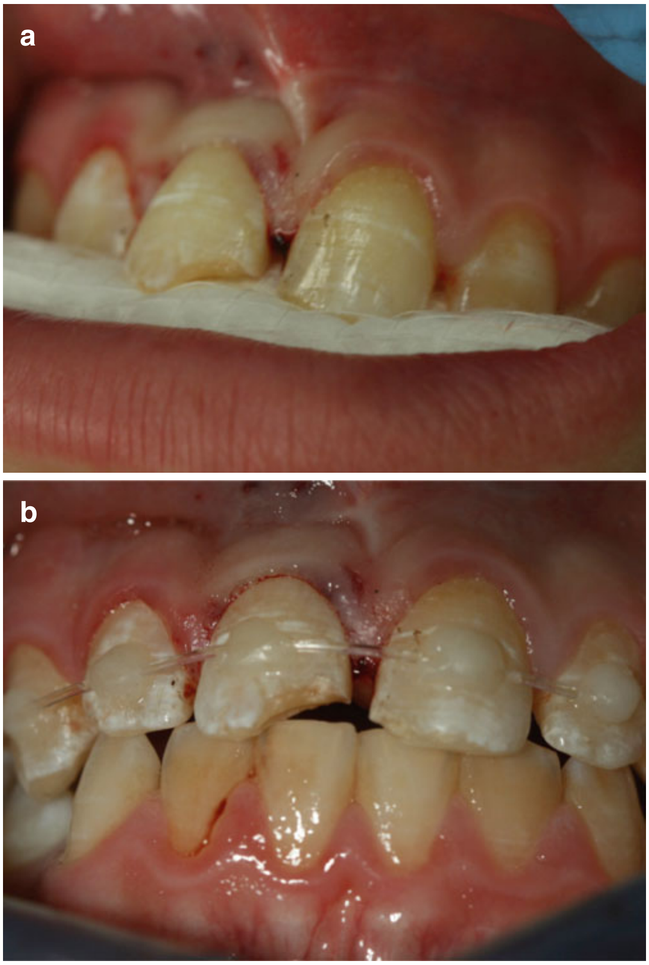

Fig. 67.5 (**a**, **b**) Splint stabilization

67.5 Complications

- Primary teeth
  - Dilaceration (bend) in the permanent tooth crown
  - Enamel defect of the lower permanent incisors as a result of avulsion of the preceding primary tooth
- Permanent teeth
  - Discoloration as a result of loss of vitality of the avulsed tooth.
  - Ankylosis of the alveolar ridge, leading to functional and aesthetic changes.
  - Replacement resorption occurs when the replanted tooth is slowly replaced with bone.
  - External inflammatory resorption is a progressive loss of tooth associated with destruction of adjacent alveolar bone.
  - Infection
  - Aspiration of an inadequately secured replanted tooth.

67.6 Pearls and Pitfalls

- Pearls
  - At the initial examination, make sure that all avulsed teeth are accounted for.
    - If not, a radiographic examination is necessary to ensure that the missing tooth is not completely intruded (pushed into the gum) or has sustained a root fracture with loss of the coronal fragment.
  - In children, always consider the likelihood of nonaccidental trauma (abuse).
  - Short-term and long-term dental follow-up cannot be emphasized enough.
- Pitfalls
  - If the avulsed tooth cannot be accounted for, aspiration is a possibility.
- Prognosis
  - Depends on extraoral dry time (length of time the tooth has been out of the mouth and not stored in an appropriate medium):
    - Ideally, tooth should be implanted within 5 min.
    - Extraoral dry time greater than 60 min has a poor prognosis for periodontal healing.
  - Depends on stage of root development of the avulsed tooth (Fig. 67.6)
- The more advanced the root development, the lower the probability of pulp healing and survival.

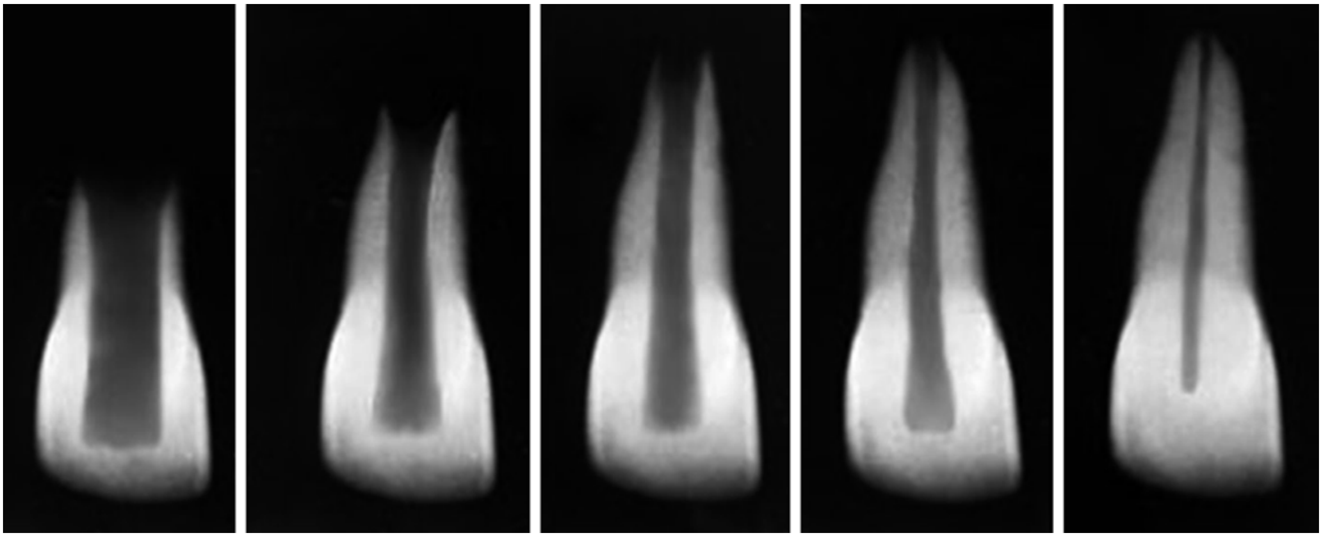

Fig. 67.6 The various stages of root development, from the less advanced (*open apex, left*) to the more advanced (*closed apex, right*)

Selected Reading

AAPD Council on Clinical Affairs. Guideline on management of acute dental trauma. AAPD reference manual. Chicago: American Academy of Pediatric Dentistry; 2010–2011. pp. 202–12.

Andreasen FM, Andreasen JO. Avulsions. In: Andreasen JO, Andreasen FM, Andersson L, editors. Textbook and color atlas of traumatic injuries to the teeth. 4th ed. Oxford: Blackwell; 2007. p. 444–88.

Andreasen JO, Jensen SS, Sae-Lim V. The role of antibiotics in preventing healing complications after traumatic dental injuries: a literature review. Endod Topic. 2006;14:80–92.

Finucane D, Kinirons MJ. External inflammatory and replacement resorption of luxated, and avulsed replanted permanent incisors: a review and case presentation. Dent Traumatol. 2003;19:170–4.

Flores MT, Andersson L, Andreasen JO, et al. Guidelines for the management of traumatic dental injuries. II. Avulsion of permanent teeth. Dent Traumatol. 2007;23:130–6.

Hile LM, Linklater DR. Use of 2-octyl cyanoacrylate for the repair of a fractured tooth. Ann Emerg Med. 2006;47:424–6.

The dental trauma guide. Available at: www.dentaltraumaguide.org.

Part X

Gastrointestinal Procedures

Excision of Thrombosed External Hemorrhoid

68

Latha Ganti

68.1 Indications

- Acute pain within 72 h of thrombosis onset
- The thrombosis will be visible as a bluish-purplish painful mass in perianal area (Fig. 68.1).

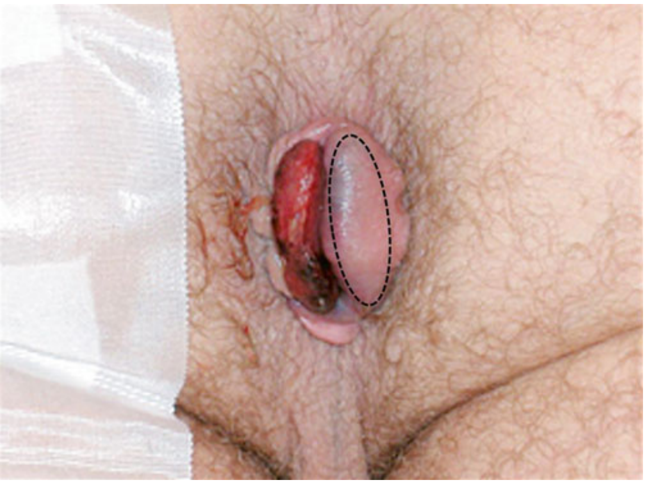

Fig. 68.1 Bluish-purplish appearance of an external thrombosed hemorrhoid. The *ellipse* denotes the area of the elliptical incision to be made (Reproduced with permission from: Fargo and Latimer [3])

68.2 Contraindications

- Absolute
  - Severe coagulopathy
  - Hemodynamic instability
  - Concurrent internal hemorrhoid with rectal prolapse
  - Painless rectal mass (external hemorrhoids are almost always painful, so a painless mass is not a thrombosed external hemorrhoid; also, the whole point of doing this procedure is to provide pain relief).
- Relative
  - Allergy to local anesthetics
  - Perianal infection
  - Inflammatory bowel disease
  - Serious systemic illness or comorbidity that could increase risk of procedure

68.3 Materials and Medications

- Sterile gloves and drape
- Alcohol swabs or pads
- 10 % povidone-iodine or chlorhexidine prep
- 2 % lidocaine with epinephrine
- 5 ml syringe with 25 or 27 gauge and 18 gauge needles
- #11 or #15 scalpel blade and handle
- Direct light source
- Forceps
- Iris scissors
- 4×4 gauze pads
- Adhesive tape
- 3-0 absorbable suture
- 1/4 inch iodoform packing
- Silver nitrate sticks
- Sterile dressing

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_68

68.4 Procedure

1. Patient should be in either prone, left lateral decubitus, or jackknife position (Fig. 68.2).
2. Place 2 vertically oriented pieces of tape down each buttock from lower back to upper thigh. Next, place a perpendicular (horizontal) strip spreading buttocks to either side, securing gurney (Fig. 68.3).
3. Place sterile drape over field and center direct lighting over field (Fig. 68.4).
4. Wipe area with alcohol.
5. Inject 1–2 cc of anesthetic into base of hemorrhoid.
6. Clean area with povidone-iodine or chlorhexidine prep.
7. Make an elliptical incision in the roof of the hemorrhoid, being careful to avoid anal sphincter muscle.
8. Remove blood clot(s); multiple clots are often present.
9. If profuse bleeding is a problem, consider cauterization with silver nitrate sticks.
10. Wound can be closed with figure-of-8 absorbable suture OR can be loosely packed with1/4 inch iodoform gauze if not suturing (do not suture wound closed with packing inside).
11. Cover wound with 4×4 gauze folded in half and taped into place (Fig. 68.5).

- Discharge medicines:
  - Antibiotics generally not necessary.
  - Prescribe ibuprofen and/or acetaminophen for analgesia. Avoid opiates, which are constipating.
  - Prescribe stool softeners, to be taken two to three times daily.
- Discharge instructions to patient:
  - Sitz baths 3–4 times daily, for 20 min, warm not hot water.
  - Packing should fall out spontaneously in 2 days.
  - Keep well hydrated.
  - Use gauze to protect underclothing from soilage/blood stains.
  - Return to ED if pain persists beyond 48 h.

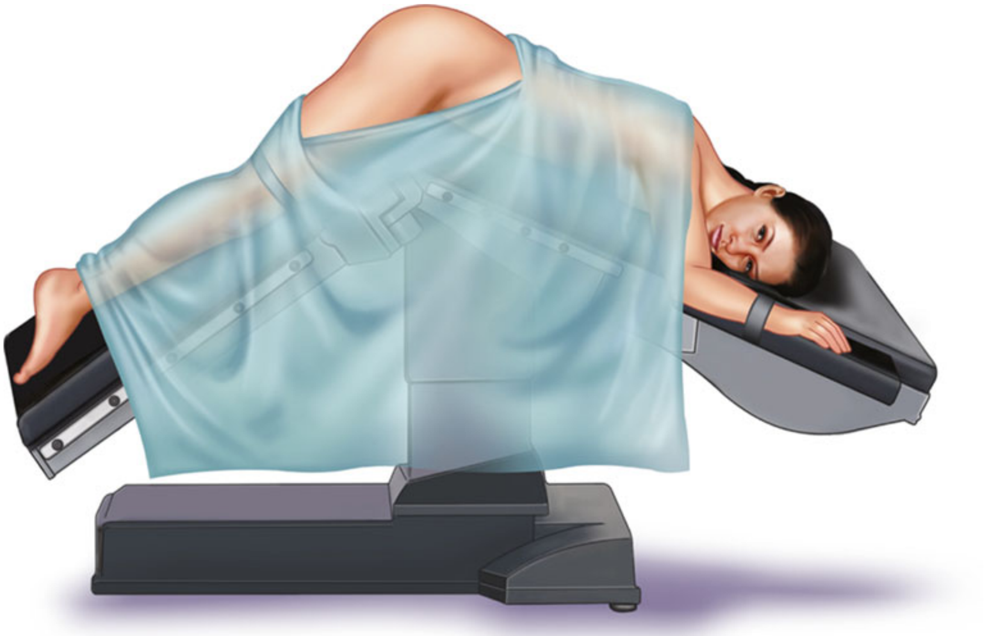

Fig. 68.2 Jackknife position

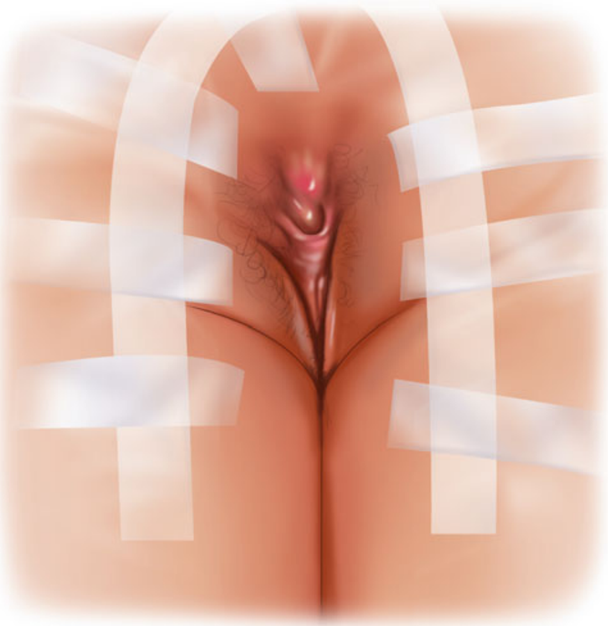

Fig. 68.3 Taping of buttocks to maximize visualization of hemorrhoid

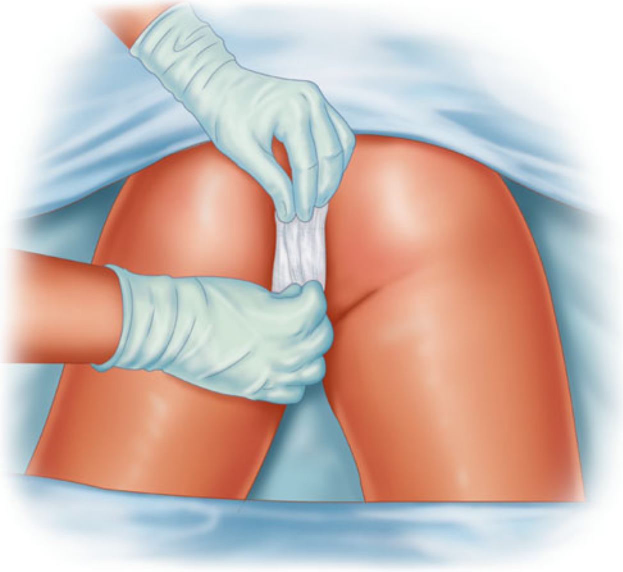

Fig. 68.5 Dress wound with sterile gauze

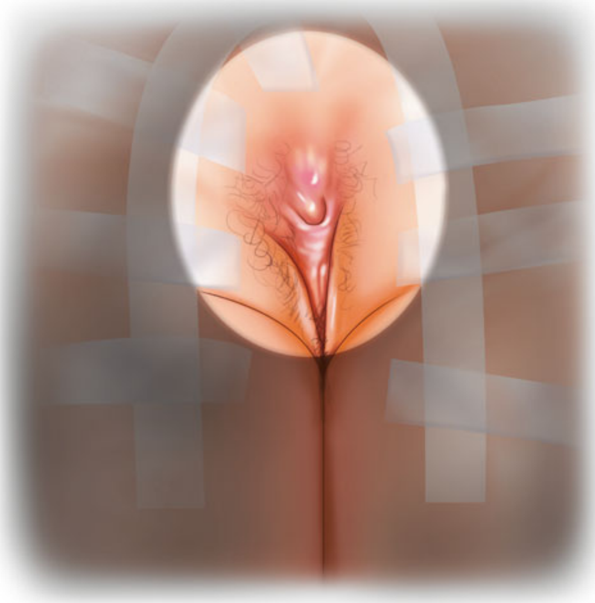

Fig. 68.4 Direct lighting over field

68.5 Complications

- Common
  - Bleeding: usually self-limited. Can apply cautery or figure-of-8 suture if not previously done
  - Pain: usually controlled with ibuprofen or acetaminophen
  - Perianal skin tag: benign
- Rare
  - Infection rate is 5 % [1].
  - Recurrence rate is 5–19 %, vs. 30 % for simple lancing [2].
  - Stricture and/or incontinence: prevented by avoiding underlying external anal sphincter muscle

68.6 Pearls and Pitfalls

- Pearls
  - Elliptical excision of the hemorrhoid results in much lower recurrence rate than simple lancing.
  - Risk factors for thrombosed external hemorrhoids include constipation, 2nd or 3rd trimester pregnancy, and traumatic vaginal delivery.
- Pitfalls
  - Excision of multiple hemorrhoids in circumferential fashion on all sides of the anal canal can cause anal stenosis.
  - Excision of a painless mass: if it is *painless*, it is *not* a thrombosed external hemorrhoid.

References

1. Lorber BW. Thrombosed external hemorrhoid excision. Medscape.com. www.emedicine.medscape.com/article/81039. Accessed 27 July 2014.
2. Rivadeneira DE. Outpatient and surgical procedures for hemorrhoids. UpToDate.com. http://www.uptodate.com/contents/outpatient-and-surgical-procedures-for-hemorrhoids. Accessed 27 July 2014.
3. Fargo MV, Latimer KM. Evaluation and management of common anorectal conditions. Am Fam Physician. 2012;85(6):624–30.

Selected Reading

Fargo MV, Latimer KM. Evaluation and management of common anorectal conditions. Am Fam Physician. 2012;85(6):624–30.

Jongen J, Bach S, Stübinger SH, Bock JU. Excision of thrombosed external hemorrhoid under local anesthesia: a retrospective evaluation of 340 patients. Dis Colon Rectum. 2003;46(9):1226–31.

Diagnostic Peritoneal Lavage

69

Latha Ganti

69.1 Indications (Table 69.1)

- Inability to perform FAST exam due to lack of equipment or operator
- Hemodynamically unstable patient in whom FAST exam is negative or equivocal

Table 69.1 Comparison parameters for DPL, FAST, and CT [1]

	DPL	FAST	CT
Speed	10–15 min	Fastest: <5 min	Variable
Repeatable	Yes, but rarely done	Yes, and frequently done	Yes, but not done often
Cost	$	$$	$$$
Invasive	Yes	No	No
Mobile	Yes	Yes	No
Advantages	Most sensitive for mesenteric and hollow viscus injuries	Highest specificity	Highly accurate but can be hampered by patient movement
Disadvantages	Misses retroperitoneal and diaphragm injuries	Hampered by subcutaneous or intra-abdominal air, obesity, and pelvic fractures Significant false-negative rate	Misses diaphragm, small bowel, and pancreatic injuries Small but significant risk of radiation-associated malignancy Cannot be done at bedside

---

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_69

69.2 Contraindications

- Absolute
  - Indication for laparotomy already exists.
- Relative
  - 2nd or 3rd trimester pregnancy
  - Previous lower abdominal surgery
  - Inexperienced operator
  - Abdominal wall infection
  - Coagulopathy
  - Cirrhosis
  - Morbid obesity

69.3 Materials and Medications

- 10 % povidone iodine prep
- 1 % lidocaine with epinephrine
- Fenestrated drape
- #10 scalpel blade and scalpel holder
- Skin retractors
- Hemostats
- Diagnostic peritoneal lavage (DPL) catheter (standard peritoneal dialysis catheter)
- 10 cc syringe
- Warmed lactated Ringer's or normal saline solution
- Skin stapler
- Simple suture tray with suture material

69.4 Procedure

69.4.1 Patient Preparation

- Place patient in supine position.
- Ensure nasogastric and urethral catheter (Foley) are in place.
- Prep and drape the area from the umbilicus to the symphysis pubis.
- Anesthetize the skin using 1 % lidocaine with epinephrine in the midline where incision will be made (Fig. 69.1).

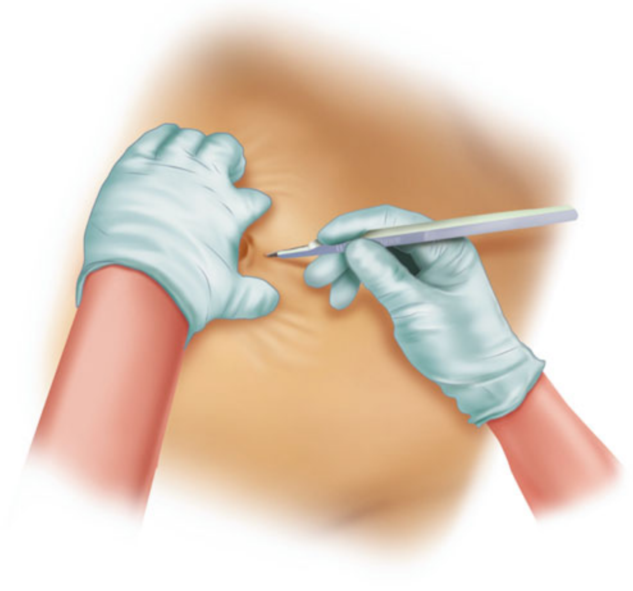

Fig. 69.1 Anesthetize skin where incision will be made (supraumbilical incision shown here, can also make infraumbilical incision)

69.4.2 Three DPL Techniques

- Semiopen (Seldinger) technique
  1. Using a #10 scalpel blade, make a 2 cm incision either superior or inferior to the umbilicus.
  2. Dissect subcutaneous fat until linea alba is exposed.
  3. Hold incision open with skin retractors (Fig. 69.2).
  4. Grasp fascia with hemostats on either side of midline.
  5. Insert 18 gauge needle at 45° angle toward pelvis (Fig. 69.3).
  6. First "pop" will be heard once fascia is penetrated.
  7. Second "pop" will be heard once peritoneum is traversed.
  8. Pass guidewire through needle into pelvis (should pass easily without resistance).
  9. Remove needle while keeping wire stable.
  10. Pass dilator over the wire through fascia, and remove (Fig. 69.4).
  11. Slip DPL catheter over guidewire aiming toward pelvis.
  12. Aspirate peritoneal contents with syringe; aspiration of blood is a positive DPL and means immediate laparotomy (can stop DPL procedure here).
  13. If no blood is immediately obvious, then connect the DPL catheter to a liter of warmed lactated Ringer's (LR) or normal saline (NS) solution for lavage (ensure setup has no one-way valves as solution and peritoneal fluid need to be able to freely mix).
  14. Place LR or NS bag on floor once it is almost empty (minimum 300–350 ml for adults or 10–15 ml/kg for children) and allow intra-abdominal fluid to return (Fig. 69.5).
  15. Send fluid for analysis (Table 69.2).
  16. Irrigate wound, and close skin only with staples or sutures.
- Open technique
  1. Make a 5 cm incision inferior to the umbilicus over linea alba and directly visualize peritoneal cavity.
  2. Both fascia (absorbable suture) and skin (nonabsorbable suture) need to be closed.
- Closed technique
  1. Access peritoneal cavity via percutaneous needle access.
  2. No surgical closure required.

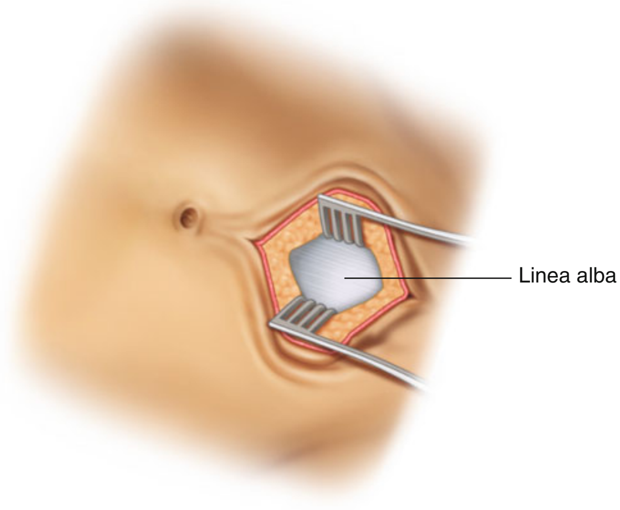

Fig. 69.2 Hold incision open with skin retractors

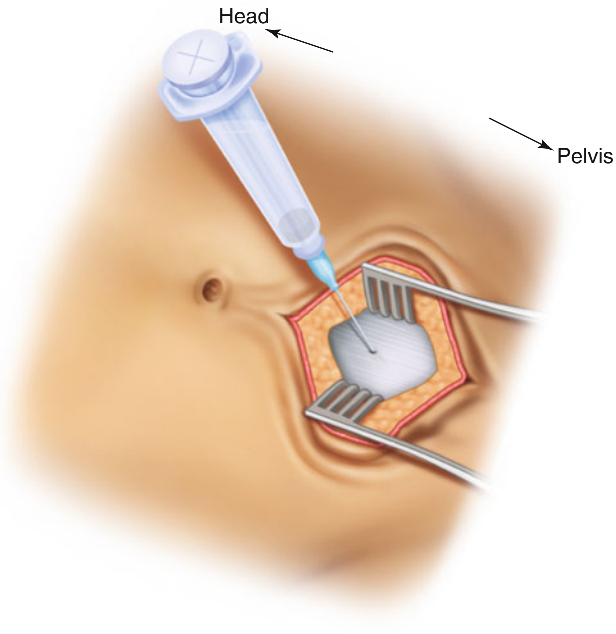

Fig. 69.3 Insert 18 gauge needle at 45° angle toward the pelvis

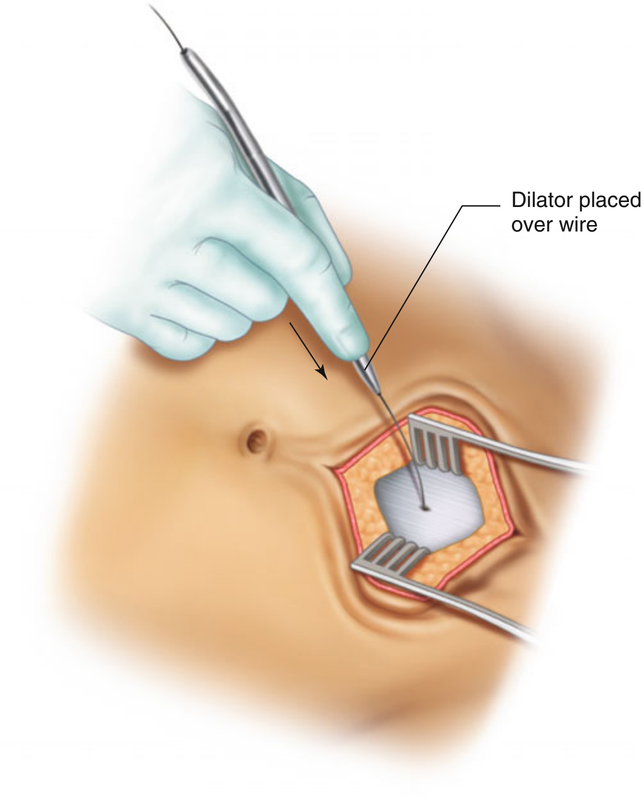

Fig. 69.4 Pass dilator over the wire through fascia and remove

Table 69.2 Diagnostic peritoneal lavage red blood cell criteria (per mm^3) [2]

	Positive	Indeterminate
Immediate gross return of blood via catheter	Any amount	
Immediate return of food particles/intestinal contents	Any amount	
Aspiration of blood	10 cc	
RBC in blunt trauma	100,000	20,000–100,000
RBC in penetrating trauma	10,000	5000–10,000
RBC in gunshot wound	5000	1000–5000
Amylase level (IU/L)	≥175	
Alkaline phosphatase level (IU/L)	≥3	
WBCs (per mm^3)	>500	250–500

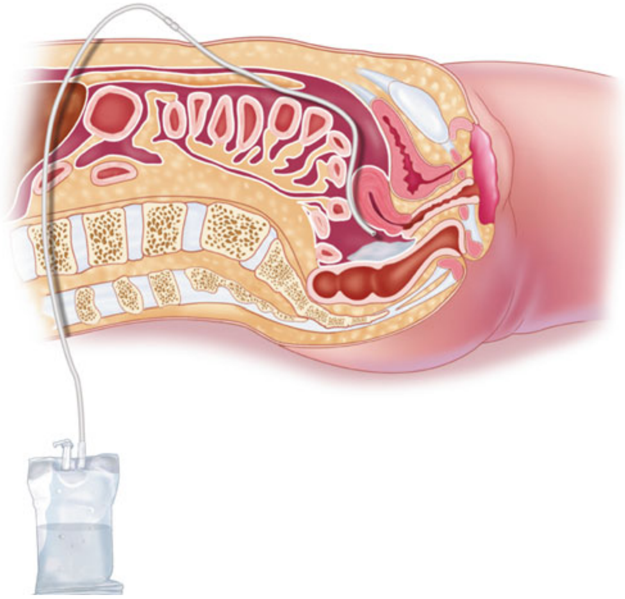

Fig. 69.5 Lavage

69.5 Complications

- Wound infection or dehiscence
- Intraperitoneal injury to organs or vessels (iatrogenic hemoperitoneum)
- Unnecessary laparotomy due to false-positive result from bleeding within rectus sheath or from site of incision
- Potential failure to recover lavage fluid due to:
  - Inadvertent placement of the catheter into the preperitoneal space
  - Compartmentalization of fluid by adhesions
  - Obstruction of fluid outflow (e.g., by omentum)
  - Fluid pooling in the intrathoracic cavity due to diaphragmatic injury
- Incisional hernia

69.6 Pearls and Pitfalls

- Pearls
  - When properly done, complication rate for DPL is low.
  - Prophylactic antibiotics are generally not indicated.
- Pitfalls
  - Inadequate decompression of stomach and urinary bladder increases the chance of injury to these organs; thus, nasogastric and Foley decompression is an important step in patient preparation.

References

1. Jagminas L. Diagnostic peritoneal lavage. Medscape.com. http://emedicine.medscape.com/article/82888-overview#a17. Accessed 28 Aug 2014.
2. Marx JA. Diagnostic peritoneal lavage. In: Ivatury RR, Cayten CG, editors. The textbook of penetrating trauma. Baltimore: Williams & Wilkins; 1996. p. 337.

Selected Reading

Whitehouse JS, Weigelt JA. Diagnostic peritoneal lavage: a review of indications, technique, and interpretation. Scand J Trauma Resusc Emerg Med. 2009;17:13.

Manual Reduction of Abdominal Hernia

70

Latha Ganti

An abdominal wall hernia is a protrusion of the intestine through an opening or area of weakness in the abdominal wall. See Table 70.1 for types of abdominal hernias and Fig. 70.1 for locations along the abdominal anatomy.

Table 70.1 Types of abdominal hernias

Type	Defect	Most commonly seen in	Notes
Inguinal	Intestine or bladder protrudes through abdominal wall or into inguinal canal in the groin	Men because of a natural weakness in this area	96 % of all groin hernias are inguinal; 4 % are femoral
Femoral	Intestine enters canal carrying femoral artery into the upper thigh	Women, especially those who are pregnant or obese	
Incisional	Intestine pushes through abdominal wall at the site of previous abdominal surgery	Elderly or overweight people who are inactive after abdominal surgery	
Umbilical	Part of the small intestine passes through abdominal wall near the navel	Newborns and obese women or those who have had many children	In children, not repaired until age five because often resolve on their own
Hiatal	Upper stomach squeezes through hiatus, an opening in the diaphragm through which the esophagus passes		

L. Ganti, MD, MS, MBA
Professor of Emergency Medicine, University of Central Florida, Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_70

Fig. 70.1 Types of abdominal wall hernias

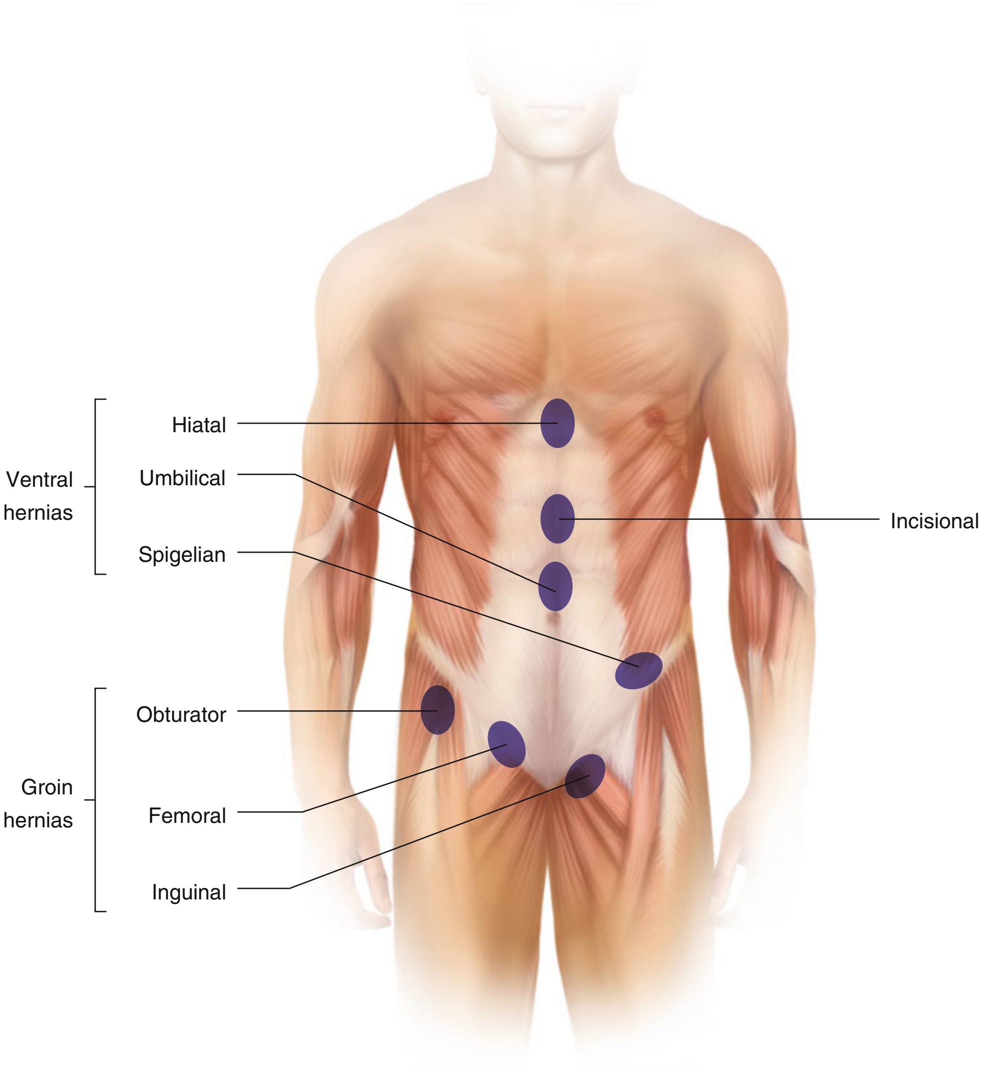
Ventral hernias
Hiatal
Umbilical
Spigelian
Incisional
Groin hernias
Obturator
Femoral
Inguinal

70.1 Indications

- Incarcerated hernia

70.2 Contraindications

- Absolute
  - Strangulated hernia (could result in placing dead bowel into abdominal cavity)
- Relative
  - Inability to get patient relaxed enough
  - Previous unsuccessful attempts

70.3 Materials and Medications

- Ice or cold compress
- Stretcher or gurney that can tilt to provide Trendelenburg position
- Moderate sedation drugs if providing moderate sedation
- Truss for post-procedure

70.4 Procedure

1. Patient positioning:
   - For abdominal hernia: place patient supine.
   - For groin hernia in adult: place in 20 ° of Trendelenburg.
   - For groin hernias in children: place in unilateral frog leg position (Fig. 70.2).
2. Apply ice or cold compress directly over hernia site to reduce swelling.
3. Administer opiate analgesia or moderate/procedural sedation.
4. Wait up to 30 min as hernia may reduce spontaneously after swelling has gone down and patient is relaxed.
5. Gently apply steady pressure distally on the tissue at the neck of the hernia with one hand and with other hand, guide hernia proximally through fascial defect. Too much pressure distally can cause hernia to balloon further, making manual reduction difficult. Takes up to 15–20 min. Do not rush procedure.
6. Once hernia is reduced, pain will improve.
7. An external support garment or truss (Fig. 70.3) can be helpful to hold reduced hernia in place and serve as a temporizing measure until surgical repair can be done.
8. Advise patient to schedule elective surgical repair.
9. If unable to reduce the hernia, obtain surgical consultation. Do not force repeated attempts.

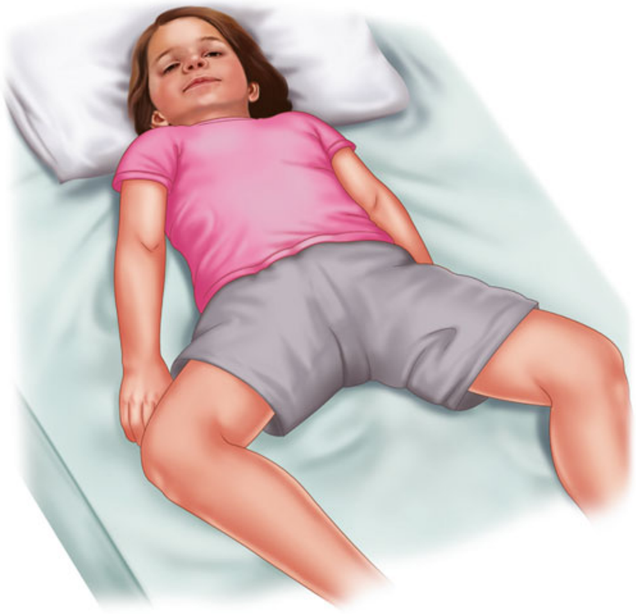

Fig. 70.2 Frog leg position in child

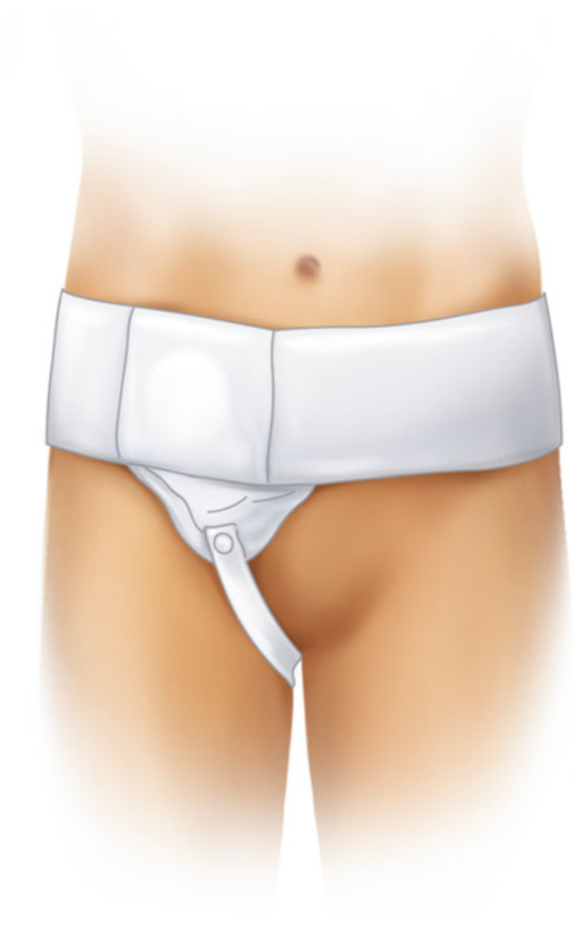

Fig. 70.3 An example of a truss, or external support, that can be useful as a temporizing measure until definitive hernia repair can be done

70.5 Complications

- Pain.
- Inability to achieve manual reduction, leading to strangulation of the hernia.
- Strangulation can result in peritonitis and sepsis.
- Recurrence.
- Hydrocele.

70.6 Pearls and Pitfalls

- Pearls
  - Definitive treatment for a hernia is surgery (herniorrhaphy). Without surgery, hernias grow larger over time; they do not disappear. Small hernias are easier to fix and result in fewer complications.
  - The only hernia that can resolve on its own is an umbilical hernia in a child.
  - Trusses, bandages, and tape may provide some comfort but do not reduce risk of incarceration or strangulation.
  - Note that if a truss is worn, it should be in place after reduction of the hernia. Also, it can be impractical in hot climates.
- Pitfalls
  - When the constricting neck and the protrusion are both reduced into the abdomen together (known as a reduction *en masse*), without actually reducing the hernia itself, strangulation ensues even though it appears one has reduced the hernia.
  - If there is still considerable pain after the reduction, it is likely the reduction was not successful or that dead bowel has been reduced into the abdominal cavity.
  - Not recognizing strangulation leads to gangrenous bowel, peritonitis, and sepsis.

Selected Reading

Campanelli G, Canziani M, Frattini F, et al. Inguinal hernia: state of the art. Int J Surg. 2008;6 Suppl 1:S26–8.

Jenkins JT, O'Dwyer PJ. Inguinal hernias. BMJ. 2008;336(7638):269–72.

Moses S. Hernia reduction. 2014. http://www.fpnotebook.com/mobile/Surgery/GI/HrnRdctn.htm. Accessed Sept 15, 2014.

Extended Focused Assessment with Sonography for Trauma

71

Coben Thorn and L. Connor Nickels

71.1 Indications

- Blunt abdominal or chest trauma
- Penetrating abdominal or chest trauma
- Undifferentiated hypotension
- The “E” in EFAST refers to the “extended” ability to detect lung pathology such as a pneumothorax or hemothorax during the otherwise standard trauma FAST exam using the same equipment with or without an additional transducer probe.
- Specific findings that can be detected on extended focused assessment with sonography for trauma (EFAST):
  - Pericardial fluid
  - Pleural fluid
  - Free intraperitoneal fluid
  - Pneumothorax
- Free fluid appears as anechoic or black.

71.2 Contraindications

- Need for immediate operative intervention

71.3 Materials and Medications

- Ultrasound machine
- Probe(s): phased array probe (5 to 1 MHz) or curved array probe (5 to 2 MHz)
  - Phased array has a smaller footprint, allowing easier access between intercostal spaces (Fig. 71.1); however, curved array provides better resolution of images

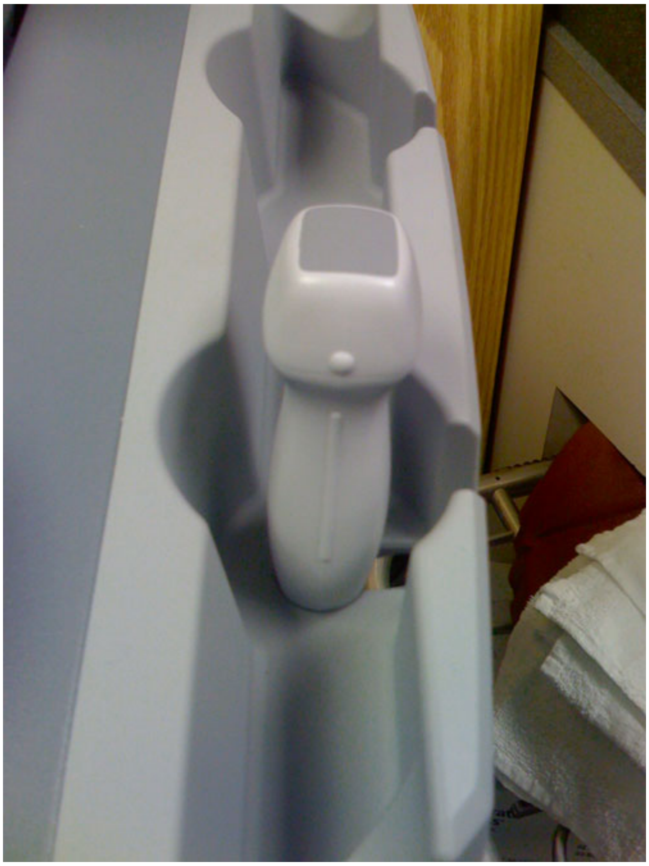

Fig. 71.1 Phased array transducer (P17) with a small footprint that is helpful to fit between the ribs and can be used for focused assessment with sonography for trauma (FAST) examination

---

C. Thorn, MD
Department of Emergency Medicine, Bon Secours St. Francis Health System, Greenville, SC, USA
e-mail: cobenthorn@gmail.com

L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_71

(Fig. 71.2). The linear array transducer (L38, 10–5 MHz) is good for lung images.

- Gel
- Skilled ultrasound operator
- ± Laboratory work, cardiac monitor, and two large-bore intravenous (IV) needles
  - All trauma alerts and unstable patients must have all of these.

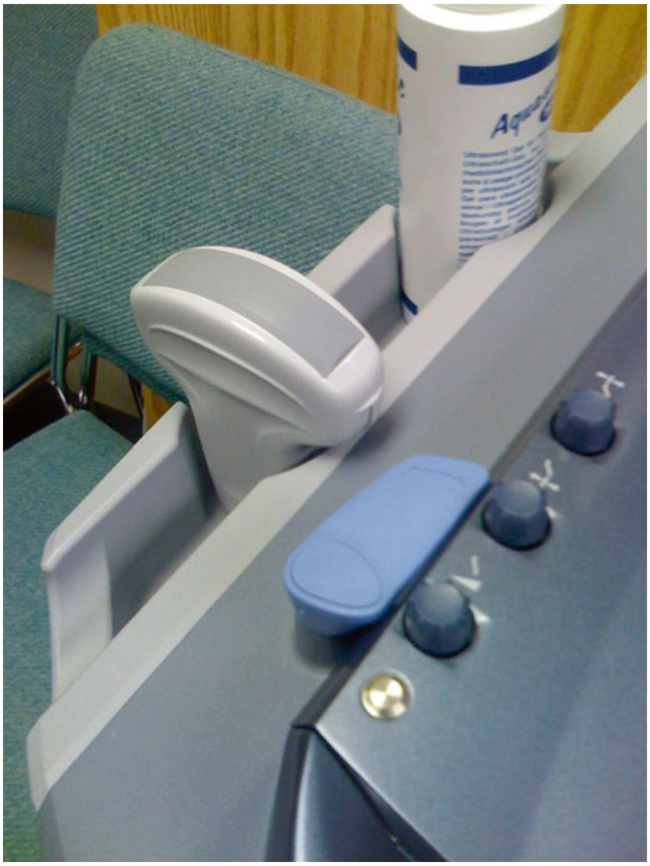

Fig. 71.2 Curved array transducer (C60) with a larger footprint and better resolution for deeper imaging that can be used for FAST examination as well as lung examination

71.4 Procedure

1. Ultrasound machine in the abdominal preset.
2. Patient in the supine position.
3. Phased array or curved array probe for focused assessment with sonography for trauma (FAST) and linear array for lung.
4. Begin scanning the patient in a systematic fashion.
   - All the views should be scanned by thoroughly sweeping through the area in question in order to maximize the information obtained.
   - All views should be obtained in the same order every time.
   - Obtain all four views, five views if pneumothorax is included.

71.4.1 Subxiphoid Four-Chamber View (Fig. 71.3)

1. Examine for free pericardial fluid.
    - Anechoic (black) stripe seen between the myocardium and the pericardium
2. Probe is placed in the subxiphoid area.
3. Indicator is to the patient's right.
4. Probe is directed toward the patient's left shoulder.
5. Use a shallow angle in the head to feet direction.
6. Should adequately visualize the following:
    - Liver edge superficially
    - Right ventricle
    - Left ventricle
    - Right atrium
    - Left atrium
7. If unable to obtain this view, proceed to *parasternal long-axis view*:
    - Probe is placed perpendicular at the left parasternal border.
    - Third to fourth intercostal space.
    - Indicator is to the patient's right shoulder.
    - Coronal section through the heart's long axis should adequately visualize the following:
        - Right ventricle most superficially
        - Left ventricle
        - Mitral valve
        - Left atrium
        - Aortic valve
        - Aortic outflow tract

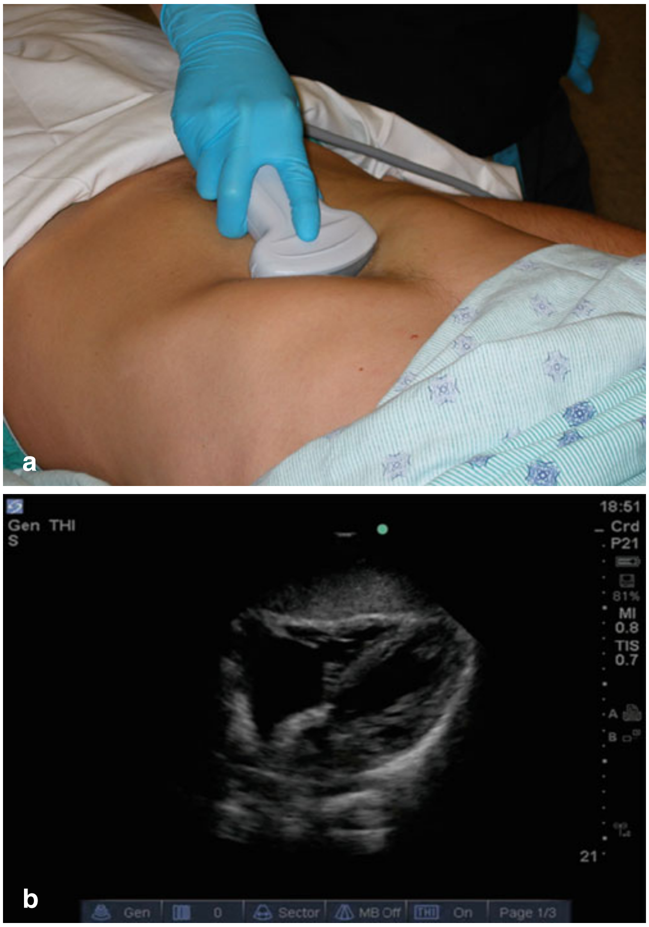

Fig. 71.3 (**a**) Image shows how to obtain the subxiphoid four-chamber view of the heart using the curved array transducer (C60) (Photograph courtesy of F. Eike Flach, MD).(**b**) Ultrasound image of four-chamber view of the heart (Used with permission from First aid for the emergency medicine clerkship 3rd Ed, McGraw Hill, 2011.) *RV* right ventricle, *LV* left ventricle, *RA* right atrium, *LA* left atrium

71.4.2 Right Upper Quadrant View (Fig. 71.4)

1. Examine for free fluid in all of the following areas:
   - Right intrathoracic space
     – Anechoic area above the diaphragm
   - Morison's pouch: hepatorenal space
     – Anechoic stripe between the liver and the kidney
   - Right paracolic gutter
     – Anechoic collection surrounding the inferior tip of the kidney
2. Probe is placed in the midaxillary line on the right.
3. Indicator is directed toward the patient's head.
4. Probe is in the coronal plane, angle can be aimed obliquely while scanning anterior to posterior.

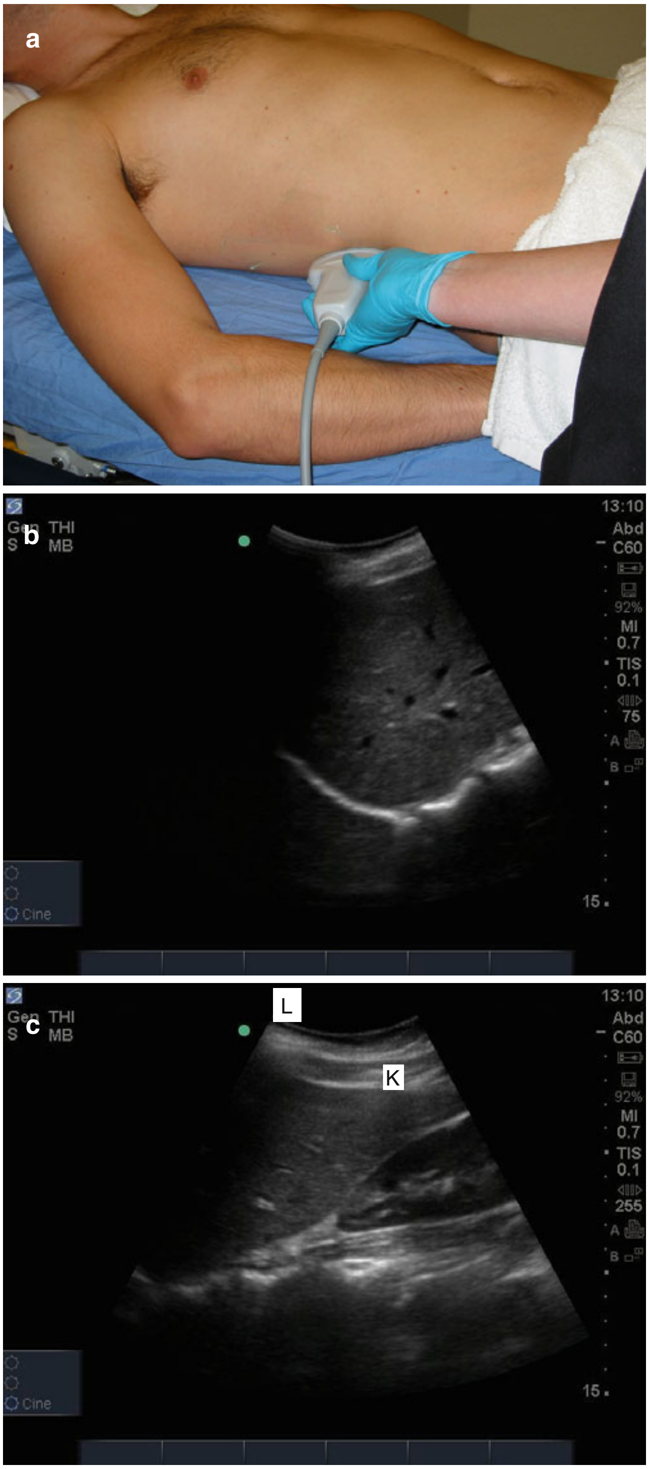
a
b
Gen S
THI MB
13:10
Abd
C60
92%
MI 0.7
TIS 0.1
75
A
B
15
Cine
c
L
K
Gen S
THI MB
13:10
Abd
C60
92%
MI 0.7
TIS 0.1
255
A
B
15
Cine

Fig. 71.4 (**a**) Image shows how to obtain the right upper quadrant view of the FAST exam using the curved array transducer (C60). The probe is aimed slightly obliquely in the coronal plane to get a better view between the ribs (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of the liver. (**c**) Ultrasonographic view of the liver–kidney interface (Morrison's pouch). *L* liver, *K* kidney

71.4.3 Left Upper Quadrant View (Fig. 71.5)

1. Examine for free fluid in all of the following areas:
   - Left intrathoracic space
     - Anechoic area above the diaphragm
   - Subphrenic space
     - Anechoic stripe below the diaphragm and above the spleen
   - Splenorenal space
     - Anechoic stripe between the spleen and kidney
   - Left paracolic gutter
     - Anechoic collection surrounding the inferior tip of the kidney
2. Probe is placed in the midaxillary line on the left.
3. Indicator is directed toward the patient's head.
4. Probe in the coronal plane, angle can be aimed obliquely while scanning anterior to posterior.

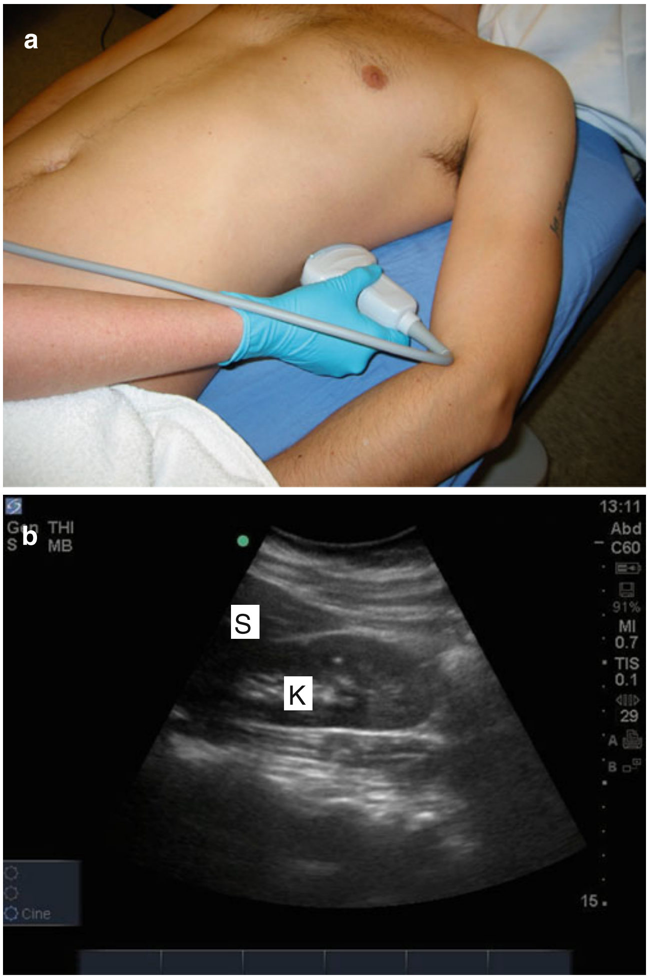

Fig. 71.5 (**a**) Image shows how to obtain the left upper quadrant view of the FAST examination using the curved array transducer (C60). Again, the probe is aimed slightly obliquely and is placed more superiorly in the midaxillary line (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of spleen–kidney interface. *S* spleen, *K* kidney

71.4.4 Pelvic View (Figs. 71.6 and 71.7)

1. Examine for intraperitoneal free fluid in the pelvis:
   - Anterior pelvis, above the bladder
     - Anechoic fluid above the bladder
   - Posterior cul-de-sac (pouch of Douglas)
     - Anechoic fluid posterior to the bladder or uterus
2. Probe is placed above the pubic symphysis over the bladder.
3. Scan through in both planes:
   - Transverse plane (Fig. 71.6)
     - Indicator is to the patient's right.
     - Scan through the bladder in the head to feet direction.
   - Sagittal plane (Fig. 71.7)
     - Indicator is aimed to the patient's head.
     - Scan through the bladder in a right to left direction.

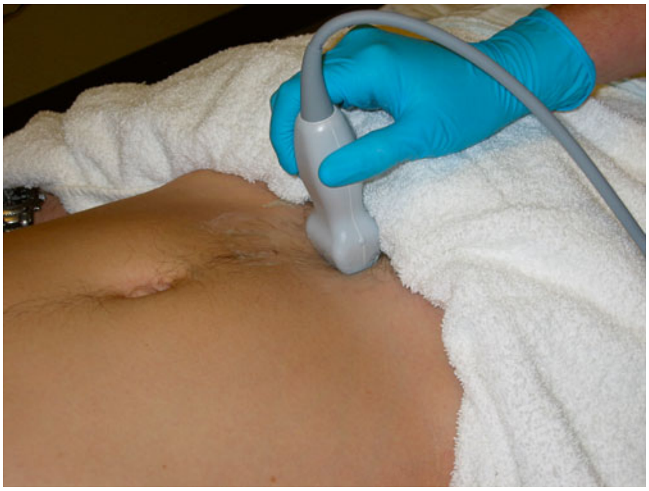

Fig. 71.6 Transverse pelvic view of the FAST examination using the phased array transducer (P17). With gentle force, the probe is pressed downward in order to look back behind the pubic symphysis and view the bladder (Photograph courtesy of F. Eike Flach, MD)

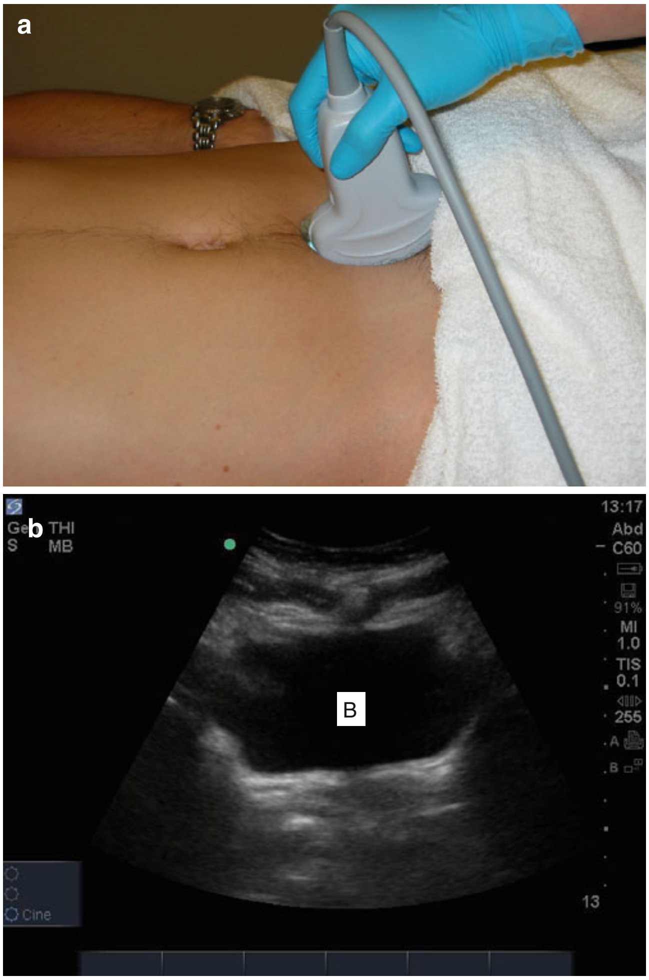

Fig. 71.7 (**a**) Sagittal pelvic view of the FAST exam using the curved array transducer (C60). With gentle force, the probe is pressed downward in order to look back behind the pubic symphysis and view the bladder (Photograph courtesy of F. Eike Flach, MD). (**b**) Ultrasonographic view of the bladder. *B* bladder

71.4.5 EFAST with Lung Views (Figs. 71.8 and 71.9)

1. Examine for pneumothorax:
   - Lung sliding
     - Absence: pneumothorax
     - Presence: normal lung
   - M mode tracing
     - Stratosphere sign: pneumothorax
     - Seashore sign: normal lung
2. Probe is placed on the anterior chest in the midaxillary line.
3. Level of the second to fourth intercostal spaces.
4. Sagittal position.
5. Center the probe over the pleural line between the ribs.
   - Find the rib and then slide the probe toward the head or feet to center the pleural line.
6. Observe for lung sliding.
7. Press M mode and move the line over the pleural line and press M mode again to get the tracing.
8. Examine multiple other areas anteriorly, moving distally, and in midaxillary line laterally, moving from superior to inferior.

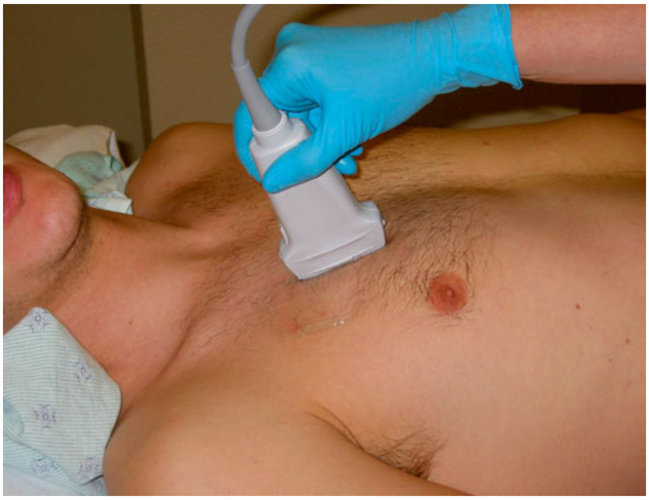

Fig. 71.8 Right lung view of the extended focused assessment with sonography for trauma (EFAST) examination using the linear array transducer (L38). The probe is placed in the sagittal plane on the anterior chest in the midaxillary line approximately at the second intercostal space and centered over the pleural line (Photograph courtesy of F. Eike Flach, MD)

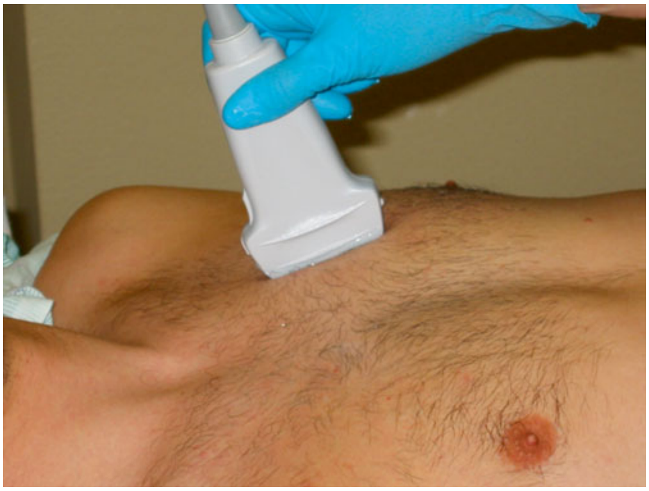

Fig. 71.9 Left lung view of the EFAST examination using the linear array transducer (L38) (Photograph courtesy of F. Eike Flach, MD)

71.5 Complications

- Overreliance on ultrasound to rule out abdominal injury:
  - FAST examinations do not detect retroperitoneal bleeding, solid organ injury, contained subcapsular hematomas, and bowel injuries.
- Not scanning through the object in question could lead to false-negative results.

71.6 Pearls and Pitfalls

- Always follow the ABCs (airway, breathing, circulation) first in any unstable patient.
- Always make sure the depth is set adequately.
  - Recommend starting deeper to make sure positive findings are not missed and then adjustments can be made from there.
- The curved array probe may be used throughout the entire EFAST for convenience if necessary.

71.6.1 Subxiphoid Four-Chamber View

- For larger body habitus, need to parallel the probe with the body in the subxiphoid area and use firm pressure to press the entire probe downward so as to look up under the xiphoid process at the heart.
- Moving the entire probe more to the patient's right in the subxiphoid area while still looking toward the left shoulder may improve visualization by using the liver as a window.
- Fat pad:
  - May be mistaken for pericardial fluid
  - Contains echoes and, therefore, is hypoechoic rather than anechoic
  - Should only be present anteriorly
    - Fluid should be gravity dependent, completely encircling the heart, and seen in multiple views.

71.6.2 Right Upper Quadrant View

- Normal artifacts of mirroring and loss of the spine are obscured when pleural fluid is present and, instead, the anechoic fluid is seen and there is loss of mirroring and continuation of the spine.

71.6.3 Left Upper Quadrant View

- Same as right upper quadrant view
- May be more difficult view to find than in right upper quadrant view for all of the following:
  - Spleen and kidney are more posterior and superior than in right upper quadrant view.
  - Spleen is smaller and less of a window for viewing.

71.6.4 Pelvic View

- Bowel can be mista0ken for free fluid or vice versa, but holding the probe still and observing can sometimes help distinguish the two.
  - Peristalsis will occur with bowel.
  - Internal echoes may be present in bowel.

71.6.5 Lung Views

- Ultrasound is more sensitive than a supine portable chest x-ray.
  - Apex anteriorly in midaxillary line.
- Rib
  - Hyperechoic horizontal line with a dense shadow posteriorly
  - Evenly spaced along the chest
- Pleural line
  - First hyperechoic line deep to the rib.
  - Actually includes the visceral and parietal pleura, but appears as one line.
  - Lung sliding is present in normal lung.
  - Comet tail artifact.
  - M mode tracing will be the same in normal lung and pneumothorax above the pleural line and different below the pleural line.
    - Seashore sign (Fig. 27.5a):
      - Appears as waves washing up on the shore.
      - Granular appearance represents movement.
    - Stratosphere sign (Fig. 27.5b):
      - Appears as straight lines
      - Bar code appearance

Selected Reading

Brunett P, Cameron P. Trauma in adults. In: Tintinalli J, Stapczynski J, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw Hill; 2012. p. 1671–5.

Ma JO, Mateer JR, Blaivas M. Trauma. In: Emergency ultrasound. Course Materials; New York, NY: McGraw Hill; 2008. pp. 7–109.

Saul T, Rivera M, Lewiss R. Ultrasound image quality. ACEP News. 2011;4:24–5.

Nasogastric Tube Placement

72

David P. Nguyen, L. Connor Nickels,
and Giuliano De Portu

72.1 Indications

- Evaluation of upper gastrointestinal (GI) bleeding (history of melena, bright red blood per rectum, or coffee-ground emesis)
  - Only in the cases in which frank blood is obtained, the sensitivity/specificity in detecting upper GI bleeding is poor.
  - It should not be used for diagnostic purposes. It is used to remove blood that is irritating the stomach and to determine whether bleeding is still occurring (lavage does not clear).
- Commonly used in decompression of the GI tract (partial/complete small bowel obstruction)
- Prevents aspiration and gastric dilation in intubated patients
- Used during gastric lavage and/or removal of toxins (activated charcoal) for acute overdose or poisonings

72.2 Contraindications

- Absolute
  - Facial trauma with possible cribriform plate fracture
- Relative
  - Severe coagulopathy (consider orogastric tube placement)
  - Esophageal strictures and alkali ingestions (possible esophageal perforation)
  - Esophageal varices (studies show that it is actually safe)

72.3 Materials and Medications

- For awake patients, consider pretreatment: lidocaine gel (2 % viscous)/nebulized lidocaine (4 or 10 %), vasoconstrictors (e.g., phenylephrine 0.5 %), and antiemetic (e.g., ondansetron 4 mg).
- 16- or 18-French sump tube lubricating jelly
- 50- or 60-mL syringe stethoscope

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_72

72.4 Procedure

- Preparation
  1. Awake patients, should receive antiemetics 15 min before procedure.
  2. Anesthetize both nares at least 5 min before placement.
     - Spray vasoconstrictor into both nares.
     - Inject about 5 mL of lidocaine gel along the floor of the nose.
     - Nebulized lidocaine via facemask also reduces both nasal and pharyngeal discomfort.
  3. Elevate the head of the bed to an upright position (when possible).
  4. Estimate tube insertion distance by measuring the tube from the xiphoid to the earlobe and then to the tip of the nose. Add 6 in. to this estimate and note the total distance. This helps with placement in the stomach and prevents esophageal placement or coiling in stomach. Mark the tube with markers or tape at the desired length.
  5. Lubricate the nasogastric (NG) tube.
- Insertion (Fig. 72.1)
  1. Always insert the tube gently into the nares along the floor of the nose under direct visualization. Always point inferiorly (do not point upward).
  2. If resistance is encountered, try to apply a small amount of pressure. STOP if unable to advance. Try the other side. It is necessary to prevent bleeding or dissecting the tissues.
  3. Have the patient flex his or her head forward when the tube is in the nasopharynx. This helps direct the tube toward the correct placement in the esophagus and not the trachea. Have the awake and cooperative patient sip water from a straw and swallow as the tube enters the oropharynx.
  4. Making the tube more rigid by placing it in cold water will help advance it because the "warmer" tube will tend to coil.
  5. Once the tube is in the esophagus, rapidly advance the tube into the stomach, taking into consideration the previously marked depth.
- Confirmation of tube placement
  1. Insufflate air into the end of the NG tube, via a 50- or 60-mL syringe, while auscultating for a rush of air (borborygmi) over the stomach.
  2. Aspiration of gastric contents (pH<4, there is >90 % gastric placement).
  3. The awake and cooperative patient should be able to talk, and if coughing or severe discomfort occurs, consider that esophageal or bronchial placement might have occurred.
  4. Radiographic evaluation:
     - "Gold standard" is to evaluate simple radiograph for position.
     - Consider in comatose patients.
- Secure the tube
  1. Tape the NG tube in place by taping both the tube and the nose. A butterfly bandage is typically used. Some companies produce a specific fixation for the tube.
  2. Secure the tube to where it does not press on the medial or lateral nostril (can lead to bleeding/necrosis).

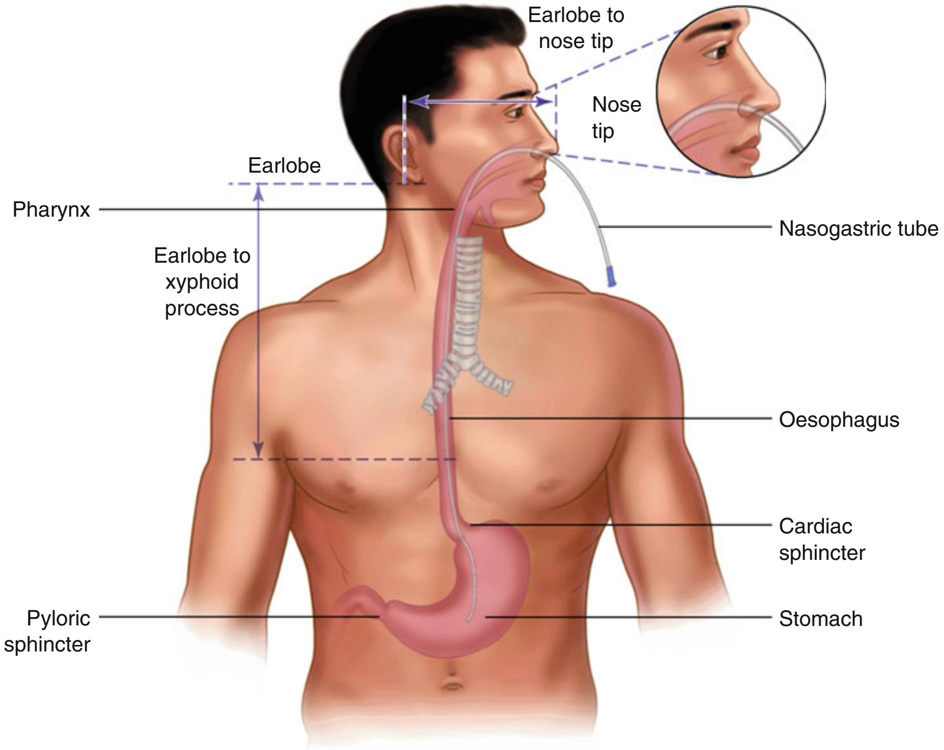

Fig. 72.1 NG tube placement

72.5 Complications

- Inability to pass the tube
- Bleeding
- Curling of the NG tube in the patient's mouth
- Pulmonary placement
- Nasal necrosis

72.6 Pearls

- Pearls
  - NG tube placement was ranked #1 as the most painful procedure in the emergency department so it is imperative to maintain patient's comfort by using anesthetics and even maybe intravenous anxiolytics.
  - Estimate the proper length of the tube before passage to avoid placing the tip of the tube in the esophagus or excessively coiling it in the stomach.
  - If leaving the tube for a prolonged period of time, make sure that the suction is set "intermittent" or "off" to prevent irritation to the gastric mucosa owing to direct pressure.

Selected Reading

Chun DH, Kim NY, Shin YS, Kim SH. A randomized, clinical trial of frozen versus standard nasogastric tube placement. World J Surg. 2009;33:1789–92.

Goff JS. Gastroesophageal varices: pathogenesis and therapy of acute bleeding. Gastroenterol Clin North Am. 1993;22:779.

Henneman PL. Gastrointestinal bleeding. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Tho PC, Mordiffi S, Ang E, Chen H. Implementation of the evidence review on best practice for confirming correct placement of nasogastric tube in patients in an acute care hospital. Int J Evid Based Healthc. 2011;9:51–60.

73

Esophageal Foreign Body Removal

David P. Nguyen, L. Connor Nickels,
and Rohit Pravin Patel

73.1 Indications

- Patient presenting with any one or combination of the following:
  - Foreign body sensation in throat, neck, substernal chest, or epigastric area
  - Clear history of ingestion
  - Dysphagia
  - Airway compromise
  - Drooling
  - Inability to tolerate fluids
  - Inability to tolerate solids
  - Evidence of perforation
  - Active bleeding
- All unstable patients should have immediate airway management and urgent endoscopy.
- FBs lodged superior to the epiglottis may be retrieved by emergency physicians in an emergent situation, but generally, a consultant should be present, whether ear, nose, and throat, gastroenterology, or general surgery.

73.2 Contraindications

- Absolute
  - None
- Relative
  - Performing rapid sequence intubation (RSI) in a patient with an FB that could compromise the airway.
  - Generally, if the patient is breathing on their own, collaborate with a consultant on the best method to secure the airway (emergency department, intensive care unit, operating room).
  - Treating with glucagon repeatedly if it is inducing vomiting and/or not working.
  - Conservatively managing a patient who should otherwise undergo a procedure for removal.

73.3 Materials

- Esophagoscopy/endoscopy is the definitive diagnostic and therapeutic procedure for impacted esophageal FBs.
  - Generally, should not be performed by an emergency physician.
  - Devices used include forceps, baskets, polypectomy snares, and nets.
  - Endoscopic techniques include push into stomach, push plus fragmentation, pull with retrieval forceps, and pull with various items (basket, snare, nets).
- Foley catheter removal
  - Widely used technique for recently ingested single, smooth, blunt, and radiopaque objects
  - #12 to #16 French Foley catheter
  - Forceps (bayonet and Magill) of various sizes
  - Often done under fluoroscopic guidance
- Bougienage
  - A single, smooth object, such as a coin, lodged less than 24 h in a patient with no respiratory distress or esophageal disease can be advanced successfully into the stomach by using bougienage.
  - Dilator size is selected based on age:
    - 1–2 years: 28 French
    - 2–3 years: 32 French
    - 3–4 years: 36 French
    - 4–5 years: 38 French
    - Longer than 5 years: 40 French
- Relaxation of the lower esophageal sphincter (LES)
  - Some FBs lodged at the LES can be medically managed by relaxation of the LES.

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_73

  - Most ingested FBs and impacted food boluses eventually pass spontaneously.
    - 1–2 mg of glucagon intravenously
    - 0.4–0.8 mg of nitroglycerin sublingually
    - 5–10 mg of nifedipine sublingually
    - Carbonated beverage

73.4 Procedure

- Push technique and push with fragmentation technique (generally performed by specialists)
  1. First accepted endoscopic method.
  2. Gentle pressure is applied with the tip of the endoscope on the esophageal food bolus after air insufflation.
  3. If pressure does not disimpact the bolus, fragmentation can be attempted but is generally avoided owing to unknown pathology behind the food bolus.
- Foley catheter removal
  1. Moderate sedation and nasopharyngeal topical anesthesia may be used.
  2. Place the patient in a head-down Trendelenburg position.
  3. Check for symmetrical balloon inflation of the Foley catheter.
  4. Under fluoroscopy, visually pass the catheter distal to the FB.
  5. Fill the balloon slowly with 3–5 mL of saline or contrast agent.
  6. Using steady, gentle traction, withdraw the catheter with the balloon inflated distal to the FB.
  7. Grasp the object with fingers, forceps, or clamp once it is visualized in the oropharynx.
- Bougienage
  1. Topical anesthesia is recommended.
  2. Blind esophageal bougienage resembles placement of an orogastric tube.
  3. Place the patient in a sitting position.
  4. Pass a well-lubricated, appropriately sized bougie posteriorly along the roof of the mouth, following the natural curve of the soft palate caudally to the hypopharynx.
  5. Encourage the patient to swallow (to help pass the dilator through the cricopharyngeus muscle).
  6. Ask the patient to phonate to help exclude accidental laryngeal intubation.
  7. Once past the cricopharyngeus muscle, extend the head to aid the bougie in passing distally to the stomach.
  8. Post-procedure radiograph is used to confirm passage into the stomach.
- Relaxation of the LES
  1. Premedicate with an antiemetic, such as ondansetron.
  2. Administer 1–2 mg of glucagon intravenously (0.02–0.03 mg/kg in children, not to exceed 0.5 mg) with the patient in a sitting position over 1–2 min.
  3. Carbonated beverages given after glucagon ingestion have shown to have higher success rates.
  4. An alternative is to use either sublingual nitroglycerin (1–2 0.4 mg tabs) or 5–10 mg of nifedipine to relieve LES tone.
  5. This procedure does not work in patients with structural abnormalities.

73.5 Complications

- Esophageal FBs may cause esophageal pressure leading to edema, necrosis, infection, laceration, and/or perforation.
- Be cognizant of time (risk of complications is higher the longer the FB is left in place) and treatment side effects (i.e., do not continue to give patient water or glucagon if these induce vomiting).
- Aspiration and perforation during procedures listed previously.
- Late complications: esophageal stricture, abscess, mediastinitis, tracheoesophageal fistula, vascular injuries, pneumothorax, pericarditis, aspiration pneumonia, and vocal cord paralysis.

73.6 Pearls

- Esophageal foreign bodies can be lodged in the upper (proximal), middle, or lower (distal) one third:
  - Proximal: cervical web, Zenker's diverticulum
  - Middle: eosinophilic esophagitis, cancer, radiation structure, spastic dysmotility
  - Distal: peptic stricture, eosinophilic esophagitis, cancer, achalasia, esophageal diverticula, spastic dysmotility
- Because food bolus impactions are generally associated with pathology, follow-up evaluation for these abnormalities should be considered.
- Esophagus foreign bodies should not be allowed to remain in the esophagus beyond 24 h from presentation.
- Button/disc batteries in esophagus (emergent removal)
  - Considered an emergency, because liquefaction necrosis and perforation can occur rapidly.
  - Most common ingestions are hearing aid batteries.
  - If in stomach, and patient is a symptomatic, can wait up to 24 h.
- Sharp objects (emergent removal)

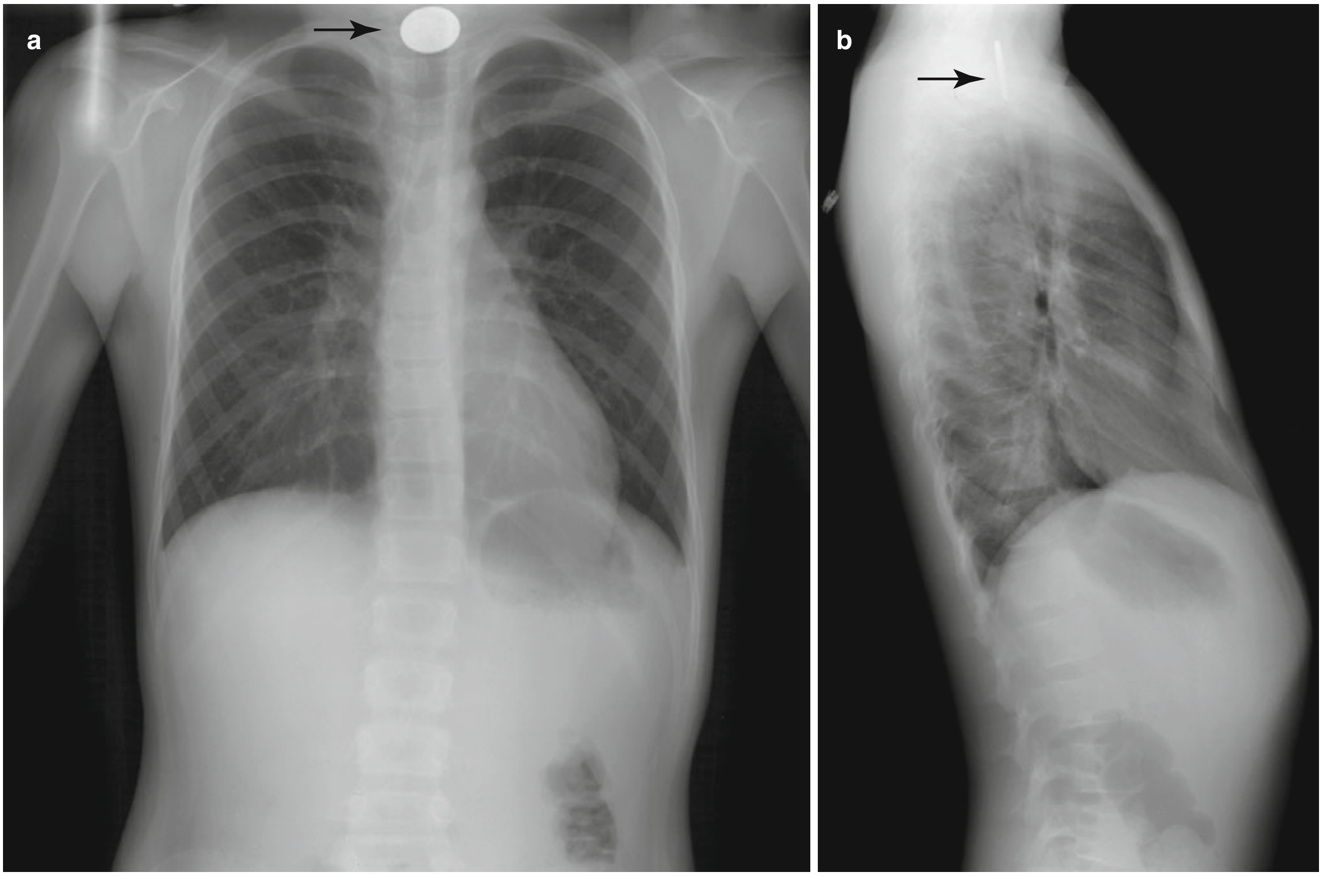

Fig. 73.1 (**a**) Anteroposterior (AP) and (**b**) lateral views demonstrating a coin in the esophagus. A coin in the trachea would present in the opposite manner—the coin would be seen on edge in the AP view and flat on the lateral view (Reproduced with permission from McGraw-Hill: Stead LG, et al. *First Aid for the Pediatrics Clerkship*. New York: McGraw-Hill, 2010)

  - Cause the majority of complications (~35 %) with esophageal FBs.
  - Direct visualization with endoscopy is the only appropriate removal technique.
- Magnets (urgent removal)
  - Can cause necrosis and fistula formation due to the way they adhere to the mucosa
- Esophageal coins (remove within 24 h) (Fig. 73.1)
  - Up to 80 % of coins at the LES will pass spontaneously within 24 h without interventions. The watchful waiting approach is used only in patients with single coins and who are asymptomatic.
  - Common complications of these procedures include mild bleeding, lip laceration, bradycardia with Foley catheter insertion, and teeth injuries.
  - Some protocols include RSI as part of the management process and should be considered if lifesaving.

Selected Reading

ASGE Standards of Practice Committee. Management of ingested foreign bodies and food impactions. Gastrointest Endosc. 2011; 73:1085.

Bhargava R, Brown L. Esophageal coin removal by emergency physicians: a continuous quality improvement project incorporating rapid sequence intubation. CJEM. 2011;13:28–33.

Conway WC, Sugawa C, Ono H, Lucas CE. Upper GI foreign body: an adult emergency hospital experience. Surg Endosc. 2007; 21:455–60.

Katsinelos P, Kountouras J, Paroutoglou G, et al. Endoscopic techniques and management of foreign body ingestion and food bolus impaction in the upper gastrointestinal tract: a retrospective analysis of 139 cases. J Clin Gastroenterol. 2006;40:784–9.

Activated Charcoal

74

Deylin I. Negron Smida and Judith K. Lucas

74.1 Indications

- Single-dose activated charcoal (AC) (Fig. 74.1)
  - Does not meet criteria for gastric emptying.
  - Gastric emptying may be too harmful.
  - Ingestion of toxic xenobiotic is known to be adsorbed by AC.
  - Ingestion occurred with a time frame amenable to adsorption by AC, or clinical factors are present that suggest that not all of the xenobiotic had already been systemically absorbed.
  - Ingestion of extended- or sustained-release formulations.
- Multiple-dose activated charcoal therapy (MDAC)
  - Life-threatening ingestion of:
    - Carbamazepine
    - Phenobarbital
    - Quinine
    - Theophylline
    - Dapsone
  - Life-threatening ingestion of another xenobiotic that undergoes enterohepatic recirculation and is adsorbed to AC
  - Ingestion of a significant amount of a slowly released xenobiotic
  - Ingestion of a xenobiotic known to form concretions or bezoars, such as aspirin

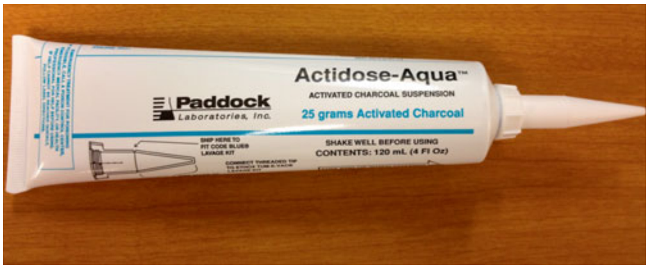
Paddock Laboratories, Inc.
Actidose-Aqua™
ACTIVATED CHARCOAL SUSPENSION
25 grams Activated Charcoal
SHAKE WELL BEFORE USING
CONTENTS: 120 mL (4 Fl Oz)

Fig. 74.1 Activated charcoal (AC)

D.I.N. Smida, MD
Department of Emergency Medicine, University of Pittsburgh Medical Center, Saint Margaret Hospital, Pittsburgh, PA, USA
e-mail: Deylin.negron@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_74

74.2 Contraindications

- Absolute
  - Gastric perforation
  - Gastrointestinal ileus, obstruction, or diminished peristalsis
  - Nonintubated patients with the potential of losing protective airway reflexes
  - Intestinal obstruction
  - Ingestion of:
    - Corrosives
    - Petroleum distillates
- Relative
  - Altered or decreased level of consciousness unless intubated.
  - Vomiting.
  - Xenobiotic has limited toxicity at almost any dose.
  - Dose ingested is less than the dose expected to produce significant illness.
  - Presentation many hours after ingestion.
  - Minimal signs or symptoms of poisoning.
  - Ingested xenobiotic has a highly efficient antidote.
  - Administration of charcoal may increase the risk of aspiration (i.e., hydrocarbons).

74.3 Materials and Medications

- Nasogastric (NG) tube/orogastric (OG) tube (Fig. 74.2)
- Baby bottle with split nipple (designed for drinking of slurry solutions, such as thickened formulas) or sippy cup without the valve
- Absorbent pad
- Basin
- Water-soluble lubricant
- Tubing connected to suction device
- Flavored syrup

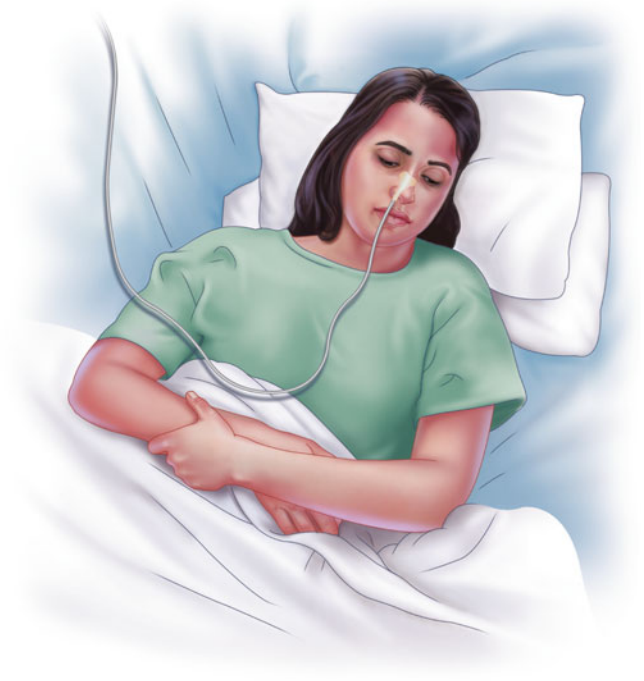

Fig. 74.2 Drinking AC (by cup, sippy cup, or bottle) is effective, but it may also be infused via nasogastric or orogastric tube

74.4 Procedure

- Single-dose administration
  - Adult
    - Can be taken via a cup and straw (drunk) if the patient is cooperative.
    - The optimal dose of AC is unknown.
    - 50–100 g/dose (1 g/kg), administered at a rate no less than 12.5 g/h or its equivalent.
    - If vomiting is anticipated, an intravenous antiemetic is recommended.
  - Children
    - 1 g/kg or 10:1 ratio of AC to drug ingested.
    - After massive ingestion, give 2 g/kg.
    - Many children will drink the suspension from a bottle or sippy cup, especially if it is mixed with juice or flavored syrup (e.g., chocolate or strawberry).
- MDAC
  - Adults: 0.5–1 g/kg every 2–4 h for 24–48 h
  - Children (<12 y old): 0.25–0.5 g/kg every 2–4 h or rate of 0.2 g/kg/h for 24–48 h
- Mixtures
  - Ready to drink
  - Powder form
    - Add eight parts water to the selected powdered form.
    - Gatorade or juices can also be used to help hide the flavor and texture.
    - In children, the AC can be mixed with cold chocolate or some other flavored syrup, which also hides the flavor.
  - Shake liquid suspension well for 1 min.
  - If the patient vomits, the dose should be repeated. Smaller, more frequent, dosing may be better tolerated, and an antiemetic may be needed.

74.5 Complications

- Aspiration pneumonitis
- Transient constipation
- Intestinal bezoars
- Bowel obstruction
- Diarrhea, dehydration, hypermagnesemia, and hypernatremia with coadministered cathartics or MDAC
- Vomiting
- Corneal abrasion if spilled in the eyes

74.6 Pearls and Pitfalls

- Pearls
  - If an OG or NG tube is used, time should be allowed for the last dose to pass through the stomach before the tube is removed. Suctioning the tube before removal may prevent subsequent AC aspiration.
  - With children, the colder and sweeter the solution and if the color is camouflaged (cup with a lid or a sippy cup), there will be increased success at oral administration (vs NG/OG).
- Pitfalls
  - No evidence-based literature supports the assertion that AC changes clinical outcome.
  - Xenobiotics and AC adsorption (Table 74.1).
  - Incorrect application (e.g., into the lungs) results in pulmonary aspiration, which can be fatal if unrecognized.
    - Incorrect placement of NG/OG tube into trachea.
    - Administration of AC to a patient with an ileus (e.g., in anticholinergic overdoses).
  - No specific contraindication for AC in pregnant women; however, diarrhea or hypernatremia in the mother may adversely affect the fetus.

Table 74.1 Absorption of xenobiotics by AC

Good absorption	Poor absorption
Acetaminophen	Alkali
Bupropion	Chlorpropamide
Caffeine	Doxepin
Carbamazepine	Ethanol or other alcohols
Chlordecone	Ethylene glycol
Dapsone	Fluoride
Digitoxin	Heavy metals
Nadolol	Imipramine
Phenobarbital	Inorganic salts
Phenylbutazone	Iron
Phenytoin	Lithium
Salicylate	Methotrexate
Theophylline	Mineral acids
	Potassium
	Tobramycin
	Valproate sodium
	Vancomycin

Selected Reading

American Academy of Clinical Toxicology; European Association of Poisons Centres and Clinical Toxicologists. Position statement and practice guidelines on the use of multi-dose activated charcoal in the treatment of acute poisoning. J Toxicol Clin Toxicol. 1999;37:731–51.

Chyka PA, Seger D. American Academy of Clinical Toxicology; European Association of Poisons Centres and Clinical Toxicologists. Position statement: single-dose activated charcoal. J Toxicol Clin Toxicol. 1997;35:721–41.

Gude A, Hoegberg LCG. Techniques to prevent gastrointestinal absorption. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 9th ed. New York: McGraw-Hill; 2011. p. 93–7,431.

Lie D. Use of activated charcoal in drug overdose. Medscape family medicine. 25 Mar 2004. www.medscape.com/viewarticle/471331

Olson KR. Emergency evaluation and treatment. In: Olson KR, Anderson IB, Benowitz NL, et al., editors. Poisoning and drug overdose. 5th ed. New York: McGraw-Hill; 2007. p. 1–56.

75 Gastric Lavage

Deylin I. Negron Smida and Judith K. Lucas

75.1 Indications

- Recent ingestion (<30–60 min).
- Life-threatening exposure where there is a high suspicion that a xenobiotic is still present in the stomach and evacuation is expected to contribute to an improved outcome (e.g., iron, tricyclic antidepressants).
- Ingested agent is not absorbed with activated charcoal (e.g., pesticides, hydrocarbons, iron, alcohols, lithium, and solvents).
- Activated charcoal is unavailable.
- Ingestion exceeds adsorptive capacity of initial activated charcoal dosing (e.g., >100 mg/kg of pills).
- Ingestion of an agent likely to form a durable mass or bezoars after overdose.

75.2 Contraindications

- Vomiting
- Unintubated patients with potential to lose airway protective reflexes
- Ingestion of a xenobiotic with aspiration potential (e.g., hydrocarbon) without intubation
- Ingestion of caustic substances (alkali or acidic)
- Ingestion of sharp metals
- Ingestion of a foreign body (e.g., drug packet)
- Risk for hemorrhagic gastrointestinal perforation
- Ingestion of xenobiotic in a form known to be too large to fit into the lumen of the orogastric tube
- Nontoxic ingestions

75.3 Materials and Medications

- Orogastric tube (Ewald tube or the Tum-E-Vac) (Fig. 75.1)
  - Adults and adolescents: 36–40 French
  - Children: 22–28 French
- Pen or tape to mark the length of the tube
- Water-soluble lubricant
- Suction
- Emesis basin
- Absorbent pad
- Catheter-tip syringe with 2 mL water/saline to check position of the tube
- Room temperature irrigation fluid
- Bite block or oral airway to prevent patients from biting down on the tube

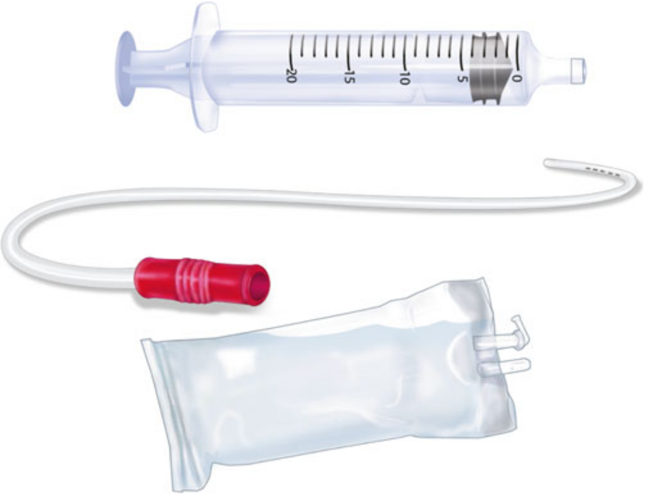

Fig. 75.1 Materials needed for gastric lavage include a large-bore nasogastric tube, a 60-cc non-Luer-Lok syringe, and a solution, typically normal saline, for lavage

---

D.I.N. Smida, MD
Department of Emergency Medicine, University of Pittsburgh Medical Center, Saint Margaret Hospital, Pittsburgh, PA, USA
e-mail: Deylin.negron@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_75

75.4 Procedure

1. If there is potential airway compromise, endotracheal or nasotracheal intubation should precede orogastric lavage.
2. Place an oral airway or a bite block to prevent biting of the endotracheal tube if the patient recovers consciousness or has convulsions during the procedure.
3. Ensure suction apparatus is available and functioning.
4. Place the patient in an upright-seated position if awake and alert.
5. Place patient in the left lateral decubitus position if obtunded.
6. Before insertion, the proper length of tubing to be passed should be measured from the mouth, back to the ear, and down anterior to the chest and abdomen, beyond the point where any side ports on the tube would be beyond the level of the estimated lower esophageal sphincter (Figs. 75.2 and 75.3).
7. If the patient is still awake, insert the gastric tube to the level of the glottis, and encourage the patient to swallow.
8. Pass the tube to the stomach.
   (a) Coughing, airflow, or fog from the tube raises the concern for inadvertent tracheal positioning.
9. After the tube is inserted, it is essential to confirm that the distal end of the tube is in the stomach, by "popping" 5–10 mL of air into the tube while someone is listening with a stethoscope over the stomach.

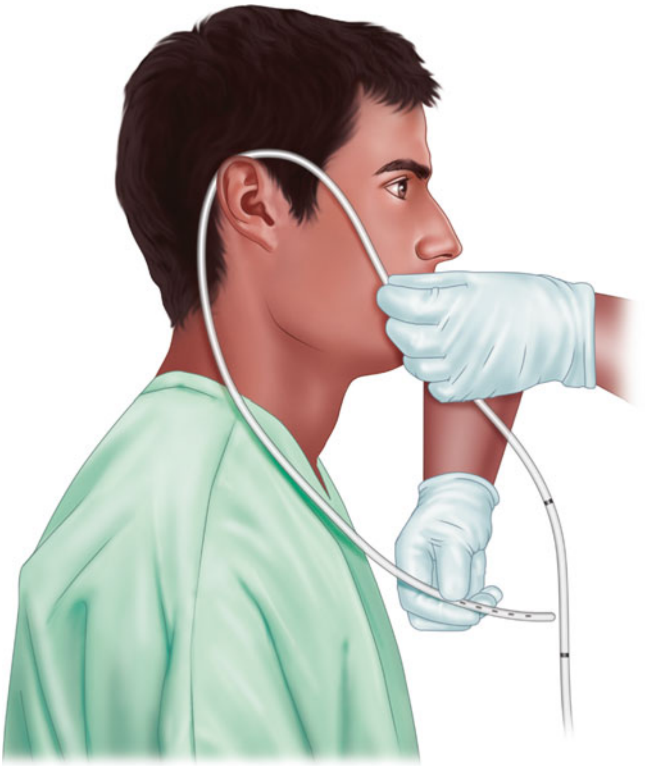

Fig. 75.3 Measuring correct placement of the tube. Place the distal tip over the stomach and wrap the tube up behind the ear (usually the right ear because the tubes generally pass easier through the right nares) and around the nares. The black line or centimeter mark at the level of the nares is the point of insertion when passage stops

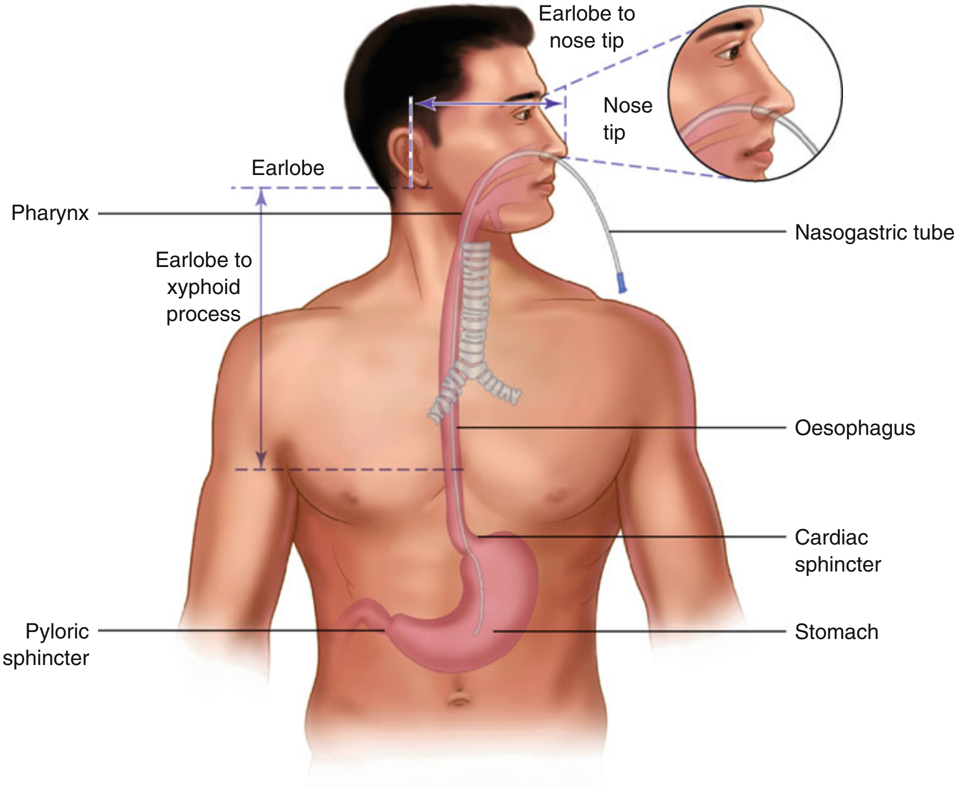

Fig. 75.2 Diagram illustrates appropriate placement of the lavage tube

10. In adults, 250-mL aliquots of a room temperature saline lavage solution are instilled via a funnel or lavage syringe. In children, aliquots should be 10–15 mL/kg to a maximum of 250 mL and suctioned back out of the tube attached to low to moderate continuous wall suction. Instillation of lavage solution and suction is repeated (Fig. 75.4).
11. Orogastric lavage should continue for at least several liters in an adult and/or at least 0.5–1 L in a child if the return is free of debris or until no particulate matter returns and the effluent lavage solution is clear.
12. Those caring for the patient must remain protected at all times, using goggles, mask, gown, and gloves. If the ingested poison is toxic via pulmonary or skin absorption, isolate the ingestant immediately in a self-contained wall suction unit.
13. Any material still in the stomach should be withdrawn, and immediate instillation of the activated charcoal should be considered for large ingestions of xenobiotics known to be adsorbed by activated charcoal.

75.5 Complications

- Vomiting
- Esophageal tears or perforation after orogastric tube insertion
- Inadvertent tracheal intubation and/or airway trauma
- Aspiration pneumonitis

75.6 Pearls and Pitfalls

- Pearls
  - You must use a large-bore orogastric tube for maximal efficacy.
  - The left lateral decubitus position is recommended because the pylorus points upward in this orientation. This position theoretically helps prevent the xenobiotic from passing through the pylorus during the procedure.
- Pitfalls
  - Large drug packets, adherent masses of pills, and plant and mushroom fragments will not pass through a 40-French lavage tube.

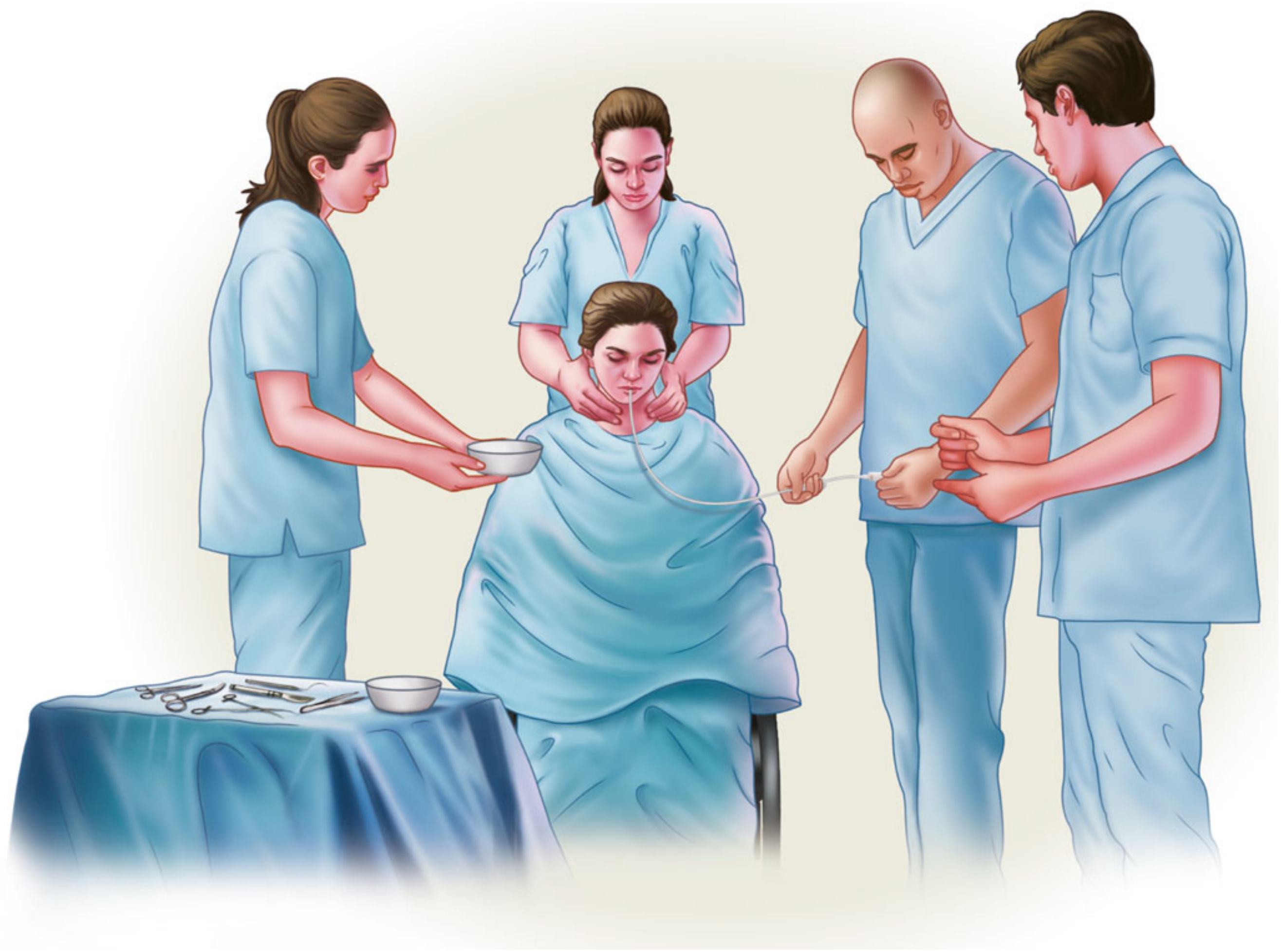

Fig. 75.4 Lavage in progress

Selected Reading

Gude A, Hoegberg LCG. Techniques to prevent gastrointestinal absorption. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 8th ed. New York: McGraw-Hill; 2006. p. 91–3.

Olson KR. Poisoning & drug overdose. In: Olson KR, Anderson IB, Benowitz NL, et al., editors. Emergency evaluation and treatment. 6th ed. New York: McGraw-Hill; 2012.

Smilktein MJ. Techniques used to prevent gastrointestinal absorption of toxic compounds. In: Nelson LS, Lewin NA, Howland MA, et al., editors. Goldfrank's toxicologic emergencies. 7th ed. New York: McGraw-Hill; 2002. p. 46–8.

Whole-Bowel Irrigation

76

Judith K. Lucas

76.1 Indications

- Whole-bowel irrigation (WBI) should not be used routinely in the management of the poisoned patient (because there is no clinical proof it will change clinical outcome).
- Ingestion of significant amount of medications.
  - Not adsorbed by activated charcoal
    - Lead, lithium, arsenic, and zinc
    - Substantial amounts of iron (high morbidity and no other effective method to gastrointestinal decontamination)
  - Sustained-release medications or enteric-coated drugs
  - Disk batteries distal to the pylorus
  - Whole transdermal patches (fentanyl, clonidine, nicotine)
  - Drug concretions
  - Ingested packets of illicit drugs

76.2 Contraindications

- Absolute
  - Bowel obstruction
  - Bowel perforation
  - Ileus
  - Hemodynamic instability
  - Compromised or unprotected airway
  - Intractable vomiting
- Relative
  - Concurrent or recent administration of activated charcoal (may decrease the effectiveness of activated charcoal)

76.3 Materials and Medications

- Topical anesthesia, although not mandatory, will reduce the pain of nasogastric (NG) tube placement.
  - 10 % lidocaine spray
  - Lidocaine gel
- Small-bore (12-French) NG tube (Fig. 76.1).
- Tape for securing the NG tube.
- Reservoir or feeding bag used for NG tube feedings (Fig. 76.2).
- Intravenous pole.
- Bedside commode or toilet (Fig. 76.3).
- Polyethylene glycol-electrolyte solution (PEG-ES) (Fig. 76.4).
- Antiemetic.
  - No absolute indication for prophylactic use
  - May be helpful if vomiting ensues during infusion
    - Metoclopramide
      - Antiemetic
      - *Increases gastric motility*

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_76

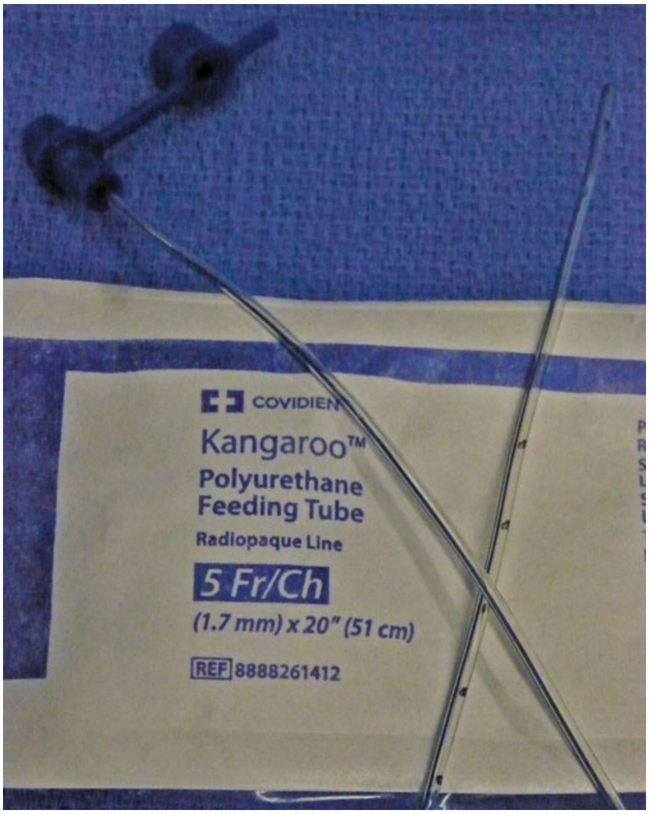
COVIDIEN
Kangaroo™
Polyurethane
Feeding Tube
Radiopaque Line
5 Fr/Ch
(1.7 mm) x 20" (51 cm)
REF 8888261412

Fig. 76.1 Nasogastric (NG) tube. Typically, the infusion of the lavage solution is too rapid to be taken orally, so an NG tube can be placed. Since the irrigation solution is of low viscosity, a small-bore NG tube should be used for comfort

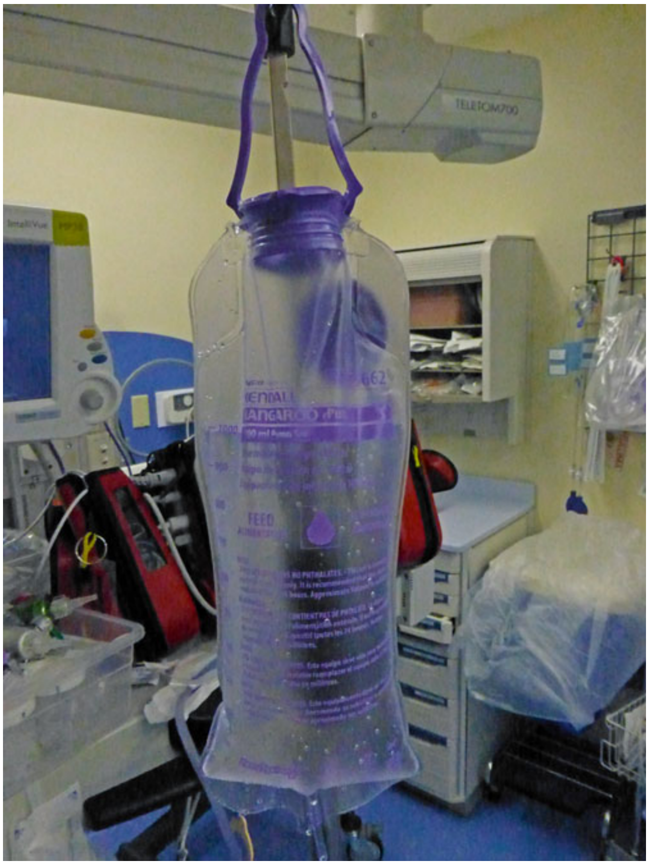

Fig. 76.2 Bag from which the lavage solution will drain; it is similar to the bags used for gastrostomy tube feeding

Fig. 76.3 Almost always, the patient will need to be seated on or very near a portable commode, as once the irrigation solution starts to move through the bowels, defecation will occur rapidly

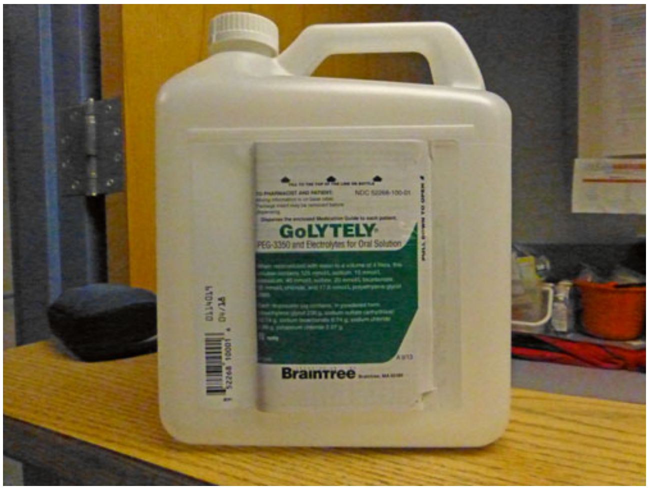

Fig. 76.4 Example brands of intestinal irrigation electrolyte solutions

76.4 Procedure

1. An NG tube is required because most patients will not drink the PEG-ES at the necessary rate.
2. Place a small-bore (12-French) NG tube to a sufficient distance that the tip lies in the central portion of the stomach.
3. Confirm NG placement with a radiograph.
4. Attach the tube to the reservoir bag of PEG-ES and hang from an elevated site (an extended intravenous pole).
5. The patient should be seated in an upright position.
   - Promotes settling of the intoxicant in the distal portion of the stomach
   - Decreases the likelihood of vomiting
6. Dosing:
   - Children 9 months to 6 years: 500 mL/h
   - Children 6–12 years: 1,000 mL/h
   - Adolescents/adults: 1,500–2,000 mL/h
7. Collect effluent.
8. Continue infusion.
   - Until the rectal effluent is the same color as the influent (i.e., clear), usually between 4 and 6 h.
   - You may continue beyond clear effluent if clinical evidence indicates ongoing effectiveness:
     - Continued pill fragments or drug packets are present in the effluent.
     - Radiographic evidence that pills, pharmacobezoars, or packets are still present.

76.5 Complications

- Nausea, vomiting, and bloating
- Misplacement of the NG tube
- Esophageal perforation owing to NG tube placement
- Aspiration pneumonitis in the unprotected airway

76.6 Pearls

- Overall, WBI is probably more effective than gastric lavage, but probably less effective than activated charcoal in preventing poison absorption (when the intoxicant can be adsorbed to charcoal).
- Vomiting.
  - Usually secondary to the ingestant (i.e., emetogenic toxins, such as iron)
  - May be due to rate of infusion
    - Slow rate by 50 % for 30–60 min.
    - Then return to original rate.
- If resistance is encountered during NG tube placement, do not force passage. Remove and redirect.

Selected Reading

Bailey B. To decontaminate or not to decontaminate? the balance between potential risks and foreseeable benefits. Clin Pediatr Emerg Med. 2008;9:17–23.

Hanhan UA. The poisoned child in the pediatric intensive care unit. Pediatr Clin North Am. 2008;55:669–86. xi.

Lheureux P, Tenenbein M. Position paper: whole bowel irrigation. American Academy of Clinical Toxicology/European Association of Poison Centres and Clinical Toxicologists. J Toxicol Clin Toxicol. 2004;42:843–54.

Othong R. Whole-bowel irrigation. MedScape Reference: drugs, diseases, and procedures. Updated: Aug 2011

Postuma R. Whole bowel irrigation in pediatric patients. J Pediatr Surg. 1982;17:350–2.

77 Sengstaken-Blakemore Tube

Thomas T. Nguyen, Etan Eitches,
and Stephanie Wetmore-Nguyen

77.1 Indications

- Life-threatening esophageal variceal bleed refractory to endoscopy and medical therapy
- Life-threatening esophageal variceal bleed refractory to medical therapy in the absence of possible endoscopy

77.2 Contraindications

- Absolute
  - Known esophageal rupture
  - Unable to intubate or maintain airway
- Relative
  - History of prior esophageal trauma or strictures
  - Recent surgery of the gastroesophageal junction
  - Resolved or resolving variceal bleeding

77.3 Materials and Medications

- Sengstaken-Blakemore (SB) tube (Fig. 77.1)
- 60-mL syringe with catheter tip (Fig. 77.2)
- Sphygmomanometer (Fig. 77.3) or cuffalator (Fig. 77.4)
- Y-Tube connector (Fig. 77.5) or 3-way stop-valve connector (Fig. 77.6)
- Vacuum suction device and tubing (Fig. 77.7)
- Tube clamps (4) (Fig. 77.8)
- Lubricant (water soluble)
- Lidocaine (Xylocaine) spray or gel
- Anchoring device such as a football helmet or catcher's mask (Fig. 77.9)
- Cup of water and straw if the patient is awake
- Scissors (Fig. 77.10)
- Intubation equipment
- Sterile water

T.T. Nguyen, MD (✉)
Department of Emergency Medicine,
Mount Sinai Beth Israel, New York, NY, USA
e-mail: tnguyen@chpnet.org

E. Eitches, MD
Department of Emergency Medicine,
Beth Israel Medical Center, New York, NY, USA
e-mail: eeitches@gmail.com

S. Wetmore-Nguyen, MD
Department of Emergency Medicine,
New York-Presbyterian Hospital/Columbia University Medical Center,
New York, NY, USA
e-mail: slw9004@nyp.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_77

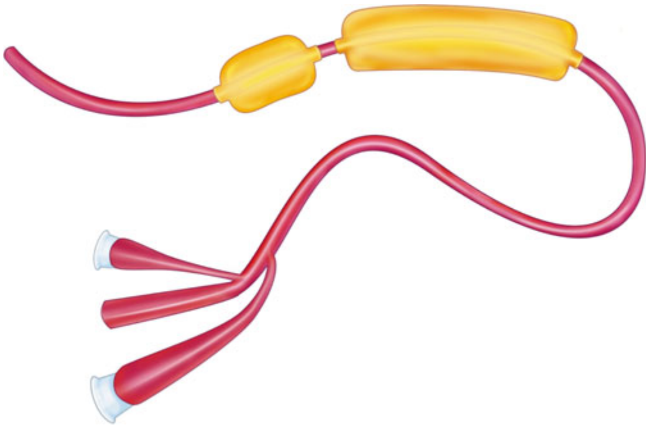

Fig. 77.1 Sengstaken-Blakemore (SB) tube

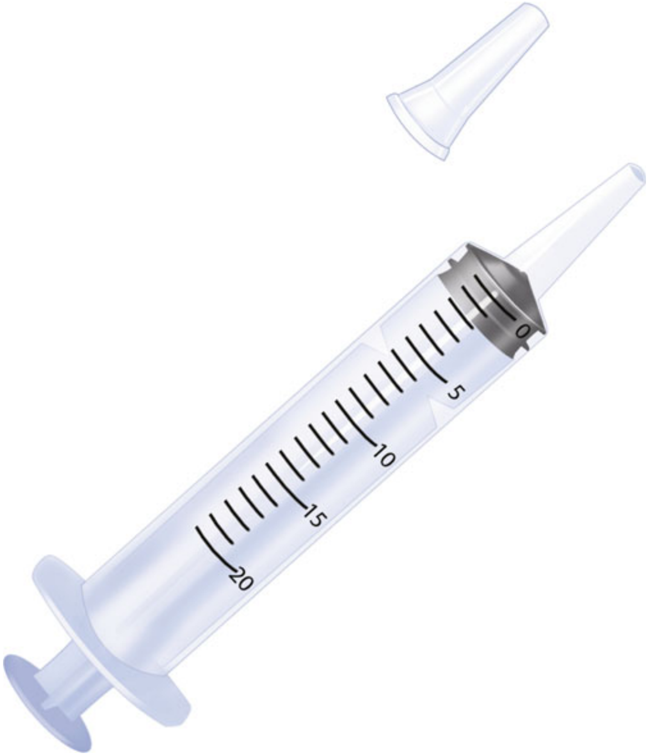

Fig. 77.2 Syringe

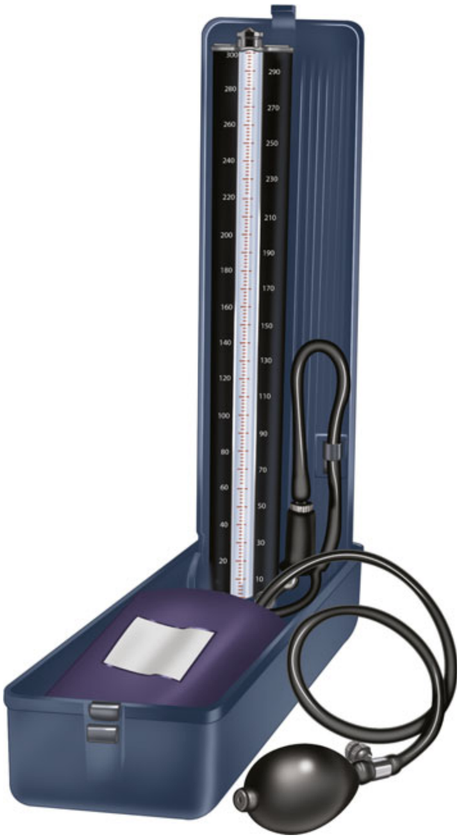

Fig. 77.3 Sphygmomanometer

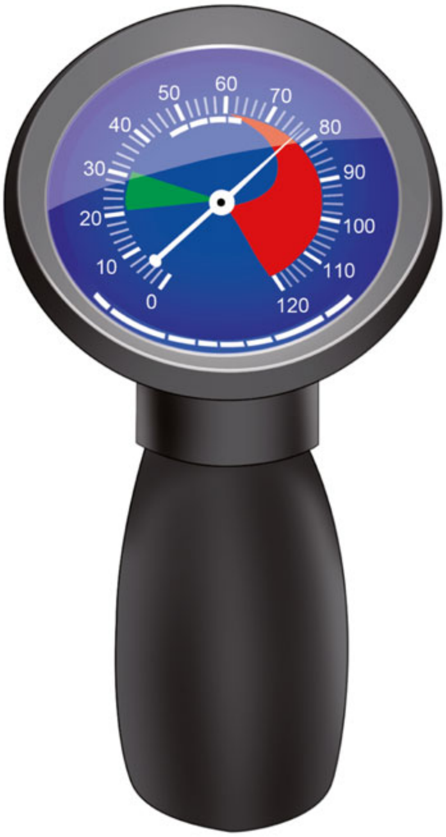

Fig. 77.4 Cuffalator

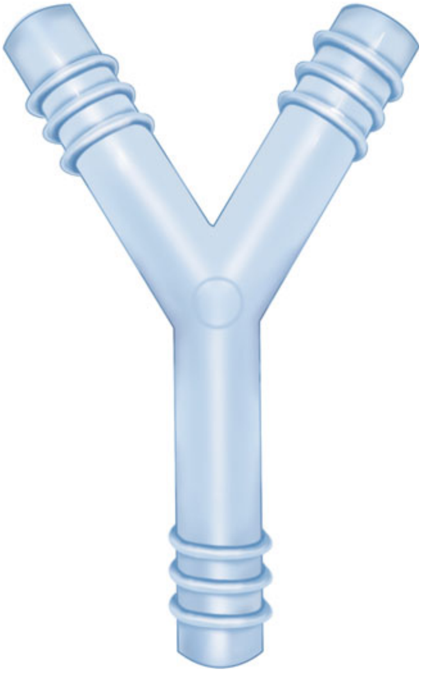

Fig. 77.5 Y-Tube connector

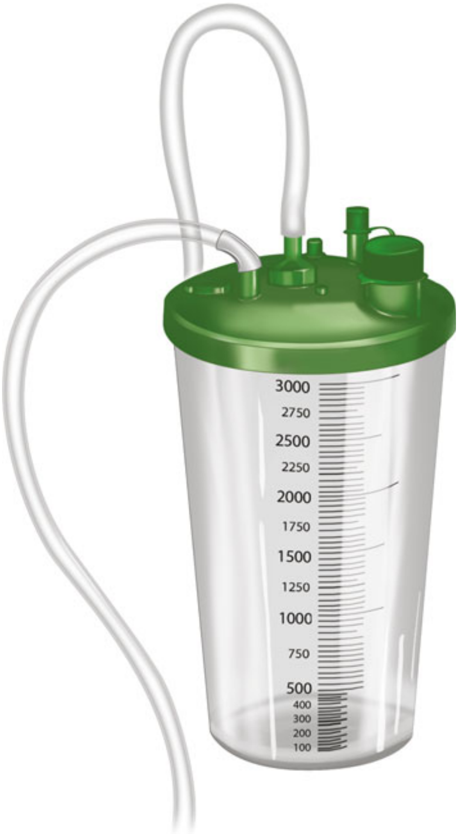
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
400
300
200
100

Fig. 77.7 Suction device and tubing

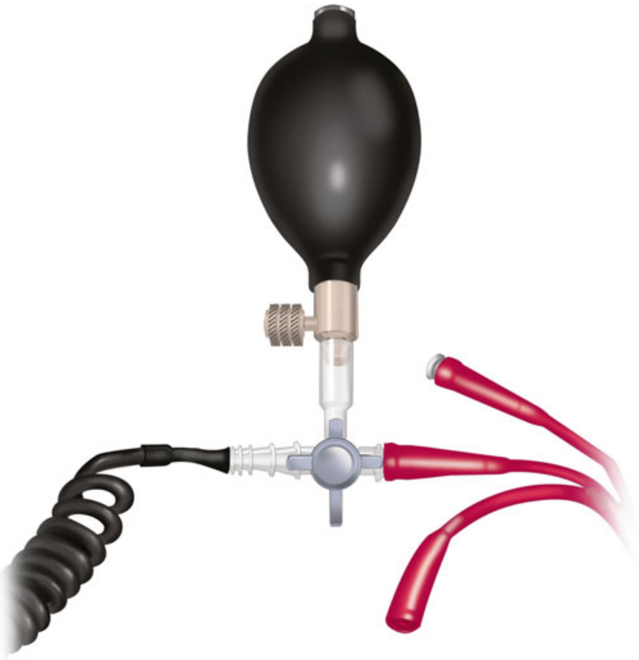

Fig. 77.6 3-way stop-valve tube connector

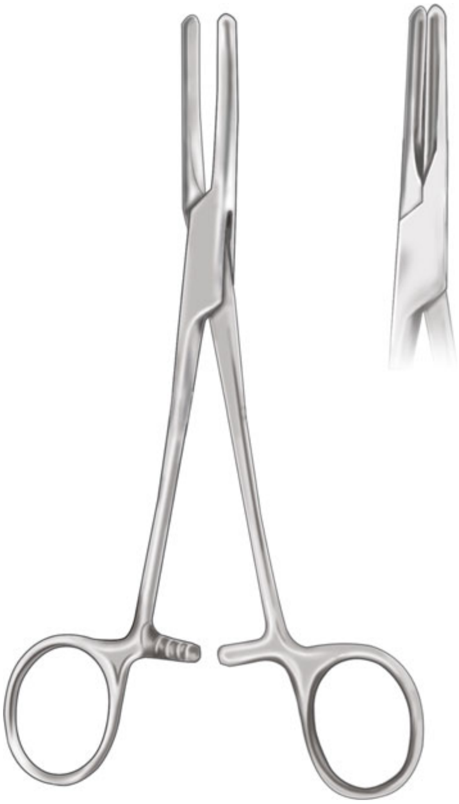

Fig. 77.8 Tube clamps

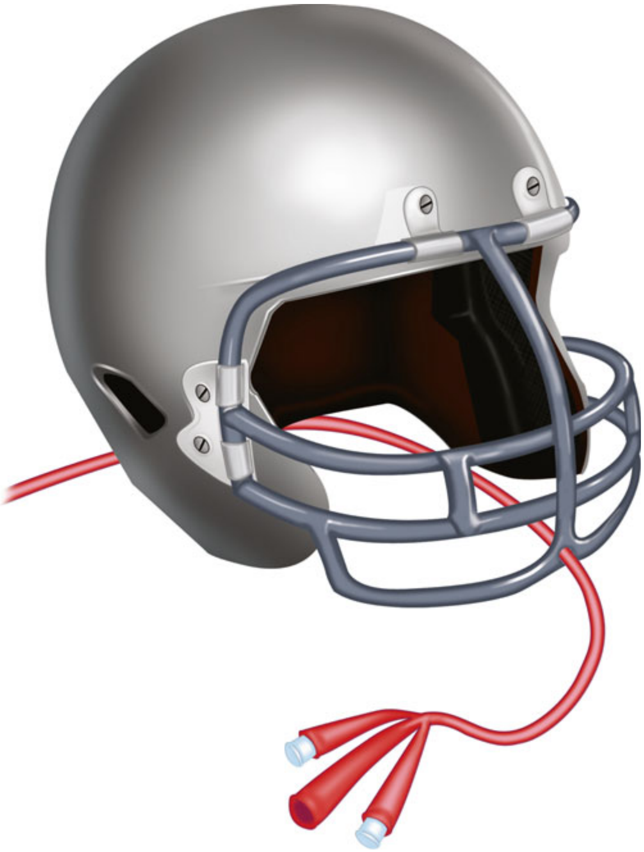

Fig. 77.9 Helmet traction setup

Fig. 77.10 Scissors

77.4 Procedure

1. Sedate and/or intubate the patient for adequate control of the patient during the procedure.
2. Ensure that the SB tube balloons are functional by inflating and deflating the balloons to ensure the absence of leaks.
3. Perform gastric lavage and irrigate the stomach with copious amount of sterile water.
4. Coat the distal and proximal portions of the SB tube with a thin layer of lubricating jelly or lidocaine gel. Spray the nasal passage with lidocaine spray.
5. Pass the SB tube via the nasogastric (NG) or the orogastric route (in intubated patients) to the 50-cm line. You may confirm placement with x-ray.
6. Inflate the gastric balloon to 200 mL of air and clamp the tube.
7. Apply gentle traction of 1–2 lb (0.4–2 kg) of force until it is felt that the gastric balloon has lodged at the gastroesophageal junction (Fig. 77.11).
8. Secure the tube to an anchor (e.g., football helmet or catcher's mask) placed on the patient's head.
9. Aspirate and lavage the gastric aspiration port. If it is clear of blood, do not inflate the esophageal balloon.
10. If it is not clear of blood, connect the esophageal tube of the SB tube to the sphygmomanometer/cuffalator using the 3-way stop-valve device (Fig. 77.12). You may use the Y-Tube connector instead.
11. Inflate the esophageal balloon to the lowest pressure determined to stop bleeding, typically 20–45 mmHg. Clamp the balloon.
12. Place an NG tube until it is felt overlying the top of the esophageal balloon of the SB tube. Check for further proximal esophageal bleed through aspiration and gentle lavage. Attach this NG tube to intermittent section to aid in the clearance of secretions.
13. Obtain a portable radiograph to confirm the position of the SB tube.
14. The esophageal tube should be at the lowest pressure that prevents bleeding and kept inflated for 24 h or until other definitive treatment is obtained.

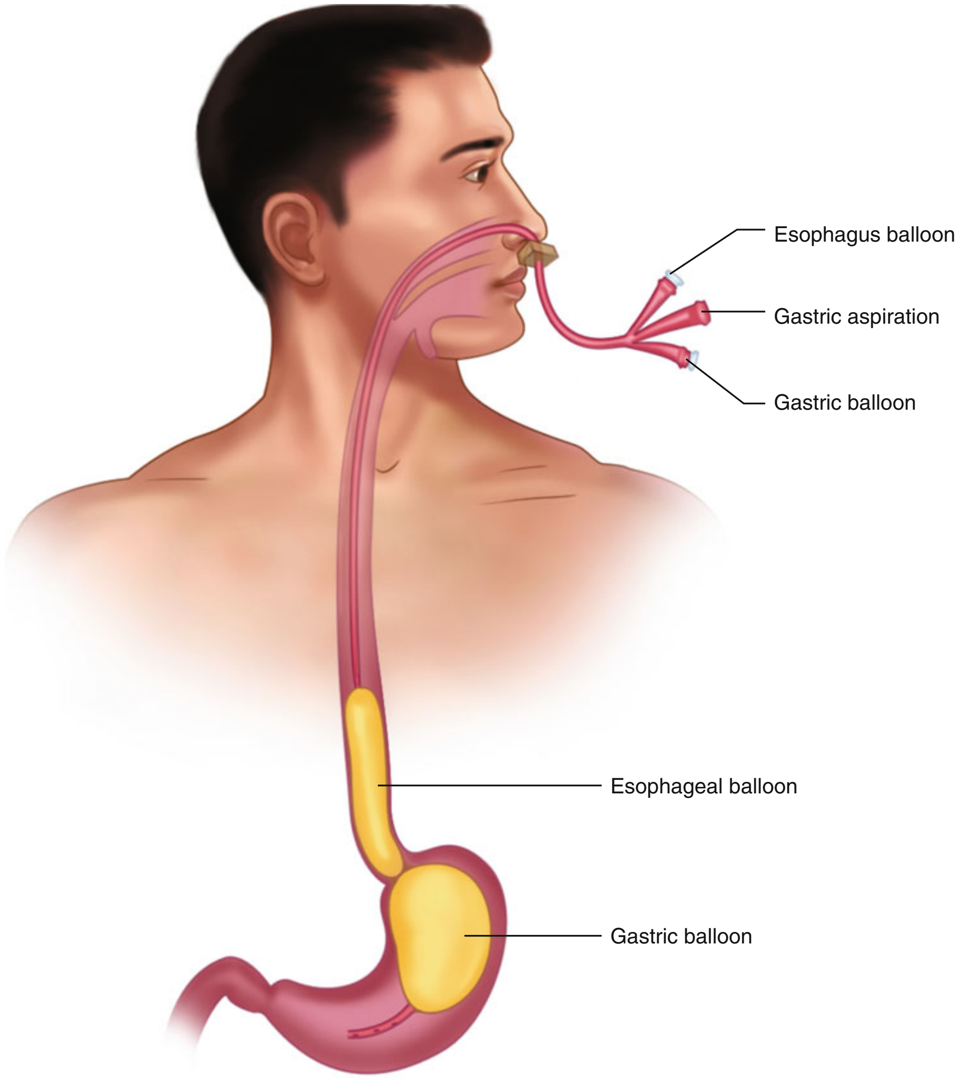
Esophagus balloon
Gastric aspiration
Gastric balloon
Esophageal balloon
Gastric balloon

Fig. 77.11 Proper placement of the balloons

Fig. 77.12 Connection of the sphygmomanometer to the esophagus balloon port

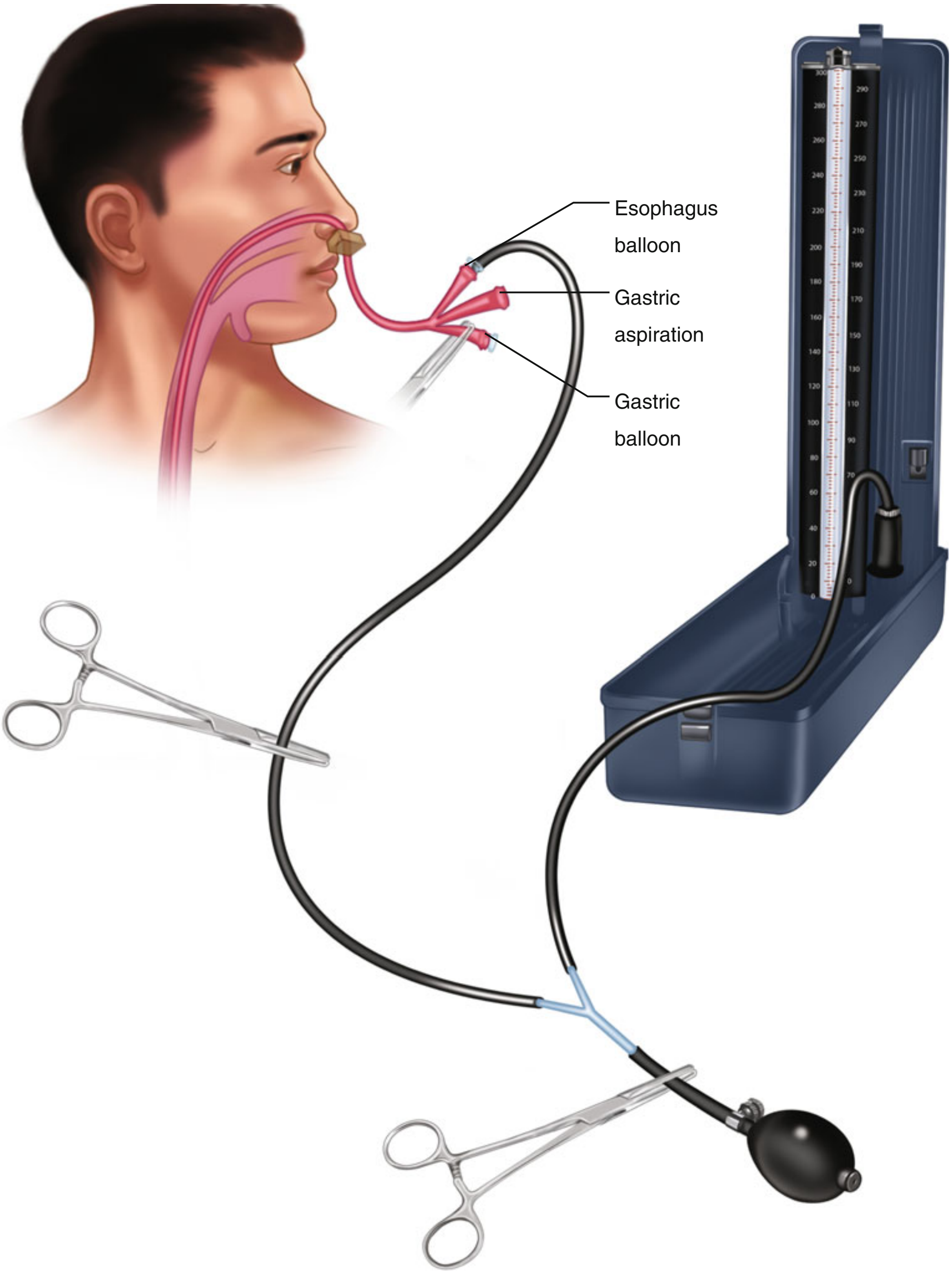
Esophagus balloon
Gastric aspiration
Gastric balloon

77.5 Complications

- Esophageal rupture occurs owing to esophageal erosion and necrosis owing to a balloon tamponade effect on tissue perfusion or overzealous balloon inflation.
- Airway obstruction owing to gastric tube deflation or failure, allowing esophageal tube to move up and occlude airway. Keep scissors near the patient to cut the SB tube lumens and remove the tube as necessary.
- Regurgitation and aspiration pneumonia from failure to adequately suction oropharyngeal secretions.

77.6 Pearls and Pitfalls

- Pearls
  - The esophageal balloon should not be inflated if the gastric balloon alone stops the bleeding.
  - Never inflate the esophageal balloon without the inflating the gastric balloon first. This will prevent it from slipping proximally into the oropharynx and obstructing the airway.
  - Nausea, vomiting, or aspiration is highly likely to occur. Use antiemetics and lavage the stomach before the procedure.
  - Intubate if there is airway compromise or risk of aspiration into the lungs.
  - Inflate the esophageal balloon only to the minimum pressure necessary to stop the variceal bleeding.
  - Using a catcher's mask may be more practical and comfortable for the recumbent patient.
- Pitfalls
  - The SB tube may induce hiccups.

Selected Reading

Bauer J, Kreel I, Kark A. The use of the Sengstaken-Blakemore tube for immediate control of bleeding esophageal varices. Ann Surg. 1974;179:273–7.

Henneman PL. Gastrointestinal bleeding. In: Rosen P, Barkin RM, editors. Emergency medicine. 6th ed. St. Louis: Mosby; 1998.

Remonda G, Morachioli N, Petruzzelli C. The use of the Sengstaken-Blakemore tube for immediate control of bleeding esophageal varices. Ann Osp Maria Vittoria Torino. 1981;24:115–20.

Sengstaken RW, Blakemore AH. Balloon tamponage for the control of hemorrhage from esophageal varices. Ann Surg. 1950;131:781–9.

Treger R, Graham T, Dea S. Sengstaken-Blakemore tube. Available at http://emedicine.medscape.com/article/81020-overview#a01. Accessed 18 May 2014.

Gastrostomy Tube Placement

78

Nathaniel Lisenbee and Latha Ganti

There are several types of gastrostomy tubes and related procedure variations:

- Procedure variations:
  - Open gastrostomy tube (G tube)
  - Percutaneous endoscopic gastrostomy (PEG) tube
  - Laparoscopic G tube
- Tube types:
  - PEG tube (Fig. 78.1)
  - Malecot tube (Fig. 78.2)
  - Balloon G tube (Fig. 78.3)
  - Low-profile G tube (nonobturated, button) (Fig. 78.4)
  - Low-profile G tube (obturated, button) (Fig. 78.5)

N. Lisenbee, MD
U. S. Air Force Medical Center Keesler, Biloxi, MI, USA
e-mail: nate4085@ufl.edu

L. Ganti, MD (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_78

Fig. 78.1 Percutaneous endoscopic gastrostomy (PEG) tube

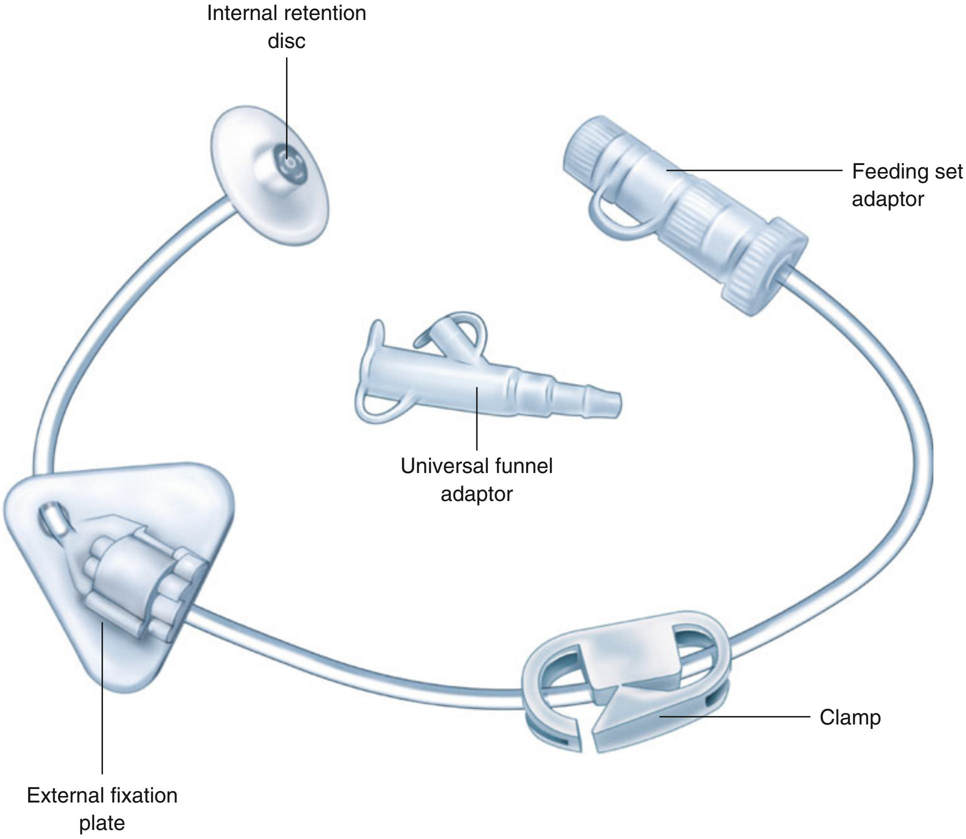
Internal retention disc
Feeding set adaptor
Universal funnel adaptor
External fixation plate
Clamp

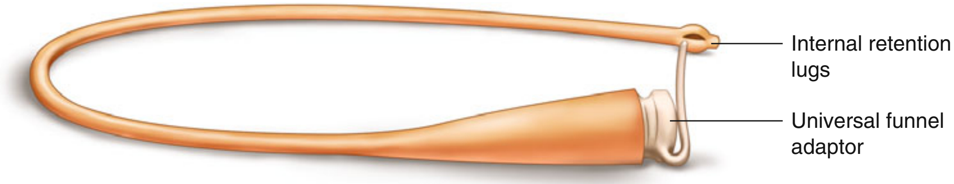
Internal retention lugs
Universal funnel adaptor

Fig. 78.2 Malecot tube

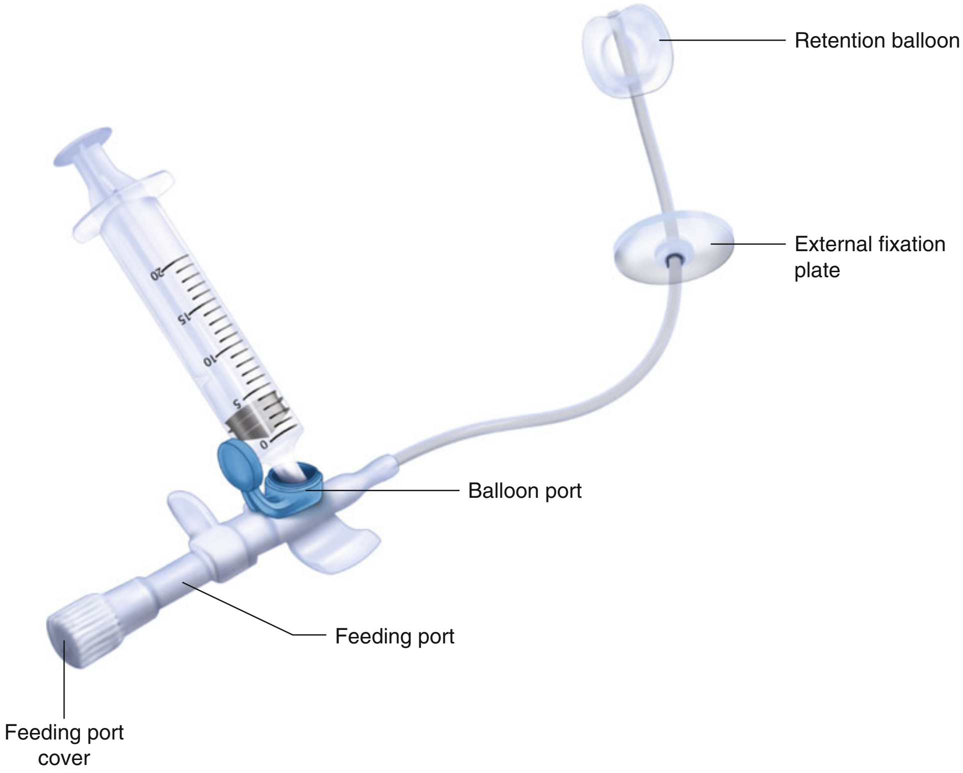
Retention balloon
20
15
10
5
0
External fixation plate
Balloon port
Feeding port
Feeding port cover

Fig. 78.3 Balloon gastrostomy tube

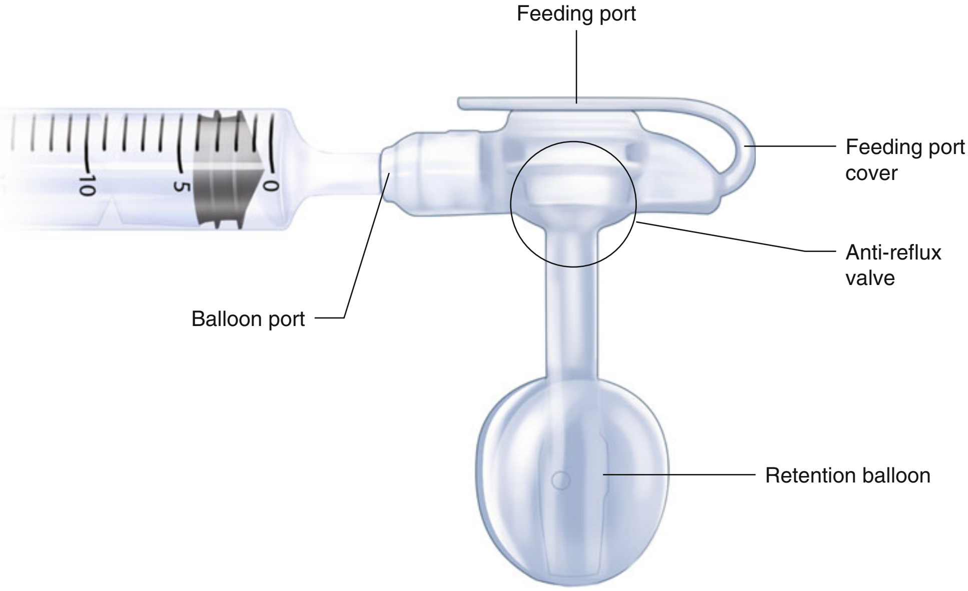
Feeding port
10
5
0
Feeding port cover
Anti-reflux valve
Balloon port
Retention balloon

Fig. 78.4 Low-profile gastrostomy tube (nonobturated, button)

Fig. 78.5 Low-profile gastrostomy tube (obturated, button)

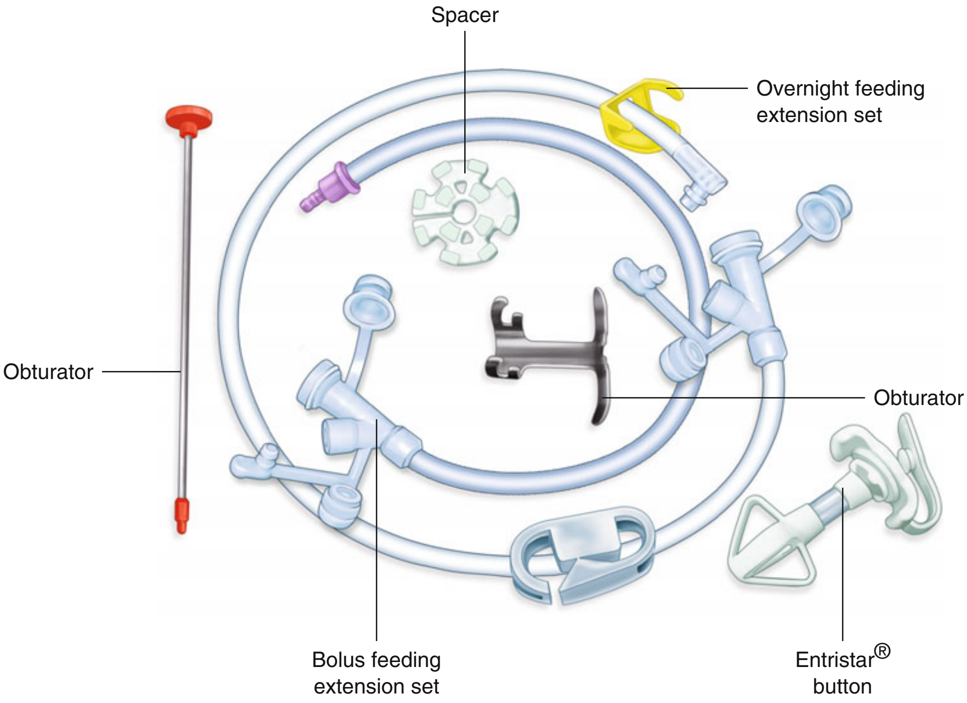
Spacer
Overnight feeding extension set
Obturator
Obturator
Bolus feeding extension set
Entristar® button

78.1 Indications

- Inability to swallow
  - Neurological deficit
  - Head trauma
  - Facial burns
  - Decreased mental status
- Need for gastric decompression
  - Gastric outflow obstruction
  - Small bowel obstruction (SBO), ileus, or volvulus
  - Intra-abdominal malignancy

78.2 Contraindications

- Absolute
  - Peritonitis
  - Ascites
- Relative
  - Hemodynamic instability
  - Coagulopathy
  - Abdominal wall infection at surgical site
  - History of gastric resection
  - Portal hypertension
  - Gastric varices

78.3 Materials and Medications

- Materials
  - Gloves
  - Lubricant
  - G tube (commercial kit) or Foley catheter
  - Sterile saline
  - External bolster
  - Multiple syringes
  - Suture material
  - Needle driver
  - Scissors
  - Stethoscope
- Medications
  - PEG tube placement
    - Moderate sedation (e.g., propofol, midazolam, fentanyl)
    - Local anesthesia (e.g., lidocaine, bupivacaine)
  - Open or laparoscopic G tube placement (not an emergency department procedure)
    - General anesthesia

78.4 Procedure (PEG Tube Replacement)

1. If the G tube is only partially removed upon patient presentation, the tube must first be removed.
2. Remove the G tube by deflating the balloon and pull gently on the tube while applying pressure to the abdominal wall at the surgical site.
3. The G tube should slide out easily with gentle traction, and the procedure should be discontinued if it does not.
4. Initially, it is important to assess the tract to determine the size and potential need for dilation.
5. If necessary for tube passage, the tract can be dilated with a cotton-tipped applicator or hemostat; however, be sure to dilate gently because it is possible to create a false tract.
6. Once the tract has been assessed, obtain the appropriate tube for replacement.
7. Initially, attempt replacement of the patient's G tube with an identical tube.
8. If an identical tube is not available, attempt placing a small Foley catheter in the tract to ensure that it stays open.
9. To initiate placement, place lubricant on the tube and carefully advance the tube into the tract.
10. Once the tube is in place, secure an air-filled syringe to the tip of the tube and insufflate a small amount of air into the stomach while auscultating to confirm passage of air into the stomach.
11. Secondary procedural confirmation can be performed by aspirating gastric contents from the tube.
12. Once placement is verified, do not forget to inflate the balloon with saline and then pull the tube backward until it abuts the inside of the stomach wall (Fig. 78.6).
13. Finally, it is very important to secure the external portion of the tube to prevent the tube from being lost into the stomach.
14. Commercial G tubes are accompanied by a bolster made specifically for the specific type of G tube in order to provide security of tube placement.

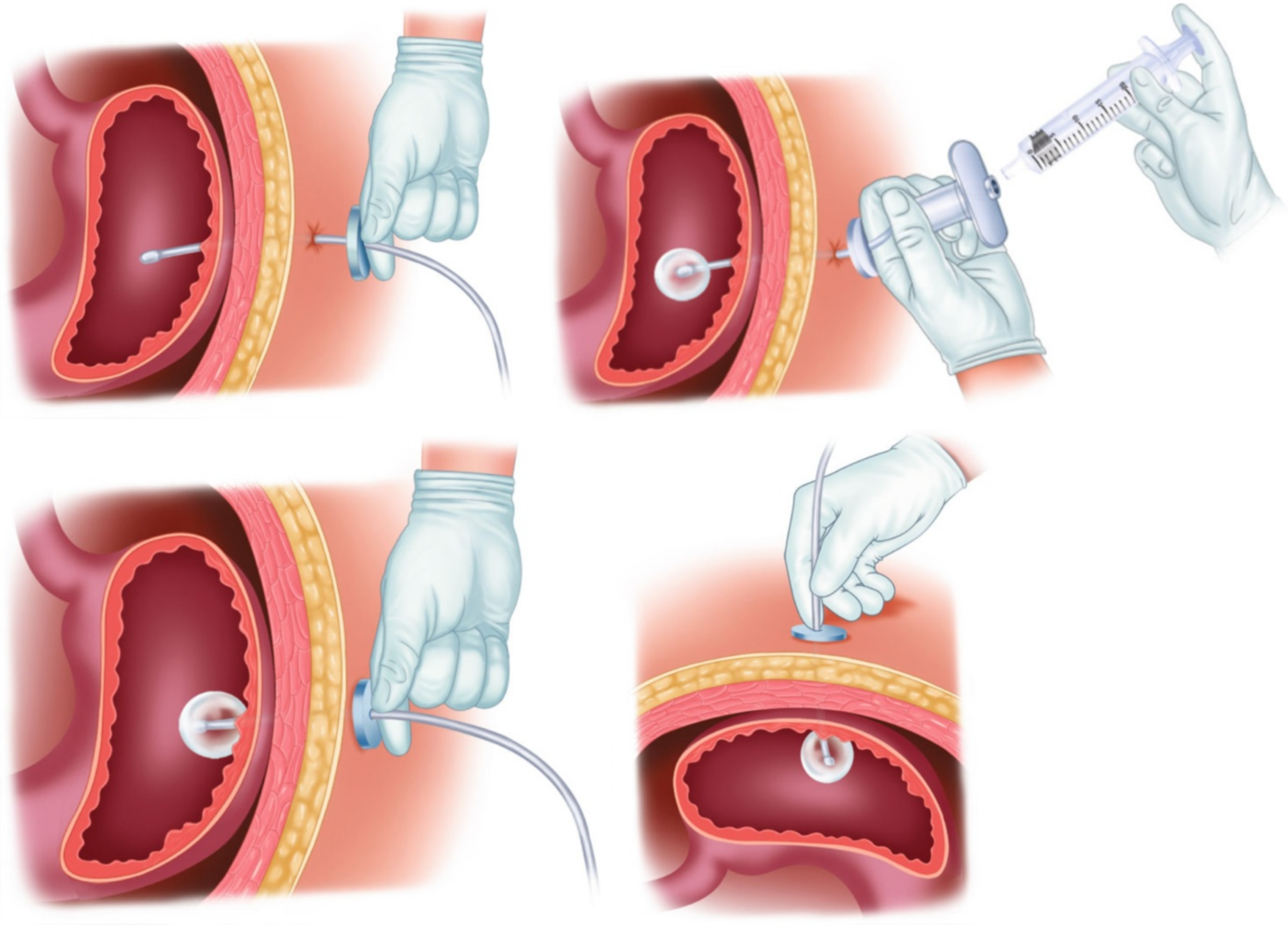

Fig. 78.6 Placement of low-profile gastrostomy tube

15. If a Foley catheter is used to maintain tract patency, an external bolster must be created using the following steps:
    - Trim a 2- to 3-in. portion from the tip of the catheter.
    - Cut two small holes just above each other on the 2- to 3-in. trimmed catheter portion.
    - Slide the external portion of the Foley catheter through both holes of the bolster.
    - Slide the bolster down the Foley catheter to the abdominal wall.
    - Secure the two ends of the bolster to the skin in order to maintain patency of the Foley catheter.
    - An interventional radiologist may also be contacted to advance the tube over a wire under fluoroscopic guidance.
16. Radiographic confirmation.
    - Typically, G tube placement is confirmed by injecting approximately 20 mL of Gastrografin® (diatrizoate meglumine, diatrizoate sodium) water-soluble contrast into the G tube, followed shortly by an abdominal x-ray.
    - Proper placement will result in an abdominal radiograph showing contrast outlining the stomach (Fig. 78.7).

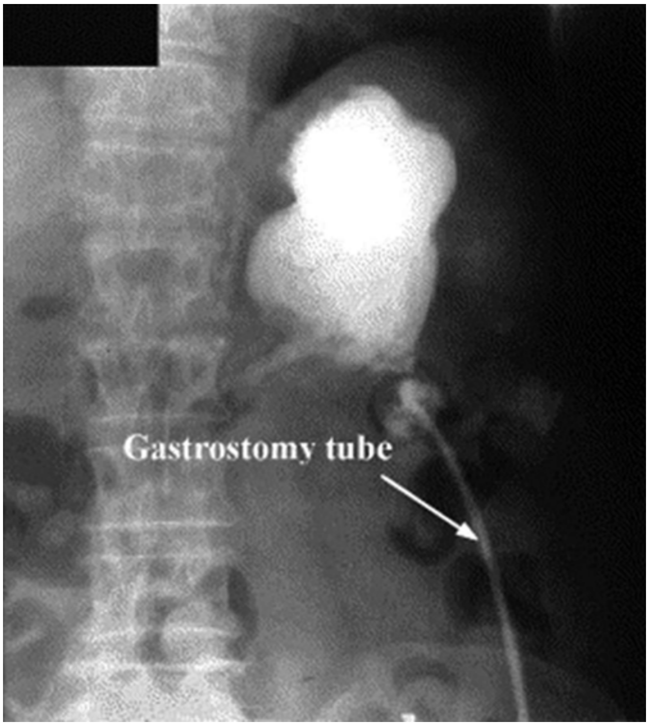

Fig. 78.7 An anteroposterior abdominal radiograph after PEG tube placement and injection of 25 mL of gastrografin. The tube can be seen projecting up the left side of the abdomen, and contrast medium appears to enter the stomach lumen. The balloon is not visualized in the stomach (Reproduced with permission from: Burke DT, El Shami A, Heinle E, and Pina BD. Comparison of gastrostomy tube replacement verification using air insufflation versus gastrografin. Archives of physical medicine and rehabilitation. 2006;87(11): 1530–3.)

78.5 Pearls and Pitfalls

- Pearls
  - Even if a patient's G tube is only partially removed upon presentation to the emergency department, the tube likely needs complete removal and replacement.
  - Most G tubes are easily removed at the bedside; however, some are not able to be safely removed in the emergency department setting. Thus, if the tube does not withdraw easily, attempt to contact the proceduralist who placed the tube in order to inquire about the best method for removal.
- Pitfalls
  - Never use barium contrast when confirming G tube placement radiographically because barium can cause significant intra-abdominal damage if accidentally injected in the intraperitoneal cavity.

78.6 Complications

- Aspiration
- Surgical site infection
- Bleeding
- Pneumoperitoneum
- Accidental perforation of the colon or small bowel
- Tube dislodgment
- Peritonitis

Selected Reading

Arora G. Medscape: percutaneous endoscopic gastrostomy tube placement. Retrieved 3 Jan 2013, from http://emedicine.medscape.com/article/149665-overview#a09.

Gauderer MW, Ponsky JL, Izant Jr RJ. Gastrostomy without laparotomy: a percutaneous endoscopic technique. J Pediatr Surg. 1980; 15:872–5.

Great Ormand street hospital for children website. Retrieved 3 Jan 2013, from www.gosh.nhs.uk/EasySiteWeb/GatewayLink.aspx?alId=102263.

Sarani B. Percutaneous endoscopic gastrostomy tube placement. In: Falter F, editor. Bedside procedures in the ICU. Philadelphia: Springer; 2012. p. 113–22.

Tawa Jr NE, Fischer JE. Metabolism in surgical patients. In: Townsend Jr CM, Beauchamp RD, Evers BM, Mattox KL, editors. Sabiston textbook of surgery. 18th ed. Philadelphia: Elsevier; 2007.

Paracentesis

79

Shalu S. Patel and Bobby K. Desai

79.1 Indications

- Diagnosis of infection in ascites
- Diagnosis of malignant ascites
- Diagnosis of hemoperitoneum in traumas
- Relief of abdominal pressure/pain or respiratory compromise secondary to ascites

79.2 Contraindications

- Severe coagulopathy
  - Prothrombin time (PT) > 21 s
  - International normalized ratio (INR) > 1.6
  - Platelets < 50,000/mm^3
- Skin infection over the needle insertion site
- Acute abdomen that requires surgery
- Pregnancy
- Distended bowel
- Intra-abdominal adhesions

79.3 Materials and Medications

- 18- to 22-gauge 1.5- to 3.5-in. needle or angiocatheter, 25-gauge needle
- Lidocaine 1 or 2 % (10 mL)
- Syringes
  - 10 mL (1), 50 mL (2)
- 1-L vacuum bottle (4) (if therapeutic tap)
- Thoracentesis kit tubing or any high-pressure connection tubing (if therapeutic tap)
- Sterile gloves
- Surgical pen (recommended)
- Povidone-iodine (Betadine) or other skin antiseptic
- Sterile drape
- Sterile gauze (4 × 4)
- Band-Aid
- Bedside ultrasound (recommended)

79.4 Procedure (Fig. 79.1)

1. Position the patient supine. If possible, adjust the head of the bed to make a 45° angle to help the fluid accumulate in the pocket. Sometimes, it may also be beneficial to have the patient lie recumbent toward the site of drainage.
2. Scan the abdomen with an ultrasound to determine whether there is a pocket of fluid that can be drained. This also allows the physician to see how far the needle needs to be inserted and how deep it can be placed without risking injury to the bowel (Fig. 79.2).
3. Mark the optimal needle insertion site with a surgical pen.
4. Prepare the skin and drape in a sterile fashion.
5. Using lidocaine, anesthetize the appropriate area subcutaneously and then continue to insert the needle, and inject anesthetic through the deeper tissues until ascitic fluid can be drawn back.
6. Withdraw the needle.
7. When ready for the paracentesis, stretch the skin caudad and insert the needle or angiocatheter (connected to a syringe) while aspirating. Then, release the skin and continue to insert the needle or angiocatheter through the peritoneal wall until fluid is retrieved. This will create a "Z-track" that will decrease leakage of peritoneal fluid through the skin.

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA
e-mail: shalu314@gmail.com

B.K. Desai, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_79

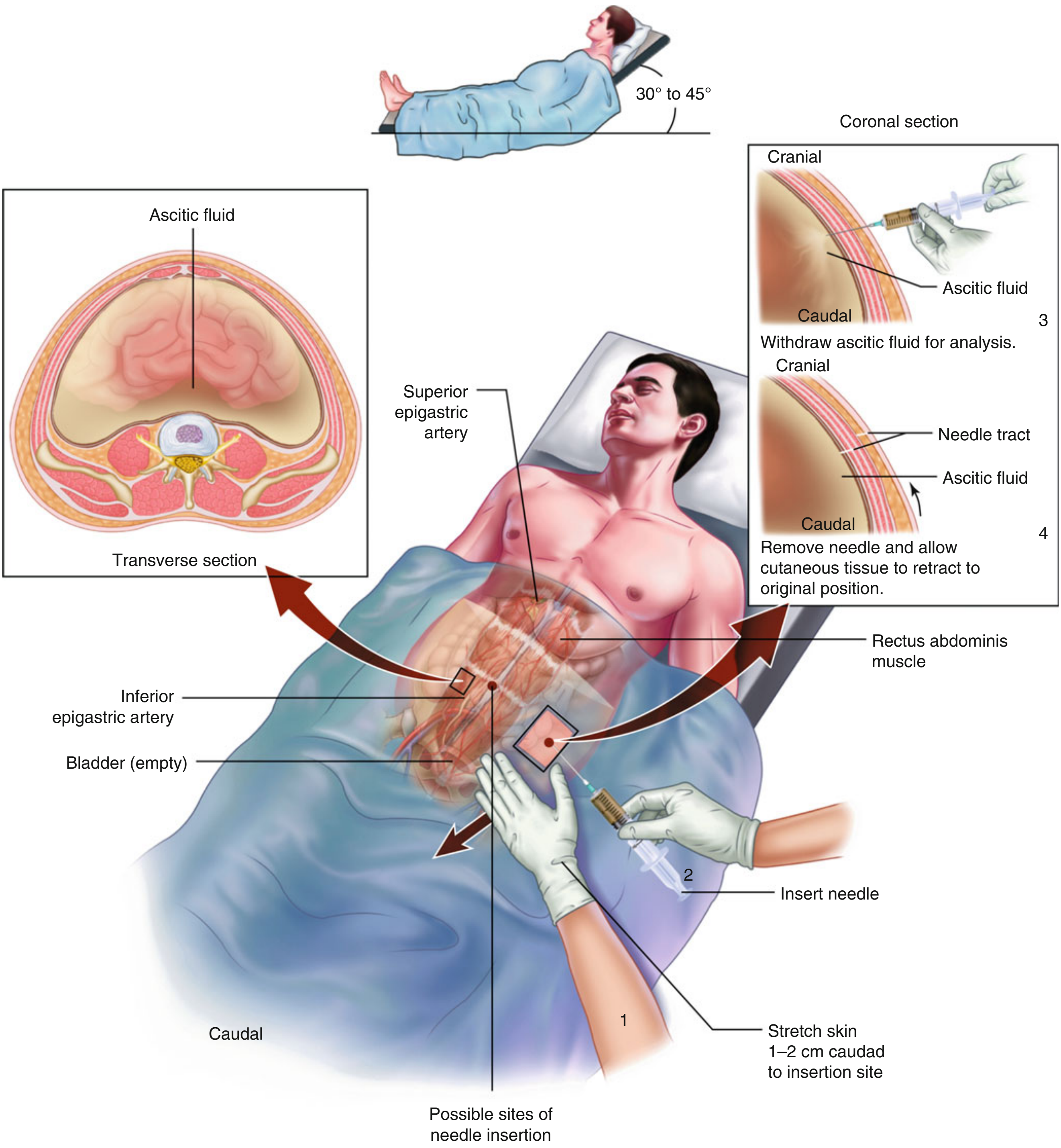
30° to 45°
Coronal section
Cranial
Ascitic fluid
Caudal
3
Withdraw ascitic fluid for analysis.
Cranial
Needle tract
Ascitic fluid
Caudal
4
Remove needle and allow cutaneous tissue to retract to original position.
Ascitic fluid
Transverse section
Superior epigastric artery
Rectus abdominis muscle
Inferior epigastric artery
Bladder (empty)
2
Insert needle
1
Caudal
Stretch skin 1–2 cm caudad to insertion site
Possible sites of needle insertion

Fig. 79.1 Paracentesis procedure

8. Once fluid is retrieved, push in the catheter and remove the needle portion (if used) or hold the needle steady.
9. Aspirate from the catheter to ensure that it is in the appropriate location.
10. If fluid easily is aspirated, unscrew the syringe and connect a 50-mL syringe to the needle or catheter and fill it with fluid. This may be done twice. Alternatively, if the procedure is done for therapeutic purposes, attach the tubing that is already connected to the vacuum bottle to the catheter and allow the vacuum to withdraw fluid into the collection bottles.
11. If fluid cannot be aspirated easily, the catheter can be repositioned further in the pocket or turned by 45° sequentially as needed.

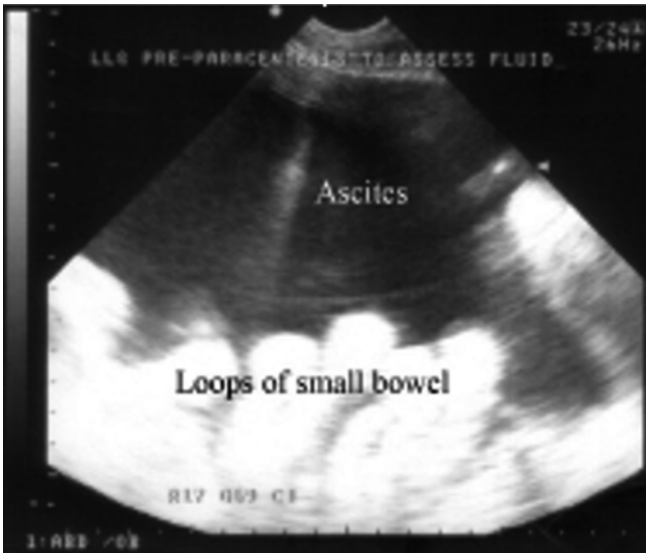

Fig. 79.2 Ultrasound to determine whether there is a pocket of fluid that can be drained

12. Once the fluid is aspirated, pull out the needle or angiocatheter and hold pressure with gauze. Bleeding should be minimal.
13. Place a Band-Aid or other dressing over the site.
14. Send the fluid to the laboratory. Generally, laboratory analyses include protein, albumin, specific gravity, glucose, bilirubin, amylase, lipase, triglyceride, lactate dehydrogenase (LDH), cell count and differential, culture and sensitivity (C&S), Gram stain, acid-fast bacillus (AFB), fungal culture, cytology, and pH.

79.5 Complications

- Persistent leakage from the needle insertion site
- Abdominal wall hematoma
- Bowel perforation
- Introduction of infection
- Hypotension (after a large-volume paracentesis)
- Dilutional hyponatremia
- Hepatorenal syndrome
- Bleeding
- Postparacentesis circulatory dysfunction

79.6 Pearls and Pitfalls

- Pearls
  - The preferred site of entry is in the midline of the abdomen, below the umbilicus.
  - The serum-ascites albumin gradient (SAAG) can be used to identify the cause of the ascites. It is calculated by subtracting the albumin concentration in the ascites from the albumin concentration in the serum. A high gradient (>1.1 g/dL) suggests portal hypertension, whereas a low gradient (<1.1 g/dL) suggests other causes.
  - Postparacentesis circulatory dysfunction (PPCD) occurs secondary to hypovolemia after large-volume paracentesis (>4 L) in cirrhotic patients. It is associated with worsening hyponatremia, renal dysfunction, shorter time to ascites recurrence, and increased mortality. Prevention of PPCD has been demonstrated with the administration of 6–8 g of albumin per liter of ascites removed.
- Pitfalls
  - Polymorphonuclear lymphocyte (PMN) count greater than 250/mm^3 is diagnostic of spontaneous bacterial peritonitis.

Selected Reading

Ginès P, Cárdenas A, Arroyo V, Rodés J. Management of cirrhosis and ascites. N Engl J Med. 2004;350:1646–54.

Ginès P, Tito L, Arroyo V, et al. Randomized comparative study of therapeutic paracentesis with and without intravenous albumin in cirrhosis. Gastroenterology. 1988;94:1493–502.

Ruiz del Arbol L, Monescillo A, Jimenéz W, et al. Paracentesis-induced circulatory dysfunction: mechanism and effect on hepatic hemodynamics in cirrhosis. Gastroenterology. 1997;113:579–86.

Runyon BA. Paracentesis of ascitic fluid. A safe procedure. Arch Intern Med. 1986;146:2259–61.

Wong CL, Holroyd-Leduc J, Thorpe KE, Straus SE. Does this patient have bacterial peritonitis or portal hypertension? How do I perform a paracentesis and analyze the results? JAMA. 2008;299:1166–78.

Anal Fissure Management

80

David P. Nguyen, L. Connor Nickels, and
Giuliano De Portu

80.1 Indications

- An anal fissure is a small ulcer of the mucosa at the anal verge (Fig. 80.1).
- It is the most common cause of intense sudden rectal bleeding.
- Posterior midline anal fissures are the most common type (90 %).
  - Mostly found in young adults (30–50 y) but can occur at any age.
  - Usually associated with constipation (firm, large-caliber, painful bowel movements) or chronic diarrhea.
  - Most uncomplicated fissures resolve in 3–4 weeks.
- Can be extremely painful, during and after defecation.
- Classified as acute or chronic.
- Now believed to be caused by reduced anal blood flow in the posterior midline, anal sphincter hypertonia, and thus mucosal ischemia.

D.P. Nguyen, DO
Department of Emergency Medicine, Rush-Copley Medical Center, Aurora, IL, USA
e-mail: davidpnguyen@yahoo.com

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_80

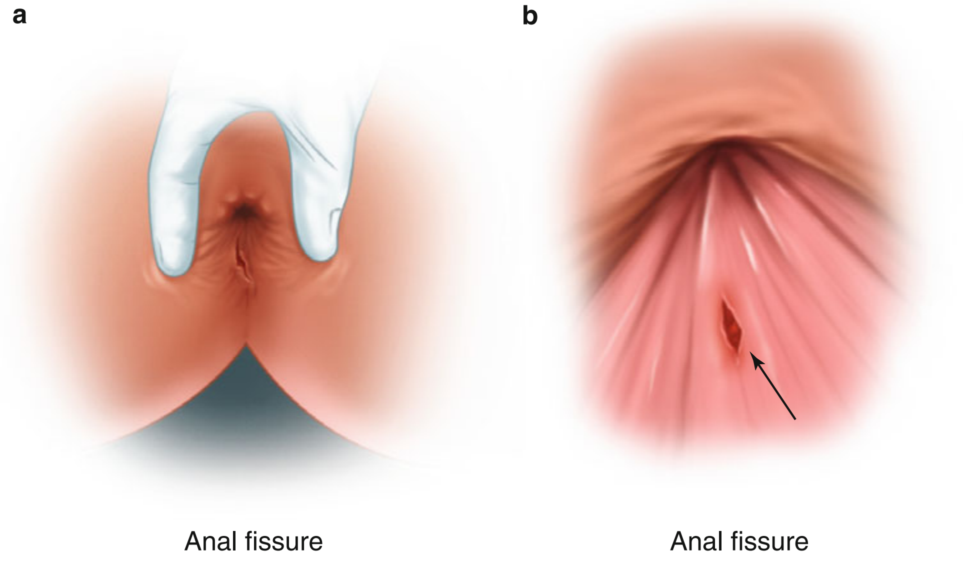
a
b
Anal fissure
Anal fissure

Fig. 80.1 (**a**, **b**) Anal fissures

80.2 Contraindications

- Digital rectal examination should be avoided unless the diagnosis is in doubt.
- Surgical procedures are generally reserved for when medical management has failed after 1–3 months of treatment.

80.3 Materials and Medications

- Standard precautions barrier protection for the provider.
- Good light source.
- Optional emergency department treatments:
  - Topical anesthetic/preparation (Anusol [pramoxine hydrochloride; zinc oxide] with cortisone).
  - Nitroglycerin (0.2 %) or nifedipine gel (2 %) is second-line therapy (relaxes muscles and promotes blood flow).

80.4 Procedure

1. In a private, calm environment, gently spread the buttocks for complete visual inspection.
   - This may cause an increase in the patient's pain and spasming.
   - If a fissure is clearly identified, stop here.
2. Apply topical anesthetic/preparation for symptomatic relief (optional, as the physician may want to just start with the treatments that follow).
3. Discharge the patient with conservative therapy management.
4. In acute anal fissures (onset of 3–6 weeks), medical management is indicated along with dietary modifications (WASH regimen [warm baths, analgesia, stool softeners, high-fiber diet]):
   - Warm sitz baths.
     - Usually 20 min soaking each time
     - Recommended after every bowel movement
     - At least twice per day if not having regular bowel movements
   - High-fiber diet with fiber supplements.
   - Increase fluid intake.
   - May add stool softeners, if needed.
   - If chronic or the previous regimen has been exhausted, one of the following may be considered:
     - 0.2–0.4 % nitroglycerin cream applied to anal area
       - May cause headache
       - Recommend wearing a glove to prevent absorption through digital skin
     - Calcium channel blockers:
       - Topical nifedipine
       - 2 % diltiazem cream
     - Botulinum toxin A injection:
       - Controversial; may have poorer success rates than surgery
5. Provide surgical referral for nonhealing wounds.
   - Lateral internal sphincterotomy is the surgical procedure of choice.

80.5 Complications

- Infection
- Abscess
- Bleeding
- Chronic fissure formation
- Postsurgical fecal incontinence

80.6 Pearls and Pitfalls

- Pearls
  - Multiple or recurrent fissures are associated with Crohn's disease, tuberculosis, syphilis, human immunodeficiency virus (HIV), and malignancy.
- Pitfalls
  - Suspect child abuse if an anal fissure is found in a child.

Selected Reading

Feldman M, Friedman LS, Brandt LJ, editors. Sleisenger and Fordtran's gastrointestinal and liver disease. 9th ed. Philadelphia: Saunders; 2010.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2010. Chapter 94: Disorders of the anorectum

Oztürk H, Onen A, Dokucu AI, Otçu S, Yağmur Y, Yucesan S. Management of anorectal injuries in children: an eighteen-year experience. Eur J Pediatr Surg. 2003;13:249–55.

Townsend CM Jr, Beauchamp RD, Evers BM, Mattox KL, editors. Sabiston textbook of surgery. 19th ed. Philadelphia: Saunders; 2012. Chapter 53: Anus

Part XI

Genitourinary Procedures

Bladder Catheterization

81

Maritza A. Plaza-Verduin and Judith K. Lucas

81.1 Indications

- Obtaining a sterile urine specimen
- Preventing or relieving urinary retention
- Close monitoring of urine output for fluid balance with an indwelling urinary catheter
- Urgent cystourethrography
- Child with contusion or burns to the perineum and at risk for meatal swelling and obstruction to urine outflow
- Temporary measure to relieve lower urinary tract obstruction
- Neurogenic bladder
- Anesthesia-induced and/or surgery-induced urinary retention has occurred.

81.2 Contraindications

- Absolute
  - Potential urethral injury from trauma
    - Pelvic fractures
    - Known trauma to the urethra
    - Blood at the meatus
- Relative
  - Recent genitourinary surgery (consult with a urologist before placing a catheter)

81.3 Materials and Medications

- Bladder catheterization kit:
  - Sterile gloves
  - Sterile drapes
  - Povidone-iodine (Betadine) solution
  - Cotton sponges or applicators for sterilizing solution
  - Lubricant
  - Specimen collection cup
  - Catheter
    - 5-French feeding tube for neonates
    - 8-French catheters for Infants
    - 10- to 12-French catheters in older children
- Local anesthetic (if desired—2 % lidocaine hydrochloride jelly)
- Absorbent pad

M.A. Plaza-Verduin, MD
Pediatric Division, Department of Emergency Medicine, Arnold Palmer Hospital for Children, Orlando, FL, USA
e-mail: mari.plazav@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_81

81.4 Procedure

1. Inspect the urinary catheterization tray for all the appropriate materials.
2. Place the patient supine with an absorbent pad under the buttocks.
   (a) Girls should be placed in the frog-leg position (Fig. 81.1).
3. Before sterilizing the field, locate the urethral opening.
4. Remove any powder, ointments, or medicated creams the child might have on the perineum.
5. If needed, apply anesthetic to the area.
   (a) Soak a cotton ball with anesthetic (2 % lidocaine hydrochloride jelly) and hold over the urethra opening for 2 min.
   (b) 0.5–2.0 mL of anesthetic can also be injected into the urethra.
6. Sterilize the area and place drapes appropriately, exposing the genitalia.
7. Catheterization of males
   (a) If uncircumcised, gently retract foreskin, *if possible*, for cleaning and visualization of the meatus.
   (b) Hold the penis using the nondominant hand at a 90° angle from the body (Fig. 81.2).
   (c) Lubricate the catheter tip.
   (d) Insert the lubricated catheter into the meatal opening and advance it while applying gentle traction to the penis from the base of the penis.
   (e) If resistance is met, maintain gentle pressure with the catheter.
      - *Do not attempt to force the catheter that could create a false tract or traumatic fistula.*
   (f) Advance the catheter until urine is obtained, approximately inserting the catheter to just beyond the penile length.
   (g) Once completed, gently withdraw the catheter.
   (h) Clean the area, wiping away the Betadine solution.
   (i) If uncircumcised, pull the foreskin over the glans to avoid paraphimosis.
8. Catheterization of females
   (a) Sterilization of the area should occur from anterior to posterior.
   (b) Have an assistant hold the labia majora apart.
      (i) If no assistant is available, use the nondominant hand to hold the labia apart.
         - Holding the labia majora with a gentle outward, lateral, and upward traction will help visualize the meatus (Fig. 81.3).

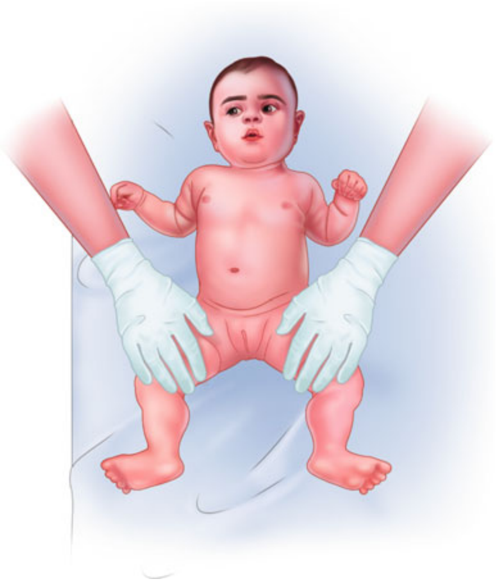

Fig. 81.1 Infant held in the frog-leg position for catheterization

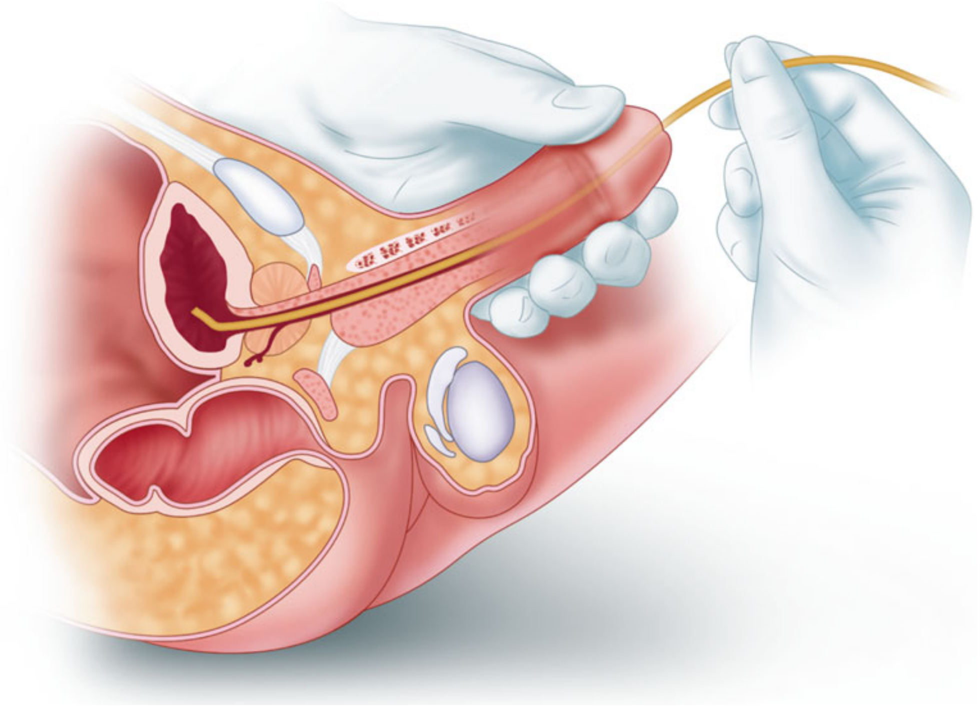

Fig. 81.2 Bladder catheterization of a male; penis should be held perpendicular to the body

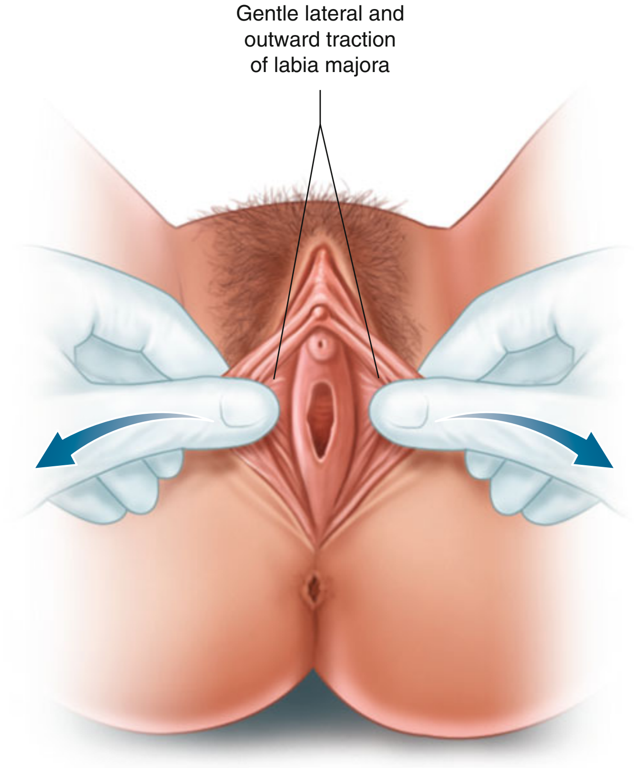
Gentle lateral and
outward traction
of labia majora

Fig. 81.3 Positioning of labia for better visualization of the meatus

- Downward displacement of the cephalad aspect of the vaginal introital fold with a cotton-tipped applicator can help visualize the urethral meatus (Fig. 81.4).

(c) Lubricate the catheter tip.

(d) Insert the lubricated catheter into the meatal opening. Advance slowly until urine is obtained (Fig. 81.5).

(e) Once completed, gently withdraw the catheter.

(f) Clean area, wiping away the Betadine solution.

Fig. 81.4 Better visualization of the meatus is achieved with downward displacement of the introital mucosa

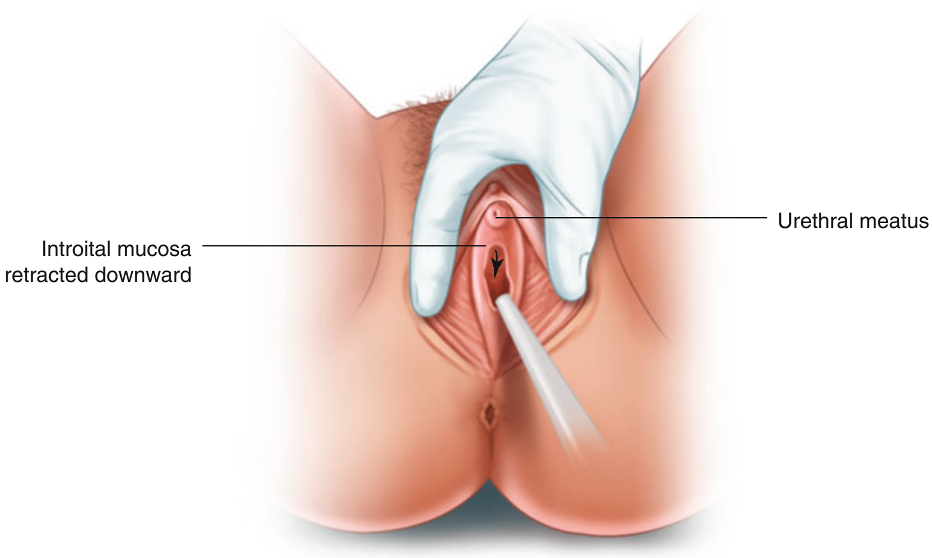

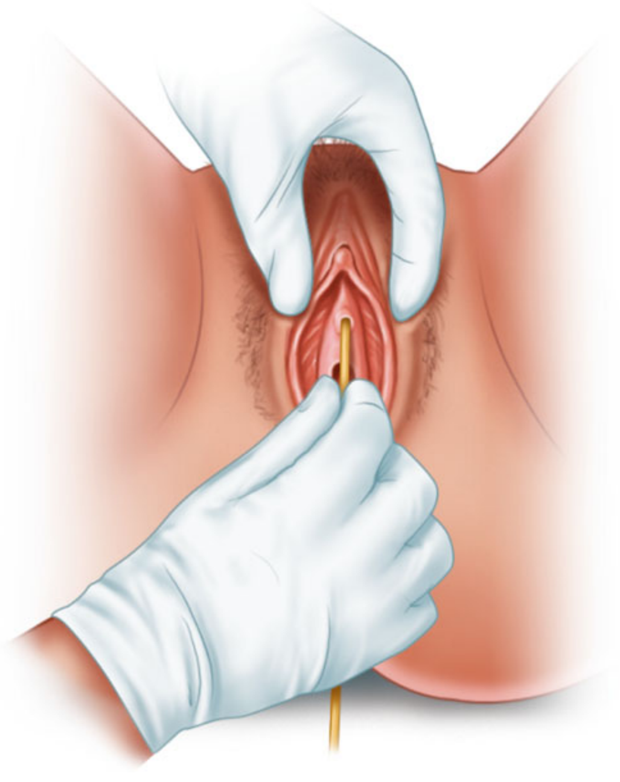

Fig. 81.5 Bladder catheterization of a female

81.5 Complications

- Urethral or bladder injury
- Infection if sterile field not maintained
- Paraphimosis owing to failure to restore a retracted foreskin to its normal position

81.6 Pearls and Pitfalls

- Pearls
  - It is not necessary to fully retract a foreskin. This only causes trauma and increases the likelihood of paraphimosis. As the infant/boy ages, the foreskin will loosen and the naturally occurring adhesions will spontaneously release.
  - The urethral meatus in an infant female is usually tucked just above the redundant hymen (as opposed to the more anteriorly located meatus in the adult woman) and often looks like a dimple or small blind pouch.
  - Using viscous lidocaine in lieu of, or blended with, lubricant anesthetizes the meatus and urethra as the catheter passes.
  - Have the specimen cup at the ready because, at times, once the (usually cold) Betadine or antiseptic solution is applied, the infant often releases the urine and it can be caught, literally, midstream.
  - In the uncircumcised male, be certain to return the foreskin over the glans to avoid paraphimoses.
- Pitfalls
  - If catheterizing a child in search of infection, send a urine culture regardless of the urinalysis results because the younger infants can have false-negative urinalysis and still have positive cultures.

Selected Reading

American Academy of Pediatrics, Subcommittee on urinary tract infection, steering committee on quality improvement and management. Urinary tract infection: clinical practice guideline for the diagnosis and management of the initial UTI in febrile infants and children 2 to 24 months. Pediatrics. 2011;128:595–609.

Beno S, Schwab S. Bladder catheterization. In: King C, Henretig FM, editors. Textbook of pediatric emergency procedures. 2nd ed. New York: Lippincott Williams & Wilkins; 2008.

Cheng YW, Wong SN. Diagnosing symptomatic urinary tract infection in infants by catheter urine culture. J Paediatr Child Health. 2005;41:437–40.

Gerard LL, Cooper CS, Duethman KS, et al. Effectiveness of lidocaine lubricant for discomfort during pediatric urethral catheterization. J Urol. 2003;170:564–7.

Kozer E, Rosenbloom E, Goldman D, et al. Pain in infants who are younger than 2 months during suprapubic aspiration and transurethral bladder catheterization: a randomized, controlled study. Pediatrics. 2006;118:e51–6.

Pelvic Examination and Wet Preparation

82

Nauman W. Rashid, Elaine B. Josephson, and Muhammad Waseem

82.1 Indications

- Lower abdominal or pelvic pain
- Vaginal bleeding or discharge
- Cancer screening
- Pregnancy
- Exposure to sexually transmitted disease
- Sexual assault

82.2 Contraindications

- Physical or mental disability
- Recent gynecological surgery
- Third-trimester pregnancy with bleeding
- Premenstrual females (may not be indicated in adolescents, who are not sexually active, unless there is discharge, bleeding, suspicion for abuse, or a foreign body)
- If a speculum examination is necessary, examination under general anesthesia should be considered.

82.3 Materials and Medications

- Examination table with stirrups (Fig. 82.1)
- Reliable light source
- Appropriately sized speculum (Fig. 82.2)
- Endocervical brush or spatula
- Culture swab for gonorrhea and chlamydia
- Large cotton swabs for vaginal discharge or bleeding (Fig. 82.2)
- pH paper
- Saline and potassium hydroxide dropper bottles for wet preparations
- Lubricating gel
- Disposable gloves (Fig. 82.2)
- Small stool or chair for examiner

82.4 Procedure

1. Obtain permission from patient before beginning examination.
2. Chaperone should be present (medical staff member).
3. Make sure the examination table is clean and appropriately draped.
4. Have the patient in a loose-fitting gown.
5. Place the patient on the examination table in the lithotomy position with both feet in the stirrups and have the patient's pelvis as close to the edge of the table as possible.
6. Turn on the light source and adjust for optimum illumination. Put on the disposable gloves.
7. Communicate the procedure well to the patient.
8. Begin the examination with inspection and palpation of the abdomen.
9. Examine the external genitalia. Evaluate the skin, labia minora and majora, clitoris, urethral meatus, vaginal

N.W. Rashid, MD
Department of Emergency Medicine, WellStar Kennestone Hospital, Marietta, GA, USA
e-mail: naumanrashid@hotmail.com

E.B. Josephson, MD
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, Weill Cornell Medical College of Cornell University, Bronx, NY, New York
e-mail: elaine.Josephson@nychhc.org

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_82

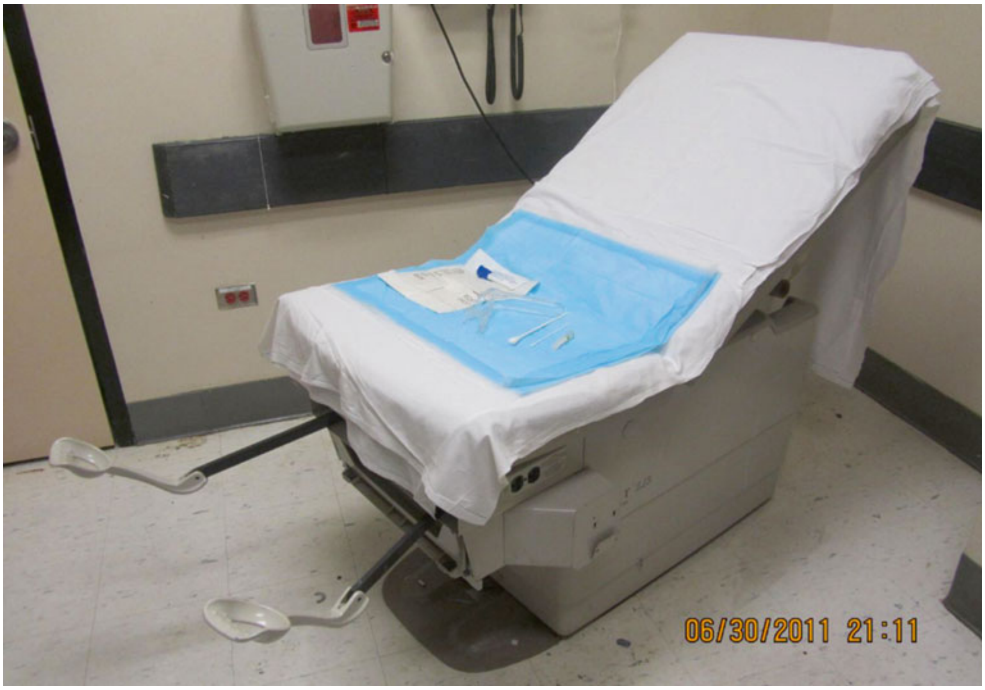

Fig. 82.1 Examination table with stirrups

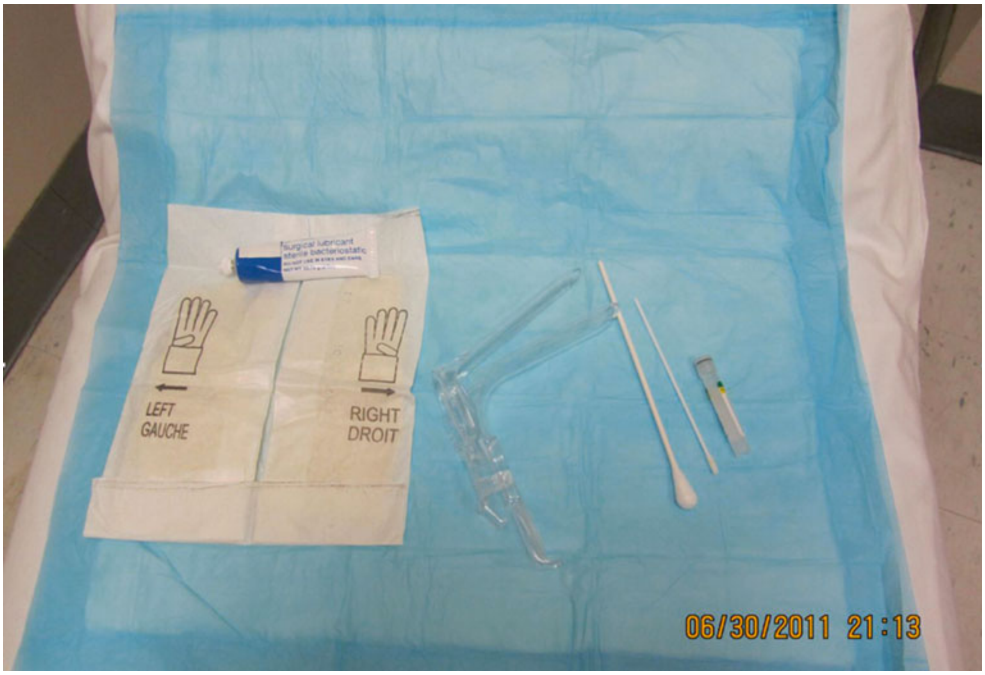

Fig. 82.2 Gloves, speculum, and swabs

canal, and Bartholin glands (Fig. 82.3). Look for skin abnormalities, lesions, masses, rashes, excoriation, abscesses, discharge, bleeding, or trauma. Palpate for tenderness.

10. Lubricate the appropriate-size speculum (mostly medium size). Insert the speculum through the vaginal opening with gentle downward pressure. The speculum should advance without any resistance until the cervical os is visualized.
11. Inspect the vaginal walls for any lesions or masses (Fig. 82.4). The cervical os is inspected to see if it is open or closed. Cervical cultures for gonorrhea and chlamydia are obtained with a cotton swab and sent for microbiology.
12. A sample of the discharge or bleeding is taken with a large cotton swab. The color, odor, and amount should be noted. The pH of the vaginal discharge can be evaluated. Normal pH is less than 4.5. An elevated pH indicates an infection (Table 82.1).

Fig. 82.3 Female external genitalia

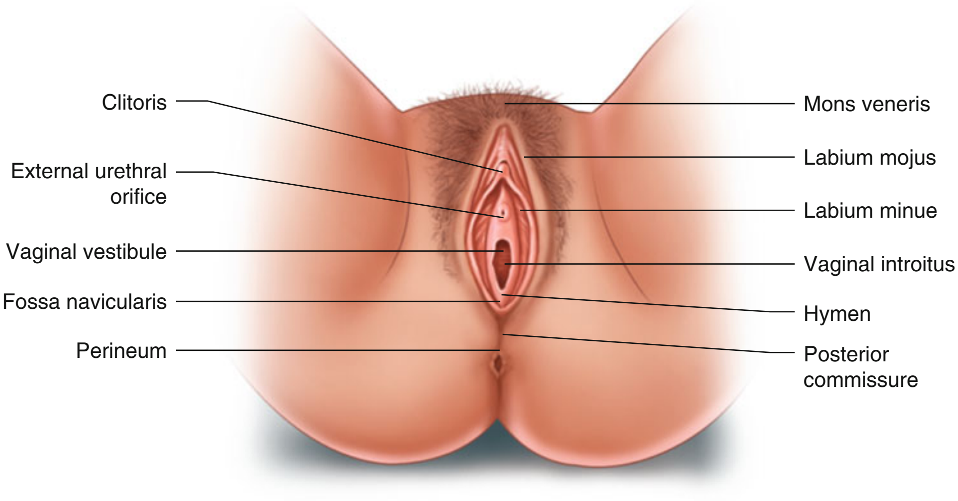
Clitoris
External urethral orifice
Vaginal vestibule
Fossa navicularis
Perineum
Mons veneris
Labium mojus
Labium minue
Vaginal introitus
Hymen
Posterior commissure

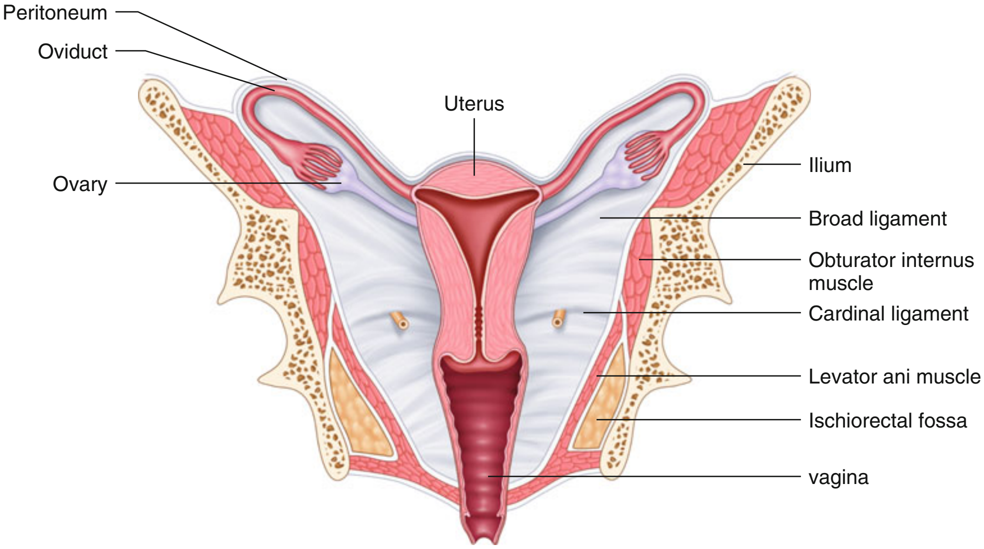
Peritoneum
Oviduct
Ovary
Uterus
Ilium
Broad ligament
Obturator internus muscle
Cardinal ligament
Levator ani muscle
Ischiorectal fossa
vagina

Fig. 82.4 Female internal genitalia

Table 82.1 Wet preparation interpretation

Organism	Preparation	pH	Microscope	Cervix	Appearance of discharge
Bacterial vaginosis	Saline	>4.5	Clue cells	Redness	Thin, milky, fishy odor
Trichomoniasis	Saline	>4.5	Motile flagella	Strawberry red	Yellow-green, foamy
Yeast	Potassium hydroxide	3.8–4.5	Budding yeast pseudohyphae	Normal	White, cottage cheese

13. Next, a bimanual examination should be performed (Fig. 82.5). Lubricating gel is applied to the nondominant gloved hand and the index and middle finger are inserted into the vagina until the cervix is felt. The other hand is placed on the abdomen to palpate the uterus and ovaries. Pressure is applied to the abdomen while the vaginal hand is elevated upward.
14. The cervix is palpated to elicit any cervical motion tenderness. The uterus is palpated and the size, position, and mobility are noted. The adnexa are examined for masses and tenderness. If a mass is palpated, the size, mobility, consistency, and tenderness are noted.
15. The final part of the pelvic examination is the rectovaginal examination. Lubricate the index and middle fingers

of the left hand. Place the index finger in the vagina and the middle finger in the rectum. Palpate for any fistulas or masses. With the finger, also palpate the uterosacral ligaments, the broad ligaments, and the pelvic side walls. The finger is then gently removed and any feces are inspected for mucous or occult blood.

16. A wet preparation is made by obtaining a sample of the vaginal discharge and placing it in a vial mixed with saline solution. A drop of the solution is placed on a microscopic slide and examined under high magnification for the presence of clue cells (Fig. 82.6) diagnostic for bacterial vaginosis and trichomonads (Fig. 82.7) diagnostic for trichomoniasis. For yeast, two drops of the solution is mixed with two drops of potassium hydroxide. Presence of hyphae is diagnostic of candida (yeast) species (Fig. 82.8)

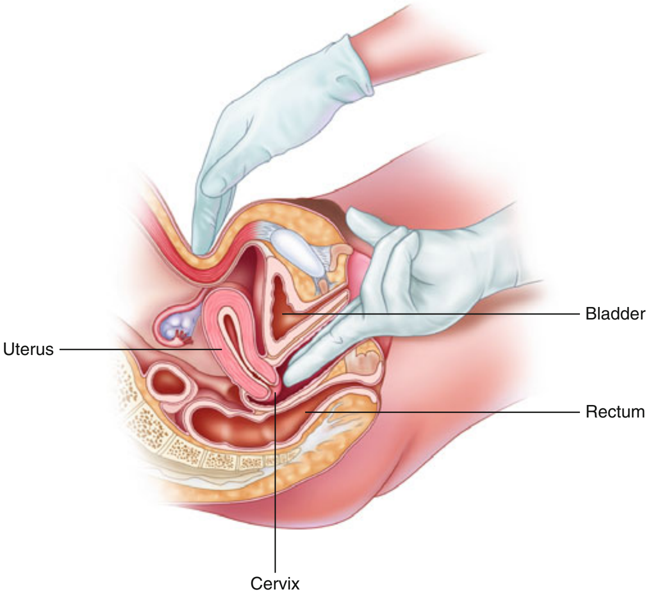

Fig. 82.5 Bimanual pelvic examination

82.5 Complications

- Urinary tract infection
- Vaginal bleeding
- Cramping

82.6 Pearls and Pitfalls

- Pearls
  - Good communication is essential to ensure the patient is comfortable and not anxious.
  - The chaperone should be a medical staff member.
  - Do not skip the pelvic examination if the patient is menstruating.
- Pitfalls
  - Do not forget to perform a complete abdominal exam along with the pelvic exam to rule out any GI etiology.
  - In older females (>50), perform the DRE for a stool occult sample as a possible source of bleeding.

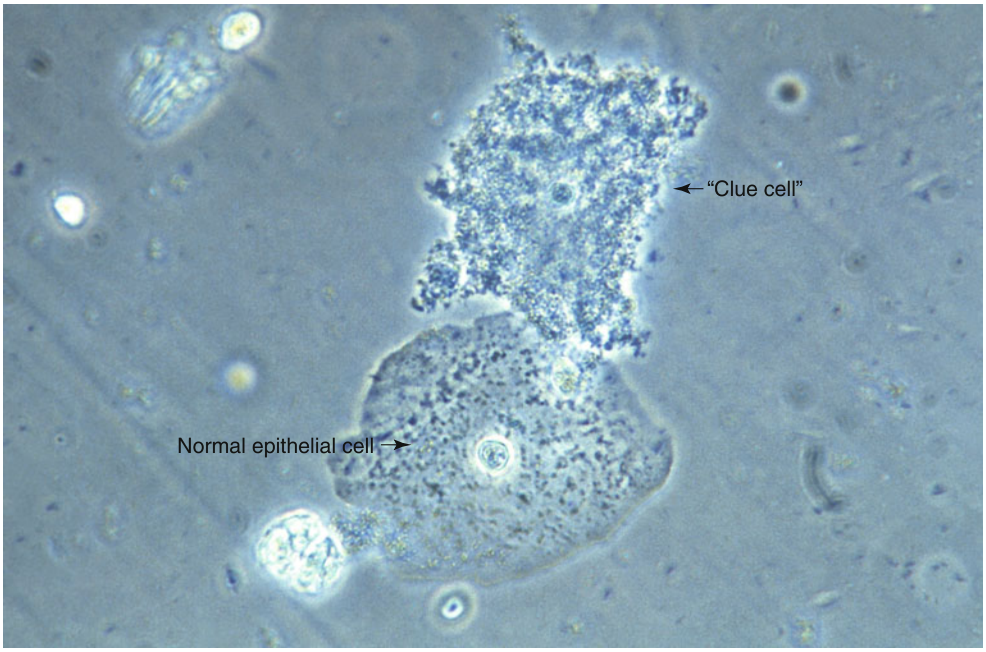

Fig. 82.6 Photomicrograph of a vaginal smear specimen depicting two epithelial cells, a normal cell, and an epithelial cell with its exterior covered by bacteria giving the cell a roughened, stippled appearance known as a "clue cell" (From the CDC Public Health Image library)

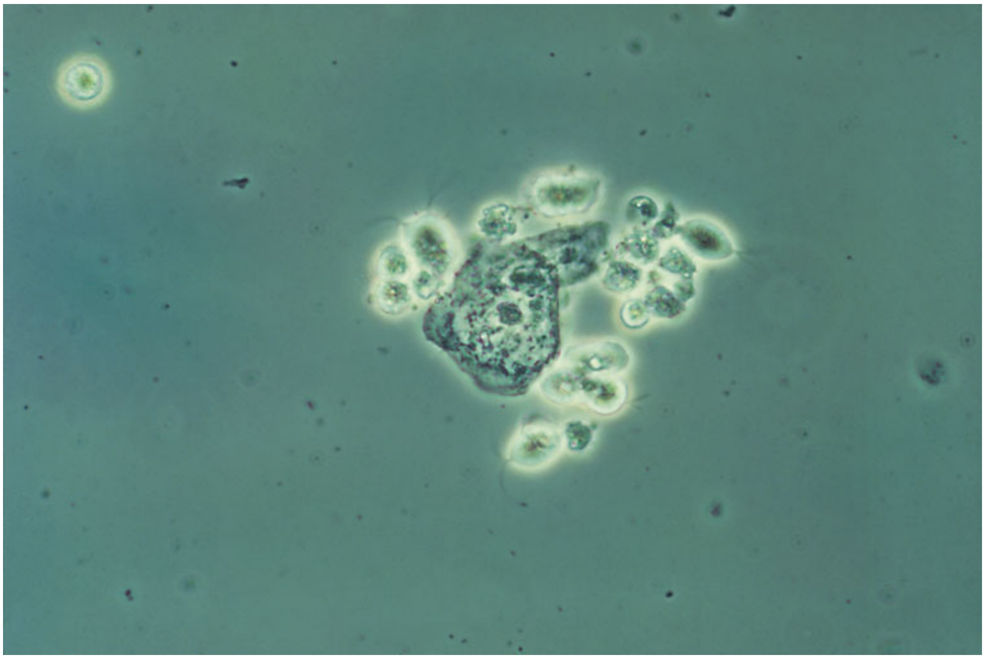

Fig. 82.7 Photomicrograph of trichomonads in wet mount prepared witvh physiological saline (From the CDC Public Health Image library)

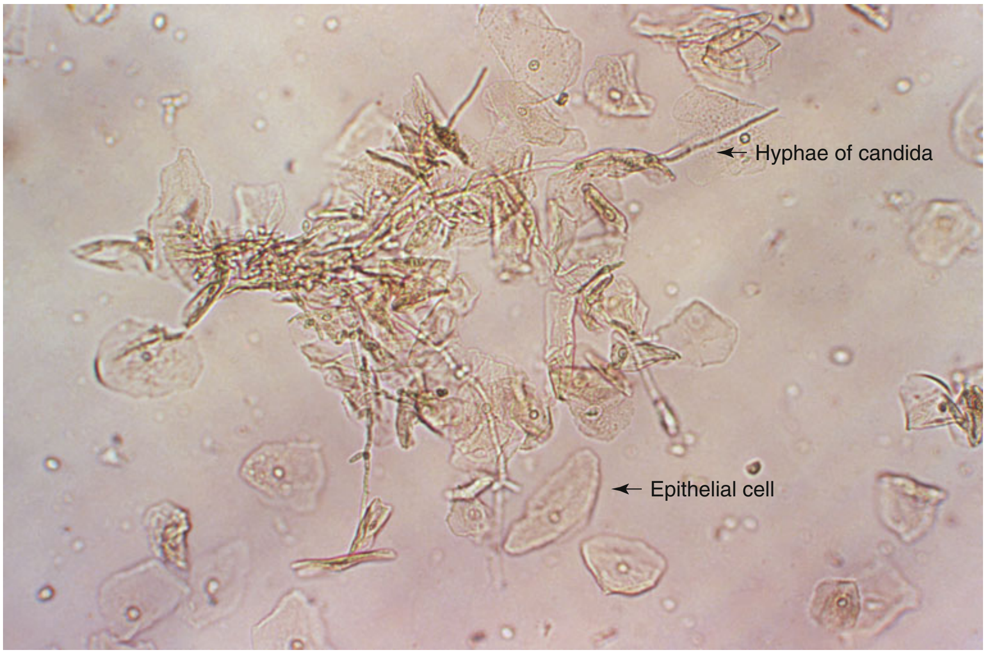

Fig. 82.8 *Candida albicans* from vaginal wet prep (From the CDC Public Health Image library)

Selected Reading

Brown J, Fleming R, Aristzabel J, Gishta R. Does pelvic exam in the emergency department add useful information? West J Emerg Med. 2011;12:208–12.

Butler J, Barton D, Shepherd J, Reynolds K, Kehoe S. Gynaecological examinations. Good not bad medicine. BMJ. 2011;342:d1760.

Carr SE, Carmody D. Outcomes of teaching medical students core skills for women's health: the pelvic examination educational program. Am J Obstet Gynecol. 2004;190:1382.

Katz VL, Lentz G, Lobo RA, Gershenson D, editors. Comprehensive gynecology. 5th ed. Philadelphia: Mosby; 2007.

Tiemstra J, Chico P, Pela E. Genitourinary infections after a routine pelvic exam. J Am Board Fam Med. 2011;24:296–303.

83 Bartholin Gland Abscess/Cysts Drainage

Holly H. Charleton, Marylin Otero, Diane F. Giorgi, and Joseph A. Tyndall

83.1 Indications

- Bartholin gland cysts larger than 1 cm or painful
- Bartholin gland abscess

83.2 Contraindications

- Absolute
  - None
- Relative
  - Recurrent/complex abscess requiring general anesthesia in operating room
  - Coagulopathy

83.3 Procedure Types

- Incision and drainage (I&D)
- Iodoform packing
- Word catheter
- Jacobi ring catheter
- Silver nitrate stick ablation (after I&D)

H.H. Charleton, MD
Emergency Department, The Brooklyn Hospital Center,
Brooklyn, NY, USA
e-mail: Hcmd08@gmail.com

M. Otero, MD
Department of Emergency Medicine, Franklin Hospital,
Valley Stream, NY, USA
e-mail: momd42@hotmail.com

D.F. Giorgi, MD
Department of Emergency Medicine, Mount Sinai Queens,
New York, NY, USA
e-mail: dgiorgi@earthlink.net

J.A. Tyndall, MD (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_83

83.4 Materials

- Sterile gloves
- Sterile skin preparatory solution/swabs and drapes
- Lidocaine 1 % (local anesthesia); may give oral or intravenous sedatives or analgesics
- Needles 25 or 27 gauge with 3-mL syringe for lidocaine injection
- Scalpel #11 for incision
- Culture swab
- Hemostat or suture kit
- Word catheter (if choosing that method) (Fig. 83.1)
- Jacobi ring catheter (if choosing that method) (Fig. 83.2)
- 3-mL syringe with saline for inflation of Word catheter balloon
- Silver nitrate stick (if choosing that method)
- Gauze pads for bleeding and effluents
- Iodoform packing

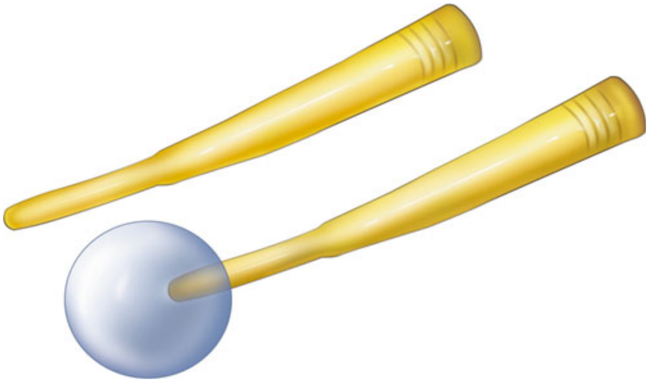

Fig. 83.1 Word catheter (inflated/deflated)

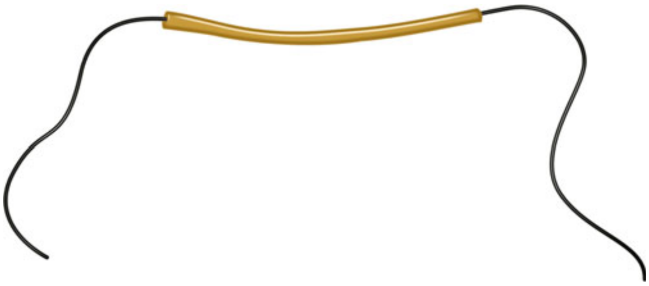

Fig. 83.2 Jacobi ring catheter

83.5 Procedures

83.5.1 Incision and Drainage

1. Obtain informed consent from the patient.
2. Place the patient in the lithotomy position.
3. Prepare the cyst/abscess and the surrounding area with sterilizing fluid/swabs and drape the area, leaving the cyst/abscess accessible for the procedure (Fig. 83.3)
4. May need to hold traction to the labia to fully expose the cyst/abscess (Fig. 83.4).
5. Inject 1–4 mL of lidocaine at the planned site of incision (Fig. 83.5).
6. Hold one side of the cyst/abscess with a forceps or hemostat to maintain traction while incising.
7. With a #11 scalpel, make an incision approximately 0.5–1 cm and 1.5 cm deep in the introitus or behind the hymnal ring to prevent vulvar scarring. The incision should be made through the fluctuant area of the abscess on the mucosal surface. The incision should be linear and large enough to fit the Word catheter (if using that method).
8. Drain the cyst/abscess completely, using the hemostat to break up the loculations.
9. Culture the abscess with a swab and send to microbiology.
10. All previous steps should be done regardless of the procedure type to follow.

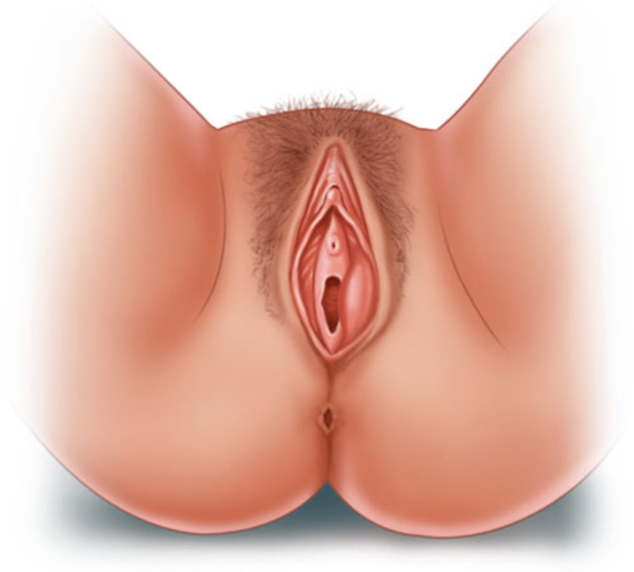

Fig. 83.3 Cyst/abscess accessible for the procedure

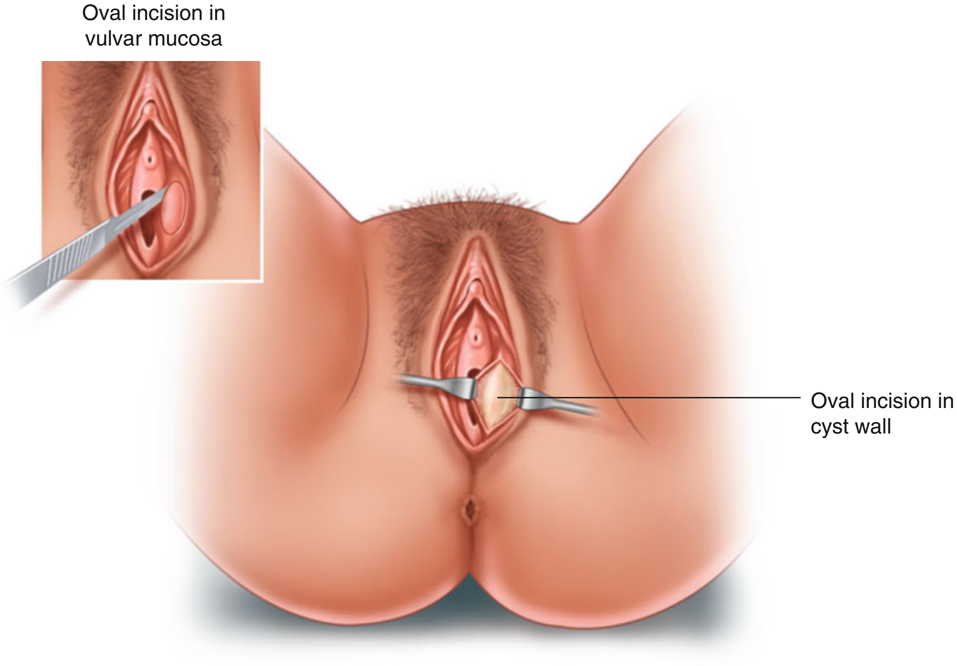
Oval incision in vulvar mucosa
Oval incision in cyst wall

Fig. 83.4 May need to hold traction to the labia to fully expose the cyst/abscess

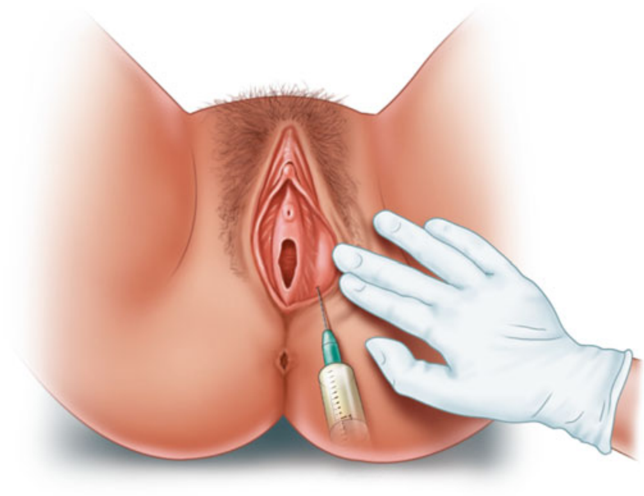

Fig. 83.5 Inject 1–4 mL of lidocaine at the planned site of incision

83.6 Iodoform Packing

1. Pack with iodoform packing, grasping the end of the packing material with a hemostat and inserting deeply within the cavity.
2. The cavity should be filled with the packing material and a small piece left exposed (for ease of retrieval during follow-up).
3. Clean the site and cover with gauze.

83.6.1 Word Catheter

- After culturing the abscess
  1. Place the Word catheter into the incision site as deep as possible (if the incision is too large, the catheter will slip out) (Fig. 83.6).
  2. Inflate the balloon of the Word catheter with 2–3 mL of saline or water injected into the hub with a needle and syringe (Fig. 83.7).
  3. Tuck the end of the Word catheter into the vagina for comfort purposes.
  4. The catheter should remain in place for 4 weeks to allow for epithelialization of the tract.

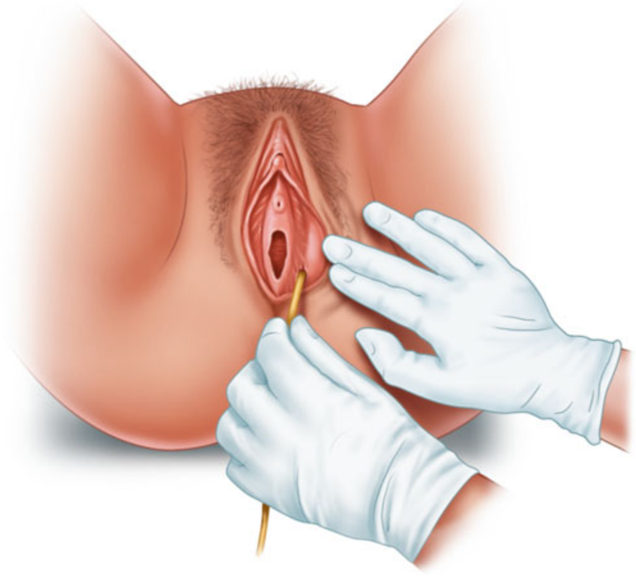

Fig. 83.6 Place the Word catheter into the incision site as deep as possible

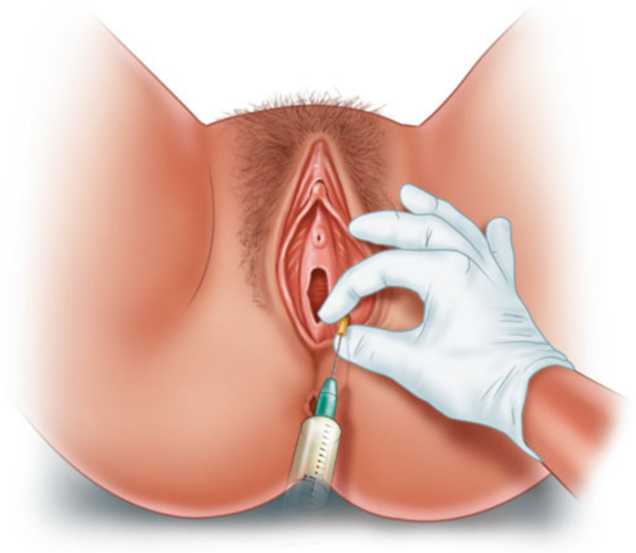

Fig. 83.7 Inflate the balloon of the Word catheter with 2–3 mL of saline or water injected into the hub with a needle and syringe

83.6.2 Jacobi Ring Catheter

1. Grasp one end of the Jacobi ring with a hemostat and pass it through the initial incision site.
2. At this time, use a hemostat to break loculations and culture the material for microbiology.
3. Pull the Jacobi ring through the abscess cavity (be careful not to pull the suture out of the catheter) and make a second incision to pull the catheter out through it.
4. The two ends of the catheter are then tied, forming a ring.

83.6.3 Nitrate Stick

- After sending the culture
    1. Take the silver nitrate stick and place it deep within the cyst/abscess cavity (Fig. 83.8).
    2. The patient is instructed to return in 48 h for removal of the remaining nitrate material and necrotic tissue and wound cleaning.
    3. The patient should be warned of side effects including pain, chemical burns of nearby tissue, edema, discharge, and scarring.

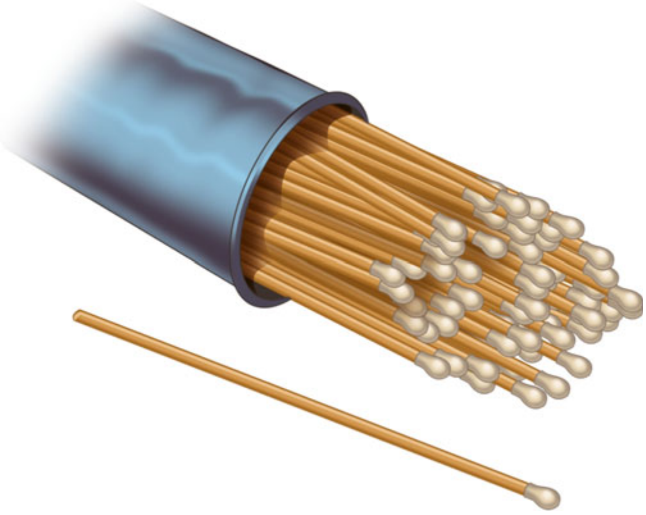

Fig. 83.8 Silver nitrate sticks

83.7 Aftercare

- Follow-up care is required after each of these procedures.
- A high-risk patient (e.g., diabetic) may need to be covered with broad-spectrum antibiotics.
- Pregnant women are also considered high risk and should be given antibiotics and followed closely.
- The patient should be instructed to remain on pelvic rest (nothing in vagina) while the catheter is in place, wear a pad owing to discharge, use sitz baths and analgesics for pain control, and follow up 1–2 days after the procedure.

Although a few other methods are available for draining Bartholin cyst/abscess including marsupialization, carbon dioxide laser vaporization, and excision, these methods are typically done by gynecologists and rarely, if ever, performed by an emergency department physician.

Selected Reading

Chen KT, Robert LB, Falk SJ. Disorders of Bartholin's gland. Available at: WWW.Uptodate.com.

Chen KT, Robert LB, Falk SJ. Word catheter placement for treatment of Bartholin's cysts and abscesses. Available at: WWW.Uptodate.com.

Gennis P, Li SF, Provataris J, et al. Jacobi ring catheter treatment of Bartholin's abscesses. Am J Emerg Med. 2005;23:414–5.

Vitaly KA, Mosquera C. Novel technique for management of Bartholin gland cysts and abscesses. J Emerg Med. 2008;36:388–90.

Wechter ME, Wu JM, Marzano D, Haefner H. Management of Bartholin duct cyst and abscesses: a systematic review. Obstet Gynecol Surv. 2009;64:395–404.

84 Sexual Assault Forensic Examination

Rajnish Jaiswal, Mary T. Ryan, and Muhammad Waseem

When the survivor of a sexual assault seeks medical care, in addition to addressing their medical needs, their forensic needs must also be addressed. This is best achieved by a specialist examiner, who is trained to conduct a Sexual Assault Forensic Examination (SAFE). When the examiner is a nurse, she or he is referred to as a Sexual Assault Nurse Examiner (SANE). In designated centers, the forensic examiner and the nurse, physician, law enforcement officials, social workers, and patient advocates work together as a Sexual Assault Response Team (SART).

The process of caring for survivors of sexual assault continues to evolve and reflects the advances in forensic science, judicial reform and our understanding of assault survivor psychology.

However, when an emergency medical condition exists, it should be addressed by the designated medical team. The role of the SAFE examiner becomes secondary in these situations. Life- or limb-threatening injuries always take priority over forensic evidence collection, although emergency medical care can often be rendered without compromising existing evidence.

84.1 Indications

- Survivors of sexual assault who seek and consent to forensic examination.
- The upper limit of time for evidence collection varies from state to state (e.g., 96 h in New York State [1]).

84.2 Contraindications

- Absolute
  - If the survivor does not consent to evidence collection
- Relative
  - If the upper time limit has been exceeded

84.3 Materials and Medications

- Ideally, a designated SAFE room should be available.
- Standardized sexual assault evidence collection kits.
- Gloves.
- Camera.
- Portable light source.
- Swab dryer.
- Wood's lamp.
- Anoscope.
- Colposcope, ideally with a camera.
- Support material for survivors: information pamphlets, clothing.
- Prophylactic medications: antibiotics, antiretrovirals, contraceptives.

R. Jaiswal, MD
Department of Emergency Medicine, New York Medical College,
Metropolitan Hospital Center, New York, NY, USA
e-mail: Rajnish.Jaiswal@gmail.com

M.T. Ryan, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: maryryanmd@aol.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_84

84.4 Procedure: "Prepare the Patient, Prepare the Room"

1. Informed Consent
    - A separate consent is required for the SAFE. Obtaining consent has important psychosocial implications for the survivor and returns "control" and "choice" to him or her at this critical time [2]. If the survivor chooses not to undergo a SAFE, the examiner must respect his or her decision. Consent is not an "all-or-none" phenomenon and survivors can chose to consent to some steps and decline others. The examiner should be respectful of their decision.
    - Consent for the SAFE should include consent for evidence collection, forensic photography, release of evidence to law enforcement, and permission to discuss the findings of the SAFE with investigators.
2. Law Enforcement Involvement
    - State laws vary in terms of reporting requirements for sexual assault. The examiner should be familiar with the requirements in the state in which she or he practices. All survivors should be offered law enforcement involvement and the benefits of doing so should be outlined to them.
3. Evidence Collection
    - Sexual assault evidence collection kits are specialized preassembled kits containing essential materials for collecting and preserving evidence (Fig. 84.1). The kit contains written instructions, swabs, envelopes, body diagrams, and an integrity seal for the examiner's use.
4. Forensic Interview and History Taking
    - The forensic interview is the first step in the SAFE process. It is a therapeutic as well as a forensic exercise, designed to establish rapport with the survivor, offer support, and gather information to help guide the medical care and direct evidence collection. Acquiring information is a continuous process that ends only when the survivor-SAFE interaction ends.
    - The survivor's exact words with quotation marks should be recorded. A simple factual account of events should be documented. Avoid biased or prejudicial language, such as "allegedly" or "claims." Relevant information includes the time of the assault, the type of contact involved (offender-survivor and survivor-offender), the number of people involved, and the survivor's activities since the assault. A basic medical and obstetrical-gynecological history is also relevant. The SAFE interview is not an investigative interview. Investigation of the sexual assault is the role of law enforcement.
5. General Physical Examination
    - The patient should be asked to undress over a paper sheet to allow any trace evidence to fall and be collected. She or he should be given a gown to wear. A systematic head-to-toe examination should be undertaken. Identify any injuries, no matter how minor. Document them in writing, on a body diagram (Fig. 84.2), and, when possible, with photography. Pay attention to areas that can be easily overlooked: in the mouth, behind the ears, under the chin, and the soles of the feet, for example. Take time to palpate the scalp for areas of tenderness.
6. Injury Documentation
    - Always take time during documentation. Describe the type of injury—abrasion, contusion, laceration, or bite mark. Document the size and site of the injury, ideally include a measuring device in the photograph. A commonly used scale is the one provided by the American Board of Forensic Odontology (ABFO) (Fig. 84.3). If an injury appears to have a shape or pattern (e.g., linear, circular, curvilinear, petechial), describe it without drawing specific conclusions.
7. Bite Marks
    - Bite marks require additional evaluation because they may have salivary trace evidence associated with them. In addition to being described and photographed, they should be swabbed and the dried swabs included in the evidence collection kit.
8. Forensic Pelvic Examination
    - The purpose of the genital examination (external and speculum) is to identify injury and collect forensic evidence.
9. Inspection
    - Visually examine the external genitalia. Separate the labia and look in skin folds and at the posterior fourchette for injury. The TEARS mnemonic (T = tear, E = ecchymosis, A = abrasions, R = redness, S = swelling) is a useful tool while inspecting and documenting. External genital injury findings can be photographed using a standard camera (digital or conventional 35 mm) or a colposcope camera for additional magnification.
10. Speculum Examination
    - Insert a moistened speculum under a good light source and inspect the vault and cervix for any injuries or possible trace evidence for collection (pooled secretions, hair, retained condom, debris). A colposcope (Fig. 84.4) is a useful adjunct and allows for magnification and assists in injury identification and photodocumentation.
    - Bimanual pelvic examination may be a part of some protocols but is not mandatory.
11. Rectal Examination
    - Inspect the area looking for fissures, bleeding, or secretions. Anoscopy, if indicated by this history and permitted by the survivor, should be performed and the findings documented and photographed.

12. Evidence Collection
    - The evidence collection kit should be opened and the contents laid out in a systematic way. Once the evidence collection kit has been opened, it cannot be left unattended at any time. Each envelope should be labeled with the survivor's name and the time and date of collection. The required swabs and slides are included in the kit.
13. Collection of Biological Material
    - Evidence collection will include oral, anal, and vaginal swabs. Swabs should be allowed to air dry before being placed back in the envelopes. Trace evidence should be collected and may include nail scrapings, dried secretions, loose hair collection, and possible foreign bodies (e.g., soil, condom). A Wood's lamp may help the examiner to identify dried secretions on skin or clothing. When each step is completed, the envelope will be closed, sealed, and signed by the examiner and returned to the box.
    - When completed, the Sexual Assault Evidence Collection Kit (SAECK) is closed, the provided evidence seal placed on the box, and the seal signed and dated by the examiner. The evidence is then given to law enforcement (if the patient consents) or maintained in a predesignated, secure locked area if law enforcement is not yet involved in the case. Each time evidence is passed from person to person, the transfer must be documented in writing to ensure it is not compromised or tampered with in any way. This is the underlying principle of maintaining a "Chain of Custody." This chain must be maintained for evidence to be admissible in court.
14. Collection of Clothing
    - Clothing may be considered "evidence" and collected in some cases. Depending on the case, this may include underwear and any feminine hygiene products. These may fit in the evidence collection kit itself. Larger items of clothing and/or shoes will need to be collected separately. They should be placed in an appropriately sized paper bag and labeled with the patient's name. The bag should be sealed, signed, and dated by the examiner in the same way as all other evidence. Any additional evidence should remain with the SAECK. The survivor should be provided with replacement clothes and underwear.
15. Forensic Photography
    - Although the examiner is not expected to be a specialized forensic photographer, photodocumentation of injuries is an important part of the SAFE. A separate consent is required. Either a conventional 35-mm camera or a high-resolution digital camera is acceptable.
    - At least one image should include the survivor's face or some form of identifying marks. Near and far images should be taken. The camera should be held at 90° to the surface to avoid distortion of the image. A tape measure should be included when an injury is being photographed. An identifier, like medical record number or case number, should be visible in the image if possible. The examiner should document in the records that photographs were taken.
16. Investigations
    - Baseline complete blood count (CBC), chemistry panel, and liver function tests are generally drawn before initiation of human immunodeficiency virus (HIV) prophylaxis. Serologic tests for syphilis, hepatitis B virus (HBV), hepatitis C virus, and HIV should be obtained. Urine should be sent for analysis and pregnancy testing. Urine for toxicology may be useful in selected cases. Testing for gonorrhea and chlamydia before starting prophylactic antibiotics may be undertaken, but this remains controversial.
17. Prophylaxis
    - Survivors should be offered prophylaxis against pregnancy, common sexually transmitted infections, HBV, and HIV. The current Centers for Disease Control and Prevention (CDC) guidelines recommend the following:

    (a) HBV vaccination should be offered to sexual assault victims at the time of the initial examination if they have not been previously vaccinated. Postexposure HBV vaccination, without hepatitis B immunoglobulin (HBIG), should adequately protect against HBV infection. Follow-up doses of vaccine should be administered 1–2 and 4–6 months after the first dose.

    (b) An empirical antimicrobial regimen for chlamydia, gonorrhea, and trichomonas should be offered.
    - Recommended regimens:
        - *Ceftriaxone* 125 mg intramuscularly in a single dose
        - *PLUS*
        - *Metronidazole* 2 g orally in a single dose
        - *PLUS*
        - *Azithromycin* 1 g orally in a single dose
        - *OR*
        - *Doxycycline* 100 mg orally twice a day for 7 days

    (c) Emergency contraception protocols are state and institution specific. A negative pregnancy test should be documented before evidence collection. Commonly prescribed regimens are "Plan B," "Ovral," and the recently approved "Ella."

    (d) Update tetanus profile if indicated.

    (e) HIV postexposure prophylaxis.
    - All patients with significant exposure should receive pretest counseling and postexposure prophylaxis as per CDC guidelines [3]. The regimens are complex (Table 84.1).

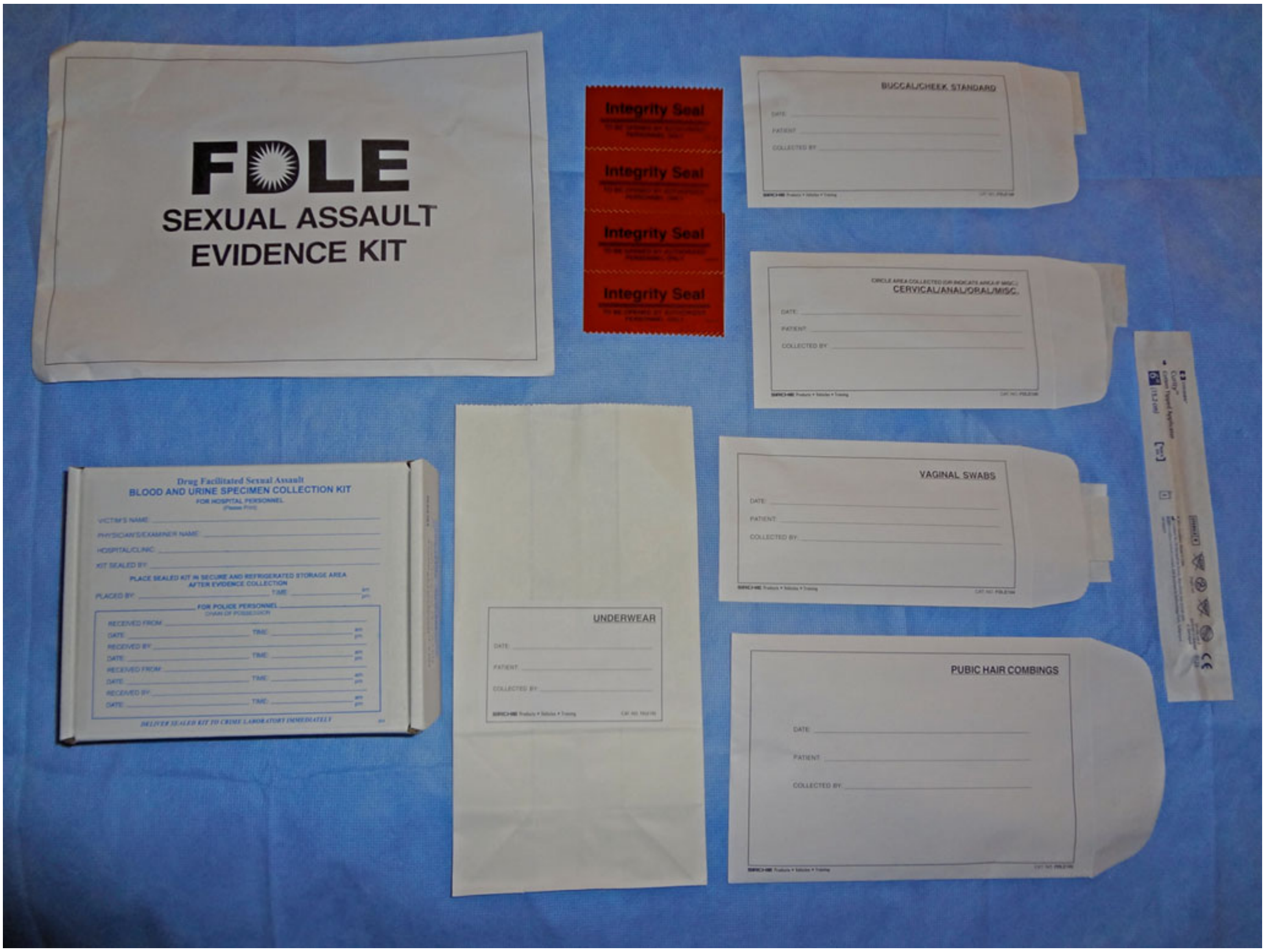
FDLE
SEXUAL ASSAULT
EVIDENCE KIT
Integrity Seal
Integrity Seal
Integrity Seal
Integrity Seal
BUCCAL/CHEEK STANDARD
CERVICAL/ANAL/ORAL/MISC.
VAGINAL SWABS
UNDERWEAR
PUBIC HAIR COMBINGS
DATE
PATIENT
COLLECTED BY
Drug Facilitated Sexual Assault
BLOOD AND URINE SPECIMEN COLLECTION KIT
FOR HOSPITAL PERSONNEL
VICTIM'S NAME
PHYSICIAN'S/EXAMINER NAME
HOSPITAL/CLINIC
KIT SEALED BY
PLACE SEALED KIT IN SECURE AND REFRIGERATED STORAGE AREA AFTER EVIDENCE COLLECTION
PLACED BY
TIME
FOR POLICE PERSONNEL
CHAIN OF POSSESSION
RECEIVED FROM
RECEIVED BY
DELIVER SEALED KIT TO CRIME LABORATORY IMMEDIATELY

Fig. 84.1 Sexual assault evidence collection kit

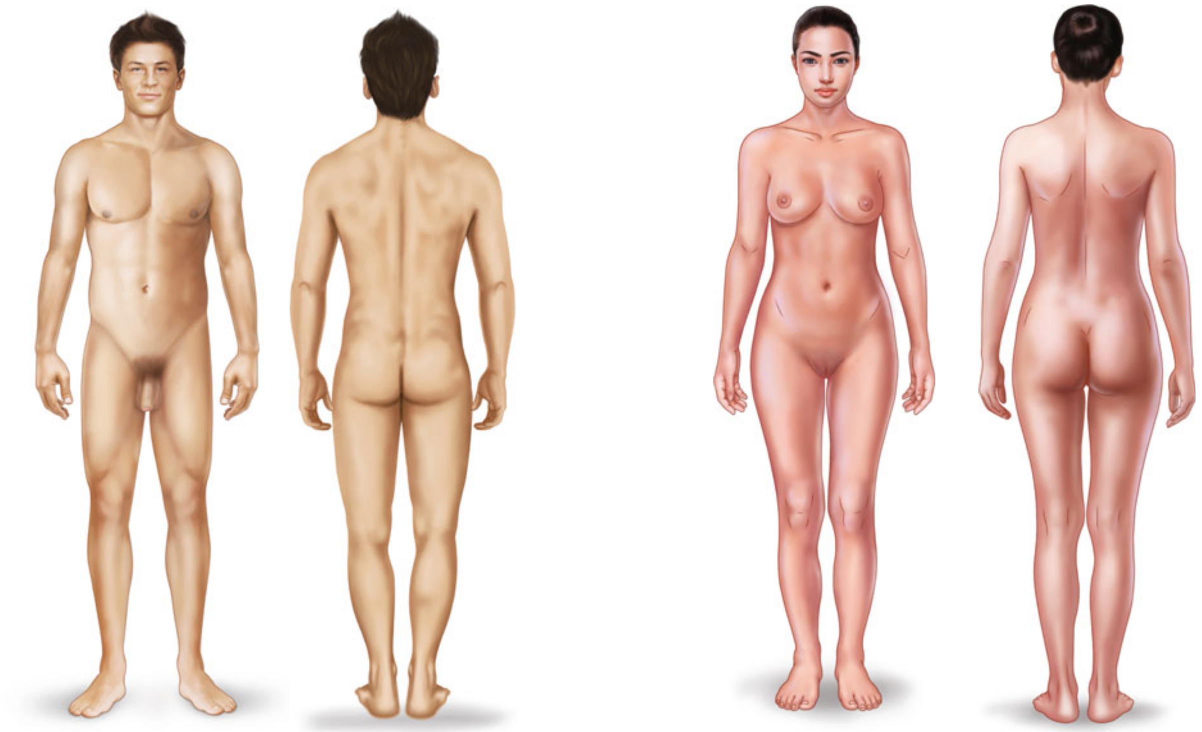

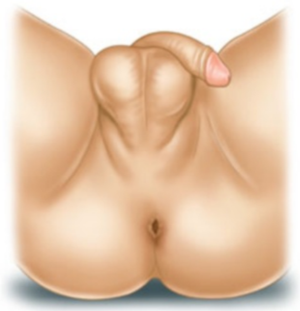

Male genitalia

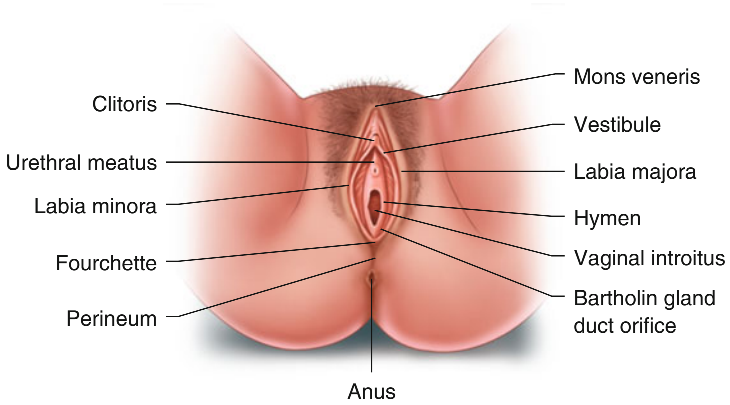
Clitoris
Urethral meatus
Labia minora
Fourchette
Perineum
Mons veneris
Vestibule
Labia majora
Hymen
Vaginal introitus
Bartholin gland duct orifice
Anus

Female genitalia

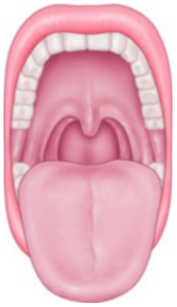

Oral

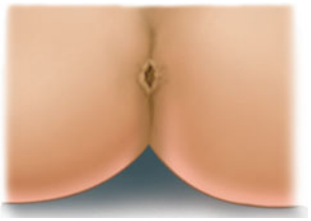

Anal

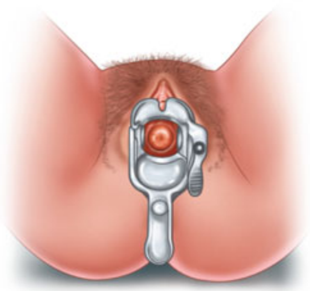

Cervical Observation

Fig. 84.2 Traumagram

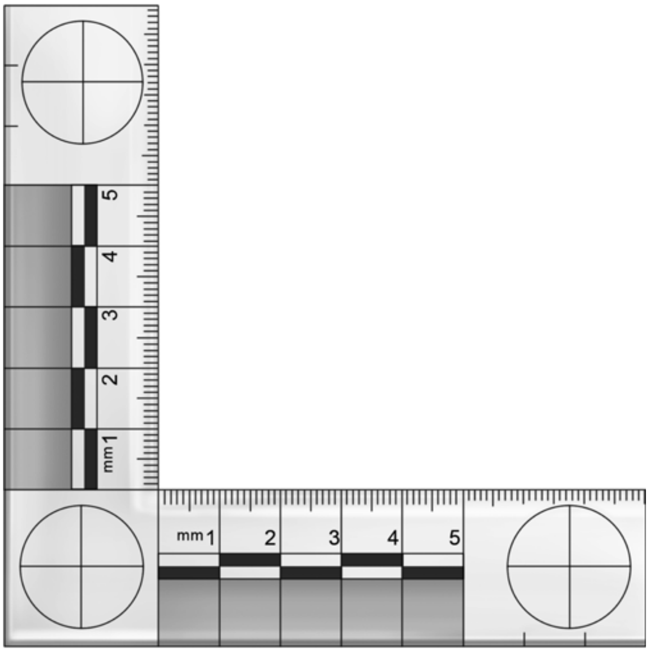

Fig. 84.3 American Board of Forensic Odontology (ABFO) scale (Courtesy Bronx SART Program)

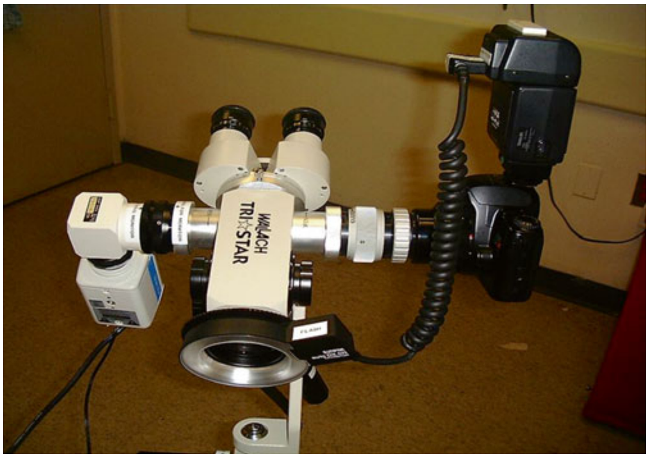

Fig. 84.4 Colposcope (Courtesy Bronx SART Program)

Table 84.1 Postexposure prophylaxis as per Centers for Disease Control and Prevention (CDC) guidelines [3]

Regimen	Dosage
Zidovudine (Retrovir, ZDV, AZT) + lamivudine (Epivir®, 3TC); *available as Combivir™*	ZDV: 300 mg twice daily or 200 mg three times daily, with food; total: 600 mg daily 3TC: 300 mg once daily or 150 mg twice daily Combivir: one tablet twice daily
Zidovudine (Retrovir, ZDV, AZT) + emtricitabine (Emtriva, FTC)	300 mg twice daily or 200 mg three times daily, with food; total: 600 mg/day, in 2 or 3 divided doses; FTC: 200 mg (one capsule) once daily
Tenofovir DF (Viread, TDF) + lamivudine (Epivir, 3TC)	300 mg once daily; 3TC: 300 mg once daily or 150 mg twice daily
Tenofovir DF (Viread, TDF) + emtricitabine (Emtriva, FTC)	TDF: 300 mg once daily; FTC: 200 mg once daily Truvada: one tablet daily

AZT azidothymidine, *FTC* emtricitabine, *3TC* lamivudine, *TDF* tenofovir disoproxil fumarate, *ZDV* zidovudine

84.5 Pearls and Pitfalls

- Pearls
  - Survivors will need to have both medical and psychosocial follow-up. Medical referrals should include gynecology and primary care for follow-up of their baseline serology, testing and completion of HBV vaccination regimen, and so on.
  - Referrals for counseling and information with 24-h hotlines should be provided**.** Recovery from a sexual assault is a process and is best achieved by a long-term support network [4].
- Pitfalls
  - It is estimated that survivors are men in fewer than 10 % of cases, although sexual assault in males appears to be greatly underreported. The same principles for evidence and prophylaxis apply for the SAFE.

References

1. Department of Health, State of New York. Acute care of the adult patient reporting sexual assault. 2004.
2. Criminal Victimization in the United States 2010. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics; 2010.
3. Varghese B, Maher JE, Peterman TA, et al. Reducing the risk of sexual HIV transmission. Sex Transm Dis. 2002;29:38–43.
4. Parekh V, Brown CB. Follow up of patients who have been recently sexually assaulted. Sex Transm Infect. 2003;79:349.

Treatment of Priapism

85

Jeffrey Kile, Katrina John, and Amish Aghera

85.1 Indications

- Ischemic (“low-flow”) priapism

85.2 Contraindications

- To cavernosal aspiration/irrigation
  - Nonischemic (“high-flow”) priapism
  - Overlying cellulitis
  - Uncontrolled bleeding disorder
  - Skin infection at the site of injection
- To intracavernosal injection of vasoactive agents (α-adrenergic sympathomimetics)
  - Severe hypertension
  - Dysrhythmias
  - Monoamine oxidase inhibitor use

85.3 Materials and Medications (Fig. 85.1)

- Sterile gloves
- Antimicrobial solution and swabs
- 4×4 gauze sponges
- Local anesthetic (1 % lidocaine 5 mL and 0.5 % bupivacaine 5 mL, without epinephrine)
- 10-mL syringe
- 20-mL syringe
- 19- or 21-gauge butterfly or straight needles (2)
- Blunt needle
- 27-gauge needle
- Normal saline, 1000 mL
- Phenylephrine 1 % solution (10 mg/mL), 1 mL

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_85

Fig. 85.1 Materials and medications

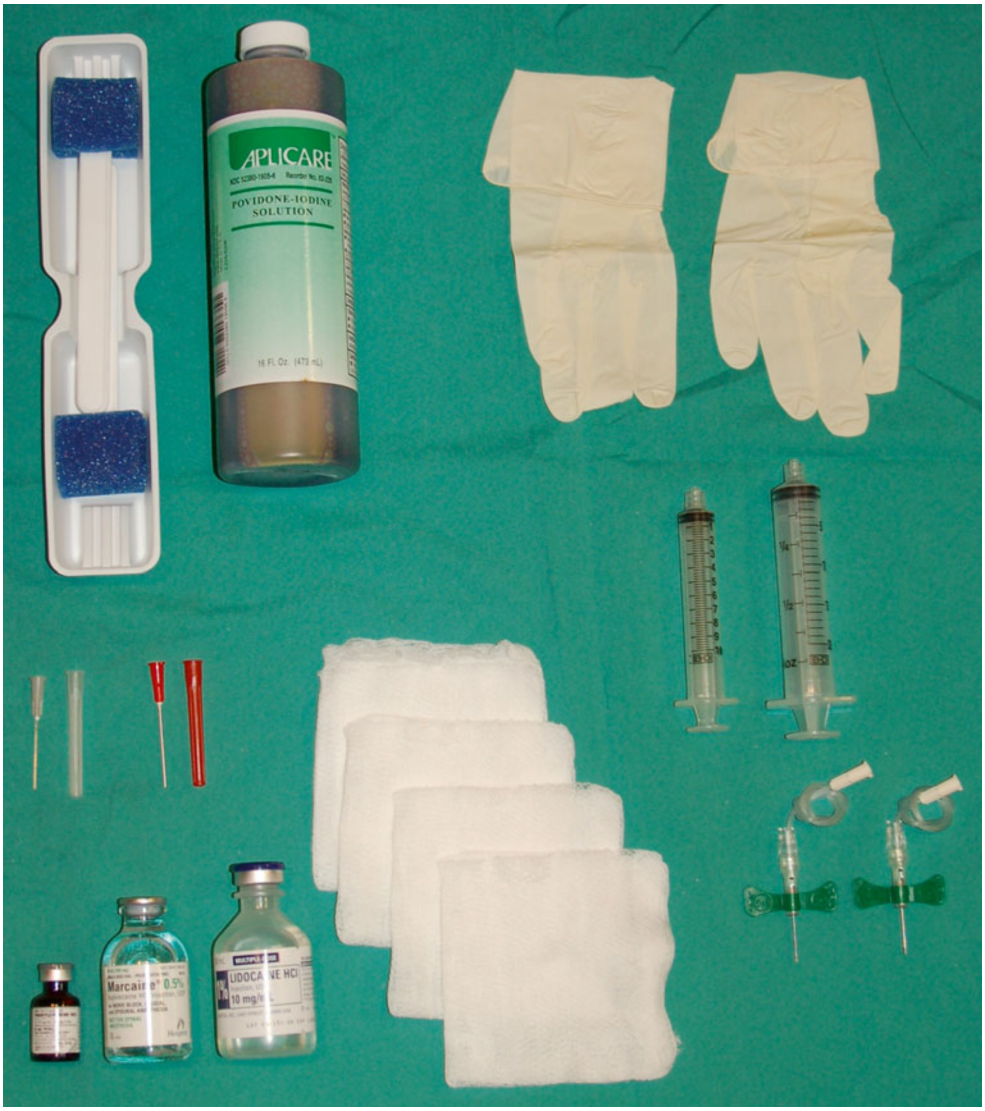

85.4 Noninvasive Therapy Preprocedure

1. Administer analgesia (e.g., parenteral opiates, benzodiazepines).
2. Administer subcutaneous terbutaline as soon as the diagnosis is suspected (0.25–0.5 mg subcutaneous in quadriceps, deltoid, or gluteus maximus) and repeat after 20 min if necessary.
   - If resolution of priapism does not occur with subcutaneous terbutaline, proceed to cavernosal aspiration.

85.5 Procedures

85.5.1 Dorsal Penile Nerve Block Procedure

1. Position the patient in the supine position.
2. Apply povidone-iodine solution liberally to the penis and scrotum using a 4×4 gauze pad.
3. Clean the glans and shaft of the penis in a circular motion.
4. Establish a sterile field by placing drapes between the scrotum and the shaft, above the shaft, and on either side (Fig. 85.2).
5. Draw up 5 mL 0.5 % bupivacaine and 5 mL 1 % lignocaine (both without epinephrine) into a single syringe.
6. Using a 27-gauge needle, inject local anesthetic superficially to raise skin wheals at the (dorsal) 2 and 10 o'clock positions as proximal to the base of the penis as possible.
7. Insert the needle through the wheal at the 2 o'clock position at the base of the penis until it contacts the pubic symphysis.
8. Withdraw the needle slightly and walk the needle in a caudal fashion down the pubis until the needle passes immediately below the symphysis and advance to a depth of 5 mm deeper than the depth of the pubic symphysis (Fig. 85.3).
   - A transmitted "pop" may be felt as the needle penetrates the superficial penile fascia beneath the symphysis.
9. Aspirate to confirm the tip of the needle is not within the lumen of a vessel.
10. Inject 4 mL of solution.
11. Repeat the injection of local anesthetic as outlined at the 10 o'clock position of the penile base to anesthetize the right dorsal penile nerve (Fig. 85.4).

Fig. 85.2 Priapism in sterile field

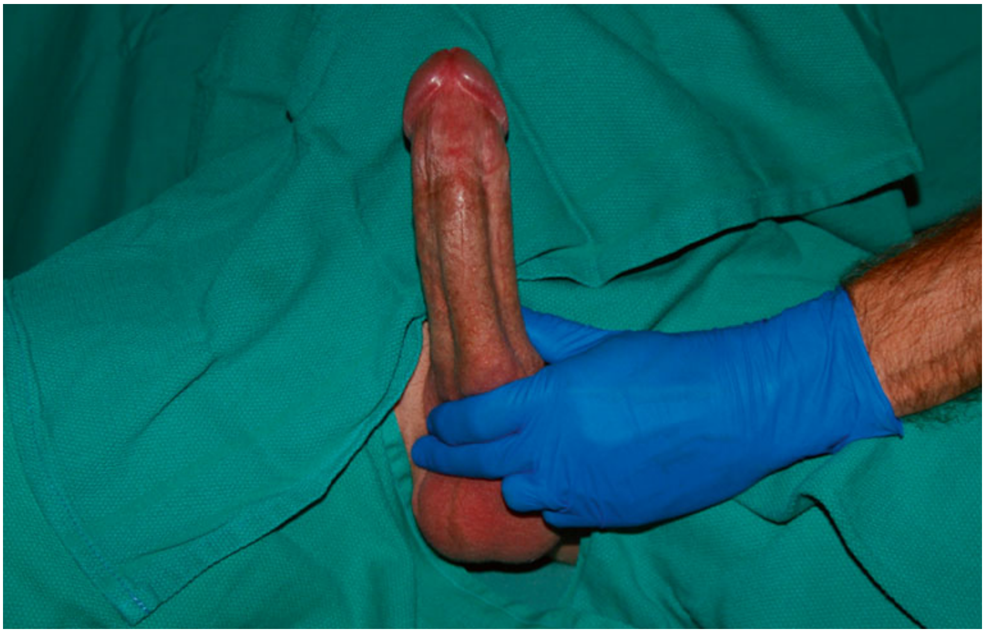

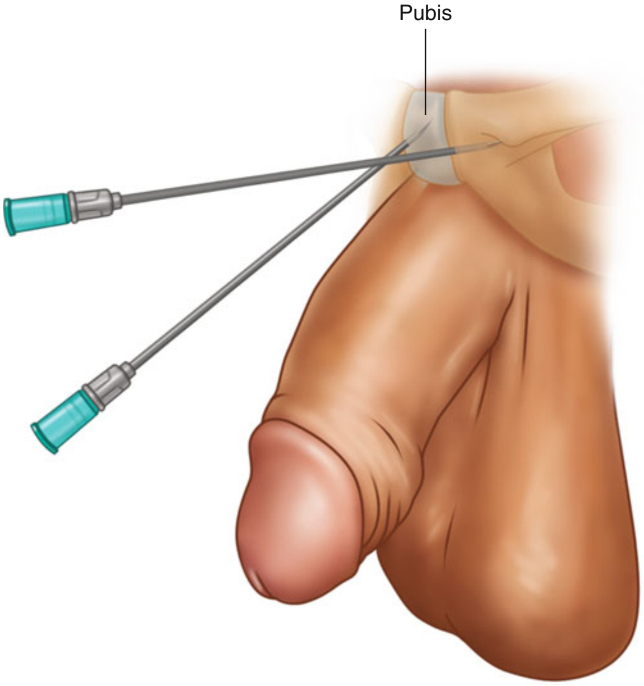

Fig. 85.3 Schematic anatomy of dorsal penile nerve block

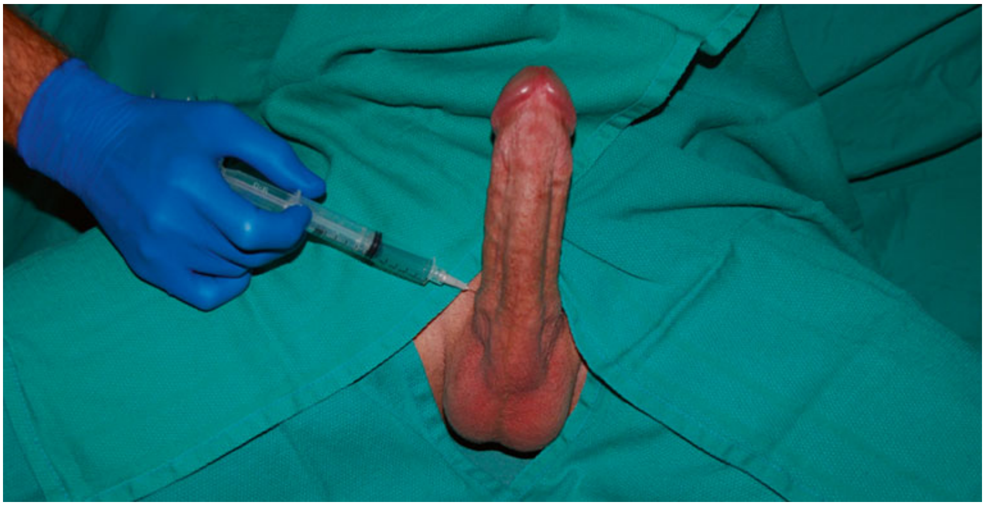

Fig. 85.4 Injection of local anesthetic

85.5.2 Cavernosal Aspiration Procedure

1. Attach a 19- or 21-gauge needle to a syringe.
2. Puncture the corpus cavernosa at the 2 o'clock or 10 o'clock position (~+60° or −60° from the midline) on the suprapubic aspect of the penis approximately 3 cm from the penile base, directing the needle straight toward the center of the ipsilateral cavernosum.
   - Never use the glans as a puncture site during this procedure.
3. Advance the needle slowly while drawing back on the plunger until blood is visible in the syringe (blood is usually easily aspirated).
4. Once blood is obtained, do not advance further, stabilize the needle, and use one hand to aspirate 20–30 mL of blood while milking the corpus with the free hand (Fig. 85.5).
   - The needle should not be advanced further once blood is visible in the syringe to minimize the risk of injury to the cavernosal artery.
   - Avoid excessive negative pressure on the plunger because this often halts aspiration.
   - If detumescence is not achieved using the above steps, proceed with the following steps.
5. Insert an irrigation needle by puncturing the corpus cavernosum on the same side of the penis punctured with the aspiration needle, approximately 1 cm from the penile base.
6. Irrigate the oxygen-depleted blood in the cavernosa by injecting 20–30 mL of 0.9 % normal saline via the proximal needle in exchange for the blood aspirated (Fig. 85.6).
7. Repeat the cycle of aspiration of 20- to 30-mL volumes of blood from the distal needle followed by irrigation with an equal volume of 0.9 % normal saline via the proximal needle until flow into the syringe of dark red (oxygen-depleted) blood ceases and bright red (oxygen-rich) blood is aspirated or until detumescence is achieved (Fig. 85.7).
   - When removing the needle after cavernosal aspiration, compress the puncture site for approximately 1 min to prevent hematoma formation.
8. Wrap the detumescent penis in gauze or an elastic bandage to prevent return of priapism and to compress the puncture site(s) (Fig. 85.8).

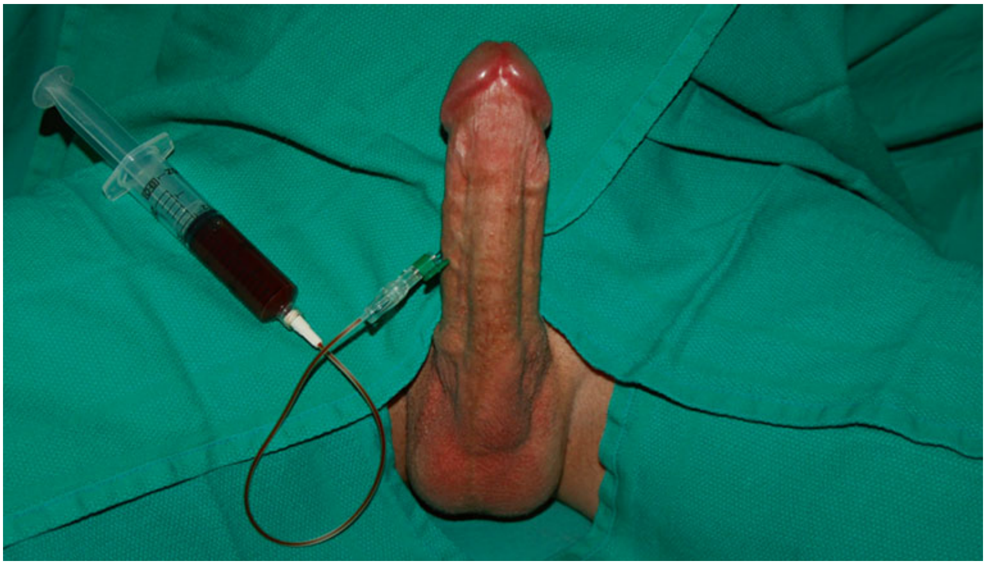

Fig. 85.5 Aspiration of cavernosal blood

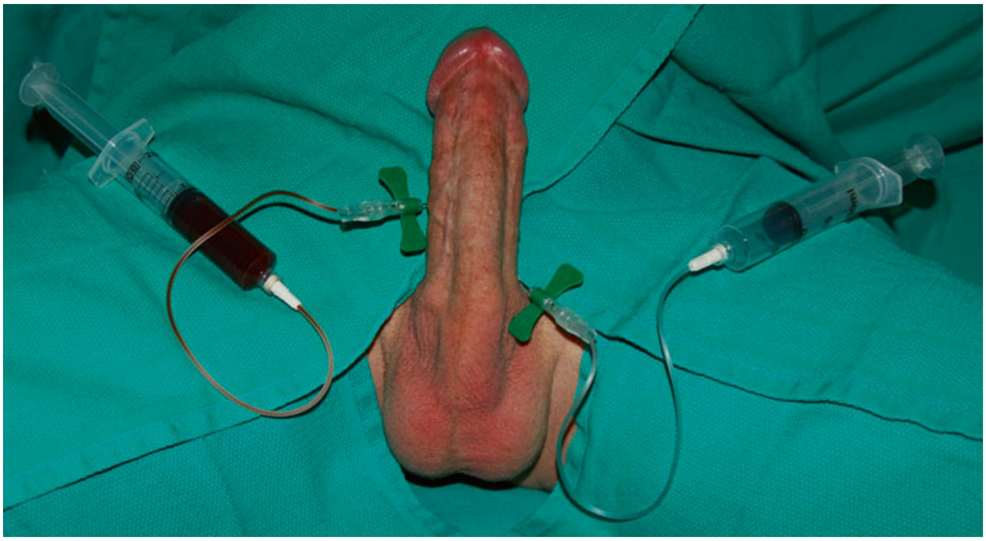

Fig. 85.6 Aspiration and irrigation of cavernosal blood

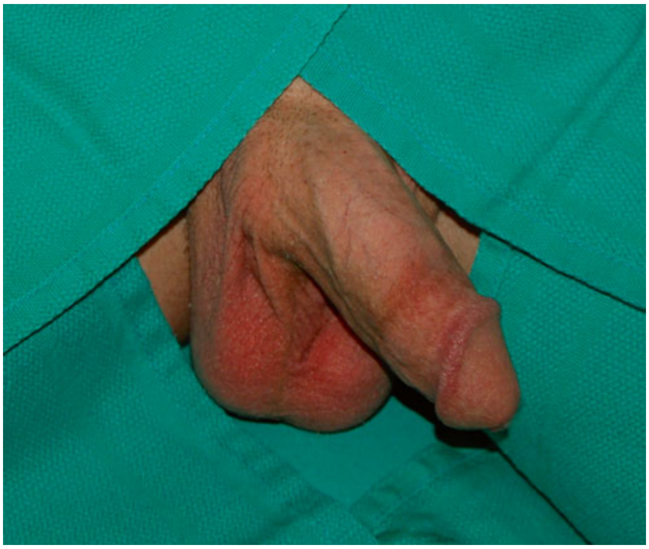

Fig. 85.7 Detumescence

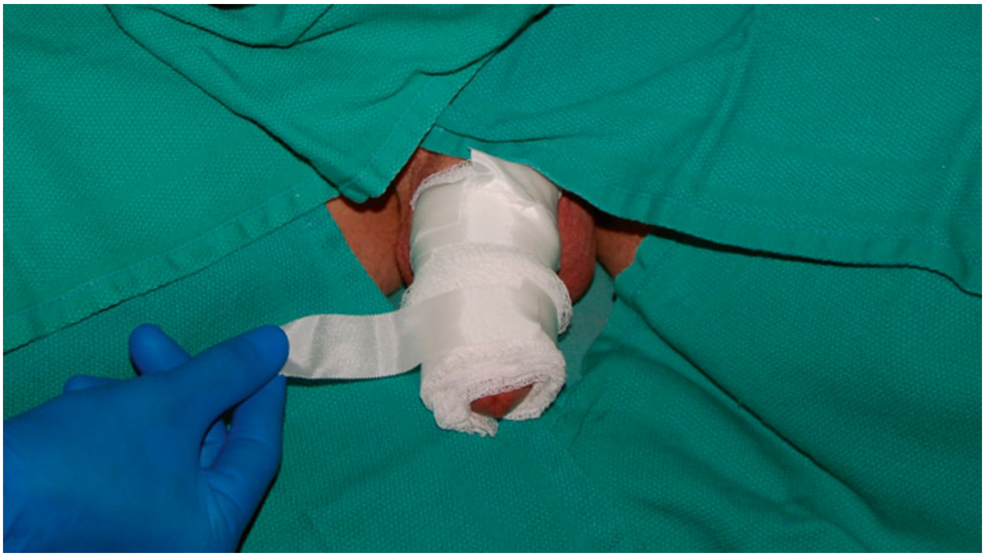

Fig. 85.8 Detumescent penis wrapped in compression dressing

85.5.3 Intracorporeal Injection Procedure

1. Prepare a diluted concentration of 100 μg/mL (1 mg/10 mL) phenylephrine solution by aspirating 0.1 mL of standard 1 % (10 mg/mL) phenylephrine solution into a 10-mL syringe and then adding normal saline to a total volume of 10 mL
2. Attach a 25- or 27-gauge needle to the syringe.
3. Puncture the corpus cavernosum at the 2 o'clock *or* 10 o'clock position (~+60° or −60° from the midline) on the suprapubic aspect of the penis approximately 1 cm from the penile base.
   - Puncture only one side of the penis.
4. Confirm the position of the needle by drawing back on the plunger to aspirate blood from the corpus cavernosa.
5. Inject 1 mL of phenylephrine solution every 3–5 min.
   - Repeat injections of phenylephrine (up to the maximum dose of 1000 μg) should be continued until the erection resolves; only thereafter should this procedure be abandoned in favor of the more invasive approach of surgical shunt.
6. Wrap the detumescent penis in gauze or an elastic bandage to prevent the return of priapism and to compress the puncture site(s).

85.6 Complications

- Of cavernosal aspiration/irrigation
  - Hematoma (at puncture site)
  - Infection (at insertion site or systemic)
  - Thrombosis
  - Arteriovenous fistula
  - Pseudoaneurysm formation
  - Traumatic puncture of dorsal penile or urethra
  - Exsanguination (secondary to dislodgement of catheter)
  - Cerebrovascular accident (secondary to air embolism)
- Of intracavernosal injection of vasoactive agents (α-adrenergic sympathomimetics)
  - Fibrosis of the corpora, pain, penile necrosis, urinary retention
  - Phenylephrine toxicity
  - Acute hypertension, headache, reflex bradycardia, tachycardia, palpitations, cardiac arrhythmia

85.7 Pearls and Pitfalls

- Pearls
  - In ischemic priapism, the penis and corpora cavernosa are rigid and tender to palpation.
  - Ischemic priapism commonly results from an underlying hypercoagulable state, tumor, infection, neurological impairment ("spinal shock"), or vasoactive drug use.
  - During intracorporeal injection, the patient should be monitored for known side effects of sympathomimetics, including hypertension, headache, reflex bradycardia, tachycardia, palpitations, and cardiac arrhythmia. In addition, in patients with elevated cardiovascular risk profiles, blood pressure and electrocardiographic monitoring should be performed.
  - Seek a urological consult as soon as possible for any patient presenting with priapism.
  - Resolution of priapism can be verified by measurement of cavernous blood gases or measurement of blood flow by color duplex ultrasonography.
  - Phenylephrine is the sympathomimetic agent of choice for intracavernosal injection because it is has a lower likelihood of causing adverse cardiovascular side effects than other agents. If this is unavailable, alternatives include epinephrine, norepinephrine, ephedrine, metaraminol, and etilephrine.
  - Intracavernous aspiration/irrigation/injection therapy is unlikely to resolve ischemic priapism lasting for 48 h or longer. In such cases, immediate surgical shunting is first-line treatment.
  - Once detumescence is achieved, any unmetabolized drugs in the corpus cavernosa enter the venous circulation, and thus, dosages of any vasoactive drugs injected must be monitored carefully.
- Pitfalls
  - The most common complication of ischemic priapism is complete erectile dysfunction.

85.8 Considerations

85.8.1 Blood Gas Analysis

This investigation provides a rapid distinction between ischemic and nonischemic priapism. Blood aspirated from the corpus cavernosum in ischemic priapism is dark in color with partial pressure of oxygen (PO_2) less than 30 mmHg, partial pressure of carbon dioxide (PCO_2) greater than 60 mmHg, and pH less than 7.25. In nonischemic priapism, respective values will be PO_2 greater than 90 mmHg, PCO_2 less than 40 mmHg, and pH of 7.4 (Table 85.1).

Table 85.1 Summary of cavernosal blood gas findings

	pH	PO_2 (mm Hg)	PCO_2 (mm Hg)
Ischemic priapism	<7.25	<30	>60
Arterial blood	7.40	>90	<40
Mixed venous blood	7.35	40	50

PCO_2 partial pressure of carbon dioxide, *PO2* partial pressure of oxygen

85.8.2 Sickle Cell Testing

The sickle-solubility test detects any sickle hemoglobin (therefore, it is positive in patients with either sickle cell train or sickle disease). Hemoglobin electrophoresis with 10 % or greater HbS suggests sickle cell disease. Anemia and increased reticulocyte count may also be present in sickle cell disease.

85.8.3 Hemoglobin Electrophoresis

Confirmatory test for sickle status after a positive sickle-solubility test.

85.8.4 Complete Blood Count

White blood cell (WBC) count may suggest infection or blood dyscrasia. Hemoglobin (Hb) and reticulocyte counts may suggest sickle cell disease.

85.8.5 Color Duplex Ultrasonography

Blood flow in cavernosal arteries is absent or minimal in ischemic priapism, whereas flow velocity is normal to high in nonischemic priapism.

85.8.6 Urine Toxicology and Psychoactive Drug Screen

The following drugs have been associated with priapism: antihypertensives, anticoagulants, antidepressants, alcohol, marijuana, cocaine, and other illegal substances. Intracavernous injection therapy using drugs such as alprostadil, papaverine, prostaglandin E1, phentolamine, and others may precipitate priapism.

Selected Reading

Burnett AL, Bivalacqua TJ. Priapism: new concepts in medical and surgical management. Urol Clin North Am. 2011;38:185–94.
Dubin J, Davis JE. Penile emergencies. Emerg Med Clin North Am. 2011;29:485–99.
Montague DK, Jarow J, Broderick GA, et al; Members of the Erectile Dysfunction Guideline Update Panel; American Urological Association. American Urological Association guideline on the management of priapism. J Urol. 2003;170:1318–24.
Shrewsberry A, Weiss A, Ritenour CW. Recent advances in the medical and surgical treatment of priapism. Curr Urol Rep. 2010;11:405–13.
Vilke GM, Harrigan RA, Ufberg JW, Chan TC. Emergency evaluation and treatment of priapism. J Emerg Med. 2004;26:325–9.

86

Reduction of Phimosis/Paraphimosis

Justin Chen and Muhammad Waseem

Phimosis occurs when the distal aspect of the prepuce cannot be retracted over the glans.

Paraphimosis is a true urological emergency; it occurs when retracted paraphimotic foreskin cannot be replaced to its normal position past the coronal sulcus, resulting in venous and lymphatic congestion leading to arterial occlusion, ischemia, and necrosis of the glans.

Both conditions commonly result from chronic infection from poor local hygiene in uncircumcised males; it can also be due to redundant skin.

86.1 Indications

- Phimosis: signs of acute urinary retention
- Paraphimosis: signs of present or impending *arterial* occlusion

86.2 Contraindications

- Absolute
  - Failure to rule out penile swelling and pain due alternative conditions (e.g., posthitis/balanoposthitis, angioedema, insect bite, constricting band)

86.3 Materials and Medications

- Latex-free gloves (sterile)
- Local anesthetic
  - 2 % lidocaine without epinephrine (preferred)
  - 2 % lidocaine gel or eutectic mixture of local anesthetics (EMLA) cream (2.5 % prilocaine and 2.5 % lidocaine)
- 25- to 27-gauge 1.5-in. needles (2)
- Small plastic syringe, 10 mL (1)
- Bag of ice (1)
- 2-inch elastic pressure dressing (1)
- Sterile gauze (1)

J. Chen, MD, MSc
Department of Emergency Medicine, North Shore University Hospital, Manhasset, NY, USA
e-mail: chen.justin@gmail.com

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_86

86.4 Procedure: Manual Reduction for Paraphimosis

1. Grasp the swollen foreskin of the penis and apply gentle compression for a few minutes.
2. Grasp the swollen foreskin and elevate upward.
3. Push the glans into the foreskin.
4. Place the patient in the supine position and carefully inspect the penis for constricting bands or foreign bodies (e.g., piercings).
5. Usage of penile block to provide analgesia to the shaft and glans penis depends on urgency, patient age, and cooperativeness (Fig. 86.1).
   (a) Use a 25- to 27-gauge needle to inject lidocaine into the base of the penis, at the junction between the penis and the suprapubic skin, away from the midline to avoid the superficial dorsal vein.
   (b) Inject lidocaine just deep to Buck's fascia (3–5 mm beneath the skin) to form wheals, where a slight "pop" is felt as the needle penetrates the fascial layer.
   (c) Between 1 and 5 mL of local anesthetic should be used, depending on the age of the patient, which can be delivered as follows:
      - Half the volume should be injected at the 10 o'clock position and the remainder at the 2 o'clock position *or*
      - Full volume of local anesthetic is injected midline through Buck's fascia, directed toward each direction, ensuring negative aspiration of blood.
6. Alternatively, use gauze soaked in topical local anesthetic to cover the penis (2 % lidocaine gel or EMLA cream [2.5 % prilocaine and 2.5 % lidocaine]), which has slower onset.
7. For phimosis
   (a) If acute urinary retention occurs, dilate the foreskin under procedural sedation or penile block to allow Foley catheterization.
8. For paraphimosis
   (a) Once penile analgesia is achieved (typically in 5 min with penile block), relieve tissue edema before attempting reduction by using:
      - Bag of ice (3 min at a time), *or*
      - Granulated sugar (*contraindicated* in emergent situations owing to time required), *or*
      - Manual compression (squeezing the glans for 5 min), *or*
      - Pressure dressing (2-in. elastic bandage over the gland for 5 min)
   (b) Place both thumbs over the glans, with both index fingers and long fingers surrounding the trapped foreskin proximal from the paraphimotic tissue.
   (c) Use the thumbs to push the glans back into the foreskin while pulling the trapped foreskin distally, which may require a few minutes of constant pressure (Fig. 86.2).
   (d) Can also attempt using Babcock (once in each quadrant) or Adson (3 and 9 o'clock positions) forceps to grasp the paraphimotic tissue (Fig. 86.3).
   (e) If ineffective owing to extreme tissue edema, seek emergent urological consultation.
   (f) Follow-up with urologist is always recommended; circumcision may be performed once infection and/or edema have resolved to prevent recurrence (Fig. 86.4).

Fig. 86.1 Penile block to provide analgesia to the shaft and glans penis

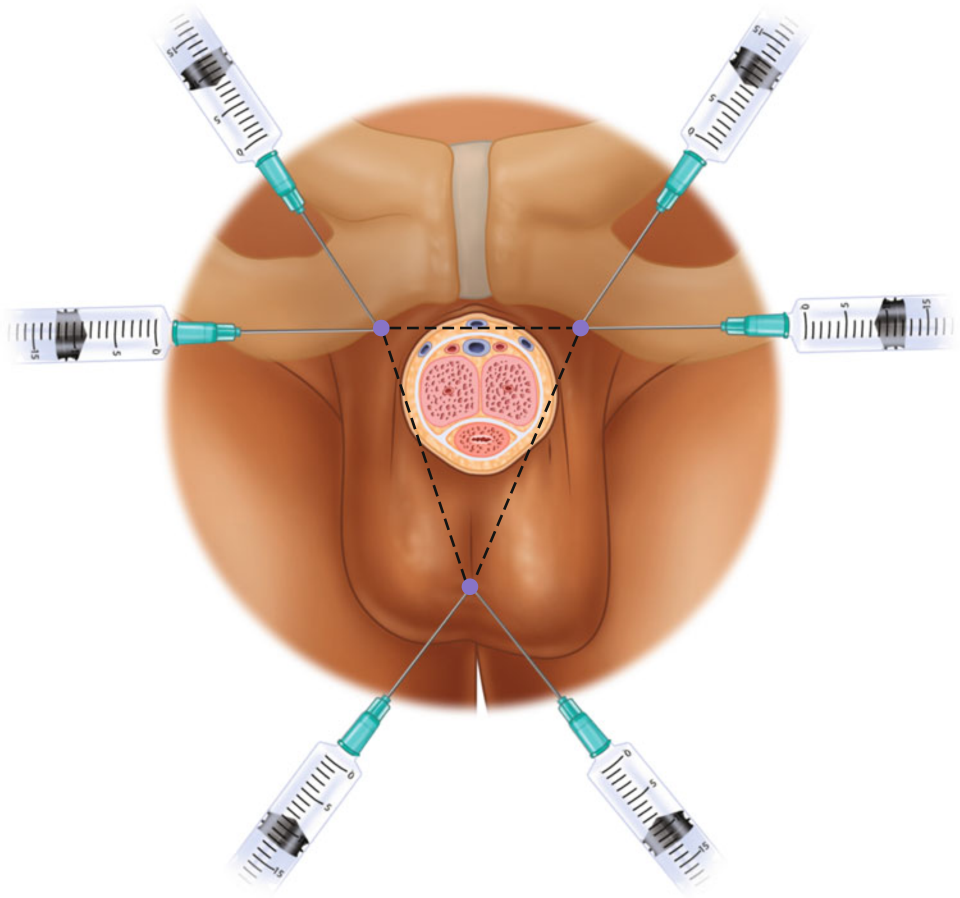

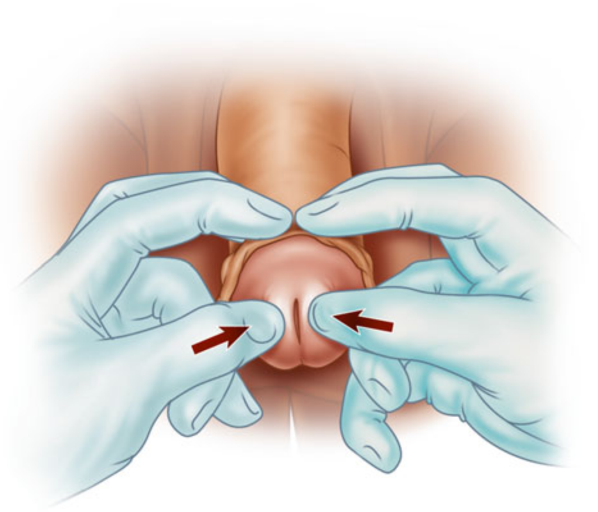

Fig. 86.2 Use the thumbs to push the glans back into the foreskin while pulling the trapped foreskin distally

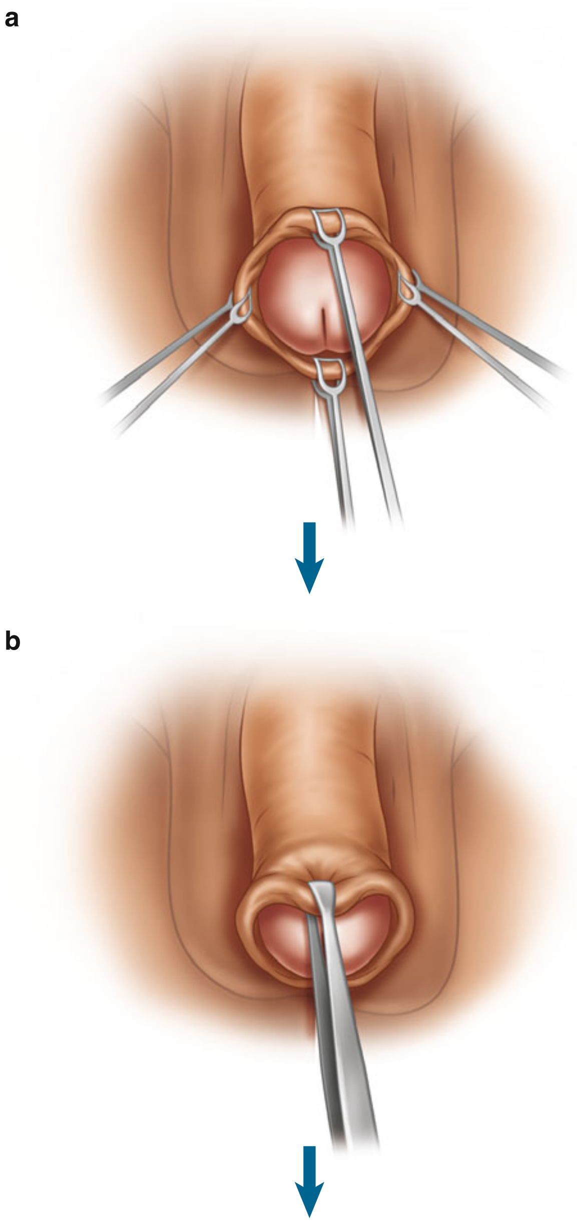

Fig. 86.3 (**a**) Babcock forceps. (**b**) Adson forceps

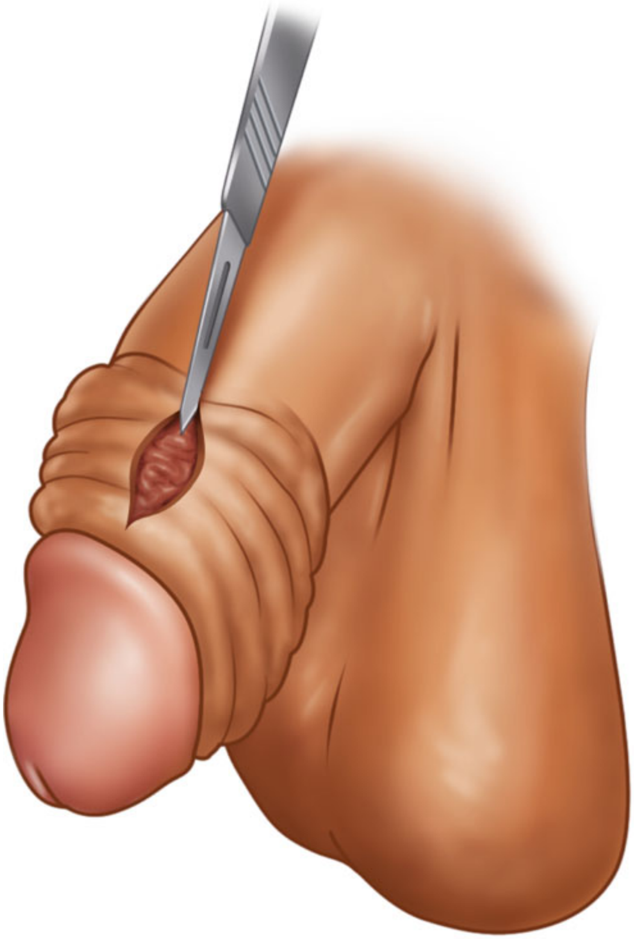

Fig. 86.4 Circumcision may be performed once infection and/or edema have resolved to prevent recurrence

86.5 Complications

- Bleeding/infection of injection site(s).
- Phimosis and scarring due to foreskin manipulation.
- Usage of Adson or Babcock forceps may result in minor bruising and abrasion to the foreskin and glans penis.

86.6 Pearls and Pitfalls

- Pearls
  - Phimosis is a normal occurrence in young males (<5–6 years) and should be treated only in the presence of acute urinary retention.
  - If arterial compromise is imminent in paraphimosis, the emergency physician should attempt reduction if urological consult is unavailable.
- Pitfalls
  - Reduction of phimotic tissue over the coronal sulcus may lead to emergent paraphimosis.

Selected Reading

Doherty GM, editor. Current diagnosis and treatment: surgery. 13th ed. New York: McGraw-Hill Medical; 2010.

King C, Henretig FM, editors. Textbook of pediatric emergency procedures. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.

Knoop KJ, editor. Atlas of emergency medicine. 3rd ed. New York: McGraw-Hill Professional; 2010.

Smith DR, Tanagho EA, McAninch JW, editors. Smith's general urology. New York: Lange Medical Books/McGraw-Hill; 2008.

Manual Testicular Detorsion

87

Brandon R. Allen and L. Connor Nickels

87.1 Indications

- Testicular torsion is a clinical diagnosis (Fig. 87.1).
  - Although no single clinical finding has 100 % sensitivity for the presence of testicular torsion, patients will have one or more of these signs and symptoms: nausea or vomiting, pain for less than 24 h, high position of the testis, and/or abnormal cremasteric reflex [1].
  - If the diagnosis is in question, radionuclide scan or ultrasonography of the testicles may be helpful to assess blood flow and to differentiate torsion from other conditions (Fig. 87.2).

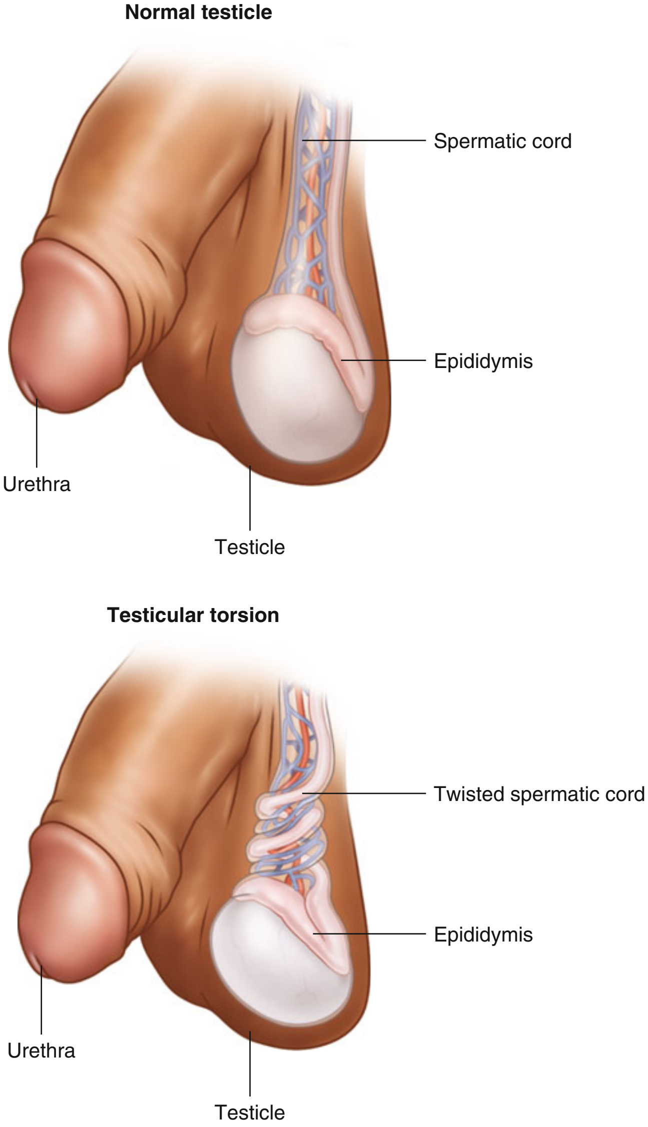
Normal testicle
Spermatic cord
Epididymis
Urethra
Testicle
Testicular torsion
Twisted spermatic cord
Epididymis
Urethra
Testicle

Fig. 87.1 Schematic of normal testicle and testicular torsion

B.R. Allen, MD • L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: brandonrallen@ufl.edu; cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_87

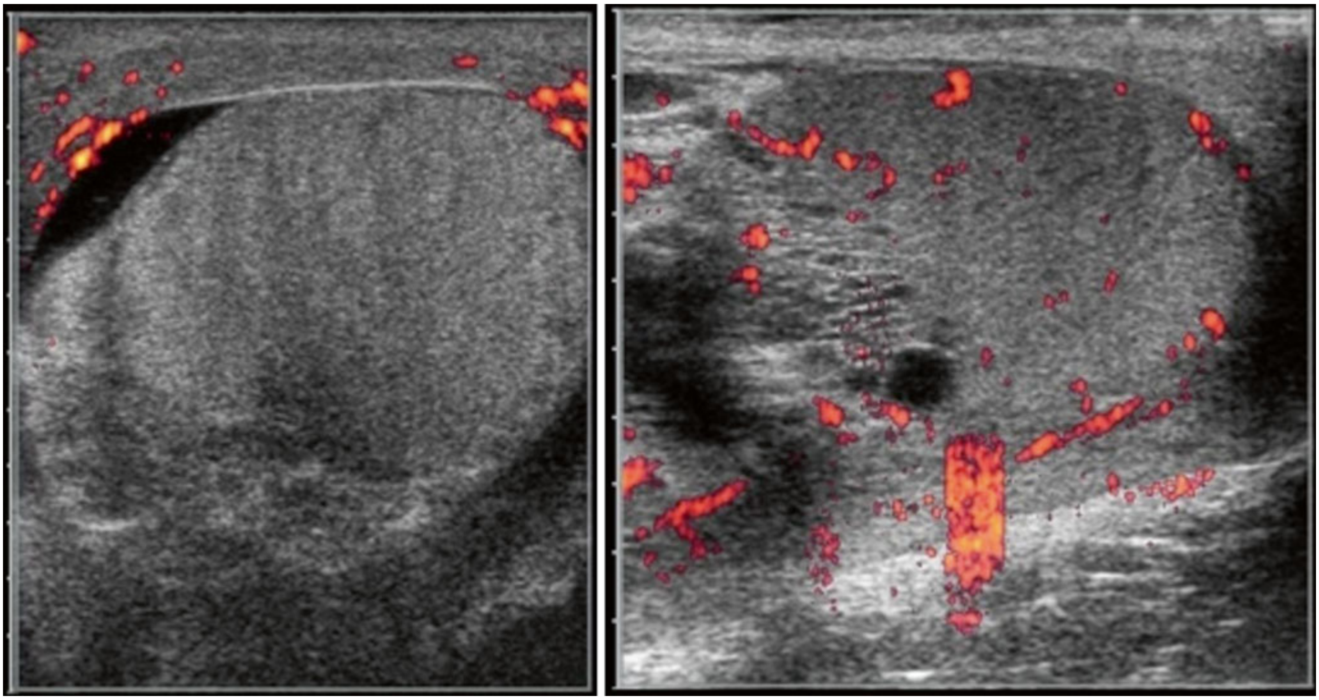

Fig. 87.2 Doppler ultrasound of bilateral testes shows swollen right testis with hypoechoic areas within and reduced arterial signal suggesting testicular torsion with necrosis (*left panel*). This is compared to the left testis which has normal flow (*right panel*) (Reproduced with permission from: Bhagra et al. [4])

87.2 Contraindications

- Do not attempt if the length of symptoms is greater than 24 h.
- Manual detorsion should not delay scrotal exploration and bilateral orchiopexy in the operating room [2].

87.3 Materials and Medications

- Standard precautions barrier protection for the provider
- Local anesthetic (optional to anesthetize the spermatic cord near the external ring)
- Intravenous sedation (optional)

87.4 Procedure

- The physician stands at the patient's feet and rotates the affected testicle away from the midline (as if opening a book) (Fig. 87.3).
  - For detorsion of the left testicle: the physician will place his or her right thumb and index finger on the affected testicle and rotate the testicle 180° from medial to lateral [2].
  - For detorsion of the right testicle: the physician will place his or her left thumb and index finger on the affected testicle and rotate the testicle 180° from medial to lateral.

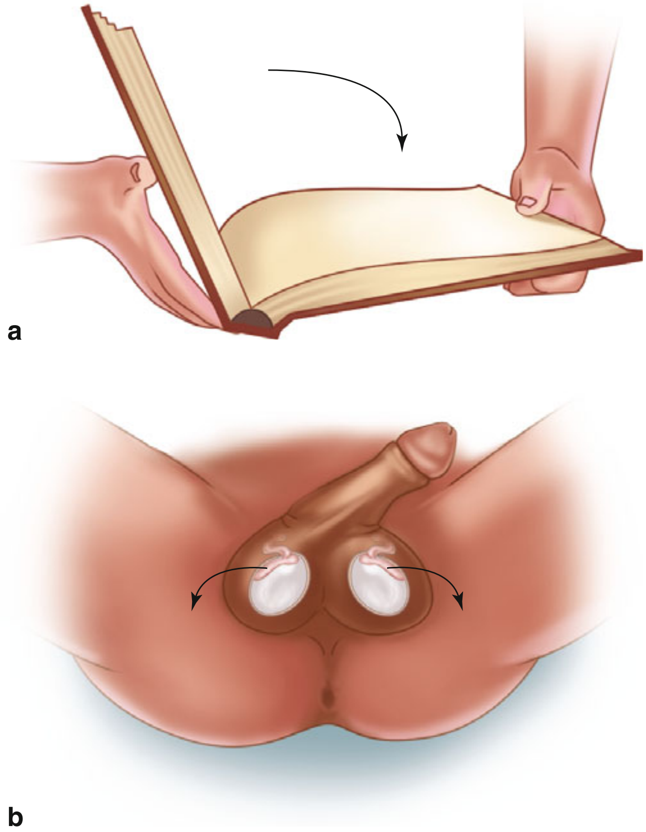

Fig. 87.3 (**a**) As if opening a book, rotate the testicle away from the midline. (**b**) Affected testicle should be rotated 180° from medial to lateral

87.5 Pearls and Pitfalls

- Pearls
  - The most common misdiagnosis is epididymitis.
  - Because torsion of greater than 360° degrees is possible, more than one rotation may be needed to fully detorse the affected testis (Fig. 87.4).
  - Only surgical exploration can provide a definitive resolution if testicular torsion is present [2].
- Pitfalls
  - The most common causes of testicular loss after torsion are delay in seeking medical attention (58 %), incorrect initial diagnosis (29 %), and delay in treatment at the referral hospital (13 %) [3]

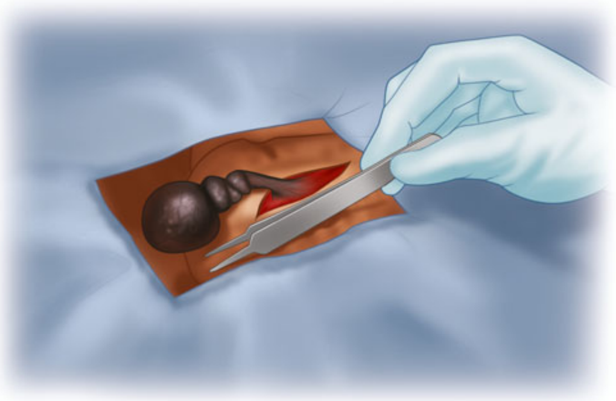

Fig. 87.4 Because torsion of greater than 360° degrees is possible, more than one rotation may be needed to fully detorse the affected testis

References

1. Beni-Israel T, Goldman M, Bar Chaim S, Kozer E. Clinical predictors for testicular torsion as seen in the pediatric ED. Am J Emerg Med. 2010;28:786–9.
2. Ringdahl E, Teague L. Testicular torsion. Am Fam Physician. 2006;74:1739–43.
3. Jones DJ, Macreadie D, Morgans BT. Testicular torsion in the armed services: twelve year review of 179 cases. Br J Surg. 1986;73:624–6.
4. Bhagra A, Suravaram S, Schears RM. Testicular torsion–a common surgical emergency. Int J Emerg Med. 2008;1(2):147.

Part XII

Skin and Soft Tissue Procedures

Local Anesthesia

88

Derek Ailes and Muhammad Waseem

88.1 Indications

- Laceration repair
- Abscess incision and drainage
- Wound exploration
- Vascular access procedures
- Foreign body removal
- Lumbar puncture

88.2 Contraindications

- History of allergy (usually to the ester class [e.g., procaine, tetracaine]), amide class (e.g., lidocaine [Xylocaine], bupivacaine, mepivacaine) may be safely substituted if true allergy to esters and vice versa. One percent diphenhydramine (4 mL normal saline: 1 mL 5 % intravenous [IV] diphenhydramine [Benadryl] mixture) can be used in patients with true allergy.
- Topical preparations on mucous membranes, burns, abraded/denuded skin, or eyes owing to potential toxicity from increased absorption and corneal injury.
- Common teaching is to avoid epinephrine-containing anesthetic solutions in nose, penis, and digits for concern of ischemia owing to end-artery constriction. Recent studies including prospective trials and comprehensive literature reviews, however, do not validate this concern [1, 2].

D. Ailes, MD • M. Waseem, MD (✉)
Department of Emergency Medicine,
Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: derekailes@yahoo.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_88

88.3 Materials and Medications (Fig. 88.1)

- 1 % Xylocaine with or without 1:200,000 epinephrine, 0.25 % bupivacaine solution. 8.4 % (1 mL/mL) sodium bicarbonate (optional)
- 18-, 25-, or 27-gauge needles, syringes up to 10 mL
- Sterile and nonsterile gloves, face shield
- Alcohol pads and povidone-iodine swabs

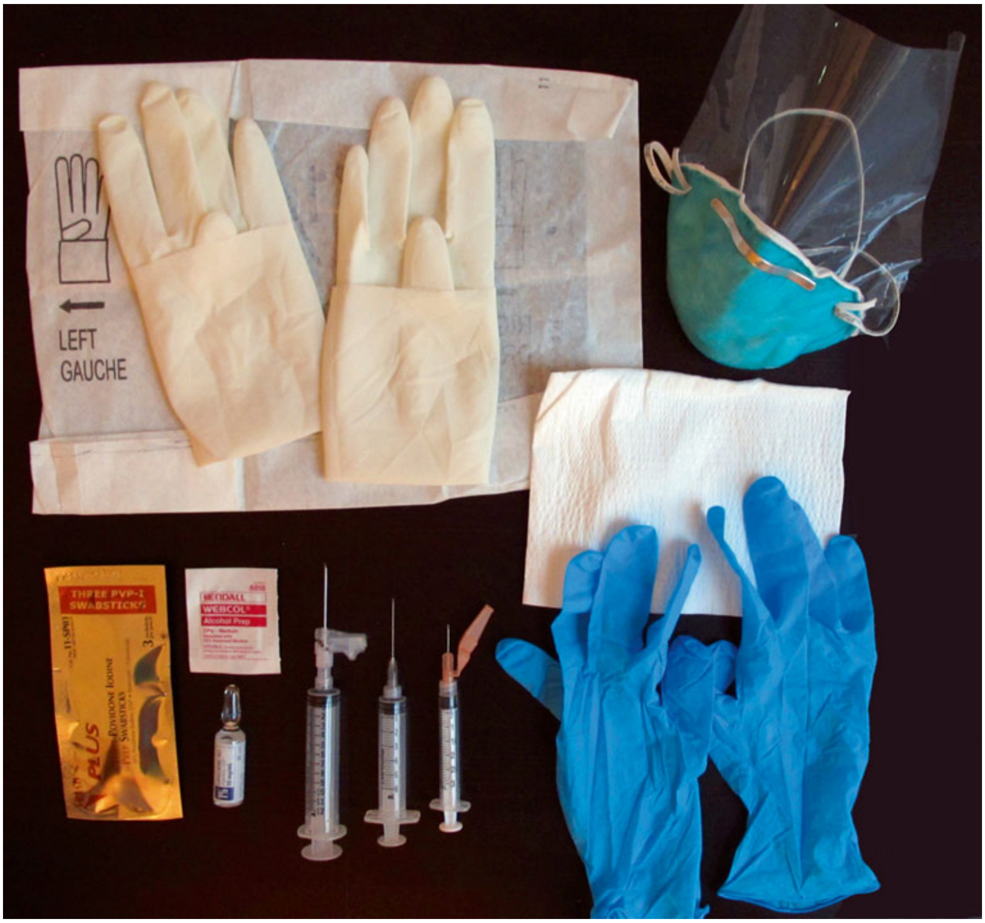
LEFT
GAUCHE
THREE PVP-I SWABSTICKS
POVIDONE IODINE
WEBCOL
Alcohol Prep

Fig. 88.1 Local anesthesia materials

88.4 Procedure

1. Position the patient in a comfortable position (supine, sitting, anticipating vasovagal response).
2. Draw the anesthetic with an 18-gauge needle into a syringe. Be aware of the maximum dose for the particular local anesthetic to avoid systemic toxicity.
3. Take steps to minimize the pain of infiltration (Table 88.1).
4. Prepare the area with povidone-iodine solution; cover surrounding areas with sterile drapes.
5. Inject subcutaneously by direct infiltration with a 25- to 27-gauge needle noting wheal and blanching. If a clean wound, may inject into wound edges. Usually no aspiration is needed because the infiltration is superficial to major blood vessels.
6. For contaminated wounds or abscess incision and drainage, perform field block by inserting the needle into clean intact skin adjacent to the wound and continuing in a circular manner around the wound, injecting into the previously anesthetized area (Fig. 88.2).
7. Wait several minutes for the local anesthetic to provide the maximum effect.
8. Test the area for sharp sensation with a needle tip or other sharp object.

Table 88.1 Common anesthetics and their characteristics

Medication	Time to onset (min)	Length of action	Maximum dose (mg/kg)	Maximum dose with epinephrine (mg/kg)
Marcaine (bupivacaine)	5–10	>3 h, up to 9 h with epinephrine	2.5	2.5–3.5
Xylocaine (lidocaine)	<2	30–75 min (longer with epinephrine)	3	5–7

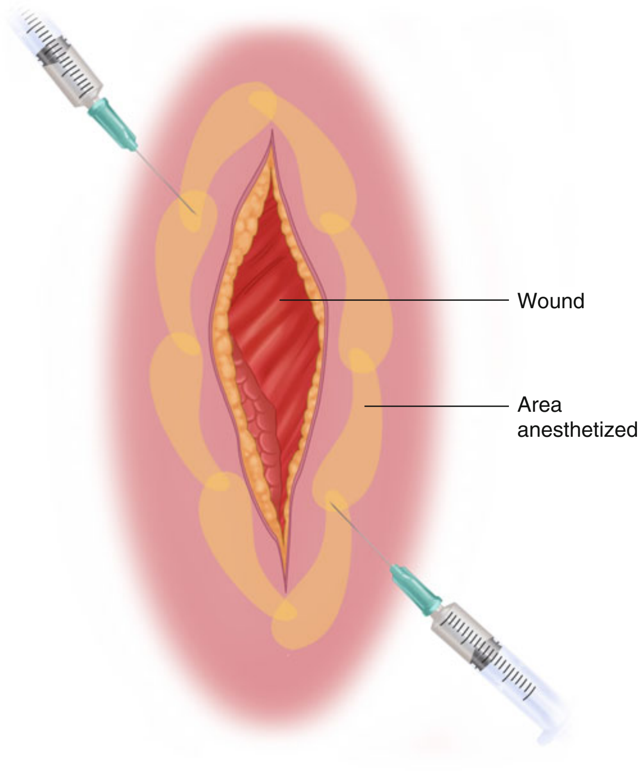

Fig. 88.2 Local infiltration should be performed in a circular fashion with each injection performed over the prior anesthetized area

88.5 Complications

- Systemic toxicity
- Allergic reaction
- Infection
- Digital artery vasospasm from accidental injection of epinephrine (can be reversed with topical nitroglycerine or subcutaneous phentolamine)
- Vasovagal response

88.6 Pearls

- Minimize or reduce the pain of infiltration by use of the following:
  - Warm Xylocaine before infiltration (blanket warmer or water bath) [3].
  - Buffer Xylocaine with 1 mL 8.4 % sodium bicarbonate for every 10 mL of Xylocaine. Buffer bupivacaine with 0.05–0.10 mL sodium bicarbonate for every 10 mL of bupivacaine (greater chance of precipitation).
  - Use a small-gauge needle (e.g., 27 gauge) and inject slowly.
  - Use a small syringe (1–3 mL) to reduce the pressure of injection.
  - Withdraw the needle and, just before exiting the skin, redirect and inject.
  - Inject in a circular manner around the wound with each subsequent injection entering a previously anesthetized area, such that the patient feels only one needle stick (Fig. 88.2).
  - Inject into the subcutaneous plane as opposed to the intradermal plane.
  - Consider using a topical anesthetic before infiltration, especially in pediatrics (lidocaine, epinephrine, tetracaine [LET]).
- Beware of toxicity by not exceeding the maximum dose, especially in large or multiple lacerations. Even at standard doses, toxicity can occur with inadvertent vascular injection, injection into highly vascular areas, or onto mucous membranes [4, 5].
  - Convert % mg/mL into mg/kg by moving the decimal one place to the right (e.g., 1 % Xylocaine becomes 10 mg/mL and 0.25 % bupivacaine becomes 2.5 mg/mL).
  - Xylocaine can be safely injected up to 3.5 mg/kg every 30 min, up to 300 mg/dose. If the mixture contains epinephrine, 5–7 mg/kg is safe.
  - Bupivacaine can be injected at 2.5 mg/kg and 3.5 mg/kg with epinephrine and can be injected every 3 h with daily maximum of 400 mg [5].
- When treating a wound, it is important to first anesthetize so debridement, cleansing, and irrigation can adequately be performed.
- Choose appropriate anesthetics. Xylocaine lasts approximately 75 min, and bupivacaine lasts several hours. Adding epinephrine to either increases vascular constriction, thereby decreasing systemic absorption and significantly increasing the duration of effect.
- Topical anesthetics have a role in pediatric populations and in conjunction with or as an alternate to local infiltrative anesthesia. TAC is a mixture of 0.5 % tetracaine, 0.05 % epinephrine, and 11.8 % cocaine. LET is 4 % lidocaine, 0.1 % epinephrine, and 0.5 % tetracaine. LET has been found to be safer and more cost-effective [6].

References

1. Muck AE, Bebarta VS, Borys DJ, Morgan DL. Six years of epinephrine digital injections: absence of significant local or systemic effects. Ann Emerg Med. 2010;56:270–4.
2. Waterbrook AL, Germann AC, Southall JC. Is epinephrine harmful when used with anesthetics for digital blocks? Ann Emerg Med. 2007;50:472–5.
3. Hogan ME, vanderVaart S, Perampalades K, Machado M, Einarson TR, Teddio A. Systematic review and meta-analysis of the effect of warming local anesthetics on injection pain. Ann Emerg Med. 2011;58:86–98.
4. Reichman EF, Simon RR, editors. Emergency medicine procedures. New York: McGraw-Hill Medical; 2004. p. 937.
5. Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010. p. 2425–7.
6. Kravitz ND. The use of compound topical anesthetics. J Am Dent Assoc. 2007;138:1333–9.

89 Regional Anesthesia (Nerve Blocks)

Derek Ailes and Muhammad Waseem

89.1 Indications

- Repair of wounds where preserving anatomical landmarks or having precise anatomical alignment is important (e.g., vermillion border of lip)
- Pain control in dislocation or fracture reductions
- Incision and drainage of abscesses
- Burn and wound care
- Extensive or multiple lacerations (reduces total amount of local anesthetic needed)
- Foreign body removal

89.2 Contraindications

- Allergic to local anesthetic (see Chap. 88)
- History of coagulopathy or bleeding disorder
- Injection through infected tissue

D. Ailes, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: derekailes@yahoo.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_89

89.3 Materials and Medications (Fig. 89.1)

- Povidone-iodine, alcohol swabs
- Sterile gloves and drapes
- Local anesthetic solution (e.g., lidocaine, Marcaine with or without epinephrine)
- 18-gauge, 20- to 30-gauge needles 2 in. in length
- 22- to 24-gauge spinal needles
- Syringes up to 60 mL

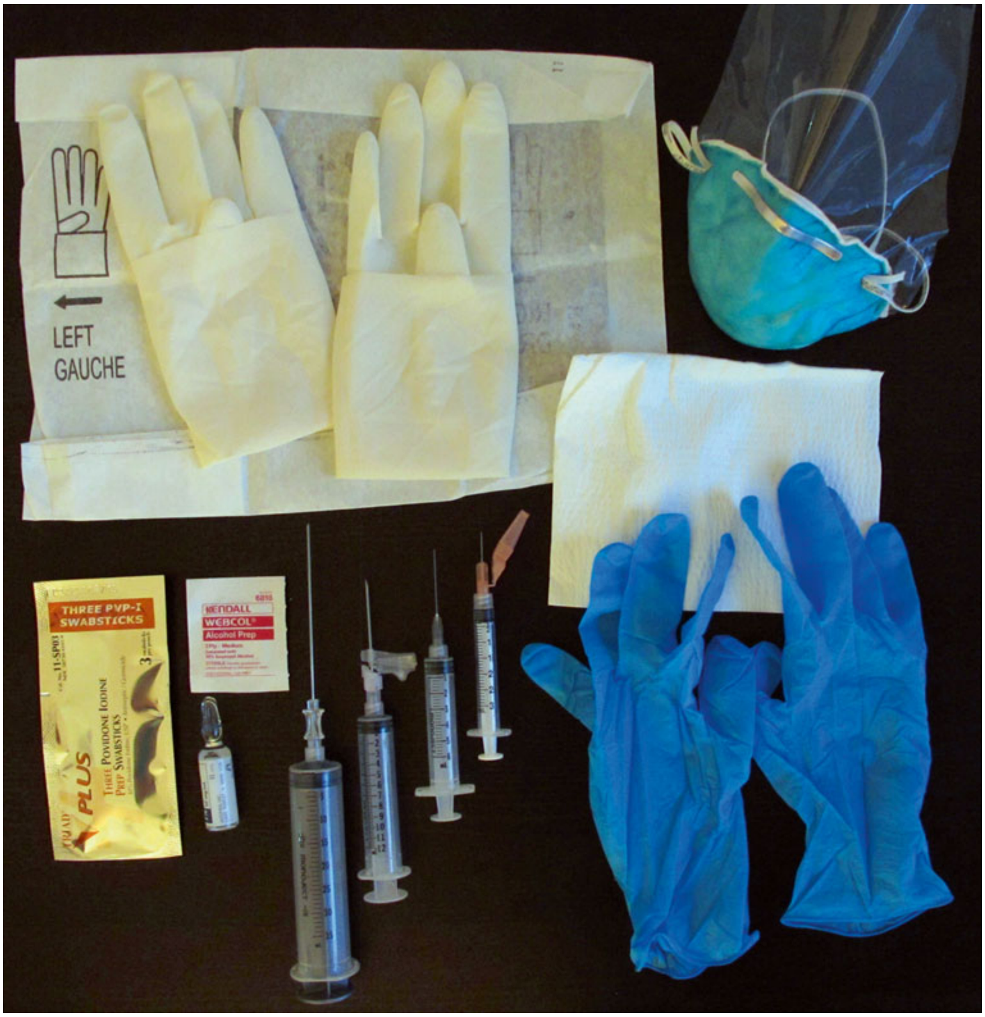
LEFT
GAUCHE
THREE PVP-I SWABSTICKS
PLUS
THREE POVIDONE IODINE PREP SWABSTICKS
KENDALL
WEBCOL
Alcohol Prep

Fig. 89.1 Materials

89.4 Procedure: General Block

1. Obtain consent after explaining the risks of procedure including temporary paresthesias and expected duration of block. Perform neurological examination before procedure, documenting any preexisting deficits.
2. Position the patient comfortably, preferably supine, anticipating vasovagal response.
3. Identify landmarks for the block. Clean area, prepare with povidone-iodine, and surround with sterile drapes.
4. A small skin wheal of local anesthetic may be placed at the site of needle entry before block.
5. Insert the needle into the site while aspirating to ensure it is not in a vessel.
6. If paresthesia is elicited, withdraw the needle slightly allowing paresthesia to improve and inject.
7. Wait 5–15 min for the block to reach full effect.
8. Test for sharp sensation in the anesthetized area and document.

89.5 Complications

- Infection
- Hemorrhage
- Hematoma
- Allergic reaction
- Systemic toxicity (exceeded maximum dose or inadvertent injection into vasculature)
- Paresthesias, pain
- Intraneural injection causing ischemia
- Intra-arterial injection of epinephrine causing vasospasm and tissue ischemia

89.6 Pearls and Pitfalls

- Pearls
  - It is important to aspirate before injecting anesthesia when performing regional anesthesia because, unlike local techniques, the needle is deeper and in proximity to larger vessels.
  - Shooting pain and/or paresthesias occurs when the needle contacts the nerve. When this happens, withdraw the needle 1 mm and wait for the paresthesia to resolve before injecting.
  - Injury can occur to a limb or digit if the patient manipulates it before the anesthesia wears off. The patient should be cautioned not to use the affected area until motor and sensation return. If an extensive block was done, monitor the patient in the emergency department until return of baseline neurological function.
- Pitfalls
  - It is traditionally taught to avoid epinephrine-containing solutions in blocks in end-artery blocks (e.g., digital blocks). However, evidence for any vascular insufficiency and necrosis as a result is lacking in standard commercially available lidocaine with epinephrine preparations. They should, however, be avoided in patients with peripheral artery disease [1, 2].

89.7 Selected Specific Blocks

- Contraindications and materials are the same as general block.

89.7.1 Facial Blocks: Trigeminal Nerve (Fig. 89.2) [2, 3]

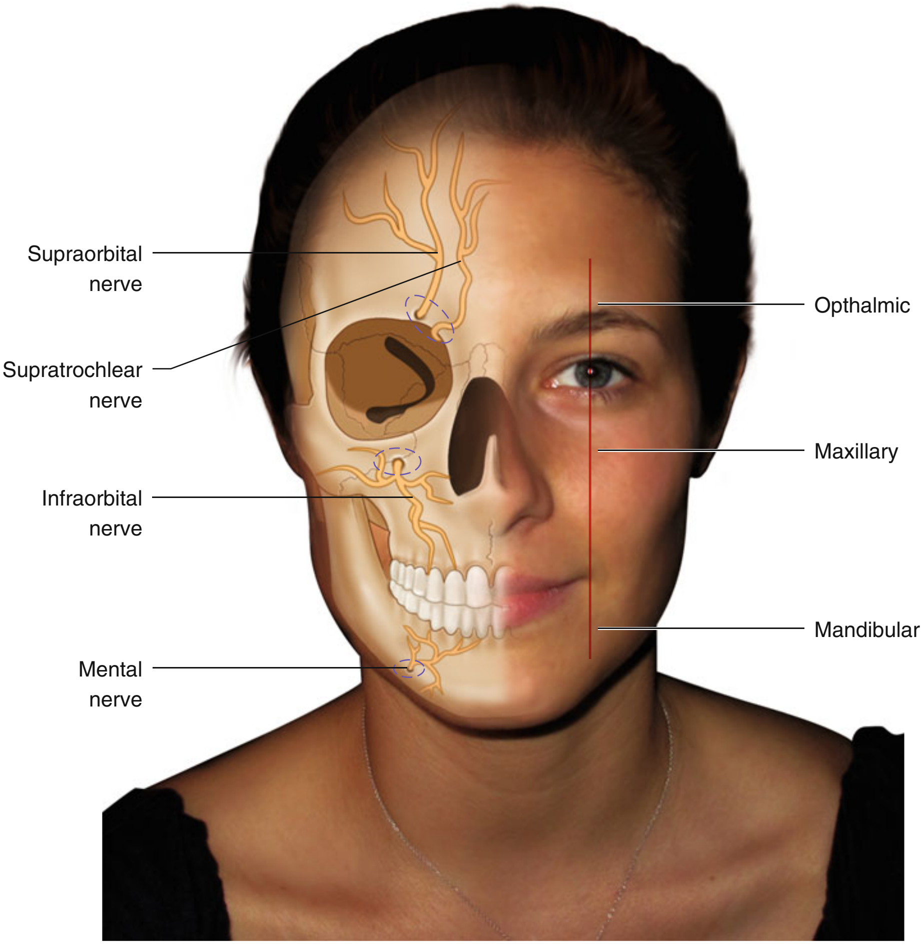
Supraorbital nerve
Supratrochlear nerve
Infraorbital nerve
Mental nerve
Opthalmic
Maxillary
Mandibular

Fig. 89.2 Vertical plane through the midposition of the pupil shows the position of the supraorbital foramen, infraorbital foremen, and mental foramen

89.7.1.1 Supraorbital (Fig. 89.3) and Supratrochlear Nerve Block

- Indications
  - To anesthetize the forehead from the orbital ridge to the vertex of the scalp. The supraorbital nerve emerges from the supraorbital foramen/notch and is a branch of the ophthalmic division of the trigeminal nerve. The supratrochlear nerve also is a branch of the ophthalmic division of the trigeminal nerve and exits through the superior medial aspect of the orbit.
- Procedure
  1. Inject local anesthetic solution over the midline of the forehead at eyebrow level.
  2. Inject a 25- or 27-gauge needle through skin wheal aimed laterally while injecting 3–5 mL local anesthetic subcutaneously. Stop infiltrating when the needle reaches the midline of the orbit.

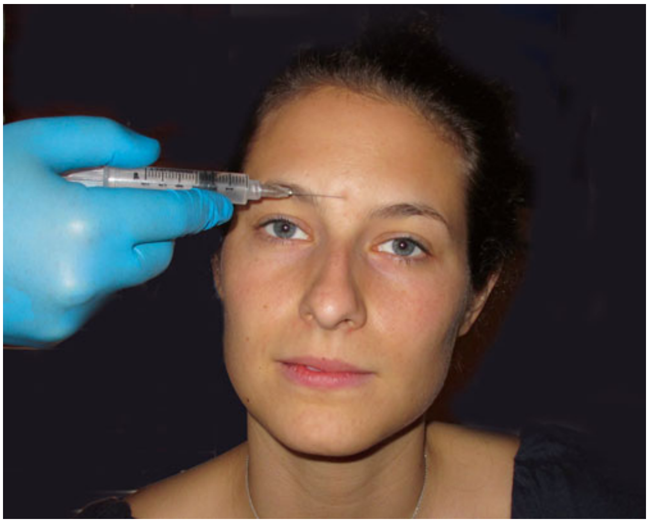

Fig. 89.3 Supraorbital nerve block

89.7.1.2 Infraorbital Nerve Block

- Indications
  - To anesthetize the medial cheek, upper lip, philtrum, skin between the lips and the nose, and nasal ala. The infraorbital nerve emerges from the infraorbital foramen and is a branch of the maxillary division of the trigeminal nerve. Anesthesia to the infraorbital nerve will also provide anesthesia to its terminus, the superior alveolar nerves.
- Procedure: Extraoral Approach (Fig. 89.4)
  1. Palpate the inferior orbital foramen in its midline position. The infraorbital nerve is often tender on palpation as it exits the foramen.
  2. Inject a 25- or 27-gauge needle just above the infraorbital foramen injecting 1–2 mL of local anesthetic.
  3. Take care not to inject into the foramen because there is an increased risk of intraneural injection.
  4. Hold a finger on the inferior orbital rim to avoid ballooning of the lower eyelid with injection.
  5. Intraoral approach is possible and preferred because it is less painful.

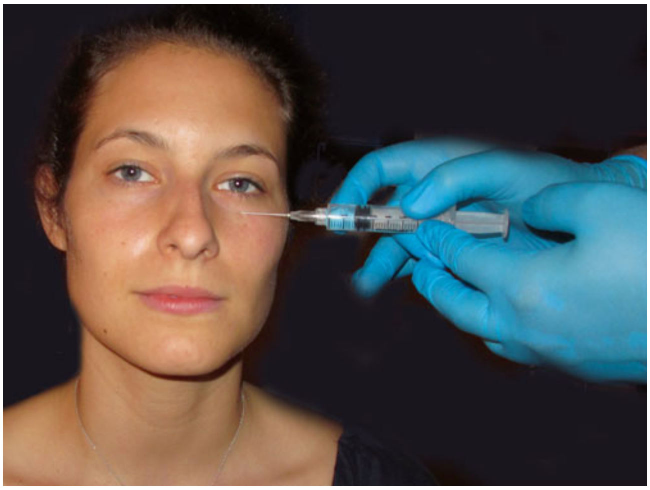

Fig. 89.4 Extraoral approach to the infraorbital nerve block

- Procedure: Intraoral Approach (Fig. 89.5) [3]
  1. Apply topical benzocaine or lidocaine gel to the point of insertion, which is the height of the mucobuccal fold over the first premolar, which is the site of insertion. Wipe off after 1–3 min.
  2. Palpate with the finger of the noninjecting hand over the inferior border of the inferior orbital rim. Retract the lip with the noninjecting hand.
  3. Using a long 25- to 27-gauge needle, with the bevel toward the bone, advance the needle at the insertion site toward the infraorbital foramen. Once the target is reached, aspirate and inject 1 mL of local anesthetic.
  4. Exert pressure on the foramen for 1 min after injection to force the anesthetic through the infraorbital foramen.
  5. If the needle is difficult to advance and the patient experiences pain on insertion, redirect the needle laterally and advance.
  6. If analgesia is attained for the lip but not the eyelid, the analgesia was placed inferior to foramen, and if analgesia is attained for the eyelid but not the lip, placement was superior to the foramen.

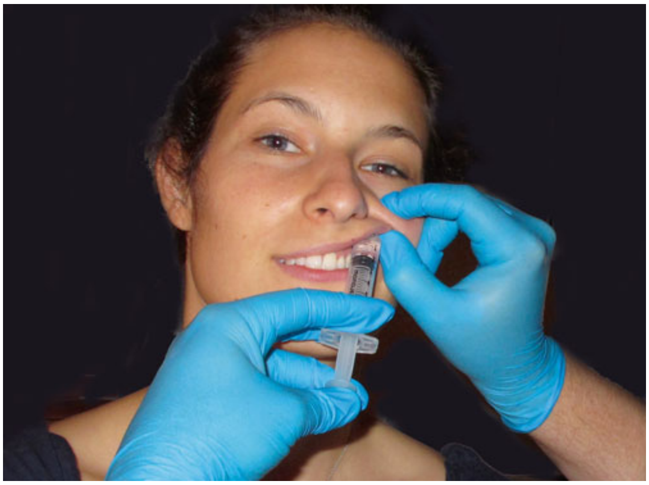

Fig. 89.5 Intraoral approach to the infraorbital nerve block

89.7.1.3 Mental Nerve Block

- Indications
  - To anesthetize the lower lip and chin and is especially useful in laceration repair at those sites. The mental nerve emerges from the mental foramen and is a branch of the mandibular division of the trigeminal nerve. Mental foramen lies in the vertical plane with the midpoint of the pupil and sits in the middle of the body of the mandible.
- Procedure: Extraoral Approach (Fig. 89.6)
  1. Inject local anesthetic solution over the identified location of the mental foramen, creating a skin wheal.
  2. Advance a 25- or 27-gauge needle through the skin wheel until the mandible is contacted, injecting 1–2 mL of local anesthetic.
  3. Intraoral approach is possible and preferred because it is less painful.
- Procedure: Intraoral Approach (Fig. 89.7) [3]
  1. Apply topical benzocaine or lidocaine gel to the point of insertion, which is the mucobuccal fold between the apices of the first and the second premolars. Wipe off after 1–3 min.
  2. Insert a 25- to 27-gauge needle, with the bevel toward the mandible, aimed toward the mental foramen.
  3. After advancing one-third the depth of the mandible and contacting the mandible, inject 1–2 mL of local anesthetic.
  4. By pressing firmly on the mental foramen for 2–3 min after the mental foramen has been blocked, an incisive nerve block is also created. This is useful if anesthesia to the lower anterior teeth is also desired.

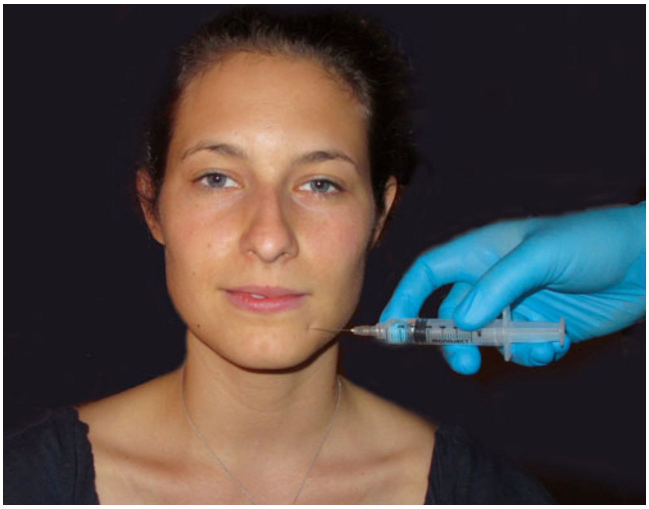

Fig. 89.6 Extraoral approach to the mental nerve block

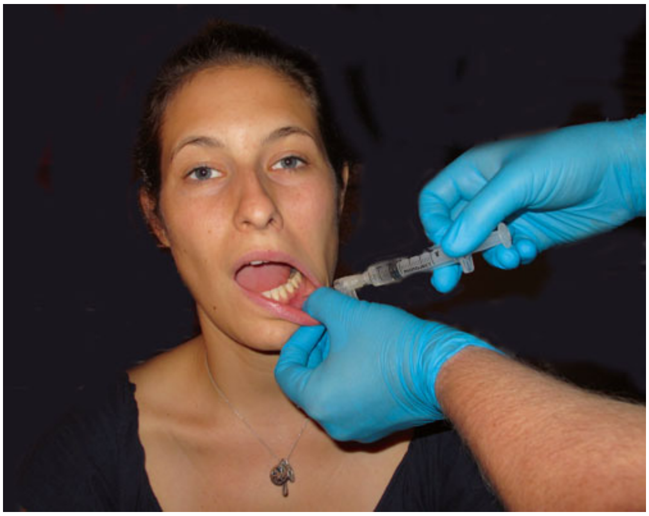

Fig. 89.7 Intraoral approach to the mental nerve block

89.7.2 External Ear Block (Fig. 89.8) [4]

- Indications
  - To anesthetize the entire external ear, excluding the external auditory canal and the concha
  - Especially useful in large lacerations of the ear and surrounding skin, hematoma evacuations, or incision and drainage of abscess.
- Procedure: Auricular Ring Block (Fig. 89.9) [4]
  1. Using a 25- to 27-gauge needle, insert the needle just inferior to the earlobe directing it toward the tragus.
  2. Aspirate and advance the needle superiorly subcutaneously injecting 3–4 mL of local anesthetic (Fig. 89.9, #1).
  3. Withdraw the needle without fully removing it and redirect it posterosuperiorly along the inferior posterior auricular sulcus, aspirating and injecting as before (Fig. 89.9, #2).
  4. Remove the needle and insert it just superior to the point of helix insertion into the scalp.
  5. Advance the needle and aspirate and inject in the direction of the tragus. Inject into the subcutaneous tissue while avoiding the ear cartilage (Fig. 89.9, #3).
  6. Withdraw and redirect the needle posteriorly and inferiorly toward the skin behind the ear, injecting as before (Fig. 89.9,#4).
  7. Beware of inadvertent cannulation of the superficial temporal artery, which crosses the zygomatic arch and crosses medial to the ear. If the artery is violated, it requires 20–30 min application of firm pressure.

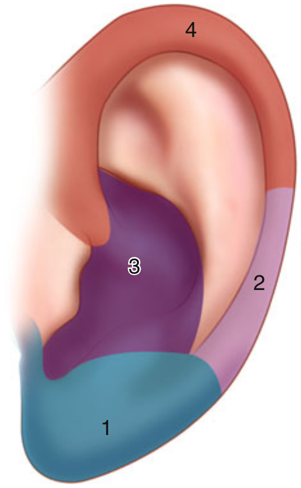

Fig. 89.8 Auricular block anesthetizes four nerves that innervate the auricle. *1* Great auricular nerve, *2* lesser occipital nerve, *3* auricular branch of vagus nerve, *4* auriculotemporal nerve

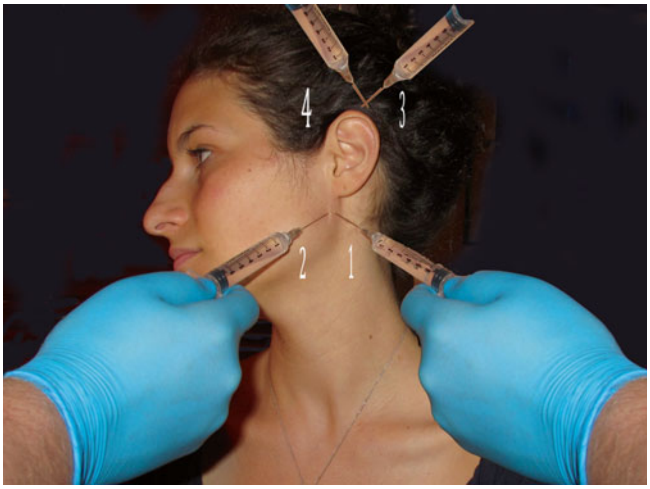

Fig. 89.9 Auricular ring block technique

89.7.3 Wrist Block (Fig. 89.10) [3, 5, 6]

- Indications
  - To anesthetize the hand in preparation for laceration repair, fracture or dislocation reduction, or pain relief

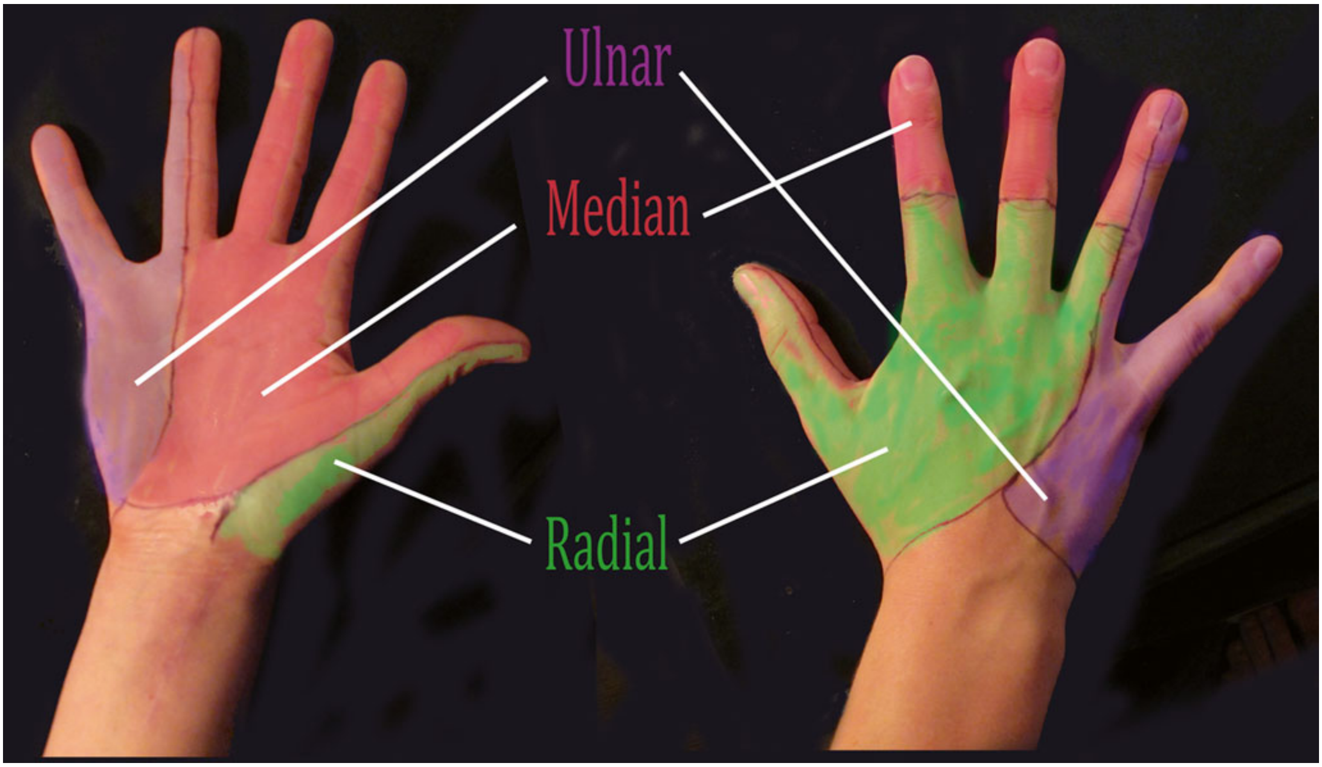

Fig. 89.10 Nerve distribution in the hand

89.7.3.1 Wrist Block: Median Nerve (Fig. 89.11)

- Procedure
  1. Position the patient supine with the palmar surface of the hand face up.
  2. Have the patient make a fist and slightly flex the wrist so that the palmaris longus and flexor carpi radialis tendons become prominent.
  3. Create a skin wheal of local anesthetic between the two tendons between the proximal skin crease and the distal skin crease at the wrist. Alternatively, anesthetic can be injected in line with the ulnar styloid process at the proximal skin crease.
  4. The palmaris longus tendon is absent normally in 10–20 % of the population. In this case, inject over the midpoint of the proximal skin crease at the level of the styloid process.

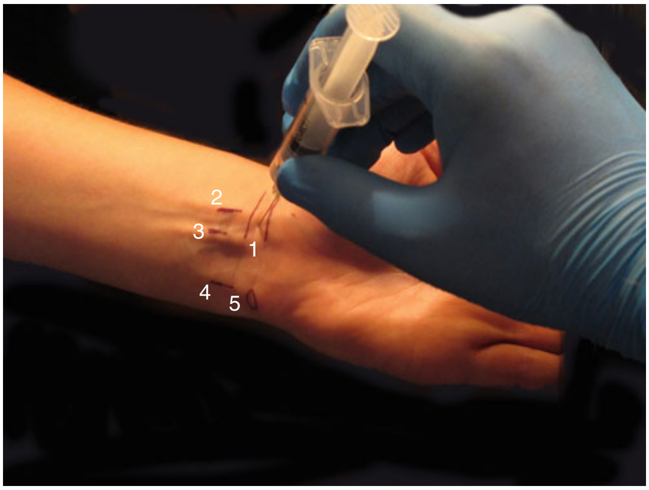

Fig. 89.11 Median nerve block at the wrist: *1* proximal and distal wrist creases radial artery, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* styloid process

89.7.3.2 Wrist Block: Radial Nerve (Figs. 89.12 and 89.13) [3, 7]

- Procedure
  1. To block the radial nerve, block multiple peripheral branches on the dorsal and the radial aspects of the lateral wrist.
  2. A field block in and around the anatomical snuffbox may be performed, requiring roughly 5–6 mL of local anesthetic. This anesthetizes the terminal branches of the radius arising from the forearm.
  3. Position the patient's palm face up and inject a skin wheal to the area 1 mm lateral to the radial pulse and in line with the proximal wrist crease. Inject 2 mL of local anesthesia with a 25- to 27-gauge needle. This anesthetizes the terminal trunk of the radial nerve.

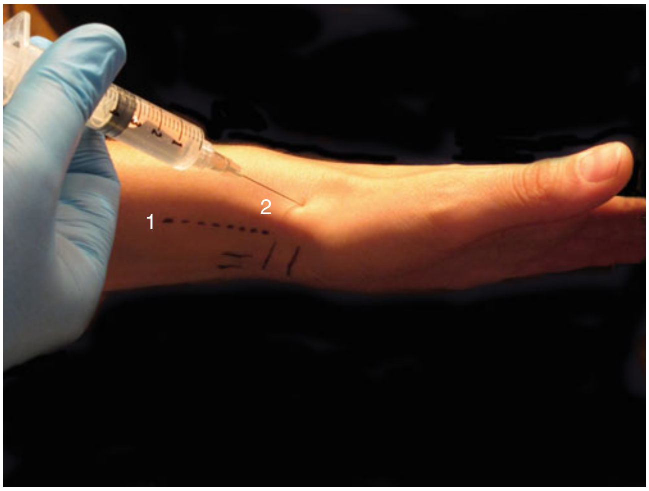

Fig. 89.12 Radial nerve block at the wrist: *1* radial artery, *2* anatomical snuffbox

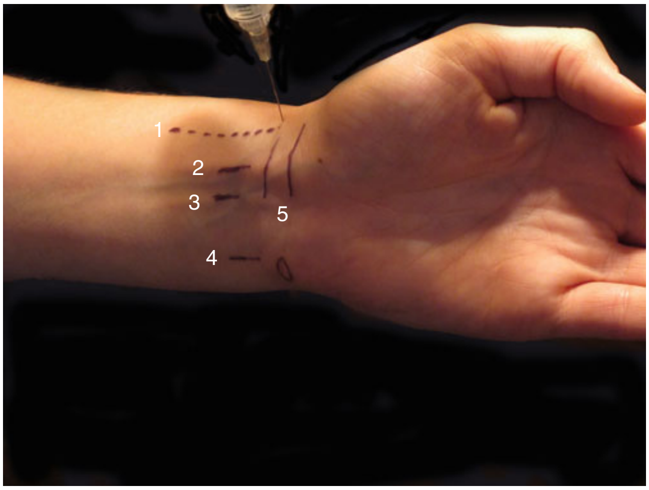

Fig. 89.13 Radial nerve block at the wrist: *1* radial artery, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* proximal and distal wrist creases

89.7.3.3 Wrist Block: Ulnar Nerve (Fig. 89.14)

- Procedure
  1. Position the patient with the hand face up.
  2. Palpate the styloid process of the ulna and pisiform and find the ulnar artery pulse.
  3. Create a skin wheal of local anesthetic between lateral to the ulnar artery and medial to the flexor carpi ulnaris tendon just proximal to the styloid process. This is at the level of the proximal wrist crease.

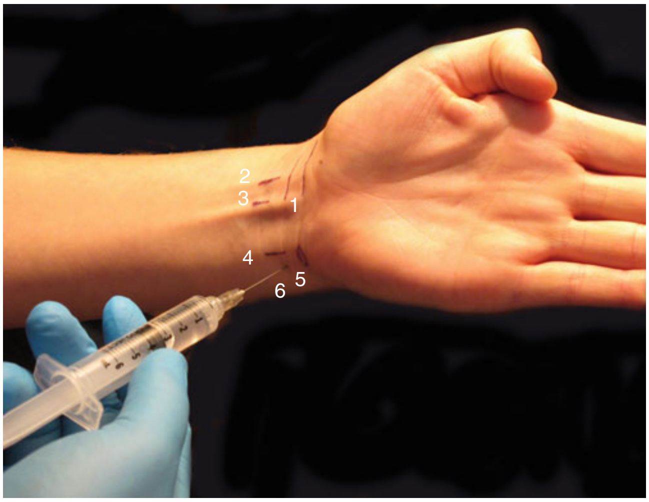

Fig. 89.14 Ulnar nerve block at the wrist: *1* proximal and distal wrist creases, *2* flexor carpi radialis tendon, *3* palmaris longus tendon, *4* ulnar artery, *5* styloid process, *6* flexor carpi ulnaris

89.7.4 Digital Nerve Blocks: Ring, Web Space, and Tendon Sheath (Fig. 89.15) [3, 8, 9]

- Indications
  - To anesthetize the digits in preparation for laceration repair, nail bed repair, joint reduction, or pain relief

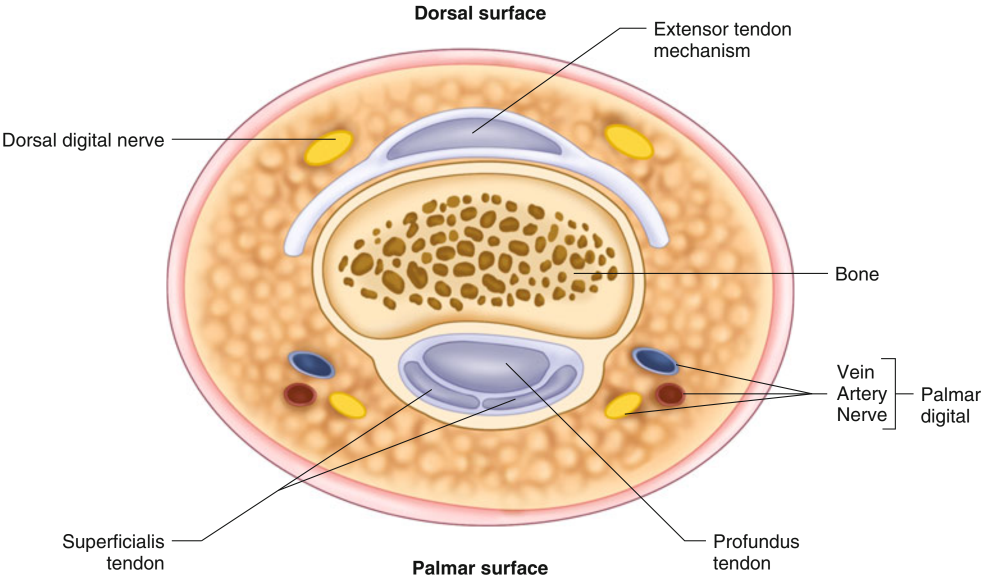
Dorsal surface
Extensor tendon mechanism
Dorsal digital nerve
Bone
Vein
Artery
Nerve
Palmar digital
Superficialis tendon
Profundus tendon
Palmar surface

Fig. 89.15 Cross section of finger

89.7.4.1 **Ring Block** (Fig. 89.16) [3]

- Procedure
  1. Insert a 25-gauge needle on the dorsal surface of the proximal phalanx of the digit to be anesthetized. Inject 1 mL along the dorsal surface and withdraw the needle.
  2. Reinsert the needle again perpendicular to the last injection and running on the lateral surface of the phalanx. Inject 1–1.5 mL of local anesthetic to just past the phalanx base.
  3. Repeat the injection in the same fashion on the medial aspect of the phalanx.
  4. Do not inject more than 5 mL into a digit.
  5. Toe blocks are similar to finger ring blocks, except that the great toe requires plantar surface injection as well, owing to its unique nerve supply.

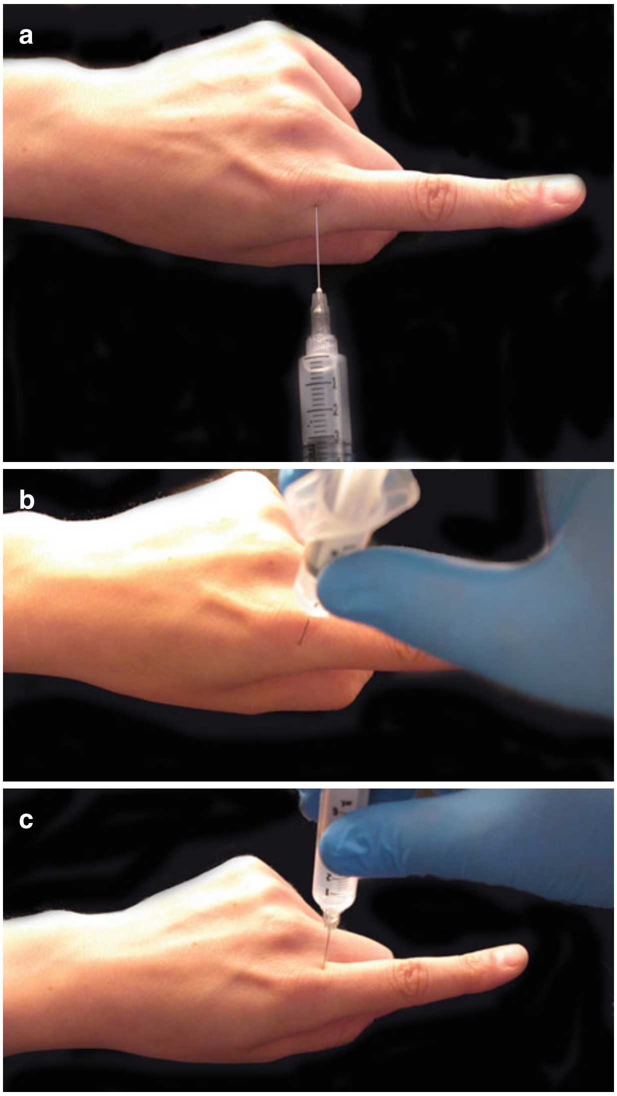

Fig. 89.16 Dorsal surface injection of digital nerve ring block. Digital nerve block is performed by injecting onto the (**a**) dorsal, (**b**) lateral, and (**c**) medial surfaces of the proximal phalanx

89.7.4.2 **Web Space Digital Block** (Fig. 89.17) [9]

- Procedure
  1. Have the patient abduct the fingers.
  2. Palpate the metacarpophalangeal joint and then insert a 25- to 27-gauge needle into the lateral web space subcutaneously, directing it dorsally. Aspirate then inject 1 mL of local anesthetic.
  3. Withdraw the needle but before the exiting skin, redirect toward the palmar aspect until the tip is next to the metacarpophalangeal joint, and inject 1 mL of local anesthetic.
  4. Repeat the procedure on the medial web space of the digit. Each digit blocked requires injection on both the lateral and the medial web spaces.

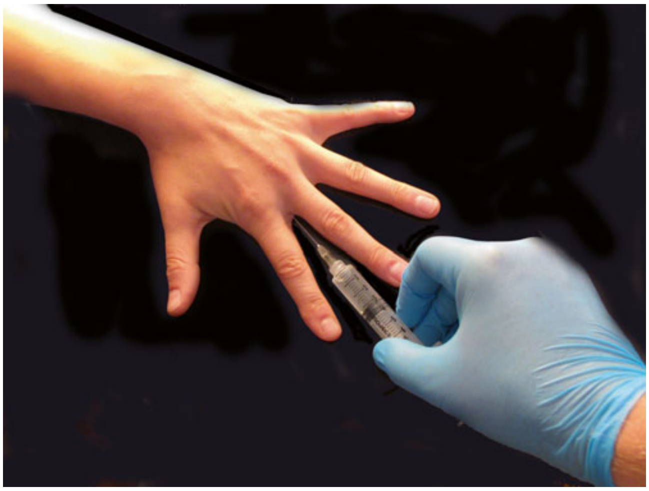

Fig. 89.17 Web space approach to the digital block, requiring injection on the medial and lateral web spaces for a blocked digit

89.7.4.3 Intrathecal Digital Block: Flexor Tendon Sheath (Fig. 89.18) [8]

- Procedure
  1. Inject anesthetic directly into the flexor tendon sheath. Palpate on the palmar surface over and proximal to the metacarpophalangeal joint. Gentle flexion of digit may better reveal the sheath. Have the patient abduct the fingers.
  2. Insert a 25-gauge needle at a 45° angle to the skin and along the long axis of the digit directly into the flexor tendon sheath at the level of the distal skin crease.
  3. Inject 2 mL of local anesthetic. The anesthetic should flow freely if it is in the sheath. If it does not, it is likely in the tendon and should be withdrawn slightly.
  4. Contraindications to intrathecal block are local infection and preexisting flexor tendon injury.
  5. Risk of tenosynovitis; sterilize the skin before introducing the needle.
  6. If laceration has involved the tendon, anesthetic may leak from the wound.

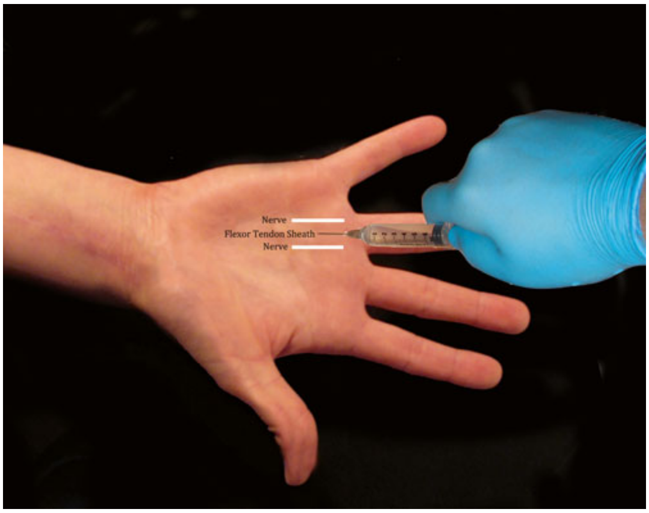

Fig. 89.18 Intrathecal/tendon sheath approach to the digital block

Acknowledgment The authors would like to thank Katy Howard for serving as the subject in many of the photographs in this chapter.

References

1. Muck AE, Bebarta VS, Borys DJ, Morgan DL. Six years of epinephrine digital injections: absence of significant local or systemic effects. Ann Emerg Med. 2010;56:270–4.
2. Waterbrook AL, Germann AC, Southall JC. Is epinephrine harmful when used with anesthetics for digital blocks? Ann Emerg Med. 2007;50:472–5.
3. Reichman EF, Simon RR, editors. Emergency medicine procedures. New York: McGraw-Hill Medical; 2004. p. 961–3.
4. Benko K. Fixing faces painlessly: facial anesthesia in emergency medicine. Available at: Emergency Medicine Practice (ebmedicine.net). 2009;11.
5. Rosh AJ. Ear anesthesia. Medscape reference. Available at: http://emedicine.medscape.com/article/82698-overview#a15.
6. Butterworth JF. Atlas of procedures in anesthesia and critical care. Philadelphia: WB Saunders; 1992. p. 160–4.
7. Brown DL. Atlas of regional anesthesia. Philadelphia: WB Saunders; 1992. p. 52.
8. Morrison WG. Transthecal digital block. Arch Emerg Med. 1993;10:35–8.
9. Mueller J, Davenport M. Digital nerve block (web space and tendon sheath). New York: McGraw-Hill's Access Emergency Medicine. Available at: http://www.accessemergencymedicine.com/videosPDF/DigitalNerveBlock.pdf.

Wound Management and Hemostasis

90

Rich Teitell and Muhammad Waseem

90.1 Indications for Primary Wound Repair

- Wounds that have been sustained less than 8 h before presentation may be closed primarily.
- If there is suspicion of broken glass or the presence of a foreign body, the patient should receive an x-ray before primary wound repair. Of note, x-ray has poor sensitivity for radiolucent foreign bodies such as wood and plastic.
- Wounds that are free of foreign bodies or show no gross evidence of infection or dead tissue may be closed primarily.

90.2 Contraindications for Primary Wound Repair

- History must address the factors that predispose a wound to greater risks of infection and not implement primary wound closure in such wounds that have had a duration greater than 8–12 h or contamination with saliva, stool, or foreign matter and a wound that has been sustained by blunt or crush mechanism.
- When wounds are highly contaminated with debris or if devitalized tissue is present, the wound should be left open for 3–4 days and then reevaluated.
- After a proper physical examination has been performed that assesses distal pulse and sensory and motor function of the surrounding region and distal extremity, if any of these are compromised, appropriate orthopedic and/or vascular consultation should be sought before wound closure, especially in the setting of a suspicion for open fracture.
- Consider delayed closure in wounds over joints.
- Wounds that have a significant loss of skin should be left to heal by secondary intention.

90.3 Procedure for Wound Closure and Hemostasis

1. Wound should be thoroughly prepared with Betadine (10 % povidone with 1 % iodine), which has a rapid onset with greatest antimicrobial effect.
2. Skin surrounding the wound should be covered with sterile drapes (Fig. 90.1) so as to reduce additional contamination within the open wound.
3. The most important determinant for preventing wound infection is adequate high-pressure irrigation.
4. Normal saline is the best solution for irrigation. Use pressures above 7 psi and an 18-gauge needle with a syringe of 30 mL or more; this gauge needle will achieve pressure greater than 7 psi.
5. It has been suggested that tap water irrigation may be just as effective. Because this approach will reduce costs, it can be considered.
6. Sharp debridement should be performed when foreign matter is found in an open wound that does not easily rinse out with saline. In this circumstance, consider wound closure by secondary intention or delayed primary closure.
7. At this point, wounds that do not meet criteria for primary closure must be left open, covered with a sterile dressing, and reevaluated in 3–4 days for consideration of delayed primary closure versus healing by secondary intention.
8. Wounds that do meet criteria as previously discussed for primary closure should be closed with either sutures or staples.
9. Staples are the preferred method of closure of scalp wounds and can be used under low wound tension.

R. Teitell, MD
Department of Emergency Medicine, Waterbury Hospital,
Waterbury, CT, USA
e-mail: rjt0013@gmail.com

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_90

10. The angle of the stapler is critical and must be positioned perpendicular to the intended area of delivery. A two-person approach can be utilized for optimal placement of the staple, using forceps to approximate the edges of the tissue.
11. Synthetic or monofilament sutures have a lower rate of infection than braided sutures (for closure techniques with sutures, see Chap. 92).
12. Dermabond should be utilized only for the most superficial linear lacerations and confers the benefit of efficiency of time and less discomfort to the patient. However, it should be mentioned that Dermabond does not offer any statistically significant difference in cosmesis when compared with suturing or other adhesives (see Chap. 93).
13. When a wound presents with deep underlying tissue involvement and direct bleeding is evident, direct pressure may be used to achieve hemostasis within a wound. Surgicel may also be applied with pressure to achieve control of small capillary and venous oozing (Fig. 90.2).
14. When a persistent bleed is visible within a wound that does not respond to direct pressure, a figure-of-eight suture (Fig. 90.3) may be used.
15. If a bleeding vessel is visible within a wound, an alternate method of hemostasis may be achieved by utilizing a hemostat to clamp the vessel and then tying an absorbable suture (Fig. 90.4).

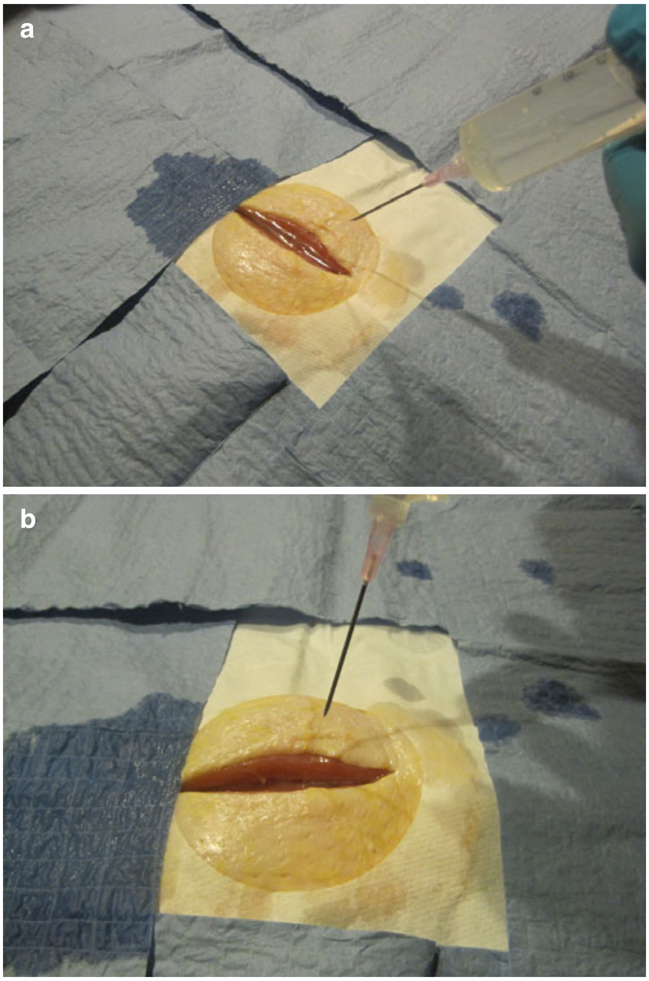

Fig. 90.1 (**a**) High-pressure irrigation using an 18-gauge needle (**b**)

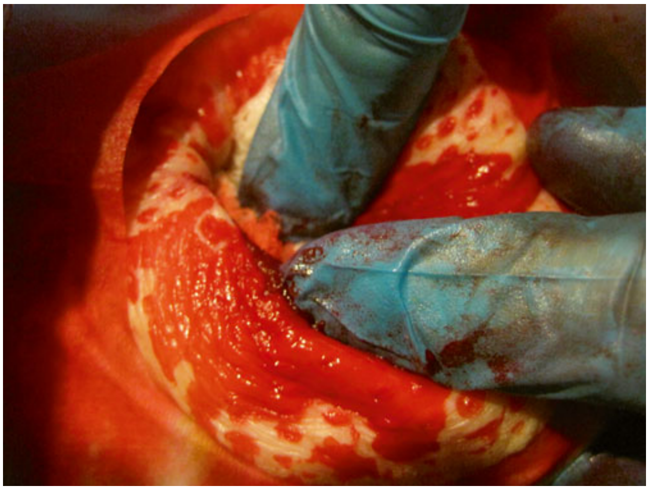

Fig. 90.2 Direct pressure with or without Surgicel may be used to achieve hemostasis within a wound and also to achieve control of small capillary and venous oozing

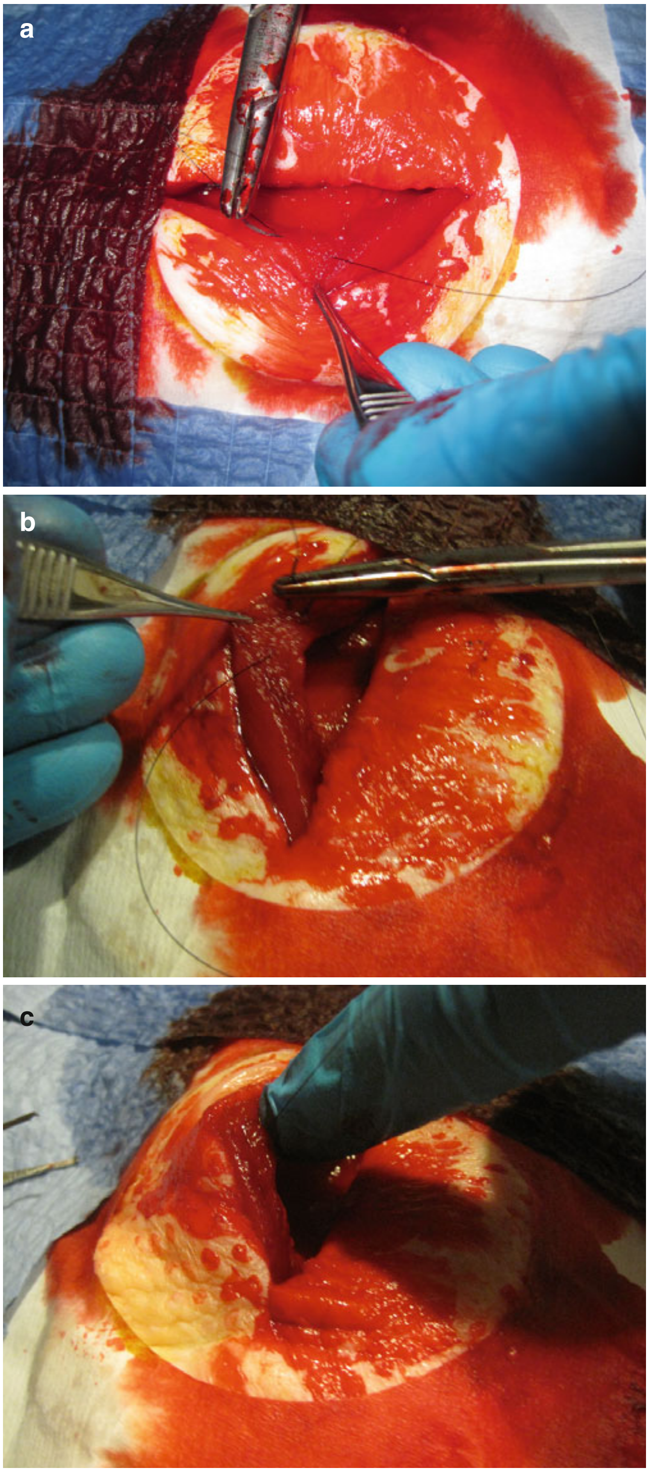

Fig. 90.3 (**a**) The location of the bleeding vessel in the subcutaneous tissue is identified and an absorbable suture is threaded through the region. (**b**) In a diagonal direction, a second suture is placed, which when tied will form the shape of an eight. (**c**) Once tied, the suture forms a figure eight in the tissue and hemostasis is achieved

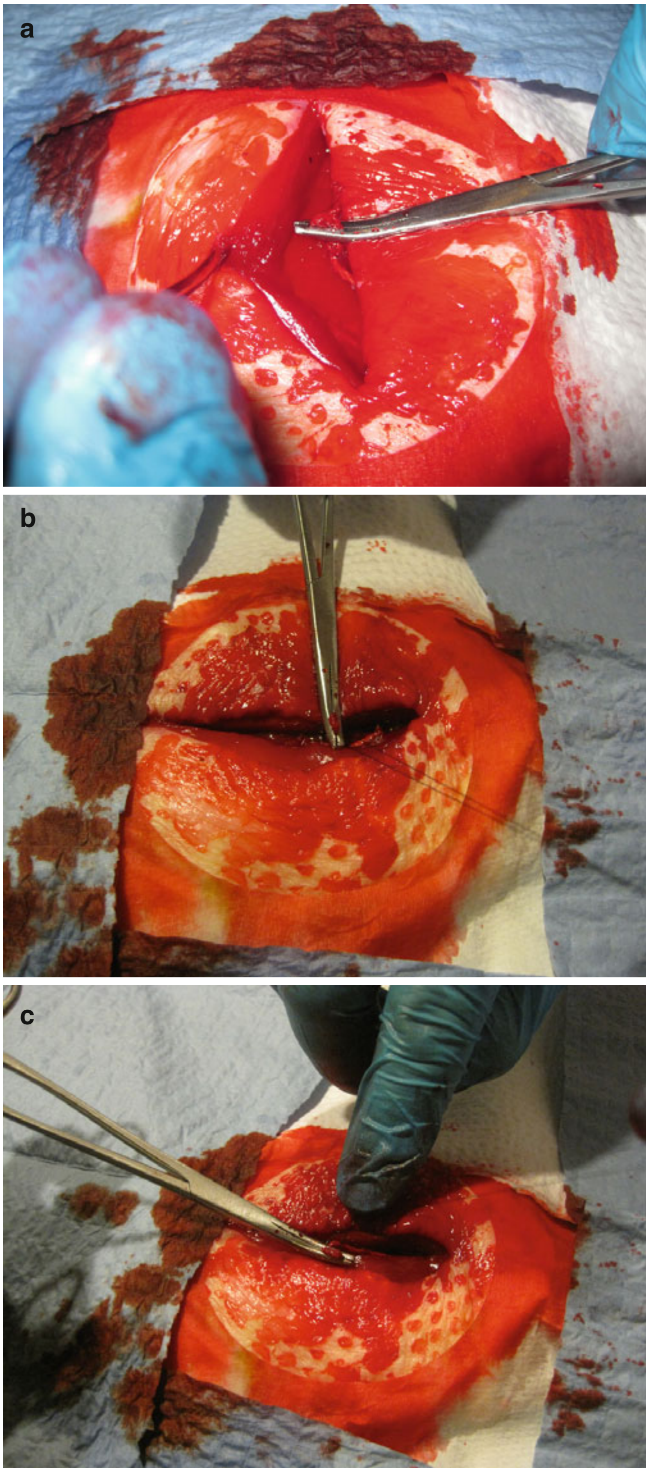

Fig. 90.4 (**a**) The culprit vessel is identified and clamped with a hemostat. (**b**) An absorbable suture is wrapped around the vessel inferior to the hemostat. (**c**) Once the first tie is placed being driven down with index finger, the hemostat may be removed and the knot can be completed for a total of three or four passes

90.4 Complications

- Hematoma formation
- Infection
- Scar formation

90.5 Pearls and Pitfalls

- Pearls
  - Antibiotic coverage should be employed judiciously depending on the wound and location.
  - Wounds to the face that are not heavily contaminated generally do not require antibiotics because the face and scalp have a significant vascular supply.
  - Patients with diabetes have higher incidences of wound infection, and the physician should consider prophylactic antibiotics in such individuals.
  - Wounds that are clean-contaminated and sutured in the emergency department should have a nonadherent dressing applied for 48 h until epithelialization has occurred. After that, the dressing may be removed, and soap and water gentle washing should commence to keep area clean. Only small amounts of bacitracin should be applied to wound in the interim period before sutures are removed.
- Pitfalls
  - Wounds located on the extremities have greater incidences of infection than those on the trunk or head/face.

Selected Reading

Blankenship RB, Baker T. Imaging modalities in wounds and superficial skin infections. Emerg Med Clin North Am. 2007;25:223–34.

Dimick AR. Delayed wound closure: indications and techniques. Ann Emerg Med. 1988;17:1303–4.

Edlich RF, Rodeheaver GT, Morgan RF, Berman DE, Thacker JG. Principles of emergency wound management. Ann Emerg Med. 1988;17(12):1284–302.

Farion K, Osmond MH, Hartling L, et al. Tissue adhesives for traumatic lacerations in children and adults. Cochrane Database Syst Rev. 2002;(3):CD003326.

Hall S. A review of the effect of tap water versus normal saline on infection rates in acute traumatic wounds. J Wound Care. 2007;16:38–41.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 6th ed. Philadelphia: Mosby-Elsevier; 2006. p. 702–3.

Ritchie AJ, Rocke LG. Staples versus sutures in the closure of scalp wounds: a prospective, double-blind, randomized trial. Injury. 1989;20:217–8.

Singer AJ, Hollander JE, Quinn JV. Evaluation and management of traumatic lacerations. N Engl J Med. 1997;337:1142–8.

Burn Care

91

Thomas Parry and Jeffrey Pepin

Burns result from exposure to heat, caustic chemicals, electricity, or radiation. Damage to the natural barrier provided by skin results in rapid fluid losses and risk of infection. Permanent scarring is a common long-term complication. They may also be complicated by sensory deficits due to loss of nerve connections and limb loss due to circulatory compromise.

91.1 Burn Description (Fig. 91.1)

- First degree or superficial: burn that remains confined to the epidermis (e.g., sunburn)
- Second degree or partial thickness: burn that extends into the dermis (e.g., blistering scald burns)
- Third degree or full thickness: burn involving the entire depth of the dermis and epidermal appendages

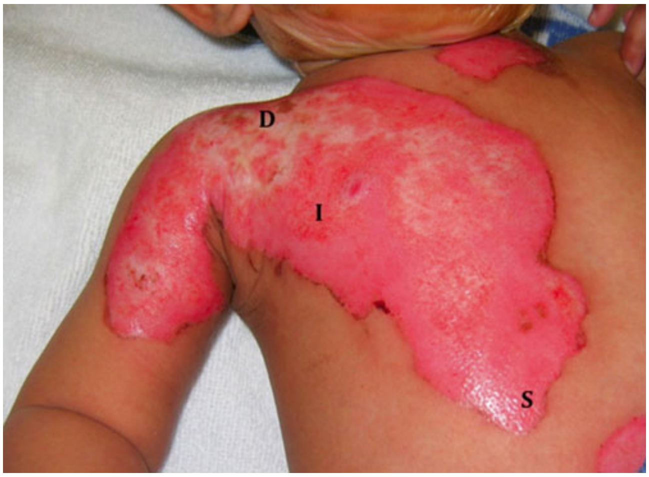

Fig. 91.1 Tea scald over the chest and shoulder of a child showing heterogeneity of burn depth. *D* deep (second or third degree), *I* intermediate (second degree), *S* superficial (first degree) (Reproduced from: Enoch S, Roshan A, Shah M. Emergency and early management of burns and scalds. *BMJ*. 2009;338:b1037, with permission from BMJ Publishing Group Ltd.)

T. Parry, MD (✉)
Department of Emergency Medicine, Lincoln Hospital and Mental Health Center, New York, NY, USA
e-mail: Thomas.parry@nychhc.org

J. Pepin, MD
Department of Emergency Medicine, University of Minnesota Medical Center Fairview, Minneapolis, MN, USA
e-mail: medicone76@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_91

91.2 Indications

- Superficial, deep second degree, and third degree.
- First-degree burns generally need only supportive care.

91.3 Materials and Medications

- Appropriate analgesia
- Antibiotic ointment
- Silver sulfadiazine (avoid on face)
- Sulfamylon for ears
- Alternative agents include bacitracin or polymyxin B ointments
- Cleansing solution such as chlorhexidine prep and water
- Basin
- Petroleum gauze
- Several 4 × 4 gauze pads
- Rolled gauze
- Tape
- Gloves

Airway, breathing, and circulation and cervical spine should be immediately assessed before any burn wound management. It is important to remember that any patient suspected of an inhalation injury or carbon monoxide poisoning should receive 100 % humidified oxygen until carboxyhemoglobin returns to normal. Inhalation injuries may not present themselves until after fluid resuscitation has already been started, so early intubation is recommended in these cases.

91.4 Sizing

Total body surface area (TBSA) of body parts is estimated by multiples of 9 (Rule of Nines)

- Adults
  - Head and neck: 9
  - Arms: 9 each
  - Legs: 18 each
  - Trunk: 18 front and 18 back
  - Perineum and palms: 1
- Infants/children
  - Head and neck: 18
  - Arms: 9 each
  - Legs: 14 each
  - Trunk: 18 front and 18 back

A second way to estimate TBSA in smaller burns is by using the palm surface area: using the patient’s palm as a guide, each palmar surface equals 1 %.

A third way to estimate TBSA is via the use of the Lund and Browder chart (Fig. 91.2).

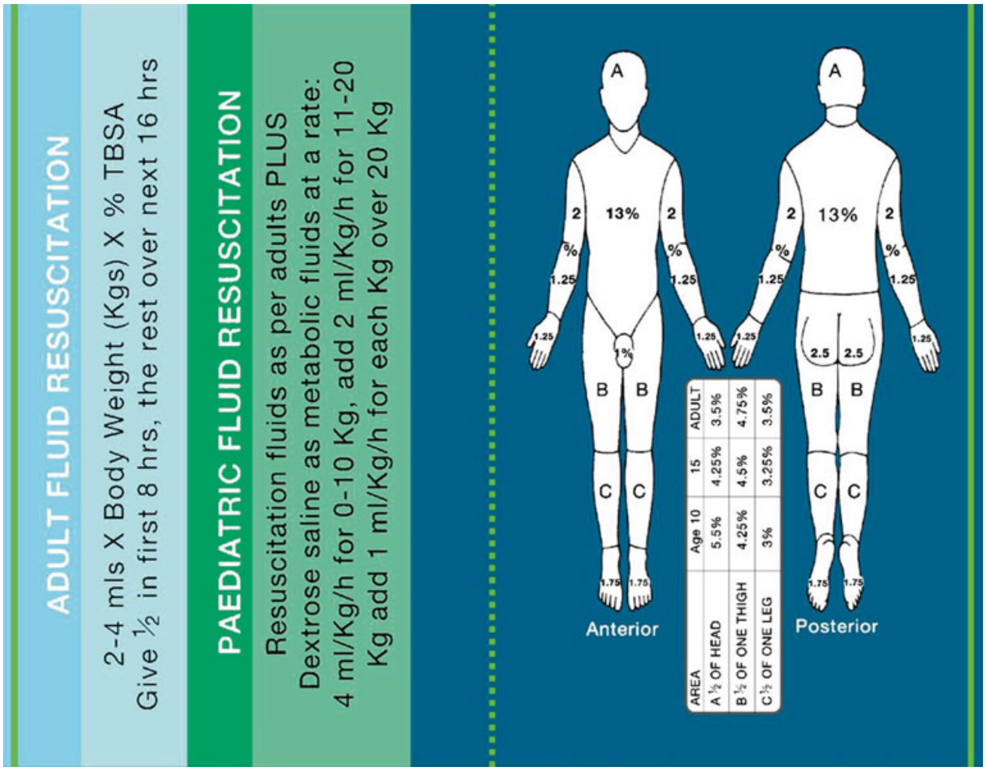

ADULT FLUID RESUSCITATION

2-4 mls X Body Weight (Kgs) X % TBSA
Give ½ in first 8 hrs, the rest over next 16 hrs

PAEDIATRIC FLUID RESUSCITATION

Resuscitation fluids as per adults PLUS
Dextrose saline as metabolic fluids at a rate:
4 ml/Kg/h for 0-10 Kg, add 2 ml/Kg/h for 11-20
Kg add 1 ml/Kg/h for each Kg over 20 Kg

AREA	Age 10	15	ADULT
A ½ OF HEAD	5.5%	4.25%	3.5%
B ½ OF ONE THIGH	4.25%	4.5%	4.75%
C ½ OF ONE LEG	3%	3.25%	3.5%

Fig. 91.2 Lund Browder chart for estimating total body surface area of burns, with suggested fluid resuscitation guidelines (Reprinted from Malic CC, Karoo RO, Austin O, Phipps A. Resuscitation burn card–a useful tool for burn injury assessment. *Burns*. 2007;33(2)195–9, with permission from Elsevier)

91.5 Fluid Resuscitation

Because fluid resuscitation is absolutely essential in the early aspects of burn care, the Parkland formula has been used to estimate fluid requirements in burn patients. The patient's weight in kilograms is multiplied by the percent body surface area (BSA) involved; this number is multiplied by 4 mL of lactated Ringer's solution. Half of this amount is given during the first 8 h and the remaining amount is given over the next 16 h of resuscitation. The goal is to keep urine output approximately 0.5–1.0 mL/kg/h.

In pediatric cases, the Galveston formula may be used instead: 5000 mL/m^2 BSA burned plus 2000 mL/m^2. TBSA of 5 % dextrose in lactated Ringer's solution intravenously over the first 24 h, half in the first 8 h, and the other half over the next 16 h.

91.6 Procedure

1. Provide appropriate analgesia.
2. Clean the wound with antiseptic solution and water (if it is a dry chemical burn, make sure to brush off as much chemical before using copious amounts of water to clean the wound).
3. Debride any loose skin or foreign debris using a dry 4×4 or rolled gauze.
4. Apply antibiotic ointment or cream to the burn.
5. Apply petroleum gauze in a single layer just over the affected skin.
6. Use loose 4×4 gauzes and "fluff" them to make a thick layer of padding to place over the petroleum gauze.
7. Either wrap the entire area with a rolled gauze or tape a small layer of 4×4 gauze over the "fluffed" layer of 4×4 gauze.

91.7 Complications

- Wound infection
- Nonhealing wound requiring skin graft (deep second-degree and third-degree burns)
- Compartment syndrome (circumferential burns may require escharotomy)
- Rhabdomyolysis

91.8 Pearls and Pitfalls

- Pearls
  - Determination of the depth of burns on initial presentation is difficult (especially when covered with petroleum). A good rule of thumb is if it blanches and/or hurts, it is a partial thickness burn.
  - First-degree burns are not included in burn size estimations for fluid resuscitation calculation.
  - Burn size determines fluid requirements and transfer decision.
  - Burn of greater than 20 % TBSA should receive intravenous fluids.
  - Patients who have inhalation injury, greater than 10 % TBSA, high-voltage electrical injury, or chemical burns should be referred to burn unit.
  - Fluid resuscitation should be adjusted according to physiological response such as urine output (30–50 mL/h in adults and 1 mL/kg/h in children).
  - Assure tetanus is up to date.
- Pitfalls
  - Keep the patient warm in the first few hours. There is no need to apply ice.
  - All jewelry and rings should be removed.
  - Prophylactic antibiotics are not recommended.
  - Silver sulfadiazine should be avoided in facial burns because of the risk of staining of the skin.

Blister care is a very controversial topic. Current research suggests that it may be beneficial to keep the blister intact unless it appears to be tense or over a joint. Most blisters will rupture in 2–4 days. Ruptured blisters should be debrided with all the extra skin removed. Most burn units will scrub everything off once they receive a patient.

Wounds should be kept clean to prevent an environment that will increase the chances of infection. Wrap in saline-soaked sterile gauze prior to transfer.

Dressing changes should be done daily with all previously applied antibiotic ointment removed before a reapplication of new ointment. It is important to provide analgesic 30 min before a dressing change.

91.9 Admission Criteria

- Partial-thickness burns of noncritical areas not including the eyes, ears, face, hands, feet, or perineum that total a BSA of 10–20 % in adults
- Partial-thickness burns of noncritical areas involving 5–10 % of BSA in children younger than 10 years
- Suspicious of non-accidental trauma
- Patients unable to care for wounds in outpatient settings
- Prompt referral to a burn specialist is required in the following cases:
  - Partial-thickness and full-thickness burns greater than 10 % of the TBSA in patients younger than 10 years or older than 50 years of age.
  - Partial-thickness and full-thickness burns greater than 20 % of the TBSA in other age groups.

- Partial-thickness and full-thickness burns involving the face, eyes, ears, hands, feet, genitalia, or perineum or the skin overlying major joints.
- Full-thickness burns greater than 5 % TBSA in any age group.
- Electrical burns, including lightning injury: significant volumes of tissue beneath the surface may be injured and result in acute renal failure and other complications.
- Significant chemical burns.
- Inhalation injury.
- Burn injury in patients with preexisting illness that could complicate management, prolong recovery, or affect mortality.
- Any burn patient in whom concomitant trauma poses an increased risk of morbidity or mortality may be treated initially in a trauma center until stable before transfer to a burn center.
- Children with burns seen in hospitals without qualified personnel or equipment for their care should be transferred to a burn center with these capabilities.
- Burn injury in patients who will require special social and emotional or long-term rehabilitative support, including cases involving suspected child abuse and neglect.

Selected Reading

Bezuhly M, Fish JS. Acute burn care. Plast Reconstr Surg. 2012;130(2):349e–58.

Rex S. Burn injuries. Curr Opin Crit Care. 2012;18(6):671–6.

Wasiak J, Cleland H, Campbell F, Spinks A. Dressings for superficial and partial thickness burns. Cochrane Database Syst Rev. 2013;(3):CD002106.

Wound Closure

92

Oliver Michael Berrett, Jeffrey Joseph Harroch,
Karlene Hosford, and Muhammad Waseem

92.1 Indications

- Open wound of skin or mucosal tissues
- Purpose
  - Preserve function
  - Control bleeding
  - Promote healing
  - Cosmesis

92.2 Contraindications

- Wounds caused by animal or human bites
- Contaminated, infected, or puncture wounds
- Complex wounds (may require operating room)

92.3 Methods

- Suture placement
- Tissue adhesives
- Adhesive tapes
- Staples

92.4 Preparation

- Obtain a complete history of injury
  - Mechanism
  - Time since injury
  - Tetanus status
  - Comorbidities
- Examine the extent of the wound, remove any contaminants, debride devitalized tissue, inspect for foreign bodies. Obtain a radiograph if foreign body is suspected.
- Copious irrigation with normal saline.
- Clean around the wound with a povidone-iodine solution.
- Drape the area in sterile manner.
- Inject the anesthetic agent into the wound edges using a 25- or 27-gauge needle. The most common agent is lidocaine 2 % with or without epinephrine 1 % (maximum dose is 3 mg/kg without epinephrine and 5 mg/kg with epinephrine).
- Close the wound using the appropriate technique (see later).
- Dress the wound.
- Update tetanus and diphtheria vaccination if needed.

O.M. Berrett, MD • K. Hosford, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: omberrett@gmail.com; krlhos@aol.com;
waseemm2001@hotmail.com

J.J. Harroch, MD
Department of Emergency Medicine, University of Miami Miller
School of Medicine, University of Miami Hospital,
Miami, FL, USA
e-mail: jharroch@med.miami.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_92

92.5 Suture Repair

92.5.1 Materials and Medications

- Commercial kits commonly contain all the following except the suture material (Fig. 92.1)
  - Povidone-iodine solution
  - Normal saline for irrigation
  - 5- to 12-mL syringe with a 25-gauge needle
  - Anesthetic agent
  - Needle holder
  - Pickups
  - Suture scissors
  - Suturing material
  - Sterile drape or sheet, gloves, gauze

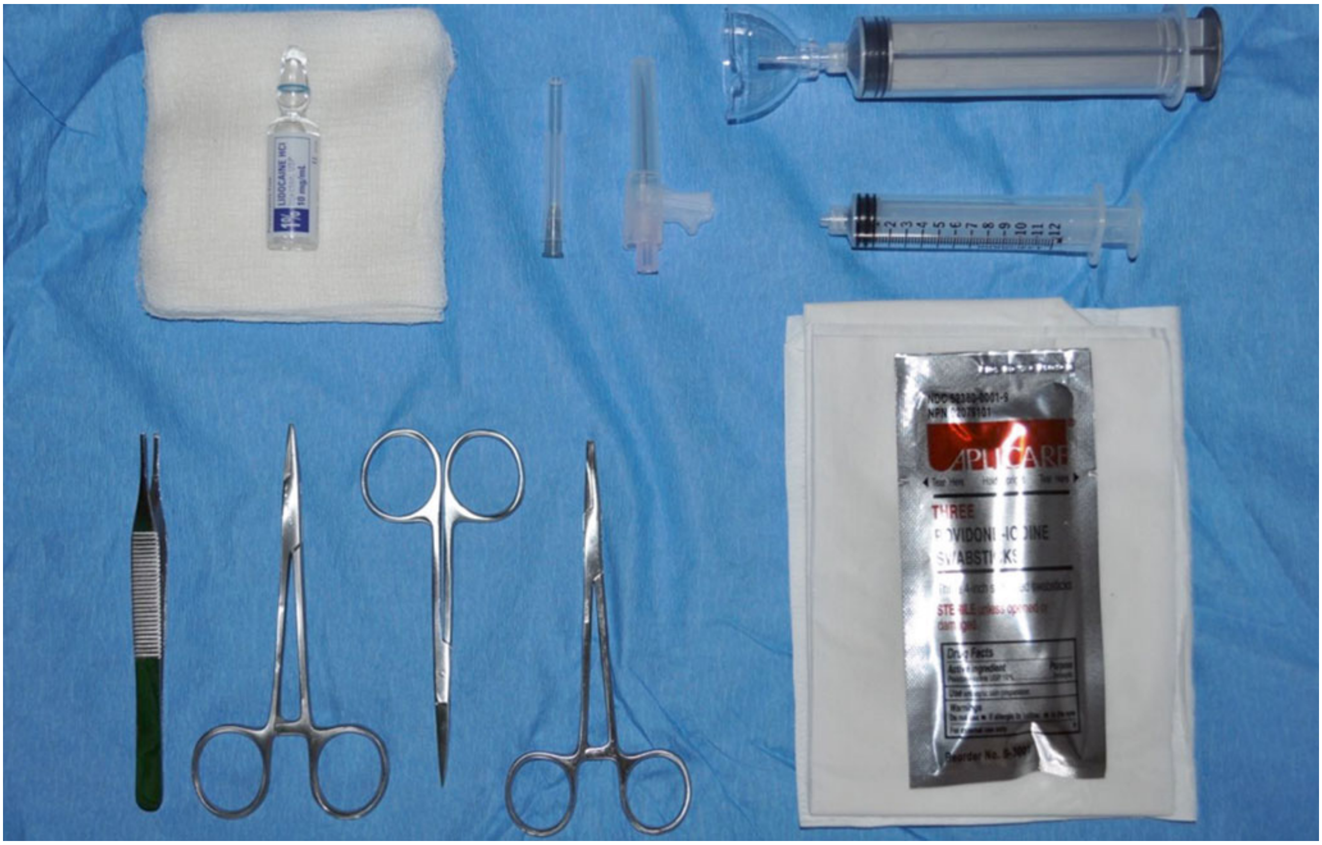
LIDOCAINE HCl
10 mg/mL
APLICARE
THREE
POVIDONE-IODINE
SWABSTICKS
STERILE
Drug Facts

Fig. 92.1 Suture kit

92.5.2 General Guidelines

- Local anesthetic lidocaine 1 % or lidocaine 1 % with epinephrine.
- Minimize direct use of instruments on the tissues.
- Wound edges should be everted to maximize healing and cosmetic effect. This is achieved by inserting the needle at 90° to the skin.
- Sutures should be evenly spaced, placed 1–3 mm apart and 2 mm from the wound edge.
- Optimal tension is achieved by tying the sutures so the edges lightly approximate.

92.5.3 Suture Material

- Nonabsorbable
  - Silk: for specialty use, reactive and weak
  - Nylon (Ethilon), polypropylene (Prolene): good strength, good overall material for cutaneous wounds
  - Polypropylene: good strength, difficult to use
  - Require removal at a specified time
- Absorbable
  - Undergo rapid degradation in tissues, losing their tensile strength within 60 days
  - Indication: buried suture to reduce wound edge tension
  - Vicryl: subcutaneous placement, mucous membranes
  - Chromic: use for intraoral lacerations
  - Removal not required

92.5.4 Suture Size

- 0.0–2.0: thick material for large wounds, trunk
- 0.3–0.4: used on medium-sized wounds, extremities, scalp
- 0.5–0.6: fine sutures, used on facial wounds

92.5.5 Suture Techniques

- Simple interrupted sutures (Fig. 92.2)
  - Most common method
  - Position needle 2 mm from the wound edge at a 90° angle. Enter the needle into the skin and arc through the wound edge and into the opposing edge at the same level, exiting the skin on the opposing side 2 mm from wound edge, tie.
- Deep dermal suture (Fig. 92.3)
  - Also known as buried sutures, used to minimize tension in a wound.
  - Use absorbable suture material.
  - The needle is placed at the base of the wound wall and arched upward, exiting the ipsilateral wall more superficially. The needle is then directed across to the opposing wound wall at the same level and directed downward, exiting deep, tie. Note: The knot will be deep in the wound.
- Simple running suture (Fig. 92.4)
  - This method provides rapid closure of long and relatively linear lacerations.
  - Place the initial suture in the same manner as a simple suture, tie, cut the free strand, and leave the needle attached. Reintroduce the needle into the skin on the opposite side so the suture crosses the wound superficially at a 65° angle. The needle is then inserted perpendicular to the skin, emerging on the opposite side about 3 mm from the wound edge. Repeat without tying until closure is complete, maintaining appropriate tension. When the last stitch is placed, leave a loose loop of suture on one side so that both ends can be tied together.
- Vertical mattress suture (Fig. 92.5)
  - Provides the benefits of both simple and deep techniques. For use on deeper, gaping wounds and wounds over high-tension areas such as joints.
  - Position the needle 1 cm from skin edge, at a 90° angle, drive a deep arc perpendicular through the wound, exiting the skin on the opposing side the same distance from the wound edge. Next, reinsert the needle on the ipsilateral side 2 mm from the edge, emerging on the opposing side and approximate, tie.
- Horizontal mattress suture (Fig. 92.6)
  - For large wounds with tension.
  - Place the initial suture in the same manner as a simple suture, only do not tie. Reposition the needle on the ipsilateral side, horizontally 5 mm to the side of the exit at the same distance from the wound edge and drive through to the other side, tie. Note: The tie lies parallel to the wound.
- Half-buried horizontal mattress suture (Fig. 92.7)
  - Used on wounds with skin flap.
  - On one side, drive the needle percutaneously, and then pass horizontally through the dermal tissue of the tip of flap, finally passing into the dermis of the opposing edge and exiting the skin, tie.

92.5.6 Recommendations

- Face
  - 5.0–6.0 nylon.
  - Remove after 3–5 days.
- Scalp
  - 2.0–3.0 nylon or staple.
  - Remove at 8–10 days.

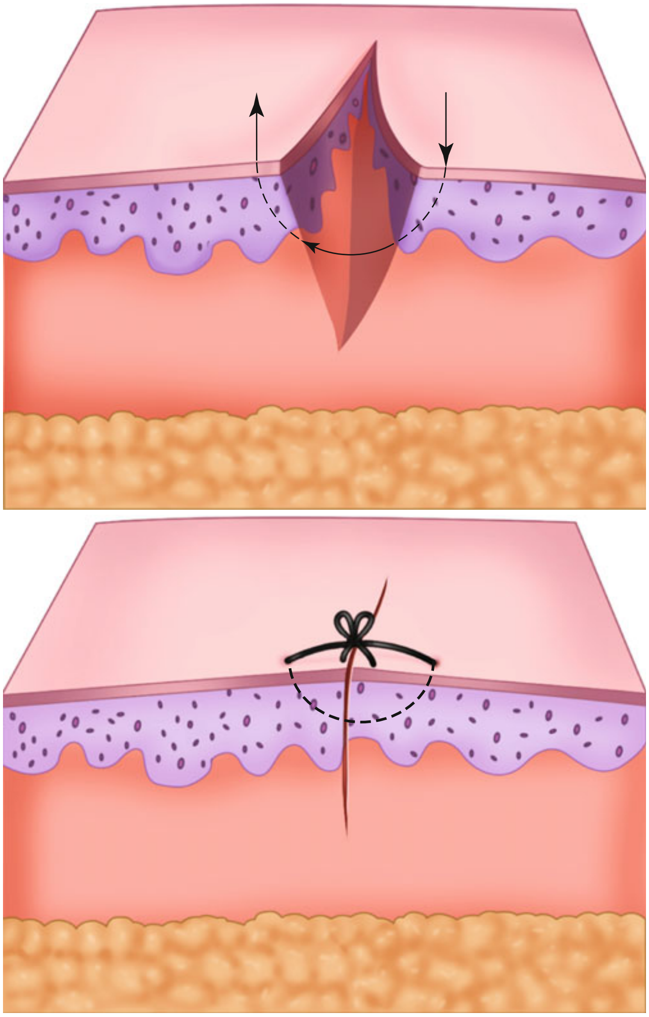

Fig. 92.2 Simple suture

- Hand
  - 4.0–5.0 nylon, consider vertical or horizontal mattress.
  - 5.0–6.0 Monocryl for nail bed.
  - Remove at 10–14 days, use for a longer time if directly over the joint.
- Extremity
  - Nonmobile skin: 3.0–4.0 nylon, remove at 8–10 days
  - Over joint: 3.0–4.0 nylon, remove at 10–14 days
- Trunk
  - Anterior trunk: 3.0–4.0 nylon, remove at 8–10 days.
  - Posterior trunk: 2.0–3.0 nylon, remove at 10–14 days.
  - Consider staples.
- Oral mucosa
  - Thin mucosa: 4.0 Vicryl will absorb; duration, 5–7 days
  - Tongue: 3.0 Vicryl will absorb; duration, 5–7 days

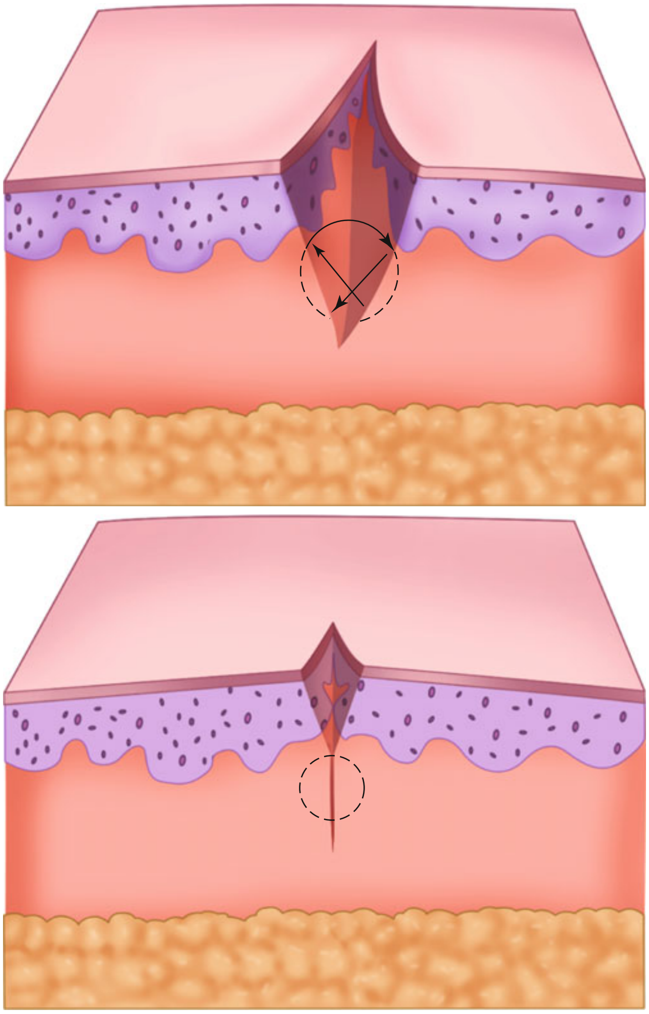

Fig. 92.3 Deep suture

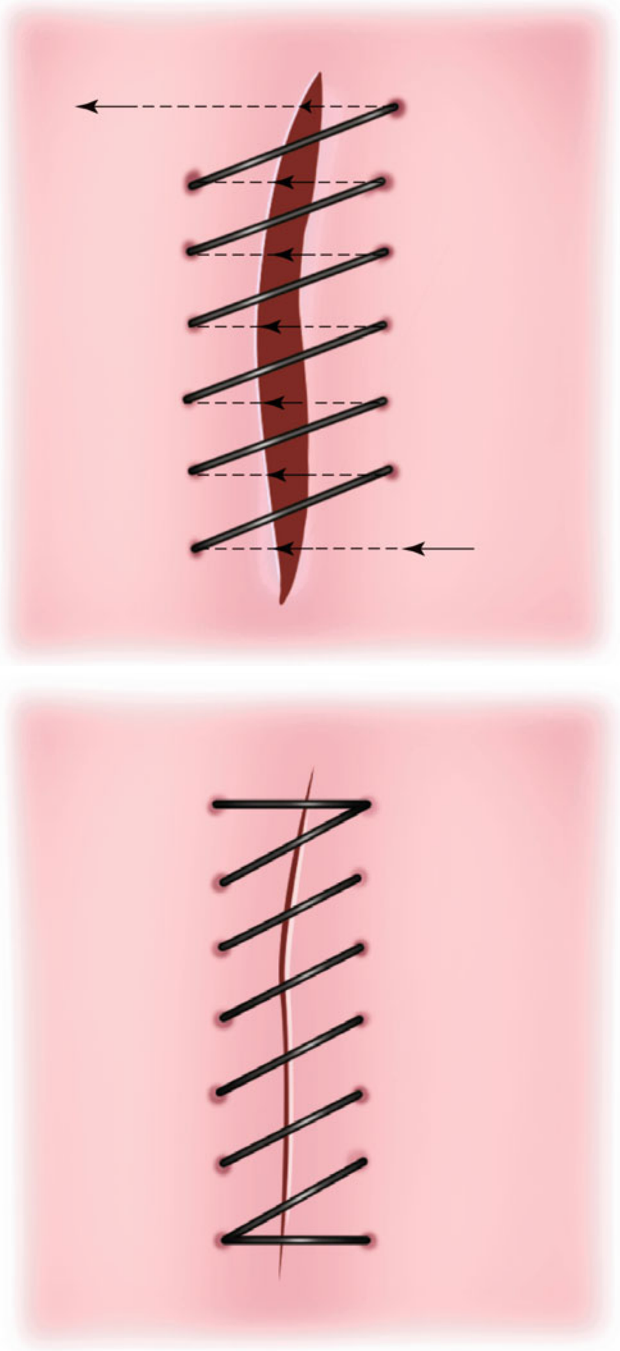

Fig. 92.4 Running suture

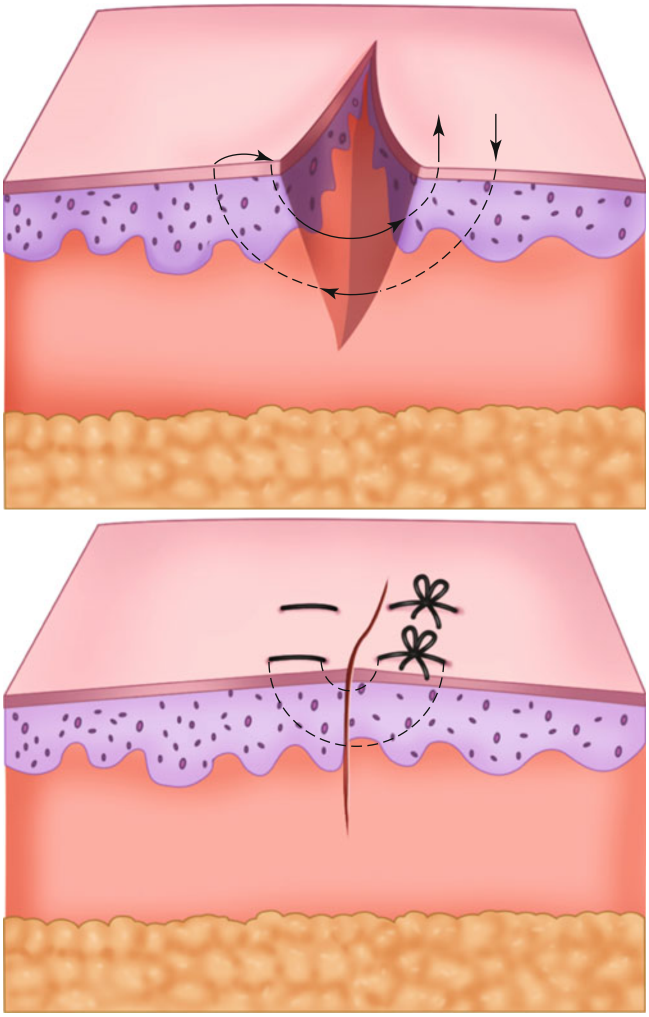

Fig. 92.5 Vertical mattress suture

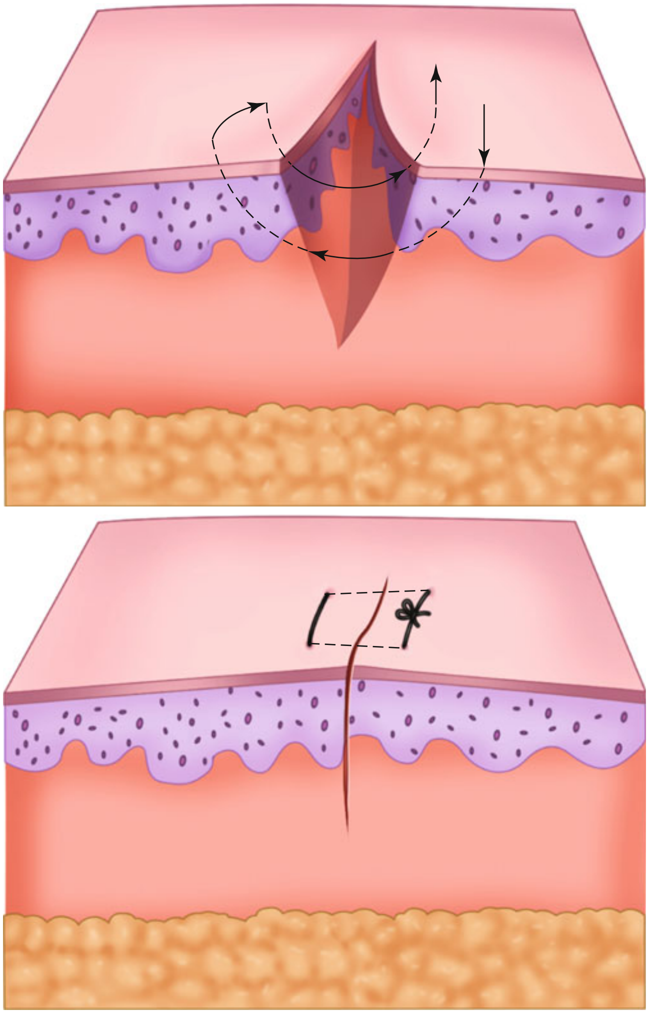

Fig. 92.6 Horizontal mattress suture

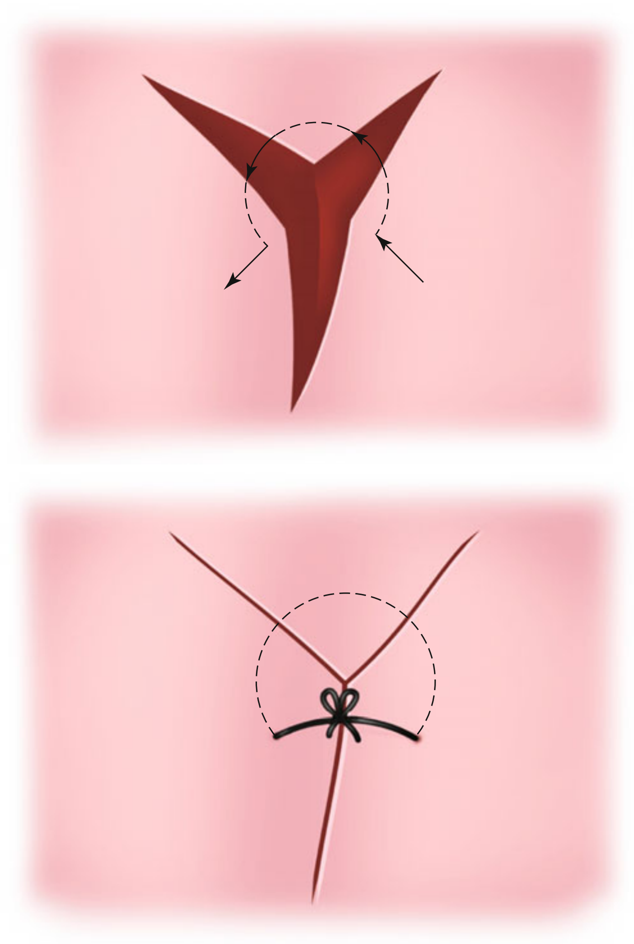

Fig. 92.7 Half-buried horizontal mattress suture

92.6 Alternative Methods of Wound Closure

92.6.1 Cyanoacrylate Tissue Adhesives

See Chap. 93.

92.6.2 Adhesive Tape

- Advantages
  - Rapid and painless application
  - Inexpensive
  - Good cosmetic result
- Disadvantages
  - Minimal strength
- Contraindications
  - Allergy to product
- Precautions
  - For use with wounds under little tension
- Procedure
  - Thoroughly clean the wound as described previously.
  - Approximate the wound edges.
  - Apply adhesive tape directly over the wound with 2–3 mm of space between strips.
  - An adjunct adhesive, such as benzoin, may be used to improve durability.

92.6.3 Staple Closure (Fig. 92.8)

- Advantages
  - Rapid
  - Inexpensive
- Disadvantages
  - Minimal strength
- Contraindications
  - Wounds on face, hands, or feet
- Procedure
  - Thoroughly clean the wound as described previously.
  - Anesthetize the wound edges.
  - Approximate the wound edges (may require an additional set of hands).
  - Apply staples.

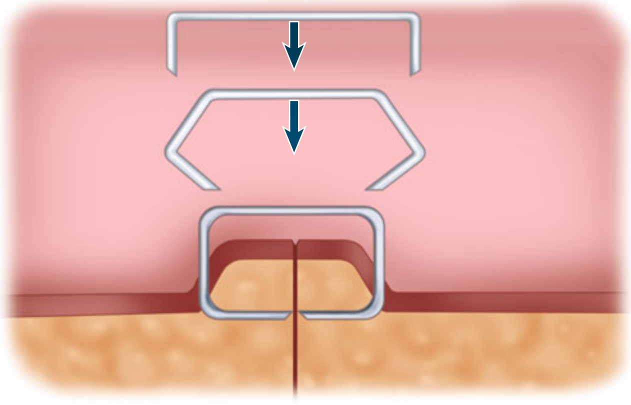

Fig. 92.8 Staple

92.7 Pearls and Pitfalls

- Pearls
  - Sutures should not remain in place longer than the recommended period. If the wound is not completely healed, sutures should be removed and an adhesive tape may be used for edge approximation.
  - Remember to verify tetanus status.
  - Antibiotics should be used judiciously.
- Pitfalls
  - Wound infection.
  - Incomplete healing may lead to wound separation.
  - Tissue reaction to suture or adhesive materials.
  - Allergy to anesthetic agent.

Selected Reading

http://apps.med.buffalo.edu/procedures/repairoflacerations.asp?p=17. Accessed 19 May 2014.

Singer AJ, Hollander JE. Methods for wound closure. In: Ma OJ, Cline DH, Tintinalli JE, Kelen GD, Stapczynski JS, editors. Emergency medicine manual. 6th ed. New York: McGraw Hill; 2003: Chap. 13, Fig. 13–14.

Singer AJ, Hollander JE, editors. Laceration and acute wounds: an evidence-based guide. Philadelphia: FA Davis; 2003. p. 122.

University of Connecticut Health Center, suturing 101. fitsweb.uchc.edu/suturing101. Accessed 19 May 2014.

Zuber TJ. The mattress sutures: vertical, horizontal and corner stitch. Am Fam Physician. 2002;66:2231–6.

Wound Closure with Tissue Adhesive

93

Pratik S. Patel and Latha Ganti

93.1 Indications

- Small superficial skin incisions or laceration repairs which require 5.0 or smaller-diameter sutures.

93.2 Contraindications

- Absolute
    - Large irregular/stellate lacerations
    - Infected/contaminated wounds
    - Animal/human bites
    - Puncture wounds
    - Crush wounds
    - Skin ulcers
    - Mucous membranes and mucocutaneous junctions
    - Axillae and perineum (owing to high moisture)
- Relative
    - Wounds on extremities (unless kept dry)
    - Joints (unless kept immobilized with a splint)

93.3 Materials and Medications

- Dermabond (2-octyl cyanoacrylate)
- Betadine (povidone-iodine) solution
- 0.9 % normal saline solution
- 20-mL sterile syringe
- Sterile gloves
- Dry 4 × 4 gauze

93.4 Optional Materials

- Topical anesthesia
- 1:1000 epinephrine solution
- Forceps
- Bacitracin ointment or sterile petroleum jelly ointment
- Gown and protective eyeglasses
- Splint

P.S. Patel, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_93

93.5 Procedure

1. Have the patient rest comfortably on a chair or bed.
2. Use universal precaution measures: sterile gloves (gown and eye-screen, if necessary for wound irrigation).
3. Wash the wound with 0.9 % normal saline irrigation.
4. Use a topical anesthetic such as LET (lidocaine, epinephrine, tetracaine) or EMLA (eutectic mixture of local anesthetics) cream (lidocaine and prilocaine), or a 1:1000 epinephrine solution soaked into gauze can be used to achieve hemostasis in a bleeding wound.
5. Approximate the edges of the wound with fingers. Toothed forceps or other skin approximation devices may be used as an adjunct. Apply bacitracin ointment or Vaseline to the tips of the forceps and wipe off the excess to prevent sticking of Dermabond glue to the forceps.
6. Crush Dermabond vial between the thumb and the finger while in the inverted position until the adhesive is seen at the applicator tip (Fig. 93.1).
7. Squeeze gently until a drop of adhesive forms at the applicator tip.
8. Gently brush the adhesive at the applicator tip over the approximated wound edges (Fig. 93.2). (Do not force or press the applicator tip over the wound.)
9. Cover the entire wound with single coat of adhesive.
10. Hold the wound edges for 30 s to 1 min until it dries.
11. Apply two or three more coats of adhesive over and around the wound in a circular or oval movement to provide extra strength.
12. Wipe off extra adhesive in the surrounding skin with gauze if needed.
13. Apply a splint (optional) to provide wound stability over the joints.
14. Recommend the patient to keep the wound dry for 4–5 days. Patients may shower but should be instructed to pat dry instead of rubbing a towel over the skin.
15. No topical antibiotic ointment is required before or after application of Dermabond.

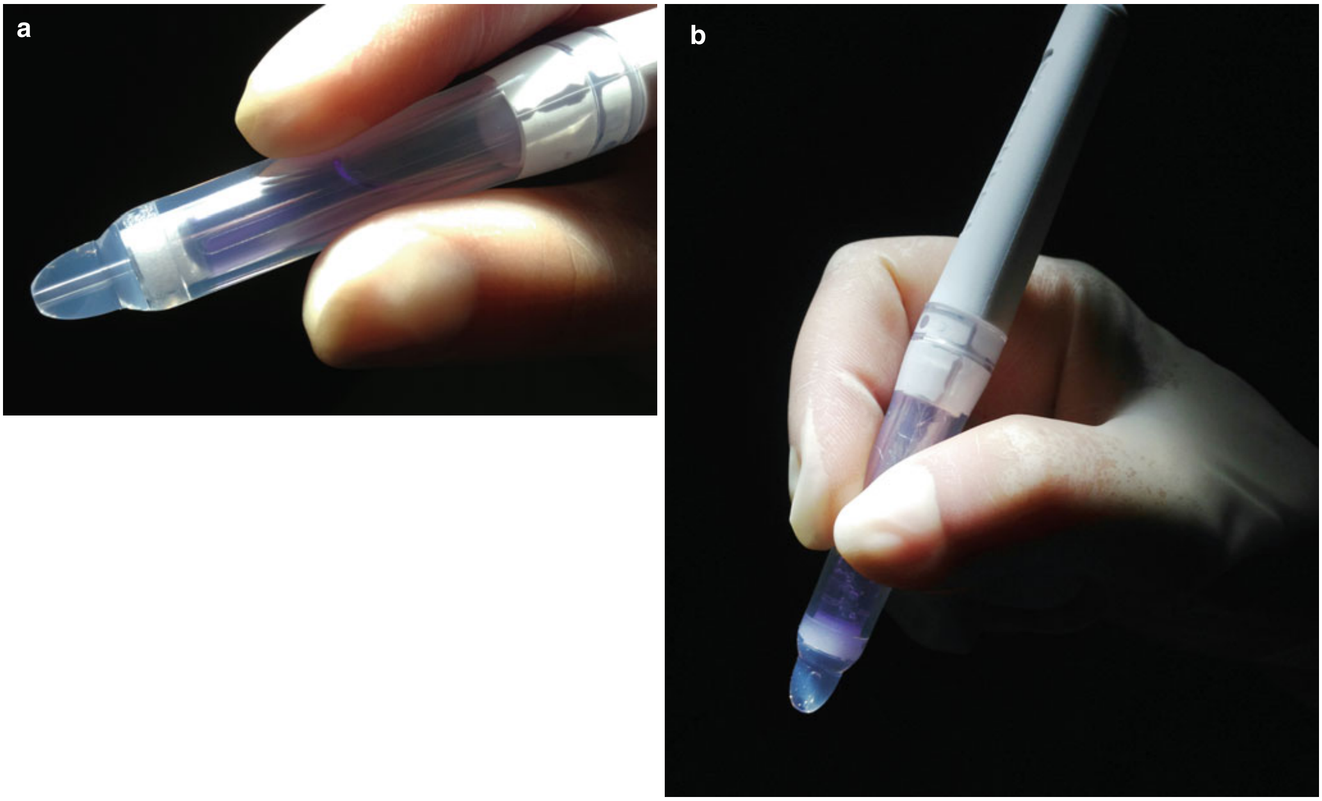

Fig. 93.1 (**a**) Before the tip has been crushed. (**b**) The purple Dermabond is in the tip

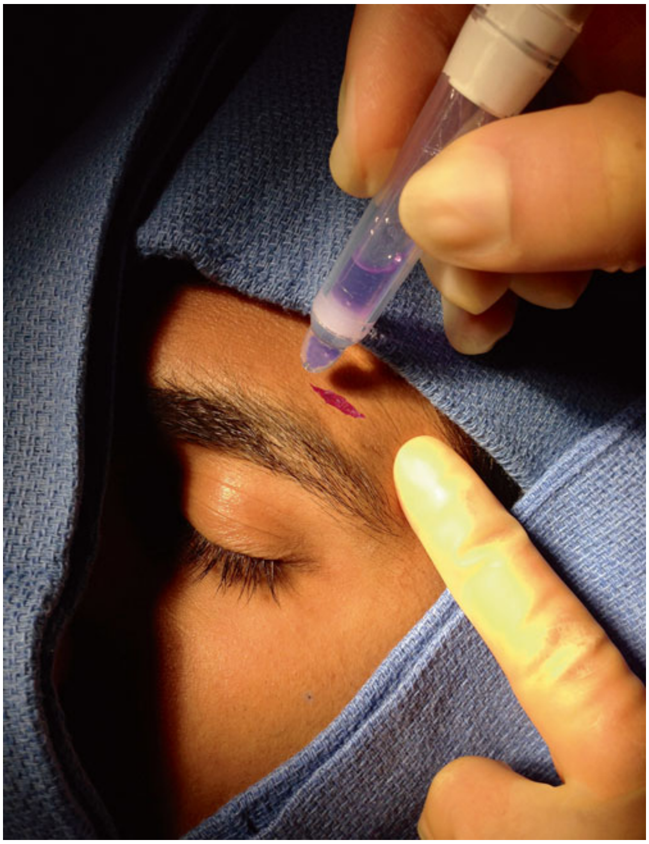

Fig. 93.2 Closing the wound: use tip of Dermabond pen to "paint" over the laceration

93.6 Complications

- Wound dehiscence
- Wound infection

93.7 Pearls and Pitfalls

- Pearls
  - Advantages of tissue adhesives for wound repair compared with sutures include faster repair time, better acceptance by patients (especially children), water-resistant covering, and no need for a second visit to remove sutures (sloughs off in 5–10 days).
- Pitfalls
  - Dermabond is a super adhesive. Take care not to have the glove, finger, drape, gauze, or instrument inadvertently stuck to the wound or the patient by having a bacitracin or petroleum jelly coating around the wound, on gloved fingers, and on forceps as needed.

Selected Reading

Bruns TB, Worthington JM. Using tissue adhesive for wound repair: a practical guide to Dermabond. Am Fam Physician. 2000;61:1383–8.

Farion K, Osmond MH, Hartling L, et al. Tissue adhesives for traumatic lacerations in children and adults. Cochrane Database Syst Rev. 2002;(3):CD003326.

Fishhook Removal

94

Judith K. Lucas

94.1 Indications

- Removal of a fishhook from nonvital structures (Fig. 94.1)

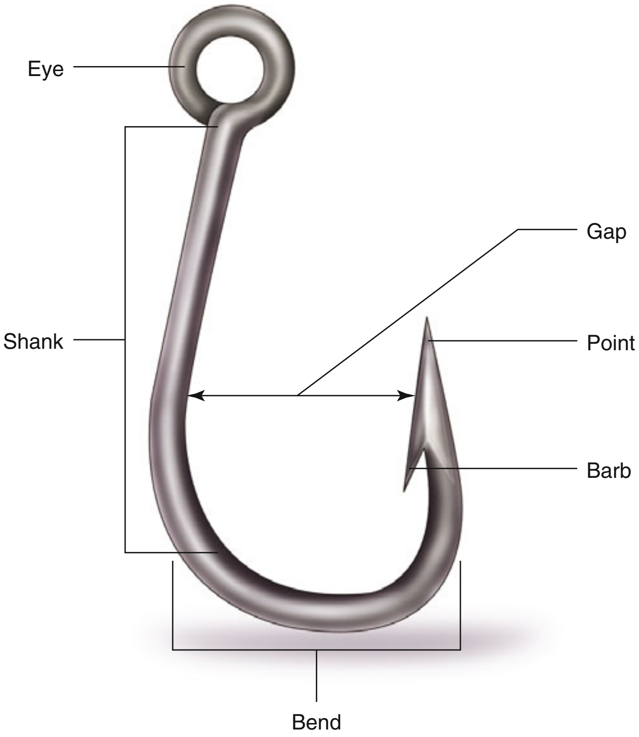

Fig. 94.1 Anatomy of a fishhook

94.2 Contraindications

- Removal of hooks located near/in eyes or eyelids, embedded near or within neurovascular structures, or embedded within vital structures such as peritoneum, testicle, or urethra.
  - Fishhooks in these areas require specialist consultation.

94.3 Materials and Medications

- Antiseptic cleansing solutions
  - Betadine (povidone-iodine)
  - Chloraprep
- Local anesthetic
  - 1 % lidocaine, with or without epinephrine
- Needle drivers or pliers
- 18- or 20-gauge needle
- 3-0 silk suture or umbilical tape
- Wire cutters
- Protective eyewear

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_94

94.4 Procedures

- Retrograde technique (Fig. 94.2): simplest, least traumatic, but least successful; good for small to medium hooks, superficially embedded hooks, and hooks with no barbs or a single barb.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the puncture site and surrounding tissue with antiseptic solution.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. Apply downward pressure on the shank or shaft of the hook, near the eye, thus disengaging the barb from the tissue.
  5. Back the hook out of the skin along the path of entry.
- String and yank technique (Fig. 94.3): Good for small to medium hooks. Often requires no anesthesia. Good for areas of deep soft tissue penetration. Cannot be used on parts of the body that are not fixed, such as earlobes.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Consider infiltrating the entry site with local anesthetic.
  4. Wrap 3-0 silk suture, fishing line, or umbilical tape several times around the bend of the hook (at the point of greatest curvature).
  5. The loose ends of the string need be held tightly. Sometimes the loose ends can be more firmly held if wrapped around a pencil or tongue depressor.
  6. The skin around the entry site should be well stabilized, while simultaneously depressing the shank, close to the eye of the hook.
  7. While stabilizing the skin and applying downward pressure on the shaft, quickly and firmly yank the string, in a parallel line to the shaft. Be certain downward pressure is applied along the shank.
  8. The hook will “fly” out quickly. Be certain to wear eye protection.
- Needle cover technique (Fig. 94.4): good method to remove large hooks with a single barb, especially if superficially embedded
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. Advance an 18-gauge needle along the entrance wound of the hook. Pass the needle parallel to the shank. The bevel should be pointed downward, toward the barb and point.
  5. The needle is advanced until it disengages the barb, entrapping it within the needle lumen.
  6. The hook and needle are advanced just enough to disengage the barb.
  7. The hook and needle are withdrawn along the track of the wound.
- Advance and cut technique (Fig. 94.5). Almost always successful but does cause additional tissue trauma.
  1. Detach extra hooks, line, or foreign materials (e.g., worms, fish, debris).
  2. Cleanse the penetration site and surrounding area with antiseptic wash.
  3. Infiltrate the entry site and surrounding area with local anesthetic.
  4. If the hook has a single barb:
     - Grip the hook on the shank, near the bend, with either a needle driver or pliers.
     - Push the hook through along its natural trajectory, until the point and barb pass completely through the skin.
     - Clip the point proximal to the barb. Be certain to wear eye protection because the point can fly off in an unpredictable direction.
     - Withdraw the hook out back along its entry path.
  5. If the hook has multiple barbs (Fig. 94.6):
     - Grip the hook on the shank, near the hook’s bend with either a needle driver or pliers.
     - Push the hook through along its natural trajectory, until the point and barb pass completely through the skin.
     - Clip the eye of the hook with wire cutters.
     - Grasp the point and withdraw the hook through the skin forward along its natural course.

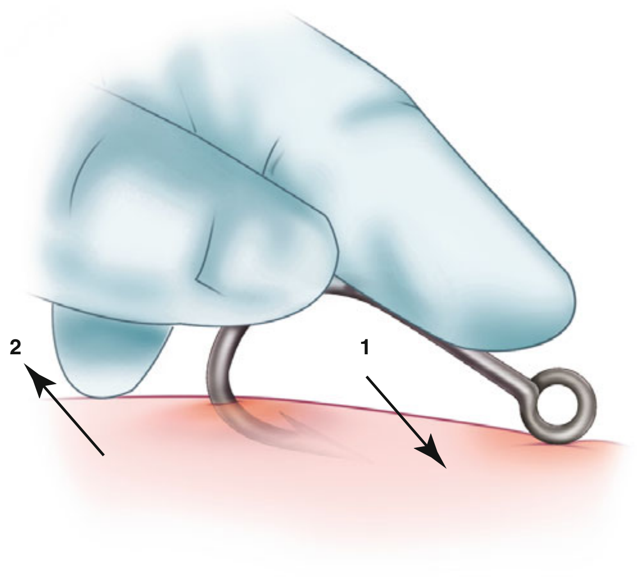

Fig. 94.2 Retrograde technique. *1* Push down along whole shank and slightly forward to release the barb. *2* Pull back and out along hook's entry path, applying steady downward pressure

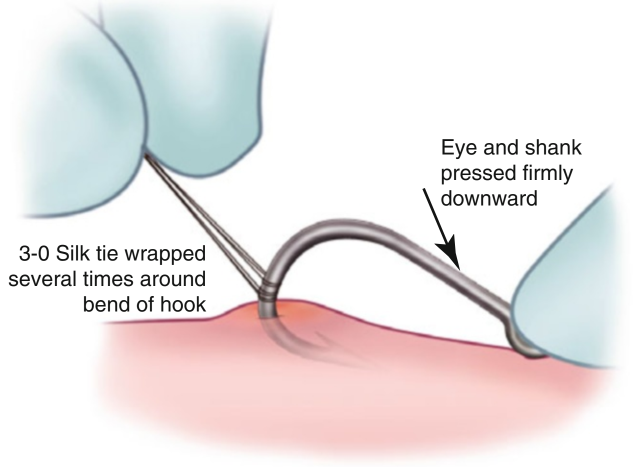
Eye and shank
pressed firmly
downward
3-0 Silk tie wrapped
several times around
bend of hook

Fig. 94.3 String and yank technique

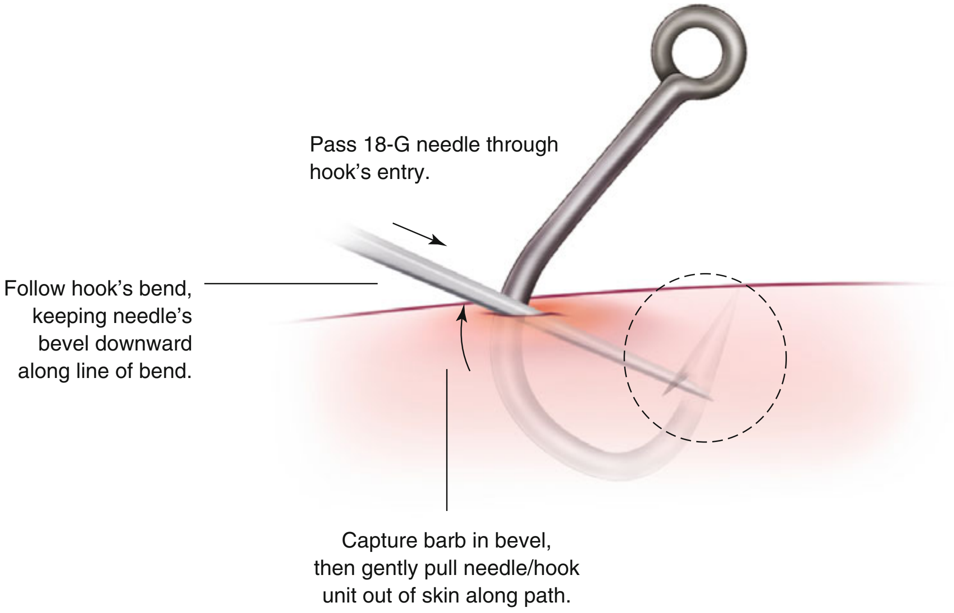
Pass 18-G needle through
hook's entry.
Follow hook's bend,
keeping needle's
bevel downward
along line of bend.
Capture barb in bevel,
then gently pull needle/hook
unit out of skin along path.

Fig. 94.4 Needle technique

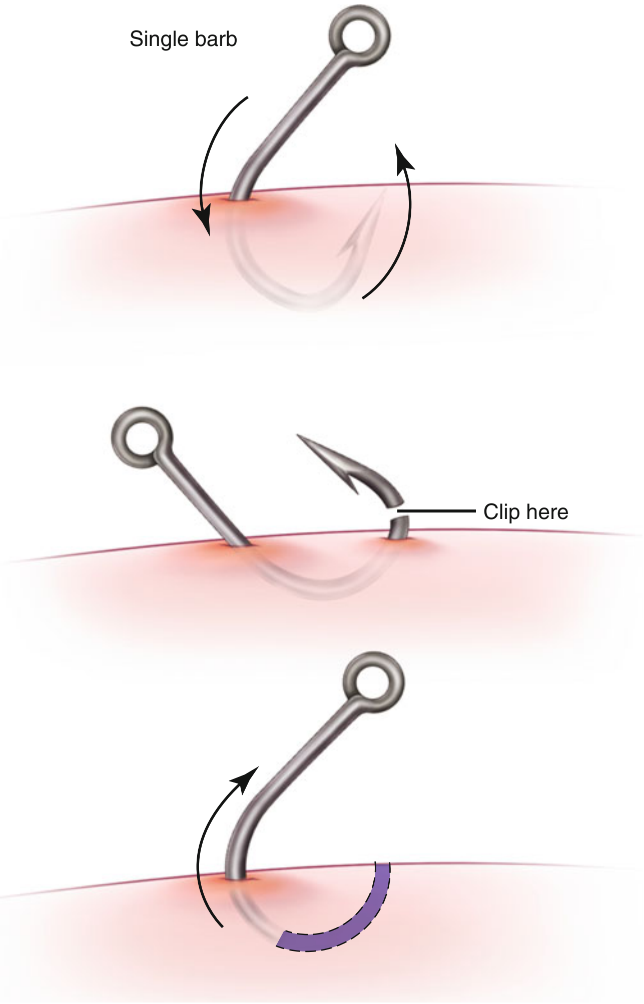
Single barb
Clip here

Fig. 94.5 Advance and cut technique

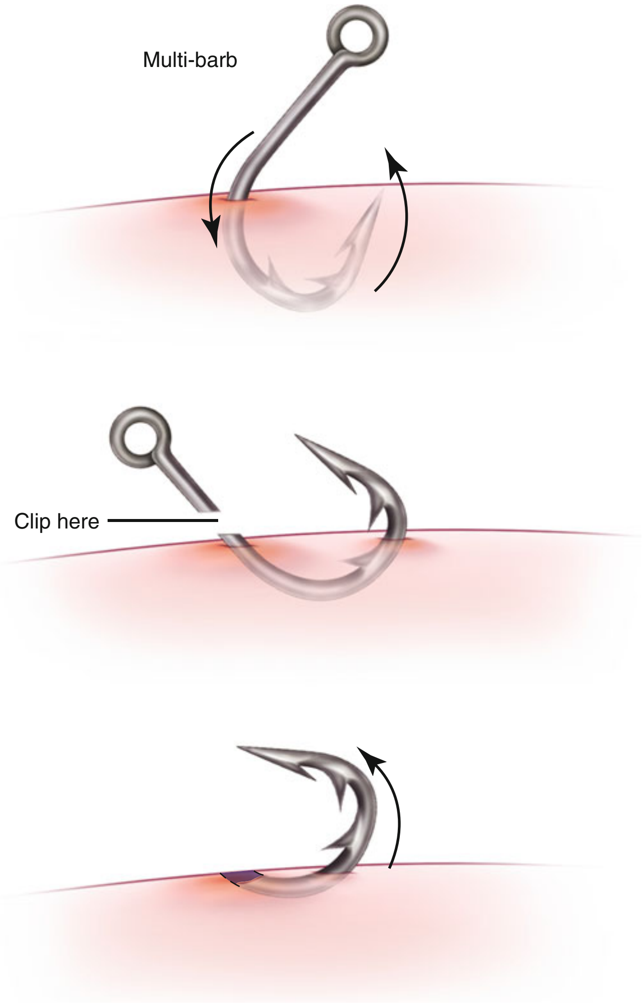
Multi-barb
Clip here

Fig. 94.6 Advance and cut a hook with multiple barbs present

94.5 Postremoval Wound Care

- Immune competent, without peripheral vascular disease:
    1. Explore the wound for possible foreign bodies.
    2. Irrigate or scrub the wound copiously with soapy solution.
    3. Apply antibiotic ointment and sterile dressing.
    4. Tdap (or Dtap) should be given to anyone in whom the last tetanus booster has been longer than 5 years.
    5. Wound check in 24–48 h with care provider.
- Immunocompromised or a patient with peripheral vascular disease:
    1. As previously.
    2. Give strong consideration to treating prophylactically with antibiotics, choosing a fluoroquinolone, third-generation cephalosporin, or aminoglycoside.

94.6 Complications

- Infection
- Retained foreign body
- Injury to neurovascular structures, if method of removal is not carefully selected
- Injury to provider if adequate (eye) protection not used

94.7 Pearls and Pitfalls

- Pearls
    - Begin with either the retrograde method or the string-yank method, because these result in less tissue damage and are the easiest to perform, although they have the lowest rate of success.
    - Eye protection is imperative, especially if utilizing the string-yank method or clipping any portion of the hook (eye or point) because the retraction of the hook, or its parts, is generally at a high velocity and travels an unpredictable path.
    - When trying the advance and cut technique, stop immediately if, when advancing the hook, impasse, or resistance is met because this may indicate bone or neurovascular structures are blocking the natural path of the hook.
    - Fishhooks that embed into or near the eye or lids should be covered with a metal patch or cup and the patient should be sent (immediately) for ophthalmologic consultation.
    - Close follow-up is imperative to watch for signs of infection.
- Pitfalls
    - When utilizing the advance and cut technique, do not cut anything until certain there is another portion of the hook on which to grasp.

Selected Reading

Bothner J. Fish-hook removal techniques. Available at: www.UpToDate.com. Literature review version 19.2, May 2011. Topic last updated: Sep 2010. Accessed 23 June 2014.

Gammons M, Jackson E. Fishhook removal. Am Fam Physician. 2001;63:2231–7.

Wakeman K. Fishhook removal. Available at: www.fishgame.com. Texas Fish and Game, LLC. July 2003. Accessed 23 June 2014.

Tick Removal

95

David N. Smith and Judith K. Lucas

95.1 Indications

- Tick attachment to the skin

95.2 Contraindications

- None

95.3 Materials and Medications

- Gloves
- Skin disinfectant (commercially available product, such as Chloraprep, isopropyl alcohol, or Betadine [povidone-iodine])
- Fine-toothed forceps

95.4 Procedure

1. Comfortably position the patient with the tick site exposed.
2. Grasp the tick as close to the skin surface as possible (e.g., grasp the mouth parts).
3. Gently pull upward with steady, nontwisting, even traction (Figs. 95.1 and 95.2).
4. After removal, clean the bite area and apply antibiotic ointment.

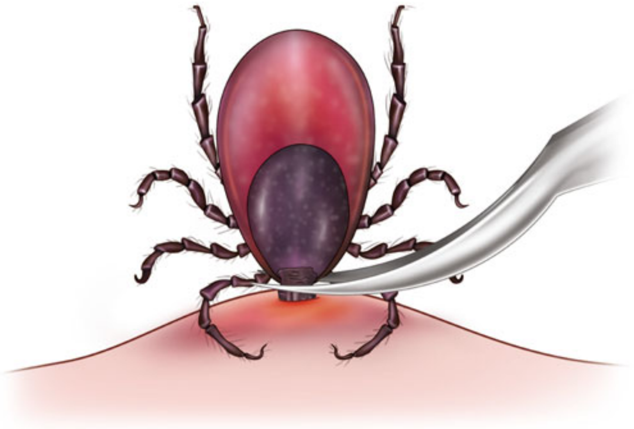

Fig. 95.1 Place the forceps as close to the mouth of the tick as possible, hold firmly, and pull straight up with steady gentle traction

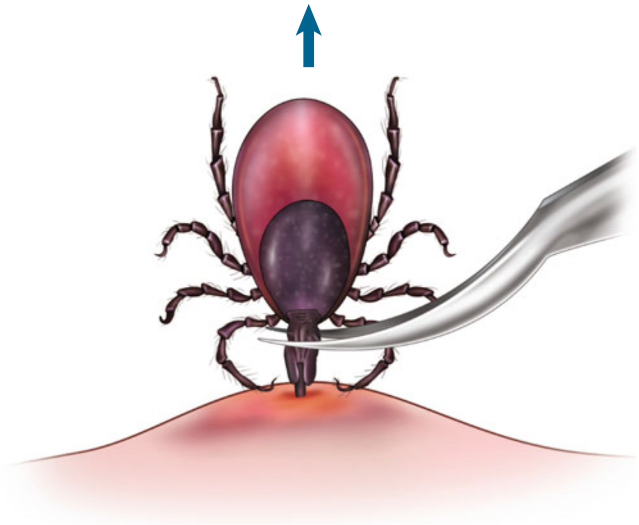

Fig. 95.2 Steady upward traction

D.N. Smith, MD
Department of Pediatrics, University of Alabama at Birmingham,
Children's of Alabama, Birmingham, AL, USA
e-mail: dsmith@peds.uab.edu

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_95

95.5 Complications

- Multiple diseases including:
  - Lyme disease
  - Human granulocytic and monocytic ehrlichiosis
  - Babesiosis
  - Relapsing fever
  - Rocky Mountain spotted fever
  - Colorado tick fever
  - Tularemia
  - Q fever
  - Tick paralysis
- Secondary infection (methicillin-resistant *Staphylococcus aureus* [MRSA] and group A streptococcus).
- Scratching can lead to lichenification.
- Rare cases of alopecia when tick located in the scalp.

95.6 Pearls and Pitfalls

- Pearls
  - Do not twist or jerk the tick out (may cause breakage of mouth parts; they may remain in the skin).
  - Do not squeeze the body of the tick.
  - Do not use a hot match, gasoline, or other noxious stimulus for removal (causes irritation of tick and release of internal contents).
  - Lyme disease transmission increases significantly after 24–48 h of attachment, so early removal is the key.
  - Patients should be monitored for up to 30 days for signs of tick-borne diseases, including erythema migrans (bull's-eye rash) indicating Lyme disease (Fig. 95.3).
- Prophylactic antibiotic treatment, a single dose of doxycycline, is used only in patients with an identified *Ixodes scapularis* tick that has been attached for longer than 36 h and if treatment can start within 72 h of tick removal.
  - Serological testing for Lyme disease is not indicated for a reported tick bite.
- Pitfalls
  - Use of lidocaine subcutaneously can irritate the tick and cause it to regurgitate its stomach contents, increasing the risk of disease transfer.

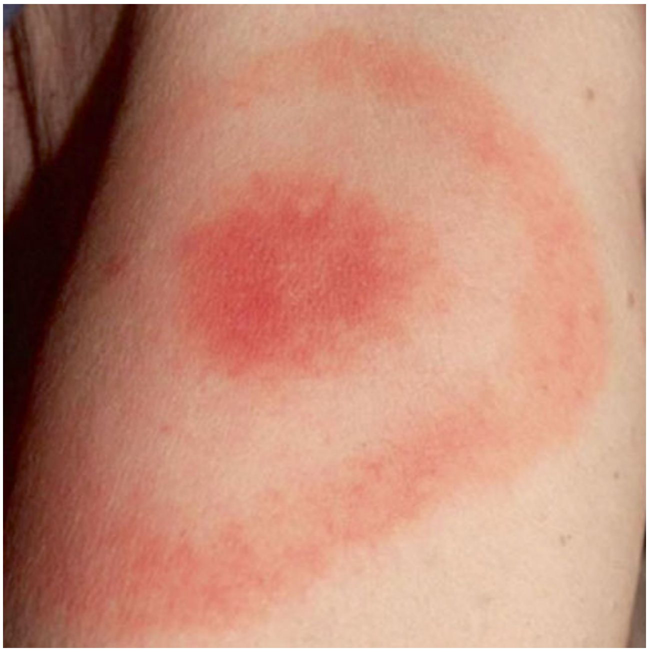

Fig. 95.3 Erythema migrans (the bull's-eye rash), the rash typically associated with Lyme disease. Often the rash is so pale as to go unnoticed

Selected Reading

Centers for Disease Control and Prevention. Tick removal. In: Ticks. 2014. http://www.cdc.gov/ticks/removing_a_tick.html. Accessed 23 Feb 2014.

Needham GR. Evaluation of five popular methods for tick removal. Pediatrics. 1985;75:997–1002.

Sexton DJ. Evaluation of a tick bite for possible Lyme disease. Up To Date; 2010. http://uptodate.com.

Sloan S. Background. In Tick removal. 2014. http://emedicine.medscape.com/article/1413603-overview. Accessed 23 Feb 2014.

96

Subungal Hematoma Drainage

Pratik S. Patel and Latha Ganti

96.1 Causes

- Crushing injury to the finger or toe
- Ill-fitted shoes/inadequate space for toes

96.2 Indications

- Pain that is not tolerable with nail edges intact

96.3 Contraindications

- Relative
  - Disrupted nail edges
  - Tolerable pain that can be managed conservatively
  - Skin infections around toe/finger
  - Bleeding disorder

96.4 Materials and Medications

- 18-gauge needle
- Betadine (povidone-iodine) solution
- Gloves
- Topical antibiotic

96.4.1 Optional materials

- Electrocautery tool
- Paper clip or sewing needle and sterilizing flame
- Finger splint
- Lidocaine (1 or 2 % with epinephrine)

P.S. Patel, MD
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: drpratik@gmail.com

L. Ganti, MD, MS, MBA (✉)
Professor of Emergency Medicine, University of Central Florida,
Orlando, FL, USA

Director, SE Specialty Care Centers of Innovation, Orlando
Veterans Affairs Medical Center, Orlando, FL, USA
e-mail: lathagantimd@gmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_96

96.5 Procedure

1. Have the patient rest the affected finger comfortably on a flat surface.
2. Use universal precaution measures: gloves, gown, and eye screen.
3. Sterile skin preparation with Betadine.
4. Optional (for complicated subungal hematoma): use 1 or 2 % lidocaine with epinephrine for providing local digital block.
5. Using the thumb and index finger gently twist the 18-gauge needle with light pressure over the base of the nailbed or in the center of the hematoma until no resistance is felt. Do not apply any further pressure in order to avoid nailbed damage. Nail penetration is confirmed by return of dark blood from the hole. This procedure is known as trephination (Figs. 96.1 and 96.2).
6. Apply light pressure around the tip of the finger and hematoma to facilitate drainage. In case of continuous bleeding elevate the digit and apply firm and continuous pressure over the nail with a gauze piece.
7. Apply topical antibiotic ointment (e.g., bacitracin) over the puncture hole.
8. Apply a gauze dressing or a bandage over the wound site/fingernail.
9. Apply a finger splint (optional) to provide additional comfort.
10. Recommend the patient to keep the finger or toe dry (avoid soaking) and elevated for 1–2 days.

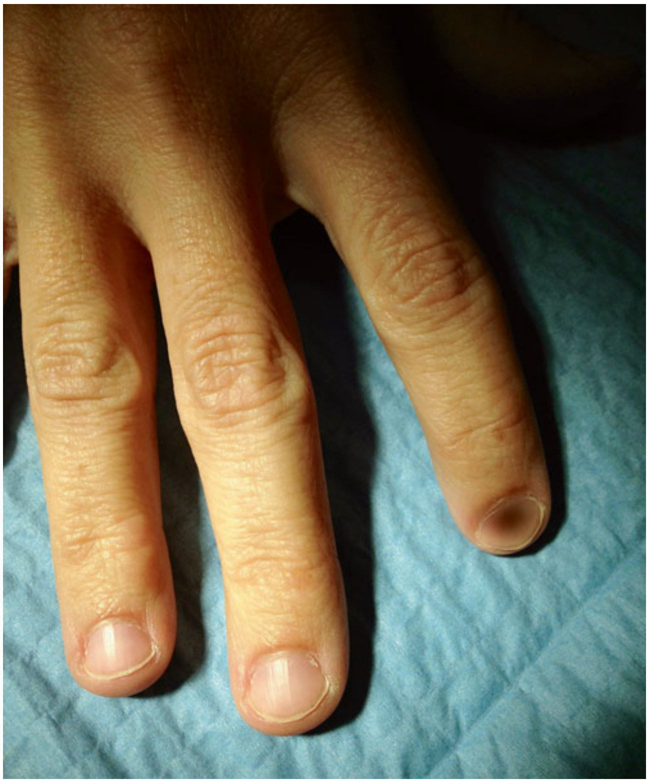

Fig. 96.1 Subungal hematoma of right index finger

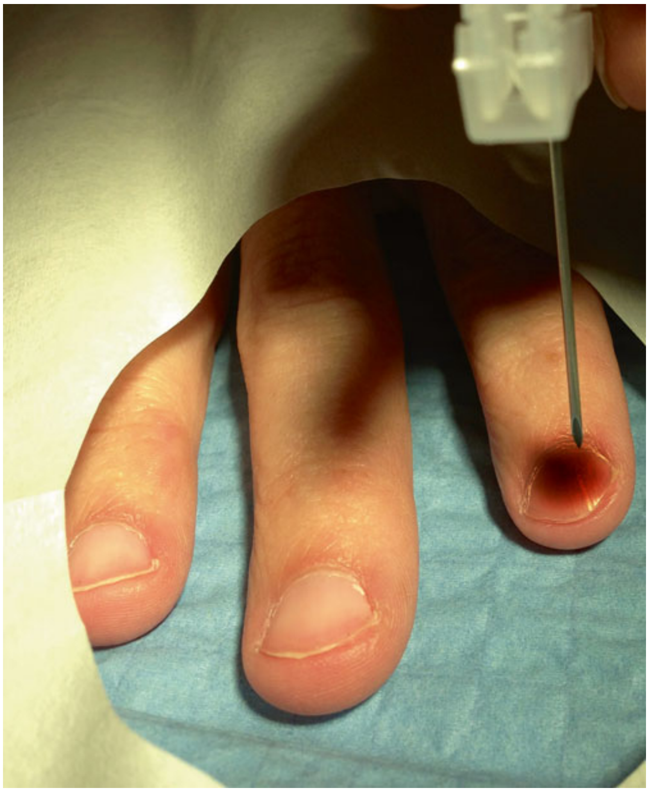

Fig. 96.2 Positioning of needle to perform subungal hematoma drainage

96.6 Complications (Rare)

- Infection
- Injury to nail bed or underlying bone if too much pressure is applied with puncture
- Onycholysis when there is incomplete drainage in rare cases

96.7 Pearls and Pitfalls

- Pearls
  - Multiple holes may be necessary for appropriate drainage of the hematoma.
  - When using electrocautery for hole puncture, execute high caution with acrylic nails because they are flammable.
  - When using a paper clip or regular needle, make sure to sterilize the tip over a flame.
  - Take a radiograph of the finger whenever necessary to rule out phalangeal fracture.
  - Check for avulsion of the extensor tendon.
  - Inform the patient that the existing nail may fall off and will regenerate in few months if the nailbed is intact.
  - Systemic or oral antibiotics are not recommended.
- Pitfalls
  - Development of a dark color change over the nailbed without any history of trauma should raise suspicion of a tumor and should be evaluated accordingly.
  - Do not remove the nail to evaluate injury to nailbed.

Selected Reading

Brown RE. Acute nail bed injuries. Hand Clin. 2002;18:561–75.

Dean B, Becker G, Little C. The management of the acute traumatic subungual haematoma: a systematic review. Hand Surg. 2012;17:151–4.

Incision and Drainage of Abscess

97

Nicholas D. Caputo, Karlene Hosford,
and Muhammad Waseem

97.1 Indications

- Abscess greater than 5 mm in diameter and in accessible areas (e.g., axilla, extremities, trunk)

97.2 Contraindications

- Absolute
  - Absence of fluctuation
  - Large, deep, and complicated (multiloculated) abscesses
  - Location
    - Perianal
    - Mastoid
- Relative
  - Location
    - Face (e.g., nose, nasolabial fold)
    - Palms
  - Coagulopathy
  - Recurrent pilonidal cysts (may mandate operative excision)
  - Area of cosmetic importance where aspiration may be preferred

97.3 Materials and Medications

- Incision and drainage tray (Fig. 97.1)
  - Drape
  - Betadine (povidone-iodine) swabs
  - 1 % lidocaine
  - 18- and 27-gauge needles
  - 12-mL syringes, gauze pads
  - #11 scalpel, mosquito clamps (hemostat)
  - Iodoform packing of appropriate size
- Ultrasound machine (Fig. 97.2)

N.D. Caputo, MD, MSc
Emergency Department Critical Care, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: Ncaputo.md@gmail.com

K. Hosford, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: krlhos@aol.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_97

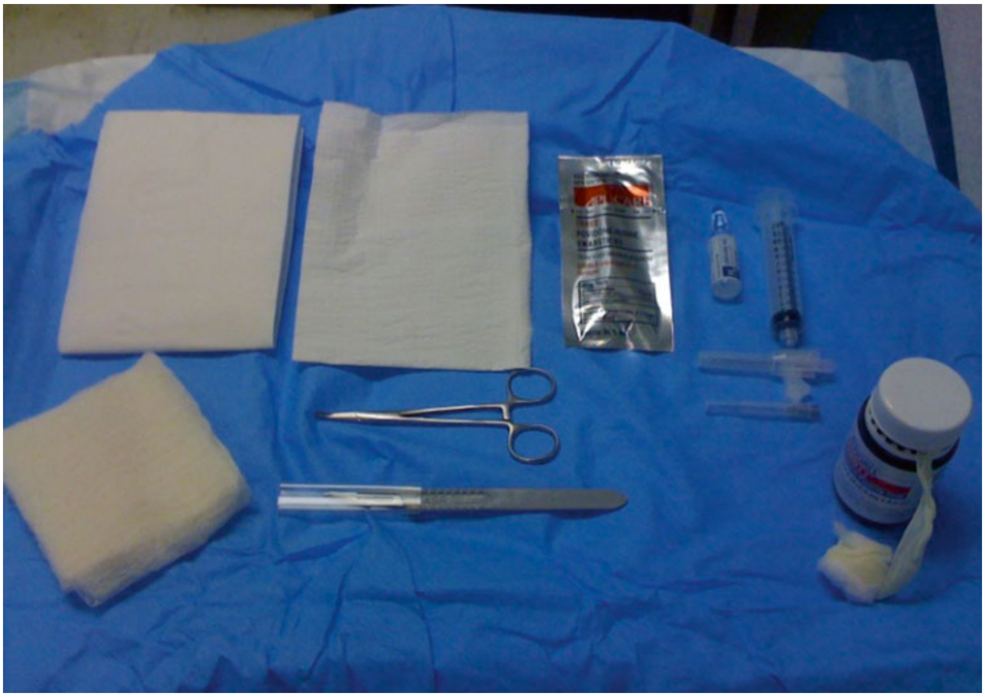

Fig. 97.1 Supplies necessary for incision and drainage

Fig. 97.2 Bedside SonoSite ultrasound

97.4 Procedure

1. Ultrasound (optional) may be helpful when abscess is suspected in the absence of fluctuation. Using the vascular probe (7 mHz), confirm the clinical suspicion of abscess and check the depth and width of the abscess (Fig. 97.3).
2. Sterile skin preparation with Betadine swab and sterile drape.
3. Anesthetize the appropriate area subcutaneously with 5 mL of 1 % lidocaine by inserting the 27-gauge needle at an acute angle into the intradermal space (Fig. 97.4a, b).
4. Using a #11 blade, make an approximately 1- to 2-cm skin incision over the desired area parallel to the Langer lines. The incision must approach into the abscess cavity (Fig. 97.4c).
   - Some physicians still advocate the technique of making a cruciate incision. This may leave a larger scar and should be discussed with patient before doing so because of cosmetic consequences.
5. Allow for spontaneous drainage. After resolution of drainage, you may express more pus with gentle downward pressure.
6. Using the hemostat, enter the incision to break any suspected loculations. This should be done with the clamps closed and curved part down. The clamps should then be opened and removed slowly (Fig. 97.4d).
7. After clearing the remaining loculations, the wound should be packed.
   - Evidence suggests that packing the abscesses does not prevent recurrence; however, this is still practiced. If packing the wound, follow the next step.
8. Take the iodoform packing with the hemostat, and place the packing into the incision site until no further packing will fit.
9. Cut the packing leaving a tail out of the incision site (Fig. 97.4e).
10. Apply dressing with 4×4 gauze and adhesive tape (2 in.).

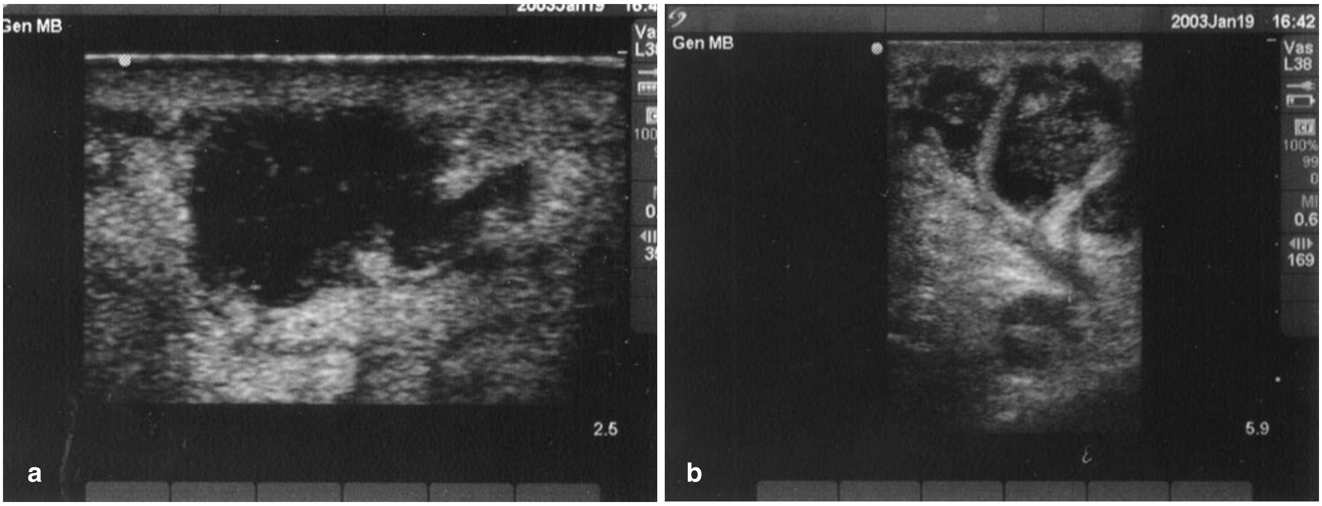

Fig. 97.3 (**a**) Example of an abscess as viewed on bedside ultrasound. (**b**) A multiloculated abscess

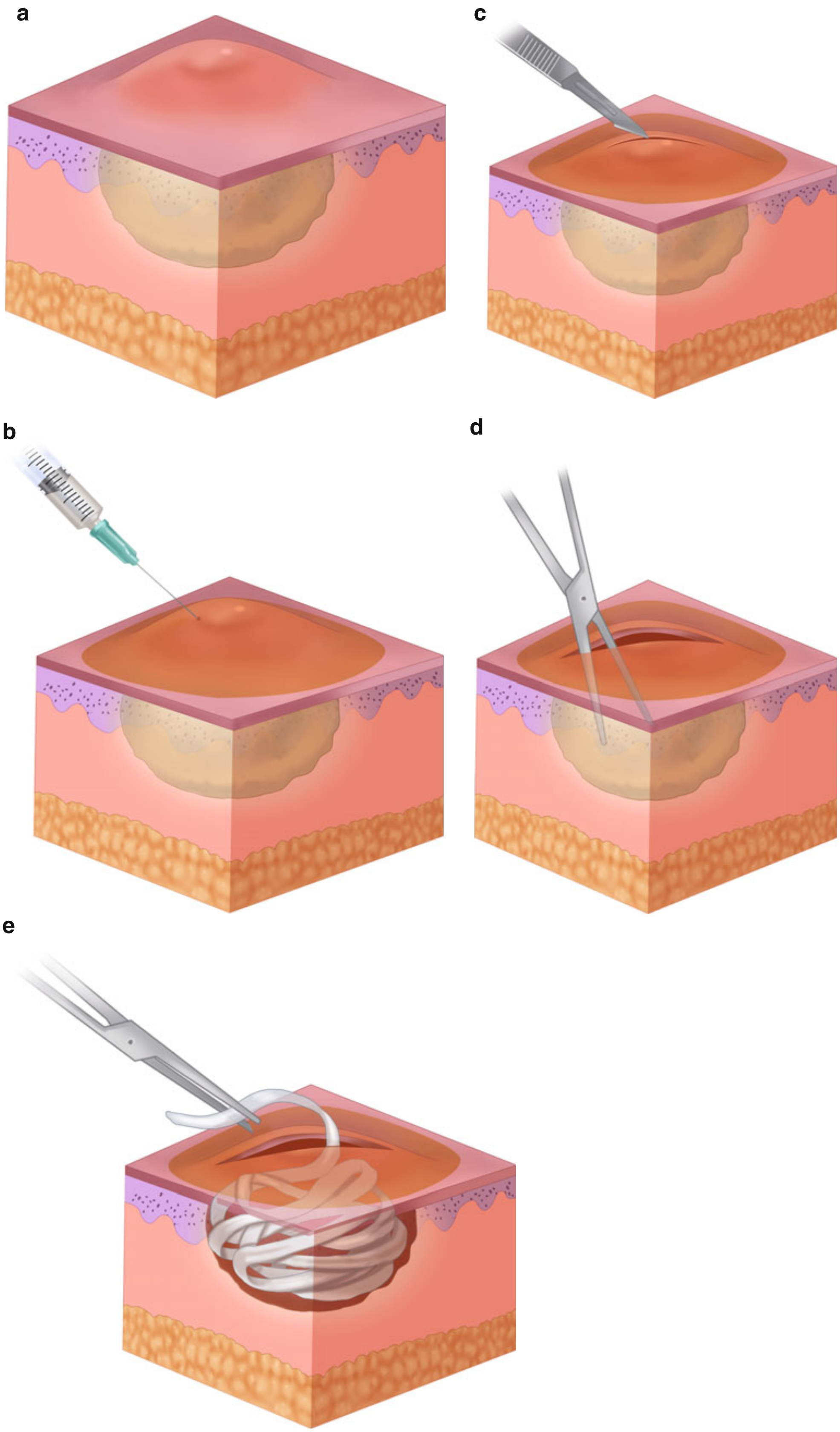

Fig. 97.4 (**a**) Abscess with overlying erythema. (**b**) Lidocaine injection in the superficial layer. (**c**) Linear incision with #11 blade. (**d**) Expression of purulent material and breaking of loculations with clamps. (**e**) Optional placement of packing

97.5 Complications

- Recurrence of abscess
- Progression of cellulitis
- Neurovascular injury to adjacent structures

97.6 Pearls

Antibiotic coverage is a controversial topic. Methicillin-resistant *Staphylococcus aureus* (MRSA) is a concern not only in the immunocompromised and diabetic patients. *S. aureus* has been detected in up to 51 % of patients with abscesses. Of these isolates, approximately 75 % were MRSA. Bactrim (trimethoprim/sulfamethoxazole) should be utilized for all prophylactic measures.

Selected Reading

Barnes SM, Milsom PL. Abscess: an open and shut case. Arch Emerg Med. 1988;5:200–5.

Burney RE. Incision and drainage procedures: soft tissue abscesses in the emergency service. Emerg Med Clin North Am. 1986;4:527–42.

Duong M, Markwell S, Peter J, Barenkamp S. Randomized, controlled trial of antibiotics in the management of community-acquired skin abscesses in the pediatric patient. Ann Emerg Med. 2010;55:401–7.

Fitch MT, Manthey DE, McGinnis HD, Nicks BA, Pariyadath M. Abscess incision and drainage. N Engl J Med. 2007;357:e20.

Frazee BW, Lynn J, Charlebois ED, Lambert L, Lowery D, Perdreau-Remington F. High prevalence of methicillin-resistant *Staphylococcus aureus* in emergency department skin and soft tissue infections. Ann Emerg Med. 2005;45:311–20.

Hankin A, Everett WW. Are antibiotics necessary after incision and drainage of a cutaneous abscess? Ann Emerg Med. 2007;50:49–51.

O'Malley GF, Dominici P, Giraldo P, et al. Routine packing of simple cutaneous abscesses is painful and probably unnecessary. Acad Emerg Med. 2009;16:470–3.

Part XIII

Orthopedic Procedures

Splinting

98

Christopher H. Stahmer and Muhammad Waseem

98.1 Indications

- Need for immobilization for fracture, dislocation, or soft tissue injury
- Suspicion for occult injury of an extremity
- Immobilization for pain management

98.2 Contraindications

- Absolute
  - Open fracture (requires operative intervention)
- Relative
  - Infection
  - Compartment syndrome

98.3 Materials and Medications (Fig. 98.1)

- Plaster of Paris
- Fast drying: 5–8 min to set
- Extra fast drying: 2–4 min to set
  - Variety of widths depending upon splint of choice:
  - Splints may take up to 2 days to dry and achieve maximum strength.
- Prefabricated splinting materials
  - Plaster OCL® (Orthopedic Casting Laboratories)
    - 10–20 sheets of plaster with padding and cover
    - Faster setup time but less customizable
  - Fiberglass splints
    - Cure rapidly
    - Less messy
    - Less moldable
    - Stronger and lighter
- Stockinette.
  - Protects the skin.
  - Variety of sizes available.
- Soft wrap (Webril™).
  - Provides padding.
  - Five to six layers depending on anticipated swelling.
    - Too much padding reduces the stability of the splint.
  - Use extra padding over bony prominences.
  - Pad between digits for splinting of digits.
  - Avoid wrinkles, which generate pressure points.
  - Do not wrap circumferentially.
    - Increased risk of ischemia.
- Ace wraps.
  - Variety of sizes depending.
    - Larger widths over legs.
    - Narrow widths around fingers and joints.
      - Avoid bunching by using narrow widths at joints.
- Water
  - Warm water and splint sets more quickly but increases the risk of burns.
    - Splint drying is an exothermic or heat-releasing reaction.
  - Hot water leaves less time to mold the splint.

C.H. Stahmer, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: stahmer@gmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_98

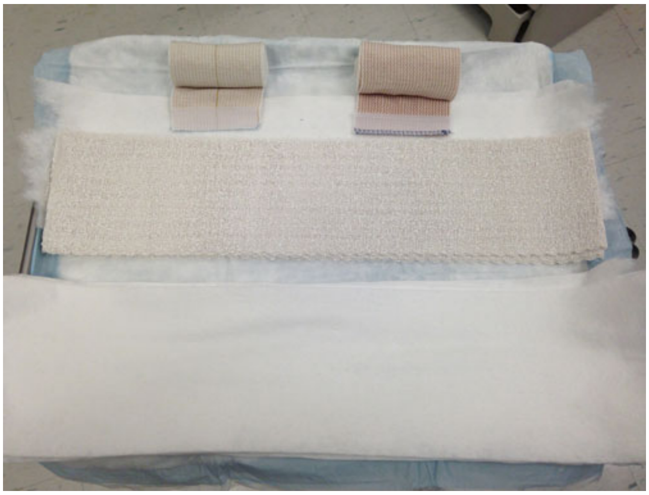

Fig. 98.1 Posterior splint materials: plaster of Paris, Ace wraps, soft roll. Note there are two layers of soft wrap: the inner layer to face the patient (eight layers) and the outer layer to pad the exterior (two layers)

98.4 Procedure

1. Completely expose and examine the afflicted body part for tissue, vascular, or neurological injury.
    - Address respective injuries before proceeding.
2. Lay out all splinting materials before initiating procedure.
    - Layer plaster of Paris.
        - Upper extremity: 8–10 layers.
        - Lower extremity: 12–15 layers.
        - Up to 20 for a large person.
        - More layers of plaster of Paris increase the risk of burn and the weight of the splint.
3. Administer appropriate anesthesia.
    - Conscious sedation
    - Hematoma block
    - Intra-articular injection
    - Intravenous pain medication
    - Oral pain medication
4. Hang fractures as indicated for improved success of reduction to relax muscles before reduction attempt.
5. Reduce afflicted extremity.
6. While maintaining reduction, apply respective splint.
7. Apply in the following order for plaster of Paris splint.
    - Stockinette (not necessary).
    - Soft wrap.
    - Select appropriate layers of plaster of Paris.
    - Prepare plaster of Paris to create splint:
        - Layer plaster with no overlap.
        - Submerge completely into water.
        - Crumple into ball without letting go of the ends of the splint.
        - Release the lower end of the splint while holding the top tightly together.
        - Run fingers in a “squeegee” manner from top to bottom to smooth the splint (Fig. 98.2).
            - This also removes excess water.
            - Repeat until the splint is smooth and free of dripping water.
        - Apply soft wrap layers to the splint.
            - Apply thicker layer to the patient’s body.
            - Apply two or three layers of soft wrap to the exterior of plaster of Paris for padding and to facilitate drying.
        - Apply Ace wrap to hold the splint and assist in contouring the splint to the patient’s extremity (Fig. 98.3).
            - Applying the Ace wrap too tightly may cause ischemia. Observe the patient after splinting for 30 min for tingling, burning, pain, or discomfort.
        - Mold the splint without making indentations with the fingertips (Fig. 98.4).
            - An indentation may cause a pressure point, which may result in an ulcer.

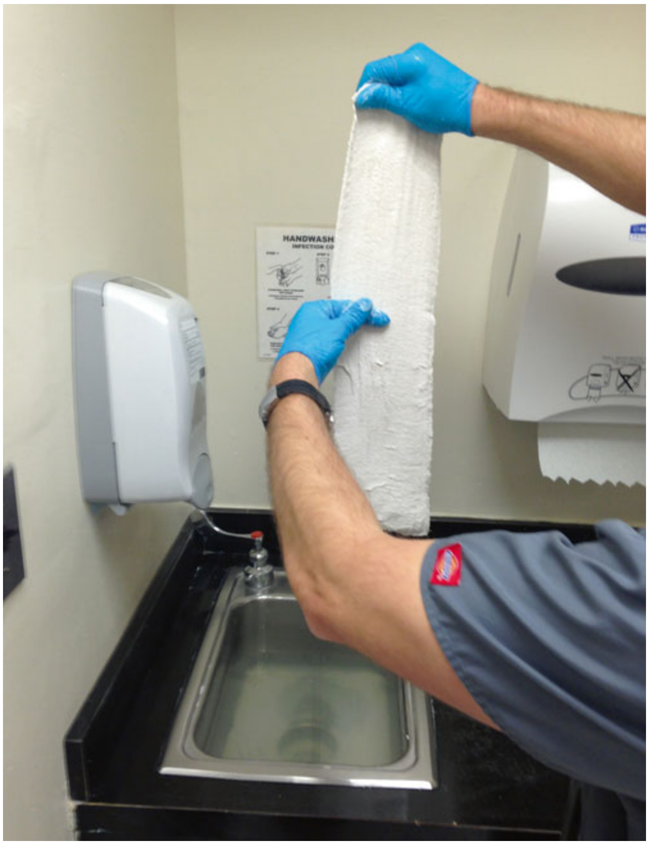

Fig. 98.2 Hold the top of the saturated plaster securely with one hand while removing excess water with the other hand

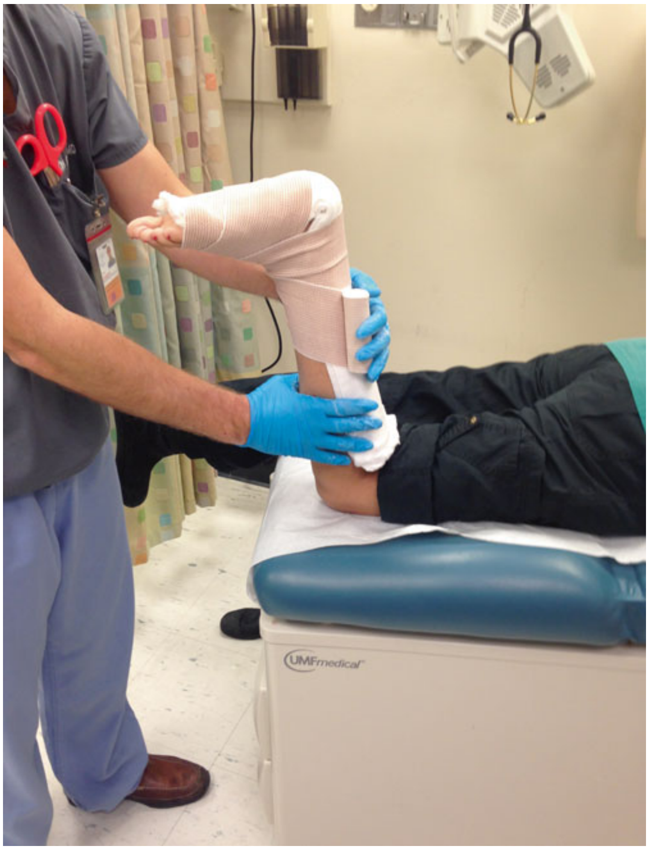

Fig. 98.3 Apply Ace wrap to hold the splint and assist in contouring the splint to the patient's extremity

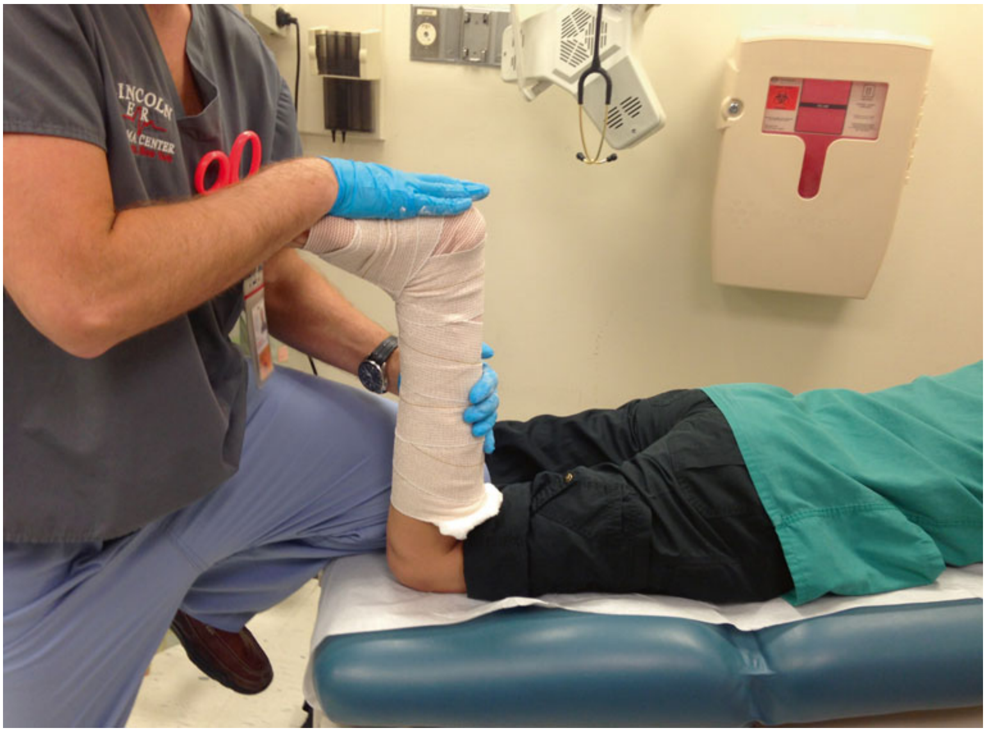

Fig. 98.4 Hold the splint in a neutral anatomical position while taking care not to make indentations with the fingertips

  - Allow the splint to cure while the practitioner maintains the appropriate position. This will take approximately 5 min depending upon water temperature and splint thickness.

98.5 Complications

- Ischemia may result in compartment syndrome.
  - Advise the patient to unwrap the splint for the following indications.
    - Increasing pain.
    - Discoloration of fingers, toes, or the splinted extremity.
    - Loss of sensation of splinted extremity.
- Burns
  - Plaster drying releases heat.
  - Increased risk with limited layers of padding.
  - If pain is troubling the patient, remove the splint and add more padding.
- Pressure sores
  - Apply ample padding.
  - Smooth all wrinkles.
  - Instruct the patient to return for increased discomfort.
- Infection
  - Clean and débride all devitalized tissue before application.
  - Requires close follow-up to reevaluate wounds.

Selected Reading

Fitch MT, Nicks BA, Pariyadath M, McGinnis HD, Manthey DE. Basic splinting techniques. N Engl J Med. 2008;359:e32.

Marx JA, Hockberger R, Walls R, editors. Rosen’s emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Simon R, Sherman S, Koenigsknecht S. Emergency orthopedics—the extremities. New York: McGraw-Hill; 2007.

Ulnar Gutter Splint

99

Jeffrey Kile, Katrina John, and Amish Aghera

99.1 Indications

- Fractures and soft tissue injuries of the ring or little finger
- Fractures of the neck, shaft, or base of the fourth or fifth metacarpal

99.2 Contraindications

- Relative
  - Evidence of compartment syndrome or any neurovascular compromise

99.3 Materials (Fig. 99.1)

- Splint roll (plaster of Paris, prefabricated foamcore, or fiberglass)
- Stockinette
- Cotton padding, such as simple cotton sheet wadding or a newer alternative such as Webril (Curity), Specialist (Johnson & Johnson), and so on
- Elastic bandages, such as Ace (3M) or similar
- Shears
- Adhesive tape
- Gloves

J. Kile, MBBS, PhD, MPH (✉) • K. John, MBBS
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: jeffrey.kile@gmail.com; trenjohn@me.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_99

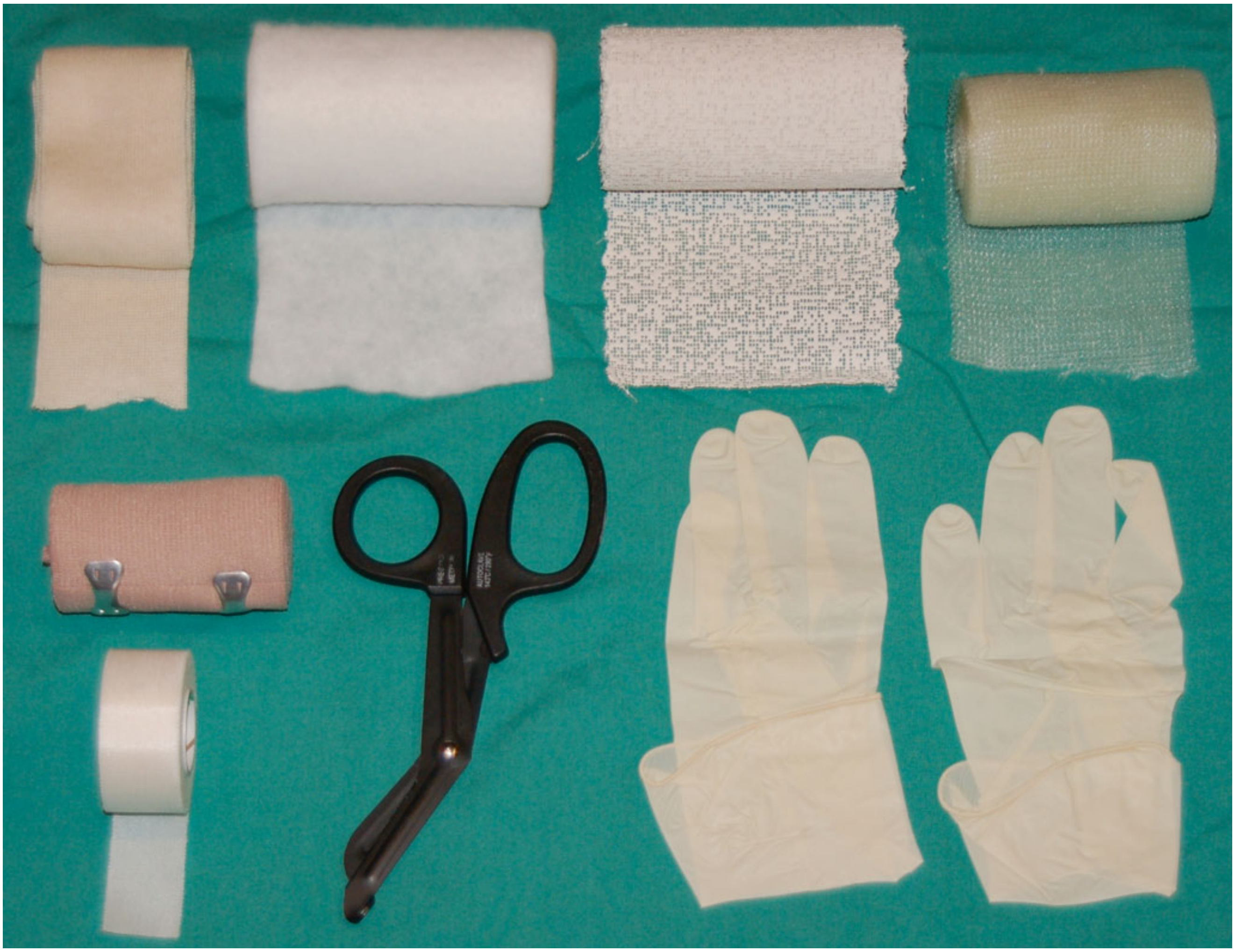

Fig. 99.1 Materials and medications

99.4 Procedure

1. Select the appropriate diameter stockinette according to the size of the forearm.
   - Generally, 3-in. diameter stockinette is used for the ulnar gutter splint.
2. Cut the stockinette to a length approximately 20 cm longer than the total length of the desired splint.
3. Apply the stockinette to the injured limb so that the stockinette extends approximately 10 cm beyond the region to be splinted proximally and distally.
4. Cut the distal aspect of the stockinette to free the thumb and also the index and middle finger (Fig. 99.2).
5. Insert cotton padding cut to the appropriate size between the ring and the little fingers to prevent skin degradation (Fig. 99.3).
6. Wrap cotton padding circumferentially around the entire region to be splinted, with each turn overlapping the previous turn by approximately 25 % of its width.
   - Generally, 3-in.-wide padding is used for the ulnar gutter splint.
   - Apply padding to an approximate thickness of 1 cm, with extra padding placed over the bony prominences of the wrist and carpometacarpal joints.
   - Extend the padding at least an inch beyond the desired splint length so that it may later be folded back over the jagged ends of the plaster or fiberglass splint roll (Fig. 99.4).
7. Unwrap and prepare an appropriate length splint.
   - The width of splinting material used should be approximately one-half the circumference of the extremity measured at the wrist, such that it is wide enough to extend approximately halfway around the distal forearm once applied.
   - Layer the splint eight sheets thick if using plaster roll and six sheets thick if using fiberglass roll.
   - It is best to begin with a generous length of splinting material because both plaster and fiberglass splint rolls shrink slightly once moistened. If the wet splint turns out too long, the ends can be either folded back or cut with shears before hardening.
8. Apply the dry splinting material to the ulnar aspect of the extremity from the midforearm to just beyond the distal interphalangeal joint of the little finger and cut the splint to approximately 20 cm longer than the desired length.
9. Submerge the dry splint in water until bubbling stops.
10. Remove the splint from water, place on a hard flat surface, and smooth out the excess water to ensure no wrinkles are present.
11. Apply to the splint to the limb over the cotton padding.
    - If the splint is too long at this stage, the ends may be folded back or cut with shears to the proper length.

12 Once positioned properly, fold the side of the splint up around the ulnar aspect of the forearm and hand to form a gutter (Fig. 99.5).

13. Fold the underlying cotton padding and stockinette back over the ends of the splint, which both protects the skin and holds the splint in place.
14. Secure the loose end(s) of the elastic bandage with adhesive tape.
    - Avoid using the metal clips often packaged with elastic bandages because these can become displaced and embed in the skin.
15. Manipulate the forearm and hand into the "neutral" position: (1) the wrist in slight extension (10–20°), (2) the metacarpophalangeal joints in 50° of flexion, and (3) the proximal and distal interphalangeal joints of the ring and little finger in slight flexion (10–15°) (Fig. 99.6).
    - When splinting a boxer's (i.e., metacarpal neck) fracture, the metacarpophalangeal joint should be flexed to 90°.
16. Mold the splint to the contour of the extremity (Fig. 99.7).
    - Use only the palms of the hands when molding because the fingertips may cause indentations, resulting in excessive skin pressure.
17. Wrap the extremity with an elastic bandage in a distal-to-proximal direction to secure the splint in place (Fig. 99.8).
18. Ensure the extremity is neurovascularly intact.
19. Instruct the patient to loosen the elastic bandage if it feels too tight.

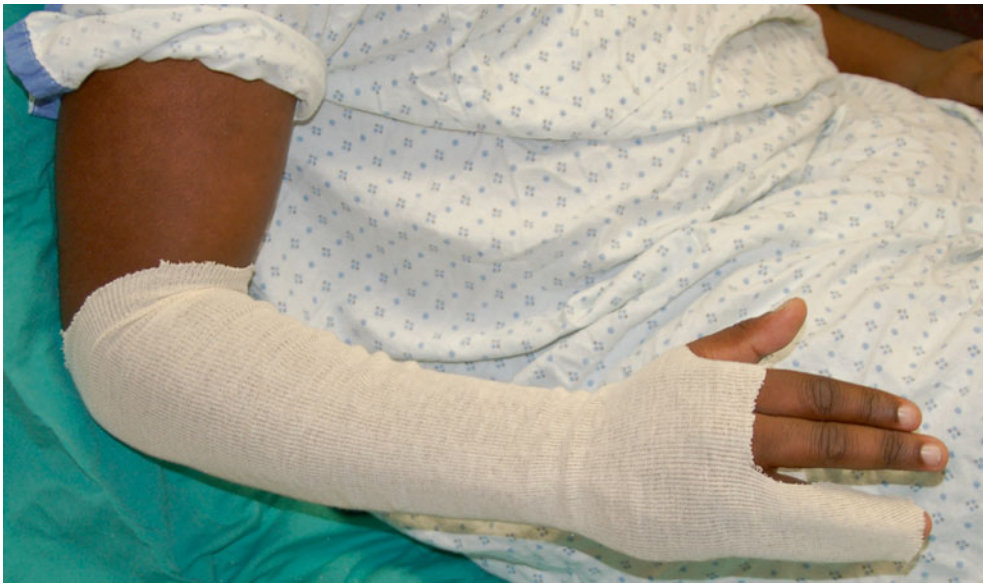

Fig. 99.2 Stockinette applied to the forearm and hand

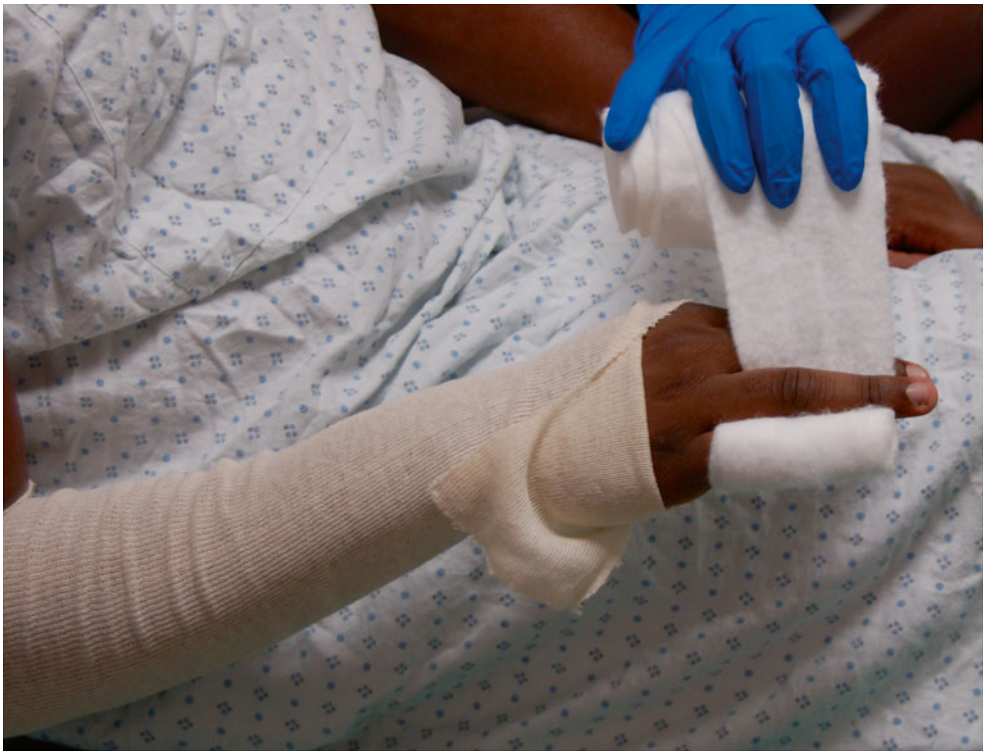

Fig. 99.3 Cotton padding applied to the ring and little fingers

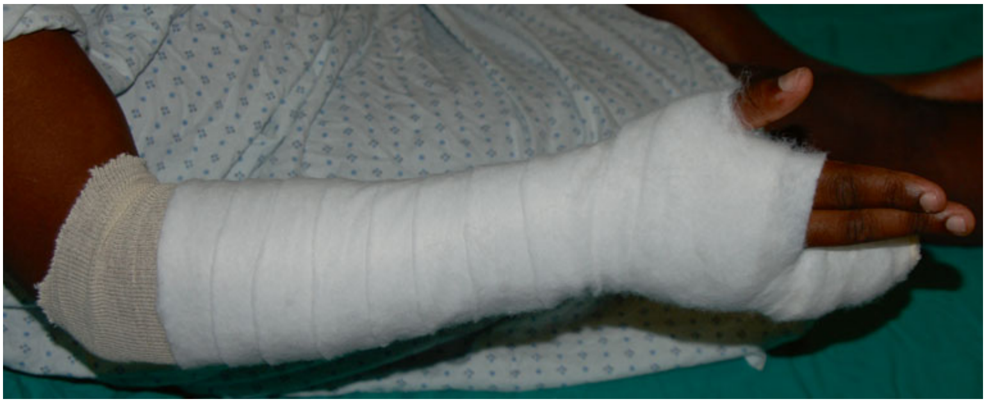

Fig. 99.4 Cotton padding applied to the forearm and hand

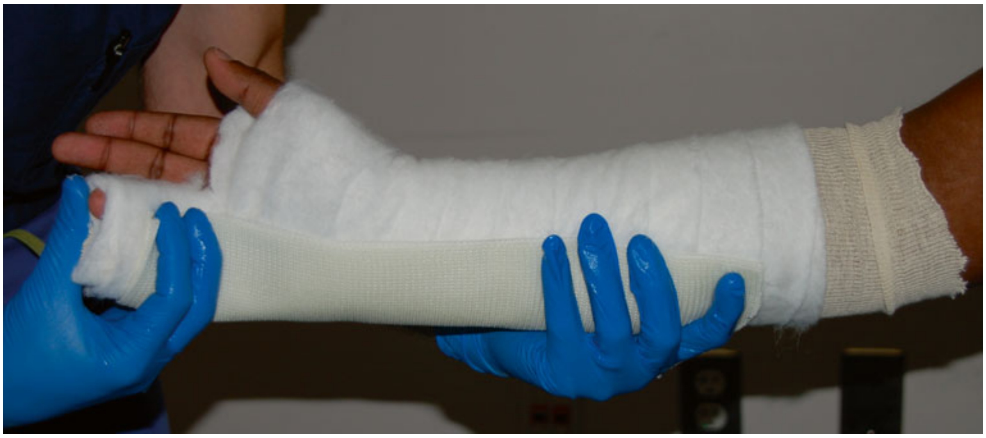

Fig. 99.5 Splint folded around the ulnar aspect of the forearm and hand

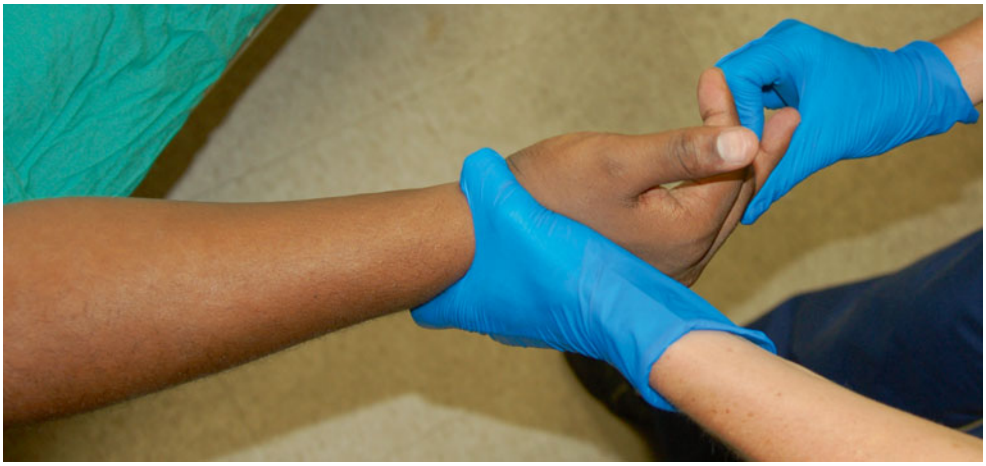

Fig. 99.6 Forearm and hand in the neutral position (without splint)

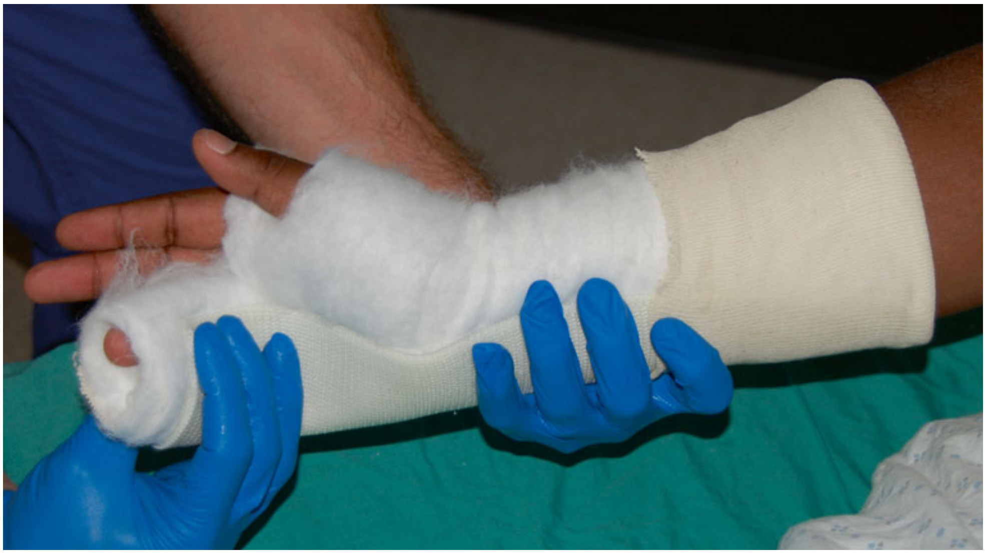

Fig. 99.7 Splint molded to maintain the forearm and hand in the neutral position

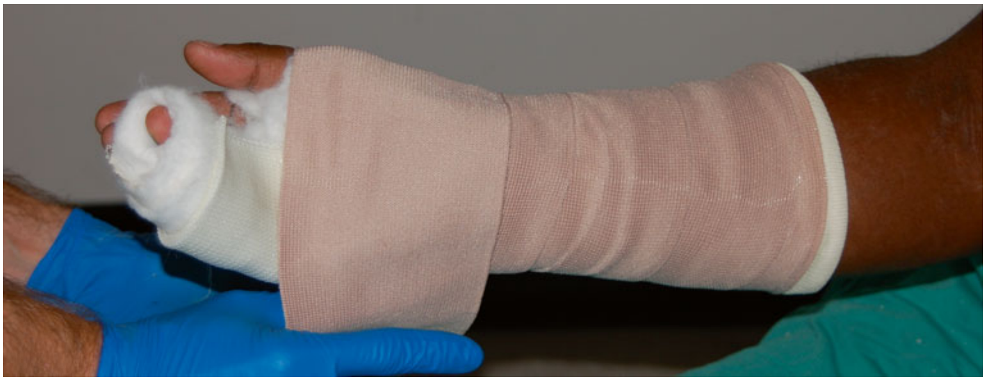

Fig. 99.8 Splint secured in place with an elastic bandage

99.5 Complications

- Neurovascular or other soft tissue compromise (if the splint/bandage is too tight)
- Soft tissue degradation (if the splint is left in place too long)

99.6 Pearls and Pitfalls

- Pearls
  - Simple plaster of Paris splints are inexpensive and allow a thoroughly customizable fit but can be damaged by water and require more time to set and more clean up than prefabricated foamcore or fiberglass splint rolls.
  - Simple fiberglass splints set quickly, are not damaged by water, are stronger and lighter than simple plaster and prefabricated splints, and offer a fully customized fit, but are not applied as quickly as prefabricated splints.
  - Prefabricated fiberglass splints are quickly applied, require virtually no clean up, and are not damaged by water, but are relatively expensive and provide a somewhat less customizable fit than simple splint rolls.
  - Stockinette protects the skin and, when folded back over the ends of the plaster or fiberglass splint roll, holds the wet splint in place before the elastic bandage is placed, and provides a padded rim with a professional appearance.
- Pitfalls
  - Avoid wrinkling the cotton padding applied between the stockinette and the splinting material because, once under the pressure of the elastic bandage, wrinkles can cause unnecessary skin pressure.
  - Avoid using more or fewer layers of splinting material than recommended. Additional layers can result in excessive heat during the setting process and a splint that is too heavy, whereas insufficient layers can result in a splint that is too weak.

Selected Reading

Harrison BP, Hilliard MW. Emergency department evaluation and treatment of hand injuries. Emerg Med Clin North Am. 1999;17: 793–822.

Henry MH. Fractures of the proximal phalanx and metacarpals in the hand: preferred methods of stabilization. J Am Acad Orthop Surg. 2008;16:586–95.

Margic K. External fixation of closed metacarpal and phalangeal fractures of digits. A prospective study of one hundred consecutive patients. J Hand Surg Br. 2006;31:30–40.

McMahon PJ, Woods DA, Burge PD. Initial treatment of closed metacarpal fractures. A controlled comparison of compression glove and splintage. J Hand Surg Br. 1994;19:597–600.

Viegas SF, Tencer A, Woodard P, Williams CR. Functional bracing of fractures of the second through fifth metacarpals. J Hand Surg Am. 1987;12:139–43.

Shoulder Dislocation Reduction Techniques

100

Katrina Skoog Nguyen, L. Connor Nickels,
and Rohit Pravin Patel

100.1 Indications

- Subjective history of new-onset dislocation or recurrent dislocations combined with clinical assessment consistent with shoulder dislocation
  - Anterior dislocations (96 %)
    - Typical mechanism of injury being indirect, with combination of abduction, extension, and external rotation. Rarely, the etiology is a direct blow to the posterior shoulder.
    - Prominent acromion with a palpable drop off below the acromion and subclavicular region fullness is consistent with anterior shoulder dislocation.
  - Posterior dislocations (4 %)
    - Mechanism of injury is indirect with a combination of internal rotation, adduction, and flexion. Precipitating events include seizure, electrical shock, and falls.
    - More subtle presentation. Patient will maintain arm locked in internal rotation and adduction; he or she cannot externally rotate. Shoulder is flattened anteriorly and rounded posteriorly.
    - Ultrasound can be used to prevent missed or delayed diagnosis (Figs. 100.1 and 100.2 show probe positioning and a diagram of abnormal ultrasound anatomy).
  - Inferior dislocations (luxation erecta)
    - Arm will be held fixed in overhead position.
- Radiographs reveal shoulder dislocation.
- Ultrasound can be used to identify the nature of the dislocation (anterior or posterior) and can be determined by the position of the humeral head relative to the transducer and glenoid. Although at this point, it should not replace radiographs owing to missed fractures. Advantages may include less radiation (decreased need for postreduction x-rays) and re-sedation if reduction is not complete.

K. Skoog Nguyen, DO
Northwest Community Hospital, Arlington Heights, IL, USA
e-mail: katrinaskoog@hotmail.com

L.C. Nickels, MD, RDMS (✉) • R.P. Patel, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; rohitpatel@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_100

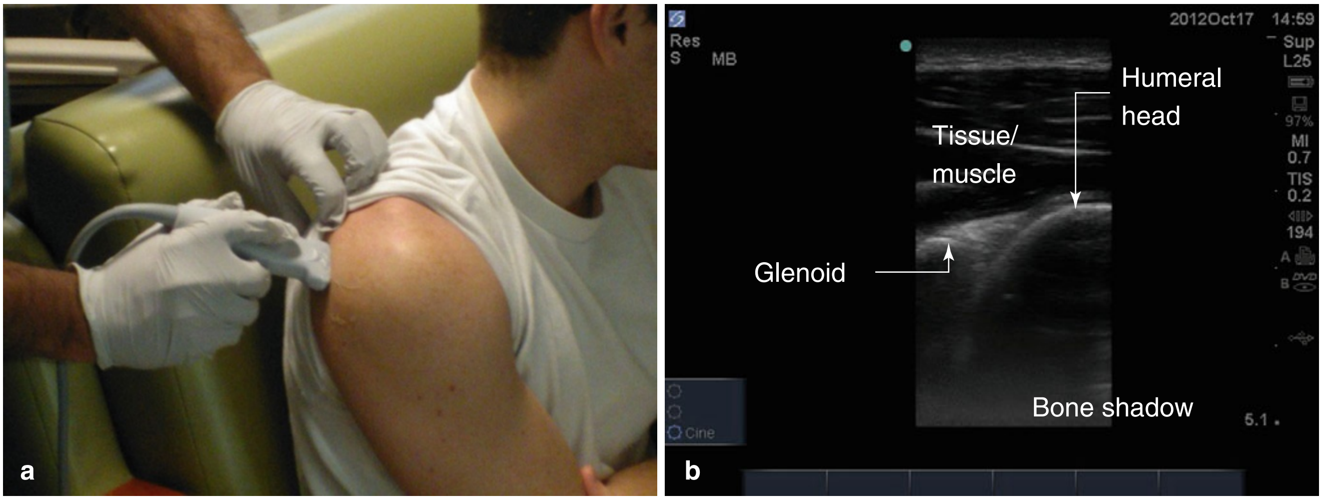
a
b
2012Oct17 14:59
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
194
Humeral head
Tissue/ muscle
Glenoid
Bone shadow
5.1
Cine

Fig. 100.1 (**a**, **b**) Ultrasound image of normal shoulder anatomy. "Dot fits the Dot" means when looking at the ultrasound machine from the sonographer's standpoint, the side of the probe marker corresponds to the side the marker is on the screen. This ensures when doing procedures, the direction of needle correction is the same as the orientation of the probe (Images courtesy of Dr. Rohit Patel)

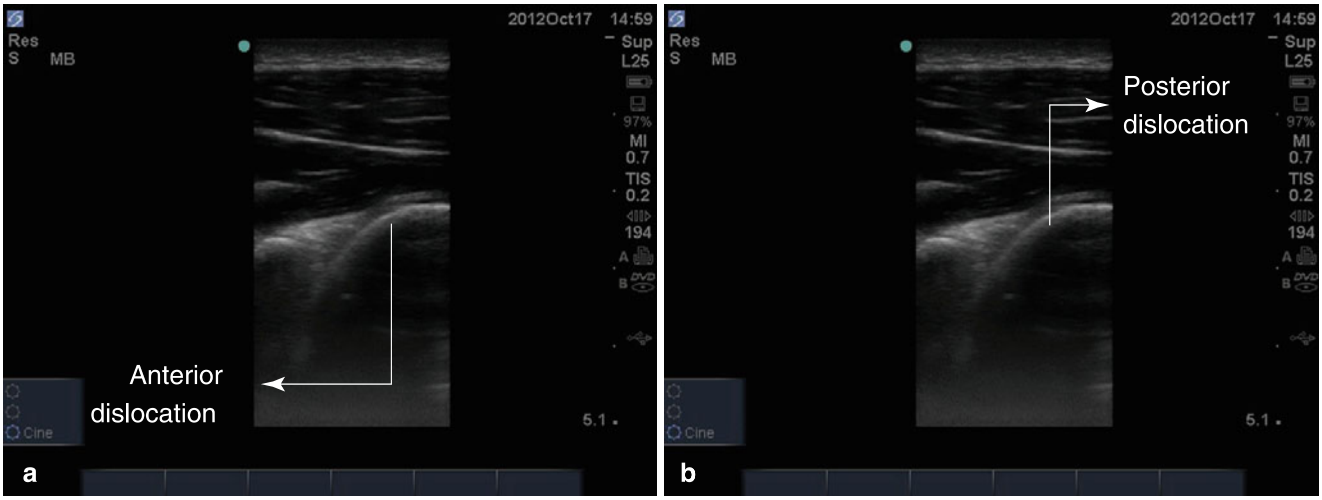
a
b
2012Oct17 14:59
Res
S MB
Sup
L25
97%
MI
0.7
TIS
0.2
194
Anterior dislocation
Posterior dislocation
5.1
Cine

Fig. 100.2 Anterior (**a**) and posterior (**b**) dislocations (Images courtesy of Dr. Rohit Patel)

100.2 Contraindications

- Associated fracture
  - This warrants orthopedic evaluation.
- Associated neurovascular deficit
  - May attempt reduction once but avoid multiple attempts.

100.3 Materials and Medications

- 1 % lidocaine, with syringe and needle and povidone-iodine prep if administering local anesthesia
- Moderate sedation medications if administering moderate sedation
- Bed sheet for traction-countertraction method
- Dangling weight for Stimson maneuver

100.4 Procedure

- Physical examination
  - Compare affected with unaffected shoulder.
  - Perform a complete neurovascular examination: test axillary, radial, ulnar, and median nerves for sensory deficit and motor function.
- Radiographs
  - Always obtain before attempting reduction for assessment of possible fracture and type and position of dislocation.
  - Obtain three views: anteroposterior, scapular Y, and axillary lateral views.
  - Anterior dislocations: humeral head appears anterior to the glenoid fossa on lateral or Y views.
  - Posterior dislocations: on anteroposterior view (vacant glenoid sign, 6-mm sign, lightbulb sign; on lateral or Y view: humeral head appears posterior to glenoid fossa).
- Pain management and sedation
  - Decide whether to use intra-articular lidocaine versus procedural sedation and analgesia.
  - For intra-articular lidocaine
    - Use 10–20 mL of 1 % lidocaine.
    - Attach a 1.5-in., 20-gauge needle.
    - Prepare the shoulder with povidone-iodine.
    - Insert the needle lateral to the acromion process and 2 cm inferiorly into the sulcus.
    - After withdrawing to ensure that the needle is not in a vessel, inject 10–20 mL lidocaine into the joint.
- Reduction techniques: it is important for the emergency department physician to be familiar with several different techniques. The following techniques are presented:

100.4.1 Stimson Maneuver (Fig. 100.3)

1. Patient is placed prone with 2.5–5 kg of weight hanging from the wrist.
2. Reduction may be facilitated by traction and external rotation of the arm.
3. A success rate of 96 % has been reported using the combined prone position, hanging weights, intravenous drug therapy, and scapular manipulation.

- Advantage: can be performed by one person only.
- Disadvantages: requires time to gather materials; the danger involved in the patient falling off the stretcher, requiring staff to monitor the patient.

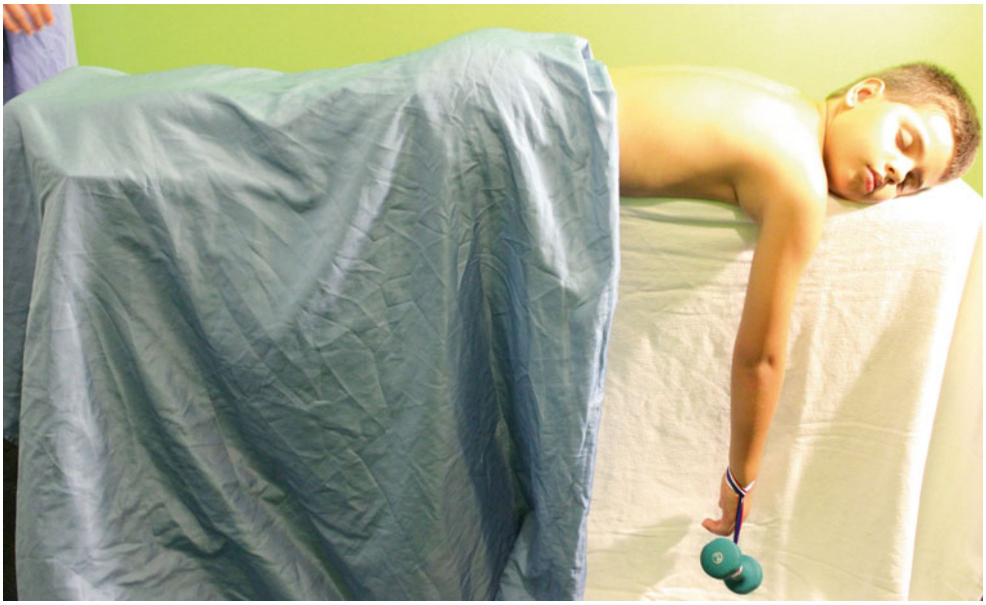

Fig. 100.3 Stimson maneuver

100.4.2 Scapular Manipulation Technique (Fig. 100.4)

1. Place the patient in the prone position with the affected arm hanging downward.
2. Apply traction down on the arm.
3. Locate the inferior tip of the scapula. Simultaneously push the inferior tip of the scapula medially toward the spine and use the other hand push the superior scapula laterally.

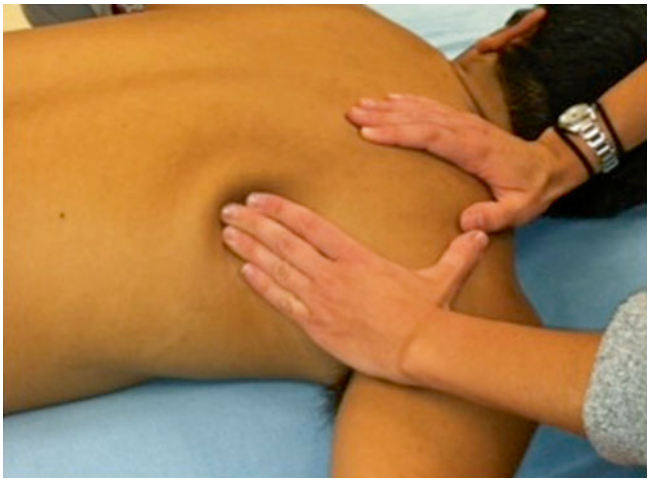

Fig. 100.4 Scapular manipulation method

- Advantages: high success rate, greater than 90 %; very safe to perform.
- Disadvantages: it requires the patient to assume the prone position; may require another person to perform traction.

100.4.3 External Rotation Method (Fig. 100.5)

1. Place the patient in the supine position with the affected arm adducted directly next to the patient's side with the elbow flexed to 90°.
2. The operator uses one hand to direct downward traction on the affected arm while maintaining it next to the patient's side.
3. The operator uses the other hand to hold the patient's wrist and guide the arm into slow external rotation.
4. Reduction usually takes place between 70° and 110° of external rotation.

- Advantages: requires no strength by operator; well tolerated by patients.
- Disadvantage: patient may have persistent dislocation during procedure, requiring operator to make adjustments.

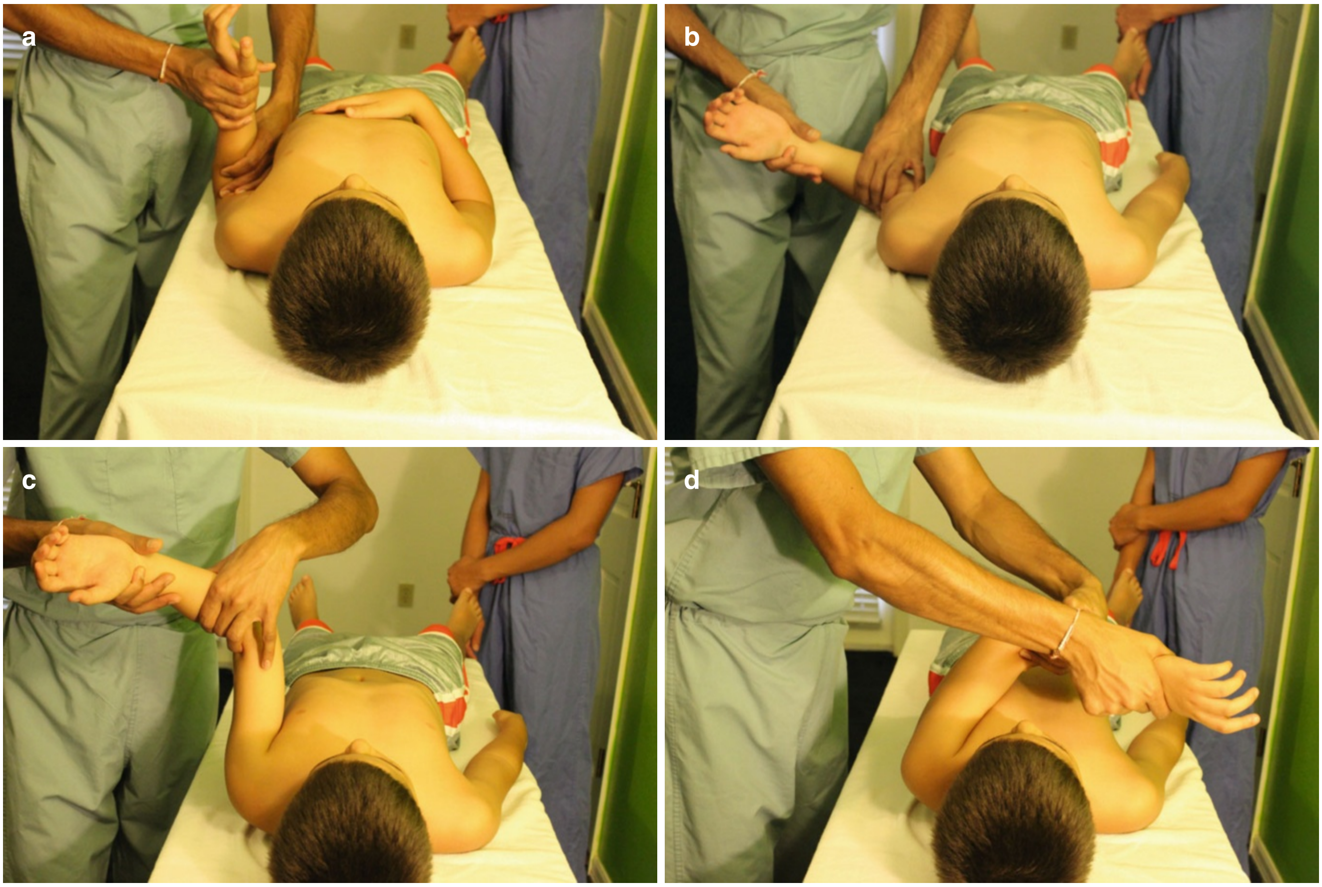

Fig. 100.5 (**a**–**d**) Kocher technique: external rotation method

100.4.4 Milch Technique (Fig. 100.6)

1. Technique looks as though one is reaching up to grab an apple from a tree.
2. Abduct the injured arm up to the overhead position.
3. Once in the overhead position, apply gentle vertical traction with external rotation.
4. An adjustment may need to be made if the reduction does not occur easily; push the humeral head upward into the glenoid fossa.

- Advantages: lack of complications; patient tolerance
- Disadvantage: variable success rate reported: 70–90 %

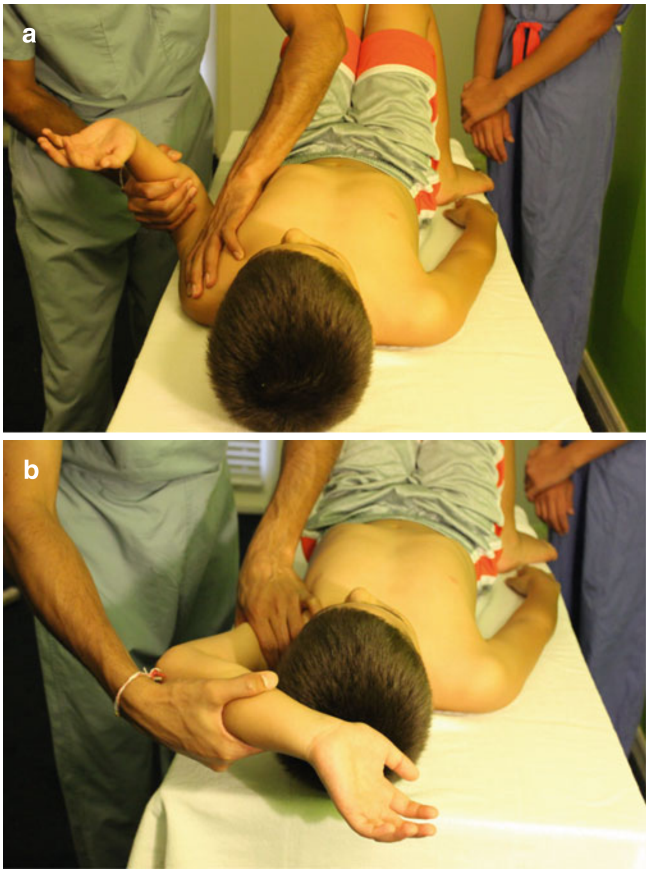

Fig. 100.6 (**a**, **b**) Milch technique

100.4.5 Spaso Technique (Fig. 100.7)

1. Place the patient in the supine position.
2. Operator grasps the affected arm at the wrist and lifts the straight arm directly upward while applying longitudinal traction.
3. Apply external rotation.

- Advantages: single operator, high level of success
- Disadvantage: may require more time to allow the shoulder muscles to relax

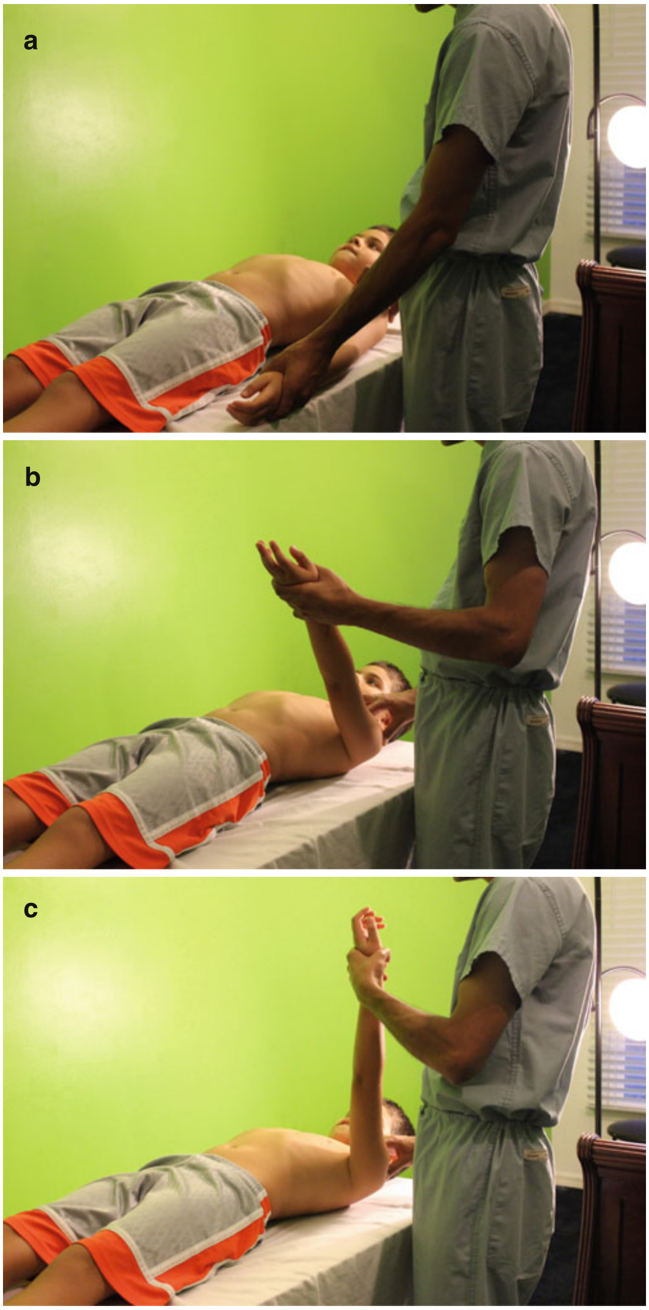

Fig. 100.7 (**a**–**c**) Spaso technique (Photographs courtesy of Dr. Pratik S. Patel)

100.4.6 Traction-Countertraction Technique (Fig. 100.8)

1. With the patient is sitting up, have an assistant wrap a sheet around the upper chest and under the axilla of the affected shoulder. Have the assistant wrap the sheet behind her or his back. Now have the patient lay supine.
2. Wrap another sheet around the flexed elbow of the affected arm and behind the operator's back.
3. Both the operator and the assistant lean back, applying gentle traction.

- Advantage: many older physicians are familiar with this method and, therefore, have a high degree of success.
- Disadvantages: requires two people; may cause skin tears on elderly patients.

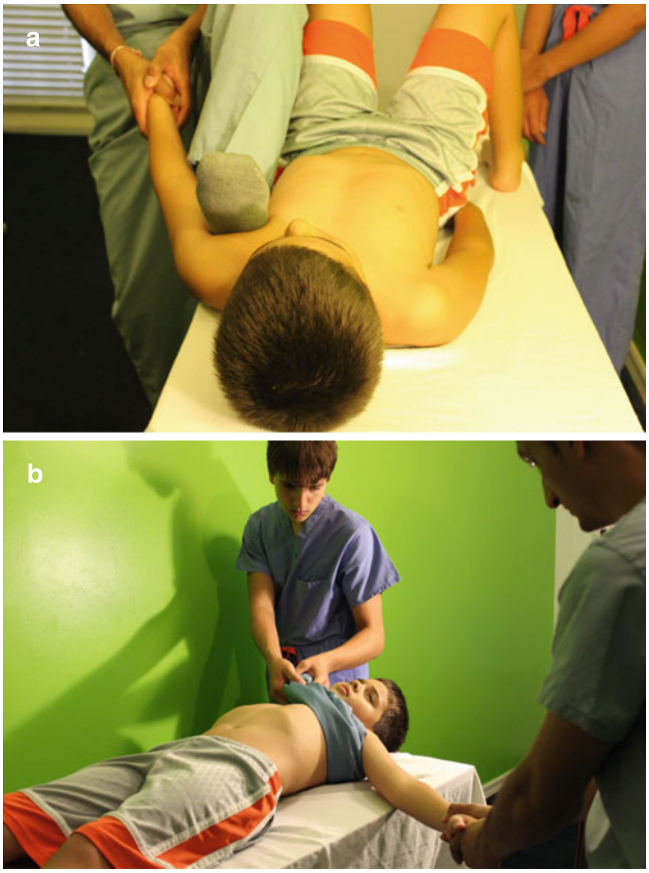

Fig. 100.8 (**a**, **b**) Hippocrates method/traction-countertraction method

100.4.7 Posterior Shoulder Dislocation Reduction

1. Give adequate premedication.
2. Place the patient supine and apply lateral traction on the proximal humerus.
3. Have an assistant apply anterior pressure to the posteriorly located humeral head.

- Advantage: logical methods for reduction
- Disadvantages: require sufficient premedication because often posterior dislocations present late; may require open reduction

100.4.8 Postreduction

- Obtain postreduction x-rays. There is some literature on using ultrasound to confirm adequate reduction, which allows repetitive assessments throughout procedure, as well as reduce radiation (see Fig. 100.2 for ultrasound of anterior and posterior dislocations).
- Do a postreduction neurovascular examination.
- Sling and swath or shoulder immobilizer for 2–3 weeks.
- Orthopedic follow-up in 1 week.

100.5 Complications

- Fractures
- Adhesive capsulitis, or frozen shoulder; especially a concern in the elderly with prolonged immobilization in sling
- Brachial plexus injury, especially of the axillary nerve
- Vascular laceration, most commonly of the axillary artery
- Rotator cuff tears

100.6 Pearls

- It is imperative to document the pre- and postreduction neurovascular status in the medical record.
- If unsure whether the reduction was successful, attempt to place the palm of the injured extremity on the contralateral shoulder. This is a good sign the reduction was successful.

Acknowledgment The authors would like to thank Karthik Stead for serving as the subject in many of the photographs in this chapter.

Selected Reading

Beck S, Chilstrom M. Point-of-care ultrasound diagnosis and treatment of posterior shoulder dislocation. Am J Emerg Med. 2013;31:449.e3–5.

Blakeley CJ, Spencer O, Newman-Saunders T, Hashemi K. A novel use of portable ultrasound in the management of shoulder dislocation. Emerg Med J. 2009;26:662–3.

Dala-Ali B, Penna M, McConnell J, Vanhegan I, Cobiella C. Management of acute anterior shoulder dislocation. Br J Sports Med. 2014;48(16):1209–15.

Simão MN, Noqueira-Barbosa MH, Muqlia VF, Barbieri CH. Anterior shoulder instability: correlation between magnetic resonance arthrography, ultrasound arthrography, and intraoperative findings. Ultrasound Med Biol. 2012;38:551–60.

Yuen CK, Chung TS, Mok KL, Kan PG, Wong YT. Dynamic ultrasonographic sign for posterior shoulder dislocation. Emerg Radiol. 2011;18:47–51.

Elbow Dislocation Reduction

101

Katrina John, Jeffrey Kile, and Amish Aghera

101.1 Indications

- Any dislocation of the elbow joint. Direction of the dislocation (i.e., anterior, posterior, lateral and divergent radius, and ulnar dislocations) is determined by the position of the ulna relative to the joint space (Fig. 101.1).

101.2 Contraindications

- Relative
  - Compound fracture dislocation

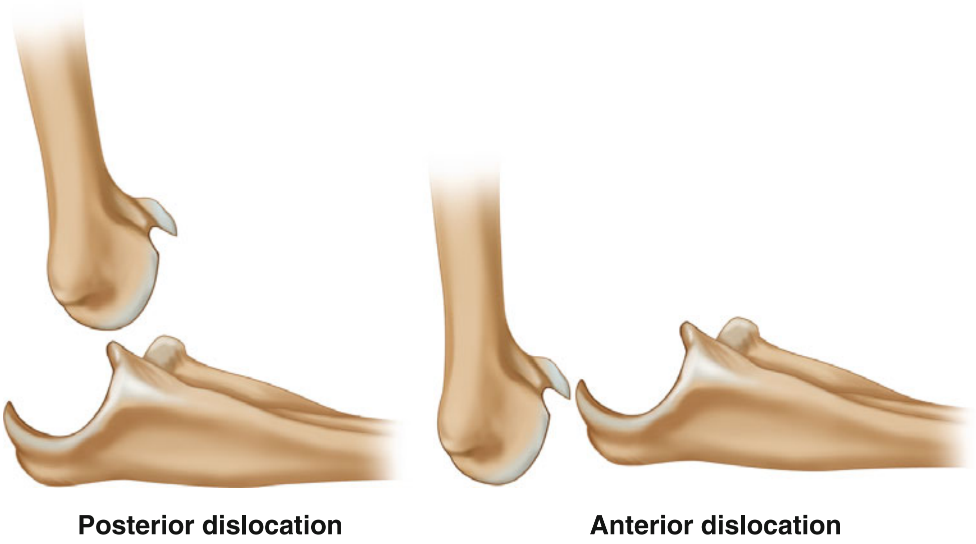

Fig. 101.1 Anatomical depiction

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_101

101.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Local anesthetic for local and intra-articular anesthesia
- Splinting material
- Stockinette
- Padding
- Elastic bandage
- Tape
- Sling

101.4 Procedure for Posterior Dislocations

1. Obtain a true lateral and anteroposterior radiographs of the affected elbow.
2. Ensure adequate sedation and analgesia.
3. Consider intra-articular analgesia.
4. Check the neurovascular status of affected extremity.
5. Follow a selected method for reduction as detailed later.
6. Following successful reduction gently flex the elbow to ensure full range of motion.
7. Place a long-arm posterior splint with the elbow in at least 90° flexion and secure the arm in a regular sling.
8. Check neurovascular status.
9. Obtain a postreduction radiograph of the elbow.

101.4.1 Method A (Fig. 101.2)

1. Position the patient on a stretcher in the supine position.
2. Apply steady traction at the supinated distal forearm keeping the elbow slightly flexed, while an assistant applies countertraction to the midhumerus with both hands.

101.4.2 Method B (Fig. 101.3)

1. Position the patient on a stretcher in the supine position.
2. Extend the affected extremity over the edge of the stretcher.
3. Apply traction to the supinated forearm slightly flexed at the elbow, while an assistant holds the distal humerus with both hands and uses thumbs to apply pressure to the olecranon as if pushing it away from the humerus.

101.4.3 Method C (Fig. 101.4)

1. Position the patient on a stretcher in the prone position.
2. Hang the affected extremity over the side of the stretcher toward the floor.
3. Apply downward traction to the pronated distal forearm and with the other hand just above the patient's antecubital fossa lift the humerus toward you.

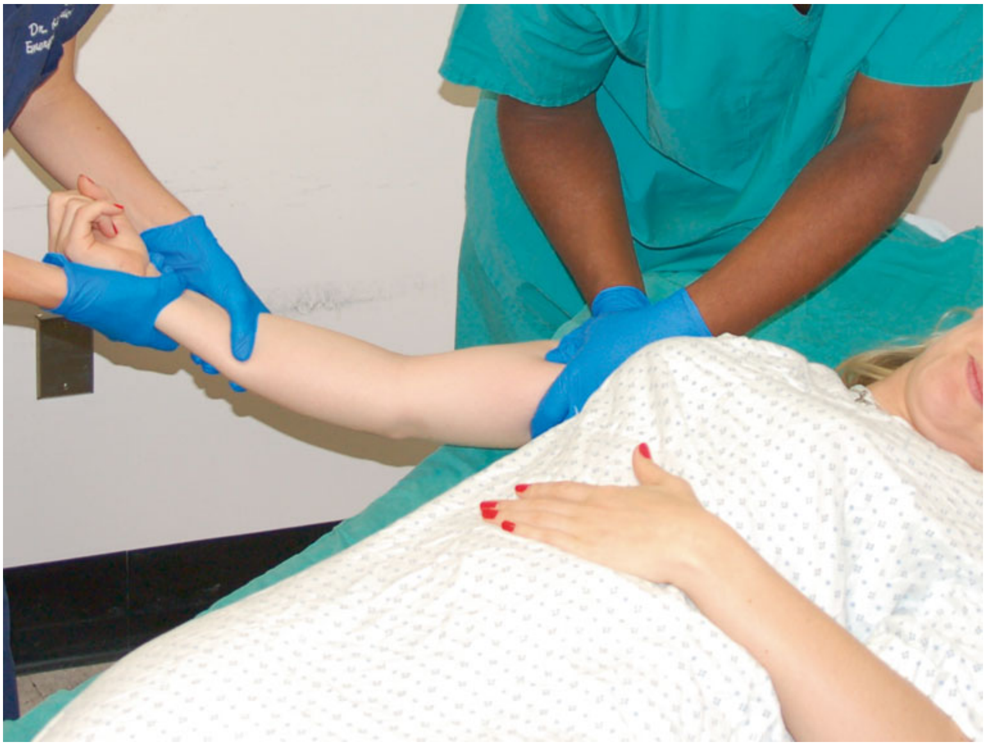

Fig. 101.2 Posterior method A

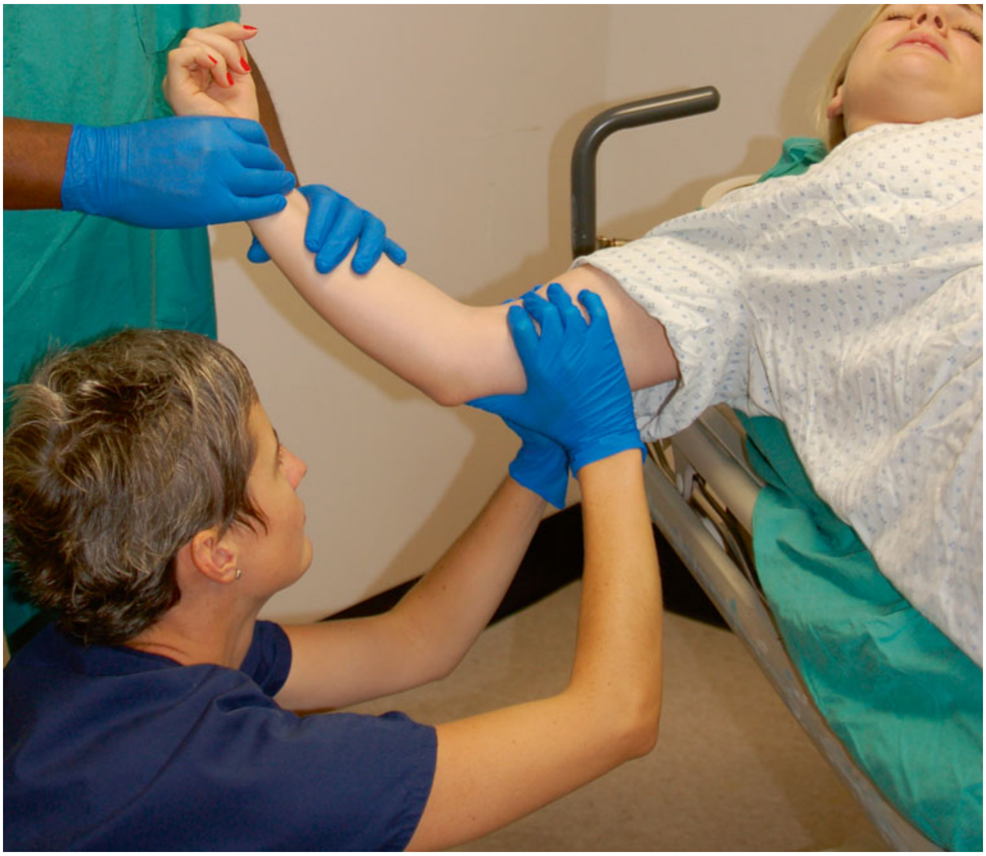

Fig. 101.3 Posterior method B

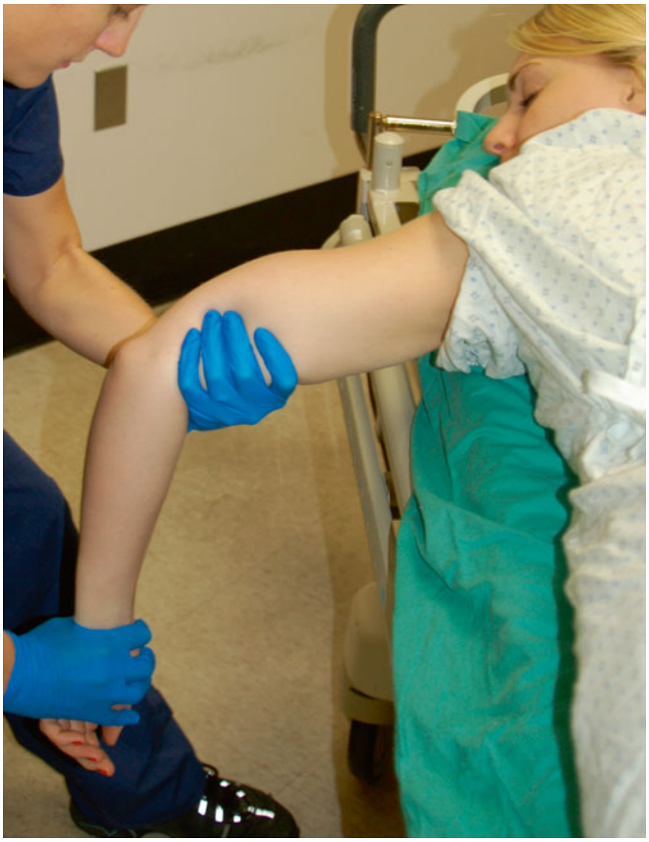

Fig. 101.4 Posterior method C

101.5 Procedure for Anterior Dislocations (Fig. 101.5)

1. Follow pre- and postprocedure steps as documented for the posterior dislocation.
2. Position the patient on a stretcher in the supine position.
3. With one hand, apply traction to the supinated distal forearm with the elbow extended, while an assistant applies countertraction with both hands around the distal humerus.
4. With the other hand apply downward and backward pressure over the proximal forearm just below the antecubital fossa.

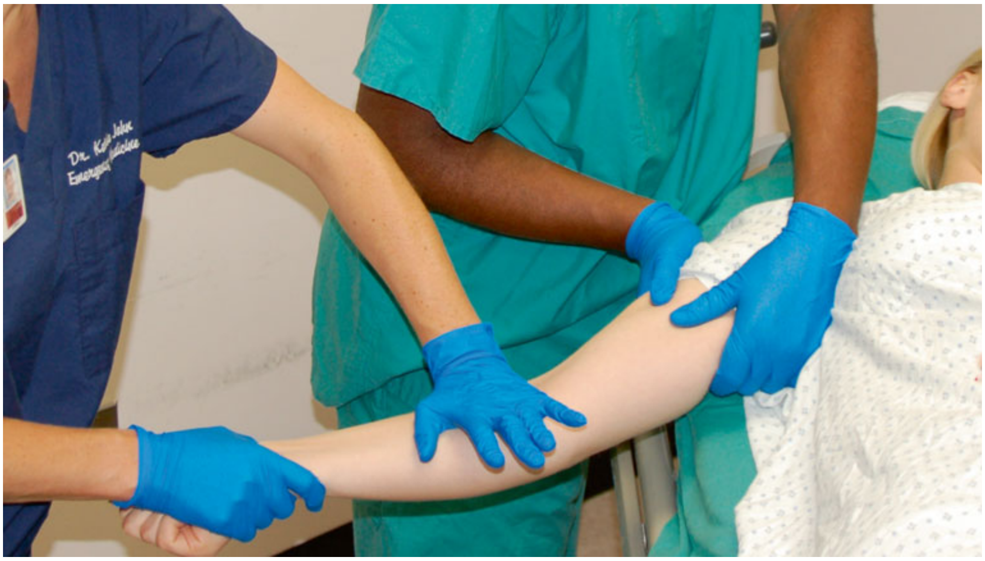

Fig. 101.5 Anterior elbow

101.6 Procedure for Radial Head Subluxations (See Also Chap. 127)

1. This procedure can normally be performed without any sedation or parenteral analgesia.
2. Position the patient, most commonly a child aged 1–3 years, facing forward on the caretaker's lap.
3. Hold the flexed elbow of the affected extremity placing your thumb firmly over the radial head.
4. With the other hand, take the child's hand and wrist, and in one continuous movement, hyperpronate and flex the forearm (Figs. 101.6 and 101.7).
5. Another method is to supinate and flex the forearm instead of hyperpronating it (Fig. 101.8).
6. Leave the room, encourage the caretaker to engage the child with distracting activities and reexamine the child in 10–20 min, at which stage, if reduction was successful, the child should be using the extremity normally again.
7. No postreduction radiograph or immobilization is required.

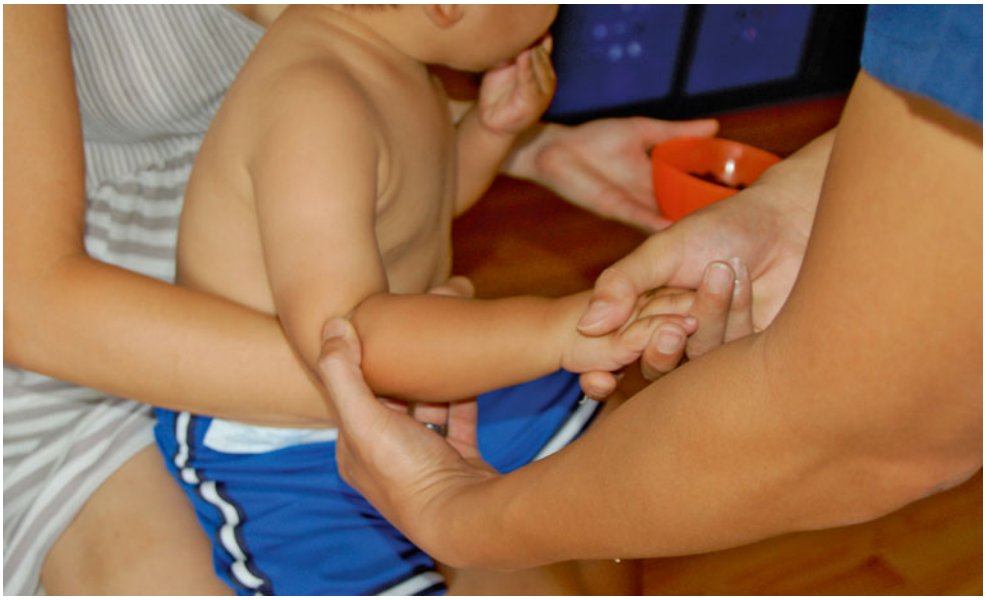

Fig. 101.6 Subluxation hyperpronated

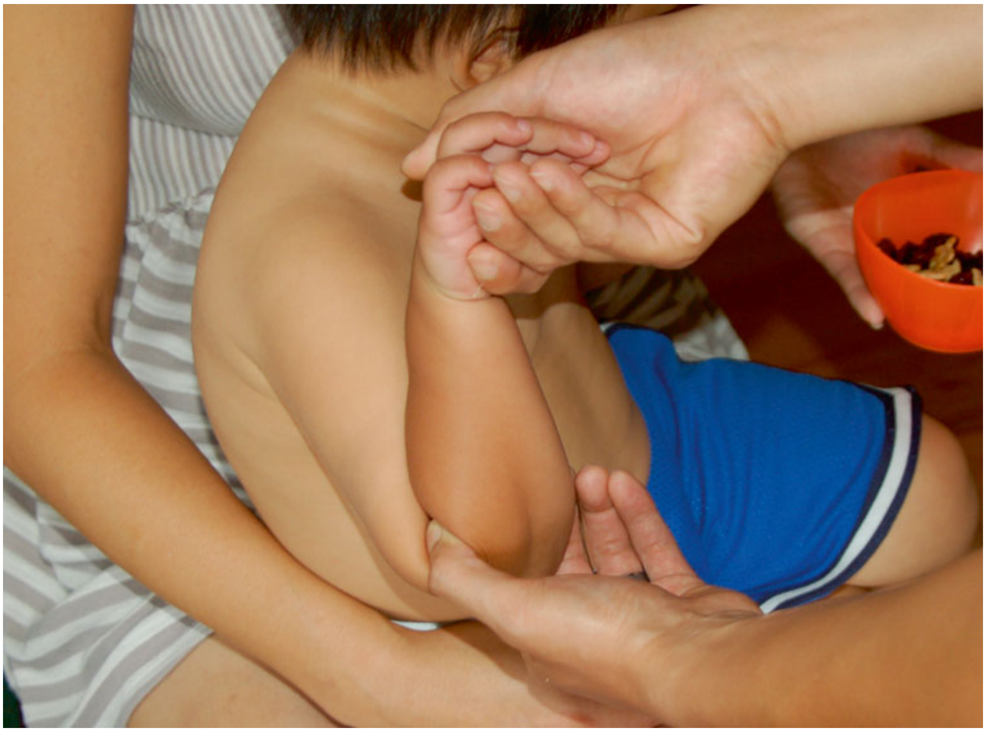

Fig. 101.7 Subluxation hyperpronated and flexed

Fig. 101.8 Subluxation supinated

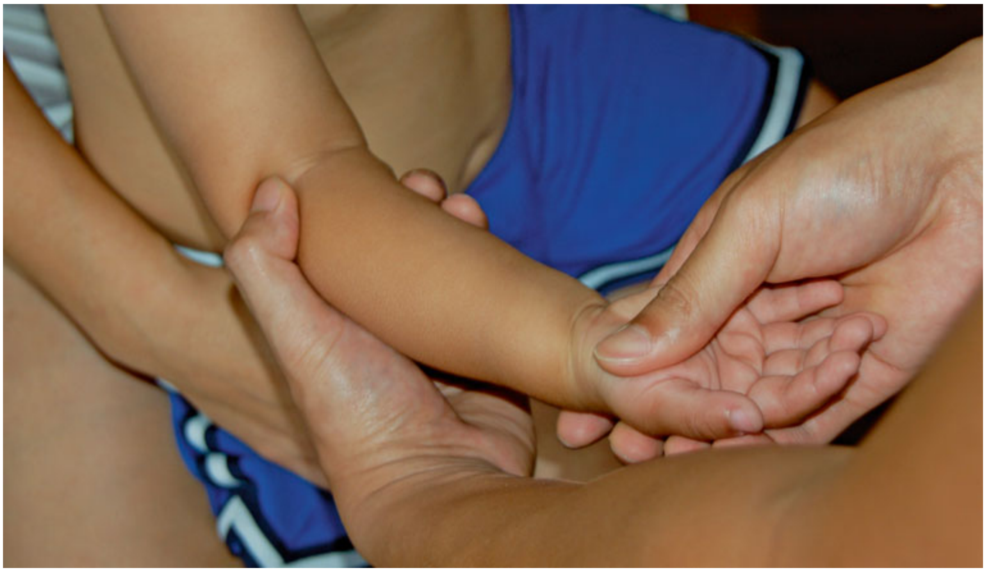

101.7 Complications

- Concomitant fractures
- Vascular injury, most commonly to the brachial artery
- Median nerve injury/entrapment
- Recurrent dislocation—rare

101.8 Pearls and Pitfalls

- Pearls
  - A true lateral radiograph is necessary to accurately detect and identify elbow fractures, dislocations, and soft tissue abnormalities (i.e., the fat pad sign). It is obtained with the patient's elbow in 90° flexion, in neutral rotation with the thumb pointing up and the arm and forearm resting on the radiograph cassette and the beam nearly perpendicular to the cassette. On a true lateral, the "hourglass" or "figure-of-eight" formation at the distal humerus should be clearly visible, and the rings of the capitellum and trochlea should be concentric.
  - During nursemaid elbow reduction, provide age appropriate distractions to divert the child's attention and minimize resistance.
- Pitfalls
  - On the pre- and postreduction radiographs, search for commonly associated fractures of the distal humerus, radial head, and coronoid process.
  - Inability to range the elbow after apparent reduction indicates possible trapped fracture fragments and the need for operative intervention.
  - Vascular or open injuries are common with anterior dislocations, and early orthopedic consultation is advised.
  - Ninety percent of simple elbow dislocations are posterior, and this injury is rarely associated with vascular injury. However, it does occur, and vascular evaluation after every reduction is good clinical practice.

Selected Reading

Jain K, Shashi Kumar Y, Mruthyunjaya, Ravishankar R, Nair AV. Posterior dislocation of elbow with brachial artery injury. J Emerg Trauma Shock. 2010;3:308.

Kuhn MA, Ross G. Acute elbow dislocations. Orthop Clin North Am. 2008;39:155–61.

McDonald J, Witelaw C, Goldsmith LJ. Radial head subluxation. Comparing two methods of reduction. Acad Emerg Med. 1999;6:715.

Sheps DM, Hildebrand KA, Boorman RS. Simple dislocations of the elbow: evaluation and treatment. Hand Clin. 2004;20:389–404.

Villarin Jr LA, Belk KE, Freid R. Emergency department evaluation and treatment of elbow and forearm injuries. Emerg Med Clin North Am. 1999;17:843–58.

Distal Interphalangeal Joint Reduction

102

Justin Chen and Muhammad Waseem

Distal interphalangeal (DIP) joint dislocation is rare. It occurs when an axial force is applied to the distal phalanx (Fig. 102.1).

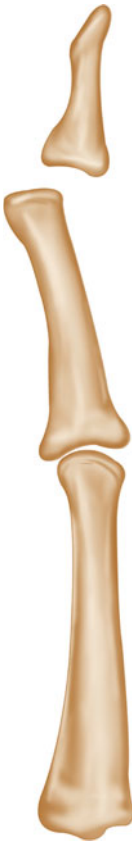

Fig. 102.1 Distal interphalangeal joint dislocation

102.1 Indications

- DIP joint reduction is performed to alleviate functional and anatomical derangements resulting from DIP joint dislocation, commonly dorsal, from axial compression.

102.2 Contraindications

- Absolute
  - Absence of radiographic confirmation (anteroposterior, true lateral, and oblique) of simple DIP joint dislocation, especially in pediatric cases
- Relative
  - Open joint dislocation, associated fracture, or entrapped volar plate
  - Digital neurovascular compromise

102.3 Materials and Medications

- Latex-free gloves
- Local anesthetic: 2 % lidocaine without epinephrine, 1.5 % mepivacaine, 0.5 % ropivacaine, or 0.5 % bupivacaine
- 25-gauge × 1.5-in. needle (can substitute with 27 or 30 gauge)
- Small plastic syringe, 10 mL
- Padded, malleable, aluminum digital splint

J. Chen, MD, MSc
Department of Emergency Medicine, North Shore University Hospital, Manhasset, NY, USA
e-mail: chen.justin@gmail.com

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_102

102.4 Procedure

1. Place the patient in the seated position with the arms at rest on a bedside table or supported by an assistant.
2. Pronate the patient's hand, remove rings if present, and rest on a flat surface.
3. Insert a 25-gauge needle at the dorsolateral aspect of the base of the finger to form a wheal to reduce patient discomfort.
4. Advance the needle and direct anteriorly toward the phalangeal base.
5. Inject 0.5–1 mL of local anesthetic as the needle is withdrawn 1–2 mm from the point of bone contact.
6. Inject an additional 1 mL of local anesthetic continuously as the needle is withdrawn.
7. The injection should never render the tissue tense nor be circumferential.
8. Hyperextend the DIP joint while applying longitudinal traction, followed by immediate joint flexion at the base of the distal phalanx.
9. Place finger(s) in an aluminum digital dorsal splint in slight flexion for 2 weeks (Fig. 102.2).
10. Postreduction radiograph is recommended for confirmation.

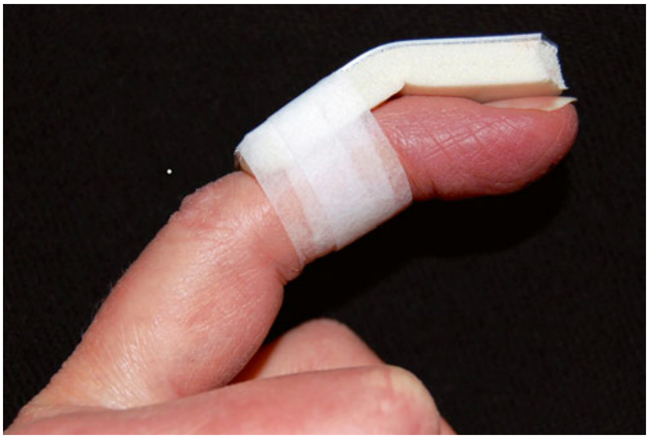

Fig. 102.2 Padded aluminum splint applied to block the DIP joint in flexion but allow further flexion, which encourages active flexion of that joint when the PIP joint flexes (Reproduced with permission from: HandLab Clinical Pearls Feb 2011, No 12. www.handlab.com)

102.5 Complications

- Irreducible dislocations
- Stiffness
- Recurrent dislocation
- Extensor lag in joints with residual subluxation
- Associated with dorsal joint prominences, swan-neck/boutonnière deformity, and degenerative arthritis

102.6 Pearls and Pitfalls

- Pearls
  - Lidocaine without epinephrine is preferred owing to the risk of vasoconstriction of the digital vessels with epinephrine.
  - Regardless of the mechanism of trauma, all joints (DIP, proximal interphalangeal, metacarpophalangeal) should be assessed for instability.
  - Joint dislocations involving volar plate entrapment may require surgical repair (open reduction internal fixation) for successful reduction.
- Pitfalls
  - Irreducible DIP joint dislocations may be due to entrapment of an avulsion fracture, the profundus tendon, or the volar plate.

Selected Reading

Calfee RP, Sommerkamp TG. Fracture-dislocation about the finger joints [review]. J Hand Surg Am. 2009;34:1140–7.

Knoop KJ. Atlas of emergency medicine. 3rd ed. New York: McGraw-Hill Professional; 2010.

Simon RR, Sherman SC, Sharieff GQ. Emergency orthopedics. 6th ed. New York: McGraw-Hill Medical; 2011.

Stone CK, Humphries RL. Current diagnosis & treatment emergency medicine. 6th ed. New York: McGraw-Hill; 2008.

Tintinalli JE, Stapczynski JS, Ma OJ, Cline D, Cydulka R, Meckler G, editors. Tintinalli's emergency medicine: a comprehensive study guide. 7th ed. New York: McGraw-Hill; 2012.

Hip Dislocation Reduction

103

Katrina John, Jeffrey Kile, and Amish Aghera

103.1 Indications

Displacement of the femoral head in relation to the acetabulum without concomitant femoral neck, head, or acetabulum fractures:

- Posterior hip dislocations make up 80–90 % of cases.
- Anterior hip dislocations make up 10–15 % of cases. These are classified into obturator, pubic, iliac, central, or inferior types. Central dislocations are associated with comminuted acetabulum fractures, and inferior dislocations are a rare occurrence normally occurring in children younger than 7 years of age.
- Prosthetic hip dislocations

103.2 Contraindications

- Absolute
  - Femoral neck fracture: attempted reduction may increase the displacement of the fracture and increase the probability of avascular necrosis.
- Relative
  - Fractures in other parts of the affected lower extremity: these may limit the pressure that can be applied necessary for traction during reduction.

103.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Sheet or belt to fix the pelvis to the stretcher
- Knee immobilizer
- Abduction pillow

103.4 Procedure

1. Check the neurovascular status of the affected extremity.
2. Obtain anteroposterior (AP) views of the pelvis and lateral views of the hip.
3. Ensure adequate parenteral sedation and analgesia.
4. Decide upon a technique, as detailed later, and position the patient accordingly.
5. Once the hip has been successfully reduced, test the joint for stability by moving it gently thought its range of motion.
6. Place a knee immobilizer and an abduction pillow between the knees.
7. Check the neurovascular status.
8. Obtain repeat AP films of the pelvis.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_103

103.4.1 Stimson Maneuver

1. Place the patient prone on the stretcher with the affected extremity hanging over the edge and the hip flexed to 90°.
2. Flex the knee and the foot to 90°.
3. Apply downward pressure to the area just distal to the popliteal fossa with a hand (Fig. 103.1) or knee (Fig. 103.2) while using the opposite hand to internally and externally rotate the hip at the ankle.
4. Have an assistant simultaneously manipulate the displaced femoral head into position with both hands, applying downward pressure over the affected buttock (Fig. 103.3).

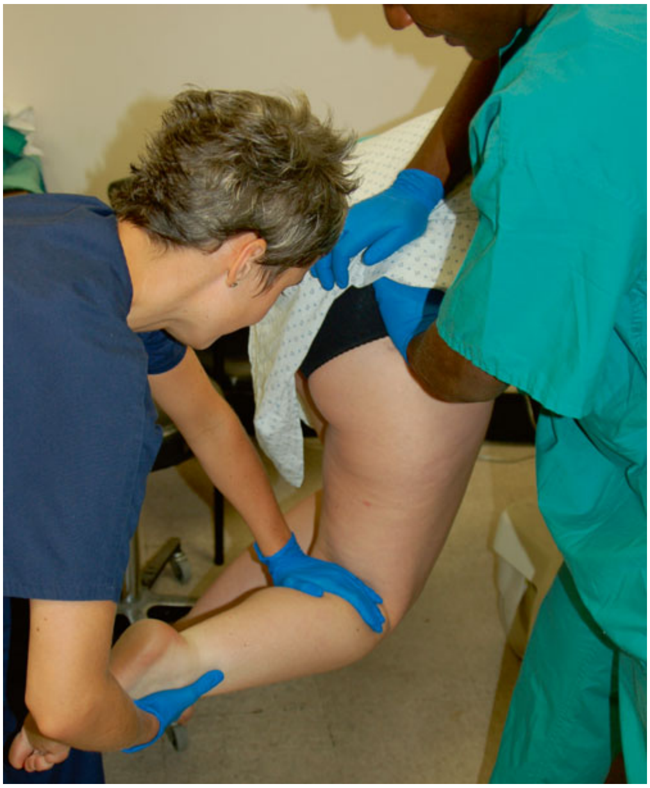

Fig. 103.1 Stimson maneuver with hand

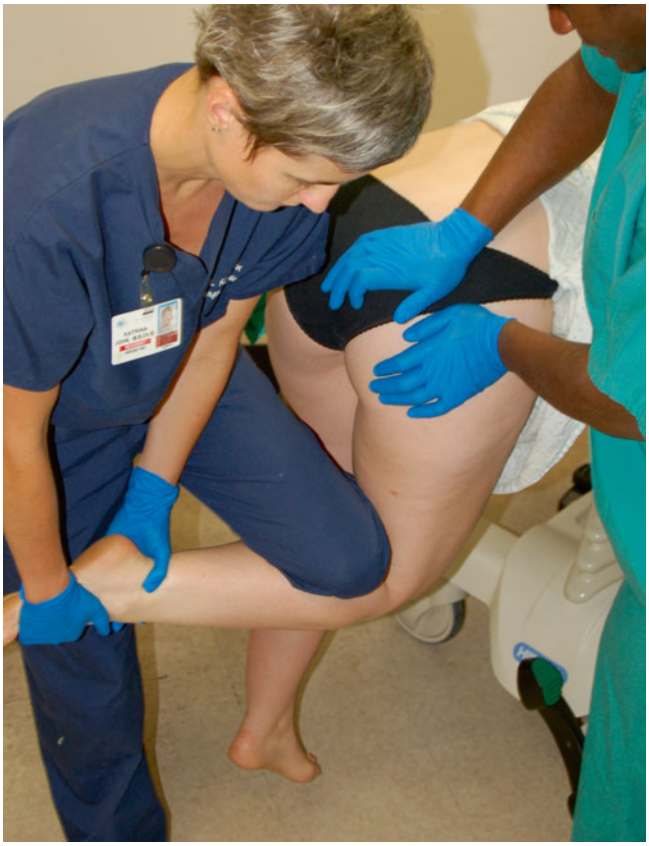

Fig. 103.2 Stimson maneuver with knee

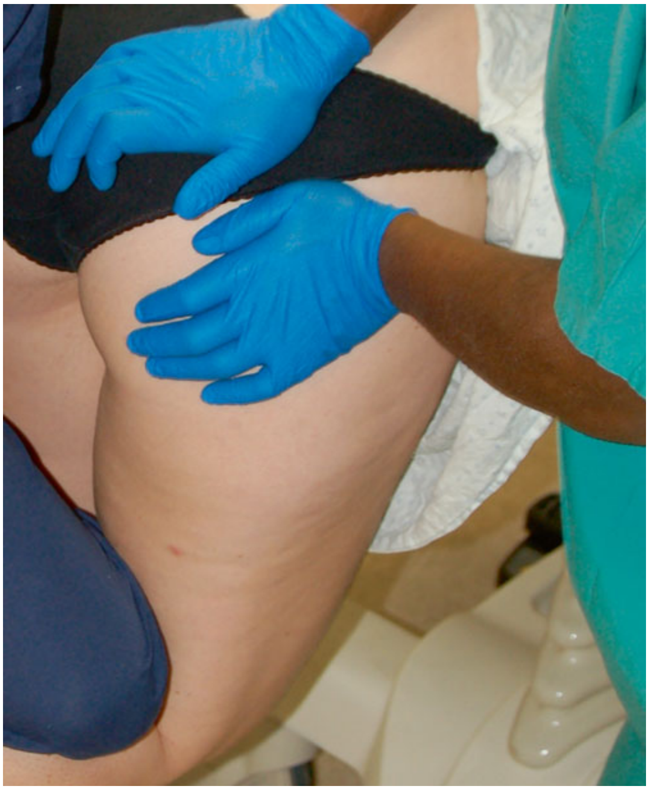

Fig. 103.3 Manipulation of the femoral head

103.4.2 Allis Maneuver

1. Position the patient supine on the stretcher.
2. The operator should stand on the stretcher to achieve maximum leverage or have the patient on a backboard on the ground.
3. Have an assistant apply downward pressure to both iliac crests.
4. Apply constant, gentle upward traction in line with the deformity while maneuvering the hip to 90° flexion and through internal and external rotation (Fig. 103.4).
5. Have a second assistant provide lateral traction to the midthigh.
6. Once the femoral head has cleared the outer lip of the acetabulum, continue traction while keeping the hip in external rotation and gently abducting and extending the hip (Fig. 103.5).

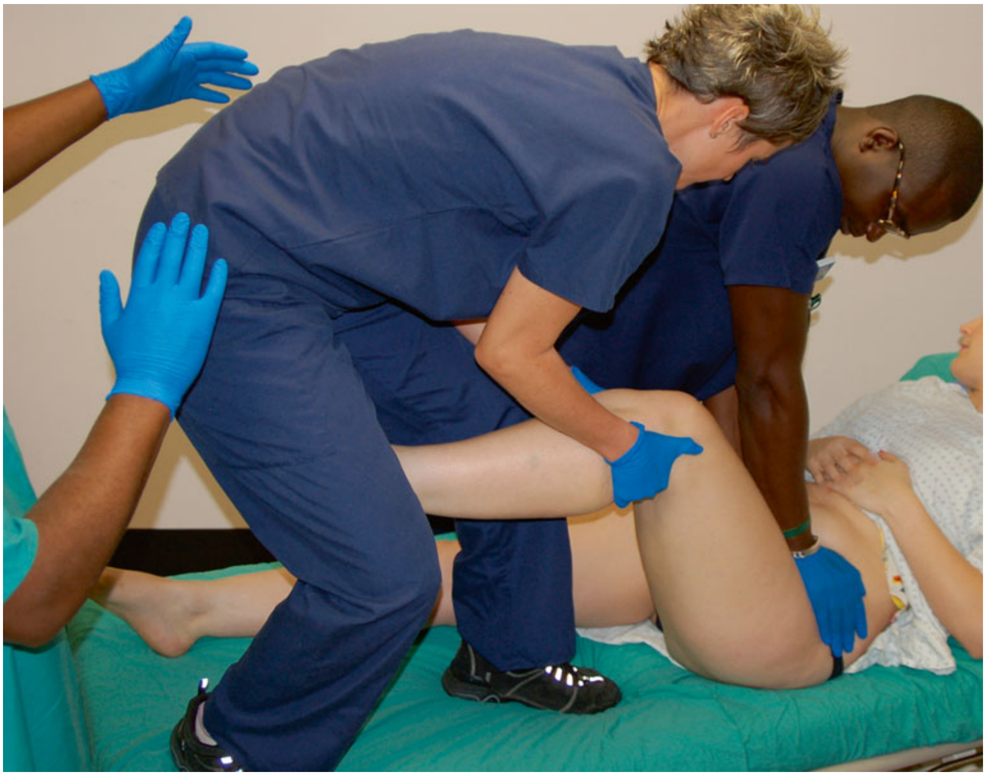

Fig. 103.4 Allis flexion

Fig. 103.5 Allis extension

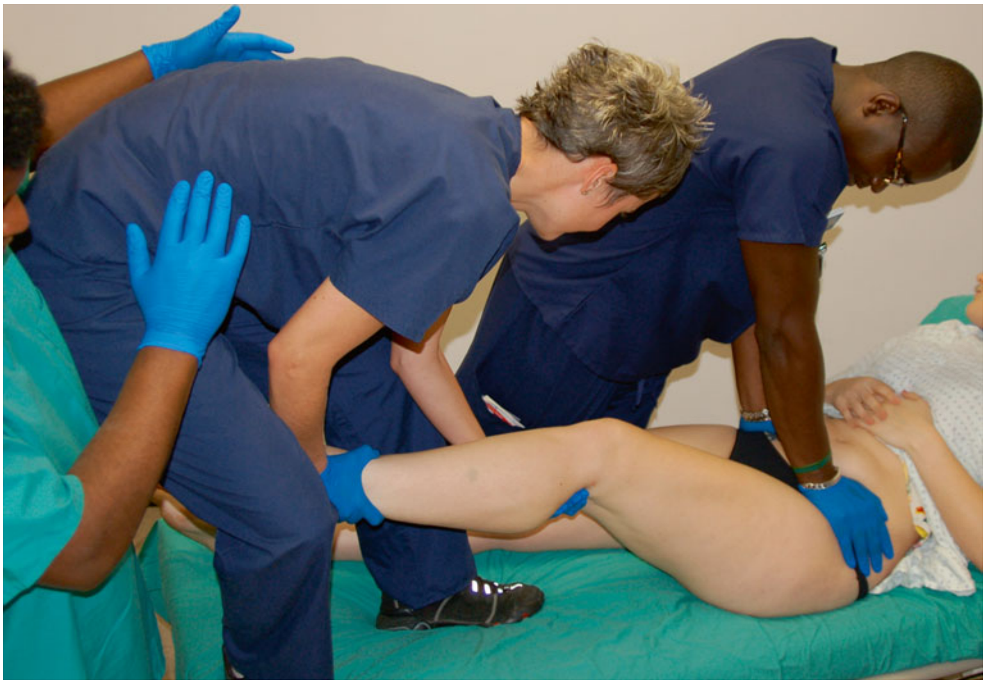

103.4.3 Whistler Technique

1. Position the patient supine on the stretcher with the knee and hip flexed to 45°.
2. Have an assistant stabilize the pelvis with downward pressure on both iliac crests.
3. Stand on the side of the affected extremity and place one arm under the knee, resting the hand on the flexed knee of the unaffected extremity.
4. Secure the ankle of the affected extremity with the other hand and elevate the shoulder of the opposite arm, providing upward traction at the distal thigh and a strong fulcrum to reduce the dislocation (Fig. 103.6).
5. Internal and external rotation can be achieved with the opposite hand at the ipsilateral ankle.

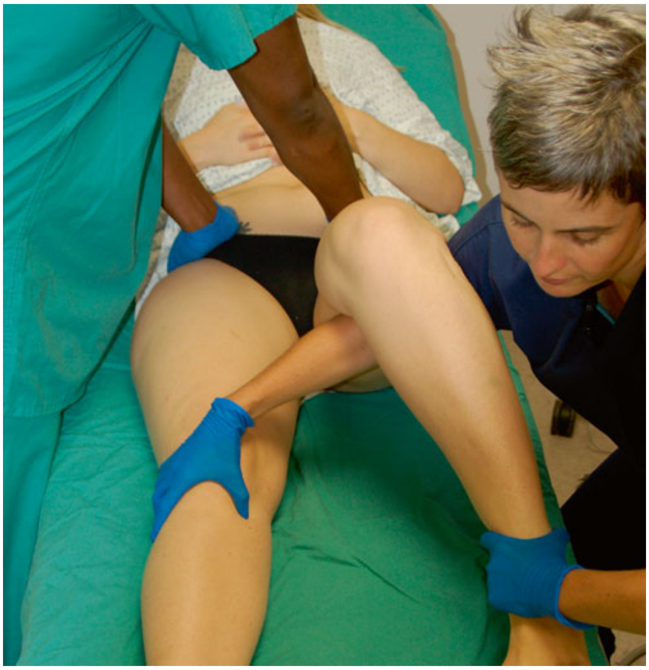

Fig. 103.6 Whistler technique

103.4.4 Captain Morgan Technique

1. Position the patient supine on the stretcher with the knee and hip flexed to 90°.
2. Stabilize and fix the pelvis with a sheet tied securely over the pelvis and under the stretcher.
3. Standing on the side of the affected extremity, the operator's foot should be resting perpendicular on the stretcher with the knee placed under the patient's knee.
4. With the opposite hand, apply downward pressure to the ankle and provide a sustained upward force to the patient's thigh by elevation of the knee through plantar flexion of the toes and upward pressure of the other hand placed behind the patient's knee.
5. Internal and external rotation can be applied simultaneously if necessary by gently twisting the ankle (Fig. 103.7).

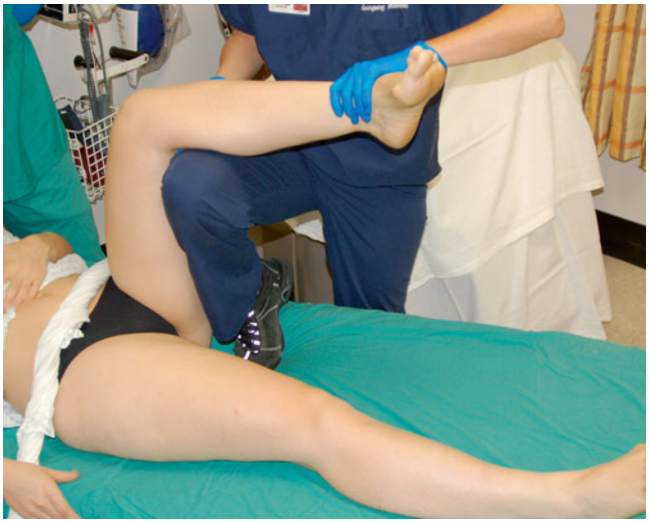

Fig. 103.7 Captain Morgan technique

103.5 Complications

- Sciatic nerve injury
- Avascular necrosis of the femoral head due to delay in adequate reduction
- Inability to perform reduction due to occult fractures and fracture fragments, incarceration of the joint capsule, or associated tendons
- Unstable or irreducible dislocations
- Traumatic arthritis and joint instability

103.6 Pearls and Pitfalls

- Pearls
  - On AP radiograph, posterior dislocations can be more easily detected by the presence of a smaller femoral head compared with the unaffected side and poor visualization of the lesser trochanter.
  - On AP radiographs, anterior dislocations can be detected by a larger femoral head and a clear lesser trochanter seen in profile alongside the femoral shaft.
  - Pay close attention to the femoral vessels and the sciatic nerve. Injury to the sciatic nerve most commonly affects the common peroneal branch, therefore causing weakness in great toe extension and foot dorsiflexion. Sensation may also be reduced over the dorsum of the foot.
  - Check the femoral head is intact and clearly in the acetabulum and for intact Shenton lines, symmetrical intra-articular spaces, and clear outlines of the lesser trochanters.
  - For any of the techniques requiring stabilization of the pelvis, an alternative is to fix it to the stretcher using a sheet or belt.
  - To overcome the powerful muscles that oppose successful reduction, it is important to provide adequate muscle relaxation and steady, prolonged traction.
  - An assistant should stand on the floor behind to support the operator if standing on a stretcher.
- Pitfalls
  - Owing to the force necessary to dislocate a native hip, this injury should serve as a red flag to the physician to consider other potentially life- or limb-threatening occult injuries.
  - Hip dislocation is a true orthopedic emergency and must be treated without delay. Delay in reduction, especially greater than 6 h, results in increased incidence of avascular necrosis of the femoral head and sciatic nerve injury.
  - Review imaging carefully because associated fractures of the femoral head, neck, and acetabulum are often present.
  - It is recommended that anterior dislocations be reduced by orthopedic surgeons under general anesthetic in the operating room. These are often more complicated and difficult to reduce, and failure at closed reduction in the operating room can be followed by an open procedure.
  - Multiple attempts at reduction should not be performed in the emergency department because these are unlikely to be successful and will only delay definitive management and lead to an increased risk of complications.

Selected Reading

Hendey GW, Avila A. The captain Morgan technique for the reduction of the dislocated hip. Ann Emerg Med. 2011;58:536–40.

Newton EJ, Love J. Emergency department management of selected orthopedic injuries. Emerg Med Clin North Am. 2007; 25:763–93.

Nordt WE. Maneuvers for reducing dislocated hips. Clin Orthop Relat Res. 1999;360:160–4.

Rupp JD, Schneider LW. Injuries to the hip joint in frontal motor-vehicle crashes: biomechanical and real-world perspectives. Orthop Clin North Am. 2004;35:493–504.

Walden PD, Hamer JR. Whistler technique used to reduce traumatic dislocation of the hip in the emergency department setting. J Emerg Med. 1999;17:441–4.

Knee Dislocation Reduction

104

Katrina John, Jeffrey Kile, and Amish Aghera

104.1 Indications

- Dislocation of the knee/fibular head/patella

104.2 Contraindications

- Absolute
  - None
- Relative
  - Immediate availability of orthopedic consultation

104.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Knee immobilizer or splinting materials

104.4 Procedure

104.4.1 Knee (Femur/Tibia) Dislocation Reduction

1. Assess neurovascular function.
2. Pretreat the patient with sedation or analgesia as appropriate.
3. Position the patient supine with the affected leg fully extended.
4. Instruct an assistant to stand near the patient's hip and, facing the patient's affected knee, grasp the distal femur firmly with both hands to fix it in place.
5. Stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction.
   - Longitudinal traction-countertraction alone, as described previously, will usually reduce the dislocation. If reduction does not occur, proceed with the following steps.
6. While applying straight traction in a distal direction to the tibia with the dominant hand, with the nondominant hand:
   (a) Anterior dislocation: push the proximal tibia in a posterior direction (Fig. 104.1)
   (b) Posterior dislocation: lift the proximal tibia in an anterior direction (Fig. 104.2)
   (c) Lateral dislocation: push the proximal tibia in a medial direction (Fig. 104.3)
   (d) Medial dislocation: push the proximal tibia in a lateral direction (Fig. 104.4)
   (e) Rotary dislocation: rotate the proximal tibia into proper linear alignment with the femoral condyles (Fig. 104.5)
      - Reduction may be facilitated by the use of two assistants rather than just one. The second assistant grasps the distal tibia and applies straight traction in a distal direction, freeing the operator to manipulate the proximal tibia as described previously using both hands.
7. After reduction, reassess neurovascular function and, if available, obtain angiography.
8. Immobilize the knee in 15° of flexion in a knee immobilizer or long-leg posterior splint.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_104

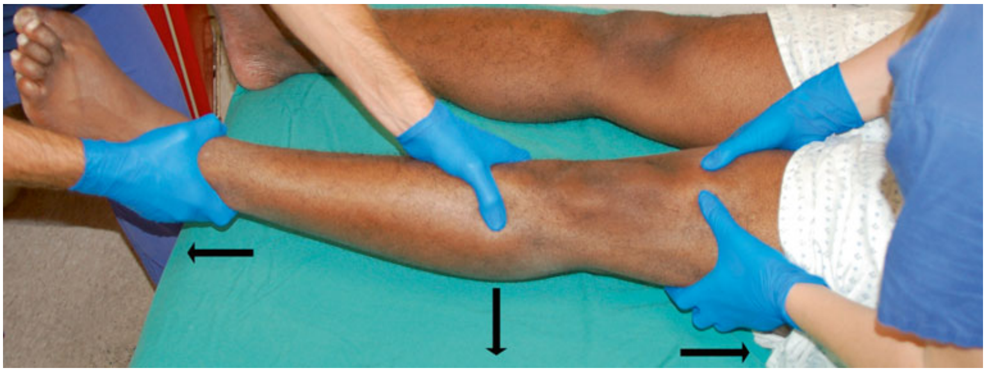

Fig. 104.1 Anterior dislocation of the knee: the proximal tibia is pushed in a posterior direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

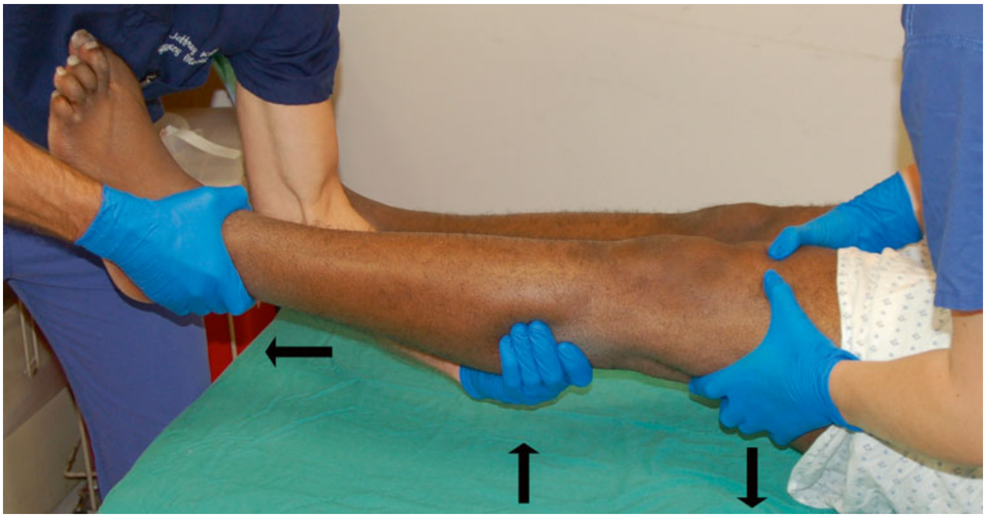

Fig. 104.2 Posterior dislocation of the knee: the proximal tibia is pushed in an anterior direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

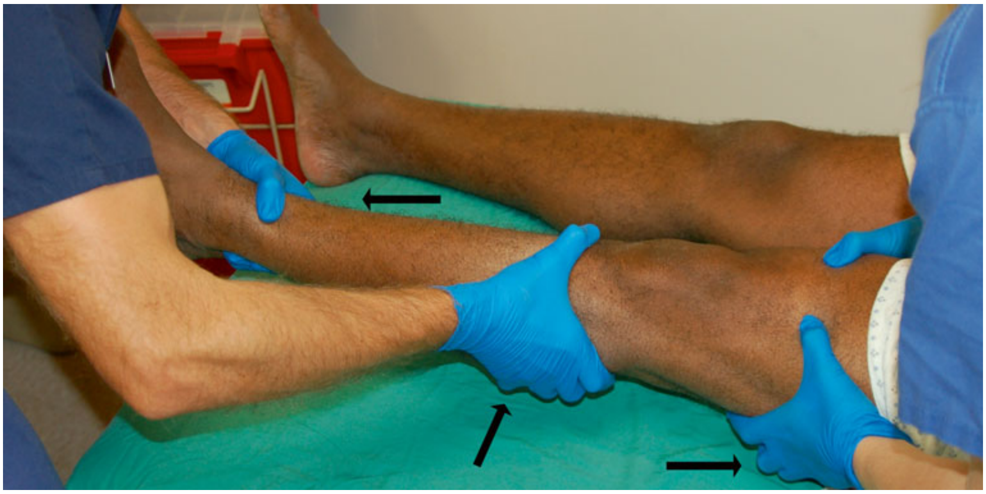

Fig. 104.3 Lateral dislocation of the knee: the proximal tibia is pushed in a medial direction. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

Fig. 104.4 Medial dislocation of the knee: the proximal tibia is pushed in a lateral direction. The *arrows* indicate the direction in which force should be applied by the operator during reduvction of dislocation

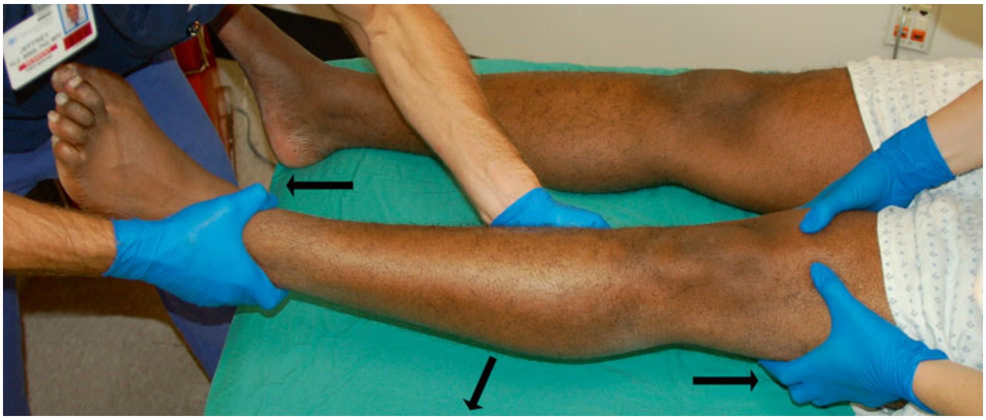

Fig. 104.5 Rotary dislocation of the knee: the proximal tibia is rotated into proper alignment with the femoral condyles. The *arrows* indicate the direction in which force should be applied by the operator during reduction of dislocation

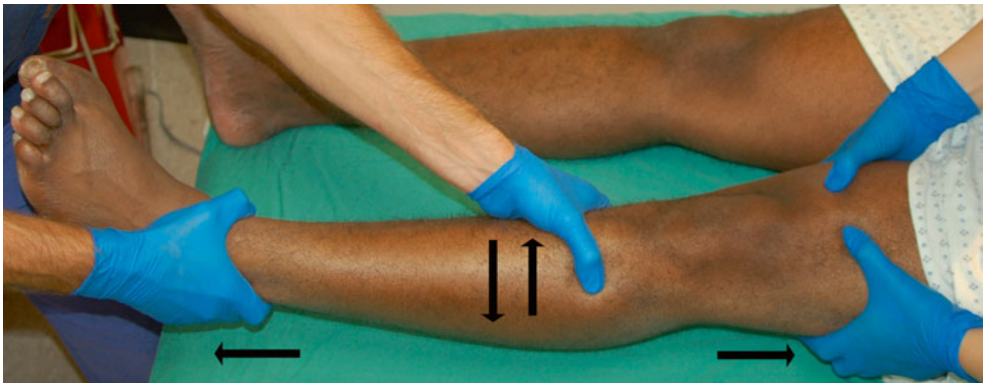

104.4.2 Fibular Head Dislocation Reduction

1. Assess neurovascular function.
2. Pretreat the patient with sedation or analgesia as appropriate.
3. Position the patient supine.
4. Flex the knee to 90° to relax the biceps femoris tendon.
5. Instruct an assistant to stand near the patient's hip and, facing the patient's affected knee, grasp the distal femur firmly with both hands to fix it in place.
6. Stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction with the dominant hand and with the nondominant hand.
   (a) Anterior dislocation: push the fibular head in a posterior direction (Fig. 104.6)
   (b) Posterior dislocation: push the fibular head in an anterior direction (Fig. 104.7)
      - Reduction may be facilitated by the use of two assistants rather than just one. If a second assistant is available, instruct the second assistant to stand near the patient's foot and, facing the patient's affected knee, grasp the distal tibia and apply straight traction in a distal direction. This enables the operator to grasp and move the proximal fibula as described previously using both hands.
      - Reduction is often signified by a palpable and audible click as the fibula snaps back into position.
7. After reduction, reassess neurovascular function and, if available, obtain angiography.
   - After reduction, patients should receive orthopedic referral, avoid weight-bearing for the first 2 weeks, and then gradually increase weight-bearing over the next 6 weeks.
   - Typically, immobilization is not required following reduction of an i.solated fibular head dislocation.

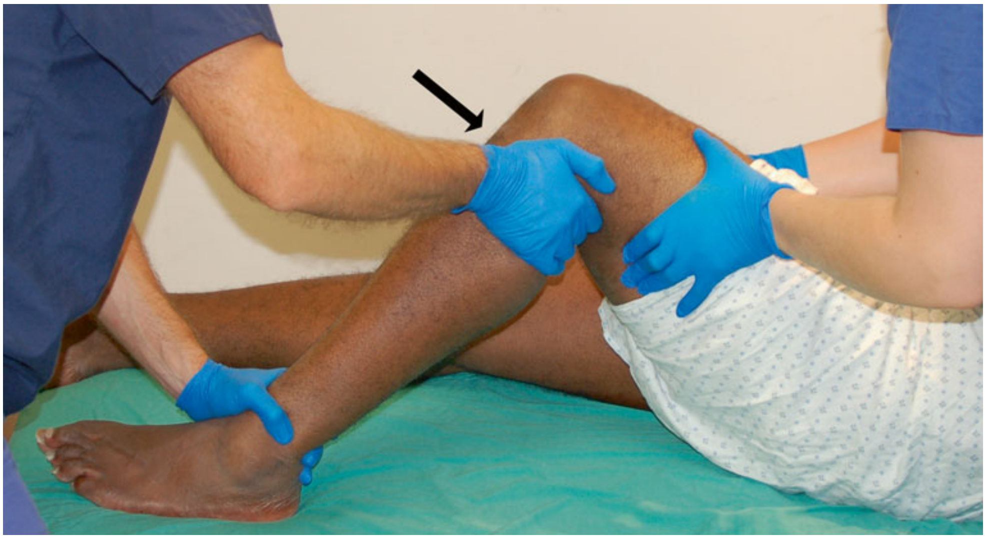

Fig. 104.6 Anterior dislocation of the fibular head: the fibular head is pushed in a posterior direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

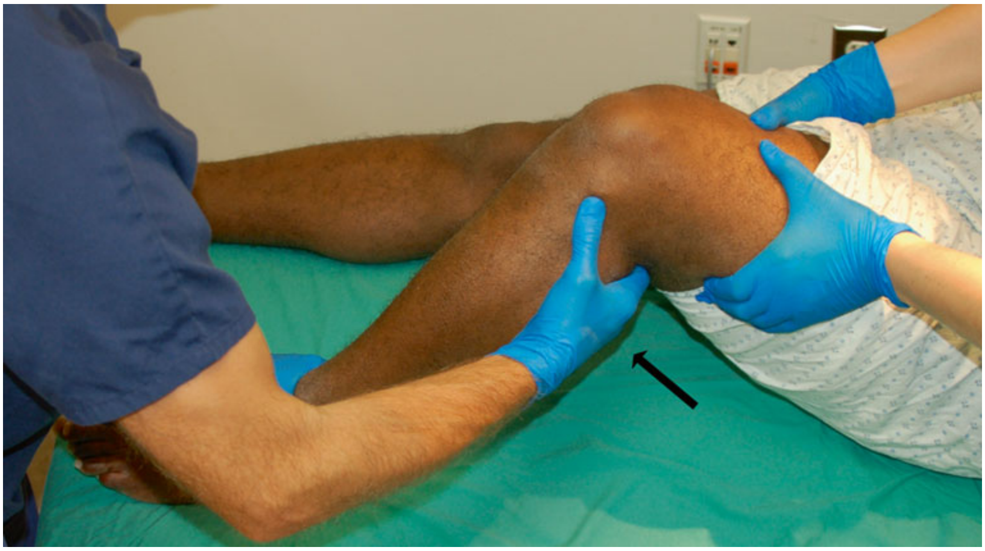

Fig. 104.7 Posterior dislocation of the fibular head: the fibular head is pushed in an anterior direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

104.4.3 Lateral Patellar Dislocation Reduction

1. Pretreat the patient with sedation or analgesia as appropriate.
2. Stand at the side of the affected knee and, facing the knee, grasp the distal tibia and slowly extend the knee with one hand, and with the other hand simultaneously apply gentle pressure to the patella in a medial direction.

- The lateral edge of the patella may be lifted slightly to facilitate its travel over the femoral condyle during reduction (Fig. 104.8).
- After reduction, the knee should be immobilized in full extension in a knee immobilizer or long-leg posterior splint, and the patient should receive orthopedic referral, avoid weight-bearing for the first 2 weeks, and then gradually increase weight-bearing over the next 6 weeks.

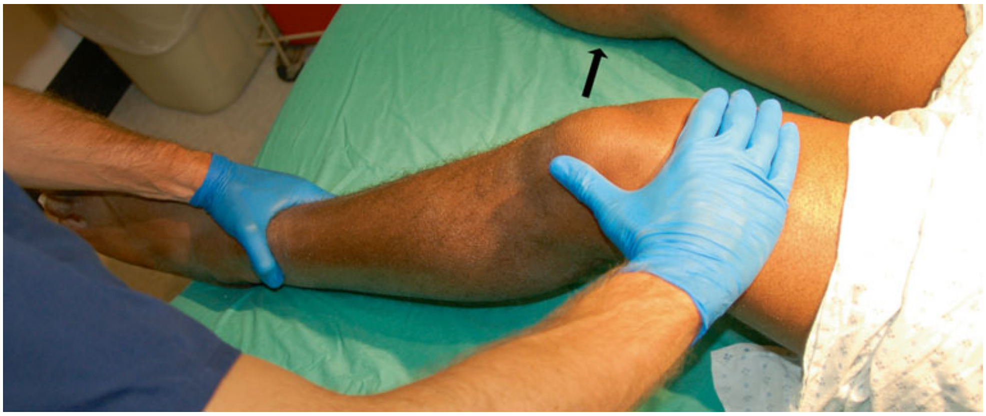

Fig. 104.8 Lateral dislocation of the patella: the patella is pushed in a medial direction. The *arrow* indicate the direction in which force should be applied by the operator during reduction of dislocation

104.5 Complications

104.5.1 Knee (Femur/Tibia) Dislocations

- Distal ischemia (even requiring amputation)
- Degenerative arthritis
- Joint instability due to ligamentous injury

104.5.2 Fibular Head Dislocations

- Peroneal nerve injury
- Fibular head instability/subluxation
- Degenerative arthritis

104.5.3 Patellar Dislocations

- Failure of reduction
- Degenerative arthritis
- Recurrent dislocation/subluxation

104.6 Pearls and Pitfalls

104.6.1 Knee (Femur/Tibia) Dislocations

- Pearls
  - Dislocations of the knee are described in terms of the tibia's position in relation to the femur.
  - All knee dislocations require orthopedic evaluation at the earliest possible opportunity.
  - Owing to the frequency of associated popliteal artery and peroneal nerve injury, a neurovascular examination should be performed before and after any attempts at reduction or manipulation of the knee.
  - Dislocations of the knee should be reduced as soon as possible, particularly if distal neurovascular compromise exists.
  - Operative ligamentous repair is often required approximately 2 weeks postreduction (once acute swelling has resolved) to achieve the maximum functional recovery.
- Pitfalls
  - If the knee hyperextends more than 30° when the horizontal leg is lifted by the foot, the knee is considered severely unstable. This is likely due to a previous dislocation, and thus, the knee should be evaluated for the neurovascular complications of dislocation.
  - Because the joint capsule is commonly disrupted during knee dislocation, synovial fluid may diffuse into the surrounding tissue, such that an effusion is not always present.
  - A posterolateral dislocation may be irreducible because the medial femoral condyle traps the medial capsule within the joint.

104.6.2 Fibular Head Dislocations

- Pearls
  - Fibular head dislocations are usually anterolateral, but these do not result in neurovascular compromise.
  - A knee joint effusion is usually not seen in a fibular head dislocation because the tibiofibular ligaments are contained within a separate synovium.
  - Anterior dislocations typically result from a fall on the flexed, adducted leg, often combined with ankle inversion.
  - Flexion of the knee relaxes the fibular collateral ligament, reducing the stability of the tibiofibular joint.
  - Superior dislocation is accompanied by interosseus membrane damage and proximal displacement of the lateral malleolus.
- Pitfalls
  - Posterior fibular head dislocations usually result from direct trauma to the flexed knee and may be accompanied by peroneal nerve injury.

104.6.3 Patellar Dislocations

- Pearls
  - Patellar dislocation occurs most frequently among adolescents.
  - Patellar dislocation typically occurs in the setting of external rotation combined with a strong valgus force and quadriceps contraction.
  - Patellar dislocations are described in terms of the patellar relationship to the normal knee joint.
  - The most common patellar dislocations are lateral.
  - If a spontaneous reduction has occurred, a knee effusion and tenderness along the medial aspect of the patella are likely to be present on examination, and the patellar apprehension test will be positive.
  - To perform the patellar apprehension test, flex the knee to 30° and push the patella laterally. If the patient senses an impending redislocation, the test is considered positive.

  - Isolated lateral patellar dislocations do not usually require hospitalization, but orthopedic follow-up is recommended owing to the likelihood of persistent instability.
  - Intracondylar and superior dislocations require surgical reduction.
  - Patients with an isolated patellar dislocation typically present with the knee in 20–30° of flexion and the patella displaced laterally.
- Pitfalls
  - Dislocations tend to be recurrent, particularly in patients with patellofemoral anatomical abnormalities.

Selected Reading

Martinez D, Sweatman K, Thompson EC. Popliteal artery injury associated with knee dislocations. Am Surg. 2001;67:165–7.

Peskun CJ, Levy BA, Fanelli GC, et al. Diagnosis and management of knee dislocations. Phys Sportsmed. 2010;38:101–11.

Rihn JA, Groff YJ, Harner CD, Cha PS. The acutely dislocated knee: evaluation and management. J Am Acad Orthop Surg. 2004;12:334–46.

Roberts DM, Stallard TC. Emergency department evaluation and treatment of knee and leg injuries. Emerg Med Clin North Am. 2000;18:67–84. v–vi.

Wascher DC, Dvirnak PC, DeCoster TA. Knee dislocation: initial assessment and implications for treatment. J Orthop Trauma. 1997;11:525–9.

Ankle Dislocation Reduction

105

Katrina John, Jeffrey Kile, and Amish Aghera

105.1 Indications

- Dislocation of the ankle joint. This is defined by the articulation of the talus with the mortise that is formed by the distal tibia and fibula. Dislocations can be posterior, anterior, superior, or lateral and are classified by the position of the talus in relation to the tibial mortise.

105.2 Contraindications

- Relative
  - Open dislocations where there is no evidence of acute neurovascular compromise are better managed definitively in the operating room to avoid further contamination.

105.3 Materials and Medications

- Parenteral sedation and analgesia medications
- Local anesthetic for local and intra-articular anesthesia
- Splinting material
- Stockinette
- Padding
- Elastic bandage
- Tape
- Sheet

105.4 Procedure

1. Check the neurovascular status of the affected foot and ankle.
2. If there is no evidence of critical neurovascular compromise, obtain a lateral and an anteroposterior radiograph of the affected ankle.
3. Ensure adequate parenteral sedation and analgesia to maximize success and limit pain and suffering.
4. Position the patient on a stretcher with the knee flexed at 90° over a folded pillow or rolled-up sheet or with the lower leg and knee hanging over the edge of the stretcher.

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_105

105.4.1 Posterior Dislocations

1. Hold the heel in one hand and pull with longitudinal traction.
2. With the other hand, hold the top of the foot and gently plantarflex it downward, while an assistant provides countertraction at the back of the midcalf (Fig. 105.1).
3. Continue longitudinal traction at the heel and countertraction at the calf.
4. Dorsiflex the foot while another assistant applies downward pressure to the distal anterior leg (Fig. 105.2).
5. Examine foot for restoration of normal anatomy and for any new lacerations or defects to the skin.
6. Recheck neurovascular integrity.
7. Place the leg in a sugar-tong splint with the foot at 90°.
8. Recheck neurovascular integrity.

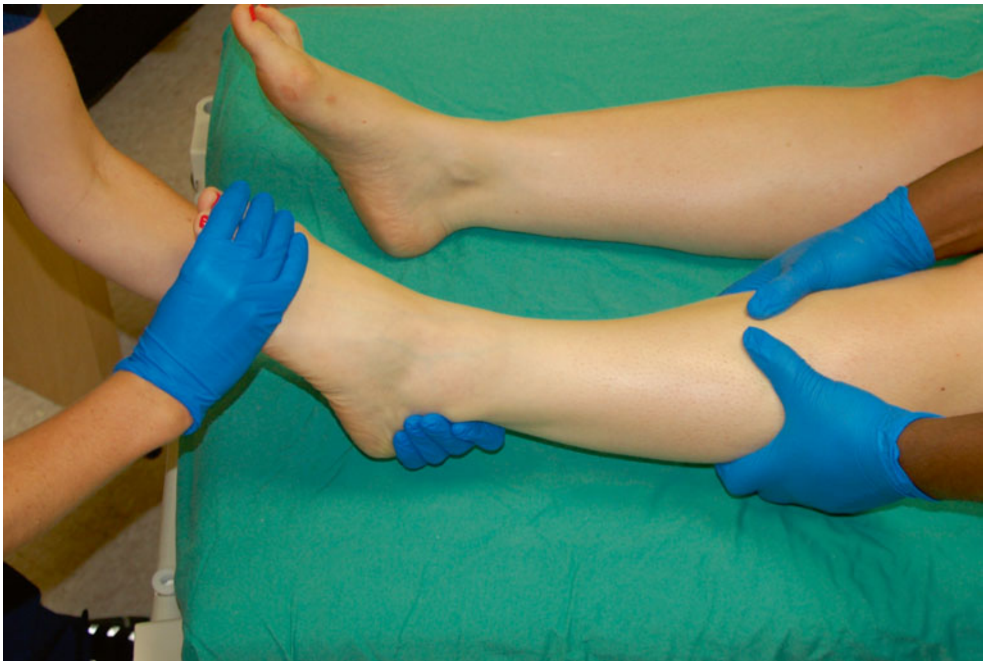

Fig. 105.1 Plantarflexion with longitudinal heel traction

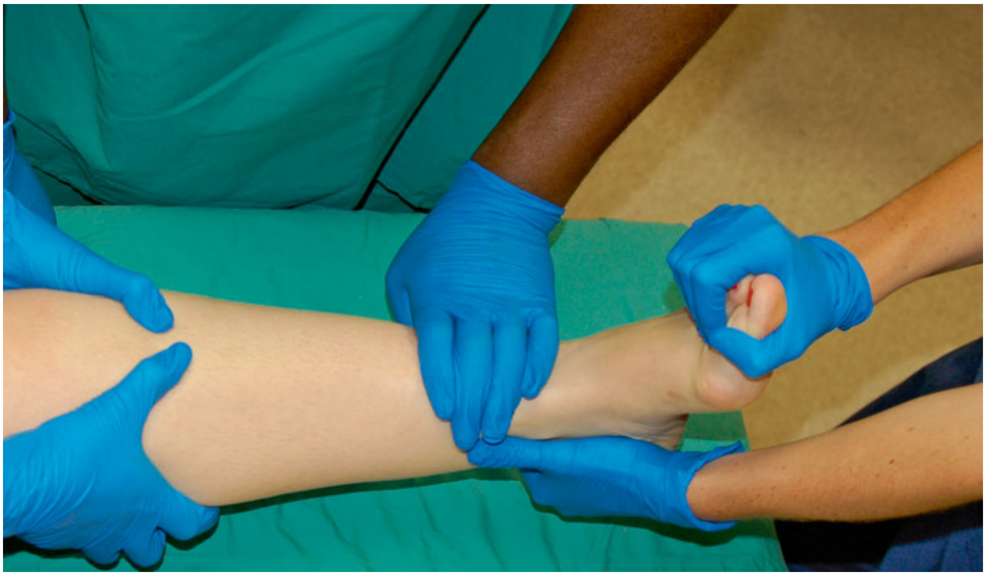

Fig. 105.2 Dorsiflexion with longitudinal heel traction

105.4.2 Anterior Dislocations

1. Hold the heel in one hand and pull with longitudinal traction.
2. With the other hand, hold the top of the foot and dorsiflex, while an assistant provides countertraction at the back of the midcalf (Fig. 105.3).
3. Continue longitudinal traction at the heel and countertraction at the calf.
4. Keeping the foot at 90° to the leg, hold the foot firmly and push the foot downward toward the floor while another assistant applies upward pressure to the distal posterior leg (Fig. 105.4).
5. Examine the foot for restoration of normal anatomy and for any new lacerations or defects to the skin.
6. Recheck neurovascular integrity.
7. Place the leg in a sugar-tong splint with the foot at 90°.
8. Recheck neurovascular integrity.

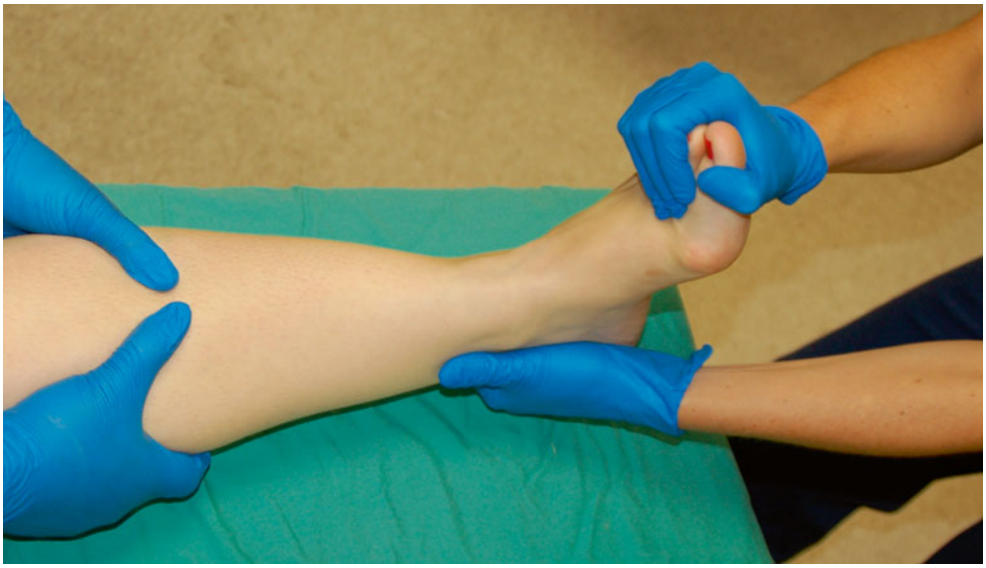

Fig. 105.3 Dorsiflexion with longitudinal heel traction

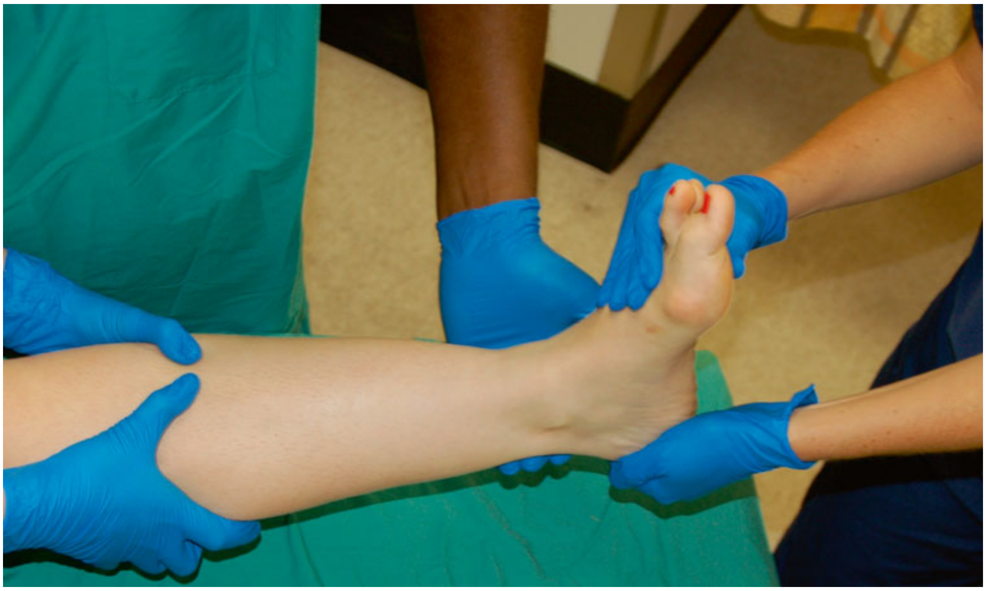

Fig. 105.4 Downward movement of foot (toward the floor) with longitudinal heel traction

105.5 Complications

- Compound fractures
- Neurovascular injury
- Skin and soft tissue damage
- Compartment syndrome

105.6 Pearls and Pitfalls

- Pearls
  - The ankle rarely dislocates without associated fractures.
- Pitfalls
  - Ankle dislocation is an orthopedic emergency, and reduction should not be delayed by imaging if there is evidence of neurovascular impairment. Complications that are exacerbated by delay in management include concomitant fractures, gross deformity of the ankle, severe stretching and tenting of the skin with resultant skin blisters, skin necrosis, and possible conversion to a compound fracture.
  - Be sure to check the radiograph carefully for commonly associated fractures notably of the malleoli.

Selected Reading

Collins DN, Temple SD. Open joint injuries: classification and treatment. Clin Orthop. 1989;243:48.

Hamilton WC. Injuries of the ankle and foot. Emerg Med Clin North Am. 1984;2:361.

Kelly PJ, Peterson FP. Compound dislocations of the ankle without fractures. Am J Surg. 1986;103:170.

Simon RR, Sherman SC, Koenigsknecht SJ, editors. Emergency orthopedics—the extremities. 5th ed. New York: McGraw-Hill; 2007. p. 264.

Wedmore IS, Charette J. Emergency department evaluation and treatment of ankle and foot injuries. Emerg Med Clin North Am. 2000;18:85.

Arthrocentesis

106

Shalu S. Patel and Bobby K. Desai

106.1 Indications

- Diagnosis of septic joint
- Diagnosis of traumatic effusion
- Diagnosis of inflammatory effusion
- Diagnosis of crystal-induced arthritis
- Therapeutic relief of pain from effusion

106.2 Contraindications

- Severe coagulopathy
- Skin infection over the needle insertion site
- Joint prosthesis
- Patients with bacteremia or sepsis (except to diagnose a septic joint)

106.3 Materials and Medications

- Betadine (povidone-iodine) or other skin antiseptic
- Sterile gloves
- Sterile towels
- Lidocaine 1 % or 2 % (5 mL) or other anesthetic of choice
- 18- to 22-gauge needle, 25-gauge needle
- Syringes (5 mL, 5–50 mL)
- Sterile gauze (4×4)
- Band-aid

106.4 Procedure

1. Informed consent may be required.
2. Position the patient appropriately. The joint should be placed in slight flexion.
3. Palpate the joint and identify anatomical landmarks.
    (a) For knee arthrocentesis, the needle should be inserted at the midpoint of either the medial or the lateral side of the patella (Fig. 106.1).
    (b) For acromioclavicular (AC) joint arthrocentesis, the needle should be inserted at the superior surface of the AC joint (Fig. 106.2).
    (c) For glenohumeral joint arthrocentesis, there are two approaches.
        (i) In the anterior approach, the needle is inserted into the groove lateral to the coracoid process (Fig. 106.3).
        (ii) In the posterior approach (preferred), the needle is inserted below the posterior border of the acromion process and lateral to the border of the scapula (Fig. 106.4).
4. Prepare the skin and drape in a sterile fashion.
5. Using lidocaine (drawn up in 5-mL syringe), anesthetize the skin with the 25-gauge needle.
6. Secure the 18- to 22-gauge needle on the 5- to 50-mL syringe (depending on the size of the joint) and insert it into the skin.
7. Advance the needle slowly into the joint space while aspirating until joint fluid can easily be withdrawn. While inserting the needle into the joint space, avoid scraping the needle against the bone.

S.S. Patel, MD
Department of Emergency Medicine, Florida Hospital Tampa,
Florida Hospital Carrollwood, Tampa, FL, USA

B.K. Desai, MD (✉)
Department of Emergency Medicine, University of Florida Health
Shands Hospita, Gainesville, FL, USA
e-mail: bdesai@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_106

8. If fluid cannot be aspirated easily, the catheter can be repositioned further in the joint space or turned by 45° sequentially as needed.
9. Once the joint fluid is aspirated, pull out the needle and hold pressure with gauze. Bleeding should be minimal.
10. Place a band-aid or other dressing over the site.
11. Send the synovial fluid to the laboratory. Generally, laboratory analyses may include crystals, protein, glucose, cell count and differential, culture and sensitivity, and Gram stain.

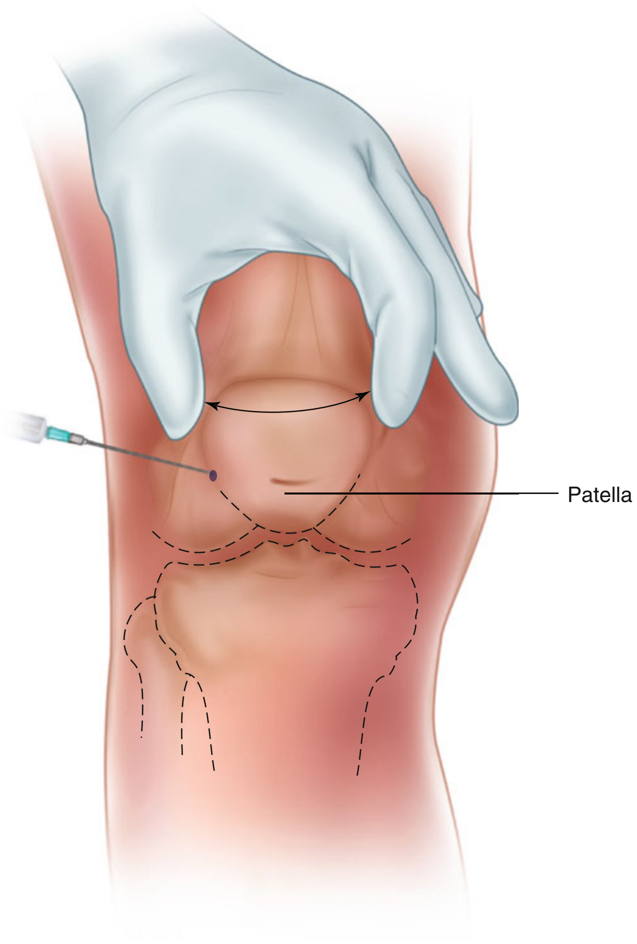

Fig. 106.1 Knee arthrocentesis

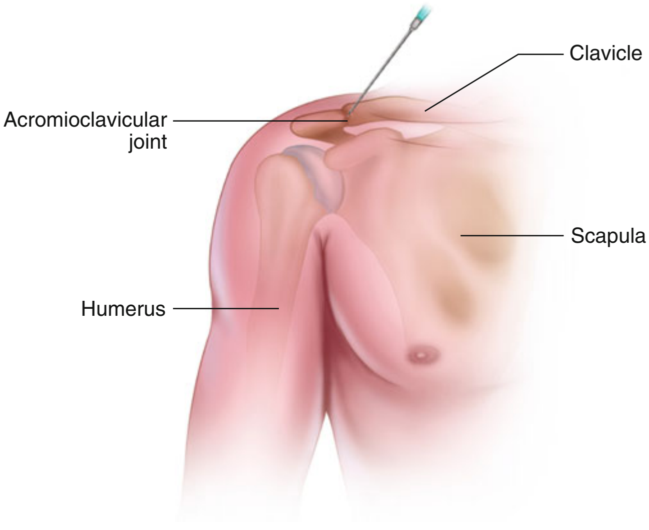

Fig. 106.2 Acromioclavicular joint arthrocentesis

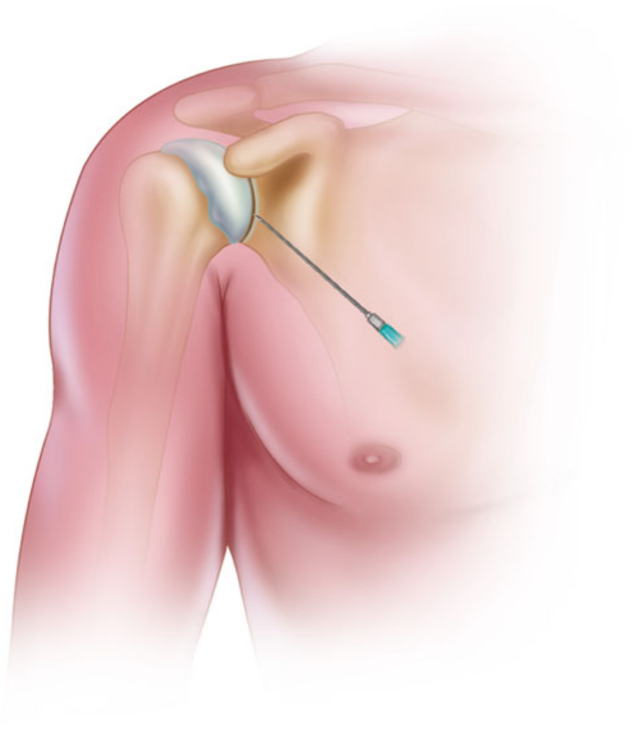

Fig. 106.3 Glenohumeral joint arthrocentesis: anterior approach

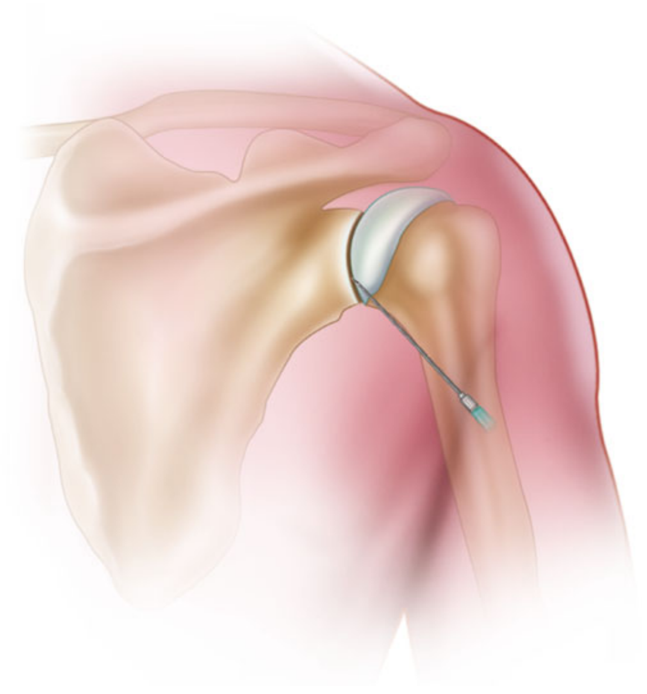

Fig. 106.4 Glenohumeral joint arthrocentesis: posterior approach

106.5 Complications

- Introduction of infection
- Bleeding

106.6 Pearls and Pitfalls

- The preferred site of entry is over the extensor surface of the joint. This will reduce the risk of damage to tendons, ligaments, and blood vessels.
- When assessing synovial fluid, the Rule of Twos may be used to differentiate among normal, inflammatory, and septic fluid. Normal synovial fluid has less than 200 white blood cells (WBCs)/mm^3. Noninflammatory synovial fluid has 200–2000 WBCs/mm^3. Inflammatory synovial fluid has greater than 2000 WBCs/mm^3 (but <50,000 WBCs/mm^3). Septic synovial fluid has greater than 75,000 WBCs/mm^3.
- Only septic synovial fluid will have a positive Gram stain and culture.

Selected Reading

Biundo JJ, Roberts N, Deodhar A. Regional musculoskeletal complaints. In: Stone JH, editor. A clinician's pearls and myths in rheumatology. New York: Springer Science; 2009. p. 433–4.

Parrillo SJ, Fisher J. Arthrocentesis. In: Roberts JR, Hedges J, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004. p. 1042–57.

Self WH, Wang EE, Vozenilek JA, del Castillo J, Pettineo C, Benedict L. Dynamic emergency medicine. Arthrocentesis. Acad Emerg Med. 2008;15:298.

Thomsen TW, Shen S, Shaffer RW, Setnik GS. Arthrocentesis of the knee. N Engl J Med. 2006;354:e19.

Intra-articular Injection

107

Bharat Kothakota and Muhammad Waseem

107.1 Indications

- Aspiration of fluid (arthrocentesis)
  - For diagnosis: to rule out infection
  - To relieve pressure from large, painful joint effusion
- Injection of joints with inflammatory arthritis
  - Tendinitis
  - Bursitis
  - Rheumatoid arthritis (RA)
- Injection of joints with osteoarthritis (OA)
  - Injection of large weight-bearing joints
  - Injection of small joints of hands
- Intra-articular anesthetic
  - Shoulder reduction
  - Ankle impingement syndrome

107.2 Contraindications

- Cellulitis
- Bacteremia
- Fracture

107.3 Materials and Medications

- Glucocorticoid
  - Duration of effect inversely proportional to solubility
    - Less soluble → longer acting
  - Choice of steroid is the personal preference of the physician
  - Methylprednisone (Depo-Medrol) and triamcinolone acetonide (Kenalog)
    - Less likely to induce postinjection flare
  - Kenalog and triamcinolone hexacetonide (Aristospan)
    - Longest-acting agents
- Local anesthetic
  - 1 % Lidocaine
- Needle
  - 18- to 22-gauge used for knee, ankle, hip, elbow, and shoulder
  - 25-gauge or smaller used for smaller joints (interphalangeal)
- Syringe

B. Kothakota, MD • M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: bharatkothakota@yahoo.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_107

107.4 Procedure

1. Selecting an injection approach.
    - Knee
        - Lateral approach: 1 cm inferior to the patella laterally (Fig. 107.1)
        - Medial approach: 1 cm inferior to the patella medially (Fig. 107.2)
    - Ankle
        - Lateral approach: just inferior to the lateral malleolus (Fig. 107.3)
        - Medial approach
            - Plantar flex the foot.
            - Angle the needle cephalad to pass between the medial malleolus and the tibialis anterior tendon (Fig. 107.4).
    - Shoulder
        - Posterior approach
            - Insert the needle 1 cm inferior and 1 cm medial to the posterolateral corner of the acromion.
            - Direct the needle anterior and medial toward the coracoid process (Fig. 107.5).
2. Skin preparation.
    - Make three separate concentric outward spirals with iodine disinfectant
    - Scrub with cyclohexidine preparation
3. Mark the injection site by impressing the skin with a hard object.
    - Sterile end of needle sheath
    - Ball point pen with tip retracted
4. Local anesthesia.
    - 1 % Lidocaine injected into the skin and subcutaneous tissue
    - Short burst of ethyl chloride spray before iodine preparation
    - Mixing lidocaine with glucocorticoid preparation
5. Always aspirate joint fluid before injecting the corticosteroids.
    - Use a 1.5-inch 18-gauge needle for aspiration.
    - Confirm that the needle is in the joint space.
    - Reduced effusion size before injection can improve outcomes.
    - Compress the opposite side of the joint to aid in aspiration (Fig. 107.6).
6. After aspirating, change the syringe.
    - Use a sterile hemostat or hand to stabilize the needle within the joint space (Fig. 107.7).
    - Avoid injecting corticosteroids if the aspirate appears purulent.
7. Injection of medication.
    - Can use the same needle used for aspiration.
    - Insert needle 0.75–1.25 inch in depth for injection.
8. Remove the needle, wipe the iodine solution clean, and apply the bandage.
9. Postinjection care.
    - First 48 h: bedrest versus minimize walking
    - Next 2–3 weeks: crutches or cane

Fig. 107.1 Knee arthrocentesis, lateral approach

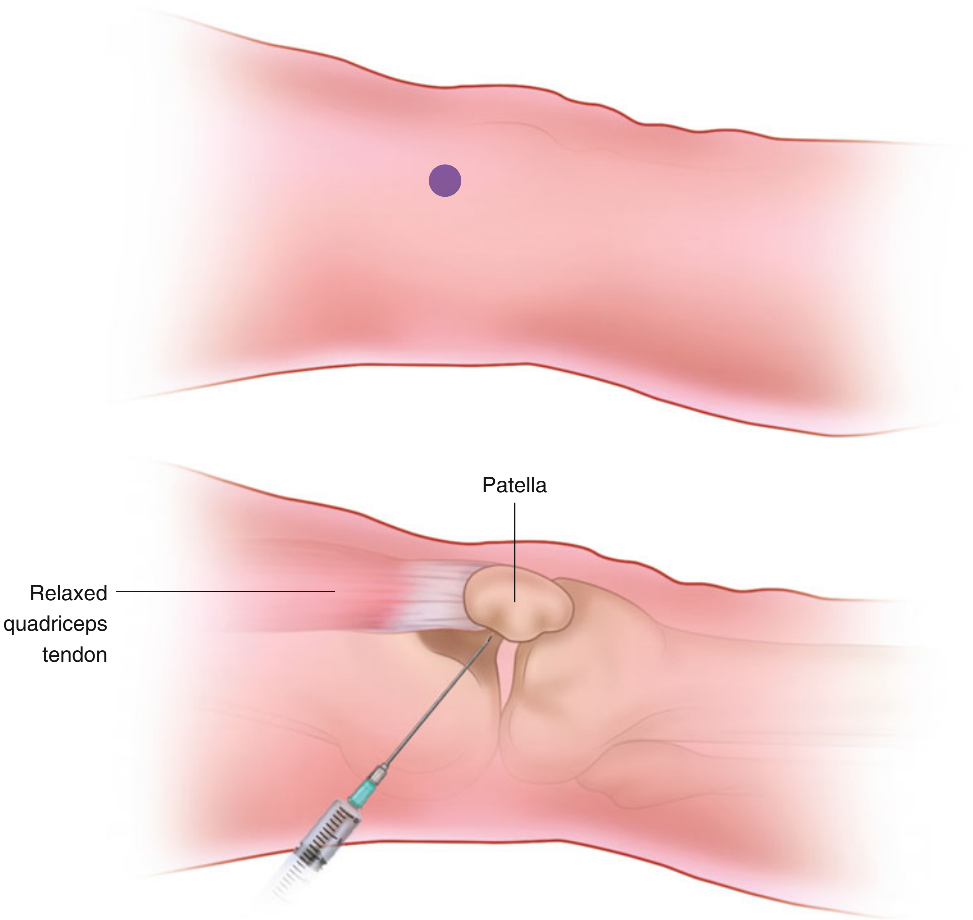
Patella
Relaxed quadriceps tendon

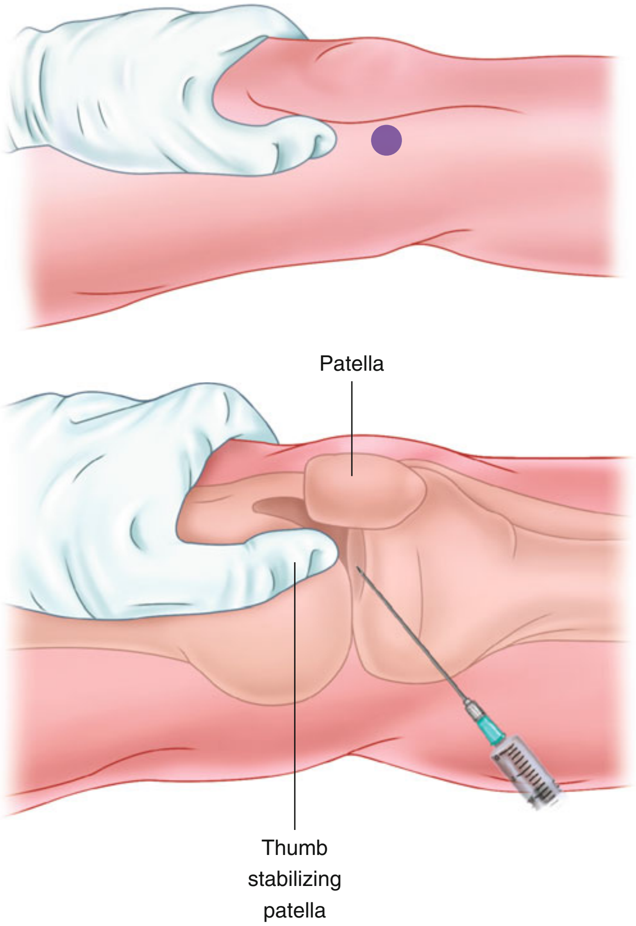
Patella
Thumb stabilizing patella

Fig. 107.2 Knee arthrocentesis, medial approach

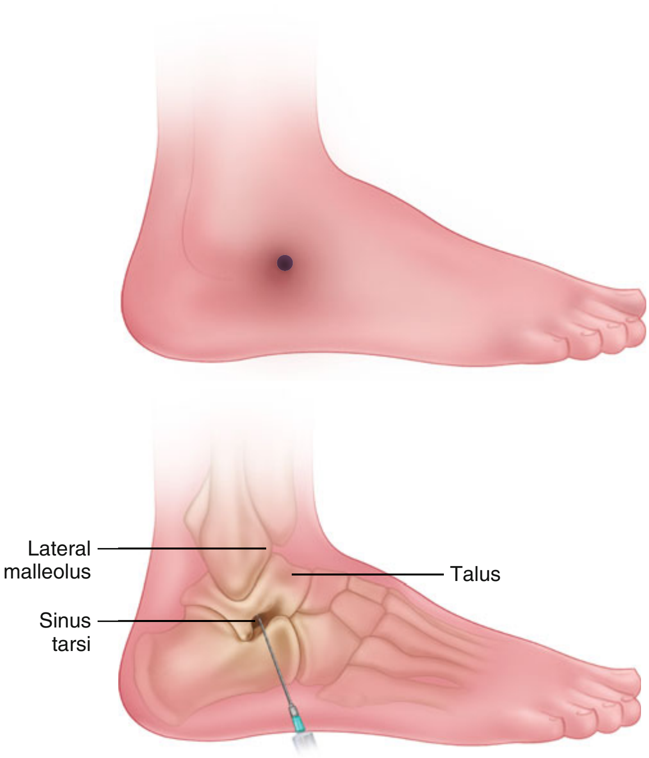
Lateral malleolus
Talus
Sinus tarsi

Fig. 107.3 Ankle arthrocentesis, lateral approach

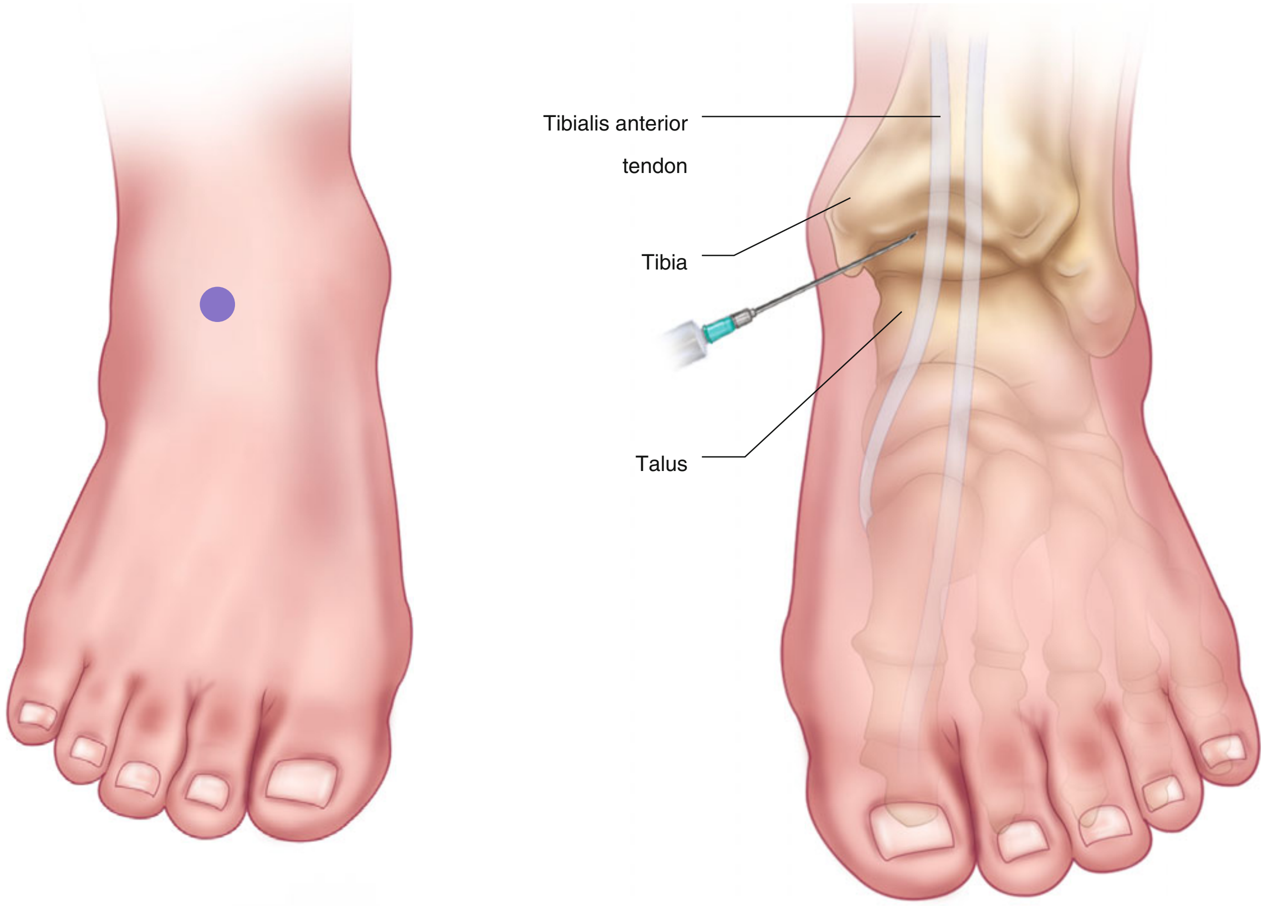
Tibialis anterior
tendon
Tibia
Talus

Fig. 107.4 Ankle arthrocentesis, medial approach

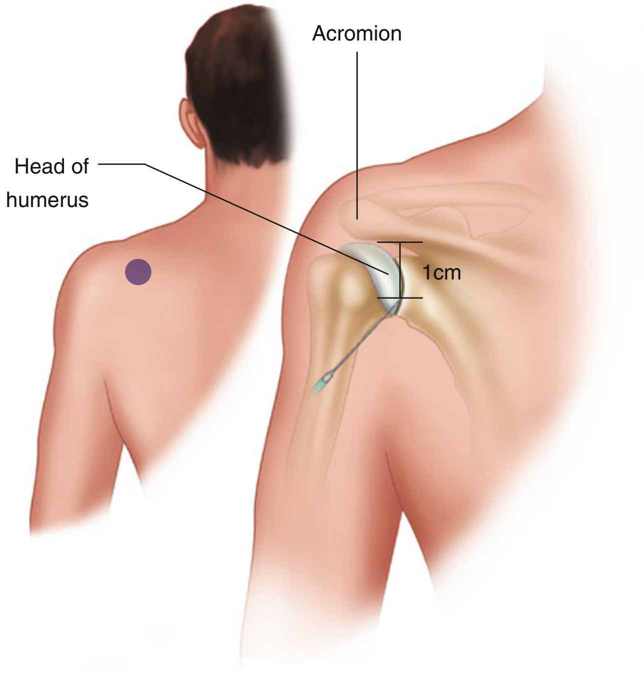
Acromion
Head of
humerus
1cm

Fig. 107.5 Shoulder arthrocentesis, posterior approach

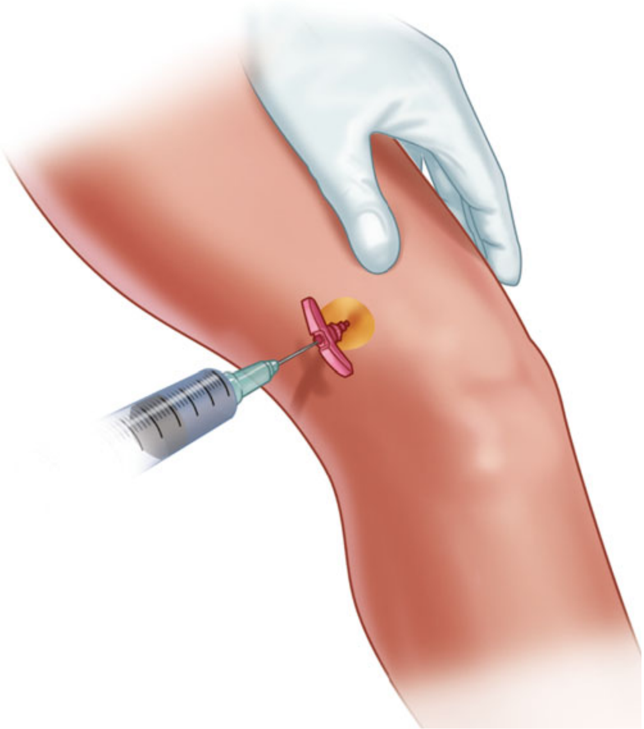

Fig. 107.6 Compress opposite side of joint to aid in aspiration

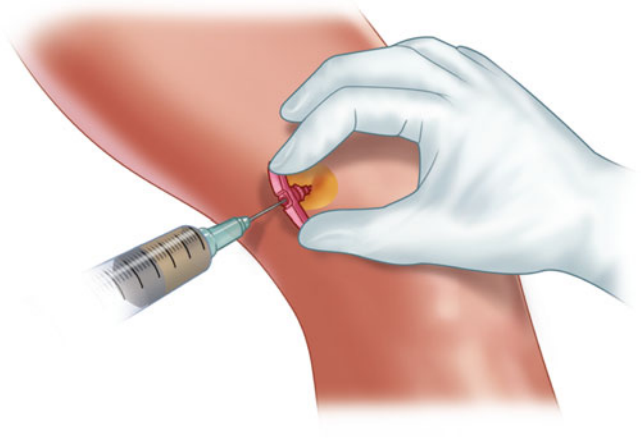

Fig. 107.7 Use hemostat or hand to stabilize needle within joint space

107.5 Complications

- Local postinjection flare
  - Irritation of the synovium by steroid microcrystals
  - Can be confused for infection
  - Occurs and resolves within 48 h after injection
  - Treat with ice and appropriate analgesics.
- Iatrogenic joint infection
  - Suspect if it begins later than, or lasts longer than, flare
  - Increasing pattern of pain
  - Fever, malaise, redness, or drainage around injection site
  - *Staphylococcus aureus* most common
- Subcutaneous atrophy and depigmentation
  - Leakage of corticosteroids into soft tissues
- Aspiration of blood
  - Indicative of trauma or bleeding disorder (hemophilia)
- Systemic absorption
  - Water-soluble preparations
  - Dose dependent
  - Injection into multiple joints
  - Transient hyperglycemia in diabetic patients
  - Avascular necrosis of the femoral head

107.6 Pearls

- Mixing lidocaine with glucocorticoids
  - Reduces pain caused by injection of steroids into joint space
  - Less likely to cause soft tissue atrophy and tendon rupture
  - Immediate relief from anesthetic indicates proper injection
- Limiting intra-articular glucocorticoid injections per joint
  - OA
    - Four injections per lifetime of the joint
    - Injections reduce the rate of accelerated degeneration in joints
  - RA
    - Limit of one injection per month
    - No evidence of glucocorticoid-induced cartilage loss

Selected Reading

Aponte EM, Schraga ED. Joint reduction, shoulder dislocation, anterior. Available at: http://emedicine.medscape.com/article/109130-overview#a08.

Cianflocco AJ. Intra-articular injections of the knee: a step-by-step guide. J Fam Pract. 2011;60(Suppl):S48–9. Available at: http://www.jfponline.com/pages.asp?id=10062.

Lavelle W, Lavelle ED, Lavelle L. Intra-articular injections. Anesthesiol Clin. 2007;25:835–62.

Molis MA, Young CC. Ankle impingement syndrome. Available at: http://emedicine.medscape.com/article/85311-overview.

Neustadt DH. Intra-articular injections for osteoarthritis of the knee. Cleve Clin J Med. 2006;73:897–911.

Roberts WN. Intraarticular and soft tissue injections: what agents(s) to inject and how frequently? Available at: www.uptodate.com.

Roberts WN. Joint aspiration or injection in adults: techniques and indications. Available at: www.uptodate.com.

Sugar-Tong Splint

108

Katrina John, Jeffrey Kile, and Amish Aghera

108.1 Indications

- Fractures to the wrist or forearm
- To prevent motion at the wrist and elbow
- To prevent supination and pronation

108.2 Contraindications

- Relative
  - Evidence of compartment syndrome or any neurovascular compromise

108.3 Materials (Fig. 108.1)

- Stockinette
- Padding
- Splint material: fiberglass/plaster of Paris or prefabricated splint rolls, 2, 3, 4 inches depending on age and body habitus
- Trauma shears/scissors
- Elastic bandage
- Tape
- Container with water
- Gloves, eyemask, sheet
- Sling

K. John, MBBS • J. Kile, MBBS, PhD, MPH (✉)
Department of Emergency Medicine, Eisenhower Medical Center,
Rancho Mirage, CA, USA
e-mail: trenjohn@me.com; jeffrey.kile@gmail.com

A. Aghera, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: aaghera@maimonidesmed.org

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_108

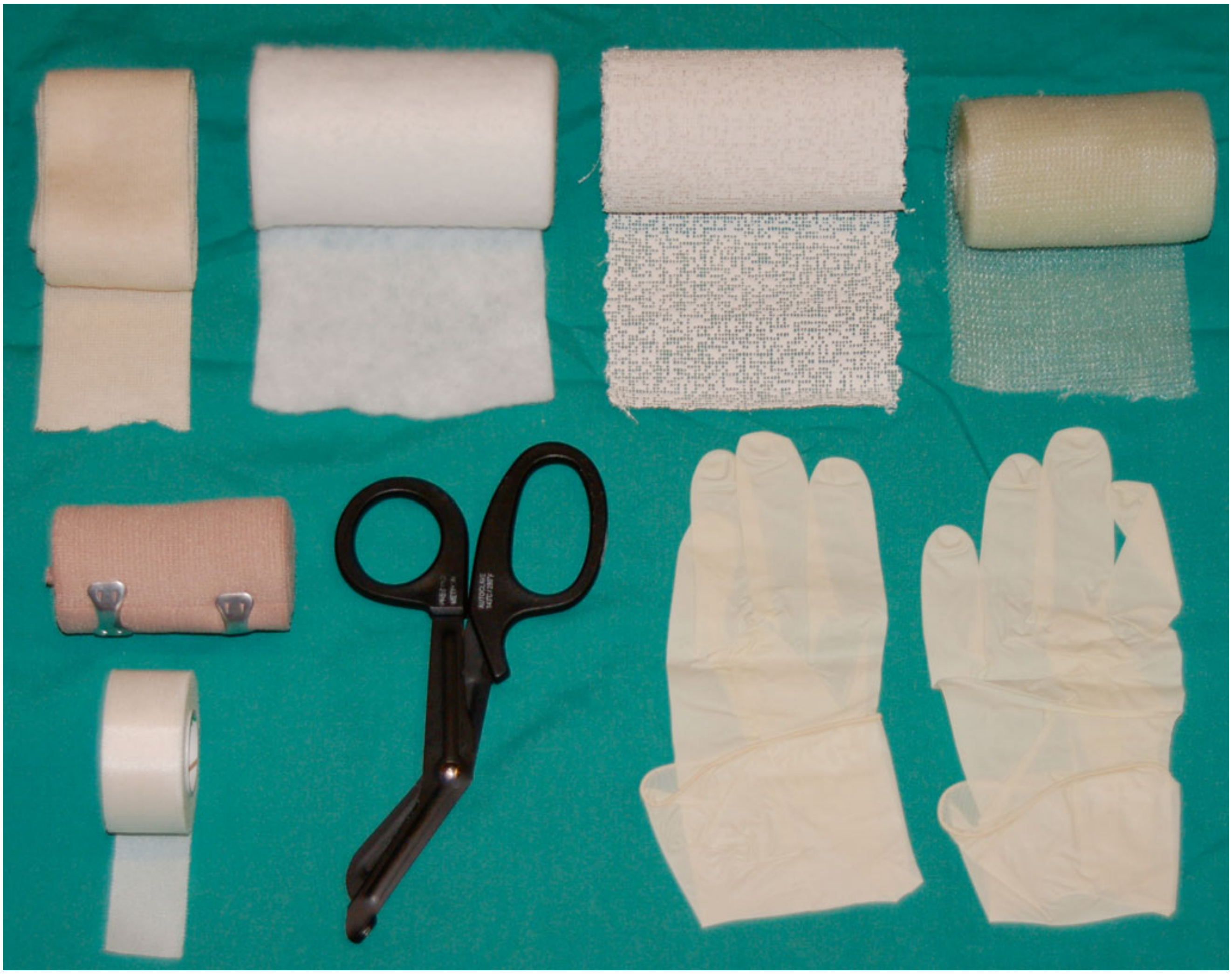

Fig. 108.1 Equipment

108.4 Procedure

1. Ensure the skin of the affected extremity is clean, dry, and intact.
2. Cover the patient with a sheet or gown to protect the patient's clothing and the surrounding area.
3. Position the patient's arm abducted at 90° at the shoulder and internally rotated with the elbow flexed at 90° (Fig. 108.2).
4. Measure the splinting material by running a single layer from the metacarpal heads of the dorsum of the hand along the extensor surface of the forearm over the elbow and humeral condyles and back down the flexor surface of the forearm to the palmar aspect of the hand to the metacarpal heads.
5. If using non-prefabricated splint rolls, lay the measured piece of splinting material out on a flat surface and multiply the layers to the same length, 6–8 layers for fiberglass and 10–12 layers for plaster.
6. Measure the stockinette from the finger tips to the mid-humerus and cut a hole for the thumb.
7. Place the stockinette on the arm (Fig. 108.3).
8. Use a 3- to 4-inch padding roll to apply several layers of circumferential padding extending from the metacarpal heads to the midhumerus below the level of the stockinette (Fig. 108.4).
9. Wet the already prepared and measured splinting material and remove the excess water.
10. Ensure the forearm is in the aforementioned position, and apply the splinting material from the metacarpal heads of the dorsum of the hand along the extensor surface of the forearm over the elbow and humeral condyles and back down the flexor surface of the forearm to the palmar aspect of the hand to the metacarpal heads (Fig. 108.5).
11. Fold each end of the stockinette down over the padding and splinting material.
12. An extra layer of padding can be added at this stage.
13. Secure the entire splint with two elastic bandages/ace wrap and apply tape to ensure the bandages stay in place (Fig. 108.6).
14. Place the arm in a sling.

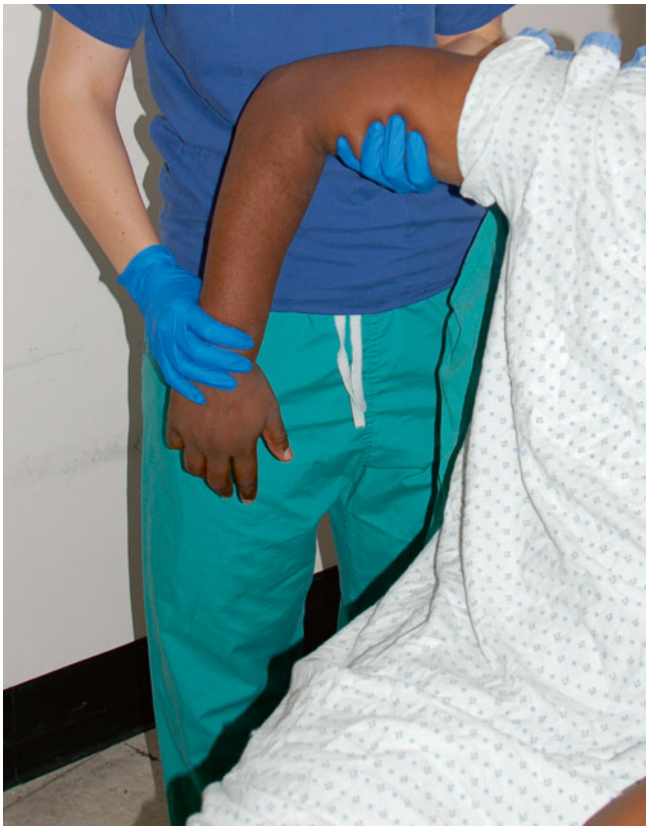

Fig. 108.2 Arm positioning

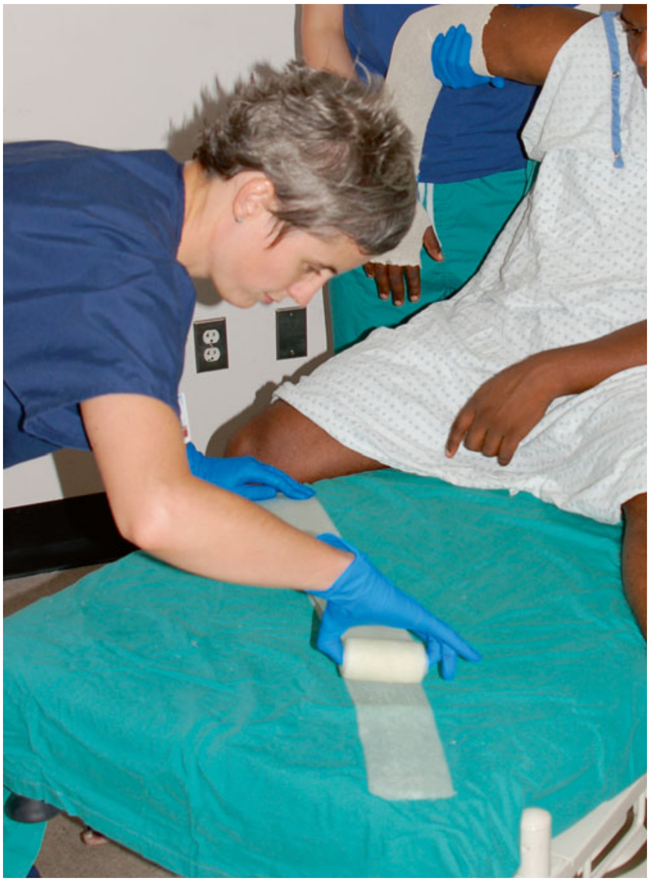

Fig. 108.3 Fiberglass being measured and stockinette on patient

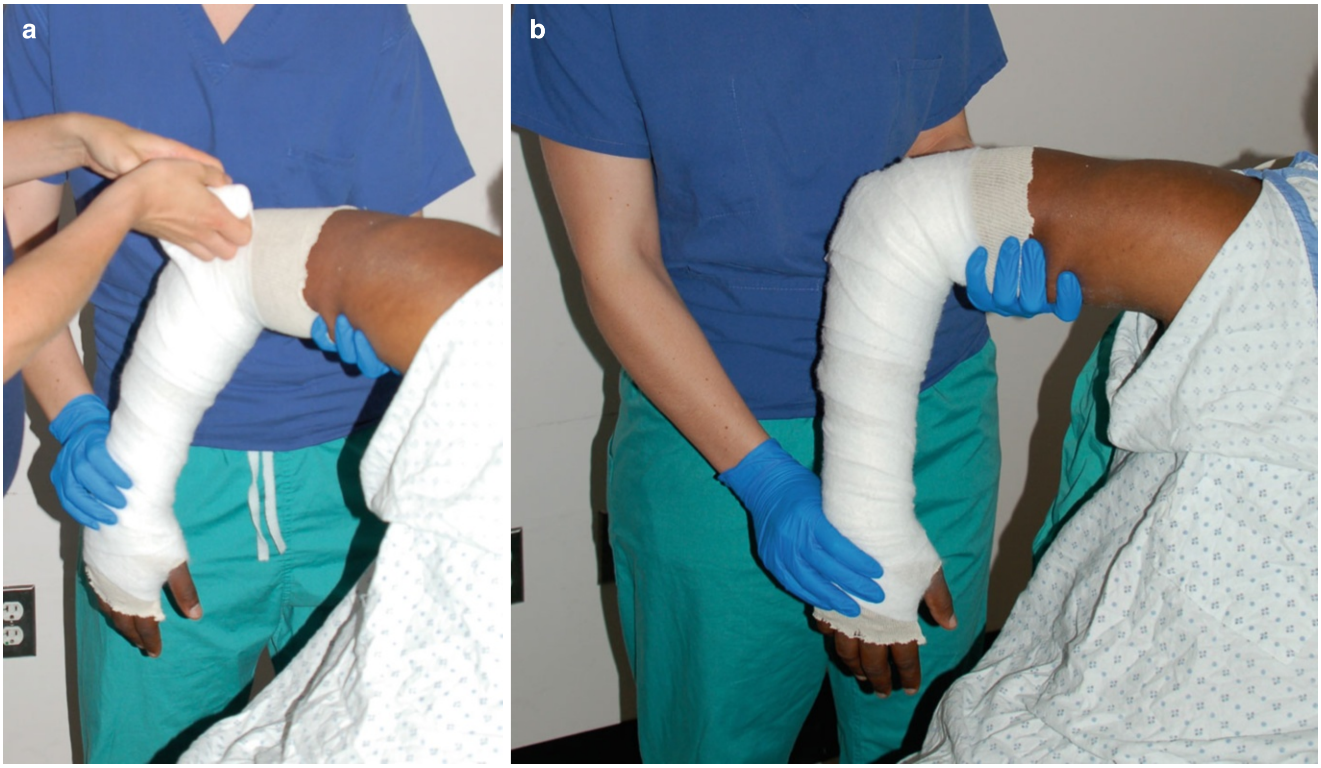

Fig. 108.4 (**a**) Application of padding over the stockinette (partial), (**b**) Application of padding over the stockinette (complete)

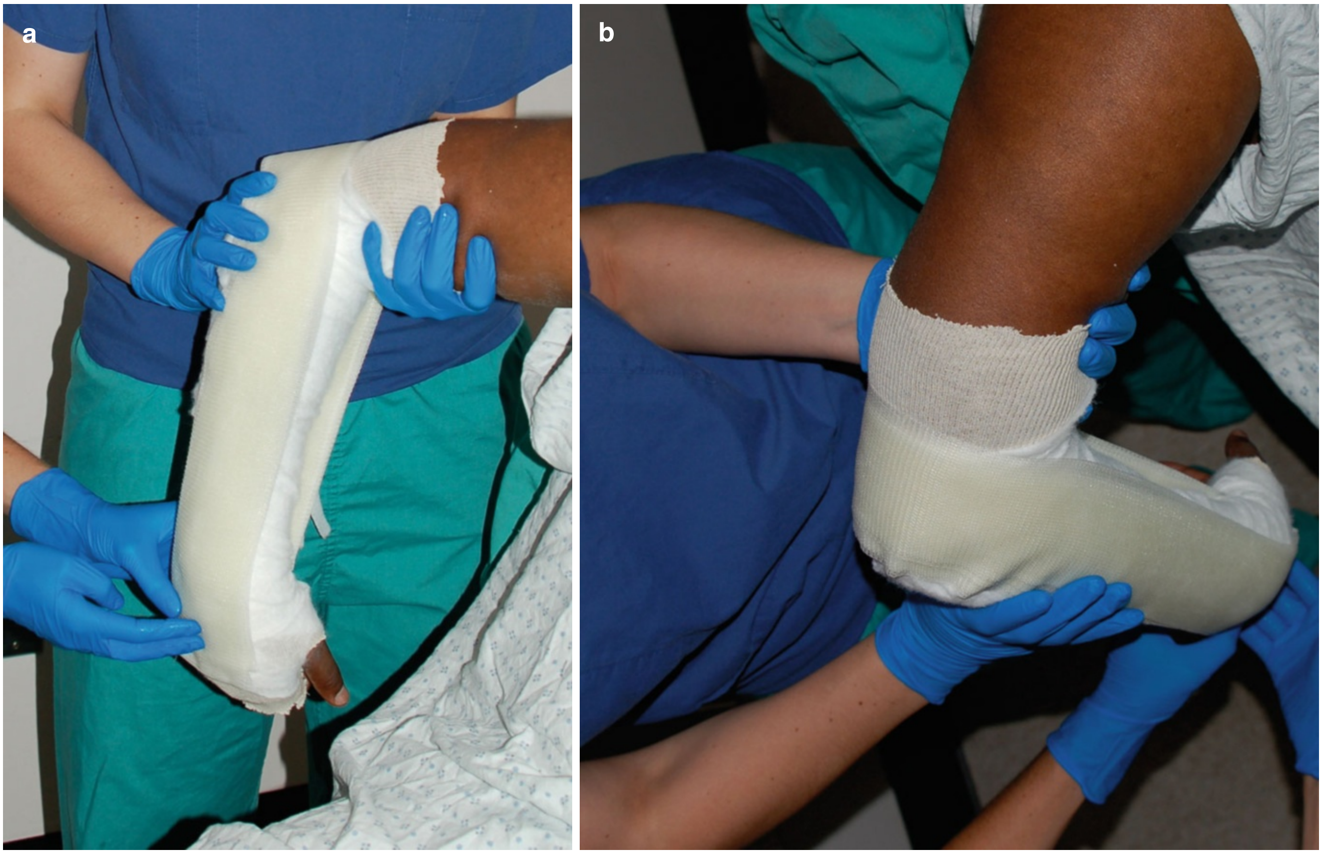

Fig. 108.5 (**a**) Application of splinting material over padding (anteromedial view), (**b**) Application of splinting material over padding (anterolateral view)

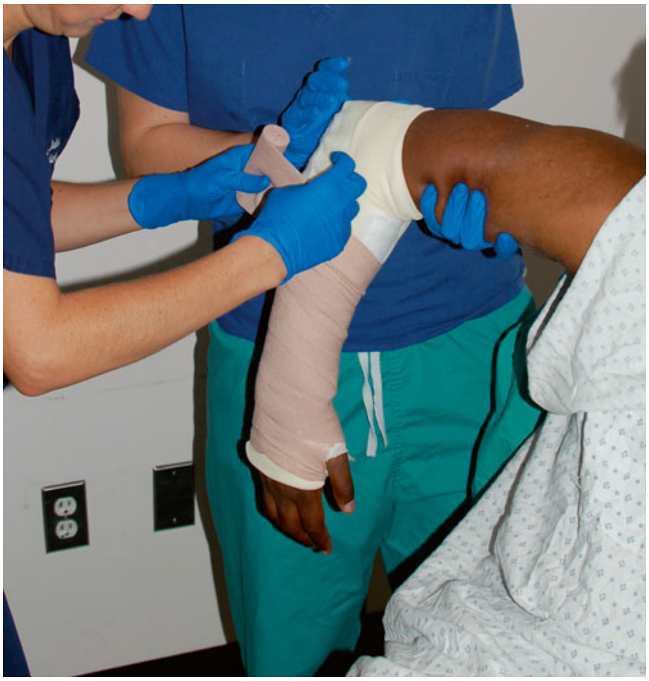

Fig. 108.6 Application of elastic bandage over splinting material

108.5 Complications

- Compartment syndrome

108.6 Pearls and Pitfalls

- Pearls
  - This procedure is best done with an assistant to hold the extremity in the desired position and to prevent the splint slipping as it is secured.
  - Having the patient in the illustrated position enables the practitioner to use gravity to hold the splint in the correct position while securing it; especially helpful if an assistant is not available.
- Pitfalls
  - If the splint is too short, it fails to immobilize the wrist.
  - If the splint is too long, it will cause reduced motion and stiffness at the metacarpophalangeal joints and swelling of the fingers due to immobility.

Selected Reading

Bong MR, Egol KA, Leibman M, Koval KJ. A comparison of immediate postreduction splinting constructs for controlling initial displacement of fractures of the distal radius: a prospective randomized study of long-arm versus short-arm splinting. J Hand Surg Am. 2006;31:766–70.

Denes AE, Goding R, Tamborlane J, Schwartz E. Maintenance of reduction of pediatric distal radius fractures with a sugar-tong splint. Am J Orthop (Belle Mead NJ). 2007;36:68–70.

Gartland JJ. The sugar tong splint. Am J Orthop. 1963;5:131.

McGeorge DD, Stilwell JH. The sugar tong splint: a reliable method of arm splintage in the child. J Hand Surg Br. 1989;14:357.

Simon RR, Koenigsknecht SJ, editors. Emergency orthopedics—the extremities. 3rd ed. New York: McGraw Hill; 1995.

Part XIV

Obstetric Procedures

109 Fetal Heart Rate Monitoring

Nathaniel Lisenbee and Joseph A. Tyndall

109.1 Indications

- Fetal heart rate (FHR) monitoring is important because it provides basic patterns that can be correlated to the acid–base status, circulatory volume, and oxygenation status of the fetus through brainstem detection and subsequent cardiac response. It has numerous indications during the antepartum and intrapartum stages [1].
- Antepartum indications include:
  - Nonstress test (consists of monitoring FHR in conjunction with fetal movements)
  - Contraction stress test (consists of monitoring FHR during contractions, which are induced pharmacologically)
  - Biophysical profile (BPP; consists of a nonstress test with an additional ultrasound)
- Intrapartum indications include monitoring FHR during:
  - Uterine contractions
  - Pain medications/anesthetic administration to the mother during labor
  - Procedures performed during labor
  - Second stage of labor
  - High-risk pregnancies, which can be defined by a number of conditions including [2, 3]:
    - Maternal diabetes, asthma, preeclampsia/eclampsia
    - Multiple gestations
    - Intrauterine growth restriction
    - Premature rupture of membranes
    - Lack of prenatal care

109.2 Contraindications

- Contraindications for internal FHR monitoring
  - Presence of placenta previa
  - Lack of ability to identify the portion of the fetal body where device application is being considered
  - Active herpes, active hepatitis, or human immunodeficiency virus (HIV) in the mother
- Contraindications for external FHR monitoring
  - None

109.3 Methods

- Two methods for FHR monitoring:
  - Auscultation monitoring
    - Defined as auscultating FHR every 15 min in the first stage of labor and auscultating every 5 min in the second stage
    - Does not provide strips with information on FHR variability or the shape of FHR accelerations and decelerations
- Electronic FHR monitoring
  - Allows for real-time continuous monitoring of FHR activity
  - Provides strips with information on FHR variability or the shape of FHR accelerations and decelerations
  - Can be performed by Doppler ultrasound or internal fetal electrocardiography (ECG)

When comparing the two methods for electronic FHR monitoring, both are equally as reliable in most settings. Thus, external monitoring is the preferred method because it is noninvasive. However, in instances in which external monitoring becomes difficult owing to poor quality or technical difficulties, invasive monitoring is indicated.

N. Lisenbee, MD
U. S. Air Force Medical Center Keesler,
Biloxi, MI, USA
e-mail: nate4085@ufl.edu

J.A. Tyndall, MD, MPH (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_109

109.4 Equipment and Procedures

Multiple methods exist for electronic FHR monitoring [4]. The most commonly used are external monitoring by Doppler ultrasound and internal monitoring by fetal ECG.

109.4.1 Doppler Ultrasound is a Noninvasive Method to Monitor FHR (Fig. 109.1)

- Equipment
  - Electronic FHR monitor
  - Contraction monitor sensor with belt
  - FHR sensor with belt (consists of ultrasound transducer and ultrasound sensor)
  - Ultrasound coupling gel
- Procedure
  1. Place the patient in a supine position.
  2. Palpate the fetal anatomy through the maternal abdomen to find the approximate location of the fetal heart.
  3. Place ultrasound coupling gel on the maternal abdomen at the sight of suspected fetal cardiac activity.
  4. Place the transducer probe on gel and locate the fetal heart tones.
  5. Once the fetal heart tones are located, secure the FHR sensor to the maternal abdomen with the attached belt.
  6. Place the contraction monitor sensor near the fundus in order to monitor uterine contractions.
  7. Attach the FHR sensor and contraction monitor to the electronic FHR monitor to obtain printouts of FHR and uterine contractions.

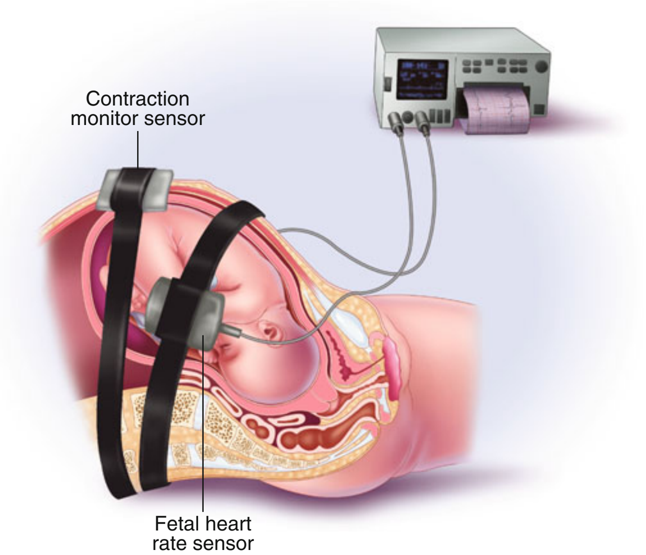

Fig. 109.1 External fetal heart rate monitoring

109.4.2 Internal Fetal ECG is an Invasive Method to Monitor FHR and is Used Only in the Intrapartum Period (Fig. 109.2)

- Equipment
  - Fetal scalp monitoring electrode
  - Leg plate electrode
  - Sterile vaginal lubricant
  - Electronic FHR monitor
- Procedure (Fig. 109.3)
  1. Place the patient in a dorsal lithotomy position.
  2. Sterilize the perineal area.
  3. Perform a bimanual vaginal examination to identify the presenting fetal head. (Note: rupture of membranes must occur before scalp electrode placement.)
  4. Place the spiral electrode guide tube on the fetal scalp and advance the electrode until it contacts the scalp.
  5. Rotate the drive tube clockwise approximately one rotation while maintaining pressure on the guide tube and drive tube.
  6. Release the electrode locking device by pressing together the arms on the drive tube grip.
  7. Carefully slide the drive and guide tubes off the electrode wires while holding the locking device open.
  8. Attach the leg plate to the inner thigh of the mother as a means to eliminate electrical interference.
  9. Attach the spiral electrode wires to the color-coded leg plate, which is then connected to the electronic fetal monitor.
  10. Do not forget to sterilize the area of electrode placement after delivery is completed and the scalp electrode is removed.

When comparing the two methods, both are equally reliable in most settings. Thus, external monitoring is the preferred method because it is noninvasive. However, in instances in which external monitoring becomes difficult owing to poor quality or technical difficulties, invasive monitoring is indicated.

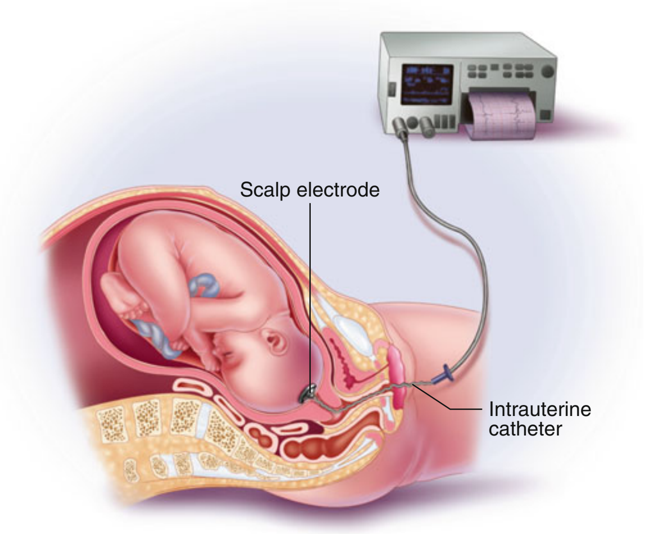

Fig. 109.2 Internal fetal heart rate monitoring

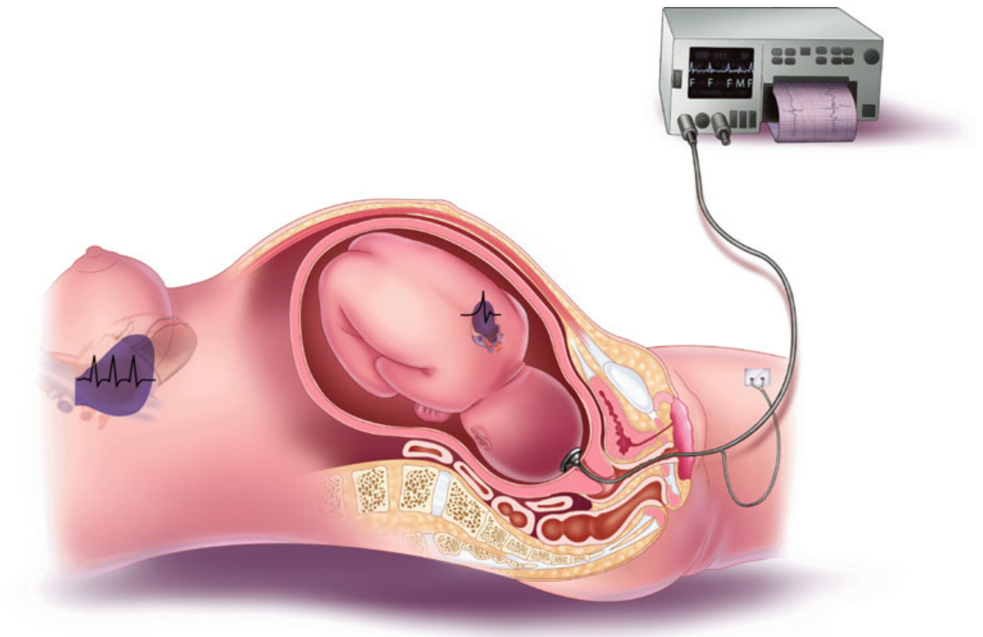

Fig. 109.3 Internal fetal heart rate monitoring

109.5 Complications

- Complications of external FHR monitoring
  - Confusing maternal aortic pulsations with FHR
  - Inability to locate FHR
- Complications of internal FHR monitoring
  - Fetal or maternal hemorrhage, fetal infection (usually scalp abscess at the site of insertion)
  - Uterine perforation
  - Subsequent fetal infection due to the invasive nature of the procedure

References

1. Hobel CJ. Intrapartum clinical assessment of fetal distress. Am J Obstet Gynecol. 1971;110:336–42.
2. Byrd JE. Intrapartum electronic fetal heart rate monitoring (EFM) and amnioinfusion. In: Advanced life support in obstetrics course syllabus. Kansas City: American Academy of Family Physicians; 1996. p. 97–106.
3. Queenan JT, Hobbins JC, Spong CY. Protocols for high-risk pregnancies. New York: Wiley; 2010. Retrieved 16 Jan 2012, from http://lib.myilibrary.com?ID=268955.
4. External and internal heart rate monitoring of the fetus. New Haven: Yale Medical Group; 2012. Retrieved from http://www.yalemedicalgroup.org/stw/.

Selected Reading

Alfirevic Z, Devane D, Gyte GM. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev. 2006;(3):CD006066.

American College of Obstetricians and Gynecologists. Fetal heart rate patterns: monitoring, interpretation, and management, ACOG technical bulletin, vol. 207. Washington, DC: ACOG; 1995.

Cunningham FG, Leveno KJ, Bloom SL, Hauth JC, Rouse DJ, Spong CY. Chapter 18: Intrapartum assessment. In: Williams obstetrics. 23rd ed. New York: McGraw Hill; 2010.

Freeman RK. Problems with intrapartum fetal heart rate monitoring interpretation and patient management. Obstet Gynecol. 2002;100:813.

Gomella LG, Haist SA. Chapter 13: Bedside procedures. In: Clinician's pocket reference. 11th ed. Columbus: McGraw-Hill; 2007.

Sweha A, Hacker TW, Nuovo J. Interpretation of the electronic fetal heart rate during labor. Am Fam Physician. 1999;59:2487–500.

110

Ultrasonography for Ectopic Pregnancy

L. Connor Nickels

An ectopic pregnancy is a pregnancy occurring outside of the uterine cavity (fundus) (Fig. 110.1).

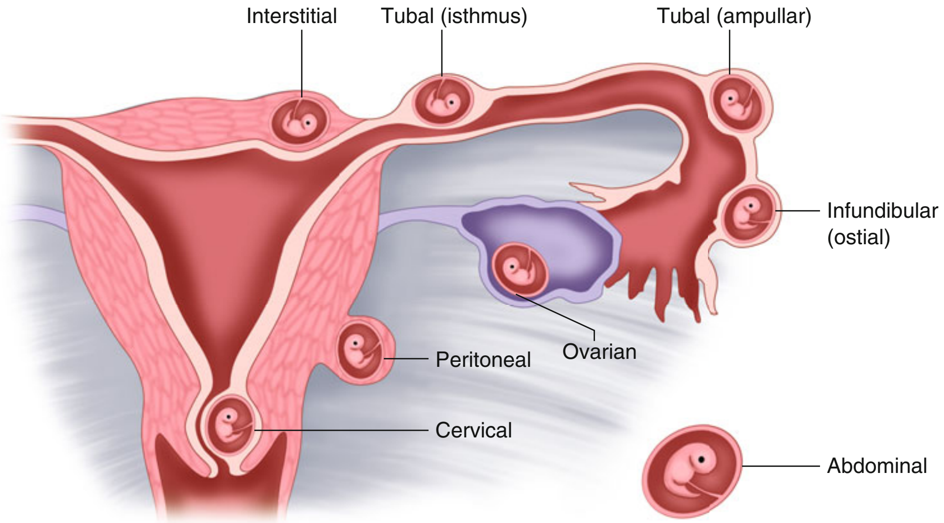
Interstitial
Tubal (isthmus)
Tubal (ampullar)
Infundibular (ostial)
Ovarian
Peritoneal
Cervical
Abdominal

Fig. 110.1 Ectopic pregnancy diagram: pregnancy occurring outside of the uterine cavity

---

L.C. Nickels, MD, RDMS
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_110

110.1 Indications

- Patient in the first trimester of pregnancy with any combination of the following:
  - Vaginal bleeding.
  - Acute pelvic pain.
  - Hypotension or shock.
  - Dizziness or syncope.
  - Positive pregnancy test.
  - Adnexal mass.
  - Cervical tenderness.
  - Abnormal rise in the serum human chorionic gonadotropin (hCG).
  - No specific sign or symptom is absolute; therefore, the index of suspicion must be high.
- Risk factors for ectopic pregnancy
  - Pelvic inflammatory disease (PID)
  - Previous or current intrauterine device
  - Infertility treatment
  - Previous tubal surgery
  - Tubal ligation
  - Advanced maternal age
  - Previous ectopic pregnancy
- Usually a sonographic diagnosis
- Sonographic signs of an ectopic can include any of the following:
  - Gestational sac seen in one of the following:
    - Adnexa with any of the following:
      - Yolk sac (Figs. 110.2 and 110.3)
      - Fetal pole with or without cardiac activity
      - Both of these
    - Low position in the cervix (cervical ectopic)
    - Seemingly in the uterus, but off to one side and with minimal surrounding myometrium (interstitial ectopic)
    - Within the peritoneal cavity, outside of the tubes (abdominal ectopic)
  - Pseudogestational sac seen in the uterus
    - Uterine enlargement or decidual reaction (single outline only) in the endometrium without a gestational sac
  - Other unidentifiable adnexa mass
  - Free fluid in the pelvis or other gravity-dependent area (i.e., Morison's pouch in the right upper quadrant in a supine or reverse Trendelenburg patient) (Fig. 110.4)
    - Small free fluid: tracks less than one third of the posterior cul-de-sac
    - Moderate free fluid: tracks less than two third of the posterior cul-de-sac
    - Large free fluid: tracks greater than two third of the posterior cul-de-sac
    - Right upper quadrant free fluid: 100 % predictability for ectopic
  - Empty uterus with serum hCG >1000 mIU/mL
- Patient in the first trimester of pregnancy with a serum hCG at or above the discriminatory zone (i.e., 1000 mIU/mL) without a sonographically normal gestational sac visualized within the uterus has an ectopic pregnancy until proven otherwise and should have an obstetrics consult in the emergency department.
  - Discriminatory zone may differ depending on the reference, but typically serum hCG between 1000 and 2000 mIU/mL.
  - For this chapter, we have used serum hCG greater than 1000 mIU/mL [1].
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.
- May be treated surgically or medically depending on findings.
  - Decision to be made by obstetrics consultant

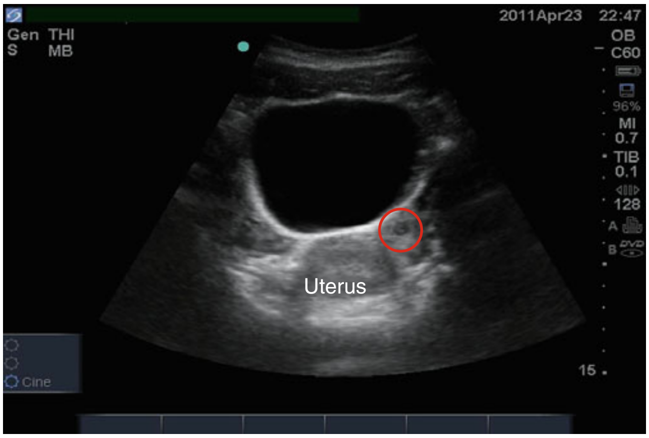
2011Apr23 22:47
Gen THI
S MB
OB
C60
96%
MI
0.7
TIB
0.1
128
Uterus
15
Cine

Fig. 110.2 Transabdominal transverse image of the uterus shows an ectopic pregnancy in the left adnexa (red circle) with the yolk sac present (Photo courtesy of L. Connor Nickels, MD, RDMS)

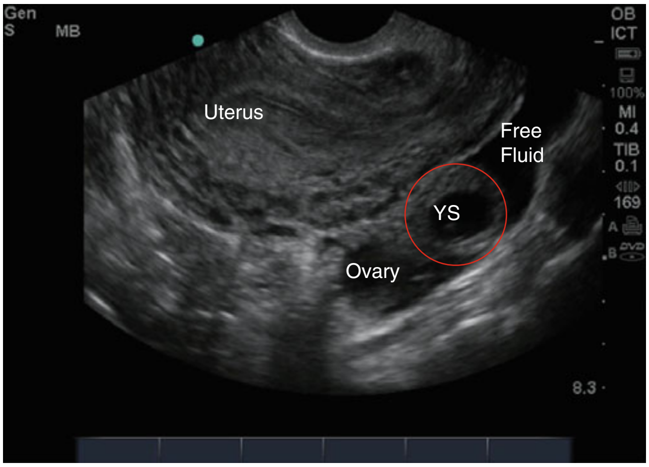
Gen
S MB
OB
ICT
100%
MI
0.4
TIB
0.1
169
Uterus
Free
Fluid
YS
Ovary
8.3

Fig. 110.3 Transvaginal sagittal image of an empty uterus with an ectopic pregnancy (in red circle) noted just posterior in the adnexa, adjacent to the ovary. The gestational sac contains a yolk sac and there is free fluid surrounding the ectopic, concerning for rupture (Photo courtesy of L. Connor Nickels, MD, RDMS). YS–yolk sac

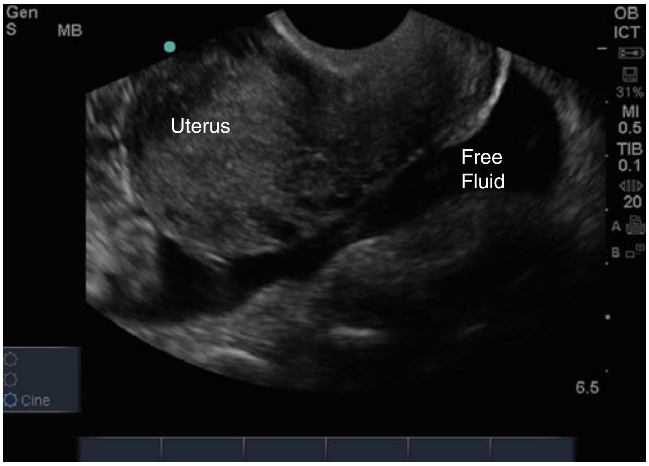
Gen
S MB
OB
ICT
31%
MI
0.5
TIB
0.1
20
Uterus
Free
Fluid
6.5
Cine

Fig. 110.4 Transvaginal sagittal image of an empty uterus and large free fluid in the posterior cul-de-sac in a patient with a presumed ectopic pregnancy (Photo courtesy of L. Connor Nickels, MD, RDMS)

110.2 Contraindications

- Treating ectopic pregnancy medically when there is a fetal pole with fetal cardiac activity present
- Failing to consult obstetrics/gynecology when the patient has a serum hCG above the discriminatory zone and no sonographic findings to diagnose intrauterine pregnancy

110.3 Materials and Medications

- Ultrasound machine
- Probes: transabdominal and transvaginal
- Gel
- Skilled ultrasound operator
- Endocavitary probe covers
- Pelvic setup (speculum and cultures)
- Cardiac monitor, two large-bore intravenous needles
- Laboratory work: serum quantitative hCG, hemoglobin, group and Rh, type and screen
  - May add additional laboratory tests depending on the stability and symptoms of the patient

110.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal
  1. Place the patient in the supine position.
  2. Ideally, the bladder will be full for a good acoustic window.
  3. Using a curvilinear probe, 3.5–5.0 MHz.
  4. Begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Imaging in the transverse plane should be done in the same fashion.
  6. Any signs of an intrauterine pregnancy should be clearly identified and measured.
     - Gestational sac diameter
     - Yolk sac diameter
     - Fetal pole with crown rump diameter
       - If present, fetal cardiac activity should be recorded by using M mode to obtain a tracing and measure the fetal heart rate.
  7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
  8. Any abnormalities identified should be noted.
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
     - On pelvic imaging or FAST examination
     - Intrauterine contents not consistent with an intrauterine pregnancy and/or not clearly, centrally visualized within the fundus of the uterus
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
  9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
- Transvaginal
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use a transvaginal probe, 5–7.5 MHz.
  4. Repeat same steps as transabdominal imaging previously.
- The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

110.5 Complications

- Bleeding: internal and/or external
- Maternal death if ectopic ruptures
  - Nine percent of pregnancy-related deaths
  - Leading cause of maternal death in the first trimester
- Sterility if tube(s) are damaged or surgically removed

110.6 Pearls and Pitfalls

- Pearls
  - Following the algorithm despite the patient being asymptomatic can help avoid missed ectopic pregnancies.
  - Cervical ectopic pregnancy can be difficult to distinguish from a spontaneous abortion. If the patient is aborting, the ultrasound findings should change quickly and the patient should have vaginal bleeding.
  - Double decidual reaction versus pseudogestational sac can be a very subtle distinction and should not be made by emergency physicians; hence, the statement made earlier, requiring a yolk sac as the earliest definitive sign of an intrauterine pregnancy.

  - Fibroids, bicornuate uterus, and eccentrically located normal pregnancy can all appear similar to a cornual pregnancy.
  - The majority of times when the patient is pregnant, has an empty uterus on sonographic imaging, and has vaginal bleeding, the final diagnosis is still unknown because it could still be an early normal pregnancy or an ectopic pregnancy.
- Pitfalls
  - Failing to obtain a pregnancy test in all reproductive-age women who have not undergone a hysterectomy
  - Failing to identify subtle signs of ectopic pregnancies

Reference

1. Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill Professional; 2007.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Philadelphia: Churchill Livingstone; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Stead LG, Behera SR. Ectopic pregnancy. J Emerg Med. 2007;32(2):205–6.

Ultrasonography for Hydatidiform Mole

111

L. Connor Nickels

Definition: molar pregnancy = hydatidiform mole = anomalous growth of trophoblastic tissue

- Complete: 46,XX or 46,XY
  - Completely paternal in origin
  - Contains no fetal tissue
  - Most recognizable by clinical symptoms
- Incomplete (partial mole): 69,XXX or 69,XXY
  - Maternal and paternal in origin
  - Contains fetal tissue
  - More subtle clinical presentation

111.1 Indications

- Clinical presentation of any combination of the following:
  - Vaginal bleeding
  - With or without vomiting, persistent hyperemesis gravidarum
  - High blood pressure
  - Uterine size large for dates
- Serum human chorionic gonadotropin (hCG) very elevated (more than would be consistent for dates)
  - Usually greater than 100,000 mIU/mL

111.2 Contraindications

- Conservative management

111.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear (abdominal) and endocavitary (transvaginal)
- Gel
- Skilled ultrasound operator
- Endocavitary probe covers
- Pelvic setup (speculum and cultures)
- Lab work: serum quantitative hCG, hemoglobin, group and Rh
  - May add additional laboratory tests depending on the stability and symptoms of the patient

111.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal imaging
  1. Place the patient in the supine position.
  2. Ideally the bladder will be full for a good acoustic window.
  3. Using a curvilinear probe, 3.5–5.0 MHz, begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  4. Rotate the probe to a transverse position (counterclockwise), with the probe indicator toward the patient's right. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Any signs of an intrauterine pregnancy should be clearly identified and measured.
  6. Sonographic findings of a molar pregnancy include:
     - Uterus filled with heterogeneous material.
       - Echogenic material interspersed with anechoic areas known as the "snowstorm" is the most common appearance of a mole (Fig. 111.1).

L.C. Nickels, MD, RDMS
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_111

    - Can be confused with missed abortion or fibroid, underscoring the importance of the serum hCG because this will be markedly elevated, which would make the alternative diagnoses less likely.
  - Theca lutein cysts are present in the adnexa.
    - Multiple cysts that occur with trophoblastic disease, multiple pregnancies (e.g., twins, triplets, quadruplets), and induced ovulation.
7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
8. Any abnormalities should be noted.
9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.

- Transvaginal imaging
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use an endocavitary probe, 5–7.5 MHz.
  4. Repeat the same steps as for transabdominal imaging previously discussed (#2).
  5. After scanning through the uterus in sagittal and transverse planes, the adnexa should be scanned in the same format bilaterally, looking for any abnormalities, as mentioned previously.
  6. Again, any abnormalities should be noted:
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) and/or ovaries
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
- The procedure is the same for all pelvic ultrasounds, transabdominal and transvaginal imaging, because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be the first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

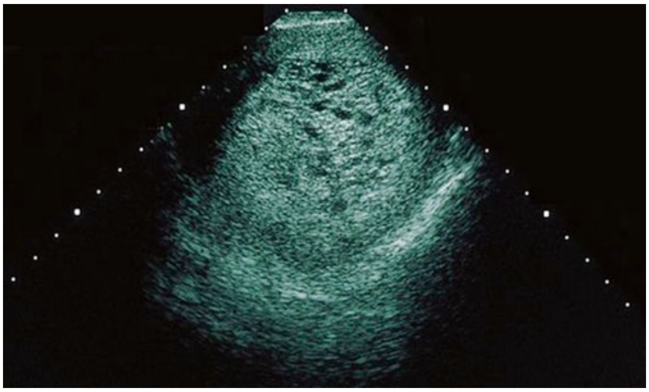

Fig. 111.1 Uterus with "cluster of grapes" represents a molar pregnancy (With kind permission from Springer Science+Business Media: Swisher E, Greer B, Montz FJ, Stenchever M. Chapter 14. In: *Atlas of Clinical Gynecology*. Vol. 4. 2002)

111.5 Complications

- Multiple complications can occur and include the following:
  - Invasive mole: when a hydatidiform mole recurs after a dilation and curettage and, subsequently, invades the muscle of the uterus.
    - Molar tissue is extremely vascular and necessitates an ultrasound with color flow to visualize any residual invasive tissue.
  - Choriocarcinoma: molar tissue develops into an aggressive malignancy, metastasizing early throughout the body.
    - In this scenario, the tumor takes on a very cystic appearance with an echogenic rim. When this is seen, one should attempt to view the liver as well for heterogeneous appearance consistent with metastasis or consider other imaging tests (more comprehensive ultrasound, computed tomography, or magnetic resonance imaging) as indicated to further assess concerns for malignancy.
    - Extremely sensitive to chemotherapy, but most favorable results are seen when diagnosed and treated early.

111.6 Pearls and Pitfalls

- Pearls
  - An obstetrics consult must be obtained in the emergency department when this diagnosis is made.
    - Patient must have a dilation and curettage to evacuate the mole.
    - Serum hCG must be followed to less than detectable levels because they have a high rate of recurrence (invasive mole) and potential for malignancy (choriocarcinoma).
  - May occur with intrauterine or ectopic pregnancies or after spontaneous abortions or full-term pregnancies
  - Qualitative and quantitative β-hCG results may be falsely negative in the setting of a molar pregnancy secondary to the “high-dose hook effect” found in sandwich immunoassays in which there is an overabundance of antigen. This error can be corrected by diluting the urine or serum sample and repeating the test.
- Pitfalls
  - Many pitfalls with potentially life-threatening consequences surround this diagnosis. Meticulous observation of the entire picture, including clinical presentation, laboratory results, and ultrasound findings, will help to avoid these outcomes.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. New York: Churchill Livingstone; 2007.

Hunter CL, Ladde J. Molar pregnancy with false negative â-hCG urine in the emergency department. West J Emerg Med. 2011;12:213–5.

Lentz GM, Lobo RA, Gershenson DM, et al., editors. Comprehensive gynecology. 6th ed. Philadelphia: Mosby; 2012.

Ma OJ, Mateer JR, Blaivas M. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2008.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Ultrasonography for Blighted Ovum (Anembryonic Gestation)

112

Katrina Skoog Nguyen and L. Connor Nickels

112.1 Indications

- A blighted ovum or anembryonic gestation is a pregnancy in which the embryo never develops in the gestational sac (Fig. 112.1).
- Presents as first-trimester vaginal bleeding and/or pelvic pain or cramping
- Criteria for defining an anembryonic pregnancy:
  - Based, in part, on mean gestational diameter (MGD), which is an average of three orthogonal measurements of the gestational sac
    - MGD greater than or equal to 25mm with no embryo
    - 5 MHz or less
      - MGD greater than or equal to 20 mm without a yolk sac
  - May have abnormally low sac position
- References differ in their measurements depending on the approach, transabdominal versus transvaginal, and the frequency of the probe (e.g., 5 MHz vs. 6.5 MHz or greater).

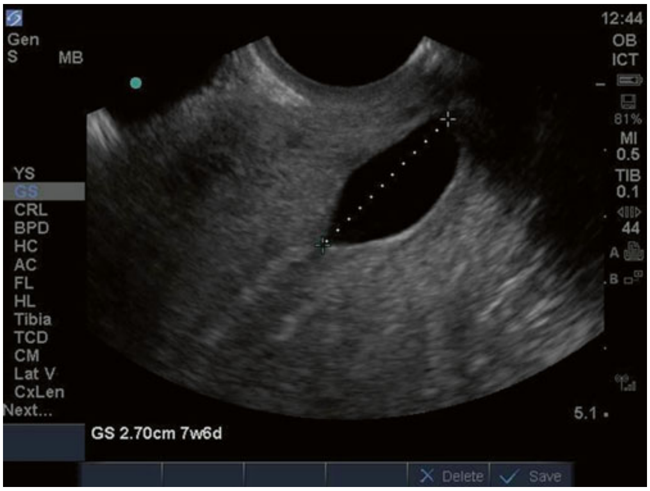

Fig. 112.1 Low-lying empty gestational sac consistent with blighted ovum (Photo courtesy of L. Connor Nickels, MD, RDMS)

K.S. Nguyen, DO • L.C. Nickels, MD, RDMS (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: katrinaskoog@hotmail.com; cnickels@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_112

112.2 Contraindications

- Misinterpreting images as blighted ovum when any of the following could potentially be present:
  - Early intrauterine pregnancy
  - Pseudogestational sac
  - Molar pregnancy
- Discussions with the patient should always involve an obstetrics consult or referral for definitive decisions and treatment.

112.3 Materials and Medications

- Ultrasound machine
- Probes: transabdominal and transvaginal
- Ultrasound gel
- Endocavitary probe covers
- Skilled ultrasound operator
- Laboratory tests: serum quantitative human chorionic gonadotropin (hCG), hemoglobin, group, and Rh
  - May add additional laboratory tests depending on the stability and symptoms of the patient

112.4 Procedure

- Ultrasound machine in obstetrics preset
- Transabdominal
  1. Place the patient in the supine position.
  2. Ideally the bladder will be full for a good acoustic window.
  3. Use a curvilinear probe, 3.5–5.0 MHz.
  4. Begin scanning the patient in a sagittal position to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane, looking for signs of intrauterine pregnancy.
  5. Imaging in the transverse plane should be done in the same fashion.
  6. Any signs of an intrauterine pregnancy should be clearly identified and measured.
     - In this case, gestational sac and any other intrauterine contents
  7. Although sometimes limited transabdominally, an attempt to identify the adnexa should be performed in both planes bilaterally.
  8. Any abnormalities should be noted.
     - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
     - Masses or contained fluid collections outside the uterus
     - Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus
  9. Transvaginal scanning should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
- Transvaginal
  1. Place the patient in the lithotomy position.
  2. Bladder is preferably empty.
  3. Use a transvaginal probe, 5–7.5 MHz.
  4. Repeat the same steps as for transabdominal imaging previously discussed (#2).
- The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.
- Yolk sac should be first sign of definitive intrauterine pregnancy for emergency physicians because the decidual reaction and gestational sac are not 100 % accurate.

112.5 Complications

- Bleeding
- Retained products of conception

112.6 Pearls and Pitfalls

- Pearls
  - See "Contraindications." If following an algorithm, a blighted ovum would fall into indeterminate category and obstetrics consult is recommended in the emergency department.
- Pitfalls
  - It is imperative to inform all pregnant patients that obstetrical ultrasonography performed by emergency physicians is limited and not used to detect fetal health and/or anatomy.

Selected Reading

Cosby K, Kendall J, editors. Practical guide to emergency ultrasound. Philadelphia: Lippincott Williams & Wilkins; 2006.

Doubilet PM, Benson CB, Bourne T, Blaivas M. Diagnostic criteria for nonviable pregnancy early in the first trimester. N Engl J Med. 2013;369(15):1443–51.

Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007.

Morin L, Van den Hof MC. Ultrasound evaluation of first trimester pregnancy complications. J Obstet Gynaecol Can. 2005;161:581–5.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Shah K, Mason C, editors. Essential emergency procedures. Philadelphia: Lippincott Williams & Wilkins; 2008.

Ultrasonography for Threatened, Incomplete, or Compete Abortion

113

L. Connor Nickels and Giuliano De Portu

Key Points

Threatened abortion: closed os, no passage of products of conception (POC).

Inevitable abortion: open os, no POC.

Incomplete abortion: open os, passage of POC.

Complete abortion: closed os, empty uterus (all POC passed).

Missed abortion: fetus has died and the uterus has failed to enlarge any further.

113.1 Indications

- Vaginal bleeding in the setting of early intrauterine pregnancy

113.2 Contraindications

- There are no contraindications to ultrasonography.

113.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear (transabdominal) and endocavitary (transvaginal)
- Gel
- Endocavitary probe covers
- Towels
- Skilled ultrasound operator
- Pelvic setup (speculum and cultures)
- Cardiac monitor and intravenous access if bleeding is significant
- Laboratory tests: serum human chorionic gonadotropin (hCG), hemoglobin, group and Rh, type, and screen

113.4 Procedure

The procedure is the same for all pelvic ultrasounds including transabdominal and transvaginal imaging because this should be performed in a systematic fashion so as to not miss pertinent findings. Therefore, the findings may change, but the examination remains the same.

1. Place the patient in the supine position.
2. Ultrasound machine in obstetrics preset with abdominal probe and gel.
3. Begin scanning the patient in a transabdominal sagittal plane to identify the uterus as it lies in position with the bladder. Scan through the uterus completely in this plane looking for signs of intrauterine pregnancy (gestational sac, yolk sac, fetal pole, fetal cardiac activity).
    - For our purposes as emergency physicians, a yolk sac (at the least) must be seen within the uterus to definitively report an intrauterine pregnancy (Fig. 113.1). The decidual reaction and/or presence of gestational sac is not 100 % accurate.
4. Transabdominal imaging in the transverse plane should be done in the same fashion.
5. If an intrauterine pregnancy is clearly identified on transabdominal scanning, appropriate measurements can be taken to estimate dates:
    - Gestational sac diameter.
    - Yolk sac diameter or crown rump length (CRL).
    - If present, fetal cardiac activity should be recorded by using M mode to obtain a tracing and measure the fetal heart rate (Fig. 113.2).
6. For a complete examination, the adnexa should be scanned in an attempt to identify the ovaries bilaterally.
7. Any abnormalities identified should be noted:
    - Free fluid surrounding the uterus (anterior or posterior cul-de-sac) or ovaries
        - On pelvic imaging or focused assessment with sonography in trauma (FAST) examination

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_113

- Intrauterine contents not consistent with an intrauterine pregnancy and/or not clearly, centrally visualized within the fundus of the uterus
- Any masses or contained fluid collections noted outside the uterus
- Yolk sac or fetal pole with or without fetal cardiac activity seen outside the uterus

8. Transvaginal scanning with the endocavitary probe should be performed if a definitive intrauterine pregnancy is not identified on transabdominal imaging.
9. Transvaginal imaging is performed in the sagittal and transverse planes to identify an intrauterine pregnancy, with bladder preferably empty.
10. Measurements should be taken of any intrauterine findings.
    - Same as #5.
    - CRL >7 mm on transvaginal ultrasound should have fetal cardiac activity.
11. Each adnexa should be scanned in the sagittal and transverse planes.
12. Any abnormalities identified should be noted.
    - Same as #7.
13. The uterus will be empty in the case of a complete abortion (Figs. 113.3 and 113.4).

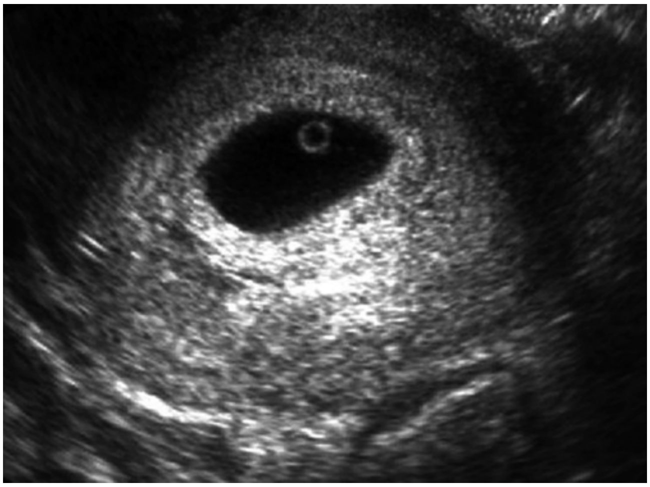

Fig. 113.1 Yolk sac present within the gestational sac with double decidual reaction (With kind permission from Springer Science+Business Media: Buja LM, Chandrasekhar C. Chapter 7: pathology of the breast and female genital tract. In: Krueger GRF, Buja LM, eds. *Atlas of Anatomic Pathology with Imaging*. 2013)

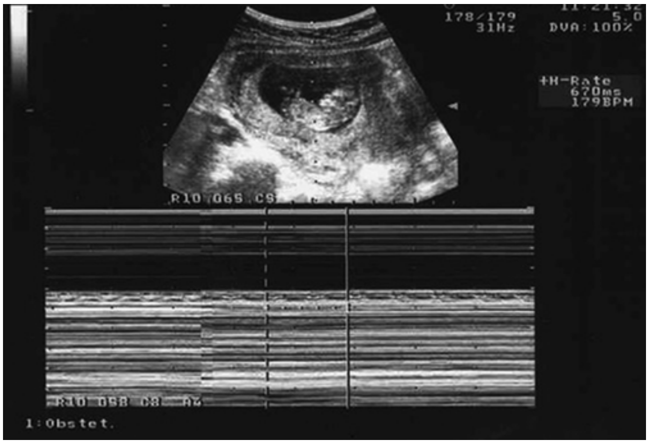

Fig. 113.2 Fetal cardiac activity recorded using M mode to obtain a tracing and measure the fetal heart rate (With kind permission from Springer Science+Business Media: Hanprasertpong T, Phupong V. First trimester embryonic/fetal heart rate in normal pregnant women. *Archives of Gynecology and Obstetrics*. 2006;274(5))

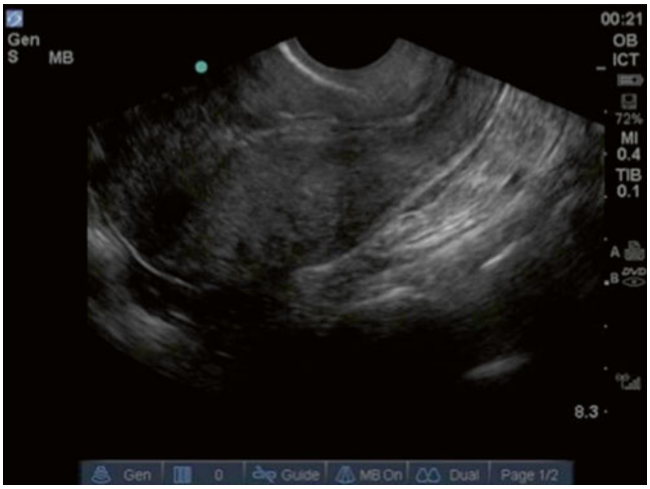

Fig. 113.3 Transvaginal sagittal image of an empty uterus (Photo courtesy of L. Connor Nickels, MD, RDMS)

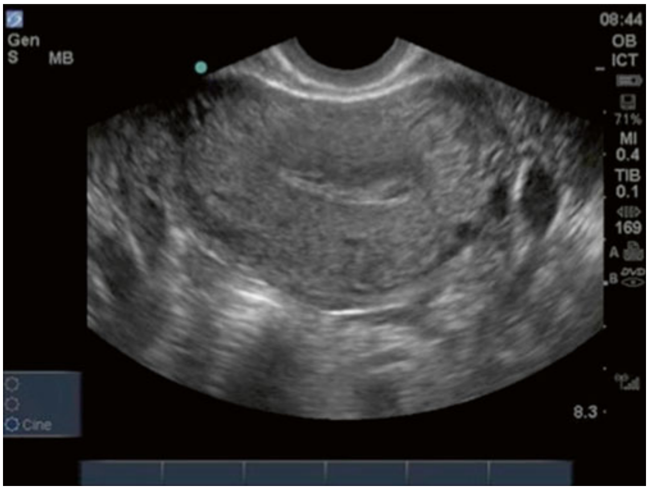

Fig. 113.4 Transvaginal transverse image of an empty uterus (Photo courtesy of L. Connor Nickels, MD, RDMS)

113.5 Complications

- Vaginal bleeding
- Miscarriage
- Rh isoimmunization

113.6 Pearls and Pitfalls

- Pearls
  - Threatened abortion:
    - Threatened abortion is not visible sonographically. It is the presumed diagnosis when a patient presents with vaginal bleeding in the first 20 weeks of pregnancy, an intrauterine pregnancy is found with dates corresponding to the patient's dates, and the cervix is closed.
    - Fetal cardiac activity suggests a better prognosis than those without.
    - Fetal heart rate less than 120 bpm suggests impending fetal death (only 6 % survival).
  - Completed abortion:
    - There are only three scenarios in which this diagnosis can confidently be made in the emergency department:
      1. Intact gestation is passed and identified in the emergency room.
      2. Ultrasound shows an empty uterus in the setting of a prior known intrauterine pregnancy.
         - Small internal echoes within the uterus may represent blood rather than retained products of conception, but this determination should be made in consultation with obstetrics.
      3. Negative pregnancy test result in the setting of a prior known intrauterine pregnancy.
    - In all other cases, the quantitative β-hCG must be followed until it is less than 2 mIU/mL.
- Pitfalls
  - Treating threatened abortion as fetal demise or abortion – many of these pregnancies actually go on to completion with a normal fetus.
  - If there is any concern for alternative diagnoses in any of the abortion types (i.e., ectopic gestation), obstetrics consult should be obtained.
  - Failing to order a quantitative serum hCG to rule out other possibilities (e.g., molar pregnancy) and to ensure that it is trending downward (if others to compare).
  - Failing to establish follow-up for an incomplete abortion to ensure appropriate management. The patient may eventually require a dilation and curettage if products do not pass naturally.

Selected Reading

Doubilet PM, Benson CB, Bourne T, Blaivas M. Diagnostic criteria for nonviable pregnancy early in the first trimester. N Engl J Med. 2013;369(15):1443–51.

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Orlando: Churchill Livingstone; 2007.

Ma OJ, Mateer JR, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw-Hill Professional; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

Ultrasonography for Placenta Previa

114

L. Connor Nickels and Giuliano De Portu

Placenta previa is a condition in which the placenta covers the internal cervical os. It is the leading cause of antepartum hemorrhage. There are four grades of the condition:

- Complete: Placenta completely covers the internal os.
- Partial: Placenta partially covers the internal os.
- Marginal : Lower margin of placenta reaches the internal os but does not cover it (within 3 cm).
- Low lying: Lower margin of placenta is located in the lower uterine segment but does not reach the internal os (Fig. 114.1).

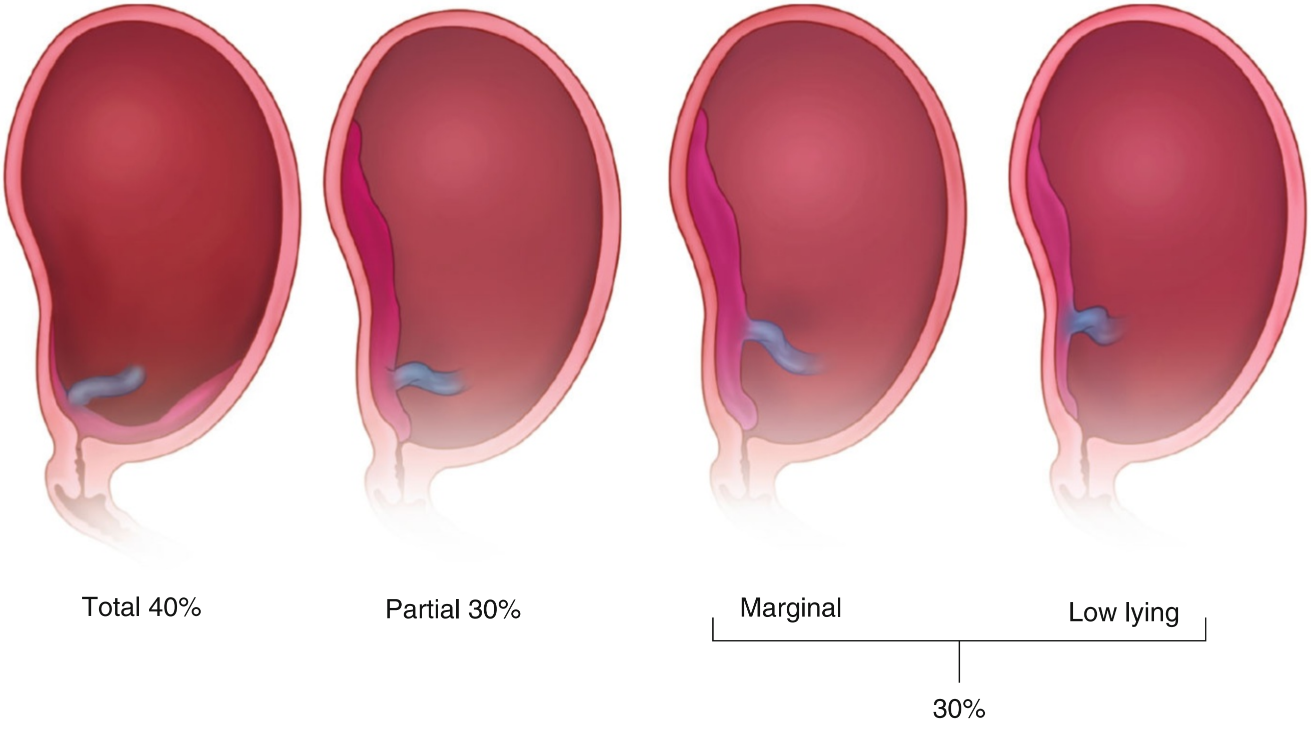

Fig. 114.1 Varying degrees of placenta previa

L.C. Nickels, MD, RDMS (✉) • G. De Portu, MD
Department of Emergency Medicine, University of Florida Health
Shands Hospital, Gainesville, FL, USA
e-mail: cnickels@ufl.edu; gdeportu@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_114

114.1 Indications

- A patient in the second half of pregnancy who presents to the emergency room with bright red, painless vaginal bleeding OR premature labor should be evaluated for this condition.

114.2 Contraindications

- Absolute:
  - There are no absolute contraindications to transabdominal ultrasonography.
- Relative:
  - Transvaginal ultrasound should be carried out with caution, so as to avoid hemorrhage.

114.3 Materials and Medications

- Ultrasound machine
- Probes: curvilinear or phased array, endocavitary
- Sterile gel
- Endocavitary probe covers
- Towels
- Sterile speculum (see previously)
- Cardiac monitor, two large-bore intravenous needles, laboratory tests

114.4 Procedure

1. Place the patient in the supine position.
2. Ultrasound machine with abdominal probe and gel.
3. Begin scanning the patient in a transabdominal sagittal position (Fig. 114.2) to determine the placental position and whether it is lying in the lower uterine segment.
4. The patient should then be scanned in a transverse fashion to further evaluate the exact position of the placenta.
5. If the placenta appears to lie in the lower uterine segment, the patient should be scanned in the oblique plane as well.
6. The bladder should initially be full for best visualization. However, if the placenta appears to reside low or lie over the internal os, the scanning should be repeated after the patient has voided.
   - An overdistended bladder may create the appearance of a placenta previa. The anterior wall of the uterus is compressed against the posterior wall by the distended bladder, shortening the distance between the placenta and the internal os (Fig. 114.3).
7. The following ultrasonographic findings exclude placenta previa [1]:
   - Direct apposition of the presenting part of the fetus and the cervix without space for interposed tissue
   - Presence of amniotic fluid between the presenting part of the fetus and the cervix, without the presence of placental tissue
   - Distance of greater than 2 cm between the inferior aspect of the placenta
   - Indirect visualization of the internal cervical os
8. If placenta previa cannot be ruled out with transabdominal ultrasound, the patient should then be scanned transvaginally with the endocavitary probe because this is more sensitive for diagnosing placenta previa.
9. A sterile speculum examination should be performed before transvaginal scanning to assess the cervix and ensure there are no presenting parts or bulging membranes. Transvaginal imaging is contraindicated if the patient has ruptured or bulging membranes.
10. The probes should be swapped out and a sterile cover placed on the endocavitary probe with gel inside the cover and sterile gel on the outside of the cover.
11. The endocavitary probe should be inserted into the vaginal canal, ensuring that caution is taken to stay off the cervix and distal to it, keeping the cervix in view on the screen.
12. The patient should be scanned in both the sagittal and the transverse planes to assess the inferior margin of the placenta.

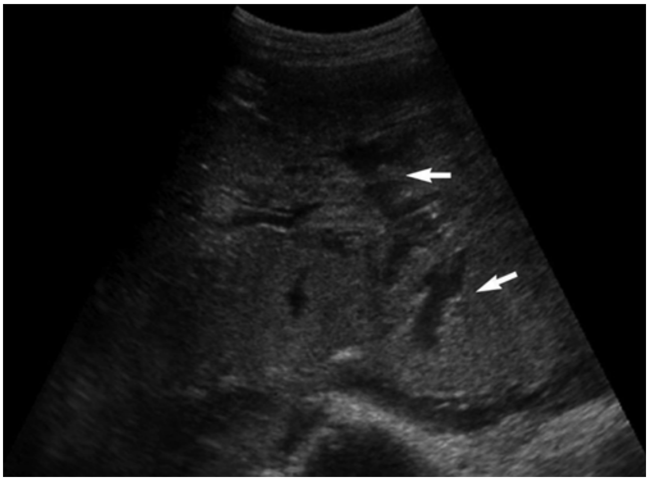

Fig. 114.2 Transabdominal ultrasound of placenta previa. Note loss of the decidual interface between the placenta and the myometrium on the lower part of the uterus and multiple intraplacental lacunae (*arrows*) (Reproduced with permission from *J Korean Med Sci*. 2010; 25(4):651–5)

13. If the inferior margin appears to be located near the internal os, then the distance should be measured.
    - Steps 8 through 13 should only be performed by the emergency physician if he or she feels confident in this ultrasound skill and there is no obstetrics available. If obstetrics is available, these steps should be performed in conjunction with them or by them to ensure the best outcome for the patient.

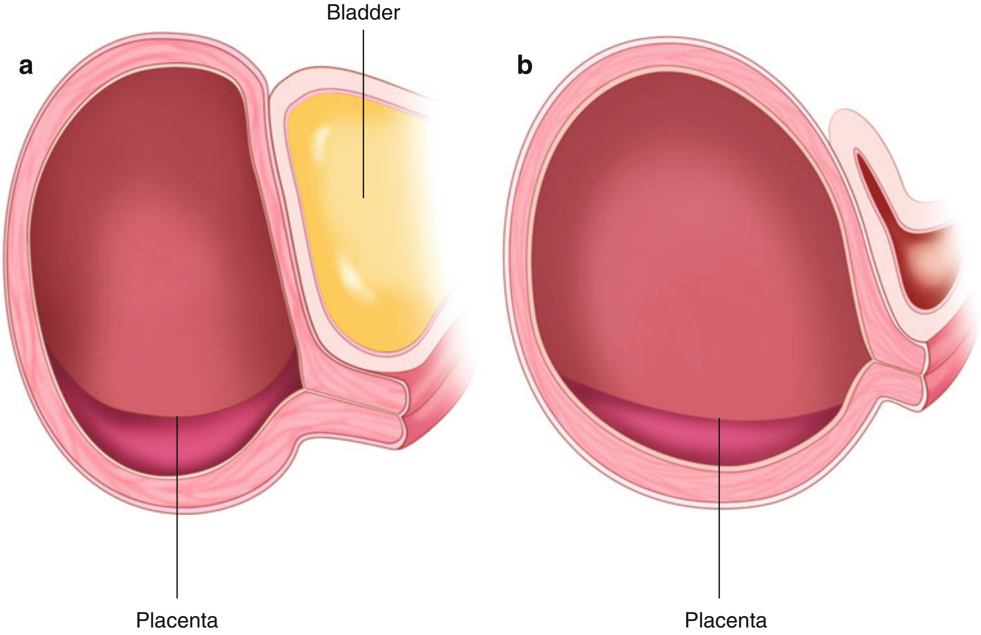

Fig. 114.3 (**a**) Overdistended bladder creating the appearance of placenta previa, (**b**) empty bladder showing the more accurate position of the placenta

114.5 Complications

- Bleeding:
  - May range from self-limited to life-threatening hemorrhage
- Maternal and/or fetal distress or death

114.6 Pearls and Pitfalls

- Pearls
  - Risk factors include:
    - Multiparity
    - Multiple pregnancy
    - Advanced maternal age
    - Previous placenta previa
    - Cigarette smoking
    - Cocaine abuse
    - Hypertension
    - Previous cesarean delivery/uterine surgeries
  - Myometrium contraction: can mimic placenta previa by temporarily displacing the placenta in the lower uterine segment.
  - If ultrasound capabilities are not available, but the patient is in the second half of pregnancy and having vaginal bleeding, do not perform a digital cervical examination.
- Pitfalls
  - Digital examination should be avoided because this may precipitate life-threatening hemorrhage and/or death.
  - Always consult obstetrics as soon as possible if this is suspected or known and the patient is symptomatic. Not consulting obstetrics could be detrimental to the mother and fetus.
  - Gentle sterile speculum examination should be done only if an obstetrician is not available. This is to ensure the bleeding is coming from the cervix. If placenta previa is suspected or known and obstetrics are available, then abdominal ultrasound evaluation alone should be sufficient for the examination of the patient.

Reference

1. American College of Radiology. Role of imaging in second and third trimester bleeding. In: ACR Appropriateness Criteri. Reston: American College of Radiology; 2001.

Selected Reading

Gabbe SG, Niebyl JR, Galan HL, et al., editors. Obstetrics: normal and problem pregnancies. 5th ed. Orlando: Churchill Livingstone; 2007.

Ma OJ, Mateer J, Blaivas M, editors. Emergency ultrasound. 2nd ed. New York: McGraw Hill Professional; 2007.

Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Sanders RC, Winter T, editors. Clinical sonography: a practical guide. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2007.

115

Vaginal Delivery

Umarfarook Javed Mirza, Christopher Shields,
and Muhammad Waseem

115.1 Indications

- Inevitable delivery of the fetus (cervix fully dilated at 10 cm and fully effaced 1 mm with crowning of the head)

115.2 Contraindications

- Absolute:
  - Indications for emergent cesarean section (C-section)
  - Prolapsed cord
  - Prior C-section with classic vertical incision
  - Placenta previa – complete/partial
  - Breech presentation – footling
- Relative:
  - Placenta previa – marginal/low lying
  - Breech presentation – complete/incomplete/frank

115.3 Materials and Medications (Fig. 115.1)

- 4 Crile clamps or Kelly clamps
- 1 Mayo scissor curved and 1 Mayo scissor straight
- 2 sponge forceps
- 2 towel clamps
- 1 Mayo-Hegar needle holder
- 1 mouse-tooth forceps
- Suction bulb
- Betadine (povidone-iodine)
- Umbilical cord clamp (Fig. 115.2)
- Incubator warmer (Fig. 115.3)
- Most importantly, help.

U.J. Mirza, DO
Department of Emergency Medicine,
Baylor University Medical Center, Dallas, TX, USA
e-mail: Umarfarook.Mirza@gmail.com

C. Shields, MD • M. Waseem, MD (✉)
Department of Emergency Medicine,
Lincoln Medical and Mental Health Center, New York, NY, USA
e-mail: cpatrickshields@hotmail.com; waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_115

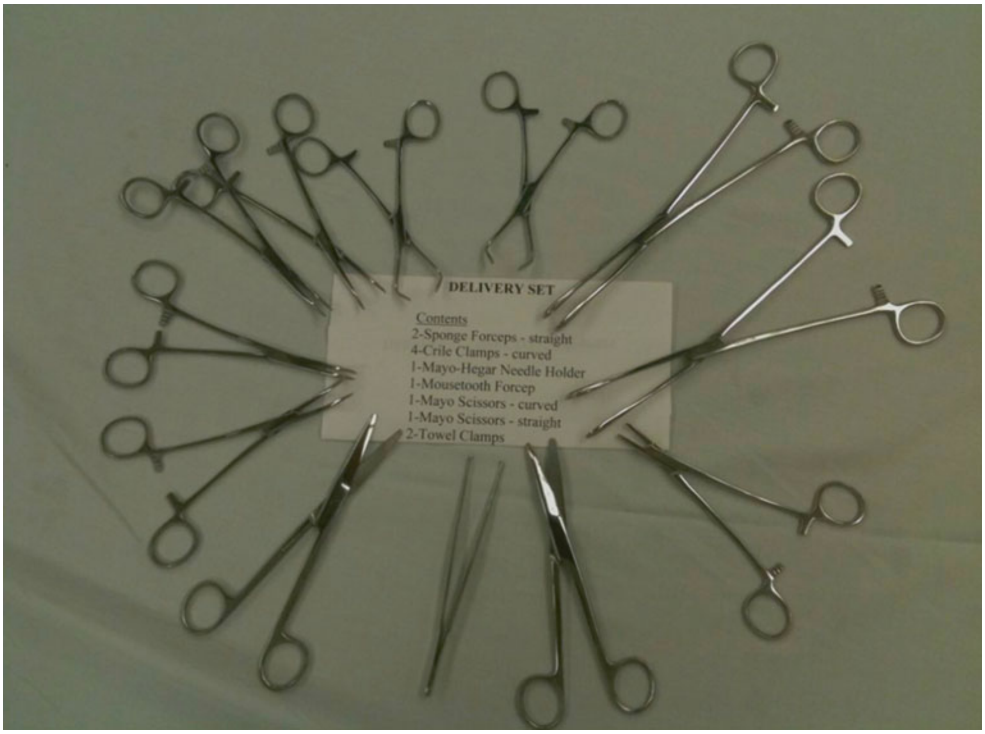
DELIVERY SET
Contents
2-Sponge Forceps - straight
4-Crile Clamps - curved
1-Mayo-Hegar Needle Holder
1-Mousetooth Forcep
1-Mayo Scissors - curved
1-Mayo Scissors - straight
2-Towel Clamps

Fig. 115.1 Equipment: delivery set

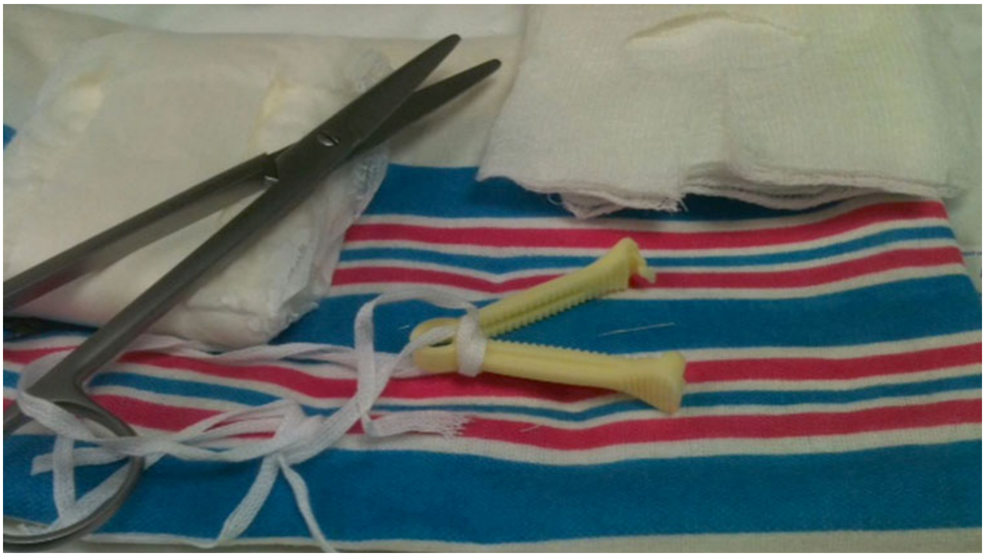

Fig. 115.2 Equipment: umbilical clamp

Fig. 115.3 Equipment: incubator warmer

115.4 Procedure

1. Determine the fetal presentation (breech or cephalic) with Leopold's maneuvers or bedside ultrasound.
2. If breech, follow the breech pathway.
3. Perform a vaginal examination.
4. Check for cord prolapse.
5. Check effacement.
6. Check dilation.
7. Prepare the site.
8. Position the mother in the extreme lithotomy position.
9. Deliver the head in a controlled manner.
10. Check to see if the cord is wrapped around the neck. If present, follow cord presentation pathway.
11. Deliver the anterior shoulder with downward traction of the head.
12. Check for shoulder dystocia. If present, follow the McRobert-Rubin maneuvers (see Chap. 116).
13. Deliver the posterior shoulder by gently pulling the trunk upward.
14. Suction the airway.
15. Clamp the cord 5 cm from the umbilicus in two places and cut.
16. Obtain Apgar score of the baby. Initiate resuscitation if required.
17. Keep the baby warm in the incubator.
18. Clamp the cord again closer to the vaginal opening.
19. Use this clamp to help deliver the placenta. Use gentle controlled traction on the clamp with one hand while placing the other hand suprapubically to push the uterus upward.
20. If the placenta does not deliver easily, stop and wait a few minutes to allow it to come away from the uterine wall naturally and then try again.
21. Placenta must be fully intact. If not, check for retained intrauterine placental products, and manually remove them; retained placenta may cause postpartum hemorrhage.
    - Massage the fundus externally to prevent uterine atony, but only after all products have been removed.
22. Repair any vaginal or cervical lacerations.

115.5 Complications

- Postpartum hemorrhage due to uterine atony or retained products
- Uterine inversion
- Rectal or urethral injuries
- Shoulder dystocia
- Meconium aspiration

115.6 Pearls and Pitfalls

- Pearls
  - Postpartum hemorrhage has a high incidence during the first hour postpartum.
  - Be aware of possible second delivery.
  - Oxytocin or Methergine (methylergonovine maleate) may be used after delivery of the placenta; both help increase uterine contractions and decrease postpartum hemorrhage.
- Pitfalls
  - Rushing delivery of the head/shoulders can lead to trauma to the mother/baby.
  - Rushing delivery of the placenta, when the placenta is not ready to detach, may cause placental or cord tearing or uterine inversion. Remember: it may take up to 20 min for the placenta to detach from the uterus.

Selected Reading

Desai S, Henderson SO, Mallon WK. Labor and delivery and their complications. In: Marx J, Hockberger R, Walls R, editors. Rosen's emergency medicine: concepts and clinical practice. 7th ed. Philadelphia: Mosby; 2010.

Liao JB, Buhimschi CS, Norwitz ER. Normal labor: mechanism and duration. Obstet Gynecol Clin North Am. 2005;32:145–64.

Norwitz ER, Robinson JN, Repke JT. Labor and delivery. In: Gabbe SG, Niebyl JR, Simpson JL, editors. Obstetrics: normal and problem pregnancies. 4th ed. Philadelphia: Saunders; 2001. p. 353–94.

Shoulder Dystocia Management

116

Irina Fox Brennan and Joseph A. Tyndall

116.1 Indications

- Retracting fetal head ("turtle sign")
- Failure of anterior shoulder delivery following gentle downward traction
- Difficult face-chin delivery
- Failure of shoulders to descend
- Factors that signal the need for cesarean section:
  - Complete/partial placenta previa
  - Cord prolapse
  - Brow/face presentation
  - Non-reassuring fetal heart rate
  - Extensive uterine surgery and previous classic uterine incision

116.2 Contraindications

- Mother with diabetes and fetus larger than 4500 g
- Mother without diabetes and fetus larger than 5000 g
- History of shoulder dystocia with prior birth

116.3 Materials and Medications

- No specialized materials/equipment required.
- Team coordination is of utmost importance.

I.F. Brennan, MD • J.A. Tyndall, MD, MPH (✉)
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: ibrennan@ufl.edu; tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_116

116.4 Procedure

1. *Accurate documentation*: Designate a team member to document the progress of labor, the position and rotation of the infant's head, the time of delivery for the head and body, the presence of episiotomy, anesthesia requirements, the duration of shoulder dystocia, whether the shoulder is anterior or posterior at the time of delivery, the onset/duration/results of maneuver(s) performed, and Apgar scores.

   The HELPERR mnemonic is commonly employed:
2. *Help recruitment*: Involve anesthesiology, pediatric resuscitation, and obstetric/gynecology colleagues either as part of advance preparations or via activation of protocol during labor.
3. *Episiotomy consideration*: Not required but might be of benefit if more space is needed for rotation maneuvers.
   - Episiotomy alone will not release the impacted shoulder.
4. *Legs positioned for the McRoberts maneuver*: Position the woman's thighs onto her abdomen by abducting and flexing her hips. Have two assistants help her, one on either side. This increases the functional size of the bony pelvis by rotating the pubic symphysis toward the mother's head, hence aiding delivery (Fig. 116.1).

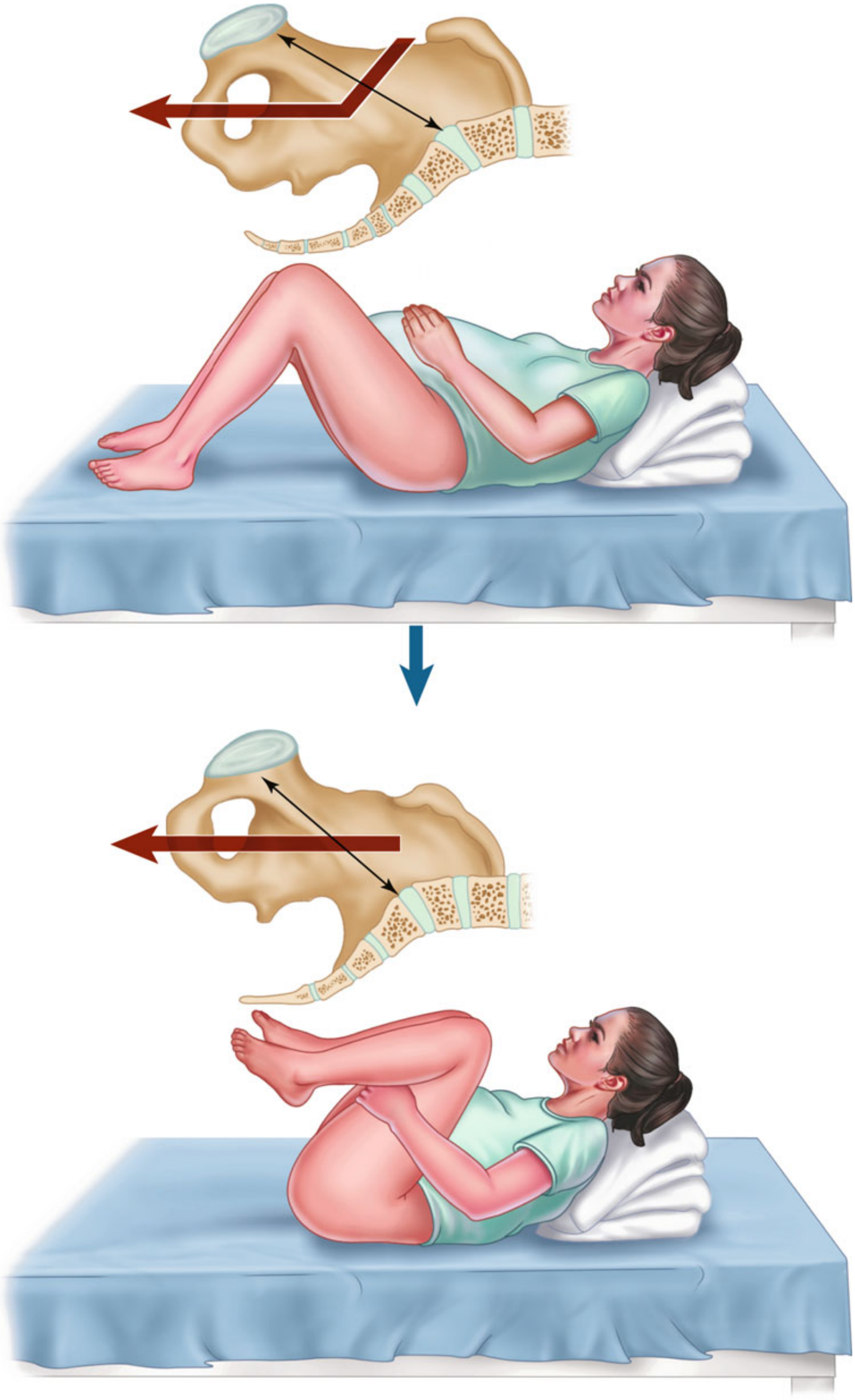

Fig. 116.1 McRoberts maneuver: hyperflexion of the maternal thighs against the abdomen

5. *Pressure suprapubically* (*Rubin I maneuver*): While continuing downward traction, have a colleague place his or her hand suprapubically on the fetal anterior shoulder and apply pressure down and laterally on its posterior aspect in 30-second increments. This allows the fetal shoulders to enter the pelvis in an oblique fashion (pelvic inlet is widest in the transverse plane) (Fig. 116.2).
6. *Enter maneuver* (*internal rotation*): Rotate the anterior shoulder into the oblique plane and under the maternal symphysis.
    - It may be necessary to push the fetus slightly up into the pelvis in order to successfully perform this maneuver.
        (a) *Rubin II maneuver*: Insert one hand vaginally behind the fetal anterior shoulder and rotate the shoulder toward the chest. This will reduce the diameter of the fetal shoulder girdle and facilitate delivery (Fig. 116.3).
        (b) *The Woods corkscrew maneuver*: Place one hand at the front of posterior shoulder and push upward gently. This can be combined with the Rubin II maneuver (Fig. 116.4).
        (c) *The reverse Woods corkscrew maneuver*: Place one hand behind the fetal posterior shoulder and rotate in the direction opposite to that of the Woods corkscrew maneuver to adduct the shoulder.
7. *Remove the posterior arm*: Flex the fetal elbow and deliver the forearm in a sweeping motion over the fetal anterior chest wall. This allows the fetus to drop into the sacral hollow, freeing the anterior shoulder impaction.
    - Avoid grasping and pulling on the fetal arm directly because it may fracture the fetal humerus.
8. *Roll the patient* (*Gaskin maneuver*): Position the patient on all fours (this acts as an upside-down McRoberts maneuver). Continue gentle traction. The turning itself, as well as gravity, will often dislodge the impacted shoulder.

    Last-resort maneuvers:
9. *Deliberate fracture of the fetal clavicle*: Apply direct pressure upward in the middle of the clavicle to reduce the shoulder-to-shoulder distance.

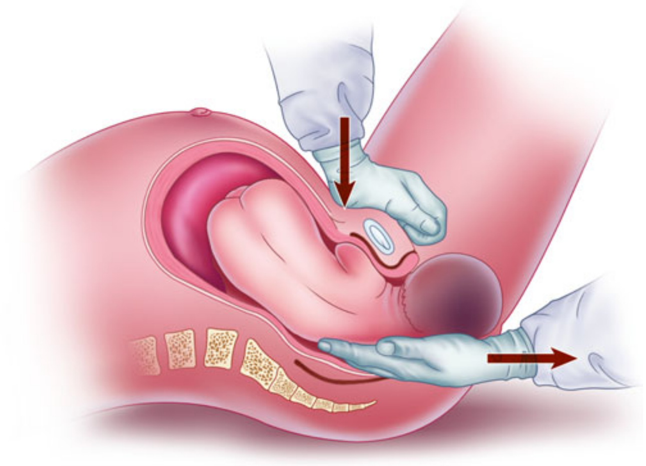

Fig. 116.2 Ruben I maneuver: suprapubic pressure is directed at the anterior shoulder

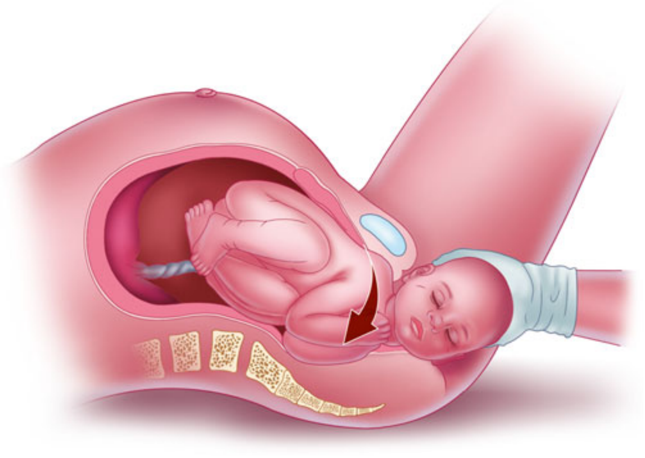

Fig. 116.3 Ruben II maneuver: pressure is applied to the most accessible part of the fetal shoulder and rotated toward the chest

Fig. 116.4 The Woods corkscrew maneuver: pressure is applied to the clavicle of the posterior arm, enabling rotation and dislodgement of the anterior shoulder

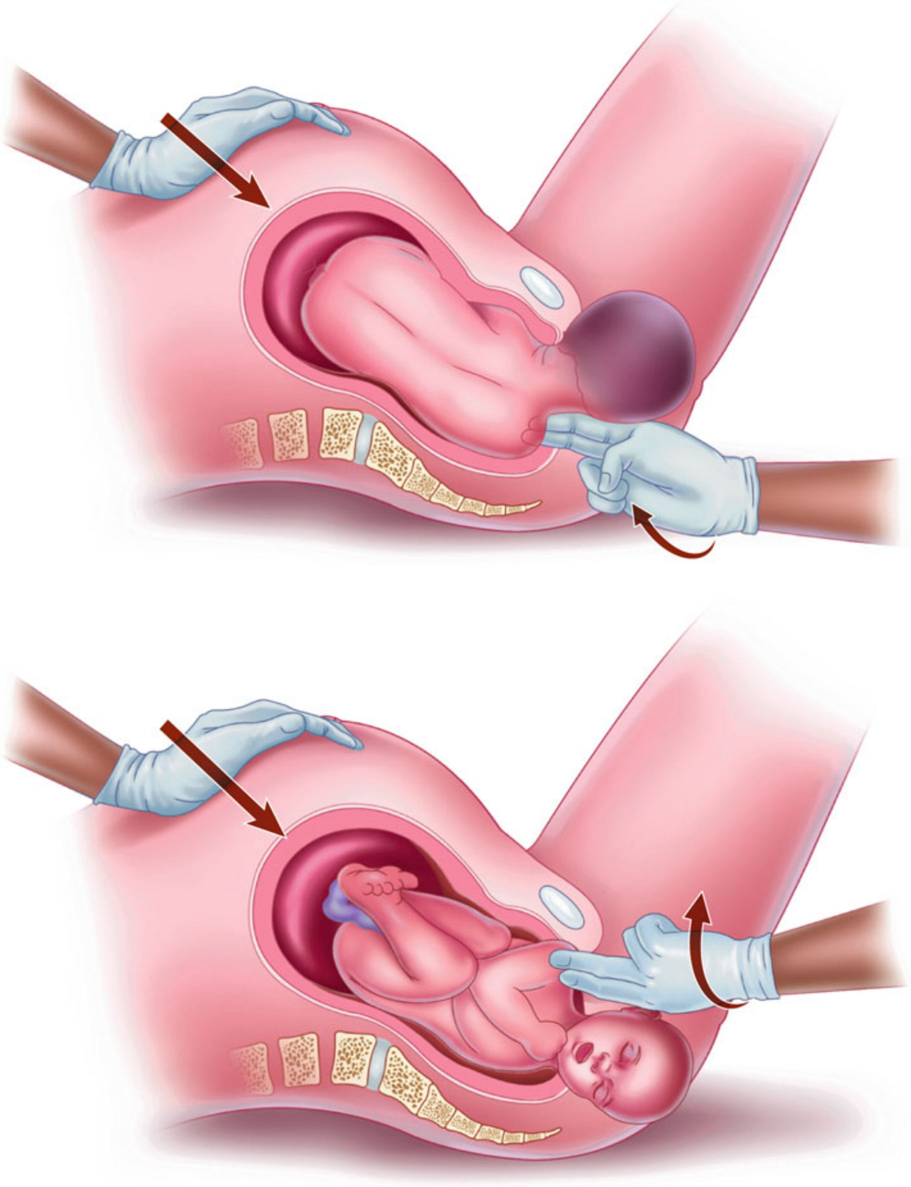

10. *Zavanelli maneuver*: Rotate the fetal head into the direct occiput anterior position, flex, and push back into the birth canal. Continue holding upward pressure until cesarean delivery. Ensure the operating and anesthesiology teams are present (Fig. 116.5).
    - Do not perform this maneuver if the nuchal cord has been clamped or cut.
11. *Symphysiotomy*: Place the patient in the exaggerated lithotomy position, insert a transurethral catheter, displace the urethra laterally, and separate the cephalad portion of fibrous cartilage of the symphysis pubis under local anesthesia with either a scalpel blade or Kelly clamp.
    - This technique is associated with significant maternal morbidity, including bladder neck injury and infection. It is truly the last resort and should be used only when all other methods have failed and cesarean delivery is unavailable.

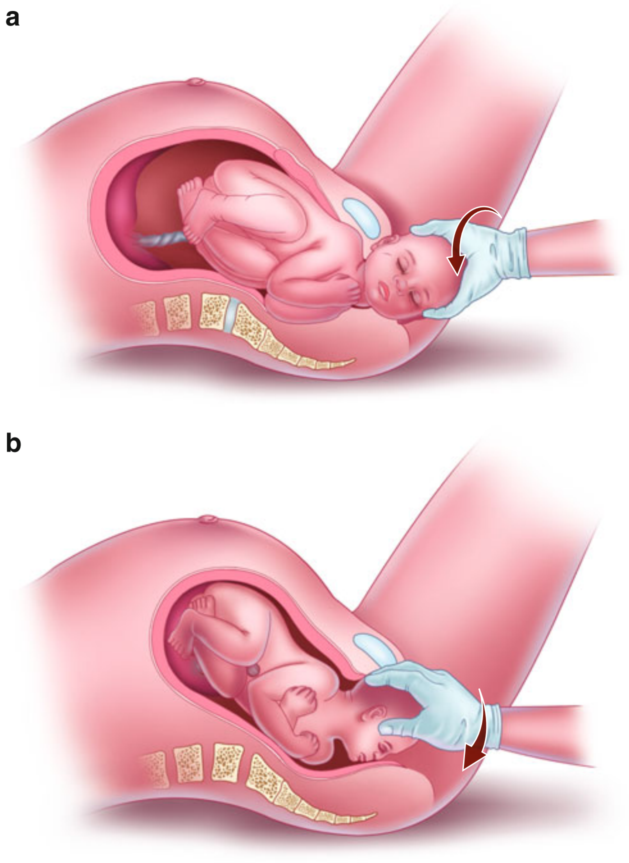

Fig. 116.5 Zavanelli maneuver: the fetal head is rotated into the direct occiput anterior position, flexed, and pushed back into the birth canal. (**a**) Rotation of the head to the occiput anterior position. (**b**) Replace head with constant firm pressure on occiput with palm of hand flexing head

116.5 Complications

- Maternal:
  - Postpartum hemorrhage
  - Fourth-degree laceration/soft tissue damage to cervix and vagina
  - Uterine rupture
  - Bladder atony
  - Symphyseal separation with or without femoral neuropathy (transient)
  - Sacroiliac dislocation
- Fetal:
  - Brachial plexus palsies (Erb, Klumpke): most resolve within 6–12 months; permanent injury occurs in less than 10 % of patients
  - Clavicular and/or humeral fractures
  - Ischemia/hypoxia/asphyxia due to umbilical cord compression with or without neurological damage
  - Significant fetal acidosis, with pH drop of 0.04 U/min between delivery of the head and trunk
  - Death

116.6 Pearls and Pitfalls

- Pearls
  - Many cases of shoulder dystocia and birth trauma are encountered in the absence of risk factors and in non-macrosomic infants. Shoulder dystocia is largely neither predictable nor preventable.
  - Presence of risk factors for shoulder dystocia should prompt advance preparation in anticipation of a difficult delivery. Risk factors include maternal diabetes, obesity, multiparity, advanced maternal age, prolonged pregnancy, macrosomia (fetal weight >4,500–5,000 g), male fetal gender, and prior birth complicated by shoulder dystocia.
- Pitfalls
  - Do not cut and clamp the nuchal cord if at all possible in order to avoid fetal hypoxia and hypotension should shoulder dystocia arise.
  - Labor induction or prophylactic cesarean delivery of macrosomic fetuses has not decreased rates of shoulder dystocia.
  - Avoid fundal pressure in shoulder dystocia because it has been shown to increase the risk of permanent neurological damage and uterine rupture.

Selected Reading

Acker DB. A shoulder dystocia intervention form. Obstet Gynecol. 1991;78:150–1.

Athukorala C, Middleton P, Crowther CA. Intrapartum interventions for preventing shoulder dystocia. Cochrane Database Syst Rev. 2006;4:CD005543.

Baxley EG, Gobbo RW. Shoulder dystocia. Am Fam Physician. 2004;69:1707–14.

Chauhan SP, Gherman R, Hendrix NW, Bingham JM, Hayes E. Shoulder dystocia: comparison of the ACOG practice bulletin with another national guideline. Am J Perinatol. 2010;27:129–36.

Gherman RB, Chauhan S, Ouzounian JG, Lerner H, Gonik B, Goodwin TM. Shoulder dystocia: the unpreventable obstetric emergency with empiric management guidelines. Am J Obstet Gynecol. 2006;195:657–72.

Gottlieb AG, Galan HL. Shoulder dystocia: an update. Obstet Gynecol Clin North Am. 2007;34:501–31.

Hoffman MK, Bailit JL, Branch DW, et al. Consortium on safe labor. A comparison of obstetric maneuvers for the acute management of shoulder dystocia. Obstet Gynecol. 2011;117:1272–8.

Patterson DA, Winslow M, Matus CD. Spontaneous vaginal delivery. Am Fam Physician. 2008;78:336–41.

Sokol RJ, Blackwell SC, American College of Obstetricians and Gynecologists. Committee on Practice Bulletins–Gynecology. ACOG practice bulletin no. 40: shoulder dystocia. November 2002 (replaces practice pattern no. 7, October 1997). Int J Gynaecol Obstet. 2003;80:87–92.

Breech Delivery in the Emergency Department

117

Kristin Stegeman, Sapnalaxmi Amin, Anton A. Wray,
and Joseph A. Tyndall

Breech presentation: the buttocks enter the pelvis before the head (Fig. 117.1)

- Frank (extended):
  - Hips flexed and knees extended, buttocks presenting
  - Most common
- Incomplete/complete (flexed):
  - One or both hips and knees flexed, buttocks presenting
- Footling:
  - One or both hips and knees extended, foot presenting

The emergent delivery of a breech baby is one of the most challenging situations for an emergency physician.

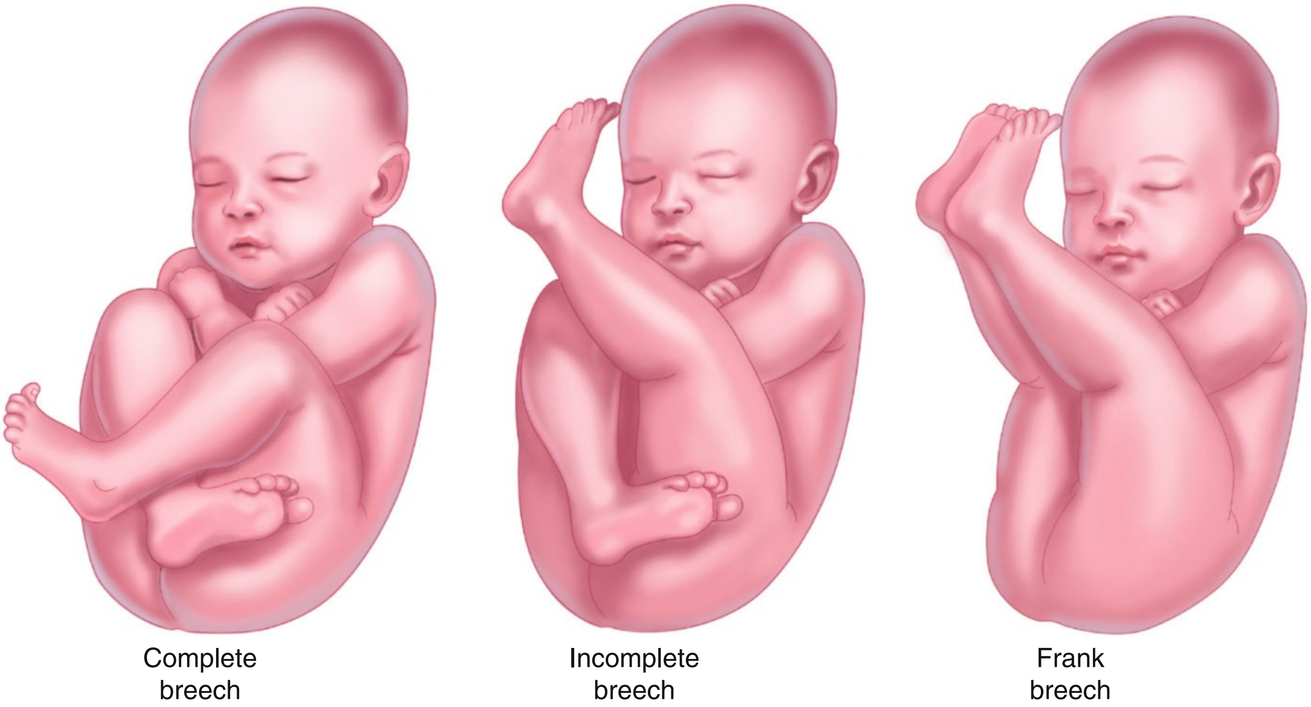

Fig. 117.1 Variations of the breech presentation

K. Stegeman, MD • A.A. Wray, MD (✉)
Department of Emergency Medicine, The Brooklyn Hospital Center, New York, NY, USA
e-mail: kristin.stegeman@gmail.com; anw9071@nyp.org

S. Amin, MD
Department of Family Medicine/Urgent Care,
Bayside Urgent Care Center, Clearwater, FL, USA
e-mail: sapna_amin59@yahoo.com

J.A. Tyndall, MD, MPH
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_117

117.1 Indications

- Inevitable delivery of fetus with complete or frank breech presentation
- Absence of vertex presentation

117.2 Contraindications

- Footling presentation (increased risk of cord prolapse and entrapment of the after-coming head).
- There is adequate time to transfer the mother safely to labor and delivery in the knee/chest position (administer subcutaneous terbutaline before transfer).

117.3 Materials and Medications

- Supplemental oxygen for mother
- Ultrasound machine
- Piper forceps
- Sterile towels/gloves
- Betadine (povidone-iodine) or another antiseptic preparing solution
- Sterile lubricant (Surgilube)
- Instruments to cut the umbilical cord/perform episiotomy/resuscitate baby (Kelly clamps/scissors/#10 scalpel/bulb suction)
- Available emergency department staff
- Pediatric, obstetrics, and anesthesia practitioners

117.4 Procedure

1. Assess the health of the mother and the baby (vitals/physical examination/history).
2. Listen to the fetal heart with stethoscope or Doppler ultrasound.
    - Should be 120–160 beats/min.
3. Identify the type of presentation by bedside ultrasound or by digital exam.
4. Perform a sterile digital examination to confirm the position of the baby and the stage of labor.
    - Membranes should not be artificially ruptured. The amniotic sac will help to dilate the cervix, lubricate the canal, and protect the umbilical cord from compression.
5. If footling presentation: await OB/general surgery for emergency C-section.
6. If frank/complete/incomplete: instruct the patient to push when the cervix is completely dilated.
7. When the breech has descended to the perineum, consider performing an episiotomy if more space is needed. Cleanse the perineum with antiseptic and sterile lubricant beforehand.
8. Allow the baby to extrude to the umbilicus with maternal efforts alone. Do not exert traction before this time.
9. If frank: deliver the posterior leg by gently guiding the sacrum anteriorly, grasping the thigh and flexing the leg at the knee.
10. Deliver the anterior leg in a similar manner while guiding the sacrum posteriorly.
11. If incomplete: deliver the extended leg as #8 or #9 as appropriate.
12. Delivery continues as it would for a complete presentation.
13. Wrap the legs/buttocks in a clean towel to decrease trauma (create grip).
14. Grasp the upper legs with the index fingers holding the anterior iliac crests. Place the thumbs on the sacrum (Fig. 117.2).
15. Apply gentle traction as the mother pushes until the scapulae and axillae are visible.
16. If there is difficulty delivering the shoulders, deliver the posterior shoulder by rotating the trunk 90° and applying gentle downward traction to rotate the shoulder anteriorly.
17. Rotate the baby 180° so as to deliver the anterior arm in a similar manner.
18. If the arms do not spontaneously deliver, a finger can be hooked over the shoulders to bring the arm down while rotating the trunk as above (Fig. 117.3).
19. Use the McRoberts position to increase the diameter of the pelvis (Fig. 117.4).

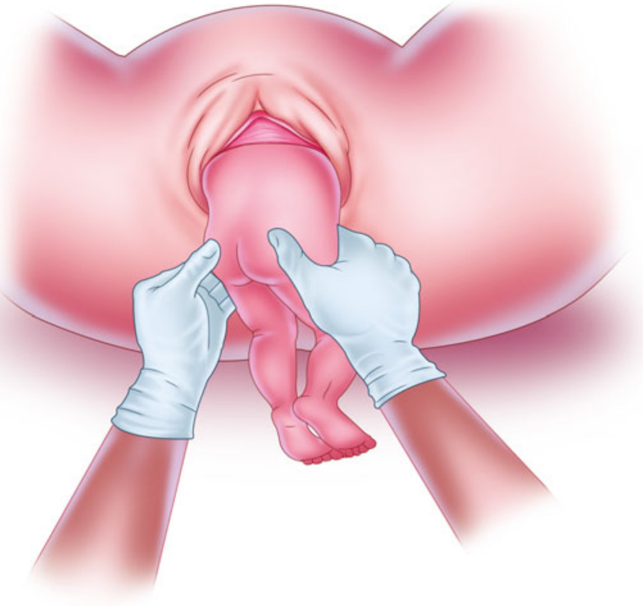

Fig. 117.2 Correct placement of hands on sacrum

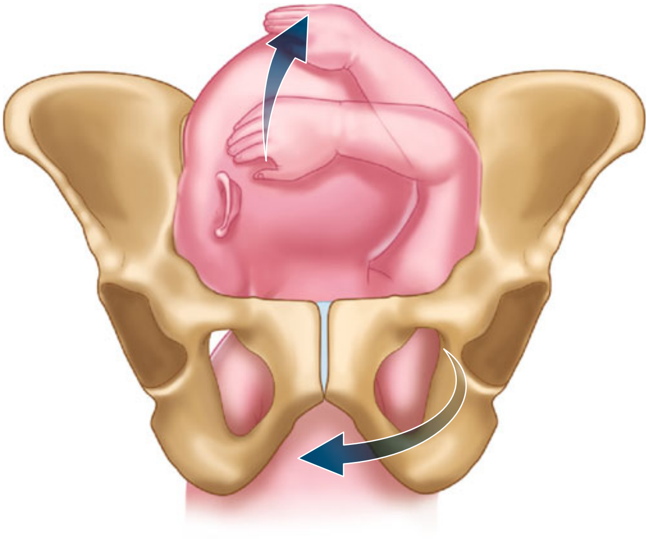

Fig. 117.3 Rotation toward maternal symphysis pubis to avoid nuchal arm

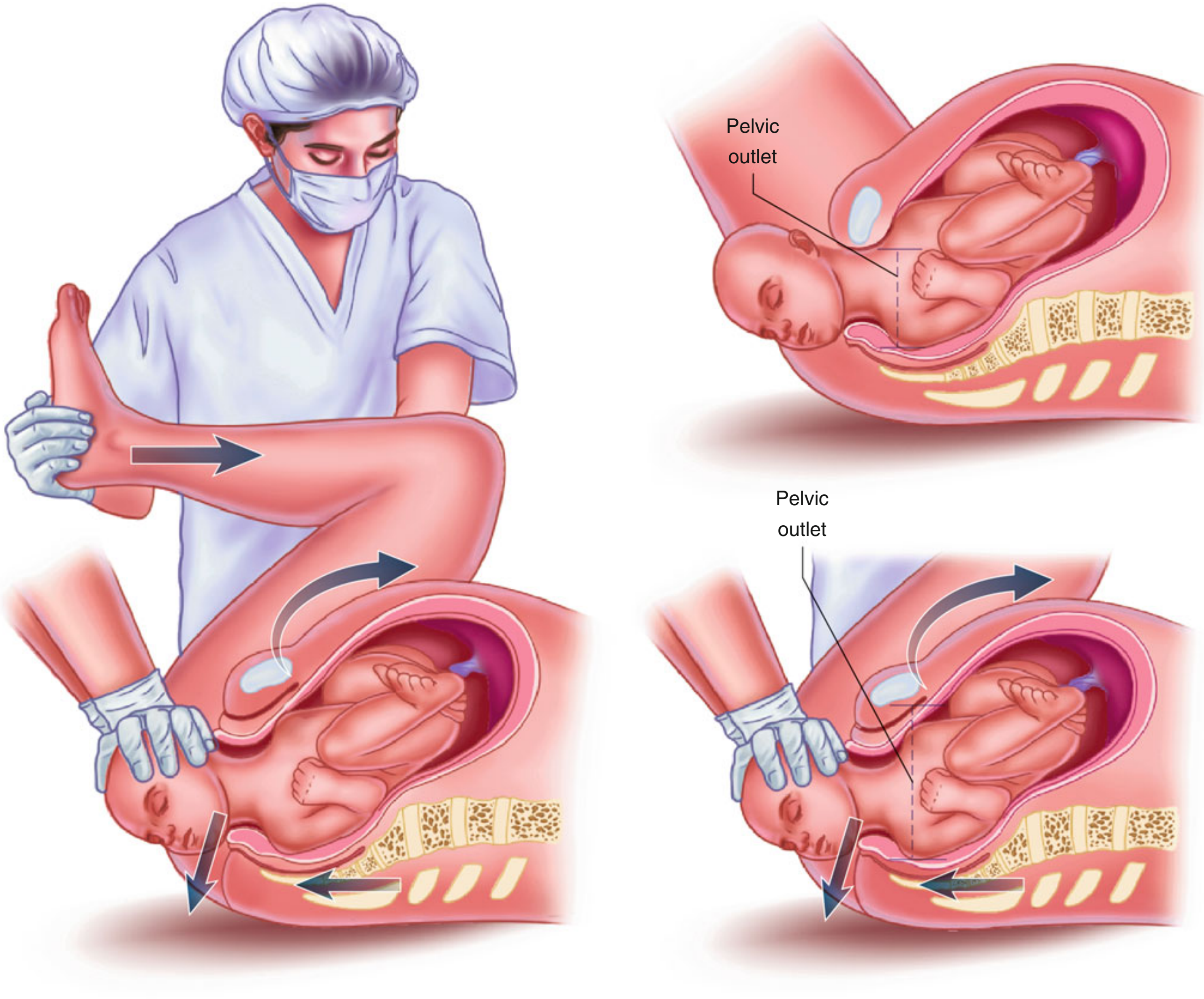
Pelvic
outlet
Pelvic
outlet

Fig. 117.4 The McRoberts maneuver in vertex (head-first) presentation

20. Maintain the baby in the same plane as the vagina (support the body with the forearm) and place the second and fourth fingers over the maxilla of the baby. Place the middle finger in the mouth or on the chin and the other hand on the upper back/occiput (Fig. 117.5).
    - Avoid extreme elevation of the fetus to prevent hyperextension and cervical spine injury. Piper forceps may also be used to promote flexion.
21. Deliver the head in the flexed position.
    - If the baby descends with the neck and abdomen facing anteriorly: Grasp the shoulders posteriorly with two fingers of one hand while the other hand flexes the abdomen and the baby's feet are brought upward (Fig. 117.6).
    - If the baby's neck remains extended: Leave the baby hanging (weight = traction). When the hairline appears under the symphysis, grab the baby by the feet and elevate upward (Fig. 117.7).
22. Clamp and cut the umbilical cord (collect arterial and venous samples for pH).
23. Suction the baby's mouth and nose, and resuscitate as indicated.
24. Deliver the maternal placenta.
25. Repair tears or episiotomy made during delivery.

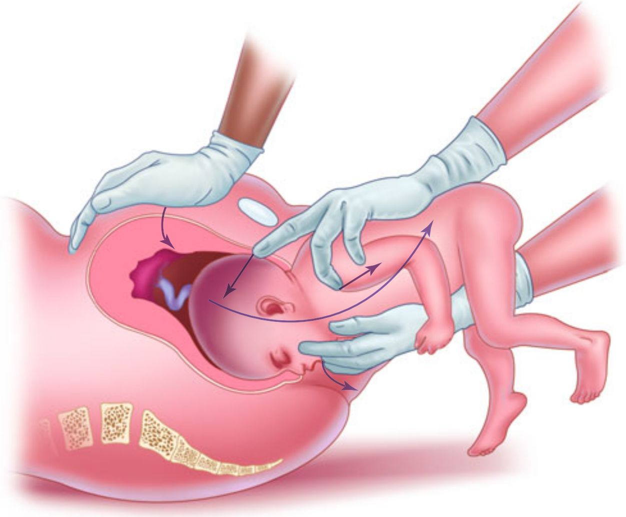

Fig. 117.5 Delivery of head in flexed position

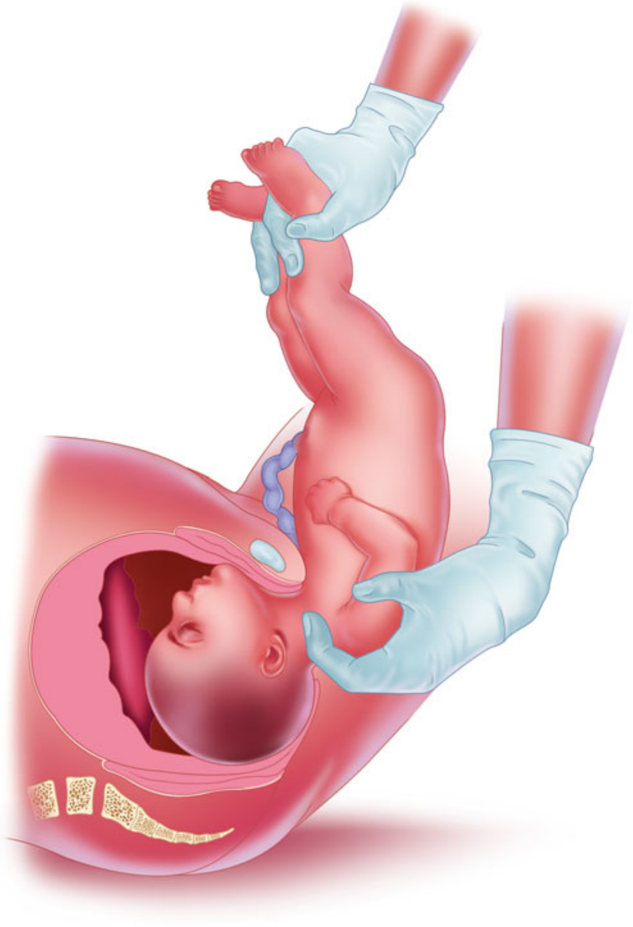

Fig. 117.6 Correct hand placement if abdomen facing anterior

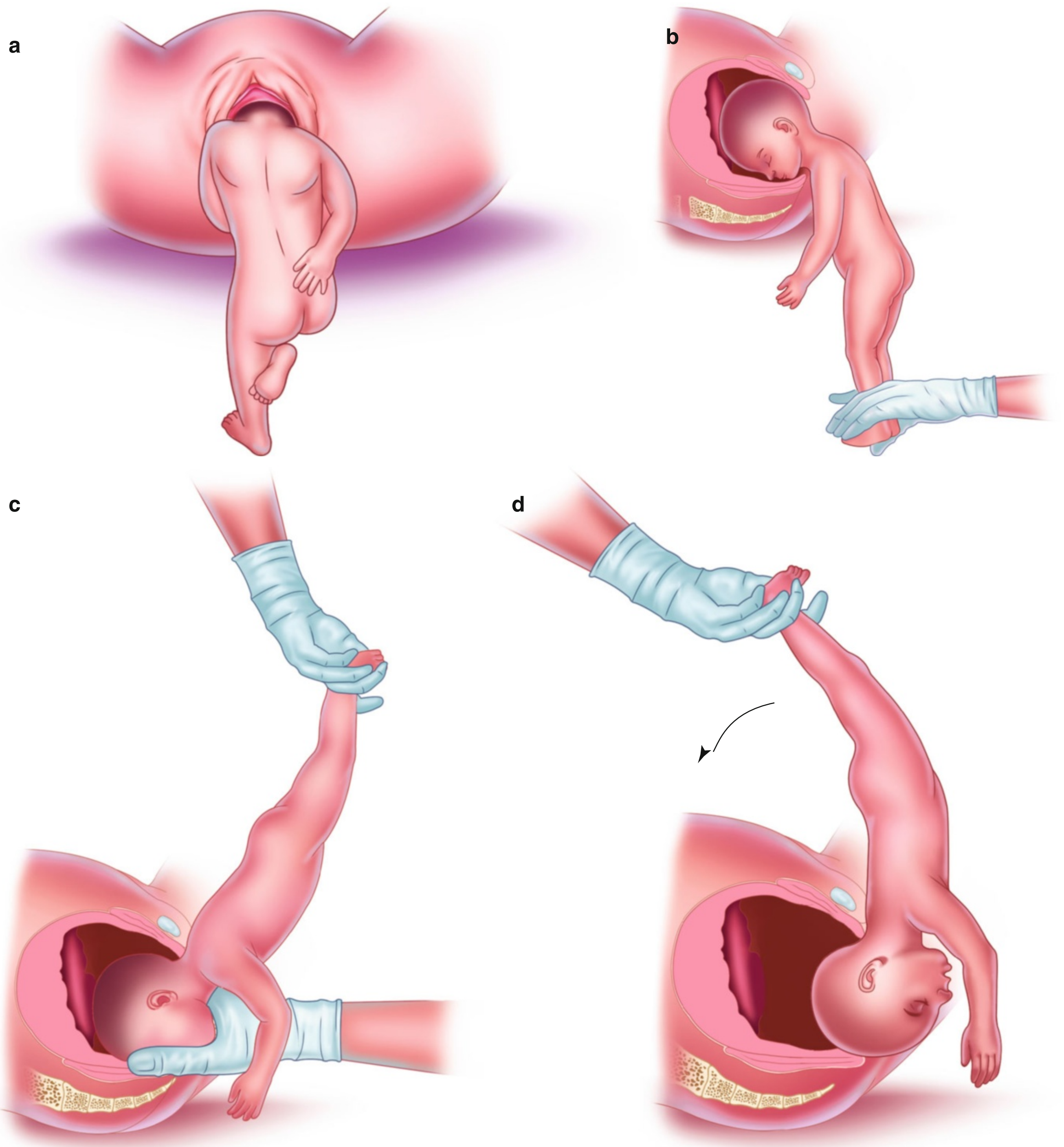

Fig. 117.7 The Burns-Marshall maneuver. (**a**) Allow baby to hang until you can see the hair at the nape of his neck; (**b**) swing the baby's head clear of the birth canal; (**c**) left hand guards and slips the perineum over fetal mouth; suction baby's air passage to clear mucus; (**d**) hold baby's feet

117.5 Complications

- Umbilical cord prolapse
- Brachial plexus injury (from the nuchal arm)
- Fetal head entrapment
- Cervical spine injury (from hyperextension of the neck)
- Birth asphyxia

117.6 Pearls and Pitfalls

- Pearls
  - Allow the uterine contractions to help deliver the baby.
- Pitfalls
  - Do not rush the delivery or use too much force. This can increase the risk of trauma to the baby and mother.
  - Beware of nuchal arm. To avoid brachial plexus injury: Rotate the face of baby toward the maternal symphysis pubis. This will reduce the tension keeping the arm around the back of the fetal head (Fig. 117.3).

Selected Reading

Auerbach PS. Gynecologic and obstetric emergencies. In: Wilderness medicine. 6th ed. Philadelphia: Elsevier; 2012.

Buckley RG, Knoop KJ. Gynecologic and obstetric conditions. In: Knoop KJ, Stack LB, Storrow AB, Thurman RJ, editors. The atlas of emergency medicine. 3rd ed. New York: McGraw Hill; 2010.

Cunningham FG, Leveno KJ, Bloom SL, Hauth JC, Rouse DJ, Spong CY. Breech presentation and delivery. In: Williams obstetrics. 23rd ed. New York: McGraw Hill; 2010.

Kish K, Collea JV. Chapter 21. Malpresentation & cord prolapse. In: DeCherney AH, Nathan L, editors. Current diagnosis & treatment: obstetrics & gynecology. 10th ed. New York: McGraw Hill; 2007.

Kotaska A, Menticoglou S, Gagnon R. Vaginal delivery of breech presentation. Int J Gynecol Obstet. 2009;107:169–76.

Probst BD. Emergency childbirth. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Elsevier; 2010.

Management of Primary Postpartum Hemorrhage

118

Megan Kwasniak, Anton A. Wray, and Joseph A. Tyndall

Postpartum hemorrhage (PPH) is defined as ≥500 ml blood loss within 24 h of vaginal delivery or 1000 ml loss within 24 h of cesarean section. It is the leading cause of maternal mortality worldwide.

118.1 Indications

- Excessive vaginal bleeding with or without pain and/or hemodynamic instability within 24 h of delivery

118.2 Contraindications

- There are no absolute contraindications to the management of PPH.

118.3 Materials and Medications

- Sterile technique
- Good lighting
- Sponge forceps
- Gauze
- Towels
- IV fluids
- Type and screen/crossmatch of blood
- Absorbable suture with curved needle
- Needle holder
- Tooth forceps

118.4 Procedure

- Standard resuscitation measures: Place IV, O_2, monitor.
- Assessment and treatment should occur simultaneously.
- All techniques should be performed under strict sterile conditions.

118.4.1 Uterine Exam

1. Assess by placing the hand on the uterine fundus and checking its size and firmness.
2. High, soft, or boggy uterus implicates retained placenta or uterine atony. Start external fundal massage.
3. Non-palpable uterus implicates uterine inversion.

118.4.2 Vaginal Exam

1. Keeping one hand on the abdomen, gently examine the vaginal canal using the other sterilely gloved hand.
2. Gently scoop out any clots/retained placenta that are easily removable.
3. Look for traumatic sources of bleeding from the perineum, vaginal walls, and cervical lacerations.
4. Gauze-wrapped ring forceps can be used to assist direct visualization and clearing of clots.

M. Kwasniak, MD
Emergency Department, The Brooklyn Hospital Center,
New York, NY, USA
e-mail: megan.kwasniak@gmail.com

A.A. Wray, MD (✉)
Department of Emergency Medicine, The Brooklyn Hospital
Center, New York, NY, USA
e-mail: anw9071@nyp.org

J.A. Tyndall, MD, MPH
Department of Emergency Medicine, University of Florida Health,
Gainesville, FL, USA
e-mail: tyndall@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_118

118.4.3 If Uterine Atony Suspected

1. Continue external fundal massage with the abdominal hand. Make a fist with the vaginal hand and start bimanual massage. Raise the uterus from the pelvis and pivot it anteriorly, compressing it between the external hand and the internal fist. This maneuver will result in expression of any clots present and decrease uterine bleeding via direct compression (Fig. 118.1).
2. Administer oxytocin (20–40 units in 1 L of normal saline or lactated Ringer's solution intravenously [IV]; alternatively give 10 units intramyometrially with a spinal needle).
3. Methylergonovine may also be used at this time if oxytocin fails to reduce uterine bleeding (100 or 125 mcg IV or intramyometrially; alternatively 200 or 250 mcg may be given intramuscularly).
4. Continue massage until bleeding slows and the uterus becomes more firm. This can take 15–30 mins.

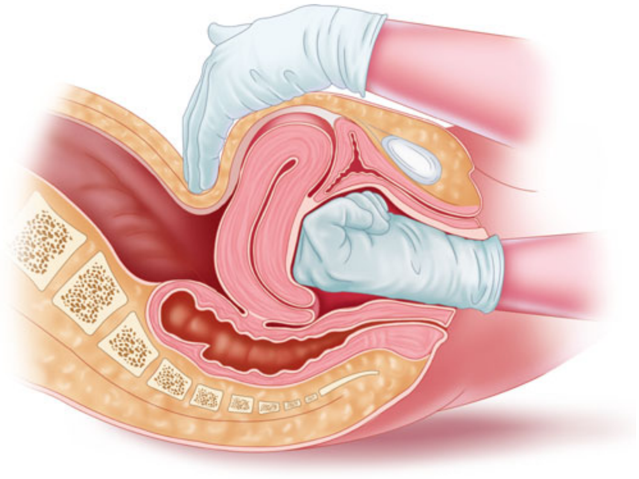

Fig. 118.1 Intrauterine massage for uterine atony

118.4.4 If Retained Placenta Suspected

1. Manual removal of the placenta or any of its retained tissue should be facilitated with sedation or additional analgesia.
2. Keeping the thumb and fingers together in a teardrop shape and using sterile technique as described previously, insert the hand through the vaginal canal and the cervix into the lower uterine segment.
3. Keep the other hand on the lower abdomen to continue gentle yet firm upward pressure and massage.
4. Find the placental edge within the uterus, grasp it gently, detach it from the uterine wall, and withdraw the hand from the patient.
5. Repeat the maneuver as necessary; remove any additional clots from the uterus and continue bimanual massage until it becomes firm and bleeding decreases.
6. If the entire placenta has been removed this way, an assistant should be available to inspect it for completeness and any torn vessels.
7. If the bleeding is particularly severe and placenta accreta is strongly suspected, pack the uterus with gauze and prepare the patient for urgent surgical intervention.

118.4.5 If Uterine Inversion Suspected

1. Using sterile technique, insert a fist through the vaginal canal and push the inverted fundus back through the cervical canal with pressure directed toward the umbilicus.
2. If the uterus has contracted, tocolytics such as IV magnesium sulfate or terbutaline can be used to aid myometrial relaxation. Alternatively, provided the patient is not overly hypotensive, nitroglycerin 50–100 mcg IV may be administered to relax the myometrium and facilitate return to normal uterine position.
3. If manual replacement is ineffective, hydrostatic reduction may be attempted. Warm fluids are run into the upper vagina under high pressure while occluding the introitus.

118.4.6 Trauma: Genital and Perineal Lacerations

1. For significant cervical lacerations: use absorbable sutures with a continuous interlocking stitch technique to close (Fig. 118.2).
2. For vaginal wound repair: place the initial and final stitch above the apices of the lacerations and grab a good amount of tissue with the needle. Small bites can lead to ongoing bleeding and hematoma formation.
3. Observe the repaired lacerations for any additional bleeding after the torn edges have been sutured.
    - Apply additional pressure to any site that continues to ooze blood; gauze-wrapped ring forceps may be used for this purpose if necessary.

If none of the above causes are apparent, consider underlying coagulopathies and treat appropriately. This may require administration of fresh-frozen plasma, platelets, or clotting factors as indicated.

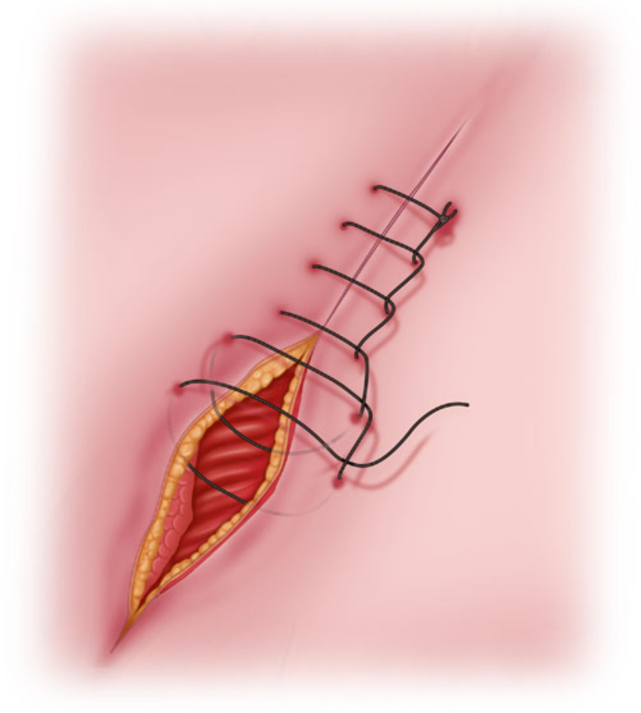

Fig. 118.2 Continuous interlocking stitch: perineal lacerations repair

118.5 Complications

- Uterine perforation and scarring
- Urinary and genital tract trauma and injury
- Genitourinary and genitointestinal fistula
- Pelvic hematoma
- Genital vascular injury
- Infection and sepsis
- Disseminated intravascular coagulation (DIC)
- Maternal death

118.6 Pearls and Pitfalls

- Pearls
    - Causes of PPH can be divided into the “5 Ts”:
        - Tone: uterine atony, occurring within the first 4 h after delivery.
        - *T*issue: retention of the placenta, especially placenta accreta and its fragments, more common at extreme preterm deliveries.
        - *T*rauma: injury to the uterus, cervix, and perineal structures after delivery of a large fetus, use of forceps and/or vacuum, frequent vaginal manipulation during delivery, and episiotomy procedures.
        - *T*hrombosis: intrinsic or acquired coagulation disorders, including idiopathic thrombocytic purpura (ITP); hemolysis, elevated liver enzymes, and low platelets (HELLP) syndrome; and disseminated intravascular coagulation (DIC), as well as preexisting conditions such as von Willebrand disease.
        - *T*raction: inversion of the uterus during placental delivery secondary to excessive traction on the umbilical cord. The uterine fundus can be within the endometrial cavity, in the cervical canal, or outside the external os and within the vaginal canal.
    - The administration of broad-spectrum antibiotics should be strongly considered following any manual removal, exploration, or instrumentation of the uterus and the genital tract.
    - Bedside ultrasonography can be very helpful for identifying uterine abnormalities, retained placental tissue, free fluid in the pelvis, and/or intrauterine hematoma.
    - Risk factors for PPH include prolonged active phase of labor, previous PPH, multiple pregnancy, and history of a bleeding disorder.
- Pitfalls
    - Failure to recognize and treat PPH early increases morbidity and mortality.
    - Underestimating the potential blood loss of PPH. The gravid uterus at term has a blood flow of 600 ml/min (non-gravid: 60 ml/h).

Selected Reading

Anderson J, Duncan E. Prevention and management of postpartum hemorrhage. Am Fam Physician. 2007;75:875–82.

Druelinger L. Postpartum emergencies. Emerg Med Clin North Am. 1994;12:219–37.

Leduc D, Senikas V, Lalonde AB, et al. Active management of the third stage of labour: prevention and treatment of postpartum hemorrhage. J Obstet Gynaecol Can. 2009;31:980–93.

Miller S, Lester F, Hensleigh P. Prevention and treatment of postpartum hemorrhage: new advances for low-resource settings. J Midwifery Womens Health. 2004;49:283–92.

Sheiner E, Sarid L, Levy A, Seidman DS, Hallak M. Obstetric risk factors and outcome of pregnancies complicated with early postpartum hemorrhage: a population-based study. J Matern Fetal Neonatal Med. 2005;18:149–54.

Tessier V, Pierre F. Risk factors of postpartum hemorrhage during labor and clinical and pharmacological prevention. J Gynecol Obstet Biol Reprod (Paris). 2004;33(8 suppl):4S29–56.

World Health Organization. World health report 2005: make every mother and child count. Available at: http://www.who.int/whr/2005/whr2005_en.pdf.

119

Perimortem Cesarean Section

Jordana J. Haber, Elaine B. Josephson,
and Muhammad Waseem

119.1 Indications

- Maternal arrest with a viable fetus (gestation >24 weeks)

119.2 Contraindications

- Stable mother
- Fetus less than 24 weeks' gestation
- Extreme fetal prematurity
- Maternal hypoxia longer than 15 min

119.3 Materials and Medications

- Cesarean section instrument tray if available
  - #10 or #11 scalpel blade, scissors, bladder retractor, 2 large retractors, gauze sponges, hemostats, suction, forceps, and straight and curved clamps
- Skin antiseptic preparing solution, such as Betadine (povidone-iodine)
- Silk suture with needle driver or skin stapler
- Sterile drapes
- Sterile gloves
- Obstetrical pack (See Chap. 115)
  - Bulb syringe and umbilical cord clamp
- Clean blanket or towels for delivery
- Neonatal resuscitation equipment

Owing to the rarity of this procedure in the emergency department, it is unlikely to have a prepared cesarean section tray available. In this case, a thorocotomy or thorocostomy tray combined with an obstetrical pack would contain all the supplies needed. At a minimum, a scalpel and an obstetrical pack are necessary.

J.J. Haber, MD
Department of Emergency Medicine, Maimonides Medical Center,
New York, NY, USA
e-mail: Jordana.haber@gmail.com

E.B. Josephson, MD
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, Weill Cornell Medical College of Cornell
University, Bronx, New York, NY, USA
e-mail: elaine.Josephson@nychhc.org

M. Waseem, MD (✉)
Department of Emergency Medicine, Lincoln Medical and Mental
Health Center, New York, NY, USA
e-mail: waseemm2001@hotmail.com

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_119

119.4 Procedure

1. Prepare skin with antiseptic solution and a sterile drape.
2. Insert a Foley catheter to empty the bladder.
3. Continue cardiopulmonary resuscitation until delivery.
4. Obtain emergent obstetrician and neonatologist consult if available, but do not delay procedure.
5. Using a #10 or #11 blade, make a vertical midline incision beginning 4–5 cm below the xiphoid process and extend the incision to the pubic symphysis (Fig. 119.1).
6. Incise through the subcutaneous fat no further than the rectus sheath.
7. Lift the rectus sheath with a toothed forceps and make an incision with scissors to expose the uterus (Fig. 119.2).
8. With forceps and scissors, lift and incise the peritoneal membrane in the midline.
9. Identify and lift the bowel and cover it with saline-soaked gauze.
10. Retract the rectus sheath and bladder with a bladder retractor or, if not available, use saline-soaked gauze or a towel.
11. Create a 2- to 4-cm midline vertical opening in the uterus.
12. Place a finger in the opening directed caudally to protect the fetus while making a superior incision through the uterine wall. Once complete, repeat this step in the inferior direction.
13. Use a clamp to rupture the amniotic membranes. Immediately deliver the fetus and clamp the umbilical cord.
14. Expulse the head by placing a hand between the pubic symphysis landmark and the fetal occiput. Then, gently flex the fetus while simultaneously moving the head superiorly and anteriorly until delivery (Fig. 119.3).
15. Suction the mouth and nose with a bulb syringe immediately.
16. Deliver the shoulders, followed by the torso and extremities. Secure the umbilical cord with a hemostat or umbilical cord clamp 10 cm distal to the fetus and a second clamp 2 cm distal to this clamp. With scissors, incise the umbilical cord between the two clamps.
17. Immediately begin resuscitation of the infant (Fig. 119.4).
18. If the patient is still alive or regains vital signs, prepare to deliver the placenta. Begin with an oxytocin infusion at 20 U in 1 L at 10 mL/h. Apply cautious traction to the umbilical cord until the placenta separates from the uterus (Fig. 119.5).
19. Following delivery, the uterus should be closed using two layers with either 2-0 or 1-0 suture. In the case of maternal death, skin staples or a running stitch is an acceptable method of skin closure.

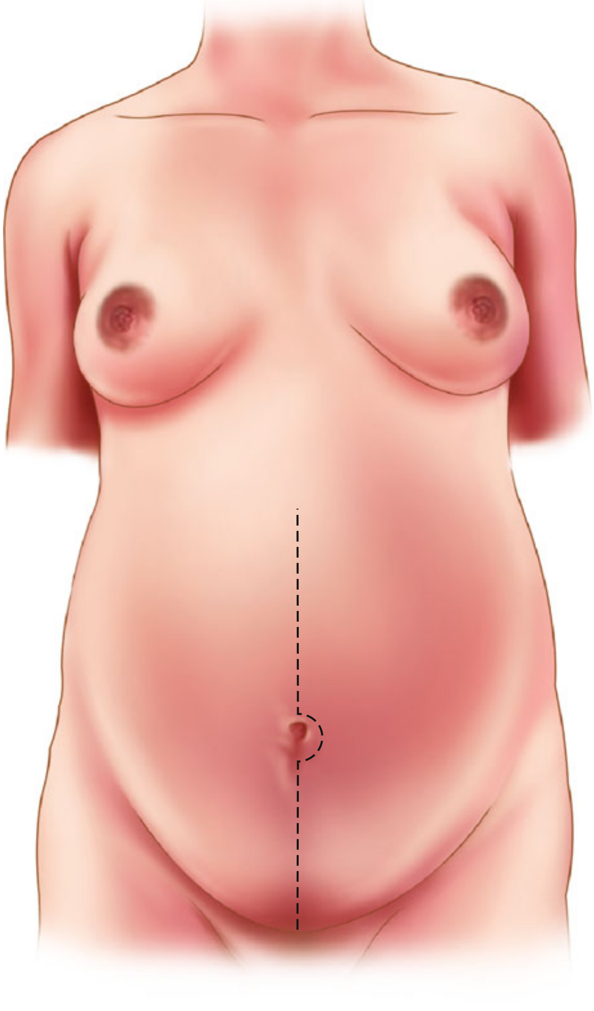

Fig. 119.1 Vertical incision

Fig. 119.2 Exposing rectus sheath

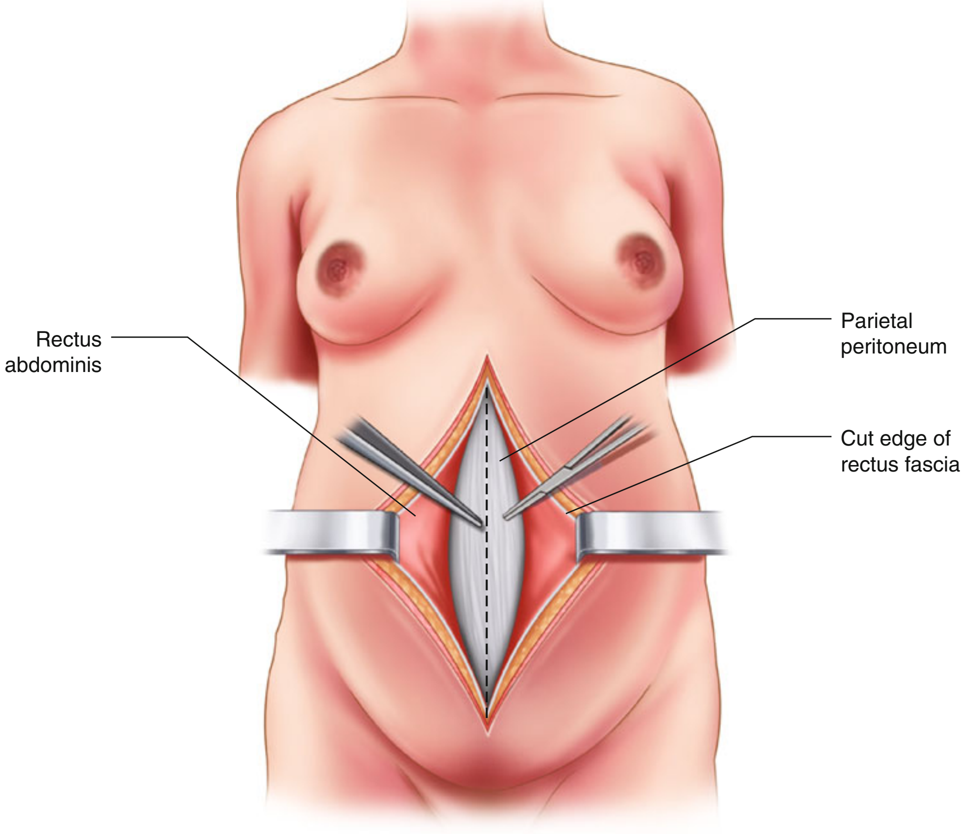
Rectus
abdominis
Parietal
peritoneum
Cut edge of
rectus fascia

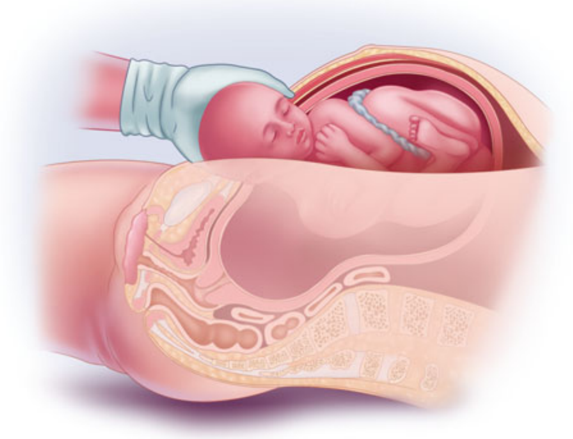

Fig. 119.3 Delivery of the fetus

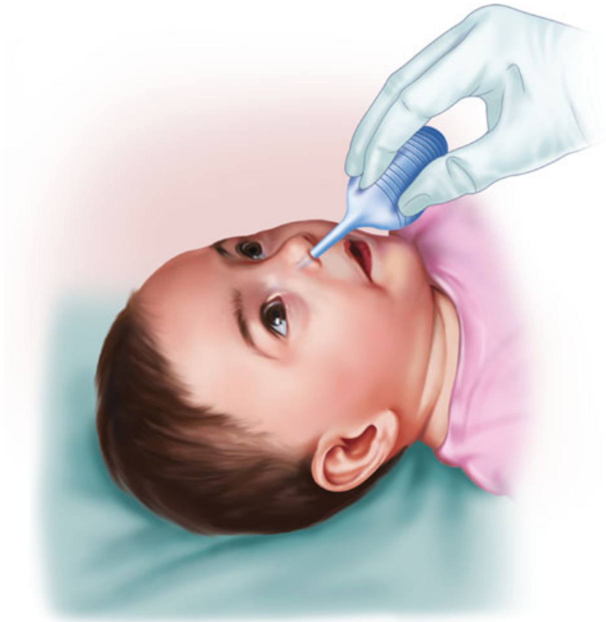

Fig. 119.4 Suctioning newborn as part of resuscitation

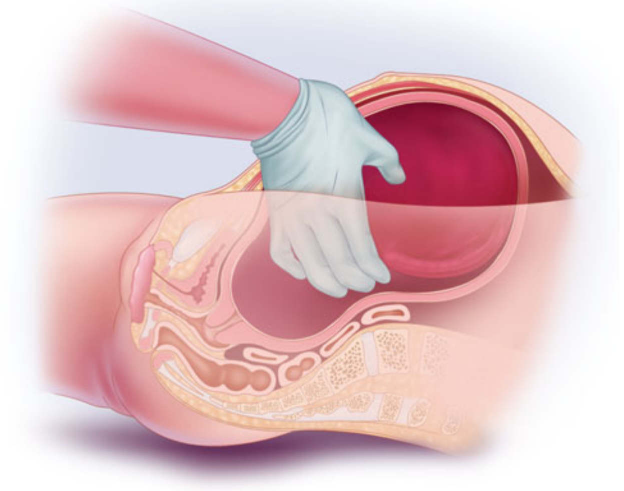

Fig. 119.5 Delivery of the placenta

119.5 Complications

- Maternal sepsis
- Maternal visceral injury
- Maternal hemorrhage
- Maternal death
- Fetal injuries and laceration
- Fetal sepsis

119.6 Pearls and Pitfalls

- Pearls
  - Perimortem cesarean section, although rarely performed, should be considered in any maternal arrest when the fetus is greater than 24 weeks' gestation.
  - In addition to saving the life of the fetus, this procedure may aid in resuscitation of the mother. Emptying of the uterus may improve thoracic compliance and, therefore, improve maternal ventilation.
- Pitfalls
  - The decision to perform an emergency cesarean section must be made early. There is a higher chance of survival if performed no more than 5 min after the onset of maternal cardiac arrest.

Selected Reading

Doan-Wiggins L. Emergency childbirth. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 4th ed. Philadelphia: Saunders; 2004. p. 1117–43.

Flippin A, Hendricks S. Perimortem cesarean section. In: Reichman EF, Simon RS, editors. Emergency medicine procedures. New York: McGraw Hill Medical; 2004. p. 1070–8.

Gianopoulos JG. Emergency complications of labor and delivery. Emerg Med Clin North Am. 1994;12:201–17.

Jeejeebhoy FM, Zelop CM, Windrim R, Carvalho JC, Dorian P, Morrison LJ. Management of cardiac arrest in pregnancy: a systematic review. Resuscitation. 2011;82:801–19.

Whitten M, Irvine LM. Postmortem and perimortem caesarean section: what are the indications? J R Soc Med. 2000;93:6–9.

Part XV

Common Pediatric Procedures

Peripheral Venous Catheterization

120

David N. Smith and Judith K. Lucas

120.1 Indications (See Also Chap. 2)

- Fluid resuscitation
- Medication administration
- Blood draws

120.2 Contraindications

- Relative
  - Avoid catheterizing areas of trauma in which extravasation of fluid is possible (e.g., burns, open wounds, or severe edema in tissue).
  - Avoid catheterizing in an area of local infection for risk of inoculating the circulation with bacteria (e.g., cellulitis).
- Absolute
  - None

120.3 Materials and Medications

- Gloves
- Skin disinfectant (isopropyl alcohol, chlorhexidine, or Betadine [povidone-iodine])
- Appropriate-sized catheter (18- to 24-gauge [IV]) (Fig. 120.1)
  - Large child: 18 to 20 gauge
  - Infant or small child: 22 to 24 gauge
- Tourniquet
- Sterile 2 × 2 gauze
- Appropriate-sized Tegaderm transparent dressing
- Adhesive tape
- IV bag with solution set (tubing flushed and ready) or saline lock
- Sharps container

D.N. Smith, MD
Department of Pediatrics, University of Alabama at Birmingham,
Children's of Alabama, Birmingham, AL, USA
e-mail: dsmith@peds.uab.edu

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_120

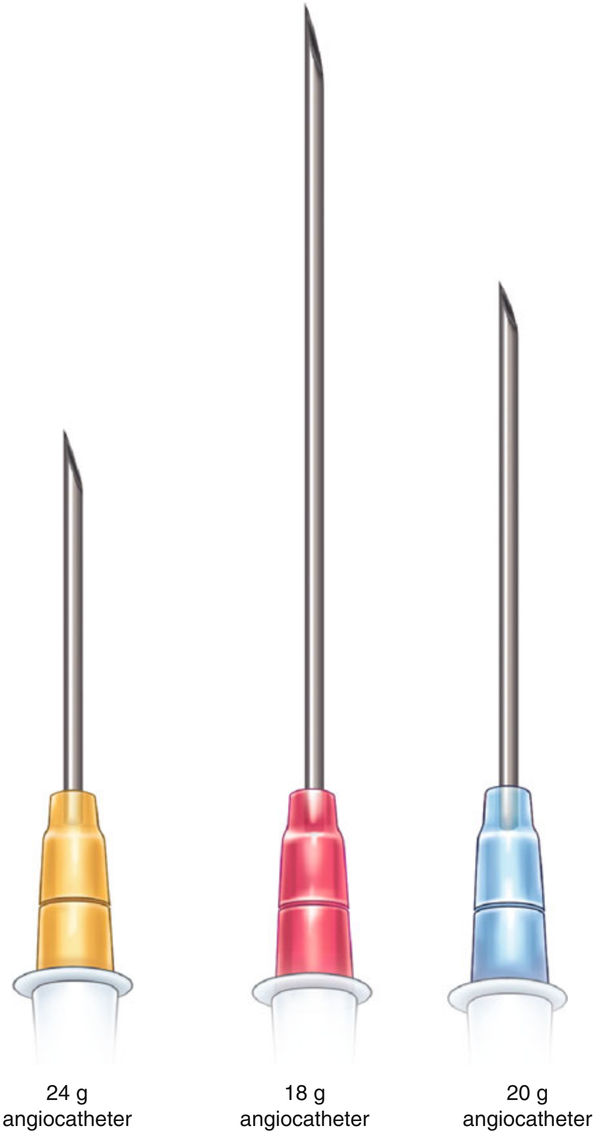

Fig. 120.1 24-Gauge angiocatheter (*yellow*), 18-gauge angiocatheter (*red*), 20-gauge angiocatheter (*blue*)

120.4 Procedure

1. Comfortably position the patient with the site exposed.
2. Assemble the equipment and don a pair of (nonlatex) examination gloves.
3. Apply the tourniquet to the extremity above the site to be catheterized (Fig. 120.2).
4. Visualize and palpate the vein.
5. Cleanse the site with a disinfectant swab using an expanding circular motion.
6. Prepare and inspect the catheter and flush the tubing; be certain that the stylet and catheter separate easily, then fit again into the notch, aligning the bevel with the hub.
7. Stabilize the vein and apply countertension to the skin, being careful not to touch the cleansed area.
8. Insert the stylet through the skin and then reduce the angle while advancing through the vein (Fig. 120.3).
9. Observe for “flashback” as blood slowly fills the flashback chamber.
10. Advance the needle approximately 1–2 mm further into the vein, depending on the gauge and age of the patient, to ensure that the catheter is within the vein.
11. Slowly advance the catheter into the vein while keeping tension on the vein and skin (Fig. 120.4).
12. While advancing the catheter, be certain to hold the stylet portion with the thumb and forefinger, so as to avoid advancing the needle portion into and through the opposite side of the vessel, thus “blowing” the vein.
13. When the catheter is advanced about halfway, slowly withdraw and remove the stylet while simultaneously continuing to advance the catheter to its hub.
14. Attach a 3-mL non-Luer-Lok syringe to the hub.
15. Remove the tourniquet.
16. Gently attempt to aspirate blood. The blood should be free flowing.
17. Secure the catheter by either placing a transparent occlusive dressing (e.g., Tegaderm) over the lower half of the catheter hub or taping over the catheter hub in a cruciate

fashion, taking care not to cover the IV tubing connection (Fig. 120.5).

18. Remove the cover from the end of the IV tubing and insert the IV tubing into the hub of the catheter (the tubing must have been flushed with IV solution before connecting with the catheter hub: the unit from the solution bag/bottle through the catheter must be air free).
19. Open up the IV roller clamp and observe for drips forming in the drip chamber.
20. Place a piece of tape over the catheter hub then make a small (kink-free) loop in the IV tubing and place a second piece of tape over the first piece to secure the loop (Fig. 120.6).
21. Place a third piece of tape over the IV tubing above the site.
22. Ensure that the IV is properly secured and infusing properly.
23. Ensure that all "sharps" are placed in the sharps container.

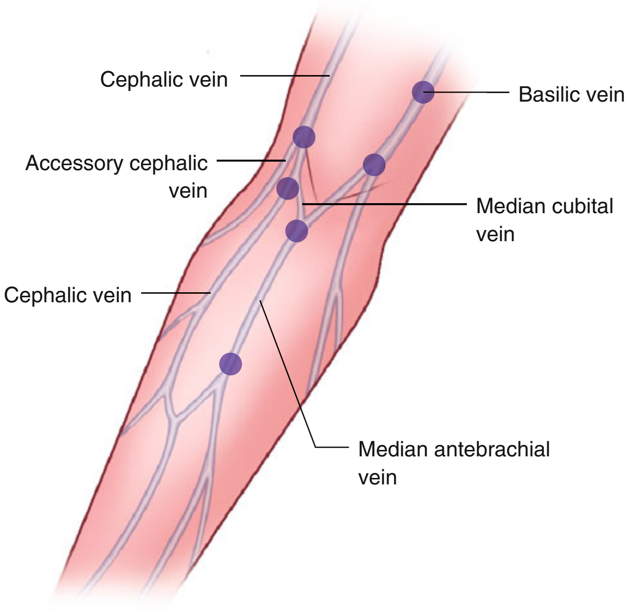
Cephalic vein
Basilic vein
Accessory cephalic vein
Median cubital vein
Cephalic vein
Median antebrachial vein

Fig. 120.2 Anatomy of the volar surface antecubital fossa and forearm

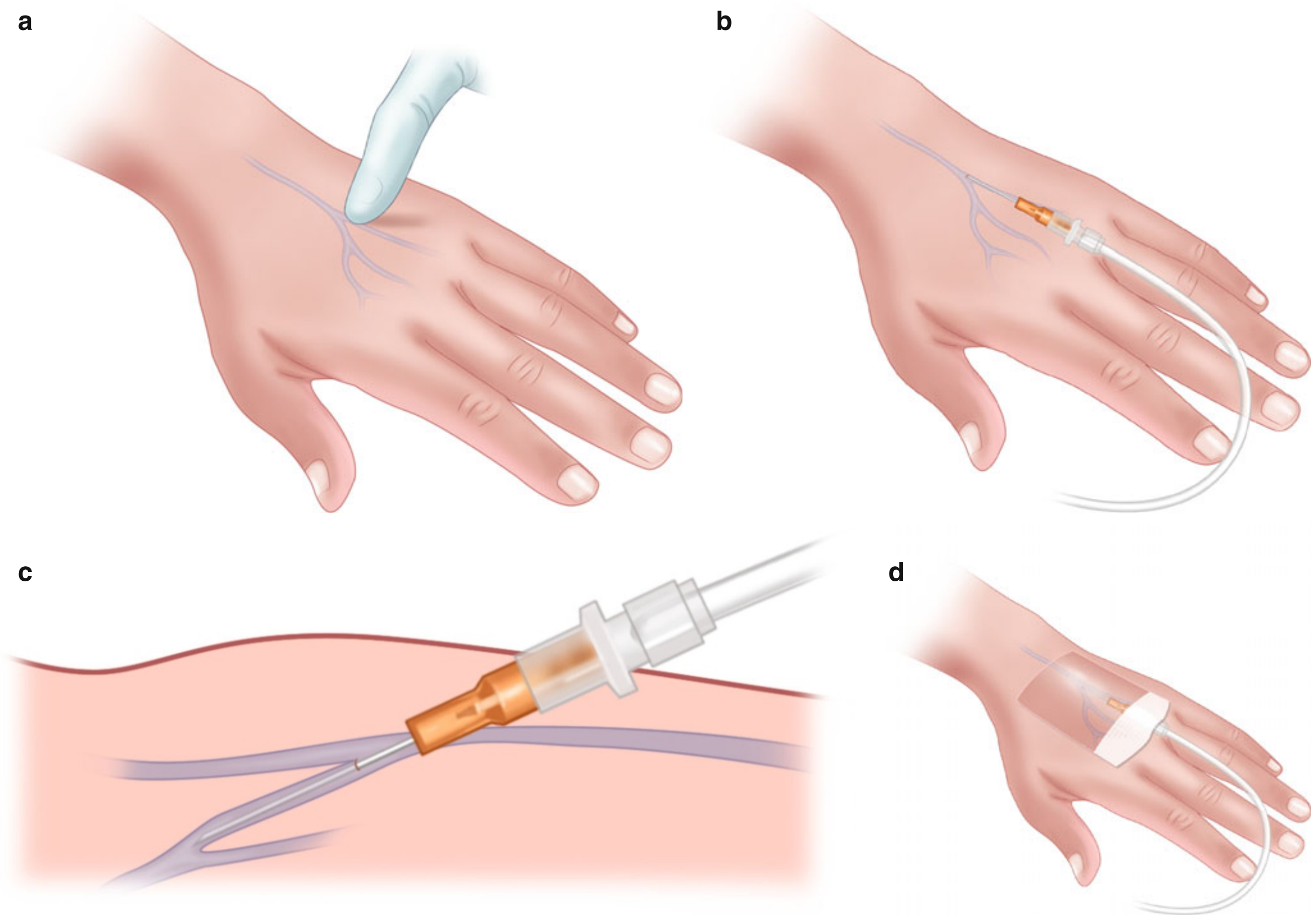

Fig. 120.3 (**a**–**d**) After applying the tourniquet, palpate a vein, as straight as possible, and ideally without many "knots" (i.e., valves)

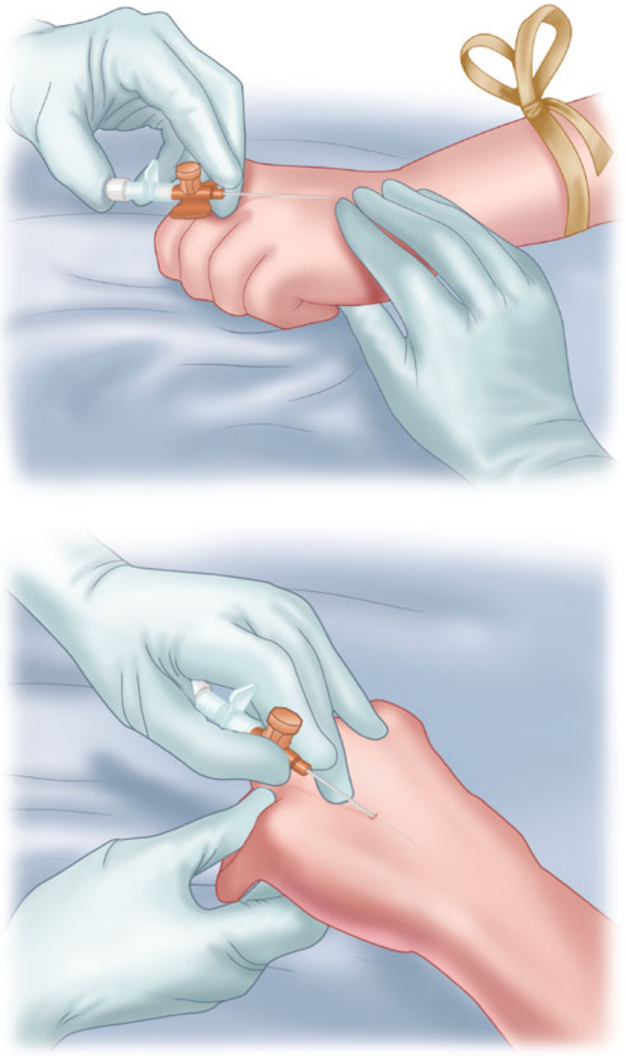

Fig. 120.4 With the tourniquet “up,” and the vein distended, apply traction to the skin, pierce the skin, and pass the catheter tip (into the vein) until blood return is noted in the catheter hub

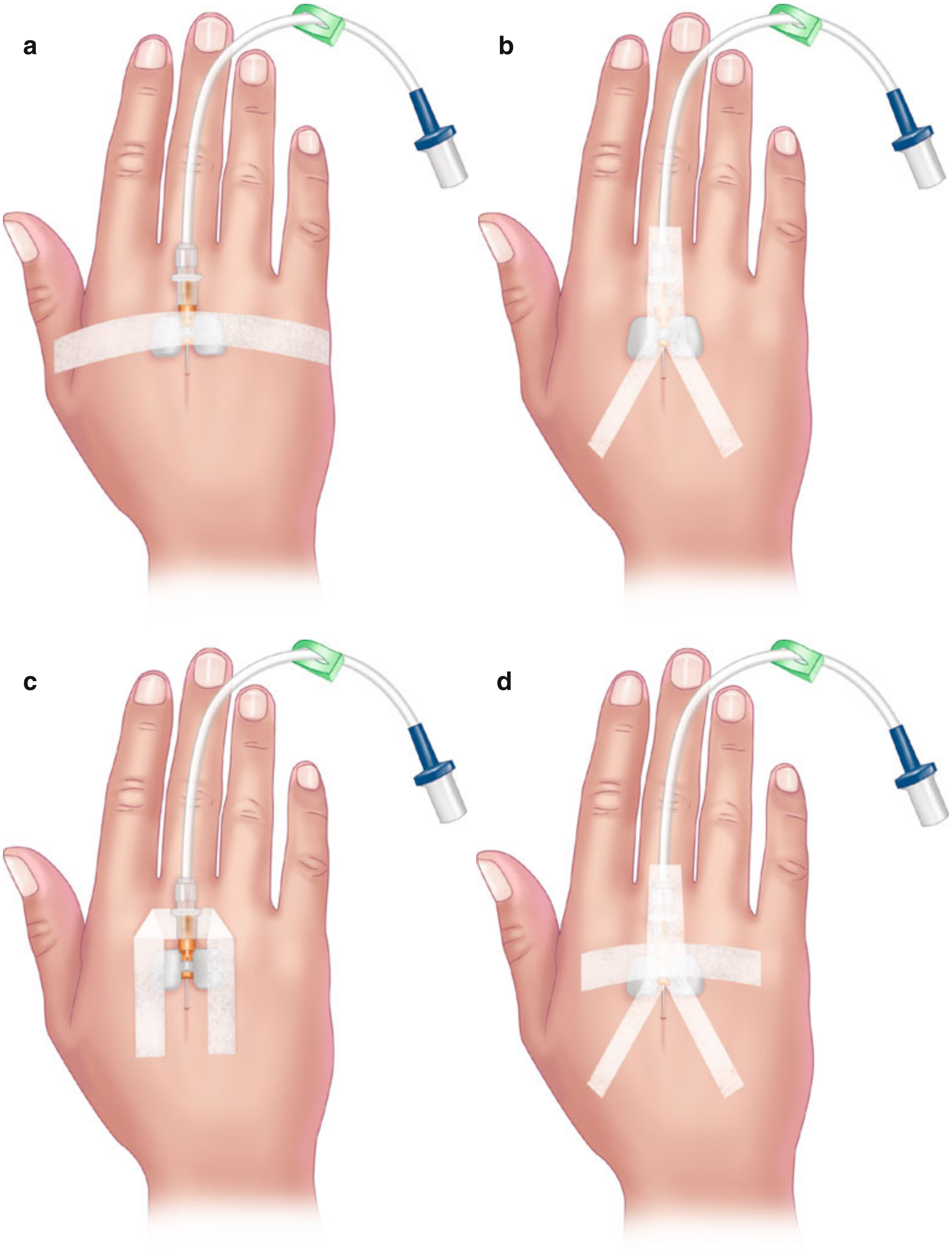

Fig. 120.5 (**a**–**d**) Securing the intravenous line utilizing the cruciate taping style

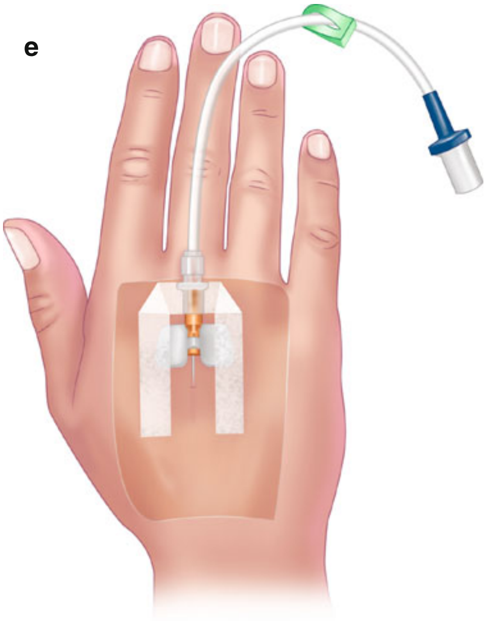

Fig. 120.5 (continued)

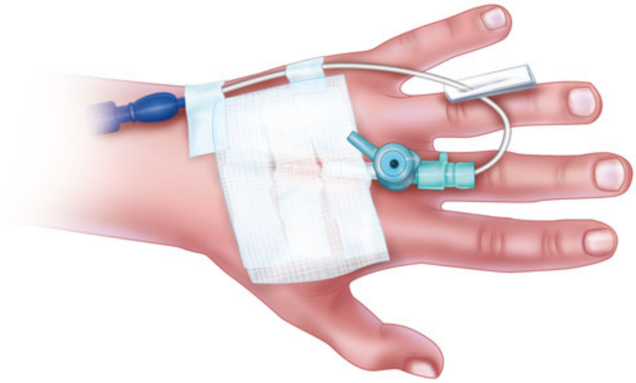

Fig. 120.6 Commercially available hub stabilizer, minimizing the need for excessive tape

120.5 Pearls and Pitfalls

- Pearls
  - Start catheter attempts distal in the extremities and move proximally with each subsequent attempt.
  - The use of ultrasound or a light source in infants can aid in location of the vessel and placement of the line.
  - In an emergent situation, in which fluids or medications are needed quickly, intraosseous access (see Chap. 122) can be obtained if venous catheterization fails.
- Pitfalls
  - The use of lidocaine subcutaneously can improve patient comfort, but it does disrupt anatomical landmarks.

Selected Reading

Bailey P. Vascular (venous) access for pediatric resuscitation and other pediatric emergencies. UpToDate. http://www.uptodate.com/contents/vascular-venous-access-for-pediatric-resuscitation-and-other-pediatric-emergencies. Accessed 23 June 2014.

Department of Emergency Medicine, University of Ottawa. Peripheral intravenous access. 2003. http://www.med.uottawa.ca/procedures/iv/. Accessed 23 June 2014.

Kost S. Ultrasound-guided vascular (venous) access. UpToDate. http://www.uptodate.com/contents/principles-of-ultrasound-guided-venous-access. Accessed 23 June 2014.

Nursing Resource Administration. Medical procedure: insertion of peripheral IV line. http://nursing-resource.com/iv-insertion/. Accessed 23 June 2014.

Umbilical Venous Catheters (Insertion and Removal)

121

Judith K. Lucas

121.1 Indications

- Temporary vascular access for infants up to roughly 10 days of life (between 7 and 14 days) with shock or cardiopulmonary failure
- Emergency vascular access in this age group, when peripheral intravenous (IV) access cannot be rapidly obtained
- Preferred vascular access in infants less than 1000 g

121.2 Contraindications

- Omphalitis
- Omphalocele
- Necrotizing enterocolitis
- Peritonitis

121.3 Materials and Medications

- Anesthetic: not necessary; procedure is painless
- Soft ties to restrain infant’s extremities
- Sterile gloves/gowns
- Antiseptic solution
- Sterile towels/drapes
- 3.5- (infants <1500 g) or 5-French (>1500 g) umbilical venous catheter
- 5-French feeding tube
- Three-way stopcock
- 10-mL syringe with heparinized saline flush (1 U/mL)
- Umbilical tape or 3-0 silk on a cutting needle
- Non-toothed forceps
- Small hemostats (2)
- #11 scalpel and blade
- Scissors
- Graph depicting length of catheter insertion, if placing umbilical venous catheter (UVC) above the diaphragm in either very small infants or infants for whom measurement of central venous pressure (CVP) is indicated (Fig. 121.1)

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_121

Fig. 121.1 (**a**) Instrumentation suggested for umbilical venous catheter placement. (**b**) Umbilical venous catheters (in an emergency, 5 F feeding tube is an acceptable alternative)

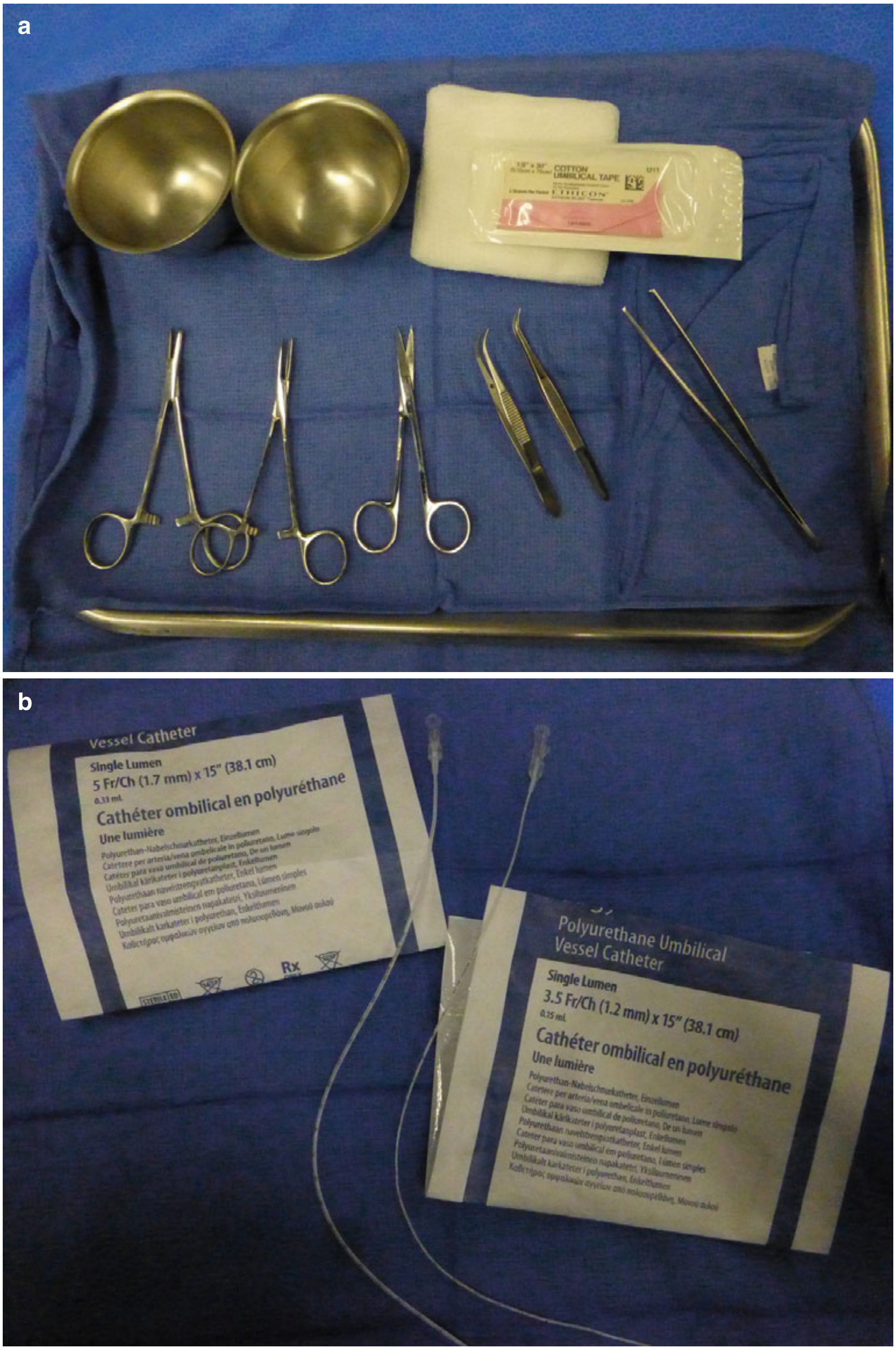
a
COTTON UMBILICAL TAPE
ETHICON
b
Vessel Catheter
Single Lumen
5 Fr/Ch (1.7 mm) x 15" (38.1 cm)
Cathéter ombilical en polyuréthane
Une lumière
Rx
Polyurethane Umbilical Vessel Catheter
Single Lumen
3.5 Fr/Ch (1.2 mm) x 15" (38.1 cm)
Cathéter ombilical en polyuréthane
Une lumière

121.4 Procedure (Insertion)

1. Place the infant under a radiant warmer.
2. Using soft ties, restrain the infant's extremities.
3. Scrub the umbilicus and surrounding abdomen with antiseptic solution.
4. Drape the umbilicus and area in sterile manner (leave the infant's head exposed).
5. Tie a loose loop with the umbilical tape around the base of the cord OR run the 3-0 silk through the skin *of the cord* in a purse-string fashion.
   - This will be used later to anchor the line after placement and to provide hemostasis should the line accidentally be pulled out and bleeding ensue.
6. Using the scalpel blade, cut the umbilical cord horizontally approximately 2 cm above the junction between the cord and the skin.
7. Identify the umbilical vessels
   - The vein is thinner walled, larger in diameter, and somewhat floppy appearing, relative to the umbilical arteries, and typically located at the 12 o'clock position.
   - The arteries are smaller, thick walled, and paired (a single umbilical artery often signifies the presence of a congenital malformation/syndrome) and located at the 4 and 8 o'clock positions (Fig. 121.2).
8. Place the stopcock on the receiving end of the umbilical catheter or the 5-French feeding tube and flush with heparinized saline solution, and then close the stopcock.
   - It is imperative that there is no air in the catheter.
9. Introduce a closed smooth-surfaced forceps into the lumen of the umbilical vein and allow the forceps to separate, allowing the vein to dilate (Fig. 121.3).
10. Insert the catheter (or feeding tube) into the lumen and gently advance, directing the catheter toward the right shoulder (Fig. 121.4).
11. Advance the catheter only until good blood flow is noted and then another 1 to 2 cm (this should be a total of only 4–5 cm in a term infant).
    - *Do not force the advancement.*
    - At this level the tip of the catheter should still be inferior to the liver (Fig. 121.5).
12. Tighten the umbilical tie or the purse-string suture.
13. Secure the catheter with a tape bridge (Figs. 121.6, 121.7, and 121.8).
14. Although if placed only up to point of blood return, the catheter tip should be below the liver; it is best to get an abdominal x-ray to confirm; many solutions are caustic to the liver and can result in complications.

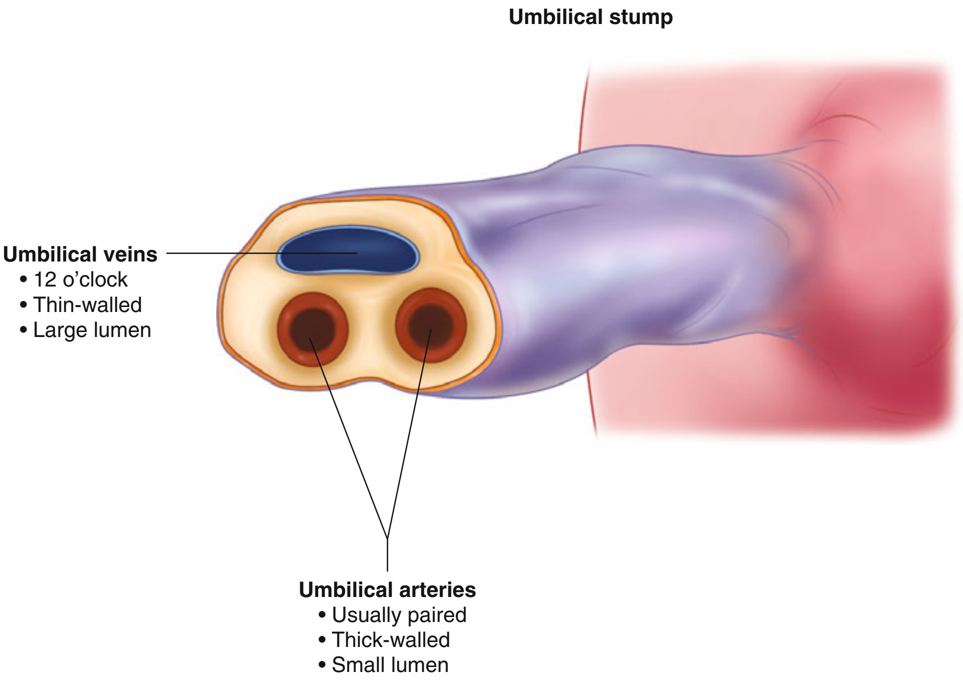

Fig. 121.2 Anatomy of the umbilical cord when cut transversely approximately 2 cm from abdominal wall

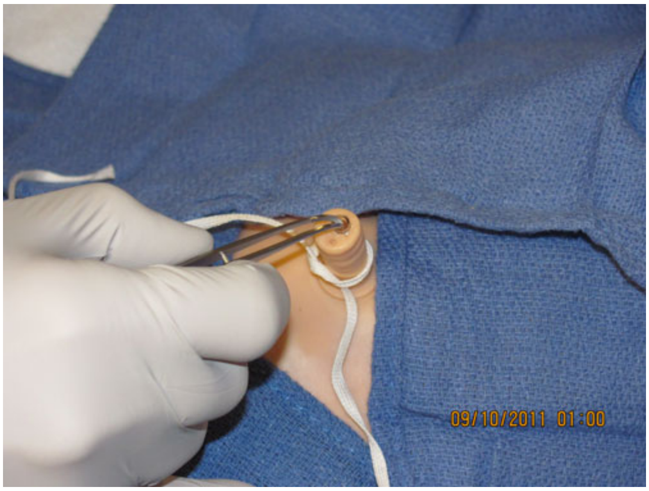

Fig. 121.3 Sterilely draped umbilicus, with umbilical tape loosely tied at base of umbilical cord. This tie will be cinched and secured once the catheter is placed, but meanwhile, can assist with homeostasis should the catheter be inadvertently dislodged during placement. Gently dilate the umbilical vein with a smooth toothed forceps. Insert the distal most couple of mm of a closed forceps into umbilical vein and relax, so the forceps tips smoothly separate

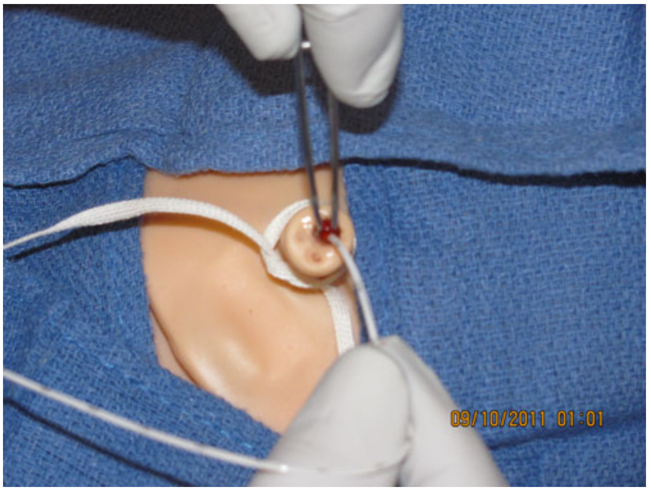

Fig. 121.4 Gently pass the umbilical venous catheter (or 5 F feeding tube) until there is easy blood return

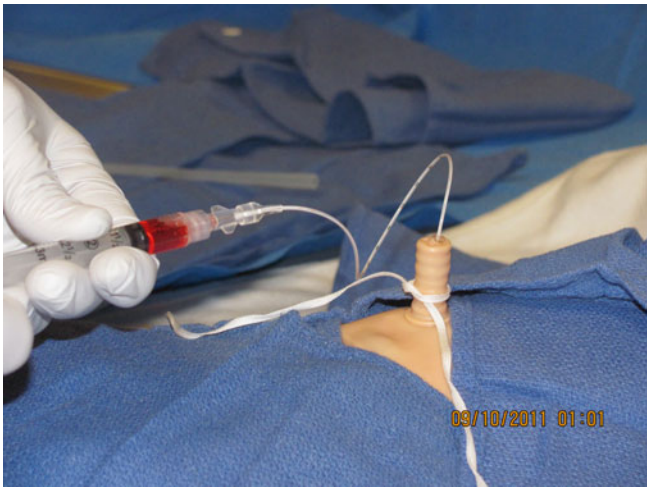

Fig. 121.5 Once there is easy blood return, pass the catheter an additional 1–2 cm. The catheter should still be inferior to the liver at this point

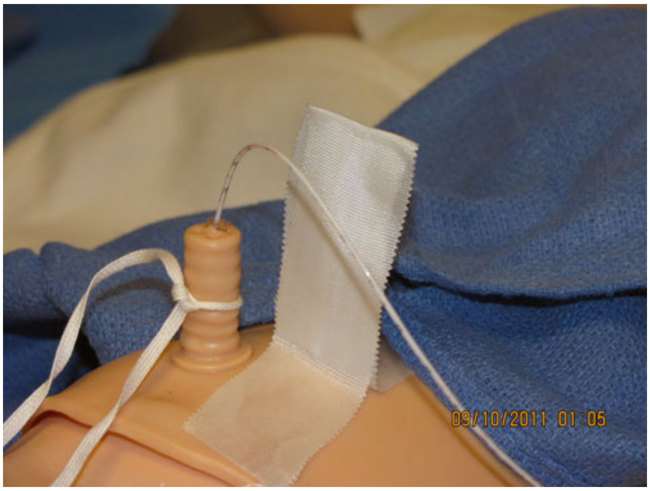

Fig. 121.6 Creating an umbilical catheter tape bridge: the uprights

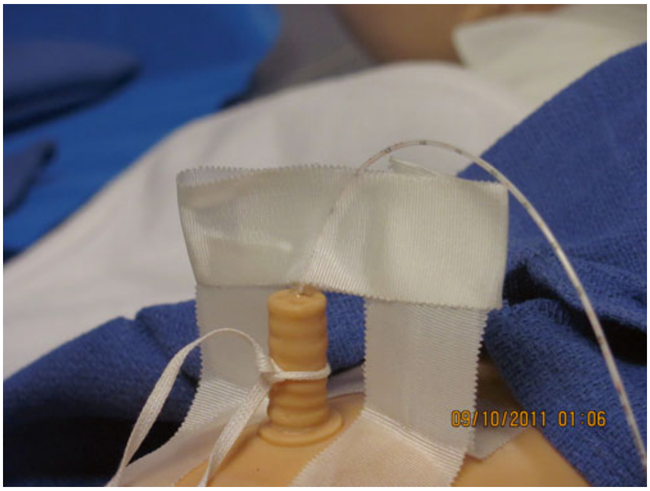

Fig. 121.7 Umbilical tape catheter bridge: the "cross bar" secures the catheter without applying tape to the umbilical stump

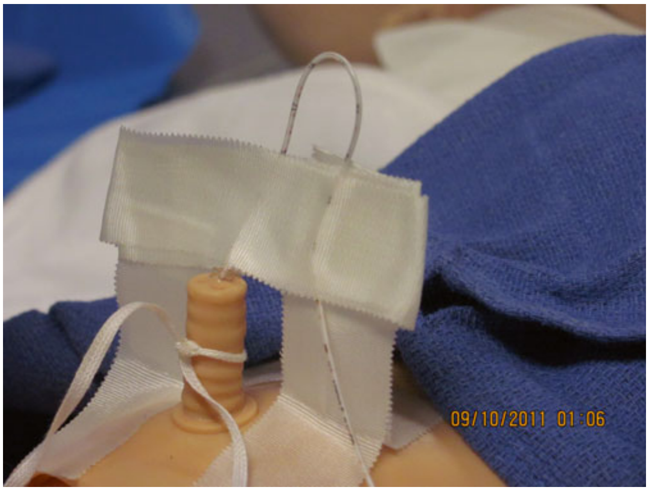

Fig. 121.8 Fold the catheter over to create a "U." Be careful not to kink the catheter, and secure with a second taped cross bar to lend additional security to the catheter and prevent displacement

121.5 Procedure (Removal)

1. The UVC should be removed as soon as adequate peripheral venous access is obtained (unless in an infant estimated to weigh <1000 g).
2. Turn infusions off.
3. Be certain the stopcock is closed to the infant.
4. It is imperative that there be no air in the catheter before withdrawal (if air is present and infant takes inspiration, the negative pressure generated can pull significant amount of air into the central vasculature).
5. Remove the securing tape from the infant.
6. Withdraw the catheter gradually as a single maneuver.

121.6 Complications

- Infection
- Bleeding due to disconnection of tubing (*always use Luer-Lok connections*) or perforation of vessels
- Arterial injury by accidental perforation
- Hepatic injury and necrosis if the catheter sits within a portal vein
- Thrombosis
- Air embolus
- Dysrhythmia or pericardial tamponade or perforation if catheter is advanced too far

121.7 Pearls and Pitfalls

- Pearls
  - The umbilical vein is 2–3 cm long before it widens into the umbilical recess, just before intersecting with the left portal vein and the ductus venosus.
  - Be certain to include the length of the umbilical stump in any calculations for placement.
  - If using the calculating graph, measure from the right shoulder to the umbilicus (Fig. 121.9).
- If choosing the above-diaphragm site for placement, corkscrewing the catheter clockwise while passing the catheter will encourage it to pass through the ductus venosus.
  - A kidney, ureter, and bladder study is mandatory in high UVC placement.
- Pitfalls
  - When preparing with Betadine (povidone-iodine), be certain to wipe away and remove any pooled Betadine along the infant's side because this will cool the infant initially; as the Betadine is warmed, it becomes highly irritating to newborn skin.

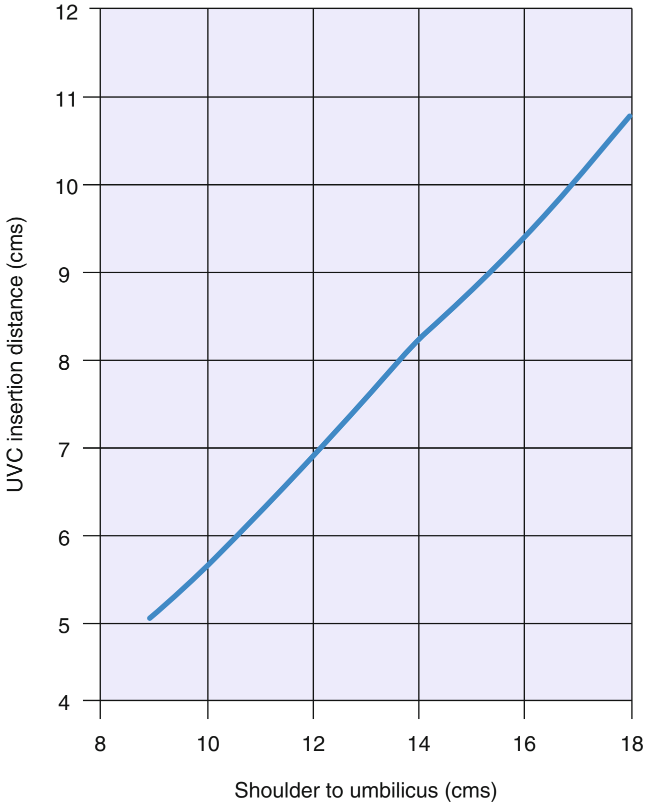

Fig. 121.9 Graph for estimation of insertion distance of the catheter; be certain to add the length (cm) of the umbilical stump to the insertion distance

Selected Reading

Sudbury, Jones, Bartlett. Emergency vascular access. In: APLS: the pediatric emergency medicine resource. 5th ed., p. 741.

Magnan JP. Umbilical Vein Catheterization. http://emedicine.medscape.com/article/80469-overview.

Schlesinger AE, Braverman RM, DiPietre MA. Neonates and umbilical venous catheters: normal appearance, anomalous positions, complications, and potential aid to diagnosis. Presented at the 2001 annual meeting of the American Roentgen Ray Society, Seattle, Apr 2001.

Intraosseous Access

122

Judith K. Lucas

122.1 Indications

- During cardiac arrest: Failure to attain vascular access after three peripheral intravenous attempts or 90 s, whichever comes first
- Inability to gain vascular access in pediatric patients presenting in shock due to hemorrhage (trauma), sepsis, profound dehydration, or cardiac failure

122.2 Contraindications

- Absolute
  - Fracture of the long bone considered for intraosseous (IO) access
- Relative
  - Previous IO access attempt in the same long bone
  - Cellulitis over the insertion site
  - Inferior vena cava injury (circulatory access proximal to the injury site is preferred).
  - Osteogenesis imperfecta

122.3 Materials and Medications

- Preparation materials, such as an antiseptic solution and sterile drapes, IF the patient's stability offers the time.
- Lidocaine without epinephrine: If it becomes necessary to place an IO in a conscious patient, use lidocaine to anesthetize the skin to the bony cortex; then once accessed, infiltrate 2–3 mL into the marrow to alleviate some of the pain of medications infusing through the marrow.
- IO needle (a few examples shown) (Figs. 122.1, 122.2, and 122.3).

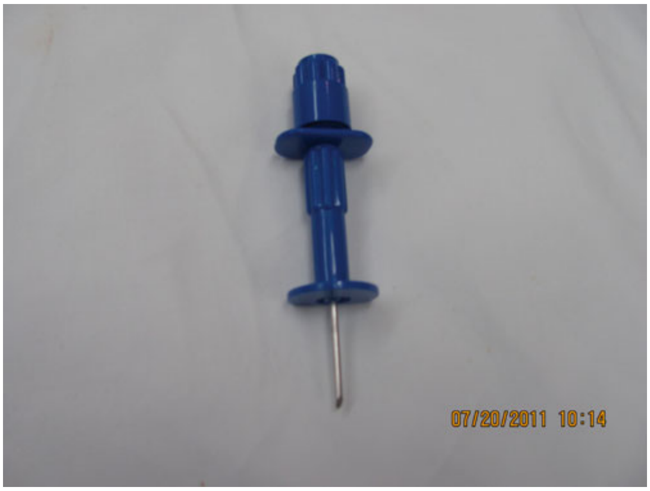

Fig. 122.1 Jamshidi disposable sternal/iliac aspiration needle (Jamshidi, Cardinal Health Dublin, OH)

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_122

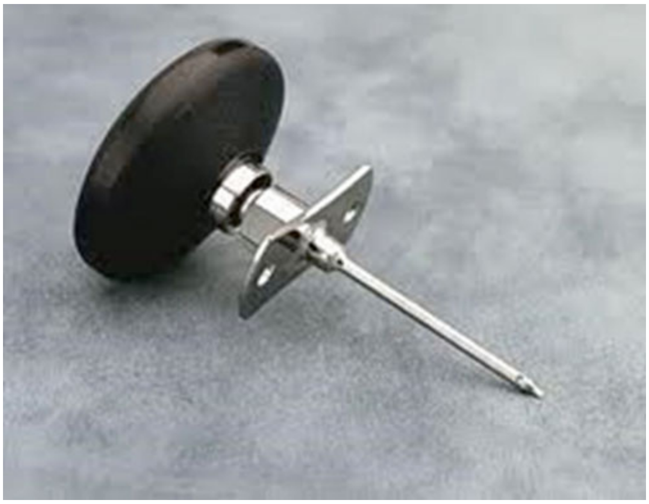

Fig. 122.2 Cook intraosseous needle (Cook Critical Care, Bloomington, IN)

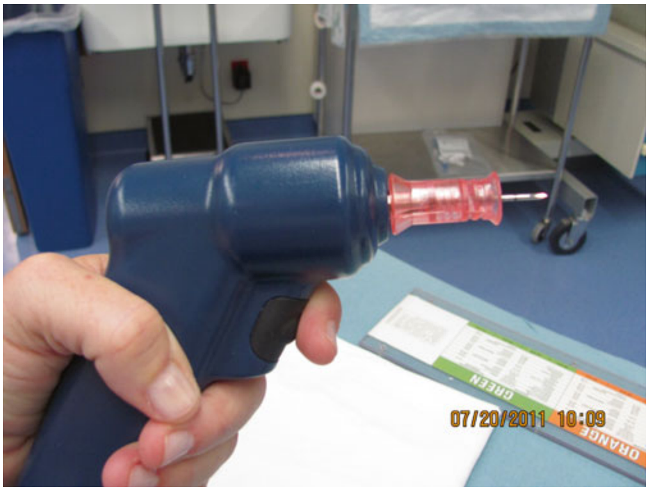

Fig. 122.3 EZ-IO (Vida-Care, San Antonio, TX)

122.4 Procedure

- Most common site: proximal tibia
  1. Patient should be supine, with the intended leg slightly externally rotated and flexed at the hip. Flex the knee about 90°.
  2. Place a towel underneath the knee and the proximal lower leg.
  3. Palpate the tibial tuberosity. Then move fingers 2 cm distal to the tuberosity (1–2 fingerbreadths) and 2 cm medial. This area is consistently flat and is distal to the growth plate (Figs. 122.4 and 122.5).
  4. If the patient is stable, create a sterile field focused on the insertion site.
  5. Insertion site should be cleaned with an antiseptic solution.
  6. Pass the needle through the skin and subcutaneous tissue to the cortex.
  7. Once bone is reached, stabilize the IO with the thumb and first fingers adjacent to where the needle penetrates the skin.
  8. Apply steady and firm pressure downward and with a twisting motion (back and forth) with the palm on the end of the device. Although inserting the needle perpendicularly to the bone is totally acceptable, ideally angling the needle caudally slightly (~15°) will avoid the growth plate (Fig. 122.6).
  9. As the cortex is fully penetrated indicating entrance into the marrow space, one *may* feel an ease in resistance. Do not advance the needle further. The needle should be able to stand up without support.
  10. Remove the inner trocar (Fig. 122.7).
  11. Attempt to aspirate marrow/blood. Not being able to do so DOES NOT mean inadequate or inaccurate placement. Instead, if marrow cannot be aspirated, affix a 10-mL syringe with normal saline and attempt to infiltrate. Resistance to flow should be minimal, and one should not appreciate either extravasation around the insertion site, coolness, or tissue expansion posterior to the site (indicating the fluid is passing through the IO).

12. Once placement is confirmed, connect an intravenous line, using a three-way stop-cock.
13. Secure the needle with tape and gauze (Fig. 122.8).

- Alternative sites
  - Distal tibia
    - The landmarks are the medial aspect of the tibia (the flat portion), 2 fingerbreadths proximal to the medial malleolus.
    - Again, externally rotate and abduct the hips, with the knee flexed about 60° (as with the proximal tibia).
    - Angle the needle toward the knee (cephalad) about 10–15° to avoid the distal tibia growth plate.
    - The remainder of the IO insertion is identical to the procedure for proximal tibia placement.
  - Distal femur
    - Slightly flex and externally rotate the hip.
    - Flex the knee enough that the quadriceps muscle group is relaxed.
    - The landmarks for the distal femur are the anterior thigh, midline, about 3 fingerbreadths proximal to the medial and lateral condyles.
    - The IO is inserted perpendicular to the bone (because there is no growth plate in the distal femur other than the condyles).
    - The remainder of the procedure is identical to the procedure for placement in the proximal tibia.
  - Proximal humerus (Fig. 122.9)
    - The patient should be supine, with the shoulder, upper arm, and elbow as close to the body as possible, yet still on the bed. The elbow should be flexed at 90°, with the forearm and palm resting on the patient's abdomen.
    - The provider slides his or her thumb up the anterior shaft of the humerus toward the shoulder until the greater humerus tubercle is palpated, which identifies the surgical neck of the humerus.
    - The insertion site is perpendicular to the humerus, approximately 1–2 fingerbreadths proximal to the tubercle.
    - The remainder of the insertion is identical to that for the proximal tibia.
  - Sternum (not usually recommended for small children).
    - Requires a special IO needle and system.
    - Method of placement is specific to the insertion system chosen.
    - Manubrium is the desired site (as opposed to the body of the sternum, which would interfere with cardiopulmonary resuscitation).
    - Risks specific to sternal placement include pneumothorax, mediastinitis, and great vessel injury.

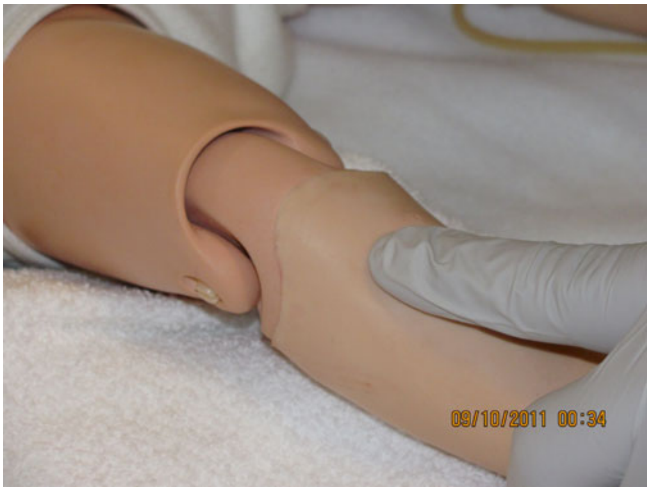

Fig. 122.4 Step 1. Identify the tibial tuberosity

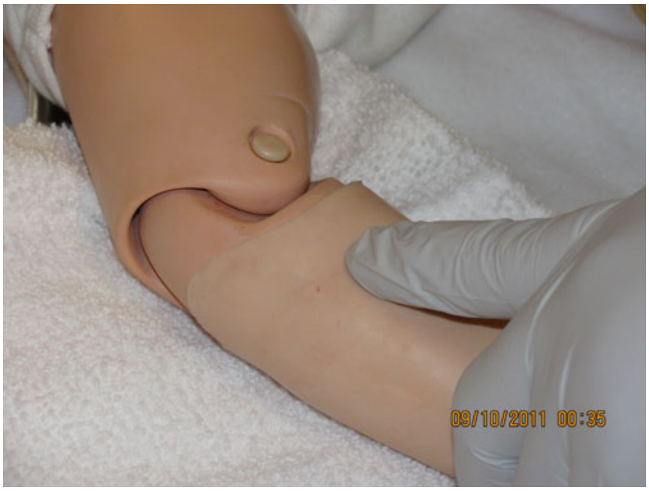

Fig. 122.5 Step 2. One to 2 fingerbreadths below the tuberosity and medially, to the flat aspect of the proximal tibia

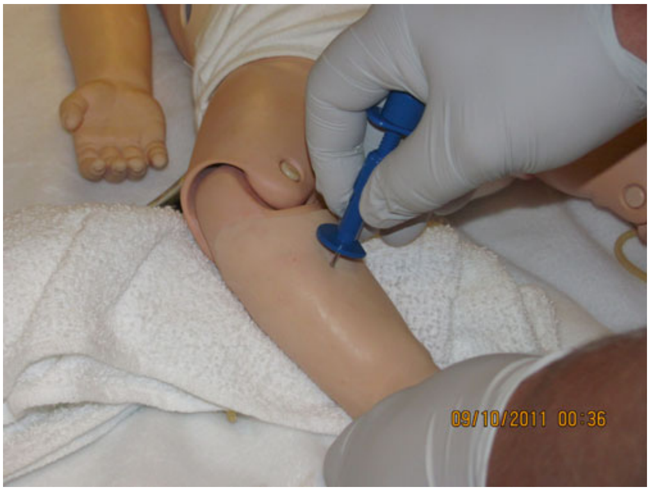
09/10/2011 00:36

Fig. 122.6 Step 3. Angle the device slightly caudal to avoid growth plate

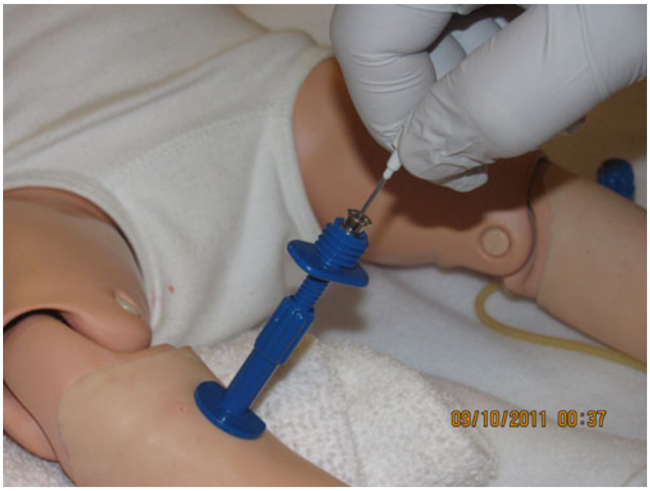
09/10/2011 00:37

Fig. 122.7 Step 4. Uncap the top of the device and remove the trocar

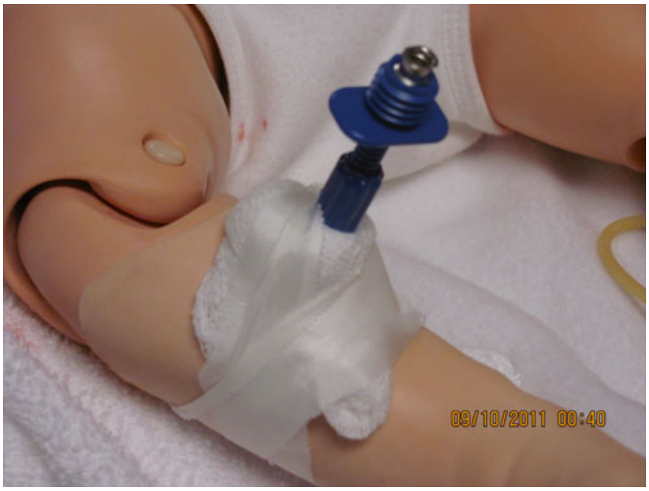
09/10/2011 00:40

Fig. 122.8 Step 5. Stabilize the intraosseous cannula

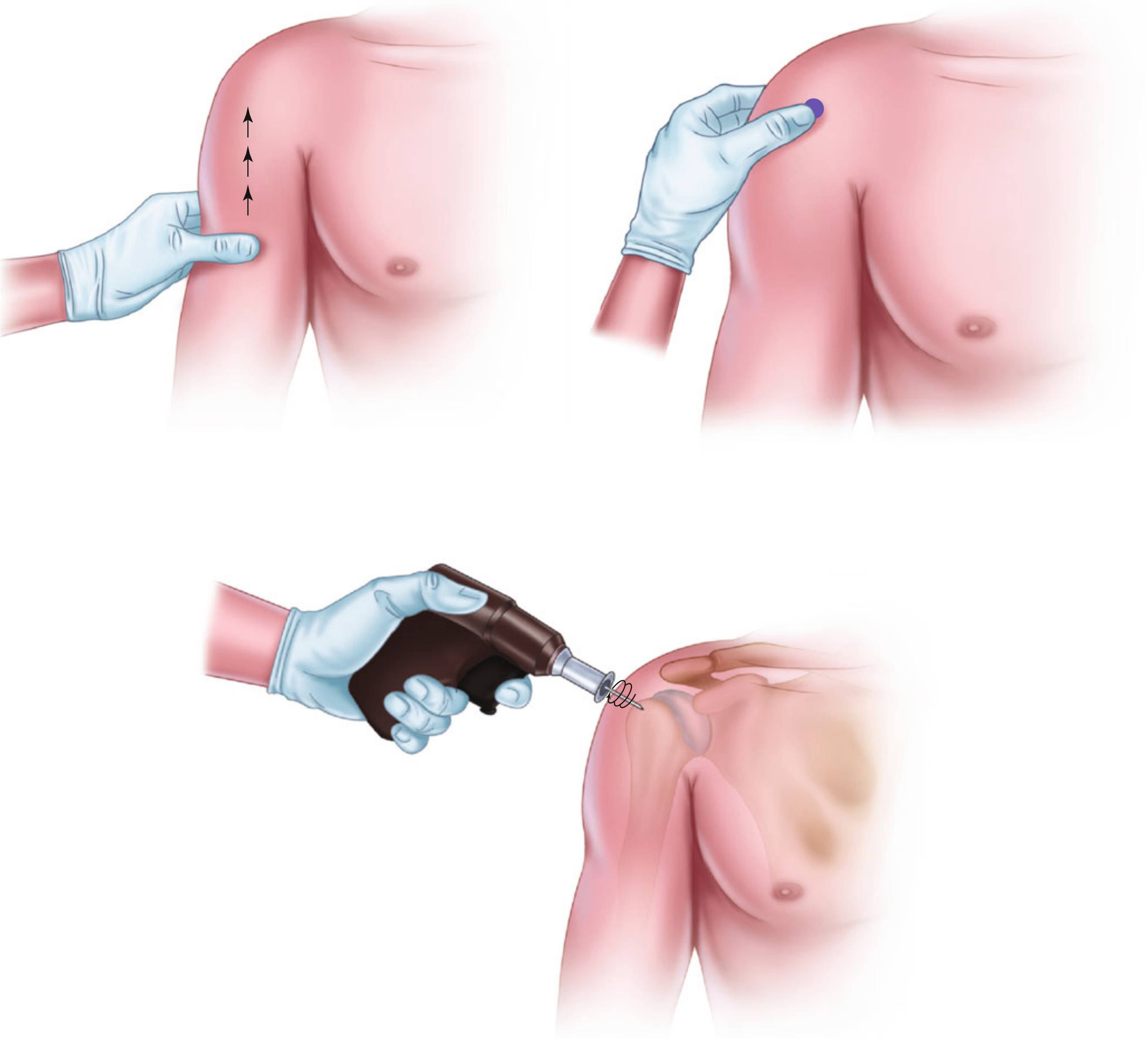

Fig. 122.9 IO insertion into proximal humerus site

122.5 Complications

- Extravasation of fluid.
  - Occurs as a result of a misplaced IO, either not completely in the marrow space anteriorly or through the cortex posteriorly.
  - Tissue necrosis.
- Compartment syndrome if extravasation is not recognized.
- Fracture and growth plate injury.
- Infection and osteomyelitis are rare complications if sterile technique is used.
- Fat embolism, considered rare and only in adult patients.

122.6 Pearls and Pitfalls

- Be certain to avoid placing a hand beneath the IO site during placement in order to prevent possible IO penetration into the provider's tissue.
- Do not rely on the sensation of a "pop" to determine appropriate penetration into the cortex, especially in infants.
- Likewise, one may not be able to aspirate bone marrow or blood, even with an appropriately placed IO. DO NOT pull out the IO if marrow cannot be aspirated. Rather, gently but firmly infiltrate 10 mL normal saline.
- Blood gases, body chemistries, and blood typing can be obtained from IO samples, but the sample CANNOT be used for hematocrit determination.

Selected Reading

Bohn D. Intraosseous vascular access: from archives to the ABC. Crit Care Med. 1999;27:1053.

DeCaen AR, Reis A, Bhutta A. Vascular access and drug therapy in pediatric resuscitation. Pediatr Clin North Am. 2008;55:909–27.

Deitch K. Intraosseous infusion. In: Roberts JR, Hedges JR, editors. Clinical procedures in emergency medicine. 5th ed. Philadelphia: Saunders; 2009. Chap. 25.

EMS World. Intraosseous infusion: not just for kids anymore. Posted Jan. 12, 2011. Updated from Mar 2005. EMSWorld.com. Cygnus Business Media Site.

Halm B, Yamamoto LG. Comparing ease of intraosseous needle placement: Jamshidi versus Cook. Am J Emerg Med. 1998;16:420.

Lumbar Puncture in Pediatrics

123

Maritza A. Plaza-Verduin and Judith K. Lucas

123.1 Indications

- Evaluation of cerebrospinal fluid (CSF) for infection or malignancy
- Measurement of opening pressure
- Treatment of pseudotumor cerebri
- Diagnosis of central nervous system (CNS) metastases
- Instillation of intrathecal chemotherapy
- Injection of radiopaque dye for spinal cord imaging

123.2 Contraindications

- Increased intracranial pressure
- Bleeding diathesis (platelet count <50,000)
- Overlying skin infection near the area of puncture site
- Spinal cord trauma or spinal cord compression
- Signs of progressive cerebral herniation
- Condition of the patient (e.g., unstable airway, potentially dangerous breathing problem, severe circulatory instability) that could cause an abrupt decompensation
- Known spinal cord deformity

123.3 Materials and Medications

- Lumbar puncture (LP) tray (Fig. 123.1)
  - Sterile drapes
  - Betadine (povidone-iodine) swabs or tray to pour Betadine
  - Sterile sponges for preparing the puncture site
  - Sterile 3-mL syringe with needle for lidocaine injection
  - Sterile collecting tubes (4)
  - Sterile spinal needle with stylet (size depending on age of patient)
    - Premature infant: 22 gauge or smaller, 1.5 inch
    - Neonate to 2 years: 22 gauge, 1.5 inch
    - 2–12 years: 22 gauge, 2.5 inch
    - Older than 12 years: 20 or 22 gauge, 3.5 inch
  - Pressure manometer column with a three-way stopcock
- Betadine solution
- Sterile gloves
- Mask
- Lidocaine (1–2 % without epinephrine)
- 4 % Lidocaine cream (LMX-4) or lidocaine and prilocaine mixture (EMLA [eutectic mixture of local anesthetics])

M.A. Plaza-Verduin, MD
Pediatric Division, Department of Emergency Medicine, Arnold Palmer Hospital for Children, Orlando, FL, USA
e-mail: mari.plazav@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_123

Fig. 123.1 Lumbar puncture (LP) tray

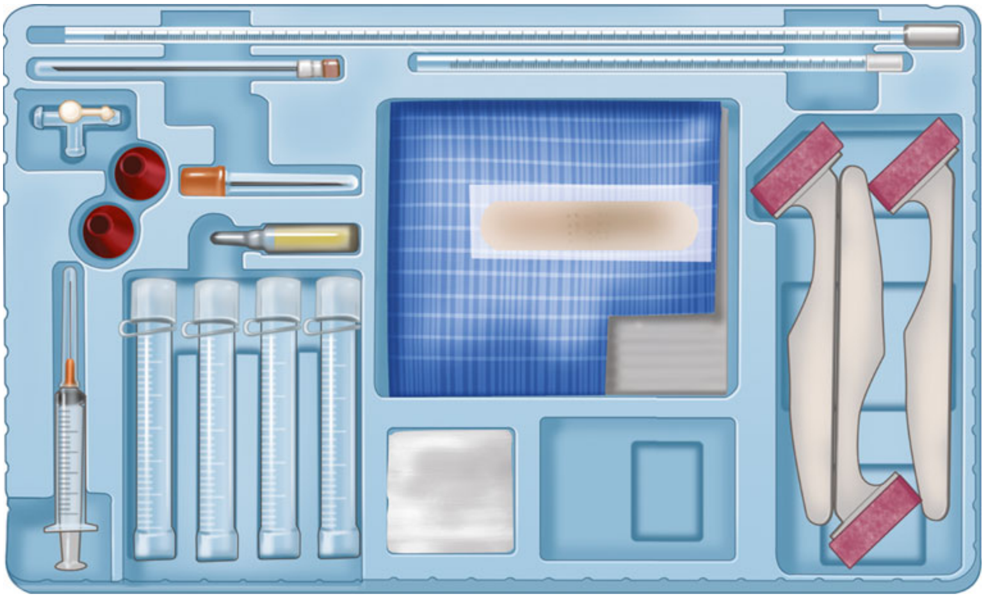

123.4 Procedure

1. Position the child (Figs. 123.2 and 123.3).
    (a) Can be positioned in either the sitting or the lateral recumbent position with the hips, knees, and neck flexed.
    (b) For small infants or any patient with any degree of cardiorespiratory compromise, keep close monitoring of heart rate, respirations, and oxygen saturation while in the flexed position.
2. Palpate the top of the iliac crest and draw an imaginary line connecting the two across the back, which should cross the midline just above the fourth lumbar spine (Fig. 123.4).
3. Palpate L3–4 or L4–5 space along this line.
4. Place EMLA or LMX-4 on the area at this time and allow some time for anesthesia to occur (can take up to 30 min to be effective).
5. Prepare the skin in sterile fashion with Betadine solution using enlarging circles that begin at the puncture site.
6. Drape the patient with sterile towels, exposing the puncture site.
    (a) If an infant, do so conservatively to be able to monitor the infant during the procedure.
7. Locate the intervertebral space (L3–4 or L4–5) once again.
    (a) Make a small mark on the chosen intervertebral space with a fingernail depression or the plastic cap of the spinal needle.
8. If desired, or if anesthetic cream was not previously used or more is needed, apply a small wheal of lidocaine at the desired puncture site using a 25-gauge needle.
9. Insert the spinal needle in the intervertebral space.
    (a) Puncture the skin in the midline just caudal to the palpated spinous process.
    (b) Bevel should be positioned so that the dura mater is pierced parallel to its fibers (which will reduce the likelihood of CSF leakage).
        (i) If in the lateral recumbent position, the bevel of the needle should be positioned horizontally.
        (ii) If in the sitting position, the bevel of the needle should be positioned vertically.
        (iii) Angle the needle slightly cephalad toward the umbilicus and parallel to the bed if the patient is in the lateral recumbent position or slightly caudal (perpendicular to the skin) if the patient is in the sitting position.
10. Advance the needle several millimeters at a time and withdraw the stylet frequently to check for CSF flow.
    (a) Advance the needle until a loss of resistance is felt or approximately 1–2 cm.
        (i) In the infants, whose dura are not so thick and a "pop" or give may be unnoticeable, after passing between the spiny processes and approximately 2 cm through the skin, remove the stylet frequently to check for CSF return. This will allow you to avoid passing through the subarachnoid space.
        (ii) After the change in resistance occurs, if resistance is met again, pull back gently on the needle to reposition in the subarachnoid space and remove the stylet to check for CSF fluid.
11. Once CSF is free flowing, attach a pressure manometer to the needle hub via the three-way stopcock (Fig. 123.5).
    (a) Make sure to hold the spinal needle with one hand in place while attaching the manometer to prevent movement of the needle.
    (b) Measure the CSF pressure once the CSF reaches the highest level in the manometer column.
    (c) CSF pressures will be best obtained in the lateral recumbent position with the neck and legs extended.
    (d) Normal CSF pressure is 5–20-cm H_2O with the neck and legs extended, 10–28-cm H_2O with the neck and legs flexed.
    (e) Have an assistant hold the top of the pressure manometer while the spinal needle is supported at the connection of the manometer and stopcock.
12. Remove the manometer and collect the CSF into sterile tubes.
    (a) Make sure to keep the needle in place when removing the manometer.
    (b) Continue to drain fluid into the collecting tubes, approximately 1 mL per tube.
    (c) The tubes should be labeled for specific studies, depending on the order of collection.
        (i) First tube for gram stain and culture.
        (ii) Second tube for quantitative glucose and protein.
        (iii) Third or fourth tube for cell count and differential.
        (iv) Leftover tube for any additional studies that may be needed.
        (v) When the LP is to assess for possible subarachnoid blood, the first and fourth tubes are sent for cell count.
13. Replace the stylet and remove the spinal needle.
14. Cleanse the puncture area to avoid staining with the Betadine solution.
15. Apply a sterile dressing to the puncture site.
16. Patients 4 years and older should remain in the supine position, with the head elevated no more than the height of a pillow (for comfort), for at least an hour to avoid spinal tap headaches.
17. Likewise, after the LP, giving a fluid bolus of normal saline (10–20 mL/kg in children; 1 L in non-volume-sensitive adults) will also assist with avoidance of spinal tap headaches.

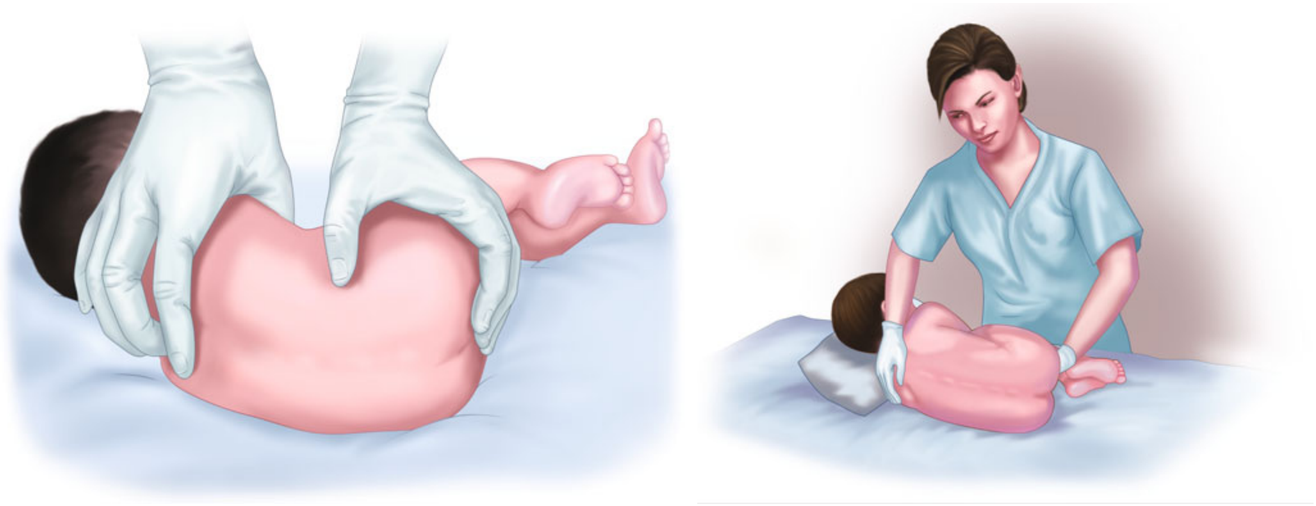

Fig. 123.2 Curling up for an LP

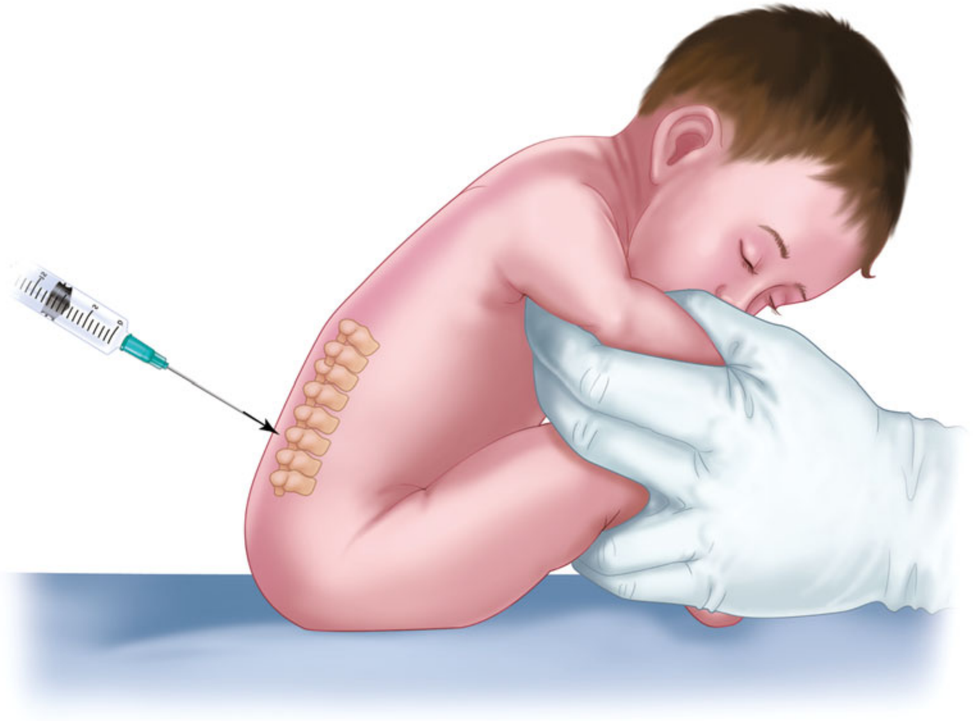

Fig. 123.3 Having the LP

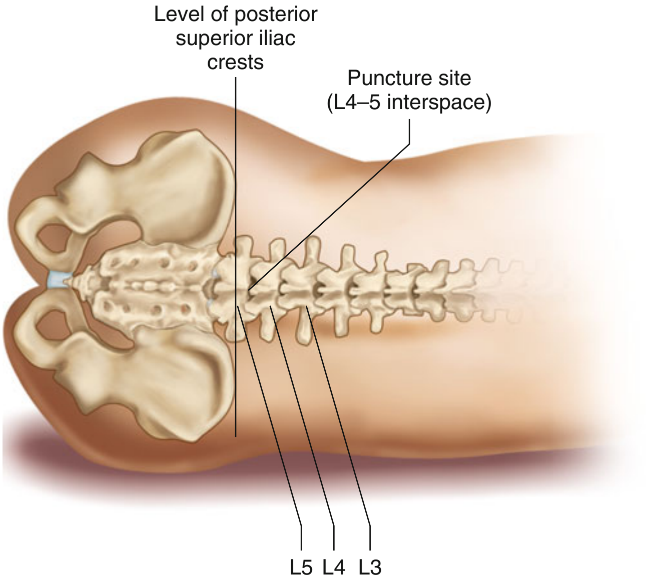
Level of posterior superior iliac crests
Puncture site (L4–5 interspace)
L5
L4
L3

Fig. 123.4 Anatomy of the lumbar spine showing the sites for dural puncture

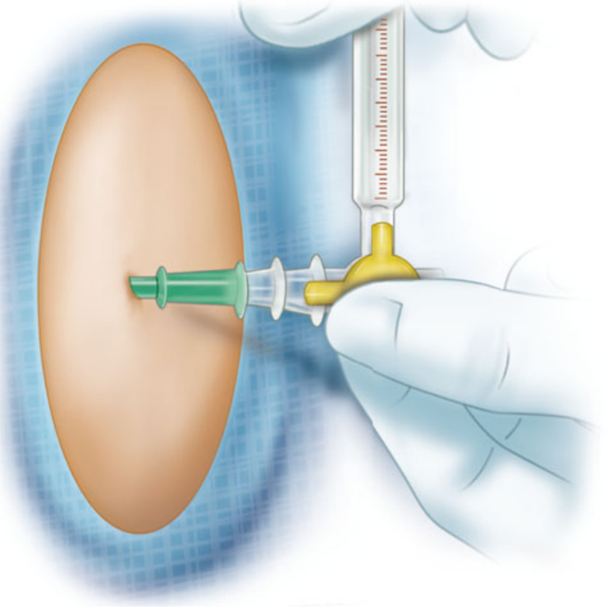

Fig. 123.5 Pressure manometer attached to the needle hub via the three-way stopcock

123.5 Complications

- Minor
  - Localized back pain
  - Transient paresthesia during procedure
  - Post-LP headache
- Major
  - Severe back pain associated with neurological signs (may be subdural or epidural spinal hematoma)
  - LP-induced meningitis
  - Cerebral herniation
  - Acquired epidermoid tumor
  - Damage to adjacent structures (disk herniation, retroperitoneal abscess, spinal cord hematoma)

123.6 Pearls and Pitfalls

- Pearls
  - Success of the LP depends on the positioning of the patient.
    - Goal of positioning is to stretch the ligamentum flavum and increase the interlaminar spaces.
    - In the recumbent position, the shoulder and hips should be perpendicular to the bed, keeping the spinal cord straight, with no rotation.
    - Sitting position is useful with older, cooperative patients or with very young infants who are unlikely to struggle and may have increased respiratory distress in the lateral recumbent position.
  - Administration of an anxiolytic (e.g., midazolam) can be used to facilitate the procedure in an older child. In some cases, procedural sedation may be required.
  - Placing the tip of the thumb on the spinous process just above the space being entered can ensure good alignment of the needle.
  - In older patients a “pop” may often be felt as a change in resistance occurs once the dura is penetrated; however, in infants and neonates, that “pop” may be exceedingly subtle or not palpable.
  - If no CSF returns, attempt the following options:
    - Ensure the needle is in the appropriate position, withdrawing the needle slowly if necessary.
    - Rotate the needle 90°.
    - If an infant, have the assistant massage the anterior fontanel to help facilitate CSF flow.
    - If the procedure continues without CSF, withdraw the needle to just under the skin and redirect it.
    - If the procedure continues without CSF, withdraw the needle and insert a new needle with stylet at an alternate site.

    - If these steps do not yield CSF, the infant may be dehydrated, not allowing for adequate CSF flow.
      - Give the patient a bolus and reattempt later.
      - Attempt putting the infant in a sitting position to increase flow.
- Pitfalls
  - Traumatic LP
    - Bloody CSF fluid, which usually clears as the CSF drains if the needle is in the correct space.
    - Occurs with improper technique (inserting the needle too far to one side into an epidural venous plexus or through the subarachnoid space into or adjacent to the vertebral body).
    - Can occur with proper technique as well.
    - If fluid does not clear and clots form in the tubes, LP should be reattempted at a different site.
  - Failed LP attempts despite proper procedure and positioning
    - Ultrasound can be used to visualize the area and determine the reason for failure or the likelihood of success with future attempts.

Selected Reading

Coley BD, Shiels WE, Hogan MJ. Diagnostic and interventional ultrasonography in neonatal and infant lumbar puncture. Pediatr Radiol. 2001;31:399–402.

Cronan KM, Wiley JF. Lumbar puncture. In: King C, Henretig FM, editors. Textbook of pediatric emergency procedures. 2nd ed. New York: Lippincott Williams & Wilkins; 2008.

Ebinger F, Kosel C, Pietz J, Rating D. Headache and backache after lumbar puncture in children and adolescents: a prospective study. Pediatrics. 2004;113:1588–92.

Friedman AG, Mulhern RK, Fairclough D, et al. Midazolam premedication for pediatric bone marrow aspiration and lumbar puncture. Med Pediatr Oncol. 1991;19:499–504.

Partin WR. Emergency procedures. In: Stone CK, Humphries RL, editors. Current diagnosis & treatment: emergency medicine. 6th ed. New York: McGraw-Hill; 2007.

Suprapubic Bladder Aspiration

124

Maritza A. Plaza-Verduin and Judith K. Lucas

124.1 Indications

- Collection of sterile urine for urinalysis and culture (avoiding urethral contamination)
- Collection of sterile urine in a child with gastroenteritis and frequent diarrheal stools
- Female child with labial adhesions or male child with minimally retractable foreskin
- Urinary retention

124.2 Contraindications

- Empty or nonpalpable bladder
- Urination within 1 h before the procedure
- Anatomical abnormalities of the gut or genitourinary tract
- Bleeding diathesis
- Intestinal obstruction
- Overlying cellulitis on abdominal wall
- Lower abdominal scars or wounds

124.3 Materials and Medications

- Sterile gloves
- Lidocaine (1–2 %) with syringe and needle
- EMLA (eutectic mixture of local anesthetics) cream
- Betadine (povidone-iodine) solution
- Sterile syringe, 5–20 mL
- Sterile needle 22 or 23 gauge, 1.5 inch
- Sterile specimen container
- Sterile towels
- Sterile gauze
- Sterile dressing
- Adhesive bandage

M.A. Plaza-Verduin, MD
Pediatric Division, Department of Emergency Medicine,
Arnold Palmer Hospital for Children, Orlando, FL, USA
e-mail: mari.plazav@gmail.com

J.K. Lucas, MD (✉)
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016
L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_124

124.4 Procedure

1. Place the infant in the supine, frog-leg position (Fig. 124.1).
2. Localize the bladder.
3. Palpate the midline between the umbilicus and the pubis symphysis to feel for bladder fullness.
    - Use a portable ultrasound device to localize the bladder and allow for approximation of bladder size (see later).
4. Localize the symphysis pubis and the imaginary line midline from the umbilicus to the pubic symphysis (Fig. 124.2).
5. Sterilize the area from the umbilicus to the urethra.
6. Drape the area with sterile towels, keeping the puncture site area exposed.
7. Area of insertion should be midline 1–2 cm above the symphysis pubis on the abdominal wall.
    - The suprapubic crease can usually be used as a guideline to the puncture site.
8. Place a small wheal of anesthetic at the intended puncture site or can apply EMLA cream.
    - For use of EMLA cream, place before sterilization and allow some time for anesthesia to occur.
9. Occlude the urethral opening to avoid spontaneous loss of urine specimen.
    - In males, apply gentle pressure to the base of the penis against the pubic symphysis.
    - In females, directly apply pressure to the urethral meatus.
10. Puncture the skin with the needle (attached to the syringe).
    - Puncture at a 10–20° angle to the perpendicular, aiming slightly cephalad (Fig. 124.3).
11. Apply negative pressure to the syringe as the needle is advanced until urine enters the syringe; do not advance more than 1 inch.
    - If unsuccessful, draw back the needle until it rests in the subcutaneous fat and redirect 10° in either direction.
    - Do not attempt more than three times.
12. Cleanse the area of antiseptic solution and apply an adhesive dressing.

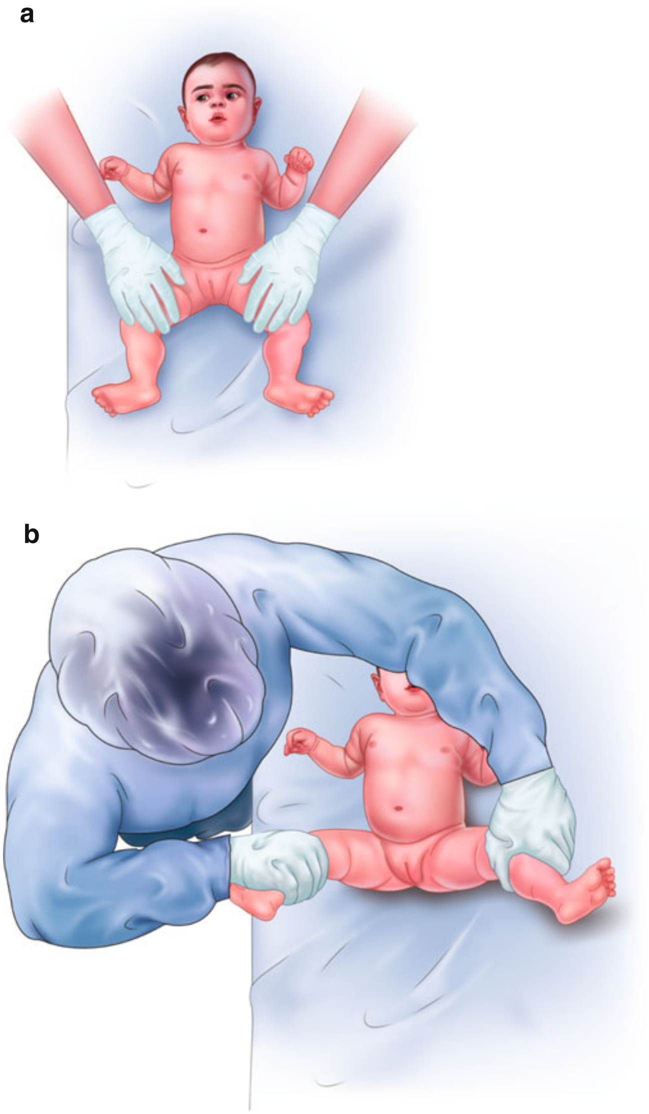

Fig. 124.1 Frog-leg position

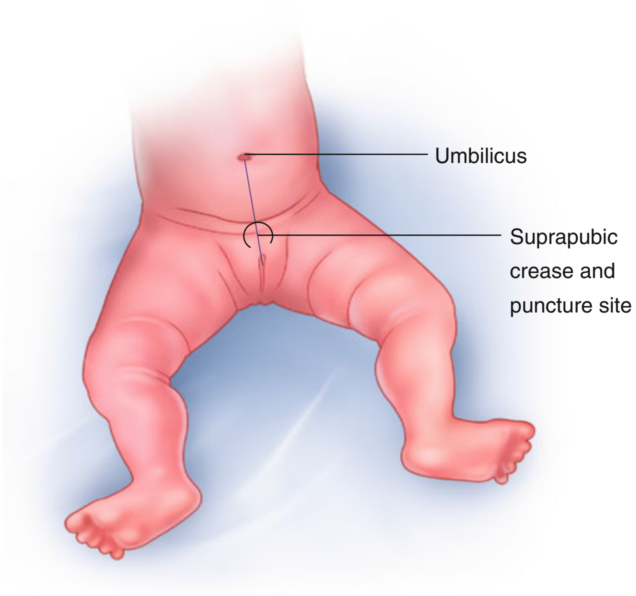
Umbilicus
Suprapubic crease and puncture site

Fig. 124.2 Landmarks

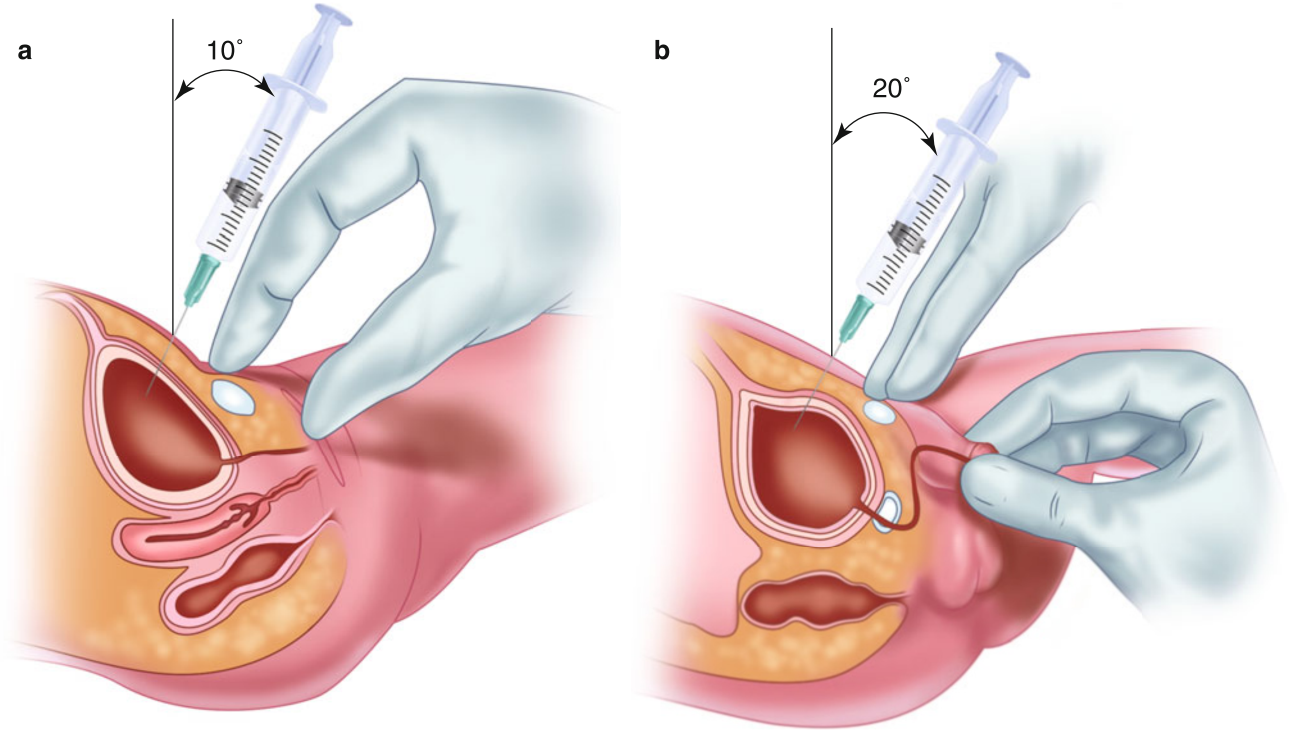

Fig. 124.3 Insert the needle 10–20° perpendicular to the skin

124.5 Complications

- Peritoneal perforation with or without bowel perforation
- Infection
- Hematuria
- Inability to aspirate urine

124.6 Pearls and Pitfalls

- If no fluid is obtained:
  - Hydrate the child and reattempt in 1 h.
- Ultrasound can be used to identify bladder size or for ultrasound-guided aspiration:
  - Use a portable ultrasound device with a standoff 7.5-MHz sector probe to allow for superficial scanning and measurement of the diameter of the bladder.
  - Bladder diameter measurement:
    1. Apply approximately 5 mL of ultrasound transmission gel to the infant's suprapubic region.
    2. Apply pressure to the ureteral meatus as previously described.
    3. Gently apply the probe to the suprapubic region in the midline and scan the area in transverse plane, with the probe directed caudad or cephalad as needed to maximize bladder image.
       - Bladder will appear anechoic below the brighter reflections of the rectus muscle and bladder wall.
    4. When maximum bladder size obtained, freeze the image.
       - Measure the anteroposterior and transverse internal bladder diameters.
       - Goal measurement is 2 cm or more of each diameter. If either diameter is less than 2 cm, the bladder is considered to be empty.
  - Ultrasound-guided bladder aspiration:
    1. After preparing the sterile field, place a sterile sheath over the ultrasound probe and proximal cable.
       (a) Gel should be placed within the sheath to eliminate the air interface between the probe and the sheath.
    2. Apply gel to the abdomen above the symphysis pubis.
    3. Locate the bladder and measure the maximum diameter as described previously.
    4. Insert the needle midline at the location where the bladder wall is closest to the probe.
    5. Continue steps for aspiration as described previously.

Selected Reading

Kozer E, Rosenbloom E, Goldman D, et al. Pain in infants who are younger than 2 months during suprapubic aspiration and transurethral bladder catheterization: a randomized, controlled study. Pediatrics. 2006;118:e51–6.

Leong Y, Tang KW. Bladder aspiration for diagnosis of urinary tract infection in infants and young children. J Singapore Paediatr Soc. 1976;18:43–7.

Loiselle JM. Ultrasound-assisted suprapubic bladder aspiration. In: King C, Henretig FM, editors. Textbook of pediatric emergency procedures. 2nd ed. New York: Lippincott Williams & Wilkins; 2008.

Pollack CV, Pollack ES, Andrew ME. Suprapubic bladder aspiration versus urethral catheterization in ill infants: success, efficiency and complication rates. Ann Emerg Med. 1994;23:225–30.

Polnay L, Fraser AM, Lewis JM. Complication of suprapubic bladder aspiration. Arch Dis Child. 1975;50:80–1.

Removal of Hair/Thread Tourniquet

125

Judith K. Lucas

125.1 Indications

- Consider the diagnosis when presented with an appendage that has a well-demarcated, circumferential, painful, edematous distal segment, adjacent to a nonedematous, nonerythematous proximal portion (Fig. 125.1).
- Removal is imperative in all cases of tourniquet syndrome and should be undertaken without delay.
- Method of removal is determined by degree of constriction.
  - Unwrapping: in situations with minimal or no edema and the ability to visualize the offending hair or thread
  - Blunt probe, cutting: mild to moderate edema
  - Incision: severe swelling or inability to visualize the constricting band; epithelialization
  - Depilatory: for areas with mild to moderate edema, but without epithelialization

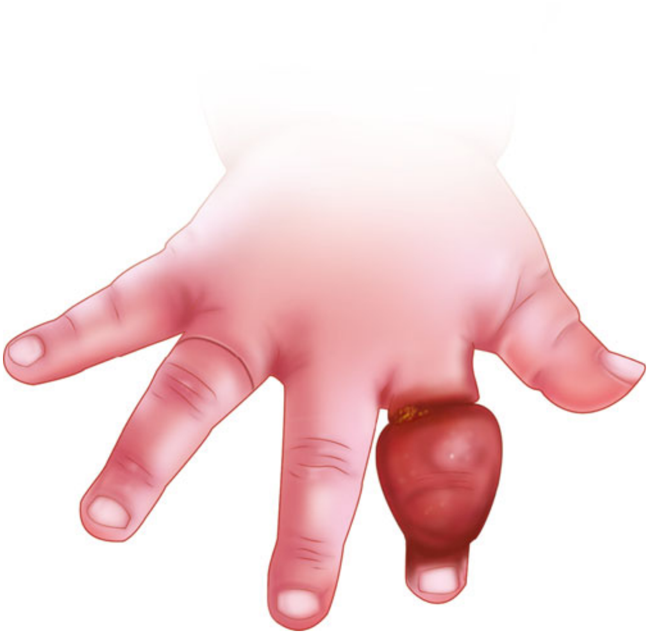

Fig. 125.1 Hair tourniquet

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_125

125.2 Contraindications

- Absolute
  - There are no absolute contraindications; the tourniquet must be removed.
- Relative
  - Based more on the specific approach (e.g., avoid incising the skin of a patient with hemophilia; avoid depilatory creams in a patient with known allergies to such)

125.3 Materials and Medications

- Unraveling technique
  - Fine-tipped, non-rat-tooth forceps or small fine hemostats
- Blunt probe and tourniquet cutting
  - Antiseptic solution of choice
  - Lidocaine 1 % without epinephrine for local or regional anesthesia
  - Scalpel blade #11 or Iris scissors
  - Blunt probe or metal earwax curette
- Incision technique
  - Antiseptic solution of choice
  - Lidocaine 1 % without epinephrine for local or regional anesthesia
  - Scalpel blade #11
  - Fine-tipped, non-rat-tooth forceps or small, fine hemostats
- Depilatory technique
  - Commercial depilatory cream

125.4 Procedures

- Unraveling method (Fig. 125.2)
  1. Place the appendage in an orientation that maximizes exposure.
  2. Apply skin traction such that as much of the tourniquet and base of the constriction can be seen.
  3. Look closely to identify a free end of the hair or thread. If no free end is visible, identify an area of bunching or a knot. Break the knot from the strand and grasp the end with the forceps or hemostat.
  4. Slowly and gently pull and unwind the hair.
  5. This may take several attempts because the hair strand or thread may break repeatedly during removal. Sometimes more than one hair strand may be involved.
- Blunt probe and cutting method (Figs. 125.3 and 125.4)
  1. Apply gentle traction to the skin to maximally expose the involved area and to make the base of the wedge caused by the tourniquet as shallow as possible.
  2. Gently impose a blunt probe or a metal ear curette between the skin and the tourniquet, starting proximally to the band and sliding distally beneath the constriction.
  3. Cut the tourniquet by sliding the scalpel blade along the edge of the probe, placing the blunt edge of the scalpel against the skin and slicing the constricting material in a movement away from the skin, in order to avoid inadvertently cutting the skin.
  4. Once an end of the tourniquet has been created, the remainder of the offending strap can be removed via unraveling.

- Incision technique for digits (Fig. 125.5)
  1. Perform a digital nerve block to anesthetize the extremity proximal to the constriction.
  2. Sterilize the involved area with antiseptic solution of choice and drape in typical protocol to allow as sterile a procedure as is possible outside a surgical suite.
  3. Incise the skin across the demarcation indicating the presence of the tourniquet at the 3 o'clock or 9 o'clock position or the midline dorsal area (12 o'clock), passing the scalpel blade proximally to distally and down to the bone (Fig. 125.6). Incision at these locations will avoid the laterally located neurovascular bundles of the digits.
- Incision technique for severe penile tourniquet
  1. Begin with a dorsal nerve block (Fig. 125.7).
  2. Sterilize the involved area with antiseptic solution of choice and drape in typical protocol to allow as sterile a procedure as is possible outside a surgical suite.
  3. A longitudinal incision is made at either the 4 o'clock or the 8 o'clock position, in order to avoid the penile neurovascular structures located dorsally.
  4. The incision should be made perpendicular to the tourniquet, but shallowly, in order to avoid penetration through the deep fascia. It will be necessary to repeat the incision, staying within the original incision, gently but repeatedly in order to cut through the whole tourniquet. In this way the constriction is relieved, but the integrity of the corpus cavernosum and the corpus spongiosum is maintained (Fig. 125.8).
  5. Once the layers of the tourniquet have been interrupted, grasp an end of the hair/thread and remove the remainder of the tourniquet utilizing the unraveling method.
- Depilatory method
  1. Depilatories work only on hair and will not work on threads.
  2. Apply the depilatory cream directly to the hair tourniquet with a saturated cotton swab, so as to avoid applying this potential irritant to surrounding tissue.
  3. Wait the recommended amount of time, as dictated by whatever brand of depilatory is used.
  4. After the appropriate time has elapsed, wash off the cream thoroughly with soap and water.

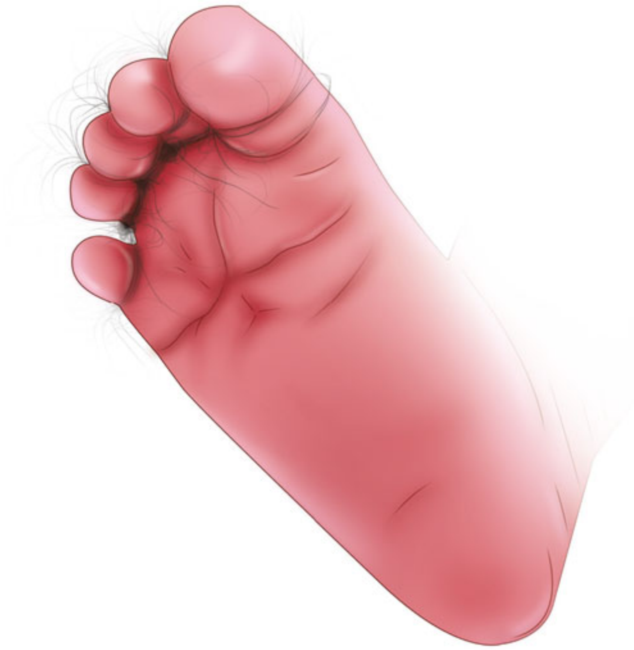

Fig. 125.2 A good example for use of the unraveling technique. The edema is minimal and the hair is easily identifiable

Fig. 125.3 Blunt and cut method. Notice the cutting edge of the scalpel blade is away from the skin as it is placed adjacent to the probe and slips beneath the tourniquet

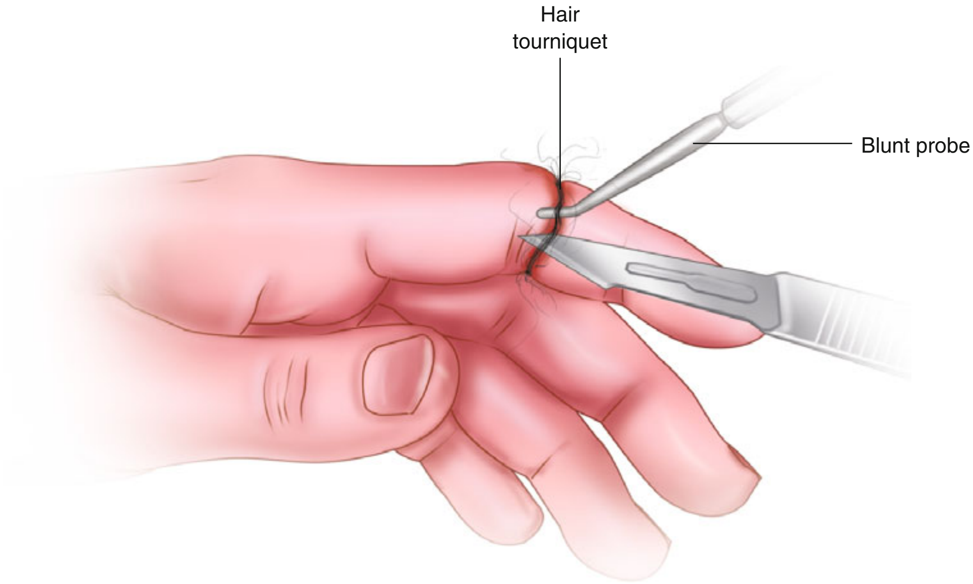
Hair tourniquet
Blunt probe

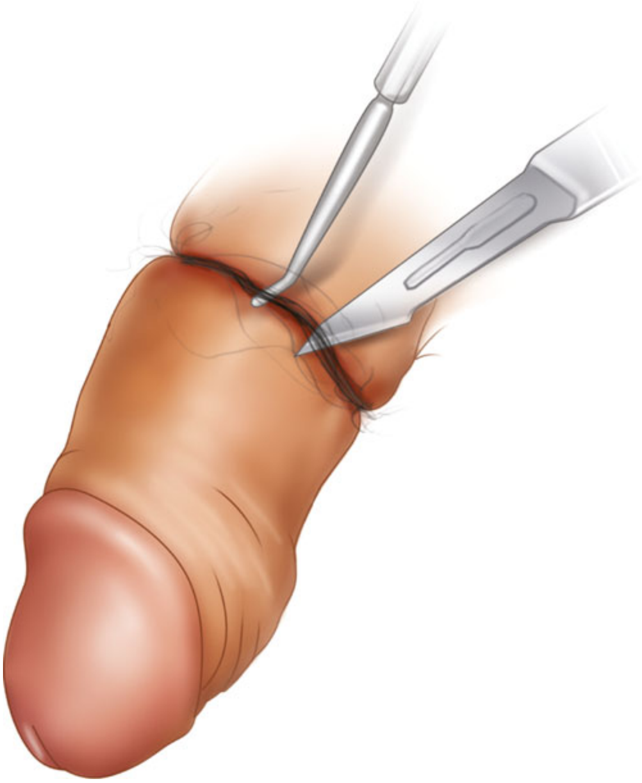

Fig. 125.4 Blunt and cut method

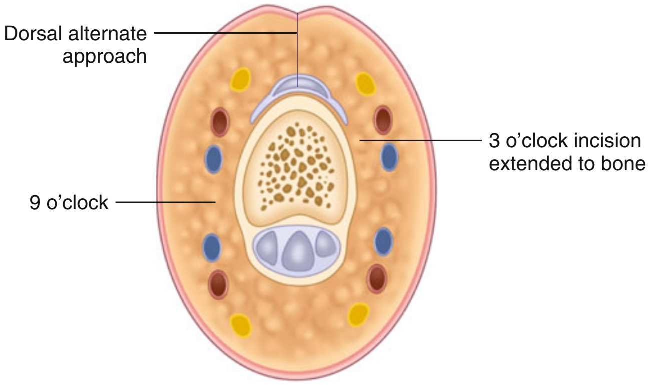
Dorsal alternate approach
3 o'clock incision extended to bone
9 o'clock

Fig. 125.5 Incision technique for digits

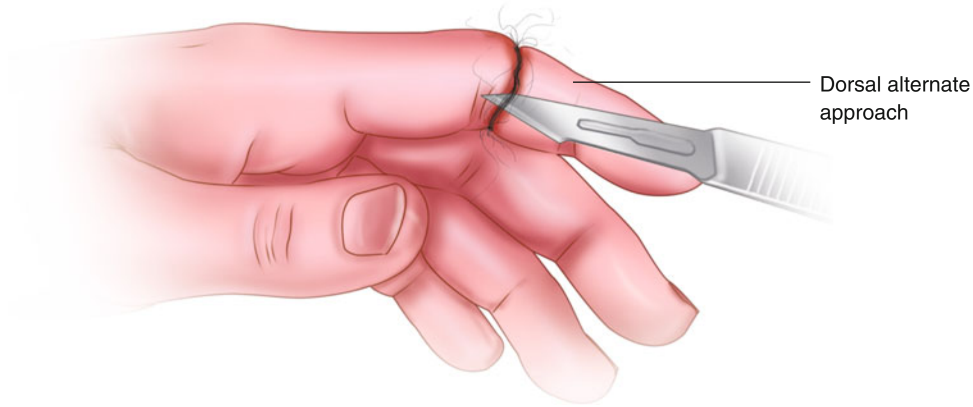
Dorsal alternate approach

Fig. 125.6 Incision at 12 o'clock preferred

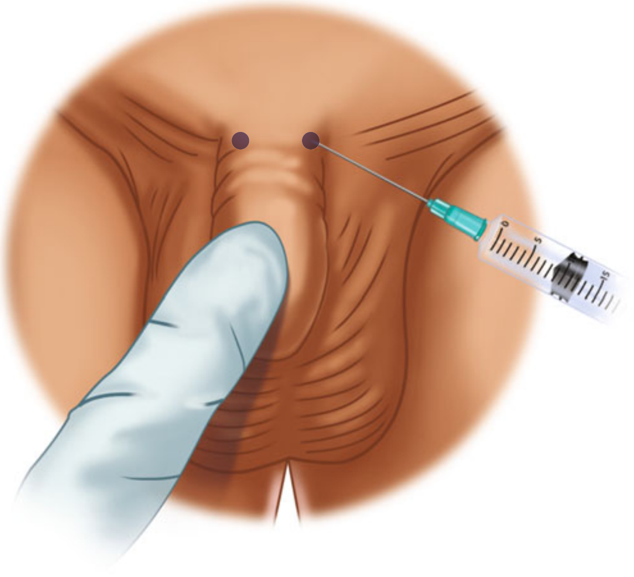

Fig. 125.7 Dorsal nerve block

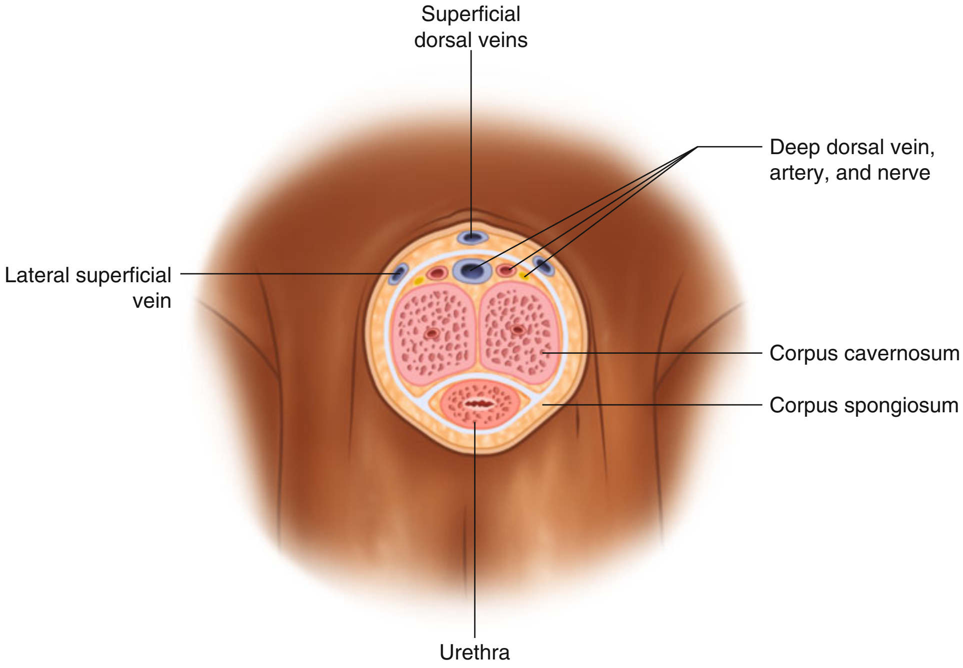
Superficial dorsal veins
Deep dorsal vein, artery, and nerve
Lateral superficial vein
Corpus cavernosum
Corpus spongiosum
Urethra

Fig. 125.8 Penile tourniquet schematic

125.5 Complications

- Necrosis of tissue distal to the tourniquet
- Neurovascular damage secondary to prolonged ischemia
- The incision technique can cut into the tendon insertions or neurovascular bundles of the digits or into the corpus callosum of the penis. Be certain to ascertain function after incision.
- Infection, especially after incision.

125.6 Pearls and Pitfalls

- Pearls
  - All tourniquets must come off.
  - The easiest way to get to the bottom of the tourniquet, regardless of method used, is to identify the knot or bunched area if possible.
  - When using the incision technique, keep the incision in the longitudinal plane of the appendage and perpendicular to the tourniquet.
  - Do not use depilatories on open wounds.
  - 24-h follow-up is mandatory, although if the tourniquet has been present for an extended period of time, it may actually take several days for return of color and blood flow.
  - Be certain to document neurovascular status before and after tourniquet removal.
  - Consider surgical consult in cases of severe tissue edema or distorted anatomy.
  - Tourniquets have been described in the child abuse literature. Consider that possibility in the preverbal child.
- Pitfalls
  - The depilatories will not work on nonorganic tourniquets.

Selected Reading

Bothner J. Hair entrapment removal techniques. UpToDate. Accessed 30 Nov 2010.

Cardriche, D. Hair tourniquet removal. Medscape reference: drugs, diseases, and procedures. Available at: http://emedicine.medscape.com/article/1348969-overview. Accessed 17 May 2009.

Klusmann A. Tourniquet syndrome—accident or abuse? Eur J Pediatr. 2004;163:495–8.

Lundquist ST, Stack LB. Genitourinary emergencies: diseases of the foreskin, penis, and urethra. Emerg Med Clin North Am. 2001;19:529–46.

Peleg D, Steiner A. The Gomco circumcision: common problems and solutions. Am Fam Physician. 1998;58:891–8.

Use of the Broselow Tape

126

Judith K. Lucas

126.1 Indications

- To determine the equipment size and the medication doses during a pediatric resuscitation, without having to take the time to perform calculations

126.2 Contraindications

- Premature infant or newborn whose heel, while the infant is fully extended, does not fall at least into the white area (corresponding to 3, 4, and 5 kg).
- The length of the child exceeds the distal end of the green area (36 kg).

126.3 Materials

- Broselow Pediatric Emergency Tape (Armstrong Medical Industries, Wilshire, IL) (Figs. 126.1 and 126.2)

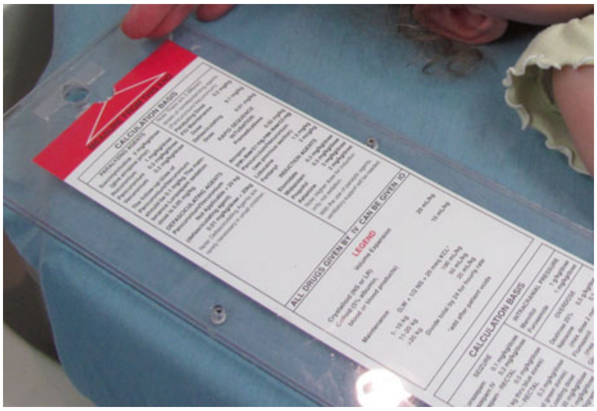

Fig. 126.1 Broselow tape, proximal end (at patient's head)

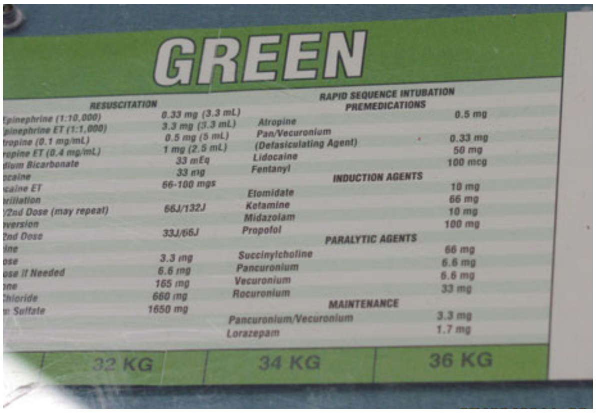

GREEN

RESUSCITATION	
Epinephrine (1:10,000)	0.33 mg (3.3 mL)
pinephrine ET (1:1,000)	3.3 mg (3.3 mL)
tropine (0.1 mg/mL)	0.5 mg (5 mL)
ropine ET (0.4 mg/mL)	1 mg (2.5 mL)
dium Bicarbonate	33 mEq
ocaine	33 mg
caine ET	66-100 mgs
rillation	
/2nd Dose (may repeat)	66J/132J
version	
2nd Dose	33J/66J
ine	
ose	3.3 mg
ose if Needed	6.6 mg
one	165 mg
Chloride	660 mg
Sulfate	1650 mg

RAPID SEQUENCE INTUBATION	
PREMEDICATIONS	
Atropine	0.5 mg
Pan/Vecuronium (Defasiculating Agent)	0.33 mg
Lidocaine	50 mg
Fentanyl	100 mcg
INDUCTION AGENTS	
Etomidate	10 mg
Ketamine	66 mg
Midazolam	10 mg
Propofol	100 mg
PARALYTIC AGENTS	
Succinylcholine	66 mg
Pancuronium	6.6 mg
Vecuronium	6.6 mg
Rocuronium	33 mg
MAINTENANCE	
Pancuronium/Vecuronium	3.3 mg
Lorazepam	1.7 mg

32 KG | 34 KG | 36 KG

Fig. 126.2 Broselow tape, distal end (the tape will accommodate a 36-kg child, at most)

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_126

126.4 Procedure (Figs. 126.3, 126.4, and 126.5)

- Place the infant or child on the bed in the supine position.
- Extend the Broselow Pediatric Emergency Tape next to the patient, placing the red line at the top of the patient's head.
- With the patient fully extended, especially through the hips and knees, align the bottom of the heel, while the ankle is flexed and toes upward, with the color on the tape adjacent to the heel.
- The color on the tape corresponds to the patient's weight.
- The side of the tape with weights noted along the bottom has medication doses for resuscitation, rapid-sequence intubation (RSI), and paralytics (Fig. 126.6).
- The top half of the other side of the tape has the appropriate dosing for managing seizures, overdoses, elevating intracranial pressure, and fluids. The bottom half of this side has the appropriate equipment sizes, such as endotracheal tubes, nasogastric tubes, chest tubes, and other equipment, that correspond to the patient's weight (Fig. 126.7).

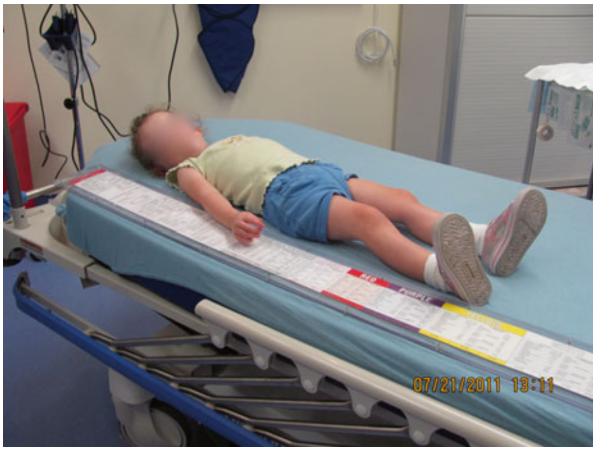

Fig. 126.3 Placement of the child on the tape

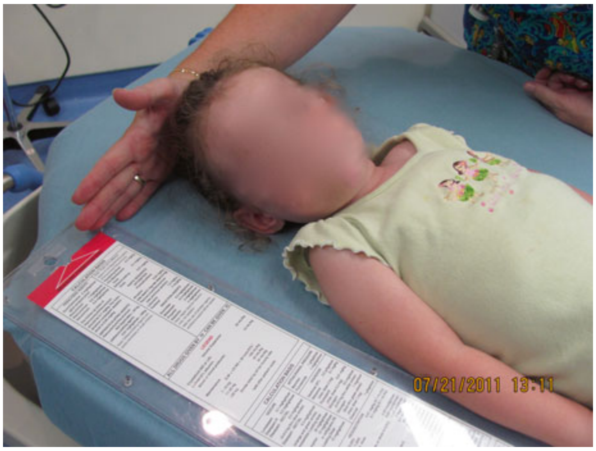

Fig. 126.4 Placement of the child on the tape

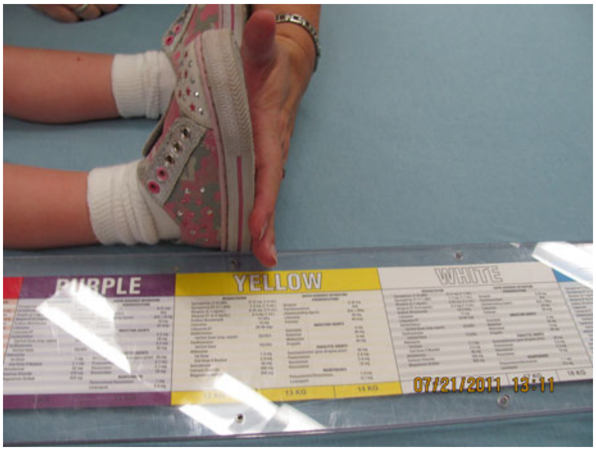

Fig. 126.5 Placement of the child on the tape

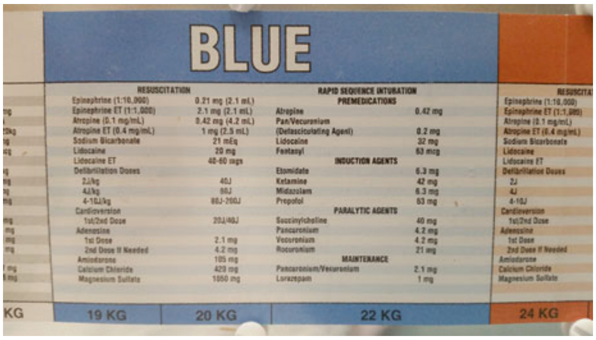

BLUE

RESUSCITATION	
Epinephrine (1:10,000)	0.21 mg (2.1 mL)
Epinephrine ET (1:1,000)	2.1 mg (2.1 mL)
Atropine (0.1 mg/mL)	0.42 mg (4.2 mL)
Atropine ET (0.4 mg/mL)	1 mg (2.5 mL)
Sodium Bicarbonate	21 mEq
Lidocaine	20 mg
Lidocaine ET	40-60 mgs
Defibrillation Doses	
2J/kg	40J
4J/kg	80J
4-10J/kg	80J-200J
Cardioversion	
1st/2nd Dose	20J/40J
Adenosine	
1st Dose	2.1 mg
2nd Dose If Needed	4.2 mg
Amiodarone	105 mg
Calcium Chloride	420 mg
Magnesium Sulfate	1050 mg

RAPID SEQUENCE INTUBATION	
PREMEDICATIONS	
Atropine	0.42 mg
Pan/Vecuronium (Defasciculating Agent)	0.2 mg
Lidocaine	32 mg
Fentanyl	63 mcg
INDUCTION AGENTS	
Etomidate	6.3 mg
Ketamine	42 mg
Midazolam	6.3 mg
Propofol	63 mg
PARALYTIC AGENTS	
Succinylcholine	40 mg
Pancuronium	4.2 mg
Vecuronium	4.2 mg
Rocuronium	21 mg
MAINTENANCE	
Pancuronium/Vecuronium	2.1 mg
Lorazepam	1 mg

KG | 19 KG | 20 KG | 22 KG | 24 KG

Fig. 126.6 Front of tape. Weight groupings are noted at the bottom. This is the side with resuscitation and RSI medications

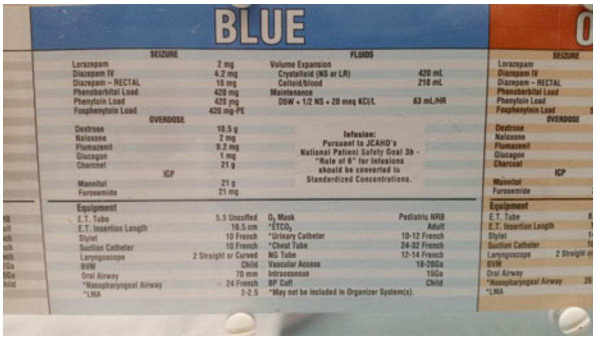

BLUE

SEIZURE	
Lorazepam	2 mg
Diazepam IV	4.2 mg
Diazepam – RECTAL	10 mg
Phenobarbital Load	420 mg
Phenytoin Load	420 mg
Fosphenytoin Load	420 mg-PE
OVERDOSE	
Dextrose	10.5 g
Naloxone	2 mg
Flumazenil	0.2 mg
Glucagon	1 mg
Charcoal	21 g
ICP	
Mannitol	21 g
Furosemide	21 mg

FLUIDS	
Volume Expansion	
Crystalloid (NS or LR)	420 mL
Colloid/blood	210 mL
Maintenance	
D5W + 1/2 NS + 20 meq KCl/L	63 mL/HR

Infusion:
Pursuant to JCAHO's
National Patient Safety Goal 3b -
"Rule of 6" for Infusions
should be converted to
Standardized Concentrations.

Equipment			
E.T. Tube	5.5 Uncuffed	O_2 Mask	Pediatric NRB
E.T. Insertion Length	16.5 cm	*$ETCO_2$	Adult
Stylet	10 French	*Urinary Catheter	10-12 French
Suction Catheter	10 French	*Chest Tube	24-32 French
Laryngoscope	2 Straight or Curved	NG Tube	12-14 French
BVM	Child	Vascular Access	18-20Ga
Oral Airway	70 mm	Intraosseous	15Ga
*Nasopharyngeal Airway	24 French	BP Cuff	Child
*LMA	2-2.5		

*May not be included in Organizer System(s).

Fig. 126.7 Reverse side of tape. This side has medications for critical interventions aside from the primary survey, as well as equipment sizes

126.5 Pearls and Pitfalls

- Pearls
  - Be certain that the top of the head is at the red line and that the patient's body is fully extended. Infants tend to lay flexed at the hips and knees.
  - Determining the appropriate color code depends on noting the bottom of the heel while the ankle is flexed and the toes are pointing upward.
- Pitfalls
  - Beware of the child who is unusually heavy compared with other children of the same height.

Selected Reading

Luten R. Error and time delay in pediatric trauma resuscitation: addressing the problem with color-coded resuscitation aids. Surg Clin North Am. 2002;82:303–14, vi.

Luten R, Broselow J. Standardization of product concentration in emergency dosing: a response to Fineberg and Arendt. Ann Emerg Med. 2008;52:477–8.

Rosenberg M, Greenberger S, Rawal A, Latimer-Pierson J, Thundiyil J. Comparison of Broselow tape measurements versus estimations of pediatric weights. Am J Emerg Med. 2011;29:482–8.

Nursemaid's Elbow

127

Judith K. Lucas

127.1 Definition

- An injury commonly seen in children between 6 months and late preteens, although generally seen between 1 and 3 years
- A subluxation of the radial head, usually resulting from sudden, longitudinal traction on an extended arm with the wrist pronated
- Often occurs when a parent/caregiver is holding the child by the hand while walking and suddenly pulls the child away from a dangerous situation (Fig. 127.1)

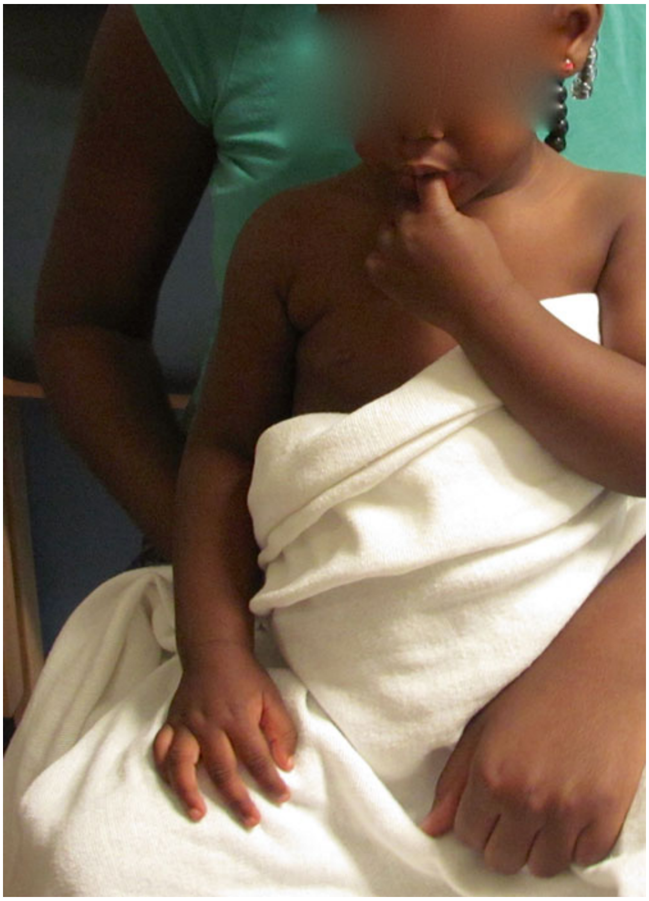

Fig. 127.1 Note that the toddler prefers to hold her right arm pronated and somewhat flexed at the elbow. When asked where it hurts, she may as often point to her wrist as to her elbow. There will be no soft tissue swelling anywhere along the upper extremity

---

J.K. Lucas, MD
Department of Emergency Medicine,
University of Florida Health Shands Hospital,
Gainesville, FL, USA
e-mail: judithklucas@ufl.edu

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0_127

127.2 Clinical Diagnosis

- History of pulling-type injury.
- Child presents with arm held slightly flexed at elbow and pronated.
- Patient may point to elbow or wrist as source of pain, but both areas are without any swelling and are nontender on palpation.
- Elbow can be flexed and extended, but the forearm cannot be supinated.
- Radiography is not helpful.

127.3 Indications

- Clinical presence of subluxed radial head

127.4 Contraindications

- Absolute
  - Radiographic evidence of elbow or forearm fracture
  - Swelling or pain about the elbow, forearm, or wrist
- Relative
  - Unknown history or witnessing of pulling-type injury

127.5 Materials and Medications

- None

127.6 Procedure

127.6.1 Method A: Superpronation (preferred method) (Fig. 127.2)

1. Explain to the caretaker that the reduction will cause the child *very brief* discomfort.
2. Seat the child on the lap of a parent/caregiver or an assistant, facing the operator and holding the child in such a way as to hold the child's humerus against her or his side.
3. The person performing the procedure holds the elbow at approximately 90° and grasps the elbow, with the physician's thumb over the region of the radial head (this is done in order to be able to palpate the reduction "clunk" or "click").
4. The physician then holds the patient's wrist firmly and rapidly hyperpronates the forearm. A palpable or audible "click" signifies successful reduction but may not be appreciable.
5. The child may cry for a few minutes after reduction; however, the provider should leave the bedside for 5–10 min and instruct the parents to allow the child to simply play, without focusing on the affected arm.
6. After approximately 10 min, return to the bedside and reevaluate the child, who should have full use of her or his arm by then.

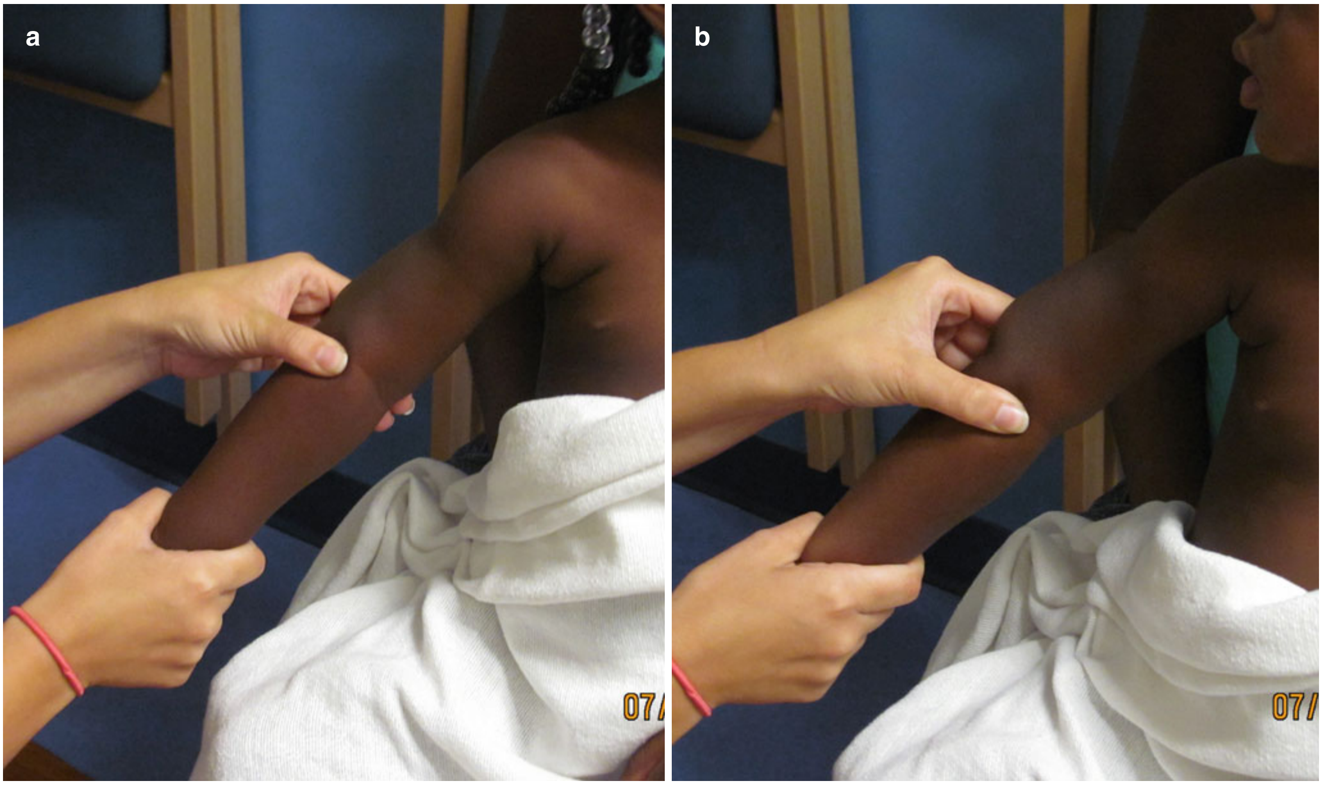

Fig. 127.2 Pronation: (**a**) The examiner's left hand holds the patient's radial head and medial condyle to better appreciate palpable reduction of the subluxation, (**b**) The examiner's right hand gently, but firmly, hyperpronates the patient's forearm from the level of the wrist. Within this maneuver, the examiner should feel or sense a "pop"

127.6.2 Method B: Supination (Fig. 127.3)

1. Follow the first three steps in Method A.
2. While holding the patient's wrist firmly, the operator steadily supinates the patient's forearm completely, followed by flexing the elbow, bringing the patient's wrist up to the shoulder.
3. Return to the fifth and sixth steps in Method A.

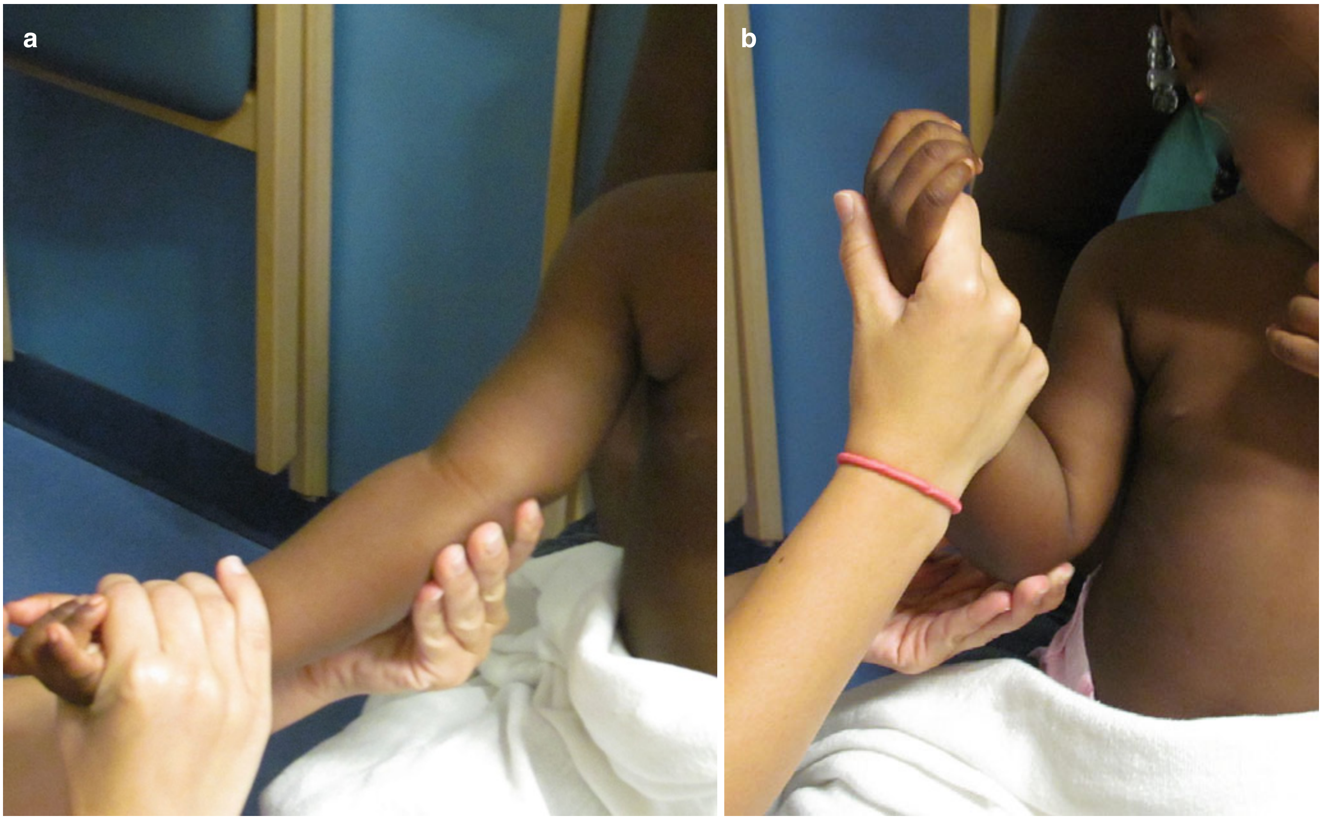

Fig. 127.3 Supination: (**a**) With the examiner's left hand in the same position as for the pronation technique, (**b**) the examiner supinates the forearm, from the level of the wrist, and fully flexes at the elbow. In either of these two maneuvers, the examiner should feel or sense a "pop"

127.7 Evaluation After Procedure

- If the child has complete use of the arm, no further intervention is required.
- If the child is still unable to supinate the forearm after 20–30 min, consider a repeat attempt at reduction.
- If there is still no return to full function after 30 min and/or repeated attempts at reduction, X-rays should be considered.
- In a child who still refuses to use the arm and X-rays are negative, the child should be reevaluated in 24 h.

127.8 Pearls and Pitfalls

- Pearls
  - Once reduced, there is rarely the need for analgesic medications. However, should the need arise, stay with ibuprofen because it is a proven anti-inflammatory.
  - If not seeing improvement within 15 min of reduction, consider this to be a fracture and obtain appropriate views of the elbow and forearm.
- Pitfalls
  - The longer the radial head has been subluxed, the longer it will take the child to return to full use.

Selected Reading

Krul M, van der Wouden JC, Koes BW, Schellevis FG, van Suijlekom-Smit LW. Nursemaid's elbow: its diagnostic clues and preferred means of reduction. J Fam Pract. 2010;59:e5–7.

Krul M, van der Wouden JC, van Suijledom-Smit LW, Koes BW. Manipulative interventions for reducing pulled elbow in young children. Cochrane Database Syst Rev. 2009;4:CD007759. doi: 10.1002/14651858.CD007759.pub2.

Macias CG, Bothner J, Wiebe R. A comparison of supination/flexion to hyperpronation in the reduction of radial head subluxations. Pediatrics. 1998;102:e10–8.

Quan L, Marcuse EK. The epidemiology and treatment of radial head subluxation. Am J Dis Child. 1985;139:1194–7.

Switzer JA, Ellis T, Swiontkowski MF. Wilderness orthopaedics. In: Auerbach PS, editor. Wilderness medicine. 5th ed. Philadelphia: Elsevier; 2007. p. 573.

Index

A
Abdominal hernias
    complications, 404
    contraindications, 403
    definition, 401
    indications, 403
    locations, 401, 402
    materials and medications, 403
    procedure, 403
    types of, 401
ABFO scale. *See* American Board of Forensic Odontology (ABFO) scale
Abortion
    complications, 659
    contraindications, 657
    indications, 657
    materials and medications, 657
    pearls and pitfalls, 659
    procedure
        checking abnormalities, 657–658
        decidual reaction, 657, 658
        fetal cardiac activity, 657, 658
        transvaginal sagittal imaging, 658
        transvaginal transverse imaging, 658
Abscess
    anal fissure, 455
    brain, 239
    difficult cricothyrotomy, 126
    incision and drainage
        Bartholin gland (*see* Bartholin gland abscess/cysts)
        complications, 567
        contraindications, 563
        external ear block, 520
        indications, 563
        local anesthesia, 509
        materials and medications, 563, 564
        parotid duct (*see* Parotid duct abscess)
        pearls, 567
        peritonsillar (*see* Peritonsillar abscess)
        procedure, 565–566
        sublingual (*see* Sublingual abscess)
    retroperitoneal, 717
    septal, 324
Acromioclavicular (AC) joint arthrocentesis, 621, 622
Activated charcoal (AC) adsorption
    complications, 423
    contraindications, 422
    indications
        MDAC therapy, 421
        single-dose activated charcoal, 421
    materials and medications, 422
    pearls, 423
    pitfalls, 423
    procedure, 423
Acute anal fissures, 453, 455
Acute dental fracture, 379
Acute respiratory distress syndrome (ARDS), 171, 172
Adson forceps, 241, 498, 500
Advance and cut technique, 550, 552
    with multiple barbs, 550, 552
    pearls and pitfalls, 553
    with single barb, 550
Airway assessment. *See* Difficult airway
Airway management. *See* Cricothyroidotomy
Allen test, 5, 7
Alligator forceps
    cerumen removal, 331, 332
    foreign body removal
        ear, 335, 336, 338
        nasal, 326–328
Allis maneuver, 605–606
Alveolar osteitis. *See* Dry socket
Ambu bags, 57, 326, 327, 329
Ambu Perfit, 219, 221, 230
American Academy of Neurology, 267
American Board of Forensic Odontology (ABFO) scale, 478, 482
Anal fissures, 454
    complications, 455
    contraindications, 455
    definition, 453
    indications, 453
    materials and medications, 455
    pearls and pitfalls, 455
    procedure, 455
Anembryonic gestation, 655
    complications, 656
    contraindications, 656
    indications
        low-lying empty gestational sac, 655
        mean gestational diameter, 655
    materials and medications, 656
    pearls and pitfalls, 656
    procedure
        transabdominal, 656
        transvaginal, 656
Anesthesia
    local (*see* Local anesthesia)
    mandibular nerve (*see* Mandibular anesthesia)
    over femoral artery, 9, 90
    regional (*see* Regional nerve block)
    for splinting, 572
Ankle arthrocentesis, 626, 629

© Springer Science+Business Media New York 2016

L. Ganti (ed.), *Atlas of Emergency Medicine Procedures*, DOI 10.1007/978-1-4939-2507-0

Ankle dislocation reduction
  complications, 620
  contraindications, 617
  indications, 617
  materials and medications, 617
  pearls and pitfalls, 620
  procedure
    anterior dislocations, 619
    posterior dislocations, 618
Antepartum hemorrhage, 661
Anterior dislocation
  ankle, 619
  elbow, 591, 594
  femur, 609, 610
  fibular head, 612
  shoulder, 583, 584
Anterior epistaxis
  complications, 314
  initial preparation, 313
  management of, 313–314
    commercial nasal pack, 313, 315
    gauze method, 314, 316
    silver nitrate stick and packaging, 313, 315
  pearls and pitfalls, 314
Aortic occlusion, 165, 168
Apnea test, 262, 266
ARDS. *See* Acute respiratory distress syndrome (ARDS)
Arrhythmia, 19, 33, 185, 188, 318
Arterial cannulation
  complications, 10
  contraindications, 3
  critical care, 3
  femoral artery (*see* Femoral artery cannulation)
  indications, 3
  pearls and pitfalls, 10–11
  radial artery (*see* Radial artery cannulation)
Arthritis diagnosis. *See* Arthrocentesis
Arthrocentesis. *See also* Intra-articular injection
  acromioclavicular joint, 621, 622
  ankle, 626, 629
  complications, 623
  contraindications, 621
  fluid aspiration, 625
  glenohumeral joint, 621, 622
  indications, 621
  knee
    lateral approach, 626, 627
    medial approach, 626, 628
    procedure, 621, 622
  materials and medications, 621
  pearls and pitfalls, 623
  procedure, 621–622
  shoulder, 626, 629
Ascites, 445, 449, 451
Aspen collar, 219–222
Atrial fibrillation, 185, 186
Atrial flutter, 185, 186
Atrial injuries, 181
Auricular hematoma
  cauliflower ear, 341
  contraindications, 342
  incision procedure, 342–348
    compression dressing, materials needed for, 343, 345
    elastic bandage, wrap head with, 343, 347
    Kerlix, wrap head with, 343, 346
    mold sterile gauze, 343, 346
    mold Xeroform gauze, 343, 345
    pad outside ear, 343, 346
    place dry cotton, 343
    secure dental rolls, 343, 348
  materials and medications, 342
  needle aspiration, 342–343
  pitfalls and complications, 348
  schematic of, 342
Auricular ring block technique, 520
Auriculotemporal block, 373
Auscultation monitoring, 641
Awake orotracheal intubation
  contraindications, 53
  indications, 53
  materials and medications, 53
  pearls and pitfalls, 56
  procedure, 53–55
    deep structures and trachea, lidocaine atomization of, 54, 55
    lidocaine lollipop, 54
    lighted stylet, 54, 55
    mucosal atomizer, 54
    nebulization of 4 % lidocaine, 53, 54
    posterior pharynx, lidocaine atomization of, 54, 55
    preparation, 53
    sedation, 54, 55

B
Babcock forceps, 498, 500, 501
Babysitter's elbow. *See* Nursemaid's elbow
Bag-valve-mask (BVM) ventilation, 127, 133, 135, 138
  airway, 89
  complications, 50
  contraindications, 47
  indications, 47
  materials, 47
  pearls and pitfalls, 50
  procedure, 48–50
    nasal airway insertion, 48, 49
    one-handed seal technique, 48, 49
    oral airway insertion, 48
    two-handed seal technique, 48, 50
Balloon gastrostomy tube, 441, 443
Balloon tamponade, 439
Bartholin gland abscess/cysts
  contraindications, 471
  follow-up care, 476
  indications, 471
  materials, 472
  procedure, 471
    incision and drainage, 472–473
    iodoform packing, 474
    Jacobi ring catheter, 475
    silver nitrate sticks, 475
    Word catheter, 474
Basilic vein (BV)
  cutdown, 41
  PIV cannulation, scanning for, 14
Benign paroxysmal positional vertigo (BPPV)
  Dix-Hallpike maneuver, 257
  Epley maneuver (*see* Epley maneuver for vertigo)
Bezoars, 421, 425
Bimanual pelvic examination, 467, 468, 478
Bladder catheterization
  complications, 463
  contraindications, 459
  indications, 459
  materials and medications, 459

  pearls and pitfalls, 463
  procedure, 460–462
    for females, 460–462
    frog-leg position, 460
    for males, 460, 461
Bleeding evaluation, gastrointestinal. *See* Nasogastric tube (NT) placement
Blighted ovum. *See* Anembryonic gestation
Blister care, 534
Blood gas analysis, 494
Blunt probe and cutting method, 724, 726
Body surface area (BSA), 534
Bonfils intubation fiberscope, 109, 110
Bougie. *See* Endotracheal tube (ETT) introducer
Bowel obstruction/perforation. *See* Whole-bowel irrigation (WBI)
BPEG. *See* British Pacing and Electrophysiology Group (BPEG)
BPPV. *See* Benign paroxysmal positional vertigo (BPPV)
Brachial artery (BrA), 14
Brachial vein (BrV), 14
Bradycardia
  transcutaneous pacing, 195, 196
  transvenous cardiac pacing, 197
Brain death examination in adults
  confirmatory testing, 267
  contraindications, 261
  indications, 261
  materials and medications, 261
  pearls and pitfalls, 267
  procedure
    apnea test, 262, 266
    bulbar paralysis, 262, 266
    motor response, pain, 261–263
    oculocephalic reflex, 261, 264
    oculovestibular reflex, 261, 265
  repeat examination, 267
Breech delivery
  breech presentation, 677
  complications, 682
  contraindications, 678
  indications, 678
  materials and medications, 678
  pearls and pitfalls, 682
  procedure
    Burns-Marshall maneuver, 680, 681
    correct hand placement, 680
    head delivery in flexed position, 680
    McRoberts maneuver, 678, 679
    placing thumbs on sacrum, 678, 679
    rotating towards maternal symphysis pubis, 678, 679
British Pacing and Electrophysiology Group (BPEG), 211, 212
Broselow Pediatric Emergency Tape
  contraindications, 729
  indications, 729
  materials, 729
  pearls and pitfalls, 731
  procedure, 730–731
Buccal nerve block, 367–368
Bulbar paralysis, 262, 266
Burn care
  admission criteria, 534–535
  blister care, 534
  complications, 534
  first degree/superficial burn, 531
  fluid resuscitation, 534
  inhalation injuries, 532
  materials and medications, 532
  pearls and pitfalls, 534
  permanent scarring, 531
  procedure, 534
  second degree/partial thickness burn, 531, 532
  sensory deficits, 531
  TBSA, 532–533
  third degree/full thickness burn, 531, 532
Burns-Marshall maneuver, 680, 681
Burr hole craniotomy
  complications, 240
  contraindications, 235
  indications, 235
  materials and medications, 235–236
  pearls and pitfalls, 240
  procedure
    anatomy, 236, 238
    automatic stop craniotomy drill, 236, 239
    landmark for, 236, 239
    patient positioning, 236, 237
    penetrator drill, 236, 239
    Penrose drain, suture in, 237, 239
    preparation, 236
    self-retaining retractors, placement of, 236, 239
    wound closure, 237, 239
BVM ventilation. *See* Bag-valve-mask (BVM) ventilation

C
Caloric reflex testing
  complications, 253
  contraindications, 253
  indications, 253
  materials and medications, 253
  pearls and pitfalls, 255
  procedure, 253
Canadian C-spine rules (CCR), 223
Canalith repositioning procedure. *See* Epley maneuver for vertigo
Candida albicans, 468, 469
Cannulation
  PIV, 14
  of radial and femoral arteries (*see* Arterial cannulation)
  ultrasound guided, 20, 26
Captain Morgan technique, 607
Cardiac arrest, 193, 194, 205
Cardiac injury
  complications, 183
  contraindications, 179
  indications, 179
  materials and medications, 179–180
  pearls and pitfalls, 183
  procedure, 181–182
    atrial injuries, 181
    direct manual pressure, 181, 182
    Finochietto retractor positioning, 181
    lacerations, 181, 182
    patient positioning, 181
    polypropylene suture, 181, 182
    surgical skin stapler, 181, 182
Cardiac massage. *See* Resuscitative thoracotomy
Cardiac output monitoring
  definitions, 37
  Edwards Vigileo System, 35, 36
    contraindications, 37
    materials, 38
    pearls and pitfalls, 38
    procedure, 38
Cardiac pacing, transvenous. *See* Transvenous cardiac pacing
Cardiac tamponade, 159, 179, 203, 207

Cardioversion
  synchronized electrical (*see* Synchronized electrical cardioversion)
  unsynchronized cardioversion
    complications, 192
    contraindications, 189
    high-energy shock, 189
    indications, 189
    materials and medications, 189
    pearls and pitfalls, 192
    procedure, 190–191
Catheterization
  bladder
    complications, 463
    contraindications, 459
    indications, 459
    materials and medications, 459
    pearls and pitfalls, 463
    procedure, 460–462
  peripheral vein
    contraindications, 693
    indications, 693
    materials and medications, 693, 694
    pearls and pitfalls, 699
    procedure, 694–699
  umbilical vein (*see* Umbilical venous catheters (UVC)
Cauliflower ear, 341
Cavernosal aspiration procedure, 490–492
CCR. *See* Canadian C-spine rules (CCR)
Centers for Disease Control and Prevention (CDC) guidelines, 479, 482
Central venous line
  contraindications, 17
  femoral vein access
    complications, 25–26
    pearls and pitfalls, 25
    procedure, 19, 21, 24–25
  indications, 17
  internal jugular vein access (*see* Internal jugular central venous line)
  materials and medications, 17–18
  removal of, 26
  subclavian vein access
    complications, 25–26
    pearls and pitfalls, 24
    procedure, 22–24
  ultrasound guided cannulation, 24, 26
Cephalic vein, 41, 695
Cerumen removal
  contraindications, 331
  indications, 331
  materials and medications
    alligator forceps, 331, 332
    bulb syringe, 331, 333
    cerumen-softening solutions, 331
    cerumen spoons, 331, 332
    ear curettes, 331, 333
    ear speculum, 331, 332
    eye and biohazard protection, 331
    headlamp/direct lighting instrument, 331
    irrigation setup, 331
    Schuknecht suction tube, 331
    specialized ear syringe, 331, 333
  pitfalls and complications, 334
  procedure, 334
Cerumen spoons
  cerumen removal, 331, 332
  ear foreign body removal, 335, 336
Cervical collar (C-collar)
  Aspen collar, 219–222
  cervical spine clearance (*see* Cervical spine clearance)
  contraindications, 219
  features of, 219
  indications, 219
  log roll, 229
  materials and medications, 219
  one-piece C-collars, 219
  pearls and pitfalls, 222
  two-piece C-collars, 219
Cervical ectopic pregnancy, 648
Cervical spine clearance
  CCR, 223
  complications, 227
  computed tomography, 224
  contraindications, 223
  flexion-extension X-rays, 224
  indications, 223
  magnetic resonance imaging, 224
  NEXUS clinical criteria, 223
  pearls and pitfalls, 227
  plain X-rays, 224
  procedure, 355–358
    C-collar placement, 225
    C spine tenderness assessment, 225, 226
Chemical burns, eye, 279
Chest trauma, 159
Chest tube thoracostomy. *See* Tube thoracostomy
Choriocarcinoma, 653
Chronic anal fissures, 453, 455
Clue cell, 458
Cobalt blue, 273
Color-coded pediatric emergency tape. *See* Broselow Pediatric Emergency Tape
Color duplex ultrasonography, 495
Colposcope, 478, 482
Combitube
  contraindications, 85
  indications, 85
  materials, 85
  pearls and pitfalls, 87
  procedure, 86–87
Complete abortion, 657–659
Complete breech, 677
Computed tomography (CT)
  brain death examination, 267
  choriocarcinoma, 653
  C-spine, 224
  EVD, 243, 245
  herniation syndromes, 250
  intraocular foreign body, 285
Cook intraosseous needle, 707, 708
Corneal foreign body removal
  complications, 285
  contraindications, 283
  indications, 283
  materials and medications, 284
  procedure, 285
Cotton tip applicator, 281, 284, 285
Cricothyroidotomy
  complications, 126
  contraindications, 119
  indications, 119
  pearls, 126
  scalpel-bougie techniques
    materials and medications, 119–120
    procedure, 121–122

    Scalpel-Trousseau technique
        materials and medications, 123
        procedure, 124–125
Cricothyrotomy. *See* Cricothyroidotomy
Cuffalator, 433, 434
Curved array transducer (C60)
    FAST examination
        with larger footprint, 406
        left upper quadrant view of, 409
        right upper quadrant view of, 408
    heart, subxiphoid four-chamber view of, 407
Curved hook
    cerumen removal, 331
    foreign body removal
        ear, 335, 336, 338
        nasal, 326–328
Cyanoacrylate, 335, 337, 339

D
Defibrillation. *See* Unsynchronized cardioversion
Delivery
    breech birth (*see* Breech delivery)
    perimortem cesarean (*see* Perimortem cesarean section)
    shoulder dystocia (*see* Shoulder dystocia management)
    vaginal, 665–668
Dental avulsion
    complications, 386
    contraindications, 384
    indications, 383–384
    materials and medications, 384
    pearls and pitfalls, 386
    procedure, 385
    prognosis, 386
    root development, stages of, 386
Dental fracture. *See* Fractured tooth
Depilatory method, hair tourniquet, 725
Dermabond, 546–547
Diagnostic peritoneal lavage (DPL)
    complications, 399
    contraindications, 396
    indications, 395
    materials and medications, 396
    pearls and pitfalls, 399
    procedure
        closed technique, 397
        open technique, 397
        patient preparation, 396
        semiopen (Seldinger) technique, 397–398
Difficult airway
    awake orotracheal intubation, 53
    indications, 89
    materials and medications, 89
    pearls and pitfalls, 92
    procedure
        6-D method, 90
        LEMON, predict difficult direct laryngoscopy, 89–90
        MOANS, BVM difficulty, 89
        obese patient, optimizing laryngoscopy in, 92
Digital nerve blocks, 523
    intrathecal/tendon sheath approach, 525
    ring block, 524
    web space approach, 524
DIP joint reduction. *See* Distal interphalangeal (DIP) joint reduction
Direct laryngoscopy, 118
    complications, 74
    contraindications, 67
    indications, 67
    materials and medications, 68
    pearls and pitfalls, 74
    procedure, 69–74
        ETT, 69, 74
        glottic view, 69, 72–73
        Macintosh blade insertion, 69, 71
        Miller blade insertion, 69, 71
        patient positioning, 69, 70
Dislocation
    ankle
        anterior dislocations, 619
        complications, 620
        contraindications, 617
        indications, 617
        materials and medications, 617
        pearls and pitfalls, 620
        posterior dislocations, 618
    elbow
        anterior dislocations, 591, 594
        complications, 597
        contraindications, 591
        indications, 591
        materials and medication, 592
        pearls and pitfalls, 597
        posterior dislocations (*see* Posterior dislocation, elbow)
        radial head subluxations, 595–596
    finger joint (*see* Distal interphalangeal (DIP) joint reduction)
    hip (*see* Hip dislocation reduction)
    knee (*see* Knee dislocation reduction)
    shoulder (*see* Shoulder, dislocation reduction)
    temporomandibular joint (*see* Temporomandibular joint (TMJ) dislocation)
Distal interphalangeal (DIP) joint reduction, 599
    complications, 601
    contraindications, 599
    indications, 599
    materials and medications, 599
    pearls and pitfalls, 601
    procedure, 600
Dix-Hallpike maneuver
    complications, 258
    contraindications, 257
    indications, 257
    materials, 257
    pearls and pitfalls, 258
    procedure, 257, 258
Doll's eyes
    brain death examination in adults, 261, 264
    contraindications, 251
    indications, 251
    pearls and pitfalls, 255
    procedure, 251, 252
Doppler ultrasonography
    fetal heart rate monitoring, 642
    testicular torsion, 503, 504
Dorsal penile nerve block, 487–489
Double-lumen airway. *See* Combitube
DPL. *See* Diagnostic peritoneal lavage (DPL)
Dry socket
    complications, 376
    contraindications, 376
    indications, 375
    materials and medications, 376
    pearls and pitfalls, 376
    procedure, 376
Durable mass, 425

E

Ear canal impaction. *See* Cerumen removal
Ear curettes, 331, 333
Ear foreign body removal
  alligator forceps, 335, 336, 338
  cerumen spoon, 335, 336
  contraindications, 335
  curved hook, 335, 336, 338
  cyanoacrylate, 335, 337, 339
  ear speculum, 335, 336
  ear syringe, 335, 337
  eye and biohazard protection, 335
  headlamp/direct lighting instrument, 335
  indications, 335
  irrigation setup, 335, 337, 339
  Schuknecht catheter, 335, 337, 338
Ear hematoma. *See* Auricular hematoma
Early pregnancy failure. *See* Anembryonic gestation
Ear speculum
  cerumen removal, 331, 332
  foreign body removal, 335, 336
Ear syringe, 335, 337
Ear wax removal. *See* Cerumen removal
Ectopic pregnancy, 645
  complications, 648
  contraindications, 648
  indications
    in first trimester, 646
    free fluid in pelvis, 646, 647
    gestational sac, 646, 647
    hCG level, 646
    pseudogestational sac, 646
    risk factors, 646
    yolk sac, 648
  materials and medications, 648
  pearls and pitfalls, 648–649
  procedure
    transabdominal, 648
    transvaginal, 648
Edwards Vigileo System, 35, 36
  contraindications, 37
  materials, 38
  pearls and pitfalls, 38
  procedure, 38
EFAST. *See* Extended focused assessment with sonography for trauma (EFAST)
Elbow dislocation reduction
  anterior dislocations, 591, 594
  complications, 597
  contraindications, 591
  indications, 591
  materials and medication, 592
  pearls and pitfalls, 597
  posterior dislocations
    anatomical depiction, 591
    method A, 592
    method B, 592, 593
    method C, 592, 593
  radial head subluxations
    subluxation hyperpronated, 595
    subluxation hyperpronated and flexed, 595
    subluxation supinated, 595, 596
Electrocardiography (ECG)
  complete left bundle branch block, 29
  fetal heart rate monitoring, 643
Electronic fetal heart rate monitoring
  vs. auscultation method, 641
  external, 642
  internal, 643
Electronic indentation tonometry. *See* Tono-Pen tonometry
Embolism. *See* Pulmonary embolism
Emergent delivery. *See* Breech delivery
Emergent thoracotomy
  aortic occlusion, 165, 168
  complications, 169
  contraindications, 163
  indications, 163
  materials and medications, 163–164
  resuscitative thoracotomy and open cardiac massage, 165–167
    Finochietto retractor, 165, 166
    Lebsche knife, 165, 166
    pericardium, 165, 167
    skin and subcutaneous tissue incision, 165, 166
Empty skull sign, 267
Encephalitis, 247
Endocardial pacing. *See* Transvenous cardiac pacing
Endotracheal tube (ETT) introducer, 116
  complications, 101
  contraindications, 99
  indications, 99
  materials and medications, 99
  pearls and pitfalls, 101
  procedure, 100–101
    airway visualization, 100
    epiglottis, 100
    placement, 100, 101
    removal, 100, 101
Endotracheal tube (ETT) intubation
  fiber-optic stylet intubation (*see* Fiber-optic stylet intubation)
  lighted stylet intubation (*see* Lighted stylet (LS) intubation)
  LMA Fastrach, 80–81
  Storz video laryngoscope (*see* Storz video laryngoscope)
Episiotomy, 670, 678, 680, 685
Epistat, 317, 318
Epistaxis control
  anterior epistaxis (*see* Anterior epistaxis)
  contraindications, 313
  indications, 313
  materials and medications, 313
  nasal septum, vascular supply to, 311, 312
  nose, lateral wall of, 311, 312
  posterior epistaxis
    management of, 317, 318
    pearls and pitfalls, 317–318
    preliminary steps, 317
Epley maneuver for vertigo
  complications, 260
  contraindications, 259
  indications, 259
  materials, 259
  pearls and pitfalls, 260
  procedure, 259–260
Erythema migrans, 556
Esophageal foreign body removal
  complications, 418
  contraindications, 417
  indications, 417
  materials
    bougienage, 417
    esophagoscopy/endoscopy, 417

        Foley catheter removal, 417
        LES, relaxation of, 417–418
    pearls, 418–419
    procedure
        bougienage, 418
        Foley catheter removal, 418
        LES, relaxation of, 418
        push technique, 418
        push with fragmentation technique, 418
Esophagoscopy, 417
EVD. *See* External ventricular drain (EVD)
Extended focused assessment with sonography for trauma (EFAST), 410–411
External ear block
    auricular ring block, 520
    indications, 520
External fetal heart rate monitoring, 642
External rotation method, 587
External thrombosed hemorrhoid. *See* Thrombosed external hemorrhoid
External ventricular drain (EVD)
    complications, 244
    contraindications, 241
    CSF, 241
    indications, 241
    materials and medications, 241, 243
    pearls and pitfalls, 244, 245
    procedure, 241–245
        advance ventriculostomy catheter with stylet, 242, 244
        anesthetize skin, 242, 244
        brain, CT scan of, 243, 245
        Kocher's point, 242, 243
        level of patient's ear tragus, 243, 245
        make burr hole, 242, 244
        make stab incision, 242, 244
Extradural hematoma (EDH), 235
Extrauterine pregnancy. *See* Ectopic pregnancy
Eye
    corneal foreign body removal (*see* Corneal foreign body removal)
    intraocular pathology (*see* Intraocular pathology)
    irrigation
        complications, 282
        contraindications, 279
        indications, 279
        materials and medications, 279–280
        pearls and pitfalls, 282
        procedure, 281
    lateral canthotomy (*see* Lateral canthotomy)
    retinal detachment (*see* Retinal detachment)
    slit lamp examination
        complications, 276
        contraindications, 272
        corneal foreign body removal, 284
        fluorescein testing (*see* Fluorescein testing)
        indications, 271, 272
        materials, 273
        pitfalls, 276
        procedure, 274–275
    tonometry (*see* Tonometry)

F
Facial nerve block
    infraorbital nerve block, 517–518
    mental nerve block, 519
    suborbital and supratrochlear nerve block, 517
FAST examination. *See* Focused assessment with sonography for trauma (FAST) examination
Feeding tube, 459, 701–704
Femoral artery cannulation. *See also* Arterial cannulation
    materials and medications, 3
    pearls and pitfalls, 10–11
    Seldinger technique, 3, 9–10
Femoral central venous line
    complications, 25–26
    pearls and pitfalls, 25
    procedure, 20, 21, 24–25
Femoral neck fracture, 603
Femur dislocation reduction. *See also* Hip dislocation reduction
    anterior dislocation, 609, 610
    complications, 614
    lateral dislocation, 609, 610
    medial dislocation, 609, 611
    pearls and pitfalls, 614–615
    posterior dislocation, 609, 610
    rotary dislocation, 609, 611
Fetal heart rate (FHR) monitoring
    complications, 644
    contraindications, 641
    electronic method
        vs. auscultation method, 641
        external, 642
        internal, 643
    indications
        antepartum, 641
        intrapartum, 641
Fiber-optic stylet intubation
    contraindications, 109
    indications, 109
    materials and medications, 109–110
    pearls and pitfalls, 113
    procedure
        advancing scope, 111, 112
        endotracheal tube placement, 110
        epiglottis visualization, 110, 112
        lens fogging prevention, 110
        scope insertion, 110, 111
        scope removal, 111, 112
Fibrinolytic osteitis. *See* Dry socket
Fibular head dislocation reduction
    anterior dislocation, 612
    complications, 612
    pearls and pitfalls, 614
    posterior dislocation, 612
Finger dislocation joint reduction. *See* Distal interphalangeal (DIP) joint reduction
Finochietto retractor, 163–166, 179, 181
Fishhook removal
    anatomy, 549
    complications, 553
    contraindications, 549
    materials and medications, 549
    from nonvital structures, 549
    pearls and pitfalls, 553
    postremoval wound care, 553
    procedures
        advance and cut technique (*see* Advance and cut technique)
        needle cover technique, 550, 551
        retrograde technique, 550, 551
        string and yank technique, 550, 551

- Fissures. *See* Anal fissures
- Flail chest
  - abdominal injuries, 162
  - blunt lung injury and contusion, 162
  - chest wall, surgical stabilization of, 162
  - diagnosis, 162
  - mortality, 159
  - posterior rib fractures, 159, 160
- Fluorescein testing
  - complications, 277
  - indications and uses, 276
  - pitfalls, 277
  - procedure, 276
- Focused assessment with sonography for trauma (FAST) examination, 179
  - abdominal/chest trauma, 405
  - complications, 411
  - contraindications, 405
  - *vs.* DPL, 395
  - free fluid, 405
  - hypotension, 405
  - materials and medications, 405–406
  - pearls and pitfalls, 411–412
  - pneumothorax/hemothorax, 405
  - procedure
    - EFAST examination, right and left lung view of, 410, 411
    - left upper quadrant view, 409
    - right upper quadrant view, 408
    - sagittal pelvic view of, 409, 410
    - subxiphoid four-chamber view, 407
    - transverse pelvic view of, 409
- Foley catheter, 417, 498, 688
  - foreign body removal
    - esophageal, 417, 418
    - nasal, 326, 328
  - large cardiac lacerations, control of, 181
  - in posterior nasopharynx, 317, 318
- Forearm fracture. *See* Sugar-tong splint
- Foreign body (FB) removal
  - corneal
    - complications, 285
    - contraindications, 283
    - indications, 283
    - materials and medications, 284
    - procedure, 285
  - ear (*see* Ear foreign body removal)
  - esophageal (*see* Esophageal foreign body removal)
  - nasal (*see* Nasal foreign body removal)
- Forensic examination, sexual assault. *See* Sexual Assault Forensic Examination (SAFE)
- Fracture
  - acetabulum, 603
  - basilar skull, 251, 253
  - concomitant, 597
  - cribriform plate, facial trauma with, 413
  - DIP joint dislocations, 601
  - elbow, 597
  - femoral neck, 603
  - fetal clavicle, 672
  - fetal humerus, 672
  - jaw, 376
  - long bone, 707
  - pelvic, 459
  - petrous bone, 253
  - posterior rib, flail chest, 159, 160
  - splinting (*see* Splinting)
- Fractured tooth
  - anatomy of tooth, 380
  - complications, 382
  - contraindications, 381
  - indications, 379
  - materials and medications, 381
  - methods of, 379
  - pearls, 382
  - procedure, 381
  - types of, 380
- Frank breech, 677

G

- Galveston formula, 534
- Gaskin maneuver, 672
- Gastric aspiration. *See* Nasogastric tube (NT) placement
- Gastric decompression. *See* Nasogastric tube (NT) placement
- Gastric irrigation. *See* Gastric lavage
- Gastric lavage
  - complications, 427
  - contraindications, 425
  - indications, 425
  - materials and medications, 425
  - pearls and pitfalls, 427
  - procedure, 426–427
- Gastrografin®, 447
- Gastrostomy tube placement
  - balloon G tube, 441, 443
  - complications, 448
  - contraindications, 445
  - general anesthesia, 445
  - indications, 445
  - laparoscopic G tube, 441
  - low-profile G tube
    - nonobturated, button, 441, 443
    - obturated, button, 441, 444
    - placement of, 446
  - Malecot tube, 441, 442
  - materials, 445
  - open gastrostomy tube, 441
  - pearls and pitfalls, 448
  - PEG tube, 441, 442
    - local anesthesia, 445
    - moderate sedation, 445
    - replacement procedure, 446–447
- Gelfoam®, 377, 378
- Genital and perineal lacerations, 685
- Glenohumeral joint arthrocentesis, 621, 622
- GlideScope
  - complications, 96
  - contraindications, 93
  - indications, 93
  - materials and medications, 93, 94
  - pearls and pitfalls, 96
  - procedure, 95–96
    - cable insertion, video monitor, 95
    - Cobalt system, 95
    - distal end, video cable connection, 95
    - ETT abuts, 96
- GlideScope Video Laryngoscope (GVL) system, 93, 94
- Globe rupture, 294
- Glucocorticoids, 625, 626, 630
- Guillain-Barré syndrome, 247

H

- Hair tourniquet
  - complications, 728
  - contraindications, 724
  - indications, 723
  - materials and medications, 724
  - pearls and pitfalls, 728
  - procedure
    - blunt probe and cutting method, 724, 726
    - depilatory method, 725
    - digits, incision technique for, 725, 726
    - severe penile tourniquet, incision technique for, 725, 727
    - unraveling method, 724, 725
- Half-buried horizontal mattress suture, 540, 543
- Headache, 248, 250, 493
- Head-hanging positioning maneuver. *See* Dix-Hallpike maneuver
- Heart failure
  - Edwards Vigileo System (*see* Edwards Vigileo System)
  - pulmonary artery catheter, 29
- Heart strain. *See* Pulmonary embolism
- Hematomas
  - auricular hematoma
    - cauliflower ear, 341
    - contraindications, 342
    - incision procedure, 342–348
    - materials and medications, 342
    - needle aspiration, 342, 343
    - pitfalls and complications, 348
    - schematic of, 342
  - septal (*see* Septal hematomas)
  - subungal hematoma (*see* Subungal hematoma)
- Hemoperitoneum, 153, 449
- Hemorrhage
  - antepartum, 661
  - postextraction (*see* Postextraction hemorrhage)
  - postpartum (*see* Postpartum hemorrhage (PPH))
  - retrobulbar, 305
  - subarachnoid, 241, 242
  - vitreous, 292
- Hepatitis B virus (HBV), 479
- Hernia. *See* Abdominal hernias
- High-frequency (5–8 MHZ) linear probe, 13
- High-risk pregnancy, 641
- Hip dislocation reduction
  - Allis maneuver, 605–606
  - Captain Morgan technique, 607
  - complications, 608
  - contraindications, 603
  - indications, 603
  - materials and medications, 603
  - pearls and pitfalls, 608
  - Stimson maneuver
    - femoral head manipulation, 604, 605
    - with hand, 604
    - with knee, 604
  - Whistler technique, 607
- Hippocrates method, 589
- Horizontal mattress suture, 540, 542
- Human immunodeficiency virus (HIV), 455, 479, 641
- Hunt and Hess scale, 242
- Hydatidiform mole
  - complications, 653
  - contraindications, 651
  - definition, 651
  - indications, 651
  - materials and medications, 651
  - pearls and pitfalls, 653
  - procedure
    - transabdominal imaging, 651–652
    - transvaginal imaging, 652
- Hydrocephalus, 241, 245, 250
- Hypotension
  - Edwards Vigileo System (*see* Edwards Vigileo System)
  - FAST examination, 405
  - pulmonary artery catheter, 29
- Hypovolemia, 19, 23, 24, 39, 451

I

- IAN. *See* Inferior alveolar nerve (IAN) block
- ICP. *See* Intracranial pressure (ICP)
- Impotency, 493
- Incision and drainage (I&D), abscess
  - Bartholin gland (*see* Bartholin gland abscess/cysts)
  - complications, 567
  - contraindications, 563
  - external ear block, 520
  - indications, 563
  - local anesthesia, 509
  - materials and medications, 563–564
  - parotid duct (*see* Parotid duct abscess)
  - pearls, 567
  - peritonsillar (*see* Peritonsillar abscess)
  - procedure, 565–566
  - sublingual (*see* Sublingual abscess)
- Incomplete abortion, 657–659
- Incomplete breech, 677
- Inevitable abortion, 657–659
- Infectious arthritis. *See* Arthrocentesis
- Inferior alveolar nerve (IAN) block
  - anesthetized areas, 363–364
  - anesthetized nerves, 363
  - contraindications, 364
  - indications, 364
  - materials and medications
    - aspirating syringe, 365
    - local anesthetic carpule, 365
    - mouth props and retractors, 365
    - needle, 365
  - procedure, 366
- Inferior vena cava (IVC), 208, 209
- Infraorbital nerve block
  - contraindications, 513
  - extraoral approach, 517
  - indications, 517
  - intraoral approach, 518
  - materials and medications, 514
- Inhalation injuries, 532
- Internal fetal heart rate monitoring, 643
- Internal jugular central venous line
  - complications, 25–26
  - procedure, 18–21
    - angiocath, 21, 23
    - informed consent, 18
    - J-tip straightening, 19, 21
    - long-and short-axis views, 18, 20
    - marking, 19, 21
    - patient positioning, 18

Intra-articular injection. *See also* Arthrocentesis
  complications, 630
  contraindications, 625
  indications, 625
  materials and medications, 625
  pearls, 630
  procedure, 626–630
    ankle arthrocentesis, 626, 629
    joint fluid aspiration, 626, 629
    knee arthrocentesis, 626–628
    shoulder arthrocentesis, 626, 629
    using sterile hemostat, 626, 630
Intracorporeal injection procedure, 493
Intracranial pressure (ICP), 241, 243, 247, 295
Intraocular pathology
  complications, 291
  foreign body, 294
  globe rupture, 294
  increased ICP, 295
  indications, 291
  materials, 291
  optic nerve sheath measurement, 295
  pitfalls, 291
  procedure, 291
  retinal detachment, 292–293
  vitreous detachment, 292
  vitreous hemorrhage, 292
Intraocular pressure (IOP), 289
  lateral canthotomy, 305, 306, 308
  tonometry (*see* Tonometry)
Intraosseous (IO) access
  complications, 712
  contraindications, 707
  indications, 707
  materials and medications
    Cook intraosseous needle, 707, 708
    EZ-IO, 707, 708
    Jamshidi disposable sternal/iliac aspiration needle, 707
  pearls and pitfalls, 712
  procedure
    distal femur, 709
    distal tibia, 709
    proximal humerus, 709, 711
    proximal tibia, 113, 708, 709
    sternum, 709
Intraosseous infusion. *See* Intraosseous access
Intrathecal digital block, 525
Intrauterine pregnancy, 657–659
IOP. *See* Intraocular pressure (IOP)
Ischemic priapism. *See* Priapism
Ixodes scapularis, 556

J
Jacobi ring catheter, 471, 472, 475
Jamshidi disposable sternal/iliac aspiration needle, 707
Joint aspiration. *See* Arthrocentesis
Joint effusion treatment. *See* Arthrocentesis
Joint pain treatment. *See* Arthrocentesis
Joint stabilization, splinting. *See* Splinting

K
Kiesselbach's plexus, 311, 312
Knee arthrocentesis
  lateral approach, 626, 627
  medial approach, 626, 628
  procedure, 621, 622
Knee dislocation reduction
  complications, 614
  contraindications, 609
  femur/tibia
    anterior dislocation, 609, 610
    complications, 614
    lateral dislocation, 609, 610
    medial dislocation, 609, 611
    pearls and pitfalls, 614
    posterior dislocation, 609, 610
    rotary dislocation, 609, 611
  fibular head
    anterior dislocation, 612
    complications, 614
    pearls and pitfalls, 614
    posterior dislocation, 612
  indications, 609
  lateral patellar
    complications, 614
    pearls and pitfalls, 614–615
    procedure, 613
  materials and medications, 609
Kocher technique, 587

L
Laparoscopic gastrostomy tube placement, 441, 445
Laryngeal mask airways (LMAs), 109, 119
  complications, 83
  contraindications, 75
  indications, 75
  LMA Classic Excel, 76–78
  LMA Fastrach, 76
    endotracheal intubation, 80–81
    insertion, 79
    removal of, 81–82
  LMA Unique, 76–78
  pearls and pitfalls, 83
  sizing, 77
Laryngotomy. *See* Cricothyroidotomy
Lateral canthotomy
  complications, 308
  contraindications, 306
  materials and medications, 306
  pearls and pitfalls, 308
  procedure, 306–307
    initial cut, 306, 307
    intraocular pressure, further cut to reduce, 307
    lateral canthal tendon, 306, 307
    patient positioning, 306
  retrobulbar hemorrhage, 305
Lateral patellar dislocation reduction
  complications, 614
  pearls and pitfalls, 614–615
  procedure, 613
LES. *See* Lower esophageal sphincter (LES)
Levitan FPS malleable fiber-optic stylet, 110
Lighted stylet (LS) intubation
  contraindications, 103
  indications, 103
  materials and medications, 103
  pearls and pitfalls, 107
  procedure, 104–106
    dim/blurred light, 104, 105

        ETT placement, 104, 106
        LS-ETT unit insertion, 104
        LS-ETT unit preparation, 104
        midline glow, use of, 104, 105
        patient positioning, 104
        preoxygenate, 104
Light Wand™, 103
LMAs. *See* Laryngeal mask airways (LMAs)
Local anesthesia
    Bartholin gland abscess/cysts drainage, 472
    bladder catheterization, 459
    characteristics, 511
    complications, 512
    contraindications, 509
    dorsal penile nerve block procedure, 487, 488
    dry socket, 376
    fishhook removal, 549
    indications, 509
    infiltration, 511
    intra-articular injection, 626
    lumbar puncture in adults, 248
    materials and medications, 510
    needle thoracostomy, 144
    pearls, 512
    PEG tube placement, 445
    phimosis/paraphimosis, 497
    postextraction hemorrhage, 377
    procedure, 511
    septal hematomas, 322
    TMJ dislocation reduction, 373
Log roll
    contraindications, 229
    indications, 229
    materials and medications, 229
    pearls and pitfall, 232
    procedure, 229–231
        Ambu Perfit one-piece collar height, 229, 230
        anatomical line of symmetry, 231
        backboard removed, 231
        C-spine stabilization, 229, 230
        thoracic and lumbar spine stabilization, 230
Lower esophageal sphincter (LES), 417–418, 426
Low-profile gastrostomy tube
    nonobturated, button, 441, 443
    obturated, button, 441, 444
    placement of, 446
Lumbar puncture (LP), in adults
    complications, 250
    contraindications, 247
    indications, 247
    materials and medications, 247
    pearls and pitfalls, 250
    procedure
        CSF fluid collection, 250
        landmarks, 248
        local anesthesia, 248
        needle insertion, 248–250
        patient positioning, 248, 249
        pressure measurement, 249, 250
        sterile preparation, 248
        ultrasound guidance, 248, 249
Lumbar puncture (LP), in pediatrics
    complications, 717
    contraindications, 713
    indications, 713
    materials and medications, 713–714
    pearls and pitfalls, 717–718
    procedure
        attaching pressure manometer, 715, 717
        child, positioning of, 715, 716
        CSF freeflow, checking for, 715
        iliac crest top, palpation of, 715, 717
        locating intervertebral space, 715
        spinal needle insertion, 715
Lund Browder chart, 532, 533
Lung ultrasonography
    contraindications, 171
    indications, 171
    materials, 171
    preparations, 171
    procedure, 172–175
        A lines identification, 172, 174
        assessment and clinical decision making, 173
        bat sign identification, 172, 174
        B lines identification, 172, 174
        consolidations identification, 173, 175
        diaphragm, liver and spleen identification, 173, 175
        lung pulse, 172
        lung sliding identification, 172, 174
        pleural effusions identification, 173, 175
        probe position, 172, 174
        sinusoid sign, 173, 175
Lyme disease, 556

M
Malecot tube, 441, 442
Malfunctioning pacemakers
    ECG evaluation, 214
    failure to capture, 213
    failure to pace, 213
    failure to sense, 213
    inappropriate rate, 213
    pacemaker-mediated tachycardia management, 214
    pacemaker syndrome, 213
Mallampati classification, 61, 90, 91
Mandibular anesthesia
    buccal nerve block, 367–368
    inferior alveolar nerve block (*see* Inferior alveolar nerve (IAN) block)
    mental nerve block, 368–369
Manual testicular detorsion. *See* Testicular torsion
McRoberts maneuver
    breech delivery, 678, 679
    shoulder dystocia management, 670, 671
Mean gestational diameter (MGD), 655
Meningitis, 240, 244, 247
Mental nerve block, 368–369
    contraindications, 513
    extraoral approach, 517
    indications, 513
    intraoral approach, 518
    materials and medications, 514
Methicillin-resistant *Staphylococcus aureus* (MRSA), 567
Miami J collar, 219
Milch technique, 588
Miscarriage, 657–659
Missed abortion, 657–659
Molar pregnancy, 651–653
Morgan lens, 279–281
Multiple-dose activated charcoal (MDAC) therapy, 421, 423
Myocardial infarction, 29
Myometrium contraction, 664

N
Nailbed injury. *See* Subungal hematoma
Nail trauma. *See* Subungal hematoma
Nasal foreign body removal
  alligator forceps, 326–328
  Ambu bags, 326, 327
  contraindications, 325
  curved hook, 326–328
  eye and biohazard protection, 326
  Foley catheter, 326, 328
  headlamp/direct lighting instrument, 326
  indications, 325
  lidocaine jelly, 326
  lidocaine without epinephrine, 326
  nasal positive pressure, 329
  nasal speculum, 326
  phenylephrine, 326
  schuknecht suction tube, 326, 327
Nasal trauma
  epistaxis control, 311
  septal hematomas (*see* Septal hematomas)
Nasogastric tube (NT) placement
  complications, 415
  contraindications, 413
  indications, 413
  materials and medications, 413
  pearls, 415
  procedure
    insertion, 414
    preparation, 414
    secure tube, 414
    tube placement, confirmation of, 414
Nasopharyngeal airways, 47–49
Nasostat, 317, 318
National Emergency X-Radiography Utilization Study (NEXUS), 223
Needle cover technique, 550, 551
Needle decompression thoracostomy
  complications, 147
  indications
    absolute, 143
    awake patient, 143
    injured patient, 143
    ventilated patient, 143
  materials, 143
  pearls and pitfalls, 147
  procedure
    anterior chest, exposure of, 144
    second intercostal space, needle insertion, 144, 145
    secure angiocatheter, 144, 146
    skin preparation, 144
Noninvasive cardiac monitoring. *See* Cardiac output monitoring
Nonischemic priapism. *See* Priapism
North American Society of Pacing and Electrophysiology (NASPE), 211, 212
Nosebleeds, 311
Nuchal arm. *See* Breech delivery
Nursemaid's elbow, 597
  clinical diagnosis, 734
  contraindications, 734
  definition, 733
  evaluation, 737
  indications, 734
  pearls and pitfalls, 737
  superpronation method, 734–735
  supination method, 736

O
Obstetric complications. *See* Breech delivery
Ocular pathology. *See* Intraocular pathology
Oculocephalic reflex. *See* Doll's eyes
Oculovestibular reflex, 261, 265, 267
One-handed seal technique, BVM ventilation, 48, 49
Open chest wounds
  sucking chest wound, 159–161
  thoracic trauma, 159, 161
Open gastrostomy tube placement, 441, 445
Open pneumothorax, 159, 161
Ophthalmic burr drill, 284
Oral airways, 47, 48
Orogastric (OG) tube
  activated charcoal, 422, 423
  gastric lavage, 425
Orotracheal intubation, 67. *See also* Awake orotracheal intubation
Osteoarthritis (OA), 625, 630

P
Pacemaker evaluation
  electrocardiogram, 214
  impulses, 211
  insulated wire, 211, 212
  laboratory studies, 214
  malfunctioning (*see* Malfunctioning pacemakers)
  NASPE/BPEG code, 211, 212
  pacemaker codes, 213
  pacemaker lead insertion sites, 211, 212
  patient history, 214
  physical examination, 214
  prevalence, 211
  pulse generator, 211, 212
Pacemaker-mediated tachycardia, 213, 214
Pacemaker syndrome, 213
Pacer pads, 194
Paracentesis
  complications, 451
  contraindications, 449
  indications, 449
  materials and medications, 449
  pearls and pitfalls, 451
  procedure, 449–451
Paraphimosis
  complications, 501
  contraindications, 497
  indications, 497
  manual reduction, procedure for, 498–500
    Adson/Babcock forceps, 498, 500
    circumcision, 498, 500
    patient positioning, 498
    penile block, usage of, 498, 499
  materials and medications, 497
  pearls and pitfalls, 501
Parasternal long-axis view (PSLA), 207
Parasternal short-axis view (PSSA), 207, 208
Parkland formula, 534
Parotid duct abscess
  complications, 360
  contraindications, 357
  materials and medications, 357–358
  pearls and pitfalls, 360
  procedure, 359–360
    intraoral stab incision, 355
    lighting, 355

    parotid papilla, 359, 360
    patient positioning, 359
    salivary glands, anatomy of, 359
Parotid papilla, 359, 360
Particle repositioning. *See* Epley maneuver for vertigo
Patellar dislocation reduction
  complications, 614
  pearls and pitfalls, 614–615
  procedure, 613
PEG tube. *See* Percutaneous endoscopic gastrostomy (PEG) tube
Pelvic examination and wet preparation
  complications, 468
  contraindications, 465
  indications, 465
  materials and medications, 465, 466
  pearls and pitfalls, 468
  procedure, 465–468
    bimanual examination, 467, 468
    Candida albicans, 468, 469
    female external and internal genitalia, 466, 467
    normal and clue cell, 458
    rectovaginal examination, 467
    trichomoniasis, 468, 469
    wet preparation interpretation, 466, 467
  SAFE, 478
Pelvic ultrasound, 657–658
Percutaneous endoscopic gastrostomy (PEG) tube, 441, 442
  local anesthesia, 445
  moderate sedation, 445
  replacement procedure, 446–447
Percutaneous TTJV. *See* Transtracheal jet ventilation (TTJV)
Pericardial effusion, 203, 205
Pericardiocentesis
  complications, 205
  contraindications, 203
  indications, 203
  materials and medications, 203
  pearls, 205
  pitfalls, 205
  procedure, 203
    apical approach, 204
    parasternal long-axis approach, 205
    subxiphoid approach, 204
Pericardium. *See* Pericardiocentesis
Perimortem cesarean section
  complications, 690
  contraindications, 687
  indications, 687
  materials and medications, 687
  pearls and pitfalls, 690
  procedure
    exposing rectus sheat, 688, 689
    fetus delivery, 688, 689
    placenta delivery, 688, 690
    resuscitation of infant, 688, 689
    vertical incision, 688
Peripheral intravenous (PIV) access, ultrasound-guided
  complications, 15
  contraindications, 13
  indications, 13
  materials and medications, 13
  pearls and pitfalls, 15
  procedure
    basilic and cephalic veins, 14
    brachial artery and vein, 14
    needle insertion, preparation for, 14
    Pythagorean Theorem, concept of, 14
    transverse and longitudinal orientation, 14
Peripheral vascular disease, 553
Peripheral venous catheterization
  contraindications, 693
  indications, 693
  materials and medications, 693, 694
  pearls and pitfalls, 699
  procedure, 694–699
    applying tourniquet, 694
    hub stabilizer placement, 695, 699
    passing catheter tip, 694, 697
    secure intravenous line, 695, 698
    vein, palpation of, 694, 696
Peripheral venous cutdown
  basilic vein, 41
  cephalic vein, 41
  complications, 44
  contraindications, 39
  indications, 39
  materials and medications, 39–40
  mini-cutdown technique, 43
  modified/guide wire technique, 43
  pearls and pitfalls, 44
  saphenous vein, 41
  standard venous cutdown technique, 41–43
    distal and proximal ligatures, 41, 42
    mobilization of vein, 41, 42
    skin incision, 41, 42
    venous catheter, 41–43
    venous incision, 41, 42
Peritonsillar abscess
  aspiration of, 349–350
  complications, 352
  contraindications, 349
  incision and drainage of, 351
  indications, 349
  materials and medications, 349
  pearls and pitfalls, 352
Phased array transducer (P17), 405
Phenylephrine
  nasal foreign body removal, 326
  priapism, 493
  septal hematoma, 322
Philadelphia collar, 219
Phimosis. *See also* Paraphimosis
  complications, 501
  contraindications, 497
  indications, 497
  materials and medications, 497
  pearls and pitfalls, 501
  procedure, 498–500
pH paper, 279, 280
Piles. *See* Thrombosed external hemorrhoid
Placenta previa
  complications, 664
  contraindications, 662
  grades, 661
  indications, 662
  materials and medications, 662
  pearls and pitfalls, 664
  procedure, 662–663
  varying degrees, 661

Placenta removal, 684
Pleural effusions, 173, 175
Pneumonia, 172, 173, 175
Pneumothorax, 156, 158, 405
Poisoning treatment. *See* Activated charcoal (AC) adsorption
Polyethylene glycol-electrolyte solution, 429, 431
Polymorphonuclear lymphocyte (PMN), 451
Posterior dislocation
  ankle, 618
  elbow
    anatomical depiction, 591
    method A, 592
    method B, 593
    method C, 593
  femur, 609, 610
  fibular head, 612
  shoulder, 583, 584
Posterior epistaxis
  management of, 317
  pearls and pitfalls, 317–318
  preliminary steps, 317
Postextraction hemorrhage
  complications, 378
  contraindications, 377
  indications, 377
  materials and medications, 377
  pearls and pitfalls, 378
  procedure, 377–378
Postparacentesis circulatory dysfunction (PPCD), 451
Postpartum hemorrhage (PPH), 668
  complications, 685
  contraindications, 683
  definition, 683
  indications, 683
  materials and medications, 683
  pearls and pitfalls, 685
  procedure
    continuous interlocking stitch technique, 685
    placenta removal, 684
    uterine atony, intrauterine massage for, 684
    uterine examination, 683
    uterine inversion management, 684
    vaginal examination, 683
  shoulder dystocia management, 675
Pregnancy
  blighted ovum (*see* Anembryonic gestation)
  complications (*see* Placenta previa)
  fetus/embryo, removal of (*see* Abortion)
  high-risk, 641
  hydatidiform mole (*see* Hydatidiform mole)
  tubal (*see* Ectopic pregnancy)
Priapism
  blood gas analysis, 494
  color Duplex ultrasonography, 495
  complete blood count, 495
  complications, 493
  contraindications, 485
  hemoglobin electrophoresis, 495
  indications, 485
  materials and medications, 485–486
  noninvasive therapy preprocedure, 487
  pearls and pitfalls, 493
  procedure
    cavernosal aspiration, 490–492
    dorsal penile nerve block, 487–489
    intracorporeal injection, 493
  psychoactive drug screen, 495
  sickle cell testing, 495
  urine toxicology, 495
Prophylaxis, 479, 482
PSLA. *See* Parasternal long-axis view (PSLA)
PSSA. *See* Parasternal short-axis view (PSSA)
Pulled elbow. *See* Nursemaid's elbow
Pulmonary artery catheter
  catheter preparation, 29
  complications, 33
  contraindications, 29
  indications, 29
  materials, 29
  pearls and pitfalls, 33
  procedure, 30–32
    balloon inflation, 30
    chest radiograph, correct placement, 30, 32
    sterile plastic sleeve, 30, 31
    thick and thin markings, 30, 31
    waveform and pressure readings, 30, 31
Pulmonary embolism
  clinical findings, 209
  complications, 209
  contraindications, 207
  indications, 207
  materials and medications, 207
  pearls and pitfalls, 209
  procedure, 207–208
    four-chamber view, 208
    inferior vena cava, 207–208
    PSLA, 207
    PSSA, 207, 208
    subxiphoid approach, 208
Pulmonary hypertension, 29, 30, 33
Pulseless electrical activity (PEA), 179, 189, 205
Pulseless ventricular tachycardia, 189

R
Radial artery cannulation. *See also* Arterial cannulation
  Allen test, 7–8
  over-the-needle catheter assembly with integrated guide wire, 3, 7
  pearls and pitfalls, 10–11
  standard over-the-needle catheter assembly
    materials and medications, 3, 4
    procedure, 5–6
Radial head subluxations. *See also* Nursemaid's elbow
  subluxation hyperpronated, 595
  subluxation hyperpronated and flexed, 595
  subluxation supinated, 595
Radial nerve block, wrist, 522
Rapid sequence intubation (RSI), 110
  complications, 65
  contraindications, 57
  indications, 57
  materials and medications, 57, 59
  pearls and pitfalls, 65
  procedure

        direct laryngoscopy (*see* Direct laryngoscopy)
        induction, 60, 62
        paralysis, 60, 62
        positioning, 60, 62
        postintubation management, 60, 64
        preoxygenation, 60, 62
        preparation, 60, 61
        pretreatment, 60, 62
        proof of placement, 60, 64
Rectal bleeding, 453
Rectal effluent, 431
Rectal pain, 453
Recurrent epistaxis, 313, 325
Re-expansion pulmonary edema, 153, 158
Regional nerve block
    complications, 515
    contraindications, 513
    digital nerve blocks, 523
        intrathecal/tendon sheath approach, 525
        ring block, 524
        web space approach, 524
    external ear block
        auricular ring block, 520
        indications, 520
    facial blocks, trigeminal nerve
        infraorbital nerve block, 517–518
        mental nerve block, 519
        suborbital and supratrochlear nerve block, 519
    general block, procedure for, 515
    indications, 513
    materials and medications, 514
    pearls and pitfalls, 515
    wrist nerve block
        indications, 521
        median nerve block, 522
        nerve distribution in hand, 521
        radial nerve block, 522
        ulnar nerve block, 523
Respiratory failure, 171, 173
Resuscitative thoracotomy, 165–168
    Finochietto retractor, 165, 166
    Lebsche knife, 165, 166
    open cardiac massage, 165, 167
    pericardium, 165, 167
    skin and subcutaneous tissue incision, 165, 167
Retinal detachment
    complications, 289
    contraindications, 287
    graphical representation, 292, 293
    indications, 287
    materials and medications, 287
    pearls and pitfalls, 289
    procedure, 287–288
        anatomy findings, 287, 288
        hyperechoic linear structure, 287, 288
        patient positioning, 287, 288
    ultrasound image, 292
Retrobulbar hemorrhage, 305
Reverse Woods corkscrew maneuver, 672
Rheumatoid arthritis (RA), 90, 625, 630
RSI. *See* Rapid sequence intubation (RSI)
Rubin II maneuver, 672
Rubin I maneuver, 672
Running suture, 540, 541

S
SAFE. *See* Sexual Assault Forensic Examination (SAFE)
Salivary glands, anatomy of, 359
Saphenous vein cutdown, 41
SB tube. *See* Sengstaken-Blakemore (SB) tube
Scalpel-bougie techniques. *See also* Cricothyroidotomy
    materials and medications, 119
    procedure, 121–122
Scalpel-Trousseau technique. *See also* Cricothyroidotomy
    materials and medications, 123
    procedure, 124–125
Scapular manipulation method, 586
Schiøtz tonometer, 297
    complications, 303
    pitfalls, 299
    procedure, 299
Schuknecht suction tube
    cerumen removal, 331
    foreign body removal
        ear, 336–338
        nasal, 326–328
Seidel test, 276
Seldinger technique
    DPL, 397–398
    femoral artery cannulation, 3, 9–10
Sengstaken-Blakemore (SB) tube
    complications, 439
    contraindications, 433
    materials and medications
        cuffalator, 433, 434
        helmet traction setup, 433, 436
        intubation equipment, 433
        lidocaine, 433
        lubricant (water soluble), 433
        scissors, 433, 436
        sphygmomanometer, 433, 434
        sterile water, 433
        suction device and tubing, 433, 435
        syringe with catheter tip, 433, 434
        3-way stop-valve tube connector, 433, 435
        tube clamps, 433, 435
        water and straw, 433
        Y-Tube connector, 433, 435
    pearls and pitfalls, 439
    procedure, 436–438
        balloons, placement of, 436, 437
        sphygmomanometer, connection of, 436, 438
    variceal GI bleed, 433
Sepsis, 685
    Edwards Vigileo System (*see* Edwards Vigileo System)
    fetal, 690
    maternal, 690
    peripheral venous cutdown, 44
Septal hematomas
    contraindications, 322
    indications, 321
    materials and medications, 322
    pitfalls and complications, 324
    procedure
        bilateral hematomas, drainage of, 322–323
        preprocedure, 322
        unilateral hematoma, drainage of, 322, 323

Serum-ascites albumin gradient (SAAG), 451
Sexual Assault Forensic Examination (SAFE)
  contraindications, 477
  indications, 477
  life/limb-threatening injuries, 477
  materials and medications, 477
  pearls and pitfalls, 483
  procedure
    biological material, collection of, 479
    bite marks, 478
    clothing, collection of, 479
    evidence collection kit, 478–480
    forensic interview and history taking, 478
    forensic pelvic examination, 478
    forensic photography, 479
    general physical examination, 478, 481
    informed consent, 478
    injury documentation, 478, 481
    inspection, 478
    investigations, 479
    law enforcement involvement, 478
    prophylaxis, CDC guidelines, 479, 482
    rectal examination, 478
    speculum examination, 478, 482
  SANE, 477
  SART, 477
  survivors, 477
Sexual Assault Nurse Examiner (SANE), 477
Sexual Assault Response Team (SART), 477
Shoulder arthrocentesis, 626, 629
Shoulder, dislocation reduction
  complications, 590
  contraindications, 585
  indications
    anterior dislocations, 583, 584
    inferior dislocations, 583
    posterior dislocations, 583, 584
  materials and medications, 585
  pearls, 590
  procedure
    external rotation method, 587
    Milch technique, 588
    pain management and sedation, 585
    physical examination, 585
    posterior shoulder, 589
    postreduction, 589
    radiographs, 585
    scapular manipulation method, 586
    Spaso technique, 588
    Stimson maneuver, 585
    traction-countertraction method, 589
Shoulder dystocia management
  complications, 675
  contraindications, 669
  indications, 669
  materials and medications, 669
  pearls and pitfalls, 675
  procedure
    accurate documentation, 670
    episiotomy consideration, 670
    Gaskin maneuver, 672
    help recruitment, 670
    McRoberts maneuver, 670, 671
    reverse Woods corkscrew maneuver, 672
    Rubin II maneuver, 672
    Rubin I maneuver, 672
    symphysiotomy, 674
    Woods corkscrew maneuver, 672, 673
    Zavanelli maneuver, 674
Sickle cell anemia, 495
Sickle-solubility test, 495
Silver nitrate stick ablation, 471, 475
Single-dose activated charcoal, 421
Slit lamp
  complications, 276
  contraindications, 272
  corneal foreign body removal, 284
  fluorescein testing (*see* Fluorescein testing)
  indications, 271, 272
  materials, 273
  pitfalls, 276
  procedure, 274–275
Spaso technique, 588
Sphenopalatine artery, 311, 312
Sphygmomanometer, 433, 434
Spine immobilization, 22, 231
Splinting
  complications, 574
  contraindications, 571
  DIP joint reduction, 600
  indications, 571
  materials and medications, 571–572
  procedure, 572–574
  sugar-tong (*see* Sugar-tong splint)
  ulnar gutter (*see* Ulnar gutter splint)
Spontaneous bacterial peritonitis, 451
Sterile urine aspiration. *See* Suprapubic bladder aspiration
Stifneck, 219
Stimson maneuver, 585
  femoral head manipulation, 604, 605
  with hand, 604
  with knee, 604
Stomach pumping. *See* Gastric lavage
Stomal closing, 131
Storz C-Mac, 116
Storz video laryngoscope
  complications, 118
  contraindications, 115
  indications, 115
  materials and medications, 116
  pearls and pitfalls, 118
  procedure
    blade, warming and insertion of, 116, 117
    ETT, placement of, 116, 117
    glotti, direct and indirect visualization of, 116, 117
String and yank technique, 550, 551
Subarachnoid hemorrhage, 241
Subclavian central venous line
  complications, 25–26
  pearls and pitfalls, 24
  procedure, 22–24
Subdural hematoma (SDH), 235
Sublingual abscess, 353
  complications, 356
  contraindications, 354
  incision and drainage, 355
  materials and medications, 354
  pearls and pitfalls, 356
Suborbital nerve block, 517
Subungal hematoma
  causes, 559

    complications, 561
    contraindications, 559
    indications, 559
    materials and medications, 559
    pearls and pitfalls, 561
    procedure
        positioning of needle, 560
        right index finger, 560
        trephination, 560
Sucking chest wound. *See* Open pneumothorax
Sugar-tong splint
    complications, 637
    contraindications, 633
    indications, 633
    materials, 633–634
    pearls and pitfalls, 637
    procedure, 635–637
        arm positioning, 635
        elastic bandage application, 635, 637
        splinting material application over padding, 635, 636
        stockinette positioning, 635
Superpronation method, nursemaid's elbow, 734
Supination method, nursemaid's elbow, 736
Supraglottic airways, 109
Suprapubic bladder aspiration
    complications, 722
    contraindications, 719
    indications, 719
    materials and medications, 719
    pearls and pitfalls, 722
    procedure
        applying negative pressure to syringe, 720
        landmark identification, 720, 721
        needle insertion, 720, 721
        occluding urethral opening, 720
        placing infant in frog-leg position, 720
Supratrochlear nerve block, 517
Surgical airway management. *See* Cricothyroidotomy
Swinging flashlight test, 305
Synchronized electrical cardioversion
    complications, 188
    contraindications, 185
    indications, 185
    materials and medications, 185
    pearls and pitfalls, 188
    procedure, 185–187
        anterolateral position, 186
        anteroposterior position, 186
        defibrillator/cardioverter, 186, 187
        pediatric pad placement, 186, 187
Synovial fluid aspiration. *See* Arthrocentesis

T
Tachyarrhythmias, 29, 193, 194
Tachycardia, 185
    awake patient, 143
    direct laryngoscopy, 74
    non-sinus-rhythm tachycardias, 185
    pacemaker-mediated, 213, 214
    ventilated patient, 143
TBSA. *See* Total body surface area (TBSA)
Temporary dental cement, 381
Temporary pacemakers, 202
Temporomandibular joint (TMJ) dislocation
    closed reduction
        contraindications for, 371
        with local anesthesia, 373
        without local anesthesia, 372
    complications, 374
    materials and medications, 371
    open lock, 371
    pearls and pitfalls, 374
Tension pneumothorax treatment. *See* Needle decompression thoracostomy; Tube thoracostomy
Testicular torsion
    contraindications, 505
    Doppler ultrasound of, 503, 505
    materials and medications, 505
    and normal testicle, schematic of, 503
    pearls and pitfalls, 512
    procedure, 505
    signs and symptoms, 503
Thoracentesis
    complications, 158
    contraindications, 155
    indications, 155
    materials and medications, 155
    pearls and pitfalls, 158
    procedure, 156–157
        intercostal neurovascular bundle, 156, 157
        needle and catheter insertion, 156, 157
        needle insertion mark, 156, 157
        patient positioning, 156
        pleural effusion location and height, 156
        pleural fluid collection via syringe pump method, 156, 157
    re-expansion pulmonary edema, 158
    ultrasound guidance, 158
Thoracic trauma, 159, 161
Thoracostomy
    needle (*see* Needle decompression thoracostomy)
    pneumothorax, 25, 162
    tube (*see* Tube thoracostomy)
Thoracotomy. *See* Emergent thoracotomy
Thread tourniquet. *See* Hair tourniquet
Threatened abortion, 657–659
Three-way stop clock, 701, 709, 713, 715, 716
Thrombin, 314, 377
Thrombosed external hemorrhoid
    complications, 394
    contraindications, 391
    indications, 391
    materials and medications, 391
    pearls and pitfalls, 394
    procedure, 392–393
        buttocks, taping of, 392, 393
        direct lighting over field, 392, 393
        discharge medicines and instructions, 392
        dressing, 392, 393
        jackknife position, 392, 393
        wound closure, 392
Tibia dislocation reduction. *See* Femur dislocation reduction
Tick removal
    complications, 556
    erythema migrans, 556
    indications, 555
    materials and medications, 555
    pearls and pitfalls, 556
    procedure, 555
TMJ dislocation. *See* Temporomandibular joint (TMJ) dislocation

- Tonometry
  - complications, 303
  - contraindications, 297
  - impression tonometry, 297
  - indications, 297
  - Schiøtz tonometer, 297
    - pitfalls, 299
    - procedure, 299
  - Tono-Pen (*see* Tono-Pen tonometry)
- Tono-Pen tonometry, 297, 298
  - applanation and indentation tonometry, 300
  - calibration position, 300
  - pitfalls, 301
  - pressure-sensitive electronics, use of, 300
  - procedure, 301
- Tooth avulsion. *See* Dental avulsion
- Total body surface area (TBSA), 532, 534
- Tourniquet
  - hair/thread (*see* Hair tourniquet)
  - peripheral venous catheterization, 694, 696, 697
  - peripheral venous cutdown, 41, 43
- Toxic ingestion treatment. *See* Activated charcoal (AC) adsorption
- Tracheal stenosis, 131
- Tracheostomy tube (TT) malfunction
  - indications, 127
  - materials and medications, 127
  - obstruction procedure, 127, 128
  - pearls and pitfalls
    - false lumen during recannulation, 131
    - stomal closing, 131
    - tracheal stenosis, 131
    - unrecognized trachea, 131
    - unrecognized tracheoesophageal fistula, 131
  - replacement procedure
    - bougie/NG tube, 128, 129
    - direct insertion, 127, 129
    - fingertip technique, 127, 130
- Trachlight™, 103, 107
- Traction-countertraction method, 585, 589, 609
- Transcranial Doppler ultrasonography, 267
- Transcutaneous pacing
  - contraindications, 193
  - indications, 193
  - materials and medications, 193
  - pearls and pitfalls, 196
  - procedure, 194–195
- Transillumination, 107
- Transtracheal jet ventilation (TTJV)
  - complications, 139
  - contraindications, 133
  - indications, 133
  - materials and medications, 133
  - pearls and pitfalls, 139
  - procedure
    - advancing catheter through cricothyroid membrane, 135, 137
    - attaching BVM, 135, 138
    - attaching catheter, 135, 136
    - attaching tubing and regulator, 135
    - operating the valve, 135, 138
    - palpating cricothyroid membrane, 135, 136
    - secure distal end of oxygen tubing, 135, 137
- Transvenous cardiac pacing
  - complications, 202
  - contraindications, 197
  - indications, 197
  - materials and medications, 198
  - pearls and pitfalls, 202
  - procedure, 199–201
    - cardiac monitor displaying, 201
    - guidewire placement, 199, 200
    - pacing wire balloon locking lever, 199, 201
    - pacing wire position, 201
    - ultrasound guidance, 199, 200
- Trauma
  - FAST examination (*see* Focused assessment with sonography for trauma (FAST) examination)
  - nasal, 311
  - thoracic, 159, 161
- Traumagram, 478, 481
- Traumatic effusion treatment. *See* Arthrocentesis
- Trichomoniasis, 468
- TTJV. *See* Transtracheal jet ventilation (TTJV)
- Tubal pregnancy. *See* Ectopic pregnancy
- Tube thoracostomy
  - complications, 153
  - contraindications, 149
  - indications, 149
  - materials and medications, 149–150
  - open pneumothorax, 161
  - pearls and pitfalls, 153
  - procedure, 151–152
- Two-handed seal technique, BVM ventilation, 48, 50
- Two-point stabilization technique, 229, 230

U

- Ulna bone fracture. *See* Sugar-tong splint
- Ulnar gutter splint
  - complications, 581
  - contraindications, 575
  - indications, 575
  - materials and medications, 575
  - pearls and pitfalls, 581
  - procedure, 577–580
- Ulnar nerve block, wrist, 523
- Ultrasonography
  - abortion (*see* Abortion)
  - abscess incision and drainage, 565,
  - bladder aspiration, 722
  - blighted ovum (*see* Anembryonic gestation)
  - ectopic pregnancy (*see* Ectopic pregnancy)
  - fluid determination, 449, 451
  - hydatidiform mole (*see* Hydatidiform mole)
  - intraocular pathology (*see* Intraocular pathology)
  - lung (*see* Lung ultrasonography)
  - pelvic, 647–648
  - pericardiocentesis (*see* Pericardiocentesis)
  - peripheral intravenous access (*see* Peripheral intravenous (PIV) access, ultrasound-guided)
  - placenta previa (*see* Placenta previa)
  - pulmonary embolism and heart strain (*see* Pulmonary embolism)
  - retinal detachment (*see* Retinal detachment)
  - shoulder anatomy, 583, 584
  - thoracentesis, 158
- Umbilical cord prolapse. *See* Vaginal delivery
- Umbilical tape, 701, 703, 705
- Umbilical venous catheters (UVC)
  - complications, 705
  - contraindications, 701
  - indications, 701
  - insertion procedure

        catheter insertion, 703, 704
        dilating umbilical vein, 703securing catheter, 703, 705
        umbilical catheter tape bridge, creation of, 703, 704
        umbilical vessels, identification of, 703
    materials and medications, 701–702
    pearls and pitfalls, 706
    removal procedure, 705
Unsynchronized cardioversion
    complications, 192
    contraindications, 189
    high-energy shock, 189
    indications, 189
    materials and medications, 189
    pearls and pitfalls, 192
    procedure, 191–192
Upper gastrointestinal (GI) bleeding evaluation. *See* Nasogastric tube (NT) placement
Urine collection. *See* Suprapubic bladder aspiration
Uterine atony
    delivery, 667
    management, 684
Uterine inversion management, 684
UVC. *See* Umbilical venous catheters (UVC)

V
Vaginal bleeding
    anembryonic gestation, 655
    early intrauterine pregnancy, 657
    ectopic pregnancy, 646, 648
    hydatidiform mole, 651
    pelvic examination and wet preparation, 465, 468
    placenta previa, 662, 664
    postpartum hemorrhage, 683
    threatened abortion, 659
Vaginal delivery
    complications, 668
    contraindications, 665
    indications, 665
    materials and medications, 665–667
    pearls and pitfalls, 668
    procedure, 667
Ventricular fibrillation (VF, 181, 185, 188, 189
Ventriculostomy tube. *See* External ventricular drain (EVD)
Vertical mattress suture, 540, 542
Vestibuloocular reflex. *See* Caloric reflex testing
Vision loss. *See* Retinal detachment
Vitreous detachment, 292
Vitreous hemorrhage, 292

W
Web space digital block, 524
Wet preparation. *See* Pelvic examination and wet preparation
Whistler technique, 607
Whole-bowel irrigation (WBI)
    contraindications, 429
    indications, 429
    materials and medications
        bedside commode, 429, 430
        lavage solution bag, 430
        nasogastric tube, 429
        polyethylene glycol-electrolyte solution, 429, 431
    pearls, 431
    procedure, 431
Woods corkscrew maneuver, 672, 673
Word catheter, 471, 472, 474Wound closure
    adhesive tape, 543
    complications, 530
    contraindications, 537
    with cyanoacrylate tissue adhesive
        complications, 547
        contraindications, 545
        indications, 545
        materials and medications, 545
        pearls and pitfalls, 547
        procedure, 546
    and hemostasis, procedure for, 527–529
        dermabond, 528
        direct pressure, Surgicel, 528
        figure-of-eight suture, 528, 529
        high-pressure irrigation, 527, 528
        staples, 527
        two-person approach, 528
        tying absorbable suture, 528, 529
    indications, 537
    methods, 537
    pearls and pitfalls, 530, 544
    preparation, 537
    primary wound repair
        contraindications, 527
        indications, 527
    staple closure, 544
    suture repair
        deep dermal suture, 540, 541
        general guidelines, 539
        half-buried horizontal mattress suture, 540, 543
        horizontal mattress suture, 540, 542
        materials and medications, 538
        recommendations, 540
        running suture, 540, 541
        simple interrupted sutures, 540
        suture material, 539
        suture size, 539
        vertical mattress suture, 540, 542
Wrist fracture, sugar-tong splint. *See* Sugar-tong splint
Wrist nerve blocks
    indications, 521
    median nerve, 522
    nerve distribution in hand, 521
    radial nerve, 522
    ulnar nerve, 523

Y
Y-Tube connector, 433, 435

Z
Zavanelli maneuver, 674

IMMUNOLOGY

Seventh Edition

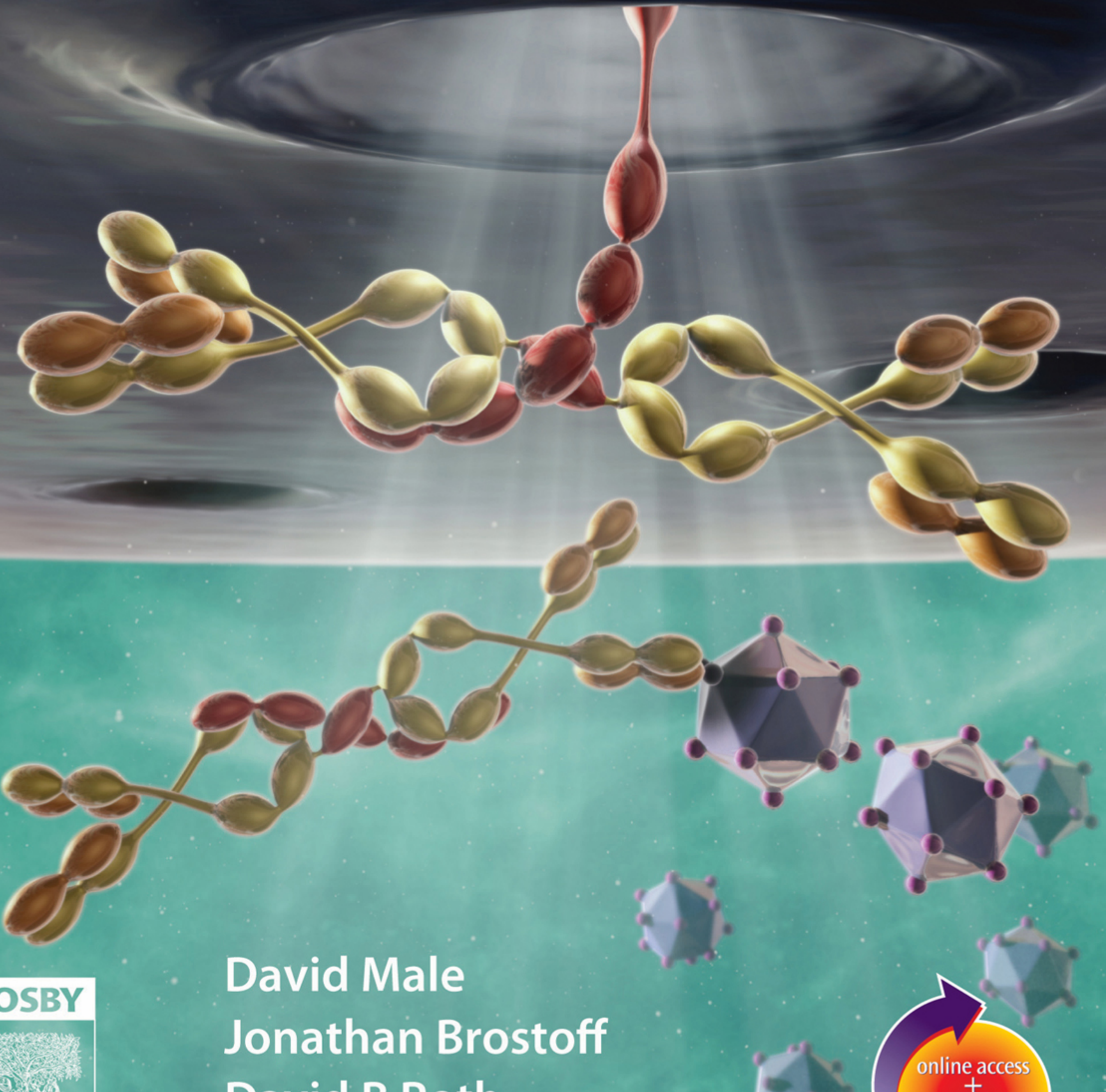

David Male
Jonathan Brostoff
David B Roth
Ivan Roitt

MOSBY
ELSEVIER

online access
+
interactive extras
studentconsult.com

IMMUNOLOGY

Commissioning Editor: Inta Ozols
Development Editor: Louise Cook
Editorial Assistant: Nani Clansey
Project Manager: Jess Thompson
Design Manager: Jayne Jones
Illustration Manager: Bruce Hogarth
Illustrator: Cactus Illustration
Marketing Manager(s) (UK/USA): Ian Jordan and John Gore

Immunology

Seventh Edition

David Male MA PhD
Professor of Biology
Department of Biological Sciences
The Open University
Milton Keynes, UK

Jonathan Brostoff MA DM(Oxon) DSc(Med) FRCP (Lond) FRCPath FIBiol
Professor Emeritus of Allergy and Environmental Health
School of Biomedical and Health Sciences
King's College London
London, UK

David B Roth MD PhD
Irene Diamond Professor of Pathology
Investigator, Skirball Institute
Program in Molecular Pathogenesis
Chairman, Department of Pathology
New York University School of Medicine
New York, NY, USA

Ivan Roitt MA DSc (Oxon) Hon FRCP (Lon) FRCPath FRS
Emeritus Professor of Immunology
Royal Free and University College Medical School
University College
London, UK

MOSBY
ELSEVIER

MOSBY
ELSEVIER
An imprint of Elsevier Limited

© 2006, Elsevier Limited. All rights reserved.

The right of Dale Male, Jonathan Brostoff, David B Roth and Ivan Roitt to be identified as authors of this work has been asserted by them in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the Publishers. Permissions may be sought directly from Elsevier's Health Sciences Rights Department, 1600 John F. Kennedy Boulevard, Suite 1800, Philadelphia, PA 19103-2899, USA: phone: (+1) 215 239 3804; fax: (+1) 215 239 3805; or, e-mail: *healthpermissions@elsevier.com*. You may also complete your request on-line via the Elsevier homepage (http://www.elsevier.com), by selecting 'Support and contact' and then 'Copyright and permission'.

First edition published by Gower Medical Publishing Ltd., 1985
Second edition published by Gower Medical Publishing Ltd., 1989
Third edition published by Mosby-Year Book Europe Ltd., 1993
Fourth edition published by Mosby, an imprint of Times Mirror International Publishers, 1996
Fifth edition published by Mosby, an imprint of Times Mirror International Publishers, 1998
Sixth edition published by Elsevier Ltd, 2001
Seventh edition published by Elsevier Ltd, 2006
Reprinted 2007

Main edition
ISBN 13: 978 0 323 03399 2
ISBN 10: 0 323 03399 7

International edition
ISBN 13: 978 0 8089 2332 9
ISBN 10: 0 808 92332 3

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress

Notice
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our knowledge, changes in practice, treatment and drug therapy may become necessary or appropriate. Readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of the practitioner, relying on their own experience and knowledge of the patient, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions. To the fullest extent of the law, neither the Publisher nor the Authors assume any liability for any injury and/or damage to persons or property arising out or related to any use of the material contained in this book.

The Publisher

ELSEVIER your source for books, journals and multimedia in the health sciences
www.elsevierhealth.com

Working together to grow
libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Printed in Canada

Last digit is the print number 9 8 7 6 5 4 3 2

The
Publisher's
policy is to use
paper manufactured
from sustainable forests

CONTENTS

Preface and user guide *vii*
List of Contributors *ix*

Section 1 Components of the Immune System

1 Introduction to the immune system *3*
David Male

2 Cells, tissues and organs of the immune system *19*
Peter Lydyard and Carlo E Grossi

3 Antibodies *59*
Roy Jefferis

4 Complement *87*
B Paul Morgan

5 T cell receptors and MHC molecules *105*
David B Roth

Section 2 Modes of Immune Response

6 Mechanisms of innate immunity *127*
David Male

7 Antigen presentation *145*
John Trowsdale

8 Cell Cooperation in the antibody response *163*
Joseph C Marini

9 Mononuclear phagocytes in immune defence *181*
Siamon Gordon

10 Cell mediated cytotoxicity *203*
Gillian M Griffiths

11 Regulation of the immune response *215*
Arne Akbar and Joanne E Cook

12 Immune responses in tissues *235*
David Male

Section 3 Defence Against Infectious Agents

13 Immunity to viruses *247*
Anthony A Nash

14 Immunity to bacteria and fungi *257*
Gregory J Bancroft

15 Immunity to protozoa and worms *277*
Janette E Bradley and Richard Pleass

16 Primary immunodeficiency *299*
Fred S Rosen

17 AIDS and secondary immunodeficiency *311*
Bruce D Walker

18 Vaccination *325*
Peter C L Beverley

Section 4 Immune Responses Against Tissues

19 Immunological tolerance *343*
D C Wraith

20 Autoimmunity and autoimmune disease *365*
Ivan Roitt

21 Transplantation and rejection *383*
Robert Lechler and Andrew J T George

22 Immunity to Cancers *401*
P K Srivastava

Section 5 Hypersensitivity

23 Immediate hypersensitivity (Type I) *423*
Thomas A E Platts-Mills

24 Hypersensitivity (Type II) *449*
David Male

25 Hypersensitivity (Type III) *461*
Frank Hay, Olwyn M R Westwood

26 Type IV hypersensitivity *477*
Warwick Britton

Explanations for critical thinking and case studies *493*

Appendices

1. Major histocompatibility complex *507*
2. Recognized CD markers of hemopoietic cells and their distribution *508*
3. The major cytokines *522*
4. Human chemokines and their receptors *524*

Glossary *527*

Index *535*

This Page Intentionally Left Blank

PREFACE

Immunology as a distinctive subject developed in the middle of the 20th century as researchers started to understand how the adaptive immune system aids in defence against pathogens. For many years, text books reflected this view of immunology with short opening sections on innate defences, before giving detailed descriptions of how lymphocytes protect the body against infection. In the last few years, the neglected but critical elements of the innate immune system have come back into sharp focus and have been the subject of much research. Phagocytes and the evolutionarily ancient mechanisms for recognising pathogen, have been retained in mammals and indeed have developed alongside the adaptive immune system; it is only as we have started to understand how innate immunity operates, that we have started to recognise the importance of these systems. In the seventh edition of *Immunology*, we have included a new chapter on innate immunity, but equally important is the assimilation of descriptions of innate defences into other parts of the book. This truly reflects the way the immune system operates with the integration of ancient and recently-evolved immune defences.

The book is organised into five sections. The opening section describes the building blocks of the immune system, - cells, organs, complement and the major receptor molecules, including antibodies, T cell receptors and MHC molecules. The second section deals with the initiation of the immune response starting with innate defences, and leading on to antigen presentation, costimulation and cell activation pathways. The following three chapters discuss the principal effector arms of the immune response, TH2 responses with antibody production, TH1 responses and mononuclear phagocytes, and cytotoxicity, including cytotoxic T cells and NK cells. In this edition, material on individual cytokines has been incorporated into each of these chapters, rather than being treated as a separate subject. The final chapters in this section look into the regulation of the immune response, and there is a new chapter on the distinctive types of immune response that develop in different tissues of the body. Although it has long been recognised that immune responses in tissues vary, the underlying reasons for the differences are only just being elucidated.

Section three describes the immune responses that develop against different types of infection, and how immunodeficiency leads to increased susceptibility to particular infections. Indeed the diversity and complexity of the immune system can only be understood in relation to the diversity of the pathogens which it protects against. In recent years, the devious strategies employed by pathogens to evade immune responses, have provided some quite startling revelations, both on the adaptability of the pathogens and on the flexibility of the immune system. Ultimately the immune system can only be understood in relation to its principal function – defence against pathogens.

Section four describes immune responses against tissues, and section five hypersensitivity. These areas are of great clinical importance. One aim of this book is to provide readers with a sound understanding of the immune responses which underlie clinically important areas including hypersensitivity states and allergy, immunopathology, tumour immunotherapy and transplantation. In these sections, we have maintained what we believe to be an important feature of the book, namely a clear description of the scientific principles of clinical immunology, integrated with histology, pathology, and clinical examples. In this edition, the two chapters on tolerance and autoimmunity have been linked and rewritten, to reflect how the traditional system-based approach of clinical immunology and the cell-based approach of basic immunology are finally producing a coherent picture of how autoimmune disease develops.

Previously, we have included a final chapter which collates immunological techniques. In this edition the methods and techniques are distributed throughout the book, in methods boxes, in the most appropriate place. We hope that this will make it easier to locate relevant techniques if they are of interest, but allow the reader to pass over such material, if it is not.

A new feature of the book is the inclusion of in-text questions. These are designed to check that the reader understands the implications of the preceeding paragraphs or can relate that material to information in earlier chapters. Although the answers follow on immediately below the questions, we would suggest that readers pause to formulate an answer, before reading on through. Another useful learning aid is the critical thinking sections at the end of each chapter. A set of solutions to the problems is included at the end of the book, although some of the questions are open ended and would form a good basis for class discussion or tutorials. Finally, we have put a lot of care into the chapter summaries, ensuring that they really do distil the key aspects of each chapter, into a manageable overview. The summary boxes make an excellent revision guide for exams, in addition to setting the framework for each chapter.

In conjuction with the book, the *Immunology* seventh edition website provides access to the full set of illustrations and text on-line. Other ancillaries include a set of 16 animations, with a total of 90 minutes playing time, that were originally developed as one component of Immunology Interactive 3.0. The site also contains a question-bank and other ancillaries, to aid learning and enhance understanding of the subject.

The contributors to this volume include many experts in different areas of immunology, but we are particularly pleased to welcome David Roth as a new editor. We hope that David's contribution has further broadened our appreciation of how immunology is taught in other countries, in addition to his scientific input. Also, we would like to thank Jane Loughlin and Lindy van den Berghe for their critical reading of the text, suggestions for changes and contribution to the in-text questions. We greatly appreciate the hard work of our Publishers and their colleagues, particularly, Inta Ozols, Jess Thompson and Louise Cook from Elsevier.

Immunology bridges basic sciences and medicine and encompasses approaches from numerous fields, including biochemistry, genetics, cell biology, structural biology and molecular biology. For the past century, immunology has fascinated and inspired some of the greatest scientific thinkers of our time and numerous Nobel prizes have been awarded for fundamental discoveries in immunology, from Paul Ehrlich's work on antibodies (1908) to the studies of Zinkernagel and Doherty (1996) elucidating mechanisms of cell-mediated immunity. We wish our readers well in their study of immunology, a subject which continues to excite and surprise us, and which underpins many other areas of biology and biomedical sciences.

David Male
Jonathan Brostoff
David B Roth
Ivan Roitt
2006

The following are standard shapes which are used consistently throughout the book

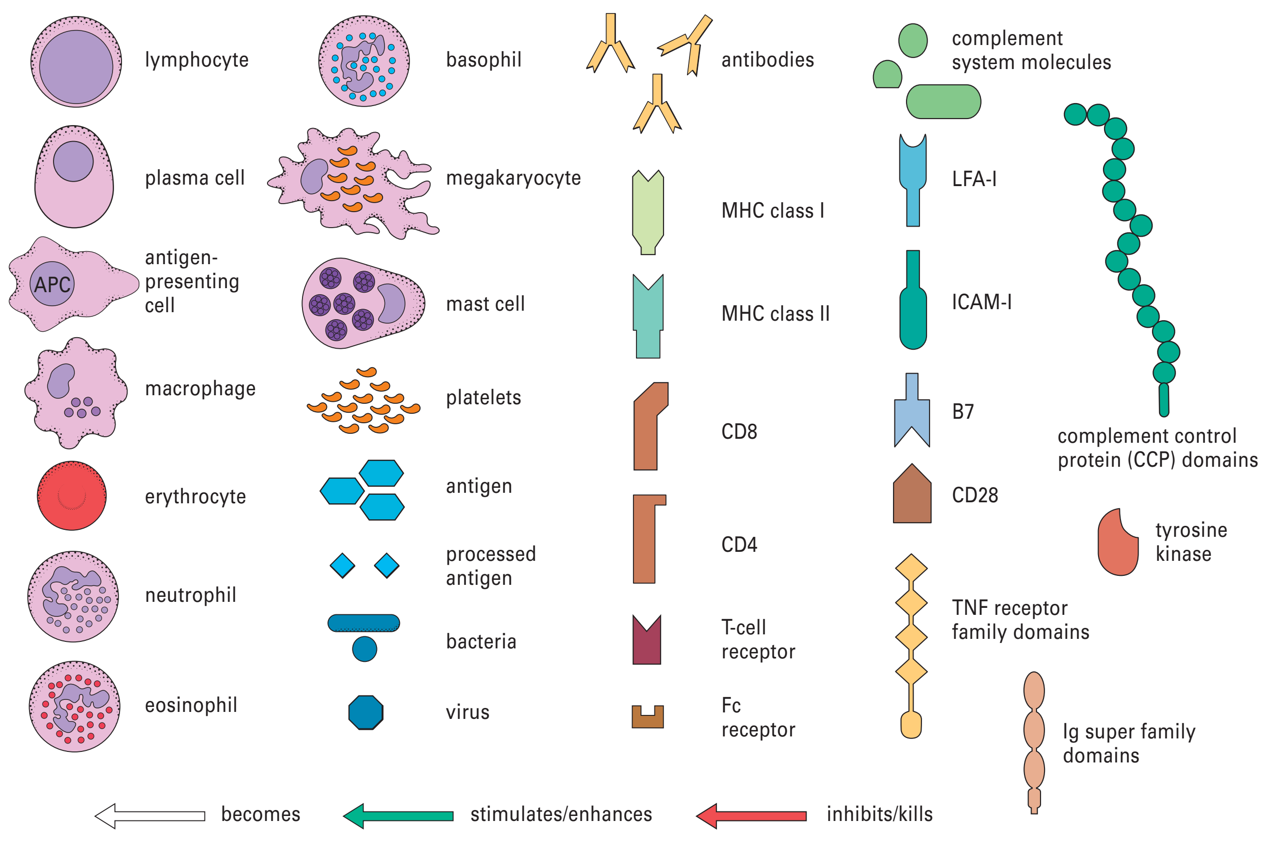
lymphocyte
plasma cell
APC
antigen-presenting cell
macrophage
erythrocyte
neutrophil
eosinophil
basophil
megakaryocyte
mast cell
platelets
antigen
processed antigen
bacteria
virus
antibodies
MHC class I
MHC class II
CD8
CD4
T-cell receptor
Fc receptor
complement system molecules
LFA-I
ICAM-I
B7
CD28
TNF receptor family domains
complement control protein (CCP) domains
tyrosine kinase
Ig super family domains
becomes
stimulates/enhances
inhibits/kills

LIST OF CONTRIBUTORS

Arne Akbar
Professor of Immunology
Department of Immunology and Molecular Pathology
Royal Free and University College Medical School
Windeyer Institute of Medical Sciences
London, UK

Gregory J Bancroft
Reader in Immunology
Department of Infectious and Tropical Diseases
London School of Hygiene and Tropical Medicine
London, UK

Peter Beverley
Principal Reseach Fellow
The Edward Jenner Institute for Vaccine Research
Compton
Berkshire, UK

Janette E Bradley
Chair of Parasitology
School of Biology
University of Nottingham
Nottingham, UK

Jonathan Brostoff
Professor Emeritus of Allergy and Environmental Health
School of Biomedical and Health Sciences
King's College London
London, UK

Warwick Britton
Professor of Medicine
Department of Medicine
Central Clinical School
University of Sydney
Sydney, NSW, Australia

Joanne E Cook
Postdoctoral Research Assistant
Department of Immunology and Molecular Pathology
Royal Free and University College Medical School
Windeyer Institute of Medical Sciences
London, UK

Andrew J T George
Professor of Molecular Immunology
Department of Immunology
Division of Medicine
Imperial College London
London, UK

Siamon Gordon
Professor of Cellular Pathology
Sir William Dunn School of Pathology
University of Oxford
Oxford, UK

Gillian M Griffiths
Professor of Experimental Pathology
Sir William Dunn School of Pathology
University of Oxford
Oxford, UK

Carlo E Grossi
Head of Department
Department of Experimental Medicine
Section of Human Anatomy
University of Genova
Genova, Italy

Frank Hay
Professor of Immunology and Society of Apothecaries
Lecturer in the History of Medicine
Department of Biochemistry and Immunology
St George's Hospital Medical School
London, UK

Roy Jefferis
Professor of Molecular Immunology
Division of Immunity and Infection
School of Medicine
University of Birmingham
Birmingham, UK

Robert I Lechler
Vice-Principal (Health)
Kings College London
London, UK

Jane Loughlin
Senior Lecturer in Cell Biology
Department of Biological Sciences
The Open University
Walton Hall
Milton Keynes
Bucks, UK

Peter Lydyard
Professor of Immunology
Department of Immunology and Molecular Pathology
Division of Infection and Immunity
Royal Free and University College Medical School
Windeyer Institute of Medical Sciences
London, UK

David Male
Professor of Biology
Department of Biological Sciences
The Open University
Walton Hall
Milton Keynes
Bucks, UK

Joseph C Marini
Principal Research Scientist
Clinical Pharmacology and Experimental Medicine
Centocor Research and Development
Radnor
PA, USA

B Paul Morgan
Professor of Medical Biochemistry and Immunology
Department of Medical Biochemistry and Immunology
School of Medicine
University of Cardiff
Cardiff, UK

Anthony A Nash
Professor of Veterinary Pathology
Centre for Infectious Diseases
University of Edinburgh
Edinburgh, UK

Thomas A E Platts-Mills
Professor of Medicine
Asthma and Allergic Diseases Center
Charlottesville, VA, USA

Richard J Pleass
Lecturer of Parasite Immunology
Scholl of Biology
Institute of Genetics
Queens Medical Centre
University of Nottingham
Nottingham, UL

Ivan Roitt
Emeritus Professor of Immunology
Royal Free and University College Medical School
University College
London, UK

The late Fred S Rosen
Formerly Professor of Pediatrics
Harvard Medical School
The Centre for Blood Research
Boston , MA, USA

David B Roth
Irene Diamond Professor of Pathology
Investigator, Skirball Institute of Biomolecular Medicine
Program in Molecular Pathogenesis
Chairman, Departmental of Pathology
New York University School of Medicine
New York, NY, USA

P K Srivastava
Professor of Immunology
Director of the Center for Immunotherapy of Cancer and Infectious Diseases
University of Connecticut School of Medicine
Farmington, CT, USA

John Trowsdale
Head of Division of Immunology
Cambridge University
Department of Pathology
Immunology Division
Cambridge, UK

Bruce D Walker
AIDS Research Center
Massachusetts General Hospital
Charlestown , MA, USA

Olwyn M R Westwood
MSc Programme Co-ordinator
School of Life Sciences
University of Surrey Roehampton
London, UK

D C Wraith
Professor of Experimental Pathology
Department of Pathology and Microbiology
University of Bristol
School of Medical Sciences
Bristol, UK

Components of the Immune System

SECTION 1

This Page Intentionally Left Blank

CHAPTER

1

Introduction to the Immune System

SUMMARY

- **The immune system has evolved to protect us from pathogens.** Intracellular pathogens infect individual cells (e.g. viruses), whereas extracellular pathogens divide extracellularly within tissues or the body cavities (e.g. many bacteria).

- **Phagocytes and lymphocytes are key mediators of immunity.** Phagocytes internalize pathogens and degrade them. Lymphocytes (B and T cells) bear receptors that recognize specific molecular components of pathogens and have specialized functions. B cells make antibodies, cytotoxic T lymphocytes (CTLs) kill virally infected cells, and helper T cells coordinate the immune response by direct cell–cell interactions and the release of cytokines.

- **Specificity and memory are two essential features of adaptive immune responses.** As a result the immune system mounts a more effective response on second and subsequent encounters with a particular antigen. Non-adaptive (innate) immune responses do not alter on repeated exposure to an infectious agent.

- **Antigens are molecules that are recognized by receptors on lymphocytes.** B cells usually recognize intact antigen molecules, whereas T cells recognize antigen fragments on the surface of other cells.

- **An immune response occurs in two phases – antigen recognition and antigen eradication.** In the first phase clonal selection involves recognition of antigen by particular clones of lymphocytes, leading to clonal expansion of specific clones of T and B cells and differentiation to effector and memory cells. In the effector phase, these lymphocytes coordinate an immune response, which eliminates the source of the antigen.

- **Vaccination depends on the specificity and memory of adaptive immunity.** Vaccination is based on the key elements of adaptive immunity, namely specificity and memory. Memory cells allow the immune system to mount a much stronger response on a second encounter with antigen.

- **Inflammation is a response to tissue damage.** It allows antibodies, complement system molecules, and leukocytes to enter the tissue at the site of infection, resulting in phagocytosis and destruction of the pathogens. Lymphocytes are also required to recognize and destroy infected cells in the tissues.

- **The immune system may fail (immunopathology).** This can lead to immunodeficiency, hypersensitivity, or autoimmune diseases.

- **Normal immune reactions can be inconvenient in modern medicine**, for example blood transfusion reactions and graft rejection.

THE IMMUNE SYSTEM PROTECTS US FROM PATHOGENS

A highly discriminatory immune system is fundamental to survival. How the immune system accomplishes this level of discrimination remains deeply enigmatic, but such questions are among the many that make immunology a fascinating discipline.

The immune system has evolved a powerful collection of defense mechanisms to protect against potential invaders that would otherwise take advantage of the rich source of nutrients provided by the vertebrate host. At the same time it:

- must be sophisticated enough to differentiate between the individual's own cells and those of harmful invading organisms and not attack the commensal flora that inhabit the gut, skin, and many other tissues to great benefit;
- should avoid rejecting tissue that is demonstrably foreign, namely the fetus.

This chapter provides an overview of the complex network of processes that form the immune system of higher vertebrates. It:

- illustrates how the components of the immune system fit together to allow students to rapidly grasp the 'big picture' before delving into the material in more depth in subsequent chapters;
- introduces the basic elements of the immune system and of immune responses, which are mediated principally by white blood cells or **leukocytes** (from the Greek for 'white cell') and are detailed in Chapters 2–18.

A description of cells and molecules important in immunology is given in Figs 1.1–1.4.

Over many millions of years, different types of immune defense, appropriate to the infecting pathogens, have evolved in different groups of organisms. In this book, we concentrate on the immune systems of mammals, especially humans. Because mammals are warm-blooded and long-lived, their immune systems have evolved particularly sophisticated systems for recognizing and destroying pathogens.

Q. Why do warm-blooded, long-lived animals require particularly complex immune defenses?

A. Infectious agents such as bacteria can divide rapidly in warm-blooded creatures. Animals have to remain healthy during their reproductive years to raise offspring.

Many of the immune defenses that have evolved in other vertebrates (e.g. reptiles, amphibians) and other phyla (e.g. sponges, worms, insects) are also present in some form in mammals and play crucial roles. Consequently the mammalian immune system consists of multi-layered, interlocking defense mechanisms that incorporate both primitive and recently evolved elements.

THE CELLS AND SOLUBLE MEDIATORS OF THE IMMUNE SYSTEM

Cells of the immune system

Immune responses are mediated by:

- a variety of cells; and
- the soluble molecules that these cells secrete (Fig. 1.1).

Although the leukocytes are central to all immune responses, other cells in the tissues also participate, by signaling to the lymphocytes and responding to the cytokines released by T cells and macrophages.

Phagocytes internalize antigens and pathogens and break them down

The most important long-lived phagocytic cells belong to the mononuclear phagocyte lineage. These cells are all derived from bone marrow stem cells, and their function is to:

- engulf particles, including infectious agents;
- internalize them; and
- destroy them.

For this purpose mononuclear phagocytes are strategically placed where they will encounter such particles. For example, the Kupffer cells of the liver line the sinusoids along which blood flows, while the synovial A cells line the synovial cavity (Fig. 1.2).

MONONUCLEAR PHAGOCYTES INCLUDE THE LONG-LIVED MONOCYTES AND MACROPHAGES – Leukocytes of the mononuclear phagocyte lineage are called **monocytes**. Monocytes migrate from the blood into the tissues, where they develop into tissue **macrophages**, which are very effective at presenting antigens to T cells (see Chapter 2, p. 21).

POLYMORPHONUCLEAR NEUTROPHILS ARE SHORT-LIVED PHAGOCYTES – Polymorphonuclear neutrophils (often just called **neutrophils** or **PMNs**) are another important group of phagocytes. Neutrophils constitute the majority

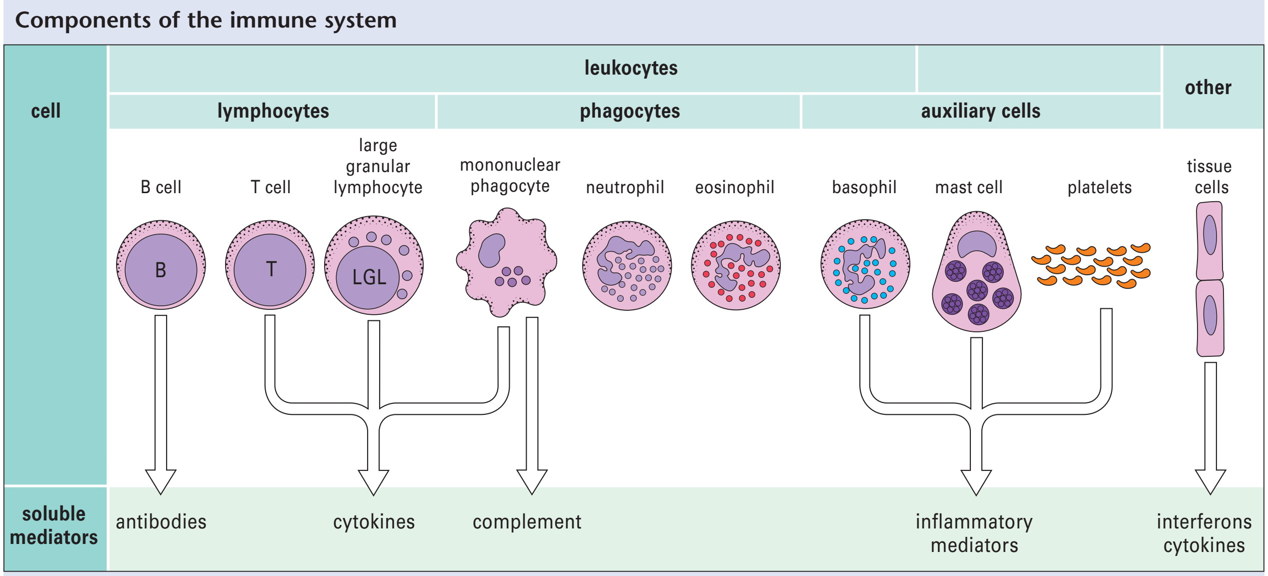

Fig. 1.1 The principal cells of the immune system and the mediators they produce are shown. Neutrophils, eosinophils, and basophils are collectively known as polymorphonuclear granulocytes (see Chapter 2). Cytotoxic cells include cytotoxic T lymphocytes (CTLs), natural killer (NK) cells (large granular lymphocytes [LGLs]), and eosinophils. Complement is made primarily by the liver, though there is some synthesis by mononuclear phagocytes. Note that each cell produces and secretes only a particular set of cytokines or inflammatory mediators.

Cells of the mononuclear phagocyte lineage

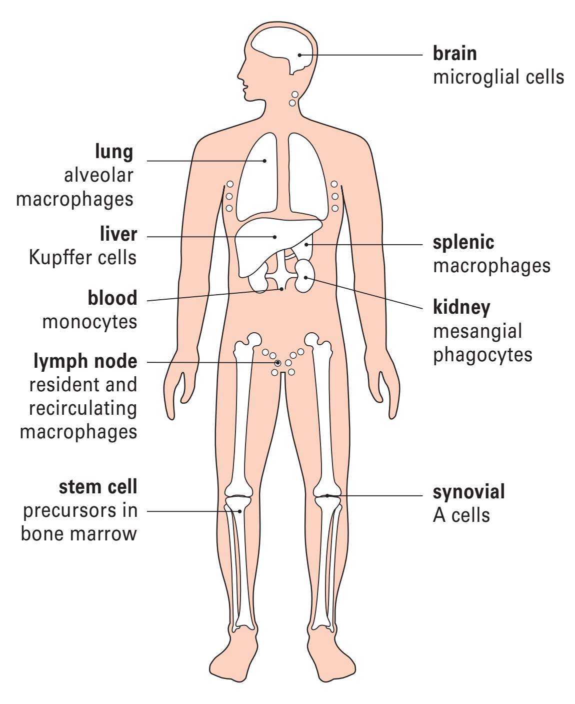

Fig. 1.2 Many organs contain cells belonging to the mononuclear phagocyte lineage. These cells are derived from blood monocytes and ultimately from stem cells in the bone marrow.

of the blood leukocytes and develop from the same early precursors as monocytes and macrophages.

Like monocytes, neutrophils migrate into tissues, particularly at sites of inflammation, However, neutrophils are short-lived cells that phagocytose material, destroy it, and then die.

B cells and T cells are responsible for the specific recognition of antigens

Lymphocytes are wholly responsible for the specific immune recognition of pathogens, so they initiate adaptive immune responses. All lymphocytes are derived from bone marrow stem cells, but T lymphocytes (T cells) then develop in the thymus, while B lymphocytes (B cells) develop in the bone marrow (in adult mammals).

B CELLS EXPRESS ANTIBODY – Each B cell is genetically programmed to express a surface receptor specific for a particular antigen. This antigen receptor molecule is called an **antibody**. If a B cell binds to its specific antigen, it will multiply and differentiate into **plasma cells** (see Fig. 1.13), which produce large amounts of the antibody, but in a secreted form.

Secreted antibody molecules are large glycoproteins found in the blood and tissue fluids. Because secreted antibody molecules are a soluble version of the original receptor molecule (antibody), they bind to the same antigen that initially activated the B cells.

T CELLS HAVE T CELL ANTIGEN RECEPTORS – There are several different types of T cell, and they have a variety of functions:

- one group interacts with mononuclear phagocytes and helps them destroy intracellular pathogens – these are called **type 1 helper T cells or TH1 cells**;
- another group interacts with B cells and helps them to divide, differentiate, and make antibody – these are the **type 2 helper T cells or TH2 cells**;
- a third group of T cells is responsible for the destruction of host cells that have become infected by viruses or other intracellular pathogens – this kind of action is called cytotoxicity and these T cells are therefore called **cytotoxic T lymphocytes (CTLs or Tc cells)**.

In every case, the T cells recognize antigens present on the surface of other cells using a specific receptor – the **T cell antigen receptor (TCR)** – which is quite distinct from, but related in structure to, the antigen receptor (antibody) on B cells.

T cells generate their effects either:

- by releasing soluble proteins, called **cytokines**, which signal to other cells; or
- by direct cell–cell interactions.

Cytotoxic cells recognize and destroy other cells that have become infected

Several cell types have the capacity to kill other cells should they become infected. Cytotoxic cells include CTLs, natural killer (NK) cells (large granular lymphocytes), and eosinophils. Of these, the CTL is especially important, but other cell types may be active against particular types of infection.

All of these cell types damage their different targets by releasing the contents of their intracellular granules close to them. Cytokines secreted by the cytotoxic cells, but not stored in granules, contribute to the damage.

LARGE GRANULAR LYMPHOCYTES ARE ALSO KNOWN AS NK CELLS – Lymphocytes known as large granular lymphocytes (LGLs) have the capacity to recognize the surface changes that occur on a variety of tumor cells and virally infected cells. LGLs damage these target cells, but use a different recognition system to CTLs. This action is sometimes called NK cell activity, so these cells are also described as NK cells.

EOSINOPHIL POLYMORPHS ENGAGE AND DAMAGE LARGE EXTRACELLULAR PARASITES – Eosinophils are a specialized group of leukocytes that have the ability to engage and damage large extracellular parasites, such as schistosomes.

Auxiliary cells control inflammation

The main purpose of inflammation is to attract leukocytes and the soluble mediators of immunity towards a site of infection. Inflammation is mediated by a variety of other cells including basophils, mast cells, and platelets.

BASOPHILS AND MAST CELLS RELEASE MEDIATORS OF INFLAMMATION – Basophils and mast cells have granules that contain a variety of mediators, which:
- induce inflammation in surrounding tissues; and
- are released when the cells are triggered.

Basophils and mast cells can also synthesize and secrete a number of mediators that control the development of immune reactions.

Mast cells lie close to blood vessels in all tissues, and some of their mediators act on cells in the vessel walls. Basophils are functionally similar to mast cells, but are mobile, circulating cells.

PLATELETS RELEASE MEDIATORS OF INFLAMMATION – Platelets are essential in blood clotting, but can also be activated during immune responses to release mediators of inflammation.

Soluble mediators of immunity

A wide variety of molecules are involved in the development of immune responses, including:
- antibodies and cytokines, produced by lymphocytes; and
- other molecules that are normally present in serum.

The serum concentration of a number of these proteins increases rapidly during infection and they are therefore called **acute phase proteins**.

One example of an acute phase protein is **C reactive protein (CRP)**, so-called because of its ability to bind to the C protein of pneumococci. This promotes the uptake of pneumococci by phagocytes, a process known as opsonization. Molecules such as antibody and CRP that promote phagocytosis are said to act as opsonins.

Another important group of molecules that can act as opsonins are components of the complement system.

Functions of complement

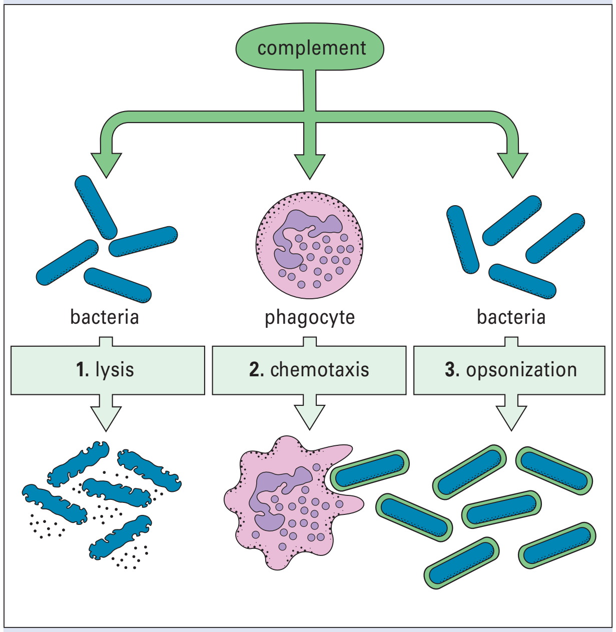

Fig. 1.3 Components of the complement system can lyse many bacterial species (**1**). Complement fragments released in this reaction attract phagocytes to the site of the reaction (**2**). Complement components opsonize the bacteria for phagocytosis (**3**). In addition to the responses shown here, activation of the complement system increases blood flow and vascular permeability at the site of activation. Activated components can also induce the release of inflammatory mediators from mast cells.

Complement proteins mediate phagocytosis, control inflammation, and interact with antibodies in immune defense

The complement system is a group of about 20 serum proteins whose overall function is the control of inflammation (Fig. 1.3 and see Chapter 4). The components interact with each other, and with other elements of the immune system. For example:
- a number of microorganisms spontaneously activate the complement system, via the so-called '**alternative pathway**', which is an innate immune defense – this results in the microorganism being opsonized (i.e. coated by complement molecules, leading to its uptake by phagocytes);
- the complement system can also be activated by antibodies or by mannose binding lectin bound to the pathogen surface via the '**classical pathway**'.

Complement activation is a cascade reaction, where one component acts enzymatically on the next component in the cascade to generate an enzyme, which mediates the following step in the reaction sequence, and so on. (The blood clotting system also works as an enzyme cascade.)

Activation of the complement system generates protein molecules or peptide fragments, which have the following effects:
- opsonization of microorganisms for uptake by phagocytes and eventual intracellular killing;
- attraction of phagocytes to sites of infection (chemotaxis);
- increased blood flow to the site of activation and increased permeability of capillaries to plasma molecules;
- damage to plasma membranes on cells, Gram-negative bacteria, enveloped viruses, or other organisms that have induced the activation, which in turn can result in lysis of the cell or virus and so reduce the infection;
- release of inflammatory mediators from mast cells.

Cytokines signal between lymphocytes, phagocytes, and other cells of the body

Cytokine is the general term for a large group of molecules involved in signaling between cells during immune responses. All cytokines are proteins or glycoproteins. The different cytokines fall into a number of categories, and the principal subgroups of cytokines are outlined below.

INTERFERONS LIMIT THE SPREAD OF CERTAIN VIRAL INFECTIONS – **Interferons (IFNs)** are cytokines that are particularly important in limiting the spread of certain viral infections:

- one group of interferons (IFNα and IFNβ) is produced by cells that have become infected by a virus;
- another type, IFNγ, is released by activated TH1 cells.

IFNs induce a state of antiviral resistance in uninfected cells (Fig. 1.4). They are produced very early in infection and are important in delaying the spread of a virus until such time as the adaptive immune response has developed.

INTERLEUKINS HAVE A VARIETY OF FUNCTIONS – The **interleukins (ILs)** are a large group of cytokines produced mainly by T cells, though some are also produced by mononuclear phagocytes or by tissue cells. They have a variety of functions. Many interleukins cause other cells to divide and differentiate.

COLONY STIMULATING FACTORS DIRECT THE DIVISION AND DIFFERENTIATION OF LEUKOCYTE PRECURSORS – **Colony stimulating factors (CSFs)** are cytokines primarily involved in directing the division and differentiation of bone marrow stem cells, and the precursors of blood leukocytes. The balance of different CSFs is partially responsible for the proportions of different immune cell types produced. Some CSFs also promote further differentiation of cells outside the bone marrow. For example, macrophage CSF (M-CSF) promotes the development of monocytes in bone marrow and macrophages in tissues.

CHEMOKINES DIRECT THE MOVEMENT OF LEUKOCYTES AROUND THE BODY – **Chemokines** are a large group of chemotactic cytokines that direct the movement of leukocytes around the body, from the blood stream into the tissues and to the appropriate location within each tissue. Some chemokines also activate cells to carry out particular functions.

Interferons

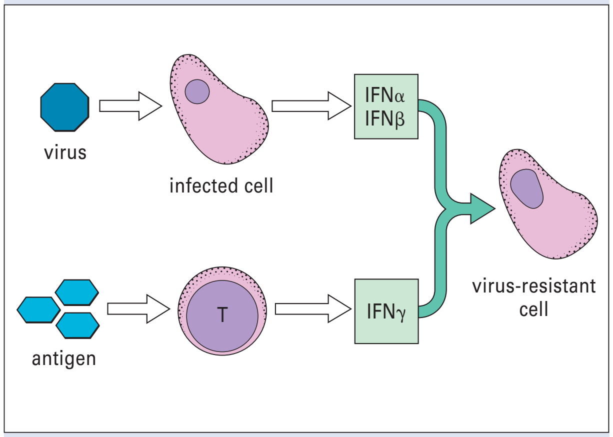

Fig. 1.4 Host cells that have been infected by virus secrete interferon-α (IFNα) and/or interferon-β (IFNβ). TH1 cells secrete interferon-γ (IFNγ) after activation by antigen. IFNs act on other host cells to induce resistance to viral infection. IFNγ has many other effects as well.

OTHER CYTOKINES INCLUDE TNFα AND TNFβ, AND TGFβ – **Tumor necrosis factors** TNFα and TNFβ, and **transforming growth factor-β** (TGFβ), have a variety of functions, but are particularly important in mediating inflammation and cytotoxic reactions.

EACH SET OF CELLS RELEASES A PARTICULAR BLEND OF CYTOKINES – Each set of cells releases a particular blend of cytokines, depending on the type of cell and whether, and how, it has been activated. For example:

- TH1 cells release one set of cytokines, which promote TH1 cell interactions with mononuclear phagocytes;
- TH2 cells release a different set of cytokines, which allow TH2 cells to activate B cells.

Some cytokines may be produced by all T cells, and some just by a specific subset.

Equally important is the expression of cytokine receptors. Only a cell that has the appropriate receptors can respond to a particular cytokine. For example:

- the receptors for interferons (see above) are present on all nucleated cells in the body;
- other receptors are much more restricted in their distribution.

In general, cytokine receptors are specific for their own individual cytokine, but this is not always so. In particular, many chemokine receptors respond to several different chemokines.

IMMUNE RESPONSES ARE ADAPTED TO DIFFERENT TYPES OF PATHOGEN

The environment contains a great variety of infectious agents, including bacteria, viruses, fungi, Protoctista, and multicellular parasites, which can enter the body and cause disease and, if they multiply unchecked, death.

The primary function of the immune system is to eliminate infectious agents and minimize the damage they cause. It ensures that most infections in normal individuals are short-lived and leave little permanent damage.

Infectious agents that cause damage are referred to as **pathogens**. However, even normally harmless organisms can pose a lethal threat to an individual who has a defect in a critical component of their immune system (see Chapters 16 and 17).

The immune system can fail to act or act inappropriately in a variety of ways, leading to immunopathological reactions. The paramount importance of immune defenses is underscored by the fact that individuals with genetic defects in various components of the immune system succumb to infections early in life (see Chapter 16).

Immune system responses vary and depend on the pathogen

Pathogens use many modes of transmission and reproduction, so the immune system has evolved many ways of responding to them.

The exterior defenses of the body (Fig. 1.5) present an effective barrier to most organisms. Very few infectious agents can penetrate intact skin. In contrast:

- many infectious agents gain access to the body across the epithelia of the gastrointestinal or urogenital tracts;

Exterior defenses

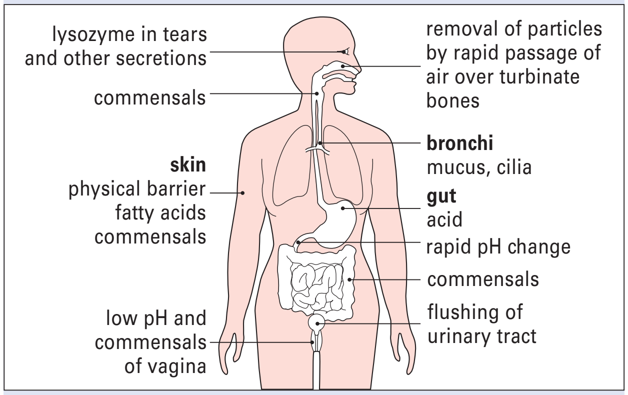
lysozyme in tears and other secretions
commensals
skin
physical barrier
fatty acids
commensals
low pH and commensals of vagina
removal of particles by rapid passage of air over turbinate bones
bronchi
mucus, cilia
gut
acid
rapid pH change
commensals
flushing of urinary tract

Fig. 1.5 Most infectious agents are prevented from entering the body by physical and biochemical barriers. The body tolerates a number of commensal organisms, which compete effectively with many potential pathogens.

- others, such as the virus responsible for the common cold, infect the respiratory epithelium of nasopharynx and lung;
- a small number of infectious agents infect the body only if they enter the blood directly (e.g. malaria and sleeping sickness).

Once inside the body, the site of the infection and the nature of the pathogen largely determine which type of immune response will be induced – most importantly (Fig. 1.6) whether the pathogen is:

Intracellular and extracellular pathogens

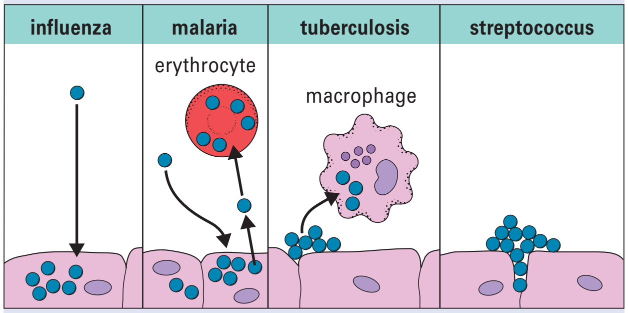

Fig. 1.6 All infectious agents spread to infect new cells by passing through the body fluids or tissues. Many are intracellular pathogens and must infect cells of the body to divide and reproduce (e.g. viruses such as influenza viruses and malaria, which has two separate phases of division, either in cells of the liver or in erythrocytes). The mycobacteria that cause tuberculosis can divide outside cells or within macrophages. Some bacteria (e.g. streptococci, which produce sore throats and wound infections) generally divide outside cells and are therefore extracellular pathogens.

- an **intracellular pathogen** (i.e. invades the host cells to divide and reproduce); or
- an **extracellular pathogen** (i.e. does not invade the host cells);

Many bacteria and larger parasites live in tissues, body fluids, or other extracellular spaces, and are susceptible to the multitude of immune defenses, such as **antibodies** (see Chapter 3) and **complement** (see Chapter 4), that are present in these areas. Because these components are present in the tissue fluids of the body (the ‘humors’ of ancient medicine), they have been classically referred to as **humoral immunity**.

Some pathogens evade ‘humoral immunity’

Many organisms (e.g. viruses, some bacteria, some parasites) evade these formidable defenses by being intracellular pathogens and replicating within host cells. To clear these infections, the immune system has developed ways to specifically recognize and destroy infected cells. This is largely the job of **cell-mediated immunity**.

Intracellular pathogens cannot, however, wholly evade the extracellular defenses (see Fig. 1.6) because they must reach their host cells by moving through the blood and tissue fluids. As a result they are susceptible to humoral immunity during this portion of their life cycle.

Any immune response involves:

- first, recognition of the pathogen or other foreign material; and
- second, a reaction to eliminate it.

Innate immune responses are the same on each encounter with antigen

Broadly speaking, immune responses fall into two categories – those that become more powerful following repeated encounters with the same antigen (**adaptive immune responses**) and those that do not become more powerful following repeated encounters with the same antigen (**innate immune responses**).

Innate immune responses (see Chapter 6) can be thought of as simple though remarkably sophisticated systems present in all animals that are the first line of defense against pathogens and allow a rapid response to invasion.

Innate immune response systems range from external barriers (skin, mucous membranes, cilia, secretions, and tissue fluids containing antimicrobial agents; see Fig. 1.5) to sophisticated receptors capable of recognizing broad classes of pathogenic organisms, for example:

- innate immune receptors on certain leukocytes recognize **pathogen-associated molecular patterns (PAMPs)**, which are common to many foreign invaders and are not normally present in the host (e.g. constituents of bacterial cell walls);
- components of the complement system (see Chapter 4), which are important soluble mediators of innate immunity, can be specifically activated by bacterial surface molecules.

The innate defenses are closely interlinked with adaptive responses.

Phagocytes internalize and kill invading organisms

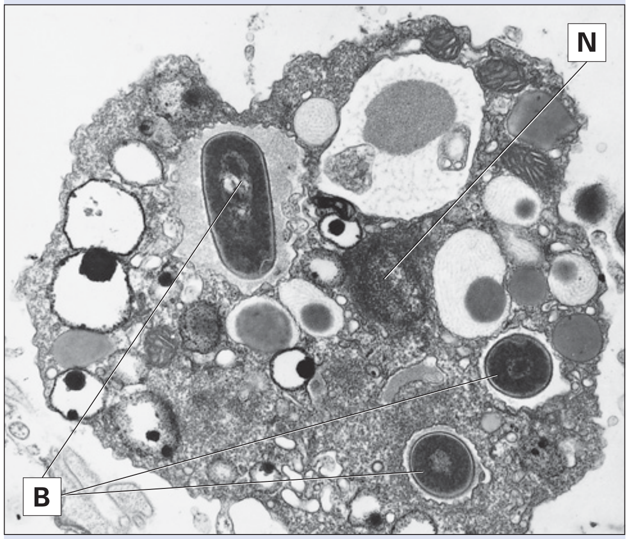

Fig. 1.7 Electron micrograph of a phagocyte from a tunicate (sea squirt) that has endocytosed three bacteria (B). (N, nucleus) (Courtesy of Dr AF Rowley)

PHAGOCYTES AND LYMPHOCYTES ARE KEY MEDIATORS OF IMMUNITY

Phagocytes are a first line of defense against infection

Phagocytes are found in all animals and are cells that ingest foreign substances. They form an important part of the innate immune system and include monocytes, macrophages, and neutrophils.

Phagocytes bind to invading organisms using a variety of receptors that recognize PAMPs and then internalize and kill the organism (Fig. 1.7). **Phagocytosis** describes the internalization (endocytosis) of large particles or microbes.

The primitive responses of phagocytes are highly effective, and people with genetic defects in certain phagocytic cells usually succumb to infections in infancy.

Lymphocytes have specialized functions

Adaptive immune responses are mediated by a specialized group of leukocytes, the **lymphocytes**, which includes T and B lymphocytes (T cells and B cells). These two classes of lymphocytes carry out very different protective functions:

- **B cells** are responsible for the humoral arm of the adaptive immune system, and thus act against extracellular pathogens;
- **T cells** are responsible for the cell-mediated arm of the adaptive immune system, and are mainly concerned with cellular immune responses to intracellular pathogens, such as viruses, and also with regulation of B cell responses.

B cells express their antigen receptor (**immunoglobulin** molecules) on their cell surface during their development and, when mature, secrete soluble immunoglobulin molecules (also known as antibodies) into the extracellular fluids.

Immunoglobulins are an essential component of humoral immunity, and, when bound to their cognate antigens, can activate the complement system and help phagocytes to take up antigens.

T cells recognize antigen via cell surface receptors (**T cell receptors**) and have a wide range of biological functions (Fig. 1.8):

- some are involved in the control of B cell development and antibody production;
- another group of T cells interacts with phagocytic cells to help them destroy pathogens they have taken up;
- a third set of T cells recognizes cells infected by virus and destroys them.

Functions of different types of lymphocyte

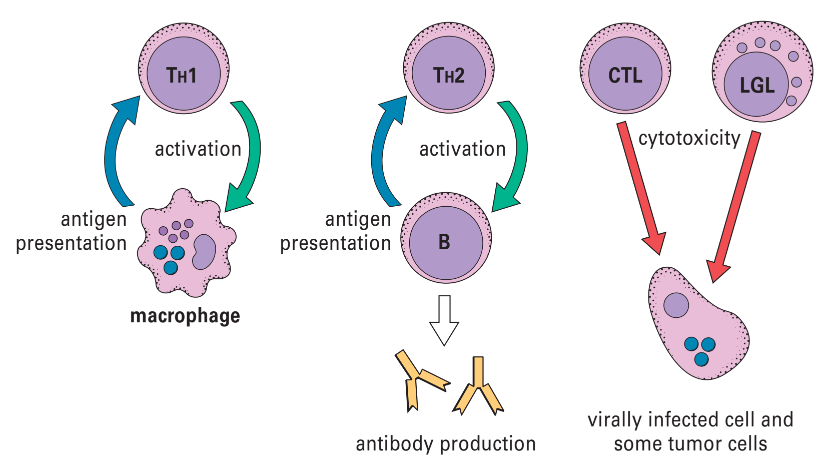

Fig. 1.8 Macrophages present antigen to TH1 cells, which then activate the macrophages to destroy phagocytosed pathogens. B cells present antigen to TH2 cells, which activate the B cells, causing them to divide and differentiate. Cytotoxic T lymphocytes (CTLs) and large granular lymphocytes (LGLs) recognize and destroy virally infected cells.

SPECIFICITY AND MEMORY ARE KEY FEATURES OF THE ADAPTIVE IMMUNE RESPONSE

In contrast to the innate immune response, which recognizes common molecular patterns such as PAMPs (see p. 140) in potential invaders, the adaptive immune system takes a highly discriminatory approach, with a very large repertoire of specific antigen receptors that can recognize virtually any component of a foreign invader (see Chapters 3 and 5). This use of highly specific antigen receptor molecules provides the following advantages:

- pathogens that lack stereotypical patterns (which might avoid recognition by the innate immune system) can be recognized;
- responses can be highly specific for a given pathogen;
- the **specificity** of the response allows the generation of immunological memory – related to its use of highly individual antigen receptors, the adaptive immune system has the capacity to 'remember' a pathogen.

These features underlie the phenomenon of specific immunity (e.g. diseases such as measles and diphtheria induce adaptive immune responses that generate life-long immunity following an infection).

Specific immunity can, very often, be induced by artificial means, allowing the development of vaccines (see Chapter 18).

ANTIGENS ARE MOLECULES RECOGNIZED BY RECEPTORS ON LYMPHOCYTES

Originally the term **antigen** was used for any molecule that induced B cells to produce a specific antibody (*anti*body *gen*erator). This term is now more widely used to indicate molecules that are specifically recognized by antigen receptors of either B cells or T cells.

Antigens, defined broadly, are molecules that initiate adaptive immune responses (e.g. components of pathogenic organisms), though purists may prefer the term **immunogen** in this context.

Antigens are not just components of foreign substances, such as pathogens. A large variety of 'self' molecules can serve as antigens as well, provoking autoimmune responses that can be highly damaging, and even lethal (see Chapter 20).

Antigens are the initiators and driving forces of all adaptive immune responses

The immune system has evolved to recognize antigens, destroy them, and eliminate the source of their production (e.g. bacteria, virally infected cells). When antigen is eliminated, immune responses switch off.

Both T cell receptors and immunoglobulin molecules (antibody) bind to their cognate antigens with a high degree of specificity. These two types of receptor molecules have striking structural relationships and are closely related evolutionarily, but bind to very different types of antigens and carry out quite different biological functions.

Antibody specifically binds to antigen

Soluble antibodies (also called immunoglobulins) are a group of serum molecules closely related to and derived from the antigen receptors on B cells.

All antibodies have the same basic Y-shaped structure, with two regions (variable regions) at the tips of the Y that bind to antigen. The stem of the Y is referred to as the constant region and is not involved in antigen binding (see Chapter 3).

The two variable regions contain identical antigen-binding sites that, in general, are specific for only one type of antigen. The amino acid sequences of the variable regions of different antibodies, however, are extremely variable. The antibody molecules in the body therefore provide an extremely large repertoire of antigen-binding sites. The way in which this great diversity of antibody variable regions is generated is explained in Chapter 3.

Each antibody binds to a restricted part of the antigen called an epitope

Pathogens typically have many different antigens on their surface. Each antibody binds to an **epitope**, which is a restricted part of the antigen. A particular antigen can have several different epitopes or repeated epitopes (Fig. 1.9). Antibodies are specific for the epitopes rather than the whole antigen molecule.

Q. Many evolutionarily related proteins have conserved amino acid sequences. What consequences might this have in terms of the antigenicity of these proteins?

A. Related proteins (with a high degree of sequence similarity) may contain the same epitopes and therefore be recognized by the same antibodies.

Fc regions of antibodies act as adapters to link phagocytes to pathogens

The constant region of the antibody (the Fc region, see Chapter 3) can bind to Fc receptors on phagocytes, so acting as an adapter between the phagocyte and the

Antigens and epitopes

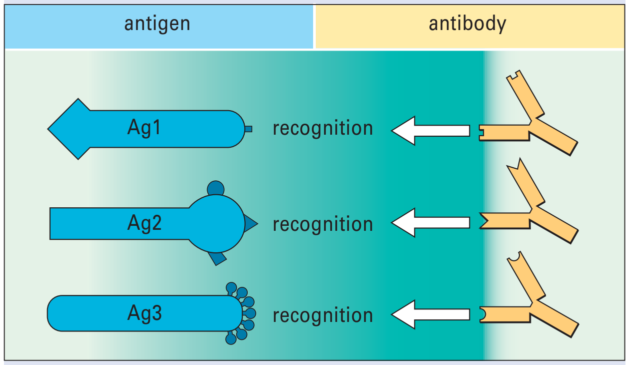

Fig. 1.9 Antibodies recognize molecular shapes (epitopes) on the surface of antigens. Each antigen (Ag1, Ag2, Ag3) may have several epitopes recognized by different antibodies. Some antigens have repeated epitopes (Ag3).

Antibody acts as an adapter that links a microbe to a phagocyte

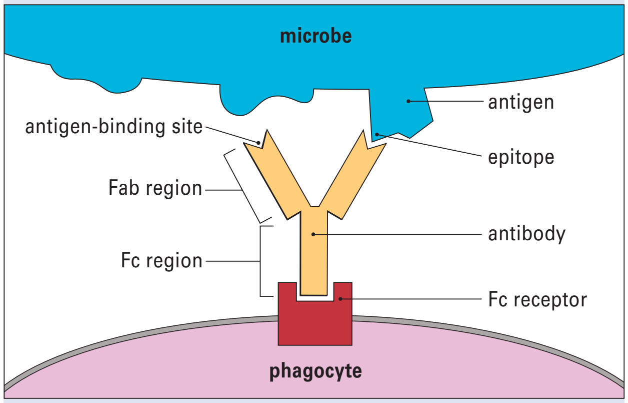

Fig. 1.10 The antibody binds to a region of an antigen (an epitope) on the microbe surface, using one of its antigen-binding sites. These sites are in the Fab regions of the antibody. The stem of the antibody, the Fc region, can attach to receptors on the surface of the phagocytes.

Opsonization

	phagocyte	opsonin	binding
1	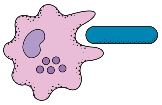	–	±
2	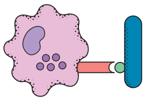	complement C3b	+ +
3	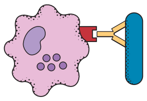	antibody	+ +
4	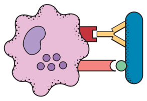	antibody and complement C3b	+ + + +

Fig. 1.11 Phagocytes have some intrinsic ability to bind to bacteria and other microorganisms (**1**). This is much enhanced if the bacteria have been opsonized by complement C3b (**2**) or antibody (**3**), each of which cross-link the bacteria to receptors on the phagocyte. Antibody can also activate complement, and if antibody and C3b both opsonize the bacteria, binding is greatly enhanced (**4**).

pathogen (Fig. 1.10). Consequently, if antibody binds to a pathogen, it can link to a phagocyte and promote phagocytosis.

The process in which specific binding of an antibody activates an innate immune defense (phagocytosis) is an important example of collaboration between the innate and adaptive immune responses.

Other molecules (such as activated complement proteins) can also enhance phagocytosis when bound to microbial surfaces. The general process is termed **opsonization** (from the Latin, opsono, 'to prepare victuals for') to "render microorganisms delectable so that the white blood corpuscles pitch into them with an appetite" (GB Shaw).

Binding and phagocytosis are most effective when more than one type of adapter molecule (**opsonin**) is present (Fig. 1.11). Note that antibody can act as an adapter in many other circumstances, not just phagocytosis.

Peptides from intracellular pathogens are displayed on the surface of infected cells

Antibodies patrol only extracellular spaces and so only recognize and target for destruction extracellular pathogens. Intracellular pathogens (such as viruses) can escape antibody-mediated responses once they are safely ensconced within a host cell. The adaptive immune system has therefore evolved a specific method of displaying portions of virtually all cell proteins on the surface of each nucleated cell in the body so they can be recognized by T cells.

For example, a cell infected with a virus will present fragments of viral proteins (peptides) on its surface that are recognizable by T cells. The antigenic peptides are transported to the cell surface and presented to the T cells by **MHC molecules** (see Chapter 5). T cells use their antigen-specific receptors (T cell receptors – TCRs) to recognize the antigenic peptide–MHC molecule complex (Fig. 1.12).

Q. Why is it necessary to have a mechanism that transports antigen fragments to the host cell surface for cytotoxic T cells to recognize infected cells?

A. The T cell cannot 'see' what is going on inside an infected cell. Its antigen receptor can only interact with and recognize what is present on the surface of cells. Therefore antigen fragments need to be transported to the cell surface for recognition, and this is the key function of MHC molecules.

T cell responses require proper presentation of antigen by MHC molecules (**antigen presentation**). To activate T cell responses this must occur on the surface of specialized **antigen-presenting cells (APCs)**, which internalize antigens by phagocytosis or endocytosis. Several different types of leukocyte can act as APCs, including dendritic cells, macrophages, and B cells.

APCs not only display antigenic peptide–MHC complexes on their surface, but also express co-stimulatory molecules that are essential for initiating immune responses (see Chapter 7).

Co-stimulatory signals are upregulated by the presence of pathogens, which can be detected by the engagement of innate immune receptors that recognize PAMPs.

T cell recognition of antigen

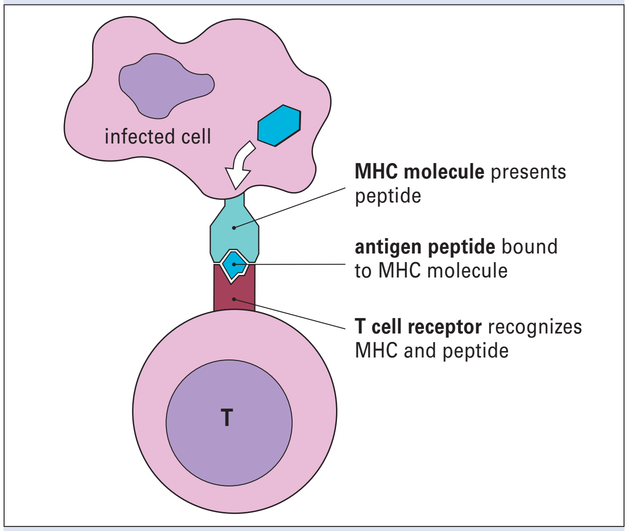
infected cell
MHC molecule presents peptide
antigen peptide bound to MHC molecule
T cell receptor recognizes MHC and peptide
T

Fig. 1.12 MHC molecules transport peptides to the surface of an infected cell where they are presented to T cells, which may recognize the MHC–peptide combination. If a cell is infected, MHC molecules present peptides derived from the pathogen, as well as the cell's own proteins.

T cells comprise the cell-mediated arm of the adaptive immune system

T cells recognize antigens only in the context of cell–cell interactions. A T cell must dock with an APC that bears the correct antigenic peptide presented by an MHC molecule. T cells therefore comprise the cell-mediated arm of the adaptive immune system. T cells function only in the context of intimate cell–cell contacts, which allow either:

- helper functions, such as control of antibody responses via contacts with APCs and B cells; or
- cytolytic functions, that is the recognition and lysis of virally infected cells requiring cell–cell contact.

These functions are carried out by different types of T cell, namely **helper T cells (TH)** and **cytotoxic T lymphocytes (CTLs or Tc cells)**.

ANTIGEN RECOGNITION AND ANTIGEN ERADICATION – THE TWO PHASES OF IMMUNE RESPONSES

Most immune responses to infectious organisms are made up of a variety of innate and adaptive components:

- in the earliest stages of infection, innate responses predominate;
- later the lymphocytes start to generate adaptive immune responses;
- after recovery, immunological memory remains within the population of lymphocytes, which can then mount a more effective and rapid response if there is a reinfection with the same pathogen at a later date.

The two major phases of any immune response are antigen recognition and a reaction to eradicate the antigen.

Lymphocytes are responsible for antigen recognition

In adaptive immune responses, lymphocytes are responsible for immune recognition, and this is achieved by **clonal selection**.

Antigen selects and activates specific clones of lymphocyte

Each lymphocyte is genetically programmed to be capable of recognizing just one particular antigen. However, the immune system as a whole can specifically recognize many thousands of antigens, so the lymphocytes that recognize any particular antigen are only a tiny proportion of the total.

How then is an adequate immune response to an infectious agent generated? The answer is that, when an antigen binds to the few lymphocytes that can recognize it, they are induced to proliferate rapidly. Within a few days there is a sufficient number to mount an adequate immune response. In other words, the antigen selects and activates the specific clones to which it binds (Fig. 1.13), a process called clonal selection. This operates for both B cells and T cells.

B cell clonal selection

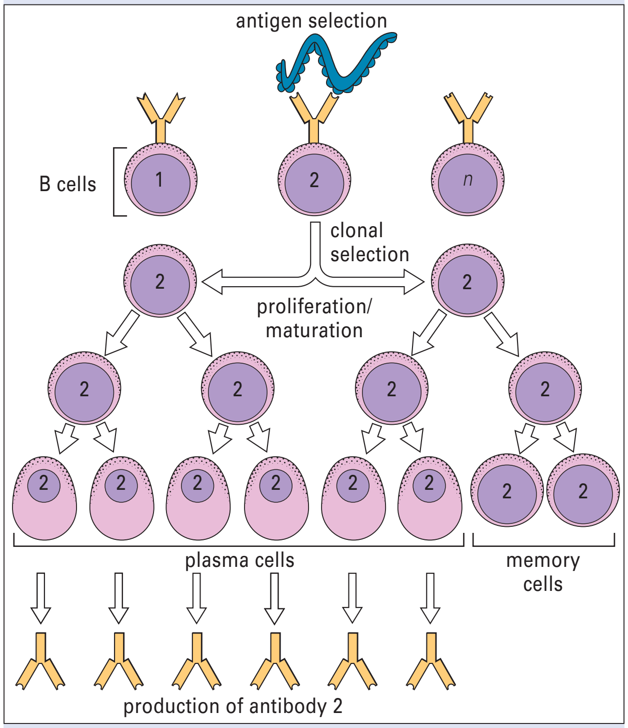

Fig. 1.13 Each B cell expresses just one antibody (i.e. with specificity for a single particular antigen), which it uses as its antigen receptor. Antigen binds only to those B cells with the specific antibody (number 2 in this example), driving these cells to divide and differentiate into plasma cells and memory cells, all having the same specificity as the original B cell. Thus an antigen selects just the clones of B cells that can react against it.

How can the immune system ‘know’ which specific antibodies will be needed during an individual’s lifetime? It does not know. The immune system generates antibodies (and T cell receptors) that can recognize an enormous range of antigens even before it encounters them. Many of these specificities, which are generated more or less at random (see Chapters 3 and 5), will never be called upon to protect the individual against infection.

Q. What advantage could there be in having an immune system that generates billions of lymphocytes that do not recognize any known infectious agent?

A. Many pathogens mutate their surface antigens. If the immune system could not recognize new variants of pathogens, it would not be able to make an effective response. By having a wide range of antigen receptors, at least some of the lymphocytes will be able to recognize any pathogen that enters the body.

Lymphocytes that have been stimulated, by binding to their specific antigen, take the first steps towards cell division. They:

- express new receptors that allow them to respond to cytokines from other cells, which signal proliferation;
- may start to secrete cytokines themselves;
- will usually go through a number of cycles of division before differentiating into mature cells, again under the influence of cytokines.

For example, proliferating B cells eventually mature into antibody-producing **plasma cells**.

Even when the infection has been overcome, some of the newly produced lymphocytes remain, available for restimulation if the antigen is ever encountered again. These cells are called **memory cells**, because they are generated by past encounters with particular antigens. Memory cells confer lasting immunity to a particular pathogen.

Antigen eradication involves effector systems

There are numerous ways in which the immune system can destroy pathogens, each being suited to a given type of infection at a particular stage of its life cycle. These defense mechanisms are often referred to as effector systems.

ANTIBODY BINDING RESULTING IN NEUTRALIZATION – In one of the simplest effector systems, antibodies can combat certain pathogens just by binding to them. For example, antibody to the outer coat proteins of some rhinoviruses (which cause colds) can prevent the viral particles from binding to and infecting host cells.

PHAGOCYTOSIS IS PROMOTED BY COMPLEMENT AND OPSONINS – More often antibody activates complement or acts as an opsonin to promote ingestion by phagocytes.

Phagocytes that have bound to an opsonized microbe, engulf it by extending pseudopodia around it. These fuse and the microorganism is internalized (endocytosed) in a phagosome. Granules and lysosomes fuse with the phagosome, pouring enzymes into the resulting phagolysosome, to digest the contents (Fig. 1.14). The mechanisms involved are described fully in Chapters 9 and 15.

Phagocytes have several ways of dealing with internalized opsonized microbes in phagosomes. For example:

- macrophages reduce molecular oxygen to form microbicidal reactive oxygen intermediates (ROIs), which are secreted into the phagosome;
- neutrophils contain lactoferrin, which chelates iron and prevents some bacteria from obtaining this vital nutrient.

CYTOTOXIC REACTIONS ARE DIRECTED AGAINST WHOLE CELLS – Cytotoxic reactions are effector systems directed against whole cells that are in general too large for phagocytosis.

The target cell may be recognized either by:

- specific antibody bound to the cell surface; or
- T cells using their specific TCRs.

In cytotoxic reactions the attacking cells direct their granules towards the target cell (in contrast to phagocytosis where the contents are directed into the phagosome). As a result granules are discharged into the extracellular space close to the target cell.

The granules of CTLs contain molecules called **perforins**, which can punch holes in the outer membrane of the target (see Fig. 10.12). (In a similar way, antibody bound to the surface of a target cell can direct complement to make holes in the cell’s plasma membrane.)

Some cytotoxic cells can signal to the target cell to initiate programmed cell death – a process called **apoptosis**.

Q. What risks are associated with discharging granule contents into the extracellular space?

A. Cells other than the target cell may be damaged. This is minimized by close intercellular contact between the CTL and the target cell.

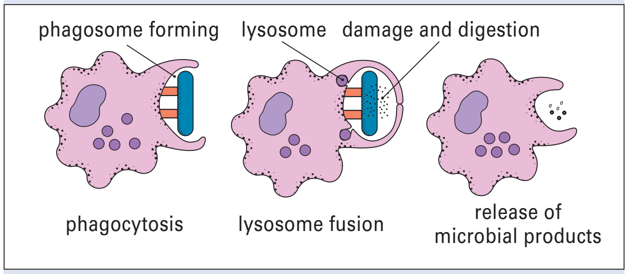

Phagocytosis

Fig. 1.14 Phagocytes attach to microorganisms using cell surface receptors for microbial products or via antibody or complement C3b. Pseudopods extend around the microorganism and fuse to form a phagosome. Killing mechanisms are activated and lysosomes fuse with the phagosomes, releasing digestive enzymes that break down the microbe. Undigested microbial products may be released to the outside.

Termination of immune responses to avoid damaging host tissues

Although it is important to initiate immune responses quickly, it is also critical to terminate them appropriately once the threat has ended.

To clear the offending pathogen immune responses are often massive, with:

- millions of activated lymphocytes;
- proliferation of large clones of specific T and B cells;
- activation of huge numbers of inflammatory cells.

These responses, if left unchecked, can also damage host tissues.

A number of mechanisms are employed to dampen or terminate immune responses. One is a passive process – that is, simple clearance of antigen should lead to a diminution of immune responses.

Q. Why would removal of antigen lead to the decline in an immune response?

A. Antigen is required to stimulate B cell proliferation and differentiation, with the consequent production of antibody. Antigen combined with antibody activates several effector systems (e.g. complement). Antigen is also required to stimulate T cells with consequent production of cytokines. Therefore removal of antigen takes away the primary stimulus for lymphocyte activation.

Antigen elimination can be a slow process, however, so the immune system also employs a variety of active mechanisms to downregulate responses, as discussed in Chapter 11.

The innate and adaptive arms of the immune system interact and collaborate

Many examples of the intimate interconnections between the innate and adaptive arms of the immune defense network (some of which are outlined below) are explored in Chapters 3, 4, 6, and 9.

A key regulatory control for initiating immune responses

The recognition of PAMPs by innate immune receptors provides a key regulatory control for initiating immune responses (see Chapter 6). The innate immune system:

- senses 'danger' (e.g. by detecting the presence of bacterial cell wall material); and
- upregulates co-stimulatory molecules on the surface of APCs, allowing them to become potent activators of T cell (and, indirectly, B cell) responses.

This is a good example of collaboration between the innate and adaptive arms of the immune system.

Adaptive responses can augment innate immune defenses

In the example of antigen presentation given above (i.e. the presentation of bacterial antigens), the activated T cells release soluble signaling molecules (**cytokines**), which activate the phagocytes and cause them to destroy the pathogens they have internalized. Thus, the collaboration between T cells and APCs can be viewed as a two-way system (see Chapter 7).

Q. Can you think of another example where soluble factors from lymphocytes help in the destruction of pathogens?

A. Opsonization of pathogens by antibodies is another classic example of collaboration between innate and adaptive immune defenses.

Antibody can activate the complement system

Antibodies bound to bacterial surfaces can also stimulate activation of the complement system – an innate defense mechanism that enhances phagocytosis and drills holes in the outer phospholipid membrane of some types of bacteria, leading to their lysis (see Chapter 4).

Immune responses to extracellular and intracellular pathogens differ

In dealing with extracellular pathogens, the immune system aims to destroy the pathogen itself and neutralize its products.

In dealing with intracellular pathogens, the immune system has two options:

- T cells can destroy the infected cell (i.e. cytotoxicity); or
- T cells can activate the infected cell to deal with the pathogen itself (e.g. helper T cells release cytokines, which activate macrophages to destroy the organisms they have internalized).

Because many pathogens have both intracellular and extracellular phases of infection, different mechanisms are usually effective at different times. For example, the polio virus travels from the gut, through the blood stream to infect nerve cells in the spinal cord. Antibody is particularly effective at blocking the early phase of infection while the virus is in the blood stream, but to clear an established infection CTLs must kill any cell that has become infected.

Consequently, antibody is important in limiting the spread of infection and preventing reinfection with the same virus, while CTLs are essential to deal with infected cells (Fig. 1.15). These factors play an important part in the development of effective vaccines.

VACCINATION DEPENDS ON THE SPECIFICITY AND MEMORY OF ADAPTIVE IMMUNITY

The study of immunology has had its most successful application in vaccination (see Chapter 18), which is based on the key elements of adaptive immunity, namely specificity and memory. Memory cells allow the immune system to mount a much stronger response on a second encounter with antigen. Compared with the primary response, the secondary response is:

- faster to appear;
- more effective.

The aim in vaccine development is to alter a pathogen or its toxins in such a way that they become innocuous without losing antigenicity. This is possible because antibodies and T cells recognize particular parts of antigens (the **epitopes**), and not the whole organism or toxin.

Take, for example, vaccination against tetanus. The tetanus bacterium produces a toxin that acts on receptors

Reaction to extracellular and intracellular pathogens

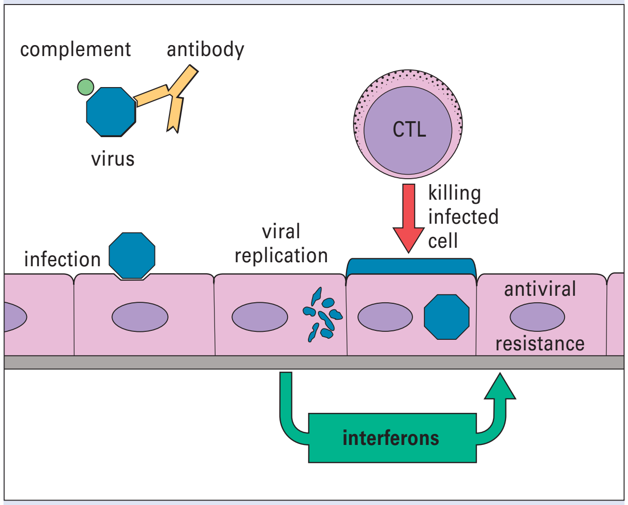

Fig. 1.15 Different immunological systems are effective against different types of infection, here illustrated as a virus infection. Antibodies and complement can block the extracellular phase of the life cycle and promote phagocytosis of the virus. Interferons produced by infected cells signal to uninfected cells to induce a state of antiviral resistance. Viruses can multiply only within living cells; cytotoxic T lymphocytes (CTLs) recognize and destroy the infected cells.

to cause tetanic contractions of muscle. The toxin can be modified by formalin treatment so that it retains its epitopes, but loses its toxicity. The resulting molecule (known as a toxoid) is used as a vaccine (Fig. 1.16).

Whole infectious agents, such as the poliovirus, can be attenuated so they retain their antigenicity, but lose their pathogenicity.

INFLAMMATION IS A RESPONSE TO TISSUE DAMAGE

Tissue damage caused by physical agents (e.g. trauma or radiation) or by pathogens results in the tissue response of **inflammation** (see p. 128), which has three principal components:

- increased blood supply to the infected area;
- increased capillary permeability due to retraction of the endothelial cells lining the vessels, permitting larger molecules than usual to escape from the capillaries;
- migration of leukocytes out of the venules into the surrounding tissues – in the earliest stages of inflammation, neutrophils are particularly prevalent, but in later stages monocytes and lymphocytes also migrate towards the site of infection or damage.

Q. What advantage could the inflammatory responses have in the defense against infection?

A. The inflammatory responses allow leukocytes, antibodies, and complement system molecules (all of which are required for the phagocytosis and destruction of pathogens) to enter the tissues at the site of infection. Lymphocytes are also required for the recognition and destruction of infected cells in the tissues.

Leukocytes enter inflamed tissue by crossing venular endothelium

The process of leukocyte migration is controlled by **chemokines** (a particular class of cytokines) on the surface of venular endothelium in inflamed tissues. Chemokines activate the circulating leukocytes causing them to bind to the endothelium and initiate migration across the endothelium (Fig. 1.17).

Once in the tissues, the leukocytes migrate towards the site of infection by a process of chemical attraction known as **chemotaxis**. For example, phagocytes will actively migrate up concentration gradients of certain (chemotactic) molecules.

Principle of vaccination

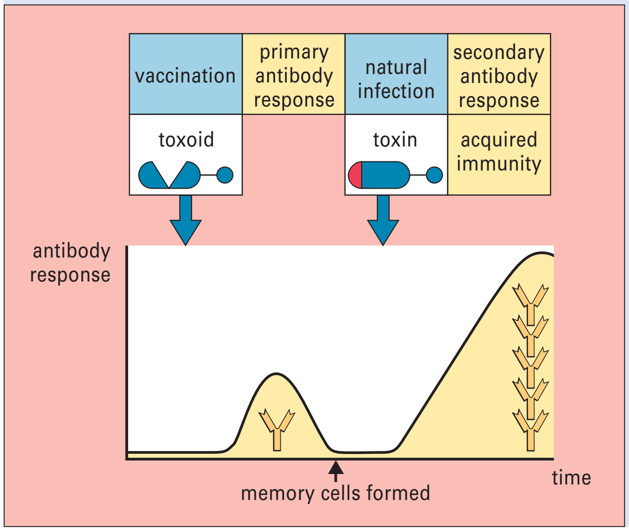

Fig. 1.16 Chemical modification of tetanus toxin produces a toxoid, which has lost its toxicity but retains many of its epitopes. A primary antibody response to these epitopes is produced following vaccination with the toxoid. If a natural infection occurs, the toxin restimulates memory B cells, which produce a faster and more intense secondary response against that epitope, so neutralizing the toxin.

Three phases in neutrophil migration across endothelium

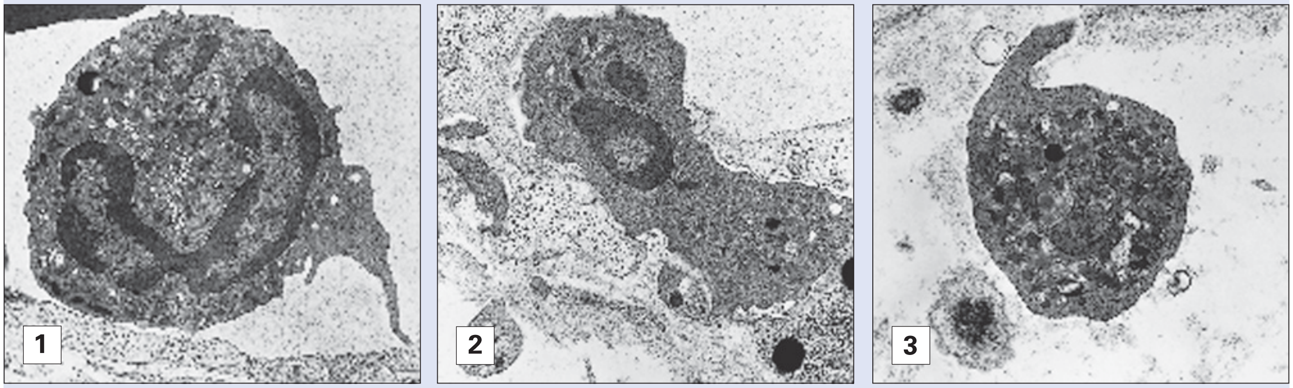

Fig. 1.17 A neutrophil adheres to the endothelium in a venule (**1**). It extends its pseudopodium between the endothelial cells and migrates towards the basement membrane (**2**). After the neutrophil has crossed into the tissue, the endothelium reseals behind (**3**). The entire process is referred to as diapedesis. (Courtesy of Dr I Jovis)

A particularly active chemotactic molecule is **C5a**, which is a fragment of one of the complement components (Fig. 1.18) and attracts both neutrophils and monocytes. When purified C5a is applied to the base of a blister in vivo, neutrophils can be seen sticking to the endothelium of nearby venules shortly afterwards. The cells then squeeze between the endothelial cells and move through the basement membrane of the microvessels to reach the tissues. This process is described more fully in Chapter 6.

THE IMMUNE SYSTEM MAY FAIL (IMMUNOPATHOLOGY)

Strong evolutionary pressure from infectious microbes has led to the development of the immune system in its

Chemotaxis

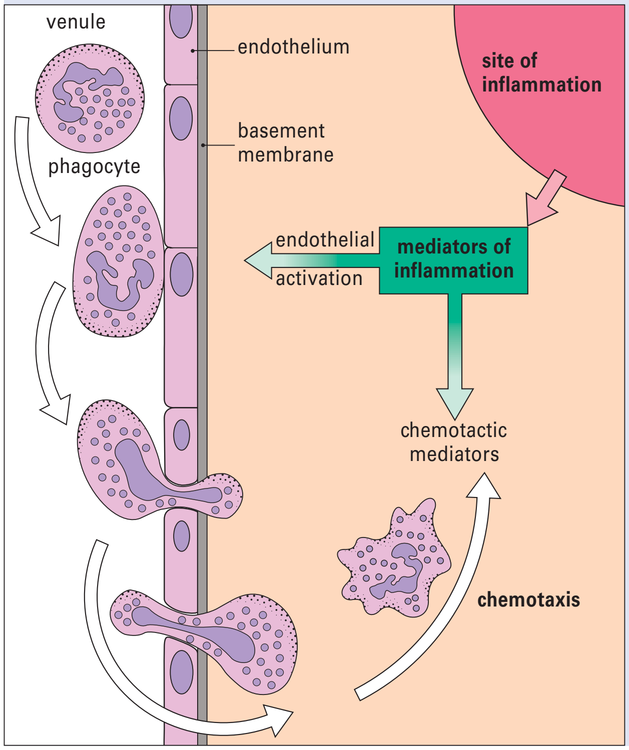
venule
phagocyte
endothelium
basement membrane
site of inflammation
endothelial activation
mediators of inflammation
chemotactic mediators
chemotaxis

Fig. 1.18 At a site of inflammation, tissue damage and complement activation cause the release of chemotactic peptides (e.g. chemokines and C5a), which diffuse to the adjoining venules and signal to circulating phagocytes. Activated cells migrate across the vessel wall and move up a concentration gradient of chemotactic molecules towards the site of inflammation.

Failure of the immune system

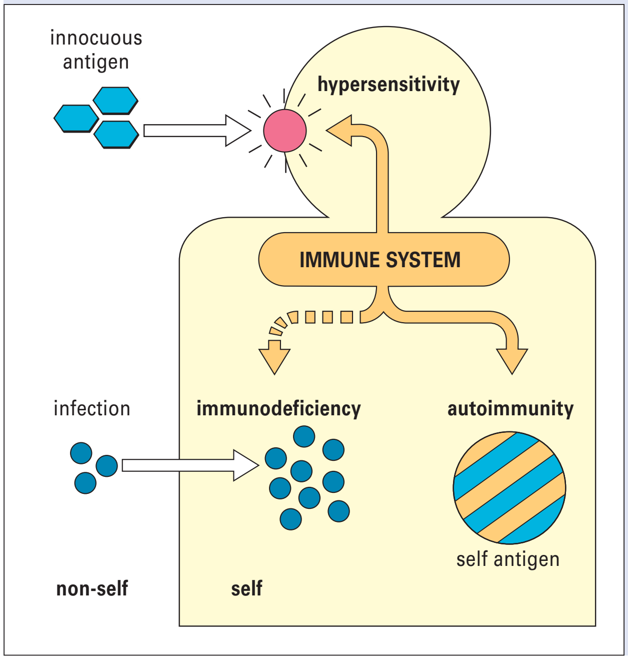

Fig. 1.19 The three principal ways in which the immune system can fail result in hypersensitivity (an overactive immune response to an antigen), immunodeficiency (an ineffective immune response to an infection), and autoimmunity (the immune system reacts against the body's own tissues).

present form. Deficiencies in any part of the system leave the individual exposed to a greater risk of infection, but other parts of the system may partly compensate for such deficiencies. However, there are occasions when the immune system is itself a cause of disease or other undesirable consequences.

In essence the immune system can fail in one of three ways (Fig. 1.19), resulting in autoimmunity, immunodeficiency, or hypersensitivity.

Inappropriate reaction to self antigens – autoimmunity

Normally the immune system recognizes all foreign antigens and reacts against them, while recognizing the body's own tissues as 'self' and making no reaction against them. The mechanisms by which this discrimination between 'self' and 'non-self' is established are described in Chapter 19.

When the immune system reacts against 'self' components, the result is an **autoimmune disease** (see Chapter 20), for example:

- rheumatoid arthritis (see p. 377);
- pernicious anemia (see p. 376).

Ineffective immune response – immunodeficiency

If any elements of the immune system are defective, the individual may not be able to fight infections adequately, resulting in **immunodeficiency** (see Chapters 16 and 17). Some immunodeficiency conditions:

- are hereditary and start to manifest shortly after birth;
- develop later in life, for example the acquired immune deficiency syndrome (AIDS).

Overactive immune response – hypersensitivity

Sometimes immune reactions are out of all proportion to the damage that may be caused by a pathogen. The immune system may also mount a reaction to a harmless antigen, such as a food molecule. Such immune reactions (**hypersensitivity**) may cause more damage than the pathogen or antigen (see Chapters 23–26). For example, molecules on the surface of pollen grains are recognized as antigens by particular individuals, leading to the symptoms of hay fever or asthma.

NORMAL IMMUNE RESPONSES CAN BE INCONVENIENT IN MODERN MEDICINE

Q. Can you think of an instance where an individual is treated to suppress an immune reaction that does not fall into one of the three categories of immunopathology described above?

A. Immunosuppression to prevent graft rejection.

The most important examples of normal immune reactions that are inconvenient in the context of modern medicine are:

- blood transfusion reactions (see Chapter 24);
- graft rejection (see Chapter 21).

In these cases it is necessary to carefully match the donor and recipient tissues so that the immune system of the recipient does not attack the donated blood or graft tissue.

Critical thinking: Specificity and memory in vaccination (see p. 493 for explanations)

The recommended schedules for vaccination against different diseases are strikingly different. Two examples are given in the table. For tetanus, the vaccine is a modified form of the toxin released by the tetanus bacterium. The vaccine for influenza is either an attenuated non-pathogenic variant of the virus, given intranasally, or a killed preparation of virus, given intradermally. Both vaccines induce antibodies that are specific for the inducing antigen.

Schedules for vaccination against tetanus and influenza A

pathogen	type of vaccine	recommended for	vaccination	effectiveness (%)
tetanus	toxoid	everyone	every 10 years	100
influenza A	attenuated virus	health workers and older people	annually	variable, 0–90

1 Why is it necessary to vaccinate against tetanus only every 10 years, though antibodies against the toxoid disappear from the circulation within a year?

2 Why is the vaccine against tetanus always effective, whereas the vaccine against influenza protects on some occasions but not others?

3 Why is tetanus recommended for everyone and influenza for only a restricted group of 'at-risk' individuals, even though influenza is a much more common disease than tetanus?

CHAPTER

2

Cells, Tissues, and Organs of the Immune System

SUMMARY

- **Most cells of the immune system** derive from hemopoietic stem cells.
- **Phagocytic cells** are found in the circulation (monocytes and granulocytes) and reside in tissues (e.g. Kupffer cells in the liver).
- **Development and differentiation of different cell lineages** depend on cell interactions and cytokines.
- **Each cell type expresses characteristic surface molecules (markers)**, which identify them.
- **Eosinophils, basophils, mast cells, and platelets take part in the inflammatory response**.
- **NK cells recognize and kill virus-infected cells and certain tumor cells** through apoptosis.
- **Antigen-presenting cells** link the innate and adaptive immune systems and are required by T cells to enable them to respond to antigens.
- **Lymphocytes are heterogeneous** phenotypically, functionally, and morphologically.
- **B and T cells express antigen receptors**, which are required for the antigen recognition.
- **There are two major subpopulations of T cells**, which have helper and cytotoxic activities.
- **Cell surface immunoglobulin and signaling molecules** form the 'B cell receptor' complex.
- **B cells can differentiate into antibody-secreting plasma cells** following activation.
- **Lymphoid organs and tissues** are either primary (central) or secondary (peripheral).
- **Lymphoid stem cells develop and mature within the primary lymphoid organs** – the thymus and bone marrow – this process is called lymphopoiesis.
- **T cells developing in the thymus** are subject to positive and negative selection processes.
- **The diverse antigen repertoires found in mature animals** are generated during lymphopoiesis by recombination of gene segments encoding the T cell receptor (TCR) and immunoglobulin.
- **Mammalian B cells develop mainly in the fetal liver and from birth onwards in the bone marrow**. This process continues throughout life. B cells also undergo a selection process at the site of B cell generation.
- **Lymphocytes** migrate to, and function in, the secondary lymphoid organs and tissues.
- **Lymphoid organs and tissues protect different body sites** – the spleen responds to blood-borne antigens; the lymph nodes respond to lymph-borne antigens; and MALT protects the mucosal surface.
- **Systemic lymphoid organs** include the spleen and lymph nodes.
- **Mucosa-associated lymphoid tissue (MALT)** includes all the lymphoid tissues associated with mucosae. Peyer's patches are a major site of lymphocyte priming to antigens crossing mucosal surfaces of the small intestine.
- **Most lymphocytes recirculate around the body**; there is continuous lymphocyte traffic from the blood stream into lymphoid tissues and back again into the blood via the thoracic duct and right lymphatic duct.

MOST CELLS OF THE IMMUNE SYSTEM DERIVE FROM HEMATOPOIETIC STEM CELLS

There is great heterogeneity in the cells of the immune system, most of which originate from hematopoietic stem cells in the fetal liver and in the postnatal bone marrow.

Innate immune system cells are monocytes/macrophages, polymorphonuclear granulocytes, NK cells, mast cells, and platelets

Phagocytic cells of the innate immune system include:

- the monocytes/macrophages;

- the **polymorphonuclear granulocytes** (polymorphonuclear neutrophils [PMNs], basophils, and eosinophils), which are mainly involved in dealing with extracellular microbes; and
- natural killer (NK) cells, which are responsible for killing virus-infected cells.

Other cells that function in innate immune protection include mast cells and platelets (Fig. 2.1).

Cells of the innate system recognize microbes through their pathogen-associated molecular patterns (PAMP) receptors. PAMP receptors have broad specificity and a non-clonal distribution (see Chapter 6).

APCs link the innate and adaptive immune systems

A specialized group of cells termed antigen-presenting cells (APCs) link the innate and adaptive immune systems by producing molecules (cytokines), which:

- enhance innate immune cell function; and
- contribute to lymphocyte function (Fig. 2.2).

The main adaptive immune system cells are lymphocytes

Lymphocytes (T and B cells) recognize antigens through clonally expressed antigen receptors (see Chapters 3 and 5). T cells are produced in the thymus (see Fig. 2.1) and require antigen to be processed and presented to them by specialized APCs.

Whereas the cells of the innate immune system are found in the blood stream and in most organs of the body, lymphocytes are localized to specialized organs and tissues.

The lymphoid organs where the lymphocytes differentiate and mature from stem cells are the **primary lymphoid organs** and include:

- the thymus – the site of T cell development;
- the fetal liver and postnatal bone marrow – the sites of B cell development.

Cells of the T and B cell lineages migrate from the primary lymphoid organs to function in the **secondary lymphoid organs** – namely:

Origin of cells of the immune system

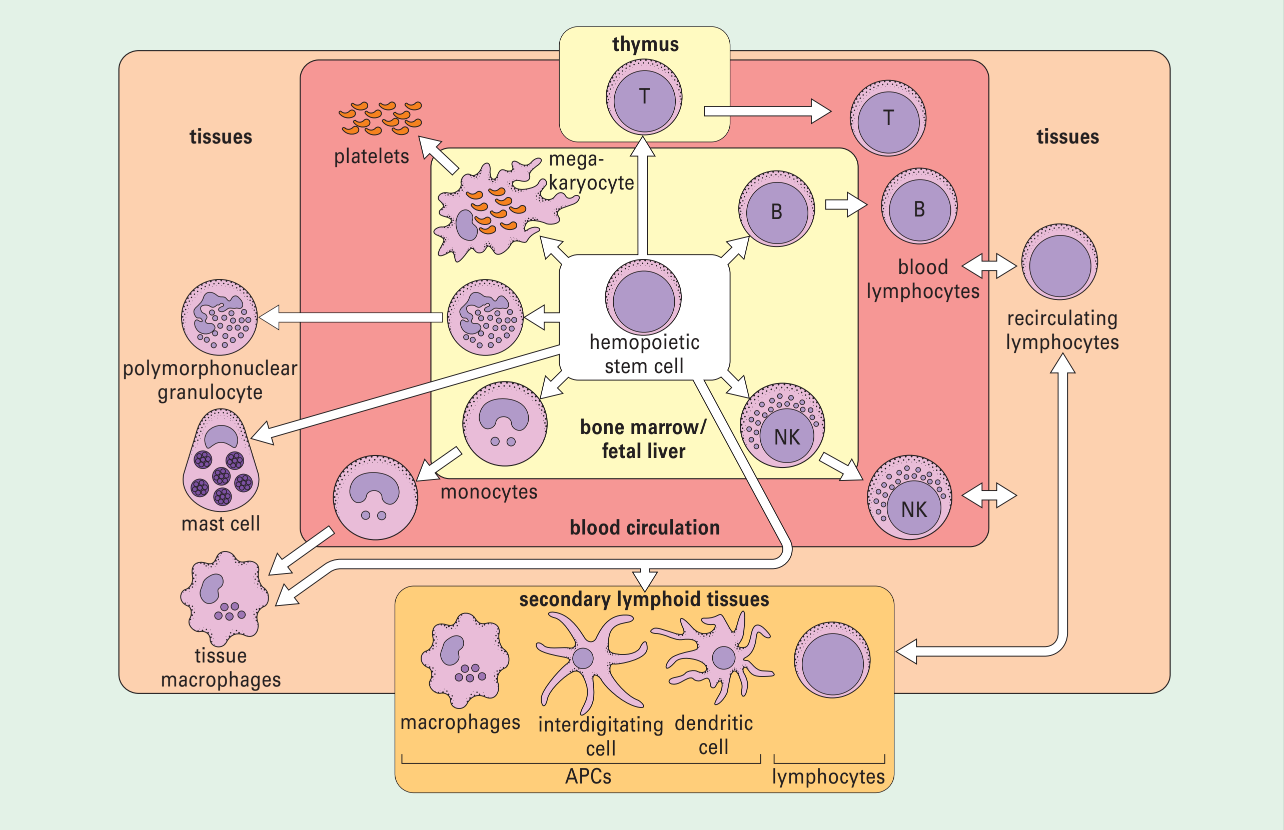
thymus
T
T
tissues
platelets
mega-
karyocyte
B
B
tissues
blood
lymphocytes
recirculating
lymphocytes
polymorphonuclear
granulocyte
hemopoietic
stem cell
bone marrow/
fetal liver
NK
mast cell
monocytes
blood circulation
NK
tissue
macrophages
secondary lymphoid tissues
macrophages
interdigitating
cell
dendritic
cell
APCs
lymphocytes

Fig. 2.1 All cells shown here arise from the hematopoietic stem cell. Platelets produced by megakaryocytes are released into the circulation. Polymorphonuclear granulocytes and monocytes pass from the circulation into the tissues. Mast cells are identifiable in all tissues. B cells mature in the fetal liver and bone marrow in mammals, whereas T cells mature in the thymus. The origin of the large granular lymphocytes with natural killer (NK) activity is probably the bone marrow. Lymphocytes recirculate through secondary lymphoid tissues. Interdigitating cells and dendritic cells act as antigen-presenting cells (APCs) in secondary lymphoid tissues.

Rapid Reference Box 2

adhesion molecules – cell surface molecules involved in the binding of cells to extracellular matrix or neighboring cells, where the principal function is adhesion rather than cell activation (e.g. integrins and selectins).

APCs (antigen-presenting cells) – a variety of cell types that carry antigen in a form that can stimulate lymphocytes.

dendritic cells – a set of cells that can be derived from either the lymphoid or mononuclear phagocyte lineages, are present in tissues, capture antigens, and migrate to the lymph nodes and spleen, where they are particularly active in presenting the processed antigen to T cells.

γδ T cells – the minor subset of T cells that express the gd form of the T cell receptor (TCR).

intercellular adhesion molecules (ICAMs) – cell surface molecules found on a variety of leukocytes and non-hematogenous cells that interact with integrins.

inflammation – a series of reactions that bring cells and molecules of the immune system to sites of infection or damage – this appears as an increase in blood supply, increased vascular permeability, and increased transendothelial migration of leukocytes.

LFA-1 (leukocyte functional antigen-1) – an integrin consisting of an α chain (CD11a) and a β chain (CD18) that acts as an intercellular adhesion molecule by binding to ICAM-1, ICAM-2, and ICAM-3. It has a role in leukocyte migration across vascular endothelium and in antigen presentation.

MHC (major histocompatibility complex) – a genetic region found in all mammals, encoding more than 100 genes. Class I and class II molecules are primarily responsible for transporting peptide antigens to the surface of a cell for recognition by T cells. Recognition of MHC class I and II molecules also underlies graft rejection.

Selected CD markers mentioned in this chapter

CD3 – a component of the T cell receptor complex.

CD4 – a marker of helper T cells.

CD5 – a marker of B-1 cells and marginal zone B cells.

CD8 – a marker of cytotoxic T lymphocytes (CTLs, Tc cells).

CD11a/CD18 – an integrin (LFA-1) found on most leukocytes.

CD11b/CD18 – an integrin (Mac-1) characteristic of mononuclear phagocytes.

CD16 – a low-affinity receptor for IgG, found on NK cells.

CD79 – a component of the B cell receptor complex.

Antigen-presenting cells (APCs) in the immune system

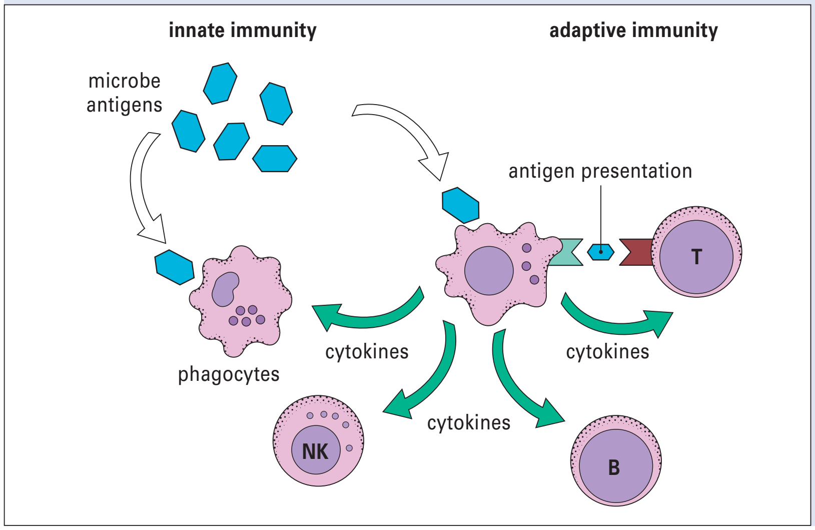

Fig. 2.2 Specialized APCs are involved in both innate and adaptive immunity to bacteria and viruses by the production of cytokines and for presentation of antigens to T cells.

- the spleen;
- lymph nodes; and
- mucosa-associated lymphoid tissues (MALT).

In the secondary lymphoid organs and tissues, the effector B and T cells account for the two major cell types participating in adaptive immune responses of humoral and cellular immunity, respectively.

As the cells of the immune system develop, they acquire molecules that are important for their function. These specific functional molecules are referred to as '**lineage markers**' because they identify the cell lineage, for example:

- myeloid cells – polymorphs and monocytes;
- lymphoid cells – T and B cells.

Other marker molecules include those involved in regulating cell proliferation and function and those involved in regulating the number of cells participating in the immune response. In both cases, these 'death receptors' mediate apoptosis of cells following their ligation.

PHAGOCYTIC CELLS ARE FOUND IN THE CIRCULATION AND IN THE TISSUES

Mononuclear phagocytes and polymorphonuclear granulocytes are the two major phagocyte lineages

Phagocytes belong to two major lineages:

- mononuclear phagocytes – monocytes/macrophages; and
- polymorphonuclear granulocytes.

The mononuclear phagocytes consist of circulating cells (the monocytes) and macrophages, which reside in a variety of organs (e.g. spleen, liver, lungs) where they display distinctive morphological features and perform diverse functions.

The other family of phagocytes, polymorphonuclear granulocytes, have a lobed, irregularly shaped (polymorphic) nucleus. On the basis of how their cytoplasmic granules stain with acidic and basic dyes they are classified into neutrophils, basophils, and eosinophils and have distinct effector functions:

- the neutrophils, also called polymorphonuclear neutrophils (PMNs), are most numerous and constitute the majority of leukocytes (white blood cells) in the blood stream (around 60–70% in adults);
- the primary actions of eosinophils and basophils, which can both function as phagocytes, involve granule exocytosis.

The mononuclear phagocytes and polymorphonuclear granulocytes develop from a common precursor (see below, Fig. 2.6).

Mononuclear phagocytes are widely distributed throughout the body

Cells of the mononuclear phagocytic system are found in virtually all organs of the body where the local microenvironment determines their morphology and functional characteristics (e.g. in the lung as alveolar macrophages and in the liver as Kupffer cells; Fig. 2.3).

The main role of the mononuclear phagocytes is to remove particulate matter of 'foreign' origin (e.g. microbes) or self origin (e.g. aged erythrocytes).

Myeloid progenitors in the bone marrow differentiate into pro-monocytes and then into circulating monocytes, which migrate through the blood vessel walls into organs to become macrophages.

The human blood monocyte:

- is large (10–18 μm in diameter) relative to the lymphocyte;
- has a horseshoe-shaped nucleus;
- contains azurophilic granules;
- possesses ruffled membranes, a well-developed Golgi complex, and many intracytoplasmic lysosomes (Fig. 2.4).

The lysosomes contain peroxidase and several acid hydrolases, which are important for killing phagocytosed microorganisms. Monocytes/macrophages actively phagocytose microorganisms or even tumor cells.

Microbial adherence, followed by ingestion, occurs through specialized receptors. These receptors (which include scavenger receptors, Toll-like receptors, and mannose receptors) attach mainly to sugars or lipids known as PAMPs on the microbial surface (see Chapter 6).

Kupffer cells

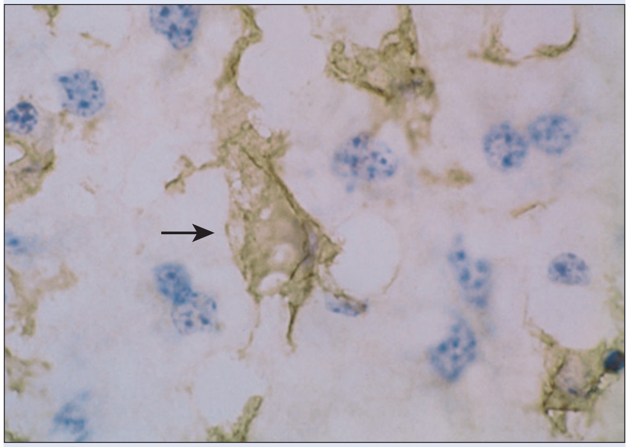

Fig. 2.3 Kupffer cells in the normal mouse liver stain strongly positive with antibody to F4/80 (arrow). Sinusoidal endothelial cells and hepatocytes are F4/80-negative. (Courtesy of Professor S Gordon and Dr DA Hume)

Morphology of the monocyte

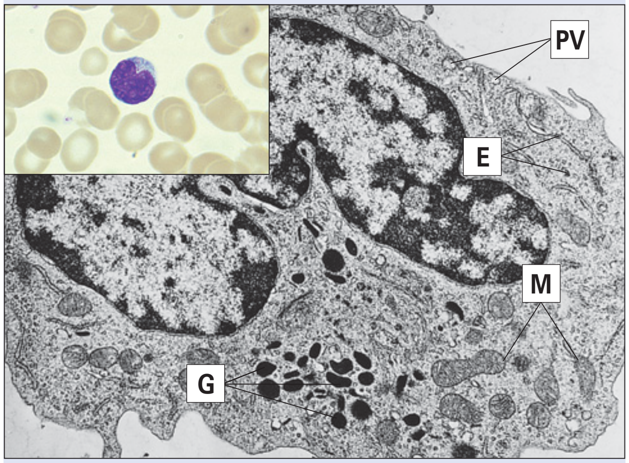

Fig. 2.4 Ultrastructure of a monocyte showing the horseshoe-shaped nucleus, pinocytotic vesicles (PV), lysosomal granules (G), mitochondria (M), and isolated rough endoplasmic reticulum cisternae (E). × 8000. (Courtesy of Dr B Nichols) Inset: Light microscope image of a monocyte from the blood. × 1200.

Coating microbes with complement components and/or antibodies (opsonization) enhances phagocytosis by monocytes/macrophages and is mediated by specialized complement receptors and antibody receptors (see Chapters 3 and 4).

Q. How do macrophages recognize microbes that have been coated with antibody?

A. Macrophages have Fc receptors that recognize the constant portion of the antibody molecule (see Figs 1.10 and 1.11).

There are three different types of polymorphonuclear granulocyte

Polymorphonuclear granulocytes (often referred to as **polymorphs** or **granulocytes**) consist mainly of neutrophils (PMNs). They:

- are released from the bone marrow at a rate of around 7 million per minute;
- are short-lived (2–3 days) relative to monocytes/macrophages, which may live for months or years.

Like monocytes, PMNs adhere to endothelial cells lining the blood vessels (marginate) and extravasate by squeezing between the endothelial cells to leave the circulation (see Fig. 1.17). This process is known as **diapedesis**. Adhesion is mediated by receptors on the granulocytes and ligands on the endothelial cells, and is promoted by chemoattractants (**chemokines**) such as **interleukin-8 (IL-8)** (see Chapter 4 and Appendix 3).

Like monocytes/macrophages, granulocytes also have receptors for PAMPs, and PMNs play an important role in acute inflammation (usually synergizing with antibodies and complement) in providing protection against microorganisms. Their predominant role is phagocytosis and destruction of pathogens.

The importance of granulocytes is evident from the observation of individuals who have a reduced number of white cells or who have rare genetic defects that prevent polymorph extravasation in response to chemotactic stimuli (see Chapter 16). These individuals have a markedly increased susceptibility to infection.

NEUTROPHILS COMPRISE OVER 95% OF THE CIRCULATING GRANULOCYTES – Neutrophils have a characteristic multilobed nucleus and are 10–20 μm in diameter (Fig. 2.5). Chemotactic agents for neutrophils include:

- protein fragments released when complement is activated (e.g. C5a);
- factors derived from the fibrinolytic and kinin systems;
- the products of other leukocytes and platelets; and
- the products of certain bacteria (see Chapter 4).

Chemotactic stimuli result in neutrophil margination and diapedesis (see Fig. 1.17).

Neutrophils have a large arsenal of antibiotic proteins stored in two main types of granule:

- the primary (azurophilic) granules are lysosomes containing acid hydrolases, myeloperoxidase, and muramidase (lysozyme);
- the secondary granules (specific to neutrophils) contain lactoferrin and lysozyme (see Fig. 2.5).

The primary granules also contain the antibiotic proteins – defensins, seprocidins, cathelicidins, and bacterial permeability inducing (BPI) protein.

The lysosomes containing the antibiotic proteins fuse with vacuoles containing ingested microbes (termed **phagosomes**) to become **phagolysosomes** where the killing takes place.

Neutrophils can also release granules and cytotoxic substances extracellularly when they are activated by immune complexes through their Fc receptors. This may be an important pathogenetic mechanism in immune complex diseases (type III hypersensitivity, see Chapter 25).

Morphology of the neutrophil

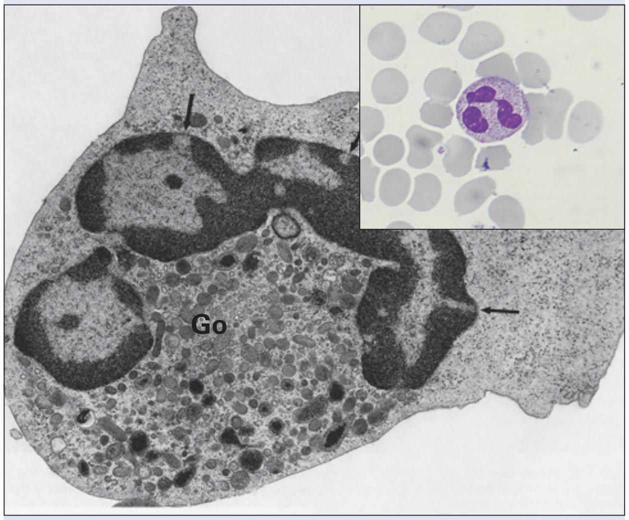

Fig. 2.5 At the ultrastructural level, the azurophilic (primary) granules are larger than the secondary (specific) granules with a strongly electron-dense matrix; the majority of granules are specific granules and contain a variety of toxic materials to kill microbes. A pseudopod (to the right) is devoid of granules. Arrows indicate nuclear pores. (Go, Golgi region) Inset: A mature neutrophil in a blood smear showing a multilobed nucleus. × 1500. (From Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

DIFFERENTIATION OF DIFFERENT CELL LINEAGES DEPENDS ON CELL INTERACTIONS AND CYTOKINES

Monocytes and neutrophils develop from a common precursor cell. **Myelopoiesis** (the development of myeloid cells) commences in the liver of the human fetus at about 6 weeks of gestation.

Studies in which colonies have been grown in vitro from individual stem cells have shown that the first progenitor cell derived from the hematopoietic stem cell (HSC) is the colony-forming unit (CFU), which can give rise to granulocytes, erythrocytes, monocytes, and megakaryocytes (CFU-GEMM).

CFU-GEMMs mature under the influence of colony-stimulating factors (CSFs) and several interleukins, including IL-1, IL-3, IL-4, IL-5, IL-6, and IL-7 (Fig. 2.6). These factors, which are relevant for the positive regulation of hemopoiesis, are:

- derived mainly from stromal cells (connective tissue cells) in the bone marrow;
- also produced by mature forms of differentiated myeloid and lymphoid cells.

Other cytokines, such as transforming growth factor-β (TGFβ) may downregulate hemopoiesis.

The CFU-granulocyte macrophage cells (CFU-GMs) are the precursors of both neutrophils and mononuclear phagocytes (see Fig. 2.6).

Development of granulocytes and monocytes

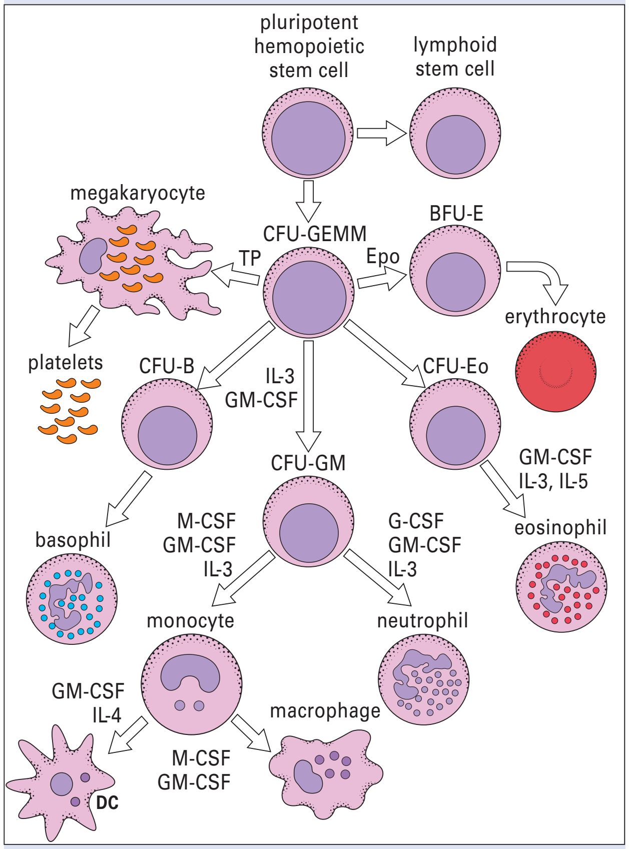

Fig. 2.6 Pluripotent hematopoietic stem cells generate colony-forming units (CFUs) that can give rise to granulocytes, erythrocytes, monocytes, and megakaryocytes (CFU-GEMMs). CFU-GEMMs therefore have the potential to give rise to all blood cells except lymphocytes. IL-3 and granulocyte–macrophage colony-stimulating factor (GM-CSF) are required to induce the CFU-GEMM stem cell to enter one of five pathways (i.e. to give rise to megakaryocytes, erythrocytes via burst-forming units, basophils, neutrophils, or eosinophils). IL-3 and GM-CSF are also required during further differentiation of the granulocytes and monocytes. Eosinophil (Eo) differentiation from CFU-Eo is promoted by IL-5. Neutrophils and monocytes are derived from the CFU-GM through the effects of G-CSF and M-CSF, respectively. Both GM-CSF and M-CSF, and other cytokines (including IL-1, IL-4, and IL-6), promote the differentiation of monocytes into macrophages. Thrombopoietin (TP) promotes the growth of megakaryocytes. (B, basophil; BFU-E, erythrocytic burst-forming unit; DC, dendritic cell; Epo, erythropoietin; G, granulocyte; M, monocyte)

EACH CELL TYPE EXPRESSES CHARACTERISTIC SURFACE MOLECULES

Monocytes express CD14 and significant levels of MHC class II molecules

CFU-GMs taking the monocyte pathway give rise initially to proliferating monoblasts. Proliferating monoblasts differentiate into pro-monocytes and finally into mature circulating monocytes (see Fig. 2.4). Circulating monocytes are a replacement pool for the tissue-resident macrophages (e.g. lung macrophages).

CD34, like other early maturation markers in this lineage, is lost in mature neutrophils and monocytes/macrophages. Other markers may be lost as differentiation occurs along one pathway, but retained in the other. For example, the common precursor of monocytes and neutrophils, the CFU-GM cell, expresses **major histocompatibility complex (MHC) class II molecules**, but only monocytes continue to express significant levels of this marker.

Q. What is the functional significance of the expression of MHC molecules on monocytes?

A. Monocytes can present antigens to helper T cells, but neutrophils generally cannot.

Monocytes/macrophages and granulocytes have different functional molecules (e.g. monocytes express CD14 which is part of the receptor complex for the lipopolysaccharide of Gram-negative bacteria). In addition, monocytes acquire many of the same surface molecules as mature neutrophils (e.g. the adhesion molecules CD11a and b and antibody Fc receptors). Examples of the distinct and common functional molecules can be seen in the list of CD markers in Appendix 2.

Neutrophils express adhesion molecules and receptors involved in phagocytosis

CFU-GMs go through several stages to become neutrophils. As the CFU-GM cell differentiates along the neutrophil pathway, several distinct morphological stages are distinguished. Myeloblasts develop into promyelocytes and myelocytes, which mature and are released into the circulation as neutrophils.

The one-way differentiation of the CFU-GM into mature neutrophils is the result of acquiring specific receptors for growth and differentiation factors at progressive stages of development. Surface differentiation markers disappear or are expressed on the cells as they develop into granulocytes. For example, MHC class II molecules are expressed on the CFU-GM, but not on mature neutrophils.

Other surface molecules acquired during the differentiation process include:

- adhesion molecules (e.g. the leukocyte integrins CD11a, b, c, and d associated with CD18 β_2 chains); and
- receptors involved in phagocytosis including complement and antibody Fc receptors (see Appendix 4).

It is difficult to assess the functional activity of different developmental stages of granulocytes, but it seems likely that the full functional potential is realized only when the cells are mature.

There is some evidence that neutrophil activity, as measured by phagocytosis or chemotaxis, is lower in fetal than in adult life. However, this may be due, in part, to the lower levels of opsonins in the fetal serum, rather than to a characteristic of the cells themselves.

To become active in the presence of opsonins, neutrophils must interact directly with microorganisms and/or with cytokines generated by a response to antigen. This limitation could reduce neutrophil activity in early life.

Activation of neutrophils by cytokines and chemokines is also a prerequisite for their migration into tissues (see Chapter 9).

EOSINOPHILS, BASOPHILS, MAST CELLS, AND PLATELETS TAKE PART IN THE INFLAMMATORY RESPONSE

Eosinophils are thought to play a role in immunity to parasitic worms

Eosinophils comprise 2–5% of blood leukocytes in healthy non-allergic individuals. Human blood eosinophils usually have a bilobed nucleus and many cytoplasmic granules, which stain with acidic dyes such as eosin (Fig. 2.7). Although not their primary function, eosinophils appear to be capable of phagocytosing and killing ingested microorganisms.

The granules in mature eosinophils are membrane-bound organelles with crystalloid cores that differ in electron density from the surrounding matrix (see Fig. 2.7). The crystalloid core contains the **major basic protein (MBP)**, which:

- is a potent toxin for helminth worms;
- induces histamine release from mast cells;
- activates neutrophils and platelets; and
- of relevance to allergy, provokes bronchospasm.

Other proteins with similar effects are found in the granule matrix, for example:

- eosinophil cationic protein (ECP); and
- eosinophil-derived neurotoxin (EDN).

Morphology of the eosinophil

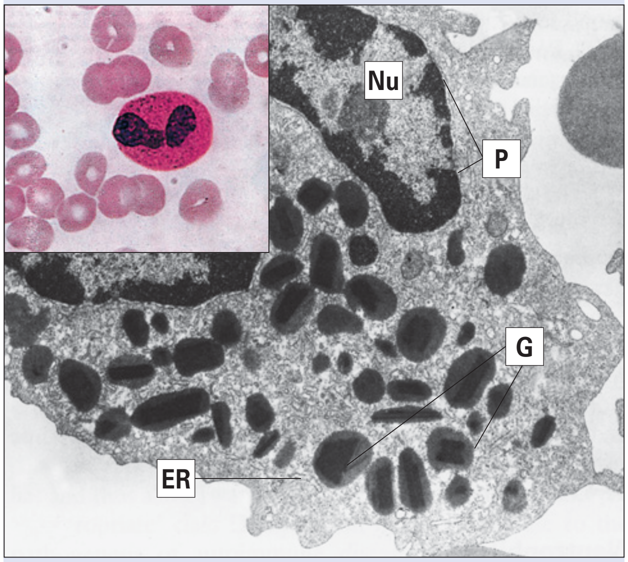

Fig. 2.7 The ultrastructure of a mature eosinophil shows granules (G) with central crystalloids. × 17 500. (ER, endoplasmic reticulum; Nu, nucleus; P, nuclear pores) Inset: A mature eosinophil in a blood smear is shown with a bilobed nucleus and eosinophilic granules. × 1000. (From Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

Release of the granules on eosinophil activation is the only way in which eosinophils can kill large pathogens (e.g. schistosomula), which cannot be phagocytosed. Eosinophils are therefore thought to play a specialized role in immunity to parasitic worms using this mechanism (see Fig. 15.15).

Basophils and mast cells play a role in immunity against parasites

Basophils are found in very small numbers in the circulation and account for less than 0.2% of leukocytes (Fig. 2.8).

The mast cell (Fig. 2.9), which is not found in the circulation, is indistinguishable from the basophil in a

Morphology of the basophil

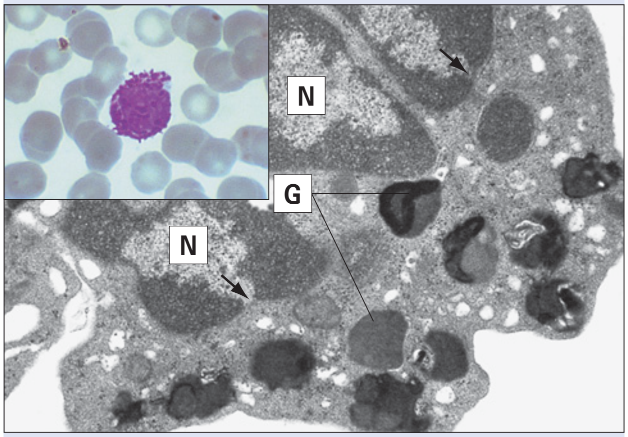

Fig. 2.8 Ultrastructural analysis shows a segmented nucleus (N) and the large cytoplasmic granules (G). Arrows indicate nuclear pores. × 11 000. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003) Inset: This blood smear shows a typical basophil with its deep violet-blue granules. × 1000.

Histological appearance of human connective tissue mast cells

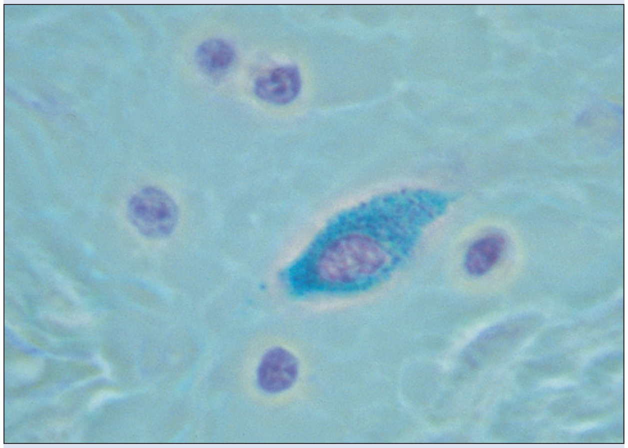

Fig. 2.9 This micrograph shows dark blue cytoplasm with purple granules. Alcian blue and safranin stain. × 600. (Courtesy of Dr TS Orr)

Some distinctive and common characteristics of basophils and mast cells

	basophils	mast cells
origin	bone marrow	bone marrow
site of maturation	bone marrow	connective tissues
presence in the circulation	yes	no
proliferative capacity	no	yes
life span	days	weeks to months
surface expression of FcεR1	yes	yes
granule content		
histamine	yes	yes
heparin	?	yes
cytokine production		
IL-4	yes	yes
IL-13	yes	yes

Fig. 2.10 Some distinctive and common characteristics of basophils and mast cells.

number of its characteristics, but displays some distinctive morphological features (Fig. 2.10).

The stimulus for mast cell or basophil degranulation is often an **allergen** (i.e. an antigen causing an allergic reaction). To be effective, an allergen must cross-link IgE molecules bound to the surface of the mast cell or basophil via its high-affinity Fc receptors for IgE (FcεRI) (see p. 77). Degranulation of a basophil or mast cell results in all contents of the granules being released very rapidly. This occurs by intracytoplasmic fusion of the granules, followed by discharge of their contents (Fig. 2.11).

Mediators such as histamine, released by degranulation, cause the adverse symptoms of allergy, but, on the positive side, also play a role in immunity against parasites by enhancing acute inflammation.

Platelets have a role in clotting and inflammation

Blood platelets (Fig. 2.12) are not cells, but cell fragments derived from megakaryocytes in the bone marrow. They contain granules, microtubules, and actin/myosin filaments, which are involved in clot contraction. Platelets are also involved in immune responses, especially in inflammation.

The adult human produces 10^{11} platelets each day. About 30% of platelets are stored in the spleen, but may be released if required.

Q. What circumstance might require the release of additional platelets into the circulation?

A. Severe blood loss.

Electron micrograph study of rat mast cells

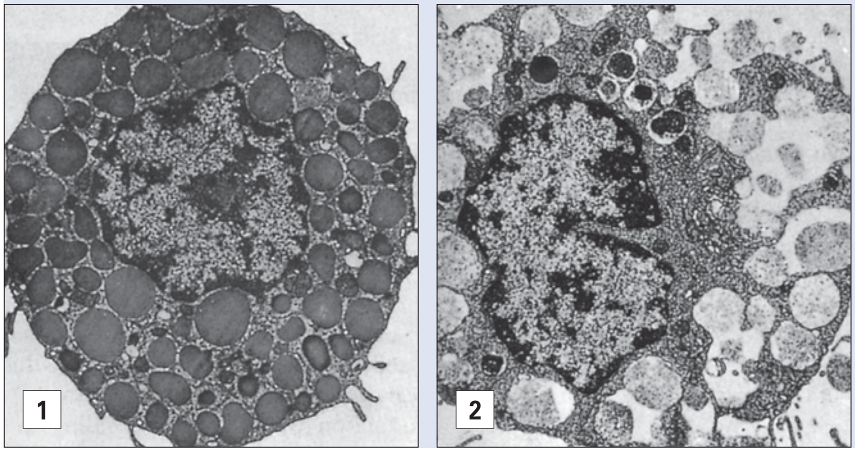

Fig. 2.11 Rat peritoneal mast cells show electron-dense granules (**1**). Vacuolation with exocytosis of the granule contents has occurred after incubation with anti-IgE (**2**). Transmission electron micrographs. × 2700. (Courtesy of Dr D Lawson)

Ultrastructure of a platelet

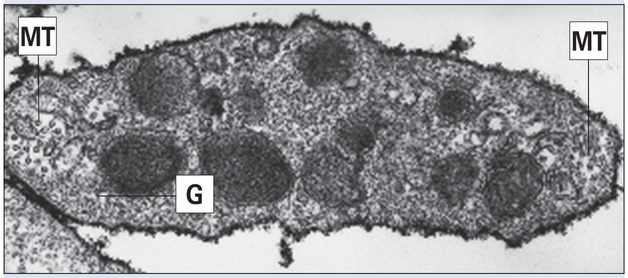

Fig. 2.12 Cross-section of a platelet showing two types of granule (G) and bundles of microtubules (MT) at either end. × 42 000. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

Platelets express class I MHC products and receptors for IgG (CD32; FcγRII), which are important in platelet activation via IgG immune complexes. In addition, megakaryocytes and platelets carry:

- receptors for clotting factors (e.g. factor VIII); and
- other molecules important for their function, such as the GpIIb/IIIa complex (CD41) responsible for binding to fibrinogen, fibronectin, vitronectin (tissue matrix), and von Willebrand factor (another clotting factor).

Both receptors and adhesion molecules are important in the activation of platelets.

Following injury to endothelial cells, platelets adhere to and aggregate at the damaged endothelial surface. Release of platelet granule contents, which include de-novo synthesized serotonin and endocytosed fibrinogen, results in:

- increased capillary permeability;
- activation of complement (and hence attraction of leukocytes); and
- clotting.

NK CELLS KILL VIRUS-INFECTED AND TUMOR CELLS

NK cells account for up to 15% of blood lymphocytes and express neither T cell nor B cell antigen receptors. They are derived from the bone marrow and morphologically have the appearance of large granular lymphocytes (see Fig. 2.19).

Functional NK cells are found in the spleen, and cells found in lymph nodes that express CD56 but not CD16 (see below) might represent immature NK cells.

Most surface antigens detectable on NK cells by monoclonal antibodies are shared with T cells or monocytes/macrophages.

CD16 and CD56 are important markers of NK cells

Monoclonal antibodies to CD16 (FcγRIII) are commonly used to identify NK cells in purified lymphocyte populations. CD16 is involved in one of the activation pathways of NK cells and is also expressed by neutrophils, some macrophages, and some T cells.

On granulocytes, CD16 is linked to the surface membrane by a glycoinositol phospholipid (GPI) linkage, whereas NK cells, macrophages, and γδ T cells express the transmembrane form of the molecule.

The CD56 molecule, a homophilic adhesion molecule of the immunoglobulin superfamily (NCAM), is another important marker of NK cells.

The absence of CD3, but the presence of CD56 and CD16, are currently the most reliable markers for NK cells in humans, though both markers can also be found on a minority of T cells (mostly with $CD3^+/CD8^+$ phenotype).

Some markers of NK cells are given in Appendix 4. Resting NK cells also express the β chain of the IL-2 receptor, which is an intermediate affinity receptor of 70 kDa, and the signal transducing common γ chain of IL-2 and other cytokine receptors. Therefore, direct stimulation with IL-2 activates NK cells.

The function of NK cells is to recognize and kill (Fig. 2.13) through apoptosis:

- virus-infected cells; and
- certain tumor cells.

The mechanism of recognition is not fully understood, but involves both activating and inhibitory receptors.

Classical and non-classical MHC class I molecules (see Fig. 5.16) are ligands for the inhibitory receptors and this explains why normal body cells (all of which normally express MHC class I molecules) are not targeted by NK cells.

Downregulation or modification of MHC molecules in virus-infected cells and some tumors makes them susceptible to NK cell-mediated killing.

Q. What advantage is there for a virus in causing the loss of MHC class I molecules in the cell it has infected?

A. The infected cell can no longer be recognized by cytotoxic T cells (see Fig. 1.12).

An NK cell attached to a target cell

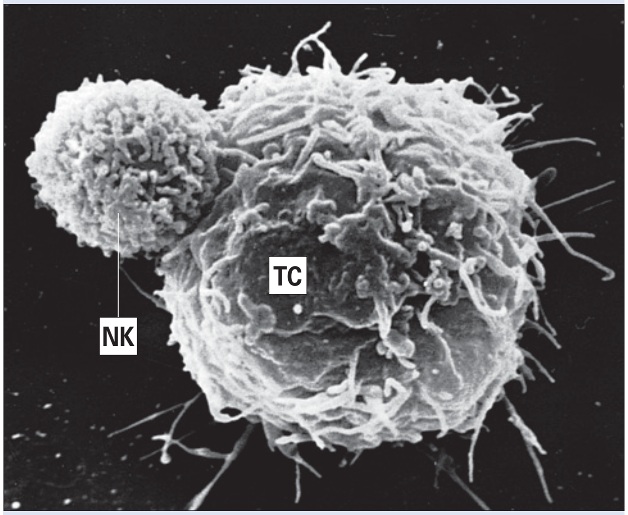

Fig. 2.13 An NK cell (NK) attached to a target cell (TC). × 4500. (Courtesy of Dr G Arancia and W Malorni, Rome)

NK cells are also able to kill targets coated with IgG antibodies via their receptor for IgG (FcγRIII, CD16). This property is referred to as **antibody-dependent cellular cytotoxicity (ADCC)**.

NK cells release interferon-γ (IFNγ) and other cytokines (e.g. IL-1 and GM-CSF) when activated, which might be important in the regulation of hemopoiesis and immune responses.

Different kinds of antigen-presenting cells (APCs)

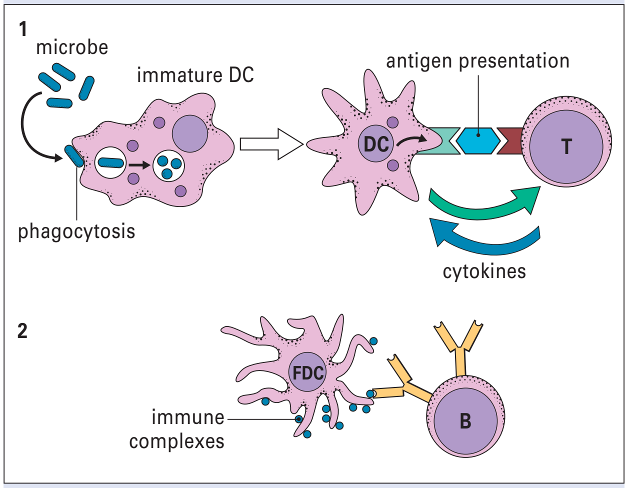

Fig. 2.14 There are two main types of specialized APCs – dendritic cells (DCs) and follicular dendritic cells (FDCs).
(1) Immature DCs are derived from bone marrow and interact mainly with T cells. They are highly phagocytic, take up microbes, process the foreign microbial antigens, and become mature APCs carrying the processed antigen on their surface with specialized MHC molecules. Specific T cells recognize the displayed antigen and, in the presence of cytokines produced by the mature DC, proliferate and also produce cytokines.
(2) FDCs are not bone marrow derived and interact with B cells. In the B cell follicles of lymphoid organs and tissues they bind small immune complexes (IC, called iccosomes). Antigen contained within the IC is presented to specific B cells in the lymphoid follicles. This protects the B cell from cell death. The B cell then proliferates and with T cell help can leave the follicle and become a plasma cell or memory cell (see Fig. 2.48)

APCs LINK THE INNATE AND ADAPTIVE IMMUNE SYSTEMS

APCs are a heterogeneous population of leukocytes that are:

- important in innate immunity (see Fig. 2.2);
- play a pivotal role in the induction of functional activity of T helper (TH) cells.

In this regard, APCs are seen as the interface between the innate and adaptive immune systems.

Functionally, APCs are divided into those that both process and present foreign protein antigens to T cells – **dendritic cells (DCs)** – and a separate type of APC that passively presents foreign antigen in the form of immune complexes to B cells in lymphoid follicles – **follicular dendritic cells** (**FDCs**; Fig. 2.14).

APCs are found primarily in the skin, lymph nodes, and spleen, and within or underneath most mucosal epithelia. They are also present in the thymus, where they present self antigens to developing T cells (Fig. 2.15).

Langerhans' cells and IDCs are rich in class II MHC molecules

Langerhans' cells in the epidermis and in other squamous epithelia migrate as 'veiled cells', via the afferent lymphatics into the paracortex of the draining lymph nodes (see Fig. 2.15). Here, they interact with T cells and are termed interdigitating cells (IDCs; Fig. 2.16). These APCs are rich in class II MHC molecules, which are important for presenting antigen to TH cells.

Q. What function could the migration of Langerhans' cells to the lymph nodes from the mucosa or skin serve?

A. The migration of Langerhans' cells provides an efficient mechanism for carrying antigen from the skin and mucosa to the TH cells in the lymph nodes, and are rich in class II MHC molecules, which are important for presenting antigen to TH cells. Lymph nodes provide the appropriate environment for lymphocyte proliferation.

APCs are also present within the germinal centers (GCs) of secondary lymphoid follicles (i.e. they are the MHC class II molecule-positive germinal center DCs [GCDCs]). In contrast to FDCs, they are migrating cells, which on arrival in the GC interact with T cells and are probably involved in antibody class switching (see Chapter 8).

Thymic APCs (also called IDCs) are especially abundant in the medulla (see Fig. 2.15). The thymus is of crucial importance in the development and maturation of T cells, and it appears that IDCs play a role in deleting T cells that react against self antigen. This process is referred to as 'negative selection'.

Most APCs derive from one of two bone marrow precursors:

- a myeloid progenitor (DC1) that gives rise to myeloid DCs; and
- a lymphoid progenitor (DC2) that develops into plasmacytoid DCs (Fig. 2.17).

DCs are not the only APCs interacting with T cells because both macrophages and classical B cells are rich in membrane class II MHC molecules, especially after activation, and are thus able to process and present specific antigens to (activated) T cells (see Chapter 7).

Somatic cells other than immune cells do not normally express class II MHC molecules, but cytokines such as IFNγ and tumor necrosis factor-α (TNFα) can induce the expression of class II molecules on some cell types, and thus allow them to present antigen (e.g. epidermis and thyroid epithelium and endothelial cells). This induction of 'inappropriate' class II expression might contribute to the pathogenesis of autoimmune diseases and to prolonged inflammation (see Chapter 20).

Migration of antigen-presenting cells (APCs) into lymphoid tissues

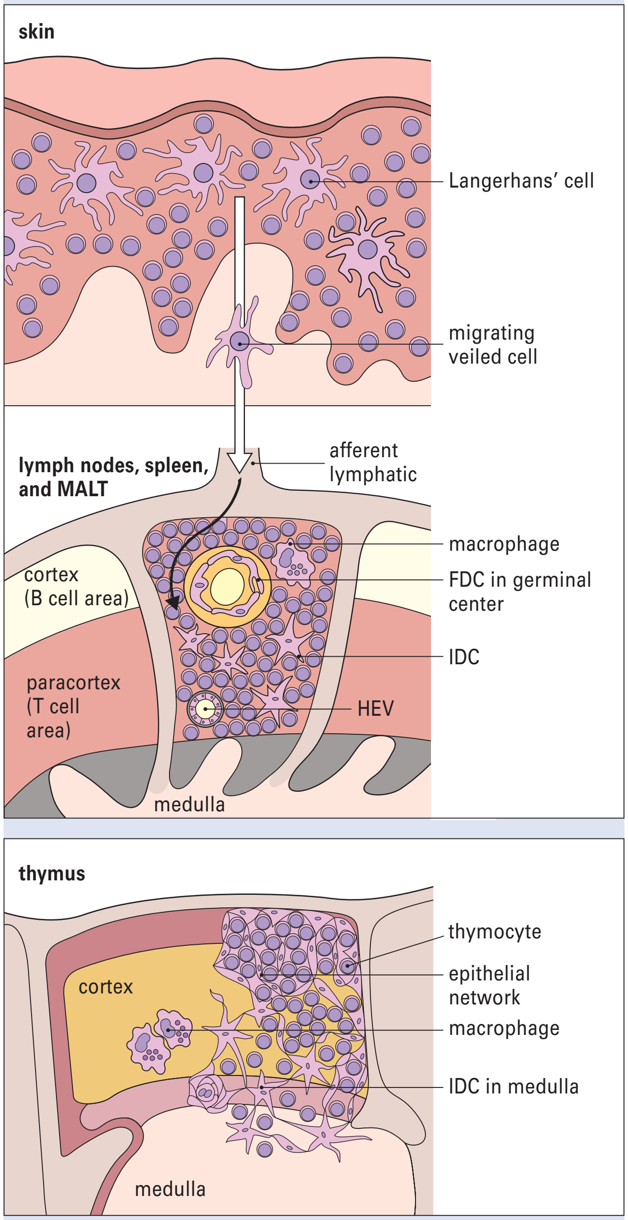

Fig. 2.15 Bone marrow-derived APCs are found especially in lymphoid tissues, in the skin, and in mucosa. APCs in the form of Langerhans' cells are found in the epidermis and are characterized by special granules (the tennis racquet-shaped Birbeck granules; not shown here). Langerhans' cells are rich in MHC class II molecules, and carry processed antigens. They migrate via the afferent lymphatics (where they appear as 'veiled' cells) into the paracortex of the draining lymph nodes. Here they make contact with T cells. These 'interdigitating dendritic cells (IDCs)', localized in the T cell areas of the lymph node, present antigen to T helper cells. Antigen is exposed to B cells on the follicular dendritic cells (FDCs) in the germinal centers of B cell follicles. Some macrophages located in the outer cortex and marginal sinus may also act as APCs. In the thymus, APCs occur as IDCs in the medulla. (HEV, high endothelial venule)

Ultrastructure of an interdigitating dendritic cell (IDC) in the T cell area of a rat lymph node

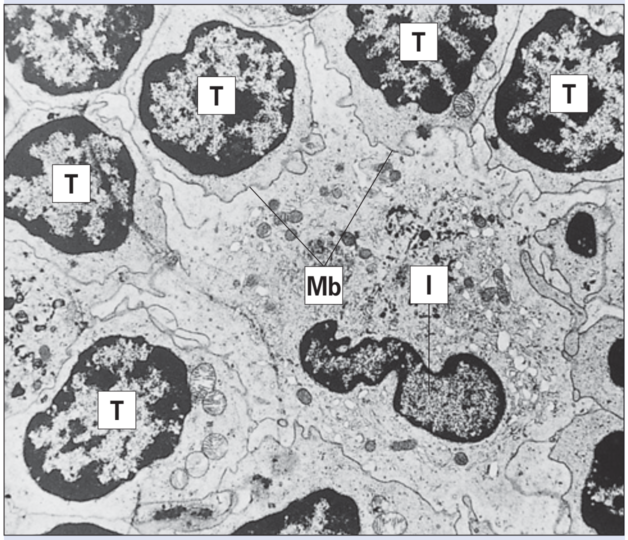

Fig. 2.16 Intimate contacts are made with the membranes of the surrounding T cells. The cytoplasm contains a well-developed endosomal system and does not show the Birbeck granules characteristic of skin Langerhans' cells. × 2000. (I, IDC nucleus; Mb, IDC membrane; T, T cell nucleus) (Courtesy of Dr BH Balfour)

FDCs lack class II MHC molecules and are found in B cell areas

Unlike the APCs that actively process and present protein antigens to T cells, FDCs have a passive role in presenting antigen in the form of immune complexes to B cells. They are therefore found in the primary and secondary follicles of the B cell areas of the lymph nodes, spleen, and MALT (see Fig. 2.15). They are a non-migratory population of cells and form a stable network (a kind of web) by establishing strong intercellular connections via desmosomes.

FDCs lack class II MHC molecules, but bind antigen via complement receptors (CD21 and CD35), which attach to complement associated with immune complexes (iccosomes; Fig. 2.18). They also express Fc receptors. These FDCs are not bone marrow derived, but are of mesenchymal origin.

LYMPHOCYTES ARE HETEROGENEOUS

Lymphocytes are phenotypically and functionally heterogeneous

Large numbers of lymphocytes are produced daily in the primary or central lymphoid organs (i.e. thymus and postnatal bone marrow). Some migrate via the circulation into the secondary lymphoid tissues (i.e. spleen, lymph nodes, and MALT).

The average human adult has about 2×10^{12} lymphoid cells and lymphoid tissue as a whole represents about 2% of total body weight. Lymphoid cells account for about 20% of the leukocytes in the adult circulation.

Myeloid and plasmacytoid dendritic cells

	myeloid DCs	plasmacytoid DCs
origin of precursor	myeloid (DC1)	lymphoid (DC2)
localization	diffuse – epidermis, mucosae, thymus, and T cell areas of secondary lymphoid organs and tissues	restricted to T cell areas of secondary lymphoid organs and tissues
myeloid markers	many	none
characteristic cytokines produced	mainly IL-8, IL-12	mainly type I interferons (on challenge with enveloped viruses)

Fig. 2.17 Myeloid and plasmacytoid dendritic cells (DCs).

Follicular dendritic cell

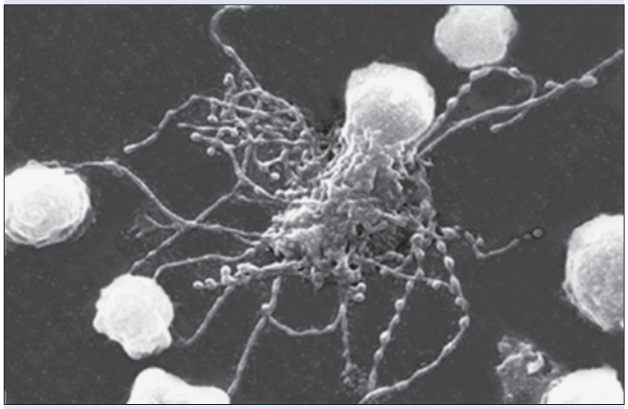

Fig. 2.18 An isolated follicular dendritic cell (FDC) from the lymph node of an immunized mouse 24 hours after injection of antigen. The FDC is of intermediate maturity with smooth filiform dendrites typical of young FDCs, and beaded dendrites, which participate in the formation of iccosomes (immune complexes) in mature FDCs. The adjacent small white cells are lymphocytes. (Electron micrograph kindly provided by Dr Andras Szakal; reproduced by permission of the Journal of Immunology)

Many mature lymphoid cells are long-lived, and persist as memory cells for many years.

Q. Given that there are roughly 10^9 lymphocytes/litre of blood and 10 litres of blood in an individual, and that roughly 2×10^9 new cells are produced each day, what can you infer about the location and life span of lymphocytes within an individual?

A. This implies that less than 1% of an individual's lymphocytes are in the circulation. This highly selective population of lymphocytes is mostly en route between tissues. The data also imply that many lymphocytes must die each day to maintain the overall balance of the lymphoid system, and that the average life span of a lymphocyte will be months to years. The actual values are, however, enormously variable, depending on the lymphocyte.

Lymphocytes are morphologically heterogeneous

In a conventional blood smear, lymphocytes vary in both size (from 6 to 10 μm in diameter) and morphology. Differences are seen in:

- nuclear to cytoplasmic (N:C) ratio;
- nuclear shape; and
- the presence or absence of azurophilic granules.

Two distinct morphological types of lymphocyte are seen in the circulation as determined by light microscopy and a hematological stain such as Giemsa (Fig. 2.19):

- the first type is relatively small, is typically agranular and has a high N:C ratio (Fig. 2.19[1]).
- the second type is larger, has a lower N:C ratio, contains cytoplasmic azurophilic granules, and is known as the large granular lymphocyte (LGL).

LGLs should not be confused with granulocytes, monocytes, or their precursors, which also contain azurophilic granules.

Most T cells express the αβ T cell receptor (see below) and, when resting, can show either of the above morphological patterns.

Most T helper (TH) cells (approximately 95%) and a proportion (approximately 50%) of cytotoxic T cells (Tc or CTL) have the morphology shown in Fig. 2.19(1).

The LGL morphological pattern displayed in Fig. 2.19(2) is shown by less than 5% of TH cells and by about 30–50% of Tc cells. These cells display LGL morphology with primary lysosomes dispersed in the cytoplasm and a well-developed Golgi apparatus, as shown in Fig. 2.19(3).

Most B cells, when resting, have a morphology similar to that seen in Fig. 2.19(1) under light microscopy.

B AND T CELLS EXPRESS ANTIGEN RECEPTORS

Lymphocytes express characteristic surface markers

Lymphocytes (and other leukocytes) express a large number of different functionally important molecules on their surfaces, which can be used to distinguish ('mark') cell subsets. Many of these cell markers can be identified by specific monoclonal antibodies (mAb) and can be used

Morphological heterogeneity of lymphocytes

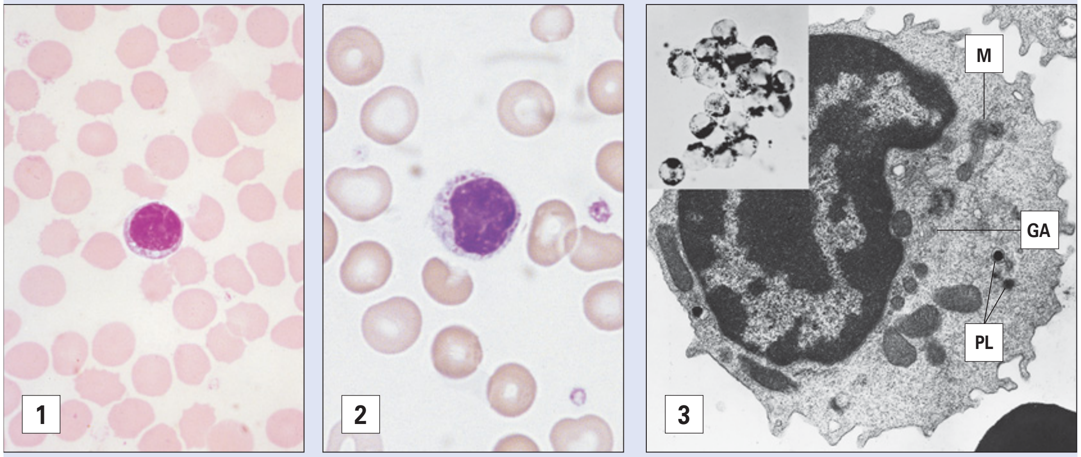

Fig. 2.19 (**1**) The small lymphocyte has no granules, a round nucleus, and a high N:C ratio. (**2**) The large granular lymphocyte (LGL) has a lower N:C ratio, indented nucleus, and azurophilic granules in the cytoplasm. Giemsa stain. (Courtesy of Dr A Stevens and Professor J Lowe) (**3**) Ultrastructure of the LGL shows characteristic electron-dense peroxidase-negative granules (primary lysosomes, PL), scattered throughout the cytoplasm, with some close to the Golgi apparatus (GA) and many mitochondria (M). ×10 000. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

Main distinguishing markers of T and B cells

CD number	T cells	B cells
antigen receptor	TCR – ($\alpha\beta$ or $\gamma\delta$)	immunoglobulin (Ig)
CD1	–	+
CD3	+ (part of the TCR complex)	–
CD4	+ (subset)	–
CD8	+ (subset)	–
CD19	–	+
CD20	–	+
CD23	+ (subset)	+
CD40	–	+
CD79a	–	+ (part of the BCR complex)
CD79b	–	+ (part of the BCR complex)

BCR, B cell receptor; TCR, T cell receptor

Fig 2.20 Main distinguishing markers of T and B cells.

to distinguish T cells from B cells (Fig. 2.20 and see Appendix 2 for more CD markers).

Human lymphocytes can be identified in tissues and their function can be measured as separated populations. Immunofluorescence techniques to identify leukocytes are shown in Method Box 2.1.

Lymphocytes express a variety of cell surface molecules that belong to different families, which have probably evolved from a few ancestral genes. These families of molecules are shared with other leukocytes and are distinguished by their structure. The major families include:

- the immunoglobulin superfamily;
- the integrin family;
- selectins;
- proteoglycans.

METHOD BOX 2.1
Identification of cell populations

Molecules on or in cells can be identified using fluorescent antibodies as probes. The antibodies can be applied to tissue sections or used in flow cytometry.

Immunofluorescence
Immunofluorescence detects antigen in situ. A section is cut on a cryostat from a deep-frozen tissue block. This ensures that labile antigens are not damaged by fixatives.

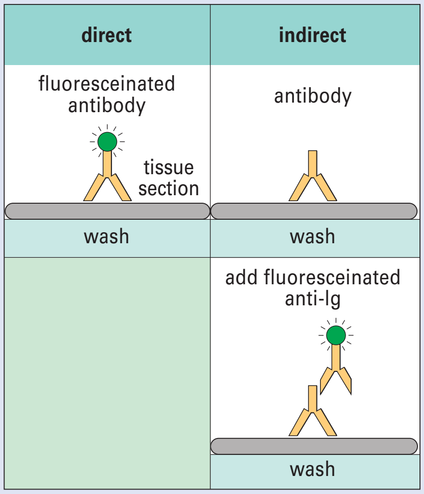
direct
indirect
fluoresceinated antibody
tissue section
antibody
wash
wash
add fluoresceinated anti-Ig
wash

Fig. 1 Immunofluorescence.

Direct immunofluorescence
The test solution of fluoresceinated antibody, e.g. to CD3, is applied to the section in a drop, incubated, and washed off. Any bound antibody is then revealed under the microscope. Ultraviolet light is directed onto the section through the objective and as a result the field is dark and areas with bound fluorescent antibody on and in the T cells fluoresce.

Indirect immunofluorescence
Antibody applied to the section as a solution is visualized using fluoresceinated anti-immunoglobulin, to the mouse anti-human CD3 antibodies in this case.

Indirect complement amplified – an elaboration of indirect immunofluorescence
This is an elaboration of the indirect immunofluorescence for the detection of complement-fixing antibody. In the second step fresh complement is added, which becomes fixed around the site of antibody binding. Due to the amplification steps in the classical complement pathway (see Chapter 4) one antibody molecule can cause many C3b molecules to bind to the section – these are then visualized with fluoresceinated anti-C3.

Flow cytometry and cell sorting
Cells are stained with specific fluorescent reagents to detect surface molecules and are then introduced into the flow chamber of the flow cytometer. The cell stream passing out of the chamber is encased in a sheath of buffer fluid. The stream is illuminated by laser light and each cell is measured for size (forward light scatter) and granularity (90° light scatter), as well as for red and green fluorescence to detect two different surface markers. In a cell sorter the flow chamber vibrates the cell stream causing it to break into droplets, which are then charged and may be steered by deflection plates under computer control to collect different cell populations according to the parameters measured. Plot (**2.1**) shows peripheral blood mononuclear cells double stained with FITC-conjugated anti-CD3 antibody (x axis) and PE-labeled anti-CD8 antibody (y axis). Four populations can be seen, and the CD8 cells appear in the right upper quadrant. The CD8 cells are then selected (gated) and the isolated CD8 population is shown in plot (**2.2**).

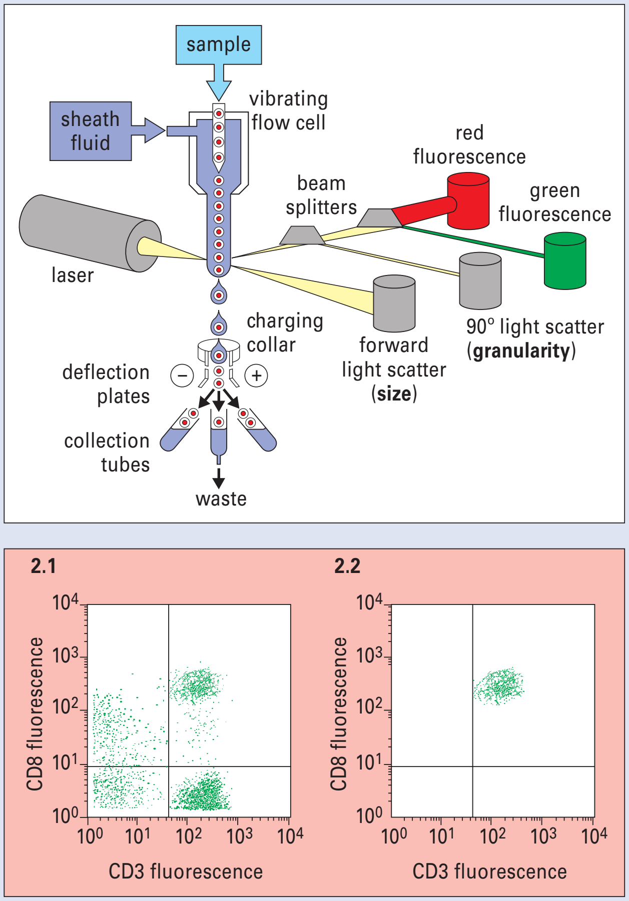
sample
sheath fluid
vibrating flow cell
red fluorescence
beam splitters
green fluorescence
laser
charging collar
forward light scatter (size)
90° light scatter (granularity)
deflection plates
collection tubes
waste
2.1
2.2
CD8 fluorescence
CD3 fluorescence
10^0
10^1
10^2
10^3
10^4

Fig. 2 Fluorescence-activated cell sorter (FACS).

METHOD BOX 2.1
Identification of cell populations *(Continued)*

ELISPOT assays

Individual B cells producing specific antibody or individual T cells secreting particular cytokines may be detected by enzyme linked immuno SPOT (ELISPOT) assay.

For detection of antibody-producing cells

The lymphocytes are plated onto an antigen-sensitized plate. Secreted antibody binds antigen in the immediate vicinity of cells producing the specific antibody. The spots of bound antibody are then detected chromatographically using enzyme coupled to anti-immunoglobulin and a chromogen.

For detection of cytokine-producing cells

The plates are coated with anti-cytokine and the captured cytokine is detected with enzyme-coupled antibody to a different epitope on the cytokine.

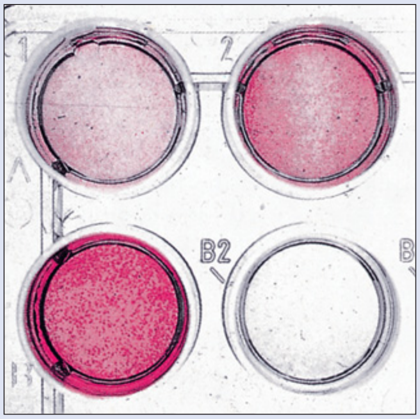

Appearance of a developed ELISPOT plate

Photograph courtesy of P. Hutchings and Blackwell Scientific.

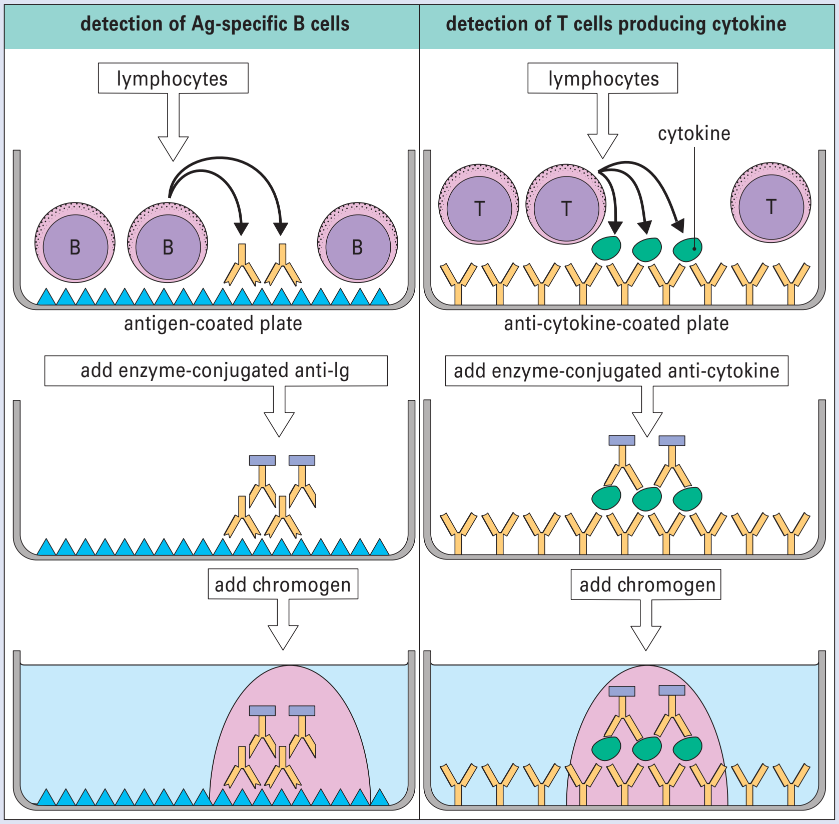
detection of Ag-specific B cells
detection of T cells producing cytokine
lymphocytes
B
B
B
antigen-coated plate
lymphocytes
cytokine
T
T
T
anti-cytokine-coated plate
add enzyme-conjugated anti-Ig
add enzyme-conjugated anti-cytokine
add chromogen
add chromogen

3

Fig. 3 ELISPOT assays.

The **immunoglobulin superfamily** comprises molecules with structural characteristics similar to those of the immunoglobulins and includes CD2, CD3, CD4, CD8, CD28, MHC class I and II molecules, and many more.

The **integrin** family consists of heterodimeric molecules with α and β chains. There are several integrin subfamilies and all members of a particular subfamily share a common β chain, but each has a unique α chain:

- one integrin subfamily (**the β_2-integrins**) uses CD18 as the β chain, which can be associated with CD11a, CD11b, CD11c, or αd – these combinations make up the lymphocyte function antigens LFA-1, Mac-1 (CR3), p150,95, and $\alpha d\beta_2$ surface molecules respectively – and are commonly found on leukocytes.
- a second subfamily (**the β_1-integrins**) has CD29 as the β chain, which again is associated with various other peptides and includes the VLA (very late activation) markers.

The **selectins** (CD62, E, L, and P) are expressed on leukocytes (L) or activated endothelial cells and platelets (E and P). They have lectin-like specificity for a variety of sugars expressed on heavily glycosylated membrane glycoproteins (e.g. CD43).

The **proteoglycans**, typically CD44, have a number of glycosaminoglycan (GAG) binding sites (e.g. for chondroitin sulfate), and bind to extracellular matrix components (typically, hyaluronic acid).

Other families include:

- the tumor necrosis factor (TNF) and nerve growth factor (NGF) receptor superfamily;
- the C-type lectin superfamily;
- the family of receptors with seven transmembrane segments (tm7); and
- the tetraspanins, a superfamily with four membrane-spanning segments (tm4), for example CD20.

Marker molecules allow lymphocytes to communicate with their environment

The major function of the families of marker molecules described above is to allow lymphocytes to communicate with their environment. They are extremely important in cell trafficking, adhesion, and activation.

Markers expressed by lymphocytes can often be detected on cells of other lineages (e.g. CD44 is commonly expressed by epithelial cells).

Marker molecules allow lymphocytes to be isolated from each other

The presence of characteristic surface molecules expressed by cell populations allows them to be identified using fluorescent antibodies as probes. These can be applied to tissue sections to identify cell populations or be used in flow cytometry to enumerate and separate cells in suspension on the basis of their size and fluorescent staining (see Method Box 2.1). These techniques together with the expression of surface molecules allowing the cell populations to be isolated from each other using cell panning and immunomagnetic beads (Method Box 2.2) have allowed a detailed dissection of lymphoid cell populations.

METHOD BOX 2.2
Isolating cell populations

Density-gradient separation of lymphocytes on Ficoll Isopaque

Lymphocytes can be separated from whole blood using a density gradient. Whole blood is defibrinated by shaking with glass beads and the resulting clot removed. The blood is then diluted in tissue culture medium and layered on top of a tube half full of Ficoll. Ficoll has a density greater than that of lymphocytes, but less than that of red cells and granulocytes (e.g. neutrophils).

After centrifugation the red cells and polymorphonuclear neutrophils (PMNs) pass down through the Ficoll to form a pellet at the bottom of the tube while lymphocytes settle at the interface of the medium and Ficoll.

The lymphocyte preparation can be further depleted of macrophages and residual PMNs by the addition of iron filings. These are taken up by phagocytes, which can then be drawn away with a strong magnet. Macrophages can also be removed by leaving the cell suspension to settle on a plastic dish. Macrophages adhere to plastic, whereas the lymphocytes can be washed off.

Isolating cell populations using their characteristic surface molecules

The presence of characteristic surface molecules expressed by cell populations allows the cell populations to be isolated from each other using cell panning and immunomagnetic beads.

Isolation of lymphocyte subpopulations – panning

Cell populations can be separated on antibody-sensitized plates. Antibody binds non-covalently to the plastic plate (as for solid-phase immunoassay) and the cell mixture is applied to the plate. Antigen-positive cells (Ag^+) bind to the antibody and the antigen-negative cells (Ag^-) can be carefully washed off.

By changing the culture conditions or by enzyme digestion of the cells on the plate, it is sometimes possible to recover the cells bound to the plate.

Often the cells that have bound to the plate are altered by their binding (e.g. binding to the plate cross-links the antigen, which can cause cell activation). The method is therefore most satisfactory for removing a subpopulation from the population, rather than isolating it.

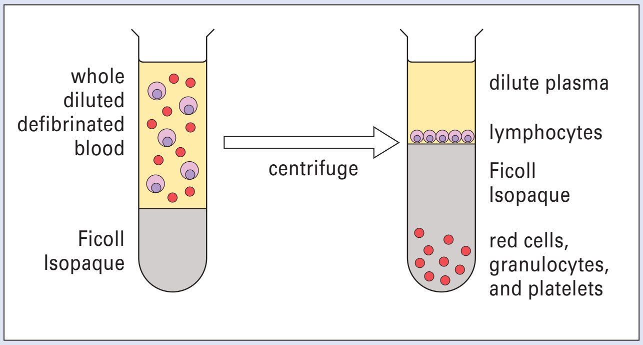
whole diluted defibrinated blood
Ficoll Isopaque
centrifuge
dilute plasma
lymphocytes
Ficoll Isopaque
red cells, granulocytes, and platelets

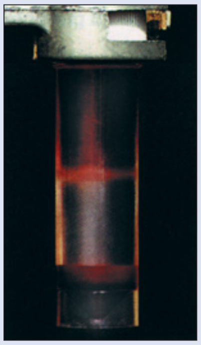

Fig. 1 Density-gradient separation of lymphocytes on Ficoll Isopaque.

METHOD BOX 2.2
Isolating cell populations *(Continued)*

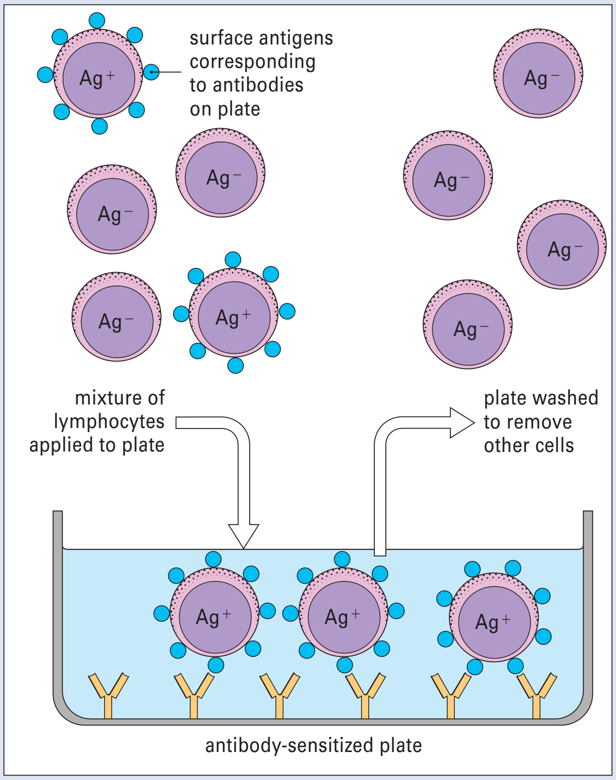
Ag+
surface antigens corresponding to antibodies on plate
Ag−
Ag−
Ag−
Ag−
Ag−
Ag−
Ag+
Ag−
mixture of lymphocytes applied to plate
plate washed to remove other cells
Ag+
Ag+
Ag+
antibody-sensitized plate

Fig. 2 Isolation of lymphocyte subpopulations – panning.

Examples of the application of this method include:

- separating TH and Tc cell populations using antibodies to CD4 or CD8; and
- separating T cells from B cells using anti-Ig (which binds to the surface antibody of the B cell).

In reverse, by sensitizing the plate with antigen, antigen-binding cells can be separated from non-binding cells.

Cell separation by immunomagnetic beads

Direct method (shown in Fig. 3)

The beads are coated with a monoclonal antibody to the cellular antigen of interest, by either:

- direct binding to the bead; or
- binding the primary antibody to secondary antibody-coated beads.

The coated beads are then incubated with the cell suspension (or even whole blood) and the cells bound by the antibody on the beads (positively selected cells) are immobilized by applying a magnetic field to the tube. The non-immobilized cells (negatively selected) are removed from the tube and the positively selected cells are recovered following washing and dissociation from the antibody-coated beads.

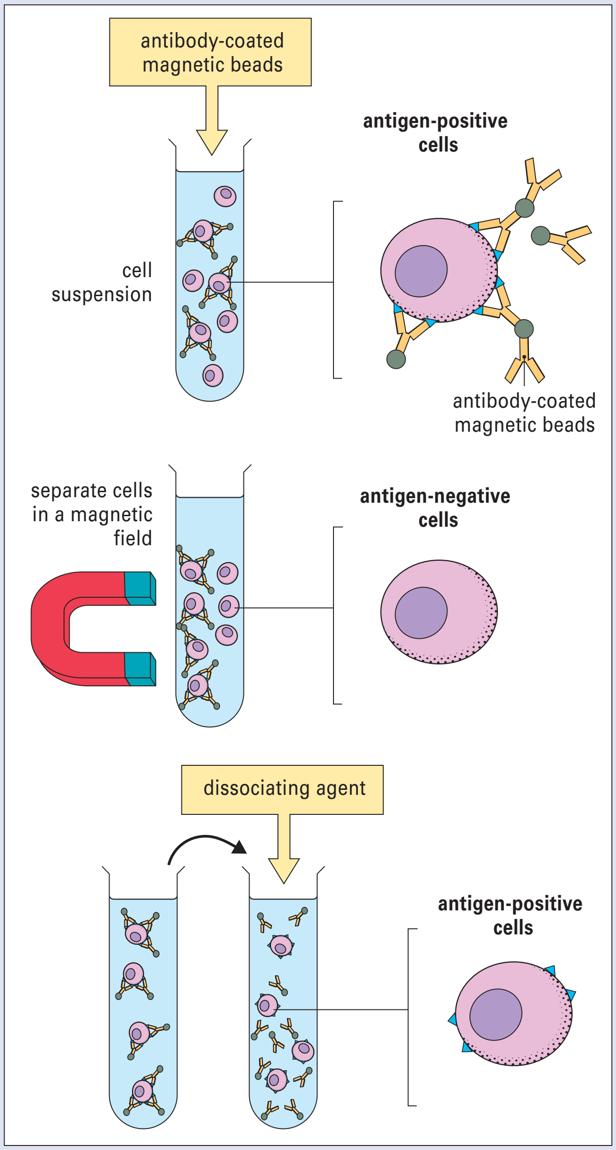
antibody-coated magnetic beads
antigen-positive cells
cell suspension
antibody-coated magnetic beads
separate cells in a magnetic field
antigen-negative cells
dissociating agent
antigen-positive cells

Fig. 3 Cell separation by immunomagnetic beads.

Indirect method

The monoclonal antibody to the target cellular antigen is first added to the cell suspension. Following incubation with the antibody, the cells are washed and mixed with beads coated with the appropriate secondary anti-Ig antibody. Cells bound to the magnetic beads (positively selected) are then immobilized with a magnetic field and the non-immobilized cells (negatively selected) removed. The positively selected cells are then washed and dissociated from the beads.

T cells can be distinguished by their different antigen receptors

The definitive T cell lineage marker is the T cell antigen receptor (TCR). The two defined types of TCR are:

- a heterodimer of two disulfide-linked polypeptides (α and β);
- a structurally similar heterodimer consisting of γ and δ polypeptides.

Both receptors are associated with a set of five polypeptides (the CD3 complex) and together form the TCR complex (TCR–CD3 complex; see Chapter 5).

Approximately 90–95% of blood T cells are αβ T cells and the remaining 5–10% are γδ T cells.

THERE ARE TWO MAJOR SUBPOPULATIONS OF T CELLS

αβ T cells are subdivided into two distinct non-overlapping populations:

- one subset carries the **CD4 marker (CD4⁺ T cells)** and mainly 'helps' or 'induces' immune responses (TH); and
- the other subset carries the **CD8 marker (CD8⁺ T cells)** and is predominantly cytotoxic (Tc).

CD4⁺ T cells recognize their specific antigens in association with MHC class II molecules, whereas CD8⁺ T cells recognize antigens in association with MHC class I molecules (see Chapter 7). Thus, the presence of CD4 or CD8 limits (restricts) the type of cell with which the T cell can interact (Fig. 2.21).

A small proportion of αβ T cells express neither CD4 nor CD8; these 'double negative' T cells might have a regulatory function.

Functional T cell subsets

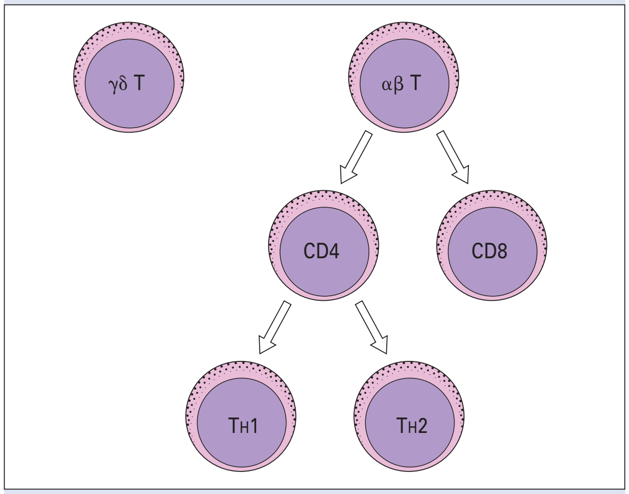

Fig. 2.21 T cells express either γδ or αβ T cell receptors (TCR). T cells are divided into CD4 and CD8 subsets and these subsets determine whether they see antigen (peptides) with MHC class II or I molecules, respectively. CD4⁺ T cells can be further subdivided into TH1 and TH2 on the basis of their cytokine profiles.

In contrast, while most circulating γδ cells are 'double negative', most γδ T cells in the tissues express CD8.

CD4⁺ T cell subsets are distinguished by their cytokine profiles

CD4⁺ T cells can be further divided into functional subsets on the basis of the spectrum of the cytokines they produce:

- TH1 cells secrete IL-2 and IFNγ;
- TH2 cells produce IL-4, IL-5, IL-6, and IL-10 (see Fig. 11.9).

TH1 cells mediate several functions associated with cytotoxicity and local inflammatory reactions. Consequently, they are important for combating intracellular pathogens including viruses, bacteria, and parasites.

Q. Which cell type do TH1 cells interact with to help combat intracellular pathogens?

A. Mononuclear phagocytes.

TH2 cells are more effective at stimulating B cells to proliferate and produce antibodies, and therefore function primarily to protect against free-living microorganisms (humoral immunity).

The number of cells producing a given cytokine can be measured using flow cytometry and antibodies that are allowed to penetrate the cells following permeabilization (see Method Box 2.1). The same technique can be used to determine the number of B cells producing a particular antibody.

The measurement of single cells secreting a particular cytokine or antibody can be achieved using an enzyme-linked method, namely ELISPOT (Method Box 2.1).

Other T cell subsets include γδ T cells and NKT cells

γδ T cells may protect the mucosal surfaces of the body

γδ T cells are relatively frequent in mucosal epithelia, but form only a minor subpopulation of circulating T cells (around 5%). Most intraepithelial lymphocytes (IELs) are γδ T cells and express CD8, a marker not found on most circulating γδ T cells.

γδ T cells have a specific repertoire of TCRs biased towards certain bacterial/viral antigens (**superantigens**, see Fig. 7.22).

Human blood γδ T cells have specificity for low molecular mass mycobacterial products (e.g. ethylamine and isopentenyl pyrophosphate).

Current opinion is that γδ T cells may play an important role in protecting the mucosal surfaces of the body.

Some γδ T cells may recognize antigens directly (i.e. with no need for MHC molecule-mediated presentation).

γδ T cells display LGL characteristics (see Fig. 2.19) and a dendritic morphology in lymphoid tissues (Fig. 2.22).

NKT cells may initiate T cell responses

NKT cells have T markers and also some NK cell markers. Characteristically they express CD3 and have a unique αβ TCR (expressing an invariant Vα and Vβ11, see Chapter 5).

Dendritic morphology of γδ T cells in the tonsil

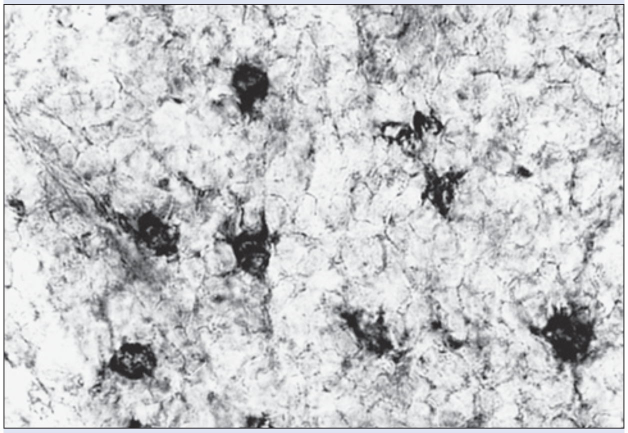

Fig. 2.22 The γδ T cell population is predominantly localized in the interfollicular T cell-dependent zones. Note the dendritic morphology of the cells. Anti-γδ T cell monoclonal antibody and immunoperoxidase. × 900. (Courtesy of Dr A Favre, from *Eur J Immunol* 1991;21:173, with permission)

Q. What markers would one normally use to distinguish T cells from NK cells?

A. CD16 and CD56 are used to distinguish NK cells. CD3 is characteristic of T cells.

NKT cells are thought to recognize glycolipid antigens presented by CD1d molecules (see Chapter 5), but not conventional MHC molecules. In response to antigen they are capable of producing large amounts of IFNγ and IL-4.

NKT cells are therefore thought to act as an interface between the innate and adaptive systems by initiating T cell responses.

NKT cells are also thought to regulate immune responses (especially dendritic cell function) through the production of cytokines (e.g. IL-10).

CELL SURFACE IMMUNOGLOBULIN AND SIGNALING MOLECULES FORM THE ‘B CELL RECEPTOR’ COMPLEX

About 5–15% of the circulating lymphoid pool are B cells, which are defined by the presence of surface immunoglobulin. Immunoglobulins are constitutively produced and are inserted into the B cell membrane, where they act as specific antigen receptors.

Most human B cells in peripheral blood express two immunoglobulin isotypes on their surface:

- IgM; and
- IgD (see Chapter 3).

On any B cell, the antigen-binding sites of these IgM and IgD isotypes are identical.

Fewer than 10% of the B cells in the circulation express IgG, IgA, or IgE, but B cells expressing IgG, IgA, or IgE are present in larger numbers in specific locations of the body (e.g. IgA-bearing cells in the intestinal mucosa).

Immunoglobulin associated with other ‘accessory’ molecules on the B cell surface forms the ‘B cell antigen receptor complex’ (BCR). These ‘accessory’ molecules consist of disulfide-bonded heterodimers of:

- Igα (CD79a); and
- Igβ (CD79b).

The heterodimers interact with the transmembrane segments of the immunoglobulin receptor (see Fig. 3.1), and, like the separate molecular components of the TCR/CD3 complex (see Fig. 5.2), are involved in cellular activation.

Other B cell markers include MHC class II antigens and complement and Fc receptors

Most B cells carry MHC class II antigens, which are important for cooperative (cognate) interactions with T cells. These class II molecules consist of I-A or I-E in the mouse and HLA-DP, DQ, and DR antigens in humans (see Fig. 5.18).

Complement receptors for C3b (CD35) and C3d (CD21) are commonly found on B cells and are associated with activation and, possibly, ‘homing’ of the cells. CD19/CD21 interactions with complement associated with antigen play a role in antigen-induced B cell activation via the antigen-binding antibody receptor.

Fc receptors for exogenous IgG (FcγRII, CD32) are also present on B cells and play a role in negative signaling to the B cell (see Chapter 11).

CD19 and CD20 are the main markers currently used to identify human B cells. Other human B cell markers are CD22 and CD72 to CD78.

Murine B cells also express CD72 (Lyb-2) together with B220, a high molecular weight (220 kDa) isoform of CD45 (Lyb-5).

CD40 is an important molecule on B cells and is involved in cognate interactions between T and B cells (see Fig. 8.8).

CD5$^+$ B-1 cells and marginal zone B cells produce natural antibodies

CD5$^+$ B-1 cells have a variety of roles

Many of the first B cells that appear during ontogeny express CD5, a marker originally found on T cells. These cells (termed B-1 cells) are found predominantly in the peritoneal cavity in mice, and there is some evidence for a separate differentiation pathway from ‘conventional’ B cells (termed **B-2 cells**).

CD5$^+$ B-1 cells express their immunoglobulins from unmutated or minimally mutated germline genes and produce mostly IgM, but also some IgG and IgA. These so-called natural antibodies are of low avidity, but, unusually, they are polyreactive and are found at high concentration in the adult serum. CD5$^+$ B-1 cells:

- respond well to TI (T-independent) antigens (i.e. antigens that can directly stimulate B cells without T cell help);
- may be involved in antigen processing and antigen presentation to T cells; and
- probably play a role in both tolerance and antibody responses.

Functions proposed for natural antibodies include:
- the first line of defense against microorganisms;
- clearance of damaged self components; and
- regulatory 'idiotype network' interactions within the immune system.

Characteristically, natural antibodies react against autoantigens including:
- DNA;
- Fc of IgG;
- phospholipids; and
- cytoskeletal components.

CD5 has been shown to be expressed by B-2 cells when they are activated appropriately, so there is some controversy about whether CD5 represents an activation antigen on B cells. Current theories therefore support the notion for two different kinds of $CD5^+$ B cells.

Although the function of CD5 on human B cells is unknown, it is associated with the BCR and may be involved in the regulation of B cell activation.

Marginal zone B cells are thought to protect against polysaccharide antigens

Much has been learned about **marginal zone B cells** over the past few years. These cells accumulate slowly in the marginal zone of the spleen – a process that takes between 1 and 2 years in humans.

Like B-1 cells, marginal zone B cells respond to thymus-independent antigens, and they are thought to be our main protection against polysaccharide antigens. They also produce natural antibodies, and together with B-1 cells have recently been termed 'innate-like B cells'.

B CELLS CAN DIFFERENTIATE INTO ANTIBODY-SECRETING PLASMA CELLS

Following B cell activation, many B cell blasts mature into **antibody-forming cells** (AFCs), which progress in vivo to terminally differentiated **plasma cells**.

Some B cell blasts do not develop rough endoplasmic reticulum cisternae. These cells are found in germinal centers and are named **follicle center cells** or **centrocytes**.

Under light microscopy, the cytoplasm of the plasma cells is basophilic due to the large amount of RNA being used for antibody synthesis in the rough endoplasmic reticulum. At the ultrastructural level, the rough endoplasmic reticulum can often be seen in parallel arrays (Fig. 2.23).

Plasma cells are infrequent in the blood, comprising less than 0.1% of circulating lymphocytes. They are normally restricted to the secondary lymphoid organs and tissues, but are also abundant in the bone marrow.

Antibodies produced by a single plasma cell are of one specificity and immunoglobulin class.

Immunoglobulins can be visualized in the plasma cell cytoplasm by staining with fluorochrome-labeled specific antibodies (Fig. 2.24).

Many plasma cells have a short life span, surviving for a few days and dying by apoptosis (Fig. 2.25). However, a subset of plasma cells with a long life span (months) has recently been described in the bone marrow.

Ultrastructure of the plasma cell

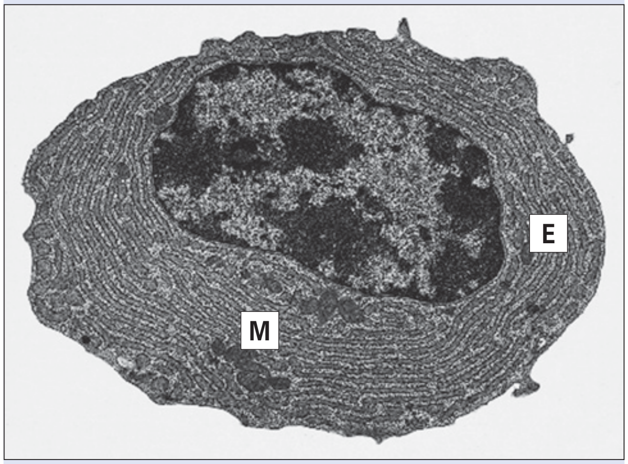

Fig 2.23 The plasma cell is characterized by parallel arrays of rough endoplasmic reticulum (E). In mature cells, these cisternae become dilated with immunoglobulins. Mitochondria (M) are also seen. × 5000. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

Immunofluorescent staining of intracytoplasmic immunoglobulin in plasma cells

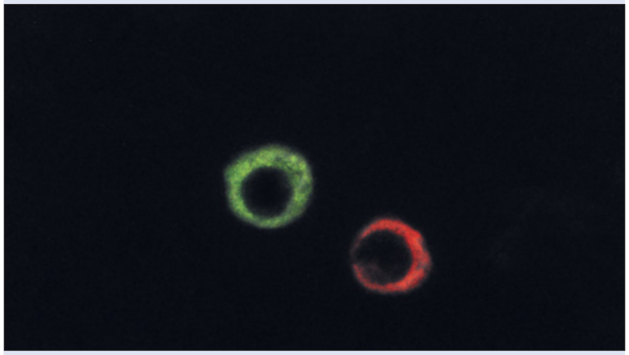

Fig. 2.24 Fixed human plasma cells, treated with fluoresceinated anti-human-IgM (green) and rhodaminated anti-human-IgG (red), show extensive intracytoplasmic staining. As the distinct staining of the two cells shows, plasma cells normally produce only one class or subclass (isotype) of antibody. × 1500. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

LYMPHOID ORGANS AND TISSUES ARE EITHER PRIMARY OR SECONDARY

Lymphocytes, the effector cells of the adaptive immune response, are the major component of organs and tissues that collectively form the lymphoid system.

Plasma cell death by apoptosis

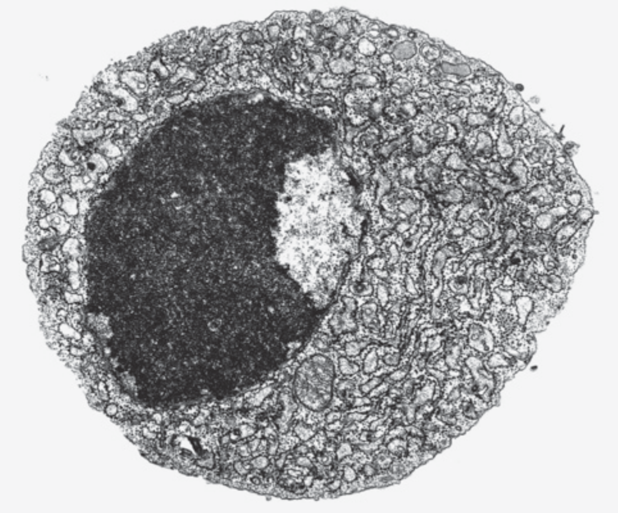

Fig. 2.25 Plasma cells are short-lived and die by apoptosis (cell suicide). Note the nuclear chromatin changes, which are characteristic of apoptosis. × 5000.

Major lymphoid organs and tissues

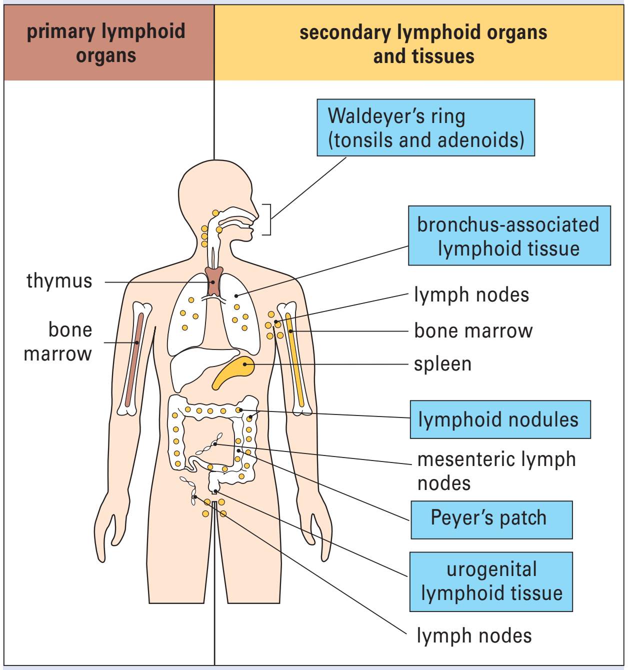
primary lymphoid organs
secondary lymphoid organs and tissues
Waldeyer's ring (tonsils and adenoids)
bronchus-associated lymphoid tissue
thymus
lymph nodes
bone marrow
bone marrow
spleen
lymphoid nodules
mesenteric lymph nodes
Peyer's patch
urogenital lymphoid tissue
lymph nodes

Fig. 2.26 Thymus and bone marrow are the primary (central) lymphoid organs. They are the sites of maturation for T and B cells, respectively. Cellular and humoral immune responses occur in the secondary (peripheral) lymphoid organs and tissues. Secondary lymphoid organs can be classified according to the body regions they defend. The spleen responds predominantly to blood-borne antigens. Lymph nodes mount immune responses to antigens circulating in the lymph, entering through the skin (subcutaneous lymph nodes) or through mucosal surfaces (visceral lymph nodes). Tonsils, Peyer's patches, and other mucosa-associated lymphoid tissues (MALT) (blue boxes) react to antigens that have entered via the surface mucosal barriers. Note that the bone marrow is both a primary and a secondary lymphoid organ because it gives rise to B and NK cells, but is also the site of B cell terminal differentiation (long-lived plasma cells).

Within the lymphoid organs, lymphocytes interact with other cell types of both hematopoietic and non-hematopoietic origin that are important for lymphocyte:

- maturation;
- selection;
- function; and
- disposal of terminally differentiated cells.

These other cell types are termed accessory cells and include:

- APCs;
- macrophages;
- reticular cells; and
- epithelial cells.

The lymphoid system is arranged into either discrete encapsulated organs or accumulations of diffuse lymphoid tissue, which are classified into primary (central) and secondary (peripheral) organs or tissues (Fig. 2.26).

In essence, lymphocytes:

- are produced, mature, and are selected in primary lymphoid organs; and
- exert their effector functions in the secondary lymphoid organs and tissues.

Tertiary lymphoid tissues are anatomical sites that under normal conditions contain sparse lymphocytes, if any, but may be selectively populated by these cells in pathological conditions (e.g. skin, synovium, and lungs).

LYMPHOID STEM CELLS DEVELOP AND MATURE WITHIN PRIMARY LYMPHOID ORGANS

In the primary lymphoid organs, lymphocytes (B and T cells):

- differentiate from lymphoid stem cells;
- proliferate;
- are selected; and
- mature into functional cells.

In mammals, T cells mature in the thymus and B cells mature in the fetal liver and postnatal bone marrow (see Chapter 8). Birds have a specialized site of B cell generation, the bursa of Fabricius.

In the primary lymphoid organs:

- lymphocytes acquire their repertoire of specific antigen receptors to cope with the antigenic challenges that individuals encounter during their lifetime;
- cells with receptors for **autoantigens** are mostly eliminated; and
- in the thymus, T cells also 'learn' to recognize appropriate **self MHC molecules**.

There is evidence that some lymphocyte development might occur outside primary lymphoid organs.

Thymus section showing the lobular organization

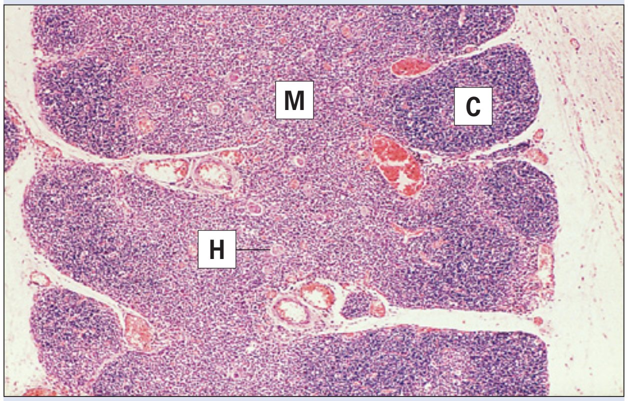

Fig. 2.27 This section shows the two main areas of the thymus lobule – an outer cortex of immature cells (C) and an inner medulla of more mature cells (M). Hassall's corpuscles (H) are found in the medulla. H&E stain. × 25. (Courtesy of Dr A Stevens and Professor J Lowe)

Q. Why do lymphocytes need to 'learn' what constitutes self MHC and self antigens?

A. Each individual is different and has a particular set of MHC molecules and particular variants of the many other molecules present in the body. The process of what constitutes immunological 'self' is different in each individual and therefore learning to recognize 'self' is a dialogue between T cells and APCs that takes place in each individual.

T CELLS DEVELOPING IN THE THYMUS UNDERGO SELECTION PROCESSES

The thymus in mammals is a bilobed organ in the thoracic cavity overlying the heart and major blood vessels. Each lobe is organized into lobules separated from each other by connective tissue trabeculae.

Within each lobule, the lymphoid cells (thymocytes) are arranged into:

- an outer tightly packed cortex, which contains the majority of relatively immature proliferating thymocytes; and
- an inner medulla containing more mature cells, implying a differentiation gradient from cortex to medulla (Fig. 2.27).

The main blood vessels that regulate cell traffic in the thymus are high endothelial venules (HEVs; see Fig. 2.29) at the corticomedullary junction of thymic lobules. It is through these veins that T cell progenitors formed in the fetal liver and bone marrow enter the **epithelial anlage** and migrate towards the cortex.

In the cortex of the thymus the T cell progenitors undergo proliferation and differentiation processes that lead to the generation of mature T cells through a corticomedullary gradient of migration.

A network of epithelial cells throughout the lobules plays a role in the differentiation and selection processes from fetal liver and bone marrow-derived prethymic cells to mature T cells.

The mature T cells probably leave the thymus through the same PCVs, at the corticomedullary junction from which the T cell progenitors entered (Fig. 2.28).

Cell migration to and within the thymus

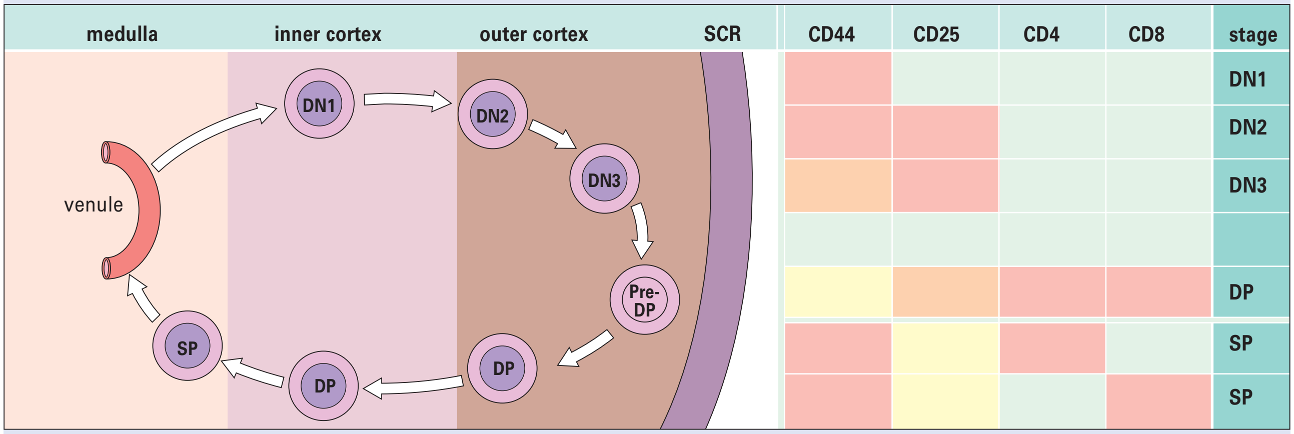

Fig. 2.28 T cell progenitors enter the thymic lobule through postcapillary venules (PCVs) at the corticomedullary junction. These double negative 1 (DN1) cells are CD4$^-$, CD8$^-$, CD25$^-$, but CD44$^+$. They move progressively towards the outer cortex and differentiate into DN2 (CD25$^+$, CD44$^+$) and DN3 cells (CD25$^+$, CD44^{lo}). Thymocytes accumulate in the subcapsular region where they actively proliferate and differentiate into double positive (DP; CD4$^+$, CD8$^+$) cells. DP thymocytes reverse their polarity and move towards the medulla. In the course of this migration, thymocytes are selected and as single positive (SP; CD4$^+$ or CD8$^+$) cells ultimately leave the thymus, presumably via HEVs at the corticomedullary junction.

Three types of thymic epithelial cell have important roles in T cell production

At least three types of epithelial cell can be distinguished in the thymic lobules according to distribution, structure, function, and phenotype:

- the epithelial nurse cells are in the outer cortex;
- the cortical thymic epithelial cells (TECs) form an epithelial network; and
- the medullary TECs are mostly organized into clusters (Fig. 2.29).

These three types of epithelial cell have different roles for thymocyte proliferation, maturation, and selection:

- **nurse cells** in the outer cortex sustain the proliferation of progenitor T cells, mainly through cytokine production (e.g. IL-7);
- **cortical TECs** are responsible for the positive selection of maturing thymocytes, allowing survival of cells that recognize MHC class I and II molecules with associated peptides via TCRs of intermediate affinity;
- **medullary TECs** display a large variety of organ-specific self peptides.

Q. What is the significance of the presence of organ-specific self peptides in the thymus?

A. An individual needs to be tolerant of antigens that are expressed in other tissues, not just the thymus. By presenting a library of self molecules, the thymus can delete or tolerize lymphocytes that might otherwise react against self molecules once they had migrated into other tissues. Thus TECs, together with other APCs (interdigitating cells and macrophages), play a role in negative selection (i.e. the deletion of self-reactive T cells).

Hassall's corpuscles (see Fig. 2.27) are found in the thymic medulla. Their function is unknown, but they appear to contain degenerating epithelial cells rich in high molecular weight cytokeratins.

The mammalian thymus involutes with age (Fig. 2.30). In humans, atrophy begins at puberty and continues throughout life. Thymic involution begins within the cortex and this region may disappear completely, whereas medullary remnants persist.

Cortical atrophy is related to a sensitivity of the cortical thymocytes to corticosteroid, and all conditions associated

Schematic structure of the thymus

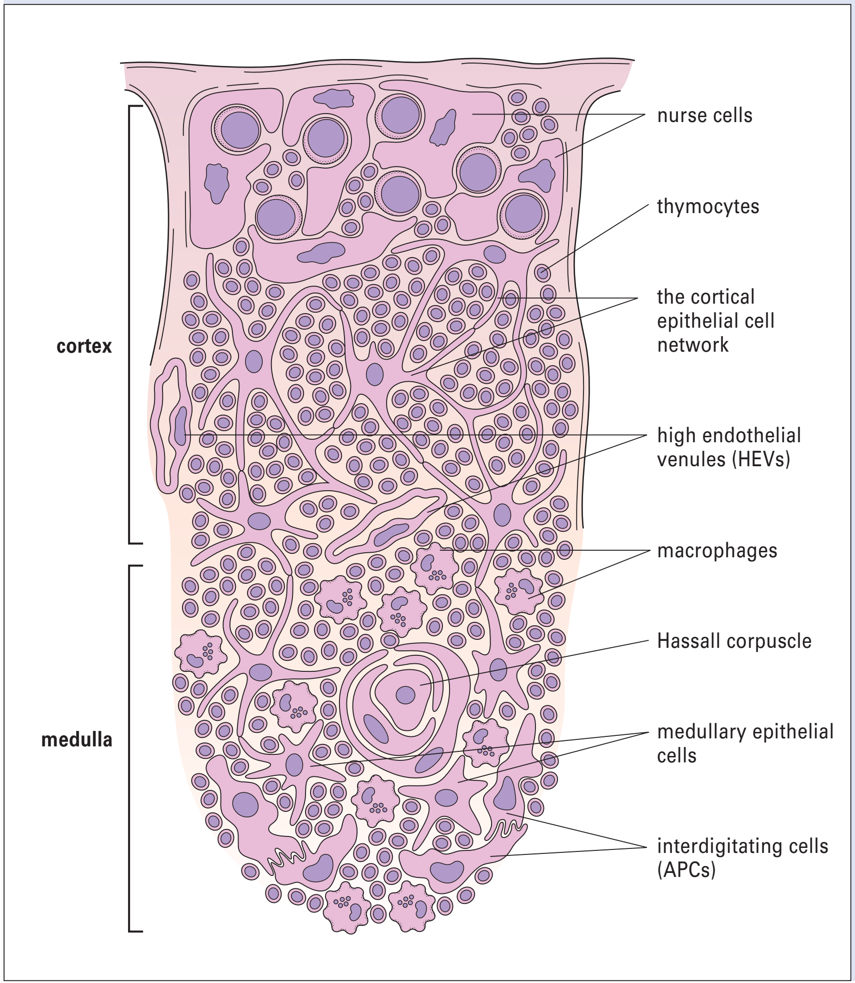
nurse cells
thymocytes
the cortical epithelial cell network
high endothelial venules (HEVs)
macrophages
Hassall corpuscle
medullary epithelial cells
interdigitating cells (APCs)
cortex
medulla

Fig. 2.29 A schematic representation of the cell types found in a fully developed thymic lobule. Subcapsular epithelial cells that produce IL-7 (nurse cells) sustain T lymphoblast proliferation in the outer cortex. Developing T cells interact with the cortical epithelial network where they are positively selected. Apoptotic cells are phagocytosed by macrophages present in the deep cortex and in the medulla. TCR$^+$ thymocytes co-expressing CD4 and CD8 undergo the process of negative selection by interacting with a variety of antigen-presenting cells (APCs), such as dendritic cells, interdigitating cells, macrophages, and epithelial cells. T cells that have survived the selection processes are exported from the thymus via high endothelial venules (HEVs) and lymphatic vessels. (From Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

Atrophic adult thymus

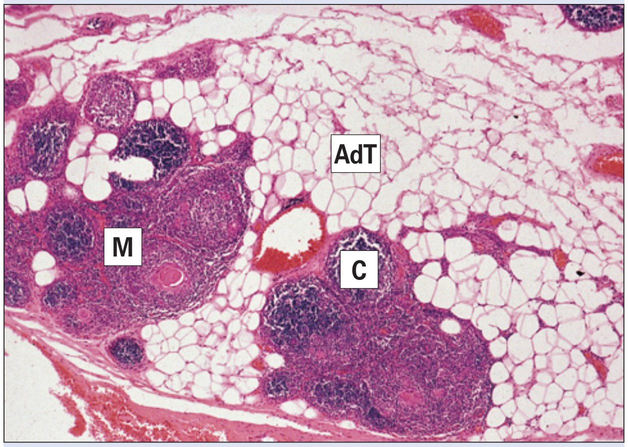

Fig 2.30 There is an involution of the thymus with replacement by adipose tissue (AdT). The cortex (C) is largely reduced and the less cellular medulla (M) is still apparent. (Courtesy of Dr A Stevens and Professor J Lowe)

with an acute increase in corticosteroids (e.g. pregnancy and stress) promote thymic atrophy.

It is conceivable that T cell generation within the thymus continues into adult life, albeit at a low rate. Evidence for de-novo T cell production in the thymus (recent thymic emigrants) has been shown in humans over the age of 76 years.

Stem cell migration to the thymus initiates T cell development

The thymus develops from the endoderm of the third pharyngeal pouch as an epithelial rudiment that becomes seeded with blood-borne stem cells. Relatively few stem cells appear to be needed to give rise to the enormous repertoire of mature T cells with diverse antigen receptor specificities.

From experimental studies, migration of stem cells into the thymus is not a random process, but results from chemotactic signals periodically emitted from the thymic rudiment. β_2-Microglobulin, a component of the MHC class I molecule, is one such putative chemoattractant.

In birds, stem cells enter the thymus in two or possibly three waves, but it is not clear that there are such waves in mammals.

Once in the thymus, the stem cells begin to differentiate into thymic lymphocytes (called thymocytes), under the influence of the epithelial microenvironment.

Whether or not the stem cells are ‘pre-T cells’ (i.e. are committed to becoming T cells before they arrive in the thymus) is controversial. Although the stem cells express CD7, substantial evidence exists that they are in fact multipotent. Granulocytes, APCs, NK cells, B cells, and myeloid cells have all been generated in vitro from hematopoietic precursors isolated from the thymus. This suggests that the prethymic bone marrow-derived cell entering the thymic rudiment is multipotent.

Epithelial cells, macrophages, and bone marrow-derived IDCs, molecules rich in MHC class II, are important for the differentiation of T cells from this multipotent stem cell. For example, specialized epithelial cells in the peripheral areas of the cortex (the thymic nurse cells, see above) contain thymocytes within pockets in their cytoplasm. The nurse cells support lymphocyte proliferation by producing the cytokine IL-7.

The subcapsular region of the thymus is the only site where thymocyte proliferation occurs. Thymocytes develop into large, actively proliferating, self-renewing lymphoblasts, which generate the thymocyte population.

There are many more developing lymphocytes (85–90%) in the thymic cortex than in the medulla, and studies of function and cell surface markers have indicated that cortical thymocytes are less mature than medullary thymocytes. This reflects the fact that cortical cells migrate to, and mature in, the medulla.

Most mature T cells leave the thymus via HEVs at the corticomedullary junction, though other routes of exit may exist, including lymphatic vessels.

T cells change their phenotype during maturation

As with the development of granulocytes and monocytes, ‘differentiation’ markers of functional significance appear or are lost during the progression from stem cell to mature T cell.

Analyses of genes encoding αβ and γδ TCRs and other studies examining changes in surface membrane antigens suggest that there are multiple pathways of T cell differentiation in the thymus. It is not known whether these pathways are distinct, but it seems more likely that they diverge from a common pathway.

Only a small proportion (< 1%) of mature T lymphocytes express the γδ TCR. Most thymocytes differentiate into αβ TCR cells, which account for the majority (> 95%) of T lymphocytes in secondary lymphoid tissues and in the circulation.

Phenotypic analyses have shown sequential changes in surface membrane antigens during T cell maturation (Fig. 2.31). The phenotypic variations can be simplified into a three-stage model.

Stage I thymocytes are CD4$^-$, CD8$^-$

There are two phases of stage I (early) thymocytes. In the first phase, the TCR genes are in the germline configuration and the cells:

- express CD44 and CD25;
- are CD4$^-$, CD8$^-$ (i.e. double negative cells).

In this early first phase, cells entering the thymus via the HEVs in the corticomedullary junction express CD44, which allows them to migrate towards the outermost cortex, the zone of thymocyte proliferation. These cells are not fully committed to the T cell lineage, because outside the thymic environment they can give rise to other hematopoietic lineages. Surface expression of CD44 is downregulated once the cells are in the external cortex.

In the second phase the cells:

- become CD44$^-$;
- are CD25$^+$;
- remain double negative for CD4 and CD8;
- rearrange the β chain of the TCR;

Expression of human T cell markers during development

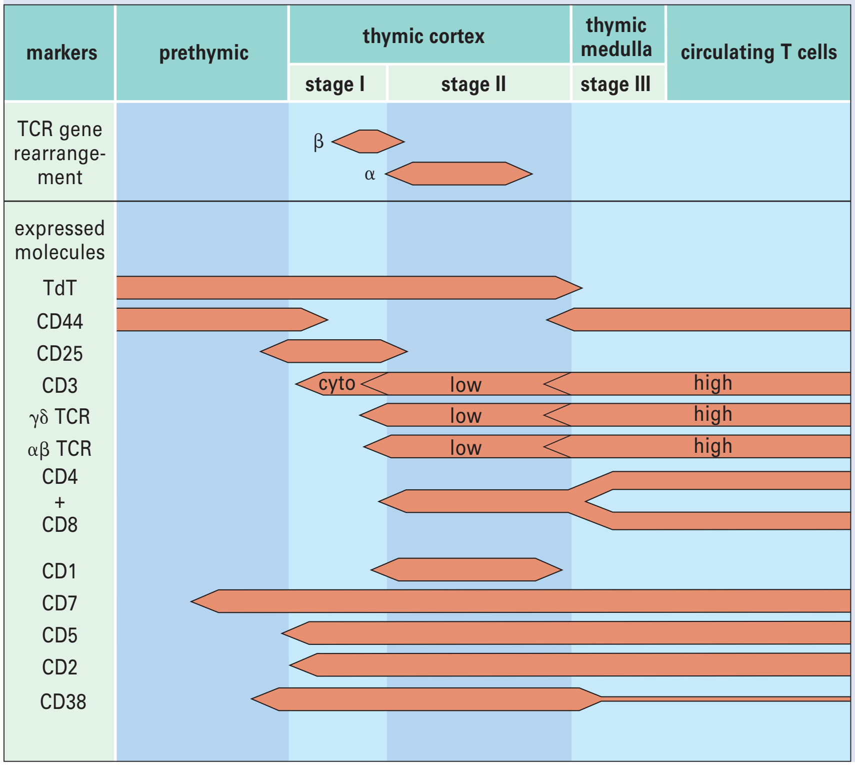

Fig. 2.31 Terminal deoxynucleotidyl transferase (TdT) is an enzyme in thymic stem cells. It decreases in stage II and is lost altogether in the medulla. Several surface glycoproteins appear during differentiation. CD1 is present on stage II cortical thymocytes and is lost in the medulla. CD2 and CD7 (the pan-T marker) appear very early in differentiation and are maintained through to the mature T cell stage. CD5 appears at an early stage and persists on mature T cells. CD3 is expressed first in the cytoplasm in stage I cells (cyto), and then on the surface simultaneously with the T cell receptor (TCR). In most stage II cells, both surface CD3 and the αβ TCR are expressed at low density, but these markers are present at high density on stage III cells. CD4 and CD8 are co-expressed on stage II cells (double positives). One of these molecules is lost during differentiation into mature stage III cells (single positives).

- express cytoplasmic but not surface TCR-associated CD3;
- are irreversibly committed to become T cells;
- continue to express CD7 together with CD2 and CD5.

Proliferation markers such as the transferrin receptor (CD71) and CD38 (a marker common to all early hematopoietic precursors) are also expressed at this stage.

Stage II thymocytes become $CD4^+$, $CD8^+$

Stage II (intermediate or common) thymocyte cells account for around 80% of thymocytes in the fully developed thymus. Characteristically they:

- are $CD1^+$, $CD44^-$, $CD25^-$;
- become CD4+, CD8+ (double positives).

Genes encoding the TCR α chain are rearranged in these intermediate thymocytes; both chains of the αβ TCR are expressed at low density on the cell surface in association with polypeptides of the CD3/antigen receptor complex.

Stage III thymocytes become either $CD4^+$ or $CD8^+$

Stage III (mature) thymocytes show major phenotypic changes, namely:

- loss of CD1;
- cell surface CD3 associated with the αβ TCR expressed at a higher density;
- the distinction of two subsets of cells expressing either CD4 or CD8 (i.e. single positives).

Most stage III thymocytes:

- lack CD38 and the transferrin receptor;
- are virtually indistinguishable from mature, circulating T cells.

All stage III cells re-express the receptor CD44, which is thought to be involved in migration and homing to peripheral lymphoid tissues. L-selectin (CD62L) is also expressed at this time.

DIVERSE ANTIGEN REPERTOIRES ARE GENERATED BY RECOMBINATION OF GENE SEGMENTS ENCODING THE TCR

T cell receptor diversity is generated in the thymus

T cells have to recognize a wide variety of different antigens. The genes of the αβ and γδ TCR undergo **somatic recombination** during thymic development to produce functional genes for the different TCRs (see Chapter 5).

Q. How are receptors with different antigenic specificity distributed among the total population of T cells?

A. Each T cell has one specificity of receptor (i.e. the receptors are clonally distributed).

Q. Are there any differences in the mechanism by which TCRs are generated in $CD4^+$ and $CD8^+$ T cells?

A. Because $CD4^+$ and $CD8^+$ T cells do not differentiate until after TCR generation, the mechanism for generating diverse receptors is the same for both major T cell subsets.

TCR gene recombination takes place within the subcapsular and outer cortex of the thymus, where there is active cell proliferation. Through a random assortment of different gene segments, a large number of different TCRs are made and thymocytes that fail to make a functional receptor die.

The TCRs associate with peptides of the CD3 complex, which transduces activating signals to the cell.

'Alternative' forms of the T cell receptor during development

Studies in transgenic mice have shown that early in ontogeny T cells can express alternative forms of the TCR which may be involved in generating signals that drive development:

- TCRβ dimers can associate with CD3 in the absence of TCRα;
- TCRβ chains can be found on the cell surface as phosphoinositol linked proteins that do not associate with CD3;
- surface TCRβ chains can be associated with an incomplete CD3 complex;
- a 33-kDa glycoprotein linked to TCRβ chains in pre-T cells acts as a 'surrogate' chain (pre-Tα chain).

These pre-T cell antigen receptors, like the surrogate pre-B cell receptors (see Fig. 8.2), are probably involved in proliferation, maturation, and selection during the early phases of lymphocyte development.

POSITIVE AND NEGATIVE SELECTION OF DEVELOPING T CELLS IN THE THYMUS

The processes involved in the education of T cells are shown in Fig. 2.32, and self tolerance is discussed fully in Chapter 11.

Positive selection ensures only TCRs with an intermediate affinity for self MHC develop further

T cells:

- recognize antigenic peptides only when presented by self MHC molecules on APCs;
- show 'dual recognition' of both the antigenic peptides and the polymorphic part of the MHC molecules.

CD4, found on a subset of T cells, also recognizes the class II MHC molecule, but its non-polymorphic portion.

Positive selection (the first stage of **thymic education**) ensures that only those TCRs with an intermediate affinity for self MHC are allowed to develop further.

T cell differentiation within the thymus

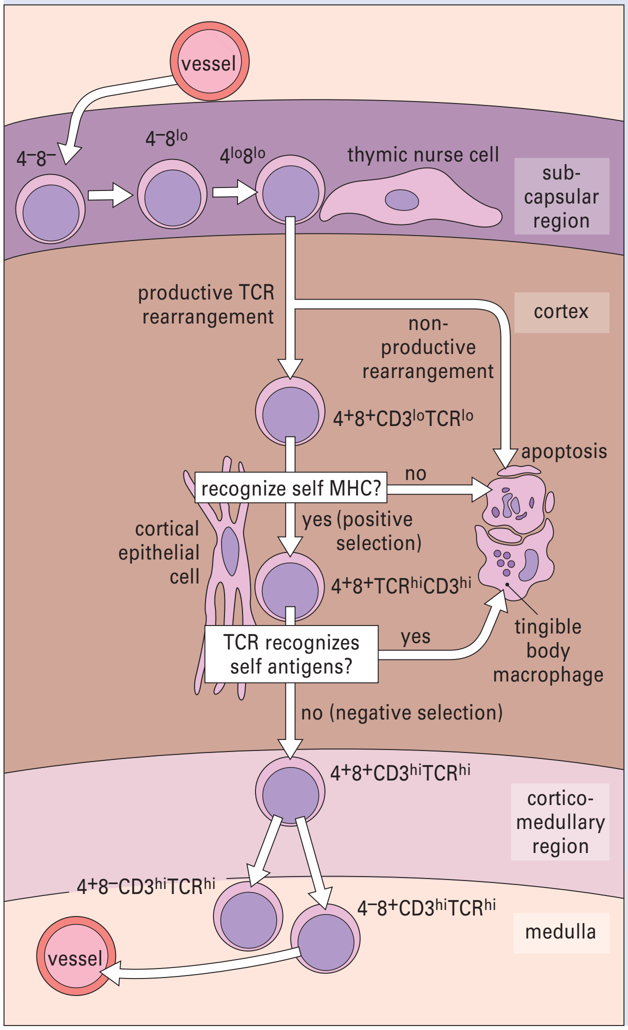
vessel
4^-8^-
4^-8^{lo}
$4^{lo}8^{lo}$
thymic nurse cell
sub-capsular region
productive TCR rearrangement
cortex
non-productive rearrangement
$4^+8^+\text{CD3}^{lo}\text{TCR}^{lo}$
apoptosis
recognize self MHC?
no
yes (positive selection)
cortical epithelial cell
$4^+8^+\text{TCR}^{hi}\text{CD3}^{hi}$
TCR recognizes self antigens?
yes
tingible body macrophage
no (negative selection)
$4^+8^+\text{CD3}^{hi}\text{TCR}^{hi}$
cortico-medullary region
$4^+8^-\text{CD3}^{hi}\text{TCR}^{hi}$
$4^-8^+\text{CD3}^{hi}\text{TCR}^{hi}$
medulla
vessel

Fig. 2.32 In this model, pre-thymic T cells are attracted to and enter the thymic rudiment at the corticomedullary junction. They reach the subcapsular region where they proliferate as large lymphoblasts, which give rise to a pool of cells entering the differentiation pathway. Many of these cells are associated with epithelial thymic nurse cells. Cells in this region first acquire CD8 and then CD4 at low density. They also rearrange their T cell receptor (TCR) genes and may express the products of these genes at low density on the cell surface. Maturing cells move deeper into the cortex and adhere to cortical epithelial cells. These epithelial cells are elongated and branched, and thus provide a large surface area for contact with thymocytes. The TCRs on the thymocytes are exposed to epithelial MHC molecules through these contacts. This leads to positive selection. Those cells that are not selected undergo apoptosis and are phagocytosed by macrophages. There is an increased expression of CD3, TCR, CD4, and CD8 during thymocyte migration from the subcapsular region to the deeper cortex. Those TCRs with self reactivity are now deleted through contact with autoantigens presented by medullary thymic epithelial cells, interdigitating cells, and macrophages at the corticomedullary junction – a process called negative selection. Following this stage, cells expressing either CD4 or CD8 appear and exit to the periphery via specialized vessels at the corticomedullary junction. (Adapted from D Zucker-Franklin, CE Grossi, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

There is evidence that positive selection is mediated by TECs acting as APCs.

T cells displaying very high or very low receptor affinities for self MHC undergo apoptosis and die in the cortex. Apoptosis is a pre-programmed 'suicide', achieved by activating endogenous nucleases that cause DNA fragmentation (Fig. 2.33).

T cells with TCRs that have intermediate affinities are rescued from apoptosis, survive, and continue along their pathway of maturation. A possible exception is provided by some T cells equipped with γδ receptors, which (like B cells) recognize native antigenic conformations with no need for APCs.

Negative selection ensures that only T cells that fail to recognize self antigen proceed in their development

Some of the positively selected T cells may have TCRs that recognize self components other than self MHC. These cells are deleted by a 'negative selection' process, which occurs:

- in the deeper cortex;
- at the corticomedullary junction; and
- in the medulla.

T cells interact with antigen presented by interdigitating cells, macrophages, and medullary TECs. The role of medullary TECs for negative selection has been emphasized recently by the finding that these cells express genes for virtually all tissue antigens in the body.

Only T cells that fail to recognize self antigen are allowed to proceed in their development. The rest undergo apoptosis and are destroyed. These, and all the other apoptotic cells generated in the thymus, are phagocytosed by (tingible body) macrophages (see Fig. 2.47) in the deep cortex.

T cells at this stage of maturation (CD4$^+$ CD8$^+$ TCR^{lo}) go on to express TCR at high density and lose either CD4 or CD8 to become 'single positive' mature T cells.

The separate subsets of CD4$^+$ and CD8$^+$ cells possess specialized homing receptors (e.g. CD44), and exit to the T cell areas of the peripheral (secondary) lymphoid tissues where they function as mature 'helper' and 'cytotoxic' T cells, respectively.

Thymic cell apoptosis

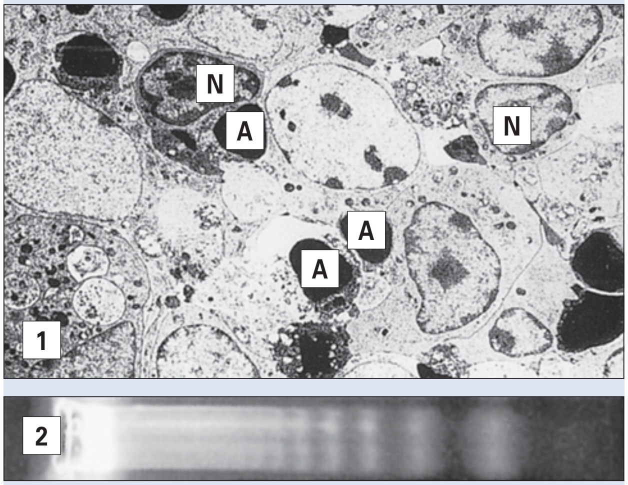

Fig. 2.33 (**1**) Fetal thymic lobes in culture were treated with anti-CD3 antibodies – this simulates activation via the TCR and therefore triggers programmed cell death (apoptosis). This electron micrograph shows the heavy condensation of nuclear chromatin in apoptotic nuclei (A) compared with the dispersed chromatin of normal cells (N). (Courtesy of Dr C Smith) (**2**) Analysis of the DNA from apoptotic cells by agarose gel electrophoresis shows the characteristically ordered, ladder-like pattern created by bands of digested DNA fragments.

Q. Which subset of T cells functions as TH and which as Tc cells?

A. CD4$^+$ T cells function mainly as TH cells whereas CD8$^+$ T cells are predominantly Tc cells.

Less than 5% of thymocytes leave the thymus as mature T cells. The rest die as the result of:

- selection processes; or
- failure to undergo productive rearrangements of antigen receptor genes.

Adhesion of maturing thymocytes to epithelial and accessory cells is crucial for T cell development

Adhesion of maturing thymocytes to epithelial and other accessory cells is mediated by the interaction of complementary adhesion molecules, such as:

- CD2 with LFA-3 (CD58); and
- LFA-1 (CD11a, CD18) with ICAM-1 (CD54) (see Appendix 2).

These interactions induce the production of the cytokines IL-1, IL-3, IL-6, IL-7, and GM-CSF (see Appendix 3), which are required for T cell proliferation and maturation in the thymus.

Early thymocytes also express receptors for IL-2, which together with IL-7 sustains cell proliferation.

Negative selection may also occur outside the thymus in peripheral lymphoid tissues

Not all self-reactive T cells are eliminated during intrathymic development, probably because not all self antigens can be presented in the thymus. The thymic epithelial barrier that surrounds blood vessels may also limit access of some circulating antigens.

Given the survival of some self-reacting T cells, a separate mechanism is required to prevent them attacking the body. Experiments with transgenic mice have suggested that peripheral inactivation of self-reactive T cells (**peripheral tolerance**, see Chapter 19) could occur via several mechanisms as follows:

- downregulation of the TCR and CD8 (in cytotoxic cells) so that the cells are unable to interact with target autoantigens;
- **anergy**, due to the lack of crucial co-stimulatory signals provided by the target cells, followed by induction of apoptosis after interaction with autoantigen;
- regulatory T cells (Tregs).

Regulatory T cells are involved in peripheral tolerance

Tregs have been the subject of intensive research over the past few years, especially in the areas of autoimmunity and vaccine development.

In addition to NKT cells and γδ T cells regulating immune responses, there is now substantial evidence that separate CD4⁺ subsets also have this function. The general consensus is that there are two main types of Treg – naturally occurring and antigen induced.

Naturally occurring Tregs:

- constitutively express CD25 (the α chain of the low-affinity receptor for IL-2);
- constitute about 5–10% of the peripheral CD4⁺ T cells;
- express the unique transcription factor FoxP3;
- constitutively express the marker CTLA4;
- do not proliferate in response to antigenic challenge;
- are thought to produce their suppressive effects through cell contact (e.g. with APCs, TH1 or TH2 cells).

Antigen-induced Tregs:

- also express CD25;
- can develop from CD25⁻, CD4⁺ T cells;
- are believed to exert their suppressive effects through cytokines such as TGFβ and IL-10.

There is some evidence for extrathymic development of T cells

The vast majority of T cells require a functioning thymus for differentiation, but small numbers of cells carrying T cell markers that are often oligoclonal in nature have been found in athymic (‘nude’) mice. Although the possibility that these mice possess thymic remnants cannot be ruled out, there is accumulating evidence to suggest that bone marrow precursors can home to mucosal epithelia and mature without the need for a thymus to form:

- functional T cells with γδ TCRs; and
- probably also T cells with αβ TCRs.

The importance of extrathymic development in animals that are euthymic (i.e. that have a normal thymus) is at present unclear.

B CELLS DEVELOP MAINLY IN THE FETAL LIVER AND BONE MARROW

Unlike birds, which have a discrete organ for the generation of B cells (the bursa of Fabricius), in mammals B cells develop directly from lymphoid stem cells in the hematopoietic tissue of the fetal liver (Fig. 2.34). This occurs at 8–9 weeks of gestation in humans, and by about 14 days in the mouse. Later, the site of B cell production moves from the liver to the bone marrow, where it continues through adult life. This fetal liver-bone marrow migration of stem cells is also true for cells of other hematopoietic lineages such as erythrocytes, granulocytes, monocytes, and platelets.

B cell progenitors are also present in the omental tissue of murine and human fetuses and are the precursors of a self-replicating B cell subset, the B-1 cells (see above).

Hemopoiesis in fetal liver

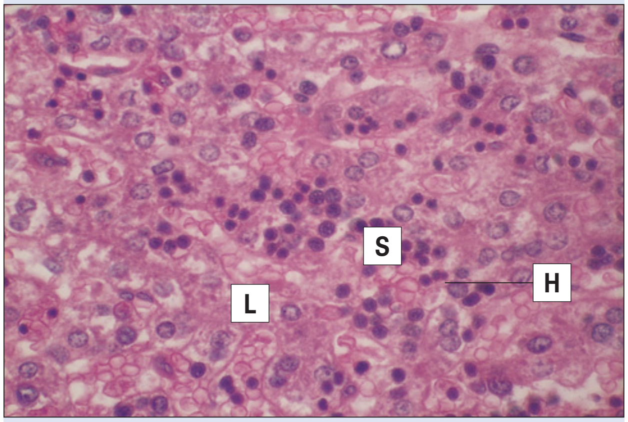

Fig. 2.34 Section of human fetal liver showing islands of hemopoiesis (H). Hematopoietic stem cells are found in the sinusoidal spaces (S) between plates of liver cells (L). (Courtesy of Dr A Stevens and Professor J Lowe)

B cell production in the bone marrow does not occur in distinct domains

B cell progenitors in the bone marrow are seen adjacent to the endosteum of the bone lamellae (Fig. 2.35). Each B cell progenitor at the stage of immunoglobulin gene rearrangement may produce up to 64 progeny. The progeny migrate towards the center of each cavity of the spongy bone and reach the lumen of a venous sinusoid (Fig. 2.36).

In the bone marrow, B cells mature in close association with **stromal reticular cells**, which are found both adjacent to the endosteum and in close association with the central sinus, where they are termed **adventitial reticular cells**.

Where the B cells differentiate, the reticular cells have mixed phenotypic features with some similarities to fibroblasts, endothelial cells, and myofibroblasts. The reticular cells produce type IV collagen, laminin and the smooth muscle form of actin. Experiments in vitro have shown that reticular cells sustain B cell differentiation, possibly by producing the cytokine IL-7.

Adventitial reticular cells may be important for the release of mature B cells into the central sinus.

B cells are subject to selection processes.

Most B cells (> 75%) maturing in the bone marrow do not reach the circulation, but (like thymocytes) undergo a process of programmed cell death (apoptosis) and are phagocytosed by bone marrow macrophages.

Q. By analogy with T cell development, infer what determines whether a B cell will die during development in the bone marrow.

A. B cells with a non-productive rearrangement of the immunoglobulin genes do not survive and many self-reactive B cells are also eliminated through negative selection.

B cell–stromal cell interactions enhance the survival of developing B cells and mediate a form of selection that

Schematic organization of B cell development in the bone marrow

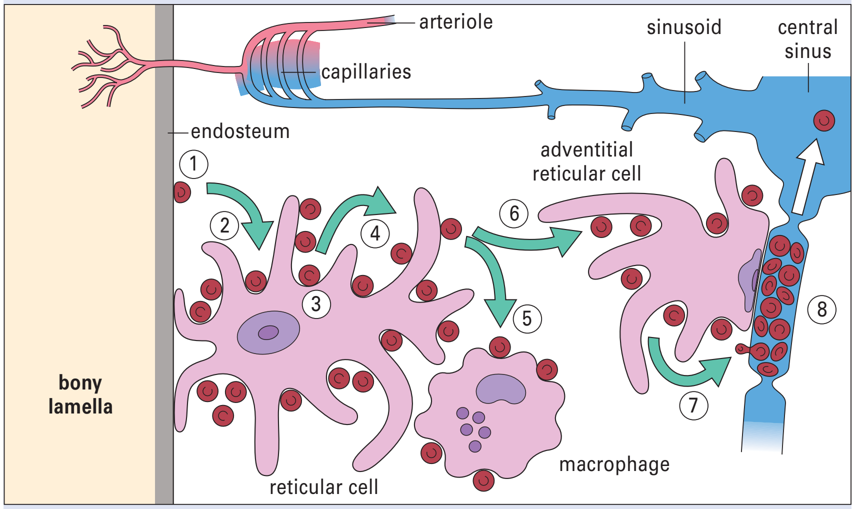

Fig. 2.35 The earliest B cell progenitors are found close to the endosteum (**1**) where they interact with stromal reticular cells (**2**). The stromal reticular cells prompt precursor B cell proliferation and maturation (**3** and **4**). During these processes, selection occurs, which implies B cell apoptosis and phagocytosis of apoptotic cells by macrophages (**5**). B cells that have survived selection mature further and interact with adventitial reticular cells (**6**), which may facilitate their ingress (**7**) into bone marrow sinusoids (**8**) and finally the central venous sinuses, from which they enter the general circulation. In this model, maturation and selection events follow a gradient from the periphery of the bone marrow tissue contained in the bony spaces towards the center.

rescues a minority of B cells with productive rearrangements of their immunoglobulin genes from programmed cell death.

Many self-reactive B cells are also eliminated through negative selection in the bone marrow.

From kinetic data, it is estimated that about 5×10^7 murine B cells are produced each day. As the mouse spleen contains approximately 7.5×10^7 B cells, a large proportion of B cells must die, probably at the pre-B cell stage, because of non-productive rearrangements of receptor genes or if they express self-reactive antibodies and are not rescued.

Immunoglobulins are the definitive B cell lineage markers

Lymphoid stem cells expressing terminal deoxynucleotidyl transferase (TdT) proliferate, differentiate, and undergo immunoglobulin gene rearrangements (see Chapters 3 and 8) to emerge as pre-B cells which express μ heavy chains in the cytoplasm. Some of these pre-B cells bear small numbers of surface μ chains, associated with 'surrogate' light chains $V_{\text{pre B}}$ and λ5 (see Fig. 8.2). **Allelic exclusion** of either maternal or paternal immunoglobulin genes has already occurred by this time. The proliferating pre-B cells are thought to give rise to smaller pre-B cells.

Once a B cell has synthesized light chains, which may be either κ- or λ-type, it becomes committed to the antigen-binding specificity of its surface IgM (sIgM) antigen receptor.

One B cell can therefore make only one specific antibody – a central tenet of the **clonal selection theory** for antibody production.

Surface immunoglobulin-associated molecules Igα and Igβ (CD79a and b) are present by the pre-B cell stage of development.

Developing B cells acquire characteristic surface molecules

A sequence of immunoglobulin gene rearrangements and phenotypic changes takes place during B cell ontogeny (see Chapter 8), similar to that described above for T cells.

Heavy-chain gene rearrangements occur in B cell progenitors and represent the earliest indication of B lineage commitment. This is followed by light-chain gene rearrangements, which occur at later pre-B cell stages.

B cells migrate to and function in the secondary lymphoid tissues

Early B cell immigrants into fetal lymph nodes (17 weeks in humans) are surface IgM^+ and are B-1 cells. CD5^+ B cell precursors are found in the fetal omentum.

Some CD5^+ B cells are also found in the marginal zone of the spleen and mantle zone of secondary follicles in adult lymph nodes (see Fig. 2.44).

Following antigenic stimulation, mature B cells can develop into memory cells or antibody-forming cells (AFCs).

Bone marrow

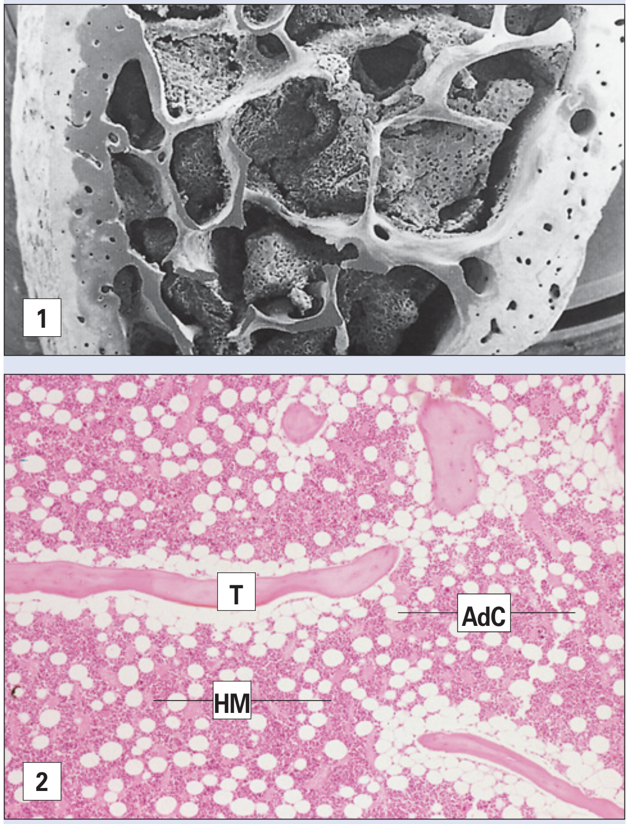

Fig. 2.36 (**1**) Low-power scanning electron micrograph showing the architecture of bone and its relationship to bone marrow. Within the cavities of spongy bone between the bony trabeculae, B cell lymphopoiesis takes place, with maturation occurring in a radial direction towards the center (from the endosteum to the central venous sinus). (**2**) The biopsy below shows hematopoietic bone marrow (HM) in the spaces between the bony trabeculae (lamellae) (T). Some of the space is also occupied by adipocytes (AdC). (Courtesy of Dr A Stevens and Professor J Lowe)

Surface immunoglobulins (sIg) are usually lost from plasma cells (the terminally differentiated form of an AFC) because their function as a receptor is no longer required. Like any other terminally differentiated hematopoietic cell, the plasma cell has a limited life span, and eventually undergoes apoptosis.

LYMPHOCYTES MIGRATE TO AND FUNCTION IN SECONDARY LYMPHOID ORGANS AND TISSUES

The generation of lymphocytes in primary lymphoid organs is followed by their migration into peripheral secondary tissues, which comprise:

- well-organized encapsulated organs, the **spleen** and **lymph nodes (systemic lymphoid organs)**; and
- non-encapsulated accumulations of lymphoid tissue.

Lymphoid tissue found in association with mucosal surfaces is called **mucosa-associated lymphoid tissue (MALT)**.

LYMPHOID ORGANS AND TISSUES PROTECT DIFFERENT BODY SITES

The systemic lymphoid organs and the mucosal system have different functions in immunity:

- the spleen is responsive to blood-borne antigens, and patients who have had their spleen removed are much more susceptible to pathogens that reach the blood stream;
- the lymph nodes protect the body from antigens that come from skin or from internal surfaces and are transported via the lymphatic vessels;
- the mucosal system protects mucosal surfaces.

Responses to antigens encountered via the spleen and lymph nodes result in the secretion of antibodies into the circulation and local cell-mediated responses.

The mucosal system is the site of first encounter (priming) of immune cells with antigens entering via mucosal surfaces, and lymphoid tissues are associated with surfaces lining:

- the intestinal tract – gut-associated lymphoid tissue (GALT);
- the respiratory tract – bronchus-associated lymphoid tissue (BALT); and
- the genitourinary tract.

The major effector mechanism at mucosal surfaces is secretory IgA antibody (sIgA), which is actively transported via the mucosal epithelial cells to the lumen of the tracts.

Q. More than 50% of the body's lymphoid tissue is in the MALT and IgA is the most abundant immunoglobulin in the body. What reason could explain this preponderance of immune defenses in mucosal tissues?

A. The mucosal surfaces present a large surface area, vulnerable to infectious agents. Most infectious agents enter the body by infecting and/or crossing mucosal surfaces.

SYSTEMIC LYMPHOID ORGANS INCLUDE THE SPLEEN AND LYMPH NODES

The spleen is made up of white and red pulp and a marginal zone

The spleen lies at the upper left quadrant of the abdomen, behind the stomach and close to the diaphragm. The adult spleen is around 13 × 8 cm in size and weighs approximately 180–250 g.

The outer layer of the spleen consists of a capsule of collagenous bundles of fibers, which enter the parenchyma of the organ as short trabeculae. These, together with a reticular framework, support two main types of splenic tissue:

- the **white pulp**; and
- the **red pulp**.

A third compartment, the **marginal zone**, is located at the outer limit of the white pulp.

White pulp of the spleen

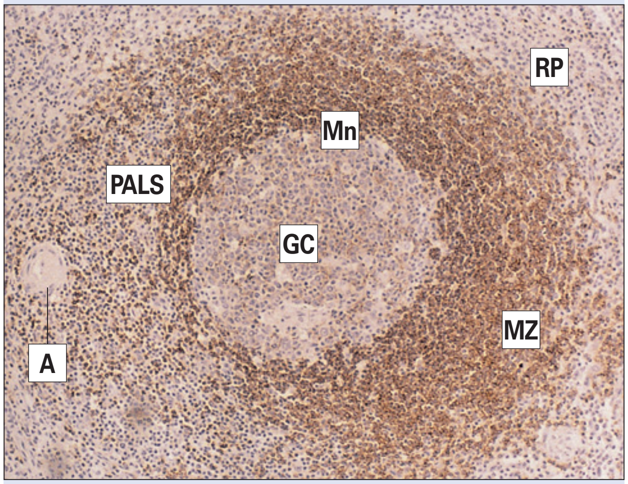

Fig. 2.37 Spleen section showing a white pulp lymphoid aggregate. A secondary lymphoid follicle, with germinal center (GC) and mantle (Mn), is surrounded by the marginal zone (MZ) and red pulp (RP). Adjacent to the follicle, an arteriole (A) is surrounded by the periarteriolar lymphoid sheath (PALS) consisting mainly of T cells. Note that the marginal zone is present only at one side of the secondary follicle. (Courtesy of Professor I Maclennan)

The white pulp consists of lymphoid tissue

The white pulp of the spleen consists of lymphoid tissue, the bulk of which is arranged around a central arteriole to form the periarteriolar lymphoid sheaths (PALS; Fig. 2.37). PALS are composed of T and B cell areas:

- the T cells are found around the central arteriole;
- the B cells may be organized into either primary 'unstimulated' follicles (aggregates of virgin B cells) or secondary 'stimulated' follicles (which possess a germinal center with memory cells).

The germinal centers also contain follicular dendritic cells (FDCs) and phagocytic macrophages. Macrophages and the FDCs present antigen to B cells in the spleen.

B cells and other lymphocytes are free to leave and enter the PALS via branches of the central arterioles, which enter a system of blood vessels in the marginal zone (see below). Some lymphocytes, especially maturing plasmablasts, can pass across the marginal zone via bridges into the red pulp.

The red pulp consists of venous sinuses and cellular cords

The venous sinuses and cellular cords of the red pulp contain:

- resident macrophages;
- erythrocytes;
- platelets;
- granulocytes;
- lymphocytes; and
- numerous plasma cells.

In addition to immunological functions, the spleen serves as a reservoir for platelets, erythrocytes, and granulocytes. Aged platelets and erythrocytes are destroyed in the red pulp in a process referred to as 'hemocatheresis'.

The functions of the spleen are made possible by its vascular organization (Fig. 2.38). Central arteries surrounded by PALS end with arterial capillaries, which open freely into the red pulp cords. Circulating cells can therefore reach these cords and become trapped. Aged platelets and erythrocytes are recognized and phagocytosed by macrophages.

Blood cells that are not ingested and destroyed can reenter the blood circulation by squeezing through holes in the discontinuous endothelial wall of the venous sinuses, through which plasma flows freely.

The marginal zone contains B cells, macrophages, and dendritic cells

The marginal zone surrounds the white pulp and exhibits two major features, namely:

- a characteristic vascular organization; and
- unique subsets of resident cells (B cells, macrophages, and dendritic cells).

The blood vessels of the marginal zone form a system of communicating sinuses, which receive blood from branches of the central artery (see Fig. 2.38).

Most of the blood from the marginal sinuses enters the red pulp cords and then drains into the venous sinuses, but a small proportion passes directly into the venous sinuses to form a closed circulation.

Cells residing in the marginal zone comprise:

- various types of APC – metallophilic macrophages, marginal zone macrophages, dendritic cells;
- a subset of B cells with distinctive phenotype and function – they express IgM brightly with low or absent IgD and are long-lived recirculating cells (non-recirculating in the mouse);
- some B-1 cells.

Q. In humans, the marginal zone does not develop fully until 2 years of age. What is the functional consequence of this delay in development of the marginal zone?

A. Marginal zone B cells and B-1 cells respond strongly to **thymus-independent antigens**, including capsular polysaccharides of bacteria, and the main function of the marginal zone is to mount immune responses to bacteria that have reached the circulation (e.g. streptococci). Infants therefore have a reduced ability to respond to blood-borne infections with certain (encapsulated) bacteria.

Lymph nodes filter antigens from the interstitial tissue fluid and lymph

The lymph nodes form part of a network that filters antigens from the interstitial tissue fluid and lymph during its passage from the periphery to the thoracic duct and the other major collecting ducts (Fig. 2.39).

Lymph nodes frequently occur at the branches of the lymphatic vessels. Clusters of lymph nodes are strategically placed in areas that drain various superficial and deep regions of the body, such as the:

Vascular organization the spleen

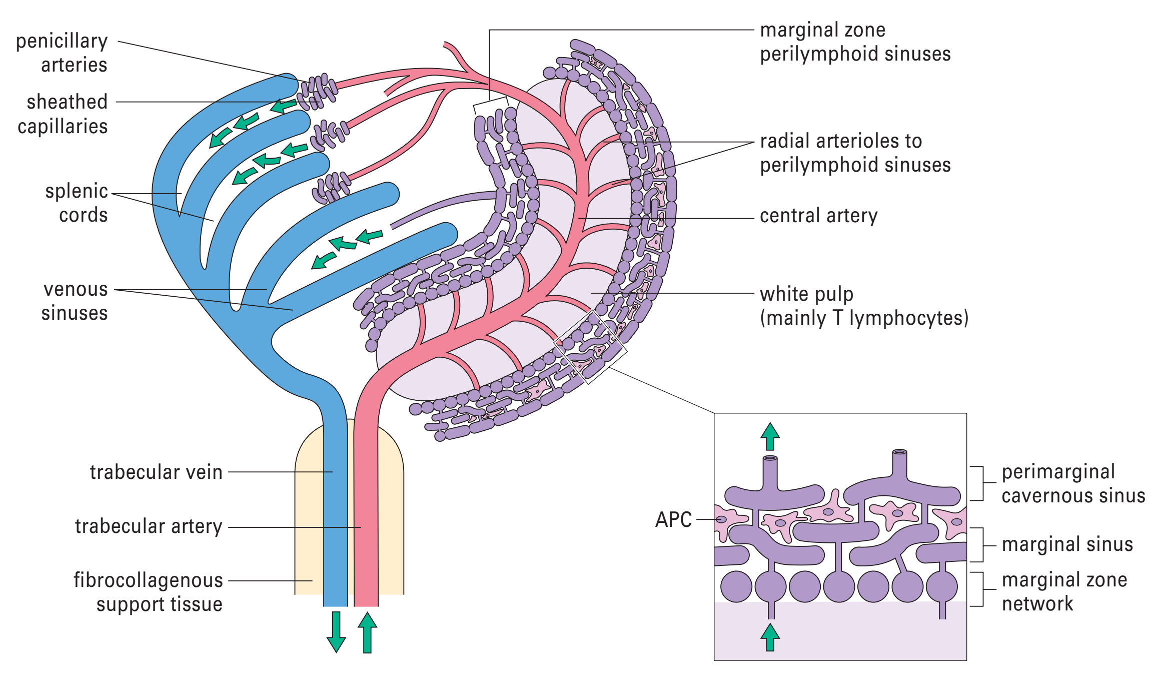
penicillary arteries
sheathed capillaries
splenic cords
venous sinuses
trabecular vein
trabecular artery
fibrocollagenous support tissue
marginal zone perilymphoid sinuses
radial arterioles to perilymphoid sinuses
central artery
white pulp (mainly T lymphocytes)
APC
perimarginal cavernous sinus
marginal sinus
marginal zone network

Fig. 2.38 The splenic artery branches to form trabecular arteries, which give rise to central arteries surrounded by periarteriolar lymphoid sheaths (PALS), which are the T cell areas of the white pulp. Leaving the PALS, the central arteries continue as penicillary arteries and sheathed capillaries, which open in the splenic cords of the red pulp. From the red pulp (where hemocatheresis takes place) the blood percolates through the wall of the venous sinuses. The central arterioles surrounded by PALS give collateral branches that reach a series of sinuses in the marginal zone. Most of the blood from the marginal sinuses enters the red pulp cords and then drains into the venous sinuses, but a proportion passes directly into the sinuses to form a closed circulation. (Courtesy of Dr A Stevens and Professor J Lowe)

- neck;
- axillae;
- groin;
- mediastinum; and
- abdominal cavity.

Lymph nodes protect the skin (superficial subcutaneous nodes) and mucosal surfaces of the respiratory, digestive, and genitourinary tracts (visceral or deep nodes).

Human lymph nodes are 2–10 mm in diameter, are round or kidney shaped, and have an indentation called the hilus where blood vessels enter and leave the node.

Lymph arrives at the lymph node via several afferent lymphatic vessels, and leaves the node through a single efferent lymphatic vessel at the hilus.

Lymph nodes consist of B and T cell areas and a medulla

A typical lymph node is surrounded by a collagenous capsule. Radial trabeculae, together with reticular fibers, support the various cellular components. The lymph node consists of:

- a **B cell area (cortex)**;
- a **T cell area (paracortex)**; and
- a central **medulla**, consisting of cellular cords containing T cells, B cells, abundant plasma cells, and macrophages (Figs 2.40–2.42).

The paracortex contains many APCs (interdigitating cells), which express high levels of MHC class II surface molecules. These are cells migrating from the skin (Langerhans' cells) or from mucosae (dendritic cells), which transport processed antigens into the lymph nodes from the external and internal surfaces of the body (Fig. 2.43). The bulk of the lymphoid tissue is found in the cortex and paracortex.

The paracortex contains specialized postcapillary vessels – **high endothelial venules (HEVs)** – which allow the traffic of lymphocytes out of the circulation into the lymph node (see 'Lymphocyte traffic' below and Fig. 2.54).

The medulla is organized into cords separated by lymph (medullary) sinuses, which drain into a terminal sinus – the origin of the efferent lymphatic vessel (see Fig. 2.42).

Scavenger phagocytic cells are arranged along the lymph sinuses, especially in the medulla. As the lymph passes across the nodes from the afferent to the efferent

The lymphatic system

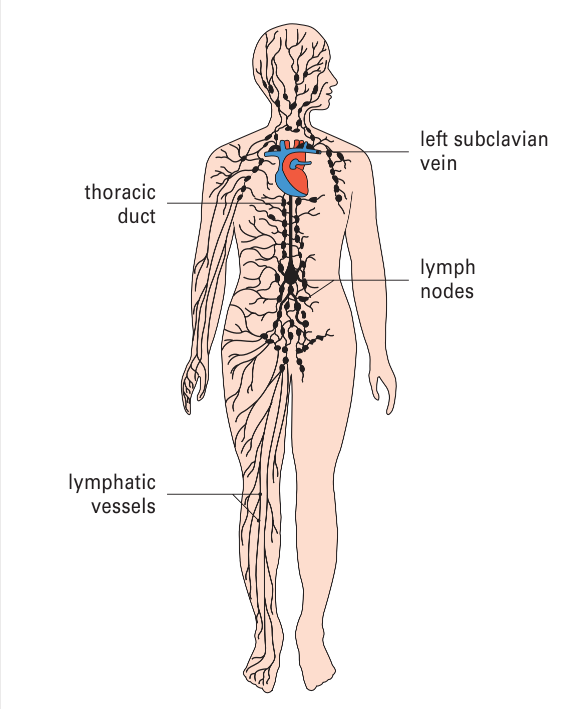

Fig. 2.39 Lymph nodes are found at junctions of lymphatic vessels and form a network that drains and filters interstitial fluid from the tissue spaces. They are either subcutaneous or visceral, the latter draining the deep tissues and internal organs of the body. The lymph eventually reaches the thoracic duct, which opens into the left subclavian vein and thus back into the circulation.

Lymph node section

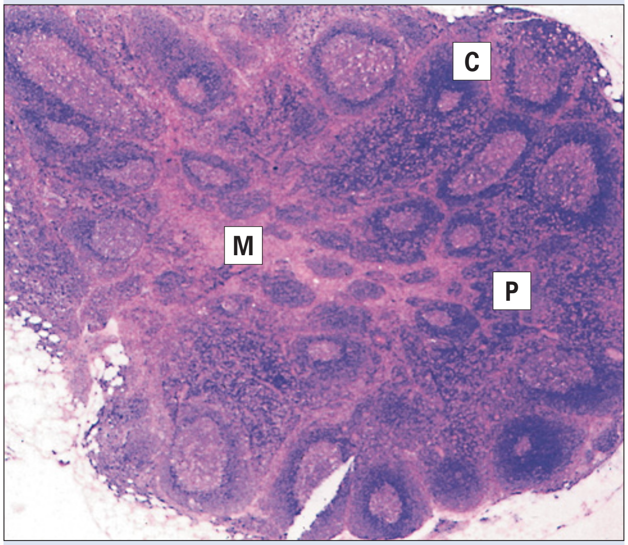

Fig. 2.40 The lymph node is surrounded by a connective tissue capsule and is organized into three main areas: the cortex (C), which is the B cell area; the paracortex (P), which is the T cell area; and the medulla (M), which contains cords of lymphoid tissue (T and B cell areas rich in plasma cells and macrophages). H&E stain. × 10. (Adapted from Zucker-Franklin D, Grossi CE, eds. Atlas of blood cells: function and pathology, 3rd edn. Milan: Edi Ermes; 2003)

lymphatic vessels, particulate antigens are removed by the phagocytic cells and transported into the lymphoid tissue of the lymph node.

The cortex contains aggregates of B cells in the form of primary or secondary follicles.

B cells are also found in the subcapsular region, adjacent to the marginal sinus. It is possible that these cells are similar to the splenic marginal zone B cells that intercept incoming pathogens primarily by mounting a rapid, IgM-based, T-independent response.

T cells are found mainly in the paracortex. Therefore, if an area of skin or mucosa is challenged by a T-dependent antigen, the lymph nodes draining that particular area show active T cell proliferation in the paracortex.

Q. How is antigen transported from the skin to the paracortex of the regional lymph nodes?

A. On Langerhans' cells/veiled cells in afferent lymph.

Further evidence for this localization of T cells in the paracortex comes from patients with congenital thymic aplasia (DiGeorge syndrome), who have fewer T cells in the paracortex than normal. A similar feature is found in neonatally thymectomized or congenitally athymic ('nude') mice or rats.

Histological structure of the lymph node

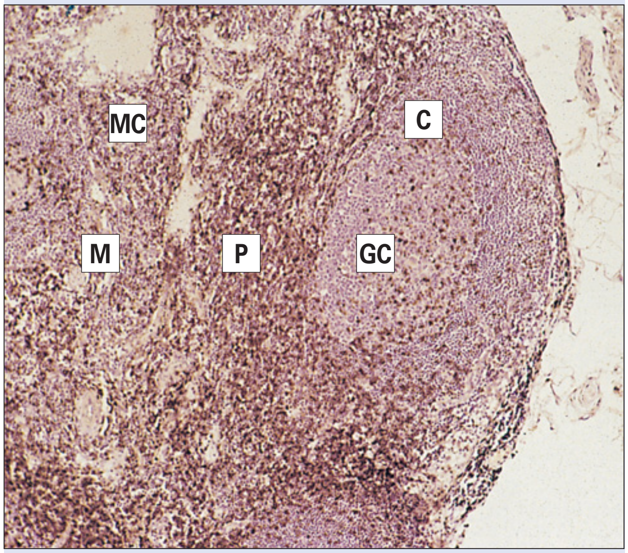

Fig. 2.41 Cortex (C), paracortex (P), and medulla (M) are shown. The section has been stained to show the localization of T cells. They are most abundant in the paracortex, but a few are found in the germinal center (GC) of the secondary lymphoid follicle, in the cortex, and in the medullary cords (MC). (Courtesy of Dr A Stevens and Professor J Lowe)

Schematic structure of the lymph node

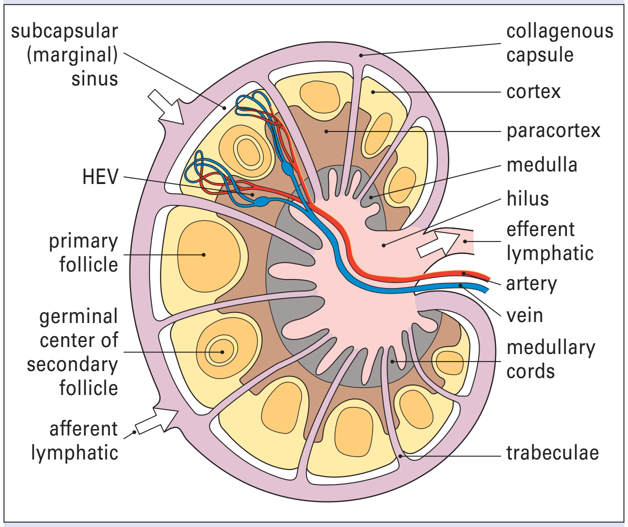

Fig. 2.42 Beneath the collagenous capsule is the subcapsular sinus, which is lined by endothelial and phagocytic cells. Lymphocytes and antigens from surrounding tissue spaces or adjacent nodes pass into the sinus via the afferent lymphatics. The cortex is mainly a B cell area. B cells are organized into primary or, more commonly, secondary follicles – that is, with a germinal center. The paracortex contains mainly T cells. Each lymph node has its own arterial and venous supply. Lymphocytes enter the node from the circulation through the highly specialized high endothelial venules (HEVs) in the paracortex. The medulla contains both T and B cells in addition to most of the lymph node plasma cells organized into cords of lymphoid tissue. Lymphocytes leave the node through the efferent lymphatic vessel.

Interdigitating cells in the lymph node paracortex

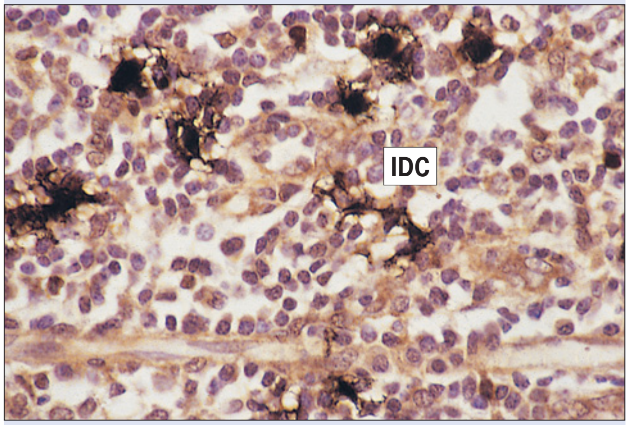

Fig. 2.43 Interdigitating dendritic cells (IDC; stained dark brown) form contacts with each other and with paracortical T cells. (Courtesy of Dr A Stevens and Professor J Lowe) (see also Fig. 2.16)

Structure of the secondary follicle

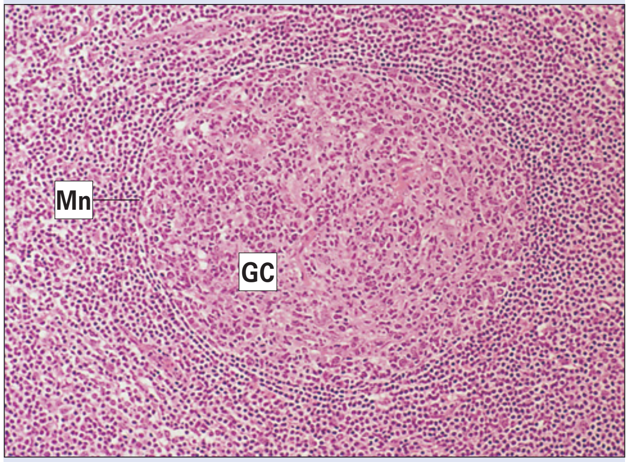

Fig. 2.44 A large germinal center (GC) is surrounded by the mantle zone (Mn).

Distribution of B cells in the lymph node cortex

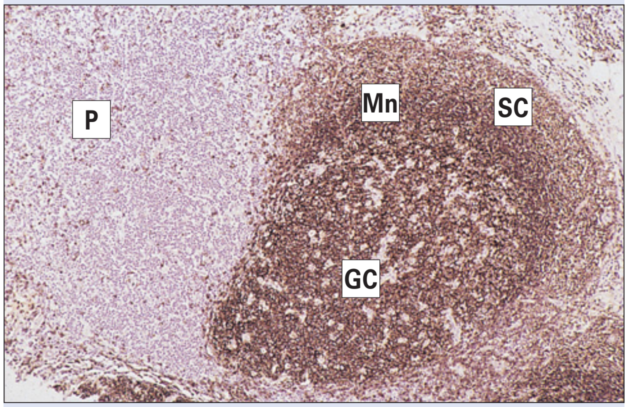

Fig. 2.45 Immunohistochemical staining of B cells for surface immunoglobulin shows that they are concentrated largely in the secondary follicle, germinal centre (GC), mantle zone (Mn), and between the capsule and the follicle – the subcapsular zone (SC). A few B cells are seen in the paracortex (P), which contains mainly T cells (see also Fig. 2.41).

Secondary follicles are made up of a germinal center and a mantle zone

Germinal centers in secondary follicles are seen in antigen-stimulated lymph nodes. These are similar to the germinal centers seen in the B cell areas of the splenic white pulp and of MALT.

Germinal centers are surrounded by a mantle zone of lymphocytes (Fig. 2.44). Mantle zone B cells (Fig. 2.45) co-express surface IgM, IgD, and CD44. This is taken as evidence that they are virgin and actively recirculating B cells.

In most secondary follicles, the thickened mantle zone or corona is oriented towards the capsule of the node. Secondary follicles contain:

- FDCs (Fig. 2.46);

Follicular dendritic cells in a secondary lymphoid follicle

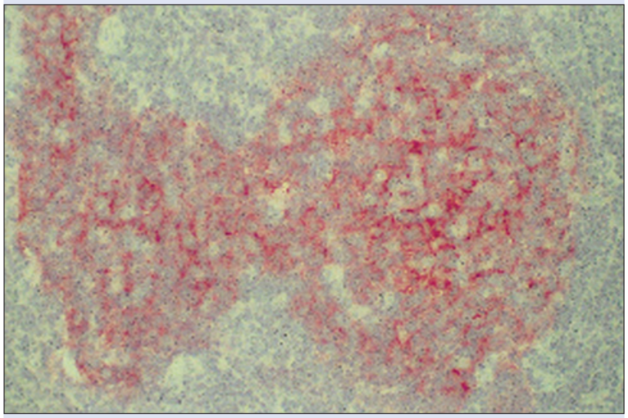

Fig. 2.46 This lymph node follicle is stained with enzyme-labeled monoclonal antibody to demonstrate follicular dendritic cells.

Germinal center macrophages

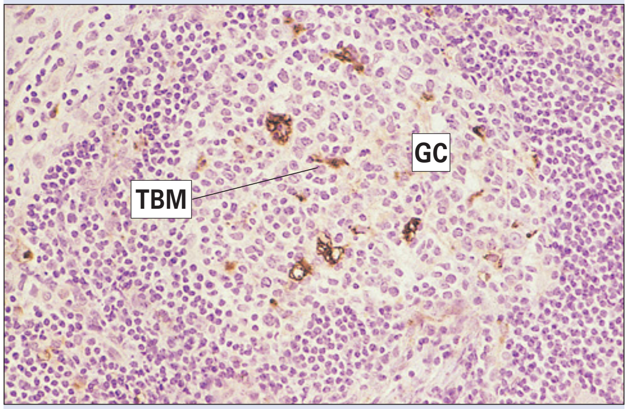

Fig. 2.47 Immunostaining for cathepsin D shows several macrophages localized in the germinal center (GC) of a secondary follicle. These macrophages, which phagocytose apoptotic B cells, are called tingible body macrophages (TBM). (Courtesy of Dr A Stevens and Professor J Lowe)

- some macrophages (Fig. 2.47); and
- a few $CD4^+$ T cells.

All the cells in the secondary follicle together with specialized marginal sinus macrophages, appear to play a role in generating B cell responses and, in particular, in the development of B cell memory.

In the germinal centers B cells proliferate, are selected, and differentiate into memory cells plasma cell precursors

The germinal center consists of a dark zone and a light zone:

- the dark zone is the site where one or a few B cells enter the primary lymphoid follicle and undergo active proliferation leading to clonal expansion – these B cells are termed **centroblasts** and undergo a process of **somatic hypermutation**, which leads to the generation of cells with a wide range of affinities for antigen;
- in the light zone, B cells (**centrocytes**) encounter the antigen on the surface of FDCs (see Fig. 2.14) and only those cells with higher affinity for antigen survive.

Cells with mutated antibody receptors of lower affinity die by apoptosis and are phagocytosed by germinal center macrophages.

Selected centrocytes interact with germinal center $CD4^+$ TH cells and undergo **class switching** (i.e.

Structure and function of the germinal center

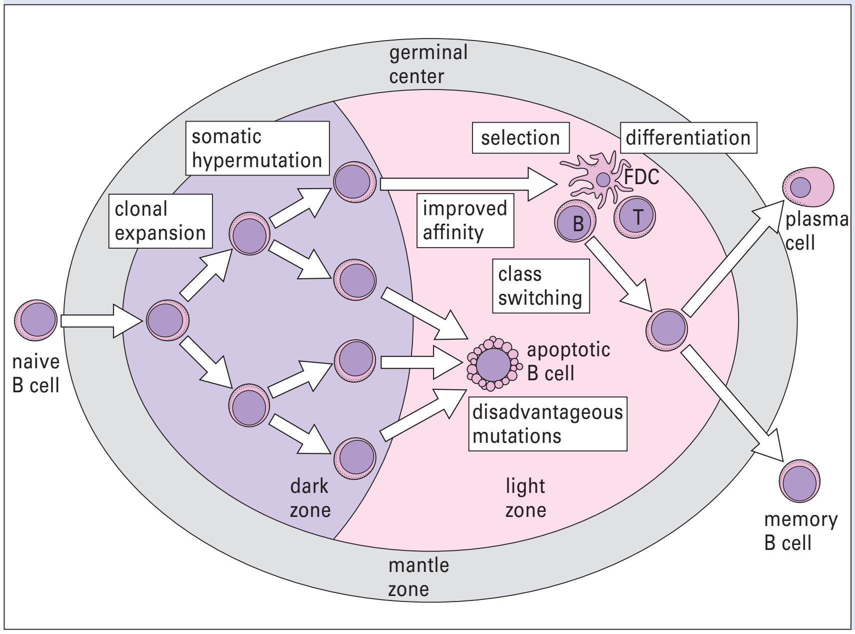

Fig. 2.48 One or a few B cells (founder cells) in the dark zone proliferate actively. This proliferation leads to clonal expansion and is accompanied by somatic hypermutation of the immunoglobulin V region genes. B cells with the same specificity, but different affinity, are therefore generated. In the light zone, B cells with disadvantageous mutations or with low affinity undergo apoptosis and are phagocytosed by macrophages. Cells with appropriate affinity encounter the antigen on the surface of the follicular dendritic cells (FDCs) and, with the help of $CD4^+$ T cells, undergo class switching, leaving the follicle as memory B cells or plasma cells precursors.

replacement of their originally expressed immunoglobulin heavy chain constant region genes by another class – for instance IgM to IgG or IgA, see Chapter 8).

The selected germinal center B cells differentiate into **memory B cells** or **plasma cell** precursors and leave the germinal center (Fig. 2.48).

MALT INCLUDES ALL LYMPHOID TISSUES ASSOCIATED WITH MUCOSA

Aggregates of encapsulated and non-encapsulated lymphoid tissue are found especially in the lamina propria and submucosal areas of the gastrointestinal, respiratory, and genitourinary tracts (see Fig. 2.26).

The tonsils contain a considerable amount of lymphoid tissue, often with large secondary follicles and intervening T cell zones with HEVs. The three main kinds of tonsil that constitute Waldeyer's ring are:

- palatine tonsil;
- pharyngeal tonsil (called adenoids when diseased); and
- lingual tonsil (Fig. 2.49).

Aggregates of lymphoid tissue are also seen lining the bronchi and along the genitourinary tract.

The digestive, respiratory, and genitourinary mucosae contain dendritic cells for the uptake, processing, and transport of antigens to the draining lymph nodes.

Lymphoid tissues seen in the lamina propria of the gastrointestinal wall often extend into the submucosa and are found as either:

- solitary nodules (Fig. 2.50); or
- aggregated nodules such as in the appendix (Fig. 2.51).

Follicle-associated epithelium is specialized to transport pathogens into the lymphoid tissue

Peyer's patches are found in the lower ileum. The intestinal epithelium overlying Peyer's patches (follicle-associated epithelium – FAE) and other mucosa-associated lymphoid aggregates (e.g. the tonsils) is specialized to allow the transport of pathogens into the lymphoid tissue. This particular function is carried out by epithelial cells termed 'M' cells, which are scattered among other epithelial cells and so called because they have numerous microfolds on their luminal surface.

M cells contain deep invaginations in their basolateral plasma membrane, which form pockets containing B and T lymphocytes, dendritic cells, and macrophages (Fig. 2.52). Antigens and microorganisms are transcytosed into the pocket and to the organized mucosal lymphoid tissue under the epithelium (Fig. 2.53).

M cells are not exclusive to Peyer's patches, but are also found in epithelia associated with lymphoid cell accumulations at 'antigen sampling' areas in other mucosal sites.

Structure of the lingual tonsil

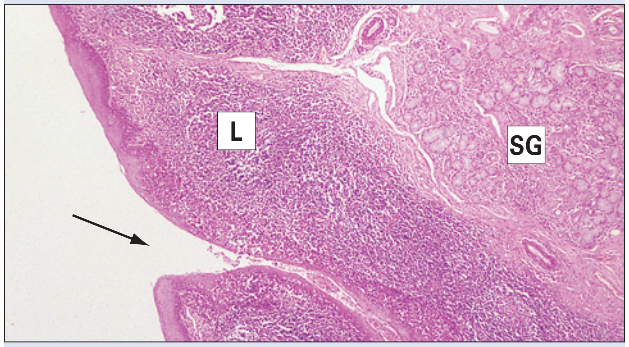

Fig. 2.49 The lingual tonsil, situated in the posterior one-third of the tongue, consists of accumulations of lymphoid tissue (L) with large secondary follicles associated with a mucosa that forms deep cleft-like invaginations (arrow). Mucus-containing salivary glands (SG) are seen around the tonsil. These are common features of all types of tonsil. (Courtesy of Dr A Stevens and Professor J Lowe)

A solitary lymphoid nodule in the large intestine

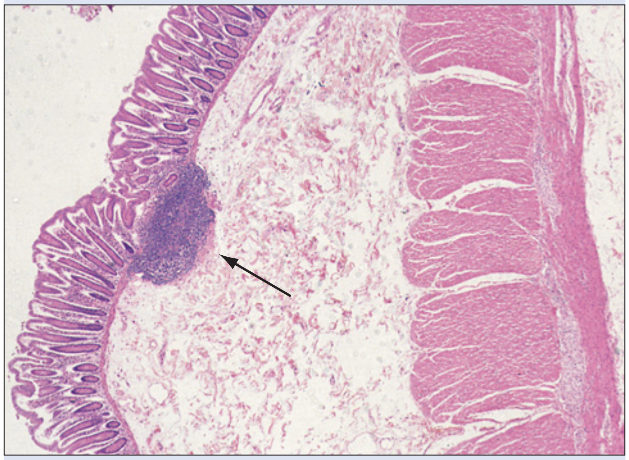

Fig. 2.50 This nodule is localized in the mucosa and submucosa of the intestinal wall (arrow). (Courtesy of Dr A Stevens and Professor J Lowe)

Lymphoid nodules in the human appendix

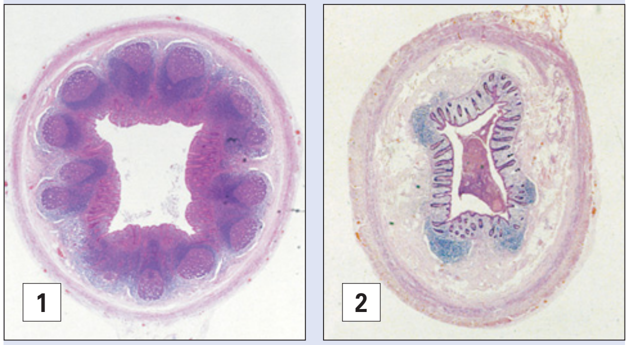

Fig. 2.51 (**1**) Appendix of a 10-year-old child showing large lymphoid nodules extending into the submucosa. (**2**) Appendix from a 36-year-old man. Note the dramatic reduction of lymphoid tissue, with the virtual disappearance of lymphoid follicles. This illustrates the atrophy of lymphoid tissues during ageing, which is not limited to the appendix. (Courtesy of Dr A Stevens and Professor J Lowe)

Location of M cells

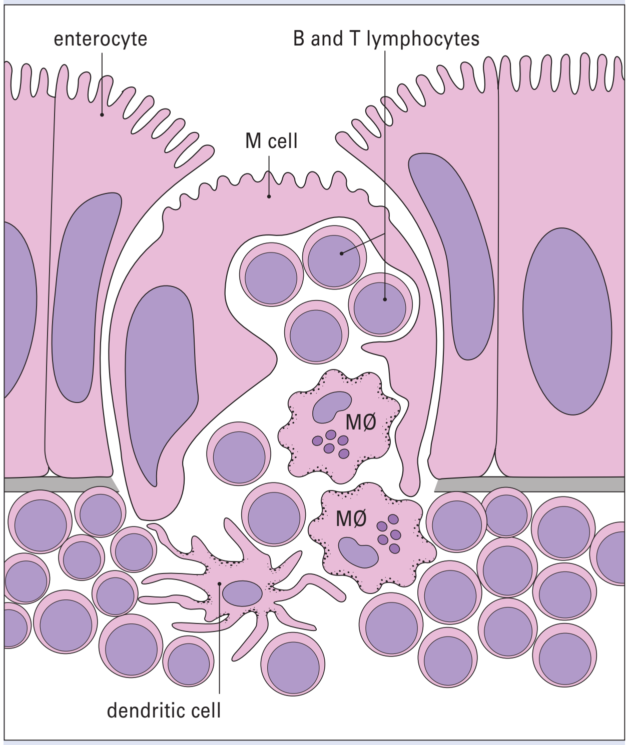
enterocyte
B and T lymphocytes
M cell
MØ
MØ
dendritic cell

Fig. 2.52 The intestinal follicle-associated epithelium contains M cells. Note the lymphocytes and occasional macrophages (MØ) in the pocket formed by invagination of the basolateral membrane of the M cell. Antigens endocytosed by the M cell are passed via this pocket into the subepithelial tissues (not shown).

The dome area of Peyer's patches and the subepithelial regions of tonsils harbor B cells that display a phenotype and function similar to that seen for the splenic marginal zone B cells (see above).

Q. The major defense at mucosal surfaces is antibody of the IgA isotype. What characteristics of this antibody would be critical in this respect?

A. The IgA isotype antibody produced at the mucosal level is a specific secretory form that can traverse epithelial membranes and helps prevent the entry of infectious microorganisms. Resistance to digestion by enzymes in the gut would also be an important feature of secretory IgA in GALT. Transport of IgA across mucosal epithelium is described in detail in Chapter 3 (see Fig. 3.11).

Lamina propria and intraepithelial lymphocytes are found in mucosa

In addition to organized lymphoid tissue forming the MALT system, a large number of lymphocytes and plasma cells are found in the mucosa of the:

- stomach;
- small and large intestine;
- upper and lower respiratory airways; and
- several other organs.

Lymphocytes are found both in the connective tissue of the lamina propria and within the epithelial layer:

- lamina propria lymphocytes (LPLs) are predominantly activated T cells, but numerous activated B cells and plasma cells are also detected – these plasma cells secrete mainly IgA, which is transported across the epithelial cells and released into the lumen;

Mucosal lymphoid tissue

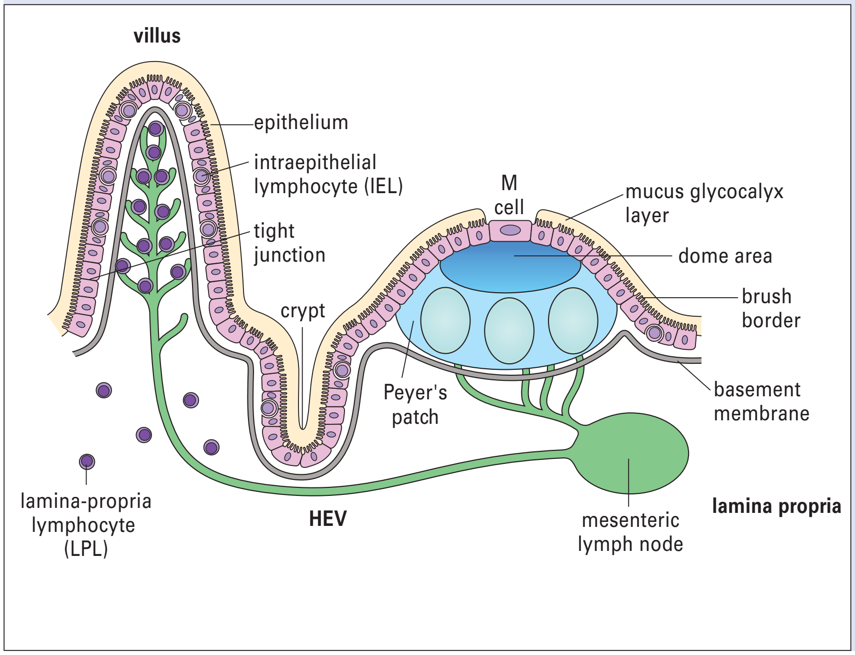
villus
epithelium
intraepithelial lymphocyte (IEL)
tight junction
crypt
M cell
mucus glycocalyx layer
dome area
brush border
Peyer's patch
basement membrane
lamina-propria lymphocyte (LPL)
HEV
mesenteric lymph node
lamina propria

Fig. 2.53 Peyer's patches, as well as tonsils and other lymphoid areas of MALT, are sites of lymphocyte priming by antigens, which are internalized by M cells in the follicle-associated epithelium (FAE). The subepithelial region, the dome, is rich in APCs and also contains a subset of B cells similar to those found in the splenic marginal zone. Lymphoid follicles and intervening T-dependent zones are localized under the dome region. Lymphocytes primed by antigens in these initiation sites of the gut mucosa migrate to the mesenteric lymph nodes and then to the effector sites (the intestinal villi), where they are found both in the lamina propria (LPLs) and within the surface epithelium (IELs)

Patterns of lymphocyte traffic

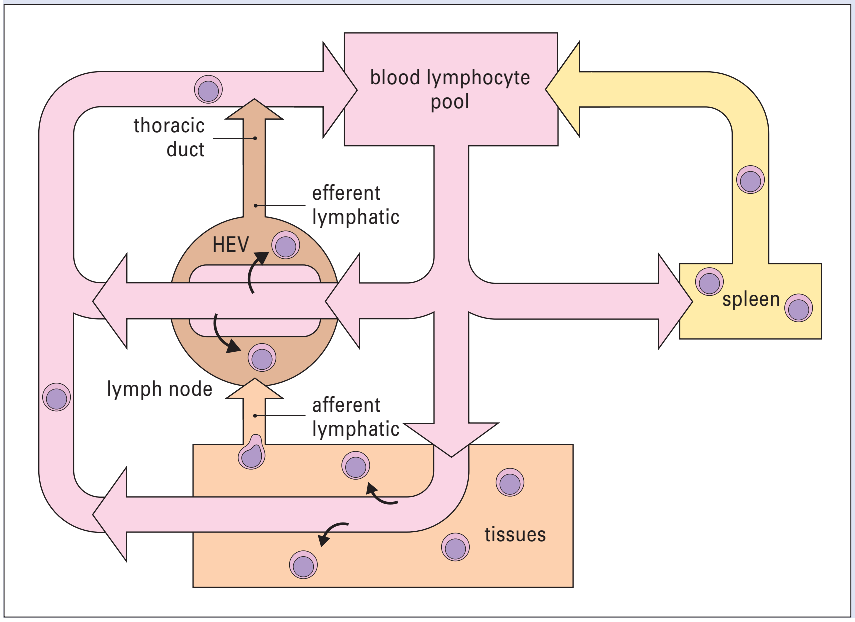

Fig. 2.54 The lymphocytes move through the circulation and enter the lymph nodes and MALT via the specialized endothelial cells of the postcapillary venules (i.e. high endothelial venules [HEVs]). They leave the lymph nodes and MALT through the efferent lymphatic vessels and pass through other nodes, finally entering the thoracic duct, which empties into the circulation at the left subclavian vein (in humans). Lymphocytes enter the white pulp areas of the spleen in the marginal zones, pass into the sinusoids of the red pulp, and leave via the splenic vein.

- intraepithelial lymphocytes (IELs) are mostly T cells – the population is different from the LPLs because it includes a high proportion of γδ T cells (10–40%) and CD8$^+$ cells (70%).

Most LPL and IEL T cells belong to the CD45RO subset of memory cells. They respond poorly to stimulation with antibodies to CD3, but may be triggered via other activation pathways (e.g. via CD2 or CD28).

The integrin αE chain HML-1 (CD103) is not present on resting circulating T cells, but is expressed following phytohemagglutinin (PHA) stimulation. Antibodies to CD103 are mitogenic and induce expression of the low-affinity IL-2 receptor α chain (CD25) on peripheral blood T cells. αE is coupled with a β_7 chain to form an $\alpha E/\beta_7$ heterodimer, which is an integrin expressed by IELs and other activated leukocytes. E-cadherin on epithelial cells is the ligand for $\alpha E/\beta_7$. Binding of $\alpha E/\beta_7$ to E-cadherin may be important in the homing and retention of $\alpha E/\beta_7$-expressing lymphocytes in the intestinal epithelium.

IELs are known to release cytokines, including IFNγ and IL-5. One function suggested for IELs is immune surveillance against mutated or virus-infected host cells.

MOST LYMPHOCYTES RECIRCULATE AROUND THE BODY

Once in the secondary tissues the lymphocytes do not simply remain there; many move from one lymphoid organ to another via the blood and lymph (Fig. 2.54).

Lymphocytes leave the blood via high endothelial venules

Although some lymphocytes leave the blood through non-specialized venules, the main exit route in mammals is through a specialized section of the PCVs known as high endothelial venules (HEVs; Figs 2.55 and 2.56). In the lymph nodes these are mainly in the paracortex, with fewer in the cortex and none in the medulla.

Some lymphocytes, primarily T cells, arrive from the drainage area of the node through the afferent lymphatics, not via HEVs – this is the main route by which antigen enters the nodes.

Besides lymph nodes, HEVs are also found in MALT and in the thymus (see Fig. 2.29).

Lymph node paracortex showing high endothelial venules (HEVs)

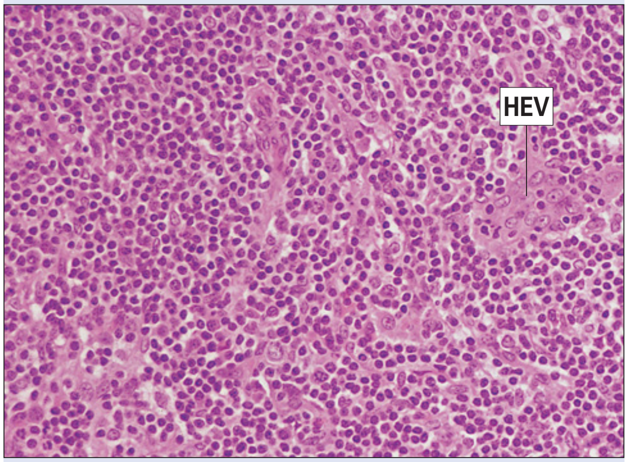

Fig. 2.55 Lymphocytes leave the circulation through HEVs and enter the node. H&E. × 200. (Courtesy of Dr A Stevens and Professor J Lowe)

Electron micrograph showing a high endothelial venule (HEV) in the paracortex of a lymph node

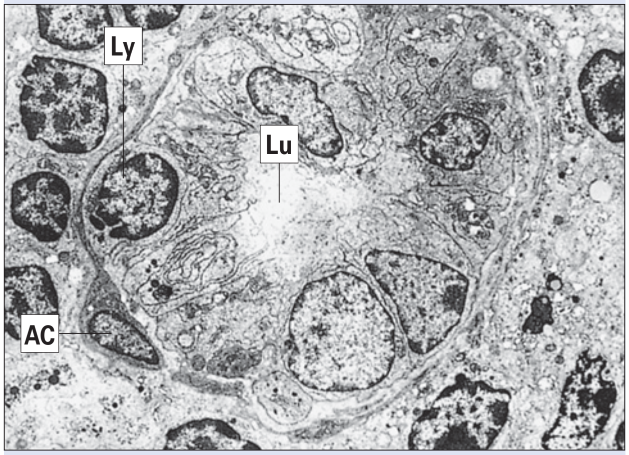

Fig. 2.56 A lymphocyte (Ly) in transit from the lumen (Lu) of the HEV can be seen close to the basal lamina. The HEV is partly surrounded by an adventitial cell (AC). × 1600.

Q. What types of molecule required for the movement of lymphocytes would you expect to be expressed on HEVs?

A. HEVs express a distinctive set of chemokines that signal lymphocytes to migrate into the lymphoid tissue. They also have a specialized set of adhesion molecules that allow the cells to attach to the endothelial cells, as they migrate.

HEVs are permanent features of secondary lymphoid tissues, but can also develop from normal endothelium at sites of chronic inflammatory reactions (e.g. in the skin and in the synovium). This, in turn, may direct specific T cell subsets to the area where HEVs have formed.

The movement of lymphocytes across endothelium is controlled by adhesion molecules (see Appendix 2) and chemokines (see Appendix 4). For example:

- the adhesion molecule MadCAM-1 is expressed on endothelial cells in intestinal tissues;
- VCAM-1 is present on endothelial cells in the lung and skin.

Homing molecules on lymphocytes selectively direct lymphocytes to particular organs by interaction with these adhesion molecules (see Chapter 6). In the case of the intestine, a critical role is played by $\alpha_4\beta_7$-integrins, which mediate adherence of lymphocytes to HEVs of Peyer's patches that express MadCAM-1.

Lymphocyte trafficking exposes antigen to a large number of lymphocytes

Lymphoid cells within lymph nodes return to the circulation by way of the efferent lymphatics, which pass via the thoracic duct into the left subclavian vein. About 1–2% of the lymphocyte pool recirculates each hour. Overall, this process allows a large number of antigen-specific lymphocytes to come into contact with their specific antigen in the microenvironment of the peripheral lymphoid organs.

Q. Why is it important that antigen can contact many lymphocytes?

A. Lymphoid cells are monospecific and only a limited number of lymphocytes are capable of recognizing any particular antigen. Lymphocyte recirculation and the movement of an antigen and APCs increase the opportunity for lymphocytes to encounter their specific antigen soon after infection.

Under normal conditions there is continuous lymphocyte traffic through the lymph nodes, but when antigen enters the lymph nodes of an animal already sensitized to that antigen there is a temporary shutdown in the traffic, which lasts for approximately 24 hours. Thus, antigen-specific lymphocytes are preferentially retained in the lymph nodes draining the source of antigen. In particular, blast cells do not recirculate but appear to remain in one site.

Antigen stimulation at one mucosal area elicits an antibody response largely restricted to MALT

One reason for considering MALT as a system distinct from the systemic lymphoid organs is that mucosa-associated lymphoid cells mainly recirculate within the mucosal lymphoid system. Thus, lymphoid cells stimulated in Peyer's patches pass via regional lymph nodes to the blood stream and then 'home' back into the intestinal lamina propria (Fig. 2.57 and see Fig. 2.53).

Specific recirculation is made possible because the lymphoid cells expressing homing molecules attach to adhesion molecules that are specifically expressed on endothelial cell adhesion molecules of the mucosal PCVs, but are absent from lymph node HEVs (see above).

Thus, antigen stimulation at one mucosal area elicits an antibody response largely, but not exclusively, restricted to mucosal tissues.

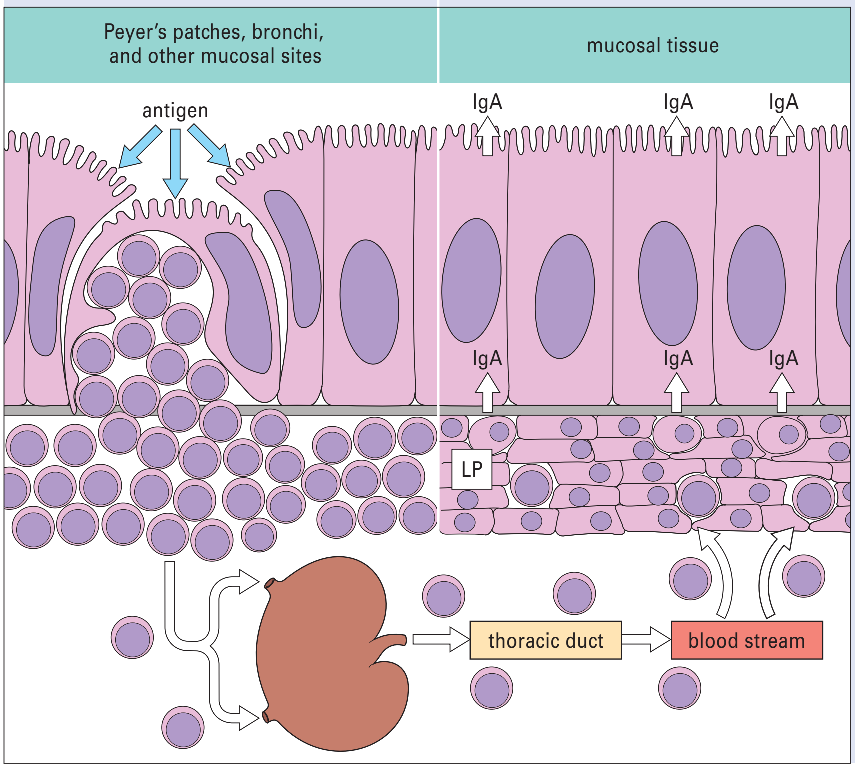

Lymphocyte circulation within the mucosal lymphoid system

Fig. 2.57 Lymphoid cells that are stimulated by antigen in Peyer's patches (or the bronchi or another mucosal site) migrate via the regional lymph nodes and thoracic duct into the blood stream and hence to the lamina propria (LP) of the gut or other mucosal surfaces, which might be close to or distant from the site of priming. Thus lymphocytes stimulated at one mucosal surface may become distributed selectively throughout the MALT system. This is mediated through specific adhesion molecules on the lymphocytes and mucosal HEV.

FURTHER READING

Goldsby RA, Kindt TJ, Kuby J, Osborne BA. Kuby Immunology, 5th edn. Oxford: WH Freeman; 2003.

Janeway Jr CA, Travers P, Wolpert M, Schlomchik MJ. Immunobiology. London: Garland Publishing; 2004.

Lamm ME, Strober W, McGhee JR, Mayer L, eds. Mestecky J, Bienenstock J. Mucosal Immunology. San Diego: Academic Press; 2005.

Liu Y-J, Kanzler H, Soumelis V, Gilliet M. Dendritic cell lineage, plasticity and cross-regulation. Nature Immunology 2001;7:585–589.

Lydyard PM, Whelan A, Fanger MW. Instant Notes in Immunology, 2nd edn. London: Garland Science/Bios Scientific Publishing, 2004.

Playfair JHL, Chain BM. Immunology at a Glance, 7th edn. Oxford: Blackwell Scientific Publications; 2004.

Reddy KV, Yedery RD, Aranha C. Antimicrobial peptides: premises and promises. Int J Antimicrob Agents 2004;24:536–547.

Roitt IM, Delves P. Essential Immunology, 10th edn. Oxford: Blackwell Scientific Publications; 2001.

Shevach E. Regulatory/suppressor T cells in health and disease. Arthritis Rheum 2004;50:2721–2724.

Von Andrian UH, Mempel TR. Homing and cellular traffic in lymph nodes. Nature Rev Immunol 2003;3:867–878.

Critical thinking: Development of the immune system (see p. 493 for explanations)

Immunodeficiencies can tell us a lot about the way the immune system functions normally. Mice that congenitally lack a thymus (and have an associated gene defect that produces hairlessness), termed 'nude mice', are often used in research.

1 What effect would you expect this defect to have on numbers and types of lymphocytes in the blood? How would this affect the structure of the lymph nodes? What effect would this have on the ability of the mice to fight infections?

Occasionally adult patients develop a tumor in the thymus (thymoma) and it is necessary to completely remove the thymus gland.

2 What effect would you expect adult thymectomy to have on the ability of such patients to fight infections?

With the development of modern techniques in molecular biology, it is possible to produce animals that completely lack individual genes. Such animals are called 'gene knockouts'. Sometimes these knockouts can have quite surprising effects on development, and sometimes only minor effects. Others, like the immunodeficiencies, are very informative. Based on the information provided in Chapter 2, what effects would you expect the following 'knockouts' to have on the development of leukocytes and/or lymphoid organs?

3 RAG-1? (*RAG-1* and *RAG-2* genes are involved in the recombination processes that generate antigen receptors on B and T cells.)

4 Interleukin-7?

5 The β_7-integrin chain?

CHAPTER

3

Antibodies

SUMMARY

- **Circulating antibodies (also called immunoglobulins) are soluble glycoproteins that recognize and bind antigens** present in serum, tissue fluids or on cell membranes. Their purpose is to help eliminate their specific antigens or microorganisms bearing those antigens. Immunoglobulins also function as membrane-bound antigen receptors on B cells, and play key roles in B cell differentiation.

- **All immunoglobulin isotypes, with the exception of IgD, are bifunctional** – they bind antigen and exhibit one or more effector functions. These biological activities are localized to sites that are distant from the antigen-binding sites.

- **Immunoglobulins are made up of a unit with four polypeptide chains** – two identical light chains and two identical heavy chains. The N terminal domains of each light and heavy chain are highly variable in sequence, are referred to as the variable regions (V_L and V_H, respectively), and form the antigen-binding sites of the antibody. The C terminal domains of the light and heavy chains together form the constant regions (C_L and C_H, respectively), which determine the effector functions of the immunoglobulin.

- **There are five classes of antibody in mammals** – IgG, IgA, IgM, IgD, and IgE.

- **Antigen-binding sites of antibodies are specific for the three-dimensional shape (conformation) of their target** – the antigenic determinant or epitope.

- **Antibody affinity** is a measure of the strength of the interaction between an antibody combining site and its epitope. The avidity (or functional affinity) of an antibody depends on its number of binding sites and its ability to engage multiple epitopes on the antigen – the more epitopes it binds, the greater the avidity.

- **Each immunoglobulin isotype (subclass) mediates a distinct set of effector functions.** In humans there are a number of different isotypes and these are determined by the constant (C) regions of their heavy chains.

- **Receptors for immunoglobulin constant regions (Fc receptors)** are expressed by mononuclear cells, neutrophils, natural killer cells, eosinophils, basophils, and mast cells. They interact with the Fc regions of different classes of immunoglobulin and promote activities such as phagocytosis, tumor cell killing, and mast cell degranulation.

- **A vast repertoire of antigen-binding sites is achieved by random selection and recombination** of a limited number of V, D, and J gene segments that encode the variable (V) domains. This process is known as VDJ recombination and generates the primary antibody repertoire.

- **Repeated rounds of somatic hypermutation and selection** act on the primary repertoire to generate antibodies with higher specificity and affinity for the stimulating antigen. Immunoglobulin class switching combines rearranged VDJ genes with various C region genes so that the same antigen receptor can serve a variety of effector functions.

IMMUNOGLOBULINS RECOGNIZE AND BIND ANTIGENS

The recognition of antigen is the hallmark of the specific adaptive immune response. Two distinct types of molecule are involved in this process:

- immunoglobulins (antibodies); and
- T cell antigen receptors (TCRs).

Structural and functional diversity are characteristic features of these molecules.

Antibody genes have diversified in different species by multiple gene duplications and subsequent divergence. In many species, including humans, diversity is further amplified by extensive gene recombination and somatic mutation during the lifetime of an individual.

Immunoglobulins function as membrane-bound antigen receptors on B cells and soluble circulating antibodies

Immunoglobulins are glycoproteins expressed as:

- membrane-bound receptors on the surface of B cells; or
- soluble molecules (secreted from plasma cells) and present in serum and tissue fluids.

Contact between the B cell receptor and the antigen it recognizes results in B cell activation and differentiation

Surface and secreted antibodies

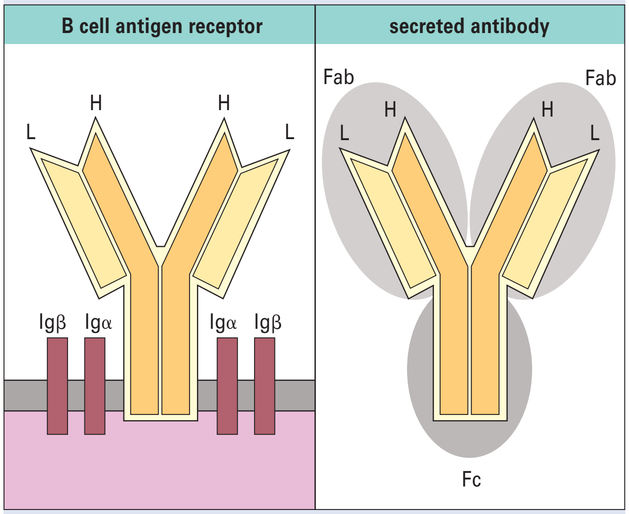

Fig. 3.1 The B cell antigen receptor (left) consists of two identical heavy (H) chains and two identical light (L) chains. In addition, secondary components (Igα and Igβ) are closely associated with the primary receptor and are thought to couple it to intracellular signaling pathways. Circulating antibodies (right) are structurally identical to the primary B cell antigen receptors except they lack the transmembrane and intracytoplasmic sections. Many proteolytic enzymes cleave antibody molecules into three fragments – two identical Fab (antigen binding) fragments and one Fc (crystallizable) fragment.

Rapid Reference Box 3

affinity – a measure of the binding strength between an antigenic determinant (epitope) and an antibody-combining site.

allotype – the protein of an allele that may be detectable as an antigen by another member of the same species.

antibody – a molecule produced in response to antigen that combines specifically with the antigen that induced its formation.

antigen – a molecule that reacts with antibody and the specific receptors on T and B cells.

avidity – the functional combining strength of an antibody with its antigen related to both the affinity of the reaction between the epitopes and paratopes, and the valencies of the antibody and antigen.

CD markers – cell surface molecules of leukocytes and platelets that are distinguishable with monoclonal antibodies and may be used to differentiate different cell populations (see Appendix 2).

cell-mediated immunity (CMI) – a term for immune reactions mediated by cells rather than by antibody or other humoral factors.

epitope – the part of an antigen that contacts the antigen-binding sites of an antibody or the T cell receptor.

idiotype – a single antigenic determinant on an antibody V region.

immunoglobulins – the serum antibodies, including IgG, IgM, IgA, IgE, and IgD.

isotype – refers to genetic variation within a family of proteins or peptides such that every member of the species will have each isotype of the family represented in its genome (e.g. immunoglobulin subclasses).

somatic mutation – a process occurring during B cell maturation and affecting the antibody gene region that permits refinement of antibody specificity.

recombination – a process by which genetic information is rearranged during meiosis; this process also occurs during the somatic rearrangements of DNA that occur in the formation of genes encoding antibody molecules and T cell receptors.

CD markers mentioned in this chapter

CD16 – FcγRIII (class of Fc receptor for IgG).

CD23 (CD23a and CD23b) – FcεRII (class of Fc receptor for IgE)

CD32 – FcγRII (class of Fc receptor for IgG).

CD64 – FcγRI (class of Fc receptor for IgG).

to generate **plasma cells**, which secrete large amounts of antibody.

The secreted antibody has the same binding specificity as the original B cell receptor (Fig. 3.1).

Immunoglobulins are a family of glycoproteins

Immunoglobulins are a family of glycoproteins. In humans, five distinct **classes of immunoglobulin** molecule are recognized namely IgG, IgA, IgM, IgD, and IgE. They differ in:

- size;
- charge;
- amino acid sequence; and
- carbohydrate content.

The IgG class can be subdivided into four **subclasses** (IgG1, IgG2, IgG3, IgG4) and the IgA into two subclasses (IgA1, IgA2).

The classes and subclasses together represent nine **isotypes** – present in all normal individuals.

Each isotype is defined by the sequence of the constant region of its heavy chain.

Some antigens provoke an antibody response in each of the immunoglobulin isotypes, whereas responses to other antigens may be relatively restricted (e.g. IgG2 responses to carbohydrate antigens).

ALL IMMUNOGLOBULIN ISOTYPES (EXCEPT IgD) ARE BIFUNCTIONAL

Each immunoglobulin isotype is bifunctional (except serum IgD) – that is, they:

- recognize and bind antigen; and then
- promote the killing and/or removal of the immune complex formed through the activation of effector mechanisms.

One part of the antibody molecule determines its antigen specificity while another determines which effector functions will be activated. Effector functions include binding of the immunoglobulin to:

- receptors expressed on host tissues (e.g. phagocytic cells);
- the first component (C1q) of the complement system to initiate the classical pathway complement cascade (Fig. 3.2).

Immunoglobulin class and subclass depend on the structure of the heavy chain

The basic structure of each immunoglobulin molecule is a unit consisting of:

- two light polypeptide chains; and
- two heavy polypeptide chains.

The amino acid sequences of the two light chains are identical; so are the sequences of the two heavy chains.

Both light and both heavy chains are folded into discrete domains, and the type of heavy chain determines the class and subclass of the antibody:

- μ (IgM);
- γ1, γ2, γ3, and γ4 (IgG1, IgG2, IgG3, IgG4);
- α1 and α2 (IgA1, IgA2);
- δ (IgD);
- ε (IgE).

There are no subclasses of IgM, IgD, or IgE (Fig. 3.3).

Different immunoglobulin classes and subclasses activate different effector systems

The human IgG subclasses (IgG1–IgG4), which are are present in serum in the approximate proportions of 66, 23, 7, and 4%, respectively, appear to have arisen after the divergence of evolutionary lines leading to humans and the mouse. Consequently there is no direct structural or functional correlation between the four human and mouse IgG subclasses, despite their similar nomenclature.

The relative proportions of IgA1 and IgA2 vary between serum and external secretion, where IgA is present in a secretory form (see Fig. 3.3).

Q. What advantage might there be in having such a variety of different antibody classes?

A. Each class and subclass of antibody can act as an adapter for different types of immune effector system. Because different types of effector system are appropriate for dealing with each pathogen, the immune system can adjust the type of response that is generated according to the type of antibodies that are produced.

IgG is the predominant immunoglobulin in normal human serum

IgG accounts for 70–75% of the total serum immunoglobulin pool and consists of a monomeric four-chain molecule of mass 146–170 kDa.

IgM accounts for about 10% of the serum immunoglobulin pool

IgM accounts for approximately 10% of the serum immunoglobulin pool. It is a pentamer of a basic four-chain structure with a mass of ~970 kDa, formation of which is aided by the inclusion of the J (joining) polypeptide chain of mass ~15 kDa.

A transmembrane monomeric form (mIgM) is present as an antigen-specific receptor on mature B cells.

IgA is the predominant immunoglobulin in seromucous secretions

IgA accounts for approximately 15–20% of the serum immunoglobulin pool. In humans over 80% of serum IgA is a monomer of the four-chain unit, but in most mammals serum IgA is predominantly polymeric, occurring mostly as a dimer.

Antibodies act as adapter molecules for immune effector systems

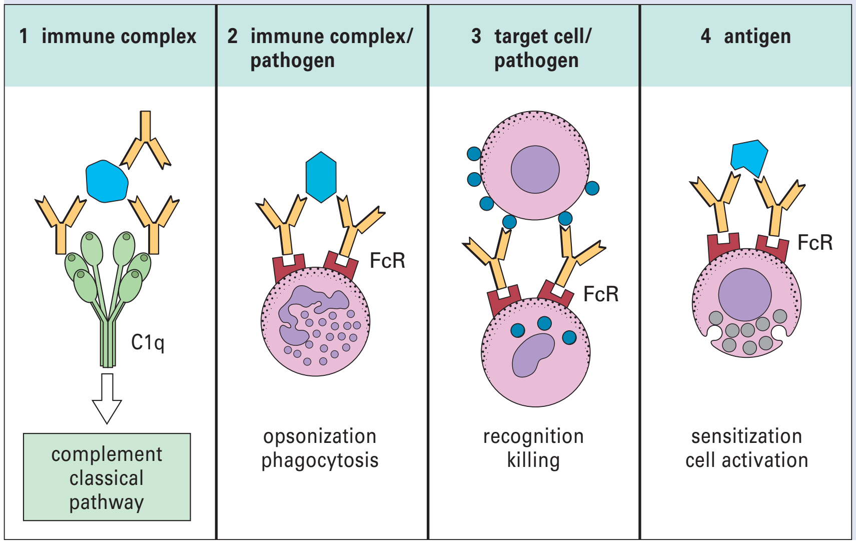
1 immune complex
2 immune complex/ pathogen
3 target cell/ pathogen
4 antigen
C1q
complement classical pathway
FcR
opsonization phagocytosis
FcR
recognition killing
FcR
sensitization cell activation

Fig. 3.2 Antibodies act as adapter molecules for different immune effector systems, linking antigens to receptor molecules (C1q and FcR) of the immune system. (**1**) Immune complexes can activate the complement classical pathway. (**2**) Antibodies bound to the surface of pathogens opsonize them for phagocytosis. (**3**) Antibodies bound to cells can promote their recognition and killing by NK cells. (Similarly recognition of some parasitic worms by eosinophils, mediated by antibodies, targets them for killing.) (**4**) Antibody bound to Fc receptors sensitizes cells so that they can recognize antigen, and the cell becomes activated if antigen binds to the surface antibody.

Physicochemical properties of human immunoglobulin classes

property	immunoglobulin type									
	IgG1	IgG2	IgG3	IgG4	IgM	IgA1	IgA2	sIgA	IgD	IgE
heavy chain	γ_1	γ_2	γ_3	γ_4	μ	α_1	α_2	α_1/α_2	δ	ϵ
mean serum conc. (mg/ml)	9	3	1	0.5	1.5	3.0	0.5	0.05	0.03	0.00005
sedimentation constant	7s	7s	7s	7s	19s	7s	7s	11s	7s	8s
mol. wt (kDa)	146	146	170	146	970	160	160	385	184	188
half-life (days)	21	20	7	21	10	6	6	?	3	2
% intravascular distribution	45	45	45	45	80	42	42	trace	75	50
carbohydrate (%)	2–3	2–3	2–3	2–3	12	7–11	7–11	7–11	9–14	12

Fig. 3.3 Each immunoglobulin class has a characteristic type of heavy chain. Thus IgG posesses γ chains; IgM, μ chains; IgA, α chains; IgD, δ chains; and IgE, ε chains. Variation in heavy chain structure within a class gives rise to immunoglobulin subclasses. For example, the human IgG pool consists of four subclasses reflecting four distinct types of heavy chain. The properties of the immunoglobulins vary between the different classes. In secretions, IgA occurs in a dimeric form (sIgA) in association with a protein chain termed the secretory component. The serum concentration of sIgA is very low, whereas the level in intestinal secretions can be very high.

IgA is the predominant immunoglobulin in seromucous secretions such as saliva, colostrum, milk, and tracheobronchial and genitourinary secretions.

Secretory IgA (sIgA) is comprised of a dimeric form of the basic unit together with the J chain and a secretory component, and has a total mass of ~385 kDa.

mIgD is an antigen-specific receptor on mature B cells

IgD accounts for less than 1% of the serum immunoglobulin pool.

A transmembrane monomeric form (mIgD) is present as an antigen-specific receptor on mature B cells. (B cells may have both mIgM and mIgD as their receptor, though on any one B cell these receptors will have the same antigen specificity.)

Basophils and mast cells are continuously saturated with IgE

Serum IgE levels are very low (< 0.05 μg/ml) relative to the other immunoglobulin isotypes, However, basophils and mast cells express an IgE-specific receptor of very high affinity with the result that they are continuously saturated with IgE.

IgE is a monomeric immunoglobulin of ~188 kDa.

IMMUNOGLOBULINS ARE MADE UP OF FOUR POLYPEPTIDE CHAIN UNITS

The basic four-chain structure of immunoglobulin molecules is illustrated by IgG1 (Fig. 3.4). This four-chain structure of antibody molecules was first demonstrated for rabbit IgG.

The **light chains** (25 kDa) are bound to the heavy chains (55 kDa) by interchain disulfide bridges and multiple non-covalent interactions.

The **heavy chains** are similarly bound to each other by interchain disulfide bridges and multiple non-covalent interactions.

Each segment of approximately 110 amino acids is folded to form a compact domain, which is stabilized through a covalent intrachain disulfide bond. Thus:

- the light chain has an intrachain disulfide bond in each of the VL and CL domains (Fig. 3.5 and see Fig. 3.4);
- there is one intrachain disulfide bond in each domain of the heavy chain.

Each disulfide bond encloses a peptide loop of 60–70 amino acid residues.

The amino acid sequence of immunoglobulin domains exhibits significant homology and this is reflected in a common conformational motif, referred to as the **immunoglobulin fold**. This characteristic fold defines the immunoglobulin superfamily members.

Q. What conditions are required to dissociate the heavy and light chains in an immunoglobulin molecule?

A. Because the heavy and light chains are linked by disulfide bonds, it is necessary to reduce these first before dissociating the chains under conditions of low pH.

VL and VH form antigen-binding sites and CL and CH determine effector functions

The light chains of most vertebrates have been shown to exist in two structurally distinct forms:

- kappa (κ); and
- lambda (λ).

These are isotypes, being present in all individuals.

Genetic variants (known as **allotypes**) of the κ chains exist in different individuals (see p. 79).

Either light chain type may combine with any of the heavy chain types, but in any individual immunoglobulin

The basic structure of IgG1

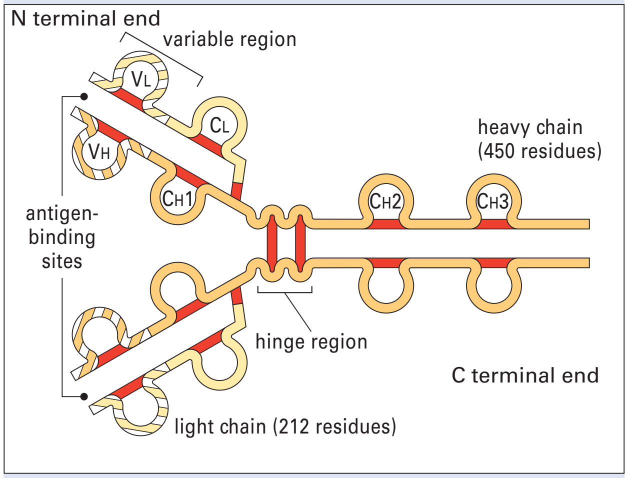

Fig. 3.4 The N terminal end of IgG1 is characterized by sequence variability (V) in both the heavy and light chains, referred to as the VH and VL regions, respectively. The rest of the molecule has a relatively constant (C) structure. The constant portion of the light chain is termed the CL region. The constant portion of the heavy chain is further divided into three structurally discrete regions: CH1, CH2, and CH3. These globular regions, which are stabilized by intrachain disulfide bonds, are referred to as 'domains'. The sites at which the antibody binds antigen are located in the variable domains. The hinge region is a segment of heavy chain between the CH1 and CH2 domains. Flexibility in this area permits the two antigen-binding sites to operate independently. There is close pairing of the domains except in the CH2 region (see Fig. 3.8). Carbohydrate moieties are attached to the CH2 domains.

molecule both light chains and heavy chains are of the same type.

Amino sequence analysis of monoclonal mouse and human light chains has revealed two structurally distinct regions:

- the sequence of the ~110 N terminal residues was seen to be unique for each protein analysed;
- the C terminal sequence (~110 residues) was constant for a given isotype (κ or λ) and allotype.

Thus, the light chain variable (VL) and constant (CL) regions were defined and these regions correspond to the domains (see Fig. 3.5).

Similarly, the ~110 N terminal residues of heavy chains were seen to be unique for each protein analyzed whereas the remaining constant domains were characteristic for each immunoglobulin isotype.

The constant domains of the heavy chains are generally designated as CH1, CH2, CH3, and CH4, or according to the isotype of the constant domains.

Hypervariable regions form the antigen-combining site

Within the variable regions of both heavy and light chains, some polypeptide segments show exceptional variability and are termed **hypervariable regions**. These segments are located around amino acid positions 30, 50, and 95 (Fig. 3.6) and are referred to as Hv1, Hv2, and Hv3 or Lv1, Lv2, and Lv3, respectively.

Basic folding in the light chain

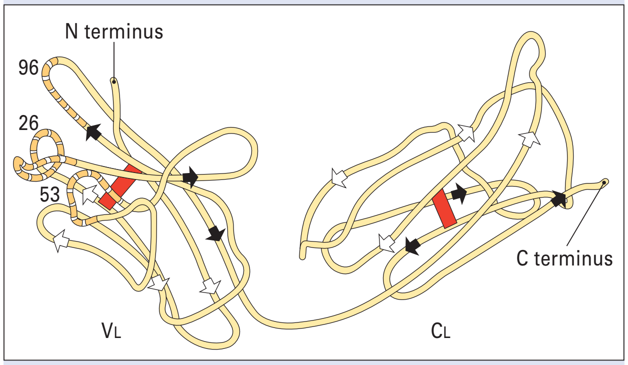

Fig. 3.5 The immunoglobulin domains in the light chain share a basic folding pattern with several straight segments of polypeptide chain lying parallel to the long axis of the domain. Light chains have two domains – one constant and one variable. Within each domain, the polypeptide chain is arranged in two layers, running in opposite directions, with many hydrophobic amino acid side chains between the layers. One of the layers has four segments (arrowed white), the other has three (arrowed black); both are linked by a single disulfide bond (red). Folding of the VL domains causes the hypervariable regions (see Fig. 3.6) to become exposed in three separate, but closely disposed, loops. One numbered residue from each hypervariable region is identified.

X-ray crystallographic studies show that the hypervariable regions are intimately involved in antigen binding and hence in creating an interaction site that is complementary in shape, charge, and hydrophobicity to the epitope it binds. Consequently the hypervariable regions are also termed the complementarity determining regions (CDR1, CDR2, and CDR3).

The intervening peptide segments are called framework regions (FRs) and determine the fold that ensures the CDRs are in proximity to each other (see Fig. 3.6).

THE FIVE CLASSES OF ANTIBODY IN MAMMALS ARE IgG, IgA, IgM, IgD, AND IgE

The overall structure of an antibody depends on its class and subclass

X-ray crystallography has provided structural data on complete IgG molecules providing computer-generated α-carbon backbone and van der Waals surface models for this class of immunoglobulin (Fig. 3.7). Mobility around the hinge region allows for the generation of the Y- and T-shaped structures visualized by electron microscopy.

In IgG, there is extensive pairing between VH/VL and CH1/CL domains through extensive non-covalent interactions to form the IgG Fab region.

Amino acid variability in the variable regions of immunoglobulins

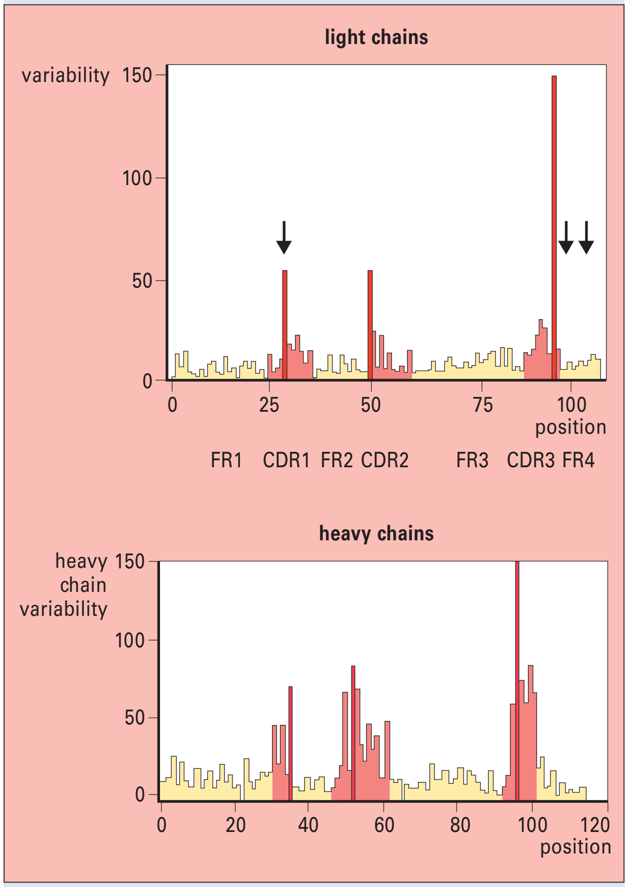

Fig. 3.6 Variability is calculated by comparing the sequences of many individual chains and, for any position, is equal to the ratio of the number of different amino acids found at that position, to the frequency of the most common amino acid. The areas of greatest variability, of which there are three in the VL domain, are termed the hypervariable regions. The areas shaded pink denote regions of hypervariability (CDR), and the most hypervariable positions are shaded red. The four framework regions (FR) are shown in yellow. (Courtesy of Professor EA Kabat)

Model of an IgG molecule

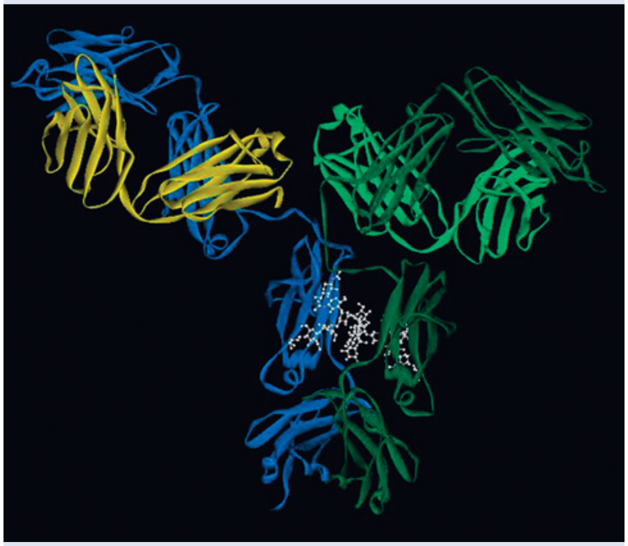

Fig. 3.7 A model of an IgG molecule showing the polypeptide backbones of the four chains as a ribbon. Heavy chains are shown in dark blue and dark green. The antigen-binding sites are at the tips of the arms of the Y-shaped molecule and are formed by domains from both the heavy and light chains. The extended, unfolded hinge region lies at the center of the molecule. Carbohydrate units are shown as ball and stick structures, covalently linked to the Fc region.

Similarly, the CH3 domains of the IgG Fc are paired.

The CH2 domains are not paired, but bear an N-linked complex oligosaccharide moiety that is integral to the conformation of the IgG Fc.

Despite the structural similarities between V and C domains, there are significant differences in the manner in which they pair.

Each IgG subclass has a unique hinge region

Each of the four human IgG subclasses exhibits a unique profile of biological activities, despite the fact that there is over 95% sequence identity between their Fc regions.

The structures of the hinge regions are particularly distinct (Fig. 3.8) and contribute to the relative mobilities of the IgG Fab and IgG Fc moieties within the intact molecule and, presumably, the structural form of immune complexes.

The heavy chains bear N-linked oligosaccharides and although the oligosaccharide accounts for only 2–3% of the mass of IgG molecules it is crucial to the expression of effector functions.

Assembled IgM molecules have a 'star' conformation

IgM is present in human serum as a pentamer of the basic four-chain unit (see Fig. 3.8). Each heavy chain is comprised of a VH and four CH domains.

The subunits of the pentamer are linked by covalent disulfide bonds between adjacent CH2 and CH3 domains, and the C terminal 18-residue peptide sequence, referred to as the 'tailpiece'.

Assembly of the pentamer is further facilitated by incorporation of the J (joining) polypeptide chain. J chain is synthesized within plasma cells, has a mass of ~15 kDa and folds to form an immunoglobulin domain. It has four oligosaccharide moieties that account for approximately 12% of the mass.

In electron micrographs the assembled IgM molecule is seen to have a 'star' conformation with a densely packed central region and radiating arms (Fig. 3.9). Electron micrographs of IgM antibodies binding to poliovirus show molecules adopting a 'staple' or 'crab-like' configuration (see Fig. 3.9), which suggests that flexion readily occurs

Structural characteristics of various human immunoglobulins

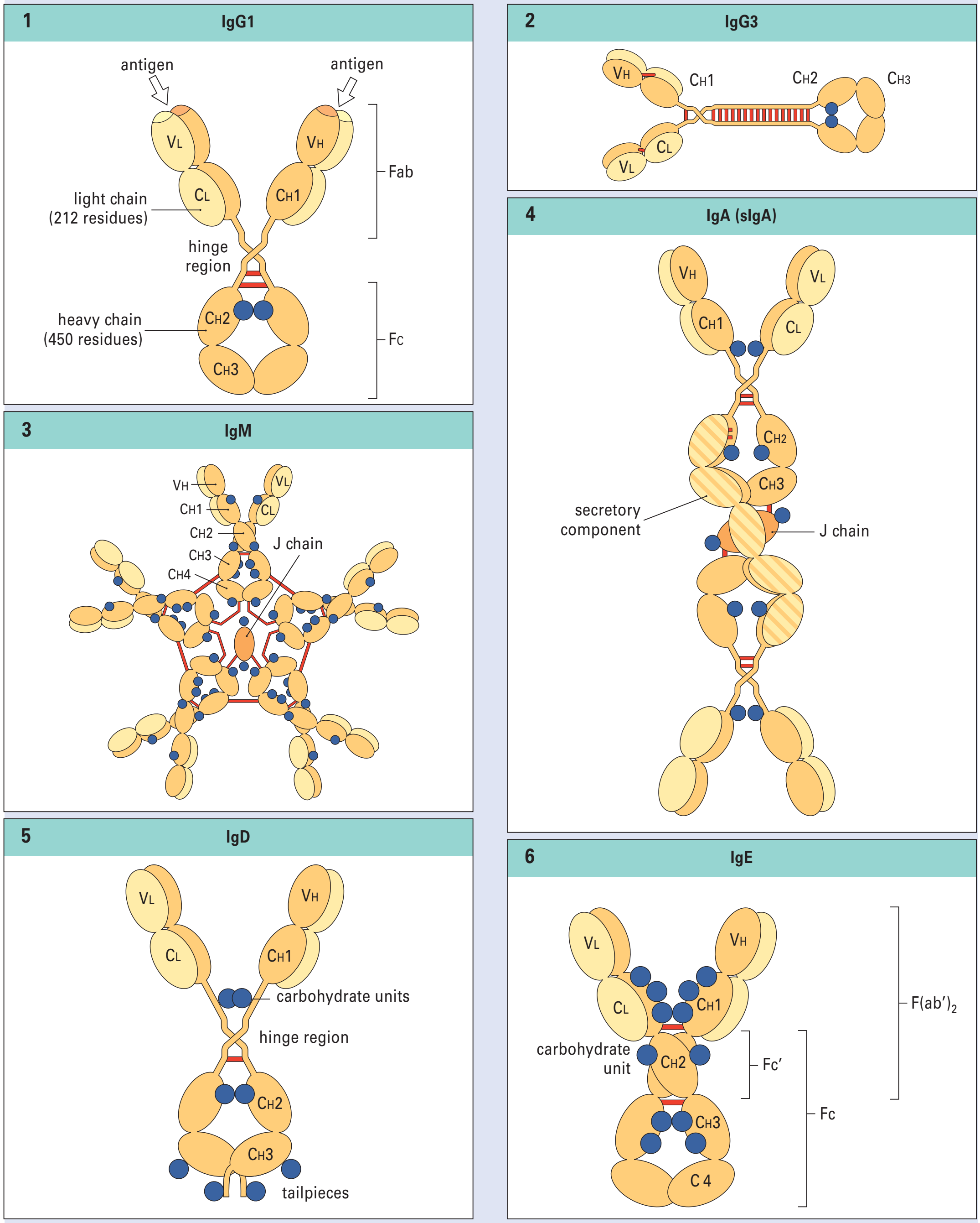
1
IgG1
antigen
antigen
VL
VH
CL
CH1
light chain
(212 residues)
hinge
region
Fab
CH2
heavy chain
(450 residues)
CH3
Fc
2
IgG3
VH
CH1
CH2
CH3
CL
VL
3
IgM
VH
VL
CH1
CL
CH2
CH3
CH4
J chain
4
IgA (sIgA)
VH
VL
CH1
CL
CH2
CH3
secretory
component
J chain
5
IgD
VL
VH
CL
CH1
carbohydrate units
hinge region
CH2
CH3
tailpieces
6
IgE
VL
VH
CL
CH1
F(ab')2
carbohydrate
unit
CH2
Fc'
Fc
CH3
C 4

Fig. 3.8 Carbohydrate side chains are shown in blue. Inter heavy (H) chain disulfide bonds are shown in red, but interchain bonds between H and L chains are omitted. (**1**) A model of IgG1 indicating the globular domains of H and L chains. Note the apposition of the CH3 domains and the separation of the CH2 domains. The carbohydrate units lie between the CH2 domains. (**2**) Polypeptide chain structure of human IgG3. Note the elongated hinge region. (**3**) IgM H chains have five domains with disulfide bonds cross-linking adjacent CH3 and CH4 domains. The possible location of the J chain is shown. IgM does not have extended hinge regions, but flexion can occur about the CH2 domains. (**4**) The secretory component of sIgA is probably wound around the dimer and attached by two disulfide bonds to the CH2 domain of one IgA monomer. The J chain is required to join the two subunits. (**5**) This diagram of IgD shows the domain structure and a characteristically large number of oligosaccharide units. Note also the presence of a hinge region and short octapeptide tailpieces. (**6**) IgE can be cleaved by enzymes to give the fragments $F(ab')_2$, Fc, and Fc'. Note the absence of a hinge region.

between the CH2 and CH3 domains, though this region is not structurally homologous to the IgG hinge. Distortion of this region, known as **dislocation**, results in the 'staple' configuration of IgM required to activate complement.

sIgA is composed of IgA, J chain, and a secretory component

IgA is present in serum as a monomer of the basic four-chain structure. Each heavy chain is comprised of a VH and three CH domains.

The IgA1 and IgA2 subclasses differ substantially in the structure of their hinge regions:

- the hinge of IgA1 is extended and bears O-linked oligosaccharides;
- the hinge of of IgA2 is truncated.

Both heavy chains bear N-linked oligosaccharides.

IgA is the predominant antibody isotype in external secretions and exists as a complex secretory form. IgA is secreted by plasma cells as a dimer in which a heavy chain 'tailpiece' is disulfide linked to a J chain (see Fig. 3.8).

Electron micrographs of IgA dimers show double Y-shaped structures, suggesting that the monomeric subunits are linked end-to-end through the C terminal Cα3 regions (Fig. 3.10).

The dimeric form of IgA binds a **poly-Ig receptor** (Fig. 3.11) expressed on the basolateral surface of epithelial cells. The complex is internalized, transported to the apical surface, and released as the secretory form of IgA (sIgA), which has a mass of ~385 kDa and is comprised of:

- two units of IgA;
- J chain; and
- a secretory component (mass 70 kDa) (see Figs 3.8 and 3.11).

Secretory component results from enzymatic cleavage of the IgA–poly-Ig receptor complex. sIgA is relatively resistant to cleavage by enzymes in the gut.

The heavy chain of IgD bears three N-linked oligosaccharides

Serum IgD accounts for less than 1% of the total serum immunoglobulin. Each heavy chain is comprised of a VH domain and three CH domains with an extended hinge

Electron micrographs of IgM molecules

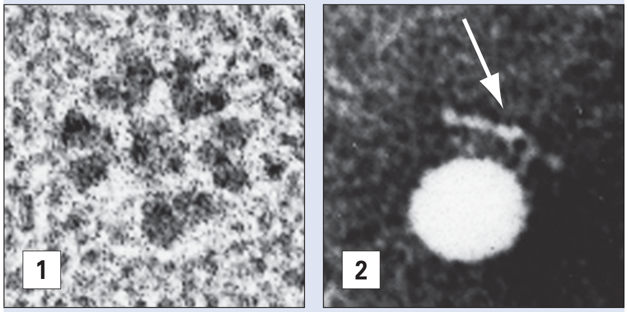

Fig. 3.9 (**1**) In free solution, deer IgM adopts the characteristic star-shaped configuration. ×195 000. (Courtesy of Drs E Holm Nielson, P Storgaard, and Professor S-E Svehag) (**2**) Rabbit IgM antibody (arrow) in 'crab-like' configuration with partly visible central ring structure bound to a poliovirus virion. × 190 000. (Courtesy of Dr B Chesebro and Professor S-E Svehag)

Electron micrographs of human dimeric IgA molecules

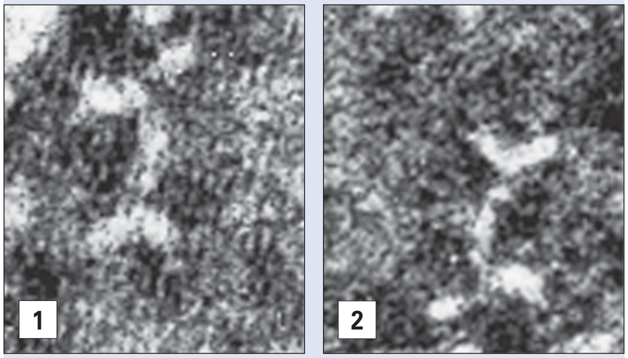

Fig. 3.10 The double Y-shaped appearance suggests that the monomeric subunits are linked end to end through the C terminal Cα3 domain. ×250 000. (Courtesy of Professor S-E Svehag)

Transport of IgA across the mucosal epithelium

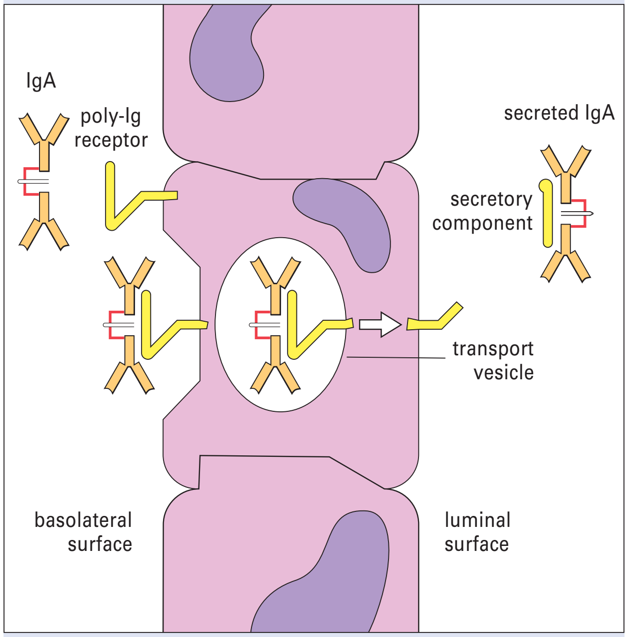

Fig. 3.11 IgA dimers secreted into the intestinal lamina propria by plasma cells bind to poly-Ig receptors on the internal (basolateral) surface of the epithelial cells. The sIgA–receptor complex is then endocytosed and transported across the cell while still bound to the membrane of transport vesicles. These vesicles fuse with the plasma membrane at the luminal surface, releasing IgA dimers with bound secretory component derived from cleavage of the receptor. The dimeric IgA is protected from proteolytic enzymes in the lumen by the presence of this secretory component.

region that is susceptible to proteolysis, at least on purification.

The heavy chain bears three N-linked oligosaccharides and the hinge region bears multiple O-linked oligosaccharides (see Fig. 3.8).

The heavy chain of IgE bears six N-linked oligosaccharides

It is estimated that serum IgE accounts for approximately 50% of total body IgE, with the rest being bound to mast cells and basophils through their high-affinity IgE Fcε receptor (FcεRI).

Each heavy chain is comprised of a VH and four CH domains and bears six N-linked oligosaccharides (see Fig. 3.8).

Immunoglobulins are prototypes of the immunoglobulin superfamily

The type of domain referred to as the immunoglobulin fold was first identified in antibodies. However, it also occurs in many other molecules that may or may not have an overt immune function. Examples include:

- the adhesion molecules ICAM-1 and VCAM-1 (see Chapter 6);
- the TCR (see Chapter 5);
- MHC molecules (see Chapter 5);
- cellular receptors for antibodies.

Such molecules are said to belong to the **immunoglobulin supergene family (IgSF)**.

The principal elements of the domain are two opposed β-pleated sheets, with stabilization by one or more disulfide bonds between the β-pleated sheets. This structure is sometimes referred to as a β **barrel**.

The immunoglobulin domain structure developed early in evolution and has been used and adapted in different molecules as vertebrates have radiated and diverged.

ANTIGEN-BINDING SITES OF ANTIBODIES ARE SPECIFIC FOR THE THREE-DIMENSIONAL SHAPE OF THEIR TARGET

Conformations of target antigen and receptor-binding site are complementary

Protein molecules are not rigid structures, but exist in a dynamic equilibrium between structures that differ in their ability to form a **primary interaction** with specific **ligands**.

Following a primary interaction, each 'partner' may influence the final **conformation** within the complex. This concept approximates to the 'induced fit' model of protein–protein interactions.

An examination of the interaction between the Fab fragment of the mouse D1.3 monoclonal antibody and hen egg white lysozyme (HEL) reveals the complementary surfaces of the epitope and the antibody's combining site. These surfaces extend beyond the hypervariable regions. In total, 17 amino acid residues on the antibody contact 16 residues on the lysozyme molecule (Fig. 3.12). All hypervariable regions in both the heavy and light chains contribute to the antibody-binding site, though the third hypervariable region in the heavy chain lying at the center of the combining site appears to be most important.

The Fab–lysozyme complex

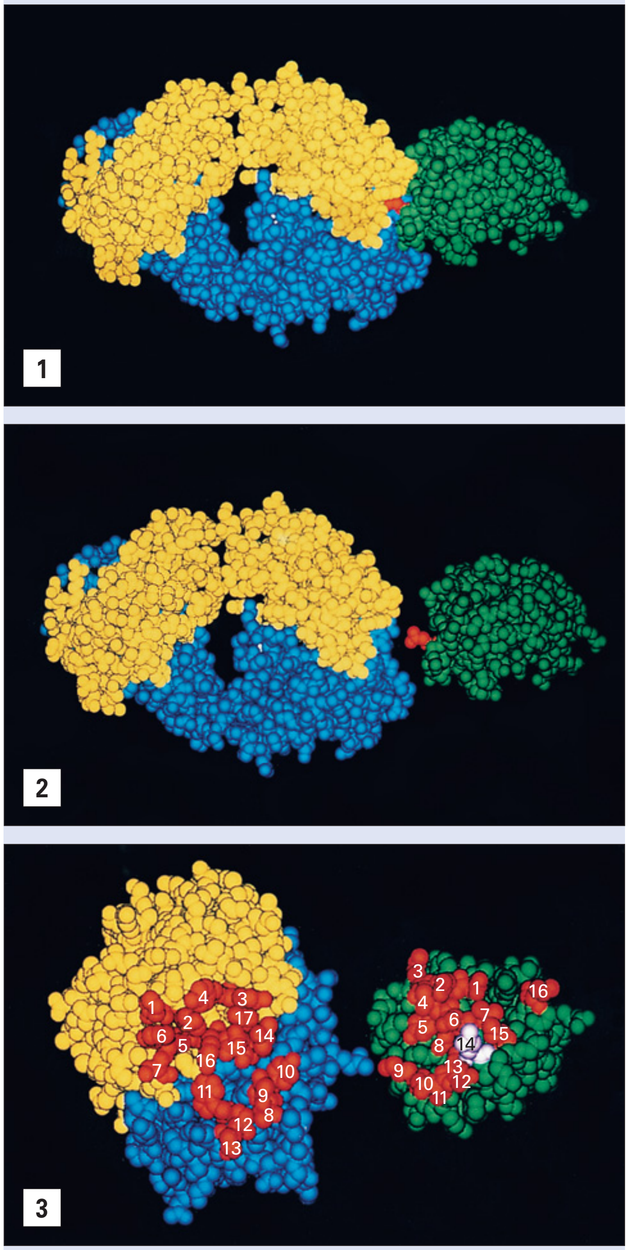

Fig. 3.12 (**1**) Lysozyme (green) binds to the hypervariable regions of the heavy (blue) and light (yellow) chains of the Fab fragment of antibody D1.3. (**2**) The separated complex with Glu121 visible (red). This residue fits into the center of the cleft between the heavy and light chains. (**3**) The same molecules rotated 90° to show the contact residues that contribute to the antigen–antibody bond. (Reprinted with permission from Poljak RJ, Science 1986;233:747–753. Copyright 1986 AAAS and reprinted with permission from Garcia KC et al, Science 1996;274:209–219. Copyright 1996 AAAS)

The binding site of this D1.3 monoclonal antibody may be regarded as 'classical'. The structures of other lysozyme– antibody complexes have been solved and show differing involvement of hypervariable and framework residues.

These studies are essential when engineering antibody molecules (e.g. when 'humanizing' an antibody to generate an antibody therapeutic, see Method Box 3.1).

ANTIBODY AFFINITY IS A MEASURE OF THE STRENGTH OF INTERACTION BETWEEN AN ANTIBODY-COMBINING SITE AND ITS EPITOPE

X-ray crystallographic studies of antibody Fab fragments show that the hypervariable regions are clustered at the end of the Fab arms to form a unique surface topography that is complementary to structures on the antigen. This complementary surface is called a **paratope** (corresponding to the antigenic determinant known as an **epitope**).

Proximal framework residues sometimes contribute to antigen binding, but their main role is to maintain the immunoglobulin fold and the integrity of the binding site.

Antibodies form multiple non-covalent bonds with antigen

The antigen–antibody interaction results from the formation of multiple non-covalent bonds. These attractive forces consist of:

- hydrogen bonds;
- electrostatic bonds;
- van der Waals forces; and
- hydrophobic forces.

Each bond is relatively weak in comparison with covalent bonds, but together they can generate a high-affinity interaction.

The strength of a non-covalent bond is critically dependent on the distance (d) between the interacting groups, being proportional to $1/d^2$ for electrostatic forces, and to $1/d^7$ for van der Waals forces.

Thus interacting groups must be in quite intimate contact before these attractive forces come into play.

For a paratope to combine with its epitope (see Fig. 3.12) the interacting sites must be complementary in shape, charge distribution, and hydrophobicity, and in terms of donor and acceptor groups capable of forming hydrogen bonds.

Close proximity of two protein surfaces can also generate repulsive forces (proportional to $1/d^{12}$) if electron clouds overlap.

In combination, the attractive and repulsive forces have a vital role in determining the specificity of the antibody molecule and its ability to discriminate between structurally similar molecules.

The great specificity of the antigen–antibody interaction is exploited in a number of widely used assays (Method Box 3.2)

All antigen–antibody interactions are reversible

The strength of the interaction between an antigen and an antibody is loosely referred to as **affinity**. It is the sum of the attractive and repulsive forces resulting from binding between the binding site of a monovalent Fab fragment and its epitope. (Note that true affinity is a thermodynamic parameter and its determination is much more demanding, though not discussed here.)

The interaction between an antigen and an antibody will be reversible, so at equilibrium the Law of Mass Action can be applied and an equilibrium constant, K (the association constant), can be determined (Fig. 3.13).

Avidity is likely to be more relevant than affinity

Because each antibody unit of four polypeptide chains has two antigen-binding sites, antibodies are potentially multivalent in their reaction with antigen.

In addition, antigen can be:

- monovalent (e.g. small chemical groups, haptens); or
- multivalent (e.g. microorganisms).

The strength with which a multivalent antibody binds a multivalent antigen is termed **avidity** to differentiate it from the **affinity**, which is determined for a univalent antibody fragment binding to a single antigenic determinant.

The avidity of an antibody for its antigen is dependent on the affinities of the individual antigen-combining sites for the epitopes on the antigen. Avidity will be greater than the sum of these affinities if both antibody-binding sites bind to the antigen because all antigen–antibody bonds would have to be broken simultaneously for the complex to dissociate (Fig. 3.14).

In physiological situations, avidity is likely to be more relevant than affinity because antibodies are minimally divalent and most naturally occurring antigens are multivalent.

In practice we determine the association constant at equilibrium when the rate of formation of complex (ka) is equal to the spontaneous rate of dissociation (kd). The association or equilibrium constant is defined as $K = ka/kd$.

It has been suggested that B cell selection and stimulation during a maturing antibody response depend upon selection for the ability of antibodies to bind to antigens both:

- rapidly (kinetic selection); and
- tightly (thermodynamic selection).

Cross-reactive antibodies recognize more than one antigen

Antigen–antibody reactions can show a high level of specificity, but can also be **cross-reactive**, binding to a structurally related but different antigen.

The monoclonal antibodies to hen egg lysozyme (HEL) may therefore also bind the structurally homologous duck egg lysozyme (DEL), though the D1.3 antibody illustrated in Fig. 3.12 does not bind DEL (due to a single amino acid substitution within the epitope).

A polyclonal antiserum to HEL will contain populations of antibodies specific for HEL and others that cross-react with DEL (Fig. 3.15).

Antibodies recognize the conformation of antigenic determinants

Antibodies are capable of expressing remarkable specificity and are able to distinguish small differences in the

METHOD BOX 3.1
Recombinant antibodies for human therapy

Recombinant antibody therapeutics (rMAbs) are likely to become the largest family of disease-modifying drugs available to clinicians. Their efficacy results from specificity for the target antigen and biological activities (effector functions) activated by the immune complexes formed. Currently, 14 rMAbs are licensed and hundreds are in clinical trials or under development. Initial trials administered mouse monoclonal antibodies specific for human targets and provided ‘proof of principle’.

Q. What problems could you envisage with the use of mouse antibodies to treat diseases in humans?

A. The antibodies might not interact appropriately with human Fc receptors. In the longer term an individual might mount an immune response against the non-self mouse antibodies.

However, individuals mounted immune responses against the non-self mouse antibodies and the development of these human anti-mouse antibody responses (HAMAs) meant that repeated dosing, which was required for chronic diseases, was not possible.

In response, scientists then produced:

- chimeric antibodies, in which the mouse V domains were linked to human antibody C domains;
- more recently, fully humanized antibodies.

This development of humanized antibodies has successively moderated, but not eliminated, such responses and sometimes human anti-human antibody (HAHA) responses are encountered.

Although such therapeutic antibodies are termed ‘human’, they are derived in a manner (phage display or Ig transgenic mice) that generates specificities that would ordinarily be eliminated by an individual’s immune system because they are directed against self.

The success of antibodies such as infliximab (anti-TNFα) and rituximab (anti-CD20) has resulted in demands for their production in metric tonnes. The biopharmaceutical industry has met the challenge with the construction of new mammalian cell culture facilities (10 000–20 000-liter capacities) to produce rMAbs; however, productivity, cost, and potency remain to be optimized.

All antibody therapeutics that are currently licensed have been produced by mammalian cell culture using Chinese hamster ovary (CHO) cells or the mouse NSO or Sp2/0 plasma cell lines; other systems under development and evaluation include transgenic animals, fungi, and plants.

The efficacy of an antibody therapeutic is critically dependent on appropriate post-translational modifications (PTMs) and each production system offers a different challenge because PTMs show species, tissue, and site specificity. Essential PTMs are relevant not only to potency, but also to the potential immunogenicity of the product.

Further improvement of the potency of rMAbs depends on optimizing the ‘downstream’ biological effector functions activated in vivo by the immune complexes formed. This offers a considerable challenge due to the difficulty of monitoring events in vivo.

Production of Fv antibodies by phage display

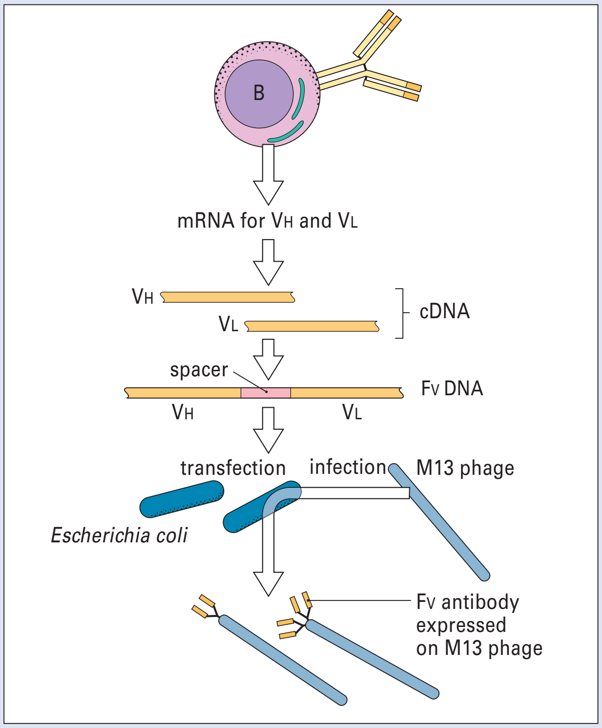

Fig. 1 Artificial antibodies such as Fv fragments, consisting of a hybrid of antibody VH and VL regions, can be generated by phage display. Antibody VH and VL genes are first amplified from B cell mRNA by the polymerase chain reaction. The genes are joined together with a spacer to give a gene for an Fv fragment. Bacteria are then transfected with the gene in a phagemid vector containing a leader sequence, a fragment of the gene expressing phage coat protein 3, and an M13 origin of replication, and then infected with M13 phage. The phages replicate and express the Fv on their tips. Phages displaying the right specificity are isolated by panning on antigen-coated plates, and amplified. The antigen-specific phage can be used to infect strains of bacteria that allow the secretion of the Fv protein into the culture medium.

A number of genetically modified animal models are under development that might mimic human biology, for example transgenic and knockout mice that produce human antibodies and express human cellular IgG Fc receptors (FcγR).

At present we extrapolate from effector functions activated in vitro to presumed activity in vivo, for example complement-dependent cytotoxicity (CDC), antibody-dependent cellular cytotoxicity (ADCC), and the induction of apoptosis.

In recent years in-vitro studies have unequivocally established that essential effector functions are dependent on appropriate glycosylation of the antibody molecule.

Glycosylation has therefore been a focus of attention for the biopharmaceutical industry for the past several years because regulatory authorities require that a consistent human-type glycosylation is achieved for rMAbs, irrespective of the system in which they are produced.

METHOD BOX 3.2
Assays for antibodies and antigens

Immunoassays use labeled reagents for detecting either antibodies or antigens. They are very sensitive and economical in the use of reagents.

Solid-phase assays for antibodies using detection reagents labeled with radioisotopes (radioimmunoassay, RIA) or with enzymes (enzyme-linked immunosorbent assay, ELISA) are probably the most widely used of all immunological assays, because large numbers can be performed in a relatively short time.

Similar assays using fluorescent or chemiluminescent reagents to detect the bound antibody are also in common use.

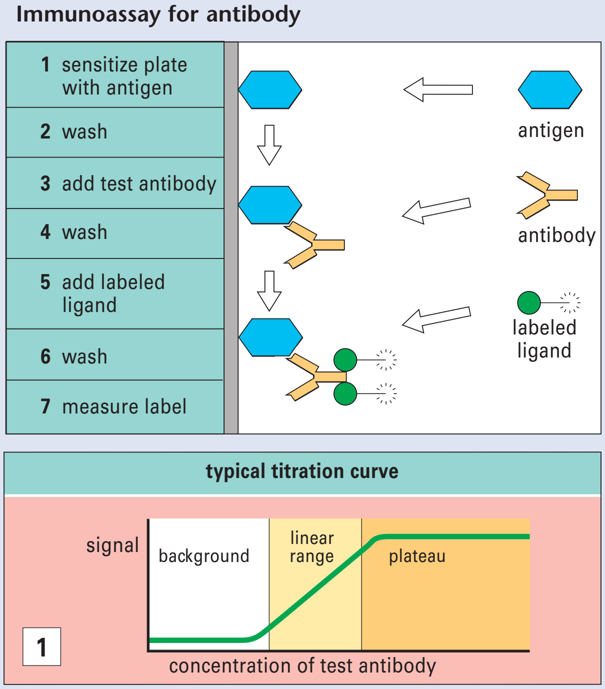
Immunoassay for antibody
1 sensitize plate with antigen
2 wash
3 add test antibody
4 wash
5 add labeled ligand
6 wash
7 measure label
antigen
antibody
labeled ligand
typical titration curve
signal
background
linear range
plateau
concentration of test antibody
1

Fig. 1 (**1**) Antigen in saline is incubated on a plastic plate or tube, and small quantities become absorbed onto the plastic surface. (**2**) Free antigen is washed away. (The plate may then be blocked with excess of an irrelevant protein to prevent any subsequent non-specific binding of proteins.) (**3**) Test antibody is added, which binds to the antigen. (**4**) Unbound proteins are washed away. (**5**) The antibody is detected by a labeled ligand. The ligand may be a molecule such as staphylococcal protein A that binds to the Fc region of IgG – more often it is another antibody specific for the test antibody. By using a ligand that binds to particular classes or subclasses of test antibody it is possible to distinguish isotypes. (**6**) Unbound ligand is washed away. (**7**) The label bound to the plate is measured. A typical titration curve is shown in the graph. With increasing amounts of test antibody the signal rises from a background level through a linear range to a plateau. Antibody titers can only be detected correctly within the linear range. Typically the plateau binding is 20–100 times the background. The sensitivity of the technique is usually about 1–50 ng/ml of specific antibody. Specificity of the assay may be checked by adding increasing concentrations of free test antigen to the test antibody at step 3; this binds to the antibody and blocks it from binding to the antigen on the plate. Addition of increasing amounts of free antigen reduces the signal.

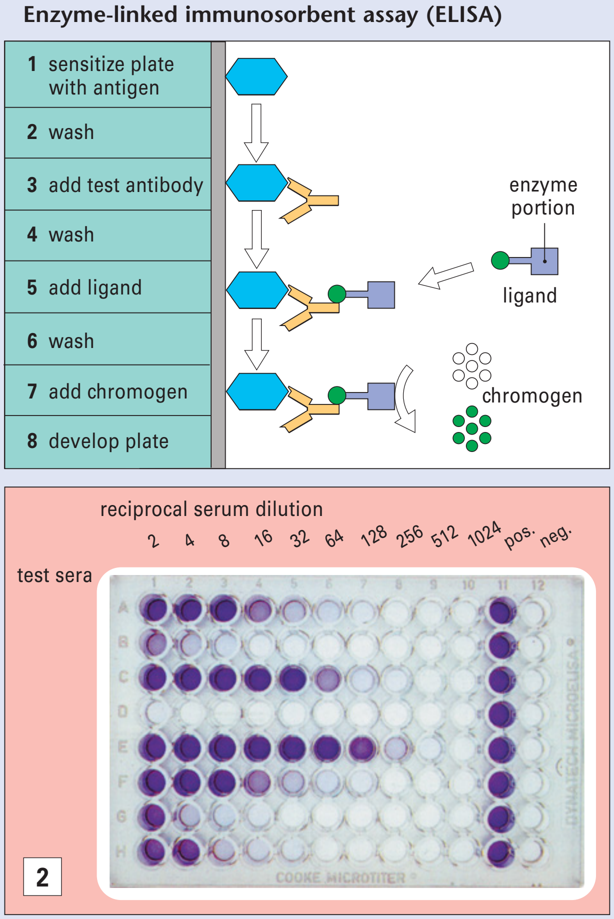
Enzyme-linked immunosorbent assay (ELISA)
1 sensitize plate with antigen
2 wash
3 add test antibody
4 wash
5 add ligand
6 wash
7 add chromogen
8 develop plate
enzyme portion
ligand
chromogen
reciprocal serum dilution
2
4
8
16
32
64
128
256
512
1024
pos.
neg.
test sera
A
B
C
D
E
F
G
H
1
2
3
4
5
6
7
8
9
10
11
12
COOKE MICROTITER
DYNATECH MICROELISA
2

Fig. 2 The ELISA plate is prepared in the same way as the immunoassay up to step 4. In this system, the ligand is a molecule that can detect the antibody and is covalently coupled to an enzyme such as peroxidase. This binds the test antibody, and after free ligand is washed away (**6**) the bound ligand is visualized by the addition of chromogen (**7**) – a colorless substrate that is acted on by the enzyme portion of the ligand to produce a colored end-product. A developed plate (**8**) is shown in the lower panel. The amount of test antibody is measured by assessing the amount of colored end-product by optical density scanning of the plate.

METHOD BOX 3.2
Assays for antibodies and antigens (*Continued*)

Related assays are used to detect and quantitate antigens. Two examples are the competition immunoassay and the two-site capture immunoassay.

Assay of antigen

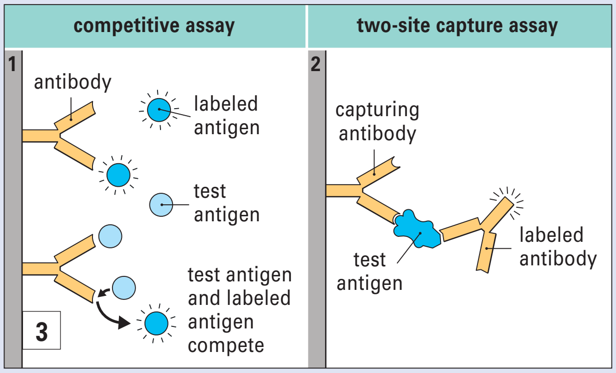
competitive assay
two-site capture assay
1
antibody
labeled antigen
test antigen
test antigen and labeled antigen compete
3
2
capturing antibody
test antigen
labeled antibody

Fig. 3 (**1**) Competitive assay. The test antigen is placed together with labeled antigen onto a plate coated with specific antibody. The more test antigen present, the less labeled standard antigen binds. This type of assay is often used to measure antigens at relatively high concentrations, or hormones that have only a single site available for combination with antibody. (**2**) Two-site capture assay. The assay plate is coated with specific antibody, the test solution then applied, and any antigen present captured by the bound antibody. After washing away unbound material, the captured antigen is detected using a labeled antibody against another epitope on the antigen. Because the antigen is detected by two different antibodies, the second in excess, such assays are both highly specific and sensitive.

Calculation of antibody affinity

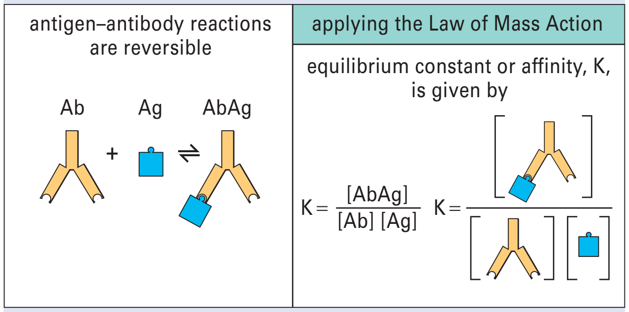
antigen–antibody reactions are reversible
Ab
Ag
AbAg
+
⇌
applying the Law of Mass Action
equilibrium constant or affinity, K, is given by
K = [AbAg] / [Ab] [Ag]
K =

Fig. 3.13 All antigen–antibody reactions are reversible. The Law of Mass Action can therefore be applied, and the antibody affinity (given by the equilibrium constant, K) can be calculated. (Square brackets refer to the concentrations of the reactants)

Affinity and avidity

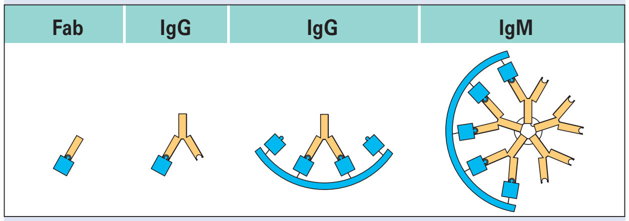
Fab
IgG
IgG
IgM

antibody	**Fab**	**IgG**	**IgG**	**IgM**
effective antibody valence	1	1	2	up to 10
antigen valence	1	1	n	n
equilibrium constant (L/mol)	10^4	10^4	10^7	10^{11}
advantage of multivalence	–	–	10^3-fold	10^7-fold
definition of binding	affinity	affinity	avidity	avidity
	intrinsic affinity		functional affinity	

Fig. 3.14 Multivalent binding between antibody and antigen (avidity or functional affinity) results in a considerable increase in stability as measured by the equilibrium constant, compared with simple monovalent binding (affinity or intrinsic affinity, here arbitrarily assigned a value of 10^4 L/mol). This is sometimes referred to as the 'bonus effect' of multivalency. Thus there may be a 10^3-fold increase in the binding energy of IgG when both valencies (combining sites) are used and a 10^7-fold increase when IgM binds antigen in a multivalent manner.

Specificity, cross-reactivity, and non-reactivity

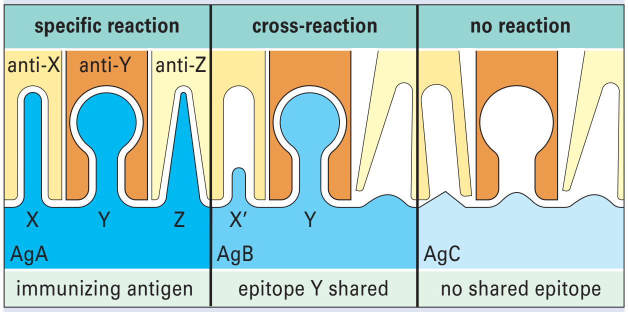
specific reaction
cross-reaction
no reaction
anti-X
anti-Y
anti-Z
X
Y
Z
AgA
X′
Y
AgB
AgC
immunizing antigen
epitope Y shared
no shared epitope

Fig. 3.15 Antiserum specificity results from the action of a population of individual antibody molecules (anti-X, anti-Y, anti-Z) directed against different epitopes (X, Y, Z) on the same or different antigen molecules. Antigen A (AgA) and antigen B (AgB) have epitope Y in common. Antiserum raised against AgA (anti-XYZ) not only reacts specifically with AgA, but cross-reacts with AgB (through recognition of epitopes Y and X′). The antiserum gives no reaction with AgC because there are no shared epitopes.

shape and chemical properties (e.g. charge, hydrophobicity) of epitopes. Small changes in the epitope, such as the position of a single chemical group, can therefore abolish binding (Fig. 3.16).

One consequence of this specificity is that many antibodies bind only to native antigens or to fragments of antigens that retain sufficient tertiary structure to permit the multiple interactions required for bond formation (Fig. 3.17).

Specificity and cross-reactivity

radical (R)	sulfonate	arsonate	carboxylate
	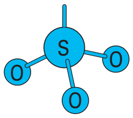 tetrahedral	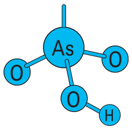 tetrahedral	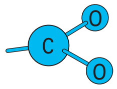 planar
ortho 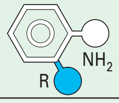	+ +	–	–
meta 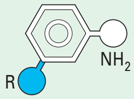	+ + +	+	±
para 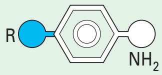	±	–	–

Fig. 3.16 Antiserum raised to the meta isomer of aminobenzene sulfonate (the immunizing hapten) is mixed with ortho and para isomers of aminobenzene sulfonate, and also with the three isomers (ortho, meta, para) of two different, but related, antigens: aminobenzene arsonate and aminobenzene carboxylate. The antiserum reacts specifically with the sulfonate group in the meta position, but will cross-react (although more weakly) with sulfonate in the ortho position. Further, weaker cross-reactions are possible when the antiserum is reacted with either the arsonate group or the carboxylate group in the meta, but not in the ortho or para, position. Arsonate is larger than sulfonate and has an extra hydrogen atom, while carboxylate is the smallest of the three groups. These data suggest that an antigen's configuration is as important as the individual chemical groupings that it contains.

Configurational specificity

antiserum	antigen		
	lysozyme	isolated 'loop' peptide	reduced 'loop'
	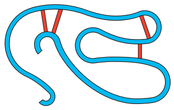	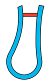	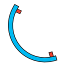
anti-lysozyme	+ +	+	–
anti-'loop' peptide	+	+ +	–

Fig. 3.17 The lysozyme molecule possesses an intrachain bond (red), which produces a loop in the peptide chain. Antisera raised against whole lysozyme (anti-lysozyme) and the isolated loop (anti-'loop' peptide) are able to distinguish between the two. Neither antiserum reacts with the isolated loop in its linear reduced form. This demonstrates the importance of tertiary structure in determining antibody specificity.

Q. Consider an antibody that recognizes an antigen on the outer envelope of a virus. If the antigen mutated, what would be the effect on its ability to bind antibody?

A. It depends on the nature of the mutation. Only mutations that affect the structure of the epitope could affect binding. However, if the mutation caused a radical change in the epitope structure (e.g. charge change or the insertion of a bulky amino acid residue), then the ability to bind might be lost or the affinity of binding drastically reduced.

Q. What effect would it have on the virus if a mutation of the epitope prevented it from being recognized by the antibody?

A. Viruses that mutate to avoid recognition by antibody will be at a selective advantage because they will not be destroyed by the immune system or prevented from attaching to host cells. Many viruses, including influenza and HIV, mutate their surface molecules and so evade recognition by antibodies.

Analysis of antibodies to protein antigens reveals that specificity may be for epitopes:

- consisting of a single contiguous stretch of amino acids (**a continuous epitope**);
- dependent on the native conformation of the antigen and formed from two or more stretches of sequence that are separated in the primary structure (**discontinuous or conformational epitopes**).

Q. How might these differing characteristics of antigen be relevant when producing antibodies for immunological assays?

A. Antibodies specific for discontinuous epitopes may not bind denatured antigen, for example on immunoblots (Method Box 3.3), whereas antibodies to continuous epitopes bind denatured antigen more often.

EACH ISOTYPE (SUBCLASS) MEDIATES A DISTINCT SET OF EFFECTOR FUNCTIONS

Antibodies are bifunctional because they must both:

- form complexes with antigens to limit their spread;
- elicit host responses to facilitate antigen removal and destruction.

The nature of the constant region determines the effector function of the antibody and the various host responses elicited, such as complement-mediated lysis or phagocytosis.

In antibody–antigen complexes the antibody molecules are essentially aggregated such that the multiple Fc regions are able to engage and activate ligands or

METHOD BOX 3.3

Assays to characterize antigens

The techniques of immunoprecipitation and immunoblotting (also known as western blotting) are both used to characterize antigens, giving information on their molecular mass and heterogeneity, and some limited information on their abundance.

These techniques may be useful for comparing the abundance of antigens, but are less well suited to the quantitation of antigens than the immunoassays described in Method Box 3.2.

Immunoblotting is a simple technique, but only antibodies that can recognize continuous epitopes can be used to detect the denatured antigens on the blots.

Immunoprecipitation is more often used where only antibodies that recognize undenatured antigens are available.

Immunoblotting

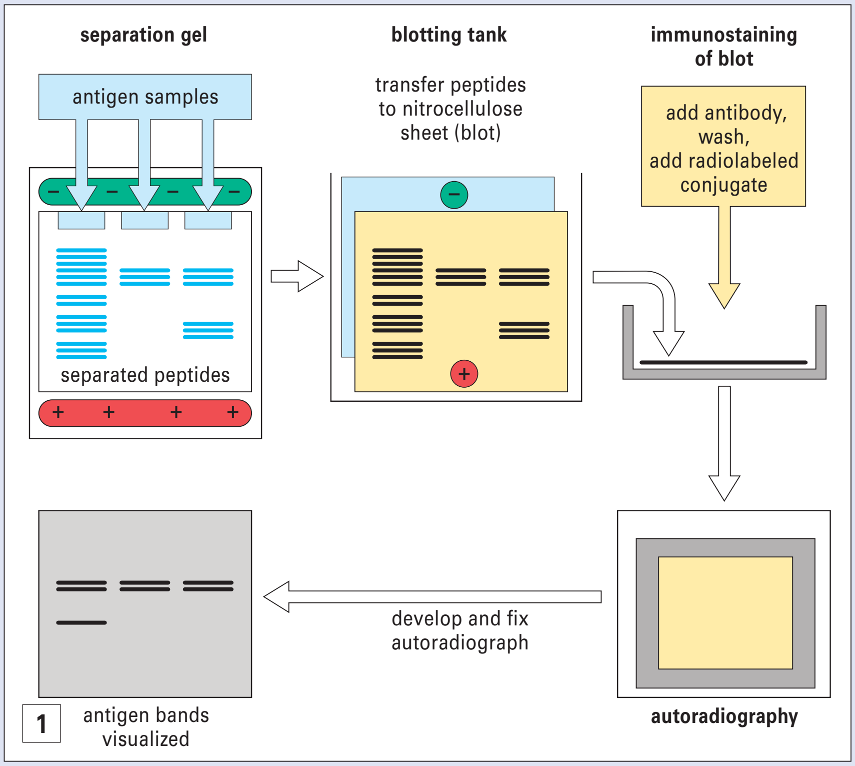

Fig. 1 In immunoblotting, antigen samples are first separated in an analytical gel (e.g. an SDS polyacrylamide gel or an isoelectric focusing gel). The resolved molecules are transferred electrophoretically to a nitrocellulose membrane in a blotting tank. The blot is then treated with antibody to the specific antigen and washed, and a radiolabeled conjugate to detect antibodies is bound to the blot. The principle is similar to that of a radioimmunoassay (RIA, see Method Box 3.2) or enzyme-linked immunosorbent assay (ELISA, see Method Box 3.2). After washing again, the blot is placed in contact with X-ray film in a cassette. The autoradiograph is developed and the antigen bands that have bound the antibody are visible. This immunoblotting technique can be modified for use with a chemiluminescent label or an enzyme-coupled conjugate (as in ELISA), where the bound material can be detected by treatment with a chromogen, which deposits an insoluble reagent directly onto the blot.

receptors (e.g. FcγR and C1q) (Fig. 3.18). Antibody is said to **opsonize** the antigen (bacterium, virus), making it a more attractive target for phagocytic cells.

IgM is the predominant antibody in the primary immune response

IgM is the first antibody formed in the primary immune response and is largely confined to the intravascular pool. It is frequently associated with the immune response to antigenically complex, blood-borne infectious organisms.

Once bound to its target, IgM is a potent activator of the classical pathway of complement.

Although the pentameric IgM antibody molecule consists of five Fcμ regions, it does not activate the classical pathway of complement in its uncomplexed form. However, when bound to an antigen with repeating identical epitopes, it forms a 'staple' structure (see Fig. 3.9), undergoing a conformational change referred to as **dislocation**, and the multiple Fcμ presented in this form are able to initiate the classical complement cascade.

IgG is the predominant antibody of secondary immune responses

IgG equilibrates between the intravascular and extravascular pools, so providing comprehensive systemic protection.

The four IgG subclasses are highly homologous in structure, but each has a unique profile of effector functions; thus, in activating the classical pathway of complement, complexes formed with:

- IgG1 and IgG3 antibodies are efficient;
- IgG2 antibodies are less effective;
- IgG4 antibodies are reported to be inactive.

METHOD BOX 3.3
Assays to characterize antigens *(Continued)*

Immunoprecipitation

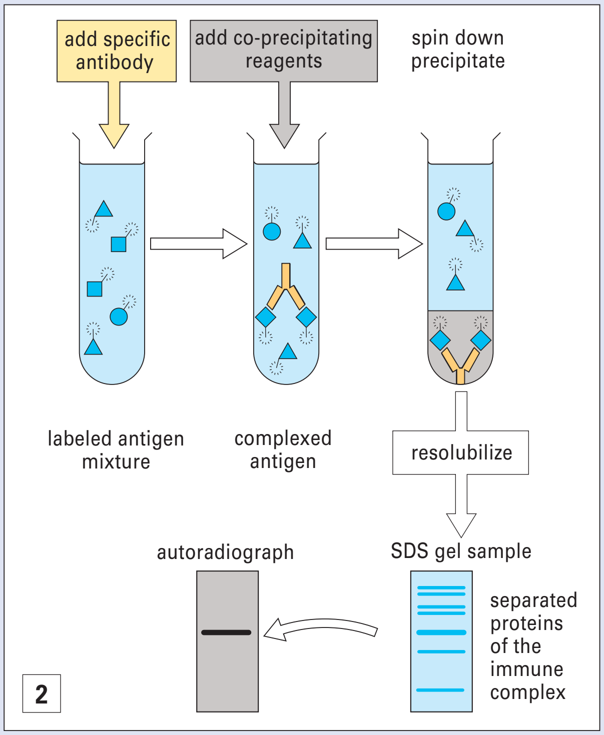

Fig. 2 In immunoprecipitation the antigens being tested are labeled with ^{125}I, and antibody is added that binds only to its specific antigen. The complexes are precipitated by the addition of co-precipitating agents, such as anti-immunoglobulin antibodies or staphylococcal protein A. The insoluble complexes are spun down and washed to remove any unbound labeled antigens. The precipitate is then resolubilized (e.g. in SDS), and the components are separated on analytical gels. After running, the fixed gels are autoradiographed, to show the position of the specific labeled antigen. Frequently the antigens are derived from the surface of radiolabeled cells, which are solubilized with detergents before the immunoprecipitation. It is also possible to label the antigens with biotin, and detect them at the end chromatographically using streptavidin (binds biotin) coupled to an enzyme such as peroxidase (see ELISA technique in Method Box 3.2).

The IgG subclasses also interact with a complex array of cellular Fc receptors (FcγR) expressed on various cell types (see Fig. 3.18 and below).

In humans, the newborn infant is not immunologically competent and the fetus must be protected by passive IgG antibody selectively transported across the placenta (Fig. 3.19). Transport is mediated by the neonatal Fc receptor (FcRn) – all IgG subclasses are transported. IgG subclasses are the only isotypes that are transported across the placenta, but the cord/maternal blood ratios differ, being approximately 1.2 for IgG1 and approximately 0.8 for IgG4.

Q. How do you interpret these ratios with respect to IgG transport across the placenta?

A. IgG1 is transported more efficiently than IgG4. Also, because the ratio for IgG1 is >1, the antibody is moving up a concentration gradient; therefore it must be active transport.

In some species (e.g. the pig), maternal immunoglobulin is transferred to the offspring in the postnatal period through selective transport of IgG across the gastrointestinal tract via a homologous FcRn receptor.

IgA is characteristic of a secondary immune response

Serum IgA is characteristic of a secondary immune response. Immune complexes of IgA opsonize antigens and activate phagocytosis through cellular Fc receptors (FcαR).

A predominant role for IgA antibody is in its secretory form, affording protection of the respiratory, gastrointestinal, and reproductive tracts.

The IgA1 subclass predominates in:

- human serum (approximately 90% of total IgA); and
- secretions such as nasal mucus, tears, saliva, and milk (70–95% of total IgA).

In the colon, IgA2 predominates (approximately 60% of the total IgA). Many microorganisms that can infect the upper respiratory and gastrointestinal tracts have adapted to their environment by releasing proteases that cleave IgA1 within the extended hinge region, whereas the short hinge region of IgA2 is not vulnerable to these enzymes.

IgD is a transmembrane antigen receptor on B cells

IgD has no known effector functions as a serum protein. It functions as a transmembrane antigen receptor on mature B lymphocytes.

IgE may have evolved to protect against helminth parasites

Despite its low serum concentration, the IgE class:

- is characterized by its ability to bind avidly to circulating basophils and tissue mast cells through the high-affinity FcεRI receptor (see below);
- sensitizes cells on mucosal surfaces such as the conjunctival, nasal, and bronchial mucosae.

IgE may have evolved to provide immunity against helminth parasites, but in developed countries it is now more commonly associated with allergic diseases such as asthma and hay fever.

Fc RECEPTORS ARE EXPRESSED BY MONONUCLEAR CELLS, NEUTROPHILS, NK CELLS, EOSINOPHILS, BASOPHILS, AND MAST CELLS

Antibodies sometimes protect just by binding to a pathogen, so preventing the pathogen from attaching to cells of the body and infecting them. More often, the biological and protective functions of antibodies are

Biological properties of human immunoglobulins

isotype	IgG1	IgG2	IgG3	IgG4	IgA1	IgA2	IgM	IgD	IgE
complement activation									
classical pathway	++	+	+++	−	−	−	+++	−	−
alternative pathway	varies with epitope density and antibody/antigen ratio								
lectin pathway	varies with glycosylation status								
Fc receptor binding									
FcγRI (monocytes)	+++	−	+++	++	−	−	−	−	−
FcγRIIa (monocytes, neutrophils, eosinophils, platelets)	+	±*	+	−	−	−	−	−	−
FcγRIIb (lymphocytes)	+	?	+	−					
FcγRIII (neutrophils, eosinophils, macrophages, LGLs, NK cells, T cells)	+	−	+	±	−	−	−	−	−
FcεRI (mast cells, basophils)	−	−	−	−	−	−	−	−	+++
FcεRII (monocytes, platelets, neutrophils, B and T cells, eosinophils)	−	−	−	−	−	−	−	−	++
FcαR (monocytes, neutrophils, eosinophils, T and B lymphocytes)	−	−	−	−	+	+	−	−	−
FcμR (T cells, macrophages)	−	−	−	−	−	−	+	−	−
FcδR (T and B cells)	−	−	−	−	+	−	−	+	−
pIgR, poly-Ig receptor; mucosal transport	−	−	−	−	+	+	+	−	−
FcRn, placental transport and catabolism	+	+	+	+	−	−	−	−	−
products of microorganisms									
SpA, staphylococcal protein A	+	+	−	+					
SpG, streptococcal protein G	+	+	+	+	−	−	−	−	−

*dependent on the allotype of FcγRIIa; LGLs, large granular lymphocytes; NK cells, natural killer cells

Fig. 3.18 The biological functions of different antibody classes and subclasses depend on which receptors they bind to and the cellular distribution of those receptors.

mediated by acting as adapters that bind via their Fc region to Fc receptors on different cell types.

The three classes of Fc receptor for IgG are FcγRI, FcγRII, and FcγRIII

Three classes of cell surface receptor for IgG (FcγR) are defined in humans:

- FcγRI (CD64);
- FcγRII (CD32); and
- FcγRIII (CD16).

Each receptor is characterized by a glycoprotein α chain that binds to antibody and has extracellular domains homologous with immunoglobulin domains (Fig. 3.20) – that is, they belong to the immunoglobulin superfamily, as do receptors for IgA (FcαR) and IgE (FcεRI).

FcγRs are expressed constitutively on a variety of cell types and may be upregulated or induced by environmental factors (e.g. cytokines).

Biological activation results from cross-linking of the FcγR and consequent aggregation of **immunoreceptor tyrosine-based activation (ITAM)** or **immunoreceptor tyrosine-based inhibitory (ITIM)** motifs in the cytoplasmic sequences.

Phosphorylation of the ITAM motif triggers activities such as:

- phagocytosis;
- **antibody-dependent cell-mediated cytotoxicity (ADCC)**;
- mediator release; and
- enhancement of antigen presentation.

Immunoglobulins in the serum of the fetus and newborn child

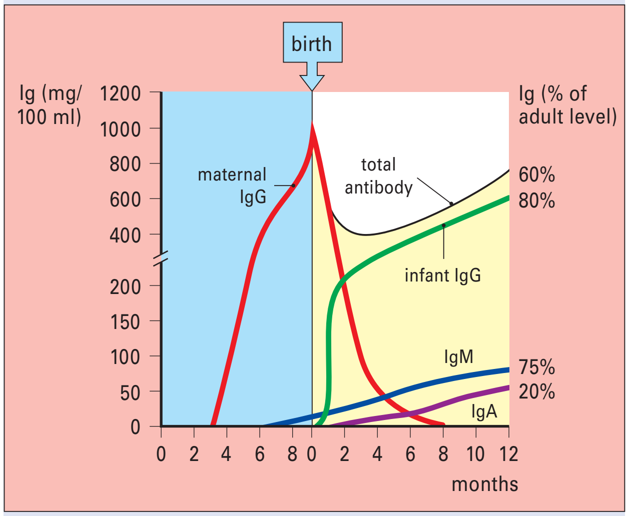

Fig. 3.19 IgG in the fetus and newborn infant is derived solely from the mother. This maternal IgG has disappeared by the age of 9 months, by which time the infant is synthesizing its own IgG. The neonate produces its own IgM and IgA; these classes cannot cross the placenta. By the age of 12 months, the infant produces 80% of its adult level of IgG, 75% of its adult IgM level, and 20% of its adult IgA level.

Fcγ receptors

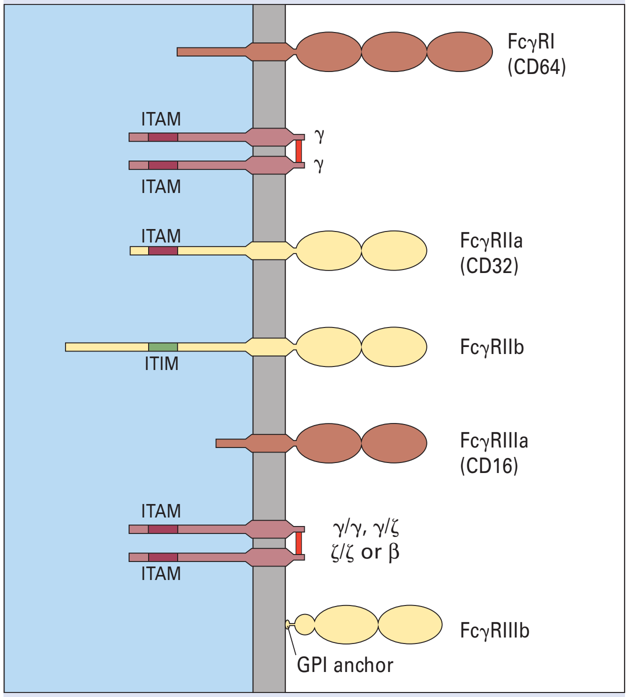

Fig. 3.20 Receptors for Fcγ in humans belong to the immunoglobulin superfamily, and have either two or three extracellular immunoglobulin domains. Motifs (ITAM, ITIM) on the intracellular segments or on associated polypeptides are targets for tyrosine kinases involved in initiating intracellular signaling pathways.

In contrast, phosphorylation of ITIM blocks cellular activation.

FcγRI is involved in phagocytosis of immune complexes and mediator release

FcγRI (CD64; 72 kDa) binds:

- monomeric IgG1 and IgG3 with high affinity (10^8–10^9 l/mol); and
- IgG4 with lower affinity (10^7–10^8 l/mol).

It has a more restricted cellular distribution than the other FcγRs, but is expressed on all cells of the mononuclear phagocyte lineage, and is involved in the phagocytosis of immune complexes, mediator release, etc.

The α chain contains three immunoglobulin domains in the extracellular portion and is associated with a γ chain that bears an ITAM motif in its cytoplasmic part.

FcγRII is expressed as FcγRIIa and FcγRIIb

FcγRII (CD32; 40 kDa) is expressed as structurally and functionally distinct FcγRIIa and FcγRIIb forms with wide but differing cellular distribution.

The α chains have:

- low affinity (< 10^7 l/mol) for monomeric IgG1 and IgG3;
- bind complexed (multivalent, aggregated) IgG with high avidity.

The FcγRIIa molecule:

- binds IgG1 and IgG3 only; and
- expresses an ITAM motif within its cytoplasmic tail.

It may also be associated with a γ chain and cross-linking results in cellular activation.

Polymorphism in the FcγRIIA gene results in the presence of histidine or arginine at position 131 in the extracellular domains – the His^{131} allotype binds and is activated by immune complexes of IgG2.

The FcγRIIb molecule:

- expresses an ITIM motif within its cytoplasmic tail; and
- when cross-linked, blocks cellular activation, particularly on B cells (see Fig. 11.5).

FcγRIII is expressed as FcγRIIIa and FcγRIIIb

FcγRIII (CD16; 50–80 kDa) is expressed as structurally and functionally distinct FcγRIIIa and FcγRIIIb, which are extensively glycosylated and have differing cellular distributions.

FcγRIIIa is a transmembrane protein (like FcγRI, FcγRIIa, and FcγRIIb), whereas FcγRIIIb is GPI (glycosyl phosphatidyl inositol) anchored (see Fig. 3.20).

The α chains of FcγRIIIa:

- have a moderate affinity for monomeric IgG (approximately 3×10^7 l/mol); and
- may be associated with γ/ξ and/or β chains bearing ITAM motifs.

FcγRIIIa is expressed on monocytes, macrophages, NK cells, and some T cells.

The FcγRIIIb form:

- is selectively expressed on neutrophils; and
- has a low affinity for monomeric IgG (<10^7 l/mol).

It appears that engagement of FcγRIIIb can result in transmembrane signaling through its association with other membrane proteins bearing signaling motifs.

POLYMORPHISM IN THE FcγRIIIA AND FcγRIIIB GENES: FcγRIIIB-NA1 AND FcγRIIIB-NA2 FORMS – Polymorphism in the *FcγRIIIA* gene results in the presence of phenylalanine (Phe) or valine (Val) at position 158 in the extracellular domains; the Val^{158} allotype is associated with more efficient NK cell activity.

Polymorphism in the *FcγRIIIB* gene results in multiple amino acid sequence differences, which affect the extent of glycosylation.

The consequent FcγRIIIb-NA1 and FcγRIIIb-NA2 forms are reported to differ in functional activity.

Q. What factors determine whether a particular IgG subclass will have a particular biological function (e.g. the ability to opsonize a bacterium for phagocytosis by a macrophage)?

A. The distribution of Fc receptors between different effector cells, the ability of each subclass of antibody to bind to these Fc receptors, and the affinity of binding.

IgG Fc interaction sites for several ligands have been identified

Application of site-directed mutagenesis, X-ray crystallography and nuclear magnetic resonance spectroscopy has allowed elucidation of the molecular topography of IgG Fc interaction sites for several ligands that bind overlapping non-identical sites at the CH2/CH3 interface, for example:

- staphylococcal protein A;
- streptococcal protein G;
- FcRn; and
- a monoclonal rheumatoid factor Fab fragment.

The interactions between maternal IgG and the MHC class I molecule-like FcRn expressed on the intestinal epithelium of the neonatal rat have now been studied at high resolution (Fig. 3.21) and are believed to mimic closely the binding of the human placental counterpart, hFcRn, with maternal IgG. Titration of IgG histidine residues in the binding site for FcRn may explain its:

- pH sensitivity;
- binding at pH 6.5 (the pH within vacuoles); and
- dissociating at pH 7.4 (the pH of blood).

Given the symmetry of the Fc region, the fragments used in the experiments above are functionally divalent and may form multimeric complexes.

If monomeric IgG were divalent for FcγR and C1q, however, it would not function properly because circulating monomeric IgG could form activating multimers. The interaction sites for these ligands (e.g. FcγR and C1q) have been 'mapped' to the CH2 domain next to the hinge.

The crystal structure of an IgG Fc/FcγRIII complex reveals an asymmetric interaction site embracing the CH2 domains of both heavy chains, thereby ensuring monovalency.

Site-directed mutagenesis indicates that common amino acid residues engage each of FcγRI, FcγRII, and FcγRIII, but care should be taken in extrapolating from isotype to isotype and from species to species.

For example, Glu318, Lys320, and Lys322 of the CH2 domain is widely quoted as the interaction motif for C1q activation, but this result was obtained for mouse IgG2b with guinea pig complement. This finding does not apply to the activation of human complement by human IgG1 – even though human IgG1 expresses the same Glu-x-Lys-x-Lys motif.

Neonatal rat intestinal Fc receptor FcRn

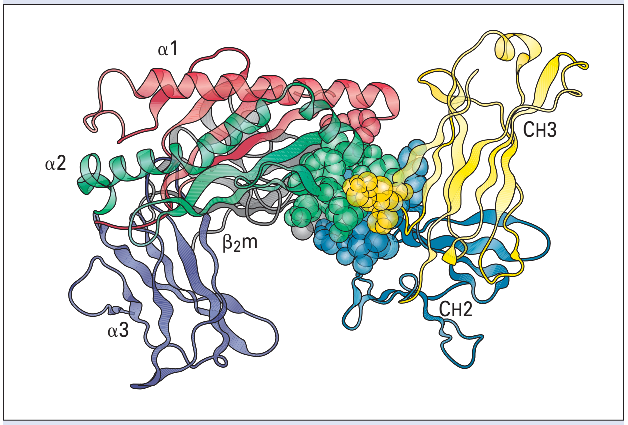

Fig. 3.21 Principal interactions between neonatal rat intestinal FcRn and the Fc of maternal IgG (derived from milk) are illustrated by ribbon diagrams of FcRn (domains α1, α2, α3, and β$_2$m are shown in red, light green, purple, and gray, respectively) and of Fc (CH2 and CH3 domains are shown in blue and yellow). The main contact residues of the FcRn (α1 domain, 90; α2, 113–119 and 131–135; β$_2$m, 1–4 and 86) are depicted as space-filling structures. (Reproduced from Ravetch JV, Margulies DH. New tricks for old molecules. Nature 1994;372:323–324)

IgM–antigen complexes are very efficient activators of the classical complement system, but the mechanism by which IgM binds C1q appears different from that of IgG. The conformational change from a 'star' to a 'staple' conformation upon binding to multivalent antigen is thought to unveil a ring of occult C1q-binding sites that are not accessible in the star-shaped configuration (see Fig. 3.9).

Glycosylation is important for receptor binding to IgG

N-linked glycosylation of the IgG Cγ2 domain is essential for the binding and activation of FcγRI, FcγRII, FcγRIII, and C1q. The oligosaccharide is of a complex type, generating multiple glycoforms, and forms multiple non-covalent interactions with the polypeptide such that it appears to be sequestered within the protein structure and so inaccessible for direct interactions with the FcγRI, FcγRII, FcγRIII, and C1q ligands.

Fidelity of glycosylation has become an important issue in the production of monoclonal antibodies for therapy because glycosylation is a species- and tissue-specific post-translational modification (see Method Box 3.1).

Chinese hamster ovary (CHO) cells and mouse myeloma cells can produce appropriately glycosylated human IgG heavy chains, though the range of glycoforms is restricted. Recent studies have shown that the glycoforms thus produced do not express optimal effector functions (such as killing of tumor cells by ADCC; see

Chapter 10). These production cell lines, and other possible production platforms, are being engineered to produce optimal human glycoforms by transfecting in human glycosyltransferases and 'knocking-out' non-human glycosyltransferases.

The two classes of Fc receptor for IgE are FcεRI and FcεRII

Two classes of Fc receptor for IgE (FcεR) are defined in humans (Fig. 3.22):

- the high-affinity FcεRI (45 kDa), which is expressed on mast cells and basophils and is the 'classical' IgE receptor; and
- the low-affinity FcεRII (CD23; 45 kDa), which is expressed on leukocytes and lymphocytes.

The α chain of FcεRI is a glycoprotein and has two extracellular domains homologous to immunoglobulin domains; it is therefore a member of the immunoglobulin superfamily.

The low-affinity FcεRII is not a member of the immunoglobulin superfamily, but has substantial structural homology with several animal C-type lectins (e.g. mannose-binding lectin [MBL]).

Cross-linking of IgE bound to FcεRI results in histamine release

The high-affinity receptor (FcεRI) is present on the surface of mast cells and basophils as a complex with a β (33 kDa) and two γ (99 kDa) chains to form the $\alpha\beta\gamma_2$ receptor unit (see Fig. 3.22).

FcεRI binds IgE with an affinity of approximately 10^{10} l/mol such that, although the serum concentration of IgE is very low, the receptors are permanently saturated. Cross-linking of the IgE bound to these receptors results in the activation and release of histamine and other vasoactive and inflammatory mediators.

Models for FcεRI and FcεRII

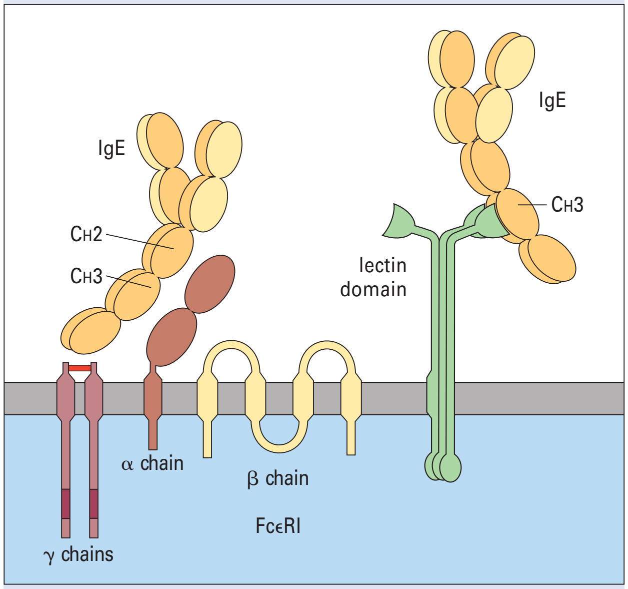

Fig. 3.22 Models for the high-affinity IgE receptor (FcεRI), which binds to IgE via its α chain, and the low-affinity receptor (FcεRII), which binds using its lectin domains. Both receptors are shown with IgE bound.

FcεRII is a type 2 transmembrane molecule

FcεRII is the low-affinity (CD23) receptor and is a type 2 transmembrane molecule (i.e. one in which the C termini of the polypeptides are extracellular, see Fig. 3.22). The two forms of human CD23 are:

- **CD23a**, which is expressed in antigen-activated B cells and influences IgE production; and
- **CD23b**, expression of which is induced in a wide range of cells by IL-4.

CD23a and CD23b differ by six or seven amino acids in their cytoplasmic N termini and contain different signaling motifs that modify their functions.

IgE receptors bind to IgE by different mechanisms

Decades of research and controversy surround the identification of the interaction site on IgE for the high-affinity FcεRI receptor.

Recent crystal structures of a Cε2Cε3Cε4 fragment and Cε3Cε4–FcεRI complex seem to have resolved the issue, and there is a striking structural homology between this site and the site on IgG Fc that binds FcγRIII.

One distinctive feature is that, unlike IgG binding to FcγRIII, binding of IgE to its receptor does not appear to depend on glycosylation of the immunoglobulin.

These studies should be interpreted with caution because the antibody fragments were not produced in mammalian cells.

Possible models for the interaction of IgE with the FcεRI and FcεRII receptors are illustrated in Fig. 3.22. The low-affinity IgE Fc receptor, FcεRII (CD23), is a C-type lectin and is therefore sensitive to glycosylation status.

A VAST REPERTOIRE OF ANTIGEN-BINDING SITES IS ACHIEVED BY RANDOM RECOMBINATION

How can a finite genome provide the information required for the vast repertoire of antibody molecules that an individual can synthesize? This question posed an intellectual and practical challenge for decades.

Ehrlich's side-chain hypothesis, put forward at the beginning of the 20th century (Fig. 3.23), proposed antigen-induced selection. His model is close to our present view of clonal selection, except that he placed receptors of several different specificities on the same cell and did not address the question of how the diverse receptors were generated.

Several decades later, Dreyer and Bennett realized that the existence of a constant region and a variable region implied that the antibody protein was encoded by at least two different genes, contrary to a central dogma of molecular biology ('one gene, one protein').

Shortly thereafter, Tonegawa performed the decisive experiments that proved the Dreyer and Bennett hypothesis and showed that somatic recombination events were able to take a limited amount of genetic material and create almost innumerable permutations, encoding a vast antigen receptor repertoire.

Ehrlich's side-chain theory

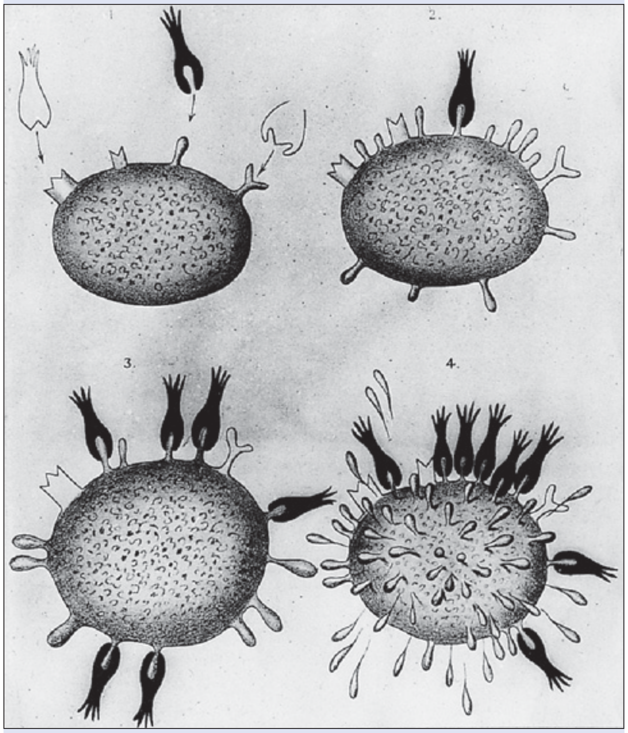

Fig. 3.23 Ehrlich proposed that the combination of antigen with a preformed B cell receptor (now known to be antibody) triggered the cell to produce and secrete more of those receptors. Although the diagram indicates that he thought a single cell could produce antibodies to bind more than one type of antigen, it is evident that he anticipated both the clonal selection theory and the idea that the immune system could generate receptors before contact with antigen.

Q. Given that one human can produce more heavy chains and more light chains than there are genes in the entire human genome, what mechanism(s) could lead to the presence of the great diversity of immunoglobulin genes seen in different B cells, each encoding a specific antibody?

A. The antibody genes could be produced by splicing together different gene segments, with a different combination of segments used in different B cells.

Another mechanism could involve mutation of the original gene(s) with different mutations occurring in different B cells. In fact, both of these mechanisms take place in human B cells.

The two mechanisms occurring within individual B cells – gene segment recombination and gene mutation – described above, are termed **somatic recombination** and **somatic mutation** to distinguish them from the related processes that occur in germ cells.

In addition to somatic recombination and somatic mutation, which produce diverse heavy and light chains, the pairing of a unique VL domain with a unique VH domain to generate an antigen-binding site also generates diversity and became an established paradigm.

Recently, however, it has been shown that functional heavy (H chain) antibodies devoid of light chains account for about half of the circulating immunoglobulins in the Camelidae family. This suggests new possibilities for using just VH domains as therapeutics.

Immunoglobulins show isotypic, allotypic, and idiotypic variation

Originally the immunoglobulin isotypes were defined serologically by the antigenic uniqueness of their constituent polypeptide chains, using antisera raised predominantly in rabbits. Allotypes were recognized using human sera from individuals that lacked the allotype.

Idiotypy (variation related to the specific structure of individual immunoglobulin domains and the antigen-binding site) was revealed with antisera raised to monoclonal IgG immunoglobulins isolated from the sera of patients with a B cell tumor. When adsorbed with polyclonal IgG, such an antiserum produced a reagent specific for the immunizing protein (the **idiotype**).

In the modern era, we identify the human immunoglobulin **isotypes** as products of defined immunoglobulin genes encoding the constant regions of heavy and light chains, and the **allotypes** as polymorphic variants of these genes.

The idiotype of an antibody molecule results from antigenic uniqueness reflecting the structural uniqueness of the VH and VL regions (Fig. 3.24).

Isotypes are the products of genes within the genome of all healthy members of a species

The human immunoglobulin heavy chain isotypes are encoded by μ, δ, γ3, γ1, α1, γ2, γ4, α2, and ε genes on chromosome 14. The single κ and multiple λ isotypes are encoded by Cκ and Cλ genes on chromosomes 2 and 22, respectively.

Variability of immunoglobulin structure

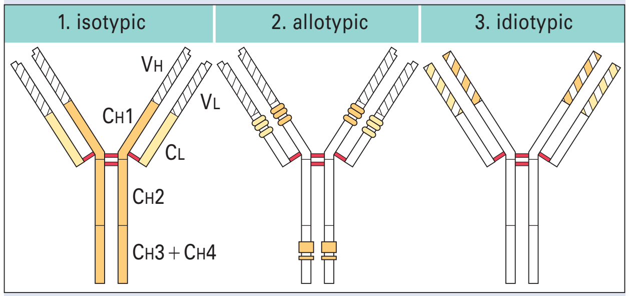

Fig. 3.24 All immunoglobulins have the basic four-chain structure. There are three types of immunoglobulin variability. (**1**) Isotypic variation is present in the germline of all members of a species, producing the heavy (μ, δ, γ, ε, α) and light chains (Igκ, λ) and the V region frameworks (subgroups). (**2**) Allotypic variation is intraspecies allelic variability. (**3**) Idiotypic variation refers to the diversity at the antigen-binding site (paratope) and in particular relates to the hypervariable segments.

Allotypes result from genetic variation at a locus within the species

Immunoglobulin allotypes refer to proteins that are the product of polymorphic immunoglobulin genes (i.e. the result of genetic variation at a given locus within the species). The WHO-adopted nomenclature – for example G3m(g) – requires:

- definition of the heavy chain class;
- followed by the subclass;
- an m (indicating 'marker'); and
- the designated number or letter in brackets.

The sequence correlates have been determined. For example G1m(f) and G1m(z) allotypes are characterized by the presence of, respectively, an arginine or lysine residue at position 214 in the CH1 or Cγ1 domain of the heavy chains.

In humans allotypes are encountered within:

- IgG;
- IgA; and
- κ light chains.

Because all heavy chain genes are encoded within the same locus, they are inherited as a complex or **haplotype**. Studies of immunoglobulin haplotypes help in the identification of different population groups of humans.

Idiotypes result from antigenic uniqueness

The structural uniqueness of antibody variable regions can be reflected in antigenic uniqueness recognized by antisera raised in a heterologous species (e.g. rabbit).

However, in addition to antigenic uniqueness, cross-reactivity may be observed for two V regions that are highly homologous. The terms private and public (or cross-reactive and recurrent) idiotypes are used to describe this property. Idiotypes:

- may be specific for individual B cell or plasma cell clones (**private idiotypes**);
- are sometimes shared between different B cell clones (**public, cross-reacting, or recurrent idiotypes**).

Each therapeutic antibody is structurally unique and has the potential to provoke an immune response in recipients. This is a current concern for this very promising class of drugs.

Monoclonal antibodies have been developed as successful therapeutics, but their unique specificity is reflected in unique structure (idiotype) such that they are potentially immunogenic. In practice a variable proportion of patients produce **anti-idiotypic antibodies**, depending on the patient's genetic background, the disease treated, the dosing regimen, etc.

HEAVY CHAIN GENE RECOMBINATION PRECEDES LIGHT CHAIN RECOMBINATION

The revealed organization of mammalian genomes invalidated the earlier dogma of 'one gene, one polypeptide sequence' and replaced it with 'genes in pieces' as coding (exons) and non-coding (introns) DNA sequences were identified. Generally:

- an open reading frame is transcribed to nuclear RNA;
- the introns are 'spliced' out of nuclear RNA to yield a continuous messenger RNA (mRNA);
- protein is translated from continuous mRNA.

Germline DNA encoding immunoglobulin polypeptide chains shows a further level of complexity. Information for the variable domains is present in two or three libraries of gene segments termed **V, D, and J gene segments** that recombine at the level of the DNA. These gene segments are widely separated from exons encoding the constant regions.

Recombination between the V, D, and J gene segments generates a continuous sequence encoding the V domain. However, in the initial nuclear RNA transcript the information for the V and C domains is still widely separated.

The DNA encoding the leader to the end of the C gene, including introns, is transcribed into heterogeneous nuclear RNA (hnRNA).

The hnRNA is processed with the 'splicing out' of introns to yield the mRNA encoding the polypeptide V and C domains within a continuous RNA sequence that is translated into protein.

The primary antibody repertoire is:

- present in the B cell population and expressed as membrane-bound IgM and IgD acting as antigen receptors;
- generated through recombination of germline gene segments, with selection to eliminate B cells producing anti-self antibodies.

Immunoglobulin genes undergo rearrangement during B cell development

Immunoglobulin gene rearrangement is initiated at the heavy chain locus. If a 'productive' rearrangement occurs (i.e. generating a functional product), then rearrangement at the κ locus is initiated:

- if the κ rearrangement is productive, the assembled antibody is expressed as a membrane receptor;
- if recombination at the κ locus is not productive, rearrangement at the λ locus proceeds.

A number of **pseudogenes** (non-functional genes with high homology to the heavy chain genes) and **orphan genes** (putative genes with no known homology to known functional genes) have been identified as related to functional heavy and light chain genes.

A number of genes present within the VH, VK, and VL loci do not have 'open reading frames' and cannot be transcribed. They do not, therefore, contribute to the generation of antibody diversity.

Pseudogenes account for a large part of what is known as 'junk DNA' in eukaryotic genomes, and are thought to be derived by mutations in functional ancestral genes.

Rearrangement at the heavy chain VH locus precedes rearrangement at light chain loci

The germline human heavy chain locus, on chromosome 14 (Fig. 3.25), contains a library of 38 to 46 functional VH gene segments that encode the N terminal 95 residues of the VH region.

The C terminal residues of the VH region are encoded within 23 DH and six JH gene segments (see Fig. 3.25).

Productive rearrangement at the VH locus is an obligatory and early event in the generation of B cells that precedes rearrangement at light chain loci.

Heavy chain VDJ recombination in humans

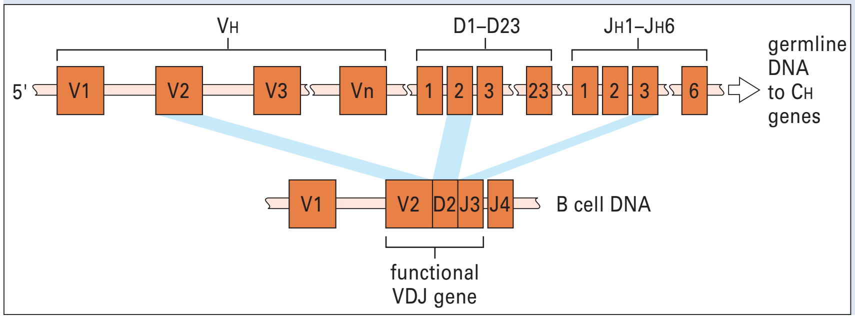

Fig. 3.25 The heavy chain gene loci recombine three segments to produce a VDJ gene, which encodes the VH domain. Of some 80 V genes, about 50 are functional and the others are pseudogenes. The V gene segment recombines with one of 23 DH segments and one of six JH segments to produce a functional VDJ gene in the B cell.

The first event is recombination between a JH gene segment and DH segments

DH segments:

- are highly variable, both in the number of codons and in the nucleotide sequence;
- may be read in three possible reading frames without generating stop codons; and
- can be used singly or in combinations.

Productive recombination between DH and JH gene segments signals recombination of this DJ sequence to a VH gene segment, forming a contiguous DNA sequence encoding the entire VH protein sequence (see Fig. 3.25).

The recombined VH, DH, and JH gene segments generate widely diverse hypervariable Hv3 (CDR3) sequences, which contribute greatly to the diversity of the primary antibody repertoire.

The heavy chain locus includes approximately 78 related pseudogenes, and orphan genes occur on chromosomes other than 14.

Rearrangement results in a Vκ gene segment becoming contiguous with a Jκ gene segment

The germline human κ light chain locus on chromosome 2 (Fig. 3.26) contains a library of 31 to 35 functional Vκ gene segments that encode the N terminal 95 residues of the Vκ region.

The C terminal residues of the Vκ region are encoded within five Jκ gene segments (see Fig. 3.26).

During B cell development, rearrangement of the DNA occurs such that one of the Vκ genes becomes contiguous with one of five Jκ genes.

Q. If there are 31 Vκ segments and five Jκ segments, what is the theoretical maximum number of rearrangements that could occur at this locus?

A. 155 (i.e. 31 × 5), but imprecise joining introduces additional diversity (see below).

The κ locus also includes over 30 related pseudogenes, and orphan genes are present on other chromosomes.

A leader or signal sequence (a short hydrophobic segment responsible for targeting the chain to the endoplasmic reticulum) precedes each Vκ segment. The leader sequence is cleaved in the endoplasmic reticulum, and the antibody molecule is then processed through the intracellular secretory pathway.

Recombination results in a Vλ gene segment becoming contiguous with a functional Jλ gene segment

The germline human λ light chain locus (Fig. 3.27) on chromosome 22 contains a library of 29 to 33 functional

κ chain production in humans

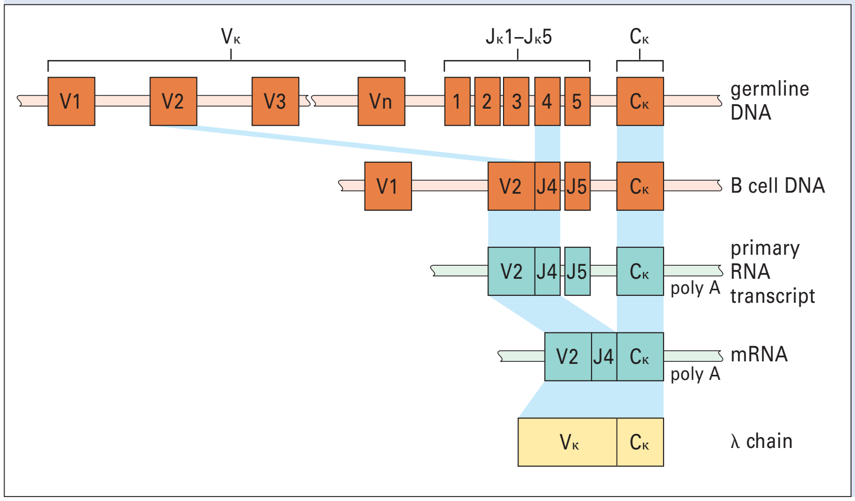

Fig. 3.26 During differentiation of the pre-B cell, one of several Vκ genes on the germline DNA (V1–Vn) is recombined and apposed to a Jκ segment (Jκ1–Jκ5). The B cell transcribes a segment of DNA into a primary RNA transcript that contains a long intervening sequence of additional J segments and introns. This transcript is processed into mRNA by splicing the exons together, and is translated by ribosomes into κ chains. The rearrangement illustrated is only one of the many possible recombinations. (B cell DNA is colored light brown; RNA is colored green; and immunoglobulin peptides are colored yellow.)

λ chain production in humans

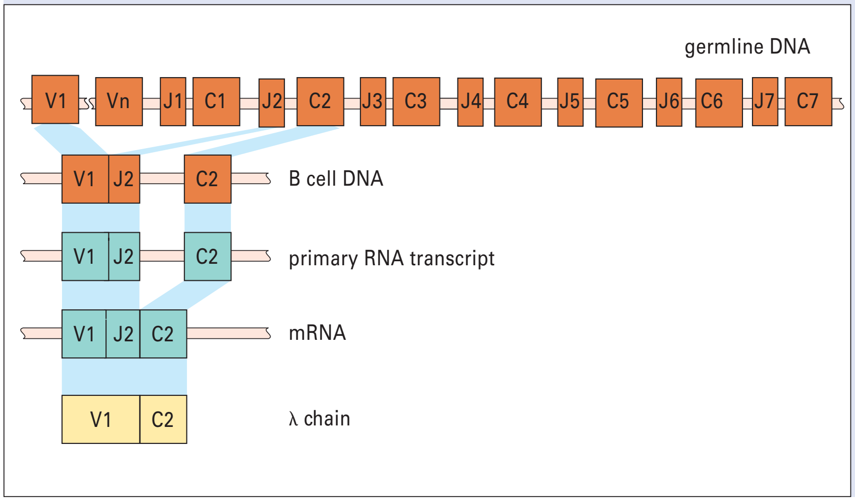

Fig. 3.27 During B cell differentiation, one of the germline Vλ genes recombines with a J segment to form a VJ combination. The rearranged gene is transcribed into a primary RNA transcript complete with introns (non-coding segments occurring between the genes), exons (which code for protein), and a poly A tail. This is spliced to form mRNA with loss of the introns, and then translated into protein.

Vλ gene segments that encode the N terminal 95 residues of the Vλ region.

There are 7 to 11 Jλ gene segments with each linked to a Cλ gene sequence (see Fig. 3.27) – the number of JλCλ sequences depends on the haplotype.

During the generation of B cells, unproductive rearrangement at the κ locus leads to recombination at the λ locus such that one of the Vλ genes becomes contiguous with one of four or five functional Jλ genes.

The number of possible λ chain variable regions that could be produced in this way is approximately 120 to 160. Imprecise joining introduces additional diversity (see below).

The λ locus also includes over 35 related pseudogenes, and orphan genes are present on other chromosomes.

Following recombination between Vλ and Jλ genes, there is still an intron (a non-coding intervening sequence) between the recombined VλJλ gene and the exon encoding the C region.

Recombination involves recognition of signal sequences by the V(DJ) recombinase

Recombination of germline gene segments is a key feature in the generation of the primary antibody repertoire. How is the recombination effected?

Each V, D, and J segment is flanked by **recombination signal sequences (RSS)**:

- a signal sequence found downstream (3′) of VH, VL, and DH gene segments consists of a heptamer CACAGTG or its analog, followed by a spacer of non-conserved sequence (12 or 23 bases), and then a nonamer ACAAAAACC or its analog (Fig. 3.28);
- immediately upstream (5′) of a germline JL, DH, and JH segment is a corresponding signal sequence of a nonamer and a heptamer, again separated by an unconserved sequence (12 or 23 bases).

The heptameric and nonameric sequences following a VL, VH, or DH segment are complementary to those preceding the JL, DH, or JH segments (respectively) with which they recombine.

Recombination sequences in immunoglobulin genes

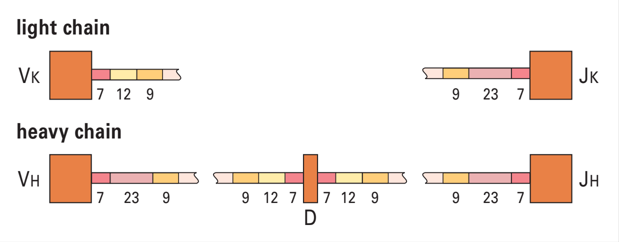

Fig. 3.28 The recombination sequences in the light chain genes (top) and heavy chain genes (bottom) consist of heptamers (7), 12 or 23 unconserved bases, and nonamers (9). The sequences of heptamers and nonamers are complementary and the nonamers act as signals for the recombination activating genes to form a synapsis between the adjoining exons. Similar recombination sequences are present in the T cell receptor V, D, and J gene segments (see Chapter 5).

The 12 and 23 base spacers correspond to either one or two turns of the DNA helix (see Fig. 3.28).

The recombination process is mediated by the protein products of the two **recombination-activating genes** (**RAG-1** and **RAG-2**):

- a RAG-1–RAG-2 complex recognizes the RSS, bringing a 12-RSS and a 23-RSS together into a synaptic complex (Fig. 3.29);
- the RAG proteins initiate cleavage by introducing a nick in the area bordering the 5′ end of the signal heptamer and the coding region;
- the RAG proteins then convert this nick into a double-strand break, generating a hairpin at the coding end and a blunt cut at the signal sequence;
- the hairpinned coding end must be opened before the joining step, and usually undergoes further processing (the addition or deletion of nucleotides), resulting in an imprecise junction.

Stages of V(D)J recombination

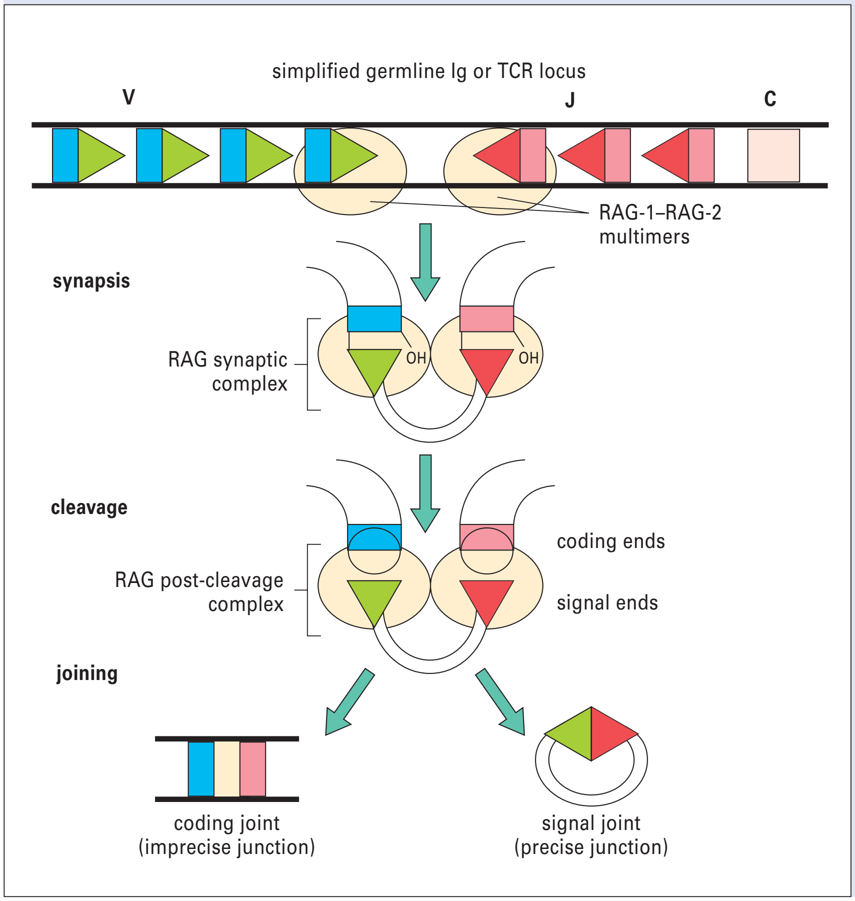
simplified germline Ig or TCR locus
V
J
C
RAG-1–RAG-2 multimers
synapsis
RAG synaptic complex
OH
OH
cleavage
coding ends
RAG post-cleavage complex
signal ends
joining
coding joint (imprecise junction)
signal joint (precise junction)

Fig. 3.29 The recombination-activating genes (RAG-1–RAG-2) bind to two recombination signal sequences (RSS) – one 12-RSS and one 23-RSS – bringing them into a synaptic complex. Synapsis stimulates cleavage: first, one DNA strand is nicked, and the resulting 3′ OH group attacks the opposite strand, leaving two hairpinned coding ends and two blunt signal ends. After cleavage, these ends are held in the RAG post-cleavage complex and joined by the non-homologous end-joining (NHEJ) factors. The coding joint forms the new variable region exon; the signal ends are joined together to form a signal joint, often (but not always) as an excised circular product that is lost from the cell. The signal joint has no known immunological function.

Signal ends, in contrast, are usually joined precisely to form circular signal joints that have no known immunological function and are lost from the cell (see Fig. 3.29)

The loss or addition of nucleotides during coding joint formation creates additional diversity in the junctions that is not encoded by the V, D, or J segments.

Furthermore, following cleavage, the enzyme terminal deoxynucleotidyl transferase may add random nucleotides to the exposed cut ends of the DNA. Nucleotides may therefore be inserted between DH and JH, and between VH and DH, without need of a template (Figs 3.30 and 3.31).

VARIABLE REGION GENES UNDERGO SOMATIC HYPERMUTATION

In the mid-20th century, following the elucidation of the structure of DNA, it was believed that the function of tissues, the cells that constitute them, and therefore the integrity of the individual depended on accurate replication of DNA sequences. The enormous size and sequence variability of the antigen-specific antibody repertoire, however, made the ‘one gene, one polypeptide chain’ dogma untenable.

Light chain diversity created by variable recombination

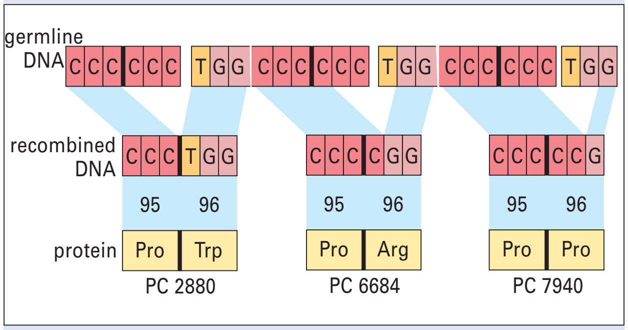
germline DNA
C C C C C C
T G G
C C C C C C
T G G
C C C C C C
T G G
recombined DNA
C C C T G G
C C C C G G
C C C C C G
95
96
95
96
95
96
protein
Pro
Trp
Pro
Arg
Pro
Pro
PC 2880
PC 6684
PC 7940

Fig. 3.30 The same Vκ21 and J1 sequences of the germline create three different amino acid sequences in the proteins PC2880, PC6684, and PC7940 by variable recombination. PC2880 has proline and tryptophan at positions 95 and 96, caused by recombination at the end of the CCC codon. Recombination one base down produces proline and arginine in PC6684. Recombination two bases down from the end of Vκ21 produces proline and proline in PC7940.

Heavy chain diversity created by variable recombination and N region diversity

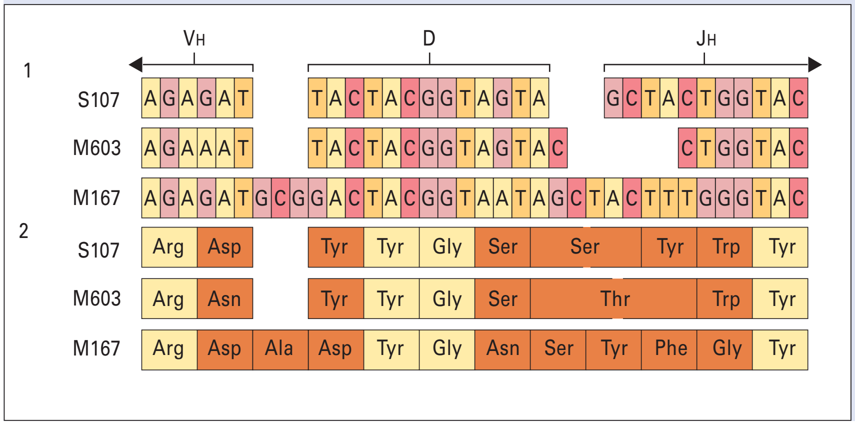
VH
D
JH
1
S107
AGAGAT
TACTACGGTAGTA
GCTACTGGTAC
M603
AGAAAT
TACTACGGTAGTAC
CTGGTAC
M167
AGAGATGCGGACTACGGTAATAGCTACTTTGGGTAC
2
S107
Arg Asp
Tyr Tyr Gly Ser Ser Tyr Trp Tyr
M603
Arg Asn
Tyr Tyr Gly Ser Thr Trp Tyr
M167
Arg Asp Ala Asp Tyr Gly Asn Ser Tyr Phe Gly Tyr

Fig. 3.31 The DNA sequence (**1**) and amino acid sequence (**2**) of three heavy chains of anti-phosphorylcholine are shown. Variable recombination between the germline, V, D, and J regions, and N region insertion causes variation (orange) in amino acid sequences. In some cases (e.g. M167) there appear to be additional inserted codons. However, these are in multiples of three, and do not alter the overall reading frame.

A period of intense intellectual and experimental activity followed as seemingly opposing germline and somatic mutation models were explored. Introduction of a theory of somatic mutation for immunoglobulin genes to account for a structural diversity that could not be accommodated within the genome led to a schism between immunologists and classical geneticists.

In due time, the argument proved unnecessary because it became evident that nature actually employs both types of process to form the antibody repertoire.

Libraries of germline gene segment sequences recombine, with a degree of junctional diversity, to generate the primary antibody repertoire, but this process does not account for the sequence diversity observed for antibodies that are generated during a secondary immune response.

It is now established that in germinal centers the recombined DNA encoding variable light and heavy chain sequences undergoes repeated rounds of random mutation (**somatic hypermutation**) to generate B cells expressing structurally distinct receptors (Fig. 3.32). Survival and expansion of these B cells requires that their receptors engage antigen presented to them by follicular dendritic cells in the lymphoid tissues:

- a majority will acquire a deleterious mutation and, in the absence of a survival signal, die;
- a minority bind antigen with increased affinity and compete with those of lower affinity (i.e. antibodies characteristic of the primary response).

This process, called **affinity maturation**, is dependent on both T cells and germinal centers.

Athymic mice lacking T cells do not form germinal centers and show no affinity maturation.

Mutations in the DNA of two heavy chain genes

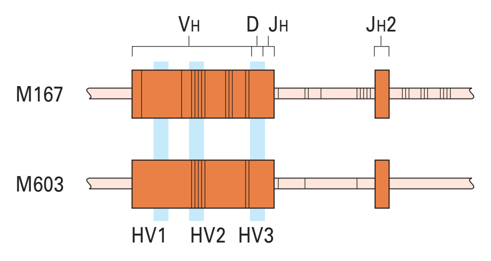

Fig. 3.32 DNA of two IgG antibodies to phosphorylcholine are illustrated. Both antibodies share the T15 idiotype. Black lines indicate positions where the sequence has mutated from the germline sequence. There are large numbers of mutations in both the introns and exons of both genes, but particularly in the second hypervariable region (HV2). By comparison, there are no mutations in the regions encoding the constant genes, implying that the mutational mechanism is highly localized.

Diversity is generated at several different levels

Antibody diversity therefore arises at several levels:

- first there are the multiple V genes recombining with D and J segments;
- the imprecision with which joining of the segments occurs produces further variation;
- as virtually any light chain may pair with any heavy chain, the combinatorial binding of heavy and light chains amplifies the diversity enormously (Fig. 3.33).

Seven mechanisms for the generation of antibody diversity

1. multiple germline V genes
2. VJ and VDJ recombinations
3. N–nucleotide addition
4. gene conversion
5. recombinational inaccuracies
6. somatic point mutation
7. assorted heavy and light chains

Fig. 3.33 Because each mechanism can occur with any of the others, the potential for increased diversity multiplies at each step of immunoglobulin production.

The structures of the first and second hypervariable regions (CDR-1 and CDR-2) of the primary antibody repertoire are encoded entirely by germline gene segments whereas CDR-3 diversity is generated by recombination events.

Different species have different strategies for generating diversity

Current understanding of the mechanisms for generating antibody diversity results primarily from studies in the mouse and human, species that show remarkable similarities in gene organization and expression. However, other species have evolved additional solutions to the problem of generating antibody diversity from a relatively small amount of genetic information. As a result, whereas mouse and human employ recombination among multiple germline genes together with somatic mutation:

- sharks rely on a large number of antibody genes and do not use somatic recombination (Fig. 3.34); and
- chickens have a very small number of antibody genes that undergo a very high level of gene conversion (Fig. 3.35).

Shark immunoglobulin VH loci

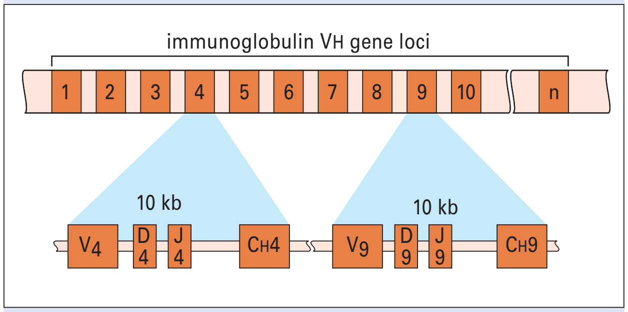

Fig. 3.34 In the shark there is a series of about 200 heavy chain gene clusters, each with a single V, D, J, and C gene segment. The fourth and ninth clusters are shown expanded. VH, DH, and JH segments are closely linked and occur within approximately 1.3 kb. Together with the CH segment, they occupy only about 10–15 kb. The arrangement of genes within a cluster appears to be encoded in the germline rather than by somatic rearrangement mechanisms, which may account for the lack of interindividual variation associated with the immune response of this species. (Based on data of Dr JJ Marchalonis and Dr GW Litman)

Genetic basis of antibody diversity in chickens

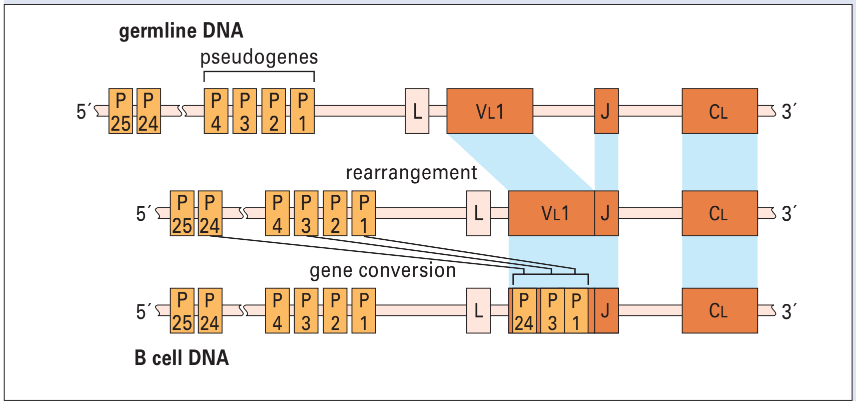

Fig. 3.35 The chicken germline immunoglobulin light chain locus has less than 30 kb of DNA. A single functional V gene (VL) lies 2 kb upstream from a single JC unit, with an adjacent cluster of 25 pseudogenes (P) in a 19-kb region. Rearrangement occurs briefly during early B cell development. Antibody diversity is achieved by gene conversion between P and the rearranged sequence. The arrangement shown (P1, P3, and P24) is illustrative; converted segments do not necessarily lie in order in the V gene segment.

FURTHER READING

Almogren A, Senior BW, Loomes LM, Kerr MA. Structural and functional consequences of cleavage of human secretory and human serum immunoglobulin A1 by proteinases from *Proteus mirabilis* and *Neisseria meningitidis*. Infect Immun 2003;71:3349–3356.

Burton D, Woof JM. Human antibody effector function. Adv Immunol 1992;51:1–84.

Glennie MJ, van de Winkel JG. Renaissance of cancer therapeutic antibodies. Drug Discov Today 2003;8:503–510.

Gould HJ, Sutton BJ, Beavil AJ, et al. The biology of IgE and the basis of allergic disease. Annu Rev Immunol 2003;21:579–628.

Jefferis R, Lund J, Pound JD. IgG-Fc-mediated effector functions: molecular definition of interaction sites for effector ligands and the role of glycosylation. Immunol Rev 1998;163:59–76.

Jefferis R, Lund J. Interaction sites on human IgG-Fc for FcγR: current models. Immunol Lett 2002;82:57–65.

Jung D, Alt FW. Unraveling V(D)J recombination; insights into gene regulation. Cell 2004;116:299–311.

Montiero RC, van de Winkel JG. IgA Fc receptors. Annu Rev Immunol 2003;21:177–204.

Padlan EA. X-ray crystallography of antibodies. Adv Protein Chem 1996;49:57–133.

Ravetch JV, Bolland S. IgG Fc receptors. Annu Rev Immunol 2001;19:275–290.

Rudd PM, Elliott T, Cresswell P, et al. Glycosylation and the immune system. Science 2001;291:2370–2376.

Woof JM, Burton D. Human antibody-Fc receptor interactions illuminated by crystal structures. Nat Rev Immunol 2004;4:89–99.

INTERNET REFERENCES

IMGT, the International ImMunoGeneTics Information System® http://imgt.cines.fr

- An integrated knowledge resource for the immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), immunoglobulin superfamily and related proteins of the immune system of human and other vertebrate species.

Mike Clark's Immunoglobulin Structure/Function Home Page http://www.path.cam.ac.uk/~mrc7/mikeimages.html

- Provides a wealth of information, of his own generation and through access to many other related sites.

National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/

- Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information – all for the better understanding of molecular processes affecting human health and disease.

Roy Jefferis Home Page http://medweb.bham.ac.uk/immunity infection/research/RoyJefferis/home.html

- Provides a basic introductory course and information on the influence of glycosylation on antibody structure and function

V BASE http://www.mrc-cpe.cam.ac.uk/COMPILATION.php?menu=901

- Compiled using all available human germline V gene sequences (V, D, or J) taken from the literature and from the EMBL/Genbank databases.
- Currently, 961 sequences from 145 different publications make up the V BASE database.

Critical thinking: The specificity of antibodies (see p. 493 for explanations)

The human rhinovirus HRV14 is formed from four different polypeptides: one of them (VP4) is associated with viral RNA in the core of the virus, while the other three polypeptides (VP1–VP3) make up the shell of the virus – the capsid.

1 When virus is propagated in the presence of neutralizing antiviral antiserum it is found that mutated forms of the virus develop. Mutations are detected in VP1, VP2, or VP3, but never in VP4. Why should this be so?

The most effective neutralizing antibodies are directed against the protein VP1 – this is termed an immunodominant antigen. Two different monoclonal antibodies against VP1 were developed and used to induce mutated forms of the virus. When the sequences of the mutated variants were compared with the original virus, it was found that only certain amino acid residues became mutated (see table below).

antibody	amino acid number	residue in wild type	observed mutations
VP1-a	91	Glu	Ala, Asp, Gly, His, Asn, Val, Tyr
VP1-a	95	Asp	Gly, Lys
VP1-b	83	Gln	His
VP1-b	85	Lys	Asn
VP1-b	138	Glu	Asp, Gly
VP1-b	139	Ser	Pro

2 What can you tell about the epitopes that are recognized by the two different monoclonal antibodies?

3 When the binding of the antibody VP1-a is measured against the different mutant viruses, it is found that it binds with high affinity to the variant with glycine (Gly) at position 138, with low affinity to the variant with Gly at position 95, and does not bind to the variant with lysine (Lys) at position 95. How can you explain these observations?

CHAPTER

4

Complement

SUMMARY

- **Complement is central to the development of inflammatory reactions** and forms one of the major immune defense systems of the body.
- **Complement activation pathways have evolved to label pathogens for elimination.** The classical pathway links to the adaptive immune system. The alternative and lectin pathways provide non-specific 'innate' immunity, and the alternative pathway is linked to the classical pathway.
- **The complement system is controlled to protect the host**. C1 inhibitor controls the classical and lectin pathways. C3 and C5 convertase activity are controlled by decay and enzymatic degradation.
- **The membrane attack pathway results in the formation of a transmembrane pore.** Regulation of the membrane attack pathway reduces the risk of 'bystander' damage to adjacent cells.
- **Many cells express one or more membrane receptors for complement products**. Receptors for fragments of C3 are widely distributed on different leukocyte populations. Receptors for C1q are present on phagocytes, mast cells, and platelets The plasma complement regulator fH binds leukocyte surfaces.
- **Complement has a variety of functions.** Its principal functions are chemotaxis including opsonization and cell activation, lysis of target cells, and priming of the adaptive immune response.
- **Complement deficiencies illustrate the homeostatic roles of complement.** Classical pathway deficiencies result in tissue inflammation. Deficiencies of mannan-binding lectin (MBL) are associated with infection in infants. Alternative pathway and C3 deficiencies are associated with bacterial infections. Terminal pathway deficiencies predispose to Gram-negative bacterial infections. C1 inhibitor deficiency leads to hereditary angioedema. Deficiencies in alternative pathway regulators produce a secondary loss of C3.

COMPLEMENT IS CENTRAL TO THE DEVELOPMENT OF INFLAMMATORY REACTIONS

The complement system was discovered at the end of the 19th century as a heat-labile component of serum that augmented (or 'complemented') its bactericidal properties.

Complement is now known to comprise some 16 plasma proteins, together constituting nearly 10% of the total serum proteins, and forming one of the major immune defense systems of the body (Fig. 4.1). The functions of the complement system include:

- the triggering and amplification of inflammatory reactions;
- attraction of phagocytes by chemotaxis;
- clearance of immune complexes;
- cellular activation;
- direct microbial killing; and
- an important role in the development of antibody responses.

In evolutionary terms the complement system is very ancient and antedates the development of the adaptive immune system: even starfish and worms have a functional complement system.

The importance of complement in immune defense is readily apparent in individuals who lack particular components – for example, children who lack the component **C3** are subject to overwhelming bacterial infections.

Like most elements of the immune system, when overactivated or activated in the wrong place the complement system can cause harm.

Complement is involved in the pathology of many diseases, provoking a search for therapies that control complement activation.

COMPLEMENT ACTIVATION PATHWAYS HAVE EVOLVED TO LABEL PATHOGENS FOR ELIMINATION

One major function of complement is to label pathogens and other foreign or toxic bodies for elimination from the host. The complement activation pathways have evolved to serve this purpose, and the multiple ways in which activation can be triggered, together with intrinsic amplification mechanisms, ensure efficient:

- recognition; and
- clearance.

Role of complement in inflammation

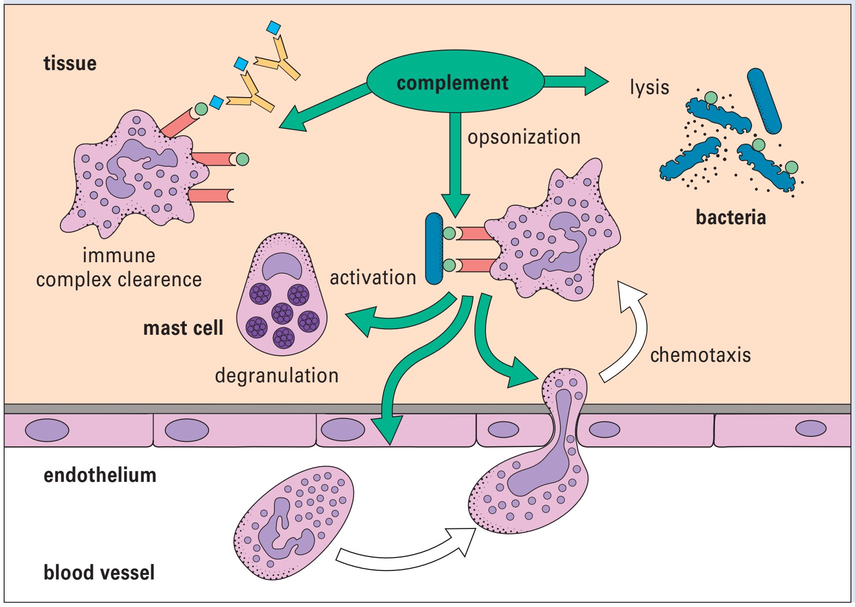
tissue
complement
lysis
opsonization
bacteria
immune
complex clearence
activation
mast cell
degranulation
chemotaxis
endothelium
blood vessel

Fig. 4.1 Complement has a central role in inflammation causing chemotaxis of phagocytes, activation of mast cells and phagocytes, opsonization and lysis of pathogens, and clearance of immune complexes.

Moreover, there are several different ways to activate the complement system, so providing a large degree of flexibility in response (Fig. 4.2).

The first activation pathway to be discovered, now termed the **classical pathway**, is initiated by antibodies bound to the surface of the target. Although an efficient means of activation, it requires that the host has previously encountered the target microorganism in order for an antibody response to be generated.

The **alternative pathway**, described in the 1950s, provides an antibody-independent mechanism for complement activation on pathogen surfaces.

The **lectin pathway**, the most recently described activation pathway, also bypasses antibody to enable efficient activation on pathogens.

All three pathways – classical, alternative, and lectin pathways:

- involve the activation of C3, which is the most abundant and most important of the complement proteins;
- comprise a proteolytic cascade in which complexes of complement proteins create enzymes that cleave other complement proteins in an ordered manner to create new enzymes, thereby amplifying and perpetuating the activation cascade.

Thus a small initial stimulus can rapidly generate a large effect. Fig. 4.3 summarizes how each of the pathways is activated.

All activation pathways converge on a common **terminal pathway** – a non-enzymatic system for causing membrane disruption and lytic killing of pathogens.

The immune defense and pathological effects of complement activation are mediated by the fragments and complexes generated during activation:

- the small chemotactic and proinflammatory fragments **C3a** and **C5a**;

Complement activation pathways

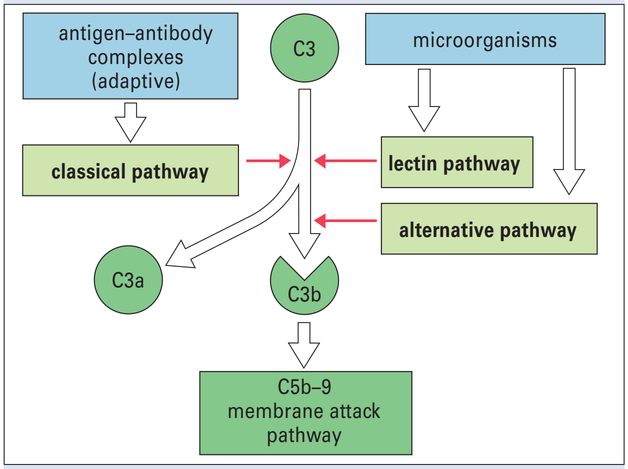
antigen–antibody complexes (adaptive)
C3
microorganisms
classical pathway
lectin pathway
alternative pathway
C3a
C3b
C5b–9
membrane attack
pathway

Fig. 4.2 Each of the activation pathways generates a C3 convertase, which converts C3 to C3b, the central event of the complement pathway. C3b in turn activates the terminal lytic membrane attack pathway. The first stage in the classical pathway is the binding of antigen to antibody. The alternative pathway does not require antibody and is initiated by the covalent binding of C3b to hydroxyl and amine groups on the surface of various microorganisms. The lectin pathway is also triggered by microorganisms in the absence of antibody, with sugar residues on the pathogen surface providing the binding sites. The alternative and lectin pathways provide non-specific 'innate' immunity, whereas the classical pathway represents a more recently evolved link to the adaptive immune system.

Rapid Reference Box 4

alternative pathway – the activation pathways of the complement system involving C3 and factors B, D, P, H, and I, which interact in the vicinity of an activator surface to form an alternative pathway C3 convertase (C$\overline{3bBb}$).

amplification loop – the alternative complement activation pathway that acts as a positive feedback loop when C3 is split in the presence of an activator surface.

anaphylatoxins – complement peptides (C3a and C5a) that cause mast cell degranulation and smooth muscle contraction.

bystander lysis – complement-mediated lysis of cells in the immediate vicinity of a complement activation site that are not themselves responsible for the activation.

C1–C9 – the components of the complement pathways responsible for mediating inflammatory reactions, opsonization of particles, and lysis of cell membranes.

C3 convertases – the enzyme complexes C$\overline{3bBb}$ and C$\overline{4b2a}$ that cleave C3.

classical pathway – the pathway by which antigen–antibody complexes can activate the complement system, involving components C1, C2, and C4, and generating a classical pathway C3 convertase (C$\overline{4b2a}$).

complement – a group of serum proteins involved in the control of inflammation, the activation of phagocytes, and the lytic attack on cell membranes.

complement receptors (CR1–CR4) – a set of four cell surface receptors for fragments of complement C3.

decay accelerating factor (DAF) – a cell surface molecule on mammalian cells that limits activation and deposition of complement C3b.

lectin pathway – a pathway of complement activation, initiated by mannan-binding lectin (MBL), that intersects the classical pathway.

membrane attack complex (MAC) – the assembled terminal complement components C5b–C9 of the lytic pathway that becomes inserted into cell membranes.

zymogen – pro-enzyme that requires proteolytic cleavage to become active; the enzymatically active form is distinguished from its precursor by a bar drawn above.

Summary of the activators of the classical, lectin, and alternative pathways

	immunoglobulins	microorganisms			other
		viruses	bacteria	other	
classical pathway	immune complexes containing IgM, IgG1, IgG2, or IgG3	HIV and other retroviruses, vesicular stomatitis virus		*Mycoplasma* spp.	polyanions, especially when bound to cations PO_4^{3-} (DNA, lipid A, cardiolipin) SO_4^{2-} (dextran sulfate, heparin, chondroitin sulfate)
lectin pathway		HIV and other retroviruses	many Gram-positive and Gram-negative organisms		arrays of terminal mannose groups
alternative pathway	immune complexes containing IgG, IgA, or IgE (less efficient than the classical pathway)	some virus-infected cells (e.g. by Epstein –Barr virus)	many Gram-positive and Gram-negative organisms	trypanosomes, *Leishmania* spp., many fungi	dextran sulfate, heterologous erythrocytes, complex carbohydrates (e.g. zymosan)

Fig. 4.3 Summary of the activators of the classical, lectin, and alternative complement activation pathways.

- the large opsonic fragments **C3b** and **C4b**; and
- the lytic **membrane attack complex (MAC)**.

The details of complement activation, the nomenclature, and the ways in which the pathways are controlled are shown in Fig. 4.4.

The classical pathway links to the adaptive immune system

The classical pathway is activated by antibody bound to antigen and requires Ca^{2+}

Only surface-bound IgG and IgM antibodies can activate complement, and they do so via the classical pathway. Surface binding is the key:

- IgM is the most efficient activator, but unbound IgM in plasma does not activate complement;
- among IgG subclasses, IgG1 and IgG3 are strong complement activators, whereas IgG4 does not activate because it is unable to bind the first component of the classical pathway.

Q. What occurs when IgM binds to the surface of a bacterium that allows it to activate complement?

A. A transition occurs from a flat planar molecule to a staple form, which exposes binding sites for the first component of the complement system, C1 (see Fig. 3.9).

Overview of the complement activation pathways

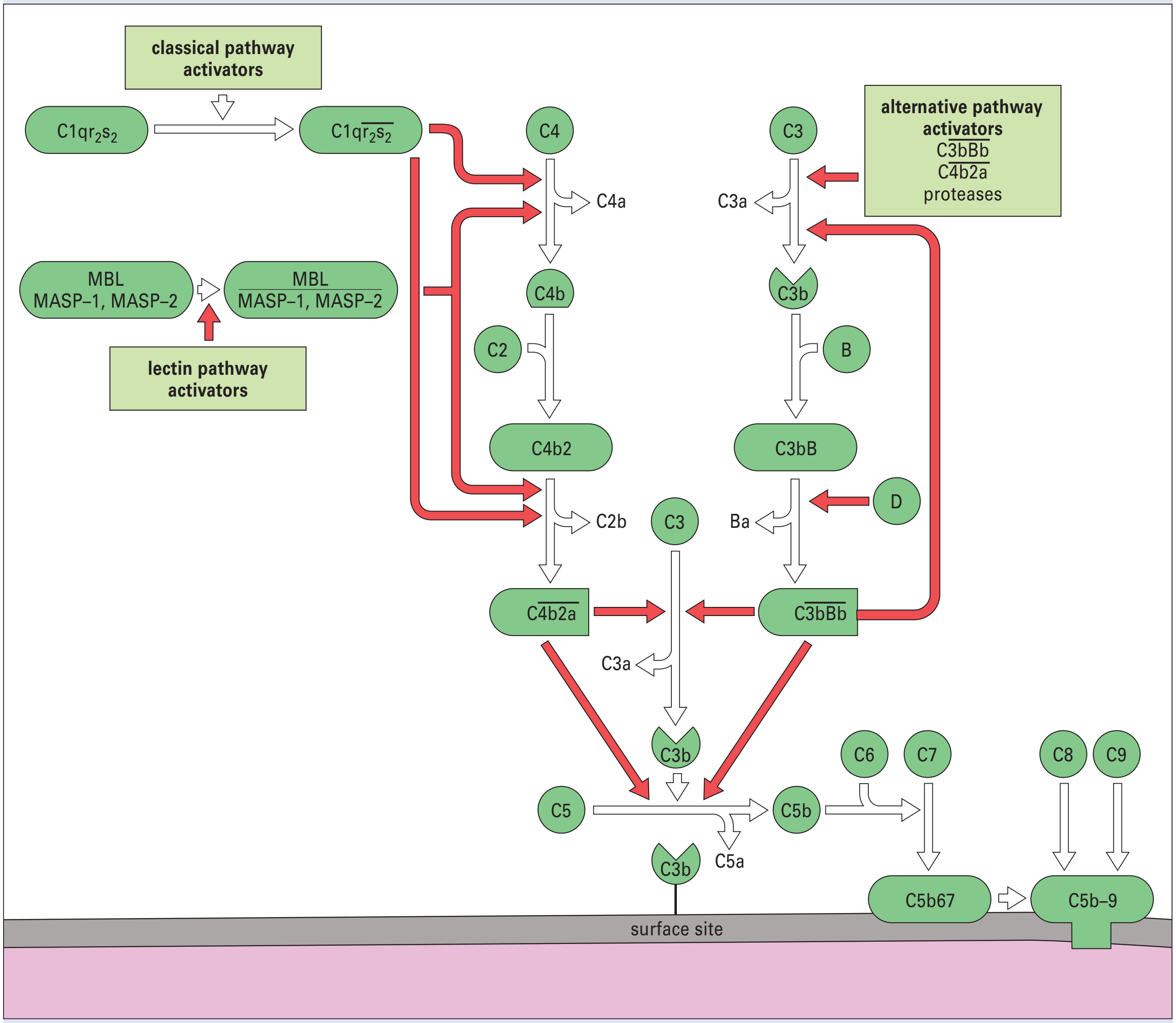
classical pathway activators
$C1qr_2s_2$
$C1\overline{qr_2s_2}$
C4
C4a
C3
C3a
alternative pathway activators
$\overline{C3bBb}$
$\overline{C4b2a}$
proteases
MBL
MASP–1, MASP–2
$\overline{\text{MBL}}$
$\overline{\text{MASP–1, MASP–2}}$
lectin pathway activators
C4b
C3b
C2
B
C4b2
C3bB
D
C2b
C3
Ba
$C\overline{4b2a}$
$C\overline{3bBb}$
C3a
C3b
C6
C7
C8
C9
C5
C5b
C3b
C5a
C5b67
C5b–9
surface site

Fig. 4.4 The proteins of the classical and alternative pathways are assigned numbers (e.g. C1, C2). Many of these are zymogens (i.e. pro-enzymes that require proteolytic cleavage to become active). The cleavage products of complement proteins are distinguished from parent molecules by suffix letters (e.g. C3a, C3b). The proteins of the alternative pathway are called 'factors' and are identified by single letters (e.g. factor B, which may be abbreviated to fB or just 'B'). Components are shown in green, conversion steps as white arrows, and activation/cleavage steps as red arrows. The classical pathway is activated by the cleavage of C1r and C1s following association of $C1qr_2s_2$ with classical pathway activators (see Fig. 4.3), including immune complexes. Activated C1s cleaves C4 and C2 to form the classical pathway C3 convertase $C\overline{4b2a}$. Cleavage of C4 and C2 can also be effected via MASP-1 and MASP-2 of the lectin pathway, which are associated with mannan-binding lectin (MBL). The alternative pathway is activated by the cleavage of C3 to C3b, which associates with factor B and is cleaved by factor D to generate the alternative pathway C3 convertase $C\overline{3bBb}$. The initial activation of C3 happens to some extent spontaneously, but this step can also be effected by classical or alternative pathway C3 convertases or a number of other serum or microbial proteases. Note that C3b generated in the alternative pathway can bind more factor B and generate a positive feedback loop to amplify activation on the surface. Note also that the activation pathways are functionally analogous, and the diagram emphasizes these similarities. For example, C3 and C4 are homologous, as are C2 and factor B. MASP-1 and MASP-2 are homologous to C1r and C1s, respectively. Either the classical or alternative pathway C3 convertases may associate with C3b bound on a cell surface to form C5 convertases, $C\overline{4b2a3b}$ or $C\overline{3bBb3b}$, which split C5. The larger fragment C5b associates with C6 and C7, which can then bind to plasma membranes. The complex of C5b67 assembles C8 and a number of molecules of C9 to form a membrane attack complex (MAC), C5b–9.

The first component of the pathway, **C1**, is a complex molecule comprising a large recognition unit termed **C1q** and two molecules each of **C1r** and **C1s**, the enzymatic units of the complex (Fig. 4.5). Assembly of the C1 complex is Ca^{2+}-dependent, and the classical pathway is therefore inactive if Ca^{2+} ions are absent.

C1 activation occurs only when several of the head groups of C1q are bound to antibody

C1q in the C1 complex binds through its globular head groups to the Fc regions of the immobilized antibody and undergoes changes in shape that trigger autocatalytic activation of the enzymatic unit C1r. Activated C1r then cleaves C1s at a single site in the protein to activate it.

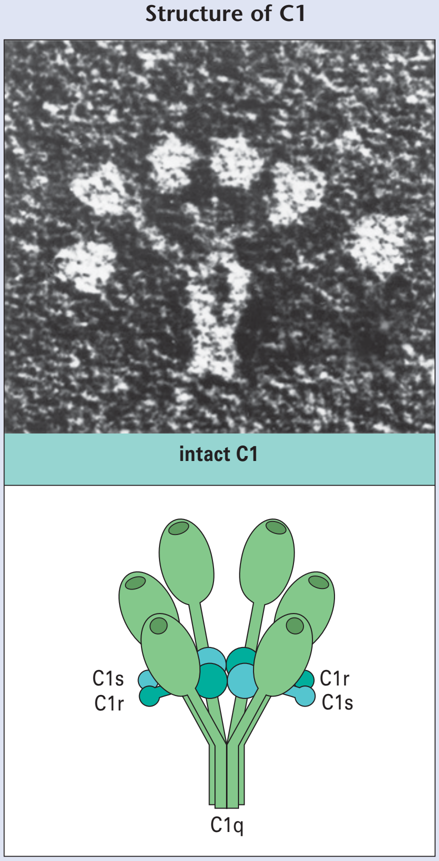
Structure of C1

Fig. 4.5 Electron micrograph of a human C1q molecule demonstrates six subunits. Each subunit contains three polypeptide chains, giving 18 in the whole molecule. The receptors for the Fc regions of IgG and IgM are in the globular heads. The connecting stalks contain regions of triple helix and the central core region contains collagen-like triple helix. The lower panel shows a model of intact C1 with two C1r and two C1s pro-enzymes positioned within the ring. The catalytic heads of C1r and C1s are closely apposed and conformational change induced in C1q following binding to complexed immunoglobulin causes mutual activation/cleavage of each C1r unit followed by cleavage of the two C1s units. The cohesion of the entire complex is dependent on Ca^{2+}. (Electron micrograph, reproduced by courtesy of Dr N Hughes-Jones)

Because C1 activation occurs only when several of the head groups of C1q are bound to antibody, only surfaces that are densely coated with antibody will trigger the process. This limitation reduces the risk of inappropriate activation on host tissues.

C1s enzyme cleaves C4 and C2

The C1s enzyme has two substrates – C4 and C2 – which are the next two proteins in the classical pathway sequence. (Note that the complement components were named chronologically, according to the order of their discovery, rather than according to their position in the reaction.) C1s cleaves the abundant plasma protein **C4** at a single site in the molecule:

- releasing a small fragment, **C4a**; and
- exposing a labile thioester group in the large fragment **C4b**.

Through the highly reactive thioester, C4b becomes covalently linked to the activating surface (Fig. 4.6).

C4b binds the next component, **C2**, in a Mg^{2+}-dependent complex and presents it for cleavage by C1s in an adjacent C1 complex:

- the fragment **C2b** is released; and
- **C2a** remains associated with C4b on the surface.

C$\overline{4b2a}$ is the classical pathway C3 convertase

The complex of C4b and C2a (termed **C$\overline{4b2a}$** – the classical pathway C3 convertase) is the next activation enzyme. C2a in the C$\overline{4b2a}$ complex cleaves **C3**, the most abundant of the complement proteins:

- releasing a small fragment, **C3a**; and

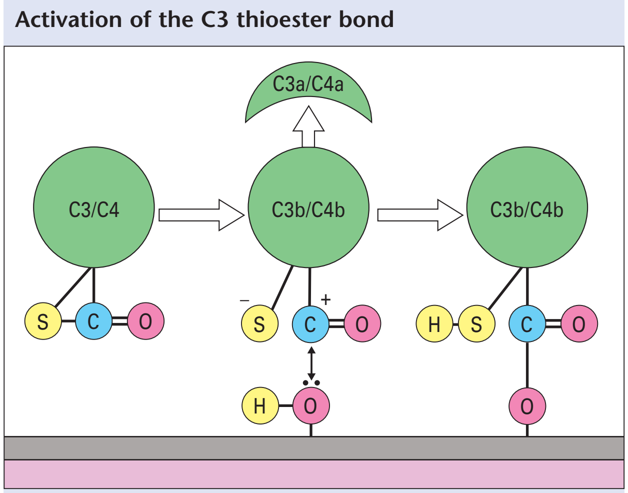
Activation of the C3 thioester bond

Fig. 4.6 The α chain of C3 contains a thioester bond formed between a cysteine and a glutamine residue, with the elimination of ammonia. Following cleavage of C3 into C3a and C3b, the bond becomes unstable and susceptible to nucleophilic attack by electrons on -OH and -NH_2 groups, allowing the C3b to form covalent bonds with proteins and carbohydrates – the active group decays rapidly by hydrolysis if such a bond does not form. C4 also contains an identical thioester bond, which becomes activated similarly when C4 is split into C4a and C4b.

- exposing a labile thioester group in the large fragment **C3b**.

As described above for C4b, C3b covalently binds the activating surface.

C$\overline{4b2a3b}$ is the classical pathway C5 convertase

Some of the C3b formed will bind directly to C$\overline{4b2a}$, and the trimolecular complex formed, **C$\overline{4b2a3b}$** (the classical pathway C5 convertase), can bind **C5** and present it for cleavage by C2a:

- a small fragment, **C5a**, is released; and
- the large fragment, **C5b**, remains associated with the C$\overline{4b2a3b}$ complex.

Cleavage of C5 is the final enzymatic step in the classical pathway.

The ability of C4b and C3b to bind surfaces is fundamental to complement function

C3 and C4 are homologous molecules that contain an unusual structural feature – an internal thioester bond between a glutamine and a cysteine residue that, in the intact molecule, is buried within the protein.

When either C3 or C4 is cleaved by the convertase enzyme, a conformational change takes place that exposes the internal thioester bond in C3b and C4b, making it very unstable and highly susceptible to attack by nucleophiles such as hydroxyl groups (-OH) and amine groups (-NH_2) in membrane proteins and carbohydrates. This reaction creates a covalent bond between the complement fragment and the membrane ligand, locking C3b and C4b onto the surface (see Fig. 4.6).

The exposed thioester remains reactive for only a few milliseconds because it is susceptible to hydrolysis by water. This lability restricts the binding of C3b and C4b to the immediate vicinity of the activating enzyme and prevents damage to surrounding structures.

The alternative and lectin pathways provide non-specific 'innate' immunity

The lectin pathway is activated by microbial carbohydrates

The lectin pathway differs from the classical pathway only in the initial recognition and activation steps. Indeed, it can be argued that the lectin pathway should not be considered a separate pathway, but rather a route for classical pathway activation that bypasses the need for antibody.

The C1 complex is replaced by a structurally similar multimolecular complex, comprising the C1q-like recognition unit **mannan-binding lectin (MBL)** and several **MBL-associated serine proteases (MASPs)**, which provide enzymatic activity. As in the classical pathway, assembly of this initiating complex is Ca^{2+}-dependent.

C1q and MBL are members of the collectin family of proteins characterized by globular head regions with binding activities and long collagenous tail regions with diverse roles (see Fig. 6.23).

MBL binds the simple carbohydrates mannose and *N*-acetyl glucosamine present in the cell walls of diverse pathogens, including bacteria, yeast, fungi, and viruses. Binding induces shape changes in MBL that in turn induce autocatalytic activation of the MASPs. These enzymes can cleave C4 and C2 to continue activation exactly as in the classical pathway.

The lectin pathway is not the only means of activating the classical pathway in the absence of antibody. Mitochondria and other products of cell damage can directly bind C1q, triggering complement activation and aiding the clearance of the dead and dying tissue.

Alternative pathway activation is accelerated by microbial surfaces and requires Mg^{2+}

The alternative pathway of complement activation also provides antibody-independent activation of complement on pathogen surfaces. This pathway is in a constant state of low-level activation (termed 'tickover').

C3 is hydrolyzed at a slow rate in plasma and the product, **C3(H_2O)**, has many of the properties of C3b, including the capacity to bind a plasma protein **factor B (fB)**, which is a close relative of the classical pathway protein C2. Formation of the complex between C3b (or C3(H_2O)) and fB is Mg^{2+}-dependent, so the alternative pathway is inactive in the absence of Mg^{2+} ions. (The differences in the ion requirements of the classical and alternative pathway are exploited in laboratory tests for complement activity.)

The C$\overline{3bBb}$ complex is the C3 convertase of the alternative pathway

Once bound to C3(H_2O) or C3b, fB can bind and activate a plasma enzyme termed **factor D (fD)**. fD cuts fB in the C3bB complex:

- releasing a fragment, **Ba**;
- while the residual portion, **Bb**, becomes an active protease.

The **C$\overline{3bBb}$** complex is the C3 cleaving enzyme (C3 convertase) of the alternative pathway. C3b generated by this convertase can be fed back into the pathway to create more C3 convertases, thus forming a positive feedback amplification loop (Fig. 4.7). Activation may occur in plasma or, more efficiently, on surfaces.

Q. What physiological advantages and problems can you see in a system with a positive feedback loop (i.e. where the presence of C3b leads to the production of an enzyme C$\overline{3bBb}$ that generates more C3b)?

A. The positive feedback amplification and 'always on' features of the alternative pathway are well suited to pathogen surveillance. For example, a small initial stimulus could produce the deposition of large amounts of C3b on a pathogen surface, thereby facilitating its phagocytosis. If unregulated, however, the system will continue to activate until all available C3 has been consumed. Self cells would also be subjected to complement activation and could be damaged or destroyed.

Specific features of host cell surfaces, including their surface carbohydrates and the presence of complement regulators (see below), act to rescue the host cell from being destroyed, and such surfaces are non-activating.

On an activating surface such as a bacterial membrane, amplification will occur unimpeded and the surface will

Regulation of the amplification loop

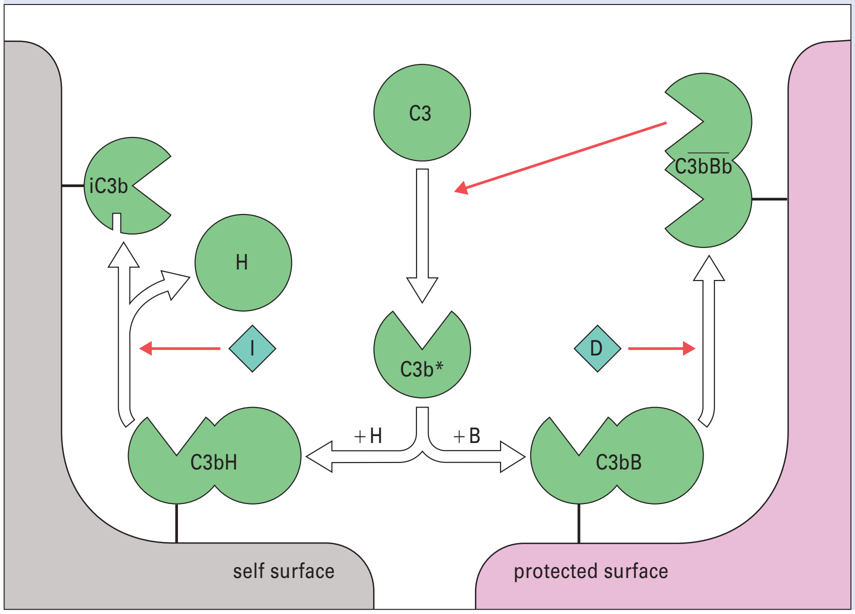

Fig. 4.7 Alternative pathway activation depends on the presence of activator surfaces. C3b bound to an activator surface recruits factor B, which is cleaved by factor D to produce the alternative pathway C3 convertase $\mathrm{C3\overline{bBb}}$, which drives the amplification loop, by cleaving more C3. However, on self surfaces the binding of factor H is favored and C3b is inactivated by factor I. Thus the binding of factor B or factor H controls the development of the alternative pathway reactions. In addition, proteins such as membrane cofactor protein (MCP) and decay accelerating factor (DAF) also limit complement activation on self cell membranes (see Fig. 4.9).

rapidly become coated with C3b (see Fig. 4.7). In a manner analogous to that seen in the classical pathway, C3b molecules binding to the C3 convertase will change the substrate specificity of the complex, creating a C5 cleaving enzyme, $\mathrm{C3\overline{bBbC3b}}$.

Cleavage of C5 is the last proteolytic step in the alternative pathway and the C5b fragment remains associated with the convertase.

The alternative pathway is linked to the classical pathway

The alternative pathway is inexorably linked to the classical pathway in that C3b generated through the classical pathway will feed into the alternative pathway to amplify activation. It therefore does not matter whether the initial C3b is generated by the classical, lectin, or alternative pathway – the amplification loop can ratchet up the reactions if they take place near an activator surface.

THE COMPLEMENT SYSTEM IS CONTROLLED TO PROTECT THE HOST

Control of the complement system is required to prevent the consumption of components through unregulated amplification and to protect the host. Complement activation poses a potential threat to host cells, because it could lead to cell lysis. To defend against this threat a family of regulators has evolved alongside the complement system to prevent uncontrolled activation and protect cells from attack.

C1 inhibitor controls the classical and lectin pathways

In the activation pathways, the regulators target the enzymes that drive amplification:

- activated C1 is controlled by a plasma serine protease inhibitor (**serpin**) termed **C1 inhibitor (C1inh)**, which removes C1r and C1s from the complex, switching off classical pathway activation;
- C1inh also regulates the lectin pathway in a similar manner, removing the MASP enzymes from the MBL complex to switch off activation.

C3 and C5 convertase activity are controlled by decay and enzymatic degradation

The C3 and C5 convertase enzymes are heavily policed with plasma and cell membrane inhibitors to control activation. In the plasma:

- **factor H (fH)** and a related protein **factor H-like 1 (fHL-1)** destroy the convertase enzymes of the alternative pathway;
- **C4 binding protein (C4bp)** performs the same task in the classical pathway.

On membranes, two proteins, **membrane cofactor protein (MCP)** and **decay accelerating factor (DAF)**, collaborate to destroy the convertases of both pathways (Fig. 4.8).

The regulators of the C3 and C5 convertases are structurally related molecules that have arisen by gene duplication in evolution. These duplicated genes are tightly linked in a cluster on chromosome 1, termed the **regulators of complement activation (RCA) locus**. This locus also encodes several of the complement receptors (see below).

Control of the convertases is mediated in two complementary ways

THE FIRST CONTROL IS DECAY ACCELERATION – The convertase complexes are labile, with a propensity to dissociate within a few minutes of creation. The regulators

C3 and C5 convertase regulators

	number of SCR domains	dissociation of C3 and C5 convertases		cofactor for factor I on		localization
		classical pathway	alternative pathway	C4b	C3b	
C4b binding protein (C4bp)	52 or 56 in 7or 8 chains	+	–	+	–	plasma
factor H (fH)	20	–	+	–	+	plasma
decay accelerating factor (DAF) (CD55)	4	+	+	–	–	blood cells, endothelia, epithelia
membrane cofactor protein (MCP) (CD46)	4	–	–	+	+	blood cells (not erythrocytes), endothelia, epithelia
complement receptor 1 (CR1) (CD35)	30	+	+	+	+	erythrocytes, B cells, follicular dendritic cells, macrophages

SCR, short consensus repeat

Fig. 4.8 The five proteins listed are widely distributed and control aspects of C3b and C4b dissociation or breakdown. Each of these proteins contains a number of short consensus repeat (SCR) domains. They act either by enhancing the dissociation of C3 and C5 convertases or by acting as cofactors for the action of factor I on C3b or C4b.

fH, fHL-1, and C4bp from the fluid phase and DAF on cell membranes bind the convertase complex and markedly accelerate decay, knocking out:

- C2a from the classical pathway convertase; and
- Bb from the alternative pathway enzyme (Fig. 4.9).

Very recently, genetic defects in fH have been implicated in patients with the non-diarrheal form of hemolytic uremic syndrome (HUS).

Point mutations in fH inhibit binding to host cells and weaken defense against complement, resulting in the hemolysis, platelet dysfunction, and renal damage that typify the syndrome.

THE SECOND CONTROL MECHANISM IS FACTOR I COFACTOR ACTIVITY – **Factor I (fI)** is a fluid-phase enzyme that, in the presence of an appropriate cofactor, can cleave and irreversibly inactivate C4b and C3b (see Fig. 4.9). MCP is a cofactor for fI cleavage of both C4b and C3b, whereas:

- C4bp specifically catalyzes cleavage of C4b; and
- fH/fHL-1 catalyzes cleavage of C3b.

It is interesting to note that, whereas the plasma regulators contain both activities in a single molecule, the two membrane regulators each contain only one activity.

Efficient regulation of the convertases on membranes therefore requires the concerted action of:

- DAF to dissociate the complex; and
- MCP to irreversibly inactivate it by catalyzing cleavage of the central component.

The alternative pathway also has a unique positive regulator, **properdin**, which stabilizes the C3 convertase and markedly increases its life span.

The two processes by which C3 convertase regulators inactivate the enzymes

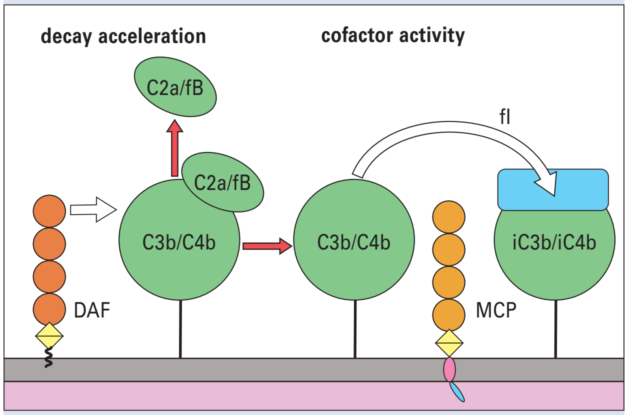

Fig. 4.9 DAF binds the enzyme complex, displacing the enzymatically active component (C2a or Bb). Membrane cofactor protein (MCP) binds the C3b/C4b unit released after decay and acts as a cofactor for factor I (fI) cleavage of C3b or C4b, resulting in the irreversible inactivation of the convertase.

Complement receptor 1 (CR1) is often included in the list of membrane regulators of C3 convertase activity, and, indeed, CR1 is a powerful regulator with both decay accelerating and cofactor activities in both pathways. Nevertheless, it is excluded from the above discussion because CR1 is primarily a receptor for complement-coated particles and does not have a role in protecting the host cell.

THE MEMBRANE ATTACK PATHWAY RESULTS IN THE FORMATION OF A TRANSMEMBRANE PORE

The **terminal or membrane attack pathway** involves a distinctive set of events whereby:

- a group of five globular plasma proteins associate with one another; and,
- in the process, acquire membrane-binding and pore-forming capacity to form the membrane attack complex (MAC).

The MAC is a transmembrane pore (Fig. 4.10). Cleavage of C5 creates the nidus for MAC assembly to begin. While still attached to the convertase enzyme, C5b binds first **C6** then **C7** from the plasma. Conformational changes occurring during assembly of this trimolecular **C5b67** complex:

- cause release from the convertase; and
- expose a labile hydrophobic site.

The membrane attack pathway

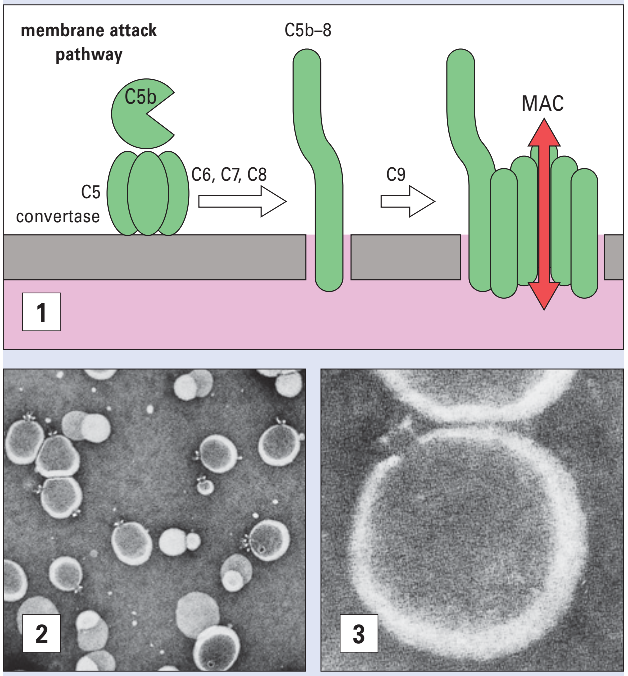

Fig. 4.10 (**1**) C5b, while still attached to the C5 convertase, binds C6 and C7 from the fluid phase. The trimolecular C5b–7 complex dissociates from the convertase and binds the target cell membrane. Binding of C8 and multiple copies of C9 generates a rigid protein-lined transmembrane channel, the membrane attack complex (MAC). (**2** and **3**) Electron micrographs of the MAC. The complex consists of a cylindrical pore in which the walls of the cylinder, formed by C9, traverse the cell membrane. In these micrographs the human C5b–9 complex has been incorporated into a lecithin liposomal membrane. × 234 000. (Courtesy of Professor J Tranum-Jensen and Dr S Bhakdi)

The complex can stably associate with a membrane through the labile hydrophobic site, though the process is inefficient and most of the C5b67 formed is inactivated in the fluid phase.

Membrane-bound C5b67 recruits **C8** from the plasma, and, finally, multiple copies of **C9** are incorporated in the complex to form the MAC.

The latter stages of assembly are accompanied by major conformational changes in the components with globular hydrophilic plasma proteins unfolding to reveal amphipathic regions that penetrate into and through the lipid bilayer.

The fully formed MAC creates a rigid pore in the membrane, the walls of which are formed from multiple (up to 12) copies of C9 arranged like barrel staves around a central cavity.

The MAC is clearly visible in electron microscopic images of complement-lysed cells as doughnut-shaped protein-lined pores, first observed by Humphrey and Dourmashkin 40 years ago (see Fig. 4.10). The pore has an inner diameter approaching 10 nm:

- allowing the free flow of solutes and electrolytes across the cell membrane; and
- because of the high internal osmotic pressure, causing the cell to swell and sometimes burst.

Metabolically inert targets such as aged erythrocytes are readily lyzed by even a small number of MAC lesions, whereas viable nucleated cells resist killing through a combination of ion pump activities and recovery processes that remove MAC lesions and plug membrane leaks.

Even in the absence of cell killing, MAC lesions may severely compromise cell function or cause cell activation.

Regulation of the membrane attack pathway reduces the risk of 'bystander' damage to adjacent cells

Although regulation in the activation pathways is the major way in which complement is controlled, there are further failsafe mechanisms to protect self cells from MAC damage and lysis.

First, the membrane binding site in C5b67 is labile. If the complex does not encounter a membrane within a fraction of a second after release from the convertase, the site is lost through:

- hydrolysis; or
- binding of one of the fluid-phase regulators of the terminal pathway – **S protein** (also termed **vitronectin**) or **clusterin** – both of which are multifunctional plasma proteins with diverse roles in homeostasis.

C8, an essential component of the MAC, also behaves as a regulator in that binding of C8 to C5b–7 in the fluid phase blocks the membrane binding site and prevents MAC formation.

The net effect of all these plasma controls is to limit MAC deposition to membranes in the immediate vicinity of the site of complement activation, so reducing risk of 'bystander' damage to adjacent cells.

Complexes that do bind are further regulated on host cells by **CD59**, a membrane protein that locks into the MAC as it assembles and inhibits binding of C9, thereby preventing pore formation (Fig. 4.11). CD59 is a broadly

Role of CD59 in protecting host cells from complement damage

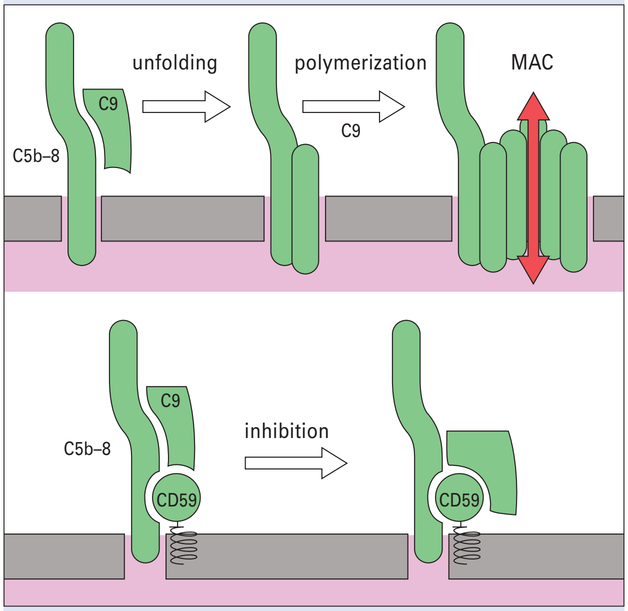

Fig. 4.11 The upper diagram models assembly of the MAC in the absence of the regulator CD59 – C9 binds C5b–8, unwinds, and traverses the membrane and recruits further C9 molecules, which in turn unfold and insert to form the MAC. In the lower diagram, CD59 binds the C5b–8 complex and prevents the unfolding and insertion of C9, which is essential for the initiation of MAC pore formation.

expressed, glycolipid-anchored protein that is structurally unrelated to the complement regulators described above.

IMPORTANCE OF CD59 IN PROTECTING HOST CELLS FROM COMPLEMENT DAMAGE – The importance of CD59 in protecting host cells from complement damage is well illustrated in the hemolytic disorder paroxysmal nocturnal hemoglobinuria (PNH), in which erythrocytes and other circulating cells are unable to make glycolipid anchors and as a consequence lack CD59 (and also DAF). The ‘tickover’ complement activation that occurs on all cells without much consequence is then sufficient in the absence of CD59 to cause low-grade hemolysis and hemolytic crises.

MANY CELLS EXPRESS ONE OR MORE MEMBRANE RECEPTORS FOR COMPLEMENT PRODUCTS

Receptors for fragments of C3 are widely distributed on different leukocyte populations

Many cells express one or more membrane receptors for complement products (Fig. 4.12). An understanding of the receptors is essential because the majority of the effects of complement are mediated through these molecules. The best characterized of the complement receptors are those binding fragments of C3.

CR1, CR2, CR3, and CR4 bind fragments of C3 attached to activating surfaces

Four different receptors, termed complement receptors 1, 2, 3, and 4 (CR1, CR2, CR3, and CR4), bind fragments of C3 attached to activating surfaces:

- **CR1**, expressed on erythrocytes and leukocytes, binds the largest fragment C3b (and also C4b), an interaction that is crucial to the processing of immune complexes (see below);
- **CR2**, expressed mainly on B cells and follicular dendritic cells (FDCs), binds fragments derived from fI-mediated proteolysis of C3b – iC3b and C3d;

These interactions aid the B cell response to complement-coated particles.

Both CR1 and CR2 are structurally related to the C3 convertase regulators fH, C4bp, MCP, and DAF, and are encoded in the RCA cluster on chromosome 1.

CR3 and **CR4**:

- belong to the integrin family of cell adhesion molecules;
- are expressed on the majority of leukocytes;
- bind the iC3b fragment, aiding adhesion of leukocytes to complement-coated particles and facilitating phagocytic ingestion of these particles.

Receptors for C3a and C5a mediate inflammation

C3a, the small fragment released during activation of C3, binds to a receptor (**C3aR**) expressed abundantly on eosinophils and basophils, and at much lower levels on neutrophils and many other cell types.

The C5a fragment released from C5 during activation is closely related to C3a and binds a distinct, but structurally related, receptor, the C5a receptor (**C5aR**), which is present on a wide variety of cell types, including all leukocytes.

The receptors for C3a and C5a are members of the large receptor family of seven-transmembrane segment receptors that associate with heterotrimeric G proteins. Receptors for chemokines belong to this same family and in many ways C3a and C5a behave like chemokines.

Together, C3aR and C5aR are important in orchestrating inflammatory responses (see below).

Receptors for C1q are present on phagocytes, mast cells, and platelets

Receptors for C1q are less well characterized than C3 receptors, but are increasingly recognized as important in homeostasis.

Receptors for the collagen tails (**cC1qR**):

- can recognize C1q attached through its globular head regions to complement-coated particles;
- are present on leukocytes, platelets, and some other cell types; and
- likely play roles in enhancing phagocytosis of C1q-labeled particles.

Receptors for the globular heads (**C1qRp**):

- bind C1q in an orientation that mimics antibody binding;
- are expressed principally on phagocytic cells, mast cells, and platelets; and
- may collaborate with cC1qR to mediate cell activation events.

Cell receptors for complement components and fragments

ligand	receptor	structure	function	location
C1q	cC1qR (C1q receptor enhancing phagocytosis)	acidic 100-kDa transmembrane glycoprotein	binds collagenous tail of C1q, enhances phagocytosis	myeloid cells, endothelia, platelets
	C1qRp (receptor for C1q globular heads)	acidic 33-kDa glycoprotein	binds globular heads of C1q, possible role in phagocytosis	all blood cells
C3, C4, and C5 fragments	CR1 (complement receptor 1 – CD35)	SCR-containing transmembrane glycoprotein, 30 SCRs	binds C3b and C4b, cofactor and decay accelerating activities, roles in immune complex handling	erythrocytes, B cells, FDCs, macrophages
	CR2 (complement receptor 2 – CD21)	SCR-containing transmembrane glycoprotein, 15 or 16 SCRs	binds C3d and iC3b, role in regulating B cell response to antigen	B cells, FDCs, some T cells, basophils, epithelia
	CR3 (complement receptor 3 – CD11b/CD18)	integrin family member, heterodimer	binds iC3b, roles in cell adhesion	myeloid cells, some B cells and NK cells
	CR4 (complement receptor 4 – CD11c/CD18)	integrin family member, heterodimer	binds iC3b, roles in cell adhesion	myeloid cells, FDCs, activated B cells
	C3aR (receptor for the C3a anaphylatoxin)	G protein-coupled 7-transmembrane spanning receptor	binds C3a, mediates cell activation	widely distributed on blood and tissue cells
	C5aR (receptor for the C5a anaphylatoxin – CD88).	G protein-coupled 7-transmembrane spanning receptor	binds C5a, mediates cell activation and chemotaxis	myeloid cells, smooth muscle, endothelia, epithelia

Fig. 4.12 Summary of information on the cell surface receptors for complement components and fragments and their biological roles. (FDC, follicular dendritic cell)

The bulk of C1q in the circulation is, however, already complexed with C1r and C1s to form intact C1, and in this complex C1q does not interact with its receptors. This ensures that C1q receptors are activated only in specific circumstances, such as during complement activation when free C1q is available.

The plasma complement regulator fH binds leukocyte surfaces

The plasma complement regulator fH binds leukocyte surfaces in a specific and saturable manner, suggesting the presence of an fH receptor, which remains uncharacterized.

It has been suggested that surface-bound fH plays important roles in modulating leukocyte activation as well as in protecting cells from complement attack. Indeed, a fascinating disease, atypical hemolytic uremic syndrome, which is typified by hemolysis, platelet destruction, and renal damage that may progress to renal failure, appears to be caused in many cases by loss of the surface binding capacity of fH. Uncontrolled complement activation due to the production of a mutant fH lacking the membrane binding site has been demonstrated in patients.

COMPLEMENT HAS A VARIETY OF FUNCTIONS

The principal functions of complement are:

- chemotaxis;
- opsonization and cell activation;
- lysis of target cells; and
- priming of the adaptive immune response.

C5a is chemotactic for macrophages and polymorphs

Polymorphs and macrophages express receptors for C3a and C5a. These small (~10 kDa) fragments (i.e. C3a and C5a) diffuse away from the site of activation, creating a chemical gradient along which the motile cells migrate to congregate at the site of activation (Fig. 4.13).

Binding of C3a and C5a to their receptors also causes cell activation:

- increasing adhesive properties;
- triggering extravasation; and
- priming phagocytes to release proinflammatory molecules including enzymes, vasoactive amines, reactive oxygen intermediates, and inflammatory cytokines.

Actions of anaphylatoxins

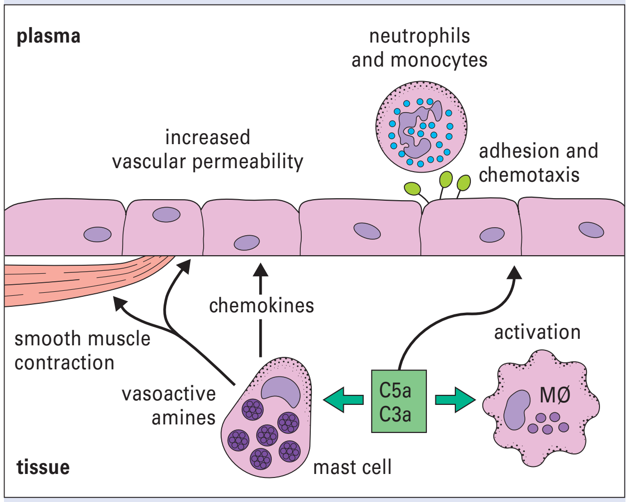

Fig. 4.13 C5a and C3a both act on mast cells to cause degranulation and release of vasoactive amines, including histamine and 5-hydroxytryptamine, which enhance vascular permeability and local blood flow. The secondary release of chemokines from mast cells causes cellular accumulation, and C5a itself acts directly on receptors on monocytes and neutrophils to induce their migration to sites of acute inflammation and subsequent activation.

C3a and C5a also enhance adhesion molecule expression on phagocytes, increasing cell stickiness, and may cause increased expression of the C3 fragment receptors CR1 and CR3.

Q. What effect would increased expression of CR1 and CR3 have on phagocytes?

A. It would enhance their ability to phagocytose particles coated with C3b and iC3b.

C3a and C5a activate mast cells and basophils

Tissue mast cells and basophils also express C3aR and C5aR, and binding of ligand triggers massive release of:

- histamine; and
- cytokines (see Fig. 4.13).

Together, these products cause local smooth muscle contraction and increased vascular permeability to generate the swelling, heat, and pain that typify the inflammatory response. These effects mirror on a local scale the more generalized and severe reactions that can occur in severe allergic or anaphylactic reactions, and for this reason C3a and C5a are sometimes referred to as **anaphylatoxins**.

The actions of C3a and C5a are limited temporally and spatially by the activity of a plasma enzyme, carboxypeptidase-N, which cleaves the carboxy terminal amino acid, arginine, from both of these fragments. The products, termed C3a-desArg and C5a-desArg (-desArg = without arginine), respectively, are either completely biologically inert (C3a-desArg) or have much reduced activity (C5a-desArg).

The retention in C5a-desArg of some chemotactic activity enables the recruitment of phagocytes even from distant sites, making C5a and its metabolite the most important complement-derived chemotactic factor.

C3b and iC3b are important opsonins

Complement activation and amplification cause complement fragments to efficiently coat activator surfaces of targets such as bacteria or immune complexes. Phagocytes and other cells carrying receptors for these complement fragments are then able to bind the target, triggering ingestion and cell activation. The key players here are the surface-bound fragments of C3 and the family of C3 fragment receptors described above.

The amplification inherent in the system ensures that bacteria and other activating surfaces rapidly become coated with C3b and its breakdown product **iC3b** (i.e. the process of **opsonization**; Fig. 4.14).

Phagocytes lured by the complement-derived chemotactic factors described above and activated to increase expression of CR1 and CR3 (receptors for C3b and iC3b, respectively) will bind the activating particle and engulf it for destruction in the phagosome system.

Opsonization, binding, and phagocytosis

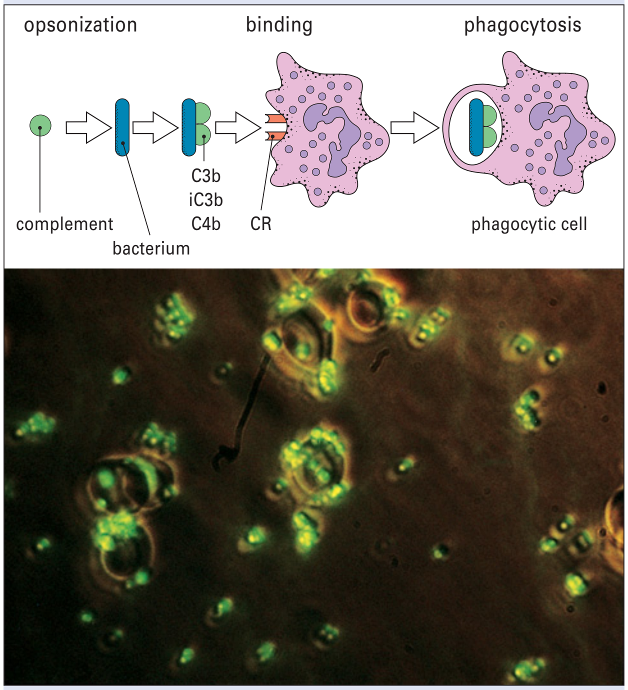

Fig. 4.14 A bacterium is sensitized by the covalent binding of C3b, iC3b, and C4b, which allow it to be recognized by complement receptors (CR) on neutrophils and mononuclear phagocytes. This promotes phagocytosis and activation of the phagocyte. In primates, erythrocytes also express CR1, which allows them to bind opsonized bacteria and immune complexes. In the lower panel fluoresceinated bacteria that have been opsonized with antibody and complement are seen adhering to human erythrocytes. (Courtesy of Professor GD Ross)

Q. In addition to C3b and iC3b, what other complement components can act as opsonins?

A. C1q via its interaction with C1q receptors and C4b by binding to CR1.

The importance of complement opsonization for defense against pathogens is illustrated in individuals deficient in complement components. C3 deficiency (see below) in particular is always associated with repeated severe bacterial infections that without adequate prophylaxis inevitably lead to early death.

C3b disaggregates immune complexes and promotes their clearance

Immune complexes containing antigens derived either from pathogens or from the death of host cells form continuously in health and disease. Because they tend to grow by aggregation and acquisition of more components, they can cause disease by precipitating in capillary beds in the skin, kidney, and other organs, where they drive inflammation.

Complement activation on the immune complex via the classical pathway efficiently opsonizes the immune complex and helps prevent precipitation in tissues:

- first, coating with C3b masks the foreign antigens in the core of the immune complex, blocking further growth;
- second, coating with C3b disaggregates large immune complexes by disrupting interactions between antigen and antibody;
- third, C3b (and C4b) on immune complexes interacts with CR1 on erythrocytes, taking the immune complex out of the plasma – the **immune adherence phenomenon**.

Immune complex adherence to erythrocytes provides an efficient means of handling and transporting the hazardous cargo to sites of disposal (i.e. the resident macrophages in spleen and liver). Here the immune complex is:

- released from the erythrocyte; and
- captured by complement and immunoglobulin receptors on the macrophage, internalized, and destroyed.

The MAC damages some bacteria and enveloped viruses

Assembly of the MAC creates a pore that inserts into and through lipid bilayers, breaching the membrane barrier (see Fig. 4.10). The consequences of MAC attack vary from target to target:

- for most pathogens, opsonization is the most important antibacterial action of complement;
- for Gram-negative bacteria, particularly organisms of the genus *Neisseria*, MAC attack is a major component of host defense, and individuals deficient in components of the MAC (e.g. patients with C6 deficiency, which is the second most common deficiency of complement) are susceptible to neisserial meningitis.

Gram-negative bacteria are protected by a double cell membrane separated by a peptidoglycan wall. Precisely how MAC traverses these protective structures to damage the inner bacterial membrane and causes osmotic lysis of these organisms remains unclear.

The MAC:

- may also play roles in the efficient dispatching of other pathogens, including some viruses;
- can also damage or destroy host cells – in some instances, such as autoimmunity, the host cell is itself the target and complement is directly activated on the cell, leading to MAC attack.

Q. What term is applied to the deposition of MACs on cells near to but not directly the cause of complement activation, and what mechanism normally limits this process?

A. This event is called **bystander lysis** and is normally limited by the presence of CD59 and fluid phase regulators, and the inefficiency of C5b6 deposition.

Erythrocytes have only a limited capacity to resist and repair damage and can be lysed, as is seen in autoimmune hemolytic anemias and some other hemolytic disorders. Although nucleated host cells may escape lysis by MAC, the insertion of pores in the membrane is not without consequence. Ions, particularly Ca^{2+}, flow into the cell and cause activation events with diverse outcomes that may contribute to disease.

Immune complexes with bound C3b are very efficient in priming B cells

Complement is a key component of the innate immune response. However, it has recently become apparent that complement also plays important roles in adaptive immunity and acts as a bridge between these two arms of immunity. This realization arose from studies in complement-depleted and complement-deficient mice in which antibody responses to foreign particles were markedly reduced. At least three linked mechanisms contribute to this effect (Fig. 4.15):

- first, immature B cells directly bind foreign particles through the B cell receptor (BCR) recognizing specific antigen in the particle, and through CR2 recognizing attached C3d – this co-ligation triggers B cell maturation, with the mature cells migrating to the lymphoid organs;
- second, while in the lymph nodes, mature B cells encounter opsonized antigen and, in the presence of B cell help, are induced to become activated and to proliferate;
- finally, in the lymphoid organs FDCs capture antigen through attached C3 fragments and use this bait to select the correct activated B cells and switch them on to further maturation and proliferation to form plasma cells and B memory cells (see Chapters 8 and 11).

The overriding principle of this ‘adjuvant’ effect of C3 opsonization is that simultaneous engagement of CR2 and BCR on the B cell, by recruiting signaling molecules to form an activation complex on the B cell surface, efficiently triggers the B cell response. As a consequence, complement-opsonized particles may be 1000-fold as active as the unopsonized particle in triggering antibody production.

Complement plays important roles in adaptive immunity

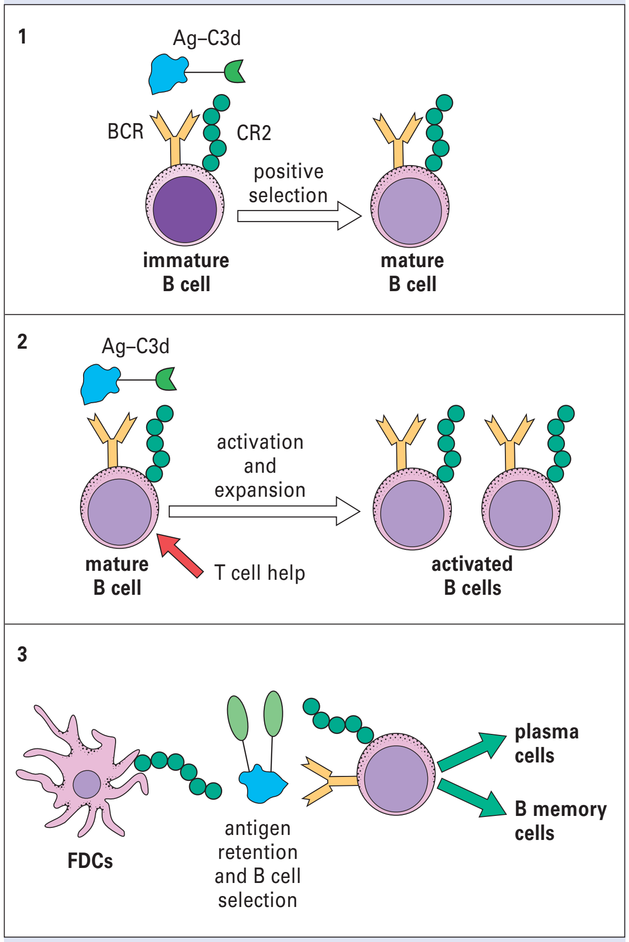
1
Ag–C3d
BCR
CR2
positive selection
immature B cell
mature B cell
2
Ag–C3d
activation and expansion
mature B cell
T cell help
activated B cells
3
FDCs
antigen retention and B cell selection
plasma cells
B memory cells

Fig. 4.15 C3 fragments bound to antigen (Ag) bind complement receptors on B cells and follicular dendritic cells (FDCs), enhancing B cell development at multiple stages in the process. (**1**) B cells bind Ag through the B cell receptor (BCR) and bind Ag-attached C3d through CR2. The combined signals, delivered through these receptors and their co-receptors, markedly enhance positive selection of Ag-reactive B cells and subsequent maturation. (**2**) Binding of C3d-opsonized Ag to mature B cells in the lymphoid follicles (with appropriate T cell help) triggers B cell activation and proliferation. (**3**) In the spleen and bone marrow, C3-opsonized Ag binds complement receptors on FDCs to retain Ag on the FDC, where it is efficiently presented to activated B cells. Ligation of BCR and C3d on the activated B cell triggers differentiation to plasma cells and B memory cells.

COMPLEMENT DEFICIENCIES ILLUSTRATE THE HOMEOSTATIC ROLES OF COMPLEMENT

Genetic deficiencies of each of the complement components and many of the regulators have been described and provide valuable 'experiments of nature' illustrating the homeostatic roles of complement. In general, complement deficiencies are rare, though some deficiencies are much more common in some racial groups.

A variety of assays (Method Box 4.1) are available for detecting:
- the activity of different complement pathways;
- the functional activity of individual components;
- the total amount of individual components (functional or non-functional).

The consequences of a deficiency in part of the complement system depend upon the pathway(s) affected (Fig. 4.16).

METHOD BOX 4.1

A variety of assays are available for detecting:
- the activity of different complement pathways;
- the functional activity of individual components;
- the total amount of individual components (functional or non-functional).

Measuring classical and alternative pathway activity

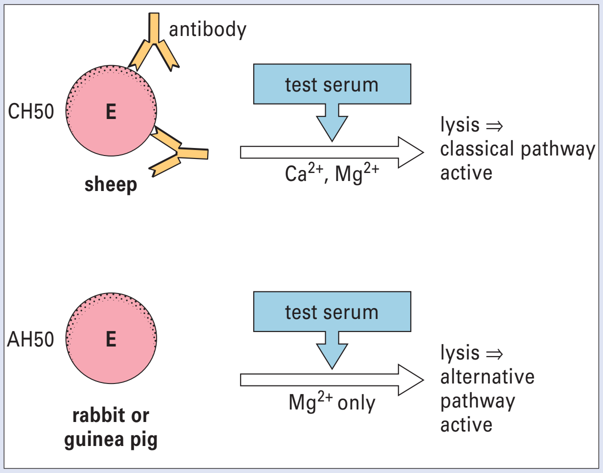
antibody
CH50
E
sheep
test serum
Ca^{2+}, Mg^{2+}
lysis ⇒ classical pathway active
AH50
E
rabbit or guinea pig
test serum
Mg^{2+} only
lysis ⇒ alternative pathway active

Fig. 1 Classical pathway activity is measured using antibody-sensitized sheep erythrocytes in a buffer containing both Ca^{2+} and Mg^{2+} ions. Alternative pathway activity is measured using rabbit erythrocytes in a buffer containing Mg^{2+} ions but no Ca^{2+} ions. In both assays, standards and samples diluted in the specific buffer are incubated with the relevant target cell and the amount of hemolysis is measured. The results are converted mathematically to standardized hemolytic units (CH50 for classical pathway, AH50 for alternative pathway).

Functional activity of individual components is measured in similar hemolytic assays by assessing the capacity of the test serum to restore hemolytic activity to sera depleted of individual components – for example, C6 activity would be tested using C6-depleted serum in a classical pathway assay.

The total amount of an individual component is measured using specific antisera in techniques such as nephelometry, radial immunodiffusion, and rocket immunoelectrophoresis.

Complement system deficiencies

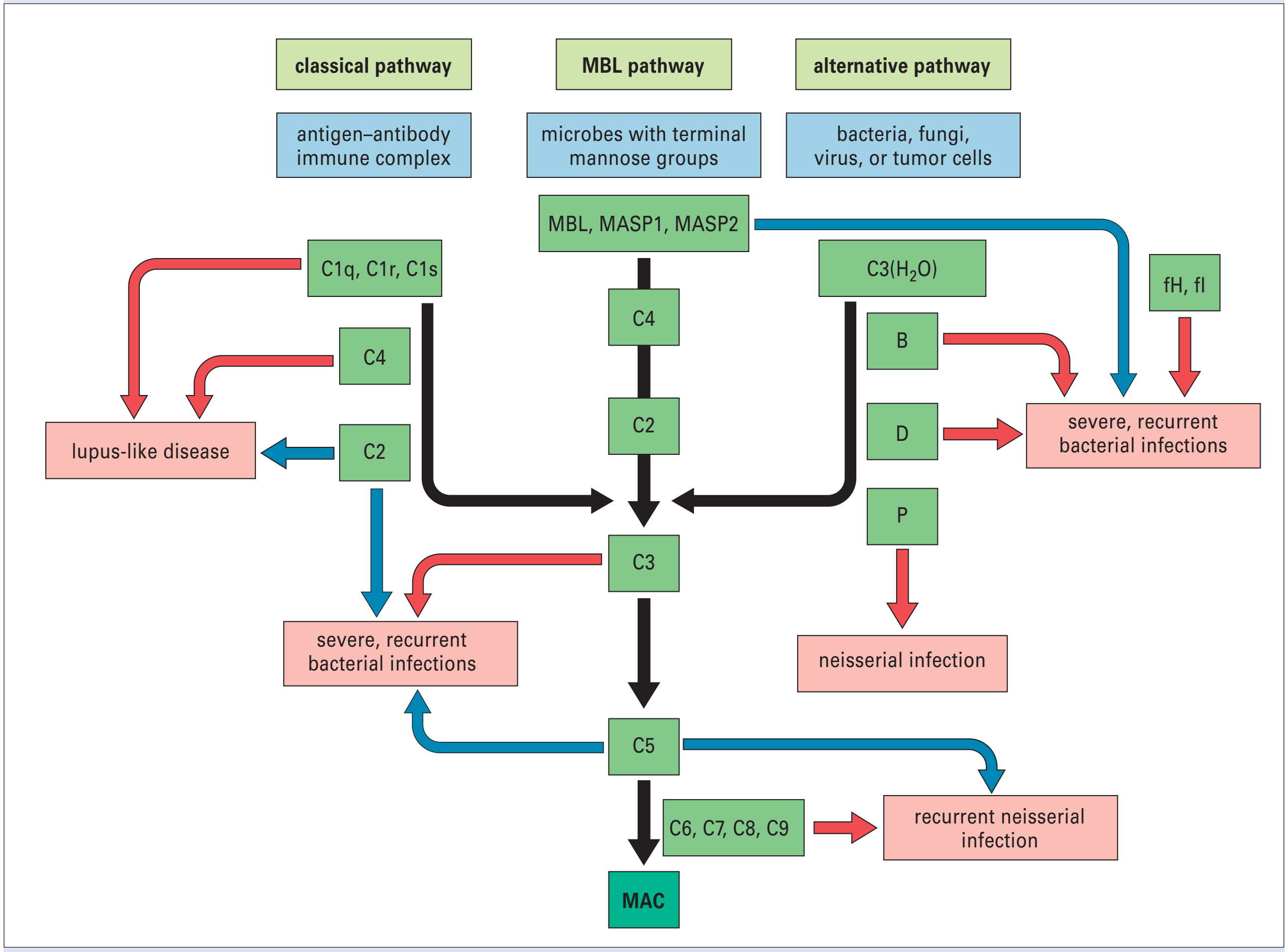
classical pathway
MBL pathway
alternative pathway
antigen–antibody immune complex
microbes with terminal mannose groups
bacteria, fungi, virus, or tumor cells
MBL, MASP1, MASP2
C1q, C1r, C1s
C3(H$_2$O)
fH, fI
C4
C4
B
lupus-like disease
C2
C2
D
severe, recurrent bacterial infections
P
C3
severe, recurrent bacterial infections
neisserial infection
C5
C6, C7, C8, C9
recurrent neisserial infection
MAC

Fig. 4.16 A summary of the clinical consequences of the various complement deficiencies. Black arrows denote pathway, red arrows show strong effects, and blue arrows indicate weak effects.

Classical pathway deficiencies result in tissue inflammation

Deficiency of any of the components of the classical pathway (C1, C4, and C2) predisposes to a condition that closely resembles the autoimmune disease systemic lupus erythematosus (SLE), in which immune complexes become deposited in the kidney, skin, and blood vessels.

Deficiency of any of the C1 subunits (C1q, C1r, or C1s) invariably causes severe disease with typical SLE features including skin lesions and kidney damage. The disease usually manifests early in childhood and few patients reach adulthood.

C4 deficiency also causes severe SLE. Total deficiency of C4 is extremely rare because C4 is encoded by two separate genes (*C4A* and *C4B*), but partial deficiencies of C4 are relatively common and are associated with an increased incidence of SLE.

C2 deficiency is the commonest complement deficiency in Caucasians. Although it predisposes to SLE, the majority of C2-deficient individuals are healthy.

The large majority of cases of SLE are, however, not associated with complement deficiencies, and autoimmune SLE is discussed below and on p. 376.

The historical view of immune complex disease in classical pathway deficiency was based upon defective immune complex handling.

Q. Why would a deficiency in the classical pathway lead to impaired handling of immune complexes?

A. Classical pathway activation and opsonization of immune complexes helps prevent precipitation in tissues and aids the carriage of immune complexes on erythrocytes. Classical pathway deficiencies would therefore result in a failure to maintain solubilization and permit the resultant precipitation of immune complexes in the tissues where they drive inflammation.

Although these mechanisms of immune complex handling undoubtedly contribute, a new perspective has recently developed that takes a different view of the role of complement in waste management.

Cells continually die by apoptosis in tissues and are removed silently by tissue macrophages. Complement contributes to this essential process because the apoptotic cell binds C1q and activates the classical pathway. In C1 deficiencies, apoptotic cells accumulate in the tissues and eventually undergo necrosis, which releases toxic cell contents and causes inflammation.

This recent observation, emerging from studies in complement deficiencies, has altered the way we think of the handling of waste in the body and moved complement to center-stage in this vital housekeeping role.

Deficiencies of MBL are associated with infection in infants

MBL is a complex multi-chain **collectin**. Each chain comprises a collagenous stalk linked to a globular carbohydrate recognition domain.

The plasma level of MBL is extremely variable in the population, and governed by a series of single nucleotide polymorphisms in the *MBL* gene, either in the promoter region or in the first exon, encoding part of the collagenous stalk:

- mutations in the promoter region alter the efficiency of gene transcription;
- mutations in the first exon disrupt the regular structure of the collagenous stalk, destabilizing complexes containing mutated chains, perhaps disrupting association with the MASP enzymes.

At least seven distinct haplotypes arise from mixing of these mutations, four of which yield very low plasma MBL levels. As a consequence, at least 10% of the population have MBL levels below 0.1 μg/ml and are considered to be MBL deficient.

MBL deficiency is associated in infants with increased susceptibility to bacterial infections. This tendency disappears as the individual ages and the other arms of immunity mature.

In adults, MBL deficiency is of little consequence unless there is an accompanying immunosuppression – for example, people with HIV infection who are MBL deficient appear to have more infections than those who have high levels of MBL.

Alternative pathway and C3 deficiencies are associated with bacterial infections

Deficiencies of either fB or fD prevent complement amplification through the alternative pathway amplification loop, markedly reducing the efficiency of opsonization of pathogens. As a consequence, deficient individuals are susceptible to bacterial infections and present with a history of severe recurrent infections with a variety of pyogenic (pus-forming) bacteria. Only a few families with each of these deficiencies has been identified, but the severity of the condition makes it imperative to identify affected families so that prophylactic antibiotic therapy can be initiated.

C3 is the cornerstone of complement, essential for all activation pathways and for MAC assembly, and is also the source of the major opsonic fragments C3b and iC3b. Individuals with **C3 deficiency** present early in childhood with a history of severe, recurrent bacterial infections affecting the respiratory system, gut, skin, and other organs. Untreated, all die before adulthood. When given broad-spectrum antibiotic prophylaxis, patients do reasonably well and survival into adulthood becomes the norm.

Terminal pathway deficiencies predispose to Gram-negative bacterial infections

Deficiencies of any of the terminal complement components (C5, C6, C7, C8, or C9) predisposes to infections with Gram-negative bacteria, particularly those of the genus *Neisseria*. This genus includes the meningococci responsible for meningococcal meningitis and the gonococci responsible for gonorrhea.

Q. Why should these deficiencies be specifically associated with infection by Gram-negative bacteria and not with all bacterial infections?

A. Gram-negative bacteria have an outer phospholipid membrane, which may be targeted by the lytic pathway. Gram-positive bacteria have a thick bacterial cell wall on the outside.

Individuals with terminal pathway deficiencies usually present with meningitis, which is often recurrent and often accompanied by septicemia. Any patients with second or third episodes of meningococcal infection without obvious cause should be screened for complement deficiencies because prophylactic antibiotic therapy can be life-saving. Terminal pathway-deficient patients should also be intensively immunized with the best available meningococcal vaccines.

It is likely that terminal pathway deficiencies are relatively common and underascertained in most countries:

- in Caucasians and most other groups investigated, C6 deficiency is the most common;
- in the Japanese population, C9 deficiency is very common, with an incidence of more than 1 in 500 of the population.

C1 inhibitor deficiency leads to hereditary angioedema

Deficiency of the classical pathway regulator C1inh is responsible for the syndrome **hereditary angioedema (HAE)**. C1inh regulates C1 and MBL in the complement system and also controls activation in the kinin pathway that leads to the generation of bradykinin and other active kinins (Fig. 4.17).

HAE is relatively common because the disease presents even in those heterozygous for the deficiency (i.e. it is an autosomal dominant disease).

The halved C1inh synthetic capacity in those with HAE cannot maintain plasma levels in the face of continuing consumption of C1inh, which is a **suicide inhibitor** that is consumed as it works. As a consequence, the plasma levels measured are often only 10–20% of normal, even in periods of health.

Episodes of angioedema are often triggered in the skin or mucous membranes by minor trauma – occasionally stress may be sufficient to induce an attack. Swelling, which may be remarkable in severity, rapidly ensues as

Pathogenesis of hereditary angioedema

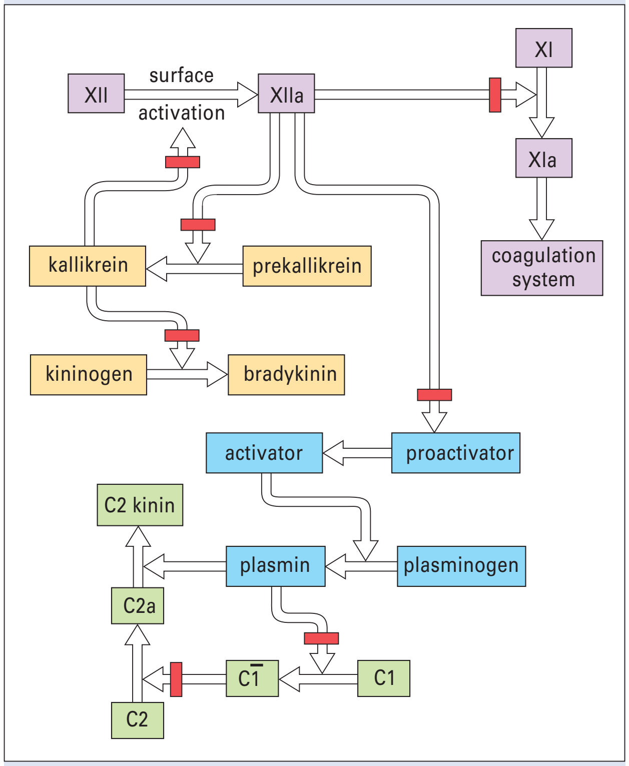

Fig. 4.17 C1 inhibitor (C1inh) is involved in the inactivation of elements of the kinin, plasmin, complement, and clotting systems, all of which may be activated following the surface-dependent activation of factor XII. The points at which C1inh acts are shown in red. Uncontrolled activation of these pathways results in the formation of bradykinin and C2 kinin, which induce edema.

Hereditary angioedema

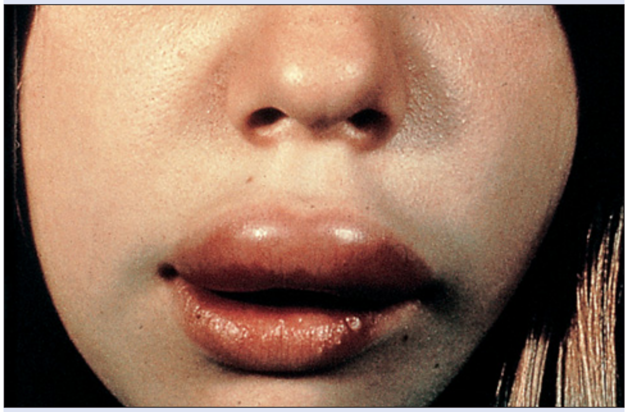

Fig. 4.18 The clinical appearance of hereditary angioedema, showing the local transient swelling that affects mucous membranes.

unregulated activation of the kinin and complement systems occurs in the affected area, inducing vascular leakiness. Swelling of mucous membranes in the mouth and throat may block the airways, leading to asphyxia (Fig. 4.18).

Episodes of angioedema are transient and usually wane over the course of a few hours without therapy. Emergency treatment for life-threatening attacks involves the infusion of a purified C1inh concentrate. Prophylactic treatment involves the induction of C1inh synthesis using anabolic steroids, or minimizing consumption of C1inh using protease inhibitors.

Although the majority of cases of HAE involve a mutation that prevents synthesis of C1inh by the defective gene (type I), in about 15% of cases the mutation results in the production of a functionally defective protein (type II). In type II HAE, the plasma levels of C1inh may be normal or even high, but its function is markedly impaired, leading to disease.

A similar syndrome to type II HAE can develop later in life. Acquired angioedema is, in most or all cases, associated with autoantibodies that:

- target C1inh;
- may arise in an otherwise healthy individual or may be associated with other autoimmune diseases, particularly SLE.

Occasionally, acquired angioedema occurs in association with some lymphoproliferative disorders – whether autoantibodies are involved in these cases remains unclear.

Deficiencies in alternative pathway regulators produce a secondary loss of C3

fH or fI deficiency predisposes to bacterial infections

fH and fI collaborate to control activation of the alternative pathway amplification loop. Deficiency of either leads to uncontrolled activation of the loop and complete consumption of C3, which is the substrate of the loop. The resultant **acquired C3 deficiency** predisposes to bacterial infections and yields a clinical picture identical to that seen in primary C3 deficiency.

Properdin deficiency causes severe meningococcal meningitis

Properdin is a stabilizer of the alternative pathway C3 convertase that increases efficiency of the amplification loop. **Properdin deficiency** is inherited in an X-linked manner and is therefore seen exclusively in males. Boys deficient in properdin present with severe meningococcal meningitis, often with septicemia. The first attack is often fatal and survivors do not usually have recurrent infections because the acquisition of anti-meningococcal antibody enables a response via the classical pathway in the next encounter. Diagnosis is nevertheless important to identify affected relatives before they get disease – administration of meningococcal vaccine and antibiotic prophylaxis will prevent infection.

FURTHER READING

Barrington R, Zhang M, Fischer M, Carroll MC. The role of complement in inflammation and adaptive immunity. Immunol Rev 2001;180:5–15.

Campbell RD, Law SKA, Reid KBM, et al. Structure, organization, and regulation of the complement genes. Annu Rev Immunol 1988;6:161–195.

Carroll MC. The complement system in B cell regulation. Mol Immunol 2004;41:141–146.

Cole DS, Morgan BP. Beyond lysis: how complement influences cell fate. Clin Sci 2003;104:455–466.

Colten HR, Rosen FS. Complement deficiencies. Annu Rev Immunol 1992;10:809–834.

Davis AE 3rd. The pathogenesis of hereditary angioedema. Transfus Apher Sci 2003;39:195–203.

Dodds AW, Ren XD, Willis AC, et al. The reaction mechanism of the internal thioester in the human complement component C4. Nature 1996;379:177–179.

Frank MM, Fries LF. The role of complement in inflammation and phagocytosis. Immunol Today 1991;12:322–326.

Gerard C, Gerard NP. C5a anaphylatoxin and its seven transmembrane-segment receptor. Annu Rev Immunol 1994;12:775–808.

Holmskov U, Malhotra R, Sim RB, et al. Collectins: collagenous C-type lectins of the innate immune defense system. Immunol Today 1994;14:67–74.

Hourcade D, Holers VM, Atkinson JP. The regulators of complement activation (RCA) gene cluster. Adv Immunol 1989;45:381–416.

Jack DL, Klein NJ, Turner MW. Mannose-binding lectin: targeting the microbial world for complement attack and opsonophagocytosis. Immunol Rev 2001;180:86–99.

Lambris JD, Reid KBM, Volanakis JE. The evolution, structure, biology and pathophysiology of complement. Immunol Today 1999;20:207–211.

Liszewski MK, Farries TC, Lublin DM, et al. Control of the complement system. Adv Immunol 1996;61:201–283.

Manderson AP, Botto M, Walport MJ. The role of complement in the development of systemic lupus erythematosus. Annu Rev Immunol 2004;22:432–456.

Moffitt MC, Frank MM. Complement resistance in microbes. Springer Semin Immunopathol 1994;15:327–344.

Morgan BP. Complement regulatory molecules: application to therapy and transplantation. Immunol Today 1995;16:257–259.

Morgan BP. Complement in inflammation. In: Rey K, ed. Physiology of inflammation. Oxford: Oxford University Press; 2001:131–145.

Morgan BP, Harris CL. Complement therapeutics; history and current progress. Mol Immunol 2003;40:159–170.

Morgan BP, Meri S. Membrane proteins that protect against complement lysis. Springer Semin Immunopathol 1994;15:369–396.

Morgan BP, Walport MJ. Complement deficiency and disease. Immunol Today 1991;12:301–306.

Muller-Eberhard HJ. The membrane attack complex of complement. Annu Rev Immunol 1986;4:503–528.

Nonaka M, Yoshizaki F. Evolution of the complement system. Mol Immunol 2004;40:879–902.

Reid KB, Turner MW. Mammalian lectins in activation and clearance mechanisms involving the complement system. Springer Semin Immunopathol 1994;15:307–326.

Walport MJ. Complement: first of two parts. N Engl J Med 2001;344:1058–1066.

Walport MJ. Complement: second of two parts. N Engl J Med 2001;344:1140–1144.

Critical thinking: Complement deficiency (see p. 494 for explanations)

A family has been identified in which three of the seven children have had repeated upper respiratory tract infections since early childhood. Of these, one has developed bacterial meningitis, and another a fatal septicemia. In all of the children the levels of antibodies in the serum are within the normal range. When an assay for hemolytic complement (CH50) is carried out, however, the three affected children are all found to be deficient in this functional assay.

1 Why would a deficiency in complement cause the children to be particularly susceptible to bacterial infections?

Measurements are made of individual complement components of the classical and alternative pathways to determine which of the components is defective. The results are shown in the table.

complement component	normal concentration (μg/ml)	levels in affected children (μg/ml)
C4	600	480–520
C2	20	15–22
C3	1300	10–80
factor B (fB)	210	not detectable
factor H (fH)	480	200–350
factor I (fI)	35	not detectable

2 Using knowledge of the complement reaction pathways, how can you explain the apparent combined deficiencies in C3, fB, and fI?

3 What is the fundamental deficiency in this family and how would you treat the affected children?

CHAPTER

5

T Cell Receptors and MHC Molecules

SUMMARY

- **The T cell antigen receptor (TCR) is located on the surface of T cells and plays a critical role in the adaptive immune system.** Its major function is to recognize antigen and transmit a signal to the interior of the T cell, which generally results in activation of T cell responses.

- **TCRs are similar in many ways to immunoglobulin molecules.** Both are made up of pairs of subunits (α and β or γ and δ), which are themselves members of the immunoglobulin superfamily, and both recognize a wide variety of antigens via N terminal variable regions. Both the αβ TCR and the γδ TCR are associated with CD3, forming TCR complexes.

- **The two types of TCR may have distinct functions.** In humans and mice, the αβ TCR predominates in most peripheral lymphoid tissues, whereas cells bearing the γδ TCR are enriched at mucosal surfaces. Critical signaling functions are performed by the invariant chains of the TCR, the CD3 complex.

- **Like immunoglobulins, TCRs are encoded by several sets of genes,** and a large repertoire of TCR antigen-binding sites is generated by V(D)J recombination during T cell differentiation. Unlike immunoglobulins, TCRs are never secreted and do not undergo class switching or somatic hypermutation.

- **Recognition by the αβ TCR requires the antigen to be bound to a specialized antigen-presenting structure known as a major histocompatibility complex (MHC) molecule.** Unlike immunoglobulins, TCRs recognize antigen only in the context of a cell–cell interaction.

- **Class I and class II MHC molecules bind to peptides derived from different sources.** Class I MHC molecules bind to peptides derived from cytosolic (intracellular) proteins, known as endogenous antigens. Class II MHC molecules bind to peptides derived from extracellular proteins that have been brought into the cell by phagocytosis or endocytosis (exogenous antigens).

- **Class I and class II MHC present peptide antigens to the TCR in a cell–cell interaction** between an antigen-presenting cell (APC) and a T cell.

- **HLA-A, HLA-B, and HLA-C gene loci encode class I MHC molecules**

- **HLA-DP, HLA-DQ, AND HLA-DR gene loci encode class II MHC molecules**.

- **An individual's MHC haplotype affects susceptibility to disease**.

- **CD1 is an MHC class 1-like molecule that presents lipid antigens**.

THE TCR PLAYS A CRITICAL ROLE IN THE ADAPTIVE IMMUNE SYSTEM

As discussed in Chapter 1, the immune system of higher vertebrates can be divided into two components – humoral immunity and cell-mediated immunity.

Humoral immunity, of which antibodies are a key component, provides protection via the extracellular fluids. Antibodies deal quite effectively with extracellular pathogens:

- targeting them for phagocytosis or complement-mediated lysis;
- neutralizing receptors on the surface of bacteria and viruses; and
- inactivating circulating toxins.

If antibodies were our only defense, however, pathogens could escape immune surveillance simply by hiding within cells. In fact, many pathogens – all viruses, some bacteria, and certain parasites – do just that, carrying out substantial portions of their life cycles within host cells. Remarkably, some bacteria even thrive within macrophages after being phagocytosed. These considerations highlight the need for a second arm of the immune response – **cell-mediated immunity** – of which T cells are critical operatives.

T cells recognize antigen via specialized cell surface antigen receptors – T cell receptors (TCRs) – which are structurally and evolutionarily related to antibodies.

TCRs recognize antigen via variable regions generated through **V(D)J recombination** (see Chapter 3), much like immunoglobulins, but are much more restrictive in their antigen recognition capabilities.

TCRs recognize peptides displayed by MHC molecules

T cells generally recognize fragments of degraded proteins (peptides), which must be bound to ('presented by') specialized antigen-presenting molecules termed **major histocompatibility complex (MHC) molecules**.

MHC molecules can display a wide range of peptides derived from intracellular proteins on the cell surface, thereby alerting the immune system to the presence of intracellular invaders.

Because MHC molecules are expressed only on the surface of antigen-presenting cells (APCs), engagement of the TCR occurs only in the context of intimate cell–cell interactions.

When a primed T cell bearing an appropriate TCR (capable of recognizing, for example, a particular viral peptide) comes into contact with an infected cell, it can rapidly kill that cell and thereby limit the spread of the viral infection.

The requirement for cell–cell interaction also allows T cells to provide critical regulatory functions. T cells, for example, can couple signals from the innate immune system, such as those arising from 'professional' APCs (e.g. dendritic cells) with adaptive (T and B cell) responses, providing a critical layer of integration, coordination, and regulation.

TCRs ARE SIMILAR TO IMMUNOGLOBULIN MOLECULES

The TCR was identified much later than immunoglobulin, even though early theoretical considerations strongly suggested that T cells must bear cell surface antigen receptors. It is now clear that there are two varieties of TCR, termed αβ and γδ, and that both molecules:

- resemble immunoglobulins in several significant ways (Fig. 5.1);
- are made up of heterodimers (either α and β or γ and δ subunits), which are disulfide-linked;
- are integral membrane proteins with large extracellular domains and short cytoplasmic tails.

The extracellular portions are responsible for antigen recognition and contain variable N terminal regions, like antibodies. Both α and β (or γ and δ) subunits contribute to the antigen-binding sites.

THE TWO TYPES OF TCR MAY HAVE DISTINCT FUNCTIONS

The two types of TCR tend to populate different tissue sites and are thought to perform distinct functions.

The αβ heterodimer is the antigen recognition unit of the αβ TCR

The αβ TCR is the predominant receptor found in the thymus and peripheral lymphoid organs of mice and humans. It is a disulfide-linked heterodimer of α (40–50 kDa) and β (35–47 kDa) subunits and its structural features have been determined by X-ray crystallography (Fig. 5.2).

Each polypeptide chain of the αβ TCR contains two extracellular immunoglobulin-like domains of approximately 110 amino acids, anchored into the plasma membrane by a transmembrane domain that has a short cytoplasmic tail.

Q. How can receptors that lack intracytoplasmic domains signal to the cell? Give some examples.

A. They signal by associating with other membrane molecules that do have intracytoplasmic domains. For example, immunoglobulin associates with Igα and Igβ (see Fig. 3.1), and FcγRI associates with its γ chain dimer (see Fig. 3.20).

The extracellular portions of the α and β chains fold into a structure that resembles the antigen-binding portion (Fab) of an antibody (see Fig. 3.12). Indeed, as in antibodies, the amino acid sequence variability of the TCR resides in the N terminal domains of the α and β (and also the γ and δ) chains.

The regions of greatest variability correspond to immunoglobulin hypervariable regions and are also known as **complementarity determining regions**

Similarities and differences between T cell receptors and immunoglobulins

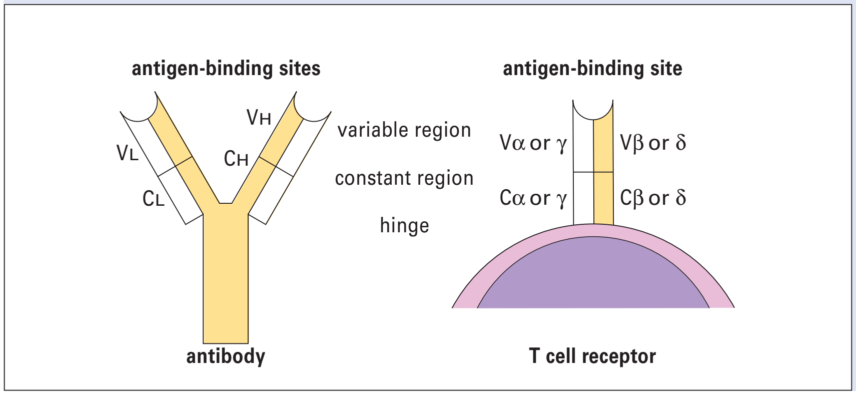

Fig. 5.1 TCRs are very similar to Fab fragments of B cell receptors. Both receptor types are composed of two different peptide chains and have variable regions for binding antigen, constant regions, and hinge regions. The principal differences are that TCRs remain membrane-bound and contain only a single antigen-binding site.

The T cell antigen receptor

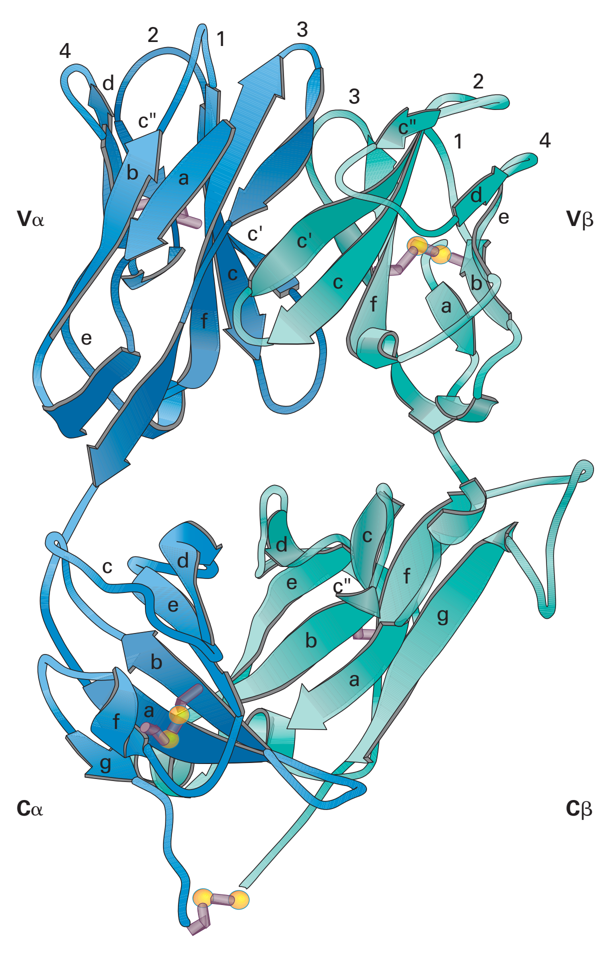

Fig. 5.2 Three-dimensional structure of an αβ TCR – only extracellular domains are shown. The α chain is colored blue (residues 1–213), and the β chain is colored green (residues 3–247). The β strands are represented as arrows and labeled according to the standard convention used for the immunoglobulin fold. The disulfide bonds (yellow balls for sulfur atoms) are shown within each domain and for the C terminal interchain disulfide. The complementarity determining regions (CDRs) lying at the top of the diagram are numerically labeled (1–4) for each chain. These form the binding site for antigen/MHC molecule. (Adapted from Garcia KC, Degano M, Stanfield RL, et al. Science 1996;274:209–219. Copyright AAAS)

Rapid Reference Box 5

CDRs (complementarity determining regions) – the sections of an antibody or T cell receptor V region responsible for antigen or antigen–MHC molecule binding.
H-2 – the mouse MHC.
histocompatibility – the ability to accept grafts between individuals.
HLA – the human MHC.
ITAMs (immunoreceptor tyrosine activation motifs) and ITIMs (immunoreceptor tyrosine inhibitory motifs) – target sequences for phosphorylation by kinases involved in cell activation or inhibition.
MHC (major histocompatibility complex) – a genetic region found in all mammals whose products are primarily responsible for the rapid rejection of grafts between individuals; it functions in the signaling between lymphocytes and cells expressing antigen.
MHC class I and II molecules – molecules coded within the MHC – class I molecules have one MHC-encoded peptide complexed with β_2-microglobulin; class II molecules have two MHC-encoded peptides, which are non-covalently associated.
MHC restriction – a characteristic of many immune reactions in which cells cooperate most effectively with other cells that share an MHC haplotype.
RAG-1 and RAG-2 – recombination activating genes, required for recombination of V, D, and J gene segments during the generation of functional antigen receptor genes.
V, D, and J genes – variable (V), joining (J), and diversity (D) genes.

(CDRs). They are clustered together to form an antigen-binding site analogous to the corresponding site on antibodies (see Fig. 5.2). Note, however, that:

- the CDR3 loops from both the α and β chains lie at the center of the antigen-binding site.

These CDR3 loops make contact with antigen, which in the case of the TCR, is a peptide (see below).

The disulfide bond that links the α and β chains is in a peptide sequence located between the constant domain of the extracellular portion of the receptor and the transmembrane domain (shown as the C terminal residue in the α and β chain in Fig. 5.2).

One remarkable feature of the transmembrane portion of the receptor is the presence of positively charged residues in both the α and β chains. Unpaired charges would be unfavorable in a transmembrane segment. Indeed, these positive charges are neutralized by assembly of the complete TCR complex, which contains additional polypeptides bearing complementary negative charges (see below).

The CD3 complex associates with the antigen-binding αβ or γδ heterodimers to form the complete TCR

The αβ or γδ heterodimers must associate with a series of polypeptide chains collectively termed the **CD3 complex** for the antigen-binding domains of the TCR to form a complete, functional receptor that is stably expressed at the cell surface and is capable of transmitting a signal upon binding to antigen.

The four members of the CD3 complex (γ, δ, ε, and ζ) are sometimes termed the **invariant chains** of the TCR because they do not show variability in their amino acid sequences. (The γ and δ chains of the CD3 complex should not be confused with the quite distinct antigen-binding variable chains of the TCR that bear the same names.)

The CD3 chains are assembled as heterodimers of γε and δε subunits with a homodimer of ζ chains, giving an overall TCR stoichiometry of $(\alpha\beta)_2$, γ, δ, ε_2, ζ_2. Current data suggest that the TCR complex exists as a dimer (Fig. 5.3).

The CD3 γ, δ, and ε chains are the products of three closely linked genes, and similarities in their amino acid sequences suggest that they are evolutionarily related.

Indeed, all three are members of the immunoglobulin superfamily, each containing an external domain followed by a transmembrane region and a substantial, highly conserved, cytopasmic tail of 40 or more amino acids.

As with the transmembrane domains of the variable chains of the TCR, the membrane-spanning regions of these CD3 chains contain charged amino acids.

The T cell receptor complex

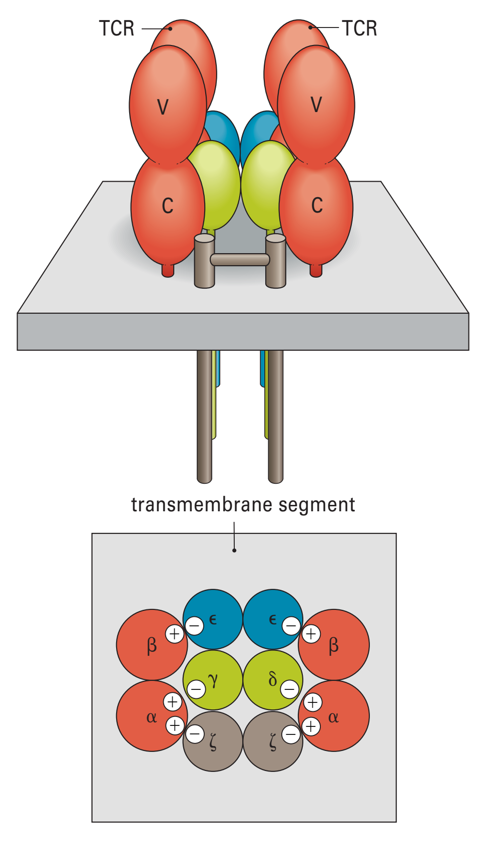

Fig. 5.3 The TCR α and β (or γ and δ) chains each comprise an external V and C domain, a transmembrane segment containing positively charged amino acids, and a short cytoplasmic tail. The two chains are disulfide linked on the membrane side of their C domains. The CD3 γ, δ, and ε chains comprise an external immunoglobulin-like C domain, a transmembrane segment containing a negatively charged amino acid, and a longer cytoplasmic tail. A dimer of ζζ, ηη, or ζη is also associated with the complex. Several lines of evidence support the notion that the TCR–CD3 complex exists at the cell surface as a dimer. The transmembrane charges are important for the assembly of the complex. A plausible arrangement that neutralizes opposite charges is shown.

It is thought that the negatively charged residues in the transmembrane region of the CD3 chains interact with (and neutralize) the positively charged amino acids in the αβ polypeptides, leading to the formation of a stable TCR complex (Fig. 5.3).

The CD3 ζ gene is on a different chromosome from the CD3 γδε gene complex, and the ζ protein is structurally unrelated to the other CD3 components. The ζ chains possess:

- a small extracellular domain (nine amino acids);
- a transmembrane domain carrying a negative charge; and
- a large cytoplasmic tail.

An alternatively spliced form of CD3ζ, called CD3η, possesses an even larger cytoplasmic tail (42 amino acids longer at the C terminus).

The cytoplasmic portions of ζ and η chains contain ITAMs

These ζ and η chains may associate in all three possible combinations (ζζ, ζη, or ηη) and play a critical role in signal transduction through the TCR. The cytoplasmic portions of these subunits contain particular amino acid sequences called **immunoreceptor tyrosine-based activation motifs (ITAMs)**, and each chain contains three of these motifs.

Q. Which other group of cell surface molecules contains ITAMs?

A. The Fcγ receptors, either as an intrinsic intracellular domain of the receptor, or because they associate with signaling molecules that have ITAMs (see Fig 3.20).

The conserved tyrosine residues in the ITAM motifs are targets for phosphorylation by specific protein kinases. When the TCR is bound to its cognate antigen–MHC complex, the ITAM motifs become phosphorylated within minutes in one of the first steps in T cell activation (see Fig. 7.21).

ITAMs:

- are essential for T cell activation, and mutational substitution of the tyrosines in the motif prevents activation;
- play critical roles in B cell activation, and are present in the B cell receptor chains, Igα and Igβ (see Fig. 3.1 and Chapter 8).

CD3ζ also functions in another signaling pathway, associating with the low-affinity FcγRIIIa receptor (CD16), which is involved in the activation of macrophages and natural killer (NK) cells (see Fig. 3.20).

Other subunits of the CD3 complex (γ, δ, ε), though lacking in ITAMs, may also become phosphorylated following TCR engagement. Phosphorylation of the CD3γ chain downregulates TCR expression on the cell surface via a mechanism involving increased receptor internalization.

The γδ TCR structurally resembles the αβ TCR but may function differently

The overall structure of the γδ TCR is similar to that of its αβ counterpart. Each chain is organized into:

- extracellular V and C domains;

- a transmembrane segment containing positively charged amino acids; and
- a short cytoplasmic tail.

One indication that the two types of T cell (i.e. T cells with αβ TCRs and T cells with γδ TCRs) might perform different functions comes from their anatomic distribution:

- in humans and mice, αβ TCRs are present on more than 95% of peripheral blood T cells and on the majority of thymocytes;
- T cells bearing γδ TCRs are relatively rare in spleen, lymph nodes, and peripheral blood but predominate at epithelial surfaces – they are common in skin and in the epithelial linings of the reproductive tract and are especially numerous in the intestine, where they are found as **intraepithelial lymphocytes**.

It is further believed that there are distinct subsets of γδ T cells that can perform different functions.

Antigen recognition by γδ T cells is unlike that of their αβ counterparts

The fact that γδ T cells are rare in anatomic locations known to support the classical mechanisms of antigen presentation and lymphocyte clonal expansion suggests the possibility that γδ cells might not need to rely upon normal antigen presentation mechanisms for their activation.

Several lines of evidence support the hypothesis that γδ T cells can recognize antigen in an MHC-independent fashion, for example:

- γδ T cells can be found in normal numbers in MHC class I and class II-deficient mice;
- their cognate antigens are not necessarily peptides, and do not require classical processing – indeed, some murine γδ T cells have been found to recognize proteins directly, including MHC molecules and viral proteins, in a manner that requires neither antigen processing nor presentation by MHC.

γδ T cells therefore appear to be able to follow a different paradigm for T cell recognition of antigen than that employed by αβ T cells.

γδ T cells recognize at least two classes of ligand:

- molecules that signal the presence of cellular stress; and
- small organic molecules that serve as signifiers of infection.

For example, human intraepithelial γδ T cells have been found to respond to MHC class I-related antigens (MICA and MICB) expressed on the surface of stressed cells.

In addition, some human γδ T cells recognize small, non-peptidic, organic compounds secreted by mycobacteria, such as monoethylphosphate and isopentenyl pyrophosphate. These ligands are secreted by a number of bacteria and may also be produced by some eukaryotic pathogens.

The γδ T cell arm of the adaptive immune system therefore appears to share some key characteristics of innate immune responses.

γδ T CELLS HAVE A VARIETY OF BIOLOGICAL ROLES – γδ T cells:

- are essential for primary immune responses to certain viral and bacterial pathogens in mouse models, but in many cases their contribution to the primary response can be substituted for by αβ T cells, and they rarely contribute to memory responses;
- interact with a variety of lymphocytes, and have been implicated in stimulating immunoglobulin class switch recombination by B cells in response to T-dependent antigens;
- provide regulatory signals to αβ T cells and have been implicated in shaping immune responses (e.g. γδ cells appear to be involved in downregulating inflammation and in this role may be responding to epithelial cells stressed by inflammatory processes rather than to specific antigens borne by pathogens).

The unique ability of γδ T cells to, on the one hand, sense tissue damage and, on the other, to recognize antigens without the normal constraints of antigen processing/MHC restriction (see below), may allow them to fill several key biological roles such as immunoregulation. In particular, γδ T cells may downregulate potentially damaging inflammatory responses, providing immunoprotection:

- when MHC function is compromised by, for example, viral infections that downregulate MHC;
- in early life when αβ T cell function is immature and when the antigen processing and antigen sampling systems have not yet matured.

TCRs ARE ENCODED BY SEVERAL SETS OF GENES

The general arrangement of the genes encoding the α, β, γ, and δ chains of the TCR is remarkably similar to that of the immunoglobulin heavy chain genes (see Chapter 3, p. 81), suggesting a common origin from a primordial rearranging antigen receptor locus.

Fig. 5.4 illustrates the murine TCR genes, which are similar to those of humans. All four TCR gene families have been strongly conserved across more than 400 million years of evolution of the jawed vertebrates, which suggests a strong selective pressure for the preservation of both αβ and γδ T cell functions.

The nomenclature of the TCR genes is simple – the *TCRA* locus encodes the α gene, *TCRB* the β gene, and so on. Interestingly, the *TCRD* locus is nested within the *TCRA* cluster.

The α and γ loci have sets of V and J gene segments (analogous to immunoglobulin light chain loci), whereas the β and δ loci have V, D, and J gene segments (analogous to immunoglobulin heavy chains).

TCR variable region gene diversity is generated by V(D)J recombination

As with antibody genes, a highly diverse repertoire of TCR variable region genes is generated during T cell differentiation by a process of somatic gene rearrangement termed V(D)J recombination (see Chapter 3). Variable (V), joining (J), and sometimes diversity (D) gene segments are joined together to form a completed variable region gene.

Junctional diversity (imprecise joining of V, D, and J with loss and/or addition of nucleotides) contributes an enormous amount of variability to the TCR repertoire in

Murine T cell receptor genes

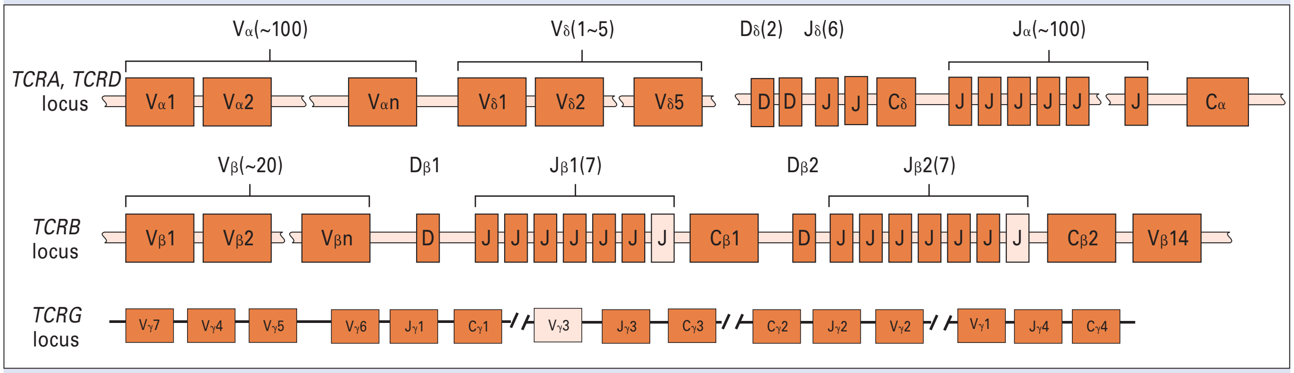

Fig. 5.4 The δ chain loci are embedded within the α loci and tandem duplication has occurred in the β chain loci. The last of each set of Jβ genes and the Vγ3 gene are pseudogenes.

addition to the variation that results from combinatorial assortment of the various gene segments. There is minor variation in detail for each locus (see Fig. 5.4).

TCRA recombination entails joining of V to J gene segments

As in the kappa light chain (IgK locus), joining of a Vα segment to a Jα segment produces a complete variable region gene. The large number of Jα segments contributes to the ultimate diversity of TCRα specificities.

*The **TCRB** locus includes two sets of D, J, and C genes*

Most of the Vβ genes are grouped together, but one (Vβ14) is present at the extreme 3′ end of the locus. The tandem duplication of Dβ, Jβ, and Cβ must have occurred early in mammalian evolution because it is present in both mice and humans. Extensive diversity is generated by recombination because not only are VDJ arrangements possible, but also VJ and VDDJ combinations.

The D segments are used in all three reading frames, further contributing to β chain diversity.

*The arrangement of the **TCRG** locus differs in mice and in humans*

The murine γ locus bears a striking similarity to the antibody light chain locus, with four Cγ genes (including a pseudogene), each associated with one J gene and from one to four Vγ genes. There are no D genes.

In humans there are eight Vγ genes, followed by three Jγ and the first Cγ, then two additional Jγ genes before Cγ2.

*The **TCRD** locus possesses only five Vδ, two Dδ, and six Jδ genes*

The *TCRD* locus was discovered during studies on the *TCRA* locus and possesses only five Vδ, two Dδ, and six Jδ genes. Despite this relative paucity of genetic material, more than 1000 different δ chains can be generated.

Q. How may the great number of δ chains be generated when there is only a limited number of V, D, and J gene segments in the *TCRD* locus?

A. As described in Chapter 3, the two D region segments can be used in all three possible reading frames, and imprecise joining of the VD and DJ junctions produces further diversity.

The mechanism of V(D)J recombination is the same in both T cells and B cells

The TCR genes are flanked by recombination signal sequences, just like their immunoglobulin cousins (see Fig. 3.29), and the same recombination machinery (the RAG proteins) operates in both B and T cells. Indeed, experiments have shown that TCR Dβ and Jγ genes can rearrange appropriately even if transfected into B cells.

Analysis of the amino acid sequences of many different TCRs shows that the greatest diversity lies within the third CDR (CDR3), which is also the case for B cell receptors. Addition of N regions (non-templated nucleotides added to the junctions by terminal deoxynucleotidyl transferase, TdT) is much more pronounced in TCRs, however. It is important to note, too, that neither somatic hypermutation nor class switching occur in T cells.

Q. Why is it that, unlike B cells, T cells have not evolved a class switching mechanism?

A. Class switching is irrelevant because there is no secreted form of the TCR and hence no interaction analogous to that of immunoglobulin and FcR.

RECOMBINATION YIELDS GREAT DIVERSITY – Hunkapiller and Hood have calculated that it is possible to construct about:

- 4.4×10^{13} different forms of TCR Vβ; and
- 8.5×10^{12} forms of TCR Vα.

They estimate that if only 1% of the sequences coded for viable proteins this would still give 2.9×10^{22} receptors. Even if 99% of these viable receptors were rejected due to autoreactivity or other defects, recombination would still yield 2.9×10^{20} possible murine TCRs. This would seem to be more than enough potential diversity, given that the thymus produces fewer than 10^9 thymocytes over the lifetime of a mouse.

TCR V genes used in the responses against different antigens

A major area of research in recent years has been to determine which sets of TCR V genes are used in the responses against different antigens. Because T cells recognize antigenic peptides bound to a particular MHC molecule this depends on:

- the antigen; and
- the MHC molecules expressed by an individual.

Once the TCRs have been generated, T cells are subjected to thymic selection and may be further selected by interactions with APCs in the periphery. For these reasons, even if TCRs are generated by random recombination of gene segments, the expressed repertoire will be skewed toward the use of particular gene segments. Furthermore, the preference for different V gene segments by distinct T cell subsets may reflect their ontogeny. For example:

- γδ T cells residing in mouse skin (dendritic epidermal cells or DECs) express only the Vγ3 and Vδ1 segments;
- intraepithelial lymphocytes from the gut express Vγ5 almost exclusively (in combination with Vδ4–7).

It is thought that these populations arise at distinct stages of intrathymic T cell development, and they appear to have distinct functions.

RECOGNITION BY THE αβ TCR REQUIRES ANTIGEN TO BE BOUND TO AN MHC MOLECULE

Intracellular pathogens and neoplasia present a special challenge to the immune system. The immediate problem is one of detection. The immune system has therefore evolved an elegant means of recognizing pathogens that have invaded or been taken into the host cell, as well as tumor antigens produced within the cytosol:

- proteins from within the cell are digested into short peptide fragments;
- the short peptide fragments are displayed on the cell surface through binding to specialized antigen-presenting molecules termed **MHC class I** proteins;
- in a similar fashion, peptides derived from proteins ingested from the extracellular environment by phagocytosis are presented by **MHC class II** molecules;
- the peptide–MHC complexes serve as ligands for TCRs.

This antigen processing and presentation pathway, upon which both activation and regulation of the immune response rests, is a complex and fascinating subject (see Chapter 7).

In humans the MHC is known as the HLA

The proteins responsible for presenting antigens to T cells, MHC class I and class II proteins, were originally discovered as histocompatibility (transplantation) antigens. Histocompatibility refers to the ability to accept tissue grafts from an unrelated donor.

The major histocompatibility complex locus (MHC) comprises over 100 separate genes and was discovered when it was recognized that both donor and recipient had to possess the same MHC haplotype to avoid graft rejection.

The principal moieties that determine rejection were identified as MHC class I and class II molecules (see below), but we know now that that the main purpose of the MHC is not to prevent graft rejection. The remaining genes in the MHC (sometimes called class III) are very diverse. Some encode:

- complement system molecules (C4, C2, factor B);
- cytokines (e.g. tumor necrosis factor);
- enzymes;
- heat-shock proteins; and
- other molecules involved in antigen processing.

There are no functional or structural similarities between these other gene products.

All mammalian species possess the MHC, though details of the complex vary from one species to the next. In humans the locus is known as the **HLA** (an abbreviation for **human leukocyte antigen**); in mice it is known as the **H-2 locus** (Fig. 5.5).

MHC molecules provide a sophisticated surveillance system for intracellular antigens

From a cell's perspective, there are two types of antigen that must be dealt with:

- antigens that are **intrinsic** to the cell (antigenic peptides from viruses or other pathogens that inhabit the cell);
- antigens that are **extrinsic** to the cell.

Q. How are antigens taken up by APCs?

A. They are internalized by phagocytosis or pinocytosis, either by directly binding to receptors on the surface of the APC (see Fig. 1.11) or following opsonization by antibody and/or complement (see Fig. 4.14).

MHC class I molecules handle intrinsic antigens, while MHC class II molecules handle extrinsic antigens. In both cases, the antigenic peptides are produced by proteolytic processing of proteins.

In general:

- MHC class I molecules present antigen to cytotoxic T cells, which are important in controlling viral infections by lysing infected cells;
- MHC class II molecules present antigen to helper T cells, which aid B cells in generating antibody responses to extracellular protein antigens.

MHC class I molecules consist of an MHC-encoded heavy chain bound to a β_2-microglobulin

The overall structure of the extracellular portion of an MHC class I molecule is depicted in Fig. 5.6. It comprises a glycosylated heavy chain (45 kDa) non-covalently associated with β_2-microglobulin (12 kDa), which is a polypeptide that is also found free in serum.

Organization of the murine and human MHCs

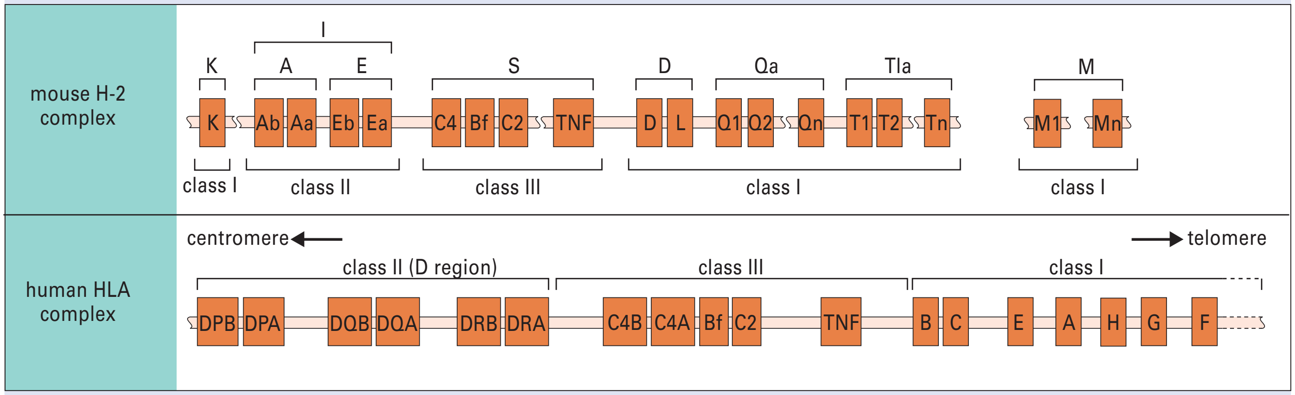

Fig. 5.5 Diagram showing the locations of subregions of the murine and human MHCs and the positions of the major genes within these subregions. The human organization pattern, in which the class II loci are positioned between the centromere and the class I loci, occurs in every other mammalian species so far examined. The regions span 3–4 Mbp of DNA.

A model of an MHC class I molecule

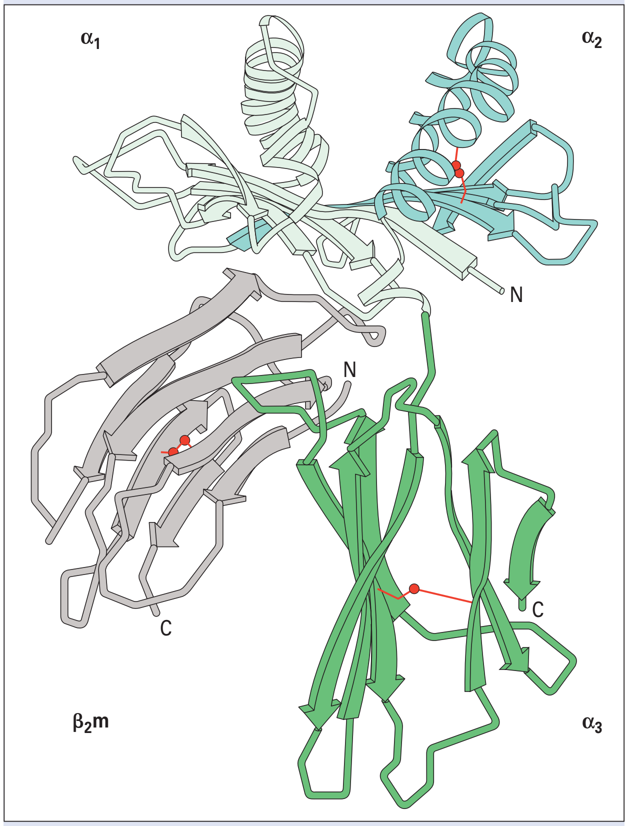

Fig. 5.6 The peptide backbone of the extracellular portion of HLA-A2 is shown. The three globular domains (α_1, α_2, and α_3) of the heavy chain are shown in green or turquoise and are closely associated with the non-MHC-encoded peptide, β_2-microglobulin (β_2m, gray). β_2-Microglobulin is stabilized by an intrachain disulfide bond (red) and has a similar tertiary structure to an immunoglobulin domain. The groove formed by the α_1 and α_2 domains is clearly visible.

The class I heavy chain consists of:

- three extracellular domains, designated α_1 (N terminal), α_2, and α_3;
- a transmembrane region; and
- a cytoplasmic tail.

The three extracellular domains each comprise about 90 amino acids:

- the α_2 and α_3 domains both have intrachain disulfide bonds enclosing loops of 63 and 86 amino acids, respectively;
- the α_3 domain is structurally homologous to the immunoglobulin constant region domain (C) and contains a site that interacts with CD8 on cytotoxic T cells.

The extracellular portion of the class I heavy chain is glycosylated, the degree of glycosylation depending on the species and haplotype.

The predominantly hydrophobic transmembrane region comprises 25 amino acid residues and traverses the lipid bilayer, most probably in an α-helical conformation.

The hydrophilic cytoplasmic domain, 30 to 40 residues long, may be phosphorylated in vivo.

β_2-Microglobulin is essential for expression of MHC class I molecules

β_2-Microglobulin:

- is non-polymorphic in humans, but dimorphic in mice (because of a single amino acid change at position 85);
- like the α_3 domain, has the structure of an immunoglobulin constant region domain;
- associates with a number of other class I-like molecules, such as the products of the *CD1* genes on chromosome 1 in humans (see below) and the Fc receptor, which mediates the uptake of IgG from milk in neonatal rat intestinal cells. These class I-like molecules, which have a structural similarity to the products of the MHC class I genes, are encoded by genes located in the class I loci and are referred to as **class Ib molecules**.

- is essential for the expression of all class I molecules at the cell surface – mutant mice lacking β_2-microglobulin do not express MHC class I molecules and are severely defective in presenting intrinsic antigens to T cells.

Heavy chain α_1 and α_2 domains form the antigen-binding groove

X-ray crystallography has shown that the α_1 and α_2 domains constitute a platform of eight antiparallel β strands supporting two antiparallel α helices (Fig. 5.7). The disulfide bond in the α_2 domain connects the N terminal β strand to the α helix of the α_2 domain. A long groove separates the α helices of the α_1 and α_2 domains.

The original crystal structure of the HLA-A2 molecule revealed diffuse extra electron density in the groove, suggesting the presence of bound peptide antigen. This interpretation was supported by the observation that the majority of polymorphic residues and T cell epitopes on class I molecules are located in or near the groove.

VARIATIONS IN AMINO ACID SEQUENCE CHANGE THE SHAPE OF THE BINDING GROOVE – Comparison of the structures of HLA-A2 and HLA-Aw68 have further refined our understanding of the structural basis for the binding of peptide to class I antigens.

The antigen-binding site of the MHC class I molecule HLA-A2

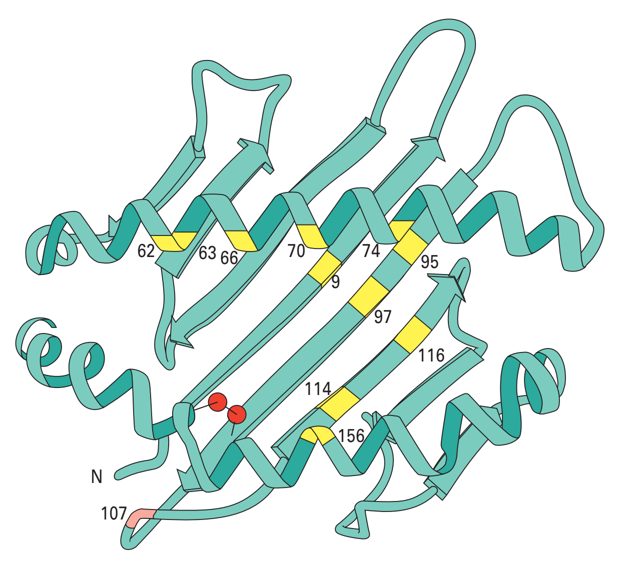

Fig. 5.7 The view of the peptide antigen-binding groove in HLA-A2 as 'seen' by the TCR. The α_1 and α_2 domains each consist of four antiparallel β strands followed by a long helical region. The domains pair to form a single eight-stranded β sheet topped by α helices. The locations of the most polymorphic residues are highlighted. Residues around the binding site are highly polymorphic. For example, HLA-2 and HLA-Aw68 differ from each other by 13 amino acid residues. Ten of these differences occur around the antigen-binding site (yellow). (Modified from Bjorkman et al. Nature 1987;329:512–516, with additional data from Parham P. Nature 1989;342:617–618)

The differences between HLA-A2 and HLA-Aw68 result from amino acid side-chain differences at 13 positions:

- six in α_1;
- six in α_2; and
- one (residue 245, which contributes to interactions with CD8) in α_3.

Ten of the twelve differences between HLA-A2 and HLA-Aw68 are at positions lining the floor and side of the peptide-binding groove (see Fig. 5.7). These differences give rise to dramatic differences in the shape of the groove and on the antigen peptides that it will bind.

Seen in detail, the peptide-binding groove forms a number of ridges and pockets with which amino acid side chains can interact. Typically the groove of an MHC class I molecule will accommodate peptides of eight or nine residues.

Amino acid variations within the peptide-binding groove can vary the positions of the pockets, providing the structural basis for differences in peptide-binding affinity that in turn govern exactly what is presented to a T cell (Fig. 5.8).

MHC class II molecules resemble MHC class I molecules in their overall structure

The products of the MHC class II genes are:

- A and E in the mouse;
- DP, DQ, and DR in humans.

These products are heterodimers of heavy (α) and light (β) glycoprotein chains, and both chains are encoded in the MHC:

- the α chains have molecular weights of 30–34 kDa;
- the β chains range from 26 to 29 kDa, depending on the locus involved.

A number of lines of evidence indicate that the α and β chains have the same overall structures. An extracellular portion comprising two domains (α_1 and α_2 or β_1 and β_2)

Peptide-binding grooves of HLA-Aw68 and HLA-A2

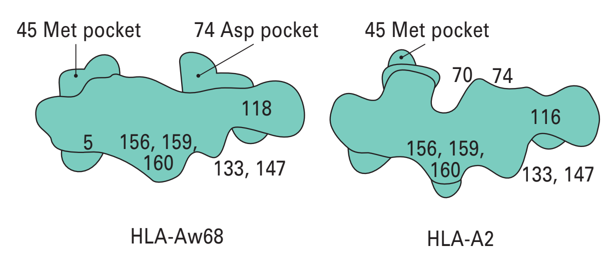

Fig. 5.8 The shapes of the antigen-binding groove on each molecule are illustrated. Differences in amino acids around the groove create different antigen-binding sites. For example, residues around position 45 produce a methionine-binding pocket in both molecules, but the aspartate-binding pocket around residue 74 is present only in HLA-Aw68.

is connected by a short sequence to a transmembrane region of about 30 residues and a cytoplasmic domain of about 10 to 15 residues.

The α_2 and β_2 domains are similar to the class I α_3 domain and β_2-microglobulin, possessing the structural characteristics of immunoglobulin constant domains.

The β_1 domain contains a disulfide bond, which generates a 64 amino acid loop.

The difference in molecular weights of the class II α and β chains is due primarily to differential glycosylation:

- the α_1, α_2, and β_1 domains are N-glycosylated;
- the β_2 domain is not N-glycosylated.

The β_2 domain does, however, contain a binding site for CD4, and MHC class II molecules on APCs interact with CD4 on T cells in a manner analogous to the interaction of MHC class I molecules with CD8. CD4 and CD8 are important elements in the recruitment of kinases that signal T cell activation.

Despite the differences in length and organization of the polypeptide chains, the overall three-dimensional structure of MHC class II molecules is very similar to that of MHC class I molecules (Fig. 5.9).

Comparison of the extracellular domains of class I and class II molecules

HLA-Aw68 (class I)

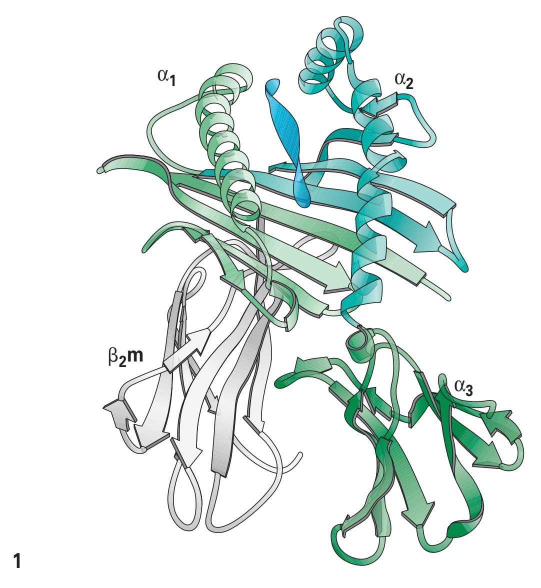

HLA-DR1 (class II)

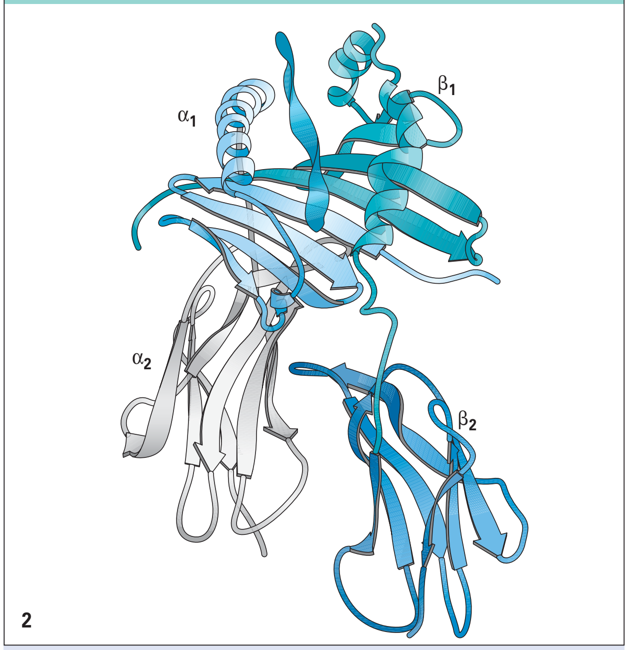

Fig. 5.9 Ribbon diagrams of the extracellular domains of class I HLA-Aw68 (**1**) and class II HLA-DR1 (**2**) MHC molecules. The binding cleft is shown with a resident peptide. These diagrams emphasize the similarity in the three-dimensional structures of class I and class II molecules. (Redrawn from Stern LJ. Structure 1994;2:245–251. Copyright 1994 with permission from Elsevier)

MHC CLASS I AND CLASS II MOLECULES BIND TO PEPTIDES DERIVED FROM DIFFERENT SOURCES

The MHC class II binding groove accommodates longer peptides than MHC class I

The structures of MHC class I and class II molecules reflect their functional differences.

The binding groove of MHC class II molecules is more open than that of MHC class I molecules, to accommodate longer peptides (Figs 5.10 and 5.11):

- MHC class I molecules bind short fragments of eight to ten amino acids; whereas
- MHC class II molecules bind peptides of 13 to 24 amino acids.

The structural features of the class II antigen-binding site have been illuminated by determination of the crystal structure of HLA-DR1 complexed with an influenza virus peptide. Pockets clearly visible within the peptide-binding site accommodate five side chains of the bound peptide and explain the peptide specificity of HLA-DR1.

The precise topology of the MHC peptide-binding groove depends partly on the nature of the amino acids within the groove, and thus varies from one haplotype to the next.

Which peptide can bind to a particular MHC molecule depends on the nature of the side chains of the peptide and their complementarity with the MHC molecule's binding groove. Some amino acid side chains of the peptide stick out of the groove and are available to contact the TCR.

Peptides are held in MHC molecule binding grooves by characteristic anchor residues

It is possible to purify and sequence peptides that have been generated by a cell and then bound by MHC molecules at the cell surface. These peptides include:

- foreign peptides from internalized antigens or viral particles; and
- self molecules produced within the cell or endocytosed from extracellular fluids.

Peptide-binding sites of class I (H-2K^b) and class II (HLA-DR1) MHC molecules

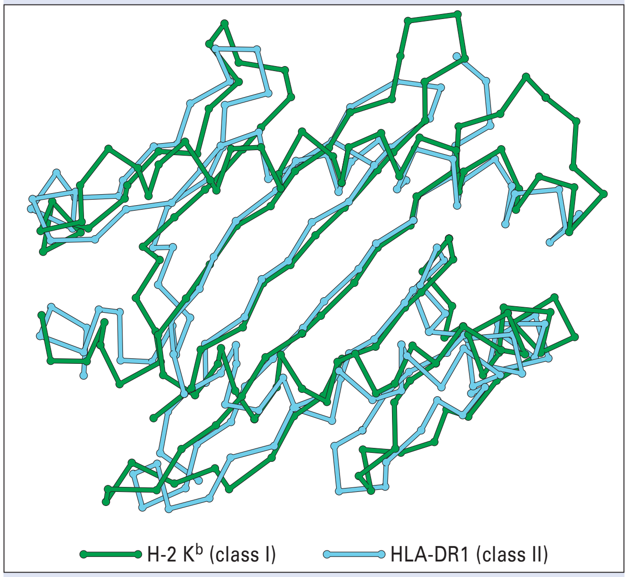

Fig. 5.10 The peptide binding sites of class I (H-2K^b) and class II (HLA-DR1) MHC molecules are shown as α carbon atom traces in a top view of the peptide-binding clefts. The similarities between the two sites can clearly be seen, but there are also some differences, some of which account for the difference in peptide length preference between class I (8–10 amino acid residues) and class II (>12 amino acid residues). (Redrawn from Stern LJ. Structure 1994;2:245–251. Copyright 1994 with permission from Elsevier)

Self peptides eluted from MHC class I molecules have been purified and sequenced.

For MHC class I molecules, interactions at the N and C terminals confine the peptide to the binding groove

A number of peptides bound by particular MHC molecules have been sequenced, and characteristic residues identified – one at the C terminus and another close to the N terminus of the peptide. These characteristic motifs distinguish sets of binding peptides for different MHC class I molecules (Fig. 5.12).

The significance of the conserved residues has become clear by analysis of the three-dimensional structures of several MHC class I molecules, which have generated a clear picture of the peptide residing in the binding groove:

- the ends of the peptide-binding groove are closed;
- the peptide is an extended (not α-helical) chain of nine amino acids and the N and C terminals are buried at the ends of the groove;
- some of the side chains extend into the pockets formed within the variable region of the class I heavy chain;
- numerous hydrogen bonds are formed between residues in the class I molecule and those of the peptide along its length;
- in particular, tyrosine residues commonly found at the N terminus of the peptide and a conserved lysine in the MHC class I molecule binding groove stabilize peptide binding (see Fig. 5.11);
- the centers of the peptides bulge out of the groove, so presenting different structures to TCRs.

Peptides bind non-covalently within the antigen-binding groove

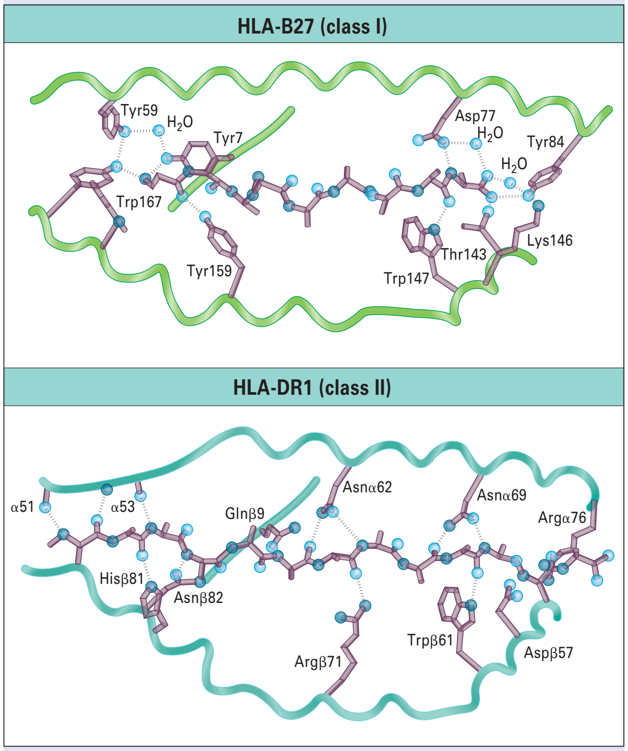
HLA-B27 (class I)

HLA-DR1 (class II)

Fig. 5.11 The hydrogen bonds made by the main chain of a bound peptide with class I (HLA-B27) or class II MHC molecules (HLA-DR1) are shown. The major difference between the two hydrogen-bonding patterns is the clustering of conserved class I hydrogen bonds at the ends of the peptide. By contrast, conserved class II hydrogen bonds are distributed throughout the length of the peptide. (Redrawn from Stern LJ. Structure 1994;2:245–251. Copyright 1994 with permission from Elsevier)

This picture is consistent with the characteristic motifs found at the ends of peptides eluted from class I molecules.

For MHC class II molecules, peptides may extend beyond the ends of the binding groove

The binding groove of the MHC class II molecule:

- also incorporates a number of binding pockets, though the locations are somewhat different from that on class I molecules;
- is not closed at the ends, so bound peptides extend out of the ends of the groove.

Consistent with this observation, peptides eluted from MHC class II molecules tend to be longer (over 15 residues).

Conserved anchor residues in peptides eluted from MHC class II molecules have been identified (Fig. 5.13), but these are more difficult to detect than those of class I

Allele-specific motifs in peptides eluted from MHC class I molecules

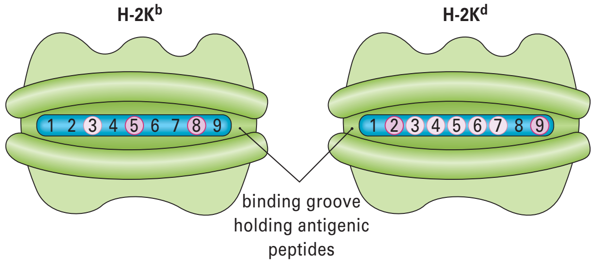

position										position								
1	2	③	4	⑤	6	7	⑧	9		1	②	③	④	⑤	⑥	⑦	8	⑨
				F Y			L		dominant anchor residue		Y							I L
		Y					M		strong anchor residue			N I L	P	M	K F	T N		

Fig. 5.12 Class I MHC molecules from either H-2K^b or H-2K^d haplotypes were immunoprecipitated. Peptides bound to these molecules were purified and sequenced. Amino acid residues commonly found at a particular position are classified as 'dominant' anchor residues. Residues that are fairly common at a site are shown as 'strong'. Positions for which no amino acid is shown could be occupied by several different amino acids with equal frequency. The one-letter amino acid code is used. The diagram represents the class I MHC molecule binding groove, viewed from above with anchor positions of each haplotype highlighted.

Allele-specific motifs in peptides eluted from MHC class II molecules

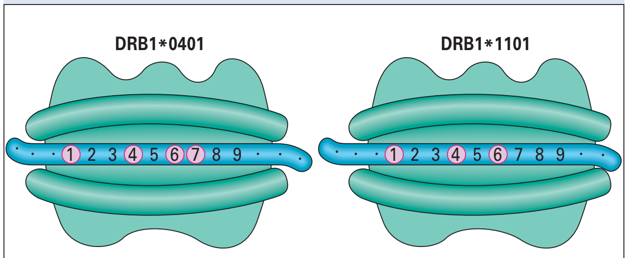

position										position								
①	2	3	④	5	⑥	⑦	8	9		①	2	3	④	5	⑥	7	8	9
W Y			M		T	L Q			dominant anchor residue	W			M L		R			
			A V L		S	M N			strong anchor residue				V		K			

Fig. 5.13 HLA-DR molecules of two haplotypes (DRB1*0401 and DRB1*1101) were purified and incubated with a library of peptides (generated in phage M13). After multiple rounds of selection, peptides that bound effectively to the MHC class II molecules were identified and sequenced. Residues having a frequency greater than 20% are shown as 'dominant' anchor residues. Other fairly common residues are shown as 'strong'. Note that the binding site on MHC class II molecules accommodates longer peptides than that on MHC class I molecules. (Data abstracted from Hammer J, Valsasnini P, Tolba K, et al. Cell 1993;74:197–203)

peptides because of the ragged ends and the more even distribution of bonds tethering the peptide to the MHC class II molecule binding groove (see Fig. 5.11).

Peptides binding MHC class II are less uniform in size than those binding MHC class I molecules

A major difference between MHC class I and class II molecules occurs at the ends of the peptide-binding groove:

- for MHC class I molecules, interactions at the N and C terminals confine the peptide to the cleft;
- for MHC class II molecules, peptides may extend beyond the ends of the cleft.

The peptides that bind to MHC class I molecules come from proteins synthesized within the cell, which are broken down and transported to the endoplasmic reticulum. The mechanism of antigen processing is explained fully in Chapter 7, but it should be noted here that the internal antigen processing pathways generally produce peptides of an appropriate size to occupy the MHC class I molecule antigen-binding groove.

Peptides that bind to MHC class II molecules come from proteins that have been internalized by the cell and then degraded. These peptides are less uniform in size than those that bind to MHC class I molecules and may be trimmed once they have found their way to the MHC class II molecule.

The MHC class II molecule antigen processing pathway is quite distinct from the MHC class I molecule pathway (see Chapter 7).

CLASS I AND CLASS II MHC MOLECULES PRESENT PEPTIDE ANTIGENS TO THE TCR IN A CELL–CELL INTERACTION

Once the structures of the TCR and the MHC–peptide complex had been established, the next question was to determine how they interacted.

The first crystallographic data were derived using a co-crystal of a mouse MHC class I molecule bound to an endogenous cellular peptide and αβ TCR (Fig. 5.14). This structure showed that the axis of the TCR was roughly aligned with the peptide-binding groove on the MHC molecule, but set at 20–30° askew. This means that:

- the first and second CDRs of the TCR α and β chains (see Fig. 5.1) are positioned over residues near the N and C terminals of the presented polypeptide; and
- the third CDRs of each chain, lying at the center of the TCR binding site, are positioned over the central residues of the peptide that protrude from the groove.

Interaction of a T cell receptor and MHC–peptide complex

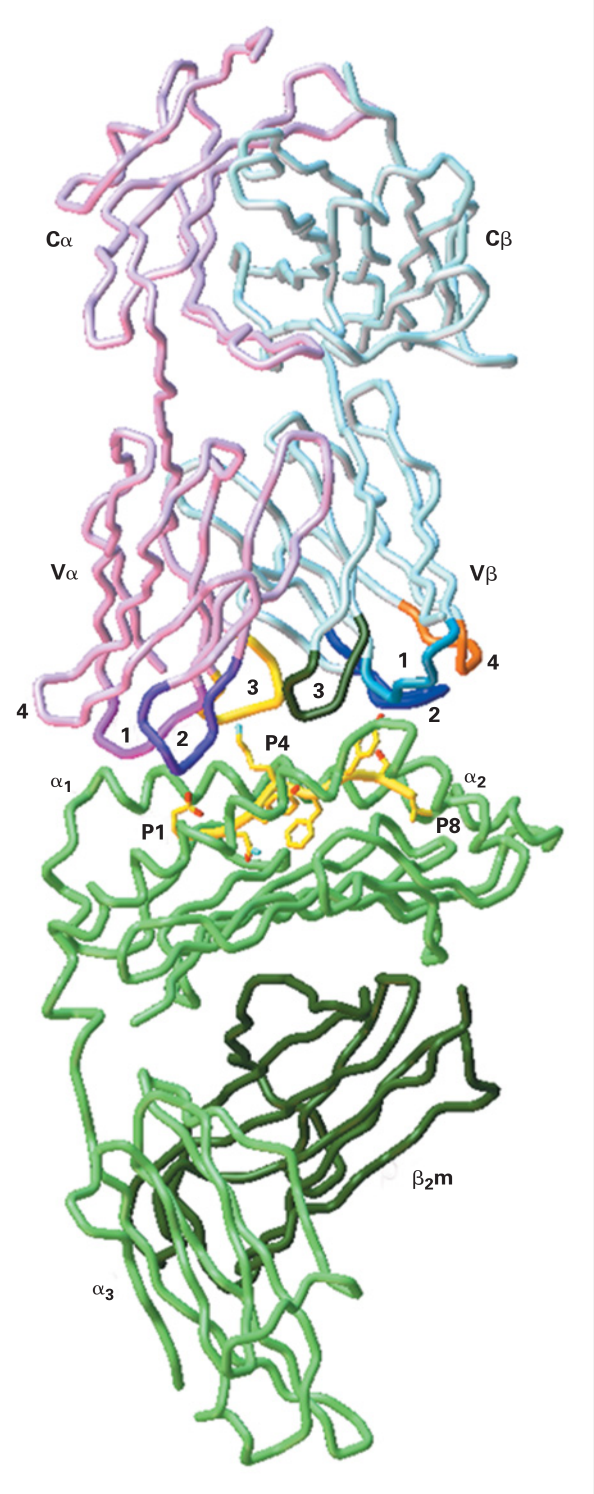

Fig. 5.14 The structure of an MHC class I molecule (H-2K^b) complexed to an octapeptide (yellow tube) is shown bound to an αβ TCR. The six CDRs that contact the peptide (1, 2, 3, 3, 1, 2) are highlighted in deeper colors. Residues from αHV4 (pink) and βHV4 (orange) are not positioned to take part in the intermolecular interactions. (Illustration kindly provided by Dr Christopher Garcia, from Science 1996;274:209–219. Copyright 1996 AAAS)

Residues from each of the CDRs are positioned to interact with residues from the MHC molecule.

Q. What advantage is there in having CDR3 segments at the center of the TCR binding site?

A. CDR3 demonstrates the greatest diversity of the TCR CDRs (because it is generated by gene recombination, see Fig. 3.6). It is therefore better suited than CDR1 and CDR2 for interactions with diverse antigenic peptides.

The molecular structure therefore underpins the experimental findings that T cells recognize antigenic peptides bound to particular MHC molecules.

This arrangement of TCR and MHC–antigen is broadly comparable to those found with the small number of receptors so far analyzed by X-ray crystallography.

Aggregation of TCRs initiates T cell activation

Although basic models of T cell activation by antigen–MHC show one receptor being triggered by one complex (e.g. Fig. 1.9), this is simplistic.

Each T cell may express 10^5 receptors and each APC has a similar number of MHC molecules. If a T cell engages an APC, only a tiny proportion of the MHC–antigen complexes on its surface will be of the correct type to be recognized by the T cell.

What, then, is the minimum signal for T cell activation? In practice:

- only a few peptide–MHC–TCR interactions are needed – perhaps 100 specific interactions, involving about 0.1% of the MHC molecules on the APC;
- moreover, the interactions can take place over a period of time – it is not necessary for all 100 TCRs to be engaged simultaneously.

The model of the TCR shown in Fig. 5.3 suggests that it can form a dimeric structure clustered around the signaling molecules of the CD3 complex.

Interestingly, there is some evidence that MHC molecules can also dimerize, and that TCRs bound to MHC–peptide complexes tend to form dimers or aggregates. These observations have led to the view that T cell activation requires the cooperative aggregation of specific TCRs with MHC–peptide complexes.

The auxiliary molecules CD4 and CD8 are also important in T cell activation, and the presence of CD4 or CD8 can help stabilize the interaction of the TCR and MHC–peptide. In addition, kinases associated with these molecules are brought into proximity with CD3 so they can phosphorylate the ζζ dimer that initiates activation (Fig. 5.15) The ensuing steps are described in Chapter 7 (see Fig. 7.21).

Antigenic peptides can induce or antagonize T cell activation

The affinity of TCRs for antigen peptide–MHC, expressed as an association constant, is typically of the order of 10^{-5} to 10^{-6} M, which is much lower than the affinity of antibody for an epitope on an antigen commonly generated by an immune response.

The role of CD4 and CD8 in T cell activation

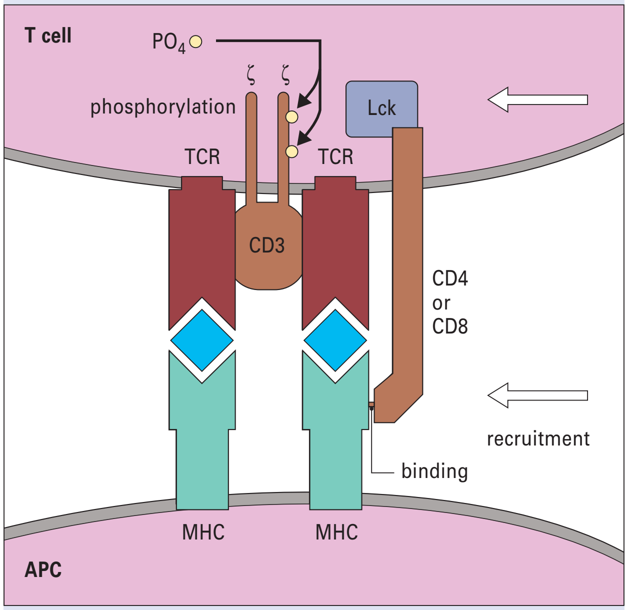

Fig. 5.15 After aggregation of MHC–peptide on the APC with the TCR, either CD4 or CD8 can join the complex. CD4 binds to MHC class II molecules and CD8 to MHC class I molecules. The kinase Lck is attached to the intracytoplasmic portion of CD4 or CD8. Binding of the CD4 or CD8 molecules to specific sites of the MHC molecules brings the kinase into proximity with the ITAMs on the CD3 ζζ dimer (see Fig. 5.3). The kinase phosphorylates the motifs, as the first step in T cell activation.

The affinity of antigen–MHC for the TCR is especially important because it determines the degree to which the T cell becomes activated. Peptides that activate a T cell are called **agonist peptides**. By changing one or two amino acids in an agonist peptide, it is possible to generate peptides that antagonize the normal activation by the original peptide. The **antagonist peptides** typically have a lower affinity for the TCR than the original agonist peptide and are thought to act by either:

- interfering with the binding of the agonist peptide to the MHC molecule; or
- binding less effectively to the TCR.

Occasionally, a modified peptide will be more effective than the agonist in activating the T cell. These **strong agonist (or superagonist) peptides** produce higher affinity binding in the TCR–MHC–peptide complex.

The distinctions between these three different types of peptide (agonist, antagonist, and strong agonist) have biological ramifications. Researchers are attempting, for example, to use antagonist peptides to block adverse immune responses.

Understanding TCR affinity could provide insight into how T cells become tolerant to self antigens during development (see Chapter 19).

What constitutes T cell specificity?

Finally we should consider the question of what constitutes T cell specificity. When the specificity of lymphocytes was first explained in Chapter 1, it was stated that each lymphocyte binds to just one antigen using its receptor. Although this is a useful starting point for understanding immune responses, it is not strictly true.

In Chapter 3 it was seen that immunoglobulins could bind to different antigens if they had epitopes that were sufficiently similar. This chapter has shown that antigenic peptides can be mutated and still bind and trigger T cell activation.

One question, then, is how far a peptide can be mutated and still bind to its own TCR. In some cases it has been possible to individually change each amino acid in a peptide without destroying its ability to bind to the MHC molecule or the TCR.

Provided the peptide can form part of the TCR–MHC–peptide complex and provide sufficient binding energy, its precise amino acid sequence does not matter. This is a very important observation.

Later, when we consider how antigenic peptides from microorganisms can trigger autoimmune diseases (see Chapter 20), we learn that one possible mechanism is that a self peptide and a foreign peptide are sufficiently similar to bind to the same T cell, so causing a breakdown in self tolerance. One conclusion from the work cited above is that such peptides do not have to be identical to be cross-reactive.

THE OVERALL ORGANIZATION OF THE MHC LOCI DIFFERS BETWEEN SPECIES

The number of gene loci for MHC class I and class II molecules varies between species and between different haplotypes within each species, and many polymorphic variants have been described at each of the loci.

Much of the original work on the MHC was done using mice and what has been learned from these studies has been broadly applicable to humans.

The three principal human MHC class I loci are HLA-A, HLA-B, and HLA-C

The human MHC class I region contains three principal class I loci – called HLA-A, HLA-B, and HLA-C. Each locus encodes the heavy chain of a classical MHC class I molecule and the whole region:

- extends over 1.8 million bases of DNA; and
- includes 118 genes (Fig. 5.16).

Closer analysis of this region has revealed multiple additional MHC class I genes.

HLA-E, HLA-F, HLA-G, *and* HLA-H *are class Ib genes*

The *HLA-E*, *HLA-F*, *HLA-G*, and *HLA-H* genes also encode MHC class I proteins, and are called class Ib genes. They are much less polymorphic than the A, B, and C locus gene products, and recent work has ascribed various functions to them. For example, *HLA-E* and *HLA-G* gene products can bind to antigenic peptides, but are involved in recognition by NK cells (see p. 207).

Genes within the human MHC class I region

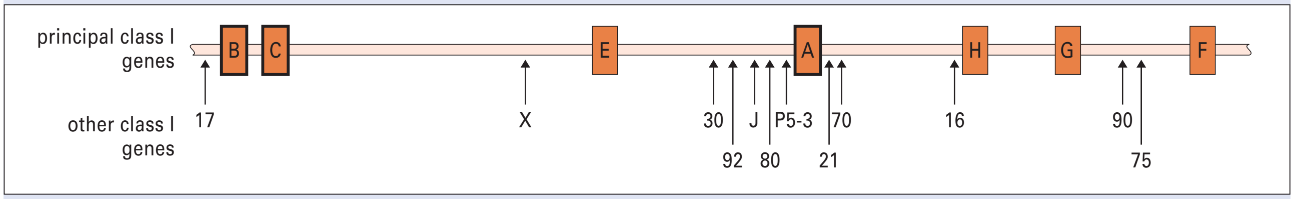

Fig. 5.16 The human MHC class I region lies telomeric to the MHC class II region (see Fig. 5.5). In addition to the genes encoding the classic transplantation antigens (*HLA-A, HLA-B,* and *HLA-C*), several other principal class I-like genes have been identified (*HLA-E, HLA-F,* and *HLA-G*). Mutations in the *HLA-H* gene are associated with hemochromatosis, a disorder that causes the body to absorb excessive amounts of iron from the diet. A number of other non-classical class I genes and pseudogenes are present, mostly with unknown functions.

Mice have two or three MHC class I loci

The mouse MHC (H-2) has two or three MHC class I loci, but the number of MHC class I genes varies between haplotypes (Fig. 5.17).

MHC class I genes involved in antigen presentation to T cells are located in H-2K, H-2D, and H-2L loci.

The organization of the H-2K region is similar in all strains that have been studied. It contains two class I genes, termed K and K2:

- the *H-2K* gene encodes the H-2K antigen expressed on most cell types and recognized serologically; whereas
- the *H-2K2* gene exhibits varied patterns of expression, depending on the strain.

The numbers of genes in the H-2D and H-2L loci are variable (see Fig. 5.17).

Qa, Tla, and M genes encode MHC class Ib molecules

The Qa, Tla, and M genes of the mouse MHC class I region (see Figs 5.5 and 5.17) encode class Ib molecules, of which the functions are mostly unknown:

- the Qa locus comprises about 200 kb of DNA distal to H-2D/L and encodes the serologically detectable molecules Qa-1, 2, 3, 4, and 5, and encompasses a cluster of eight (BALB/c) to ten (B10) MHC class I genes;
- Qa-1b corresponds to HLA-E, which is involved in NK cell recognition (see Fig. 10.4);
- the Tla region, although defined initially as encoding the TL (thymus leukemia) antigen, has subsequently been shown to contain the largest number of class I genes and the greatest number of differences in organization between the B10 and BALB/c haplotypes;
- the M region contains a number of new MHC class I genes, termed M1–M7.

The class Ib genes exhibit a low degree of polymorphism.

Human MHC class II genes are located in the HLA-D region

The human MHC class II region spans about 1000 kb of DNA, and the order and orientation of the various loci are similar to that of the homologous loci in the mouse MHC class II region (see below).

The HLA-D region encodes at least six α and ten β chain genes for MHC class II molecules (Fig. 5.18). Three loci (DR, DQ, and DP) encode the major expressed prod-

Genes within the murine class I region

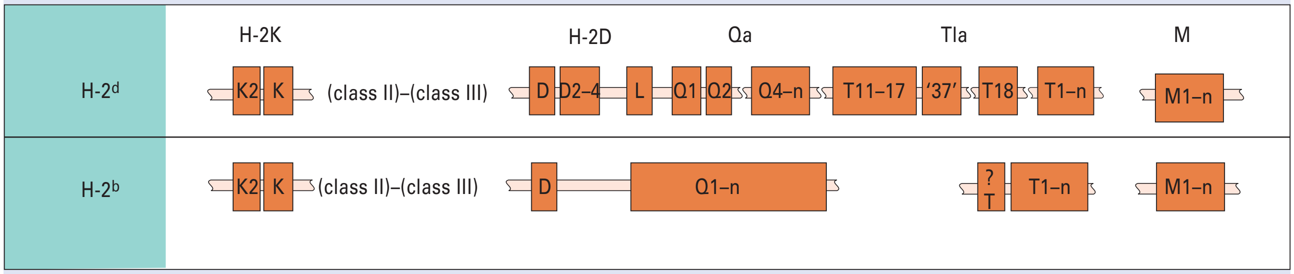

Fig. 5.17 The organization of the MHC class I region of two haplotypes, BALB/c (H-2^d) and B10 (H-2^b), is shown. The class II region lies between the H-2K and H-2D regions. The brackets denote gaps added to align alleles between the two haplotypes. H-2^d and H-2^b haplotypes have different numbers of class I genes in the H-2D/H-2L region. Five class I genes map to the D/L region of BALB/c (H-2^d) mice. Two of these genes encode the serologically detectable H-$2D^d$ and H-$2L^d$ antigens. Three additional class I genes are found in the region between the proximal H-$2D^d$ and distal H-$2L^d$ genes. These genes, called $D2^d$, $D3^d$, and $D4^d$, are of unknown function. Only one class I gene has been identified in the H-2D region of B10 mice. The Q, T, and M regions contain a large number of class Ib genes, the functions of which are mostly unknown.

Genes within the human and mouse class II regions

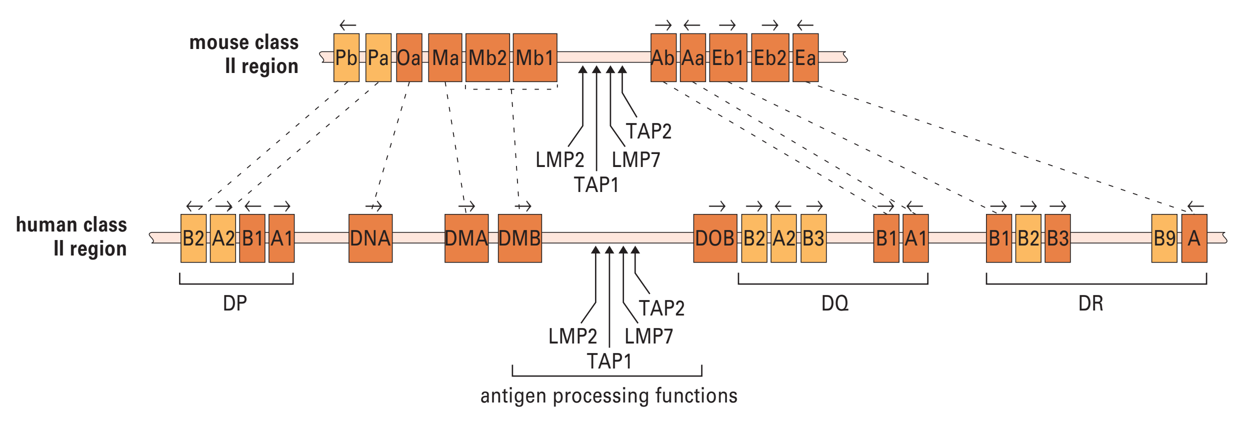
mouse class II region
Pb
Pa
Oa
Ma
Mb2
Mb1
Ab
Aa
Eb1
Eb2
Ea
LMP2
TAP1
LMP7
TAP2
human class II region
B2
A2
B1
A1
DNA
DMA
DMB
DOB
B2
A2
B3
B1
A1
B1
B2
B3
B9
A
DP
DQ
DR
LMP2
TAP1
LMP7
TAP2
antigen processing functions

Fig. 5.18 The arrangement of the genes within the human and murine MHCs is shown. Homologous genes between the two species are indicated. Expressed genes are colored orange and pseudogenes are shown yellow. Mice of the b, s, f, and q haplotypes fail to express class II-E molecules. The b and s haplotypes fail to transcribe the Ea gene, but make normal cytoplasmic levels of Eb chain. Mice of f and q haplotypes fail to make both Ea and Eb chains.

ucts of the human MHC class II region, but additional genes have also been identified:

- the DR family comprises a single α gene (DRA) and up to nine β genes (DRB1–9), including pseudogenes and several different gene arrangements occur within the locus;
- the DQ and DP families each have one expressed gene for α and β chains and an additional pair of pseudogenes.

DR, DQ, and DP α chains associate in the cell primarily with β chains of their own loci. For example:

- the *DPA1* and *DPB1* gene products associate to generate the HLA-DP class II molecules detected using specific antibodies;
- similarly, *DQA1* and *DQB1* encode the HLA-DQ antigens.

The organization and length of the DRB region varies in different haplotypes (Fig. 5.19), with different numbers of β chains expressed.

The MHC class II region also contains genes that encode proteins involved in antigen presentation that are not expressed at the cell surface (see Chapter 7).

Mouse MHC class II genes are located in the H-2I region

The α and β chains of mouse MHC class II molecules are encoded by separate genes located in the I region of the H-2 complex (see Figs 5.5 and 5.18). The gene nomenclature indicates first the locus and then the type of chain it encodes:

- the Ab and Aa genes thus encode the β and α chains of the A molecule; and
- the Eb and Ea genes likewise encode the two chains of the E molecule (Fig. 5.5).

Several other class II b and a genes for which no gene product is known have been cloned:

- one of these – Pb – is a pseudogene;
- two others – Ob and Eb2 – may be functional.

The number of DRB loci varies with different haplotypes

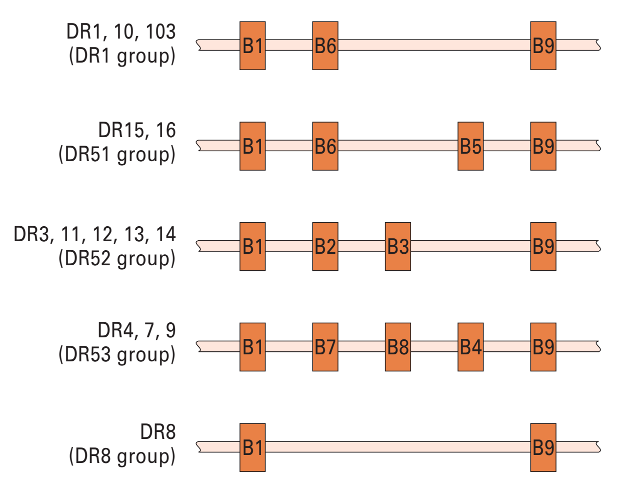
DR1, 10, 103
(DR1 group)
B1
B6
B9
DR15, 16
(DR51 group)
B1
B6
B5
B9
DR3, 11, 12, 13, 14
(DR52 group)
B1
B2
B3
B9
DR4, 7, 9
(DR53 group)
B1
B7
B8
B4
B9
DR8
(DR8 group)
B1
B9

Fig. 5.19 The numbers of DRB loci varies between individuals. For example, a person who has a haplotype producing molecules of the type DR1 (see Appendix 1) has three loci for DRB (top line). Not all of these loci produce mRNA for DR β chains.

The Ob and Eb2 genes display a low level of polymorphism and are transcribed, but it is not known whether they are translated.

Different strains of mice vary in their expression of some of the MHC class II genes (see Fig. 5.18).

MHC polymorphism is concentrated in and around the peptide-binding cleft

A hallmark of the MHC is the extreme degree of polymorphism (structural variability) of the molecules encoded within it. The class Ib molecules are much less polymorphic than the classical class I and II molecules. A list of allelic variants of HLA class I and II molecules is given in Appendix 1.

Within a particular MHC class I or class II molecule, the structural polymorphisms are clustered in particular regions. The amino acid sequence variability in class I molecules is clustered in three main regions of the α_1 and α_2 domains. The α_3 domain appears to be much more conserved.

Q. Correlate the domains that show structural variability and structural conservation with the molecules that they interact with.

A. The α_1 and α_2 domains interact with the variable antigen peptides (see Fig. 5.7) and the TCR, while the α_3 domain interacts with the monomorphic CD8 molecule (see Fig. 5.15).

In MHC class II molecules, the extent of variability depends on the subregion and on the polypeptide chain. For example:

- most polymorphism occurs in DRβ and DQβ chains, whereas DPβ chains are slightly less polymorphic;
- DQα is polymorphic whereas DRα chains are virtually invariant, being represented by just two alleles.

In outbred populations in which individuals have two MHC haplotypes, hybrid class II molecules with one chain from each haplotype can be produced. This generates additional structural diversity in the expressed molecules.

Most of the polymorphic amino acids in MHC class I and class II molecules are clustered on top of the molecule around the peptide-binding site (see Fig. 5.7). Variation is therefore centered in the base of the antigen-binding groove or pointing in from the sides of the α helices. This polymorphism affects the ability of the different MHC molecules to bind antigenic peptides.

AN INDIVIDUAL'S MHC HAPLOTYPE AFFECTS SUSCEPTIBILITY TO DISEASE

Genetic variations in MHC molecules affect:

- the ability to make immune responses, including the level of antibody production;
- resistance or susceptibility to infectious diseases;
- resistance or susceptibility to autoimmune diseases and allergies.

Knowing this, we can start to answer the question why the MHC is so polymorphic. The immune system must handle many different pathogens. By having several different MHC molecules, an individual can present a diverse range of antigens and is therefore likely to be able to mount an effective immune response. There is therefore a selective advantage in having different MHC molecules.

Going beyond this, we know that different pathogens are prevalent in different areas of the world, so evolutionary pressures from pathogens will tend to select for different MHC molecules in different regions.

Q. The haplotype HLA-B53 is associated with protection against childhood malaria, a disease that is prevalent in equatorial regions. In which country would you expect to find the highest frequency of the HLA-B53 allele – China, Ghana, or South Africa?

A. The gene frequency is around 40% in Ghana and 1–2% in China and South Africa, which are outside the equatorial regions affected by malaria.

All nucleated cells of the body express MHC class I molecules

The function of MHC class I molecules is to present antigens that have entered the cell, such as viral peptides. Because any cell of the body may become infected with a virus or intracellular pathogen, all cells need to sample their internal molecules and present them at the cell surface to cytotoxic T cells.

By contrast, MHC class II molecules are used by APCs to present antigens to helper T cells. Consequently the distribution of class II molecules is much more limited (see Fig. 7.4).

MHC molecules are co-dominantly expressed

This means that, in one individual, all of the principal MHC gene loci are expressed from both the maternal and paternal chromosomes. As there are three MHC class I loci in humans (HLA-A, HLA-B, and HLA-C), each of which is highly polymorphic, most individuals will have genes for six different class I molecules, all of which will be present at the cell surface. Each MHC molecule will have:

- a slightly different shape; and
- present a different set of antigenic peptides.

A similar logic applies to MHC class II molecules. There are three principal class II loci in humans (HLA-DP, HLA-DQ, and HLA-DR), all of which are polymorphic. At first sight, it would appear that an APC could express six different class II molecules as well as its class I molecules. However, this is probably an underestimate. As noted above, hybrid class II molecules (using one polypeptide encoded by the maternal chromosome and one by the paternal chromosome) also occur.

The specificity of the TCR and MHC explains genetic restrictions in antigen presentation

Much of the original work on antigen presentation was carried out using strains of mice that had been inbred to the point where both maternal and paternal chromosomes were identical. Any offspring therefore inherited the same set of autosomes from each parent, and the offspring were genetically identical to their parents. Clearly the level of diversity in the MHC molecules was much less than in an outbred human population.

The artificial simplicity of the inbred mouse system, however, allowed immunologists to dissect how antigens were presented to T cells in a whole animal, when the molecular structures of the MHC molecules and the TCR were completely unknown.

The key experiment that demonstrated the importance of the MHC in antigen presentation revealed a phenomenon called **genetic restriction** (also known as **MHC restriction**). In essence, it was noted that cytotoxic T cells from a mouse infected with a virus are primed to kill cells of the same H-2 haplotype infected with that virus; they do not kill cells of a different haplotype infected with the same virus (Fig. 5.20).

These data, and similar experiments using APCs and helper T cells, showed that T cells that have been primed to recognize antigen presented on MHC molecules of one haplotype will normally respond again only when they see the same antigen on the same MHC molecule.

Q. Interpret these findings in relation to the way in which T cells recognize antigen.

A. The TCR interacts with residues from both the antigenic peptide and the associated MHC molecule. In other words, the T cell recognizes the specific combination of MHC molecule plus peptide.

Peter Doherty and Rolf Zinkernagel performed the key experiments delineating the phenomenon of MHC restriction of T cell responses in the mid 1970s and were awarded the Nobel Prize in Physiology or Medicine in 1996 for this work.

CD1 IS AN MHC CLASS 1-LIKE MOLECULE THAT PRESENTS LIPID ANTIGENS

CD1 molecules are structurally related to MHC class I molecules and are non-covalently bound to β_2-microglobulin.

MHC restriction of cytotoxic T cells

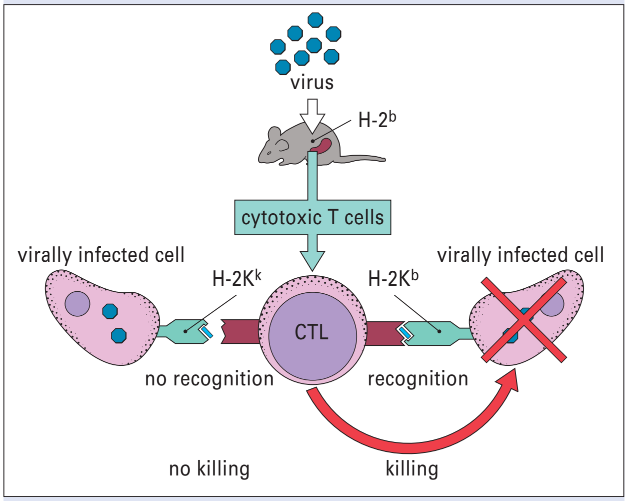

Fig. 5.20 A mouse of the H-2^b haplotype is primed with virus and the cytotoxic T cells thus generated are isolated and tested for their ability to kill H-2^b and H-2^k cells infected with the same virus. The cytotoxic T cells kill H-2^b, but not H-2^k cells. In this instance, it is the H-2K class I gene product presenting the antigen to the T cells. The T cell is recognizing a specific structure produced by the association of a specific MHC molecule with a specific viral antigen.

The genes encoding CD1 molecules are located outside the MHC and are not polymorphic. In humans, they consist of five closely linked genes, of which four are expressed (Fig. 5.21), encoding proteins that fall into two separate groups:

- group 1 molecules in humans include CD1a, CD1b, and CD1c;
- CD1d proteins form the second group.

Murine CD1b has been crystallized and analyzed by X-ray crystallography. This shows that the molecule has a deep electrostatically neutral antigen-binding groove, which is highly hydrophobic and can accommodate lipid or glycolipid antigens (Fig. 5.22). One model for the binding places hydrophobic acyl groups of the lipids into the large hydrophobic pockets, leaving the more polar groups of the antigens such as phosphate and carbohydrate on the top, where they can interact with the TCR.

The binding requirements of the hydrophobic pockets on CD1 are fairly tolerant because they will accommodate acyl groups of different lengths, but the interactions with the TCR are much more specific – small changes in the structure of the carbohydrate moiety will destroy the ability to stimulate a T cell.

The antigens presented by the group I CD1 molecules and CD1d are different. For example, group I molecules present lipoarabinomannan, a component of the cell wall of mycobacteria (see Fig. 14.1), whereas CD1d cannot do this.

Another difference between CD1 and conventional MHC molecules is the way in which antigen is loaded into the antigen-binding groove:

- MHC class I molecules are loaded with antigenic peptides in the endoplasmic reticulum, and this requires transport of the peptides from the cytoplasm (see Chapter 7).
- group I CD1 molecules appear to be loaded in an acidic endosomal compartment because they do not bind to lipid antigens unless they are partially unfolded at low pH.

There is some debate about the physiological functions of the CD1 molecules in host defense:

- group I CD1 molecules present lipids from mycobacteria and *Haemophilus influenzae* and can stimulate both CD4$^+$ and CD8$^+$ cytotoxic T cells, and therefore appear to have a role in antimicrobial defense.

Human CD1 genes

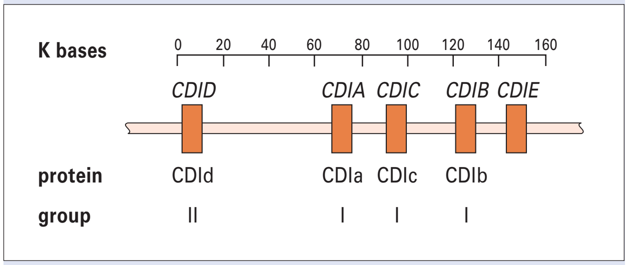

Fig. 5.21 The genes of the human CD1 cluster extend over 160 kilobases on chromosome 1. A gene product for *CD1E* has not yet been identified.

Glycolipid antigens presented by CD1

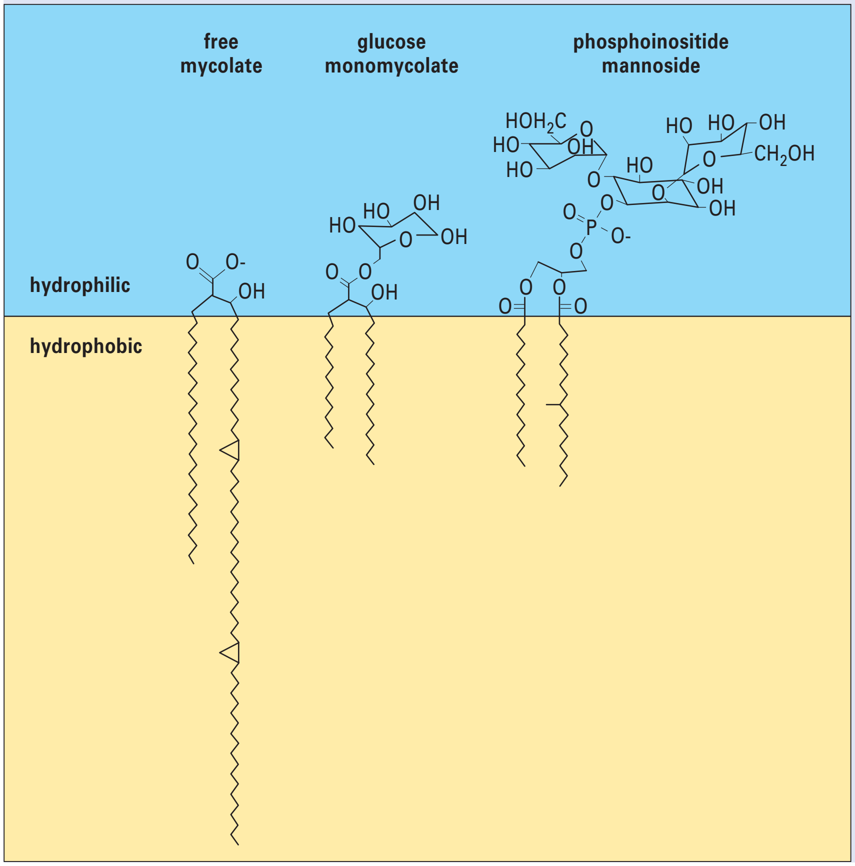

Fig. 5.22 Some of the antigens presented by CD1b are shown, each of which is a component of mycobacterial cell walls. Each of the antigens has two aliphatic tails, which are thought to be buried in the two hydrophobic binding pockets on CD1b. This leaves the hydrophilic segments exposed in the center of the binding groove, where they can contact the TCR. (Information based on Porcelli SA, Segelke BW, Sugita M, et al. Immunol Today 1998;19:362–368)

- most CD1d molecules appear to bind self antigens, though they also present lipids from parasites such as *Plasmodium falciparum* and *Trypanosoma brucei* (see Chapter 15) to T cells that use a restricted group of TCRs, indicating a role in defense against single-celled protozoal parasites.

FURTHER READING

Bentley GA, Mariuzza RA. The structure of the T cell antigen receptor. Annu Rev Immunol 1996;14:563–590.

Bjorkman PJ, Parham P. Structure, function and diversity of class I major histocompatibility complex molecules. Annu Rev Biochem 1990;59:253–288.

Bjorkman PJ, Saper MA, Samraoui B, et al. The structure of the human class I histocompatibility antigen HLA-A2. Nature 1987;329:506–512.

Bjorkman PJ, Samraoui B, Bennett WS, et al. The foreign antigen binding site and T-cell recognition regions of class I histocompatibility antigens. Nature 1987;329:512–516.

Bodmer JG, Marsh, SE, Albert ED, et al. Nomenclature for factors of the HLA system, 1994. Tissue Antigens 1994;44:1–18.

Brenner MB, MacLean J, Dialynas DP, et al. Identification of a putative second T-cell receptor. Nature 1986;322:145–149.

Brown JH, Jardetzky TS, Gorga JC, et al. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 1993;364:33–39.

Burdin N, Kronenberg M. CD1-mediated immune responses to glycolipids. Curr Opin Immunol 1999;111:326–331.

Carosella ED, Dausett J, Kirzenbaum H. HLA-G revisited. Immunol Today 1996;17:407–409.

Chien Y-H, Jores R, Crowley MP. Recognition by γδ T cells. Annu Rev Immunol 1996;14:511–532.

Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. Nature 1988;334:395–402.

Davis MM, Boniface JJ, Reich Z, et al. Ligand recognition by αβ T cells. Annu Rev Immunol 1998;16:523–544.

Garcia KC, Degano M, Stanfield RL, et al. An αβ T cell receptor structure at 2.5Å and its orientation in the TCR–MHC. Complex Sci 1996;274:209–219.

Garcia KC, Teyton GL, Wilson IA. Structural basis of T cell recognition. Annu Rev Immunol 1999;17:369–398.

Garratt TPJ, Saper MA, Bjorkman PJ, et al. Specificity pockets for the side chains of peptide antigens in HLA-w68. Nature 1989;342:692–696.

Hass W, Pereira P, Tonegawa S. Gamma/delta cells. Annu Rev Immunol 1993;11:637–685.

Hayday AC. γδ cells: a right time and a right place for a conserved third way of protection. Annu Rev Immunol 2000;18:975–1026.

Hunkapiller T, Hood L. Diversity of the immunoglobulin gene superfamily. Adv Immunol 1989;44:1–63.

Lefranc M-P, Rabbitts TH. The human T-cell receptor γ (TRG) genes. Trends Biochem Sci 1989;14:214–218.

Leiden JM. Transcriptional regulation of T cell receptor genes. Annu Rev Immunol 1993;11:539–570.

Madden DR, Gorga JC, Strominger L, et al. The structure of

HLA-B27 reveals nonamer self-peptides bound in an extended conformation. Nature 1991;353:321–325.

Manning TC, Kranz DM. Binding energetics of T-cell receptors: correlation with immunological consequences. Immunol Today 1999;20:417–422.

Porcelli SA, Segelke BW, Sugita M, et al. The CD1 family of lipid antigen-presenting molecules. Immunol Today 1998;19:362–368.

Powis SH, Trowsdale J. Human major histocompatibility complex genes. Behring Inst Mitt 1994;94:17–25.

Raulet DH. How γδ T cells make a living. Curr Biol 1994;4:246–251.

Roth DB. Generating antigen receptor diversity. Lewin B, ed. www.ergito.com Immunology module of Virtual Text. 2005: Chapter 5.

Salter RD, Benjamin RJ, Wesley PK, et al. A binding site for the T-cell co-receptor CD8 on the α_3 domain of HLA-A2. Nature 1990;345:41–46.

San José E, Sahuquillo AG, Bragado R, Alarcón B. Assembly of the TCR/CD3 complex: CD3 epsilon/delta and CD3 epsilon/gamma dimers associate indistinctly with both TCR alpha and TCR beta chains. Evidence for a double TCR heterodimer model. Eur J Immunol 1998;28:12–21.

Sloan-Lancaster J, Allen PM. Altered peptide–ligand induced partial T cell activation: molecular mechanisms and role in T cell biology. Annu Rev Immunol 1996;14:1–27.

Stern LJ, Wiley DC. Antigenic peptide binding by class I and class II histocompatibility proteins. Structure 1994;2:245–251.

Stern LJ, Brown JH, Jardetzky TS, et al. Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. Nature 1994;368:215–221.

Weiss A, Littman DR. Signal transduction by lymphocyte antigen receptor. Cell 1994;76:263–274.

Zeng ZH, Castaño LH, Segelke B, Stura EA, Peterson PA, Wilson IA. The crystal structure of mouse CD1: an MHC-like fold with a large hydrophobic binding groove. Science 1997;277:339–345.

Critical thinking: Somatic hypermutation (see p. 494 for explanation)

1 Can you think of reasons why TCRs do not undergo somatic hypermutation?

The specificity of T cells (see p. 494 for explanations)

SM/J mice of the haplotype H-2^{v} were immunized with the λ repressor protein, a molecule with 102 amino acid residues. After 1 week, T cells were isolated from the animals and set up in culture with APCs and antigen. The ability of APCs to activate the T cells was determined in a lymphocyte proliferation assay (see Method Box 7.2, p. 160).

It was found that when APCs from SM/J mice were used in the culture the T cells were activated, but that when APCs from Balb/c mice were used (H-2^{d}) they were not activated. APCs from F1 (SM/J•Balb/c) mice were able to activate the T cells just as well as the APCs from the parental SM/J strain.

2 Explain why the SM/J cells and the F1 cells can present antigen to the T cells, but the Balb/c cells cannot.

Using the primed T cells from the SM/J mouse and APCs from the same strain, the investigation continues, but peptides of the λ repressor protein are used instead of intact antigen. It is found that a peptide corresponding to residues 80–94 of the intact protein is able to stimulate the T cells, but that other peptides are much less effective or ineffective. The table below shows the sequences of some of these peptides and their ability to activate the T cells when included in the culture at a concentration of 10 μM.

peptide	amino acid sequence	T cell activation
12–36	QLEDARRLKAIYEKKKNELGLSQESV	–
80–102	SPSIAREIYEMYEAVSMQPSLRS	+++
73–88	ILKVSVEEFSPSIAREIY	–
80–94	SPSIARE**I**YEMYEAVS	++
84–98	AREIYEMYEAVSMQP	–

3 Explain why peptides 80–102 and 80–94 activate the T cells while the others do not.

In a final experiment the T cells are stimulated with a mutated variant of peptide 80–94 with aspartate (D) substituted for isoleucine (I) at position 87 (bold type). It is found that the mutated peptide is able to stimulate the T cells as well as the original peptide, even when present at lower concentrations (1 μM).

4 What term is used to describe this kind of mutated peptide? What would you predict about the binding affinity of this peptide within the TCR–MHC–peptide complex?

Modes of Immune Response

SECTION

2

This Page Intentionally Left Blank

CHAPTER 6

Mechanisms of Innate Immunity

SUMMARY

- **Innate immune responses do not depend on immune recognition by lymphocytes**, but have co-evolved with and are functionally integrated with the adaptive elements of the immune system.

- **The body's responses to damage include inflammation, phagocytosis, and clearance of debris and pathogens, and remodeling and regeneration of tissues.** Inflammation is a response that brings leukocytes and plasma molecules to sites of infection or tissue damage.

- **The phased arrival of leukocytes in inflammation depends on chemokines and adhesion molecules expressed on the endothelium.** Adhesion molecules fall into families that are structurally related. They include the cell adhesion molecules (CAMs) of the immunoglobulin supergene family (which interact with leukocyte integrins), and the selectins (which interact with carbohydrate ligands).

- **Leukocyte migration to lymphoid tissues is also controlled by chemokines.** Chemokines are a large group of signaling molecules that initiate chemotaxis and/or cellular activation. Most chemokines act on more than one receptor, and most receptors respond to more than one chemokine.

- **Plasma enzyme systems modulate inflammation and tissue remodeling.** The kinin system and mediators from mast cells including histamine contribute to the enhanced blood supply and increased vascular permeability at sites of inflammation.

- **Pathogen-associated molecular patterns (PAMPs) are distinctive biological macromolecules that can be recognized by the innate immune system.** Innate antimicrobial defenses include molecules of the collectin, ficolin, and pentraxin families, which can act as opsonins, either directly or by activating the complement system. Macrophages also have surface lectins, which allow them to directly bind to pathogens. The Toll-like receptors recognize various PAMPs and cause macrophage activation. Their signaling systems and actions are closely related to those used by inflammatory cytokines TNFα and IL-1.

- **Microbicidal proteins are part of the innate immune system.** Many cell types synthesize and secrete antimicrobial proteins.

INNATE IMMUNE RESPONSES DO NOT DEPEND ON IMMUNE RECOGNITION BY LYMPHOCYTES

The immune system deals with pathogens by means of a great variety of different types of immune response, but these can be broadly divided into:

- adaptive responses; and
- innate immunity.

The adaptive immune responses depend on the recognition of antigen by lymphocytes, a cell type that has evolved relatively recently – lymphocytes are present in all vertebrates, but not invertebrates, though lymphocyte-like cells are present in closely related phyla, including the tunicates and echinoderms (Fig. 6.1).

Q. What are the two key characteristics of adaptive immune responses?

A. They display a high level of specificity for the particular pathogen, and the responses show long-lasting memory.

Electron micrographs of lymphocyte-like cells

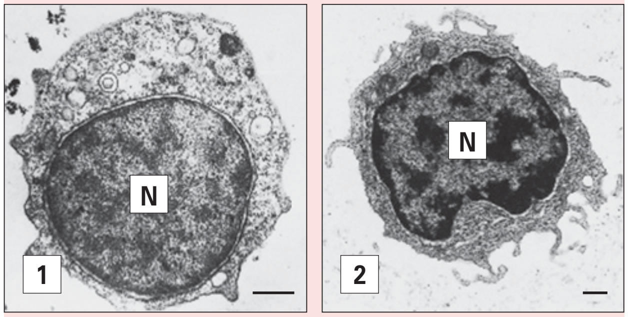

Fig. 6.1 Electron micrographs of lymphocyte-like cells from the tunicate *Ciona intestinalis* (**1**), and from a fish, the blenny, *Blennius pholis* (**2**). Note the similar morphology – both cells have a large nucleus and a thin rim of undifferentiated cytoplasm. Scale bar 0.5 µm. (Courtesy of Dr AF Rowley from Endeavour 1989:13;72–77. Copyright 1989 with permission from Elsevier)

Before the evolution of lymphocytes, and the emergence of specific antigen receptors – antibodies (see Chapter 3) and the T cell receptor (TCR, see Chapter 5) – we can infer that a number of different types of immune defense were already present in precursor organisms. Many of these systems have been retained in vertebrates and have continued to evolve alongside the adaptive immune system. Hence, in present-day mammals we see an integrated immune system in which different types of defense work in concert.

In reality it is quite artificial to try to segregate adaptive and innate immune responses. For example a macrophage:

- displays the very primitive immune defense of phagocytosis; but also
- expresses MHC molecules and acts as an antigen-presenting cell, a function that makes sense only in relation to the evolution of T cells.

We can identify some of the ancient innate immune defense systems because related systems are seen in distant phyla. For example, the family of **Toll-like receptors** (**TLRs**, see p. 143) present on mammalian macrophages and involved in phagocytosis were first identified in insects. We can therefore infer that the distant ancestor of mammals and insects had a receptor molecule of this type that probably recognized microbial components.

Q. Why would it be a mistake to think that the immune system seen in insects or worms was the precursor of the immune system seen in present-day vertebrates.

A. Both have been separately evolved for millions of years. The immune systems of worms and insects have developed to cope with the pathogens that they encounter in the context of their life cycles.

Having stated how the functional distinction between adaptive and immune systems is essentially artificial, this chapter outlines some of the immune defenses that do not depend on immune recognition by lymphocytes.

THE BODY'S RESPONSES TO DAMAGE INCLUDE INFLAMMATION

The body's response to tissue damage depends on:

- what has caused the damage;
- its location; and
- its severity.

In many cases damage can be caused by physical means, and does not involve infection or an adaptive immune response.

However, if an infection is present, the body's innate systems for limiting damage and repairing tissues work in concert with the adaptive immune responses. The overall process involves a number of overlapping stages, which typically take place over a number of days or weeks. These may include some or all of the following:

- stopping bleeding;
- acute inflammation;
- killing of pathogens, neutralizing toxins, limiting pathogen spread;
- phagocytosis of debris, pathogens, and dead cells;
- proliferation and mobilization of fibroblasts or other tissue cells to contain an infection and/or repair damage;
- removal or dissolution of blood clots and remodeling of the components of the extracellular matrix;
- regeneration of cells of the tissue and re-establishing normal structure and function.

Inflammation brings leukocytes and plasma molecules to sites of infection or tissue damage

Many immune responses lead to the complete elimination of a pathogen (sterile immunity), followed by resolution of the damage, disappearance of leukocytes from the tissue and full regeneration of tissue function – the response in such cases is referred to as **acute inflammation**.

Q. What three principal changes occur in the tissue during an acute inflammatory response?

A. An increased blood supply to the affected area, an increase in capillary permeability allowing larger serum molecules to enter the tissue, and an increase in leukocyte migration into the tissue (see Chapter 1).

In some cases an infection is not cleared completely. Most pathogenic organisms have developed systems to deflect the immune responses that would eliminate them. In this case the body often tries to contain the infection or minimize the damage it causes; nevertheless, the persistent antigenic stimulus and the cytotoxic effects of the pathogen itself lead to ongoing **chronic inflammation**.

The cells seen in acute and chronic inflammation are quite different, and reflect the phased arrival of different populations of leukocytes into a site of infection (Fig. 6.2). Consequently:

- sites of acute inflammation tend to have higher numbers of neutrophils and activated helper T cells; whereas
- sites of chronic inflammation have a higher proportion of macrophages, cytotoxic T cells, and even B cells.

The phased arrival of different populations of leukocytes at a site of inflammation is dependent on **chemokines** expressed on the endothelium (see below). These chemokines activate distinct leukocyte populations causing them to migrate into the tissue.

The cell types seen in sites of damage and the capacity of the tissue for repair and regeneration also depend greatly on the tissue involved. For example, in the brain the capacity for cell regeneration is very limited, so in chronic inflammatory diseases, such as multiple sclerosis, the area of damage often becomes occupied by scar tissue formed primarily by a specialized CNS cell type, the astrocyte.

The following sections explain the general principles of how inflammation develops, though the specific details depend on:

- the type of infection;
- the tissue; and
- the immune status of the individual.

The phased arrival of different populations of leukocytes into a site of infection

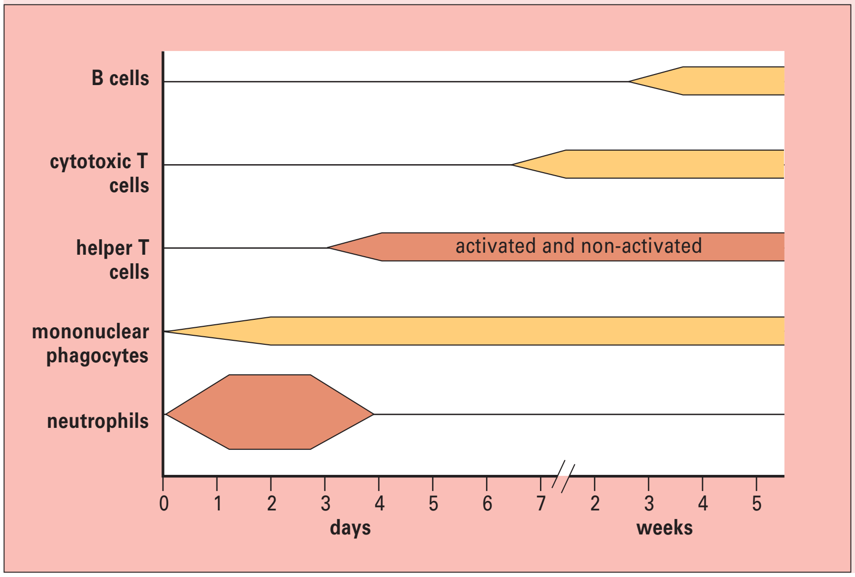

Fig. 6.2 Leukocytes enter sites of infection in phases. Sites of chronic inflammation have more macrophages and T cells.

Cytokines control the movement of leukocytes into tissues

Tissue damage leads to the release of a number of inflammatory cytokines, either from:

- patrolling leukocytes; or
- cells within the tissue, including resident mononuclear phagocytes.

The cytokine **tumor necrosis factor-α (TNFα)** is particularly important in this respect.

TNFα has many functions in the development of inflammation

TNFα is produced primarily by macrophages and other mononuclear phagocytes and has many functions in the development of inflammation and the activation of other leukocytes (Fig. 6.3).

Notably, TNFα induces the adhesion molecules and chemokines on the endothelium (which are required for the accumulation of leukocytes), and activates the microbicidal systems of phagocytes. In addition TNFα can induce **apoptosis** in susceptible cells.

TNFα and the related cytokines, the **lymphotoxins**, act on a family of receptors. One of the effects of these cytokines is to cause the activation of the transcription factor **NF-κB** (Fig. 6.4), which has been described as a master-switch of the immune system (see Fig. 9.21).

The signaling pathways that lead to NF-κB activation are evolutionarily very ancient, and are found in species as diverse as fruit flies, humans, and sea urchins.

IL-1 has wide-ranging functions that switch the body into a state to combat disease

Interleukin-1 (IL-1) is another important inflammatory cytokine. It is produced by many cell types in the body and shares some functions with TNFα, though both TNFα and IL-1 and their receptors are quite different. For example, IL-1 induces adhesion molecules on endothelium that promote leukocyte migration.

IL-1 also has a number of distinct functions, of which the most important is the induction of fever. IL-1 acts directly on centers in the hypothalamus that control the body's temperature.

It is thought that the increase in temperature that accompanies inflammation in mammals gives the immune system a slight edge on the pathogens. A small increase in body temperature leads to much faster lymphocyte division, which means that the adaptive immune system (which relies on rapid expansion of pathogen-specific clones) can be mobilized more quickly.

IL-1 has an interesting variety of other constitutional effects. For example, it suppresses appetite and promotes one type of sleep. Another pro-inflammatory cytokine – interleukin-6 (IL-6), which is produced in the liver and by vascular endothelium in inflammation – also produces these constitutive symptoms, but this effect is due to the secondary induction of IL-1.

Taken together, we can see that in addition to its local effects in inflammation, IL-1 has wide-ranging functions that switch the body into a state to combat disease.

Interferons and NK cells delay viral spread

Viruses and many bacteria can replicate at an enormous rate and potentially have the capacity to overwhelm the individual before the adaptive immune system gets going – it takes several days to activate and expand clones of antigen-specific lymphocytes. During this time **interferons** (see Chapter 13, p. 247) and **NK cells** (see Chapter 10, p. 204) are particularly important in slowing the spread of infection:

- interferons act by inducing antiviral proteins, which limit the ability of a virus to replicate within a cell;
- NK cells have some capacity to recognize and kill virus-infected cells.

TNFα is a cytokine with many functions

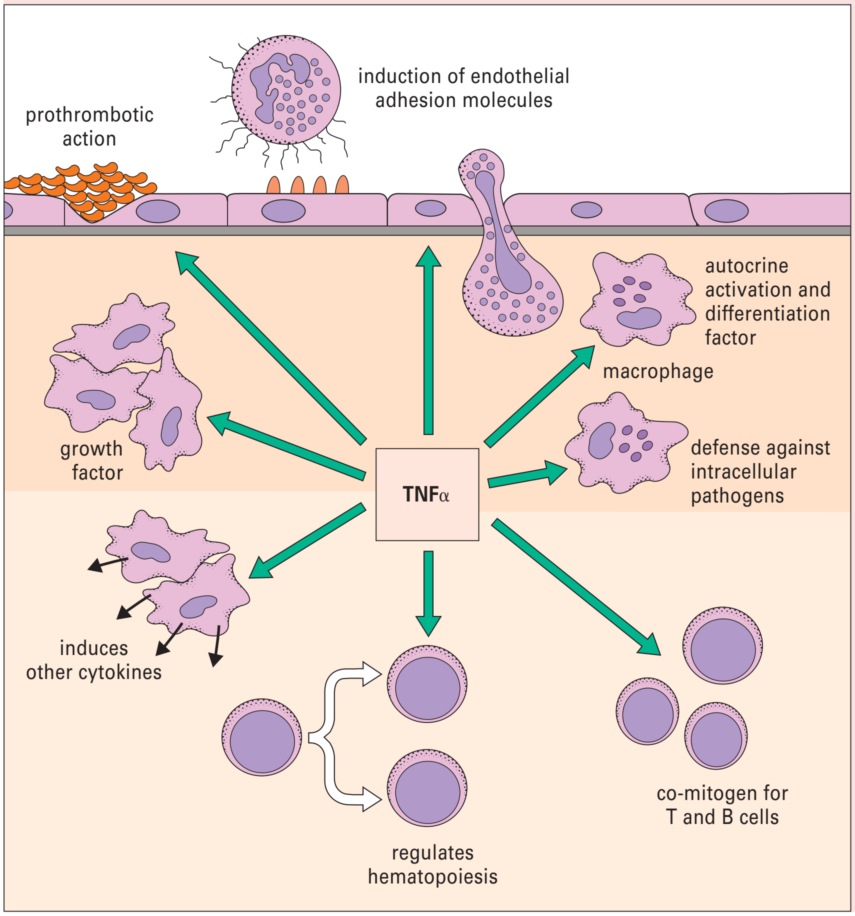
prothrombotic action
induction of endothelial adhesion molecules
autocrine activation and differentiation factor
macrophage
growth factor
TNFα
defense against intracellular pathogens
induces other cytokines
regulates hematopoiesis
co-mitogen for T and B cells

Fig. 6.3 TNFα has several functions in inflammation. It is prothrombotic and promotes leukocyte adhesion and migration (top). It has an important role in the regulation of macrophage activation and immune responses in tissues (center), and it also modulates hematopoiesis and lymphocyte development (bottom).

THE PHASED ARRIVAL OF LEUKOCYTES IN INFLAMMATION DEPENDS ON CHEMOKINES AND ADHESION MOLECULES

The mechanisms that control leukocyte migration into inflamed tissues have been carefully studied because of their biological and medical importance. These mechanisms are also applicable in principle to the cell movement that occurs between lymphoid tissues as the immune system develops and during normal life.

The routes that leukocytes take as they move around the body are determined by interactions between:

- circulating cells; and
- the endothelium of blood vessels.

Leukocyte migration is controlled by signaling molecules, which are expressed on the surface of the endothelium, and occurs principally in venules (Fig. 6.5). There are three reasons for this:

- the signaling molecules and adhesion molecules that control migration are selectively expressed in venules;
- the hemodynamic shear force in the venules is relatively low, and this allows time for leukocytes to receive signals from the endothelium and allows adhesion molecules on the two cell types to interact effectively;
- the endothelial surface charge is lower in venules (Fig. 6.6).

Although the patterns of leukocyte migration are complex, the basic mechanism appears to be universal. The initial interactions are set out in a three-step model (Fig. 6.7):

- leukocytes are slowed as they pass through a venule and roll on the surface of the endothelium before being halted – this is mediated primarily by adhesion molecules called **selectins** interacting with carbohydrates on glycoproteins;
- the slowed leukocytes now have the opportunity to respond to signaling molecules held at the endothelial surface – particularly important is the large group of cytokines called **chemokines,** which activate particular populations of leukocytes expressing the appropriate chemokine receptors;
- activation upregulates the affinity of the leukocytes' **integrins**, which now engage the cellular adhesion molecules on the endothelium to cause firm adhesion and initiate a program of migration.

Transendothelial migration is an active process involving both leukocytes and endothelial cells and occurs near the junctions between endothelial cells (Fig. 6.8).

There has been much controversy concerning whether leukocytes migrate through the junctions between endothelial cells or across the endothelium itself. In specialized tissues such as the brain and thymus, where the

Intracellular signaling pathways induced by TNFα

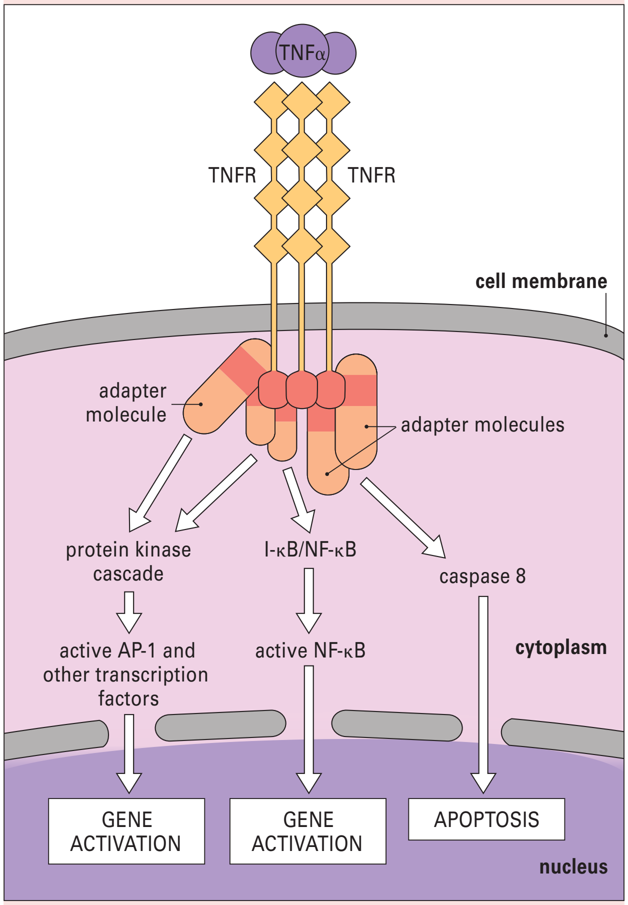
TNFα
TNFR
TNFR
cell membrane
adapter molecule
adapter molecules
protein kinase cascade
I-κB/NF-κB
caspase 8
active AP-1 and other transcription factors
active NF-κB
cytoplasm
GENE ACTIVATION
GENE ACTIVATION
APOPTOSIS
nucleus

Fig. 6.4 TNFα induces the trimerization of the TNF receptor (TNFR) on the cell surface, which causes adapter molecules to be recruited to the receptor complex. One pathway leads to the activation of caspase 8 and apoptosis. Other pathways lead to the activation of transcription factors AP-1 and NF-κB, which activate many genes involved in adaptive and innate immune responses.

endothelium is connected by continuous tight junctions, it is clear that lymphocytes migrate across the endothelium in vacuoles and that the junctions do not break apart.

Migrating cells extend pseudopods down to the basement membrane and move beneath the endothelium using new sets of adhesion molecules. Enzymes are now released that digest the collagen and other components of the basement membrane, allowing cells to migrate into the tissue. Once there, the cells can respond to new sets of chemotactic stimuli, which allow them to position themselves appropriately in the tissue.

Leukocyte traffic into tissues is determined by where and when adhesion and signaling molecules are expressed

Intercellular adhesion molecules are membrane-bound proteins that allow one cell to interact with another. Often these molecules traverse the membrane and are linked to the cytoskeleton.

Leukocytes adhering to the wall of a venule

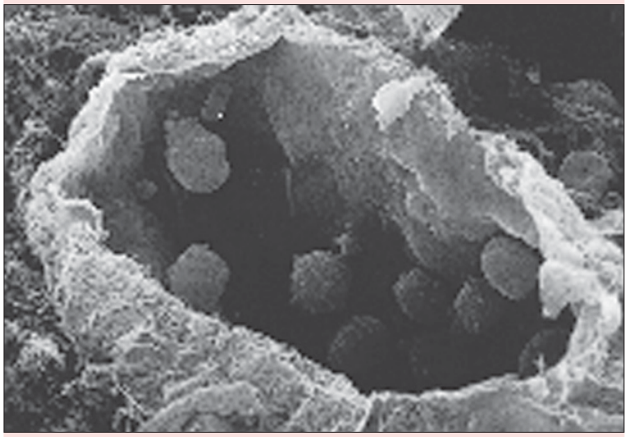

Fig. 6.5 Scanning electron micrograph showing leukocytes adhering to the wall of a venule in inflamed tissue. ×16 000. (Courtesy of Professor MJ Karnovsky)

Leukocyte migration across endothelium

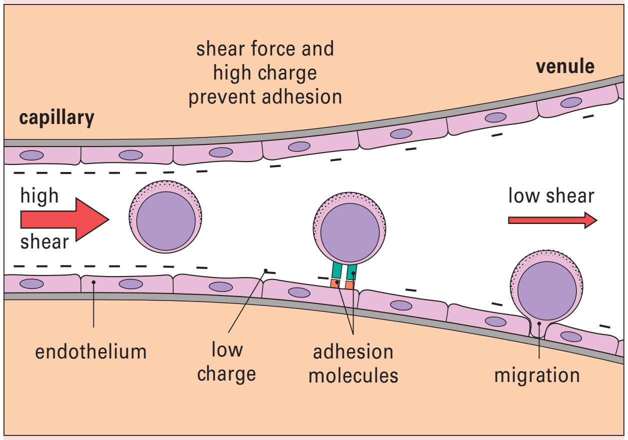
shear force and high charge prevent adhesion
venule
capillary
high shear
low shear
endothelium
low charge
adhesion molecules
migration

Fig. 6.6 Leukocytes circulating through a vascular bed may interact with venular endothelium via sets of surface adhesion molecules. In the venules, hemodynamic shear is low, surface charge on the endothelium is lower, and adhesion molecules are selectively expressed.

Q. Why is it important that cell adhesion molecules (CAMs) interact with the cytoskeleton?

A. By binding to the cytoskeleton the adhesion molecules allow a cell to gain traction on another cell or on the extracellular matrix, which allows the cell to move through tissues.

In many cases, a particular adhesion molecule can bind to more than one ligand, using different binding sites. Although the binding affinity of individual adhesion molecules to their ligands is usually low, clustering of the molecules in patches on the cell surface, means that the avidity of the interaction can be high.

Three-step model of leukocyte adhesion

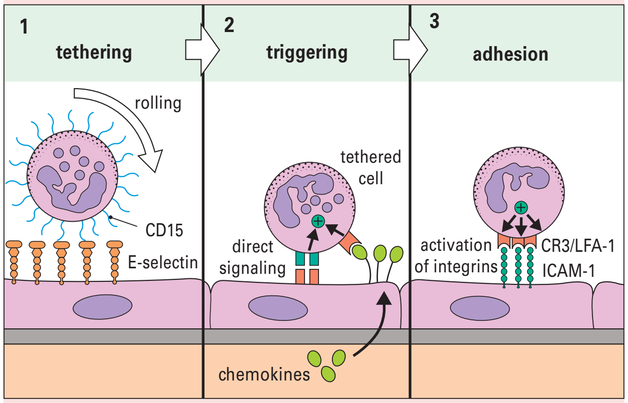

Fig. 6.7 The three-step model of leukocyte adhesion and activation is illustrated by a neutrophil, though different sets of adhesion molecules would be used by other leukocytes in different situations. (**1**) Tethering – the neutrophil is slowed in the circulation by interactions between E-selectin and carbohydrate groups on CD15, causing it to roll along the endothelial surface. (**2**) Triggering – the neutrophil can now receive signals from chemokines bound to the endothelial surface or by direct signaling from endothelial surface molecules. The longer a cell rolls along the endothelium, the longer it has to receive sufficient signal to trigger migration. (**3**) Adhesion – the triggering upregulates integrins (CR3 and LFA-1) so that they bind to ICAM-1 induced on the endothelium by inflammatory cytokines.

Lymphocyte migration

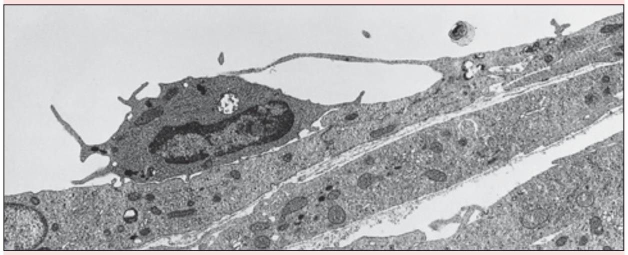

Fig. 6.8 Electron micrograph showing a lymphocyte adhering to brain endothelium close to the interendothelial cell junction in an animal with experimental allergic encephalomyelitis. Adhesion precedes transendothelial migration into inflammatory sites. (Courtesy of Dr C Hawkins)

Cells can modulate their interactions with other cell types by increasing the numbers of adhesion molecules on the surface or altering their affinity and avidity. They can alter the level of expression of adhesion molecules in two ways:

- many cells retain large intracellular stores of these molecules in vesicles, which can be directed to the cell surface within minutes following cellular activation;
- alternatively, new molecules can be synthesized and transported to the cell surface, a process that usually takes several hours.

Selectins bind to carbohydrates to slow the circulating leukocytes

The selectins include the molecules:

- E-selectin and P-selectin, which are expressed predominantly on endothelium and platelets; and
- L-selectin, which is expressed on some leukocytes (Fig. 6.9).

Selectins are transmembrane molecules with a number of extracellular domains homologous to those seen in complement control proteins (see Fig. 4.8). The extracellular region also has a domain related to the epidermal growth factor receptor and an N terminal domain that has lectin-like properties (i.e. it binds to carbohydrate residues), hence the name selectins.

The carbohydrate ligands for the selectins may be associated with several different proteins:

- at sites of inflammation, E-selectin and P-selectin, which are induced on activated endothelium, bind to the sialyl Lewis-X carbohydrate associated with CD15, present on many leukocytes;
- some of the selectin ligands are selectively expressed on particular populations of leukocytes, for example the molecule PSGL-1 (P-selectin glycoprotein ligand) present on TH1 cells binds to E- and P-selectin, but a variant found on TH2 cells does not.

When selectins bind to their ligands the circulating cells are slowed within the venules. Video pictures of cell migration show that the cells stagger along the endothelium. During this time the leukocytes have the opportunity of receiving migration signals from the endothelium. This is a process of signal integration – the

Selectins

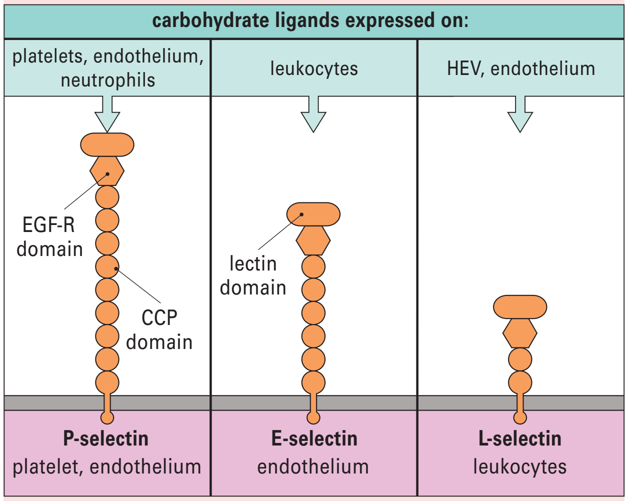

Fig. 6.9 The structures of three selectins are shown. They have terminal lectin domains, which bind to carbohydrates on the cells listed. The EGF-R domain is homologous to a segment in the epidermal growth factor receptor. The CCP domains are homologous to domains found in complement control proteins, such as factor H, decay accelerating factor, and membrane cofactor protein (see Fig. 4.8).

more time the cell spends in the venule, the longer it has to receive sufficient signals to activate migration. If a leukocyte is not activated it detaches from the endothelium and returns to the venous circulation. A leukocyte may therefore circulate hundreds of times before it finds an appropriate place to migrate into the tissues.

Chemokines and other chemotactic molecules trigger the tethered leukocytes

Chemokines are small cytokines

The chemokines are a group of at least 40 small cytokines involved in cell migration, activation, and chemotaxis. They determine which cells will cross the endothelium and where they will move within the tissue. Most chemokines have two binding sites:

- one for their specific receptors; and
- a second for carbohydrate groups on proteoglycans (such as heparan sulfate), which allows them to attach to the luminal surface of endothelium (blood side), ready to trigger any tethered leukocytes (Fig. 6.10).

The chemokines may be produced by the endothelium itself. This depends on several factors including:

- the tissue;
- the presence of inflammatory cytokines; and
- hemodynamic forces.

In addition chemokines produced by cells in the tissues can be transported to the luminal side of the endothelium. Immune reactions or events occurring within the tissue can therefore induce the release of chemokines, which signal the inward migration of populations of leukocytes.

Q. What advantage is there in having different types of inflammation occurring in different tissues?

A. What constitutes an appropriate immune response depends on the pathogen, the amount of damage it is causing in a particular tissue, and the capacity of that tissue to repair and regenerate.

It is easiest to understand what chemokines do by considering their receptors

Chemokines fall into four different families, based on the spacing of conserved cysteine (C) residues. For example:

- α-chemokines have a CXC structure; and
- β-chemokines a CC structure.

All chemokines act via receptors that have seven transmembrane segments (**7tm receptors**) linked to GTP-binding proteins (**G-proteins**), which cause cell activation.

Most chemokines act on more than one receptor, and most receptors will respond to several chemokines. Because of this complexity, it is easiest to understand what chemokines do by considering their receptors:

- the receptors for the CXC chemokines are called CXCR1, CXCR2, and so on; while
- the receptors for the CC chemokines are called CCR1, CCR2, etc.

Until recently most chemokines had a descriptive name and acronym such as macrophage chemotactic protein-1 (MCP-1). The current nomenclature describes them according to their type, hence MCP-1 is CCL2, meaning that it is a **ligand** for the CC family of chemokine receptors (Fig. 6.11). A full list of chemokines is given in Appendix 4.

The chemokine receptors are selectively expressed on particular populations of leukocytes (see Fig. 6.11) and this determines which cells can respond to signals coming from the tissues.

The profile of chemokine receptors on a cell depends on its type and state of differentiation. For example:

- all T cells express CCR1;
- TH2 cells preferentially express CCR3; and
- TH1 cells preferentially express CCR5 and CXCR3.

After activation in lymph nodes, the levels of CXCR3 on a T cell increase, so that it becomes more responsive for the chemokines CXCL9, CXCL10, and CXCL11, which

Chemokines

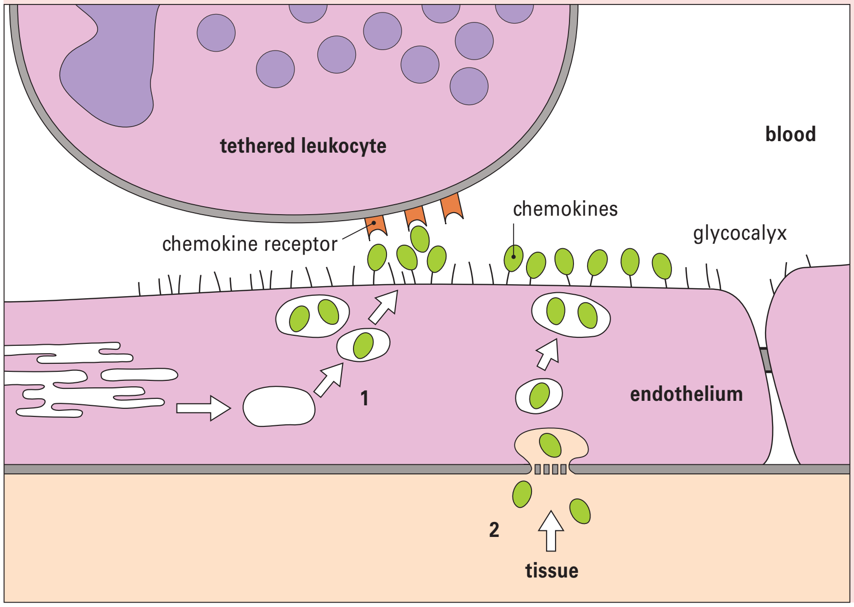

Fig. 6.10 Chemokines bind to glycosaminoglycans on endothelium via one binding site while a second site interacts with chemokine receptors expressed on the surface of the leukocyte. Chemokines may be synthesized by the endothelial cell and stored in vesicles (Weibel–Palade bodies) to be released to the luminal surface (blood side) following activation (**1**). Alternatively chemokines may be produced by cells in the tissues and transported across the endothelium (**2**). Cells in the tissue can therefore signal either directly or indirectly to circulating leukocytes.

Some chemokine receptors and their principal ligands

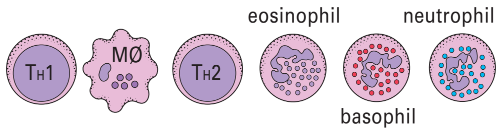

	TH1	MØ	TH2	eosinophil	basophil	neutrophil
	(CCR1) CCR2 CCR5 CXCR3	CCR1 CCR2 CCR5	CCR2 CCR3	(CCR1) CCR2 CCR3	CCR1 CCR2 CCR3	(CCR1) CXCR1 CXCR2
CCL3	+	+		+	+	+
CCL4	+	+				
CCL5	+	+	+	+	+	+
CCL2	+	+	+	+	+	
CCL7,8	+	+	+	+	+	+
CCL13	+	+	+	+	+	
CCL11			+	+	+	
CXCL8						+
CXCL1,2,3						+
CXCL9,10,11	+					

Fig. 6.11 Some of the chemokine receptors found on particular leukocytes and the chemokines they respond to are listed. The cells are grouped according the principal types of effector response. Note that TH1 cells and mononuclear phagocytes both express chemokine receptor CCR5, which allows them to respond to chemokine CCL3, while TH2 cells, eosinophils, and basophils express CCR3, which allows them to respond to CCL11. This allows selective recruitment of sets of leukocytes into areas with particular types of immune/inflammatory response. Both groups of cells express chemokine receptors CCR1 and CCR2, which allow responses to macrophage chemotactic proteins (CCL2, CCL7, CCL8, and CCL13). Neutrophils express chemokine receptors CXCR1 and CXCR2, which allow them to respond to CXCL8 (IL-8) and CXCL1 and CXCL2. Bracketed entries indicate that only a subset of cells express that receptor. Full details are given in Appendix 4.

activate that receptor. As a consequence, antigen-activated lymphocytes are more readily triggered to enter sites of inflammation where these chemokines are expressed.

Once a leukocyte has crossed the endothelium, it is capable of responding to a new set of chemokines, which direct its migration through the tissues.

Chemokines thus work in a hierarchical way, with a cell receiving successive signals for the next destination.

Other molecules are also chemotactic for neutrophils and macrophages

Several other molecules are chemotactic for neutrophils and macrophages, both of which have an f.Met-Leu-Phe (f.MLP) receptor. This receptor binds to peptides blocked at the N terminus by formylated methionine. Because prokaryotes (e.g. bacteria) initiate all protein translation with this amino acid, whereas eukaryotes do not, this provides a simple specific signal for the presence of bacteria, towards which phagocytes should move.

Neutrophils and macrophages also have receptors for:

- C5a, which is a fragment of a complement component generated at sites of inflammation following complement activation; and
- LTB4, which is generated at sites of inflammation following activation of a variety of cells, particularly macrophages and mast cells.

In addition, molecules generated by the blood clotting system, notably fibrin peptide B and thrombin, attract phagocytes, though many molecules such as these only act indirectly by inducing chemokines.

The first leukocytes to arrive at a site of inflammation, if activated, are able to release chemokines that attract others. For example:

- CXCL8 released by activated monocytes can induce neutrophil and basophil chemotaxis;
- similarly, macrophage activation leads to metabolism of arachidonic acid with release of LTB4.

Like chemokines, these other chemotactic molecules, including C5a and fMLP, act via 7tm receptors.

Q. The innate immune system has the capacity to recognize pathogen-associated molecular patterns (PAMPs). Which of the chemotactic molecules described above has a PAMP?
A. fMLP. It is the only molecule that comes from bacteria; all other chemotactic molecules are produced by cells of the body.

Integrins on the leukocytes bind to CAMs on the endothelium

Activation of leukocytes via their chemokine receptors initiates the next stage of migration.

Leukocytes and many other cells in the body interact with other cells and components of the extracellular matrix using a group of molecules called integrins.

In the third step of leukocyte adhesion (see Fig. 6.7), the leukocytes develop a firm adhesion to the endothelium using their surface integrins. Cell activation promotes this step in three ways:

- it can cause integrins to be released from intracellular stores;
- it can cause clustering of integrins on the cell surface into high-avidity patches;
- most importantly, the cell activation induced by the chemokines causes the integrins to become associated with the cytoskeleton and can switch them into a high-affinity form (Fig. 6.12) – this type of signaling is called inside-out signaling because activation inside the cell causes a change in the affinity of the extracellular portion of the integrin.

Normally the binding affinity of integrins for the CAMs on the endothelium is relatively weak, but when sufficient interactions take place, the cells adhere firmly.

Many of the CAMs on the endothelium have immunoglobulin-like domains (i.e. they belong to the immunoglobulin superfamily).

The affinity of integrins is controlled by inside-out signaling

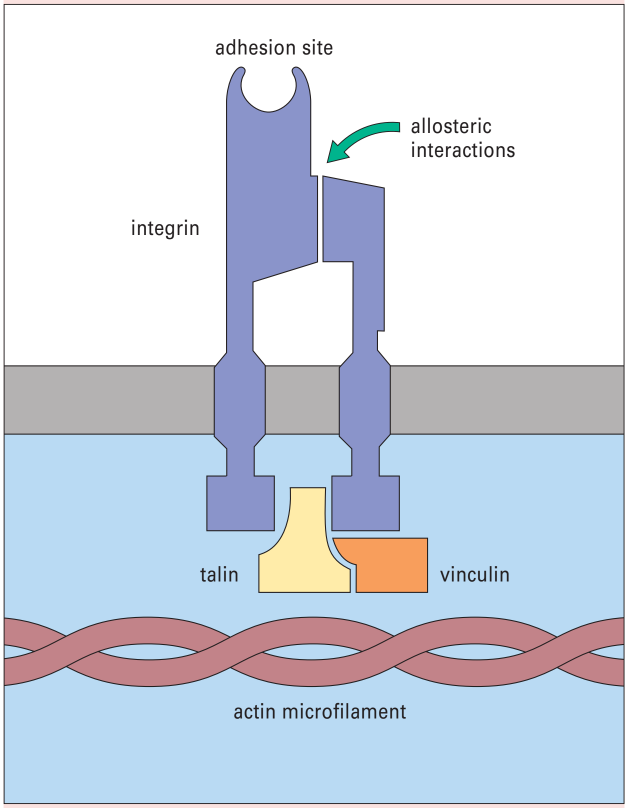

Fig. 6.12 Activation of the cell causes a change in the position of the two chains of the integrin, which become linked to the cytoskeleton via the adapter molecules vinculin and talin. The association produces an allosteric change in the extracellular portion of the molecule, causing the binding site to open and allowing the integrin to attach to its ligand.

Some of the CAMs are constitutively expressed, while others are specifically induced on particular endothelia or at sites of inflammation.

Integrins and CAMs are very important groups of molecules

INTEGRINS COMPRISE A MAJOR GROUP OF ADHESION MOLECULES – Integrins are present on many cells, including leukocytes. Each member of this large family of molecules consists of two non-covalently bound polypeptides (α and β), both of which traverse the membrane.

Integrins fall into three major families depending on which β chain they have, because any β chain can associate with several α chains. Broadly speaking:

- the β_1-integrins are involved in binding of cells to extracellular matrix;
- the β_2-integrins are involved in leukocyte adhesion to endothelium or to other immune cells; and
- the β_3-integrins (cytoadhesins) are involved in the interactions of platelets and neutrophils at inflammatory sites or sites of vascular damage.

However, there are several exceptions to this rule, and some α chains can associate with more than one β chain (Fig. 6.13).

The ability of integrins to bind to their ligands depends on divalent cations. For example, LFA-1 ($\alpha_L\beta_2$-integrin) has a Mg^{2+} ion coordinated at the center of a binding site, which accommodates an aspartate residue from the ligand. Activation of the integrin causes the binding site to open, allowing access for the aspartate side chain.

Many of the integrins can bind to more than one ligand. For example:

- LFA-1 present on most lymphocytes binds to both intercellular CAM-1 (ICAM-1) and ICAM-2, which are expressed on endothelium;
- VLA-4 ($\alpha_4\beta_1$-integrin) binds to vascular CAM-1 (VCAM-1) expressed on endothelium, or to fibronectin (an extracellular matrix component).

ENDOTHELIAL CAMS INCLUDE ICAM-1, ICAM-2, VCAM-1, AND MAdCAM-1 – The CAMs on the endothelium that interact with integrins are all members of the immunoglobulin supergene family. They include ICAM-1 and ICAM-2, VCAM-1, and mucosal addressin CAM (MAdCAM-1) (Fig. 6.14). All members of this family are expressed or inducible on vascular endothelium:

- ICAM-1 and VCAM-1 are both induced by inflammatory cytokines;
- ICAM-2 is constitutively expressed at low levels on some endothelia and is downregulated by inflammatory cytokines.

Integrins

integrin		ligands	expression
VLA-1	$\alpha_1\beta_1$	?	T cells, fibroblasts
VLA-2	$\alpha_2\beta_1$	collagen	activated T cells, platelets
VLA-3	$\alpha_3\beta_1$	laminin, collagen, fibronectin	kidney, thyroid
VLA-4	$\alpha_4\beta_1$	VCAM-1, fibronectin	lymphocytes, some phagocytes
VLA-5	$\alpha_5\beta_1$	fibronectin	some leukocytes, platelets
VLA-6	$\alpha_6\beta_1$	laminin	widely distributed
LPAM-1	$\alpha_4\beta_7$	MAdCAM-1 (VCAM-1)	some T cells
	$\alpha_E\beta_7$	E-cadherin	intraepithelial T cells
LFA-1	$\alpha_L\beta_2$	ICAM-1, ICAM-2 (ICAM-3)	most leukocytes
CR3	$\alpha_M\beta_2$	C3b, C4b, ICAM-1	mononuclear phagocytes, neutrophils
CR4	$\alpha_X\beta_2$	C3b, C4b, ICAM-1 ?	macrophages
leukointegrin	$\alpha_D\beta_2$	ICAM-3, VCAM-1	macrophages

Fig. 6.13 The table gives the properties of some of the major integrins involved in leukocyte binding to endothelium or extracellular matrix.

Endothelial cell adhesion molecules

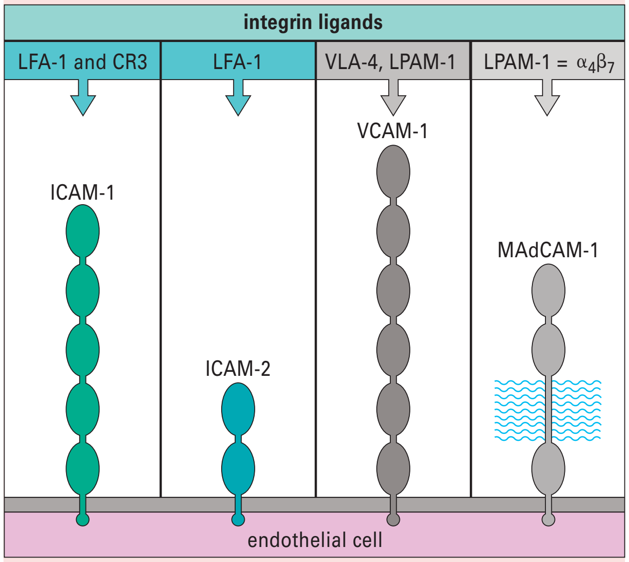

Fig. 6.14 The molecules ICAM-1, ICAM-2, VCAM-1, and MAdCAM-1 are illustrated diagrammatically with their immunoglobulin-like domains. Their integrin ligands (see Fig. 6.13) are listed above. MAdCAM-1 also has a heavily glycosylated segment, which binds L-selectin.

Mucosal addressin cell adhesion molecule of endothelium

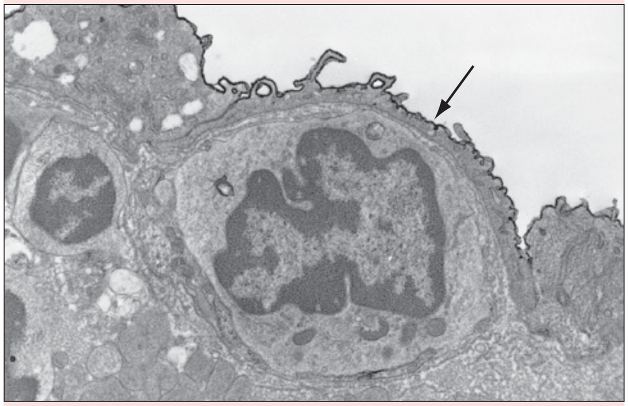

Fig. 6.15 The immunoelectron micrograph has been stained to show MAdCAM-1 as a dark border (arrow) on the luminal surface of endothelium. In this instance the molecule is expressed on brain endothelium in chronic relapsing experimental allergic encephalomyelitis induced by immunization of Biozzi AB/H mice with myelin basic protein. (Reproduced from Immunology, O'Neill JK, Butter C 1991;72:520–525. Copyright 1991 with permission from Blackwell Publishing)

The two N terminal domains of ICAM-1 are homologous to those of ICAM-2 and both molecules interact with LFA-1. (There are additional ligands for LFA-1, ICAM-3, and ICAM-4, which are expressed on lymphocytes.)

MAdCAM-1 is a composite molecule that includes:

- two immunoglobulin-like domains that interact with an integrin; and
- a glycosylated segment that can interact with selectins.

MAdCAM-1 was first identified on mucosal lymph node endothelium, but can also be induced at sites of chronic inflammation (Fig. 6.15)

Leukocyte migration in inflammation is controlled by both leukocytes and endothelium

Although the mechanisms described above are applicable to all leukocyte migration, the following discussion concerns just inflammation. Several factors account for the pattern of leukocyte accumulation, including:

- the state of activation of the lymphocytes or phagocytes – the expression of adhesion molecules and their functional affinity vary depending on the type of cell and whether it has been activated by antigen, cytokines, or cellular interactions;
- the types of adhesion molecule expressed by the vascular endothelium, which is related to its anatomical site and whether it has been activated by cytokines;
- the particular chemotactic molecules and cytokines present – receptors vary between leukocyte populations so that particular chemotactic agents act selectively on some cells only.

Chemokines and adhesion molecules determine which leukocytes migrate where

Different sets of adhesion molecules and chemotactic agents are used by each type of cell.

E-selectin and P-selectin slow circulating leukocytes

When tissue is damaged, cells, including mononuclear phagocytes, release inflammatory cytokines such as TNFα or IL-1, which induce synthesis and expression of E-selectin.

In vitro, cells transfected with the gene for E-selectin bind neutrophils strongly, and this suggests that the slowing of neutrophils by E-selectin is a critical first step in neutrophil migration.

P-selectin acts similarly to E-selectin, but is held ready-made in the Weibel–Palade bodies of endothelium and released to the cell surface if the endothelium becomes activated or damaged. Both E-selectin and P-selectin can slow circulating platelets or leukocytes.

Different chemokines cause different types of leukocyte to accumulate

In the second step of migration, neutrophils are triggered by chemokines such as CXCL8 synthesized by cells in the tissue, or by the endothelium itself. CXCL8 acts on two different chemokine receptors – CXCR1 and CXCR2 (see Fig. 6.11) – to initiate neutrophil migration.

In some tissues different sets of chemokines cause the local accumulation of other groups of leukocytes. For example:

- in the bronchi of individuals with asthma, CCL11 (eotaxin) is released, which causes the accumulation of eosinophils – CCL11 acts on CCR3, which is also present on TH2 cells and basophils, so by releasing one chemokine, the tissue can signal to three different kinds of cell to migrate into the tissue and this particular set of cells is characteristic of the cellular infiltrates in asthma;

- in sites of chronic inflammation, the chemokines CXCL10 and CCL2 are released by endothelium in response to interferon-γ (IFNγ) and TNFα – CXCL10 acts on activated TH1 cells (via CXCR3), while CCL2 acts on macrophages (via CCR5); consequently macrophages and TH1 cells tend to accumulate at sites of chronic inflammation.

Q. In what type of inflammatory reaction would you expect to see the production of CXCL10 and the accumulation of activated macrophages?

A. In chronic inflammatory reactions.

Some endothelial adhesion molecules are induced by cytokines

Some endothelial adhesion molecules are induced at sites of inflammation by cytokines, particularly ICAM-1 and VCAM-1 (Fig. 6.16). Their level of expression is also tissue-specific. For example:

- ICAM-1 is expressed at higher levels on brain endothelium than VCAM-1;
- ICAM-I and VCAM-1 are equally expressed in skin endothelium.

By contrast ICAM-2 is not induced by inflammatory cytokines, and it has been suggested that ICAM-2 determines the basal level of binding of leukocytes to different types of endothelium, while ICAM-1 and VCAM-1 mediate migration into sites of inflammation.

ICAM-1 and ICAM-2 both interact with LFA-1 ($\alpha_L\beta_2$-integrin, see Fig. 6.13), which is present on most leukocytes, whereas VCAM-1 interacts with VLA-4 ($\alpha_4\beta_1$-integrin). Both of these integrins are upregulated on activated lymphocytes, so the process of cell activation induces both:

- the chemokine receptors that allow activated lymphocytes to respond to inflammatory cytokines; and
- the integrins that allow activated lymphocytes to migrate.

Expression and induction of endothelial adhesion molecules

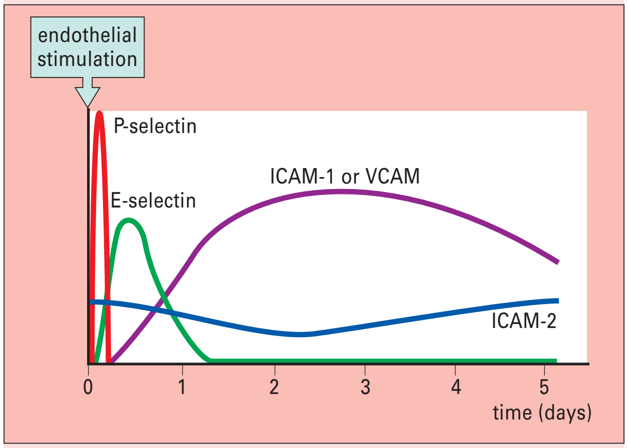

Fig. 6.16 The graph shows the time course of induction of different endothelial molecules on human umbilical vein endothelium in vitro following stimulation by TNFα.

CR3 is an important integrin expressed on monocytes

Another important integrin is CR3 ($\alpha_M\beta_2$-integrin, see Fig. 6.13), which is expressed on monocytes. Its function in phagocyte accumulation has been pinpointed by studies in vivo using antibodies to CR3, which inhibit phagocyte migration.

It is notable that a group of patients who have leukocyte adhesion deficiency (LAD) syndrome and suffer from severe infections due to poor phagocyte accumulation, are deficient in all β_2-integrins (LFA-1, CR3, etc., see Fig. 6.13).

CR3 recognizes a site on ICAM-1 separate from that recognized by LFA-1.

Lymphocyte binding to endothelium can be modulated using antibodies to adhesion molecules or inactive analogs of chemokines, and this holds out prospects for therapy of diseases where immunopathological inflammation occurs.

VLA molecules allow interaction with extracellular matrix and belong to the β_1-integrin group

Once leukocytes have crossed the endothelium and entered the tissues, they interact with the proteins of the extracellular matrix (collagen, laminin, fibronectin, etc.), as well as the tissue cells.

As lymphocytes leave the blood vessel they lose some of their surface molecules (e.g. L-selectin), which are no longer required. The functional phenotype changes from that of a circulating cell to one adapted to move through tissues.

Many of the molecules that allow interaction with extracellular matrix belong to the β_1-integrin group, and are known as very late antigens (VLAs), so-called because they were first identified on the T cell surface at a late stage after T cell activation. The whole group of β_1-integrins are now referred to as VLA molecules, though most of them are not just expressed on lymphocytes. This group includes:

- receptors for collagen (VLA-2 and VLA-3);
- receptors for laminin (VLA-3 and VLA-6); and
- receptors for fibronectin (VLA-3, VLA-4, and VLA-5).

The fact that some of these molecules appear late after lymphocyte activation suggests that cells go through a program of differentiation, and that the ability to interact with extracellular matrix is one of the last functions to develop.

LEUKOCYTE MIGRATION TO LYMPHOID TISSUES IS ALSO CONTROLLED BY CHEMOKINES

Migration of leukocytes into lymphoid tissues is also controlled by chemokines and adhesion molecules on the endothelium.

Q. What is distinctive about the endothelium in secondary lymphoid tissues?

A. These tissues have high endothelial venules (HEVs) with columnar endothelial cells (see Figs 2.54 and 2.55) that express high levels of sulfated glycoproteins and distinctive sets of adhesion molecules.

Up to 25% of lymphocytes that enter a lymph node via the blood may be diverted across the HEV. In contrast, only a tiny proportion of those circulating through other tissues will cross the regular venular endothelium at each transit. HEVs are therefore particularly important in controlling lymphocyte recirculation. Normally they are only present in the secondary lymphoid tissues, but they may be induced at sites of chronic inflammation.

In addition to their peculiar shape, HEV cells express distinct sets of heavily glycosylated sulfated adhesion molecules, which bind to circulating T cells and direct them to the lymphoid tissue.

The HEVs in different lymphoid tissues have different sets of adhesion molecules. In particular, there are separate molecules controlling migration to:

- Peyer's patches;
- mucosal lymph nodes; and
- other lymph nodes.

These molecules were previously called vascular addressins, and their expression on different HEVs accounts for the way in which lymphocytes relocalize to their own lymphoid tissue.

Naive lymphocytes express L-selectin, which contributes to their attachment to carbohydrate ligands on HEVs in mucosal and peripheral lymph nodes. Once they have stopped on the HEV, migrating lymphocytes may use the integrin $\alpha_4\beta_7$ (LPAM-1, see Fig. 6.13) to bind to MAdCAM on the HEV of mucosal lymph nodes or Peyer's patches.

Because the expression of $\alpha_4\beta_7$ allows migration to mucosal lymphoid tissue, whereas $\alpha_4\beta_1$ allows attachment to VCAM-1 on activated endothelium or fibronectin in tissues, expression of one or other of these molecules can alternately be used by naive lymphocytes migrating to lymphoid tissue or activated T cells to inflammatory sites.

Chemokines are important in controlling cell traffic to lymphoid tissues

Chemokines are also important in controlling cell traffic to lymphoid tissues. Naive T cells:

- express chemokine receptors CCR7 and CXCR4, which allow them to respond to chemokines expressed in lymphoid tissues;
- also respond to chemokines CCL18 and CCL19 produced by dendritic cells, which is thought to direct them to the appropriate T cell areas of the lymph node where dendritic cells can present antigen to them.

Once T cells have been activated they lose CXCR4 and CCR7, but gain new chemokine receptors (see Fig. 6.11), which allow them to respond to different chemokines (e.g. those produced at sites of inflammation).

Naive B cells also express:

- CCR7 and CXCR4, involved in migration to secondary lymphoid tissues;
- CXCR5, a receptor for CXCL13, which appears to be required for localization to lymphoid follicles within the lymph nodes.

Cells moving into lymphoid tissue therefore respond sequentially to signals on the endothelium and signals from the different areas within the tissue (Fig. 6.17).

PLASMA ENZYME SYSTEMS MODULATE INFLAMMATION AND TISSUE REMODELING

The change in vascular permeability is another important component of inflammation. However, whereas cell migration occurs across venules, serum exudation occurs primarily across capillaries where blood pressure is higher and the vessel wall is thinnest. This event is controlled in two ways:

- blood supply to the area increases;
- there is an increase in capillary permeability caused by retraction of the endothelial cells and possibly also by increased vesicular transport across the endothelium – this permits larger molecules to traverse the endothelium than would ordinarily be capable of doing so, and so allows antibody and molecules of the plasma enzyme systems to reach the inflammatory site.

Chemokines and cell migration into lymphoid tissue

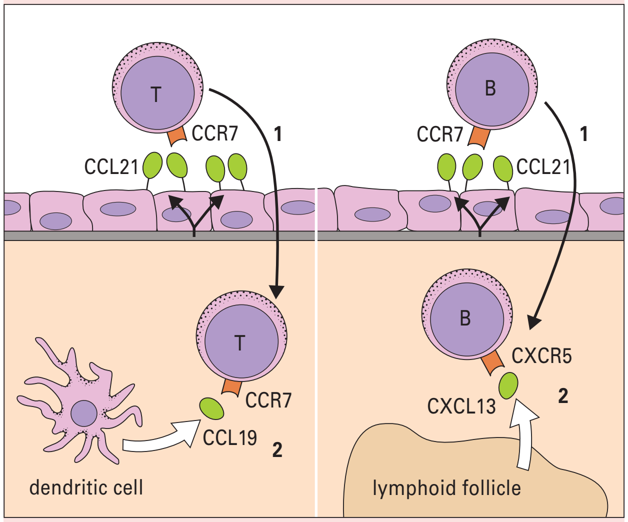

Fig. 6.17 Cell migration occurs in stages. Naive T cells express CCR7, which allows them to respond to CCL21 expressed by secondary lymphoid tissues (**1**). Once the cells have migrated across the endothelium, the same receptor can respond to signals from CCL19 produced by dendritic cells (**2**), which promotes interactions with the T cells in the paracortex (T cell area) of the lymph node. B cells also express CCR7 and use similar mechanisms to migrate into the lymphoid tissues (**1**). However, they also express CXCR5, which allows them to respond to CXCL13, a chemokine produced in lymphoid follicles (**2**) – B cells are therefore directed to the B cell areas of the node. Mice lacking CXCR5 do not develop normal lymphoid follicles.

The four major plasma enzyme systems that have an important role in hemostasis and control of inflammation are the:

- clotting system;
- fibrinolytic (plasmin) system;
- kinin system; and
- complement system (see Chapter 4).

The kinin system generates powerful vasoactive mediators

The kinin system generates the mediators bradykinin and lysyl-bradykinin (kallidin).

Bradykinin is a very powerful vasoactive nonapeptide that causes:

- venular dilation;
- increased vascular permeability; and
- smooth muscle contraction.

Bradykinin is generated following the activation of Hageman factor (XII) of the blood clotting system, whereas tissue kallikrein is generated following activation of the plasmin system or by enzymes released from damaged tissues (Fig. 6.18).

Activation of the kinin system

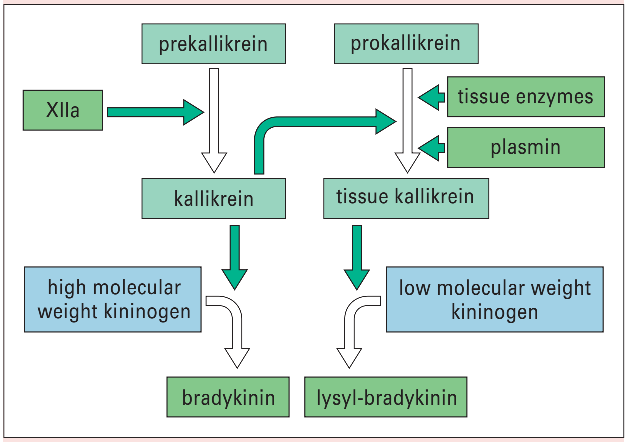

Fig. 6.18 Activated Hageman factor (XIIa) acts on prekallikrein to generate kallikrein, which in turn releases bradykinin from high molecular weight kininogen (HMWK). Prekallikrein and HMWK circulate together in a complex. Various enzymes activate prokallikrein to tissue kallikrein, which releases lysyl-bradykinin from low molecular weight kininogen. Bradykinin and lysyl-bradykinin are both extremely powerful vasodilators.

The plasmin system has an important role in tissue remodeling and regeneration

The plasmin system can be activated by a soluble or a tissue-derived plasminogen activator, which leads to the enzymatic conversion of plasminogen into plasmin. Plasmin itself:

- was originally identified by its ability to dissolve fibrin;
- has several other activities, in particular it activates some metalloproteases – enzymes that are required for the breakdown and remodeling of collagen (Fig. 6.19);
- additionally can promote angiogenesis (the formation of new blood vessels) by causing the release of cytokines that induce proliferation and migration of endothelial cells.

The plasmin system

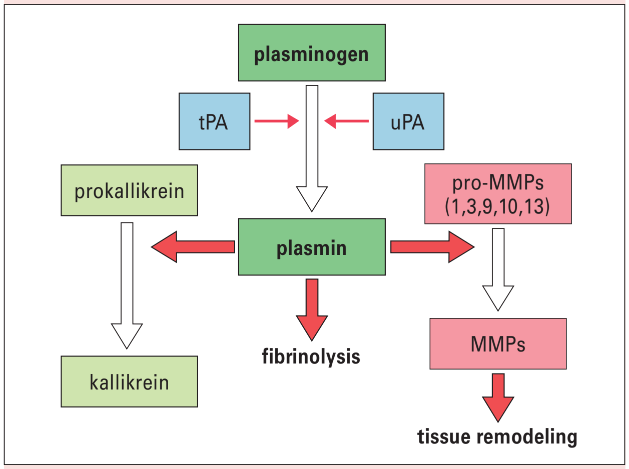

Fig. 6.19 Plasmin is generated by the enzymatic activity of plasminogen activators. (MMP, matrix metalloproteinase; tPA, tissue plasminogen activator; uPA, urokinase plasminogen activator)

Auxiliary cells are also sources of inflammatory mediators

Auxiliary cells, including mast cells, basophils, and platelets, are also very important in the initiation and development of acute inflammation. They act as sources of the vasoactive mediators histamine and 5-hydroxytryptamine (serotonin), which produce vasodilation and increased vascular permeability.

Mast cells, basophils, and platelets release a variety of mediators

Many of the proinflammatory effects of C3a and C5a result from their ability to trigger mast cell granule release, because they can be blocked by antihistamines.

Mast cells and basophils are also:

- a route by which the adaptive immune system can trigger inflammation – IgE sensitizes these cells by binding to their IgE receptors and the cells can then be activated by antigen;
- an important source of slow-reacting inflammatory mediators, including the leukotrienes and prostaglandins, which contribute to a delayed component of acute inflammation and are synthesized and act some hours after mediators like histamine, which are pre-formed and released immediately following mast cell activation (see Chapter 23).

Fig. 6.20 lists the principal mediators of acute inflammation, and the interaction of the immune system with complement and other inflammatory systems is shown in Fig. 6.21.

Inflammatory mediators

mediator	main source	actions
histamine	mast cells, basophils	increased vascular permeability, smooth muscle contraction, chemokinesis
5-hydroxytryptamine (5HT – serotonin)	platelets, mast cells (rodent)	increased vascular permeability, smooth muscle contraction
platelet activating factor (PAF)	basophils, neutrophils, macrophages	mediator release from platelets, increased vascular permeability, smooth muscle contraction, neutrophil activation
IL-8 (CXCL8)	mast cells, endothelium, monocytes and lymphocytes	polymorph and monocyte localization
C3a	complement C3	mast cell degranulation, smooth muscle contraction
C5a	complement C5	mast cell degranulation, neutrophil and macrophage chemotaxis, neutrophil activation, smooth muscle contraction, increased capillary permeability
bradykinin	kinin system (kininogen)	vasodilation, smooth muscle contraction, increased capillary permeability, pain
fibrinopeptides and fibrin breakdown products	clotting system	increased vascular permeability, neutrophil and macrophage chemotaxis
prostaglandin E_2 (PGE_2)	cyclo-oxygenase pathway, mast cells	vasodilation, potentiates increased vascular permeability produced by histamine and bradykinin
leukotriene B_4 (LTB_4)	lipoxygenase pathway, mast cells	neutrophil chemotaxis, synergizes with PGE_2 in increasing vascular permeability
leukotriene D_4 (LTD_4)	lipoxygenase pathway	smooth muscle contraction, increasing vascular permeability

Fig. 6.20 The major inflammatory mediators that control blood supply and vascular permeability or modulate cell movement.

Platelets may be activated by:

- immune complexes; or
- platelet activating factor (PAF) from neutrophils, basophils, and macrophages.

Activated platelets release mediators. This is thought to be important in type II and type III hypersensitivity reactions (see Chapter 24).

Lymphocytes and monocytes release mediators that control the accumulation and activation of other cells

Once lymphocytes and monocytes have arrived at a site of infection or inflammation, they can also release mediators, which control the later accumulation and activation of other cells. For example:

- activated macrophages release the chemokine CCL3 (MIP-1α) and the leukotriene LTB_4, both of which are chemotactic and encourage further monocyte migration;
- lymphocytes can modulate later lymphocyte traffic by the release of chemokines and inflammatory cytokines, particularly IFNγ.

Q. What determines whether an immune response is acute or chronic?

A. Ultimately the outcome of an acute inflammatory response is related to the fate of the antigen. If the initiating antigen or pathogen persists, then leukocyte accumulation continues and a chronic inflammatory reaction develops. If the antigen is cleared then no further leukocyte activation occurs and the inflammation resolves.

In recurrent inflammatory reactions and in chronic inflammation the patterns of cell migration are different from those seen in an acute response.

We now know that the patterns of inflammatory cytokines and chemokines vary over the time course of an inflammatory reaction and this can be related to the successive waves of migration of different types of leukocyte into the inflamed tissue.

Chronic inflammation is characteristic of sites of persistent infection (see Chapter 9) or in autoimmune reactions where the antigen cannot ultimately be eradicated (see Chapter 20).

PAMPs ARE DISTINCTIVE BIOLOGICAL MACROMOLECULES THAT CAN BE RECOGNIZED BY THE INNATE IMMUNE SYSTEM

The molecules in this section are divided into families according to structure. Some of each family are also **acute phase proteins** (i.e. they are present in the blood and their levels increase during infection).

The immune system in acute inflammation

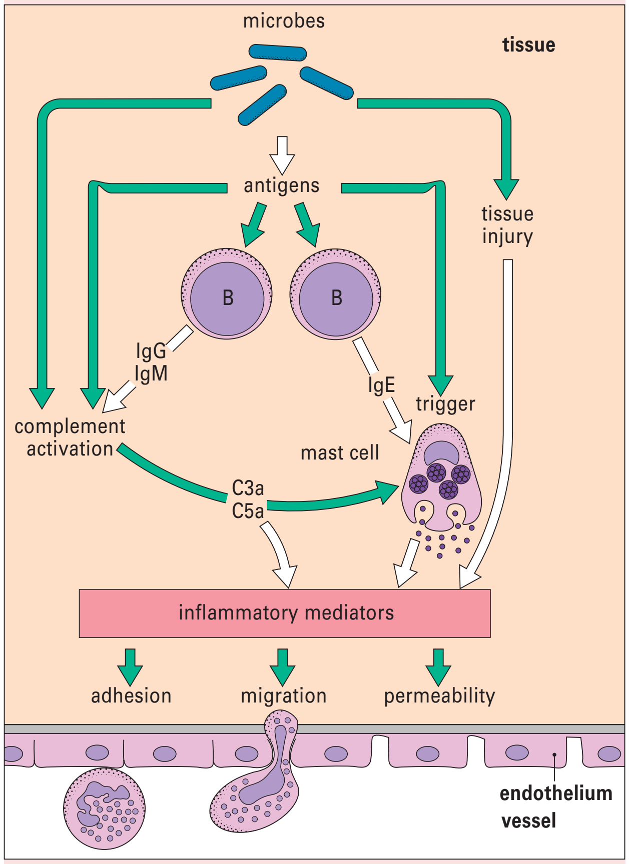

Fig. 6.21 The adaptive immune system modulates inflammatory processes via the complement system. Antigens (e.g. from microorganisms) stimulate B cells to produce antibodies including IgE, which binds to mast cells, while IgG and IgM activate complement. Complement can also be activated directly via the alternative pathway (see Fig. 4.7). When triggered by antigen, sensitized mast cells release their granule-associated mediators and eicosanoids (products of arachidonic acid metabolism, including prostaglandins and leukotrienes). In association with complement (which can also trigger mast cells via C3a and C5a), the mediators induce local inflammation, facilitating the arrival of leukocytes and more plasma enzyme system molecules.

Once phagocytes have entered the tissues, they bind, internalize, and destroy pathogens. Many pathogens have distinctive biological macromolecules on their surface, which can be recognized by elements of the innate immune system. These distinctive molecular structures are called **pathogen-associated molecular patterns (PAMPs)**.

In many cases, the innate recognition mechanisms allow phagocytes to bind and internalize the pathogens and this is often associated with activation of the phagocytes, which enhances their microbicidal activity.

The binding of the pathogen to the phagocyte can be direct or indirect:

- direct recognition involves the surface receptors on the phagocyte directly recognizing surface molecules on the pathogen;
- indirect recognition involves the deposition of serum-derived molecules onto the pathogen surface and their subsequent binding to receptors on the phagocyte (i.e. the process of opsonization) (Fig. 6.22)

Q. Give an example of an innate immune recognition system that allows macrophages to recognize and phagocytose bacteria.

A. Complement C3b can become deposited on bacteria following activation of either the lectin pathway or the alternative pathway (see Fig. 4.4). The deposited C3b acts as an opsonin and is recognized by the receptors CR1, CR3, and CR4 on macrophages (see Fig. 4.12).

Collectins and ficolins opsonize pathogens and inhibit invasiveness

The **mannan-binding lectin (MBL)** activates the lectin pathway and is one member of a group of serum proteins that recognize carbohydrates and act as opsonins (see Fig. 4.4). Each member of the group has subunits formed of a triple-helical collagenous tail and a lectin head (Fig. 6.23). A number of subunits may be linked together in the complete molecule and the overall structure of these molecules is similar to that of C1q (see Fig. 4.5), though C1q is not itself a lectin. Most of these molecules are produced by the liver and circulate in the blood, but the **surfactant proteins** are produced by lung epithelium.

MBL and the surfactant proteins are acute phase proteins

The synthesis of acute phase proteins is greatly increased in response to infection. Some acute phase proteins such as MBL increase 2–3-fold following infection, but others, for example C-reactive protein (CRP, see below), may increase several 100-fold.

By binding to the surface of bacteria or viruses the acute phase proteins can inhibit their ability to invade the tissues. For example surfactant protein-A (SP-A) binds to the glycosylated hemagglutin molecule on the surface of influenza virus and reduces the ability of the virus to infect cells.

In addition binding of collectins and ficolins can directly opsonize pathogens for uptake by phagocytes, though the receptors on the cells that bind to these molecules are not well defined.

Because some collectins and ficolins promote activation of the complement alternative pathway, their opsonizing activity may ultimately be mediated by the interaction of C3b with complement receptors.

The pentraxins CRP and SAP also opsonize bacteria

Two further acute phase proteins, **CRP** and **serum amyloid-P (SAP)**, both produced by the liver, are important opsonins. These molecules belong to a family called **pentraxins**, which have a characteristic protein fold.

CRP is a pentameric protein that binds to phosphorylcholine present on, for example, pneumococci and promotes their phagocytosis by binding to C1q

The binding of pathogen to the macrophage can be direct or indirect

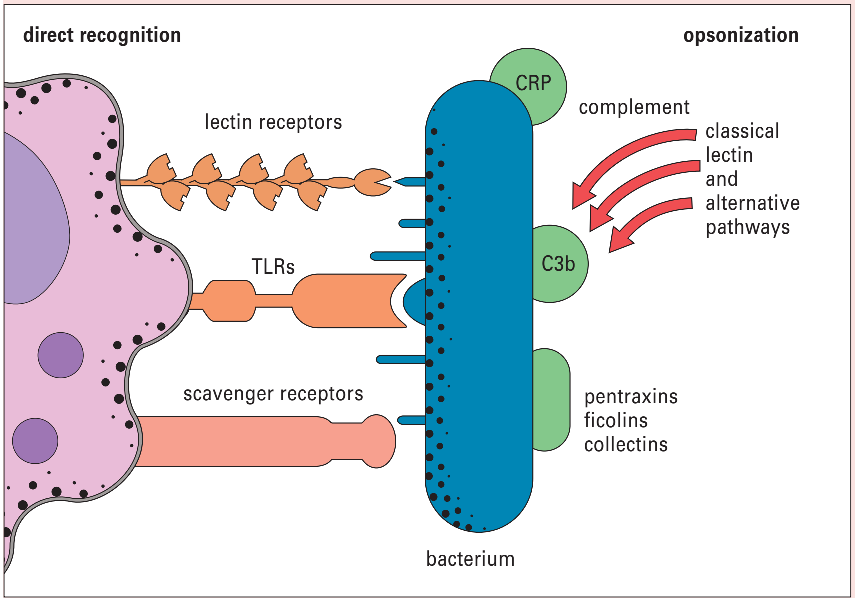
direct recognition
opsonization
CRP
complement
lectin receptors
classical
lectin
and
alternative
pathways
TLRs
C3b
scavenger receptors
pentraxins
ficolins
collectins
bacterium

Fig. 6.22 Macrophages can recognize PAMPs either directly or following opsonization with serum molecules.

Structure of collectins and ficolins

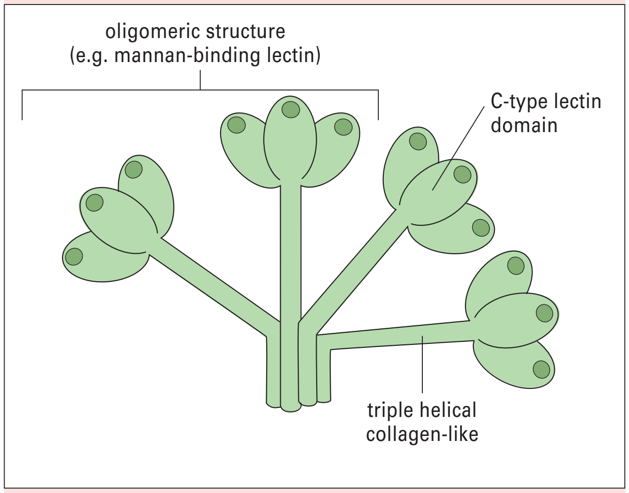
oligomeric structure
(e.g. mannan-binding lectin)
C-type lectin
domain
triple helical
collagen-like

Fig. 6.23 Collectins and ficolins are oligomers of triple-helical molecules with C-type lectin domains. They recognize a variety of microbial PAMPs.

and activating the complement classical pathway (see Fig. 4.4).

Because of its rapid synthesis in response to IL-6, CRP levels are often used clinically as a marker of infection or inflammation.

SAP also binds to bacteria and has a long evolutionary history (a homolog is present in horseshoe crabs), though its importance in host protection is still debated.

Other members of the pentraxin family are produced in different tissues and appear to have important roles in recognizing and clearing apoptotic cells.

Phagocytes have receptors that recognize pathogens directly

Even in the absence of opsonins, phagocytes have a number of receptors that allow them to recognize PAMPs. These include:

- scavenger receptors;
- carbohydrate receptors; and
- Toll-like receptors (TLRs).

Scavenger receptors and carbohydrate receptors (lectins) allow phagocytes to bind directly to microbes. Important examples include:

- the mannose receptor (CD206) on mononuclear phagocytes, which binds to mannan present in bacterial cell walls; and
- the dendritic cell receptors dectin-1 (which binds β1,3-glycan found in fungal cell walls) and dectin-2.

Each of these receptors has a C-type lectin domain, related to that found in the collectins and ficolins (see Fig. 6.23).

The scavenger receptors fall into two classes, members of which bind to lipopolysaccharide and lipoteichoic acid characteristic of Gram-negative bacteria. They are important, not just for recognition of pathogens, but also for clearing cell debris and apoptotic cells. The functions of these receptors in phagocytosis are described more fully in Chapter 9.

Q. Why do bacteria not just mutate their PAMPs so that they cannot be recognized by the innate immune system – after all, protein antigens of pathogens often mutate?

A. This is a difficult question because it requires deep insight into the evolutionary history of microbes, but many of the PAMPs are so fundamental to the structure of the bacterial or fungal cell walls that it is difficult to see how they could be altered without destroying the integrity of the microbe.

Toll-like receptors activate phagocytes and inflammatory reactions

The transmembrane protein Toll was first identified in the fruit fly *Drosophila* as a molecule required during embryogenesis. It was also noted that mutants lacking Toll were highly susceptible to infections with fungi and Gram-positive bacteria, suggesting that the molecule might be involved in immune defense. Subsequently a series of Toll-like receptors (TLRs) was identified in mammals that had very similar intracellular portions to the receptor in the flies.

The intracellular signaling pathways activated by the TLRs and the receptor for IL-1 are very similar and lead to activation of the transcription factor NK-κB (see Chapter 9).

The family of TLRs includes more than ten different receptors, many of which are capable of recognizing different microbial components (Fig. 6.24).

All of the TLRs are present on phagocytic cells, and some are also expressed on dendritic cells, mast cells, and B cells. Indeed, most tissues of the body express at least one TLR.

Q. TLR5 is present on the basal surface of epithelia in the gut, but not on the apical surface. What consequence might this have for immune reactions in the gut?

A. Bacteria within the gut would not stimulate the cell, whereas any microbes that have invaded the epithelial layer may do so. This may explain why non-pathogenic commensal bacteria in the gut do not elicit an inflammatory reaction, whereas pathogenic bacteria that have invaded the tissue do.

Expression of many of the TLRs is increased by inflammatory cytokines (e.g. TNFα, IFNγ) and elevated expression is seen in conditions such as inflammatory bowel disease.

The functional importance of the TLRs has been demonstrated in knockout mouse strains lacking individual receptors. Depending on the TLR involved, such animals fail to secrete inflammatory cytokines in response to pathogens and the microbicidal activity of phagocytes is not stimulated. These results show that the TLRs are primarily important in activating phagocytes in addition to any role they may have in endocytosis.

The TLRs make an important link between the innate and adaptive immune systems because their activation leads to the expression of co-stimulatory molecules on the phagocytes, which convert them into effective antigen-presenting cells.

The binding of microbial components to the TLRs effectively acts as a danger signal to increase the microbicidal activity of the phagocytes and allows them to activate T cells.

MICROBICIDAL PROTEINS ARE PART OF THE INNATE IMMUNE SYSTEM

Once a microbe has been endocytosed, it becomes subject to a number of microbicidal mechanisms.

In *Drosophila* fruit flies infection causes the release of various microbicidal proteins from the fat-body into the hemolymph.

In mammals, related proteins belonging to the **defensin** and **cathelicidin** families limit microbial spread by disrupting their outer membranes. For example the outer membrane of Gram-negative bacteria contains a higher level of negatively charged phospholipids than the mammalian plasma membrane. The positively charged defensins therefore preferentially insert into the bacterial membrane, particularly in conditions of low salt concentration.

Defensins constitute one of the major killing mechanisms of neutrophils and eosinophils. However, individual defensins are also produced by several cell types in tissues, particularly at epithelial surfaces in the lung, gut, and bladder. Some of these microbicidal peptides also cause mast cell degranulation with chemokine release and/or have direct chemotactic activity. They can therefore attract leukocytes to sites of infection.

The importance of these ancient antimicrobial defenses has been illustrated in gene-knockout animals, which often show increased susceptibility to particular types of

The Toll-like receptors

Toll-like receptor	ligand	pathogen
TLR1	lipopeptides	Gram-negative bacteria mycobacteria
TLR2	lipoteichoic acid lipoarabinomannan zymosan	Gram-positive bacteria mycobacteria fungi
TLR3	dsRNA	viruses
TLR4	lipopolysaccharide	Gram-negative bacteria
TLR5	flagellin	bacteria
TLR6	di-acyl lipopeptides	mycobacteria
TLR9	CpG-DNA	bacteria

Fig. 6.24 The Toll-like receptors recognize a variety of PAMPs. Gene duplication of the TLR precursor and divergence of function has led to a family of molecules capable of recognizing different types of pathogen.

infection. There is, however, little information on naturally occurring deficiencies of these molecules in humans. The functions of the microbicidal proteins are described in more detail in Chapter 9.

FURTHER READING

Akira S. Mammalian Toll-like receptors. Curr Opin Immunol 2003;15:5–11.

Binnerts ME, van Kooyk Y. How LFA-1 binds to different ligands. Immunol Today 1999;20:234–239.

Holmskov U, Thiel S, Jensenius JC. Collectins and ficolins: humoral lectins of innate immune defense. Annu Rev Immunol 2003;21;547–578.

Hynes RO. Integrins: bidirectional allosteric signalling machines. Cell 2002;110:673–678.

Lindhout E, Vissers JLM, Figdor CG, Adema GJ. Chemokines and lymphocyte migration. The Immunologist 1999;7:147–152.

Rossi D, Zlotnik A. The biology of chemokines and their receptors. Annu Rev Immunol 2000;18:217–242.

Springer TA. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. Cell 1994;76:301–314.

Underhill M, Ozinsky A. Phagocytosis of microbes: complexity in action. Annu Rev Immunol 2002;20:825–852.

Yang D, Biragyn A, Hoover DM, et al. Multiple roles of antimicrobial defensins. Cathelicidins and eosinophil-derived neurotoxin in host defense. Annu Rev Immunol 2004;22:181–216.

Critical thinking: The role of adhesion molecules in T cell migration (see p. 494 for explanations)

An experiment has been carried out to determine which CAMs mediate the migration of antigen-activated T cells across brain endothelium using a monolayer of endothelium overlaid with lymphocytes in vitro. The endothelium is either unstimulated or has been stimulated for 24 hours before the experiment with IL-1. In some cases the co-cultures were treated with blocking antibodies to different adhesion molecules. The results shown in the table below indicate the percentage of T cells that migrate across the endothelium in a 2-hour period.

blocking antibody	percentage of T cells migrating in 2 hours	
	unstimulated endothelium	IL-1-stimulated endothelium
none	18	48
anti-ICAM-1	3	16
anti-VCAM-1	19	28
anti-$\alpha_L\beta_2$-integrin (LFA-1)	2	14
anti-$\alpha_4\beta_1$-integrin (VLA-4)	17	32

1 Why does treatment of the endothelium with IL-1 cause an increase in the percentage of migrating cells in the absence of any blocking antibody?

2 Why does it require 24 hours of treatment with IL-1 to enhance the migration (1 hour of treatment does not produce this effect)?

3 Which adhesion molecules are important in mediating T cell migration across unstimulated endothelium?

4 Which adhesion molecules are important in mediating T cell migration across IL-1-activated endothelium?

CHAPTER

7

Antigen Presentation

SUMMARY

- **T cells recognize peptide fragments that have been processed and become bound to major histocompatibility complex (MHC) class I or II molecules.** These MHC–antigen complexes are presented at the cell surface.

- **MHC class I molecules associate with endogenously synthesized peptides, binding to peptides produced by degradation of the cells' internal molecules.** This type of antigen processing is carried out by proteasomes (which cleave the proteins) and transporters (which take the fragments to the endoplasmic reticulum [ER]).

- **MHC class II molecules bind to peptides produced following the breakdown of proteins that the cell has endocytosed.** The peptides produced by degradation of these external antigens are loaded onto MHC class II molecules in a specialized endosomal compartment called MIIC.

- **Cross-presentation allows APCs to acquire antigens from infected cells.** A specialized pathway allows the acquisition of antigens from infected cells by APCs. This pathway, called cross-presentation, allows the display of exogenous antigens by MHC class I molecules.

- **Co-stimulatory molecules are essential for T cell activation.** Molecules such as B7 (CD80/86) on the APC bind to CD28 on the T cell to cause activation. Antigens presented without co-stimulation usually induce T cell anergy. Intercellular adhesion molecules also contribute to the interaction between a T cell and an antigen-presenting cell (APC). Interactions between intercellular cell adhesion molecule-1 (ICAM-1) and leukocyte functional antigen-1 (LFA-1) and between CD2 and its ligands extend the interaction between T cells and APCs.

- **CD4 binds to MHC class II and CD8 to MHC class I molecules.** These interactions increase the affinity of T cell binding to the appropriate MHC–antigen complex and bring kinases to the TCR complex.

- **The highly ordered area of contact between the T cell and APC is an immunological synapse.**

- **T cell activation induces enzyme cascades, leading to the production of interleukin-2 (IL-2) and the high-affinity IL-2 receptor on the T cell.** IL-2 is required to drive T cell division.

- **Antigen presentation affects the subsequent course of an immune response.** The immune system responds to clues that an infection has taken place before responding strongly to antigens.

T CELLS RECOGNIZE PEPTIDE FRAGMENTS BOUND TO MHC MOLECULES

Antigens recognized by T cells are degraded or processed in some way so that the determinant recognized by the T cell receptor (TCR) is only a small fragment of the original antigen.

Antigen processing refers to the degradation of antigen into peptide fragments, which become bound to major histocompatibility complex (MHC) class I or class II molecules (see Chapter 5). These are the critical fragments involved in triggering T cells. TCRs are sensitive to the sequences of amino acids in the MHC molecule peptide-binding groove rather than the conformational determinants recognized by antibodies.

Antigen presentation plays a central role in initiating and maintaining an appropriate immune response to antigen. The process is tightly controlled at several levels as follows:

- different types of antigen-presenting cell (APC) are brought into play depending on the situation, dendritic cells (DCs) being crucial for initiating responses;
- a complex series of molecular interactions takes place to ensure that small fragments of antigens are recognized in a highly specific manner by T cells;
- another level of control is exerted by **co-stimulatory molecules** on APCs, resulting in T cell activation only when appropriate, such as in an infection;
- adhesion molecules on the interacting cells also contribute to the stable binding of the cells, which promotes effective antigen presentation;
- signals from the cell surface are then transmitted by a series of signal transduction pathways that regulate gene expression in the nucleus;

- in the final stages the actions of cytokines on the lymphocytes drive cell division.

The four stages of antigen presentation are outlined in Fig. 7.1.

Different types of APC are involved depending on the situation

Interaction with APCs is essential for T cell activation

The interaction between T cells and the heterogeneous group of cells collectively termed APCs is the most extensively studied example of cell interaction in the immune system. It is the first such interaction to occur after antigen challenge and its outcome largely dictates the subsequent course of events:

- if a sufficient number of $CD4^+$ helper T (TH) cells are triggered, then the activation of B cells or the development of cell-mediated immunity follows;
- if TH cells are not triggered, a form of immunological tolerance can develop, so that no other immunological events follow.

Several cell types can act as APCs

A wide spectrum of cells can present antigen, depending on how and where the antigen is first dealt with by cells of the immune system. In a lymphoid organ, the three main types of APC are:

- DCs;
- macrophages; and
- B cells (Fig. 7.2).

Activation of naive T cells on first encounter with antigen on the surface of an APC is called **priming**, to distinguish it from the responses of effector T cells to antigen on the surface of their target cells and the responses of primed memory T cells.

Summary of the key intercellular signals in T cell activation

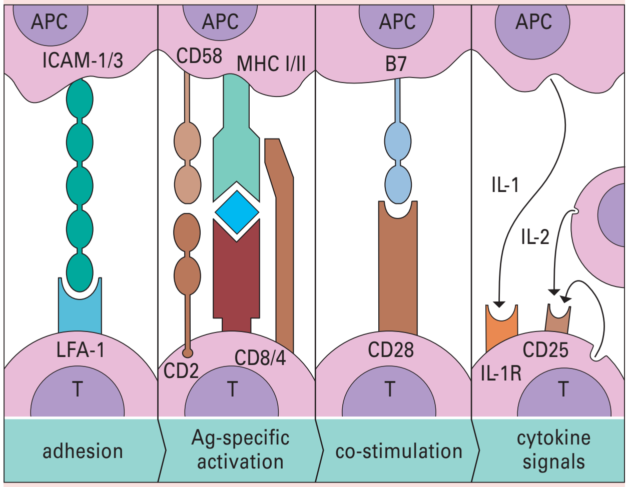

Fig. 7.1 Adhesion – association of APCs and T cells first involves non-specific, reversible binding through adhesion molecules, such as LFA-1 with ICAM-1 or ICAM-3. Antigen-specific recognition – recognition of the peptide antigen in the MHC molecule by the TCR, which provides the specificity of the interaction, results in prolonged cell–cell contact. Co-stimulation – a second signal (co-stimulation) is necessary for the T cell to respond efficiently, otherwise tolerance may result. Cytokine signals – activation results in upregulation of cytokines and their receptors, which boosts the activating signals and helps to decide the cell's fate.

Localization of antigen-presenting cells (APCs) in lymph nodes

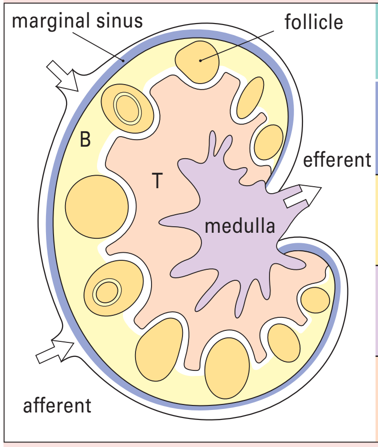

area	antigen-presenting cells	antigen	persistence of antigen
subcapsular (marginal) sinus	marginal zone macrophages	polysaccharides Ficoll (T_{ind})	+ + + +
follicles and B cell areas	follicular dendritic cells	immune complexes that fix complement	+ + +
medulla	classic macrophages	most antigens	+
T cell areas	interdigitating dendritic cells	most antigens	+ +

Fig. 7.2 A lymph node represented schematically showing afferent and efferent lymphatics, follicles, the outer cortical B cell area, and the paracortical T cell area. Different APCs predominate in these areas and selectively take up different types of antigen, which then persist on the surface of the cells for variable periods. Polysaccharides are preferentially taken up by marginal zone macrophages and may persist for months or years, whereas antigens on recirculating macrophages in the medulla may last for only a few days or weeks. Note that recirculating 'veiled' cells (Langerhans' cells), which originally come from the skin, change their morphology to become interdigitating dendritic cells within the lymph node. Both these cells and the follicular dendritic cells have long processes, which are in intimate contact with lymphocytes.

Rapid Reference Box 7

CD2 – Receptor for CD58 and CD48, expressed by T cells.
CD11a – LFA-1 expressed by T, B, and NK cells, monocyte/macrophages, and granulocytes.
CD28 – binds CD80 and CD86, expressed by subpopulations of T cells and activated B cells.
CD40 – binds CD154, co-stimulatory, expressed by B cells, monocyte/macrophages, and FDCs.
CD45 – leukocyte common antigen expressed by populations of T, B, and NK cells, and monocyte/macrophages.
CD54 – ICAM-1 expressed by T, B, and NK cells, monocyte/macrophages, granulocytes, platelets, DCs, and other cells.
CD58 – a ligand for CD2.
CD80 – B7-1, expressed by subpopulations of activated B and T cells and monocyte/macrophages, and binds CD28 and CD152.
CD86 – B7-2, expressed by activated B cells and monocyte/macrophages, and binds CD28 and CD152.
CD152 – CTLA-4, binds CD80 and CD86, a downregulatory signaling molecule that competes with CD28 for ligation of B7 on APCs, expressed by a subpopulation of activated T cells.
CD154 – CD40 ligand (CD40L).

(APC, antigen-presenting cell; CR, complement receptor; DC, dendritic cell; FDC, follicular dendritic cell; ICAM-1, intercellular adhesion molecule-1; LCA, leukocyte common antigen; LFA, lymphocyte functional antigen)

DCs ARE CRUCIAL FOR INITIATING RESPONSES – DCs, which are found in abundance in the T cell areas of lymph nodes and spleen, are the most effective cells for the initial activation of naive T cells. They pick up antigens in peripheral tissues, then migrate to lymph nodes, where they express high levels of adhesion and co-stimulatory molecules, as well as MHC class II molecules, which interact with the TCR and CD4 on TH cells.

Once they have migrated, DCs stop synthesizing MHC class II molecules, but maintain high levels of stable expression of MHC class II molecules containing peptides from antigens derived from the tissue where the DCs originated.

Interdigitating DCs are believed to be the major APCs involved in primary immune responses because they induce T cell proliferation more effectively than any other APC.

Remarkably, DCs are able to present internalized antigens to T cells via MHC class I as well as class II molecules, in a phenomenon called **cross-presentation**. This results in activation of cytotoxic T lymphocytes (CTLs), which are then available for killing of infected cells.

Macrophages and B cells express appropriate co-stimulatory molecules for activation of naive T cells only upon infection. Macrophages:

- ingest microbes and particulate antigens;
- digest them in phagolysosomes; and
- present fragments at the cell surface on MHC molecules.

Q. A number of bacterial components enhance the expression of MHC molecules and co-stimulatory molecules on macrophages. What effect would you expect this to have on the immune response? Would it be advantageous for the individual?

A. In the presence of infection, the action of the microbial components would enhance the ability of macrophages to present antigen to T cells. This would generally be advantageous because it would allow the immune system to respond more effectively to the infection. However, in some circumstances it might be disadvantageous because microbial components would also enhance unwanted immune responses such as autoimmune reactions.

B cells can:

- bind to a specific antigen through surface IgM or IgD;
- internalize it; and
- then degrade it into peptides, which associate with MHC class II molecules.

If antigen concentrations are very low, B cells with high-affinity antigen receptors (IgM or IgD) are the most effective APC because other APCs simply cannot capture enough antigen. Therefore, for secondary responses, when the number of antigen-specific B cells is high, B cells may be a major type of APC.

B cells do not normally express co-stimulatory molecules such as B7, but these can be induced by bacterial constituents.

The properties and functions of some APCs are summarized in Figs 7.3 and 7.4.

A series of molecular interactions ensures highly specific antigen recognition by T cells

Antigens are processed before they are presented to T cells

Antigen processing involves degrading the antigen into peptide fragments. The vast majority of epitopes recognized by T cells are fragments from a peptide chain (see Chapter 5).

Only a minority of peptide fragments from a protein antigen are able to bind to a particular MHC molecule. Furthermore, different MHC molecules bind different sets of peptides (see Chapter 5). For example, studies using a viral antigen that is recognized by mouse strains of several different haplotypes (that is, having different MHC molecules) showed that TH cells from each haplotype recognized a distinct peptide from that antigen (Fig. 7.5). This depended largely on whether the strain possessed an MHC class II molecule that could bind to the particular peptide.

Q. The APCs of BALB/c mice present peptide 10–26 of the λ-repressor, whereas APCs of the C57BL/6 mouse present peptide 70–86 (see Fig. 7.5). Which of these peptides would be presented by APCs of an F1 mouse derived from a cross of these two strains?

A. Because MHC molecules are co-dominantly expressed, the APCs in the F1 strain will express both types of APC, and will therefore be able to present both peptides.

Antigen presentation

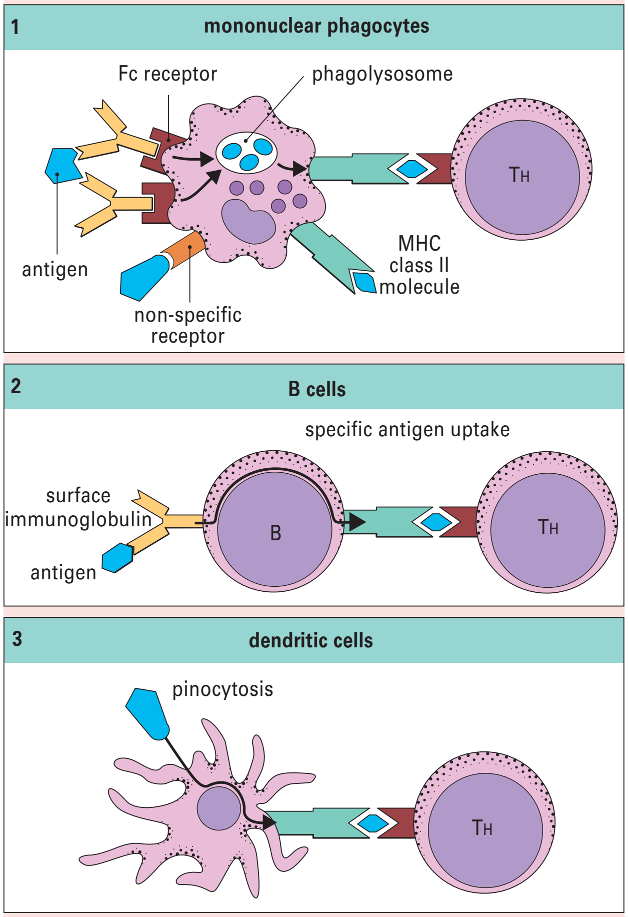

Fig. 7.3 Mononuclear phagocytes (**1**), B cells (**2**), and DCs (**3**) can all present antigen to MHC class II restricted TH cells. Macrophages take up bacteria or particulate antigen via non-specific receptors or as immune complexes, process it, and return fragments to the cell surface in association with MHC class II molecules. Activated B cells can take up antigen via their surface immunoglobulin and present it to T cells associated with their MHC class II molecules. DCs constitutively express MHC class II molecules and take up antigen by macropinocytosis.

Antigen-presenting cells

	phagocytosis	type	location	class II expression
phagocytes (monocyte/ macrophage lineage)	+	monocytes	blood	(+)→+ + + inducible
		macrophages	tissue	
		marginal zone macrophages	spleen and lymph node	
		Kupffer cells	liver	
		microglia	brain	
non-phagocytic constitutive APCs	–	Langerhans' cells	skin	+ + constitutive
		interdigitating DCs (IDCs)	lymphoid tissue	
		follicular dendritic cells	lymphoid tissue	–
lymphocytes	–	B cells and T cells	lymphoid tissues and at sites of immune reactions	– →+ + inducible
facultative APCs	+	astrocytes	brain	inducible
	–	follicular cells	thyroid	inducible
		endothelium	vascular and lymphoid tissue	– →+ + inducible
		fibroblasts	connective tissue	
		other types in appropriate tissue		

Fig. 7.4 Many APCs are unable to phagocytose antigen, but can take it up in other ways, such as by pinocytosis. Endothelial cells (not normally considered to be APCs) that have been induced to express MHC class II molecules by interferon-γ (IFNγ) are also capable of acting as APCs, as are some epithelial cells. Another example is the thyroid follicular cell, which acts as an APC in the pathogenesis of Graves' autoimmune thyroiditis.

T cell antigenic peptides in the λ repressor protein

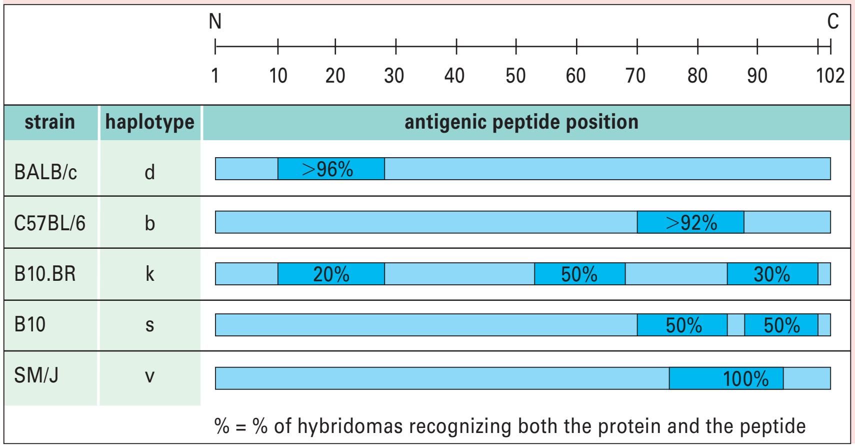

Fig. 7.5 Diversity of antigenic peptides in relation to MHC class II molecules. Mice were immunized with λ repressor protein. T cell hybridomas generated from mouse cells were tested against a panel of overlapping peptides, which spanned the entire protein. Positions of antigenic peptides are shown in dark blue. One peptide was always immunodominant, though more than one peptide was antigenic in some mouse strains. Usually only a few peptides from a protein are stimulators. (Adapted from data of Roy S, Scherer MT, Briner TJ, et al. Science 1989;244:572–575)

Originally researchers carrying out these experiments would have to use particular antigenic peptides or whole molecules presented by APCs of the appropriate type. Now it is possible to use synthetic MHC–peptide combinations (Method Box 7.1).

METHOD BOX 7.1
Tetramers

Tetrameric complexes of MHC molecules may be used to identify clones of T cells specific to presentation of a particular peptide in a particular MHC molecule. These are usually tetramers of MHC class I molecules, though MHC class II tetramers are being developed.

The tetramer assay allows the detection of antigen-specific cytotoxic T lymphocytes (CTLs) in ex-vivo cell preparations without the need for in-vitro expansion.

The basis of this technique is to directly stain antigen-specific T cells with the MHC molecule–peptide complex that they recognize via their T cell receptors (TCRs).

Because the affinity of single TCR molecules for the MHC molecule–peptide is low, direct staining is not possible because of rapid dissociation. The formation of an MHC tetramer with four copies of the MHC molecule–peptide complex results in a very enhanced avidity for the TCR.

MHC molecules of the appropriate allotype are fused to a peptide (bsp), which acts as a substrate for biotinylation. MHC-bsp and β_2-microglobulin are expressed in *Escherichia coli* mixed with the specific CTL epitopic peptide under folding conditions, and the MHC molecule–peptide complexes thus formed are purified. The complexes are biotinylated and mixed at a ratio of 4:1 with fluorescence-labeled streptavidin to produce the tetramers (Fig. 1). The binding of these labeled tetramers by antigen-specific CTLs can then be analyzed by the fluorescence-activated cell sorter (FACS, see Method Box 2.1, p. 32).

A class I MHC tetramer

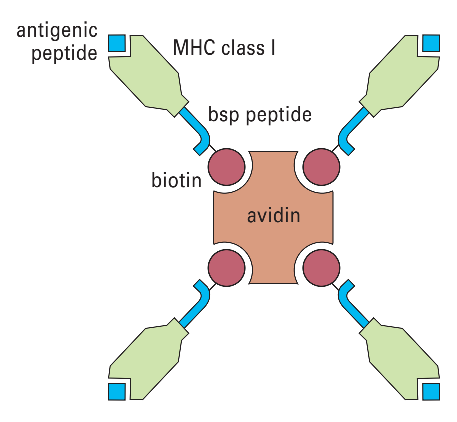

Fig. 1 A tetrameric MHC molecule is produced by biotinylating the MHC molecules via a linking peptide (bsp) and allowing them to bind to the tetravalent molecule avidin – a natural receptor for biotin.

Antigens are partially degraded into peptides before binding to MHC molecules

The processing of antigens to generate peptides that can bind to MHC class II molecules occurs in intracellular organelles (Fig. 7.6). Phagosomes containing endocytosed proteins fuse with lysosomes where a number of proteases are involved in breaking down the proteins to smaller fragments. The proteases include:

- cathepsins B and D;
- an acidic thiol reductase, γ-interferon-inducible lysosomal thiol reductase (GILT), which acts on disulfide-bonded proteins.

Alkaline agents such as chloroquine or ammonium chloride diminish the activity of proteases in the phagolysosomes and therefore interfere with antigen processing.

In laboratory studies, the requirement for internal degradation of antigen by APCs can be circumvented by the use of synthetic peptides. This ability to use synthesized peptides of known sequences has enabled researchers to readily identify epitopes recognized by T cells with different specificities.

The relative importance of different amino acids within a defined epitope can also be investigated by amino acid replacements at different sites. Comparison of the effects of amino acid substitution on MHC molecule binding and T cell reactivity has enabled conclusions to be drawn as to:

- which amino acid residues contact the MHC molecule; and
- which amino acid residues contact the TCR

MHC CLASS I MOLECULES ASSOCIATE WITH ENDOGENOUS PEPTIDES

MHC class I-restricted T cells (CTLs) recognize endogenous antigens synthesized within the target cell, whereas

Antigen processing

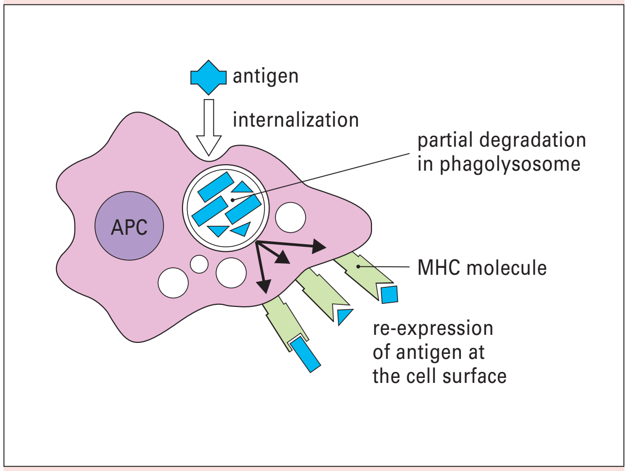
antigen
internalization
partial degradation in phagolysosome
APC
MHC molecule
re-expression of antigen at the cell surface

Fig. 7.6 Exogenous antigens are internalized by APCs and are then degraded by proteolytic enzymes in specialized intracellular compartments. Antigenic peptides associate with MHC class II molecules in vesicles that intersect the endocytic pathway on their way to the cell surface.

class II-restricted T cells (TH) recognize exogenous antigen.

Manipulation of the location of a protein can determine whether it elicits an MHC class I- or class II-restricted response. For example:

- influenza virus hemagglutinin (HA), a glycoprotein associated with the membrane of an infected host cell, normally elicits only a weak CTL response, but influenza virus HA can be generated in the cytoplasm by deleting that part of its sequence that encodes the N terminal signal peptide (required for translation across the membrane of the endoplasmic reticulum [ER]) and, when this is done, there is a strong CTL response to HA;
- the introduction of ovalbumin into the cytoplasm of a target cell (using an osmotic shock technique) generates CTLs recognizing ovalbumin, whereas the addition of exogenous ovalbumin generates an exclusively TH cell response.

Generation of immunoproteasomes by replacement of active subunits

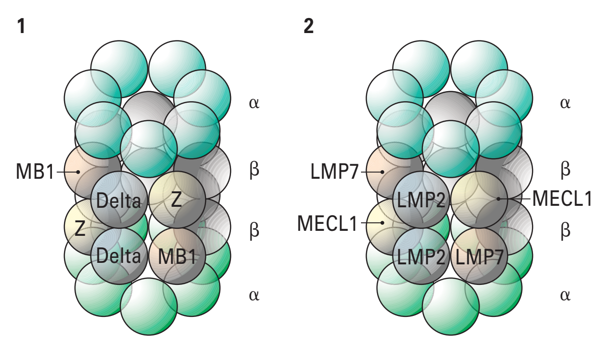

Fig. 7.7 The 20S proteasome, shown in cartoon form, is composed of four stacked disks, two identical outer disks of α subunits, and two similar inner disks comprised of β subunits. Each disk has seven different subunits. Peptides enter into the body of the proteasome for cleavage into peptides. Only three of the β subunits are active. In normal proteasomes, these are called MB1, delta, and Z (**1**). Interferon-γ treatment of cells results in replacement of these three subunits by the two MHC-encoded proteins, LMP2 and LMP7, as well as a third inducible protein, MECL1 (**2**). These subunits are shown adjacent to each other here, whereas they are actually in separate parts of the β ring and some would be hidden at the back of the structure shown.

Proteasomes are cytoplasmic organelles that degrade cytoplasmic proteins

Although the assembly of MHC class I molecules occurs in the ER of the cell, peptides destined to be presented by MHC class I molecules are generated from cytosolic proteins. The initial step in this process involves an organelle called the proteasome – a complex of 14–17 different subunits, which form a barrel-like structure (Fig. 7.7).

Proteasomes provide the major proteolytic activity of the cytosol. They have a range of different endopeptidase activities and they degrade denatured or ubiquitinated proteins to peptides of about 5–15 amino acids (ubiquitin is a protein that tags other proteins for degradation).

Two genes, *LMP2* and *LMP7*, located in the class II region of the MHC (Fig. 7.8), encode proteasome components that subtly modify the range of peptides produced by proteasomes. The expression of these genes is induced by interferon-γ (IFNγ), and the protein products:

- displace constitutive subunits of the proteasome. LMP2 and LMP7 along with a third inducible proteasome component encoded on a different chromosome;
- influence processing of peptides by creating peptide fragments suitable for binding MHC class I molecules from pathogens such as viruses, in addition to self proteins.

Additional subunits associate with the ends of core (20S) proteasomes and may influence antigen processing. These include interferon-inducible PA28 molecules as well as a complex of proteins that result in a larger 26S particle.

Proteasomes may not be the only proteases involved in producing peptides for presentation by MHC class I molecules. There is evidence for the involvement of enzymes, such as the giant tripeptidyl aminopeptidase II (TPPII) complex.

MHC genes involved in antigen processing and presentation

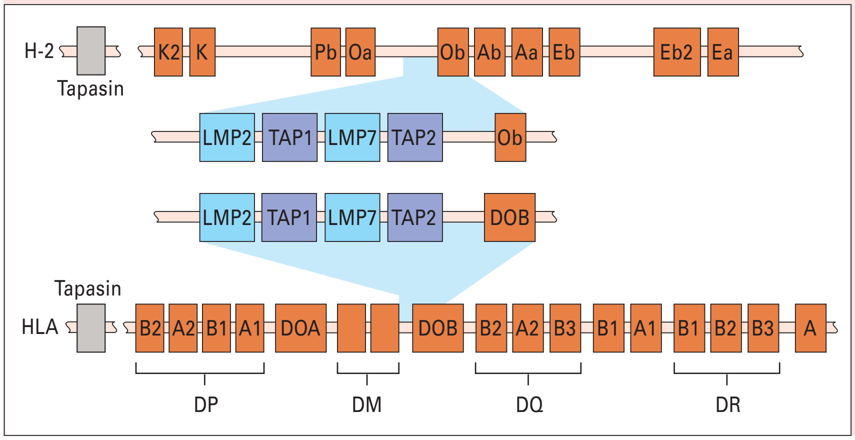

Fig. 7.8 Genes encoding the two subunits of the peptide transporter (TAP) and two components (LMP) of the multisubunit proteasome (see Fig. 7.7) are located in the murine and human class II regions. The Tapasin gene is located just centromeric (left) of the MHC.

Ribosomes are error prone during the course of rapid synthesis of viral proteins and there is evidence that defective ribosomal products (also known as DRiPs) are a direct source of peptides for MHC class I molecules.

Q. What advantage might there be in presenting DriPs on MHC class I molecules rather than fragments of mature viral peptides?

A. DriP-derived peptides will be generated earlier in infection than fragments produced by the degradation of mature viral proteins. This allows the CTLs to recognize virus-infected cells more quickly.

How do the peptides produced by proteasomes traverse the ER?

The products of two genes, *TAP1* and *TAP2*, that map adjacent to *LMP2* and *LMP7* in the MHC (see Fig. 7.8), function as a heterodimeric transporter that translocates peptides into the lumen of the ER. TAP is a member of the large ATP-binding cassette (ABC) family of transporters localized in the ER membrane. Microsomes from cells lacking TAP1 or TAP2 could not take up peptide in experiments in vitro. Using a similar system it was shown that the most efficient transport occurred with peptide substrates of 8–15 amino acids. Although this size is close to the length preference of MHC class I molecule binding sites, it suggests that some additional trimming may be required by enzymes in the lumen of the ER. These include an ER amino peptidase, ERAP, and enzymes such as the TPPII complex may also contribute to trimming the N termini of peptides for class I binding.

Recent evidence has suggested that some of the components of the processing pathway are physically associated in the ER (Fig. 7.9). For example:

- newly synthesized MHC class I molecule–β_2-microglobulin complexes are associated with TAP in the ER;
- another MHC-encoded protein, called **tapasin**, forms a bridge between the MHC class I molecule and TAP, until peptide is associated with the class I molecule; dissociation occurs upon transport to the cis-Golgi.

Peptide loading involves other proteins such as calnexin, calreticulin, and ERp57. These chaperones promote and guide the assembly of stable MHC class I molecule–β_2-microglobulin–peptide complexes.

Assembly of endogenous peptides with MHC class I molecules

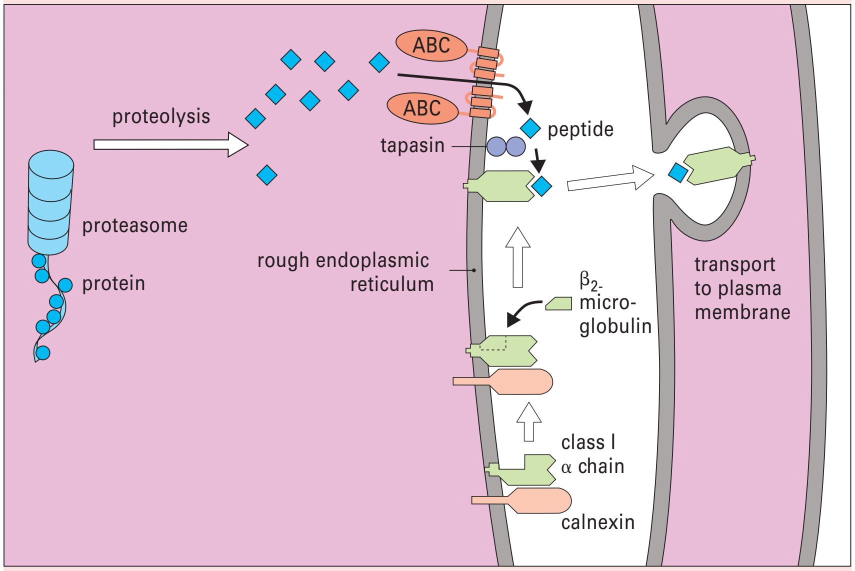

Fig. 7.9 Proposed assembly pathway of antigen–MHC class I molecule complex. Cytoplasmic antigens are processed by proteasomes (see Fig. 7.7). Peptides are transported by two members of the 'ABC' superfamily of transporters, also encoded within the MHC (TAP1 and TAP2). Antigenic peptides associate with class I heavy chains (see Fig. 5.6) and β_2-microglobulin in the ER. Molecular chaperones, such as calnexin and calreticulin, associate with partially assembled class I complexes. The class I molecule waiting to be loaded with peptide is engaged in a complex of proteins. The complex includes calreticulin, a lectin-like chaperone that binds to a sugar residue on the class I molecule. ERp57, a thiol oxido-reductase non-covalently associates with calreticulin and is disulfide bonded to tapasin. The immunoglobulin superfamily molecule tapasin forms a bridge between TAP and the MHC class I molecule, and may also act as a peptide 'editor'. An intact loading complex is essential for efficient MHC class I association with peptide. The fully assembled class I molecules are transported to the cell surface.

MHC class I molecule complexes lacking peptide are unstable, ensuring that only functionally useful complexes are available for interaction with TCRs.

A number of class I-like molecules can present very limited sets of antigens

In addition to the standard MHC class I molecules (class Ia), a number of class I-like molecules (class Ib), encoded in the MHC or elsewhere on the genome, can present very limited sets of antigens.

HLA-E–signal peptide complex interacts with the NKG2A inhibitory receptor on NK cells

HLA-E molecules bind a restricted set of peptides consisting of hydrophobic leader sequence peptides from class Ia molecules. Intriguingly, though these leader sequences are generated by signal peptidase within the ER, HLA-E is dependent on TAP transporters. By binding and presenting sequences from class Ia molecules HLA-E signals the fact that MHC class Ia expression has not been downregulated (e.g. by a virus).

The HLA-E–signal peptide complex interacts with the NKG2A inhibitory receptor on NK cells (see Fig. 10.4). A cell that expresses HLA-E is therefore not killed by NK cells.

Q. Why would conventional HLA-A, HLA-B, or HLA-C molecules be less well suited to the presentation of signal peptide to receptors on NK cells than HLA-E?

A. The conventional MHC molecules have evolved as highly diverse molecules that present the great range of microbial polypeptides to the diverse repertoire of T cell receptors. In contrast HLA-E molecules have a single function – to present well-defined signal peptides to a monomorphic receptor.

CD1 molecules present lipids and glycolipids

CD1 molecules, encoded on chromosome 1, present lipids and glycolipids. Humans have five *CD1* genes and mice have two. CD1b presents the bacterial lipid mycolic acid to T cells with αβ TCRs. Other CD1 molecules are recognized by γδ T cells.

The location of antigen-processing genes in the MHC may not be fortuitous

The finding of a cassette of antigen-processing genes such as the LMPs and TAPs in the class II region of the MHC is striking.

There is some evidence, especially from studies in rats, that particular alleles of *TAP* are genetically linked with alleles of class I genes that are most suited to receive the kind of peptides preferentially transported by the products of that *TAP* allele (Fig. 7.10).

The rat data suggest that localization of some antigen-processing genes in the MHC provides a selective advantage. In fact, the clustering of antigen processing and presenting genes in the MHC of most vertebrate species may not be fortuitous. It may help to coordinate co-evolution of some molecules as well as facilitating exchange of sequences between loci.

In rats, *TAP* genes are polymorphic and different alleles are linked in *cis* to the appropriate class I allele

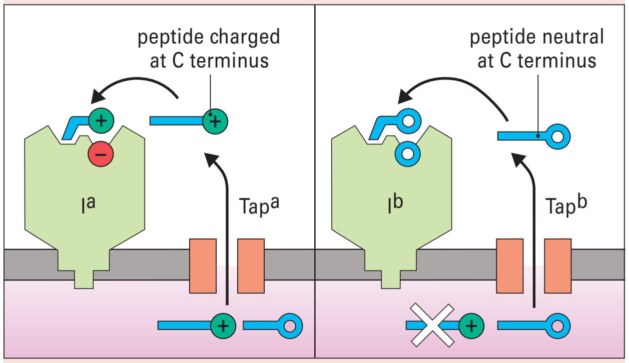

Fig. 7.10 Different MHC class I molecules in rats can accommodate peptides (blue bars) with either a positive charge at the C terminus (+) or a neutral amino acid (o). Similarly, TAP molecules (orange) come in two forms, which differ in the types of peptide they preferentially transport into the ER. Most rat strains have the appropriate *TAP* allele on the same haplotype as the class I gene that it serves best.

MHC CLASS II MOLECULES ARE LOADED WITH EXOGENOUS PEPTIDES

MHC class II molecule α and β chains (see Chapter 5) are found in the ER complexed to a polypeptide called the invariant chain (Ii). This protein is encoded outside the MHC. The αβ–Ii complex is transported through the Golgi complex to an acidic endosomal or lysosomal compartment called MIIC. These MIIC vesicles appear to be specialized for the transport and loading of MHC class II molecules. They have characteristics of both endosomes and lysosomes and have an onion-skin appearance under the electron microscope, comprising multiple membrane structures. The αβ complex spends 1–3 hours in this compartment before reaching the cell surface. The Ii chain is cleaved to small fragments, one of which, termed CLIP (class II-associated invariant peptide), is located in the groove of the class II molecule until replaced by peptides destined for presentation (Figs 7.11 and 7.12).

How do antigenic peptides derived from exogenous proteins meet MHC molecules in the appropriate compartment?

The answer to this question lies in the intracellular traffic routes of MHC molecules. After synthesis in the ER both types of MHC molecule are transported through the Golgi compartment, class I in association with antigenic peptide and class II bound to invariant chain Ii. Class II molecules segregate from class I molecules in the trans-Golgi network. They then join the endosomal/lysosomal MIIC compartment en route to the plasma membrane.

Exogenous antigen can also enter APCs via an endocytic route (either receptor mediated or fluid phase, see Fig. 7.11) where in some cells, such as DCs, it can load onto:

- MHC class II molecules in MIIC vesicles; and

Proposed routes of intracellular trafficking of MHC molecules involved in antigen presentation

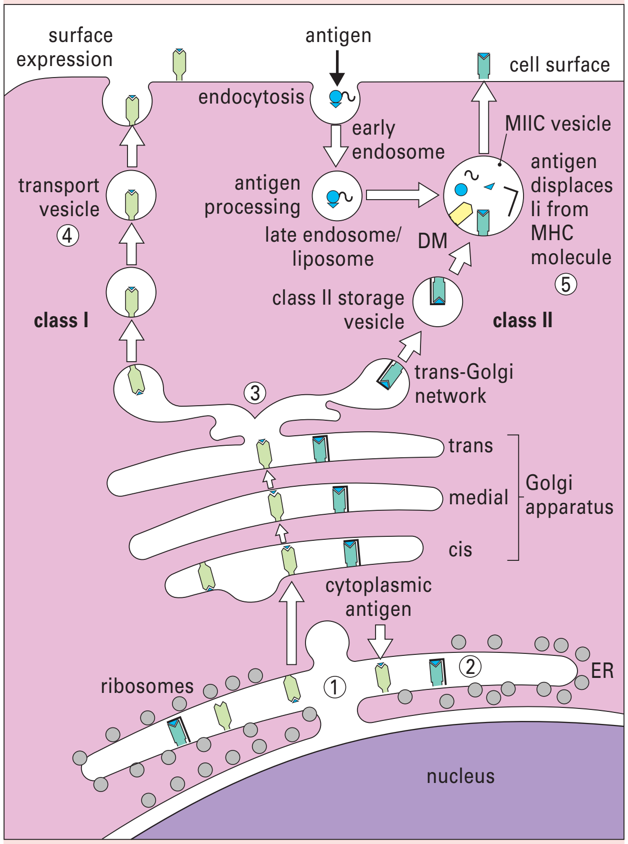
surface expression
antigen
cell surface
endocytosis
early endosome
MIIC vesicle
transport vesicle
④
antigen processing
antigen displaces Ii from MHC molecule
⑤
late endosome/ liposome
DM
class II storage vesicle
class I
class II
③
trans-Golgi network
trans
medial
cis
Golgi apparatus
cytoplasmic antigen
ribosomes
①
②
ER
nucleus

Fig. 7.11 Newly synthesized MHC class I molecules are loaded with peptide (**1**). MHC class II molecules associate with Ii in the ER (**2**). Ii prevents loading with peptide and contains sequences that enable the MHC class II molecule to exit from the rough endoplasmic reticulum (RER). MHC class I and class II molecules segregate after transit through the Golgi (**3**). Class I molecules go directly to the cell surface (**4**). Class II molecules enter an acidic compartment called MIIC, where they are loaded with peptide derived from exogenous antigen, and the CLIP peptide that occupies the binding groove dissociates (**5**).

MHC class II molecule processing compartment

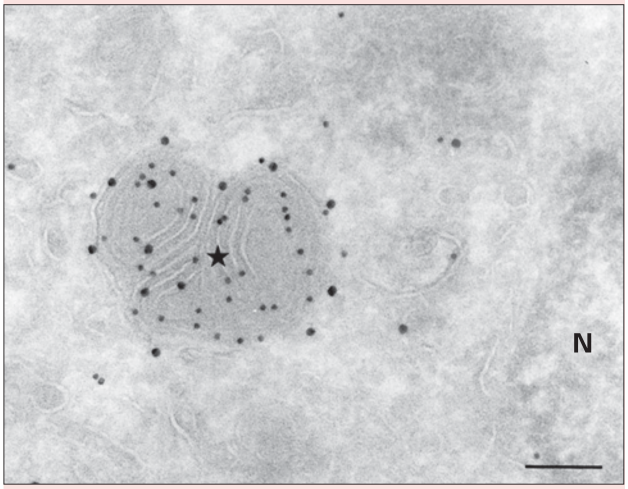

Fig. 7.12 Electron micrograph of ultrathin cryosections from B cells showing a multilaminar MIIC vesicle. The bar represents 100 nm. MHC class II molecules are revealed by antibodies coupled to 10 nm gold particles and HLA-DM by large gold particles (15 nm). (Courtesy of Dr Monique Kleijmeer)

- MHC class I molecules in the ER, in a TAP-dependent manner.

The exchange of CLIP for other peptides is orchestrated by a class II-related molecule called HLA-DM (Fig. 7.13). This glycoprotein consists of an α chain and a β chain, both of which are encoded in the class II region of the MHC (see Fig. 7.8).

HLA-DM acts by stabilizing empty MHC class II molecules so that when CLIP is released other peptides get a chance to associate. The DM molecule itself has a closed groove and it is not capable of binding peptide.

In cell lines lacking *DMA* and *DMB* genes, class II molecules are unstable and the cells no longer process and present proteins. Their class II molecules end up at the cell surface occupied by CLIP fragments of the invariant chain (Fig. 7.14).

HLA-DM acts like a catalyst to influence binding of peptides in exchange for CLIP

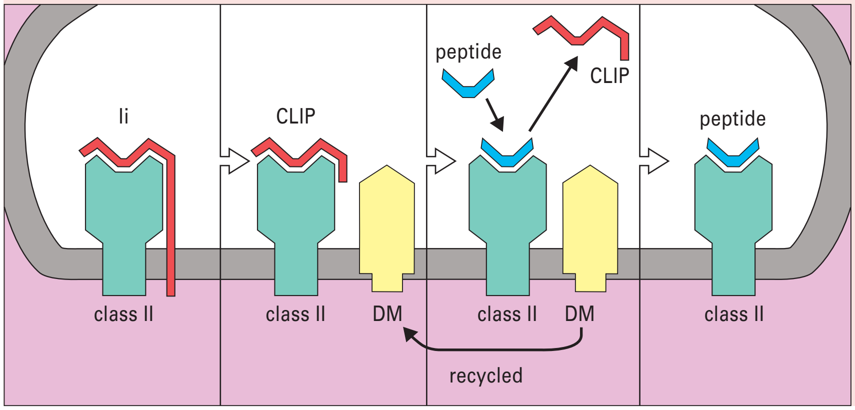
Ii
class II
CLIP
class II
DM
peptide
CLIP
class II
DM
recycled
peptide
class II

Fig. 7.13 An MHC class II molecule is shown loaded with the Ii chain. The Ii chain is cleaved to the CLIP fragment. Association of the complex with HLA-DM then allows CLIP to be exchanged for other peptides derived from endocytosed proteins present in MIIC vesicles.

Absence of HLA-DM leads to failure of MHC class II molecules to bind appropriate peptides

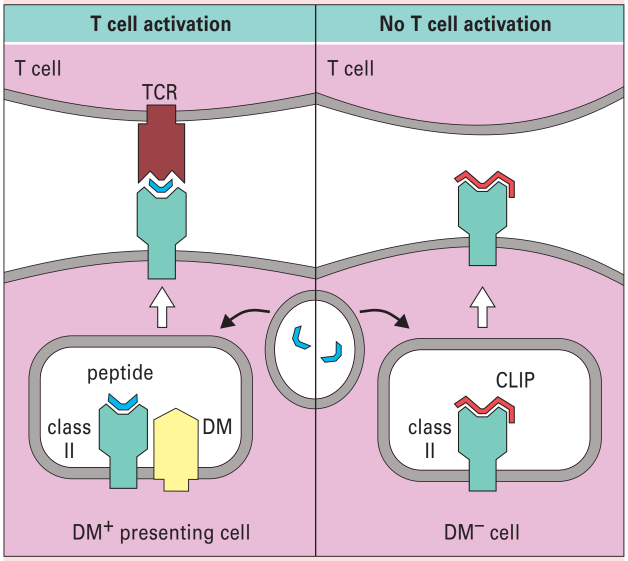

Fig. 7.14 In normal APCs (left) peptides are loaded onto class II molecules, as shown in Fig. 7.13. In the absence of HLA-DM (right), the CLIP fragment is not removed, and the blocked MHC molecule is unable to activate T cells.

A further MHC-encoded molecule, HLA-DO (see Fig. 7.8), which associates with DM, is also involved in peptide loading. Like conventional MHC class II molecules, HLA-DO is a heterodimer (see Chapter 5), consisting of the DOA (formerly DNA) and DOB chains. Similar genes and molecules to DM and DO are found in other species, for example in mouse they are encoded by the Ma, Mb, and Oa, Ob genes, respectively.

CROSS-PRESENTATION ALLOWS APCs TO ACQUIRE ANTIGENS FROM INFECTED CELLS

Cross-presentation occurs when exogenous peptides are presented by MHC class I molecules. The principle that endogenously synthesized proteins are the source of peptides for MHC class I molecules appears to be violated by the phenomenon of cross-presentation. There is a specialized pathway in which T cells recognize MHC class I molecule–peptide complexes formed from exogenous protein antigens endocytically acquired by DCs.

Q. What function could cross-presentation fulfil in defense against viral infections?

A. Presumably such a mechanism is required so that, even if a virus fails to infect an APC directly, a successful cytotoxic CD8 T cell-mediated immune response can be generated.

In this process of cross-presentation, APCs acquire remnants of virus-infected cells by phagocytosis and present them on MHC class I molecules. Internalized proteins appear to be transferred from the endocytic pathway into the cytosol where they are then subject to processing and transport by proteasomes and TAP, respectively, as for other peptides (Fig. 7.15).

CO-STIMULATORY MOLECULES ARE ESSENTIAL FOR T CELL ACTIVATION

The process of activating T cells generally takes place in the lymph node nearest to the infection. The TCR recognizes a specific peptide lodged in the peptide-binding groove of the MHC molecule. This interaction dictates immunological specificity because a peptide associated with an MHC molecule forms a unique structure to be recognized by the TCR. Other molecules have a complementary role in this interaction.

The initial encounter of T cells with APCs is by nonspecific binding through adhesion molecules. This transient binding by adhesion molecules permits the T cell to encounter a large number of different MHC molecule–peptide combinations on different APCs. In the absence of a specific interaction, the APC and T cell rapidly dissociate.

Cross-presentation when exogenous peptides are presented by MHC class I molecules

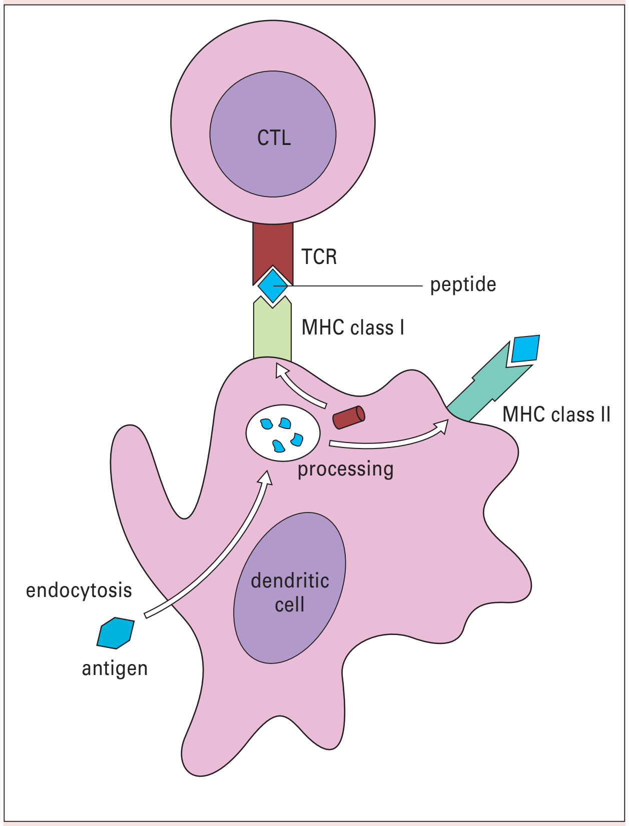

Fig. 7.15 DCs can present internalized antigen via both the MHC class I and class II pathways. This is atypical – external antigen is usually presented via the class II pathway. Peptides are transferred from the endosome to the cytosol where they are processed by the proteasome to enter the class I pathway.

Crucially, co-stimulatory molecules act together with the antigen-specific signals before the T cell is sanctioned for proliferation. Co-stimulatory and antigen-specific signals must be present simultaneously on the same cell.

Overall, antigen presentation through MHC class I or class II molecules can be split into four stages – adhesion, antigen-specific activation, co-stimulation, and cytokine signaling (see Fig. 7.1).

Multiple cell surface molecules interact during antigen presentation to T cells

Intercellular adhesion molecules (ICAMs), particularly ICAM-1 (CD54), interact with the integrin, lymphocyte functional antigen-1 (LFA-1 or CD11a/CD18), present on all immune cells.

If mouse cells are transfected with both human MHC and human ICAM-1, their capacity to act as human APCs is augmented.

When the T cell encounters the appropriate MHC molecule–peptide, which happens rarely except during an ongoing infection, a conformational change in LFA-1 on the T cell, signaled via the TCR, results in tighter binding to ICAM-1, which results in prolonged cell–cell contact. The joined cells can exist as a pair for long periods, allowing time for the T cell to proliferate and differentiate.

The specific MHC molecule–peptide–TCR interaction, though necessary, is not sufficient to fully activate the T cell. A second signal is required, otherwise the T cell will become unresponsive. This second signal, also referred to as co-stimulation, is of crucial importance.

Some co-stimulatory molecules that interact with ligands on the T cell's surface are shown in Fig. 7.16.

The most potent co-stimulatory molecules known are B7s, which are members of the immunoglobulin superfamily molecules; they include B7-1 (CD80) and B7-2 (CD86).

Several other B7-related molecules are beginning to emerge.

B7s exist as homodimers on the cell surface. These proteins are constitutively expressed on DCs, but can be upregulated on monocytes, B cells, and probably other APCs.

Upregulation of co-receptors is stimulated by inflammation and by interaction of microbial products with Toll-like receptors (TLRs, see Fig. 6.24) on the APC.

Co-receptors are the ligands for other immunoglobulin superfamily molecules – CD28 and its homolog CTLA-4 (CD152), which is expressed after T cell activation. CD28 is the main co-stimulatory ligand expressed on naive T cells. CD28 stimulation:

- has been shown to prolong and augment the production of IL-2 and other cytokines; and
- is probably important in preventing the induction of tolerance.

Although the CD28–B7 interaction is extremely important, CD28 knockout mice do respond to antigen, but require higher doses, so CD28 triggering is not obligatory, even for naive T cells. In CD28 knockout mice other co-stimulatory signals probably replace that delivered by CD28–B7.

CTLA-4, the alternative ligand for B7, is an inhibitory receptor limiting T cell activation, resulting in less IL-2 production. Thus CD28, constitutively expressed, initially interacts with B7, leading to T cell activation. Once this has peaked, the upregulation of CTLA-4 with its higher affinity limits the degree of activation because available B7 will interact with CTLA-4 (Fig. 7.17).

Q. What effect would you expect to see in mice that have the CTLA-4 gene knocked out?

A. They suffer from an aggressive lymphoproliferative disorder, because they do not inactivate dividing T cells efficiently (see Fig. 7.17).

Critical molecules involved in antigen presentation

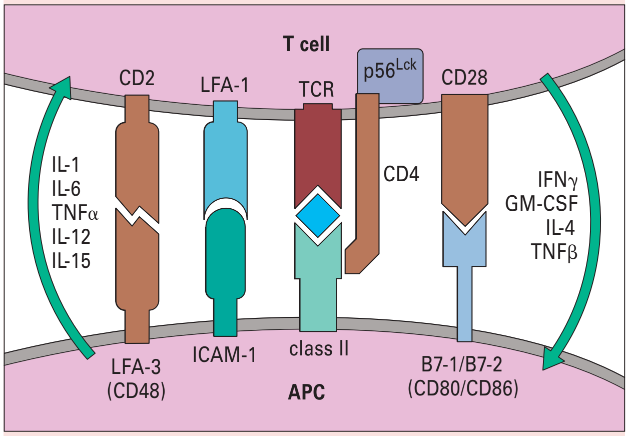

Fig. 7.16 The molecules involved in the interaction between T cells and APCs. The various cytokines and their direction of action are also shown. In humans LFA-3 (CD58) acts as a ligand for CD2, but in rodents CD48 performs this function.

Role of CTLA-4 in controlling T cell activation

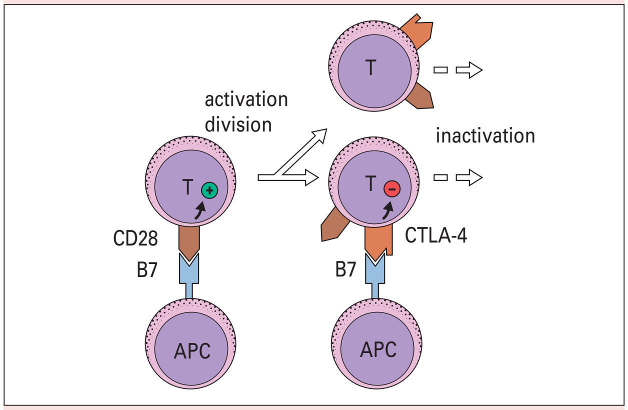

Fig. 7.17 Before activation, T cells express CD28, which ligates B7-1 and B7-2 on APCs (e.g. B cells). After activation, CTLA-4 is expressed, which is an alternative high-affinity ligand for B7. CTLA-4 ligates B7, so the T cells no longer receive an activation signal.

The CD2 molecule on T cells is also involved in T cell activation, in conjunction with the TCR. CD2 is a receptor for LFA-3 (CD58), which is widely distributed on cells and is present on all APCs. In rodents CD48 binds to CD2 and appears to be functionally equivalent to LFA-3 in humans. Both CD2 and LFA-3 are members of the immunoglobulin superfamily.

Co-stimulation via B7 is a necessary signal for T cell activation

Resting T cells cannot respond optimally without co-stimulation via B7, and if they recognize their antigen in a non-stimulating manner they become inactivated, producing a state of immunological tolerance (Fig. 7.18). This tolerance is specific, only affecting TH cells that respond to a particular antigen. Persistent tolerance without cell death is known as **clonal anergy**.

Duel signaling is necessary for full T cell activation

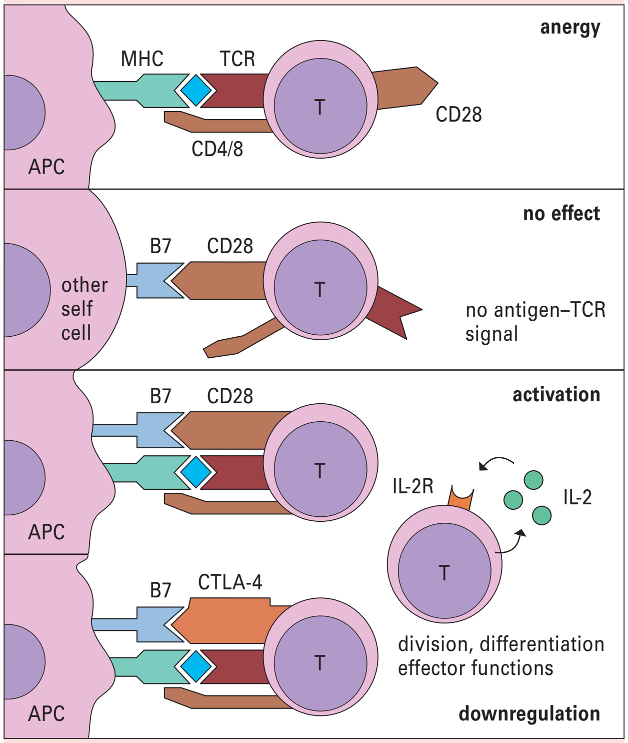

Fig. 7.18 A T cell requires signals from both the TCR and CD28 for activation. Anergy – in the absence of co-stimulatory molecules inactivation or anergy results. This situation would prevail to tolerize T cells not removed by central tolerance to self antigens expressed on peripheral tissues. No effect – in the absence of an antigen-specific signal (e.g. wrong peptide) there is no effect on the T cell. Activation – co-reception of both signals, from the surface of a professional APC, activates the T cell to produce IL-2 and its receptor (IL-2R). The cell divides and differentiates into an effector T cell, which no longer requires signal 2 for its effector function. Downregulation – at the termination of the immune response, CTLA-4 replaces CD28 and downregulates T cell function.

As discussed in Chapters 2 and 12, many T cells are made tolerant to self molecules during their development in the thymus. However, the process is not perfect and many autoreactive T cells are present in the circulation. These cells may still be tolerized outside the thymus because, although they are specific for self MHC–peptide, this is often first encountered on cells that lack the appropriate co-stimulatory signals.

CD4 AND CD8 BINDING TO MHC MOLECULES INCREASE T CELL BINDING AFFINITY

CD4 and CD8 co-receptors for MHC molecules increase the sensitivity of T cells to antigen-presenting molecules by about 100-fold. The two major subsets of T cells, TH and CTLs, express either CD4 or CD8, respectively.

CD4 and CD8 are only distantly related, but serve similar functions as co-receptors for MHC molecule–peptide–TCR interactions. In association with the MHC molecule–peptide–TCR complex:

- CD8 interacts with MHC class I molecules and consists of two chains with immunoglobulin-like domains on long stalks, coupled by a disulfide bridge;
- CD4 interacts with MHC class II molecules and consists of four immunoglobulin domains external to the cell, and its cytoplasmic tail has a docking site for the tyrosine kinase, Lck (see below);

CD8 co-receptor binds to conserved portions of the α_3 domains of class I molecules (see Fig. 5.6), leaving the portion enclosing peptide free to interact with the TCR. Like CD4, CD8 associates with Lck.

Both CD4 and CD8 co-receptors increase the sensitivity of T cells to antigen-presenting molecules by about 100-fold.

THE HIGHLY ORDERED AREA OF CONTACT BETWEEN A T CELL AND AN APC IS AN IMMUNOLOGICAL SYNAPSE

The junction between a T cell and an APC does not just consist of interacting molecules arranged randomly on the cell surfaces. There is a clustering of TCRs in a patch on the cell surface in an area of membrane with distinct properties, surrounded by a ring of adhesion molecules.

Immunological synapse formation is thought to be an active dynamic process that helps T cells to distinguish potential antigenic ligands.

Experiments show that initially TCR ligands are relegated to the outer ring of the nascent synapse. Subsequently, there is transport of the complexes to a central cluster, which is dependent, in an unknown way, on TCR–ligand interaction kinetics.

Productive T cell proliferation appears to depend on the formation of a stable central cluster of TCRs interacting with MHC molecules (Fig. 7.19). The final configuration of the synapse has MHC molecule–peptide–TCR interactions at the hub of the synapse ringed by ICAM-1–LFA-1 interactions (Fig. 7.20). Large

Diagram of an immunological synapse

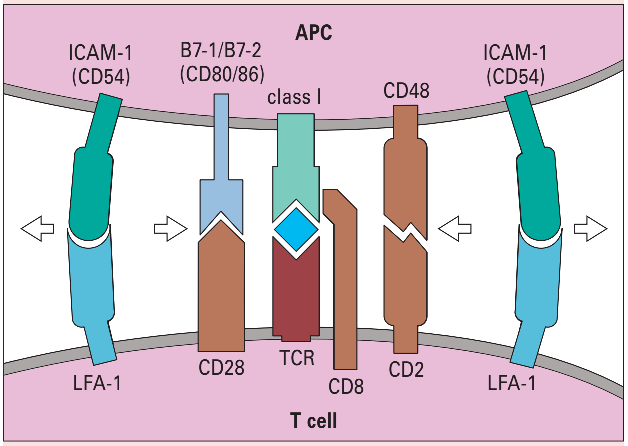

Fig. 7.19 Sets of receptors and their ligands – TCR–peptide–MHC molecule; CD2–CD48; CD28–CD80 – all contribute to the formation of a synapse. The TCR assembles with MHC molecule–antigen peptide molecules and this moves to form a cluster. LFA-1–ICAM-1 interactions are relegated to the perimeter of the raft.

Color-enhanced reconstruction of an immunological synapse

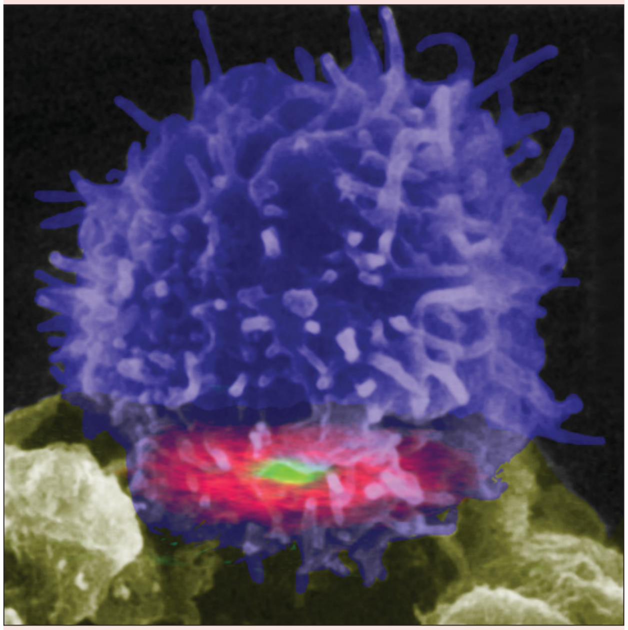

Fig. 7.20 A live-cell fluorescence image, showing the peripheral zone of adhesion molecules (red) surrounding the core containing TCRs (green), is superimposed on a scanning electron micrograph of a T cell (purple) interacting with a DC (dark green). (Courtesy of Dr Mike Dustin and Science)

CD45 molecules, also known as the leukocyte common antigen, are excluded to an outer ring. The precise arrangement of the immunological synapse is cell specific and may be different on T cells and other cells such as NK cells.

Although the interaction between CD4$^+$ T cells and APCs has been reasonably well studied, that between CD8$^+$ T cells and APCs is less well understood. CD4$^+$ cells help in the activation of most CD8$^+$ T cells. Because a single interdigitating DC (IDC) can bind many T cells, it has been proposed that activation takes place in a cluster of CD4$^+$ and CD8$^+$ T cells gathered on the surface of IDCs.

T CELL ACTIVATION LEADS TO PRODUCTION OF IL-2 AND HIGH-AFFINITY IL-2R

Receptors on the surface of T cells signal to the interior of the cell using signal transduction pathways that are common to many other cell types, including B cells. A key principle is the clustering of receptors upon ligand binding. This leads to activation of associated tyrosine kinases, which phosphorylate tyrosine residues in the cytoplasmic tails of the clustered receptors, followed by recruitment of additional kinases and signaling molecules, in cascades. The end-result of these complex pathways is the induction of gene synthesis by the activation of transcription factors (Fig. 7.21).

Activation is initiated by the actions of tyrosine kinases

TCR α and β chains are associated with the CD3γ, δ, and ε molecules, the ζ and η chains (see Fig. 5.3), and the enzyme Lck (p56^{lck}), which is attached to the intracellular portions of CD4 or CD8 (see Fig. 7.21). The label p56^{lck} signifies a lymphocyte-specific tyrosine kinase of 56 kDa.

Recognition of an antigen–MHC molecule complex by the TCR and ligation of co-stimulatory molecules initiates the signaling events.

The earliest events involve tyrosine kinases of the Src family, particularly Lck associated with CD4 and CD8, and Fyn, which phosphorylate target sequences found in the ζ chain (also present on Igα, Igβ, and FcγR, see Fig. 3.20), termed **immunoreceptor tyrosine-based activation motifs (ITAMs)**.

Phosphorylation of the ITAMs initiates a series of steps (see Fig. 7.21), that lead to the activation of transcription factors and their translocation to the nucleus. The transcription factors act on genes required for T cell activation, including the IL-2 and the **IL-2 receptor (IL-2R)** genes.

Q. What occurs when a cell generates both a cytokine and the receptor for that cytokine?

A. This allows autocrine stimulation, where the cell can activate itself as well as similar cells in the immediate vicinity.

Cell division occurs after IL-2 production and is secondary to ligation of the IL-2R.

Intracellular signaling in T cell activation

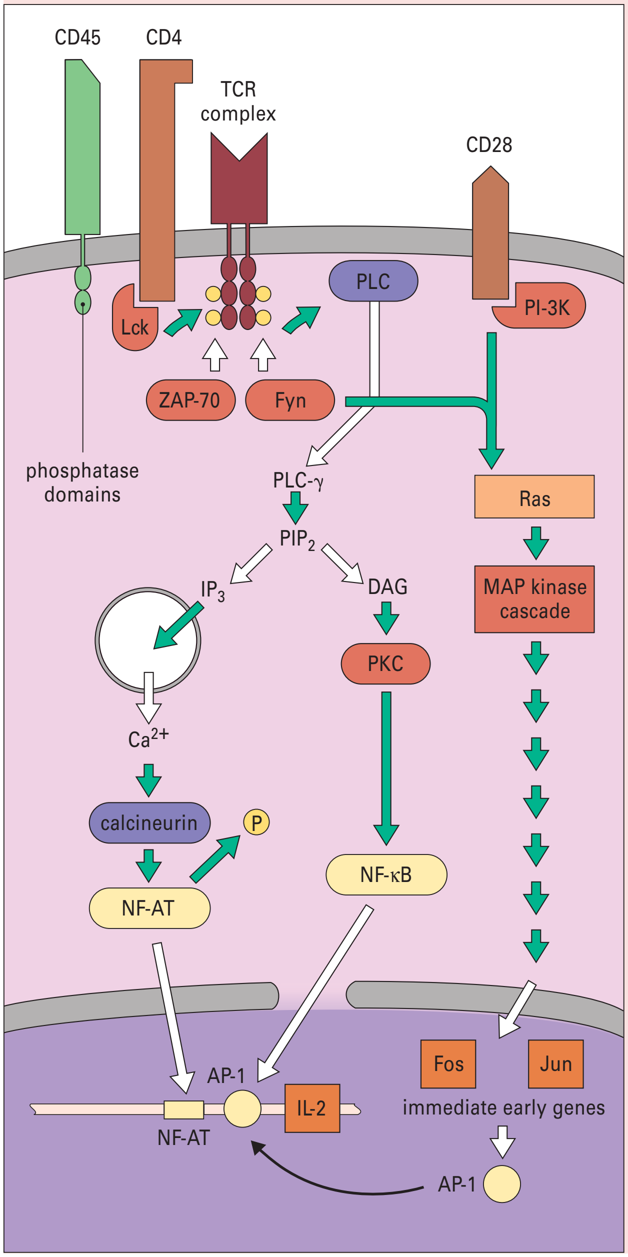

Fig. 7.21 T cell activation involves the transduction of signals from both the TCR and CD28. Clustering of surface receptors such as TCR, CD4, CD28, and CD45 results in activation of tyrosine kinases Fyn and Lck. CD4, which is associated with the TCR complex, binds to the kinase, Lck. Such kinases become activated by dephosphorylation, possibly by phosphatase domains on CD45 (leukocyte common antigen). Lck can now phosphorylate the ITAM domains on the γ chains of CD3 (see Figs 5.3 and 5.15), which allows them to associate with other kinases including Fyn and ZAP-70. Fyn activates phospholipase C (PLCγ), which leads to two pathways by the cleavage of phosphatidylinositol bisphosphate (PIP$_2$) into diacylglycerol (DAG) and inositol trisphosphate (IP$_3$). IP$_3$ releases Ca^{2+} from intracellular (ER) stores to activate calcium-dependent enzymes such as calcineurin. Calcineurin removes phosphate from the transcription factor NF-AT (nuclear factor of activation of T cells), which causes its translocation to the nucleus. DAG activates protein kinase C (PKC), which then activates the transcription factor NF-κB. Meanwhile, ZAP-70, Fyn, and PI-3 kinase (PI-3K) (associated with CD28) integrate signals via kinase cascades in the cytoplasm that activate specific transcription factors. Adapter proteins are used to link the various receptors to the common intracellular signaling components. GTP-binding proteins (G proteins) are involved in activating a set of protein kinases in the MAP kinase cascade. The transcription factors translocate to the nucleus to activate genes, including 'immediate early genes' Fos and Jun for cell division and the promoter AP-1, which acts with NF-AT on the IL-2 gene.

In summary, TCR stimulation results in activation of a variety of tyrosine kinases and downstream effectors that regulate cellular responses, resulting in the regulation of IL-2 gene expression, which is largely responsible for activating the T cell.

B cell activation and T cell activation follow similar patterns

In B cells, the function of CD3 is replaced by Igα and Igβ, which also carry ITAM motifs in their cytoplasmic tails. Cross-linking of surface immunoglobulin leads to activation of the Src family kinases, which in B cells are Fyn, but also Lyn and Blk. A series of analogous activation steps occurs in B cells (see Chapter 8).

An activating signal is integrated from both the antigen receptor and co-stimulatory molecules

It is still not clear exactly what is an effective antigenic signal. For a T cell, interaction at a single TCR is not sufficient, but exactly how many interactions are necessary may depend on other lymphocyte stimulatory signals and the type and activation state of the T cell being stimulated.

Murine T cell hybrids (derived from the fusion of a normal T cell with a tumor T cell) are known to be easily triggered. To stimulate a T cell hybrid:

- an effective APC, such as a macrophage, need carry no more than 60 MHC class II molecule–antigen fragment complexes;

- a weak APC, such as a class II-transfected fibroblast, needs 5000 such complexes.

Recent work has suggested that about 8000 TCR molecules need to interact with MHC molecule–peptide complexes and become activated, as judged by their loss from the T cell surface, for T cell clones to be triggered.

As TCR–MHC molecule interactions are of low affinity, one MHC molecule–peptide complex could activate multiple TCRs.

In the presence of a co-stimulatory signal such as CD28–B7 ligation (see Fig. 7.16), 1500 activated and internalized TCRs are sufficient for the activation of T cell clones.

As discussed above, interaction at the TCR or membrane immunoglobulin alone cannot mediate a positive activation signal for T or B cells. On the contrary it may induce a negative or tolerogenic signal.

Co-stimulatory molecules such as CD2 and CD11a/CD18 (LFA-1) (see Fig. 7.16) are not responsible only for binding – their cytoplasmic domains are also involved in signaling. For example, experimental deletion of the intracytoplasmic domain of CD2 interfered with activation, but left its adhesion function unaltered.

CD28 ligation appears to result in induction of IL-2 synthesis by a pathway distinct from that regulated by calcineurin.

As shown in Fig. 7.21, the induction of IL-2 gene expression requires integrated signals from the activation of PLC, Ras, and the calcium-dependent serine phosphatase calcineurin.

The IL-2 enhancer contains a binding site for a nuclear factor, **NF-AT (nuclear factor of activation of T cells)**, that is induced upon T cell activation. NF-AT is activated by calcineurin. It then translocates to the nucleus where it interacts with Fos and Jun to induce IL-2 gene expression.

IL-2 expression is dependent on the additional signals initiated upon ligation of CD28 by B7 (see Figs 7.19 and 7.21). How the signal from CD28 complements the TCR signal is not established, but it may act by stabilizing IL-2 mRNA as well as increasing transcription.

Ciclosporin, an immunosuppressive drug, causes inhibition of the phosphatase activity of calcineurin.

Rapamycin, another immunosuppressive agent, blocks signaling downstream of IL-2 activation.

Mitogens and superantigens can also activate T cells

T cells can also be activated by some mitogens. Most T cells are stimulated by:

- phytohemagglutinin (PHA), a lectin isolated from red kidney beans; or
- concanavalin-A, extracted from castor beans.

These molecules are able to bind to T cell surface molecules including the TCR complex and CD2, causing them to cluster on the cell surface, thereby mimicking the clustering caused by antigen presentation. Such mitogens, however, will activate the T cells regardless of their antigen specificity.

Another group of molecules that can activate T cells non-specifically are the **superantigens**. These are mostly of bacterial origin and include:

- staphylococcal enterotoxins (responsible for some types of acute food poisoning);
- toxic shock syndrome toxin (responsible for tampon sepsis-induced shock); and
- exfoliative dermatitis toxin.

Superantigens bind to MHC class II molecules on APCs and are recognized by TCRs, but not in the same way as an MHC molecule–peptide complex. Binding is to the Vβ chain of the TCR (see Fig. 5.2) alone. Depending on experimental conditions, the effects are the same as with antigen in that either activation or clonal anergy may occur (Fig. 7.22). A key point about superantigens is that they stimulate families of clones using common Vβ gene segments.

CTLA-4 expression downregulates T cells

The role of the B7 counter-receptor CTLA-4 (CD152) in downregulating T cell activation has been referred to above (see Fig. 7.17).

The negative signal generated by CTLA-4 may be propagated through the cytoplasmic tail of the molecule, which bears a motif known as an **ITIM (immunoreceptor tyrosine-based inhibitory motif)**, which has the characteristic sequence IxYxxL. ITIM motifs are thought to work by recruiting inhibitory phosphatases such as SHIP, which remove phosphate groups added by tyrosine kinases.

T cell stimulation by superantigens

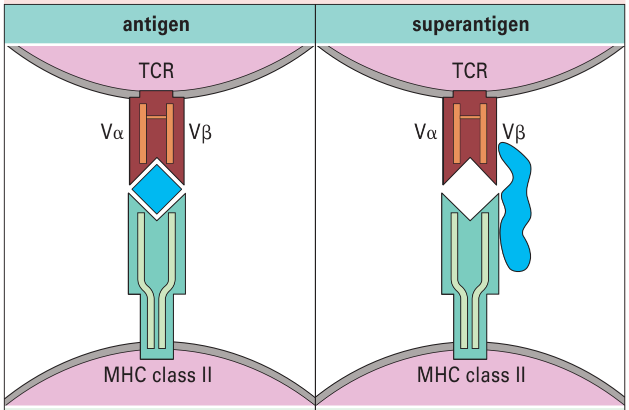

antigen	superantigen
recognition of antigenic peptide in MHC groove – requires Vα, Jα, Vβ, Dβ, Jβ	recognition of Vβ alone
low frequency of antigen-specific cells (<0.01%)	high frequency of antigen-specific cells (≤10%)

Fig. 7.22 Antigenic peptides must normally be processed and presented on MHC molecules to trigger the TCR. However, superantigens, such as staphylococcal enterotoxins, are not processed, but bind directly to MHC class II molecules and Vβ of the TCR. Each superantigen activates a distinct set of Vβ-expressing T cells, depending on which Vβ gene segment the T cell is expressing.

Q. Which other receptors have ITIMs?

A. The inhibitory Fc receptor FcγRIIb expressed on B cells has ITIMs (see Fig. 3.20). Other inhibitory receptors with similar motifs include the killer inhibitory receptors (KIRs) expressed on NK cells as well as T cells (see Figs 10.3 and 10.5).

Interleukin-2 drives T cell division

Apart from the cell surface interactions, cytokines, acting locally, are required for promoting T cell division (see Fig. 7.18).

Triggering of the TCR along with co-receptors results in IL-2 synthesis by the T cell itself. In most CD4$^+$ cells and some CD8$^+$ T cells, there is a transient production of IL-2 for 1–2 days. During this time the interaction of IL-2 with the high-affinity IL-2R results in T cell division.

On resting T cells, the IL-2R is predominantly present as a low-affinity form consisting of two polypeptide chains, a β chain (p75) that binds IL-2 and a γ chain that signals to the cell. When the T cell is activated, it produces an α chain (p55 = CD25), which contributes to IL-2 binding and, together with the β and γ chains, forms the high-affinity receptor (Fig. 7.23).

The transient expression of the high-affinity IL-2R for about 1 week after stimulation of the T cell, together with the induction of CTLA-4, helps limit T cell division. In the absence of positive signals, the T cells will start to die by apoptosis.

An understanding of the requirements for T cell division has allowed researchers to generate long-lived T cell lines, which are used in many areas of immunological research (Method Box 7.2).

Other cytokines contribute to activation and division

Other cytokines may also contribute to T cell proliferation. For example, IL-1 and IL-6 induce the expression of IL-2R on resting T cells.

Q. Which cells generate IL-1 and IL-6?

A. Mononuclear phagocytes among others. Thus the production of these cytokines enhances their antigen-presenting function.

Additionally IL-4 and IL-15 may provide stimulatory signals through their own receptors.

IL-12 is also of major importance in T cell activation:

- helping to enhance IFNγ production; and
- directing naive T cells to develop into TH1 cells.

IL-15 made by APCs can also induce T cell proliferation and may be very important before IL-2 is produced.

ANTIGEN PRESENTATION AFFECTS THE COURSE OF AN IMMUNE RESPONSE

APCs may be activated rapidly in an immune response, for example by:

- the immunogenic entity itself, in the case of bacteria and some viruses; or
- the antigen in conjunction with the adjuvant component of a vaccine.

METHOD BOX 7.2
T cell lines

To test the response of lymphocytes to a specific antigen that is being presented, the lymphocyte stimulation test (Fig. 1) may be used. This test measures the response of the T cells to antigen as indicated by their entering the cell cycle and incorporating precursors of DNA synthesis.

The lymphocyte stimulation test

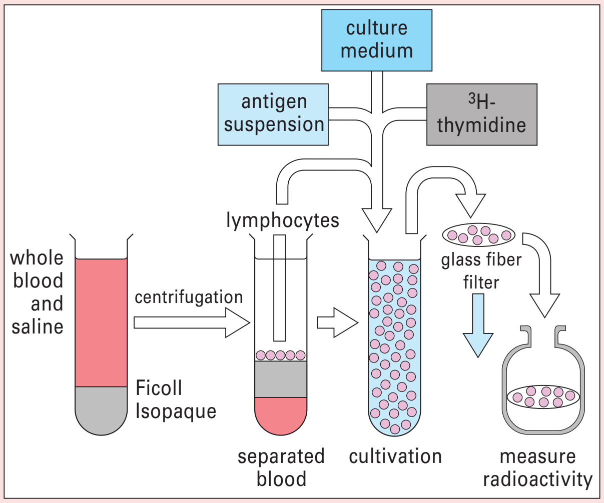

Fig. 1 In the lymphocyte stimulation test, whole blood in saline solution is first layered on Ficoll Isopaque (which has a density between, and therefore separates, white cells and red cells) and centrifuged (400 *g*). This separates the lymphocytes from the other cell and serum constituents. The cells are washed (to remove contaminants such as antigen) and then put into test tubes with a suspension of antigen and culture medium. Tritiated thymidine (^{3}H-thymidine) is added 16 hours before the cells are harvested. The cells are harvested on a glass-fiber filter disk and their radioactivity measured by placing the disk in a liquid scintillation counter. A high count indicates that the lymphocytes have undergone transformation and confirms their responsiveness to the antigen. This test can also be used for cells from lymphoid tissue.

Antigen presentation is not a unidirectional process. T cells, as they become activated:

- release cytokines such as IFNγ and granulocyte–macrophage colony stimulating factor (GM–CSF);
- express surface molecules such as CD40 ligand, which enhance antigen presentation.

When APCs are activated, they express more MHC class I and II molecules, Fc receptors, and co-stimulatory adhesion molecules, including B7-1 and B7-2, CD11a/b/c, ICAM-1, and ICAM-3. They also produce numerous cytokines (e.g. IL-1, IL-6, TNFα), enzymes, and other mediators.

Activation of lymphocytes leads to two partially competing processes:

- cell proliferation; and
- cell differentiation into effector cells.

Expression of the high-affinity IL-2 receptor on T cells

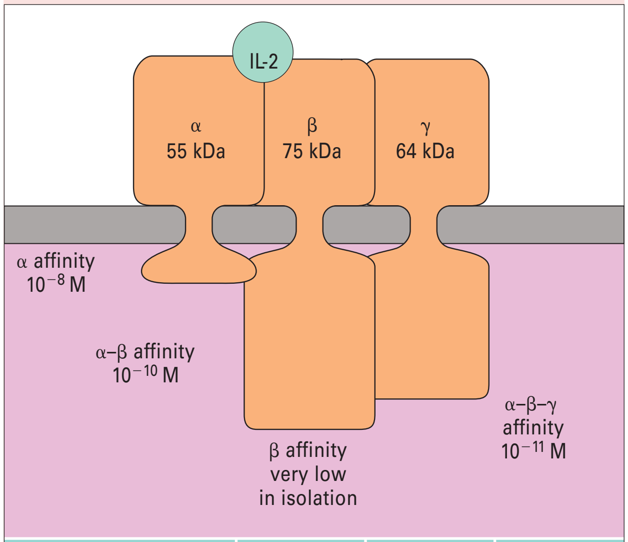

approximate numbers per cell	α chain	β chain	γ chain
at rest	0	500	5000
maximum activation	50 000	5000	~5000

Fig. 7.23 The high-affinity IL-2R consists of three polypeptide chains, shown schematically. Resting T cells do not express the α chain, but after activation they may express up to 50 000 α chains per cell. Some of these associate with the β chain to form the high-affinity IL-2R.

Cells at the end stage of differentiation, such as plasma cells, may become so specialized that they lose surface molecules such as MHC class II, and are unable to respond to regulatory signals or to proliferate.

The fate of lymphocytes responding to antigen is varied:

- some can persist for a long time as memory cells – the life span of memory cells can be more than 40 years in humans, as judged by the chromosome abnormalities (e.g. cross-linking of DNA which would prevent mitosis) found in the blood cells of Hiroshima survivors;
- other lymphocytes have a short life span, which explains why moderate antigenic stimulation does not lead to lymphoid enlargement – this is nevertheless sufficient for generating effective cell-mediated and antibody responses.

Apoptosis is critically important for disposing of unwanted cells after an immune response.

The immune system responds to clues that an infection has taken place before responding strongly to antigens

In recent years it has become appreciated that APCs must respond appropriately to an infection, for example, but not to high levels of harmless substances that may fluctuate in the environment.

APC activation is generally a response to infection, or at least the presence of substances, such as constituents of bacterial cell walls, characteristic of infection. This requirement neatly explains the need for **adjuvants**, which are typically derived from bacterial components.

Adjuvants are generally necessary in vaccines to stimulate a robust immune response. The concept of immune activation only in response to infection (or adjuvant as a surrogate for infection), and not to other antigens, has been popularized as the **'danger' hypothesis**. This idea proposes that the immune system does not merely distinguish self from non-self, but responds to clues that an infection has taken place before responding strongly to antigens.

In other words, foreign substances may be innocuous or invisible to the immune system unless accompanied by danger signals, such as infection. These danger signals are provided by receptors for microbial products on APCs, such as the Toll-like receptors (TLRs, see Fig. 6.24).

FURTHER READING

Ackerman, AL, Cresswell P. Cellular mechanisms governing cross-presentation of exogenous antigens. Nat Immunol 2004;5:678–684.

Alberola IJ, Takaki S, Kerner JD, Perlmutter RM. Differential signaling by lymphocyte antigen receptors. Annu Rev Immunol 1997;15:125–154.

Bell D, Young JW, Banchereau J. Dendritic cells. Annu Rev Immunol 1999;17:255–305.

Boes M, Ploegh HL. Translating cell biology in vitro to immunity in vivo. Nature 2004;430:264–271.

Brocke P, Garbi N, Momburg F, Hammerling GJ. HLA-DM, HLA-DO and tapasin: functional similarities and differences. Curr Opin Immunol 2002;14:22–29.

Clements JL, Boerth NJ, Ran Lee J, Koretzky GA. Integration of T cell receptor-dependent signaling pathways by adapter proteins. Annu Rev Immunol 1999;17:89–108.

Germain RN, Stefanova I. The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation. Annu Rev Immunol 1999;17:467–522.

Grakoui A, Bromley SK, Sumen C, et al. The immunological synapse: a molecular machine controlling T cell activation. Science 1999;285:221–227.

Healy JI, Goodnow CC. Positive versus negative signaling by lymphocyte antigen receptors. Annu Rev Immunol 1998;16:645–670.

Kloetzel PM. Generation of MHC class I antigens: functional interplay between proteasomes and TPPII. Nat Immunol 2004;5:661–669.

Lehner PJ, Cresswell P. Recent developments in MHC class I-mediated antigen presentation. Curr Opin Immunol 2004;16:82–89.

Lehner PJ, Trowsdale J. Antigen processing: coming out gracefully. Curr Biol 1998;8:R605–R608.

Mellman I, Turley SJ, Steinman RM. Antigen processing for amateurs and professionals. Trends Cell Biol 1998;8:231–237.

Nelson CA, Fremont DH. Structural principles of MHC class II antigen presentation. Rev Immunogenet 1999;1:47–59.

Pamer E, Cresswell P. Mechanisms of MHC class I-restricted antigen processing. Annu Rev Immunol 1998;16:323–358.

Parham P. Accessory molecules in the immune response. Immunol Rev 1996;153.

Parham P. Mechanisms of antigen-processing. Immunol Rev 1996;151.

Rodriguez A, Regnault A, Kleijmeer M, et al. Selective transport of internalized antigens to the cytosol for MHC class I presentation in dendritic cells. Nat Cell Biol 1999;1:362–368.

Terhorst C, Spits H, Staal F, Exley M. T lymphocyte signal transduction. In: Hames BD, Glover DM, eds. Molecular Immunology. Oxford: IRL Press; 1996.

Watts C, Powis S. Pathways of antigen processing and presentation. Rev Immunogenet 1999;1:60–74.

Critical thinking: Antigen processing and presentation (see p. 495 for explanations)

Two T cell clones have been produced from a mouse infected with influenza virus. One of the clones reacts to a virus peptide when it is presented on APCs that have the same MHC class I (H-2K) locus as the original mouse (i.e. the clone is MHC class I restricted). The other clone is MHC class II restricted. The two clones are stimulated in tissue culture using syngeneic macrophages as APCs. The macrophages have been either infected with live influenza virus or treated with inactivated virus. The patterns of reactivity of the two clones are shown in the table below. In the last two lines of the table the macrophages are pretreated with either emetine or chloroquine before they are infected with virus. Emetine is a protein synthesis inhibitor. Chloroquine inhibits the fusion of lysosomes with phagosomes.

antigen	**APCs treated with**	**reactivity of clone**	
		clone 1	**clone 2**
none	–	–	–
live virus	–	+	+
inactivated virus	–	–	+
live virus	emetine	–	+
live virus	chloroquine	+	–

1 Why are macrophages used as APCs in this experiment? Would you get the same results if you used infected fibroblasts?

2 Why does the live influenza virus stimulate both clones whereas the inactivated virus stimulates only the MHC class II-restricted clone?

3 Why does emetine prevent the macrophages from presenting antigen to the MHC class I-restricted T cells whereas chloroquine prevents them from presenting to MHC class II-restricted cells?

4 One of these clones expresses CD4 and the other CD8. Which way round is it?

CHAPTER

8

Cell Cooperation in the Antibody Response

SUMMARY

- **The primary development of B cells is antigen-independent.** Pre-B cells recombine genes for immunoglobulin heavy and light chains to generate their surface receptor for antigen.
- **T-independent (TI) antigens activate B cells without requiring T cell help.** They can be divided into two groups. TI-1 antigens can act as polyclonal stimulators, while TI-2 antigens are polymers that activate by cross-linking the B cell receptor.
- **T-dependent (TD) antigens are taken up by B cells, processed, and presented to helper T (TH) cells.** T cells and B cells usually recognize different parts of an antigen.
- **B and T cell activation follow similar patterns.** B cell activation requires signals from the B cell receptor (BCR) and co-stimulation. CD40 is the most important co-stimulatory molecule on B cells. Ligation of B cell co-receptor complex can lower the threshold of antigen needed to trigger the B cell. Intracellular signaling pathways are analogous in B and T cells.
- **Cytokine secretion from $CD4^+$ T cells is important in B cell proliferation and differentiation.** Activated B cells proliferate and differentiate into antibody-forming cells (AFCs).
- **B cell affinity maturation takes place in the germinal centers.** Mutation of immunoglobulin genes followed by selection of high-affinity clones is the basis of affinity maturation.
- **B cells switch to another immunoglobulin class (class switching) by recombining heavy chain genes.** Differential splicing of long RNA transcripts is a second mechanism by which B cells can produce more than one type of antibody.

THE PRIMARY DEVELOPMENT OF B CELLS IS ANTIGEN INDEPENDENT

Over a lifetime, an individual may encounter a vast array of unique and different pathogens. To protect the individual, the immune system has evolved a complex system to be able to respond to any conceivable antigen that it encounters. One of the tactics our immune system uses to fight off these foreign antigens is the production of protective proteins, or antibodies.

Antibodies are produced by B cells, and are exceptional in their almost unlimited potential for diversity.

The development of an antibody response is the culmination of a series of cellular and molecular interactions occurring in an orderly sequence between a B cell and a variety of other cells of the immune system.

In this chapter, the principles of B cell development, activation, proliferation, and differentiation leading to the generation of antibody-producing plasma and memory cells are discussed in the context of immune cell cooperation. In addition, some of the critical molecules involved are described and the consequences of the interactions, including affinity maturation and class switching, are examined.

Within the bone marrow, a sequence of immunoglobulin rearrangements and phenotypic changes takes place during B cell ontogeny that leads to the production of the B cell receptor (BCR). This process is analogous to the generation of the T cell receptor (TCR) in thymic T cells (see Chapter 5).

The BCR complex includes the:

- membrane-bound immunoglobulin (mIg);
- Igα; and
- Igβ.

The molecular processes involved in immunoglobulin gene rearrangement are described in Chapter 3, and this section relates these events to B cell development.

The earliest stage of antigen-independent B cell development is the progenitor B (pro-B) cell (Fig. 8.1).

Pro-B cells are divided into early and late pro-B cell groups based on the expression of:

- terminal deoxynucleotidyltransferase (TdT), which is an intranuclear enzyme uniquely expressed during VH gene rearrangement (see Chapter 3); and
- a marker called B220, which identifies a particular variant of leukocyte common antigen (CD45R), a tyrosine phosphatase that appears to be important in regulating BCR signaling.

Early pro-B cells express TdT alone whereas late pro-B cells express B220 and have reduced TdT (intermediate pro-B cells express both markers).

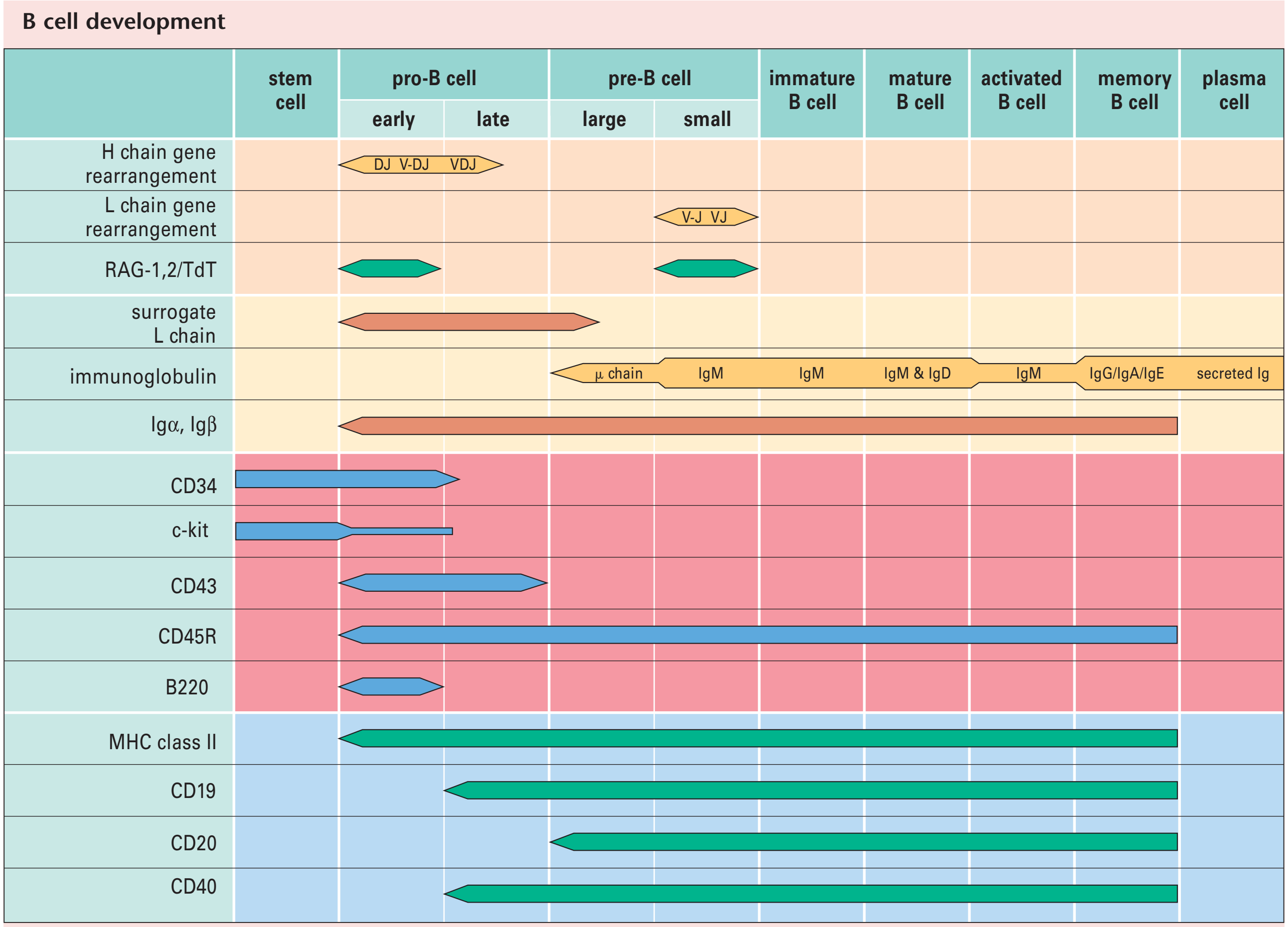
B cell development
stem cell
pro-B cell
early
late
pre-B cell
large
small
immature B cell
mature B cell
activated B cell
memory B cell
plasma cell
H chain gene rearrangement
DJ V-DJ VDJ
L chain gene rearrangement
V-J VJ
RAG-1,2/TdT
surrogate L chain
immunoglobulin
μ chain
IgM
IgM
IgM & IgD
IgM
IgG/IgA/IgE
secreted Ig
Igα, Igβ
CD34
c-kit
CD43
CD45R
B220
MHC class II
CD19
CD20
CD40

Fig. 8.1 B cells differentiate from lymphoid stem cells into naive B cells and may then be driven by antigen to become memory cells or plasma cells. Upon antigen stimulation, the B cell proliferates and develops into a plasma cell or a memory cell after a phase of proliferation, activation, and blast transformation. During early B cell development, the heavy chain gene (IgH) undergoes a DJ recombination, followed by a VDJ recombination. The recombined heavy chain is initially expressed with a surrogate light chain and the Igα and Igβ chains (CD79) to produce the pre-B cell receptor. Later, light chain genes undergo a VJ recombination and surface IgM is produced as the B cell receptor (BCR). Pre-B cells express cytoplasmic μ chains only. The immature B cell has surface IgM, and the mature B cell other immunoglobulin isotypes. The diagram also shows the time of expression of a number of B cell markers.

B220 remains expressed on the surface throughout the remainder of B cell ontogeny.

Genes for immunoglobulin H and L chains are recombined to generate the surface receptor for antigen

Cells progressing through the pro-B cell stage rearrange their immunoglobulin heavy chain genes and express the **recombination-activating genes *RAG-1* and *RAG-2*** (see Chapter 3). As late pro-B cells pass into the pre-B cell stage, they downregulate TdT, RAG-1, and RAG-2 (see Fig. 8.1).

Pre-B cells can be divided into:

- mitotically active pre-B cells; and
- small non-dividing pre-B cells.

Both large and small pre-B cells express immunoglobulin μ heavy chains (see Chapter 3) in the cytoplasm (cμ) and the pre-B cell receptor complex on their surface (Fig. 8.2).

Large pre-B cells have successfully rearranged their immunoglobulin heavy chain genes. As these cells pass from the large pre-B cell group into the small pre-B cell group, they begin to rearrange their immunoglobulin light chain genes and express *RAG-1* and *RAG-2* again.

The final stage of B cell development is the immature B cell stage. Immature B cells have successfully rearranged their light chain genes and express surface IgM as their antigen receptor, at which time *RAG-1* and *RAG-2* gene expression is again turned off.

Q. What is the significance of the observation that RAG genes are expressed at two distinct times during B cell development?

A. The heavy and the light chain genes undergo rearrangements at different times and the RAGs are expressed at the times when recombination is taking place in either the heavy or the light chain gene loci.

Rapid Reference Box 8

activation-induced cytidine deaminase (AID) – an enzyme expressed in activated B cells that causes somatic mutation.
affinity maturation – the increase in average antibody affinity frequently seen during a secondary immune response.
B cell coreceptor complex – a complex of three cell surface molecules, including CD19, CD81, and the complement receptor CD21, which enhances the sensitivity of the B cell to activation by low quantities of antigen.
CD40 – a critical co-stimulatory receptor on B cells that is ligated by CD40L (= CD154) on helper T cells.
class switching – the process by which an individual B cell can link immunoglobulin heavy chain C genes to its recombined VDJ gene to produce a different class of antibody with the same specificity; this process is also reflected in the overall class switch seen during the maturation of an immune response.
iccosomes – immune complexes in the form of small inclusion bodies found on follicular dendritic cells.
interleukin-4 (IL-4) – previously called B cell growth factor, a key cytokine for B cell division and differentiation.
receptor editing – a process in which the genes for the B cell antigen receptor (antibody) undergo a secondary process of recombination, which allows production of non-self reactive antibodies.
somatic mutation – a process during B cell maturation affecting the antibody gene region, that permits refinement of antibody specificity.
T-dependent and T-independent antigens – describe the requirements of B cells in producing secreted antibody in response to antigenic stimulation; T-independent antigens can induce antibody production by B cells without T cell help.

The pre-B cell receptor

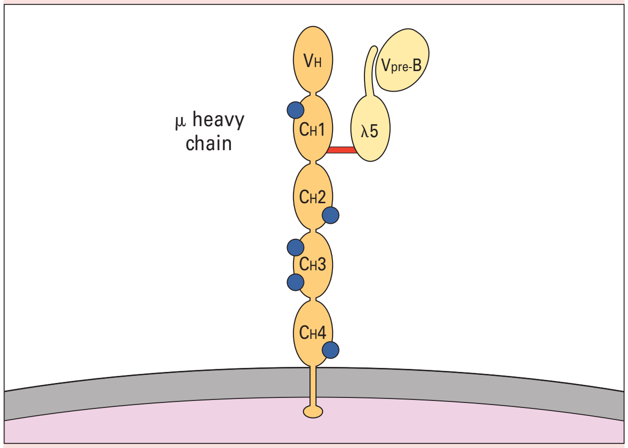

Fig. 8.2 The surrogate BCR complex is composed of a μ heavy chain and $V_{pre\text{-}B}$ and λ5 proteins (surrogate light chains). The receptor has a role in early differentiation and development of B cells. The unit shown here associates with the signaling molecules Igα and Igβ.

As immature B cells develop further into mature B cells, they begin to express both IgM and IgD on their surface. These mature B cells are then free to exit the bone marrow and migrate into the periphery.

Other phenotypic markers can also help to identify particular populations of pro-B, pre-B, or immature B cells (see Fig. 8.1).

In addition, a number of growth and differentiation factors are required to drive the B cells through the early stages of development. Receptors for these factors are expressed at various stages of B cell differentiation:

- IL-4, IL-3, and low molecular weight B cell growth factor (L-BCGF) are important in initiating the process of B cell differentiation;
- other factors are active in the later stages.

TI ANTIGENS DO NOT REQUIRE T CELL HELP TO ACTIVATE B CELLS

Naive mature B cells are free to exit the bone marrow and migrate into the periphery. If these cells do not encounter antigen, they die within a few weeks by apoptosis. If, however, these mature B cells encounter specific antigen, they undergo activation, proliferation, and differentiation, leading to the generation of plasma cells and memory B cells.

The immune response to most antigens depends on both T cells and B cells recognizing the antigen in a linked fashion. This type of antigen is called a **T-dependent (TD) antigen**.

A small number of antigens, however, can activate B cells without MHC class II-restricted T cell help and are referred to as **T-independent (TI) antigens** (Fig. 8.3).

T-independent antigens

antigen	polymeric	polyclonal activation	resistance to degradation
lipopolysaccharide (LPS)	+	+ + +	+
Ficoll	+ + +	–	+ + +
dextran	+ +	+	+ +
levan	+ +	+	+ +
poly-D amino acids	+ + +	–	+ + +
polymeric bacterial flagellin	+ +	+ +	+

Fig. 8.3 The major common properties of some of the main TI antigens are listed. TI antigens induce the production of cytokines interleukin-1 (IL-1), tumor necrosis factor-α (TNFα), and IL-6 by macrophages. (Note that both poly-L amino acids and monomeric bacterial flagellin are TD antigens, demonstrating the role of antigen structure in determining TI properties.)

Importantly, many TI antigens are particularly resistant to degradation. TI antigens can be divided into two groups (TI-1 and TI-2) based on the manner in which they activate B cells:

- TI-1 antigens are predominantly bacterial cell wall components – the prototypical TI-1 antigen is lipopolysaccharide (LPS), a component of the cell wall of Gram-negative bacteria;
- TI-2 antigens are predominantly large polysaccharide molecules with repeating antigenic determinants (e.g. Ficoll, dextran, polymeric bacterial flagellin, and poliomyelitis virus).

Q. What common characteristic can you see in many of the TI antigens, and how are such antigens recognized by the immune system?

A. Many TI antigens have **pathogen-associated molecular patterns (PAMPs)** recognized by **Toll-like receptors** (e.g. TLR4 recognizes LPS and TLR5 recognizes flagellin, see Fig. 6.24). They therefore have an intrinsic ability to activate the immune system irrespective of their ability to bind to the specific antigen receptor on individual B cell clones.

Many TI-1 antigens possess the ability in high concentrations to activate B cell clones that are specific for other antigens – a phenomenon known as polyclonal B cell activation. However, in lower concentrations they only activate B cells specific for themselves. TI-1 antigens do not require a second signal.

TI-2 antigens, on the other hand, are thought to activate B cells by clustering and cross-linking immunoglobulin molecules on the B cell surface, leading to prolonged and persistent signaling. TI-2 antigens require residual non-cognate T cell help, such as cytokines.

Several signal transduction molecules are necessary for mediating TI antigen responses in B cells. These include:

- CD19;
- HS1 protein;
- Lyn;
- IL-5Rα;
- receptors for TNFα and lymphotoxin α.

TI antigens induce poor memory

Primary antibody responses to TI antigens in vitro are generally slightly weaker than those to TD antigens. They peak fractionally earlier and both generate mainly IgM.

The secondary responses to TD and TI antigens differ greatly:

- the secondary response to TI antigens resembles the primary response; whereas
- the secondary response to TD antigens is far stronger and has a large IgG component (Fig. 8.4).

It seems, therefore, that TI antigens do not usually induce the maturation of a response leading to **class switching** or to an increase in antibody affinity, as seen with TD antigens. This is most likely due to a lack of co-stimulation mediated via CD40 (see below) and a lack of key cytokines released by T cells (IL-2, IL-4, IL-5).

Memory induction to TI antigens is also relatively poor.

Comparison of the secondary immune response to T-dependent and T-independent antigens in vitro

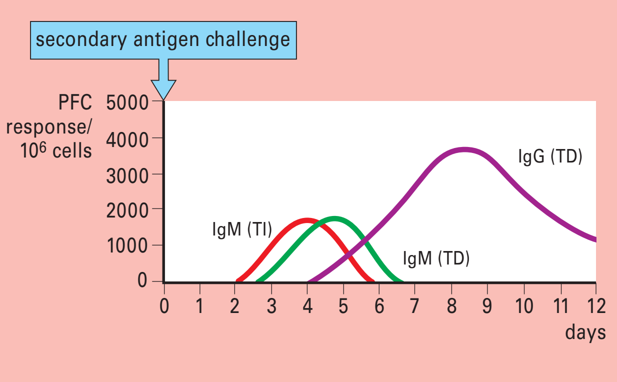

Fig. 8.4 The secondary response to TD antigens is stronger and induces a greater number of IgG-producing cells, as measured in a plaque-forming cell (PFC) assay (see Method Box 8.1).

A number of assays for measuring antibody production by B cells are outlined in Method Box 8.1.

There are potential survival advantages if the immune response to bacteria does not depend on complex cell interactions because it could be more rapid.

Many bacterial antigens bypass T cell help because they are very effective inducers of cytokine production by macrophages – they induce IL-1, IL-6, and TNFα from macrophages. Some such antigens directly stimulate B cells.

TI antigens tend to activate the CD5$^+$ subset of B cells

TI antigens predominantly activate the B-1 subset of B cells found mainly in the peritoneum. These B-1 cells can be identified by their expression of CD5, which is induced upon binding of TI antigens. In contrast to conventional B cells, B-1 cells have the ability to replenish themselves.

TD ANTIGENS REQUIRE T CELL HELP TO ACTIVATE B CELLS

T cells and B cells recognize different parts of antigens

In the late 1960s and early 1970s, studies by Mitchison and others, using chemically modified proteins, led to significant advances in understanding of the different functions of T cells and B cells.

To induce an optimal secondary antibody response to a small chemical group or **hapten** (which is immunogenic only if bound to a protein carrier), it was found that the experimental animal must be immunized and then challenged using the same hapten–carrier conjugate – not just the same hapten. This was referred to as the **carrier effect**.

METHOD BOX 8.1
Measuring antibody production – the PFC assay and ELISPOT

Various methods have been developed for assaying antibody production. Two such methods are the plaque-forming cell (PFC) assay and the enzyme-linked immunospot assay (ELISPOT).

Antibody-forming cells (AFCs) are measured by means of a quantitative PFC assay (Fig. 1), originally developed by Niels Jerne in the 1960s. B cells (e.g. spleen cells) are plated in agar with sheep red blood cells sensitized by binding the specific antigen to their surface. Antibody produced by any B cell will coat the red blood cells. Following the addition of complement, these coated cells may be lyzed, causing the appearance of a zone of lyzed cells (plaque) in the agar. The panel on the right in Fig. 1 shows the appearance of such a plaque, with a B cell in the center, under the microscope. The plaques are then counted to give a quantitative measure of the number of PFCs.

Another way of detecting antibody-producing cells is by means of an enzyme-linked immunospot assay (ELISPOT, Method Box 2.1, Fig. 3). An ELISPOT assay starts out by coating a plastic well with antigen and adding a known quantity of B cells. The antigen coated onto the plastic will then capture any antibody in the vicinity of the activated B cell that is producing the antibody. After a period of time, the B cells are removed, and the specific antibody can be detected by adding an enzyme-labeled anti-immunoglobulin plus chromogen. Development of the label in this assay results in a spot surrounding the active B cell. Counting each spot and knowing the quantity of B cells originally added to the well allows one to enumerate the frequency of B cells producing the specific antibody. Method Box 2.1, Fig. 3 shows the method of detecting antibodies and the appearance of the spots on the developed plates. In addition to analyzing specific antibody-secreting B cells, the ELISPOT assay has been adapted to measure the frequency of cytokine-secreting T cells and various other cell types (right panel). With the improvement in ELISPOT assay plate design and in ELISPOT detection equipment, antibody- or cytokine-secreting cells can now be detected at the single cell level.

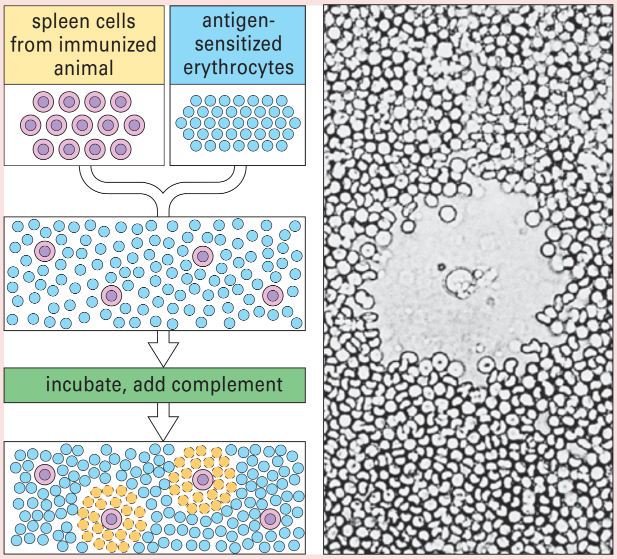
spleen cells from immunized animal
antigen-sensitized erythrocytes
incubate, add complement

Fig. 1 Plaque-forming cell (PFC) assay.

By manipulating the cell populations in these experiments, it was shown that:

- helper T cells (TH) cells are responsible for recognizing the carrier; whereas
- the B cells recognize hapten.

These experiments were later reinforced by details of how:

- B cells use antibody to recognize epitopes; while
- T cells recognize processed antigen fragments.

One consequence of this system is that an individual B cell can receive help from T cells specific for different antigenic peptides provided that the B cell can present those determinants to each T cell.

In an immune response in vivo, it is believed that the interactions between T and B cells that drive B cell division and differentiation involve T cells that have already been stimulated by contact with the antigen on other antigen-presenting cells (APCs), for example dendritic cells.

This has led to the basic scheme for cell interactions in the antibody response set out in Fig. 8.5. It is proposed that antigen entering the body is processed by cells that present the antigen in a highly immunogenic form to the TH and B cells. The T cells recognize determinants on the antigen that are distinct from those recognized by the B cells, which differentiate and divide into antibody-forming cells (AFCs). Therefore two processes are required to activate a B cell:

- antigen interacting with B cell immunoglobulin receptors – this involves 'native' antigen;
- stimulating signal(s) from TH cells that respond to processed antigen bound to MHC class II molecules.

B AND T CELL ACTIVATION FOLLOW SIMILAR PATTERNS

In B cells the signaling function of CD3 in T cells is carried out by a heterodimer of Igα and Igβ. Two molecules of the Igα–Igβ heterodimer associate with surface immunoglobulin to form the BCR. The cytoplasmic tails of Igα and Igβ carry **immunoreceptor tyrosine activation motifs (ITAMs)**.

Cross-linking of surface immunoglobulin leads to activation of the Src family kinases, which in B cells are Fyn, but also Lyn, and Blk. Syk is analogous to ZAP-70 in T cells, and binds to the phosphorylated ITAMs of Igα and Igβ (Fig. 8.6). This leads to activation of a kinase cascade and translocation of nuclear transcription factors

Cell cooperation in the antibody response

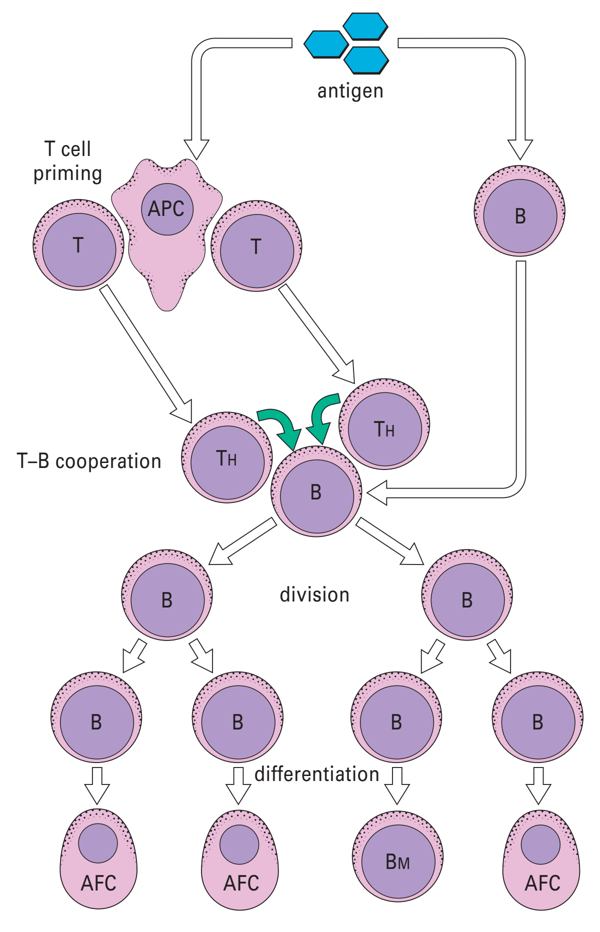

Fig. 8.5 Antigen is presented to naive T cells by APCs such as dendritic cells. B cells also take up antigen and present it to the T cells, receiving signals from the T cells to divide and differentiate into antibody-forming cells (AFCs) and memory B cells (BM).

Intracellular signaling in B cell activation

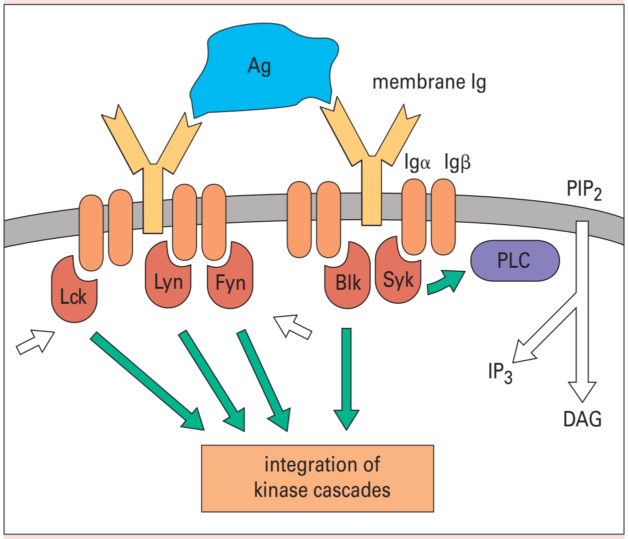

Fig. 8.6 B cell activation is similar to T cell activation. If membrane immunoglobulin becomes cross-linked (e.g. by a TI antigen), tyrosine kinases, including Lck, Lyn, Fyn, and Blk, become activated. They phosphorylate the ITAM domains in the Igα and Igβ chains of the receptor complex. These can then bind another kinase, Syk, which activates phospholipase C. This acts on membrane phosphatidylinositol bisphosphate (PIP_2) to generate inositol trisphosphate (IP_3) and diacyl glycerol (DAG), which activates protein kinase C. Signals from the other kinases are transduced to activate nuclear transcription factors.

analogous to the process that occurs in T cells (see Fig. 7.21).

B cell activation is also markedly augmented by the '**co-receptor complex**' comprising three proteins:

- CD21 (complement receptor-2, CR2);
- CD19; and
- CD81 (target of antiproliferative antibody, TAPA-1) (Fig. 8.7).

Follicular dendritic cells are known to retain antigen on their surface for prolonged periods of time as immune complexes (**iccosomes**). The antigen in such complexes can bind to:

- CD21 (via the complement fragment C3d); and
- surface immunoglobulin on the B cells.

Phosphorylation of the cytoplasmic tail of CD19 can then occur, leading to binding and activation of Lyn. It is likely that these kinases enhance the activation signal through the phospholipase C and phosphatidylinositol 3-kinase pathways, particularly when antigen concentration is low.

Q. If an animal was depleted of complement C3 during the primary immune response, what effect do you think this would have on the development of the secondary antibody response?

A. Lack of C3 means that immune complexes containing C3b and C3d do not form. Therefore they cannot bind to follicular dendritic cells (FDCs) via CR2 or engage with B cells via the B cell co-receptor complex (see Fig. 8.7); consequently B cell activation and the development of the secondary immune response is greatly impaired. (These experiments have been done by depleting mice of C3 using cobra venom factor.)

Direct interaction of B cells and T cells involves co-stimulatory molecules

Antigen-specific T cell populations can be obtained by growing and cloning T cells with antigens, APCs, and IL-2. It is therefore possible to visualize directly B cell and T cell clusters interacting in vitro:

- the T cells become polarized, with the TCRs concentrated on the B cell side;
- the B cells also become polarized and express most of their MHC class II molecules and intercellular adhesion molecule-1 (ICAM-1) in proximity to the T cells.

B cell co-receptor complex

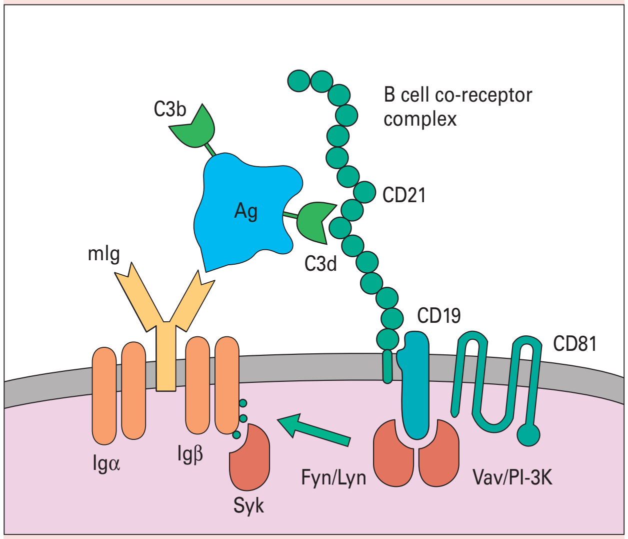

Fig. 8.7 The B cell co-receptor complex consists of CD21 (the complement receptor type 2), CD19, and CD81 (a molecule with four transmembrane segments). Antigen with covalently bound C3b or C3d can cross-link the membrane immunoglobulin to CD21 of the co-receptor complex. This greatly reduces the cell's requirement for antigen to activate it. CD19 can associate with tyrosine kinases including Lyn, Fyn, Vav, and PI-3 kinase (PI-3K). Compare this with CD28 on the T cell (see Fig. 7.21). Receptor cross-linking causes phosphorylation of the Igα and Igβ chains of the antigen–receptor complex and recruitment and activation of Syk.

The interactions in these clusters strongly suggest an intense exchange of information, which leads to two important events in the B cell life cycle:

- induction of proliferation; and
- differentiation into AFCs.

The initial interaction between a naive B cell and a **cognate antigen** via the BCR in the presence of cytokines or other growth stimuli induces activation and proliferation of the B cell. This then leads to processing of the TD antigen and presentation to T cells (see Chapter 7).

The interaction between B and T cells is a two-way process in which:

- B cells present antigen to T cells; and
- receive signals from the T cells for division and differentiation (Fig. 8.8).

The central antigen-specific interaction is that between the MHC class II molecule–antigen complex and the TCR. This interaction is augmented by interactions between pairs of adhesion molecules and co-stimulatory molecules.

Q. Which co-stimulatory molecules, present on B cells and other APCs, promote T cell proliferation? What mechanisms are involved?

A. B7-1 (CD80) and B7-2 (CD86) act on antigen-activated T cells by ligating CD28, leading to expression of the high-affinity IL-2 receptor (see Fig. 7.16).

Cell surface molecules involved in the interaction between B and TH cells

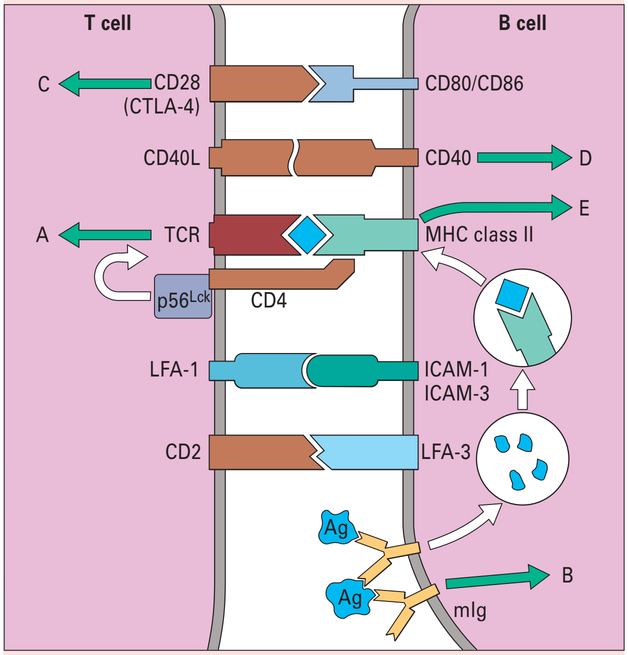

Fig. 8.8 Membrane immunoglobulin (mIg) takes up antigen (Ag) into an intracellular compartment where it is degraded and peptides can combine with MHC class II molecules. Other arrows show the discrete signal transduction events that have been established. A and B are the antigen receptor signal transduction events involving tyrosine phosphorylation and phosphoinositide breakdown. The antigen receptors also regulate LFA-1 affinity for ICAM-1 and ICAM-3, possibly through the signal transduction events. In the T cell, CD28 also sends a unique signal to the T cell (C). In the later phases of the response CTLA-4 can supplant CD28 to cause downregulation (see Fig. 7.18). In the B cell, stimulation via CD40 is the most potent activating signal (D). In addition, MHC class II molecules appear to induce distinct signaling events (E). Not shown is the exchange of soluble interleukins and binding to the corresponding receptors on the other cell. (Adapted from DeFranco A. Nature 1991;351:603–605)

The interaction between B and T cells is a two-way event as follows:

- CD40, a member of the TNF receptor family, delivers a strong activating signal to B cells, more potent even than signals transmitted via surface immunoglobulin;
- upon activation, T cells transiently express a ligand, termed CD40L (a member of the TNF family), which interacts with CD40;
- CD40–CD40L interaction helps to drive B cells into cell cycle;
- transduction of signals through CD40 induces upregulation of CD80/CD86 and therefore helps to provide further co-stimulatory signals to the responding T cells.

Signaling through CD40 is also essential for germinal center development and antibody responses to TD antigens.

Q. Some individuals have a mutation that produces a non-functional variant of CD40L. In its homozygous form, what effect would you expect this mutation to have on serum antibody levels?

A. The mutation produces hyper-IgM syndrome, with high serum IgM and low IgG, IgA, and IgE levels, due to lack of germinal centers and failure of the B cells to switch isotype.

CYTOKINE SECRETION FROM CD4⁺ T CELLS CONTROLS B CELL PROLIFERATION AND DIFFERENTIATION

The cytokine milieu is critical for determining the outcome of an immune response following B cell–T cell interaction. Recent work has shown that CD4⁺ T cells in both mouse and human can be divided into different subsets, depending on their cytokine profile (Fig. 8.9):

- TH1 cells – CD4⁺ T cells that produce IL-2 and interferon-γ (IFNγ), but not IL-4, are designated TH1 and are chiefly responsible for delayed-type hypersensitivity responses, but can also help B cells produce IgG2a (mouse), but not much IgG1 or IgE;
- TH2 cells – CD4⁺ T cells that produce IL-4, IL-5, IL-10, and IL-13, but not IL-2 or IFNγ, are designated TH2 and are very efficient helper cells for production of antibody, especially IgG1 and IgE;
- TH0 cells – many CD4⁺ T cells, especially in humans, have cytokine profiles intermediate between TH1 cells and TH2 cells, and are known as 'TH0' cells. These cells are capable of producing both the TH1 cytokine IFNγ and the TH2 cytokine IL-4. However, classic TH1 and TH2 cells have been well documented in humans, especially in diseased tissues.

During B cell–T cell interaction, T cells can secrete a number of cytokines that have a powerful effect on B cells (see Fig. 8.9). IL-2, for example, is an inducer of proliferation for B cells.

The specific cytokines produced by TH2 cells also affect B cells. These cytokines include IL-4, IL-5, IL-6, IL-10, and IL-13:

- IL-4 (previously known as B cell activating or differentiation factor-1) acts on B cells to induce activation and differentiation, and acts on T cells as a growth factor and promotes differentiation of TH2 cells, thus reinforcing the antibody response – excess IL-4 plays a part in allergic disease, causing production of IgE;
- IL-5 in humans is chiefly a growth and activation factor for eosinophils and is responsible for the eosinophilia of parasitic disease (in the mouse it also acts on B cells to induce growth and differentiation);
- IL-6 is produced by many cells including T cells, macrophages, B cells, fibroblasts, and endothelial cells, and acts on most cells, but is particularly important in inducing B cells to differentiate into AFCs – IL-6 is considered to be an important growth factor for multiple myeloma, a malignancy of plasma cells;
- IL-10 acts as a growth and differentiation factor for B cells in addition to modulating cytokine production by TH1 cells;
- IL-13, which shares a receptor component and signaling pathways with IL-4, acts on B cells to produce IgE.

Other cytokines such as IL-7, originally isolated from a stromal cell line as a factor supporting pre-B cell growth, have been shown to be indispensable for B cell development.

Q. Give three different ways in which IL-4 can reinforce the TH2-type immune response.

A. It promotes differentiation of TH0 cells into TH2 cells. It acts on B cells to promote their division, differentiation, and antibody synthesis. It also acts on endothelium and tissue cells to promote the synthesis of chemokines that selectively attract TH2 cells.

Cytokines and B cell development

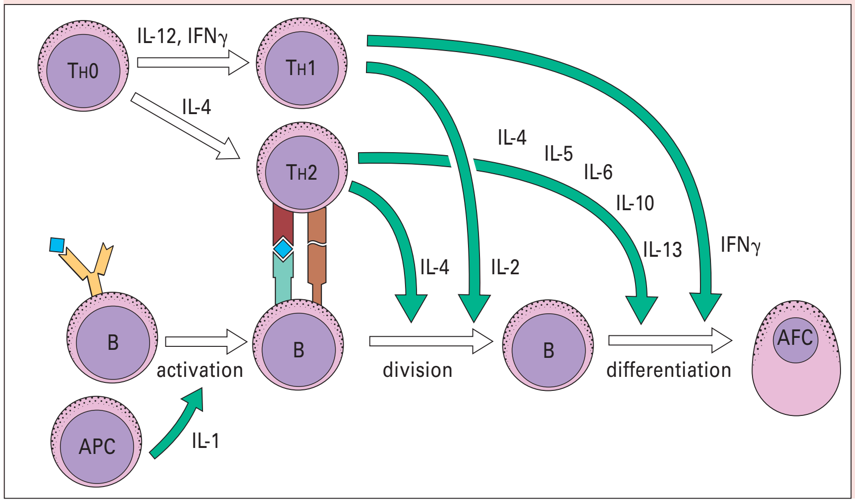

Fig. 8.9 B cell development is influenced by cytokines from T cells and APCs, and by direct interactions with TH2 cells. IL-4 is most important in promoting division, and and a variety of cytokines including IL-4, IL-5, IL-6, IL-10, and IFNγ influence development into antibody-forming cells (AFCs), and affect the isotype of antibody that will be produced.

Cytokines can also influence antibody affinity. Antibody affinity to most TD antigens increases during an immune response, and a similar effect can be produced by certain immunization protocols. For example, high-affinity antibody subpopulations are potentiated after immunization with antigen and IFNγ (Fig. 8.10).

A number of adjuvants are capable of enhancing levels of antibody, but few have this characteristic of also potentiating affinity. As affinity markedly influences the biological effectiveness of antibodies, IFNγ may be an important adjuvant for use in vaccines.

Q. Why do adjuvants not generally enhance antibody affinity, though they do enhance antibody titers? – think of when and how adjuvants act.

A. Most adjuvants act as depots for antigen or have components that promote antigen presentation to T cells. They do not act directly on B cells.

In addition to the effects of cytokines on B cell proliferation and differentiation, cytokines are capable of influencing the class switch from IgM to other immunoglobulin classes (see below).

Cytokine influence on antibody affinity

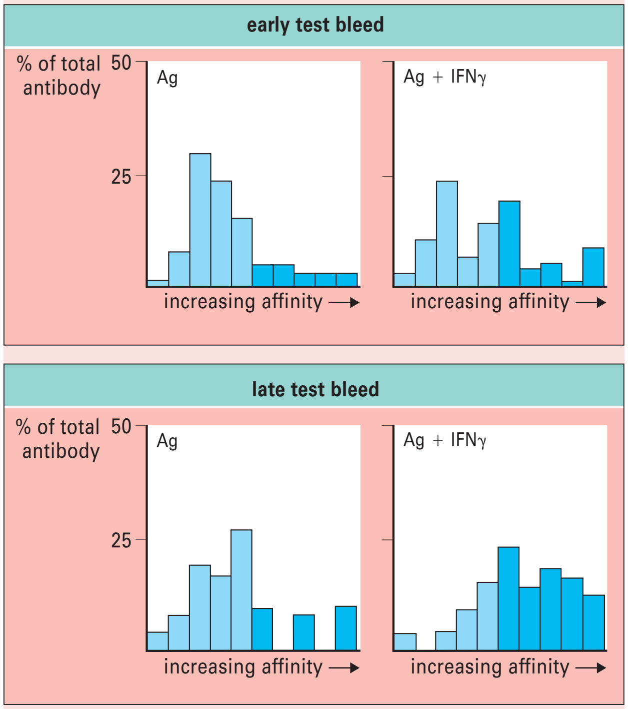

Fig. 8.10 Mice were immunized either with antigen alone (Ag) or with antigen plus 30 000 units of IFNγ (Ag + IFNγ). The affinity of the antibodies was measured either early or late after immunization. Mice receiving IFNγ showed more high-affinity antibody (darker bars) in both the early and late bleeds than mice that received antigen alone. (Adapted from Holland GP, Holland N, Steward MW. Clin Exp Immunol 1990;82:221–226)

Cytokine receptors aid in B cell growth and differentiation

Receptors for the many growth and differentiation factors required to drive the B cells through their early stages of development are expressed at various stages of B cell differentiation. Receptors for IL-7, IL-3, and low molecular weight B cell growth factor are important in the initial stages of B cell differentiation whereas other receptors are more important in the later stages (Fig. 8.11).

B cell–T cell interaction may either activate or inactivate (anergize)

The above description of B cell–T-cell interaction suggests that the only possible outcome is activation of the B cell. However, this is not the case.

APC–T cell interaction may yield two diametrically opposing results, namely activation or inactivation (**clonal anergy**, see Fig. 7.18).

In the same way, B cells frequently become anergic. This is an important process because affinity maturation of B cells (see below) during the immune response as a result of rapid mutation in the genes encoding the antibody variable regions could easily result in high-affinity autoantibodies.

Clonal anergy and other forms of tolerance in the periphery are important for silencing these potentially damaging clones. However, the molecular details of this process are still unclear. Moreover, the respective roles of IgM and IgD, the two cell surface receptors for antigen on B cells, are also not understood in terms of activation or inactivation. Both IgM and IgD appear to be capable of transmitting signals for both functions.

Following activation, antigen-specific B cells can follow either of two separate developmental pathways:

- the first pathway involves proliferation and differentiation into AFCs in the lymph nodes or in the periarteriolar lymphoid sheath of the spleen – these AFCs function to rapidly clear antigen, but the great majority of these cells die via apoptosis within 2 weeks, so it is unlikely that these AFCs are responsible for long-term antibody production;
- in the second pathway some members of the expanded B cell population migrate into adjacent follicles to form germinal centers before differentiating into memory B cells.

The mechanism that determines which path a B cell takes is unknown. However, it is likely that the decision can be influenced by the nature of the naive B cells initially recruited into the response, the affinity and specificity of the BCR, the type of antigen driving the response, and the levels of T cell help.

Q. Which type of APC is specifically located in the germinal center?

A. FDCs (see Fig. 7.2), and these turn out to be very important in driving B cell development.

Cytokine receptor expression during B cell development

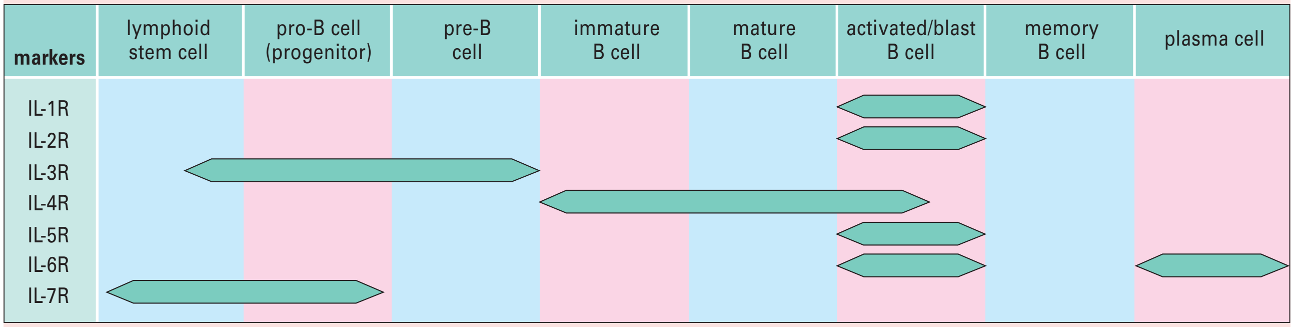

Fig. 8.11 The whole life history of B cells from stem cell to mature plasma cell is regulated by cytokines present in their environment. Receptors for these cytokines are selectively expressed by B cells at different stages of development. Many of these receptors now have CD nomenclature (see Appendix 2).

B CELL AFFINITY MATURATION TAKES PLACE IN THE GERMINAL CENTERS

The germinal center is important in that it provides a microenvironment whereby B cells can undergo developmental events that ultimately result in an affinity-matured, long-lived memory B cell compartment of the immune system (Fig. 8.12). These developmental events come about due to complex interactions between B cells, $CD4^+$ helper T cells, and FDCs. These events include:

- clonal proliferation;
- antibody variable region somatic hypermutation;
- receptor editing;
- isotype switch recombination;
- affinity maturation; and
- positive selection.

The germinal center initially contains only dividing centroblasts. Shortly thereafter, the germinal center polarizes itself into a dark zone containing centroblasts and a light zone containing non-dividing (resting) centrocytes (Figs 8.13 and 8.14). Centroblasts proliferate rapidly in the dark zone and downregulate the expression of their surface immunoglobulin. **Somatic hypermutation** then occurs to diversify the rearranged variable region genes (see Chapter 3, p. 83). Somatic hypermutation allows a single B cell to give rise to a clone that contains variants with different affinities for the antigen.

B cell development in germinal centers

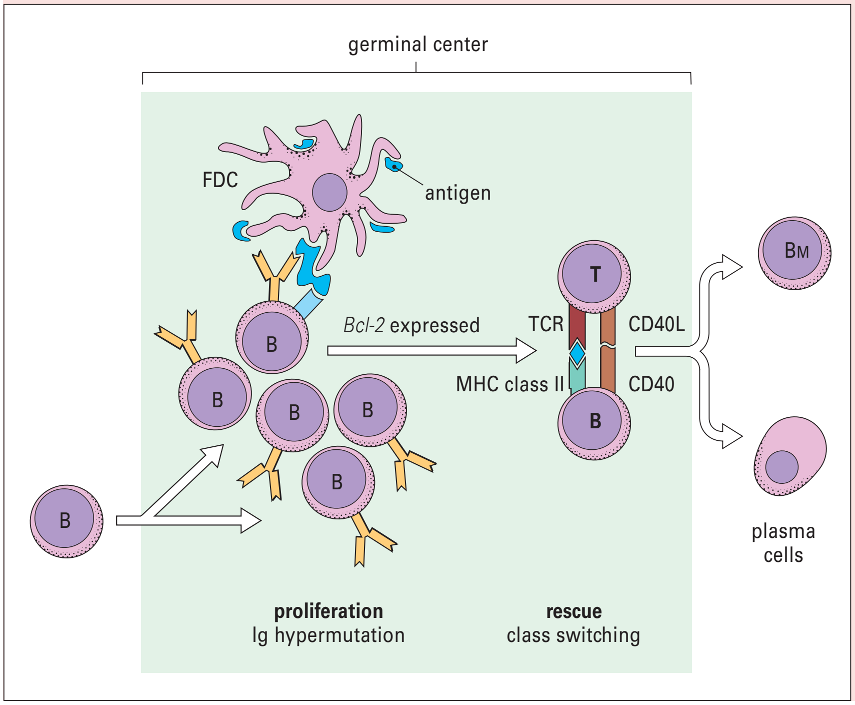

Fig. 8.12 A B cell enters a germinal center and undergoes rapid proliferation and hypermutation of its immunoglobulin genes. Antigen is presented by the follicular dendritic cell (FDC), but only B cells with high-affinity receptors will compete effectively for this antigen. B cells that do have a higher-affinity immunoglobulin express *Bcl-2* and are rescued from apoptosis by interaction with T cells (i.e. the B cell presenting antigen to the T cell). Interaction with T cells promotes class switching. The class switch that takes place depends on the T cells present, which partly relates to the particular secondary lymphoid tissue and the type of current immune response (TH1 versus TH2). B cells leave the germinal center to become either plasma cells or B memory cells (BM).

Schematic organization of the germinal center

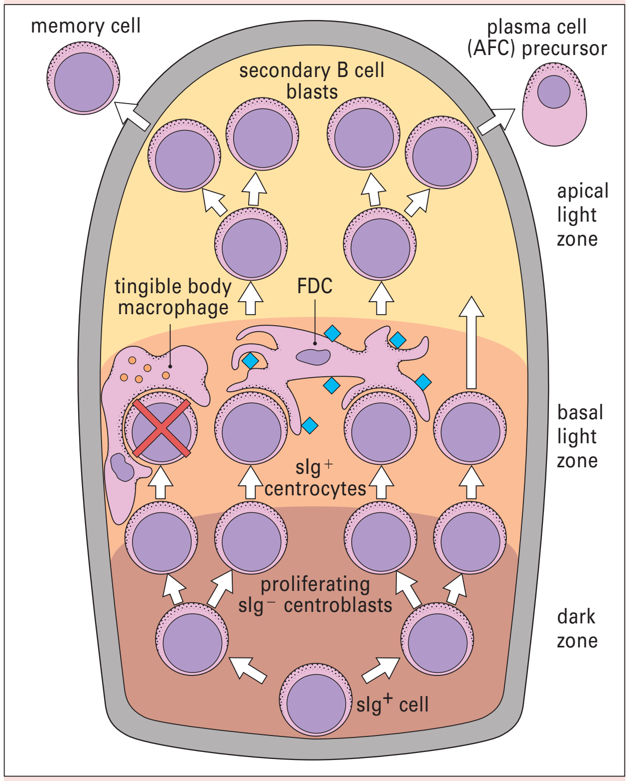

Fig. 8.13 The functions of the germinal center are clonal proliferation, somatic hypermutation of immunoglobulin receptors, receptor editing, isotype class switching, affinity maturation, and selection by antigen. In this model, the germinal center is composed of three major zones – a dark zone, a basal light zone, and an apical light zone. These zones are predominantly occupied by centroblasts, centrocytes, and secondary blasts, respectively. Primary B cell blasts carrying surface immunoglobulin receptors (sIg$^+$) enter the follicle and leave as memory B cells or AFCs. Antigen-presenting follicular dendritic cells (FDCs) are mainly found in the two deeper zones, and cell death by apoptosis occurs primarily in the basal light zone where tingible body macrophages are also located. Blue squares are iccosomes on the FDC. (Adapted from Roitt IM. Essential Immunology, 7th edn. Oxford: Blackwell Scientific, 1991. Copyright 1991 with permission from Blackwell Publishing)

Isotype switch recombination (see below) occurs following somatic hypermutation and requires cell cycling. Receptor editing of immunoglobulin light chain genes also occurs in centroblasts.

Following these developmental changes, the centroblasts migrate to the FDC light zone of the germinal center and give rise to centrocytes, which then re-express surface immunoglobulin BCR. In the light zone, centrocytes encounter antigen bound to the FDCs and antigen-specific TH2 cells. FDCs and T cells interact with centrocytes through:

- surface molecules such as the BCR, CD40, CD80 (B7-1), CD86 (B7-2), LFA-1, VLA-4, CD54 (ICAM-1), and CD106 (VCAM-1); and

Zoning of the germinal center of a lymph node

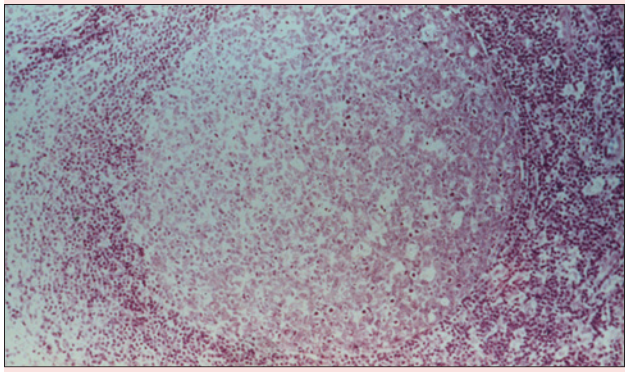

Fig. 8.14 This Giemsa-stained section (× 40) shows the light part (left) and the more actively proliferating dark part (right). There is a well-developed mantle of small resting lymphocytes, which have less cytoplasm and therefore appear more densely packed. (Courtesy of Dr K McLennan)

- cytokines such as IL-2, IL-4, IL-5, IL-6, IL-10, IL-13, and lymphotoxin-α.

After the centrocytes have stopped dividing, they are selected according to their ability to bind antigen. Those with high-affinity receptors for foreign antigen are positively selected whereas those without adequate affinity undergo Fas-dependent apoptosis (see Chapter 10).

Self-reactive B cells generated by somatic mutation are deleted

Centrocytes that respond to soluble antigen or do not receive T cell help are negatively selected and undergo **Fas-independent apoptosis**. In this way, selection provides a mechanism for elimination of self-reactive antibodies that may be generated during somatic hypermutation.

Positively selected centrocytes can re-enter the dark zone for successive rounds of expansion, diversification, and selection.

Somatic hypermutation and selection improve the average affinity of the germinal center B cell population for presented antigen.

Following these B cell developmental stages, the centrocytes exit the germinal center and lose their susceptibility to apoptosis by:

- reducing Fas; and
- increasing the expression of Bcl-2.

Three possible outcomes are associated with exit from the germinal center:

- antibody-secreting bone marrow homing effector B cells;
- marginal zone memory B cells;
- recirculating memory B cells.

The factors that regulate the decision to exit the germinal center are poorly understood.

In-vivo antibody responses show isotype switching, affinity maturation, and memory

The earliest studies on antibody responses followed the development of specific antibodies in animals immunized with TD or TI antigens. With our improved knowledge of B cell development and maturation, it is now possible to understand the features of immune responses in vivo in terms of the underlying cellular events. Features of antibody responses in vivo include:

- the enhanced secondary response;
- isotype switching;
- affinity maturation;
- the development of memory.

However, some of these events can be understood only by viewing the B cell population as a whole, rather than as a collection of individual cells. The elements of the antibody response in vivo are detailed as follows.

Following primary antigenic challenge, there is an initial lag phase when no antibody can be detected. This is followed by phases in which the antibody titer increases logarithmically to a plateau and then declines. The decline occurs because the antibodies are either:

- naturally catabolized; or
- bind to the antigen and are cleared from the circulation (Fig. 8.15).

An examination of the responses following primary and secondary antigenic challenge shows that the responses differ in the following four major respects (Fig. 8.16):

- time course – the secondary response has a shorter lag phase and an extended plateau and decline;
- antibody titer – the plateau levels of antibody are much greater in the secondary response, typically 10-fold or more than plateau levels in the primary response;
- antibody class – IgM antibodies form a major proportion of the primary response, whereas the secondary response consists almost entirely of IgG, with very little IgM;
- antibody affinity – the affinity of the antibodies in the secondary response is usually much higher; this is referred to as ‘affinity maturation’.

The functions of the different antibody classes are detailed in Chapter 3.

The four phases of a primary antibody response

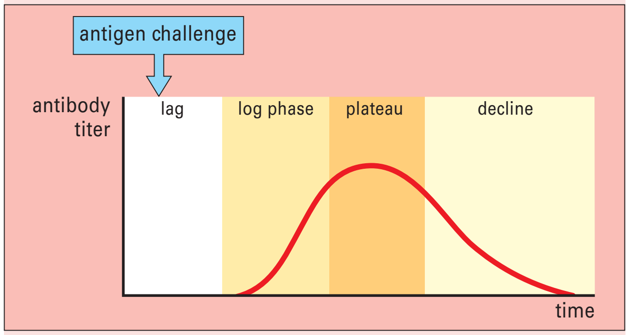

Fig. 8.15 After antigen challenge, the antibody response proceeds in four phases – a lag phase when no antibody is detected; a log phase when the antibody titer increases logarithmically; a plateau phase during which the antibody titer stabilizes; and a decline phase during which the antibody is cleared or catabolized. In a primary immune response, IgM initially predominates, followed by IgG. The actual time course and titers reached depend on the nature of the antigenic challenge and the nature of the host.

Primary and secondary antibody responses

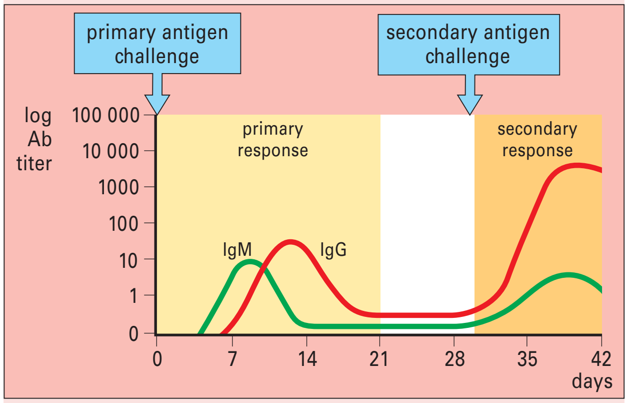

Fig. 8.16 In comparison with the antibody response after primary antigenic challenge, the antibody level after secondary antigenic challenge in a typical immune response appears more quickly, persists for a longer period of time, attains a higher titer, and consists predominantly of IgG antibodies.

Affinity maturation of B cells depends on cell selection

The antibodies produced in a primary response to a TD antigen generally have a low average affinity. However, during the course of the response, the average affinity of the antibodies increases or ‘matures’. As antigen becomes limiting, the clones with the higher affinity have a selective advantage. This process is called **affinity maturation**.

The degree of affinity maturation is inversely related to the dose of antigen administered. High antigen doses produce poor maturation compared with low antigen doses (Fig. 8.17). It is thought that:

- in the presence of low antigen concentrations only B cells with high-affinity receptors bind sufficient antigen and are triggered to divide and differentiate;
- in the presence of high antigen concentrations, there is sufficient antigen to bind and trigger both high- and low-affinity B cells.

Although individual B cells do not usually change their overall specificity, the affinity of the antibody produced by a clone may be altered.

Affinity maturation is achieved through two processes

Affinity maturation is achieved through two processes:

- somatic mutation of the variable (V), diversity (D), and joining (J) gene segments encoding the variable regions of antibodies (see Chapter 3);

Affinity maturation

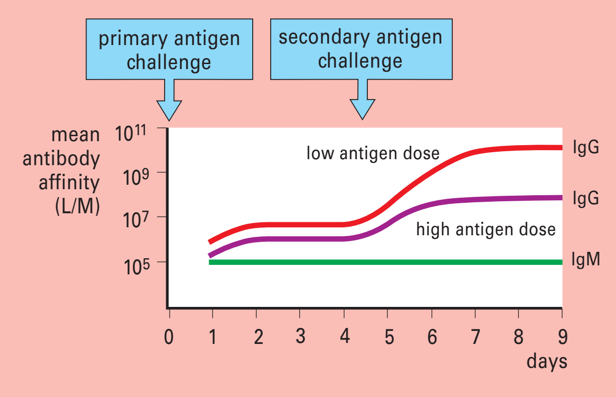

Fig. 8.17 The average affinity of the IgM and IgG antibody responses after primary and secondary challenge with a TD antigen is shown. The affinity of the IgM response is constant throughout. The affinity maturation of the IgG response depends on the dose of the secondary antigen. Low antigen doses produce higher-affinity immunoglobulin than high antigen doses because the high-affinity clones compete effectively for the limiting amount of antigen.

- antigen-driven selection and expansion of mutant clones expressing higher-affinity antibodies.

The mechanism by which affinity maturation occurs is thought to involve B cell progeny binding to antigen on FDCs in order to proliferate and differentiate further. It is thought that unprocessed antigen in immune complexes is captured by FDCs via their Fc and complement receptors (see Fig. 2.18) and held there. As B cells encounter the antigen, there is competition for space on the surface of the FDC, leading to selection. When a cell with higher affinity arises, it will stay there longer and presumably be given a stronger signal. B cells with higher affinity will therefore have a selective advantage.

An alternative theory is that B cells with higher-affinity receptors compete more effectively to bind and internalize antigen and therefore have a greater potential of presenting antigen to T cells, and receiving T cell help.

Somatic hypermutation and receptor editing lead to changes in the BCR

Somatic hypermutation is a common event in AFCs during TD responses and is important in the generation of high-affinity antibodies.

Somatic hypermutation introduces point mutations at a very high rate into the variable regions of the rearranged heavy and light chain genes (see Fig. 3.32). This results in mutated immunoglobulin molecules on the surface of the B cell. Mutants that bind antigen with higher affinity than the original surface immunoglobulin provide the raw material for the selection processes mentioned above.

The mechanism by which somatic hypermutation occurs involves an enzyme called **activation-induced cytidine deaminase (AID)**, which converts cytidine bases in DNA into uracil. The enzyme is specifically expressed in active B cells and is induced by IL-4 and ligation of CD40. Animals lacking this enzyme have deficient somatic hypermutation and class switch recombination (see below).

The abnormal uracil bases produced in the DNA are processed by a number of DNA repair systems, which fix the mutation in the immunoglobulin gene sequence.

AID appears to be targeted to the actively transcribing DNA genes by association with RNA polymerase, though it acts preferentially on the non-transcribing strand. The mechanism that preferentially targets AID to immunoglobulin genes is still uncertain, but it may involve secondary structures in the displaced (non-coding) DNA, or additional proteins expressed in B cells that are associated with the immunoglobulin V region gene.

Receptor editing is another mechanism by which diversity can be introduced into B cells during affinity maturation. Secondary V(D)J recombination can occur in immature B cells whose antigen receptors bind self antigen. The resulting immunoglobulin rearrangement converts these cells into non-self-reactive cells. In this way, specificity for foreign antigens can be improved and self-reactivity avoided.

B CELLS SWITCH IMMUNOGLOBULIN CLASS BY RECOMBINING HEAVY CHAIN GENES

B cells produce antibodies of five major classes – IgM, IgD, IgG, IgA, and IgE. In humans, there are also four subclasses of IgG and two of IgA (see Chapter 3). Each terminally differentiated plasma cell is derived from a specific B cell and produces antibodies of just one class or subclass.

All classes of immunoglobulin use the same set of variable region genes.

The first B cells to appear during development carry surface IgM as their antigen receptor. Upon activation, other classes of immunoglobulin are seen, each associated with different effector functions.

When a mature AFC switches antibody class, all that changes is the constant region of the heavy chain. The expressed V(D)J region and light chain do not change. Antigen specificity is therefore retained.

This has been shown by the analysis of double myelomas, in which two monoclonal antibodies are present in the serum at the same time:

- IgM and IgG antibodies from a patient with multiple myeloma have been found to have identical light chains and VH regions – only the constant regions are switched, from μ to γ;
- similarly, IgM and IgD are often found on the surface of a lymphocyte at the same time.

Again, although the classes are different, the antigen-binding specificities are identical.

The constant region genes encoding the different heavy chains (CH) are responsible for the generation of antibody classes and subclasses. These genes are clustered at the 3′ end of the immunoglobulin heavy chain (IGH) locus, downstream of the J segment genes.

Constant region genes of mouse

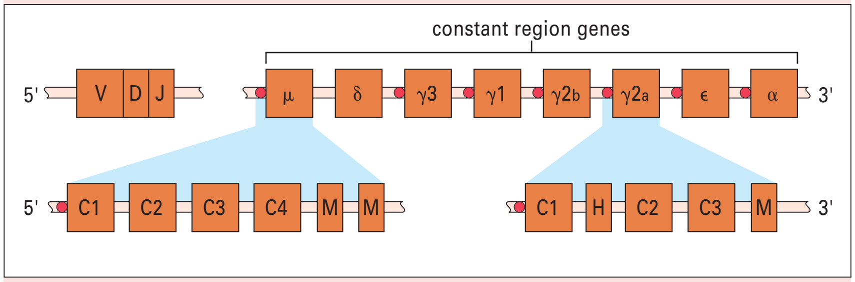

Fig. 8.18 The constant region genes of the mouse are arranged 8.5 kb downstream from the recombined VDJ segment. Each C gene (except Cδ) has one or more switching sequences at its start (red circles), which correspond to a sequence at the 5′ end of the μ gene. This allows any of the C genes to be expressed together with the VDJ segment. δ genes appear to use the same switching sequences as μ, but the μ gene transcript is lost in RNA processing to produce IgD. The C genes (expanded below for μ and γ2a) contain introns separating the exons for each domain (C1, C2, etc.). The γ genes also have a separate exon coding for the hinge (H) and all the genes have one or more exons coding for membrane-bound immunoglobulin (M).

The arrangement of the constant genes in mouse and humans is shown in Figs 8.18 and 8.19. Upstream of the μ genes is a switch sequence (S), which is repeated upstream to each of the other constant region genes except δ. These sequences are important in the recombination events that occur during class switching, as explained below.

Class switching occurs during maturation and proliferation

Most class switching occurs during proliferation. However, it can also take place before encounter with exogenous antigen during early clonal expansion and maturation of the B cells (Fig. 8.20). This is known because some of the progeny of immature B cells synthesize antibodies of other immunoglobulin classes, including IgG and IgA.

Further B cell differentiation results in synthesis of surface IgD, an antibody class found almost exclusively on B cell membranes.

Different classes of surface immunoglobulin on the same B cell have the same antigen specificity, that is, they express the same V region genes, though later additional diversity within a single clone may be generated by somatic mutation after class switching.

Evidence that some class switching occurs independently of antigen comes from experiments with vertebrates

Constant region genes and class switching in humans

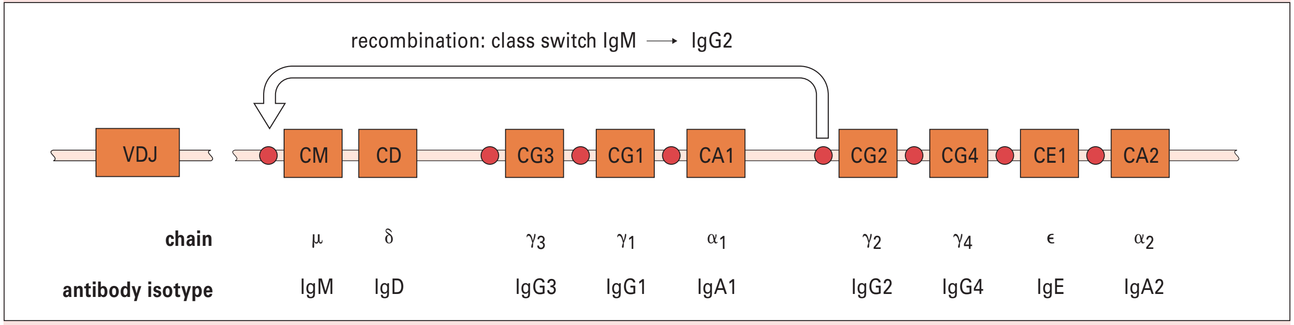

Fig. 8.19 The human immunoglobulin heavy chain gene locus (IGH) is shown. Initially, B cells transcribe a VDJ gene and a μ heavy chain that is spliced to produce mRNA for IgM. Under the influence of T cells and cytokines, class switching may occur, illustrated here as a switch from IgM to IgG2. Each heavy chain gene except CD (which encodes IgD) is preceded by a switch region. When class switching occurs, recombination between these regions takes place, with the loss of the intervening C genes – in this case CM, CD, CG3, CG1, and CA1. (Note that pseudogenes have been omitted from this diagram.)

raised in gnotobiotic (virtually sterile) environments where exposure to exogenous antigens is severely restricted.

Class switching is achieved by gene recombination

B cells switch from IgM to the other classes or subclasses by an intrachromosomal deletion process in which the intervening genetic material between highly repetitive switch regions 5′ to each CH gene is excised as a circle (Fig. 8.21).

Switching involves cytokine-dependent transcription of DNA in the region of the new constant region, reflecting changes in the chromatin in that region. This occurs before recombination of the 5′ switch regions that precede the genes for each of the heavy chain isotype constant region domains.

Class switching may also be achieved by differential splicing of mRNA

Class switching is important in the maturation of the immune response and may be accompanied or preceded by somatic mutation.

Initially, a complete section of DNA that includes the recombined VDJ region and the δ and μ constant regions is transcribed. Two mRNA molecules may then be produced by differential splicing, each with the same VDJ segment, but having either μ or δ constant regions.

It is suggested that much larger stretches of DNA are sometimes also transcribed together, with differential splicing giving other immunoglobulin classes sharing VH regions (Fig. 8.22). This has been observed in cells simultaneously producing IgM and IgE.

Production of membrane and secreted immunoglobulin results from differential transcription of the H chain region

Membrane-bound immunoglobulin (BCR) is identical to secreted immunoglobulin (antibody) except for an extra stretch of amino acids at the C terminus of each heavy chain.

B cell differentiation – class diversity

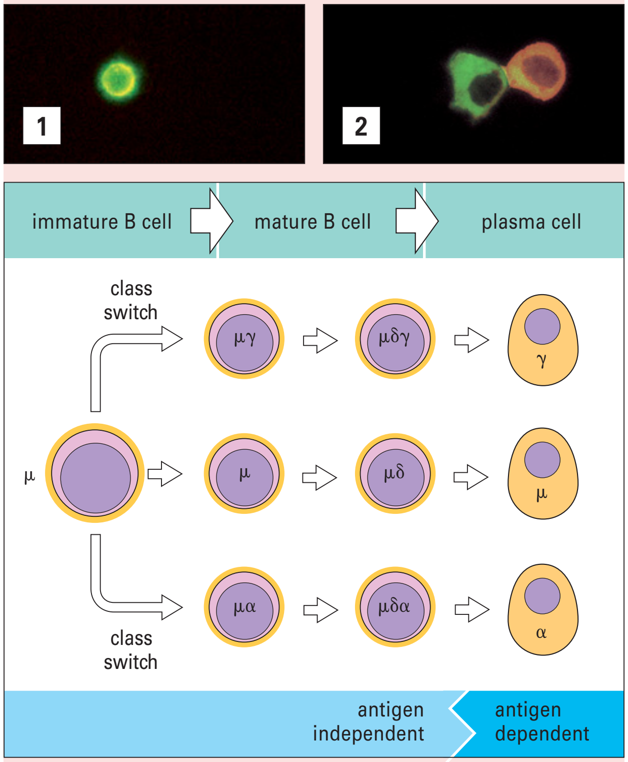

Fig. 8.20 Immature B cells produce IgM only, but mature B cells can express more than one cell surface antibody because mRNA and cell surface immunoglobulin remain after a class switch. IgD is also expressed during clonal maturation. Maturation can occur in the absence of antigen, but the development into plasma cells (which have little surface immunoglobulin, but much cytoplasmic immunoglobulin) requires antigen and (usually) T cell help. The photographs show B cells stained for surface IgM (green – **1**) and plasma cells stained for cytoplasmic IgM and IgG (green and red – **2**). IgM is stained with fluorescent anti-μ chain and IgG with rhodaminated anti-γ chain.

Class switching by gene recombination

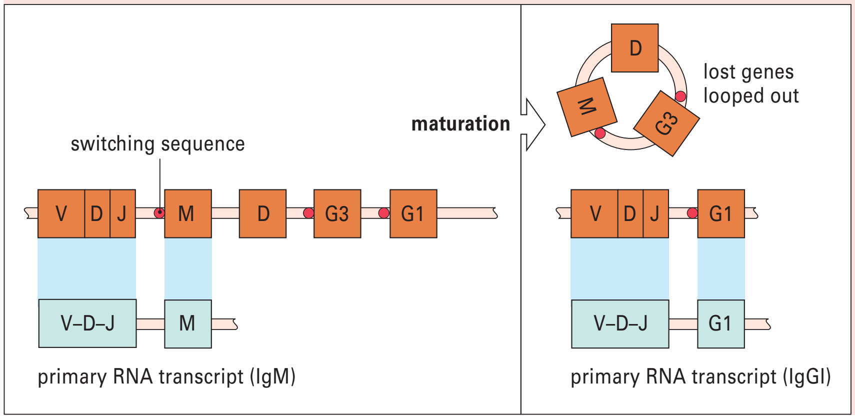

Fig. 8.21 Initially the VDJ region is transcribed together with the M gene for the IgM heavy chain (left). After removal of introns during processing, mRNA for secreted IgM is produced. During B cell maturation, class switch recombination occurs between the Sμ (the switch region of the μ heavy chain gene) recombination region and a downstream switch region (G1 in this example). The intervening region (containing genes for IgM, IgD, and IgG3 in this instance) is looped out and then cut, with deletion of the intervening regions and joining of the two switch regions.

Isotype switching by differential RNA splicing

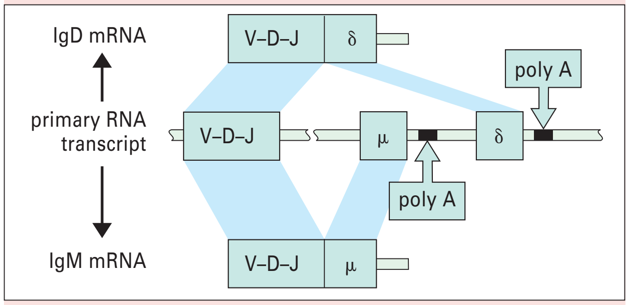

Fig. 8.22 Single B cells produce more than one antibody isotype from a single long primary RNA transcript. A transcript containing μ and δ is shown here. Polyadenylation can occur at different sites, leading to different forms of splicing, producing mRNA for IgD (top) or IgM (bottom). Even within this region, there are additional polyadenylation sites that determine whether the translated immunoglobulin is the secreted or membrane-bound form.

Membrane immunoglobulins are therefore slightly larger than their secreted counterparts. Their additional amino acids traverse the cell membrane and anchor the molecule in the lipid bilayer. In membrane IgM, for example, a section of hydrophobic (lipophilic) amino acids is sandwiched between hydrophilic amino acids, which lie on either side of the membrane (Fig. 8.23).

Membrane immunoglobulins:

- exist only as the basic four-chain unit;
- do not polymerize further; and
- are associated with molecules involved in signal transduction (e.g. Igα and Igβ).

Production of the two forms of immunoglobulin (membrane and secreted) occurs by differential transcription of the germline C region (Fig. 8.24). It is thought that the poly A sequence is important in determining which RNA transcript is produced, but exactly how this is controlled is uncertain.

Immunoglobulin class expression is influenced by cytokines and antigenic stimulus

During a TD immune response, there is a progressive change in the predominant immunoglobulin class of the specific antibody produced, usually to IgG. This class switch is not seen in TI responses, in which the predominant immunoglobulin usually remains IgM.

There is now considerable evidence for the involvement of T cells and their cytokines in the de-novo isotype switching:

- in mice, T cells in mucosal sites have been shown to stimulate IgA production;
- IL-4 preferentially switches B cells that have been either polyclonally activated (by lipopolysaccharide) or specifically activated by antigen to the IgG1 or IgE isotype, with concomitant suppression of other isotypes, and in a similar system IL-5 induces a 5–10-fold increase in IgA production with no change in other isotypes;

Membrane and secreted IgM

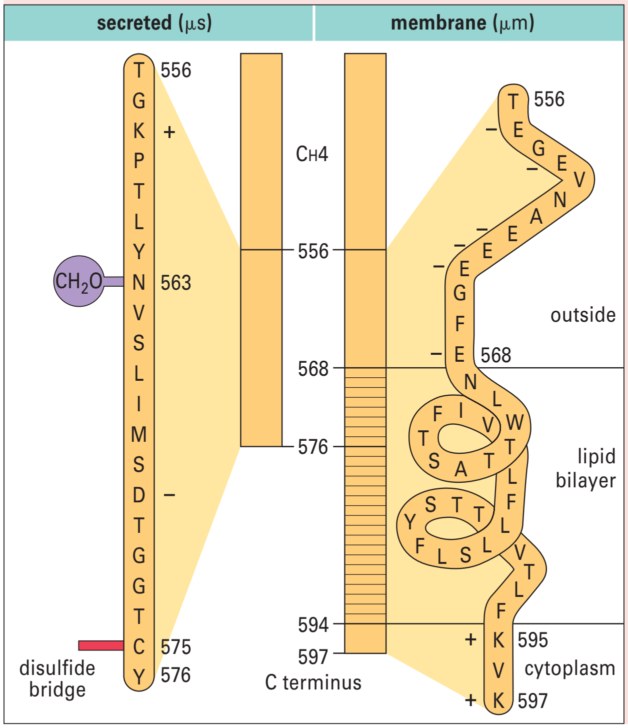

Fig. 8.23 C terminal amino acid sequences for both secreted and membrane-bound IgM are identical up to residue 556. Secreted IgM has 20 further residues. Residue 563 (asparagine) has a carbohydrate unit attached to it, while residue 575 is a cysteine involved in the formation of interchain disulfide bonds. Membrane IgM has 41 residues beyond 556. A stretch of 26 residues between 568 and 595 contains hydrophobic amino acids sandwiched between sequences containing charged residues. This hydrophobic portion may traverse the cell membrane as two turns of a helix. A short, positively charged, section lies inside the cytoplasm. Mouse IgM is shown in this example.

Production of membrane and secreted IgM

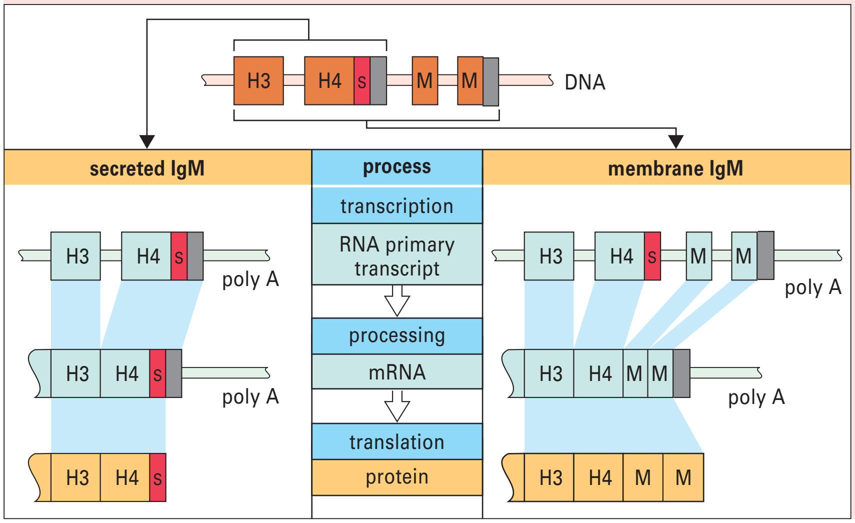

Fig. 8.24 Part of the DNA coding for IgM is shown diagrammatically. The exons for the µ3 and µ4 domains (H3 and H4) and the transmembrane and cytoplasmic segment of membrane IgM (M) are indicated. The 3′ untranslated sequence is present at the end of the H4 and second membrane segments (S = stop codons). The DNA can be transcribed in two ways. If transcription stops after S, the transcript with a poly A tail is processed to produce mRNA for secreted IgM. If transcription runs through to include the membrane segments, processing removes the codons for the terminal amino acids and the stop signal of H4, so translation yields a protein with a different C terminus.

Isotype regulation by mouse T cell cytokines

TH	cytokines	immunoglobulin isotypes					
		IgG1	IgE	IgA	IgG3	IgG2b	IgG2a
TH2	IL-4	↑	↑	↓	↓	↓	↓
	IL-5	=	=	↑	=	=	=
TH1	IFNγ	↓	↓	↓	↓	↓	↑

Fig. 8.25 The effects of IFNγ (product of TH1 cells) and IL-4 and IL-5 (products of TH2 cells), which result in an increase (↑), a decrease (↓), or no change (=) in the frequency of isotype-specific B cells after stimulation with the polyclonal activator lipopolysaccharide (LPS) in vitro.

- IFNγ enhances IgG2a responses in mice, but suppresses all other isotypes (Fig. 8.25).

It is interesting that IL-4 and IFNγ, which act as reciprocal regulatory cytokines for the expression of antibody isotypes, are derived from different TH subsets. In addition, IL-12 and IL-18 stimulation of mouse T cells can induce the production of IFNγ. These cells can therefore act as immunoregulatory cells by differentially inducing IgG2a expression, while inhibiting IgG1, IgE, and IgG2b expression. Transforming growth factor-β (TGFβ) induces the switch to IgA or IgG2b. In humans, the situation is somewhat different:

- IL-4 induces the expression of IgG4 and IgE; and
- TGFβ induces the expression of IgA alone.

FURTHER READING

Hardy RR, Hayakawa K. B cell development pathways. Annu Rev Immunol 2001;19:595–621.

Henderson A, Calamé K. Transcriptional regulation during B cell development. Annu Rev Immunol 1998;16:163–200.

Lane P. Development of B cell memory and effector function. Curr Opin Immunol 1996;8:331–336.

Lipsky PE, Attrep JF, Grammer AC, et al. Analysis of CD40–CD40 ligand interactions in the regulation of human B cell function. Ann NY Acad Sci 1997;815:372–383.

Liu Y-J, de Bouteiller O, Fugier-Vivier I. Mechanisms of selection and differentiation in germinal centers. Curr Opin Immunol 1997;9:256–262.

McHeyzer-Williams LJ, McHeyzer-Williams MG. Antigen-specific memory B cell development. Annu Rev Immunol 2005;23:487–513.

Osmond DG, Rolink A, Melchers F. Murine B lymphopoiesis: towards a unified model. Immunol Today 1998;19:65–68.

Przylepa J, Himes C, Kelsoe G. Lymphocyte development and selection in germinal centers. Curr Top Microbiol Immunol 1998;229:85–104.

Smith KG, Fearon DT. Receptor modulators of B cell receptor signaling – CD19/CD22. Curr Top Microbiol Immunol 2000;245:195–212.

Snapper CM, Mond JJ. A model for induction of T cell-independent humoral immunity in response to polysaccharide antigens. J Immunol 1996;157:2229–2233.

Stavnezer J. Antibody class switching. Adv Immunol 1996;61:79–146.

Tarlington D. Germinal centers: form and function. Curr Opin Immunol 1998;10:245–251.

Critical thinking: Development of the antibody response (see p. 495 for explanations)

A project is under way to develop a vaccine against mouse hepatitis virus, a pathogen of mice that may become a serious problem in colonies of mice. The vaccine consists of capsid protein of the virus, which is injected subcutaneously as a depot in alum on day 0. At days 5 and 14, the group of six mice is bled and the serum is tested for the presence of antibodies against the viral capsid protein. Separate assays are done for each of the immunoglobulin classes, IgM, IgG, and IgA. The amounts, expressed in mg/ml of antibody are shown in the figure.

When the data are analyzed it appears that two of the animals have high titers of antibody, particularly of IgG and IgA, at both days 5 and 14.

1 Why do the titers of IgG antibodies increase more rapidly between days 5 and 14 than the titers of IgM antibodies, in all animals?

2 Propose an explanation for the high titers of IgG antibodies in the two animals indicated at day 5. Can this explanation also account for the relatively high levels of IgA antibodies also seen in these mice?

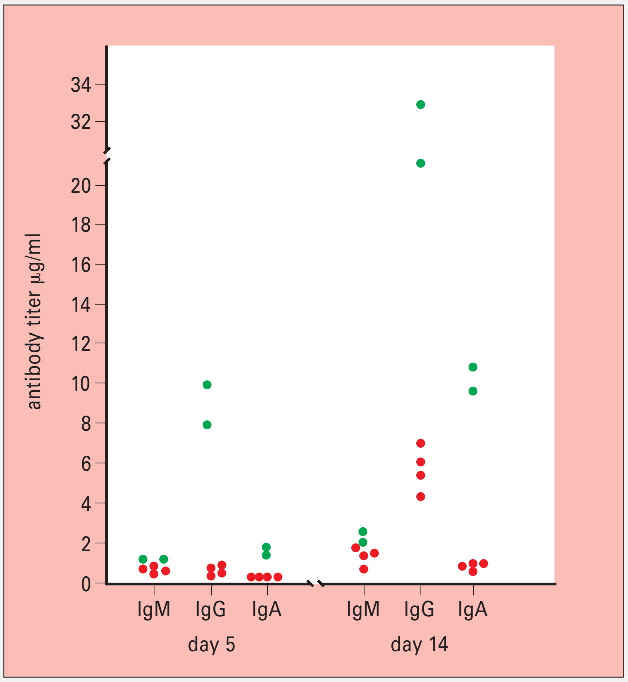

The spleens from mice taken at day 14 are used to produce B cells making monoclonal antibodies against the viral protein (see Method Box 3.1). Of the clones produced, 15 produce IgG, three produce IgM, and none produces IgA.

3 Why do you suppose there are no IgA-producing clones, despite the good IgA response?

4 You want a high-affinity antibody for use in an assay. Which of the clones you have produced are likely to be of higher affinity?

CHAPTER

9

Mononuclear Phagocytes in Immune Defense

SUMMARY

- **Macrophages, myeloid dendritic cells, and osteoclasts all differentiate from circulating blood monocytes**.

- **Resident macrophages are widely distributed throughout the body.** Phenotypically distinct populations are present in each organ and within the different zones of spleen and lymph nodes.

- **Resident and recruited macrophages respond to injury and immune stimuli.** TH1 cytokines such as IFNγ enhance inflammation and antimicrobial activity. TH2 cytokines induce an alternate activation with efficient antigen presentation to B cells. TGFβ, corticosteroids, and IL-10 can induce an anti-inflammatory phenotype.

- **Resident macrophages clear apoptotic cells using scavenger receptors and the vitronectin receptor.** Endocytosis by this pathway does not activate the macrophage killing mechanisms.

- **Macrophages internalize pathogens using a variety of specific and opsonic receptors.** These include the LPS receptor and Toll-like receptors, the β-glucan receptor, the mannose receptor (MR), the Fc receptors, and complement receptors CR1, CR3, and CR4.

- **Activated macrophages secrete cytokines, enzymes, complement system molecules, and procoagulant.** Resting macrophages have only a limited killing activity.

- **Activated macrophages produce reactive oxygen and nitrogen intermediates which are highly toxic for endocytosed bacteria and fungi.** Recently recruited cells are most effective at antimicrobial killing.

- **Macrophages can initiate, promote, prevent, suppress, and terminate immune responses**.

MACROPHAGES DIFFERENTIATE FROM CIRCULATING BLOOD MONOCYTES

Mononuclear phagocytes comprise a family of cells that share common hematopoietic precursors and are distributed via the blood stream as monocytes to all tissues of the body, including secondary lymphoid organs, even in the absence of an overt inflammatory stimulus.

Within tissues the mononuclear phagocytes undergo maturation, adapt to their local microenvironment, and differentiate into various cell types:

- macrophages;
- myeloid-derived dendritic cells; and
- osteoclasts.

These various cell types perform specific housekeeping, trophic, and immunologic functions (Fig. 9.1).

The macrophages for the most part become and remain highly efficient phagocytes and play an important role in pathogen recognition and clearance during infection, as well as removal of senescent and dying cells and immune complexes (Fig. 9.2).

In response to particulate and other potential antigenic stimuli:

- the dendritic cells are uniquely efficient inducers of primary immune responses by naive T cells; whereas
- macrophages produce a range of secretory products that affect the migration and activation of other immune cells.

Macrophages are actively endocytic and degrade antigens, but can present peptides to already primed T cells in secondary responses after MHC and other accessory molecules have been induced by cytokines such as IFNγ derived from TH1 cells and NK cells. Once recruited and activated by such interactions, the macrophage plays a major role in the effector limb of cell-mediated immunity to intracellular pathogens such as mycobacteria, often within focal accumulations of cells known as granulomas.

Macrophages are extremely heterogeneous in their gene expression and cellular activities, with beneficial as well as destructive roles in tissue homeostasis and host defense.

Through their varied plasma membrane receptor repertoire and secretory responses, macrophages interact with other leukocytes as well as non-hematopoietic cells in all tissues, so regulating both innate and acquired immunity.

Although the concept of the mononuclear phagocyte system is central to our understanding of immunity, this is but one role in a wider context of physiologic and pathologic contributions to tissue maintenance, response to injury, and repair.

Differentiation of monocytes

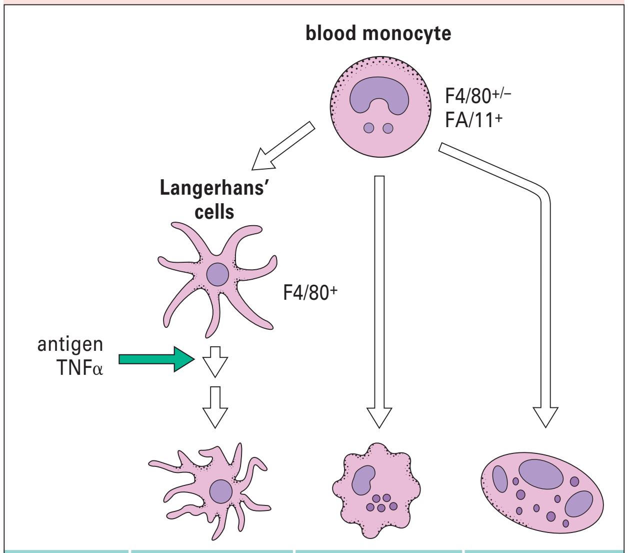
blood monocyte
F4/80+/–
FA/11+
Langerhans' cells
F4/80+
antigen
TNFα

marker	mature dendritic cell	tissue macrophage	osteoclast
F4/80	–	+/–	–
FA/11	+	+	+

Fig. 9.1 Circulating monocytes give rise to myeloid dendritic cells (distinct from plasmacytoid dendritic cells, which are thought to originate from distinct precursors), tissue macrophages, and osteoclasts. F4/80 and FA/11 (macrosialin, the murine homolog of CD68) are differentiation antigens of mouse macrophages and closely related cells.

Rapid Reference Box 9

chemokines – a large group of cytokines falling into four separate families. The main families are the CC group and the CXC group. They include interleukin-8 (CXCL8), monocyte chemotactic protein-1 (MCP-1, CCL2), RANTES, (CCL5) macrophage inflammatory protein-1α (MIP-1α, CCL3) interferon-inducible protein 10 (IP-10, CCL10), and many others. They act on seven-transmembrane pass receptors and have a variety of chemotactic and cell-activating properties, acting on selected populations of target cells (see Appendix 4).

CSFs (colony stimulating factors) – a group of cytokines that control the differentiation of hematopoietic stem cells.

cytokines – a generic term for soluble molecules that mediate interactions between cells (see Appendix 3).

IFNs (interferons) – a group of molecules involved in signaling between cells of the immune system and in protection against viral infections.

IL-1–IL-27 (interleukins) – a group of molecules involved in signaling between cells of the immune system.

LTs (leukotrienes) – a collection of metabolites of arachidonic acid that have powerful pharmacological effects.

LPS (lipopolysaccharide) – a product of some Gram-negative bacterial cell walls that can act as a B cell mitogen.

PGs (prostaglandins) – pharmacologically active derivatives of arachidonic acid. Different prostaglandins are capable of modulating cell mobility and immune responses.

TGFs (transforming growth factors) – a group of cytokines identified by their ability to promote fibroblast growth; they are also generally immunosuppressive.

RESIDENT MACROPHAGES ARE WIDELY DISTRIBUTED THROUGHOUT THE BODY

Some of the key features of resident macrophage populations in tissues are shown in Fig. 9.3. The cells vary in their life span, morphology, and phenotype, for example the microglial cells in the brain appear quite unlike mononuclear phagocytes in other tissues (Fig. 9.4).

Resident cells have usually ceased to proliferate, but retain the ability to express a range of mRNA and proteins, often as relatively long-lived cells, with low turnover, unlike neutrophils.

Plasma membrane differentiation markers such as the F4/80 antigen have proved useful in defining the distribution of mature macrophages in many (but not all)

Role of macrophages and dendritic cells in immune defense

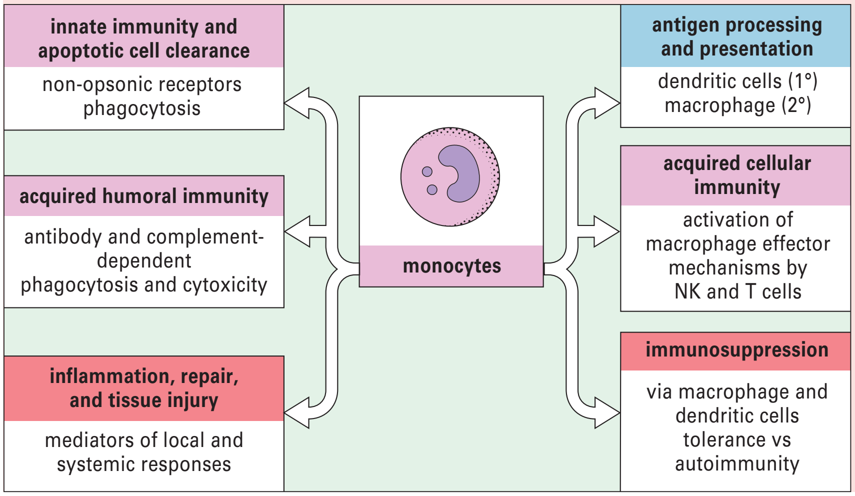
innate immunity and apoptotic cell clearance
non-opsonic receptors phagocytosis
acquired humoral immunity
antibody and complement-dependent phagocytosis and cytoxicity
inflammation, repair, and tissue injury
mediators of local and systemic responses
monocytes
antigen processing and presentation
dendritic cells (1°)
macrophage (2°)
acquired cellular immunity
activation of macrophage effector mechanisms by NK and T cells
immunosuppression
via macrophage and dendritic cells tolerance vs autoimmunity

Fig. 9.2 Both macrophages and dendritic cells play important roles in innate immune responses, inflammation, and tissue remodeling, as well as in specific immune responses.

Resident tissue macrophage populations

organ	name/site	functions/properties
bone marrow	stromal macrophage	interacts with hematopoietic cells, removes erythroid nuclei
liver	Kupffer cells	clearance of cells and complexes from the blood
spleen	red pulp macrophages white pulp tingible body macrophages marginal zone macrophages	clearance of senescent blood cells phagocytosis of apoptotic B cells interface between circulation and immune system
lymph node	subcapsular sinus macrophages medullary macrophages	interface with afferent lymph interface with efferent lymph
thymus	thymic macrophage	clearance of apoptotic cells
gut	lamina propria	endocytosis
lung	alveolar macrophage	clearance of particulates and infectious agents
brain	microglia in neuropil choroid plexus	interacts with neurons interface with cerebrospinal fluid
skin	Langerhans' cells	antigen capture
reproductive tract	ovary, testis	clearance of dying cells
endocrine organs	adrenal, thyroid, pancreas, etc.	metabolic homeostasis
bone	osteoclasts	bone remodeling

Fig. 9.3 Resident tissue macrophage populations.

Microglia in mouse brain stained for F4/80

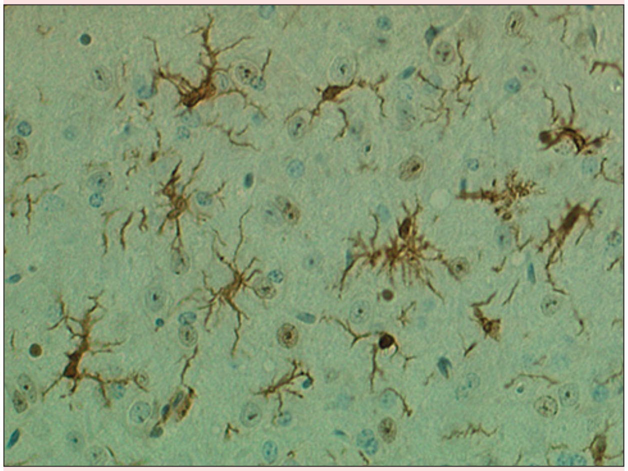

Fig. 9.4 Microglia are widely distributed in non-overlapping fields and show a dendritic morphology. (Courtesy of Dr Payam Rezaie)

murine tissues. In humans, the CD68 antigen, an intracellular vacuolar marker, is widely expressed; the murine homolog (macrosialin) is a pan-macrophage marker, also present in many myeloid dendritic cells and osteoclasts, unlike F4/80.

It is possible to reconstruct a constitutive migration pathway in which monocytes become endothelial-like and line vascular sinusoids, as in the liver (Kupffer cells, see Fig. 2.3), or penetrate between endothelial cells. They underlie endothelia or epithelia, which can also be penetrated by, for example, Langerhans' cells (F4/80$^+$ precursors of dendritic cells), or enter the interstitial space or serosal cavities (Fig. 9.5).

Although macrophages are often regarded as sessile cells compared with freely migratory dendritic cells, they readily migrate to draining lymph nodes after an inflammatory stimulus and become arrested there. They are therefore absent from efferent lymph and do not, as a rule, re-enter the circulation.

Q. Why is there no advantage for the immune system in the recirculation of macrophages, whereas recirculation of lymphocytes is a central element in immune defense?

A. Macrophages do not develop an immunological memory, and do not undergo the selective clonal expansion of antigen-stimulated lymphocytes that occurs in lymphoid tissues. Therefore there is no advantage in having macrophages return to the circulation from lymphoid tissues.

Mature macrophages are themselves part of the stromal microenvironment in bone marrow. They associate with developing hematopoietic cells to perform poorly defined non-phagocytic trophic functions, as well as removing effete cells and erythroid nuclei.

Secondary lymphoid organs contain several distinct types of macrophage and related cells, which are still poorly understood. Fig. 9.6 illustrates the complexity of

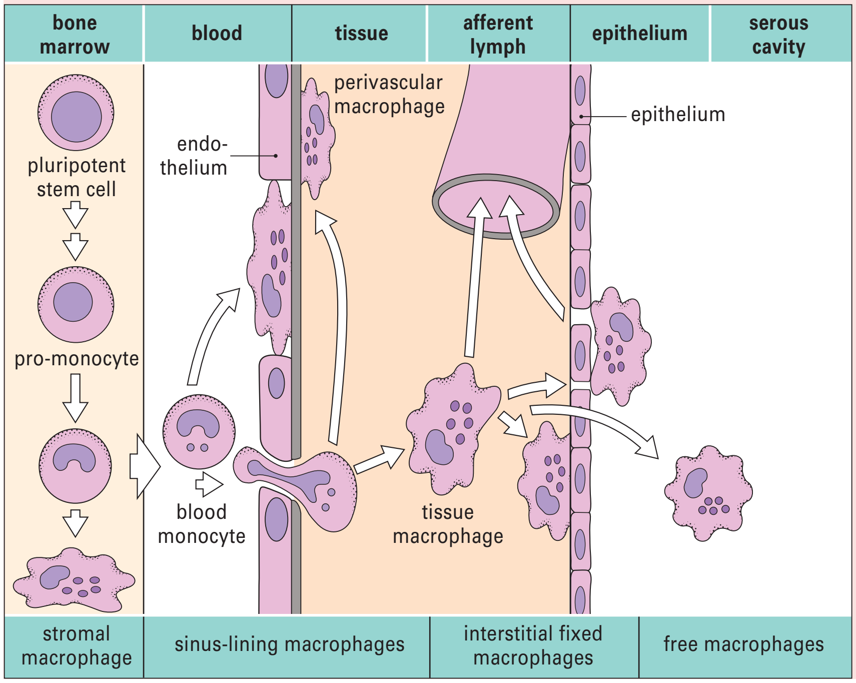

Differentiation and distribution of macrophages

Fig. 9.5 Blood monocytes are derived from bone marrow in the adult host and enter tissues initially as sinus-lining or extravascular mature cells. Interstitial or intraepithelial adherent macrophages and serosal macrophages can enter afferent lymphatics. Monocyte-derived dendritic cells can arise from tissue macrophages by reverse migration across endothelium. Stromal macrophages in bone marrow may also derive from the circulating monocyte population.

macrophage-like cells in rodent spleen (red and white pulp, marginal zone) and normal lymph nodes (subcapsular sinus, medulla).

Differentiation antigens such as sialoadhesin, a lectin-like receptor for sialylated glycoconjugates, are particularly strongly present on marginal metallophils (inner marginal zone) and on subcapsular sinus macrophages. These cells are present at the interface between:

- blood (spleen);
- afferent lymph (lymph nodes); and
- organized lymphoid structures.

They play a role in the capture of organisms, particulates, polysaccharides, and soluble antigens, as well as of circulating host cells or migrating dendritic cells, which have their own characteristic distribution in T and B cell-dependent areas.

The molecular mechanisms of constitutive macrophage distribution and induced migration are beginning to be defined, and involve cellular adhesion molecules, cytokines, and growth factors, as well as chemokines and chemokine receptors, as summarized in Fig. 9.7.

Macrophage colony stimulating factor (M-CSF or CSF-1) is a major growth, differentiation, and survival factor selective for macrophages, whereas granulocyte–macrophage colony stimulating factor (GM-CSF) regulates myeloid cell production and function.

Chemokines are produced by a range of hematopoietic and tissue cells, including macrophages themselves, and act on various leukocyte subpopulations depending principally on their chemokine receptor profile.

RESIDENT AND RECRUITED MACROPHAGES RESPOND TO INJURY AND IMMUNE STIMULI

Inflammatory stimuli (e.g. local infection) enhance the recruitment of monocytes (and often other myeloid cells) from blood, and ultimately from bone marrow stores.

Monocytes adhere to activated endothelium through a series of well-described interactions involving selectins, β_2 integrins, and CD31, as well as chemotactic molecules acting on their transmembrane G protein-coupled receptors (Fig. 9.8, see also Chapter 6). When ligands bind to these receptors it causes the recruitment of intracellular proteins to the area of membrane activation of integrins and polarization of the cells (Fig. 9.9).

Diapedesis results in local tissue interactions and accumulation of macrophages with enhanced turnover and an altered phenotype (e.g. the upregulation of endocytic receptors and production of proinflammatory mediators).

Phagocytosis by recently recruited monocytes profoundly alters their differentiation into:

- migratory dendritic cells; or
- more sessile macrophages.

Recent studies have confirmed the potential of monocytes to differentiate into either of these cell types in vivo, as they do in cell culture.

It is convenient to distinguish the macrophages 'elicited' by a non-specific inflammatory stimulus from 'immunologically activated' macrophages. The latter cells:

- respond to IFNγ by acquiring enhanced antimicrobial properties; and

Macrophages in secondary lymphoid tissues

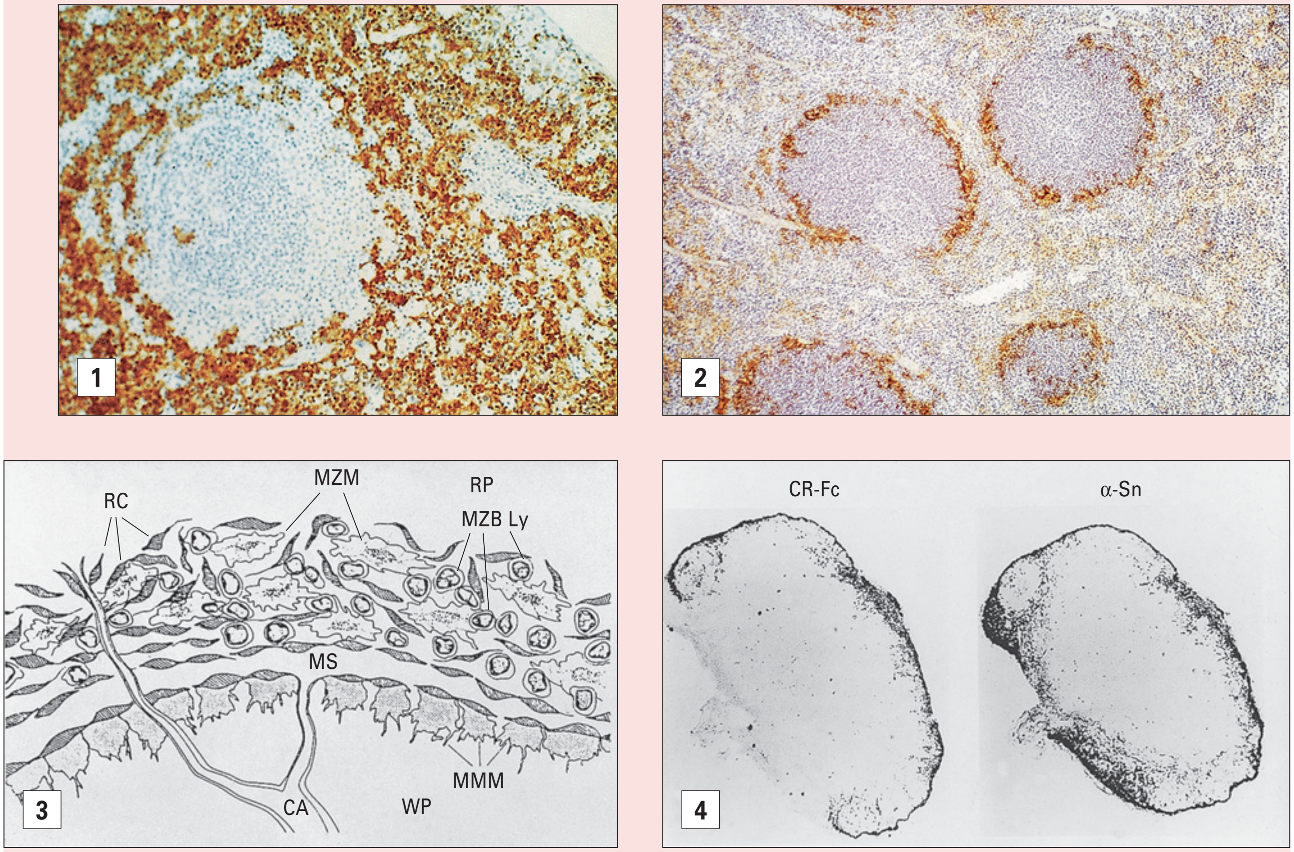

Fig. 9.6 Heterogeneity of macrophages in secondary lymphoid organs of mouse. (**1**) Red pulp of spleen stained for F4/80. Macrophages stain strongly positive. (Courtesy of Dr DA Hume) (**2**) Mouse spleen stained with antibody to sialoadhesin. The marginal metallophils of spleen are strongly sialoadhesin positive. (Courtesy of Dr PR Crocker) (**3**) The marginal zone contains inner metallophilic and outer zones. In the diagram of the marginal zone of the spleen, the central arteriole (CA) branches into small capillaries, which either pass through the marginal zone and end in the red pulp (RP) or open into the marginal sinus (MS). The marginal zone is composed of reticular cells (RCs). Within this reticular framework, the large marginal zone macrophages (MZMs), dendritic cells, and marginal zone B cells (MZBs) are localized. The marginal metallophilic macrophages (MMM) are situated at the inner border of the marginal sinus and the white pulp (WP) (Courtesy of Dr G Kraal) (**4**) Lymph node contains subcapsular sinus macrophages, which are sialoadhesin positive and bind CR-Fc, a chimeric probe of the mannose receptor cysteine-rich domain, and human Fc (left); medullary macrophages express only sialoadhesin (right). (Courtesy of Dr L Martinez-Pomares)

- express enhanced MHC class II molecules, whereas other genes (e.g. the mannose receptor) are downregulated.

Q. What will be the functional effect of the changes in cell surface molecules described above?

A. The increase in MHC class II molecules will enhance the ability to present antigen to CD4$^+$ T cells, while the reduction in the MR may mean a reduced ability to take up some microbes by non-immune mechanisms.

Cytokines produced by NK cells, lymphocytes, and various antigen-presenting cells (APCs) influence the pattern of gene expression by macrophages. Work in vitro, and some studies in vivo, especially with cytokine/receptor knockout mouse strains, have made it possible to classify macrophages differing in their functional state (Fig. 9.10).

The initial and further interactions with endocytic and phagocytic stimuli give rise to further heterogeneity in phenotype. It should be emphasized that, although the evidence for individual molecules as mediators of activation is good, we have little insight into the complex interactions that pertain in vivo.

Fig. 9.11 illustrates a mycobacterium-induced granuloma, rich in macrophages. Cytokines, for example TNFα, and receptors such as CR3 play an important role in granuloma formation.

Environmental influences on macrophages

stimulus	example/receptors	response
growth factors	M-CSF, GM-CSF	growth, differentiation, survival
cytokines	IFNγ TNFα IL-4, IL-13 IL-10	activation activation alternate activation deactivation
chemokines	CCL2 (macrophage chemotactic protein-1, MCP-1)	migration
extracellular matrix	fibronectin / β_1-integrin	adhesion, phagocytosis
peptides	vasoactive peptide (VIP)	modulation of various functions
eicosanoids	prostaglandin E_2 (PGE_2) leukotriene B_4 (LTB_4)	modulation of various functions
cellular interactions: endothelium T cells	fibronectin / β_1- integrin accessory molecules (CD80, CD86)	recruitment antigen presentation
microbial interactions	LPS (CD14)	secretion, activation
proteases	neutral proteinases (e.g. plasmin)	adhesion, secretion

M-CSF, macrophage colony stimulating factor; GM-CSF, granulocyte—macrophage colony stimulating factor; IFNγ, interferon-γ; IL, interleukin; LPS, lipopolysaccharide; TNFα, tumor necrosis factor-α

Fig. 9.7 Some of the mediators that act on macrophages.

Macrophage 7tm G protein-coupled receptors for chemotactic stimuli

receptors	ligands	sources
fMLP	fMet-Leu peptides	prokaryote protein synthesis
C5a receptor	complement C5a	lytic activation
LTB_4 receptor	leukotriene B_4	mast cells, macrophages
PAF receptor	platelet-activating factor	platelets, neutrophils, macrophages
CCR1, CCR2, CCR5	chemokines CCL2, 3, 4, 5, 7, 8	leukocytes, tissue cells
CXCR1, CXCR2	CXCL8 (IL-8)	endothelium, lymphocytes
CXCR3	CXCL9, 10, 11	endothelium, tissue cells
CX3CR1	CX3CL1 (fractalkine)	endothelium, epithelium, neurons, smooth muscle cells

7tm, seven-transmembrane segments; CCL, CC chemokine ligand; CXCL, CXC chemokine ligand; CCR, CC chemokine receptor; CXCR, CXC chemokine receptor; fMLP, f.Met-Leu-Phe; PAF, platelet-activating factor

Fig. 9.8 Macrophages express a variety of receptors for chemotactic agents. Fractalkine (CX3CL1) is unique as a transmembrane ligand – its receptor, CX3CR1, and CCR2 may be differentially expressed by monocytes, giving rise to inflammatory and resident macrophages.

Intracellular signaling via chemotactic agents such as f-met-leu-phe (fMLP)

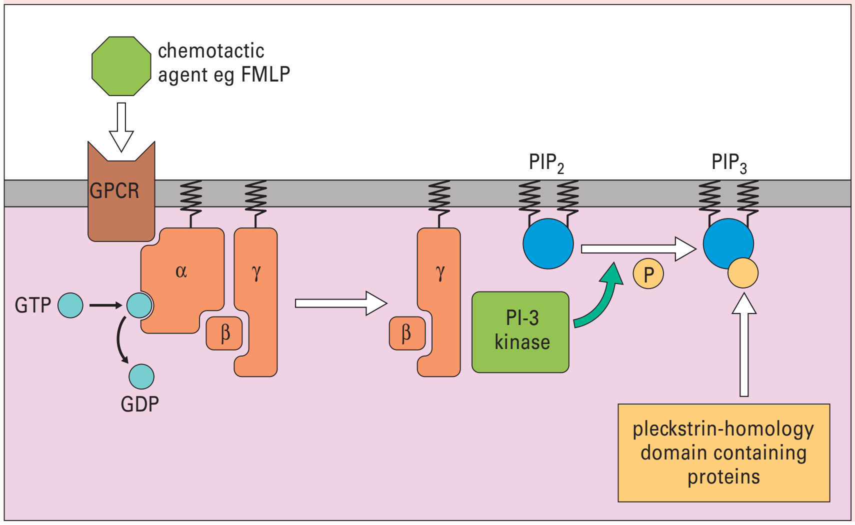
chemotactic agent eg FMLP
GPCR
PIP$_2$
PIP$_3$
α
γ
γ
GTP
β
β
P
PI-3 kinase
GDP
pleckstrin-homology domain containing proteins

Fig. 9.9 fMLP binds to its receptor and causes guanosine triphosphate (GTP) to displace guanosine diphosphate (GDP) on the heterotrimeric G protein. This causes the β and γ chains to dissociate from the α chain and they in turn recruit phosphatidyl inositol (PI)-3 kinase to the membrane. The kinase phosphorylates phosphatidyl inositol diphosphate (PIP$_2$) to the triphosphate PIP$_3$. A number of proteins that have pleckstrin homology (PH) domains are then recruited to the membrane at this point, causing focal adhesion, activation of integrins, and polarization of the cell, in preparation for migration.

Activation of macrophages

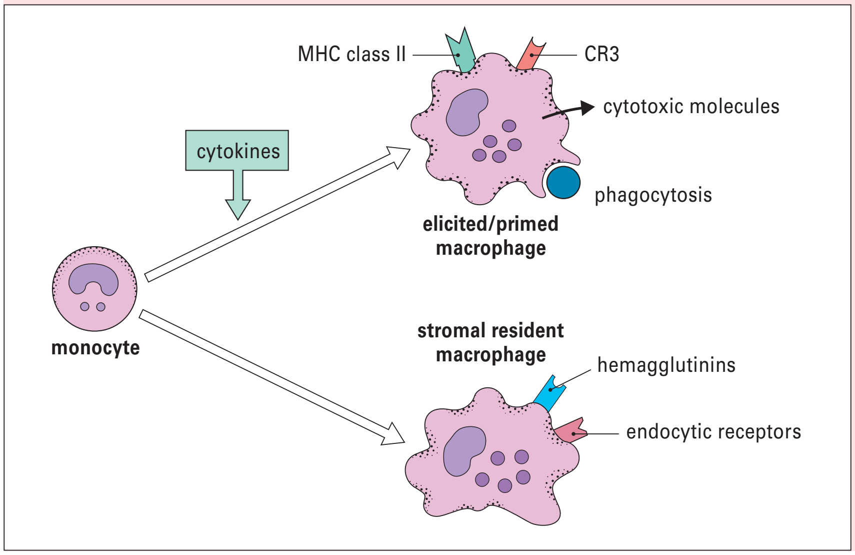

Fig. 9.10 Elicited and immunologically primed macrophages differ from resident macrophages. Monocyte recruitment is enhanced and yields macrophages with proinflammatory and cytotoxic properties. Activation by cytokines enhances expression of MHC class II molecules and the complement receptor CR3. Phagocytosis and the production of proinflammatory mediators and cytotoxic products are increased. By contrast, resident macrophages (e.g. in bone marrow) lack inflammatory functions, but participate in trophic reactions (e.g. with developing hematopoietic cells) as well as performing endocytosis. Microbial stimuli can act directly on resident macrophages to induce distinct surface and secretory properties.

Immunologically activated macrophages express the capacity to produce antimicrobial products including lysozyme and reactive oxygen and nitrogen species (see below). However, characteristic morphologic features such as epithelioid cell and giant cell formation, hallmarks of mycobacterial granulomas, remain mysterious in origin.

Cytokines modulate the phenotype of macrophages

A further level of heterogeneity derives from the distinct effects of TH1- and TH2-type cytokines on monocyte/macrophage differentiation (Fig. 9.12).

In-vitro studies with macrophages treated with different cytokines have indicated that there is a spectrum of gene expression induced by different cytokines (Fig. 9.13):

- IFNγ characteristically enhances proinflammatory and antimicrobial activities ('immune activation');
- IL-10 efficiently counteracts proinflammatory and antimicrobial activities ('deactivation');
- IL-4 and IL-13 exert distinctive effects on MHC class II and MR expression, which we have termed 'alternative activation', rather than focusing on the modest inhibition of proinflammatory products.

Broadly speaking:

- 'activated' macrophages mediate cellular immunity;
- 'alternatively activated' macrophages may promote humoral immunity, including repair processes.

It is well established that macrophages and myeloid dendritic cells produce IL-12 and IL-18, which enhance the production of IFNγ by NK cells and T cells. There is, as yet, no comparable IL-4/IL-13-inducing activity that can be ascribed to the APCs, though subsets of dendritic cells have been implicated in TH2 differentiation, which has also been postulated to constitute a 'default' pathway.

A granulomatous reaction in pulmonary tuberculosis

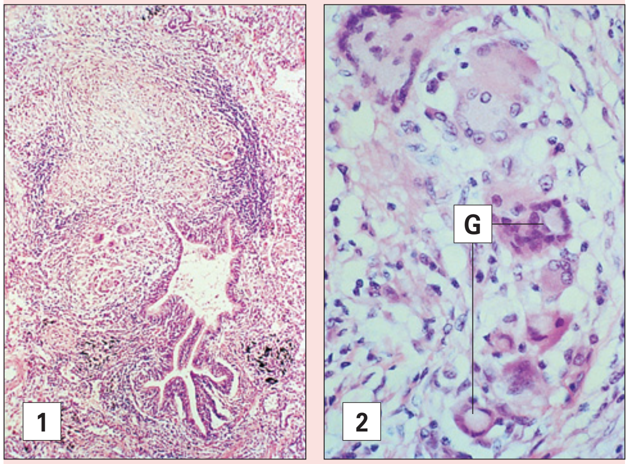

Fig. 9.11 The central area of caseous necrosis in which much of the cellular structure is destroyed is characteristic of tuberculosis in the lung. Apart from this necrosis, the histology is typical of chronic T cell-dependent 'tuberculoid' granulomas. The lesion is surrounded by a ring of epithelioid cells and mononuclear cells. Multinucleate giant cells, believed to be derived from the fusion of epithelioid cells, are also present (left, × 170). Giant cells (G) are illustrated at a higher magnification (right). Hematoxylin and eosin stain. (Courtesy of Dr G Boyd)

The links between innate and acquired immune responses are therefore understood only in part.

MACROPHAGES CLEAR APOPTOTIC CELLS USING SCAVENGER RECEPTORS AND THE VITRONECTIN RECEPTOR

Macrophages express a very wide range of plasma membrane receptors, which underlie their interactions with:

- other cells;
- extracellular ligands derived from plasma;
- extracellular matrix; and
- microorganisms.

It is useful to distinguish:

- opsonic receptors (e.g. for antibody and complement); from
- 'non-opsonic' receptors.

The processes of phagocytosis mediated by opsonic receptors are shown in Fig. 9.14. Whereas neutrophils express similar opsonic receptors, they lack most of the opsonin-independent receptors for phagocytosis.

Q. What groups of receptors do not require opsonins for their actions?

A. The Toll-like receptors, lectin receptors, and scavenger receptors, which recognize pathogen-associated molecular patterns (PAMPs) (see Fig. 6.24).

Modulation of macrophage activation

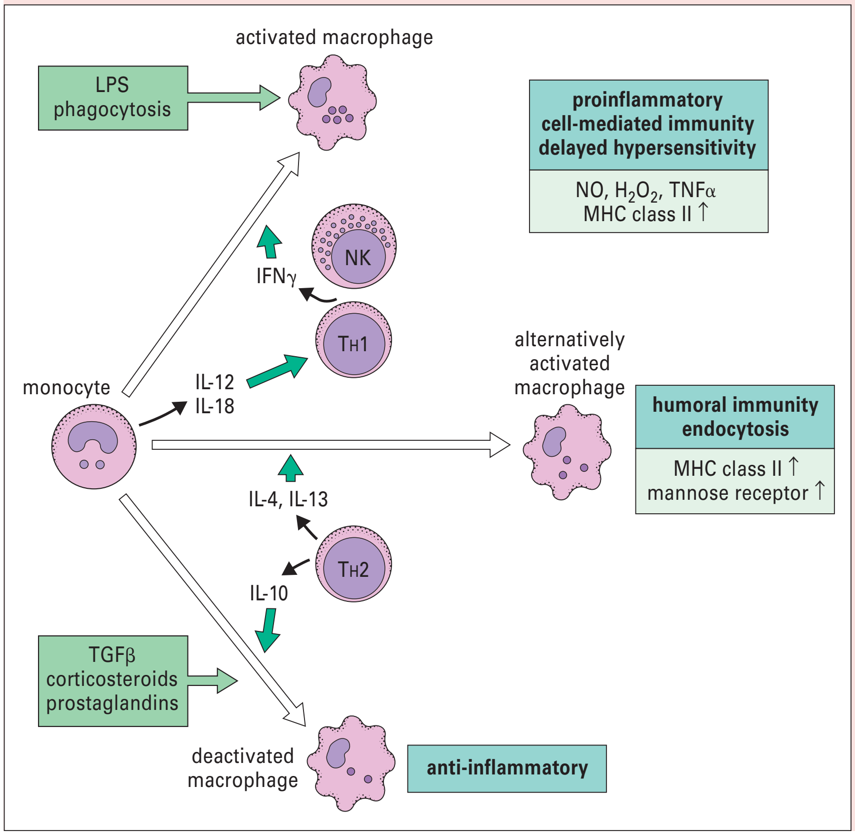
activated macrophage
LPS
phagocytosis
proinflammatory
cell-mediated immunity
delayed hypersensitivity
NO, H_2O_2, TNFα
MHC class II ↑
NK
IFNγ
TH1
IL-12
IL-18
monocyte
alternatively
activated
macrophage
humoral immunity
endocytosis
MHC class II ↑
mannose receptor ↑
IL-4, IL-13
TH2
IL-10
TGFβ
corticosteroids
prostaglandins
deactivated
macrophage
anti-inflammatory

Fig. 9.12 Signals from microbial products, phagocytosis, and cytokines result in changes to the surface and secretory properties of macrophages, which can be classified as activated, deactivated, or alternatively activated. (NO, nitric oxide; TGFβ, transforming growth factor-β)

Regulation of macrophage phenotype

TH1 type		TH2 types		significance
IFNγ		IL-4/IL-13	IL-10	
MHC class II molecule (Ia)	++	+	–	immune cell interactions
respiratory burst NO	++ ++	(–) (–)	– –	cell-mediated immunity tissue injury (e.g. tuberculosis)
TNFα IL-1 IL-6	++	(–)	–	proinflammatory
mannosyl receptor	–	++	0	phagocytosis/endocytosis (e.g. antigens)
growth	–	+	0	local growth of MØ in immune lesions
fusion	0	++	0	giant cell formation–granulomas (e.g. tuberculosis)
growth factor secretion	–	++	+	healing of lesions

Fig. 9.13 TH1 and TH2 cytokines act on macrophages to induce distinctive functions, which can be described as 'activation' (TH1-type), 'alternative activation', or deactivation (TH2-type). (+, increase; -, decrease; 0, no effect)

Phagocytosis mediated by opsonic receptors

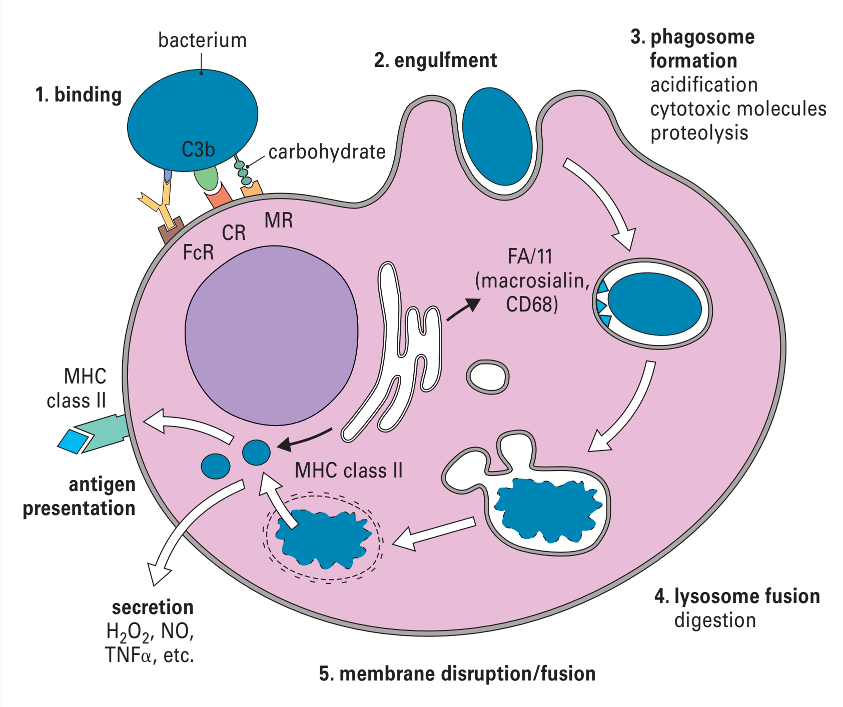
bacterium
1. binding
C3b
carbohydrate
2. engulfment
3. phagosome formation
acidification
cytotoxic molecules
proteolysis
FcR
CR
MR
FA/11
(macrosialin, CD68)
MHC class II
MHC class II
antigen presentation
secretion
H_2O_2, NO, TNFα, etc.
4. lysosome fusion
digestion
5. membrane disruption/fusion

Fig. 9.14 (**1**) Pathogens, such as bacteria, are taken up by binding to opsonic receptors including the Fc receptor, complement receptors, and receptors for carbohydrate (MR). (**2**) The particle is engulfed and the phagosome forms (**3**). Acidification of the phagosome follows as toxic molecules (reactive oxygen and nitrogen intermediates) are pumped into the phagosome. The marker FA/11 is located in the phagosome membrane. (**4**) Lysosomes fuse with the phagosome, releasing proteolytic enzymes into the phagolysosome, which digest the bacteria. (**5**) On completion the membrane of the phagolysosome is disrupted. Antigenic fragments may become diverted to the acidic endosome compartment for interaction with MHC class II molecules and antigen presentation. The process induces secretion of toxic molecules and cytokines.

However, pattern recognition clearly extends to endogenous, host-derived molecules (self, which may be modified), though there may be intriguing differences in the outcome of these receptor–ligand interactions regarding cytokine responses and the inflammatory consequences.

In particular, macrophage receptors play a dual role in clearance of both macromolecules and apoptotic cells, which differs from their role in uptake of microbial invaders or foreign antigens:

- in the first case there is no or even a suppressed inflammatory response;
- in the second, the immune system may need to be alerted.

The roles of cell type, receptor profile, signal pathway, and mediator release in these dual responses are all under investigation.

Q. In what ways can macrophages alert lymphocytes that microbial antigens are present?

A. They can express co-stimulatory molecules and inflammatory cytokines, which are induced following for example ligation of Toll-like receptors by microbial products (see Chapters 6 and 7).

To maintain appropriate numbers of all cell types in development and normal tissue homeostasis, as well as during immune, inflammatory, and other pathologic responses, cells die naturally by:

- programmed death (apoptosis); or
- necrosis, the latter releasing potentially injurious products.

Cellular and biochemical pathways resulting in apoptosis are conserved in evolution; apoptotic cells are rapidly and efficiently cleared by macrophages (e.g. in thymus; Fig. 9.15), though they can also be engulfed by nonprofessional phagocytes.

The receptors implicated in apoptotic cell recognition show considerable redundancy and include a range of scavenger receptors (e.g. SR-AI; Fig. 9.16), vitronectin receptor, an ATP-transporter, and CD14 (see below). Thrombospondin and C1q can act as opsonins.

The ligands expressed by apoptotic cells are not well defined. Recently a conserved, widely distributed receptor has been postulated for phosphatidylserine, a phospholipid displayed on the outer leaflet of apoptotic cells.

Ligation of the vitronectin receptor results in production of anti-inflammatory products (e.g. PGE_2 and TGFβ), which can overcome proinflammatory responses induced by potent stimuli such as LPS. Indeed, intracellular pathogens can induce and exploit this downregulation, which is presumably a host protective reaction, to promote their own survival.

Phagocytosis of apoptotic thymocyte by thymic macrophage

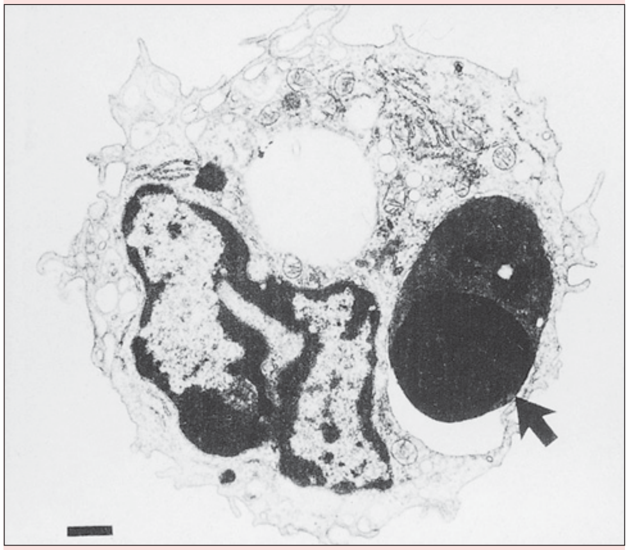

Fig. 9.15 Thymic macrophages phagocytose the large numbers of thymocytes that die during T cell development. The arrow indicates the nucleus of a phagocytosed thymocyte.

Inefficient clearance of apoptotic cells may also contribute to autoimmune disorders such as systemic lupus erythematosus (see Chapter 20), and may explain their association with genetic deficiencies of complement components.

Macrophages are more efficient than mature dendritic cells in the capture and digestion of apoptotic cells. Macrophages and dendritic cells may express different receptor profiles, but other differences could also contribute to a different fate for ingested dying cells.

Class A and related scavenger receptors

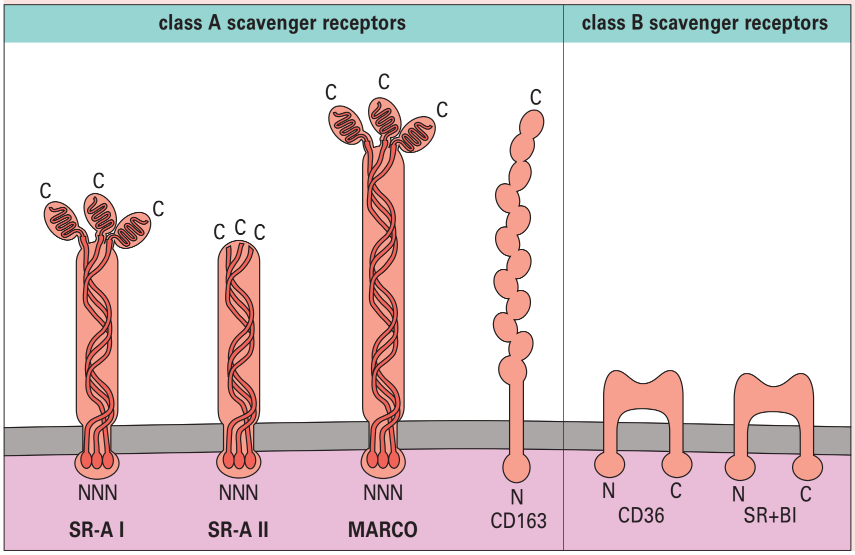

Fig. 9.16 Selected scavenger receptors are shown. Scavenger receptors of macrophages are responsible for the uptake of apoptotic cells, modified lipoproteins, and other polyanionic ligands (e.g. LPS and lipoteichoic acids [LTA]), as well as selected bacteria such as *Neisseria* spp. CD163 is involved in endocytosis of hemoglobin–haptoglobin complexes.

Clearance of dead cells

	apoptotic cells	necrotic cells
receptors	scavenger receptors phosphatidyl serine receptors	Fc receptors complement receptors
cytokines	TGFβ1	TNFα, IL-1β, IL-8

Fig. 9.17 Receptors involved in the uptake of apoptotic and necrotic cells, and the cytokines released following uptake.

Q. Give examples in the immune system where macrophages phagocytose apoptotic cells.

A. Thymocytes that fail the processes of positive and negative selection die by apoptosis and are phagocytosed by thymic macrophages (see Chapter 2). B cells that die within lymphoid follicles are taken up by tingible body macrophages (see Chapter 8).

The recognition mechanisms of normally senescent hematopoietic cells (e.g. erythrocytes, platelets) and of necrotic cells are still obscure.

Fig. 9.17 compares the differential recognition and cytokine production following the uptake of apoptotic and necrotic cells.

MACROPHAGES INTERNALIZE PATHOGENS USING A VARIETY OF SPECIFIC AND OPSONIC RECEPTORS

The mannose receptor is the best studied lectin-like receptor

Although macrophages can express a range of lectins specific for various sugar ligands, the MR is the best studied and may play a unique role in tissue homeostasis as well as host defense (Fig. 9.18). Endogenous ligands include lysosomal hydrolases and myeloperoxidase, and a range of pro- and eukaryotic organisms express mannose-rich structures, which may serve as ligands.

An N terminal cysteine-rich domain of the MR is a distinct lectin for sulfated glycoconjugates highly

Lectin receptors

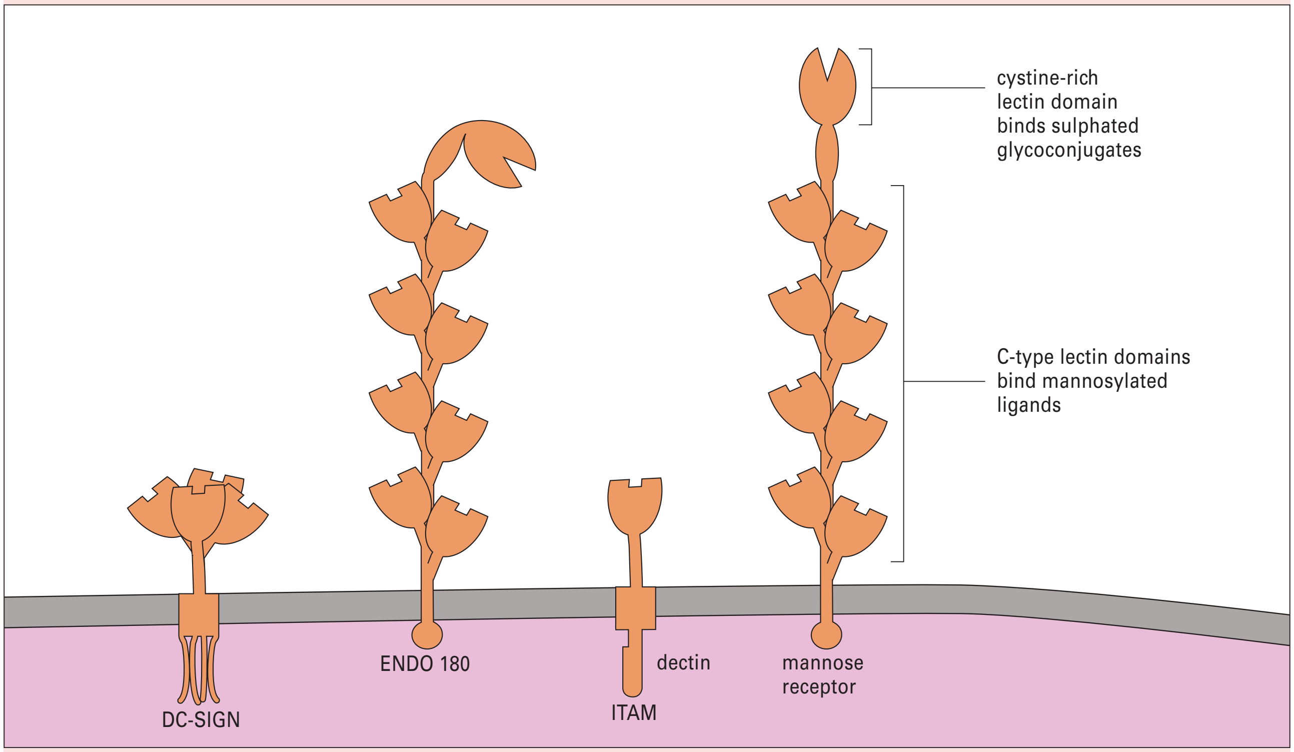

Fig. 9.18 The mannose receptor contains eight C-type lectin domains involved in binding to mannosylated carbohydrates and related glycoconjugates. It is related to a more broadly expressed endocytic receptor, ENDO 180. A distinct lectin domain located in the distal (C terminal) segment binds sulfated glycoconjugates. The β-glucan receptor (dectin-1) contains a single lectin domain and an intracellular immunoreceptor tyrosine-based activation motif (ITAM). DC-SIGN (another mannose-binding C-type lectin) has four associated lectin domains.

expressed in secondary lymphoid organs (marginal metallophils, subcapsular sinus macrophages, and follicular dendritic cells).

A novel antigen transport pathway has been postulated by which cell-associated or soluble MRs could deliver antigen to B and/or T cells, modulating their immune response.

The cysteine-rich domain also contributes to the clearance of hormones such as lutropin, which contains similar sulfated structures.

DC-SIGN, another mannose-binding C-type lectin, can be expressed on selected macrophages. It has been implicated in interactions between APCs and T cells and in microbial recognition (see Fig. 9.18).

Dectin-1 is a β-glucan receptor with a single lectin-like domain and an intracellular ITAM, responsible for the uptake of fungi and intracellular signaling (Fig. 9.19, see Fig. 9.18).

CD14 interacts with Toll-like receptors to initiate signal transduction

Monocytes and to a lesser extent tissue macrophages express a glycosyl phosphatidylinositol (GPI)-anchored molecule – CD14 – implicated in the cellular response to bacterial LPS. A plasma LPS-binding protein enhances LPS responses markedly; genetic ablation of CD14 in mice renders them highly resistant to septic shock, mainly mediated by TNFα release into the circulation. Toll-like receptors (TLRs, see Fig. 6.24) play a major role in LPS signaling in macrophages.

Zymosan particles phagocytosed by a macrophage

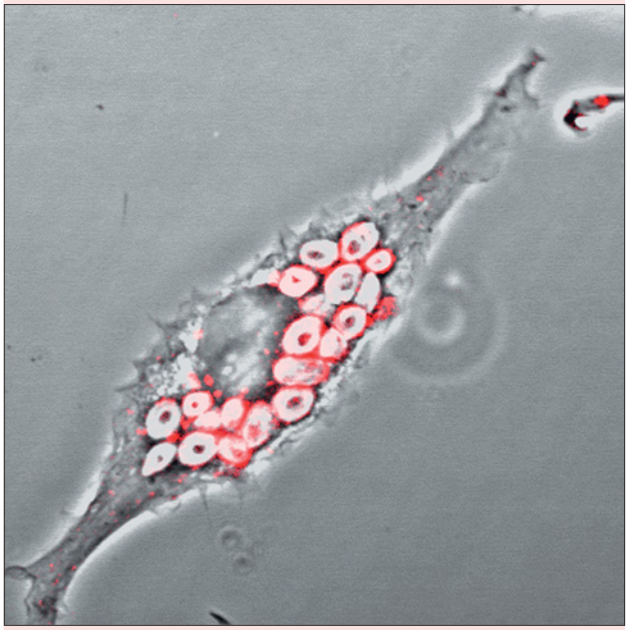

Fig. 9.19 The micrograph shows zymosan (yeast) particles phagocytosed by a macrophage, a process dependent on dectin-1. Truncation of the receptor cytoplasmic tail prevents phagocytosis.

Activation of macrophages by lipopolysaccharide and IL-1

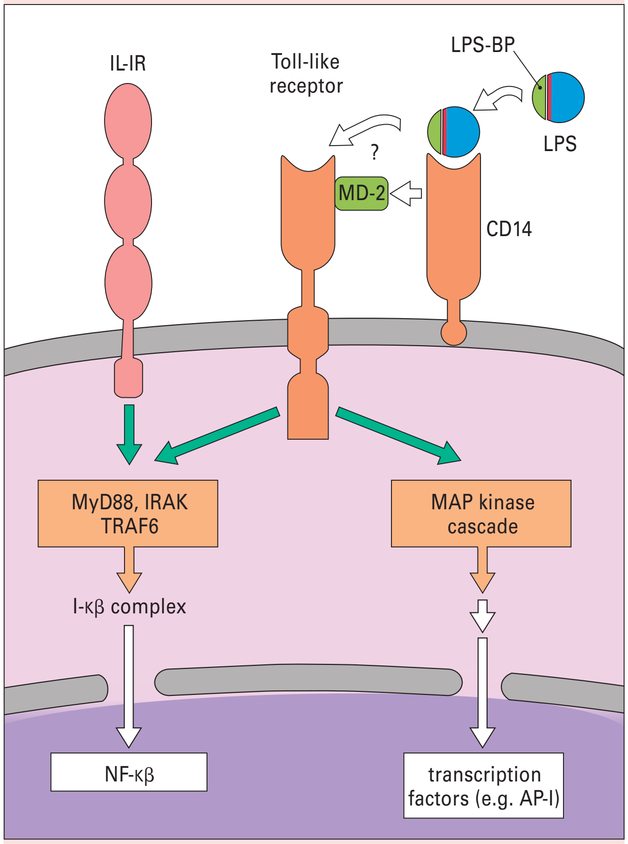

Fig. 9.20 LPS binds leukocyte CD14 (which is GPI-anchored to the membrane) through LPS binding protein (LPS-BP) and interacts with transmembrane TLRs by interaction with MD-2 to initiate signal transduction. Signaling pathways of the receptor share elements with the IL-1R pathway (e.g. IRAK – IL-1R-associated kinase). Cell activation proceeds via both the mitogen-activated protein (MAP) kinase pathways and the induction of NF-κB. (IL-1R, IL-1 receptor)

The signaling pathway for IL-1 and TLR shows many similarities (Fig. 9.20), resulting in nuclear factor kappa B (NF-κB) activation. Details of the NF-κB signaling pathway, which is activated following clustering of the TLRs, are shown in Fig. 9.21.

A distinct pathway leads to the induction of type 1 interferon (IFN) genes (*IFNα*, *IFNβ*), which can then act on type 1 interferon receptors in an autocrine pathway that induces many other genes (Fig. 9.22).

Q. What can you infer about the role of IFNs in immune defense against pathogens from their intracellular signaling pathways?

A. IFNs are involved in both antibacterial and antiviral defenses because they can activate many genes involved in inflammation.

Other receptors involved in LPS clearance, such as SR-A (see Fig. 9.16), may serve to downregulate responses induced via the CD14 pathway (see Fig. 9.20) and

The nuclear factor kappa B activation pathway

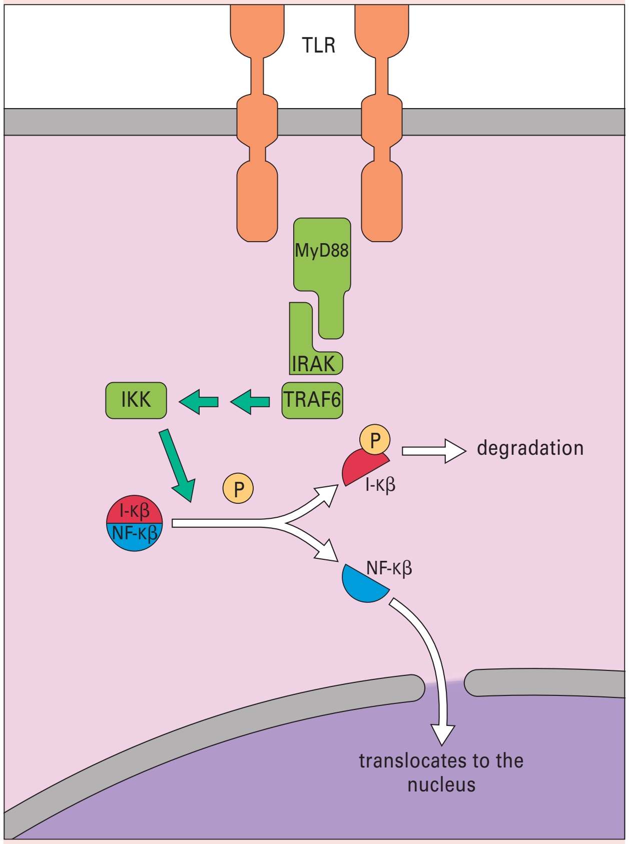

Fig. 9.21 Activation of NF-κB can be induced by ligation of a number of cell surface receptors including the Toll-like receptors (TLRs) illustrated. Apart from MyD88, a small number of other adapters transmit signals from the TLRs to cause NF-κB activation. An intracellular signaling cascade leads to the activation of inhibitor of kappa B kinase (IKK). NF-κB is normally retained in the cytosol by interaction with an inhibitor (I-κB). Phosphorylation of I-κB by IKK leads to dissociation and degradation of the inhibitor and NF-κB translocates to the nucleus to act as a transcription factor that activates a set of genes involved in inflammatory reactions.

therefore limit the systemic release of TNFα and resultant septic shock.

Monocytes and macrophages express a range of complement receptors

Monocytes and macrophages express a range of heterodimeric receptors for C3 cleavage products (CR1, CR3, CR4) and interact with other components of the classical, alternative, or lectin-induced pathways of complement activation (see Fig. 4.12).

CR3 contributes to myelomonocytic cell recruitment by adhesion to induced endothelial ligands, including intercellular adhesion mlecule-1 (ICAM-1), as shown in human and murine genetic deficiency syndromes.

CR3 is a promiscuous receptor, with ligands other than iC3b, including fibrinogen. Its role in regulated phagocytosis has been well studied, and the mechanism of CR3-mediated ingestion differs strikingly from that mediated by Fc receptors (see below). Inflammatory and secretory responses also differ after complement and antibody-dependent uptake.

Complement receptors:

- contribute to apoptotic cell clearance as well as to host defense to infection; but also
- serve as a 'safe' entry receptor for organisms such as mycobacteria by direct interaction with microbial ligands or after opsonization.

Fc receptor-mediated uptake proceeds by a zipper-like process

Fc receptors are opsonic receptors for a variety of immunoglobulin subclasses, especially IgG. The receptors are themselves immunoglobulin superfamily members, with two or three domains (see Fig. 3.20).

Fc receptors contain immunoreceptor tyrosine-based activation motifs (ITAMs) or immunoreceptor tyrosine-based inhibitory motifs (ITIMs), which, by interaction with other membrane molecules (e.g. γ chain) and cytoplasmic kinases (e.g. Syk) and/or phosphatases, regulate complex signaling pathways.

Apart from activation of effector responses (phagocytosis, endocytosis, antibody-dependent cytotoxicity), different Fc receptors can also downregulate inflammatory cascades and may provide a link between innate and adaptive immunity.

The mechanism for ingestion of antibody-coated particles is distinct from that mediated by CR3 (Fig. 9.23). Fc receptor-mediated uptake proceeds by a zipper-like process where sequential attachment between receptors and ligands guides pseudopod flow around the circumference of the particle. CR3 contact sites are discontinuous for complement-coated particles which 'sink' into the macrophage cytoplasm. Small GTPases play distinct roles in actin cytoskeleton engagement by each receptor-mediated process.

Macrophages express several receptors with ITIMs

Recent studies have discovered several new receptors with ITIMs expressed on macrophages, which deactivate their effector functions (e.g. signal regulatory proteins [SIRPs]) or resemble NK inhibitory receptors in their ability to ligate MHC class I molecules (see Chapter 10).

Transmembrane glycoproteins such as DAP-12 promote the surface expression of C-type lectins analogous to NK cell receptors. These may play an important regulatory function in monocytes and macrophages, for example in deactivation of cellular cytotoxicity.

Macrophages express a range of G protein-coupled receptors

Macrophages express a range of G protein-coupled receptors in addition to multispan receptors for chemokines. The F4/80 antigen (see Fig. 9.1) contains:

- a seven-transmembrane spanning portion homologous to a family of peptide receptors (e.g. vasoactive intestinal peptide, VIP); and

Activation of the *IFNβ* gene by Toll-like receptors

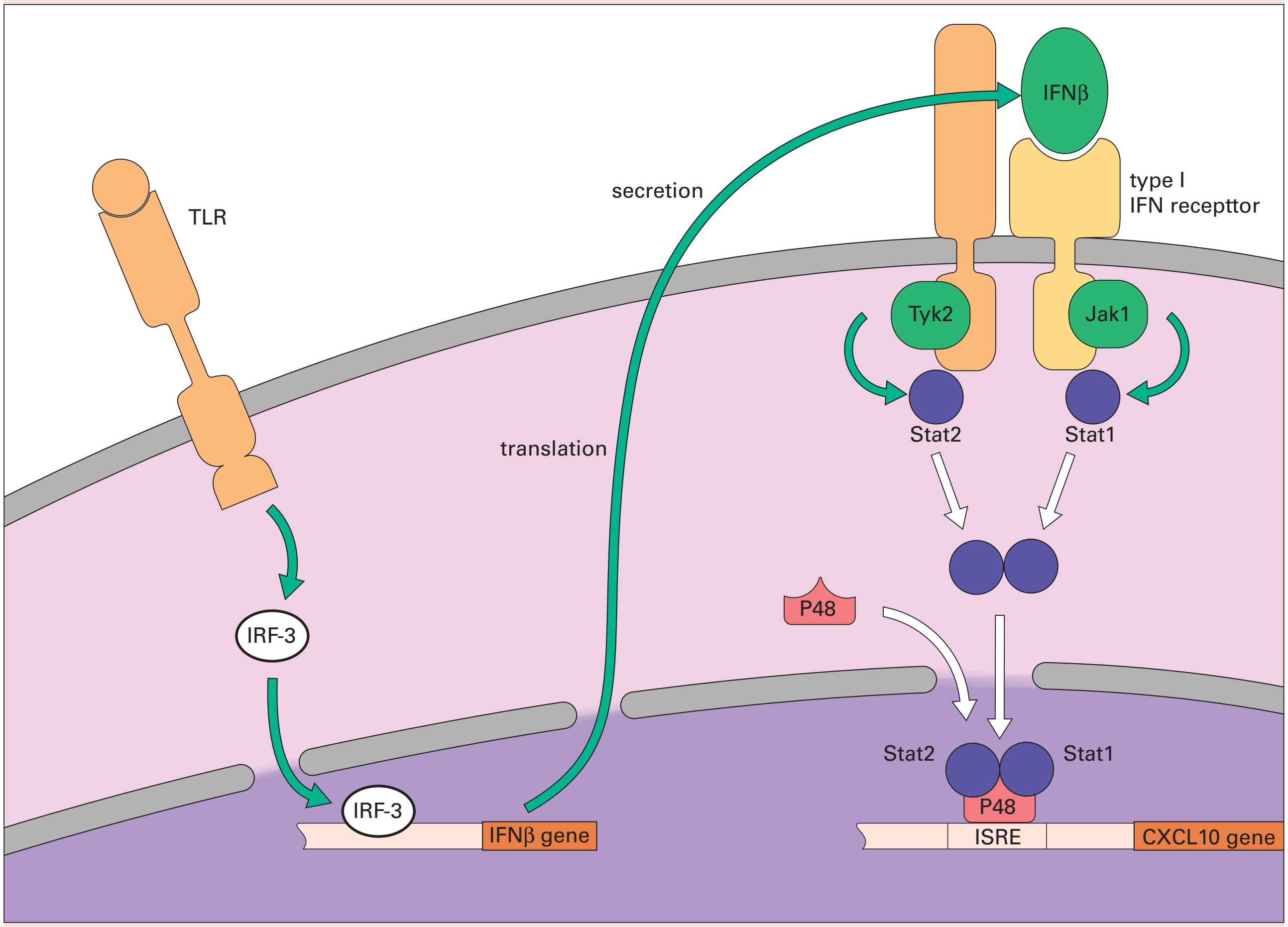

Fig. 9.22 TLRs can activate the *IFNβ* gene via a pathway that is independent of MyD88, but involves activation of the transcription factor interferon regulatory factor-3 (IRF-3). Released IFNβ acts on type 1 IFN receptors causing the aggregation of the two subunits of the receptor. This leads to activation and phosphorylation of two Jak kinases, Jak1 and Tyk2. These two transcription factors form a complex with a DNA-binding protein, P48. The complex moves to the nucleus and induces transcription of genes bearing an IFN-specific response element (ISRE), such as the chemokine CXCL10.

- a large extracellular domain consisting of multiple epidermal growth factor (EGF) domains.

The F4/80 molecule has been implicated in the induction of peripheral tolerance.

Non-opsonic phagocytosis can lead to intracellular infection

In the absence of opsonins, macrophages use multiple receptors to recognize and engulf a range of microorganisms, parasites, and viruses directly. These include:

- CR3;
- SR-As;
- MR;
- the β-glucan receptor; and
- CD14.

SR-As use selected polyanionic ligands including LPS (Gram-negative) and lipoteichoic acid (Gram-positive) for bacterial recognition.

MR ligands are found on mycobacteria, HIV gp120, *Pneumocystis carinii*, *Klebsiella* spp., as well as yeasts (*Candida albicans*).

CR3 ligands include phosphatidylinositol mannoside (PIM), and a direct saccharide binding site on CR3 has been implicated in opsonin-independent recognition.

Different receptors often collaborate (e.g. CR3 and MR in the uptake of *Leishmania* promastigotes).

Non-opsonic uptake or invasion results in a range of survival strategies by intracellular pathogens enabling them to:

- avoid phagosome maturation;
- inhibit fusion with lysosomes and acidification (mycobacteria); or conversely
- promote invasion by rapid recruitment of lysosomes (*C. albicans*, *Trypanosoma cruzi*).

Legionella pneumophila induces a coiling phagocytosis, and recruits various intracellular organelles (endoplasmic

Zipper model of phagocytosis

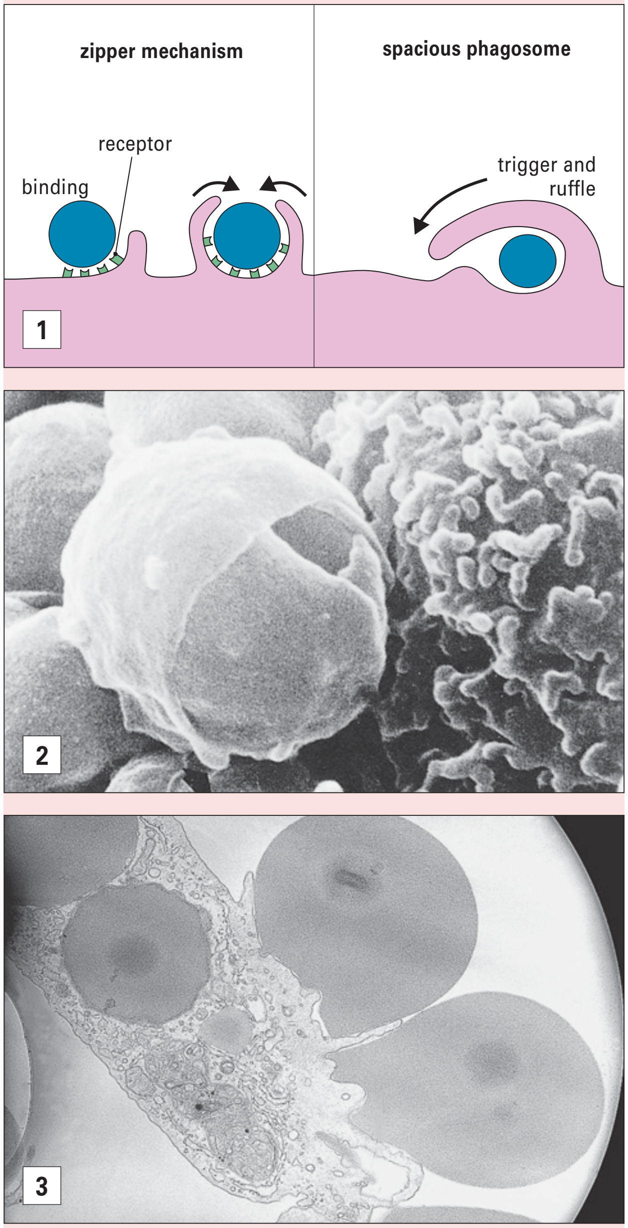

Fig. 9.23 (**1**) During phagocytosis, receptor–ligand interactions guide the extension of tightly apposed pseudopods around the particle’s total circumference until a fusion of the plasma membrane occurs at the tip. This is known as the zipper mechanism. Alternative ‘trigger’ mechanisms, in which spacious phagosomes result from flipping over of ruffles back onto the plasma membrane, have also been described. The cytoskeleton of phagocytes plays a key role in engulfment, during which there is extensive remodeling of actin filaments. Some microorganisms and intracellular parasites induce novel mechanisms to recruit cell membranes during entry into phagocytes. (**2 and 3**) Electron micrographs of ingestion of antibody (IgG)-coated sheep erythrocytes by peritoneal macrophage by the zipper mechanism. (Scanning electron micrograph (**2**), courtesy of Dr GG MacPherson. Transmission electron micrograph (**3**), courtesy of Dr SC Silverstein)

reticulum, mitochondria), whereas salmonellae induce the formation of spacious phagosomes.

Once within vacuoles:

- the organism can replicate in secondary lysosomes (*Leishmania* spp.); or
- escape into the cytosol by membrane dissolution (*Listeria monocytogenes*).

IFNγ can overcome many of these evasion strategies, resulting in killing or stasis of the organisms by mechanisms described below.

Opsonization by antibody and/or complement can alter the fate of the organism by extracellular lysis or by targeting it to lysosomes.

Some organisms induce apoptosis in macrophages (e.g. *Salmonella* spp.); others spread between cells by fusion (e.g. HIV) or by intercellular infection (e.g. *Listeria* spp.).

The ability of macrophages to present processed peptide antigens derived from intracellular pathogens to T cells is poorly defined. It is possible that they collaborate with dendritic cells by releasing breakdown products, so activating naive T cells. Once MHC class II molecules are induced by IFNγ, exogenous antigens are presented by the MHC class II pathway (see Fig. 7.11). Endogenous antigens are presented by the MHC class I pathway (see Fig. 7.9), so infected macrophages can become targets for CD8$^+$ cytotoxic T cells.

Q. How are lipid antigens such as mycolic acids of mycobacteria presented to T cells?

A. CD1 molecules present these antigens, which are inducible on selected macrophage populations.

Macrophages contain intracellular receptors for pathogens

Recently a number of intracellular proteins have been discovered in macrophages and other cells, that are able to sense microbial wall degradation products. These proteins contain leucine-rich repeats, pyrin, caspase-associated recruitment domains (CARD), and **nucleotide oligomerization domains (NOD)**. Ligation induces expression of proinflammatory cytokines (e.g. IL-1).

As illustrated for NOD-1 (Fig. 9.24), signal transduction involves the formation of protein–protein complexes that are involved in IL-1 processing (**inflammasomes**). Mutations in the pyrin domain have been implicated in rare hyperinflammatory syndromes such as familial Mediterranean fever. IL-1 receptor antagonists can be remarkably effective as therapy in related conditions.

The *NOD-2* gene is frequently mutated in people with Crohn’s disease, a chronic granulomatous inflammatory bowel disease (Fig. 9.25). It is expressed by monocytes and macrophages and recognizes cytosolic muramyl dipeptides, resulting from the degradation of both Gram-negative and Gram-positive organisms.

NOD-2 deficiency leads to increased susceptibility to bacterial infection via the oral route, and may be required for expression of antimicrobial peptides. Other genes have also been implicated in susceptibility to inflammatory bowel disease.

NOD-1 signal transduction

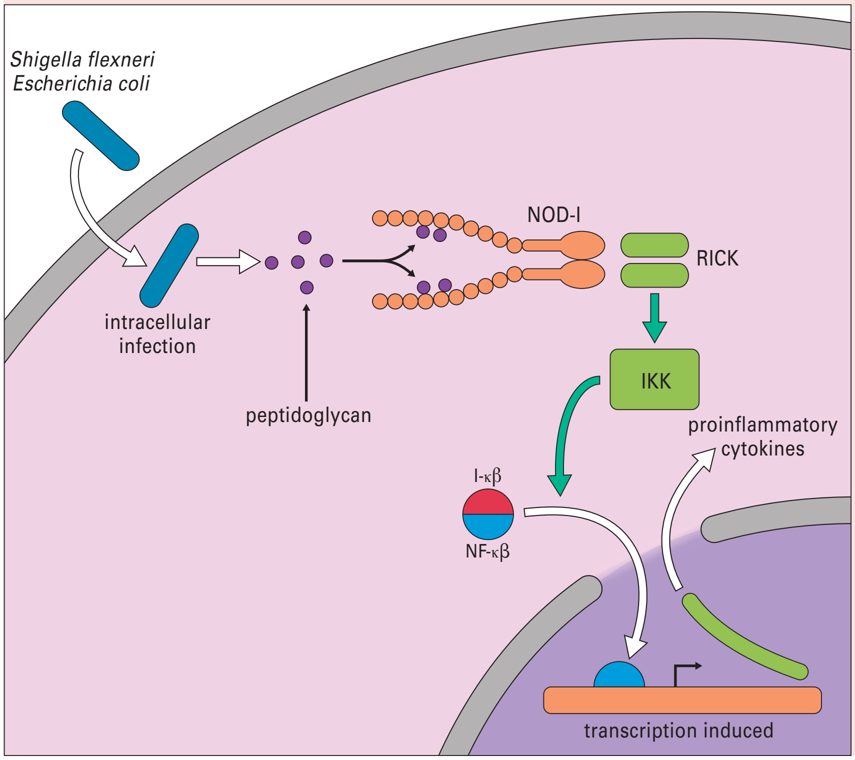

Fig. 9.24 NOD-1 is an intracellular pattern recognition molecule. Peptidoglycan from, for example, *Shigella flexneri* and some strains of *Escherichia coli* binds to the protein. A domain (RICK) then activates IKK (see Fig. 9.21) leading to the transcription of proteins dependent on NF-κB.

Crohn's disease

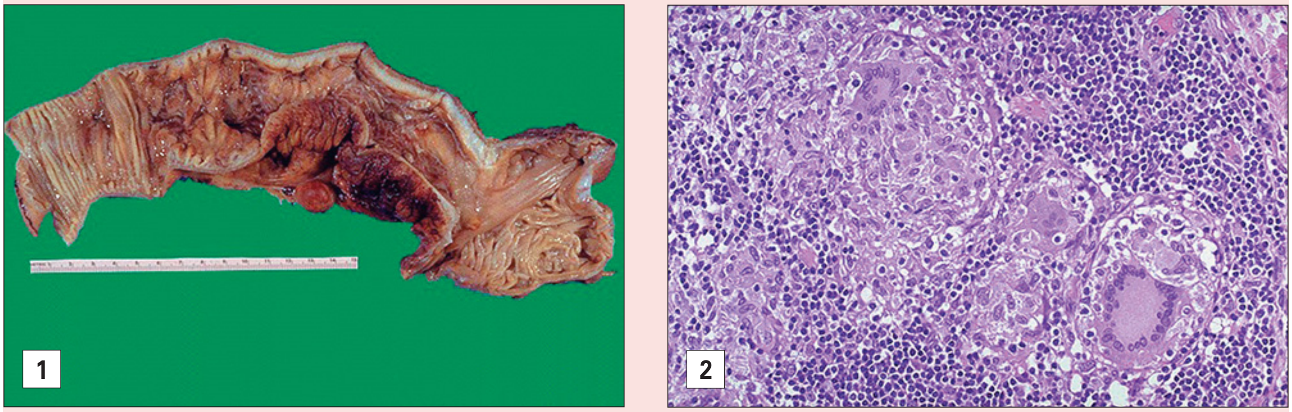

Fig. 9.25 A section of the gut wall from a patient with Crohn's disease, showing the intense inflammation of the tissue (**1**). The histological section shows a granulomatous reaction with lymphocyte and macrophage infiltration and the formation of giant cells (**2**).

ACTIVATED MACROPHAGES SECRETE A VARIETY OF MOLECULES

Following encounters with microorganisms and antigens, resident macrophages are able to enhance their transcription and translation of a wide range of gene products, including secreted molecules, which often act locally close to the cell surface.

Secretory products include:

- eicosanoids;
- cytokines;
- complement proteins; and
- enzymes, especially lysozyme (Fig. 9.26).

After priming by IFNγ, macrophages triggered by surface-acting stimuli such as LPS release increased levels of various products, which promote an inflammatory reaction, recruit other cells, regulate their activities, and induce adhesion.

Macrophages also release inhibitory molecules (e.g. TGFβ, PGE_2, IL-10), which suppress inflammatory and immune responses, including T cell proliferation.

Secretory products of macrophages

category	example	function
low molecular weight metabolites	reactive oxygen intermediates	killing, inflammation
	reactive nitrogen intermediates	killing, inflammation
	eicosanoids – prostaglandins, leukotrienes	regulation of inflammation
	platelet-activating factor (PAF)	clotting
cytokines	IL-1β, TNFα, IL-6	local and systemic inflammation
	IFNα/IFNβ	antiviral, innate immunity, immunomodulation
	IL-10	deactivation of MØ, B cell activation
	IL-12, IL-18	IFNγ production by NK and T cells
	TGFβ	repair, modulation, inflammation
	CCL2, CCL3, CCL4, CCL5, CXCL8	chemokines
adhesion molecules	fibronectin	opsonization, matrix adhesion,
	thrombospondin	phagocytosis of apoptotic cells
complement	C3, all others	local opsonization
procoagulant	tissue factor	clotting cascade
enzymes	lysozyme	Gram-positive bacterial lysis
	urokinase (plasminogen activator)	fibrinolysis
	collagenase	matrix catabolism
	elastase	matrix catabolism

Fig. 9.26 Macrophages produce a wide range of secreted molecules.

IFNγ, IL-1, and TNFα are potent regulators of leukocyte and other cellular activities.

Q. Describe four roles of IFNγ in the development of the immune response, each modulating a different group of cells.
A. IFNγ induces MHC molecules and co-stimulatory molecules on APCs, and induces adhesion molecules and chemokines on endothelium. It also activates macrophage microbicidal activity, modulates class switching by B cells, and regulates the balance between TH1 and TH2 populations of T cells.

IL-6 and IL-1 act as circulating mediators of the acute phase response. Distant targets of these macrophage-derived cytokines include thermoregulatory centers in the central nervous system, muscle and fat stores, and the neuroendocrine system.

Cytotoxic products and powerful neutral proteinases such as elastase, collagenase, and urokinase (generating plasmin) (Fig. 9.27) are able to induce tissue injury and contribute to destructive chronic inflammation in joints and lung.

Monocyte-derived procoagulant/tissue factor can also induce vascular occlusion and tissue damage.

Macrophages themselves express receptors for cytokines, as well as producing molecules that are cleaved by metalloproteinases and shed into the circulation. By contrast, soluble receptors and receptor antagonists (eg. IL-1Ra, soluble TNFR) are potential inhibitors of receptor–ligand interactions.

mRNA microarray methods to analyze gene expression generate complex profiles of selective induction and inhibition of large numbers of genes, some highly restricted to macrophages, others shared with different cell types. Patterns of macrophage mRNA and protein expression now provide insights into intrinsic and extrinsic regulation of macrophage differentiation and activation.

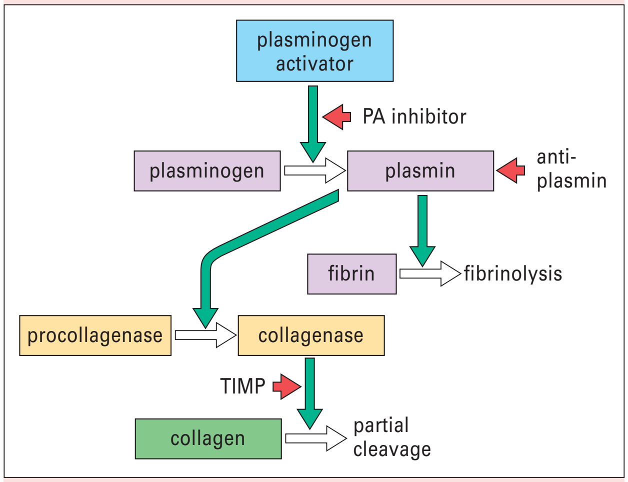
Actions of macrophage neutral proteinases

Fig. 9.27 Macrophage-derived neutral proteinases interact with plasma and tissue enzymes and their inhibitors (which can also be generated by macrophages) to regulate fibrinolysis and matrix catabolism. (PA, plasminogen activator; TIMP, tissue inhibitor of matrix metalloproteinase)

Resting macrophages have limited antimicrobial killing activity

Following phagocytosis, pathogens are subjected to a variety of killing mechanisms, which are generally more diverse and more effective in activated macrophages.

Following lysosome fusion, there is a transient rise in the pH of the phagolysosome, followed by a fall in pH, which occurs within 10–15 minutes. The acidification of the phagolysosome may, by itself, contribute to the killing of some organisms, but the additional killing mechanisms, described below, may be more effective at specific pH values. For example:

- the production of reactive oxygen intermediates (ROIs) in activated macrophages occurs immediately after internalization;
- cationic proteins are most active during the early alkaline phase; and
- many digestive lysosomal enzymes are more active in the later acidic phase (Fig. 9.28).

Lysozyme acts directly on the bacterial cell wall proteoglycans

The resident macrophage has a relatively limited killing capacity, which may be sufficient to restrict growth of many, if not most, organisms, including viruses. Lysozyme acts directly on the bacterial cell wall proteoglycans, present especially in the exposed cell wall of Gram-positive bacteria (Fig. 9.29).

The cell walls of Gram-negative bacteria may also become exposed to lysozyme if they have been damaged by complement membrane attack complexes.

Lysozyme is constitutively produced by macrophages.

Defensins also contribute to macrophage antibacterial activities

A group of highly cationic proteins and polypeptides called defensins also contribute to macrophage antibacterial activities.

Mechanism involved in bacterial killing

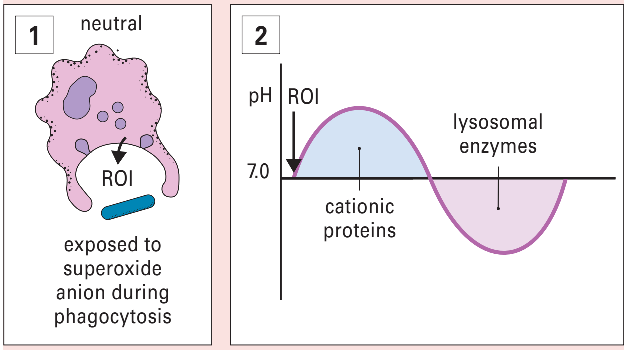

Fig. 9.28 During phagocytosis there is immediate exposure to reactive oxygen intermediates (ROIs) (**1**). This leads to a transient increase in pH, when cationic proteins may be most effective (**2**). Subsequently the pH falls as H^+ ions are pumped into the phagolysosome and neutral proteinases and lysosomal enzymes with low pH optima become effective. Lactoferrin acts by chelating free iron and can do so at alkaline or acidic pH.

Action of lysozyme on the cell wall of *Staphylococcus aureus*

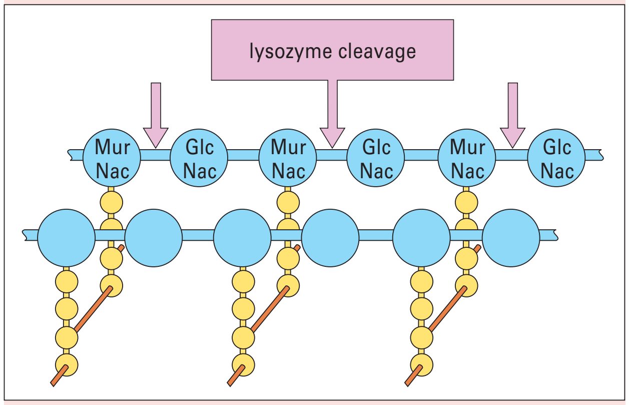

Fig. 9.29 The structure of the cell wall of Gram-positive bacteria such as *Staphylococcus aureus* includes a backbone of N-acetylglucosamine (GlcNac) alternating with N-acetylmuramic acid (MurNac) cross-linked by amino acid side chains (yellow) and bridges of five glycine residues (orange). Lysozyme splits the molecules at the places indicated.

The defensins:

- are small peptides (30–33 amino acids) found in some macrophages of many species and specifically in human neutrophils, where they comprise up to 50% of the granule proteins;
- form ion-permeable channels in lipid bilayers and probably act before acidification of the phagolysosome;
- are able to kill a range of pathogens, including bacteria (*S. aureus*, *Pseudomonas aeruginosa*, *E. coli*), fungi (*Cryptococcus neoformans*), and enveloped viruses (herpes simplex).

Q. What other molecule can generate ion-permeable channels in lipid bilayers?

A. The final component of the lytic complement pathway, C9 (see Chapter 4).

In resident macrophages, endocytosis, lysosome delivery and fusion, and acidification favor destruction of phagocytosed pathogens, but may facilitate escape of an organism from vacuole to cytosol.

The **N-ramp multispanning glycoprotein** found in macrophage phagosomes contributes to cellular and ultimately host resistance or susceptibility to a range of intracellular pathogens (mycobacteria, *Leishmania* spp., *Salmonella* spp.) by influencing iron transport into or possibly out of the vacuole.

ACTIVATED MACROPHAGES PRODUCE REACTIVE OXYGEN AND NITROGEN INTERMEDIATES

Elicited (i.e. recruited by a non-immune inflammatory stimulus) and immunologically activated macrophages produce higher levels of lysozyme than resting macrophages,

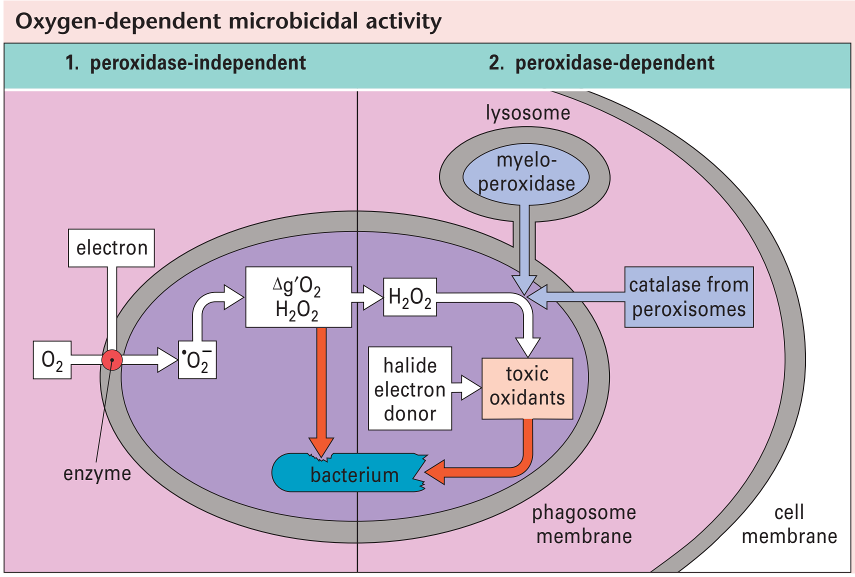
Oxygen-dependent microbicidal activity
1. peroxidase-independent
2. peroxidase-dependent
lysosome
myelo-peroxidase
electron
$\Delta g' O_2$
H_2O_2
H_2O_2
catalase from peroxisomes
O_2
$^\bullet O_2^-$
halide electron donor
toxic oxidants
enzyme
bacterium
phagosome membrane
cell membrane

Fig. 9.30 (**1**) An enzyme (NADPH oxidase, see p. 307) in the phagosome membrane reduces oxygen to the superoxide anion ($^\bullet O_2^-$). This can give rise to hydroxyl radicals ($^\bullet OH$), singlet oxygen ($\Delta g' O_2$), and hydrogen peroxide (H_2O_2), all of which are potentially toxic. Lysosome fusion is not required for these parts of the pathway, and the reaction takes place spontaneously following formation of the phagosome. (**2**) If lysosome fusion occurs, myeloperoxidase (or, under some circumstances, catalase from peroxisomes) acts on peroxides in the presence of halides (preferably iodide). Then additional toxic oxidants, such as hypohalite (HIO, HClO), are generated.

as well as proinflammatory cytokines, chemokines, growth factors, and proteases.

One major difference between resting and activated macrophages is the ability to generate hydrogen peroxide (H_2O_2), and other metabolites generated by the respiratory burst (Fig. 9.30). The phagocyte oxidase (Phox) membrane and cytosolic proteins found in neutrophils and activated macrophages assemble to form an NADPH oxidase complex, which contains a novel cytochrome.

In the presence of myeloperoxidase, released by myelomonocytic cells, the Klebanoff reaction generates even higher levels of reactive oxygen metabolites. Patients with chronic granulomatous disease lack essential oxidase components and suffer from repeated infections.

Macrophages in mouse as well as human can be activated by IFNγ to express high levels of **inducible nitric oxide synthase (i-NOS)**, which catalyzes the production of nitric oxide (NO) from arginine (Fig. 9.31). Release of NO by macrophages affects neighboring vessels and leukocytes, and contributes to the killing of intracellular pathogens such as *Leishmania* spp. either directly or by production of peroxynitrites.

Failure of macrophage activation in AIDS contributes to opportunistic pathogen infections and persistence of HIV, as well as reactivation of latent tuberculosis (e.g. by anti-TNFα therapy).

Rare inborn errors in humans for IL-12 and IFN receptors have confirmed experimental gene ablation studies in the mouse, which render the host susceptible to infection by intracellular bacteria.

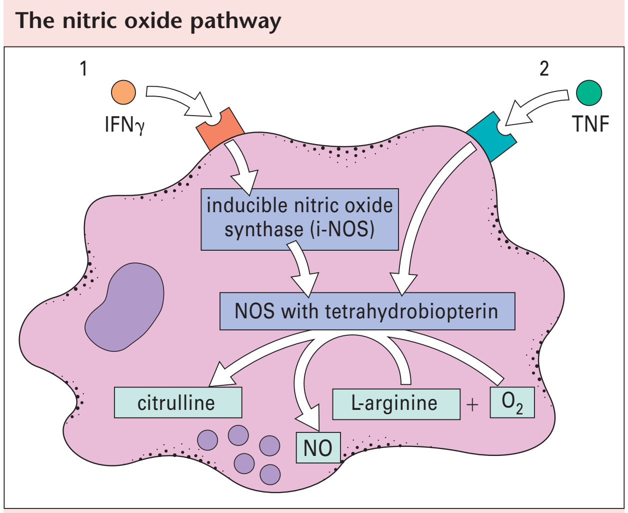
The nitric oxide pathway
1
2
IFNγ
TNF
inducible nitric oxide synthase (i-NOS)
NOS with tetrahydrobiopterin
citrulline
NO
L-arginine
+
O_2

Fig. 9.31 The inducible nitric oxide synthase (i-NOS) combines oxygen with guanidino nitrogen of L-arginine to give nitric oxide (NO), which is toxic for bacteria and tumor cells. Toxicity may be increased by interactions with products of the oxygen reduction pathway, leading to the formation of peroxynitrites. Tetrahydrobiopterin is needed as a cofactor. In murine macrophages IFNγ activates the pathway (**1**), which is then optimally triggered by TNF (**2**). Triggering release of NO from human macrophages is more complex and usually involves cross-linking of membrane CD23. Human macrophages can sometimes express i-NOS but they contain little tetrahydrobiopterin, and maximal NO release may require interaction with other cell types.

Recently recruited cells are particularly effective in microbial killing

Macrophages vary considerably in their antimicrobial activity. Monocytes produce more reactive oxygen metabolites and myeloperoxidase than differentiated tissue macrophages.

It is the recently recruited and further activated cell that is primarily responsible for host defense. In-situ analysis shows that granuloma macrophages are the major source of newly produced proteins such as lysozyme, TNFα, and IL-1. The latter cytokines are more tightly regulated than lysozyme, and further exposure to microbial components such as LPS is needed to induce the majority of macrophages to express these mediators.

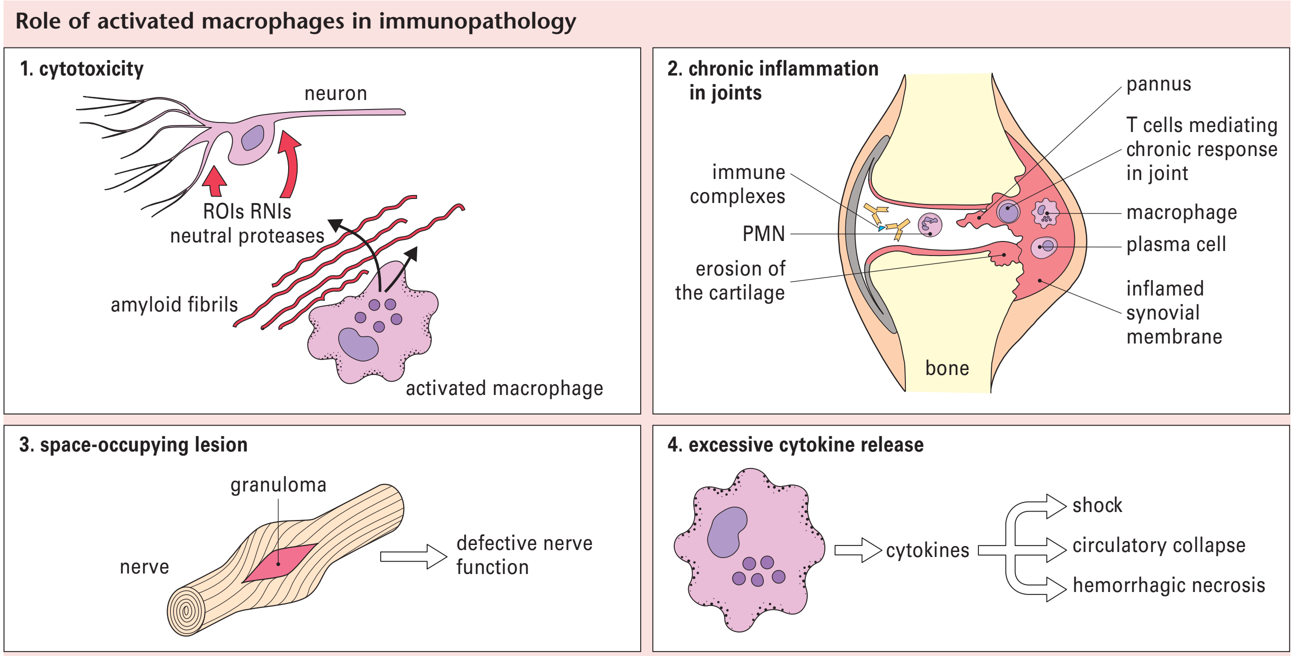
Role of activated macrophages in immunopathology
1. cytotoxicity
neuron
ROIs RNIs
neutral proteases
amyloid fibrils
activated macrophage
2. chronic inflammation in joints
immune complexes
PMN
erosion of the cartilage
bone
pannus
T cells mediating chronic response in joint
macrophage
plasma cell
inflamed synovial membrane
3. space-occupying lesion
granuloma
nerve
defective nerve function
4. excessive cytokine release
cytokines
shock
circulatory collapse
hemorrhagic necrosis

Fig. 9.32 (**1**) Activated microglia may recognize, but be unable to degrade, extracellular β-amyloid fibrils and release cytotoxic products that affect neuronal function and survival. (**2**) The response may be directed towards autoantigens (or perhaps to unidentified cryptic infections or commensal organisms), leading to chronic inflammation, as seen in rheumatoid arthritis. Non-internalizable immune complexes may perpetuate inflammation and antibody/complement-dependent cytotoxicity. (**3**) A granuloma may cause a bulky space-occupying lesion, impairing the function of sensitive tissues such as brain, retina, and nerve. (**4**) Excessive release of cytokines can lead to several tissue-damaging syndromes, especially toxic shock syndromes with TNFα.

Once effector mechanisms are activated, local surface reactions also influence the release of potentially injurious macrophage molecules. If ligands are presented on a non-internalizable surface (‘frustrated phagocytosis’), secretory products can be released into the extracellular environment.

Activated macrophages contribute considerably to tissue damage in autoimmune and chronic inflammatory diseases in joints, lung, and the nervous system (Fig. 9.32).

MACROPHAGES INITIATE, PROMOTE, PREVENT, SUPPRESS, AND TERMINATE IMMUNE RESPONSES

The macrophage is able to initiate, promote, prevent, suppress, or terminate an immune response. Its actions may seem antigen non-specific, though antigen dependent.

Apart from a major role in host defense, macrophages:

- are participants in autoimmunity (failure to prevent immunity) and peripheral tolerance;
- may contribute to T cell downregulation in placenta (a macrophage-rich organ) and at immunologically privileged sites such as the anterior chamber of the eye;
- can nurse or kill neighboring cells and partners, and we need to learn more about how they discriminate friend from foe, and danger from safety.

The host is not viable without macrophages, which perform vital functions within and beyond the immune system.

FURTHER READING

Reviews

Aderem A, Underhill DM. Mechanisms of phagocytosis in macrophages. Annu Rev Immunol 1999;17:593–623.

Akira S, Takeda K. Toll-like receptor signaling. Nat Rev Immunol 2004;4:499–511.

Brown GD, Gordon S. Fungal β-glucans and mammalian immunity. Immunity 2003;19:311–315.

Desjardins M. ER-mediated phagocytosis: a new membrane for new functions. Nat Rev Immunol 2003;3:280–291.

Ezekowitz RAB, Hoffmann JA, eds. Innate immunity. New Jersey: Humana Press; 2003.

Geijtenbeek TBH, Engering A, Van Kooyk Y. DC-SIGN, a C-type lectin on dendritic cells that unveils many aspects of dendritic cell biology. J Leuk Biol 2002;71:921–931.

Girardin SE, Philpott DJ. The role of peptidoglycan recognition in innate immunity. Eur J Immunol 2004;34:1777–1782.

Gordon S. Macrophages and the immune response In: Paul W, ed. Fundamental immunology, 5th ed. Philadelphia: Lippincott Raven; 2003:481–495.

Gordon S. Alternative activation of macrophages. Nat Rev Immunol 2003;3:23–35.

Gordon S. Pattern recognition receptors: doubling up for the innate immune response. Cell 111:927–930.

Hull KM, Shoham N, Chae JJ, et al. The expanding spectrum of systemic autoinflammatory disorders and their rheumatic manifestations. Curr Opin Rheumatol 2003;15:61–69.

Kaufmann SHE, Medzhitov R, Gordon S, eds. The innate immune response to infection. Washington, DC: ASM Press; 2004.

Kuijpers TW, Roos D. Neutrophils: the power within. In: Kaufmann SHE, Medzhitov R, Gordon S, eds. The innate immune response to infection. Washington, DC: ASM Press; 2004:47–70.

Luster AD. Chemokines linking innate and adaptive immunity. Curr Opin Immunol 2002;14:129–135.

O'Neill LAJ. TLRs: Professor Mechnikov, sit on your hat. Trends Immunol 2004;25:687–693.

Peiser L, Mukhopadhyay S, Gordon S. Scavenger receptors in innate immunity. Curr Opin Immunol 2002;14:123–128.

Portnoy DA, Auerbuch V, Glomski IJ. The cell biology of *Listeria monocytogenes* infection. J Cell Biol 2002;158:409–414.

Puel A, Picard C, Cheng-Lung K, et al. Inherited disorders of NF-kappaB-mediated immunity in man. Curr Opin Immunol 2004;16:34–41.

Rabinovitch M. Professional and non-professional phagocytes: an introduction. Trends Cell Biol 1995;5:85–87.

Randolph GJ. Dendritic cell migration to lymph nodes: cytokines, chemokines and lipid mediators. Semin Immunol 2001;13:267–274.

Russell DG. *Mycobacterium tuberculosis*: here today and here tomorrow. Nat Rev Mol Cell Biol 2001;2:569–577.

Savill J, Dransfield I, Gregory C, Haslett C. A blast from the past: clearance of apoptotic cells regulates immune responses. Nat Rev Immunol 2002;2:965–975.

Taylor PR, Martinez-Pomares L, Stacey M, et al. Macrophage receptors and immune recognition. Annu Rev Immunol 2005;23:901–944.

Tschopp J, Martinon F, Burns K. NALPs: a novel protein family involved in inflammation. Nat Rev Mol Cell Biol 4:95–104.

Selected papers

Brown GD, Herre J, Williams DL, et al. Dectin-1 mediates the biological effects of β-glucans. J Exp Med 2003;197:1119–1124.

Dalton DK, Pitts-Meek S, Keshav S, et al. Multiple defects of immune cell function in mice with disrupted interferonγ genes. Science 1993;259:1739–1742.

Geissmann F, Jung S, Littman DR. Blood monocytes consist of two principal subsets with distinct migratory properties. Immunity 2003;19:71–82.

Kindler Y, Sappino AP, Grau GE, et al. The inducing role of TNF in the development of bactericidal granulomas during BCG infection. Cell 1981;56:731–740.

Lin H-H, Faunce DE, Stacey M, et al. The macrophage F4/80 receptor is required for the induction of antigen-specific efferent regulatory T cells in peripheral tolerance. J Exp Med 2005;201:1615–1625.

Medzhitov R, Preston-Hurlburt P, Janeway CA Jr. A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. Nature 1997;388:394–397.

Poltorak A, He X, Smirnova I, et al. Defective LPS signaling in C3H/HeJ and C57 Bl/10 ScCr mice: mutations in the *TLR4* gene. Science 1998;282:2085–2088.

Critical thinking: The role of macrophages in toxic shock syndrome (see p. 495 for explanations)

In an experimental model of septic shock, mice are infected systemically with bacille Calmette–Guérin (BCG), a non-lethal vaccine strain of mycobacteria. After 12 days, the mice are challenged intraperitoneally with graded doses of lipopolysaccharide (LPS). Blood samples are taken at 2 hours and the clinical condition of the mice is monitored for up to 24 hours. Experiments are terminated earlier if mice show severe signs of distress.

1 What cytokines would you measure in the 2-hour serum sample?
2 What clinical signs would be indicative of incipient septic shock?
3 What mechanisms contribute to septic shock?
4 What outcome would you expect if the following knockout mouse strains were used instead of wild-type controls: CD14, scavenger receptor class A (SR-A), IFNγ?
5 Interpret your results.
6 Suggest further experiments.
7 What is the clinical significance of this experiment?

The distinction between inflammatory responses and house-keeping of apoptotic cells (see p. 495 for explanations)

Mouse peritoneal macrophages are fed a meal of apoptotic cells in culture, and then challenged with an intracellular pathogen, *Trypanosoma cruzi*. Subsequent single cell analysis of parasite survival shows that *T. cruzi* growth is enhanced by prior uptake of apoptotic cells, but not of necrotic cells or control particles.

8 How would you investigate the macrophage surface receptors responsible for apoptotic cell uptake?
9 Suggest a possible mechanism by which apoptotic cell uptake promotes *T. cruzi* survival.
10 How would you investigate this experimental model?
11 What is the possible in vivo and clinical significance of this observation?

The role of macrophages in TH1 and TH2 responses (see p. 496 for explanations)

Isolated mouse peritoneal macrophages are treated for 2 days with selected cytokines (IFNγ, IL-4, IL-13, IL-10) in culture and a range of assays for cell activation is employed. It is found that IFNγ enhances the respiratory burst (after LPS challenge), MHC class II expression, and proinflammatory cytokine production, but downregulates mannose receptor (MR)-mediated endocytosis. IL-10 is an efficient antagonist of the above effects. IL-4 and IL-13 are weak antagonists of respiratory burst and proinflammatory cytokine production and markedly induce MHC class II molecules and MR activity.

12 Interpret the significance and possible functional relevance of these results in relation to concepts of TH1/TH2 differentiation.

13 What further work could be done to investigate the possibility that macrophage activation could by analogy be classified as M1/M2?

14 How would you investigate the role of macrophages and dendritic cells as possible inducers of CD4^{+} T cell subset differentiation?

CHAPTER

10

Cell-mediated Cytotoxicity

SUMMARY

- **Cell-mediated cytotoxicity is an essential defense against intracellular pathogens, including viruses, some bacteria, and parasites.** CTLs recognize antigen presented on MHC molecules. Most CTLs are CD8$^+$ and recognize antigenic peptides presented on MHC class I molecules. NK cells react against cells that do not express MHC class I molecules. They can interact with these cells using a variety of receptors.

- **NK cells express a variety of receptors.** The lectin-like receptor CD94 interacts with HLA-E. KIRs (killer immunoglobulin-like receptors) are members of the immunoglobulin superfamily – those with short tails are activating, and those with long tails are inhibitory. Immunoglobulin-like transcripts (ILTs) have a wider cell distribution than other NK cell receptors.

- **Interactions with NK receptors determine NK cell action.** NK cells use several different receptors to positively identify their targets, and intracellular signaling pathways coordinate inhibitory and activating signals.

- **Cytotoxicity is effected by direct cellular interactions, cytokines, and granule exocytosis.** Fas ligand and TNF can signal apoptosis to the target cell. Granules containing perforin and granzymes contribute to target cell damage. Ligation of Fas or the type 1 TNF receptor on the target cell leads to the activation of caspases, which are the ultimate mediators of apoptosis in the target.

- **Macrophages, neutrophils, and eosinophils are non-lymphoid cytotoxic effectors.** Macrophages damage targets using their non-specific toxic effector systems or via cytokines. Eosinophils mediate cytotoxicity by exocytosis of their granules.

CELL-MEDIATED CYTOTOXICITY IS AN ESSENTIAL IMMUNE DEFENSE MECHANISM

Cytotoxicity describes the ways in which leukocytes recognize and destroy other cells.

Cell-mediated cytotoxicity is an essential defense against:

- intracellular pathogens, including viruses;
- some bacteria;
- some parasites.

Tumor cells, eukaryotic pathogens, and even cells of the body may also become the target of cytotoxic cells.

Cytotoxicity is also important in the destruction of allogeneic tissue grafts.

Several types of cell have cytotoxic activity including:

- cytotoxic T lymphocytes (CTLs);
- natural killer (NK) cells; and sometimes
- myeloid cells.

The mechanisms of recognition and killing used by the lymphoid cells are quite distinct from those of the myeloid cells, and will be considered first.

CTLs recognize antigen presented on MHC molecules and NK cells react against cells that do not express MHC class I

CTLs and NK cells recognize their targets in different ways (Fig. 10.1):

Recognition of target cells by cytotoxic T lymphocytes (CTLs) and NK cells

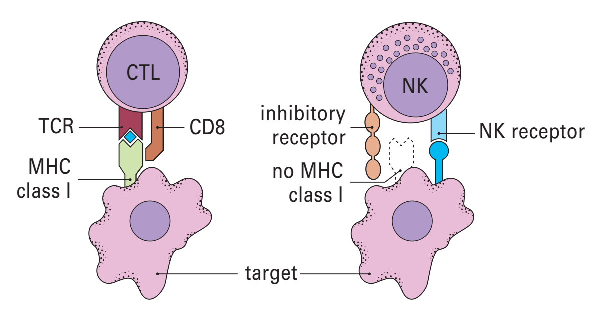

Fig. 10.1 CTLs recognize processed antigen presented on the target cell by MHC molecules using their T cell receptor (TCR). Most CTLs are CD8$^+$ and recognize antigen presented by MHC class I molecules, but a minority are CD4$^+$ and recognize antigen presented by MHC class II molecules. In contrast, NK cells have receptors that recognize MHC class I on the target and signal inhibition of cytotoxicity. They use a number of different receptors (NK receptors) to identify their targets positively, including CD2, CD69, or antibody bound to their Fc receptor (CD16).

Rapid Reference Box 10

ADCC (antibody-dependent cell-mediated cytotoxicity) – a cytotoxic reaction in which Fc receptor-bearing cells recognize target cells via specific antibodies.

Caspases – a group of enzymes that initiate and act as effectors of apoptosis.

CD94 – lectin-like receptor that recognizes MHC class I peptides presented by HLA-E molecules.

CD95 – Fas, a molecule expressed on a variety of cells that acts as a target for ligation by CD95L on cytotoxic cells.

Granzymes – granule associated enzymes that are released by cytotoxic cells and which induce apoptosis in target cells.

HLA-E molecules – class Ib molecules encoded within the MHC which present leader peptides of MHC class I molecules for recognition by CD94/NKG2 receptors on NK cells.

HLA-G molecules – class Ib molecules encoded within the MHC which are expressed on cells of the placenta and inhibit NK cell-mediated cytotoxicity.

ITAMs and ITIMs (immunoreceptor tyrosine activation/inhibitory motifs) – these are target sequences for phosphorylation by kinases involved in cell activation or inhibition.

KIRs (killer immunoglubulin-like receptors) – a group of molecules present on NK cells that, when ligated can activate or inhibit cytotoxicity.

Lymphotoxins – a group of cytokines that act on TNF/LT receptors and which may induce apoptosis in target cells.

Perforin –a molecule contained in the granules of cytotoxic cells which is related to complement C9, and when released can form pores in the plasma membrane of a target cell.

TNFα – a cytokine encoded within the MHC, that may induce apoptosis in target cells.

- CTLs recognize specific antigens (e.g. viral peptides on infected cells) presented by MHC molecules – most CTLs cells are $CD8^+$ and recognize antigen presented on MHC class I molecule, though some $CD4^+$ cells are cytotoxic and recognize antigen presented on MHC class II molecules;
- NK cells recognize cells that fail to express MHC class I molecules and also use a variety of receptors to recognize their targets positively (e.g. they can bind to antibody already attached to antigen on a target cell using their Fc receptors (CD16) – a process known as **antibody-dependent cell-mediated cytotoxicity [ADCC]**).

The most important role of CTLs is the elimination of cells infected with virus (see Chapter 13).

Nearly all nucleated cells express MHC class I molecules and if they become infected can therefore present antigen to $CD8^+$ CTLs (see Chapters 5 and 7).

Cellular molecules that have been partly degraded by proteasomes are transported to the endoplasmic reticulum to become associated with MHC class I molecules and are then transported to the cell surface.

Q. How does a virus-infected cell know which molecules are viral antigens and should therefore be presented to CTLs?
A. It does not. Each cell samples its own molecules and presents them for review by $CD8^+$ CTLs. Both the cell's own molecules and those of intracellular pathogens will be presented in this way. It is the T cells that can distinguish self molecules from non-self antigens, not the infected cell.

Additional interactions may be required to stabilize the bond between the CTL and the target (Fig. 10.2), and can even help trigger the killing event. For example, by adding antibodies against CD3 or CD2 on the CTL in vitro, it is possible to trigger the killing of target cells that are bound to the CTL. It is probable that binding of physiological ligands to these molecules (CD3 and CD2) can also trigger CTLs in this way.

CTLs and NK cells are complementary in the immune defense against virally infected cells

Several viruses (particularly herpes viruses) have evolved mechanisms to avoid recognition by CTLs. They reduce the expression of MHC molecules or even produce proteins that pick MHC molecules out of the endoplasmic reticulum to reduce the likelihood that processed viral peptides will be presented at the cell surface.

Because NK cells specifically recognize cells that have lost their MHC class I molecules, we can see that CTLs and NK cells act in a complementary way to protect the body. In effect:

- the NK cells check that cells of the body are carrying their identity card (MHC class I);
- the CTL checks the specific identity (antigen specificity) on the card.

Experiments involving the micromanipulation of individual cells have shown that a single CTL can sequentially kill several target cells. To do this CTLs must be:

- resistant to their own killing mechanisms; and
- able to detach effectively from dying target cells.

Interactions between CTLs and target cells

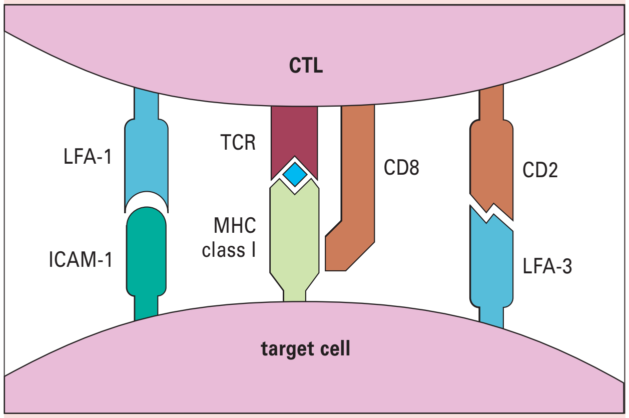

Fig. 10.2 Some of the ligands involved in the interaction between CTLs and their targets. (ICAM-1, intercellular adhesion molecule-1; TCR, T cell receptor; LFA-3, leukocyte functional antigen-1)

Cytokine-activated killer (LAK) cells are related to NK cells

Immunologists have experimented with several potential treatments for cancer. One approach has been to activate the patient's own lymphocytes in vitro with interleukin-2 (IL-2), and then to reinfuse them. These cells, which are initially derived from blood or spleen, are called **cytokine-activated killer cells** or, originally, **lymphokine-activated killers** (**LAKs**). They:

- show enhanced MHC non-restricted cytotoxicity; and
- appear to be largely derived from precursor cells that are indistinguishable from NK cells.

Therefore LAK cells probably do not represent a separate lineage, but rather a consequence of activation. This type of cell is undergoing trials for the treatment of cancer in humans.

NK CELLS EXPRESS A VARIETY OF RECEPTORS

NK cells are mostly derived from large granular lymphocytes (LGLs), which comprise about 5% of human peripheral blood lymphocytes. The majority of NK cells are:

- $CD3^-$, $CD16^+$, $CD56^+$, $CD94^+$ (see Appendix 2); and
- do not contain productive rearrangements of the TCR genes.

Initial experiments on the specificity of NK cells showed that MHC class I expression protected cells from NK cell-mediated cytotoxicity and that particular allotypes of HLA-C were dominant genes for producing resistance.

This led to a search for receptors on NK cells that could inhibit cytotoxicity and that might be expressed in a variety of different forms capable of:

- reacting with MHC class I; and
- signaling the presence of MHC class I molecules to the NK cell.

Two major types of molecule were identified and termed **killer inhibitory receptors** (**KIRs**):

- one group of molecules was identified as type 2 membrane glycoproteins (C terminus outside) with a C-type lectin domain (Ca^{2+}-dependent);
- the other group of molecules were members of the immunoglobulin superfamily.

Subsequently it was realized that, although some members of each group did indeed act as killer inhibitory receptors, others could actually activate the killer cell. So it was not correct to call all members of both groups 'inhibitory receptors'. It was therefore proposed that the term '**KIR**' now meaning '**killer immunoglobulin-like receptor**', should be used for just the second group of molecules.

The lectin-like receptor CD94 interacts with HLA-E

The lectin-like receptor **CD94** is a characteristic marker of human NK cells, but is also found on a subset of CTLs. It covalently assembles with different members of another group of type 2 membrane molecules called **NKG2**, and the dimers are expressed at the cell membrane (Fig. 10.3).

There are at least six members of the NKG2 family (NKG2A–NKG2F). The dimer of CD94–NKG2A is an inhibitory receptor that blocks NK cell-mediated cytotoxicity. By contrast CD94–NKG2C is an activating receptor (see Fig. 10.3). Although these two NKG2 molecules are very similar, they differ in their intracellular segments, and this determines whether the receptor is inhibitory or activating.

Lectin-like receptors of NK cells

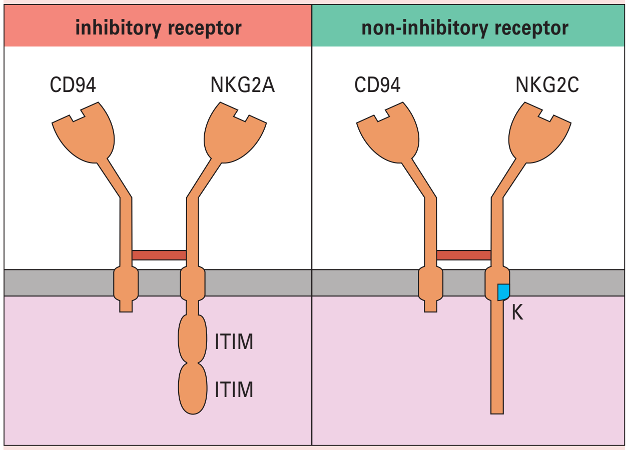

Fig. 10.3 The inhibitory receptors consist of CD94 disulfide-bonded (red) to peptides from the *NKG2* locus, such as NKG2A, which have intracellular domains carrying ITIM motifs (immunoreceptor tyrosine inhibitory motif). Non-inhibitory receptors (such as CD94/NKG2C) lack the ITIMs, but have a charged lysine residue (K) in the transmembrane segment that allows them to interact with signal transducing molecules.

In vivo, the role of the inhibitory version is clear, but that of the activating variant is not. Possibly, it may act as one of the receptors by which NK cells carrying CD94–NKG2C actively engage their targets.

CD94 is distantly related to the mouse molecule Ly-49, which is present in the NK cell gene complex (NKC). Indeed, it was partly this homology of CD94 to mouse NK cell receptors that alerted researchers to the possibility that CD94 was an NK cell receptor in humans.

The ligand for both CD94–NKG2A and CD94–NKG2C is the **HLA-E** molecule.

HLA-E molecules present peptides from other MHC class I molecules

The *HLA-E* gene locus encodes an **MHC class I-like molecule**. These are sometimes called **class Ib molecules** to distinguish them from the classical MHC molecules that present antigen to CTLs.

The extraordinary function of HLA-E is to present peptides from other MHC class I molecules. The leader peptides from other MHC molecules are transported to the endoplasmic reticulum and are required to stabilize functional HLA-E molecules (Fig. 10.4). Cells lacking classical MHC class I molecules do not express HLA-E at the cell surface. Hence an inhibitory signal is not passed to the NK cell.

In effect HLA-E is a sensitive mechanism for monitoring whether viruses or tumors have downregulated MHC class I molecule expression in a cell.

HLA-E presents peptides of other MHC class I molecules

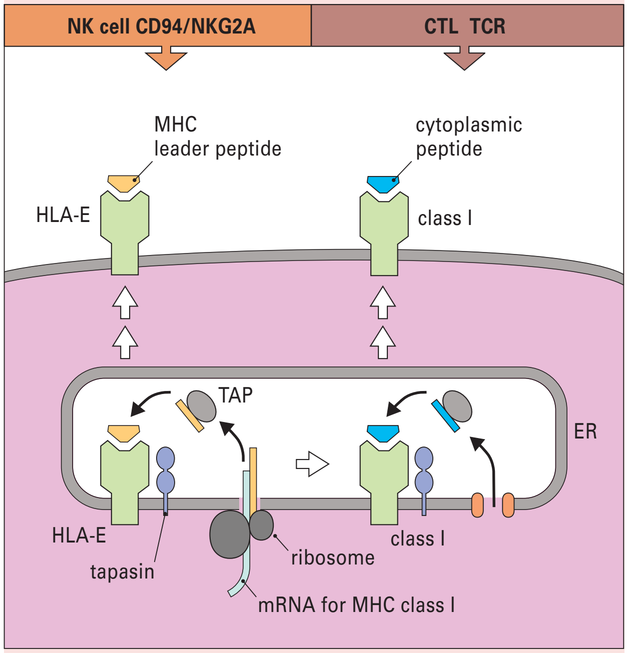

Fig. 10.4 Leader peptides from MHC class I molecules are loaded onto HLA-E molecules in the endoplasmic reticulum, a process that requires TAP transporters and tapasin to assemble functional HLA-E molecules. These are presented at the cell surface for review by the CD94 series of receptors on NK cells (left). The MHC class I molecules meanwhile present antigenic peptides from cytoplasmic proteins that have been transported into the endoplasmic reticulum. These complexes are presented to the TCR on CD8$^+$ CTLs.

Q. Unlike the MHC molecules themselves, their leader peptides are very similar, regardless of the haplotype of the molecules. From this observation, what can you infer about immune recognition, mediated by NK cells?

A. The combination of HLA-E plus leader peptide is similar regardless of the haplotype, therefore the CD94–NKG2 receptor does not need to be highly polymorphic to recognize MHC molecules of different haplotypes. Compare this with antigen presentation by HLA-A and -B and immune recognition by the TCR.

Although CD94 has a lectin domain, the recognition of the HLA-E–peptide occurs via interaction with residues from the peptide and the α1 and α2 domains of the MHC molecule. The lectin domain could, however, reinforce this interaction by binding to carbohydrate associated with the MHC molecule.

KIRs are members of the immunoglobulin superfamily

The second group of NK cell receptors are members of the immunoglobulin superfamily. They fall into two subsets, having either two or three immunoglobulin domains (Fig. 10.5).

Killer immunoglobulin-like receptors

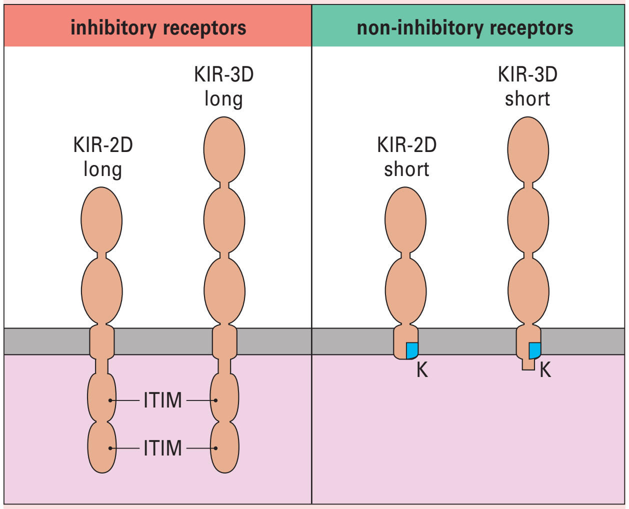

Fig. 10.5 KIRs consist of either two or three extracellular immunoglobulin superfamily domains. The inhibitory forms are longer and have intracellular ITIMs (immunoreceptor tyrosine inhibitory motifs), whereas the non-inhibitory forms have the charged residue in the membrane comparable to the non-inhibitory forms of CD94–NKG2 (see Fig. 10.3).

Thirteen groups of KIRs have been identified encoded in a gene cluster on chromosome 19q13.4. The molecules are highly diverse:

- the two-domain members (KIR-2D) are defined as CD158;
- the three-domain members are KIR-3D.

The nomenclature also indicates whether the variant has a short or long cytoplasmic tail (e.g. KIR-2DL1 for a two-domain variant with a long cytoplasmic tail).

KIRs with short tails are activating, and those with long tails are inhibitory

Two of the two-domain members have been shown to bind to allelic variants of HLA-C, and these isoforms have inhibitory cytoplasmic domains (see below). The specificity of these receptors can:

- partly explain why particular HLA-C allotypes inhibited some NK cells;
- explain why these allotypes produce dominant resistance – if a cell expresses a particular HLA-C allotype, then that is sufficient to inhibit an NK cell that has engaged it.

Interestingly the expression of KIRs varies on individual NK cells. Not all cells express all KIRs, though the total population of NK cells does express all of the KIR genes present in that individual.

Other isoforms of the KIRs have activating domains. These generally engage HLA molecules with a lower affinity than the inhibitory forms, and, like the activating forms of CD40–NKG2, their physiological functions are uncertain.

Immunoglobulin-like transcripts (ILTs) have a wider cell distribution than other NK cell receptors

A third group of receptors is clustered in the same region as the KIRs and have either two or four immunoglobulin-like domains. These receptors, called **immunoglobulin-like transcripts** (**ILTs**), have a wider cellular distribution than the other NK cell receptors. Some of these interact with a broad spectrum of MHC molecules, and others with none at all – their functions are therefore still uncertain.

INTERACTIONS WITH NK RECEPTORS DETERMINE NK CELL ACTION

HLA-G inhibits NK cell action against the placenta

An intriguing recent discovery is that the HLA-G molecule, which is expressed only on placental trophoblasts, is a dominant NK cell inhibitor that confers resistance to all types of NK cell (HLA-G is another MHC class Ib molecule). Trophoblast cells are derived from the fetus and invade the maternal circulation as the placenta is established.

Q. What can you say about trophoblast cells with regard to their susceptibility to damage by CTLs?

A. The fetal cells contain paternal MHC genes so are allogeneic in the mother, and one might therefore expect them to be susceptible to destruction by CTLs.

Q. Because trophoblasts are not destroyed by maternal CTLs, what can you predict about their expression of MHC molecules, and what is the consequence of this?

A. *HLA-A* and *HLA-B* genes are downregulated on trophoblasts, and consequently one would expect the cells to become susceptible to damage by NK cells (*HLA-C* is still expressed.) In practice, therefore, *HLA-G* expression is required to protect the placenta from attack by NK cells.

There is some debate as to which of the inhibitory receptors recognize HLA-G. One ligand for HLA-G is ILT-2. The leader peptide of HLA-G can also be presented in complex with HLA-E to CD94. It is possible that HLA-G can also be directly recognized by a KIR (KIR-2DL4).

NK cells use several different receptors to positively identify their targets

NK cells may engage their targets using a variety of receptors, including CD2, CD16, and CD69, and receptors related to those that inhibit cytotoxicity.

The Fc receptor (CD16) binds antibody bound to target cells and mediates ADCC (Fig. 10.6). Historically this was referred to as killer (K) cell activity, but this function may also be performed by several other cell types with Fc receptors, including T cells.

Myeloid cells expressing Fc receptors can also show antibody-dependent cytotoxicity, but probably use different killing mechanisms to those of T cells and NK cells.

K cell activity

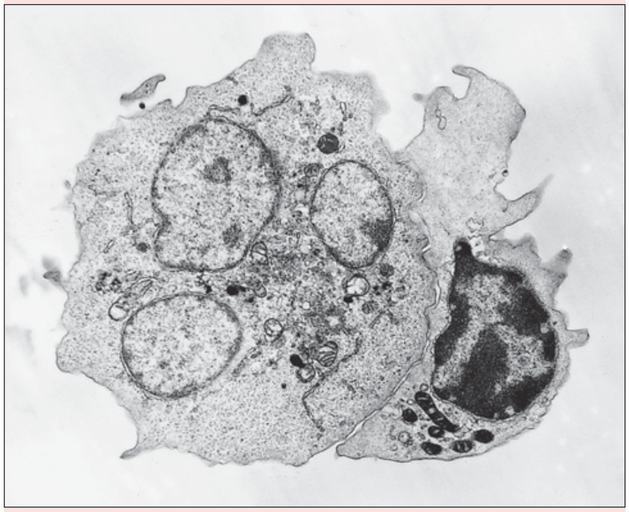

Fig. 10.6 Electron micrograph of a large granular lymphocyte (right) engaging a target cell sensitized with antibody (left). × 2500. (Courtesy of Dr P Penfold)

Potential targets for ADCC include:

- viral antigens on cell surfaces;
- MHC molecules; and
- some epitopes present on tumors.

Therefore monocytes and (according to some controversial reports) polymorphs may also be active against antibody-coated tumor targets. Some myeloid cells (monocytes and eosinophils) are certainly important effectors of damage to antibody-coated schistosomulae (see Chapter 15).

Intracellular signaling pathways coordinate inhibitory and activating signals

The next question is how does an NK cell decide between cytotoxic action or inaction. This decision is thought to depend on the coordination of intracellular signaling pathways, and may involve the balance between activating and inhibitory signals.

Both the lectin-like receptor and the KIRs occur as inhibitory or activating molecules. The key difference is the presence of **immunoreceptor tyrosine inhibitory motifs (ITIMs)**:

- If ITIM motifs become phosphorylated, they can recruit phosphatases, which downregulate the activity of the NK cell (Fig. 10.7).
- In contrast, other KIRs that lack the ITIMs can associate with a molecule (DAP12) which is related to the ζ chain of the TCR complex. DAP12 has **immunoreceptor tyrosine activation motifs (ITAMs)**, which allow it to phosphorylate and recruit tyrosine kinases including ZAP-70 (see Figs 10.7 and 7.21), which lead to cell activation.

At present, it is not known how the balance of activation and inhibition is resolved. In particular it is uncertain how a ligand such as HLA-C will act when it can bind to both inhibitory and activating receptors on the same cell.

Activation and inhibition by NK cell receptors

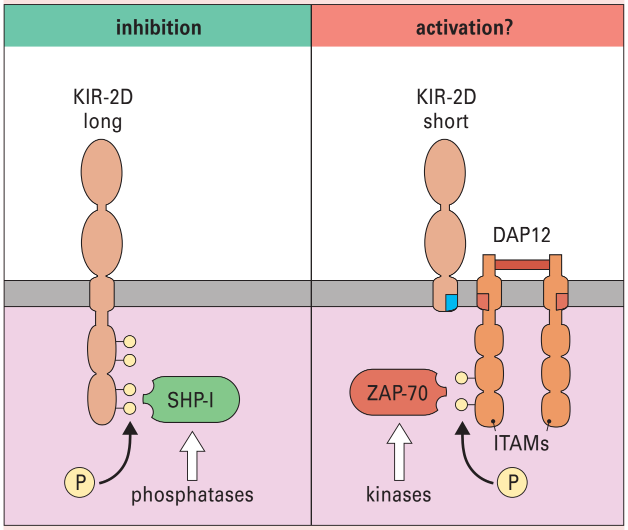

Fig. 10.7 Following phosphorylation of its ITIMs, the inhibitory receptors of NK cells can bind to phosphatases, including SHP-1 and SHP-2, which inhibit killing. The non-inhibitory forms of the receptor associate with a dimeric molecule DAP12, via the complementary charged residues in their membranes. DAP12 has immunoreceptor tyrosine activation motifs (ITAMs). When phosphorylated this recruits kinases of the Syk family or ZAP-70 (see Chapter 7). Whether this leads to NK cell activation or whether it modulates the inhibitory signals is not known.

TNF receptor family of molecules

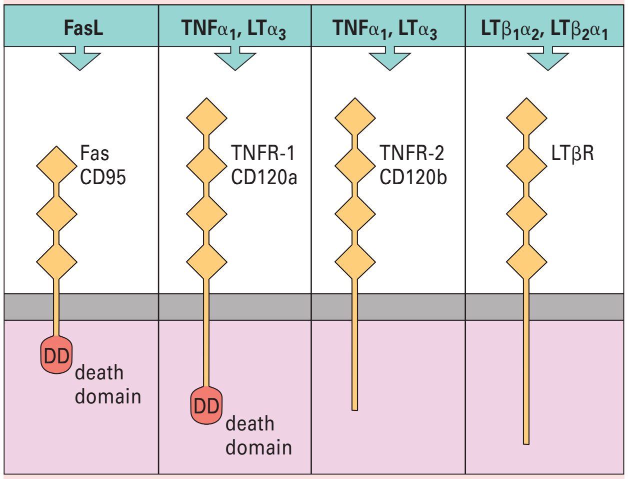

Fig. 10.8 The molecule Fas (CD95), the two TNF receptors, and the lymphotoxin receptor are illustrated. The extracellular domains are similar to those found in the NGF (nerve growth factor) receptor. Both Fas and TNFR-1 have death domains, which are involved in the recruitment of caspases. The ligands for these receptors are indicated at the top. Lymphotoxin-α can form homotrimers or heterotrimers with lymphotoxin-β. Up to 25 other members of these families have been identified by database searching.

CYTOTOXICITY IS EFFECTED BY DIRECT CELLULAR INTERACTIONS, CYTOKINES, AND GRANULE EXOCYTOSIS

CTLs and NK cells use a variety of different mechanisms to kill their targets. These include:

- direct cell–cell signaling via surface molecules; and
- indirect signaling via cytokines.

In addition, CTLs and large granular lymphocytes (NK cells) have granules that contain proteins that damage target cells if they are released directly against the target cell plasma membrane.

Exactly which combination of these three mechanisms is used depends on the CTL involved.

Cytotoxicity may be signaled via Fas or a TNF receptor on a target cell

CTLs signal to their targets using members of the **tumor necrosis factor (TNF)** receptor group of molecules, which include:

- Fas (CD95); and
- the TNF receptors (Fig. 10.8).

Both CD4 and CD8 T cells can initiate cell death via expression of Fas ligand, a member of the TNF family (Fig. 10.9). Fas ligand recognizes the widely expressed cell surface protein Fas. Cross-linking leads to trimerization of Fas and recruitment of FADD to the death domain in the cytoplasmic tail of Fas. FADD recruits caspase 8 or 10, leading to Ca^{2+}-dependent apoptosis.

Activated caspase 8 can cleave and activate other caspases, in addition to its own direct actions in the pathways of apoptosis.

Other members of the TNF receptor family with death domains in their cytoplasmic tails, such as TNFR-1, can also trigger caspase-dependent cell death upon engagement with their ligand by a similar mechanism.

This ability to induce cell death is dependent on the presence of the death domain in the cytoplasmic tail. TNF receptor members lacking the death domain do not mediate caspase-dependent cell death.

Granules of CTLs contain perforin and granzymes

Activated CTL and NK cells contain numerous cytoplasmic granules termed **lytic granules**. Upon recognition of a target cell, these granules polarize to the site of contact, an area referred to as the 'immunological synapse', releasing their contents into a small cleft between the two cells (Fig. 10.10).

The lytic granules contain the pore-forming protein, **perforin**, as well as a series of granule-associated enzymes, **granzymes**, all of which are secreted upon recognition of a target.

Perforin is a monomeric pore-forming protein

Although NK cells and CTLs use different mechanisms to recognize their targets, the lytic mechanisms used to destroy the targets are the same. The key protein is perforin, a monomeric pore-forming protein that is

Mechanisms of cell killing

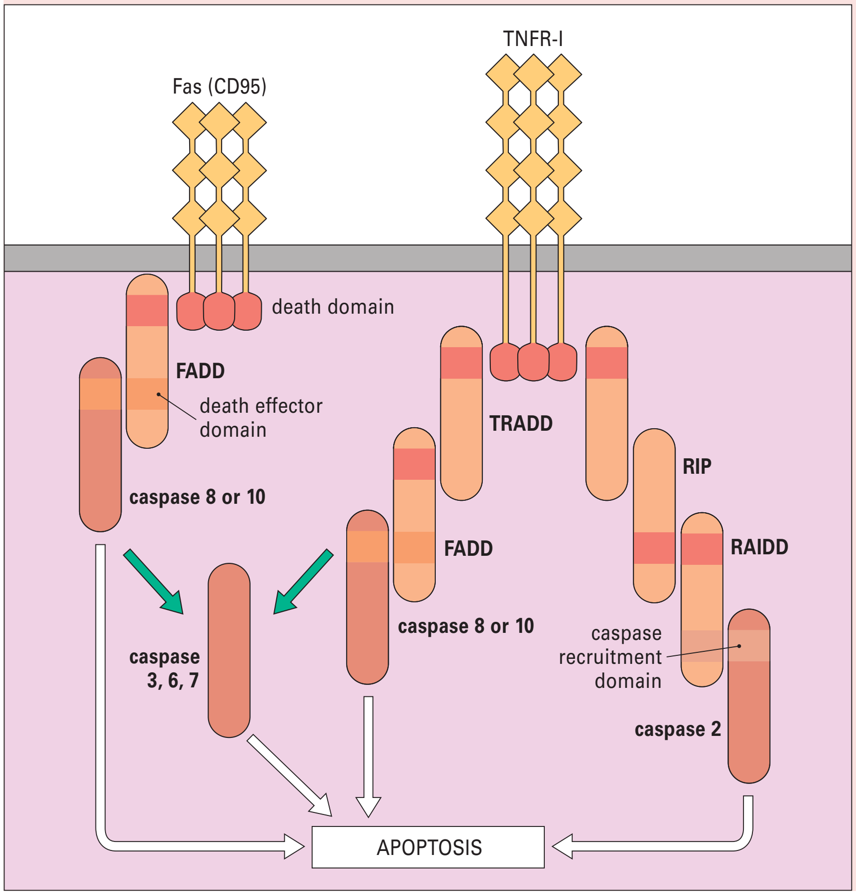

Fig. 10.9 Ligation of CD95 or TNFR-1 causes trimerization of the receptors. Death domains in the cytoplasmic portion of CD95 bind to the adapter protein FADD (MORT1), which recruits caspase 8 or 10. TNFR-1 can activate either caspase 8 or 10, via TRADD and FADD, or caspase 2 via RIP and RAIDD. Caspase 8 can further activate other caspases, and these in concert lead to apoptosis of the target cell.

related both structurally and functionally to the complement component C9 (see Chapter 4).

Q. What function does C9 perform?

A. It polymerizes to form channels across membranes as the final step in the lytic pathway.

Perforin is inactive when located within the granules, but undergoes a conformational activation, which depends on the concentration of Ca^{2+} ions (Fig. 10.11).

Like C9, perforin is able to form homopolymers, inserting into the membrane to form a circular pore of approximately 16 nm in diameter.

Unlike C9, perforin possesses a domain that can bind phospholipid membranes directly in the presence of Ca^{2+} ions.

Perforin-deficient mice show greatly reduced cytotoxicity. The fact that some cytotoxicity remains demonstrates that other mechanisms contribute to CTL and NK cell-mediated death. Some of this residual killing is likely to come from the Fas ligand and TNF pathways.

Granzymes cleave substrates in the target cell, leading to the rapid initiation of apoptosis

Also contained within the lytic granules are the **granzymes**, which are a series of serine proteases that require perforin to enter the cytoplasm of the target cell (Fig. 10.12). Once in the cytoplasm, granzymes can cleave a number of substrates, leading to the rapid initiation of apoptosis:

- Granzyme B cleaves a number of pro-caspases, leading to the activation of caspases 10, 3, and 7, so triggering apoptosis of the target cell. Granzyme B-deficient mice show delayed but not ablated cytotoxicity, revealing the role of other pathways.
- Another mechanism is provided by granzyme A, which triggers apoptosis via a caspase-independent pathway. Granzyme A acts on one of its substrates, SET (an endoplasmic reticulum-associated protein complex), activating a DNAse, which nicks the DNA and leads to nuclear breakdown.
- Granzyme C also contributes to apoptosis via a caspase-independent pathway, though the details of this pathway are not yet known.

Eleven granzymes have been identified in the mouse and five in humans, though the functions of all of these are not yet known. The large number of granzymes is likely to provide multiple pathways to trigger apoptosis, ensuring that cell death ensues.

In summary, recognition of a target leads to rapid polarization of the lytic granules to the immunological synapse. The contents of the granules are released,

Intracellular reorganizations during effector–target cell interaction

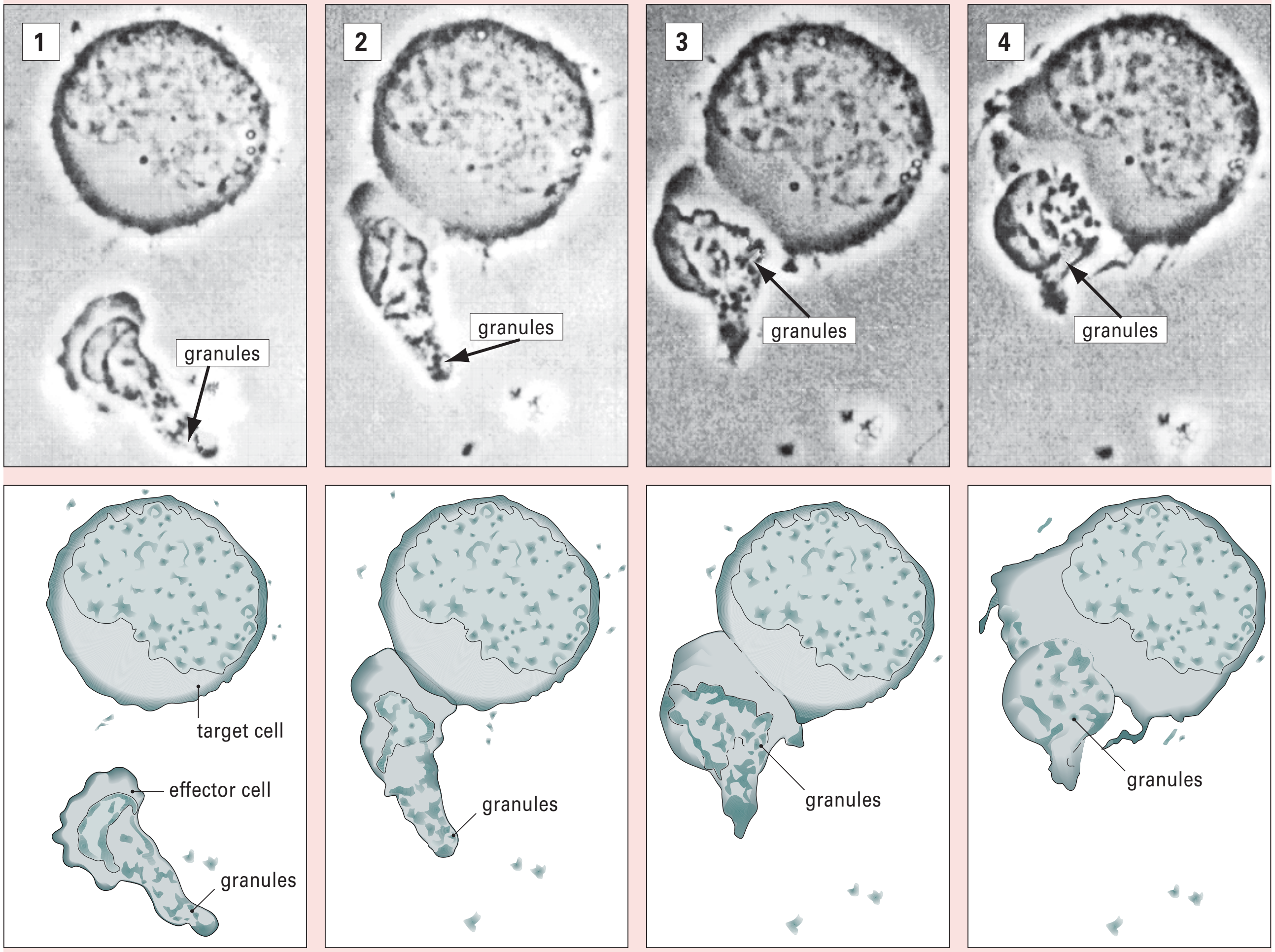

Fig. 10.10 Early events in the interaction of CTLs with specific targets were studied with high-resolution cinematographic techniques. The figure shows four frames (together with interpretative drawings), taken at different times, of a CTL interacting with its target. The location of the granules within the effector cell is indicated in each case. Before contact with the target (**1**), the effector has granules located in a uropod at the rear, and is seen to move randomly by extending pseudopods from the organelle-free, broad, leading edge of the cell. Within 2 minutes of contacting the target (**2**), the CTL has begun to round up and initiate granule reorientation (**3**). After 10 minutes (**4**), the granules occupy a position in the zone of contact with the target, where they appear to be in the process of emptying their contents into the intercellular space between the two cells. (Courtesy of Dr VH Engelhard)

including the pore-forming protein perforin, which allows granzymes to enter the cytoplasm of the target cell and rapidly trigger apoptosis.

CTL and NK cells kill targets without killing themselves

Although CTLs can be killed by other CTLs they do not destroy themselves when they kill another target. A number of mechanisms contribute to this:

- Both perforin and granzymes are synthesized as inactive precursors that need to be activated by removal of a small pro-piece.
- Activation takes place only after perforin and granzymes have been released from the granules, where the pH, Ca^{2+} levels, and the presence of proteoglycans that bind both perforin and granzymes are thought to keep both perforin and granzymes inactive.

Nevertheless the CTL is still resistant to its own perforin as it releases the granule contents into the immunological synapse. Some recent work has shown that a membrane-bound form of cathepsin B contributes to the CTL resistance to lysis because it lines the granule membranes and cleaves perforin on the CTL side of the synapse.

Additional mechanisms render CTLs more resistant to lysis, for example CTLs express:

- cFLIP, a protein that inhibits the cleavage of caspase 8 and prevents apoptosis triggering via the caspase 8 pathway;
- the serpin, protease inhibitor 8 (PI-8), which can prevent the action of granzyme B.

Perforin undergoes a conformational activation

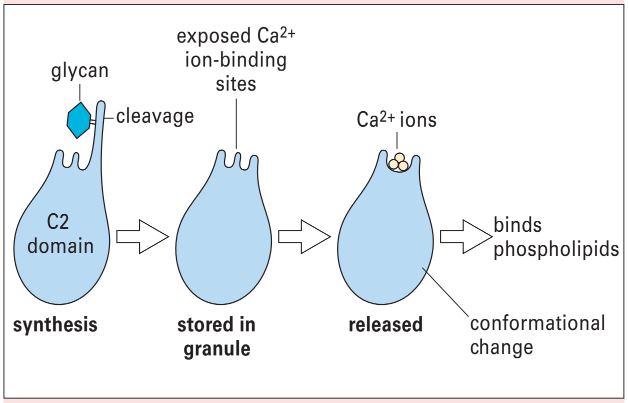

Fig. 10.11 Perforin is synthesized with a tailpiece of 20 amino acid residues with a large glycan residue attached. In this form it is inactive. Cleavage of the tailpiece within the granules allows Ca^{2+} ions to access a site associated with the C2 domain of the molecule, but the molecule is thought to be maintained in an inactive state by low Ca^{2+} concentrations in the granules. Following secretion the increased Ca^{2+} concentration permits a conformational change, which exposes a phospholipid-binding site, allowing the perforin to bind to the target membrane as a precursor to polymerization.

Diseases resulting from a loss of lytic granule proteins

A number of genetic diseases have been identified that result from loss of function of several different effector proteins involved in CTL-mediated killing. Strikingly, these give rise to very different clinical syndromes.

Mutations that cause loss of perforin expression or function give rise to **familial hemophagocytic lymphohistiocytosis**, a disease characterized by massive infiltration of organs with activated macrophages engulfing lymphocytes.

Q. What effect would you expect lack of perforin to have in affected individuals?

A. CTLs are unable to kill virally infected cells, and it is thought that the associated macrophage infiltration is a consequence of the inability to regulate the response to viral infections.

Loss of Fas ligand expression gives rise to **autoimmune lymphoproliferative syndrome type 2 (ALPS2)**, which results in aberrant lymphoproliferation in affected patients. The gld mouse model also lacks Fas ligand and has a similar pathology. The symptomatology of ALPS2 is very similar to that caused by mutations that result in loss of Fas (ALPS1).

Mutations that prevent activation of granzymes have now been identified. Cathepsin C is required to activate granzymes by proteolytic cleavage. Patients lacking cathepsin C present with **Papillon–Lefèvre syndrome**. This rare condition appears in young children and is characterized by severe inflammation around the teeth and dry scaly skin, which may become infected. Only a small number of the patients show increased susceptibility to infections and, interestingly, do not appear to have an increased susceptibility to viral infections.

MACROPHAGES, NEUTROPHILS, AND EOSINOPHILS ARE NON-LYMPHOID CYTOTOXIC EFFECTORS

Macrophages can damage targets using their non-specific toxic effector systems or via cytokines

A number of non-lymphoid cells may be cytotoxic to other cells or invading microorganisms, such as bacteria or parasites. Cytotoxicity may be triggered specifically to a

Granule-associated killing mechanisms

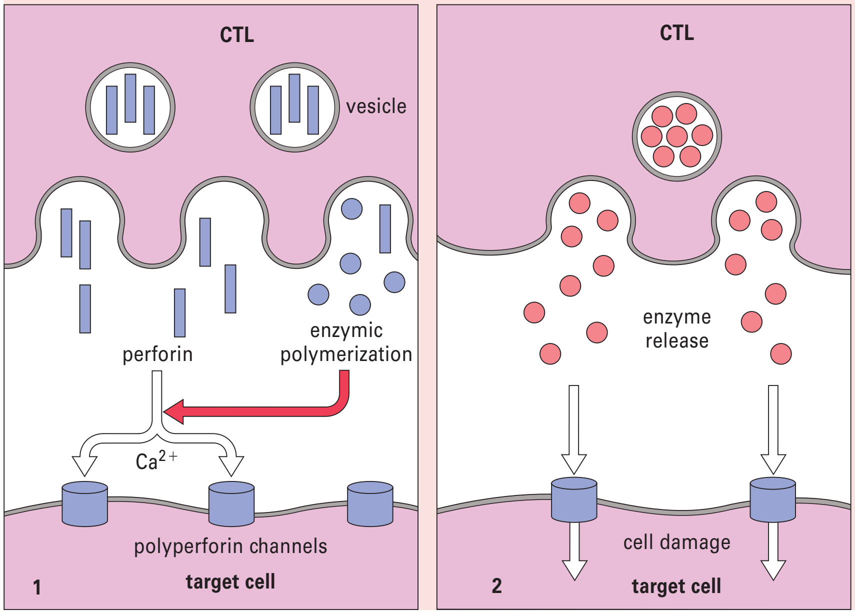

Fig. 10.12 The CTL degranulates, releasing perforin and various enzymes (granzymes) into the immediate vicinity of the target cell membrane. In the presence of Ca^{2+}, perforin binds to the target cell membrane and forms polyperforin channels (**1**). Enzymes that activate the apoptosis pathways, degradative enzymes, and other toxic substances released from the cytotoxic cell may pass through the channels in the target cell membrane and cause cell damage or killing (**2**).

target by ADCC or may involve a range of non-specific toxic mediators. For example, macrophages and neutrophils both express FcγRI and FcγRII which allows them to engage tumors by ADCC.

In general macrophages and neutrophils aim to destroy pathogens by internalizing them and subjecting them to toxic molecules and enzymes within the phagolysosome. These include:

- the production of reactive oxygen intermediates, toxic oxidants, and nitric oxide (Fig. 10.13, see Chapter 9); as well as
- the secreted molecules such as neutrophil defensins, lysosomal enzymes, and cytostatic proteins.

If the phagocyte fails to internalize its target, then these mediators may be released into the extracellular environment and contribute to localized cell damage. This action is referred to as '**frustrated phagocytosis**' and occurs when the target is engaged by surface receptors, but is too large to phagocytose. The actions of the mediators produced by the phagocyte damage the target rather than induce apoptosis. For this reason, the cytotoxic action of phagocytes tends to produce necrosis and inflammation.

Additionally, activated macrophages secrete TNF, which can induce apoptosis in a similar way to NK cells and CTLs. Macrophages can therefore induce necrosis, apoptosis, or a combination of both, depending on the state of activation of the macrophages and the target involved.

Eosinophils mediate cytotoxicity by exocytosis of their granules

Mature eosinophils are characterized by their granules, which have a crystalloid core that binds the dye eosin. Generally, eosinophils are only weakly phagocytic – they ingest some bacteria following activation, but are less efficient than neutrophils at intracellular killing.

The major function of eosinophils appears to be the secretion of various toxic granule constituents following activation. They are therefore effective for the extracellular killing of microorganisms, particularly large parasites such as schistosomes (see Chapter 15).

Mechanisms that may contribute to the cytotoxicity of myeloid cells

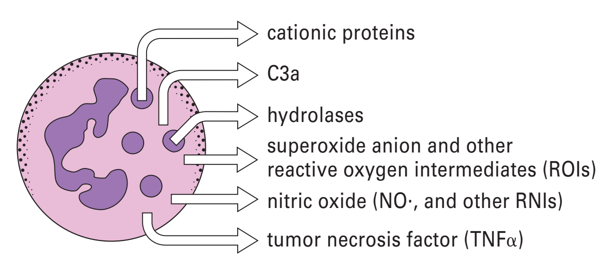
cationic proteins
C3a
hydrolases
superoxide anion and other reactive oxygen intermediates (ROIs)
nitric oxide (NO·, and other RNIs)
tumor necrosis factor (TNFα)

Fig. 10.13 Reactive oxygen intermediates (ROIs) and reactive nitrogen intermediates (RNIs), cationic proteins, hydrolytic enzymes, and complement proteins released from myeloid cells may damage the target cell in addition to cytokine-mediated attack.

The components of the eosinophil granule include the following:

- Major basic protein (MBP) – not to be confused with myelin basic protein of oligodendrocytes. MBP is the major component of eosinophil granules, forming the crystalloid core, and has been shown to damage, and sometimes kill, parasites, but also damages host tissue cells.
- Eosinophil peroxidase (EPO) – a highly cationic heterodimeric 71–77-kDa hemoprotein that is distinct from the myeloperoxidase of neutrophils and macrophages. In the presence of H_2O_2, also produced by eosinophils, EPO will oxidize a variety of substrates, including halide ions to produce hypohalite, and this may represent the eosinophil's most potent killing mechanism for some parasites.
- Eosinophil cationic protein (ECP) – an eosinophil-specific toxin that is very potent at killing many parasites, particularly the schistosomulae of *Schistosoma mansoni*. The molecule is a ribonuclease, which, because of its high charge, binds avidly to negatively charged surfaces and it is possible that it forms membrane channels, which allow other mediators access to the target organism.

Other molecules produced by eosinophils are lysophospholipase and eosinophil-derived neurotoxin (EDN), which is also a ribonuclease but with strong neurotoxic activity.

Degranulation of eosinophils can be triggered in a number of ways:

- binding to IgG-coated parasites via surface FcγRII triggers the release of some mediators including ECP, but not EPO;
- in contrast triggering via FcεRII leads to the release of EPO, but not ECP.

Parasite killing may involve contact-dependent degranulation or may simply require deposition of toxins within the local tissue.

Degranulation may also be triggered directly in vitro by several cytokines, including:

- interleukin-3 (IL-3);
- IL-5;
- granulocyte–macrophage colony stimulating factor (GM-CSF);
- TNF;
- interferon-β (IFNβ); and
- platelet-activating factor (PAF).

These mediators also enhance ADCC-mediated degranulation.

Eosinophils are prominent in the inflammatory lesion of a number of diseases, particularly atopic disorders of the gut, skin, and respiratory tract, where they are often closely associated with fibrotic reactions. Examples are atopic eczema, asthma, and inflammatory bowel disease.

Although eosinophils may play some regulatory role in these conditions, such as inactivating histamine, their toxic products and cytotoxic mechanisms are a major cause of the tissue damage. For example, in asthma, eosinophil granule proteins are detectable in the blood and lungs following asthmatic attacks. MBP can kill some pneumocytes and tracheal epithelial cells while EPO kills type II pneumocytes. MBP can also induce mast cells to secrete histamine, so exacerbating allergic inflammation.

FURTHER READING

Berke G. The binding and lysis of target cells by cytotoxic lymphocytes. Annu Rev Immunol 1994;12:735.

Bleackle RC, Lobe CG, Duggan B, et al. The isolation of a family of serine protease genes expressed in activated cytotoxic T lymphocytes. Immunol Rev 1988;103:5.

Kagi D, Ledermann B, Burki K, et al. Cytotoxicity mediated by T cells and natural killer cells is greatly impaired in perforin-deficient mice. Nature 1994;369:31–37.

Lanier LL. NK cell recognition. Annu Rev Immunol 2005;23:225–274.

Lee N, Llano M, Carreto M, et al. HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. Proc Natl Acad Sci 1998;95:5199–204.

López-Botet M, Bellón T. Natural killer cell activation and inhibition by receptors for MHC class I. Curr Opin Immunol 1999;11:301–307.

Navarro F, Llano M, Bellón T, et al. The ILT2 (LIR-1) and CD94NKG2A cell receptors respectively recognise HLA-G1 and HLA-E molecules on coexpressed targets. Eur J Immunol 1999;29:277–283.

Rathmell JC, Thompson CB. The central effectors of cell death in the immune system. Annu Rev Immunol 1999;17:781–828.

Wallach D, Varfolomeev EE, Malanin NL, et al. Tumor necrosis factor receptors and Fas signalling mechanisms. Annu Rev Immunol 1999;17:331–368.

Watanabe-Fukunga R, Brannan CI, Copeland NG, et al. Lymphoproliferation disorder in mice explained by defects in Fas antigen that mediates apoptosis. Nature 1992;356:314.

Yanelli JR, Sullivan JA, Mandell GL, Engelhard VH. Reorientation and fusion of cytotoxic T cell granules after interaction with target cells as determined by high resolution cinematography. J Immunol 1986;136:377.

Yokoyama WM, Kim S, French AR. The dynamic life of natural killer cells. Annu Rev Immunol 2004;22:405–430.

Critical thinking: Mechanisms of cytotoxicity (see p. 496 for explanations)

Lymphocytes from a normal individual were stimulated in vitro by co-culture with irradiated T lymphoma cells. (Irradiation of these stimulator cells prevents them from dividing in the co-culture.) After 7 days the lymphocytes were harvested and fractionated to obtain a population of cytotoxic cells ($CD8^+$) and a population of NK cells ($CD94^+$, $CD16^+$). These effector cells were set up in a cytotoxicity assay with the tumor cells as targets. The tumor cells were labeled to detect both DNA fragmentation and cell lysis. The results shown in the table were obtained in these assays.

treatment	lysis (%)	DNA fragmentation (%)
no effector cells	4	1
$CD8^+$ CTLs	82	80
NK cells	12	11
anti-tumor cell antibody	5	1
NK cells + anti-tumor cell antibody	28	28
purified perforin	95	2

1 Why do CTLs lyze the targets and induce DNA fragmentation? What are these cells recognizing on the tumor cell surface?

2 Why might the NK cells cause some damage to the tumor? Why does the presence of antibody to the tumor enhance the cytotoxic capacity of the NK cells?

3 Explain the result with purified perforin.

This Page Intentionally Left Blank

CHAPTER 11

Regulation of the Immune Response

SUMMARY

- **Many factors govern the outcome of any immune response.** These include the antigen itself, its dose and route of administration, and the genetic background of the individual responding to antigenic challenge. A variety of control mechanisms serve to restore the immune system to a resting state when the response to a given antigen is no longer required.
- **The APC may affect the immune response** through its ability to provide co-stimulation to T cells. Different types of APC promote different modes of immune response.
- **Immunoglobulins can influence the immune response,** positively as an anti-idiotype or through immune complex formation. They may also negatively influence immune responses by reducing antigenic challenge or by feedback inhibition of B cells.
- **T cells regulate the immune response.** Cytokine production by T cells influences the type of immune response elicited by antigen. CD4$^+$ T cells can deviate immune responses to TH1- or TH2-type responses. Regulatory T cells may belong to the CD4 or CD8 subpopulations. They can inhibit responses by the production of suppressive cytokines such as IL-10 and TGFβ.
- **Senescence of cells is regulated by telomere erosion.** Upregulation of telomerase by cells during an immune response compensates for telomere shortening caused by cell division.
- **Selective migration of lymphocyte subsets to different sites can modulate the local type of immune response** because TH1 cells and TH2 cells respond to different sets of chemokines.
- **The neuroendocrine system influences immune responses.** Corticosteroids in particular downregulate TH1 responses and macrophage activation.
- **Genetic factors influence the immune system and include both MHC-linked and non-MHC-linked genes.** They affect the level of immune response and susceptibility to infection.

MANY FACTORS GOVERN THE OUTCOME OF ANY IMMUNE RESPONSE

The immune response, like all biological systems, is subject to a variety of control mechanisms. These mechanisms restore the immune system to a resting state when responsiveness to a given antigen is no longer required.

An effective immune response is an outcome of the interplay between antigen and a network of immunologically competent cells.

The nature of the immune response, both qualitatively and quantitatively, is determined by many factors, including:

- the form and route of administration of the antigen;
- the antigen-presenting cell (APC);
- the genetic background of the individual; and
- any history of previous exposure to the antigen in question or to a cross-reacting antigen.

Specific antibodies may also modulate the immune response to an antigen.

T cells and B cells are triggered by antigen after effective engagement of their antigen-specific receptors together with appropriate co-stimulation. In the case of the T cell, this engagement is not with antigen itself, but with processed antigenic peptide bound to MHC class I or class II molecules on APCs (see Chapter 7).

The nature of an antigen, the dose, and the route of administration have all been shown to have a profound influence on the outcome of an immune response. An effective immune response removes antigen from the system.

Repeated antigen exposure is required to maintain T and B cell proliferation, and during an effective immune response there is often a dramatic expansion of specifically reactive effector cells.

At the end of an immune response reduced antigen exposure results in a reduced expression of IL-2 and its receptor leading to **apoptosis** (or programmed cell death, see Chapter 10) of the antigen-specific T cells. The majority of antigen-specific cells therefore die at the end of an immune response leaving a minor population of long-lived T and B cells to survive and give rise to the memory population.

Q. What process leads to the inactivation of T cells and loss of the IL-2 receptor?

A. Ligation of CTLA-4, the alternative receptor for the co-stimulatory molecule B7 (see Fig. 7.18).

Different antigens elicit different kinds of immune response

Intracellular organisms such as some bacteria, parasites, or viruses induce a cell-mediated immune response. Cell-mediated immune responses are also induced by agents such as silica.

In contrast extracellular organisms and soluble antigens induce a humoral response, with the polysaccharide capsule antigens of bacteria generally inducing IgM responses.

In some situations, antigens (e.g. those of intracellular microorganisms) may not be cleared effectively leading to a sustained immune response. This has pathological consequences for autoimmunity and hypersensitivity (see Chapters 20 and 23–26).

Large doses of antigen can induce tolerance

Very large doses of antigen often result in specific T and sometimes B cell tolerance.

Administration of antigen to neonatal mice often results in tolerance to the antigen. It was speculated that this might be the result of immaturity of the immune system. However, more recent studies have shown that neonatal mice can develop efficient immune responses (Fig. 11.1) and that non-responsiveness may in some cases be attributable, not to the immaturity of T cells, but to immune deviation. In this case a non-protective TH2-type cytokine response would dominate a protective TH1-type cytokine response.

T-independent polysaccharide antigens have been shown to generate tolerance in B cells after administration in high doses. Tolerance and its underlying mechanisms are discussed in Chapter 19.

Effect of antigen dose on the outcome of the immune response to murine leukemia virus

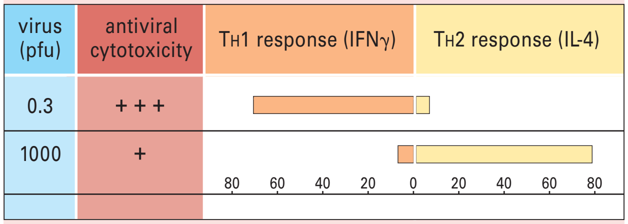

Fig. 11.1 Newborn mice were infected with either 0.3 or 1000 plaque-forming units (pfu) of virus and the CTL response against virally infected targets was assessed together with the production of interferon-γ (IFNγ – a helper T cell type 1 [TH1] cytokine) or interleukin-4 (IL-4 – a TH2 cytokine) in response to viral challenge. Mice infected with a low dose of virus make a TH1-type response and are protected. The results are presented as arbitrary units.

Antigen route of administration can determine whether an immune response occurs

The route of administration of antigen has been shown to influence the immune response:

- antigens administered subcutaneously or intradermally evoke an active immune response; whereas
- antigens given intravenously, orally, or as an aerosol may cause tolerance or an immune deviation from one type of CD4$^+$ T cell response to another.

For example, rodents that have been fed ovalbumin or myelin basic protein (MBP) do not respond effectively to a subsequent challenge with the corresponding antigen. Moreover, in the case of MBP, the animals are protected from the development of the autoimmune disease experimental allergic encephalomyelitis (EAE).

Q. Considering the type of immune responses generated in mucosal tissues why might oral administration of antigen deviate the immune response, and in what way?

A. Mucosal tissues have high levels of antibody-producing B cells. Consequently antigens that first encounter the immune system at these sites may be presented by B cells, which tend to induce TH2-type responses.

This phenomenon may have some therapeutic value in allergy. Recent studies have shown that oral administration of a T cell epitope of the Der p1 allergen of house dust mite (*Dermatophagoides pteronyssimus*) could tolerize to the whole antigen. The potential mechanisms of such tolerance induction include anergy, immune deviation, and the generation of regulatory T cells that act through the production of cytokines such as TGFβ and IL-10.

Similar observations have been made when antigen is given as an aerosol. Studies in mice have shown that aerosol administration of an encephalitogenic peptide inhibits the development of experimental allergic encephalomyelitis (EAE) that would normally be induced by a conventional (subcutaneous) administration of the peptide (Fig. 11.2). This may also have therapeutic implications because the inhibition of the response is not limited to the antigen administered as an aerosol, but also includes other antigens capable of inducing EAE, such as proteolipid protein.

A clear example of how different routes of administration affect the outcome of the immune response is provided by studies of infection with lymphocytic choriomeningitis virus (LCMV). Mice primed subcutaneously with peptide in incomplete Freund's adjuvant develop immunity to LCMV. However, if the same peptide is repeatedly injected intraperitoneally the animal becomes tolerized and cannot clear the virus (Fig. 11.3).

THE APC MAY AFFECT THE IMMUNE RESPONSE

The nature of the APC initially presenting the antigen may determine whether immune responsiveness or tolerance ensues.

Effective activation of T cells requires the expression of co-stimulatory molecules on the surface of the APC. Therefore presentation by dendritic cells or activated

Aerosol administration of antigen modifies the immune response

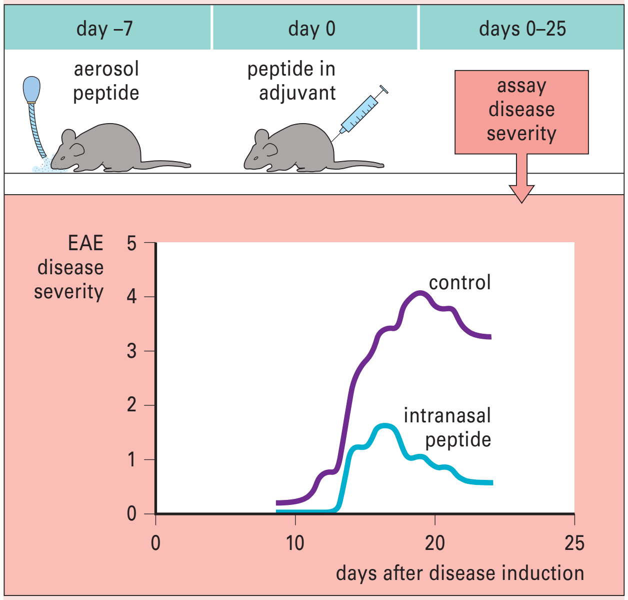

Fig. 11.2 Mice were treated with a single aerosol dose of either 100 µg peptide (residues 1–11 of MBP) or just the carrier. Seven days later the same peptide, this time in adjuvant, was administered subcutaneously. The subsequent development of experimental allergic encephalomyelitis (EAE) was significantly modified in pretreated animals.

Peptide-induced inactivation of LCMV-specific T cells

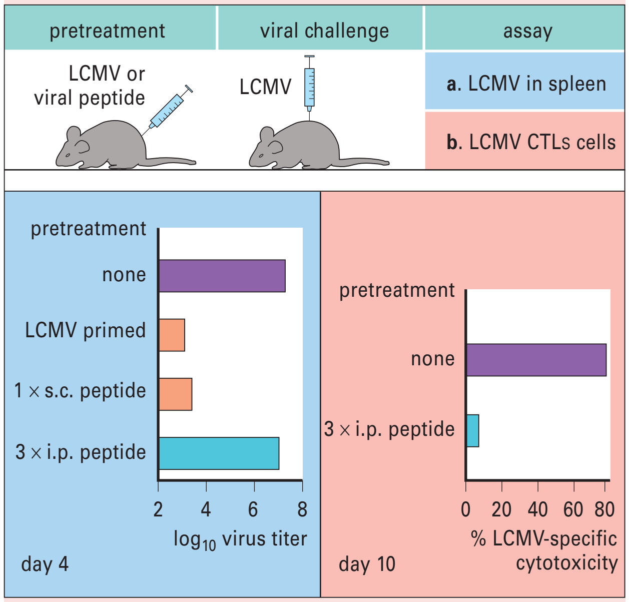

Fig. 11.3 Mice were either primed with LCMV or injected with 100 mg LCMV peptide. The peptide was given either subcutaneously (s.c.) or three times intraperitoneally (i.p.) with incomplete Freund's adjuvant. The animals were later infected with LCMV (day 0). The titer of virus in the spleen was measured on day 4. Animals that had been pretreated with subcutaneous peptide or with LCMV developed neutralizing antibody and protective immunity against the virus; animals pretreated with peptide i.p. did not develop immunity. Cytotoxic T lymphocyte (CTL) activity was assessed in the mice on day 10. Mice that had received no pretreatment demonstrated CTLs specific for the LCMV peptide. Mice pretreated with peptide i.p. failed to show such activity.

macrophages that express high levels of MHC class II molecules in addition to co-stimulatory molecules results in highly effective T cell activation (see Fig. 7.18).

Furthermore, the interaction of CD40L on activated T cells with CD40 on dendritic cells is important for the high-level production of IL-12 necessary for the generation of an effective TH1 response.

If antigen is presented to T cells by a 'non-professional' APC that is unable to provide co-stimulation, unresponsiveness or immune deviation results. For example, when naive T cells are exposed to antigen by resting B cells they fail to respond and become tolerized. Recent experimental observations illustrate this point.

Neonatal animals are more susceptible to tolerance induction. Therefore mice administered MBP in incomplete Freund's adjuvant during the neonatal period are resistant to the induction of EAE. This is due to the development of a dominant TH2 response (see Fig. 11.1). The prior TH2 response to MBP prevents the development of the TH1 pathological response, which mediates EAE.

Adjuvants may facilitate immune responses by inducing the expression of high levels of MHC and co-stimulatory molecules on APCs. Furthermore, their ability to activate Langerhans' cells leads to the migration of these skin dendritic cells to the local draining lymph nodes where effective T cell activation can occur.

The importance of dendritic cells in initiating a cytotoxic T lymphocyte (CTL) response is illustrated by experiments showing that newborn female mice injected with male spleen cells fail to develop a CTL response to the male antigen, H-Y. However if male dendritic cells are injected into female newborn mice, a good H-Y-specific CTL response develops.

IMMUNOGLOBULINS CAN INFLUENCE THE IMMUNE RESPONSE

Antibody has been shown to exert feedback control on the immune response.

Passive administration of IgM antibody together with an antigen specifically enhances the immune response to that antigen, whereas IgG antibody suppresses the response. This was originally shown with polyclonal antibodies, but has since been confirmed using monoclonal antibodies (Fig. 11.4).

The ability of passively administered antibody to enhance or suppress the immune response has certain clinical consequences and applications:

- certain vaccines (e.g. mumps and measles) are not generally given to infants before 1 year of age because levels of maternally derived IgG remain high for at least

Feedback control by antibody

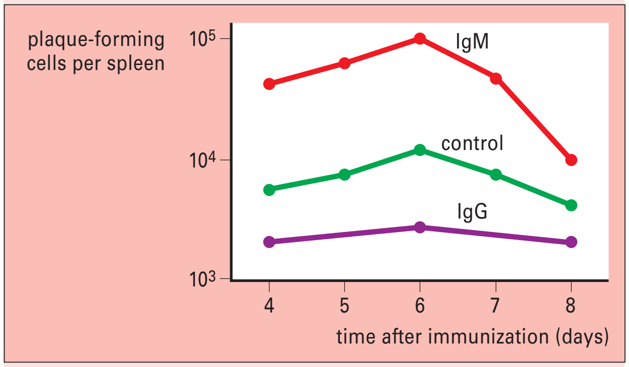

Fig. 11.4 Mice received either a monoclonal IgM anti-SRBC (sheep red blood cells), IgG anti-SRBC, or medium alone (control). Two hours later all groups were immunized with SRBC. The antibody response measured over the following 8 days was enhanced by IgM and suppressed by IgG.

Antibody-dependent B cell suppression

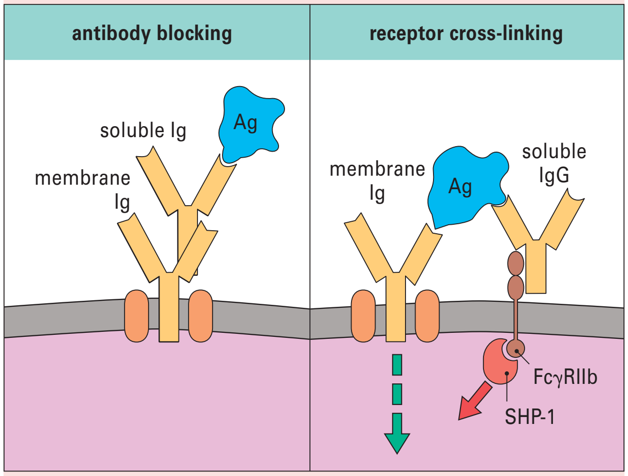

Fig. 11.5 Antibody blocking – high doses of soluble immunoglobulin (Ig) block the interaction between an antigenic determinant (epitope) and membrane immunoglobulin on B cells. The B cell is then effectively unable to recognize the antigen. This receptor-blocking mechanism also prevents B cell priming, but only antibodies that bind to the same epitope to which the B cell's receptors bind can do this. Receptor cross-linking – low doses of antibody allow cross-linking by antigen of a B cell's Fc receptors and its antigen receptors. The FcγRIIb receptor associates with a tyrosine phosphatase (SHP-1), which interferes with cell activation by tyrosine kinases associated with the antigen receptor. This allows B cell priming, but inhibits antibody synthesis. Antibodies against different epitopes on the antigen can all act by this mechanism.

6 months after birth and the presence of such passively acquired IgG at the time of vaccination would result in the development of an inadequate immune response in the baby;
- in cases of Rhesus (Rh) incompatibility, the administration of anti-RhD antibody to Rh$^-$ mothers prevents primary sensitization by fetally derived Rh$^+$ blood cells, presumably by removing the foreign antigen (fetal erythrocytes) from the maternal circulation (see Chapters 24 and 25).

The mechanisms by which antibody modulates the immune response are not completely defined. In the case of IgM-enhancing plaque-forming cells, there are thought to be two possible interpretations:
- IgM-containing immune complexes are taken up by Fc or C3 receptors on APCs and are processed more efficiently than antigen alone;
- IgM-containing immune complexes stimulate an **anti-idiotypic response** to the IgM, which amplifies the immune response.

IgG antibody can suppress specific IgG synthesis

For IgG-mediated suppression there are also various ways in which the antibody is known to act.

Passively administered antibody binds antigen in competition with B cells (antibody blocking)

The impact of the IgG in antibody blocking (Fig. 11.5) in this case is highly dependent on the concentration of the antibody, and on its affinity for the antigen compared with the affinity of the B cell receptors. Only high-affinity B cells compete successfully for the antigen. This mechanism is independent of the Fc portion of the antibody.

Immunoglobulin can inhibit B cell differentiation by cross-linking (receptor cross-linking)

IgG antibody is also known to have an effect that is Fc dependent. Immunoglobulin can inhibit B cell differentiation by cross-linking the antigen receptor with the Fc receptor (FcγRIIb) on the same cell (see Fig. 11.5). In this case, the antibodies may recognize different epitopes.

Increasing average antibody affinity

Doses of IgG that are insufficient to inhibit the production of antibodies completely have the effect of increasing the average antibody affinity because only those B cells with high-affinity receptors can successfully compete with the passively acquired antibody for antigen. For this reason, antibody feedback is believed to be an important factor driving the process of **affinity maturation** (Fig. 11.6).

Immune complexes may enhance or suppress immune responses

One of the ways in which antibody (either IgM or IgG) might act to modulate the immune response involves an Fc-dependent mechanism and immune complex formation with antigen.

Immune complexes can inhibit or augment the immune response (Fig. 11.7). By activating complement, immune complexes may become localized via interactions with CR2 on follicular dendritic cells (FDCs). This could facilitate the immune response by maintaining a source of antigen.

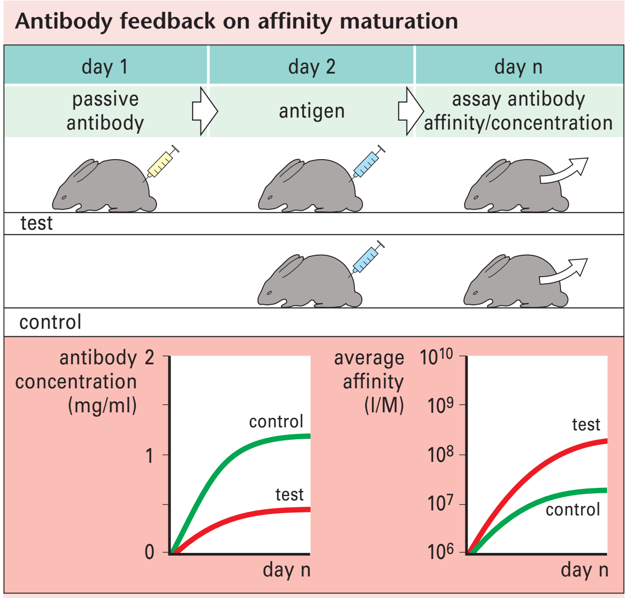

Fig. 11.6 The effect of passive antibody on the affinity and concentration of secreted antibody. One of two rabbits was injected with antibody (passive antibody) on day 1. Both rabbits were immunized with antigen on day 2 and the affinity and concentration of antibody raised to this antigen were assayed at a later time (day n). The antibody assay results show that passive antibody reduces the concentration, but increases the affinity of antibody produced.

Q. How does the presence of antigen on FDCs facilitate the immune response?

A. The antigen is available for uptake by B cells and presentation to T cells within the follicle, a process required for **class switching** and affinity maturation (see Fig. 8.12).

CR2 is also expressed on B cells and, as co-ligation of CR2 with membrane IgM has been shown to activate B cells, immune complex interaction with CR2 of the B cell–co-receptor complex and membrane Ig might lead to an enhanced specific immune response.

The immune response of patients with malignant tumors is often depressed, and it has been postulated that this is the result of the presence of circulating immune complexes composed of antibody and tumor cell antigens.

Idiotypic interactions may enhance or suppress antibody responses

Individual T cell receptors (TCRs) and immunoglobulins are immunogenic by virtue of unique sequences within their variable regions known as **idiotypes**. Antibodies formed against these antigen-binding sites are called **anti-idiotypic antibodies**, and are capable of influencing the outcome of an immune response.

Idiotypic determinants may be:

- encoded in the germline V region genes; or
- generated by the process of recombination and mutation involved in producing functional V region elements (see Chapter 5).

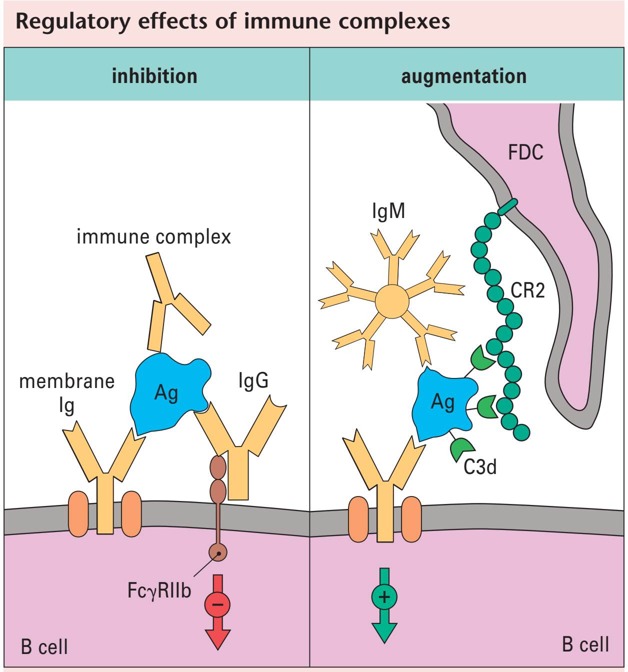

Fig. 11.7 Immune complexes can act either to inhibit or to augment an immune response. Inhibition – when the Fc receptor of the B cell is cross-linked to its antigen receptor by an antigen–antibody complex, a signal is delivered to the B cell, inhibiting it from entering the antibody production phase. Passive IgG may have this effect. Augmentation – antibody encourages presentation of antigen to B cells when it is present on an APC, bound via Fc receptors or, in this case, complement receptors (CR2) on a follicular dendritic cell (FDC). Passive IgM may have this effect.

Immunogenic epitopes in or around the binding site are termed **idiotopes** (Fig. 11.8).

Jerne proposed that an immune network existed within the body that interacted by means of idiotype recognition. According to this proposition, when an antibody response is induced by antigen, this antibody will in turn evoke an anti-idiotypic response to itself. This hypothesis is conceptually interesting, but the role of such an idiotype network in controlling a normal immune response remains unclear.

There is good evidence that anti-idiotypes can affect the representation of recognized idiotypes in an immune response. For example, when strain C57Bl/6 mice are challenged with the hapten nitrophenyl (NP), they produce antibodies that are largely restricted to a few defined idiotypes. The anti-idiotype 146 can enhance or suppress the production of idiotype 146 when the mice are subsequently challenged with NP on a carrier protein. The observed effect depends on the amount of anti-idiotype given and is idiotype specific; the overall level of anti-NP antibody is hardly affected. Most importantly, the amounts of anti-idiotype used are within the normal physiological range for particular idiotype-bearing antibodies, which suggests that idiotypic regulation may occur in vivo.

Idiotopes associated with the antibody-combining site

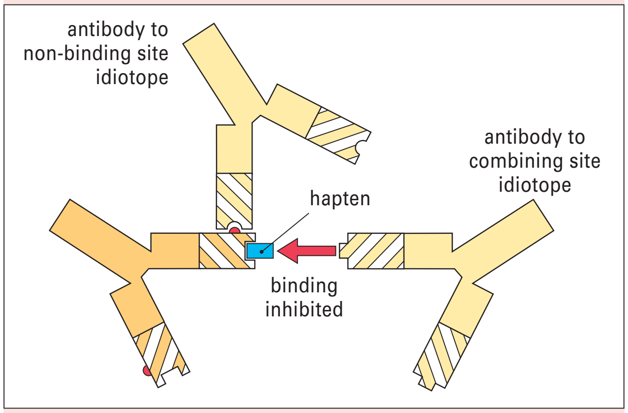

Fig. 11.8 An anti-idiotype serum may contain antibodies directed to various sites on the immunoglobulin molecule. The sites associated with the combining site are site-associated idiotopes. Binding to these can be inhibited by hapten. Antibodies to non-binding site idiotopes (non-site associated) are not inhibited by hapten.

T CELLS REGULATE THE IMMUNE RESPONSE

Cytokines are part of an extracellular signaling network that controls every function of the innate and adaptive immune system and have:

- numerous effects on cell phenotypes; and
- the ability to regulate the type of immune response generated and its extent.

Differentiation into $CD4^+$ TH subsets is an important step in selecting effector functions

A single TH cell precursor is able to differentiate into either a TH1 or a TH2 phenotype. The TH1/TH2 decision is crucial to the effective immunity and it is likely that many factors contribute to that decision. Factors that may influence the differentiation of TH cells include:

- the sites of antigen presentation;
- co-stimulatory molecules involved in cognate cellular interactions;
- peptide density and binding affinity – high MHC class II peptide density favors TH1, low densities favor TH2;
- APCs and the cytokines they produce;
- the cytokine profile and balance of cytokines evoked by antigen;
- activity of co-stimulatory molecules and hormones present in the local environment;
- host genetic background.

The cytokine balance is one of the major stimuli.

IL-12 is a potent initial stimulus for IFNγ production by T cells and natural killer (NK) cells and therefore regulates TH1 differentiation.

IFNα, a cytokine produced early during viral infection, induces IL-12 and can also switch cells from a TH2 to a TH1 profile.

By contrast, early production of IL-4 favors the generation of TH2 cells.

If polarizing signals are not present, $CD4^+$ TH0 cells have a less differentiated cytokine profile and represent a heterogeneous population with individual clones that can differentiate along the TH1 or TH2 pathway.

Cytokines from TH1 cells inhibit the actions of TH2 cells and vice versa. Thus, cross-regulation of TH subsets has been demonstrated whereby IFNγ secreted by TH1 cells can inhibit the responsiveness of TH2 cells (Fig. 11.9) whereas IL-10 produced by TH2 cells downregulates B7 and IL-12 expression by APCs, which in turn inhibits TH1 activation.

The TH1/TH2 balance is modulated not only by the level of expression of IL-12, but also by expression of the IL-12 receptor (IL-12R). The high-affinity IL-12R is composed of two chains, β1 and β2, with both chains being expressed only in TH1 cells. Both TH1 and TH2 cells express the β1 chain, but expression of the β2 chain is induced by IFNγ and inhibited by IL-4 (Fig. 11.10).

Differentiation of $CD4^+$ TH cells

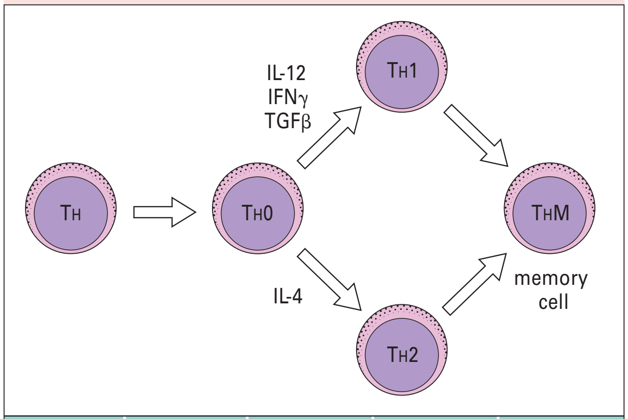

TH	TH0	TH1	TH2	THM
IL-2	IL-2	IL-2		IL-2
	IFNγ	IFNγ		
	TNFβ	TFNβ		
	IL-3	IL-3		
	IL-4		IL-4	
	IL-5		IL-5	
	IL-6		IL-6	
	IL-9		IL-9	
	IL-10		IL-10	
	IL-13		IL-13	
	GM-CSF	GM-CSF	GM-CSF	
	TNFα	TNFα	TNFα	

Fig. 11.9 The diagram illustrates the differentiation of murine TH cells. IL-12, IFNγ, and transforming growth factor-β (TGFβ) favor differentiation of TH1 cells, and IL-4 favors differentiation of TH2 cells. (GM-CSF, granulocyte–macrophage colony stimulating factor; TNF, tumor necrosis factor)

Regulation of the TH1 response by IL-12

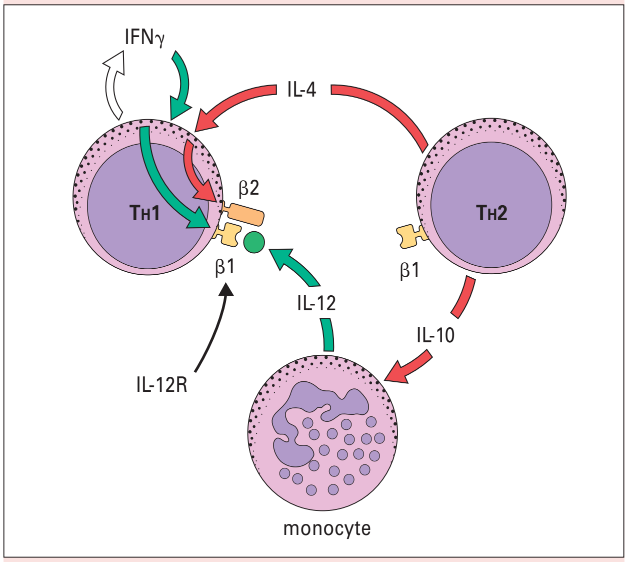

Fig. 11.10 The high-affinity IL-12R consisting of the β1 and β2 chains is expressed only on TH1 cells. IL-12 released by mononuclear phagocytes promotes the development and activation of TH1 cells, but IL-12 production is inhibited by IL-10 released by TH2 cells. IFNγ from TH1 cells promotes production of the β1 chain and therefore production of the high-affinity IL-12R. However, this is inhibited by IL-4.

An immune response therefore tends to settle into a TH1 or TH2 type of response.

Immune responses are not always strongly polarized in this way and the phenomenon works more closely in mice than humans.

TH1 and TH2 cells are best considered as extremes on a scale, but TH1 and TH2 responses do play different protective roles and may contribute to immunopathology.

Q. Apart from the cross-regulation by cytokines described above, in what other ways can a TH1-type or TH2-type immune response reinforce itself?

A. The chemokines induced by TH1 and TH2 cells tend to induce the accumulation of the same sets of lymphocytes and their associated effector cells (see Chapter 6).

TH cell subsets determine the type of immune response

It is clear that:

- local patterns of cytokine and hormone expression help to select lymphocyte effector mechanism; and
- the polarized responses of $CD4^+$ TH cells seem to be based on their profile of cytokine secretion (Fig. 11.11).

Q. In what way do cytokines that are normally associated with TH1 responses (IFNγ, IL-2) affect B cell differentiation?

A. IL-2 promotes B cell division. IFNγ promotes affinity maturation and class switching to IgG2a (see Chapter 8, Figs 8.9 and 8.10) which acts as an opsonin and fixes complement.

Selection of effector mechanisms by TH1 and TH2 cells

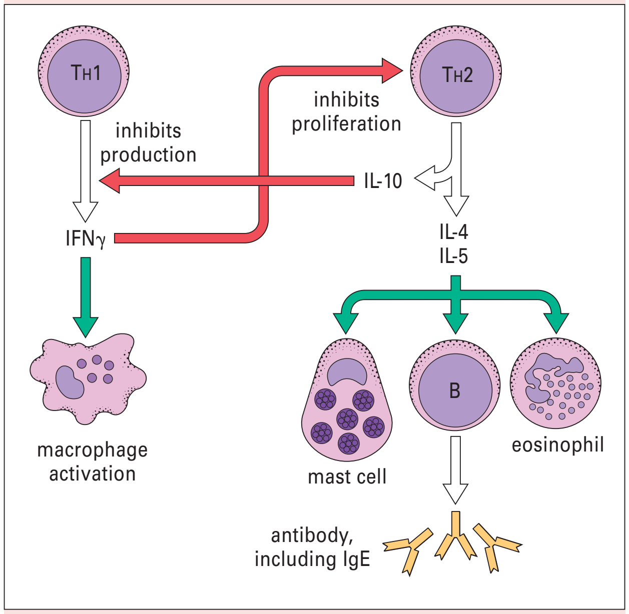

Fig. 11.11 The cytokine patterns of TH1 and TH2 cells drive different effector pathways. TH1 cells activate macrophages and are involved in antiviral and inflammatory responses. TH2 cells are involved in humoral responses and allergy.

TH1 cytokines including IFNγ, TNFβ, and IL-2 (see Fig. 11.9) also promote:

- macrophage activation;
- antibody-dependent cell-mediated cytotoxicity; and
- delayed-type hypersensitivity.

TH2 clones are typified by production of IL-4 and IL-5, with IL-6, IL-9, IL-10, and IL-13 also commonly produced (see Fig. 11.9). These cells provide optimal help for humoral immune responses biased towards:

- IgG1 and IgE isotype switching;
- mucosal immunity;
- stimulation of mast cell and eosinophil growth and differentiation; and
- IgA synthesis.

Therefore, in essence, TH1 cells are associated with cell-mediated inflammatory reactions and TH2 cells are associated with strong antibody and allergic responses.

$CD8^+$ T cells can be divided into subsets on the basis of cytokine expression

Many $CD8^+$ CTLs make a spectrum of cytokines similar to TH1 cells and are termed CTL1 cells. $CD8^+$ cells that make TH2-associated cytokines are associated with regulatory functions. The differentiation of these cells may be affected by the $CD4^+$ cell cytokine profile, with CTLs commonly associated with TH1 responses and less commonly found when TH2 cells are present. Thus:

- IFNγ and IL-12 may encourage CTL1 generation; and
- IL-4 may encourage CTL2 generation.

However, both CTL1 and CTL2 cells can be cytotoxic and kill mainly by a Ca^{2+}/perforin-dependent mechanism (see Figs 10.11 and 10.12).

CD4^{+} Tregs can be derived in two ways

Although T cells modulate the immune response in a positive sense by providing T cell help, which may favor either cell-mediated (TH1) or humoral (TH2) immunity as discussed above, T cells are also capable of down-regulating immune responses, and both CD4^{+} and CD8^{+} T cell subsets can induce this inhibition.

A naturally occurring population of CD4^{+}CD25^{+} regulatory T cells (Tregs) is generated in the thymus. Additionally, CD4^{+} Tregs can be induced from non-regulatory T cells in the periphery. These induced Tregs are a heterogeneous population of cells with a regulatory function.

Although Tregs do not prevent initial T cell activation, they inhibit a sustained response and prevent chronic and potentially damaging immunopathology.

The generation and mechanisms of suppression by the different CD4^{+} Treg populations vary, but:

- both populations are anergic (non-responsive) in vitro;
- neither have characteristics of TH1 or TH2 CD4^{+} cells; and
- they both inhibit cellular responses.

The characteristics of both cell types are compared in Fig. 11.12.

CD4^{+}CD25^{+} Tregs are naturally occurring

The immunosuppressive functions of CD4^{+} cells were initially observed by adoptively transferring T cells depleted of CD25^{+} cells into immunodeficient mice. This resulted in multiorgan autoimmunity suggesting that CD25^{+} cells play an important role in preventing self-reactivity. When the CD25^{+} T cells were replaced, autoimmune disease was prevented.

CD4^{+}CD25^{+} Tregs constitute 5–10% of peripheral CD4^{+} cells in both mice and humans. The presence of these cells in cord blood suggests that they have a thymic origin. In support of this notion, CD4^{+}CD8^{-}CD25^{+} cells from the thymus can suppress T cell proliferation, suggesting that they are a naturally occurring, distinct lineage of cells.

Q. What function does CD25 serve on developing thymocytes and on activated T cells?

A. It is part of the IL-2R (see Chapter 7 and Fig. 7.23), which is required for cell proliferation.

Naturally occurring CD4^{+}CD25^{+} Tregs are educated in the thymus during the negative selection phase and then exit the thymus as a regulatory population.

Comparison of CD4^{+}CD25^{+} Tregs with naive and activated CD4^{+} T cells shows that regulatory cells selectively express **Foxp3**, a member of the forkhead/winged helix transcription factors (Fig. 11.13) essential for the development and function of CD4^{+}CD25^{+} Tregs. Mutations in the *Foxp3* gene cause immune dysregulation, polyendocrinopathy enteropathy, X-linked syndrome (IPEX). Individuals with this disease have increased autoimmune and inflammatory diseases.

The importance of Foxp3 in the development of CD4^{+}CD25^{+} Tregs was underlined following transduction of Foxp3 into naive T cells (which do not express Foxp3). This increased expression of CD25 and induced suppressor function.

The study of CD4^{+}CD25^{+} Tregs has been difficult because of problems in separating these cells from activated T cells, which also express CD25.

Comparison of CD4^{+} regulatory T cells

CD4^{+} CD25^{+} Treg	Tr1/TH3
differences	
generated in the thymus	generated in the periphery
express Foxp3	variable expression of Foxp3
CD25 high, CD45R low	CD45R0 high, CD45RB low variable CD25 expression
cell contact dependent	cell contact independent
require IL-2 for suppression	IL-2-independent suppression
similarities	
triggered by specific antigen	triggered by specific antigen
secrete IL-10/TGFβ	secrete IL-10/TGFβ
unresponsive to anti-CD3 in vitro	unresponsive to anti-CD3 in vitro

Fig. 11.12 This table outlines the similarities and differences in generation and function of the subsets of naturally occurring thymic and peripherally induced CD4^{+} regulatory T cells.

Foxp3 is selectively expressed in Tregs

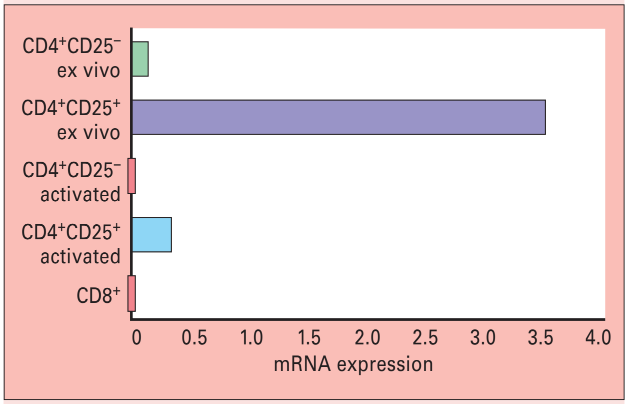

Fig 11.13 Cell populations were collected and real-time polymerase chain reaction (PCR) was performed to assess levels of Foxp3. Foxp3 mRNA is expressed selectively in CD4$^+$CD25$^+$ cells ex vivo. It was expressed at low levels in activated CD4$^+$CD25$^+$ cells and was not present in CD25$^-$ cells or CD8$^+$ cells. (Adapted from Fontenot JD, Gavin MA, Rudensky AY. Nat Immunol 2003;4:330–336)

The expression of a number of cell surface molecules on CD4$^+$CD25$^+$ Tregs has been examined and include:

- CTLA-4;
- CD45RB (which is weakly expressed on CD4$^+$CD25$^+$ cells); and
- GITR (glucocorticoid-induced TNF-related receptor).

However, none of these cell surface molecules exclusively identifies the naturally occurring Treg cell population, and CTLA-4, for example, is expressed in both activated T cells and Tregs.

Thymic-derived CD4$^+$CD25$^+$ Tregs work in a cell contact-dependent manner

Naturally occurring CD4$^+$CD25$^+$ Tregs respond to numerous environmental stimuli.

IL-2 is required for suppression, suggesting that CD25 (the IL-2Rα chain, see Fig. 7.23) has a functional role in these cells.

CD4$^+$CD25$^+$ Treg-induced suppression in vitro is cell contact dependent. It is thought that the GITR might be involved in this interaction, but the exact nature is not known.

Induced CD4$^+$ regulatory T cells can be generated following exposure to antigen

In addition to naturally occurring CD4$^+$CD25$^+$ Tregs, which are generated in the thymus, some populations of CD4$^+$ regulatory T cells can be induced to develop in the periphery. Thymectomized mice are able to generate regulatory cells, therefore there must be a thymic-independent induction of these cells in the periphery. For example they arise:

- following allograft transplantation; and
- after oral administration of antigen.

CD4$^+$ regulatory T cells can also be generated from antigen-experienced cells in the presence of certain cytokines. Upon stimulation with IL-10, antigen-experienced cells can differentiate into **Tr1 cells** – these cells do not express CD25 or Foxp3 and they mediate suppression through the secretion of IL-10.

Alternatively, if induced CD4$^+$ regulatory T cells are generated in the presence of TGFβ, they express Foxp3 and upregulate expression of CD25. These induced regulatory cells mediate suppression via production of TGFβ (Fig. 11.14).

Although induced CD4$^+$CD25$^+$ Tregs have an overlapping phenotype with thymus-derived CD4$^+$CD25$^+$ Tregs, they are a different population of cells and have different mechanisms for inducing suppression.

The role of CD4$^+$ Tregs in infections is not clear

Although CD4$^+$ Tregs play a vital role in the prevention of autoimmune diseases (see Chapter 20), their role in infections is less clear.

CD4$^+$ Tregs have a protective role against immune-mediated pathology and their ability to suppress both TH1 and TH2 responses is important in reducing inflammation. For example, lesions in the eye in stromal keratitis are less severe in the presence of CD4$^+$CD25$^+$ Tregs.

CD4$^+$ Tregs can also suppress virus-specific responses. Many pathogens induce high levels of IL-10 and TGFβ, which promote the induction of CD4$^+$CD25$^+$ Tregs. In chronic viral infections, including HIV, cytomegalovirus

Regulatory T cells can be generated in the periphery

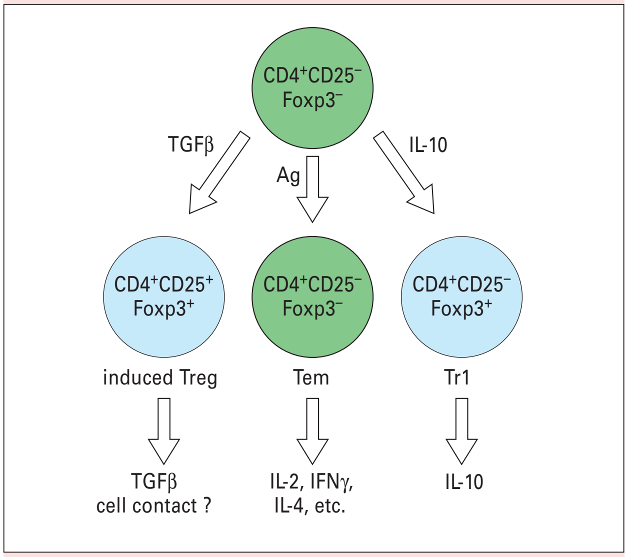
CD4$^+$CD25$^-$ Foxp3$^-$
TGFβ
Ag
IL-10
CD4$^+$CD25$^+$ Foxp3$^+$
CD4$^+$CD25$^-$ Foxp3$^-$
CD4$^+$CD25$^-$ Foxp3$^+$
induced Treg
Tem
Tr1
TGFβ
cell contact ?
IL-2, IFNγ,
IL-4, etc.
IL-10

Fig 11.14 CD4$^+$CD25$^-$ cells in the periphery can be stimulated to become suppressive. If TGFβ is present cells upregulate Foxp3 and CD25. In the presence of IL-10 cells become suppressive, but they do not express Foxp3 or CD25. Induced Tregs suppress via cytokine-dependent mechanisms. In the presence of antigen and co-stimulation, effector T cells are generated. (Tem, T effector memory cells)

(CMV), and herpes simplex virus (HSV) infections, increased numbers of Tregs are responsible for decreased antigen-specific responses by $CD4^+$ and $CD8^+$ T cells. This can lead to disease development.

$CD8^+$ T cells negatively regulate secondary immune responses

$CD8^+$ cells can be induced to become suppressive.

$CD8^+$ Tregs are generated in vitro by stimulating highly differentiated $CD8^+$ cells with antigen and IL-10. Like $CD4^+CD25^+$ cells, $CD8^+$ Tregs express the forkhead protein Foxp3.

Many $CD8^+$ Tregs work in a cell contact-dependent manner. They cause the downregulation of co-stimulatory molecules on dendritic cells and endothelial cells. This induces tolerance because upon the interaction with the TCR and peptide there is insufficient co-stimulation to generate a functional immune response.

In mice the MHC class Ib molecule, Qa-1, is crucial for suppression. Mice lacking Qa-1 are unable to suppress responses. However when Qa-1 is replaced into these cells their ability to suppress immune responses is restored (Fig. 11.15).

$CD8^+$ Tregs:

- need to be primed by $CD4^+$ cells during a primary response; they then suppress secondary immune responses;
- can downregulate both TH1 and TH2 responses and also control $CD4^+$ responses to superantigen;
- can suppress immune responses through the secretion of cytokines including IFNγ, IL-6, and IL-10.

NK cells produce immunoregulatory cytokines and chemokines

NK cells make cytokines and chemokines and therefore play an important role in the innate immune response to infections and tumors.

CD4 Qa-1 is essential for $CD8^+$ T suppressor function

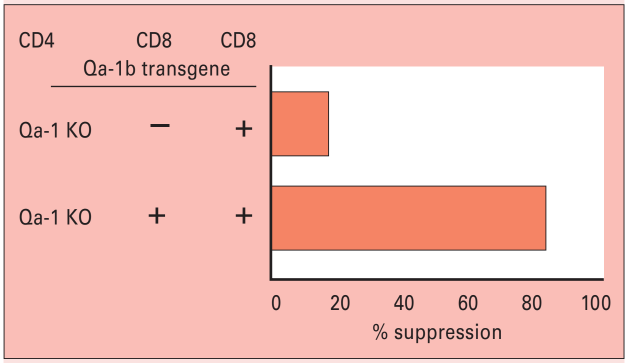

Fig. 11.15 Qa-1 knockout (KO) mice were generated. Qa-1 KO $CD4^+$ cells were mixed with $CD8^+$ cells and the level of suppression was assessed. $CD8^+$ cells could not suppress CD4 cells lacking Qa-1 expression. Insertion of the Qa-1b allele into $CD4^+$ cells resulted in effective suppression of responses. (Adapted from Hu D, Ikizawa K, Lu L, et al. Nat Immunol 2004;5:516–523)

The production of immunoregulatory cytokines and chemokines at early stages in the immune response influences the characteristics of the subsequent adaptive immune response and can therefore influence the outcome of the immune response.

NK cells play a key role in the early immune response to intracellular pathogens, largely through their production of IFNγ, which activates macrophages and facilitates differentiation of TH1 cells.

NK cell activity itself is induced by a variety of cytokines including:

- IFNα/β;
- IL-15;
- IL-18; and
- IL-12.

NK cells in turn are negatively regulated by cytokines such as IL-10 and TGFβ.

Q. Which cells will tend to promote the development of NK cells?

A. The TH1 population and macrophages through the production of IL-12, IL-15, and IL-18, whereas the TH2 population will tend to inhibit their development through the production of IL-10.

NK T cells produce cytokines when their TCR engages glycolipids in association with CD1d (see Chapter 10). It has been suggested that these cells play an immunoregulatory role in the control of autoimmunity, parasite infection, and tumor cell growth.

Recent experiments suggest that NK T cells secreting IFNγ are able to induce NK cell activation, increasing both NK proliferation and cytotoxicity. They are capable of making both TH1-type (IFNγ) and TH2-type (IL-4) cytokines depending on the cytokines present in the microenvironment when they are activated (Fig. 11.16).

The presence of IL-7, for example, has been shown to elicit IL-4 production in NK T cells and therefore to promote a TH2 response. This ability to make IL-4, particularly in the thymus, has been associated with the prevention of autoimmunity. For example, non-obese diabetic (NOD) mice have a deficit in NK T cells, and injection of NK T cells into these mice prevents the spontaneous development of autoimmune diabetes.

Deficiencies of NK T cells have also been reported in human autoimmune diseases including:

- rheumatoid arthritis;
- psoriasis;
- ulcerative colitis; and
- multiple sclerosis.

SENESCENCE OF CELLS IS REGULATED BY TELOMERE EROSION

Cells age by the process of cellular senescence, causing growth arrest and limiting their proliferative life span. This influences the extent of the immune response. The mechanisms that control cell senescence are an important part of immune regulation.

NK T cells produce both TH1- and TH2-type cytokines

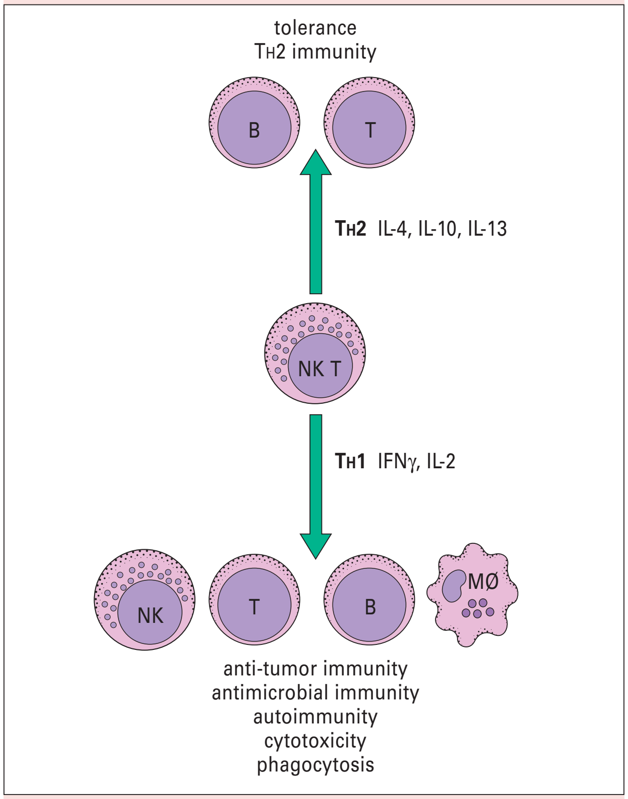

Fig. 11.16 NK T cells exert effects on many cell types. They can produce TH1 cytokines and TH2 cytokines and are involved in all aspects of the immune response.

Telomeres visualized by confocal microscopy

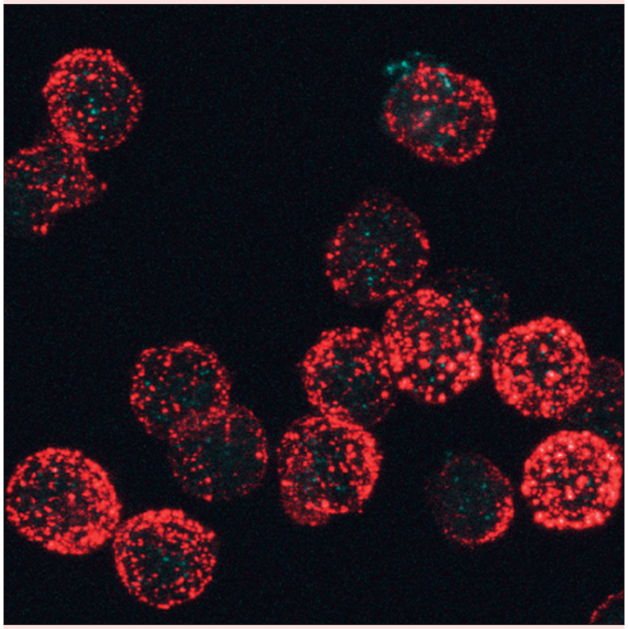

Fig. 11.17 The surface of living CD4 cells was stained with a red anti-CD4 antibody and then treated to destroy subcellular membranes, following which the telomeres were hybridized with a blue probe specific for the telomere sequences and visualized by confocal microscopy. Telomeres cap the ends of eukaryotic chromosomes and are repeating GGGTTA sequences at the ends of chromosomes with up to 2000 copies per cell. They protect chromosome stability and integrity, and are a mechanism for replicating DNA ends. (Courtesy of Dr J Skok and Dr JM Fletcher)

Telomeres are repeats of the DNA sequence GGGTTA and protein located at the end of chromosomes (Fig. 11.17).

Telomeres provide stabilization to chromosomes and protect the chromosomal ends from damage; they are also important in regulating the replicative capacity of cells.

During every cell division telomeres shorten by 50–100 base pairs because DNA polymerase cannot replicate the ends of chromosomes. The mechanism of telomere shortening prevents uncontrolled cell proliferation, which could lead to cancer.

When telomeres become too short the chromosome becomes unstable and DNA damage can occur. To prevent damaged cells being replicated, cells with unstable chromosomes:

- die by **apoptosis**; or
- enter cell cycle arrest, also known as **cellular senescence**.

Studies on T cells in elderly individuals show that they have significantly shorter telomeres than young individuals (Fig. 11.18).

In addition individuals with some premature ageing syndromes have short telomeres, are more susceptible to infections, and usually have a low life expectancy.

The decrease in telomere length with age is reflected in the increased exposure to antigens that elderly people have encountered. Memory T cells have shorter telomeres than naive T cells because they can have undergone many cell divisions (see Fig. 11.18). However, recent studies have shown that the extent of telomere erosion is dependent on which antigens are encountered. For example, CMV-specific cells have shorter telomeres than those of other antigen-specific cells, including EBV and VZV. Too much telomere erosion leads to senescence (i.e. effective loss of cells through cell cycle arrest) and this may cause the loss of specific populations of cells during ageing.

Cellular senescence can be delayed by maintaining telomere length; this occurs through the action of the enzyme **telomerase**. Telomerase contains an RNA template for the telomeric sequence GGGTTA. The enzyme is a reverse transcriptase that adds multiple telomere sequences to the 3′ end of replicated DNA and prevents telomere shortening (Fig. 11.19).

T cells can upregulate the expression of telomerase in certain situations, including during antigen stimulation. This prevents telomere loss during clonal expansion of antigen-specific cells and ensures that cells entering the memory pool retain their function. However, after repeated stimulation the ability to upregulate telomerase

Telomere erosion in ageing and antigen-specific memory

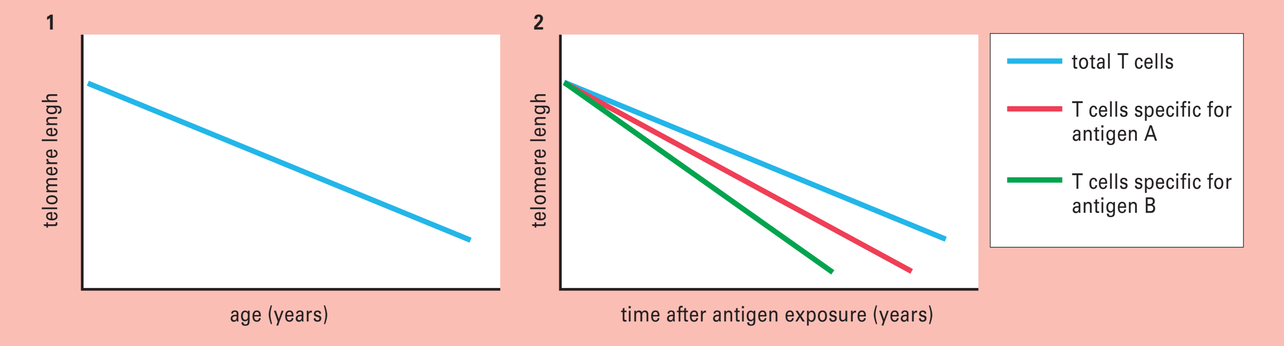

Fig. 11.18 (1) There is a reduction in telomere length in total T cells during ageing. (2) Continuous antigen exposure results in telomere erosion. Cells specific for CMV (antigen A) have shorter telomeres than those specific for other antigens including Epstein–Barr virus (EBV) and varicella zoster virus (VZV). (Adapted from Akbar AN, Beverley PCL, Salmon M. Nat Rev Immunol 2004;4:737–743. Copyright 2004 with permission from Nature Reviews. www.nature.com/reviews)

Telomere elongation by telomerase

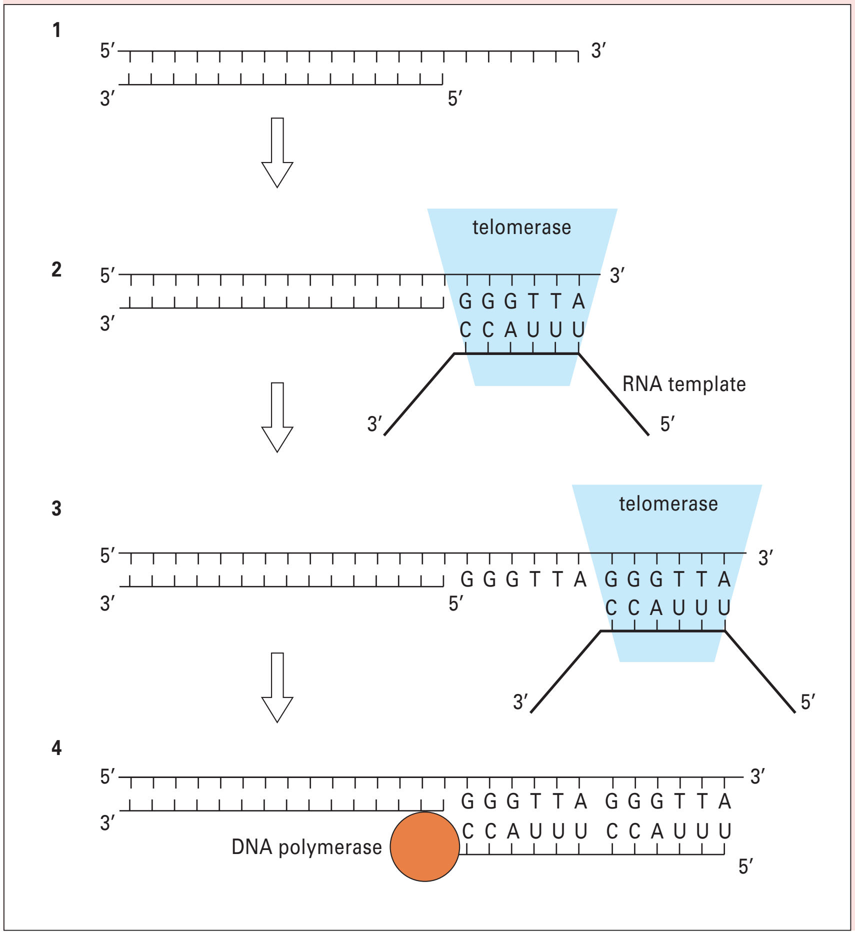
1
5′
3′
3′
5′
2
telomerase
5′
3′
3′
G G G T T A
C C A U U U
RNA template
3′
5′
3
telomerase
5′
3′
3′
5′
G G G T T A G G G T T A
C C A U U U
3′
5′
4
5′
3′
3′
G G G T T A G G G T T A
C C A U U U C C A U U U
DNA polymerase
5′

Fig. 11.19 (**1**) DNA replication results in incomplete transcription of DNA. (**2**) Telomerase is a ribonucleoprotein. The RNA strand encodes CCAUUU and acts as a template for extension of the telomere. (**3**) Telomerase translocates along the DNA from a 5′ to 3′ direction with addition of multiple telomeric sequences. (**4**) Transcription of the opposite strand is then completed by DNA polymerase.

decreases, leading to the telomere erosion observed in memory T cell populations.

Apoptosis maintains homeostasis in the immune system

Apoptosis is a cellular clearance mechanism through which homeostasis is maintained.

Unlike cell damage-induced death (i.e. **necrosis**), which can trigger immune responses, apoptosis maintains intracellular structures within the cell.

Apoptopic cells undergo nuclear fragmentation and the condensation of cytoplasm, plasma membranes, and organelles into **apoptopic bodies** (Fig. 11.20).

Apoptopic cells are rapidly phagocytosed by macrophages, which prevents the release of toxic cellular components into tissues, so avoiding immune responses to the dead cells.

Apoptosis is programmed cell death

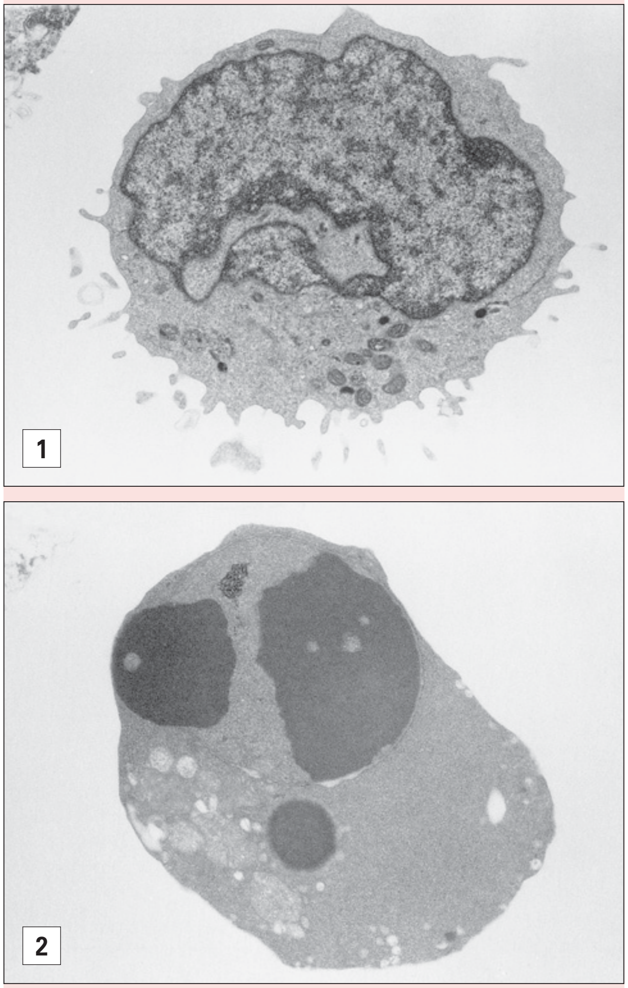

Fig. 11.20 (**1**) Normal cell. (**2**) Apoptotic cell. Apoptosis involves nuclear fragmentation, with 'blebbing' often observed. There is also condensation of cytoplasm and plasma membranes.

Apoptosis is:

- involved in clearing cells with a high avidity for antigen in the thymus and is an important mechanism of immunological tolerance (see Chapter 19);
- an important mechanism in maintaining homeostasis in the immune system.

Following resolution of an immune response the majority of antigen-specific cells die by apoptosis. This ensures that no unwanted effector cells remain and also maintains a constant number of cells in the immune system.

A small number of cells are prevented from undergoing apoptosis and enter the memory T cell pool.

Apoptosis is controlled by a number of factors in the cell and depends on expression of the death trigger molecule **CD95 (Fas)**.

Q. How does ligation of Fas lead to apoptosis?

A. It causes activation of a series of inducer and effector caspases (see Chapters 10 and 12 and Fig. 10.9).

Expression of the anti-apoptotic molecule **Bcl-2** makes cells more resistant to cell death. Memory T cells generally express high levels of Bcl-2, which may contribute to the rescue of memory populations from apoptosis.

Different cell populations can express both pro- and anti-apoptotic molecules with the balance of these molecules determining whether a cell survives to participate in immune responses.

SELECTIVE MIGRATION OF LYMPHOCYTE SUBSETS CAN MODULATE THE IMMUNE RESPONSE

Chemokines regulate the migration of lymphocytes

The spatial and temporal production of chemokines by different cell types is an important mechanism of immune regulation.

There is good evidence to suggest that the recruitment of TH1 and TH2 cells is differentially controlled, thereby ensuring the maintenance of locally polarized immune responses.

The expression of different chemokine receptors on TH1 cells (CXCR3 and CCR5) and TH2 cells (CCR3, CCR4, CCR8) allows chemotactic signals to produce the differential localization of T cell subsets to sites of inflammation (see Fig. 6.11).

Chemokines can be induced by cytokines released at sites of inflammation, so providing a mechanism for local reinforcement of particular types of response (Fig. 11.21). Once a response is established the T cells can induce further migration of appropriate effector cells. This is clearly illustrated in TH1 responses where the secondary production of CCL2, CCL3, CXCL10, and CCL5 serves to attract mononuclear phagocytes to the area of inflammation.

The ability of cytokines such as TGFβ, IL-12, and IL-4 to influence chemokine or chemokine receptor expression provides a further level of control on cell migration or recruitment.

Mechanisms for local reinforcement of different modes of immune response

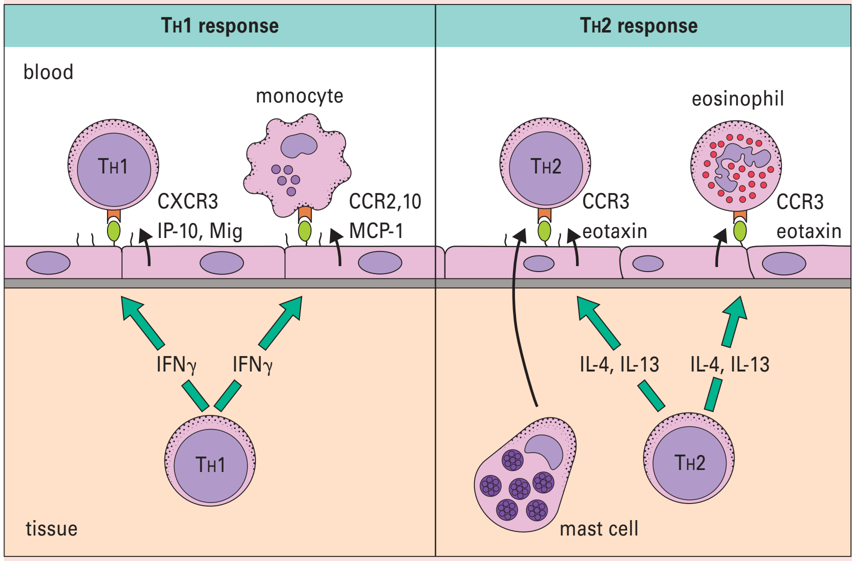

Fig. 11.21 Activated TH1 cells release IFNγ, which induces the chemokines IP-10 and Mig. These act on the chemokine receptors CXCR3, which are selectively expressed on TH1 cells, thereby reinforcing this type of response. Macrophage chemotactic protein-1 (MCP-1), which attracts macrophages and monocytes, is also induced by IFNγ. Mast cells release eotaxin when activated, and endothelial cells and bronchial epithelium can also synthesize this chemokine in response to IL-4 and IL-13 from TH2 cells. Eotaxin acts on CCR3, which is selectively expressed on TH2 cells, thereby reinforcing the TH2 response. Eosinophils and basophils, which mediate allergic responses in airways, also express CCR3. Chemokines can therefore potentiate both the initiation and effector phases of a specific type of immune response.

Immune responses do not normally occur in immune-privileged sites

Immune responses do not normally occur at certain sites in the body such as the anterior chamber of the eye and the testes. These sites are called **immune privileged**.

The failure to evoke immune responses in these sites is partly due to the presence of inhibitory cytokines such as TGFβ and IL-10, which inhibit inflammatory responses. The presence of migration inhibition factor (MIF) in the anterior chamber of the eye also inhibits NK cell activity.

The constitutive expression of FasL in cells of the testes and the eye has been proposed as an additional means of eliminating Fas-expressing lymphocytes that reach these sites through apoptosis.

T cell expression of different molecules can mediate tissue localization

Most studies on the human immune system are performed on blood due to the ethical issues in obtaining tissues. However, only a small proportion of the lymphocyte pool is circulating in the blood, most being resident in lymphoid or effector tissues.

The expression of molecules on T cells can mediate circulation through different tissues. Loss of the **lymph node homing molecules CCR7 and CD62L** on the surface of T cells prevents cells from circulating through lymphoid tissue – in this way a population of highly differentiated memory cells migrate to non-lymphoid sites where they can exert effector functions.

It has recently been suggested that there are two types of memory cell:

- **central memory cells** express CCR7, home to lymphoid tissues, and do not have immediate effector function;
- **effector memory cells** do not express CCR7, migrate to non-lymphoid tissues, and produce effector cytokines – effector cells in non-lymphoid tissues, such as the skin, can proliferate and senesce and this has a direct effect on the level of local immune responses.

Q. Why do cells that express CCR7 migrate to lymphoid tissues?

A. They respond to **CCL21** expressed only in secondary lymphoid tissues (see Fig. 6.17).

Mucosal surfaces contain specialized cells

Mucosal surfaces are the main point of contact for antigens entering the immune system.

There is an increased proportion of T cells expressing the γδ TCR at mucosal sites, and intraepithelial lymphocytes (IELs) in the gut express CD8 in addition to their γδ TCRs whose repertoire specificity is skewed towards bacterial antigens (see Chapter 5).

T cells and B cells in the lamina propria are thought to play a role in immune surveillance. They also enhance the humoral IgA response, which is an important aspect of mucosal immunity (see Chapter 12).

THE NEUROENDOCRINE SYSTEM INFLUENCES IMMUNE RESPONSES

It is known that stressful conditions can lead to the suppression of immune functions, for example reducing the ability to recover from infection.

The nervous, endocrine, and immune systems are interconnected (Fig. 11.22). There are two routes by which the central nervous system could modulate immune function:

- most lymphoid tissues receive direct sympathetic innervation to the blood vessels passing through the tissues and directly to lymphocytes;
- the nervous system directly and indirectly controls the output of various hormones, in particular corticosteroids, growth hormone, prolactin, α-melanocyte-stimulating hormone, thyroxine, and epinephrine (adrenaline).

Neuroendocrine interactions with the immune system

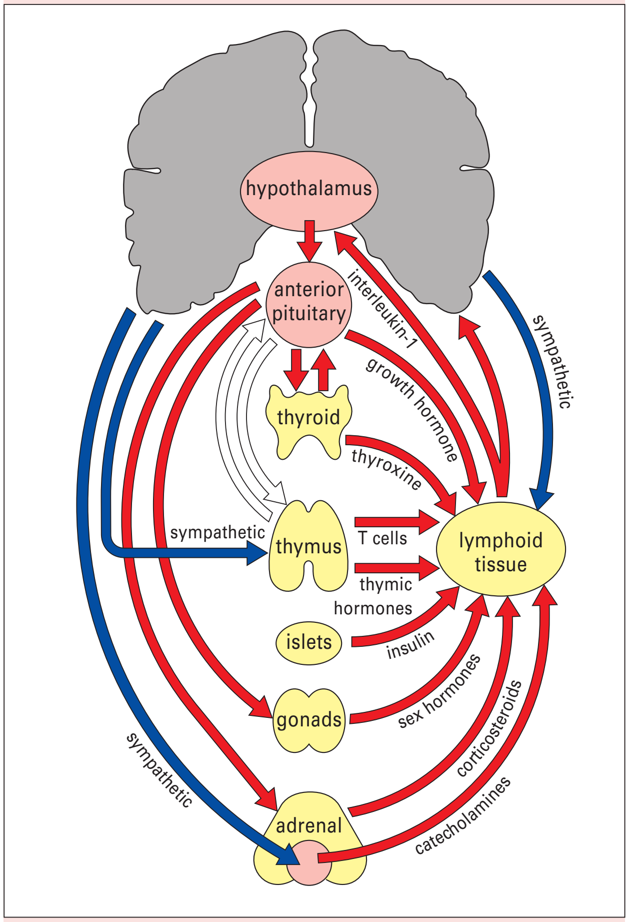

Fig. 11.22 The diagram indicates some of the potential connections between the endocrine, nervous, and immune systems. Blue arrows indicate nervous connections, red arrows indicate hormonal interactions, and white arrows indicate postulated connections for which the effector molecules have not been established.

Lymphocytes express receptors for many hormones, neurotransmitters, and neuropeptides – expression and responsiveness vary between different lymphocyte and monocyte populations, such that the effect of different transmitters may vary in different circumstances.

Corticosteroids, endorphins, and enkephalins, all of which may be released during stress, are immunosuppressive in vivo. The precise in-vitro effects of endorphins vary depending on the system and on the doses used – some levels are suppressive and others enhance immune functions.

It is certain, however, that corticosteroids act as a major feedback control on immune responses. It has been found that lymphocytes themselves can respond to corticotrophin releasing factor to generate their own adrenocorticotropic hormone (ACTH), which in turn induces corticosteroid release. Corticosteroids:

- inhibit TH1 cytokine production while sparing TH2 responses; and
- induce the production of TGFβ, which in turn may inhibit the immune response.

The low levels of plasma corticosteroids found in Lewis rats are believed to contribute to the susceptibility of this strain to a variety of induced autoimmune conditions.

The importance of corticosteroids in the overall susceptibility to disease induction is demonstrated in rat models of autoimmune disease where animals normally resistant to EAE become susceptible if adrenalectomized.

The interplay between the neuroendocrine system and the immune system is not unidirectional. Cytokines, in particular IL-1 and IL-6 produced by T cells, neurons, glial cells, and cells in the pituitary and adrenal glands, are bidirectional modulators of neuroendocrine–immune communication. These cytokines are potent stimulators of adrenal corticosteroid production through their influence on corticotrophin releasing hormone (CRH).

Q. In what other ways does IL-1 mediate interactions between the immune and nervous systems?

A. It causes an increase in body temperature, suppresses appetite, and enhances the duration of slow-wave sleep (see Chapter 6).

GENETIC FACTORS INFLUENCE THE IMMUNE SYSTEM

Familial patterns of susceptibility to infectious agents suggest that resistance or susceptibility might be an inherited characteristic. Such patterns of resistance and susceptibility are also shown with autoimmune diseases.

Often many genes are involved in governing susceptibility or resistance to disease and the disease is said to be under polygenic control.

Considerable advances have been made in mapping and identifying the genes governing the response to some diseases as a result of:

- the development of techniques such as microsatellite mapping;
- increased availability of DNA samples; and
- sequencing of the human genome.

In most cases these studies have led to the identification of potential candidate genes, but their real role in disease susceptibility remains to be clarified. In other cases single mutations in genes of known function have been found and the mechanism by which they contribute to disease identified.

MHC haplotypes influence the ability to respond to an antigen

The development of inbred mouse strains conclusively demonstrated that genetic factors have a role in determining immune responsiveness. For example, strains of mice with different MHC haplotypes vary in their ability to mount an antibody response to specific antigens (Fig. 11.23). This function depends on MHC class II molecules (see Chapter 5) and is specific for each antigen – a high responder strain for some antigens may be a low responder strain for others.

Strain differences in the antibody response

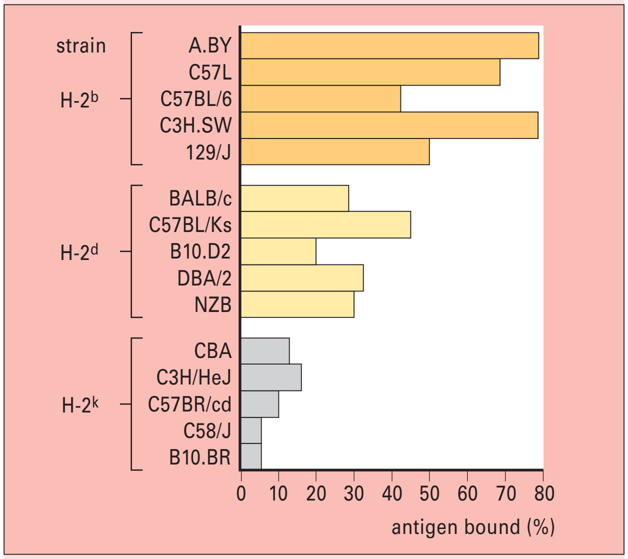

Fig. 11.23 Fifteen strains of mice were given a standard dose of the synthetic antigen (TG)-A-L. Antibody responses are expressed as the antigen-binding capacity of the sera. Animals of the H-2^b haplotype are high responders, whereas those of the H-2^d and H-2^k haplotypes are intermediate and low responders, respectively. However, there is some overlap between the levels of response in different haplotypes, indicating that H-2-linked genes are not the only ones controlling the antibody response.

MHC-linked genes control the response to infections

MHC-linked genes (see Chapter 5) are involved in the immune response to infectious agents. In some cases the gene involved is the MHC gene itself, but in others it can be a gene linked to the MHC.

Susceptibility to infection by Trichinella spiralis *is affected by the I-E locus in mice*

The first observation that genes within the MHC could influence the response to parasites involved the susceptibility to *Trichinella spiralis* infection.

The response of different recombinant mouse strains to infection with *T. spiralis* is affected by the I-E locus. Mouse strains that express I-E (B10.BR, B10.P) appear to be susceptible whereas those that do not (B10.S, B10.M) are resistant to infection (Fig. 11.24). The response to *T. spiralis* is also influenced by the MHC-linked gene *Ts-2*, which maps close to the *TNF* genes.

The I-E locus also influences susceptibility to *Leishmania donovani*

Using H-2 congenic mice, it was shown that mice expressing the I-E locus were susceptible to visceral leishmaniasis. Parasite clearance was enhanced by anti-I-E antibody, but not by anti-I-A antibody, showing direct involvement of the I-E product. Furthermore, insertion of an I-E transgene into mice lacking the locus prevented effective clearance of parasites from the liver and spleen when compared to the original strain.

Certain HLA haplotypes confer protection from infection

Comparison of HLA haplotypes in humans revealed that certain MHC class I and class II alleles (HLA-B*5301 and DRB1*1302, respectively) are associated with a reduced risk of severe malaria.

Susceptibility to *Trichinella spiralis*

mouse strain	H-2 haplotype	I-E expression	resistance index	resistance phenotype
B10.BR	k	+	0	sus
B10.P	p	+	–22	sus
B10.RIII	r	+	33	sus
B10	b	–	63	res/int
B10.S	s	–	100	res
B10.M	f	–	104	res
B10.Q	q	–	105	res

Fig. 11.24 Association of H-2 haplotype, expression of cell surface I-E molecules, and susceptibility to infection with *T. spiralis*. The resistance index is measured as the number of parasites present after a constant challenge relative to strains B10.BR (susceptible = 0% resistance) and B10.S (resistant = 100% resistance). B10 shows intermediate resistance.

DRB1*1302 binds different peptides from those bound by DRB1*1301 as a result of a single amino acid difference in the β chain. This change is sufficient to influence the response to the malaria parasite.

HLA-DRB1*1302 has also been associated with an increased clearance of the hepatitis B virus and consequently a decreased risk of chronic liver disease.

The MHC class I type, HLA-A*02, is associated with a reduced risk of disease development following human T-lymphotropic virus-1 (HTLV-1) infection. The viral load was lower in HLA-A*02-positive healthy carriers of HTLV-1 correlating with the presence of high levels of virus-specific CTLs.

In HIV-1 infection a selective advantage against disease has been noted in individuals expressing maximal HLA heterozygosity of class I loci (A, B, and C) and lacking expression of HLA-B*35 and HLA-Cw*04.

Polymorphisms in the promoter region of the *TNF*α gene, which lies within the MHC, influence its level of expression through altered binding of the transcription factor OCT-1. One of these polymorphisms, which is commonly associated with cerebral malaria, results in high levels of TNF expression. This may lead to upregulation of intercellular adhesion molecule-1 (ICAM-1) on vascular endothelium, increased adherence of infected erythrocytes, and subsequent blockage of blood flow.

This polymorphism in the TNFα promoter has also been associated with:

- lepromatous leprosy;
- mucocutaneous leishmaniasis; and
- death from meningococcal disease.

Many non-MHC genes also modulate immune responses

Some genes outside the MHC region also govern the immune response. However, these genes are generally less polymorphic than MHC genes and contribute less to variations in disease susceptibility. Nevertheless, their effects are found in autoimmune diseases, allergy, and infection. For example:

- individuals with defects in the complement components C1q, C1r, and C1s (see Chapter 4) are predisposed to develop systemic lupus erythematosus (SLE) and lupus nephritis – the development of SLE-like symptoms in C1q knockout mice parallels the human situation;
- deficiency in C3 leads to an increased susceptibility to bacterial infections and a predisposition to immune complex disease – this is also seen with a deficiency in C2 and C4, both of which are located within the MHC region;
- high IgE production in some allergy-prone families associates with the presence of an 'atopy gene' on human chromosome 11q;
- Biozzi generated two lines of mice based on their responsiveness to erythrocyte antigens. The high responder and low responder Biozzi mice make quantitatively different amounts of antibody in response to antigenic challenge and the basis for these differences has in part been attributed to genetic differences in macrophage activity; the high and low responder strains also differ in their ability to respond to parasitic infections, but this does not correlate with the level of antibody produced.

Q. Why would deficiency in the components of the classical pathway lead to immune complex disease?

A. C3b deposition is required to transport complexes on erythrocytes and their ultimate phagocytosis by macrophages in the liver and spleen (see Chapter 4).

Non-MHC-linked genes affect susceptibility to infection

Genes regulating macrophage activity may determine the outcome of immune responses

Macrophages have a key role in the immune system. Therefore genes regulating their activity may determine the outcome of many immune responses.

The *Lsh/Ity/Bcg* gene provides a good example of such genetic control of macrophage function. This gene governs the early response to infection with *Leishmania donovani*, *Salmonella typhimurium*, *Mycobacterium bovis*, *Mycobacterium lepraemurium*, and *Mycobacterium intracellulare*. Its influence is on the early phase of macrophage priming and activation, and it has wide-ranging effects, including:

- upregulation of the oxidative burst;
- enhanced tumoricidal activity;
- enhanced antimicrobial activity; and
- upregulation of MHC class II molecule expression.

Recent congenic studies have identified the **natural resistance-associated macrophage protein-1 (*Nramp1*)** as the *Bcg* gene. *Nramp1* encodes a membrane protein with homology to known transport proteins and may be implicated in the transport of nitrite (NO_2^-) into the phagolysosome, so facilitating the killing of intracellular organisms.

The human homolog of the mouse gene *Nramp* (*NRAMP1*) has been cloned and several different alleles identified. Polymorphisms in this gene may contribute to resistance to tuberculosis in humans, though the data so far are not as convincing as in the mouse.

Polymorphisms in cytokine and chemokine genes affect susceptibility to infections

Polymorphisms in the genes encoding cytokine receptors have been shown to correlate with an increased susceptibility to:

- infection;
- severe combined immune deficiency (SCID); and
- inflammatory conditions.

The outcome of the mutation is dependent on which cytokine gene is affected.

Q. What effect would you expect a deficiency in the IL-7 receptor (IL-7R) to produce?

A. It causes a reduction in T cell numbers because IL-7 is required for early thymocyte development (see Fig. 2.29).

For example, humans with:

- mutations in the IL-7R α chain have a reduced number of T cells;
- whereas those with deficiency in the common cytokine receptor γ chain (γc), a component of IL-2, IL-4, IL-7, IL-9, and IL-15 receptors, have reduced numbers of T and NK cells and impaired B cell function, in part attributable to the lack of T cell help (Fig. 11.25).

Further examples are the mutations in the IFNγ receptor (IFNγR) or IL-12 receptor (IL-12R), which increase susceptibility to mycobacterial infection.

A list of genetic defects that contribute to impaired immune responses is given in Fig. 11.26.

Mutations in the cytokine promoters influence the levels of expression of cytokines. Polymorphisms such as these have been linked to certain autoimmune conditions and also to susceptibility to infections.

Susceptibility to severe malaria is under complex genetic control with other genes in addition to MHC genes playing an important role. Recent studies have linked the development of cerebral malaria to a polymorphism in the promoter region of the *TNFα* gene. Other studies have implicated polymorphisms in the promoter region of the inducible NO synthase gene (*NOS2*).

Some genes involved in immune responses affect disease susceptibility, but do not affect immune responsiveness. For example, disease progression to AIDS has been shown to be associated with polymorphisms in the chemokine receptor gene-5 (CCR5).

CCR5 is a co-receptor used in the entry of macrophage-tropic strains of HIV-1 into cells. A mutation that inactivates this receptor is found in some individuals of European origin, but is rare in populations of Asian or sub-Saharan African descent. Individuals homozygous for this CCR5 mutation are very resistant to HIV-1 infection. In this case resistance is related to the reduced primary spread of the virus rather than an enhanced immune response against it.

Role of mutations in cytokine receptor genes in severe combined immune deficiency (SCID)

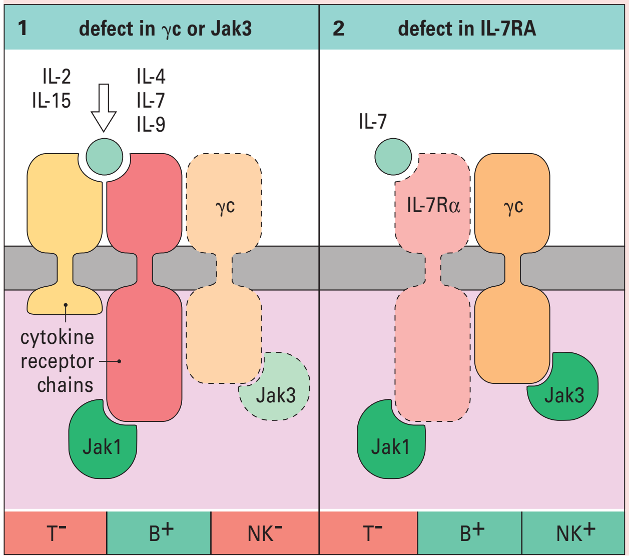

Fig. 11.25 (**1**) A defect in the common chain (γc) of the cytokine receptors IL-2, IL-15, IL-4, IL-7, and IL-9 leads to a SCID, with loss of both T and NK cells. A similar deficiency results from mutation in the janus kinase (Jak3), which transduces signals from the γc chain. Note that IL-2 and IL-15 have three chains in their high-affinity receptor, whereas IL-4, IL-7, and IL-9 have only two chains. (**2**) Absence of the specific IL-7R chain also produces a severe immunodeficiency, but this primarily affects T cell development.

Genetic defects associated with immune deficiency or abnormalities

condition	defective gene	result
SCID	*γc*	failure of signal transduction by cytokines
	IL-2Rα	failure of IL-2 signal in activation and development
	IL-7Rα	failure of IL-7 signal in lymphocyte development
	Jak3	lack of signal transduction by cytokines
	CD3γ	no signal transduced from TCR
	CD3ε	no signal transduced from TCR
	ZAP70	no signal transduced from TCR
	ADA	T cell toxicity
	RAG1/2	failure in TCR and BCR gene recombination
T cell deficiency	*PNP*	T cell development failure
class II deficiency	*CIIT*	failure to express MHC class II molecules
class I deficiency	*TAP1/2*	failure to load MHC class I molecules
X-linked hyper-IgM	*CD40L*	no maturation of antibody response
X-linked-agammaglobulinemia	*Btk*	failure of B cell development
X-linked lymphoproliferative syndrome	*SH2DIA/SAP*	impaired negative signals to B cells
Autoimmune lymphoproliferative syndrome	*Fas(CD95)* or *FasL*	extended lymphocyte life span due to reduced apoptosis
mycobacterial infection	*IFNγ*R1/2, IL-12R	impaired TH1 responses

ADA, adenosine deaminase gene; BCR, B cell receptor; *Btk,* Bruton's tyrosine kinase gene; TCR, T cell receptor; *PNP*, purine nucleoside phosphorylase gene; *RAG*, recombination activating gene

Fig. 11.26 Genetic defects associated with immune deficiency or abnormalities. Based on a review by Leonard, Curr Opin Immunol 2000;12:465–467. See also Chapter 16.

FURTHER READING

Agnello D, Lankford CS, Bream J, et al. Cytokines and transcription factors that regulate T helper cell differentiation: new players and new insights. J Clin Immunol 2003;23:147–161.

Akbar AN, Beverley PCL, Salmon M. Will telomere erosion lead to a loss of T-cell memory. Nat Rev Immunol 2004;4:737–743.

Blalock JE, Bost KL, eds. Shared ligands and receptors as a molecular mechanism for communication between immune and neuroendocrine systems. Ann NY Acad Sci 1994;741:292–298.

Biron C, Nguyen KB, Pien GC, et al. Natural killer cells in antiviral defense: function and regulation by innate cytokines. Annu Rev Immunol 1999;17:189–220.

Bodnar AG, Ouellette M, Frolkis M, et al. Extension of lifespan by introduction of telomerase into normal human cells. Science 1998;279:349–352.

Chess L, Jiang H. Resurrecting $CD8^+$ suppressor cells. Nat Immunol 2004;5:569–571.

Cottrez F, Groux H. Specialization in tolerance: innate $CD4^+CD25^+$ versus acquired Tr1 and TH3 regulatory T cells. Transplantation 2004;1:S12–S15.

Heymann B. Regulation of antibody responses via antibodies, complement, and Fc receptors. Annu Rev Immunol 2000;18:709–738.

Lenardo M, Chan FK-M, Hornung F, et al. Mature T lymphocyte apoptosis–immune regulation in a dynamic and unpredictable antigenic environment. Annu Rev Immunol 1999;17:221–253.

Leonard WJ. Genetic effects on immunity. Curr Opin Immunol 2000;12:465–467.

Metzler B, Wraith DC. Inhibition of experimental autoimmune encephalomyelitis by inhalation but not oral administration of the encephalitogenic peptide: influence of MHC binding affinity. Int Immunol 1993;5:1159–1165.

Mills KHG, McGuirk P. Antigen-specific regulatory T cells – their induction and role in infection. Semin Immunol 2004;16:107–117.

Reed JR, Vukmanovic-Stejic M, Fletcher JM, et al. Telomere erosion in memory T cells induced by telomerase inhibition at the site of antigenic challenge in vivo. J Exp Med 2004;199:1433–1443.

Romagnani S. Th1/Th2 cells. Inflamm Bowel Dis 1999;5:285–294.

Van der Vliet HJJ, Molling JW, von Blomberg BM, et al. The immunoregulatory role of CD1d-restricted natural killer T cells in disease. Clin Immunol 2004;112:8–23.

Zlotnik A, Yoshie O. Chemokines: a new classification system and their role in immunity. Immunity 2000;12:121–127.

Critical thinking: Regulation of the immune response (see p. 496 for explanations)

1 Why is there a need to regulate the extent of immune activation and lymphocyte proliferation?

2 In what ways can antibody regulate the immune response?

3 How would you use regulatory T cells to modulate aberrant immune responses?

4 What is the advantage of inducing replicative senescence in lymphocytes that are excessively stimulated?

CHAPTER

12

Immune Responses in Tissues

SUMMARY

- **A tissue can influence local immune responses,** promoting some classes of immunity and suppressing others. The tissue can accomplish this by producing cytokines that influence both T cells and APCs. Direct interactions between cells of the tissue can also modulate immune responses.
- **Certain sites in the body are immunologically privileged** and fully allogeneic tissue can be transplanted into them without risk of rejection. These sites, which include the anterior chamber of the eye, the brain, and the testes, are able to promote certain kinds of beneficial classes of immune responses while suppressing classes that can do irreparable local damage.
- **Immune responses in gut, lung, and skin distinguish between pathogens and innocuous organisms and antigens.** Gut enterocytes influence the local immune response. Intraepithelial lymphocytes (IELs) respond to stress-induced class Ib molecules and produce many immunomodulatory cytokines.
- **Endothelium controls which populations of lymphocytes enter a tissue.** Endothelium within each tissue provides signals for the selective accumulation of specific tissue-resident lymphocytes, many of which do not recognize foreign antigens, but are activated by stress-induced self molecules.
- **A number of principles govern the immunological characteristics of each tissue.** As well as the principles outlined in the summary points above, cells of the tissue can promote the accumulation of tissue-resident lymphocytes and exert their effects via cytokines or by direct cell–cell interactions. A tissue can have several microenvironments, each of which has its own physiology and preferred immune response, and the local immune response can produce systemic changes.

A TISSUE CAN INFLUENCE LOCAL IMMUNE RESPONSES

Any immune system faced with a potential threat must answer two questions:

- The first question is ‘Shall I respond?’, and the mechanisms used to determine that answer are the topics of much of this book.
- Once made, however, there is a second question – ‘What type of response shall I make?’

What determines whether an immune response should be comprised of, for example, activated cytotoxic T lymphocytes (CTLs) or a particular class of antibodies?

Although it has long been taught that the effector class of an immune response is tailored to the pathogen being fought, accumulating evidence suggests that there may also be a strong influence from the local tissue milieu (Fig. 12.1).

This chapter focuses on:

- the features of immune responses that are unique to individual tissues; and
- the mechanisms by which the tissues influence local and systemic characteristics.

There are several reasons why a particular organ may need to modify local immunity.

Q. Herpes simplex virus infects sensory neurons and causes cold sores. Why might a CTL response be inappropriate for controlling such an infection?

A. An effective cytotoxic response would kill the infected neuron, which cannot be replaced. This would be worse for the host than the moderate inconvenience caused by the sporadically reactivating viral infection.

This is an example of how an immune response summoned to clear an infection can interfere with a tissue’s physiology as seriously as the infection itself.

A similar outcome can occur in both the eye and the gut, which may be damaged by cytokines such as tumor necrosis factor-α (TNFα) and interferon-γ (IFNγ) produced locally during cell-mediated immune reactions.

Indeed, when TNFα and IFNγ reach high systemic levels, they can result in shock and rapid death. An example is seen in Dengue shock syndrome. Individuals who are immune to the Dengue virus or infants with maternal antibodies may develop rapid circulatory collapse. It is thought that interaction between activated T cells and macrophages causes the release of TNFα, which acts on endothelium leading to an increase in capillary permeability and consequent fall in blood pressure.

Factors controlling the characteristic immune response of a tissue

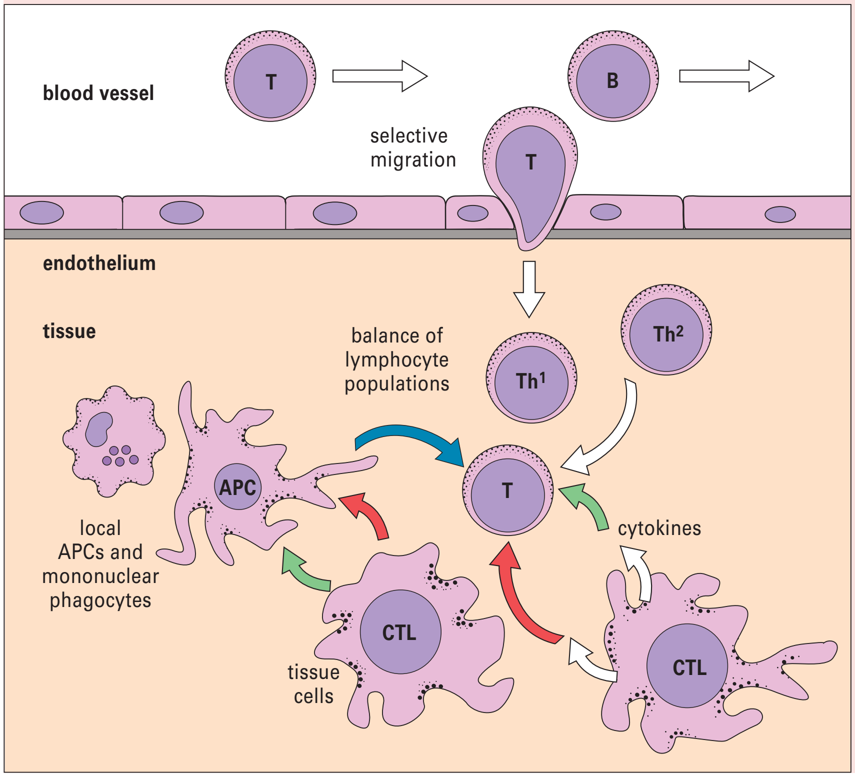

Fig. 12.1 The characteristic immune response of a tissue is controlled both by the leukocyte populations present and by the direct and indirect signals from the endogenous cells of the tissue. The population of lymphocytes that enter a particular tissue is controlled by the vascular endothelium in that tissue. Antigen presentation within the tissue depends on the populations of antigen-presenting cells (APCs), which include resident mononuclear phagocytes. APCs and T cells are influenced directly by endogenous cells of the tissue, and indirectly via cytokines.

Another reason for specific local immune responses is that certain types of immune response are more effective in some tissues than in others.

Q. What type of immune response is suited to the gut and mucosal surfaces?

A. The production of IgA antibodies, promoted by a TH2-type immune response. Accumulation and recirculation of IgA-producing B cells within submucosal tissues means that IgA antibodies are produced near the point of secretion (Fig. 12.2 and see Fig. 3.11)

These observations suggest that the immune system might normally receive guidance from each tissue to make a response that is both appropriate for and effective at that site.

Consequently, tissues have evolved regulatory mechanisms that influence the immune response that occurs within them. The detailed study of these regulatory mechanisms is in its infancy, but we will attempt to assemble some of the broad operating rules, and provide examples of the types of immune response seen in different tissues.

IgA synthesis in the lamina propia

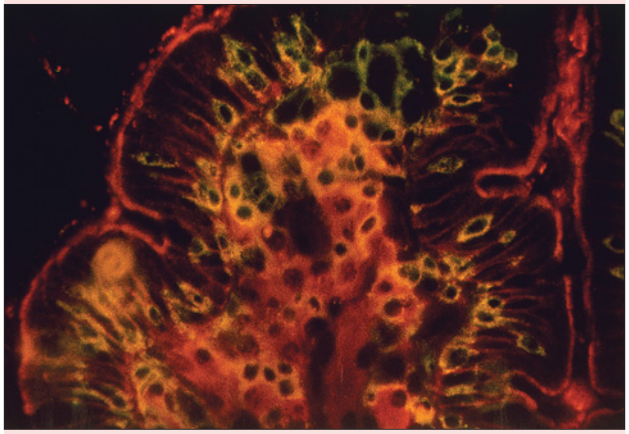

Fig. 12.2 Lymphoid cells in the epithelium and lamina propria fluoresce green using antibody to leukocyte common antigen (CD45). Red cytoplasmic staining is obtained with anti-IgA antibody, which detects plasma cells in the lamina propria and IgA in the mucus. (Courtesy of Professor G Janossy)

CERTAIN SITES IN THE BODY ARE IMMUNOLOGICALLY PRIVILEGED

There are certain sites in the body where fully allogeneic tissue can be transplanted without risk of rejection. These include the:

- anterior chamber of the eye (Fig. 12.3);
- brain;
- testes; and
- (oddly) the hamster cheek pouch.

Several factors may contribute to the **immunological privilege** of these sites, some of which affect the initiation of the immune response and some affect the effector phase. Some of the factors that limit the development of immune reactions in the central nervous system (CNS) include the following:

- the blood–brain barrier (endothelium plus astrocytes) prevents the movement of over 99% of large serum proteins into the brain tissue (IgG, complement, etc.);
- low levels of MHC molecule expression, and co-stimulatory molecules result in inefficient antigen presentation;
- there is no conventional lymphatic drainage system from brain tissue to local lymph nodes;
- there are low levels of leukocyte traffic into the CNS in comparison with other tissues;
- neurons have direct immunosuppressive actions on glial cells;
- astrocytes and neurons produce immunosuppressive cytokines.

Similar factors apply in other privileged sites.

The 'privileged sites' were long thought to be locations where adaptive immune responses are so dangerous that the immune system is either not allowed entry, is destroyed upon arrival, or is prevented from functioning.

From an evolutionary point of view, however, this explanation for the acceptance of grafts seems rather unlikely because an unprotected tissue that is warm, wet, and full of nutrients would quickly become a haven for parasites.

Recent evidence suggests that the concept of privileged sites may have been a misconception based on the limitations of the experimental assay systems. Once the experiments were expanded to include a wider variety of assays it became apparent that privileged sites are not immunologically impaired. They are simply sites that are able to promote certain kinds of beneficial classes of immune responses while suppressing classes that can do irreparable local damage.

The eye has very limited self-regenerative capacity and can be completely destroyed by a cell-mediated immune response with the concomitant local production of TNFα and IFNγ (see Fig. 12.3). The same response to the same antigen within the skin does not have such a severe effect. The principal difference is the regenerative capacities of the two organs. The eye has therefore evolved two major mechanisms to actively suppress cell-mediated responses:

- first, the epithelial cells lining the anterior chamber express Fas ligand;
- second, the fluid of the anterior chamber contains cytokines such as transforming growth factor-β (TGFβ).

Q. What would be the functional effect of epithelial cells lining the anterior chamber expressing Fas ligand and the fluid of the anterior chamber containing TGFβ have on an immune response?

A. Expression of Fas ligand (CD178) on the epithelial cells permit engagement of incoming cells (e.g. T cells) that express Fas (CD95) and thereby induce their apoptosis (see Figs 10.8 and 10.9) – a case of the tables being turned on the lymphocytes. Production of TGFβ limits T cell proliferation and tends to induce TH2 and possibly Treg lymphocytes. Additionally, the expression of Fas ligand may also favor a switch to a TH2-type response because Fas ligand has been shown to induce the death of TH1 cells more easily than TH2 cells.

Corneal allografts

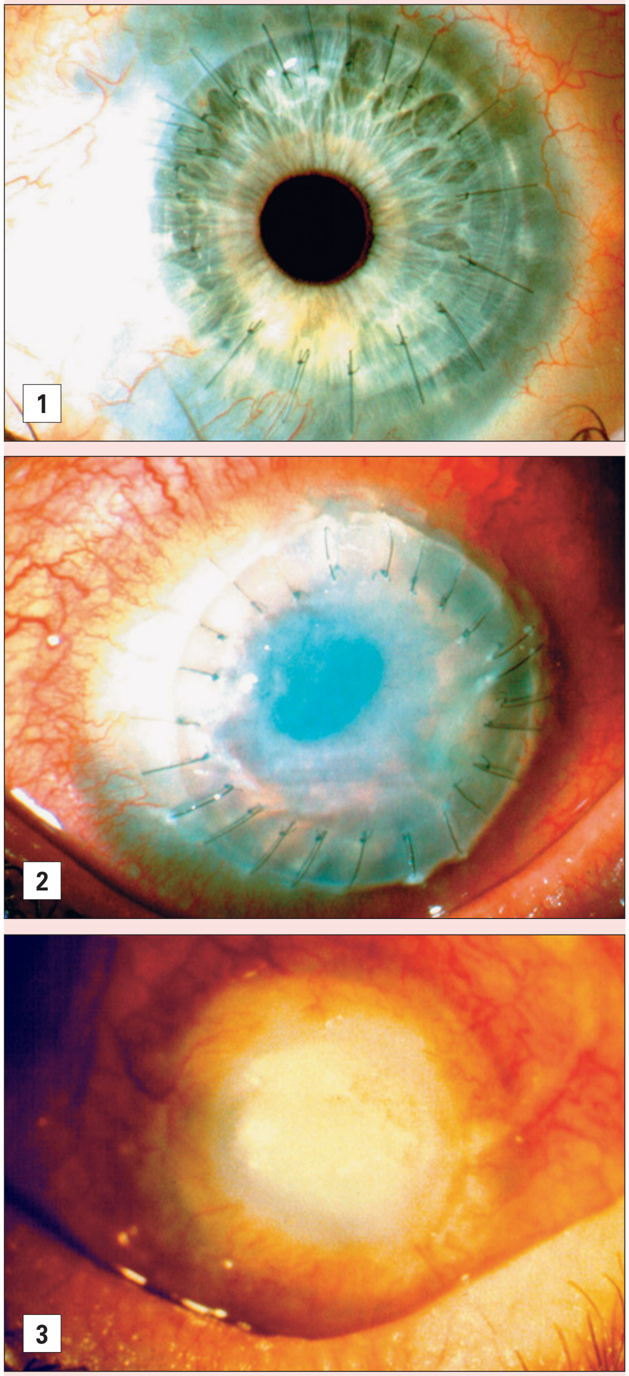

Fig. 12.3 (**1**) A successful corneal allograft. More than 90% of such grafts remain clear after 1 year. The sutures are still visible in the periphery. Note that vascularization does not extend into the graft. (**2**) An acutely rejected graft. (**3**) A chronically rejected graft. The rejected grafts indicate that effective cell-mediated responses can develop within the eye. Note the increased vascularization in the rejected grafts. (With permission from the University of Aberdeen)

Local immune privilege may extend to other systems

When first discovered, the features of immune privilege described above were thought to fit with the idea of the eye as a privileged site. However, an ingenious set of experiments by Streilein and his colleagues showed that the eye actually does something much more interesting than merely suppress local immunity and shift immune responses into a class of immunity that is non-destructive to the eye. In their experiments they transplanted tumors into the eye, which grew unhindered as expected. Unexpectedly, however, the tumor recipients subsequently became more generally tolerant of any other tissue that shared some of the tumor antigens, no matter the site of transplantation (Fig. 12.4).

Following up on this finding, it was discovered that spleen APCs that had been incubated with fluid from the anterior chamber of the eye were no longer able to induce T cell proliferative responses, but instead promoted production of IL-4, IL-10, and TGFβ. Streilein called this switch anterior chamber associated immune deviation (ACAID) and it was found to be controlled by the cells of the iris and ciliary body, which secrete immunomodulatory cytokines, including TGFβ, vasoactive intestinal peptide (involved in control of the synthesis of the secretory piece of IgA), γ-melanocyte-stimulating hormone (γMSH), and calcitonin gene-related peptide.

These cytokine and molecular controls on immunity are underpinned by the cellular organization of the eye, which has barriers that limit the movement of molecules into the retina, anterior chamber, and vitreous (Fig. 12.5).

Altogether, the picture that emerged is of an organ that actively promotes a type of immunity that is protective and beneficial (the IgA response) while suppressing a type of immunity that can be destructive.

Three principles are illustrated by ACAID:

- a tissue can influence its local immune responses, promoting some classes and suppressing others;
- the tissue can accomplish this by directly producing cytokines that influence both T cells and APCs;
- local influence can cause systemic changes.

The blood–brain barrier shields the CNS from immune reactions

Many of the factors that lead to limited immune responses in the eye also occur in the CNS. Indeed the blood–retinal barrier described in Fig. 12.5 is related to the blood–brain barrier present in most areas of the CNS.

The blood–brain barrier is a composite structure formed by the specialized brain endothelium and the foot processes of astrocytes. Astrocytes are required to induce the special properties of brain endothelial cells, which have continuous belts of tight junctions connecting them to other endothelial cells.

The blood–brain barrier has many physiological properties, but of interest to immunologists is the very low permeability of the endothelium to serum proteins (Fig. 12.6).

For example, the level of IgG found in the CNS is normally approximately 0.2% of the level found in serum. The level may rise during an immune reaction as the endothelial barrier becomes more permeable in response to inflammatory cytokines. Additionally, in some conditions, such as multiple sclerosis, there is often local synthesis of antibody within the CNS, which may contribute to the immune reaction.

Nevertheless, even in the most severe immune reactions, the levels of antibody, complement components, and other serum molecules is much lower in the brain than in other tissues.

Tolerance to allografts transplanted into the eye

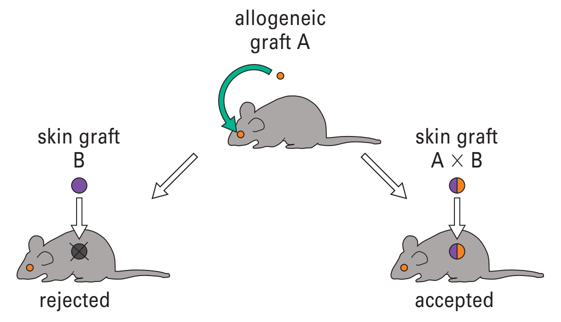

Fig. 12.4 Tumor cells of tissue type A transplanted into the eye are tolerated. Subsequently skin grafts of tissue type A × B are also tolerated in these animals. Tolerance to A, induced within the eye, extends to other tissues, and can turn off the normal immune response that would be expected against a type B skin allograft.

Anatomic and immunological basis of immune privilege in the eye

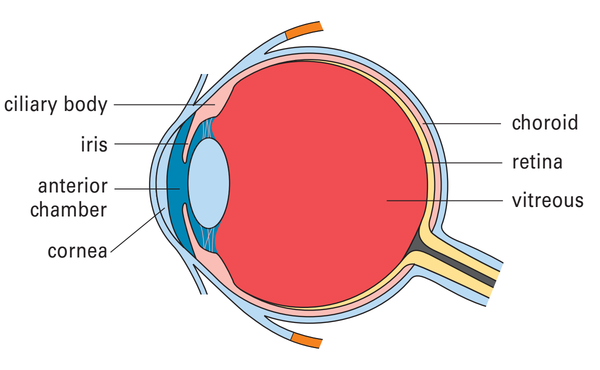

Fig. 12.5 The uvea, consisting of the iris, ciliary body, and chorea, is highly vascularized, but lacks draining lymphatic vessels. Tight junctions between the vessels in the iris and between non-pigmented ciliary epithelial cells maintain a blood–aqueous barrier. Similar junctions on the retinal pigment epithelium and the retinal endothelium maintain a blood–retinal barrier. The iris and ciliary body secrete immunomodulatory cytokines.

The blood–brain barrier

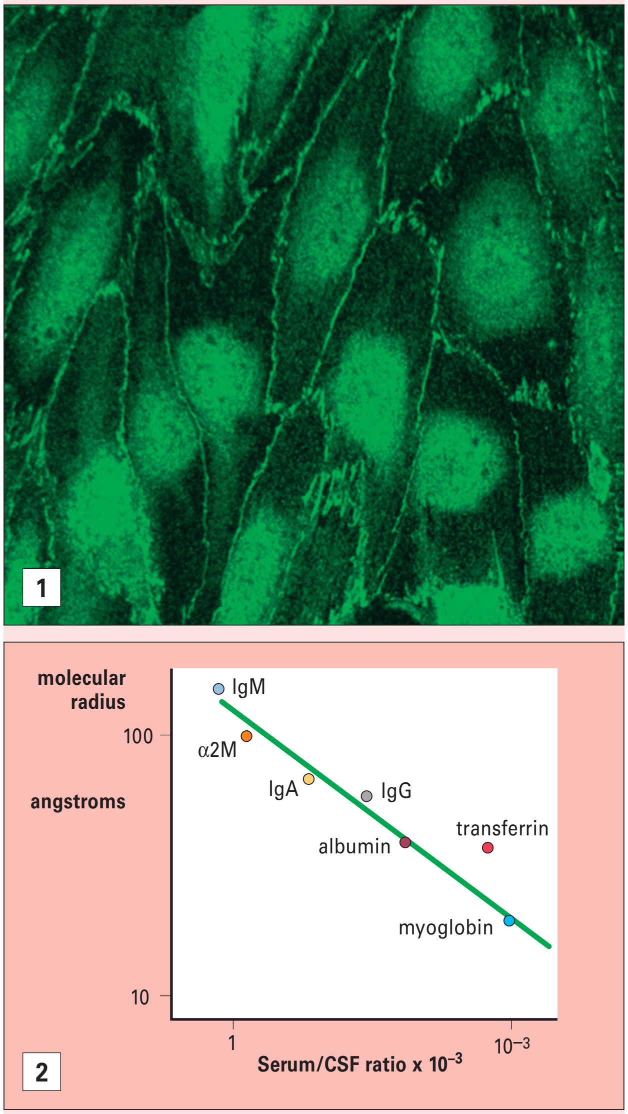

Fig. 12.6 (**1**) Brain endothelial cells have a continuous ring of tight junctions, identified in this micrograph by the junctional marker ZO-1. The resulting permeability barrier excludes serum proteins from the brain tissue. (**2**) The graph shows the serum/cerebrospinal fluid (CSF) ratio for different proteins. There is an inverse relationship between molecular size and the level in CSF. Molecules such as transferrin, which are transported into the brain, are present at higher levels than would be expected from their size. Immunoglobulins are not transported.

Neurons suppress immune reactivity in neighboring glial cells

When researchers first started to examine individual populations of cells from the CNS in vitro (e.g. astrocytes), they found that astrocytes would respond to IFNγ by increasing their expression of MHC class I molecules and inducing expression of the normally absent MHC class II molecules. However, in vivo, astrocytes rarely respond in this way. It appeared that the local environment of the CNS could in some way suppress the ability of astrocytes to respond to IFNγ.

Subsequently it was found that when neurons are co-cultured in contact with astrocytes, then the ability to induce MHC molecules was suppressed, but if the cells were not in contact, they were not repressed (Fig. 12.7).

Q. How do you interpret the repression of MHC induction described in Fig. 12.7?

A. Because repression of MHC induction requires contact between the cells it implies that the effect is not caused by a cytokine.

The results are even more interesting because it subsequently became clear that electrically active neurons are required. This means that:

- neurons that are functioning normally can suppress their neighboring glial cells (and downregulate their own MHC molecules); whereas
- damaged neurons will lift the local immunosuppression to allow an immune response to develop.

The examples of the eye and the CNS illustrate a number of mechanisms that lead to immune privilege. In practice, a combination of mechanisms is active in each tissue.

IMMUNE RESPONSES IN GUT, LUNG, AND SKIN DISTINGUISH BETWEEN PATHOGENS AND INNOCUOUS ANTIGENS

The gut, lung, and skin are examples of tissues that are continuously in contact with high levels of harmless

Immunosuppression by neurons

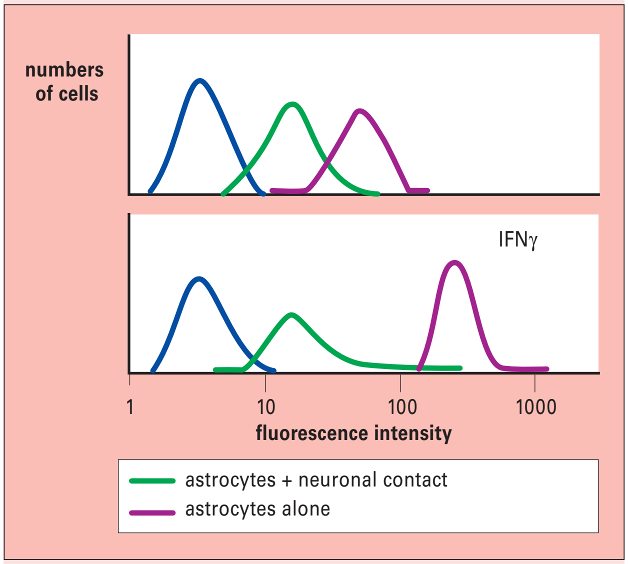

Fig. 12.7 The top fluorescence graph shows the expression of MHC class I molecules on astrocytes cultured in contact with or separate from neurons. Note that neurons suppress MHC class I expression. If IFNγ is added to the cultures (lower graph), astrocytes cultured alone increase their expression of MHC class I molecules, but those cultured with neurons are still suppressed. (Based on data from Tonsch U, Rott O. Immunology 1993;80:507)

commensal organisms and innocuous antigens as well as potential pathogens. The responses in these tissues are concerned not just with whether an immune response takes place, but also on the quality of that response.

Q. What factors allow the immune system to distinguish between pathogens and innocuous organisms and antigens?

A. Many pathogenic organisms have components that are recognized by **pathogen-associated molecular patterns (PAMP) receptors** (see Chapter 6 and Figs 6.22–6.24), which when ligated can lead to efficient antigen presentation. Additionally, pathogenic organisms cause tissue damage. The proteins released from damaged cells can directly activate APCs, while released enzymes such as tissue plasminogen activator (see Fig. 6.19) and tissue kallikrein (see Fig. 6.18) lead to activation of the plasmin and kinin systems.

Oral tolerance or regional immunity in the gut?

Occasionally we hear that a native American Indian child has eaten poison ivy but has not developed the cell-mediated hypersensitivity reaction that normally follows within 48 hours of contact with the plant. This is **oral tolerance**.

Oral tolerance, and the related phenomenon of nasal tolerance, illustrate two points:

- it is systemic in that it can influence responses at non-mucosal sites;
- it is dominant in that it is transferable to naive individuals by CD4 (or occasionally CD8) T cells.

Tolerance is not the only form of immunity that arises from ingestion of antigens. **Oral vaccination** has been recognized since 1919 when Besredka noticed that rabbits were protected from fatal dysentery by oral immunization with killed shigellae. The attenuated polio vaccine that was developed in the 1950s was also given as an oral vaccination. In both case these vaccines would not be viewed as a harmless antigen by the immune system:

- shigellae contain antigens that activate the intracellular PAMP receptor NOD-1 (see Fig. 9.24);
- the attenuated polio vaccine proliferates in the gut and lymphoid tissue, and therefore induces some tissue damage.

In each case, however, the vaccine is effective in protecting the host against disease.

Q. Can you give another example of tolerance induced by administration of antigen across mucosal tissues?

A. Tolerance in mice administered myelin basic protein (MBP) by aerosol can suppress the induction of experimental allergic encephalomyelitis (EAE) in response to the same antigen given intradermally in adjuvant (see Fig. 11.2).

Oral vaccination and tolerance are two sides of the same coin

In essence, they are very similar responses – the main difference lies with the assays used to measure them.

Antigens introduced orally tend to invoke an immune response that is appropriate for the gut and other mucosal surfaces, namely the production of local IgA, and some systemic IgG. There is generally little production of TH1 cells or CTLs, and no cell-mediated immune response.

Measurement of the IgA response would lead to the conclusion that oral antigens induce immunity. However, if the cell-mediated response (lymphocyte proliferation in response to antigen) is measured, one can erroneously conclude that the animal has become tolerant.

The success of the polio vaccine was measured by its ability to protect against the virally induced disease and was therefore labeled 'vaccination'. In contrast, reducing the autoimmune reaction in EAE (a model of multiple sclerosis) is viewed as a success, and this procedure is therefore labeled 'tolerance'. These outcomes – vaccination and tolerance – are, however, merely two sides of the same coin.

Gut enterocytes influence the local immune response

The enterocytes of the gut considerably influence the local immune response by secreting a variety of immunomodulatory factors such as TGFβ, VIP, IL-1, IL-6, IL-7, CXCL8, and CCL3.

Structure of the villi and crypts of the small intestine

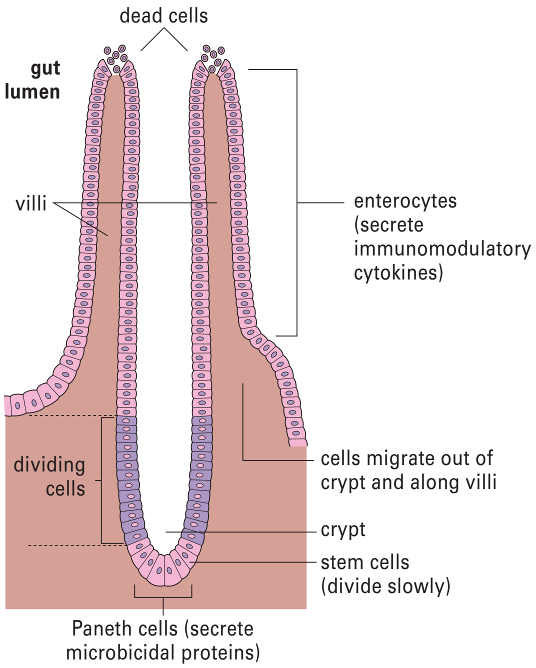
dead cells
gut lumen
villi
enterocytes (secrete immunomodulatory cytokines)
dividing cells
cells migrate out of crypt and along villi
crypt
stem cells (divide slowly)
Paneth cells (secrete microbicidal proteins)

Fig. 12.8 Stem cells give rise to Paneth cells at the base of the crypt, which secrete antimicrobial peptides. The enterocytes are derived from the rapidly dividing transit-amplifying population and will eventually be shed from the tips of the villi. These cells secrete immunomodulatory cytokines.

The Paneth cells at the bottoms of the crypts meanwhile produce several natural antibiotic peptides (Fig. 12.8), which help prevent bacterial overgrowth in these sensitive sites of cellular differentiation.

Enterocytes may also promote migration of intraepithelial lymphocyte (IEL) populations.

Q. What are the characteristics of IELs?

A. IELs have a dendritic morphology and express the γδ form of the T cell receptor (TCR) and often CD8 (see Fig. 2.22).

By expressing **E-cadherin**, a ligand for **αEβ7-integrin**, and by secreting TGFβ, which upregulates the expression of αEβ7, enterocytes may provide a signal for the selective accumulation of IELs that express this integrin. In this way, the tissue cells invite the residency of cells that promote certain types of immunity and aid in repair (see below).

IELs respond to stress-induced class Ib molecules

Many IELs have an activated or memory phenotype and recognize the ancient conserved MHC class I-like molecules, **MICA** and **MICB**, which are upregulated by cellular stress.

When activated by MICA and MICB some of the mucosa-associated lymphocytes secrete epidermal cell growth factor (ECGF) and may therefore function to repair and renew damaged intestinal epithelial cells.

Hence, by expressing molecules, the tissue cells can activate their local resident lymphocytes regardless of the specific antigen that is the target of the immune response. In this way the tissue can influence local immunity against many different antigens.

IELs produce many immunomodulatory cytokines

IELs also produce a wealth of immunomodulatory cytokines, such as IL-1, TGF, lymphotoxin (LT), IFNγ, and TNFα.

Although the conditions under which each set of cytokines is produced have not yet been elucidated, it is clear that certain cytokines, and their attendant antibody subclasses, are associated with certain intestinal diseases.

For example in ulcerative colitis, celiac disease, and Crohn's disease there is some associated production of TNFα and IFNγ, as there is in some other mucosal diseases such as chronic gastritis, chronic rhinitis, and chronic sialadenitis (i.e. situations where there is dysregulation of the normal mucosal immune responses).

ENDOTHELIUM CONTROLS WHICH POPULATIONS OF LYMPHOCYTES ENTER A TISSUE

Migration of leukocytes into different tissues of the body is dependent on the vascular endothelium in each tissue.

For many years, it was thought that the endothelium in different tissues was essentially similar, with the possible exception of tissues such as the brain and retina, which have barrier properties.

It was also acknowledged that the endothelium in lymphoid tissues – high endothelial venules (HEVs) – was specialized to support lymphocyte traffic specific for different areas.

Despite this common belief, it was well known that inflammation in different tissues had different characteristics, even when the inducing agents were similar. Compare the two type IV hypersensitivity reactions seen in the CNS and dermis (Figs 12.9 and 12.10):

- in the brain the lymphocyte infiltrate is confined to the region around the venules and there is little overt tissue disruption;
- in the dermis there is extensive infiltration with macrophages and lymphocytes, which is accompanied by severe edema and tissue disruption.

It has recently become clear that a major element controlling inflammation and the immune response is the vascular endothelium in each tissue, which has its own characteristics. Different endothelia produce distinctive blends of chemokines (Fig. 12.11).

In normal lung there is a high level of macrophage migration, which may be related to the high expression of CCL2 (macrophage chemotactic protein-1) by lung endothelium.

Type IV hypersensitivity in CNS

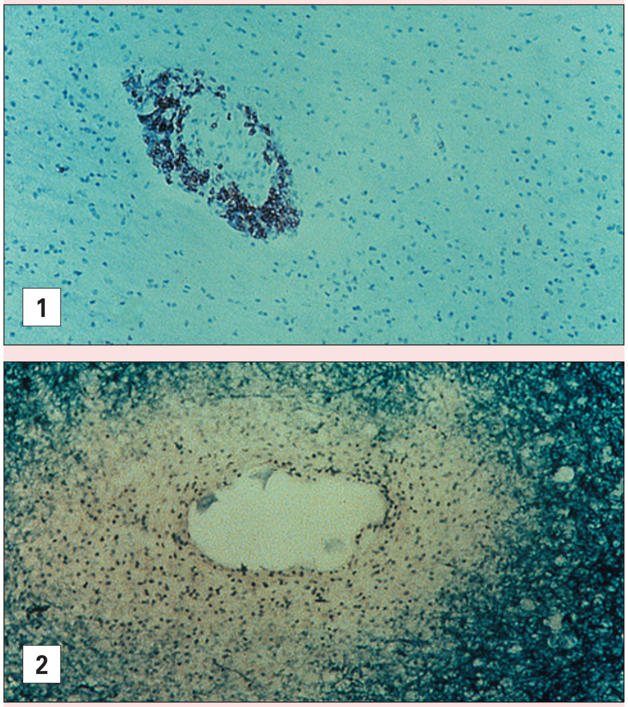

Fig. 12.9 (**1**) The perivenous infiltrations of lymphocytes in parainfectious encephalomyelitis are confined to the area around the venule top. There is little disruption of the tissue, but a stain for myelin shows that demyelination (**2**) extends out beyond the initial area of infiltration. (Courtesy of Dr N Woodroofe and Dr H Okazaki)

Type IV hypersensitivity in dermis

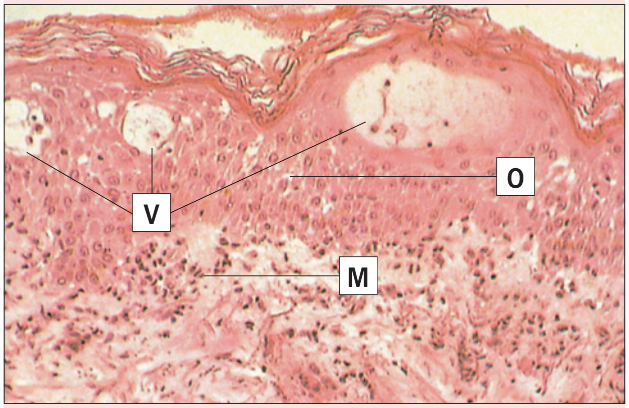

Fig. 12.10 A histological section of a cell-mediated immune reaction (type IV hypersensitivity) in the dermis. Mononuclear cells (M) infiltrate both dermis and epidermis. The epidermis is pushed out and microvesicles (V) form within it due to edema (O). × 130.

Production of chemokines by endothelium derived from different human tissues

	endothelium		
	lung	dermis	brain
CXCL8	++	++	+++
CXCL10	+	+	+++
CCL2	+++	++	+
CCL5	+	++	+

Fig. 12.11 Brain endothelium produces high levels of CXCL8 and CXCL10 associated with a TH1-type immune response. Dermal endothelium produces high levels of CCL5 associated with T cell migration, whereas lung endothelium produces high levels of CCL2, a chemokine that causes macrophage migration.

Q. How would you expect the high production of CXCL10 to affect immune reactions in the CNS?

A. CXCL10 is induced in response to IFNγ and acts on CXCR3, which is selectively expressed on TH1 cells (see Fig. 6.11). Therefore brain endothelium produces chemokines that promote a TH1-type reaction.

The cellular infiltrate in the CNS can be contrasted with that seen in the bronchus during allergic asthma where eosinophils predominate (Fig. 12.12). In this case the production of IL-5 and CCL3 (eotaxin) are characteristic of both the TH2 response and this particular tissue.

Histological section of an airway from a case of fatal asthma

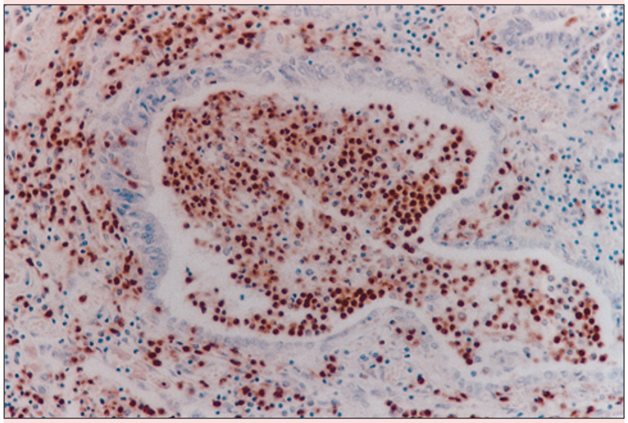

Fig. 12.12 The lumen of an alveolus of the lung is heavily infiltrated with inflammatory exudates, fibrin, and cellular debris. Immunohistochemical staining with monoclonal antibody against EG2 indicates that the majority of cells are eosinophils. × 275. (Courtesy of Arshad SH. Allergy: An Illustrated Colour Text. Philadelphia: Churchill Livingstone, 2002. With permission from Elsevier)

A NUMBER OF PRINCIPLES GOVERN THE IMMUNOLOGICAL CHARACTERISTICS OF EACH TISSUE

Knowledge of the distinctive immunological characteristics of each tissue is only just emerging. Nevertheless a number of principles can be discerned as follows.

- A tissue can influence the local immune response, promoting some classes of immunity and suppressing others.
- The local endothelium plays a major role in determining which leukocytes will enter the tissue by the secretion of distinct blends of chemokines.
- Cells of the tissue can promote the accumulation of tissue-resident lymphocytes, many of which do not recognize foreign antigens, but instead bind to stress-induced self molecules.
- Cells in the tissue can exert their effects via cytokines or by direct cell–cell interactions. In effect, cells of the tissue can signal damage or distress.
- A tissue can have several microenvironments, each of which has its own physiology and preferred immune response.
- The local immune response can produce systemic changes.

FURTHER READING

Brandtzaeg P. History of oral tolerance and mucosal immunity. Ann NY Acad Sci 1996;778:1–27.

Cheroutre H. Starting at the beginning: new perspectives on the biology of mucosal T cells. Annu Rev Immunol 2004;22:217–246.

Debendictis C, Joubeh S, Zhang G, et al. Immune functions of the skin. Clin Dermatol 2001;19:573–585.

Engelhardt B, Ransohoff RM. The ins and outs of T-lymphocyte trafficking to the CNS: anatomical sites and molecular mechanisms. Trends Immunol 2005;26:485–495.

Greenwood J, Begley DJ, Segal MB, eds. New concepts of a blood brain barrier. New York: Plenum Press; 1995.

Mowat AM, Viney JL. The anatomical basis of intestinal immunity. Immunol Rev 1997;156:145–166.

Spellberg B. The cutaneous citadel: a holistic view of skin and immunity. Life Sci 2002;67:477–502.

Wayne Streilein J. Regional immunology of the eye. In: Pepose JS, Holland GN, Wilhelmus KR, eds. Ocular infection and immunity. Oxford: Elsevier; 1996:19–33.

Weiler-Normann C, Rehermann B. The liver as an immunological organ. J Gastroenterol Hepatol 2004;19:S279–S283.

Wilbanks GA, Streilein JW. Fluids from immune privileged sites endow macrophages with the capacity to induce antigen-specific immune deviation via a mechanism involving transforming growth factor-β. Eur J Immunol 1992;22:1031–1036.

Williams IR, Kupper TS. Immunity at the surface: homeostatic mechanisms of the skin immune system. Life Sci 1996;58:1485–1507.

Zhang X, Brunner T, Carter L, et al. Unequal death in T helper cell (Th)1 and Th2 effectors: Th1, but not Th2, effectors undergo rapid Fas/FasL-mediated apoptosis. J Exp Med 1997;185:1837–1849.

Critical thinking: Immune reactions in the gut (see p. 496 for explanations)

Oyster poisoning occurs when an individual eats an oyster that has concentrated bacteria or Protoctista in itself from sea water. Often an individual who has eaten an infected oyster will then be unable to eat oysters again – even good oysters make them ill. Construct a logical explanation for this observation, based on your understanding of immune reactions in the gut.

This Page Intentionally Left Blank

SECTION

3

Defence Against Infectious Agents

This Page Intentionally Left Blank

CHAPTER 13

Immunity to Viruses

SUMMARY

- **Innate immune mechanisms (interferon, NK cells, and macrophages) restrict the early stages of infection and delay spread of virus.** Interferons exert antiviral activity by a variety of mechanisms. NK cells are cytotoxic for virally infected cells. Macrophages act at three levels to destroy virus and virus-infected cells. DC2 dendritic cells produce IFNα in herpesvirus and influenza virus infection.
- **As a viral infection proceeds, the adaptive (specific) immune response unfolds.** Antibodies and complement can limit viral spread or reinfection. T cells mediate viral immunity in several ways – $CD8^+$ CTLs destroy virus-infected cells; $CD4^+$ T cells are a major effector cell population in the response to many virus infections.
- **Viruses have evolved strategies to evade the immune response.** Virus latency and antigenic variation are the most effective mechanisms. Many viruses deviate the immune response by the production of cytokine analogs and cytokine receptor analogs. Many DNA viruses have strategies to control the expression of MHC molecules.
- **Responses to viral antigens can cause tissue damage** from the formation of immune complexes and by causing immunosuppression, immunodeficiency, or autoimmunity.

EARLY IMMUNE DEFENSES AGAINST VIRUSES INCLUDE INTERFERON, NK CELLS, AND MACROPHAGES

The early stage of a viral infection is often a race between the virus and the host's defense system. The initial defense against virus invasion is the integrity of the body surface. Once breached, early 'non-specific' or innate immune defenses such as interferon (IFN), natural killer (NK) cells, and macrophages become active.

Interferons exert antiviral activity by a variety of mechanisms

There are two major families of IFN – type I and type II IFNs.

Type I IFNs include:

- **IFNα** (leukocyte IFN) – a family of 13 genes on chromosome 9 – produced following virus infection of cells;
- **IFNβ** (fibroblast IFN) – a single gene on chromosome 9 – produced following virus infection of cells;
- **IFNτ/ε** (trophoblast IFN); and
- two new families, IFNλ and IFNκ.

Type II IFN or **IFNγ** – a single gene on chromosome 12 – is only produced following antigenic or mitogenic stimulation of T and NK cells.

Virus infection of a cell leads to the production of IFNα/β, which activates antiviral mechanisms in neighboring cells enabling them to resist virus infection (Fig. 13.1).

Interferons activate a number of genes with direct antiviral activity. A key mechanism involves the dsRNA-dependent protein kinase R (PKR), which:

- phosphorylates the α-subunit of eIF-2 leading to a block in the translation of viral mRNA;
- triggers apoptosis via Bcl-2 and caspase-dependent mechanisms thereby short circuiting virus infection through suicide of the infected cell.

Another mechanism activated by dsRNA involves 2′,5′-oligoadenylate synthetase, which activates a latent endonuclease (RNaseL) involved in degrading viral RNA.

Q. What role does dsRNA play in the normal metabolism of a mammalian cell?

A. dsRNA is formed as part of the replication/transcription of RNA viruses. However, because the mammalian genome is DNA, dsRNA is not produced during transcription of mRNA; consequently dsRNA is a signature of viral replication. At least, this was the view until recently. It is now known that dsRNA is formed in eukaryotic cells as part of the process that controls mRNA activity by small inhibitory RNAs (siRNA). Mammalian cells therefore have an intrinsic normal mechanism that leads to the recognition and degradation of dsRNA.

A third mechanism for controlling viral infection involves Mx proteins, which inhibit viral transcription of a range of RNA viruses, but have little effect on DNA viruses.

The molecular basis of interferon action

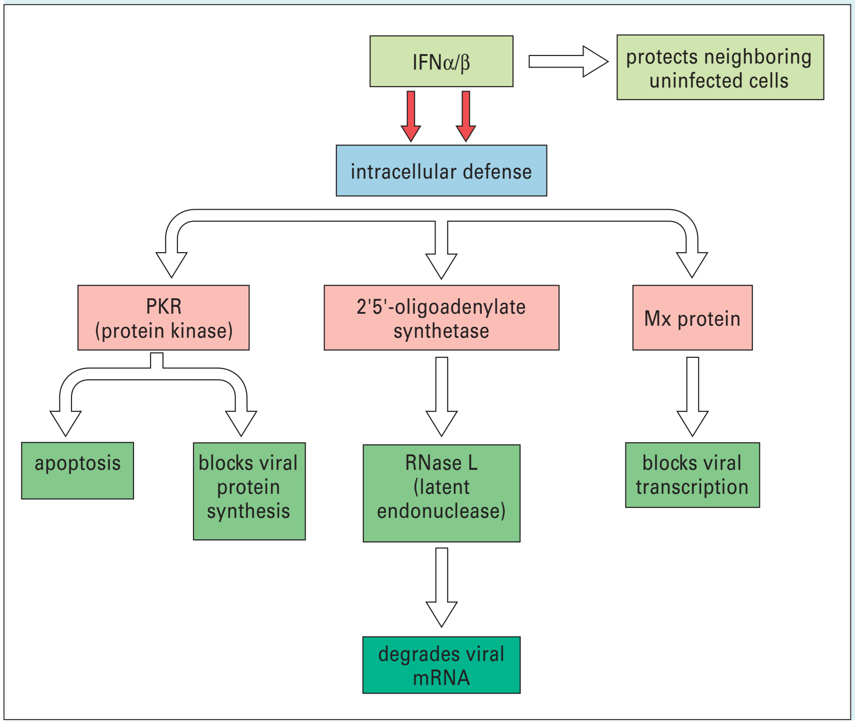
IFNα/β
protects neighboring uninfected cells
intracellular defense
PKR (protein kinase)
2'5'-oligoadenylate synthetase
Mx protein
apoptosis
blocks viral protein synthesis
RNase L (latent endonuclease)
blocks viral transcription
degrades viral mRNA

Fig. 13.1 The antiviral state develops within a few hours of interferon stimulation and involves various intracellular antiviral defenses.

Other mechanisms of IFN-mediated intracellular defense exist because mice deficient in *PKR*, *RnaseL*, and *Mx* genes (triple knockouts) still retain antiviral activity.

In addition to the direct inhibition of virus replication, IFNγ and IFNα/β enhance the efficiency of the adaptive immune response by stimulating increased expression of MHC class I and II, along with the antigen processing machinery. These IFNs also activate macrophages and NK cells, promoting their antiviral activity (see below).

The importance of IFNs in vivo is underlined by the increased susceptibility of mice lacking the IFNα/β receptor to virus infection. Similarly, depletion of IFNs by specific antibody treatment also augments virus infection.

NK cells are cytotoxic for virally infected cells

Active NK cells are detected within 2 days of a virus infection. They have been identified as major effector cells against herpesviruses and, in particular, murine cytomegalovirus (MCMV). An absence or reduction of NK cell activity, as seen in beige mutant mice or following depletion with antibodies against NK1.1 and asialo GM1, correlates with an increased susceptibility to MCMV infection.

During the initial stages of infection, NK cells undergo non-specific proliferation mediated by type I IFN and interleukin-15 (IL-15). In MCMV infection there then follows:

- a selection for NK cells expressing Ly49H (an activator of NK cell function); and
- expansion following interaction with the viral protein ml57 present on infected cells.

This interaction is the basis of resistance to MCMV infection in resistant mouse strains.

In susceptible mouse strains the activating Ly49H is absent and instead ml57 interacts with Ly49I (an inhibitory receptor) that inhibits NK cell responses. In these mice m157 acts as a decoy protein mimicking the host inhibitory receptors, thereby favoring survival of the virus.

The MCMV NK cell model represents a paradigm of how we envisage NK cells operating in other virus infections where an inverse correlation exists between MHC class I expression and NK cell killing. This is an interesting feature because a number of viruses are now known to downregulate MHC class I expression – presumably a strategy to evade T cell recognition.

In executing the antiviral response, NK cells function by:

- mediating direct cytolysis of infected cells through a perforin–granzyme mechanism (see Chapter 10); and
- through the production of IFNγ, which protects cells from infection and activates macrophage antiviral mechanisms.

NK cells are also one of the main mediators of antibody-dependent cellular cytotoxicity (ADCC).

Macrophages act at three levels to destroy virus and virus-infected cells

Macrophages are ever present in the tissues of the body and act as a first line of defense against many pathogens. In virus infection they act at three levels to destroy virus and virus-infected cells:

- phagocytosis of virus and virus-infected cells;
- killing of virus-infected cells; and

- production of antiviral molecules such as tumor necrosis factor-α (TNFα), nitric oxide, and IFNα.

Phagocytosis of infected cells and virus complexes is part of the normal housekeeping role of macrophages at a site of infection.

As with many pathogens the phagolysosome represents a hostile environment for viruses in which oxygen-dependent and oxygen-independent destructive mechanisms prevail. The induction of nitric oxide synthetase and the generation of nitric oxide is a potent inhibitor of herpesvirus and poxvirus infection.

DC2 dendritic cells produce IFNα in herpesvirus and influenza virus infection

A major producer of IFNα in herpesvirus and influenza virus infection is the macrophage-like plasmacytoid-derived DC2 dendritic cell. In influenza virus infection DC2 dendritic cells make large amounts of interferon following the endosomal recognition of viral RNA.

Q. How can macrophages recognize dsRNA?

A. Toll-like receptor-3 (TLR-3) (and TLR-7) recognizes dsRNA (see Fig. 6.23). These cells, along with type I IFN production from infected cells, act as major coordinators of innate immune responses during the early stages of a virus infection. This network is summarized in Fig. 13.2.

AS A VIRAL INFECTION PROCEEDS THE ADAPTIVE (SPECIFIC) IMMUNE RESPONSE UNFOLDS

An absence of T cells renders the host highly susceptible to virus attack. For example, cutaneous infection of congenital athymic 'nude' mice (which lack mature T cells) with herpes simplex virus (HSV) results in a spreading lesion – the virus eventually travels to the central nervous system, resulting in death of the animal. The transfer of HSV-specific T cells shortly after infection is sufficient to protect the mice.

Antibodies and complement can limit viral spread or reinfection

Antibodies can neutralize the infectivity of viruses

As a viral infection proceeds, the adaptive (specific) immune response unfolds, with the appearance of:

- cytotoxic T lymphocytes (CTLs);
- helper T (TH) cells; and
- antiviral antibodies.

Antibodies provide a major barrier to virus spread between cells and tissues and are particularly important in restricting virus spread in the blood stream. IgA production becomes focused at mucosal surfaces where it serves to prevent reinfection.

An alternative mechanism of IgA-mediated neutralization occurs intracellularly as IgA passes from the luminal to the apical surface. During this transcytosis vesicles containing IgA interact with those containing virus, leading to neutralization.

Antibodies may be generated against any viral protein in the infected cell.

Q. Which proteins are likely to be the most important targets of antibody-mediated defenses?

A. Only antibodies directed against glycoproteins that are expressed on the virion envelope or on the infected cell membrane are of importance in controlling infection.

Antibody-mediated immunity can be achieved in a number of ways, involving quite diverse mechanisms.

Defense against free virus particles involves neutralization of infectivity, which can occur in various ways (Fig. 13.3). Such mechanisms are likely to operate in vivo because injection of neutralizing monoclonal antibodies is highly effective at inhibiting virus replication. Clearly the

Interferon activates NK cells and macrophages

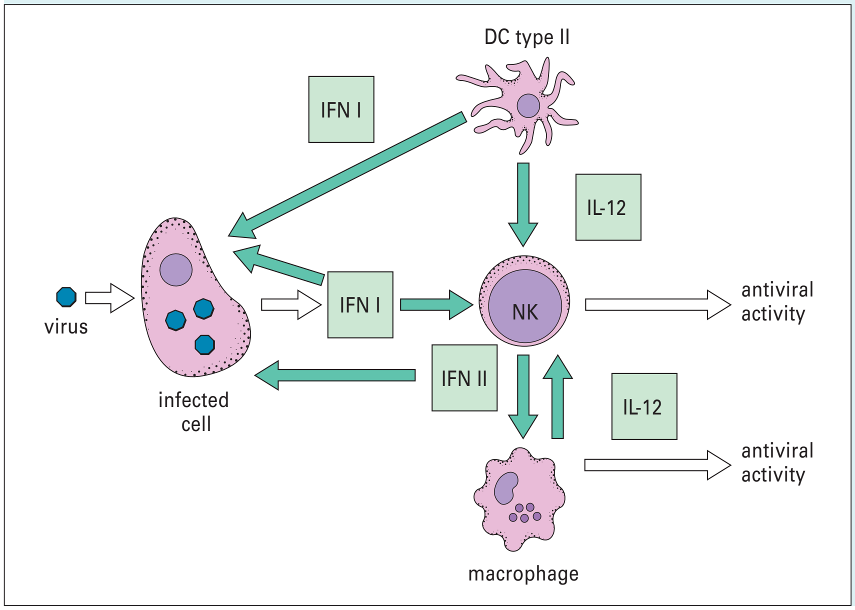

Fig. 13.2 Interferon (IFN) has an important role in orchestrating innate antiviral defenses. (DC type II, plasmacytoid-derived DC2 dendritic cell; IFN I, type I IFN; IFN II, type II IFN)

Antiviral effects of antibody

target	agent	mechanism
free virus	antibody alone	blocks binding to cell blocks entry into cell blocks uncoating of virus
	antibody + complement	damage to virus envelope blockade of virus receptor
virus-infected cells	antibody + complement	lysis of infected cell opsonization of coated virus or infected cells for phagocytosis
	antibody bound to infected cells	ADCC by NK cells, macrophages, and neutrophils

ADCC, antibody-dependent cellular cytotoxicity

Fig. 13.3 Antibody acts to neutralize virus or kill virally infected cells.

presence of circulating virus-neutralizing antibodies is an important factor in the prevention of reinfection.

Complement is involved in the neutralization of some free viruses

Complement can also damage the virion envelope, a process known as **virolysis**, and some viruses can directly activate the classical and alternative complement pathways. However, complement is not considered to be a major factor in the defense against viruses because individuals with complement deficiencies are not predisposed to severe viral infection.

This should be contrasted with those herpesviruses and poxviruses that carry viral homologs of complement regulatory proteins (CD46, CD55) which regulate complement activation. Presumably these viruses perceive complement to be a threat.

Antibodies mobilize complement and/or effector cells to destroy virus-infected cells

Antibodies are also effective in mediating the destruction of virus-infected cells. This can occur by antibody-mediated activation of the complement system, leading to the assembly of the membrane attack complex and lysis of the infected cell (see Chapter 4). This process requires a high density of viral antigens on the membrane (about 5×10^6/cell) to be effective. In contrast, ADCC mediated by NK cells can recognize as few as 10^3 IgG molecules in order to bind and kill the infected cell.

Q. How can NK cells use antibody to recognize and destroy virus-infected cells?

A. The IgG-coated target cells are bound using the NK cell's FcγRIII (CD16; see Chapter 3), and are rapidly destroyed by a perforin-dependent killing mechanism (see Figs 10.1 and 10.12).

Just how important these mechanisms for destroying virus-infected cells are in vivo is difficult to resolve. The best evidence in favor of ADCC comes from studying the protective effect of non-neutralizing monoclonal antibodies in mice. Although these antibodies fail to neutralize virus in vivo, they can protect C5-deficient mice from a high-dose virus challenge. (C5-deficient mice are used in this study to eliminate the role of the late complement components.)

T cells mediate viral immunity in several ways

T cells exhibit a variety of functions in antiviral immunity:

- most of the antibody response is T-dependent, requiring the presence of $CD4^+$ T cells for class switching and affinity maturation;
- $CD4^+$ T cells also help in the induction of $CD8^+$ CTLs and in the recruitment and activation of macrophages at sites of virus infection;
- $CD8^+$ T cells are effective in the prevention of re-infection (following vaccination) by viruses such as influenza virus and respiratory syncytial virus – however, even memory T cells need time to evolve a response to a reinfection virus, and therefore antibodies assume a more dominant role by neutralizing incoming virus and containing the infection by preventing spread to other tissues (see above).

Q. How do $CD4^+$ T cells help induce and recruit $CD8^+$ T cells?

A. Cytokines, including IL-2 released by $CD4^+$ T cells, are required for division of $CD8^+$ T cells. $CD4^+$ T cells can recruit $CD8^+$ T cells to sites of infection by the release of chemokines.

$CD8^+$ CTLs destroy virus-infected cells

The principal T cell surveillance system operating against viruses is highly efficient and selective. MHC class I-restricted $CD8^+$ CTLs cells focus at the site of virus replication and destroy virus-infected cells.

$CD8^+$ T cells:

- not only kill infected cells through the release of perforin and granzymes or through Fas–FasL interactions; but

- can 'cure' some types of persistent virus infection (e.g. hepatitis B virus) involving the release of IFNγ and/or TNF, resulting in clearance of the virus without death of the cell.

Virtually all cells in the body express MHC class I molecules, making this an important mechanism for identifying and eliminating or curing virus-infected cells.

Because of the central role played by MHC class I in targeting CD8$^+$ T cells to infected cells, some viruses have evolved elaborate strategies to disrupt MHC class I expression, thereby interfering with T cell recognition and favoring virus persistence (see below).

Virtually any viral protein can be processed in the cytoplasm to generate peptides, which are then transported to the endoplasmic reticulum where they interact with MHC class I molecules.

Q. Why might it be advantageous to the host to present viral peptides that are produced early in the replication cycle, and which may not be part of the assembled virus?

A. Viral proteins expressed early in the replication cycle can be targeted on infected cells, enabling T cell recognition to occur long before new viral progeny are produced. For example, CD8$^+$ T cell-mediated immunity against murine cytomegalovirus (CMV) is mediated predominantly by immediate early protein pp89 (80–90% of the T cell response is directed against pp89). The epitope has been identified as a nonamer peptide presented by the MHC class I molecule L^d. Immunization of mice with a recombinant vaccinia virus containing pp89 is sufficient to confer complete protection from murine CMV-induced disease; deletion of the DNA sequence encoding the nonapeptide abolishes the protective effect of the protein.

The importance of T cell mechanisms in vivo has been identified using various techniques:

- the adoptive transfer of specific T cell subpopulations or T cell clones to infected animals and monitoring of viral clearance;
- depletion of T cell populations in vivo using monoclonal antibodies to CD4 or CD8;
- creation of 'gene knockout' mice, in which genes encoding cell surface receptors (CD8, CD4), transcription factors (T-bet), and signal transduction molecules (STAT) are removed from the germline.

The continued ability of knockout mice that lack particular lymphocyte populations to mount a response against virus infections is a good illustration of the redundancy that can occur in the immune system. For example, in the absence of CD8$^+$ T cells, CD4$^+$ T cells or other mechanisms are able to compensate and bring the infection under control.

CD4$^+$ T cells are a major effector cell population in the response to many virus infections

CD4$^+$ T cells have now been identified as a major effector cell population in the immune response to many virus infections. A good example is in HSV-1 infection of epithelial surfaces. In this instance recruitment of macrophages occurs in delayed-type hypersensitivity (see Chapter 26), resulting in an accelerated clearance of virus. Macrophages are an important component in this process, inhibiting virus infection probably through the generation and action of nitric oxide (Fig. 13.4). Key cytokines in this response include IFNγ, important in the activation of monocytes, and tumour necrosis factor (TNFα).

TNFα has several antiviral activities, including the induction of intracellular interferon defense mechanisms and apoptotic cell death following interaction with the apoptotic TNF receptor.

In measles virus and Epstein–Barr virus (EBV) infection, CD4$^+$ CTLs are generated that recognize and kill MHC class II-positive cells infected with the virus. This suggests that measles virus and EBV peptides are generated by normal pathways of antigen presentation (i.e. following phagocytosis and degradation, see Chapter 7). However, other pathways have been implicated in which some measles proteins/peptides enter class II vesicles from the cytosol.

A summary of antiviral defense mechanisms is illustrated in Fig. 13.5, and the kinetics of their induction is shown in Fig. 13.6.

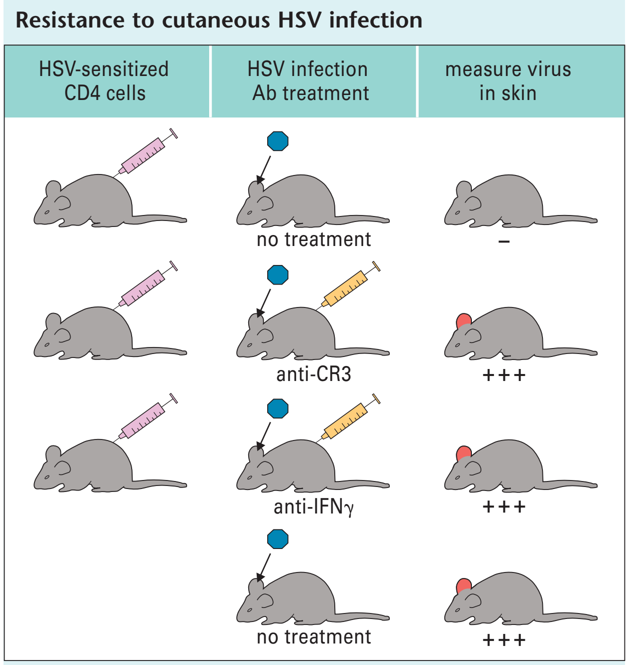

Fig. 13.4 CD4$^+$ T cells, macrophages, and IFNγ all have a protective role in cutaneous infections with HSV-1. CD4$^+$ T cells were obtained from mice infected with HSV-1 8 days previously. The cells were transferred to syngeneic mice infected with HSV-1 in the skin. These mice were treated with anti-CR3 (to block macrophage migration to the site of infection), or anti-IFNγ (to block the activation of macrophages), or were untreated. An additional control group was infected, but did not receive CD4$^+$ T cells. The amount of infectious virus remaining after 5 days was then determined. The results demonstrate that the protective effects of CD4$^+$ T cells are mediated by macrophages and IFNγ.

Effector mechanisms against virus and virus-infected cells

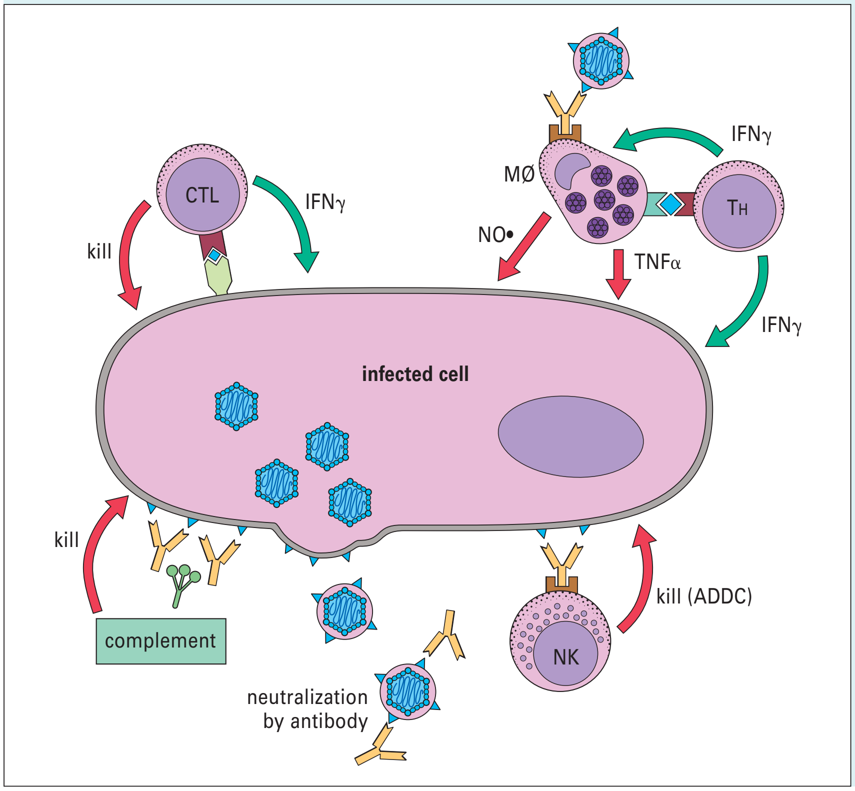

Fig. 13.5 Entry of virus at mucosal surfaces is inhibited by IgA. Following the initial infection, the virus may spread to other tissues via the blood stream. Interferons produced by the innate (IFNα and IFNβ) and adaptive (IFNγ) immune responses make neighboring cells resistant to virus infection. Antibodies are important in controlling free virus, whereas T cells and NK cells are effective at killing infected cells.

Response to a typical acute virus infection

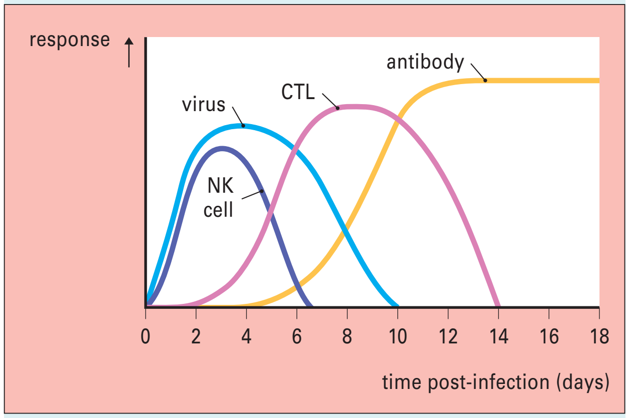

Fig. 13.6 Kinetics of host defenses in response to a typical acute virus infection. Following an acute virus infection (e.g. by influenza or herpes virus), NK cells and interferon are detected in the blood stream and locally in infected tissues. CTLs then become activated in local lymph nodes or spleen, followed by the appearance of neutralizing antibodies in serum. Although activated T cells are absent by the second to third week, T cell memory is established and may last for many years.

VIRUSES HAVE EVOLVED VARIOUS STRATEGIES TO EVADE RECOGNITION BY ANTIBODY AND T CELLS

Virus latency and antigenic variation are the most effective mechanisms

Antigenic variation involves mutating regions on proteins that are normally targeted by antibody and T cells. Antigenic variation is seen in human immunodeficiency virus (HIV) and in foot and mouth disease virus, and is responsible for the antigenic shift and drift seen with influenza virus (Fig. 13.7). Humoral immunity to such diseases lasts only until the new virus strain emerges, making effective longlasting vaccinations difficult to produce.

In HIV infection, sequence changes (mutations) can arise in those viral peptides that bind to MHC class I molecules to which the initial T cell response arose. This results in a failure of T cell surveillance and the emergence of new pathogenic variant viruses.

All viruses trigger an interferon response and consequently most viruses have evolved a strategy for disrupting the interferon system. This occurs at three levels:

- disruption of PKR or 2′,5′-oligoadenylate synthetase activity (adenovirus and herpesviruses);
- production of soluble interferon receptors (pox viruses); and
- interference with interferon signaling (paramyxoviruses).

Antigenic shift in influenza virus

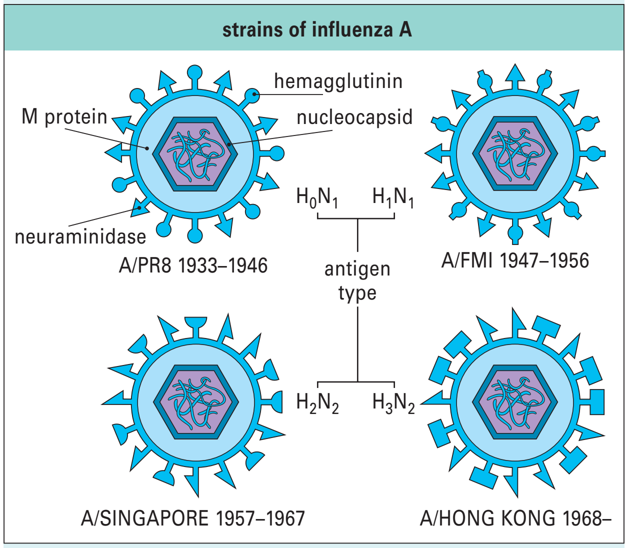

Fig. 13.7 The major surface antigens of influenza virus are hemagglutinin and neuraminidase. Hemagglutinin is involved in attachment to cells, and antibodies to hemagglutinin are protective. Antibodies to neuraminidase are much less effective. The influenza virus can change its antigenic properties slightly (antigenic drift) or radically (antigenic shift). Alterations in the structure of the hemagglutinin antigen render earlier antibodies ineffective and new virus epidemics therefore break out. The diagram shows strains that have emerged by antigenic shift since 1933. The official influenza antigen nomenclature is based on the type of hemagglutinin (H_1, H_2, etc.) and neuraminidase (N_1, N_2, etc.) expressed on the surface of the virion. Note that, although new strains replace old strains, the internal antigens remain largely unchanged.

Viruses also:
- encode other cytokine homologs such as vIL-10 and vIL-6 (herpesviruses); as well as
- receptors to interfere with TNF function.

Q. From this observation, what can you infer about an effective immune response to herpesviruses?

A. The vIL-10 will deviate the immune response, inhibiting a TH1-type reaction with macrophage activation. One can therefore infer that this type of immune response is the key to anti-herpes immunity and the type of response the viruses aims to deflect.

Chemokines represent an important traffic-light system for cell migration and viruses have evolved elaborate strategies for disrupting the chemokine network. The herpesviruses encode:
- chemokine homologs (e.g. CCL3);
- chemokine receptor homologs; and
- chemokine-binding proteins, which have powerful effects on delaying or inhibiting cell migration during inflammation.

Additionally herpes and pox viruses encode homologs not only for CD46 and CD55 – complement regulatory proteins that block C3 activation – but also for CD59.

HIV makes use of cellular CD59, which is incorporated in the viral envelope, thereby blocking complement-mediated lysis of the virion.

Many DNA viruses have strategies to control the expression of MHC molecules

MHC class I expression can be disrupted by:
- blocking peptide uptake into the endoplasmic reticulum (e.g. HSV-1); and
- preventing maturation assembly and migration of the trimolecular MHC complex (e.g. human CMV).

Similar mechanisms apply for MHC class II molecules where:
- some herpesviruses block transcription; whereas
- other mechanisms involve premature targeting of MHC class II for degradation.

Downregulation of MHC class I may disrupt $CD8^+$ T cell recognition, but NK cells are more efficient killers in the absence of MHC class I. Human and murine CMV have tried to redress the balance by encoding their own MHC class I homolog, which is expressed on infected cells and can inhibit NK cell activation.

Examples of virus-encoded homologs of the host defense system are shown in Fig. 13.8.

RESPONSES TO VIRAL ANTIGENS CAN CAUSE TISSUE DAMAGE

Damage may result from the formation of immune complexes

Immune complexes may arise in body fluids or on cell surfaces and are most common during persistent or chronic infections (e.g. with LCMV or hepatitis B virus). Antibody is ineffective (non-neutralizing) in the presence of large amounts of the viral antigen; instead, immune complexes form and are deposited in the kidney or in blood vessels, where they evoke inflammatory responses leading to tissue damage (e.g. as in glomerulonephritis, see Chapter 25).

An unusual pathological consequence of some virus interactions with weakly neutralizing antibody is **antibody-dependent enhancement of virus infection (ADE)**. This involves Fc receptor-mediated uptake of antibody–virus complexes by macrophages and subsequent enhancement of virus infectivity. This is seen in many persistent or viremic virus infections where the common target for replication is the macrophage. An example is Dengue virus infection where cross-reactive antibodies from different Dengue virus subtypes can result in ADE with the consequence of initiating:
- Dengue hemorrhagic fever; and
- Dengue shock syndrome, which results in excessive procoagulant release by monocytes.

CTL responses can cause severe tissue damage

In any virus infection some tissue damage is likely to arise from the activity of infiltrating T cells. However, in some situations this damage may be considerable, resulting in

Examples of viral products that interfere with host defenses

host defense affected	virus	virus product	mechanism
IFN	EBV	EBERS (small RNAs)	blocks protein kinase R activation
	vaccinia	IFNα/β receptor homolog	secreted and binds to IFNα/β
	SV5 – paramyxovirus	V protein	targets STAT-1 degradation
complement	KSHV	ORF-4-RCA homolog	inhibition of C3 activation
	HVS	CD59 homolog	inhibits formation of membrane attack complex
cytokines	EBV	IL-10 homolog	interferes with IFNγ function
	coxpox	TNF receptor homolog	secreted, binds TNF
chemokines	KSHV	MIP-1 homolog	CCR8 agonist and attracts T$_H$2 cells
	HCMV	US29-CK receptor homolog	binds CC CK, mediates cell migration
	MHV-68	M3-CK binding protein	binds C, CC, CXC, CX3C CK blocks cell migration
MHC	adenovirus	E3	blocks transport of MHC class I to surface
	HSV	ICP-47	stops peptide binding to TAP
	HIV	Nef protein	interferes with MHC class II processing
apoptosis	MHV-68	Bcl-2 homolog	inhibits apoptosis
NK cells	HCMV	UL18-MHC class I homolog	inhibits NK cell lysis

Fig. 13.8 Viruses use diverse strategies to outwit host defenses. (CC, chemokine; CXCR and CCR, chemokine receptors; CK, chemokine; HCMV, human cytomegalovirus; HVS, herpesvirus Saimiri; KSHV, Kaposi's sarcoma-associated herpesvirus; MHV-68, murine gammaherpesvirus)

the death of the animal. The best example of this is the CTL responses to LCMV in the central nervous system (Fig. 13.9). Removal of T cells protects the animal from death, indicating that they, rather than the virus, are damaging the brain.

A similar mechanism has been postulated for chronic active hepatitis in humans, whereby CTLs target hepatitis B virus-infected cells and may also participate in a non-viral autoimmune disease.

Viruses can infect cells of the immune system

Some viruses (e.g. HIV) directly infect lymphocytes or macrophages, resulting in pathogenic effects.

Immunocompetent cells are also favored sites of virus persistence. To facilitate persistence in cells of the lymphoid system, some herpesviruses carry a Bcl-2 homolog.

Q. Where have you encountered Bcl-2 previously, and what effect does it have on lymphocytes?

A. Bcl-2 is induced in germinal center B cells and is required to protect the cell from apoptosis (see Fig. 8.12).

In this case the Bcl-2 homolog also prevents apoptosis thereby preserving the life of the infected cell and favoring virus persistence/latency (see above). In the resting state, leukocytes harbor the virus in a non-infectious form, but on activation of the infected cells the virus may also be reactivated to produce infectious virus particles. Example of viruses infecting B cells, T cells, and macrophages are shown in Fig. 13.10.

HIV infects CD4$^+$ cells

Many of the points raised above are illustrated by HIV, the retrovirus that causes AIDS. Infection with HIV is characterized by:

- prolonged clinical latency;
- ineffective immunity;
- continuous virus mutation;
- neuropathology; and
- a tendency to infect bone marrow-derived cells and lymphocytes (see Chapter 19).

HIV is taken up by CD4$^+$ T cells and macrophages following binding of a viral glycoprotein (gp120) to CD4 and certain chemokine receptors (CXCR4 and CCR5). It also enters other antigen-presenting cells by this route. However, entry into cells bearing Fc and complement

Lymphocytic choriomeningitis virus infection (LCMV) in mice

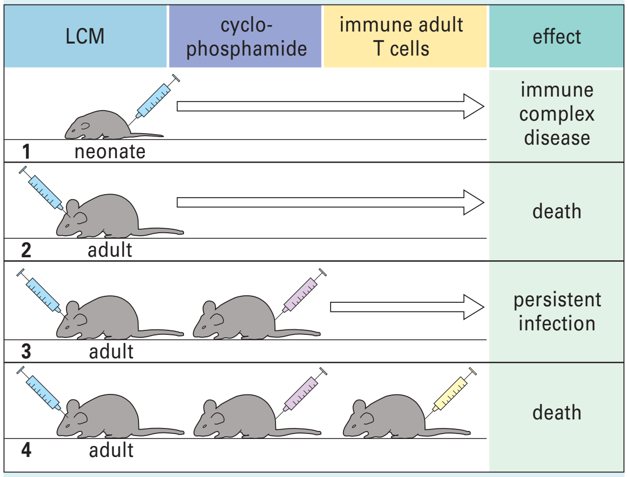

Fig. 13.9 The different effects of LCMV are related to differences in immune status. Infection of neonatal mice (**1**) produces chronic virus shedding and immune complex disease, manifesting itself as glomerulonephritis and vasculitis. Intracerebral infection of adult mice (**2**) results in death. This is due to a T cell reaction because suppression of immunity with cyclophosphamide (**3**) leads to persistent infection, but prevents death. This 'protective' effect produced by cyclophosphamide can be reversed by T cells from an immune animal (**4**).

Virus infection of immunocompetent cells

B lymphocytes	**Epstein–Barr virus** **murine gammaherpesvirus** **infectious bursal disease virus**
T lymphocytes	**human T lymphotropic virus types 1 and 2** **HIV** **measles virus** **herpesvirus Saimiri** **human herpesvirus 6**
macrophages	**Visna virus** **HIV** **lactate dehydrogenase virus** **cytomegalovirus**

Fig. 13.10 Some viruses persist indefinitely in immunocompetent cells. Periodically, this infection may lead to pathological consequences, involving the death of the cell (HIV) or transformation leading to neoplasia (Epstein–Barr virus, HTLV-1).

receptors can be enhanced by antibody (ADE), suggesting that this provides an alternative route into phagocytic cells or enhances entry when CD4 is scarce.

There is a long but variable period of clinical latency, and in about 50% of patients progression to AIDS does not occur for 10 years. During this latent period, HIV can exist as a provirus, integrated within the host's genomic DNA, without any transcription occurring.

Numerous factors can lead to the activation of transcription. In vitro both TNF and IL-6 cause increased production of infectious virus from latently infected T cell lines. This may be important in vivo because monocytes from individuals carrying HIV tend to release abnormally large quantities of these cytokines. It is possible that there is a cycle of TNF and IL-6 release leading to enhanced virus transcription (Fig. 13.11). This could lead to infection of further cells, and release of more cytokine. Production is increased in vitro by other cytokines and lymphokines, and by mitogens and phorbol esters.

Elimination of the HIV virus does not occur for a variety of reasons, including:
- latency;
- viral mutation (giving rapid antigenic drift); and
- progressive immunodeficiency.

Viral infection may provoke autoimmunity

Viruses may trigger autoimmune disease in a number of ways, as follows.
- Virus-induced damage – during the course of some virus infections tissues become damaged, provoking an inflammatory response during which 'hidden' antigens become exposed and can be processed and presented to the immune system. Examples of this include Theiler's virus (a murine picornavirus) and murine hepatitis virus infection of the nervous system, in which the constituents of myelin (the insulating material of axons) become targets for antibody and T cells.

Infection of lymphocytes and macrophages by HIV

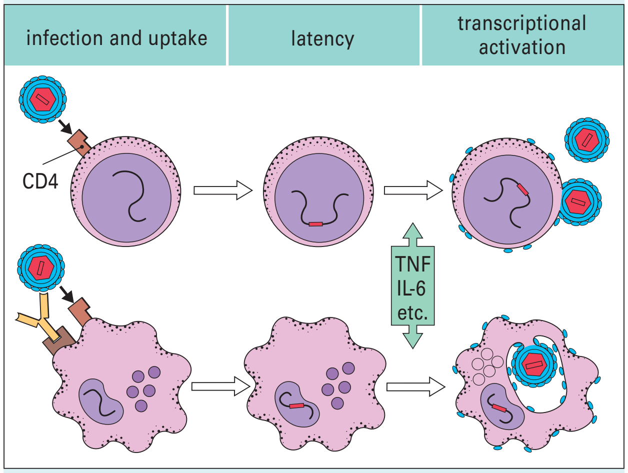

Fig. 13.11 The gp120 on the surface of HIV virions binds to CD4 and chemokine receptors on the lymphocyte membrane, and this triggers uptake. The virus can enter macrophages, which express relatively low levels of CD4, but this may be assisted by binding through antibody to Fc receptors. The virus remains latent, integrated in the host cell's genomic DNA, until some stimulus (e.g. cytokines) causes transcriptional activation. Assembled viruses bud from the outer membrane of T cells or into intracytoplasmic vacuoles of macrophages, where a large reservoir of potentially infectious particles can accumulate.

- Molecular mimicry – a sequence in a viral protein that is homologous to a 'self' protein can become recognized, leading to a breakdown in immunological tolerance to cryptic self antigens in the consequent attack on host tissues by the immune system (see Chapter 26). A good example is Coxsackie B virus-induced myocarditis. Patients with inflammatory cardiomyopathy have antibodies that cross-react with peptides derived from coxsackie B3 protein and with peptides derived from cellular adenine nucleotide translocator.

FURTHER READING

Alcami A, Koszinowski UH. Viral mechanisms of immune evasion. Immunol Today 2000;21:36–50.

Arase H, Mocarski ES, Campbell AE, et al. Direct recognition of cytomegalovirus by activating and inhibitory NK cell receptors. Science 2002;296:1323–1326.

Biron CA, Nguyen KB, Pien GC, et al. Natural killer cells in antiviral defence: function and regulation by innate cytokines. Annu Rev Immunol 1999;17:189–220.

Clements JE, Gidovin SL, Montelaro RC, et al. Antigenic variation in lentiviral disease. Annu Rev Immunol 1988;6:139–159.

Doherty PC, Allan W, Eichelberger M, et al. Roles of α/β and γ/δ T cell subsets in viral immunity. Annu Rev Immunol 1992;10:123–151.

Douek DC, Picker LJ, Koup RA. T cell dynamics in HIV-1 infection. Annu Rev Immunol 2003;21:265–304.

Goodbourn S, Didcock L, Randall RE. Interferons: cell signaling, immune modulation, antiviral responses and virus counter measures. J Gen Virol 2000;81:2341–2364.

Guidotti LG, Chisari FV. Noncytolytic control of viral infections by the innate and adaptive immune response. Annu Rev Immunol 2001;19:65–91.

Klasse PJ, Sattentau QJ. Occupancy and mechanism in antibody-mediated neutralization of animal viruses. J Gen Virol 2002;83:2091–2108

Koszinowski UH, del Val M, Reddehasse MJ. Cellular and molecular basis of the protective immune response to cytomegalovirus infection. Curr Top Microbiol Immunol 1990:154:189–220.

Levy JA. Pathogenesis of human immunodeficiency virus infection. Microbiol Rev 1993;57:183–289.

McMichael A. How viruses hide from T cells. Trends Microbiol 1997;5:211–214.

Mims CA. Interactions of viruses with the immune system. Clin Exp Immunol 1986;66:1–16.

Nash AA, Cambouropoulos P. The immune response to herpes simplex virus. Semin Virol 1993;4:181–186.

Schwimmbeck PL, Fitzgerald NA, Schulze K, et al The humoral immune response in viral heart disease: characterization and pathophysiological significance of antibodies. Med Microbiol Immunol (Berl) 2004;193:115–119.

Tirado SMC, Yoon K-J. Antibody-dependent enhancement of virus infection and disease. Virol Immunol 2003;16:69–86.

Tortorella D, Gewurz BE, Furman MH, et al. Viral subversion of the immune system. Annu Rev Immunol 2000;18:861–926.

Critical thinking: Virus–immune system interactions (see p. 496 for explanations)

1 What are the features of a virus that enable it to evade host defense mechanisms?

A series of IgG monoclonal antibodies were developed against glycoprotein D of herpes simplex virus. When tested in vitro for virus neutralizing activity the antibodies could be divided into two groups: those capable of neutralizing virus infectivity and non-neutralizing antibodies. However, when individual neutralizing or non-neutralizing antibodies were injected into mice infected with herpes simplex virus, both sets of antibodies protected the animals from an overwhelming infection.

2 How do you explain the protection achieved by the non-neutralizing monoclonal antibodies?

3 What experiments would you propose to support some of your conclusions?

CHAPTER

14

Immunity to Bacteria and Fungi

SUMMARY

- **Mechanisms of protection from bacteria can be deduced from their structure and pathogenicity.** There are four main types of bacterial cell wall and pathogenicity varies between two extreme patterns. Non-specific, phylogenetically ancient recognition pathways for conserved bacterial structures trigger protective innate immune responses and guide the development of adaptive immunity.

- **Lymphocyte-independent bacterial recognition pathways have several consequences.** Complement is activated via the alternative pathway. Release of proinflammatory cytokines increases the adhesive properties of the vascular endothelium. Pathogen recognition generates signals that regulate the lymphocyte-mediated response.

- **Antibody provides an antigen-specific protective mechanism.** Neutralizing antibody may be all that is needed for protection if the organism is pathogenic only because of a single toxin or adhesion molecule. Opsonizing antibody responses are particularly important for resistance to extracellular bacterial pathogens. Complement can kill some bacteria, particularly those with an exposed outer lipid bilayer, such as Gram-negative bacteria.

- **Ultimately most bacteria are killed by phagocytes** following a multistage process of chemotaxis, attachment, uptake, and killing. Macrophage killing can be enhanced on activation. Optimal activation of macrophages is dependent on TH1 CD4 T cells. Persistent macrophage recruitment and activation can result in granuloma formation, which is a hallmark of cell-mediated immunity to intracellular bacteria.

- **Successful pathogens have evolved mechanisms to avoid phagocyte-mediated killing** and have evolved a startling diversity of mechanisms for avoiding other aspects of innate and adaptive immunity.

- **Infected cells can be killed by CTLs.** Other T cell populations and some tissue cells can contribute to antibacterial immunity.

- **The response to bacteria can result in immunological tissue damage.** Excessive release of cytokines caused by microorganisms can result in immunopathological syndromes, such as endotoxin shock and the Schwartzman reaction.

- **Fungi can cause life-threatening infections.** Immunity to fungi is predominantly cell mediated and shares many similarities with immunity to bacteria.

MECHANISMS OF PROTECTION FROM BACTERIA CAN BE DEDUCED FROM THEIR STRUCTURE AND PATHOGENICITY

Bacterial infections have had an enormous impact on human society and despite the discovery of antibiotics continue to be a major threat to public health.

Plague caused by *Yersinia pestis* is estimated to have killed one-quarter of the European population in the Middle Ages, whereas infection with *Mycobacterium tuberculosis* is currently a global health emergency.

The immune defense mechanisms elicited against pathogenic bacteria are determined by their:

- surface chemistry;
- mechanism(s) of pathogenicity; and
- whether they are predominantly extracellular or also have the ability to survive inside mammalian cells.

There are four main types of bacterial cell wall

The four main types of bacterial cell wall (Fig. 14.1) belong to the following groups.

- Gram-positive bacteria;
- Gram-negative bacteria;
- mycobacteria;
- spirochetes.

The outer lipid bilayer of Gram-negative organisms is of particular importance because it is often susceptible to lysis by complement. However, killing of most bacteria usually requires uptake by phagocytes. The outer surface of the bacterium may also contain fimbriae or flagellae, or it may be covered by a protective capsule. These can impede the functions of phagocytes or complement, but they also act as targets for the antibody response, the role of which is discussed later.

Bacterial cell walls

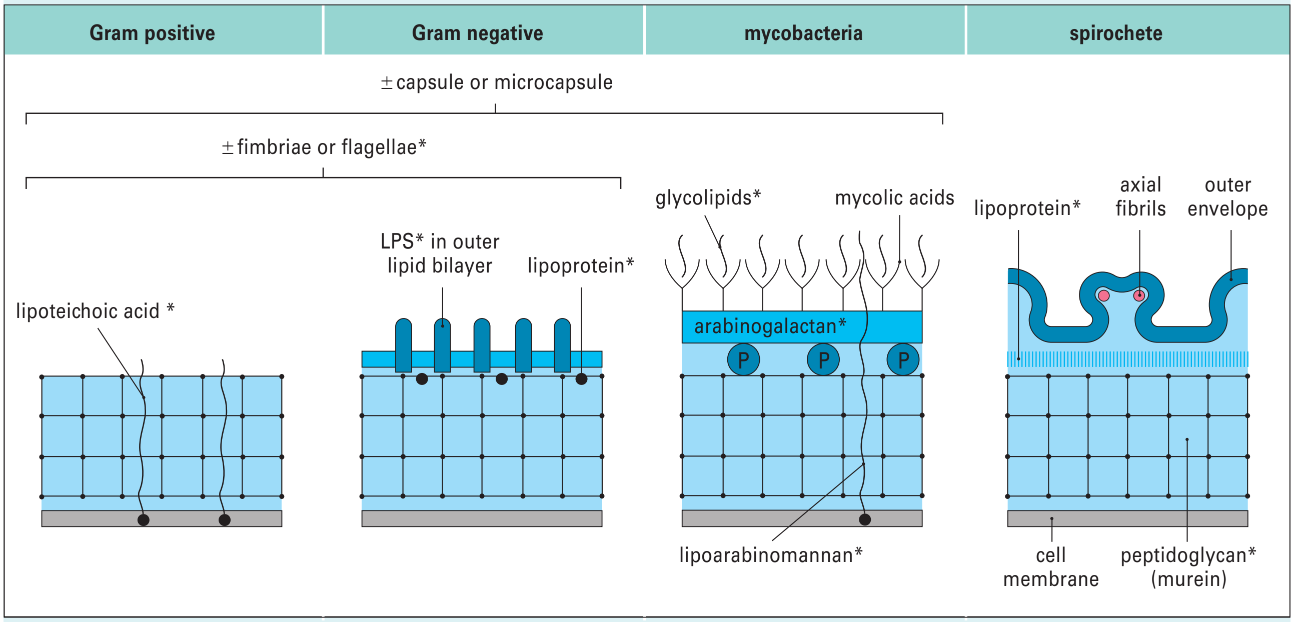

Fig. 14.1 Different immunological mechanisms have evolved to destroy the cell wall structure of the different groups of bacteria. All types have an inner cell membrane and a peptidoglycan wall. Gram-negative bacteria also have an outer lipid bilayer in which lipopolysaccharide (LPS) is embedded. Lysosomal enzymes and lysozyme are active against the peptidoglycan layer, whereas cationic proteins and complement are effective against the outer lipid bilayer of the Gram-negative bacteria. The compound cell wall of mycobacteria is extremely resistant to breakdown, and it is likely that this can be achieved only with the assistance of the bacterial enzymes working from within. Some bacteria also have fimbriae or flagellae, which can provide targets for the antibody response. Others have an outer capsule, which renders the organisms more resistant to phagocytosis or to complement. The components indicated with an asterisk (*) are recognized by the innate immune system as a non-specific 'danger' signal that selectively boosts some aspects of immune activity. (Gram staining is a method that exploits the fact that crystal violet and iodine form a complex that is more abundant on Gram-positive bacteria. The complex easily elutes from Gram-negative bacteria.)

Pathogenicity varies between two extreme patterns

The two extreme patterns of pathogenicity are:

- toxicity without invasiveness;
- invasiveness without toxicity (Fig. 14.2).

However, most bacteria are intermediate between these extremes, having some invasiveness assisted by some locally acting toxins and spreading factors (tissue-degrading enzymes).

Corynebacterium diphtheriae and *Vibrio cholerae* are examples of organisms that are toxic, but not invasive. Because their pathogenicity depends almost entirely on toxin production, neutralizing antibody to the toxin is probably sufficient for immunity, though antibody binding to the bacteria and so blocking their adhesion to the epithelium could also be important.

In contrast, the pathogenicity of most invasive organisms does not rely so heavily on a single toxin, so immunity requires killing of the organisms themselves.

The first lines of defense do not depend on antigen recognition

The body's first line of defense against pathogenic bacteria consists of simple barriers to the entry or establishment of the infection. Thus, the skin and exposed epithelial surfaces have non-specific or innate protective systems, which limit the entry of potentially invasive organisms (see Fig. 1.2).

Intact skin is impenetrable to most bacteria. Additionally, fatty acids produced by the skin are toxic to many organisms. Indeed, the pathogenicity of some strains correlates with their ability to survive on the skin. Epithelial surfaces are cleansed, for example, by ciliary action in the trachea or by flushing of the urinary tract.

Many bacteria are destroyed by pH changes in the stomach and vagina, both of which provide an acidic environment. In the vagina, the epithelium secretes glycogen, which is metabolized by particular species of commensal bacteria, producing lactic acid.

Commensals can limit pathogen invasion

Commensals can limit pathogen invasion through the production of antibacterial proteins termed **colicins**. Commensals may therefore occupy an ecological niche that would otherwise be occupied by something more unpleasant.

When the normal flora are disturbed by antibiotics, infections by *Candida* spp. or *Clostridium difficile* can occur,

Mechanisms of immunopathogenicity

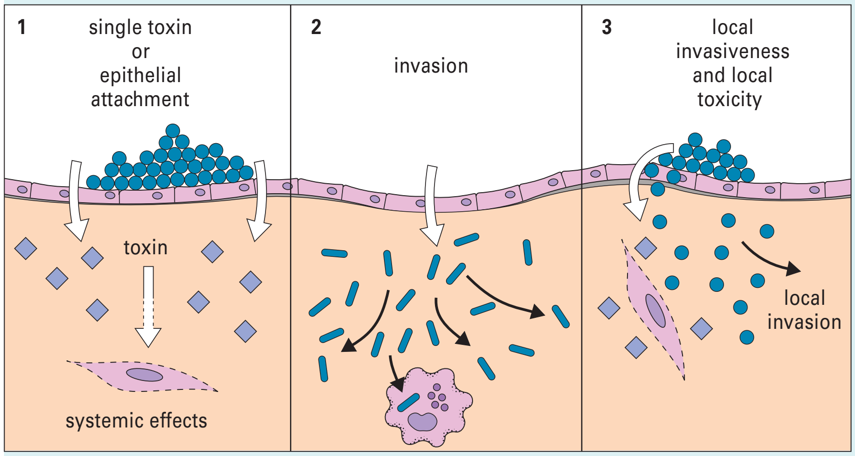
1
single toxin or epithelial attachment
toxin
systemic effects
2
invasion
3
local invasiveness and local toxicity
local invasion

Fig. 14.2 (**1**) Some bacteria cause disease as a result of only a single toxin (e.g. *Corynebacterium diphtheriae, Clostridium tetani*) or because of an ability to attach to epithelial surfaces (e.g. in group A streptococcal sore throat). Immunity to such organisms may require only antibody to neutralize this critical function. (**2**) At the other extreme there are organisms that are not toxic, and cause disease by invasion of tissues and sometimes cells, where damage results mostly from the bulk of organisms or from immunopathology (e.g. lepromatous leprosy). Where organisms invade cells, they must be destroyed and degraded by the cell-mediated immune response. (**3**) Most organisms fall between the two extremes, with some local invasiveness assisted by local toxicity and enzymes that degrade extracellular matrix (e.g. *Staphylococcus aureus, Clostridium perfringens*). Antibody and cell-mediated responses are both involved in resistance.

and the latter is a major cause of antibiotic-induced colitis and diarrhea.

Several studies suggest that the reintroduction of non-pathogenic 'probiotic' organisms such as lactobacilli into the intestinal tract can alleviate the symptoms, presumably by replacing those killed by the antibiotics.

In practice, only a minute proportion of the potentially pathogenic organisms around us ever succeed in gaining access to the tissues.

The second line of defense is mediated by recognition of bacterial components

If organisms do enter the tissues, they can be combated initially by further elements of the innate immune system.

Numerous bacterial components are recognized in ways that do not rely on the antigen-specific receptors of either B cells or T cells. These types of recognition are phylogenetically ancient 'broad-spectrum' mechanisms that evolved before antigen-specific T cells and immunoglobulins, allowing protective responses to be triggered by common microbial components bearing so-called **'pathogen-associated molecular patterns' (PAMPs)**. The host molecules that recognize these microbial components are referred to as the **'pattern recognition molecules'** of the innate immune system.

Q. List some examples of soluble molecules, cell surface receptors, and intracellular molecules that recognize PAMPs.

A. Collectins and ficolins (see Fig. 6.23), the Toll-like receptors (see Fig. 6.24), the mannose receptor (see Fig. 9.18), and the NOD proteins (see Fig. 9.24) all recognize PAMPs.

Many organisms, such as non-pathogenic cocci, are probably removed from the tissues as a consequence of these pathways, without the need for a specific adaptive immune reaction. Fig. 14.3 shows some of the microbial components involved and the host responses that are triggered.

The immune system has selected these structures for recognition because they are not only characteristic of microbes, but are essential for their growth and cannot be easily mutated to evade discovery (though, as might be predicted, there are increasing examples of pathogen strategies that attempt to subvert this process).

It is interesting to note that the 'Limulus assay', which is used to detect contaminating lipopolysaccharide (LPS) in preparations for use in humans, is based on one such recognition pathway found in an invertebrate species. In *Limulus polyphemus* (the horseshoe crab), tiny quantities of LPS trigger fibrin formation, which walls off the LPS-bearing infectious agent.

Bacterial PAMPs activate cells via Toll-like receptors

Many bacterial PAMPs activate cells via Toll-like receptors (TLRs). These are homologs of a receptor mediating antimicrobial immune responses in the fruit fly (*Drosophila* spp.).

The TLR family is made up of at least ten different TLR molecules which differ in the microbial structures they recognize. The most prominent TLRs involved in recognition of bacterial components are TLR1, 2, 4, 5, 6, and 9 (see below).

TLRs are preferentially expressed on phagocytes, dendritic cells, and epithelial cells at sites of bacterial entry

Protective mechanisms not involving antigen-specific B or T cells

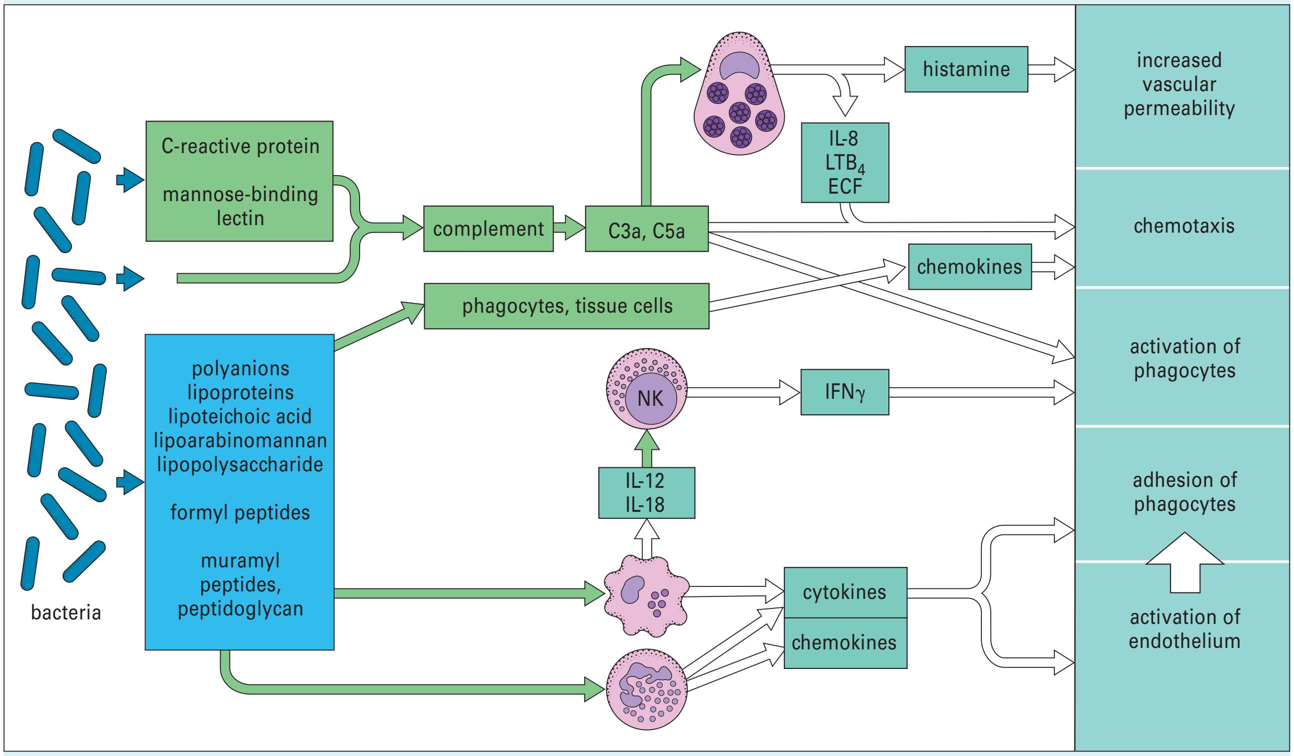
histamine
increased vascular permeability
C-reactive protein
mannose-binding lectin
IL-8
LTB_4
ECF
complement
C3a, C5a
chemotaxis
chemokines
phagocytes, tissue cells
activation of phagocytes
polyanions
lipoproteins
lipoteichoic acid
lipoarabinomannan
lipopolysaccharide
formyl peptides
muramyl peptides, peptidoglycan
NK
IFNγ
IL-12
IL-18
adhesion of phagocytes
cytokines
chemokines
activation of endothelium
bacteria

Fig. 14.3 Several common bacterial PAMPs are recognized by molecules present in serum and by receptors on cells. These recognition pathways result in activation of the alternative complement pathway (factors C3, B, D, P), with consequent release of C3a and C5a; activation of neutrophils, macrophages, and NK cells; triggering of cytokine and chemokine release; mast cell degranulation, leading to increased blood flow in the local capillary network; and increased adhesion of cells and fibrin to endothelial cells. These mechanisms, plus tissue injury caused by the bacteria, may activate the clotting system and fibrin formation, which limit bacterial spread.

to the host. Each cell type can express a different combination of receptors and this repertoire can be altered by inflammatory stimuli, allowing the greatest possible recognition coverage to a diverse range of pathogens.

Other important pattern recognition receptors include:

- the mannose receptor; and
- scavenger receptors.

Of course, recognition of bacteria also occurs in the absence of cells via:

- complement;
- C-reactive protein (CRP);
- mannose-binding lectin in the blood;
- surfactant protein A in the lungs.

LPS is the dominant activator of innate immunity in Gram-negative bacteria

Injection of pure LPS into mice or even humans is sufficient to mimic most of the features of acute Gram-negative infection, including massive production of proinflammatory cytokines, such as interleukin-1 (IL-1), IL-6, and tumor necrosis factor (TNF), leading to severe shock.

Q. How does release of proinflammatory cytokines cause shock?

A. These proinflammatory cytokines act directly on endothelium to increase vascular adhesiveness, and indirectly activate other plasma enzyme systems to release vasoactive peptides and amines leading to a drop in blood pressure.

Recognition of LPS is a complex process involving molecules that bind LPS and pass it on to cell membrane-associated receptors on leukocytes, and endothelial and other cells, which initiate this proinflammatory cascade; these events are illustrated in Fig. 14.4.

Binding of LPS to TLR4 is a critical event in immune activation. TLR4 knockout mice are resistant to LPS-induced shock and there is some evidence that polymorphisms in human TLR4 may influence the course of infection with these bacteria.

Other bacterial components are also potent immune activators

Gram-positive bacteria do not possess LPS yet still induce intense inflammatory responses and severe infection.

Effects of lipopolysaccharide

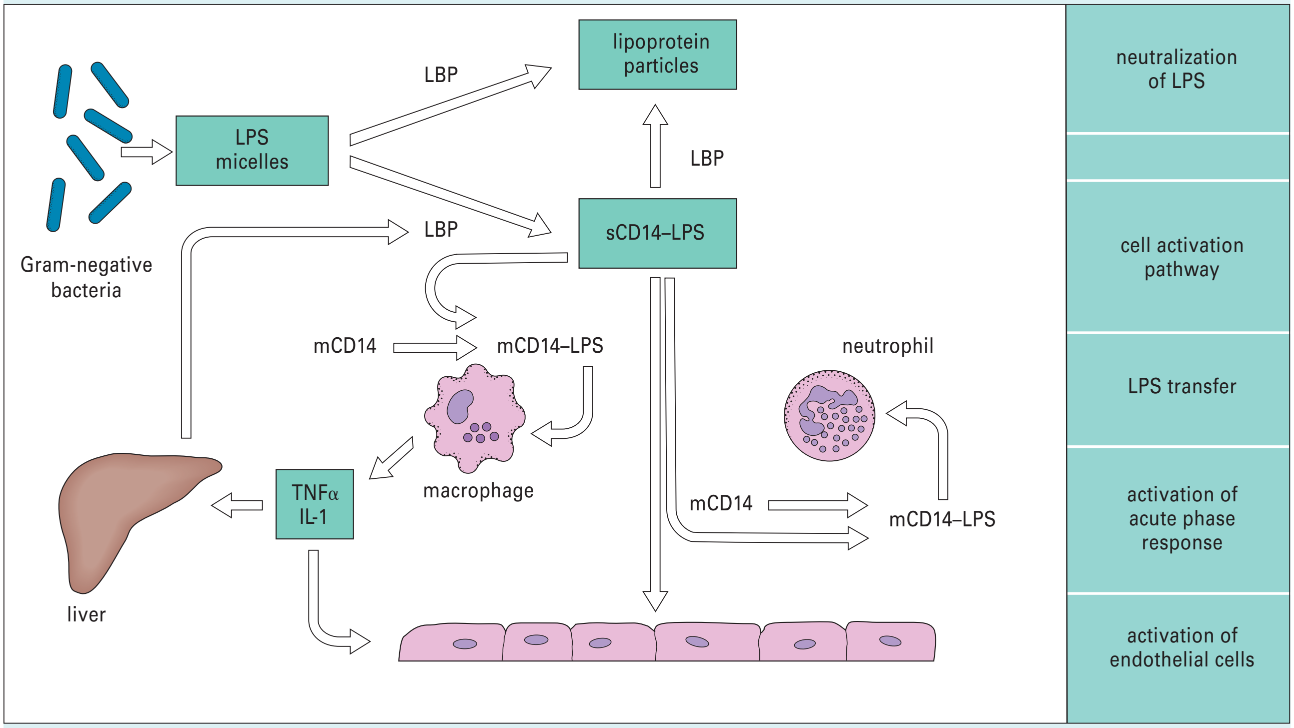

Fig. 14.4 LPS released from Gram-negative bacteria becomes bound to LPS-binding protein (LBP) which promotes transfer of LPS to either soluble CD14 (sCD14) or to a GPI-linked membrane form of the protein (mCD14) expressed on neutrophils and macrophages. LBP then dissociates and the mCD14–LPS complex, in association with Toll-like receptor 4 (TLR4) and MD2, transduces signals that cause increased expression of integrins (adhesion molecules) and increased release of many proinflammatory cytokines including TNFα and IL-1. These in turn activate endothelial cells and drive the acute phase response in the liver. One product of the acute phase response is further LBP.

Most capsular polysaccharides are not potent activators of inflammation (though some can activate macrophages) and instead attempt to shield the bacterium from host immune defenses.

Components of the cell wall, including peptidoglycans and lipoteichoic acids, are the dominant activators of innate immunity, and TLR2, often in cooperation with TLR1 or TLR6, is the major TLR involved.

The LBP and CD14, which bind LPS, are also involved in recognition of lipid-containing bacterial components from mycoplasmas, mycobacteria, and spirochetes.

Other molecules that trigger innate immunity include mycoplasma lipoproteins (via TLR 2/6), flagellin (via TLR5), and DNA (due to its distinct **CpG motifs**) via TLR9.

Most pattern recognition receptors are expressed on the plasma membrane of cells, making contact with microbes during the process of binding and/or phagocytosis.

However, others are designed to detect intracellular pathogens and their products inside phagosomes (such as TLR9) or in the cytosol.

Q. Which proteins can recognize pathogens in the cytosol, and which pathogen components?

A. NOD-1 and NOD-2 proteins recognize peptidoglycans of both Gram-positive and Gram-negative bacteria (see Fig. 9.24).

Epithelial cells of the gut and lung have few TLRs on their luminal surface, but can be triggered by pathogens that:

- actively invade the cell (such as *Listeria* spp.);
- inject their components (such as *Helicobacter pylori*); or
- actively reach the basolateral surface (e.g. *Salmonella* spp.).

This may explain why constant exposure to non-pathogenic microbes in the intestine and airways does not induce a chronic state of inflammation – the host waits until they move beyond the lumen, signifying the presence of a real pathogenic threat.

LYMPHOCYTE-INDEPENDENT BACTERIAL RECOGNITION PATHWAYS HAVE SEVERAL CONSEQUENCES

Complement is activated via the alternative pathway

Complement activation can result in the killing of some bacteria, particularly those with an outer lipid bilayer susceptible to the **lytic complex (C5b–9)**.

Q. To which strains of bacteria are individuals with C9 deficiency more susceptible?

A. *Neisseria* spp. (see Fig. 4.16).

Perhaps more importantly, complement activation releases C5a, which attracts and activates neutrophils and causes degranulation of mast cells (see Chapter 3). The consequent release of **histamine** and **leukotriene (LTB_4)** contributes to further increases in vascular permeability (see Fig. 14.3).

Opsonization of the bacteria, by attachment of **cleaved derivatives of C3**, is also critically important in subsequent interactions with phagocytes.

Release of proinflammatory cytokines increases the adhesive properties of the vascular endothelium

The rapid release of cytokines such as TNF and IL-1 (see Fig. 14.4) from macrophages increases the adhesive properties of the vascular endothelium and facilitates the passage of more phagocytes into inflamed tissue. Combined with the release of chemokines such as CCL2, CCL3, and CXCL8 (see Chapter 6), this directs the recruitment of different leukocyte populations.

Epithelial cells, neutrophils, and mast cells are also important sources of proinflammatory cytokines.

IL-1, TNF, and IL-6 also initiate the **acute phase response**, increasing the production of complement components as well as other proteins involved in scavenging material released by tissue damage and, in the case of CRP, an opsonin for improving phagocytosis of bacteria.

When NK cells are stimulated by the phagocyte-derived cytokines **IL-12** and **IL-18** they rapidly release large quantities of interferon-γ (IFNγ). This response happens within the first day of infection, well before the clonal expansion of antigen-specific T cells, and provides a rapid source of IFNγ to activate macrophages. This T cell-independent pathway helps to explain the considerable resistance of mice with SCID (severe combined immune deficiency, a defect in lymphocyte maturation) to infections such as with *Listeria monocytogenes*. In mice, CD1d-restricted NK T cells also secrete IFNγ in response to IL-12 and IL-18 and other ligands, and help to further activate both NK cells and macrophages.

Pathogen recognition generates signals that regulate the lymphocyte-mediated response

The signals generated following the recognition of pathogens not only generate a cascade of innate immune events, but also regulate the development of the appropriate lymphocyte-mediated response.

Dendritic cells (DCs) are crucial for the initial priming of naive T cells specific for bacterial antigens. Contact with bacteria in the periphery induces immature DCs to migrate to the draining lymph nodes and augments their antigen-presenting ability by increasing their:

- display of MHC molecule–peptide complexes;
- expression of co-stimulatory molecules (such as CD40, CD80, and CD86); and
- secretion of T cell differentiating cytokines.

Some of this DC activation occurs secondary to their production of cytokines such as type I IFN.

Activated macrophages also act as antigen-presenting cells (APCs), but probably function more at the site of infection, providing further activation of effector rather than naive T cells.

Binding of bacterial components to pattern recognition receptors such as TLRs induces a local environment rich in cytokines such as IFNγ, IL-12, and IL-18, which promote T cell differentiation down the TH1 rather than TH2 pathway.

Immunologists have made use of these effects for many decades (even without knowing their true molecular basis) in the use of **adjuvants** in vaccination. 'Adjuvant' is derived from the Latin *adiuvare*, to help. When given experimentally, soluble antigens evoke stronger T and B cell-mediated responses if they are mixed with bacterial components that act as adjuvants. Components with this property are indicated in Fig. 14.1. This effect probably reflects that the antigen-specific immune response evolved in a tissue environment that already contained these pharmacologically active bacterial components.

With the exception of proteins such as **flagellin**, which itself stimulates TLR5 and is also a strong T cell immunogen, the response to pure bacterial antigens, injected without adjuvant-active bacterial components, is essentially an artificial situation that does not occur in nature.

The best known adjuvant in laboratory use, **complete Freund's adjuvant**, consists of killed mycobacteria suspended in oil, which is then emulsified with the aqueous antigen solution.

New-generation adjuvants based on bacterial components (and safe to use in humans, unlike Freund's adjuvant) include synthetic TLR activators such as CpG motifs and monophosphoryl lipid A (MPL) as well as recombinant cytokines such as IL-12, IL-1, and IFNγ.

ANTIBODY PROVIDES AN ANTIGEN-SPECIFIC PROTECTIVE MECHANISM

The relevance to protection of interactions of bacteria with antibody depends on the mechanism of pathogenicity.

Antibody clearly plays a crucial role in dealing with bacterial toxins:

- it neutralizes diphtheria toxin by blocking the attachment of the binding portion of the molecule to its target cells;
- similarly it may block locally acting toxins or extracellular matrix-degrading enzymes, which act as spreading factors;

Antibody can also interfere with motility by binding to flagellae.

An important function on external and mucosal surfaces, often performed by secretory IgA (sIgA, see Chapter 3), is to stop bacteria binding to epithelial cells – for instance, antibody to the M proteins of group A streptococci gives type-specific immunity to streptococcal sore throats.

It is likely that some antibodies to the bacterial surface can block functional requirements of the organism such as binding of iron-chelating compounds or intake of nutrients (Fig. 14.5).

An important role of antibody in immunity to non-toxigenic bacteria is the more efficient targeting of complement.

The antibacterial roles of antibody

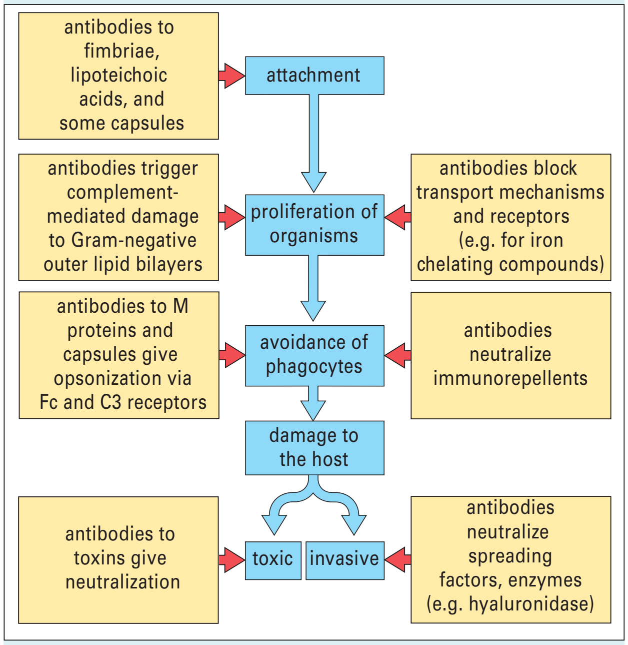
antibodies to fimbriae, lipoteichoic acids, and some capsules
attachment
antibodies trigger complement-mediated damage to Gram-negative outer lipid bilayers
proliferation of organisms
antibodies block transport mechanisms and receptors (e.g. for iron chelating compounds)
antibodies to M proteins and capsules give opsonization via Fc and C3 receptors
avoidance of phagocytes
antibodies neutralize immunorepellents
damage to the host
antibodies to toxins give neutralization
toxic
invasive
antibodies neutralize spreading factors, enzymes (e.g. hyaluronidase)

Fig. 14.5 This diagram lists the stages of bacterial invasion (blue) and indicates the antibacterial effects of antibody (yellow) that operate at the different stages. Antibodies to fimbriae, lipoteichoic acid, and some capsules block attachment of the bacterium to the host cell membrane. Antibody triggers complement-mediated damage to Gram-negative outer lipid bilayers. Antibody directly blocks bacterial surface proteins that pick up useful molecules from the environment and transport them across the membrane. Antibody to M proteins and capsules opsonizes the bacteria via Fc and C3 receptors for phagocytosis. Bacterial factors that interfere with normal chemotaxis or phagocytosis are neutralized. Bacterial toxins may be neutralized by antibody, as may bacterial spreading factors that facilitate invasion (e.g. by the destruction of connective tissue or fibrin).

Naturally occurring IgM antibodies, which bind to common bacterial structures such as phosphorylcholine, are important for protection against some bacteria (particularly streptococci) via their complement fixing activity.

Specific, high-affinity IgG antibodies elicited in response to infection are most important – children with primary immune deficiencies in B cell development or in T cell help have increased susceptibility to extracellular rather than intracellular bacteria.

With the aid of antibodies, even organisms that resist the alternative (i.e. innate) complement pathway (see below) are damaged by complement or become coated with C3 products, which then enhance the binding and uptake by phagocytes (Figs 14.6 and 14.7).

Q. How do C3 products attach to pathogens?

A. Following activation by cleavage the larger fragment, **C3b**, attaches covalently to hydroxyl and amine groups on the target (see Fig. 4.6).

Effect of antibody and complement on rate of clearance of virulent bacteria from the blood

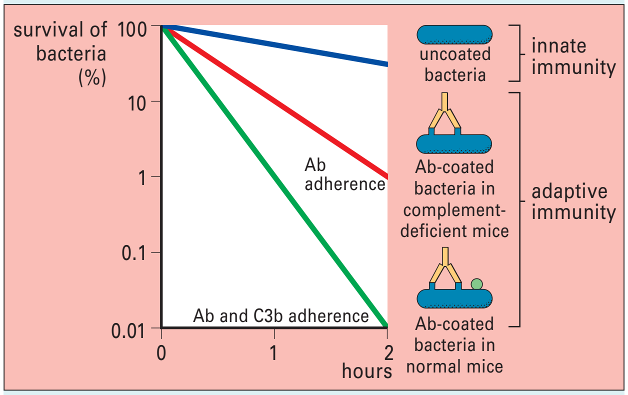

Fig. 14.6 Uncoated bacteria are phagocytosed rather slowly (unless the alternative complement pathway is activated by the strain of bacterium); on coating with antibody (Ab), adherence to phagocytes is greatly increased. The adherence is somewhat less effective in animals temporarily depleted of complement.

The interaction between bacteria and phagocytic cells

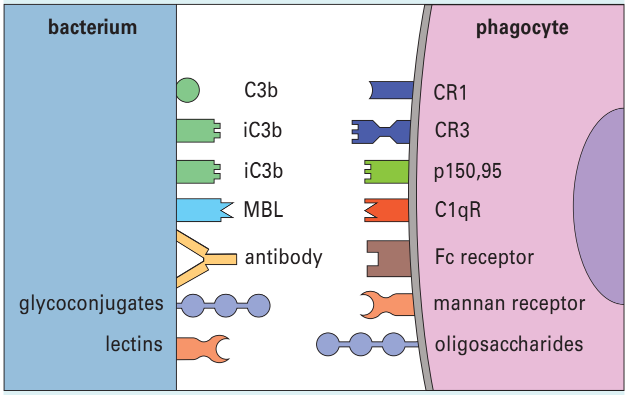
bacterium
phagocyte
C3b
CR1
iC3b
CR3
iC3b
p150,95
MBL
C1qR
antibody
Fc receptor
glycoconjugates
mannan receptor
lectins
oligosaccharides

Fig. 14.7 A variety of molecules facilitate the binding of the organisms to the phagocyte membrane. These are in addition to the TLR system (e.g. TLR4 for LPS, TLR5 for flagellin, and TLR2 [plus TLR1/TLR6] for bacterial lipoproteins and peptidoglycans). The precise nature of the interaction will determine whether uptake occurs and whether cytokine secretion and appropriate killing mechanisms are triggered. Recognition invariably involves combinations of different receptor families. Note that apart from complement, antibody, and mannose-binding lectin (MBL), which bind to the bacterial surface, the other components are constitutive bacterial molecules.

The most efficient **complement-fixing antibodies** in humans are IgM, then IgG3 and to a lesser extent IgG1, whereas IgG1 and IgG3 are the subclasses with the highest affinity for Fc receptors.

Pathogenic bacteria may avoid the effects of antibody

Neisseria gonorrhoeae is an example of a pathogenic bacterium that uses several immune evasion strategies

Mechanisms used by *Neisseria gonorrhoeae* to avoid the effects of antibody

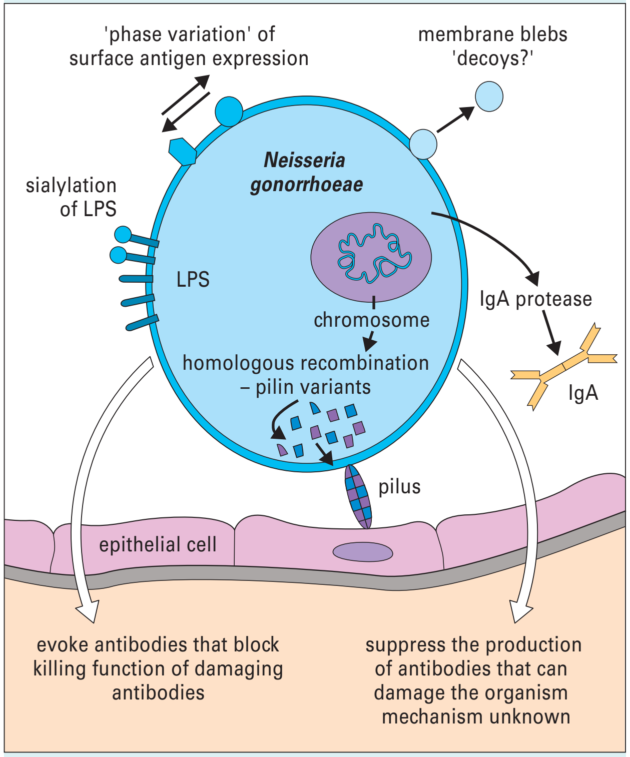
'phase variation' of surface antigen expression
membrane blebs 'decoys?'
sialylation of LPS
Neisseria gonorrhoeae
LPS
chromosome
IgA protease
homologous recombination – pilin variants
IgA
pilus
epithelial cell
evoke antibodies that block killing function of damaging antibodies
suppress the production of antibodies that can damage the organism mechanism unknown

Fig. 14.8 *N. gonorrhoeae* is an example of a bacterium that uses several strategies to avoid the damaging effects of antibody. First, it fails to evoke a large antibody response, and the antibody that does form tends to block the function of damaging antibodies. Second, the organism secretes an IgA protease to destroy antibody. Third, blebs of membrane are released, and these appear to adsorb and so deplete local antibody levels. Finally, the organism uses three strategies to alter its antigenic composition: (i) the LPS may be sialylated, so that it more closely resembles mammalian oligosaccharides and promotes rapid removal of complement; (ii) the organism can undergo phase variation, so that it expresses an alternative set of surface molecules; (iii) the gene encoding pilin, the subunits of the pilus, undergoes homologous recombination to generate variants. *N. gonorrhoeae* also impairs T cell activation by engaging a co-inhibitory receptor CEACAM-1 on the lymphocyte surface by one of its OPA proteins.

Avoidance of complement-mediated damage

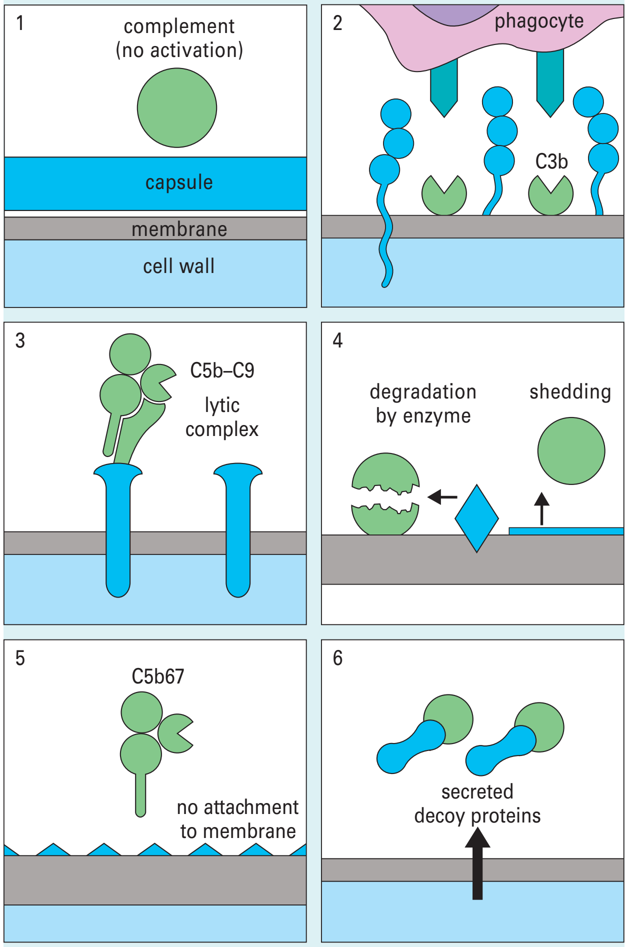
1
complement (no activation)
capsule
membrane
cell wall
2
phagocyte
C3b
3
C5b–C9 lytic complex
4
degradation by enzyme
shedding
5
C5b67
no attachment to membrane
6
secreted decoy proteins

Fig. 14.9 Bacteria avoid complement-mediated damage by a variety of strategies. (**1**) An outer capsule or coat prevents complement activation. (**2**) An outer surface can be configured so that complement receptors on phagocytes cannot obtain access to fixed C3b. (**3**) Surface structures can be expressed that divert attachment of the lytic complex (MAC) from the cell membrane. (**4**) Membrane-bound enzyme can degrade fixed complement or cause it to be shed. (**5**) The outer membrane can resist the insertion of the lytic complex. (**6**) Secreted decoy proteins can cause complement to be deposited on them and not on the bacterium itself.

(Fig. 14.8) and humans can be repeatedly infected with *N. gonorrhoeae* with no evidence of protective immunity.

Antibodies may also be important for effective immunity against some intracellular bacteria such as *Legionella* and *Salmonella* spp.

Pathogenic bacteria can avoid the detrimental effects of complement

Some bacterial capsules are very poor activators of the alternative pathway (Fig. 14.9).

For other bacteria, long side chains (O antigens) on their LPS may fix C3b at a distance from the otherwise vulnerable lipid bilayer. Similarly, smooth-surfaced Gram-negative organisms (*Escherichia coli*, *Salmonella* spp., *Pseudomonas* spp.) may fix but then rapidly shed the C5b–C9 membrane lytic complex.

Other organisms exploit the physiological mechanisms that block destruction of host cells by complement. When C3b has attached to a surface it can interact with factor B leading to further C3b amplification or it can become inactivated by factors H and I. Capsules rich in sialic acid (as host cell membranes are) seem to promote the interaction with factors H and I.

Neisseria meningitidis, *E. coli* K1, and group B streptococci all resist complement attachment in this way.

The M protein of group A streptococci acts as an acceptor for factor H, thus potentiating C3bB dissociation. These bacteria also have a gene for a C5a protease.

Q. What value is a C5a protease to bacteria?

A. C5a is a major chemotactic molecule generated by complement activation that acts on specific receptors on macrophages, neutrophils, and mast cells (see Fig. 4.13).

ULTIMATELY MOST BACTERIA ARE KILLED BY PHAGOCYTES

A few, mostly Gram-negative, bacteria are directly killed by complement. However, immunity to most bacteria, whether considered as extracellular or intracellular pathogens, ultimately needs the killing activity of phagocytes. This process involves several steps.

Bacterial components attract phagocytes by chemotaxis

Unlike neutrophils, which in the uninfected host are found almost entirely in the blood, resident macrophages are constitutively present in tissues where exposure to pathogens first occurs (such as alveolar macrophages in the lung and Kupffer cells in the liver). These macrophages have some killing activity, but invariably need to be supplemented by recruitment of neutrophils and/or monocytes across the blood vessel wall.

Phagocytes are attracted by:

- bacterial components such as **f-Met-Leu-Phe** (which is chemotactic for leukocytes);
- complement products such as C5a; and
- locally released chemokines and cytokines derived from resident macrophages and epithelial cells (see Chapter 6).

The cellular composition of this inflammatory response varies according to the pathogen and the time since infection. For instance:

- acute infection with encapsulated bacteria such as *Streptococcus pyogenes* give rise to tissue lesions rich in neutrophils (typical of so-called pyogenic or pus-forming infections);
- at the other extreme, chronic infections with *M. tuberculosis* result in granulomas rich in macrophages, macrophage-derived multinucleated giant cells, and T cells;
- other organisms, such as *Listeria* and *Salmonella* spp., result in lesions of more mixed composition.

The choice of receptors is critical

The choice of receptors used for attachment of the phagocyte to the organism is critical and will determine:

- the efficiency of uptake;
- whether killing mechanisms are triggered;
- whether the process favors the pathogen by subverting immunity.

The binding can be mediated by lectins on the organism (e.g. on the fimbriae *of E. coli*), but receptors on the phagocyte are the most important. These either bind directly to the bacterium or indirectly via host complement and antibody deposited on the bacterial surface (**opsonization**).

- direct binding is mediated by pattern recognition molecules including Toll-like receptors and scavenger receptors (such as SRA, MARCO), mannose receptor, and dectin-1b;
- opsonization is mediated through complement receptors such as CR1, CR3, and CR4, which recognize complement fragments deposited on the organism via the alternative or classic complement pathways.

Complement can also be fixed by MBL present in serum, which can itself bind to **C1q receptors** and CR1.

Additionally, Fc receptors on the phagocyte (**FcγRI, FcγRII, and FcγRIII**, see Chapter 3) bind antibody that has coated bacteria (see Fig. 14.7), whereas various integrins can bind **fibronectin** and **vitronectin** opsonized particles.

Uptake can be enhanced by macrophage-activating cytokines

The binding of an organism to a receptor on the macrophage membrane does not always lead to its uptake. For example, zymosan particles (derived from yeast) bind via the glucan-recognizing lectin-like site on the CR3 of the macrophage and are taken up, whereas erythrocytes coated with iC3b are not, even though the iC3b also binds to CR3. This can, however, be enhanced by macrophage-activating cytokines such as **granulocyte–macrophage colony stimulating factor (GM-CSF)**.

Different membrane receptors vary in their efficiency at inducing a microbicidal response

Just as the binding of an organism to membrane receptors does not guarantee uptake, so different membrane receptors vary in their efficiency at inducing a microbicidal response – for example, mannose receptors and Fc receptors are particularly efficient at inducing the respiratory burst, but complement receptors are not, providing an evasion strategy for some organisms.

Phagocytic cells have many killing methods

The killing pathways of phagocytic cells can be:

- oxygen dependent (see Chapter 9); or
- oxygen independent.

In summary, one oxygen-dependent pathway involves the reduction of oxygen to superoxide anion (which is molecular oxygen to which a single unpaired electron has been added). This then interacts with numerous other molecules to give rise to a series of free radicals and other toxic derivatives, which can kill bacteria and fungi.

Recent studies in neutrophils suggests that the oxidative burst may also act indirectly, by promoting the flux of K^+ ions into the phagosome and activating microbicidal proteases.

A second oxygen-dependent pathway involves the creation of nitric oxide ($NO^\bullet$) from the guanidino nitrogen of L-arginine. This in turn leads to further toxic substances such as the peroxynitrites, which result from interactions of $NO^\bullet$ with the products of the oxygen reduction pathway.

Q. What pathway leads to the production of NO by macrophages?

A. Cytokine activation by IFNγ and TNFα leads to production of inducible NO synthase, which generates NO from L-arginine (see Fig. 9.31).

Oxygen-independent killing mechanisms may be more important than previously thought. Many organisms can be killed by cells from patients with chronic granulomatous disease (CGD), which cannot produce reactive oxygen intermediates, and from patients with myeloperoxidase (MPO) deficiency, which cannot produce hypohalous acids. Some of this killing may be due to NO$^\bullet$, but many organisms can be killed anaerobically, so other mechanisms must exist. Some have been identified and are discussed below.

Some cationic proteins have antibiotic-like properties

The **defensins** (Fig. 14.10) are cysteine- and arginine-rich cationic peptides of 30 to 33 amino acids found in phagocytes such as neutrophils, where they comprise 30–50% of the granule proteins.

Q. Which other cells secrete antimicrobial peptides?

A. Paneth cells of the intestine (see Fig. 12.8) and airway epithelial cells (i.e. sites of primary contact with pathogens).

Defensins evolved early in evolution and similar molecules are found in insects. They act by integrating into microbial lipid membranes (in some cases forming ion-permeable channels) and disrupting membrane function and structure, resulting in lysis of the pathogen. Defensins can kill organisms as diverse as *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *E. coli*, *Cryptococcus neoformans*, and the enveloped virus herpes simplex.

Cationic host defense peptides in immunity to fungi and bacteria

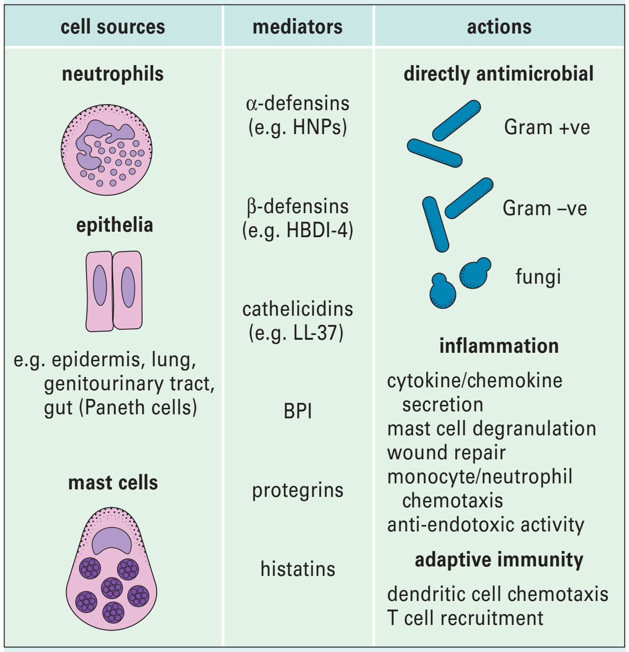

Fig. 14.10 Numerous cationic host defense peptides are produced by neutrophils, epithelia, and mast cells. Their synthesis is usually constitutive but is also enhanced by proinflammatory cytokines such as TNF generated following infection. Originally defined by their direct killing activity against pathogens, they are now known also to act on immune cells, having multiple immunomodulatory effects on inflammation and adaptive immunity.

Defensins also have important immunostimulatory properties including:

- promoting chemotaxis and phagocytosis;
- regulating cytokine production; and
- acting as adjuvants for adaptive immunity.

Other antibacterial peptides include the **cathelicidins** and **protegrins**, which can bind LPS and also form membrane pores.

There are also cationic proteins with different pH optima, including **cathepsin G** and **azurocidin**, both of which are related to elastase, but have activity against Gram-negative bacteria – this is unrelated to their enzyme activity.

Other antimicrobial mechanisms also play a role

Following lysosome fusion there is a transient rise in pH before acidification (a fall in pH) of the phagolysosome takes place. This occurs within 10–15 minutes.

The acidification of phagosomes containing bacteria following their fusion with lysosomes is an important step in the killing process and is related to the low pH optima of lysosomal enzymes.

Certain Gram-positive organisms may be killed by **lysozyme**, which is active against their readily exposed peptidoglycan layer.

The availability of intracellular iron is another important factor in the interplay between host and pathogen. Iron is essential for the growth of many bacteria and also influences their expression of key virulence genes. Sequestration of iron can therefore be an effective antimicrobial strategy, particularly for intracellular bacteria.

Lactoferrin is a mammalian iron-binding protein released by degranulating neutrophils that sequesters iron from pathogens, inhibiting their growth, and in the case of *P. aeruginosa* also reducing biofilm formation, a key event in the pathogenesis of infection in cystic fibrosis patients. **Lactoferricin**, an antimicrobial peptide derived from lactoferrin, kills other bacteria.

Iron is also required for many host immune functions including the respiratory burst, the generation of NO$^\bullet$, and the development of pathogen-specific T cells.

Both iron excess and iron deficiency can therefore have complex effects on the outcome of infection. For example, individuals with iron overload syndromes resulting from genetic defects (such as thalassemia or hemochromatosis), nutritional excess, or following iron or red cell supplementation (such as in the treatment of anemias) have increased susceptibility to infection with *Yersinia* and *Salmonella* spp., and *M. tuberculosis*.

Macrophage killing can be enhanced on activation

Unlike neutrophils, which have a short life span but are efficient killers even in their normal state, macrophages are long-lived cells that without appropriate activation can actually provide a haven for microbial growth.

Macrophage activation occurs most effectively by the combination of exposure to:

- microbial products (through the receptors described above); and
- cytokines (particularly IFNγ) derived from cells of the innate and adaptive immune system.

Optimal activation of macrophages is dependent on TH1 CD4 T cells

Microbial products can directly activate monocytes and resident macrophages to secrete proinflammatory cytokines and thus initiate the immune process. However, complete activation, including the ability to kill intracellular microbes, requires the action of IFNγ – IFNγ knockout mice are extremely susceptible to infection and children with deficiencies in either the IFNγ receptor or the cytokines necessary for its production (such as IL-12, IL-18, and IL-23) have increased susceptibility to intracellular bacteria such as bacille Calmette–Guérin (BCG), *Salmonella* spp., and atypical mycobacteria.

IFNγ is so potent because it enhances several different microbicidal pathways, including both the respiratory burst and the generation of $NO^{\bullet}$.

As described above, NK cells, NK T cells, and even macrophages themselves can produce IFNγ during the innate immune response. However, the additional actions of antigen-specific T cells are necessary for optimal cell-mediated immunity.

The most important source of IFNγ during the adaptive immune response to intracellular bacteria is from TH1 $CD4^+$ T cells (Fig. 14.11).

Patients who have AIDS and a reduced CD4 T cell number and function have dramatically increased susceptibility to *M. tuberculosis* as well as other bacteria such as *Mycobacterium avium* and atypical salmonella.

As mentioned above, many bacterial components activate the TLR pattern recognition system, ensuring the preferential expression of TH1 rather than TH2 $CD4^+$ T cell responses in most cases.

TH1 T cells provide both IFNγ for macrophage activation and B cell help to produce IgG subclasses for opsonization of bacteria, rather than the eosinophilia and IgE responses typical of helminth infections.

There is mutual antagonism between the TH1 and TH2 pathways at the level of both T cell differentiation and also directly on the macrophage:

Overview of $CD4^+$ T cell-mediated immunity to bacteria and fungi

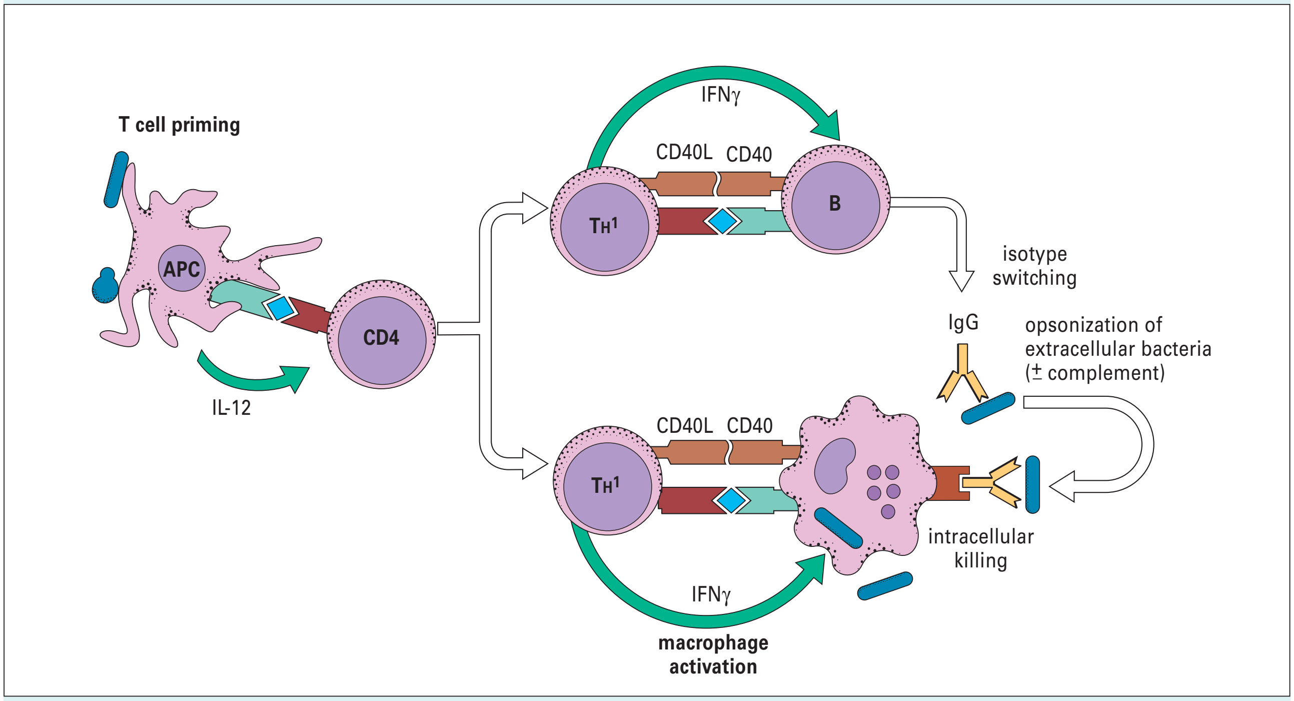

Fig. 14.11 Naive $CD4^+$ T cells are stimulated by class II MHC positive antigen-bearing dendritic cells (DCs) via the TCR, in conjunction with co-stimulatory molecules such as CD80/86 and CD28, which induce T cell activation and proliferation. Differentiation into either TH1 or TH2 effector cells is strongly influenced by the cytokine environment during this interaction – microbial pattern recognition events that favor production of IL-12 promote TH1 development, whereas low IL-12 favors TH2 responses. Although not shown here, conditions with high levels of IL-10 or TGFβ can induce regulatory T cells, rather than TH1 or TH2. Optimal T cell help for either B cells or macrophage responses involves T cell-derived cytokines and direct cell contact. TH1 cells also promote opsonizing antibody production, which complements their activation of phagocytes by IFNγ.

- IFNγ upregulates induced NO$^{\bullet}$ synthetase expression; whereas
- IL-4 and 13 promote the expression of arginase, which inhibits NO$^{\bullet}$ production, reducing the macrophage killing potential and diverting it to a profibrotic phenotype.

Other cytokines such as GM-CSF and TNF can also contribute to macrophage activation.

Macrophage activation is also promoted by direct contact with CD4 T cells via **CD40–CD40L interactions**.

Thus T cell-mediated help for macrophages and B cells shares the common themes of soluble and cell-contact mediated activation by CD4 TH1 cells.

Persistent macrophage recruitment and activation can result in granuloma formation

If intracellular pathogens are not quickly eliminated, the persistent recruitment and activation of macrophages and T cells to an infected tissue can result in the formation of **granulomas**. These are generally associated with chronic bacterial infections such as tuberculosis and syphilis, but similar (although not identical) structures are also induced in parasitic diseases such as schistosomiasis and in response to non-infectious materials such as asbestos.

In the classical example of tuberculosis, granulomas are composed of a core of infected (and uninfected) macrophages, epithelioid cells, and multinucleated giant cells (derived from the fusion of activated macrophages), and a peripheral accumulation of T cells. Neutrophils and dendritic cells can also be found in granulomas, along with extracellular matrix components such as collagen. In human tuberculosis, the center of granulomas undergoes caseating necrosis. The presence of activated macrophages and the fibrosis that ensues is believed to control bacterial growth and prevent dissemination to other organs. Generating these new immunological structures is a highly complex event involving multiple adhesion molecules, chemokines, and cytokines. Once formed, their continued existence also requires active immunological input.

AIDS and diabetes mellitus are important risk factors for loss of control of *M. tuberculosis* growth. TNF is also critical for granuloma maintenance – some patients given TNF-blocking antibodies to alleviate the symptoms of rheumatoid arthritis rapidly reactivate tuberculosis that had otherwise been controlled for many years.

SUCCESSFUL PATHOGENS HAVE EVOLVED MECHANISMS TO AVOID PHAGOCYTE-MEDIATED KILLING

Because most organisms are ultimately killed by phagocytes, it is not surprising that successful pathogens have evolved an array of mechanisms to counteract this risk (Fig. 14.12).

Intracellular pathogens may 'hide' in cells

Some organisms may thrive inside metabolically damaged host phagocytes, or escape killing by moving out of phagosomes into the cytoplasm.

Evasion mechanisms of bacteria (and some fungi)

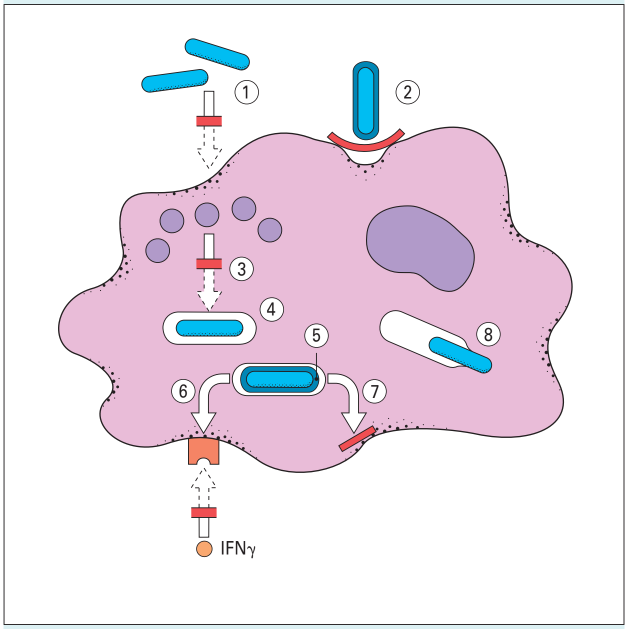

Fig. 14.12 Evasion mechanisms of bacteria (and some fungi), particularly those that are successful intracellular parasites, have evolved the ability to evade different aspects of phagocyte-mediated killing. (**1**) Some can secrete repellents or toxins that inhibit chemotaxis. (**2**) Others have capsules or outer coats that inhibit attachment by the phagocyte (e.g. *Streptococcus pneumoniae* or the yeast *C. neoformans*. (**3**) Others permit uptake, but release factors that block subsequent triggering of killing mechanisms. Once ingested, some, such as *M. tuberculosis*, inhibit lysosome fusion with the phagosome. They also inhibit the proton pump that acidifies the phagosome, so the pH does not fall. (**4**) They may also secrete catalase (e.g. staphylococci), which breaks down hydrogen peroxide. (**5**) Organisms such as *M. leprae* have highly resistant outer coats. *M. leprae* surrounds itself with a phenolic glycolipid, which scavenges free radicals. (**6**) Mycobacteria also release a lipoarabinomannan, which blocks the ability of macrophages to respond to the activating effects of IFNγ. (**7**) Cells infected with *Salmonella enterica, M. tuberculosis*, or *Chlamydia trachomatis* have impaired antigen-presenting function. (**8**) Several organisms (e.g. *Listeria* and *Shigella* spp.) can escape from the phagosome to multiply in the cytoplasm. Finally, the organism may kill the phagocyte via either necrosis (e.g. staphylococci) or induction of apoptosis (e.g. *Yersinia* spp.).

Listeria monocytogenes, *Shigella* spp., and *Burkholderia pseudomallei* achieve this by releasing enzymes that lyze the phagosome membrane. These organisms also illustrate the point that bacteria are not just inert particles, but have evolved strategies for taking control of functions of the host cell.

Other organisms, such as *M. leprae* and salmonellae, cause themselves to be taken up by cells that are not normally considered phagocytic and have little antibacterial potential such as Schwann cells, hepatocytes, and epithelial cells.

Before they can be taken up by activated phagocytes or exposed to other killing mechanisms, the organisms may need to be released from such cells.

INFECTED CELLS CAN BE KILLED BY CTLs

CD8$^+$ cytotoxic T lymphocytes (CTLs) can release intracellular organisms by killing the infected cell.

Even macrophages may be targets of CTLs because when infected with intracellular bacteria they can become refractory to activation and less able to kill the organism themselves.

Mice become strikingly susceptible to *M. tuberculosis* if class I MHC genes are knocked out so that antigen-specific CD8$^+$ T cells do not develop.

This is consistent with an essential role for CTLs in resistance to intracellular bacteria, and inducing these responses is now a primary goal of new vaccines against bacteria such as *M. tuberculosis* as well as other pathogens.

Dendritic cells appear to be particularly important in the generation of strong CD8 T cell responses to bacteria such as *L. monocytogenes* and *Salmonella* spp.

Although antigen processing and presentation via the class I MHC pathway (see Chapter 5) is most efficient for microbial antigens derived from the cytosol, nevertheless, CTLs are also clearly induced by bacteria that never escape the phagosome such as *M. tuberculosis*, salmonellae, and chlamydiae. This occurs either by **cross-presentation** of antigens within the same cell or by **cross-priming** where antigens are released from infected cells undergoing apoptosis and then transferred to nearby DCs for efficient presentation via the MHC I pathway. In some cases, lysis of infected host cells by CTLs can result in killing of the organism inside. This can be due to the action of **granulysin** – an antibacterial peptide stored in the cytotoxic granules and released during the cytotoxic process.

CTLs can also secrete IFNγ when they recognize infected targets, providing an additional pathway of macrophage activation and protective immunity (Fig. 14.13).

Other T cell populations can contribute to antibacterial immunity

In addition to the classical MHC class I- and MHC class II-mediated recognition of bacterial proteins by αβ CD4 and CD8 T cells, other 'non-conventional' T cell populations allow the host to respond to other microbial chemistries.

Pathways of CD8 T cell activation and function

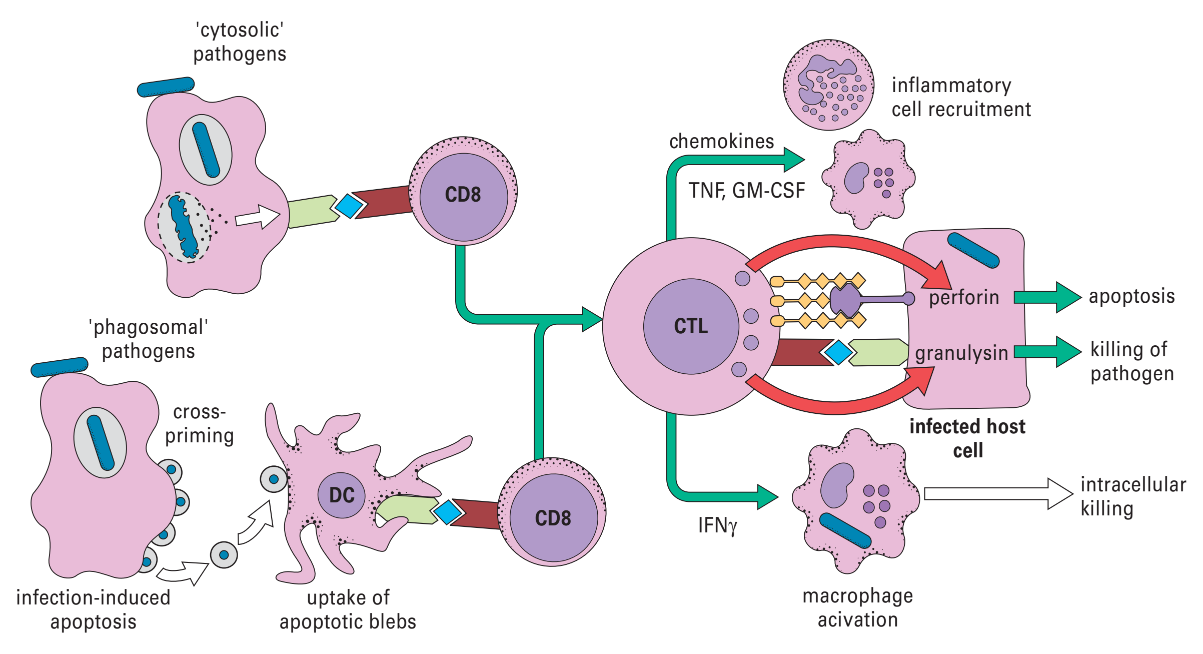

Fig. 14.13 Naive CD8 T cells are activated by peptides presented via MHC class I molecules, primarily derived from microorganisms that reside in the cytoplasm, such as viruses and some intracellular bacteria that escape the phagosome such as *Listeria* spp. Other pathogens that do not escape the phagosome (such as *M. tuberculosis*) can still induce CD8 T cell responses via cross-priming, in which infected and apoptotic host cells release antigenic fragments that are taken up by dendritic cells (DCs). Effector CD8 T cells (CTLs) provide protection by releasing proinflammatory and macrophage-activating cytokines and killing infected host cells via perforin release and Fas (see Figs 10.9 and 10.12). In some cases, the release of granulysin from the CTL can also result in killing of the pathogen.

T cells bearing γδ (rather than αβ) receptors (see Chapter 5) proliferate in response to bacterial infection.

Q. Where in the body are γδ T cells located?

A. They preferentially home to epithelial surfaces (see Chapter 2).

Some γδ T cells recognize small phospholigands derived from *M. tuberculosis* and possibly other bacteria, whereas others are triggered in an antigen-independent manner by the presence of pathogen-activated dendritic cells expressing high levels of co-stimulatory molecules and IL-12.

There are also αβ T cells (either CD4 or CD8) that recognize not proteins, but microbial glycolipids such as the **lipoarabinomannan** from *M. tuberculosis* or the capsular LPS of *H. influenzae*. These are presented via CD1a/b/c molecules, which are non-polymorphic homologs of MHC class I molecules that specialize in the binding of lipophilic antigens.

Such γδ and CD1-restricted αβ T cells have cytotoxic activity and can also secrete IFNγ suggesting a potential role in host defense. In animal models of infection these non-conventional T cells can be protective or immunoregulatory, but their relative importance in human immunity is not resolved.

Some tissue cells can express antimicrobial mechanisms

Tissue cells that are not components of the immune system can also harbor bacteria such as *M. leprae*, invasive *Shigella* and *Salmonella* spp., and *Rickettsia* and *Chlamydia* spp. As mentioned earlier, these infected cells may be sacrificed by CTLs. On the other hand, intracellular organisms such as *Rickettsia* and *Chlamydia* spp. may be starved by reduction in the availability of L-tryptophan, due to increased expression of indoleamine 2,3-dioxygenase in response to inflammatory cytokines such as IFNγ.

The secretion of antimicrobial peptides such as defensins by epithelial cells also provides an example of the protective effects of cells not strictly considered as part of the immune system.

THE RESPONSE TO BACTERIA CAN RESULT IN IMMUNOLOGICAL TISSUE DAMAGE

The events described so far are generally beneficial to the host and critical for resistance against pathogenic bacteria. However, all immune responses designed to kill invading pathogens have the potential for causing collateral damage to the host.

Excessive cytokine release can lead to endotoxin shock

If cytokine release is sudden and massive, several acute tissue-damaging syndromes can result and are potentially fatal.

One of the most severe examples of this is **endotoxin (septicemic) shock**, when there is massive production of cytokines, usually caused by bacterial products released during septicemic episodes. Endotoxin (LPS) from Gram-negative bacteria is usually responsible, though Gram-positive septicemia can cause a similar syndrome. There can be life-threatening fever, circulatory collapse, diffuse intravascular coagulation, and hemorrhagic necrosis, leading eventually to multiple organ failure (Fig. 14.14).

Paradoxically, individuals who recover from the initial life-threatening phase often overcompensate and switch from a hyper- to a hyporesponsive phase, in which excessive production of endogenous immune regulators such as IL-10 and TGFβ (and possibly other mechanisms) results in immune paralysis, making them susceptible to secondary infection.

The Schwartzman reaction is a form of cytokine-dependent tissue damage

Schwartzman observed that if Gram-negative organisms were injected into the skin of rabbits, followed by a second dose given intravenously 24 hours later, hemorrhagic necrosis occurred at the prepared skin site. This is known as the **Schwartzman reaction** (Fig. 14.15).

Many other organisms are now known to 'prepare' the skin in the same way, including streptococci, mycobacteria, *Haemophilus* spp., corynebacteria, and vaccinia virus.

Schwartzman also noted that two intravenous injections 24 hours apart caused a systemic reaction, commonly involving circulatory collapse and bilateral necrosis of the renal cortex. Sanarelli had made similar observations and this is now known as the systemic Schwartzman, or Sanarelli–Schwartzman, reaction.

These reactions can also be accompanied by necrosis in the pancreas, pituitary, adrenals, and gut. There is marked diffuse intravascular coagulation and thrombosis.

Endotoxin (LPS) is the active component of the intravenous 'triggering' injection.

Early work implicated endothelial changes, fibrin deposition, neutrophil accumulation and degranulation, and platelets as mediating the damage. This is correct, but it is now clear that tissues are primed by the induction of IFNγ (involving IL-12) derived from either NK cells or NK T cells, whereas TNF is critical in the effector phase of systemic tissue damage.

These phenomena may contribute to the characteristic hemorrhagic rash seen in children with meningococcal meningitis and the systemic effects observed in Gram-negative septic shock.

The Koch phenomenon is necrosis in T cell-mediated mycobacterial lesions and skin test sites

The **Koch phenomenon** is a necrotic response to antigens of *M. tuberculosis*, originally demonstrated by Robert Koch in tuberculous guinea pigs (Fig. 14.16). It may be related to the necrosis that also occurs in the lesions in tuberculosis. It is at least partly due to the release of cytokines into a T cell-mediated inflammatory site (**delayed hypersensitivity site**). Such sites can be extremely sensitive to the tissue-damaging effects of cytokines, as seen in the Schwartzman reaction, particularly when there is mixed TH1 and TH2 activity.

Examples to illustrate the relationship between the nature of an organism, the disease, and immunopathology

Endotoxin shock

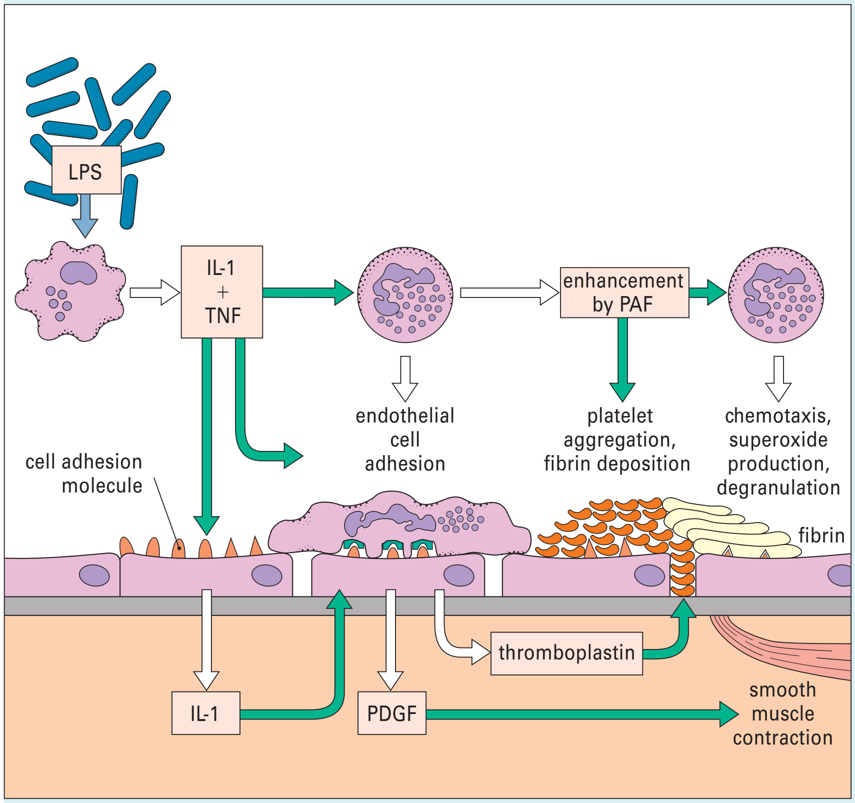
LPS
IL-1 + TNF
enhancement by PAF
endothelial cell adhesion
platelet aggregation, fibrin deposition
chemotaxis, superoxide production, degranulation
cell adhesion molecule
fibrin
thromboplastin
IL-1
PDGF
smooth muscle contraction

Fig. 14.14 Excessive release of cytokines, often triggered by the endotoxin (LPS) of Gram-negative bacteria, can lead to diffuse intravascular coagulation with consequent defective clotting, changes in vascular permeability, loss of fluid into the tissues, a fall in blood pressure, circulatory collapse, and hemorrhagic necrosis, particularly in the gut. This figure illustrates some important parts of this pathway at the cellular level. The cytokines TNF and IL-1 cause endothelial cells to express cell adhesion molecules and tissue thromboplastin. These promote adhesion of circulating cells and deposition of fibrin, respectively. Platelet activating factor (PAF) enhances these effects. In experimental models, shock can be blocked by neutralizing antibodies to TNF, and greatly diminished by antibodies to tissue thromboplastin, or by inhibitors of PAF or of nitric oxide production, but these have not been successful clinically. Gram-positive bacteria can induce shock, for example by massive release of cytokines mediated by superantigens. (PDGF, platelet-derived growth factor, produced by both platelets and endothelium)

The Schwartzman reaction

induction in rabbit		human clinical equivalent
endotoxin injected into skin	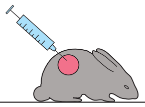	a minor septicemic episode results in dissemination of meningococci to the skin
24 h later endotoxin given intravenously		24 h later a further larger septicemic episode results in systemic release of cytokines and activation of leukocytes
hemorrhagic necrosis in the 'prepared' skin site	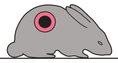	this results in necrosis in the sites where bacteria lodged after the first episode

Fig. 14.15 In the Schwartzman reaction there is cytokine-mediated tissue damage in a site of previous inflammation. This phenomenon is related to several clinical situations in humans. The first injection into the skin prepares the site by inducing inflammation and upregulating cytokine receptors, which are now the target for systemic cytokines released by the later intravenous injection of endotoxin (LPS).

caused, and the mechanism of immune response that leads to protection, are given in Fig. 14.17.

Some individuals suffer from excessive immune responses

Why some individuals suffer from excessive immune responses and others control infection with little or no damage is still unclear. The host constantly attempts to regulate immune responses to avoid the scenarios described above. For instance, immunoregulatory cytokines such as IL-10 and transforming growth factor-β (TGFβ) are produced from:

- macrophages during episodes of inflammation; and
- regulatory T cells during adaptive immunity (see Chapter 11).

The influence of both 'natural' and 'inducible' regulatory T cell populations has been reported in infections with bacteria such as *L. monocytogenes*, *Helicobacter* spp.,

The Koch phenomenon

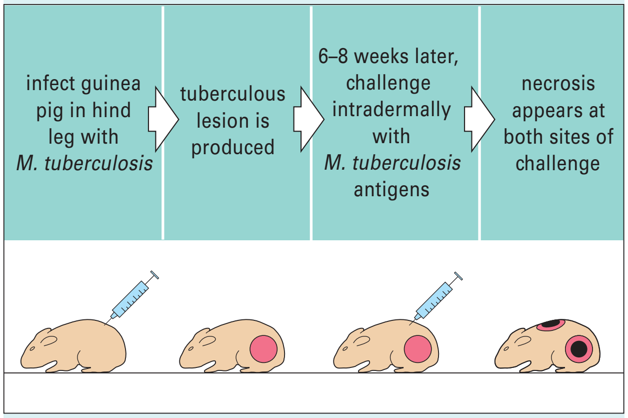
infect guinea pig in hind leg with *M. tuberculosis*
tuberculous lesion is produced
6–8 weeks later, challenge intradermally with *M. tuberculosis* antigens
necrosis appears at both sites of challenge

Fig. 14.16 Robert Koch observed that injection of *M. tuberculosis* or soluble antigens from *M. tuberculosis* into the skin of tuberculous guinea pigs resulted in a necrotizing reaction both at the challenge site and in the original tuberculous lesion. This is at least partly due to the fact that the delayed hypersensitivity reaction to mycobacterial antigens can, like the LPS-injected site in the Schwartzman reaction, be very sensitive to the toxicity of cytokines. These may be released locally in the skin test site. A similar reaction is seen in humans who have or have had tuberculosis. Responses to the same antigen in individuals who are skin-test positive as a result of BCG (bacille Calmette–Guérin) vaccination do not usually show necrosis.

Bordetella pertussis, and *M. tuberculosis*. Their presence may be beneficial to the host in preventing unwanted tissue damage, but can also be subverted by pathogens as a means of immune evasion. For example, filamentous hemagglutinin and adenylate cyclase toxin, two known virulence factors of *B. pertussis*, promote the induction of IL-10-secreting regulatory T cells in mice, dampening protective TH1 immunity and promoting persistence in the host.

Excessive immune responses can occur during treatment of severe bacterial infections

Paradoxically, excessive immune responses can also occur during treatment of humans with severe bacterial infections.

The simultaneous destruction of large numbers of organisms by antibiotics can cause increased cytokine production and systemic pathology following treatment of relapsing fever, Lyme disease, and tuberculosis.

Also, when patients who have AIDS and underlying infections such as tuberculosis and cryptococcal meningitis are treated with effective antiretroviral drugs (highly active antiretroviral therapy – HAART), the recovering T cells induce an ‘**immune reconstitution inflammatory syndrome**’ with severe tissue damage.

THE TOXICITY OF SUPERANTIGENS RESULTS FROM MASSIVE CYTOKINE RELEASE

Certain bacterial components called **superantigens** bind directly to the variable regions of β chains (Vβ) of antigen receptors on subsets of T cells, and cross-link them to the MHC molecules of APCs, usually outside the normal antigen-binding groove.

Q. What effect do superantigens have on T cells?

A. All T cells bearing the relevant Vβ gene product are activated without the processing and presentation of the antigen as peptides in the cleft of the MHC molecule that is normally required for T cell activation (see Fig. 7.22).

Immunity in some important bacterial infections

infection	pathogenesis	major defense mechanisms
Corynebacterium diphtheriae	non-invasive pharyngitis – toxin	neutralizing antibody
Vibrio cholerae	non-invasive enteritis – toxin	neutralizing and adhesion-blocking antibodies
Neisseria meningitidis (Gram-negative)	nasopharynx →bacteremia →meningitis →endotoxemia	killed by antibody and lytic complement; opsonized and phagocytosed
Staphylococcus aureus (Gram-positive)	locally invasive and toxic in skin, etc.	osponized by antibody and complement; killed by phagocytes
Mycobacterium tuberculosis	invasive, evokes immunopathology	macrophage activation by cytokines from T cells, CTLs
Mycobacterium leprae	invasive, space-occupying and/or immunopathology	

Fig. 14.17 This table provides examples of how a knowledge of the organism, and the mechanism of disease, can lead to a prediction of the relevant protective mechanism.

Between them staphylococci and streptococci have some 21 different superantigens and these molecules can also be found in other bacteria such as mycoplasmas. The full biological significance of this bacterial adaptation is not yet clear – it could be to the organism's advantage to exhaust or deplete T cells that would otherwise be protective.

One certain effect is the toxicity of the massive release of cytokines (including IL-2, TNFα, and TNFβ, together with IL-1β from activated macrophages) due to the simultaneous stimulation of up to 20% of the entire T cell pool.

The staphylococcal toxins responsible for the **toxic shock syndrome** (toxic shock syndrome toxin-1 [TSST-1], etc.) operate in this way, though not all shock syndromes caused by staphylococci are the result of T cell activation.

Recent evidence suggests that streptococcal M protein, a known virulence factor of *S. pyogenes*, forms a complex with fibrinogen, which then binds to β-integrins on neutrophils, causing the release of inflammatory mediators, which also result in massive vascular leakage and shock.

Heat-shock proteins are prominent targets of immune responses

Heat-shock proteins (HSPs or stress proteins) are found in all eukaryotic and prokaryotic cells, where they have essential roles in the assembly, folding, and transport of other molecules.

Mammalian cells exposed to abnormally elevated temperatures (or to other stresses including infection) express higher levels of these proteins, which reflects their role in the stabilization of protein structure.

Bacteria also increase HSP expression, particularly as they adapt to the intracellular environment of the host.

Immune responses are generated against both mammalian and pathogen-derived HSPs during infection.

Mammalian (and probably also bacterial) HSPs are able to activate innate immunity via multiple cell surface receptors including the α-globulin receptor CD91, CD40, CD14, and TLRs 2 and 4.

HSPs can enhance the presentation of antigens via the class I MHC/CD8 pathway (see Chapter 5) and are themselves immunogenic for both CD4 and CD8 T cells.

Because the amino acid sequences of HSPs are very highly conserved between humans and bacteria there is still considerable speculation about their role in the initiation of autoimmunity, though it seems their contribution is more complicated than simply antigenic mimicry.

The 'hygiene hypothesis'

Several groups of diseases, all characterized by defects in the regulation of the immune system, are becoming more common, particularly in developed countries. These diseases include:

- allergies;
- inflammatory bowel diseases (Crohn's disease and ulcerative colitis); and
- autoimmune conditions such as multiple sclerosis.

This may have many causes but the 'hygiene hypothesis' suggests that increasing immunological dysregulation correlates with decreasing exposure to environmental microorganisms. Decreased exposure could be due to hygiene, vaccines, and antibiotic use. If proved correct, the solution would clearly not be the abandonment of the most important achievements of medicine (hygiene, vaccines, antibiotics), but rather the identification of the environmental factors that are lacking from the modern lifestyle so that they can be replaced as vaccines or probiotics.

However, recent data suggest that the correlation does not always hold true – some viral infections (such as influenza and respiratory syncytial virus) seem to promote rather than decrease allergy and asthma in animal models.

Whatever their cause, prevention or treatment of these diseases will need to address:

- the balance of TH1/TH2 responses; and
- manipulation of the regulatory T cell circuits that normally control allergy and autoimmunity.

FUNGI CAN CAUSE LIFE-THREATENING INFECTIONS

Fungi are eukaryotes with a rigid cell wall enriched in complex polysaccharides such as chitin, glucans, and mannan.

Among the 70 000 or so species of fungi, only a small number are pathogenic for humans. However, because there are no approved vaccines and antifungal drugs often have severe side effects, fungi can cause serious and sometimes life-threatening infections.

Q. Why has it been more difficult to identify antifungal antibiotics than antibacterial antibiotics?

A. Fungi are eukaryotic organisms and therefore have similar protein synthesis machinery and mechanisms to organize and replicate the genome as mammalian cells.

Fungi can exist as:

- single cells (yeasts) small enough to be ingested by host phagocytes; or
- long slender, branching hyphae, which may require extracellular killing processes.

Some pathogenic fungi are dimorphic, in that they switch from a hyphal form in the environment to a yeast form as they adapt to life in the host. Both phases possess important virulence determinants and pose different problems to the immune system.

There are four categories of fungal infection

Although some fungi can cause disease in otherwise healthy individuals, severe fungal infections are a growing problem because of the markedly increased numbers of immunologically compromised hosts. Fungal infections are therefore regularly seen in:

- patients with untreated AIDS;
- patients with cancer and undergoing chemotherapy;
- patients with transplants on immunosuppressive agents; and
- some patients taking long-term corticosteroids.

These clinical findings point to the key roles of neutrophils and macrophage activating TH1 cell responses in antifungal immunity.

Human fungal infections fall into the following four major categories:

- **superficial mycoses** caused by fungi known as dermatophytes, usually restricted to the non-living keratinized components of skin, hair and nails, and including infection by *Trichophyton* and *Microsporum* spp. (which cause ringworm and athlete's foot) and *Malassezia* spp. (which causes pityriasis);
- **subcutaneous mycoses** in which saprophytic fungi cause chronic nodules or ulcers in subcutaneous tissues following trauma (e.g. chromomycosis, sporotrichosis, and mycetoma);
- **systemic mycoses** caused by soil saprophytes, which are inhaled from the environment and produce subclinical or acute lung infections that can disseminate to almost any tissue in the immunocompromised host – *Histoplasma*, *Blastomyces*, *Coccidioides*, and *Paracoccidioides* spp. can all cause primary disease in otherwise immunocompetent individuals, whereas *Aspergillus* spp., *Pneumocystis carinii*, and *C. neoformans* act more as opportunists;
- **candidiasis** caused by *Candida albicans*, a ubiquitous commensal and the most common opportunistic fungal pathogen – disturbance of normal physiology by immunosuppressive drugs, of normal flora by antibiotics, or T cell function (as in severe combined immune deficiency, thymic aplasia, and AIDS), results in superficial infections of the skin and mucous membranes, and systemic disease can occur in intravenous drug users and patients with lymphoma or leukemia.

Innate immune responses to fungi include defensins and phagocytes

The basic protective features of the skin and normal commensal flora described against bacterial infections above are also important in resistance to fungi.

Defensins have antifungal as well as antibacterial properties, and collectins such as MBL and the surfactant proteins A and D can bind, aggregate, and opsonize fungi for phagocytosis.

Phagocytes, particularly neutrophils (Fig. 14.18) and macrophages, are essential for killing fungi, either by:

- degranulation and release of toxic materials onto large indigestible hyphae; or
- ingestion of yeast or conidia.

Evidence for neutrophil-mediated immunity to mucormycosis

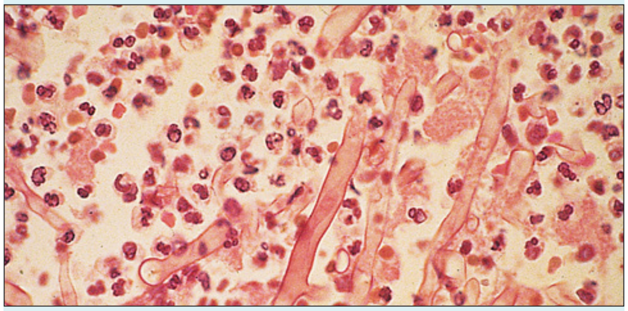

Fig. 14.18 This is a section from the lung of a patient suffering from mucormycosis – an opportunistic infection in an immunosuppressed subject. The inflammatory reaction consists almost entirely of neutrophil polymorphs around the fungal hyphae. The disease is particularly associated with neutropenia (lack of neutrophils). Silver stain. × 400. (Courtesy of Professor RJ Hay)

The oxidative burst plays a crucial role in some antifungal responses, as seen in the susceptibility to severe aspergillosis by patients with CGD who have defects in the NADPH oxidase system. However, phagocytes from such patients with defective oxygen reduction pathways nevertheless kill other yeast and hyphae with near normal efficiency, so demonstrating the role of other killing mechanisms (Fig. 14.19). For instance, $NO^{\bullet}$ and its derivatives are important for resistance to *C. neoformans*.

These responses rely on the recognition of PAMPs in the fungal cell wall by either soluble or cell-bound pattern recognition molecules. The TLR family again plays an important role in this process, along with the mannose receptor and complement receptors:

- TLR2 (which can cooperate with the β-glucan receptor dectin-1) recognizes fungal phospholipomannans, *C. albicans* yeasts, and *A. fumigatus* hyphae and conidia;

Monocyte/macrophage killing of fungi

organism	source of monocytes/macrophages		
	normal	CGD	MPO deficiency
Candida albicans	killed	sometimes killed	sometimes killed
Candida parapsilosis	killed	not killed	unknown
Cryptococcus neoformans	killed	unknown	killed
Aspergillus fumigatus conidia	killed	sometimes killed	killed
Aspergillus fumigatus hyphae	killed	killed	killed

Fig. 14.19 Many fungi are killed by monocytes or macrophages. Individuals with chronic granulomatous disease (CGD) are highly susceptible to *Aspergillus* spp. infections whereas myeloperoxidase (MPO) deficiency does not usually lead to opportunistic infection, suggesting that non-oxygen-dependent mechanisms are also important in host defense.

- TLR4/CD14 recognizes *C. albicans*, *Aspergillus fumigatus*, and the glucuronoxylomannan capsule of *C. neoformans*.

Not all of these events are to the host's advantage, for example recognition of *Candida albicans* mannan via TLR4 induces proinflammatory chemokine responses, whereas ligation of candidal phospholipomannan and glucans with TLR2/dectin-1 generates a strong IL-10 response, which may inhibit immune function.

T cell-mediated immunity is critical for resistance to fungi

Most fungi are highly immunogenic and induce strong antibody and T cell-mediated immune responses, which can be detected by serology and delayed-type (type IV) hypersensitivity skin reactions (see Chapter 26).

Considerable evidence points to the dominant protective role of TH1 T cells and macrophage activation, rather than antibody-mediated responses.

Patients with T cell deficiencies, rather than defects in antibody production, are more at risk of disseminated fungal disease, and antibody titers, though useful as an epidemiological tool to determine exposure, do not necessarily correlate with prognosis. Nevertheless, fungi can elicit both protective and non-protective antibodies and the protection afforded by some experimental vaccines can be adoptively transferred by immune sera.

Resistance to most pathogenic fungi (including dermatophytes and most systemic mycoses including *C. neoformans*, *Histoplasma capsulatum*, etc., but not *Aspergillus* spp.) is clearly dependent upon T cell-mediated immunity, particularly $CD4^+$ TH1 cells secreting IFNγ (Fig. 14.20). As in the case of bacteria, dendritic cells are necessary for this response and produce IL-12 after engulfing fungi.

The clinical relevance of TH1 versus TH2 responses is also clear for some human mycoses, for example:

- individuals with mild paracoccidioidomycosis have TH1-biased immune responses; whereas
- individuals with severe, disseminated infection have high levels of TH2 cytokines such as IL-4 and IL-10, and eosinophilia.

Children with the primary immunodeficiency hyper IgE syndrome with defects in the production of IFNγ also have increased susceptibility to fungal infections.

An increased level of IL-10 (with concomitant reductions in IFNγ) is also a marker of impaired immunity to systemic mycoses, *C. albicans*, and in neutropenia-associated aspergillosis.

Evidence for T cell immunity in chromomycosis

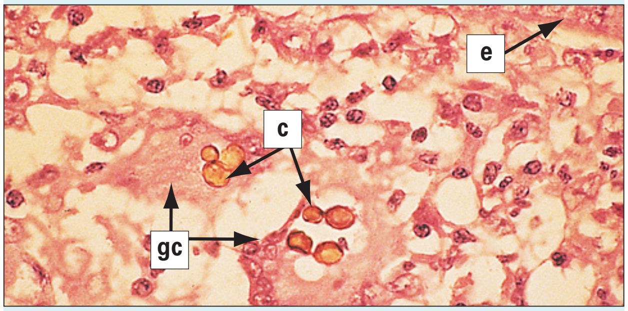

Fig. 14.20 The pigmented fungal cells of chromomycosis (a subcutaneous mycosis) (c) are visible inside giant cells (gc) in the dermis of a patient. The area is surrounded by a predominantly mononuclear cell infiltrate. The basal layer of epidermis (e) is visible at the top of the frame. H & E stain. × 400. (Courtesy of Professor RJ Hay)

Fungi possess many evasion strategies to promote their survival

Evasion strategies used by fungi to promote their survival include the following:

- *Cryptococcus neoformans* produces a polysaccharide capsule, which inhibits phagocytosis (similar in principle to that of encapsulated bacteria), though this can be overcome by the opsonic effects of complement and antibodies;
- *Histoplasma capsulatum* is an obligate intracellular pathogen that evades macrophage killing by entering the cell via CR3 and then altering the normal pathways of phagosome maturation, in parallel to the strategies of intracellular bacteria such as *M. tuberculosis*;
- dermatophytes suppress host T cell responses to delay cell-mediated destruction.

Immune responses to fungi are therefore as complex and interesting as those against bacteria and for many infections (such as the subcutaneous mycoses) these responses remain poorly understood.

New immunological approaches are being developed to prevent and treat fungal infections

Unlike many antibiotics, which are directly microbicidal, antifungal drugs need significant assistance from the immune system to be most effective.

Reducing the underlying immunosuppression that leads to susceptibility to fungi is an important goal and generic immunotherapies such as cytokine administration (using IFNγ in patients with CGD and granulocyte colony stimulating factor [G-CSF] therapy to reduce neutropenia in patients with cancer) have had some success. There is also considerable interest in dendritic cell-based vaccine strategies to promote TH1-mediated immunity.

FURTHER READING

Bacteria

Collins HL, Kaufmann SH. Prospects for better tuberculosis vaccines. Lancet Infect Dis 2001;1:21–28.

De Gregorio E, Rappuoli R. Inside sensors detecting outside pathogens. Nat Immunol 2004;5:1099–1100.

Harty JT, Tvinnereim AR, White DW. $CD8^+$ T cell effector mechanisms in resistance to infection. Annu Rev Immunol 2000;18:275–308.

Kaufmann SH. How can immunology contribute to the control of tuberculosis? Nat Rev Immunol 2001;1:20–30.

Kaufmann SH, Schaible UE. Antigen presentation and recognition in bacterial infections. Curr Opin Immunol 2005;17:79–87.

MacLennan C, Fieschi C, Lammas DA, et al. Interleukin (IL)-12 and IL-23 are key cytokines for immunity against *Salmonella* in humans. J Infect Dis 2004;190:1755–1757.

Merrell DS, Falkow S. Frontal and stealth attack strategies in microbial pathogenesis. Nature 2004;430:250–256.

Monack DM, Mueller A, Falkow S. Persistent bacterial infections: the interface of the pathogen and the host immune system. Nat Rev Microbiol 2004;2:747–765.

Philpott DJ, Girardin SE. The role of Toll-like receptors and Nod proteins in bacterial infection. Mol Immunol 2004;41:1099–1108.

Reis e Sousa C. Toll-like receptors and dendritic cells: for whom the bug tolls. Semin Immunol 2004;16:27–34.

Rouse BT, Suvas S. Regulatory cells and infectious agents: detentes cordiale and contraire. J Immunol 2004;173:2211–2215.

Schaible UE, Kaufmann SH. Iron and microbial infection. Nat Rev Microbiol 2004;2:946–953.

Stewart GR, Young DB. Heat-shock proteins and the host–pathogen interaction during bacterial infection. Curr Opin Immunol 2004;16:506–510.

Umetsu DT. Revising the immunological theories of asthma and allergy. Lancet 2005;365:98–100.

van de Vosse E, Hoeve MA, Ottenhoff TH. Human genetics of intracellular infectious diseases: molecular and cellular immunity against mycobacteria and salmonellae. Lancet Infect Dis 2004;4:739–749.

Voyich JM, Musser JM, DeLeo FR. *Streptococcus pyogenes* and human neutrophils: a paradigm for evasion of innate host defense by bacterial pathogens. Microbes Infect 2004;6:1117–1123.

Weber JR, Moreillon P, Tuomanen EI. Innate sensors for Gram-positive bacteria. Curr Opin Immunol 2003;15:408–415.

Wills-Karp M, Santeliz J, Karp CL. The germless theory of allergic disease: revisiting the hygiene hypothesis. Nat Rev Immunol 2001;1:69–75.

Fungi

Netea MG, Van der Graaf C, Van der Meer JW, Kullberg BJ. Recognition of fungal pathogens by Toll-like receptors. Eur J Clin Microbiol Infect Dis 2004;23:672–676.

Romani L. Immunity to fungal infections. Nat Rev Immunol 2004;4:1–23.

Critical thinking: Immunoendocrine interactions in the response to infection (see p. 497 for explanations)

Humans subclinically infected with tuberculosis (about one-third of the world's population) may harbor live organisms for the rest of their lives. Similarly tuberculosis can establish a latent non-progressive infection in mice. If animals with such latent infection are subjected to a period of restraint stress (placed in a tube that limits movement) each day for several days, the infection may reactivate. This also happens if cattle with latent disease are transported in trucks. Similarly tuberculosis increases in human populations in war zones, probably due to reactivation of latent disease.

1 What is the physiology of this reactivation?

When American military trainees were subjected to an extremely stressful training schedule their serum IgE levels rose and they lost their previously positive delayed hypersensitivity skin-test responses. The levels of mRNA encoding IFNγ in the peripheral blood mononuclear cells of medical students were lower during the examination period than at other times of the year.

2 Do these observations suggest changes in cytokine profile? If so, why did it happen?

CHAPTER

15

Immunity to Protozoa and Worms

SUMMARY

- **Parasites stimulate a variety of immune defense mechanisms**.
- **Parasitic infections are often chronic and affect many people.** They are generally host specific and most cause chronic infections. Many are spread by invertebrate vectors and have complicated life cycles. Their antigens are stage specific.
- **Innate immune responses are the first line of immune defense**.
- **T and B cells are pivotal in the development of immunity.** Both CD4 and CD8 T cells are needed for protection from some parasites, and cytokines, chemokines, and their receptors have important roles.
- **Effector cells such as macrophages, neutrophils, eosinophils, and platelets can kill both protozoa and worms.** They secrete cytotoxic molecules such as reactive oxygen radicals and nitric oxide (NO$^\bullet$). All are more effective when activated by cytokines. Worm infections are usually associated with an increase in eosinophil number and circulating IgE, which are characteristic of TH2 responses. TH2 cells are necessary for the elimination of intestinal worms.
- **Parasites have many different escape mechanisms.** Evasion of the host's immune response by parasites occurs in various ways. Some exploit the host response for their own development. It is becoming clear that many parasites, in particular helminths, are able to modulate the host immune response.
- **Inflammatory responses can be a consequence of eliminating parasitic infections**.
- **Parasitic infections have immunopathological consequences.** Parasitic infections are associated with pathology, which can include autoimmunity, splenomegaly, and hepatomegaly. Much immunopathology may be mediated by the adaptive immune response.
- **Vaccines against human parasites are not yet available**.

PARASITES STIMULATE A VARIETY OF IMMUNE DEFENSE MECHANISMS

Parasitic infections typically stimulate a number of immune defense mechanisms, both antibody and cell mediated, and the responses that are most effective depend upon the particular parasite and the stage of infection. Some of the more important parasitic infections of humans (Fig. 15.1) affect the host in diverse ways.

Parasitic protozoa may live:

- in the gut (e.g. amebae);
- in the blood (e.g. African trypanosomes);
- within erythrocytes (e.g. *Plasmodium* spp.);
- in macrophages (e.g. *Leishmania* spp., *Toxoplasma gondii*);
- in liver and spleen (e.g. *Leishmania* spp.); or
- in muscle (e.g. *Trypanosoma cruzi*).

Parasitic worms that infect humans include trematodes or flukes (e.g. schistosomes), cestodes (e.g. tapeworms), and nematodes or roundworms (e.g. *Trichinella spiralis*, hookworms, pinworms, *Ascaris* spp., and the filarial worms).

Tapeworms and adult hookworms inhabit the gut, adult schistosomes live in blood vessels, and some filarial worms live in the lymphatics (Fig. 15.2). It is clear that there is widespread potential for damaging pathological reactions.

Many parasitic worms pass through complicated life cycles, including migration through various parts of the host's body:

- hookworms and schistosome larvae invade their hosts directly by penetrating the skin;
- tapeworms, pinworms, and roundworms are ingested; and
- filarial worms depend upon an intermediate insect host or vector to transmit them from person to person.

Most protozoa rely upon an insect vector, apart from *Toxoplasma* and *Giardia* spp. and amebae, which are transmitted by ingestion. Thus:

- malarial parasites are spread by mosquitoes;
- trypanosomes by tsetse flies;
- *T. cruzi* by triatomine bugs; and
- *Leishmania* by sandflies.

Important parasitic infections of humans

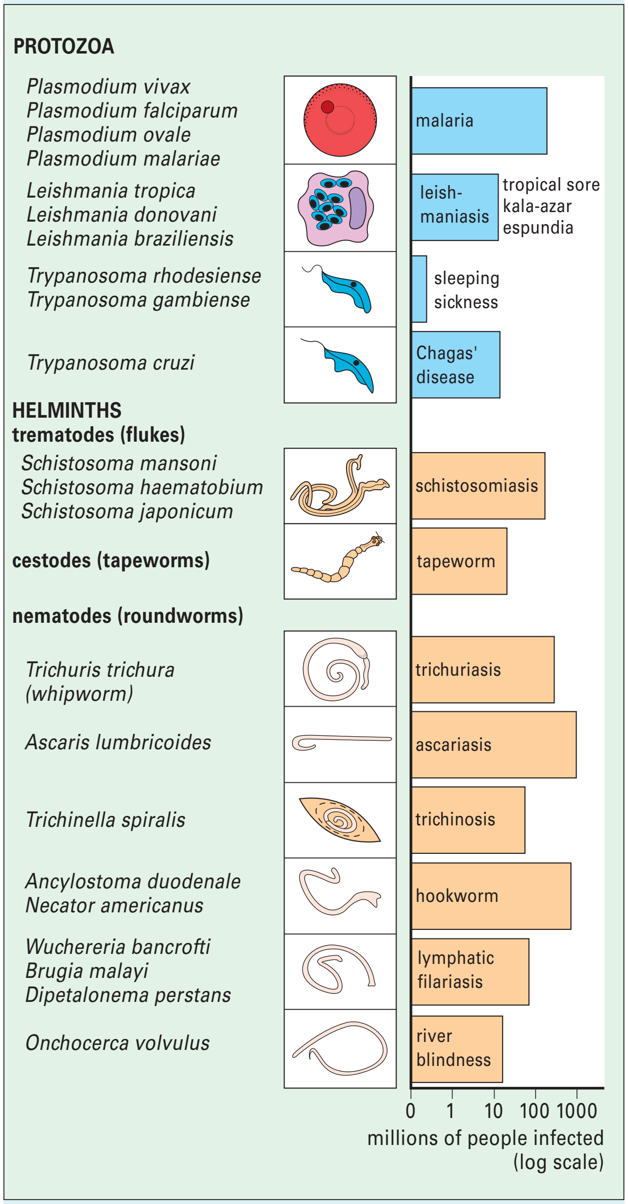

Fig. 15.1 Important parasitic infections, including data from the World Health Organization (1993).

Sites of infection of medically important parasites

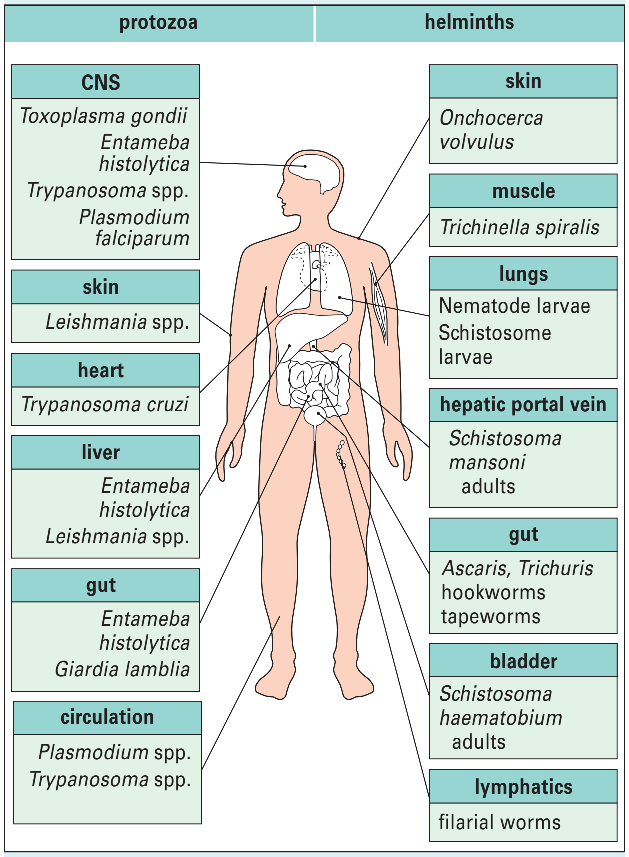
protozoa
helminths
CNS
Toxoplasma gondii
Entameba histolytica
Trypanosoma spp.
Plasmodium falciparum
skin
Leishmania spp.
heart
Trypanosoma cruzi
liver
Entameba histolytica
Leishmania spp.
gut
Entameba histolytica
Giardia lamblia
circulation
Plasmodium spp.
Trypanosoma spp.
skin
Onchocerca volvulus
muscle
Trichinella spiralis
lungs
Nematode larvae
Schistosome larvae
hepatic portal vein
Schistosoma mansoni adults
gut
Ascaris, Trichuris
hookworms
tapeworms
bladder
Schistosoma haematobium adults
lymphatics
filarial worms

Fig. 15.2 Sites of infection of medically important parasites.

PARASITIC INFECTIONS ARE OFTEN CHRONIC AND AFFECT MANY PEOPLE

Parasitic infections present a major medical problem, especially in tropical countries (see Fig. 15.1), for example:

- malaria kills 1–2 million people every year;
- intestinal worms infect one-third of the world's population – the severity of disease depends upon the worm burden, but in children even moderate intensities of infection may be associated with stunted growth and slow mental development.

Anemia and malnutrition are also associated with parasitic disease.

Over millions of years of evolution, parasites have become well adapted to their hosts and show marked host specificity. For example, the malarial parasites of birds, rodents, or humans can each multiply only in their own particular kind of host.

There are some exceptions to this general rule, for example:

- the protozoan parasite *T. gondii* is not only able to invade and multiply in all nucleated mammalian cells, but can also infect immature mammalian erythrocytes, insect cell cultures, and the nucleated erythrocytes of birds and fish;
- similarly, the tapeworm of the pig can also infect humans.

Protozoan parasites and worms are considerably larger than bacteria and viruses (Fig. 15.3), and have very different strategies for avoiding the host immune response.

Q. Apart from size, what is the basic biological difference between bacterial pathogens and parasites, and how would this affect they way they are recognized by the immune system?

A. Bacteria are prokaryotes, whereas parasites are eukaryotes. The plasma membrane and cell wall structures in bacteria are different, so they have distinct **pathogen-associated molecular patterns** (**PAMPs**, see Chapter 6). In addition, protein synthesis (initiated with formyl methionine) is different in prokaryotes.

Comparative size of various parasites

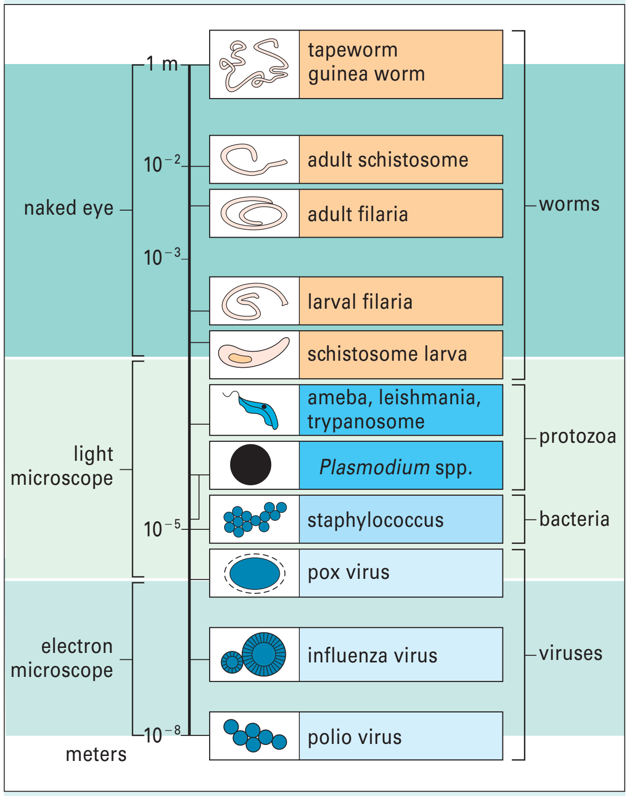

Fig. 15.3 Comparative size of various parasites.

Some species can also change their surface antigens, a process known as **antigenic variation** (see below).

Parasites that have complicated life histories may express certain antigens only at a particular stage of development, giving rise to a stage-specific response. Thus, the protein coat of the sporozoite (the infective stage of the malarial parasite transmitted by the mosquito) induces the production of antibodies that do not react with the erythrocytic stages. The different stages of the worm *T. spiralis* also display different surface antigens.

Protozoa that are small enough to live inside human cells have evolved a special mode of entry:

- the merozoite, the invasive form of the blood stage of the malarial parasite, binds to certain receptors on the surface of the erythrocyte and uses a specialized organelle, the rhoptry, to enter the cell;
- *Leishmania* spp. parasites, which inhabit macrophages, use complement receptors to encourage the cells to engulf them.

Leishmania can also gain entry to the cell by using the mannose receptor (see Fig. 9.18) on the macrophage surface.

Host resistance to parasite infection may be genetic

The resistance of individual hosts to infection varies, and may be controlled by a number of genes. These may be MHC, non-MHC, or other genes (Fig. 15.4).

One should not assume that host genetic background is the only reason determining the outcome of infection. There may be many factors involved. In most helminth infections, for example, a heavy worm burden occurs in comparatively few individuals, but may cluster in families, implying a genetic basis. On the other hand, studies have shown that human behavior can account for large variation in exposure between families.

Many parasitic infections are long-lived

It is not in the interest of a parasite to kill its host, at least not until transmission to another host has been ensured. During the course of a chronic infection the type of immune response may change and immunosuppression and immunopathological effects are common.

Host defense depends upon a number of immunological mechanisms

The development of immunity is a complex process arising from the activation of both adaptive and innate immune responses and the switching on of many different kinds of cell over a period of time. Effects are often local and many cell types secreting different mediators may be present at sites of immune rejection. Moreover, the

Human gene polymorphisms that affect the outcome of parasite infection

parasite	genetic trait
Plasmodium spp.	sickle cell hemoglobin (HbS) protects from malaria certain MHC genes common in West Africans, but rare in Caucasians, protect from malaria (e.g. HLA-B53) Duffy blood group antigen (Fy/Fy)-negative erythrocytes protect from *Plasmodium vivax*
Leishmania spp.	polymorphisms in *Nramp1* govern susceptibility to macrophage invasion
Schistosoma spp.	candidate polymorphic genes on chromosome 5q31-q33, a region that includes key cytokines IL-4 and IL-5
Ascaris spp.	candidate polymorphic genes on chromosomes 1 and 13, a region that includes the TNF family of cytokines

Fig. 15.4 Human gene polymorphisms that affect the outcome of parasite infection.

processes involved in controlling the multiplication of a parasite within an infected individual may differ from those responsible for the ultimate development of resistance to further infection.

In some helminth infections a process of ‘concomitant immunity’ occurs, whereby an initial infection is not eliminated, but becomes established, and the host then acquires resistance to invasion by new worms of the same species.

In very general terms, humoral responses are important to eliminate extracellular parasites such as those that live in blood (Fig. 15.5), body fluids, or the gut.

However, the type of response conferring most protection varies with the parasite. For example, antibody, alone or with complement, can damage some extracellular parasites, but is better when acting with an effector cell.

As emphasized above, within a single infection different effector mechanisms act against different developmental stages of parasites. Thus in malaria:

- antibody against extracellular forms blocks their capacity to invade new cells;
- cell-mediated responses prevent the development of the liver stage within hepatocytes.

Protective immunity to malaria does not correlate simply with antibody levels and can even be induced in the absence of antibody.

INNATE IMMUNE RESPONSES ARE THE FIRST LINE OF IMMUNE DEFENSE

The innate and adaptive immune responses are co-evolving to allow mammals to identify and eliminate parasites.

The innate immune system provides the first line of immune defense by detecting the immediate presence and nature of infection.

Many different cells are involved in generating innate responses including phagocytic cells and NK cells. It is also becoming clear that early recognition of parasites by antigen-presenting cells (APCs), for example dendritic cells, determines the phenotype of the adaptive response (Fig. 15.6).

Innate immune recognition relies on a growing number of receptors, termed **pattern recognition receptors (PRRs)** that have evolved to recognize pathogen-associated molecular patterns (PAMPs).

Adult schistosome worm pairs in mesenteric blood vessels

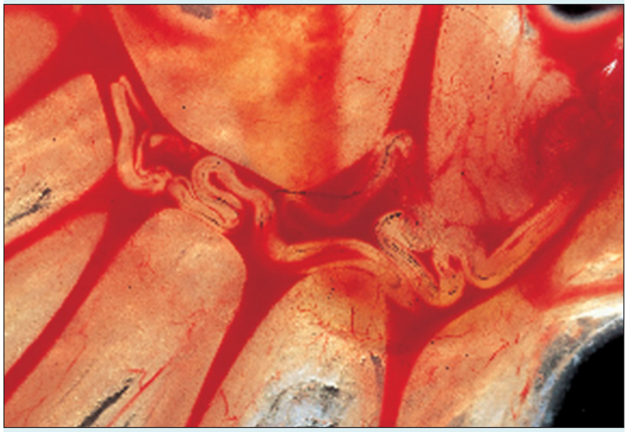

Fig. 15.5 Although very exposed to immune effectors, adult schistosomes are highly resistant and can persist for an average of 3–5 years. (Courtesy of Dr Alison Agnew)

Q. Which groups of receptors and soluble molecules recognize PAMPs?

A. Toll-like receptors (see Fig. 6.24), the mannose receptor (see Fig. 9.18), and scavenger receptors (see Fig. 9.16) allow phagocytes to directly recognize pathogens. Ficolins, collectins, and pentraxins act as soluble opsonins by binding to pathogen surfaces (see Chapter 6 and Fig. 6.23).

Development of the immune response to protozoan and helminth infection

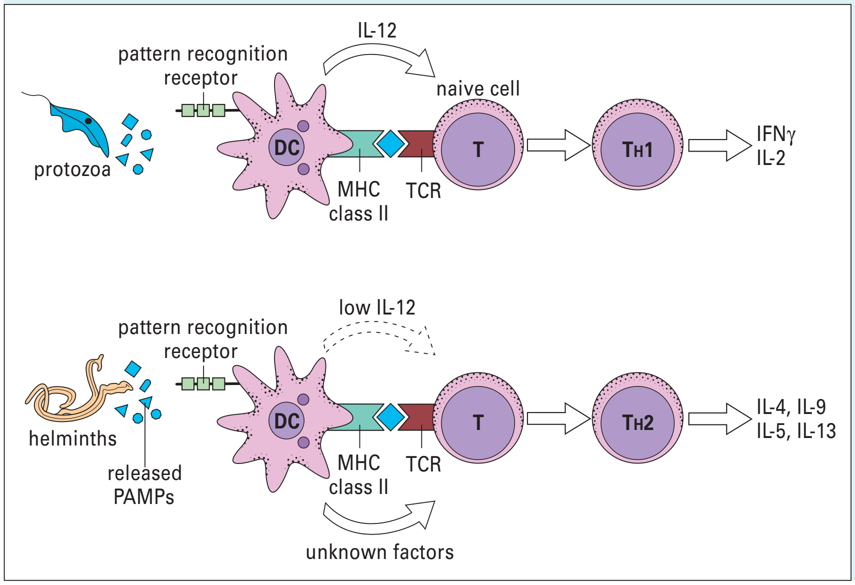

Fig. 15.6 The cytokines secreted by the different subsets of T cells are shown. Note the importance of dendritic cell (DC)-derived IL-12 in driving maturation of helper T cell subsets.

A unifying feature of these targets is their highly conserved structures, which are invariant between parasites of a given class.

Although many parasites are known to activate the immune system in a non-specific manner shortly after infection, it is only recently that attention has been given to the mechanisms involved.

While major advances are being achieved in the area of microbial recognition by PRRs, a small but growing number of studies show that parasites also possess specific molecular patterns capable of engaging PRRs. Examples of some parasite PAMPs along with their receptors are given in Fig. 15.7.

TLR domain-containing PRRs

The discovery of the TLR family (see Fig. 15.7), an evolutionarily conserved group of thirteen mammalian PRRs involved in antimicrobial immunity, has enriched our understanding of how innate and adaptive immunity are mutually dependent.

Very few studies have examined the role of TLRs in immunity to parasites.

Lyso-phosphatidylserine (lyso-PS) from *S. mansoni*, glycophosphatidylinositol (GPI) anchors, and Tc52 from *T. cruzi* are capable of signaling through TLR2. Interestingly, TLR2 triggering by these diverse parasite patterns leads to different immune outcomes:

- with *S. mansoni*, triggering leads to the development of fully mature dendritic cells capable of inducing a Treg response (see Chapter 11), characterized by elevated IL-10 levels;
- with *T. cruzi*, mature dendritic cells induce a TH1 response (see Chapter 7) with raised levels of IL-12.

This dichotomous response could, in part, be explained by the cooperation between TLR2 and other TLRs, including TLR1 and TLR6.

The formation of heterodimers comprising TLR2 and either TLR1 or TLR6 has not yet been observed, but could explain the observed differences in T cell phenotypes.

Innate immune receptors involved in parasite recognition

family	member	parasite ligand(s)
collectins (see Fig. 6.23)	MBL	mannose-rich sugars from numerous protozoans and helminths
pentraxins	CRP	phospholipids and phosphosugars
		Leishmania spp. LPG
C-type lectins (see Fig. 9.18)	macrophage mannose receptor, DC-SIGN	*Trypanosoma cruzi*, *Schistosoma* spp. (LewisX)
scavenger receptors (see Fig. 9.16)	SR-B (CD36)	*Plasmodium falciparum* (PfEMP1)
complement receptors	CR1/CR3	*Leishmania* spp. LPG
		Necator NIF
		Plasmodium PfEMP1
Toll-like receptors (see Fig. 6.24)	TLR2 (with TLR1/ TLR6)	GPI anchors from many protozoa
		lyso-PS from *Schistosoma* spp.
	TLR3	double-stranded RNA from *Schistosoma* spp.
	TLR4 (with CD14)	LPS-like molecules from the filarial endosymbiont *Wolbachia* spp.
	TLR9	protozoal DNA
		malarial pigment hemozoin

CRP, C-reactive protein; GPI, glycosylphosphatidylinositol; LPG, lipophosphoglycan; lyso-PS, lyso-phosphatidylserine; MBL, mannose-binding lectin; NIF, neutrophil inhibition factor (NIF); PfEMP1, *P. falciparum* erythrocyte membrane protein-1

Fig. 15.7 Innate immune receptors involved in parasite recognition.

TLR9 mediates innate immune activation by the malaria pigment hemozoin.

Classical human PRRs

Classical PRRs play important roles in the innate response to parasite infection (see Fig. 15.7) and include:

- **collectins** (e.g. MBL);
- **pentraxins** (e.g. CRP);
- **C-type lectins** (e.g. macrophage mannose receptor); and
- **scavenger receptors** (e.g. CD36).

For example MBL binds mannose-rich LPG from *Leishmania*, *Plasmodium*, trypanosomes, and schistosomes; and polymorphisms in the *MBL* gene are associated with increased susceptibility to severe malaria.

Complement receptors are archetypal PRRs

Complement receptors, in particular CR3, are archetypal PRRs involved in innate immune responses (see Fig. 15.7). They are truly multifunctional, being involved in phagocyte adhesion, recognition, migration, activation, and microbe elimination.

Accumulating evidence indicates that complement has pivotal roles to play in diverse biological processes, ranging from early hematopoiesis to skeletal, vascular, and reproductive development.

Why then is CR3, a linchpin of phagocyte responses, the favored portal of entry for diverse intracellular parasites, including leishmania via LPG?

- first, CR3 offers a multiplicity of binding sites, enabling opsonic or non-opsonic binding;
- second, phagocytosis by CR3 alone does not generate an oxidative burst in phagocytic cells;
- third, binding of CR3 suppresses the secretion of IL-12.

Q. What effect would suppression of IL-12 secretion have on the immune response?

A. Because IL-12 promotes the development of the TH1 response, reduced expression will tend to reduce macrophage-mediated immunity.

In isolation, CR3 is not an activating receptor. It requires cooperation from other receptors, most notably Fc receptors, for pathogen killing. Helminths have also exploited this chink in the immune armoury. Hookworm NIF has been shown to bind a domain in the α subunit of CR3, presumably to downregulate cell-mediated immunity.

T AND B CELLS ARE PIVOTAL IN THE DEVELOPMENT OF IMMUNITY

In most parasitic infections, protection can be conferred experimentally on normal animals by the transfer of spleen cells, especially T cells, from immune animals.

The T cell requirement is also demonstrable because nude (athymic) or T-deprived mice fail to clear otherwise non-lethal infections of protozoa such as *T. cruzi* or *Plasmodium yoelii*, and T cell-deprived rats fail to expel the intestinal worm *Nippostrongylus brasiliensis* (Fig. 15.8).

Counter-intuitively, many parasites require signals from immune cells to thrive – for example, schistosomes fail to develop in the absence of hepatic CD4$^+$ lymphocytes.

B cells also play key roles in regulating and controlling immunity to parasites. For example:

- B cells and antibodies are required for resistance to the parasitic gastrointestinal nematode *Trichuris muris*; and
- passive transfer of IgG can protect people from malaria.

Both CD4 and CD8 T cells are needed for protection from some parasites

The type of T cell responsible for controlling an infection varies with the parasite and the stage of infection, and depends upon the kinds of cytokine they produce. For example, CD4$^+$ and CD8$^+$ T cells protect against different phases of *Plasmodium* infection:

- CD4$^+$ T cells mediate immunity against blood-stage *P. yoelii*;
- CD8$^+$ T cells protect against the liver stage of *Plasmodium berghei.*

The action of CD8$^+$ T cells is twofold:

- they secrete IFNγ, which inhibits the multiplication of parasites within hepatocytes;
- they are able to kill infected hepatocytes, but not infected erythrocytes.

Q. By what mechanism are the hepatocytes killed, and why do the CD8$^+$ cells not recognize infected erythrocytes?

A. The CD8$^+$ T cells recognize parasite antigens presented by MHC class I molecules and induce apoptosis in the target, via Fas ligand, tumor necrosis factor (TNF), lymphotoxin, and granzymes. Erythrocytes do not express MHC molecules.

The immune response against *T. cruzi* depends not only upon CD4$^+$ and CD8$^+$ T cells, but also on NK cells and antibody production; the same is true for the immune response against *T. gondii.*

CD8$^+$ T cells confer protection in mice depleted of CD4$^+$ T cells, both through their production of IFNγ and because they are cytotoxic for infected macrophages.

NK cells, stimulated by IL-12 secreted by the macrophages, are another source of IFNγ.

Chronic infections are associated with reduced production of IFNγ.

These observations probably underlie the high incidence of toxoplasmosis in patients with AIDS, who are deficient in CD4$^+$ T cells.

CD4$^+$ T cells are critical for the expulsion of intestinal nematodes and as immunity to *T. muris* can be transferred to a SCID (severe combined immune deficiency) mouse by the transfer of CD4$^+$ T cells alone there is no evidence for a role of CD8$^+$ T cells.

The cytokines produced by CD4$^+$ T cells can be important in determining the outcome of infection. Helper T cells have been phenotypically divided into TH1 and TH2 and more recently regulatory T cell subsets based on the cytokines produced.

As TH1 and TH2 cells have contrasting and cross-regulating cytokine profiles, the roles of TH1 or TH2 cells

Parasitic infections in T cell-deprived mice

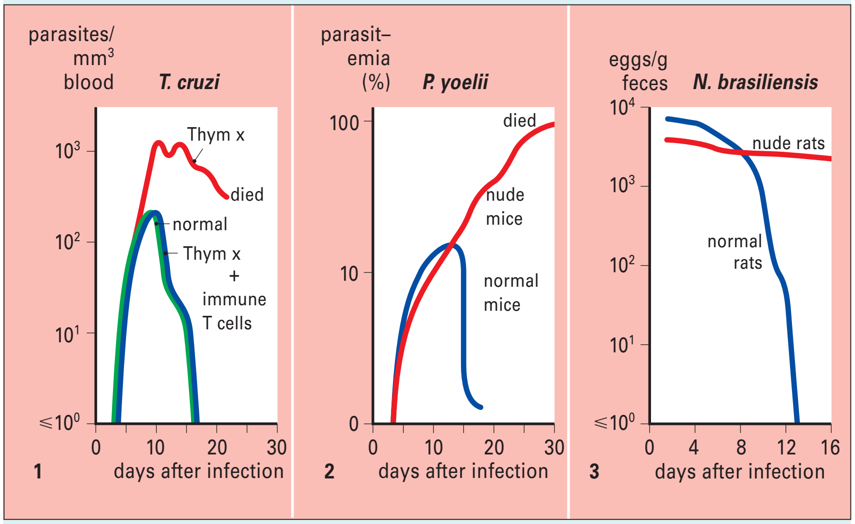

Fig. 15.8 The first two graphs plot the increase in number of blood-borne protozoa (parasitemia) following infection. (**1**) *T. cruzi* multiplies faster (and gives fatal parasitemia) in mice that have been thymectomized and irradiated to destroy T cells (Thym x). In normal mice, parasites are cleared from the blood by day 16. Reconstitution of T cell-deprived mice with T cells from immune mice restores their ability to control the parasitemia. In these experiments both thymectomized groups were given fetal liver cells to restore vital hematopoietic function. (**2**) *P. yoelii* causes a self-limiting infection in normal mice and the parasites are cleared from the blood by day 20. In nude mice the parasites continue to multiply, killing the mice after about 30 days. (**3**) This graph illustrates the time course of the elimination of the intestinal nematode *N. brasiliensis* from the gut of rats. In normal rats the worms are all expelled by day 13, as determined by the number of worm eggs present in the rats' feces. T cells are necessary for this expulsion to occur, as shown by the establishment of a chronic infection in the gut of nude rats.

in determining the outcome of parasitic infections have been extensively investigated.

As a result of early studies, predominantly in mouse infections, certain dogmas have arisen suggesting that:

- TH1 responses mediate killing of intracellular pathogens; and
- TH2 responses eliminate extracellular ones.

However, this is very much an oversimplification of the true picture.

Although the TH1/TH2 paradigm may be a useful tool in some situations, it is probably more realistic to consider that TH1 and TH2 phenotypes represent the extremes of a continuum of cytokine profiles and that perhaps it may be more accurate to look at the role of the cytokines themselves in the resolution of infectious disease.

Regulatory T cells are able to modulate the extremes of both TH2 and TH1 responses.

Cytokines, chemokines, and their receptors have important roles

Q. What experimental methods can be used to elucidate whether a particular cytokine is needed to clear a parasite infection?

A. Experiments with cytokine-knockout animals or cytokine blocking with specific antibody can determine whether a particular cytokine is required. This can be confirmed by observing whether exogenously administered cytokine speeds recovery.

Cytokines not only act on effector cells to enhance their cytotoxic or cytostatic capabilities, but also act as growth factors to increase cell numbers, while chemokines attract cells to the sites of infection. Thus in malaria, the characteristic enlargement of the spleen is caused by an enormous increase in cell numbers.

Other examples include:

- the accumulation of macrophages in the granulomas that develop in the liver in schistosomiasis;
- the eosinophilia characteristic of helminth infections; and
- the recruitment of eosinophils and mast cells into the gut mucosa that occurs in worm infections of the gastrointestinal tract.

Mucosal mast cells and eosinophils are both important in determining the outcome of some helminth infections and proliferate in response to the products of T cells – IL-3 and granulocyte–macrophage colony stimulating factor (GM-CSF), and IL-5, respectively.

However, an increase in cell number can itself harm the host. Thus administration of IL-3 to mice infected with *Leishmania major* can exacerbate the local infection and increase the dissemination of the parasites, probably through the proliferation of bone marrow precursors of the cells the parasites inhabit.

IL-10 and transforming growth factor-β (TGFβ), the regulatory cytokines (see Chapter 11), downregulate the proinflammatory response and thus minimize pathological damage.

Chemokines are key molecules in recruiting immune cells by chemotaxis, but also act in leukocyte activation, hematopoiesis, inflammation, and antiparasite immunity.

Protozoan parasites have been most studied in the context of chemokines and their diverse roles in the parasite–host relationship. For example, *T. gondii* possesses cyclophilin-18, which binds to the chemokine receptor CCR5 and induces IL-12 production by dendritic cells.

T cell responses to protozoa depend on the species

T cell-mediated immunity operating to control protozoan parasites depends on the species of animal infected and the location and complexity of the parasite life cycle within the host.

For example, in mouse models, the induction of TH1 cells with concomitant upregulation of IFNγ and nitric oxide (NO$^{\bullet}$) is crucial for protection of mice from leishmania. Strains of mice driving TH2 responses on infection, manifested by high levels of IL-4, IL-13, IL-10, and antibody, develop progressive and ultimately lethal disease (Fig. 15.9).

The polarization of helper T cell responses in murine models does not conveniently translate to humans, where both TH1 and TH2 responses appear to be involved in protection.

The importance of TH1 cells to protection from toxoplasmosis is also evident in murine models.

For malaria, the TH1/TH2 paradigm is also somewhat contradictory and less helpful in understanding immunity. This is because the type of immune response mounted and the ensuing risk of pathology depends on whether the first exposure to the parasite occurs during infancy or adulthood.

As a consequence immunity to malaria is best thought of in the context of regulated TH1 responses. Thus, in endemic populations, primary malaria infections in infants induce low levels of IFNγ and TNFα via an innate pathway (potentially involving dendritic cells), which leads to T cell priming.

The infection induces minimal pathology and the parasites can be cleared, either immunologically via maternal antibody or because parasites fail to thrive in fetal hemoglobin.

On reinfection, the malaria-primed T cells produce massive amounts of IFNγ and TNFα leading to an increased risk of unwanted pathology, including cerebral malaria.

Development of the immune response to *Leishmania major* and *Trichuris muris* infection

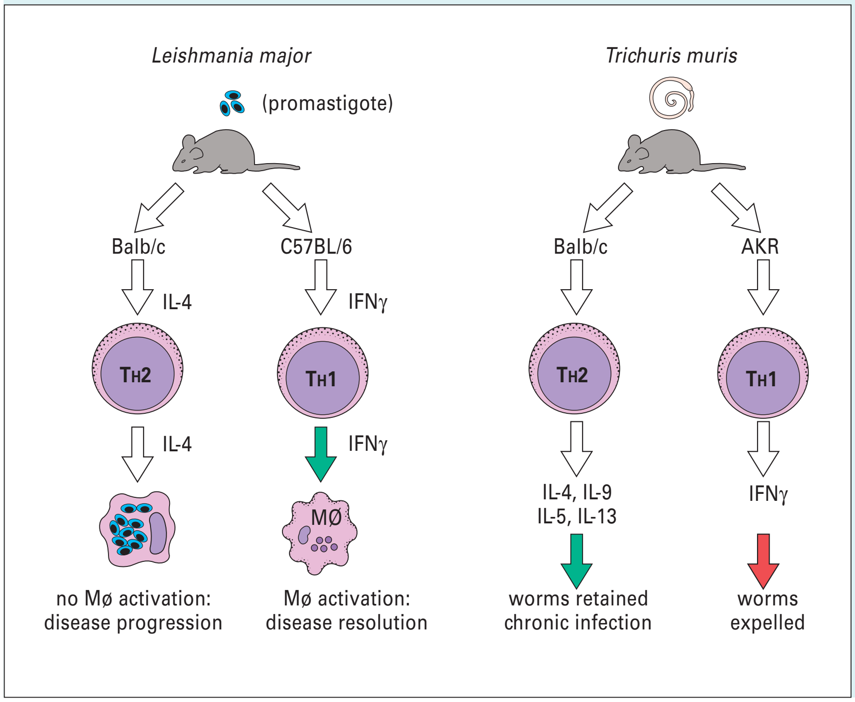

Fig. 15.9 The cytokines secreted by the different subsets of T cells and their effect on the resolution of the disease are shown. Note that resolution of infection is dependent on mouse strain. (Mø, macrophage)

Q. What effects will IFNγ and TNFα have on cerebral blood vessels?

A. These cytokines cause an increase in adhesion molecules (ICAM-1, VCAM-1) and synthesis of inflammatory chemokines (CCL2, CXCL10) producing leukocyte migration into the brain (see Chapter 6). They also cause an increase in the permeability of the vessels so that large serum molecules enter the CNS, and ionic equilibria are disturbed. This is referred to as a breakdown in the blood–brain barrier.

Further infections induce effective anti-parasite immunity principally through the development of an individual's own repertoire of high-affinity antibodies, which inhibit parasite development.

This change in immune environment ultimately leads to a switch in T cell phenotype from TH1 to a regulatory T cell phenotype in which raised levels of IL-10 and TGFβ can be detected.

By contrast, non-immune individuals who contract malaria for the first time in adulthood are unable to control their infections and are more likely to develop severe pathology.

This is believed to arise from cross-reactively primed T cells generated against other microbes that appear to contribute to the development of severe disease.

The immune response to worms depends upon TH2-secreted cytokines

IgE and eosinophilia are the hallmarks of the immune response to worm infections, and depend upon cytokines secreted by TH2 cells (see Fig. 15.9).

In humans schistosomiasis and infection with gastrointestinal nematodes, resistance to reinfection after drug treatment is correlated with the production of IgE and high pre-treatment levels of TH2 cytokines such as IL-4, IL-5, and IL-13.

The primary stimuli for TH2 development in schistosomiasis are egg antigens. Similarly the excretory and secretory products of nematodes have been shown to polarize cells towards TH2 responses. Again the control of T cell phenotype seems to be exerted by the dendritic cell after exposure to these substances.

The mechanisms of induction of TH2 responses are less well understood than TH1 responses.

- one hypothesis, the default hypothesis, suggests that unless the triggers for TH1 responses are received (including high IL-12), TH2 responses occur;
- more recent evidence, however, suggests that specific signals induce the T cell to make TH2 cytokines, probably including cell–cell interactions.

The pattern of cytokine production in infected hosts may be different from that in vaccinated hosts. For example:

- in mice infected with *S. mansoni*, IL-5-producing TH2 cells predominate;
- in mice that have been immunized, IgE levels and eosinophil numbers are low and TH1 cells predominate.

IFNγ activates effector cells that destroy lung stage larvae, via the production of NO$^{\bullet}$.

Q. How does IFNγ lead to the production of NO$^{\bullet}$?

A. It causes the production of inducible NO synthase in macrophages (see Fig. 9.31).

However, when adult worms start to produce eggs, a soluble egg antigen is released that has an effect only in susceptible mice. The antigen reduces levels of IFNγ and increases production of IL-5.

TH2 cytokines control effector mechanisms important in controlling intestinal worm infections. Perhaps the example that demonstrates this most clearly is *T. muris* infection in mice.

- animals normally resistant to infection develop persistent infections in IL-4 and/or IL-13 knockout mice;
- conversely, susceptible mice expel the worms if IL-4 activity is promoted by administration of neutralizing antibody against IFNγ.

The role of IFNγ in promoting chronic infection is again shown by the administration of IL-12 to mice soon after infection with the intestinal worm *Nippostrongylus brasiliensis* (Fig. 15.10).

N. brasiliensis stimulates IFNγ production, which delays expulsion of the worms.

IL-12 acts by inhibiting the production of TH2 cytokines – in particular IL-4 and IL-5, thereby preventing the production of IgE, eosinophilia, and mast cell hypertrophy. IL-9 is another TH2 cytokine that seems to be important in resistance to intestinal nematode infection

Course of infection with *Nippostrongylus brasiliensis* in mice

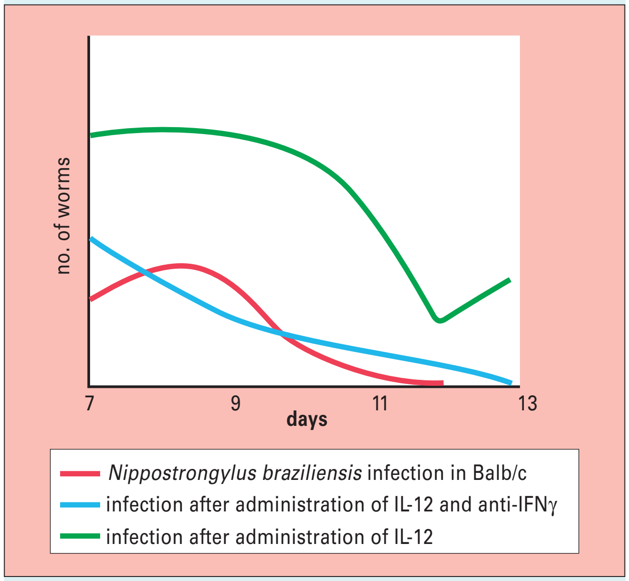

Fig. 15.10 The normal course of infection in Balb/c mice is that worms will be expelled by day 11. This is associated with a TH2 response. If IL-12 is administered straight after infection a TH1 response is promoted and as a result more worms establish and the infection is maintained for longer. If IL-12 is administered, but IFNγ is neutralized, this result is reversed.

and is involved in the production of mucosal mast cell responses and the production of IgE.

IL-9 transgenic mice that produce higher levels of this cytokine have enhanced expulsion of *T. muris*.

What is clear from a number of studies is that there is no single mechanism by which a TH2 response mediates expulsion of all intestinal worms.

The species of worm, its anatomical position within the gut, and the immune status of the host are all factors likely to influence whether a particular immune mechanism will be effective at promoting worm loss.

The host may isolate the parasite with inflammatory cells

In some parasitic infections, the immune system cannot completely eliminate the parasite, but reacts by isolating the organism with inflammatory cells. The host reacts to locally released antigen, which stimulates the production of cytokines that recruit cells to the region. An example of this has been shown in mice vaccinated with radiation-attenuated schistosome cercariae. Infiltrating cells, which are mostly TH1-type lymphocytes, surround the lung-stage larvae as early as 24 hours after intravenous challenge infection. This prevents subsequent migration to the site necessary for development into the adult parasite.

The schistosome egg granuloma in the liver is another example of the host reacting by ‘walling off’ the parasite. This reaction is a chronic cell-mediated response to soluble antigens released by eggs that have become trapped in the liver. Macrophages accumulate and release fibrogenic factors, which stimulate the formation of granulomatous tissue and, ultimately, fibrosis. Although this reaction may benefit the host, in that it insulates the liver cells from toxins secreted by the worm eggs, it is also the major source of pathology, causing irreversible changes in the liver and the loss of liver function. In the absence of T cells, there is no granuloma formation and no subsequent fibrous encapsulation.

Different mechanisms may affect:

- worms that inhabit different anatomical sites, such as the gut (e.g. *T. trichura*) or the tissues (e.g. *Onchocerca volvulus*); and
- different stages of the life cycle (e.g. schistosome larvae in the lungs and adult worms in the veins).

Parasites induce non-specific and specific antibody production

Many parasitic infections provoke a non-specific hypergammaglobulinemia, much of which is probably due to substances released from the parasites acting as B cell mitogens.

Levels of total immunoglobulins are raised:

- IgM in trypanosomiasis and malaria;
- IgG in malaria and visceral leishmaniasis.

The relative importance of antibody-dependent and antibody-independent responses varies with the infection and host (Fig. 15.11).

Relative importance of antibody-dependent and antibody-independent responses in protozoal infections

parasite and habitat		antibody-dependent			antibody-independent	
		importance	mechanism	means of evasion	importance	mechanism
T. brucei free in blood	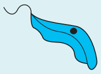	+ + + +	lysis with complement, which also opsonizes for phagocytosis	antigenic variation	–	
Plasmodium spp. inside red cell	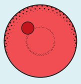	+ + +	blocks invasion, opsonizes for phagocytosis	intracellular; antigenic variation	liver stage + + + blood stage + + +	cytokines macrophage activation
T. cruzi inside macrophage	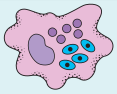	+ +	limits spread in acute infection, sensitizes for ADCC	intracellular	+ + + (chronic phase)	macrophage activation by IFNγ and TNFα, and killing by $NO^\bullet$ and metabolites of O_2
Leishmania spp. inside macrophage	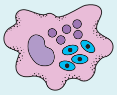	+	limits spread	intracellular	+ + + +	

Fig. 15.11 This table summarizes the relative importance of the two immune responses, the mechanisms involved, and, for antibody, the means by which the protozoon can evade damage by antibody. Antibody is the most important part of the immune response against those parasites that live in the blood stream, such as African trypanosomes and malarial parasites, whereas cell-mediated immunity is active against those like leishmania that live in the tissues. Antibody can damage parasites directly, enhance their clearance by phagocytosis, activate complement, or block their entry into their host cell and so limit the spread of infection. Once inside the cell the parasite is safe from the effects of antibody. *Trypanosoma cruzi* and *Leishmania* spp. are both susceptible to the action of oxygen metabolites release by the respiratory burst of macrophages, and to $NO^\bullet$. Treating macrophages with cytokines enhances release of these products and diminishes the entry and survival of the parasites. (ADCC, antibody-dependent cell-mediated cytotoxicity)

Mechanisms by which specific antibody controls some parasitic infections

parasite	*Plasmodium* spp. sporozoite, intestinal worms, trypanosome	*Plasmodium* spp. sporozoite and merozoite, *T. cruzi*, *T. gondii*	*Plasmodium* spp. trypanosome	schistosomes, *T. spiralis*, filarial worm larvae
mechanism	**1** 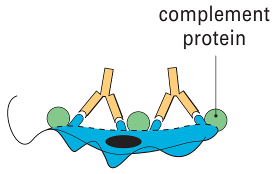	**2** 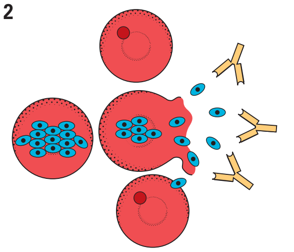	**3** 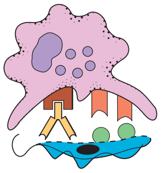	**4** 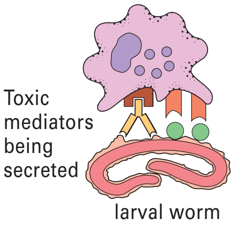
effect	direct damage or complement-mediated lysis	prevents spread by neutralizing attachment site, prevents escape from lysosomal vacuole, prevents inhibition of lysosomal fusion	enhancement of phagocytosis	antibody-dependent cell-mediated cytotoxicity (ADCC)

Fig. 15.12 (**1**) Direct damage. Antibody activates the classical complement pathway, causing damage to the parasite membrane and increasing susceptibility to other mediators. (**2**) Neutralization. Parasites such as *Plasmodium* spp. spread to new cells by specific receptor attachment; blocking the merozoite binding site with antibody prevents attachment to the receptors on the erythrocyte surface and prevents further multiplication. (**3**) Enhancement of phagocytosis. Complement C3b deposited on the parasite membrane opsonizes it for phagocytosis by cells with C3b receptors (e.g. macrophages). Macrophages also have Fc receptors. (**4**) Eosinophils, neutrophils, platelets, and macrophages may be cytotoxic for some parasites when they recognize the parasite via specific antibody (ADCC). The reaction is enhanced by complement.

The mechanisms by which specific antibody can control parasitic infections and its effects are summarized in Fig. 15.12. Antibody:

- can act directly on protozoa to damage them, either by itself or by activating the complement system (Fig. 15.13);
- can neutralize a parasite directly by blocking its attachment to a new host cell, as with *Plasmodium* spp., whose merozoites enter red blood cells through a special receptor – their entry is inhibited by specific antibody (Fig. 15.14);
- may prevent spread (e.g. in the acute phase of infection by *T. cruzi*);
- can enhance phagocytosis by macrophages – phagocytosis is increased even more by the addition of complement; these effects are mediated by Fc and C3 receptors on macrophages, which may increase in number as a result of macrophage activation;
- is involved in antibody-dependent cell-mediated cytotoxicity (ADCC), for example in infections caused by *T. cruzi*, *T. spiralis*, *S. mansoni*, and filarial worms – cytotoxic cells such as macrophages, neutrophils, and eosinophils adhere to antibody-coated worms by means of their Fc and C3 receptors and degranulate, spilling their toxic contents onto the worm.

Different antibody isotypes may have different effects. In individuals infected with schistosomes, parasite-specific IgE and IgA are associated with resistance to infection and there is an inverse relationship between the amount of IgE in the blood and reinfection.

Direct effect of specific antibody on sporozoites of malaria parasites

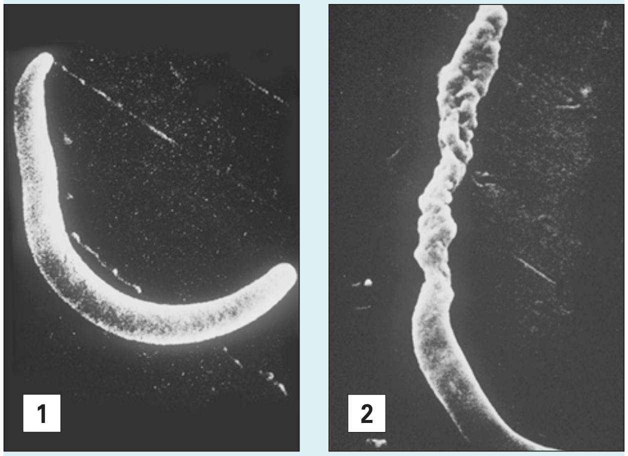

Fig. 15.13 These scanning electron micrographs show a sporozoite of *P. berghei*, which causes malaria in rodents, before (**1**) and after (**2**) incubation in immune serum. The surface of the sporozoite is damaged by the antibody, which perturbs the outer membrane, causing leakage of fluid. Specific antibody protects against infection with *Plasmodium* spp. at several of the extracellular stages of the life cycle. The antibody is stage specific in each case. (Courtesy of Dr R Nussenzweig)

Effect of antibody on malarial parasites

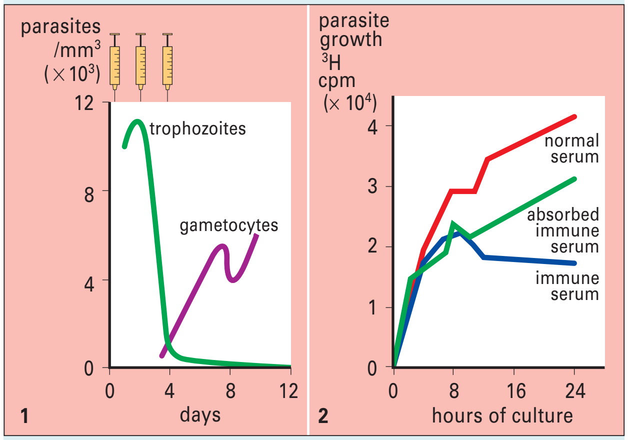

Fig. 15.14 (**1**) Transfer of γ-globulin from immune adults to a child infected with *Plasmodium falciparum* caused a sharp drop in parasitemia. Specific antibody acts at the merozoite stage in the life of the parasite and prevents the initiation of further cycles of multiplication in the blood. The development of gametocytes from existing intracellular forms is unaffected. (**2**) In culture, the presence of immune serum blocks the continued increase in number of *Plasmodium knowlesi* (a malarial parasite of monkeys), as measured by incorporation of ^{3}H-leucine. It stops multiplication at the stage after schizont rupture by preventing the released merozoites from invading fresh red blood cells. The inhibitory activity of the immune serum can be reduced by prior absorption of the specific antibody with free schizonts.

Q. What roles does IgE have in immune defense?

A. IgE mediates inflammation by binding to mast cells and basophils, sensitizing them to parasite antigens. Additionally it can act as an opsonin for eosinophils (see Chapter 3).

IgG4 appears to block the action of IgE; reinfection is more likely in children who have high levels of IgG4. Class switching to IgG4 appears to occur in the context of a modified TH2 response involving the induction of Tregs.

The development of immunity seems to depend upon a switch from IgG4 to IgE that occurs with age – infection rates are highest in 10–14-year-olds when IgG4 levels are also at their highest.

In many infections it is difficult to distinguish between cell-mediated and antibody-mediated responses because both can act in concert against the parasite. This is illustrated in Fig. 15.15, which summarizes the immune reaction that can be mounted against schistosome larvae.

EFFECTOR CELLS SUCH AS MACROPHAGES, NEUTROPHILS, EOSINOPHILS, MAST CELLS, AND PLATELETS CAN ATTACK PARASITES

Antibody and cytokines produced specifically in response to parasite antigens enhance the anti-parasitic activities of all these effector cells, though tissue macrophages, monocytes, and granulocytes have some intrinsic activity before enhancement. The point of entry of the parasite is obviously important, for example:

- the cercariae of *S. mansoni* enter through the skin – experimental depletion of macrophages, neutrophils, and eosinophils from the skin of mice increases their susceptibility to infection;
- trypanosomes and malarial parasites entering the blood are removed from the circulation by phagocytic cells in the spleen and liver;
- comparison of strains of mice with various immunological defects for their resistance to infection by *Trypanosoma rhodesiense* shows that the African trypanosomes are destroyed by macrophages, and later in infection, when opsonized with antibodies and complement C3b, they are taken up by macrophages in the liver more quickly still.

Before acting as APCs initiating an immune response, macrophages act as effector cells to inhibit the multiplication of parasites or even to destroy them. They also secrete molecules that regulate the inflammatory response:

- some of these molecules – IL-1, IL-12, TNFα, and the colony stimulating factors (CSFs) – enhance immunity by activating other cells or stimulating their proliferation;
- others, like IL-10, prostaglandins, and TGFβ, may be anti-inflammatory and immunosuppressive.

Macrophages can kill extracellular parasites

Phagocytosis by macrophages provides an important defense against the smaller parasites. Macrophages also secrete many cytotoxic factors, enabling them to kill parasites without ingesting them.

When activated by cytokines, macrophages can kill both relatively small extracellular parasites, such as the erythrocytic stages of malaria, and also larger parasites, such as the larval stages of the schistosome. Macrophages also:

- act as killer cells through ADCC – specific IgG and IgE, for instance, enhance their ability to kill schistosomules;
- secrete cytokines, such as TNFα and IL-1, which interact with other types of cell, for example rendering hepatocytes resistant to malarial parasites.

Reactive oxygen intermediates (ROIs) are generated by macrophages and granulocytes following phagocytosis of *T. cruzi*, *T. gondii*, *Leishmania* spp., and malarial parasites, for instance. Filarial worms and schistosomes also stimulate the respiratory burst.

When activated by cytokines, macrophages release more superoxide and hydrogen peroxide than normal resident macrophages, and their oxygen-independent killing mechanisms are similarly enhanced.

Nitric oxide, a product of L-arginine metabolism, is one of the potent oxygen-independent toxins. Its synthesis by macrophages in mouse experimental systems is induced by the cytokines IFNγ and TNFα and is greatly increased when they act synergistically. NO$^\bullet$ can also be produced by endothelial cells. It contributes to host resistance in leishmaniasis, schistosomiasis, and malaria, and is probably important in the control of most parasitic infections (see Fig. 15.15). For instance, the innate resistance to infection by *T. gondii* that is lost in immuno-

Possible effector responses to schistosomules

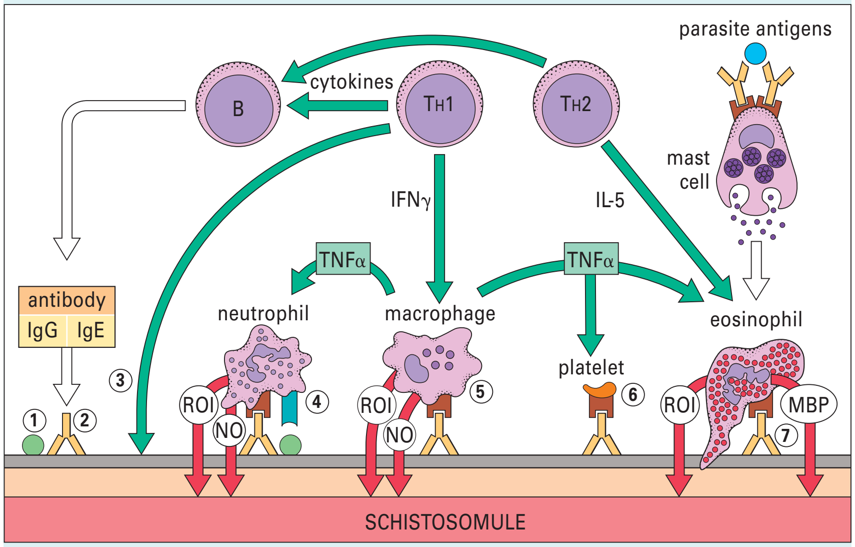

Fig. 15.15 The various effector mechanisms damaging schistosomes in vitro are shown. Complement alone damages worms (**1**) and also does so in combination with antibody (**2**). TH1 cells may act directly, reducing the number of larvae in the lungs (**3**). Antibody sensitizes neutrophils (**4**), macrophages (**5**), platelets (**6**), and eosinophils (**7**) for ADCC. Neutrophils and macrophages probably act by releasing toxic oxygen and nitrogen metabolites, whereas eosinophils damage the worm tegument by the release of major basic protein (MBP) plus reactive oxygen intermediates (ROI). The response is potentiated by cytokines (e.g. TNFα). IgE antibody is important in sensitizing both eosinophils and local mast cells, which release a variety of mediators, including those that activate the eosinophils.

compromised individuals appears to be due to the inhibition of parasite multiplication by such an oxygen-independent mechanism.

Activation of macrophages is a feature of early infection

All macrophage effector functions are enhanced soon after infection. Although their specific activation is by cytokines secreted by T cells (e.g. IFNγ, GM-CSF, IL-3, and IL-4), they can also be activated by T cell-independent mechanisms, for example:

- NK cells secrete IFNγ when stimulated by IL-12 produced by macrophages;
- macrophages secrete TNFα in response to some parasite products (e.g. phospholipid-containing antigens of malarial parasites and some *T. brucei* antigens) – this TNFα then activates other macrophages.

Although TNFα may be secreted by several other cell types, activated macrophages are the most important source of TNFα, which is necessary for protective responses to several species of protozoa (e.g. *Leishmania* spp.) and helminths. Thus TNFα activates macrophages, eosinophils, and platelets to kill the larval form of *S. mansoni*, its effects being enhanced by IFNγ.

TNFα may have harmful as well as beneficial effects on the infected host, depending upon the amount produced and whether it is free in the circulation or locally confined. Serum concentrations of TNFα in falciparum malaria correlate with the severity of the disease. Administration of TNFα cures a susceptible strain of mice infected with the rodent malarial parasite *Plasmodium chabaudi*, but kills a genetically resistant strain. Presumably the latter can already make enough TNFα to control parasite replication, and any more has toxic effects.

Neutrophils can kill large and small parasites

The effector properties displayed by macrophages are also seen in neutrophils. Neutrophils are phagocytic and can kill by both oxygen-dependent and oxygen-independent mechanisms, including NO$^{\bullet}$. They produce a more intense respiratory burst than macrophages and their secretory granules contain highly cytotoxic proteins.

Q. Which groups of cytotoxic protein are found in neutrophil granules?

A. Defensins, seprocidins, and cathelicidins (see Chapter 2).

Neutrophils can be activated by cytokines, such as IFNγ, TNFα, and GM-CSF.

Extracellular destruction by neutrophils is mediated by hydrogen peroxide, whereas granular components are involved in the intracellular destruction of ingested organisms.

Neutrophils are present in parasite-infected inflammatory lesions and probably act to clear parasites from bursting cells.

Like macrophages, neutrophils bear Fc and complement receptors and can participate in antibody-dependent cytotoxic reactions to kill the larvae of *S. mansoni*, for example. In this mode, they can be more destructive than eosinophils against several species of nematode, including *T. spiralis*, though the relative effectiveness of the two types of cell may depend upon the isotype and specificity of antibody.

Eosinophils are characteristically associated with worm infections

It has been suggested that:

- the eosinophil evolved specifically as a defense against the tissue stages of parasites that are too large to be phagocytosed; and
- the IgE-dependent mast cell reaction has evolved primarily to localize eosinophils near the parasite and enhance their anti-parasitic functions.

The importance of eosinophils in vivo has been shown by experiments using antiserum against eosinophils. Mice infected with *T. spiralis* and treated with the antiserum develop more cysts in their muscles than the controls – without the protection offered by eosinophils, the mice cannot eliminate the worms and so encyst the parasites to minimize damage.

However, recent work has shown that although eosinophils can help the host to control a worm infection, particularly by limiting migration through the tissues, they do not always do so. For instance, their removal does not abolish the immunity of mice infected with *S. mansoni*, nor does this increase the parasite load in a tapeworm infection.

Removal of IL-5, which is important in the generation and activation of eosinophils, did not change the outcome of *T. spiralis* or *T. muris* infection. In contrast the infectivity of *Strongyloides venezuelensis* is enhanced in IL-5-deficient mice. Although *T. spiralis* worm burdens were not affected in a primary infection of IL-5 deleted mice, the worm numbers were significantly higher after challenge infection.

The role of IL-5 and therefore eosinophils has also been suggested from human epidemiological studies on gastrointestinal nematode infections where, after drug treatment, low reinfection worm burdens were associated with high pre-treatment levels of IL-5.

Elevated eosinophilia is often associated with high levels of IgE, both of which are hallmarks of infection with parasites. Although eosinophils express FcεR1, most of the protein is confined to the cytoplasm, and there is little evidence for IgE-dependent function.

Eosinophils can kill helminths by oxygen-dependent and independent mechanisms

Eosinophils are less phagocytic than neutrophils. They degranulate in response to perturbation of their surface membrane and their activities are enhanced by cytokines such as TNFα and GM-CSF. Most of their activities, however, are controlled by antigen-specific mechanisms. Thus their binding in vitro to the larvae of worms coated with IgE or IgA (e.g. *S. mansoni* and *T. spiralis*) increases the release of their granular contents onto the surface of the worms (see Fig. 15.15).

Damage to schistosomules can be caused by the major basic protein (MBP) of the eosinophil crystalloid core. MBP is not specific for any particular target, but because it is confined to a small space between the eosinophil and the schistosome, there is little damage to nearby host cells.

Eosinophils and mast cells can act together

The killing of *S. mansoni* larvae by eosinophils is enhanced by mast cell products, and when studied in vitro eosinophils from patients with schistosomiasis are found to be more effective than those from normal subjects. The antigens released cause local IgE-dependent degranulation of mast cells and the release of mediators. These selectively attract eosinophils to the site and further enhance their activity. Other products of eosinophils later block the mast cell reactions. These effector mechanisms may function in vivo, as has been shown in monkeys, where schistosome killing is associated with eosinophil accumulation.

Mast cells control gastrointestinal helminths

In the case of *T. spiralis* and *Heligmosomoides polygyrus* there is good evidence to suggest the involvement of mucosal mast cells (Fig. 15.16).

Section through the gut of a mouse infected with *Heligmosomoides polygyrus*

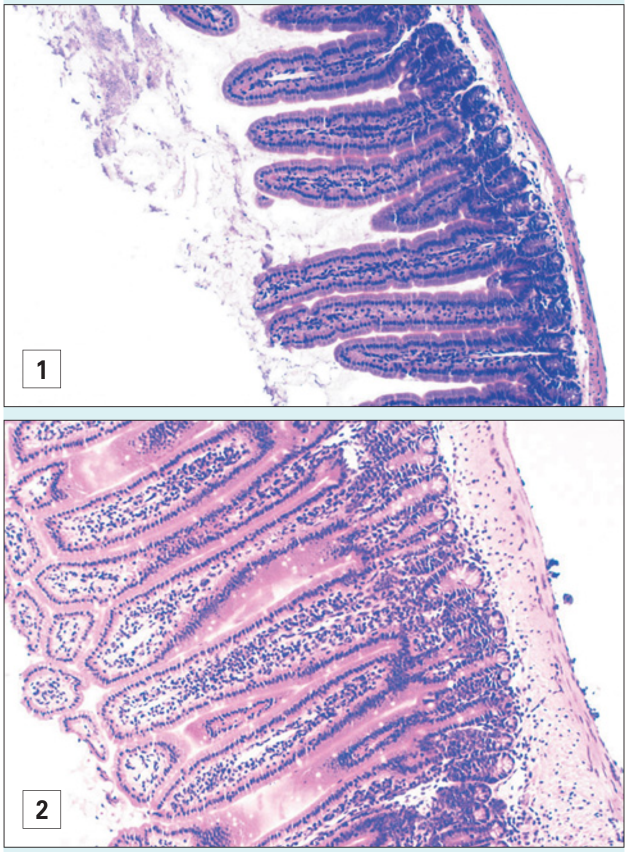

Fig. 15.16 (**1**) Gut of an uninfected mouse. (**2**) Gut of an infected mouse. The crypts have shortened and a large influx of mast cells can be clearly seen.

Mast cells contain a number of lipid mediators, such as prostaglandins, proteases, and histamine. In addition, they also represent a source of cytokines such as IL-3, IL-4, IL-5, GM–CSF, and TNFα.

Following mast cell activation, the mast cell contents are released resulting in changes to the permeability of the intestinal epithelium and ultimately an environment that appears hostile for continued *T. spiralis* survival. By contrast, expulsion of *N. brasiliensis* and *T. muris* still proceeds normally following depression of mastocytosis, suggesting that the mast cell is not the major effector cell type.

Therefore, although TH2 cytokines are critical for the elimination of worms from the gut, the exact effector mechanism operating may vary.

Platelets can kill many types of parasite

Potential targets for platelets include the larval stage of flukes, *T. gondii*, and *T. cruzi.*

Like other effector cells, the cytotoxic activity of platelets is enhanced by treatment with cytokines (e.g. IFNγ and TNFα). In rats infected with *S. mansoni*, platelets become larvicidal when acute phase reactants appear in the serum but before antibody can be detected. Incubation of normal platelets in such serum can cause their activation.

Platelets, like macrophages and the other effector cells, also bear Fcε receptors on their surface membrane, by which they mediate antibody-dependent cytotoxicity associated with IgE.

PARASITES HAVE MANY DIFFERENT ESCAPE MECHANISMS

It is a necessary characteristic of all successful parasitic infections that they can evade the full effects of their host's immune responses. Parasites have developed many different ways of doing this. Some even exploit cells and molecules of the immune system to their own advantage – *Leishmania* parasites, by using complement receptors to effect their entry into macrophages, avoid triggering the oxidative burst and thus destruction by its toxic products.

Despite their protective role in the immune response to many different parasites:

- host TNFα actually stimulates egg production by adult worms of *S. mansoni*;
- IFNγ is used as a growth factor by *T. brucei*.

Parasites can resist destruction by complement

In the case of *Leishmania*, resistance correlates with virulence:

- *L. tropica*, which is easily killed by complement, causes a localized self-healing infection in the skin; whereas
- *L. donovani*, which is ten times more resistant to complement, becomes disseminated throughout the viscera, causing a disease that is often fatal.

The mechanisms whereby parasites can resist the effect of complement differ:

- the lipophosphoglycan (LPG) surface coat of *L. major* activates comple-ment, but the complex is then shed so the parasite avoids lysis;
- the trypomastigotes of *T. cruzi* bear a surface glycoprotein that has activity resembling the decay accelerating factor (DAF) that limits the complement reaction.

The resistance that schistosomules acquire as they mature is also correlated with the appearance of a surface molecule similar to DAF.

Intracellular parasites can avoid being killed by oxygen metabolites and lysosomal enzymes

Intracellular parasites that live inside macrophages have evolved different ways of avoiding being killed by oxygen metabolites and lysosomal enzymes (Fig. 15.17):

- *T. gondii* penetrates the macrophage by a non-phagocytic pathway and so avoids triggering the oxidative burst;
- *Leishmania* spp. can enter by binding to complement receptors – another way of avoiding the respiratory burst.

Leishmania organisms also possess enzymes such as superoxide dismutase, which protects them against the action of oxygen radicals.

It can be demonstrated that the vacuole in which *Leishmania* organisms survive is lysosomal in nature (Fig. 15.18), but the parasites have evolved mechanisms that protect it against enzymatic attack. The LPG surface coat not only acts as a scavenger of oxygen metabolites and affords protection against enzymatic attack, but a glycoprotein, Gp63 (Fig. 15.19), inhibits the action of the macrophage's lysosomal enzymes.

Leishmania spp. can also downregulate the expression of MHC class II molecules on the macrophages they inhabit, thus reducing their capacity to stimulate TH cells.

These escape mechanisms, however, are less efficient in the immune host.

Q. Why would reduction in MHC molecule expression be less effective in an immune host?

A. In immune hosts, the release of IFNγ enhances MHC molecule expression by APCs. In addition the level of stimulation required by a primed T cell is less than for a naive T cell, due in part to its enhanced level of receptors for co-stimulatory signals (see Chapter 7).

Parasites can disguise themselves

Parasites that are vulnerable to specific antibody have evolved different methods of evading its effects.

The African trypanosome undergoes antigenic variation

The molecule that forms the surface coat of the African trypanosome, the variable surface glycoprotein (VSG), changes to protect the underlying surface membrane from the host's defense mechanisms. New populations of parasites are antigenically distinct from previous ones (Figs 15.20 and 15.21).

Several antigens of malarial parasites also undergo antigenic variation.

For example, the *P. falciparum* erythrocyte membrane protein-1 (PfEMP1) is extremely polymorphic and

The different ways by which protozoa that multiply within macrophages escape digestion by lysosomal enzymes

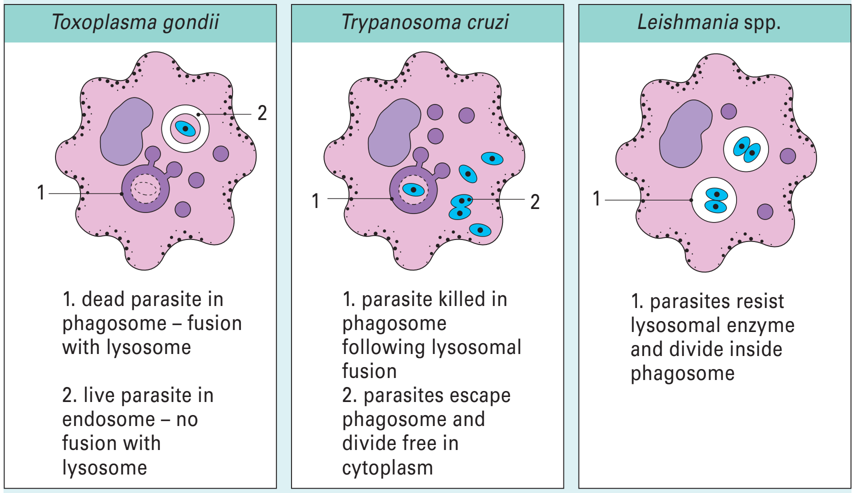
Toxoplasma gondii
Trypanosoma cruzi
Leishmania spp.
1
2
1
2
1
1. dead parasite in phagosome – fusion with lysosome
2. live parasite in endosome – no fusion with lysosome
1. parasite killed in phagosome following lysosomal fusion
2. parasites escape phagosome and divide free in cytoplasm
1. parasites resist lysosomal enzyme and divide inside phagosome

Fig. 15.17 *T. gondii* – Live parasites enter the cell actively into a membrane-bound vacuole. They are not attacked by enzymes because lysosomes do not fuse with this vacuole. Dead parasites, however, are taken up by normal phagocytosis into a phagosome (by interaction with the Fc receptors on the macrophage if they are coated with antibody) and are then destroyed by the enzymes of the lysosomes that fuse with it. *T. cruzi* – Survival of these parasites depends upon their stage of development; trypomastigotes escape from the phagosome and divide in the cytoplasm whereas epimastigotes do not escape and are killed. The proportion of parasites found in the cytoplasm is decreased if the macrophages are activated. *Leishmania* spp. – These parasites multiply within the phagosome and the presence of a surface protease helps them resist digestion. If the macrophages are first activated by cytokines, the number of parasites entering the cell and the number that replicate diminish.

The leishmanial vacuole is lysosomal in nature

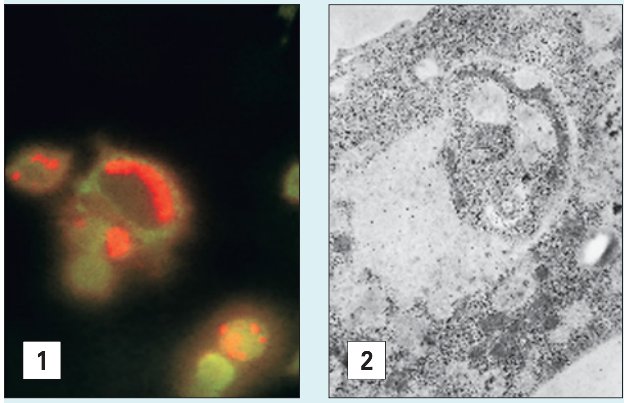

Fig. 15.18 (**1**) Immunofluorescence of *Leishmania mexicana*-infected murine macrophages probed with a rhodamine-conjugated anti-tubulin antibody to illustrate the parasite (stained yellow/red) and a fluorescein-conjugated monoclonal antibody, which reacts with the late endosomal/lysosomal marker LAMP-1 (stained green). (**2**) Immunoelectron micrograph of *L. mexicana*-infected murine macrophage probed with gold-labeled anti-cathepsin D demonstrating the lysosomal aspartic proteinase in the leishmanial vacuole. (Courtesy of Dr David Russell)

variable between different strains of the parasite because it is perpetually exposed to the immune system by its location on the red cell membrane.

Other parasites acquire a surface layer of host antigens

Other parasites, such as schistosomes, acquire a surface layer of host antigens so that the host does not distinguish them from 'self'.

Schistosomules cultured in medium containing human serum and red blood cells can acquire surface molecules containing A, B, and H blood group determinants. They can also acquire MHC molecules and immunoglobulins. However, schistosomules maintained in a medium devoid of host molecules also become resistant to attack by antibody and complement, as mentioned above.

Some extracellular parasites hide from immune attack

Some species of protozoa (e.g. *Entameba histolytica*) and helminths (e.g. *T. spiralis*) form protective cysts, while adult worms of *Onchocerca volvulus* in the skin induce the host to surround them with collagenous nodules.

Intestinal nematodes and tapeworms are preserved from many host responses simply because they live in the gut.

Two surface antigens of *Leishmania*

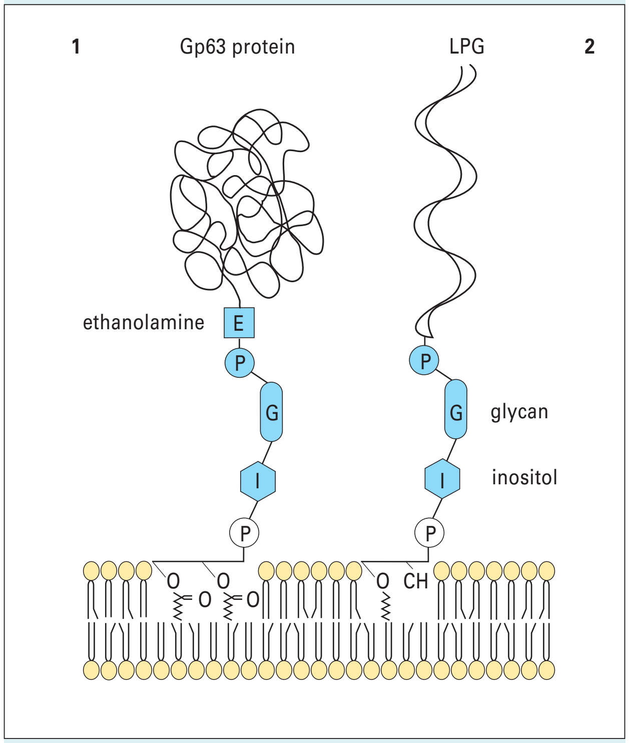

Fig. 15.19 Schematic representation of two surface antigens of *Leishmania* that are anchored to the membrane by phosphatidylinositol tails (GPI anchors). (**1**) This protein antigen, Gp63, has protease activity. That of *L. mexicana*, together with lipophosphoglycan (LPG), binds complement. This enables the promastigote to enter the macrophage through the C3 complement receptor. (**2**) This glycolipid antigen, a LPG, imparts resistance to complement-mediated lysis. That of *L. major* binds C3b, the third component of complement, enabling the promastigote to enter through the CR1 complement receptor. Antibodies to both antigens confer protection against murine cutaneous leishmaniasis. Note that many coat proteins of parasites, such as the variable surface glycoprotein (VSG) of *T. brucei*, are now known to be bound to the surface membrane by a GPI anchor.

Antigenic variation in trypanosomes

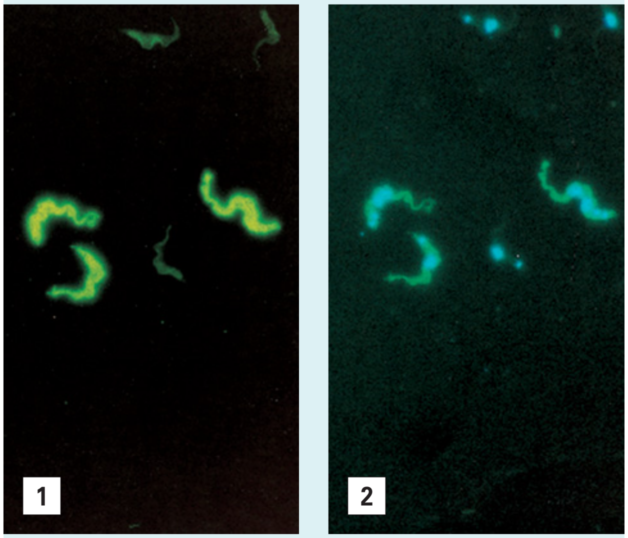

Fig. 15.20 Immunofluorescent labeling of trypanosomes with a variant antigen-type specific monoclonal antibody (**1**). Panel (**2**) shows the same field of view where the nuclei and kinetoplasts of all the parasites are stained with a dye that binds to DNA. Only some of the parasites express a given antigen variant. (Courtesy of Dr Mike Turner)

Some extracellular parasites can withstand immune attack

There are numerous examples of simple, physical, protective strategies in parasites:

- nematodes have a thick extracellular cuticle, which protects them from the toxic effects of an immune response;
- the tegument of schistosomes thickens during maturation to offer similar protection;
- the loose surface coat of many nematodes may slough off under immune attack;
- tapeworms actually prevent attack by secreting an elastase inhibitor, which stops them attracting neutrophils.

Many parasitic worms have evolved methods of resisting the oxidative burst. For instance, schistosomes have surface-associated glutathione S-transferases, and *Onchocerca* spp. can secrete superoxide dismutase.

Some nematodes and trematodes have evolved an elegant method of disabling antibodies by secreting proteases, which cleave immunoglobulins, removing the Fc portion, and preventing their interaction with Fc receptors on phagocytic cells; for example, schistosomes can cleave IgE.

Most parasites interfere with immune responses for their benefit

Parasites produce molecules that interfere with host immune function

Parasites produce molecules that can affect the phenotype of the adaptive response, which may be to their own advantage (Figs 15.22 and 15.23).

In leishmaniasis, T cells from patients infected with *L. donovani* when cultured with specific antigen do not secrete IL-2 or IFNγ. Their production of IL-1 and expression of MHC class II molecules is also decreased, whereas secretion of prostaglandins is increased. IL-2, characteristic of TH1 responses, is also deficient in other protozoal infections including malaria, African trypanosomiasis, and Chagas' disease. In mice infected with *T. cruzi*, a parasite product appears to interfere with expression of the IL-2 receptor.

Filarial worms secrete a protease inhibitor that has been shown to affect the proteases critical in the processing of antigens to peptides resulting in the reduction of class II molecule presentation in filariasis. Onchocystatin – one such protease inhibitor – is also able to modulate T cell proliferation and elicit the upregulation of IL-10

Antigenic variation in African trypanosomes

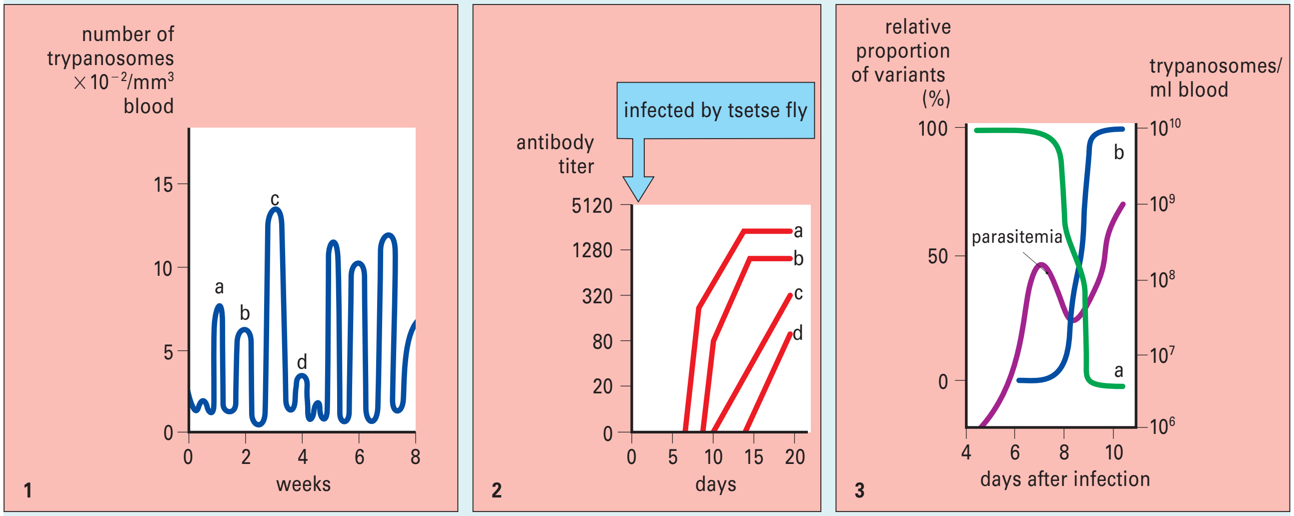

Fig. 15.21 Trypanosome infections may run for several months giving rise to successive waves of parasitemia. Graph (**1**) shows a chart of the fluctuation in parasitemia in a patient with sleeping sickness. Although infection was initiated by a single parasite, each wave is caused by an immunologically distinct population of parasites (a, b, c, d); protection is not afforded by antibody against any of the preceding variants. There is a strong tendency for new variants to appear in the same order in different hosts. Variation does not occur in immunologically compromised animals (i.e. animals treated to deprive them of some aspect of immune function). Graph (**2**) shows the time course of production of antibody against four variants in a rabbit bitten by a tsetse fly carrying *Trypanosoma brucei*. Antibody to successive variants appears shortly after the appearance of each variant and rises to a plateau. The appearance of antibody drives the parasite towards another variant type. Graph (**3**) shows the kinetics of one cycle of antigenic variation. A rat was infected with a homogeneous population of one variant (a) of *T. brucei*. The second wave of parasitemia develops as the new variant (b) emerges and predominates.

expression and is therefore able to modulate the T cell phenotype. Prostaglandins (PGs) produced by helminth parasites may also perform a similar role by modulating APC function. PGE$_2$ is produced by filarial parasites and tapeworms and blocks the production of IL-12 by dendritic cells and thus may direct responses towards TH2.

Phosphorylcholine (PC)-containing molecules are commonly found in infectious organisms and experiments using a nematode PC-bearing glycoconjugate, ES-62, have been shown to desensitize APCs to subsequent exposure to LPS and may therefore also skew against a TH1 response (LPS is a classical inducer of TH1 responses). ES-62 is also able to inhibit proliferation by both T cells and B cells, causes a decrease in the level of protein kinase C rendering both B and T cells anergic.

Parasites also produce cytokine-like molecules mimicking TGFβ, migration inhibitory factor (MIF), and a histamine-releasing factor.

Genes encoding possible cytokine homologs are being found as part of the genome sequencing projects that are under way for many parasites. Although the sequences are related to cytokines or cytokine receptors, their functions remain to be established.

Soluble parasite antigens released in huge quantities may impair the host's response by a process termed **immune distraction**. Thus the soluble antigens (S or heat-stable antigens) of *P. falciparum* are thought to mop up circulating antibody, providing a 'smokescreen' and diverting the antibody from the body of the parasite.

Many of the surface antigens that are shed are soluble forms of molecules inserted into the parasite membrane by a GPI anchor, including the VSG of *T. brucei*, the LPG or 'excreted factor' of *Leishmania* (see Fig. 15.19), and several surface antigens of schistosomules. These are released by endogenous phosphatidylinositol-specific phospholipases.

The hypergammaglobulinemic immunoglobulins produced by malaria parasites can bind to FcγRIIB, which may benefit the parasite.

Q. Which cells express FcγRIIB and what effect does ligation of this receptor have on them?

A. B cells express the receptor and ligation inhibits antibody synthesis.

INFLAMMATORY RESPONSES CAN BE A CONSEQUENCE OF ELIMINATING PARASITIC INFECTIONS

Immunosuppression is a common feature of chronic helminth infections, both parasite-specific and general-

Some mechanisms by which parasites avoid host immunity

parasite	habitat	main host effector mechanism	method of avoidance
Trypanosoma brucei	blood stream	antibody + complement	antigenic variation
Plasmodium spp.	hepatocyte bloo dstream	T cells, antibody	antigenic variation, sequestration
Toxoplasma gondii	macrophage	ROI, $NO^{\bullet}$, lysosomal enzymes	suppresses IL-12, inhibits fusion of lysozymes
Trypanosoma cruzi	many cells	ROI, $NO^{\bullet}$, lysosomal enzymes	escapes to cytoplasm so avoiding digestion
Leishmania spp.	macrophage	ROI, $NO^{\bullet}$, lysosomal enzymes	induction of Tregs, resists digestion by phagolysosome
Schistosoma spp.	skin, blood, lungs, portal veins	myeloid cells antibody + complement	acquisition of host antigens (e.g. IgG), proteolytic cleavage of immune proteins, inhibition of dendritic cell maturation
filariasis	lymphatics	myeloid cells	induction of Tregs, secretion of cytokine mimics, interference with antigen processing

$NO^{\bullet}$, nitric oxide; ROI, reactive oxygen intermediates

Fig. 15.22 A summary of the various methods that parasites have evolved to avoid host defense mechanisms.

ized. For example, patients with schistosomiasis and filariasis have diminished responsiveness to antigens from the infecting parasite. Studies have also shown diminished responses to bystander infections and vaccinations. This spillover suppression may in fact be beneficial to the host in some situations. Reduced inflammatory responses have been observed in *Helicobacter pylori* infection and malaria.

Parasites have co-evolved with humans over millions of years and until recently it was normal for people to carry worms, a fact that argues for their importance in the '**hygiene hypothesis**', which proposes that the rise in immune disorders, including allergies and autoimmune illness, is due to cleaner living conditions and the almost complete elimination of parasitic infections in westernized societies.

The ability of parasites to suppress hyperactive immune responses is believed to be due to the induction of regulatory T cells (see Fig. 15.23). Understanding how regulatory T cells, also known as suppressor T cells, are induced and how they dampen immune responses is the focus of intensive research. It seems to be the dendritic cell that polarizes the T cell towards a regulator phenotype after exposure to parasite extracts. The schistosome lyso-PS with acyl chains not present on mammalian lyso-PS was shown to activate dendritic cells via TLR2, which when incubated with T cells induced IL-10-producing regulatory T cells.

PARASITIC INFECTIONS HAVE IMMUNOPATHOLOGICAL CONSEQUENCES

Apart from the directly destructive effects of some parasites and their products on host tissues, many immune responses themselves have pathological effects.

In malaria, African trypanosomiasis, and visceral leishmaniasis, the increased number and heightened activity of macrophages and lymphocytes in the liver and spleen lead to enlargement of those organs. In schistosomiasis much of the pathology results from the T cell-dependent granulomas forming around eggs in the liver. The gross changes in individuals with elephantiasis are probably caused by immunopathological responses to adult filariae in the lymphatics.

The formation of immune complexes is common – they may be deposited in the kidney, as in the nephrotic syndrome of quartan malaria, and may give rise to many other pathological changes. For example, tissue-bound immunoglobulins have been found in the muscles of mice infected with African trypanosomes and in the choroid plexus of mice with malaria.

Some immunomodulatory effects of parasites

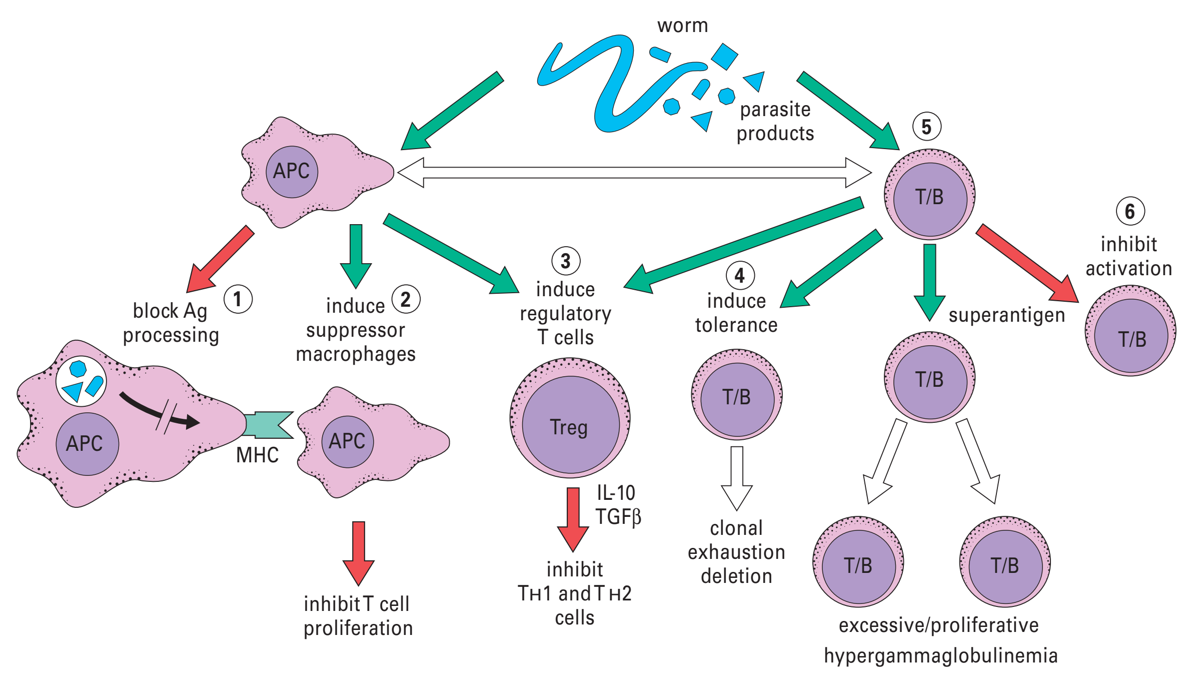

Fig. 15.23 Interference with the host's immune response by molecules released by protozoa or worms. Parasite products act via the APC to (**1**) interfere with antigen processing or presentation (e.g. protease inhibitors from filarial parasites interfere with proteases in the MHC signaling pathway and block antigen presentation); (**2**) induce suppressor macrophages, which can inhibit T cell proliferation; (**3**) induce regulatory T cells (e.g. lyso-PS from schistosomes acts via TLR2 on dendritic cells to induce T cells that secrete IL-10, which inhibits the inflammatory response). Parasite products may also affect lymphocytes to (**3**) make them become regulatory; (**4**) induce T or B cell tolerance by clonal exhaustion or by the induction of anergy; (**5**) cause polyclonal activation (many parasite products are mitogenic to T or B cells, and the high serum concentrations of non-specific IgM [and IgG] commonly found in parasitic infections probably result from this polyclonal stimulation – its continuation is believed to lead to impairment of B cell function, the progressive depletion of antigen-reactive B cells, and thus immunosuppression); (**6**) directly inhibit the activation of T or B cells (e.g. ES-62, a secreted product of a filarial parasites, is able to inhibit the proliferation of both T and B cells).

The IgE of worm infections can have severe effects on the host due to release of mast cell mediators. Anaphylactic shock may occur when a hydatid cyst ruptures. Asthma-like reactions occur in *T. canis* infections and in tropical pulmonary eosinophilia when filarial worms migrate through the lungs.

Autoantibodies, which probably arise as a result of polyclonal activation, have been detected against red blood cells, lymphocytes, and DNA (e.g. in trypanosomiasis and in malaria).

Antibodies against the parasite may cross-react with host tissues. For example, the chronic cardiomyopathy, enlarged esophagus, and megacolon that occur in Chagas' disease are thought to result from the autoimmune effects on nerve ganglia of antibody and cytotoxic T cells that cross-react with *T. cruzi*. Similarly *O. volvulus*, the cause of river blindness, possesses an antigen that cross-reacts with a protein in the retina.

Excessive production of cytokines may contribute to some of the manifestations of disease. Thus the fever, anemia, diarrhea, and pulmonary changes of acute malaria closely resemble the symptoms of endotoxemia and are probably caused by TNFα. The severe wasting of cattle with trypanosomiasis may also be mediated by TNFα.

A single parasite protein may produce multiple pathological effects, as seen with PfEMP1, coded by the *var* genes, and expressed on the surface of infected erythrocytes (Fig. 15.24).

Lastly, the non-specific immunosuppression that is so widespread probably explains why people with parasitic infections are especially susceptible to bacterial and viral infections (e.g. measles). It may also account for the association of Burkitt's lymphoma with malaria because malaria-infected individuals are less able to control infection with the Epstein–Barr virus that causes Burkitt's lymphoma.

Malarial pathology resulting from interactions with PfEMP1

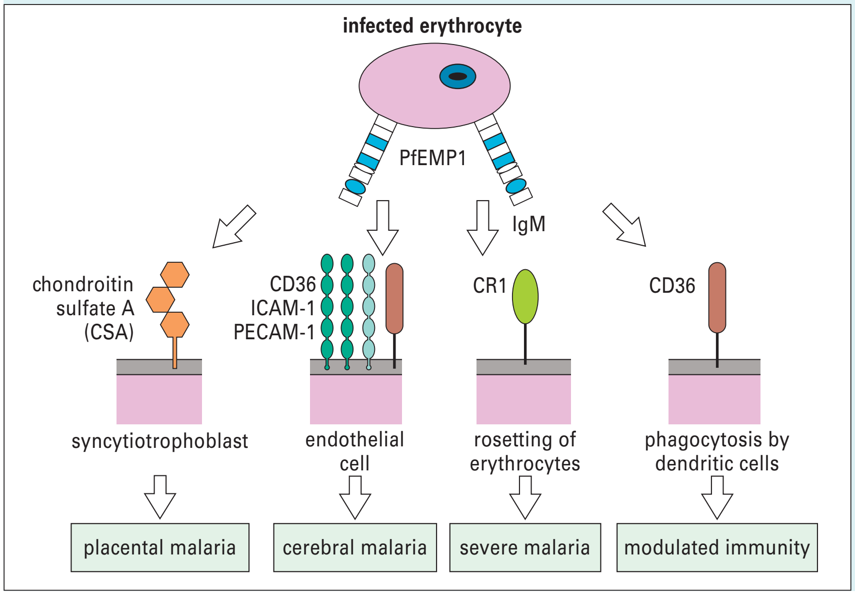

Fig. 15.24 Malaria-infected erythrocytes cause disease through many mechanisms involving the interaction between *P. falciparum* erythrocyte membrane protein-1 (PfEMP1) and diverse host receptors.

VACCINES AGAINST HUMAN PARASITES ARE NOT YET AVAILABLE

Some vaccines composed of attenuated living parasites have proved successful in veterinary practice. However, so far there is none in use against human parasites, though much effort has been directed towards the development of subunit vaccines against malarial parasites and schistosomes in particular. Some clinical trials of vaccines against malaria, based on combinations of putatively protective peptides, are in progress (see Chapter 18).

Parasite genome sequencing ventures and analysis of whole parasite proteomes are highlighting novel targets for both drug and vaccine design. These modern technologies are providing startling insights into how parasite and host immune systems interact. For example, the malaria protein PfEMP1 encoded by *var* genes, known to be expressed in the red blood cell stage within the human host and implicated in immune evasion, has been identified in the mosquito sporozoite stage, indicating that it may have several alternative functions (see Fig. 15.24).

FURTHER READING

Crabb BS, Cooke BM. Molecular approaches to malaria. Trends Parasitol 2004;20:547–614.

Maizels RM, Yazdanbakhsh M. Immune regulation by helminth parasites: cellular and molecular mechanisms. Nat Rev Immunol 2003;3:733–744.

McGuiness DH, Dehal PK, Pleass RJ. Pattern recognition molecules and innate immunity to parasites. Trends Parasitol 2003;19:312–319.

Scott P. Immunoparasitology. Immunol Rev 2004;201:5–8.

Sher A, Pearce E, Kaye P. Shaping the immune response to parasites: role of dendritic cells. Curr Opin Immunol 2003;15:421–429.

Websites

Various newsgroups on the www (though not all dedicated to immunology) can be accessed by exploring keywords

Several discussion groups operate through the bionet, for example http://www.bio.net/archives.html (a parasitology mail newsgroup) and http://www.parasitology.org.uk (British Society for Parasitology)

Critical thinking: Immunity to protozoa and helminths (see p. 497 for explanations)

1. In general, protozoa and helminths adopt different strategies for survival and for transmission to the subsequent host. How do they differ?
2. Many parasites have evolved to live in host cells. Consider the advantages and disadvantages to this mode of existence. Consider the different cell types and how parasites have to adapt this environment to their advantage. In particular *Toxoplasma gondii, Trypanosoma cruzi*, and *Leishmania* spp. have adapted to live in the macrophage and can escape destruction by lysosomal enzymes, but the way in which they do this differs. How have these adaptations helped parasite survival?
3. Extracellular parasites have evolved sophisticated mechanisms to avoid the immune response. Consider examples of how they do this.

This Page Intentionally Left Blank

CHAPTER

16

Primary Immunodeficiency

SUMMARY

- **Primary immunodeficiency diseases** result from intrinsic defects in immune system cells, complement components, and phagocytic cells.
- **Defects in B cell function** result in recurrent pyogenic infections. Defective antibody responses are due to failure of B cell function, as occurs in X-linked agammaglobulinemia, or failure of proper T cell signals to B cells, as occurs in hyper-IgM (HIgM) syndrome, common variable immunodeficiency (CVID), and transient hypogammaglobulinemia of infancy.
- **Poor T cell function** results in susceptibility to opportunistic infections. Defective cell-mediated immunity is due to failure of T cell function, seen for example in severe combined immunodeficiency (SCID), MHC class II deficiency, ataxia telangiectasia (AT), the Wiskott–Aldrich syndrome (WAS), and the DiGeorge anomaly.
- **Hereditary complement component defects** are found in a number of clinical syndromes, the most common of which is that of the C1 inhibitor, which results in hereditary angiedema (HAE). Hereditary complement deficiencies of the terminal complement components (C5, C6, C7, and C8) and the alternative pathway proteins (factor H, factor I, and properdin) lead to extraordinary susceptibility to infections with the two *Neisseria* species, *N. gonorrheae* and *N. meningitidis*.
- **Genetic defects of phagocytes** can result in overwhelming infection. Defects in the oxygen reduction pathway of phagocytes, such that the phagocytes cannot assemble NADPH oxidase and produce the hydrogen peroxide and oxygen radicals that kill bacteria, are the basis of chronic granulomatous disease (CGD). The resulting persistence of bacterial products in phagocytes leads to abscesses or granulomas, depending on the pathogen.
- **Leukocyte adhesion deficiency (LAD)** is associated with a persistent leukocytosis because phagocytic cells with defective integrin molecules cannot migrate through the vascular endothelium from the blood stream into the tissues.

PRIMARY IMMUNODEFICIENCY RESULTS FROM INTRINSIC DEFECTS IN IMMUNE SYSTEM CELLS

Immunodeficiency disease results from the absence or failure of normal function of one or more elements of the immune system:

- specific immunodeficiency diseases involve abnormalities of T or B cells (the cells of the adaptive immune system);
- non-specific immunodeficiency diseases involve abnormalities of elements such as complement or phagocytes, which act non-specifically in immunity.

Primary immunodeficiency diseases are due to intrinsic defects in cells of the immune system and are for the most part genetically determined.

Immunodeficiency diseases cause increased susceptibility to infection in patients.

- the infections encountered in patients who are immunodeficient fall, broadly speaking, into two categories: patients with defects in immunoglobulins, complement proteins, or phagocytes are very susceptible to recurrent infections with encapsulated bacteria such as *Haemophilus influenzae*, *Streptococcus pneumoniae*, and *Staphylococcus aureus* – these are called **pyogenic infections**, because the bacteria give rise to pus formation;
- patients with defects in cell-mediated immunity (i.e. in T cells) are susceptible to overwhelming, even lethal, infections with microorganisms that are ubiquitous in the environment and to which normal people rapidly develop resistance – for this reason, these are called **opportunistic infections**, and opportunistic microorganisms include yeast and common viruses such as chickenpox.

B CELL DEFICIENCIES

Defects in B cell function result in recurrent pyogenic infections

Patients with common defects in B cell function (Fig. 16.1) have recurrent pyogenic infections such as:

- pneumonia;
- otitis media; and
- sinusitis.

Primary B cell deficiencies
X-linked agammaglobulinemia (X-LA)
IgA deficiency
IgG subclass deficiency
immunodeficiency with increased IgM (HIgM)
common variable immunodeficiency (CVID)
transient hypogammaglobulinemia of infancy

Fig. 16.1 The range of B cell deficiencies varies from a delayed maturation of normal immunoglobulin production, through single isotype deficiencies to X-linked agammaglobulinemia, where affected male children have no B cells and no serum immunoglobulins.

If untreated, they develop severe obstructive lung disease (bronchiectasis) from recurrent pneumonia, which destroys the elasticity of the airways.

Early B cell maturation fails in X-LA

X-linked agammaglobulinemia (X-LA) is the model B cell deficiency and was the first immunodeficiency disease to be understood in detail, the underlying deficiency being discovered in 1952.

Affected males have few or no B cells in their blood or lymphoid tissue – consequently their lymph nodes are very small and their tonsils are absent. Their serum usually contains no IgA, IgM, IgD, or IgE, and only small amounts of IgG (less than 100 mg/dl).

For the first 6–12 months of life, affected males are protected from infection by the maternal IgG that crossed the placenta into the fetus. As this supply of IgG is exhausted, they develop recurrent pyogenic infections. If they are infused intravenously with large doses of gammaglobulin they remain healthy.

The X-LA gene lies on the long arm of the X chromosome (Fig. 16.2). This is the site of many other hereditary immunodeficiency diseases and the localization of these genes facilitates prenatal diagnosis.

The gene that is defective in X-LA has recently been identified as a B cell cytoplasmic tyrosine kinase (*btk*) belonging to the *src* oncogene family. Its role in B cell maturation is not yet understood, but it is obviously vital for the process. The bone marrow of males with X-LA contains normal numbers of pre-B cells, but as a result of mutations in the *btk* gene these cells cannot mature into B cells (Fig. 16.3).

Terminal differentiation of B cells fails in IgA and IgG subclass deficiency

IgA deficiency is the most common immunodeficiency. One in 700 Caucasians has the defect, but it is not found, or is found only rarely, in other ethnic groups.

People with IgA deficiency tend to develop immune complex disease (type III hypersensitivity).

The X-linked immunodeficiencies

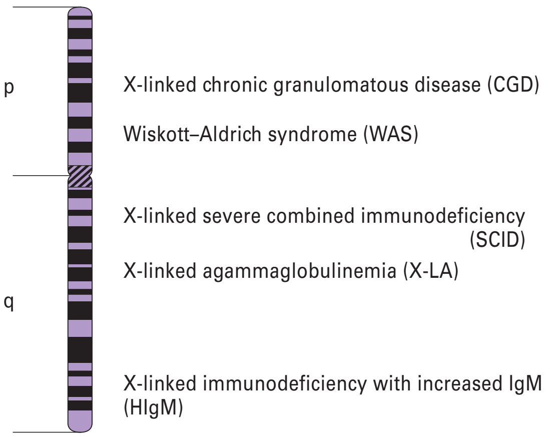
p
q
X-linked chronic granulomatous disease (CGD)
Wiskott–Aldrich syndrome (WAS)
X-linked severe combined immunodeficiency (SCID)
X-linked agammaglobulinemia (X-LA)
X-linked immunodeficiency with increased IgM (HIgM)

Fig. 16.2 The genes for many immunodeficiency diseases are located on the X chromosome. The genetic defects have been identified for all these diseases. (Adapted from Schwaber J, Rosen FS. X chromosome linked immunodeficiency. Immunodefic Rev 1990;2:233–251)

About 20% of IgA-deficient individuals also lack IgG2 and IgG4, and are therefore susceptible to pyogenic infections. In humans, most antibodies to the capsular polysaccharides of pyogenic bacteria are in the IgG2 subclass. A deficiency in IgG2 alone therefore also results in recurrent pyogenic infections. Individuals with deficiency of only IgG2 are also susceptible to recurrent infections.

These class and subclass deficiencies result from failure in the terminal differentiation of B cells (see Fig. 16.3).

Isotype switching does not occur in HIgM

A peculiar immunodeficiency – **immunodeficiency with increased IgM (hyper-IgM, HIgM)** – results in a situation where individuals are IgG and IgA deficient, but synthesize large amounts (more than 200 mg/dl) of polyclonal IgM.

People with HIgM are susceptible to pyogenic infections and should be treated with intravenous gammaglobulin. They tend to form IgM autoantibodies to neutrophils, platelets, and other elements of the blood, as well as to tissue antigens, thereby adding the complexities of autoimmune disease to the immunodeficiency. The tissues, particularly of the gastrointestinal tract, become infiltrated with IgM-producing cells (Fig. 16.4).

In 70% of cases, HIgM is inherited as an X-linked recessive condition that results from mutations in the **CD40 ligand**, whose gene maps to precisely the same location on the long arm of the X chromosome as HIgM.

Q. What is the function of CD40 and the CD40 ligand (CD154)?

A. Interaction between the CD40 molecule on the B cell surface and the CD40 ligand on activated T cells is a potent co-stimulatory signal required for class switching and affinity maturation (see Fig. 8.8).

B cell maturation in X-linked immunodeficiencies

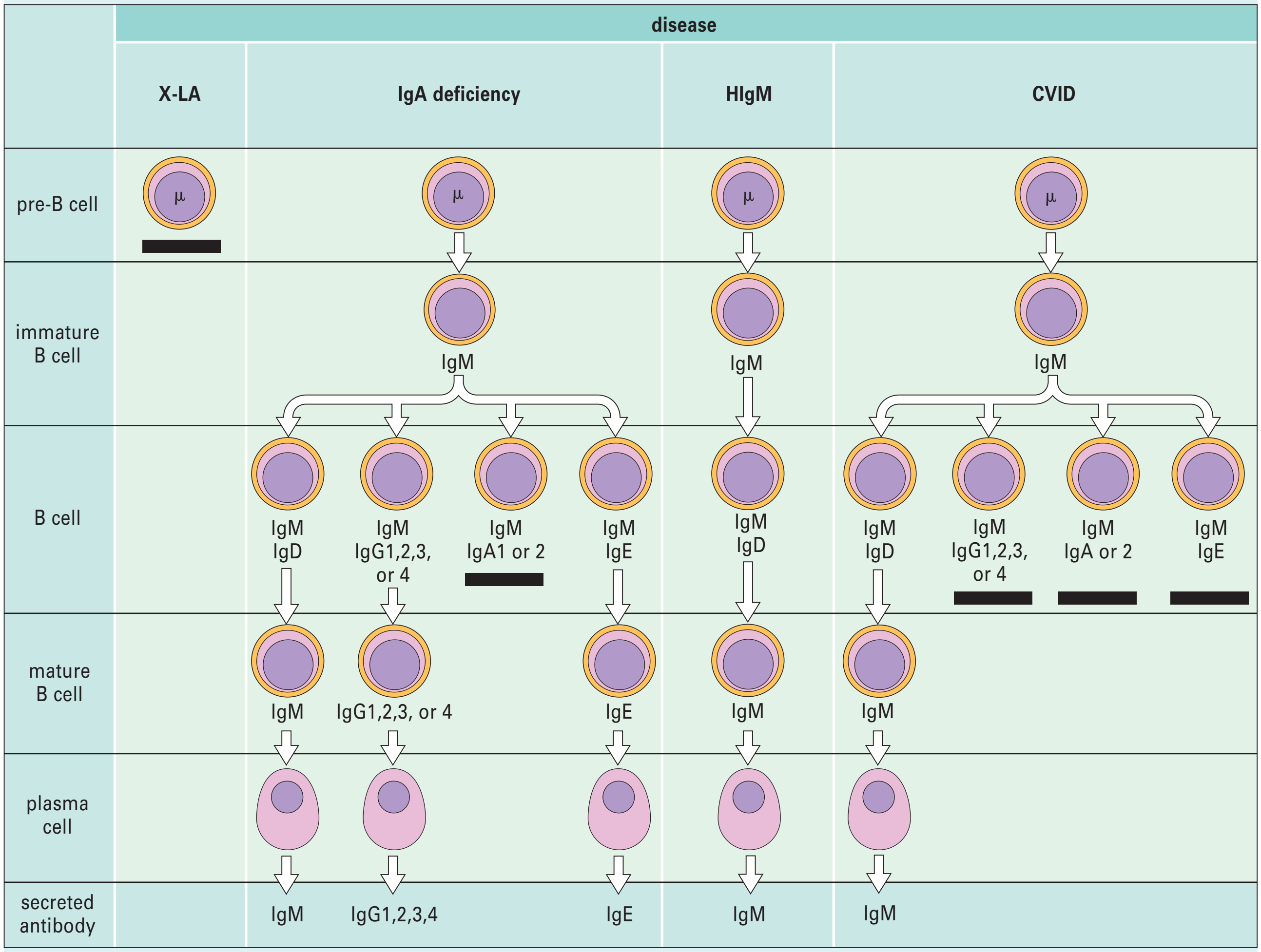
disease
X-LA
IgA deficiency
HIgM
CVID
pre-B cell
immature B cell
B cell
mature B cell
plasma cell
secreted antibody
μ
μ
μ
μ
IgM
IgM
IgM
IgM IgD
IgM IgG1,2,3, or 4
IgM IgA1 or 2
IgM IgE
IgM IgD
IgM IgD
IgM IgG1,2,3, or 4
IgM IgA or 2
IgM IgE
IgM
IgG1,2,3, or 4
IgE
IgM
IgM
IgM
IgG1,2,3,4
IgE
IgM
IgM

Fig. 16.3 In X-LA, affected male infants have no B cells and no serum immunoglobulins, except for small amounts of maternal IgG. In IgA deficiency, IgA-bearing B cells, and in some cases IgG2- and IgG4-bearing B cells, are unable to differentiate into plasma cells. People with HIgM lack IgG and IgA. In CVID, B cells of most isotypes are unable to differentiate into plasma cells. Black bars denote points of inhibition.

Gall bladder from a patient with HIgM

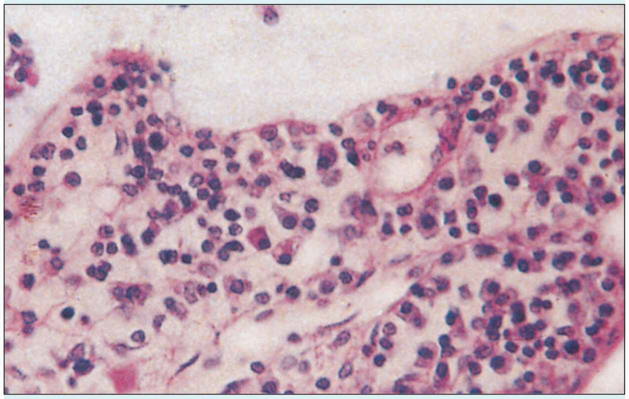

Fig. 16.4 The submucosa is filled with cells with pink-staining cytoplasm and eccentric nuclei. The cells are synthesizing and secreting IgM.

In HIgM the B cells cannot make the switch from IgM to IgG, IgA, and IgE synthesis that normally occurs in B cell maturation.

HIgM may also be transmitted as an autosomal recessive trait, affecting females as well as males. Autosomal recessive HIgM is due to a genetic defect in a B cell enzyme called **activation-induced cytidine deaminase (AID)**. This enzyme, activated in stimulated B cells, converts cytidine to uridine in single-stranded DNA. The uridine is further degraded to uracil and this causes breaks in the DNA strand, allowing downstream recombination in the heavy chain genes – or class switch recombination.

T cell signaling to B cells is defective in CVID

Individuals with **common variable immunodeficiency (CVID)** have acquired agammaglobulinemia in the second or third decade of life, or later.

Both males and females are equally affected, and the cause is generally not known, but may follow infection with viruses such as Epstein–Barr virus (EBV).

Patients with CVID, like males with X-LA, are very susceptible to pyogenic organisms and to the intestinal protozoon, *Giardia lamblia* (Fig. 16.5), which causes severe diarrhea.

Most patients (80%) with CVID have B cells that do not function properly and are immature. The B cells are not defective; instead, they fail to receive proper signals from the T cells. The T cell defects have not been well defined in CVID.

Patients with CVID should be treated with intravenous gammaglobulin because it provides protection against recurrent pyogenic infections. Many patients develop autoimmune diseases, most prominently pernicious anemia, but the reason for this is not known.

CVID is not hereditary, but is commonly associated with the MHC haplotypes HLA-B8 and HLA-DR3.

Q. In what other immunodeficiency do patients often develop autoantibodies to elements of the blood?
A. In HIgM (see above).

IgG production is delayed in transient hypogammaglobulinemia of infancy

Infants are protected initially by their mother's IgG. The maternal IgG is catabolized, with a half-life of approximately 30 days. By 3 months of age, normal infants begin to synthesize their own IgG, though formation of antibody to bacterial capsular polysaccharides does not commence in earnest until the second year of life. In some infants, the onset of normal IgG synthesis can be delayed for as long as 36 months and, until then, such infants are susceptible to pyogenic infections. The B cells of these infants are normal, but they appear to lack help from $CD4^+$ T cells in synthesizing antibodies.

T CELL DEFICIENCIES

Poor T cell function results in susceptibility to opportunistic infections

The major T cell deficiencies are shown in Fig. 16.6.

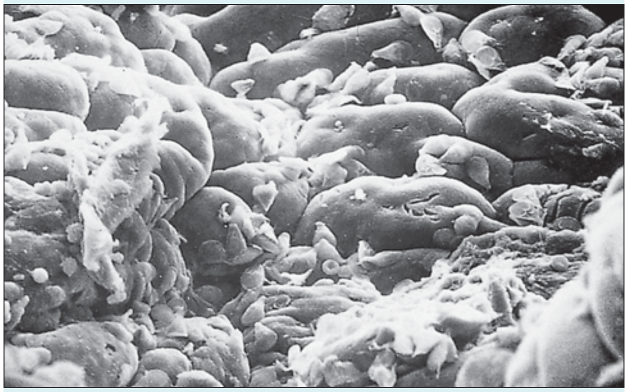

Giardia lamblia

Fig. 16.5 Numerous *Giardia lamblia* parasites can be seen swarming over the mucosa of the jejunum of a patient with CVID.

Primary T cell deficiencies
severe combined immunodeficiency (SCID)
adenosine deaminase deficiency
purine nucleoside phosphorylase deficiency
MHC class II deficiency
DiGeorge anomaly
hereditary ataxia telangiectasia (AT)
Wiskott–Aldrich syndrome (WAS)

Fig. 16.6 There is a wide range of causes for T cell deficiencies, ranging from absence of lymphocytes, to enzyme deficiency, through to MHC deficiency. All affect the ability of T cells to function, which leads to combined T and B cell deficiency.

People with no T cells or poor T cell function are susceptible to opportunistic infections.

Because B cell function in humans is largely T cell dependent, T cell deficiency also results in humoral immunodeficiency, that is, it leads to a combined deficiency of both humoral and cell-mediated immunity.

In SCID there is lymphocyte deficiency and the thymus does not develop

The most profound hereditary deficiency of cell-mediated immunity occurs in infants with **severe combined immunodeficiency (SCID)** who develop recurrent infections early in life (in contrast to X-LA). They:

- have prolonged diarrhea due to rotavirus or bacterial infection of the gastrointestinal tract; and
- develop pneumonia, usually due to the protozoon, *Pneumocystis carinii*.

The common yeast organism *Candida albicans* grows luxuriantly in the mouth or on the skin of patients with SCID (Fig. 16.7).

If patients with SCID are vaccinated with live organisms, such as poliovirus or bacille Calmette–Guérin (BCG) (used for immunization against tuberculosis), they die from progressive infection with these ordinarily benign organisms.

SCID is incompatible with life and affected infants usually die within the first 2 years unless they are rescued with transplants of bone marrow. In this case they become **lymphocyte chimeras** and can survive and live normally.

Infants with SCID have very few lymphocytes in their blood (fewer than 3000/ml). Their lymphoid tissue also contains few or no lymphocytes. The thymus has a fetal appearance (Fig. 16.8), containing the endodermal stromal cells derived embryonically from the third and fourth pharyngeal pouch. Lymphoid stem cells, which normally populate the thymus by 6 weeks of human gestation (see Chapter 2), fail to appear and the thymus does not become a lymphoid organ.

***Candida albicans* in the mouth of a patient with SCID**

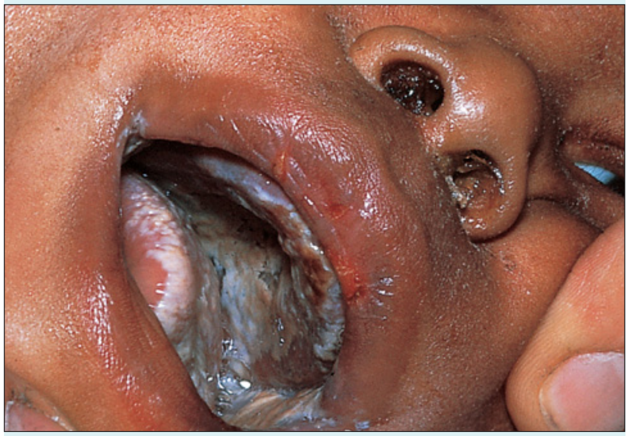

Fig. 16.7 *C. albicans* grows luxuriantly in the mouth and on the skin of patients with SCID.

Thymus of SCID

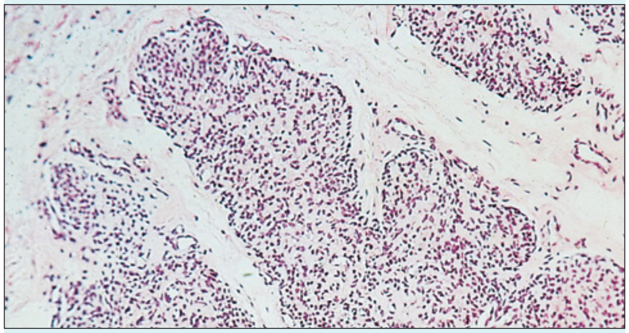

Fig. 16.8 Note that the thymic stroma has not been invaded by lymphoid cells and no Hassall's corpuscles are seen. The gland has a fetal appearance.

Possible role of adenosine deaminase and purine nucleoside phosphorylase deficiency in SCID

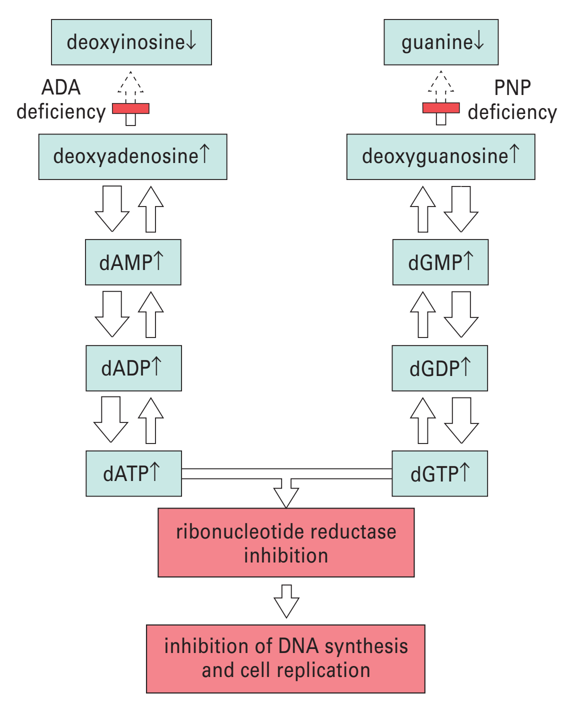

Fig. 16.9 It is thought that deficiencies of ADA and PNP lead to accumulations of dATP and dGTP, respectively. Both of these metabolites are powerful inhibitors of ribonucleotide reductase, which is an essential enzyme for DNA synthesis.

SCID is more common in male than female infants (3:1) because over 50% of cases are caused by a gene defect on the X chromosome. The defective gene encodes the γ chain of the interleukin-2 (IL-2) receptor. This γ chain also forms part of the receptors for IL-4, IL-7, IL-11, and IL-15.

Q. Which of these cytokines is most critically important in early T cell development?

A. Of these, the binding of IL-7 to the IL-7 receptor is most important for T cell maturation. Thus, the lymphoid stem cells are incapable of receiving a number of signals for growth and maturation (see Fig. 2.29).

The remaining cases of SCID are due to recessive genes on other chromosomes. Of these, half have a genetic deficiency of **adenosine deaminase (ADA)** or **purine nucleoside phosphorylase (PNP)**. Deficiency of these purine degradation enzymes results in the accumulation of metabolites that are toxic to lymphoid stem cells, namely dATP and dGTP (Fig. 16.9). These metabolites inhibit the enzyme ribonucleotide reductase, which is required for DNA synthesis and, therefore, for cell replication.

ADA and PNP are found in all mammalian cells, so why should these defects affect only lymphocytes? The explanation appears to lie in the relative deficiency of 5′-nucleotidase in lymphoid cells; in other cells, this enzyme compensates for defective ADA or PNP by preventing dAMP and dGMP accumulation.

An autosomal recessive form of SCID results from a mutation in either of the genes encoding **RAG-1 or RAG-2** .

Q. What are the functions of the genes encoding RAG-1 or RAG-2, and why would loss of function lead to combined immunodeficiency?

A. The two recombinase activation genes, *RAG-1* and *RAG-2*, are absolutely required for cleaving double-stranded DNA before recombination of DNA to form the immunoglobulin genes and the genes encoding the T cell receptor. If these gene rearrangements do not occur, B and T cells do not develop (see Fig. 3.29).

The optimal treatment for SCID is a bone marrow transplant from a completely histocompatible donor, usually a normal sibling. About 70% of patients do not have a histocompatible sibling, in which case parental marrow, which would have one haplotype identical, has been transplanted successfully.

Q. Why is it necessary to match the bone marrow graft in a patient with SCID when the immune system of the SCID recipient is unable to reject the donor bone marrow cells?

A. The donor bone marrow cells are fully immunocompetent and would generate a reaction against the recipient's cells. This is called **graft versus host disease** (see Fig. 21.8).

Recently a retroviral vector, into which the *ADA* gene had been inserted, has been used to transfect the lymphocytes of children who are ADA deficient. This was the first example of successful '**gene therapy**'. Males with X-linked SCID have since been successfully treated by transducing their bone marrow cells with the gene of the common γ chain of the IL-2 receptor incorporated into a viral vector.

TH cell deficiency results from MHC class II deficiency

The failure to express class II MHC molecules on antigen-presenting cells (macrophages and B cells) is inherited as an autosomal recessive characteristic, which is not linked to the MHC locus on the short arm of chromosome 6.

Affected infants have recurrent infections, particularly of the gastrointestinal tract.

Because the development of $CD4^+$ helper T cells (TH) depends on positive selection by MHC class II molecules in the thymus (see Chapter 2), MHC class II molecule-deficient infants have a deficiency of $CD4^+$ T cells. This lack of TH cells leads to a deficiency in antibodies as well. The MHC class II deficiency results from defects in promoter proteins that bind to the 5′ untranslated region of the class II genes.

The DiGeorge anomaly arises from a defect in thymus embryogenesis

The thymic epithelium is derived from the third and fourth pharyngeal pouches by the sixth week of human gestation. Subsequently the endodermal anlage is invaded by lymphoid stem cells, which undergo development into T cells. The parathyroid glands are also derived from the same embryonic origin.

A congenital defect in the organs derived from the third and fourth pharyngeal pouches results in the **DiGeorge anomaly**. The T cell deficiency is variable, depending on how badly the thymus is affected. Affected infants have distinctive facial features (Fig. 16.10) in that their eyes are widely separated (hypertelorism), the ears are low set, and the philtrum of the upper lip is shortened. They also have congenital malformations of the heart or aortic arch and neonatal tetany from the hypoplasia or aplasia of the parathyroid glands. These infants have **partial monosomy of 22q11-pter** or **10p**.

DiGeorge anomaly

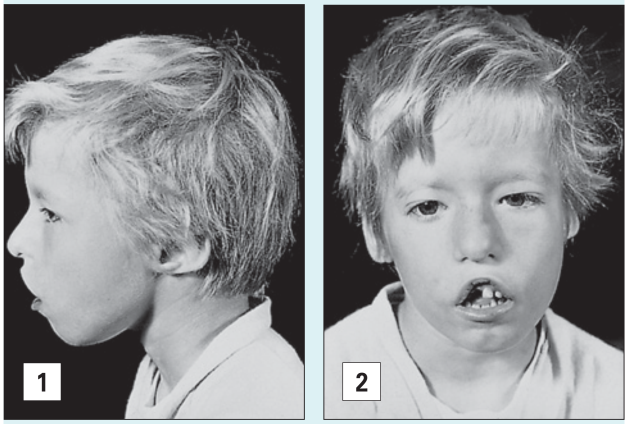

Fig. 16.10 Note the wide-set eyes, low-set ears, and shortened philtrum of upper lip. Congenital malformations of the cardiovascular system may also occur.

XLP follows infection with EBV

X-linked proliferative syndrome (XLP) results from a failure to control the normal proliferation of cytotoxic T cells following an infection with Epstein–Barr virus (EBV), which causes infectious mononucleosis.

Affected males appear normal until they encounter EBV, when they develop either:

- fatal infectious mononucleosis;
- complete destruction of their B cells so that agammaglobulinemia ensues;
- a fatal lymphoid malignancy; or
- aplastic anemia.

The defective gene on the X chromosome encodes an adapter protein of T and B cells called **SAP** or the **SLAM-associated protein**. SLAM is expressed on the surface of T and B cells. Its intracellular tail interacts with the adapter protein, SAP. By a mechanism that is not understood, SAP controls the limitless proliferation of cytotoxic T cells. A genetic defect in SAP results in the destruction of lymphoid and other hematopoietic tissue by uncontrolled proliferation of cytotoxic T cells and maternal killer cells.

Chromosomal breaks occur in TCR and immunoglobulin genes in AT

Hereditary ataxia telangiectasia (AT) is inherited as an autosomal recessive trait.

Affected infants with AT develop a wobbly gait (ataxia) at about 18 months. Dilated capillaries (telangiectasia) appear in the eyes and on the skin by 6 years of age.

AT is accompanied by a variable T cell deficiency. About 70% of patients with AT are also IgA deficient and some also have IgG2 and IgG4 deficiency.

Q. In what other ways are IgA, IgG2, and IgG4 related (as opposed to IgM, IgG1, and IgG3)?

A. The IgA1, IgA2, IgG2, and IgG4 heavy chain genes all lie further downstream of the recombined VDJ gene than IgM, IgG1, and IgG3 (see Fig. 8.19). This may account for the selective deficiency in making the class switch to IgA, IgG2, and IgG4. Class switching involves the production and resolution of double-stranded DNA breaks.

The number and function of circulating T cells are greatly diminished, so cell-mediated function is depressed.

Patients develop severe sinus and lung infections. Their cells exhibit chromosomal breaks, usually in chromosome 7 and chromosome 14, at the sites of the T cell receptor (TCR) genes and the genes encoding the heavy chains of immunoglobulins.

The cells of patients with AT, as well as those from patients with AT in vitro, are very susceptible to ionizing irradiation because the defective gene in AT encodes a protein involved in the repair of double-strand breaks in DNA.

T cell defects and abnormal immunoglobulin levels occur in Wiskott–Aldrich syndrome

Wiskott–Aldrich syndrome (WAS) is an X-linked immunodeficiency disease.

Affected males with WAS:

- have small and profoundly abnormal platelets, which are also few in number (thrombocytopenia);
- develop severe eczema as well as pyogenic and opportunistic infections;
- have increased amounts of serum IgA and IgE, normal levels of IgG, and decreased amounts of IgM;
- have T cells with defective function.

The malfunction of cell-mediated immunity gets progressively worse. The T cells have a uniquely abnormal appearance, as shown by scanning electron microscopy, reflecting a cytoskeletal defect. They have fewer microvilli on the cell surface than normal T cells.

During collaboration of T and B cells in antibody formation, the cytoskeleton of the T cells reorientates itself or becomes polarized towards the B cells. This fails to occur in WAS, resulting in faulty collaboration between immune cells.

GENETIC DEFICIENCIES OF COMPLEMENT PROTEINS

The proteins of the complement system and their interactions with the immune system are discussed in Chapter 4. Genetic deficiencies of almost all the complement proteins have been found in humans (Fig. 16.11) and these deficiencies reveal much about the normal function of the complement system (see Figs 4.16 and 4.17).

Immune complex clearance, inflammation, phagocytosis, and bacteriolysis can be affected by complement deficiencies

Deficiencies of the classical pathway components, C1q, C1r, C1s, C4, or C2, result in a propensity to develop immune complex diseases such as systemic lupus erythematosus.

Q. Why should these deficiencies result in immune complex disease?

A. The classical complement pathway is required for the dissolution of immune complexes by covalent binding of C4b and C3b to components of the complex. It is also required for the transport of complexes on erythrocytes in humans (see Fig. 4.6).

Deficiencies of C3, factor H, or factor I result in increased susceptibility to pyogenic infections – this correlates with the important role of C3 in the opsonization of pyogenic bacteria.

Deficiencies of the terminal components, C5, C6, C7, and C8, and of the alternative pathway components, factor D and properdin, result in remarkable susceptibility to infection with the two pathogenic species of the *Neisseria*

Genetic deficiencies of human complement

group	type	deficiency	hereditary		
			AR	AD	XL
I	immune complex disease	C1q	●		
		C1s, or C1r + C1s	●		
		C2	●		
		C4	●		
II	angiedema	C1 inhibitor		●	
III	recurrent pyogenic infections	C3	●		
		factor H	●		
		factor I	●		
IV	recurrent *Neisseria* infections	C5	●		
		C6	●		
		C7	●		
		C8	●		
		properdin			●
		factor D	●		
V	asymptomatic	C9	●		

Fig. 16.11 Genetic deficiencies of human complement. (AR, phenotypically autosomal recessive; AD, autosomal dominant; XL, X-linked recessive)

genus: *N. gonorrhoeae* and *N. meningitidis*. This clearly demonstrates the importance of the alternative pathway and the macromolecular attack complex in the bacteriolysis of this genus of bacteria.

All of these genetic complement component deficiencies are inherited as autosomal recessive traits, except:

- properdin deficiency, which is inherited as an X-linked recessive; and
- C1 inhibitor deficiency, which is inherited as an autosomal dominant.

HAE results from C1 inhibitor deficiency

Clinically, the most important deficiency of the complement system is that of **C1 inhibitor**. This molecule is responsible for dissociation of activated C1, by binding to $\overline{C1r_2C1s_2}$. The deficiency results in the well-known disease, **hereditary angioneurotic edema (HAE)** (Fig. 16.12).

HAE is inherited as an autosomal dominant trait. Patients have recurrent episodes of circumscribed swelling of various parts of the body (angiedema):

- when the edema involves the intestine, excruciating abdominal pains and cramps result, with severe vomiting;
- when the edema involves the upper airway, the patients may choke to death from respiratory obstruction – angiedema of the upper airway therefore presents a medical emergency, which requires rapid action to restore normal breathing.

C1 inhibitor inhibits not only the classical pathway of complement, but also joint elements of the kinin, plasmin, and clotting systems.

The edema is mediated by two peptides generated by uninhibited activation of the complement and surface contact systems:

- a peptide derived from the activation of C2, called C2 kinin; and
- bradykinin derived from the activation of the contact system (Fig. 16.13).

Hereditary angioneurotic edema

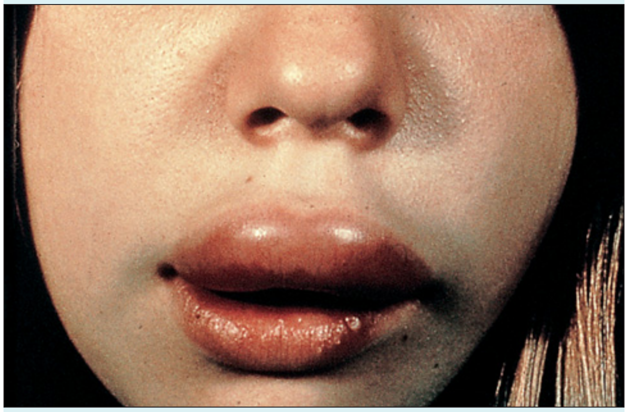

Fig. 16.12 This clinical photograph shows the transient localized swelling that occurs in this condition.

The effect of these peptides is on the postcapillary venule, where they cause endothelial cells to retract, forming gaps that allow leakage of plasma (see Chapter 2).

There are two genetically determined forms of HAE.

- in type I, the C1 inhibitor gene is defective and no transcripts are formed;
- in type II, there are point mutations in the C1 inhibitor gene resulting in the synthesis of defective molecules.

The distinction between type I and type II is important because the diagnosis of type II disease cannot be made by quantitative measurement of serum C1 inhibitor alone. Simultaneous measurements of **C4** must also be done. C4 is always decreased in the serum of patients with HAE because of its destruction by uninhibited activated C1.

C1 inhibitor deficiency may be acquired later in life. In some cases an autoantibody to C1 inhibitor is found. In others, there is a monoclonal B cell proliferation, such as occurs in chronic lymphocytic leukemia, multiple myeloma, or B cell lymphoma. Such patients make an anti-idiotype to their over-produced immunoglobulin; the idiotype–anti-idiotype interaction, for unknown reasons,

Pathogenesis of hereditary angioneurotic edema

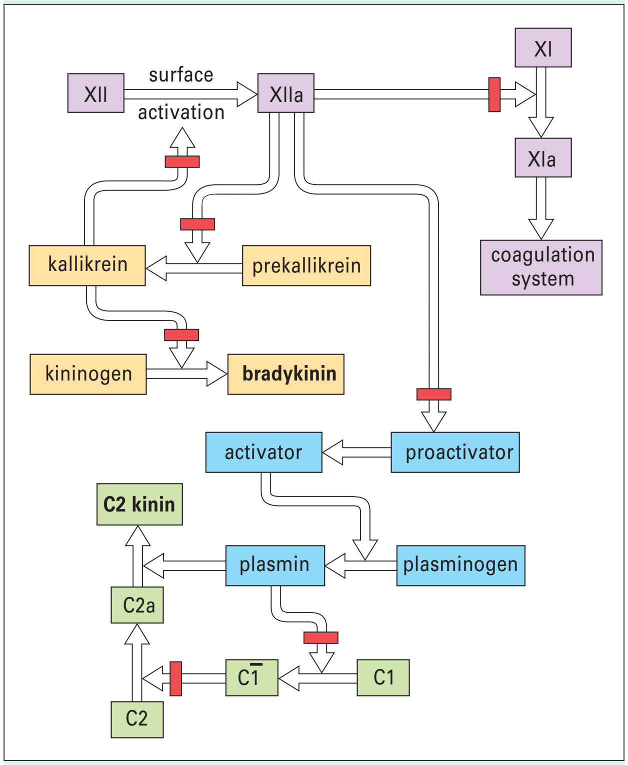

Fig. 16.13 C1 inhibitor is involved in inactivation of elements of the clotting, kinin, plasmin and complement systems, which may be activated following the surface-dependent activation of factor XII (Hageman factor). The points at which C1 inhibitor acts are shown in red. Uncontrolled activation of these pathways results in the formation of bradykinin and C2 kinin, which induce edema formation.

results in consumption of C1, C4, and C2, and C1 inhibitor without formation of an effective C3 convertase (which would cause C3 deposition and removal of the complement complex).

Purified C1 inhibitor is available for treatment of acute attacks of angiedema. Intravenous administration of C1 inhibitor brings about prompt resolution of laryngeal edema or gastrointestinal symptoms. It is not approved for use in the United States.

GENETIC DEFECTS OF PHAGOCYTES CAN RESULT IN OVERWHELMING INFECTION

Phagocytic cells – polymorphonuclear leukocytes and cells of the monocyte/macrophage lineage – are important in host defense against pyogenic bacteria and other intracellular microorganisms.

A severe deficiency of polymorphonuclear leukocytes (**neutropenia**) can result in overwhelming bacterial infection.

Two genetic defects of phagocytes are clinically important in that they result in susceptibility to severe infections and are often fatal:

- **chronic granulomatous disease (CGD)**; and
- **leukocyte adhesion deficiency (LAD)**.

CGD results from a defect in the oxygen reduction pathway

Patients with CGD have defective **NADPH oxidase**, which catalyzes the reduction of O_2 to $^{\bullet}O_2^-$ by the reaction: $NADPH + 2O_2 \rightarrow NADP^+ + 2^{\bullet}O_2^- + H^+$. They are therefore incapable of forming superoxide anions (ΣO_2^-) and hydrogen peroxide in their phagocytes following the ingestion of microorganisms.

Q. What would be the consequence of a failure of superoxide generation?

A. Phagocytes cannot readily kill ingested bacteria or fungi, particularly catalase-producing organisms (see Fig. 9.30 and Chapter 14).

As a result, microorganisms remain alive in phagocytes of patients with CGD. This gives rise to a cell-mediated response to persistent intracellular microbial antigens, and granulomas form.

Children with CGD develop pneumonia, infections in the lymph nodes (lymphadenitis), and abscesses in the skin, liver, and other viscera.

The diagnosis of CGD is made by the inability of phagocytes to reduce **nitroblue tetrazolium (NBT)** dye after a phagocytic stimulus. NBT is a pale, clear, yellow dye taken up by phagocytes when they are ingesting a particle. When NBT accepts H and is reduced as a result of NADPH oxidation it forms a deep purple precipitate inside the phagocytes; precipitation does not occur in the phagocytes of patients with CGD (Fig. 16.14).

The NADPH oxidase reaction is complicated and the enzyme complex has many subunits. In resting phagocytes the membrane contains a phagocyte-specific cytochrome, cytochrome b_{558}. This cytochrome is composed of two chains:

- one of 91 kDa, encoded by a gene on the short arm of the X chromosome; and
- one of 22 kDa, encoded by a gene on chromosome 16.

When phagocytosis occurs, several proteins from the cytosol become phosphorylated, move to the membrane, and bind to cytochrome b_{558}. The complex formed acts as an enzyme, NADPH oxidase, catalyzing the NADPH oxidation reaction and thereby activating oxygen radical production (Fig. 16.15).

Nitroblue tetrazolium test

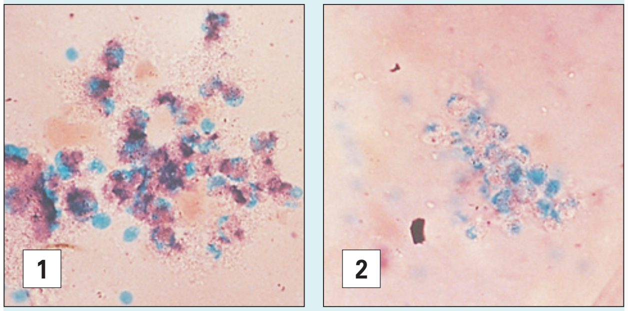

Fig. 16.14 In normal polymorphs and monocytes, reactive oxygen intermediates (ROIs) are activated by phagocytosis, and yellow NBT is converted to purple-blue formazan (**1**). Patients with CGD cannot form ROIs and so the dye stays yellow (**2**). (Courtesy of Professor AR Hayward)

The most common form of CGD is X-linked and involves a defect in the 91-kDa chain of cytochrome b_{558}.

Three types of CGD are autosomal recessive and result from defects in the 22-kDa chain of the cytochrome b_{558}, or from defects in one or other of two proteins, called $p47^{phox}$ or $p67^{phox}$.

LAD is due to integrin gene defects

The receptor in the phagocyte membrane that binds to C3bi on opsonized microorganisms is critical for the ingestion of bacteria by phagocytes. This receptor – an integrin called **complement receptor 3 (CR3)** – is deficient in patients with LAD, who consequently develop severe bacterial infections, particularly of the mouth and gastrointestinal tract.

CR3 is composed of two polypeptide chains:

- an α chain of 165 kDa (CD11b); and
- a β chain of 95 kDa (CD18).

In LAD, there is a genetic defect of the β chain, encoded by a gene on chromosome 21.

Two other integrin proteins share the same β chain as CR3 – namely lymphocyte functional antigen (LFA-1) and p150,95 (see Chapter 3). Although they have unique α chains (CD11a and CD11c, respectively), these proteins are also defective in LAD.

LFA-1 is important in cell adhesion and interacts with intercellular adhesion molecule-1 (ICAM-1) on endothelial cell surfaces and other cell membranes. Because of the defect in LFA-1, phagocytes from patients with LAD cannot adhere to vascular endothelium and cannot therefore migrate out of blood vessels into areas of

NADPH oxidase and its components

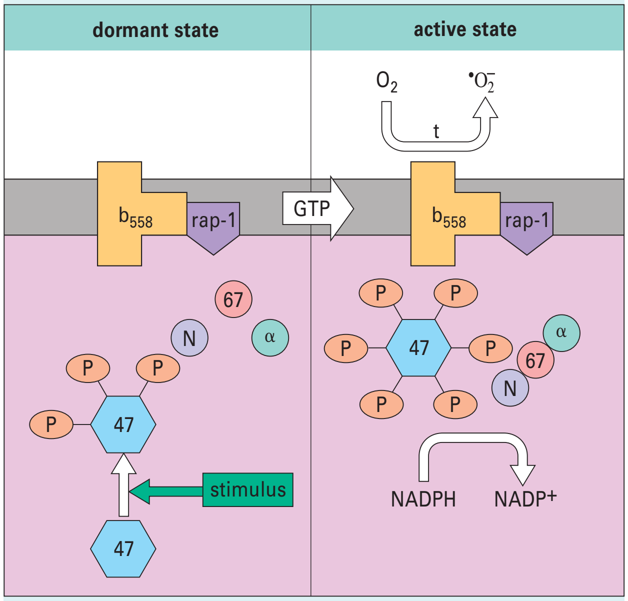

Fig. 16.15 Prevailing knowledge of the NADPH oxidase suggests that in its dormant state some of its component parts are in the membrane (cytochrome b$_{558}$ and possibly rap-1) while others are in the cytosol (p47^{phox}, p67^{phox}, the NADPH-binding component, N, and a putative fourth component, α). The initiation of phagocytosis provides a stimulus for the cytosolic components to associate and move to the membrane, an event possibly mediated by phosphorylation (P) of p47^{phox}. Once the cytosol components are associated with the membrane components, the oxidase becomes catalytically active and p47^{phox} is phosphorylated further. Activation of the cytochrome b$_{558}$ catalyzes the acceptance of an electron by O$_2$ to form the superoxide anion $^{\bullet}O_2^-$. In the different forms of CGD, there are defects in the genes for different components of the oxidase. (phox is an abbreviation for phagocytic oxidase. t, electron) (Adapted from Smith RM, Curnutte JT. Molecular basis of chronic granulomatous disease. Blood 1997;77:673–686)

infection. As a result patients with LAD cannot form pus efficiently and this allows the rapid spread of bacterial invaders.

LAD type 2 results from a failure of normal synthesis of selectin ligands

When leukocytes in the circulation enter an area of inflammation their speed of movement is greatly retarded by the interaction of **selectins** (see Fig. 6.9), which are expressed on the surface of the leukocytes, with ligands that are expressed on the surface of the endothelium in areas of inflammation. The leukocytes start to roll on the endothelial surface before the interaction of the leukocyte integrins with adhesion molecules such as ICAM-1. The ligands with which the selectins interact are glycoproteins that contain fucosylated sugars, such as the blood group sialyl Lewis^x. A genetic defect in the conversion of mannose to fucose results in failure of normal synthesis in these selectin ligands. Consequently the leukocytes of such patients cannot roll on the endothelium. This causes a second form of LAD, called **LAD type 2**.

Macrophage microbicidal activity is impaired by defects in IFNγ signaling

The destruction of intracellular microorganisms that flourish in macrophages depends on the activation of microbicidal activity in macrophages by interferon-γ (IFNγ).

When microorganisms are taken up by macrophages, the macrophages secrete IL-12, which then binds to the IL-12 receptor on T cells and induces secretion of IFNγ.

Children with genetic defects in the genes encoding IL-12, the IL-12 receptor, or the IFNγ receptor sustain recurrent infection with non-pathogenic mycobacteria and, to a lesser extent, with salmonella. These various defects are inherited as autosomal recessive traits and can be fatal unless treated with IFNγ.

FURTHER READING

Buckley RH. Primary immunodeficiency diseases due to defects in lymphocytes. N Engl J Med 2000;343:313.

Conley ME. Molecular approaches to analysis of X-linked immunodeficiencies. Ann Rev Immunol 1992;10:215.

Curnutte JT, Orkin SH, Dinauer MC. Genetic disorders of phagocyte function. In: Stamatoyannopoulos G, Nienhuis AW, Majerus PW, Varmus H, eds. The molecular basis of blood diseases. Philadelphia: PA Saunders; 1994:443.

Lekstrom-Himes JA, Gallin JI. Immunodeficiency diseases caused by defects in phagocytes. N Engl J Med 2000;343:1703.

Rosen FS, Cooper MD, Wedgwood RJP. The primary immunodeficiencies. N Engl J Med 1995;333:43.

Rosen FS, Seligman M, eds. Immunodeficiencies. Switzerland: Harwood Academic Publishers; 1993.

Snapper SB, Rosen FS. The Wiskott–Aldrich syndrome protein (WASP): roles in signaling and cytoskeletal organization. Annu Rev Immunol 1999;17:905–929.

Von Andrian UH, Berger EM, Chambers JD, et al. In vivo behaviour of neutrophils from two patients with distinct inherited leukocyte adhesion deficiency syndromes. J Clin Invest 1993;91:2893.

Critical thinking: Hyper-IgM immunodeficiency (see p. 497 for explanations)

A 3-year-old girl was brought to the emergency room because of fever and rapid respiration. She had had pneumonia once before at age 25 months. She had also had otitis media on ten different occasions, each time successfully treated with antibiotics. A chest radiograph resulted in the diagnosis of left lower lobe pneumonia. Blood and sputum cultures contained *Streptococcus pneumoniae*. The white blood count was 13 500/mcl of which 81% were neutrophils and 14% lymphocytes. Serum IgM was 470 mg/dl, IgG 40 mg/dl, and IgA and IgE were undetectable. Antibody to tetanus toxoid was undetectable. Blood typing was A positive, anti-B 1:320.

1. What clinical and laboratory tests lead to the suspicion that this child has hyper-IgM (HIgM) immunodeficiency and how do you conclude that it is not due to a mutation in CD40 ligand?
2. What further laboratory testing should be done to establish the diagnosis?
3. How do you explain that this child had no response to tetanus immunization and yet has a high titer for her age of antibody to blood group substance B?
4. What treatment would you recommend to the parents for this child and what would you say is her prognosis?
5. If the mother of this baby were to be tested for random X-chromosome inactivation, what results would you expect?
6. What would you tell the parents about this child's prognosis?

This Page Intentionally Left Blank

CHAPTER 17

AIDS and Secondary Immunodeficiency

SUMMARY

- **Some drugs selectively alter immune function.** Immunomodulatory drugs can severely depress immune functions. Steroids affect cell traffic, induce leukocytopenia, and inhibit cytokine synthesis. Cyclophosphamide, azathioprine, and mycophenolate mofetil act directly on DNA or its synthesis.

- **Nutrient deficiencies are generally associated with impaired immune responses.** Malnutrition increases the risk of infant mortality from infection through reduction in cell-mediated immunity, reduced CD4 helper cells, reduced T cell help, and a reduction of secretory IgA. Trace elements, iron, selenium, copper, and zinc are important in immunity. Lack of these elements can lead to diminished neutrophil killing of bacteria and fungi, susceptibility to viral infections, and diminished antibody responses. Vitamins A, B6, C, E, and folic acid are important in overall resistance to infection.

- **Antioxidant activity.** Carotenoids are antioxidants like vitamin C and E and can enhance NK cell activity, stimulate the production of cytokines and increase the activity of phagocytic cells. Diet and nutrition are powerful innovative tools to reduce illness and death caused by infection.

- **The most significant global cause of immunodeficiency is HIV infection.** AIDS is caused by HIV, which is a double-stranded RNA retrovirus that infects CD4 T cells. Severe CD4 depletion results from a variety of mechanisms, with drastic functional impairment of cell-mediated immunity and death from opportunistic infections.

- **Combination therapy for AIDS** with inhibitors of reverse transcriptase, protease, and viral entry are reasonably successful, but associated with long-term toxicities in almost 50% of persons. An effective vaccine remains an elusive goal.

SOME DRUGS SELECTIVELY ALTER IMMUNE FUNCTION

There have been substantial advances over the past decade in understanding how the immune system is regulated and how drugs may selectively alter function, producing not only **immunodeficiency**, but also, in some circumstances, **immune enhancement**. This chapter examines the most important agents commonly used for systemic immunotherapy.

Corticosteroids are powerful immune modulators

The immune system is regulated by at least four fundamental mechanisms:

- hormones (e.g. glucocorticoids);
- the cytokine system (including interleukins [ILs] and interferons [IFNs]);
- network connectivity (through idiotypic–anti-idiotypic responses); and
- antigens.

Glucocorticoids are among the most powerful naturally occurring modulators of the immune response and have profound effects at most levels and on most components.

As lipid-soluble molecules, steroids gain entry into cells by diffusion through the plasma membrane where they bind to their receptors in the cytoplasm.

The steroid–receptor complex translocates to the nucleus where it binds promoter regions of steroid-responsive genes.

In addition to their direct hormonal action on immune cell traffic and function, steroids have a substantial influence on cytokine synthesis, thereby also exerting a powerful indirect effect.

Significant changes in cell traffic are produced

Administration of steroids causes striking changes in circulating leukocyte populations, even when quite small quantities are used – for example, to produce physiological concentrations in previously adrenalectomized patients. These effects vary between cell types (Fig. 17.1).

Steroid treatment causes circulating **lymphocytopenia**, which is maximal at 4–6 hours and returns to normal by 24 hours. T cells are affected more than B cells, and, within the T cell subsets, CD4 cells are more depleted than CD8 cells. Experimental studies suggest these cells are redistributed to marrow and spleen.

Effects of glucocorticoids on circulating leukocytes

cell type	hours after injection		
	0	6	24
neutrophils	4000	10 000	4000
lymphocytes	2000	500	2000
eosinophils	400	100	400
monocytes	300	50	300
basophils	100	0	100

Fig. 17.1 Effect of a dose of glucocorticoid (40 mg/kg), given once at time zero, on circulating human leukocyte numbers (per mm^3).

Monocytopenia occurs after steroid treatment, is most evident at 2 hours, and recovers by 24 hours, but unlike effects on lymphocyte traffic, further repeated daily dosage does not cause subsequent cycles of depletion.

Neutrophilia is a feature of steroid treatment, due partly to the release of mature stored cells from bone marrow and partly to a reduction in cells leaving the circulation. The rapid and prolonged **decrease in circulating eosinophils and basophils** that occurs after steroid treatment in normal individuals contrasts markedly with the neutrophilia seen at the same time.

T cell activation and B cell maturation are inhibited

T cell activation and proliferation are inhibited by steroids, which make T cells unresponsive to IL-1 and therefore unable to synthesize IL-2, but steroids have little effect on mature B cells.

Q. After prolonged high steroid dosage there is a modest decrease in all immunoglobulin isotypes. Why should this be?

A. Lack of T cell help for B cells will result in a general reduction in the numbers of mature B cells that develop.

Steroids inhibit production of IL-1 and tumor necrosis factor (TNF) by monocytes (see below), but do not block the effect of cytokines on phagocytosis – indeed, they can promote it. Thus the binding of IFNγ and the subsequent expression of HLA-DR molecules and Fc receptors may be increased by low-dose steroids. However, the function of polymorphonuclear leukocytes is resistant to the levels of steroids achievable pharmacologically, as judged by chemotaxis, phagocytosis, and cytotoxicity.

Cytokine synthesis is inhibited

Studies in vitro have shown that physiological and pharmacological concentrations of steroids inhibit the synthesis of cytokines, but have little effect on their function. More impressively, after in-vivo administration, reduced production of IL-1, IL-2, IL-4, IL-6, IL-10, TNFα, and IFNγ has been demonstrated.

Q. What can you infer about which populations of T cells are targeted by steroids?

A. These cytokines are characteristic of different T cell populations, including helper T cells (TH1, TH2) and cytotoxic T cells (CTLs) (see Chapters 2 and 10).

Several different mechanisms may be involved:

- attachment to potential glucocorticoid response elements in the promoter region of the cytokine genes (IL-4, IL-6, and IL-10);
- direct binding, which antagonizes transcription-activating factors for IL-2, IL-8, and TNFα; or
- accelerating breakdown of mRNA (IL-1 and IL-3).

The major consequences of this are inhibition of:

- T cell activation, both TH1 and TH2 cells of the CD4 subpopulation being similarly affected; and
- cells of the monocyte/macrophage system.

Cyclophosphamide acts by covalent alkylation

Together with chlorambucil, cyclophosphamide belongs to the group of immunomodulatory drugs that act by covalent alkylation of other molecules.

Cyclophosphamide has no alkylating ability itself, but many of its metabolites are active, each having two active sites to effect the cross-linking of DNA strands, thereby interfering with strand separation during replication.

The main side effect of cyclophosphamide is marrow toxicity, and so leukopenia must be monitored.

Both T and B cell functions are affected

Cyclophosphamide mainly affects lymphocyte numbers and function, particularly after low-dose daily oral therapy. Polymorphonuclear cell numbers may remain relatively unchanged.

Low-dose oral therapy may have a greater impact on cell-mediated responses, and bolus intermittent treatment more effect on antibody production.

In both humans and experimental animals, after a low-dose bolus (600 mg/m^2 body surface area), the numbers of B lymphocytes are reduced more than T cells and, among T cell subsets, CD8 more than CD4; however, with higher dosage, all cell types are reduced similarly.

Experimental studies have shown that this differential effect of low-dose depletion of CD8 cells allows a paradoxical increase in some CD4-controlled functions, such as antibody production. Evidence that low-dose cyclophosphamide has corresponding clinical relevance in humans remains equivocal.

As cyclophosphamide interferes with both B and T cell function, it is effective in controlling both antibody-mediated and cell-mediated immune responses in experimental animals and humans, and therefore has a major role in the management of both:

- autoantibody-mediated disease; and
- allograft rejection.

Azathioprine is active only on dividing cells

Azathioprine is converted rapidly and non-enzymatically to 6-mercaptopurine in vivo. It exerts its effect after metabolism to thioinosinic acid by competitive inhibition of purine metabolism and by incorporation into DNA as a fraudulent base. Its main effects are therefore on DNA synthesis.

Unlike cyclophosphamide, which is cytotoxic, azathioprine is cytostatic and is active only on dividing cells, exerting maximal effect if given soon after antigenic challenge.

Allopurinol, which inhibits xanthine oxidase, increases the effective dose of azathioprine fourfold. Therefore, if allopurinol is clinically essential (e.g. to treat gout) the dose of azathioprine should be decreased.

T and B cell numbers are reduced

Azathioprine is moderately immunosuppressive and produces modest reductions in both T and B cells after prolonged oral therapy at 2–3 mg/kg daily. Natural killer (NK) cell activity also appears to be suppressed after its use. Humoral immunity and delayed-type hypersensitivity are not affected at the doses given clinically, though there is a reduction in mitogenic responses to pokeweed mitogen in lymphocytes taken from patients receiving the drug.

Mycophenolate inhibits de-novo purine biosynthesis

Mycophenolate mofetil was developed to target selectively the final stage of purine synthesis, along a pathway used specifically by lymphocytes proliferating in response to antigenic challenge. Therefore, unlike nucleoside analogs such as azathioprine, it does not inhibit DNA repair enzymes or incorporate fraudulent purine analogs into DNA.

Mycophenolate is rapidly hydrolyzed in vivo to mycophenolic acid, which is an inhibitor of the enzyme inosine monophosphate dehydrogenase, and therefore results in inhibition of de-novo purine biosynthesis.

Q. In what other circumstances does aberrant purine metabolism selectively damage T cells?

A. Deficiency of purine nucleotide phosphorylase (PNP) or adenosine deaminase (ADA), produces SCID due to the toxic build up of purine metabolites in T cells (see Fig. 16.9).

Lymphocyte proliferation is blocked

Mycophenolate blocks both T and B cell proliferative responses in doses that appear to have no effect on other cell types. It also inhibits glycosylation of adhesion molecules involved in leukocyte traffic, thus restricting amplification of inflammatory injury.

Q. How would inhibiting glycosylation reduce lymphocyte traffic to sites of inflammation?

A. Selectins on the endothelium bind to glycoproteins on lymphocytes to slow them in the circulation before transendothelial migration (see Chapter 6).

Methotrexate blocks DNA synthesis

Methotrexate is a structural analog of folic acid and blocks folic acid-dependent synthetic pathways essential for DNA synthesis.

Immunoglobulin synthesis is reduced after prolonged treatment

Several reports note a reduction in immunoglobulin synthesis, with significantly lowered levels of all isotypes after 3 months of treatment with methotrexate. No consistent change has been noted in T cell subsets, in either the short or long term, or in the function of the monocyte/macrophage system. However, inhibition of dihydrofolate reductase involved in purine synthesis releases adenosine, which is a powerful inhibitor of activated polymorphonuclear leukocytes; hence methotrexate is anti-inflammatory.

Other effects of methotrexate on inflammation may be mediated by its inhibitory effect on arachidonic acid metabolism. More anti-inflammatory actions are indicated by the rapid decrease in indices of inflammatory activity such as C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR), without affecting immune cell function or immunoglobulin synthesis.

Q. Where is CRP synthesized and what is its function?

A. CRP is synthesized primarily in the liver, and it acts as an opsonin by binding to pneumococcal carbohydrates.

Leflunomide inhibits T and B cell proliferation and antibody production

In vivo, **leflunomide** is converted to A77 1726, which non-cytotoxically and reversibly inhibits proliferation of both T and B cells stimulated by a variety of mitogens, and inhibits antibody production. A77 1726 inhibits **dihydro-orotate dehydrogenase**, an enzyme required for pyrimidine nucleotide synthesis.

Ciclosporin, tacrolimus (FK506), and rapamycin affect T cell signaling

Ciclosporin, tacrolimus (FK506), and rapamycin have complicated effects on T cell signaling and hence T cell function. They all bind to a class of cytoplasmic proteins (named **immunophilins**) with peptidyl prolyl isomerase (rotamase) activity, which they inhibit. Immunophilins are believed to have a critical role in transducing signals from the cell surface to the cell nucleus.

Ciclosporin binds to one family of immunophilins, **cyclophilins**, whereas **tacrolimus** and **rapamycin** bind to the **FK binding proteins**.

The ciclosporin–cyclophilin complex targets a serine threonine phosphatase called **calcineurin**, as does the tacrolimus–FK binding protein complex. Both inhibit signal transduction pathways, characteristically producing an increase in intracellular free calcium, and inhibiting transcriptional activation of cytokines and other genes essential for T cell proliferation and function.

Rapamycin blocks T cell proliferation by a different mechanism, through inhibition of IL-2-dependent signal transduction pathways, which function independently of calcium concentration and do not affect cytokine gene transcription (Fig. 17.2). Rapamycin arrests cells in the G_1 phase of the cell cycle, resulting in cell death by apoptosis.

Effects of ciclosporin, tacrolimus, and rapamycin

	ciclosporin	tacrolimus	rapamycin
lymphokine secretion (IL-2, IL-3, IL-4, IL-6, GM–CSF, IFNγ)	↓	↓	– or ↑
IL-2 receptor expression	↓	↓	–
response to IL-2	–	–	↓

GM-CSF, granulocyte–macrophage colony stimulating factor

Fig. 17.2 Differential effects of ciclosporin, tacrolimus, and rapamycin immunophilins on cytokine activity.

T cell proliferation is inhibited

Ciclosporin has:

- a marked inhibitory effect on the early events of T cell proliferation induced by mixed lymphocyte reactions, concanavalin A, or phytohemagglutinin;
- a specific effect on B cells, inhibiting antiglobulin-driven proliferative responses, but not those due to stimulation with lipopolysaccharide;
- effects on antigen presentation by monocytes and Langerhans' cells.

The effect of ciclosporin, while profound on T cells, therefore extends to other cells of the immune system.

It is believed that tacrolimus has a mode of action similar to that of ciclosporin, albeit by attachment to a different immunophilin.

Rapamycin, however, also affects cells of non-hematopoietic origin – for example, it inhibits proliferation of vascular smooth muscle cells after balloon catheter injury and may then be useful in preventing restenosis after angioplasty. Also, because rapamycin inhibits T lymphocyte proliferation at a later stage than ciclosporin or tacrolimus, it may be used synergistically with these agents or as an alternative in conditions refractory to the use of one or the other.

NUTRIENT DEFICIENCIES ARE GENERALLY ASSOCIATED WITH IMPAIRED IMMUNE RESPONSES

The relationship between nutrition and resistance to infection has been suggested by historical accounts of famine and pestilence, clinical observations, and epidemiological data.

Generally, nutrient deficiencies are associated with impaired immune responses. The five aspects of immunity most consistently affected by malnutrition are:

- cell-mediated immunity;
- phagocyte function;
- the complement system;
- secretory antibody; and
- cytokine production.

Worldwide, nutritional deficiency is the most common cause of immunodeficiency.

Economically underprivileged countries have a high prevalence of nutritional deficiencies, as do communities that are economically poor in many industrialized countries.

In addition, nutritional deficiency can result from primary systemic disorders – patients with cancer, chronic renal disease, burns, multiple trauma, and chronic infection show a high prevalence of malnutrition.

Paradoxically, obesity and an excess intake of nutrients are also associated with reduced immune responses.

Infection and malnutrition usually aggravate each other

Nutritional deficiency does not affect all infections equally:

- the clinical course and final outcome of pneumonia, diarrhea, measles, and tuberculosis are affected adversely by nutritional deficiency;
- for some infections (e.g. tetanus and viral encephalitis), the effect of nutritional deficiency is minimal;
- for other infections (e.g. influenza virus and human immunodeficiency virus [HIV]), nutrition exerts a moderate influence.

Many factors predispose to the development of infection in the malnourished individual, including poor sanitation, contaminated food and water, lack of nutritional and health knowledge, illiteracy, and overcrowding.

Lymphoid tissues are damaged by nutrient deficiencies

Lymphoid tissues are very vulnerable to the damaging effects of malnutrition.

The extent and severity of lymphoid dysfunction caused by nutrient deficiencies depend upon several factors, including:

- the rate of cell proliferation;
- the amount and rate of protein synthesis; and
- the role of individual nutrients in critical metabolic pathways.

Numerous enzymes with key roles in immune processes require zinc, iron, vitamin B_6, and other micronutrients to function.

Lymphoid atrophy is a prominent morphological feature of malnutrition.

The thymus, in particular, is a sensitive barometer in young children and the profound reduction in weight and size of the organ in several malnourished subjects has been termed '**nutritional thymectomy**'. Histologically:

- the lobular architecture is ill defined;
- there is a loss of corticomedullary demarcation;
- there are fewer lymphoid cells;
- Hassall's corpuscles are enlarged and degenerate – some may be calcified.

Atrophy is observed in the thymus-dependent periarteriolar areas of the spleen and in the paracortical section of the lymph nodes.

Protein–energy malnutrition affects cell-mediated immunity and phagocytosis

Protein-energy malnutrition (PEM) is associated with a significant reduction in cell-mediated immunity, indicated by:

- a reduced number of $CD4^+$ helper T cells; and
- a lower $CD4^+/CD8^+$ ratio.

Co-culture experiments indicate a reduction in T cell help available to B cells. Lymphocyte proliferative responses to mitogen are decreased. The immaturity of circulating T cells is reflected in increased leukocyte deoxynucleotidyltransferase activity. Reduced **thymulin** activity may underlie these changes in T cell number and function. There is a reduction in the secretory IgA antibody response to common vaccine antigens, which may contribute to a higher incidence of mucosal infections.

Phagocytosis is affected in PEM:

- opsonization is decreased, largely because of a reduction in levels of various complement components – C3, C5, and factor B;
- ingestion of microorganisms is intact, but the ability of phagocytes to kill intracellular organisms is impaired;
- the production of certain cytokines, such as IL-2 and TNF, is decreased.

Some innate mechanisms of immunity are also affected by nutrition:

- the production of lysozyme is slightly decreased;
- a larger number of bacteria bind to epithelial cells of malnourished subjects;
- wound healing is impaired.

There are very few data on the quality and quantity of mucus produced in PEM.

Zinc and iron deficiencies have a variety of effects on immunity

The profound effect of **zinc deprivation** on immune responses has been documented extensively. There is:

- a reduction in delayed cutaneous hypersensitivity,
- lower $CD4^+/CD8^+$ ratios; and
- T cell dysfunction.

A striking and pathognomonic feature of zinc deficiency is reduction in the activity of serum thymulin (a nonapeptide that contains zinc as an integral part of its molecule) and lymphoid atrophy.

There is also an intergeneration effect of zinc on immunity as demonstrated by the surprising finding of the effect in mice of zinc deficiency in pregnancy. Even the third-generation progeny have impaired antibody responses, as shown by diminished numbers of plaque-forming cells (PFC) and lower levels of IgM (Fig. 17.3).

Iron is a double-edged sword:

- it is required by most microorganisms for their growth;
- iron-dependent enzymes have crucial roles in lymphocyte and phagocyte function.

Iron deficiency is therefore generally associated with a reduced ability of neutrophils to kill bacteria and fungi, decreased lymphocyte response to mitogens and antigens, and impaired NK cell activity.

Q. In what ways can neutrophils and macrophages restrict iron availability for microorganisms?

A. Soluble proteins such as lactoferrin produced by neutrophils reduce the availability of free iron in the phagolysosome. Macrophages have ion pumps (e.g. nRAMP) in their phagosomal membrane that remove iron from the phagosome (see Chapter 15).

Intergeneration effects of zinc deficiency on immunity

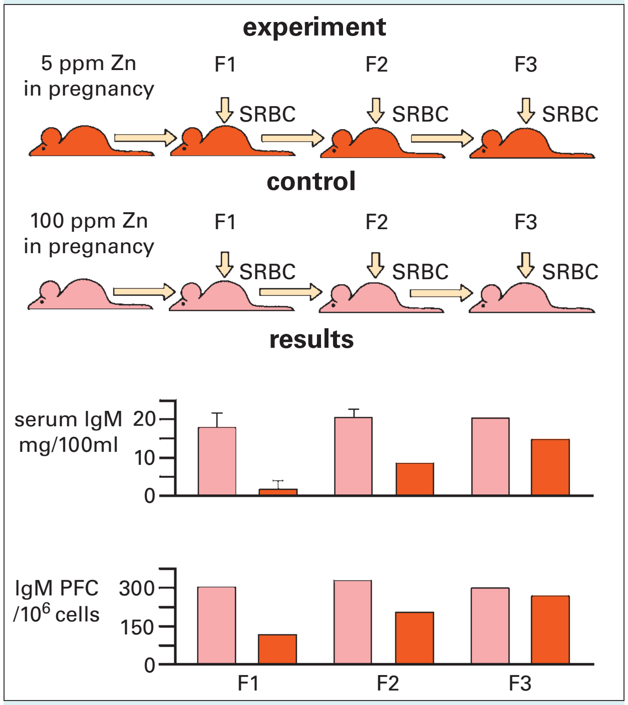

Fig. 17.3 An experimental group of animals was fed a zinc-deficient diet (5 ppm Zn) during the later two-thirds of pregnancy and a control group was fed a zinc-adequate diet (100 ppm Zn) during the same period. The former regimen reduced serum Zn to 60–70% of control values. F1, F2, and F3 generations were fed sufficient zinc throughout. Both the PFC response and serum concentration of IgM were assessed at 6 weeks. Results showed that both were low in at least three generations of offspring.

Selenium and copper are also important for immune responses

An exciting recent observation indicates that viruses can mutate and show altered virulence in malnourished hosts. Coxsackie virus recovered from selenium-deficient mice produced heightened myocardial damage – there were six nucleotide changes between the avirulent input virus strain and the virulent virus recovered from selenium-deficient animals.

The clinical counterpart of coxsackie myocarditis and nutritional deficiency is Keshan disease, which was endemic in some parts of China. Selenium supplementation has virtually eradicated this condition.

Vitamin supplementation is of value in severe measles

Vitamin A deficiency alters epithelial structure, leading to metaplasia and increased binding of bacteria. There is a reduction in the numbers of certain lymphocyte subsets and in the response to mitogen.

Vitamin supplementation is of value in preventing the complications of severe measles and reducing the mortality rate from this disease.

Vitamin B_6 and folate deficiencies reduce cell-mediated immunity, particularly lymphocyte proliferation responses, and also impair antibody production.

Obesity is associated with altered immune responses

Obese subjects and animals show alteration in various immune responses, including:

- cytotoxicity;
- NK activity; and
- the ability of phagocytes to kill ingested bacteria and fungi.

Altered levels of some micronutrients, lipids, and hormones may explain these immunological changes.

Some nutrients in moderate excess enhance immune responses

Some nutrients given in moderate excess enhance selected aspects of immune responses, particularly cell-mediated immunity. These include:

- vitamin E;
- vitamin A;
- zinc; and
- selenium.

However, for most nutrients, there is an upper limit of intake beyond which immune responses are impaired.

Nutritional intervention may prevent infection

There are exciting new possibilities for nutritional intervention for both primary and secondary prevention of infection in high-risk groups.

Hospital inpatients who are malnourished are at high risk of complicated opportunistic infections. Nutrient-enriched feeding formulas:

- enhance immunity; and
- reduce the risk of complications such as sepsis and poor wound healing.

However, the administration of such nutritional supplements, often delivered intravenously, can make the patient susceptible to nosocomial infections associated with indwelling catheters.

In the elderly, respiratory infection is a common cause of illness. Modest amounts of micronutrient supplements improve immune responses and, more significantly, reduce the incidence of respiratory infections and antibiotic usage. Furthermore, post-vaccination immune responses are higher in subjects given nutritional supplements than in untreated controls.

Probiotics benefit health and immunity

In both clinical and veterinary medicine, the value of probiotics is being recognized. These ‘desirable’ bacteria such as *Lactobacillus acidophilus*, *L. casei*, cocci such as *Enterococcus fecium*, and bifidobacteria are given orally to replace or increase their presence in the gut microflora. Benefits to health and immunity come from:

- the ‘barrier effect’ in the gut;
- production of bacteriocidins;
- alteration of the local immune response via a change in cytokine profile in the gut mucosa; and
- increased antibody production.

ACQUIRED IMMUNE DEFICIENCY SYNDROME

The most significant global cause of immunodeficiency is HIV infection

HIV is a retrovirus that is transmitted sexually, in blood or blood products, and perinatally. There are two main variants, HIV-1 and HIV-2:

- HIV-2 is endemic in West Africa and appears to be less pathogenic;
- HIV-1 has several subtypes (or clades), which are designated by the letters A through K, and the prevalence of the different clades varies by geographical region – over 90% of people infected with HIV-1 live in developing countries and spread is 80% by the sexual route.

Over 25 million people have died from AIDS since the first cases were described in 1981.

Although sub-Saharan Africa has the highest prevalence of HIV-1 infection in the world (7.4% for the region), the largest increase of new infections is occurring in East Asia, where the number of HIV-positive people increased by 50% between 2002 and 2004.

As of the end of 2004, The World Health Organization (WHO) estimates that, approximately 40 million people are living with HIV infection worldwide, with approximately five million new infections and three million deaths due to AIDS each year.

CD4 antigen is the main receptor for HIV entry

HIV is an enveloped retrovirus that contains two copies of a single-stranded RNA (ssRNA) genome (Fig. 17.4). Upon entry into a cell, the genome is reverse transcribed into complementary DNA (cDNA), which is integrated into the host cell genome (provirus).

The 10-kilobase genome encodes nine genes flanked at each end by long terminal repeat (LTR) sequences, which are:

- essential for integration of viral DNA into the host cell DNA; and
- contain binding sites for regulatory proteins involved in viral replication.

The basic gene structure contains *gag* (core protein), *pol* (reverse transcriptase, protease, and integrase enzymes), and *env* (envelope protein) genes.

In addition to these three main gene products, the virus encodes six regulatory and accessory proteins (Tat, Rev, Vpr, Vpu, Vif, and Nef), which regulate viral protein synthesis.

The viral genes overlap in different reading frames and the proteins are generated by alternative splicing mechanisms.

CD4 antigen is the main receptor for viral entry – it is present on $CD4^+$ T lymphocytes, monocytes, dendritic cells (DCs), and brain microglia. The envelope glycoprotein gp120 is also capable of binding DC-SIGN.

Human immunodeficiency virus and its life cycle

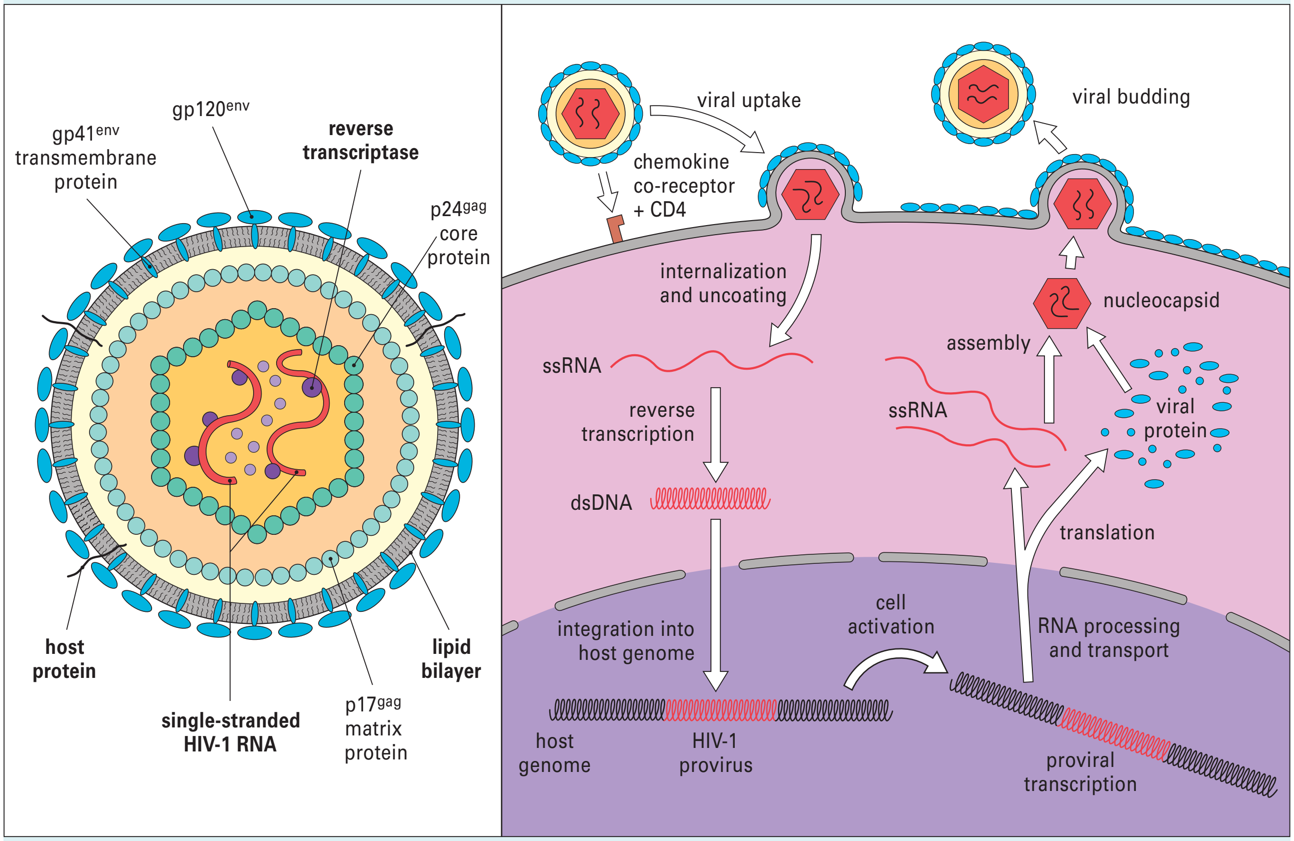
gp41^{env}
transmembrane protein
gp120^{env}
reverse transcriptase
p24^{gag}
core protein
host protein
lipid bilayer
single-stranded HIV-1 RNA
p17^{gag}
matrix protein
viral uptake
chemokine co-receptor + CD4
viral budding
internalization and uncoating
nucleocapsid
assembly
ssRNA
reverse transcription
ssRNA
viral protein
dsDNA
translation
integration into host genome
cell activation
RNA processing and transport
host genome
HIV-1 provirus
proviral transcription

Fig. 17.4 After attachment to CD4 and a chemokine co-receptor (usually CCR5), the virus membrane fuses with the cellular membrane to allow entry into the cell. Following uncoating, reverse transcription of viral RNA results in the production of double-stranded DNA (dsDNA). This is inserted into the host genome as the HIV provirus, by a virally coded integrase enzyme, which is a target for new antiviral medications currently in development. Cell activation leads to transcription and the production of viral mRNAs. Structural proteins are produced and assembled. Free HIV viruses are produced by viral budding from the host cell, after which further internal assembly occurs with the cleavage of a large precursor core protein into the small core protein components by a virally coded protease enzyme, producing mature virus particles, which are released and can go on to infect additional cells bearing CD4 and the chemokine co-receptor.

Q. What is DC-SIGN?

A. DC-SIGN is a type II transmembrane protein on DCs that has a C-type lectin extracellular domain (see Fig. 9.18).

Although HIV-1 does not use DC-SIGN to infect DCs, this interaction allows DCs to internalize HIV-1 at mucosal sites and then migrate to lymph nodes where they deliver HIV-1 to CD4 T cells.

The viral envelope glycoprotein gp120 binds to CD4

The binding of the viral envelope glycoprotein gp120 to CD4 results in conformational changes in gp120 that expose binding sites for chemokine receptors, which serve as co-receptors for viral entry.

The chemokine receptor used largely determines which cell type different HIV-1 variants can infect:

- most primary isolates of HIV-1 use CCR5 as a co-receptor (R5 viruses) – rare individuals with a genetic 32-base pair deletion in CCR5 that results in a frameshift mutation and truncation of the protein are relatively resistant to HIV-1 infection;
- the other major variant of HIV-1 uses CXCR4 as a co-receptor (X4 viruses).

In approximately 50% of individuals the viral phenotype switches from R5 to X4 during the later stages of infection.

Q. How will the switch in the viral phenotype affect which cells are infected?

A. CCR5 is selectively expressed on mononuclear phagocytes. CXCR4 is present on a number of cell types, including naive T cells, B cells, and monocytes; consequently there is a shift towards T cell infectivity (see Chapter 6 and Appendix 4).

Immune dysfunction results from the direct effects of HIV and impairment of CD4 T cells

There is wide-ranging immune dysfunction in HIV-1 infection, with immune suppression within a milieu of immune activation, resulting from:

- direct effects of HIV; and
- depletion and functional impairment of the CD4 T cell subset over time.

How HIV kills its target cells is not well understood. Several different mechanisms have been proposed, including accumulation of RNA and unintegrated DNA in the cell cytoplasm and intracellular binding of CD4 and gp120. Infected cells may bind to uninfected cells by gp120–CD4 linkages, with multinucleate giant cell and syncytium formation.

gp120 bound to the surface of uninfected CD4 T cells also makes them vulnerable to antibody-dependent cell-mediated cytotoxicity (ADCC), while infected cells may be killed by HIV-1-specific CD8 cytotoxic T lymphocytes (CTLs).

HIV proteins may act as superantigens, resulting in vast expansion and then exhaustive depletion of cells. In addition, HIV may induce T cell apoptosis and viral budding may lead to cell membrane weakening and lysis.

The spectrum of immune dysfunction is characterized by:

- depletion of the CD4 T cell subset; and
- decreased responses to antigens, mitogens, alloantigens, and anti-CD3 antibody, associated with decreased IL-2 production and other changes in cytokine production.

The HIV-1 protein Nef selectively downregulates expression of HLA-A and -B alleles, so reducing surface expression of MHC class I molecule–peptide complexes and decreasing detection by CTLs. To avoid killing of infected cells by NK cells, which kill cells devoid of MHC class I molecules, Nef does not decrease expression of HLA-C and HLA-E.

Q. What is the functional significance of the observation that Nef does not decrease HLA-C and HLA-E?

A. HLA-C and HLA-E are inhibitory signals for NK cells. Therefore Nef can decrease cell recognition by CTLs while not increasing susceptibility to NK cells (see Chapter 10).

Nef has also been shown to decrease the expression of CD4 on infected cells such that gp120 on progeny virions is free to bind to CD4 and chemokine receptors on recipient cells.

There is an:

- increase in activated and unresponsive CD8 T cells;
- increased β_2-microglobulin and neopterin in serum;
- polyclonal B cell activation with B cells refractory to T cell-independent B cell activators; and
- an increase in autoantibodies and immune complexes.

Modeling of the plasma virus and CD4 T cell responses to antiviral therapy suggest that the average half-life of the virus and infected cells in the circulation is less than 2 days; 10^9–10^{10} viruses are released from infected cells and similar numbers of new cells are infected and die daily.

Antibody response appears to be ineffective in controlling HIV infection

The development of antibodies is an important component of the adaptive immune response to prevent or modulate infection. However, in the case of HIV-1 infection, neutralizing antibodies appear to be ineffective in controlling infection and viral replication.

Antibody responses to envelope glycoproteins can be detected in the serum of infected individuals 2–3 weeks after infection, though these lack the ability to inhibit viral infection. It is not until several more weeks after virus infection, and usually after the initially high levels of viremia have subsided, that neutralizing antibody responses are detected.

The main targets of neutralizing antibodies are:

- gp41;
- the second and third hypervariable loops (V2 and V3) of gp120; and
- the CD4-binding domain of gp120.

Antibodies to the V2 and V3 loop tend to be isolate-specific and therefore of doubtful use as potential vaccine-elicited antibodies. Furthermore, the genes encoding the V2 and V3 loops of gp120 are highly mutatic.

Q. What is likely to be the effect on the virus of antibodies to V2 and V3 during an HIV infection?

A. Specific antibodies will act as selective pressure on the virus to mutate these segments of the antigen to avoid immune recognition (see Chapter 13).

It is noteworthy that broadly cross-reactive neutralizing V3-specific antibodies have been described later in infection after prolonged antigen exposure, most likely resulting from somatic hypermutation in the variable region immunoglobulin genes.

Furthermore, studies in simian immunodeficiency virus (SIV) infection in macaques have indicated that pre-existing circulating neutralizing antibodies may alter the clinical outcome of infection. This indicates that the induction of high-titre neutralizing antibodies will be an important component in strategies to prevent HIV-1 infection.

Cellular immune responses play a role in controlling HIV viremia

In addition to antibody responses, cellular immune responses are generated in infected individuals and are thought to play an important role in control of viremia, though, like antibody responses, they are unable to eliminate infection.

The importance of virus-specific CD8 T cells in HIV-1 infection has been best demonstrated using an animal model of AIDS, where macaques depleted of CD8 T cells for more than 28 days during primary SIV infection failed to control viremia, experienced an accelerated disease course, and developed an AIDS-like syndrome resulting in death.

In humans, HIV-1-specific CD8 T cells are readily detected in:

- peripheral blood;
- lymph nodes; and
- gastrointestinal mucosal tissues.

HIV-1-specific CTLs have been shown to lyze infected cells and suppress viral replication in vitro. All expressed proteins have been shown to be targeted by HIV-1-specific CTLs, and in some individuals greater than 25% of total CD8 T cells are specific for HIV-1 epitopes.

Despite the presence of high numbers of HIV-1-specific CTLs in peripheral blood, no significant correlations have been established between the total number of IFNγ-producing CTLs and viral load.

HIV-1-specific CTLs in chronic progressive HIV-1 infection are impaired

There is evidence that HIV-1-specific CTLs in chronic progressive infection are functionally impaired because they:

- typically contain low levels of intracellular perforin (see Figs 10.11 and 10.12), a cytotoxic protein; and
- fail to exhibit strong proliferative capacity upon exposure to cognate antigens in vitro in individuals with high viral loads.

HIV-1-specific CTLs do, however, exert strong selection pressure on the virus, resulting in mutations in CTL epitopes that are incapable of binding to specific MHC class I molecules.

In addition, mutations affecting antigen processing have recently been described that interfere with intracellular proteasome processing of viral proteins such that specific epitopes are no longer generated and presented by MHC class I molecules on the surface of infected cells.

Depending on where in the HIV-1 genome these occur:

- mutations may result in loss of recognition of epitopes by HIV-1-specific CTLs and subsequent loss of control of viremia; or
- the mutations may occur in functionally important regions of viral proteins that impose a significant fitness cost to the virus and ultimately result in lower infectivity of the virus.

HIV-1-specific CD4 T cells are preferentially infected by the virus

HIV-1-specific CD4 T cell responses are complicated by the fact that CD4 T cells are the major population of cells infected by HIV-1, and recent evidence indicates HIV-1-specific CD4 T cells are preferentially infected by the virus.

Although HIV-1-specific CD4 T cells are present in peripheral blood at much lower levels than their CD8 counterparts, increasing evidence supports an essential role for these cells in containing viremia.

The major targets of HIV-1-specific CD4 T cells are the Gag and Nef proteins, and an inverse correlation between proliferative responses to Gag p24 antigen and viral load has been described.

Furthermore individuals with low to undetectable viral loads have been shown to have a higher percentage of HIV-1-specific cells capable of secreting both IFNγ and IL-2 in response to cognate antigen, in contrast to CD4 T cells from individuals with higher viral loads who secrete IFNγ only, with little or no detectable IL-2 production.

Evidence that HIV-1 exerts a progressive functional impairment of CD4 T cells over time has come from recent data indicating that autologous CD4 T cells isolated from individuals during acute infection can restore the proliferative capacity of CD8 T cells from chronically infected individuals when added back to in-vitro proliferation assays.

Moreover, vaccine-induced HIV-1-specific CD4 T cell responses have been correlated with restoration of the proliferative capacity of HIV-1-specific CD8 T cells.

The long-term clinical benefits of restored HIV-1-specific CD4 and CD8 T cell function remain to be determined, and such studies will provide important information regarding correlates of immune protection in HIV-1 infection.

Acute HIV infection is associated with a transient depletion of peripheral CD4 T cells

Acute HIV infection is associated with a transient illness similar to glandular fever (infectious mononucleosis) in up to 80% of persons, with clinical symptoms including malaise, muscle pains, swollen lymph nodes, sore throat, and rash. There is:

- transient depletion of peripheral CD4 T cells;
- expansion of CD8 T cells; and
- levels of plasma virus that reach 10 million virus particles per ml of plasma in most persons (Fig. 17.5).

Gut-associated lymphoid tissue is a major site of early virus replication, where there is large percentage of CCR5-positive CD4 cells, which become virtually eliminated in the early weeks following infection.

Because it takes 2–6 weeks on average for antibodies to core and surface proteins to be detectable by enzyme-linked immunoassays (ELISAs, see Method Box 3.2, Fig. 2), acute infection cannot be diagnosed by standard tests. Instead, this early acute phase can be detected only by assays that detect virus, such as plasma viral RNA levels.

Symptoms of acute infection have usually resolved by the time antibodies are detected, and a period of asymptomatic infection ensures.

The rate of disease progression in HIV infection is associated with both viral load and CD4 count at 6 months after infection.

Viral load 6 months after acute infection averages 30 000 RNA copies/ml plasma, with normal CD4 cell numbers in most persons.

AIDS is defined by a CD4 count < 200/μl

On average it takes 10 years to develop AIDS, defined by the presence of an AIDS-defining illness or a CD4 count of less than 200 cells/μl.

Strikingly, some persons progress to AIDS in less than 2 years, whereas other untreated persons have not developed AIDS in over 25 years of documented infection.

In untreated infection, there is a gradual increase in viral load and decline in CD4 count. Early indicators of immune function decline include non-specific constitutional symptoms such as fevers, night sweats, diarrhea, and weight loss.

Natural history of HIV infection

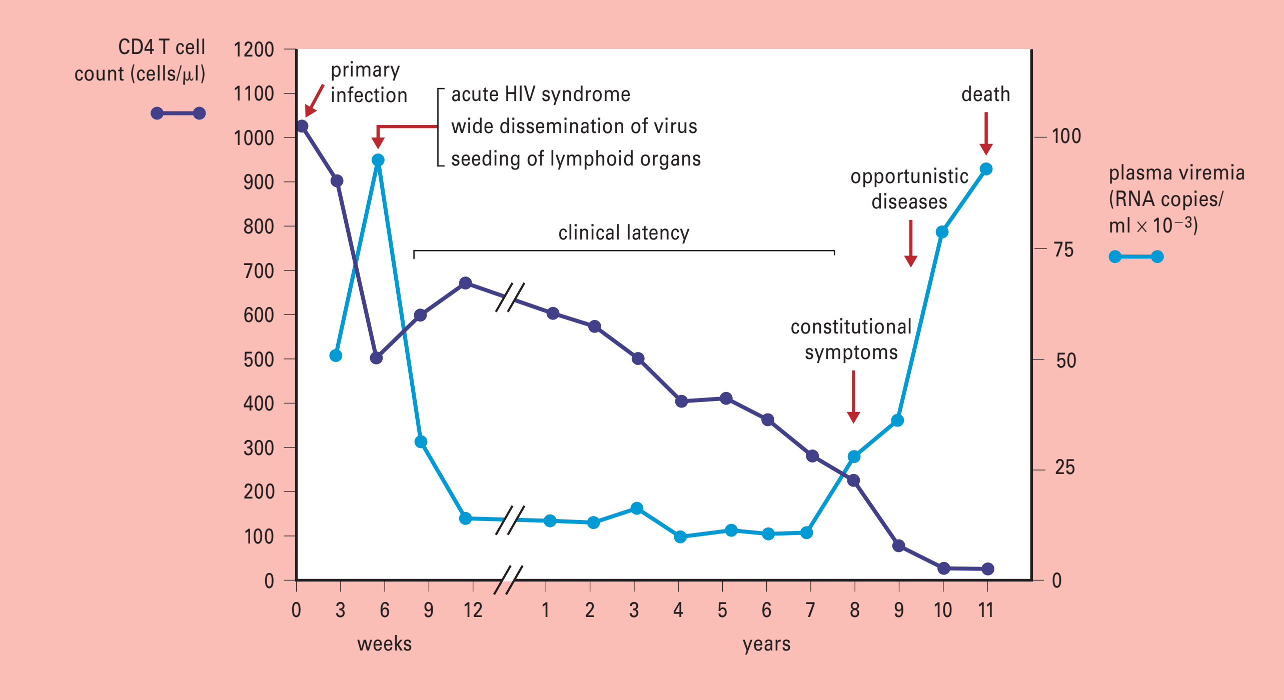

Fig. 17.5 A typical course of HIV infection. (Courtesy of Dr AS Fauci. Modified with permission from Pantaleo G, Graziosi C. N Engl J Med 1993;328:327–335. Copyright 1993 Massachusetts Medical Society. All rights reserved)

Q. Which mediators lead to the symptoms described above?

A. Inflammatory cytokines, including IL-1, which induces fever, and TNFα, which induces cachexia (weight loss), generate these symptoms (see Chapter 6).

Other less severe conditions that occur largely affect the mucous membranes and skin, for example:

- oral candidiasis (thrush);
- dermatomal varicella-zoster infection (shingles);
- recurrent anogenital herpes simplex virus infection; and
- a variety of other cutaneous skin infections.

These conditions often precede the development of serious opportunistic infections and tumors, which are AIDS-defining illnesses that typically occur when the CD4 count is less than 200/μl.

More severe infections are associated with a low CD4 count

Because of the strong relationship between CD4 count and risk of more severe infections, reflecting severe immune deficiency, AIDS is now also defined on the basis of CD4 count alone even in the absence of symptomatic disease, when the $CD4^+$ T cell count is below 200/ml (see Fig. 17.5).

Kaposi's sarcoma, a multifocal tumor of endothelial cells (Fig. 17.6), is the commonest tumor. Widespread skin, mucous membrane, visceral (gut and lungs), and lymph node disease occurs. Infection with **human herpes virus 8 (HHV8)** is associated with development of the tumor. B cell lymphomas also occur, affecting the brain, gut, and bone marrow.

Most of the opportunistic infections are due to reactivation of latent organisms in the host or, in some cases, ubiquitous organisms to which we are continually exposed. They are difficult to diagnose and treatment often suppresses rather than eradicates them. Relapses are common and continuous suppressive or maintenance treatment is necessary, using drugs that cause side effects. The three main organ systems affected are the:

- gastrointestinal tract;
- respiratory system; and
- nervous system.

Q. Why should these systems be most commonly affected in HIV infection?

A. The respiratory and gastrointestinal systems are sites that are normally colonized by non-pathogenic microbes. A shift in the balance of the immune defense of these systems means that such microbes are not controlled. Immune responses are also limited and viruses are often tolerated in host cells, rather than destroyed. Again a shift in the immune balance can leave such latent infections uncontrolled (see Chapter 6).

Pneumonia is common and ***Pneumocystis carinii*** is the commonest infection (see Fig. 17.6), but bacterial infections, including ***Mycobacterium tuberculosis***, and **fungal infections** also occur.

Common features of late-stage HIV infection

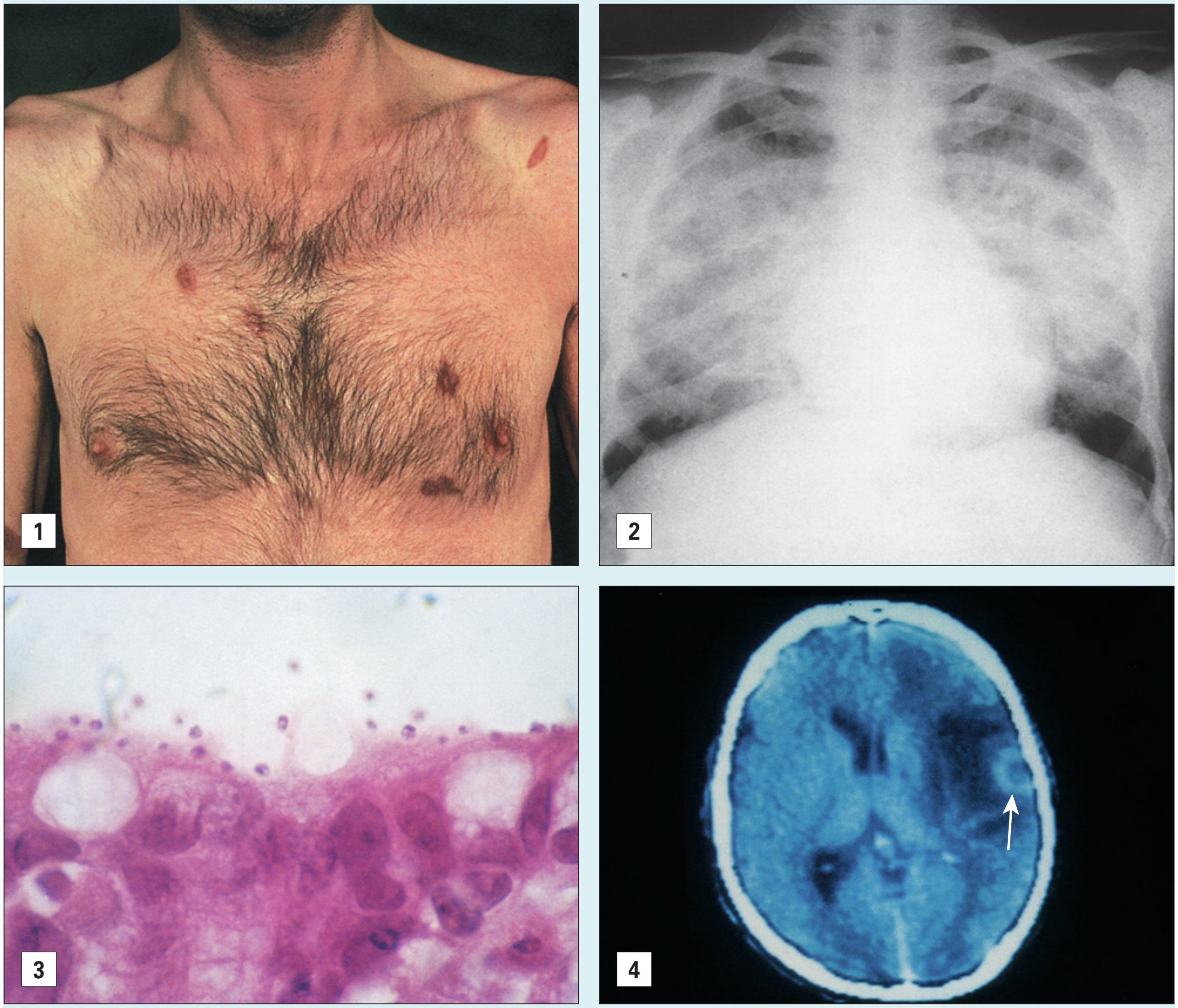

Fig. 17.6 (**1**) Multiple Kaposi's sarcoma lesions on the chest and abdomen. (**2**) Chest radiograph of a patient with *Pneumocystis carinii* pneumonia, showing bilateral interstitial shadowing. (**3**) Small bowel biopsy from a patient with diarrhea caused by cryptosporidia, showing intermediate forms of cryptosporidia (small pink dots) on the surface of the mucosa. (**4**) Computed tomography scan of the head of a patient with cerebral toxoplasmosis. The patient presented with a history of fits and weakness of the left arm and leg. Injection of contrast revealed a ring-enhancing lesion in the right hemisphere (arrow), with surrounding edema (dark area).

Discomfort on swallowing is usually caused by **candidiasis** (thrush), but cytomegalovirus can cause esophageal ulceration.

Protozoa (cryptosporidia and microsporidia) are the commonest pathogens isolated in patients with diarrhea and weight loss (see Fig. 17.6), but **enteric bacteria** such as *Salmonella* and *Campylobacter* spp. may also be found.

Neurological complications in AIDS are due to:

- direct effects of HIV infection;
- opportunistic infections; or
- lymphoma.

AIDS-related dementia once affected between 10% and 40% of patients with other manifestations of AIDS, but with more effective antiviral treatment has become less common.

Spinal cord and peripheral nerve disease also occur.

Toxoplasmosis, a protozoal infection, causes cysts in the brain and neurological deficit (see Fig. 17.6).

Cryptococcus neoformans is a fungus that causes meningitis.

Cytomegalovirus may cause inflammation of the retina, brain, and spinal cord and its nerve roots, and a

polyomavirus (JC virus), which infects oligodendrocytes in the brain, produces a rapidly fatal demyelinating disease – **progressive multifocal leukoencephalopathy**.

Treatment with highly active antiretroviral therapy (HAART) controls HIV infection

In 1987, zidovudine (AZT) was licensed as the first **nucleoside analog reverse transcriptase inhibitor (NRTI)** to be used in HIV infection. Together with targeted therapies to prevent opportunistic infections, even single drug therapy had an impact on the mortality rate.

Since then there have been considerable advances, with the development of:

- other NRTIs;
- **non-nucleoside reverse transcriptase inhibitors (NNRTIs)**;
- **protease inhibitors**; and
- **entry inhibitors** (see Fig. 17.4).

The introduction of **combination chemotherapy** for the treatment of HIV infection, which occurred in 1996, resulted in a dramatic decrease in deaths due to AIDS. HIV infection became a chronic controllable illness for persons fortunate enough to have access to these medications, which typically consist of:

- two NRTIs; and
- a NNRTI or protease inhibitor.

Treatment with these combinations, termed **highly active antiretroviral therapy (HAART)**, typically results in:

- lowering of plasma viral load to levels below detection by the most sensitive assays (< 50 RNA copies/ml plasma); and
- dramatic increases in CD4 cell counts, though these often do not return to normal levels of greater than 800 cells/μl;
- a decreased risk of opportunistic infections as CD4 counts rise.

The optimal time to start therapy continues to be debated, but most guidelines recommend HAART when:

- CD4 count is below 350 or 200 cells/μl;
- in all patients who have an AIDS defining illness.

Serum HIV RNA levels also influence this decision, though again there is debate on when to act, with some recommending therapy when the viral load is greater than 50 000 RNA copies/ml plasma, and others deferring until viral load is greater than 100 000.

Until recently, the cost of these drugs was so high that they were not accessible in resource-constrained settings, which is where over 90% of global infections are found. The combination of international donor efforts, reduced drug pricing, and generic medications have all contributed to an increasing access to these life-saving medications.

Early studies show that persons living in poverty not only can adhere to the complicated regimens, but that they are quite effective.

An effective vaccine remains an elusive goal

To date there have been clinical trials of numerous candidate HIV vaccines, but most would agree that an effective vaccine remains an elusive goal.

Despite increasing characterization of adaptive immune responses, the correlates of protection remain to be fully defined, and this has left the field with mostly empiric approaches.

Encouraging early reports of persons repeatedly exposed to HIV who never became infected suggested that adaptive immune responses, particularly HIV-specific CD8 T cell responses, might be responsible for apparent protection, but this remains somewhat controversial.

A vaccine that prevents infection altogether will almost certainly require the induction of broadly neutralizing antibody responses, something that is yet to be achieved.

Q. Why is it difficult to identify suitable antigens that could be used in a neutralizing vaccine?

A. The very rapid rate of HIV mutation and the high rate of virus production mean that the virus can readily mutate to evade a specific immune response, and yet may develop variants that retain infectivity.

Many in the field believe that a vaccine that protects from disease progression, while not necessarily protecting from the initial infection, is a more realistic short-term goal. The approaches being used for this goal focus on the induction of cellular immune responses, particularly CD8 T cell responses, though increasing data indicate that robust CD4 T cell responses to HIV will also be needed.

As no cure or vaccine is currently available, our main weapon is prevention through health education and control of infection.

FURTHER READING

Altfeld M, Allen TM, Yu XG, et al. HIV-1 superinfection despite broad CD8$^+$ T cell responses containing replication of the primary virus. Nature 2002;420:434–439.

Brenchley JM, Schacker TW, Ruff LE, et al. CD4$^+$ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. J Exp Med 2004;200:749–759.

Day CL, Walker BD. Progress in defining CD4 helper cell responses in chronic viral infections. J Exp Med 2003;198:1773–1777.

Migueles SA, Laborico AC, Shupert WL, et al. HIV-specific CD8$^+$ T cell proliferation is coupled to perforin expression and is maintained in nonprogressors. Nat Immunol 2002;3:1061–1068.

Piot P, Feachem RG, Lee JW, Wolfensohn JD. Public health. A global response to AIDS: lessons learned, next steps. Science 2004;304:1909–1910.

Robbins GK, De Gruttola V, Shafer RW, et al. Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. N Engl J Med 2003;349:2293–2303.

Critical thinking: Secondary immunodeficiency (see p. 498 for explanations)

A 52-year-old record producer developed a severe cough with increasing shortness of breath. He also had a fever, chest pain, and malaise. For the week before presentation he complained of pain on swallowing. His past medical history included gonorrhea and genital herpes within the previous 3 years. Over the previous 2 months he had suffered from persistent diarrhea and lost 9 kg in weight from a baseline of 68 kg. He lived with his male partner with whom he had been having unprotected intercourse for several years. There was no history of intravenous drug abuse.

On examination he was underweight and had enlarged lymph nodes in the neck, axillae, and groin. Plaques of *Candida albicans* were visible in his throat. There were abnormal breath sounds in his lungs. The results of his blood tests are shown in Table 1.

Table 1 Results of investigations on presentation

investigation	result (normal range)
hemoglobin (g/dl)	12.8 (13.5–18.0)
platelet count (× 10^9/l)	128 (150–400)
white cell count (× 10^9/l)	6.2 (4.0–11.0)
neutrophils (× 10^9/l)	5.4 (2.0–7.5)
eosinophils (× 10^9/l)	0.24 (0.4–0.44)
total lymphocytes (× 10^9/l)	0.75 (1.6–3.5)
T lymphocytes CD4$^+$ (× 10^9/l)) CD8$^+$ (× 10^9/l)	 0.12 (0.7–1.1) 0.42 (0.5–0.9)
B lymphocytes (× 10^9/l)	0.11 (0.2–0.5)
arterial blood gases PaO_2 (kPa) $PaCO_2$ (kPa) pH HCO_3^- base excess	 7.8 (> 10.6) 5.52 (4.7–6.0) 7.39 (7.35–7.45) 25.6 –0.9
ECG	normal
chest radiography	bilateral diffuse interstitial shadowing
bronchoscopy with bronchoalveolar lavage	positive for *Pneumocystis carinii*

Because of his sexual history, the patient was counseled about having a human immunodeficiency virus (HIV) test, and consented. An enzyme-linked immunosorbent assay (ELISA, see Method Box 3.2, Fig. 2) was positive for anti-HIV antibodies and a polymerase chain reaction (PCR) demonstrated HIV-1 RNA in the plasma.

Examination of an induced sputum specimen revealed cysts of *Pneumocystis carinii*, which together with the positive HIV ELISA is an AIDS-defining illness. Thus a clear diagnosis of acquired immune deficiency syndrome (AIDS) was made and the patient's *P. carinii* pneumonia was treated with oxygen by mask and parenteral co-trimoxazole. He was discharged from hospital taking oral co-trimoxazole.

Within 3 months he was seen again in accident and emergency with blurred vision and 'flashing lights' in his eyes. He was shown to have an infection of his retina with cytomegalovirus and was treated with injections of ganciclovir. The CD4 count at this time was 0.04 × 10^9/l. While receiving this treatment the patient became increasingly unwell and semiconscious. Investigations at this time are shown Table 2.

Critical thinking: Secondary immunodeficiency *(Continued)*

Table 2 Results of investigations 3 months after presentation

investigation	result (normal range)
hemoglobin (g/dl)	10.4 (13.5–18.0)
platelet count (× 10^9/l)	104 (150–400)
white cell count (× 10^9/l)	4.1 (4.0–11.0)
neutrophils (× 10^9/l)	4.2 (2.0–7.5)
eosinophils (× 10^9/l)	0.24 (0.4–0.44)
total lymphocytes (× 10^9/l)	0.62 (1.6–3.5)
T lymphocytes	
CD4$^+$ (× 10^9/l)	0.03 (0.7–1.1)
CD8$^+$ (× 10^9/l)	0.40 (0.5–0.9)
B lymphocytes (× 10^9/l)	0.09 (0.2–0.5)
chest radiography	minimal areas of diffuse shadowing
blood culture	negative
blood glucose (mmol/l)	7.6 (< 10.0)
CSF from lumbar puncture	
appearance	turbid
white cells (polymorphs/mm^3)	2500
protein (g/l)	4.2 (0.15–0.45)
glucose (mmol/l)	4.5 (> 60% blood glucose)
Indian ink stain	positive for cryptococcus

A diagnosis of cryptococcal meningitis was made and intravenous amphotericin was started. The patient did not respond to treatment and died shortly afterwards. At autopsy, *P. carinii* was isolated from his lungs and evidence of early cerebral lymphoma was noted.

1 What diagnostic tests are available for HIV infection?

2 Which of these tests should be used if HIV infection is suspected in a mother and her child infected vertically?

3 What serological and cellular indices can be used to monitor the course of HIV infection?

CHAPTER

18

Vaccination

SUMMARY

- **Vaccination applies immunological principles to human health.** Adaptive immunity and the ability of lymphocytes to develop memory for a pathogen's antigens underlie vaccination. Active immunization is known as vaccination.
- **A wide range of antigen preparations are in use as vaccines,** from whole organisms to simple peptides and sugars. Living and non-living vaccines have important differences, living vaccines being generally more effective.
- **Adjuvants enhance antibody production,** and are usually required with non-living vaccines. They concentrate antigen at appropriate sites or induce cytokines.
- **Most vaccines are still given by injection,** but other routes are being investigated.
- **Vaccine efficacy needs to be reviewed from time to time**.
- **Vaccine safety is an overriding consideration.** The MMR controversy resulted in measles epidemics.
- **Vaccines in general use have variable success rates.** Some vaccines are reserved for special groups only and vaccines for parasitic and some other infections are only experimental.
- **Passive immunization can be life-saving.** The direct administration of antibodies still has a role to play in certain circumstances, for example when tetanus toxin is already in the circulation.
- **Non-specific immunotherapy can boost immune activity.** Non-specific immunization, for example by cytokines, may be of use in selected conditions.
- **Immunization against a variety of non-infectious conditions is being investigated.** Recombinant DNA technology will probably be the basis for the next generation of vaccines.

VACCINATION APPLIES IMMUNOLOGICAL PRINCIPLES TO HUMAN HEALTH

Vaccination is the best known and most successful application of immunological principles to human health.

The first vaccine was named after vaccinia, the cowpox virus. Jenner pioneered its use 200 years ago. It was the first deliberate scientific attempt to prevent an infectious disease (smallpox), but it was done in complete ignorance of viruses (or indeed any kind of microbe) and immunology.

It was not until the work of Pasteur 100 years later that the general principle governing vaccination emerged – altered preparations of microbes could be used to generate enhanced immunity against the fully virulent organism.

Thus Pasteur's dried rabies-infected rabbit spinal cords and heated anthrax bacilli were the true forerunners of today's vaccines, whereas Jenner's animal-derived (i.e. 'heterologous') vaccinia virus has had no real successors.

Even Pasteur did not have a proper understanding of immunological memory or the functions of the lymphocyte, which had to wait another half century.

Finally, with Burnet's clonal selection theory (1957) and the discovery of T and B lymphocytes (1965), the key mechanism became clear.

In any immune response, the antigen(s) induces clonal expansion in specific T and/or B cells, leaving behind a population of memory cells. These enable the next encounter with the same antigen(s) to induce a secondary response, which is more rapid and effective than the normal primary response (Fig. 18.1).

While for many infections the primary response may be too slow to prevent serious disease, if the individual has been exposed to antigens from the organism in a vaccine before encountering the pathogenic organism, the expanded population of memory cells and raised levels of specific antibody are able to protect against disease.

Q. Rabies is one of the few diseases in which active immunization may be carried out after the individual becomes infected. What particular feature of rabies infection makes this a reasonable treatment?

A. The time between infection and the development of the disease is long, so an effective immune response has time to develop before virus reaches the CNS to produce symptoms.

Vaccines have two major effects

First, vaccines protect individuals against disease, and second, if there are sufficient immune individuals in a

Principles of vaccination
vaccination stimulates protective adaptive immune response – this priming expands the pool of specific memory cells
subsequent natural infection induces fast, vigorous responses
harmless forms of the immunogen are used to vaccinate – these can be killed or modified living organisms, subcellular fragments, or toxoids
vaccine adjuvants enhance immune responses
vaccines must be safe, affordable, and produce herd immunity

Fig. 18.1 Principles of vaccination.

population, transmission of the infection is prevented. This is known as **herd immunity**.

The proportion of the population that needs to be immune to prevent epidemics occurring depends on the nature of the infection:

- if the organism is highly infectious so that one individual can rapidly infect several non-immune individuals, as is the case for measles, a high proportion of the population must be immune to maintain herd immunity;
- if the infection is less readily transmitted, immunity in a lower proportion of the population may be sufficient to prevent disease transmission.

Effective vaccines must be safe to administer, induce the correct type of immunity, and be affordable by the population at which they are aimed. For many diseases, this has been achieved with brilliant success, but for others there is no vaccine whatsoever. This chapter is mainly concerned with the reasons for this disparity.

A WIDE RANGE OF ANTIGEN PREPARATIONS ARE IN USE AS VACCINES

In general, the more antigens of the microbe retained in the vaccine, the better, and living organisms tend to be more effective than killed organisms (see below). Exceptions to this rule are:

- diseases where a toxin is responsible for the pathology – in this case the vaccine can be based on the toxin alone;
- a vaccine in which microbial antigens are expressed in another type of cell, which acts as a vector.

Fig. 18.2 lists the main antigenic preparations currently available.

Live vaccines can be natural or attenuated organisms

Natural live vaccines have rarely been used

Apart from vaccinia, no other completely natural organism has ever come into standard use. However:

- bovine and simian rotaviruses have been tried in children;
- the vole tubercle bacillus was once popular against tuberculosis; and
- in the Middle East and Russia *Leishmania* infection from mild cases is reputed to induce immunity.

It is possible that another good heterologous vaccine will be found, but the safety problems will be considerable.

Attenuated live vaccines have been highly successful

The preferred strategy has been to attenuate a human pathogen, with the aim of diminishing its virulence while retaining the desired antigens.

This was first done successfully by Calmette and Guérin with a bovine strain (*Mycobacterium bovis*) of *Mycobacterium tuberculosis*, which during 13 years (1908– 1921) of culture in vitro changed to the much less virulent form now known as BCG (bacille Calmette– Guérin), which has at least some protective effect against tuberculosis.

The real successes have been with viruses, starting with the 17D strain of yellow fever virus obtained by passage in mice and chicken embryos (1937), and followed by a roughly similar approach with polio, measles, mumps, and rubella (Fig. 18.3).

Q. Why would passage of a virus in a non-human species be a rational way of developing a vaccine for use in humans?

A. The selective pressure on the virus to retain genes needed for virulence in humans and for human-to-human transfer is removed. Consequently some variants may lose these genes and are at no disadvantage in the animal, but will retain antigenicity for use as attenuated vaccine strains.

Just how successful the vaccines for polio, measles, mumps, and rubella are is shown by the decline in these four diseases between 1950 and 1980 (Fig. 18.4).

ATTENUATED MICROORGANISMS ARE LESS ABLE TO CAUSE DISEASE IN THEIR NATURAL HOST – Attenuation ‘changes’ microorganisms to make them less able to grow and cause disease in their natural host.

In early attenuated organisms, ‘changed’ meant a purely random set of mutations induced by adverse conditions of growth.

Vaccine candidates were selected by constantly monitoring for retention of antigenicity and loss of virulence – a tedious process.

When viral gene sequencing became possible it emerged that the results of attenuation were widely

The main antigenic preparations

type of antigen		vaccine examples
living organisms	natural	vaccinia (for smallpox) vole bacillus (for tuberculosis; historical)
	attenuated	polio (Sabin; oral polio vaccine)*, measles*, mumps*, rubella*, yellow fever 17D, varicella-zoster (human herpesvirus 3), BCG (for tuberculosis)*
intact but non-living organisms	viruses	polio (Salk)*, rabies, influenza, hepatitis A, typhus
	bacteria	*pertussis, typhoid, cholera, plague
subcellular fragments	capsular polysaccharides	pneumococcus, meningococcus, *Haemophilus influenzae*
	surface antigen	hepatitis B*
toxoids		tetanus*, diphtheria*
recombinant DNA-based	gene cloned and expressed	hepatitis B (yeast-derived)*
	genes expressed in vectors	experimental
	naked DNA	experimental
anti-idiotype		experimental

*standard in most countries

Fig. 18.2 A wide range of antigenic preparations are used as vaccines.

Live attenuated vaccines

disease		remarks
viruses	polio	types 2 and 3 may revert; also killed vaccine
	measles	80% effective
	mumps	
	rubella	now given to both sexes
	yellow fever	stable since 1937
	varicella-zoster	mainly in leukemia
	hepatitis A	also killed vaccine
bacteria	tuberculosis	stable since 1921; also some protection against leprosy

Fig. 18.3 Attenuated vaccines are available for many, but not all, infections. In general it has proved easier to attenuate viruses than bacteria.

divergent. An example is the divergence between the three types of live (Sabin) polio vaccine:

- type 1 polio has 57 mutations and has almost never reverted to wild type;
- types 2 and 3 vaccines depend for their safety and virulence on only two key mutations – frequent reversion to wild type has occurred, in some cases leading to outbreaks of paralytic poliomyelitis.

VACCINIA VIRUS IS BEING USED AS A VECTOR FOR ANTIGENS OF MICROORGANISMS SUCH AS HIV AND MALARIA – Vaccinia virus is not an attenuated smallpox virus (present-day vaccinia virus strains being most closely related to rodent pox viruses), nor is vaccination against smallpox carried out. However, vaccinia virus is returning to use as a vector for antigens of other microorganisms such as HIV and malaria and is yielding very interesting information on the effects of attenuation.

Pox viruses contain approximately 200 genes of which about two-thirds are essential for growth in mammalian cells in vitro.

The modified vaccinia Ankara strain (MVA) was grown for a prolonged period in avian cells and has deleted genes required to complete the replicative cycle in mammalian cells, but it can still infect them and is a safe and immunogenic smallpox vaccine. Those genes not essential for replication of the virus are mostly concerned with evasion of host responses and virulence.

Virulence is the ability to replicate efficiently and disseminate widely within the body, with pathological consequences.

All pox viruses contain virulence genes and many of these mimic or interfere with cytokine and chemokine function. Some of these have sequence homology to their mammalian counterparts and others do not. Most vaccinia

Effect of vaccination on the incidence of viral disease

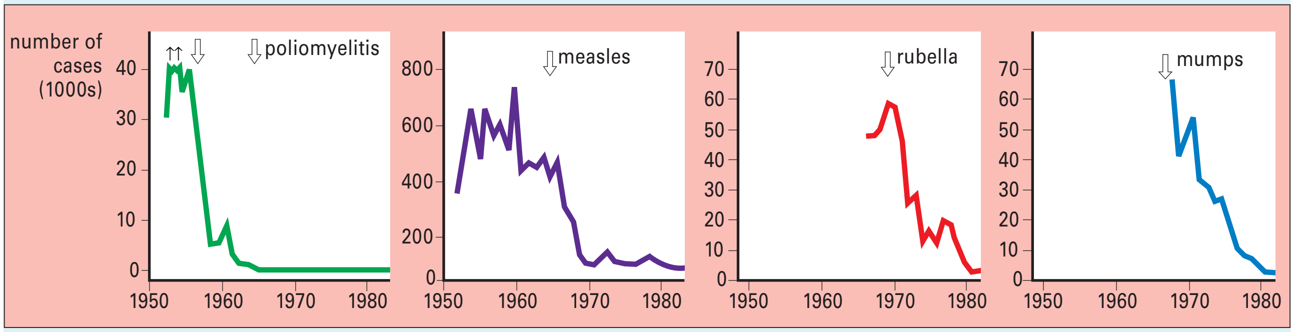

Fig. 18.4 The effect of vaccination on the incidence of various viral diseases in the USA has been that most infections have shown a dramatic downward trend since the introduction of a vaccine (arrows).

strains lack some of these genes while smallpox virus itself contains almost all of them.

It is now possible to use recombinant DNA technology to remove virulence genes completely or to introduce site-directed mutations to inactivate them. A problem remains due to the large size of pox viruses, where deriving an ideal vector is a major enterprise that has not yet been achieved.

Q. It is often found that attenuated viruses that are avirulent do not make good vaccines. Why should this be?

A. Inflammation induced by damage to the host has an adjuvant effect, leading to more effective presentation of the vaccine antigens.

Killed vaccines are intact but non-living organisms

Killed vaccines are the successors of Pasteur's killed vaccines mentioned above:

- some are very effective (rabies and the Salk polio vaccine);
- some are moderately effective (typhoid, cholera, and influenza);
- some are of debatable value (plague and typhus); and
- some are controversial on the grounds of toxicity (pertussis).

Fig. 18.5 lists the main killed vaccines in use today. Some of these will undoubtedly be replaced by attenuated or subunit vaccines.

Inactivated toxins and toxoids are the most successful bacterial vaccines

The most successful of all bacterial vaccines – tetanus and diphtheria – are based on inactivated exotoxins (Fig. 18.6), and in principle the same approach can be used for several other infections.

Subunit vaccines and carriers

Aside from the toxin-based vaccines, which are subunits of their respective microorganisms, a number of other vaccines are in use that make use of antigens either purified from microorganisms or produced by recombinant DNA technology (Fig. 18.7).

Killed (whole organism) vaccines

disease		remarks
viruses	polio	preferred in Scandinavia; safe in immunocompromised
	rabies	can be given post-exposure, with passive antiserum
	influenza	strain-specific
	hepatitis A	also attenuated vaccine
bacteria	pertussis	potential to cause brain damage (controversial)
	typhoid	about 70% protection
	cholera	protection dubious; may be combined with toxin subunit
	plague	short-term protection only
	Q fever	good protection

Fig. 18.5 The principal vaccines using killed whole organisms.

Acellular pertussis vaccine consisting of a small number of proteins purified from the bacterium is available and has been shown to be effective and less toxic.

Although hepatitis B surface antigen is immunogenic when given with alum adjuvant (see below), the bacterial capsular polysaccharides of *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Haemophilus influenzae B* are relatively poorly immunogenic and often do not induce IgG responses or long-lasting protection.

Q. Why do polysaccharide antigens not induce IgG responses or lasting immunity?

A. Polysaccharide antigens are not presented to helper T cells, so they do not induce class switching, affinity maturation, or generate memory T cells (see Chapters 7 and 8).

A significant improvement in the efficacy of subunit vaccines has been obtained by conjugating the purified

Toxin-based vaccines

organism	vaccine	remarks
Clostridium tetani	inactivated toxin (formalin)	three doses, alum-precipitated; boost every 10 years
Corynebacterium diphtheriae		usually given with tetanus
Vibrio cholerae	toxin, B subunit	sometimes combined with whole killed organisms
Clostridium perfringens	inactivated toxin (formalin)	for newborn lambs

Fig. 18.6 The principal toxin-based vaccines. Note that there are no vaccines against the numerous staphylococcal and streptococcal exotoxins, or against bacterial endotoxins such as lipopolysaccharides.

Subunit vaccines

	organism	remarks
virus	hepatitis B virus	surface antigen can be purified from blood of carriers or produced in yeast by recombinant DNA technology
bacteria	*Neisseria meningitidis*	capsular polysaccharides or conjugates of group A and C are effective; group B is non-immunogenic
	Streptococcus pneumoniae	84 serotypes; capsular polysaccharide vaccines contain 23 serotypes; conjugates with five or seven bacterial serotypes are being tested
	Haemophilus influenzae B	good conjugate vaccines now available

Fig. 18.7 Conjugate vaccines are replacing pure polysaccharides. *N. meningitidis* type B is non-immunogenic in humans because the capsular polysaccharide cross-reacts with self carbohydrates towards which the host is immunologically tolerant.

polysaccharides to carrier proteins such as tetanus or diphtheria toxoid. These protein carriers are presumed to recruit helper T cells and the conjugates induce IgG antibody responses and more effective protection.

Subunit vaccines are more expensive than conventional killed vaccines so initially their use is likely to be restricted to the developed world.

Antigens can be made synthetically or by gene cloning

Where it can be shown that a small peptide is protective it may be more convenient to make it synthetically. However, although there are animal models of viral infection or tumor immunotherapy where immunization with a synthetic MHC-binding peptide epitope can be effective, so far there is no human vaccine based on this approach.

Q. What fundamental problem associated with a peptide antigen vaccine arises when it is used in humans as opposed to animal models of disease?

A. In an outbred population (e.g. humans) a single peptide will usually be effective as an immunogen in only a proportion of the population because only certain MHC types can present the antigen and some people may not have a suitable MHC molecule (i.e. MHC restriction, see Chapter 7).

So far, therefore, the alternative approach of producing antigens by cloning their genes into a suitable expression vector has been much more useful because all possible MHC-binding epitopes within the antigen are retained:

- this approach has been highly successful with the hepatitis B surface (HBs) antigen, cloned into yeast and now replacing the first-generation HBs vaccine, which was laboriously purified from the blood of hepatitis B carriers; it has also brought down the cost of the vaccine;
- a second example of this approach is the recently tested vaccine against human papilloma virus types 16 and 11 – the L1 capsid antigen is the main component of the vaccine and, when expressed in yeast, the recombinant protein assembles into protein polymers, forming virus-like particles (VLPs) that are highly immunogenic.

An attractive feature of this approach is that further sequences can be added – for example, selected B and T cell epitopes can be combined in various ways to optimize the resulting immune response.

It is important to remember that, whereas B cells respond to the three-dimensional shape of antigens, T cells recognize linear sequences of amino acids (see Chapters 3 and 5). Peptides can therefore function well as T cell epitopes, but cannot readily mimic discontinuous B cell epitopes.

Even where a B cell determinant is linear, antibodies raised against the free flexible peptide, which adopts many configurations, do not bind optimally to the sequence in the way that they do when it is present as a more rigid structure within the native protein molecule.

Anti-idiotype vaccines could be used when the original antigen was unsuitable

Anti-idiotype vaccine is the only type of vaccine for which immunological thinking has been entirely responsible. The idea is to use monoclonal antibody (mAb) technology to make large amounts of anti-idiotype (anti-Id) against the V region (idiotype) of an antibody of proven protective value.

The anti-Id, if properly selected, would then have a three-dimensional shape similar to the original immunizing antigen and could be used in place of it (Fig. 18.8).

Although often dismissed as 'armchair immunology', this strategy could have real value where the original antigen is not itself suitable (i.e. is not immunogenic or is toxic). Polysaccharides are one example, and the lipid A region of bacterial endotoxin (lipopolysaccharide [LPS]) is another. The advantage of the mAb would be that because it is a protein it should induce memory, which polysaccharides and lipids normally do not, unless conjugated to a carrier.

Anti-idiotype antibodies as vaccines

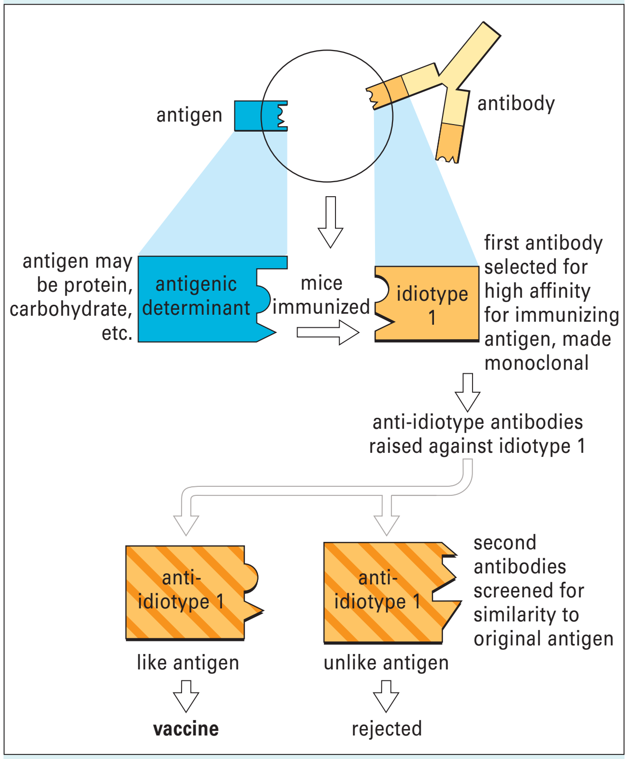
antigen
antibody
antigen may be protein, carbohydrate, etc.
antigenic determinant
mice immunized
idiotype 1
first antibody selected for high affinity for immunizing antigen, made monoclonal
anti-idiotype antibodies raised against idiotype 1
anti-idiotype 1
anti-idiotype 1
second antibodies screened for similarity to original antigen
like antigen
unlike antigen
vaccine
rejected

Fig. 18.8 Monoclonal antibody technology and the discovery of the 'idiotype network' have meant that immunoglobulins can now be used as 'surrogate' antigens. In the case of a carbohydrate or lipid antigen, this allows a protein 'copy' to be made, which may have some advantages as a vaccine. Note also that a monoclonal anti-idiotype can act as a mimic of a discontinuous B cell epitope.

ADJUVANTS ENHANCE ANTIBODY PRODUCTION

During work in the 1920s on the production of animal sera for human therapy, it was discovered that certain substances, notably aluminum salts, added to or emulsified with an antigen, greatly enhance antibody production – that is, they act as adjuvants.

Aluminum hydroxide is still widely used with, for example, diphtheria and tetanus toxoids.

With modern understanding of the processes leading to lymphocyte triggering and the development of memory, considerable efforts have been made to produce better adjuvants, particularly for T cell-mediated responses. Fig. 18.9 gives a list of these, but it should be stressed that none of the new adjuvants is yet accepted for routine human use.

Adjuvants concentrate antigen at appropriate sites or induce cytokines

It appears that the effect of adjuvants is due mainly to two activities:

- the concentration of antigen in a site where lymphocytes are exposed to it (the 'depot' effect); and
- the induction of cytokines that regulate lymphocyte function.

Aluminum salts probably have a predominantly depot function, inducing small granulomas in which antigen is retained.

Newer formulations such as liposomes and immune-stimulating complexes (ISCOMs) achieve the same purpose by ensuring that antigens trapped in them are delivered to antigen-presenting cells (APCs).

Particulate antigens such as virus-like particles (polymers of viral capsid proteins containing no viral DNA or RNA) are highly immunogenic and have the useful property that they may also induce cross-priming (i.e. enter the MHC class 1 processing pathway though not synthesized within the APC, see Chapter 7).

Q. Many bacterial carbohydrates and glycolipids are good adjuvants, even though they are not good immunogens. Why should this be so?

A. The discovery of Toll-like receptors (TLRs, see Fig. 6.24) and other pattern recognition receptors (PRRs, see Chapter 15), such as lectin-like receptors for carbohydrates (see Fig. 9.18), has provided an explanation for the long-known efficacy of many bacterial products as adjuvants. It is clear that they act mainly by binding to PRRs and stimulating the formation of appropriate cytokines by APCs.

Ligation of different PRRs may bias the response toward TH1 or TH2 cytokine production.

Not surprisingly cytokines themselves have been shown to be effective adjuvants, particularly when coupled directly to the antigen. Cytokines may be particularly useful in immunocompromised patients (see below, Fig. 18.11), who often fail to respond to normal vaccines. It is hoped that they might also be useful in directing the immune response in the desired direction – for example in diseases where only TH1 (or TH2) cell memory is wanted.

Adjuvants

adjuvant type	routinely used in humans	experimental* or too toxic for human use†
inorganic salts	aluminum hydroxide (alhydrogel) aluminum phosphate calcium phosphate	beryllium hydroxide
delivery systems		liposomes* ISCOMs* block polymers slow-release formulations*
bacterial products	*Bordetella pertussis* (with diphtheria, tetanus toxoids)	BCG *Mycobacterium bovis* and oil† (complete Freund's adjuvant) muramyl dipeptide (MDP)†
natural mediators (cytokines)		IL-1 IL-2 IL-12 IFNγ

ISCOMs, immune-stimulating complexes

Fig. 18.9 A variety of foreign and endogenous substances can act as adjuvants, but only aluminum and calcium salts and pertussis are routinely used in clinical practice.

MOST VACCINES ARE GIVEN BY INJECTION

Most vaccines are still delivered by injection, but this is a risky method in developing countries, where re-use of needles and syringes may transmit disease, particularly HIV.

Because most organisms enter via mucosal surfaces, mucosal immunization makes logical sense and the success of oral polio vaccine indicates that it can be made to work. However, although live polio works when delivered orally, most killed vaccines do not.

Q. What key problems would one expect to be associated with oral vaccines?

A. Antigens may be broken down by passage through the stomach and digestive system, but, more problematically, the intestinal immune system is designed to generate tolerance rather than an immune response against food antigens.

Immunization only occurs when pathogenic organisms invade the gut wall. This can be mimicked by providing an adjuvant. Toxins from pathogenic intestinal organisms (cholera and *Escherichia coli*) have been the most studied intestinal adjuvants. Because the native toxins are extremely potent, partially inactivating mutations have been introduced to prevent excessive intestinal stimulation. Although these adjuvants work in experimental models, it is difficult to achieve a reproducible balance between:

- adequate stimulation of an immune response; and
- excessive gut inflammation.

An alternative is to use recombinant bacteria engineered to express antigens of interest, but the same difficulty applies:

- if the bacteria are non-pathogenic they may not immunize;
- if the bacteria are too pathogenic they may cause unpleasant symptoms.

Several recombinant and partially attenuated salmonella strains have been used experimentally to explore this vaccine strategy.

Similar problems relate to nasal immunization, usually tried against upper respiratory infections such as influenza or respiratory syncytial viruses (RSV). No nasal vaccine has entered routine use because of:

- difficulties in balancing attenuation against immunogenicity in the case of live RSV vaccine strains;
- the need for an adjuvant for an inactivated nasal influenza virus;
- safety worries because of the proximity of the nasal mucosa to the brain though the cribriform plate.

More recently the transdermal route has attracted attention. The ‘gene gun’ propels microscopic gold particles coated with DNA into the epidermis using a pressurized gas device (see below), but can be used with a variety of particles and proteins or other immunogens. The target is epidermal Langerhans’ cells. Many antigens are also well absorbed and immunogenic if applied in a suitable vehicle to lightly abraded skin or even through the intact epidermis. As yet these methods are experimental, but it is hoped that eventually injection will be replaced by less invasive and safer methods.

VACCINE EFFICACY NEEDS TO REVIEWED FROM TIME TO TIME

To be introduced and approved, a vaccine must obviously be effective, and the efficacy of all vaccines is reviewed from time to time. Many factors affect it.

An effective vaccine must **induce the right sort of immunity**:

- antibody for toxins and extracellular organisms such as *Streptococcus pneumoniae*;
- cell-mediated immunity for intracellular organisms such as the tubercle bacillus.

Where the ideal type of response is not clear (as in malaria, for instance), designing an effective vaccine becomes correspondingly more difficult.

An effective vaccine must also be:

- **stable on storage** – this is particularly important for living vaccines, which normally require to be kept cold (i.e. a complete 'cold chain' from manufacturer to clinic, which is not always easy to maintain).
- have **sufficient immunogenicity** – with non-living vaccines it is often necessary to boost their immunogenicity with an adjuvant (see above, p. 330).

Live vaccines are generally more effective than killed vaccines.

Antibody responses to live and killed polio vaccine

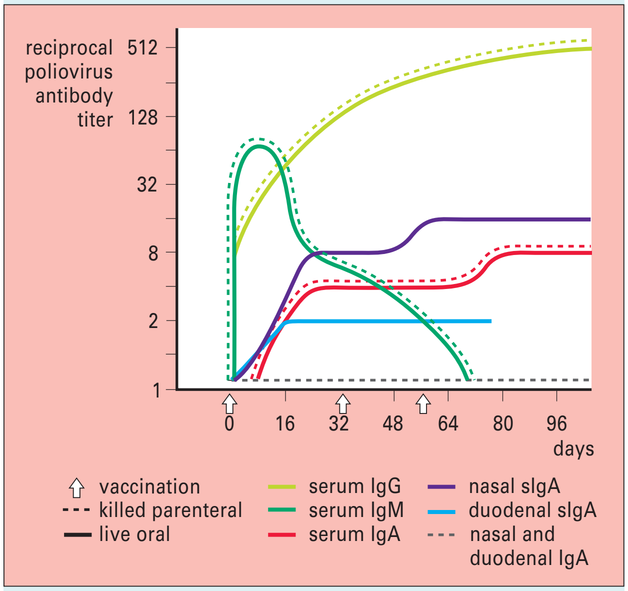

Fig. 18.10 The antibody response to orally administered live attenuated polio vaccine (solid lines) and intramuscularly administered killed polio vaccine (broken lines). The live vaccine induces the production of secretory IgA (sIgA) in addition to serum antibodies, whereas the killed vaccine induces no nasal or duodenal sIgA. As sIgA is the immunoglobulin of the mucosa-associated lymphoid tissue (MALT) system (see Chapter 2), the live vaccine confers protection at the portal of entry of the virus, the gastrointestinal mucosa. (Courtesy of Professor JR Pattison, Ch. 26 in Brostoff J, et al., eds. Clinical immunology. London: Mosby; 1991)

Induction of appropriate immunity depends on the properties of the antigen

Living vaccines have the great advantage of providing an increasing antigenic challenge that lasts days or weeks, and inducing it in the right site – which in practice is most important where mucosal immunity is concerned (Fig. 18.10).

Live vaccines are likely to contain the greatest number of microbial antigens, but safety is an issue in a time of increasing concern about the side effects of vaccines.

Vaccines made from whole killed organisms have been used, but because a killed organism no longer has the advantage of producing a prolonged antigenic stimulus, killed vaccines now frequently replaced by purified components of the organism (subunit vaccines). These may suffer from several problems:

- purified subunits may be relatively poorly immunogenic and require adjuvants;
- the smaller the antigen, the more histocompatibility complex (MHC) restriction may be a problem (see Chapters 5 and 7); and
- purified polysaccharides are typically thymus independent – they do not bind to MHC and therefore do not immunize T cells.

These problems have been overcome in vaccines that are routinely used in humans by the use of adjuvants and by coupling polysaccharides either to:

- a standard protein carrier such as tetanus toxoid; or
- to a protein from the immunizing organism such as the outer membrane protein of pneumococci.

MHC restriction is probably more of a hypothetical than real difficulty because most candidate vaccines are large enough to contain several MHC-binding epitopes. Nevertheless, even the most effective vaccines often fail to immunize every individual – for example, about 5% of individuals fail to seroconvert after the full course of hepatitis B vaccine.

Most of the vaccines in routine use in humans depend on the induction of protective antibody. However, for many important infections, particularly of intracellular organisms (e.g. tuberculosis, malaria, and HIV infection), cellular immune responses are important protective mechanisms.

In recent years there has therefore been much effort to develop vaccines to induce immunity of both CD4 and CD8 T cells. So far the use of DNA and viral vectors have been the routes most commonly explored because both of these strategies lead to the production of antigens within cells and therefore the display of processed peptide epitopes on MHC molecules.

Although these methods, particularly combined in prime-boost regimens, have been highly effective in experimental animal models, so far in humans it has proved difficult to induce high frequencies of long-lasting antigen-specific T memory cells.

Even in experimental animals the duration of protection may not be long-lasting, perhaps because protection by cellular mechanisms requires activated effector cells rather than resting memory cells. Such cells are not well maintained in the absence of antigen.

VACCINE SAFETY IS AN OVERRIDING CONSIDERATION

Vaccine safety is of course a relative term, with minor local pain or swelling at the injection site, and even mild fever, being generally acceptable. More serious complications may stem from the vaccine or from the patient (Fig. 18.11):

- vaccines may be contaminated with unwanted proteins or toxins, or even live viruses;
- supposedly killed vaccines may not have been properly killed;
- attenuated vaccines may revert to the wild type;
- the patient may be hypersensitive to minute amounts of contaminating protein, or immunocompromised, in which case any living vaccine is usually contraindicated.

Although serious complications are very rare, vaccine safety has now become an overriding consideration, in part because of the very success of vaccines:

- because many childhood infectious diseases have become uncommon in developed countries, the populations of these countries are no longer aware of the potentially devastating effects of infectious diseases;
- unlike most drugs, vaccinations are given to people who have previously been perfectly well;
- the public is becoming increasingly aware of the possibilities of profitable litigation and companies correspondingly more defensive.

Q. Why is vaccine safety perceived as a less important issue when a vaccine is first introduced?

A. At the time of vaccine introduction the prevalence and danger associated with the infection is so great that any risks associated with the vaccine are relatively small.

The MMR controversy resulted in measles epidemics

The difficulties concerning vaccine safety are well illustrated by the recent controversy over MMR (measles, mumps, and rubella triple vaccine), though anti-vaccine movements in the UK date from a few years after the introduction of smallpox vaccination by Jenner in 1796.

In 1998 a paper was published that received wide publicity in the UK media, purporting to support an association between MMR vaccination and the development of autism and chronic bowel disease. Although a large amount of subsequent work failed to substantiate these findings and the interpretation of the original paper was withdrawn, take-up of MMR in the UK and Ireland declined over several years and epidemics of measles occurred because of declining herd immunity.

At the present time, the introduction of a five-valent vaccine containing diphtheria and tetanus toxoids, acellular pertussis, *Haemophilus influenzae* type B, and inactivated polio virus threatens to lead to decreased take-up in the UK. Here the argument seems to be that giving five immunogens simultaneously may be 'too much' for the delicate immune system of infants, though the vaccine has been shown to be safe and effective. Of course, because most of the vaccines within it are subunits, except for inactivated polio, the whole vaccine contains fewer antigens than the live bacteria and other organisms that the infant will encounter every day.

A rotavirus vaccine will prevent many infants from dying in developing countries

The story of a recent rotavirus vaccine is also instructive. In 1998 a live rotavirus vaccine was tested on large numbers of infants in the USA. Approximately 1 in 2500 vaccinated infants developed intussusception (a potentially fatal bowel condition) and the vaccine was withdrawn by the manufacturer. However, although this complication rate was unacceptable in the USA, the vaccine was highly

Safety problems with vaccine

type of vaccine	potential safety problems	examples
attenuated vaccines	reversion to wild type	especially polio types 2 and 3
	severe disease in immunodeficient patients	vaccinia, BCG, measles
	persistent infection	varicella-zoster
	hypersensitivity to viral antigens	measles
	hypersensitivity to egg antigens	measles, mumps
killed vaccines	vaccine not killed	polio accidents in the past
	yeast contaminant	hepatitis B
	contamination with animal viruses	polio
	contamination with endotoxin	pertussis

Fig. 18.11 The potential safety problems encountered with vaccines emphasize the need for continuous monitoring of both production and administration.

effective and would have saved many lives if used in developing countries where as many as 1 in 200 infants die from rotavirus diarrhea.

Complicating the debate further are more recent findings that the real risk of intussusception in vaccinated infants may in fact be much lower than appeared in the original trial and even lower if the vaccine is given in the first 2 months of life. Two newer rotavirus vaccines produced by other manufacturers are now in late-phase trials. Clearly vaccine safety raises many practical and ethical issues.

New vaccines can be very expensive

Although vaccination can safely be considered the most cost-effective treatment for infectious disease, new vaccines may be very expensive. The initial high cost is necessary to recoup the enormous development costs (US$200–400 million).

A good example is the recombinant hepatitis B vaccine, which was initially marketed in 1986 at US$150 for three doses. Although the cost has decreased greatly, even $1 is beyond the health budget of many of the world's poorer nations.

By contrast, the cost of the six vaccines included in the World Health Organization Expanded Program on Immunization (diphtheria, tetanus, whooping cough, polio, measles, and tuberculosis) is less than $1. The actual cost of immunizing a child is several times greater than this because it includes the cost of laboratories, transport, the cold chain, personnel, and research.

The Children's Vaccine Initiative, set up in 1990, is a global forum that aims to bring together vaccine researchers, development agencies, governments, donors, commercial and public sector vaccine manufacturers to seek means of delivering vaccines to the world's poorest populations who most need them.

VACCINES IN GENERAL USE HAVE VARIABLE SUCCESS RATES

The vaccines in standard use worldwide are listed in Fig. 18.12. Four of them – polio, measles, mumps, and rubella – are so successful that these diseases are earmarked for eradication early in the 21st century. If this happens, it will be an extraordinary achievement, because mathematical modeling suggests that they are all more 'difficult' targets for eradication than smallpox was.

In the case of polio, where reversion to virulence of types 2 and 3 can occur, it has been suggested that it will be necessary to switch to the use of killed virus vaccine for some years, so that virulent virus shed by live virus-vaccinated individuals is no longer produced (as mentioned above, the new five-valent vaccine recently introduced in the UK and elsewhere contains inactivated poliovirus in accordance with this suggestion).

For a number of reasons, other vaccines are less likely to lead to eradication of disease. These include:

- **the carrier state** – eradication of hepatitis B would be a major triumph, but it will require the breaking of the carrier state, especially in the Far East, where mother to child is the normal route of infection;
- **suboptimal effectiveness** – effectiveness of BCG varies markedly between countries, possibly due to variation in environmental mycobacterial species (tuberculosis is increasing, especially in patients with immune deficiency or AIDS), and the pertussis vaccine is only about 70% effective;
- **side effects** – the MMR vaccine was suspected of having side effects, reducing the public's willingness to be vaccinated;
- **free-living forms and animal hosts** – the free-living form of tetanus will presumably survive indefinitely and it will not be possible to eradicate diseases that also have an animal host, such as yellow fever.

One of the future problems is going to be maintaining awareness of the need for vaccination against diseases that seem to be disappearing, while, as the reservoir of infection diminishes, cases tend to occur at a later age, which with measles and rubella could actually lead to worse clinical consequences.

Some vaccines are reserved for special groups only

In the developed world BCG and hepatitis B fall into this category, but some vaccines will probably always be confined to selected populations – travelers, nurses,

Vaccines in general use

disease	vaccine	remarks
tetanus diphtheria pertussis polio (DTPP)	toxoid toxoid killed whole killed (Salk) or attenuated (Sabin)	given together in three doses between 2 and 6 months; tetanus and diphtheria boosted every 10 years
measles mumps rubella	attenuated	given together (MMR) at 12–18 months
Haemophilus influenzae type b	polysaccharide	new; may be given with DTPP

Fig. 18.12 Vaccines that are currently given, as far as is possible, to all individuals.

Vaccines restricted to certain groups

disease	vaccine	eligible groups
tuberculosis	BCG	tropics – at birth; UK – 10–14 years; USA – at-risk only
hepatitis B	surface antigen	at risk (medical, nursing staff, etc.); drug addicts; male homosexuals; known contacts of carriers
rabies	killed	at risk (animal workers); postexposure
meningitis yellow fever typhoid, cholera hepatitis A	polysaccharide attenuated killed or mutant killed or attenuated	travelers
influenza	killed	at risk; elderly
pneumococcal pneumonia	polysaccharide	elderly
varicella-zoster	attenuated	leukemic children

Fig. 18.13 Vaccines that are currently restricted to certain groups.

the elderly, etc. (Fig. 18.13). In some cases this is because of:

- geographical restrictions (e.g. yellow fever);
- the rarity of exposure (e.g. rabies);
- problems in producing sufficient vaccine in time to meet the demand (e.g. each influenza epidemic is caused by a different strain, requiring a new vaccine).

A vaccine effective against all strains of influenza would be of tremendous value. However, both the hemagglutinin and neuraminidase antigens, which together make up the outer layer of the virus and are the antigens of importance in the vaccine, are subject to variation.

Vaccines for parasitic and some other infections are only experimental

Some of the most intensively researched vaccines are those for the major tropical protozoal and worm infections (see Chapter 15). However, none has come into standard use, and some have argued that none will because none of these diseases induces effective immunity and 'you cannot improve on nature'.

Nevertheless, extensive work in laboratory animals has shown that vaccines against malaria, leishmaniasis, and schistosomiasis are perfectly feasible, and there is a moderately effective vaccine against babesia in dogs. In cattle an irradiated vaccine against the lungworm has been in veterinary use for decades.

Q. Why, even with effective vaccines, will it be impossible to eradicate many parasitic diseases?

A. Many parasites have alternative species to humans as their host (see Chapter 15).

It remains possible, however, that the parasitic diseases of humans are uniquely difficult to treat, partly because of the polymorphic and rapidly changing nature of many parasitic antigens. For example:

- none of the small animal models of malaria shows such extensive antigenic variation as *Plasmodium falciparum*, the protozoon causing malignant tertian malaria in humans;
- similarly, rats appear to be much easier to immunize against schistosomiasis than other animals, including possibly humans.

Part of the problem is that these parasites are usually not in their natural host in the laboratory.

Several trials of clinical malaria vaccine have been published, using antigens derived from either the liver or the blood stage, with only very moderate success. Malaria is unusual in that its life cycle offers a variety of possible targets for vaccination (Fig. 18.14).

A problem with these chronic parasitic diseases is that of immunopathology. For example, the symptoms of *Trypanosoma cruzi* infection (Chagas' disease) are largely due to the immune system (i.e. autoimmunity). A bacterial parallel is leprosy, where the symptoms are due to the (apparent) overreactivity of TH1 or TH2 cells. A vaccine that boosted immunity without clearing the pathogen could make these conditions worse. Another example of this unpleasant possibility is with dengue, where certain antibodies enhance the infection by allowing the virus to enter cells via Fc receptors. Enhancing antibodies have also been reported in an experimental model of a transmission-blocking anti-malarial vaccine.

Some viral and bacterial vaccines are also in the experimental category

Other viral and bacterial vaccines that are also experimental are:

- cholera toxin;
- attenuated shigella;
- Epstein–Barr virus surface glycoprotein.

A recombinant vaccine against human papilloma virus type 16 consisting of virus-like particles has recently undergone trials and appears to be highly effective in preventing genital infection. It remains to be determined whether such a vaccine will also reduce the incidence of cervical carcinoma.

Malaria vaccine strategies

stage	vaccine strategy
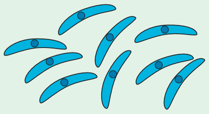 sporozoites	sporozoite vaccine to induce blocking antibody, already field-tested in humans
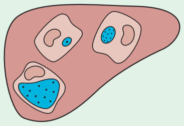 liver stage	sporozoite vaccine to induce cell-mediated immunity to liver stage
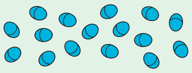 merozoites	merozoite (antigen) vaccine to induce blocking antibody
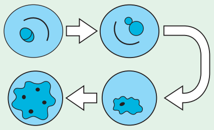 asexual erythrocyte stage	asexual stage (antigen) vaccine to induce other responses to red cell stage, and against toxic products ('anti-disease' vaccine)
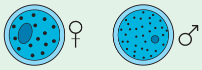 gametocytes	vaccines to interrupt sexual stages – 'transmission-blocking' vaccine
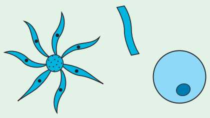 gametes	

Fig. 18.14 A number of different approaches to malaria vaccines are being investigated, reflecting the complexity of both the life cycle of malaria and immunity to it.

For many diseases there is no vaccine available

No vaccine is currently available for many serious infectious diseases (Fig. 18.15). Headed by HIV infection (see Chapter 17), these represent the major challenge for research and development in the coming decade.

PASSIVE IMMUNIZATION CAN BE LIFE-SAVING

Driven from use by the advent of antibiotics, the idea of injecting preformed antibody to treat infection is still valid for certain situations (Fig. 18.16). It can be life-saving when:

- toxins are already circulating (e.g. in tetanus, diphtheria, and snake-bite);
- high-titer specific antibody is required, generally made in horses, but occasionally obtained from recovered patients.

At the opposite end of the scale, normal pooled human immunoglobulin contains enough antibody against common infections for a dose of 100–400 mg IgG to protect hypogammaglobulinemic patients for a month. Over 1000 donors are used for each pool, and the sera must be screened for HIV and hepatitis B and C.

The use of specific monoclonal antibodies, though theoretically attractive, has not yet proved to be an improvement on traditional methods, and their chief application to infectious disease at present remains in diagnosis. This may change as human monoclonal antibodies become more readily available (and less expensive), either through cell culture or protein engineering.

In tumor immunotherapy a monoclonal antibody to CD20 has found a place in the treatment of B cell lymphomas. Immunotherapy using similar means of B cell depletion has also had beneficial effects in rheumatoid arthritis and systemic lupus erythematosus (SLE, see Chapter 20).

Antibody genes can now be engineered to form Fab, single chain Fv, or VH fragments (see Chapter 3). Libraries of these can be expressed in recombinant phages and screened against antigens of interest. Selected antibody fragments can be produced in bulk in bacteria, yeasts, or mammalian cells, for use in vitro or in vivo.

Differently sized antibody fragments may have different uses – in general the smaller the fragment the better the penetration into tissues, but also the shorter the half-life. Bacteria engineered to produce antibody fragments may in the future be used to target protective antibody to a particular microenvironment, such as the intestine.

NON-SPECIFIC IMMUNOTHERAPY CAN BOOST IMMUNE ACTIVITY

Many of the same compounds that act as adjuvants for vaccines have also been used on their own in an attempt to boost the general level of immune activity (Fig. 18.17). The best results have been obtained with cytokines, and among these IFNα is the most widely used, mainly for its antiviral properties (but also for certain tumours – see below).

Perhaps the most striking clinical effect of a cytokine has been that of granulocyte colony stimuling factor (G-CSF) in restoring bone marrow function after anti-cancer therapy, with benefit to both bleeding and infection.

Finally cytokine inhibitors can be used for severe or chronic inflammatory conditions. Various ways of inhibiting TNF and IL-1 have proved valuable:

- in rheumatoid arthritis; and
- more controversially, in septic (Gram-negative) shock and severe malaria.

In a few years, one would expect the clinical pharmacology of cytokines and cytokine inhibitors to be clarified so that these communication molecules of the immune system can be fully exploited, in the same way as vaccination has exploited the properties of the lymphocyte.

IMMUNIZATION AGAINST A VARIETY OF NON-INFECTIOUS CONDITIONS IS BEING INVESTIGATED

The idea of non-specifically stimulating the immune system to reject tumors goes back almost a century to the work of Coley, who used bacterial filtrates with considerable success, possibly through the induction of

Major diseases for which no vaccines are available

	disease	problems
viruses	HIV	antigenic variation; immunosuppression?
	herpes viruses	risk of reactivation? (but varicella-zoster appears safe)
	adenoviruses, rhinoviruses	multiple serotypes
bacteria	staphylococci group A streptococci	early vaccines ineffective (antibiotics originally better)
	Mycobacterium leprae	(BCG gives some protection)
	Treponema pallidum (syphilis)	ignorance of effective immunity
	Chlamydia spp.	early vaccines ineffective
fungi	*Candida* spp. *Pneumocystis* spp.	ignorance of effective immunity
protozoa	malaria	antigenic variation
	trypanosomiasis – sleeping sickness; Chagas' disease	extreme antigenic variation; immunopathology; autoimmunity; trials encouraging
	leishmaniasis	
worms	schistosomiasis	(trials in animals encouraging)
	onchocerciasis	ignorance of effective immunity

Fig. 18.15 For some serious diseases there is currently no effective vaccine. The predominant problem is the lack of understanding of how to induce effective immunity.

Passive immunization

disease	source of antibody	indication
diphtheria, tetanus	human, horse	prophylaxis, treatment
varicella-zoster	human	treatment in immunodeficiencies
gas gangrene, botulism, snake bite, scorpion sting	horse	post-exposure
rabies	human	post-exposure (plus vaccine)
hepatitis B	human	post-exposure
hepatitis A, measles	pooled human immunoglobulin	prophylaxis (travel), post-exposure

Fig. 18.16 Although not so commonly used as 50 years ago, injections of specific antibody can still be a life-saving treatment in specific clinical conditions.

cytokines such as TNF and IFN. However, attempts to equal his results with purified cytokines or immunostimulants (e.g. BCG) have been successful in only a restricted range of tumors, and current efforts are mainly directed at the induction of specific immunity – just as for infectious microbes – encouraged by the evidence that tumors may sometimes be spontaneously rejected as if they were foreign grafts (see Chapter 21).

In principle, conception and implantation can be interrupted by inducing immunity against a wide range of pregnancy hormones. The target of the most successful experimental trials has been human chorionic gonadotropin (hCG), the embryo-specific hormone responsible for maintaining the corpus luteum.

Vaccines based on the β chain of hCG, coupled to tetanus or diphtheria toxoid, have been extremely

Non-specific immunotherapy

source		remarks
microbial	filtered bacterial cultures	used by Coley (1909) against tumors
	BCG	some activity against tumors
cytokines	IFNα	effective for chronic hepatitis B, hepatitis C, herpes zoster, wart virus, prophylactic against common cold (also some tumors)
	IFNγ	effective in some cases of chronic granulomatous disease, lepromatous leprosy, leishmaniasis (cutaneous)
	IL-2	leishmaniasis (cutaneous)
	G-CSF	bone marrow restoration after cytotoxic drugs
cytokine inhibitors	TNF antagonists	septic shock
	IL-1 antagonists	severe (cerebral) malaria?
	IL-10	

G-CSF, granulocyte colony stimulating factor; IFN, interferon; IL, interleukin; TNF, tumor necrosis factor

Fig. 18.17 Non-specific stimulation or inhibition of particular components of the immune system may sometimes be of benefit.

successful in preventing conception in baboons and, more recently, humans. In the human trial, infertility was only temporary, and no serious side effects were observed. Clearly this represents a powerful new means of safely limiting family size, though there are of course cultural and ethical aspects to consider too.

Other uses of immunization that are being explored include:

- the treatment of drug dependency, because it is possible to neutralize the effect of a drug by pre-immunization with the drug coupled to a suitable carrier (the 'hapten–carrier' effect);
- lowering cholesterol by 'neutralizing' lipid-binding proteins; and
- preventing Alzheimer's disease by immunizing against components of amyloid plaque.

So far these vaccines are largely experimental.

Q. What problems can one foresee associated with the types of treatment described above?

A. Breaking tolerance to self molecules raises safety issues. These are well-founded concerns because in some primate and early-phase human trials of Alzheimer's vaccines encephalitis has been observed following immunization.

Future vaccines will use genes and vectors to deliver antigens

Vaccinia is a convenient vector

A development of the use of gene cloning is to insert the desired gene into a vector, which can then be injected into the patient, allowed to replicate, express the gene, and produce the antigen in situ (Fig. 18.18). Vaccinia is a convenient vector that is large enough to carry several antigens. Modified vaccinia Ankara (MVA) may be particularly safe because it does not replicate in human cells (see above, p. 330).

Because immunization against smallpox ceased some years ago, the problem of pre-existing immunity is also decreasing. A number of experimental vaccines using recombinant vaccinia have been tested, though none is yet in routine use. Many other viruses have also been proposed and tested experimentally as vaccine vectors.

BCG and salmonellae have been favored for experimental recombinant bacterial vaccines

Attenuated bacteria have the advantage that they have genomes large enough to incorporate many genes from other organisms. BCG and salmonellae have been favored organisms for experimental recombinant bacterial vaccines.

Mutant salmonellae can be given by mouth and immunize the gut-associated lymphoid tissue before being eliminated – a very useful property because diarrhea is one of the world's major killers in infancy. The mutant bacteria are also capable of inducing systemic immunity. However, as yet no bacterial recombinant vaccines are routinely used.

Transgenic plants can be genetically engineered to express vaccine antigens

An even more innovative approach is to construct transgenic plants expressing vaccine antigens. Mice have been successfully immunized by eating genetically engineered raw potatoes. Many problems with this approach need to be overcome:

- protein antigens may be rapidly degraded in the digestive tract;
- consistency and dosage will be difficult to control;

Generation of recombinant vaccinia virus for expression of a foreign gene

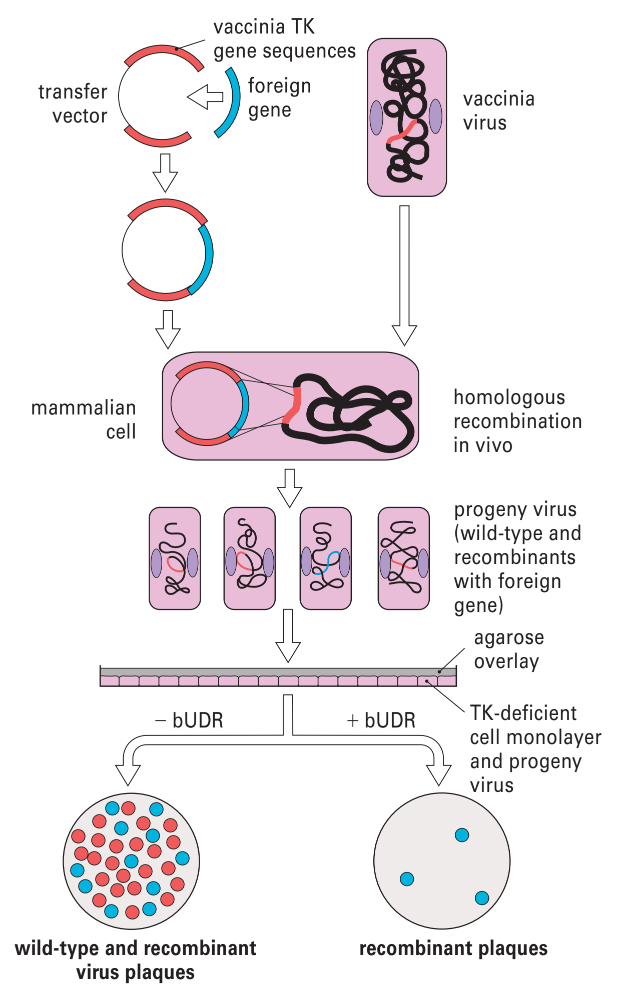
vaccinia TK gene sequences
transfer vector
foreign gene
vaccinia virus
mammalian cell
homologous recombination in vivo
progeny virus (wild-type and recombinants with foreign gene)
agarose overlay
– bUDR
+ bUDR
TK-deficient cell monolayer and progeny virus
wild-type and recombinant virus plaques
recombinant plaques

Fig. 18.18 Recombinant vaccinia virus can be generated to express a foreign gene. The foreign gene is inserted into vaccinia's thymidine kinase (TK) gene so that recombinant virus plaques can be distinguished from wild type. TK is used by virus to take up thymidine from the culture medium or intracellular pool for DNA synthesis; the recombinant virus cannot produce TK, because the gene has been interrupted and so must use the separate pathway for de-novo synthesis of thymidine. In the presence of bromodeoxyuridine (bUDR), a thymidine analog that blocks DNA synthesis when it is incorporated into DNA, wild-type virus replication is blocked but recombinant virus replication continues, using de-novo synthesis of thymidine. The cell monolayer must be TK deficient so that recombinant virus cannot use the cells' TK to take up bUDR. (Courtesy of Dr DJ Rowlands, Ch. 26 in Brostoff J., et al., eds. Clinical Immunology, London: Mosby; 1991)

- antigens administered orally often tolerize rather than immunize, and oral adjuvants may need to be incorporated into the plants.

Even if edible plant vaccines present problems, however, recombinant plants may provide a cheap and convenient way of producing vaccine antigens, which may be purified from the leaves or fruit of the plant and used in conventional subunit vaccines.

A recent development is the use of 'naked' DNA

Genes of interest coupled with a suitable promoter are injected directly into muscle or coated onto gold microparticles and 'shot' into the skin by pressurized gas – the **gene gun**. Surprisingly this can induce long-lasting cellular and humoral immunity in experimental animals. The mechanism appears to be through uptake and expression of the DNA in APCs.

This method has the advantage that immunomodulatory genes (cytokines or co-stimuli) can be incorporated into the DNA construct along with the genes coding for antigens, to generate and amplify the desired immune response. It has also been found that bacterial olignucleotide (CpG DNA) sequences have adjuvant properties and these can be included in the plasmids used to produce the DNA.

DNA vaccines have not yet fulfilled in humans the promise they have shown in animal model systems. This may be because doses of DNA used in animals have generally been relatively higher than in humans, and it is clear that different bacterial CpG sequences are needed to stimulate optimally the dendritic cells of experimental animals and humans.

At present DNA immunization is largely being tested in life-threatening situations (tumors, HIV infection) rather than in routine vaccination of infants.

Priming with DNA and boosting with recombinant vaccinia virus is significantly more effective than DNA alone, at least in animals. However, so far clinical trials of a prime boost regimen of DNA followed by MVA for an HIV vaccine have not been particularly encouraging.

Nevertheless DNA vaccines have the important potential advantages of cheapness and stability so that many means of overcoming these difficulties are being tested, such as:

- the use of in-vivo electroporation to increase the efficiency of DNA entry into cells;
- the optimization of immunostimulatory CpG sequences; and
- the incorporation of carrier proteins or adjuvant cytokines into the vaccines.

FURTHER READING

Beverley PCL, Borysiewicz L, Hill AVS, et al., eds. Vaccination. Br Med Bull 2002;62:1–230.

Fife KH, Wheeler CM, Koutsky LA, et al. Dose-ranging studies of the safety and immunogenicity of human papillomavirus type 11 and type 16 virus-like particle candidate vaccines in young healthy women. Vaccine 2004;22:2943–2952.

Mononego A, Weiner HL. Immunotherapeutic approaches to Alzheimer's disease. Science 2003;302:834–838.

Naz RK. Vaccine for contraception targeting sperm. Immunol Rev 1999;171:193–202.

Roberts L. Rotavirus vaccines' second chance. Science 2004;305:1890–1893.

Critical thinking: Vaccination (see p. 499 for explanations)

1 Why have attenuated vaccines not been developed for all viruses and bacteria?

2 'A vaccine cannot improve on nature'. Is this unduly pessimistic?

3 'The smallpox success story is unlikely to be repeated.' Is this true?

4 Will vaccines eventually replace antibiotics?

5 BCG: vaccine, adjuvant, or non-specific stimulant?

6 Why could an anti-worm vaccine do more harm than good?

7 By what means, other than their reaction with antibodies, might you identify antigens that could be used as vaccines?

SECTION

4

Immune Responses Against Tissues

This Page Intentionally Left Blank

CHAPTER

19

Immunological Tolerance

SUMMARY

- **Immunological tolerance is a state of unresponsiveness for a particular antigen.** Tolerance mechanisms are needed because the immune system randomly generates a vast diversity of antigen-specific receptors and some of these will be self reactive; tolerance prevents harmful reactivity against the body's own tissues.
- **Immunological tolerance was demonstrated by some key historical experiments.**
- **The study of tolerance involves inducing tolerance experimentally.**
- **Central tolerance takes place during T cell development in the thymus.** Central thymic tolerance to self antigens (autoantigens) results from deletion of differentiating T cells that express antigen-specific receptors with high binding affinity for intrathymic self antigens. Low-affinity self-reactive T cells, and T cells with receptors specific for antigens that are not represented intrathymically, differentiate and join the peripheral T cell pool.
- **A variety of mechanisms maintain tolerance in peripheral lymphoid organs.** Post-thymic tolerance to self antigens has five main mechanisms: (i) self-reactive T cells in the circulation may ignore self antigens, for example when the antigens are in tissues sequestered from the circulation; (ii) their response to a self antigen may be suppressed if the antigen is present in a privileged site; (iii) self-reactive cells may under certain conditions be deleted or (iv) rendered anergic and unable to respond; (v) finally a state of tolerance to self antigens can also be maintained by regulatory T cells.
- **Tolerance is also imposed on B cells.** B cell deletion takes place in both bone marrow and peripheral lymphoid organs. Differentiating B cells that express surface immunoglobulin receptors with high binding affinity for self membrane-bound antigens will be deleted soon after their generation in the bone marrow. A high proportion of short-lived, low-avidity, autoreactive B cells appear in peripheral lymphoid organs. These cells may be recruited to fight against infection.
- **Tolerance can be induced artificially** by various regimens that may eventually be exploited clinically to prevent rejection of foreign transplants and to manipulate autoimmune and allergic diseases.

IMMUNOLOGICAL TOLERANCE IS A STATE OF UNRESPONSIVENESS FOR A PARTICULAR ANTIGEN

Immunological tolerance is a state of unresponsiveness that is specific for a particular antigen; it is induced by previous exposure to that antigen.

Active tolerance mechanisms are required to prevent inflammatory responses to the many innocuous air-borne and food antigens that are encountered at mucosal surfaces in the lung and gut.

The most important aspect of tolerance, however, is **self tolerance**, which prevents the body from mounting an immune attack against its own tissues. There is potential for such attack because the immune system randomly generates a vast diversity of antigen-specific receptors, some of which will be self reactive. Cells bearing these receptors therefore must be eliminated, either functionally or physically, or regulated.

Self reactivity is prevented by processes that occur during development, rather than being genetically preprogrammed. Thus:

- homozygous animals of histoincompatible strains A and B reject each other's skin;
- their F_1 hybrid offspring (which express the antigens of both the A and B parents) reject neither A skin nor B skin;
- the ability to reject A and B skin reappears in homozygotes of the F_2 progeny.

It is therefore clear that self–non-self discrimination is learned during development – immunological 'self' must encompass all **epitopes** (antigenic determinants) encoded by the individual's DNA, all other epitopes being considered as non-self.

Q. Why is it necessary for every individual's immune system to learn self–non-self discrimination – we do not just inherit a library of immune receptors from our parents that recognize non-self?

A. The immediate reason is that each individual's immune system is unique. The diversity of MHC molecules (see Chapter 5) and allogeneic variants of self molecules means that new epitopes are present in each generation as a result of gene reassortment. The ultimate reason is that the immune system has to be able to recognize newly arising non-self antigens as new pathogens evolve to threaten each new generation of individuals.

However it is not the structure of a molecule per se that determines whether it will be distinguished as self or non-self. Factors other than the structural characteristics of an epitope are also important. Among these are:

- the stage of differentiation when lymphocytes first confront their epitopes;
- the site of the encounter;
- the nature of the cells presenting epitopes;
- the number of lymphocytes responding to the epitopes.

IMMUNOLOGICAL TOLERANCE WAS DEMONSTRATED BY SOME KEY HISTORICAL EXPERIMENTS

Soon after the existence of antibody specificity was established, it was realized that there must be some mechanism to prevent autoantibody formation. As early as the turn of the 19th century, Ehrlich coined the term 'horror autotoxicus', implying the need for a 'regulating contrivance' to stop the production of autoantibodies.

In 1938, Traub induced specific tolerance by inoculating mice in utero with lymphocytic choriomeningitis virus, producing an infection that was maintained throughout life. Unlike normal mice, these inoculated mice did not produce neutralizing antibodies when challenged with the virus in adult life.

In 1945, Owen reported an 'experiment of nature' in non-identical cattle twins that showed that cells carrying self and non-self antigens could develop within a single host. These animals exchanged hematopoietic (stem) cells via their shared placental blood vessels and each animal carried the erythrocyte markers of both calves. They exhibited life-long tolerance to the otherwise foreign cells, in being unable to mount antibody responses to the relevant erythrocyte antigens.

Following this observation, Burnet and Fenner postulated that the age of the animal at the time of first encounter was the critical factor in determining responsiveness, and hence recognition, of non-self antigens. This hypothesis seemed logical because the immune system is usually confronted with most self components before birth and only later with non-self antigens.

Experimental support came in 1953, when Medawar and his colleagues induced immunological tolerance to skin allografts (grafts that are genetically non-identical, but are from the same species) in mice by neonatal injection of allogeneic cells (Fig. 19.1).

Induction of specific tolerance in mice

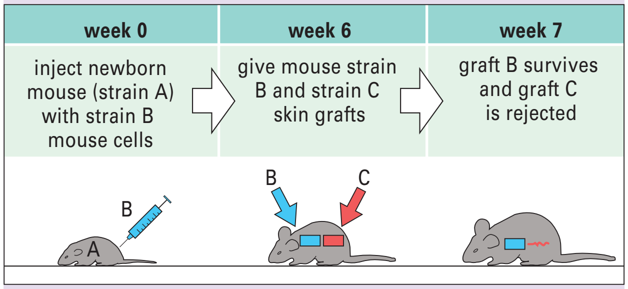
week 0
week 6
week 7
inject newborn mouse (strain A) with strain B mouse cells
give mouse strain B and strain C skin grafts
graft B survives and graft C is rejected
B
A
B
C

Fig. 19.1 The experiment demonstrates the induction of specific tolerance to grafted skin, induced by neonatal injection of spleen cells from a different strain. Mice of strain A normally reject grafts from strain B. However, if newborn mice of strain A receive cells from strain B, they show tolerance to skin grafts from this donor at 6 weeks of age, but reject grafts from other strains (C). This phenomenon is due to immune deviation.

This phenomenon was easily accommodated in Burnet's clonal selection theory (1957), which states that a particular immunocyte (a particular B or T cell) is selected by antigen and then divides to give rise to a clone of daughter cells, all with the same specificity.

According to this theory, antigens encountered after birth activate specific clones of lymphocytes, whereas when antigens are encountered before birth the result is the deletion of the clones specific for them, which Burnet termed 'forbidden clones'. Implicit in the theory is the need for the entire immune repertoire to be generated before birth, but in fact lymphocyte differentiation continues long after birth.

The key factor in determining responsiveness is therefore not the developmental stage of the individual, but rather the state of maturity of the lymphocyte at the time it encounters antigen. This was suggested by Lederberg in 1959, in his modification of the clonal selection theory: immature lymphocytes contacting antigen would be subject to 'clonal abortion', whereas mature cells would be activated.

It is now established that the neonate is in fact immunocompetent. The reason that one can induce tolerance to certain antigens in the neonate is simply that the type of immune response to antigen can be functionally different in the neonate compared with that in the adult. Past descriptions of neonatal tolerance may therefore have been early examples of this type of '**immune deviation**' (see below, p. 355).

Key discoveries in the 1960s established:

- the immunological competence of the lymphocyte;
- the crucial role of the thymus in the development of the immune system; and
- the existence of two interacting subsets of lymphocytes – T and B cells.

This set the scene for a thorough investigation of the cellular mechanisms involved in tolerance.

THE STUDY OF TOLERANCE INVOLVES INDUCING TOLERANCE EXPERIMENTALLY

Transgenic technology has allowed the study of tolerance to authentic self antigens

Until recently, only artificially induced tolerance was amenable to experimental study: antigens or foreign cells were inoculated into an animal and the fate of responding T or B cells was investigated under a variety of circumstances. It was not clear, however, to what extent these experimental models resembled natural self tolerance.

Transgenic methods have now made possible the direct investigation of self tolerance. These methods allow:

- introduction of a specific gene into mice of defined genetic background; and
- analysis of its effects upon the development of the immune system.

Furthermore, if the introduced gene is linked to a tissue-specific promoter, its expression can be confined to specific cell types.

The protein product encoded by a '**transgene**' is treated by the immune system essentially as an authentic self antigen (**autoantigen**), and its effects can be studied in vivo without the trauma and inflammation associated with grafting foreign cells or tissues.

In addition, the parent strain and the transgenic strain are ideal for control experiments and lymphocyte transfer studies because they are **congenic** – that is, they differ at only one locus.

Transgenic mice can also be created in which all of either their B or T lymphocytes express a single antigen receptor. By increasing the frequency of antigen-specific precursor cells in this way, tolerance mechanisms can be readily dissected.

Finally, the use of targeted mutagenesis has allowed immunologists to 'knock out' specific genes to study the role of their gene products in the process of immunological tolerance.

Ways in which self-reactive lymphocytes may be prevented from responding to self antigen

There are five possible ways in which self-reactive lymphocytes may be prevented from responding to self antigens:

- self-reactive T cells in the circulation may ignore self antigens (e.g. when the antigens are in tissues sequestered from the circulation);
- their response to a self antigen may be suppressed if the antigen is in a **privileged site** (see Chapter 12 and p. 352);
- self-reactive cells may be deleted at certain stages of development;
- self-reactive cells may be rendered anergic and unable to respond;
- a state of tolerance to self antigens can also be maintained by immune regulation.

Which of these fates awaits the self-reactive lymphocyte depends on numerous factors, including:

- the stage of maturity of the cell being silenced;
- the affinity of its receptor for the self antigen;
- the nature of this antigen;
- its concentration;
- its tissue distribution; and
- its pattern of expression.

CENTRAL TOLERANCE TAKES PLACE DURING T CELL DEVELOPMENT IN THE THYMUS

The process of generating new T cell receptors (TCRs) involves gene rearrangement in addition to N-region modifications. This allows the immune system to generate a vast array of TCRs. Such a broad repertoire is clearly necessary to provide protection against the multitude of different infectious agents that any individual in the species is likely to encounter.

T lymphocytes are not, however, simply effector cells of the immune system. They also function as regulators of the system through provision of help for some and suppression of other responses.

For effective control, lymphocytes must interact with other cells of the immune system and this is one reason why MHC restriction of T cell recognition has evolved. Central tolerance among T lymphocytes revolves around a schooling process in which key cells are educated so that they become dependent on self MHC for survival, while at the same time potentially rebellious lymphocytes are identified and eliminated. This process of central tolerance among T lymphocytes takes place during their development within the thymus and depends on a number of checkpoints through which the cells have to pass in order to develop further.

T cell development involves positive and negative selection and lineage commitment

T lymphocytes develop from precursors in the bone marrow and are derived from a common lymphoid progenitor cell that gives rise to:

- B cells;
- natural killer (NK) cells; and
- both αβ and γδ subsets of T lymphocytes (Fig. 19.2).

Signals exchanged between cells via the Notch receptor determine the fate of lymphoid progenitor cells. Thus:

- deletion of Notch 1 leads to a block in T cell development and results in ectopic development of B cells in the thymus;
- conversely, expression of activated forms of Notch leads to selective development of T cells from lymphoid progenitor cells.

The function of αβ T lymphocytes in the maintenance of self tolerance is now well understood

When immature T cells enter the thymus they express neither CD4 nor CD8 co-receptor molecules (Fig. 19.3). These so-called double-negative (DN) cells constitute approximately 3% of total thymocytes. At this stage the TCR β chain genes start their recombination.

A common lymphoid progenitor gives rise to B and T cells

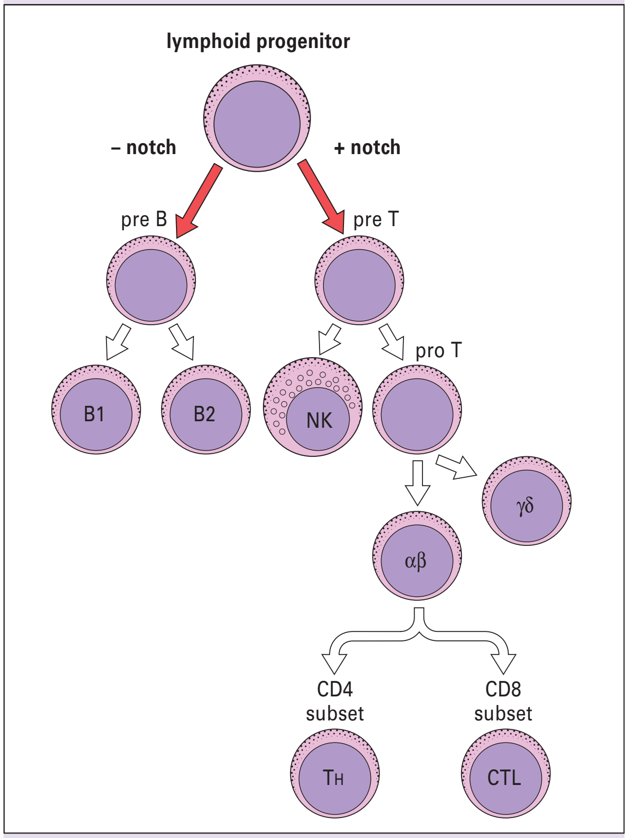

Fig. 19.2 Both B and T cells develop from a common precursor found in the fetal liver or adult bone marrow. T cell progenitors give rise to both the αβ and γδ lineages as well as NK T cells. Recent studies have shown that the decision of lymphocyte precursors to differentiate down the B or T cell pathway depends on ligation of Notch 1. Notch 1 is essential for development of both thymus-dependent and -independent T cells.

Q. What processes are involved in the generation of the diverse genes for the TCR?

A. This process involves sequential rearrangement of variable (V), diversity (D), and junctional (J) region genes within the TCR β chain locus, and expression of a β chain. Subsequently recombination of V and J segments from the TCR α chain locus generates diversity in the α chains (see Figs 5.3–5.5).

First the β chain diversity and junctional genes rearrange and this is followed by rearrangement of the DJ gene with a variable region gene. The VDJ then combines with a constant region by alternative splicing of RNA to give the complete β chain gene.

At this point the α chain genes remain in their genomic configuration, but the transcribed and translated β chain nevertheless appears at the cell surface. This is possible only because the β chain can pair with a 'surrogate' α chain and other components of the CD3 signaling complex to migrate from the endoplasmic reticulum to the cell surface.

Surface expression of this complex allows DN cells to:

- switch off their *RAG* genes (recombination activating genes);
- begin to proliferate; and
- mature into CD4 and CD8 double-positive (DP) cells (see Fig. 19.3).

There is no evidence that this 'checkpoint' (the β selection checkpoint) involves recognition of antigen.

T cells undergo a degree of 'receptor editing' of the α chain

Newly formed DP cells reactivate *RAG* genes allowing rearrangement of the α chain.

Like the immunoglobulin light chain, the TCR α chain has no D segment and the first event is to direct rearrangement of Vα to Jα region genes. Suitable pairing of α and β chains at the α-selection checkpoint allows T cells to proceed to the next selection stage.

Evidence shows, however, that unlike the β chain, which largely permits rearrangement of only one β gene through allelic exclusion, α chain rearrangement can continue to generate a second chain.

In fact, up to 30% of mature human T cells express more than one rearranged α chain.

This implies that T cells, like B cells, undergo a degree of 'receptor editing' of the α chain to increase the likelihood of positive selection of cells selected to interact with self MHC.

T cells are positively selected for 'usefulness' (MHC restriction)

The potential for α–β pairing in combination with TCR gene rearrangement allows for a massive repertoire of TCR structures. Interestingly, however, some 95% of these structures fail to contribute to the T cell repertoire found in peripheral lymphoid tissues. This is because thymocytes undergo a rigorous education before they exit the thymus.

Education requires preliminary selection of cells for survival and their subsequent commitment to a particular lineage (**positive selection**). This is then followed by death of those cells that interact strongly with MHC (**negative selection**).

In other words, T cells are positively selected for 'usefulness' (MHC restriction) and negatively selected against 'dangerous' autoreactivity.

The controlling element in thymic education is the MHC expressed by antigen-presenting cells (APCs) in the thymus (Fig. 19.4). This is such that T cell development is blocked at the DP stage in a thymus that does not express MHC molecules. In fact, cells at this stage of development need to be nurtured by cells expressing MHC molecules.

Cells whose TCR fails to engage either a class I or a class II MHC molecule undergo programmed cell death (death by neglect), while cells that recognize MHC

Developmental pathways of murine thymocytes

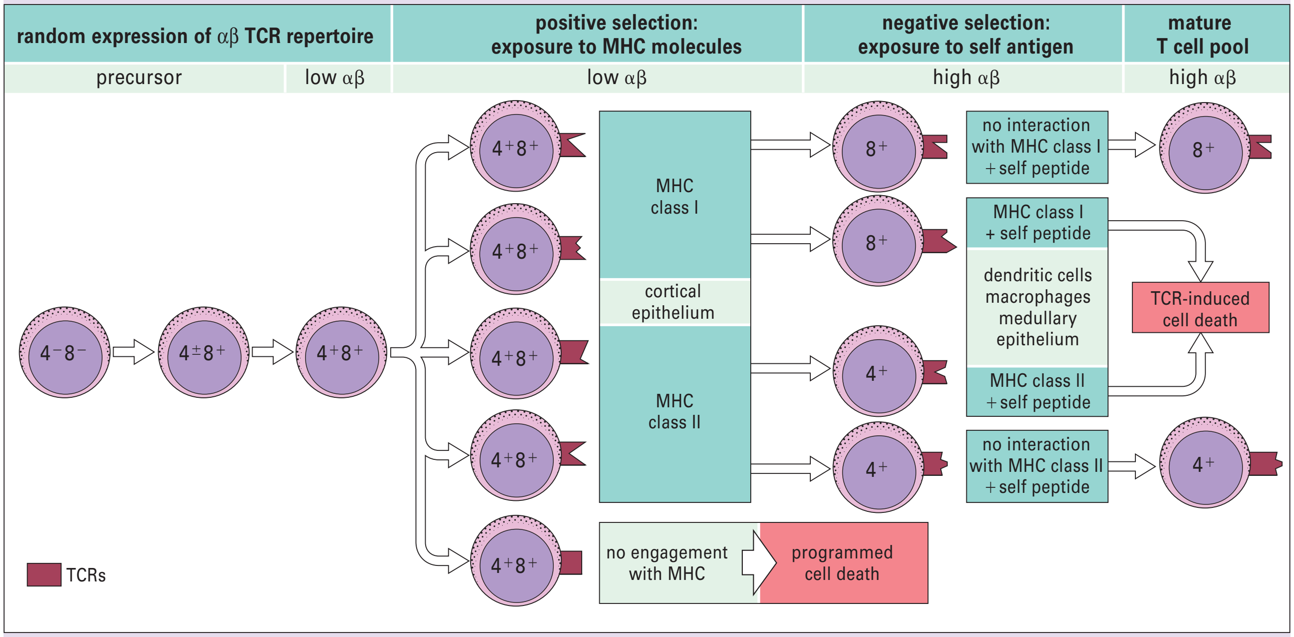
random expression of αβ TCR repertoire
positive selection: exposure to MHC molecules
negative selection: exposure to self antigen
mature T cell pool
precursor
low αβ
low αβ
high αβ
high αβ
4−8−
4±8+
4+8+
4+8+
4+8+
4+8+
4+8+
4+8+
MHC class I
cortical epithelium
MHC class II
8+
8+
4+
4+
no interaction with MHC class I + self peptide
MHC class I + self peptide
dendritic cells macrophages medullary epithelium
MHC class II + self peptide
no interaction with MHC class II + self peptide
TCR-induced cell death
8+
4+
no engagement with MHC
programmed cell death
TCRs

Fig. 19.3 Precursor thymocytes develop into 'double-positive' cells expressing low levels of the αβ TCR. These undergo positive selection for interaction with self MHC class I or class II molecules on cortical epithelium. Unselected cells (the majority) undergo programmed cell death by apoptosis. Cells undergoing positive selection lose one or the other of their co-receptor molecules (CD4 or CD8). Finally, self-reactive cells are eliminated by their interaction with self peptides presented on cells at the corticomedullary junction and in the thymic medulla.

T cell–MHC restriction occurs in the thymus

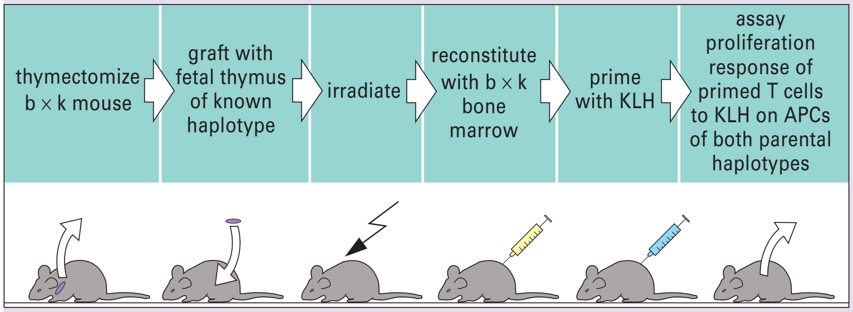
thymectomize b × k mouse
graft with fetal thymus of known haplotype
irradiate
reconstitute with b × k bone marrow
prime with KLH
assay proliferation response of primed T cells to KLH on APCs of both parental haplotypes

mouse type	engrafted fetal thymus type	proliferation response to APCs from each parental strain	
		H-2^b	H-2^k
b × k	b × k	++	++
b × k	b	++	−
b × k	b(dG-treated)	++	−
b × k	k	−	++
b × k	k(dG-treated)	−	++

Fig. 19.4 Host mice (F_1[H-2^b × H-2^k]) were thymectomized, then engrafted with 14-day fetal thymuses of various genotypes. They were subsequently irradiated to remove their resident T cell populations, then reconstituted with F_1 bone marrow to provide stem cells. After priming with antigen (keyhole limpet hemocyanin, KLH), the proliferative response of lymph node T cells to KLH on APCs from each parental strain was evaluated. In some experiments, thymus lobes were incubated before grafting with deoxyguanosine (dG), which destroys intrathymic cells of macrophage/dendritic cell lineage. The results show (i) that the thymic environment is necessary for T cells to learn to recognize MHC; and (ii) that bone marrow-derived cells (removed by dG treatment) are not required for this process to occur. (Based on data from Lo D, Sprent J. Nature 1986;319:672)

molecules with moderate affinity on cortical epithelial cells survive.

Cells mature from DP cells to the single-positive (SP) cells where they express either CD4 or CD8. It is clear that MHC molecules play a role in this selection process.

Q. What do you predict would occur in mice that lack either MHC class I or class II molecules?

A. Mice lacking MHC class I molecules have few CD8 SP thymocytes; mice that lack MHC class II molecules have few CD4 SP cells.

It is likely that evolution has shaped complementarity-determining regions (CDRs) 1 and 2 of the TCR so that the TCR preferentially matches MHC molecules.

Signaling via CD4 and CD8 drives lineage commitment

Why has the immune system evolved two separate types of T cell? Would it not be more economical to have just one DP cell that could interact with either class I or class II expressing cells?

Selection of cells expressing either CD4 or CD8 has evolved just as the need for two pathways of antigen processing has been driven by the encounter of vertebrates with increasingly sophisticated pathogens.

Class I and class II pathways have evolved to allow the immune system to recognize either cytoplasmic or extracellular/intravacuolar infectious agents, respectively. This has then driven the evolution of two subsets of T cell equipped with the means to help eradicate these infectious agents.

The remaining question is how CD4 and CD8 cells develop from one common precursor?

Q. Transgenic mice that have a TCR that recognizes antigen in association with MHC class I molecules and express a hybrid CD4/CD8 molecule (CD4 intracellular segment/CD8 extracellular domains) generate predominantly CD4$^+$ thymocytes. Interpret this experiment.

A. This implies that intracellular signaling via the CD4 intracellular domain leads to selective commitment along the CD4 lineage with inactivation of CD8 expression.

Theories to explain why MHC class II-restricted cells carry CD4 whereas MHC class I-restricted cells express CD8

The mechanisms of thymic selection and CD4/CD8 lineage commitment have many uncertainties. A number of theories have arisen to explain why MHC class II-restricted cells carry CD4 whereas MHC class I-restricted cells express CD8 co-receptors.

Stochastic selection

Lineage choice is independent of MHC–TCR signal, resulting in some thymocytes being mismatched between TCR specificity and co-receptor expression. These mismatched cells subsequently die through lack of an effective survival signal. This theory has been superseded by the following proposals.

Instructional selection

This theory states that coincident engagement of both TCR and co-receptor by MHC leads to lineage choice. Experimental evidence suggests that this is an over-simplified view of selection.

Strength of signal

The instructional model has been extended by experiments suggesting that lineage commitment is influenced by strength or duration of signal. The src family kinase Lck is associated with the cytoplasmic tail of CD4 and CD8, but is known to associate better with CD4 than CD8. It was therefore suggested that ligation of MHC plus peptide by TCR plus CD4 leads to more effective, stronger signaling and hence selection of CD4 SP cells, while TCR plus CD8 generates a weaker signal hence leading to CD8 selection. Interestingly, however, CD8 is downregulated in both CD4 and CD8 lineage-committed cells during the process of selection and this has led to revision of the validity of the signal model.

The kinetic signaling model of lineage commitment in thymic selection

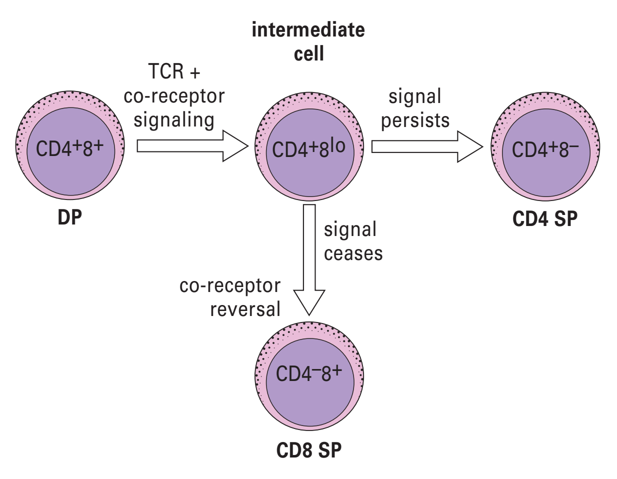

Fig. 19.5 The kinetic signaling model of lineage commitment proposes that cell fate is determined by the duration of signaling during positive selection. The model suggests that double-positive thymocytes are programmed to respond to TCR plus co-receptor signals (regardless of MHC specificity) by terminating CD8 transcription and converting into CD4$^+$CD8^{lo} 'intermediate' thymocytes. CD4$^+$CD8^{lo} intermediate thymocytes still have the potential to differentiate into either mature CD4 SP or CD8 SP T cells. Sustained signaling in intermediate thymocytes results in their differentiation into mature CD4 SP T cells, whereas cessation of signaling in intermediate thymocytes results in 'co-receptor reversal' and differentiation into CD8 SP T cells. (Based on the model described by Singer A. Curr Opin Immunol 2002;14:207)

Kinetic signaling

A further model accommodates the observation that TCR signaling initially downregulates CD8 but not CD4 expression. As such, TCR signaling of short duration leads to CD8 commitment with subsequent downregulation of CD4. Longer TCR signaling leads to CD4 commitment and further downregulation of CD8. Clearly the strength of signaling and kinetic signaling models overlaps, because a strong signal will normally be of longer duration (Fig. 19.5).

In mice, the HD mutation causes redirection of class II-restricted thymocytes to the CD8 lineage

Interestingly, a spontaneous mutation in mice, the HD mutation, causes redirection of class II-restricted thymocytes to the CD8 lineage. The mutation lies in the gene *Th-POK*. The expression of *Th-POK* is confined to cells expressing a class II-restricted TCR and CD4 suggesting that it may be induced by strong TCR signaling. Furthermore, the transcription factor GATA-3 is specific for and involved in CD4 lineage differentiation, whereas the transcriptional regulator RUNX3 functions as a CD4-silencing factor during CD8 lineage differentiation.

Antigen recognition is important for development of the T cell repertoire

Do T cells need to see antigen for positive selection and, if so, does this have to be a specific MHC-bound peptide?

Mice deficient in the proteins required to transport peptides into the endoplasmic reticulum (TAP proteins) do not allow selection of CD8 cells. This proves that peptide antigen in conjunction with MHC class I is required for CD8 cell differentiation. But how many peptides are required for a completely functional T cell repertoire?

Q. Transgenic mice that have MHC class II molecules of a single type occupied by a single peptide generate CD4 SP cells and are able to respond to a number of different antigens. T cell numbers are reduced by 50%, but the repertoire of TCRs that can develop in these animals is highly restricted. Interpret this observation.

A. A diverse range of MHC–peptide complexes is required to select a full repertoire of TCRs.

Thymic cells involved in negative selection

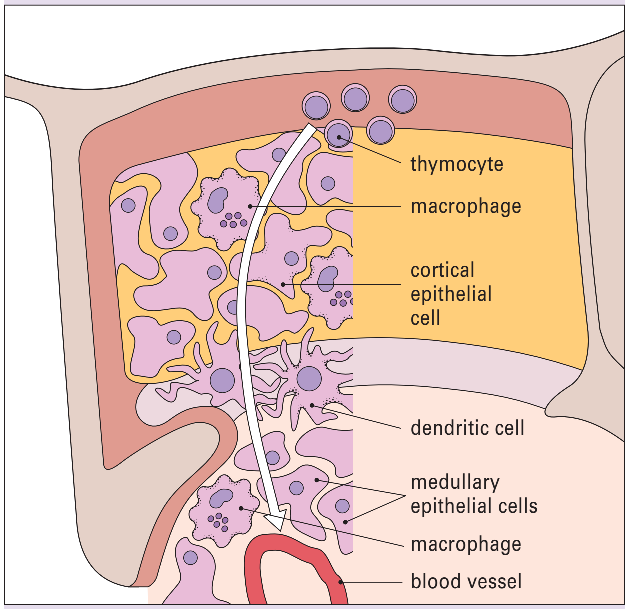

Fig. 19.6 The deleting population includes bone marrow-derived macrophages and dendritic cells, which are located at the corticomedullary junction. Other cells involved in deletion may be the thymocytes themselves, through their veto function, and some types of thymic epithelial cells, especially those in the medulla.

T cell selection is compartmentalized in the thymus

The thymus is made up of lobes, each of which is organized into outer cortical and inner medullary regions (Fig. 19.6).

Immature lymphocytes are found in the cortical region associated with cortical epithelial cells. Cells in the outer cortex are rapidly proliferating immature cells. Cells in the inner cortex are more mature DP cells probably undergoing positive selection.

The medulla contains:

- mature SP lymphocytes;
- medullary epithelial cells; and
- bone marrow-derived macrophages and dendritic cells.

It is a matter of hot debate as to whether spatial separation of MHC molecules on different APCs and their isolation in different thymic regions influence positive and negative selection. Clearly hematopoietic cells are restricted to the medulla.

Two important considerations are:

- the fact that the thymic cortex is relatively inaccessible to large circulating proteins because of its vascular supply; and
- the observation that cortical epithelial cells are inefficient at presenting exogenous proteins – these cells would thus be predicted to present endogenous antigens only and not antigens carried to the thymus in the blood supply.

Cortical epithelial cells definitely play a role in positive selection because mice expressing MHC class II molecules only on these cells show normal levels of positive selection, but impaired negative selection. By contrast, bone marrow-derived macrophages and dendritic cells account for the removal of at least 50% of all positively selected cells.

Medullary thymic epithelial cells can express antigens whose expression was previously thought to be limited to specific organs

A further question relates to how the thymus could ever possibly express all of the antigens that a T cell might encounter outside the thymus. There seems little doubt that the thymus does not express all potential self antigens. Nevertheless there is clear evidence that medullary thymic epithelial cells (mTECs) can express antigens whose expression was previously thought to be limited to specific organs (Fig. 19.7). Importantly, expression of many of these proteins is controlled by a single gene called autoimmune regulatory (*Aire*).

Aire was first identified by mapping the gene responsible for autoimmune polyendocrinopathy syndrome type 1 (APS-1) in humans. This disease is characterized by the coincidence of two or three major clinical symptoms: Addison's disease, hypoparathyroidism and chronic mucocutaneous candidiasis.

Antibodies against a variety of endocrine glands have been detected in patients with APS-1 and the disease is associated with lymphocytic infiltrates in the affected organs. The comparison of protein expression in wild-type and *Aire*-deficient mTECs shows that expression of organ-specific genes is preferentially downregulated in

The role of *Aire* in immune tolerance

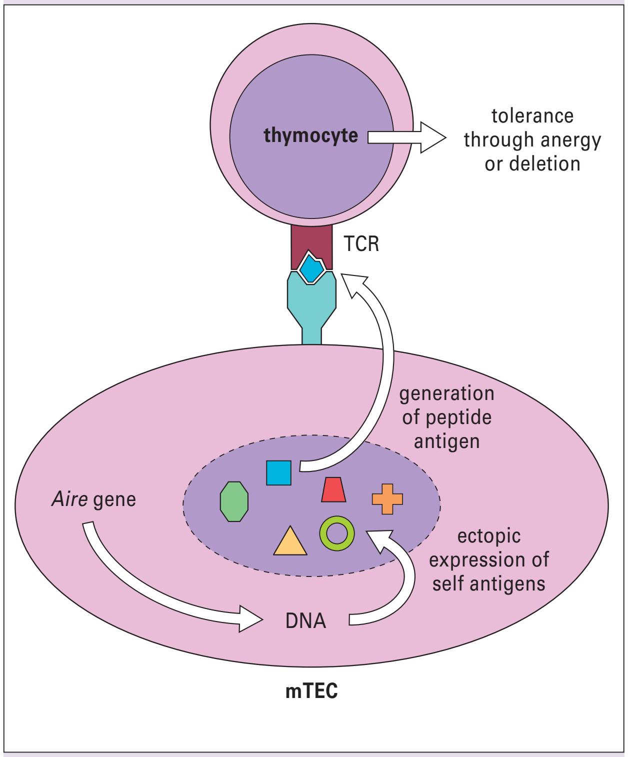
thymocyte
tolerance through anergy or deletion
TCR
generation of peptide antigen
Aire gene
ectopic expression of self antigens
DNA
mTEC

Fig. 19.7 *Aire* promotes the expression of organ-specific genes in medullary thymic epithelial cells (mTECs). These organ-specific proteins are presented on the surface of mTECs by MHC molecules to T cells developing in the thymus. Thymocytes that recognize these organ-specific proteins in the context of MHC molecules undergo negative selection. The role of *Aire* is therefore to limit the generation of self-reactive T cells. Mutations in the *Aire* gene affect expression of many but not all ectopically expressed proteins, suggesting that other genes may play a similar role. (Based on Su MA, Anderson MS. Curr Opin Immunol 2004;16:746. Copyright 2004 with permission from Elsevier)

Aire-deficient mTECs. Among these are preproinsulin and cytochrome P-4501A2, which are known targets for autoantibodies in APS-1. From this we can conclude that APS-1 and a similar disease found in *Aire*-deficient mice results from the failure to delete or anergize developing thymocytes specific for these organ-specific proteins.

T cell development includes a series of checkpoints

In conclusion, the architecture of the thymus appears to be designed to compartmentalize thymic selection. Cortical epithelial cells present a wide range of endogenous antigens and contribute to positive selection.

Interestingly, it is estimated that a developing thymocyte might only ever interact with a single cortical epithelial cell. The imprint that this leaves on the T cell clearly has a profound effect on the resulting T cell repertoire.

Medullary APCs have access to circulating antigens and are largely responsible for negative selection.

It is now also clear that antigens from a wide variety of tissues are expressed by mTECs where they induce central tolerance.

Furthermore, there is evidence that antigen recognition in the thymus may contribute to the generation of regulatory T lymphocytes, which play such an important role in peripheral tolerance (see below, p. 354).

Checkpoints in central T cell tolerance include:

- β selection checkpoint – only cells with a rearranged β chain mature from DN to DP cells – this process is not dependent on MHC proteins;
- α selection checkpoint – cells expressing an αβ complex must interact with MHC molecules to survive;
- lineage commitment checkpoint – cells are instructed to repress expression of either CD4 or CD8 and to develop into SP cells;
- negative selection checkpoint – cells that interact strongly with MHC molecules and antigen in the thymus are deleted (Fig. 19.8).

It is important to consider how a cell with one TCR can receive signals instructing it to survive and undergo lineage commitment without coincidentally receiving a signal to undergo negative selection. The most likely explanation is that this relies on the avidity of the cell's interaction with MHC and peptide.

Experiments have shown that the decision to undergo positive and negative selection is directly related to the half-life of TCR binding to the MHC–peptide complex (Fig. 19.9).

The correlation between avidity and thymocyte selection

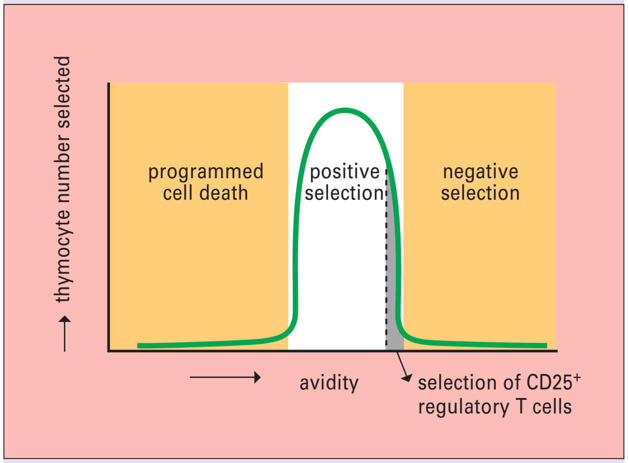

Fig. 19.8 The avidity of a T cell's interaction with antigenic peptide presented by an APC will depend on the level of expression of the MHC–peptide complex [MHC + peptide] on the APC and both the affinity and surface expression of TCR on the T cell. [MHC + peptide] depends on the affinity of peptide for MHC and the stability of the complex once formed. Current evidence suggests that $CD25^+$ regulatory T cells are selected in the thymus and have a relatively high affinity for MHC and peptide.

TCR affinity for MHC–peptide complex influences positive selection

nature of peptide	half-life	affinity of complex	thymocyte selection
agonist	++	++	negative
antagonist	±	+	positive
irrelevant	–	–	no effect

Fig. 19.9 The affinity of a soluble TCR for complexes between various peptides and the appropriate MHC restriction element can be measured by biophysical techniques such as surface plasmon resonance. There is a direct correlation between the half-life of TCR binding to the MHC–peptide complex and the response made by mature T cells expressing the same receptor (i.e. agonist > antagonist > irrelevant peptide). In thymocyte organ cultures, however, addition of the agonist peptide causes deletion of the developing cells (negative selection), whereas the antagonist stimulates positive selection. This demonstrates that low-avidity interaction promotes positive selection, whereas high-avidity interaction leads to negative selection. (Data summarized from Alam SM, Travers PJ, Wung JL, et al. Nature 1996;381:616)

Selection also depends on:

- the architecture of the thymus;
- the nature of APCs in the cortex versus the medulla of the thymus; and
- the types of antigen that these cells are able to present.

A VARIETY OF MECHANISMS MAINTAIN TOLERANCE IN PERIPHERAL LYMPHOID ORGANS

There is no doubt that many potentially autoreactive T cells escape central tolerance. This reflects the fact that many antigens are either not present or are present at insufficiently high levels to induce tolerance in the thymus. Thus, for example, peripheral blood lymphocytes from healthy individuals respond vigorously to purified myelin basic protein, a major constituent of myelin in the brain, following their culture in vitro (Fig. 19.10).

So, how are these cells kept at bay in healthy individuals and why are autoimmune diseases directed to such proteins so incredibly rare? This is because various mechanisms have evolved to maintain tolerance in peripheral lymphoid organs (Fig. 19.11).

Sequestration of antigen occurs in some tissues

Both developing and mature lymphocytes may never encounter self antigens. Many of these are sequestered away from the immune system by physical or immunological barriers. In this way, tissue antigens may never be available to T lymphocytes, either because:

- of their location; or
- they may never be processed by functional APCs.

Peripheral blood cells from a healthy individual respond to the self antigen myelin basic protein

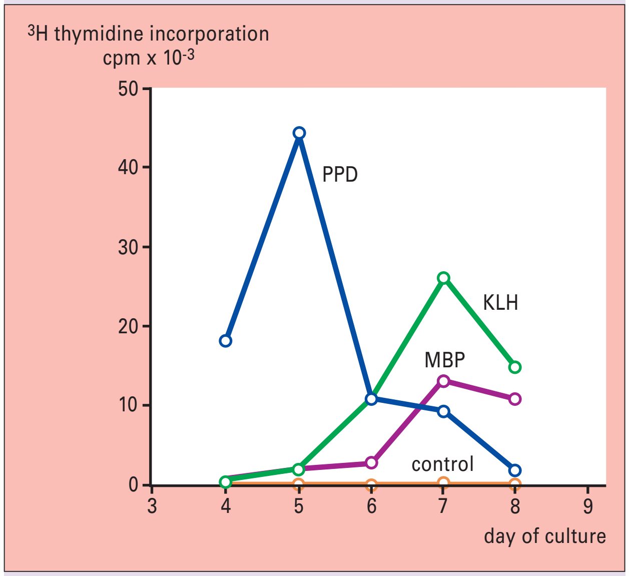

Fig. 19.10 Purified white blood cells can be stimulated in tissue culture with antigens such as purified protein derivative (PPD) from *Mycobacterium tuberculosis* or keyhole limpet hemocyanin (KLH). The rate at which the cells respond (measured by incorporation of 3H-thymidine) reflects whether the response is primary/naive (KLH) or secondary/memory (PPD). Note that blood cells from a healthy individual respond with primary kinetics to purified human myelin basic protein (MBP). This experiment shows how cells that have escaped tolerance induction in the thymus can nevertheless respond to self antigens under artificial conditions. (Based on data from Ponsford M, Mazza G, Coad J, et al. Clin Exp Immunol 2001;124:315)

Privileged sites are protected by regulatory mechanisms

Cells that have escaped tolerance in the thymus can also ignore self antigens if they are expressed in a privileged site.

Q. Name some privileged sites, their characteristics, and the underlying basis of the privilege.

A. Immunologically privileged sites include the brain, anterior chamber of the eye, and testes, because transplanted tissues have an enhanced chance of survival within them. Within these sites proinflammatory lymphocytes are controlled either by apoptosis (Fas ligand [FasL] expression) or by cytokine (transforming growth factor-β/interleukin-10 [TGFβ/IL-10]) secretion (see Chapter 12, p. 236 and Figs 12.3–12.7).

Does immune privilege contribute to self tolerance or is it a phenomenon seen only by the introduction of foreign antigen to the privileged site?

Immune privilege is clearly designed to dampen down inflammatory responses in certain vital organs. The same suppressive mechanisms would equally apply to inflam-

Mechanisms of central and peripheral tolerance

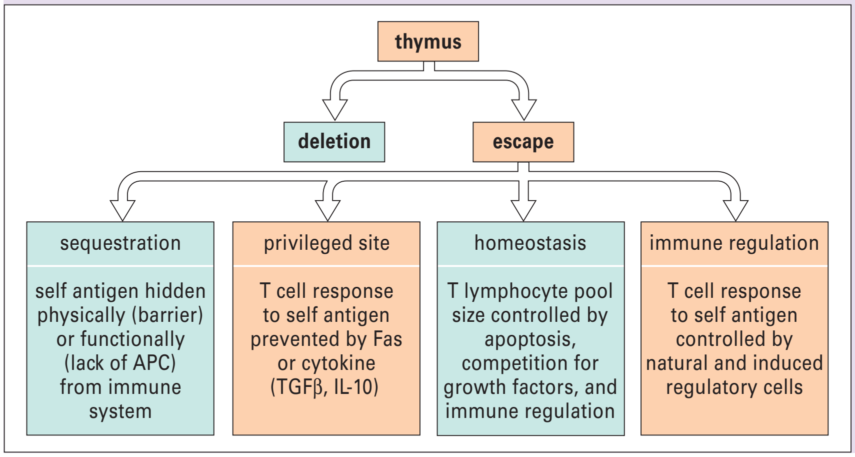
thymus
deletion
escape
sequestration
self antigen hidden physically (barrier) or functionally (lack of APC) from immune system
privileged site
T cell response to self antigen prevented by Fas or cytokine (TGFβ, IL-10)
homeostasis
T lymphocyte pool size controlled by apoptosis, competition for growth factors, and immune regulation
immune regulation
T cell response to self antigen controlled by natural and induced regulatory cells

Fig. 19.11 Cells that have escaped negative selection in the thymus are still subject to control in the periphery. Most healthy individuals maintain self tolerance by a variety of mechanisms of peripheral tolerance involving sequestration, expression at a privileged site, deletion or immune regulation. (Based on Anderton S, Burkhart C, Metzler B, Wraith D. Immunol Rev 1999;169:123)

mation caused by an immune response to either an infectious or a self antigen.

Immune privilege does not discriminate and therefore contributes to peripheral tolerance, at least in these particular organs.

T cell death can be induced by persistent activation or neglect

Apoptotic death of lymphocytes is an extremely important mechanism of immune control and is essential for the maintenance of immune homeostasis in healthy individuals. It contributes both to the deletion of cells with high avidity for antigen and to the death of lymphocytes when the immune response is no longer required. These functions are fulfilled by two distinct mechanisms:

- **activation-induced cell death (AICD)**; and
- **programmed cell death (PCD)**.

Fas is the most important death receptor

Cells repeatedly stimulated with antigen undergo AICD by mechanisms involving so-called 'death receptors' of the tumor necrosis factor receptor (TNFR) family (see Fig. 10.8). Among these, the most important molecule is Fas, which on cross-linking by its ligand (FasL) leads to activation of the caspase cascade via caspase-8 and subsequent apoptotic death of the cell (Fig. 19.12 and see Fig. 10.9). This can occur by cell–cell interactions and there is, in addition, evidence that T lymphocytes can kill themselves through 'fratricidal cell death' following the secretion of soluble FasL (Fig. 19.13).

Many activated cells die by PCD

In addition, many activated cells die by PCD because their antigen is simply eliminated, as happens, for example, following clearance of an infection. Removal of the antigen then deprives cells of essential survival stimuli including growth factors. Under these conditions mitochondria in the cell respond by releasing cytochrome c. This, in combination with apoptosis activating factor-1, leads to activation of the caspase cascade following cleavage and activation of caspase-9 (see Fig. 19.12).

The Fas pathway is required for peripheral tolerance

Presumably, the survival rate of T cells that cross-react with self antigens, but which are generated during the immune response to infection, will be increased in the absence of AICD.

The importance of the Fas pathway for AICD has been revealed by genetic defects in both mouse and human. For example, the lpr mouse has a mutation in Fas and the gld mouse has a mutation in FasL. Both mutations lead to lymphadenopathy (expanded secondary lymphoid tissue). Importantly this lack of regulation also leads to the generation of autoimmunity, autoantibody production, and nephritis with similarities to systemic lupus erythematosus in humans.

Note that thymus selection in the lpr mouse is normal, showing that the Fas pathway is not essential for central tolerance but is clearly required for peripheral tolerance.

Recent studies have shown that analogous mutations lead to a similar form of disease known as human **autoimmune lymphoproliferative syndrome (ALPS)**, characterized by:

- defective lymphocyte apoptosis;
- lymphocyte accumulation; and
- humoral autoimmunity.

The ALPS phenotype is associated with inherited mutations in the Fas gene (ALPS type 1a) or the Fas ligand gene (ALPS type 1b).

Both AICD and PCD are tightly regulated but the two apoptopic pathways are under independent regulation.

Bcl family members, for example, block PCD by inhibiting the release of cytochrome c, but do not affect AICD.

AICD is inhibited by proteins binding to the death receptor complex. Of these, the most important is FLIP (**FLICE inhibitory protein** where FLICE is the FADD-like IL-1β converting enzyme). FLIP binds to the adapter

Two distinct mechanisms of lymphocyte apoptosis

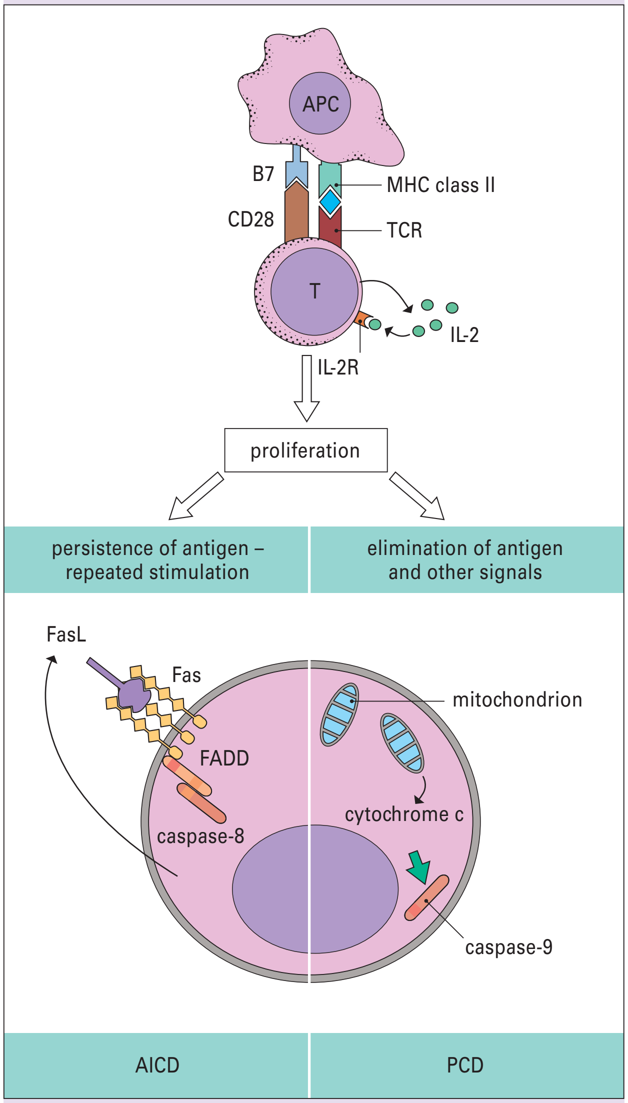

Fig. 19.12 Activated T lymphocytes will die by passive cell death (PCD) when deprived of an antigenic stimulus. This mechanism is designed to maintain homeostasis in the immune system. Activated T lymphocytes will die by activation-induced cell death (AICD) if repeatedly stimulated with antigen. This mechanism is designed to limit hypersensitivity reactions to allergens and autoantigens.

protein FADD or to a precursor form of caspase-8 and blocks generation of the Fas-associated death receptor complex.

AICD is also regulated by IL-2. This cytokine stimulates Fas-mediated AICD by enhancing transcription and expression of FasL while inhibiting transcription of FLIP.

Note that disruption of the genes for IL-2, IL-2Rα, or IL-2Rβ leads to lymphadenopathy and autoimmunity in mice in which any one of these genes has been inactivated. This is consistent with a role for IL-2-driven apoptotic death (**propriocidal death**) in homeostasis and peripheral tolerance.

The role of the Fas system in T cell death

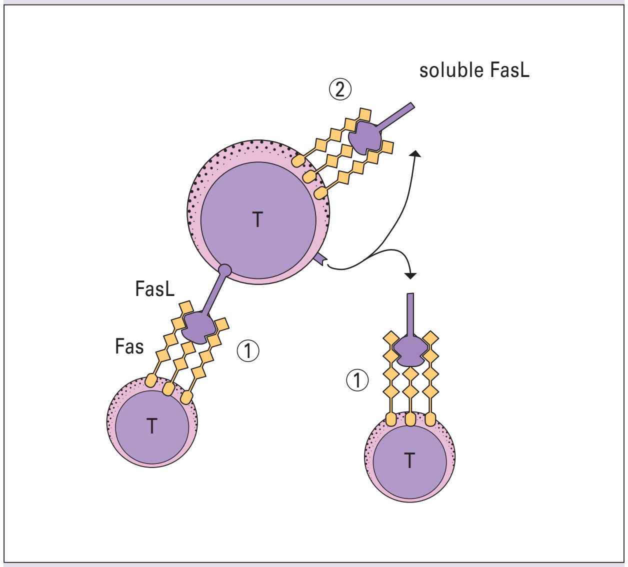

Fig. 19.13 Activated T cells express both Fas (CD95) and the ligand for this molecule (FasL). Fratricide (**1**) can result either from direct cell contact or from cleavage of FasL and the ligation of Fas by soluble FasL. Autocrine suicide (**2**) can result from the interaction of soluble FasL with Fas.

The balance of co-stimulatory signals affects immune homeostasis and self tolerance

Naive T lymphocytes require two signals to proliferate and differentiate:

- the first signal is triggered by TCR recognition of the appropriate peptide–MHC complex;
- the second signal is delivered by CD80 (B7.1) and CD86 (B7.2) co-stimulatory molecules expressed by APCs.

How a T cell interprets co-stimulation depends on which co-stimulation receptor it uses.

The CD28 molecule is constitutively expressed on T cells and signaling via CD28 enhances cell survival, prevents anergy induction, and enhances CD40L expression.

Q. What effect does ligation of CTLA-4 (CD152), the alternative ligand for B7, have on the T cell?

A. CTLA-4 ligation inhibits early T cell activation including expression of the IL-2 receptor α chain and secretion of IL-2 (see Fig. 7.18).

The **CTLA-4 pathway** inhibits IL-2 messenger RNA accumulation and progression through the cell cycle. CTLA-4 has a higher avidity (100 ×) for CD80 and CD86 than CD28, but is normally restricted to the perinuclear Golgi apparatus. On T cell contact with an APC, CTLA-4 traffics to the plasma membrane at the TCR–APC interface (Fig. 19.14).

The role of CTLA-4 in normal homeostasis is revealed in the CTLA-4 knockout mouse. These mice show normal thymus selection, but suffer from polyclonal T cell expansion and die from a fatal lymphoproliferative disease.

Interestingly, CTLA-4 influences CD4 cells more than CD8 cells because depletion of CD4 cells from the CTLA-4 knockout mouse prevents the lymphoproliferative disease in this mouse.

Dendritic cells contribute to peripheral tolerance

Thymic dendritic cells contribute to self tolerance by deleting self antigen-responsive T cells.

Dendritic cells also contribute to tolerance extrathymically. Subsets of dendritic cells are distinguished according to their cell surface phenotype and in mice have been defined as:

- myeloid (precursors shared with macrophages); or
- lymphoid (precursors shared with lymphocytes).

Plasmacytoid dendritic cells are found in various organs and are distinguished by their secretion of type I interferon in response to viruses.

All dendritic cells originate, however, from hematopoietic precursor cells and there is considerable plasticity between the different lineages.

The general rule governing dendritic cells is that they readily take up antigen in the immature state, but present antigen effectively to naive T cells only when they have undergone differentiation to a more mature state (Fig. 19.15):

- mature dendritic cells presenting specific antigen both activate T cells and promote their survival;
- immature dendritic cells presenting antigen activate T cells, but the outcome is different, resulting in apoptosis, anergy, or the generation of regulatory T cells.

In the steady state, the immature dendritic cell can take up self antigens even when these are contained in dying cells and induce tolerance towards these antigens rather than immunity.

T cell activation is controlled by co-stimulatory signals

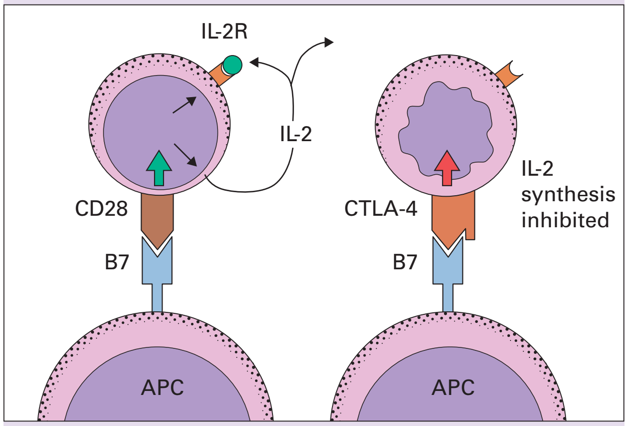

Fig. 19.14 Ligation of CD28 stimulates intracellular signals that lead to IL-2 production, IL-2 receptor (IL-2R) expression, and cell cycle progression in activated T cells. Ligation of cell surface CTLA-4 blocks these CD28-dependent responses and inhibits IL-2 synthesis.

The difference between self antigens and the antigens contained in microbes lies in the ability of microbes to induce maturation of the dendritic cell by ligating Toll-like and other receptors (see Chapters 15 and 16).

Targeting of antigens to dendritic cells, either in the form of synthetic peptides or antigens coupled to antibodies, against antigen-uptake receptors of dendritic cells (i.e. DEC-205) will therefore induce peripheral tolerance if the targeted dendritic cell remains in an immature state.

Comparison of immature and mature dendritic cells

inducing agents		properties
		antigen capture
pathogens (e.g. LPS, bacterial DNA, CpG) cytokines (e.g. TNF, GM-CSF) T cells (e.g CD40L) viral dsRNA	immature DC 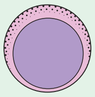	high intracellular MHCII (MIICs) endocytosis phagocytosis high CCR1, CCR5, CCR6 low CCR7 low CD54, 58, 80, 86 low CD40 low CD83 high CD68 no DC-LAMP
	mature DC	**antigen presentation**
	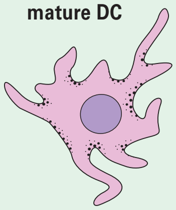	high surface MHCII low endocytosis low phagocytosis low CCR1, CCR5, CCR6 high CCR7 high CD54, 58, 80, 86 high CD40 high CD83 low CD68/high DC-LAMP high p55

Fig. 19.15 On the left are listed factors known to promote maturation of dendritic cells (DCs). These include baterial and viral products, cytokines, and interaction with T cells expressing CD40L. Immature DCs are dedicated to antigen capture whereas mature DCs take up antigen poorly, but are highly efficient APCs. (Based on Banchereau J, Briere F, Caux C, et al. Annu Rev Immunol 2000;18:767, with permission. Copyright 2000 by Annual Reviews www.annualreviews.org)

Homeostatic balance is required to prevent lymphoproliferative responses

In summary, homeostatic balance in the immune system is required to prevent lymphoproliferative responses. Lymphoproliferative disorders are associated with responses to both foreign and self antigens. Lack of homeostatic control leads to autoimmunity and therefore the molecules involved in homeostatic control are important regulators of peripheral tolerance. These molecules include:

- CTLA-4, which acts as a brake on the normal immune response to both foreign and self antigens;
- members of the TNFR family, particularly Fas and FasL;
- components of the caspase cascade;
- IL-2 and the IL-2 receptor (IL-2R), which together regulate sensitivity to Fas-mediated apoptosis.

Regulatory T cells play a crucial role in controlling autoimmune responses

Peripheral tolerance to antigens can be 'infectious'. The experimentally induced tolerance to one antigen can thus maintain tolerance or suppress the immune response to a second antigen as long as the two antigens are structurally or physically associated (e.g. within the same tissue). This implies that mechanisms other than ignorance and cell death must be involved in tolerance.

One explanation for such phenomena depends on the existence of populations of T lymphocytes that produce distinct cytokines.

Many inflammatory autoimmune diseases are caused by TH1 cells that produce cytokines such as interferon-γ (IFNγ) and tumour necrosis factor-α (TNFα). Cytokines derived from TH2 cells (IL-4, IL-5, IL-6, IL-10) support antibody production. A major additional effect of TH2-derived cytokines such as IL-10, however, is downregulation of macrophage effector functions, including antigen presentation to TH1 and naive T cells. TH2 cells are therefore able to suppress inflammatory (delayed-type hypersensitivity, DTH) responses.

TH1 cell-derived IFNγ can prevent the differentiation of TH0 to TH2 cells. This type of immune deviation was defined more than 30 years ago to describe how an individual animal could respond to the same antigen in two completely different ways:

- guinea pigs primed with antigen in alum produced high levels of IgG1 antibody, but did not support a DTH response;
- animals primed with the same antigen in complete Freund's adjuvant developed strong DTH responses.

It was subsequently suggested that the ability of the same antigen to induce either 'humoral' or 'cellular' immunity could reflect the distinct activation of two mutually antagonistic arms of the immune system.

The results of these experiments were undoubtedly a form of immune deviation resulting from the selective induction of TH2 rather than TH1 cells.

Immune deviation can influence hypersensitivity conditions. Diabetes in the NOD (non-obese diabetic) mouse is known to be caused by TH1 cells and can be prevented by antigen-primed TH2 cells, whereas certain allergic disorders can be treated by induction of TH1 cells.

Q. Transgenic mice that express the antigen influenza hemagglutinin [HA]) in pancreatic islet cells and a TCR on their T cells that recognizes this antigen have been bred into mice with different genetic backgrounds. If bred into a strain that produces high levels of IL-4 and IFNγ, the mice have no inflammatory disease in the pancreas. If bred into a strain that produces predominantly IFNγ and IL-12, the T cells infiltrate the pancreatic islets and cause diabetes. Interpret these observations.

A. Peripheral T cell tolerance depends on the genetic make-up of the individual. Immune deviation is clearly controlled by background genes, many of which combine to control the susceptibility of an individual to autoimmune disease.

In another mouse strain (B10.D2 background), the HA-reactive cells produce only TH1 cytokines. T cell-mediated diseases (such as insulin-dependent diabetes, thyroiditis, and gastritis) can be produced in otherwise normal mice simply by eliminating a subpopulation of CD4$^+$ T cells expressing CD5, CD25, or a particular isoform of CD45.

This is best illustrated by the transfer of subsets of murine cells isolated from healthy mice into *Rag*$^{-/-}$ mice that otherwise do not contain T cells of their own. Naive and activated murine T cells can be distinguished according to the level of cell surface expression of CD45 RB – activated cells have low levels of this isoform and naive cells have high levels. Activated or naive populations are then transferred into recipient *Rag*$^{-/-}$ mice:

- transfer of naive cells leads to inflammatory bowel disease (IBD) in the *Rag*$^{-/-}$ recipients; but
- co-transfer of relatively few activated cells prevents disease (Fig. 19.16).

The activated cells either produce or induce production of the immune suppressive cytokine TGFβ.

Another cytokine, IL-10, undoubtedly plays a role in the function of the activated, regulatory cells because transfer of these regulatory cells from an IL-10 knockout mouse fails to suppress the induction of IBD by naive cells.

Regulatory T cells suppress the activation of other T cells

Although the nature of regulatory T cells was discussed extensively in Chapter 11, the subject is of such relevance to tolerance induction that it is worth recapitulating the features of these cells. As described above (Fig. 19.8), a naturally occurring subset of regulatory lymphocytes is distinguished by expression of the α chain of the IL-2R (CD25). Elimination of these cells from normal mice leads to the generation of various organ-specific autoimmune conditions. Similar cells exist in humans and there is clear evidence that they play a role in protecting against autoimmunity. These natural regulatory T cells are anergic to TCR-mediated activation, but potently suppress the activation of other T cells.

CD25$^+$ AND IL-10 REGULATORY CELLS CONSTITUTE DISTINCT REGULATORY CELL SUBSETS – CD25$^+$ regulatory cells are generated in the thymus. It seems likely

IL-10 is required for regulation of inflammatory bowel disease mediated by activated T cells

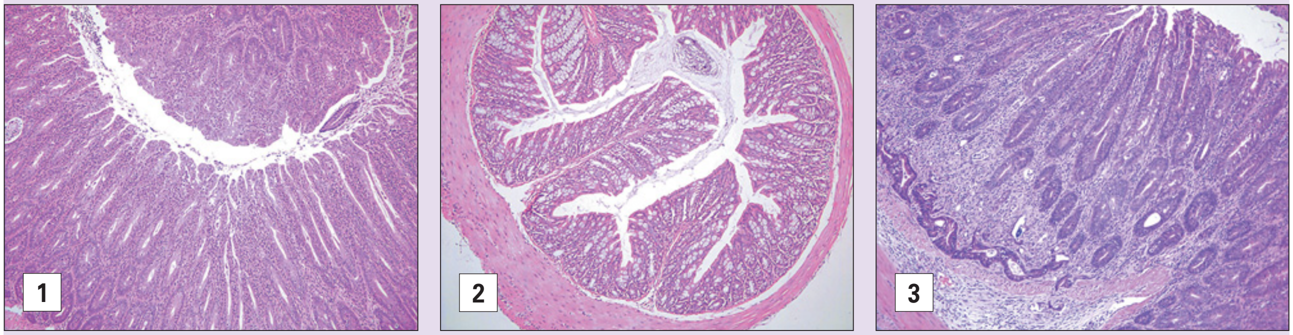

Fig. 19.16 (**1**) Severe colitis in a mouse injected with CD45RB^{high} CD4$^+$ T cells from normal mice. (**2**) Normal appearance of the colon in a mouse restored with both CD45RB^{high} and CD45RB^{low} cells. This shows that the CD45RB^{low} population of T cells is able to inhibit inflammation caused by normal CD45RB^{high} cells. (**3**) Severe colitis in a mouse restored with both CD45RB^{high} and IL-10$^{-/-}$ CD45RB^{low} cells. This experiment demonstrates that CD45RB^{low} cells that cannot produce IL-10 fail to serve as regulators of disease. (Based on Asseman C, Mauze S, Leach MW, et al. Reproduced from The Journal of Experimental Medicine 1999;190:995 by copyright permission of The Rockefeller University Press)

that the cells destined to become CD25$^+$ regulatory cells have a relatively high affinity for their MHC restriction element and only just escape negative selection in the thymus (see Fig. 19.8). Indeed their high affinity for MHC molecules could explain their continuous expression of CD25, a marker of T cell activation, as well as their anergic state. There is also evidence that CD25$^+$ regulatory cells can be generated in peripheral lymphoid tissues.

As discussed previously, the forkhead/winged helix transcription factor FoxP3 appears to control the generation and function of CD25$^+$ Treg cells (see Figs 11.13 and 11.14).

In mice, FoxP3 is preferentially expressed by CD25$^+$ cells. Furthermore, mutations of the *FoxP3* gene are associated with a broad range of hypersensitivity conditions in both mouse and humans. Importantly, CD25$^+$ FoxP3$^+$ cells from normal mice prevent disease when transferred into the FoxP3 mutant 'scurfy' mouse and it will be recalled from Chapter 11, that point mutations and microdeletions of the *FoxP3* gene were found in the affected members of families with the IPEX (immune dysfunction, polyendocrinopathy, enteropathy, X-linked) syndrome.

Precisely how CD25$^+$ cells regulate immune function in vivo is not clear. The cells express molecules such as CTLA-4 and this may function by ligating B7 molecules on neighboring T cells.

The ability of CD25$^+$ regulatory cells to control different types of immune pathology may or may not require them to secrete immunosuppressive cytokines such as TGFβ and IL-10. These cytokines can suppress antigen presentation by professional APCs and have directly suppressive effects on naive T cells.

In addition to the FoxP3$^+$ CD25$^+$ regulatory cells there are other important subsets of cells with regulatory properties. These include the CD25$^-$ TGFβ-producing Tr1 cells elicited by oral administration of proteins, and IL-10-secreting regulatory cells (see Fig. 11.14).

IL-10 regulatory cells:

- can be selected in vivo by repeated antigen encounter;
- differentiate from CD25$^-$ precursors; and
- do not necessarily express FoxP3.

CD25$^+$ and IL-10 regulatory cells therefore constitute distinct regulatory cell subsets (Fig. 19.17).

THE LIGANDS FOR TLRs AND CYTOKINES OVERRIDE SUPPRESSION IN POTENTIALLY OVERWHELMING INFECTION – The balance between immune response and immune suppression must be under fine control to ensure an effective response to infection while preventing an aberrant response to self antigens. How then does the immune system mount an effective response to infection in the face of such immune regulation? Here the nonspecific receptors for infectious agents, including the Toll-like receptors (TLRs), may sense the presence of an infectious agent. The ligands for TLRs and cytokines, secreted by both the innate and adaptive immune systems, combine to override suppression in the face of potentially overwhelming infection. When the immune system has eradicated the infection, however, the regulatory cells more than fully recover their suppressive function and hence control autoimmunity (Fig. 19.18).

Q. Which CD25$^+$ cells are not regulators?

A. The CD25 marker is upregulated on naive T cells in response to antigen as part of an active immune response (see Fig. 7.23).

In conclusion, regulatory lymphocytes play a crucial role in the control of autoimmune responses. Scientists are only just beginning to reveal the mechanisms by which these cells mediate their suppressive activity. Furthermore we know very little about how these cells are generated in the normal immune repertoire of healthy individuals. There is little doubt that clarification of these questions will help in the control of many hypersensitivity conditions including allergic and autoimmune diseases.

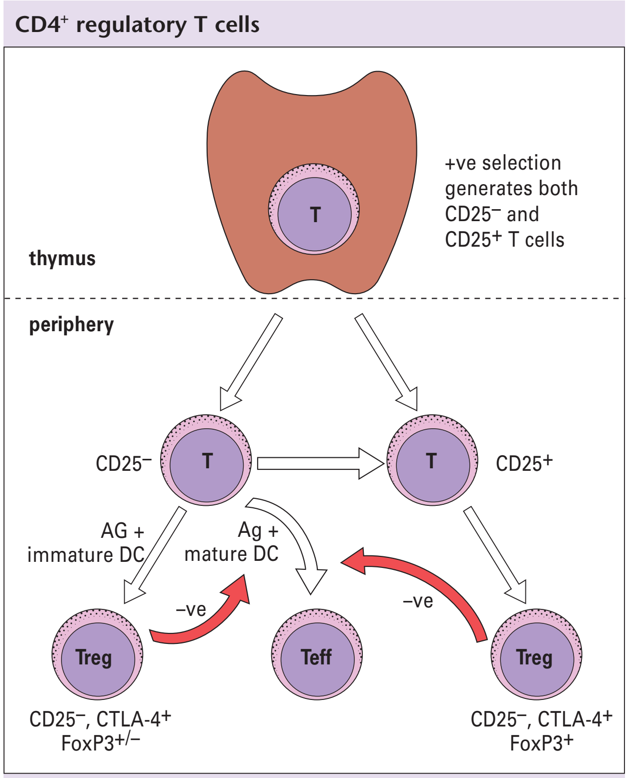
CD4$^+$ regulatory T cells
thymus
+ve selection generates both CD25– and CD25+ T cells
T
periphery
CD25–
T
T
CD25+
AG + immature DC
Ag + mature DC
–ve
–ve
Treg
Teff
Treg
CD25–, CTLA-4+ FoxP3+/–
CD25–, CTLA-4+ FoxP3+

Fig. 19.17 The thymus generates both CD25$^-$ and CD25$^+$ regulatory T (Treg) cells. T cells respond to antigen presented by mature dendritic cells by differentiating into effector T cells (Teff), secreting cytokines and providing help for cytotoxic T and B cells. Antigen presented by immature dendritic cells drives the differentiation of CD25$^-$ cells, such as IL-10$^+$ Treg cells that suppress effector cell generation. Thymus-derived CD25$^+$ Treg cells block expansion of the effector cell population. CD25$^+$ FoxP3$^+$ Treg cells can also be generated from CD25$^-$ precursors in peripheral lymphoid tissues. Both these and other regulatory populations depend on the production of cytokines, such as IL-10 and TGFβ, for suppression in vivo, although the requirement for cytokines depends on the nature of the effector T cell response. (Based on Wraith DC, Nicolson KS, Whitley NT. Curr Opin Immunol 2004;16:695. Copyright 2004 with permission from Elsevier)

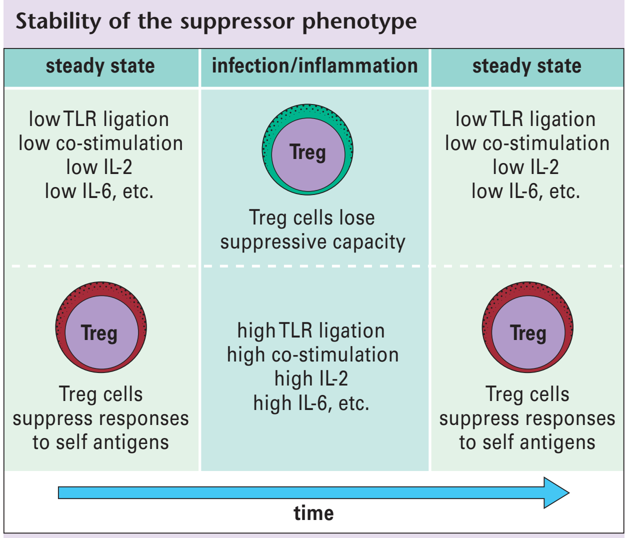
Stability of the suppressor phenotype
steady state
infection/inflammation
steady state
low TLR ligation
low co-stimulation
low IL-2
low IL-6, etc.
Treg
Treg cells lose suppressive capacity
low TLR ligation
low co-stimulation
low IL-2
low IL-6, etc.
Treg
Treg cells suppress responses to self antigens
high TLR ligation
high co-stimulation
high IL-2
high IL-6, etc.
Treg
Treg cells suppress responses to self antigens
time

Fig. 19.18 Treg cells suppress the response of autoreactive (red) cells to self antigens in the steady state by ensuring that they do not reach the threshold for activation (dashed line). Their ability to suppress responses can be overcome in the face of infection, as a result of TLR ligation or increasing levels of co-stimulation and cytokines secreted in response to infection. This might permit localized activation of some autoreactive (green) cells through bystander activation or molecular mimicry. Their expansion will, however, be prevented by clonal competition from cells specific for antigens carried by the infectious agent. Available evidence suggests that Treg cells retain their suppressive capacity and will therefore continue to suppress the response to self antigens when the infection has been cleared. (Based on Wraith DC, Nicolson KS, Whitley NT. Curr Opin Immunol 2004;16:695. Copyright 2004 with permission from Elsevier)

TOLERANCE IS ALSO IMPOSED ON B CELLS

High-affinity IgG production is T cell dependent. For this reason, and because the threshold of tolerance for T cells is lower than that for B cells, the simplest explanation for non-self reactivity by B cells is a lack of T cell help.

Until quite recently it was therefore felt that immunological tolerance should be the responsibility of T lymphocytes alone, the logical argument being: Why evolve a complex process of immunological tolerance among cells such as B cells when they are then allowed to hypermutate? All nature would have to do would be to evolve a system in which T cells never reacted with self antigens; a 'perfect' system would delete all self-reactive T cells. The immune system could then allow the B cell repertoire to expand widely and generate as many autoreactive B cells as possible by random rearrangement. Without help from T cells these B cells would remain harmless.

Now we appreciate that the immune system is not allowed to be 'perfect'. If it were, it would almost inevitably develop 'holes' in the immunological repertoire through which faster evolving microorganisms would inevitably break. We now appreciate that autoreactive T cells exist in all of us and that the balance between health and autoimmunity is a fine one.

Q. Give one piece of evidence that autoimmunity occurs in normal individuals.

A. It is possible to generate autoimmune conditions in normal animals by injection of self molecules in adjuvant (e.g. experimental autoimmune encephalomyelitis). Autoreactive T cells can be generated in vitro from lymphoid tissues of healthy individuals by appropriate stimulation with antigen–APCs and cytokines.

The B lymphocyte pool is subject to analogous but subtly different mechanisms of immunological tolerance to those that apply to the T lymphocyte pool.

There are clearly circumstances in which B cells must be tolerized directly. For example, some microorganisms have cross-reactive antigens that have both foreign T cell-reactive epitopes and other epitopes that resemble self epitopes and are capable of stimulating B cells. Such antigens could provoke a vigorous antibody response to self antigens (Fig. 19.19).

Furthermore, in contrast to TCRs, the immunoglobulin receptors on mature, antigenically stimulated B cells can undergo hypermutation and may acquire anti-self reactivities at this late stage. Tolerance thus needs to be imposed on B cells both during their development and after antigenic stimulation in secondary lymphoid tissues.

Self-reactive B cells may be deleted or anergized, depending on the affinity of the B cell antigen receptor and the nature of the antigen.

Tolerance induction by self antigens can lead to one of several results, such as deletion or anergy. The outcome depends on:

- the affinity of the B cell antigen receptor; and
- the nature of the antigen it encounters, whether this is an integral membrane protein or a soluble and largely monomeric protein in the circulation.

The fate of self-reactive B cells has been determined using transgenic technology (Figs 19.20 and 19.21).

Breakdown of B cell tolerance caused by cross-reactive antigen

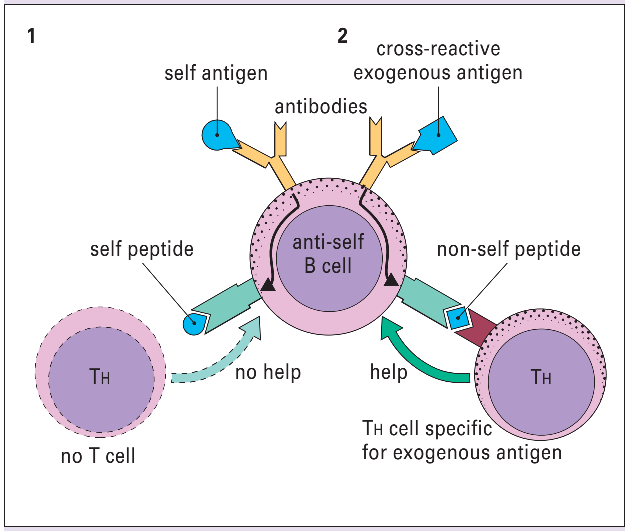

Fig. 19.19 (**1**) If TH cells are not available, either because of a hole in the T cell repertoire or because of deletion resulting from self tolerance achieved intrathymically, any B cells that are self-reactive will be unable to mount an anti-self antibody response. (**2**) Autoantibodies can be produced if an anti-self B cell collaborates with an anti-non-self TH cell in response to cross-reactive antigens containing both self and non-self determinants.

B cells pass through several developmental checkpoints

B cell development shows similar features to T cell development, but takes place largely in the bone marrow. Development is marked by ordered expression of surrogate receptor molecules.

The first notable event takes place in the late pro-B stage when the CD79a and CD79b (Igα and Igβ) molecules appear at the cell surface (see Figs 8.1 and 8.3). Progression to the pre-B stage is accompanied by VDJ recombination at the heavy chain locus. The rearranged heavy chain can then appear at the cell surface in combination with both CD79a and CD79b and the $V_{\text{pre-B}}$ and $\lambda 5$ molecules that act as surrogate light chains (see Fig. 8.2).

There are clear analogies with T cell development here – T cells rearrange the heavy chain genes first and

Tolerance induction in peripheral B cells by clonal deletion

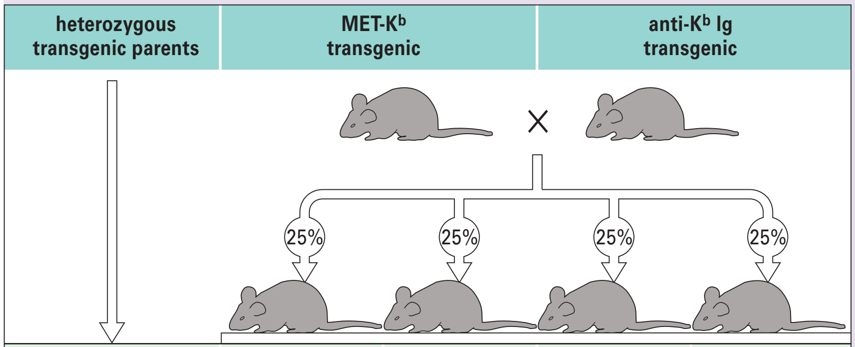

F$_1$ generation	non-transgenic	MET-K^b transgenic	anti-K^b Ig transgenic	double transgenic
anti-K^b B cell reactivity	normal	no IgG anti-K^b reactivity	Ig anti-K^b antibody present	anti-K^b B cells deleted from lymph nodes and spleen

Fig. 19.20 Non-b haplotype mice were given the gene for H-2K^b, which is a foreign MHC class I molecule. The gene was controlled by the metallothionein promoter, specific for such sites as the liver (MET-K^b transgenic). These mice were crossed with other non-b mice, which had been given the genes for anti-H-2K^b antibodies (anti-K^b Ig transgenic). Double transgenic offspring expressed H-2K^b in the liver and exported B cells specific for H-2K^b from the bone marrow. However, these self-reactive B cells were partially deleted in the spleen and entirely deleted in the lymph nodes and thus no autoantibody was produced – no idiotype corresponding to the anti-K^b Ig was detectable.

B cell tolerance induction to a soluble protein by clonal anergy

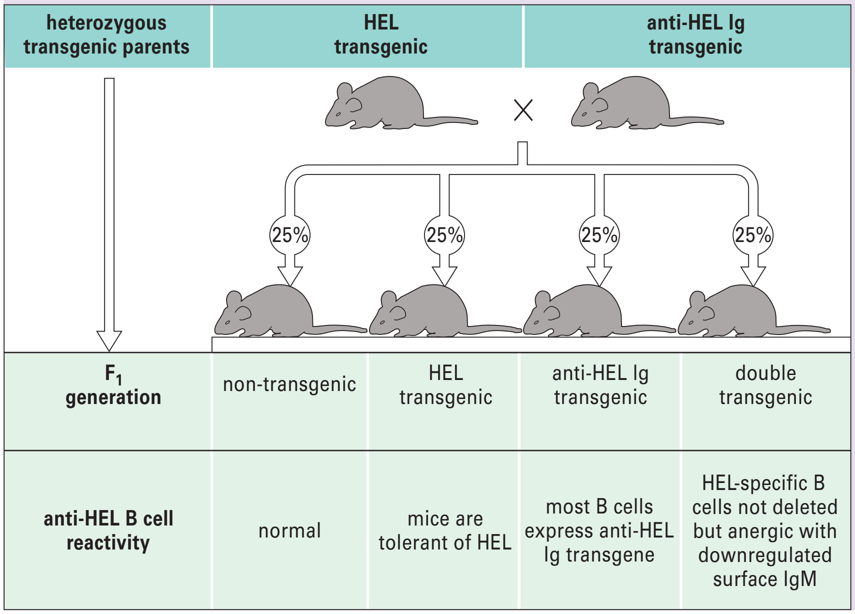
heterozygous transgenic parents
HEL transgenic
anti-HEL Ig transgenic
25%
25%
25%
25%
F_1 generation
non-transgenic
HEL transgenic
anti-HEL Ig transgenic
double transgenic
anti-HEL B cell reactivity
normal
mice are tolerant of HEL
most B cells express anti-HEL Ig transgene
HEL-specific B cells not deleted but anergic with downregulated surface IgM

Fig. 19.21 A mouse was given the hen egg lysozyme gene (HEL) linked to a tissue-specific promoter. The (largely soluble) HEL induced B and T cell tolerance. A second transgenic line (anti-HEL Ig) carried rearranged heavy and light chain genes encoding a high-affinity HEL antibody. An allotypic marker (IgH^a) distinguished this from endogenous immunoglobulin (IgH^b). The majority of B cells in these transgenics carried IgM and IgD of the 'a' allotype. Double transgenic offspring were highly HEL tolerant, producing neither anti-HEL antibody nor antibody-secreting B cells. HEL-binding (self reactive) B cells were not, however, deleted, but had downregulated surface IgM, but not IgD, receptors. They behaved as anergic cells.

also employ a surrogate light chain. In addition, it appears that expression of the pre-B cell receptor is necessary for successful allelic exclusion at the heavy chain locus.

As pre-B cells develop, recombination at the light chain locus proceeds and the cells become immature B cells. At this point the cells are IgD^-.

B cells then go through a 'transitional' stage in development becoming IgD^{low} as well as IgM^+. Mature B cells express IgD at higher levels than IgM.

Checkpoints in B cell development include:

- successful expression of CD79a and CD79b in late pro-B cells;
- successful rearrangement at the heavy chain locus in pre-B cells;
- successful rearrangement at the light chain locus and receptor editing.

Self tolerance begins when IgM first appears at the surface of the developing cell.

Immature cells are resistant to apoptosis, though the development of immature B cells can still be blocked by interaction with self antigen. However, immature B cells can edit their receptor, thus allowing development to progress.

Receptor editing allows potentially self-reactive B cells to continue development

When the IgM receptor on an immature bone marrow B cell reacts with self antigen further cell differentiation is blocked, but light chain rearrangement can continue.

If the new IgM receptor does not react with a self antigen in bone marrow, B cell development can proceed.

Interestingly the 'immature' B cell is relatively resistant to apoptotic cell death, whereas the later 'transitional' cell is sensitive.

Allowing light chain rearrangement to continue among immature cells permits the B cell to edit its receptor and thus rescue potentially autoreactive cells from inevitable death. This mechanism of altering offending receptors before the B cell becomes sensitive to antigen-mediated cell death clearly allows the immune system to optimize the generation of its repertoire.

In fact it appears that mechanisms have evolved to promote receptor editing in B cells. The κ light chain locus can be inactivated by recombination of a recombining sequence (RS). Recombination of the RS element results in deletion of CK and other sequences required for transcription of the κ allele. It is important to appreciate that 40–60% of $IgM^+\lambda^+$ B cells carry a VκJκ rearrangement inactivated by this RS recombination. Receptor editing therefore plays an important role in the generation of the normal B cell repertoire.

Self-reactive B cells are usually deleted

As B cells mature through the transitional stage they become poor at reactivating the *RAG* genes. So, if receptor editing has failed to eliminate autoreactive B cells, they are likely to be eliminated by apoptotic death because they can no longer select a new receptor. Interestingly these transitional $IgM^{hi}IgD^{lo}$ cells can emigrate to the periphery where their high IgM level ensures that apoptotic cell death can still take place.

The spleen eliminates unwanted B cells as effectively as the thymus removes unwanted T cells

Recent studies have shown that the bone marrow exports a higher proportion of new B cells than the thymus does new T cells. Also, unlike T cells leaving the thymus, B cells leaving the bone marrow are relatively immature. These cells express the heat-stable antigen (HSA) and migrate from the bone marrow to the outer T cell zone of the spleen.

It is among these cells that the most significant amount of negative selection takes place. This splenic environment eliminates unwanted B cells as effectively as the thymus removes unwanted T cells (Fig. 19.22). Self-reactive B cells are purged by a process that:

- induces anergy;
- prevents migration into B cell follicles; and
- rapidly leads to cell death.

Self-reactive B cells die in peripheral lymphoid tissues

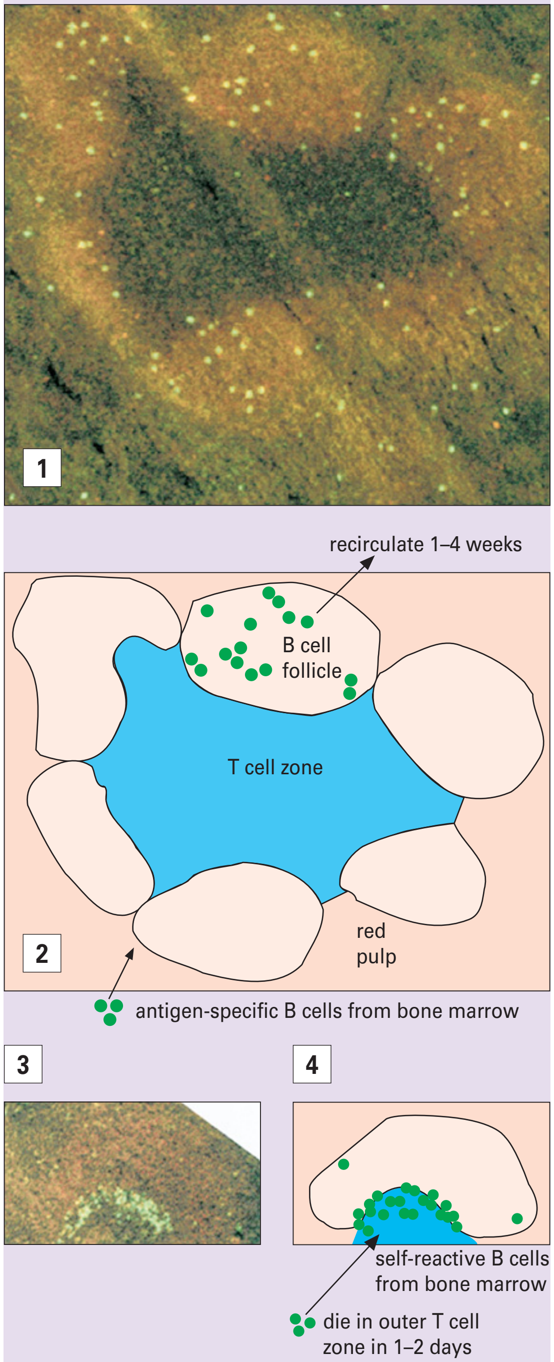
1
recirculate 1–4 weeks
B cell follicle
T cell zone
red pulp
2
antigen-specific B cells from bone marrow
3
4
self-reactive B cells from bone marrow
die in outer T cell zone in 1–2 days

Fig. 19.22 Transgenic B cells expressing a B cell receptor specific for lysozyme were stained green and adoptively transferred into either normal mice (**1** and **2**) or mice expressing soluble lysozyme as a self antigen (**3** and **4**). In the absence of antigen the B cells migrate to B cell follicles. In the presence of the self antigen the B cells are excluded from B cell follicles in the outer T cell zone where they die in 1–2 days. (From Townsend SE, Weintraub BC, Goodnow CC. Immunol Today 1999;20:217. Copyright 1999 with permission from Elsevier)

Anergy in an autoreactive B cell can be overcome by high-avidity antigen

Self-reactive B cells in the outer T cell zone are short-lived (1–3 days), whereas cells selected to enter B cell follicles become long-lived and recirculate for from 1 to 4 weeks.

Short-lived, autoreactive B cells may, however, contribute to immune responses to foreign antigens. Anergy in such an autoreactive B cell can be overcome by high-avidity antigen.

Self-reactive B cells may then be recruited into the functional immune repertoire because their B cell receptor cross-reacts sufficiently strongly with foreign antigen.

At first sight, allowing B cell tolerance to self antigen to be overwhelmed by foreign antigen would seem risky. Why promote a pluripotent B cell repertoire at the risk of autoimmunity? This reflects a balance between the risk of autoimmunity and protection of the species from infectious pandemics.

The immune system has evolved to ensure that among the population there will be individuals equipped to fight almost any infection.

Diversity among T cells is generated by rearrangement of their receptors in concert with MHC polymorphism. MHC polymorphism and TCR selection together provide sufficient diversity across the population to protect against almost any infection. MHC polymorphism has evolved to fill holes in the protective repertoire of the species.

Likewise the fact that the B cell pool permits a proportion of short-lived and potentially autoreactive B cells to emigrate to the periphery provides an additional level of protection from infection. In the absence of an infection these cells die rapidly. This short-lived pool of cells can, however, contribute to broadening the potential B cell repertoire. Such repertoire diversity and further differences between the repertoires of individuals within the population will ensure that there will always be some individuals ready to mount an effective immune response to infection.

A second window of susceptibility to tolerance occurs transiently during the generation of B cell memory. Secondary B cells (derived from memory B cells produced by T cell-dependent stimulation) are highly susceptible to tolerance by epitopes presented multivalently in the absence of T cell help. Such a tolerance-susceptible stage probably ensures that newly derived memory B cells that have acquired self reactivity (as a result of accumulated somatic mutations) are purged from the repertoire.

TOLERANCE CAN BE INDUCED ARTIFICIALLY IN VIVO

Chimerism is associated with tolerance

Tolerance can be induced by the inoculation of allogeneic cells into hosts that lack immunocompetence, for example

neonatal hosts or adult hosts after immunosuppressive regimens such as:
- total body irradiation;
- drugs (e.g. ciclosporin);
- anti-lymphocytic antibodies (anti-lymphocyte globulin, anti-CD4 antibodies, etc.).

For tolerance to be maintained, a certain degree of chimerism – the coexistence of cells from genetically different individuals – must be maintained. This is best achieved if the inoculum contains cells capable of self renewal (e.g. bone marrow cells).

If mature T cells are present in the injected cell population, they may react against the histocompatibility antigens of their host and induce a severe and often fatal disease known as graft versus host disease.

Antibodies to co-receptor and co-stimulatory molecules induce tolerance to transplants

Tolerance of transplanted tissues can be achieved in adult animals by monoclonal antibodies directed against the T cell molecules, CD4 and CD8 (the antibodies can be either the T cell-depleting or the non-depleting type). In this situation, tolerance of skin allografts is obtained even in the absence of cellular chimerism.

Q. How could an anti-CD4 antibody have its effect, if it does not deplete CD4 T cell numbers?

A. The antibody may inhibit the interaction of T cells with APCs, and the assembly of the immunological synapse (see Figs 7.19 and 7.20).

Another highly promising approach to transplantation tolerance has arisen through the use of agents designed to blockade co-stimulatory molecules.

As mentioned above, T cells require co-stimulatory signals for effective priming. CD28 and CD154 both play important roles in co-stimulation.

The CD28 pathway of activation can be inhibited by blocking both B7 molecules with a soluble form of CTLA-4 (CTLA-4–Ig). In combination with an antibody to the ligand for CD40 (CD154), CTLA-4–Ig has been shown to block recognition of allografts and allow long-term skin allograft survival in mice.

Antibodies to CD154 alone will prolong renal allograft survival in non-human primates. It is thought that anti-CD154 prevents the three-cell interplay between CD4, CD8, and dendritic cells that is required for the maturation of CD8 cells (Fig. 19.23).

Soluble antigens readily induce tolerance

Tolerance is inducible in both neonatal and adult animals by administering soluble protein antigens in deaggregated form. T and B cells differ in their susceptibility to tolerization by these antigens. Thus:
- tolerance is achieved in T cells from spleen and thymus after very low antigen doses and within a few hours;
- tolerance of spleen B cells requires much more time and higher doses of antigen (Fig. 19.24).

The antigen levels that will produce B cell tolerance in neonates are about one-hundredth of those required in adults.

CD40–CD40L interaction leads to activation of APC

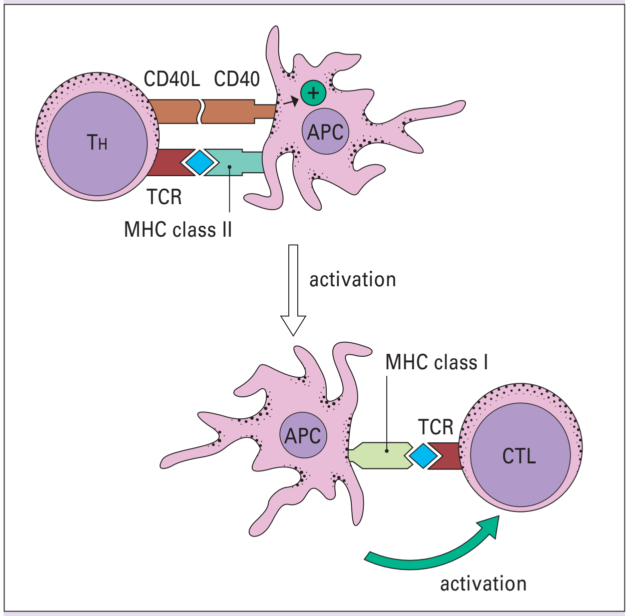

Fig. 19.23 CD8 T killer cells will not respond to resting APCs. T helper (TH) cell recognition of antigen is one way that APCs such as dendritic cells can be armed to present antigen to CD8 cells. The molecules responsible for the interaction between CD4 TH cells and dendritic cells are called CD40L and CD40. Ligation of CD40 on the surface of dendritic cells leads to their activation. Blocking this interaction is one way to prevent the presentation of alloantigens to CD8 cells and hence prevent allograft rejection. (Based on Lanzavecchia A. Nature 1998;393:413)

Oral administration of antigens induces tolerance by a variety of mechanisms

High doses of orally administered antigen can cause anergy or deletion; this could be one situation in which the term anergy is being used to describe a state of paralysis preceding cell death.

Lower doses of orally administered antigen can, however, induce the priming of T cells in the gut. As an effective antibody response in the gut requires class switching to the IgA isotype, it is no surprise that antigen feeding induces T cells that support IgA production. These 'Tr1-like' cells produce cytokines including IL-10 and TGFβ, and serve to inhibit inflammatory responses mediated by TH1 cells. TGFβ:
- inhibits the proliferation and function of B cells, cytotoxic T cells, and NK cells;
- inhibits cytokine production in lymphocytes; and
- antagonizes the effects of TNF.

Although the induction of mucosal Tr1 cells is antigen specific, the suppressive activity of cytokines such as TGFβ is not. Hence, the induction of oral tolerance to one antigen is able to suppress the immune response to a second, associated antigen.

The effect of '**bystander suppression**' by mucosal Tr1 cells allows the suppression of complex organ-specific

Relative susceptibilities of B cells and T cells to tolerance induction in vivo

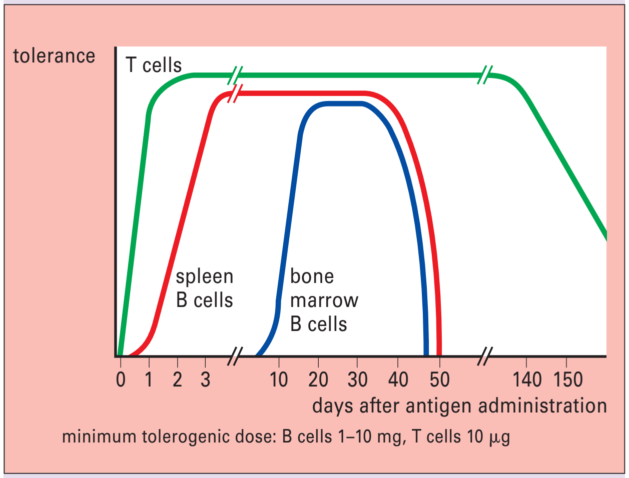

Fig. 19.24 A mouse was given a T-dependent antigen (human γ-globulin) at tolerance-inducing doses, and the duration of tolerance was measured. T cell tolerance was more rapidly induced and more persistent than B cell tolerance. Bone marrow B cells may take considerably longer to tolerize than splenic B cells. Typically, much less antigen is needed to tolerize T cells – 10 µg as opposed to 1–10 mg, a 1000-fold difference.

autoimmune diseases by feeding with a single antigen derived from the affected tissue.

Other mucosal surfaces are proving effective routes for inducing antigen-specific tolerance

Other mucosal surfaces are proving equally effective as routes for the induction of antigen-specific tolerance.

Nasal deposition of MHC class II-restricted peptides has been used to control both ‘humoral’ and ‘cellular’ immune responses. In addition, administration of aerosolized antigen to the lung can be used to control allergic responses to foreign antigens and autoimmune responses to self antigens.

Q. Why is the treatment described above not an example of immune deviation?

A. The fact that one can inhibit both autoimmune (TH1 type) and allergic (TH2 type) responses argues against immune deviation as the mechanism of tolerance induction.

Recent evidence suggests that other cells producing IL-10 may be responsible for nasal peptide-induced tolerance. It could be that other cell types, including CD8 cells, are responsible for suppressing the immune response that follows the aerosol administration of protein antigens.

Extensive clonal proliferation can lead to exhaustion and tolerance

Tolerance in T cells, and to a lesser extent in B cells, can be due to clonal exhaustion – the end result of a powerful immune response. Repeated antigenic challenge may stimulate all the antigen-responding cells to differentiate into short-lived end cells, leaving no cells that can respond to a subsequent challenge with antigen.

Anti-idiotypic responses can be associated with tolerance

An antibody’s combining site may act as an antigen and induce the formation of ‘anti-idiotypic antibodies’. By cross-linking immunoglobulin on B cells, these antibodies can block B cell responsiveness of the cell. Because most of the antibodies produced in response to particular antigens in some animals bear a particular idiotype, suppression of this idiotype by anti-idiotypic antibody can significantly alter the response. This type of tolerance will be partial, however, because it affects only those B cells carrying the idiotype.

The idiotype of a T cell is represented by the polymorphic regions of the TCR α and β chains. TCR-specific regulatory cells, induced after vaccination with self-reactive T cells, TCR peptides, and even DNA-encoding TCR molecules, have been shown to prevent autoimmunity in animal models.

Tolerance in vivo relates to persistence of antigen

Persistence of antigen plays a major part in maintaining a state of tolerance to it in vivo – when the antigen concentration decreases below a certain threshold, responsiveness is restored.

If the tolerance results from clonal deletion, recovery of responsiveness is related to the time required to generate new lymphocytes from their precursors. This tolerance can be prevented by measures such as thymectomy.

If, on the other hand, the tolerance is maintained by a suppressive mechanism resulting, for example, from the induction of regulatory T cells, then the state of tolerance can be relatively long-lived.

Tolerance could be used to control damaging immune responses

A better understanding of tolerogenesis could be valuable in many ways. It could be used to:

- promote tolerance of foreign tissue grafts;
- control the damaging immune responses in hypersensitivity states and autoimmune diseases.

The various ways of establishing artificial tolerance in adult animals are being investigated for their potential clinical applications.

Some success has been obtained in the case of transplants associated with chimerism and performed under the umbrella of immunosuppressive agents.

Treatment with monoclonal non-depleting anti-CD4 and anti-CD8 or anti-CD154 antibodies has also been used successfully with foreign tissue or organ transplants.

The possibility that tolerance can be induced either by mucosal administration of the target antigen or through the use of peptide drugs awaits extensive clinical trials in humans.

It is also important to learn how to activate T cells that ignore certain antigens to enable the immune system to mount an appropriate active response. This could be exploited to limit the growth of tumors that may express their own unique tumor-specific genes.

FURTHER READING

Gotter J, Kyewski B. Regulating self-tolerance by deregulating gene expression. Curr Opin Immunol 2004;16:741–745.

Radtke F, Wilson A, MacDonald HR. Notch signaling in T- and B-cell development. Curr Opin Immunol 2004;16:174–179.

Singer A. New perspectives on a developmental dilemma: the kinetic signaling model and the importance of signal duration for the CD4/CD8 lineage decision. Curr Opin Immunol 2002;14:207–215.

Su MA, Anderson MS. *Aire*: an update. Curr Opin Immunol 2004;16:746–752.

Venanzi ES, Benoist C, Mathis D Good riddance: thymocyte clonal deletion prevents autoimmunity. Curr Opin Immunol 2004;16:197–202.

Verkoczy LK, Martensson AS, Nemazee D. The scope of receptor editing and its association with autoimmunity. Curr Opin Immunol 2004;16:808–814.

Wraith DC, Nicolson KS, Whitley NT. Regulatory CD4⁺ T cells and the control of autoimmune disease. Curr Opin Immunol 2004;16:695–701.

Zamoyska R, Lovatt M. Signalling in T-lymphocyte development: integration of signalling pathways is the key. Curr Opin Immunol 2004;16:191–196.

Critical thinking: Tolerance (see p. 499 for explanations)

Lineage commitment in thymocyte development

Thymocytes bearing MHC class I-restricted TCRs differentiate into CD8 T cells, while those recognizing MHC class II molecules become CD4 T cells. It was previously thought that this process might be random or ‘stochastic’, but it now appears that this is an ‘instructive’ process.

AND mice express a transgenic TCR specific for a peptide derived from pigeon cytochrome c in the context of I-E^k (class II). Normally the majority of thymocytes in these mice develop into CD4 single positive (SP) cells. OT-1 mice bear a transgenic TCR specific for an ovalbumin peptide that is H-2K^b (class I) restricted. Normally the majority of thymocytes in these mice develop into CD8 SP cells. TCR$^{+/-}$ heterozygotes were crossed with heterozygotes expressing either a constitutively active (dLGF) or catalytically inactive (dLGKR) form of Lck.

Thymocytes from AND, AND/dLGKR, OT-1, or OT-1/dLGF were stained with antibodies specific for CD4 or CD8 and analysed by flow cytometry. Table 1 shows the absolute number of CD4 and CD8 SP cells obtained from three individual experiments (median values) in which littermates were studied.

experimental mouse	CD4 cells	CD8 cells	experimental mouse	CD4 cells	CD8 cells
AND	487	8	OT-1	6	113
AND/dLGKR	4	39	OT-1/dLGF	380	37

Data summarized from Hernandez-Hoyos G, Sohn SJ, Rothenberg EV, Alberola-Ila J. Immunity 2000;12:313–322.

Table 1

1. What drives expression of CD4 cells in a mouse expressing a class I-restricted TCR?
2. What effect does inhibiting the activity of Lck have on thymocyte development?
3. Are the results seen with the four different strains of mice mutually consistent?

The molecular basis of activation-induced cell death (AICD)

The 3A9 transgenic mouse expresses a TCR specific for a peptide derived from hen egg white lysozyme. These mice were crossed onto the lpr, gld, or TNF-R1 knockout background. Remember that the lpr and gld mice have defective Fas and FasL genes, respectively.

Previously activated T cells may undergo AICD on further restimulation in vitro. T cells from the 3A9-backcrossed mice were activated in vitro and then restimulated with anti-CD3 (to induce AICD), IL-2, or IL-2 with anti-Fas antibody (to induce death via the Fas pathway). The cells were then assayed for apoptosis by propidium iodide staining (Table 2).

Critical thinking: Tolerance *(Continued)*

mice	apoptotic T cells		
	anti-CD3	IL-2	IL-2 and anti-Fas
3A9	++	–	++
3A9.lpr knockout	–	–	–
3A9.TNF-R1 knockout	++	–	++
3A9.gld knockout	–	–	++

Data summarized from Refaeli Y, Van Parijs L, Abbas AK. Immunol Rev 1999;169:273–282.

Table 2

4 Is AICD mediated by signaling through the Fas or TNF-R1 pathway?
5 What evidence is there to support your choice?
6 Why are T cells from the gld mouse susceptible to anti-Fas-mediated death?
7 Why has the immune system developed a way of killing off activated T cells?
8 If T cells die on reactivation, how does immune memory arise?

The role of regulatory T cells in peripheral tolerance

Certain mice are congenitally athymic and do not have mature T lymphocytes. The 'nude' mouse carries a mutation in the gene for the Wnt transcription factor. This results in lack of hair and absence of the thymus. Nude mice can, however, be reconstituted with lymphocytes from normal mice. In the experiment for which results are shown in Table 3, either thymocytes or spleen and lymph node (SP/LN) cells from normal mice were treated with various antibodies and complement to deplete cell subsets and then transferred into 'nude' recipients by intravenous injection. The recipient mice were left for 3 months and then examined for histological and serological evidence of autoimmune disease.

inoculated cells	number of mice with autoimmune disease					
	number of mice	OOP	THR	SIAL	ADR	GN
A whole thymocytes	12	0	0	0	0	0
B CD25^{-} thymocytes	12	12	4	3	1	3
C CD4^{+}8^{-} thymocytes	5	0	0	0	0	0
D CD25^{-}4^{+} thymocytes	5	5	2	1	2	1
E whole SP/LN cells	8	0	0	0	0	0
F CD25^{-} SP/LN cells	8	8	5	5	2	2
G CD25^{-} thymocytes and CD4^{+} SP/LN cells	8	0	0	0	0	0

Data summarized from Itoh M, Takahashi T, Sakaguchi N, et al. J Immunol 1999;162:5317–5326. (ADR, adrenalitis; GN, glomerulonephritis; OOP, oophoritis; SIAL, sialoadenitis; THR, thyroiditis)

Table 3

9 Why do CD25^{-} lymphocytes cause widespread autoimmunity?
10 Where do CD25^{+} cells arise?
11 Why are certain tissues more susceptible to autoimmune disease than others?

CHAPTER

20

Autoimmunity and Autoimmune Disease

SUMMARY

- **Autoimmunity is associated with disease.** Autoimmune mechanisms underlie many diseases, some organ-specific, others systemic in distribution, and autoimmune disorders can overlap – an individual may have more than one organ-specific disorder, or more than one systemic disease.
- **Genetic factors play a role in the development of autoimmune diseases.** Factors such as HLA type are important, and it is probable that each disease involves several factors.
- **Self-reactive B and T cells persist even in normal subjects.** Autoreactive B and T cells persist in normal subjects, but in disease are selected by autoantigen in the production of autoimmune responses.
- **Controls on the development of autoimmunity can be bypassed.** Microbial cross-reacting antigens and cytokine dysregulation can lead to autoimmunity.
- **In most diseases associated with autoimmunity, the autoimmune process produces the lesions.** The pathogenic role of autoimmunity can be demonstrated in experimental models. Human autoantibodies can be directly pathogenic. Immune complexes are often associated with systemic autoimmune disease. Autoantibody tests are valuable for diagnosis and sometimes for prognosis.
- **Treatment of autoimmune disease has a variety of aims.** Treatment of organ-specific diseases usually involves metabolic control. Treatment of systemic diseases includes the use of anti-inflammatory and immunosuppressive drugs. Future treatment will probably focus on manipulation of the pivotal autoreactive T cells by antigens or peptides, by anti-CD4, and possibly by T cell vaccination.

AUTOIMMUNITY IS ASSOCIATED WITH DISEASE

The immune system has tremendous diversity and, because the repertoire of specificities expressed by the B and T cell populations is generated randomly, it is bound to include many that are specific for self components. The body must therefore establish self-tolerance mechanisms to distinguish between self and non-self determinants to avoid autoreactivity (see Chapter 19). However, all mechanisms have a risk of breakdown. The self-recognition mechanisms are no exception, and a number of diseases have been identified in which there is autoimmunity, due to copious production of autoantibodies and autoreactive T cells.

One of the earliest examples in which the production of autoantibodies was associated with disease in a given organ is Hashimoto's thyroiditis.

Among the autoimmune diseases, thyroiditis has been particularly well studied, and many of the aspects discussed in this chapter will draw upon our knowledge of it. It is a disease of the thyroid that is most common in middle-aged women and often leads to formation of a goiter and hypothyroidism. The gland is infiltrated, sometimes to an extraordinary extent, with inflammatory lymphoid cells. These are predominantly mononuclear phagocytes, lymphocytes and plasma cells, and secondary lymphoid follicles are common (Fig. 20.1).

In Hashimoto's disease, the gland often shows regenerating thyroid follicles, but this is not a feature of the thyroid in the related condition, primary myxedema, in which comparable immunological features are seen and where the gland undergoes almost complete destruction and shrinks.

The serum of patients with Hashimoto's disease usually contains antibodies to thyroglobulin. These antibodies are demonstrable by agglutination and by precipitin reactions when present in high titer.

Most patients also have antibodies directed against a cytoplasmic or microsomal antigen, also present on the apical surface of the follicular epithelial cells (Fig. 20.2), and now known to be thyroid peroxidase, the enzyme that iodinates thyroglobulin.

Hashimoto's thyroiditis and SLE represent the extremes of the spectrum of autoimmune diseases

The antibodies associated with Hashimoto's thyroiditis and primary myxedema react only with the thyroid, so the resulting lesion is highly localized.

By contrast, the serum from patients with diseases such as systemic lupus erythematosus (SLE) reacts with many,

Pathological changes in Hashimoto's thyroiditis

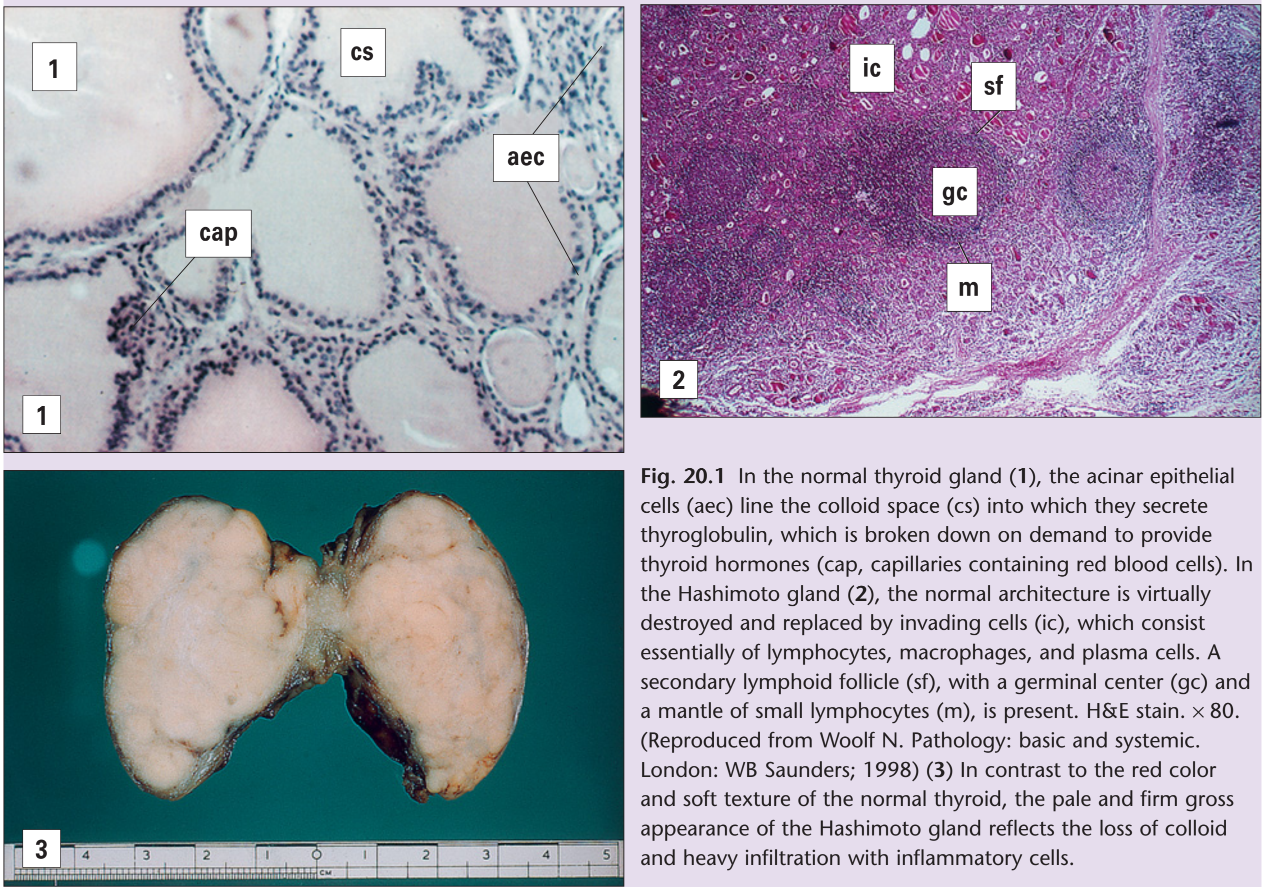

Fig. 20.1 In the normal thyroid gland (**1**), the acinar epithelial cells (aec) line the colloid space (cs) into which they secrete thyroglobulin, which is broken down on demand to provide thyroid hormones (cap, capillaries containing red blood cells). In the Hashimoto gland (**2**), the normal architecture is virtually destroyed and replaced by invading cells (ic), which consist essentially of lymphocytes, macrophages, and plasma cells. A secondary lymphoid follicle (sf), with a germinal center (gc) and a mantle of small lymphocytes (m), is present. H&E stain. × 80. (Reproduced from Woolf N. Pathology: basic and systemic. London: WB Saunders; 1998) (**3**) In contrast to the red color and soft texture of the normal thyroid, the pale and firm gross appearance of the Hashimoto gland reflects the loss of colloid and heavy infiltration with inflammatory cells.

Autoantibodies to thyroid

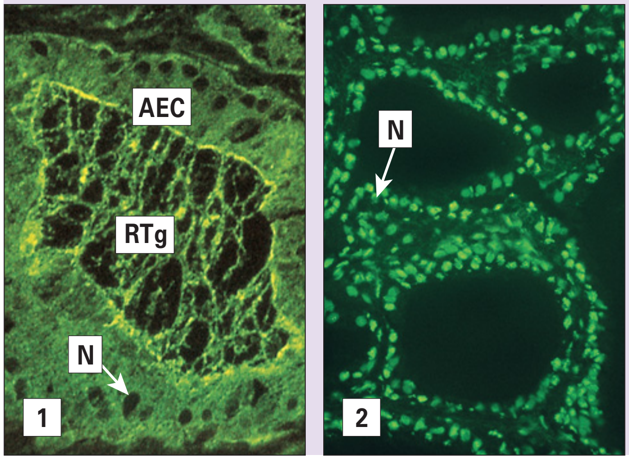

Fig. 20.2 Healthy, unfixed human thyroid sections were treated with patients' serum, and then with fluoresceinated rabbit anti-human immunoglobulin. (**1**) Some residual thyroglobulin in the colloid (RTg) and the acinar epithelial cells (AEC) of the follicles, particularly the apical surface, are stained by antibodies from a patient with Hashimoto's disease, which react with the cells' cytoplasm but not the nuclei (N). (**2**) In contrast, serum from a patient with systemic lupus erythematosus (SLE) contains antibodies that react only with the nuclei of acinar epithelial cells and leave the cytoplasm unstained. (Courtesy of Mr G Swana)

if not all, of the tissues in the body. In SLE, one of the dominant antibodies is directed against the cell nucleus (see Fig. 20.2[2]).

Hashimoto's thyroiditis and SLE represent the extremes of the autoimmune spectrum (Fig. 20.3):

- the common target organs in **organ-specific disease** include the thyroid, adrenal, stomach, and pancreas;
- the non-organ-specific diseases, often termed **systemic autoimmune diseases**, which include the rheumatological disorders, characteristically involve the skin, kidney, joints, and muscle (Fig. 20.4).

An individual may have more than one autoimmune disease

Interestingly, there are remarkable overlaps at each end of the autoimmune disease spectrum:

- thyroid antibodies occur with a high frequency in patients with pernicious anemia who have gastric autoimmunity,

The spectrum of autoimmune diseases

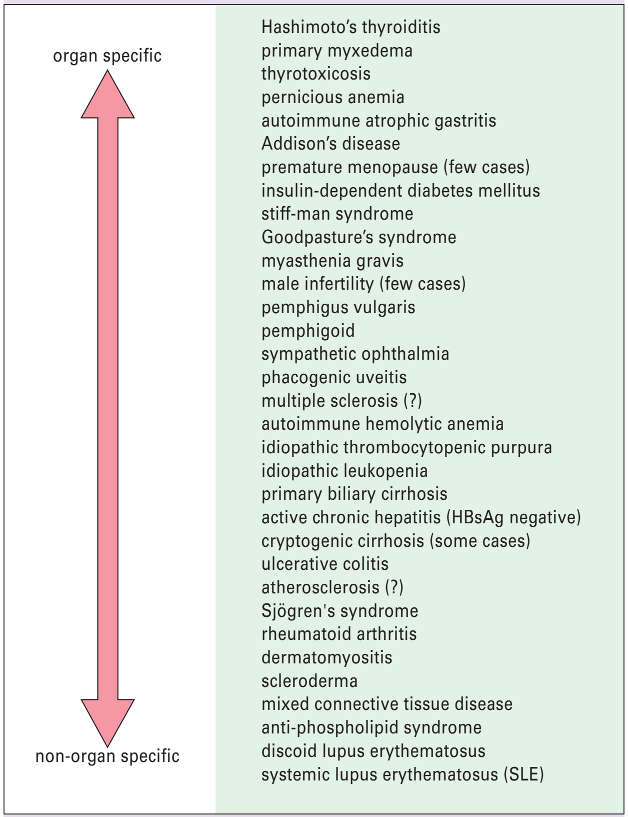

Hashimoto's thyroiditis
primary myxedema
thyrotoxicosis
pernicious anemia
autoimmune atrophic gastritis
Addison's disease
premature menopause (few cases)
insulin-dependent diabetes mellitus
stiff-man syndrome
Goodpasture's syndrome
myasthenia gravis
male infertility (few cases)
pemphigus vulgaris
pemphigoid
sympathetic ophthalmia
phacogenic uveitis
multiple sclerosis (?)
autoimmune hemolytic anemia
idiopathic thrombocytopenic purpura
idiopathic leukopenia
primary biliary cirrhosis
active chronic hepatitis (HBsAg negative)
cryptogenic cirrhosis (some cases)
ulcerative colitis
atherosclerosis (?)
Sjögren's syndrome
rheumatoid arthritis
dermatomyositis
scleroderma
mixed connective tissue disease
anti-phospholipid syndrome
discoid lupus erythematosus
systemic lupus erythematosus (SLE)

Fig. 20.3 Autoimmune diseases may be classified as organ-specific or non-organ-specific depending on whether the response is primarily against antigens localized to particular organs or against widespread antigens.

Two types of autoimmune disease

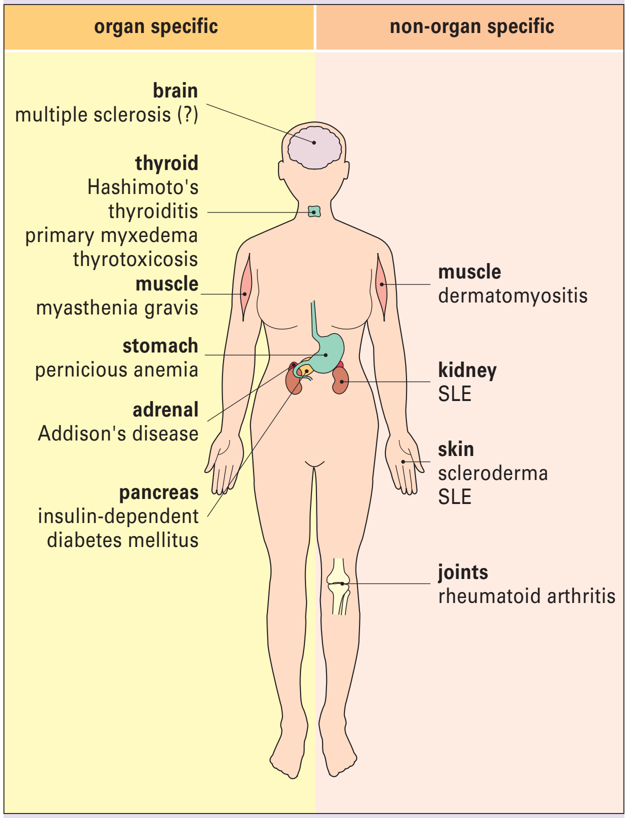

Fig. 20.4 Although the non-organ-specific diseases characteristically produce symptoms in the skin, joints, kidney, and muscle, individual organs are more markedly affected by particular diseases, for example the kidney in SLE and the joints in rheumatoid arthritis.

and these patients have a higher incidence of thyroid autoimmune disease than the normal population;

- similarly, patients with thyroid autoimmunity have a high incidence of stomach autoantibodies and, to a lesser extent, the clinical disease itself, namely pernicious anemia.

The cluster of rheumatological disorders at the other end of the spectrum also shows considerable overlap. Features of rheumatoid arthritis, for example, are often associated with the clinical picture of SLE. In these diseases immune complexes are deposited systemically, particularly in the kidney, joints, and skin, giving rise to widespread lesions.

By contrast, overlap of diseases from the two ends of the spectrum is relatively rare.

The mechanisms of immunopathological damage vary depending on where the disease lies in the spectrum:

- where the antigen is localized in a particular organ, type II hypersensitivity (e.g. autoimmune hemolytic anemia) and type IV cell-mediated reactions, as in type 1 insulin-dependent diabetes, are most important (see Chapters 24–26);
- in non-organ-specific autoimmune disorders such as SLE, type III immune complex deposition leads to inflammation through a variety of mechanisms, including complement activation and phagocyte recruitment (see Chapters 24 and 25).

GENETIC FACTORS PLAY A ROLE IN THE DEVELOPMENT OF AUTOIMMUNITY

There is an undoubted familial incidence of autoimmunity. This is largely genetic rather than environmental, as may be seen from studies of identical and non-identical twins, and from the association of thyroid autoantibodies with abnormalities of the X chromosome.

Within the families of patients with organ-specific autoimmunity, not only is there a general predisposition to develop organ-specific antibodies, it is also clear that other genetically controlled factors tend to select the organ that is mainly affected.

Thus, although relatives of patients with Hashimoto's thyroiditis and families of patients with pernicious anemia both have a higher than normal incidence and titer of thyroid autoantibodies, the relatives of patients with pernicious anemia have a far higher frequency of gastric autoantibodies (Fig. 20.5), indicating that there are genetic factors that differentially select the stomach as the target within these families.

Overlap between thyroid and gastric autoimmunity

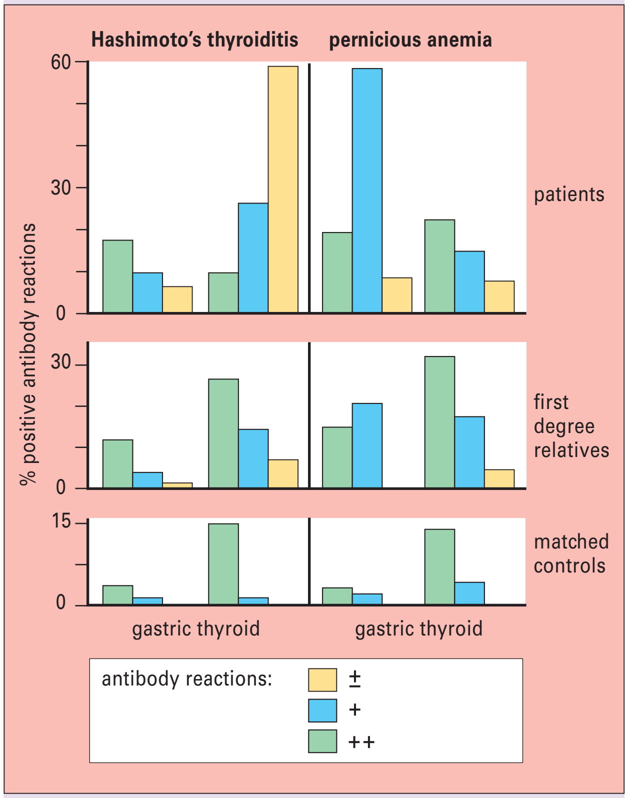

Fig. 20.5 Genetic factors affect the predisposition to autoimmunity and selection of the target organ in the relatives of patients with Hashimoto's thyroiditis and pernicious anemia. (Data adapted from Doniach D, Roitt IM, Taylor KB. Ann NY Acad Sci 1965;124:605)

Certain HLA haplotypes predispose to autoimmunity

Further evidence for the operation of genetic factors in autoimmune disease comes from their tendency to be associated with particular HLA specificities (Fig. 20.6).

Rheumatoid arthritis shows no associations with the HLA-A and -B loci haplotypes, but is associated with a nucleotide sequence (encoding amino acids 70–74 in the DRβ chain) that is common to DR1 and major subtypes of DR4. This sequence is also present in the dnaJ **heat-shock proteins** of various bacilli and EBV gp110 proteins, presenting an interesting possibility for the induction of autoimmunity by a microbial cross-reacting epitope (see below).

The plot gets even deeper, though, with the realization that HLA-DR molecules bearing this sequence can bind to another bacterial heat-shock protein, dnaK, and to the human analog, namely hsp73, which targets selected proteins to lysosomes for antigen processing.

The haplotype B8, DR3 is particularly common in the organ-specific diseases, though Hashimoto's thyroiditis tends to be associated more with DR5.

It is notable that for insulin-dependent (type 1) diabetes mellitus, DQ2/8 heterozygotes have a greatly increased risk of developing the disease (see Fig. 20.6).

HLA associations in autoimmune disease

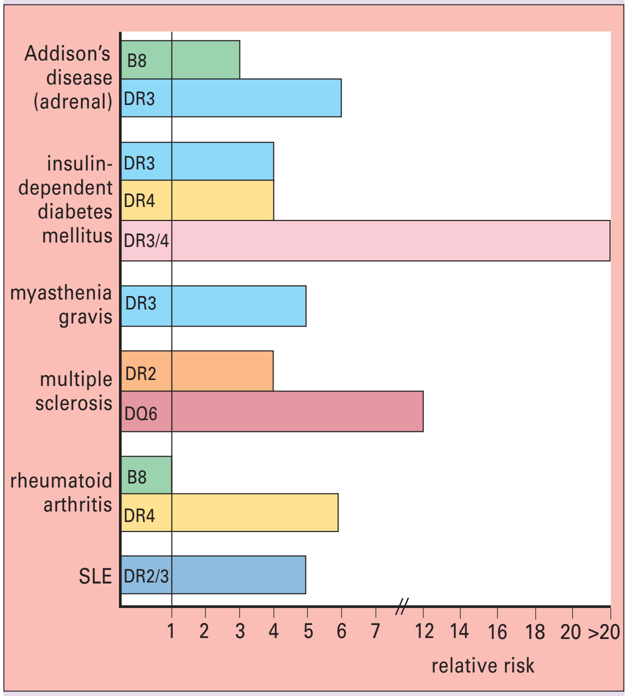

Fig. 20.6 The relative risk is a measure of the increased chance of contracting the disease for individuals bearing the HLA antigen, relative to those lacking it. Virtually all autoimmune diseases studied have shown an association with some HLA specificity. The greater relative risk for Addison's disease associated with HLA-DR3 as compared with HLA-B8 suggests that DR3 is closely linked to or even identical with the 'disease susceptibility gene'. In this case it is not surprising that B8 has a relative risk greater than 1, because it is known to occur with DR3 more often than expected by chance in the general population, a phenomenon termed linkage disequilibrium. Both DQ2 and DQ8 are also associated with type 1 diabetes mellitus, and are usually found in conjunction with the extended haplotype containing the DR3/4 disease-associated alleles. Rheumatoid arthritis is linked to a pentamer sequence in DR1 and certain subtypes of DR4, but not to any HLA-A or HLA-B alleles.

Usually, a multiplicity of genes underlies susceptibility to autoimmunity

Although HLA risk factors tend to dominate, these disorders are genetically complex and genome-wide searches for mapping the genetic intervals containing genes for predisposition to disease reveal a plethora of genes affecting:

- loss of tolerance;
- sustained inflammatory responses; and
- end-organ targeting.

A notable example is the association of autoimmune disease with a single nucleotide polymorphism (SNP) linked to CTLA-4 (Fig. 20.7), a molecule present on Tregs (see Chapter 19). It should be stressed, however, that this is a single risk factor, one of many that act in concert to provoke autoimmunity, and only rarely have

CTLA-4-linked single nucleotide polymorphism associated with autoimmune disease

signal exon1	B7 bonding domain exon 2	transmembrane domain exon 3	phosphorylation exon 4	CTLA-4 splice variant	mRNA expression of CTLA-4 isoform in:	
					susceptible human genotype	susceptible mouse (NOD) genotype
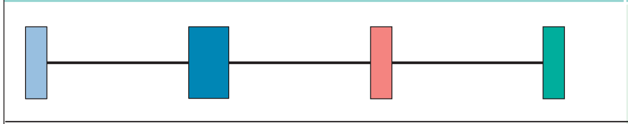				full lengh	–	–
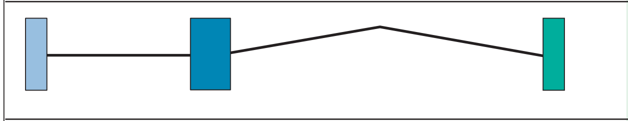				soluble	↓	–
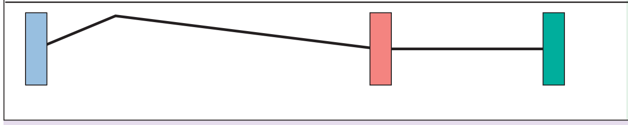				ligand independent (NOD)	–	↓

Fig. 20.7 The CT60 SNP variant in the non-coding 6.1 kb 3′ region of *CTLA-4* is responsible for the association of autoimmune disease with lower messenger RNA levels of the soluble variant splice isoforms of *CTLA-4*. (Data adapted from Ueda H, et al. Nature 2003;423:506)

single gene mutations leading to autoimmune disease been identified.

Recently it was found that mutations in the transcriptional autoimmune regulator (*Aire*) gene is associated with the multiorgan disorder known as autoimmune polyglandular syndrome type 1, characterized by Addison's disease, hypoparathyroidism, and insulin-dependent diabetes mellitus (IDDM). Deficiency in *Aire* prejudices the establishment of self tolerance to a number of organ-specific antigens such as glutamic acid decarboxylase, GAD65, one of the key autoantigens in IDDM, by preventing its expression in the thymus.

SELF-REACTIVE B AND T CELLS PERSIST EVEN IN NORMAL SUBJECTS

Despite the complex selection mechanisms operating to establish self tolerance during lymphocyte development, the body contains large numbers of lymphocytes, which are potentially autoreactive.

Thus, many autoantigens, when injected with adjuvants, make autoantibodies in normal animals, demonstrating the presence of autoreactive B cells, and it is possible to identify a small number of autoreactive B cells (e.g. anti-thyroglobulin) in the normal population.

Autoreactive T cells are also present in normal individuals, as shown by the fact that it is possible to produce autoimmune lines of T cells by stimulation of normal circulating T cells with the appropriate autoantigen (e.g. myelin basic protein [MBP]) and IL-2.

Autoimmunity results from antigen-driven self-reactive lymphocytes

Given that autoreactive B cells exist, the question remains whether they are stimulated to proliferate and produce autoantibodies by interaction with autoantigens or by some other means, such as non-specific polyclonal activators or idiotypic interactions (see below and Fig. 20.9).

Evidence that B cells are selected by antigen comes from the existence of high-affinity autoantibodies, which arise through somatic mutation, a process that requires both T cells and autoantigen. In addition, patients' serum usually contains autoantibodies directed to epitope clusters occuring on the same autoantigenic molecule. Apart from the presence of autoantigen itself, it is very difficult to envisage a mechanism that could account for the co-existence of antibody responses to different epitopes on the same molecule. A similar argument applies to the induction, in a single individual, of autoantibodies to organelles (e.g. nucleosomes and spliceosomes, which appear as blebs on the surface of apoptotic cells) or antigens linked within the same organ (e.g. thyroglobulin and thyroid peroxidase).

The most direct evidence for autoimmunity being antigen driven comes from studies of the Obese strain chicken, which spontaneously develops thyroid autoimmunity. If the thyroid gland (the source of antigen) is removed at birth, the chickens mature without developing thyroid autoantibodies (Fig. 20.8). Furthermore, once thyroid autoimmunity has developed, later removal of the thyroid leads to a gross decline of thyroid autoantibodies, usually to undetectable levels.

Comparable experiments have been carried out in the non-obese diabetic (NOD) mouse, which models human autoimmune diabetes – chemical destruction of the β cells leads to decline in pancreatic autoantibodies.

DNase treatment of lupus mice ameliorates the disease, presumably by destroying potentially pathogenic immune complexes.

In organ-specific disorders, there is ample evidence for T cells responding to antigens present in the organs under attack. But in non-organ-specific autoimmunity, identification of the antigens recognized by T cells is often inadequate. True, histone-specific T cells are generated in patients with SLE and histone could play a 'piggyback' role in the formation of anti-DNA antibodies by

Effect of neonatal thyroidectomy on Obese chickens

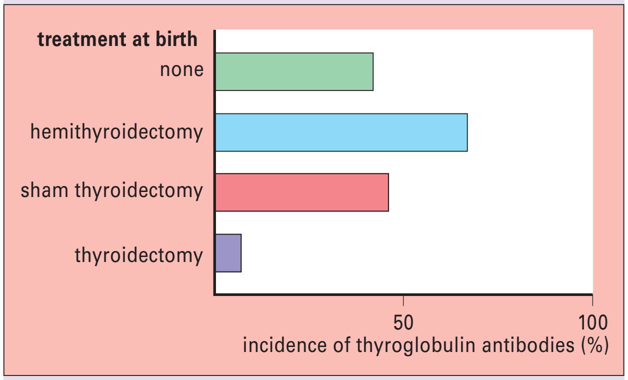

Fig. 20.8 Because removal of the thyroid at birth prevents the development of thyroid autoantibodies, it would appear that the autoimmune process is driven by the autoantigen in the thyroid gland. (Based on data from de Carvalho et al. J Exp Med 1982;155:1255)

Possible model of T cell help via processing of intermolecular complexes in the induction of autoimmunity

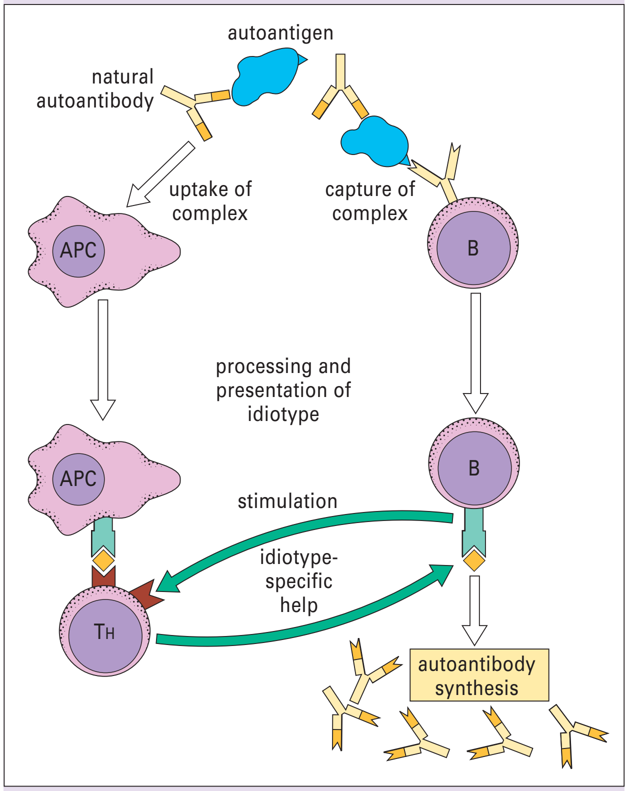

Fig. 20.9 An immune complex consisting of autoantigen (e.g. DNA) and a naturally occurring (germline) autoantibody is taken up by an antigen-presenting cell (APC), and peptides derived by processing of the idiotypic segment of the antibody (Id) are presented to TH cells. B cells that express the 'pathogenic' autoantibody can capture the complex and so can receive T cell help via presentation of the processed Id to the TH cell. Similarly, an anti-DNA-specific B cell that had endocytosed a histone–DNA complex could be stimulated to autoantibody production by histone-specific TH cells.

substituting for natural antibody in the mechanism outlined in Fig. 20.9.

Another possibility is that the T cells do not see conventional peptide antigen (possibly true of anti-DNA responses), but instead recognize an antibody's idiotype (an antigenic determinant on the V region of antibody).

In this view SLE, for example, might sometimes be initiated as an 'idiotypic disease', like the model presented in Fig. 20.9. In this scheme, autoantibodies are produced normally at low levels by B cells using germline genes. If these then form complexes with the autoantigen, the complexes can be taken up by APCs (including B cells) and components of the complex, including the antibody idiotype, presented to T cells. Idiotype-specific T cells would then help the autoantibody-producing B cells.

Evidence for the induction of anti-DNA and glomerulonephritis by immunization of mice with the idiotype of germline 'natural' anti-DNA autoantibody lends credence to this hypothesis.

CONTROLS ON THE DEVELOPMENT OF AUTOIMMUNITY CAN BE BYPASSED

Molecular mimicry by cross-reactive microbial antigens can stimulate autoreactive B and T cells

Normally, naive autoreactive T cells recognizing cryptic self epitopes are not switched on because the antigen is presented only at low concentrations on 'professional' APCs or it may be presented on 'non-professional' APCs such as pancreatic β-islet cells or thyroid epithelial cells, which lack B7 or other co-stimulator molecules.

However, infection with a microbe bearing antigens that cross-react with the cryptic self epitopes (i.e. have shared epitopes) will load the professional APCs with levels of processed peptides that are sufficient to activate the naive autoreative T cells. Once primed, these T cells are able to recognize and react with the self epitope on the non-professional APCs because they:

- no longer require a co-stimulatory signal; and
- have a higher avidity for the target, due to upregulation of accessory adhesion molecules (Fig. 20.10).

Cross-reactive antigens that share B cell epitopes with self molecules can also break tolerance, but by a different mechanism. Many autoreactive B cells cannot be activated because the $CD4^+$ TH cells they need are unresponsive, either because:

- these helper TH cells are tolerized at lower concentrations of autoantigens than the B cells; or
- because they recognize only cryptic epitopes.

However, these 'helpless' B cells can be stimulated if the cross-reacting antigen bears a 'foreign' carrier epitope to which the T cells have not been tolerized (Fig. 20.11). The autoimmune process may persist after clearance of the foreign antigen if the activated B cells now focus

Cross-reactive antigens induce autoimmune TH cells

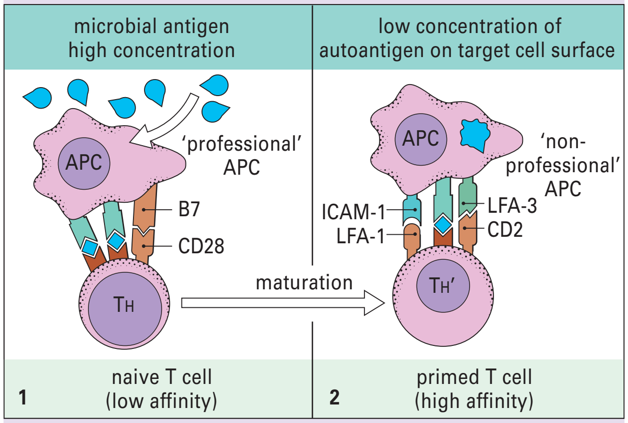

Fig. 20.10 The inability of naive TH cells to recognize autoantigen on a tissue cell, whether because of low concentration or low affinity, can be circumvented by a cross-reacting microbial antigen at higher concentration or with higher innate affinity, together with a co-stimulator such as B7 on a 'professional' APC; this primes the TH cells (**1**). Due to increased expression of accessory molecules (e.g. LFA-1 and CD2) the primed TH cells now have high affinity and, because they do not require a co-stimulatory signal, they can interact with autoantigen on 'non-professional' APCs such as organ-specific epithelial cells to produce autoimmune disease (**2**).

Induction of autoantibodies by cross-reactive antigens

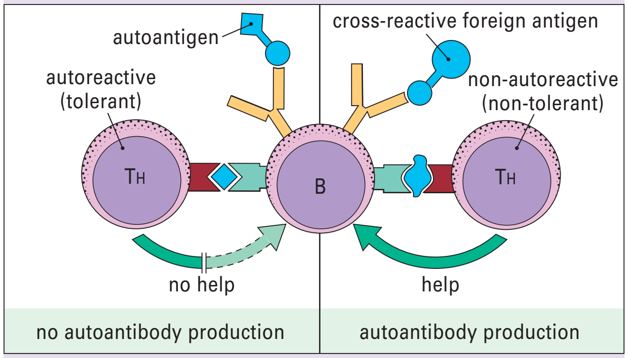

Fig. 20.11 The B cell recognizes an epitope present on autoantigen, but coincidentally present also on a foreign antigen. Normally the B cell presents the autoantigen, but receives no help from autoreactive TH cells, which are functionally deleted. If a cross-reacting foreign antigen is encountered, the B cell can present peptides of this molecule to non-autoreactive T cells and thus be driven to proliferate, differentiate, and secrete autoantibodies.

the autoantigen on their surface receptors and present it to normally resting autoreactive T cells, which will then proliferate and act as helpers for fresh B cell stimulation.

Molecular mimicry operates in rheumatic fever

A disease in which such molecular mimicry operates is rheumatic fever, in which autoantibodies to heart valve antigens can be detected. These develop in a small proportion of individuals several weeks after a streptococcal infection of the throat. Carbohydrate antigens on the streptococci cross-react with an antigen on heart valves, so the infection may bypass T cell self tolerance to heart valve antigens.

Shared B cell epitopes between *Yersinia enterolytica* and the extracellular domain of the thyroid stimulating hormone (TSH) receptor have recently been described.

There may also be cross-reactivity between HLA-B27 and certain strains of *Klebsiella* in connection with ankylosing spondylitis, and cross-reactivity between bacterial heat-shock proteins and DR4 in relationship to rheumatoid arthritis.

In some cases foreign antigen can directly stimulate autoreactive cells

Another mechanism to bypass the tolerant autoreactive TH cell is where antigen or another stimulator directly triggers the autoreactive effector cells.

For example, lipopolysaccharide or Epstein–Barr virus causes direct B cell stimulation and some of the clones of activated cells will produce autoantibodies, though in the absence of T cell help these are normally of low titer and affinity. However, it is conceivable that an activated B cell might pick up and process its cognate autoantigen and present it to a naive autoreactive T cell.

Cytokine dysregulation, inappropriate MHC expression, and failure of suppression may induce autoimmunity

It appears that dysregulation of the cytokine network can also lead to activation of autoreactive T cells.

One experimental demonstration of this is the introduction of a transgene for interferon-γ (IFNγ) into pancreatic β-islet cells. If the transgene for IFNγ is fully expressed in the cells, MHC class II genes are upregulated and autoimmune destruction of the islet cells results. This is not simply a result of a non-specific chaotic IFNγ-induced local inflammatory milieu because normal islets grafted at a separate site are rejected, implying clearly that T cell autoreactivity to the pancreas has been established.

The surface expression of MHC class II in itself is not sufficient to activate the naive autoreactive T cells, but it may be necessary to allow a cell to act as a target for the primed autoreactive TH cells. It was therefore most exciting when cells taken from the glands of patients with Graves' disease were found to be actively synthesizing class II MHC molecules (Fig. 20.12) and so were able to be recognized by $CD4^+$ T cells.

In this context it is interesting that isolated cells from several animal strains that are susceptible to autoimmunity are also more readily induced by IFNγ to express MHC class II molecules than cells from non-susceptible strains.

The argument that imbalanced cytokine production may also contribute to autoimmunity receives further support from the unexpected finding that tumor necrosis factor (TNF; introduced by means of a TNF transgene) ameliorates the spontaneous SLE of F_1 (NZB × NZW) mice.

Human thyroid sections stained for MHC class II

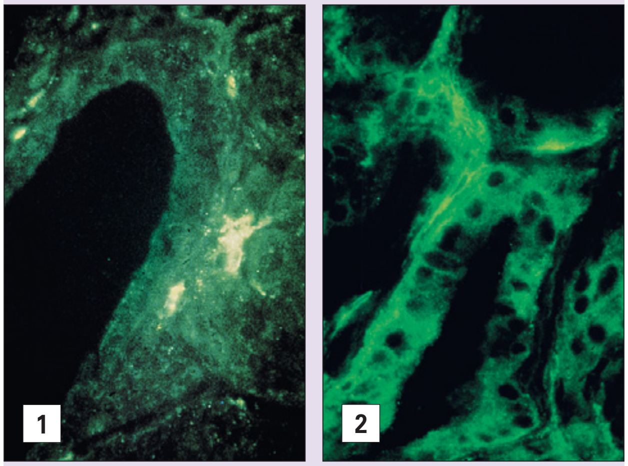

Fig. 20.12 (**1**) Normal thyroid with unstained follicular cells, and an isolated dendritic cell that is strongly positive for MHC class II. (**2**) Thyrotoxic (Graves' disease) thyroid with abundant MHC class II molecules in the cytoplasm, indicating that rapid synthesis of MHC class II molecules is occurring.

Defects in the cytokine/hypothalamic–pituitary–adrenal feedback loop in autoimmunity

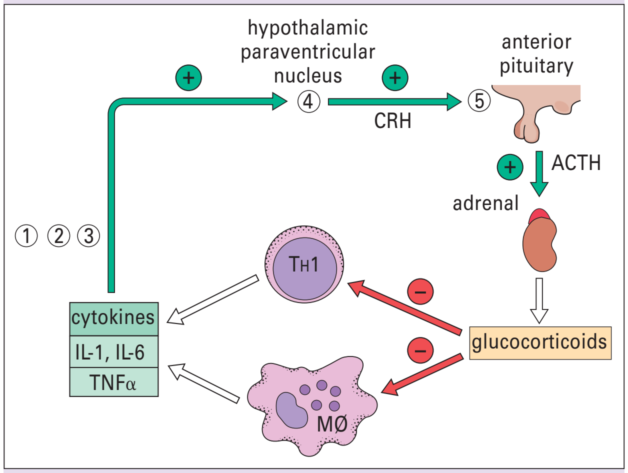

Fig. 20.13 Production of IL-1 is defective in the NOD mouse (**1**) and diabetes-prone BB rat (**2**); the disease can be corrected by injection of the cytokine. The same is true for the production of TNFα by the NZB × W lupus mouse (**3**). Patients with rheumatoid arthritis have a poor hypothalamic response to IL-1 and IL-6 (**4**). The hypothalamic–pituitary axis is defective in the Obese strain chicken and in the Lewis rat, which is prone to the development of Freund's adjuvant-mediated experimental autoimmune disease (**5**). (CRH, corticotropin releasing hormone; MØ, macrophage)

Aside from the normal 'ignorance' of cryptic self epitopes, other factors that normally restrain potentially autoreactive cells may include:

- regulatory T cells;
- hormones (e.g. steroids);
- cytokines (e.g. transforming growth factor-β [TGFβ]); and
- products of macrophages.

Deficiencies in any of these factors may increase susceptibility to autoimmunity.

The feedback loop on TH cells and macrophages through the pituitary–adrenal axis is particularly interesting because defects at different stages in the loop turn up in a variety of autoimmune disorders (Fig. 20.13).

For example, patients with rheumatoid arthritis have low circulating corticosteroid levels compared with controls. After surgery, although they produce copious amounts of IL-1 and IL-6, a defect in the hypothalamic paraventricular nucleus prevents the expected increase in adrenocorticotropin (ACTH) and adrenal steroid output.

There is currently intense interest focused on the role of Tregs. Patients with rheumatoid arthritis, for example, reveal a deficiency of Treg function (see below and Fig. 20.24).

A subset of CD4 regulatory cells present in young healthy mice of the NOD strain, which spontaneously develop IDDM, can prevent the transfer of disease provoked by injection of spleen cells from diabetic animals into NOD mice congenic for the severe combined immunodeficiency trait; this regulatory subset is lost in older mice. Prevention of disease in lupus-prone mice by targeting the FcγRIIb inhibitory receptor on B cells points to yet another potential regulatory defect predisposing to immune complex disease.

Pre-existing defects in the target organ may increase susceptibility to autoimmunity

We have already alluded to the undue sensitivity of target cells to upregulation of MHC class II molecules by IFNγ in animals susceptible to certain autoimmune diseases. Other evidence also favors the view that there may be a pre-existing defect in the target organ.

In the Obese strain chicken model of spontaneous thyroid autoimmunity, not only is there a low threshold of IFNγ induction of MHC class II expression by thymocytes, but when endogenous TSH is suppressed by thyroxine treatment, the uptake of iodine into the thyroid glands is far higher in the Obese strain than in a variety of normal strains. Furthermore, this is not due to any stimulating effect of the autoimmunity because immunosuppressed animals show even higher uptakes of iodine (Fig. 20.14).

Interestingly, the Cornell strain (from which the Obese strain was derived by breeding) shows even higher uptakes of iodine, yet these animals do not develop spontaneous thyroiditis. This could be indicative of a type of abnormal thyroid behavior, which in itself is insufficient to induce autoimmune disease, but does contribute to susceptibility in the Obese strain.

Other situations in which the production of autoantigen is affected are:

- diabetes mellitus, in which one of the genetic risk factors is associated with a microsatellite marker lying

Thyroid ^{131}I uptake in TSH-suppressed chickens

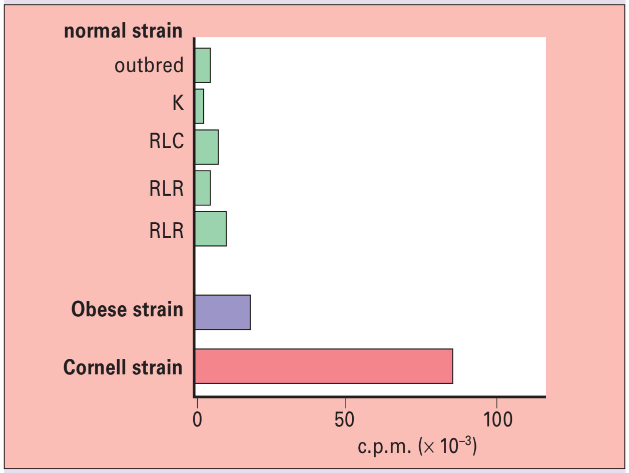

Fig. 20.14 Thyroid ^{131}I uptake in Obese strain chickens and in the related Cornell strain is abnormally high compared with that in normal strains. Endogenous TSH production was suppressed by administration of thyroxine; therefore the experiment measured TSH-independent ^{131}I uptake. Values were far higher than normal in Obese strain chickens, which spontaneously develop thyroid autoimmunity, and even higher in the non-autoimmune Cornell strain from which the Obese strain was bred. This abnormality was not due to immune mechanisms because immunosuppression actually increased ^{131}I uptake into the thyroid gland.

Upregulation of heat-shock protein 60 in endothelial cells at a site of hemodynamic stress

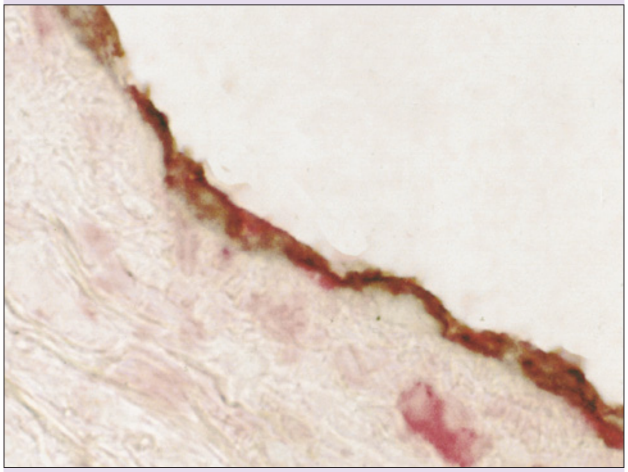

Fig. 20.15 Hsp60 expression (red) co-localized with intercellular adhesion molecule-1 (ICAM-1) expression (black) by endothelial cells and cells in the intima (macrophages) at the bifurcation of the carotid artery of a 5-month-old child. × 240. (Photograph kindly provided by Professor G Wick)

within a transcription factor controlling the rate of insulin production; and
- rheumatoid arthritis, in which the agalacto IgG glycoform is abnormally abundant.

The post-translational modification of arginine to citrulline, producing a new autoantigen in rheumatoid arthritis, represents yet another mechanism by which autoimmunity can be evoked.

Q. What evidence from rheumatoid arthritis supports the view that autoimmune diseases have a multifactorial etiology?

A. Defects in hypothalamic feedback, Tregs and post-translational modification of antigens (glycosylation and citrulline formation) as well as a strong MHC association have all been identified.

The following intriguing observations hint strongly at an autoimmune contribution to the pathology of the atherosclerosis:
- immunization with mycobacterial heat-shock protein 65 (hsp65) elicits atherosclerotic lesions at classical predilection sites subject to major hemodynamic stress; and
- patients with atherosclerosis produce antibodies to human hsp60, which react with heat or TNFα-stressed endothelial cells.

Particularly relevant to the present discussion is the finding of upregulated hsp60 expression at such critical sites even in a 5-month-old child (Fig. 20.15).

Again, one must re-emphasize the considerable importance of multiple factors in the establishment of prolonged autoimmunity.

IN MOST DISEASES ASSOCIATED WITH AUTOIMMUNITY, THE AUTOIMMUNE PROCESS PRODUCES THE LESIONS

Autoimmune processes are often pathogenic. When autoantibodies are found in association with a particular disease there are three possible inferences:
- the autoimmunity is responsible for producing the lesions of the disease;
- there is a disease process that, through the production of tissue damage, leads to the development of autoantibodies;
- there is a factor that produces both the lesions and the autoimmunity.

Autoantibodies secondary to a lesion (the second possibility) are sometimes found. For example, cardiac autoantibodies may develop after myocardial infarction.

However, sustained production of autoantibodies rarely follows the release of autoantigens by simple trauma. In most diseases associated with autoimmunity, the evidence supports the first possibility, that the autoimmune process produces the lesions.

The pathogenic role of autoimmunity can be demonstrated in experimental models

Examples of induced autoimmunity

The most direct test of whether autoimmunity is responsible for the lesions of disease is to induce autoimmunity deliberately in an experimental animal and see if this leads to the production of the lesions.

Autoimmunity can be induced in experimental animals by injecting autoantigen (self antigen) together with complete Freund's adjuvant, and this does indeed produce organ-specific disease in certain organs. For example:

- thyroglobulin injection can induce an inflammatory disease of the thyroid;
- MBP can cause encephalomyelitis.

In the case of thyroglobulin-injected animals, not only are thyroid autoantibodies produced, but the gland becomes infiltrated with mononuclear cells and the acinar architecture crumbles, closely resembling the histology of Hashimoto's thyroiditis.

THE ABILITY TO INDUCE EXPERIMENTAL AUTOIMMUNE DISEASE DEPENDS ON THE STRAIN OF ANIMAL USED – For example, it is found that the susceptibility of rats and mice to MBP-induced encephalomyelitis depends on a small number of gene loci, of which the most important are the MHC class II genes.

MBP-induced encephalomyelitis can be induced in susceptible strains by injecting T cells specific for MBP. These pathogenic T cells belong to the CD4/TH1 subset. It has been found that induction of disease can be prevented by treating the recipients with antibody to CD4 just before the expected time of disease onset, blocking the interaction of the TH cells' CD4 with the class II MHC molecules of antigen-presenting target cells (see Chapter 5).

The results indicate the importance of class II-restricted autoreactive TH cells in the development of these conditions, and emphasize the prominent role of the MHC.

Examples of spontaneous autoimmunity

It has proved possible to breed strains of animals that are genetically programmed to develop autoimmune diseases closely resembling their human counterparts.

One well-established example mentioned above is the Obese strain chicken, which parallels human autoimmune thyroid disease in terms of:

- the lesion in the gland;
- the production of antibodies to different components in the thyroid; and
- the overlap with gastric autoimmunity.

So it is of interest that, when the immunological status of Obese strain chickens is altered, quite dramatic effects are seen on the outcome of the disease. For example:

- removal of the thymus at birth appears to exacerbate the thyroiditis, suggesting that the thymus exerts a controlling effect on the disease through Tregs; but
- if the entire T cell population is abrogated by combining thymectomy with massive injections of anti-chick T cell serum, both autoantibody production and the attack on thyroid are completely inhibited.

T cells therefore play a variety of pivotal roles as mediators and regulators of this disease.

More directly, a diabetogenic CD4$^+$ T cell clone can induce the chronic leukocytic infiltrate of T cells and macrophages, which damages the insulin-producing β cells of the pancreatic islets of Langerhans in the NOD murine model of IDDM (Fig. 20.16).

The role of the T cells in mediating this attack is further emphasized by the amelioration and prevention of disease by treatment of the mice with a non-depleting anti-CD4 monoclonal antibody, which in the presence of the pancreatic autoantigens, insulin, and glutamic acid decarboxylase (GAD) induces specific T cell anergy.

The dependence of yet another spontaneous model, the F_1 hybrid of New Zealand Black and White strains, on the operation of immunological processes is aptly revealed by the suppression of the murine SLE, which characterizes this strain, by treatment with anti-CD4 (Fig. 20.17).

Human autoantibodies can be directly pathogenic

When investigating human autoimmunity directly rather than using animal models, it is of course more difficult to carry out experiments. Nevertheless, there is much evidence to suggest that autoantibodies may be important in pathogenesis, and we will discuss the major examples here.

Autoantibodies can give rise to a wide spectrum of clinical thyroid dysfunction

A number of diseases have been recognized in which autoantibodies to hormone receptors may actually mimic the function of the normal hormone concerned and produce disease. Graves' disease (thyrotoxicosis) was the first disorder in which such **anti-receptor antibodies** were clearly recognized.

The phenomenon of neonatal thyrotoxicosis provides us with a natural 'passive transfer' study, because the IgG antibodies from the thyrotoxic mother cross the placenta and react directly with the TSH receptor on the neonatal thyroid. Many babies born to thyrotoxic mothers and showing thyroid hyperactivity have been reported, but the problem spontaneously resolves as the antibodies derived from the mother are catabolized in the baby over several weeks.

Whereas autoantibodies to the TSH receptor may stimulate cell division and/or increase the production of thyroid hormones, others can bring about the opposite effect by inhibiting these functions, a phenomenon frequently observed in receptor responses to ligands that act as agonists or antagonists.

Different combinations of the various manifestations of thyroid autoimmune disease – chronic inflammatory cell destruction, and stimulation or inhibition of growth and thyroid hormone synthesis – can give rise to a wide spectrum of clinical thyroid dysfunction (Fig. 20.18).

A variety of other diseases are associated with autoantibodies

A parallel with neonatal hyperthyroidism has been observed – **antibodies to acetylcholine receptors** from mothers who have myasthenia gravis cross the placenta into the fetus and may cause transient muscle weakness in the newborn baby.

Somewhat rarely, **autoantibodies to insulin receptors and to α-adrenergic receptors** can be found, the latter associated with bronchial asthma.

Destruction of β cells in the pancreatic islets of Langerhans in the non-obese diabetic mouse by a diabetogenic T cell clone

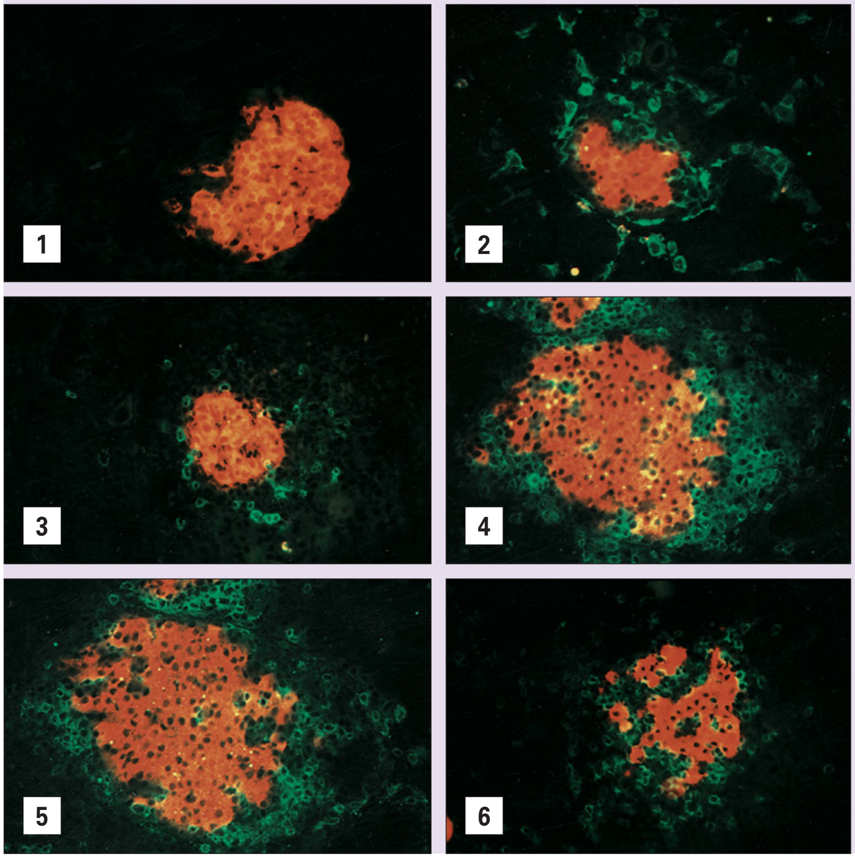

Fig. 20.16 NOD neonates were injected with the diabetogenic CD4$^+$ T cell clone, BDC2.5. The cells infiltrated the pancreas and there was recruitment of other NOD cells. Of the neonates that receive this clone 80% become diabetic within 20 days. Snap-frozen pancreas sections taken from representative neonates killed 10 days after transfer of BDC2.5, before the onset of overt diabetes, were stained for insulin (red fluorescence) and infiltrating cells (green fluorescence). (**1**) Section from a normal NOD neonate showing an islet, but no infiltration of CD3$^+$ T cells. Other sections show islets infiltrated with (**2**) macrophages (F4/80$^+$), (**3**) B cells (B220$^+$), (4) CD3$^+$ T cells, (5) CD4$^+$ T cells, and (**6**) CD8$^+$ T cells, this section showing serious loss of insulin producing cells. × 400. (Figures generously provided by Drs Jenny Phillips and Anne Cooke)

Neuromuscular defects can be elicited in mice injected with serum containing **antibodies to presynaptic calcium channels** from patients with the Lambert–Eaton syndrome, while **sodium channel autoantibodies** have been identified in the Guillain–Barré syndrome.

Yet another example of autoimmune disease is seen in rare cases of male infertility where **antibodies to spermatozoa** lead to clumping of spermatozoa, either by their heads or by their tails, in the semen.

In pernicious anemia an autoantibody interferes with the normal uptake of vitamin B_{12}

Vitamin B_{12} is not absorbed directly, but must first associate with a protein called intrinsic factor; the vitamin–protein complex is then transported across the intestinal mucosa.

Early passive transfer studies demonstrated that serum from a patient with pernicious anemia (PA), if fed to a healthy individual together with intrinsic factor–B_{12} complex, inhibited uptake of the vitamin.

Subsequently, the factor in the serum that blocked vitamin uptake was identified as **antibody against intrinsic factor**. It is now known that plasma cells in the gastric mucosa of patients with PA secrete this antibody into the lumen of the stomach (Fig. 20.19).

Antibodies to the glomerular capillary basement membrane cause Goodpasture's syndrome

In Goodpasture's syndrome, **antibodies to the glomerular capillary basement membrane** bind to the kidney in vivo (see Fig. 25.3). To demonstrate that the antibodies can have a pathological effect, a passive transfer experiment was performed. The antibodies were eluted from the kidney of a patient who had died from this disease, and injected into primates whose kidney antigens were sufficiently similar for the injected antibodies to localize on the glomerular basement membrane. The injected monkeys subsequently died from glomerulonephritis.

Blood and vascular disorders caused by autoantibodies include AHA and ITP

Autoimmune hemolytic anemia (AHA) and idiopathic thrombocytopenic purpura (ITP) result from the synthesis of **autoantibodies to red cells and platelets**, respectively.

The primary antiphospholipid syndrome characterized by recurrent thromboembolic phenomena and fetal loss is triggered by the reaction of **autoantibodies with a complex of β_2-glycoprotein 1 and cardiolipin**.

The β_2-glycoprotein turns up again as an abundant component of atherosclerotic plaques, and there is

Suppression of autoimmune disease

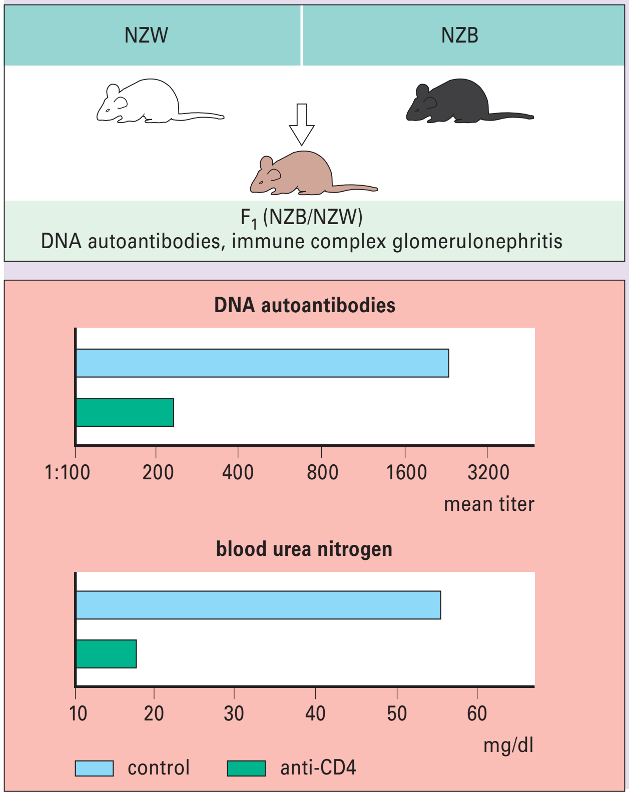

Fig. 20.17 The New Zealand Black (NZB) mouse spontaneously develops autoimmune hemolytic anaemia. The hybrid between this and the New Zealand White (NZW) strain develops DNA autoantibodies and immune complex glomerulonephritis, as in patients with SLE. Immunosuppression with monoclonal antibodies to the TH cell marker CD4 considerably reduced the severity of the glomerulonephritis and the titer of double-stranded DNA autoantibodies at 8 months of age, showing the relevance of the immune processes to the generation of the disease. (Based on data from Wofsy D, Seaman WE. J Exp Med 1985;161:378)

The spectrum of autoimmune thyroid disease

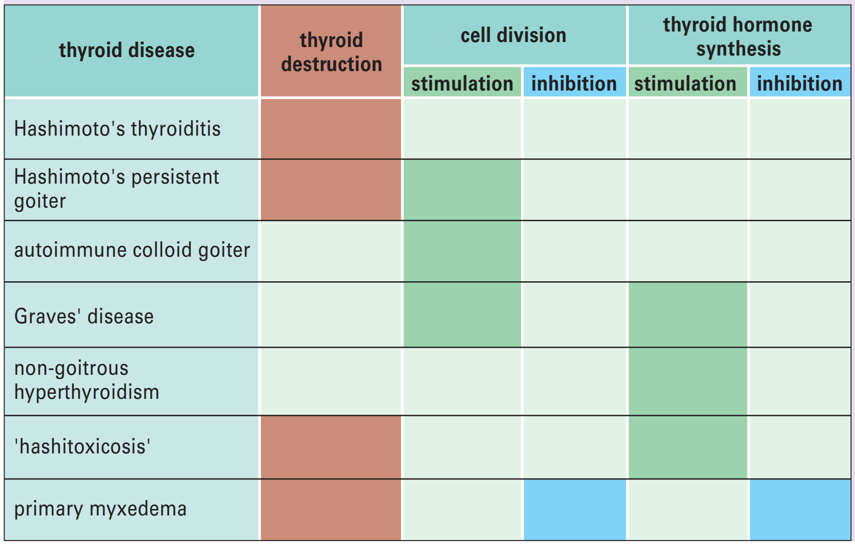

thyroid disease	thyroid destruction	cell division		thyroid hormone synthesis	
		stimulation	inhibition	stimulation	inhibition
Hashimoto's thyroiditis	🟥				
Hashimoto's persistent goiter	🟥	🟩			
autoimmune colloid goiter		🟩			
Graves' disease		🟩		🟩	
non-goitrous hyperthyroidism				🟩	
'hashitoxicosis'	🟥			🟩	
primary myxedema	🟥		🟦		🟦

Fig. 20.18 Responses involving thyroglobulin and the thyroid peroxidase (microsomal) surface microvillous antigen lead to tissue destruction, whereas autoantibodies to TSH (and other?) receptors can stimulate or block metabolic activity or thyroid cell division. 'Hashitoxicosis' is an unconventional term that describes a gland showing Hashimoto's thyroiditis and Graves' disease simultaneously.

increasing attention to the idea that autoimmunity may initiate or exacerbate the process of lipid deposition and plaque formation in this disease, the two lead candidate antigens being **heat-shock protein 60** (cf. Fig. 20.15) and the low density lipoprotein, **apoprotein B**.

The necrotizing granulomatous vasculitis that characterizes Wegener's granulomatosis is associated with **antibodies to neutrophil cytoplasmic proteinase III (cANCA)**, but their role in the pathogenesis of the vasculitis is ill defined.

Immune complexes appear to be pathogenic in systemic autoimmunity

In the case of SLE, it can be shown that complement-fixing complexes of antibody with DNA and other nucleosome components such as histones are deposited in the

Failure of vitamin B_{12} absorption in pernicious anemia

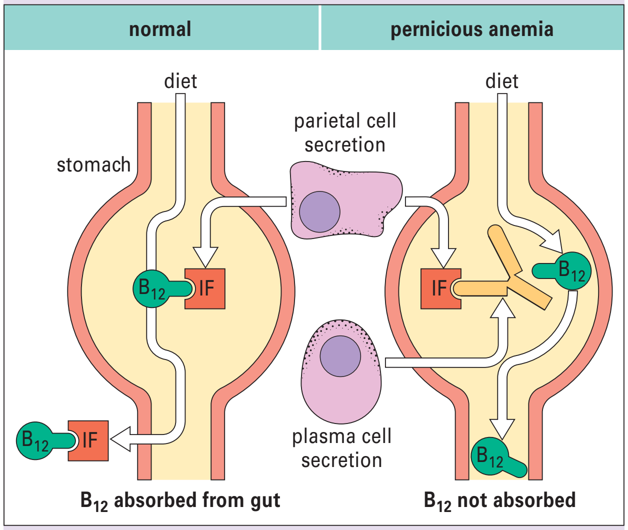

Fig. 20.19 Normally, dietary vitamin B_{12} is absorbed by the small intestine as a complex with intrinsic factor (IF), which is synthesized by parietal cells in gastric mucosa. In PA, locally synthesized autoantibodies, specific for intrinsic factor, combine with intrinsic factor to inhibit its role as a carrier for vitamin B_{12}.

kidney (see Fig. 25.3), skin, joints, and choroid plexus of patients, and must be presumed to produce type III hypersensitivity reactions as outlined in Chapters 24 and 25.

Cationic anti-DNA antibodies and histones facilitate the binding to heparin sulfate in the connective tissue structures.

Individuals with genetic deficiency of the early classical pathway complement components clear circulating immune complexes very poorly and are unduly susceptible to the development of SLE.

Turning to the experimental models, we have already mentioned the (NZB × NZW) F_1, which spontaneously develops murine SLE associated with immune complex glomerulonephritis and anti-DNA autoantibodies as major features (see Fig. 20.17). Measures that suppress the immune response in these animals (e.g. treatment with azathioprine or anti-CD4) also suppress the disease and prolong survival, adding to the evidence for autoimmune reactions causing such disease.

The erosions of cartilage and bone in rheumatoid arthritis are mediated by macrophages and fibroblasts, which become stimulated by cytokines from activated T cells and immune complexes generated by a vigorous immunological reaction within the synovial tissue (Fig. 20.20).

The complexes can arise through the self-association of IgG rheumatoid factors specific for the Fcγ domains – a process facilitated by the striking deficiency of terminal galactose on the biantennary N-linked Fc oligosaccharides (Fig. 20.21). This agalacto glycoform of IgG in complexes can exacerbate inflammatory reactions through reaction with mannose-binding lectin and production of TNF.

Evidence for directly pathogenic T cells in human autoimmune disease is hard to obtain

Adoptive transfer studies have shown that TH1 cells are responsible for directly initiating the lesions in experimental models of organ-specific autoimmunity.

In the human, factors that make it abundantly clear that T cells are utterly pivotal for the development of autoimmune disease include:

- the production of high-affinity, somatically mutated IgG autoantibodies characteristic of T-dependent responses;
- the isolation of thyroid-specific T cell clones from the glands of patients with Graves' disease;
- the beneficial effect of ciclosporin in prediabetic individuals; and
- the close associations with certain HLA haplotypes.

However, it is difficult to identify a role for the T cell as a pathogenic agent as distinct from a T helper function in the organ-specific disorders.

Indirect evidence from circumstances showing that antibodies themselves do not cause disease, such as in babies born to mothers with IDDM, may be indicative.

Autoantibodies have diagnostic and prognostic value

Whatever the relationship of autoantibodies to the disease process, they frequently provide valuable markers for diagnostic purposes. A particularly good example is the test for mitochondrial antibodies, used in diagnosing primary biliary cirrhosis (Fig. 20.22). Exploratory laparotomy was previously needed to obtain this diagnosis, and was often hazardous because of the age and condition of the patients concerned.

Autoantibodies often have predictive value. For instance, individuals testing positively for antibodies to both insulin and glutamic acid decarboxylase have a high risk of developing IDDM.

MANY AUTOIMMUNE DISEASES CAN BE TREATED SUCCESSFULLY

Often, in organ-specific autoimmune disorders, the symptoms can be corrected by metabolic control. For example:

- hypothyroidism can be controlled by administration of thyroxine;
- thyrotoxicosis can be controlled by administration of antithyroid drugs;
- in pernicious anemia, metabolic correction is achieved by injection of vitamin B_{12};
- in myasthenia gravis metabolic correction is achieved by administration of cholinesterase inhibitors.

If the target organ is not completely destroyed, it may be possible to protect the surviving cells by transfection with *FasL* or *TGFβ* genes.

Where function is completely lost and cannot be substituted by hormones, as may occur in lupus nephritis or chronic rheumatoid arthritis, tissue grafts or mechanical

Pathology of rheumatoid arthritis

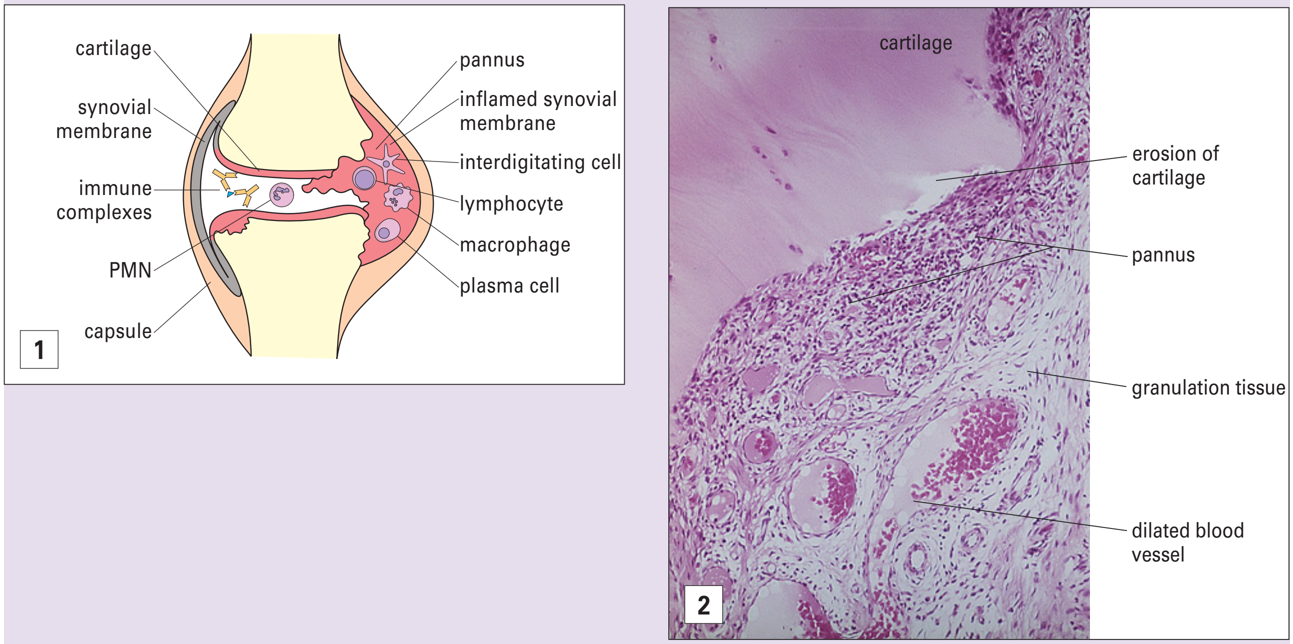
cartilage
synovial membrane
immune complexes
PMN
capsule
1
pannus
inflamed synovial membrane
interdigitating cell
lymphocyte
macrophage
plasma cell
cartilage
erosion of cartilage
pannus
granulation tissue
dilated blood vessel
2

Fig. 20.20 In the rheumatoid arthritis joint an inflammatory infiltrate is found in the synovial membrane, which hypertrophies forming a 'pannus' (**1** and **2**). This covers and eventually erodes the synovial cartilage and bone. Immune complexes and neutrophils (PMNs) are detectable in the joint space and in the extra-articular tissues where they may give rise to vasculitic lesions and subcutaneous nodules. (Histological section reproduced from Woolf N. Pathology: basic and systemic. London: W.B. Saunders; 1998)

Self-associated IgG rheumatoid factors forming immune complexes

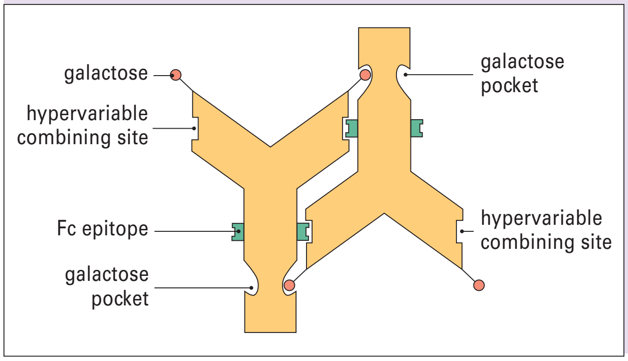
galactose
hypervariable combining site
Fc epitope
galactose pocket
galactose pocket
hypervariable combining site

Fig. 20.21 The binding between the Fab on one IgG rheumatoid factor and the Fc of another involves the hypervariable region of the combining site. As it has been established that the Fab oligosaccharides, which occur on approximately one in three different immunoglobulin molecules, are not defective with respect to glycosylation in rheumatoid arthritis, a Fab galactose residue could become inserted in the Fc pocket left vacant by a galactose-deficient Cγ2 oligosaccharide, so increasing the strength of intermolecular binding. The stability and inflammatory potency of these complexes is increased by binding IgM rheumatoid factor and Clq.

substitutes may be appropriate. In the case of tissue grafts, protection from the immunological processes that necessitated the transplant may be required.

Conventional immunosuppressive therapy with anti-mitotic drugs at high doses can be used to damp down the immune response, but, because of the dangers involved, tends to be used only in life-threatening disorders such as SLE and dermatomyositis.

The potential of ciclosporin and related drugs such as rapamycin has yet to be fully realized, but quite dramatic results have been reported in the treatment of type 1 diabetes mellitus.

Anti-inflammatory drugs are, of course, prescribed for rheumatoid diseases, with the introduction of selective cyclo-oxygenase-2 (COX-2) inhibitors representing a new development.

Diagnostic value of anti-mitochondrial antibodies

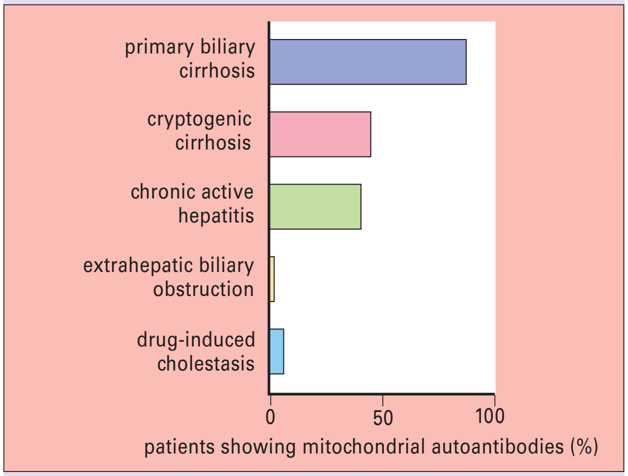

Fig. 20.22 Mitochondrial antibody tests using indirect immunofluorescence, together with percutaneous liver biopsy, can be used to assist in the differential diagnosis of these diseases. Many patients with primary biliary cirrhosis, but less than half of patients with cryptogenic cirrhosis or chronic active hepatitis, have anti-mitochondrial antibodies. The antibodies are rare in the other diseases.

Encouraging results are being obtained by the treatment of patients with rheumatoid arthritis with low corticosteroid doses at an early stage to correct the apparently defective production of these corticosteroids by the adrenal feedback loop. For those with more established disease, attention is now focused on the striking remissions achieved by synergistic treatment with anti-TNFα monoclonal antibody plus methotrexate (Fig. 20.23). It is fascinating to record that the compromised regulatory T cell function in the patients is reversed by such therapy (Fig. 20.24).

Less well-established approaches to treatment may become practicable

As we understand more about the precise defects, and learn how to manipulate the immunological status of the patient, some less well-established approaches may become practicable (Fig. 20.25):

- several centers are trying out autologous stem cell transplantation following hematoimmunoablation with cytotoxic drugs for patients with severe SLE, scleroderma, and rheumatoid arthritis;
- a draconian reduction in the T cells in multiple sclerosis by Campath-1H (anti-CD52) and of the B cell population with anti-CD20 in rheumatoid arthritis and SLE (Fig. 20.26) are giving particularly encouraging results;
- treatment with Campath-1H followed by a non-depleting anti-CD4 has produced excellent remissions in patients with Wegener’s granulomatosis who were refractory to normal treatment;
- in an attempt to establish antigen-specific suppression, considerable clinical improvement has been achieved in exacerbating–remitting multiple sclerosis by repeated injection of Cop 1, a random copolymer of alanine, glutamic acid, lysine, and tyrosine meant to simulate the postulated ‘guilty’ autoantigen, MBP.

Some experimental autoimmune diseases have been treated successfully by:

- feeding antigen to induce oral tolerance;
- the inhalation of autoantigenic peptides and their analogs; and

Synergy of anti-TNFα and methotrexate in the treatment of rheumatoid arthritis

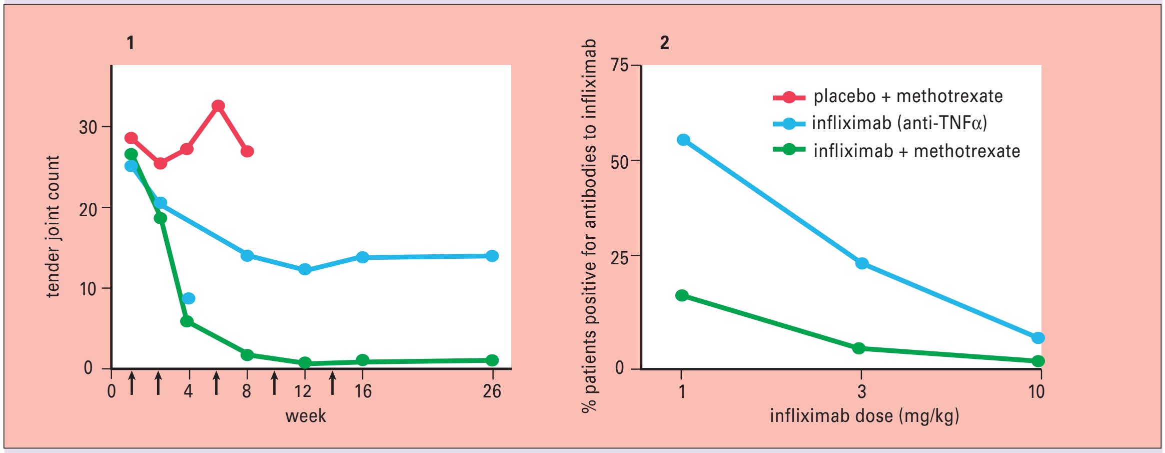

Fig. 20.23 (1) Infusions were given at times indicated by the arrows. Median joint scores were more effectively reduced by a combination of anti-TNFα with methotrexate, which (2) eliminated the anti-idiotypic response to infliximab (a humanized anti-TNFα monoclonal antibody). (Data reproduced from Maini RN, et al. Arthritis Rheum 1998;41:1552, with permission of the authors and publishers)

Reversal of compromised regulatory T cell function in rheumatoid arthritis by anti-TNFα therapy

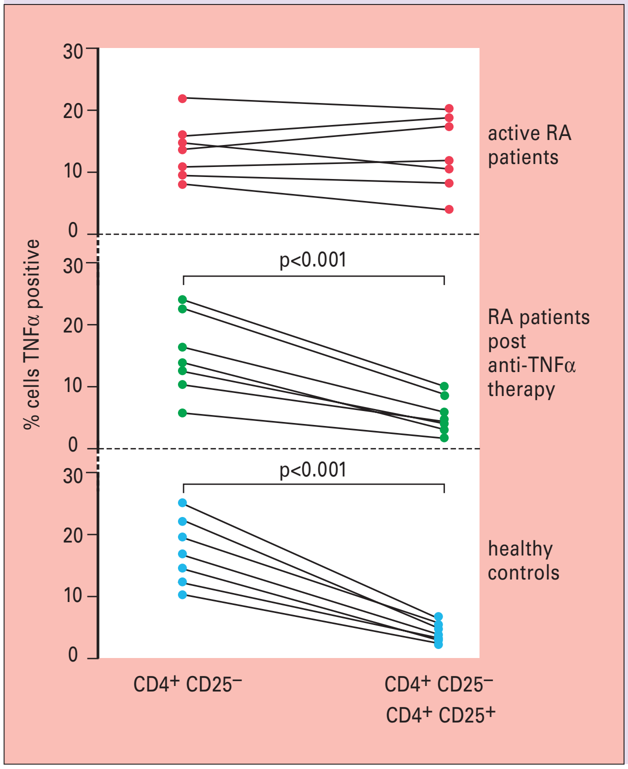

Fig. 20.24 CD4$^+$ CD25$^-$ cells isolated from peripheral blood were stimulated with a mixture of anti-CD3 and anti-CD28 alone or after adding back the CD4$^+$ CD25$^+$ putative regulatory T cells. Activation of the T cells was assessed by staining for intracellular TNFα. The lines join values for individual patients after adding back regulatory CD4$^+$ CD25$^+$ T cells. Although the regulatory T cells were dysfunctional in active rheumatoid arthritis (RA), they were able to suppress activation of the CD25$^-$ population after successful anti-TNFα therapy. (Data adapted from Ehrenstein MR, et al. J Exp Med 2004;200:277–285)

Current and potential treatment of autoimmune disease

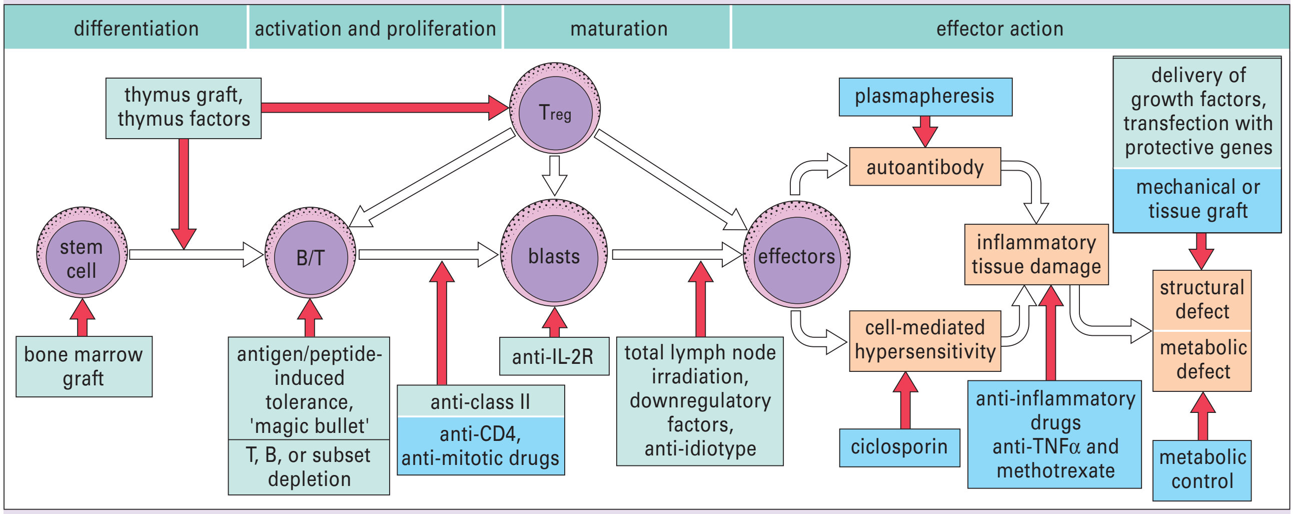
differentiation
activation and proliferation
maturation
effector action
thymus graft, thymus factors
Treg
plasmapheresis
delivery of growth factors, transfection with protective genes
autoantibody
mechanical or tissue graft
stem cell
B/T
blasts
effectors
inflammatory tissue damage
structural defect
metabolic defect
cell-mediated hypersensitivity
bone marrow graft
antigen/peptide-induced tolerance, 'magic bullet'
T, B, or subset depletion
anti-class II
anti-CD4, anti-mitotic drugs
anti-IL-2R
total lymph node irradiation, downregulatory factors, anti-idiotype
ciclosporin
anti-inflammatory drugs anti-TNFα and methotrexate
metabolic control

Fig. 20.25 Current treatments for arresting the pathological developments in autoimmune disease are given in blue boxes, and those that may become practicable in green boxes. Anti-mitotic drugs are given in severe cases of SLE or chronic active hepatitis, and anti-inflammatory drugs are widely prescribed in rheumatoid arthritis. Organ-specific disorders (e.g. primary myxedema) can be treated by supplying the defective component (e.g. thyroid hormone). When a live graft is necessary, immunosuppressive therapy can protect the tissue from damage.

Efficacy of B cell depletion in systemic lupus erythematosus

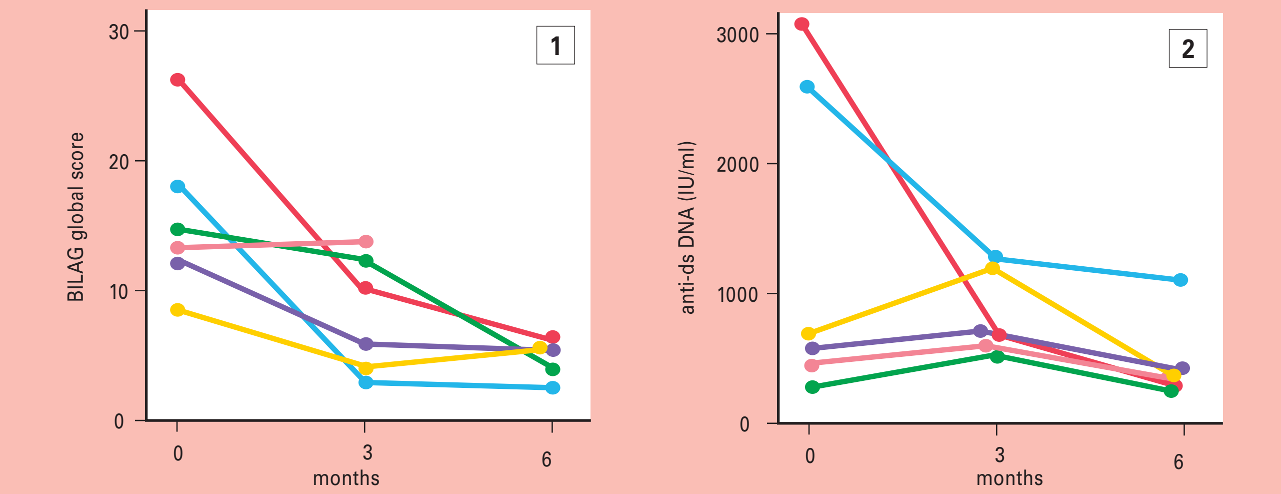

Fig. 20.26 Six female patients with active SLE resistant to standard immunosuppressive therapy were treated on an open-label basis with two infusions of monoclonal anti-CD20 to deplete the B cells, cyclophosphamide, and corticosteroids. Five of the six showed marked clinical improvement as evidenced by reduction of the British Isles Lupus Assessment Group (BILAG) global score (**1**) and significant falls in anti-dsDNA in two patients (**2**). (Data reproduced from Leandro MJ, et al. Arthritis Rheum 2002:46;2673. Copyright 2002 with permission from John Wiley & Sons Inc)

- 'vaccination' with peptides from heat-shock protein 70 or the antigen-specific receptor of autoreactive T cells.

This suggests that stimulating normally suppressive functions, including the idiotype network, could be promising.

FURTHER READING

Alt F, Marrack P, eds. Curr Opin Immunol 2003;15(6) and 2004;16 (several critical essays in each annual volume).

Arbuckle MR, McClain MT, Rubertone MV, et al. Development of autoantibodies before the clinical onset of systemic lupus erythematosus. N Engl J Med 2003;349:1526–1533.

Betterle C, Greggio NA, Volpato M. Clinical Review 93: Autoimmune polyglandular syndrome type 1. J Clin Endocrinol Metab 1998;83:1049–1055.

Chapel H, Haeney M, Misbah S, Snowden N. Essentials of clinical immunology, 4th edn. Oxford: Blackwell Science; 1999.

Edwards JC, Szczepanski L, Szechinski J, et al. Efficacy of the novel B cell targeted therapy, rituximab, in patients with active rheumatoid arthritis. N Engl J Med 2004;350:2572–2581.

Keymeulen B, Vandemeulebroucke E, Ziegler AG, et al. Insulin needs after CD3-antibody therapy in new-onset type 1 diabetes. N Engl J Med 2005;352:2598–2608.

McGaha TL, Sorrentino B, Ravetch JV. Restoration of tolerance in lupus targeted inhibiting receptor expression. Science 2005;307:590–593.

Peter JB, Shoenfeld Y, eds. Autoantibodies. Amsterdam: Elsevier; 1996.

Wraith DC, Nicolson KS, Whitely NT. Regulatory CD4^{+} T cells and the control of autoimmune disease. Curr Opin Immunol 2004;16:695–708.

Critical thinking: Autoimmunity and autoimmune disease (see p. 499 for explanations)

Miss Jacob, a 30-year-old Caribbean woman, was seen in a rheumatology clinic with stiff painful joints in her hands, which were worse first thing in the morning. Other symptoms included fatigue, a low-grade fever, a weight loss of 2 kg, and some mild chest pain. Miss Jacob had recently returned to the UK from a holiday in Jamaica and was also noted to be taking the combined oral contraceptive pill. Past medical history of note was a mild autoimmune hemolytic anemia 2 years previously.

On examination Miss Jacob had a non-specific maculopapular rash on her face and chest, and patchy alopecia (hair loss) over her scalp. Her mouth was tender and examination revealed an ulcer on the soft palate. She had moderately swollen and tender proximal interphalangeal joints. Her other joints were unaffected, but she had generalized muscle aches. The results of investigations are shown in the table.

investigation	result
radiograph of hands	soft tissue swelling, but no bone erosions
chest radiograph	a small pleural effusion at the right lung base
full blood count	a mild normocytic, normochromic anemia and mild lymphocytopenia
C-reactive protein levels	normal
erythrocyte sedimentation rate	raised
rheumatoid factor	negative
serum IgG levels	raised
anti-nuclear antibodies (ANA)	positive by immunofluorescence
anti-double-stranded DNA, anti-RNA, and anti-histone	positive by ELISA antibodies
complement (C3 and C4) levels	low
skin biopsy from an area unaffected by the rash	deposition of IgG and complement components at the junction between dermis and epidermis (lupus ‘band’ test)

A diagnosis of systemic lupus erythematosus (SLE) was made. Miss Jacob was treated with chloroquine, an anti-malarial, for the rash on her face and chest.

At a follow-up appointment urinalysis showed protein and red cells. Serum creatinine was mildly elevated as was her blood pressure. A renal biopsy showed membranous lupus nephritis. She was prescribed oral corticosteroids and an antihypertensive agent, which improved her renal function. Her physician also gave advice regarding birth control and pregnancy, and regular check-ups were arranged.

1. What is the immunological mechanism leading to the glomerulonephritis?
2. Are immune complexes the main mediator of systemic damage?
3. What is the mechanism for the vasculitis seen in SLE?
4. Are anti-double-stranded DNA (anti-dsDNA) antibodies pathognomonic of SLE?

CHAPTER

21

Transplantation and Rejection

SUMMARY

- **Transplantation is the only form of treatment** for most end-stage organ failure.
- **The barrier to transplantation** is the genetic disparity between donor and recipient.
- **The immune response in transplantation depends on a variety of factors.** Host versus graft responses cause transplant rejection. Histocompatibility antigens are the targets for rejection. Minor antigens can be targets of rejection even when donor and recipient MHC are identical. Graft versus host reactions result when donor lymphocytes attack the graft recipient.
- **Rejection results from a variety of different immune effector mechanisms.** Hyperacute rejection is immediate and caused by antibody. Acute rejection occurs days to weeks after transplantation. Chronic rejection is seen months or years after transplantation.
- **HLA matching is one of two major methods for preventing rejection of allografts.** The better the HLA matching of donor and recipient, the less the strength of rejection.
- **Successful organ transplantation depends on the use of immunosuppressive drugs.** 6-MP, azathioprine, and MPA are antiproliferative drugs. Ciclosporin, tacrolimus, and sirolimus are inhibitors of T cell activation. Corticosteroids are the anti-inflammatory drugs used for transplant immunosuppression. FTY 720 is a novel immunosuppressive agent under investigation.
- **The ultimate goal in transplantation is to induce donor-specific tolerance.** There is evidence for the induction of tolerance in humans and novel methods for inducing tolerance are being developed. Alloreactive cells can be made anergic. Immune privilege can be a property of the tissue or site of transplant.
- **Shortage of donor organs and chronic rejection limit the success of transplantation.** Living donation is one way to overcome the shortage of donor organs. Alternative approaches are being investigated. The favored animal for xenotransplantation is the pig.

TRANSPLANTATION IS THE ONLY FORM OF TREATMENT FOR MOST END-STAGE ORGAN FAILURE

Transplantation is a central topic for immunologists for two reasons:

- the first is that transplantation is an important clinical procedure;
- the second is that transplantation has proved an important tool for understanding immunological mechanisms – for example, the major histocompatibility complex (MHC; see Chapter 5) was first described in the context of transplantation, and transplantation models continue to be widely used as tools in basic as well as applied immunology.

As a clinical procedure, transplantation is used to replace tissues or organs that have failed. The first successful transplants were those of the cornea, first described in 1906.

Q. Why in retrospect was corneal transplantation more likely to be successful than transplantation of other tissues?
A. The cornea is an immunologically privileged site (see Figs 12.3–12.5).

World War II provided an important impetus, with the problems of skin grafting airmen who had extensive burns motivating a number of scientists, most notably Peter Medawar, to investigate the immunological basis of graft rejection.

The subsequent demonstration by the Medawar group that it was possible to manipulate a recipient animal so that it accepted grafts from an unrelated donor animal encouraged the subsequent clinical development of transplantation. The discovery (by Calne and others) of immunosuppressive drugs and agents then allowed the

Clinical transplantation

organ transplanted	examples of disease
cornea	keratoconus, dystrophies, keratitis
kidney	end-stage renal disease
heart	heart failure
lung/heart–lung	pulmonary hypertension, cystic fibrosis
liver	cancer, cirrhosis, biliary atresia
stem cells (bone marrow/ peripheral blood)	leukemia, immunodeficiency
skin (autografts)	burns
pancreas	diabetes mellitus
pancreatic islets	diabetes mellitus
small bowel	cancer, intestinal failure
neuronal cells	Parkinson's disease

Fig. 21.1 Organs and tissues shown in blue are routinely transplanted to treat various conditions. Transplantation of other organs (shown in yellow) is being developed, but has yet to be routinely applied in most centers.

widespread practice of transplantation in the last three to four decades of the 20th century.

In modern practice many transplants are performed routinely (Fig. 21.1). In addition to routine transplantation of the cornea, kidney, heart, lungs, and liver there is increasing interest in transplanting other organs, such as whole pancreas or islet cells for diabetes mellitus and also small bowel.

In general most transplants use organs from dead donors (**cadaveric transplants**), though there is an increasing number of living donors (usually related to the recipients) for kidney transplantation.

In addition to transplantation of organs, there is a large program of transplanting hematopoietic stem cells (cells capable of regenerating blood cells), for example in patients with leukemia or with primary immune deficiencies. These stem cells were previously normally harvested from the bone marrow of (living) donors, though increasingly peripheral stem cells obtained from the blood are used. As discussed below, stem cell transplantation has its own particular problems.

THE BARRIER TO TRANSPLANTATION IS THE GENETIC DISPARITY BETWEEN DONOR AND RECIPIENT

The main immunological problem with transplantation is that the grafted organ or tissue is seen by the immune system as 'foreign' and is recognized and attacked – leading to rejection of the organ.

Transplantation is normally performed between individuals of the same species who are not genetically identical, and the antigenic differences are known as **allogeneic differences**, and result in an **allospecific immune response** (Fig. 21.2).

However, it is also possible in experimental circumstances (and possibly in the future in the clinical setting) to perform grafting between different species. This is termed **xenotransplantation**, and the antigenic differences between donor and recipient form the **xenogeneic barrier**.

Transplantation can also be performed within an individual (e.g. skin grafting), when it is known as an **autograft**.

Syngeneic or **isografts** can be performed between genetically identical individuals. This can occur clinically for identical twins, but is more commonly seen in experimental settings with inbred strains of animals.

In the case of autografts and isografts there should be no antigenic differences between donor and recipient, and so no immune response. This can be readily illustrated using transplantation of skin or organs between inbred strains of animals (Fig. 21.3).

Genetic barriers to transplantation

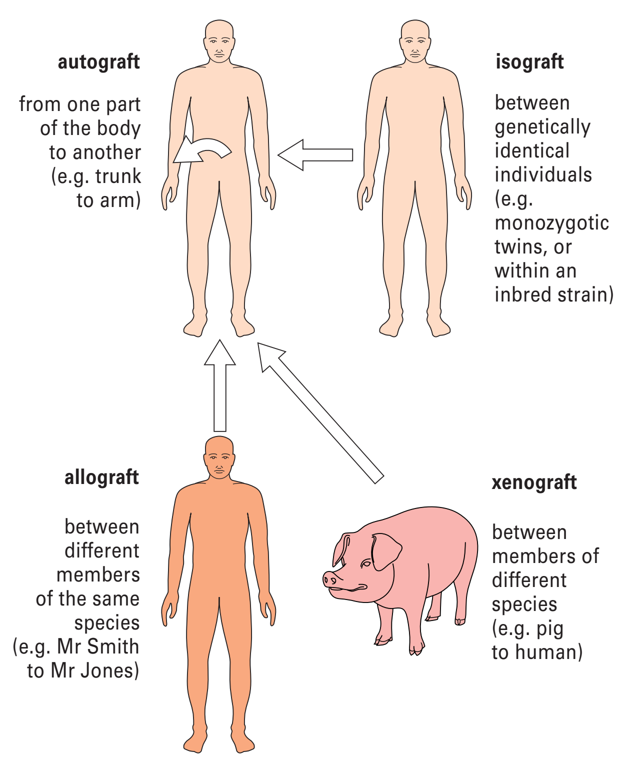
autograft
from one part of the body to another (e.g. trunk to arm)
isograft
between genetically identical individuals (e.g. monozygotic twins, or within an inbred strain)
allograft
between different members of the same species (e.g. Mr Smith to Mr Jones)
xenograft
between members of different species (e.g. pig to human)

Fig. 21.2 The genetic relationship between the donor and recipient determines whether or not rejection will occur. Autografts or isografts are usually accepted, whereas allografts and xenografts are not.

Host versus graft reactions

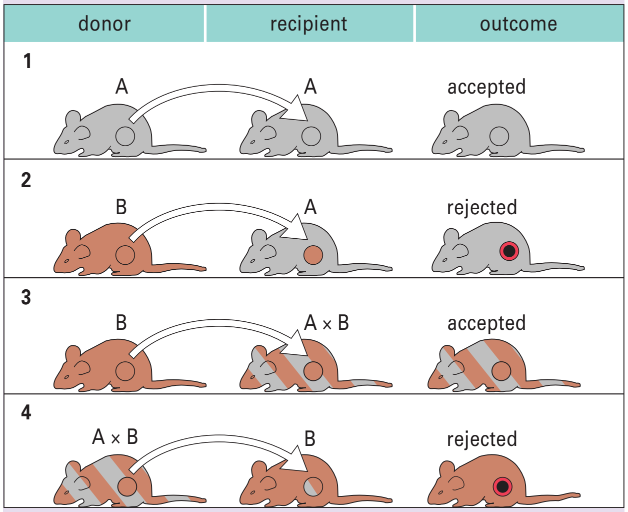

Fig. 21.3 Grafts between genetically identical animals are accepted. Grafts between genetically non-identical animals are rejected with a speed that is dependent on where the genetic differences lie. For example, syngeneic animals that are identical at the MHC locus accept grafts from each other (**1**). Animals that differ at the MHC locus reject grafts from each other (**2**). The ability to accept a graft is dependent on the recipient sharing all the donor's histocompatibility genes; this is illustrated by the difference between grafting from parental to (A × B) F_1 animals (**3**) and vice versa (**4**). Animals that differ at loci other than the MHC reject graft from each other, but much more slowly.

THE IMMUNE RESPONSE RESULTS IN GRAFT REJECTION

Host versus graft responses cause transplant rejection

Immune recognition of the antigenic differences between the donor organ and the recipient will, unless treated, lead to an immune response in which the host immune system responds to, and attacks, the donor tissue.

Q. What are the main genetic differences that are recognized by the host, and why should this be so?

A. Differences in MHC molecule allotypes are most important. The reason is that all nucleated cells express MHC molecules, and the T cell receptor on host T cells has a basic structure that interacts with and recognizes MHC molecules. Also the MHC class I and class II molecules have an extremely high level of genetic variability (see Chapter 5).

The nature of the host versus graft response is discussed in more detail below.

Similar to any other adaptive immune response, the immune response against a graft shows memory of previous encounters with an antigen. Therefore, once an animal has rejected an organ graft for the first time, if a second graft is performed from the same species or donor then it is rejected more rapidly (**second set rejection**).

There is a high frequency of T cells recognizing the graft

One of the main features of the immune response against a transplanted organ is that it is much more vigorous and strong than the response seen against a pathogen, such as a virus. This is largely reflected by the frequency of T cells that recognize the graft as foreign and react against it.

Thus, in a naive or unimmunized individual fewer than 1/100 000 T cells respond upon exposure to a virus or a protein immunization; however, 1/100–1/1000 T cells respond to allogeneic antigen-presenting cells (APCs).

This difference can be seen using a mixed lymphocyte reaction, in which T cells are taken from one individual and mixed with APCs from another. The proliferation of the T cells can then be measured (usually by measuring the incorporation of radioactive ^{3}H-thymidine) (Fig. 21.4). This will usually result in a strong proliferation of the T cells.

Measuring the strength of the alloresponse

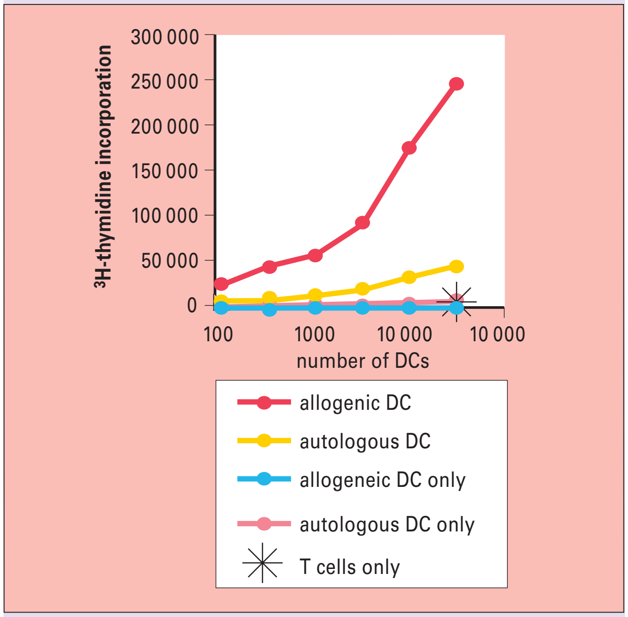

Fig. 21.4 The strength of the alloresponse can be measured in a mixed lymphocyte response. In this assay T cells from an individual were mixed with varying numbers of dendritic cells (DCs) (the most potent form of APCs) from either the same (autologous dendritic cells) or a different (allogeneic dendritic cells) donor. The dendritic cells were irradiated to prevent their proliferation. As a control cultures were included that contained just the dendritic cells or just the T cells. Five days later the cultures were pulsed with ^{3}H-thymidine, which is incorporated into the DNA of dividing cells. After 16 hours the amount of radioactive thymidine in the cells, a measure of the amount they have proliferated, is determined. As can be seen there is strong proliferation of T cells exposed to allogeneic dendritic cells, even though these T cells have not been exposed to the allogeneic cells, so this represents a primary immune response.

However, if the same experiment is performed using APCs and T cells from the same individual, and if a virus or a protein antigen is added, then no or little T cell proliferation is seen unless that person has been previously exposed to the antigen and so has a higher frequency of (memory) T cells capable of recognizing it.

Histocompatibility antigens are the targets for rejection

Early experiments showed that the bulk of the allospecific response is against molecules of the MHC. We now know that these molecules are the MHC class I or class II molecules (see Chapter 5), which are responsible for presenting antigen (in the form of peptides) to either:

- CD8 T cells (MHC class I); or
- CD4 cells (MHC class II).

As discussed in Chapter 7, MHC molecules are highly polymorphic, and it is these polymorphic differences that are seen by alloreactive T cells.

Why is there such a high frequency of allospecific T cells?

Several models seek to explain why there is such a high frequency of allospecific cells in the T cell repertoire.

The first model (**high determinant density model**) (Fig. 21.5) suggests that allospecific T cells recognize the foreign MHC molecules directly, with a low affinity, in a peptide-independent manner. The affinity of the interaction would normally be too low to activate the T cells; however, because the T cells see the MHC molecules directly, and are not recognizing the peptide, this low affinity is compensated for by the high concentration of MHC molecules on allogeneic cells.

The second model (**multiple determinant model**) states that what the allospecific T cells are recognizing are peptides derived from normal, non-polymorphic, host proteins that bind to and are presented by the foreign MHC molecules, but are not presented by self MHC. Due to lack of presentation by self MHC, the T cell repertoire is not tolerant to such peptides. The high frequency of the response is due to the large number of such antigens that can be presented by the graft.

Indirect recognition is important in chronic rejection

In a primary alloresponse, as detected by the mixed lymphocyte reaction, most of the alloreactive T cells recognize the MHC antigens by the direct pathway, as described above. In this interaction the T cell receptors on recipient T cells directly recognize the donor MHC molecules.

However, there are other forms of alloresponse, including:

- those against **minor MHC antigens** (see below); or
- the **indirect response** in which the recipient T cells recognize donor MHC molecules that have been processed by recipient donor APCs and are presented as peptides in the context of recipient MHC class II molecules (Fig. 21.6).

Why is there a high frequency of allospecific cells?

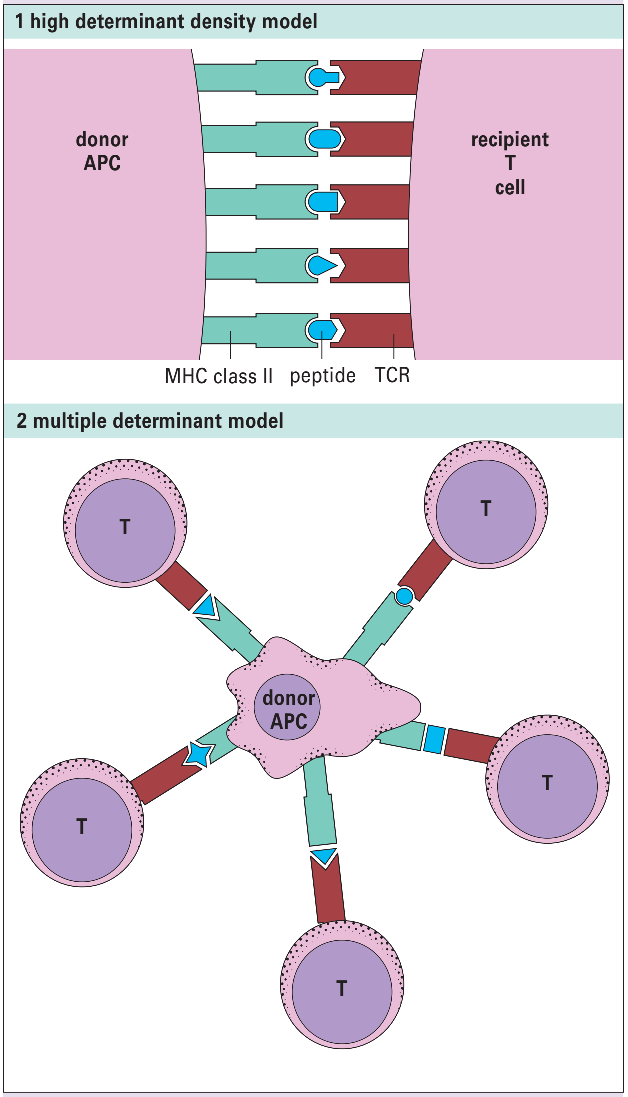

Fig. 21.5 The high precursor frequency of alloreactive T cells can be explained by one of two models. In the first (high determinant density model) the T cell receptor recognizes the MHC molecules on the surface of the APC with a moderate to low affinity. However, this recognition is largely independent of the type of peptide present in the groove of the MHC molecule. The low affinity of the interaction is compensated for by the multiple interactions that form, which provide an avidity advantage. In the multiple determinant model the T cells recognize, with a high affinity, peptide and MHC molecule complexes. The peptides are derived from normal self proteins. As the T cells have not been tolerized in the thymus to those antigens in the context of the donor MHC (though any T cells reactive with the same antigens in the context of self MHC would have been deleted), there is a high frequency of alloreactivity. Which form of allorecognition is most important will depend on the MHC combination of the donor and recipient.

Direct and indirect allorecognition

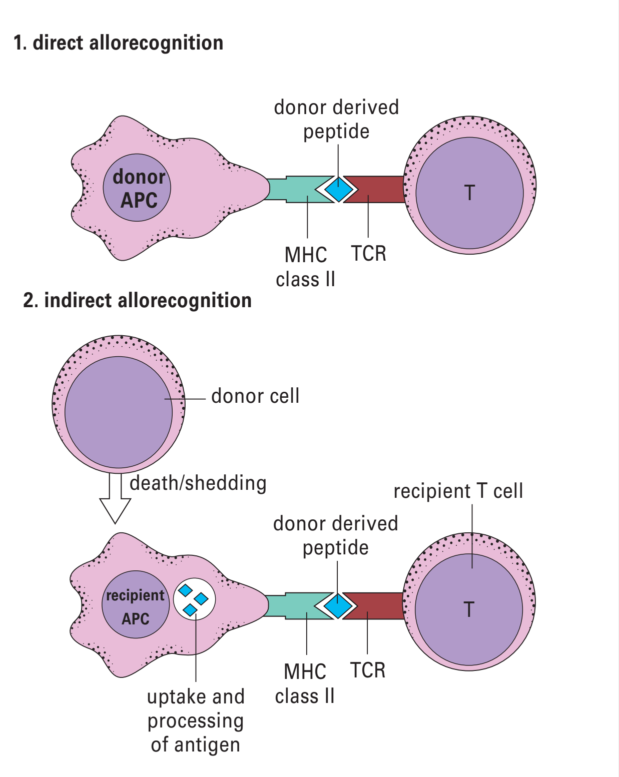

Fig. 21.6 (**1**) In direct allorecognition the T cell directly recognizes the donor APC, with the T cell receptor (TCR) binding donor MHC molecules bearing donor peptide. (**2**) Indirect allorecognition is more similar to conventional T cell recognition, in that the recipient T cells recognize foreign (donor derived) antigen that has been taken up and processed by recipient APCs. The TCR therefore recognizes recipient MHC bearing donor peptide. Although the frequency of the direct pathway allorecognition is high, that for the indirect pathway is no higher than that seen for normal antigens.

The indirect response is very similar to conventional T cell recognition of normal antigens, such as those from a pathogen, which are processed by host APCs and presented in the context of host MHC molecules.

Q. How frequent do you expect the T cells that recognize graft allogeneic molecules encoded outside the MHC to be?
A. The frequency of allospecific T cells in an unimmunized individual capable of recognizing donor tissue by the indirect pathway is low, equivalent to the frequency of reactivity to any normal antigen.

Nevertheless, the indirect pathway of recognition is important during chronic rejection, when the number of donor-derived professional APCs is no longer high enough to stimulate a direct immune response. It is also important in the rejection of corneal grafts because the cornea lacks large numbers of APCs.

Minor antigens can be targets of rejection even when donor and recipient MHC are identical

Although the MHC is the major target of the alloimmune response, there are also minor histocompatibility antigens. These can serve as targets of rejection even when the MHC is identical between donor and recipient.

The nature of most minor histocompatibility antigens is unknown, though they are assumed to be normal polymorphic molecules, peptides from which bind to host MHC and induce an immune response. In some cases they are expressed in a tissue-specific manner.

Perhaps the best studied minor histocompatibility antigen system is the H-Y system. These are antigens encoded by the Y chromosome, and so are expressed only on male cells. Thus, following immunization, it is possible to demonstrate immune responses and rejection of male organs or skin following transplantation mediated by female animals (2X chromosomes) against male cells (X and Y chromosome).

It is not possible to show responses against female antigens by male animals because the male animals carry one X chromosome in addition to their Y chromosome, and so are tolerant to all antigens encoded for by the X chromosome (Fig. 21.7).

Graft versus host reactions result when donor lymphocytes attack the graft recipient

Although it is usual to think of the immune response recognizing and destroying the transplanted organ, the situation is different when competent immune cells are transplanted into a recipient. This can happen during bone marrow transplantation, when normal donor T cells may be infused into the recipient. In such circumstances the T cells can recognize the MHC molecules and/or minor histocompatibility antigens of the recipient as foreign, and produce an immune response against the recipient. This is known as **graft versus host disease (GvHD)**.

GvHD can be lethal, causing damage in particular to the skin and gut. It can be demonstrated in animal models by transfer of bone marrow to irradiated recipient animals (Fig. 21.8). It can be avoided by:

- careful matching of the donor and recipient;
- removal of all T cells from the graft; and
- immunosuppression.

REJECTION RESULTS FROM A VARIETY OF DIFFERENT IMMUNE EFFECTOR MECHANISMS

The time of rejection is important

Rejection of organs or tissues can occur at various times, each of which is associated with different immune effector mechanisms (Fig. 21.9). These are:

- hyperacute rejection, which occurs within minutes to hours and is principally mediated by antibody;
- acute rejection, which usually occurs in days to weeks in animal models and is initiated by alloreactive T cells; and

H-Y minor histocompatibility antigens

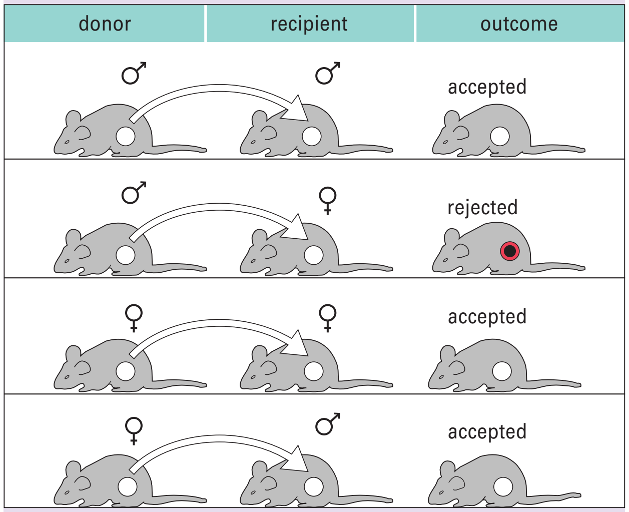

Fig. 21.7 The H-Y antigens are minor histocompatibility antigens expressed on male animals and not on females. Their existence can be demonstrated by skin grafting between animals of the same strain. As would be expected grafts from female mice are always accepted, while those from male mice are accepted on male recipients, but not female recipients.

- chronic rejection, which is seen months to years following transplantation.

Hyperacute rejection is immediate and mediated by antibody

Hyperacute rejection is seen when the recipient animal has pre-existing antibodies that are reactive with the donor tissue. This may be because:

- the individual has been sensitized to the donor MHC, for example by previous transplants, multiple blood transfusions, or pregnancy;
- the animal may have natural pre-existing antibodies (e.g. as a result of ABO blood group incompatibility).

A special case is seen in xenotransplantation, where humans and Old World monkeys and apes all have pre-existing antibodies to a carbohydrate antigen α-galactosyl. This carbohydrate is expressed on cell surface proteins of all other donors. Therefore, a xenotransplant of a cellular organ from a pig (or most other species) into a primate is at risk of hyperacute rejection.

Graft versus host disease

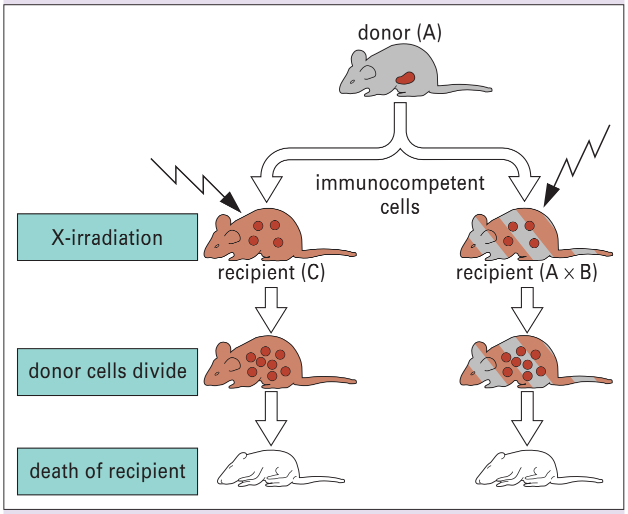

Fig. 21.8 Immunocompetent cells from a donor of type A are injected into an immunosuppressed (X-irradiated) host of type C, or a normal (A × B) F_1 recipient. The immunosuppressed individual is unable to reject the cells and the F_1 animal is fully tolerant to parental type A cells. In both cases the donor cells recognize the foreign tissue types B or C of the recipient. They divide and react against the recipient tissue cells and recruit large numbers of host cells to inflammatory sites. Very often the process leads to the death of the recipient.

Hyperacute rejection is seen within minutes of connecting the circulation into the transplanted organ. It is caused by the pre-existing antibodies binding to the endothelial cells lining the blood vessels and by initiating immune effector functions.

Q. Which immune effector functions would be activated following binding of antibody to the endothelium?

A. Complement will be activated, by the classical pathway (see Chapter 4).

Complement activation can lead to death of the endothelium, or, when the damage is sub-lethal, activation of the endothelial cells. This not only causes an inflammatory response, increasing vascular leakage, but can also cause blood coagulation (Fig. 21.10). The result is rapid destruction of the graft.

Tempo of rejection response

type of rejection	time taken	mechanisms of rejection
hyperacute rejection	minutes to hours	preformed anti-donor antibodies
acute rejection	days to weeks	activation of alloreactive T cells
chronic rejection	months to years	slow cellular response, response of organ to injury, unknown causes

Fig. 21.9 Rejection of organs or tissues can occur at various times, each of which is associated with different immune effector mechanisms, ranging from antibody mediated to acute and chronic cellular responses.

Renal histology showing hyperacute graft rejection

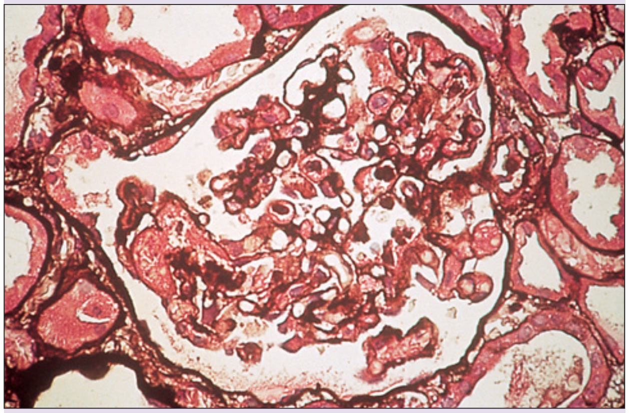

Fig. 21.10 There is extensive necrosis of the glomerular capillary associated with massive interstitial hemorrhage. This extensive necrosis is preceded by an intense polymorphonuclear infiltration, which occurs within the first hour of the graft's revascularization. The changes shown here occurred 24–48 hours after this. H&E stain. × 200.

Prevention of hyperacute rejection is performed by carefully avoiding transplanting an organ into an individual with pre-existing antibodies to that tissue. This is done by:

- ABO matching individuals; and
- cross-matching the donor and recipient.

This involves incubating donor leukocytes with recipient serum in the presence of complement; cell death indicates the presence of anti-donor antibody and is a contraindication to proceeding with transplantation. Such cross-matching is normally performed immediately before surgery.

In the case of xenotransplantation, one solution being developed is the generation of pigs that lack the α-galactosyl carbohydrate epitopes. This is done by knocking out the gene encoding the enzyme (α-galactosyl transferase) that generates this carbohydrate. In preclinical models in which pig organs are transplanted into primates, this approach has led to prolonged graft survival with no evidence of hyperacute rejection.

Acute rejection occurs days to weeks after transplantation

Acute rejection is normally seen days to weeks after transplantation, and is caused by activation of allospecific T cells capable of damaging the graft.

Donor dendritic cells (sometimes called passenger leukocytes) play an important role in triggering acute rejection. Dendritic cells that are present in the organ, following transplantation into the recipient, migrate to the lymph nodes draining the organ and stimulate a primary alloimmune response.

The importance of these dendritic cells can be shown by 'parking experiments' in which:

- a kidney is transplanted from one strain of rat to another under cover of immunosuppression (to prevent rejection) (Fig. 21.11);
- the kidney is kept in that animal long enough to ensure that all the resident dendritic cells have migrated out of the organ;
- the kidney is then transplanted into a third animal, of the same strain as the original recipient, where it shows prolonged graft survival.

However, if the third animal is injected with donor-derived dendritic cells there is rapid graft rejection. These data highlight the contribution of donor dendritic cells in initiating the alloresponse.

Although the direct pathway is thought to predominate in acute rejection, the indirect alloresponse, though significantly weaker, can also cause acute rejection in some animal models.

Once activated the T cells migrate to the organ and lead to tissue damage by standard immunological effector mechanisms (Fig. 21.12). These include:

- the generation of cytotoxic T cells; and
- the induction of delayed-type hypersensitivity reactions.

The role of the T cells in graft rejection can be demonstrated by depletion studies in which antibodies against T cell subsets are administered in vivo. Both of the major T cell subsets, $CD4^+$ and $CD8^+$, can cause graft rejection.

If the animal or patient has already been exposed to the alloantigens expressed by the graft, and as a consequence has been immunized, there will be alloreactive memory cells. This will lead to a much more rapid (accelerated) rejection of the graft (Fig. 21.13).

Chronic rejection is seen months or years after transplantation

In vascularized organs chronic rejection presents as occlusion of blood vessels, which on histological analysis show a thickening of the intima similar in some respects to the thickening seen as a result of atherosclerosis (Fig. 21.14). Smooth muscle cell proliferation is often seen, together with a macrophage infiltrate (together with some lymphocytes). This eventually leads to blockage of the blood vessels and subsequent ischemia of the organ.

A number of mechanisms can lead to chronic rejection. They include:

- a low-grade T cell response (mainly of the indirect allospecific pathway as a result of the loss of passenger leukocytes that activate T cells with direct pathway specificity, see p. 385);
- antibody can also be involved in chronic rejection, as indicated by the deposition of complement components (C4d) in tissues.

Non-immunological processes are also important, such as:

- the response of the graft to injury caused at the time of transplantation or by acute rejection episodes;
- recurrence of the original underlying disease; and
- drug-related toxicities (e.g. the immunosuppressive drug ciclosporin A is nephrotoxic and can damage the kidneys).

In some cases, initiation of chronic rejection may be immunological in nature, but its progression is due to non-immunological mechanisms.

Chronic rejection responds poorly to current immunosuppressive therapy. Therefore, although there has been

Importance of passenger leukocytes in graft sensitization

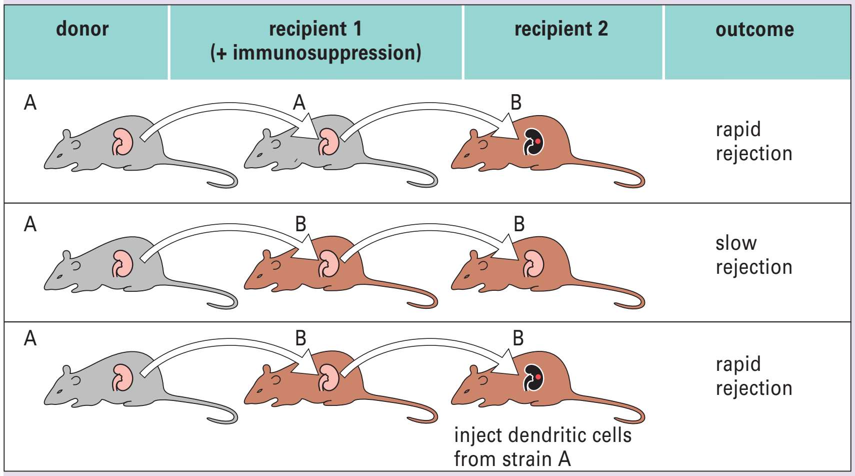

Fig. 21.11 The role of passenger leukocytes (dendritic cells) can be shown by 'parking' experiments in which kidneys are grafted from a rat of strain A into a recipient of strain B. Immunosuppression is used to prevent the animal from rejecting the graft. After a period the grafts are then retransplanted into a fresh strain B rat. There is very slow rejection of the graft (when compared to the rapid rejection seen when the kidney is transferred first into a strain A rat), which is thought to be due to the inability of the kidney from the strain A rat to immunize the strain B recipient due to the loss of dendritic cells during the period when the graft was 'parked' in the first strain A animal. The slow rejection probably occurs via the indirect pathway. The rejection occurs at the normal rapid tempo if strain A dendritic cells are injected into the recipient animal at the same time as the graft, suggesting that dendritic cells are capable of sensitizing the animal to the graft.

Renal histology showing acute graft rejection

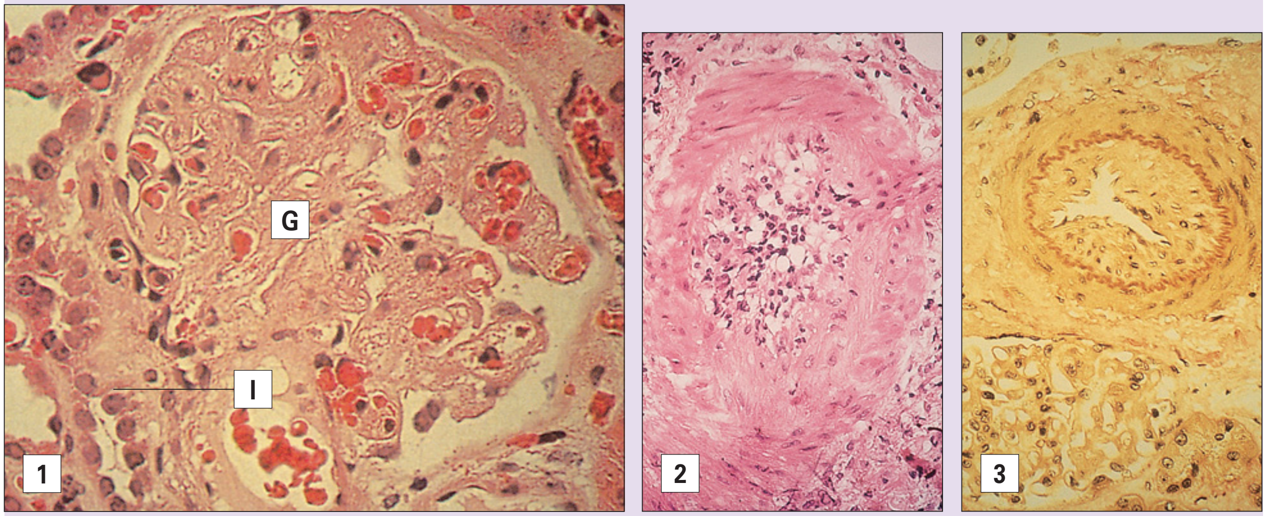

Fig. 21.12 (**1**) Small lymphocytes and other cells are accumulating in the interstitium of the graft. Such infiltration (I) is characteristic of acute rejection and occurs before the appearance of any clinical signs. H&E stain. (**2**) H&E stain of acutely rejecting kidney showing vascular obstruction. (**3**) van Gieson's stain of acutely rejecting kidney showing the end stage of this process. (G, glomerulus)

considerable improvement in overall graft survival over the past decades, this improvement is mostly seen in the first year following transplantation – the subsequent survival of grafts has hardly altered over the past 20 to 30 years (Fig. 21.15). This indicates the need to improve the treatment of chronic rejection.

HLA MATCHING IS IMPORTANT FOR PREVENTING REJECTION

The two major methods for preventing rejection of allografts are:

- to match the donor and recipient to minimize the antigenic differences;

Graft rejection displays immunological memory

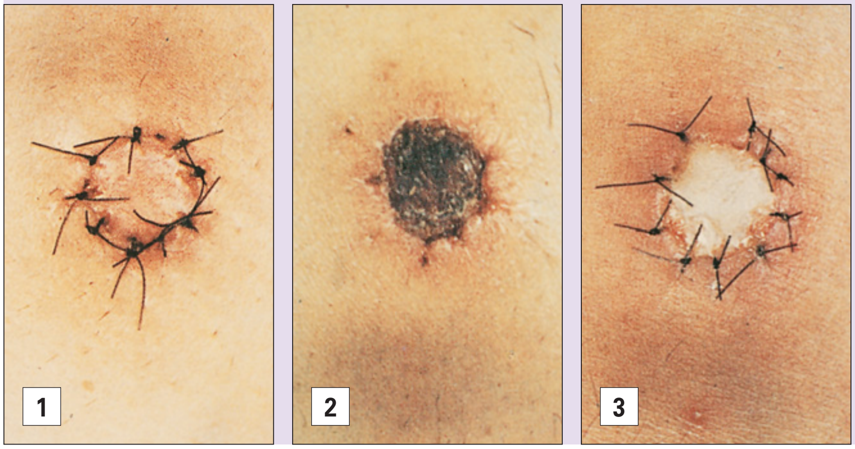

Fig. 21.13 A human skin allograft at day 5 (**1**) is fully vascularized and the cells are dividing, but by day 12 (**2**) is totally destroyed. A second graft ('second-set' graft) from the same donor shown here on day 7 (**3**) does not become vascularized and is destroyed rapidly. This indicates that sensitization to the first graft produces immunological memory.

Chronic rejection

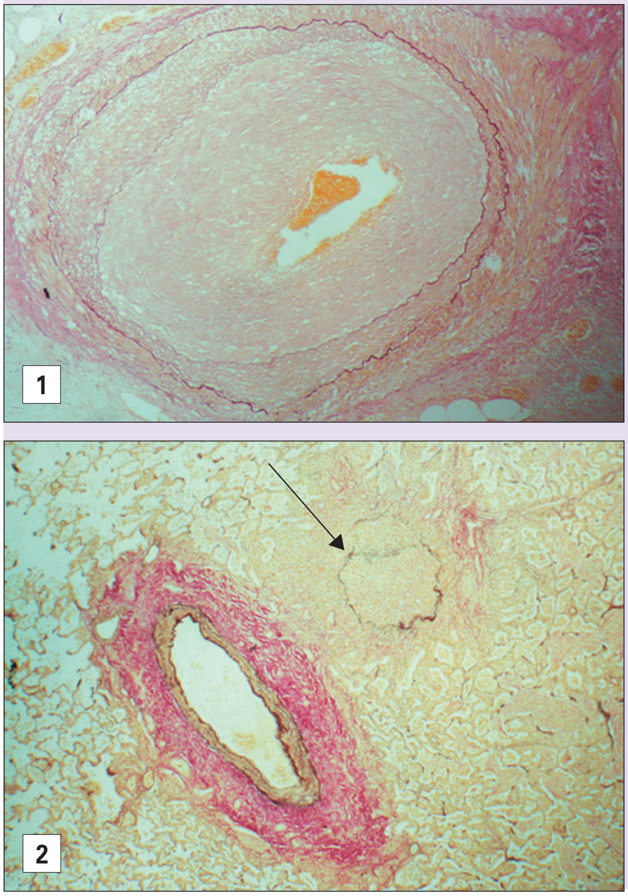

Fig. 21.14 Grafts that survive acute rejection are still capable of undergoing chronic rejection. (**1**) Section taken from a patient with chronic rejection of their heart graft. The lumen of the blood vessel in the heart has been narrowed as a result of thickening of the wall of the vessel, limiting the blood supply to the heart. (**2**) Section taken from a patient with chronic rejection of the lung, showing obliterative bronchiolitis (arrow) blocking the airways. (Kindly supplied by Professor Marlene Rose, Imperial College London, Harefield Hospital, and Dr Margaret Burke, Pathology Department, Royal Brompton Hospital and Harefield Hospital)

- to use immunosuppressive regimens that block the immune response against the organ.

However, in animal models (and hopefully in the clinic) there are techniques that can induce tolerance to an organ such that the immune system of the recipient 'learns' to treat the donor organ as 'self' and not destroy it. As discussed below, the ability to induce donor-specific tolerance is the Holy Grail of transplantation immunology (see p. 394).

The better the HLA matching of donor and recipient, the less the strength of rejection

The major antigenic differences recognized by the alloimmune response are found on the MHC molecules

Long-term survival of kidney grafts

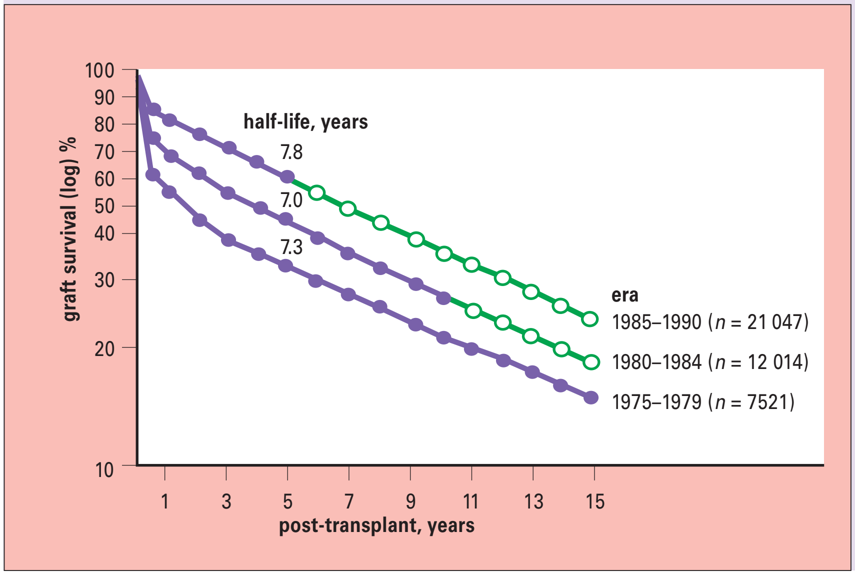

Fig. 21.15 This graph shows the survival of kidney grafts in patients transplanted in three periods: 1975–1979, 1980–1984, and 1985–1990 (with the projections for graft survival shown for later periods). There has been a considerable improvement in graft survival since 1975, but most of the improvement is seen in the first year. After that time the half-life of graft survival is similar for all patient groups. (With permission from UK Transplant, Bristol)

(known as HLA in humans, see Chapter 5). These highly polymorphic molecules have a vital role in presenting antigens to T cells.

There are many different alleles of the MHC molecules, and one way to reduce the strength of a rejection response is to match the donor and recipient so that they share as many alleles as possible. In general matching is now performed using molecular techniques, with polymerase chain reactions (PCR) that are specific for the different alleles.

In humans HLA matching is rarely perfect between unrelated donors because of the difficulty in matching all MHC class I and class II gene loci and the high level of polymorphism at each locus.

Q. How many MHC loci would need to be matched to obtain a perfect match in a human population?

A. There are three HLA class I loci (HLA-A, HLA-B, and HLA-C), and one individual can express up to six different HLA class I molecules (for each antigen there will be a maternally derived and a paternally derived version). Similarly, there are three class II loci (HLA-DR, HLA-DP, and HLA-DQ), and again the maternal and paternal chromosomes would normally be different. In other words there are 12 potentially different loci.

The extensive polymorphism means that there can be more than 600 possible versions (alleles) of each antigen (there are at least 626 different alleles for HLA-B, other antigens are less polymorphic, with 470 DRB and 349 HLA-A alleles, see Appendix 1).

Therefore, even before one considers other polymorphic molecules associated with the HLA locus, there is considerable complexity, which makes it highly unlikely for there to be a complete match.

In cases where the transplant is between living related donors (such as brothers and sisters) there is a greater opportunity for a match because in general the HLA locus is inherited en bloc as a single set (or haplotype) from each parent.

Loci outside the MHC can also lead to rejection (the minor histocompatibility antigens, see above). However, there is no attempt to match for these antigens because there is little possibility of getting a good match and the effect of matching any single minor antigen is too small to be significant.

It should be noted that even when siblings are perfectly matched at the MHC locus they will not be matched (unless they are identical twins) for the minor histocompatibility antigens.

The importance of HLA matching is not always crucial

HLA matching is very important in bone marrow transplantation (where a large potential donor pool can reduce the risk of GvHD) and has a significant influence on the outcome in kidney transplantation. For other organs the importance is less clear, for example:

- in corneal transplantation there is no benefit of HLA matching;
- for those organs where transplantation is essential to maintain life (such as heart and liver) there is no possibility of waiting for a well-matched organ to become available.

SUCCESSFUL ORGAN TRANSPLANTATION DEPENDS ON THE USE OF IMMUNOSUPPRESSIVE DRUGS

The success of organ transplantation is entirely dependent on the use of immunosuppressive drugs that control the alloimmune response. Although rejection episodes still occur, they are usually kept in check by the drugs so that lasting damage is minimized.

Over recent decades there has been a marked improvement in short-term success rates, such that over 90% of kidney transplants are functioning 1 year after transplantation. The major reason for these improved success rates is the advent of more powerful immunosuppressive agents.

Q. What problems would you expect to result from long-term immunosuppression?

A. These drugs cause blanket suppression of the immune system so transplant recipients are more prone to opportunistic infections and have a raised incidence of malignancy. This is the reason for continuing interest in strategies to promote specific immunological tolerance (see below).

Despite the continuing interest in strategies to promote specific immunological tolerance, clinical transplantation is likely to require non-specific immunosuppression for some years to come. The present challenge is to use the currently available agents intelligently to minimize side effects while preserving graft function.

The commonest cocktail of drugs used for kidney transplant patients involves three agents, each of which has a distinct mode of action:

- a drug that inhibits T cell activation;
- an antiproliferative; and
- an anti-inflammatory agent.

Usually three agents are used in the early post-transplant period while the anti-donor immune response is at its peak.

Numerous clinical trials are addressing the safety of withdrawing one of these three agents within weeks or months of transplantation. It appears that maintenance immunosuppression with two drugs is safe and has an improved side effect profile.

6-MP, azathioprine, and MPA are antiproliferative drugs

The first antiproliferative drug to be used in patients was **6-mercaptopurine (6-MP)**. This was followed by the introduction of **azathioprine**, the parent compound that is converted in vivo to 6-MP. Azathioprine was given to all patients until a more potent alternative arrived on the scene, **mycophenolic acid (MPA)**.

These antiproliferative drugs have related modes of action, namely inhibition of the synthesis of purines that are required for cell division:

- azathioprine is a purine antagonist and competes with inosine monophosphate;
- MPA inhibits an enzyme, inosine monophosphate dehydrogenase, which is essential in the de-novo synthesis of purines.

Clearly purine synthesis is needed in all cell types. Consequently, the risk of such agents is inhibition of other populations of rapidly dividing cell systems, such as the bone marrow. Indeed, close monitoring of white cell and platelet numbers is necessary in patients on antiproliferative drugs.

T cells are particularly sensitive to antiproliferative agents, partly because they are largely dependent on the de-novo synthesis of purines, whereas other cell types have a more efficient salvage pathway. Consequently, at conventional doses there is selective inhibition of T cell-dependent immunity.

Ciclosporin, tacrolimus, and sirolimus are inhibitors of T cell activation

Ciclosporin inhibits the production of IL-2

Drugs that inhibit T cell activation are the mainstay of immunosuppressive regimens. The first such drug to be discovered was **ciclosporin**, a fungal metabolite (Fig. 21.16).

The introduction of ciclosporin revolutionized clinical transplantation in that it led to a marked improvement in early success rates, which rose from approximately 70% to over 90%. Furthermore, the doses of other drugs that were required were substantially lower, so drug side effects were less troublesome.

Structure of immunosuppressive fungal macrolides

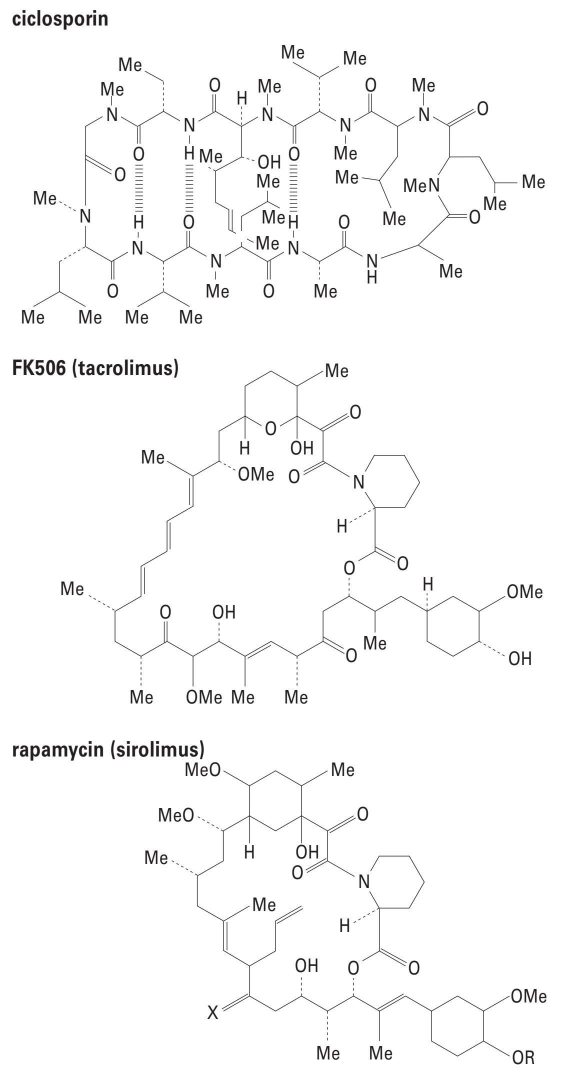

ciclosporin

FK506 (tacrolimus)

rapamycin (sirolimus)

Fig. 21.16 The immunosuppressive fungal macrolides, ciclosporin, FK506 (tacrolimus), and rapamycin (sirolimus), have quite different structures. They act on lymphocytes in different ways – ciclosporin and FK506 affect cytokine production and rapamycin interferes with signaling through the IL-2 receptor (IL-2R).

Because of the much greater potency of ciclosporin, and its ability to prevent early acute rejection, transplantation of organs other than kidney became routine. Programs of heart, liver, and more recently lung transplantation were established in major transplant centers around the world.

The key effect of ciclosporin is to inhibit the production of the major growth factor for T cells, interleukin-2 (IL-2).

The intracellular mechanism of action of ciclosporin involves binding to **cyclophilin**, and the consequent inhibition of the calcium-dependent phosphatase, **calcineurin**, which would otherwise activate the **NFAT complex**, and lead to IL-2 gene transcription.

Tacrolimus binds to an intracellular protein, FK-binding protein 12

A few years after the introduction of ciclosporin, a Japanese company discovered another fungal metabolite with a similar mode of action and greater molar potency, **tacrolimus** (**FK506**).

Tacrolimus binds to another intracellular protein, FK-binding protein 12 (FKBP12). The resulting complex then inhibits calcineurin, and the consequences of calcium-dependent signaling, in the same manner as ciclosporin.

Ciclosporin and tacrolimus are used almost interchangeably, and each has its advocates. Perhaps the major factor that is in favor of tacrolimus is that is less nephrotoxic than ciclosporin. On the other hand, tacrolimus is diabetogenic in some patients. As is usually the case, no drug is perfect, and the choice is one of balancing potency against toxicity.

Sirolimus inhibits signals transmitted by IL-2 binding to IL-2R

The third drug in the 'inhibitors of T cell activation' category is **sirolimus** (**rapamycin**).

Sirolimus is a fungal metabolite that was discovered on the Easter Island, Rapa ui, and its mechanism of action is quite distinct. Rather than inhibiting the production of IL-2, it inhibits the response to IL-2, by inhibiting some of the signals transmitted as a result of IL-2 binding to its receptor.

Indeed, sirolimus inhibits signaling through several growth factor receptors – consequently it has quite widespread antiproliferative effects, and was written off as being too non-specific when first studied in vitro.

Sirolimus has attracted much attention recently, for two reasons:

- first, because it appears to be a tolerance-permissive drug – that is, if T cells are exposed to antigen in tolerance promoting conditions, such as co-stimulatory blockade, sirolimus allows tolerance to occur, though in contrast there are some data to suggest that the calcineurin inhibitors may inhibit the development of tolerance under similar circumstances;
- second, the broader antiproliferative effects of sirolimus that almost confined it to the waste basket at the outset are now being claimed as beneficial because they reduce the incidence of some malignancies.

Corticosteroids are anti-inflammatory drugs used for transplant immunosuppression

Corticosteroids, given in combination with azathioprine, were the mainstay of immunosuppression for several decades. Only now is the possibility of rapid corticosteroid withdrawal, after 1 or 2 weeks, being explored.

Corticosteroids are pharmacological derivatives of the glucocorticoid family of steroid hormones, and act through intracellular receptors that are almost ubiquitously expressed.

The anti-inflammatory effects of steroids are highly complex, reflecting the fact that as many as 1% of genes may be regulated by glucocorticoids. Some of the most important effects are:

- inhibition of proinflammatory cytokine secretion (IL-1, IL-3, IL-4, IL-5, IL-8 (CXCL8), TNFα);
- inhibition of nitric oxide synthase;
- inhibition of adhesion molecule expression leading to reduced inflammatory cell migration;
- induction of endonucleases leading to apoptosis in lymphocytes and eosinophils.

The problem of side effects is a major issue with the use of corticosteroids, particularly at high doses. Central weight gain, fluid retention, diabetes mellitus, bone mineral loss, and thinning of the skin, all result from protracted corticosteroid use. It is for this reason that much attention is being devoted to protocols that allow steroid minimization.

FTY 720 is a novel immunosuppressive agent under investigation

Some newer immunosuppressives are under investigation and are the subject of clinical trials. One is FTY 720, which has a dramatic effect on lymphocyte migration, leading to marked sequestration of lymphocytes in lymphoid tissues, and a sharp fall in the circulating numbers of these cells. It remains to be seen whether this will find its way into the armamentarium of clinical immunosuppression.

In light of the continuing pursuit of donor-specific transplant tolerance, an interesting question is whether FTY 720 will favor or impede the development of tolerance.

THE ULTIMATE GOAL IN TRANSPLANTATION IS TO INDUCE DONOR-SPECIFIC TOLERANCE

Although generalized immunosuppression has been highly successful in preventing graft rejection, it comes at a price. This includes:

- the non-specific toxicity of the drugs;
- the need to stay indefinitely on medication; and
- the consequences of generalized immunosuppression such as the increased incidence of cancer and infection.

It would therefore be desirable to induce tolerance to the graft whereby the immune system specifically becomes non-responsive to the donor antigens, yet is still capable of responding normally to other antigens.

Tolerance to grafts was first demonstrated by Peter Medawar's group. They showed that, if allogeneic cells were injected into a neonatal animal, when the animal became adult it would be tolerant to tissue from the donor and would accept grafts without the need for immunosuppression. There have been numerous examples of inducing tolerance to grafts in animal models since, but this has yet to translate into the clinical setting.

One of the difficulties in the clinical setting is to demonstrate that tolerance really exists. In an animal model it is relatively easy to demonstrate tolerance by:

- performing a second graft (so showing that the immune system will no longer respond to the antigen); or
- demonstrating that an irrelevant third party graft is still rejected (demonstrating that graft survival is not due to a generalized immunosuppression).

However, this is more difficult in humans.

There is evidence for the induction of tolerance in humans

There are two sources of evidence for the induction of tolerance in humans.

First there are patients who have received grafts, but are no longer on immunosuppressive regimens because they cannot tolerate the drugs. They can show long-term graft survival. This is not formal evidence of tolerance, but it is highly suggestive that some people can have an operational tolerance whereby they fail to destroy their organ graft.

Second, it is possible to look at the frequency of alloreactive T cells in patients with grafts. In some groups there is a reduced frequency of these cells, but the response to other antigens remains normal (Fig. 21.17). Again, this is not a formal proof of tolerance because it is not yet known how the in-vitro assays relate to the response in patients. However, it does indicate that it might be possible to develop tests that will allow us to monitor the development of tolerance in patients, and so know how to tailor treatment to the individual (e.g. removing them from immunosuppression when indicated).

Novel methods for inducing tolerance are being developed

There are various ways in which tolerance (or the appearance of tolerance) can occur. Most of these are discussed in more detail in Chapters 11 and 19. An understanding of the mechanisms by which tolerance is induced and maintained allows the development of novel methods for inducing tolerance.

Central tolerance results from deletion of T cells in the thymus and is the most important form of tolerance induction for preventing autoimmunity. It has been harnessed for the induction of tolerance in experimental systems by transplanting the thymus from the donor into the recipient. This approach may be particularly useful in the context of xenotransplantation, where there is an opportunity to manipulate the donor and/or recipient before grafting of the organ. Limited clinical trials of inducing central tolerance using donor bone marrow in kidney transplant recipients are also under way.

Alloreactive cells can be made anergic

In the peripheral organs tolerance induction can result from deletion. However, it is also possible for alloreactive cells to be anergized. Anergy describes a state in which the cell is not deleted, but has been rendered unresponsive to further stimulation by the same antigen.

Q. What mechanism causes T cells to become anergic in the presence of an antigen presented on an appropriate MHC molecule?

A. Presentation of the antigen by a non-professional APC that lacks the ability to provide co-stimulatory signals (see Fig. 7.18).

Blockade of co-stimulatory molecules such as CD80 and CD86 with agents like CTLA-4–Ig (a fusion protein between CTLA-4, a ligand for CD80 and CD86, and the Fc part of an antibody molecule) can be used to induce anergy in alloreactive cells (Fig. 21.18). However, it

Reduction in allospecific T cells in patients with kidney grafts

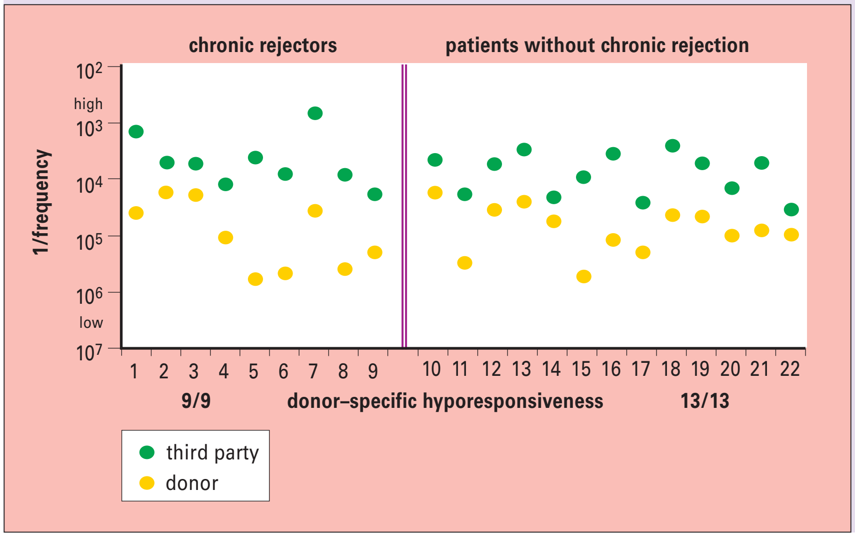

Fig. 21.17 The graph shows the frequency of T cells capable of producing IL-2 from patients with long-term kidney grafts that react with cells bearing the donor alloantigen or control (third party) alloantigens. The data show two groups of patients – those with evidence of chronic rejection and those without. Both groups of patient show a reduced frequency of T cells capable of recognizing alloantigen from the donor of the organ when compared to the frequency against third party alloantigen. This indicates that the patients show a degree of reduced reactivity to alloantigens.

Co-stimulation blockade

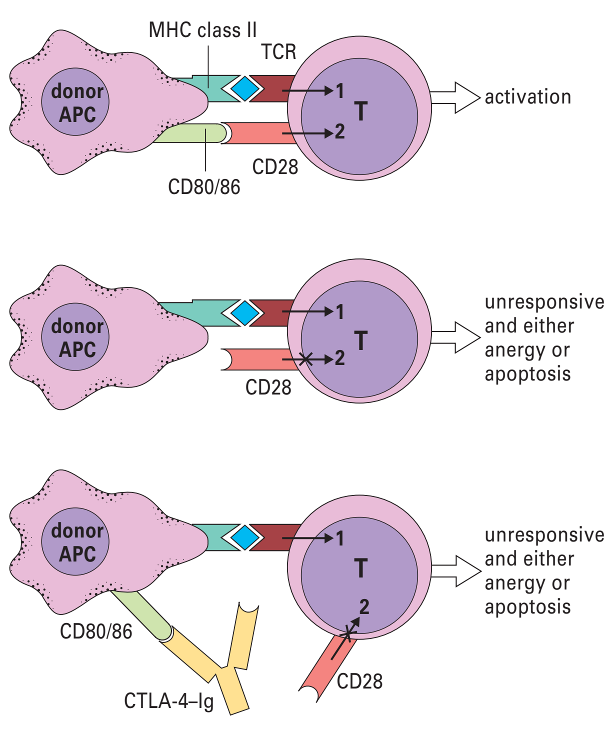

Fig. 21.18 For full T cell activation to occur the T cell needs to receive two signals – signal 1 through the T cell receptor binding the appropriate peptide–MHC complex (which provides antigenic specificity) and signal 2 through co-stimulatory molecules (the most important of which for naive T cells is CD28 on the T cell binding CD80, or CD86 on the APC). Interaction of a T cell with an APC that lacks CD80 or CD86, so that signal 1 only is received, fails to activate the T cell and can result in apoptosis or anergy of the lymphocyte. In experimental or clinical settings it is possible to block the CD28 interaction with CD80/86 by addition of a soluble molecule, CTLA-4–Ig, which consists of the extracellular part of CTLA-4 (an alternative ligand for CD80/86) fused to the Fc of immunoglobulin. This molecule prevents the interaction of CD28 with CD80/86. It can also result in upregulation of immunoregulatory enzyme, indoleamine 2,3-dioxygenase (IDO), as described in the text.

should be noted that the situation can be more complex than this. In many APCs cross-linking of CD80 and CD86 with CTLA-4–Ig results in upregulation of an immunomodulatory enzyme indoleamine 2,3-dioxygenase (IDO). This enzyme catabolizes tryptophan, and as a result prevents T cell activation both as a result of:

- depriving T cells of this essential amino acid; and
- the products of tryptophan breakdown acting directly on the T cells.

The role of IDO in immune regulation was first recognized in the placenta, where it protects the fetus from immunological rejection – inhibition of IDO causes rejection of histoincompatible fetuses.

Another alternative is to induce a regulatory response to the alloantigen. The phenomenon of T cell regulation has long been recognized (Fig. 21.19), and can be shown in experimental models by transferring T cells from a tolerant animal to a naive recipient, and showing that this results in a transfer of the tolerance. Several types of T cell are capable of regulating the immune response, and strategies that seek to expand these cells may be one method to induce tolerance in vivo.

Immune privilege can be a property of the tissue or site of transplant

Although a failure to reject a graft could be the result of tolerance (or immunosuppression), an alternative is that the graft is immune privileged and protected from the immune response against it:

- when corneal transplantation is performed in an animal model the rejection response is weaker than for other forms of transplantation, indicating the immune privileged nature of the graft;
- skin transplanted onto the anterior chamber (where the cornea is) shows a lower rate of rejection than skin transplanted into other sites.

These data indicate that the anterior chamber is an immune privileged site, and any tissue transplanted into it is protected from the immune response.

Other immune privileged sites include the posterior chamber of the eye, the testes, the brain, and the hamster cheek pouch (see Chapter 12).

It is important to note that immune privilege is not an absolute term. Corneal grafts are rejected, albeit less vigorously than other grafts. It is probably best to think of immune privilege as a spectrum ranging from grafts that show a high degree of privilege (cornea) to those where the immune response against them is very strong (skin), with other grafts somewhere inbetween.

Several mechanisms can be responsible for immune privilege. These include ‘ignorance’ – the immune system does not see the graft. The cornea is not vascularized, and has a poor lymphatic drainage. It also has a very low concentration of dendritic cells in the center. These features reduce the ability of the graft to stimulate an immune response against it. If this ignorance is disrupted (e.g. by inducing vascularization in the graft bed) then the cornea is more rapidly rejected.

In addition the tissue can deviate the immune response. In the anterior chamber of the eye there is a cytokine environment (high in TGFβ and α-melanocyte stimulating hormone) that deviates the immune response away from a tissue destructive to a non-inflammatory response.

Finally, the tissue being transplanted can itself be privileged. This is also seen in the cornea, which expresses high levels of Fas ligand.

Q. What is the significance of the expression of Fas ligand in the cornea?

A. It results in apoptosis of inflammatory cells expressing Fas that enter the graft (see Chapter 12).

Tolerance to grafts can be transferred by regulatory cells

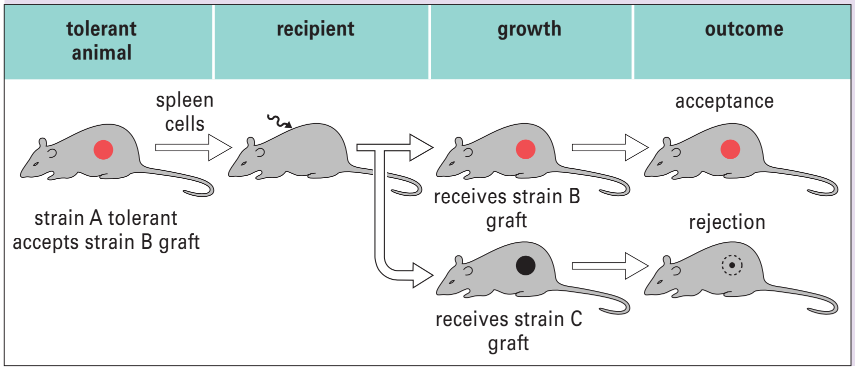

Fig. 21.19 An animal of strain A that has been made tolerant to a graft B can contain regulatory cells. If these cells are removed from the animal and transferred to a lightly irradiated strain A recipient, and this animal is then given a skin graft from strain B, then that graft may be accepted. In most cases this tolerance is specific, so if the second mouse received a graft of strain C it would be rejected.

SHORTAGE OF DONOR ORGANS AND CHRONIC REJECTION LIMIT THE SUCCESS OF TRANSPLANTATION

Two major issues limit success of transplantation:

- the first is the shortage of donor organs – this means that not everyone who would benefit from a transplant receives one, and, given the high success rate of transplantation, an increasing number of patients would benefit;
- the second problem is chronic rejection, which results in a continual loss of transplanted organs and necessitates patients remaining on immunosuppression, with the consequences of drug toxicity, systemic immunosuppression, increased incidence of malignancy, and infection.

The second of these problems, chronic rejection, would be solved if we were able to induce tolerance to grafts in patients. To do this it is necessary that:

- we develop assays that will allow us to determine when tolerance has been induced; and
- translate the therapies developed in animal models to the clinic.

These assays may involve measuring the frequency of alloreactive and/or regulatory T cells in patients. At present it is not clear which tolerance induction procedure is most likely to work in clinical transplantation, and further work in primate models is needed to address this issue.

Living donation is one way to overcome the shortage of donor organs

The shortage of donor organs can be addressed to some extent by increasing the number of potential donors, by increased advertising or changing the legislation. In addition more use could be made of living donations.

Living donation is possible for:

- kidney transplantation because the donor can survive with the one remaining kidney; and
- stem cell transplantation (either bone marrow or peripheral blood) because the donor can replace cells taken for donation.

Although some groups have performed other forms of living donation (e.g. with lung and liver lobes), this is not routine.

In most cases living donation is performed between relatives. When transplants from unrelated subjects are performed it is necessary to ensure that the donation is not coerced and that the necessary ethical standards are upheld.

Alternative approaches to overcoming the shortage of donor organs are being investigated

Although it is very important to increase the donor pool, the approaches discussed above will never provide all the organs needed. Alternatives include:

- the development of artificial mechanical organs;
- in the longer term, use of cloning and/or tissue engineering strategies to make artificial biological organs;
- xenotransplantation (the use of animal organs).

The favored animal for xenotransplantation is the pig

The reasons for the pig being the most suitable animal include the physiological and anatomical compatibility between pigs and humans and the ability to breed large numbers of animals rapidly.

There are several barriers to xenotransplantation, including public acceptability, safety, and scientific issues.

The main safety concern is the risk of transmission of viruses from the pig to the human. It should be noted that acellular xenografts have been performed for many years and are highly successful (e.g. in the case of pig heart valves, which can be used to replace diseased valves). These are not subject to immunological rejection, nor are they potential sources of viral infection.

The scientific issues revolve around preventing graft rejection. As indicated above, a pig organ transplanted into a human would undergo hyperacute rejection as a result of preformed anti-α-galactosyl antibodies in the recipient's circulation. One solution to this has been to engineer pigs (by nuclear cloning from cell lines) that lack the galactosyl transferase enzyme responsible for creating α-galactosyl residues. Organs from these pigs show reasonably long-term survival in primates.

As well as offering scientific challenges, there are also opportunities for using pig organs. The ability to genetically engineer the pig should make it easier to develop tolerance to the organ. This, as well as the ability to carry out transplants in a pre-planned manner, offers considerable advantages over conventional transplantation.

Critical thinking: Kidney transplantation (see p. 500 for explanations)

Mrs X has diabetes mellitus, and this caused severe damage to her kidneys. This complication is called diabetic nephropathy, and is one of the major indications for kidney transplantation. Mrs X was on dialysis treatment, but this was not working well for her and she was advised that she would benefit from renal transplantation. However, it proved very difficult to find a suitable cadaveric donor for Mrs X and it was suggested that a family member might donate an organ. All her immediate family – her husband, five children, and two brothers – agreed to be considered as donors.

The HLA types and blood groups of the family members are shown in the table. On the basis of these tests a donor was selected and the transplant was performed. Despite successful surgery, the kidney soon turned dark and swelled. This started to happen within a few minutes of the restoration of blood flow through the transplant, and necessitated the immediate removal of the graft.

Four years later Mrs X was still very ill on dialysis, no cadaveric donor was available, and it was decided to try again with a living related transplant. Another member of the family was selected to donate a kidney and it functioned well from the onset. Mrs X was given triple immunosuppression. She had only one rejection episode at about 3 weeks after grafting, and this was treated successfully with anti-rejection therapy. There were no other problems.

person	age	relationship to patient	HLA genotypes			blood group genes phenotype
			A	B	DR	
Mrs X	46	the patient	1 2	8 44	3 4	BODd BRh$^+$
Mr X	52	husband	2 3	14 7	8 2	AOdd ARh$^-$
Anne	25	daughter	2 3	44 7	4 2	AOdd ARh$^-$
Bert	24	son	1 2	8 14	3 8	ABDd ABRh$^+$
Chas	21	son	1 3	8 7	3 2	BOdd BRh$^-$
Dave	15	son	2 1	44 60	4 9	BODd BRh$^+$
Edna	13	daughter	1 2	8 14	3 8	AODd ARh$^+$
Fred	48	brother	1 2	8 44	3 4	ABDd ABRh$^+$
Gary	56	brother	2 2	44 14	4 15	BODD BRH$^+$

The kidney continued to work for 8 years, but its function gradually declined from the fourth year onwards. It seemed there was little the doctors could do to prevent this worsening situation, and Mrs X eventually had to return to dialysis.

1. What are the difficulties in finding a donor organ?
2. Comment on the HLA relationships between Mrs X and her brothers.
3. Comment on the relationships between the children of Mrs X.
4. Classify each member of the family in terms of their HLA relationship to Mrs X (HLA identical, HLA haplotype match, complete HLA mismatch).
5. In terms of HLA matching alone, who was the best donor for Mrs X?
6. Consider what effect the blood group antigens had on the choice of donor. From whom could kidneys have been transplanted, and who would not have been suitable?
7. Of those who had a compatible blood group, who would you have chosen as the best donor? Explain your reasoning.
8. The outcome of the transplantation was a disaster! By what mechanism was the graft attacked?
9. Why was Mrs X at a greater risk of this untoward reaction?
10. What laboratory tests are used to avoid this rejection reaction, and what seems to have gone wrong on this occasion?
11. Four years after the first transplant it was decided to try again with a living related donor. Of all the family members, who would you have chosen as the donor and whose kidney was most likely to survive in Mrs X?

Critical thinking: Kidney transplantation *(Continued)*

12 What is triple therapy immunosuppression?
13 What type of rejection occurred at 3 weeks after transplantation, and what immunological mechanisms were involved?
14 What is anti-rejection therapy?
15 There were no other problems with Mrs X. Can you think of some of the problems that might arise in a transplant recipient?
16 Why did the function of the transplant gradually decline, and why could the doctors not stop this process?

Discussion points

1 What are the ethical issues involved in this form of transplantation?
2 What novel forms of immunotherapy might be available in the future to prevent rejection of grafts?

This Page Intentionally Left Blank

CHAPTER

Immunity to Cancers 22

SUMMARY

- **Can the immune system protect the host from cancer?** The immune surveillance hypothesis has profound implications in cancer immunology.

- **Tumors can induce immunity.** Mice, rats, hamsters, and frogs can be immunized against tumors. In most tumors of animals, tumor immunity elicited by immunization is specific (or strongest) to the individual tumor that was used to immunize.

- **Tumor antigens have been characterized by three means – immunization-challenge experiments, T cell reactivity, and antibody reactivity**. Immunization-challenge experiments have uncovered the immunogenicity of heat-shock protein-chaperoned antigenic peptides. Antigens identified by reactivity to T cells and antibodies include individual tumor-specific mutated antigens, cancer testes antigens, differentiation antigens, and viral tumor antigens.

- **Vigorous anti-tumor immune responses are compromised by regulatory mechanisms**. Tumors elicit immunity in their primary host, and such immunity is downregulated. Regulatory cells such as CD25$^+$ CD4$^+$ T cells and inhibition of activated anti-tumor T cells through T cell molecules such as CTLA-4 are involved in downregulation of tumor immunity.

- **Three major categories of experimental immunotherapy are used for human cancer.** Vaccination with the idiotypes of B lymphomas, irradiated intact tumor cells, tumor lysate, or HSP–peptide complexes has been used. Vaccination with dendritic cells pulsed with peptides and adoptive immunotherapy using T cells is also being tested.

CAN THE IMMUNE SYSTEM PROTECT THE HOST FROM CANCER?

The immune surveillance hypothesis has profound implications for cancer immunology

Cancer appears to have engaged the minds of immunologists almost since the beginning of immunology itself.

Ehrlich, who opined on all things immunological, believed that the immune system could protect the host from cancer.

Burnet and Thomas refined that idea into the immune surveillance hypothesis of cancer. The hypothesis lived in limbo for several decades, failing to thrive and failing to die, until very recently when the work of Old, Schreiber, and their colleagues demonstrated that mice with a compromised immune status were more prone than immunocompetent mice to developing an array of cancers.

The immune surveillance hypothesis is often regarded as the intellectual underpinning of cancer immunology. Although the hypothesis itself has contributed little to our attempts to treat cancer through immunological means, it has profound implications for understanding the functions of the immune system.

Does infection have an anti-cancer effect?

This idea is an even more ancient stream of thought and effort in cancer immunity, and one that may be relevant to the immune surveillance hypothesis.

German physicians in the 19th century noted dramatic regressions of cancer in occasional patients who developed streptococcal infections. These anecdotes led to a systematic exploration of infections/fever as an immunotherapeutic modality by William Coley, a New York city surgeon.

The idea that a non-specific stimulation of the immune system, such as by an infection, can have an anti-cancer effect, has its origins in Coley's clinical studies. Never entirely applicable, and never to be ignored, the work of Coley and his predecessors continues to remain the permanent Rorscharch's inkblot in cancer immunology. It is only a small exaggeration to say that every idea in cancer immunity has been interpreted at some time or another, in terms of Coley's observations.

TUMORS ELICIT IMMUNITY IN THE PRIMARY HOST

Entirely unrelated to these two lines of enquiry, the study of cancer immunity saw a revival at the hands of those who were transplanting chemically induced tumors into the many inbred mice that began to be available in the 1950s. These investigators 'showed "highly successful" immunization against the "transplantable tumors" and expressed great hopes about cancer vaccination'.

In hindsight, these successful immunizations were simply a result of allogeneic differences between tumors and the host strain of mouse, a theme that played a seminal role in definition of the MHC, but had no relevance for cancer immunity.

Q. What will happen when a tumor from a mouse of one MHC haplotype is transplanted into a recipient with a different haplotype on the first occasion or the second?

A. A transplantation rejection reaction will ensue, and these reactions display specificity and memory (see Chapter 21).

Nonetheless, amidst the barrage of experiments where MHC-mismatched tumors were transplanted into mice, were the experiments of Ludwik Gross, and later those of Prehn and Main, and of George and Eva Klein, who showed that, even when MHC-matched tumors were used to immunize mice, protection against subsequent tumor growth could be achieved (Fig. 22.1). These studies led to two principles, which have informed much of cancer immunology since and are discussed below.

Cancers elicit protective immunity in the primary and syngeneic host

Mice and rats of a given haplotype can be immunized with irradiated cancer cells that arose in animals of the same haplotype. When they are challenged with live cancer cells, they are able to resist the tumor challenge. The following further observations and deductions have been derived from these results.

- The immunogenicity of tumors has provided the foundation stone for the idea of **tumor-specific antigens**. If one could immunize, then antigens must exist.
- Tumor immunity depends on many factors. The degree of tumor immunity depends upon the type of cancer and the method of its induction, or lack of induction. UV-induced cancers are highly immunogenic, methylcholanthrene-induced tumors less so, and spontaneous tumors even less so. Nonetheless, immunogenicity of tumors has been demonstrated in all model systems tested.

Immunogenicity of chemically induced tumors

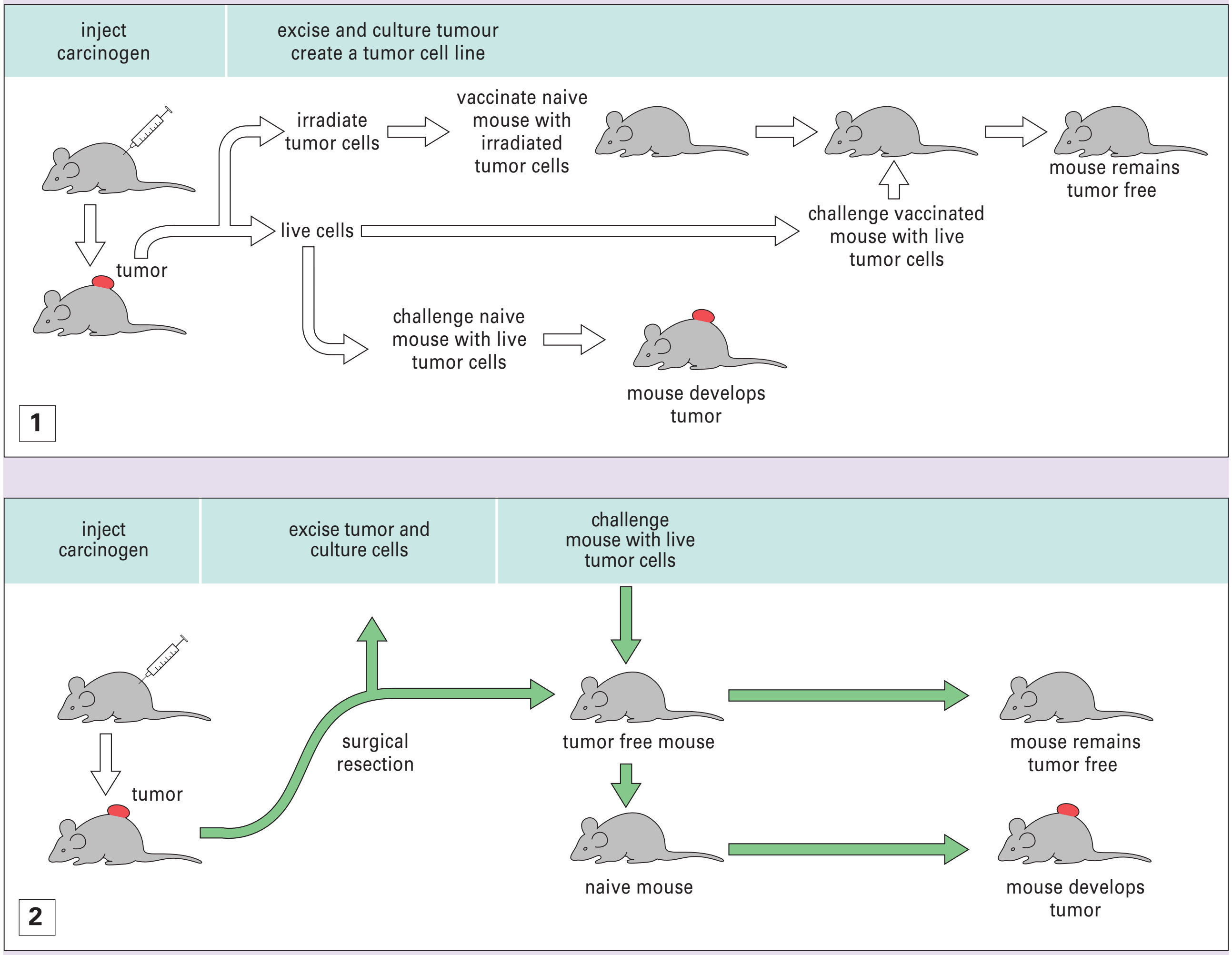
inject carcinogen
excise and culture tumour create a tumor cell line
tumor
irradiate tumor cells
vaccinate naive mouse with irradiated tumor cells
live cells
challenge vaccinated mouse with live tumor cells
mouse remains tumor free
challenge naive mouse with live tumor cells
mouse develops tumor
1
inject carcinogen
excise tumor and culture cells
challenge mouse with live tumor cells
tumor
surgical resection
tumor free mouse
naive mouse
mouse remains tumor free
mouse develops tumor
2

Fig. 22.1 (**1**) Mice immunized with tumor cells are protected against subsequent tumor growth. (**2**) The primary animal that develops the tumor is also immune to subsequent challenges with the same tumor.

- The primary animal develops immunity to subsequent tumor challenge. The Kleins made the seminal observation that not only can the syngeneic animals be immunized, but that the primary animal that develops the tumor is also immune to subsequent challenges with the same tumor (see Fig. 22.1).
- Protective immunity is only seen in prophylactic immunization and not therapeutic immunization – once a mouse has been implanted with a tumor, immunization with irradiated cancer cells (derived from the growing tumor) does nothing to mitigate tumor growth (Fig. 22.2). Exploration of differences between prophylaxis and therapy has led to fundamental insights into tumor immunity (see p. 413).
- It is not possible to test immunogenicity of tumors in humans. The transplantation-challenge experimental paradigm is not applicable to humans and no clearer method of determination of immunogenicity exists. It is therefore impossible to comment on the immunogenicity of human tumors, though references to such immunogenicity are often made. Much of the work in human cancer immunity has been done with melanomas leading to suggestions that, among human tumors, melanomas are particularly immunogenic. This is an erroneous belief – melanomas are simply the easiest human tumors to culture in vitro and hence the easiest to study.

Immunity to a fibrosarcoma is specific to the individual tumor

When mice are immunized against a given fibrosarcoma, they are rendered immune to that fibrosarcoma, but only to that individual fibrosarcoma or lines derived from it. If the mice are challenged with another fibrosarcoma, even one induced by the same carcinogen, tumor growth is unaffected by the prior immunization (Fig. 22.3).

Q. What can you infer about the nature of the immune response to tumors from this observation?

A. It must involve the adaptive immune system.

Difference between prophylaxis against future tumors and therapy of pre-existing tumors

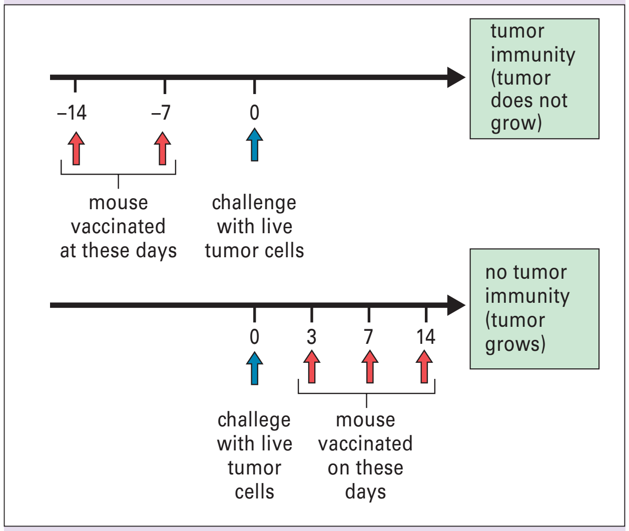
–14
–7
0
mouse vaccinated at these days
challenge with live tumor cells
tumor immunity (tumor does not grow)
0
3
7
14
challege with live tumor cells
mouse vaccinated on these days
no tumor immunity (tumor grows)

Fig. 22.2 Protective immunity is only seen in prophylactic immunization and not therapeutic immunization – once a mouse has been implanted with a tumor, immunization with irradiated cancer cells (derived from the growing tumor) does nothing to mitigate tumor growth.

Chemically induced tumors are individually distinct. The most extensive interrogation of the individually distinct antigenicity of chemically induced tumors was carried out by Basombrio and Prehn. They induced fibrosarcomas in 25 syngeneic BALB/c mice using methylcholanthrene and tested the immunogenicity of each one against all other tumors in an immunization-challenge model. They concluded that, even though all tumors were induced at the same time, in the same strain of mice of the same age by the same carcinogen and were histologically all fibrosarcomas, they were antigenically individually distinct. Independent tumors induced by the same carcinogen in the same mouse are also antigenically distinct.

Individually distinct antigenicity of chemically induced tumors

tumor used to challenge	tumor used to immunize			
	A	B	C	D
A	+	–	–	–
B	–	+	–	–
C	–	–	+	–
D	–	–	–	+

+ denotes tumor immunity or lack of tumor growth; – denotes lack of tumor immunity or tumor growth

Fig. 22.3 When mice are immunized against a given fibrosarcoma they are rendered immune to that fibrosarcoma, but only to that individual fibrosarcoma or lines derived from it. If the mice are challenged with another fibrosarcoma, even one induced by the same carcinogen, they are sensitive to it.

Spontaneous tumors are also antigenically distinct.

Cross-reactivity among tumors does occur occasionally. Typically, such cross-reactive immunity has been observed to be significantly weaker than the individually specific antigenicity. Efforts at characterization of such cross-reactive tumor-protective antigens have not made much headway, except in the case of virally induced tumors (see p. 409). In contrast, there has been considerable success in identification of individually distinct antigens.

In addition to these experimental studies, several clinical observations point to the existence of tumor-protective immunity in humans. These include the increased relative risk of cancers in patients who are immunosuppressed because they are kidney transplant recipients (Fig. 22.4) or for a variety of other reasons (Fig. 22.5).

Relative risk of tumors in immunosuppressed kidney transplant

tumor type	approximate relative risk
Kaposi's sarcoma	50–100
non-Hodgkin's lymphoma	25–45
carcinoma of the liver	20–35
carcinoma of the skin	20–50
carcinoma of the cervix	2.5–10
melanoma	2.5–10
lung	1–2

Fig. 22.4 In all forms of immunodeficiency the relative risk of developing tumors in which viruses are known to play a role is greatly increased. This is the case for all those listed except cancer of the lung. The relative risks vary in different studies according to the duration of follow-up and the presence of co-factors such as sunlight for skin cancer.

Q. The observation that immunosuppressed individuals are more susceptible to tumors can be explained in at least two different ways. Suggest what these explanations might be.

A. The individual may be unable to control virus infections that lead to cancer (oncogenic viruses), or they may be unable to survey and control primary tumors that arise by mutation. The explanations are not mutually exclusive.

TUMOR ANTIGENS HAVE BEEN CHARACTERIZED BY THREE MEANS

The two broad approaches to the identification of tumor antigens are shown in Fig. 22.6 and discussed individually below. Not surprisingly, the approaches have yielded results that are not fully concordant with each other. These differences have helped highlight a fascinating interplay between immunity and tolerance to tumors, as discussed in a later section (see p. 413).

Tumor-specific antigens defined by immunization and challenge all belong to the family of HSPs

When tumors were biochemically fractionated and individual protein fractions tested for their ability to elicit protective tumor immunity, a number of tumor-protective antigens were identified in diverse tumor models, such as mouse sarcomas, melanomas, colon and lung carcinomas, and rat hepatomas.

Interestingly, regardless of the tumor models used, all antigens were found to belong to the family of proteins known as the **heat-shock proteins** (HSPs), which:

- constitute a very large family of intracellular soluble proteins that are organized into HSP families based on molecular size (Fig. 22.7);
- were shown to elicit protective immunity;
- were of the HSP90 (gp96 and HSP90), HSP70 (HSP110 and HSP/c70), calreticulin, and HSP170 (also known as grp170) families.

HSPs have to be isolated directly from tumors to be immunologically active

Two aspects of HSP-elicited tumor immunity are notable.

- First, HSPs are present in normal tissues as well as tumors, and normal tissue-derived HSPs do not elicit

Tumor viruses and immunodeficiency

cause of immunodeficiency	common tumor types	viruses involved
inherited immunodeficiency	lymphoma	EBV
immunosuppression for organ transplants or due to AIDS	lymphoma cervical cancer skin cancer liver cancer Kaposi's sarcoma	EBV papilloma viruses probably papilloma viruses hepatitis B and C viruses human herpes virus 8
malaria	Burkitt's lymphoma	EBV
autoimmunity	lymphoma	EBV

Fig. 22.5 Skin cancer is the most common form of tumor in absolute numbers in organ transplant recipients. In other forms of immunodeficiency, tumors of the immune system dominate. Most normal adults carry both Epstein–Barr virus (EBV) and many papilloma viruses throughout life with no ill effects because they have antiviral immunity.

Comprehensive list of major types of tumor antigen

	category of antigen	individual antigens	tumors in which identified	activity in mouse models	human vaccine trials
tumor transplantation (mouse models)	heat-shock protein–peptide complexes	Gp96	fibrosarcoma, lung carcinoma, lymphoma, prostate carcinoma, etc.	+	phase III trials in renal carcinoma, melanoma underway
		Hsp70	fibrosarcoma, colon carcinoma	+	phase II trials in CML underway
		Hsp90, CRT Hsp110, 170	fibrosarcoma colon carcinoma	+	none yet
T cell/antibody (SEREX) reactivity (mostly human studies, some mouse models)	cancer/testes antigens	NY-ESO-1 MAGE, etc.	melanomas other tumors	N/A**	a number of phase I, II trials completed and underway
	differentiation antigens (lineage-specific)	melan A tyrosinase Gp100 PSA, etc.	melanomas (except PSA) (prostate carcinoma)		
	antigens with broader expression	human – CEA, MUC, HER2, G250	many tumors, except G250 (renal carcinoma only)		
		murine – P1A, gp70	many tumors	–	N/A
	common tumor-specific antigens	mutated p53, mutated ras, BCR-ABL, etc. papilloma virus?	many tumors	N/T*** exc. a single positive study with p53	no significant ongoing clinical effort
	unique tumor-specific antigens (mutations)	human – MUM-1, 2, β-catenin, HLA-A2-R170I, ELF2m, myosin-m, caspase-8, KIAA0205, HSP70-2m CDK4*, TRP2*, NA17A*	melanoma	N/A	N/A
		murine – Erk2, RNA helicase ribosomal proteins L9, L11	fibrosarcomas squamous cell carcinomas	+	none
	viral antigens	(see Fig. 22.5)			

CML, chronic myeloid leukemia; SEREX, serological analysis of recombinant cDNA expression libraries; *expressed on >1 melanoma, but not on any normal tissues; N/A**, not applicable; N/T***, not tested

Fig. 22.6 A comprehensive list of the major types of tumor antigen.

tumor rejection. They have to be isolated from tumors to be immunologically active.

- Second, HSPs elict immunity specifically against the individual tumors from which they are isolated.

These two observations suggested that HSPs in tumors differ from those in normal tissues and that HSPs in each tumor differ from the same molecules in other tumors.

This conundrum was resolved by the demonstration that the HSP molecules chaperone peptides in a peptide-binding pocket, much as the MHC molecules do (though the structural details of the pockets in HSP and MHC differ) (Fig. 22.8). The specificity of immunogenicity derives from the peptides rather than the HSP itself – dissociation of HSP-associated peptides from HSPs abrogate the tumor rejection activity.

What are the HSP-chaperoned tumor antigenic peptides?

HSPs chaperone a large repertoire of peptides generated within the cells from which they are isolated.

Major heat-shock proteins*

HSP family	members	intracellular location
small HSPs	HSP10, GROES, HSP16, α-crystallin, HSP25, HSP26, HSP27	cytosol
HSP40	HSP40, DNAJ, SIS1	cytosol
HSP47	HSP47	endoplasmic reticulum
calreticulin	calreticulin, calnexin	endoplasmic reticulum
HSP60	HSP60, HSP65, GROEL	cytosol and mitochondria
HSP70	HSP72, HSP/c70/(HSP73), HSP110/SSE, DNAK, SSC1, SSQ1, ECM10, GRP78(BiP), grp170	cytosol mitochondria endoplasmic reticulum
HSP 90	HSC84, HSP86, HTPG gp96 (grp94, Erp99, endoplasmin)	cytosol endoplasmic reticulum
HSP100	HSP104, HSP110** CLP proteins HSP78	cytosol cytosol mitochondria

*the list is not all-inclusive; **distinct from the HSP member with the same name; HSP, heat-shock protein

Fig. 22.7 HSPs constitute a very large family of intracellular soluble proteins, which are organized into HSP families based on molecular size. (Redrawn from Srivastava P. Nat Rev Immunol 2002;2:185–194. Copyright 2002, Nature Reviews Immunology www.nature.com/reviews)

HSP-chaperoned peptide pools contain cytotoxic T lymphocyte (CTL) epitopes and CTL epitope precursors for any antigens that the cell expresses (Fig. 22.9). This evidence comes from mouse and human tumors, normal tissues, and virus-infected cells. Hence, the HSP-chaperoned peptides contain among them any tumor-specific antigenic epitopes present in the tumor cell or the antigenic fingerprint of the tumor from which the HSPs are isolated.

Unique immunogenicity of HSP–peptide complexes

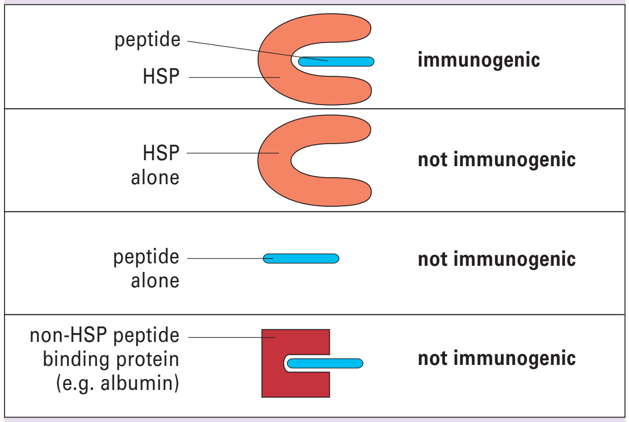

Fig. 22.8 HSP molecules chaperone peptides in a peptide-binding pocket. The specificity of immunogenicity derives from the peptides rather than the HSP itself – dissociation of HSP-associated peptides from HSPs abrogated the tumor rejection activity.

HSP molecules bind to APCs and target peptides with high efficiency

The HSP molecule itself plays at least two crucial roles other than chaperoning peptides:

- HSPs bind antigen-presenting cells (APCs) such as macrophages and dendritic cells (DCs) through HSP receptors such as CD91 and thus target the peptides chaperoned by them into the APCs with high efficiency;
- further, the HSP-chaperoned peptides, once introduced into the APC, follow the endogenous as well as the exogenous pathway of antigen presentation and are processed and re-presented by the MHC I and MHC II molecules of the APCs (Fig. 22.10).

Q. How can an APC present internalized antigen via MHC class I molecules?

A. It can do this via the mechanism of cross-presentation (see Fig. 7.15).

It is by this mechanism that immunization with tumor-derived HSPs elicits a CD8 as well as CD4 response against the tumors. In addition, the HSP molecules stimulate the APCs to mediate maturation of DCs and secretion of an array of cytokines that provide the innate milieu for the adaptive response.

CTL epitopes isolated from HSP–peptide complexes

antigen source	HSP with which associated	MHC class I molecule that presents epitope
tumor antigens		
PRL1e mouse leukemia	HSP70, HSP90, gp96	H-2L^{d}
human melanoma MART-1	HSP70	HLA-A2
human melanoma tyrosinase	HSP70	HLA-A2
human melanoma gp100	HSP70	HLA-A2
viral antigens		
vesicular stomatitis	gp96	H-2K^{b}
herpes simplex 2	gp96	H-2^{d}
influenza	gp96, HSP70	H-2K^{b}
simian virus 40	gp96	H-2D^{b} , H-2K^{b}
hepatitis B	gp96	
intracellular bacterial antigens		
Mycobacterium tuberculosis	gp96	H-2^{d}
Listeria monocytogenes	gp96	H-2^{d}
model antigens		
β-galactosidase	gp96	H-2L^{d}
ovalbumin	gp96	H-2K^{b}
normal cellular antigens		
minor histocompatibility antigens	gp96	H-2K^{b}, H-2K^{b}

Fig. 22.9 HSP-chaperoned peptide pools contain CTL epitopes and CTL epitope precursors for any antigens that the cell expresses. (Redrawn from Srivastava P. Nat Rev Immunol 2002;2:185–194. Copyright 2002, Nature Reviews Immunology www.nature.com/reviews)

The immunization-challenge model has led to identification of antigen carriers (i.e. HSPs), but not of the antigens themselves. As a result of exogenous administration, immunization with antigens themselves would generally be expected to elicit CD4 response, which in and of itself is not sufficient to elicit tumor immunity. The HSP-chaperoned antigens were picked up in this assay because they have the ability to elicit CD8 as well as CD4 responses despite being exogenously introduced. As discussed above, this response is derived from the ability of HSPs to interact with APCs through receptors such as CD91.

The status of clinical trials using tumor-derived HSP–peptide complexes for immunotherapy of human cancer is discussed below (see p. 417).

‘Tumor-specific antigens’ defined by T cells show a wide spectrum of specificity

Many studies have identified tumor-reactive T cells in blood or within a tumor, and these findings have thus supported the idea of tumor antigens.

The work of Thierry Boon and his colleagues first made it technically possible to identify the CTL epitopes of cancer cells being recognized by the tumor-reactive T cells.

Although the idea of tumor-specific antigens in mouse models of cancer was based on tumor rejection in vivo and thus had connotations of tumor specificity, the tumor antigens defined by tumor-reactive T cells show a wider spectrum of specificity, and their connection with tumor immunity in vivo, has been tenuous thus far.

The T cell-defined tumor antigens of murine tumors have been defined in a mastocytoma, two fibrosarcomas, a squamous cell carcinoma, and a colon carcinoma because these tumor lines are in popular use.

Similarly, much of the corresponding work in human tumors has been carried out in melanomas because melanoma cell lines are easier to establish in culture, rather than because of any unique immunogenicity of human melanomas.

The source of T cells is critically different in human and mouse studies

The murine T cells used for definition of tumor antigens have been obtained typically from splenocytes of mice that have been immunized against the tumor and are tumor-immune.

The human T cells have been obtained typically from patient’s blood or from the tumor tissue itself.

These differences in the sources of T cells used may lie at the root of some of the differences in the nature of tumor antigens defined in the murine and human systems, as discussed below. The tumor antigens identified as T cell epitopes fall into the following categories (see Fig. 22.6).

Cancer/testes antigens are expressed only in testes

Cancer/testes (CT) antigens, as the name suggests, are expressed on cancer cells and in testes, but not in other normal adult tissues. MAGE, BAGE, GAGE, and NY-ESO1 are examples of this class of antigen. Individual epitopes within CT antigens have been defined for CD8 as well as CD4 lymphocytes. Most CT antigens have been identified in humans, although murine examples such as PEM exist as well.

Differentiation antigens are lineage specific and not tumor specific

Differentiation antigens, which are lineage specific, but not tumor specific, are expressed on normal tissues (melanocytes) as well as tumors (melanomas). Examples of such antigens include MART-1/Melan-A, tyrosinase, gp100, Trp1, and Trp2. Individual epitopes within

Mechanism of specific immunogenicity of HSP–peptide complexes

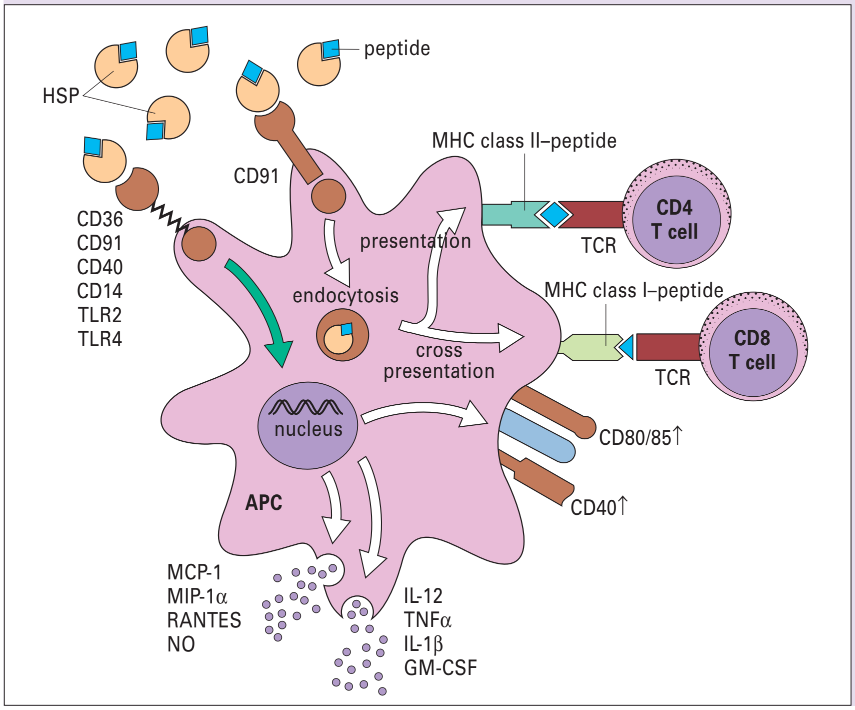

Fig. 22.10 Once introduced into the APC, the HSP-chaperoned peptides follow the endogenous class I as well as the exogenous class II pathway of antigen presentation and are processed and re-presented by the MHC class I and MHC II molecules of the APCs. HSP ligation induces production of chemokines and cytokines, and upregulation of CD40 and CD80/85. (MCP-1, monocyte chemotactic protein-1; MIP-1α, macrophage inflammatory protein-1α) (Redrawn from Srivastava P. Nat Rev Immunol 2002;2:185–194. Copyright 2002, Nature Reviews Immunology www.nature.com/reviews)

differentiation antigens have been defined for CD8 as well as CD4 lymphocytes. Most differentiation antigens have been identified in human melanomas.

Widely expressed antigens are expressed in cancers and normal tissue

These antigens are expressed on normal tissues as well as a wide array of tumors; in some instances differences in glycosylation may confer source specificity. Examples include:

- carcinoembryonic antigen (CEA), HER2/Neu, and Mucin (MUC) in humans; and
- AH1 in a mouse colon carcinoma.

Also in this category is a hypoxia-inducible carbonic anhydrase (MN/CA IX), which has been reported to be expressed in a majority of renal carcinomas, but in very few normal tissues tested.

Common tumor-specific antigens are not expressed on normal tissue

Common tumor-specific antigens can be expressed on tumors, but not on normal tissue. Such antigens, which have only been reported from the human system thus far, include antigenic epitopes generated by apparently tumor-specific alternate splicing such as TRP-2 and NA17-A (both in melanomas). This category also includes the mutated oncogenes, which are discussed in a separate category (see p. 410), because they were not identified on the basis of T cell reactivity.

Unique tumor-specific antigens have been characterized in melanomas

Individually tumor-specific or unique antigens presented by MHC class I or class II molecules have been characterized in:

- human melanomas (CDK-4, MUM-1, MUM-2, β-catenin, HLA-A2-R170I, ELF2m, myosin-m, caspase-8, KIAA0205, HSP70-2m); and
- murine tumors (Erk2, RNA helicase, ribosomal proteins L9, L11) alike (see Fig. 22.6).

In terms of their specificity, these antigens are the most similar to the individually tumor-specific antigens of murine tumors defined by tumor transplantation studies. Immunization with these antigens renders the mice resistant to tumor challenge. These are the only tumor antigens that have been shown to possess this essential property. The defined unique antigens of human tumors have not been tested for clinical activity.

T cell epitopes of viral antigens have been identified

T cell epitopes of viral antigens of virus-induced tumors have been identified such as:

- the T antigen of SV40 and the polyoma viruses;
- the E6 and E7 antigens of the human papilloma viruses that cause cervical cancer; and
- a number of antigens of the EBV (Fig. 22.11).

Some of these, such as the HPV antigens, are being tested for clinical activity against human tumors.

Microorganisms and human tumors

tumor	organism
adult T cell leukemia	human T leukemia virus-I (HTLV-I)
Burkitt's lymphoma and lymphoma in immunosuppression	EBV
cervical cancer	human papilloma viruses (HPV 16 and 18, and others)
liver cancer	hepatitis B and C
nasopharyngeal cancer	EBV
skin cancer	probably human papilloma viruses
stomach cancer	*Helicobacter pylori*

Fig. 22.11 EBV causes Burkitt's lymphoma in endemic malaria areas of Africa and nasopharyngeal carcinoma in China, suggesting that co-factors, either genetic or environmental, are required for tumor development. *Helicobacter pylori* is the only bacterium so far known to be involved in the etiology of human cancer.

'Tumor-specific antigens' defined by antibodies are rarely tumor specific

The search for antibodies that discriminate between cancer cells and normal cells has a long pedigree and has been carried out with the whole range of tools starting from antisera to panning antibody libraries.

This search has helped identify a whole range of molecules central to modern immunology, such as thymus leukemia antigen, CD8, and the T cell receptor (TCR), but, with very rare exceptions, has been largely unable to identify what it set out to do (i.e. identify cancer-specific antigens). The search has been largely unsuccessful and rarely have tumor-specific antibodies been generated.

Most anti-tumor antibodies, like anti-tumor T cells, happen to recognize CT antigens, differentiation antigens, and even more broadly distributed common antigens (see Fig. 22.6).

A recent development in this approach has focused attention on the fact that the tumor-bearing host develops antibodies to a repertoire of antigens that is distinct from the corresponding repertoire of antibodies seen in a non-tumor-bearing host. In this approach (known as SEREX for serological analysis of recombinant cDNA expression libraries), sera of cancer patients or tumor-bearing animals are used to screen expression libraries of tumors and an antibody repertoire is characterized. Although the individual antibodies are not tumor specific, the total repertoire shows relative specificity for cancer. The physiological meaning of this repertoire is unclear, but it appears to be a potent tool for diagnostic as well as monitoring purposes.

Antibodies used in the diagnosis of cancer may not be tumor specific

It is worthwhile at this point to comment on two types of antibodies that are used in diagnosis or treatment of cancers (Fig. 22.12).

As serum antibodies are technically easy to measure, they have always attracted the attention of diagnosticians. Thus:

- antibody to carcinoembryonic antigen (CEA) is often used as a marker for progression or status of certain carcinomas;
- similarly, a SEREX protein array may be used some day to monitor progression of a given cancer.

However, CEA or SEREX antigens are not tumor-specific antigens and diagnostic markers need not necessarily have the specificity required of a 'tumor-specific antigen'.

Antibodies to a B cell surface antigen, CD20, epidermal growth factor receptor, and HER2/Neu have now been approved for treatments, respectively, of B lymphoma, colorectal cancers, and breast cancers. Although these antibodies have shown some efficacy in the treatment of certain cancers at certain stages, they:

- do not recognize tumor-specific antigens;
- are actually used as pharmacological rather than immunological reagents.

This distinction does not diminish their utility and is highlighted only to draw attention to the fact that the term 'tumor-specific antigen' is used here to illustrate primarily the idea of host immune response to cancer, how it is elicited, how it is successful, and how it is not.

Tumor-specific antigens generated by mutated oncogenes may not be tumor protective

The discussion thus far has dealt with tumor antigens defined by experimental tools. Several additional antigens have been tested as tumor antigens because of their restricted or relatively restricted expression on cancer cells, with the reasonable assumption that they would elicit tumor-specific and perhaps tumor-protective immune responses. Antigenic epitopes generated by mutated oncogenes such as *ras* and *p53* as well as gene

(1) Markers for detection of cancers

marker	cancer
carcinoembryonic antigen (CEA)	adenocarcinomas of colon, breast, lung, breast, ovary
Ca 19-9	colon, pancreatic, and breast cancers
Ca 125	ovarian cancer
prostate-specific antigen (PSA)	prostate cancer
prostatic acid phosphatase (PAP)	prostate cancer
alpha-fetoprotein (AFP)	hepatocarcinoma, gonadal germ cell tumor
lactate dehydrogenase (LDH)	lymphoma, Ewing's sarcoma, melanoma
human chorionic gonadotropin (hCG)	gonadal germ cell tumor
calcitonin	medullary cancer of thyroid
catecholamines	pheochromocytoma
common acute lymphoblastic leukemia antigen (CALLA)	acute lymphoblastic leukemia
monoclonal immunoglobulin	myeloma
CD25	hairy cell leukemia, adult T cell lymphoma/leukemia
CD30	Hodgkin's lymphoma

(2) Antibodies used for therapy of selected cancers

antibody	antigen	conjugated to radioactivity/toxin	cancer
Herceptin®	HER2/Neu receptor	none	breast cancers
Rituxan®	CD20	none	non-Hodgkin's B cell lymphoma
Erbitux®	EGFR	none	metastatic colorectal cancers
Campath®	CD52	none	CLL
Zevalin®	CD20	indium-111, yttrium-90	refractory B cell lymphoma
Bexaar®	CD20	iodine-131	refractory B cell lymphoma
Mylotarg®	CD33	calicheamicin toxin	AML

Fig. 22.12 Neither the 'antigens' used as diagnostic cancer markers (**1**) nor the antibodies used for therapy of selected cancers (**2**) are tumor specific. CLL, chronic lymphatic leukemia; AML, acute myeloid leukemia.

fusion-generated oncogenes such as *bcr-abl*, which are, by definition, tumor specific, come into this category. These antigens could, in principle, be broadly applicable tumor antigens.

The idea that mutant proteins are recognized by the immune system has been explored to a degree in murine and human systems, and immune reactivity to peptides derived from such mutated oncogenes has been detected. However, no evidence of their general tumor-protective immunogenicity has emerged.

It is conceivable that the immune response to these alterations is tolerized early in tumorigenesis and, indeed, such tolerization may be a pre-condition for tumorigenesis.

The same may be said for telomerase, expression of which in cancer cells allows them to circumvent the rules of cellular mortality, and which too has been conceived of and tested as a tumor antigen with tantalizing but not compelling results.

The non-mutated oncogene HER2/Neu as an immunizing agent has been the subject of extensive studies in HER2/Neu-transgenic mouse models, which develop spontaneous mammary carcinomas. These experiments have shown that immunization with the oncogene inhibits spontaneous mammary carcinogenesis in these mice and may also promote regression of pre-existing lesions. The reasons for the differences between tumor-protective immunogenicities of these various antigens are unclear at present.

VIGOROUS ANTI-TUMOR IMMUNE RESPONSES ARE COMPROMISED BY REGULATORY MECHANISMS

Successful tumor immunity is rare in patients who have cancer

Mechanisms of tumor immunity have been examined mostly in mouse models, partly because successful tumor immunity is rare in patients who have cancer.

Further, as successful tumor immunity is rare in the tumor-bearing setting in mouse models as well, much of the work has been done in a prophylactic setting, which is not applicable to the human situation.

Not surprisingly, the pathways to elicitation of immune response to tumors are straightforward (Fig. 22.13). The tumor inoculum (with its antigenic load) is taken up by the APCs at the site of immunization and is cross-presented by them to the naive CD8 cells in the draining lymph nodes. Both responses are generally necessary in the mouse models tested and both responses have been shown to be present in the cancer patients studied.

Antibodies have not generally been shown to be protective in the natural setting.

NK cell activity has been demonstrated most commonly, but its necessity has been rarely examined critically. In the few studies where it has been examined, it appears that NK cells play a crucial role in the immune response to cancers. Clearly, the cytokines necessary for the effector functions of CD4, CD8, and NK cells, such as IL-2, IFNγ, IL-12, and others, are necessary as well.

In selected models, significant roles for innate components such as neutrophils and eosinophils in tumor rejection have been shown.

Despite an immune response, tumors continue to grow

Despite clear evidence for the existence of tumor-specific antigens and the immune response elicited by them,

A working diagram of the possible mechanism of tumor immunity

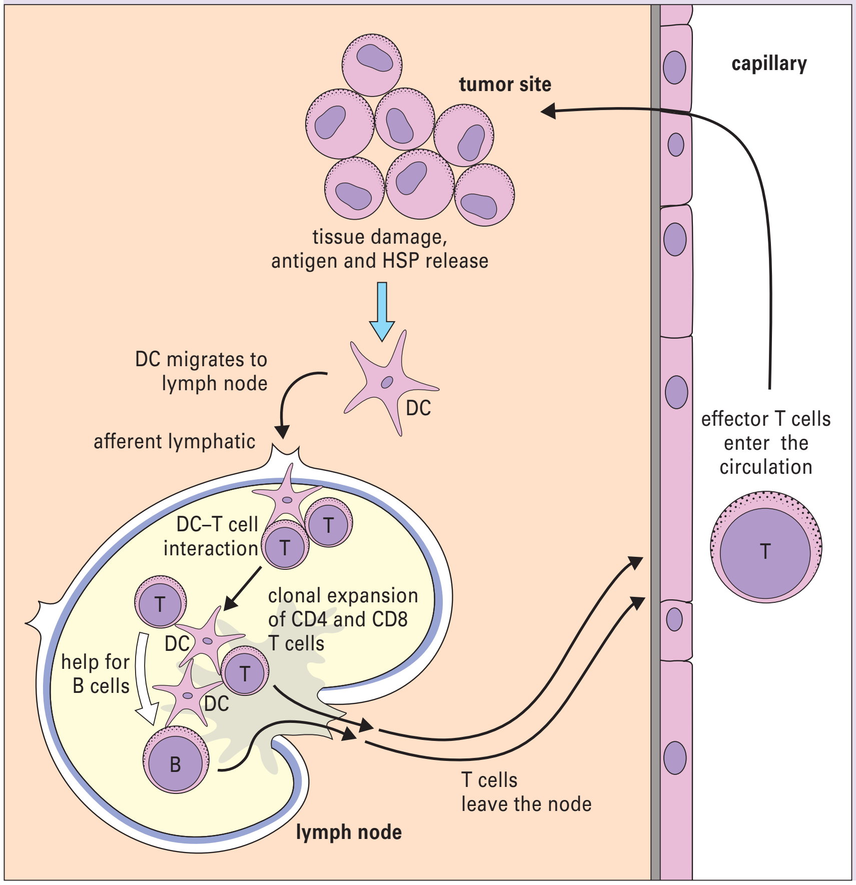

Fig. 22.13 The tumor inoculum (with its antigenic load) is taken up by APCs at the site of immunization and is cross-presented by them to the naive CD8 cells in the draining lymph nodes.

Concomitant immunity

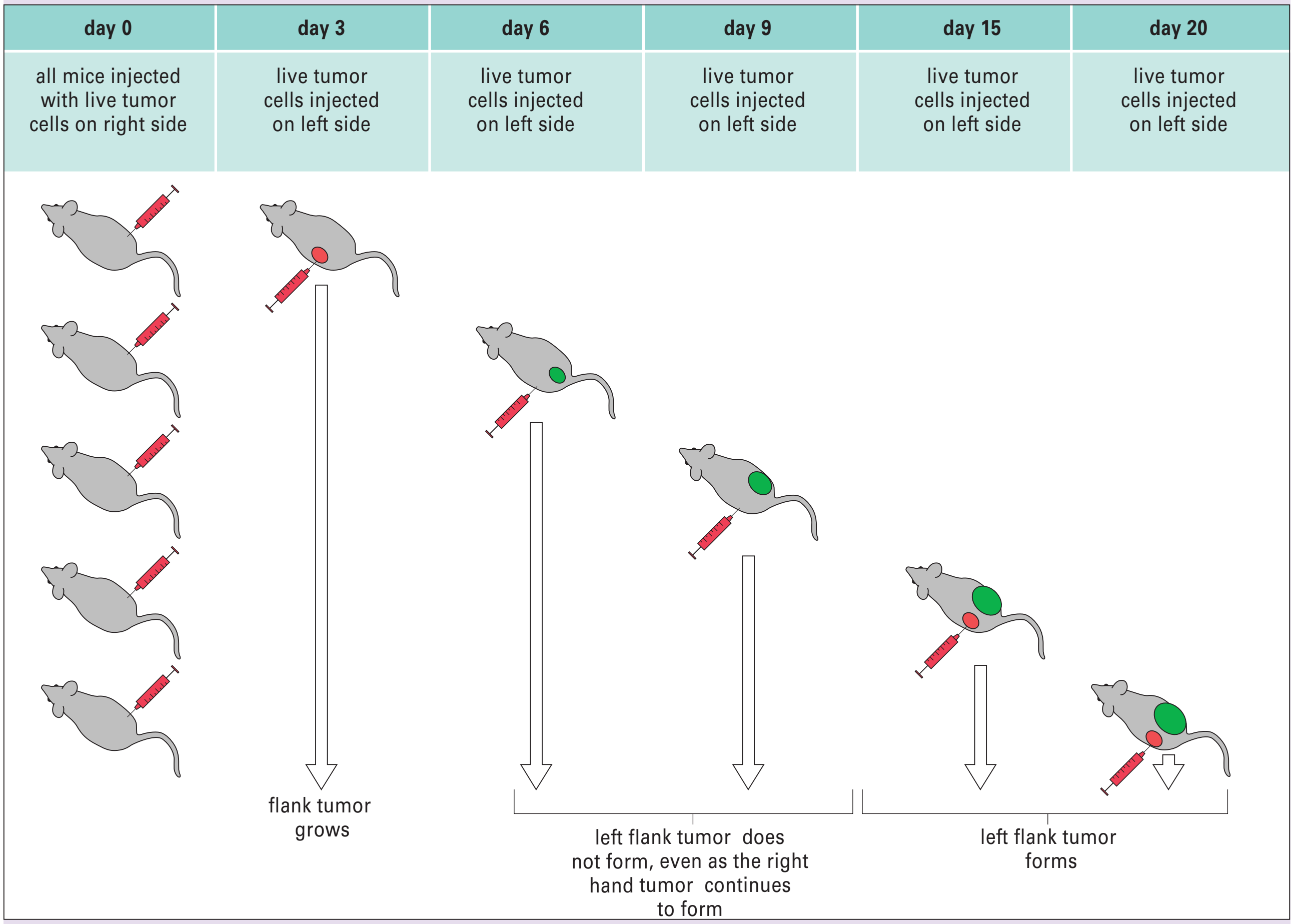
day 0
all mice injected with live tumor cells on right side
day 3
live tumor cells injected on left side
day 6
live tumor cells injected on left side
day 9
live tumor cells injected on left side
day 15
live tumor cells injected on left side
day 20
live tumor cells injected on left side
flank tumor grows
left flank tumor does not form, even as the right hand tumor continues to form
left flank tumor forms

Fig. 22.14 Animals with already growing tumors are resistant to a second tumor challenge even as the first tumor continues to grow.

tumors generally continue to grow. In this regard, Ehrlich made a curious observation that remains at the center of cancer immunity. He noted that animals with already growing tumors were strangely resistant to a second tumor challenge even as the first tumor kept growing (Fig. 22.14). This phenomenon, termed **concomitant immunity** as early as 1908, remained relatively unexamined until recently.

Concomitant immunity shows two aspects of tumor immunity

Concomitant immunity shows two aspects of tumor immunity:

- first, a growing tumor elicits in the primary host a tumor-protective immune response;
- second, although this response is sufficient to eliminate a nascent tumor, it fails to eliminate the tumor that elicited the response.

It was shown that concomitant immunity was tumor specific and operational only within a narrow window of 7–10 days after tumor implantation; if the second tumor was implanted beyond this time, it was not rejected. The lack of immunity beyond the narrow window was attributable to a new population of suppressor T cells that appeared at that time (Fig. 22.15). Similar to the phenomenon of concomitant immunity, it was also noted that mice in the process of rejecting an allograft were unable to concomitantly reject a growing tumor bearing the same alloantigens as the allograft.

Q. How do you interpret the observation above?

A. One would expect an allograft to be rejected in these circumstances, therefore there is something about the tumor that allows it to evade a normal allospecific rejection reaction.

Immunization is effective prophylactically but rarely as therapy

The findings discussed above fit in well with the observations that, although naive mice can be immunized successfully using irradiated tumor cell vaccines in almost any tumor model, the same irradiated cell vaccine is ineffective at treating established tumors.

Effector CD8 and suppressor CD4 cells in concomitant immunity

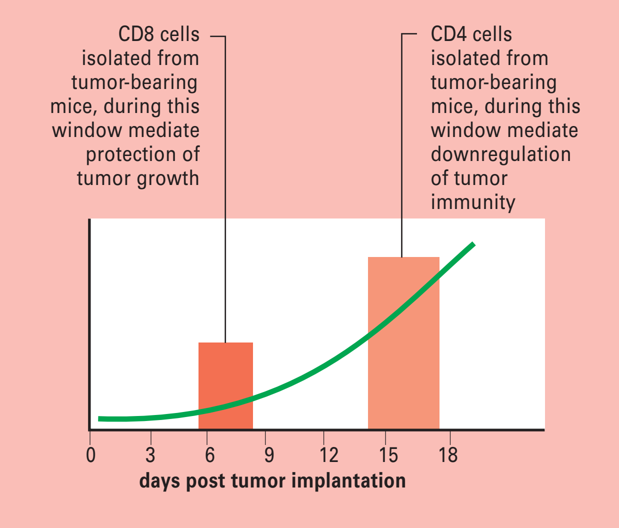

Fig. 22.15 Concomitant immunity is tumor specific and operational only within a narrow window of 7–10 days after tumor implantation. If the second tumor is implanted beyond this time, it is not rejected. The lack of immunity beyond the narrow window is attributable to a new population of suppressor T cells that appear at that time.

Prophylaxis is relatively effective, even if begun as late as 3 days before tumor implantation, but therapy is ineffective even if begun as early as 2 days after tumor implantation (see Fig. 22.2).

It is easy, and incorrect, to infer from these observations that the tumor-bearing mice or cancer patients are generally immunosuppressed. Tumor-bearing mice generally mount a vigorous immune response to model antigens and even unrelated tumors, almost until the very end of their lives. Similarly, patients with cancer do not generally succumb to the opportunistic infections that are the hallmark of general immunosuppression.

What are the mechanisms behind the change from naive to tumor-bearing states?

Collectively, these and similar other observations have shaped the thinking that, immunologically speaking, the tumor-bearing host is in a radically different state compared with the naive host.

What are the mechanisms behind this change of status? They are exactly the same as envisaged for the mechanisms for the initiation and maintenance of peripheral tolerance. None of the explanations is fully satisfactory by itself, but each perhaps contributes to the final state of tolerance in some measure. Immune unresponsiveness per se is addressed in detail in Chapters 12 and 19. Only some aspects that are of specific relevance to tumor immunity are discussed here (Fig. 22.16).

Downregulation of MHC I molecule expression may result in resistance to recognition and lysis

Loss of MHC class I molecules by tumor cells has often been suggested to be one such mechanism.

A number of studies have shown:

- downregulation of expression of one or more MHC class I alleles;
- loss of β_2-microglobulin; or
- loss or downregulation of any of the several components of the antigen processing machinery.

Although these alterations are clearly likely to inhibit recognition of tumor cells by T cells, they are also more likely to make tumors more susceptible to become targets of NK cells.

On the whole, it is not clear what effect, if any, these alterations have on the net ability of a tumor to escape the immune response in vivo.

Generation of antigen-loss variants is another mechanism that no doubt plays a role to some degree in the immunological escape of tumor cells. However, because we do not yet have a significant knowledge of the identity of truly tumor-protective antigens, the role of antigen-loss variants cannot be critically examined.

Inhibitory cytokines such as TGFβ and IL-10, and others, have often been proposed to downregulate antitumor immune responses.

Major mechanisms of suppression of anti-tumor immune response

at the level of generation of response
inhibitory cytokines such as transforming growth factor-β (TGFβ) and interleukin-10 (IL-10)
attenuation of response
general mechanisms of peripheral tolerance inhibition of T cell activity through CTLA-4 downregulation of immune response through $CD25^+$ and other suppressor cells
at the effector level
resistance to recognition and lysis by downregulation of MHC class I molecules generation of antigen-loss variants

Fig. 22.16 Only mechanisms of suppression of specific relevance to tumor immunity are listed here.

T cell activity is inhibited through CTLA-4

An exciting explanation for unresponsiveness is the role of inhibition of T cell activity through CTLA-4. CTLA-4 and CD28 belong to the same family of molecules and both are expressed on CD8 T cells.

Q. What are the functions of CD28 and CTLA-4 (CD152) on T cells?

A. CD28 transduces co-stimulatory signals to the T cell, while CTLA-4 also ligates B7, but is inhibitory (see Fig. 22.17).

CTLA-4 inhibits the T cell by raising the stimulatory threshold or by inhibiting the proliferative drive of T cells (Fig. 22.17). The biological role of CTLA-4 appears to lie in limiting the T cell response to foreign antigens as well as to autoantigens.

Administering antibodies to CTLA-4 (that inhibit CTLA-4:B7 interactions) to mice bearing a broad array of tumors inhibited tumor growth, even when the antibody was administered after the tumors were visible and palpable (Fig. 22.18). Such activity was generally seen only against the more immunogenic tumors and not against a poorly immunogenic melanoma (e.g. B16). In that instance, combination of anti-CTLA-4 antibody with a vaccine consisting of irradiated melanoma cells that were also transfected with the cytokine granulocyte–macrophage colony stimulating factor (GM-CSF), resulted in a stronger anti-tumor response than by anti-CTLA-4 antibody or the vaccine alone. Clinical development of this idea is discussed below.

Similar results have been observed in other tumor models. These results support the notion derived from studies on concomitant immunity that progressive tumor growth results in the generation of inhibitory influences on the anti-tumor immune response.

Mechanism of action of anti-CTLA-4 antibody

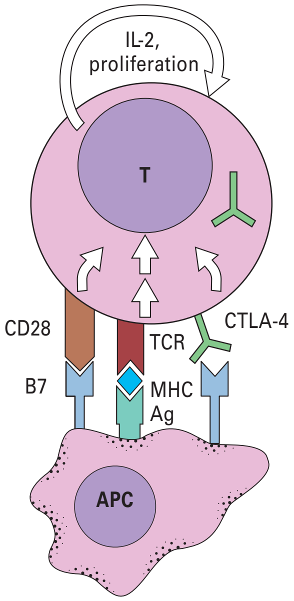

Fig. 22.17 CTLA-4 inhibits the T cell by raising the stimulatory threshold or by inhibiting the proliferative drive of T cells.

Mouse tumor rejection using anti-CTLA-4 antibody

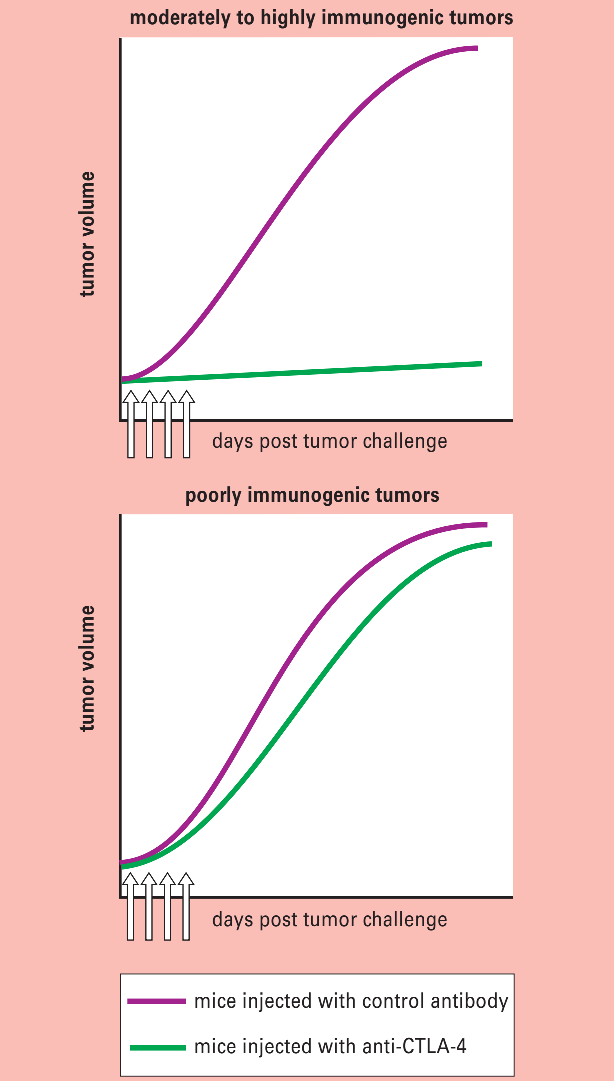

Fig. 22.18 Administering anti-CTLA-4 to mice bearing a broad array of tumors inhibited tumor growth. Such activity was generally seen only against the more immunogenic tumors and not against poorly immunogenic tumors. The arrows denote the administration of antibodies.

Abrogation of CD25$^+$ cells leads to protective tumor immunity

The notion that progressive tumor growth results in the generation of inhibitory influences on the anti-tumor immune response is further supported by the recent work on the CD25$^+$ CD4$^+$ Treg cells (see Chapter 19). These cells have been shown to play a regulatory (i.e. suppressive) role in CD8 T cell responses in general, including autoimmune responses.

Recent studies have shown that abrogation of the CD25$^+$ subpopulation (through anti-CD25 antibodies or through genetic manipulation) in tumor-bearing mice leads to a robust T cell response and protective tumor immunity even in an aggressive tumor model such as the B16 melanoma. Conversely, re-addition of these cells can suppress the anti-tumor immune response.

Results consistent with these have been observed in patients with melanoma whose regulatory CD4$^+$ T cells specifically inhibit the CD8$^+$ activity against autologous melanoma cells, but not against other targets.

More recent studies have examined this paradigm through the prism of the CD25$^+$ T cells. In a study of patients with ovarian cancer, such cells were shown to be associated with a higher risk of death and reduced survival. Interestingly, the CD25$^+$ T cells were shown to migrate preferentially to the solid tumors and the ascites, but rarely to the draining lymph nodes.

Such results have begun to point inexorably to the general theme that, despite the poor clinical outcomes in advanced cancer, the host does mount a vigorous anti-tumor immune response, which is compromised by regulatory mechanisms. Manipulation of such regulatory mechanisms for enhancing cancer immunity is bound to influence the fine balance between tolerance and autoimmunity. In certain contexts, that may be a reasonable price to pay.

IMMUNOTHERAPY FOR HUMAN CANCER

Animal models are limited in the translation of therapy

There exists a ‘common wisdom’ that treatment of cancers of mice is easy and has no bearing on treatment of human cancers. This belief is as common as it is incorrect. Mouse tumors are hard to cure – there does not exist a single publication that reports curing a mouse with stage IV disseminated disease.

Most approaches study prophylactic vaccination in mice, a smaller number begin treatment on the day of tumor challenge, and only a handful of studies begin treatment more than 10 days after tumor challenge. In most mouse tumor models, mice die within 4–6 weeks of tumor challenge, and thus the window of treatment is extremely narrow.

Most approaches to immunotherapy of human cancer have either never been tested in appropriate mouse models or have failed to show anti-tumor activity when tested. It is important to bear this in mind while examining the three major categories of approaches to immunotherapy of human cancer, discussed below.

Antibodies have been used successfully

As antibodies are the oldest known immunological reagents, it is only to be expected that the first, and thus far among the most successful, approaches to human cancer immunotherapy has been made using these reagents.

Nearly 20 years ago, Ronald Levy and colleagues treated patients with B cell lymphomas using individual patient’s tumor-specific anti-idiotypic antibodies on the premise that the antibodies will recognize and help eliminate their targets – the surface immunoglobulin on the monoclonal lymphomas. The treatment was successful clinically, leading to significant objective tumor regressions, but was limited by the re-emergence of escape variants that did not express the idiotype. This approach has not been pursued further, but remains a powerful reminder of what true tumor-specific antibodies can do to real-life tumors. The transformation of this approach to vaccination is discussed below.

Selected antibodies are now approved for clinical use, but these represent pharmacological rather than immunological use of antibodies (see Fig. 22.12) and include antibodies to:

- CD20 – rituximab – against B cell lymphoma;
- HER2/Neu – trastuzumab (Herceptin®) – against breast cancer;
- epidermal growth factor receptor – cetuximab (Erbitux®) – against colon cancer.

It is ironic that the anti-tumor antibodies that may recognize truly or relatively tumor-specific molecules, and that were the earliest hopes of much of the efforts in this area, have yet to enter the phase of randomized clinical testing. Such antibodies are difficult to characterize and therefore have been slow in development. Nonetheless, they exist and may be expected to be tested by themselves or with radioactive or toxic tags in the near future. Unfortunately, a number of antibodies that were thus tagged and already tested in clinical trials were not very specific and did not show compelling clinical results.

Vaccination can be used to treat cancer

Although the term vaccination is typically used to indicate prophylactic vaccination, cancer researchers use it to indicate the treatment of someone who already has cancer, with agents that stimulate anti-cancer immune response. Several vaccination approaches are currently being pursued (Fig. 22.19).

The idiotypes of B lymphomas have been used as vaccines

Following the use of anti-idiotypic antibodies and the attendant limitations discussed above, the idiotypes of B lymphomas have been used as vaccines. In this approach, a patient’s idiotype is determined by polymerase chain reactions from the tumor tissues, and a synthetic idiotype, conjugated to a carrier such as keyhole limpet hemocyanin, is administered along with GM-CSF. Phase I and II trials with this approach have yielded promising results and phase III trials in patients with follicular B lymphoma are now in progress. For obvious reasons, this approach is limited to B lymphoma and related hematological malignancies.

Q. Which other group of cells is (theoretically) susceptible to anti-idiotypic targeting of tumor therapy?

A. The recombined T cell receptor is also expressed on specific clones of cells. The limitations that apply to the therapy of B cell lymphomas (mutation escape) also apply here, and it is in any case more difficult to generate antibodies to the T cell receptor.

Vaccination approaches to immunotherapy of human cancer currently being tested

vaccine	cancer
unique antigen (individually patient-specific) approaches	
idiotypes + KLH + GM-CSF	B lymphoma
intact autologous irradiated tumor cells + BCG	colon cancer
intact autologous haptenated irradiated tumor cells	melanoma
intact autologous GM-CSF transfected tumor cells	melanoma, renal cancer, lung cancer
autologous tumor lysates	renal cancer
HSP–peptide complexes	melanoma, renal cancer, colon cancer, chronic myeloid leukemia
shared tumor antigen approaches	melanoma
allogeneic irradiated cell lines	melanoma
differentiation antigens	melanoma
cancer/testes antigens	

Fig. 22.19 Several vaccination approaches are currently being pursued.

Immunization with irradiated intact tumor cells has been tested against many cancers

The first randomized trials of immunizing cancer patients with irradiated intact tumor cells were carried out in patients with colorectal cancer, where each patient in the treatment arm received autologous irradiated tumor cells mixed with BCG. The trials showed detectable clinical activity in the vaccine arm, but the activity failed to achieve statistical significance.

Autologous haptenated tumor cells and autologous GM-CSF transfected tumor cells have also been used as vaccines in patients with melanoma, renal carcinoma, pancreatic cancer, and non-small cell lung cancer. Some of the early trials with these approaches have been suggestive of clinical activity.

A non-autologous whole cell strategy has been pursued in stage III and IV melanoma where the vaccines were not made from autologous tumors, but from a collection of irradiated allogeneic lines. The rationale offered for this approach was that melanomas may be expected to contain a number of shared immunogenic antigens and immunization with a collection of melanoma lines increases the odds of immunizing against the antigens expressed by the melanoma of any given patient. However, analysis of the phase III trial indicated lack of any clinical benefit to immunized patients and the trial has now been discontinued.

Q. What evidence is there that this approach to therapy is unlikely to be successful?

A. The result is consistent with the earliest observations in the history of cancer immunity that showed that the immunoprotective antigens of tumors are individually unique and not shared among tumors (see Fig. 22.3).

Tumor lysate immunization is still being tested

Immunization with whole tumor lysates (as opposed to intact irradiated cells) also has a long pedigree and is still being tested.

An allogeneic melanoma lysate named Melacine® was initially tested without significant success.

Autologous tumor lysates have been tested in patients with renal carcinoma in the post-nephrectomy adjuvant setting and a randomized trial using such lysates has shown statistically significant benefit for the immunized patients. The manufacturers of this vaccine have presently applied for it to be approved as standard treatment for patients with renal carcinoma. In principle, this method would be applicable to any cancer.

Defined MHC I epitope immunotherapy does not lead to clinical benefit

The antigenic epitopes of differentiation antigens and cancer/testes antigens have been used extensively and several hundred patients with melanoma have been treated with these. Although anti-peptide CD8 responses have been detected in most studies, these have not generally translated into clinical benefit even for those patients who have shown good CD8 responses. Attempts to understand this anomaly and to build upon these studies are in progress.

Particularly interesting are efforts to vaccinate with altered peptide epitopes with higher affinity for the cognate T cell receptors. Immunization with such altered peptides is able to break tolerance against these self antigens where immunization with native peptides is not.

Q. What problem is intrinsic to this type of tumor therapy?

A. Immunization with differentiation antigens is premised on breaking tolerance and hence carries the near certainty of generating autoimmune responses (see Chapter 19).

This approach relies on the existence of a therapeutic window, which will permit anti-cancer immunity without debilitating autoimmunity.

Immunization with HSP–peptide complexes demonstrates clinical benefit

The role of HSP–peptide complexes in eliciting tumor immunity has been discussed earlier.

The highly successful murine studies have been translated into a series of phase I and II trials involving patients with pancreatic, gastric, stomach, and renal cancers, and with melanoma, B lymphoma, and chronic myelogenous leukemia.

In these trials, surgically obtained tumor specimens (or leukemia cells obtained by leukopheresis) from a given patient are used as the starting material for preparation of gp96–peptide or hsp70–peptide complexes specifically for that patient (Fig. 22.20).

The early trials have shown suggestions of clinical activity in immunized patients and have led to randomized phase III trials in renal carcinoma and melanoma. Interim results from a randomized trial in patients with stage IV melanoma have shown that patients with early stage IV disease, who were treated with gp96, survived longer than those in the control arm. The final results of this trial, and of the phase III renal cancer trial, are expected in early 2006.

Adoptive immunotherapy using T cells: the clinical benefits

Adoptive immunotherapy using T cells has a successful pedigree in murine models of cancer (Fig. 22.21). Clinical experience with bone marrow transplant recipients also provides a strong rationale for the approach.

Patients undergoing high-dose chemotherapy lose their bone marrow and are re-constituted with allogeneic stem cells, which engraft in the recipient. However, the T cells from the donor may see the normal tissues of the host as foreign, thus causing graft versus host disease (GVHD).

Interestingly, patients who develop GVHD also have a lower cancer relapse or graft versus tumor (GVT) incidence. The clinical experience with GVHD and GVT has long remained a compelling piece of evidence for the premise that T cells can eliminate human cancers in vivo (Fig. 22.22).

A number of studies have isolated tumor-infiltrating T cells from cancer patients, expanded them in vitro and infused the expanded cells back into the patients. Such studies have occasionally shown remarkable shrinkage of tumors. However, the fact that such responses are seen only occasionally points to the fact that there remain inherent hurdles that need to be identified and rectified.

Treatment of cancer patients with HSP–peptide complexes

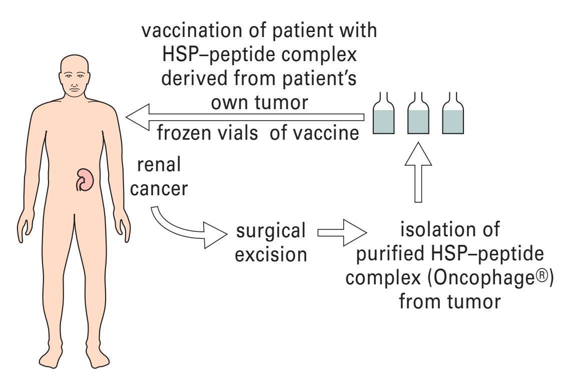
vaccination of patient with HSP–peptide complex derived from patient's own tumor
frozen vials of vaccine
renal cancer
surgical excision
isolation of purified HSP–peptide complex (Oncophage®) from tumor

Fig. 22.20 Surgically obtained tumor specimens from a given patient are used as the starting material for preparation of gp96–peptide or HSP70–peptide complexes specifically for that patient.

Adoptive immunotherapy with T cells

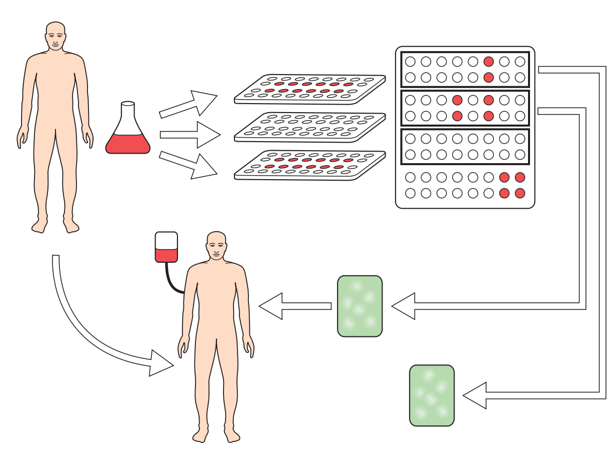

Fig. 22.21 Lymphocytes removed from a patient with a tumor are expanded in vitro. Cells that recognise the tumor are selected and reinfused into the original patient. Adoptive therapy with allogeneic lymphocytes may also be carried out. (Redrawn from Dudley ME, Rosenberg SA. Adoptive-cell-transfer therapy for the treatment of patients with cancer. Nat Rev Cancer 2003;3:666–675. Copyright 2002, Nature Reviews Cancer, Macmillan Magazines Ltd)

In a variation of this approach, cloned T cells with defined specificity have been expanded to very large numbers and infused into patients with melanoma, also to show dramatic tumor shrinkage.

However, hurdles to expansion of T cells as well as their effector functions in vivo remain, and there is considerable ongoing experimental effort to engineer T cells that will retain specificity and autonomy of growth and will be relatively refractory to downregulatory influences of the host.

Immunotherapy with peptide pulsed DCs has shown early promise

In such studies, DCs are isolated from a cancer patient, expanded in vitro, pulsed with antigenic peptides, whole proteins, or tumor lysates and infused back into the patient.

The most advanced randomized phase III clinical trials with this approach have been carried out in patients with prostate cancer who have received autologous DCs pulsed with prostatic acid phosphatase. Patients in the control arm received a placebo. Patients who received acid phosphatase pulsed DCs survived measurably longer than patients who received placebo. This trial may form the basis of approval of this treatment for patients with prostate cancer.

Inhibition of downregulation of immune modulation may be clinically valuable

The role of downregulation of tumor immunity in progressive tumor growth in mouse models has been discussed earlier. Inhibition of such downregulation is an

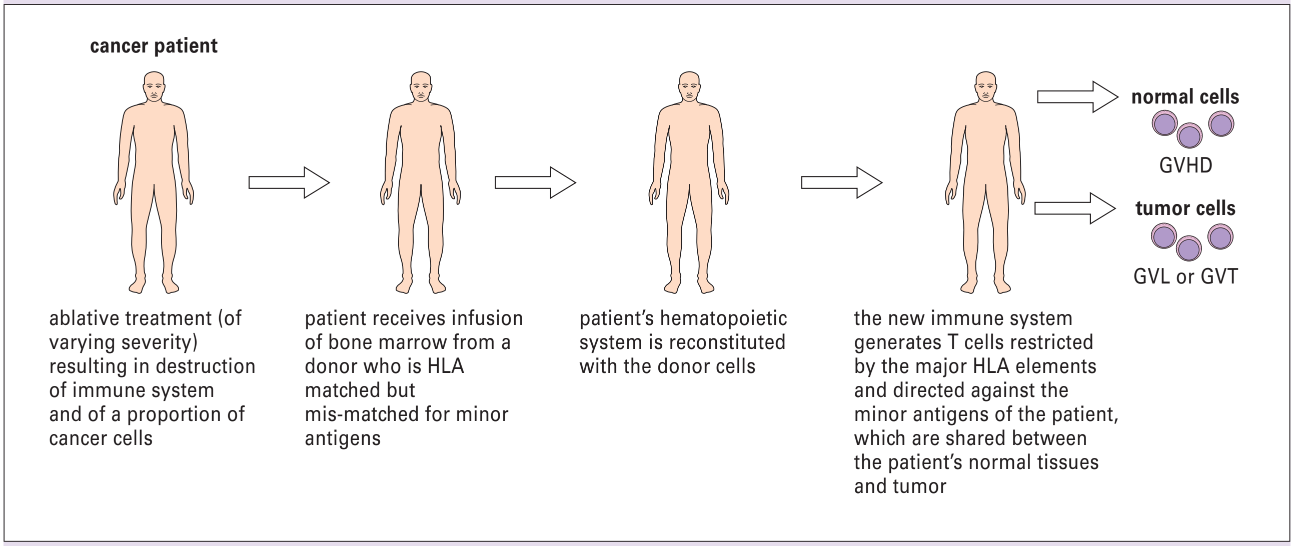
Graft versus host disease and graft versus tumor

cancer patient
normal cells
GVHD
tumor cells
GVL or GVT
ablative treatment (of varying severity) resulting in destruction of immune system and of a proportion of cancer cells
patient receives infusion of bone marrow from a donor who is HLA matched but mis-matched for minor antigens
patient's hematopoietic system is reconstituted with the donor cells
the new immune system generates T cells restricted by the major HLA elements and directed against the minor antigens of the patient, which are shared between the patient's normal tissues and tumor

Fig. 22.22 Graft versus host disease (GVHD) and graft versus leukemia (GVL) effects. Both responses are restricted by the MHC alleles and directed against the minor antigens. The minor antigens are the same between the patient's normal tissues and tumor, hence the GVHD (immune response against normal tissues) and GVL (immune response against leukemia or other tumor) are generally inextricably linked. (GVT, graft versus tumor)

attractive target of translation and antibodies to CTLA-4 are under clinical testing in this regard.

A number of phase I and II clinical trials using antibodies to CTLA-4 have been completed in patients with melanoma and ovarian and renal cancer.

The most dramatic results have been obtained in patients who had previously been vaccinated with autologous GM-CSF transfected tumor cells as part of another study and had not shown significant response to that vaccination. When these patients were treated with anti-CTLA-4 antibody, they showed dramatic tumor shrinkage and infiltration of the tumors with lymphocytes and granulocytes.

Interestingly, patients previously immunized with differentiation antigen vaccines and who received the same anti-CTLA-4 antibody did not show clinical responses.

The synergy between vaccination with GM-CSF transfected tumor cells and administration of anti-CTLA-4 antibody was also previously seen in a mouse model of melanoma. Further clinical exploration of this attractive strategy is now in progress.

Most successful treatment approaches are based on individually distinct tumor antigens

A perusal of the various approaches discussed above will show that some approaches are based on the premise that immunoprotective antigens of tumors are individually distinct, whereas others pursue the idea of immunoprotective antigens shared among individual tumors (see Fig. 22.19).

The most successful clinical data obtained thus far have been seen in trials with:

- autologous cancer cells plus BCG (colon cancers);
- idiotype vaccines (follicular B lymphoma);
- autologous tumor lysates (renal carcinoma);
- autologous tumor vaccines plus anti-CTLA-4 antibody (melanoma, ovarian cancers);
- acid phosphatase pulsed DCs; and
- gp96–peptide complexes from autologous tumors.

Of these, all approaches except the acid phosphatase pulsed DCs are based on individually distinct rather than shared immunoprotective tumor antigens. That bias is certainly consistent with the individually unique antigenicity seen in mouse tumors, as discussed at the beginning of this chapter.

FURTHER READING

Belli F, Testori A, Rivoltini L, et al. Vaccination of metastatic melanoma patients with autologous tumor-derived heat shock protein gp96–peptide complexes: clinical and immunologic findings. J Clin Oncol 2002;20:4169–4180.

Coulie PG, Karanikas V, Lurquin C, et al. Cytolytic T-cell responses of cancer patients vaccinated with a MAGE antigen. Immunol Rev 2002;188:33–42.

Egen JG, Kuhns MS, Allison JP. CTLA-4: new insights into its biological function and use in tumor immunotherapy. Nat Immunol 2002;3:611–618.

Fehervari Z, Sakaguchi S. CD4$^+$ Tregs and immune control. J Clin Invest 2004;114:1209–1217.

Ho WY, Blattman JN, Dossett ML, et al. Adoptive immunotherapy:

engineering T cell responses as biologic weapons for tumor mass destruction. Cancer Cell 2003;3:431–437.

Klein G. The strange road to the tumor-specific transplantation antigens (TSTAs). Cancer Immun 2001;1:6.

North RJ. Down-regulation of the antitumor immune response. Adv Cancer Res 1985;45:1–43.

Scanlan MJ, Gure AO, Jungbluth AA, et al. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev 2002;188:22–32.

Srivastava PK. Do human cancers express shared protective antigens? or the necessity of remembrance of things past. Semin Immunol 1996;8:295–302.

Srivastava PK. Interaction of heat shock proteins with peptides and antigen presenting cells: chaperoning of the innate and adaptive immune responses. Annu Rev Immunol 2002;20:395–425.

Van Der Bruggen P, Zhang Y, Chaux P, et al. Tumor-specific shared antigenic peptides recognized by human T cells. Immunol Rev 2002;188:51–64.

Wick M, Dubey P, Koeppen H, et al. Antigenic cancer cells grow progressively in immune hosts without evidence for T cell exhaustion or systemic anergy. J Exp Med 1997;186:229–238.

Critical thinking: Immunity to cancers (see p. 501 for explanations)

1. Animals can be easily immunized against cancers. Why then do the same cancers grow progressively and kill their hosts?
2. Heat-shock proteins do not differ between tumors and normal tissues. Yet, HSP preparations isolated from tumors elicit tumor immunity, whereas similar preparations from normal tissues do not. Explain the mechanistic basis of this phenomenon.
3. How have tumor antigens been defined? How many classes of tumor antigen have been defined thus far?
4. Explain the phenomenon of concomitant immunity.
5. Discuss the evidence that the immune response to tumors is downregulated in tumor-bearing mice and in patients with cancer.

This Page Intentionally Left Blank

Hypersensitivity

SECTION 5

This Page Intentionally Left Blank

CHAPTER

23

Immediate Hypersensitivity (Type I)

SUMMARY

- **Hypersensitivity reactions are based on the classification of Coombs and Gell.**
- **Historical observations have shaped our understanding of immediate hypersensitivity.** The severity of symptoms depends on IgE antibodies, the quantity of allergen, and also a variety of factors that can enhance the response including viral infections and environmental pollutants.
- **Most allergens are proteins.**
- **IgE is distinct from the other dimeric immunoglobulins.** Production of IgE in genetically predisposed (i.e. atopic) individuals occurs in response to repeated low-dose exposure to inhaled allergens such as dust mite, cat dander, or grass pollen.
- **Allergens are the antigens that give rise to immediate hypersensitivity and contribute to asthma.**
- **Mast cells and basophils contain histamine.** IgE antibodies bind to a specific receptor, FcεRI, on mast cells and basophils. When bound IgE is cross-linked by specific allergen, mediators including histamine, leukotrienes, and cytokines are released.
- **Multiple genes have been associated with asthma in different populations.** Multiple genetic loci influence the production of IgE, the inflammatory response to allergen exposure, and the response to treatment. Polymorphisms have been identified in the genes, in promoter regions, and in the receptors for IgE, cytokines, leukotrienes, and the β_2-receptors.
- **Skin tests are used for diagnosis and investigation.**
- **Several different pathways contribute to the chronic symptoms of allergy.**
- **Allergens contribute to asthma.**
- **Immunotherapy can be used for hayfever and anaphylactic sensitivity.**
- **New approaches are being investigated for treating allergic disease.**
- **IgE antibodies play a critical role in defense against helminths.** The biological role of immediate hypersensitivity is to control helminth infections such as schistosomiasis, hookworm, or ascariasis. However, it is likely to be a combination of effector TH2 cells, basophils, and eosinophils, as well as IgE antibodies on mast cells, that control these worms.

HYPERSENSITIVITY REACTIONS ARE BASED ON THE CLASSIFICATION OF COOMBS AND GELL

The adaptive immune response provides specific protection against infection with bacteria, viruses, parasites, and fungi. In particular, it is able to provide rapid protection against a repeated challenge with the same or a similar foreign organism or toxin. Some immune responses, however, give rise to an excessive or inappropriate reaction – this is usually referred to as **hypersensitivity**.

The term hypersensitivity evolved from the observations of Richet and Portier one hundred years ago, who described the catastrophic result of exposing a presensitized animal to systemic antigen. The resulting outcome, termed **anaphylaxis**, became the prototype of immediate hypersensitivity responses.

Coombs and Gell in 1963 proposed a classification scheme in which allergic hypersensitivity of the type described by Portier and Richet was termed type I, and broadened the definition of hypersensitivity to include:

- reactions that resulted from antibody–antigen reactions, including cytotoxic reactions which they termed type II (see Chapter 24);
- immune complex reactions of the Arthus type, which were termed type III reactions (see Chapter 25);
- delayed-type hypersensitivity responses were given the classification type IV to describe reactions typified by cutaneous contact reactions upon exposure to certain agents such as nickel (see Chapter 26).

The original Coombs and Gell classification is shown in Fig. 23.1.

The Coombs and Gell classification of the four types of hypersensitivity reaction

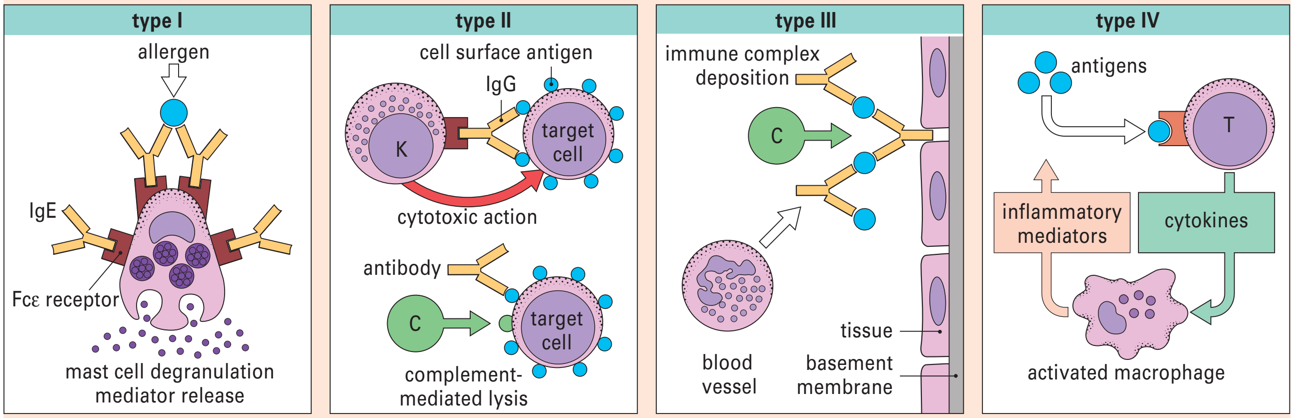

Fig. 23.1 In type I, mast cells bind IgE via their Fc receptors. On encountering allergen the IgE becomes cross-linked, inducing degranulation and release of mediators that produce allergic reactions. In type II, antibody is directed against antigen on an individual's own cells (target cell) or foreign antigen, such as transfused red blood cells. This may lead to cytotoxic action by K cells, or complement-mediated lysis. In type III, immune complexes are deposited in the tissue. Complement is activated and polymorphs are attracted to the site of deposition, causing local tissue damage and inflammation. In type IV, antigen-sensitized T cells release lymphokines following a secondary contact with the same antigen. Cytokines induce inflammatory reactions and activate and attract macrophages, which release inflammatory mediators.

In the past several years it has become apparent that the Combs and Gell classification artificially divided mechanistically related antibody reactions (such as types I, II, and III), which contribute to the pathophysiology of many common immune-mediated diseases, while including the T cell-mediated reactions of delayed-type hypersensitivity (DTH) in a common classification (termed type IV).

Based on our current understanding of the underlying pathways of inflammation triggered by antigen exposure and the disease conditions observed, common mechanisms appear to operate in types I, II, and III hypersensitivity. These common mechanisms involve the engagement by antibody–antigen complexes with cellular receptors for the Fc region of antibodies (termed **Fc receptors**).

Understanding the mechanisms underlying these categories has resulted in the development of novel therapeutics based on the common pathways activated by these reactions and may result in the broader applicability of such therapeutics to multiple disease states.

Immediate hypersensitivity responses:

- have traditionally been referred to as type I hypersensitivity; and
- are characterized by the production of IgE antibodies against foreign proteins that are commonly present in the environment (e.g. pollens, animal danders, or house dust mites).

Antibody-mediated reactions:

- were previously referred to as type II;
- occur when IgG or IgM antibodies are produced against surface antigens on cells of the body.

These antibodies can trigger either by activating complement (e.g. autoimmune hemolytic anemia) or by facilitating the binding of natural killer cells (NK).

Type III hypersensitivity:

- involves the formation of immune complexes in the circulation that are not adequately cleared by macrophages or other cells of the reticuloendothelial system.

The formation of immune complexes requires significant quantities of antibody and antigen (typically microgram quantities of each). The classical diseases in which immune complexes are thought to be involved are systemic lupus erythematosus (SLE), chronic glomerulonephritis, and serum sickness.

Cell-mediated reactions:

- were previously referred to as type IV;
- are those in which specific T cells are the primary effector cells.

Examples of T cells causing unwanted responses are:

- contact sensitivity (e.g. to nickel or plants such as poison ivy);
- graft rejection;
- the hypersensitivity skin responses of leprosy or tuberculosis;
- the exaggerated response to viral infections such as measles; and
- the persistent symptoms of allergic disease.

HISTORICAL OBSERVATIONS HAVE SHAPED OUR UNDERSTANDING OF IMMEDIATE HYPERSENSITIVITY

The classical allergic disease is seasonal hayfever caused by pollen grains entering the nose (rhinitis) and eyes (conjunctivitis). In severe cases patients also get seasonal asthma and seasonal dermatitis.

Charles Blackley in 1873 demonstrated that pollen grains placed into the nose could induce symptoms of rhinitis. He also demonstrated that pollen extract could produce a wheal and flare skin response in patients with hayfever.

The **wheal and flare skin response** is an extremely sensitive method of detecting specific IgE antibodies. The timing and form of the skin response is indistinguishable from the local reaction to injected histamine. Furthermore, the immediate skin response can be effectively blocked with antihistamines.

In 1903 Portier and Richet discovered that immunization of guinea pigs with a toxin from the jellyfish *Physalia* could sensitize them so that a subsequent injection of the same protein would cause rapid onset of breathing difficulty, influx of fluid into the lungs, and death. They coined the term **anaphylaxis** (from the Greek *ana*, non, and *phylaxos*, protection) and speculated about the relationship to other hypersensitivity diseases. They noted that:

- human anaphylaxis had no familial characteristics (unlike most of the other allergic diseases); and
- natural exposure to inhaled allergens did not cause anaphylaxis or urticaria.

Subsequently, it became clear that injection of any protein into an individual with immediate hypersensitivity to that protein can induce anaphylaxis. Thus, anaphylaxis occurs when a patient with immediate hypersensitivity is exposed to a relevant allergen in such a way that antigen enters the circulation rapidly.

Q. In what circumstances can large amounts of allergen enter the circulation rapidly?

A. Following direct injection of the antigen into the tissue, such as a bee sting, a therapeutic injection for hyposensitization, or injection of a drug.

Anaphylaxis may also occur as a result of eating an allergen such as peanut or shellfish, or following the rupture of hydatid cysts with the rapid release of parasite antigens (Fig. 23.2).

The term **allergen** was first used by von Pirquet to cover all foreign substances that could produce an immune response. Subsequently, the word 'allergen' came to be used selectively for the proteins that cause 'supersensitivity'. Thus, **an allergen is an antigen that gives rise to immediate hypersensitivity**.

Anaphylaxis and urticaria

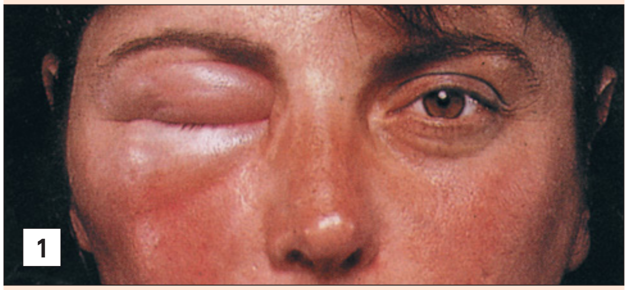

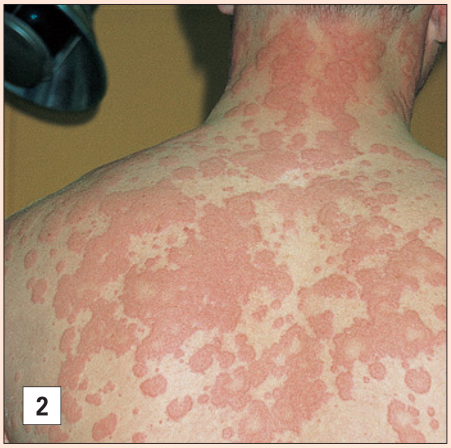

Fig. 23.2 (**1**) The anaphylactic response to bee venom in a patient who has IgE antibodies to the venom protein, phospholipase A. The immediate reaction occurs within 20 minutes and is mediated by the release of histamine and other mediators from mast cells. The patient shown had been stung on the face, but the reaction can become generalized, leading to a fall in blood pressure, generalized urticaria, and/or bronchospasm (i.e. anaphylaxis). (**2**) Diffuse urticaria on a patient with severe chronic urticaria. The lesions have a raised edge and come up within minutes or hours. The lesions almost always resolve within 12 hours leaving no trace on the skin.

MOST ALLERGENS ARE PROTEINS

Substances that can give rise to wheal and flare responses in the skin and to the symptoms of allergic disease are derived from many different sources (see www.allergen.org). When purified they are almost all found to be proteins and their size ranges in molecular weight from 10 000–40 000 Da. These proteins are all freely soluble in aqueous solution, but have many different biological functions including digestive enzymes, carrier proteins, calycins, and pollen recognition proteins.

Any allergen can be described or classified by its source, route of exposure, and nature of the specific protein (Fig. 23.3).

Extracts used for skin testing or in-vitro measurement of IgE antibodies are made from the whole material, which contains multiple different proteins, any of which can be an allergen. Indeed, it is clear that individual patients can react selectively to one or more different proteins within an extract.

Q. Name one genetic factor that determines whether an individual can make an immune response to a specific allergen.

A. The individual's MHC haplotypes determine which antigens and antigen fragments are presented.

Estimates of exposure can be made either by visual identification of particles (e.g. pollen grains or fungal spores) or by immunoassay of the major allergens (e.g. Fel d 1 or Der p 1).

Properties of allergens

source	airborne particles	dimension of airborne particle (μm)	allergen		
			name	MW (kDa)	function/homologies
dust mite – *Dermatophagoides pteronyssinus*	feces	10–40	Der p 1 Der p 2	25 13	cysteine protease (epididymal protein)
cats – *Felis domesticus*	dander particles	2–15	Fel d 1	36	uteroglobin
German cockroach – *Blattella germanica*	frass saliva	≥ 5	Bla g 2 Bla g 4 Bla g 5	36 21 23	aspartic protease calycin glutathione-*S*-transferase
rat – *Rattus norvegicus*	urine on bedding ?	2–20	Rat n 1	19	pheromone binding protein
grass	pollen	30	Lol p 1	29	not known
fungi – *Alternaria alternata, Aspergillus fumigatus*	spores spores	14 × 10 2	Alt a 1 Asp f 1	28 18	not known mitogillin

Fig. 23.3 Patients who become 'allergic' to one of the well-recognized sources of allergens have actually produced an IgE antibody response to one or more of the proteins produced by mites, trees, grass, cats, or fungi. The proteins are predominantly water soluble with a molecular weight (MW) ranging from 10 000 to 40 000 kDa. In many cases the function of the proteins is known, but it is not clear whether function such as enzymic activity alters the ability of these proteins to induce an allergic response. The properties of the particles carrying these allergens are very important because they influence both how much becomes airborne, and also where the allergen is deposited in the respiratory tract. The dimensions of the particles airborne vary from ≤ 2 μm for *Aspergillus or Penicillium* spores to ≥ 20 μm for mite fecal pellets and some pollen grains. (Sizes are given as diameter in μm)

IgE IS DISTINCT FROM THE OTHER DIMERIC IMMUNOGLOBULINS

In 1921 Küstner, who was allergic to fish, injected his own serum into the skin of Prausnitz, who was allergic to grass pollen but not fish, and demonstrated that it was possible to passively transfer immediate hypersensitivity (the Prausnitz–Küstner or P–K test). Prausnitz would also notice that an immediate wheal and flare occurred at the site of passive sensitization when he ate fish. This showed that intact fish allergen can be absorbed into the circulation.

Over the next 30 years it was established that P–K activity was a general property of the serum of patients with immediate hypersensitivity and that it was allergen specific (i.e. behaved like an antibody).

In 1967 Ishizaka and his colleagues purified the P–K activity from a patient with ragweed hayfever and proved that this was a novel isotype of immunoglobulin – IgE. However, it was obvious that the concentration of this immunoglobulin isotype in serum was very low.

The initial antisera to IgE made it possible to identify a patient with multiple myeloma whose serum contained a very high concentration of IgE (~10 mg/ml). Purification of this myeloma protein by Johansson and Bennich led to the full structure of IgE and also to the production of potent antisera.

Antisera to IgE are used in the radioallergosorbent test (RAST) to measure IgE antibodies in serum, as well as for measuring total serum IgE.

IgE is distinct from the other dimeric immunoglobulins (see Fig. 3.8) because it has:

- an extra constant region domain;
- a different structure to the hinge region; and
- binding sites for both **high-** and **low-affinity IgE receptors**, **FcεRI** and **FcεRII**, respectively (Fig. 23.4).

The primary cells that bear FcεRI are **mast cells** and **basophils**, which are the only cells in the human that contain significant amounts of histamine.

Low-affinity receptors for IgE – FcεRII or CD23 – are also present on B cells and may play a role in antigen presentation.

In addition in atopic dermatitis dendritic cells in skin can express a high-affinity receptor for IgE, but this receptor lacks the β chain of FcεRI.

The properties of IgE can be separated into three areas:

- the characteristics of the molecule including its half-life and binding to IgE receptors;
- the control of IgE and IgG antibody production by T cells; and
- the consequences of allergen cross-linking IgE on the surface of mast cells or basophils.

IgE molecules

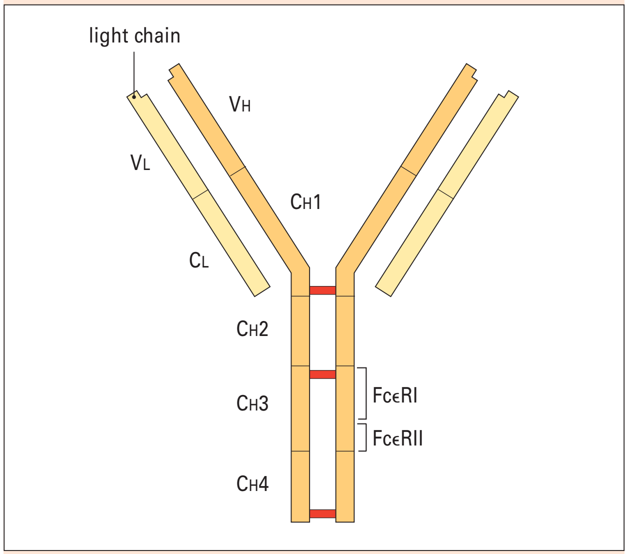

Fig. 23.4 The IgE molecule has four heavy chain constant regions. The binding sites for the high-affinity IgE receptor FcεRI and for the low-affinity receptor FcεRII, or CD23, are shown. Monoclonal antibodies to the binding site for FcεRI also block FcεRII.

The half-life of IgE is short compared with that of other immunoglobulins

The concentration of IgE in the serum of normal individuals is very low compared to all the other immunoglobulin isotypes. Values range from < 10 to 10 000 IU/ml, and the international unit (IU) is equivalent to 2.4 ng. Most sera contain < 400 IU/ml (i.e. < 1 µg/ml). The reasons why serum IgE is so low include:

- serum IgE has a much shorter half-life than other isotypes (≈ 2 days compared with 21–23 days for IgG);
- IgE is produced in small quantities and is only produced in response to a select group of antigens (allergens and parasites); and
- IgE antibodies are sequestered on the high-affinity receptor on mast cells and basophils.

Q. What fundamental reason explains the low production of IgE?

A. The class switch from IgM to IgE happens infrequently and is controlled by T cells. The position of the IgE constant region gene is towards the distal end of the Ig gene stack (see Fig. 8.19), but the position alone cannot explain the infrequency of the switch to IgE.

The half-life of IgE in the serum has been measured both by injecting radiolabeled IgE and by infusing plasma from allergic patients into normal and immune-deficient patients.

The half-life of IgE in serum is less than 2 days; by contrast, IgE bound to mast cells in the skin has a half-life of approximately 10 days.

The low quantities of IgE in the serum must reflect a more rapid breakdown of IgE, as well as removal from the circulation by binding onto mast cells.

The most important site of breakdown of IgE is thought to be within **endosomes** where the low pH facilitates breakdown of free immunoglobulin by cathepsin.

Serum is constantly being taken up by endocytosis. Most macromolecules including IgE degrade in the endosome. One major exception is IgG, which is protected by binding to the neonatal Fc gamma receptor, FcγRn (Fig. 23.5).

IgG4 is transferred across the placenta, but IgE is not

In cord blood the concentration of IgE is very low indeed, generally < 1 IU/ml (i.e. < 2 ng/ml). Thus, there appears to be almost no transfer across the placenta.

Endocytosis of plasma

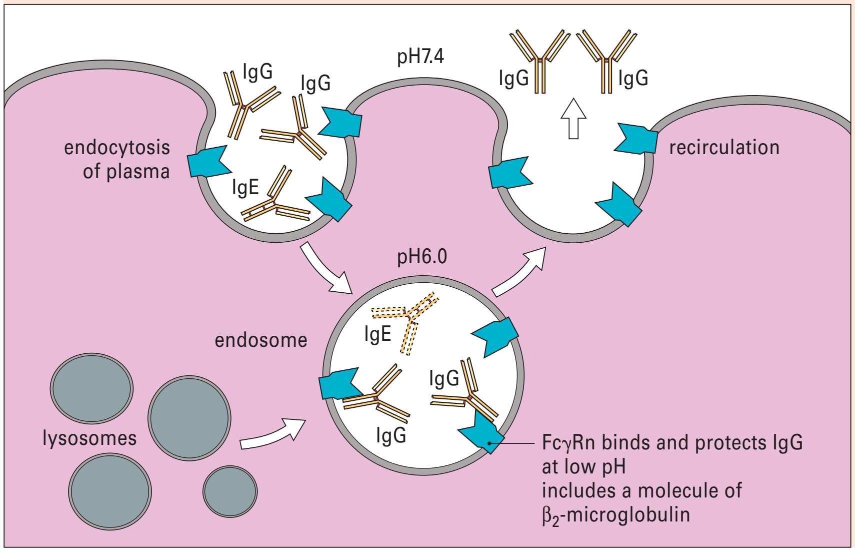

Fig. 23.5 Endocytosis of plasma contributes to the short half-life of IgE as plasma proteins are taken up and the pH falls because of lysosomes combining with the endosome. At low pH IgG including IgG4 molecules bind to the neonatal Fc γ receptor (FcγRn). By contrast, IgE molecules do not bind to FcγRn so are not protected and are digested by cathepsin. As the endosomes recirculate, the pH rises to 7.4 and the undamaged IgG molecules are released into the circulation. The FcγRn includes a molecule of β_2-microglobulin. In keeping with this model the half-life of IgG is shorter than normal in mice that have had the gene for β_2-microglobulin removed (or knocked out).

By contrast, IgG, including IgG4 antibodies to allergens such as those from dust mite or cat, are very efficiently transferred across the placenta. This process also involves endocytosis and receptor-mediated transport.

Passive transfer of IgE to the fetus may be blocked because IgE is broken down in the endosomes, or because an Fc receptor that is essential for transport is absent on the cells that comprise the placental tissues. In prenatal transfer IgG is protected in endosomes by binding to FcγRn. Although IgE can be seen on fetal villous stroma, this is not through FcεRI or FcγRn, but may be through FcγRI on fetal macrophages.

ROLE OF T CELLS IN THE IMMUNE RESPONSE TO INHALANT ALLERGENS

IgE production is dependent on TH2 cells

Experiments in animals have established that the production of IgE is completely dependent on T cells. It is also clear that T cells can suppress IgE production.

T cells that suppress IgE production:

- act predominantly by producing interferon-γ (IFNγ); and
- are produced when the animal (e.g. mouse, rat, or rabbit) is primed in the presence of Freund's complete adjuvant.

This adjuvant, which includes bacterial cell walls and probably bacterial DNA, is a very potent activator of macrophages.

With the discovery of TH1 and TH2 cells, it became clear that IgE production is dependent on TH2 cells and that any priming that generates a TH1 response will inhibit IgE production.

The main cytokines that are specifically relevant to a TH1 response include:

- interleukin-12 (IL-12) produced by macrophages; and
- IFNγ produced by T cells.

By contrast, the primary cytokines relevant to a TH2 response are:

- IL-4 (IL13);
- IL-5; and
- IL-10 (Fig. 23.6).

It is clear from experiments in mice and humans that the expression of the gene for IgE is dependent on IL-4. Thus, if immature human B cells are cultured with anti-CD40 and IL-4, they will produce IgE antibodies.

Cytokines regulate the production of IgE

In humans IgE antibodies are the dominant feature of the response to a select group of antigens and most other immune responses do not include IgE.

T cell differentiation during human immune responses

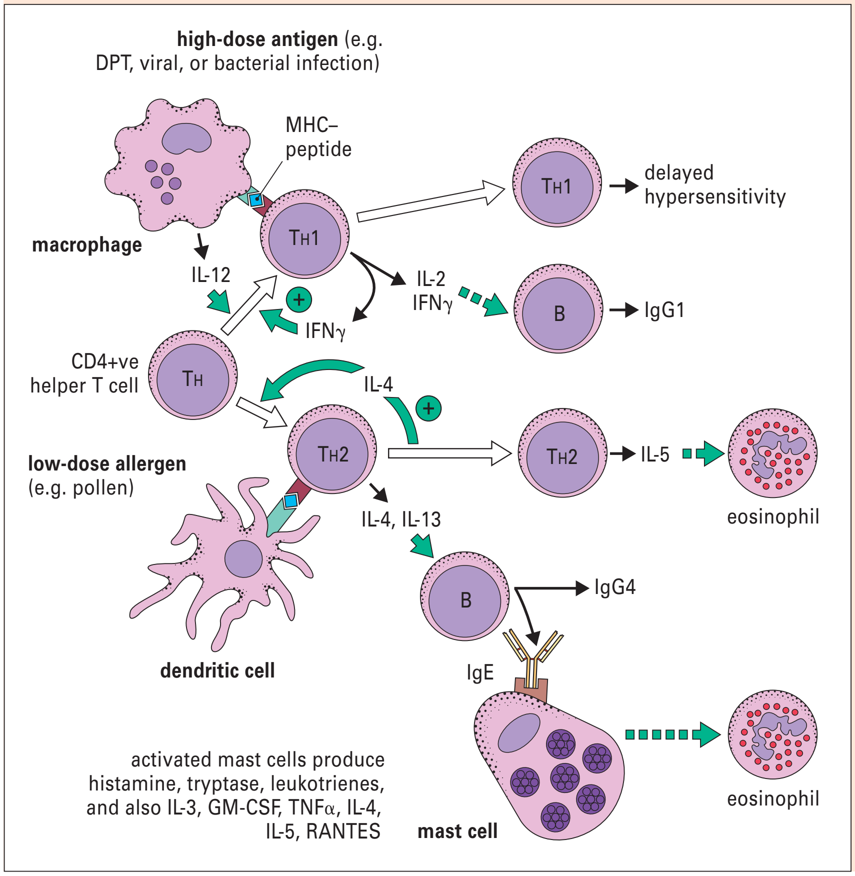
high-dose antigen (e.g. DPT, viral, or bacterial infection)
MHC–peptide
macrophage
TH1
TH1
delayed hypersensitivity
IL-12
+
IFNγ
IL-2
IFNγ
B
IgG1
CD4+ve helper T cell
TH
IL-4
+
low-dose allergen (e.g. pollen)
TH2
TH2
IL-5
eosinophil
IL-4, IL-13
dendritic cell
B
IgG4
IgE
activated mast cells produce histamine, tryptase, leukotrienes, and also IL-3, GM-CSF, TNFα, IL-4, IL-5, RANTES
mast cell
eosinophil

Fig. 23.6 The differentiation of TH cells depends on the antigen source, the quantity of allergen, and the cytokines produced. Bacterial antigens or a high dose of antigen will induce IL-12 from macrophages. In addition, the developing TH1 cells produce IFNγ, which further enhances the production of TH1 cells. Low-dose antigen without adjuvant will induce TH2 cells, which produce both IL-4 and IL-5. IL-4 plays a role in (i) enhancing the growth of TH2 cells; (ii) the expression of the gene for IgE. In turn IgE binds to the high-affinity receptor for IgE (FcεRI) on mast cells.

The classical allergens are inhaled in very small quantities (5–20 ng/day) either perennially indoors or over a period of weeks or months outdoors. Immunization of mice with repeated low-dose antigen is a very effective method of inducing IgE responses.

By contrast, the routine immunization of children with diphtheria and tetanus toxoid does not induce persistent production of IgE antibodies. This is clear because we do not routinely take precautions against anaphylaxis when administering a booster injection of tetanus.

Q. Which cytokines promote the production of IgE, and which inhibit production?

A. TH2 cytokines, including IL-4, promote production of IgE, whereas TH1 cytokines (e.g. IL-12) inhibit production.

As T cells differentiate, TH1 cells express the functional IL-12 receptor with the IL-12 β2 chain. By contrast, TH2 cells express only part of the IL-12 receptor and this part is non-functional.

IL-4 is important in the differentiation of TH2 cells and is also a growth factor for these cells. Because it is produced by TH2 cells, it is at least in part acting on the cell that produced it (i.e. in an autocrine fashion). The interaction of IL-4 with T cells can be blocked either with:

- an antibody to IL-4; or
- a soluble form of the IL-4 receptor (IL-4R).

The release of soluble IL-4R from T cells may be a natural mechanism for controlling T cell differentiation.

It follows that inhaling recombinant soluble IL-4R is a potential therapeutic strategy to control allergic responses in the lung. However, recent evidence suggests that in-vivo responses are controlled by T cells producing either IL-10 or transforming growth factor-β (TGFβ).

Both IgE and IgG4 are dependent on IL-4

The genes for immunoglobulin heavy chains are in sequence on chromosome 14. The gene for ε occurs directly following the gene for γ4. Both of these isotypes are dependent on IL-4 and they may be expressed sequentially (Fig. 23.7).

The mechanisms by which IgG4 is controlled separately from IgE are not well understood, but this may include a role for IL-10. Thus, immunotherapy for patients with anaphylactic sensitivity to honey bee venom will induce IL-10 production by T cells, decreased IgE, and increased IgG4 antibodies to venom antigens.

Recently, it has been shown that children raised in a house with a cat can produce an IgG response, including IgG4 antibody, without becoming allergic. A modified TH2 response (increased IgG4 and decreased IgE) therefore represents an important mechanism of tolerance to allergens (Fig. 23.8). IgG4 antibody responses without IgE antibody are a feature of immunity/tolerance to insect venom, rat urinary allergens, and food antigens as well as cat allergens.

ALLERGENS ARE THE ANTIGENS THAT GIVE RISE TO IMMEDIATE HYPERSENSITIVITY

Allergens have similar physical properties

In mice a wide range of proteins can be used to induce an IgE antibody response. The primary factors that influence the response are:

- the strain of mouse;
- the dose; and
- adjuvants used.

Thus, repeated low-dose immunization with alum or pertussis (but not complete Freund's adjuvant) will produce IgE responses. However, the dose necessary to induce a optimal response varies greatly from one strain to another.

The allergens that have been defined have similar physical properties (i.e. freely soluble in aqueous solution with a molecular weight between 10 000 and 40 000 Da), but are very diverse biologically.

Cloning has revealed sequence homology between allergens and diverse proteins including calycins, pheromone binding proteins, enzymes, and pollen recognition proteins.

Although many of the allergens have homology with known enzymes, this is not surprising because enzymic activity is an important property of proteins in general.

Some important allergens, for example Der p 2 from mites, Fel d 1 from cats, and Amb a 5 from ragweed pollen, have neither enzymic activity nor homology with known enzymes. Thus, enzymic activity is not essential for immunogenicity.

Nevertheless, the group I allergens of dust mites are cysteine proteases and in several model situations it has been shown that this enzymic activity influences the immunogenicity of the protein. Thus cleavage of CD23 or CD25 on lymphocytes by Der p 1 can enhance immune responses.

Chromosome 14

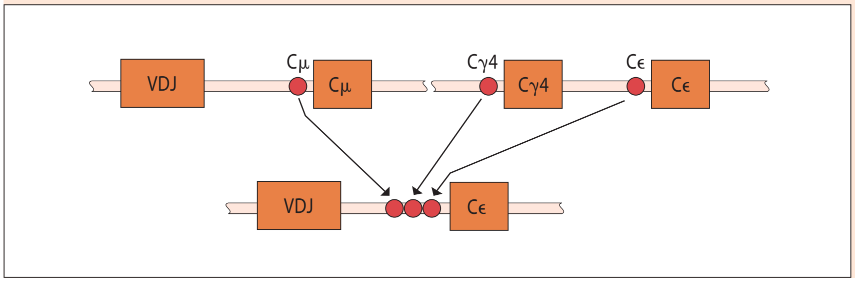

Fig. 23.7 Switch regions and heavy chain genes for immunoglobulin are arranged sequentially on chromosome 14. Both Cγ4 and Cε expression are dependent on IL-4 produced by T cells. The switch region of IgE often includes elements from Sγ4 indicating that the switching occurs sequentially. However, IgG4 responses can occur without IgE antibody responses.

Modified TH2 response

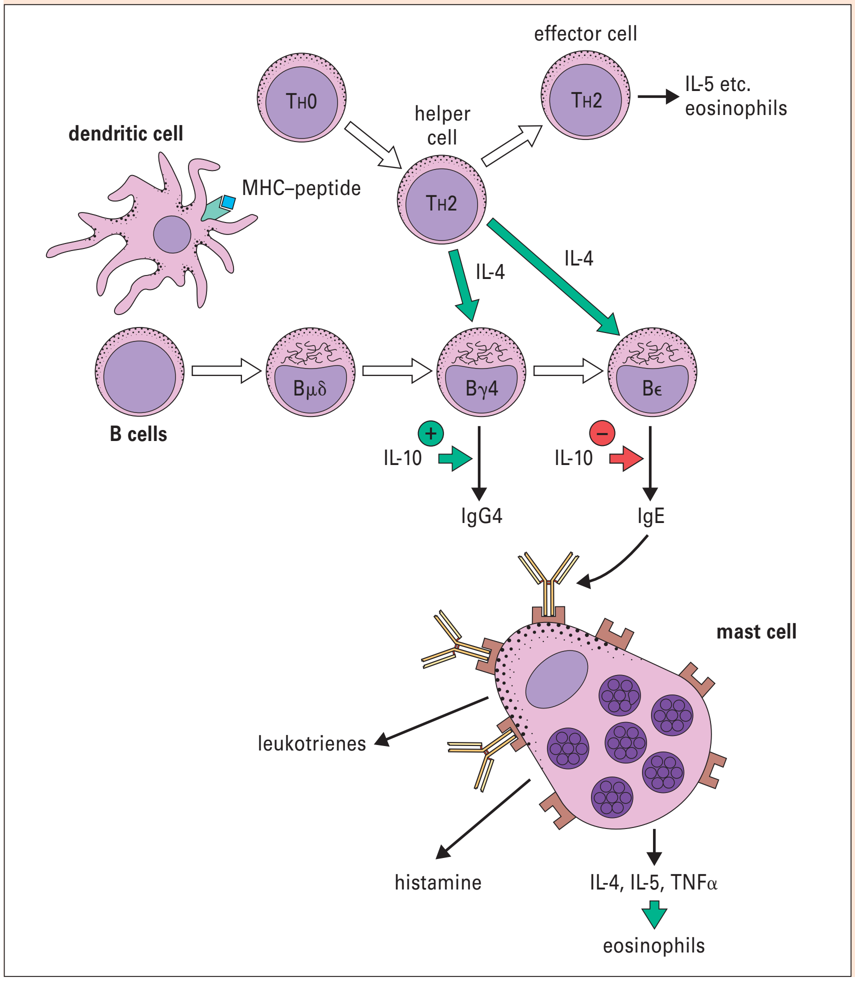

Fig. 23.8 The TH2 response includes T effector cells as well as help for IgE and IgG4 antibody production. In turn IgE plays a major role in triggering mast cells. However, increasing evidence shows that higher doses of allergen (e.g. bee venom, cat dander, or rat urine) can induce a modified or tolerant TH2 response. This response includes IgG4 antibodies, but not IgE. The cytokine IL-10 may well play a role in enhancing IgG4 antibody while suppressing production of IgE.

Alternatively, it has been shown that Der p 1 can disrupt epithelial junctions and alter the entry of proteins through the epithelial layer. The interest in this property is increased because many different mite allergens are inhaled together in the fecal particle so the enzymic activity of one protein (i.e. Der p 1) could facilitate either the physical entry or the immune response to other mite proteins. However, the lungs contain many different naturally occurring proteinases (as well as anti-proteases), which are just as potent as these allergens.

The primary characterization of allergens relates to their route of exposure. The routes includes:

- inhaled allergens;
- foods;
- drugs;
- antigens from fungi growing on the body (e.g. *Aspergillus* spp.); and
- venoms.

The routes are important because they define the ways in which the antigens are presented to the immune system. Because antigen presentation may well be the site at which genetic influences play the biggest role, the properties of the different groups of allergen need to be considered separately.

The inhalant allergens cause hayfever, chronic rhinitis, and asthma

The inhalant allergens are:

- the primary causal agents in hayfever, chronic rhinitis, and asthma among school-aged children and young adults;
- play an important role in atopic dermatitis.

Almost all evidence about the genetics of allergic disease relates to inhalant allergens.

Allergens can only become airborne in sufficient quantity to cause an immune response or symptoms when they are carried on particles. Pollen grains, mite fecal particles, particles of fungal hyphae or spores, and animal skin flakes (or dander) are the best defined forms in which allergens are inhaled (Fig. 23.9).

In each case it is possible to define the approximate particle size and the quantity of protein on the particle as well as the speed with which the proteins in the particle dissolve in aqueous solution (see Fig. 23.3).

Thus, for grass pollen, mite fecal pellets, and cat dander:

- the relevant allergens are present in high concentrations (up to 10 mg/cm^3);
- the particles are ‘large’ (i.e. 3–30 μm diameter); and
- the allergens elute rapidly in aqueous solution.

Particles carrying airborne allergens – mite fecal pellets and pollen grains

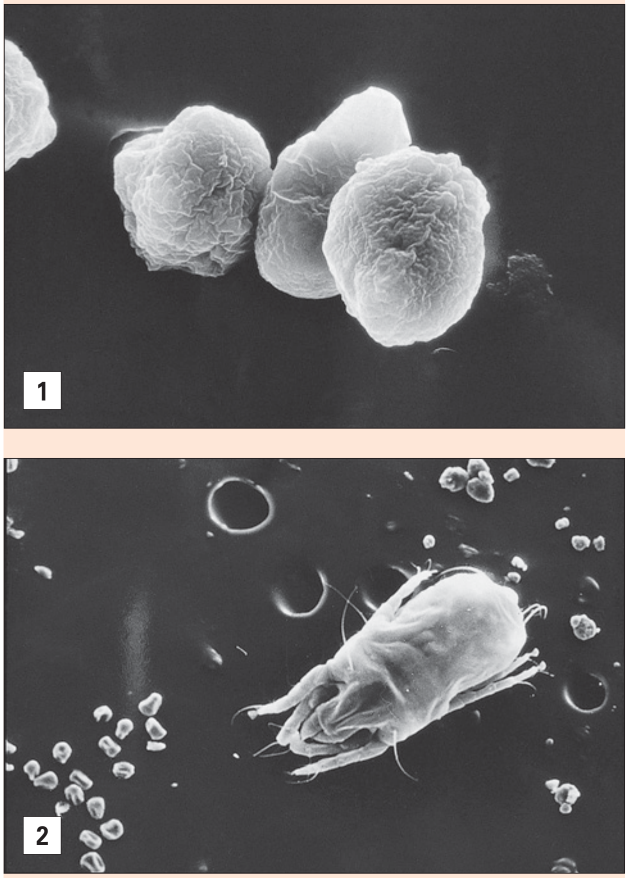

Fig. 23.9 The dust mite is the most important source of allergen in house dust, largely as fecal particles (**1**). A mite is shown in (**2**) with pollen grains lower left and fecal particles upper right. The mite is approximately 300 μm in length (i.e. just visible but not small enough to become airborne). Mite fecal particles are approximately 10–40 μm in diameter and become airborne during domestic disturbance. Pollen grains are similar in size to mite fecal particles (i.e. approximately 30 μm in diameter). The important allergic sources of pollen (i.e. grass, ragweed, and trees) are wind pollinated and the grains are designed to travel in the air for long distances.

The allergens within these particles will be delivered to the nasal epithelium because a large proportion of particles of this size will impact on the mucous membrane during passage of inhaled air through the nose.

Small quantities of inhalant allergen cause immediate hypersensitivity

Estimates of the quantity of mite or pollen-derived proteins inhaled vary from 5 to 50 ng/day. Thus exposure to some allergens may be as little as 1 μg/year. This is very important because it probably explains:

- why the immune response is consistently of this one kind (i.e. immediate hypersensitivity); and
- why no respiratory diseases, other than asthma, have been associated with these allergens.

The quantities inhaled also seriously restrict the models about how allergens contribute to asthma. Inhaling between 10 and 100 particles per day will produce localized areas of inflammation in the lungs, but would not be expected to induce acute bronchospasm. Inhaling a small number of ‘large’ particles (10–30 μm in diameter) could produce local areas of inflammation but could not give rise to alveolitis, acute bronchospasm, or progressive lung fibrosis. Equally, the quantities inhaled severely restrict the quantity that enters the blood stream and make it extremely unlikely that the fetus is sensitized or primed to inhalant allergens in utero.

Only a small number of food proteins are common causes of allergic responses

Although many food proteins can occasionally give rise to IgE responses, only a small number are common causes of allergic responses. These include egg, milk, wheat, soy, tree nuts, peanut, fish, and shellfish. In contrast to inhaled allergens, these proteins are often eaten in very large quantities (i.e. ≈10–100 g/day).

In general only a small fraction of the food proteins are absorbed. However, small peptides can be freely absorbed and may be recognized by T cells and even by IgE antibodies in a minority of individuals. Nevertheless, the bulk of the allergic and anaphylactic responses to foods are thought to be related to food proteins that have not been digested, either triggering mast cells in the intestine or entering the circulation.

Q. Recent evidence from Dr Sampson and his colleagues has shown that some children produce IgE antibodies against linear epitopes on food allergens. What implication does this have for induction of allergy in cooked food?

A. Denaturation of a protein by heat will not destroy linear epitopes, consequently cooking will not alter the allergenicity.

IgE binding sites can be identified on the tertiary structure of allergens

Many different allergens have been cloned, and for a few the tertiary structure is now known either from X-ray crystallography (e.g. Bet v 1), by nuclear magnetic resonance (e.g. Der p 2), or by modeling relationships to known homologs (Fig. 23.10).

Knowledge of the tertiary structure makes it possible to predict surface residues and to define IgE binding sites using site-directed mutagenesis. This approach provides the potential to design molecules that have decreased IgE binding properties, but with preserved T cell epitopes.

Given the importance of T cells to the control of IgE antibody production and their potential role in the recruitment of inflammatory cells, it is logical to try to use molecules that will directly ‘desensitize’ T cells. The approaches used include:

- allergen molecules modified by glutaraldehyde (allergoids);
- site-directed mutagenesis;
- allergens combined with two to four molecules of CpG;
- peptides.

Tertiary structure of two important allergens – Der p 2 and Bla g 2

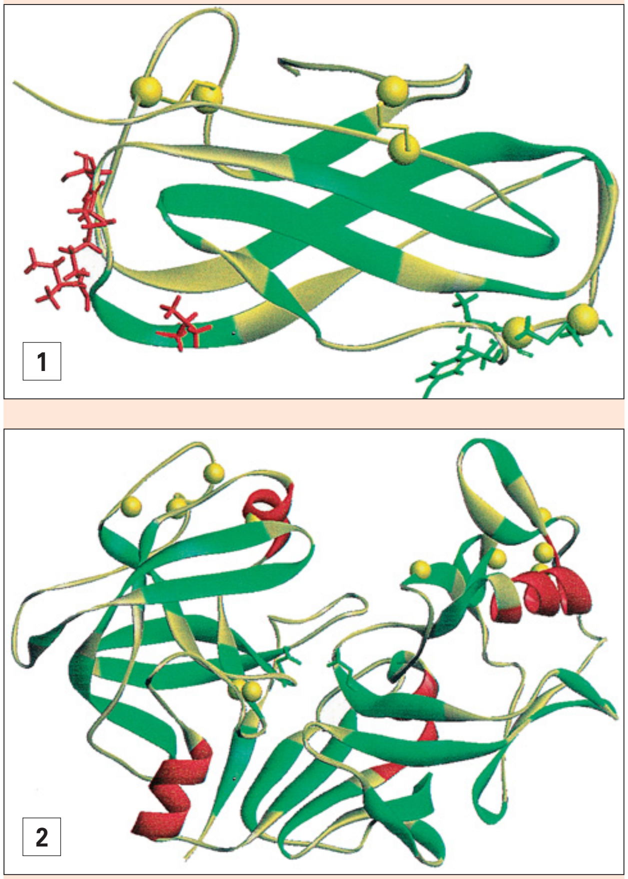

Fig. 23.10 The tertiary structure of the dust mite allergen, Der p 2 (**1**), was derived by nuclear magnetic resonance. This protein has no known function, and the structure shows no enzymic sites. On this structure the amino acids that are part of the binding sites for two monoclonal antibodies are shown (DPX green and 7A1 red). Knowledge of the structure makes it possible to alter the antibody binding sites using site-directed mutagenesis. (Courtesy of Drs Alisa Smith and Geoffrey Mueller) (**2**) The tertiary structure of a cockroach allergen, Bla g 2, shows the enzymic cleft and three residues that are considered to form the active site of an aspartic protease. However, the molecule has no enzymic activity due to other changes in the primary amino acid sequence. Some allergens (e.g. Der p 1 and Bla g 5) are active enzymes, but it remains an open question whether this activity contributes to their activity as allergens. (Courtesy of Drs A Pomes and M Chapman) On each structure the β sheets are shown as green bands, the α-helical structures in Bla g 2 are shown in red. In addition, cysteines are shown as yellow spheres.

The approach with peptides has been to test overlapping peptides of different lengths of between 12 and 35 amino acids.

Therapeutic trials have been carried out with peptides from ragweed pollen antigens and the cat allergen Fel d 1. The results show that peptide recognition is restricted by the HLA-DR type of the patient, which means that a wide range of peptides are necessary for treatment. In addition, there is clear evidence that peptides can produce a significant response in the lungs (Fig. 23.11). This is the clearest evidence yet that T cells in the lung can contribute to an asthmatic response.

MAST CELLS AND BASOPHILS CONTAIN HISTAMINE

The only human cell types that contain histamine are mast cells and basophils. In addition, these are the only cells that express the high-affinity receptor for IgE (FcεRI) under resting conditions.

Under most circumstances the primary and most rapid consequence of allergen exposure in an allergic individual is cross-linking of IgE receptors on mast cells and basophils:

- basophils are circulating polymorphs that are not present in normal tissue, but can be recruited to a local site by cytokines released from either T cells or mast cells;

Late asthmatic response to peptides from cat allergen

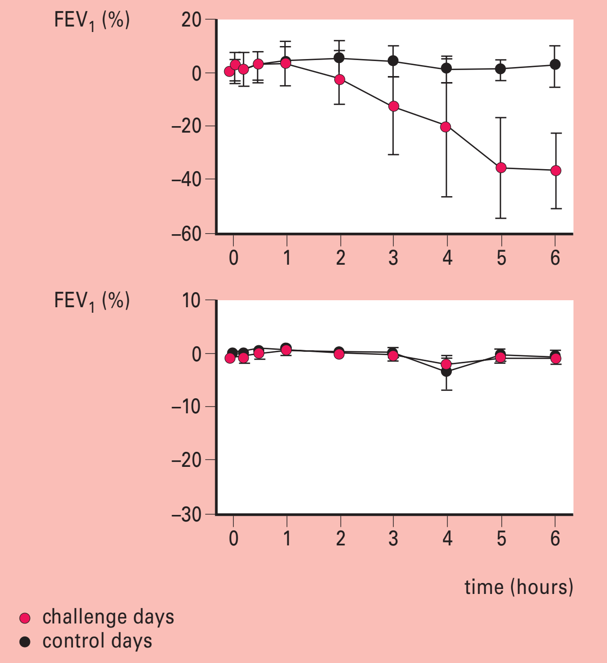

Fig. 23.11 Late asthmatic reactions induced in cat allergic patients by the intradermal injection of peptides derived from the cat allergen, Fel d 1. The nine responders show a mean fall in forced expiratory volume in 1 second (FEV$_1$) of approximately 30%. The response to the peptides is MHC-restricted and correlated with the ability of the patients' T cells to respond to these peptides in vitro. On challenge days (red filled circle) injection of peptides was associated with a fall in FEV$_1$ which did not occur on the control days (black filled circle). Data are shown for nine responders (upper graph) and 31 non-responders (lower graph). (Courtesy of Dr M Larché from J Exp Med 1999;189:1885)

Rat peritoneal mast cells

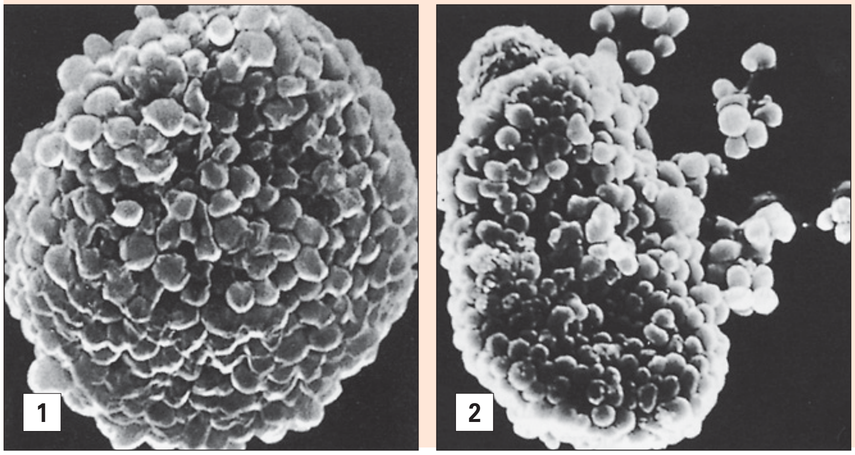

Fig. 23.12 Scanning electron micrograph (SEM) of rat peritoneal mast cells. (**1**) An intact mast cell with the cell membrane shrunk onto the granules (see Fig. 2.11). (**2**) A rat peritoneal mast cell degranulating following incubation with anti-IgE for 30 seconds. SEM. × 1500. (Courtesy of Dr TSC Orr)

- mast cells cannot be identified in the circulation, but are present in connective tissue and at mucosal surfaces throughout the body (Fig. 23.12).

Mast cells in different tissues are morphologically and cytogenetically distinct.

Both the cells that contain histamine and the biology of these cells may be very different in other species. For example:

- in the rabbit the histamine content of the peripheral blood is almost all in platelets;
- in the mouse there are few if any circulating basophils; and
- in rats the degranulation of mast cells appears to be one granule at a time (see Fig. 23.12).

By contrast, human granules tend to fuse and release their contents together (Fig. 23.13).

MMCs and CTMCs have distinct granule proteases

Mast cells were originally identified by Ehrlich who named them based on the distinctive, tightly packed granules. (*Mast* means well fed, or fattening, in German.)

Mast cells in different tissues can be distinguished by staining for proteases, and the content of these enzymes may be relevant to their role in allergic diseases.

The granule proteases of mast cells have been cloned and sequenced and are distinct for two types of mast cell (Fig. 23.14):

- mucosal mast cells (MMCs) are characterized by the presence of tryptase without chymase;
- by contrast, connective tissue mast cells (CTMCs) contain both chymase and tryptase.

These enzymes may play a direct role in the lung inflammation of asthma, either by breaking down mediators or, in the case of tryptase, by acting as a fibroblast growth factor. Basophils contain very little of either of these proteases.

Staining of basophils in tissue sections requires special fixation and staining. Without this staining the granules in basophils cannot be identified and the cells appear as neutrophils (i.e. polymorphs without special granules).

In allergic individuals mast cells can be recruited to the skin and to the nose

Although mast cells are present in normal non-inflamed tissue, their numbers are increased in response to inflammation. It is assumed that this accumulation is T cell dependent because in rats infected with *Nippostrongylus brasiliensis* accumulation of mast cells in the gut is dependent on T cells and can be suppressed by corticosteroids.

In guinea pigs the immune response to tick bites includes a large local accumulation of basophils. Indeed, the tick is thought to be killed by basophils that it ingests.

In allergic individuals mast cell recruitment has been demonstrated both:

- in the skin in response to repeated allergen exposure; and
- in the nose during the pollen season.

In both situations basophils are also recruited.

In the nose the recruitment of cells represents a shift so that mast cells move from the subepithelium into the epithelium while basophils appear in the nasal mucus. This process, which brings histamine-containing cells closer to the site of entry of allergen, is one of the ways in which allergic individuals become more sensitive. It is likely but less well established that equivalent processes occur in the human lung and gut.

Cross-linking of two FcεRI receptors results in degranulation

The process of degranulation in human mast cells and basophils involves fusing of the membrane of the granules containing histamine with the exterior cell membrane (see Fig. 23.13). The granule membrane becomes part of the cell membrane. The granule contents rapidly dissolve and are secreted, leaving behind a viable degranulated or partially degranulated cell. This process is initiated in most cases by cross-linking of two specific IgE molecules by their relevant allergen.

When two IgE receptors (FcεRI) are cross-linked, signal transduction through the γ chains of the receptor

Human basophils

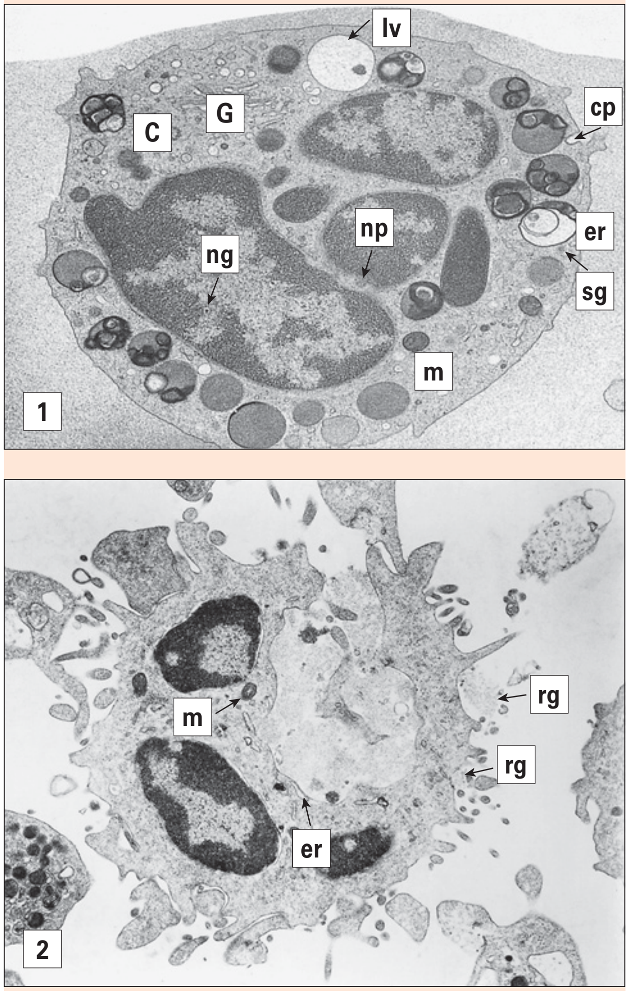

Fig. 23.13 Basophils are circulating mononuclear cells that have multilobed nuclei and distinctive granules that stain with metachromatic stains (**1**). Basophils can be recruited into local tissues such as the skin, nose, lungs, or gut by allergic and other immune responses. (**2**) A basophil degranulating 4 minutes after adding allergen. The degranulation that releases histamine occurs by fusion of the granule membrane with the external membrane of the cell. The arrows show connections between granule and the exterior of the cell. (Courtesy of Robin Hastie) (C, centriole; cp, coated pit; er, endoplasmic reticulum; np, nuclear pore; G, Golgi apparatus; lv, lucent vesicle; m, mitochondria; ng, nuclear granule; rg, residual material from granules; sg, small granules)

1 Differences between mast cell populations

	MMC	CTMC
location in vivo	gut and lung	ubiquitous
life span	< 40 days (?)	> 40 days (?)
T cell dependent	+	–
number of Fcε receptors	25×10^5	3×10^4
histamine content	+	++
cytoplasmic IgE	+	–
major AA metabolite LTC$_4$: PGD$_2$ ratio	25:1	1:40
DSGG/theophylline inhibits histamine release	–	+
major proteoglycan	chondroitin sulfate	heparin

Fig. 23.14 (**1**) There are at least two subpopulations of mast cell, the mucosal mast cells (MMCs) and the connective tissue mast cells (CTMCs). The differences in their morphology and pharmacology suggest different functional roles in vivo. MMCs are associated with parasitic worm infections and, possibly, allergic reactions. In contrast to the CTMC, the MMC is smaller, shorter lived, T cell dependent, has more Fcε receptors, and contains intracytoplasmic IgE. Both cells contain histamine and serotonin in their granules; the higher histamine content of the CTMC may be accounted for by the greater number of granules. Major arachidonic acid (AA) metabolites (prostaglandins [PGs] and leukotrienes [LTs]) are produced by both mast cell types, but in different amounts. For example, the ratios of production of the leukotriene LTC$_4$ to the prostaglandin PGD$_2$ are 25:1 in the MMC and 1:40 in the CTMC. The effect of drugs on degranulation is different between the two cell types. Sodium cromoglycate (DSCG) and theophylline both inhibit histamine release from the CTMC, but not from the MMC. (This may have important implications in the treatment of asthma.) Note that some of these data come from rodent studies and may not apply to humans.

(see Fig. 3.22) leads to influx of calcium, which initiates both degranulation and the synthesis of newly formed mediators (Fig. 23.15).

Cross-linking of IgE antibodies on FcεRI by allergens is the primary method by which mediators are released from basophils and mast cells; however, other mechanisms can be involved. Experimentally, degranulation can be triggered through FcεRI by using:

- lectins such as phytohemagglutinin (PHA) or concanavalin A (Con A); or
- antibodies to the α chain of the receptor.

Histamine release can also be triggered by agents that act on other receptors on the cell surface.

Q. Give an example of a mediator that can cause mast cell degranulation without cross-linking FcεRI.

A. Typical examples include the complement components C5a and C3a (see Fig. 6.21).

2 Differences between mast cell populations

cell type	location	amount per cell (pg)	
		tryptase	chymase
MC_T (MMC)	lung and nasal cavity, intestinal mucosa	10	< 0.04
MC_{TC} (CTMC)	skin, blood vessels, intestinal submucosa	35	4.5
basophil	circulation	0.04	< 0.04

Fig. 23.14 Cont'd (**2**) Tryptase is a tetramer of 134 kDa that may comprise as much as 25% of the mast cell protein. Chymase is a monomer of 30 kDa. The relative proportions of these proteases in mast cells define MC_T and MC_{TC} populations, which have different distributions in human tissues. Basophils have very low amounts of both proteases. (The suffixes T and TC represent tryptase and chymase present in the respective cells.)

Drugs such as codeine or morphine, the antibiotic vancomycin, and contrast media used for imaging the kidneys also degranulate mast cells. Acute reactions to these agents, which are not thought to involve IgE antibodies, are referred to as **anaphylactoid**.

Mast cell mediator release

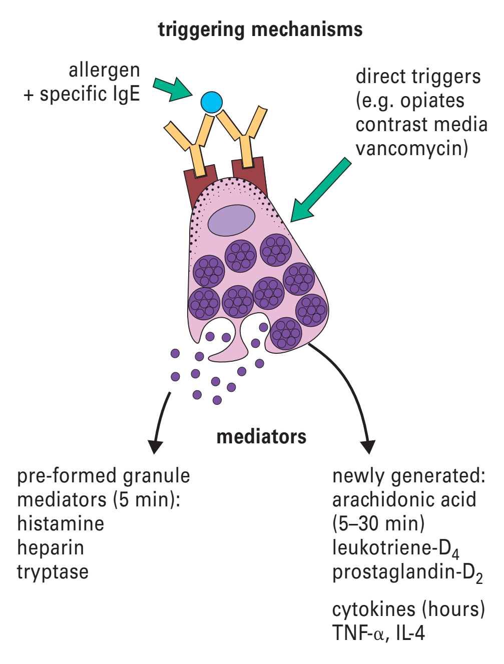

Fig. 23.15 Mast cells release mediators after cross-linking of the IgE receptors on their surface. Pre-formed mediators are released rapidly while arachidonic acid metabolites such as leukotriene D_4 and prostaglandin D_2 are released more slowly. Mast cells can also be triggered by opiates, contrast media, vancomycin, and the complement components C3a and C5a. The mediators, which are also released by basophils, include histamine, TNFα, and IL-4. Histamine released by mast cells can be measured in serum following anaphylaxis or extensive urticaria, but it has a half-life in minutes. By contrast, tryptase can be measured in serum for many hours after an anaphylactic reaction.

MULTIPLE GENES HAVE BEEN ASSOCIATED WITH ASTHMA IN DIFFERENT POPULATIONS

Children with one allergic parent have a 30% chance of developing allergic disease; those who have two allergic parents have as high as a 50% chance.

Systematic studies of allergic diseases are difficult because the phenotypes for diseases, such as hayfever and asthma, are not well defined and depend on the approach used to make the diagnosis (Fig. 23.16).

Asthma defined by a patient questionnaire is therefore less specific than asthma defined by testing of specific or non-specific bronchial hyperreactivity. Furthermore, studies on asthma are complicated because both IgE antibody responses and bronchial reactivity are genetically controlled.

Indeed, it is important not to confuse simple genetic diseases like cystic fibrosis or hemophilia with complex traits such as asthma or type II diabetes mellitus.

It is therefore not surprising that multiple genes have been associated with asthma in different populations.

A further major problem in genetic analyses of allergic disease comes from the progressive increase in the incidence of asthma between 1960 and 2000. Clearly this increase cannot be attributed to genetic change and implies that some of the genes identified would influence asthma only in the presence of other changes either in the environment or in lifestyle. This is referred to as a gene–environment interaction.

Multiple genes or genetic regions are associated with asthma. Analyses of the genetics of immediate hypersensitivity have identified both allergen-specific and non-specific influences. There are therefore HLA associations with atopy in general and also with sensitization to specific allergens. These genetic studies have given the clearest associations when purified allergens are used to test sensitization (Fig. 23.17).

The genetics of asthma has been studied both by genomic screening and by using candidate genes. Genomic screening identifies regions of the genome that link to asthma so that this region can be examined to identify specific genes.

If a candidate gene is identified, it is possible to examine the gene for polymorphisms that link to asthma. However, a brief consideration of the possible targets (Fig. 23.18)

IgE levels and atopic disease

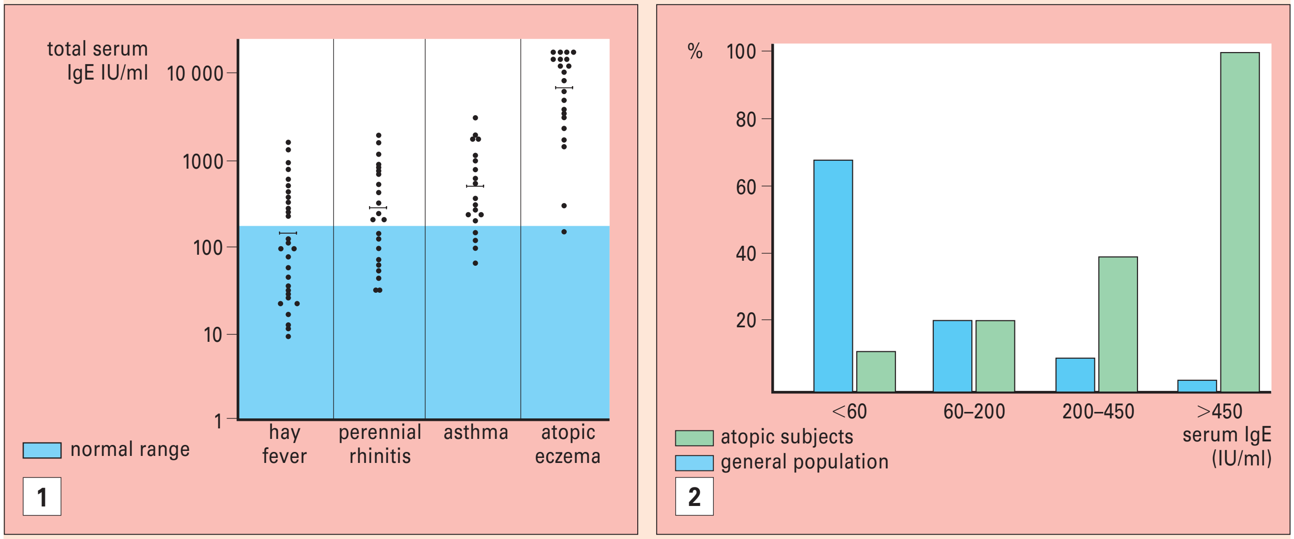

Fig. 23.16 (**1**) The serum concentration of IgE (which is around 100 IU/ml) is only approximately 0.001% that of IgG (around 10 mg/ml) and comprises less than 0.001% of the total immunoglobulin. Levels in atopic patients tend to be raised, and this is especially so in atopic eczema (1 IU = 2.4 ng). (**2**) The higher the level of IgE the smaller the percentage of the population affected, but the greater the likelihood of atopy. Where the level is greater than 450 IU/ml the majority of subjects are atopic.

Allergens and HLA associations

systematic name	old name	mol. wt (Da)	primary association	*p* value
Ambrosia (ragweed) spp.				
Amb a I	AgE	37 800	none	–
Amb a III	Ra3	12 300	A2	0.01
Amb a VI	Ra6	11 500	DR5	$< 10^{-7}$
Amb a V	Ra5	5000	DR2/Dw2	$< 10^{-9}$
Amb t V	Ra5G	4400	DR2/Dw2	$< 10^{-3}$
Lolium (ryegrass) spp.				
Lol p I	rye I	27 000	DR3/Dw3	$< 10^{-3}$
Lol p II	rye II	11 000	DR3/Dw3	$< 10^{-3}$
Lol p III	rye III	11 000	DR3/Dw3	$< 10^{-4}$

Fig. 23.17 HLA association of IgE responses to allergens from ragweed and ryegrass. (Courtesy of Dr D Marsh)

makes it clear how complex the analysis of asthma is likely to be, and indeed is proving to be. Typical examples include polymorphisms of the promoter region for IL-4 and polymorphisms of the gene for IL-5.

Q. What effect might polymorphism in these genes have?

A. The level of production of IL-4, and the activity of IL-5, will vary between individuals, each of which will affect the type of immune response that develops – TH1 versus TH2 – either of which could directly influence the inflammatory response that occurs as a result of exposure to allergens.

Alternatively, a series of polymorphisms have been identified that influence the response of asthma to treatment. These include:

- variants of the β_2-adrenergic receptor α chain; and
- genetic differences that influence the therapeutic response to leukotriene antagonists.

In the past 10 years multiple 'genes' or genetic regions associated with asthma have been identified. At present, it appears that the overall effects are too complex to be of any practical significance. Certainly it is most unlikely that gene transfer will ever be of significance. However, as genetic screening becomes easier, pharmacogenetics may

Genetic influences over asthma and allergic disease

allergen specific	HLA related
IgE	total production
	FcεRI
	FcεRII
cytokines	IL-4 promoter and receptor
	IL-5
	IL-10
	IFNγ
	TGFβ promoter
	IL-11
	IL-13 and receptor
leukotriene pathway	five lipoxygenase activating protein (FLAP)
	lipoxygenase
	LTC_4 synthase
β_2-adrenergic receptor	leukotriene receptors – LTRI, LTRII
chemokines	polymorphisms
	CCR3 receptor

Fig. 23.18 Allergic diseases run in families, but the inheritance is not simple. Population-based studies have established that the inheritance of allergic diseases is influenced by multiple genes. Some of these, such as HLA-linked control of the response to pollen antigens or genes controlling total IgE, are related to the immune response. However, many others are related to the mechanisms of inflammation (e.g. IL-4 and IL-5 gene polymorphisms) or to the response to treatment (e.g. leukotriene receptor genes or polymorphisms of the β_2-adrenergic receptor).

well become an important method for identifying the best drugs for individual patients.

SKIN TESTS ARE USED FOR DIAGNOSIS AND INVESTIGATION

The primary method for diagnosing immediate hypersensitivity is skin testing. The characteristic response is a **wheal and flare** (Fig. 23.19):

- the wheal is caused by extravasation of serum from capillaries in the skin, which occurs as a direct effect of histamine and is accompanied by pruritus (also a direct effect of histamine);
- the larger erythematous flare is mediated by an axon reflex.

This skin response takes 5–15 minutes to develop and may persist for 30 minutes or more. Techniques for skin testing include:

- a prick test, in which a 25-gauge needle or a lancet is used to introduce 0.1 μl of extract into the dermis;
- alternatively, an intradermal injection of 0.02–0.03 ml is used.

All allergen injections have the potential to cause anaphylaxis and for safety reasons the intradermal test, which introduces approximately 100 times more extract, should always be preceded by a prick test.

Skin tests are evaluated by the size of the wheal compared to a positive (histamine) and negative (saline) control. In general, a 3 × 3 mm wheal in children and a 4 × 4 mm wheal in adults can be considered a positive response to a prick test.

A positive skin test indicates that the patient has specific IgE antibodies on the mast cells in their skin. In turn this implies that bronchial or nasal challenge would also be positive if sufficient antigen were administered.

In most cases (i.e. ≈ 80%) where the skin prick test is positive, IgE antibody will be detectable in the serum. However, blood tests for IgE antibody are generally less sensitive than intradermal skin tests.

Q. Suggest explanations for the higher sensitivity of the skin test.

A. Most IgE is located on the high-affinity receptor on mast cells. Its half-life is 10 days on the mast cell as opposed to 2 days in blood. The skin test has intrinsic amplification steps caused by mast cell degranulation and mediator release.

Positive skin tests are common

Epidemiologically, sensitization to a relevant inhalant allergen is a 'risk factor' for allergic disease. An individual with a positive skin test to grass pollen is therefore up to ten times more likely to have hayfever during the grass pollen season (odds ratio ≥ 10) than a skin test-negative individual.

Equally an individual with a positive skin test to dust mite or cat allergen is more likely to have asthma (odds ratios 2–6).

It is assumed that allergen exposure contributes to the risk, but that relationship is not simple.

Positive skin tests are common, and in individual cases may not be relevant, perhaps because the patient is not exposed to the allergen, for example:

- skin tests to grass pollen are not relevant to understanding symptoms occurring during autumn;
- equally skin tests to cat dander or cockroach allergens may not be relevant if the patient has moved to an area or house where those allergens are not present.

In addition up to one-third of skin test-positive individuals do not experience symptoms when they are exposed to the relevant allergen.

Late skin reactions probably include several different events

Late reactions can occur following an immediate response to allergen, in either the skin or the lungs.

Skin tests

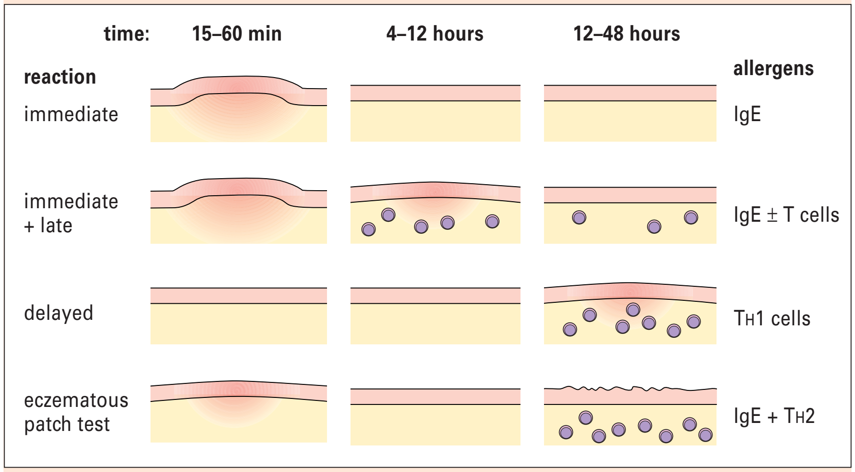

Fig. 23.19 Skin tests are carried out by introducing 0.02 ml of extract intradermally. With allergens such as pollen, cat, or dust mite, the positive reaction is an immediate (i.e. within 20 min) wheal, which in some cases is followed by an indurated response occurring late (i.e. at 4–12 hours). Non-allergic individuals make no discernible reaction to testing with these allergens. A delayed skin response is the commonest form of positive response to tuberculin, tetanus, and mumps, or to fungi such as *Trichophyton* and *Candida* spp. The skin typically shows no reaction up to 12 hours and then gradually develops an erythematous, indurated, delayed hypersensitivity response, which is maximal at 24–48 hours. Patch tests are performed by applying a gauze pad with allergen to a patch of skin that has been mildly abraded. This procedure may give an immediate wheal response, but this is followed at 24–48 hours by an indurated, erythematous response, which has many of the features of eczema. The patch test is not a diagnostic test, but has provided extensive information about the role of allergens in atopic dermatitis.

A late skin response is only common following a large immediate response (i.e. wheal size 10 × 10 mm). The late response, which is diffuse, erythematous, and indurated, generally starts 2–3 hours after the wheal and may last for up to 24 hours. The late reaction is considered to be a model of the events that lead to persistent inflammation in the nose, lungs, or skin.

Late reactions probably include several different events:

- the direct effects of prostaglandins, leukotrienes, and cytokines released by mast cells following the initial release of histamine;
- infiltration of lymphocytes, eosinophils, basophils, and neutrophils into the local site mediated by chemokines and other cytokines released from mast cells;
- release of products from the infiltrating cells.

In general, these events are occurring in parallel over a period of hours.

True delayed skin test responses (i.e. without an immediate response) are:

- characteristic of the response to tuberculin; and
- common with fungal antigens, particularly to the yeast *Candida albicans* or the dermatophyte fungus *Trichophyton* spp.

By contrast true delayed responses are very rare following skin testing with pollen, animal dander, or dust mites.

Epidermal spongiosis and a dermal infiltrate are features of a positive patch (atopy) test

The infiltration of cells into the skin that occurs in the 24 hours after an allergen is applied can be studied in several ways:

- by local intradermal injections;
- by applying a patch of allergen on gauze that stays on the skin for 2 days; or
- by fixing a chamber containing allergen over a denuded area of skin.

The skin chamber allows repeated sampling whereas the other two techniques require biopsy of the skin.

In the **patch test** 10 μg allergen is applied on a gauze pad 2.5 cm^2, and the biopsy is carried out at 24 or 48 hours. A positive patch response induces:

- macroscopic eczema;
- spongiosis of the epidermis (a hallmark of eczema); and
- an infiltrate of cells into the dermis (Fig. 23.20).

The cellular infiltrate includes eosinophils, basophils, and lymphocytes.

With persistent allergen at a site (i.e. 6 days), the eosinophils degranulate locally. This is in keeping with the

Eczema induced by patch test with mite allergen

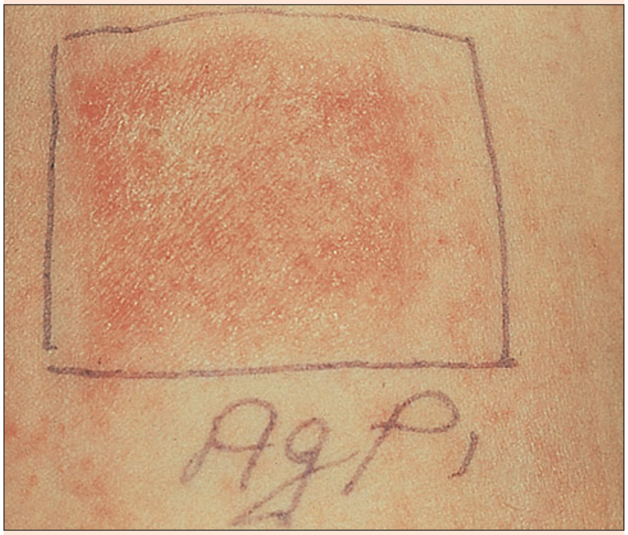

Fig. 23.20 An erythematous and eczematous skin response 48 hours after the application of 5 μg of the mite allergen Der p 1 to the skin of a patient with atopic dermatitis who had 56 IU/ml of IgE antibody to the dust mite *Dermatophagoides pteronyssinus*. Biopsy of the patch site revealed an infiltrate of eosinophils, basophils, and lymphocytes.

evidence that the skin of patients with eczema contains large quantities of the eosinophil granule major basic protein (MBP), even though very few whole eosinophilic cells are visible (Fig. 23.21).

Biopsy of patch tests also yields T cells that are specific for the allergen used, which in most cases has been dust mite, thus establishing that antigen-specific T cells are present in the skin after antigen challenge.

Some groups have succeeded in cloning allergen-specific T cells from the skin of patients with eczema without a preceding patch. This is never possible with skin from normal individuals or allergic patients without eczema.

Answering whether allergen-specific T cells are present at local sites is important because T cells could play a role both as effector cells and in the recruitment of other cells.

Establishing whether T cells play an effector role is also relevant to the nose in rhinitis, the lungs in asthma, the conjunctiva in hayfever, as well as the skin in atopic dermatitis.

Biopsy of patch test sites has also established that the Langerhans' cells in the skin of patients with eczema express FcεRI. Thus, these antigen-presenting cells use IgE antibodies to help capture allergens and to increase the efficiency of antigen presentation.

Part of the patch test response can be passively transferred by serum

Following injection of serum from a patient with atopic dermatitis into the skin of a non-allergic individual, a patch test can be carried out. Biopsy of this passively transferred response reveals large numbers of eosinophils. At least part of the eczematous response can therefore be passively transferred. The mechanisms for this response are thought to be:

- passive sensitization of IgE antibodies onto mast cells in the dermis;
- triggering of the local mast cells with allergen to release histamine, leukotrienes, and cytokines;
- recruitment of eosinophils by IL-5 as well as by chemokines such as CCL5 and CCL11 (eotaxin).

Eosinophil major basic protein in the skin of atopic dermatitis

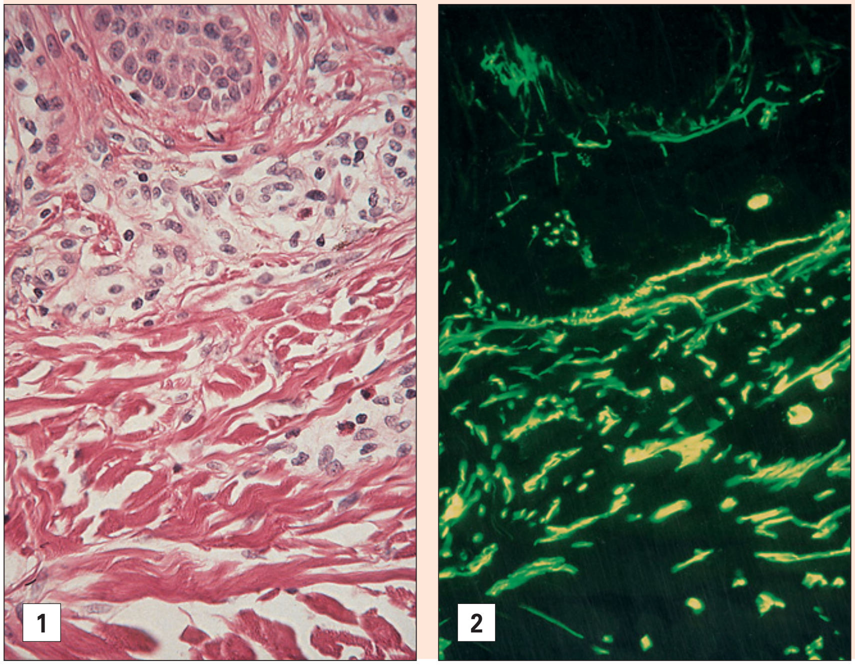

Fig. 23.21 Skin biopsy from a patient with severe atopic dermatitis. The hematoxylin and eosin (H&E) stain (**1**) shows an inflammatory infiltrate, but very few intact eosinophils are present. The same section stained with antibodies to eosinophil major basic protein (MBP) (**2**) shows extensive deposition of MBP in the dermis, demonstrating that eosinophils had degranulated in the skin. (Courtesy of Dr K Lieferman)

SEVERAL DIFFERENT PATHWAYS CAN CONTRIBUTE TO THE CHRONICITY OF SYMPTOMS OF ALLERGY

The diagnosis of allergy is made by:

- skin tests; or
- serum assays of IgE antibodies.

These antibodies form part of an immune response, which also includes antibodies of other isotypes (IgG1, IgG4, and IgA), as well as T cells, which are characteristically TH2.

The release of histamine within 15 minutes after allergen exposure can only explain a small proportion of allergic disease. In particular, the chronic inflammation in the lungs of patients with asthma and/or in the skin of patients with atopic dermatitis has many features that cannot be explained by histamine:

- first, the time course is too long;
- second, there is a cellular infiltrate in these tissues; and
- third, there are major differences in disease between patients who have apparently similar IgE antibodies in their serum and skin (Fig. 23.22).

Several different pathways contribute to chronic symptoms and can alter the severity of allergic disease as follows:

- local recruitment of mast cells and basophils, combined with increased 'releasability' of these cells, allows an increased response to the same allergen challenge – this mechanism plays a major role in the increased symptoms in the nose during the pollen season;
- release of leukotrienes, chemokines, and cytokines from mast cells or basophils – the mediators can have direct effects on blood vessels and smooth muscles – IL-5, tumor necrosis factor (TNF), and chemokines are each thought to contribute to the recruitment of inflammatory cells;
- the local actions of T effector cells – T cells release a wide range of cytokines, which can have direct inflammatory effects.

Inflammatory response in asthmatic bronchi

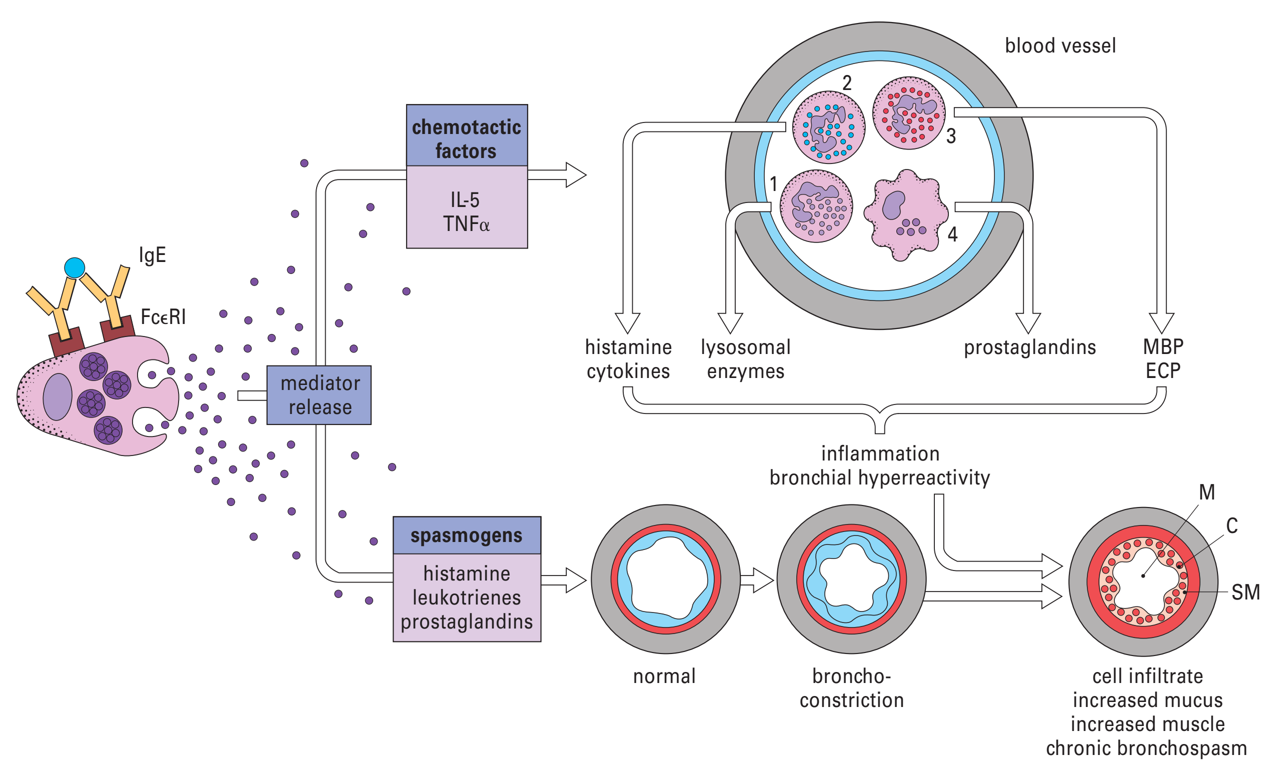
blood vessel
2
3
1
4
chemotactic factors
IL-5
TNFα
IgE
FcεRI
mediator release
histamine cytokines
lysosomal enzymes
prostaglandins
MBP ECP
inflammation
bronchial hyperreactivity
spasmogens
histamine
leukotrienes
prostaglandins
M
C
SM
normal
broncho-constriction
cell infiltrate
increased mucus
increased muscle
chronic bronchospasm

Fig. 23.22 Mast cells release factors that can induce immediate bronchospasm (e.g. histamine and LTD$_4$), but also release chemotactic factors such as LTB$_4$, IL-5, and TNFα. The spasmogens can induce edema, increased mucus, and smooth muscle constriction, resulting in an immediate decrease in airway conductance and a fall in FEV$_1$. By contrast, chemotactic factors recruit cells out of the circulation including eosinophils, neutrophils, lymphocytes, and macrophages. These cells can chronically modify the lung with goblet cell hyperplasia, collagen deposition below the basement membrane, and possibly smooth muscle hyperplasia. In addition, these cells and their products produce non-specific bronchial hyperreactivity. Thus, chronic bronchospasm includes elements of hypersecretion, inflammatory infiltrate thickening the walls of the small bronchi, and bronchial smooth muscle spasm. Evidence for this inflammatory response can be obtained from increased exhaled nitric oxide; increased eosinophils or eosinophil cationic protein (ECP) in induced sputum; and experimentally from biopsies of the lung. (1, neutrophils; 2, basophils; 3, eosinophils; 4, monocytes; C, cells; M, mucus; SM, smooth muscle)

Separating out the different factors influencing chronic allergic symptoms is difficult. Information may be obtained from:

- the effects of different drugs, particularly disodium cromoglycate, leukotriene antagonists, and anti-IgE; and
- passive transfer experiments.

The fact that cromoglycate can inhibit both the immediate and the late response of the lung to allergen challenge has been taken as evidence that the late reaction is also dependent on mast cell triggering. However, cromoglycate can control only part of the chronic inflammation.

By contrast, corticosteroids, which are an effective treatment for most of the inflammation in asthma, act selectively on the late response. The response to corticosteroid treatment cannot be used to distinguish the delayed effects of mast cell triggering from the direct effects of T cells because corticosteroids can inhibit both of these mechanisms (Fig. 23.23).

Local or systemic injection of IgE antibodies can passively transfer the wheal and flare skin response. Passive transfer of serum from an allergic patient (i.e. containing IgE antibodies) into the skin of a non-allergic individual can also transfer some aspects of the delayed or cellular response.

Both late reactions following an intradermal injection of allergen and the patch test response at 48 hours can be passively transferred. This experiment demonstrates that cross-linking of IgE antibody on mast cells in the tissue of a non-allergic individual can lead to local recruitment of eosinophils in the absence of antigen-specific T cells.

Therefore, in any analysis of the factors influencing the severity of allergic disease (e.g. response to pharmacological treatment or response to immunotherapy), it is necessary to consider the relevance of both mast cells and effector T cells.

ALLERGENS CONTRIBUTE TO ASTHMA

The causal role of bee venom in anaphylaxis or grass pollen in seasonal hayfever is obvious because:

- these diseases occur in individuals who have positive skin tests; and
- the symptoms are directly related to increased exposure.

By contrast, the role of inhaled allergens in chronic asthma is less obvious because exposure is perennial, the patients are often not aware of the relationship, and only a proportion of skin test-positive individuals develop asthma.

The evidence that allergens derived from dust mites, cats, dogs, the German cockroach, or the fungus *Alternaria* spp. contribute to asthma comes from several different lines of evidence:

- the epidemiological evidence that positive skin tests or serum IgE antibodies are a major risk factor for asthma;
- bronchial challenge with nebulized extracts can produce both rapid bronchospasm, within 20 minutes, and a late reaction, in 4–8 hours, which is characterized by renewed mediator production and a cellular infiltrate;
- reduced exposure to allergens can lead to decreased symptoms and decreased non-specific bronchial reactivity – this avoidance can be achieved either by moving patients to an allergen-free unit or by controlling exposure in the home.

Bronchi in the lungs of patients with asthma are characterized by increased mast cells, lymphocytes of the TH2 type, eosinophils, and products of eosinophils. In addition, there is increased mucus production secondary to goblet cell hyperplasia, epithelial desquamation, and collagen deposition below the basement membrane. These changes are a reflection of chronic inflammation, and it is generally considered that eosinophils play a major role in these events (see Fig. 23.22).

Q. Some recent evidence has shown that anti-IL-5 treatment has limited effects on asthma, though it decreases circulating eosinophils. How do you interpret this result?

A. Other cells, including basophils, mast cells, effector T cells, and macrophages, all contribute to the non-specific bronchial reactivity.

BAL analysis after allergen challenge demonstrates mast cell and eosinophil products

Analysis of bronchoalveolar lavage (BAL) after an allergen challenge demonstrates the presence of products derived from mast cells and eosinophils.

Actions of corticosteroids in allergic disease

- act primarily on the delayed or chronic effects of allergen exposure and are seen as anti-inflammatory
- in allergen challenge situations, block the delayed or late response in the lungs and inhibit the influx of eosinophils, basophils, and lymphocytes into local sites
- following systemic administration, circulating eosinophils decrease rapidly because of margination
- prevent eosinophil production in the bone marrow
- bind to a receptor that leads to inhibition of the transcription of the genes for many cytokines, including IL-5, TNFα, and also some chemokines
- are thought to have limited effects on leukotriene production and do not inhibit histamine release
- downregulate T cell activity as shown by the result of their clinical effect on contact sensitivity and blockade of delayed hypersensitivity skin tests

Fig. 23.23 Locally active corticosteroids are widely used in seasonal rhinitis, perennial rhinitis, asthma, and atopic dermatitis. In addition, courses of systemic corticosteroid are used for the treatment of exacerbations of asthma.

Q. What products would you expect to detect in BAL following challenge?

A. Histamine, prostaglandins, and leukotrienes from mast cells; major basic protein (MBP) and eosinophil cationic protein (ECP) from eosinophils (see Chapter 2).

Furthermore, MBP is present in biopsies of the lungs and can produce epithelial change typical of asthma in vitro (Fig. 23.24).

The subepithelial collagen deposition present in many patients with asthma is probably a reflection of fibroblast responses to local inflammation.

Although it has been suggested that these changes, which are referred to as ‘remodeling’, can lead to progressive decreases in lung function, the evidence for this view is not clear. In particular, progressive loss of lung function is unusual in asthma and there are no studies showing a correlation between the extent of collagen deposition and changes in lung function. Nonetheless, inhaled corticosteroids, which can block many different aspects of inflammation, are an effective long-term treatment that can control asthma.

Bronchial hyperreactivity is a major feature of asthma

Non-specific bronchial hyperreactivity (BHR) is present in patients with asthma and is a major feature of the disease. Thus, airway obstruction, induced by cold air or exercise, and nocturnal asthma all correlate with non-specific bronchial reactivity.

BHR can be demonstrated by challenging the lungs with histamine, methacholine, or cold air.

The mechanism by which exercise or cold air induces a bronchial response is thought to be evaporation of water with associated cooling of the epithelium. However, it is unclear whether this process triggers nerve endings directly or by causing local mediator release.

Evidence for inflammation of the lungs of patients with asthma is indirect

Bronchoscopy is not possible in patients with asthma except as a research procedure. Therefore the only evidence for inflammation of the lungs that can be obtained routinely is indirect:

- peripheral blood or nasal smear eosinophils are increased in most patients presenting with an acute episode of asthma (Fig. 23.25);
- nasal secretions may contain increased ECP and CXCL8.

Additional evidence about inflammation in the lungs can be obtained either from exhaled air or from condensates of exhaled air. Nitric oxide ($NO^{\bullet}$) gas is increased in patients with asthma, and this decreases following systemic or local corticosteroid treatment.

Q. Why should $NO^{\bullet}$ be increased in the lungs of asthmatic patients?

A. Macrophage activation causes synthesis of inducible nitric oxide synthase (iNOS) and consequent $NO^{\bullet}$ production in the lungs (see Fig. 9.31).

Localization of MBP in the lung of a severe asthmatic

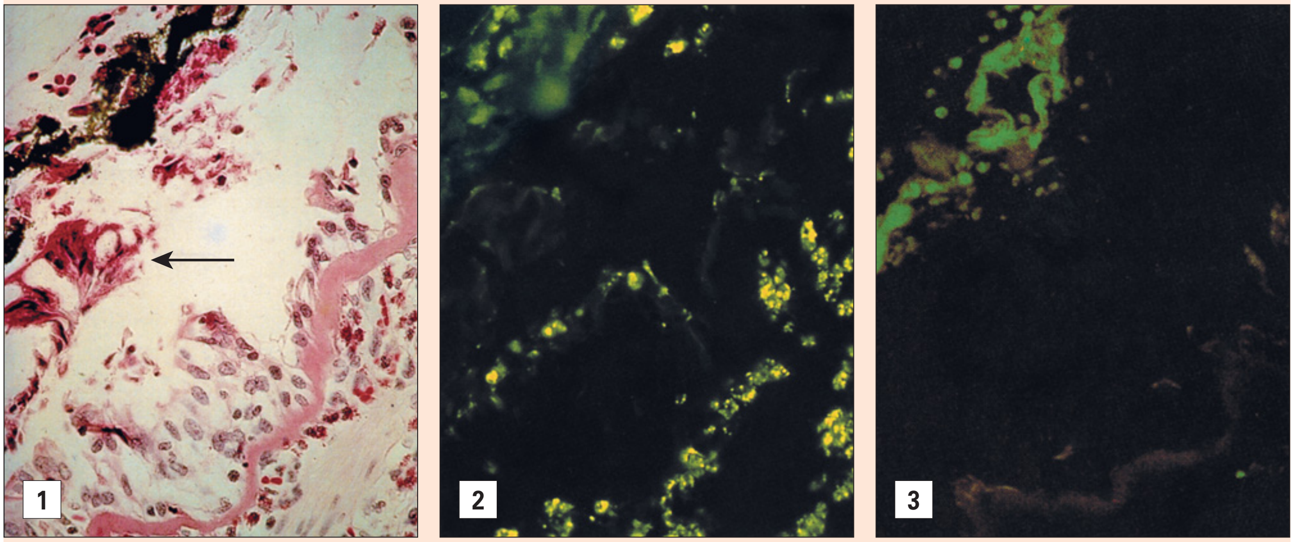

Fig. 23.24 (**1**) Respiratory epithelium showing striking submucosal eosinophil infiltration and a cluster of desquamated epithelial cells in the bronchial lumen (arrowed) next to a ‘stringy’ deposit of soot. H&E stain. (**2**) The same section stained for major basic protein (MBP) showing immunofluorescent localization in infiltrating eosinophils. MBP deposits are also seen on desquamated epithelial cells on the luminal surface. (**3**) A control section stained with normal rabbit serum does not stain eosinophils or bronchial tissue, but does show some non-specific staining of the sooty deposit. (Courtesy of Dr G Gleich, reprinted from J Allergy Clin Immunol 1982;70:160–169, with permission from American Academy of Allergy Asthma and Immunology)

Nasal eosinophils

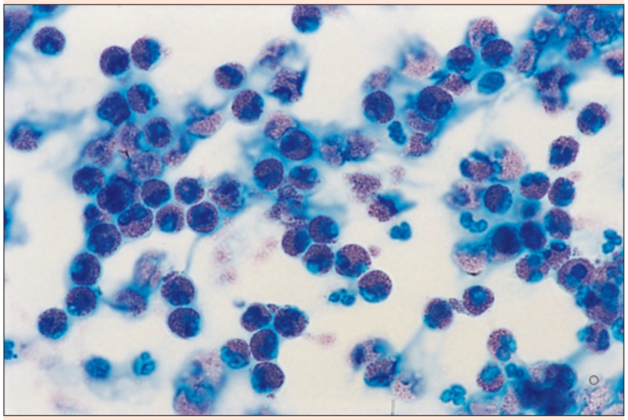

Fig. 23.25 Nasal smear from an 8-year-old boy presenting with acute asthma. Most of the cells are eosinophils – polymorphonuclear cells with a cytoplasm that stains red using H&E stain. He was known to be allergic to dust mites and had recently had a rhinovirus infection as judged by polymerase chain reaction on nasal secretions.

In addition, the pH of the condensate decreases during acute episodes. The increased exhaled $NO^{\bullet}$ may reflect upregulation of the enzyme NOS. Alternatively, the release of $NO^{\bullet}$ gas could also increase acutely as a consequence of airway acidification.

In adults further information about the inflammation in the 'respiratory tract' can be obtained from computed tomography (CT) of the nasal sinuses. Extensive opacification of the sinuses is present in approximately one-third of patients presenting with acute asthma. This reflects both:

- chronic sinusitis, which is a major feature of late-onset asthma; and
- sinus inflammation secondary to acute rhinovirus infection.

Whether the changes in the sinuses are a reflection of similar effects occurring in the lungs, or a source of mediators, or T cells that contribute to lung inflammation, is not clear.

IMMUNOTHERAPY IS AN EFFECTIVE TREATMENT FOR HAYFEVER AND ANAPHYLACTIC SENSITIVITY TO VENOM

Immunotherapy (or hyposensitization) with allergen extracts was introduced in 1911 by Noon and Freeman. At that time they were trying to establish immunity against pollen toxin.

Immunotherapy requires regular injections of allergen over a period of months. It is an established treatment for:

- seasonal hayfever; and
- anaphylactic sensitivity to bees, wasps, and hornets.

In addition, immunotherapy is an effective treatment for selected cases of other allergic diseases including asthma.

The dose is increased progressively, starting with between 1 and 10 ng and increasing up to approximately 10 µg allergen per dose.

The response to treatment includes:

- an increase in serum IgG antibodies;
- a striking decrease in the response of peripheral blood T cells to antigen in vitro; and
- a marked decrease in late reactions in the skin.

Over a longer period of time there is a progressive decrease in IgE antibodies in the serum (Fig. 23.26).

The change in antibodies, lymphocyte responses, and symptoms could all be secondary to changes in T cells. Given the known mechanisms of allergic inflammation, a response of T cells to allergen injections could influence symptoms in several ways as follows:

- decreased local recruitment of mast cells and basophils;
- decreased recruitment of eosinophils to the nose or lungs;
- increased IgG including IgG4 antibodies with progressive decreases in IgE – the IgG antibodies may act as blocking antibodies by binding allergen before it cross-links IgE on mast cells.

Some studies of cytokine RNA have suggested that immunotherapy produces a shift in T cells from a TH2 profile (i.e. IL-4 and IL-5) towards a profile that is more typical of TH1 (i.e. IFNγ). Although this could explain decreased help for IgE, and decreased eosinophil recruit-

Effects of immunotherapy on allergic rhinitis

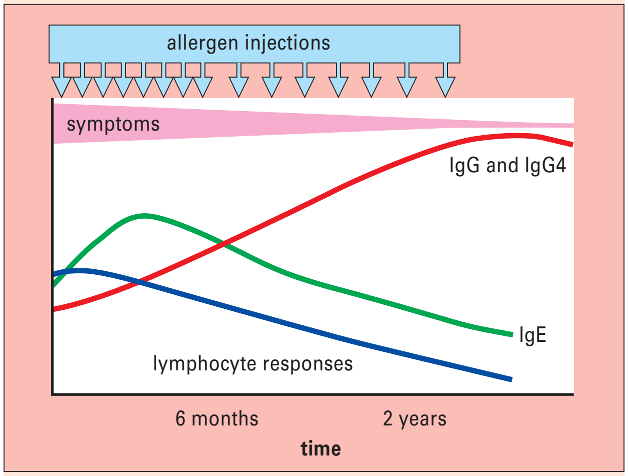

Fig. 23.26 During desensitization or immunotherapy the allergic patient receives regular subcutaneous injections of the relevant allergen. The immunological changes that occur include an initial increase in IgE antibodies followed by a gradual decline, which in pollen-allergic patients is largely due to a blunting of the seasonal increase. Antibodies of the IgG and specifically IgG4 isotype increase progressively and may reach concentrations of ten times those present before treatment. Symptoms decline, starting as early as 3 months, but generally not maximally until 2 years. Changes in T cells are less well defined, but include decreased in-vitro response to allergens and increased production of IL-10.

ment, this would not explain the production of IgG4. The expression of the gene for IgG4 is dependent on IL-4, and may also require the cytokine IL-10. The response to immunotherapy is therefore better seen as a modification of the TH2 response.

NEW APPROACHES FOR TREATING ALLERGIC DISEASE

Modified forms of allergen-specific immunotherapy

Peptides from the primary sequence of an allergen that can stimulate T cells in vitro

Peptides from the primary sequence of an allergen, usually approximately 20 amino acids in length, stimulate T cells in vitro.

In theory, peptides provide a mechanism for stimulating or desensitizing T cells without the risk of anaphylaxis, which is always present with traditional allergens.

Whether incomplete stimulation of T cells by peptides can lead to 'tolerance' or a change in the cytokine profile is not clear. Problems include:

- significant reactions in the lung with a fall in FEV_1; and
- the fact that multiple peptides are necessary to allow presentation of antigen in patients with different HLA types.

MODIFIED RECOMBINANT ALLERGENS HAVE DECREASED BINDING – Genetically modified recombinant allergens that have decreased binding to IgE antibodies can be produced. Their advantage is that the primary sequence with the T cell epitopes is preserved. Even if the molecule is extensively modified, any full-length protein has the potential to induce anaphylaxis in allergic individuals. Thus, the use of genetically modified molecules would always require precautions similar to those for traditional immunotherapy. A potential but unlikely problem is that patients would develop IgE antibodies against new epitopes.

ADJUVANTS CAN SHIFT THE IMMUNE RESPONSE TO TH1 – Adjuvants 'attached' to allergen molecules have been designed to shift the immune response from TH2 towards TH1. Possible co-molecules that act like an adjuvant include:

- the cytokine IL-12; or
- immunostimulatory sequences (ISSs).

ISSs are DNA sequences such as cytosine phosphoguanidine (CpG) that are common in bacterial DNA and have a profound effect on the mammalian immune system.

In mice combining an antigen with two or three molecules of CpG can induce a TH1 response or downregulate IgE responses.

Combining CpG with allergen not only influences the response, but also reduces the reactivity of the allergen with IgE.

Thus immunization with allergen and CpG may produce a greater immune response with less potential for an acute allergic reaction.

Preliminary trials in allergic individuals using ragweed antigen linked to CpG have been successful.

DNA vaccines are being designed to change the immune response

The concept of immunizing with the gene for an antigen is well established (i.e. DNA vaccines).

This approach has potential for the treatment of allergy because the DNA vector can be designed to change the immune response.

Prokaryotic DNA includes CpG motifs so that, as the antigen is expressed, it will induce a TH1 response.

Experiments with DNA vaccines have been very successful in mice, both in inducing a TH1 response initially and in controlling an existing IgE antibody response. However, the consequences of expressing an allergen within the tissue of an allergic individual are not known. Equally, it is not clear whether inducing a TH1 response to a ubiquitous allergen such as cat dander would give rise to other forms of inflammatory disease.

New forms of non-specific therapy

Humanized monoclonal anti-IgE

Humanized monoclonal anti-IgE treatment can reduce sensitivity and significantly decrease the number of acute episodes of asthma per year.

Antibodies directed against the binding site for FcεRI on IgE can bind to IgE in the circulation, but not when it is attached to mast cells or basophils. An antibody of this kind can therefore remove IgE from the circulation but will not induce anaphylaxis. It also downregulates IgE production by simultaneously binding to the B cell IgE surface receptor and the FcγRIIb receptor.

A mouse monoclonal antibody to IgE has been progressively humanized so that the molecule can be safely injected into patients and will bind IgE with very high affinity.

Treatment with anti-IgE antibodies has reduced the symptoms of both asthma and hayfever. In addition, continued treatment that controls free IgE below 10 ng/ml leads to a progressive decrease in the number of IgE receptors on mast cells. Thus, the treatment may achieve a secondary effect, further decreasing the sensitivity of histamine-containing cells to allergen.

The role of anti-IgE in treating food allergy, atopic dermatitis, and drug allergy remains to be established.

Recombinant soluble IL-4R can block the biological activity of IL-4

Given the central role of IL-4 in the TH2 response, it is not surprising that several efforts have been made to block its action. These include:

- a mutated IL-4 (Y124D);
- antibodies to IL-4; and
- recombinant soluble IL-4 receptor (sIL-4R).

Treatment with sIL-4R has proved moderately effective in clinical trials of allergic asthma. The mechanism is that sIL-4R binds to IL-4 before it can react with the receptor on T cells or B cells, and thus blocks its biological activity. However, it is less clear which of the many actions of IL-4 is relevant to the clinical effects:

- blocking the action of IL-4 on B cells may reduce IgE production, but would probably require several weeks to produce a clinical effect;

- the autocrine effect of IL-4 on TH2 cells may be an essential growth factor.

The efficacy of sIL-4R provides indirect evidence for the role of T cells in allergic disease.

Humanized monoclonal anti-IL-5 decreases circulating eosinophils

Anti-IL-5 (like anti-IgE) is a humanized mouse monoclonal antibody.

Following successful studies in baboons, anti-IL-5 has been shown to decrease circulating eosinophils in patients. It is therefore assumed that binding IL-5 produced by T cells (or mast cells) can decrease the production of eosinophils in the bone marrow. However, the results do not answer whether the treatment acts on IL-5 in the circulation or on IL-5 produced by T cells (or mast cells) locally in the bone marrow and/or the respiratory tract.

To the surprise of most investigators in the field, treatment with anti-IL-5 was not effective as a treatment for asthma despite an approximately 90% decrease in circulating eosinophils.

Some new treatment approaches may not be practical

The treatment of allergic disease is based on:

- allergen avoidance;
- pharmacological management including disodium cromoglycate, leukotriene antagonists, and local corticosteroids; and
- immunotherapy.

The treatment approaches using peptides, modified allergens, or allergens linked to CpG may not be practical because each allergen would have to go through clinical trials.

Although specific antagonists to other cytokines appear to be an attractive target for treatment, it is increasingly unlikely that they will be clinically successful in competition with anti-IgE, corticosteroids, and leukotriene antagonists.

IgE ANTIBODIES PLAY A CRITICAL ROLE IN DEFENSE AGAINST HELMINTHS

The biological role of IgE has been studied both in relation to human disease and using animal models. In tropical countries total serum IgE is usually much higher than in the West. Typical mean serum IgE levels among rural populations in tropical countries are 1000– 2000 IU/ml compared to approximately 100–200 IU/ml in the West.

The best established cause of increased IgE is infection with helminths (e.g. ascaris, hookworm, or schistosomiasis). Elevated serum IgE is not a feature of protozoal infections such as malaria and trypanosomiasis, or of bacterial infections such as tuberculosis or leprosy.

The hypothesis is that IgE antibodies play a critical role in the defense against helminths, primarily by acting as a gate-keeper.

A good example is to consider the possible role of immediate skin sensitivity in the protection against schistosomiasis. As the schistosomules enter through the skin of a sensitized individual, they trigger mast cell degranulation with release of histamine, leukotrienes, and cytokines. These mediators lead to the local accumulation of serum (which contains IgG antibodies) and eosinophils. IgG antibodies bound to schistosomules interact with FcγR on eosinophils leading to degranulation on the surface of the worm, which kills it (Fig. 23.27).

Another important protective mechanism against helminths is the expulsion of worms from the gut. This involves increased mucus production and activated peristalsis, as well as a role for eosinophils and mast cells.

Eosinophil adhesion and degranulation

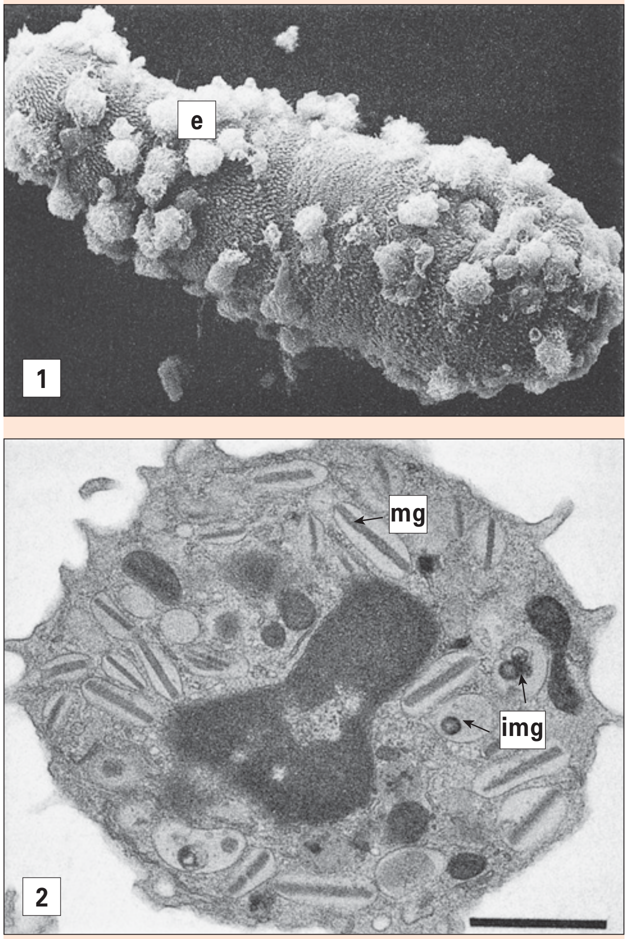

Fig. 23.27 (**1**) A schistosomule being killed by eosinophils that had been cultured from mouse bone marrow in the presence of IL-5. The larval helminth had first been treated with IgG antibodies and the eosinophils adhere by means of their Fcγ receptors. (Courtesy of Dr C Sanderson) (**2**) Transmission electron micrograph of an eosinophil. Degranulation of eosinophils usually leads to death of the cells. The mediators released include major basic protein (MBP), eosinophil cationic protein (ECP), eosinophil peroxidase (EPO), and eosinophil derived neurotoxin (EDN). In addition, eosinophils can produce leukotrienes and the cytokines IL-5, granulocyte–macrophage colony stimulating factor (GM-CSF) and TGFα. Scale bar 1 μm. (img, immature granules; mg, mature granules with crystalline core) (Courtesy of Dr C Sanderson)

Whether IgE and IgE antibodies are essential for this expulsion is not clear. However, the best analysis is that protection against worm infection is a primary role of the TH2 system. This includes roles for effector T cells, mast cells, basophils, and IgE antibodies. Furthermore, the primary mechanisms are focused on preventing entry of new worms.

Given that one third of the world's population is infested with helminths, it is clear that a protective system could have sufficient survival advantage to explain the presence of an effective TH2 system.

Perhaps as a consequence of living in the West without helminth infection, a large proportion of the population generates a TH2-based allergic response against irrelevant antigens such as those on pollen grains, cat dander, or mite fecal particles.

FURTHER READING

Akdis CA, Blaser K. IL-10-induced anergy in peripheral T cell and reactivation by microenvironmental cytokines: two key steps in specific immunotherapy. FASEB J 1999;13:603–609.

Beaven MA, Metzger H. Signal transduction by Fc receptors: the FcεRI case. Immunol Today 1993;14:222–226.

Borish L, Rosenwasser L. TH1/TH2 lymphocytes: doubt some more. J Allergy Clin Immunol 1997;99:161–164.

Bruynzeel-Koomen C, Wichen D, Toonstra J, et al. The presence of IgE molecules on epidermal Langerhans' cells in patients with atopic dermatitis. Arch Dermatol Res 1986;278:199–205.

Coca AF, Cooke RA. On the classification of the phenomenon of hypersensitiveness. J Immunol 1923;8:163.

Coyle AJ, Wagner K, Bertrand C, et al. Central role of immunoglobulin (Ig) E in the induction of lung eosinophil infiltration and T helper 2 cell cytokine production: inhibition by a non-anaphylactogenic anti-IgE antibody. J Exp Med 1996;183:1303–1310.

Galli SJ. New concepts about the mast cell. N Engl J Med 1993;328:257–265.

Geha RF. Regulation of IgE synthesis in humans. J Allergy Clin Immunol 1992;90:143–150.

Haselden BM, Kay AB, Larch M. Immunoglobulin E-independent major histocompatibility complex-restricted T cell peptide epitope-induced late asthmatic reactions. J Exp Med 1999;189:1885–1894.

Marsh DG, Neely JD, Breezeale DR, et al. Linkage analysis of IL-4 and other chromosome 5q31.1 markers and total serum immunoglobulin E concentrations. Science 1994;264:1152–1156.

Miller JS, Schwartz LB. Human mast cell proteases and mast cell heterogeneity. Curr Opin Immunol 1989;1:637–642.

Montford S, Robinson HC, Holgate ST. The bronchial epithelium as a target for inflammatory attack in asthma. Clin Exp Immunol 1992;22:511–520.

Platts-Mills TAE, Vervloet D, Thomas WR, et al. Indoor allergens and asthma: report of the Third International Workshop. J Allergy Clin Immunol 1997;100:S2–S24.

Platts-Mills TAE, Vaughan JW, Squillace S, Woodfolk JA, Sporik RB. Sensitisation, asthma and a modified TH2 response in children exposed to cat allergen. Lancet 2001;357:752–756.

Prausnitz C, Kustner H. In: Gell PGH, Coombes RRA, eds. Clinical aspects of immunology. Oxford: Blackwell Scientific Publications; 1962:808–816.

Sporik R, Holgate ST, Platts-Mills TAE, Cogswell JJ. Exposure to house-dust mite allergen (Der p I) and the development of asthma in childhood. A prospective study. N Engl J Med 1990;323:502–507.

Spry CJF, Kay AB, Gleich GJ. Eosinophils 1992. Immunol Today 1992;13:384–387.

Wan H, Winton HL, Soeller C, et al. Der p 1 facilitates transepithelial allergen delivery by disruption of tight junctions. J Clin Invest 1999;104:123–133.

Wide L, Bennich H, Johansson SGO. Diagnosis of allergy by an in vitro test for allergen antibodies. Lancet 1967;ii:1105.

Critical thinking: Severe anaphylactic shock (see p. 501 for explanations)

Sixty-two-year-old Mrs Young was stung by a bee from a hive in her back garden. Harvesting the honey had left her with several stings during the course of the summer. Several minutes after the recent sting she complained of an itching sensation in her hands, feet, and groin accompanied by cramping abdominal pain. Shortly afterwards she felt faint and acutely short of breath. Moments later she collapsed and lost consciousness. Her husband, a doctor, noticed that her breathing was rapid and wheezy and that she had swollen eyelids and lips. She was pale and had patchy erythema across her neck and arms.

On examination her apex beat could be felt, but her radial pulse was weak. Her husband immediately administered 0.5 ml of 1/1000 epinephrine (adrenaline) intramuscularly and 10 mg of chlorpheniramine (chlorphenamine) (an H_1-receptor antihistamine) intravenously with 100 mg of hydrocortisone. She regained consciousness and her respiratory rate dropped. By the following day she had recovered completely. Results of investigations at this time are shown in the table.

Mrs Young had no previous history of adverse reactions to bee venom, foods, or antibiotics. In addition there was no history of asthma, allergic rhinitis, food allergy, or atopic dermatitis. A diagnosis of anaphylactic shock due to bee venom sensitivity was made based on the history and investigations, and a decision taken to commence desensitization therapy.

investigation	**result (normal range)**
hemoglobin (g/dl)	14.2 (11.5–16.0)
white cell count (× 10^9/l)	7.5 (4.0–11.0)
neutrophils (× 10^9/l)	4.4 (2.0–7.5)
eosinophils (× 10^9/l)	0.40 (0.04–0.44)
total lymphocytes ((× 10^9/l)	2.4 (1.6–3.5)
platelet count (× 10^9/l)	296 (150–400)
serum immunoglobulins IgG (g/l) IgM (g/l) IgA (g/l) IgE (IU/ml)	 10.2 (5.4–16.1) 0.9 (0.5–1.9) 2.1 (0.8–2.8) 320 (3–150)
RAST bee venom wasp venom	 class 4 class 0
skin prick tests	grade (0–5)
bee venom (10 µg/ml)	3+

Mrs Young was made aware of the possible risk of the procedure and consented to it. She was injected subcutaneously with gradually increasing doses of bee venom, the procedures being performed in hospital with access to resuscitation apparatus. No further allergic reactions occurred and she was maintained on a dose of bee venom at 1-month intervals for the next 2 years. She was stung by a bee the following summer and had no adverse reaction.

1. What mechanisms are involved in anaphylaxis?
2. What are the clinical features and management of acute anaphylaxis?
3. How may such sensitivity be detected and what can be done to desensitize patients?

This Page Intentionally Left Blank

CHAPTER

24

Hypersensitivity (Type II)

SUMMARY

- **Type II hypersensitivity is mediated by antibodies binding to specific cells.** Type II hypersensitivity reactions are caused by IgG or IgM antibodies against cell surface and extracellular matrix antigens. The antibodies damage cells and tissues by activating complement, and by binding and activating effector cells carrying Fcγ receptors.

- **Type II hypersensitivity reactions may target cells.** Transfusion reactions to erythrocytes are produced by antibodies to blood group antigens, which may occur naturally or may have been induced by previous contact with incompatible tissue or blood following transplantation, transfusion or during pregnancy.

- **Hemolytic disease of the newborn** occurs when maternal antibodies to fetal blood group antigens cross the placenta and destroy the fetal erythrocytes.

- **Type II hypersensitivity reactions may target tissues.** Damage to tissues may be produced by antibody to functional cell surface receptors through Fab regions. In doing so, they may enhance or inhibit the normal activity of the receptor. Examples include myasthenia gravis, pemphigus, and Goodpasture's syndrome.

- **The role of autoantibodies in disease is not always clear.** Antibodies to intracellular components are not normally pathogenetic, but they may be diagnostically useful.

- **Cytotoxic antibodies are increasingly being used for therapeutic indications.**

TYPE II HYPERSENSITIVITY REACTIONS ARE CAUSED BY IgG OR IgM ANTIBODIES AGAINST CELL SURFACE AND EXTRACELLULAR MATRIX ANTIGENS

Type II hypersensitivity reactions are mediated by IgG and IgM antibodies binding to specific cells or tissues. The damage caused is therefore restricted to the specific cells or tissues bearing the antigens. In general:

- antibodies directed against cell surface antigens are usually pathogenic;
- antibodies directed against internal antigens are usually not pathogenic.

Type II reactions therefore differ from type III reactions, which involve antibodies directed against soluble antigens in the serum, leading to the formation of circulating antigen–antibody complexes. Damage occurs when the complexes are deposited non-specifically onto tissues and/or organs (see Chapter 25).

Cells engage their targets using Fc and C3 receptors

In type II hypersensitivity, antibody directed against cell surface or tissue antigens interacts with the **Fc receptors (FcR)** on a variety of effector cells and can activate complement to bring about damage to the target cells (Fig. 24.1).

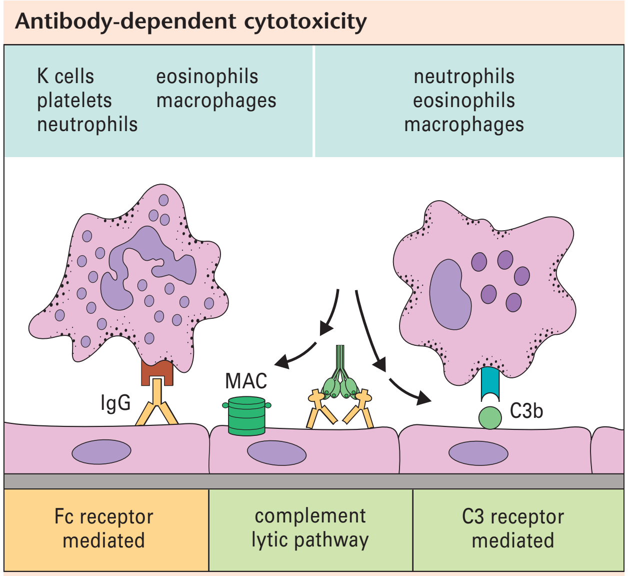

Fig. 24.1 Effector cells – K cells, platelets, neutrophils, eosinophils, and cells of the mononuclear phagocyte series – all have receptors for Fc, which they use to engage antibody bound to target tissues. Activation of complement C3 can generate complement-mediated lytic damage to target cells directly, and also allows phagocytic cells to bind to their targets via C3b, C3bi, or C3d, which also activate the cells. (MAC, membrane attack complex)

Once the antibody has attached itself to the surface of the cell or tissue, it can bind and activate complement component C1, with the following consequences:

- complement fragments (C3a and C5a) generated by activation of complement attract macrophages and polymorphs to the site, and also stimulate mast cells and basophils to produce chemokines that attract and activate other effector cells;
- the classical complement pathway and activation loop lead to the deposition of C3b, C3bi, and C3d on the target cell membrane;
- the classical complement pathway and lytic pathway result in the production of the C5b–9 membrane attack complex (MAC) and insertion of the complex into the target cell membrane.

Effector cells – in this case macrophages, neutrophils, eosinophils, and NK cells – bind to either:

- the complexed antibody via their Fc receptors; or
- the membrane-bound C3b, C3bi, and C3d, via their C3 receptors (CR1, CR3, CR4).

The mechanisms by which these antibodies trigger cytotoxic reactions in vivo have been investigated in FcR-deficient mice. Anti-red blood cell antibodies trigger erythrophagocytosis of IgG-opsonized red blood cells in an FcR-dependent manner. Fc receptor γ chain-deficient mice were protected from the pathogenic effect of these antibodies whereas complement-deficient mice were indistinguishable from wild-type animals in their ability to clear the targeted red cells.

Q. How do you interpret this observation?

A. These experiments indicate that phagocytosis of antibody-sensitized red cells depends on FcR-mediated binding to the phagocytes and is not dependent on opsonization by C3b deposited via the classical pathway.

Cells damage targets by exocytosis of their normal immune effector molecules

The mechanisms by which neutrophils and macrophages damage target cells in type II hypersensitivity reactions reflect their normal methods of dealing with infectious pathogens (Fig. 24.2).

Normally pathogens would be internalized and then subjected to a barrage of microbicidal systems including defensins, reactive oxygen and nitrogen metabolites, hypohalites, enzymes, altered pH, and other agents that interfere with metabolism (see Chapters 9 and 14).

If the target is too large to be phagocytosed, the granule and lysosome contents are released in apposition to the sensitized target in a process referred to as **exocytosis**. Cross-linking of the Fc and C3 receptors during this process causes activation of the phagocyte with production of reactive oxygen intermediates, as well as activation of phospholipase A2 with consequent release of arachidonic acid from membrane phospholipids.

Q. What inflammatory mediators are synthesized from arachidonic acid?

A. This metabolite is the precursor of eicosanoids – prostaglandins and leukotrienes (see Fig. 6.21).

Damage mechanisms

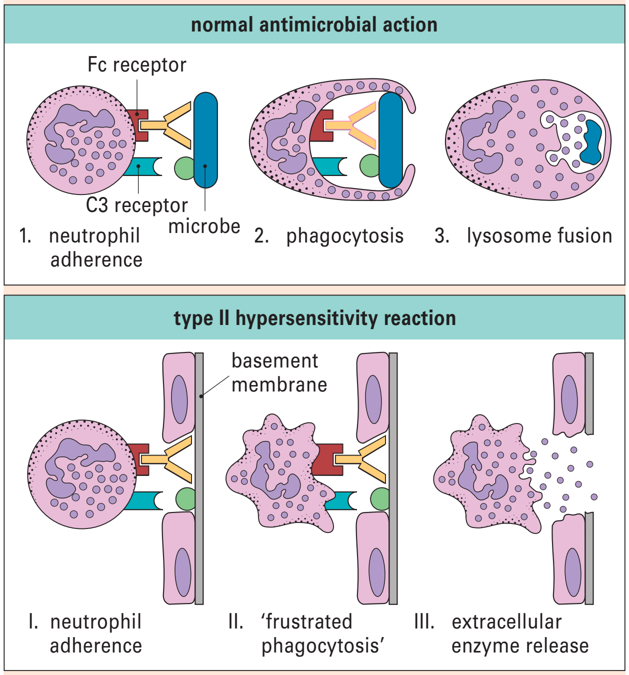

Fig. 24.2 Neutrophil-mediated damage is a reflection of normal antibacterial action. (**1**) Neutrophils engage microbes with their Fc and C3 receptors. (**2**) The microbe is then phagocytosed and destroyed as lysosomes fuse to form the phagolysosome (**3**). In type II hypersensitivity reactions, individual host cells coated with antibody may be similarly phagocytosed, but where the target is large, for example a basement membrane (**I**), the neutrophils are frustrated in their attempt at phagocytosis (**II**). They exocytose their lysosomal contents, causing damage to cells in the vicinity (**III**).

In some situations, such as the eosinophil reaction against schistosomes (see Chapter 15), exocytosis of granule contents is normal and beneficial. However, when the target is host tissue that has been sensitized by antibody, the result is damaging (Fig. 24.3).

Antibodies may also mediate hypersensitivity by NK cells. In this case, however, the nature of the target, and whether it can inhibit the NK cells' cytotoxic actions, are as important as the presence of the sensitizing antibody.

The resistance of a target cell to damage varies. Susceptibility depends on:

- the amount of antigen expressed on the target cell's surface; and
- the inherent ability of different target cells to sustain damage.

For example, an erythrocyte may be lyzed by a single active C5 convertase site, whereas it takes many such sites to destroy most nucleated cells – their ion-pumping capacity and ability to maintain membrane integrity with anti-complementary defenses is so much greater.

Phagocytes attacking a basement membrane

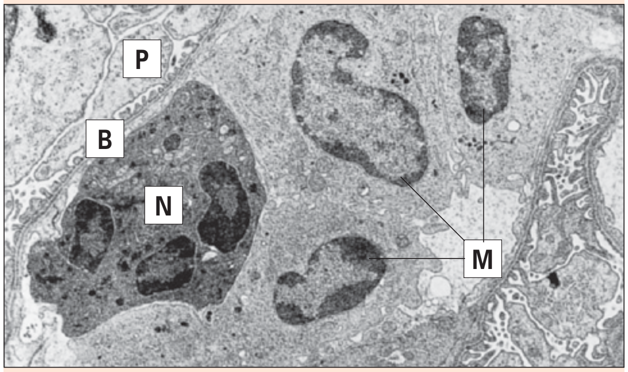

Fig. 24.3 This electron micrograph shows a neutrophil (N) and three monocytes (M) binding to the capillary basement membrane (B) in the kidney of a rabbit containing anti-basement membrane antibody. × 3500. (P, podocyte) (Courtesy of Professor GA Andres)

Q. What molecules protect the surface of nucleated cells from complement-mediated damage?

A. Decay accelerating factor (CD55) and CD59 (see Figs 4.9 and 4.11).

The rest of this chapter examines some of the instances where type II hypersensitivity reactions are thought to be of prime importance in causing target cell destruction or immunopathological damage.

TYPE II HYPERSENSITIVITY REACTIONS MAY TARGET BLOOD CELLS AND PLATELETS

Some of the most clearcut examples of type II reactions are seen in the responses to erythrocytes. Important examples are:

- incompatible blood transfusions, where the recipient becomes sensitized to antigens on the surface of the donor's erythrocytes;
- hemolytic disease of the newborn, where a pregnant woman has become sensitized to the fetal erythrocytes;
- autoimmune hemolytic anemias, where the patient becomes sensitized to his or her own erythrocytes.

Reactions to platelets can cause thrombocytopenia, and reactions to neutrophils and lymphocytes have been associated with systemic lupus erythematosus (SLE).

Transfusion reactions occur when a recipient has antibodies against donor erythrocytes

More than 20 blood group systems, generating over 200 genetic variants of erythrocyte antigens, have been identified in humans.

A blood group system consists of a gene locus that specifies an antigen on the surface of blood cells (usually, but not always, erythrocytes).

Within each system there may be two or more phenotypes. In the ABO system, for example, there are four phenotypes (A, B, AB, and O), and therefore four possible blood groups.

An individual with a particular blood group can recognize erythrocytes carrying allogeneic (non-self) blood group antigens, and will produce antibodies against them. However, for some blood group antigens such antibodies can also be produced 'naturally' (i.e. without previous sensitization by foreign erythrocytes, see below).

Transfusion of allogeneic erythrocytes into an individual who already has antibodies against them may produce erythrocyte destruction and symptoms of a 'transfusion reaction'.

Some blood group systems (e.g. ABO and Rhesus) are characterized by antigens that are relatively strong immunogens; such antigens are more likely to induce antibodies.

When planning a blood transfusion, it is important to ensure that donor and recipient blood types are compatible with respect to these major blood groups, otherwise transfusion reactions will occur.

Some major human blood groups are listed in Fig. 24.4.

The ABO blood group system is of primary importance

The epitopes of the **ABO blood group system** occur on many cell types in addition to erythrocytes and are located on the carbohydrate units of glycoproteins. The structure of these carbohydrates, and of those determining the

Five major blood group systems involved in transfusion reactions

system	gene loci	antigens	phenotype frequency (%)
ABO	1	A, B, or O	A 42 B 8 AB 3 0 47
Rhesus	2 closely linked loci: major antigen=RhD	C or c D or d E or e	RhD^+ 85 RhD^- 15
Kell	1	K or k	K 9 k 91
Duffy	1	Fy^a, Fy^b, or Fy	Fy^aFy^b 46 Fy^a 20 Fy^b 34 Fy 0.1
MN	1	M or N	MM 28 MN 50 NN 22

Fig. 24.4 Not all blood groups are equally antigenic in transfusion reactions – thus RhD evokes a stronger reaction in an incompatible recipient than the other Rhesus antigens; and Fy^a is stronger than Fy^b. Frequencies stated are for Caucasian populations – other races have different gene frequencies.

ABO blood group antigens

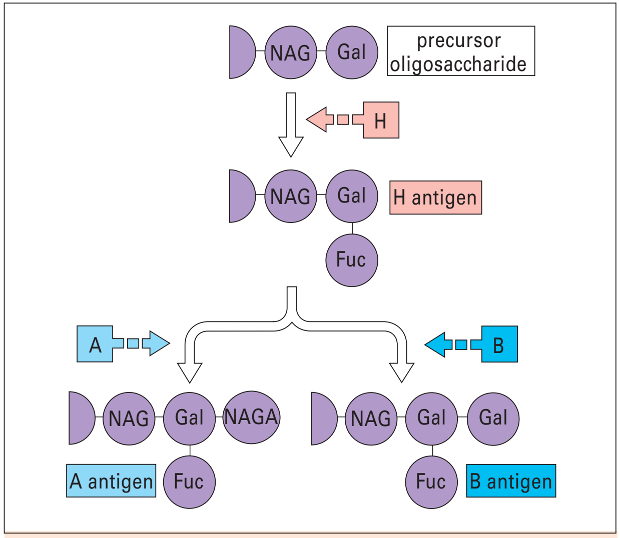

blood group (phenotype)	genotypes	antigens	antibodies to ABO in serum
A	AA, AO	A	anti-B
B	BB, BO	B	anti-A
AB	AB	A and B	none
O	OO	H	anti-A and anti-B

Fig. 24.5 The diagram shows how the ABO blood groups are constructed. The enzyme produced by the H gene attaches a fucose residue (Fuc) to the terminal galactose (Gal) of the precursor oligosaccharide. Individuals possessing the A gene now attach *N*-acetylgalactosamine (NAGA) to this galactose residue, whereas those with the B gene attach another galactose, producing A and B antigens, respectively. People with both genes make some of each. The table indicates the genotypes and antigens of the ABO system. Most people naturally make antibodies to the antigens they lack. (NAG, *N*-acetylglucosamine)

related Lewis blood group system, is determined by genes coding for enzymes that transfer terminal sugars to a carbohydrate backbone (Fig. 24.5).

Most individuals develop antibodies to allogeneic specificities of the ABO system without previous sensitization by foreign erythrocytes. This sensitization occurs through contact with identical epitopes, coincidentally and routinely expressed on a wide variety of microorganisms.

Antibodies to ABO antigens are therefore extremely common, making it particularly important to match donor blood to the recipient for this system. However, all people are tolerant to the O antigen, so O individuals are **universal donors** with respect to the ABO system.

The Rhesus system is a major cause of hemolytic disease of the newborn

The Rhesus system is also of great importance because it is a major cause of **hemolytic disease of the newborn (HDNB)**.

Erythrocyte blood group antigens

erythrocyte surface glycoprotein	blood groups expressed	number of epitopes per cell
anion transport protein	ABO, Ii	10^6
glycophorin A	MN	10^6
glucose transporter	ABO, Ii	5×10^5
Mr 45 000–100 000	ABO	
Mr 30 000	ABO, Rh	1.2×10^5
glycophorin B	N, Ss	2.5×10^5
glycophorins C & D	Gerbich (Ge)	10^5
DAF	Cromer	<10 000
CD44 (80 kDa)	Ina/Inh	3000–6000
zinc endopeptidase	Kell	3000–6000
DARC	Fy	12 000
laminin-binding glycoprotein	lutheran	1500–4000

Fig. 24.6 Note that blood group epitopes based on carbohydrate moieties, such as ABO and Ii (expressed on the precursor of the ABO polysaccharide), can appear on many different proteins, including Rhesus antigens. Antigens such as Rhesus and Duffy are proteins, so the epitope appears on only one type of molecule. In general, the most important blood group antigens are present at high levels on the erythrocytes, thus providing plenty of targets for complement-mediated lysis or Fc receptor-mediated clearance. (DAF, decay accelerating factor; DARC, Duffy antigen receptor for chemokines)

Rhesus antigens are associated with membrane proteins of 30 kDa, which are expressed at moderate levels on the erythrocyte surface. The antigens are encoded by two closely linked loci, RhD and RhCcEe, with 92% homology.

RhD is the most important clinically due to its high immunogenicity, but in RhD^- individuals the RhD locus is missing completely. The RhCcEe locus encodes a molecule that expresses the RhC/c and RhE/e epitopes.

Transfusion reactions can be caused by minor blood groups

MN system epitopes are expressed on the *N*-terminal glycosylated region of glycophorin A, a glycoprotein present on the erythrocyte surface. Antigenicity is determined by polymorphisms at amino acids 1 and 5.

The related **Ss system** antigens are carried on glycophorin B.

Proteins expressed on erythrocytes that display allelic variation can also act as blood group antigens. Examples of these include:

- the Kell antigen, a zinc endopeptidase;
- the Duffy antigen receptor for chemokines (DARC); and
- variants of decay accelerating factor (DAF).

The relationship of the blood groups to erythrocyte surface proteins is listed in Fig. 24.6.

Transfusion reactions caused by the minor blood groups are relatively rare unless repeated transfusions are given. The risks are greatly reduced by accurately cross-matching the donor blood to that of the recipient.

Cross-matching ensures that a recipient does not have antibodies against donor erythrocytes

The aim of cross-matching is to ensure that the blood of a recipient does not contain antibodies that will be able to react with and destroy transfused (donor) erythrocytes. For example:

- antibodies to ABO system antigens cause incompatible cells to agglutinate in a clearly visible reaction;
- minor blood group systems cause weaker reactions that may only be detectable by an indirect Coombs' test (see Fig. 24.10).

If the individual is transfused with whole blood, it is also necessary to check that the donor's serum does not contain antibodies against the recipient's erythrocytes. However, transfusion of whole blood is unusual – most blood donations are separated into cellular and serum fractions, to be used individually.

Transfusion reactions involve extensive destruction of donor blood cells

Transfusion of erythrocytes into a recipient who has antibodies to those cells produces an immediate reaction. The symptoms include:

- fever;
- hypotension;
- nausea and vomiting; and
- pain in the back and chest.

The severity of the reaction depends on the class and the amounts of antibodies involved:

Antibodies to ABO system antigens are usually IgM, and cause agglutination, complement activation, and intravascular hemolysis.

Other blood groups induce IgG antibodies, which cause less agglutination than IgM. The IgG-sensitized cells are usually taken up by phagocytes in the liver and spleen, though severe reactions may cause erythrocyte destruction by complement activation. This can cause circulatory shock, and the released contents of the erythrocytes can produce acute tubular necrosis of the kidneys.

These transfusion reactions are often seen in previously unsensitized individuals and develop over days or weeks as antibodies to the foreign cells are produced. This can result in anemia or jaundice.

Q. Why do antibodies to the ABO blood group that are normally present in a mother not damage the erythrocytes that are present in a fetus she is carrying when the fetus has a different blood group?

A. Antibodies to the ABO blood group are usually IgM, which is not transported across the placenta (see Fig. 3.19).

Transfusion reactions to other components of blood may also occur, but their consequences are not usually as severe as reactions to erythrocytes.

Hyperacute graft rejection is related to the transfusion reaction

Hyperacute graft rejection occurs when a graft recipient has preformed antibodies against the graft tissue. It is only seen in tissue that is revascularized directly after transplantation, in kidney grafts for example.

The most severe reactions in this type of rejection are due to the ABO group antigens expressed on kidney cells. The damage is produced by antibody and complement activation in the blood vessels, with consequent recruitment and activation of neutrophils and platelets.

Donors and recipients are now always cross-matched for ABO antigens, and this reaction has become extremely rare. Antibodies to other graft antigens (e.g. MHC molecules) induced by previous grafting can also produce this type of reaction.

HDNB is due to maternal IgG reacting against the child's erythrocytes in utero

Hemolytic disease of the newborn (HDNB) occurs when the mother has been sensitized to antigens on the infant's erythrocytes and makes IgG antibodies to these antigens. These antibodies cross the placenta and react with the fetal erythrocytes, causing their destruction (Figs 24.7 and 24.8). Rhesus D (RhD) is the most commonly involved antigen.

A risk of HDNB arises when a Rh^+-sensitized Rh^- mother carries a second Rh^+ infant. Sensitization of the

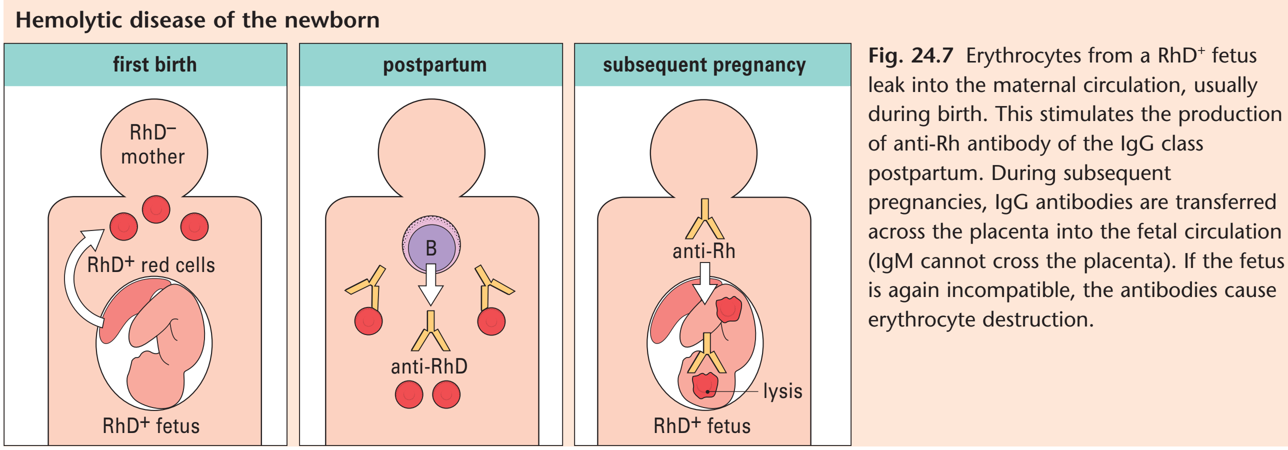

Fig. 24.7 Erythrocytes from a RhD^+ fetus leak into the maternal circulation, usually during birth. This stimulates the production of anti-Rh antibody of the IgG class postpartum. During subsequent pregnancies, IgG antibodies are transferred across the placenta into the fetal circulation (IgM cannot cross the placenta). If the fetus is again incompatible, the antibodies cause erythrocyte destruction.

A child with hemolytic disease of the newborn

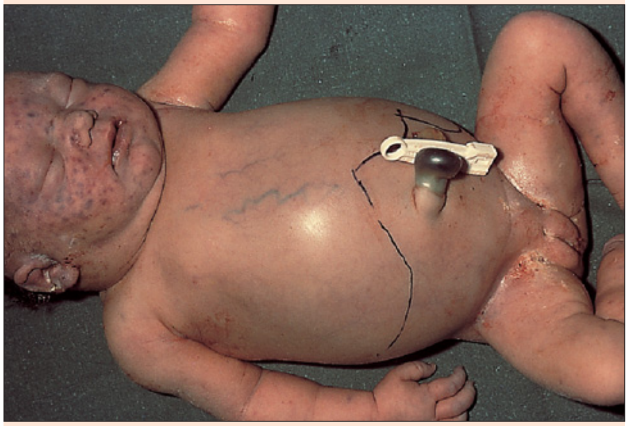

Fig. 24.8 There is considerable enlargement of the liver and spleen associated with erythrocyte destruction caused by maternal anti-erythrocyte antibody in the fetal circulation. The child had elevated bilirubin (breakdown product of hemoglobin). The facial petechial hemorrhaging was due to impaired platelet function. The most commonly involved antigen is RhD. (Courtesy of Dr K Sloper)

Rh^- mother to the Rh^+ erythrocytes usually occurs during the birth of the first Rh^+ infant, when some fetal erythrocytes leak back across the placenta into the maternal circulation and are recognized by the maternal immune system. The first incompatible child is therefore usually unaffected, whereas subsequent children have an increasing risk of being affected, as the mother is resensitized with each successive pregnancy.

Reactions to other blood groups may also cause HDNB, the second most common being the Kell system K antigen. Reactions due to anti-K are much less common than reactions due to RhD because of the relatively low frequency (9%) and weaker antigenicity of the K antigen.

The risk of HDNB due to Rhesus incompatibility is known to be reduced if the father is of a different ABO group to the mother. This observation led to the idea that these Rh^- mothers were destroying Rh^+ cells more rapidly because they were also ABO incompatible. Consequently, fetal Rh^+ erythrocytes would not be available to sensitize the maternal immune system to RhD antigen.

This notion led to the development of **Rhesus prophylaxis** – preformed anti-RhD antibodies are given to Rh^- mothers immediately after delivery of Rh^+ infants, with the aim of destroying fetal Rh^+ erythrocytes before they can cause Rh^- sensitization. This practice has successfully reduced the incidence of HDNB due to Rhesus incompatibility (Fig. 24.9). Although the number of cases of HDNB has fallen dramatically and progressively, the proportion of cases caused by other blood groups, including Kell and the ABO system, has increased.

Autoimmune hemolytic anemias arise spontaneously or may be induced by drugs

Reactions to **blood group antigens** also occur spontaneously in the autoimmune hemolytic anemias, in which patients produce antibodies to their own erythrocytes.

Autoimmune hemolytic anemia is suspected if a patient gives a positive result on a **direct antiglobulin test** (Fig. 24.10), which identifies antibodies present on the patient's erythrocytes. These are usually antibodies directed towards erythrocyte antigens, or immune complexes adsorbed onto the erythrocyte surface.

The direct antiglobulin test is also used to detect antibodies on red cells in mismatched transfusions, and in HDNB (see above).

Autoimmune hemolytic anemias can be divided into three types, depending upon whether they are caused by:

- warm-reactive autoantibodies, which react with the antigen at 37°C;
- cold-reactive autoantibodies, which can only react with antigen at below 37°C;
- antibodies provoked by allergic reactions to drugs.

Warm-reactive autoantibodies cause accelerated clearance of erythrocytes

Warm-reactive autoantibodies are frequently found against Rhesus system antigens, including determinants of the RhC and RhE loci as well as RhD. They differ from

Rhesus prophylaxis

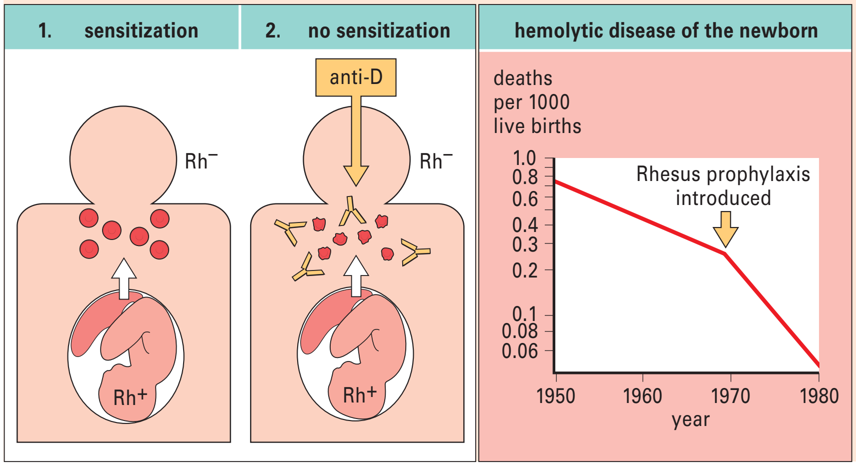

Fig. 24.9 (**1**) Without prophylaxis, Rh^+ erythrocytes leak into the circulation of a Rh^- mother and sensitize her to the Rh antigen(s). (**2**) If anti-Rh antibody (anti-D) is injected immediately postpartum it eliminates the Rh^+ erythrocytes and prevents sensitization. The incidence of deaths due to HDNB fell during the period 1950–1966 with improved patient care. The decline in the disease was accelerated by the advent of Rhesus prophylaxis in 1969.

Direct antiglobulin test

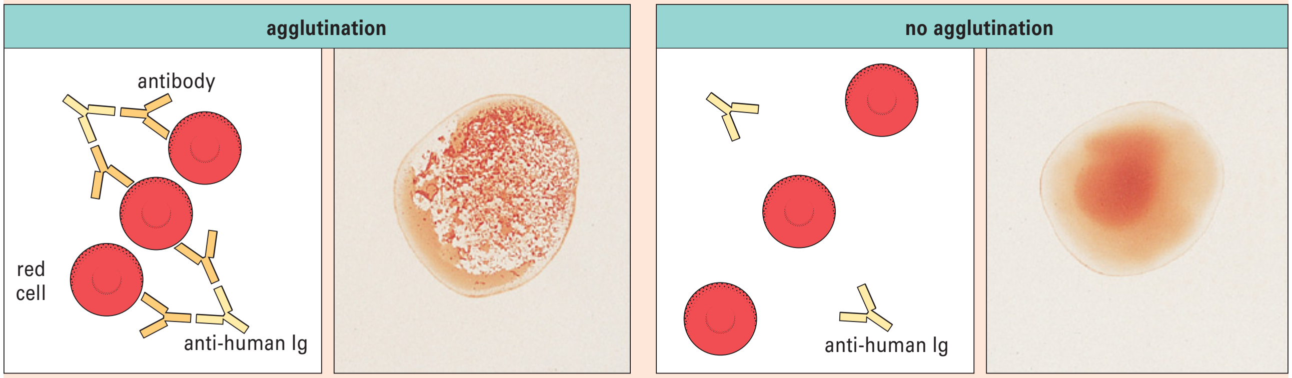

Fig. 24.10 This test, also called a Coombs' test, is used to detect antibody on a patient's erythrocytes. If antibody is present the erythrocytes can be agglutinated by anti-human immunoglobulin. If no antibody is present on the red cells, they are not agglutinated by anti-human immunoglobulin.

the antibodies responsible for transfusion reactions in that they appear to react with different epitopes.

Warm-reactive autoantibodies to other blood group antigens exist, but are relatively rare.

Most of these hemolytic anemias are of unknown cause, but some are associated with other autoimmune diseases.

Q. What mechanisms could lead to the anemia associated with these autoantibodies?

A. Fc receptor-mediated erythrophagocytosis or complement-dependent lysis.

The anemia appears to be a result of accelerated clearance of the sensitized erythrocytes by spleen macrophages more often than being caused by complement-mediated lysis.

Cold-reactive autoantibodies cause erythrocyte lysis by complement fixation

Cold-reactive autoantibodies are often present in higher titers than the warm-reactive autoantibodies. The antibodies are primarily IgM and fix complement strongly. In most cases they are specific for the Ii blood group system. The I and i epitopes are expressed on the precursor polysaccharides that produce the ABO system epitopes, and are the result of incomplete glycosylation of the core polysaccharide.

The reaction of the antibody with the erythrocytes takes place in the peripheral circulation (particularly in winter), where the temperature in the capillary loops of exposed skin may fall below 30°C. In severe cases, peripheral necrosis may occur due to aggregation and microthrombosis of small vessels caused by complement-mediated destruction in the periphery.

The severity of the anemia is therefore directly related to the complement-fixing ability of the patient's serum. (Fc-mediated removal of sensitized cells in the spleen and liver is not involved because these organs are too warm for the antibodies to bind.)

Most cold-reactive autoimmune hemolytic anemias occur in older people. Their cause is unknown, but it is notable that the autoantibodies produced are usually of very limited clonality.

Q. Why would you expect the autoimmune response to Ii antigens to be of limited clonality? There are two reasonable explanations, which are not mutually exclusive.

A. The numbers of autoreactive clones may have been curtailed by negative selection during B cell development. In addition, the antigens are carbohydrate and induce an IgM response (T independent) which has not been subject to diversification.

However, some cases may follow infection with *Mycoplasma pneumoniae*, and these are acute-onset diseases of short duration with polyclonal autoantibodies. Such cases are thought to be due to cross-reacting antigens on the bacteria and the erythrocytes, producing a bypass of normal tolerance mechanisms (see Chapter 19).

Drug-induced reactions to blood components occur in three different ways

Drugs (or their metabolites) can provoke hypersensitivity reactions against blood cells, including erythrocytes and platelets. This can occur in three different ways (Fig. 24.11).

- The drug binds to the blood cells and antibodies are produced against the drug – In this case it is necessary for both the drug and the antibody to be present to produce the reaction. This phenomenon was first recorded by Ackroyd, who noted thrombocytopenic purpura (destruction of platelets leading to purpuric rash) following administration of the drug Sedormid.

  Hemolytic anemias have been reported following the administration of a wide variety of drugs, including:
  - penicillin;
  - quinine; and
  - sulfonamides.

  All these conditions are rare.

Drug-induced reactions to blood cells

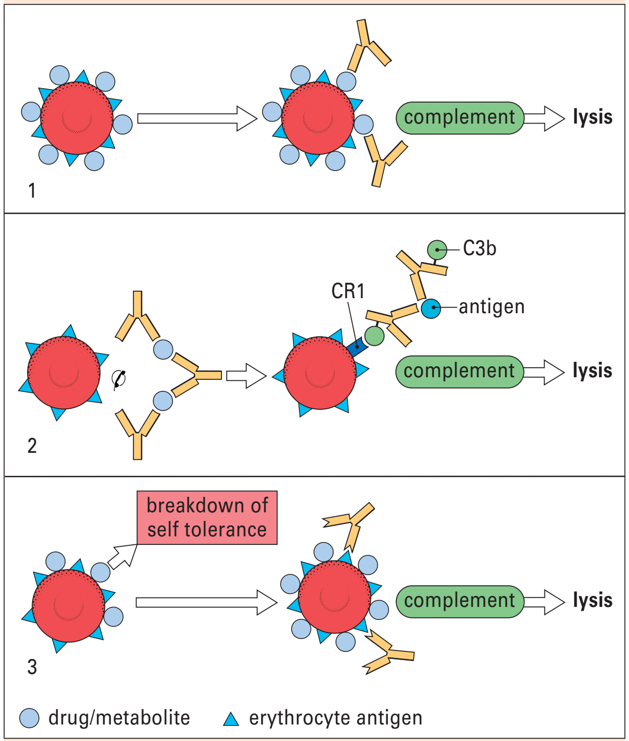

Fig. 24.11 Three ways that drug treatment can cause damage are illustrated. (**1**) The drug adsorbs to cell membranes. Antibodies to the drug bind to the cell and complement-mediated lysis occurs. (**2**) Immune complexes of drugs and antibody become adsorbed to the red cell. This could be mediated by an Fc receptor, but is more probably via the C3b receptor CR1. Damage occurs by complement-mediated lysis. (**3**) Drugs, presumably adsorbed onto cell membranes, induce a breakdown of self tolerance, possibly by stimulating T_H cells. This leads to the formation of antibodies to other blood group antigens on the cell surface. Note that in examples 1 and 2 the drug must be present for cell damage to occur, whereas in 3 the cells are destroyed whether they carry adsorbed drug or not.

- Drug–antibody immune complexes are adsorbed onto the erythrocyte cell membrane – When drug–antibody immune complexes are adsorbed on to the erythrocyte cell membrane damage occurs by complement-mediated lysis.
- The drug induces an allergic reaction – The drug induces an allergic reaction and autoantibodies are directed against the erythrocyte antigens themselves, as occurs in 0.3% of patients given a-methyldopa. The antibodies produced are similar to those in patients with warm-reactive antibody. However, the condition remits shortly after the cessation of drug treatment.

Reactions against other blood cells have been associated with SLE and with thrombocytopenia

Autoantibodies to neutrophil cytoplasmic antigens (ANCAs) are associated with a number of diseases. For example:

- antibody to proteinase-3, a cytoplasmic antigen (c-ANCA), is associated with Wegener's granulomatosis;
- antibodies to myeloperoxidase are seen more commonly in systemic lupus erythematosus (SLE), and are located in perinuclear granules (p-ANCA).

Other granule components may also act as antigens in SLE, but less commonly than myeloperoxidase. Such autoantigens are generally neutrophil specific and antibodies can be detected by immunofluorescent staining (Fig. 24.12).

By contrast, antibodies to MHC antigens, which are also seen in SLE, are highly non-tissue specific.

p-ANCAs are particularly characteristic of vasculitis and glomerulonephritis. Their contribution to disease pathogenesis appears to be relatively small, although they have some diagnostic use.

Autoantibodies to platelets are seen in ITP

Autoantibodies to platelets are seen in up to 70% of cases of **idiopathic thrombocytopenic purpura** (ITP), a disorder in which there is accelerated removal of platelets from the circulation, mediated primarily by splenic macrophages. The mechanism of removal is via the immune adherence receptors on these cells.

ITP most often develops after bacterial or viral infections, but may also be associated with autoimmune diseases including SLE.

Antibodies to neutrophils

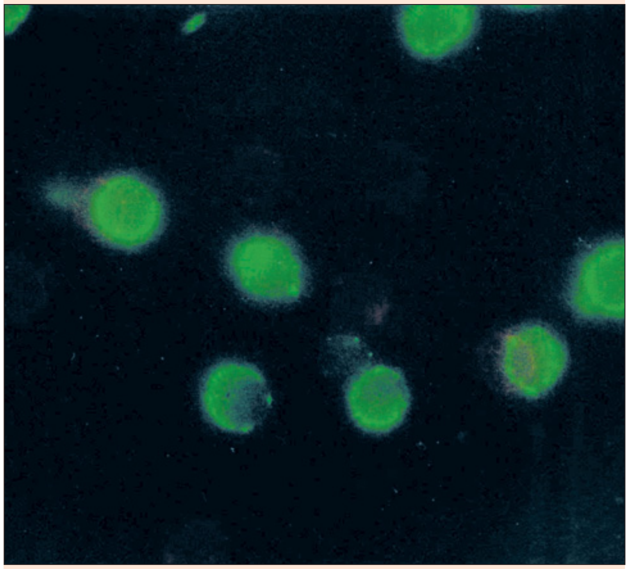

Fig. 24.12 Antibodies to neutrophils are demonstrated by immunofluorescence. In this instance, normal neutrophils were stained with antibodies from a neonate suffering from alloimmune neonatal neutropenia.

In SLE, antibodies to cardiolipin, which is present on platelets, can sometimes be detected. Autoantibodies to cardiolipin and other phospholipids can inhibit one aspect of blood clotting (**lupus anticoagulant**) and can be associated, in some cases, with venous thrombosis and recurrent abortions.

Thrombocytopenia may also be induced by drugs by similar mechanisms to those outlined in Fig. 24.11.

TYPE II HYPERSENSITIVITY REACTIONS TARGET TISSUES

A number of autoimmune conditions occur in which antibodies to tissue antigens cause immunopathological damage by activation of type II hypersensitivity mechanisms. The antigens are extracellular, and may be expressed on structural proteins or on the surface of cells. The resulting diseases are discussed here and include:

- Goodpasture's syndrome;
- pemphigus;
- myasthenia gravis.

It is often possible to demonstrate autoantibodies to particular cell types, but in these cases the antigens are intracellular, and the importance of the type II mechanisms is less well established. In these cases, recognition of autoantigen by T cells is probably more important pathologically, and the autoantibodies are of secondary importance.

Antibodies against basement membranes produce nephritis in Goodpasture's syndrome

A number of patients with nephritis are found to have **antibodies to collagen type IV**, which is a major component of basement membranes (see Fig. 25.3).

Collagen type IV undergoes alternate RNA splicing, which produces a number of variant proteins (**Goodpasture antigen**), but the antibodies appear to bind just those forms that retain the characteristic N terminus. The antibody is usually IgG and, in at least 50% of patients, it appears to fix complement.

Goodpasture's syndrome usually results in severe necrosis of the glomerulus, with fibrin deposition. The association of this type of nephritis with lung hemorrhage was originally noticed by Goodpasture (hence Goodpasture's syndrome). Although the lung symptoms do not occur in all patients, the association of lung and kidney damage is due to cross-reactive autoantigens in the basement membranes of the two tissues.

Animal models for Goodpasture's syndrome have been developed

In nephrotoxic serum nephritis (Masugi glomerulonephritis), heterologous antibodies to glomerular basement membrane are injected into rats or rabbits. The injected antibody is deposited onto the basement membranes, and this is followed by further deposition of host antibodies to the injected antibody – this precipitates acute nephritis. Development of nephritis and proteinuria depends on the accumulation of neutrophils, which bind via complement-dependent and complement-independent mechanisms. Similar lesions can be induced by immunization with heterologous basement membrane (Steblay model).

Another animal model (Heymann nephritis), caused by raising autoantibodies to a protein present in the brush border of glomerular epithelial cells, resembles human membranous glomerulonephritis. In this model, the damage is mostly complement mediated – complement depletion of the animals alleviates the condition.

Pemphigus is caused by autoantibodies to an intercellular adhesion molecule

Pemphigus is a serious blistering disease of the skin and mucous membranes. Patients have **autoantibodies against desmoglein-1 and desmoglein-3**, components of desmosomes, which form junctions between epidermal cells (Fig. 24.13). The antibodies disrupt cellular adhesion, leading to breakdown of the epidermis.

Clinical disease profiles can be related to the specificity of the antibodies. For example:

- patients with only anti-desmoglein-3 tend to show mucosal disease;
- those with anti-desmoglein-1 and anti-desmoglein-3 have skin and mucosal involvement.

Disease has been correlated with the incidence of IgG4 antibodies against a different part of the molecule.

Pemphigus is strongly linked to a rare haplotype of HLA-DR4 (DRB1*0402), and this molecule has been shown to present a peptide of desmoglein-3, which other DR4 subtypes cannot. This is therefore a clear example of an autoimmune disease producing pathology by type II mechanisms.

Autoantibodies in pemphigus

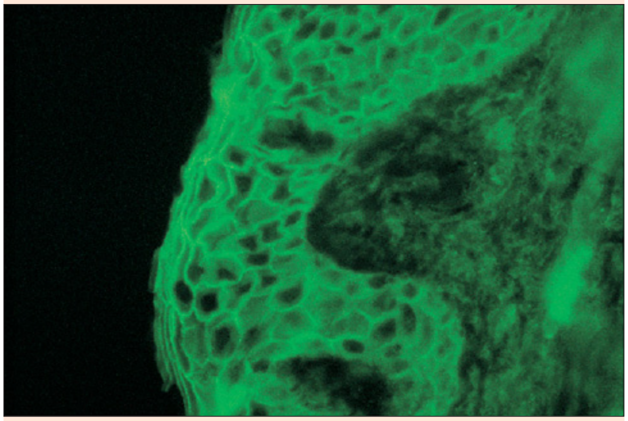

Fig. 24.13 The antibodies in pemphigus bind to components of the desmosome involved in cell adhesion. Desmoglein-1 and desmoglein-3 are most commonly involved, but other molecules, including the plakins and desmocollin, may also act as autoantigens. Immunofluorescence of human skin stained with anti-IgA. (Courtesy of Dr R Mirakian and Mr P Collins)

In myasthenia gravis and Lambert–Eaton syndrome autoantibodies to acetylcholine receptors cause muscle weakness

Myasthenia gravis, a condition in which there is extreme muscular weakness, is associated with **antibodies to the acetylcholine receptors** on the surface of muscle membranes. The acetylcholine receptors are located at the motor endplate where the neuron contacts the muscle. Transmission of impulses from the nerve to the muscle takes place by the release of acetylcholine from the nerve terminal and its diffusion across the gap to the muscle fiber.

It was noticed that immunization of experimental animals with purified acetylcholine receptors produced a condition of muscular weakness that closely resembled human myasthenia. This suggested a role for antibody to the acetylcholine receptor in the human disease.

Analysis of the lesion in myasthenic muscles indicated that the disease was not due to an inability to synthesize acetylcholine, nor was there any problem in secreting it in response to a nerve impulse – the released acetylcholine was less effective at triggering depolarization of the muscle (Fig. 24.14).

Examination of neuromuscular endplates by immunochemical techniques has demonstrated IgG and the complement proteins, C3 and C9, on the postsynaptic folds of the muscle (Fig. 24.15).

Further evidence for a pathogenetic role for IgG in this disease was furnished by the discovery of transient muscle weakness in babies born to mothers with myasthenia gravis. This is significant because it is known that IgG can and does cross the placenta, entering the blood stream of the fetus.

IgG and complement are thought to act in two ways:

- by increasing the rate of turnover of the acetylcholine receptors; and
- by partial blocking of acetylcholine binding.

Cellular infiltration of myasthenic endplates is rarely seen, so it is assumed that damage does not involve effector cells.

Lambert–Eaton syndrome is caused by defective release of acetylcholine

In a Lambert–Eaton syndrome, a condition related to myasthenia gravis, muscular weakness is caused by defective release of acetylcholine from the neuron. If serum or

Myasthenia gravis

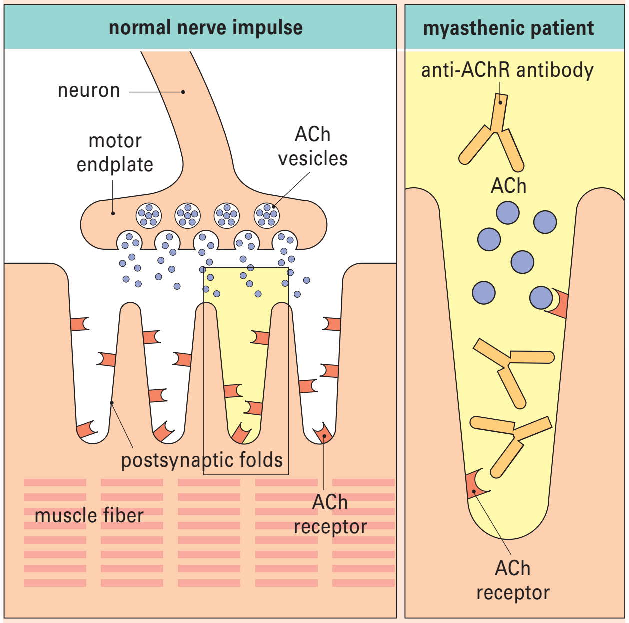

Fig. 24.14 Normally a nerve impulse passing down a neuron arrives at a motor endplate and causes the release of acetylcholine (ACh). This diffuses across the neuromuscular junction, binds ACh receptors (AChR) on the muscle, and causes ion channels in the muscle membrane to open, which in turn triggers muscular contraction. In myasthenia gravis, antibodies to the receptor block binding of the ACh transmitter. The effect of the released vesicle is therefore reduced, and the muscle can become very weak. Antibody blocking receptors are only one of the factors operating in the disease.

Motor endplate in myasthenia gravis

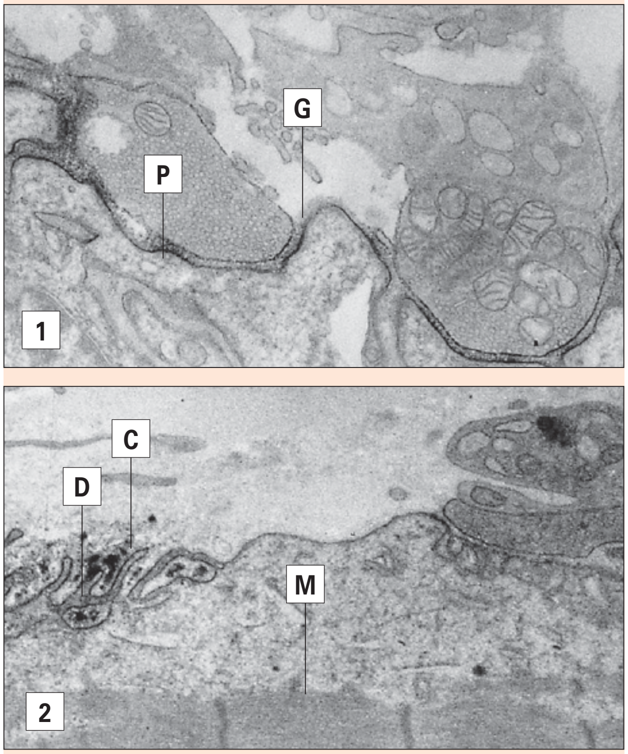

Fig. 24.15 (**1**) Electron micrograph showing IgG deposits (G) in discrete patches on the postsynaptic membrane (P). × 13 000. (**2**) Electron micrograph illustrating C9 (C) shows the postsynaptic region denuded of its nerve terminal – it consists of debris and degenerating folds (D). There is a strong reaction for C9 on this debris. × 9000. (M, muscle fiber) (Courtesy of Dr AG Engel)

IgG from patients with Lambert–Eaton syndrome is transfused into mice, the condition is also transferred, indicating the presence of autoantibody. The autoantibodies are directed against components of **voltage-gated calcium channels or the synaptic vesicle protein synaptotagmin**.

The different forms of Lambert–Eaton syndrome are thought to relate to the target antigen and the class and titer of antibodies involved.

Q. Predict whether inhibitors of acetylcholinesterase will have any value in alleviating the symptoms of myasthenia gravis or Lambert–Eaton syndrome.

A. These drugs are often useful in myasthenia gravis because acetylcholine is present and an inhibitor of acetylcholinesterase will prolong its activity. They are less useful in Lambert–Eaton syndrome because acetylcholine is not released in the first place.

Autoantibodies may block receptors or mimic receptors agonists

Myasthenia gravis and Lambert–Eaton syndrome exemplify conditions where autoantibodies to receptors block the normal function of the receptor. There are other diseases, however, where the autoantibody has an opposite effect, for example in some forms of autoimmune thyroid disease antibodies to the thyroid-stimulating hormone (TSH) receptor mimic TSH, thereby stimulating thyroid function (see Chapter 21).

AUTOANTIBODIES MAY BE PRESENT IN THE ABSENCE OF DISEASE

Although many autoantibodies react with tissue antigens, their significance in causing tissue damage and pathology in vivo is not always clear. For example, although autoantibodies to pancreatic islet cells can be detected in vitro using sera from some diabetic patients (Fig. 24.16), most of the immunopathological damage in autoimmune diabetes is thought to be caused by autoreactive T cells.

Islet cell autoantibodies

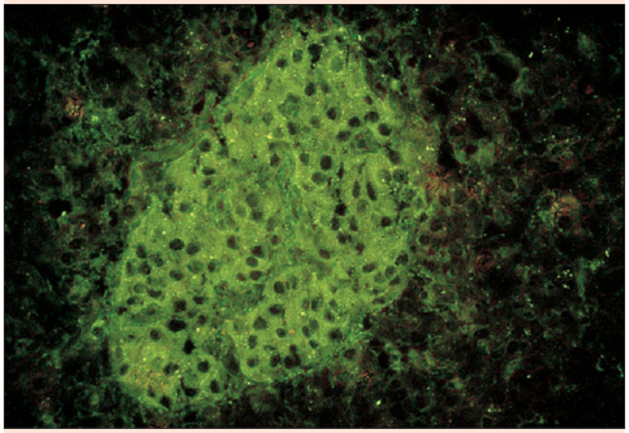

Fig. 24.16 Autoantibodies to the pancreas in diabetes mellitus may be demonstrated by immunofluorescence. The antibodies are diagnostically useful, and may contribute to the pathology. (Courtesy of Dr B Dean)

Until recently it was thought that autoantibodies against intracellular antigens would not usually cause immunopathology because they could not reach their antigen within a living cell. However, it now appears that antibodies such as anti-ribonucleoprotein (anti-RNP) and anti-DNA can reach the cell nucleus and modulate cell function – in some cases, they can induce apoptosis.

Although the relative importance of antibody in causing cell damage is still debated, autoantibodies against internal antigens of cells often make excellent disease markers because they are frequently detectable before immunopathological damage occurs.

CYTOTOXIC ANTIBODIES ARE INCREASINGLY USED FOR THERAPEUTIC INDICATIONS

It is now realized that high-specificity monoclonal antibodies can be used to induce cell death as a therapeutic procedure mediated via type II hypersensitivity mechanisms.

Some of antibodies increasingly being used for therapeutic indications work by blocking signaling pathways important in disease pathophysiology, for example:

- the **antihuman epidermal growth factor receptor 2 (anti-HER2) antibody, trastuzumab,** used in breast cancer; and
- the **anti-TNF antibody, infliximab,** used in Crohn's disease.

However, of relevance to type II hypersensitivity reactions, other monoclonal antibodies work by inducing antibody-dependent cell-mediated cytotoxicity (ADCC) and/or activation of complement pathways. The most widely used of these is the **anti-CD20 antibody, rituximab**. CD20 is expressed on normal B cells as well as most B cell malignancies and rituximab is now an established part of the treatment of B lineage non-Hodgkin's lymphomas.

Anti-CD20 therapy is also proving useful for treatment of autoimmune disease resistant to standard therapies including immune thrombocytopenia, rheumatoid arthritis, and SLE, where its activity is due to elimination of host autoreactive B cells.

Other antibodies in clinical use include the **anti-CD52 antibody, alemtuzumab**, which targets an antigen found on immune cells (both B and T cells). Alemtuzumab is used for the treatment of chronic B and T cell leukemias as well as for immune suppression in patients undergoing allogeneic hematopoietic stem cell transplantation.

FURTHER READING

Alarçon-Segovia D, Ruiz-Argüelles A, Llorente L. Broken dogma: penetration of autoantibodies into living cells. Immunol Today 1996;17:163–164.

Amagai M. Autoantibodies against desmosomal cadherins in pemphigus. J Dermatol Sci 1999;20:92–102.

Anstee DJ. Blood group active substances of the human red blood cell. Vox Sang 1990;58:1.

Bhol K, Natarajan K, Nagarwalla N, et al. Correlation of peptide specificity and IgG subclass with pathogenic and non-pathogenic auto-antibodies in pemphigus vulgaris: a model for autoimmunity. Proc Natl Acad Sci USA 1995;92:5239–5243.

Black M, Mignogna MD, Scully C. Pemphigus vulgaris. Oral Dis 2005;11:119–130.

Bloy C, Blanchard D, Lambin P, et al. Characterization of the D, C, E and G antigens of the Rh blood group system with human monoclonal antibodies. Mol Immunol 1988;25:926–930.

Dean FG, Wilson GR, Li M, Edgtton KL, et al. Experimental autoimmune Goodpasture's disease: a pathogenetic role for both effector cells and antibody injury. Kidney Int 2005;67:566–575.

Druet P, Glotz D. Experimental autoimmune nephropathies: induction and regulation. Adv Nephrol 1984;13:115.

Engelfriet CP, Reesink HW, Judd WJ, et al. Current status of immunoprophylaxis with anti-D immunoglobulin. Vox Sang 2003;85:328–337.

Le van Kim C, Mouro I, Cherif-Zahar B, et al. Molecular cloning and primary structure of the human blood group RhD polypeptide. Proc Natl Acad Sci USA 1992;89:10925–10929.

King MJ. Blood group antigens on human erythrocytes – distribution, structure and possible functions. Biochim Biophys Acta 1994;1197:14–44.

Lang B, Newsom-Davis J. Immunopathology of the Lambert–Eaton myasthenic syndrome. Springer Semin Immunopathol 1995;17:3–15.

Lindstrom J. Immunobiology of myasthenia gravis, experimental autoimmune myasthenia gravis and Lambert–Eaton syndrome. Annu Rev Immunol 1985;3:109–131.

Mauro I, Colin Y, Chenif-Zahar B, et al. Molecular genetic basis of the human Rhesus blood group system. Nature Genet 1993;5:62–65.

Naparstek Y, Plotz PH. The role of autoantibodies in autoimmune disease. Annu Rev Immunol 1993;11:79–104.

Race R, Sanger R. Blood groups in man. 6th edn. Oxford: Blackwell Scientific Publications; 1975.

Russo D, Redman C, Lee S. Association of XK and Kell blood group proteins. J Biol Chem 1998;273:13950–13956.

Schulz DR, Tozman EC. Anti-neutrophil cytoplasmic antibodies: major autoantigens, pathophysiology, and disease associations. Semin Arthritis Rheum 1995;25:143–159.

Vincent A. Antibody-mediated disorders of neuro-muscular transmission. Clin Neurophysiol Suppl 2004;57:147–158.

Yamamoto F-I, Clausen H, White T, et al. Molecular genetic basis of the histo-blood group ABO system. Nature 1990;345:229.

Critical thinking: Blood groups and hemolytic disease of the newborn (see p. 502 for explanations)

Mrs Chareston has the blood group O, Rhesus negative, and her husband Mr Chareston is A, Rhesus positive. They have had four children, of which two have been affected by hemolytic disease of the newborn (HDNB), as follows:

- first child born 1968 – unaffected;
- second child born 1974 – mildly affected;
- third child born 1976 – seriously affected, required intrauterine blood transfusion;
- fourth child born 1980 – unaffected.

In both affected cases (second and third), the cause of the hemolytic disease was identified as antibodies to Rhesus D binding to the child's red cells. Following the second, third, and fourth deliveries, Mrs Chareston was given antibodies to the Rhesus D blood group (Rhesus prophylaxis was introduced in the UK in 1972).

1. From this information, what can you deduce about the blood group of the first child?
2. Why does HDNB usually become more serious with successive pregnancies?
3. What is the reason for giving anti-Rhesus D antibodies to the mother?
4. Why are the antibodies given postpartum and not earlier?
5. Give an explanation of why the Rhesus prophylaxis after the second delivery failed to prevent HDNB in the third child.
6. What explanation can be given to account for the fact that the fourth child is unaffected?

When the blood groups of the children are examined it is found that they are:

- first child – O, Rh^+;
- second child – B, Rh^+;
- third child – A, Rh^+;
- fourth child – A, Rh^-.

7. As Mrs Chareston has antibodies to blood group A, why was the fourth child not affected by HDNB caused by these antibodies?
8. One of these children was definitely not fathered by Mr Chareston – which child?

CHAPTER

25

Hypersensitivity (Type III)

SUMMARY

- **Diseases caused by immune complexes can be divided into three groups.** Immune complexes are formed every time antibody meets antigen and are removed by the mononuclear phagocyte system following complement activation. Persistence of antigen from continued infection or in autoimmune disease can lead to immune complex disease.
- **Immune complexes can trigger a variety of inflammatory processes.** Fc–FcR interactions are the key mediators of inflammation. Most importantly, Fc regions within immune deposits within tissues engage Fc receptors on activated neutrophils, lymphocytes, and platelets to induce inflammation. During chronic inflammation B cells and macrophages are the predominant infiltrating cell type, and activation of endogenous cells within the organ participates in fibrosis and disease progression.
- **Experimental models demonstrate the main immune complex diseases.** Serum sickness can be induced with large injections of foreign antigen. Autoimmunity causes immune complex disease in the NZB/NZW mouse. Injection of antigen into the skin of presensitized animals produces the Arthus reaction.
- **Immune complexes are normally removed by the mononuclear phagocyte system.** Complement helps to disrupt antigen–antibody bonds and keeps immune complexes soluble. Primate erythrocytes bear a receptor for C3b and are important for transporting complement-containing immune complexes to the spleen for removal. Complement deficiencies lead to the formation of large, relatively insoluble complexes, which deposit in tissues.
- **The size of immune complexes affects their deposition.** Deposition of circulating, soluble immune complexes is limited by physical factors, such as the size and charge of the complexes. Small, positively charged complexes have the greatest propensity for deposition within vessels. Large immune complexes are rapidly removed in the liver and spleen.
- **Immune complex deposition in the tissues results in tissue damage.** Immune complexes can form both in the circulation, leading to systemic disease, and at local sites such as the lung. Charged cationic antigens have tissue-binding properties, particularly for the glomerulus, and help to localize complexes to the kidney. Factors that tend to increase blood vessel permeability enhance the deposition of immune complexes in tissues.
- **Deposited immune complexes can be visualized using immunofluorescence**.

DISEASES CAUSED BY IMMUNE COMPLEXES CAN BE DIVIDED INTO THREE GROUPS

Immune complexes are formed every time antibody meets antigen, and generally they are removed effectively by the mononuclear phagocyte system.

Occasionally, however, immune complexes persist and eventually deposit in a range of tissues and organs. The complement and effector cell-mediated damage that follows is known as a type III hypersensitivity reaction or immune complex disease.

The sites of immune complex deposition are partly determined by the localization of the antigen in the tissues and partly by how circulating complexes become deposited.

Diseases resulting from immune complex formation can be divided broadly into three groups:

- those due to persistent infection;
- those due to autoimmune disease; and
- those caused by inhalation of antigenic material (Fig. 25.1).

Persistent infection with a weak antibody response can lead to immune complex disease

The combined effects of a low-grade persistent infection and a weak antibody response lead to chronic immune complex formation, and eventual deposition of complexes in the tissues (Fig. 25.2). Diseases with this etiology include:

- leprosy;
- malaria;
- dengue hemorrhagic fever;
- viral hepatitis; and
- staphylococcal infective endocarditis.

Three categories of immune complex disease

cause	antigen	site of complex deposition
persistent infection	microbial antigen	infected organ(s), kidney
autoimmunity	self antigen	kidney, joint, arteries, skin
inhaled antigen	mold, plant, or animal antigen	lung

Fig. 25.1 This table indicates the source of the antigen and the organs most frequently affected.

Immunofluorescence study of immune complexes in infectious disease

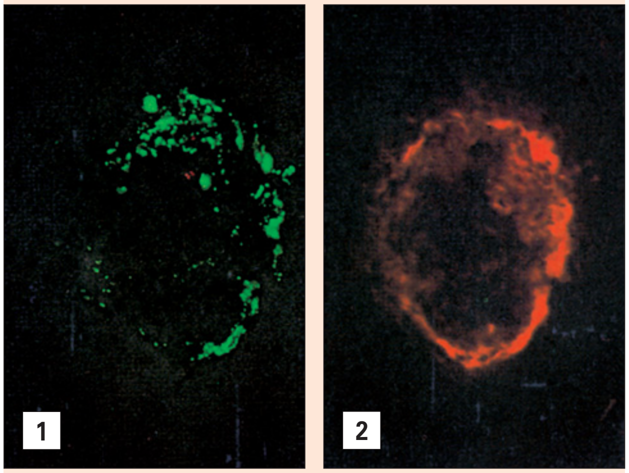

Fig. 25.2 These serial sections of the renal artery of a patient with chronic hepatitis B infection are stained with fluoresceinated anti-hepatitis B antigen (**1**) and rhodaminated anti-IgM (**2**). The presence of both antigen and antibody in the intima and media of the arterial wall indicates the deposition of complexes at this site. IgG and C3 deposits are also detectable with the same distribution. (Courtesy of Dr A Nowoslawski)

Immune complex disease is a frequent complication of autoimmune disease

Immune complex disease is common in autoimmune disease, where the continued production of autoantibody to a self antigen leads to prolonged immune complex formation. As the number of complexes in the blood increases, the systems responsible for the removal of complexes (mononuclear phagocyte, erythrocyte, and complement) become overloaded, and complexes are deposited in the tissues (Fig. 25.3). Diseases with this etiology include:

- rheumatoid arthritis;
- systemic lupus erythematosus (SLE); and
- polymyositis.

Immunofluorescence study of immune complexes in autoimmune disease

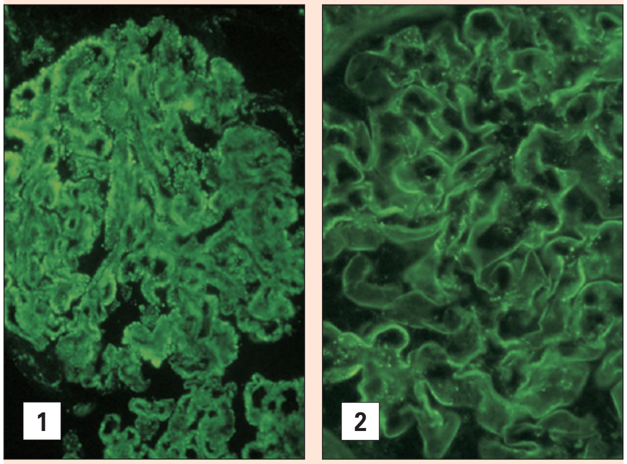

Fig. 25.3 These renal sections compare the effect of systemic lupus erythematosus (type III hypersensitivity) (**1**) and Goodpasture's syndrome (type II hypersensitivity) (**2**). In each case the antibody is detected with fluorescent anti-IgG. Complexes, formed in the blood and deposited in the kidney, form characteristic 'lumpy bumpy' deposits (**1**). The anti-basement membrane antibody in Goodpasture's syndrome forms an even layer on the glomerular basement membrane (**2**). (Courtesy of Dr S Thiru)

Immune complex following inhalation of antigen

Immune complexes may be formed at body surfaces following exposure to extrinsic antigens.

Such reactions are seen in the lungs following repeated inhalation of antigenic materials from molds, plants, or animals. This is exemplified in:

- farmer's lung, where there are circulating antibodies to actinomycete fungi (found in moldy hay); and
- pigeon fancier's lung, where there are circulating antibodies to pigeon antigens.

Both diseases are forms of **extrinsic allergic alveolitis**, and occur only after repeated exposure to the antigen. Note that the antibodies induced by these antigens are primarily IgG, rather than the IgE seen in type I hypersensitivity reactions. When antigen again enters the body by inhalation, local immune complexes are formed in the alveoli leading to inflammation and fibrosis (Fig. 25.4).

Precipitating antibodies to actinomycete antigens are found in the sera of 90% of patients with farmer's lung. However, they are also found in some people with no disease, and are absent from some patients, so it seems that other factors are also involved in the disease process, including type IV hypersensitivity reactions.

IMMUNE COMPLEXES CAN TRIGGER A VARIETY OF INFLAMMATORY PROCESSES

Immune complexes are capable of triggering a wide variety of inflammatory processes:

Extrinsic allergic alveolitis

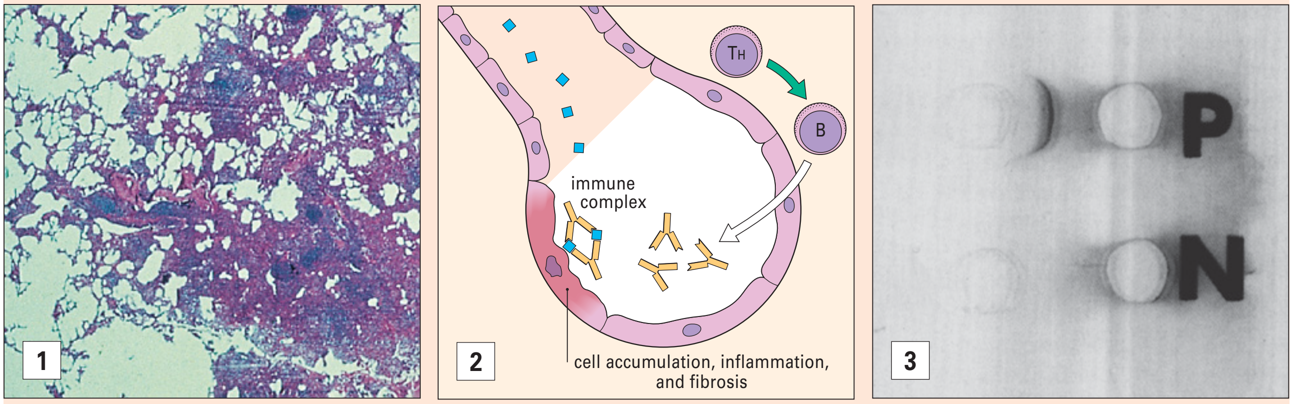

Fig. 25.4 The histological appearance of the lung in extrinsic allergic alveolitis (**1**) shows consolidated areas due to cell accumulation. When fungal antigen is inhaled into the lung of a sensitized individual, immune complexes are formed in the alveoli (**2**). Complement fixation leads to cell accumulation, inflammation, and fibrosis. Precipitin antibody (P) present in the serum of a patient with pigeon fancier's lung (**3**) is directed against the fungal antigen *Micropolyspora faeni*. Normal serum (N) lacks antibodies to this fungus.

- they interact directly with basophils and platelets (via Fc receptors) to induce the release of vasoactive amines (Fig. 25.5);
- macrophages are stimulated to release cytokines, particularly tumor necrosis factor-α (TNFα) and interleukin-1 (IL-1), which have important roles in inflammation;
- they interact with the complement system to generate **C3a and C5a (anaphylatoxins)** – these complement fragments stimulate the release of vasoactive amines (including histamine and 5-hydroxytryptamine) and chemotactic factors from mast cells and basophils; and C5a is also chemotactic for basophils, eosinophils, and neutrophils.

Recent work with knockout mice indicates that complement has a less proinflammatory role than previously thought, whereas cells bearing Fc receptors for IgG and IgE appear to be critical for developing inflammation, with complement having a protective effect.

The vasoactive amines released by platelets, basophils, and mast cells cause endothelial cell retraction and thus increase vascular permeability, allowing the deposition of immune complexes on the blood vessel wall (Fig. 25.6). The deposited complexes continue to generate C3a and C5a.

Platelets also aggregate on the exposed collagen of the vessel basement membrane to form microthrombi.

Q. Aggregation may be directly enhanced by the presence of immune complexes on the basement membrane. How can platelets recognize immune complexes?

A. They have an Fc receptor, FcγRIIa (see Fig. 3.18).

The aggregated platelets continue to produce vasoactive amines and to stimulate the production of C3a and C5a. (Platelets are also a rich source of growth factors – these may be involved in the cellular proliferation seen in immune complex diseases such as **glomerulonephritis** and **rheumatoid arthritis**.)

Polymorphs are chemotactically attracted to the site by C5a. They attempt to engulf the deposited immune complexes, but are unable to do so because the complexes are bound to the vessel wall. They therefore exocytose their lysosomal enzymes onto the site of deposition (see Fig. 25.6). If simply released into the blood or tissue fluids these lysosomal enzymes are unlikely to cause much inflammation, because they are rapidly neutralized by serum enzyme inhibitors. But if the phagocyte applies itself closely to the tissue-trapped complexes through Fc binding, then serum inhibitors are excluded and the enzymes may damage the underlying tissue.

Type II and type III hypersensitivity cause damage by similar mechanisms

It is important to understand that the inflammatory pathways involved in type II and type III reactions are identical. The distinction is based on where the antigens are derived from and how the immune complexes form:

- in type II hypersensitivity reactions, the autoantibodies bind to autoantigens where they normally reside;
- by contrast, in type III hypersensitivity reactions, the antigens may be either self (i.e. auto) or foreign and the immune deposits may form by either deposition of soluble immune complexes or in situ by initial deposition of unbound circulating antigen followed by binding of circulating antibody.

Nevertheless, the effector mechanisms leading to inflammation and tissue injury are similar in type II and type III reactions and dependent on the capacity of deposited antibodies to engage FcRs and activate complement.

Immune complexes as a trigger for increasing vascular permeability

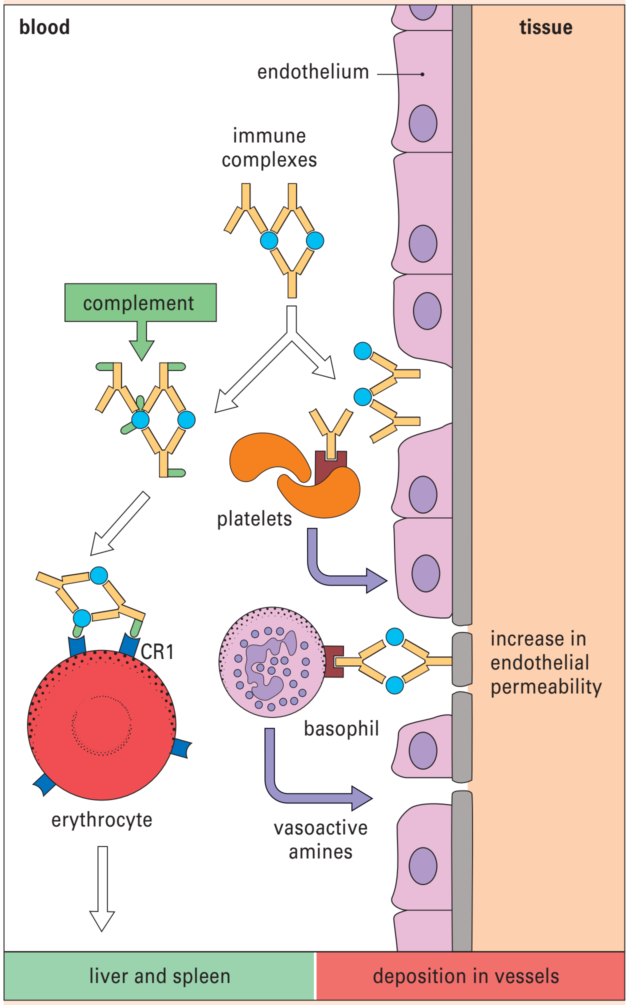
blood
tissue
endothelium
immune complexes
complement
platelets
CR1
erythrocyte
basophil
vasoactive amines
increase in endothelial permeability
liver and spleen
deposition in vessels

Fig. 25.5 Immune complexes normally bind complement and are removed to the liver and spleen after binding to CR1 on erythrocytes. In inflammation, immune complexes act on basophils and platelets (in humans) to produce vasoactive amine release. The amines released (e.g. histamine, 5-hydroxytryptamine) cause endothelial cell retraction and thus increase vascular permeability.

Complement is an important mediator of type III hypersensitivity

In many diseases, complement activation is triggered inappropriately and drives a vicious cycle, causing:

- further tissue damage;
- increased inflammation; and
- perpetuation of the disease.

This scenario is particularly evident in autoimmune diseases where immune complexes deposit in tissues and there activate complement, causing damage and destruction of host cells. Examples include:

- the kidney in various autoimmune glomerular diseases; and
- the joint in rheumatoid arthritis.

Deposition of immune complexes in blood vessel walls

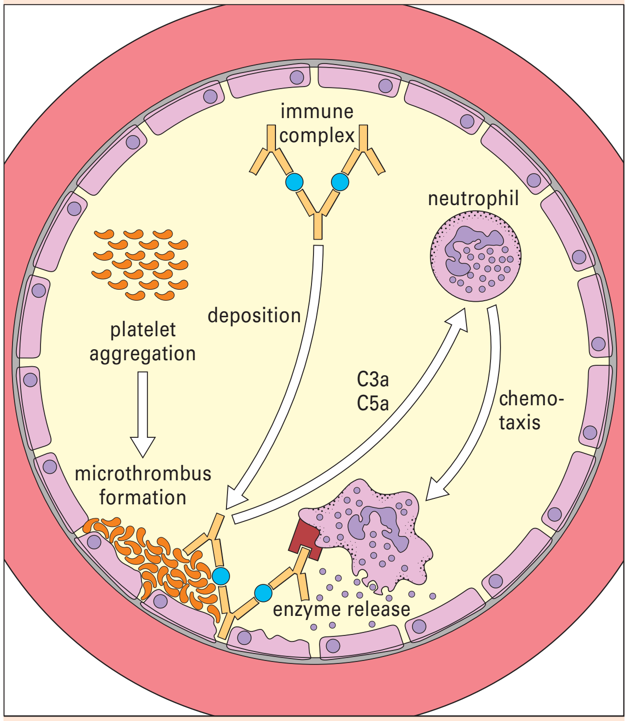
immune complex
neutrophil
deposition
platelet aggregation
C3a
C5a
chemo-taxis
microthrombus formation
enzyme release

Fig. 25.6 Increased vascular permeability allows immune complexes to be deposited in the blood vessel wall. This induces platelet aggregation and complement activation. The aggregated platelets form microthrombi on the exposed collagen of the basement membrane of the endothelium. Neutrophils are attracted to the site by complement products, but cannot ingest the complexes. They therefore exocytose their lysosomal enzymes, causing further damage to the vessel wall.

Staining of these tissues for complement deposits reveals the full extent of involvement. The tissues are often packed with C3 fragments and other complement proteins. Complement activation is also evident in the blood in these diseases:

- complement activity and the plasma concentrations of the major components C3 and C4 are reduced due to consumption in the tissues; and
- levels of complement activation fragments are increased.

In SLE, autoantibodies are generated against cell contents including DNA, cytoplasmic proteins, and mitochondria. Immune complexes form whenever cell contents are released and these deposit in capillary beds in organs such as skin, kidney, joint, and brain where they activate complement causing further tissue damage. Here complement is playing dual roles:

- the important immune complex solubilizing roles will prevent immune complex deposition until the capacity of the system is exceeded;
- beyond this threshold complexes deposit and activate complement in the tissues, causing pathology.

Complement changes from hero to villain. Patients with SLE often have markedly decreased plasma levels of complement activity and the components C3 and C4 due to the massive and widespread activation of the system. Levels of these parameters can be used to monitor disease activity.

Autoantibodies to complement components can modulate complement activity

Autoantibodies may also develop that directly target the complement components and complexes.

Autoantibodies against C1q are commonly found in SLE, correlating particularly with renal involvement.

Antibodies against the alternative pathway C3 convertase bind and stabilize the complex, markedly increasing its functional half-life and thus consuming C3.

These autoantibodies were first identified in patients with **membranoproliferative glomerulonephritis (MPGN)** and were therefore termed **C3 nephritic factors (C3NeF)**, but may also be found in SLE.

C3NeF are also associated with a bizarre 'wasting' disorder characterized by loss of adipose tissue, termed **partial lipodystrophy**, which may occur with or without associated MPGN. The syndrome was, until recently, unexplained, but recent evidence suggests that activation of the alternative pathway is a feature of normal lipid handling in adipose tissue. Stabilization of the convertase by C3NeF will then cause local increased complement activation and fat cell destruction.

Rarely, autoantibodies against the classical pathway C3 convertase, termed **C4 nephritic factors (C4NeF)**, are found in association with MPGN or SLE. These too stabilize the convertase and cause increased complement activation to drive disease.

EXPERIMENTAL MODELS DEMONSTRATE THE MAIN IMMUNE COMPLEX DISEASES

Experimental models are available for each of the three main types of immune complex disease described above:

- **serum sickness**, induced by injections of foreign antigen, mimics the effect of a persistent infection;
- the NZB/NZW mouse demonstrates **autoimmunity**;
- the **Arthus reaction** is an example of local damage by extrinsic antigen.

Care must be taken when interpreting animal experiments because the erythrocytes of rodents and rabbits lack the receptor for C3b (known as CR1), which readily binds immune complexes that have fixed complement. This receptor is present on primate erythrocytes.

Serum sickness can be induced with large injections of foreign antigen

In serum sickness, circulating immune complexes deposit in the blood vessel walls and tissues, leading to increased vascular permeability and thus to inflammatory diseases such as glomerulonephritis and arthritis.

In the pre-antibiotic era, serum sickness was a complication of serum therapy, in which massive doses of antibody were given for diseases such as diphtheria. Horse anti-diphtheria serum was usually used, and some individuals made antibodies against the horse proteins.

Serum sickness is now commonly studied in rabbits by giving them an intravenous injection of a foreign soluble protein such as bovine serum albumin (BSA). After about 1 week antibodies are formed, which enter the circulation and complex with antigen. Because the reaction occurs in antigen excess, the immune complexes are small (Fig. 25.7). These small complexes are removed only slowly by the mononuclear phagocyte system and therefore persist in the circulation.

Q. Suggest two reasons why small immune complexes would be cleared more slowly than large complexes.

A. Small complexes activate complement less effectively than large complexes because C1 must bind to several Fc regions before it is activated. The avidity of complexes for binding to phagocytes will also be increased when several Fc regions are available to bind to the Fc receptors on a phagocyte. If this were not so, monomeric antibody would block binding of complexes to Fc receptors.

The formation of complexes is followed by an abrupt fall in total hemolytic complement.

The clinical signs of serum sickness that develop are due to granular deposits of antigen–antibody and C3 forming along the glomerular basement membrane (GBM) and in small vessels elsewhere. As more antibody is formed and the reaction moves into antibody excess, the size of the complexes increases and they are cleared more efficiently, so the animals recover. Chronic disease is induced by daily administration of antigen.

Time course of experimental serum sickness

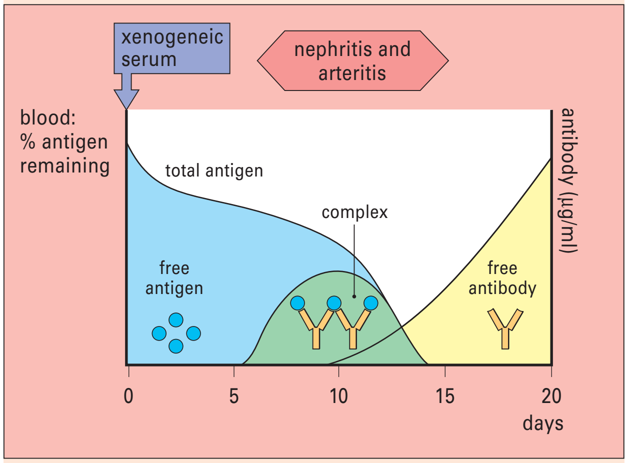

Fig. 25.7 Following an injection of xenogeneic serum there is a lag period of approximately 5 days, in which only free antigen is detectable in serum. After this time, antibodies are produced to the foreign proteins and immune complexes are formed in serum; it is during this period that the symptoms of nephritis and arteritis appear. To begin with, small soluble complexes are found in antigen excess; with increasing antibody titers, larger complexes are formed, which are deposited and subsequently cleared. At this stage the symptoms disappear.

Q. Why would immune complexes deposit in the glomerulus when the complexes are formed outside the kidney?

A. The glomerulus has fenestrated endothelium and is a site of filtration.

Autoimmunity causes immune complex disease in the NZB/NZW mouse

The F$_1$ hybrid NZB/NZW mouse produces a range of autoantibodies (including anti-erythrocyte, anti-nuclear, anti-DNA, and anti-*S*m) and suffers from an immune complex disease similar in many ways to SLE in humans.

A NZB/NZW mouse is born clinically normal, but within 2–3 months shows sign of hemolytic anemia. Tests for anti-erythrocyte antibody (the Coombs' test), anti-nuclear antibodies, lupus cells, and circulating immune complexes are all positive, and there are deposits in the glomeruli and choroid plexus of the brain. The disease is much more marked in the females, who die within a few months of developing symptoms (Fig. 25.8).

Autoimmune disease in NZB/NZW mice

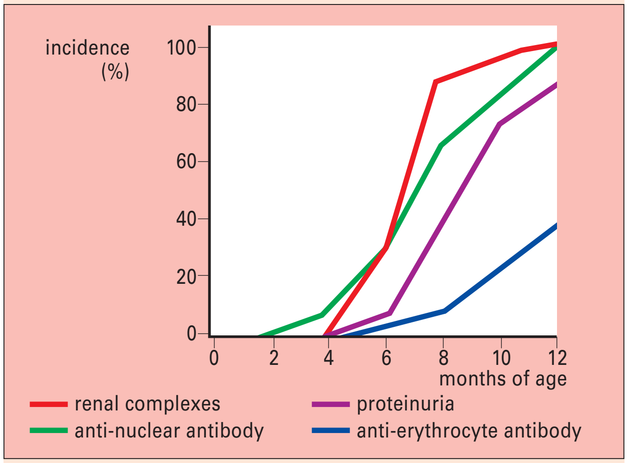

Fig. 25.8 The graph shows the onset of autoimmune disease in female NZB/NZW mice with advancing age. Incidence refers to the percentage of mice with the features identified. Immune complexes were detected by immunofluorescent staining of a kidney section. Anti-nuclear antibodies were detected in serum by indirect immunofluorescence. Proteinuria reflects kidney damage. Autoantibodies to erythrocytes develop later in the disease and so are less likely to relate to kidney pathology. Onset of autoimmune disease is delayed in male mice by approximately 3 months.

The role of Fc receptors in the 'treatment' of immune complex disease

The restoration of 'tolerance' in murine lupus disease has been achieved by the transduction of the inhibitory Fc receptor FcγRIIb. Animals so treated showed:

- a prolonged life span;
- lower levels of kidney disease as measured by reduced proteinuria; and
- lower levels of autoantibody (see Chapter 20).

Injection of antigen into the skin of presensitized animals produces the Arthus reaction

The Arthus reaction takes place at a local site in and around the walls of small blood vessels. It is most frequently demonstrated in the skin.

An animal is immunized repeatedly until it has appreciable levels of serum antibody (mainly IgG). Following subcutaneous or intradermal injection of the antigen a reaction develops at the injection site, sometimes with marked edema and hemorrhage, depending on the amount of antigen injected. The reaction reaches a peak after 4–10 hours, then wanes and is usually minimal by 48 hours (Fig. 25.9).

Immunofluorescence studies have shown that initial deposition of antigen, antibody, and complement in the vessel wall is followed by neutrophil infiltration and

The appearance of the three main skin test reactions

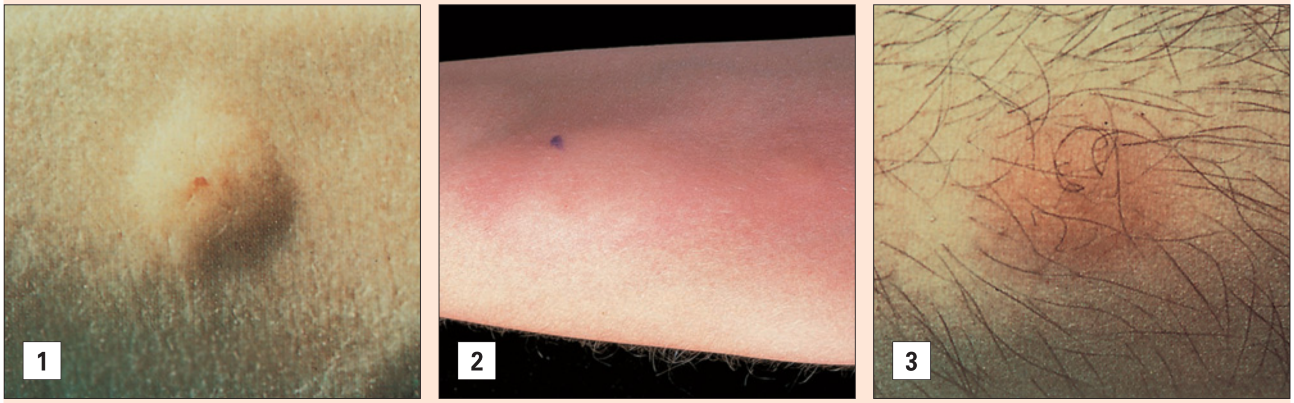

Fig. 25.9 A type I hypersensitivity reaction (**1**) produces a raised wheal, 5–7 mm in diameter and with a well defined edge after about 15 minutes. A type III hypersensitivity Arthus reaction (**2**) produces a reaction after 5–12 hours that is larger (≥ 50 mm) and has a less well defined edge. A type IV (delayed) hypersensitivity reaction shows as a red indurated lesion, about 5 mm in diameter at 24–48 hours (**3**).

intravascular clumping of platelets (Fig. 25.10). This platelet reaction can lead to vascular occlusion and necrosis in severe cases. After 24–48 hours the neutrophils are replaced by mononuclear cells, and eventually some plasma cells appear.

Complement activation via either the classical or alternative pathways was thought to be essential for the Arthus reaction to develop. But C3, C4, or C5 deficient mice were able to mount a normal Arthus reaction. However, when mice were made deficient in FcγRI or FcγRIII they were unable to produce the reaction. Furthermore, when recombinant soluble FcγRII receptors were given they inhibited the development of the Arthus reaction.

TNFα enhances cell-mediated immune responses in various ways (see Chapter 9). Treatment with antibodies to TNF can reduce severity in the Arthus reaction and, interestingly, anti-TNF is useful in treating rheumatoid arthritis.

The ratio of antibody to antigen is directly related to the severity of the ensuing reaction. Complexes formed in either antigen or antibody excess are much less toxic than those formed at equivalence.

IMMUNE COMPLEXES ARE NORMALLY REMOVED BY THE MONONUCLEAR PHAGOCYTE SYSTEM

Immune complexes are opsonized with C3b following complement activation, and removed by the mononuclear phagocyte system, particularly in the liver and spleen. Removal is mediated by the complement C3b receptor, CR1.

In primates, the bulk of CR1 in blood is found on erythrocytes. (Non-primates do not have erythrocyte CR1, and must therefore rely on platelet CR1.) There are about 700 receptors per erythrocyte, and their effectiveness is enhanced by the grouping of receptors in patches, allowing high-avidity binding to the large complexes.

CR1 readily binds immune complexes that have fixed complement, as has been shown by experiments with animals lacking complement (Fig. 25.11).

In normal primates the erythrocytes provide a buffer mechanism, binding complexes that have fixed complement and effectively removing them from the plasma. In small blood vessels 'streamline flow' allows the erythrocytes to travel in the center of the vessel surrounded by the flowing plasma. Thus it is only the plasma that makes contact with the vessel wall. Only in the sinusoids of the liver and spleen, or at sites of turbulence, do the erythrocytes make contact with the lining of the vessels.

The complexes are transported to the liver and spleen, where they are removed by fixed tissue macrophages (Fig. 25.12). Most of the CR1 is also removed in the process, so in situations of continuous immune complex formation the number of active receptors falls rapidly, impairing the efficiency of immune complex handling.

In patients with SLE, for example, the number of receptors may well be halved. With fewer complement receptors the complexes are cleared rapidly to the liver, but these complexes, which arrive directly rather than on red cells, are later released into the circulation again and may then deposit in the tissues elsewhere and lead to inflammation.

The Arthus reaction

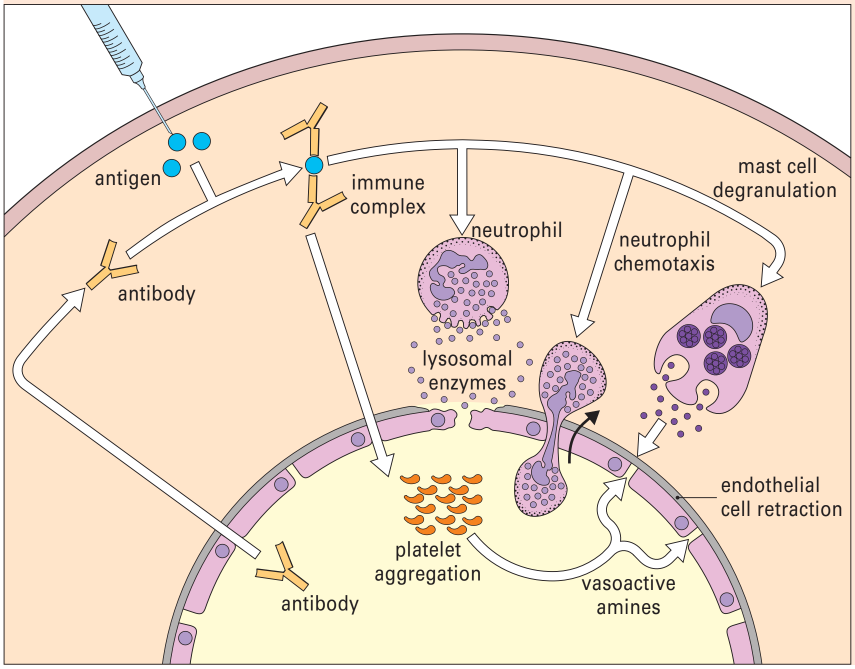

Fig. 25.10 Antigen injected intradermally combines with specific antibody from the blood to form immune complexes. The complexes act on platelets and mast cells, which release vasoactive amines. Immune complexes also induce macrophages to release TNF and IL-1 (not shown). Mast cell products, including histamine and leukotrienes, induce increased blood flow and capillary permeability. The inflammatory reaction is potentiated by lysosomal enzymes released from the polymorphs. The Arthus reaction can be seen in patients with precipitating antibodies, such as those with extrinsic allergic alveolitis associated with farmer's lung disease.

Effects of complement depletion on handling of immune complexes

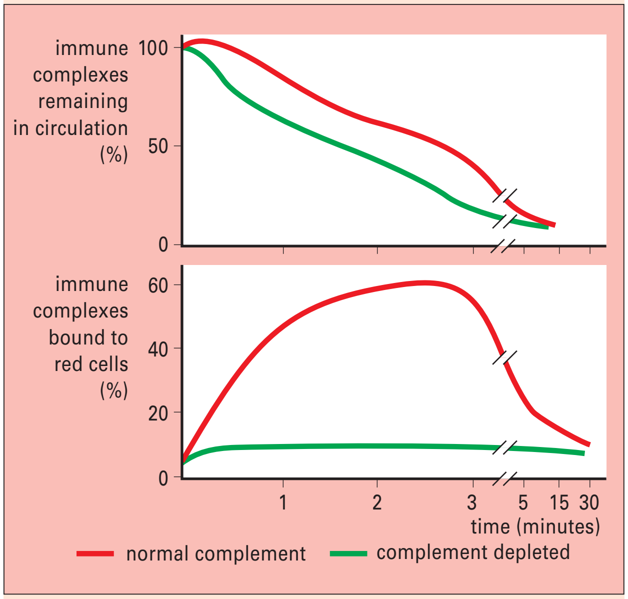

Fig. 25.11 A bolus of immune complexes was infused into the circulation of a primate. In animals with a normal complement system the complexes were bound quickly by the CR1 on erythrocytes. In animals whose complement had been depleted by treatment with cobra venom factor, the erythrocytes hardly bound immune complexes at all. Paradoxically, this results in slightly faster removal of complexes in the depleted animals, with the complexes being deposited in the tissues rather than being removed by the spleen. (Based on data from Waxman FJ, et al. J Clin Invest 1984;74:1329–1340)

Complexes can also be released from erythrocytes in the circulation by the enzymatic action of factor I.

Q. What action does factor I have in the complement system?

A. It cleaves C3b and C4b into fragments (see Figs 4.7–4.9).

This action leaves a small fragment (C3dg) attached to the CR1 on the cell membrane. These soluble complexes are then removed by phagocytic cells, particularly those in the liver, bearing receptors for IgG Fc (Fig. 25.13).

Complement solubilization of immune complexes

It has been known since Heidelberger's work on the precipitin curve in the 1930s that complement delays precipitation of immune complexes, though this information was forgotten for a long time.

The ability to keep immune complexes soluble is a function of the classical complement pathway. The complement components reduce the number of antigen epitopes that the antibodies can bind (i.e. they reduce the valency of the antigen) by intercalating into the lattice of the complex, resulting in smaller, soluble complexes. In primates these complement-bearing complexes are readily bound by the C3b receptor (CR1) on erythrocytes.

Clearance of immune complexes in the liver

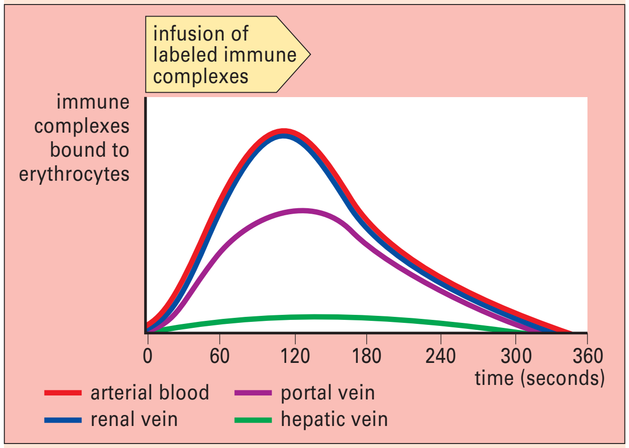

Fig. 25.12 ^{125}I-BSA/anti-BSA complexes were infused into a primate over a period of 120 seconds. Blood was sampled from renal, portal, and hepatic veins, and the level of immune complexes bound to the erythrocytes was measured by radioactive counting. The levels of complexes in the renal and portal veins were similar to that in arterial blood. However, complexes were virtually absent from hepatic venous blood throughout, indicating that complexes bound to erythrocytes are removed during a single transit through the liver. (Based on data from Cornacoff JB, et al. J Clin Invest 1983;71:236–247)

Complement can rapidly resolubilize precipitated complexes through the alternative pathway (Fig. 25.14). The solubilization appears to occur by the insertion of complement C3b and C3d fragments into the complexes.

It may be that complexes are continually being deposited in normal individuals, but are removed by solubilization. If this is the case, then the process will be inadequate in hypocomplementemic patients and lead to prolonged complex deposition.

Solubilization defects have indeed been observed in sera from patients with systemic immune complex disease, but whether the defect is primary or secondary is not known.

Q. What evidence implies that defective solubilization of complexes is a primary cause of immune complex disease?

A. Genetic deficiency of classical pathway components (i.e. primary defects) are associated with SLE and some other immune complex diseases (see Figs 4.16 and 16.11).

Complement deficiency impairs clearance of complexes

In patients with low levels of classical pathway components there is poor binding of immune complexes to erythrocytes. The complement deficiency may result from:

- depletion, caused by immune complex disease; or
- a hereditary disorder, as is the case in C2 deficiency.

Immune complex clearance

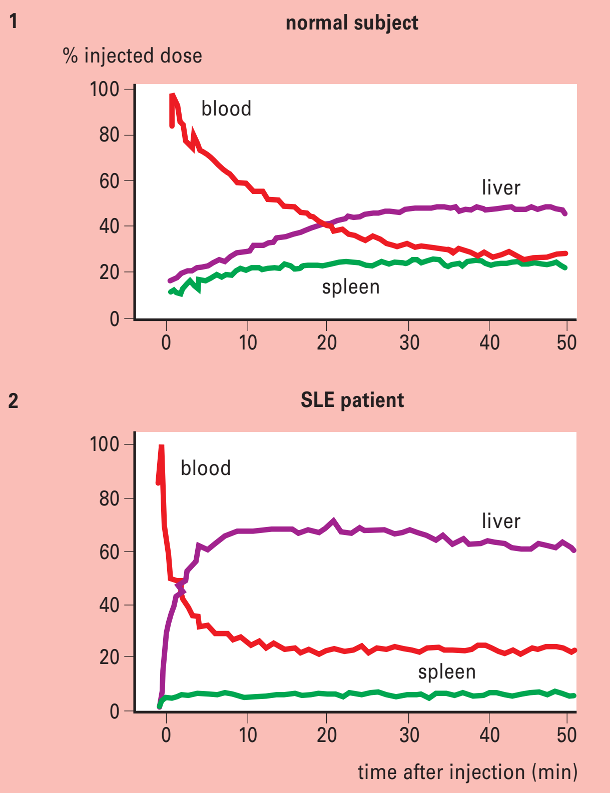

Fig. 25.13 (**1**) Immune complex clearance in a healthy normal subject. (**2**) Immune complex clearance in a patient with SLE. Radiolabeled soluble complexes were injected intravenously and immune complex localization monitored by dynamic imaging. In the normal subject complexes remained longer in the blood through binding to CR1 on red cells, followed by clearance to the liver and the spleen, where immune complexes take part in immunoregulation. In the hypocomplementemic patient with SLE there was little binding to red cells, but rapid clearance to organs such as the liver, with little localizing to the spleen, leading to impaired immunoregulation, which may be a factor in the persistence of autoimmunity.

Solubilization of immune complexes by complement

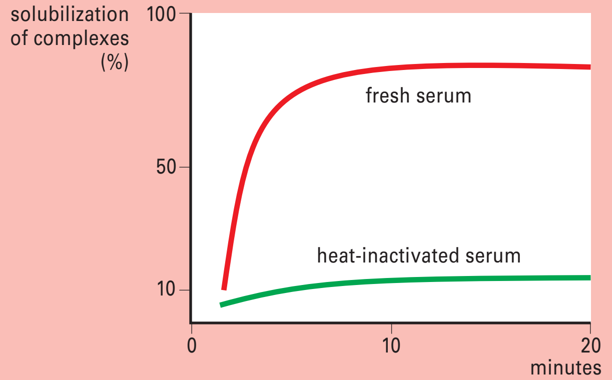

Fig. 25.14 Complement can solubilize precipitable complexes in vitro. Addition of fresh serum containing active complement to insoluble complexes induces solubilization over about 15 minutes at 37°C. Some of the complexes resist resolubilization. Heated serum (56°C for 30 minutes) lacks active complement and cannot resolubilize the complexes. Intercalation of complement components C3b and C3d into the complex causes their solubilization by disrupting antigen–antibody bonds. Complexes that have been artificially connected by covalent bonds cannot be solubilized by complement.

This might be expected to result in persistent immune complexes in the circulation, but in fact the reverse occurs, with the complexes disappearing rapidly from the circulation. These non-erythrocyte-bound complexes are taken up rapidly by the liver (but not the spleen) and are then released to be deposited in tissues such as skin, kidney, and muscle, where they can set up inflammatory reactions (Fig. 25.15).

Infusion of fresh plasma, containing complement, restores the clearance patterns to normal, illustrating the importance of complement in the clearance of immune complexes.

Failure to localize in the spleen not only results in immune complex disease, but may also have important implications for the development of appropriate immune responses. This is because the spleen plays a vital role in antigen processing and the induction of immune responses (see Chapter 2).

The size of immune complexes affects their deposition

In general, larger immune complexes are rapidly removed by the liver within a few minutes, whereas smaller complexes circulate for longer periods (Fig. 25.16). This is because larger complexes are:

- more effective at binding to Fc receptors and at fixing complement, so binding better to erythrocytes;
- released more slowly from the erythrocytes by the action of factor I.

Anything that affects the size of complexes is therefore likely to influence clearance.

It has been suggested that a genetic defect that favors the production of low-affinity antibody could lead to the formation of smaller complexes, and so to immune complex disease.

Antibodies to self antigens may have low affinity and recognize only a few epitopes. This results in small complexes and long clearance times because the formation of large, cross-linked lattices is restricted.

Affinity maturation is dependent on efficient somatic mutation and selection of B cells within germinal centers following binding of antigen. This process is far more effective when B cells are stimulated by antigen or immune complexes coated with complement. Patients with complement deficiencies are particularly prone to develop immune complex disease and recent evidence indicates that another way that this is brought about is through poor targeting of antigen complexes to germinal centers, so preventing affinity maturation.

Immune complex transport and removal

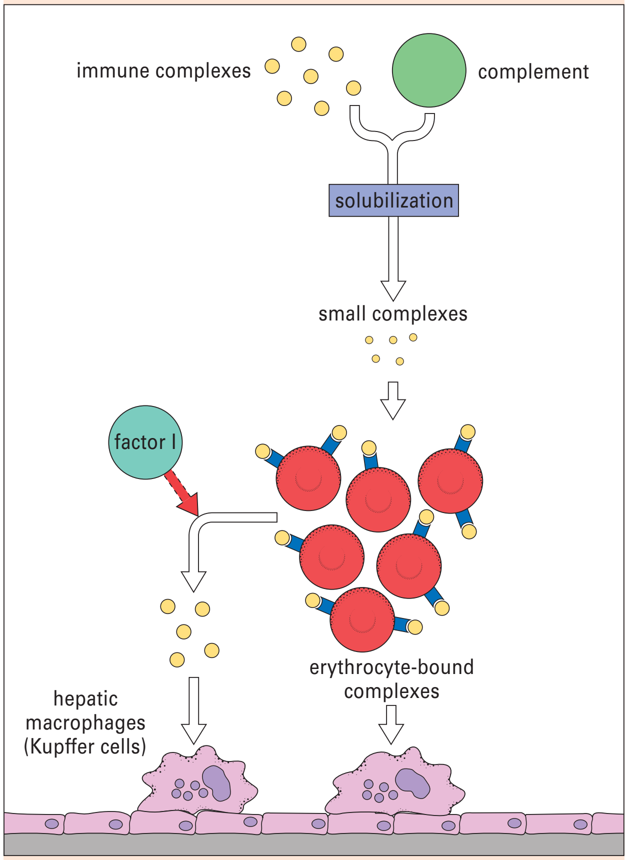

Fig. 25.15 In primates, complexes solubilized by complement are bound by CR1 on erythrocytes and transported to the liver where they are removed by hepatic macrophages. Complexes released from erythrocytes by factor I are taken up by cells (including macrophages) bearing receptors for Fc and complement.

Immunoglobulin classes affect the rate of immune complex clearance

Striking differences have been observed in the clearance of complexes with different immunoglobulin classes:

- IgG complexes are bound by erythrocytes and are gradually removed from the circulation;
- IgA complexes bind poorly to erythrocytes, but disappear rapidly from the circulation, with increased deposition in the kidney, lung, and brain.

Q. Provide an explanation for the different patterns of localization of immune complexes containing IgG and those containing IgA.

A. IgG-containing immune complexes activate the complement classical pathway and can bind to CR1 on erythrocytes. IgA does not activate the classical pathway, but can bind to Fcα receptors on mononuclear phagocytes (see Figs 3.18 and 4.3).

Complex clearance by mononuclear phagocytes

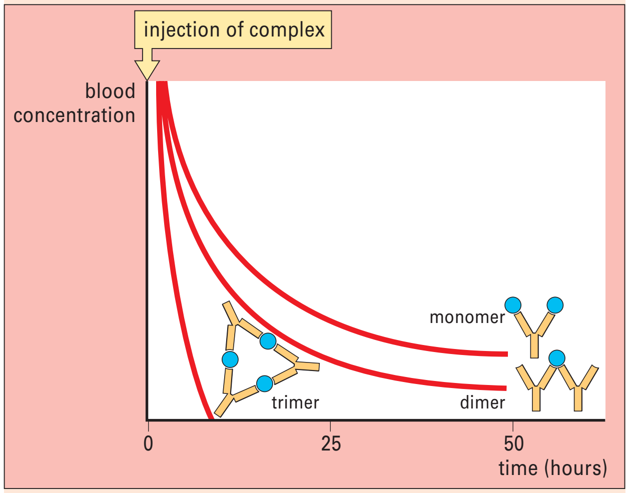

Fig. 25.16 Large immune complexes are cleared most quickly because they present an IgG–Fc lattice to mononuclear phagocyte cells with Fc receptors, permitting higher avidity binding to these cells. They also fix complement better than small complexes.

Phagocyte defects allow complexes to persist

Opsonized immune complexes are normally removed by the mononuclear phagocyte system, mainly in the liver and spleen. However, when large amounts of complex are present, the mononuclear phagocyte system may become overloaded, leading to a rise in the level of circulating complex and increased deposition in the glomerulus and elsewhere.

Defective mononuclear phagocytes have been observed in human immune complex disease, but this may be the result of overload rather than a primary defect.

Carbohydrate on antibodies affects complex clearance

Carbohydrate groups on immunoglobulin molecules have been shown to be important for the efficient removal of immune complexes by phagocytic cells.

Abnormalities of these carbohydrates occur in immune complex diseases such as rheumatoid arthritis, thus aggravating the disease process. IgG Fc oligosaccharides lack the normally terminating galactose residue, enhancing rheumatoid factor binding. Recently, mannan-binding protein has been shown to bind agalactosyl IgG and subsequently activate complement.

IMMUNE COMPLEX DEPOSITION IN TISSUES RESULTS IN TISSUE DAMAGE

Immune complexes may persist in the circulation for prolonged periods of time. However, simple persistence is not usually harmful in itself; the problems start only when complexes are deposited in the tissues.

Two questions are relevant to tissue deposition:

- Why are complexes deposited?
- Why do complexes show affinity for particular tissues in different diseases?

The most important trigger for immune complex deposition is probably an increase in vascular permeability

Animal experiments have shown that inert substances such as colloidal carbon will be deposited in vessel walls following the administration of vasoactive substances, such as histamine or serotonin. Circulating immune complexes are deposited in a similar way following the infusion of agents that cause the liberation of mast cell vasoactive amines (including histamine). Pretreatment with antihistamines blocks this effect.

In studies of experimental immune complex disease in rabbits, long-term administration of vasoactive amine antagonists, such as chlorpheniramine and methysergide, has been shown to reduce immune complex deposition considerably (Fig. 25.17). More importantly (from the point of view of disease prevention), young NZB/NZW mice treated with methysergide show less renal pathology than controls (Fig. 25.18).

Effect of a vasoactive amine antagonist on immune complex disease

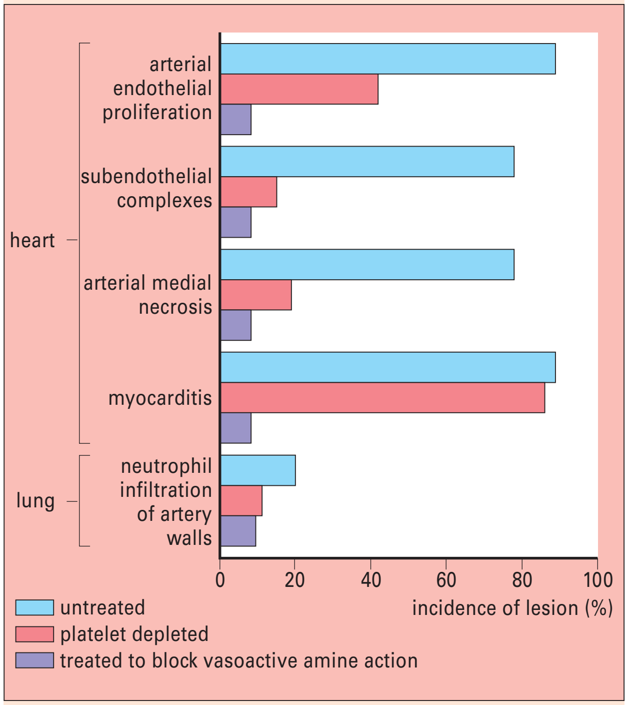

Fig. 25.17 Serum sickness was induced in rabbits with a single injection of bovine serum albumin. The animals were either untreated, platelet depleted, or treated with drugs to block vasoactive amine action. The incidence of serum sickness lesions in the heart and lung was scored. Drug treatment considerably reduced the signs of disease by lowering vascular permeability and thus minimizing immune complex deposition.

Effect of the vasoactive amine antagonist methysergide on kidney damage

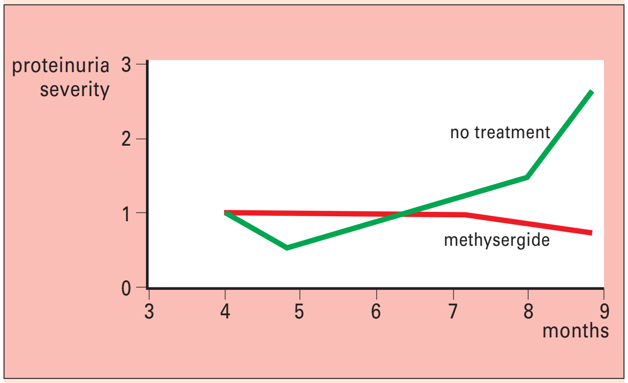

Fig. 25.18 Kidney damage, assessed by proteinuria, was measured in NZB/NZW mice over 5 months. Untreated animals developed severe proteinuria, whereas methysergide-treated animals did not. Methysergide blocks the formation of the vasoactive amine 5-hydroxytryptamine (5-HT), and thus blocks a variety of inflammatory events (e.g. deposition of complexes, neutrophil infiltration of capillary walls, and endothelial proliferation), all of which produce the glomerular pathology.

Increases in vascular permeability can be initiated by a range of mechanisms, which vary in importance, depending on the diseases and species concerned. This variability makes interpretation of some of the animal models difficult. In general, however, complement, mast cells, basophils, and platelets must all be considered as potential producers of vasoactive amines.

Immune complex deposition is most likely where there is high blood pressure and turbulence

Many macromolecules deposit in the glomerular capillaries, where the blood pressure is approximately four times that of most other capillaries (Fig. 25.19).

If the glomerular blood pressure of a rabbit is reduced by partially constricting the renal artery or by ligating the ureter, deposition is also reduced. If the glomerular blood pressure is increased by experimentally induced hypertension, immune complex deposition is enhanced as shown by the development of serum sickness. Elsewhere, the most severe lesions also occur at sites of turbulence:

- at turns or bifurcations of arteries; and
- in vascular filters such as the choroid plexus and the ciliary body of the eye.

Q. Why should the site of bifurcation in an artery be more susceptible to damage than other sites?

A. This site is subject to high pressure and more erratic shear forces, which may affect the integrity of the endothelium (see Fig. 25.19). In addition blood cells and platelets are not segregated from the vessel wall by laminar flow (see above).

Hemodynamic factors affecting complex deposition

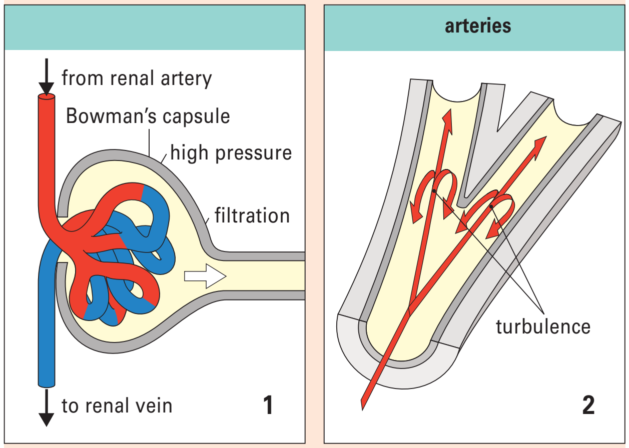

Fig. 25.19 Factors that affect complex deposition include filtration and high blood pressure, both of which occur in the formation of ultrafiltrate in the renal glomerulus (**1**). Turbulence at curves or bifurcations of arteries (**2**) also favors deposition of immune complexes.

Affinity of antigens for specific tissues can direct complexes to particular sites

Local high blood pressure explains the tendency for deposits to form in certain organs, but does not explain why complexes are deposited on specific organs in certain diseases. In SLE, the kidney is a particular target, whereas in rheumatoid arthritis, although circulating complexes are present, the kidney is usually spared and the joints are the principal target.

It is possible that the antigen in the complex provides the organ specificity, and a convincing model has been established to support this hypothesis. In the model, mice are given endotoxin causing cell damage and release of DNA, which then binds to healthy glomerular basement membrane. Anti-DNA is then produced by polyclonal activation of B cells, and is bound by the fixed DNA leading to local immune complex formation (Fig. 25.20). The production of rheumatoid factor (IgM anti-IgG) allows further immune complex formation to occur in situ.

It is possible that in other diseases antigens will be identified with affinity for particular organs.

The charge of the antigen and antibody may be important in some systems. For example, positively charged antigens and antibodies are more likely to be deposited in the negatively charged glomerular basement membrane.

The degree of glycosylation also affects the fate of complexes containing glycoprotein antigens because certain clearance mechanisms are activated by recognition of sugar molecules (e.g. mannan-binding protein).

In certain diseases the antibodies and antigens are both produced within the target organ. The extreme of this is reached in rheumatoid arthritis, where IgG anti-IgG rheumatoid factor is produced by plasma cells within the synovium; these antibodies then combine with each other (self-association), so setting up an inflammatory reaction.

Tissue binding of antigen with local immune complex formation

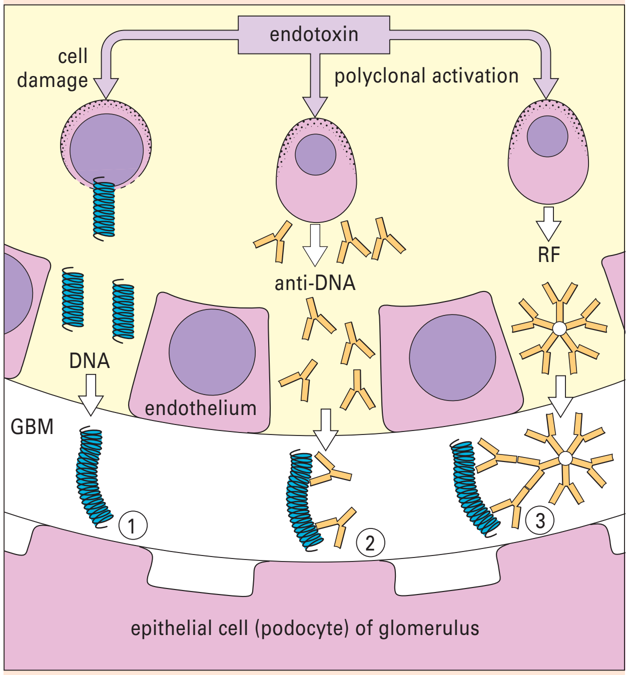

Fig. 25.20 Endotoxin injected into mice increases vascular permeability and induces cell damage and release of DNA. The DNA can then become deposited (**1**) on the collagen of the glomerular basement membrane (GBM) in the kidney. Endotoxin can also induce a polyclonal stimulation of B cells, some of which produce autoantibodies such as anti-DNA and anti-IgG – the latter are known as rheumatoid factors (RFs). Anti-DNA antibody can then bind to the deposited DNA forming a local immune complex (**2**). RFs have a low affinity for monomeric IgG, but bind with high avidity to the assembled DNA–anti-DNA complex (**3**). Thus further immune complex formation occurs in situ.

The site of immune complex deposition depends partly on the size of the complex

The fact that the site of immune complex deposition depends partly on the size of the complex is exemplified in the kidney:

- small immune complexes can pass through the glomerular basement membrane, and end up on the epithelial side of the membrane;
- large complexes are unable to cross the membrane and generally accumulate between the endothelium and the basement membrane or the mesangium (Fig. 25.21).

The size of immune complexes depends on the valency of the antigen, and on the titer and affinity of the antibody.

The class of immunoglobulin in an immune complex can also influence its deposition

There are marked age- and sex-related variations in the class and subclass of anti-DNA antibodies seen in SLE.

Immune complex deposition in the kidney

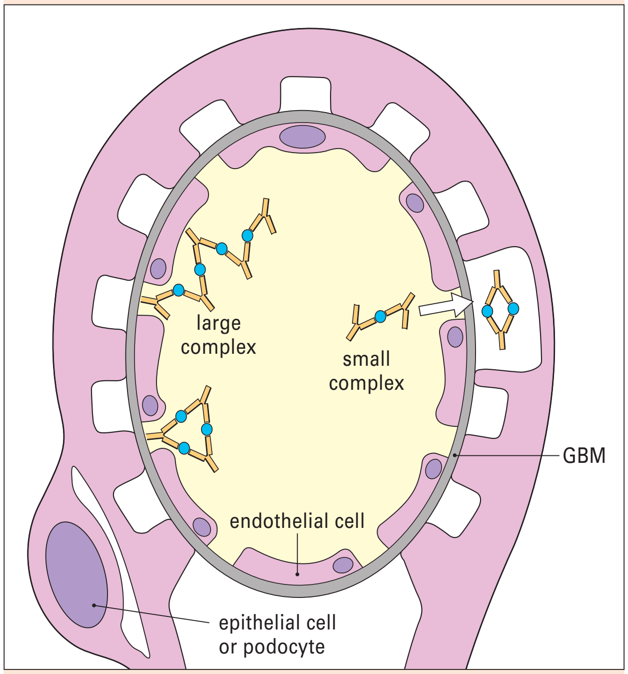

Fig. 25.21 The site of complex deposition in the kidney is dependent on the size of the complexes in the circulation. Large complexes become deposited on the glomerular basement membrane, whereas small complexes pass through the basement membrane and are seen on the epithelial side of the glomerulus.

Antibody classes in immune complex disease

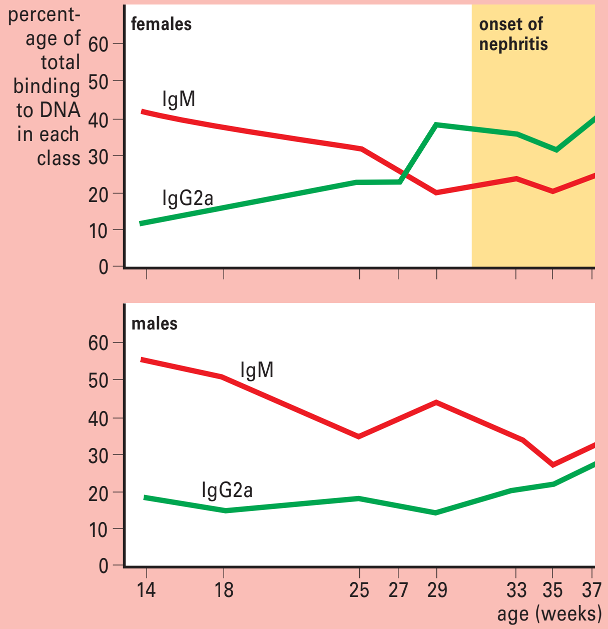

Fig. 25.22 Immune complex disease is automatic in the NZB/NZW mouse and follows a class switch during early development, from IgM to IgG2a. The graphs show the proportions of anti-DNA antibodies of the IgM and IgG2a isotypes in females and males. Both the class switch and fatal renal disease occur earlier in the female mice of this strain.

Similarly, as NZB/NZW mice grow older there is a class switch, from predominantly IgM to IgG2a. This occurs earlier in females than in males and coincides with the onset of renal disease, indicating the importance of antibody class in the tissue deposition of complexes (Fig. 25.22).

DEPOSITED IMMUNE COMPLEXES CAN BE VISUALIZED USING IMMUNOFLUORESCENCE

The ideal place to look for immune complexes is in the affected organ.

Tissue samples may be examined by immunofluorescence for the presence of immunoglobulin and complement. The composition, pattern, and particular area of tissue affected all provide useful information on the severity and prognosis of the disease. For example:

- patients with the continuous, granular, subepithelial deposits of IgG found in membranous glomerulonephritis have a poor prognosis;
- in contrast, those whose complexes are localized in the mesangium have a good prognosis.

Not all tissue-bound complexes give rise to an inflammatory response; for example, in SLE complexes are frequently found in skin biopsies from normal-looking skin, as well as from inflamed skin.

Assays for immune complexes in serum are more readily performed than in-situ immunofluorescence, although the results have to be interpreted carefully (Method box 25.1)

METHOD BOX 25.1
Assays for circulating immune complexes

Circulating complexes are found in two separate compartments:

- bound to erythrocytes; and
- free in plasma.

Erythrocyte-bound complexes are less likely to be damaging, so it is of more interest to determine the level of free complexes. Care is required when collecting the sample – bound complexes can easily be released during clotting by the action of factor I. To obtain accurate assays of free complexes, the erythrocytes should be rapidly separated from the plasma to prevent the release of bound complexes.

Circulating complexes are often identified by their affinity for complement C1q, using either radiolabeled C1q or solid-phase C1q. Complexes can also be detected by their low solubility in polyethylene glycol (Figs 1 and 2).

An assay for soluble immune complexes based on polyethylene glycol (PEG)

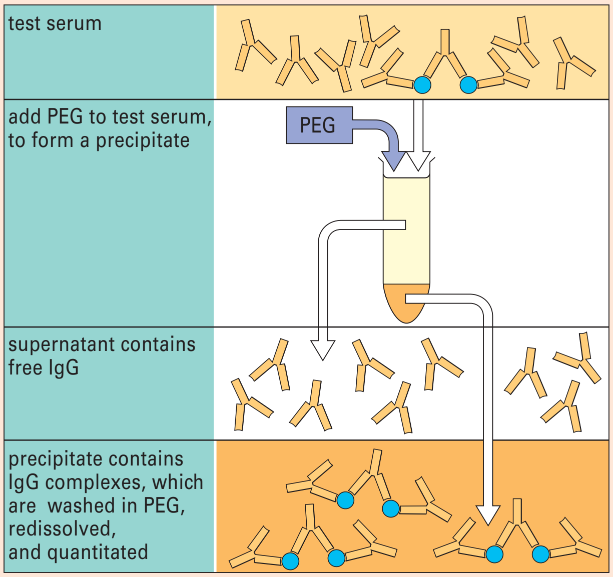
test serum
add PEG to test serum, to form a precipitate
PEG
supernatant contains free IgG
precipitate contains IgG complexes, which are washed in PEG, redissolved, and quantitated

Fig. 1 Polyethylene glycol (PEG) is added to the test serum containing IgG complexes and IgG monomer. When the concentration of PEG reaches 2%, complexes are selectively precipitated; the free antibody remains in solution. The test tube is then centrifuged and the complexes form a pellet at the bottom. The supernatant containing free antibody is removed. The precipitate is washed and redissolved so that the amount of complexed IgG can be measured (e.g. by single radial immunodiffusion, nephelometry, or radioimmunoassay (Fig. 2).

Radioimmunoassay for soluble immune complexes

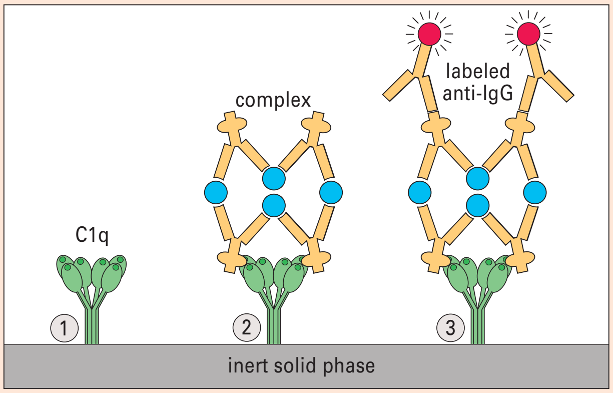

Fig. 2 A three-layer radioimmunoassay for immune complexes based on the use of C1q. (**1**) C1q is linked to an inert solid-phase support, usually a polystyrene tube or plate. (**2**) Serum containing complexes is added. The complexes bind to the solid-phase C1q by means of the array of Fc regions presented to the C1q. (**3**) Radiolabeled anti-IgG antibody is added. The amount of radioactivity remaining on the solid phase after washing is measured in a gamma-counter, and is used to calculate the amount of complex bound to the C1q.

FURTHER READING

Agnello V. Immune complex assays in rheumatic diseases. Hum Pathol 1983;14:343–349.

Arthus M. Injections répétées de sérum de cheval chez le lapin. C R Seances Soc Biol Filiales 1903;55:817.

Birmingham DJ, Herbert LA, Cosio FG, et al. Immune complex erythrocyte complement receptor interactions in vivo during induction of glomerulonephritis in non-human primates. J Lab Clin Med 1990;116:242–252.

Boackle SA, Holer VM, Karp DR. CD21 augments antigen presentation in immune individuals. Eur J Immunol 1997;27:122–129.

Boruchov AM, Heller G, Veri MC, et al. Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. J Clin Invest 2005; 115:2914–2923.

Bruhns P, Samuelsson A, Pollard JW, Ravetch JV. Colony-stimulating factor-1-dependent macrophages are responsible for IVIG protection in antibody-induced autoimmune disease. Immunity 2003;18:573–581

Clynes R, Maizes JS, Guinamard R, et al. Modulation of immune complex-induced inflammation in vivo by the co-ordinate expression of activation and inhibitory Fc receptors. J Exp Med 1999;189:179–185.

Cornacoff JB, Hebert LA, Smead WL, et al. Primate erythrocyte immune complex clearing mechanism. J Clin Invest 1983;71:236– 247.

Czop J, Nussenzweig V. Studies on the mechanism of solubilization of immune precipitates by serum. J Exp Med 1976;143:615–630.

Davies KA, Hird V, Stewart S, et al. A study of in vivo immune complex formation and clearing in man. J Immunol 1990;144:4613–4620.

Davies KA, Peters AM, Beynon HLC, Walport MJ. Immune complex processing in patients with systemic lupus erythematosus – in vivo imaging and clearance studies. J Clin Invest 1992;90:2075–2083.

Davies KA, Chapman PT, Norsworthy PJ, et al. Clearance pathway of soluble immune complexes in the pig. Insights into the adaptive nature of antigen clearance in humans. J Immunol 1995;155:5760–5768.

Davies KA, Schifferli JA, Walport MJ. Complement deficiency and immune complex diseases. Springer Semin Immunopathol 1994;15:397–416.

Dixon FJ, Joseph D, Feldman JD, et al. Experimental glomerulonephritis: the pathogenesis of a laboratory model resembling the spectrum of human glomerulonephritis. J Exp Med 1961;113:899–919.

Dixon FJ, Vazquez JJ, Weigle WO, et al. Pathogenesis of serum sickness. Arch Pathol 1958;65:18–28.

Emlen W, Carl V, Burdick CG. Mechanism of transfer of immune complexes from red blood cell CR1 to monocytes. Clin Exp Immunol 1992;89:8–17.

Finbloom DS, Magilvary DB, Harford JB, et al. Influence of antigen on immune complex behaviour in mice. J Clin Invest 1981;68:214–224.

Fukuyama H, Nimmerjahn F, Ravetch JV. The inhibitory Fc gamma receptor modulates autoimmunity by limiting the accumulation of immunoglobulin G+ anti-DNA plasma cells. Nat Immunol 2005;6:99–106.

Heidelberger M. Quantitative chemical studies on complement or alexin. J Exp Med 1941;73:681–709.

Inman RD. Immune complexes in SLE. Clin Rheum Dis 1982;8:49–62.

Johnston A, Auda GR, Kerr MA, et al. Dissociation of primary antigen–antibody bonds is essential for complement mediated solubilization of immune complexes. Mol Immunol 1992;29:659–665.

Kijlstrea H, Van Es LA, Daha MR. The role of complement in the binding and degradation of immunoglobulin aggregates by macrophages. J Immunol 1979;123:2488–2493.

Lachmann PJ. Complement deficiency and the pathogenesis of autoimmune complex disease. Chem Immunol 1980;49:245–263.

Lucisano Valim M, Lachmann PJ. The effects of antibody isotype and antigenic epitope density on the complement-fixing activity of immune complexes: a systematic study using chimaeric anti-NIP antibodies with human Fc regions. Clin Exp Immunol 1991;84:1–8.

McGaha TL, Sorrentino B, Ravetch JV. Restoration of tolerance in lupus by targeted inhibitory receptor expression. Science 2005 Jan 28;307:590–593.

McKenzie SE, Taylor SM, Malladi P, et al. The role of the human Fc receptor FcγRIIA in the immune clearance of platelets: a transgene mouse model. J Immunol 1999;162:4311–4318.

Miller GW, Nussenzweig V. A new complement function: solubilization of antigen–antibody aggregates. Proc Natl Acad Sci 1975;72:418–422.

Moll T, Nitschke L, Carroll M, et al. A critical role for Fc gamma RIIB in the induction of rheumatoid factors. J Immunol 2004;173:4724–4728.

Olsson M, Bruhns P, Frazier WA, et al. Platelet homeostasis is regulated by platelet expression of CD47 under normal conditions and in passive immune thrombocytopenia. Blood 2005;105:3577–3582.

Park SY, Ueda S, Ohno H, et al. Resistance of Fc receptor-deficient mice to fatal glomerulonephritis. J Clin Invest 1998;102:1229–1238.

Qiao J-H, Castellani LW, Fishbein MC, et al. Immune complex-mediated vasculitis increases coronary artery lipid accumulation in autoimmune-prone MRL mice. Arterioscler Thromb 1993;13:932–943.

Ravetch JV. Fc receptors. Curr Opin Immunol 1997;9:121–125.

Ravetch JV. A full complement of receptors in immune complex diseases. J Clin Invest 2002;110:1759–1761.

Schifferli JA, Ng YC, Peters DK. The role of complement and its receptor in the elimination of immune complexes. N Engl J Med 1986;315:488–495.

Sylvestre DL, Ravetch JV. A dominant role for mast cell Fc receptors in the Arthus reaction. Immunity 1996;5:387–390.

Takata Y, Tamura N, Fujita T. Interaction of C3 with antigen–antibody complexes in the process of solubilisation of immune precipitates. J Immunol 1984;132:2531–2537.

Terino FL, Powell MS, McKenzie IF, Hogarth PM. Recombinant soluble human FcγRII: production, characterization, and inhibition of the Arthus reaction. J Exp Med 1993;178:1617– 1628.

Theofilopoulos AN, Dixon FJ. The biology and detection of immune complexes. Adv Immunol 1979;28:89–220.

Warren JS, Yabroff KR, Remick DG, et al. Tumour necrosis factor participates in the pathogenesis of acute immune complex alveolitis in the rat. J Clin Invest 1989; 84:1873–1882.

Waxman FJ, Hebert LE, Cornacoff JB, et al. Complement depletion accelerates the clearance of immune complexes from the circulation of primates. J Clin Invest 1984;74:1329–1340.

Whaley K. Complement and immune complex diseases. In: Whaley K, ed. Complement in health and disease. Lancaster: MTP Press Ltd; 1987.

Williams RC. Immune complexes in clinical and experimental medicine. Massachusetts: Harvard University Press; 1980.

World Health Organization Scientific Group. Technical report 606. The role of immune complexes in disease. Geneva: WHO; 1977.

Critical thinking: Type III serum sickness after factor IX administration (see p. 503 for explanations)

An 8-year-old boy with factor IX deficiency has had repeated episodes of bleeding into his joints and skin, despite requiring administration of factor IX. Ten days after receiving a dose, he developed fever, swelling of multiple joints, and a skin rash. On physical exam, his temperature was 39°C, he had a diffuse maculopapular skin rash involving his torso and extremities, and both elbows and knees were red, warm, and appeared inflamed. His mother thought the appearance and distribution were very different from the typical appearance after either minor trauma or bleeding into his joints, which he had sustained on multiple previous occasions. His pediatrician ordered the following tests (results shown in Table 1) and prescribed a short course of corticosteroids.

1. What immunologic mechanisms are involved in this inflammatory reaction after the boy received the factor IX?
2. Why were corticosteroids prescribed?
3. What is the likelihood that this type of reaction will develop again?
4. What measures would you take to prevent this reaction from occurring again?

variable	result	normal
C3 (mg/dl)	38	85–155
C4 (mg/dl)	4	12–45
anti-nuclear antibody	negative	
hemoglobin (g/dl)	11.2	
white cell count (cells/mm^3)	11 000	
eosinophils (%)	1	

Table 1

He responds to treatment and his symptoms resolve, but 1 year later his mother notices that his face is swollen in the morning and his feet are swollen at the end of the day. Otherwise the boy feels well.

On physical exam, his blood pressure is elevated at 140/90 mmHg and his ankles are very edematous. His joints do not appear inflamed and the skin does not show either evidence of recent bleeding or inflammation. Results of tests are shown in Table 2.

5. What immunologic mechanisms are involved in this inflammatory reaction after the boy received the factor IX? How do they differ from the previous episode?
6. What is the likelihood that this type reaction will develop again?
7. What measures would you take to prevent this reaction again?

variable	result	normal
C3 (mg/dl)	142	85–155
C4 (mg/dl)	44	12–45
anti-nuclear antibody	negative	
hemoglobin (g/dl)	11.6	
white cell count (cells/mm^3)	8600	
eosinophils (%)	< 1	
albumin (g/dl)	2.5	3.5–5.5
urine protein (g/24 h)	8	< 0.2

Table 2

CHAPTER

26

Hypersensitivity (Type IV)

SUMMARY

- **DTH reflects the presence of antigen-specific CD4 T cells.**
- **There are three variants of type IV hypersensitivity reaction** – contact, tuberculin, and granulomatous.
- **Contact hypersensitivity occurs at the point of contact with an allergen.** Langerhans' cells internalize and process epicutaneously applied hapten and migrate to the draining lymph nodes where they present it to antigen-specific T cells. Cytokines produced by immune-competent skin cells (e.g. keratinocytes, Langerhans' cells, T cells) recruit antigen-non-specific T cells and macrophages.
- **Tuberculin-type hypersensitivity is induced by soluble antigens from a variety of organisms.** It is useful as a diagnostic test for exposure to a number of infectious agents.
- **Granulomatous hypersensitivity is clinically the most important form of type IV hypersensitivity.** Persistence of antigen leads to differentiation of macrophages to epithelioid cells, and fusion to form giant cells. This pathological response is termed a granulomatous reaction and it results in tissue damage. Granuloma formation is driven by T cell activation of macrophages, and is dependent on TNF. Inhibition of TNF leads to breakdown in granulomas.
- **Many chronic diseases manifest type IV granulomatous hypersensitivity.** These include tuberculosis, leprosy, schistosomiasis, sarcoidosis, and Crohn's disease.

DTH REFLECTS THE PRESENCE OF ANTIGEN-SPECIFIC CD4 T CELLS

Delayed-type hypersensitivity (DTH) is a T cell-mediated inflammatory response in which the stimulation of antigen-specific effector T cells leads to macrophage activation and localized inflammation and edema within tissues. This effector T cell response is:

- a normal component of adaptive immunity; and
- essential for the control of intracellular and other pathogens.

However, if the response is excessive it can damage host tissues.

The T cell response may be made to exogenous agents, such as microbial antigens or sensitizing chemicals, or to self-antigens. Typically T cells are sensitized to the foreign antigen during infection with the pathogen or by absorption of a contact sensitizing agent across the skin.

Q. Where in the body are T cells sensitized and how?

A. Typically, T cells are sensitized in the T cell areas of secondary lymphoid tissues, by dendritic cells.

Subsequent exposure of the sensitized individual to the exogenous antigen, either injected intradermally or applied to the epidermis, results in the recruitment of antigen-specific T cells to the site and the development of a local inflammatory response over 24–72 hours.

If the foreign antigen persists in the tissues, chronic activation of T cells and macrophages may lead to granuloma formation and tissue damage.

If the antigen is an organ-specific self antigen, autoreactive T cells may produce localized cellular inflammation resulting in autoimmune disease, such as type I diabetes mellitus.

According to the Coombs and Gell classification, type IV or DTH reactions take more than 12 hours to develop and involve cell-mediated immune reactions rather than antibody responses to antigens. Some other hypersensitivity reactions may straddle this definition with:

- a rapid antibody-mediated phase; and
- a later cell-mediated phase.

For example, the late-phase IgE-mediated reaction may peak 12–24 hours after contact with an allergen, and cells, such as helper T (TH2) cells and eosinophils, contribute to the inflammation as well as IgE (see Chapter 22).

In contrast to other forms of hypersensitivity, type IV hypersensitivity is transferred from one animal to another by T cells, particularly CD4 TH1 cells in mice, rather than by serum. Therefore it can occur in antibody-deficient humans, but is lost with the decline in CD4 T cells in HIV infection and AIDS.

Type IV hypersensitivity reflects the presence of antigen-specific CD4 T cells and is associated with protective immunity against intracellular and other pathogens.

However, there is not a complete correlation between type IV hypersensitivity and protective immunity, and progressive infections can develop in the presence of strong DTH reactivity.

THERE ARE THREE VARIANTS OF TYPE IV HYPERSENSITIVITY REACTION

Three variants of type IV hypersensitivity reaction are recognized (Fig. 26.1):

- **contact hypersensitivity** and **tuberculin-type hypersensitivity** both occur within 72 hours of antigen challenge;
- **granulomatous hypersensitivity reactions** develop over a period of 21–28 days – the granulomas are formed by the aggregation of macrophages and lymphocytes and may persist for weeks – this is the most important type of type IV hypersensitivity response for producing clinical consequences.

These three types of delayed hypersensitivity were originally distinguished according to the reaction they produced when antigen was applied directly to the skin (epicutaneously) or injected intradermally. The degree of the response is usually assessed in animals by measuring thickening of the skin.

Q. What causes the skin to become thickened during a chronic immune response?

A. The migration of lymphocytes and macrophages into the dermis, the proliferation of cells in the dermis in response to cytokines, and the deposition of new extracellular matrix components can all contribute to skin thickening.

The local response is also accompanied by a variety of systemic immune responses, such as T cell proliferation and synthesis of cytokines including interferon-γ (IFNγ).

CONTACT HYPERSENSITIVITY OCCURS AT THE POINT OF CONTACT WITH AN ALLERGEN

Contact hypersensitivity is characterized by an eczematous reaction in the skin at the point of contact with an allergen (Fig. 26.2). It is often seen following contact with agents such as nickel, chromate, rubber accelerators, and pentadecacatechol, which is the sensitizing agent in poison ivy. This is distinct from the non-immune-mediated inflammatory response to irritants.

The variants of delayed hypersensitivity

delayed reaction	maximal reaction time
contact	48–72 hours
tuberculin	48–72 hours
granulomatous	21–28 days

Fig. 26.1 Contact and tuberculin-type hypersensitivity have a similar time course and are maximal at 48–72 hours. In certain circumstances (e.g. with insoluble antigen) granulomatous reactions also develop at 21–28 days (e.g. skin testing in leprosy).

Clinical and patch test appearances of contact hypersensitivity

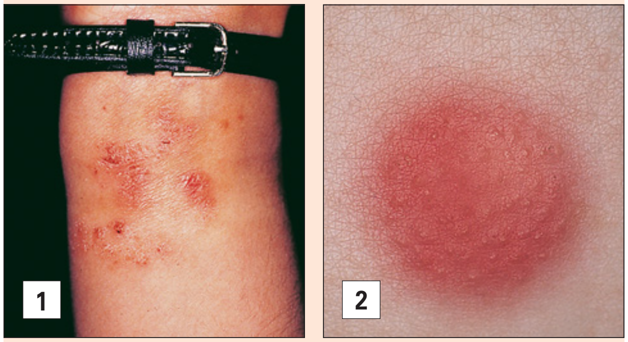

Fig. 26.2 (**1**) The eczematous area at the wrist is due to sensitivity to nickel in the watch-strap buckle. (**2**) The suspected allergy may be confirmed by applying potential allergens, in the relevant concentrations and vehicles, to the patient's upper back (patch testing). A positive reaction causes a localized area of eczema at the site of the offending allergen 2–4 days after application.

The immunologically active components of sensitizing agents are called **haptens**. Haptens are:

- too small to be antigenic by themselves, having a molecular weight often less than 1 kDa;
- lipophilic and penetrate the epidermis where they conjugate, most often covalently, to self proteins to form **neo-antigens**.

The sensitizing potential of a hapten cannot reliably be predicted from its chemical structure, though there is some correlation with the ability of the molecule to penetrate the skin and the number of haptens attached to the carrier protein.

Some potent haptens, such as dinitrochlorobenzene (DNCB), sensitize nearly all individuals and can be used to assess cell-mediated immunity. DNCB applied to the skin binds to epidermal proteins and to MHC-linked peptides through the -NH_2 groups of lysine.

Langerhans' cells and keratinocytes have key roles in contact hypersensitivity

The Langerhans' cell is the principal APC

Contact hypersensitivity is primarily an epidermal reaction, and the dendritic **Langerhans' cell**, located in the suprabasal epidermis, is the principal antigen-presenting cell (APC) involved (Fig. 26.3).

Langerhans' cells (see Chapter 2) are specialized dendritic cells derived from bone marrow and express:

- CD1;
- MHC class II molecules;
- langerin; and
- surface receptors for Fc and complement.

Electron microscopy shows **Birbeck granules**, which are organelles derived from cell membrane and are characteristic of Langerhans' cells (see Fig. 26.3).

Langerhans' cells

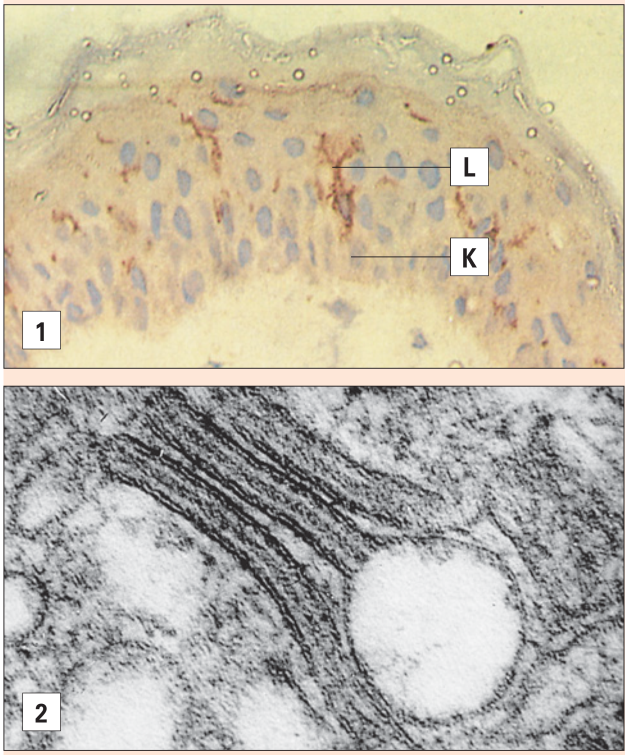

Fig. 26.3 (**1**) These dendritic cells constitute 3% of all cells in the epidermis. They express a variety of surface markers, which allow them to be visualized. Here they have been revealed in a section of normal skin using a monoclonal antibody that reacts with the CD1 antigen (counterstained with Mayer's hemalum). (L, Langerhans' cell; K, keratinocyte) × 312. (**2**) Electron micrograph of a Langerhans' cell showing the characteristic 'Birbeck granule'. This organelle is a plate-like structure with a distinct central striation and often has a bleb-like extension at one end. × 132 000.

Langerhans' cells are inactivated by ultraviolet B, which can therefore prevent or alleviate the effects of contact hypersensitivity.

Langerhans' cells take up hapten-modified proteins by micropinocytosis, and under the influence of interleukin-1 (IL-1) and tumor necrosis factor (TNF) from keratinocytes and other cells:

- undergo maturation;
- increase the expression of MHC and co-stimulatory molecules; and
- migrate to draining lymph nodes.

Keratinocytes produce a range of cytokines important to the contact hypersensitivity response

Keratinocytes provide the structural integrity of the epidermis and have a central role in epidermal immunology.

Keratinocytes can be activated by a number of stimuli, including sensitizing agents and irritants. They may express MHC class II molecules and intercellular adhesion molecule-1 (ICAM-1) in the cell membrane.

Activated keratinocytes produce a wide range of cytokines, including:

- TNF, IL-1, and granulocyte–macrophage colony stimulating factor (GM-CSF), which may activate Langerhans' cells;
- IL-3, which can also activate Langerhans' cells, co-stimulate proliferative responses, recruit mast cells, and induce the secretion of immunosuppressive cytokines, such as IL-10 and transforming growth factor-β (TGFβ) – the latter dampen the immune response and may induce clonal anergy or immunological unresponsiveness in TH1 cells.

A contact hypersensitivity reaction has two stages – sensitization and elicitation

Sensitization stimulates a population of memory T cells

Sensitization takes 10–14 days in humans.

Once absorbed, the hapten combines with a protein and is internalized by epidermal Langerhans' cells, which leave the epidermis and migrate as veiled cells through the afferent lymphatics to the paracortical areas of regional lymph nodes. Here they present processed hapten–protein conjugates in association with MHC class II molecules to CD4$^+$ lymphocytes, producing:

- effector/memory CD4$^+$ T cells; and
- regulatory CD4$^+$ T cells (Fig. 26.4).

In addition, MHC class I-restricted CD8$^+$ T cells are important in contact hypersensitivity responses in humans and mice and are the major effector cells for some allergens. For example, lipid-soluble urushiol from poison ivy can enter the cytoplasm of APCs and access the MHC class I processing pathway, leading to the activation of allergen-specific CD8$^+$ T cells. Activated T cells change the pattern of adhesion molecules on their surface by down-regulating the chemokine receptor, CCR7, and CD62L.

Q. What effect will loss of CCR7 and CD62L have on T cell function?

A. CD62L promotes adhesion of lymphocytes to high endothelial venules and CCR7 allows the cells to respond to CCL21 expressed in secondary lymphoid tissues (see Appendix 4). Hence cells lacking these receptors will lose their propensity to traffic into lymph tissues.

The expression of leukocyte functional antigen-1 (LFA-1), very late antigen-4 (VLA-4), and the chemokine receptors CXCR3 and CCR5 is increased. As a result the activated/memory T cells remain within the circulation rather than trafficking through lymphoid tissue, and are able to bind to adhesion molecules on the endothelium of inflamed tissues.

Elicitation involves recruitment of CD4$^+$ lymphocytes and monocytes

The application of a contact allergen leads to:

- rapid expression proinflammatory cytokines; and
- recruitment of effector T cells and monocytes to the site (Fig 26.5).

There is induction of mRNA for TNF, IL-1β, and GM-CSF in Langerhans' cells within 30 minutes of exposure

Sensitization phase of contact hypersensitivity

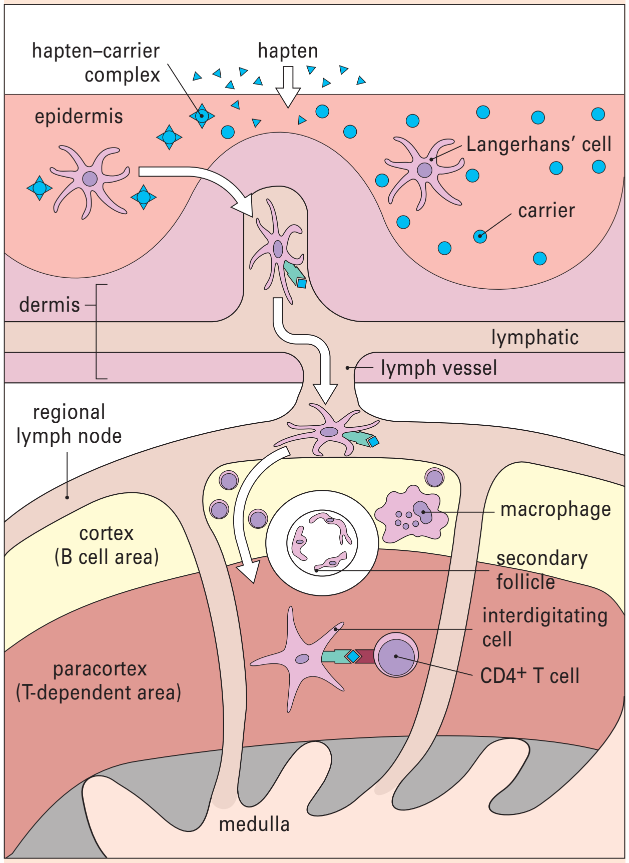

Fig. 26.4 The hapten forms a hapten–carrier complex in the epidermis. Langerhans' cells internalize the antigen, undergo maturation, and migrate via afferent lymphatics to the paracortical area of the regional lymph node where peptide–MHC molecule complexes on the surface of the Langerhans' cell can also be directly haptenated. As interdigitating cells, they present antigen to CD4$^+$ T cells.

to allergen, and increased transcription of mRNA for IL-1α, macrophage inflammatory protein-2 (CXCL2), and interferon-induced protein-10 (CXCL10) by keratinocytes.

TNF and IL-1 are potent inducers of endothelial cell adhesion molecules, including:

- E-selectin and vascular cell adhesion molecule-1 (VCAM-1) within 2 hours; and
- ICAM-1 within 8 hours (Fig. 26.6).

VCAM-1 and ICAM-1 are the receptors for VLA-4 and LFA-1, respectively, on the surface of effector/memory T cells and contribute to their recruitment across the endothelium. These locally released cytokines and chemokines also produce a gradient signal for the movement of mononuclear cells towards the dermoepidermal junction and epidermis.

During this initial response to allergen, Langerhans' cells migrate from the epidermis to the dermis, where they may activate any resident memory T cells in the dermis, and to the draining lymph nodes.

The earliest histological change, seen after 4–8 hours, is the appearance of mononuclear cells around blood vessels. Macrophages and lymphocytes invade the dermis and epidermis, peaking at 48–72 hours (Fig. 26.7).

The recruitment of memory T cells is antigen non-specific, with less than 1% of infiltrating lymphocytes bearing hapten-specific αβ T cell receptors. However, these hapten-specific T cells are stimulated by dermal Langerhans' cells expressing hapten–peptide complexes to expand and to increase the expression of adhesion molecules. This leads to the retention of hapten-specific T cells at the inflamed site.

Most infiltrating lymphocytes are CD4$^+$ TH1-like T cells secreting IFNγ, with a small proportion of CD8$^+$ T cells. The CD8$^+$ T cells contribute to inflammation by both a direct cytolytic effect on epidermal cells and the release of IFNγ.

Q. What is the name of this class of cytotoxic cells?
A. They are Tc1 cells (see Chapter 11).

Effector αβ T cells are essential for experimental contact sensitivity in mice, but NK T cells and γδ T cells also contribute to the induction and elicitation of the response.

Interestingly, hapten-specific IgM antibodies from B-1 cells are also important during the elicitation phase in mice by activating complement and recruiting T cells to the challenge site.

Experiments in gene-targeted mice show that selectins, ICAM-1, and the integrins, LFA-1 and VLA-4, are all required for the elicitation of contact and delayed hypersensitivity.

Suppression of the inflammatory reaction is mediated by multiple mechanisms

The reaction to cutaneous application of sensitizer wanes after 48–72 hours. This is due to the removal of antigenic stimulus following degradation of the hapten–conjugate and a variety of inhibitory mechanisms (see Fig. 26.6) including:

- macrophages and keratinocytes produce PGE, which inhibits IL-1 and IL-2 production;
- TGFβ from dermal mast cells, activated keratinocytes, and lymphocytes inhibits inflammation and blocks the proliferative effects of IL-1 and IL-2;
- IL-10 downregulates MHC class II molecule expression, and suppresses cytokine production and antigen-specific proliferation by TH1 cells;
- CD4$^+$ regulatory T cells directly inhibit effector T cells;
- external factors may also be involved – in mice UV light induces a specific inhibitor of IL-1 activity.

TUBERCULIN-TYPE HYPERSENSITIVITY

Tuberculin-type hypersensitivity is induced by soluble antigens from a variety of organisms. Tuberculin-type hypersensitivity was originally described by Koch. He observed that if patients with tuberculosis were injected subcutaneously with a tuberculin culture filtrate (antigens

Elicitation phase of contact hypersensitivity

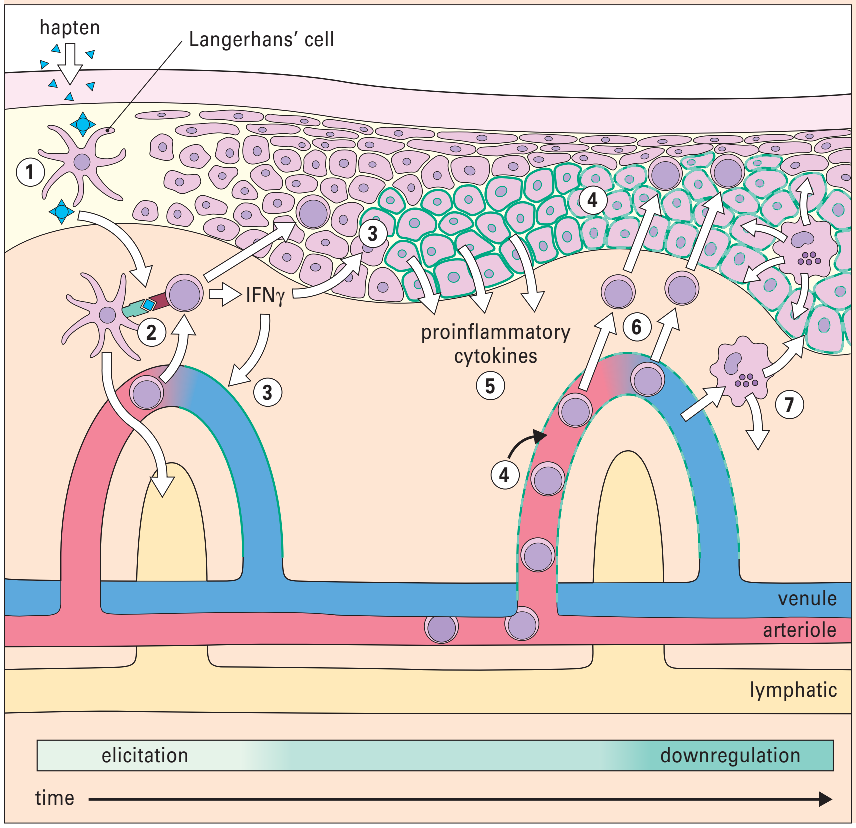

Fig. 26.5 Langerhans' cells carrying the hapten–carrier complex (**1**) move from the epidermis to the dermis, where they present the hapten–carrier complex to memory CD4$^+$ T cells (**2**). Activated CD4$^+$ T cells release IFNγ, which induces expression of ICAM-1 (**3**) and, later, MHC class II molecules (**4**) on the surface of keratinocytes and on endothelial cells of dermal capillaries, and activates keratinocytes, which release proinflammatory cytokines such as IL-1, IL-6, and GM-CSF (**5**). Non-antigen-specific CD4$^+$ T cells are attracted to the site by cytokines (**6**) and may bind to keratinocytes via ICAM-1 and MHC class II molecules. Activated macrophages are also attracted to the skin, but this occurs later. Thereafter the reaction starts to downregulate. This downregulation may be influenced by eicosanoids such as prostaglandin E (PGE) produced by activated keratinocytes and macrophages (**7**).

derived from the causative agent, *Mycobacterium tuberculosis*) they reacted with fever and generalized sickness. An area of hardening and swelling developed at the site of injection.

Soluble antigens from a number of other organisms, including *Mycobacterium leprae* and *Leishmania tropica*, induce similar reactions in sensitized people.

The skin reaction is frequently used to test for T cell-mediated responses to the organisms following previous exposure (Fig. 26.8).

This form of hypersensitivity may also be induced by T cell responses to non-microbial antigens, such as beryllium and zirconium.

The tuberculin skin test reaction involves monocytes and lymphocytes

The tuberculin skin test is an example of the recall response to soluble antigen previously encountered during infection. Dendritic cells infected with *M. tuberculosis* in the lung undergo maturation and migrate to the draining mediastinal lymph nodes where they activate CD4$^+$ and CD8$^+$ T cells.

Q. How can dendritic cells activate CD8$^+$ T cells?

A. This involves the process of cross-presentation (see Fig. 7.15).

Following intradermal tuberculin challenge in a previously infected individual, mycobacteria-specific memory T cells are recruited and activated to secrete IFNγ, which activates macrophages to produce TNFα and IL-1. These proinflammatory cytokines and chemokines from T cells and macrophages act on endothelial cells in dermal blood vessels to induce the sequential expression of the adhesion molecules E-selectin, ICAM-1, and VCAM-1. These molecules bind receptors on leukocytes and recruit them to the site of the reaction.

The initial influx at 4 hours is of neutrophils, but this is replaced at 12 hours by monocytes and T cells. The infiltrate, which extends outwards and disrupts the collagen bundles of the dermis, increases to a peak at 48 hours. CD4$^+$ T cells outnumber CD8$^+$ cells by about 2 to 1. CD1$^+$ cells (Langerhans'-like cells, but lacking Birbeck granules) are also found in the dermal infiltrate at 24 and 48 hours, and a few CD4$^+$ cells infiltrate the epidermis between 24 and 48 hours.

Monocytes constitute 80–90% of the total cellular infiltrate. Both infiltrating lymphocytes and macrophages express MHC class II molecules, and this increases the efficiency of activated macrophages as APCs. Overlying keratinocytes express HLA-DR molecules 48–96 hours after the appearance of the lymphocytic infiltrate. These events are summarized in Fig. 26.9.

Macrophages are probably the main APCs in the tuberculin hypersensitivity reaction. However, there are CD1$^+$ cells in the dermal infiltrate, which suggests that Langerhans' cells or other indeterminate dendritic cells may also participate.

Cytokines, prostaglandins, and cellular interactions in contact hypersensitivity

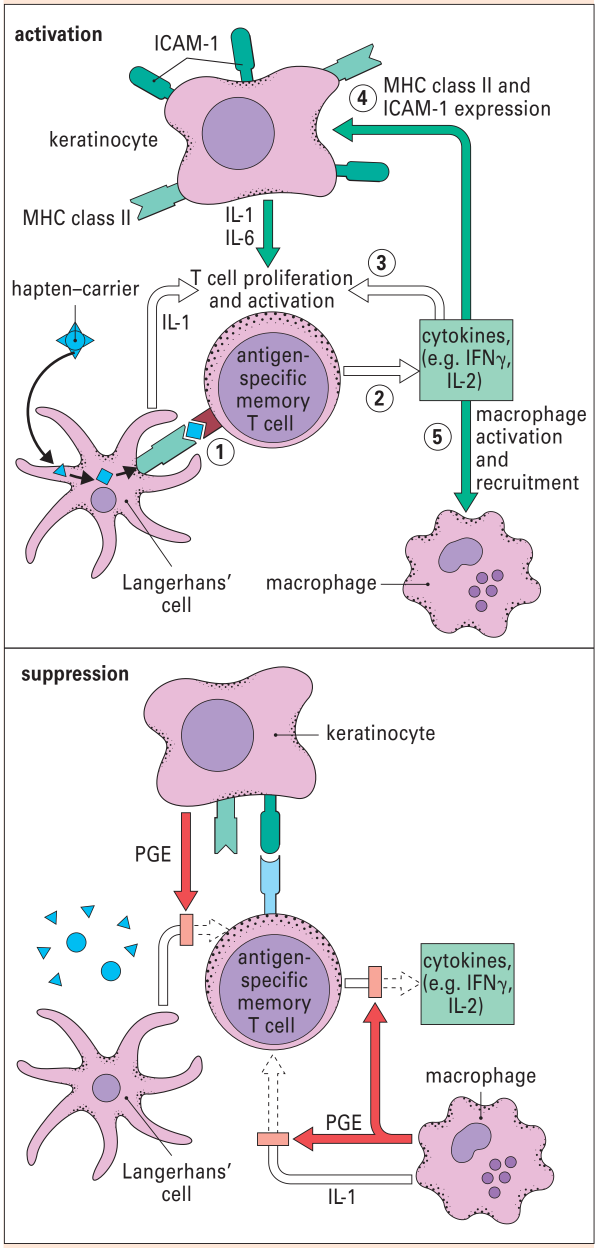
activation
ICAM-1
keratinocyte
④ MHC class II and ICAM-1 expression
MHC class II
IL-1
IL-6
③
hapten–carrier
T cell proliferation and activation
IL-1
antigen-specific memory T cell
cytokines, (e.g. IFNγ, IL-2)
②
①
⑤ macrophage activation and recruitment
Langerhans' cell
macrophage
suppression
keratinocyte
PGE
antigen-specific memory T cell
cytokines, (e.g. IFNγ, IL-2)
macrophage
PGE
Langerhans' cell
IL-1

Fig. 26.6 Cytokines and prostaglandins are central to the complex interactions between Langerhans' cells, CD4$^+$ T cells, keratinocytes, macrophages, and endothelial cells in contact hypersensitivity. The act of antigen presentation (**1**) causes the release of a cascade of cytokines (**2**). This cascade initially results in the activation and proliferation of CD4$^+$ T cells (**3**), the induction of expression of ICAM-1 and MHC class II molecules on keratinocytes and endothelial cells (**4**), and the attraction of further T cells and macrophages to the skin (**3**, **5**). Subsequent PGE production by keratinocytes and macrophages may have an inhibitory effect on IL-1 and IL-2 production. Production of PGE, binding of activated T cells to keratinocytes, and enzymatic and cellular degradation of the hapten–carrier complex all contribute to the downregulation of the reaction.

Histological appearance of the lesion in contact hypersensitivity

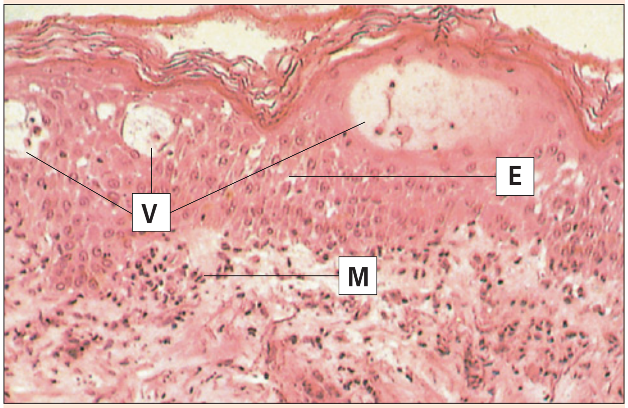

Fig. 26.7 Mononuclear cells (M) infiltrate both dermis and epidermis. The epidermis is pushed outwards and microvesicles (V) form within it due to edema (E). H&E stain. × 130.

Clinical and histological appearances of tuberculin-type sensitivity

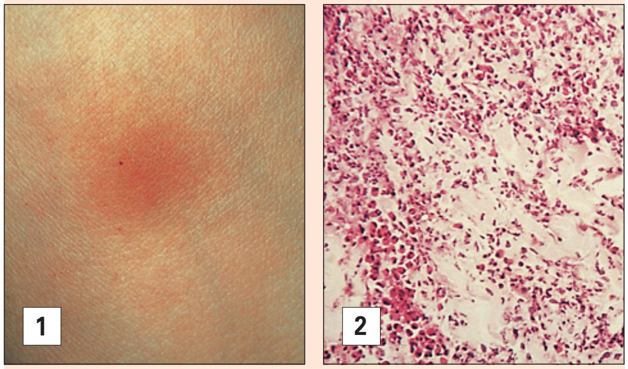

Fig. 26.8 The response to an injection of leprosy bacillus into a sensitized individual is known as the Fernandez reaction. The reaction is characterized by an area of firm red swelling of the skin and is maximal 48–72 hours after challenge (**1**). Histologically (**2**), there is a dense dermal infiltrate of leukocytes. H&E stain. × 80.

Q. What is the function of CD1?

A. It can present lipoprotein and glycolipid antigens to T cells (see Fig. 5.22).

The circulation of immune cells to and from the regional lymph nodes is thought to be similar to that for contact hypersensitivity. The tuberculin lesion normally resolves within 5–7 days, but if there is persistence of antigen in the tissues it may develop into a granulomatous reaction.

Tuberculin-type hypersensitivity

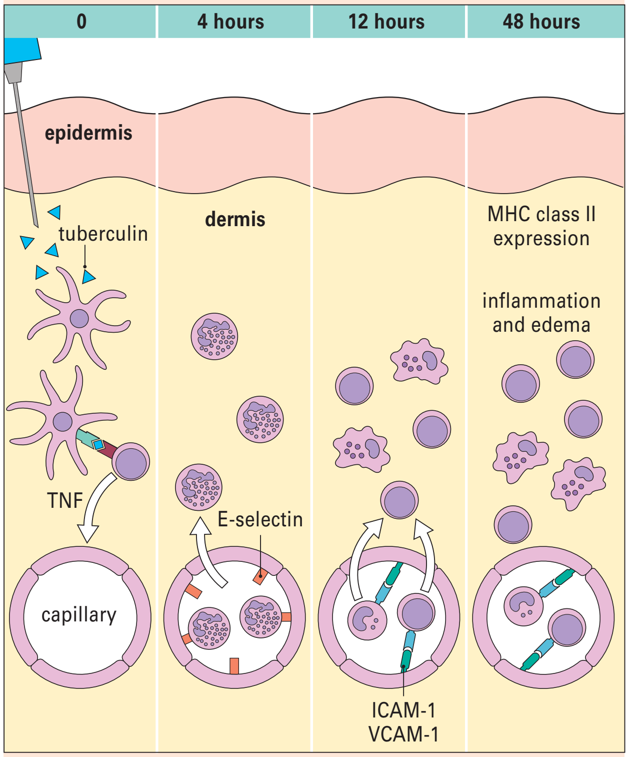

Fig. 26.9 This diagram illustrates cellular movements following intradermal injection of tuberculin. Within 1–2 hours there is expression of E-selectin on capillary endothelium leading to a brief influx of neutrophil leukocytes. By 12 hours ICAM-1 and VCAM-1 on endothelium bind the integrins LFA-1 and VLA-4 on monocytes and lymphocytes, leading to accumulation of both cell types in the dermis. This peaks at 48 hours and is followed by expression of the MHC class II molecules on keratinocytes. There is no edema of the epidermis.

Tuberculin-like DTH reactions are used practically in two ways

First, reaction to soluble antigens from a pathogen demonstrates past infection with that pathogen. Thus, tuberculin reactivity confirms past or latent infection with *M. tuberculosis*, but not necessarily active disease. However, subjects with latent tuberculosis infection have an increased lifelong risk of 7–10% for the reactivation of active tuberculosis.

Second, DTH responses to frequently encountered microbes are a general measure of cell-mediated immunity. This can be tested with intradermal injection of single antigens from common pathogens or vaccine antigens, such as *Candida albicans* or tetanus toxoid, or a multipuncture device, which delivers multiple common microbial antigens in a standardized fashion. Loss of recall responses to specific antigens occurs in a wide range of diseases and infections, including HIV infection, which impair T cell function, and during therapy with corticosteroids or immunosuppressive agents.

GRANULOMATOUS HYPERSENSITIVITY

Granulomatous hypersensitivity is clinically the most important form of type IV hypersensitivity. Granulomatous hypersensitivity causes many of the pathological effects in diseases that involve T cell-mediated immunity. It usually results from the persistence within macrophages of:

- intracellular microorganisms, which are able to resist macrophage killing; or
- other particles that the cell is unable to destroy.

This leads to chronic stimulation of T cells and the release of cytokines. The process results in the formation of **epithelioid cell granulomas** with a central collection of epithelioid cells and macrophages surrounded by lymphocytes.

The histological appearance of the granuloma reaction is quite different from that of the tuberculin-type reaction. However, both types of reaction may be caused by T cells sensitized to similar microbial antigens, for example the antigens of *M. tuberculosis* and *M. leprae* (Fig. 26.10).

Role of the antigen-specific TH lymphocyte in type IV hypersensitivity

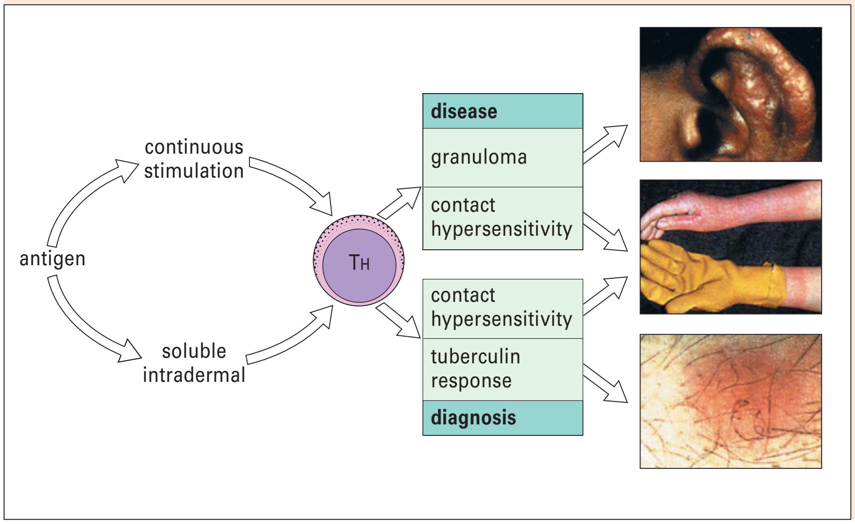

Fig. 26.10 The tuberculin skin reaction (lower photograph, courtesy of Professor JHL Playfair) is the classic diagnostic test for cell-mediated immunity in tuberculosis. If there is continuous antigenic stimulation instead of a single injection of soluble antigen, a granulomatous reaction (upper photograph, courtesy of Dr A du Vivier) or contact hypersensitivity (middle photograph, courtesy of Dr D Sharvill) follows. This granulomatous reaction can also occur if the macrophages cannot destroy the antigen.

Electron micrograph of an epithelioid cell

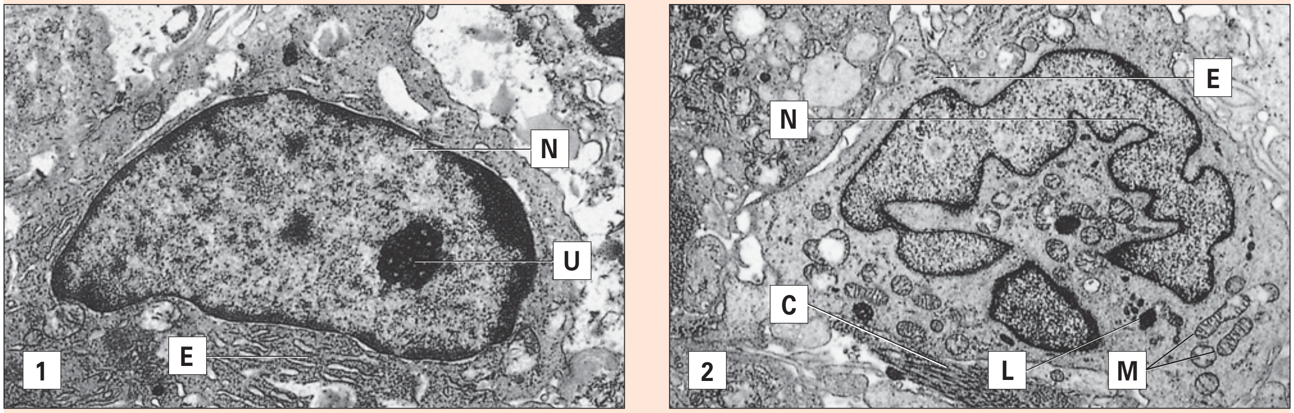

Fig. 26.11 The epithelioid cell is the characteristic cell of granulomatous hypersensitivity. Compare the extent of the endoplasmic reticulum (E) in the epithelioid cell (**1**, × 4800) with that of a tissue macrophage (**2**, × 4800). (C, collagen; L, lysosome; M, mitochondria; N, nucleus; U, nucleolus) (Courtesy of MJ Spencer)

Granulomas occur with chronic infections associated with predominantly TH1-like T cell responses, such as tuberculosis, leprosy, and leishmaniasis, and with TH2-like T cells, as in schistosomiasis.

Immune-mediated granuloma formation also occurs in the absence of infection, as in the sensitivity reactions to zirconium and beryllium, and in sarcoidosis and Crohn's disease where the antigens are unknown.

Foreign body granuloma formation occurs in response to talc, silica, and a variety of other particulate agents, when macrophages are unable to digest the inorganic matter. These non-immunological granulomas may be distinguished by the absence of lymphocytes in the lesion.

Epithelioid cells and giant cells are typical of granulomatous hypersensitivity

Epithelioid cells are large and flattened with increased endoplasmic reticulum (Fig. 26.11). They:

- are derived from activated macrophages under the chronic stimulation of cytokines;
- continue to secrete TNF and thus potentiate continuing inflammation.

Giant cells are formed when epithelioid cells fuse to form multinucleate giant cells (Fig. 26.12), sometimes referred to as **Langhans' giant cells** (not to be confused with the Langerhans' cell discussed earlier). Giant cells have several nuclei, but these are not at the center of the cell. There is little endoplasmic reticulum, and the mitochondria and lysosomes appear to be undergoing degeneration. The giant cell may therefore be a terminal differentiation stage of the monocyte/macrophage line.

The granuloma contains epithelioid cells, macrophages, and lymphocytes

An immunological granuloma typically has a core of epithelioid cells and macrophages, sometimes with giant cells. In some diseases, such as tuberculosis, this central area may have a zone of necrosis, with complete destruction of all cellular architecture.

Clinical and histological appearances of the Mitsuda reaction in leprosy seen at 28 days

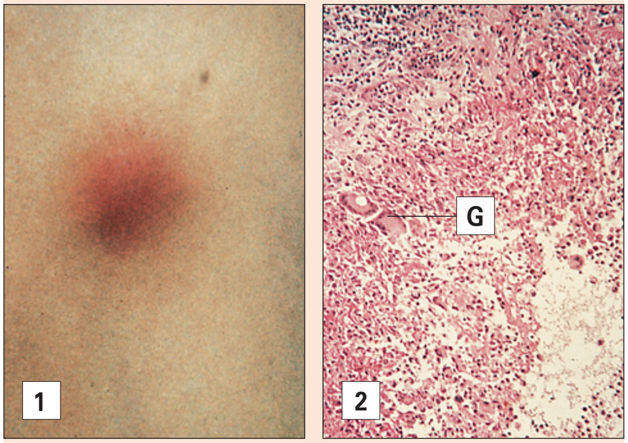

Fig. 26.12 (**1**) The resultant skin swelling (which may be ulcerated) is much harder and better defined than at 48 hours. (**2**) Histology shows a typical epithelioid cell granuloma (H&E stain. × 60). Giant cells (G) are visible in the center of the lesion, which is surrounded by a cuff of lymphocytes. This response is more akin to the pathological processes in delayed hypersensitivity diseases than the self-resolving tuberculin-type reaction. The reaction is due to the continued presence of mycobacterial antigen.

The macrophage/epithelioid core is surrounded by a cuff of lymphocytes, and there may also be considerable fibrosis (deposition of collagen fibers) caused by proliferation of fibroblasts and increased collagen synthesis.

Examples of granulomatous reactions are:

- the **Mitsuda reaction** to *M. leprae* antigens (see Fig. 26.12); or

Delayed hypersensitivity reactions

type	reaction time	clinical appearance	histology	antigen
contact	48–72 hours	eczema	lymphocytes, later macrophages; edema of epidermis	epidermal (e.g. antigen, nickel, rubber, poison ivy)
tuberculin	48–72 hours	local induration	lymphocytes, monocytes, macrophages	intradermal (e.g. tuberculin)
granuloma	21–28 days	hardening (e.g. skin of lung)	macrophages, epithelioid cells, giant cells, fibrosis	persistent antigen or antibody complexes or non-immunoglobulin stimuli (e.g. talc)

Fig. 26.13 The characteristics of type IV reactions comparing contact, tuberculin, and granulomatous reactions.

- the **Kveim test**, where patients who have sarcoidosis react to unknown splenic antigens derived from other patients with sarcoidosis.

The three types of delayed hypersensitivity are summarized in Fig. 26.13.

CELLULAR REACTIONS IN TYPE IV HYPERSENSITIVITY

T cells bearing αβ TCRs are essential

Experiments with gene knockout mice have confirmed that T cells bearing αβ TCRs rather than γδ TCRs are essential for initiating delayed hypersensitivity reactions in response to infection with intracellular bacteria.

Sensitized αβ T cells, stimulated with the appropriate antigen and APCs, undergo lymphoblastoid transformation before cell division (Fig. 26.14). This forms the basis of the lymphocyte stimulation test as a measure of T cell function. Lymphocyte stimulation is accompanied by DNA synthesis and this can be measured by assaying the uptake of radiolabeled thymidine, a nucleoside required for DNA synthesis. Lymphocytes from a patient are stimulated in culture with the suspect antigen to determine whether it induces transformation. It is important to stress that this is a test for T cell memory only, and does not necessarily imply the presence of protective immunity.

Transformed lymphocytes

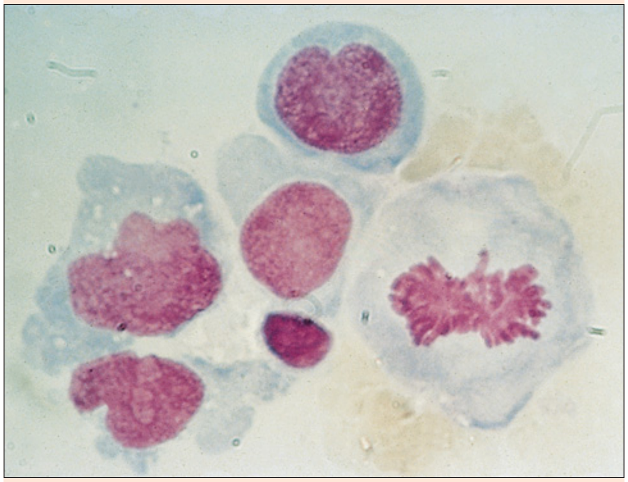

Fig. 26.14 Following stimulation with appropriate antigen, T cells undergo lymphoblastoid transformation before cell division. Blast cells with expanded nuclei and cytoplasm (as well as one lymphocyte in the metaphase of cell division) are shown.

Following activation by APCs, T cells release a number of proinflammatory cytokines, which attract and activate macrophages. These include IFNγ, lymphotoxin-α, IL-3, and GM-CSF.

Q. How can IFNγ cause the attraction of macrophages to an inflammatory site?

A. It causes the production of chemokines, including CCL2, CCL5, and CXCL10, and induces adhesion molecules ICAM-1 and VCAM-1 on the endothelium (see Chapter 6).

This TH1-like pattern of cytokines is enhanced by activation of the T cells in the presence of IL-12 and IL-23, which are released by dendritic cells on exposure to bacterial products. IL-12 suppresses the cytokine response of TH2 cells.

IFNγ is required for granuloma formation in humans

The role of individual cytokines can be analyzed in gene knockout mice deficient for a single cytokine. For example, IFNγ gene knockout mice are unable to activate macrophages and control infection with *M. tuberculosis* (Fig. 26.15).

The absolute requirement of IFNγ for granuloma formation in humans is confirmed by the syndrome of Mendelian susceptibility to mycobacterial disease. Subjects deficient in the IFNγ receptor have markedly increased susceptibility to environmental mycobacteria and the vaccine strain, BCG, and fail to develop granulomas.

TNF and lymphotoxin-α are essential for granuloma formation during mycobacterial infections

TNF and the related cytokine, lymphotoxin-α, are both essential for the formation of granulomas during mycobacterial infections (Fig. 26.16), and act in part through the regulation of chemokine production.

The importance of IFNγ in the activation of macrophages

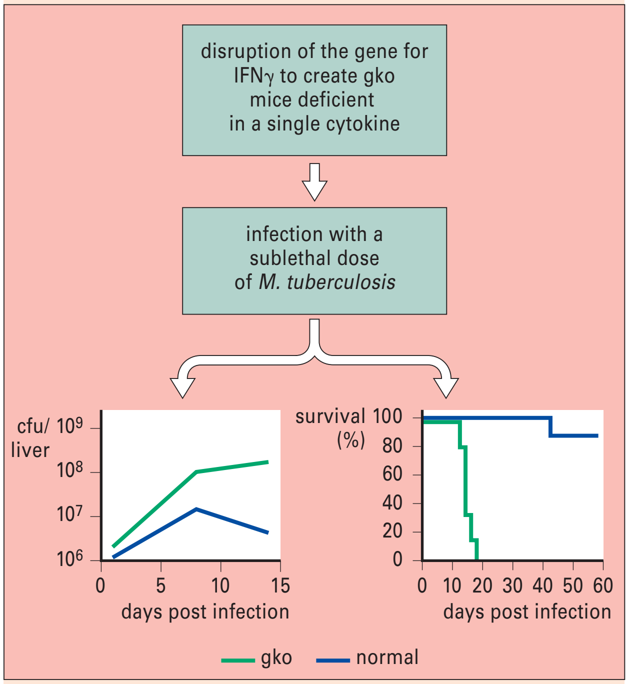

Fig. 26.15 Mice deficient in IFNγ (gene knockout [gko] mice) are unable to activate macrophages in response to infection with an intracellular bacterium. Macrophages initially accumulate at the site of infection, but do not form typical granulomas. Uncontrolled infection (graph, left) causes widespread tissue necrosis and death (graph, right). (cfu, colony forming units of infectious agent in the liver)

Both macrophage- and T cell-derived TNF contribute to this process, but within granulomas activated macrophages become the major source of TNF, driving the differentiation of macrophages into epithelioid cells and the fusion of epithelioid cells to form giant cells (Figs 26.17 and 26.18). The maintenance of granulomas is also dependent on TNF. Consequently, inhibition of TNF activity suppresses the granulomatous inflammation in Crohn's disease and sarcoidosis.

MANY CHRONIC DISEASES MANIFEST TYPE IV GRANULOMATOUS HYPERSENSITIVITY

There are many chronic diseases in humans that manifest type IV hypersensitivity. Most are due to infectious agents such as mycobacteria, protozoa, and fungi, though in other granulomatous diseases such as sarcoidosis and Crohn's disease no infectious agent has been established. Important diseases in this respect include:

- leprosy;
- tuberculosis;
- schistosomiasis;
- sarcoidosis;
- Crohn's disease.

Macrophage differentiation

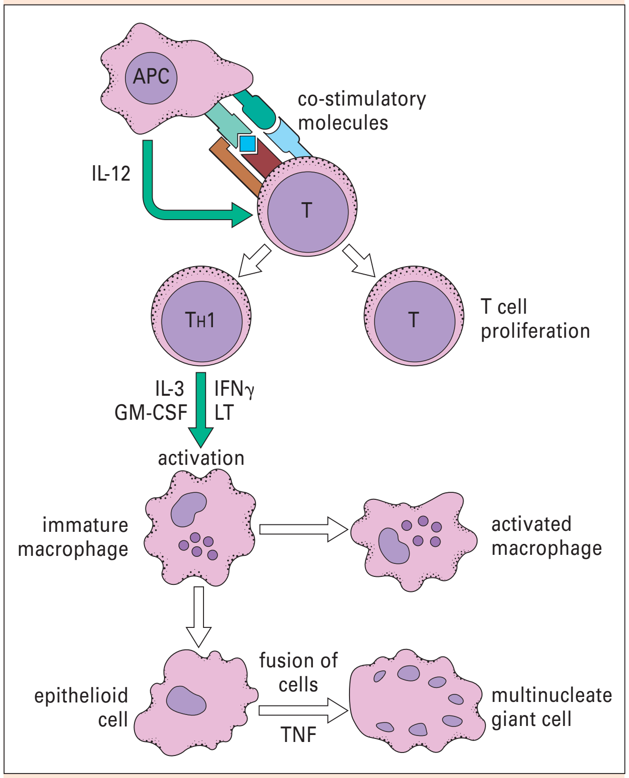
APC
co-stimulatory molecules
IL-12
T
TH1
T
T cell proliferation
IL-3
GM-CSF
IFNγ
LT
activation
immature macrophage
activated macrophage
epithelioid cell
fusion of cells
TNF
multinucleate giant cell

Fig. 26.16 Bacterial products stimulate macrophages to secrete IL-12. Activation of T cells in the presence of IL-12 leads to the release of IFNγ and other cytokines, lymphotoxin (LT), IL-3, and GM-CSF. These cytokines activate macrophages to kill intracellular parasites. Failure to eradicate the antigenic stimulus causes persistent cytokine release and promotes differentiation of macrophages into epithelioid cells, which secrete large amounts of TNFα. Some fuse to form multinucleate giant cells.

The importance of TNF in the formation of granulomas

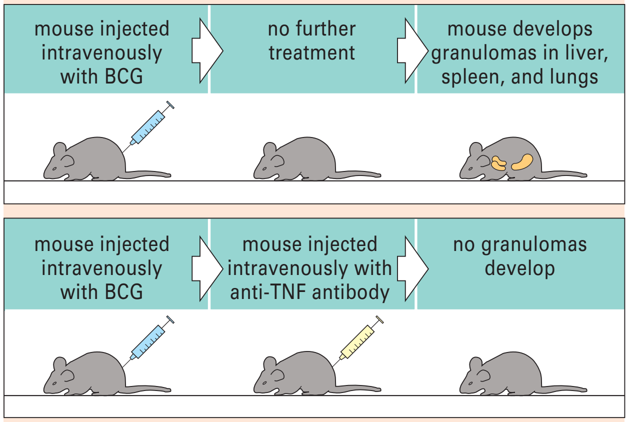
mouse injected intravenously with BCG
no further treatment
mouse develops granulomas in liver, spleen, and lungs
mouse injected intravenously with BCG
mouse injected intravenously with anti-TNF antibody
no granulomas develop

Fig. 26.17 TNF is essential for the development of epithelioid cell granulomas. If BCG-injected mice are injected with anti-TNFα antibodies, they do not develop granulomas.

Epithelioid cells in a granuloma from the lung of a patient with sarcoidosis

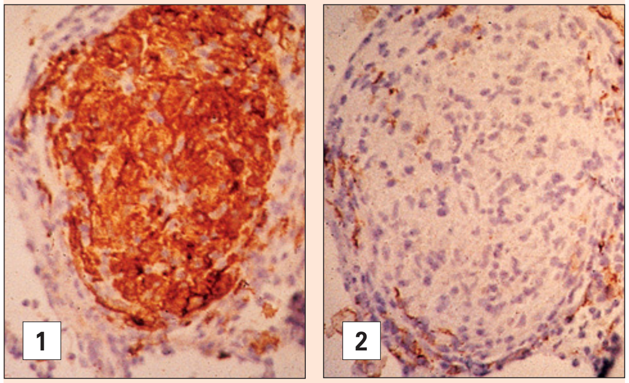

Fig. 26.18 (**1**) The epithelioid cells and giant cells in the center have been stained with the specific antibody RFD-9. (**2**) Mature tissue macrophages surrounding the granuloma are stained with the antibody RFD-7; the exact specificity of the antibody is not known. (Courtesy of CS Munro)

The immunological spectrum of leprosy

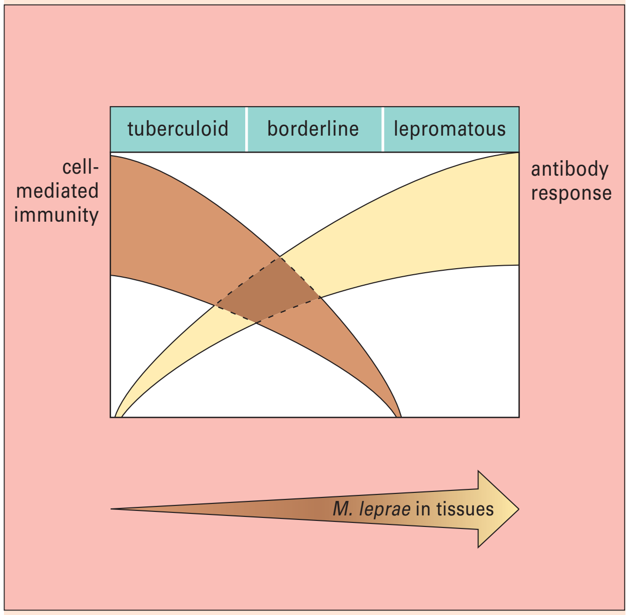

Fig. 26.19 The clinical spectrum of leprosy ranges from tuberculoid disease with few lesions and bacteria to lepromatous leprosy, with multiple lesions and uncontrolled bacterial proliferation. This range reflects host immunity as measured by specific cellular and antibody responses to *M. leprae*, and the tissue expression of cytokines.

A common feature of these infections is that the pathogen presents a persistent, chronic, antigenic stimulus. Activation of macrophages by lymphocytes may limit the infection, but continuing stimulation may lead to tissue damage through the release of macrophage products including reactive oxygen intermediates and hydrolases.

Although delayed hypersensitivity is a measure of T cell activation, the infection is not always controlled, with the result that protective immunity and delayed hypersensitivity do not necessarily coincide. Therefore some subjects showing delayed hypersensitivity may not be protected against disease in the future.

Leprosy is caused by infection with *M. leprae*

Leprosy is a chronic granulomatous disease of skin and nerves caused by infection with *M. leprae*. It is divided clinically into three main types – tuberculoid, borderline, and lepromatous:

- in tuberculoid leprosy, the skin may have a few well-defined hypopigmented patches, which show an intense lymphocytic and epithelioid cell infiltrate and no microorganisms;
- by contrast, the polar reaction of lepromatous leprosy shows multiple confluent skin lesions characterized by numerous bacilli, ‘foamy’ macrophages, and a paucity of lymphocytes;
- borderline leprosy has characteristics of both tuberculoid and lepromatous leprosy (Fig. 26.19).

In leprosy, protective immunity is usually associated with cell-mediated immunity, but this declines across the leprosy spectrum towards the lepromatous pole with a rise in non-protective anti-*M. leprae* antibodies.

The borderline leprosy reaction is a dramatic example of delayed hypersensitivity. Borderline reactions occur either spontaneously or following drug treatment. In these reactions, hypopigmented skin lesions containing *M. leprae* become swollen and inflamed (Fig. 26.20) because the patient is now able to mount a T cell response to the mycobacteria resulting in a delayed-type hypersensitivity reaction. The histological appearance shows a more tuberculoid pattern with an infiltrate of IFNγ-secreting lymphocytes. The process may occur in peripheral nerves, where Schwann cells contain *M. leprae*; this is the most important cause of nerve destruction in this disease. The lesion in borderline leprosy is typical of granulomatous hypersensitivity (see Fig. 26.20).

In patients with a tuberculoid-type reaction, T cell sensitization may be assessed in vitro by lymphocyte proliferation or the release of IFNγ following stimulation with *M. leprae* antigens (Fig. 26.21).

Tuberculosis is caused by *M. tuberculosis*

In tuberculosis, the granuloma provides the microenvironment in which lymphocytes stimulate macrophages to kill the intracellular *M. tuberculosis*. The formation and maintenance of granulomas are essential to control the infection.

In most (>90%) subjects with latent tuberculosis infection, the mycobacteria remain dormant within small granulomas in the lung. There is, however, a balance between the effects of activated macrophages:

- controlling the infection on the one hand; and
- causing tissue damage in infected organs on the other.

In those who progress to clinical tuberculosis, granulomatous reactions lead to cavitation in the lung and spread of bacteria. The reactions are frequently accompanied by

Leprosy

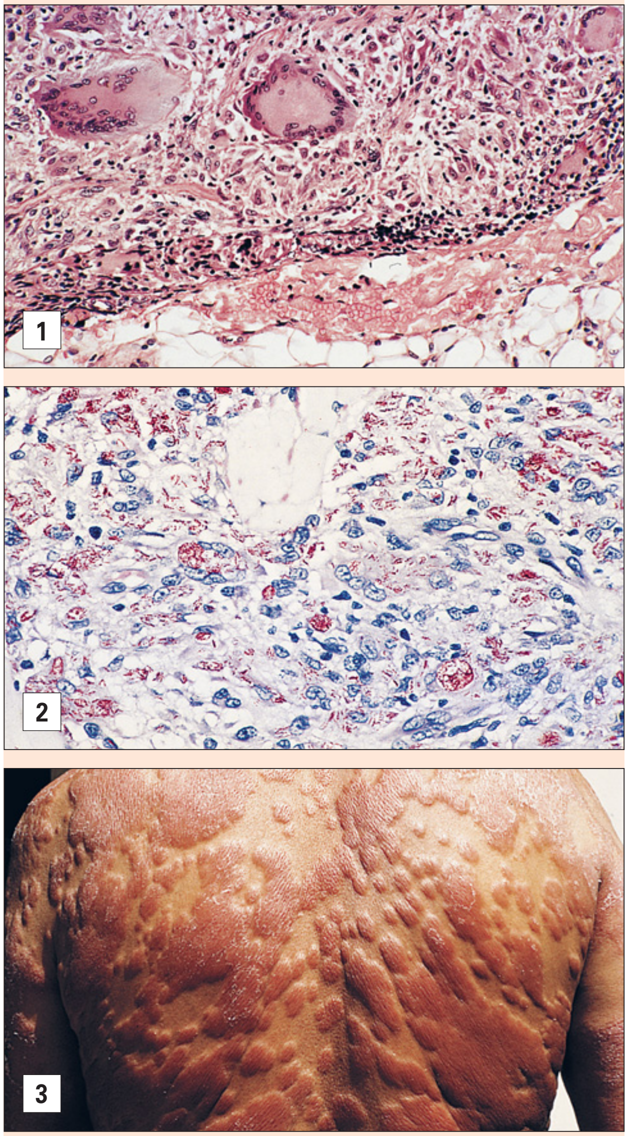

Fig. 26.20 (**1**) A borderline leprosy reaction. This small nerve is almost completely replaced by the granulomatous infiltrate. (**2**) Lepromatous leprosy. Large numbers of bacilli are present. (Courtesy of Dr Phillip McKee) (**3**) Borderline lepromatous leprosy. There are gross infiltrated erythematous plaques with well-defined borders. (Courtesy of Dr S Lucas)

extensive fibrosis and the lesions may be seen in the chest radiographs of affected patients (Fig. 26.22).

Q. What factors might affect the balance that controls a latent infection with tuberculosis?

A. Immunosuppression by drugs or infection (e.g. AIDS) (see Chapter 17) can allow reactivation of infection with tuberculosis.

The histological appearance of the lesion is typical of a granulomatous reaction, with central caseous (cheesy) necrosis (Fig. 26.23). This is surrounded by an area of

Lymphocyte stimulation test in leprosy

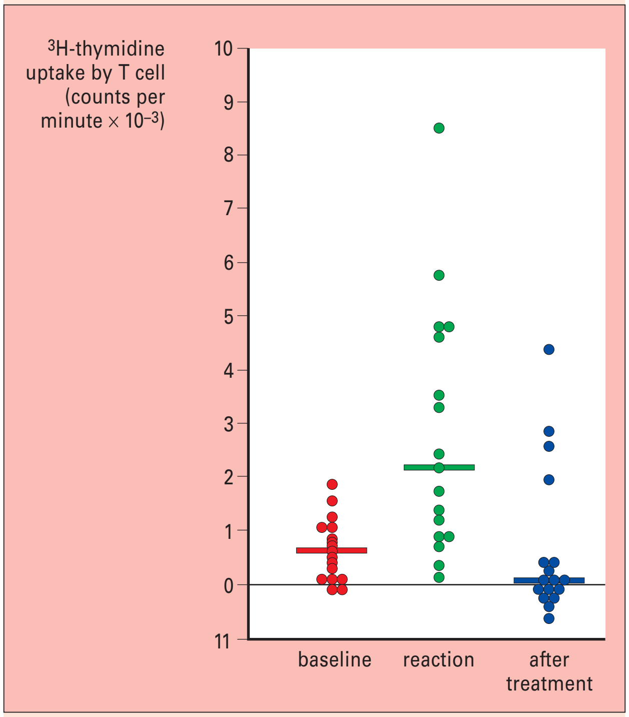

Fig. 26.21 During a borderline leprosy reaction, the lymphocyte stimulation response to *M. leprae* rises. There is a fall in response when the reaction is treated successfully with corticosteroids. The lymphocyte stimulation responses to sonicated *M. leprae* (measured by uptake of ^{3}H-thymidine) are shown for 17 patients who developed such reactions before starting treatment with anti-leprosy drugs (baseline); during the reaction; and following successful treatment with corticosteroids. Medians are indicated by horizontal bars.

Chest radiograph of a patient with pulmonary tuberculosis

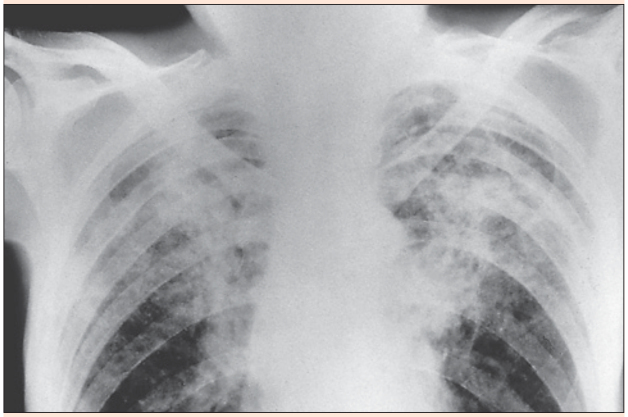

Fig. 26.22 There is extensive parenchymal streaking, predominantly in the upper fields of the lungs. These changes are typical of chronic bilateral pulmonary tuberculosis. Some enlargement of the heart is also evident.

Histological appearance of a tuberculous section of lung

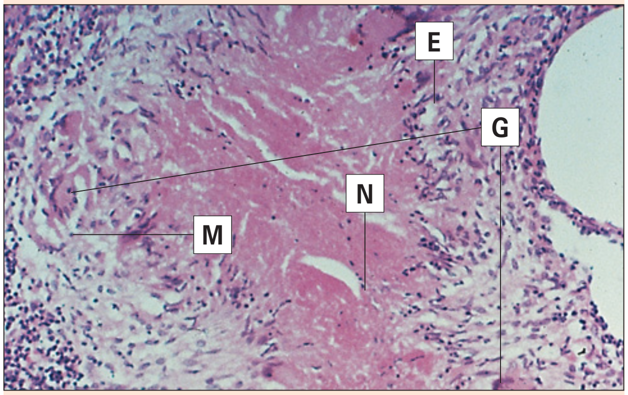

Fig. 26.23 This shows an epithelioid cell granuloma (E) with giant cells (G). Mononuclear cell infiltration can be seen (M). There is also marked caseation and necrosis (N) within the granuloma. H&E stain. × 75.

epithelioid cells, with a few giant cells. Mononuclear cell infiltration occurs around the edge.

Schistosomiasis is caused by parasitic trematode worms

In schistosomiasis, which is caused by parasitic trematode worms (schistosomes), the host becomes sensitized to the ova of the worms, leading to a typical granulomatous reaction in the parasitized tissue mediated essentially by TH2 cells (Fig. 26.24; see also Chapter 15). In this case the cytokines IL-5 and IL-13 are responsible for the recruitment of eosinophils and the formation of the granulomas around the ova. When the eggs have been deposited in the liver, the subsequent IL-13-dependent fibrosis causes hepatic scarring and portal hypertension.

Histological appearance of the liver in schistosomiasis

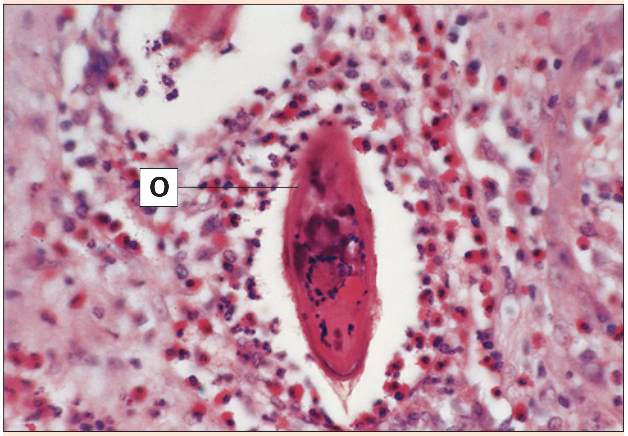

Fig. 26.24 The epithelioid cell granuloma surrounds the schistosome ovum (O) and eosinophils are prominent. H&E stain. × 300. (Courtesy of Dr Phillip McKee)

The cause of sarcoidosis is unknown

Sarcoidosis is a chronic disease of unknown etiology in which activated macrophages and non-caseating granuloma accumulate in many tissues, frequently accompanied by fibrosis (Fig. 26.25). The disease particularly affects lymphoid tissue and the lungs, and enlarged lymph nodes may be detected in chest radiographs of affected patients (Fig. 26.26). No infectious agent has been isolated, though mycobacteria have been implicated because of the similarities in the pathology.

Histological appearance of sarcoidosis in a lymph node biopsy

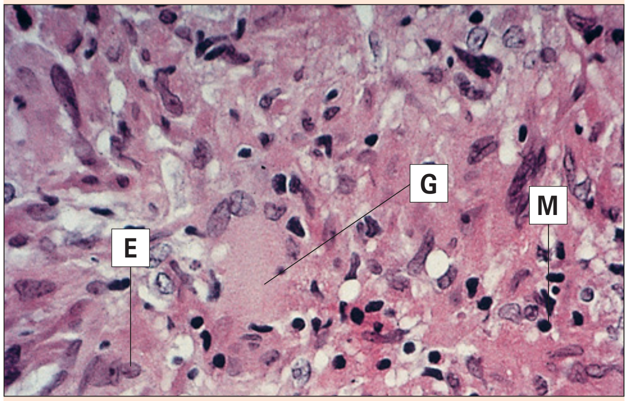

Fig. 26.25 The granuloma of sarcoidosis is typically composed of epithelioid cells (E) and multinucleate giant cells (G), but without caseous necrosis. There is only a sparse mononuclear cell infiltrate (M) evident at the periphery of the granuloma. H&E stain. × 240.

The chest radiograph of a patient with sarcoidosis

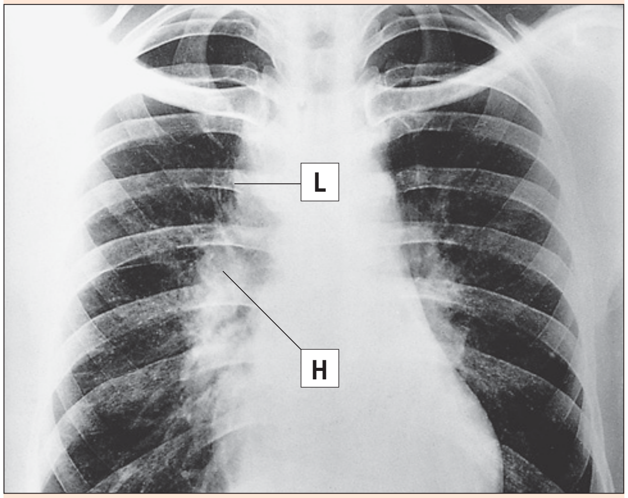

Fig. 26.26 There is enlargement of the lymph nodes adjacent to the hilar (H) and paratracheal (L) areas of the lungs, with diffuse pulmonary infiltration characteristic of the disease.

One of the paradoxes of clinical immunology is that this disease is usually associated with depression of delayed hypersensitivity both in vivo and in vitro. Patients with sarcoidosis are anergic on testing with tuberculin; however, when cortisone is injected with tuberculin antigen the skin tests are positive, suggesting that cortisone-sensitive T inhibitory cells are responsible for the anergy.

Q. What effect does cortisone normally have on an immune response, and why does its effect here appear paradoxical?

A. Cortisone would normally suppress delayed hypersensitivity, principally by its actions on macrophages.

In sarcoidosis, granulomas develop in a variety of organs, most commonly the lungs, lymph nodes, bone, nervous tissue, and skin. Patients may present acutely with fever and malaise, though in the longer term those with pulmonary involvement develop shortness of breath caused by lung fibrosis. The diagnosis is often suggested by the clinical pattern and radiographic changes and confirmed by tissue biopsy. Angiotensin converting enzyme (ACE) and serum calcium levels are sometimes elevated because activated macrophages are a source of both ACE and 1,25-dihydroxy-cholecalciferol (the active metabolite of vitamin D_3).

The cause of Crohn's disease is unknown

Crohn's disease is another non-infectious disease in which granulomas are prominent.

Crohn's disease is a chronic inflammatory disease of the ileum and colon, in which lymphocytes and macrophages accumulate in all layers of the bowel. The granulomatous reaction and fibrosis cause stricture of the bowel and penetrating fistulas into other organs. The natures of the antigens or infectious agents initiating and perpetuating this granulomatous reaction are unknown.

Activated T cells, showing a restricted T cell receptor repertoire and a TH1-like profile of cytokine production are responsible for macrophage activation and the release of inflammatory cytokines, such as TNF, reactive oxygen metabolites, and nitric oxide. These initiate and maintain the transmural intestinal inflammation.

Inhibition of TNF activity with antibody or soluble TNF receptor reduces inflammation in patients with Crohn's disease, but this therapy may be associated with reactivation of tuberculosis in subjects with latent tuberculosis infection and with other granulomatous infectious diseases.

FURTHER READING

American Thoracic Society. Statement on sarcoidosis. Am J Respir Crit Care Med 1999;160:736–755.

Askenase PW. Yes T cells, but three different T cells (αβ, γδ and NK T cells) and also B-1 cells mediate contact sensitivity. Clin Exp Immunol 2001;125:345–350.

Bean AGD, Roach DR, Briscoe H, et al. Structural deficiencies in granuloma formation in tumor necrosis factor gene-targeted mice underlie the heightened susceptibility to aerosol *Mycobacterium tuberculosis* infection which is not compensated for by lymphotoxin. J Immunol 1999;162:3504–3511.

Bjune G, Barnetson RStC, Ridley DS, et al. Lymphocyte transformation test in leprosy; correlation of the response with inflammation of lesions. Clin Exp Immunol 1976;25:85–94.

Britton WJ, Garsia RJ. Mycobacterial infections. In: Bradley J, McCluskey J, eds. Clinical immunology, vol. 38. London: Oxford University Press; 1997:483–498.

Britton WJ, Lockwood DN. Leprosy. Lancet 2004;363:1209–1219.

Casanova J-L, Abel L. Genetic dissection of immunity to mycobacteria: the human model. Annu Rev Immunol 2002;40:581–620.

Daniel H, Present MD, Rutgeerts P, et al. Infliximab for the treatment of fistulas in patients with Crohn's disease. N Engl J Med 1999;18:1398–1405.

Enk AH, Katz SI. Contact hypersensitivity as a model for T-cell activation in skin. J Invest Dermatol 1995;105:805–835.

Flynn JL, Chan J. Immunology of tuberculosis. Annu Rev Immunol 2001;19:93–129.

Flynn JL, Chan J, Triebold KJ, et al. An essential role for interferon-γ in resistance to *Mycobacterium tuberculosis* infection. J Exp Med 1993;178:2249–2254.

Girolomoni G, Sebastiani S, Albanesi C, Cavani A. T cell subpopulations in the development of atopic and contact allergy. Curr Opin Immunol 2001;13:733–737.

Grabbe S, Schwarz, T. Immunoregulatory mechanisms involved in elicitation of allergic contact hypersensitivity. Immunol Today 1998;19:37–43.

Kalish RS, Wood JA, LaPorte A. Processing of urushiol (poison ivy) hapten by both endogenous and exogenous pathways for presentation to T cells in vitro. J Clin Invest 1994;93:2039–2047.

Kindler V, Sappino A-P, Gran GE, et al. The inducing role of tumour necrosis factor in the development of bactericidal granulomas during BCG infection. Cell 1989;56:731–740.

Klimas N. Delayed hypersensitivity skin testing. In: Manual of clinical laboratory immunology, 5th edn. Washington: ASM Press; 1997:276–280.

Roach DR, Briscoe H, Saunders B, et al. Secreted lymphotoxin-alpha is essential for the control of intracellular bacterial infection. J Exp Med 2001;193:239–246.

Roach DR, Bean AGD, Demangel C, et al. Tumor necrosis factor regulates chemokine induction essential for cell recruitment, granuloma formation and clearance of mycobacterial infection. J Immunol 2002;168:4620–4628.

Romagnani P, Annunziato F, Baccari M, et al. T cells and cytokines in Crohn's disease. Curr Opin Immunol 1997;9:793–799.

Salgame P. Host innate and TH1 responses and the bacterial factors that contain *Mycobacterium tuberculosis* infection. Curr Opin Immunol 2005;17:374–380.

Trinchieri G. Interleukin-12: a cytokine at the interface of inflammation and immunity. Adv Immunol 1998;70:133–195.

Von Andrian UH, Mackay CR. T cell function and migration: two sides of the same coin. N Engl J Med 2000;343:1020–1034.

Wallis RS, Broder MS, Wong JY, et al. Granulomatous infectious disease associated with tumor necrosis factor. Clin Infect Dis 2004;38:1261–1265.

Wang B, Eshe C, Mamelak A, et al. Cytokine knockouts in contact hypersensitivity research. Cytokine Growth Factor Rev 2003;14:381–389.

Wynn TA, Thompson RW, Cheever AW, Mentink-Kane MM. Immunopathogenesis of schistosomiasis. Immunol Rev 2004;201:156–167

Yamamura M, Uyemura K, Deans RJ, et al. Defining protective immune responses to pathogens: cytokine profiles in leprosy lesions. Science 1991;254:277–279.

Websites

http://users.path.ox.ac.uk/~cholt/ – a home page describing macrophage biology and the role of macrophages in the host response to infectious disease

http://www.who.int/lep/disease/disease.htm – a home page describing leprosy infection

Critical thinking: A hypersensitivity type IV reaction (see p. 503 for explanations)

An 8-year-old boy with recent weight loss and mild fever is found to have an enlarged lymph node on the right side of the neck. He has no cough and his chest radiograph is normal. Surgical biopsy of the lymph node reveals a granulomatous infiltrate with no evident acid-fast bacilli. The result of microbiological culture for *Mycobacterium tuberculosis* is awaited. Intradermal skin testing with tuberculin causes swelling and erythema of 20 mm in diameter after 48 hours.

1 What cell types make up the granulomas in the lymph node and what cytokines are involved in their formation?

2 What is the pathology at the site of the skin testing and how does it differ from that in the lymph node?

3 What type of lymphocyte is responsible for the skin test reactivity?

4 What other conditions cause granulomas in lymph nodes and how are they diagnosed?

5 When the family members are tested, the boy's 5-year-old brother is found to have a positive tuberculin reaction (18 mm at 48 hours), but he is well with a normal chest radiograph. What does this result indicate about his immune responses and what is its significance?

This Page Intentionally Left Blank

Critical Thinking: Explanations

1. SPECIFICITY AND MEMORY IN VACCINATION

1.1 The immunological ‘memory’ induced by vaccination does not depend just on the antibodies. Memory is due to long-lived memory lymphocytes, which persist in the lymphoid tissues for many years. They will be reactivated if the individual encounters the toxin or the vaccine on a later occasion.

1.2 The tetanus toxoid is a stable molecule – it does not change or mutate, so antibodies and lymphocytes that recognize it continue to be effective. By contrast, influenza A mutates every year. Last year’s antibodies are marginally effective or ineffective against this year’s virus. Researchers must identify newly emerging virus strains and prepare vaccine from those strains they think will produce new epidemics. Often they get it right, but not always.

1.3 Recommendations are based on practicality. It is impossible to prepare sufficient vaccine each year to immunize everyone against influenza. There is not enough time to do it and not enough laboratory resources available. So the highest risk groups are targeted – health workers because they will likely be in contact with the disease and old people because the disease can lead to serious complications.

Discussion point

If we could immunize every person in the world against influenza A in 1 year, do you think that this would lead to total eradication of the disease?

2. DEVELOPMENT OF THE IMMUNE SYSTEM

2.1 The total numbers of blood lymphocytes are drastically reduced, with T cells being virtually absent and B cells significantly reduced – B cells require T cells to complete their own development. The lymph nodes are much reduced in size, and this particularly affects the paracortex (T cell areas). Compare this with diGeorge syndrome. The animals have a reduced ability to fight infections, but this is selective, affecting particularly some viruses and parasites – possibly because there is still good NK cell activity and macrophage-mediated antibacterial defenses.

2.2 Adult thymectomy has very little effect on the individual’s ability to fight infection. By adulthood, there is a large pool of peripheral T cells that may to some extent self renew. The thymus progressively involutes and becomes less important as a site of T cell development in the adult.

2.3 Because the lymphocyte precursors fail to make productive rearrangements of their antigen receptor genes, they die by apoptosis during development. This leads to a profound immune deficiency of all lymphocytes, which is analogous to severe combined immunodeficiency (SCID) in humans.

2.4 Interleukin-7 is required for lymphocyte development in primary lymphoid organs. There is a profound reduction in thymocytes and peripheral lymphocytes and a total absence of $\gamma\delta$ T cells.

2.5 The $\alpha_4\beta_7$-integrin is required for binding of cells to adhesion molecules on the high endothelial venule (HEV) of gut-associated lymphoid tissue (GALT), so this knockout results in drastically reduced lymphocyte numbers in these tissues.

3. THE SPECIFICITY OF ANTIBODIES

3.1 In the presence of the antibodies, mutated variants of the virus are selected that do not bind those antibodies. By detecting which of the virus proteins are mutated, one can infer that these are the proteins that normally would bind to the antibody. Neutralizing antibodies against viruses are generally directed against proteins in the capsid of the virus, particularly against the proteins that the virus uses to attach to the surface of its target cell. Antibodies cannot gain access to the inside of the virus, so neutralizing antibodies do not bind the core protein VP4.

3.2 The antibody VP1-a binds to an epitope that includes two closely spaced residues (91 and 95). This is a ‘continuous epitope’ and is located on a single external loop of polypeptide. By contrast, the epitope recognized by VP1-b is located in at least two distinct areas of the polypeptide chain (83–85 and 138–139). This is a ‘discontinuous epitope’: examination of the VP1 antigen shows that these residues are located on two adjacent areas of β-pleated sheet.

3.3 A mutation of residue 138 does not affect the epitope recognized by antibody VP1-a, so it continues to bind with high affinity to the antigen. This confirms that the epitopes recognized by VP1-a and VP1-b are physically separate. The mutant with Gly at position 95 still binds the VP1-a antibody weakly. Glycine is a smaller amino acid than aspartate, which is present in the wild type, hence the antibody can still bind to the epitope, although the ‘fit’ is less good, so the affinity of binding is lower. By contrast, lysine (Lys) is a larger residue than aspartate. It protrudes further out into the antibody’s binding site and completely disrupts the antigen–antibody bond.

4. COMPLEMENT DEFICIENCY

4.1 Deficiencies of components of the classical or alternative pathways, particularly of C3, produce a reduced ability to opsonize bacteria, resulting in impaired phagocytosis by macrophages and neutrophils. Patients suffer from repeated bacterial infections from Gram-positive bacteria (e.g. staphylococci, streptococci). These children are unable to clear bacterial infections because their phagocytes do not take up bacteria efficiently. Deficiencies in the lytic pathway components (C5–C9) can render patients more susceptible to neisserial infections because the lytic pathway can damage the outer membrane of Gram-negative bacteria such as *Neisseria* spp.

4.2 There is a clear deficiency in C3 and components of the alternative pathway. Components of the classical pathway are on the lower end of normal. At first this looks surprising, because the initial assay for lytic complement required the activity of the classical and lytic pathways. Nevertheless, both the bacterial infections and the lack of total hemolytic complement can be explained by the very low levels of C3. Note that the genes for C3, fB, and fI are not genetically linked, so we cannot explain this apparent multiple deficiency of alternative pathway components by some multiple gene deletion. The explanation lies in the alternative pathway amplification loop. Because the children lack fI, they cannot break down the alternative pathway C3 convertase C3bBb. Therefore C3 is continuously activated and binds fB. All the fB is consumed, as is most of the free C3. The genetic deficiency of fI therefore leads to secondary deficiencies in the components of the alternative pathway and this then affects C3 and the function of the classical and lytic pathways.

4.3 The children have a homozygous fI deficiency – both copies of the gene are missing. Replacing fI, either by an infusion of normal serum or by providing pure fI, restores all the other components to normal levels and allows the children to clear bacterial infections. Antibiotic prophylaxis will help prevent bacterial infections.

Discussion point

What problem might occur if you inject a protein such as fI into an individual who lacks it due to a genetic deficiency?

5. SOMATIC HYPERMUTATION

5.1 There are at least three theories why TCRs do not undergo somatic hypermutation.

1. T cells control most antibody responses – they therefore represent an important mechanism for maintaining self tolerance. According to this view, the immune system can afford to let immunoglobulin genes undergo somatic hypermutation because (non-mutating) T cells retain control.
2. TCRs must retain the ability to recognize self (MHC), and therefore cannot be allowed to undergo somatic hypermutation.
3. The main purpose of somatic hypermutation of antibodies is to give a more robust secondary response.

In the case of T cells, as opposed to antibodies, it is possible to increase the efficacy/avidity of the memory T cell by affecting TCR density, presence of co-stimulatory/inhibitory molecules, or other aspects of the wiring of T cell signaling. Mutating the receptor is therefore simply not necessary because it is potentially dangerous as well. According to this theory, it is best avoided.

5. THE SPECIFICITY OF T CELLS

5.2 This is an example of genetic restriction in antigen presentation. The SM/J T cells are primed with antigen on MHC molecules of the SM/J haplotype and will only respond to this combination of antigen–MHC. They do not recognize the same antigen presented by other MHC molecules. Because MHC molecules are co-dominantly expressed, the H-2v MHC molecules are present on the APCs from the F1 animal and so they too stimulate the T cells.

5.3 The minimum peptide needed to activate the T cells appears to be 80–94, which is 15 residues long and therefore corresponds well to the expected size of antigen peptides that can fit into the MHC class II binding site. This peptide is included within peptide 80–102, which also stimulates strongly. Peptides 84–98 and 73–88 lack the N and C terminals of the antigenic peptide, respectively, and therefore lack some of the anchor residues needed to hold them in the MHC peptide-binding groove.

5.4 This is called a superagonist or a strong agonist peptide. Typically such a peptide will have a stronger binding affinity for the MHC molecule and/or the TCR.

6. THE ROLE OF ADHESION MOLECULES IN T CELL MIGRATION

6.1 IL-1 induces the expression of a number of adhesion molecules, including ICAM-1 and VCAM-1, both of which can potentially mediate leukocyte migration by their interaction with the integrins LFA-1 and VLA-4, respectively.

6.2 Because it takes several hours to increase migration, and ICAM-1 and VCAM-1 appear to be involved in the process, one can infer that their expression is increased as a result of protein synthesis (which takes several hours) rather than by a relatively rapid release from intracellular stores.

6.3 Antibodies to ICAM-1/LFA-1 reduce migration of cells across unstimulated endothelium. Therefore, this pair of adhesion molecules is required for migration across resting endothelium.

6.4 Antibodies to both ICAM-1/LFA-1 and VCAM-1/VLA-4 reduce migration across IL-1-activated cells, therefore both pairs of adhesion molecules control this event. In practice it is known that ICAM-1 is present on unstimulated brain endothelium and is increased by IL-1, whereas VCAM-1 is virtually absent from unstimulated brain endothelium, but may be synthesized following stimulation with inflammatory cytokines.

7. ANTIGEN PROCESSING AND PRESENTATION

7.1 Macrophages express both MHC class I and class II molecules, and can therefore present antigen to either of the clones. Fibroblasts do not generally express MHC class II molecules and one would not expect them to stimulate the MHC class II-restricted clone.

7.2 Live flu virus infects the macrophages and flu virus polypeptides are synthesized in the cytoplasm of the cell, so the viral antigens are presented by the internal (MHC class I) pathway as well as the external (MHC class II) pathway. Inactivated virus is taken up by the macrophage, processed and presented via the class II pathway only – because there is no viral protein synthesis there is no presentation via the MHC class I pathway.

7.3 Emetine blocks protein synthesis, so no protein fragments are fed into the MHC class I pathway by the proteasomes. Chloroquine prevents phagosome/lysosome fusion so endocytosed virus cannot be broken down into peptides. Consequently no peptides are available for the MHC class II pathway.

7.4 The MHC class I-restricted T cells express CD8 and the MHC class II-restricted cells express CD4 because CD8 and CD4 are co-receptors for MHC class I and class II molecules, respectively.

8. DEVELOPMENT OF THE ANTIBODY RESPONSE

8.1 In a developing immune response to a TD antigen, B cells will switch from IgM production to IgG. Because the antigen is continuously present as a depot, by day 14 the response has the characteristics of a secondary response – IgG antibody titers are climbing rapidly.

8.2 Perhaps the two mice have already been infected by mouse hepatitis virus. By day 5 they are already making a secondary IgG response. This could be a problem in the colony, though usually all animals housed together would become infected. If these mice have been naturally infected it would be through the gut (unlike the vaccine) and one would therefore expect a stronger IgA response.

8.3 IgA-producing clones tend to be located in the mucosa-associated lymphoid tissues (see Chapter 2) and it is not surprising that no IgA-producing clones were generated from the spleen.

8.4 IgG-producing clones at day 14 are likely to be of higher affinity than IgM producers.

9. THE ROLE OF MACROPHAGES IN TOXIC SHOCK SYNDROME

9.1 TNFα, IL-1, IL-6, IL-10.

9.2 Lack of activity, ruffling of fur, respiratory distress, possibly leading to death within 24 hours.

9.3 BCG activates macrophages via infection of APCs and induction of IFNγ by NK and $CD4^+$ T cells, which primes macrophages. LPS delivers stimulus via LPS binding protein, CD4, Toll-like receptors, and NFκB activation, to enhance proinflammatory cytokine release. TNFα and IL-1, especially, act locally and systemically on vascular endothelium, neutrophils, and central nervous centers, causing hypotension and circulatory collapse.

9.4 CD14 knockout mice are extremely resistant to septic shock. Scavenger receptor A knockout mice are more susceptible to septic shock. IFNγ knockout mice are relatively resistant to septic shock.

9.5 CD14 is central to the LPS recognition and signaling pathway. SR-A clears LPS from the circulation to protect the host. IFNγ is needed to prime macrophages.

9.6 Evaluate the kinetics of pro- and anti-inflammatory cytokine production to establish the endogenous regulation of macrophage activation. Use blocking antibodies for TNFα and other cytokines, and receptor knockout mice to establish the roles of each. Evaluate cytokine production by peritoneal macrophages taken from BCG-primed mice after LPS challenge in vitro.

9.7 Septic shock is a major complication of Gram-negative (e.g. *Neisseria meningitidis*) infection. Therapeutic approaches include circulatory support, antibiotics, and possibly combinations of cytokine and receptor antagonists (blocking antibodies, inhibitors of TNFα cleavage, soluble receptors).

Relevant references

Haworth R, Platt N, Keshav S, et al. The macrophage scavenger receptor type A (SR-A) is expressed by activated macrophages and protects the host against lethal endotoxic shock. J Exp Med 1997;186:1431–1439.

Haziot A, Ferrero E, Kontgen F, et al. Resistance to endotoxin shock and reduced dissemination of Gram-negative bacteria in CD14 deficient mice. Immunity 1996;4:407–414.

9. THE DISTINCTION BETWEEN INFLAMMATORY RESPONSES AND HOUSE-KEEPING OF APOPTOTIC CELLS

9.8 Use blocking antibodies for phagocytic receptors (e.g. vitronectin receptors) or cells from knockout mice, if available.

9.9 Ligation and cross-linking of phagocytic receptors by apoptotic cells induce signaling pathways resulting in suppression of inflammatory and antimicrobial responses.

9.10 Use antibodies and antagonists to receptors to study candidate inhibitory responses such as production of prostaglandin E_2 and TGFβ.

9.11 Pathogens can exploit and induce the downregulation of inflammation by apoptotic cells to evade killing by host cells. This may be counteracted by the use of drugs to prevent inhibitory pathways, even in vivo.

Relevant references

Stein M, Keshav S, Harris N, Gordon S. IL-4 potently enhances murine macrophage MR activity; a marker of alternative immunologic macrophage activation. J Exp Med 1992;176:287–292.

Freire-de-Lima CG, Nascimento DO, Soares MBP, et al. Nature 2000;403:199–203.

9. THE ROLE OF MACROPHAGES IN TH1 AND TH2 RESPONSES

9.12 Macrophage activation involves a complex pattern of altered gene expression, covering a spectrum of activities and not just polar opposites between activation (TH1, IFNγ) and deactivation (TH2, IL-10). IL-4 and IL-13, TH2 cytokines, use common receptor chains to induce an alternative pathway of macrophage activation involved in humoral immunity and possibly repair (enhanced APC function via MHC class II expression and MRs, as well as other effects on B cell production of antibody). IFNγ and IL-10 regulate cellular immune effector functions.

9.13 Broaden the range of macrophage markers examined, ultimately by DNA gene chip analysis, and look for consistency and reproducibility of similar patterns of altered gene expression by the cytokines above. Analyze macrophage functions in mice with knockouts of cytokines or their receptors.

9.14 Find model antigens (e.g. parasites) that induce TH2 responses in vivo, and establish whether these are recognized by APC receptors that enhance IL-4/IL-13 or inhibit IFNγ production by appropriate cells.

10. MECHANISMS OF CYTOTOXICITY

10.1 CTLs kill their targets using granule-associated mechanisms such as perforin, and by activating pathways of apoptosis – hence they produce both cell lysis and DNA fragmentation. CTLs recognize allogeneic MHC class I molecules on the targets – the effectors and targets come from different individuals. Note that there is always a low level of cell damage in the controls that contain no effector cells.

10.2 Tumor cells typically show reduced expression of MHC molecules by comparison with normal cells, and as they do they become susceptible to damage mediated by NK cells. Because NK cells recognize several different types of MHC molecule, the fact that the targets are allogeneic is not so relevant. Antibody can cross-link the NK cells to their targets (K cell activity) via their Fc receptor (CD16). This moves the balance of activity towards activation of the NK cell. Note that antibody itself does not damage the targets in this type of assay because the culture medium does not contain functional complement.

10.3 Purified perforin is extremely efficient at causing cell lyis, but does not activate the pathways of apoptosis, so there is no DNA fragmentation.

11. REGULATION OF THE IMMUNE RESPONSE

11.1 Overexcessive immune activation is harmful to the host. An incorrect or excessive immune response can lead to autoimmunity or allergy. Anergy, cell death, and suppression by cytokines and regulatory cells are three ways of inducing tolerance and preventing this.

11.2 Antibody regulates the immune system through a number of feedback mechanisms. Antibodies can form immune complexes with antigen and these can lead to expansion or suppression of immune responses. Passive IgM can enhance an immune response, whereas passive IgG suppresses the response through antibody blocking or receptor cross-linking. In addition, some immunoglobulins are idiotypic and have an immunogenic sequence within them. During an immune response antibodies can be generated against this idiotypic sequence leading to downregulation of the response.

11.3 Transfer of regulatory T cells into mice with autoimmunity can cause a reduction in the disease, but it is not known how this would affect other immune responses if the system was used in the treatment of human autoimmunity. It is possible that there could be a decreased immune response to viral and bacterial infections leading to an increase in these diseases.

11.4 If lymphocytes did not senesce they might begin to proliferate in an uncontrolled manner – this could cause cancers of the immune system and also lead to exaggerated immune responses causing autoimmunity or allergy.

12. IMMUNE REACTIONS IN THE GUT

Oysters are a food item that is normally tolerated by the body. Eating an infected shellfish causes the antigens in the food to be presented in association with components of the pathogen that can activate PAMP receptors, induce co-stimulatory molecules, and break tolerance. Additionally bacterial toxins may damage the gut epithelium again, enhancing antigen presentation and allowing antigens to access the gut-associated lymphoid tissues more readily. Once tolerance is broken the TH2-type response, which is characteristic of the gut, will lead to production of antigen-specific IgE antibodies. Consequently, eating another oyster, even a good one, will lead to an allergic reaction to antigen in the food.

13. VIRUS–IMMUNE SYSTEM INTERACTIONS

13.1 Features of a virus that enable it to evade host defense mechanisms include the following.

- Antigenic variation, as seen for example in HIV and influenza virus. Evasion of antibody and T cell recognition by mutation arising in genes encoding major antigenic targets.
- Production of ‘decoy’ molecules that interfere with host immune defenses, in particular molecules targeting the interferon system, downregulating MHC class I expression, subverting the function of cytokines and chemokines, and the action of complement.
- Establishment of virus latency, thereby avoiding detection by the immune system (e.g. HSV-1 in sensory ganglia, EBV in B lymphocytes).

13.2 Non-neutralizing antibodies could function by mediating antibody-dependent complement-mediated lysis of infected cells, complement-mediated neutralization, antibody-dependent cellular cytotoxicity (ADCC), or antibody-enhanced phagocytosis by macrophages.

13.3 Experiments to support the conclusions could involve:

- examination in vitro of whether complement mediates neutralization of virus in the presence of non-neutralizing antibodies;
- use of mice deficient in late complement components (i.e. C5-deficient mice);
- analysis of whether IgG subclasses were able to mediate ADCC in vitro using virus-infected cells.

14. IMMUNOENDOCRINE INTERACTIONS IN THE RESPONSE TO INFECTION

14.1 The immune system cannot be understood in isolation from the rest of mammalian physiology. One of the many effects of stress is increased production of adrenocorticotropic hormone (ACTH) from the pituitary. This in turn drives increased production of cortisol by the adrenal. The proof is that, in the animal models mentioned, the stressor can be replaced by mimicking the stress-induced levels of cortisol (or the rodent equivalent, corticosterone) with implanted slow-release cortisol pellets. The cortisol downregulates cell-mediated immunity to tuberculosis.

14.2 These observations provide clues as to why increased cortisol levels can lead to reduced immunity to tuberculosis. Raised cortisol levels cause APCs to release more IL-10 and less IL-12, so newly recruited T cells tend to develop a TH2 cytokine profile. Moreover, cortisol actually synergizes with some functions of TH2 cytokines, and enhances the ability of IL-4 to drive IgE production. It is interesting that BCG vaccination does not lead to protective immunity if the BCG is given to animals bearing cortisol pellets that mimic stress levels of cortisol. Cortisol also reduces the antimycobacterial functions of macrophages.

These points emphasize the need for a physiological approach to the understanding of infection. A narrowly immunological approach has solved some infections, but global emergencies such as tuberculosis, HIV infection and septic shock may require integrated physiological thinking as well as pure immunology.

15. IMMUNITY TO PROTOZOA AND HELMINTHS

15.1 Protozoa replicate within the host, so there is usually a balance between the effectiveness of the immune response and the virulence of the parasite. With certain parasites the infections may be short-lived and may kill the host, but this may not be a disadvantage to the parasite if it has already been transmitted to a new host. A good example is falciparum malaria, which is potentially fatal, particularly in children in endemic areas who have not developed any immunity, but are likely to have been bitten during the course of the infection by mosquitos, which will ensure further transmission. Helminths, by contrast, do not replicate within the host and are generally long-lived chronic infections. Transmission is by the release of eggs and larvae from an adult parasite, which may be excreted or be taken up by a vector.

15.2 By adopting an intracellular mode of existence parasites may be able to 'hide' from the immune response. A good example is falciparum malaria, which lives in mature red blood cells. Because this cell type has no nucleus it cannot express MHC class I molecules on its surface, so the parasite is invisible to $CD8^+$ cytotoxic T cells. Other parasites live in nucleated cells, which will express class I MHC molecules, but experiments have shown this to be downregulated in cells infected with some parasites. *T. gondii* avoids being killed by the macrophage by inhibiting the fusion of the lysosome with the phagosome; *T. cruzi* escapes from the phagosome into the cytoplasm of the cell; and *Leishmania* spp. can resist the low pH of the phagolysosomes and are resistant to lysosomal enzymes.

15.3 Extracellular parasites can adopt a number of ways of avoiding immune attack. Parasites may 'disguise' themselves, for example by undergoing antigenic variation (African trypanosomes) or by adsorbing host molecules or undergoing molecular mimicry of the host (schistosomes). Parasites may 'hide' from the host immune response by becoming cysts (*Entameba* spp.) or by living in an immunoprivileged location (*Toxoplasma* in brain). They may 'resist' attack by having a physical barrier (helminths) or by producing enzymes that resist the oxidative burst or disable antibodies. Many parasites are able to 'modulate' the host immune response to their advantage.

16. HYPER-IgM IMMUNODEFICIENCY

16.1 Normal infants do not become infected with *Pneumocystis carinii*. The occurrence of this type of pneumonia suggests the presence of an immunodeficiency disorder. Her immunoglobulin levels point to an elevated IgM level whereas the IgG, IgA, and IgE levels are very low. These findings almost certainly rule out the diagnosis of X-linked agammaglobulinemia because in that case the IgM would be undetectable. The normal white blood cell count probably rules out severe combined immunodeficiency (SCID) because an infant with SCID would have a very low lymphocyte count ($< 3000/mm^3$). Had the IgD level been measured it would probably also be high, as is found in males with hyper-IgM immunodeficiency.

16.2 A blood sample should be obtained to isolate the B lymphocytes. In hyper-IgM immunodeficiency the B cells would stain only with fluorescent anti-IgM and anti-IgD. In a normal infant B cells that stain with anti-IgA and anti-IgG would also be present in addition to IgM^+ and IgD^+ cells. Furthermore, the mononuclear cells from the blood sample should be stimulated with phytohemagglutinin. After several hours of stimulation, the cells should be stained with fluorescent anti-CD40 ligand. Some 50% of normal T cells would stain positively whereas none would be positive in hyper-IgM immunodeficiency.

16.3 Antibodies formed to tetanus toxoid immunization are of the IgG class. Because this child is incapable of undergoing isotype switching she cannot make IgG antibodies. However, antibodies to the blood group substances are predominantly of the IgM class, which this child can synthesize. Furthermore the isohemagglutinins are so called T-independent antibodies; their formation does not require T cell help. Isotype switching to IgG, IgA, or IgE requires that the CD40 ligand on activated T cells engages CD40 on B cells. This is a critical first step in providing T cell help to antibody formation by B cells. This infant with a genetic defect in the CD40 ligand cannot effect isotype switching.

16.4 The child should be given intravenous gamma-globulin at regular monthly intervals to protect her against pyogenic infections. Intravenous gamma-globulin is virtually pure IgG. As this child is incapable of making IgG she will need lifelong infusions for passive protection. Intravenous gamma-globulin does not protect against intracellular microorganisms such as *P. carinii*. These and other microorganisms that reside in macrophages are eliminated by activating macrophages. Macrophages, like B cells, express CD40 and require interaction with the CD40 ligand on activated T cells to become highly microbicidal. This child therefore remains susceptible to recurrent *P. carinii* infection despite the gamma-globulin therapy. If it recurs she would need treatment again with pentamidine.

16.5 A normal result. The mother would exhibit random inactivation of her X chromosomes. In X-linked agammaglobulinemia non-random inactivation of B cells is observed in females who carry this X-linked disorder because the B cells that have the mutated *btk* gene cannot expand clonally; only the B cells carrying the normal allele can. The same is true in X-linked severe combined immunodeficiency (SCID). The T cells of heterozygous females which have a mutation in the common gamma chain cannot expand clonally; only the T cells with the normal allele can. The CD40 ligand is not required for clonal expansion of T cells. This molecule is only expressed when T cells are activated. Thus females who must be heterozygous for hyper-IgM immunodeficiency are immunologically normal and the clonal expansion of T cells bearing the wild-type allele and the mutated allele is normal.

16.6 It depends on the severity of the mutation in the child. If she has a missense mutation and expresses some mutated CD40 ligand on her activated T cell the prognosis is good if she continues to receive intravenous gamma-globulin. However, if her mutation results in an inability to express any CD40 ligand, her future is uncertain because these affected females usually develop an extensive polyclonal expansion of IgM-producing B cells that invade the liver and other parts of the gastrointestinal tract and cause fatal complications (see Fig. 16.4). In this case a bone marrow transplant from a histoidentical donor should be recommended as this would have a favorable outcome and would essentially be a cure of the disease.

17. SECONDARY IMMUNODEFICIENCY

17.1 Approximately 95% of HIV-positive individuals seroconvert within 3 months of infection. ELISAs for antibodies to gp41, an HIV surface glycoprotein, and p24, a core protein, are the most widely used to detect HIV infection. Confirmation is obtained by Western blot analysis to decrease the rate of false-positive results. The Centres for Disease Control and Prevention recommends that the blot should be positive for two of the p24, gp41, and gp120/160 markers (gp160 is the precursor form of gp41 and gp120, the envelope protein). ELISAs for p24 antigen can also be used, though the false-negative rate is higher. The PCR is a technique for amplifying specific sequences of DNA or RNA to produce quantities that are readily detectable. The test in the context of HIV is highly sensitive and specific, but is more costly than ELISA techniques.

17.2 The mother's serological state should be tested by ELISA and confirmed by Western blot if positive. Around 20–30% of infants born to HIV-positive mothers are infected with the virus. Transmission can occur in utero or very rarely by breastfeeding. Diagnosis presents a problem because maternal IgG specific for HIV antigens crosses the placenta and can be detected in the infant even if the infant has not become infected. The presence of HIV-specific antibodies of IgA and IgM classes in the infant should imply infection because they do not cross the placenta. Current tests lack sensitivity and remain in development. The method most widely used in the UK and USA is the PCR, which demonstrates the virus directly. Below the age of 1 month the PCR may be negative in infected children. It has been shown that, in many children, HIV is sequestered into regional lymph nodes at this age. After establishing infection at these sites a viremia follows.

17.3 The figure shows the change in a variety of indices of HIV infection over time. Acute seroconversion causes an infectious mononucleosis-like illness in up to 50% of those infected with HIV. Common symptoms are fever, lymphadenopathy, pharyngitis, rashes, and myalgia. At this point there is a drop in the CD4 (and also CD8) lymphocyte count and a rise in plasma viremia and p24 antigen concentration. Antibodies to HIV surface glycoproteins gp120 and gp41 are produced from approximately 6 weeks after infection and are initially of the IgM class. IgG antibodies of the same specificities follow the IgM response and persist during the latent phase. Viremia

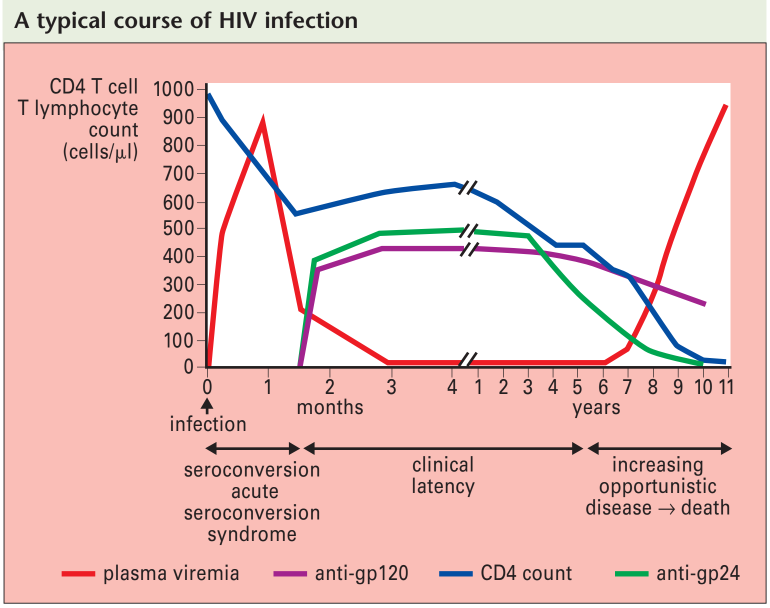
A typical course of HIV infection

and p24 antigenemia are generally low during this period. Disease progression is heralded by a declining CD4 lymphocyte count and a rise in plasma viremia. Clinically, CD4 counts have become a widely used index of progression. Plasma viremia is the most accurate measure of disease progression, and is becoming a more commonly used method.

18. VACCINATION

18.1 Successful attenuation results in an organism still capable of generating an immune response against the wild-type virulent organism, but no longer capable of causing disease. This is a delicate balance to achieve. In some instances (e.g. hepatitis B and C viruses) the organisms cannot be cultured, so attenuation by repeated in-vitro passage is impossible.

18.2 There is no reason to believe that a vaccine cannot improve on nature. Many organisms express gene products that interfere with immune responses. Removal of these from the vaccine may allow protective responses to be generated.

18.3 Although smallpox was eliminated because there was no animal reservoir or carrier state, this is also the case for some other microorganisms and it should be possible to eliminate them. For other organisms, elimination will be difficult and maintenance of herd immunity will remain important.

18.4 Vaccines are unlikely ever to replace antibiotics completely. Microorganisms evolve very rapidly and vaccine production against complex organisms, such as mycobacteria, has proved very difficult. In addition, immunodeficient individuals and the elderly remain at risk even after vaccination.

18.5 BCG clearly has multiple effects. Specific immune responses to BCG antigens can be detected, but cell wall and other components of the bacterium have potent immunomodulatory effects. The efficacy of BCG as a vaccine has been suggested to depend on cross-reactions of BCG with environmental mycobacteria as well as *Mycobacterium tuberculosis* itself.

18.6 Strong immune responses may cause tissue damage in individuals with parasites already present. However, a vaccine that prevented establishment of infection would be unlikely to be damaging.

18.7 Antigens that induce T cell immunity may be useful vaccine antigens. Molecules that contribute to virulence, such as toxins, may be the best targets for vaccines. Because the genomes of many pathogens have been or are being sequenced, searching homologous gene products may identify potential targets.

19. TOLERANCE

Lineage commitment in thymocyte development

19.1 The presence of a constitutively active Lck molecule in developing thymocytes. This is revealed in the OT-1/dLGF experimental mouse.

19.2 The presence of a catalytically inactive Lck causes a massive reduction in the number of CD4 SP cells that are selected in the AND mouse. There is a limited increase in the number of CD8 SP cells.

19.3 Yes, the results are consistent. Overexpression of Lck promotes CD4 cells and reduced expression of CD8 cells, regardless of the MHC restriction of the majority population of TCRs.

The molecular basis of activation-induced cell death (AICD)

19.4 AICD is mediated by signaling through the Fas pathway.

19.5 Activation of T cells with anti-CD3 fails to induce apoptosis in mice bearing defective Fas and FasL genes.

19.6 Despite having a defective FasL, the gld mouse has a normal Fas gene and ligation of this with the anti-Fas antibody is sufficient to cause apoptosis.

19.7 This is an important way of controlling immune homeostasis. If cells were to keep on proliferating after antigen encounter the lymph nodes would explode!

19.8 This is a good question! The answer is not yet known and obviously requires a lot more critical thinking.

The role of regulatory T cells in peripheral tolerance

19.9 This is because autoreactive lymphocytes exist as part of the normal T cell repertoire. CD25$^+$ cells clearly suppress these cells and prevent them from causing disease. Precisely how the CD25$^+$ cells work is not yet known.

19.10 It seems likely that these cells arise in the thymus. Comparison of groups A–D shows that CD25$^-$ thymocytes can cause disease and that their activity is normally regulated by CD25$^+$ thymocytes.

19.11 It looks like tissues such as the adrenal gland are relatively resistant to disease. This could be: (a) because the antigenic targets for autoimmune disease are more sequestered in these tissues; (b) because mechanisms that create immunologically privileged sites operate in these tissues to some extent.

20. AUTOIMMUNITY AND AUTOIMMUNE DISEASE

20.1 It is thought that free DNA filtered in the kidney fixes to the glomerular basement membrane and can then bind anti-DNA antibodies, which then form an immune complex in situ. Complement is then fixed, resulting in local damage.

20.2 This is a vexed question. Although DNA–anti-DNA complexes are found in tissues, efforts to find these complexes in the serum have failed. In addition, immunizing lupus-prone animals with DNA does not produce clinical lupus. However, introduction of transgenes encoding anti-ds DNA in mice can produce lupus.

20.3 A possible explanation is that the mononuclear–phagocyte system becomes saturated and is therefore unable to clear the soluble complexes, which are thought to be most likely pathogenic. It is also possible that the reduction in the complement receptors on red cells (complement receptor-1, CR1) might also predispose to poor clearance of complexes.

20.4 Over 95% of patients with SLE have ANA as the major autoantibody. Antibodies to extractable nuclear antigens are also seen, but much less frequently. Anti-dsDNA antibodies are the most specific to SLE because anti-single-stranded antibodies are found in a variety of other situations, such as other autoimmune disease, a variety of infections, and inflammatory conditions.

21. KIDNEY TRANSPLANTATION

21.1 There is a great shortage of donor organs. The supply of cadaveric organs depends on the deaths of healthy individuals and the willingness of their relatives to allow donation. In addition, the available organs are given to recipients with the best HLA tissue match. At random there is less than a 1 in 20 000 chance of finding a perfect match. In the case of kidney transplantation the blood group antigens must be taken into account. Mrs X is blood group B, which is uncommon (<10%), and will therefore have antibodies to tissues from blood group A donors. Blood group A is the most common blood group (about 45% of individuals).

21.2 Mrs X is HLA identical to her brother Fred and shares one HLA haplotype with her brother Gary. There is a 1 in 4 chance of siblings inheriting the same mendelian characteristics from their parents.

21.3 Like Mrs X and her brother Fred, Bert and Edna are HLA identical to each other. Of the five children of Mrs X, four have Mr X as their biological father. However, it is clear that Dave was fathered by a man other than Mr X! This is not uncommon. Approximately 5–10% of children may be like Dave!

21.4 HLA identical: Fred; HLA haplotype match: Anne, Bert, Chas, Dave, Gary; complete HLA mismatch: Mr X.

21.5 If only HLA matching is considered, Fred would be the best donor because there would be no HLA mismatch between the donor and recipient.

21.6 Mrs X is blood group B, and therefore has antibodies to blood group A, thus excluding Mr X, Anne, Bert, Edna, and Fred. Only Chas, Dave, and Gary would be suitable. Because Mrs X is Rhesus positive (has antigen D) she will not have anti-Rh antibodies, so typing for this blood group can be ignored in this instance.

21.7 Chas is first choice. He is HLA haplotype matched, he is ABO compatible, and is a young man. In general, younger kidneys last for longer than older kidneys, and younger people respond better to surgery. Gary might be considered. He too is haplotype identical and ABO compatible. In fact, in terms of his ABO blood group, being blood group O he is the 'universal donor', having no A or B blood group antigens on his tissues. However, he is older and therefore not an ideal donor. Dave is like his brother Chas, HLA haplotype identical and ABO compatible with his mother. However, he is only 15 years of age and minors are generally excluded from this kind of surgery.

21.8 The organ from Chas suffered hyperacute rejection. This is mediated by preformed antibodies in the recipient binding to antigens in the graft. The antibodies fix complement and initiate the process of graft thrombosis. Platelet aggregation in the blood vessels blocks blood flow and the graft dies from lack of oxygen (ischemia).

21.9 The main stimuli for antibodies that cause hyperacute rejection are rejection of a previous graft, blood transfusions, and multiple pregnancies by the same partner. Mrs X has had four children by the same father, Mr X, and is likely to have become sensitized towards class I HLA antigens that Mr X has but Mrs X does not possess. In this case antibodies to either HLA-A3 or HLA-B7 could have been responsible for the hyperacute rejection.

21.10 Cross-match tests are used to detect anti-donor antibodies before transplantation. These tests are becoming more and more sensitive, but occasionally an antibody in low titer may go undetected. It appears that the cross-match test used at the time of the transplantation failed to detect an antibody to a HLA class I antigen. Fortunately this seldom happens today and hyperacute rejection is now very rare.

21.11 There were three family members who were originally suitable for donation, Chas, Dave, and Gary. The kidney from Chas was transplanted and rejected by Mrs X. Gary is now 60 and is less attractive as a donor. In any case he shares the HLA-B14 antigen with Mr X, which, because Mrs X is sensitized to Mr X, may be a target for hyperacute rejection. The possibility that Mrs X has an antibody to HLA-B14 would have to be investigated very carefully! Dave has by now reached the age of majority and could be considered as a donor. He is HLA haplotype identical to his mother, and is ABO compatible. Furthermore, because he has a different father, Dave is less likely to express antigens that might be the target of hyperacute rejection. It seems that Dave would be the most suitable donor. You might like to consider the emotional pressure this puts on Dave, given the seriousness of his mother's condition and how his family are likely to react if he wants to change his mind about donation.

21.12 Triple immunosuppression is standard in most transplant centers. It consists of corticosteroids plus either ciclosporin or tacrolimus and either azathioprine or mycophenolate. Some centers add other agents such as anti-lymphocyte serum or a monoclonal antibody to T cells (OKT3).

21.13 About one-third of all transplant recipients suffer an episode of acute rejection in the first few weeks. This is mainly a cell-mediated immune response.

21.14 The anti-rejection therapy often used is to give three large doses of corticosteroid on three consecutive days (totaling 2–3 g of corticosteroid). This causes apoptosis (programmed cell death) of activated lymphocytes and stops the rejection episode very effectively. Other anti-rejection therapies are used, such as OKT3 anti-T cell monoclonal antibody.

21.15 Because of the immunosuppression, patients are more prone to infection. The immunosuppressive drugs used can be toxic, so a dose reduction or a change in medication may be required. In addition, transplant patients on high doses of immunosuppression may develop a post-transplant lymphoproliferative disorder caused by Epstein–Barr virus infection. Reduction of the dose usually helps. In the longer term transplant patients have a higher risk of cancer because of depressed immune surveillance.

21.16 Mrs X finally lost her kidney transplant to chronic rejection. This may involve both immunological and non-immunological damage. The damage initiates a repair process involving the production of growth factors in the graft. None of the drugs in current use controls this process very well, so although the doctors change drugs and doses it is very difficult to control chronic rejection.

22. IMMUNITY TO CANCERS

22.1 When animals are immunized prophylactically, an adaptive immune response to the tumors is generated. When the same animal is challenged with live cancer cells later, the pre-existing immune response is amplified rapidly and can eliminate the tumor challenge. In contrast, when a naive animal is challenged, the tumor has generally already grown to a considerable size by the time an immune response to it is generated. At the same time, this developing immune response is compromised by a number of immune downregulatory mechanisms (immunosuppressive cytokines, $CD4^+$ suppressor T cells, etc.) that get activated in a host with a growing tumor. Thus, even as the tumor keeps growing, the immune response to it is compromised, and the host succumbs to the tumor.

22.2 The immunity elicited by HSP preparations does not derive from the HSP molecules per se, but from the antigenic peptides chaperoned by them. The HSPs isolated from tumors are associated with antigenic tumor-specific peptides in addition to the normal non-immunogenic peptides. In contrast, HSPs isolated from normal tissues are associated with only the non-immunogenic, normal peptides. This difference accounts for the lack of immunogenicity of normal tissue-derived HSPs and immunogenicity of tumor-derived HSPs.

22.3 Tumor antigens have been defined by a number of methods, such as by their ability to elicit tumor rejection in animal models or by recognition by T cells or antibodies in animals or in patients with cancer. Altogether, tumor antigens may be classified into the following classes:

- differentiation antigens;
- cancer testes antigens;
- unique mutated antigens;
- common mutated antigens (including oncogene-encoded antigens);
- viral antigens; and
- heat-shock proteins (which act as carriers of all of the above categories of antigen).

22.4 Concomitant immunity is defined as the presence in a host of an anti-tumor immune response that is able to successfully eradicate a tumor at one site in the concomitant presence of another tumor, which is apparently unresponsive to it. This phenomenon is a dramatic demonstration of the two opposing forces in cancer immunity – the immune response to a growing cancer, and the subversion of that response.

22.5 Several observations can be cited in support of downregulation of immune response in the tumor-bearing state, as follows:

(a) The phenomenon of concomitant immunity is clear evidence in this regard. Tumor-reactive and tumor-protective T cells can be isolated early during tumor growth. At a later stage, one can show the presence of immune-suppressive T cells in the same animal. The $CD4^+CD25^+$ regulatory T cells have also been shown to play a downregulatory influence on the anti-cancer immune response.

(b) The phenomenon that the same animal can reject a skin allograft while not rejecting a tumor expressing the same allo-MHC shows that there are local factors that make the tumor apparently resistant to the existing systemic anti-tumor immune response.

(c) The extensive murine and emerging clinical data that blocking of the CTLA-4 molecule (which transmits an inhibitory signal to activated T cells) leads to tumor regression or other forms of protection from tumor growth suggest strongly that the anti-tumor immune response in a cancer-bearing host is downregulated.

23. SEVERE ANAPHYLACTIC SHOCK

23.1 Traditionally, the term anaphylaxis has been used to describe a systemic clinical syndrome caused by IgE-mediated degranulation of mast cells and basophils. Susceptible individuals exposed to a sensitizing antigen produce specific IgE antibodies, which bind to high-affinity IgE receptors (FcεRI) found on mast cells and basophils. The receptor binds the Fc portion of the antibody, leaving the Fab binding sites available to interact with antigen. The avidity of this Fc binding reaction is high and therefore the dissociation of IgE from the receptors is slow, with a long half-life. On subsequent exposure, the antigen is bound by the IgE-receptor complexes, which causes receptor-mediated activation of the cells with release of preformed and de-novo synthesized mediators. Degranulation is rapid and completed within 30 minutes. These mediators, released on a large scale, are responsible for the clinical manifestations of anaphylaxis. The IgE-mediated mechanism of mast cell degranulation has been implicated in the pathogenesis of anaphylaxis triggered by a variety of agents. These include antibiotics (e.g. penicillins, cephalosporins), foods (e.g. milk, nuts, shellfish), foreign proteins (e.g. insulin, bee venom, latex), and pharmacological agents (e.g. streptokinase, vaccines). Patients who have anaphylaxis may or may not have a history of atopy. Natural exposure to common allergens such as pollen or dust mites is only very rarely a cause of anaphylaxis. However, when patients who also have asthma develop anaphylaxis due to venom, penicillin, or food antigens, the reactions are more dangerous because they can include rapid onset of bronchospasm.

Mast cell degranulation can occur by IgE-independent pathways. In these cases prior exposure is not a prerequisite because specific IgE antibodies are not involved. Three putative mechanisms of anaphylactoid reactions are given below.

- Blood, blood products, and immunoglobulins can cause an anaphylactoid reaction. The suggested mechanism is the formation of immune complexes with subsequent complement activation and production of C3a and C5a. Both of these complement components (anaphylatoxins) are capable of degranulating mast cells directly. In addition, both components increase vasopermeability and may induce hypotension.
- Certain therapeutic and diagnostic agents such as opiates, muscle relaxants, and contrast media are also capable of directly causing mast cell degranulation and anaphylaxis.
- 5–10% of asthmatic subjects produce a reaction to non-steroidal anti-inflammatory drugs (NSAIDs), such as aspirin or indometacin. Symptoms commonly include bronchospasm, rhinorrhea, and, rarely, vascular collapse. The ability of these agents to cause anaphylaxis appears to correlate with their effectiveness in inhibiting prostaglandin synthesis. The mechanism of this sensitivity is unknown, but increased leukotriene production occurs, which suggests that triggering of mast cells is part of the reaction.

23.2 There is a great variation in the timing and nature of anaphylactic symptoms. The onset is usually within seconds or minutes of exposure, though delays of an hour have been reported. The following are common presentations, which may occur singly or in combination:

- cutaneous – erythema, pruritus of hands, feet, and abdomen, urticaria, angioedema;
- respiratory – laryngeal edema causing hoarseness, which may progress to asphyxia, bronchoconstriction causing wheezing, rhinorrhea;
- cardiovascular – hypotension, arrhythmias, tachycardia, vascular collapse;
- gastrointestinal – cramping abdominal pain, nausea, vomiting, diarrhea.

The majority of cases of anaphylactic reaction are not fatal. It has been estimated that 1–2% of courses of penicillin therapy are complicated by systemic reactions, but only 10% of these are serious. In the USA some 400–800 people die annually from penicillin anaphylaxis, with a similar figure for contrast media – 70% of deaths result from respiratory complications (laryngeal edema and/or bronchospasm) with 25% resulting from cardiovascular dysfunction. Prompt treatment of anaphylaxis is essential because death may occur rapidly. The patient is placed in the recovery position, oxygen is given by mask, and 0.5–1.0 ml epinephrine 1:1000 w/v is injected intramuscularly. This has the effect of raising the blood pressure, relaxing bronchial smooth muscle, and preventing further mediator release. Intravenous antihistamines (e.g. 10 mg chlorpheniramine) can be useful because histamine can cause vasodilation, cardiac arrhythmias, and bronchospasm. Corticosteroids (e.g. 100 mg hydrocortisone) intravenously may help to reduce any late-phase response.

23.3 The first step is to obtain a thorough history of previous adverse reactions. The timing and nature of such reactions should be noted. Skin prick testing with insect venom is a fast and sensitive method of detecting anti-venom IgE. Radioallergosorbent tests can detect venom-specific IgE, but are positive in only 80% of those with significant reactions to venom skin prick tests. Immunotherapy is best reserved for those with life-threatening systemic reactions to insect venom. The patient is given increasing subcutaneous dosages and then a monthly maintenance dose of 100 µg. The clinical protection rate is in the order of 98% for both adults and children.

24. BLOOD GROUPS AND HEMOLYTIC DISEASE OF THE NEWBORN

24.1 Because Mrs Chareston has clearly becomed sensitized to Rhesus D, it is most likely that her first child is RhD$^+$. The alternative explanation, that she has become sensitized by a blood transfusion, is highly unlikely, because of the routine matching of this blood group when carrying out transfusions.

24.2 HDNB usually becomes more serious with successive pregnancies because the mother has become sensitized to the fetal red cells and successive sensitizations produce progressively stronger responses in the mother, and more serious disease in susceptible children.

24.3 Anti-Rhesus D antibodies are given to the mother to clear the fetal RhD$^+$ erythrocytes before they have a chance to sensitize the mother's immune system.

24.4 If the antibodies are given pre-partum, they would cross the placenta and produce or exacerbate HDNB in the fetus.

24.5 Rhesus prophylaxis is not always successful, but in this case there was no treatment after the first pregnancy and Mrs Chareston was already sensitized to RhD. Preventing further responses in an individual who is already sensitized is less likely to succeed because of the nature of secondary immune responses – less antigen is required to trigger the response.

24.6 The most likely explanation is that the child is Rh$^-$. Indeed this turns out to be the case. A Rh$^-$ child must have received a Rh$^-$ gene from both parents. Assuming that Mr Chareston is the father of the fourth child, we can say that his genotype is RhD$^+$/RhD$^-$ (heterozygote), and that the child received RhD$^-$ genes from both parents.

24.7 Antigens of the ABO blood group are carbohydrates and tend to induce IgM antibodies, which do not undergo affinity maturation or class switching (see Chapter 8). IgM does not cross the placenta (see Chapter 3) and so does not produce HDNB.

24.8 The second child was definitely not fathered by Mr Chareston because ABO blood groups are co-dominantly expressed. For a child to have the blood group B, one or both parents must have blood group B. As neither Mr or Mrs Chareston has this blood group, the B gene must have come from someone else.

24. TYPE II AHA OR AHA PLUS ITP WITH LUPUS

24.9 There is clinical evidence for a systemic disease, with arthritis, thrombocytopenia, anemia, and dermatitis. The positive Coombs' test indicates that there are serum antibodies to RBCs, and she also has anti-platelet antibodies. These circulating antibodies bind to cell surface antigens. When autoantibody–RBCs and autoantibody–platelets enter the liver and spleen, their Fc regions engage FcRs on resident macrophages, and are removed. This is a consuming process, which is accelerated by complement activation on the surface of RBCs and platelets, so that erythropoiesis and thrombopoiesis do not keep up with consumption, thus lowering the RBC and platelet counts.

Immune complex formation in the joints results from deposition of autoantibody–autoantigen complexes (e.g. anti-DNA antibody–nucleosome complexes). This is facilitated by altered vascular permeability. In-situ immune complex formation may also be operative, with free nucleosomes entering the joint space and binding to cell surface antigens or matrix components through charge–charge interactions. In either case the tissue-fixed immune complexes engage activating FcR (e.g. FcRγIII) to initiate inflammation.

24.10 Because local vascular permeability in other tissues was not altered, so immune complexes and activated inflammatory cells could not traverse vessels walls (because of size restriction). Alternatively, because autoantibodies specific for either cell surface antigens or matrix components were not present, inflammation was not observed within them. In some cases immune deposits may form in tissues, but, because vascular permeability is not altered or/and inflammatory cells are not stimulated, inflammation is not observed.

24.11 Ideally one could target only those B cells that produce the autoantibodies, using therapy directed at the antigen binding regions of the antibodies. However, because the antigen-binding regions are likely to vary from patient to patient, designing therapy for each individual is impractical. More general immunosuppression (e.g. against B cells) is reasonable and should be initiated. Using drugs that alter the clearance of RBCs and platelets is also a good strategy and this could be accomplished by either enhancing the expression of the inhibitory FcRs or decreasing the clearance of the activating FcR. Intravenous immunoglobulin (IVIg) works through this mechanism and would be a reasonable therapeutic choice.

25. TYPE III SERUM SICKNESS AFTER FACTOR IX ADMINISTRATION

25.1 This is a classic example of serum sickness induced by a 'foreign protein' factor IX because the boy does not produce it.

25.2 Corticosteroids were prescribed to decrease inflammation and for immunosuppression.

25.3 The likelihood that this type of reaction will develop again is relatively high because a memory response will be produced; T cells and plasma cells will respond more rapidly during re-exposure to the same foreign antigen (i.e. factor IX).

25.4 Short courses of corticosteroids with/without transient immunosuppression.

25.5 In this case, the clinical presentation is associated with an apparent non-inflammatory process (inferred because there is only proteinuria and inflammation has not reduced the glomerular filtration rate). Although the precise pathogenesis is uncertain, it is thought that the administered factor IX gets modified (becomes more positively charged) in the circulation. Because of its charge, it becomes trapped during normal filtration between the negatively charged glomerular basement membrane and glomerular epithelial cells, where it serves as a 'planted antigen' for circulating anti-factor IX antibodies. The Fc regions of local or 'in-situ' formed immune complexes activate the classical complement system, whereby C5b–9 causes sublytic injury to the epithelial cells, leading to their detachment. Because these cells normally participate in maintaining the glomerular integrity and limiting protein filtration, this pathologic process causes loss of large amounts of protein in the urine. Inflammation is not observed because FcR engagement on circulating inflammatory cells is prevented by the intact basement membrane. Pathologically there is epithelial cell 'effacement' from the basement membrane, but a conspicuous absence of inflammation. By contrast, in acute serum sickness, when immune deposits form between the endothelium and the basement membrane (or on the basement membrane), the deposited IgG engages FcR on circulating cells, and inflammation (with cellular infiltration) is the predominant feature.

25.6 The likelihood that this type reaction will develop again is relatively high because a memory response will be produced; T cells and plasma cells will respond more rapidly during re-exposure to the same foreign antigen (i.e. factor IX).

25.7 Immunosuppression until the proteinuria resolves to treat the present episode. Because the probability of immunologic recall is high, transient immunosuppression with each factor IX therapy is indicated to reduce the production of antibodies to factor IX.

26. A HYPERSENSITIVITY TYPE IV REACTION

26.1 Granulomas are composed of lymphocytes, macrophages, and epithelioid cells. The latter develop from macrophages following chronic antigenic stimulation and may fuse to form multinucleate giant cells, typical of granulomas. Cytokines involved in this process include T cell-derived IFNγ and TNF, both for the activation of macrophages and for the organization of the granuloma.

26.2 Histological examination at the site of a DTH reaction reveals edema of the dermis with an infiltrate of monocytes and lymphocytes. This resolves over 1–2 weeks. Granulomas do not form at the sites of DTH reactions if soluble antigen, such as tuberculin, is used. By contrast, in the lymph node a chronic granulomatous response develops as the mycobacteria survive within macrophages, leading to persistent stimulation of T cells and chronic inflammation.

26.3 CD4 T lymphocytes are the major cells responsible for the recognition of soluble recall antigens and the stimulation of DTH reactions.

26.4 Other infections, such as cat scratch fever due to *Bartonella henselae*, histoplasmosis, and tularemia, may cause granulomas in lymph nodes. These are diagnosed by the clinical pattern and microbial cultures. Sarcoidosis causes non-caseating granulomas and is diagnosed by clinical features, histology, and the absence of an infectious cause. Granulomas may also develop in response to foreign bodies, such as talc and silica, or exposure to beryllium.

26.5 The brother's DTH reaction is evidence of a strong T cell response to soluble antigens from *M. tuberculosis*. This indicates that he has been infected with *M. tuberculosis*, but does not mean that he has active tuberculosis disease at present. Normally he would have investigations to exclude active tuberculosis, and if this is not present he would be considered for chemoprophylaxis to eradicate the infection and prevent progression to disease in later life.

Appendix

This Page Intentionally Left Blank

APPENDIX

1

Major Histocompatibility Complex

The major histocompatibility complex (MHC) is the most polymorphic gene locus known in humans.

Originally the MHC class I and class II molecules from the principal loci (HLA-A, HLA-B, HLA-C, and HLA-D) were distinguished using specific antibodies, which could recognize polymorphic variants at these loci. For example, HLA-A2 identifies a particular variant at the HLA-A locus. These are termed serological HLA specificities.

Later it became clear that the HLA-D region contained several class II loci (HLA-DP, HLA-DQ, and HLA-DR), and that some serological specificities were specific for variants at these loci.

As serological analysis became more refined, it was found that some specificities could be subdivided into two or more types, using newer antibodies. For example, HLA-A9 is made up of two subgroups, HLA-A23 and HLA-A24.

In some cases the designation of the serologically defined MHC molecules was not fully certain, so they were given provisional or workshop designations ‘w’ (e.g. HLA-DPw1).

Later, sequence analysis of MHC genes was undertaken and the sequences could be related to the serological specificities. For example, the sequences termed HLA-B*1301, HLA-B*1302, HLA-B*1303, HLA-B*1304, HLA-B*1305, and HLA-B*1306 are six different sequences, all of which encode HLA-B molecules, which can be serologically defined by anti-HLA-B13 antibody.

Following the ‘*’, the first two digits indicate the serological specificity, and the subsequent digits the sequence number associated with that specificity. Hence several genetic variants can have the same serological specificity.

Finally, an additional two-digit figure may be added to identify different alleles that generate synonymous variants.

The number of genetic variants is very large. A current list (January 2006) can be viewed at http://www.anthonynolan.com/HIG/lists. This site also gives information on the relationship of serological specificities and genetically defined variants.

APPENDIX

2

Recognized CD Markers of Hemopoietic Cells and their Distribution

Filled box, molecule is expressed; open box, molecule not expressed or not determined yet; half-filled box, subpopulation only

★, activated cells only

B, basophil; G, granulocyte; B, B cell; Carb, carbohydrate; CCP, complement control protein domains; Clec, C-type lectin; CytR, hemopoietic cytokine receptor family; DC, dendritic cell; Ect, ectoenzyme; End, endothelium; Eo, eosinophil; IgSF, immunoglobulin supergene family; Intα, integrin alpha chain; Intβ, integrin beta chain; LC, Langerhans' cell; LSC, lymphoid stem cell; M, mononuclear phagocyte lineage; MC, mast cell; MSC, myeloid stem cell; Muc, mucin; N, neutrophil; FDC, follicular dendritic cell; NK, natural killer cell; P, platelet; Thy, thymocyte; p, precursor; RBC, red blood cell; RTK, receptor tyrosine kinase; TM4, tetramembrane pass family; PC, plasma cell; TM7, seven- transmembrane pass; G protein- coupled receptor; SC, stem cell; TNF, TNF-like; Scav, scavenger receptors; TNFR, TNF receptor/NGF receptor family; Sial, sialoglycan; T, T cell; ZnMP, zinc metalloproteinase.

This gives principal family relationships only. Some molecules are composite. The designation 'w' (workshop) indicates a provisional assignment. Detailed information on many of the CD molecules is distributed as 'CD guides' from http://www.ncbi.nlm.nih.gov/prow. A further set of CD numbers were assigned at the 7th International Workshop in June 2000, numbered CD167–CD247. Minor refinements and corrections were also made to existing designations. The new designations can be seen at http://gryphon.jr2.ox.ac.uk/.

Recognized CD markers of hemopoietic cells and their distribution

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD1a	presentation of lipids/glycolipids	IgSF	49	Thy						LC	
CD1b	presentation of lipids/glycolipids	IgSF	45	Thy						LC	
CD1c	presentation of lipids/glycolipids	IgSF	43	Thy	◢					DC	
CD1d	presentation of lipids/glycolipids	IgSF	55	Thy						■	■
CD1e	presentation of lipids/glycolipids	IgSF	50		■		■				
CD2	LFA-3 receptor (adhesion)	IgSF	50	■							
CD3	TCR signaling complex (γ,δ,ε) (ζ,υ)	IgSF IgSF	25,20,19, 16,22	■							
CD4	MHC class II receptor	IgSF	55	◢							
CD5	co-stimulator (activation)	Scav	67	■	◢						
CD6	co-stimulator (activation)	Scav	105	◢	◢						
CD7	signal transduction	IgSF	40	◢							LSC
CD8	MHC class I receptor	IgSF	36,32	◢							
CD9	platelet activation	TM4	24		p		■	Eo, B	■		
CD10	neutral endopeptidase	ZnMP	100	p	p			■			
CD11a	LFA-1	Intα	180	■	■	■	■	■			
CD11b	CR3	Intα	165			■	■	■			

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD11c	CR4	Intα	150				🟩	🟩			
CDw12		P-prot	120			🟩	🟩	🟩 Eo, N	🟩		
CD13	aminopeptidase N	ZnMP	150			◢	🟩	🟩			🟩 End
CD14	LPS-binding protein receptor		53–55				🟩	🟩			
CD15	sialyl Lewis	Carb					🟩	🟩			
CD16	FcγRIIIA/FcγRIIIB	IgSF	50–65			🟩	🟩				
CD16b	FcγRIIIB	IgSF	48					🟩			
CDw17	lactosyl ceramide	Carb	120		◢		🟩	🟩	🟩		
CD18	LFA-1/CR3/CR4	Intβ	95	🟩	🟩	🟩	🟩	🟩			
CD19	B cell co-receptor subunit	IgSF	95		🟩						🟩 FDC
CD20	Ca^{2+} channel	TM4	33–37		🟩						
CD21	CR2 B cell co-receptor subunit	CCP	140		🟩						🟩 FDC
CD22	adhesion molecule	IgSF	130–140		🟩						
CD23	FcεRII	Clec	45		◢ ★		◢ ★	🟩 Eo			
CD24	co-stimulator (activation)		41,38		🟩			🟩			
CD25	IL-2R	CytR	55	◢ ★	◢ ★						
CD26	dipeptidyl peptidase IV		120	◢ ★	◢ ★	🟩	🟩				
CD27	binds CD70	TNFR	55	🟩	◢	🟩					
CD28	binds CD80, CD86	IgSF	44	◢	◢						
CD29	VLA-1–VLA-6	Intβ	130	🟩	🟩	🟩	🟩	🟩			

CD30	signal transduction (apoptosis)	TNFR	120	★ ◢	★ ◢	★ ◢	🟩				
CD31	PECAM-1	IgSF	140	◢	🟩	🟩	🟩	🟩	🟩		🟩 End
CD32	FcγRII	IgSF	40		🟩		🟩	🟩			
CD33		IgSF	67				🟩				🟩 MSC
CD34	binds L-selectin	Muc	105–120								🟩 End MSC
CD35	CR1	CCP	160–260	◢	🟩	◢	🟩	Eo N			🟩 FDC
CD36	adhesion molecule	Scav	88				🟩		🟩		🟩 End
CD37	signal transduction (activation)	TM4	40–52	🟩	🟩		🟩	🟩			
CD38	ribosyl cyclase	Ect	45	★ ◢	PC ◢						🟩 LSC
CD39	ecto-apyrose	Ect	70–100		★ ◢	★ ◢	🟩				🟩 FDC
CD40	binds CD154, co-stimulatory	TNFR	48		🟩		🟩				🟩 FDC
CD41	GPIIb adhesion to matrix	Intα	120,25						🟩		
CD42a	GPIX adhesion		23						🟩		
CD42b	GPIbα adhesion		135,23						🟩		
CD42c	GPIbβ adhesion		22						🟩		
CD42d	GPV adhesion	85							🟩		
CD43	leukosialin (anti-adhesive)	Muc	95	🟩	◢	🟩	🟩	🟩			
CD44	Pgp-1 adhesion to matrix		80–95	🟩	🟩	🟩	🟩	🟩			🟩
CD45	leukocyte common antigen (LCA)		200	🟩	🟩	🟩	🟩	🟩		🟩	

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD45RA	restricted LCA		220	◢	🟩	🟩	🟩			🟩	
CD45RB	restricted LCA		190,205	◢	🟩	🟩	🟩	🟩		🟩	
CD45RO	restricted LCA		190	◢	🟩	🟩	🟩			🟩	
CD46	MCP (membrane cofactor protein)	CCP	66,56	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD47		IgSF	47–52	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD48	binds CD2 (rodents)	IgSF	41	🟩	🟩	🟩	🟩	🟩		🟩	
CD49a	VLA-1 adhesion to matrix	Intα	210	★ ◢							
CD49b	VLA-2 adhesion to matrix	Intα	160	🟩	🟩	🟩	🟩	🟩			
CD49c	VLA-3 adhesion to matrix	Intα	125		🟩						🟩
CD49d	VLA-4 adhesion	Intα	150,80,70	🟩	🟩		🟩	🟩		LC	
CD49e	VLA-5 adhesion to matrix	Intα	135,25	◢			🟩		🟩		
CD49f	VLA-6 adhesion to matrix	Intα	120,25	◢			🟩		🟩		
CD50	ICAM-3	IgSF	124	🟩	🟩	🟩	🟩	🟩		LC	
CD51	vitronectin receptor α	Intα	120,24						🟩		
CD52	Campath-1		21–28	🟩	🟩	🟩	🟩	🟩			🟩
CD53	signal transduction	TM4	32–40	🟩	🟩	🟩	🟩	🟩		🟩	
CD54	ICAM-1	IgSF	75–115	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD55	DAF (decay accelerating factor)	CCP	70	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

CD56	NCAM (neural cell adhesion molecule)	IgSF	220,135	◢		🟩					🟩
CD57	HNK-1	Carb	110	◢	◢	🟩	◢				
CD58	LFA-3 binds CD2	IgSF	40–65	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD59	protectin inhibits MAC		19	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CDw60	NeuAc-NeuAc-Gal	Carb		◢					🟩		
CD61	vitronectin receptor subunit	Intβ	105						🟩		🟩 End
CD62E	E-selectin	Clec	115								🟩 ★End
CD62L	L-selectin	Clec	75–80	🟩			🟩	🟩			
CD62P	P-selectin	Clec	150						🟩		🟩 ★End
CD63	adapter	TM4	53				🟩	🟩 ★	◢ ★		🟩 End
CD64	FcγRI	IgSF	70				🟩				
CD65	ceramide dodecasaccharide	Carb					🟩	🟩			
CD66a	BGB-1 adhesion	IgSF	180–200					🟩			
CD66b	was CD67 adhesion	IgSF	95–100					🟩			
CD66c	NCA adhesion	IgSF	90–95					🟩			
CD66d	CGMI	IgSF	30					🟩			
CD66e	CEA (carcinoembryonic antigen)	IgSF	180–200					🟩			
CD66f	pregnancy-specific glycoprotein	IgSF	54–72								🟩
CD68	macrosialin	Scav	110		◢		🟩	🟩 N, B			🟩 MC Sav
CD69	activation induction	Clec	32,28	◢ ★	◢ ★	◢ ★	◢ ★				◢ Thy

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD70	binds CD27	TNF	175,95,75	★ ◢	★ ◢		★ ◢				
CD71	transferrin receptor		95	★ ◢	★ ◢	★ ◢	★ ◢				★ 🟩
CD72		Clec	43,39		🟩						
CD73	ecto-5′-nucleotidase	Ect	69	◢	◢						End FDC 🟩
CD74	Ii (MHC class II invariant chain)		41,35,33		🟩		🟩			🟩	
CD75	Lactosamine	Carb			◢						
CD75s	α-2,6 sialyl lactosamine	Carb			◢						
CD77	globotriaosylceramide	Carb			◢						
CD79a	Igα	IgSF	33		🟩						
CD79b	Igβ	IgSF	39		🟩						
CD80	B7-1, binds CD28 and CD152	IgSF	60	★ ◢	★ ◢		🟩				
CD81	TAPA B cell co-receptor unit	TM4	26	🟩	🟩	🟩					End 🟩
CD82	signal transduction	TM4	60	🟩	🟩	🟩	🟩	🟩			
CD83	? antigen presentation	IgSF	43		◢					🟩	🟩
CDw84	? co-stimulation	IgSF	74	◢	◢		🟩		🟩		
CD85				◢	◢	◢	🟩			◢	
CD86	B7-2 binds CD28 and CD152	IgSF	80		★ 🟩		🟩				
CD87	urokinase plasminogen activator	35–39		🟩		🟩	🟩	🟩			🟩

CD88	C5αR	TM7	43				🟩	🟩			MC
CD89	FcαR	IgSF	50–70	◢	◢		🟩	🟩			
CD90	Thy-1	IgSF	25–35	Thy							
CD91	α2-macroglobulin receptor	LDLR	515,85				🟩				🟩
CD92			70				🟩	N	🟩		End
CD93			120				🟩	N			End
CD94	binds MHC class I	Clec	70	◢		🟩					
CD95	Fas signal transduction (apoptosis)	TNFR	43	★	★	🟩	🟩				🟩
CD96		IgSF	160	★		🟩					
CD97	binds CD55	EGFR	75–85	★	★		🟩	🟩		🟩	
CD98	modulates intracellular Ca^{2+}		80,45	🟩	🟩	🟩		🟩			
CD99		Muc	32	🟩	🟩	🟩					
CD100	? proliferation	IgSF	150	🟩	🟩	🟩	🟩	🟩	🟩	🟩	
CD101		IgSF	120	★			🟩	🟩		🟩	
CD102	ICAM-2 binds LFA-1	IgSF	55,65								End
CD103	αEb7 integrin	Intα	150,25	🟩	🟩	🟩					
CD104	β4 integrins	Intβ	220	◢							◢ End
CD105	endoglin binds TGFβ		90				★				End
CD106	VCAM-1 binds VLA-4	IgSF	90–110								★End
CD107a	LAMP-1		110	★				★	★		★End
CD107b	LAMP-2		120	★				★	★		★End

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD108	? adhesion		80	🟩	🟩	🟩					
CD109	platelet activation factor		170	★					★		End
CD114	G-CSFR		110–130				🟩	N	🟩		
CD115	M-CSFR	IgSF	150–130				🟩				
CD116	GM-CSFR α chain	CytR	60				🟩	N, Eo			
CD117	c-kit stem cell factor receptor	IgSF	145	Thy							SC
CD118	IFNα/β receptor			🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD119	IFNγ receptor		90–100	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD120a	TNFR-I	TNFR	60	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD120b	TNFR-II	TNFR	75–85	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD121a	IL-1R type I	IgSF	80	🟩							End
CDw121b	IL-1R type II	IgSF	60–70		🟩	🟩	🟩				
CD122	IL-2R β chain	CytR	75	◩	◩	🟩	🟩				
CD123	IL-3R	CytR	70				🟩	🟩			SC
CD124	IL-4R	CytR	130–150	◩	◩						
CD125	IL-5R	CytR	55–60		★			Eo, B			
CD126	IL-6R	IgSF	80	🟩	★		🟩				End
CD127	IL-7R		68–80	🟩	Pre B		◩				LSC
CD128	IL-8R chemokine receptor CXCR1	TM7	58	◩				🟩			
CD130	IL-6R, IL-11R common subunit	IgSF	130		★						🟩

CDw131	IL-3R, IL-5R common subunit	CytR	140					🟩			🟩 SC
CD132	IL-2R, IL-4R, IL-7R, IL-4R, IL-15R common subunit	CytR	64	🟩	🟩	🟩		🟩			🟩
CD133	AC133		120								◩ SC
CD134	? adhesion molecule	TNFR	50	◩ ★							🟩
CD135	growth factor receptor		155,130		🟩 p		🟩 p				SC
CDw136	macrophage-stimulating protein receptor	180					🟩				
CDw137	co-stimulatory (activation)	TNFR	30	🟩	🟩		🟩				
CD138	syndecan binds type I collagen				🟩						
CD139			228,209		🟩		🟩	🟩			🟩 FDC
CD140	PDGF receptor		180,180								🟩 End
CD141	thrombomodulin	Clec	105				🟩	🟩 N			🟩 End
CD142	tissue factor procoagulant		46								🟩
CD143	ACE (angiotensin converting enzyme)	ZnMP	170–180	◩ ★		◩ ★					🟩 End
CD144	VE-cadherin adhesion		135								🟩 End
CDw145			110,90,25								🟩 End
CD146	? adhesion	IgSF	130	◩ ★							🟩 End
CD147	neurothelin, basigin ? adhesion	IgSF	55–65	🟩	🟩	🟩	🟩	🟩	🟩		🟩 End
CD148	contact inhibition		250	🟩		🟩	🟩	🟩		🟩	🟩

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CDw149	≡ CD47	IgSF	47–52	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CDw150	SLAM ? co-stimulation		75–95	Thy	🟩					🟩	End
CD151	PETA3 ? signaling (adhesion)	TM4	32						🟩		End
CD152	CTLA-4 binds CD80 and CD86	IgSF	33	★							
CD153	CD30L binds CD30	TNF	38–40	★	🟩		★	🟩			
CD154	CD40L binds CD40	TNF	32–39	★							
CD155	polio virus receptor	IgSF	80–90	Thy			🟩				🟩
CD156	ADAM8	ZnMP	60–70				🟩	N			
CD157	ADP-ribosyl cyclase	Ect	42–50	◢	p		🟩	🟩			FDC End
CD158a	p58.1,p50.1 binds MHC class I (KIR)	IgSF	58,50			◢					
CD158b	p58.2,p50.2 binds MHC class I (KIR)	IgSF	58,50			◢					
CD159			43			🟩					
CD160			26	◢		◢					
CD161	NKRP-1 modulates cytotoxicity	Clec	44	◢		🟩					
CD162	PSGL-1 binds selectins	Muc	240	🟩	◢		🟩	🟩			
CD163	M130	Scav	130				🟩				
CD164	MGC-24 adhesion to stromal cells	Muc	80								SC
CD165	AD2 adhesion to thymic epithelium		37	Thy		🟩			🟩		🟩
CD166	ALCAM binds CD6	IgSF	100	★			★				🟩

CD167a	DDR1	RTK	120								
CD168	RHAMM		84–88	◢ Thy			🟩				
CD169	Sialoadhesin	Siglec	185				◢	N			
CD170	Siglec-5	Siglec	140				◢				
CD171	L1	FN/IgSF	200	◢	🟩		🟩				
CD172a			110	◢			🟩				SC
CD173	Blood gr. H (2)	Carb							🟩		
CD174	Lewis Y	Carb									SC
CD175	TN	Carb									SC
CD175s	Sialyl-TN	Carb									RBC
CD176	Thomson-Fr. Antigen	Carb									RBC
CD177	NB1		56–62								
CD178	Fas ligand	TNF	38–42	★				◢ N			
CD179a	V pre beta	IgSF	16		p						
CD179b	Lambda-5	IgSF	22		p						
CD180	Rp105/Bgp95	TLR	95–105		◢		🟩			🟩	
CD183	CXCR3	TM7	40	★		★					End
CD184	CXCR4	TM7	45	◢	🟩		🟩			🟩	
CD195	CCR5	TM7	45	◢			🟩				
CDw197	CCR7	TM7	45	◢							
CD200	Ox-2		45–50	◢ ★	🟩						
CD201	EPCR	Ect	50								◢ End

Continued

Listing of all recognized serological and cellular HLA specificities—*cont'd*

CD	identity/function	family	mol. wt (kDa)	T cell	B cell	NK cell	monocyte/ macrophage	granulocyte	platelet	dendritic cell	other
CD202b	TEL/Tie2	RTK	150								🟩 SC, End
CD203c	E-NPP3	Ect	130–150					B	MegaK		
CD204	Macrophage Scavenger R.	Scav	220				🟩				
CD205	DEC205	Clec	205							🟩	
CD206	Macrophage mannose R.	Clec	180				🟩			◢	
CD207	Langerin	Clec	40							LC	
CD208	DC-LAMP		70–90							🟩	
CD209	DC-SIGN		44							◢	
CD210	IL-10R		90–110	🟩	🟩	🟩	🟩				
CD212	IL-12R		100	🟩 ★		🟩 ★					
CD213a1	IL-13R A1		65		🟩		🟩				
CD213a2	IL-13R A2		65		🟩		🟩				
CD217	IL-17R		120	🟩	🟩	🟩	🟩	🟩		🟩	
CD220	Insulin receptor	RTK	140 + 70	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD221	IgF1 receptor	RTK	140 + 70	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD222	IgF2 receptor		250	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD223	LAG3	IgSF	70	🟩 ★		🟩 ★					
CD224	γ-glutamyl transferase	Ect	27, 68	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD225	Leu13		17	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD226	DNAM-1	IgSF	65	🟩		🟩	🟩		🟩		
CD227	MUC1	Muc	300								🟩
CD228	Melanotransferrin		80–95								🟩 SC

CD229	Ly-9	IgSF	95, 110	🟩	🟩						
CD230	PrPc		35	🟩	🟩	🟩	🟩	🟩		🟩	🟩
CD231	TALLA-1	TM4	30–45	🟩							
CD232	VESPR		200	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
CD233	Band3		90								RBC
CD234	Duffy	TM7	35–45								RBC
CD235a	Glycophorin-A		36								RBC
CD235b	Glycophorin-B		20								RBC
CD236	Glycophorin-CD		32/23								RBC
CD236R	Glycophoein-X		32								RBC
CD238	Kell		93								RBC
CD239	Lutheran		78–85								RBC
CD240CE	Rhesus CE		30–32								RBC
CD240D	Rhesus D		30–32								RBC
CD241	Rhesus 50gp		50								RBC
CD242	ICAM-4	IgSF	42								RBC
CD243	MDR1		180								SC
CD244	2B4	IgSF	70	🟩		🟩					
CD245	P220/240		220–240	◢				🟩	🟩		
CD246		RTK	80	🟩							
CD247			16	🟩		🟩					

For the most recent additions to the list of recognized CD markers, see Protein Reviews on the Web at http://mpr.nci.nih.gov/prow.

APPENDIX

3

The Major Cytokines

cytokine	immune system source	other cells	principal targets	principal effects
IL-1α IL-1β	macrophages, LGLs, B cells	endothelium, fibroblasts, astrocytes, etc.	T cells, B cells, macrophages, endothelium, tissue cells	lymphocyte activation, macrophage stimulation, ↑ leukocyte/endothelial adhesion, pyrexia, acute phase proteins
IL-2	T cells		T cells	T cell proliferation and differentiation, activation of cytotoxic lymphocytes and macrophages
IL-3	T cells	stem cells		multilineage colony stimulating factor
IL-4	T cells		B cells, T cells	B cell growth factor, isotype selection, IgE, IgG1
IL-5	T cells		B cells	B cell growth and differentiation, IgA selection
IL-6	T cells, B cells	fibroblasts, macrophages	B cells, hepatocytes	B cell differentiation, induces acute phase proteins
IL-7		bone marrow stromal cells	pre-B cells, T cells	B cell and T cell proliferation
IL-8	monocytes	fibroblasts	neutrophils, basophils, T cells, keratinocytes	chemotaxis, angiogenesis, superoxide release, granule release
IL-9	T cells			enhances T cell survival, mast cell activation, synergy with erythropoietin
IL-10	T cells		TH1 cells	inhibition of cytokine synthesis
IL-11		bone marrow stromal cells, fibroblasts	hemopoietic progenitors, osteoclasts	osteoclast formation, colony stimulating factor, elevates platelet count in vivo, inhibits proinflammatory cytokine production
IL-12		monocytes	T cells	induction of TH1 cells
IL-13	activated T cells		monocytes, B cells	B cell growth and differentiation, inhibits proinflammatory cytokine production
IL-14	T cells			stimulates proliferation of activated B cells, inhibits Ig secretion
IL-15	monocytes	epithelium, muscle	T cells, activated B cells	proliferation

cytokine	immune system source	other cells	principal targets	principal effects
IL-16	eosinophils, CD8$^+$ T cells		CD4$^+$ T cells	chemoattraction of CD4$^+$ cells
IL-17	CD4$^+$ T lymphocytes		epithelium, fibroblasts, endothelium	release of IL-6, IL-8, G-CSF, PGE$_2$, enhances ICAM-1, stimulates fibroblasts to sustain CD34$^+$ progenitors
IL-18	macrophages	hepatocytes, keratinocytes	PBMC, co-factor in TH1 induction	induces IFNγ production, enhances NK activity
IL-21	T cells, mast cells		T cells, B cells, mast cells, eosinophils, hepatocytes	induces acute phase reactants, levels raised after LPS
IL-22	activated T cells		TH2 cells	inhibits IL-4 production
TGFβ1			most cell types	downregulates inflammatory cytokine production, promotes wound healing responses and scar tissue, growth inhibition
TNFα	macrophages, mast cells, lymphocytes		macrophages, granulocytes, tissue cells	activation of macrophages, granulocytes and cytotoxic cells, leukocyte/endothelial cell adhesion, cachexia, pyrexia, induction of acute phase protein, stimulation of angiogenesis, enhanced MHC class I production
TNFβ (LT)	lymphocytes			as for TNFα
IFNα	leukocytes	epithelia, fibroblasts	tissue cells	MHC class I induction, antiviral state, stimulation of NK cells, antiproliferative, stimulates IL-12 production and TH1 cells
IFNβ	fibroblasts, epithelia		tissue cells, leukocytes	MHC class I induction, antiviral state, antiproliferative
IFNγ	T cells, NK cells	epithelia, fibroblasts	leukocytes, tissue cells, TH2 cells	MHC class I and II induction, macrophage activation, ↑ endothelial cell/lymphocyte adhesion, MØ cytokine synthesis, antiviral state, antiproliferative (TH1 cells)
M-CSF	monocytes	endothelium, fibroblasts		proliferation of macrophage precursors
G-CSF	macrophages	fibroblasts	stem cells	stimulates division and differentiation
GM-CSF	T cells, macrophages	endothelium, fibroblasts		proliferation of granulocyte and macrophage precursors and activators
MIF	T cells, macrophages		macrophages	migration inhibition, macrophage activation, enhances T cell activation

APPENDIX

4

Human Chemokines and their Receptors

The table below lists the major human chemokine receptors and their principal ligands. Individual chemokines (e.g. CCL5, RANTES) often bind to more than one receptor. Similarly, individual receptors (e.g. CXCL1) often bind to more than one chemokine. The systematic nomenclature has now superceded the descriptive names, although the older names are widespread in the scientific literature and are still sometimes encountered in recent papers. The table lists only the most-frequently used names; some chemokines have several names.

BCA, B cell attracting chemokine; BRAK, breast and kidney expressed chemokine; CTACK, cutaneous T cell attracting chemokine; ENA, epithelial-cell derived neutrophil activating protein; GCP, granulocyte chemotactic protein; Gro, growth related oncogene; HCC, hemofiltrate CC-chemokine; I-309, inducible protein-309; IL-8, interleukin-8; IP-10, inducible protein-10; I-TAC, inducible T cell α chemoattractant; LARC, liver and activation regulated chemokine; MCP, monocyte chemotactic protein; MDC, macrophage-derived chemokine; MEC, mammary enriched chemokine; Mig, monokine induced by interferon-γ; MIP, macrophage inflammatory protein; NAP, neutrophil activating peptide; MPIF, myeloid progenitor inhibitory factor; PARC, pulmonary and activation regulated chemokine; PF4, platelet factor-4; RANTES, regulated on activation, normal T cell expressed and secreted; SCM, single-C-motif; SDF, stromal cell-derived factor; SLC, secondary lymphoid tissue chemokine; TARC, thymus and activation regulated chemokine; TECK, thymus-expressed chemokine.

		Receptors, CCL1–CCL10									
Chemokine	**Name**	1	2	3	4	5	6	7	8	9	10
CCL1	I-309								■		
CCL2	MCP-1	■	■								
CCL3	MIP-1α	■			■						
CCL4	MIP-1β	■							■		
CCL5	RANTES	■		■	■						
CCL7	MCP-3	■	■								
CCL8	MCP-2	■	■								
CCL11	Eotaxin		■								
CCL13	MCP-4	■	■								
CCL14	HCC-1	■									
CCL15	MIP-1δ	■									
CCL16	HCC-4	■	■								
CCL17	TARC				■						
CCL18	PARC			■					■		
CCL19	MIP-3β							■			

Chemokine	Name	Receptors, CCL1–CCL10									
		1	2	3	4	5	6	7	8	9	10
CCL20	MIP-3α						■				
CCL21	6Ckine							■			
CCL22	MDC										
CCL23	MPIF-1	■									
CCL24	Eotaxin-2			■							
CCL25	TECK									■	
CCL26	Eotaxin-3		■	■							■
CCL27	CTACK										■
CCL28	MEC			■							■

Chemokine	Name	Receptors CXCR1-6					
		1	2	3	4	5	6
CXCL1	Groα		■				
CXCL2	Groβ		■				
CXCL3	Groγ		■				
CXCL4	PF4			■			
CXCL5	ENA-78	■	■				
CXCL6	GCP-2	■	■				
CXCL7	NAP-2		■				
CXCL8	IL-8	■	■				
CXCL9	Mig			■			
CXCL10	IP-10			■			
CXCL11	ITAC			■			
CXCL12	SDF-1				■		
CXCL13	BCA-1					■	
CXCL14	BRAK						
CXCL16	–						■

Chemokine	Name	Receptor
XCL1	Lymphotactin	XCR1
XC:2	SCM-1b	XCR1
CX3CL1	Fractalkine	CX3CR1

This Page Intentionally Left Blank

Glossary

Acquired immune deficiency syndrome (AIDS). A progressive immune deficiency caused by infection of CD4 T cells with the human retrovirus HIV.

Activation-induced cytidine deaminase (AID). An enzyme expressed in activated B cells that causes somatic mutation in the immunoglobulin gene locus.

Acute phase proteins. Serum proteins whose levels increase during infection or inflammatory reactions.

ADCC (antibody-dependent cell-mediated cytotoxicity). A cytotoxic reaction in which Fc receptor-bearing killer cells recognize target cells via specific antibodies.

Adhesion molecules. Cell surface molecules involved in the binding of cells to extracellular matrix or to neighboring cells, where the principal function is adhesion rather than cell activation (e.g. integrins and selectins).

Adjuvant. A substance that non-specifically enhances the immune response to an antigen.

AFCs (antibody-forming cells). Functionally equivalent to plasma cells.

Affinity. A measure of the binding strength between an antigenic determinant (epitope) and an antibody-combining site.

Affinity maturation. The increase in average antibody affinity frequently seen during a secondary immune response.

Allelic exclusion. Occurs when the use of a gene from the maternal or paternal chromosome prevents the use of the other. This is seen with antibody and T cell receptor genes.

Allergen. An agent (e.g. pollen, dust, animal dander) that causes IgE-mediated hypersensitivity reactions.

Allergy. Originally defined as altered reactivity on second contact with antigen; now usually refers to a type I hypersensitivity reaction.

Allotype. The protein of an allele that may be detectable as an antigen by another member of the same species.

Alternative pathway. The activation pathways of the complement system involving C3 and factors B, D, P, H, and I, which interact in the vicinity of an activator surface to form an alternative pathway C3 convertase.

Amplification loop. The alternative complement activation pathway that acts as a positive feedback loop when C3 is split in the presence of an activator surface.

Anaphylatoxins. Complement peptides (C3a and C5a) that cause mast cell degranulation and smooth muscle contraction.

Anaphylaxis. An antigen-specific immune reaction mediated primarily by IgE that results in vasodilation and constriction of smooth muscle, including those of the bronchus, and may result in death.

Anergy. Failure to make an immune response following stimulation with a potential antigen.

Antagonist peptides. Analogs of antigenic peptides that bind to MHC molecules and prevent stimulation of specific clones of T cells.

Antibody. A molecule produced by animals in response to antigen that has the particular property of combining specifically with the antigen that induced its formation.

Antigen. A molecule that reacts with preformed antibody and the specific receptors on T and B cells.

Antigen receptors. The lymphocyte receptors for antigens including the T cell receptor (TCR) and surface immunoglobulin on B cells, which acts as the B cell's antigen receptor (BCR).

Antigen presentation. The process by which certain cells in the body (antigen-presenting cells) express antigen on their cell surface in a form recognizable by lymphocytes.

Antigen processing. The conversion of an antigen into a form in which it can be recognized by lymphocytes.

Antigenic determinants. See 'epitopes'.

Antigenic peptides. Peptide fragments of proteins that bind to MHC molecules and induce T cell activation.

Antiviral proteins. Proteins whose synthesis is induced by interferons. They become activated if the cell is infected by virus and limit viral replication.

APCs (antigen-presenting cells). A variety of cell types that carry antigen in a form that can stimulate lymphocytes.

Apoptosis. Programmed cell death that involves nuclear fragmentation and condensation of cytoplasm, plasma membranes, and organelles into apoptotic bodies.

Arthus reaction. Inflammation seen in the skin some hours following injection of antigen. It is a manifestation of a type III hypersensitivity reaction.

Atopy. The clinical manifestation of type I hypersensitivity reactions, including eczema, asthma, rhinitis, and food allergy.

Autocrine. This refers to the ability of a cytokine to act on the cell that produced it.

Autoimmunity. Immune recognition and reaction against the individual's own tissue.

Avidity. The functional combining strength of an antibody with its antigen, which is related to both the affinity of the reaction between the epitopes and paratopes, and the valencies of the antibody and antigen.

β_2-Microglobulin. A polypeptide that constitutes part of some membrane proteins including the class I MHC molecules.

B7-1 (CD80), B7-2 (CD86). Two molecules that are present on antigen-presenting cells. They ligate CD28 on T cells and act as powerful co-stimulatory signals.

B cells. Lymphocytes that develop in the bone marrow in adults and produce antibody. They can be subdivided into two groups, B1 and B2. B1 cells use minimally mutated receptors, which are close to the germline immunoglobulin sequences, whereas B2 cells are the major responding population in conventional immune responses to protein antigens.

B cell co-receptor complex. A group of cell surface molecules consisting of complement receptor type 2 (CD21), CD81, and CD19, which act as a co-stimulatory receptor on mature B cells.

B cell receptor complex (BCR). B cell surface immunoglobulin and its associated signaling molecules, CD79a and CD79b.

Basophil. A population of polymorphonuclear leukocytes that stain with basic dyes and have important roles in the control of inflammation.

BCG (bacille Calmette–Guérin). An attenuated strain of *Mycobacterium tuberculosis* used as a vaccine, an adjuvant, or a biological response modifier in different circumstances.

Bcl-2. A molecule expressed transiently on activated B cells that have been rescued from apoptosis.

Biozzi mice. Lines of mice bidirectionally bred to produce low or high antibody responses to a variety of antigens (originally sheep erythrocytes).

Blood groups. Sets of allelically variable molecules expressed on red cells and sometimes other tissues that may be the target of transfusion reactions.

Bradykinin. A vasoactive nonapeptide that is the most important mediator generated by the kinin system.

Bursa of Fabricius. A lymphoepithelial organ that is the site of B cell maturation and is found at the junction of the hindgut and cloaca in birds.

Bystander lysis. Complement-mediated lysis of cells in the immediate vicinity of a complement activation site that are not themselves responsible for the activation.

C domains. The constant domains of antibody and the T cell receptor. These domains do not contribute to the antigen-binding site and show relatively little variability between receptor molecules.

C genes. The gene segments that encode the constant portion of the immunoglobulin heavy and light chains and the α, β, γ, and δ chains of the T cell antigen receptor.

c-Kit (CD117). A receptor for stem cell factor, which is required for the early development of leukocytes.

C1–C9. The components of the complement classical and lytic pathways, which are responsible for mediating inflammatory reactions, opsonization of particles, and lysis of cell membranes.

C3 convertases. The enzyme complexes C3b,Bb and C4b2a that cleave complement C3.

Capping. A process by which cell surface molecules are caused to aggregate (usually using antibody) on the cell membrane.

Carrier. An immunogenic molecule or part of a molecule that is recognized by T cells in an antibody response.

Caspases. A group of enzymes that are particularly involved in the transduction of signals for apoptosis.

Cathelicidins. A group of cytotoxic peptides produced by granulocytes.

CD markers. Cell surface molecules of leukocytes and platelets that are distinguishable with monoclonal antibodies and may be used to differentiate different cell populations.

CDRs (complementarity determining regions). The sections of an antibody or T cell receptor V region responsible for antigen or antigen–MHC molecule binding.

Cell adhesion molecules (CAMs). A group of proteins of the immunoglobulin supergene family involved in intercellular adhesion, including ICAM-1, ICAM-2, ICAM-3, vascular cell adhesion molecule-1 (VCAM-1), mucosal addressin cell adhesion molecule-1 (MAdCAM-1), and platelet endothelial cell adhesion molecule (PECAM).

Central tolerance. Tolerance of T cells or B cells induced during their development in the thymus or bone marrow.

Chemokines. A large group of cytokines falling into four families, of which the main families are the CC and the CXC group. Chemokines are designated as ligands belonging to a particular family (e.g. CCL2). Many chemokines have older descriptive names, for example CCL2 is macrophage chemotactic protein-1 (MCP-1). They act on G protein-linked, seven-transmembrane pass receptors and have a variety of chemotactic and cell-activating properties.

Chemokinesis. Increased random migratory activity of cells.

Chemotaxis. Increased directional migration of cells, particularly in response to concentration gradients of certain chemotactic factors.

Ciclosporin. A T cell suppressive drug that is particularly useful in suppression of graft rejection.

Class I/II/III MHC molecules. Three major classes of molecule are coded within the MHC. Class I molecules have one MHC-encoded peptide complexed with β_2-microglobulin, class II molecules have two MHC-encoded peptides which are non-covalently associated, and class III molecules are other molecules including complement components.

Class I/II restriction. The observation that immunologically active cells will cooperate effectively only when they share MHC haplotypes at either the class I or class II loci.

Class switching. The process by which an individual B cell can link immunoglobulin heavy chain C genes to its recombined V gene to produce a different class of antibody with the same specificity. This process is also reflected in the overall class switch seen during the maturation of an immune response.

Classical pathway. The pathway by which antigen–antibody complexes can activate the complement system, involving components C1, C2, and C4, and generating a classical pathway C3 convertase.

Clonal selection. The fundamental basis of lymphocyte activation in which antigen selectively causes activation, division, and differentiation only in those cells that express receptors with which it can combine.

CMI (cell-mediated immunity). A term used to refer to immune reactions that are mediated by cells rather than by antibody or other humoral factors.

Collectins. A group of large polymeric proteins, including conglutinin and mannan-binding lectin (MBL), that can opsonize microbial pathogens.

Complement. A group of serum proteins involved in control of inflammation, activation of phagocytes, and lytic attack on cell membranes. The system can be activated by interaction with the antibodies of the immune system (classical pathway).

Complement control protein (CCP) domains (also called **short consensus repeats**). A domain structure found in many proteins of the complement classical and alternative pathways and in some complement receptors and control proteins.

Complement receptors (CR1–CR4 and C1qR). A set of four cell surface receptors for fragments of complement C3. CR1 and CR2 have numerous complement control protein (CCP) domains, and CR3 and CR4 are integrins. C1qR binds C1q.

ConA (concanavalin A). A mitogen for T cells.

Congenic. Animals that are genetically constructed to differ at one particular locus.

Conjugate. A reagent that is formed by covalently coupling two molecules together, such as fluorescein coupled to an immunoglobulin molecule.

Constant regions. The relatively invariant parts of immunoglobulin heavy and light chains, and the α, β, γ, and δ chains of the T cell receptor.

Contact hypersensitivity. A delayed inflammatory reaction on the skin seen in type IV hypersensitivity.

Co-stimulation. The signals required for the activation of a lymphocyte, in addition to the antigen-specific signal delivered via their antigen receptors. CD28 is an important co-stimulatory molecule for T cells and CD40 for B cells.

Cross-reaction. The sharing of antigenic determinants by two different antigens.

CSFs (colony stimulating factors). A group of cytokines that control the differentiation of hematopoietic stem cells.

CTLA-4 (CD152). A downregulatory signaling molecule of T cells that competes with CD28 for ligation by B7 on antigen-presenting cells.

Cytokines. A generic term for soluble molecules that mediate interactions between cells.

Cytotoxic T cells. Cells that can kill virally infected targets expressing antigenic peptides presented by MHC class I molecules.

D genes. Sets of gene segments lying between the V and J genes in the immunoglobulin heavy chain genes, and in the T cell receptor β and δ chain genes, which are recombined with V and J genes during ontogeny.

Decay accelerating factor (DAF). A cell surface molecule on mammalian cells that limits activation and deposition of complement C3b.

Defensins. A group of small antibacterial proteins produced by neutrophils.

Degranulation. Exocytosis of granules from cells such as mast cells and basophils.

Dendritic cells. A set of cells present in tissues that capture antigens and migrate to the lymph nodes and spleen, where they are particularly active in presenting the processed antigen to T cells. Dendritic cells can be derived from either the lymphoid or mononuclear phagocyte lineages.

DM molecules. Molecules related to MHC class II molecules that are required for loading antigenic peptides onto class II molecules.

Domain. A region of a peptide having a coherent tertiary structure. Both immunoglobulins and MHC class I and II molecules have immunoglobulin supergene family domains.

Dominant idiotypes. Individual idiotypes that are present on a large proportion of the antibodies generated by a particular antigen.

DTH (delayed-type hypersensitivity). This term includes the delayed skin reactions associated with type IV hypersensitivity.

Education of T cells. The process by which developing thymocytes are selected for those that recognize peptides on self MHC molecules, but not for those that recognize self antigenic peptides.

Effector cells. A functional concept, which in context means those lymphocytes or phagocytes that produce an end effect.

Eicosanoids. Products of arachidonic acid metabolism including prostaglandins, leukotrienes, and thromboxanes.

Endocytosis. Internalization of material by a cell by phagocytosis or pinocytosis.

Endothelium. Cells lining blood vessels and lymphatics.

Endotoxin. Lipolysaccharide produced by Gram-negative bacteria that activates B cells and macrophages.

Enhancement. Prolongation of graft survival by treatment with antibodies directed towards the graft alloantigens.

Eosinophils. A population of polymorphonuclear granulocytes that stain with acidic dyes and are particularly involved in reactions against parasitic worms and in some hypersensitivity reactions.

Epithelioid cells. A population of activated mononuclear phagocytes present in granulomatous reactions.

Epitopes. The parts of an antigen that contact the antigen-binding sites of an antibody or the T cell receptor.

Epstein–Barr virus (EBV). Causal agent of Burkitt's lymphoma and infectious mononucleosis that has the ability to transform human B cells into stable cell lines.

Fab. The part of an antibody molecule that contains the antigen-combining site consisting of a light chain and part of the heavy chain; it is produced by enzymatic digestion.

Factors B, P, D, H, and I. Components of the alternative complement pathway.

Fas (CD95). A molecule expressed on a variety of cells that acts as a target for ligation by Fas ligand (FasL) on the surface of cytotoxic lymphocytes.

Fc. The portion of an antibody that is responsible for binding to antibody receptors on cells and the C1q component of complement.

Fc receptors. Surface molecules on a variety of cells that bind to the Fc regions of immunoglobulins. They are antibody class specific and isotype selective.

Ficolins. A group of opsonins that recognize carbohydrate PAMPs.

Flow cytometry. Analysis of cell populations in suspension according to each individual cell's expression of selected surface markers.

Fluorescence-activated cell sorter (FACS). A machine that analyzes cells by flow cytometry and then allows them to be sorted into different populations and collected separately.

Follicular dendritic cells (FDCs). Antigen-presenting cells in the B cell areas of lymphoid tissues that retain stores of antigen.

Formyl-methionyl peptides. Prokaryotes initiate protein synthesis with f-Met. Peptides such as f-Met-Leu-Phe are highly chemotactic for mononuclear phagocytes and neutrophils.

Framework segments. Sections of antibody V regions that lie between the hypervariable regions.

Freund's adjuvant. An emulsion of aqueous antigen in oil. Complete Freund's adjuvant contains killed *Mycobacterium tuberculosis,* whereas incomplete Freund's adjuvant does not.

Frustrated phagocytosis. A term to describe the events that occur when a phagocyte attempts to internalize an antigen or antigenic particle, but is unable to do so (e.g. because of its size).

γδ T cells. The minor subset of T cells that express the γδ form of the T cell receptor.

GALT (gut-associated lymphoid tissue). The accumulation of lymphoid tissue associated with the gastrointestinal tract.

Genetic association. The condition where particular genotypes are associated with other phenomena, such as particular diseases.

Genetic restriction. The term used to describe the observation that lymphocytes and antigen-presenting cells cooperate most effectively when they share particular MHC haplotypes.

Genome. The total genetic material contained within the cell.

Genotype. The genetic material inherited from parents; not all of it is necessarily expressed in the individual.

Germinal centers. Areas of secondary lymphoid tissue in which B cell differentiation and antibody class switching occur.

Germline. The genetic material that is passed down through the gametes before it is modified by somatic recombination or maturation.

Giant cells. Large multinucleated cells sometimes seen in granulomatous reactions and thought to result from the fusion of macrophages.

GPI (glycosylphosphatidylinositol) – linkage. A way in which proteins become attached to the outer leaflet of the phospholipid bilayer that forms the plasma membrane.

Granulocytes. Neutrophils, eosinophils, and basophils.

Granulomatous reactions. Chronic inflammatory reactions (often a manifestation of type IV hypersensitivity) caused by a failure to clear antigen.

Granzymes. Granule-associated enzymes of cytotoxic T cells and large granular lymphocytes.

GVHD (graft versus host disease). A condition caused by allogeneic donor lymphocytes reacting against host tissue in an immunologically compromised recipient.

H-2. The mouse major histocompatibility complex.

Hemagglutination. Clumping of erythrocytes caused by antibody. This forms the basis of a number of immunoassays and blood group typing.

Haplotype. A set of genetic determinants located on a single chromosome.

Hapten. A small molecule that can act as an epitope, but is incapable by itself of eliciting an antibody response.

Helper (T_H) cells. A functional subclass of T cells that can help to generate cytotoxic T cells and cooperate with B cells in the production of antibody responses. Helper cells recognize antigen in association with MHC class II molecules.

Heterologous. Refers to interspecies antigenic differences.

HEV (high endothelial venule). An area of venule from which lymphocytes migrate into lymph nodes.

Hinge. The portion of an immunoglobulin heavy chain between the Fc and Fab regions that permits flexibility within the molecule and allows the two combining sites to operate independently. The hinge region is usually encoded by a separate exon.

Histamine. A major vasoactive amine released from mast cell and basophil granules.

Histocompatibility. The ability to accept grafts between individuals.

HIV (human immunodeficiency virus). The causative agent of acquired immune deficiency syndrome (AIDS).

HLA. The human major histocompatibility complex.

Homologous restriction factors. Complement components that restrict the action of the membrane attack complex on cells of the host.

Humoral. Pertaining to the extracellular fluids, including the serum and lymph.

Hybridoma. Cell line created in vitro by fusing two different cell types, usually lymphocytes, one of which is a tumor cell.

5-Hydroxytryptamine (serotonin). A vasoactive amine present in platelets and a major mediator of inflammation in rodents.

Hypersensitivity. An inordinately strong immune response that causes more damage than the antigen or pathogen that induced the response.

Hypervariable region. The most variable areas of the V domains of immunoglobulin and T cell receptor chains. These regions are clustered at the distal portion of the V domain and contribute to the antigen-binding site.

ICAM-1 (CD54), ICAM-2 (CD102), and ICAM-3 (CD50) (intercellular adhesion molecules). Cell surface molecules found on a variety of leukocytes and non-hematogenous cells that interact with leukocyte functional antigen-1 (LFA-1).

Iccosomes. Immune complexes in the form of small inclusion bodies found in follicular dendritic cells.

Idiotope. A single antigenic determinant on an antibody V region.

Idiotype. The antigenic characteristic of the V region of an antibody.

IELs (intraepithelial lymphocytes). A population of lymphocytes defined according to location in which γδ T cells are strongly represented.

Immune complex. The product of an antigen–antibody reaction that may also contain components of the complement system.

Immune response (Ir) genes. Genes that affect the level of immune responses. MHC class II genes are very important in controlling responses to specific antigens.

Immunoblotting (Western blotting). A technique for identifying and characterizing proteins using antibodies.

Immunofluorescence. A technique used to identify particular antigens microscopically in tissues or on cells by the binding of a fluorescent antibody conjugate.

Immunogenic. Having the ability to evoke B cell- and/or T cell-mediated immune reactions.

Immunoglobulins. The serum antibodies, including IgG, IgM, IgA, IgE, and IgD.

Immunoglobulin supergene family (IgSF). Molecules that have domains homologous to those seen in immunoglobulins, including MHC class I and II molecules, the T cell receptor, CD2, CD3, CD4, CD8, ICAMs, VCAM, and some of the Fc receptors.

Immunological synapse. The closely apposed region plasma membrane in the interaction between T cells and antigen-presenting cells, centered on interacting TCRs and MHC–peptide complexes.

Induced fit. A description of the way in which an antigen can alter the normal tertiary structure of the binding site on a receptor following binding, by displacing amino acids.

Inducible nitric oxide synthase (iNOS). An enzyme induced by inflammatory cytokines in macrophages that catalyzes the synthesis of nitric oxide (NO).

Inflammation. A series of reactions that bring cells and molecules of the immune system to sites of infection or damage. This appears as an increase in blood supply, increased vascular permeability, and increased transendothelial migration of leukocytes.

Integrins. A large family of cell surface adhesion molecules, some of which interact with cell adhesion molecules (CAMs), others with complement fragments, and others with components of the extracellular matrix.

Interferons (IFNs). A group of molecules involved in signaling between cells of the immune system and in protection against viral infections.

Interleukins (IL-1–IL-22). A group of molecules involved in signaling between cells of the immune system.

Ir gene. A group of immune response (Ir) genes determining the level of an immune response to a particular antigen or foreign stimulus. A number of them are found in the major histocompatibility complex.

Isotype. Refers to genetic variation within a family of proteins or peptides such that every member of the species will have each isotype of the family represented in its genome (e.g. immunoglobulin classes).

ITAMs (immunoreceptor tyrosine activation motifs) and ITIMs (immunoreceptor tyrosine inhibitory motifs). These are target sequences for phosphorylation by kinases involved in cell activation or inhibition.

JAKs (Janus kinases). A group of enzymes with two catalytic domains. They activate by cross-phosphorylation and are particularly involved in signaling from type I and II cytokine receptors.

J chain. A monomorphic polypeptide present in polymeric IgA and IgM, and essential to their formation.

J genes. Sets of gene segments in the immunoglobulin heavy and light chain genes and in the genes for the chains of the T cell receptor, which are recombined during lymphocyte ontogeny and contribute toward the genes for variable domains.

K cells. A group of lymphocytes that are able to destroy their target by antibody-dependent cell-mediated cytotoxicity. They have Fc receptors.

κ (kappa) chains. One of the immunoglobulin light chain isotypes.

Karyotype. The chromosomal constitution of a cell that may vary between individuals of a single species depending on the presence or absence of particular sex chromosomes or on the incidence of translocations between sections of different chromosomes.

Killer immunoglobulin-like receptors (KIRs). Receptors on NK cells that belong to the Ig superfamily, having either two or three extracellular domains. They may either inhibit or activate cytotoxicity, depending on their intracellular domains.

Kinins. A group of vasoactive mediators produced following tissue injury.

Knockout. An animal whose endogenous gene for a particular protein has been deleted or mutated to be non-functional.

Kupffer cells. Phagocytic cells that line the liver sinusoids.

λ (lambda) chains. One of the immunoglobulin light chain isotypes.

Langerhans' cells. Antigen-presenting cells of the skin that emigrate to local lymph nodes to become dendritic cells; they are very active in presenting antigen to T cells.

Large granular lymphocytes (LGLs). A group of morphologically defined lymphocytes containing the majority of K cell and NK cell activity. They have both lymphocyte and monocyte/macrophage markers.

Lectin pathway. A pathway of complement activation initiated by mannan-binding lectin (MBL) that intersects the classical pathway.

Leukotrienes. A collection of metabolites of arachidonic acid that have powerful pharmacological effects.

LFAs (leukocyte functional antigens). A group of three molecules that mediate intercellular adhesion between leukocytes and other cells in an antigen non-specific fashion. LFA-1 is CD11a/CD18, LFA-2 is CD2, and LFA-3 is CD58.

Ligand. A linking (or binding) molecule.

Line. A collection of cells produced by continuous growth of a particular cell culture in vitro. Such a cell line will usually contain a number of individual clones.

Linkage. The condition where two genes are both present in close proximity on a single chromosome and are usually inherited together.

Linkage disequilibrium. A condition where two genes are found together in a population at a greater frequency than that predicted simply by the product of their individual gene frequencies.

LPS (lipopolysaccharide). A product of some Gram-negative bacterial cell walls that can act as a B cell mitogen.

Ly antigens. A group of cell surface markers found on murine T cells that relate to the differentiation of T cell subpopulations. Many are now assigned to the CD system.

Lymphokines. A generic term for molecules other than antibodies that are involved in signaling between cells of the immune system and are produced by lymphocytes (cf. interleukins).

Lymphokine activated killer cells (LAKs). Cytotoxic cells generated ex vivo by stimulation with interleukin-2 (IL-2) and possibly other cytokines.

Lymphotoxins. A group of cytokines, related to tumor necrosis factors, that act on TNF receptors and mediate inflammatory reactions and leukocyte activation.

Lysosomes. Intracellular vesicles containing stored enzymes, adhesion molecules, or toxic molecules depending on the cell type.

Lysozyme. An enzyme secreted by mononuclear phagocytes that hydrolyzes bonds present in bacterial cell walls.

Lytic pathway. The complement pathway effected by components C5–C9 that is responsible for lysis of sensitized cell plasma membranes.

MALT (mucosa-associated lymphoid tissue). Generic term for lymphoid tissue associated with the gastrointestinal tract, bronchial tree, and other mucosas.

Mannose receptor (MR). A lectin-like receptor found on mononuclear phagocytes.

MAP kinases. A group of intracellular enzymes involved in signaling cascades that lead to the activation of transcription factors.

Marginal zone. An area surrounding the splenic white pulp that separates the lymphoid areas from the surrounding red pulp.

Mast cells. Cells found distributed near blood vessels in most tissues. These cells are full of granules containing inflammatory mediators.

Matrix metalloproteases. A group of zinc-containing degradative enzymes that can break down components of the extracellular matrix (e.g. collagenase).

MCPs (macrophage chemotactic proteins). A group of chemokines.

Membrane attack complex (MAC). The assembled terminal complement components C5b–C9 of the lytic pathway that becomes inserted into cell membranes.

Memory cells. Long-lived lymphocytes that have already been primed with their antigen, but have not undergone terminal differentiation into effector cells. They react more readily than naive lymphocytes when restimulated with the same antigen.

MHC (major histocompatibility complex). A genetic region found in all mammals, the products of which are primarily responsible for the rapid rejection of grafts between individuals and function in signaling between lymphocytes and cells expressing antigen.

MHC restriction. A characteristic of many immune reactions in which cells cooperate most effectively with other cells that share a MHC haplotype.

Microglia. Mononuclear phagocytes resident in the brain and spinal cord. Microglial precursors colonize the human central nervous system early in gestation.

MIF (migration inhibition factor). A group of peptides produced by lymphocytes that are capable of inhibiting macrophage migration.

MIIC compartment. An endosomal compartment where MHC class II molecules are loaded with antigenic peptides.

MIPs (macrophage inflammatory proteins). A group of chemokines.

Mitogens. Substances that cause cells, particularly lymphocytes, to undergo cell division.

MLR/MLC (mixed lymphocyte reaction/mixed lymphocyte culture). Assay system for T cell recognition of allogeneic cells in which response is measured by proliferation in the presence of the stimulating cells.

Mononuclear phagocyte system. The lineage of fixed and mobile long-lived phagocytic cells related to blood monocytes and tissue macrophages.

Myeloid cells. The lineages of bone marrow-derived phagocytes, including neutrophils, eosinophils, and monocytes.

Myeloma. A lymphoma produced from cells of the B cell lineage that can invade bone.

N regions. Gene segments present in recombined antigen receptor genes that are not present in the germline DNA.

Neoplasm. A synonym for cancerous tissue.

Neutrophils. Polymorphonuclear granulocytes, which form the major population of blood leukocytes.

NF-κB. A transcription factor that is widely used by different leukocyte populations to signal activation – sometimes called the master-switch of the immune system.

NK (natural killer) cells. A group of lymphocytes that have the intrinsic ability to recognize and destroy some virally infected cells and some tumor cells.

Nucleotide oligomerization domains (NODs). Domains in intracellular receptors for pathogen-associated molecular patterns.

Nude mouse. A genetically athymic mouse that lacks a transcription factor, which is also required for hair production.

Opsonization. A process by which phagocytosis is facilitated by the deposition of opsonins (e.g. antibody and C3b) on the antigen.

PAF (platelet activating factor). A factor released by basophils that causes platelets to aggregate.

PALS (periarteriolar lymphatic sheath). The accumulations of lymphoid tissue constituting the white pulp of the spleen.

Paneth cells. Cells present in the crypts of the small intestine that secrete antimicrobial peptides.

Paracrine. The action of a cytokine on a cell distinct from the cell that produced it.

Passenger cells. Donor leukocytes present in a tissue graft that may sensitize the recipient to the graft.

Patch test. Application of antigen to skin on a patch to test for type IV hypersensitivity reactions.

Pathogen. An organism that causes disease.

Pathogen-associated molecular patterns (PAMPs). Biological macromolecules produced by microbial pathogens that are recognized by receptors on mononuclear phagocytes and some opsonins in serum and tissue fluids.

PC (phosphorylcholine). A commonly used hapten that is also found on the surface of a number of microorganisms.

PCA (passive cutaneous anaphylaxis). The technique used to detect antigen-specific IgE, in which the test animal is injected intravenously with the antigen and dye, the skin having previously been sensitized with IgE antibody.

Pentraxins. A group of acute-phase pentameric molecules present in serum that recognize PAMPs and opsonize bacteria for phagocytosis.

Perforin. A granule-associated molecule of cytotoxic cells, homologous to complement C9. It can form pores on the membrane of a target cell.

Peyer's patches. Collections of lymphoid cells in the wall of the gut that form a secondary lymphoid tissue.

PFC (plaque forming cell). An antibody-producing cell detected in vitro by its ability to lyze antigen-sensitized erythrocytes in the presence of complement.

PHA (phytohemagglutin). A mitogen for T cells.

Phagocytosis. The process by which cells engulf material and enclose it within a vacuole (phagosome) in the cytoplasm.

Phenotype. The expressed characteristics of an individual (cf. genotype).

Plasma cell. An antibody-producing B cell that has reached the end of its differentiation pathway.

Plasmin system. One of the plasma enzyme systems that generates the fibrinolytic enzyme, plasmin, and also contributes to inflammation and tissue remodeling.

Pokeweed mitogen. A mitogen for B and T cells.

Polymorphs. A common acronym for polymorphonuclear leukocytes, including basophils, neutrophils, and eosinophils.

Prick test. Introduction of minute quantities of antigen into the skin to test for type I hypersensitivity.

Primary lymphoid tissues. Lymphoid organs in which lymphocytes complete their initial maturation steps; they include the fetal liver, adult bone marrow, and thymus, and bursa of Fabricius in birds.

Primary response. The immune response (cellular or humoral) following an initial encounter with a particular antigen.

Prime. To induce an initial sensitization to antigen.

Privileged tissues/sites. In the context of transplantation these are tissues that induce weak immune responses or sites of the body that are partly shielded from graft rejection reactions.

Prostaglandins. Pharmacologically active derivatives of arachidonic acid. Different prostaglandins are capable of modulating cell mobility and immune responses.

Proteasomes. Organelles that degrade cellular proteins tagged for breakdown by ubiquitination.

Protein A and protein G. Components of the cell wall of some strains of staphylococci that bind to Fc of most IgG isotypes.

Pseudoalleles. Tandem variants of a gene: they do not occupy a homologous position on the chromosome (e.g. C4).

Pseudogenes. Genes that have homologous structures to other genes, but are incapable of being expressed (e.g. *Jk3* in the mouse).

Radioimmunoassay (RIA). A number of different sensitive techniques for measuring antigen or antibody titers using radiolabeled reagents.

RAG-1 and RAG-2. Recombination activating genes required for recombination of V, D, and J gene segments during generation of functional antigen receptor genes.

Receptor. A cell surface molecule that binds specifically to particular extracellular molecules.

Receptor editing. A process by which immunoglobulin genes can undergo a secondary recombination event, in order to rescue B cells that are producing non-functional or autoreactive antibodies.

Recombination. A process by which genetic information is rearranged during meiosis. This process also occurs during the somatic rearrangements of DNA, which occur in the formation of genes encoding antibody molecules and T cell antigen receptors.

Recurrent idiotype. An idiotype present in the immune response of different animals or strains to a particular antigen.

Relative risk. A number that expresses how much more likely (>1) or less likely (<1) an individual is to develop a particular disease if they possess a particular genotype.

Respiratory burst. Increase in oxidative metabolism of phagocytes following uptake of opsonized particles.

Reticuloendothelial system. A diffuse system of phagocytic cells derived from the bone marrow stem cells that are associated with the connective tissue framework of the liver, spleen, lymph nodes, and other serous cavities. An old-fashioned term that is rarely used – mononuclear phagocyte system is the preferred term.

ROIs/RNIs (reactive oxygen intermediates/reactive nitrogen intermediates). Bactericidal metabolites produced by phagocytic cells, including hydrogen peroxide, hypohalites, and nitric oxide.

Rosetting. A technique for identifying or isolating cells by mixing them with particles or cells to which they bind (e.g. sheep erythrocytes to human T cells). The rosettes consist of a central cell surrounded by bound cells.

Scavenger receptors. A group of receptors that recognize cell debris and are involved in the phagocytosis of apoptotic cells and some pathogens.

SCID (severe combined immunodeficiency). A group of genetic conditions leading to major deficiencies or absence of both B cells and T cells.

Secondary response. The immune response that follows a second or subsequent encounter with a particular antigen.

Secretory component. A polypeptide produced by cells of some secretory epithelia that is involved in transporting secreted polymeric IgA across the cell and protecting it from digestion in the gastrointestinal tract.

Selectins. Three adhesion molecules – P-selectin (CD62P), E-selectin (CD62E), and L-selectin (CD62L) – involved in slowing leukocytes during their transit through venules.

Serotonin. 5-Hydroxytryptamine.

SLE (systemic lupus erythematosus). An autoimmune disease (non-organ specific) of humans usually involving anti-nuclear antibodies.

Somatic mutation. A process occurring during B cell maturation and affecting the antibody gene region that permits refinement of antibody specificity.

Spleen. A major secondary lymphoid organ in the peritoneal cavity next to the stomach.

STATs. A group of proteins that form components of transcription factors following activation by kinases.

Stem cell factor (SCF). Also called Steel factor. A cytokine required for the earliest stages of leukocyte development in bone marrow.

Superantigens. Antigens that stimulate clones of T cells with different antigen specificity but using the same TCR V genes.

Suppressor (Ts) cell. Functionally defined populations of T cells that reduce the immune responses of other T cells or B cells, or switch the response into a different pathway to that under investigation.

Surface plasmon resonance. A biophysical phenomenon that can be used to measure the association and binding constants of proteins in solution interacting with bound ligands.

Synergism. Cooperative interaction.

Syngeneic. Strains of animals produced by repeated inbreeding so that each pair of autosomes within an individual is identical.

T cells. Lymphocytes that differentiate primarily in the thymus and are central to the control and development of immune responses. The principal subgroups are cytotoxic T cells (CTLs) and T helper cells (TH0, TH1, and TH2).

TAP transporters. A group of molecules that transport proteins and peptides between intracellular compartments.

TCR (T cell receptor). The T cell antigen receptor consisting of either an αβ dimer (TCR-2) or a γδ dimer (TCR-1) associated with the CD3 molecular complex.

T-dependent/T-independent antigens. T-dependent antigens require immune recognition by both T and B cells to produce an immune response. T-independent antigens can directly stimulate B cells to produce specific antibody.

TGFs (transforming growth factors). A group of cytokines identified by their ability to promote fibroblast growth; they are also generally immunosuppressive.

Thoracic duct. Drains efferent lymph into the venous system.

Thromboxanes. Products of arachidonic acid metabolism, some of which are involved in inflammation.

Thymus. A primary lymphoid organ in the thoracic cavity over the heart.

Tissue inhibitors of matrix metalloproteases (TIMPs). A group of proteins released by cells in tissues, that limit the activity of matrix metalloproteases.

Tissue typing. Determination of an individual's allotypic variants of MHC molecules.

TNF (tumor necrosis factor). A cytokine released by activated macrophages that is structurally related to lymphotoxin released by activated T cells.

Tolerance. A state of specific immunological unresponsiveness.

Toll-like receptors (TLRs). A group of cell surface receptors that recognize molecules from pathogens (e.g. LPS). Some TLRs transduce signals for inflammation.

Tonsils. Paired lymphoid organs in the throat that form part of the mucosa-associated lymphoid tissue (MALT).

Transformation. Morphological changes in a lymphocyte associated with the onset of division. Also used to denote the change to the autonomously dividing state of a cancer cell.

Transgenic animal. An animal in which one or more new genes have been incorporated. These are often placed under specific promoters so that they are expressed only in particular tissues for limited periods.

Tumor necrosis factors (TNFs). A group of proinflammatory cytokines encoded within the MHC.

V domains. The N terminal domains of antibody heavy and light chains and the α, β, γ, and δ chains of the T cell receptor that become recombined with appropriate sets of D and J genes during lymphocyte ontogeny.

Vaccination. A general term for immunization against infectious disease, originally derived from immunization against smallpox, which uses the vaccinia virus.

Vasoactive amines. Products such as histamine and 5-hydroxytryptamine (serotonin) released by basophils, mast cells, and platelets that act on the endothelium and smooth muscle of the local vasculature.

Veiled cells. Cells of the dendritic cell lineage as seen in afferent lymph. They may be derived from Langerhans' cells or other dendritic cell types.

VLA-1–VLA-6 (very late antigens). The set of integrins that share a common β1 chain (CD29).

Waldeyer's ring. The secondary lymphoid tissues of the nasopharynx that includes the tonsils and adenoids.

Western blotting. A technique for identifying and characterizing proteins using antibodies. Synonymous with immunoblotting.

White pulp. The lymphoid component of spleen consisting of periarteriolar sheaths of lymphocytes and antigen-presenting cells.

Xenogeneic. Refers to interspecies antigenic differences.

Index

A
γABO blood group system, 451–2
acetylcholine receptor
  antibodies, 374
  autoantibodies, 458
acquired C3 deficiency, 103
acquired immune deficiency syndrome *see* AIDS
activation-induced cell death, 352, 353
  molecular basis of, 364
activation-induced cytidine deaminase, 165, 175, 301, 527
active chronic hepatitis, 367
acute phase proteins, 6, 140, 527
acute phase response, 262
acute rejection, 388, 389, 390
adaptive immunity, 8, 14, 127
  role of complement, 99–100
Addison’s disease, 367
  HLA associations, 368
adenosine deaminase deficiency, 302, 303
adhesion molecules, 21, 24–5, 130–7, 527
  induction by cytokines, 137
  leukocyte migration, 136–7
  T cell migration, 494
adjuvant, 161, 262, 330–1, 527
  allergic disease, 444
  Freund’s, 262, 529
adoptive immunotherapy, 417
adult T cell leukemia, 409
adventitial reticular cells, 46
affinity, 60, 68–72, 527
  calculation of, 71
affinity maturation, 84, 165, 218, 527
  antibody feedback, 219
  B cells, 172–5
    cell selection, 174
  dual processes of, 174–5
agonist peptides, 118
AIDS, 316–22, 527
  CD4 count, 319–20
  vaccination, 322
  see also HIV
Aire, 349
  and autoimmune disease, 369
  role in immune tolerance, 350
alemtuzumab, 459
allelic exclusion, 47, 527
allergen-specific immunotherapy, 444
allergens, 26, 425, 527
  and asthma, 441–3
  food proteins, 431
  HLA associations, 436
  IgE binding sites on, 431–2
  inhalant, 430–1
  modified recombinant, 444
  properties, 426, 429–30
  proteins, 425–6
  T cells in immune response, 428–9
allergy, 527
  genetic inheritance, 437
  pathways contributing to, 440–1
  skin tests, 437–9
allogeneic differences, 384
allografts, 344
allopurinol, 313
allorecognition
  direct, 385–6, 387
  indirect, 386–7
alloresponse, 384, 385
allotype, 60, 62, 79, 80, 527
α-adrenergic receptor antibodies, 374
α-fetoprotein, 410
α4β7 integrin, 493
αβ T cell receptors, 111–14
αβ T cells, 106–8
  antigen recognition, 109
alternative pathway, 6, 88, 89, 92–3, 527
  C3 convertase, 92–3
  deficiencies, 102
  regulator deficiencies, 103
Amb a 5 antigen, 429
amplification loop, 89, 527
anaphylactic shock, 296, 501–2
anaphylactoid reactions, 435
anaphylatoxins, 89, 98, 463, 527
anaphylaxis, 423, 425, 527
  immunotherapy, 443–4
  passive cutaneous, 532
ANCAs, 446–7
Ancylostoma duodenale, 278
anergy, 395–6, 527
  clonal, 165, 171, 359
angiedema, 305
antagonist peptides, 118, 527
anterior chamber associated immune deviation, 238
anti-CTLA-4 antibody, 414
anti-idiotype antibodies, 80, 219
  as vaccines, 327, 330
anti-idiotype response, 218
  and tolerance, 362
anti-mitochondrial antibodies, 379
anti-phospholipid syndrome, 367
anti-receptor antibodies, 374
anti-RhD antibody, 218
anti-ribonucleoprotein, 459
anti-TNFα therapy, 380
anti-tumor immune response, suppression of, 413
antibodies, 5, 59–86, 527
  acetylcholine receptors, 374
  as adapter molecules for immune effector systems, 61
  affinity, 68–72
  anti-CTLA-4, 414
  anti-idiotype, 80, 219, 327, 330
  anti-IgE, 444
  anti-IL-5, 445
  anti-mitochondrial, 379
  anti-neutrophil, 456
  anti-RhD, 218
  antibacterial action, 262–5
  antiviral, 249–50
  assays, 70–1
  auto-receptor, 374
  avidity, 60, 68
  cancer therapy, 410, 415
  cardiolipin, 375–6
  classes, 63–7
  co-receptor molecules, 361
  co-stimulatory molecules, 361
  collagen type IV, 457
  complement mobilization, 250
  complement-fixing, 263
  cross-reactivity, 68
  diversity, 80–5
  evasion of, 263–4
  to glomerular capillary basement membrane, 375
  β_2-glycoprotein 1, 375–6
  IgG, 218
  intrinsic factor, 375
  to intrinsic factor, 375
  isotype, 60–2, 72–4, 79, 531
  measurement of production, 168
  monoclonal, 80, 330
  parasitic infections, 286–8
  presynaptic calcium channels, 375
  recombinant, 69
  sodium channel, 375
  sperm, 375
  VP1-a, 493
  see also autoantibodies; immunoglobulins
antibody blocking, 218
antibody feedback, 219
antibody response, 163–80, 495
  cell cooperation, 168
  in HIV, 318

antibody response *(cont'd)*
  primary, 174, 533
  secondary, 174, 533
  to vaccines, 332
antibody-dependent cell-mediated cytotoxicity, 27, 75, 204, 248, 527
  parasitic infection, 287
antibody-dependent enhancement of virus infection, 253
antibody-dependent response, 286–8
antibody-forming cells, 38, 168, 527
antibody-independent response, 286–8
antibody-mediated hypersensitivity, 424, 449–60
  mechanism of damage, 463–4
  tissue targeting, 457–9
  transfusion reactions, 450–3
antibody-mediated immunity, 249–50
antigen binding groove, 112
antigen binding sites, 67–8
  random recombination, 78–80
antigen persistence, 362
antigen presentation, 11, 145–62, 494–5, 527
  genetic restrictions, 121–2
  and immune response, 160–1
  T cells, 145–9
antigen processing, 145–9, 494–5, 527
  cell types, 146–7
antigen receptors, 527
antigen recognition, 14, 59–60, 68–9
  αβ T cells, 109
  extracellular vs intracellular pathogens, 14
  γδ T cells, 109
  innate and adaptive immunity, 14
  lymphocytes, 12–13
  and T cell development, 349
  termination of immune response, 14
antigen-antibody reactions, 10
  reversibility, 68
antigen-presenting cells, 11, 20, 21, 28–9, 145, 148, 262, 527
  and immune response, 216–17
  location of, 146
antigen-processing genes, 152
antigen-specific memory, 226
antigenic determinants *see* epitopes
antigenic peptides, 527
antigenic variation, 252–3, 279
  parasites, 291–2, 294
antigens, 10–12, 527
  aerosol administration, 217
  anti-tumor *see* anti-tumor antigens
  assays, 70–1
  and B cell development, 163–5
  blood group, 454
  cancer/testes, 407
  carcinoembryonic, 405, 409, 410
  cell-mediated immunity, 12
  characterization of, 73, 74
  cloned, 329
  cognate, 169
  cross-reactivity, 68, 71, 358, 370–1, 529
  differentiation, 407–8
  eradication of, 13
  extrinsic, 14
  F4/80, 182
  Fc receptors, 10–11
  histocompatibility, 386, 387, 488
  and immune response, 216
  immunoglobulin class expression, 178–9
  intrinsic, 14
  orally administered, 361–2
  peptides, 11–12
  RhD, 453–4
  route of administration, 216, 361–2
  self, 345
  sequestration by lymphoid organs, 351
  soluble, inducing tolerance, 361
  synthetic, 329
  T-dependent, 165, 167–8, 533
  T-independent, 165, 533
  tissue binding, 472
  tolerance, 216
  tumor, 404–11
  tumor-specific, 127, 402, 407, 408
  as vaccines, 326–30
  very late, 534
  widely expressed, 408
  see also vaccination; vaccines; and individual antigens
antiviral antibodies, 249–50
antiviral proteins, 527
APCs *see* antigen-presenting cells
apoprotein B, 376
apoptosis, 13, 39, 45, 129, 209, 215, 225, 527
  activation-induced cell death, 352, 353, 364
  Fas-independent, 173
  and homeostasis, 227
  role of complement, 102
  T cells, 352–3
  virus effects on, 254
apoptotic bodies, 227
apoptotic cells, clearance of, 188–91
Arthus reaction, 465, 466–7, 527
artificially induced tolerance, 360–3
ascariasis, 278
Ascaris spp., 277
  genetic polymorphisms affecting outcome, 279
Ascaris lumbricoides, 278
Aspergillus spp., 274
Aspergillus fumigatus, monocyte/macrophage killing, 27
asthma, 242, 430–1
  allergens causing, 441–3
  bronchial hyperreactivity, 442
  genetic associations, 435–7
  inflammatory response, 440
  major basic protein, 442
astrocytes, 239
atherosclerosis, 367, 373
atopic disease, 436
atopy, 527
attenuated vaccines, 326–7
  safety, 333
autoantibodies, 296, 358, 459
  to acetylcholine receptors, 458
  to α-adrenergic receptors, 374
  cold-reactive, 455
  desmoglein-1/3, 457
  diagnostic and prognostic value, 377
  induction by cross-reactive antigens, 371
  to insulin receptors, 374
  islet cell, 459
  to neutrophil cytoplasmic antigens *see* ANCAs
  pathogenicity of, 374–5
  in pernicious anemia, 375
  to platelets, 375, 456–7
  to red cells, 375
  to sodium channels, 375
  thyroid, 366
  warm-reactive, 454–5
  see also autoimmune disease; autoimmunity; tolerance
autoantigens, 39, 345
autocrine, 527
autografts, 384
autoimmune atrophic gastritis, 367
autoimmune colloid goiter, 376
autoimmune disease, 365–82, 499–500
  and HLA, 368
  and immune complex disease, 462
  non-organ-specific, 367
  organ-specific, 366, 367
  spectrum of, 366–7
  suppression of, 376
  systemic, 366
  thyroid, 376
  treatment, 377–81
  see also individual conditions
autoimmune hemolytic anemia, 367, 375–6, 454–5, 502–3
autoimmune lymphoproliferative syndrome, 233, 352
  type 2, 211
autoimmune polyglandular syndrome type 1, 369
autoimmunity, 17, 216, 365–82, 499–500, 527
  bypassing development of, 370–3
  CTLA-4 SNP, 368–9
  cytokine/hypothalamic-pituitary-adrenal feedback loop, 372
  experimentally induced, 373–4
  genetic factors, 367–9
  pathogenesis caused by, 373–7
  role of Tregs in, 355–6
  spontaneous, 374
  susceptibility to, 372–3
  virus-induced, 255–6
  see also autoimmune disease
auxiliary cells, 5–6
  inflammatory mediators in, 139–40
avidity, 60, 68, 350, 360, 527
azathioprine, 312–13, 393
azurocidin, 266

B

B cells, 5, 9, 20, 528
  affinity maturation, 172–5
    and cell selection, 174
  autoreactive, 370–1
  class diversity, 177
  deficiency, 299–302
  depletion, 382
  development, 46–8, 80, 164, 346

    antigen dependence, 163–5
    checkpoints, 358–9
    cytokine receptors, 171
    cytokines, 170–2
  elimination by spleen, 359–60
  immunoglobulins as lineage markers, 47
  immunosuppressants affecting
    azathioprine, 313
    ciclosporin, 314
    cyclophosphamide, 312
    leflunomide, 313
    rapamycin, 314
    steroids, 312
    tacrolimus, 314
  marginal zone, 38, 185
  markers, 31, 37–8
  priming of, 99–100
  receptor editing, 359
  in secondary lymphoid tissues, 47–8
  selection, 46–7
  self-reactive, 173, 358
    deletion of, 359, 360
    persistence in normal subjects, 369–70
  spleen, 49
  stimulation, 371
  surface molecules, 47
  switching of immunoglobulin class, 175–9
  T cell signaling defects, 301–2
  tolerance induction, 216, 357–60
    susceptibility to, 362
  virus infection, 255
B cell activation, 158, 168–70
  intracellular signalling, 168
  T-independent antigens, 165–6
B cell antigen receptors, 60
B cell coreceptor complex, 165, 168, 169, 528
B cell maturation, 301
  inhibition by steroids, 312
B cell receptor complex, 37–8, 528
B cell receptors, 163
  somatic hypermutation, 175
B cell-T cell interaction, 171
B lymphoma idiotype, 415–16
B7-1, 147, 155, 514, 527
  blockage of, 395–6
B7-2, 147, 155, 514, 527
  blockage of, 395–6
B220, 163
bacteria, 257–73
  antibody evasion, 263–4
  avoidance of phagocytosis, 268–9
  cell wall, 257–8
  complement evasion, 264–5
  Gram-negative, 198, 257, 260
  Gram-positive, 257
  host defense, 258–61
  mycobacteria, 257
  pathogenicity, 258, 259
  phagocytosis of, 265–8
  spirochetes, 257
bacterial killing, 198, 199–200
bacterial recognition, 261–2
BALT *see* bronchus-associated lymphoid tissue
β barrel, 67
basophils, 6, 25–6, 434, 528
  activation by C3a and C5a, 98
  Fc receptor expression, 74–8
    FcεRI, 426
  glucocorticoid effects, 312
  histamine, 432–3
  IgE on, 62
  inflammatory mediators, 139–40
BCG, 267, 302, 326, 335, 338, 528
Bcl-2, 227, 528
Bcl-2 homolog, 254
BCR-ABL, 405
Bexaar, 410
Biozzi mouse, 231, 528
Birbeck granules, 478, 479
Bla g 2 antigen, 432
Blastomyces spp., 274
blood cells, drug reactions, 455–7
blood groups, 502, 528
  ABO system, 451–2
  antigens, 454
  Duffy system, 451, 452
  Kell system, 451, 452
  Lewis system, 452
  MN system, 451, 452
  Rhesus system, 451, 452
blood-brain barrier, 238–9
bone marrow transplantation, 384
Bordetella pertussis, 272
bradykinin, 139, 140, 528
bronchial hyperreactivity, 442
bronchoalveolar lavage, 441–2
bronchus-associated lymphoid tissue (BALT), 48
Brugia malayi, 278
Burkholderia pseudomallei, 268
Burkitt's lymphoma, 404, 409
bursa of Fabricius, 528
bystander lysis, 89, 95, 528
bystander suppression, 361–2

C
C domains, 528
C genes, 528
c-Kit, 528
C-reactive protein, 6, 141–2, 260
  parasite recognition, 281
C-type lectins, parasite recognition, 281
C1, 91
  deficiency, 101
C1 inhibitor, 93
  deficiency, 102–3, 306–7
C1q, 91, 231
C1q receptors, 96–7, 265
C1qRp, 96
C1r, 91, 231
C1s, 91, 231
C2
  cleavage, 91
  deficiency, 101
C3, 87, 97
  acquired deficiency, 103
  cleaved derivatives, 262
  deficiency, 102, 231
  receptors, 96
C3 convertase, 89, 91, 92–3, 528
  control of, 93–4
C3 nephritic factors, 465
C3a, 88, 91, 140, 463
  basophil activation, 98
  mast cell activation, 98
  receptors, 96
C3b, 88, 89, 92
  disaggregation of immune complexes, 99
  as opsonin, 98–9
  surface binding, 92
C3bBb complex, 92–3
C4, 97
  cleavage, 91
  deficiency, 101, 231, 306
C4 nephritic factors, 465
C4a, 91
C4b, 89, 91
  surface binding, 92
C4b binding protein, 93, 94
C4b2a, 91–2
C4b2a3b, 92
C5, 97
  deficiency, 102
C5 convertase, 92
  control of, 93–4
C5a, 88, 92, 140, 186, 463
  basophil activation, 98
  chemotaxis, 97–8
  mast cell activation, 98
  receptors, 96
C5a protease, 264
C5b, 92
C5b-9 *see* lytic complex
C5b67 complex, 95
C6, 95
  deficiency, 102
C7, 95
  deficiency, 102
C8, 95
  deficiency, 102
C9, 95
  deficiency, 102
C117 *see* c-Kit
Ca 19-9, 410
Ca 125, 410
CA1, 176
cadaveric transplantation, 384
calcineurin, 159, 313, 394
calcitonin, 410
calnexin, 151
calreticulin, 151, 406
Campath, 410
Campath-1H, 379
CAMs *see* cell adhesion molecules
cANCA, 376
cancer detection markers, 410
cancer immunotherapy, 410, 415–18
  animal models, 415
  antibodies, 410, 415
  peptide pulsed dendritic cells, 417
  T cells, 417
  vaccination, 415–17
cancer/testes antigens, 407
cancers, immunity to, 401–19, 501
  fibrosarcoma, 403–4
  immune surveillance hypothesis, 401
  prophylaxis, 403
  protective immunity, 402–3
  tumor immunogenicity, 401–2

cancers, immunity to *(cont'd)*
  tumor-specific antigens, 402
  see also tumors
Candida spp., 258
Candida albicans, 194, 274
  monocyte/macrophage killing, 274
  in SCID, 303
Candida parapsilosis, monocyte/macrophage killing, 274
candidiasis, 274, 321
capping, 528
carcinoembryonic antigen *see* CEA
cardiolipin antibodies, 375–6
carrier effect, 167
carriers, 528
caspase 8, 208, 405
caspase-associated recruitment domains (CARD), 194
caspases, 204, 528
catecholamines, 410
β-catenin, 405
cathelicidins, 143, 266, 528
cathepsin G, 266
cC1qR, 96
CCL1, 524
CCL2, 134, 137, 241, 242, 524
CCL3, 134, 140, 240, 524
CCL4, 134, 524
CCL5, 134, 242, 524
CCL6, 524
CCL7, 524
CCL7,8, 134
CCL8, 524
CCL9, 524
CCL11, 134, 136, 524
CCL13, 134, 524
CCL14, 524
CCL15, 524
CCL16, 524
CCL17, 524
CCL18, 138, 524
CCL19, 138, 524
CCL20, 525
CCL21, 525
CCL22, 525
CCL23, 525
CCL24, 525
CCL25, 525
CCL26, 525
CCL27, 525
CCL28, 525
CCR1, 186, 523
CCR2, 186, 523
CCR3, 136, 227, 523
CCR4, 227, 523
CCR5, 137, 186, 232, 523
CCR6, 523
CCR7, 138, 228, 523
CCR8, 227, 523
CCR9, 523
CCR10, 524
CCR11, 524
CD markers, 176, 510–20, 528
  see also individual markers
CD1, 31, 122–3
CD1a, 509
CD1b, 509
CD1c, 509
CD1d, 509
CD1e, 509
CD2, 147, 159, 509
CD3, 21, 31, 107–8, 509
CD3-, 205
CD4+, 36, 43, 220, 222, 357
  antibacterial and antifungal activity, 267
  antiviral activity, 250–1
  HIV infection, 254–5
  role in infections, 223–4
CD4+CD25+ Tregs, 222, 223, 380
CD4, 21, 31, 267, 509
  antibodies to, 361
  concomitant immunity, 413
  count in AIDS, 319–20
  delayed-type hypersensitivity, 477–8
  HIV infection, 316–17, 319
  HIV-specific, 3189
  lineage commitment, 348
  MHC binding, 156
  parasitic infection, 282–3
CD4-, 42–3
CD5+, 37–8
  activation by T-independent antigens, 167
CD5, 21, 509
CD6, 509
CD7, 509
CD8+, 36, 43, 221
  antiviral activity, 250–1
  immune suppression, 224
CD8, 21, 31, 509
  activation, 269
  antibodies to, 361
  concomitant immunity, 413
  lineage commitment, 348
  MHC binding, 156
  parasitic infection, 282–3
CD8-, 42–3
CD9, 509
CD10, 509
CD11a, 147, 509
CD11a/CD18 *see* lymphocyte functional antigen-1
CD11b, 509
CD11b/CD18, 21
CD11c, 510
CD13, 510
CD14, 24, 261, 510
  initiation of signal transduction, 192–3
CD15, 510
CD16+, 205
CD16 *see* FcγRIII
CD16b, 510
CD18, 159, 510
CD19, 31, 37, 51
CD20, 31, 410, 510
CD21, 37, 510
CD22, 510
CD23, 31, 60, 510
CD24, 510
CD25+, 222, 355–6
  protective tumor immunity, 414–15
CD25, 410, 510
CD26, 510
CD27, 510
CD28, 147, 155, 159, 510
CD29, 510
CD30, 410, 511
CD31, 511
CD32 *see* FcγRII
CD33, 410, 511
CD34, 511
CD35 *see* CR1
CD36, 513
  parasite recognition, 281
CD37, 511
CD38, 511
CD39, 511
CD40, 31, 37, 147, 165, 169, 300, 511
CD40-CD40L interaction, 268
CD40L, 169
CD41, 511
CD42a, 511
CD42b, 511
CD42c, 511
CD42d, 511
CD43, 511
CD44, 511
CD45, 147, 236, 511
CD45RA, 512
CD45RB, 512
CD45RO, 512
CD46, 250, 512
CD47, 512
CD48, 512
CD49a, 512
CD49b, 512
CD49c, 512
CD49d, 512
CD49e, 512
CD49f, 512
CD50 *see* ICAM-3
CD51, 512
CD52, 410, 512
CD53, 512
CD54 *see* ICAM-1
CD55 *see* decay accelerating factor
CD56+, 205
CD56, 27–8, 513
CD57, 513
CD58, 147, 513
CD59, 95, 96, 513
CD61, 513
CD62E *see* E-selectin
CD62L *see* L-selectin
CD62P *see* P-selectin
CD63, 513
CD64 *see* FcγRI
CD65, 513
CD66a, 513
CD66b, 513
CD66c, 513
CD66d, 513
CD66e, 513
CD66f, 513
CD68, 513
CD69, 513
CD70, 514
CD71, 514
CD72, 37, 514
CD73, 514
CD74, 514

CD77, 514
CD79, 21
CD79a, 31, 514
CD79b, 31, 514
CD80 *see* B7-1
CD81, 514
CD82, 514
CD83, 514
CD85, 514
CD86 *see* B7-2
CD87, 514
CD88, 515
CD89, 515
CD90, 515
CD91, 515
CD94+, 205
CD94, 204, 205–6, 515
CD95 *see* Fas
CD96, 515
CD97, 515
CD98, 515
CD99, 515
CD100, 515
CD101, 515
CD102 *see* ICAM-2
CD103, 515
CD104, 515
CD105, 515
CD106, 515
CD107a, 515
CD107b, 515
CD108, 516
CD109 *see* platelet activation factor
CD114, 516
CD115, 516
CD116, 516
CD117, 516
CD118, 516
CD119, 516
CD120a, 516
CD120b, 516
CD121a, 516
CD122, 516
CD123, 516
CD124, 516
CD125, 516
CD126, 516
CD127, 516
CD128, 516
CD130, 516
CD132, 517
CD134, 517
CD135, 517
CD138, 517
CD139, 517
CD140, 517
CD141, 517
CD142, 517
CD143, 517
CD144, 517
CD146, 517
CD147, 517
CD148, 517
CD151, 518
CD152, 518
CD152 *see* CTLA-4
CD153, 518
CD154, 147, 518
  antibodies to, 361
CD155, 518
CD156, 518
CD157, 518
CD158a, 518
CD158b, 518
CD158c, 518
CD161, 518
CD162, 518
CD163, 518
CD164, 518
CD165, 518
CD166, 518
CD167A, 519
CD168, 519
CD169, 519
CD170, 519
CD171, 519
CD172a, 519
CD173, 519
CD174, 519
CD175, 519
CD175s, 519
CD176, 519
CD177, 519
CD178, 519
CD179a, 519
CD179b, 519
CD180, 519
CD183, 519
CD184, 519
CD195, 519
CD200, 519
CD201, 519
CD202b, 520
CD203c, 520
CD204, 520
CD205, 520
CD206, 520
CD207, 520
CD208, 520
CD209, 520
CD210, 520
CD212, 520
CD213a1, 520
CD213a2, 520
CD217, 520
CD220, 520
CD221, 520
CD222, 520
CD223, 520
CD224, 520
CD225, 520
CD226, 520
CD227, 520
CD228, 520
CD229, 521
CD230, 521
CD231, 521
CD232, 521
CD233, 521
CD234, 521
CD235a, 521
CD235b, 521
CD236, 521
CD236R, 521
CD238, 521
CD239, 521
CD240CE, 521
CD240D, 521
CD241, 521
CD242, 521
CD243, 521
CD244, 521
CD245, 521
CD246, 521
CD247, 521
CDK4*, 405
CDRs, 106–7, 348, 528
CDw12, 512
CDw17, 512
CDw60, 513
CDw75, 514
CDw76, 514
CDw92, 515
CDw93, 515
CDw108, 516
CDw121b, 516
CDw131, 517
CDw136, 517
CDw137, 517
CDw145, 517
CDw149, 518
CDw150, 518
CDw197, 519
CEA, 405, 409, 410
celiac disease, 241
cell adhesion molecules, 134–6, 528
cell killing *see* apoptosis
cell lines, 531
cell populations
  identification of, 32–3
  isolation of, 34–5
cell senescence, 224–7
cell-mediated cytotoxicity, 203–13
  as immune defense mechanism, 203–5
  see also cytotoxic T lymphocytes
cell-mediated immunity, 8, 12, 60, 105, 528
  protein-energy malnutrition, 314–15
cell-mediated reactions see delayed-type hypersensitivity
central memory cells, 228
central tolerance, 345–51, 352, 395, 528
  mechanisms of, 352
centroblasts, 53
centrocytes, 38, 53
cervical cancer, 404, 409
cFLIP, 210
CFU, 23, 24
CFU-GEMMs, 23
CG1, 176
CG3, 176
Chagas' disease, 278
checkpoints, 350–1, 358–9
chemokine receptors, 134, 524–5
  see also individual receptors
chemokines, 7, 15, 23, 128, 182, 196, 524, 528
  effect on macrophages, 186
  gene polymorphism, 231–2
  leukocyte migration, 136–8
  leukocyte triggering, 133–4

chemokines *(cont'd)*
  parasitic infection, 283–4
  production
    endothelium, 242
    NK cells, 224
  regulation of lymphocyte migration, 227–8
  role in inflammation, 130–7
  virus effects on, 253, 254
chemokinesis, 528
chemotaxis, 15, 88, 97–8, 265, 528
chicken, Obese, 369, 370, 372–3
chimerism, 360–1
Chlamydia trachomatis, 268
cholera vaccine, 329, 335
chromomycosis, 274
  T cell immunity, 275
chromosome 14, 429
chronic granulomatous disease, 266, 307
chronic rejection, 388, 389–90, 391
chymase, 435
ciclosporin, 159, 313–14, 393, 528
class switching, 53, 165, 171, 528
  differential splicing of mRNA, 177, 178
  gene recombination, 177
  immunoglobulins, 175–9
classical pathway, 6, 88, 89–92, 528
  activation, 89, 91
  C3 convertase, 91
  C5 convertase, 92
  control by C1 inhibitor, 93
  deficiencies in, 101–2
CLIP, 153
clonal anergy, 165, 171, 359
clonal deletion, 358
clonal selection, 12, 47, 344, 528
Clostridium difficile, 258
Clostridium perfringens vaccine, 329
Clostridium tetani, 259
clusterin, 95
CM, 176
CNS
  IgG in, 238
  type 4 hypersensitivity, 241–2
co-receptor molecules, 155
  antibodies to, 361
co-stimulation, 529
  blockade, 396
co-stimulatory molecules, 145, 158–9
  antibodies to, 361
  B and T cell interaction, 168–9
  T cell activation, 154–6
Coccidioides spp., 274
cognate antigen, 169
cold-reactive autoantibodies, 455
colicins, 258
collagen receptors, 137
collagen type IV antibodies, 457
collectins, 102, 140, 141, 528
  parasite recognition, 281, 282
colony stimulating factors, 7, 182, 523, 529
colony-forming units *see* CFU
commensals, 258–9
common acute lymphoblastic leukemia antigen, 410
common variable immunodeficiency, 300, 301–2
complement, 6, 13, 87–104, 528
  activation, 75, 261–2
  activation pathways, 87–93
    alternative, 6, 88, 89, 92–3, 261–2
    classical *see* classical pathway
    lectin *see* lectin pathway
    terminal *see* terminal pathway
  adaptive immunity, 99–100
  antiviral effects, 249–50
  autoantibodies, 465
  bacterial evasion, 264–5
  DAF limiting, 93, 291, 451
  deficiency, 100–3, 468–9, 494
  depletion, 468
  fixation, 455
  functions, 97–100
  genetic deficiency, 305–7
  homeostatic roles, 100–3
  host protection, 93–5
  immune complex disease, 464–5
  inflammatory reactions, 87, 88
  innate immunity, 92–3
  parasite evasion, 291
  solubilization of immune complexes, 468, 469
  virus effects on, 254
complement control protein (CCP) domains, 528
complement receptors, 89, 97, 529
  expression, 192
  parasite recognition, 281, 282
  see also individual receptors
complement-fixing antibodies, 263
complementarity determining regions *see* CDRs
concanavalin A, 529
concomitant immunity, 412, 413
congenic, 345, 529
conjugate, 529
conjunctivitis, 424
connective tissue mast cells, 433, 434, 435
constant region genes, 176
constant regions, 529
contact hypersensitivity, 478–80, 482, 529
  elicitation, 479–80, 481
  sensitization, 479–80
continuous epitopes, 72
Coombs' test, 454, 455
copper deficiency, 315
cornea transplantation, 384
corneal allografts, 237
corticosteroids
  allergic disease, 441
  as immune modulators, 311–12
Corynebacterium diphtheriae, 258, 259
  immunity, 272
Coxsackie B virus-induced myocarditis, 256
CpG motifs, 261
CR1, 94–5, 96, 467, 513
  parasite recognition, 281
CR2, 96, 219
CR3, 96, 135, 137, 193
  gene defect, 307–8
  parasite recognition, 281
CR4, 96, 135
Crohn's disease, 195, 196, 241, 490
cross-linking, 433–4
cross-matching, 453
cross-presentation, 147, 154, 269
cross-priming, 269
cross-reacting idiotypes, 80
cross-reactivity, 68, 71, 358, 370–1, 529
  among tumors, 403–4
Cryptococcus neoformans, 266, 274, 321
  evasion strategies, 275
  monocyte/macrophage killing, 27
cryptogenic cirrhosis, 367
CSFs *see* colony stimulating factors
CTLA-4, 147, 155, 159–60, 353, 529
  and autoimmunity, 368–9
  inhibition of T cell activity, 414
CTLA-4-Ig, 395–6
CTLs, 5, 12, 203, 235, 496, 529
  antigen recognition, 203–4
  epitopes, 407
  granules of, 208
  HIV-specific, 319
  killing by, 210–11, 269–70
  signaling by, 208
  and tissue damage, 253–4
  viral infection, 204, 249
CX3CL1, 525
CX3CR1, 186, 524
CXC1, 524
CXCL1, 134, 525
CXCL2, 134, 525
CXCL3, 134, 525
CXCL4, 525
CXCL5, 525
CXCL6, 525
CXCL7, 525
CXCL8, 134, 136, 240, 242, 525
CXCL9, 134, 525
CXCL10, 134, 137, 242, 525
CXCL11, 134, 525
CXCL12, 525
CXCL13, 525
CXCL14, 525
CXCL16, 525
CXCR1, 136, 186
CXCR2, 136, 186, 524
CXCR3, 137, 186, 227, 524
CXCR4, 138, 524
CXCR5, 227, 524
cyclophilin, 394
cyclophosphamide, 312
cytoadhesins, 135
cytokine activated killer cells, 205
cytokine receptors, and B cell development, 171
cytokines, 6–7, 182, 522–3, 529
  and antibody affinity, 171
  contact hypersensitivity, 482
  control of B cell development, 170–2
  control of leukocyte movement, 129
  dysregulation, and autoimmunity, 371–2
  and endotoxic shock, 270, 271
  gene polymorphism, 231–2
  in immune response to worms, 285–6
  isotype regulation, 178–9
  ligands, 356
  macrophage-activating, 265
  and macrophages
    gene expression, 185

    phenotype, 187–8
  parasitic infection, 283–4
  regulation of IgE production, 428–9
  release from keratinocytes, 479
  and superantigen toxicity, 272–3
  synthesis
    by IELs, 241
    by NK cells, 224
    inhibition by steroids, 312
  T cell division, 159
  virus effects on, 254
  see also chemokines; and individual cytokines
cytomegalovirus
  in AIDS, 321–2
  murine, 248
cytotoxic T lymphocytes *see* CTLs

D

D genes, 80, 107, 529
DAF *see* decay accelerating factor
danger hypothesis, 161
DAP-12, 193, 208
DC-SIGN, 191, 192, 317
DC2 dendritic cells, 249
decay accelerating factor, 89, 94, 250, 452, 512, 529
  limitation of complement reaction, 93, 291, 451
decay acceleration, 93–4
dectin-1, 192, 274
defensins, 143, 198–9, 529
  bacterial infection, 266
  fungal infection, 274–5
degranulation, 529
delayed-type hypersensitivity, 270, 477–91, 503–4, 529
  cellular reactions in, 485–6
  CNS, 241–2
  dermis, 242
dendritic cells, 21, 28, 50, 145, 262, 529
  antiviral activity, 249
  follicular, 28, 29, 30, 172, 175, 218, 529
  graft sensitization, 390
  in immune defense, 182
  peptide pulsed, 417
  in peripheral tolerance, 353
  plasmacytoid, 354
  spleen, 49, 53
  thymus, 354
Dengue shock syndrome, 235
Dengue virus, 253
density-gradient separation, 34
Der p 2 antigen, 429, 430, 432
dermatomyositis, 367
Dermatophagoides pteronyssimus, 216
desmoglein-1 autoantibodies, 457
desmoglein-3 autoantibodies, 457
dextran, 165
diabetes mellitus, insulin-dependent, 367
  HLA associations, 368
diapedesis, 23, 184
differentiation antigens, 407–8
DiGeorge syndrome, 51, 302, 304
dihydro-orotate dehydrogenase, 313
Dipetalonema perstans, 278
diphtheria toxoid, 329, 334
direct allorecognition, 375–6, 387
direct antiglobulin test, 454, 455
direct immunofluorescence, 32
discoid lupus erythematosus, 367
discontinuous (conformational) epitopes, 72
dislocation, 66, 73
DM molecules, 529
DNA
  naked, 339
  viruses, 253
domain, 529
dominant idiotypes, 529
donor-specific tolerance, 394–7
double-negative cells, 345
double-positive cells, 346
DRB1*1302, 231
drug-induced hypersensitivity, 455–7
dsRNA, 247
Duffy blood group system, 451, 452
dust mite allergen, 431, 432

E

E-cadherin, 241
E-selectin, 136, 513
αEβ7-integrin, 241
effector cells, 529
effector memory cells, 228
effector T cells, 357
EGFR, 410
eicosanoids, 196, 529
  effect on macrophages, 186
electrostatic bonds, 68
ELF2m, 405
ELISA, 70
ELISPOT assay, 33, 168
emetine, 495
ENDO 180, 191
endocytosis, 427, 529
endosomes, 427
endothelium, 241–2, 529
  chemokine production, 242
endotoxic shock, 270, 271
endotoxin, 270, 529
  see also lipopolysaccharide
enhancement, 529
Entameba histolytica, 292–3
enterocytes, 240–1
entry inhibitors, 322
enzyme-linked immunosorbent assay (ELISA), 70
eosinophil cationic protein, 25, 212, 440
eosinophil peroxidase, 212
eosinophil-derived neurotoxin, 25, 212
eosinophils, 5, 25, 529
  adhesion, 445
  antiparasitic activity, 288, 290
  degranulation, 212–13, 445
  Fc receptor expression, 74–8
  glucocorticoid effects, 312
  major basic protein *see* major basic protein
epidermal cell growth factor, 241
epidermal spongiosis, 438–9
epithelial anlage, 40
epithelioid cell granuloma, 483, 487
epithelioid cells, 484, 529
epitopes, 10, 14, 60, 68, 344, 529
  continuous, 72
  discontinuous (conformational), 72
  T cell epitopes of viral antigens, 408
Epstein-Barr virus, 251, 301, 404, 409, 529
  direct stimulation of autoreactive cells, 371
  and X-linked proliferative syndrome, 304
Erbitux, 410
Erk2, 405
ERp57, 151
Escherichia coli, 196, 264, 266
exocytosis, 450–1
exons, 80
experimental allergic encephalomyelitis, 216
extracellular matrix, effect on macrophages, 186
extracellular pathogens, 14
extrathymic development, 46
extrinsic allergic alveolitis, 462, 463
eye, immune privilege, 236–9

F

f-met-leu-phe, 187, 265
f-met-leu-phe receptor, 134
F4/80 antigen, 182
Fab, 529
factor B, 529
factor D, 529
factor H, 93, 94, 97, 529
factor H-like 1, 93
factor P, 529
FADD, 208, 209
familial hemophagocytic lymphohistiocytosis, 211
familial Mediterranean fever, 195
farmer's lung, 462
Fas, 204, 209, 227, 352, 514, 529
  role in peripheral tolerance, 352–3
Fas ligand, 208
Fas-independent apoptosis, 173
Fc, 529
Fc receptors, 10, 74–8, 207, 424, 529
  antibody-mediated hypersensitivity, 449–50
  IgE, 78
  IgG, 75–6, 77
  immune complex disease, 466
Fc region, 10–11
FcεRI, 78, 426
  cross-linking, 433–4
  histamine release, 78
FcεRII, 78, 426
FcγRI, 60, 75, 265, 513
  phagocytosis, 76
FcγRII, 60, 75, 265, 513
  subtypes, 76
FcγRIIb, 218, 294
FcγRIII, 21, 27–8, 60, 75, 265, 512
  genetic polymorphism, 77
  subtypes, 76–7
Fel d 1 antigen, 429
fetus, immunoglobulins, 76
fibrin breakdown products, 140
fibrinopeptides, 140
fibronectin, 265

fibronectin receptors, 137
fibrosarcoma, 403–4
ficolins, 140, 141, 529
Ficoll, 165
fish, lymphocyte-like cells, 127
FK binding proteins, 313
flagellin, 261, 262
FLICE inhibitory protein, 352–3
FLIP, 352–3
flow cytometry, 529
flukes, 278
fluorescence-activated cell sorter, 32, 529
fMLP, 186
folate deficiency, 316
follicle center cells, 38
follicular dendritic cells, 28, 29, 30, 172, 175, 218, 529
  CR2 expression, 96
  spleen, 49, 53
food allergy, 431
foot and mouth disease, 252
forbidden clones, 344
formyl-methionyl peptides, 529
Foxp3, 222, 356
  selective expression, 223
fractalkaline *see* CX3CR1
framework segments, 529
fratricidal cell death, 352
Freund's adjuvant, 262, 529
frustrated phagocytosis, 200, 212, 450, 530
FTY 720, 394
fungi, 273–6
  categories of, 273–4
  evasion strategies, 275

G

G protein-coupled receptors, 193–4
G-proteins, 133
G250, 405
GAD65, 369
GALT *see* gut-associated lymphoid tissue
γδ cells, 530
γδ T cells, 21, 36, 107–9
  antigen recognition, 109
gene gun, 339
gene recombination, 177
gene therapy, 304
genetic association, 530
genetic factors, 229–33
  allergy, 437
  antigen presentation, 121–2
  asthma, 435–7
  autoimmunity, 367–9
  complement deficiency, 305–7
  immune system, 229–33
  phagocyte defects, 307–8
genetic polymorphisms
  chemokines, 231–2
  cytokines, 231–2
  FcγRIII, 77
  MHC, 120–1
  parasitic infection, 279
genetic restriction, 107, 121–2, 530
genome, 530
genotype, 530
germinal centers, 52–4, 530
  B cell affinity maturation in, 172–5
  organization of, 173
  zoning of, 173
germline, 530
giant cells, 484, 530
Giardia spp., 277
Giardia lamblia, 302
glial cells, 239
glomerular capillary basement membrane antibodies, 375
glomerulonephritis, 463
glucocorticoid-induced TNF-related receptor, 223
glucocorticoids, 311
  effects on circulating leukocytes, 312
β_2-glycoprotein 1 antibodies, 375–6
glycosylation in receptor binding, 77–8
glycosylphosphatidylinositol linkage, 530
Golgi complex, 152
Goodpasture antigen, 457
Goodpasture's syndrome, 367, 375, 457
  animal models, 457
gp70, 405
gp96, 405
gp100, 405
gp120, 317
graft rejection, 385–90
  acute, 388, 389, 390
  chronic, 388, 389–90, 391
  hyperacute, 388–9
  immune privilege, 396–7
  immunological memory, 391
  prevention by HLA matching, 390–2
  time of, 387–8
graft versus host disease, 304, 387, 388, 417, 418, 530
graft versus host reactions, 385
Gram-negative bacteria, 198, 257
  lipopolysaccharide, 260
Gram-positive bacteria, 257
granule proteases, 433
granule-associated killing, 211
granulocyte colony stimulating factor, 336
granulocyte-macrophage colony stimulating factor, 184, 265
granulocytes, 19–20, 22, 23, 530, 532
granuloma, 268
  tuberculoid, 188
granulomatous hypersensitivity, 478, 483–5, 530
  in chronic disease, 486–90
granulysin, 269
granzymes, 204, 530
  in CTLs, 208
  substrate cleavage, 209–10
granzyme A, 209
granzyme B, 209
granzyme C, 209
Graves' disease *see* thyrotoxicosis
growth factors, effect on macrophages, 186
guanosine diphosphate, 187
guanosine triphosphate, 187
Guillain-Barré syndrome, 375
gut
  crypts, 240, 241
  enterocytes, 240–1
  immune responses, 239–41, 496
  villi, 240, 241
gut-associated lymphoid tissue (GALT), 48, 530

H

H-2, 530
H-Y system, 387
HAART therapy, 322
Haemophilus influenzae see influenza virus
Hageman factor, 139, 306
haplotype, 80, 530
hapten-carrier effect, 338
haptens, 167, 478
Hashimoto's persistent goiter, 376
Hashimoto's thyroiditis, 365–6, 367, 376
hashitoxicosis, 376
Hassall's corpuscles, 41
hayfever, 430–1
  immunotherapy, 443–4
heart transplantation, 384
heart-lung transplantation, 384
heat-shock proteins see HSPs
heavy chains see immunoglobulin heavy chains
Helicobacter pylori, 261, 409
Heligmosomoides polygyrus, 290
helminths, 278
  IgE antibodies, 445–6
  immunity, 497
helper T cells *see* TH cells
hemagglutination, 530
hemagglutinin, 253
hematopoietic stem cells, 19
hemolytic disease of newborn, 452, 453–4, 502
hepatitis A vaccine, 327, 335
hepatitis B vaccine, 329, 335
HER2, 405
HER2/Neu receptor, 410
Herceptin, 410
herd immunity, 326
hereditary angioedema, 102–3
hereditary angioneurotic edema, 306
hereditary ataxia telangiectasia, 302, 304–5
herpes simplex virus, 249, 251
herpesvirus, 249
heterologous, 530
hiding
  intracellular pathogens, 268–9
  parasites, 292–3
high endothelial venules, 50, 56–7, 138, 241, 530
highly active antiretroviral therapy *see* HAART therapy
hinge, 530
histamine, 140, 262, 530
  mast cells and basophils, 432–3
histocompatibility, 107, 530
histocompatibility antigens, 386
  minor, 387, 488
Histoplasma spp., 274
Histoplasma capsulatum, evasion strategies, 275
HIV, 316, 530
  antigenic variation, 252–3
  CD4 antigen, 316–17

  cellular immune response, 318–19
  ineffective antibody response, 318
  infection of CD4+ cells, 254–5
  life cycle, 317
  mechanism of immune dysfunction, 318
  natural history, 320
  see also AIDS
HLA, 107, 111, 530
  allergen associations, 436
  and autoimmune disease, 368
  haplotypes, 230–1
  specificities, 507–9
HLA matching, 390–2
HLA-A*02, 231
HLA-A, 118
HLA-A2-R170I, 405
HLA-B*35, 231
HLA-B, 118
HLA-C, 118, 206
HLA-Cw*04, 231
HLA-D, 119–20
HLA-DM, 153, 154
HLA-DO, 154
HLA-DRB1*1302, 231
HLA-E, 118, 152, 204
HLA-F, 118
HLA-G, 118, 204
  inhibition of NK cell action, 206
  resistance to NK cells, 207
HLA-H, 118
homologous restriction factors, 530
hookworm, 277, 278
host defense peptides, 266
HSPs, 273, 368, 404, 406
  binding to APCs, 406–7
  direct isolation from tumors, 404–5
HSP-chaperoned tumor antigenic peptides, 405–6
HSP-peptide complexes
  cancer immunotherapy, 417
  CTL epitopes isolated from, 407
  specific immunogenicity, 408
HSP40, 406
HSP47, 406
HSP60, 373, 376, 406
HSP70, 404, 405, 406
HSP70-2m, 405
HSP90, 404, 406
HSP100, 406
HSP110, 405
HSP170, 404
human chorionic gonadotropin, 410
human herpes virus 8, 320
human immunodeficiency virus *see* HIV
human leukocyte antigen *see* HLA
human T leukemia virus-I, 409
humanized monoclonal anti-IgE, 444
humanized monoclonal anti-IL-5, 445
humoral immunity, 8, 105, 530
hybridoma, 530
hydrogen bonds, 68
hydrophobic forces, 68
5-hydroxytryptamine, 140, 530
hygiene hypothesis, 273, 295
hyper-IgM immunodeficiency, 300, 301, 497–8
hyperacute graft rejection, 388–9, 453
hypersensitivity, 17, 216, 530
  antibody-mediated (type II), 424, 449–60
    mechanism of damage, 463–4
    tissue targeting, 457–9
    transfusion reactions, 450–3
  classification of, 423–4
  contact, 478–80, 529
  delayed-type (type IV), 241–2, 270, 477–91, 503–4, 529
    cellular reactions in, 485–6
    CNS, 241–2
    dermis, 242
  drug-induced, 455–7
  historical aspects, 424–5
  immediate (type 1), 423–47
  immune complex disease (type III), 424, 461–76
  tuberculin-type, 478, 480–3
hypervariable region, 63, 64, 530
hypogammaglobulinemia, transient of infancy, 300, 302

I

iC3b, 98–9
ICAM-1, 135, 136, 137, 147, 154, 480, 512, 530
ICAM-2, 135, 136, 137, 515, 530
ICAM-3, 512, 530
iccosomes, 29, 30, 165, 168, 530
idiopathic leukopenia, 367
idiopathic thrombocytopenic purpura, 367, 375–6, 456–7
idiotope, 219, 530
idiotype, 60, 79, 80, 219, 530
  B lymphoma, 415–16
  dominant, 529
  private, 80
  public, 80
  recurrent, 80, 533
IELs *see* intraepithelial lymphocytes
IFN *see* interferon
IgA, 61, 66, 74
  antiviral activity, 249, 252
  biological properties, 75
  deficiency, 300
  J chain, 66
IgD, 62, 74, 175
  biological properties, 75
  structure, 66–7
IgE, 62, 74, 426–8
  antibodies, 445–6
  and atopic disease, 436
  binding sites on allergens, 431–2
  biological properties, 75
  Fc receptors, 78
  half-life, 427
  IL-4 dependence, 429
  production
    cytokines, 428–9
    TH2 cells, 428
  role in human defense, 288
  structure, 67, 427
IgG, 61, 73–4, 175
  antibody, 218, 263
  antibody-mediated hypersensitivity, 449
  biological properties, 75
  CNS, 238
  Fc receptors, 75–6, 77
  glycosylation in receptor binding, 77–8
  hemolytic disease of newborn, 453–4
  rheumatoid factors, 378
  structure, 63, 64, 65
  subclass, 64
    deficiency, 300
IgG4
  IL-4 dependence, 429
  placental transfer, 427–8
IgM, 61, 73, 175
  antibody, 263
  antibody-mediated hypersensitivity, 449
  biological properties, 75
  conformation, 64–6
  hyper-IgM immunodeficiency, 300, 301, 497–8
  J chain, 64, 65
IL *see* interleukin
immediate hypersensitivity, 423–47
immune activation, 187
immune adherence, 99
immune complex disease, 305, 376–7, 424, 461–76
  and autoimmune disease, 462
  categories of, 462
  complement in, 464–5
  increased vascular permeability, 464
  inflammatory processes, 462–5
  inhalation of antigen, 462
  mechanism of damage, 463–4
immune complexes, 218–19, 530
  antigen tissue binding, 472
  clearance, 305–6, 469
    carbohydrates on antibodies, 470
    and immunoglobulin class, 470
  complement depletion, 468
  deposition
    and immunoglobulin glass, 472–3
    and size, 472
  regulatory effects, 219
  size of, 469, 472
  solubilization by complement, 468, 469
  and tissue damage, 253–4, 470–1
  transport and removal, 467–71
    mononuclear phagocytes, 470
  visualization of, 473–4
  see also immune complex disease
immune deviation, 344
immune distraction, 294
immune enhancement, 311
immune homeostasis
  prevention of lymphoproliferative response, 355
  role of CTLA-4 in, 353–4
immune privilege, 236–9, 351–2
  and graft rejection, 396–7
  see also privileged sites
immune reconstitution inflammatory syndrome, 272
immune response, 7–9
  allospecific, 384
  antigen presentation, 160–1
  excessive, 271–2
    control by tolerance, 362–3
  and graft rejection, 385–7
  gut, 239–41

- immune response *(cont'd)*
  - immunoglobulins affecting, 217–20
  - inconvenience of, 17
  - innate immunity, 8–9
  - lung, 239–41
  - neuroendocrine modulation, 229
  - pathogen dependence, 7–8
  - regulation, 215–34, 496
    - T cells, 220–4
  - role of APCs in, 216–17
  - skin, 239–41
  - termination of, 14
  - in tissues, 235–43
  - to viral infections, 249–52
- immune response genes, 530
- immune surveillance hypothesis, 401
- immune system, 3–18
  - activation, 260–1
    - lipopolysaccharide, 260, 261
  - antigen eradication, 13
  - antigen recognition
    - extracellular vs intracellular pathogens, 14
    - innate and adaptive immunity, 14
    - lymphocytes, 12–13
    - termination of immune response, 14
  - antigens, 10–12
    - antigen-antibody reactions, 10
    - cell-mediated immunity, 12
    - Fc receptors, 10–11
    - peptides, 11–12
  - cells, 4–6, 19–58
    - auxiliary, 5–6
    - B cells and T cells see B cells; T cells
    - cytotoxic cells, 5
    - phagocytes see phagocytes; phagocytosis
  - failure of, 16–17
  - function of, 3–4
  - genetic factors affecting, 229–33
  - phagocytes and lymphocytes, 9
  - soluble mediators, 6–7
    - complement proteins, 6
    - cytokines, 6–7
  - specificity and memory, 10
  - virus infection of cells, 254–5
  - virus reactions, 496
- immune-privileged sites, 228
- immunization, 412–13
- immunoassay, 70
- immunoblotting, 73, 530
- immunodeficiency, 17
  - common variable, 300, 301–2
  - HIV, 316, 530
    - antigenic variation, 252–3
    - CD4 antigen, 316–17
    - cellular immune response, 318–19
    - ineffective antibody response, 318
    - infection of CD4+ cells, 254–5
    - life cycle, 317
    - mechanism of immune dysfunction, 318
    - natural history, 320
  - hyper-IgM, 300, 497–8
  - primary, 299–309
    - B cell deficiencies, 299–302
    - genetic defects of phagocytes, 307–8
    - genetic deficiency of complement proteins, 305–7
    - T cell deficiencies, 302–5
  - secondary, 311–24, 498–9
  - severe combined, 231, 232, 300, 302–3, 533
    - thymus in, 303
    - X-linked, 300
  - and tumor viruses, 404
- immunofluorescence, 32, 530
- immunogenicity, 530
- immunogens, 10
- immunoglobulin fold, 62
- immunoglobulin superfamily, 33, 67
- immunoglobulin-like transcripts, 207
- immunoglobulins, 9, 530
  - as B cell lineage markers, 47
  - biological properties, 75
  - class switching, 175–9
  - in fetus and neonate, 76
  - gene rearrangement, 80–3
  - heavy chains, 62
    - class switching, 176–9
    - recombination, 80–3
  - and immune response, 217–20
  - immunosuppressants affecting, methotrexate, 313
  - isotypes, 60–2
    - regulation by cytokines, 178–9
  - light chains, 62, 63
    - recombination, 80–3
  - membrane-bound, 177–8, 179
  - physicochemical properties, 62
  - recognition/binding of antigens, 59–60
  - secreted, 177–8, 179
  - structure, 62–3, 79
  - subclasses, 60, 61
  - *see also Ig types*
- immunological memory, 10, 173–4, 391, 493
- immunological synapse, 156–7, 530
- immunological tolerance *see* tolerance
- immunomagnetic beads, 35
- immunopathogenicity, 259
- immunophilins, 313
- immunoprecipitation, 73, 74
- immunoreceptor tyrosine-based activation motifs *see* ITAMs
- immunoreceptor tyrosine-based inhibitory motifs *see* ITIMs
- immunosuppressants, 311–24, 392–4
  - non-specific, 336, 338
  - *see also individual drugs*
- immunotherapy
  - allergen-specific, 444
  - anaphylactic sensitivity, 443–4
  - cancer, 410, 415–18
    - animal models, 415
    - antibodies, 410–15
    - peptide pulsed dendritic cells, 417
    - T cells, 417
    - vaccination, 415–17
  - hayfever, 443–4
- indirect allorecognition, 386–7
- indirect complement amplified, 32
- indirect immunofluorescence, 32
- induced fit, 530
- inducible nitric oxide synthase, 199, 531
- infection
  - anti-cancer effect, 410
  - bacterial, 257–73
  - CD4+ in, 223–4
  - fungal, 273–6
  - hygiene hypothesis, 273, 295
  - immunoendocrine reactions, 486–7
  - and malnutrition, 314
  - MHC-linked genes in, 230
  - mycobacterial, 233
  - non-MHC-linked genes affecting, 231–3
  - opportunistic, 299
  - parasitic, 277–97
  - pyogenic, 299
  - viral *see* viruses
- inflammasomes, 194
- inflammation, 15–16, 21, 128–30, 495, 531
  - acute, 128
  - causes
    - classical pathway deficiencies, 101–2
    - complement receptor mediation, 96
  - chronic, 128
  - complement deficiency, 305–6
  - factors involved in
    - auxiliary cells, 5–6
    - chemokines in, 130–7
    - complement, 87, 88
    - leukocytes, 15–16, 128, 130–7
    - platelets, 26–7
  - and parasite elimination, 294–5
  - plasma enzyme systems modulating, 138–40
- inflammatory cells, 286
- inflammatory mediators, 140
  - *see also individual components*
- inflammatory response, 128
- influenza virus
  - antigenic variation, 253
  - vaccine, 329, 334, 335
- inhalant allergens, 430–1
- innate immunity, 8, 14, 127–44
  - alternative and lectin pathways, 92–3
  - fungal infection, 274–5
  - microbial proteins, 143–4
  - parasitic infection, 280–2
- instructional selection, 348
- insulin receptor antibodies, 374
- integrins, 33, 130, 134–6, 531
- intercellular adhesion molecule-1 (ICAM-1), 193
- intercellular adhesion molecules (ICAMs), 21, 155
- interdigitating cells, 28–9, 50, 52
- interferons, 6–7, 182, 531
  - anti-viral activity, 247–8, 249
  - delay of viral spread, 129
  - virus effects on, 254
- interferon-α, 247, 523
- interferon-α/β, 224
- interferon-β, 247, 523
- interferon-γ, 150, 197, 224, 235, 247, 262, 267, 523
  - granuloma formation, 485
- interferon-τ/ε, 247
- interleukins, 7, 182, 531

interleukin-1, 129, 197, 240, 260, 264, 522
interleukin-2, 522
  receptors, 157
  T cell division, 160
interleukin-3, 522
interleukin-4, 165, 227, 429, 522
  B cell development, 170
  recombinant soluble, 444–5
interleukin-5, 522
  B cell development, 170
interleukin-6, 129, 197, 240, 522
  B cell development, 170
interleukin-7, 240, 493, 522
  B cell development, 170
interleukin-8, 23, 140, 522
interleukin-9, 522
interleukin-10, 355–6, 522
  B cell development, 170
interleukin-11, 522
interleukin-12, 224, 227, 262, 522
  regulation of TH1 cell response, 221
interleukin-13, 522
  B cell development, 170
interleukin-14, 522
interleukin-15, 224, 248, 522
interleukin-16, 523
interleukin-17, 523
interleukin-18, 224, 262, 523
interleukin-21, 523
interleukin-22, 523
intracellular adhesion molecules *see* ICAMs
intracellular pathogens, 14, 194–5
  hiding, 268–9
intracellular signaling, 131
  activating and inhibitory signals, 207–8
  B cell activation, 168
  chemotactic agents, 187
intraepithelial lymphocytes, 56, 109, 228–9, 241, 530
  cytokine production, 241
  response to class Ib molecules, 241
intrinsic factor antibodies, 375
introns, 80
invariant chains, 107
Ir gene, 531
iron, 266
iron deficiency, 315
ISCOMs, 330
islet cell autoantibodies, 459
isografts, 384
isotype, 60–2, 72–4, 79, 531
isotype switch recombination, 173
isotype switching, 173–4, 300–1
ITAMs, 75, 107, 108, 157, 167, 193, 204, 207, 531
ITIMS, 107, 159, 193, 204, 207

J
J chain, 64, 65, 66, 531
J genes, 80, 107, 531
Janus kinases, 531
junk DNA, 80

K
K cells, 531
kallidin, 139
kallikrein, 139
Kaposi's sarcoma, 320, 321
  risk of, 404
κ chains, 531
Kell blood group system, 451, 452
keratinocytes
  contact hypersensitivity, 478–9
  cytokine release by, 479
Keshan disease, 315
KIAA0205, 405
kidney transplantation, 384, 398, 500–1
  long-term survival, 392
  reduction of allospecific T cells, 395
  risk of tumors, 404
killed vaccines, 328
  safety, 333
killer immunoglobulin-like receptors, 204, 205, 206, 531
killer inhibitory receptors, 205
  inhibitory and activating, 207–8
kinetic signaling, 348
kinin system, 139
kinins, 531
KIRs *see* killer inhibitory receptors
Klebanoff reaction, 199
Klebsiella spp., 194
knockout, 248, 251, 345, 531
Koch phenomenon, 270–1, 272
Kupffer cells, 22, 183, 531
Kveim test, 485

L
L-selectin, 137, 228, 513
L9, 405
L11, 405
lactate dehydrogenase, 410
lactobacilli, 259
lactoferricin, 266
lactoferrin, 266
lambda chains, 531
lambda repressor protein, 148
Lambert-Eaton syndrome, 375
  autoantibodies to acetylcholine receptors, 458
  defective acetylcholine release, 458–9
lamina propria, 55–6, 236
laminin receptors, 137
Langerhans' cells, 28–9, 50, 146, 183, 531
  contact hypersensitivity, 478–9
Langhans' giant cells, 484, 530
large granular lymphocytes, 531
late reactions, 437–8
LCMV *see* lymphocytic choriomeningitis virus
lectin pathway, 88, 89, 92, 531
  control by C1 inhibitor, 93
lectin receptors, 191–2
lectin-like receptors, 205
  see also individual receptors
leflunomide, 313
Legionella spp., 264
Legionella pneumophila, 194
Leishmania spp., 195, 277, 279
  avoidance of host immunity, 295
  complement evasion, 291
  genetic polymorphisms affecting outcome, 279
  host response, 286
Leishmania braziliensis, 278
Leishmania donovani, 230, 231, 278
Leishmania tropica, 278
leishmaniasis, 278
leprosy, 487, 488
  borderline leprosy reaction, 488
leukocyte adhesion, 132
  deficiency, 137, 307–8
leukocyte functional antigens, 531
leukocyte migration, 131, 132, 136
  chemokines, 136–8
  into lymphoid tissue, 137–8
leukocytes, 3
  cytokine control of movements, 129
  role in inflammation, 15–16, 128
  triggering by chemokines, 133–4
leukointegrin, 135
leukopenia, 312
leukotrienes, 182, 531
leukotriene B4, 140, 262
leukotriene D4, 140
levan, 165
Lewis blood group system, 452
LFA-1, 21, 135, 137
ligands, 67, 133, 531
Limulus assay, 259
lineage markers, 21
linkage, 531
linkage disequilibrium, 531
lipoarabinomannan, 270
lipophosphoglycan, 291
lipopolysaccharide, 165, 166, 182, 259, 531
  direct stimulation of autoreactive cells, 371
  as immune activator, 260, 261
lipoteichoic acids, 261
Listeria spp., 261, 265
Listeria monocytogenes, 195, 262, 268
liver carcinoma, 404, 409
liver transplantation, 384
LMP2, 150
LMP7, 150
LPAM-1, 135
Lsh/Ity/Bcg gene, 231
LTB4, 186
lung
  cancer, 404
  immune responses, 239–41
lupus anticoagulant, 457
lupus nephritis, 231
Ly49H, 248
Ly49I, 248
lymph node homing molecules, 228
lymph nodes, 49–54
  structure, 50–2
lymphatic filariasis, 278
lymphocyte chimeras, 302
lymphocyte functional antigen-1, 21, 155
lymphocyte migration
  regulation by chemokines, 227–8
  regulation by epithelium, 241–2
lymphocyte stimulation test, 160
lymphocytes, 9, 20–1, 148
  glucocorticoid effects, 312
  heterogeneity of, 29–30
  intraepithelial, 56, 109, 228–9

lymphocytes *(cont'd)*
  large granular, 531
  mediator release, 140
  recirculation of, 56–8
  in secondary lymphoid tissue, 48
  self-reactive, 345
  surface markers, 30–5
  transformed, 485
  tuberculin skin test reaction, 481–3
  see also B cells; T cells
lymphocytic choriomeningitis virus, 216
lymphocytopenia, 311
lymphoid organs, 38–9
  peripheral tolerance, 45–6, 351–7, 499
  primary, 20
  secondary, 20
  systemic, 48–54
lymphoid stem cells, 39–40
lymphoid tissue
  BALT, 48
  GALT, 48, 530
  leukocyte migration, 137–8
  macrophages in, 185
  in malnutrition, 314
  MALT, 48, 54–6, 531
  primary, 532
lymphokine activated killer cells (LAKS), 205, 531
lymphokines, 531
lymphoma, 404, 409
lymphotoxin, 129, 204, 241, 531
lymphotoxin-a, granuloma formation, 485–6
lysosomes, 531
lysozyme, 67, 198, 266, 531
lytic complex, 261
lytic granules, 208
  diseases caused by loss of, 211
lytic pathway, 531

M
M cells, 54
M protein, 264
macrophage 7tm G protein-coupled receptors, 186
macrophage chemotactic proteins, 532
macrophages, 19–20, 181–2
  activated, 187, 196–8, 198–200
  activation by interferons, 249
  antiparasitic activity, 288–9
  antiviral activity, 248–9
  bacterial killing, 266–8
  C5a chemotaxis, 97–8
  complement receptor expression, 192
  defects in IFN-γ signaling, 308
  differentiation, 184
  distribution, 182–4
    spleen, 49
  environmental influences on, 186
  fungal killing, 274
  heterogeneity of, 185
  intracellular pathogen receptors, 195–6
  marginal metallophilic, 185
  marginal zone, 185
  modulation of activity, 188
  phenotype, 187–8, 189
  receptor expression
    G protein-coupled receptors, 193–4
    receptors with ITIMs, 193
    vitronectin receptors, 188–91
  response to injury and immune stimuli, 184–8
  resting, 198
  role in immune response, 200
  scavenger receptors, 188–91
  secondary lymphoid tissues, 185
  secretory products, 197
  targeting by, 211–12
  in TH1 and TH2 response, 495–6
  thymic, 170, 189
  in toxic shock syndrome, 495
  virus infection, 255
macrophage colony stimulating factor, 184
macrophage inflammatory proteins, 532
macrophage neutral proteinases, 197
macrosialin, 183
MAdCAM-1, 135, 136
MAGE, 405
major basic protein, 25, 212, 290, 439
  asthma, 442
  experimental encephalomyelitis, 374
major histocompatibility complex *see* MHC
malaria, 278
  vaccine, 336
Malassezia spp., 274
male infertility, 367
MALT *see* mucosa-associated lymphoid tissue
mannan-binding lectin *see* MBL
mannose receptor, 191–2, 531
MAP kinases, 531
marginal metallophilic macrophages, 185
marginal zone, 532
  B cells, 38, 185
  macrophages, 185
mast cells, 6, 19–20, 25–6, 532
  activation, 88
    by C3a and C5a, 98
  antiparasitic activity, 288, 290–1
  C1q receptors, 96–7
  connective tissue, 433, 434, 435
  degranulation, 433–4, 435
  Fc receptor expression, 74–8
    FcεRI, 426
  histamine, 432–3
  IgE on, 62
  inflammatory mediators, 139–40
  mucosal, 433, 434, 435
  recruitment to skin and nose, 433
matrix metalloproteases, 532
MBL, 92, 141
  as acute phase protein, 141
  deficiency, 102
  parasite recognition, 281
MBL-associated serine proteases (MASPs), 92
measles vaccine, 327
medullary thymic epithelial cells, 349–50
melan A, 405
γ-melanocyte-stimulating hormone, 238
melanoma, 404
  tumor-specific antigens in, 408
membrane attack complex, 89, 99, 532
membrane attack pathway *see* terminal pathway
membrane cofactor protein, 93, 94
membrane receptors, 265
membrane-bound immunoglobulin, 177–8
membranoproliferative glomerulonephritis, 465
memory, 10
  antigen-specific, 226
  immunological, 10, 173–4, 391, 493
memory B cells, 13, 54
memory cells
  central, 228
  effector, 228
memory T cells, 228, 479–80, 532
meningitis vaccine, 335
meningococcal meningitis, 103
menopause, premature, 367
6-mercaptopurine, 393
merozoites, 279, 287
metalloproteases, 139
methotrexate, 313
MHC, 11, 21, 24, 107, 507, 532
  antigen binding groove, 112
  antigen-processing genes, 152
  and autoimmunity, 371–2
  disruption by DNA viruses, 253
  haplotypes, 230
  polymorphism, 120–1
  recognition of bound peptides by T cell receptors, 106
  surveillance of intracellular antigens, 111
  tetramers, 149
MHC class I, 107, 111–13, 118–19, 205, 528
  antiviral activity, 251
  association with endogenous peptides, 149–52
  CD8 carriage by, 348–9
  cell-cell interaction, 116–18
  codominant expression, 121
  deficiency, 233
  expression by nucleated cells, 121
  peptide binding, 114–18
    anchor residues, 114–15
    binding groove, 114
    confinement to binding groove, 115
    uniformity of size, 116
  structure, 113–14
MHC class I-like molecules *see* class Ib
MHC class I/II restriction, 528
MHC class Ib, 111–13
MHC class II, 107, 119–20, 528
  CD4 carriage by, 348–9
  cell-cell interaction, 116–18
  codominant expression, 121
  deficiency, 233, 302
  exogenous peptides, 152–4
  peptide binding, 114–18
    anchor residues, 114–15
    binding groove, 114
    extension beyond binding groove, 115–16
    uniformity of size, 116
  structure, 113–14
  thyroid, 372
MHC class III, 528

MHC haplotype, 121–2
MHC restriction, 107, 121–2, 346–7, 347, 532
MHC-linked genes, 230
MHC-peptide complex, 351
MICA, 241
MICB, 241
microbial proteins, 143–4
microglia, 183, 532
β_2-microglobulin, 111–12, 527
Microsporum spp., 274
MIF *see* migration inhibition factor
migration inhibitory factor, 228, 294, 523, 532
MIIC compartment, 532
minor histocompatibility antigens, 387, 488
mitogens, 532
  T cell activation, 159
Mitsuda reaction, 484
mixed connective tissue disease, 367
ml57, 248
MLR/MLC, 532
MMR vaccine, 327, 333, 334
MN blood group system, 451, 452
modified vaccinia Ankara strain, 327, 338
molecular mimicry, 256, 370–1
  rheumatic fever, 371
monoclonal antibodies, 80, 330, 459
  allergic disease, 444
  humanized monoclonal anti-IgE, 444
  humanized monoclonal anti-IL-5, 445
monoclonal immunoglobulin, 410
monocytes, 4, 19–20, 184
  circulating, 181–2
  complement receptor expression, 192
  fungal killing, 274
  glucocorticoid effects, 312
  mediator release, 140
  tuberculin skin test reaction, 481–3
monocytopenia, 312
monomeric pore-forming proteins, 208–9
mononuclear cells, Fc receptor expression, 74–8
mononuclear phagocytes, 4, 22, 181–202, 532
mouse
  Biozzi, 231
  cytomegalovirus, 248
  hepatitis virus, 255
  NOD, 355, 369, 375
  nude, 249, 532
  NZB/NZW, 465, 466
  Qa-1, 224
  T-cell deprived, 283
  thymocytes, 347
  transgenic, 355
  tumor antigens, 405
mRNA, differential splicing, 177, 178
MUC, 405
mucosa-associated lymphoid tissue (MALT), 48, 54–6, 531
mucosal mast cells, 433
mucosal surfaces, 228–9
  antigen-specific tolerance, 362
multiple sclerosis, 224, 367
  HLA associations, 368
MUM-1, 405
MUM-2, 405
murine *see* mouse
Mx proteins, 247
myasthenia gravis, 367, 458
  autoantibodies to acetylcholine receptors, 458
  HLA associations, 368
mycetoma, 274
mycobacteria, 233, 257
Mycobacterium bovis, 231
Mycobacterium intracellulare, 231
Mycobacterium leprae, 268, 487
  immunity, 272
Mycobacterium lepraemurium, 231
Mycobacterium tuberculosis, 257, 265, 268, 320, 487–9
  immunity, 272
mycophenolate mofetil, 313, 393
Mycoplasma pneumoniae, 455
mycoses
  subcutaneous, 274
  superficial, 274
  systemic, 274
myelin basic protein, 216
myeloid cells, 203, 207, 532
myeloma, 532
myeloperoxidase deficiency, 266
myelopoiesis, 23
Mylotarg, 410
myosin-m, 405
myxedema, primary, 367, 376

N
N regions, 532
N-ramp multispanning glycoprotein, 198
NA17A*, 405
NADPH oxidase, 199, 307, 308
naked DNA, 339
nasopharyngeal cancer, 409
natural killer cells, 5, 19–20, 27–8, 36–7, 203, 532
  activation by interferons, 249
  antigen recognition, 203–4
  antiviral activity, 248
  chemokine production, 224
  cytokine production, 224
  deficiency, 224
  delay of viral spread, 129
  Fc receptor expression, 74–8
  killing by, 210–11
  markers of, 27–8
  receptor expression, 205–7
  resistance conferred by HLA-G, 207
  targeting by, 207
  virus effects on, 254
  virus protection by, 204
natural resistance-associated macrophage protein-1 *see* Nramp1
Necator americanus, 278
necrosis, 191, 227
Nef protein, 318
negative selection, 45–6, 346, 349
Neisseria gonorrhoeae, 263–4
Neisseria meningitidis, 264
  immunity, 272
  vaccine, 329
neo-antigens, 478
neonate
  hemolytic disease of newborn, 452, 453–4, 502
  immunocompetence of, 344
  immunoglobulins, 76
  thyrotoxicosis, 374
  transient hypogammaglobulinemia of infancy, 300, 302
neoplasm, 532
neuroendocrine system, 229
neuronal cell transplantation, 384
neurons, immunosuppression by, 239
neutropenia, 307
neutrophilia, 312
neutrophils, 4–5, 22, 24–5, 532
  antibodies, 456
  antiparasitic activity, 288, 289–90
  C5a chemotaxis, 97–8
  Fc receptor expression, 74–8
  glucocorticoid effects, 312
  target damage, 450
NF-κB, 129, 532
NFAT complex, 159, 394
Nippostrongylus brasiliensis, 282, 285
nitric oxide, 249, 265, 288
  in asthma, 442–3
nitric oxide pathway, 199
nitric oxide synthetase, 249
nitroblue tetrazolium, 307
NK cells *see* natural killer cells
NKG2, 205
NKG2A, 152
NOD mouse, 355, 369, 375
NOD, 195, 532
NOD-2, 194, 195
non-goitrous hyperthyroidism, 376
non-Hodgkin's lymphoma, 404
non-nucleoside reverse transcriptase
inhibitors, 322
non-organ-specific autoimmune disease, 367
non-reactivity, 71
non-specific immunotherapy, 336, 338
Notch receptor, 345
Nramp1, 231
nuclear factor κ B, 192–3
nucleoside analog reverse transcriptase inhibitors, 322
nucleotide oligomerization domains *see* NOD
nude mouse, 249, 532
nurse cells, 41
nutrient deficiencies, 314–16
  infection and malnutrition, 314
  lymphoid tissue damage, 314
  protein-energy malnutrition, 314–15
  selenium and copper, 315
  vitamins, 315–16
  zinc and iron, 315
nutritional intervention, 316
nutritional thymectomy, 314
NY-ESO-1, 405
NZB/NZW mouse, 465, 466

O
Obese chicken, 369, 370, 372–3
obesity, 316

Onchocerca volvulus, 278
opportunistic infections, 299
  T cell deficiency, 302–5
opsonic receptors, 189
opsonins, 11, 13, 98–9
opsonization, 11, 73, 88, 98, 265, 532
oral tolerance, 240
oral vaccination, 240
organ-specific disease, 366
orphan genes, 80

P
10p, partial monosomy, 302
p-ANCA, 456
P-selectin, 136, 513
P1A, 405
p53, mutated, 405
PAMPs, 8, 140–3, 532
  bacterial infection, 259–60
  parasitic infection, 278, 280
pancreas transplantation, 384
Paneth cells, 240, 241, 266, 532
panning, 34
papilloma virus, 405, 409
Papillon-Lefèvre syndrome, 211
Paracoccidioides spp., 274
paracrine, 532
parasites, 277–97
  antibody production induced by, 286–8
  antigenic variation, 291–2, 294
  avoidance of host immunity, 295
  chronic infection, 278–80
  escape mechanisms, 291–4
  extracellular, 292–3
  genetic polymorphisms, 279
  hiding, 292–3
  host resistance, 279
  immunity against, 25–6, 74
  immunological effects, 295–7
  incidence, 278–80
  interference with host immune function, 293–4
  intracellular, 291
  resistance to immune attack, 293
  sites of infection, 278
  size, 279
  vaccines, 297, 335
paratopes, 68
partial lipodystrophy, 465
passenger cells, 532
passenger leukocytes *see* dendritic cells
passive cutaneous anaphylaxis, 532
passive immunity, 217–18
passive immunization, 336, 337
patch test, 438–9, 532
pathogen recognition, 262
pathogen-associated molecular patterns *see* PAMPs
pathogens, 7, 532
  extracellular, 8, 14
  intracellular, 8, 14
pattern recognition molecules, 259
pattern recognition receptors, 280
  human, 282
  TLR domain-containing, 281–2
pemphigoid, 367
pemphigus vulgaris, 367, 457
pentraxins, 141–2, 532
  parasite recognition, 281, 282
peptides
  antagonist, 118, 527
  antigenic, 527
  effect on macrophages, 186
  host defense, 266
  tumor antigenic, 405–6
  viral, 251
peptidoglycans, 261
perforin, 13, 204, 532
  conformational activation, 211
  in CTLs, 208
  as monomeric pore-forming protein, 208–9
periarteriolar lymphatic sheath, 532
peripheral tolerance, 45–6, 351–7, 499
  mechanisms of, 352
  role of dendritic cells, 353
pernicious anemia, 367, 375
  failure of vitamin B12 absorption, 377
peroxynitrites, 265
pertussis vaccine, 328, 334
Peyer's patches, 54–5, 532
PfEMP1, 297
phacogenic uveitis, 367
phagocyte oxidase, 199
phagocytes, 4–5, 9, 22–3, 148, 181–202
  activation, 88, 143
  bacterial killing, 265–6
  C1q receptors, 96–7
  defects, 470
  fungal killing, 274–5
  genetic defects, 307–8
  pathogen recognition receptors, 142–3
  see also individual cells
phagocytosis, 9, 13, 24–5, 98, 184, 532
  avoidance of, 268–9
  complement deficiency, 305–6
  frustrated, 200, 212, 450, 530
  immune complexes, 467–71
  non-opsonic, 194–5
  opsonic receptors, 189
  protein-energy malnutrition, 314–15
phagolysosome, 23, 198, 249
phagosomes, 23
phenotype, 532
phosphatidyl inositol diphosphate, 187
phosphorylcholine, 532
phytohemagglutin, 532
pig, in xenotransplantation, 397
pigeon fancier's lung, 462
pinworm, 277
plaque-forming cells, 315, 532
  assay, 168
plasma cells, 5, 13, 38, 54, 532
plasma endocytosis, 427
plasma enzyme systems, 138–40
plasmacytoid dendritic cells, 354
plasmin system, 139, 532
plasminogen activator, 197
Plasmodium spp., 277
  avoidance of host immunity, 295
  genetic polymorphisms affecting outcome, 279
  host response, 286
Plasmodium berghei, 282
Plasmodium falciparum, 278
  antibody effect on, 288
  erythrocyte membrane protein-1, 291
Plasmodium malariae, 278
Plasmodium ovale, 278
Plasmodium vivax, 278
Plasmodium yoelii, 282
platelet-activating factor, 140, 516, 532
platelet-activating factor receptor, 186
platelets, 6, 19–20, 26–7
  aggregation, 463
  antiparasitic activity, 288, 291
  autoantibodies, 375, 456–7
  C1q receptors, 96–7
  inflammatory mediators, 139–40
pleckstrin homology, 187
pneumococcal pneumonia vaccine, 335
Pneumocystis carinii, 194, 274, 302, 320, 321
pokeweed mitogen, 532
polio vaccine, 327, 332, 334
poly-D amino acids, 165
poly-Ig receptor, 66
polyclonal B cell activation, 166
polyendocrinopathy enteropathy, 222
polyethylene glycol, 474
polymeric bacterial flagellin, 165
polymorphonuclear granulocytes *see* granulocytes
polymorphonuclear neutrophils *see* neutrophils
polymorphs *see* granulocytes
positive selection, 44–5, 346, 351
poxvirus, 249, 327
Prausnitz-Küstner (P-K) test, 426
pre-B cell, 164
pre-B cell receptor, 165
presynaptic calcium channel antibodies, 375
prick test, 437, 532
primary biliary cirrhosis, 367
primary immunodeficiency, 299–309
  B cell deficiencies, 299–302
  genetic defects in phagocytes, 307–8
  genetic deficiency of complement proteins, 305–7
  T cell deficiencies, 302–5
primary response, 174, 533
priming, 146, 533
private idiotypes, 80
privileged sites, 345, 533
  protection of, 351–2
pro-B cells, 163
probiotics, 259, 316
programmed cell death, 352
properdin, 94
  deficiency, 103
propriocidal death, 353
prostaglandin E2, 140, 294
prostaglandins, 182, 294, 533
  contact hypersensitivity, 482
prostate-specific antigen, 410
prostatic acid phosphatase, 410
protease, 186
protease inhibitor 8, 210
protease inhibitors, 322
proteasomes, 150, 533

protective immunity, 402–3
protegrins, 266
protein A, 533
protein G, 533
protein kinase R, 247
protein-energy malnutrition, 314–15
proteoglycans, 33–4
protozoa, 277–97
  in AIDS, 321
  immunity, 497
  T cell responses, 284–5
provirus, 316
pseudoalleles, 533
pseudogenes, 80, 533
Pseudomonas spp., 264
Pseudomonas aeruginosa, 266
psoriasis, 224
public idiotypes, 80
purine nucleoside phosphorylase deficiency, 302, 303
pyogenic infections, 299
  B cell deficiency, 299–300
pyrin, 194

Q
22q11-pter, partial monosomy, 302
Qa-1, 224

R
rabies vaccine, 335
radioallergosorbent test *see* RAST
radioimmunoassay, 474, 533
RAG, 82–3, 164
RAG-1, 107, 164, 303, 533
RAG-2, 107, 164, 303, 533
ragweed pollen, 429
RAIDD, 209
random recombination, 78–80
rapamycin, 159, 313–14, 393, 394
ras, mutated, 405
RAST, 426
reactive nitrogen intermediates, 198–200, 212, 533
reactive oxygen intermediates, 198, 198–200, 212, 288, 533
receptor cross-linking, 218
receptor editing, 165, 175, 346, 359, 533
receptors
  acetylcholine, 374, 458
  αβ T cell, 111–14
  antigen, 527
  B cell, 163, 175
  B cell antigen, 60
  C1q, 96–7, 265
  C3, 96
  C3a, 96
  C5a, 96
  chemokines, 134, 523–4
  collagen, 137
  complement, 89, 97, 192, 281, 282, 529
  cytokines, 171
  f-met-leu-phe, 134
  Fc *see* Fc receptors
  fibronectin, 137
  G protein-coupled, 193–4
  glucocorticoid-induced TNF-related, 223
  HER2/Neu, 410
  insulin, 374
  interleukin-2, 157
  killer immunoglobulin-like, 204, 205, 206, 531
  killer inhibitory, 205, 207–8
  laminin, 137
  lectin, 191–2
  lectin-like, 205
  macrophage 7tm G protein-coupled, 186
  mannose, 191–2, 531
  membrane, 265
  Notch, 345
  opsonic, 189
  PAF, 186
  pattern recognition, 280, 281–2
  poly-Ig, 66
  pre-B cell, 165
  scavenger, 188–91, 281, 533
  T cell *see* T cell receptors
  tumor necrosis factor, 352
  vitronectin, 188–91
recombinant antibodies, 69
recombinant DNA-based vaccines, 327
recombination, 60, 80–3, 533
recombination signal sequences, 82
recombination-activating genes *see* RAG
recurrent idiotype, 80, 533
red cells
  autoantibodies, 375
  lysis, 455
regulators of complement activation locus, 93
regulatory T cells *see* Tregs
rejection *see* graft rejection
relative risk, 533
respiratory burst, 533
respiratory syncytial viruses, 331
reticuloendothelial system, 533
RhD antigen, 453–4
Rhesus blood group system, 451, 452
Rhesus incompatibility, 218
Rhesus prophylaxis, 454
rheumatic fever, molecular mimicry in, 371
rheumatoid arthritis, 224, 367, 463
  anti-TNFα therapy, 380
  HLA associations, 368
  pathology, 378
rhinitis, 424, 430–1
Rituxan, 410
rituximab, 459
river blindness, 278
RNaseL, 247
rosetting, 533
rotavirus vaccine, 333–4
roundworms, 278
rubella vaccine, 327

S
S protein, 95
Salmonella spp., 195, 261, 264, 265
Salmonella enterica, 268
Salmonella typhimurium, 231
sarcoidosis, 489–90
scavenger phagocytic cells, 50
scavenger receptors, 188–91, 533
  parasite recognition, 281
Schistosoma spp.
  avoidance of host immunity, 295
  genetic polymorphisms affecting outcome, 279
Schistosoma haematobium, 278
Schistosoma japonicum, 278
Schistosoma mansoni, 212, 278
schistosomiasis, 278, 489
Schwartzman reaction, 270, 271
SCID *see* severe combined immunodeficiency
scleroderma, 367
second set rejection, 385
secondary immunodeficiency, 311–24, 498–9
secondary response, 174, 533
secretory components, 533
selectins, 33, 130, 308, 533
  carbohydrate binding, 132–3
  see also individual selectins
selenium deficiency, 315
self antigen, 345
self MHC molecules, 39
self tolerance, 343
  role of CTLA-4 in, 353–4
self-reactive lymphocytes, 345
SEREX antigens, 409
serotonin *see* 5-hydroxytryptamine
serpin, 93, 210
serum amyloid-P, 141–2
serum sickness, 465, 503
severe combined immunodeficiency, 231, 232, 302–3, 533
  thymus in, 303
  X-linked, 300
Shigella spp., 268
Shigella flexneri, 196
SHIP, 159
short consensus repeats see complement control protein (CCP) domains
sialoadhesin, 184, 185
side-chain hypothesis, 78, 79
signal transduction
  CD14, 192–3
  NOD-2, 196
simian immunodeficiency virus, 318
single nucleotide polymorphism, 368
single-positive T cells, 347
sirolimus *see* rapamycin
Sjögren's syndrome, 367
skin
  cancer, 404, 409
  immune responses, 239–41
  transplantation (autograft), 384
skin tests, 437–9
  late reactions, 437–8
  patch test, 438–9, 532
  positive, 437
  prick test, 437, 532
  tuberculin skin test reaction, 481–3
  wheal and flare, 425, 437
SLAM-associated protein, 304
sleeping sickness, 278
small bowel transplantation, 384
SNP *see* single nucleotide polymorphism
sodium channel antibodies, 375
soluble mediators, 6–7

soluble mediators *(cont'd)*
  complement proteins, 6
  cytokines, 6–7
somatic hypermutation, 53, 83–5, 172, 494
  and B cell receptor changes, 175
somatic mutation, 60, 79, 165, 533
somatic recombination, 43, 79
specificity, 10, 71
sperm antibodies, 375
spirochetes, 257
spleen, 48–9, 533
  blood supply, 50
  elimination of B cells, 359–60
  marginal zone, 48, 49
  red pulp, 48, 49
  white pulp, 48, 49
spontaneous autoimmunity, 374
sporotrichosis, 274
Src kinases, 167
Ss blood group system, 452
Staphylococcus aureus, 266
  immunity, 272
STATs, 251, 533
stem cell factor, 533
stiff-man syndrome, 367
stochastic selection, 348
stomach cancer, 409
Streptococcus pneumoniae, 268
  vaccine, 329
Streptococcus pyogenes, 265
stress proteins *see* heat shock proteins
stromal reticular cells, 46
Strongyloides venezuelensis, 290
subunit vaccines, 328–9
suicide inhibitors, 103
superagonist peptides, 118
superantigens, 36, 533
  HIV proteins, 318
  T cell activation, 159
  toxicity, 272–3
superoxide anion, 199
suppressor cells, 533
surface markers, 30–5
surface plasmon resonance, 533
surfactant proteins, 141
sympathetic ophthalmia, 367
synaptotagmin, 459
synergism, 533
syngeneic grafts, 384, 533
systemic autoimmune disease, 366
systemic lupus erythematosus, 231, 365–6, 367, 370, 533
  B cell depletion, 382
  complement deficiency, 101
  HLA associations, 368

T
T cells, 5, 9, 20, 533
  allospecific, 386, 395
  αβ, 345–6
  antigen presentation, 156–7
  antigen recognition, 145–9
  apoptosis, 352–3
  autoreactive, 370–1
  cancer immunotherapy, 417
  CD4+ *see* CD4+
  CD8+ *see* CD8+
  deficiency, 233, 275, 302–5
  development, 40–3, 346
    antigen recognition, 349
    checkpoints, 350–1
  differentiation, 428
  double-negative, 345
  double-positive, 346
  education of, 529
  effector, 357
  graft recognition
    direct, 385–6, 387
    indirect, 386–7
  immunosuppressants affecting
    azathioprine, 313
    ciclosporin, 314
    cyclophosphamide, 312
    leflunomide, 313
    rapamycin, 314
    steroids, 312
    tacrolimus, 314
  markers, 31
  phenotype, 42–3
  positive and negative selection, 44–6
  receptor editing, 346
  regulation of immune response, 220–4
  regulatory *see* Tregs
  response to allergens, 428–9
  selection, 45–6, 346–7
    thymus, 349
  self-reactive, 369–70
  single-positive, 347
  specificity, 118, 494
  suppression by CTLA-4, 414
  tolerance induction, 345–57
    susceptibility to, 362
  viral immunity, 250
  virus infection, 255
T cell activation, 157–60, 168–70
  co-stimulation, 154–6
  inhibition by steroids, 312
  intracellular signalling, 158
  mitogens, 159
  role of antigenic peptides, 117–18
  superantigens, 159
T cell antigen receptors, 5, 36, 59, 107
T cell antigenic peptides, 148
T cell division, 160
T cell migration, 494
T cell receptors, 9, 105–14, 533
  αβ, 106–7, 111–14
  alternative forms, 44
  development, 345
  diversity, 43–4
  genes encoding, 109–11
  ligand interactions, 156–7
  and positive selection, 351
  recognition of MHC-bound peptides, 106
  similarity to immunoglobulin molecules, 106
T cell signaling defects, 301–2
T cell tolerance, 216
T cell-mediated immunity, 275
T-bet, 251
T-dependent antigens, 165, 167–8, 533
T-independent antigens, 165, 165–6, 533
  CD5+ activation by, 167
  memory induction by, 167
tacrolimus, 313–14, 393, 394
TAP genes, 152
TAP transporters, 533
tapasin, 206
tapeworms, 277, 278
telomerase, 225, 226
telomere, 225
  elongation, 226
  erosion, 226
terminal deoxynucleotidyl transferase, 43, 163
terminal pathway, 89, 95–6
  deficiencies, 102
  regulation of, 95–6
tetanus toxoid, 329, 334, 493
tetramers, 149
TH cells
  autoimmune, 12, 371
  differentiation, 220–1
  subsets, 221
  viral infection, 249
TH0 cells, 170
TH1 cells, 5, 170, 189, 220–1, 267
  effector mechanisms, 221
  regulation by IL-12, 221
  resistance to fungal infection, 275
TH2 cells, 5, 170, 189, 220–1
  effector mechanisms, 221
  IgE production, 428, 430
  in immune response to worms, 285–6
  resistance to fungal infection, 275
  response, role of macrophages, 495–6
Theiler's virus, 255
thoracic duct, 533
thrombocytopenic purpura, 455
thromboxanes, 533
thymectomy, 493
  nutritional, 314
thymic epithelial cells, 41–2
thymic macrophages, 170, 189
thymocytes
  adhesion of, 45
  apoptosis, 45
  developmental pathways, 347
  lineage commitment, 364, 499
  selection, 350
  stage I 42-3, II 43, III, 43
thymulin, 315
thymus, 40–3, 533
  defective embryogenesis, 304
  medullary thymic epithelial cells, 349–50
  MHC restriction, 346–7
  in SCID, 303
  T cell selection, 349
thyroid
  autoimmune disease, 376
    see also individual conditions
  MHC class II, 372
thyrotoxicosis, 367, 372, 374, 376
TIMPs, 197, 534
tissue cells
  antimicrobial mechanisms, 270
  immune responses, 235–43

tissue damage
  collateral, 270–2
  CTLs, 253–4
  immune complexes, 253–4
  virus-induced, 253–6
tissue inhibitors of matrix metalloproteases *see* TIMPs
tissue remodeling, 138–40
tissue typing, 534
TLR1, 261
TLR2
  bacterial infection, 261
  fungal infection, 274–5
  parasite recognition, 281
TLR3, parasite recognition, 281
TLR4, 260
  parasite recognition, 281
TLR6, 261
TLR9, 261
  parasite recognition, 281
7tm receptors, 133
tolerance, 216, 343–64, 499, 534
  anti-idiotype response, 362
  antigen persistence, 362
  artificially induced, 360–3
  B cells, 357–60
  central, 345–51, 352, 395, 528
  control of damaging immune response by, 362–3
  donor-specific, 394–7
  experimental demonstration of, 344
  induction of
    experimental, 345
    novel methods, 395
    post-transplantation, 395
  oral, 240
  peripheral, 45–6, 351–7, 352, 353, 499
  self, 343
  T cells, 345–57
tolerogenesis *see* tolerance
Toll-like receptors, 128, 534
  activation by bacterial PAMPs, 259–60
  activation of *IFNβ* gene, 194
  CD14 interaction, 192–3
  ligands, 356
  parasite recognition, 281
  phagocyte activation by, 143
tonsils, 534
toxic shock syndrome, 273, 495
toxin-based vaccines, 328, 329
toxoids, 327, 328
Toxoplasma gondii, 277
  avoidance of host immunity, 295
toxoplasmosis, 321
Tr1 cells, 223
TRADD, 209
transformation, 534
transformed lymphocytes, 485
transforming growth factor, 182, 533
transforming growth factor-β, 7, 237, 240, 523
transfusion reactions, 451–3
  cross-matching, 453
  destruction of donor cells, 453
  and hyperacute graft rejection, 453
transgenes, 345
transgenic animals, 534
transgenic plants, 338–9
transgenics, 345
transient hypogammaglobulinemia of infancy, 300, 302
transplantation, 383–99
  barriers to, 384–5
  cadaveric, 384
  donor-specific tolerance, 394–7
  graft rejection *see* graft rejection
  immunosuppressants, 392–4
  live donors, 397
  shortage of donor organs, 397–8
  xenotransplantation, 384, 397
  see also individual organs
trastuzumab, 459
Tregs, 222, 355–6
  autoimmune disease, 372
  CD4+ *see* CD4+
  CD4+CD25+, 222, 223, 380
  generation of, 223
  selective Foxp3 expression, 223
  suppression of other T cells by, 355–6, 357
Trichinella muris, 290
Trichinella spiralis, 230, 277, 278, 290
  hiding, 292–3
trichinosis, 278
Trichophyton spp., 274
trichuriasis, 278
Trichuris trichura, 278
tripeptidyl aminopeptidase II, 150
TRP2*, 405
Trypanosoma brucei
  avoidance of host immunity, 295
  host response, 286
Trypanosoma cruzi, 194, 277, 278
  avoidance of host immunity, 295
  host response, 286
Trypanosoma gambiense, 278
Trypanosoma rhodesiense, 278
tryptase, 435
tuberculin skin test reaction, 481–3
tuberculin-type hypersensitivity, 478, 480–3
tuberculoid granuloma, 188
tuberculosis, 487–9
  vaccine *see* BCG
tumor antigenic peptides, 405–6
tumor antigens, 404–11
tumor growth factor β_1, 223
tumor immunity, 411
  CD25+, 414–15
  concomitant, 412, 413
  downregulation, 417–18
tumor necrosis factor, 208, 260, 262, 534
  granuloma formation, 485–6
tumor necrosis factor-α, 7, 129, 130, 204, 235, 523
  antiparasitic activity, 289
  antiviral activity, 251
  intracellular signaling pathways, 131
tumor necrosis factor-β, 7, 523
tumor necrosis factor receptor, 352
tumor-bearing status, 413
tumor-specific antigens, 127, 402, 408
  lack of anti-tumor effect, 409–11
  in melanoma, 408
  specificity, 407
tumors
  distinct antigenicity of, 403
  immunogenicity, 401–21
  isolation of HSPs from, 404–5
  microorganisms causing, 409
tunicates, lymphocyte-like cells, 127
type I hypersensitivity see immediate hypersensitivity
type II hypersensitivity see antibody-mediated hypersensitivity
type III hypersensitivity *see* immune complex disease
type IV hypersensitivity *see* delayed-type hypersensitivity
typhoid vaccine, 335
tyrosinase, 405
tyrosine kinases, 157–8
tyrosine-based inhibitory motif (ITIM), 75

U
ulcerative colitis, 224, 241, 367
universal donors, 452
urticaria, 425

V
V domains, 534
V genes, 80, 107
vaccination, 14–15, 325–40, 493, 499, 534
  AIDS, 322
  cancer immunotherapy, 415–17
  and incidence of viral disease, 328
  non-infectious conditions, 336–7
  principles of, 325–6
  vectors, 338
  see also vaccines
vaccines
  anti-idiotype, 327, 330
  attenuated, 326–8, 333
  cost, 334
  diseases lacking, 337
  efficacy, 332–3
  experimental, 335–6
  injected, 331
  killed, 328, 333
  live, 326
  oral, 240
  parasitic infections, 297, 335
  safety, 333–4
  special groups, 334–5
  subunit, 328–9
  success rates, 334–6
  toxin-based, 328, 329
  toxoid, 327, 328
vaccinia virus, 325, 327
van der Waals forces, 68
variable surface glycoprotein, 291
varicella-zoster vaccine, 327, 335
vasoactive amines, 534
vasoactive intestinal peptide, 240
VCAM-1, 135, 136, 137, 480
V(D)J recombination, 105, 109–11
veiled cells, 534
very late antigens, 137, 534

Vibrio cholerae, 258
  immunity, 272
viral peptides, 251
virolysis, 250
virulence, 327
virus latency, 252–3
viruses, 204, 496
  and autoimmunity, 255–6
  host defences affected by, 254
  immunity to, 247–56
  infection of immune system cells, 254–5
  and tissue damage, 253–6
  see also individual viruses
vitamin A deficiency, 315
vitamin B_6 deficiency, 316
vitamin B_{12}, in pernicious anemia, 377
vitronectin, 95, 265
vitronectin receptors, 188–91
VLA *see* very late antigen
VLA-1, 135
VLA-2, 135, 137
VLA-3, 135, 137
VLA-4, 135, 137
VLA-5, 135
VLA-6, 135, 137
VP1-a, 493

W
Waldeyer's ring, 54, 534
warm-reactive autoantibodies, 454–5
Wegener's granulomatosis, 376, 456
Weibel-Palade bodies, 133, 136
western blotting *see* immunoblotting
wheal and flare, 425, 437
whipworm, 278
white pulp, 534
widely expressed antigens, 408
Wiskott-Aldrich syndrome, 300, 302, 305
worms, 277–97
  immune response to, 285–6
Wuchereria bancrofti, 278

X
X-linked agammaglobulinemia, 300
X-linked chronic granulomatous disease, 300
X-linked hyper-IgM, 233
X-linked immunodeficiency with increased HIgM *see* hyper-IgM
X-linked lymphoproliferative syndrome, 233
X-linked proliferative syndrome, 304
X-linked severe combined immunodeficiency, 300
X-linked syndrome, 222
X-linked-agammaglobulinemia, 233
XC:2, 525
XCL1, 525
XCR1, 524
xenogeneic barrier, 384, 534
xenotransplantation, 384, 397

Y
yellow fever vaccine, 327, 335
Yersinia enterolytica, 371
Yersinia pestis, 257

Z
Zevalin, 410
zidovudine, 322
zinc deficiency, 315
zymogen, 89
zymosan, 265

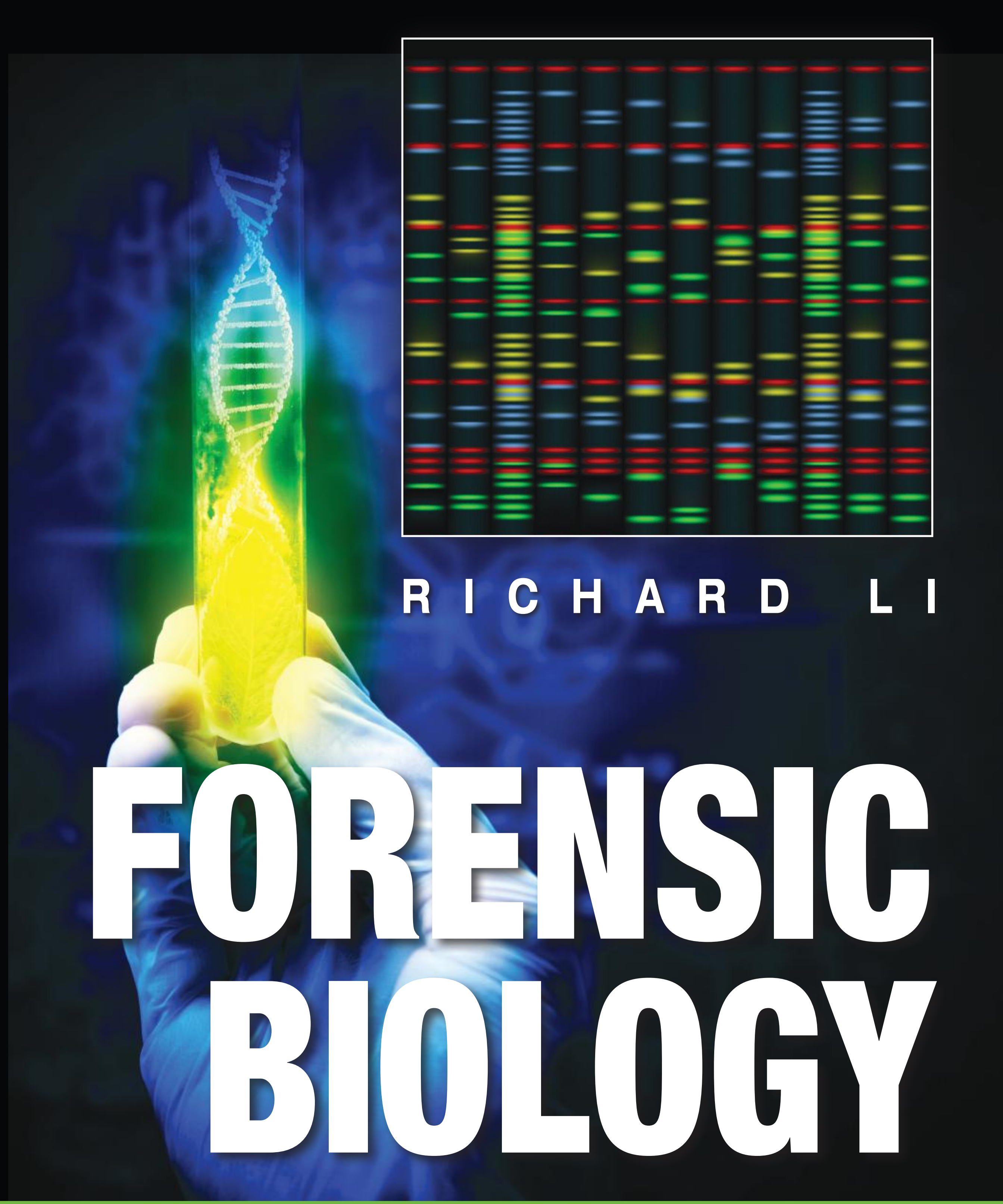

RICHARD LI

FORENSIC BIOLOGY

SECOND EDITION

CRC Press
Taylor & Francis Group

FORENSIC BIOLOGY

SECOND EDITION

RICHARD LI
John Jay College of Criminal Justice
New York, New York, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2015 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20140818

International Standard Book Number-13: 978-1-4398-8972-5 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface.. xv
Acknowledgments...xvii
Author... xix
Introduction.. xxi

SECTION I Biological Evidence

Chapter 1 Crime Scene Investigation of Biological Evidence...3

- 1.1 Protection of Crime Scene...3
- 1.2 Recognition of Biological Evidence...3
- 1.3 Searches...7
- 1.4 Documentation..16
- 1.5 Chain of Custody...17
- 1.6 Collection of Biological Evidence..18
- 1.7 Marking Evidence.. 20
- 1.8 Packaging and Transportation 20
- 1.9 Final Survey and the Release of the Crime Scene 25
- 1.10 Crime Scene Reconstruction... 29
- Bibliography 30

Chapter 2 Crime Scene Bloodstain Pattern Analysis 35

- 2.1 Basic Biological Properties of Human Blood...........................35
- 2.2 Formation of Bloodstains 35
- 2.3 Chemical Enhancement and Documentation of Bloodstain Evidence 37
- 2.4 Analyzing Spatter Stains...39
  - 2.4.1 Velocity of Blood Droplets...39
  - 2.4.2 Determining the Directionality of the Stains......................39
  - 2.4.3 Determining Angles of Impact.. 40
  - 2.4.4 Determining Area of Origin 41
- 2.5 Types of Bloodstain Patterns.. 44
  - 2.5.1 Passive Bloodstains... 44
  - 2.5.2 Transfer Bloodstains 46
  - 2.5.3 Projected Bloodstains.. 46
- Bibliography 51

Chapter 3 Forensic Biology: A Subdiscipline of Forensic Science......53

3.1 Common Disciplines of Forensic Laboratory Services......53
3.2 Laboratory Analysis of Biological Evidence......53
    3.2.1 Identification of Biological Evidence......53
    3.2.2 Comparison of Individual Characteristics of Biological Evidence......54
    3.2.3 Reporting Results and Expert Testimony......54
3.3 Forensic Science Services Related to Forensic Biology......57
    3.3.1 Forensic Pathology......58
    3.3.2 Forensic Anthropology......58
    3.3.3 Forensic Entomology......60
    3.3.4 Forensic Odontology......64
3.4 Brief History of the Development of Forensic Biology......65
    3.4.1 Antigen Polymorphism......65
    3.4.2 Protein Polymorphism......66
    3.4.3 DNA Polymorphism......66
        3.4.3.1 Genes and Related Sequences......66
        3.4.3.2 Intergenic Noncoding Sequences......66
        3.4.3.3 Human DNA Polymorphic Markers......68
        3.4.3.4 Forensic DNA Polymorphism Profiling......69
Bibliography......70

Chapter 4 Sources of Biological Evidence......77

4.1 Bodily Fluids......77
    4.1.1 Extracellular Nucleic Acids......77
4.2 Cells......81
    4.2.1 Cell Surface Markers......81
    4.2.2 Nucleated Cells......81
    4.2.3 Mitochondria and Other Organelles......83
    4.2.4 Cytosol......84
        4.2.4.1 Messenger RNAs......84
        4.2.4.2 MicroRNAs......85
4.3 Tissues......87
    4.3.1 Skin......87
        4.3.1.1 Biology of Skin......87
        4.3.1.2 Skin as Source of DNA Evidence......87
    4.3.2 Hair......88
        4.3.2.1 Biology of Hair......89
        4.3.2.2 Hair as Source of DNA Evidence......91
    4.3.3 Bone......92
        4.3.3.1 Biology of Bone......92
        4.3.3.2 Bone as Source of DNA Evidence......94
    4.3.4 Teeth......98
        4.3.4.1 Biology of Teeth......98
        4.3.4.2 Teeth as Source of DNA Evidence......100
Bibliography......102

SECTION II Basic Techniques in Forensic Biology

Chapter 5 Nucleic Acid Extraction 111

- 5.1 Basic Principles of DNA Extraction 111
  - 5.1.1 Cell and Tissue Disruption 111
  - 5.1.2 Lysis of Cellular and Organelle Membranes 114
  - 5.1.3 Removal of Proteins and Cytoplasmic Constituents 114
  - 5.1.4 Storage of DNA Solutions 115
  - 5.1.5 Contamination 115
- 5.2 Methods of DNA Extraction 115
  - 5.2.1 Extraction with Phenol–Chloroform 115
    - 5.2.1.1 Cell Lysis and Protein Digestion 115
    - 5.2.1.2 Extraction with Organic Solvents 115
    - 5.2.1.3 Concentrating DNA 115
  - 5.2.2 Extraction by Boiling Lysis and Chelation 117
    - 5.2.2.1 Washing 117
    - 5.2.2.2 Boiling 117
    - 5.2.2.3 Centrifugation 117
  - 5.2.3 Silica-Based Extraction 118
    - 5.2.3.1 Cell Lysis and Protein Digestion 118
    - 5.2.3.2 DNA Adsorption onto Silica 118
    - 5.2.3.3 Washing 119
    - 5.2.3.4 Elution of DNA 119
  - 5.2.4 Differential Extraction 121
- 5.3 Essential Features of RNA 122
- 5.4 Methods of RNA Extraction 124
  - 5.4.1 RNA–DNA Coextraction 124
  - 5.4.2 miRNA Extraction 124
- Bibliography 127

Chapter 6 DNA Quantitation 133

- 6.1 Slot Blot Assay 133
- 6.2 Fluorescent Intercalating Dye Assay 134
- 6.3 Quantitative PCR Assay 137
  - 6.3.1 Real-Time Quantitative PCR 137
    - 6.3.1.1 TaqMan Method 138
- Bibliography 140

Chapter 7 Amplification by Polymerase Chain Reaction 143

- 7.1 Denaturation and Renaturation of DNA 143
- 7.2 Basic Principles of Polymerase Chain Reaction 143
- 7.3 Essential PCR Components 145
  - 7.3.1 Thermostable DNA Polymerases 145
  - 7.3.2 PCR Primers 146
  - 7.3.3 Other Components 147

7.4 Cycle Parameters 147
7.5 Factors Affecting PCR 149
  7.5.1 Template Quality 149
  7.5.2 Inhibitors 149
  7.5.3 Contamination 149
7.6 Reverse Transcriptase PCR for RNA-Based Assays 150
  7.6.1 Reverse Transcription 150
  7.6.2 Oligodeoxynucleotide Priming 151
  7.6.3 Reverse Transcriptase PCR 151
Bibliography 153

Chapter 8 DNA Electrophoresis 159

8.1 Basic Principles 159
8.2 Supporting Matrices 159
  8.2.1 Agarose 160
  8.2.2 Polyacrylamide 161
    8.2.2.1 Polymerization Reaction 161
    8.2.2.2 Cross-Linking Reaction 161
    8.2.2.3 Denaturing Polyacrylamide Electrophoresis 163
8.3 Apparatus and Forensic Applications 163
  8.3.1 Slab Gel Electrophoresis 163
    8.3.1.1 Agarose Gel Electrophoresis 163
    8.3.1.2 Polyacrylamide Gel Electrophoresis 164
  8.3.2 Capillary Electrophoresis 164
  8.3.3 Microfluidic Devices 165
    8.3.3.1 Modular Microfluidic Devices 167
    8.3.3.2 Integrated Microfluidic Devices 168
8.4 Estimation of DNA Size 168
  8.4.1 Relative Mobility 168
  8.4.2 Local Southern Method 170
Bibliography 171

Chapter 9 Detection Methods 175

9.1 Direct Detection of DNA in Gels 175
  9.1.1 Fluorescent Intercalating Dye Staining 175
  9.1.2 Silver Staining 175
9.2 Detection of DNA Probes in Hybridization-Based Assays 177
  9.2.1 Radioisotope Labeled Probes 177
  9.2.2 Enzyme-Conjugated Probe with Chemiluminescence Reporting System 180
  9.2.3 Biotinylation of DNA with Colorimetric Reporting Systems 180
    9.2.3.1 Biotin 180
    9.2.3.2 Enzyme-Conjugated Avidin 180
    9.2.3.3 Reporter Enzyme Assay 181
9.3 Detection Methods for PCR-Based Assays 182
  9.3.1 Fluorescence Labeling 182

9.3.1.1 Fluorescent Dyes 182
9.3.1.2 Labeling Methods....182
9.3.1.3 Fluorophore Detection 185
Bibliography 186

Chapter 10 Serology Concepts 189

10.1 Serological Reagents 189
10.1.1 Immunogens and Antigens....189
10.1.2 Antibodies....189
10.1.2.1 Polyclonal Antibodies....191
10.1.2.2 Monoclonal Antibodies....191
10.1.2.3 Antiglobulins 192
10.2 Strength of Antigen–Antibody Binding 193
10.3 Antigen–Antibody Binding Reactions....193
10.3.1 Primary Reactions 194
10.3.2 Secondary Reactions 194
10.3.2.1 Precipitation....195
10.3.2.2 Agglutination.... 197
Bibliography 198

Chapter 11 Serology Techniques: Past, Current, and Future.... 199

11.1 Introduction to Forensic Serology.... 199
11.1.1 The Scope of Forensic Serology.... 199
11.1.2 Class Characteristics and Individual Characteristics of Biological Evidence 199
11.1.3 Presumptive and Confirmatory Assays....200
11.1.4 Primary and Secondary Binding Assays....200
11.2 Primary Binding Assays 201
11.2.1 Enzyme-Linked Immunosorbent Assay (ELISA) 201
11.2.2 Immunochromatographic Assays 202
11.3 Secondary Binding Assays.... 204
11.3.1 Precipitation-Based Assays.... 204
11.3.1.1 Immunodiffusion.... 204
11.3.1.2 Immunoelectrophoretic Methods.... 206
11.3.2 Agglutination-Based Assays.... 208
11.4 DNA Methylation Assays for Bodily Fluid Identification....210
11.5 Forensic Applications of RNA-Based Assays and RNA Profiling..... 212
11.5.1 Messenger RNA-Based Assays....212
11.5.2 MicroRNA-Based Assays....214
11.6 Proteomic Approaches Using Mass Spectrometry for Bodily Fluid Identification 215
11.6.1 Mass Spectrometric Instrumentation for Protein Analysis 215
11.6.2 Analysis Strategies for Protein Identification....216
11.7 Microbial DNA Analysis for Bodily Fluid Identification....217
11.8 Nondestructive Assays for the Identification of Bodily Fluids....... 220
Bibliography 221

SECTION III Identification of Biological Evidence

Chapter 12 Identification of Blood 231

- 12.1 Biological Properties 231
  - 12.1.1 Red Blood Cells 231
  - 12.1.2 White Blood Cells 231
  - 12.1.3 Platelets 231
- 12.2 Presumptive Assays for Identification 232
  - 12.2.1 Mechanisms of Presumptive Assays 232
    - 12.2.1.1 Oxidation–Reduction Reactions 234
  - 12.2.2 Colorimetric Assays 234
    - 12.2.2.1 Phenolphthalin Assay 234
    - 12.2.2.2 Leucomalachite Green (LMG) Assay 234
    - 12.2.2.3 Benzidine and Derivatives 234
  - 12.2.3 Chemiluminescence and Fluorescence Assays 236
    - 12.2.3.1 Luminol (3-Aminophthalhydrazide) 237
    - 12.2.3.2 Fluorescin 238
  - 12.2.4 Factors Affecting Presumptive Assay Results 238
    - 12.2.4.1 Oxidants 238
    - 12.2.4.2 Plant Peroxidases 239
    - 12.2.4.3 Reductants 239
- 12.3 Confirmatory Assays for Identification 239
  - 12.3.1 Microcrystal Assays 239
    - 12.3.1.1 Hemochromagen Crystal Assay 239
    - 12.3.1.2 Hematin Crystal Assay 240
  - 12.3.2 Other Assays 240
- Bibliography 241

Chapter 13 Species Identification 245

- 13.1 General Considerations 245
  - 13.1.1 Types of Antibodies 245
  - 13.1.2 Titration of Antibodies 246
  - 13.1.3 Antibody Specificity 246
  - 13.1.4 Optimal Conditions for Antigen–Antibody Binding 247
- 13.2 Assays 247
  - 13.2.1 Immunochromatographic Assays 247
    - 13.2.1.1 Identification of Human Hemoglobin Protein 247
    - 13.2.1.2 Identification of Human Glycophorin A Protein 248
  - 13.2.2 Double Immunodiffusion Assays 251
    - 13.2.2.1 Ring Assay 251
    - 13.2.2.2 Ouchterlony Assay 251
  - 13.2.3 Crossed-Over Electrophoresis 251
- Bibliography 255

Chapter 14 Identification of Semen 257

14.1 Biological Characteristics 257
  14.1.1 Spermatozoa 257
  14.1.2 Acid Phosphatase 259
  14.1.3 Prostate-Specific Antigen 259
  14.1.4 Seminal Vesicle–Specific Antigen 259
14.2 Analytical Techniques for Identifying Semen 260
  14.2.1 Presumptive Assays 260
    14.2.1.1 Lighting Techniques for Visual Examination of Semen Stains 260
    14.2.1.2 Acid Phosphatase Techniques 261
  14.2.2 Confirmatory Assays 264
    14.2.2.1 Microscopic Examination of Spermatozoa 264
    14.2.2.2 Identification of Prostate-Specific Antigen 265
    14.2.2.3 Identification of Seminal Vesicle–Specific Antigen 269
    14.2.2.4 RNA-Based Assays 270
Bibliography 271

Chapter 15 Identification of Saliva 277

15.1 Biological Characteristics of Saliva 277
  15.1.1 Amylases 277
15.2 Analytical Techniques for Identification of Saliva 279
  15.2.1 Presumptive Assays 279
    15.2.1.1 Visual Examination 279
    15.2.1.2 Determination of Amylase Activity 279
  15.2.2 Confirmatory Assays 281
    15.2.2.1 Identification of Human Salivary α-Amylase 281
    15.2.2.2 RNA-Based Assays 285
Bibliography 286

Chapter 16 Identification of Vaginal Secretions and Menstrual Blood 289

16.1 Identification of Vaginal Stratified Squamous Epithelial Cells 289
  16.1.1 Lugol's Iodine Staining and Periodic Acid–Schiff Method 290
  16.1.2 Dane's Staining Method 293
16.2 Identification of Vaginal Acid Phosphatase 293
16.3 Identification of Vaginal Bacteria 294
16.4 Outlook for Confirmatory Assays of Vaginal Secretions 295
16.5 Menstruation 295
  16.5.1 Uterine Cycle 296
  16.5.2 Uterine Endometrial Hemostasis 298
16.6 D-dimer Assay 299
16.7 Lactate Dehydrogenase Assay 301
16.8 RNA-Based Assays 302
Bibliography 303

Chapter 17 Identification of Urine, Sweat, Fecal Matter, and Vomitus........................ 307
17.1 Identification of Urine.. 307
  17.1.1 Urine Formation... 307
  17.1.2 Presumptive Assays.. 308
    17.1.2.1 The Identification of Urea 308
    17.1.2.2 Identification of Creatinine 314
  17.1.3 Confirmative Assays...315
    17.1.3.1 Identification of Tamm–Horsfall Protein.............315
    17.1.3.2 Identification of 17-Ketosteroids 315
17.2 Identification of Sweat...316
  17.2.1 Biology of Perspiration...316
  17.2.2 Sweat Identification Assays 317
17.3 Identification of Fecal Matter..318
  17.3.1 Fecal Formation 319
  17.3.2 Fecal Matter Identification Assays 319
    17.3.2.1 Macroscopic and Microscopic Examination........ 320
    17.3.2.2 Urobilinoids Tests 322
    17.3.2.3 Fecal Bacterial Identification................................ 324
17.4 Identification of Vomitus.. 324
  17.4.1 Biology of Gastric Fluid 324
  17.4.2 Vomitus Identification Assays 325
Bibliography 326

SECTION IV Individualization of Biological Evidence

Chapter 18 Blood Group Typing and Protein Profiling 331
18.1 Blood Group Typing...331
  18.1.1 Blood Groups...331
  18.1.2 ABO Blood Group System...331
    18.1.2.1 Biosynthesis of Antigens 333
    18.1.2.2 Molecular Basis of the ABO System 333
    18.1.2.3 Secretors 335
    18.1.2.4 Inheritance of A and B Antigens 336
  18.1.3 Forensic Applications of Blood Group Typing................... 336
  18.1.4 Blood Group Typing Techniques.. 336
    18.1.4.1 Lattes Crust Assay.. 336
    18.1.4.2 Absorption–Elution Assay.................................... 338
18.2 Forensic Protein Profiling.. 338
  18.2.1 Methods 339
    18.2.1.1 Matrices Supporting Protein Electrophoresis...... 340
    18.2.1.2 Separation by Molecular Weight.......................... 340
    18.2.1.3 Separation by Isoelectric Point............................. 341
  18.2.2 Erythrocyte Protein Polymorphisms................................. 342
    18.2.2.1 Erythrocyte Isoenzymes 342
    18.2.2.2 Hemoglobin 342

18.2.3 Serum Protein Polymorphisms 344
Bibliography 345

Chapter 19 Variable Number Tandem Repeat Profiling 353

19.1 Restriction Fragment Length Polymorphism 353
  19.1.1 Restriction Endonuclease Digestion 354
  19.1.2 Southern Transfer 356
  19.1.3 Hybridization with Probes 357
    19.1.3.1 Multilocus Probe Technique 357
    19.1.3.2 Single-Locus Probe Technique 358
  19.1.4 Detection 359
  19.1.5 Factors Affecting RFLP Results 360
    19.1.5.1 DNA Degradation 360
    19.1.5.2 Restriction Digestion–Related Artifacts 361
    19.1.5.3 Electrophoresis and Blotting Artifacts 362
19.2 Amplified Fragment Length Polymorphism 363
Bibliography 364

Chapter 20 Autosomal Short Tandem Repeat Profiling 369

20.1 Characteristics of STR Loci 369
  20.1.1 Core Repeat and Flanking Regions 369
  20.1.2 Repeat Unit Length 370
  20.1.3 Repeat Unit Sequences 370
20.2 STR Loci Commonly Used for Forensic DNA Profiling 371
20.3 Forensic STR Analysis 375
  20.3.1 Determining the Genotypes of STR Fragments 376
  20.3.2 Interpretation of STR Profiling Results 376
    20.3.2.1 Inclusion (Match) 376
    20.3.2.2 Exclusion 376
    20.3.2.3 Inconclusive Result 376
20.4 Factors Affecting Genotyping Results 376
  20.4.1 Mutations 377
    20.4.1.1 Mutations at STR Core Repeat Regions 377
    20.4.1.2 Chromosomal and Gene Duplications 380
    20.4.1.3 Point Mutations 381
  20.4.2 Amplification Artifacts 382
    20.4.2.1 Stuttering 382
    20.4.2.2 Nontemplate Adenylation 383
    20.4.2.3 Heterozygote Imbalance 383
    20.4.2.4 Allelic Dropout 383
  20.4.3 Electrophoretic Artifacts 384
    20.4.3.1 Pull-Up Peaks 384
    20.4.3.2 Spikes 384
20.5 Genotyping of Challenging Forensic Samples 384
  20.5.1 Degraded DNA 384

20.5.2 Low Copy Number DNA Testing ... 386
20.5.3 Mixtures ... 386
Bibliography ... 389

Chapter 21 Sex Chromosome Haplotyping and Gender Identification ... 407

21.1 Y Chromosome Haplotyping ... 407
21.1.1 Human Y Chromosome Genome ... 407
21.1.1.1 Pseudoautosomal Region ... 407
21.1.1.2 Male-Specific Y Region ... 407
21.1.1.3 Polymorphic Sequences ... 408
21.1.2 Y-STR ... 409
21.1.2.1 Core Y-STR Loci ... 410
21.1.2.2 Multilocal Y-STR Loci ... 411
21.1.2.3 Rapidly Mutating Y-STR ... 411
21.2 X Chromosome Haplotyping ... 412
21.3 Sex Typing for Gender Identification ... 416
21.3.1 Amelogenin Locus ... 416
21.3.2 Other Loci ... 418
Bibliography ... 419

Chapter 22 Single Nucleotide Polymorphism Profiling ... 437

22.1 Basic Characteristics of SNPs ... 437
22.2 Forensic Applications of SNP Profiling ... 438
22.2.1 *HLA-DQA1* Locus ... 438
22.2.1.1 DQα AmpliType and Polymarker Assays ... 438
22.2.1.2 Allele-Specific Oligonucleotide Hybridization ... 440
22.2.2 Current and Potential Applications of SNP Analysis ... 443
22.2.2.1 Application of SNP Analysis for Forensic Identification ... 443
22.2.2.2 Potential Applications of SNPs for Phenotyping ... 444
22.3 SNP Techniques ... 444
22.3.1 Next-Generation Sequencing Technologies ... 444
22.3.2 DNA Samples, Sequencing Library, and Template Preparation ... 446
22.3.3 NGS Chemistry ... 447
22.3.4 NGS Coverage ... 447
Bibliography ... 451

Chapter 23 Mitochondrial DNA Profiling ... 461

23.1 Human Mitochondrial Genome ... 461
23.1.1 Genetic Contents of Mitochondrial Organelle Genomes ... 461
23.1.2 Maternal Inheritance of mtDNA ... 463
23.2 mtDNA Polymorphic Regions ... 463
23.2.1 Hypervariable Regions ... 463
23.2.2 Heteroplasmy ... 463

23.3 Forensic mtDNA Testing 465
  23.3.1 General Considerations 465
  23.3.2 mtDNA Screen Assay........ 466
  23.3.3 mtDNA Sequencing........ 467
    23.3.3.1 PCR Amplification 467
    23.3.3.2 DNA Sequencing Reactions........ 468
    23.3.3.3 Electrophoresis, Sequence Analysis, and Mitotype Designations 471
  23.3.4 Interpretation of mtDNA Profiling Results 472
    23.3.4.1 Exclusion 474
    23.3.4.2 Cannot Exclude 474
    23.3.4.3 Inconclusive Result 474
Bibliography 474

SECTION V Forensic Issues

Chapter 24 Forensic DNA Databases: Tools for Crime Investigations........ 485

24.1 Brief History of Forensic DNA Databases........ 485
24.2 Infrastructure of CODIS........ 485
24.3 Indexes of CODIS 487
24.4 Database Entries........ 489
24.5 Database Expansion 490
24.6 DNA Profiles........491
24.7 Routine Database Searches for Forensic Investigations 492
  24.7.1 Case-to-Offender Searches........ 492
  24.7.2 Case-to-Case Searches 493
  24.7.3 Search Stringency and Partial Matches........ 493
24.8 Familial Searches 495
  24.8.1 Legal and Ethical Issues of Familial Search........ 495
  24.8.2 Familial Search Strategies........ 496
    24.8.2.1 Identity-by-State and Kinship Index Method...... 496
    24.8.2.2 Focusing on Rare Alleles........ 496
    24.8.2.3 Excluding Candidates through Y-STR Screening 498
Bibliography 498

Chapter 25 Evaluation of the Strength of Forensic DNA Profiling Results........ 503

25.1 A Review of Basic Principles of Genetics 503
  25.1.1 Mendelian Genetics........ 503
  25.1.2 Population Genetics 504
    25.1.2.1 Allele Frequency........ 504
    25.1.2.2 Genotype Frequency........ 504
    25.1.2.3 Heterozygosity........ 504
    25.1.2.4 Hardy–Weinberg Principle 505
    25.1.2.5 Testing for HW Proportions of Population Databases........ 505
    25.1.2.6 Probability of Match 511

25.2 Statistical Analysis of DNA Profiling Results......512
  25.2.1 Genotypes......513
    25.2.1.1 Profile Probability......513
    25.2.1.2 Likelihood Ratio......516
  25.2.2 Haplotypes......516
    25.2.2.1 Mitotypes Observed in Database......518
    25.2.2.2 Mitotype Not Observed in Database......518
Bibliography......518

Chapter 26 Quality Assurance and Quality Control......523

26.1 US Quality Standards......523
26.2 International Quality Standards......524
26.3 Laboratory Accreditation......525
26.4 Laboratory Validation......525
26.5 Proficiency Testing......526
26.6 Certification......526
26.7 Forensic DNA Analyst Qualifications......527
26.8 Code of Ethics of Forensic Scientists......528
Bibliography......529

Index......533

Preface

Since the first edition of this book, new developments in forensic biology have led to a rapid expansion of the knowledge of the field. Therefore, it is necessary to create a new edition. This edition provides updates in most chapters of the original edition. Additionally, three new chapters (Chapters 2, 16, and 17) have been added and approximately 200 new figures have been created for this edition. Just like the first edition of this book, the new edition aims to inspire an undergraduate audience to tackle new challenges in the forensic biology field. It is written specifically to provide a general understanding of forensic biology and assist students in becoming more knowledgeable about the field of forensic biology and the wealth of available information. My readers should find that this edition of the book contains useful information, presented in a way that is more easily understood. Hopefully, it will be utilized by students, particularly those interested in forensic biology, to further enhance their education and training. I will continue to be open to suggestions in the future.

Richard C. Li
John Jay College of Criminal Justice
City University of New York
New York, NY

Acknowledgments

Many people contributed to this book. I would like to thank my family members for their support during the many hours I spent sitting in front of my computer. I would like to thank Jan Li for spending numerous hours assisting with the completion of this book by handling tasks such as image preparation and text editing. I would also like to thank Zully Santiago and Robert Greco for their proofreading of the manuscript. Additionally, I must acknowledge that my students at John Jay College of Criminal Justice provided excellent comments, suggestions, and even a few criticisms. The editorial and production staff at Taylor & Francis Group should also be acknowledged for their contributions. I would like to thank Mark Listewnik and Becky Masterman, senior acquisitions editors, for excellent suggestions and support, as well as project editor Richard Tressider for helping to keep the project on track. I would also like to thank Michelle van Kampen, project manager at Deanta Global Publishing Services, for overseeing the production of the book.

Author

Richard Li earned his MS in forensic science from the University of New Haven and his PhD in molecular biology from the University of Wisconsin–Madison. After completing his PhD, Dr. Li was awarded a postdoctoral fellowship at Weill Medical College of Cornell University and subsequently worked as a research faculty member at the School of Medicine of Yale University. Dr. Li has served as a criminalist at the Department of Forensic Biology, the Office of Chief Medical Examiner of New York City. For the past decade, he has held faculty positions in forensic science programs in the United States. Currently, Dr. Li is an associate professor of forensic biology at John Jay College of Criminal Justice. Additionally, he serves as a faculty member for the PhD program in forensic science for the college. Dr. Li's current research interests include the identification and analysis of biological specimens that are potentially useful for forensic investigations.

Introduction

This text defines forensic biology as analyses performed in the forensic biology sections of forensic laboratories and thus focuses on forensic serology and forensic DNA analysis. The aim of this book is to emphasize the basic science and its application to forensic science in an effort to make the principles more understandable. In addition, it introduces the language of forensic biology, thus enabling students to become comfortable with its use, and it provides clear explanations of the principles of forensic analysis.

To convey a general understanding of the concepts of forensic biology, it is necessary to include explanations of various techniques that are utilized in the field. The intent is to provide students with a scientific grounding in the area of forensic biology by offering an introduction to methods and techniques utilized by forensic biology laboratories. The techniques introduced in this text are accompanied by brief background descriptions and discussions of basic principles and techniques. Schematic illustrations are included where necessary. The text also acknowledges the benefits and limitations that apply to forensic biology techniques. Forensic techniques that were used in the past are also described. Learning past examples of forensic tests can help students to review historical forensic cases.

This text contains five modules, organized by section: Section I, Biological Evidence (Chapters 1–4); Section II, Basic Techniques in Forensic Biology (Chapters 5–11); Section III, Identification of Biological Evidence (Chapters 12–17); Section IV, Individualization of Biological Evidence (Chapters 18–23); and Section V, Forensic Issues (Chapters 24–26). The 26 chapters are designed to be covered in a single-semester course.

SECTION I

Biological Evidence

1

Crime Scene Investigation of Biological Evidence

A forensic investigation involving biological evidence usually begins at the crime scene. The crime scene investigation process includes maintaining scene security, preparing documentation, and collecting and preserving physical evidence. A crime scene investigation requires teamwork and effort. Each team member should be assigned specific tasks (Figure 1.1).

1.1 Protection of Crime Scene

A crime scene investigation begins with the initial response to a scene (Figure 1.2). Securing and protecting the scene are important steps in a crime scene investigation (Figures 1.3 through 1.6), and this task is usually carried out by the first responding officer arriving on the scene. The entry of authorized personnel admitted to the scene should be documented using a log sheet (Figure 1.7). Suspects, witnesses, and living victims should be evacuated from the scene. If a victim is wounded, medical attention should be sought.

Appropriate supplies and devices should be used to prevent the contamination of evidence by investigators. Protective wear and devices including a face mask or shield, safety eyeglasses, a disposable coverall bodysuit, gloves, shoe covers, and a hairnet should be used (Figures 1.8 and 1.9). Exposure to bodily fluids may occur during a crime scene investigation. An investigator can be exposed to bodily fluids through the mucous membranes, skin exposure, and needlestick injuries (especially when investigating a clandestine drug laboratory scene). Therefore, biosafety procedures must be followed for the protection of personnel from infectious blood-borne pathogens such as the human immunodeficiency virus (HIV), hepatitis B virus (HBV), and hepatitis C virus (HCV); infectious aerosol tuberculosis pathogens; and other biohazardous materials.

1.2 Recognition of Biological Evidence

A preliminary survey should be carried out to evaluate potential evidence. In particular, the recognition of evidence plays a critical role in solving or prosecuting crimes. The priority of the potential evidence at crime scenes should be assessed based on each item's relevance to the solution of the case. Higher priority should be assigned to evidence with probative value to the case. For example, the evidence related to a *corpus delicti* is considered to be of the highest priority. *Corpus delicti* is a Latin term meaning "body of crime." In Western law, it primarily refers to the principle that in order for an individual to be convicted, it is necessary to prove the occurrence

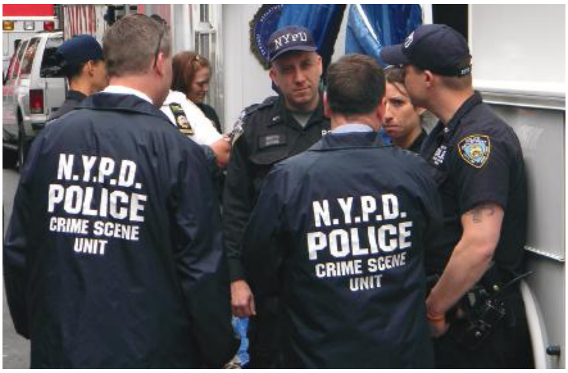
N.Y.P.D.
POLICE
CRIME SCENE
UNIT
N.Y.P.D.
POLICE
CRIME SCENE
UNIT

Figure 1.1 A crime scene investigation team. (© Richard C. Li.)

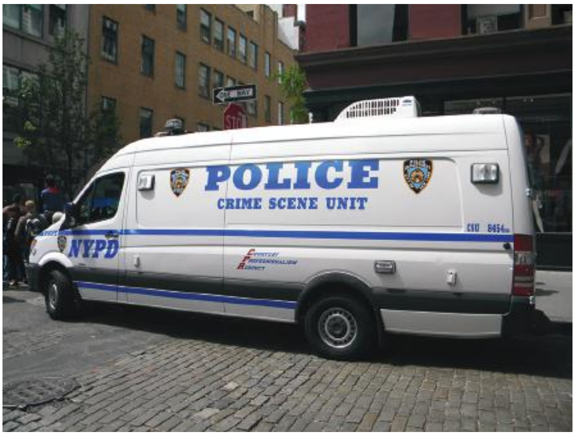
POLICE
CRIME SCENE UNIT
NYPD

Figure 1.2 A crime scene unit vehicle that is used to respond to crime scenes. This type of vehicle is usually outfitted with devices and supplies that investigators need when processing a crime scene, as well as evidence packaging materials, fingerprint collection kits, and DNA collection kits. Additionally, it can be equipped with a workstation for computer access, a refrigerator for storing chemicals, and a compact fuming hood for processing latent fingerprints, as well as equipment cabinets and drawers. (© Richard C. Li.)

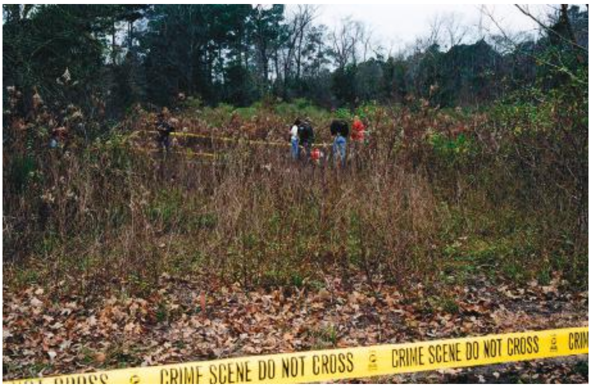
CRIME SCENE DO NOT CROSS
CRIME SCENE DO NOT CROSS

Figure 1.3 Crime scene barrier tape is used to ensure that only investigators are admitted to the scene. (Courtesy of H. Brewster.)

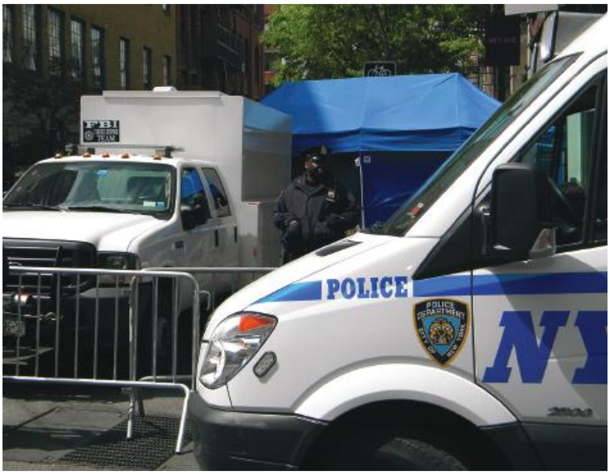
FBI
POLICE
POLICE DEPARTMENT CITY OF NEW YORK
NY

Figure 1.4 A police officer guards the crime scene. (© Richard C. Li.)

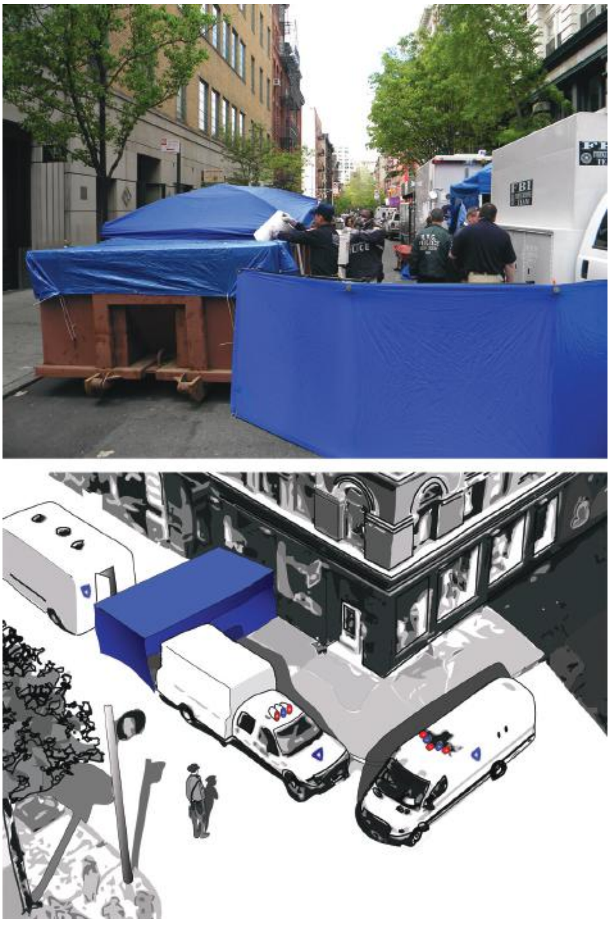

Figure 1.5 Crime scene privacy screen and tent. The screen (top) and the tent (bottom) are useful devices for shielding the evidence or body from viewing by unauthorized personnel. (© Richard C. Li.)

FBI
NYPD
POLICE

Figure 1.6 Barricades are set up to keep crowds at a distance from the scene. (© Richard C. Li.)

of the crime. In a forensic investigation, it also refers to the physical evidence proving that a crime was committed. For example, when an individual is missing, a missing persons investigation is usually initiated. If *corpus delicti*, such as a dead body or a victim's blood at a crime scene (Figure 1.10), is discovered during the investigation, a homicide case can be established and a suspect can be charged with homicide. Higher priority should also be attached to evidence that can establish connections such as *victim-to-perpetrator linkage*. For example, items found in a perpetrator's possession may be linked to a victim. This also applies to transfer evidence based on the principles of transfer theory, also known as the *Locard exchange principle*, which theorizes that the cross-transfer of evidence occurs when a perpetrator has any physical contact with an object or another person (Figures 1.11 through 1.14). Thus, trace evidence, such as hairs and fibers, may be transferred from a perpetrator to a victim or vice versa. This explains why it is important to ensure that perpetrators and their belongings are thoroughly searched for trace evidence. Likewise, victims and their belongings should be examined for the same reason.

Victim-to-scene and perpetrator-to-scene linkages can also be established. Blood belonging to a perpetrator or a victim found at a crime scene can establish such a linkage (Figure 1.15). Additionally, reciprocal transfers of trace evidence from crime scenes can be used to link a suspect or a victim to a crime scene. A perpetrator may present a unique *modus operandi* (MO). *Modus operandi*, a Latin term commonly used in criminal investigations, refers to a particular pattern of characteristics and the manner in which a crime is committed. For example, Richard Cottingham, a serial killer known as "the torso killer," dismembered his victims and took their limbs and heads with him but left their torsos at the scene. He then set the rooms on fire before

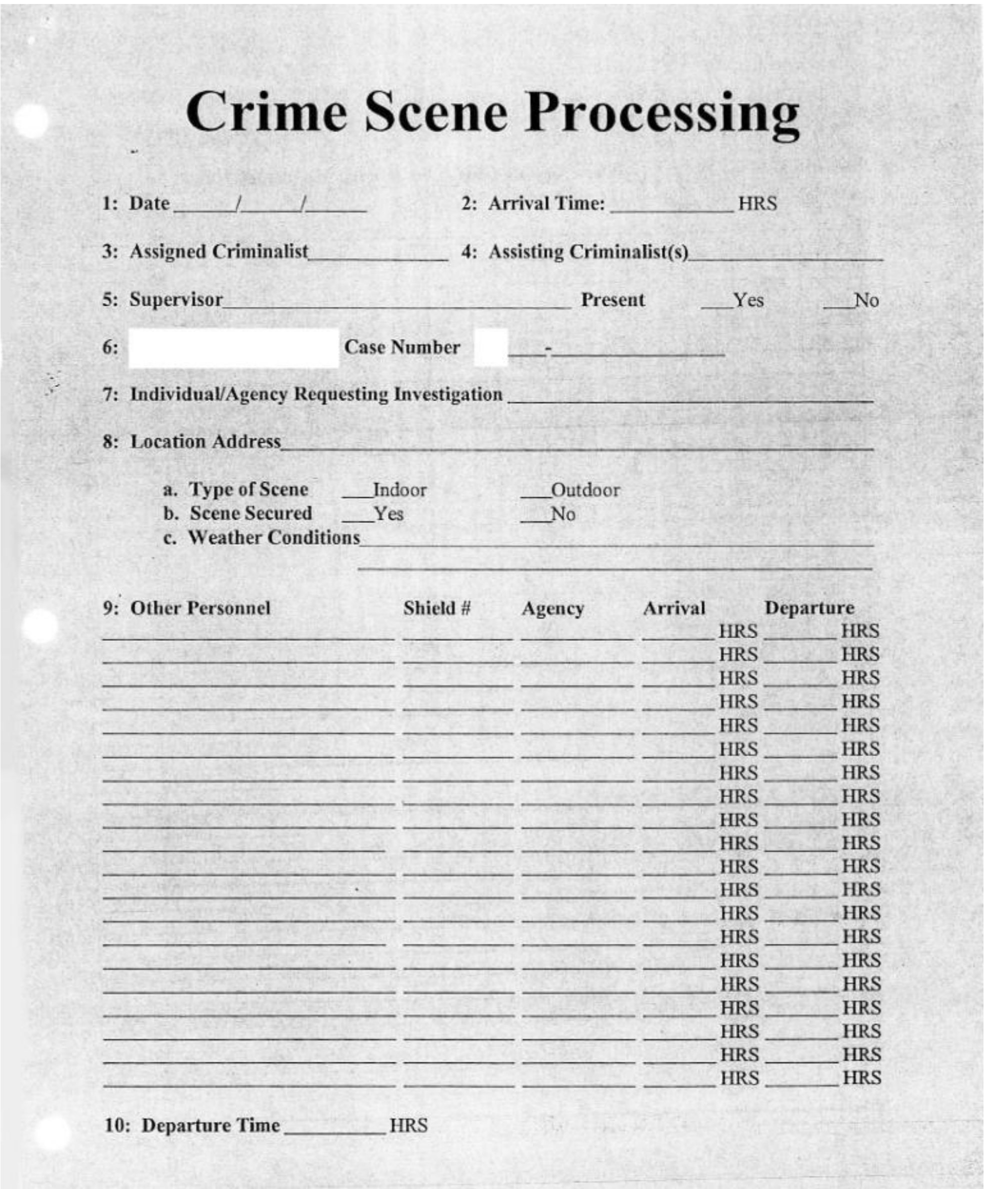

Crime Scene Processing

1: Date _____/_____/_____ **2: Arrival Time:** __________ HRS

3: Assigned Criminalist __________ **4: Assisting Criminalist(s)** __________

5: Supervisor ____________________ **Present** ___Yes ___No

6: [redacted] **Case Number** [redacted] ___-__________

7: Individual/Agency Requesting Investigation ____________________

8: Location Address ____________________

a. Type of Scene ___Indoor ___Outdoor
b. Scene Secured ___Yes ___No
c. Weather Conditions ____________________

9: Other Personnel	**Shield #**	**Agency**	**Arrival**	**Departure**
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS
			HRS	HRS

10: Departure Time __________ HRS

Figure 1.7 An example of a log sheet for documenting authorized personnel at a crime scene. (© Richard C. Li.)

fleeing the scenes. Evidence that provides information on the MO is also vital to an investigation. A distinct MO can establish a *case-to-case linkage* for serial offender cases.

1.3 Searches

Some investigations require a search for specific items of evidence such as biological stains, human remains, and all relevant evidence. A search usually has a specific purpose. Thus, the use of search patterns can be helpful, especially in cases involving large outdoor crime scenes.

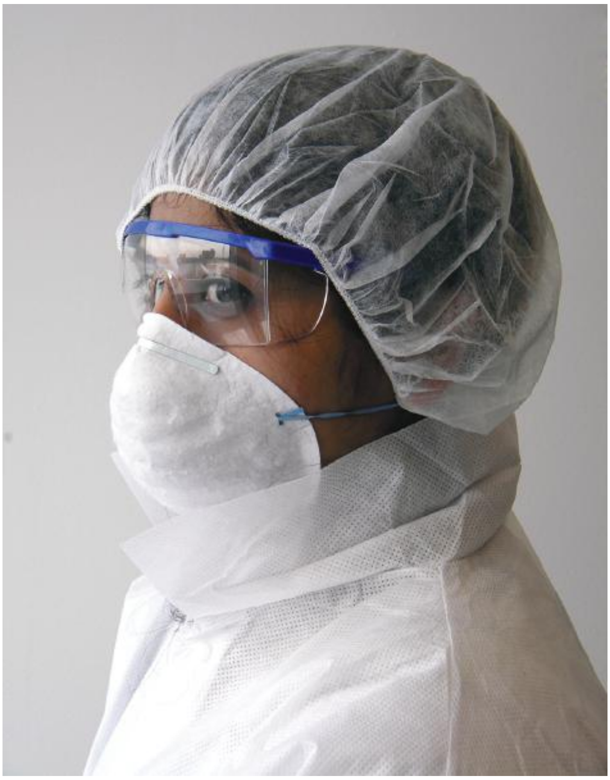

Figure 1.8 Personal protection wear and devices that are used at crime scenes. (© Richard C. Li.)

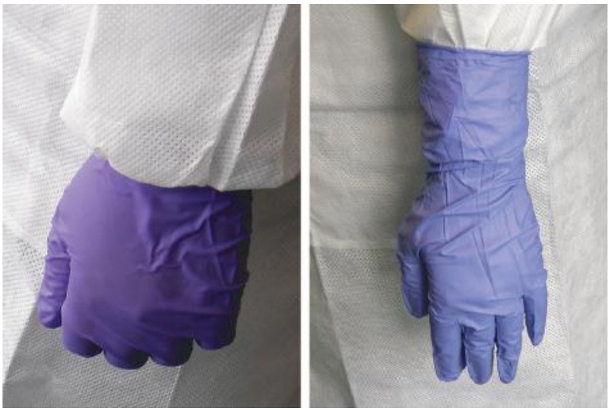

Figure 1.9 Disposable glove (left) and glove with extended cuff (right). (© Richard C. Li.)

Search patterns may include a grid, line, or zone (Figures 1.16 and 1.17). The method that is ultimately used depends on the type and size of the scene (Figures 1.18 through 1.20). Additionally, the points of entry and exit and the paths followed by a perpetrator should also be searched.

Searching for biological stains usually utilizes devices such as an alternate light source (ALS); see Figures 1.21 and 1.22. An ALS either produces a single specific wavelength of light or a desired wavelength by using specific filters. Biological materials such as blood, semen, and saliva emit fluorescent light under an ALS, which can facilitate the locating of biological materials. Additionally,

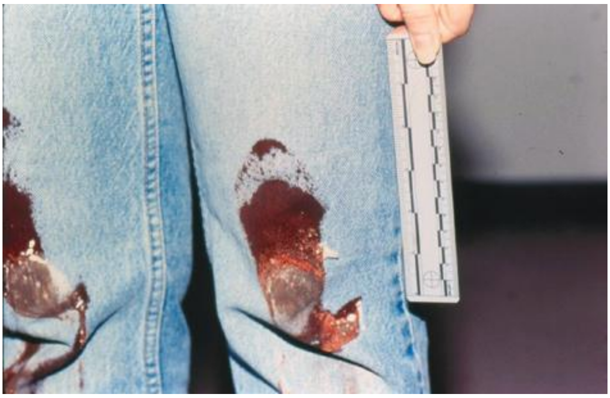

Figure 1.10 Photographic documentation of bloodstains on clothing. (© Richard C. Li.)

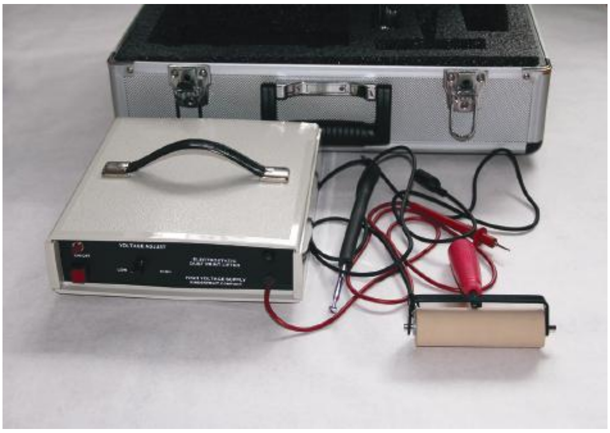

Figure 1.11 An electrostatic dust print lifting device can be utilized for processing impression evidence such as footprints and tire tracks. (© Richard C. Li.)

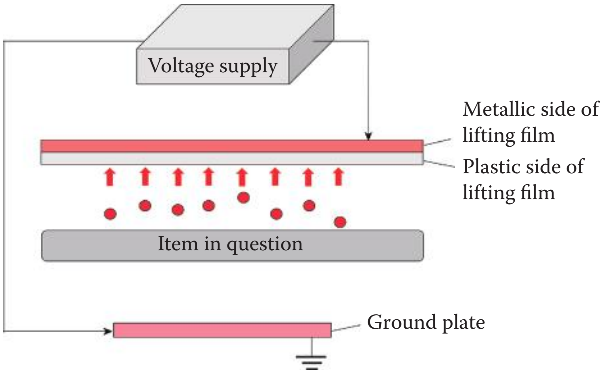
Voltage supply
Metallic side of lifting film
Plastic side of lifting film
Item in question
Ground plate

Figure 1.12 Basic components of an electrostatic dust print lifting device. The lifting film is placed on top of the item in question with the plastic side against the surface and the metallic side facing up. A ground plate is placed directly on the ground. The lifting film and ground plate are connected to the voltage supply apparatus. Once the charging voltage is turned on, the static charge transfers the dust particles from the surface to the plastic side of the lifting film. (© Richard C. Li.)

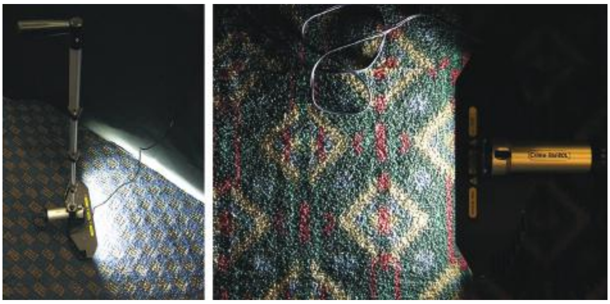

Figure 1.13 A high-intensity light-emitting diode (LED) device for locating evidence at a crime scene is particularly effective in highlighting trace evidence such as hairs, fibers, and shoe prints. (© Richard C. Li.)

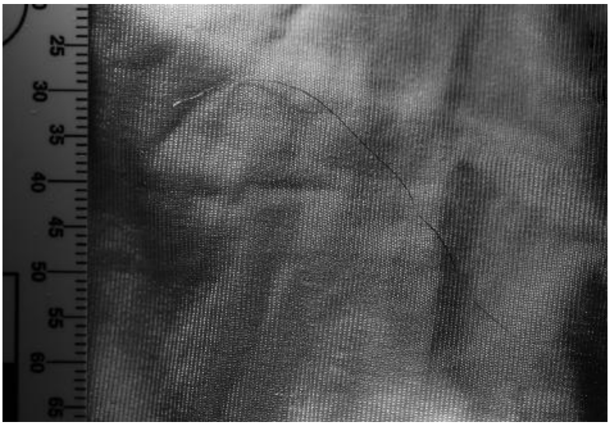

Figure 1.14 A hair found on a victim's clothing can be transferred evidence from a suspect. (© Richard C. Li.)

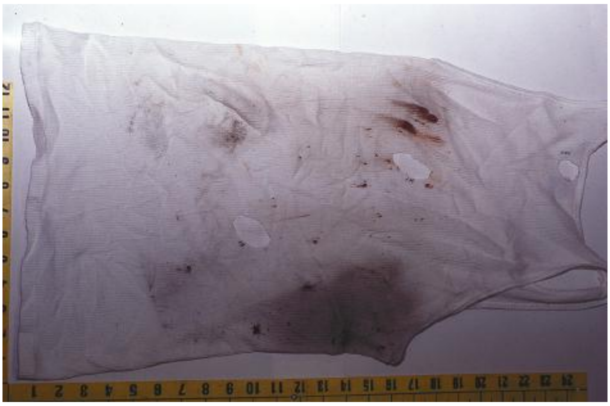

Figure 1.15 Finding a victim's blood on a suspect's clothing can establish a link between them. (© Richard C. Li.)

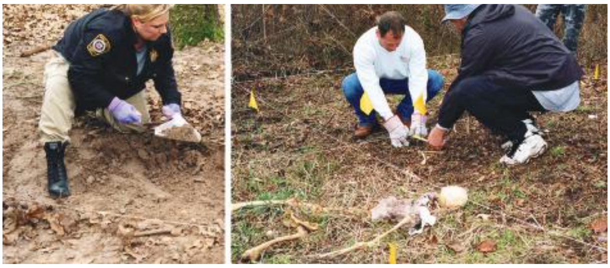

Figure 1.16 Grid search pattern for an outdoor scene. The investigators and anthropologists present are searching for human bone evidence within the grid. (Courtesy of H. Brewster.)

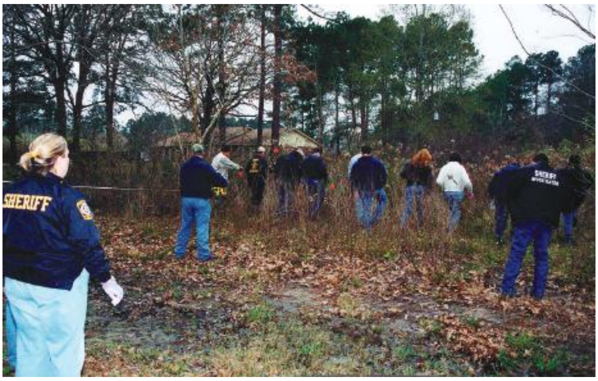

Figure 1.17 Line search pattern for an outdoor scene. (Courtesy of H. Brewster.)

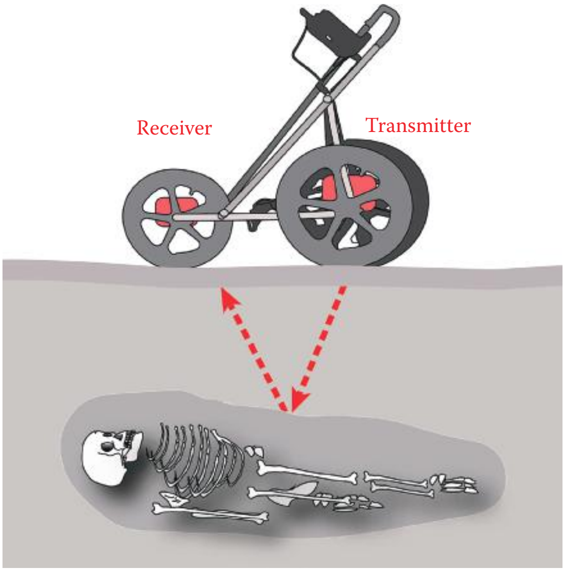

Figure 1.18 Using ground-penetrating radar (GPR) to locate clandestine graves of homicide victims. GPR uses electromagnetic waves emitted from a transmitter, which are detected by a receiver to locate clandestine burials and buried objects such as weapons embedded in soils. Images of the potential evidence are typically obtained by moving the antenna of the GPR device over the surface of the ground. (© Richard C. Li.)

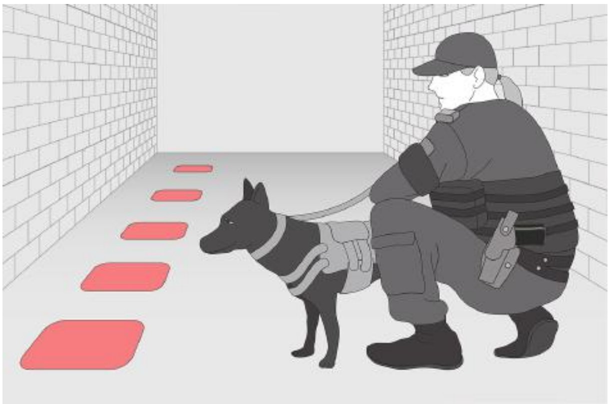

Figure 1.19 Using cadaver-sniffing dogs to alert investigators to the presence of buried bodies. The odor produced by the decomposition of the human body may be sensed by cadaver-sniffing dogs. Odor-absorbing pads absorb and retain the scent of decomposed remains and can be placed at the scene for several days. Upon sniffing the pads, a cadaver dog may indicate the presence of buried human remains or may indicate that human remains were once buried in that location. However, it is not clear that this technique is reliable. (© Richard C. Li.)

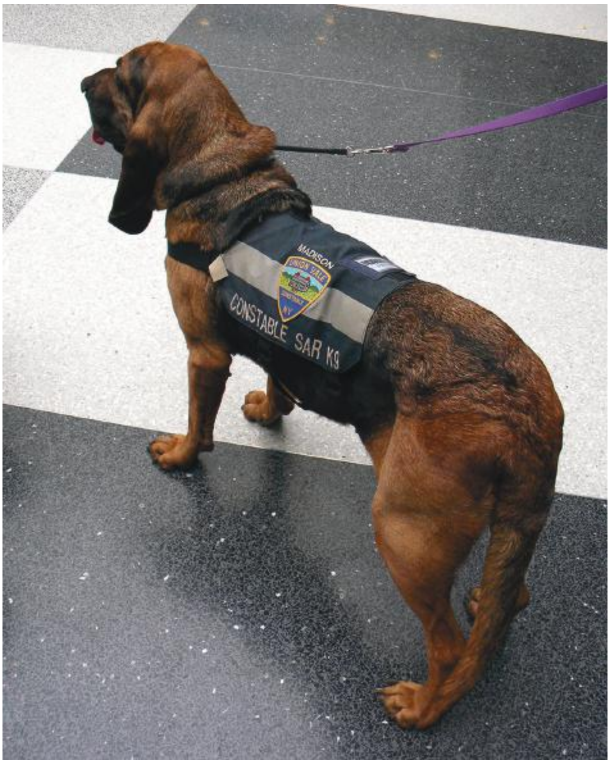

Figure 1.20 Tracking dog for searching suspects. In a situation of processing a recent crime scene when the suspect may be in close vicinity to the scene, a tracking dog can be used. A tracking dog, such as a bloodhound, can potentially follow the scent from items left at the scene to locate a suspect nearby. (© Richard C. Li.)

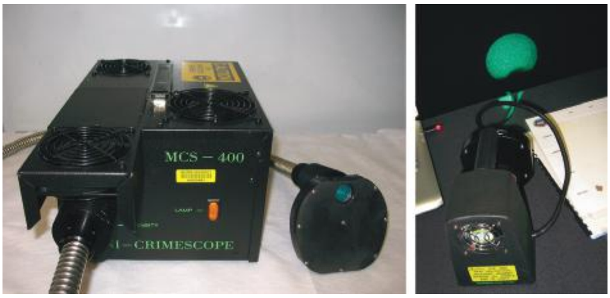

Figure 1.21 Compact alternate light source devices, which are intended specifically for use at crime scenes, reduce search time and improve the recovery of evidence such as biological evidence and chemically enhanced latent fingerprints. (© Richard C. Li.)

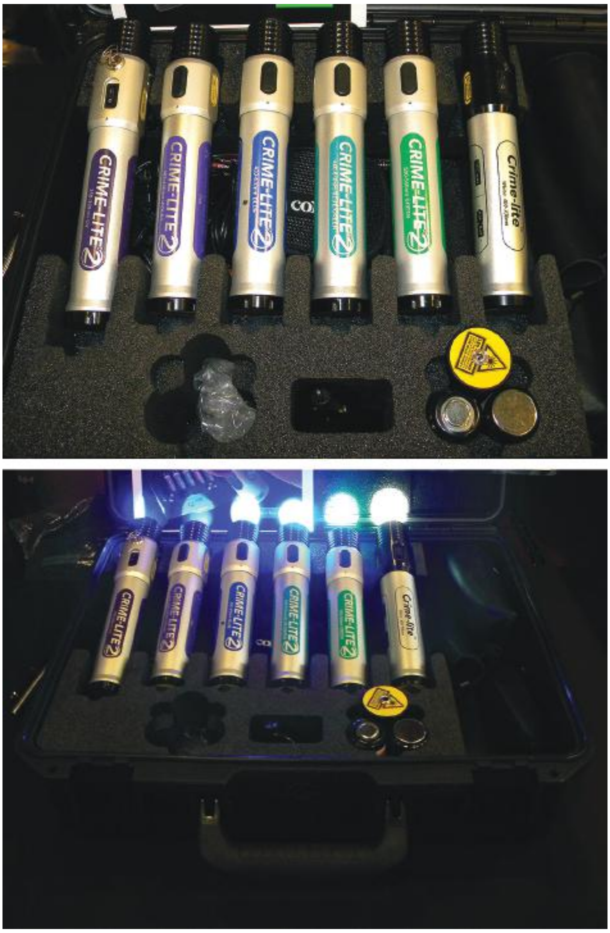

Figure 1.22 LED light sources provide illumination for locating evidence such as bodily fluid stains, hair, or fibers. (© Richard C. Li.)

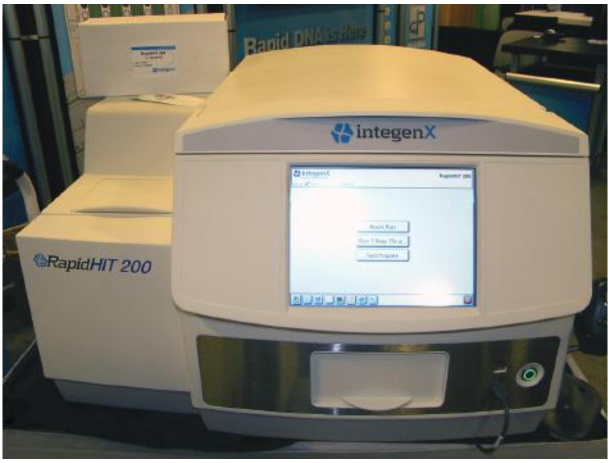

Figure 1.23 An example of a compact Rapid DNA device for processing DNA evidence in the field. (© Richard C. Li.)

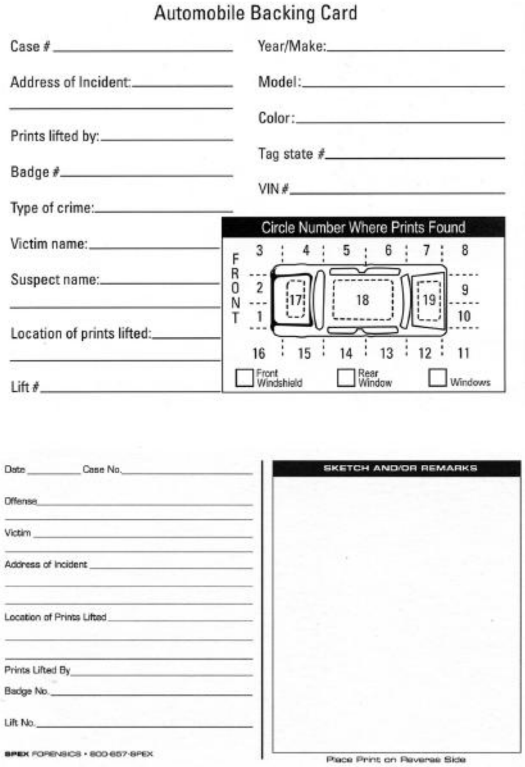

Automobile Backing Card

Case # __________
Address of Incident: __________
Prints lifted by: __________
Badge # __________
Type of crime: __________
Victim name: __________
Suspect name: __________
Location of prints lifted: __________
Lift # __________

Year/Make: __________
Model: __________
Color: __________
Tag state # __________
VIN # __________

Circle Number Where Prints Found

FRONT
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

☐ Front Windshield ☐ Rear Window ☐ Windows

Date ______ Case No. __________
Offense __________
Victim __________
Address of Incident __________
Location of Prints Lifted __________
Prints Lifted By __________
Badge No. __________
Lift No. __________

SPEX FORENSICS • 800-657-SPEX

SKETCH AND/OR REMARKS

Place Print on Reverse Side

Figure 1.24 An example of sketch documentation. (© Richard C. Li.)

field tests and enhancement reagents can be used to facilitate crime scene searching (Chapter 12). These reagents can detect and identify biological evidence. The tests are very simple, rapid, and sensitive, and thus can be used at crime scenes. For example, phenolphthalin and leucomalachite green tests can be used for detecting blood evidence. Sometimes, minute amounts of blood may be

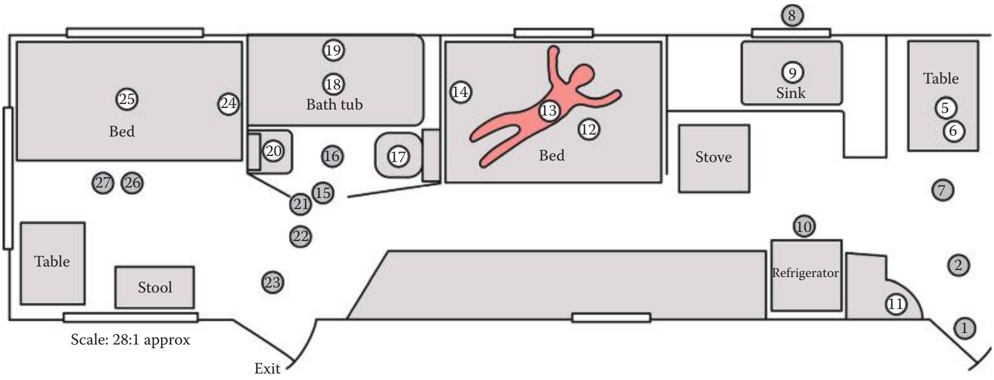

Figure 1.25 Sketches for documenting the sites of fingerprints. (© Richard C. Li.)

present at the scene as a result of attempts to clean up blood prior to the investigation. These stains may not be visible with the naked eye. Enhancement reagents such as luminol and fluorescein, which emit chemiluminant and fluorescent light upon reacting with certain biological materials, respectively, can be used. Additionally, the enhancement reagents can detect faint blood-containing pattern evidence such as faint bloody fingerprints, footprints, and other pattern evidence of physical contact such as drag marks in blood. However, precaution should be taken since these reagents are not usually very specific to blood. Certain substances such as bleach, various metals, and plants may also lead to chemical reactions with the field tests and the enhancement reagents. In these cases, the evidence collected is further tested with laboratory examination and analysis.

Recently, portable and field-deployable instruments have been developed that are capable of processing buccal swabs and potentially other evidence to produce a DNA profile on-site (Figure 1.23). It is a fully automated process, using the Rapid DNA technology (Chapter 8), that

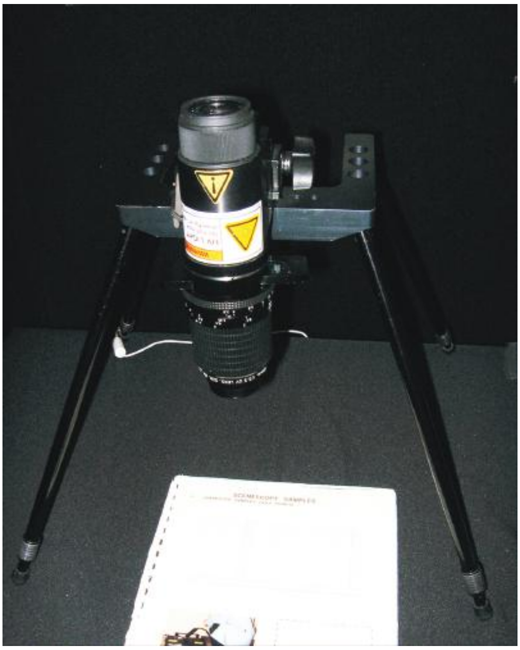

Figure 1.26 A reflected ultraviolet imaging system (RUVIS) imager for documenting close-up views of evidence such as latent fingerprints. (© Richard C. Li.)

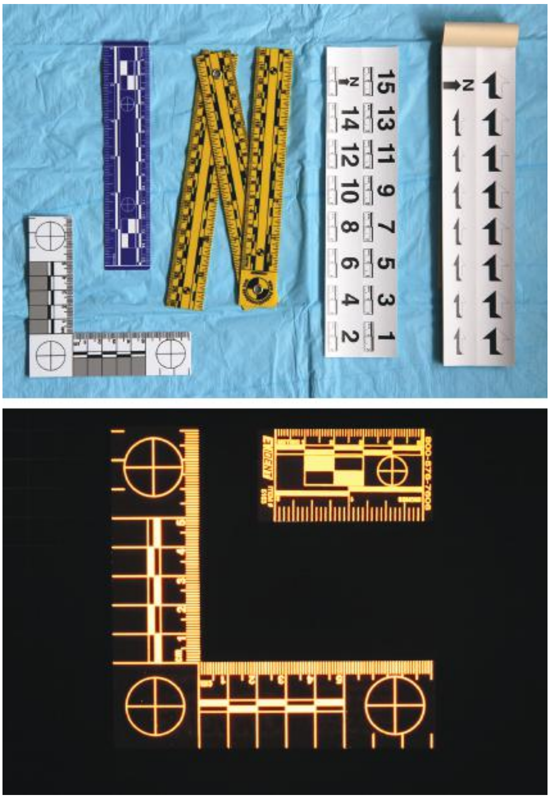

Figure 1.27 Scales for photographic documentation. Regular scales (top) can be used with visible light sources. Fluorescence scales (bottom) can be used for the documentation of fluorescing evidence after certain treatment, such as chemical enhancement, using alternate light sources. (© Richard C. Li.)

can be completed within 2 h by a trained crime scene investigator or police officer. These instruments may provide a new tool for expediting the identification of suspects and developing investigative leads at the scene. Additionally, this technology can enable law enforcement agents to rapidly determine whether the crimes were isolated incidents or part of serial crimes committed by the same offender, such as in serial burglary and arson cases. It can also be used in the identification of human remains in mass disasters.

1.4 Documentation

The conditions at a crime scene, including both the individual items of evidence and the overall scene, must be documented to provide vital information for investigators and for the courts. The most common documentation methods are drawing sketches and taking photographs and videographs. The sketch is to reflect the positions and the spatial relationships of items and persons with measurements using a scale. An investigator usually prepares a rough sketch first and later converts the rough sketch into a finished sketch (Figures 1.24 and 1.25). If bloodstains are present at the scene, the location of bloodstain patterns should be emphasized. Prior to handling and moving evidence, photographs should be taken with different views: an overall view of the entire scene, a medium-range view showing the positions and the relationships of items, and a close-up

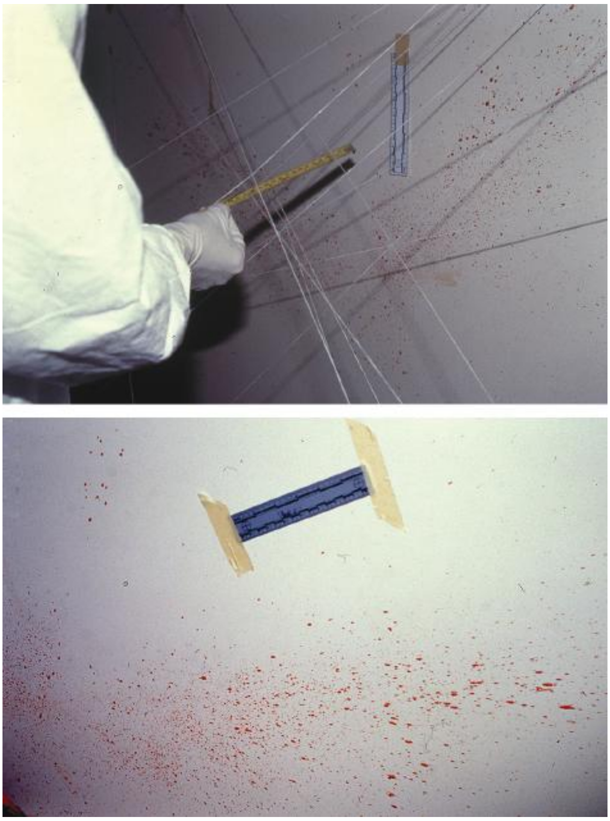

Figure 1.28 Photographic documentation of bloodstain patterns. (© Richard C. Li.)

view showing details of the evidence (Figure 1.26). Photographs should also include a measuring device such as a scale (Figure 1.27) to accurately depict the sizes of items such as bloodstains or bite marks. This can be achieved simply by placing a ruler adjacent to the evidence when it is photographed (see Figure 1.28). A photograph log sheet can be used to record the chronological order of crime scene photographs and to note filming conditions and any additional relevant information (Figure 1.29). Similar documentation should be prepared for videographs when appropriate. Additionally, written or audio-recorded notes can be used. Notes should include complete and accurate information of a crime scene investigation, such as the case identifier number, the identities of the investigators, and a description of the scene or items (e.g., location, size, and shape). Additionally, any disturbance of evidence occurring during crime scene processing should be noted.

1.5 Chain of Custody

Custody information should be recorded at each event when evidence is handled or transferred by authorized personnel. Usually, a custody form listing a specific evidence item is used to document the chain (Figure 1.30). Each individual who acquires custody of the evidence must sign a chain of custody document. An incomplete chain of custody may lead to an inference of possible tampering or contamination of evidence. As a result, the evidence may not be admissible in court.

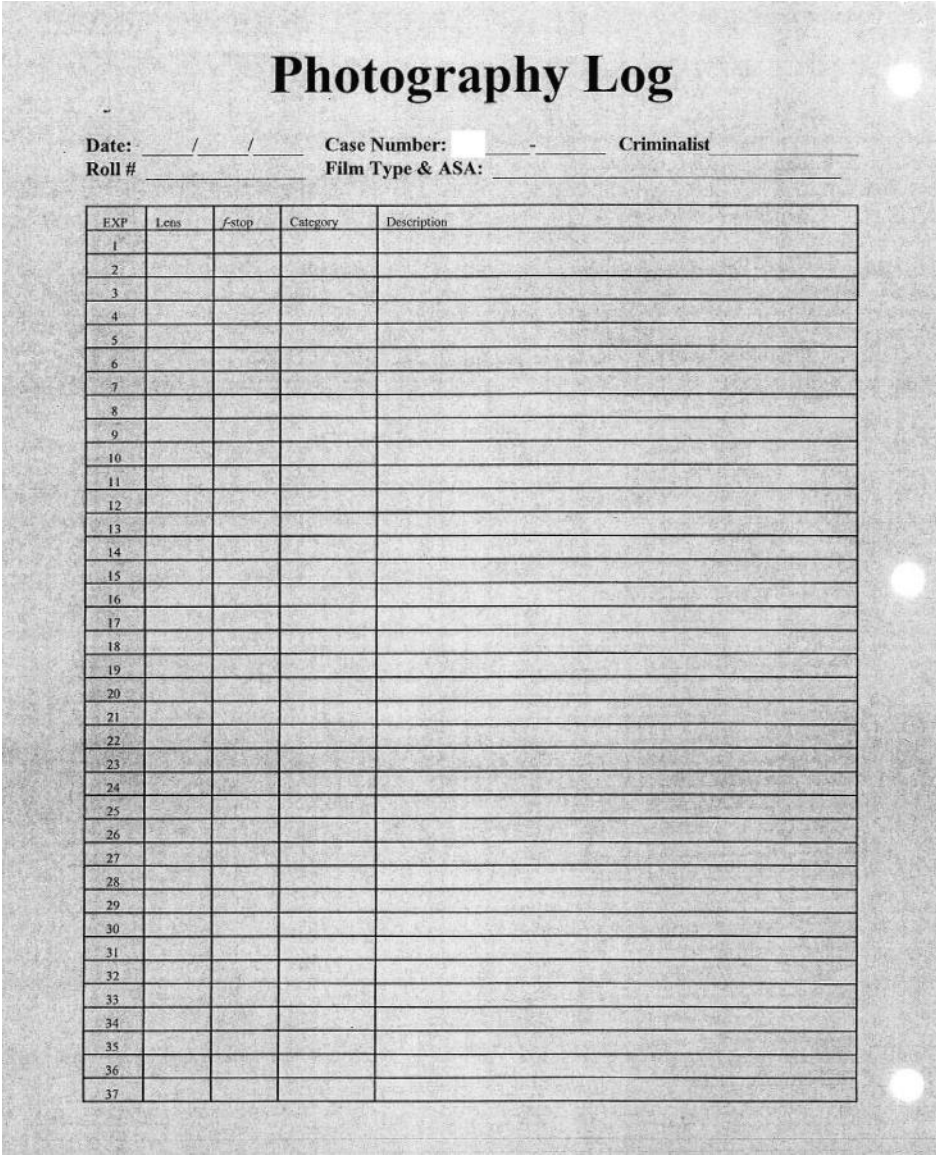

Photography Log

Date: ____/____/____ **Case Number:** [redacted] ____-____ **Criminalist** ________

Roll # ________ **Film Type & ASA:** ________

EXP	Lens	*f*-stop	Category	Description
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				

Figure 1.29 Crime scene photography log. (© Richard C. Li.)

1.6 Collection of Biological Evidence

After the crime scene documentation is completed, the collection of evidence can be initiated. Small or portable items, such as bloodstained knives, can be collected and submitted to a crime laboratory (Figure 1.31). Large or unmovable items of evidence (Figure 1.32) can be collected and submitted in sections, such as a section of wall where bloodstains are located. Table 1.1 and Figures 1.33 through 1.38 summarize and illustrate representative collection techniques. Specific care is required for the collection of biological evidence in the following situations:

- *Bloodstain pattern evidence*: It is especially important to thoroughly document the bloodstain pattern evidence at a crime scene prior to collection. Bloodstain patterns can be especially useful in crime scene reconstruction.

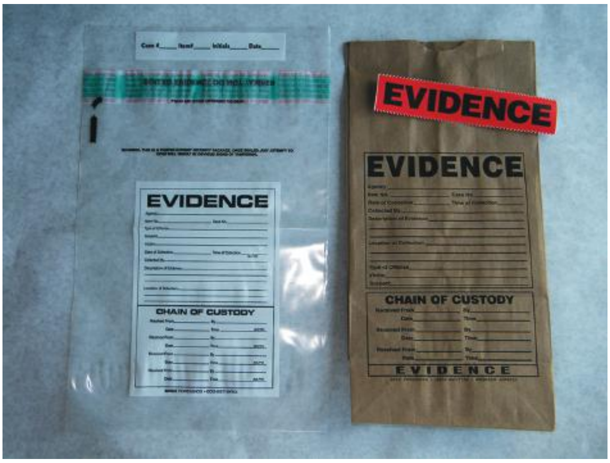
EVIDENCE
CHAIN OF CUSTODY
EVIDENCE
EVIDENCE
CHAIN OF CUSTODY
EVIDENCE

Figure 1.30 Labels with the chain of custody that are used for marking the evidence contained in the packaging. (© Richard C. Li.)

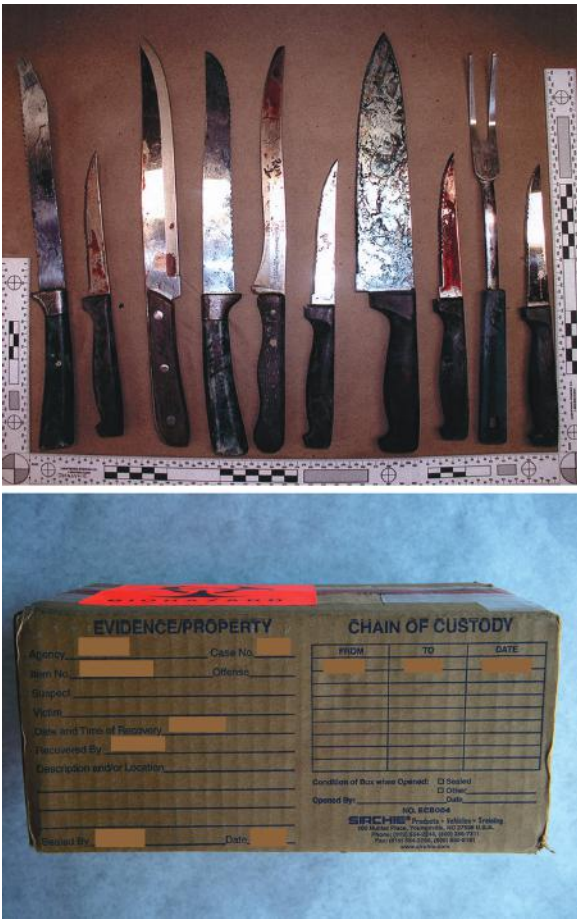
EVIDENCE/PROPERTY
Agency
Case No.
Item No.
Offense
Suspect
Victim
Date and Time of Recovery
Recovered By
Description and/or Location
Sealed By
Date
CHAIN OF CUSTODY
FROM
TO
DATE
SIRCHIE

Figure 1.31 Handling sharp objects. Bloodstained knives collected and submitted to laboratories (top) and a box for packing sharp objects (bottom). (© Richard C. Li.)

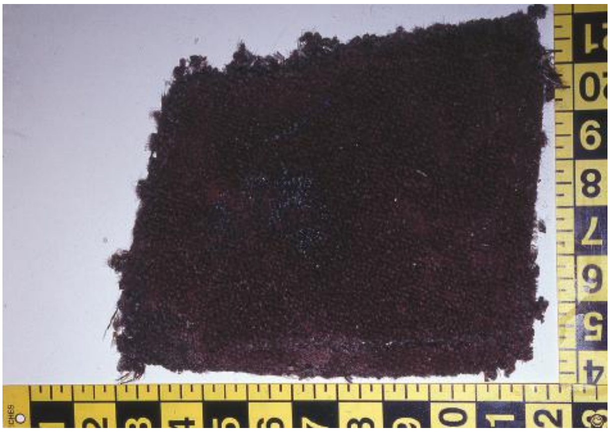

Figure 1.32 A section of bloodstained carpet is collected. (© Richard C. Li.)

- *Multiple analysis of evidence*: If multiple analyses are needed for a single item of evidence, nondestructive analyses should be carried out first. For example, a bloody fingerprint should be collected for ridge detail analysis prior to collecting blood for DNA analysis.
- *Trace evidence*: Trace evidence such as hairs and fibers can be present in bloodstained evidence and should be identified and properly collected.
- *Control samples*: Control (known or blank) samples should be collected from a control area (e.g., unstained area near a collected stain).
- *Size of stain*: Polymerase chain reaction (PCR)-based forensic DNA techniques are highly sensitive and allow for the successful analysis of very small bloodstains. All bloodstains, even if they are barely visible, should be collected at a crime scene.
- *Wet evidence*: Wet evidence should be air-dried (without heat) prior to packaging to prevent the degradation of proteins and nucleic acids, which are used for forensic serological and DNA analysis.

1.7 Marking Evidence

The marking of evidence is necessary for identification purposes so that it can be quickly recognized even years later (Figure 1.39). An investigator's initials, the item number, and the case number are usually included in marking. Information can be marked on a tag, a label attached to the item, or directly on garment evidence. The marking of evidence should not be proximal to bullet holes or biological stains to prevent the mark from interfering with analyses.

1.8 Packaging and Transportation

Packaging is intended to protect and preserve evidence. All evidence should be secured and protected from possible contamination. Fragile items should be protected to prevent any damage during transportation. Exposure to heat and humidity should be avoided to protect biological evidence from degradation during transport. Various packaging methods are available

Table 1.1 Methods for Collecting Biological Evidence

Type of Evidence	Condition	Method of Collection	Procedure
Blood	Dry	Swab	Best on nonabsorbent surfaces; lightly moisten sterile swab with distilled or sterile water, rub over stain while rotating; allow to air-dry; a combination of first a moistened swabbing followed by second a dry swabbing (both swabs submitted) is recommended
		Cutting	Cut stain from item
		Scraping	Scrape bloodstain into a clean piece of paper using a clean blade; wrap sample using druggist's fold
		Lifting	Works for nonabsorbent surfaces; use fingerprint lifting tape that does not interfere with DNA testing to lift stain; lifted stain should be covered with a piece of lifter's cover
		Collect entire item	Collect if item contains bloodstain pattern; difficult to swab; requires multiple exams
	Wet	Swab	Absorb blood sample onto sterile cotton swabs; stain should be concentrated on tip and allowed to air-dry
		FTA paper	Use a sterile disposable pipet to collect liquid blood; spot on FTA paper; allow to air-dry
	Reference liquid blood sample	Venous blood collection	Collect blood in a purple-topped vacutainer tube containing ethylenediaminetetraacetic acid (EDTA) anticoagulant; refrigerate but never freeze
Semen	Dry or wet	See "Blood" above	See "Blood" above
	Condom	Collect entire item	Secure condom with a tie and place in a container in a refrigerator; submit to laboratory as soon as possible
	Various conditions	Victim sexual assault kit	Standardized kit to collect biological evidence from the body of a victim includes swabs, microscope slides, and envelopes
	Various conditions	Suspect standard kit	Standardized kit to collect biological evidence from the body of a suspect includes swabs, microscope slides, and envelopes

(*continued*)

Table 1.1 (Continued) Methods for Collecting Biological Evidence

Type of Evidence	Condition	Method of Collection	Procedure
Victim vaginal fluid	Dry or wet	See "Blood" above	Often collected from a suspect's pubic area or fingers
Saliva	Dry or wet	See "Blood" above	Often collected from a bite mark or a victim's pubic area
	Reference saliva (buccal) samples	Swab	Swab the inside of the cheek using two swabs, rotating them during collection; allow swabs to air-dry
		Filter paper	Place donor saliva sample on marked area of filter paper; allow to air-dry
Hairs	Head and pubic hairs	Lifting	Refer to dry bloodstain lifting method
		Transfer	Use forceps to transfer hair onto a piece of paper that can be folded
		Vacuum	Vacuum can be used if necessary; generally not recommended
Fingernails and scrapings	Various conditions	Clipping	Use a clean clipper to clip nails onto clean paper; wrap samples using druggist's fold
		Scraping	Scrape undersides of nails onto clean paper; wrap samples using druggist's fold
Bones	Various conditions	Freeze in container	Collect if blood is not available; collect bone with marrow if available; rib bone and vertebrae are preferred; place specimen in a container and freeze if wet
Teeth	Various conditions	Container	Collect teeth with dental pulp if possible into a container
Tissues and organs	Wet	Freeze in a container	Collect if blood is not available; place specimen in a container and freeze
DNA database samples	Various conditions	Various	Follow jurisdiction protocol for collecting samples from arrested individuals or convicted offenders or both

Source: Fisher, B., *Techniques of Crime Scene Investigation*, 7th edn., 2004, CRC Press, Boca Raton; Lee, H.C. et al., *Henry Lee's Crime Scene Handbook*, 2001, Academic Press, San Diego; National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002, US Department of Justice, Office of Justice Programs, Washington, DC.

Note: Samples requiring refrigeration or freezing are noted; other samples may be stored at room temperature. Dry evidence should be packed in a porous material such as paper (envelope, bag, box) as described in the text.

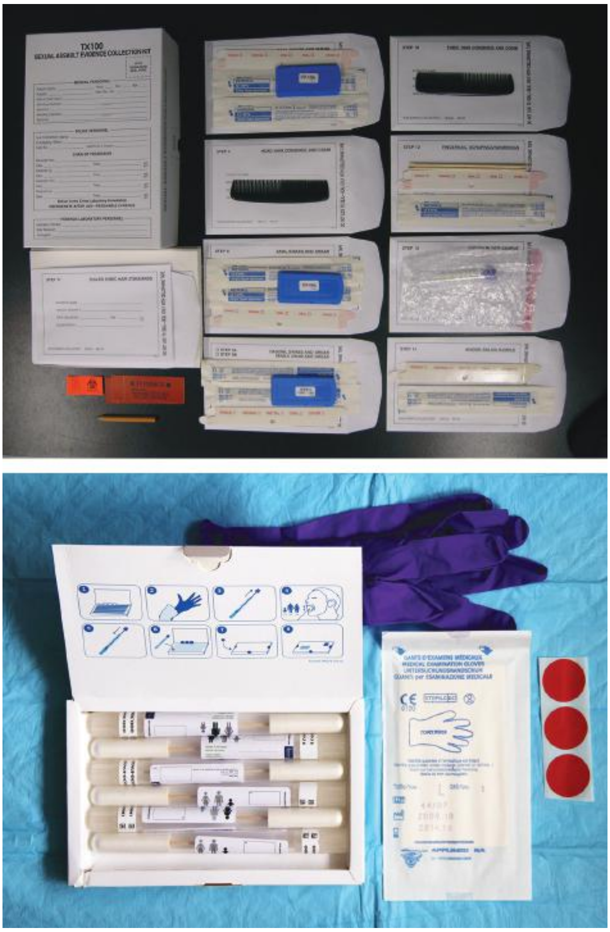
TX100
SEXUAL ASSAULT EVIDENCE COLLECTION KIT

Figure 1.33 Evidence collection kits. Sexual assault evidence collection kit (top). Paternity evidence collection kit (bottom). (© Richard C. Li.)

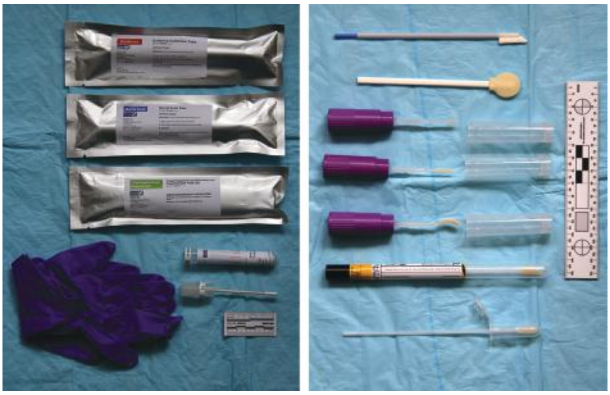

Figure 1.34 Various types of swabs that are used for collecting biological evidence. (© Richard C. Li.)

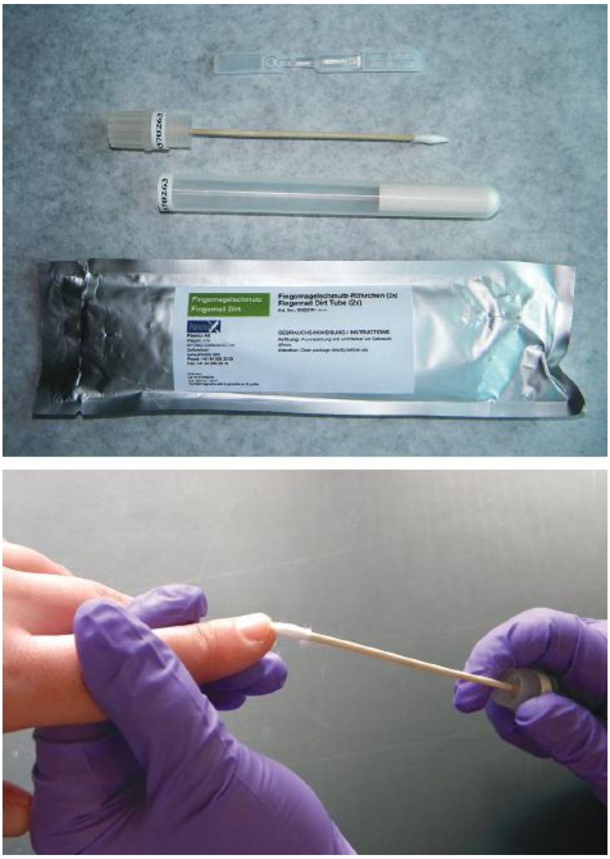

Figure 1.35 Fingernail swabbing for recovering evidence. Fingernail swabs are often collected from individuals who are involved in a struggle in violent crimes and in digital penetration in sexual assault cases. Fingernail swabs are sampled separately from both hands. (© Richard C. Li.)

depending on the type of evidence handled (Figures 1.40 through 1.42). The following are general considerations related to the packaging of evidence:

- *Evidence from different sources*: To prevent the transfer of evidence from different sources, items of evidence should not be grouped in a single package. However, evidence may be packed in a single container if the items were found together.
- *Folding of evidence*: Folding of clothing, especially items with wet bloodstains, can transfer evidence from one part of a garment to another. If a large, dry garment must be folded, a piece of clean paper should be placed between different parts of the garment to avoid direct contact between the different parts of the garment, thereby preventing the transfer of evidence.
- *Packing materials*: Envelopes, bags, and boxes that are made of porous materials such as paper are appropriate for packaging dry biological evidence. Dry, bloodstained evidence should not be sealed in plastic bags or containers that trap moisture.
- *Liquid evidence*: Tubes containing liquid such as blood should not be frozen because the volume of a liquid expands in freezing temperatures and this expansion may lead to cracking. Tubes should be placed in plastic bags to prevent leaks in case of

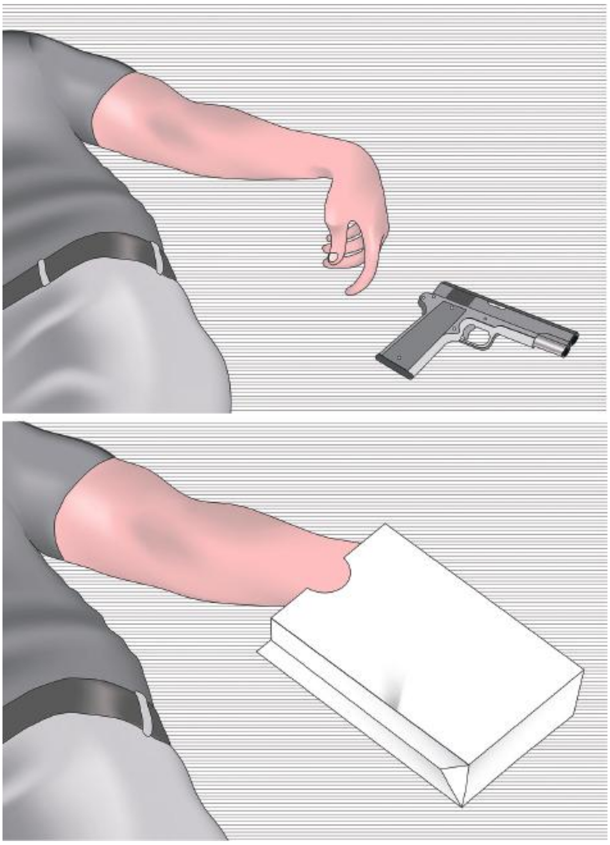

Figure 1.36 Hand bags for protecting the hands of a decedent in an alleged suicide. Gunshot residue can be found on hands after firing a weapon (top). The bagging of the hands, using paper bags and wide rubber bands, prevents the loss of gunshot residue during the transportation of the body (bottom). (© Richard C. Li.)

accidental breakage. Liquid evidence should be transported and submitted to a laboratory as soon as possible after the collection of evidence.

- *Trace evidence*: All such evidence should be wrapped in paper with a druggist's fold (Figure 1.43). The wrapped trace evidence can be packed in an envelope.

Packaged evidence should be properly labeled with a description of the evidence and sealed with evidence tape. It is important for the person packaging the evidence to initial and date across the seal to show authenticity (Figure 1.44). A seal should not be cut when a sealed evidence bag is opened. Instead, an opening can be created by cutting at an area distal from the existing seal. After analysis is complete, the evidence packaging should be resealed. Table 1.1 summarizes additional steps for packaging evidence.

1.9 Final Survey and the Release of the Crime Scene

During a final survey, a discussion with all personnel in the crime scene investigation team should be carried out to thoroughly review all aspects of the search. It is important to ensure

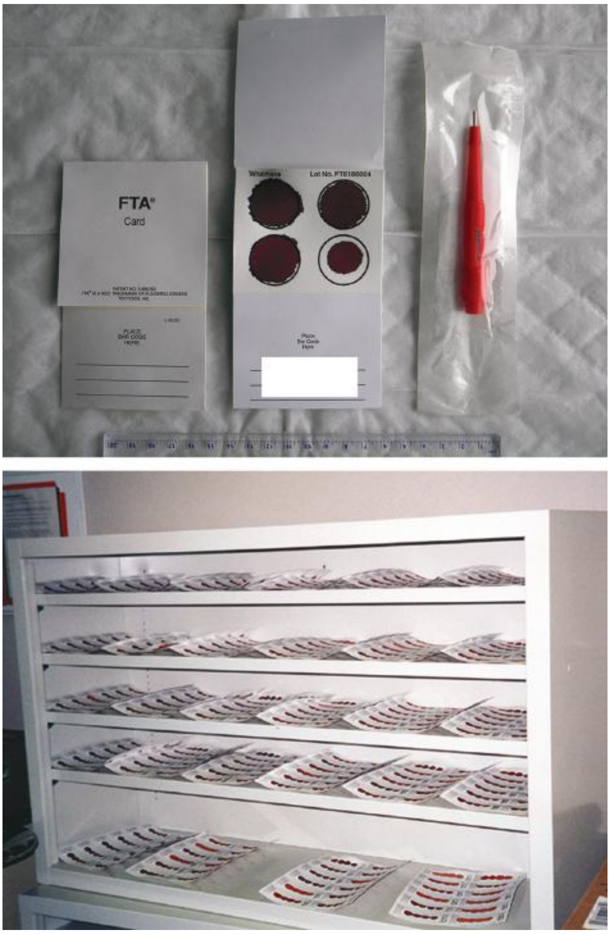

Figure 1.37 Blood cards are typically used for collecting blood evidence from a known source such as a suspect or a victim (top). A manual hole punch can be used to create a blood card punch for DNA extraction. Blood samples are air-dried on blood cards for storage (bottom). (© Richard C. Li.)

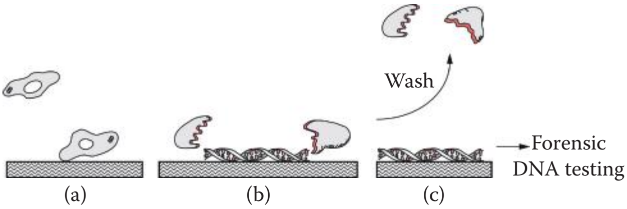

Figure 1.38 Application of Flinders Technology Associates (FTA) filter paper for the collection of biological evidence. (a) Biological fluid with cells is applied to FTA paper. (b) Cells are lysed and DNA is immobilized on FTA paper. (c) Cellular materials are washed away and DNA remaining on the FTA paper can be used for forensic testing. (© Richard C. Li.)

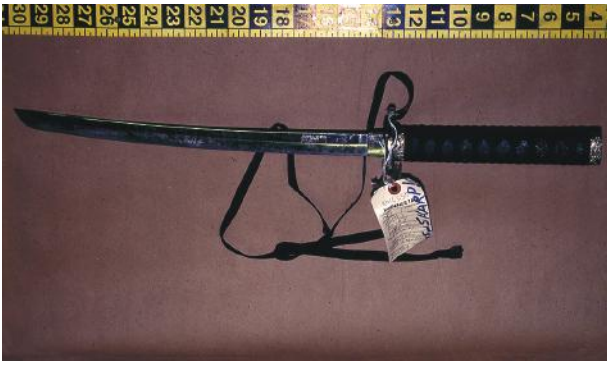

Figure 1.39 Photographic documentation of a knife. Note the evidence tag. (© Richard C. Li.)

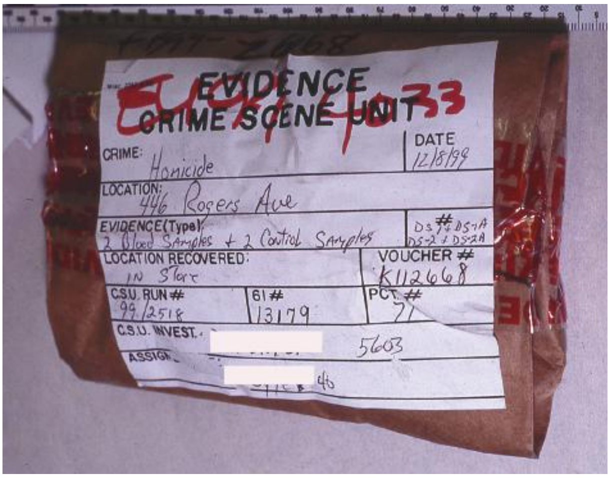

EVIDENCE
CRIME SCENE UNIT

CRIME: Homicide
DATE 12/8/99
LOCATION: 446 Rogers Ave
EVIDENCE(Type): 2 Blood Samples + 2 Control Samples
DS # DS-1 + DS-1A DS-2 + DS-2A
LOCATION RECOVERED: in Store
VOUCHER # K112668
C.S.U. RUN # 99/2518
61 # 13179
PCT. # 77
C.S.U. INVEST. [redacted] 5603
ASSIGN [redacted] 40

Figure 1.40 Evidence containing dried bodily fluid stains is packed in a paper bag. (© Richard C. Li.)

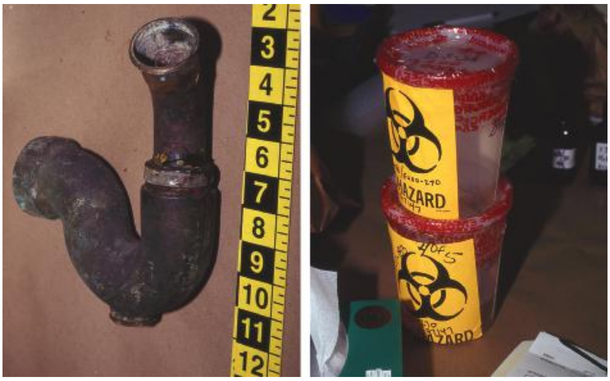

Figure 1.41 Alleged diluted blood collected from a pipe (left) and placed in plastic containers (right). (© Richard C. Li.)

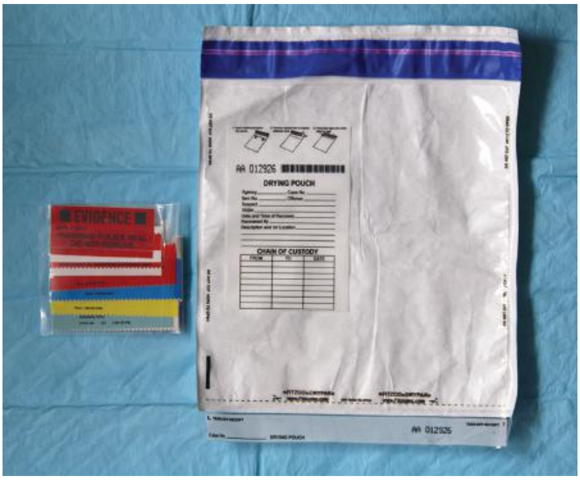
EVIDENCE
AA 012926
DRYING POUCH
CHAIN OF CUSTODY

Figure 1.42 Evidence pouch. The front of the pouch is transparent for viewing the content. The back of the pouch is made of breathable materials allowing wet evidence, such as swabs, to dry inside the pouch. (© Richard C. Li.)

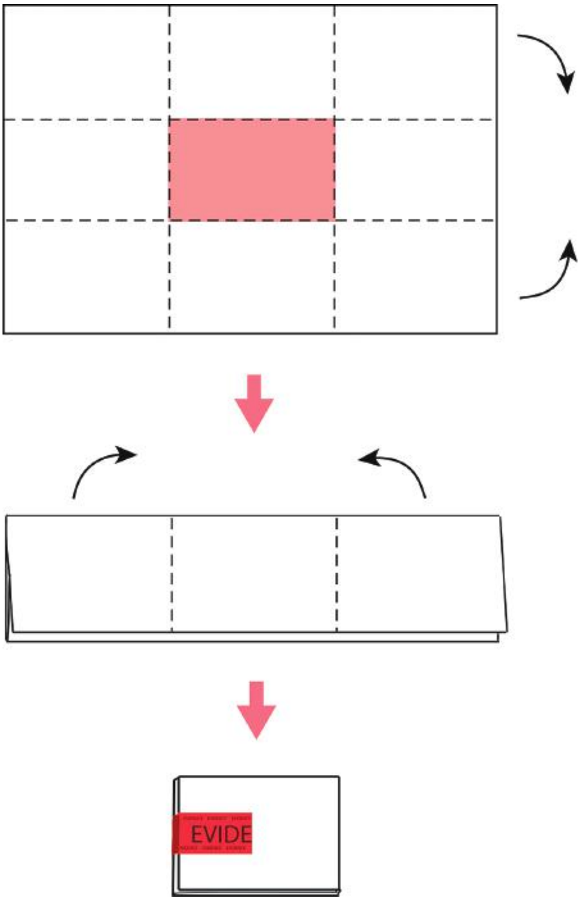

Figure 1.43 An example of a druggist's fold. Trace evidence should be deposited in the center (colored area) of the paper. (© Richard C. Li.)

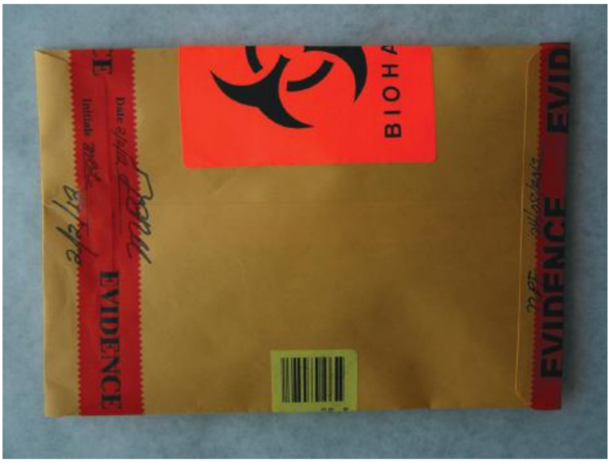

Figure 1.44 Proper marking of sealed evidence. Note that the evidence packaging was cut a second time and resealed at a different location than that of the preexisting seal. (© Richard C. Li.)

that the scene has been searched correctly and completely, and that no area has been missed or overlooked. All documentation including the chain of custody document must be complete and all evidence should be collected, packed, documented, and marked. Photographs of the final condition of the scene should be taken. Once the final survey is completed, the crime scene can be released. Reentry into the crime scene may require a search warrant after the scene is released. Crime scene release documentation usually includes the time and date of release, to whom it is released, and by whom it is released.

1.10 Crime Scene Reconstruction

Crime scene reconstruction is the scientific process of determining the sequence of events and actions that occurred prior to, during, and after a crime. Reconstruction is carried out based on the information from the crime scene observations and the laboratory examination of physical evidence. The overall scientific process in reconstruction usually involves several steps. The process usually begins with the formulation of questions related to the problems that need to be solved. The questions can refer to the explanation of the specifics of the crime, for instance, “Where was the shooter’s position when the shooting occurred?,” “Where was the victim’s position when shot?,” and “What is the muzzle-to-target distance during the shooting?” In order to conduct a thorough crime scene reconstruction, all useful information is collected for review, such as photographs, videotapes, notes, sketches, autopsy reports, and analysis reports of the physical evidence. A hypothesis is then constructed based on the information obtained, which may explain the events and actions involved in a crime. The next step is making predictions that determine the logical consequences of the hypothesis. One or more predictions are selected for testing. The hypothesis is tested by conducting reconstruction experiments. One example of a reconstruction test is bloodstain pattern reconstruction in violent crimes (Chapter 2). Other examples of reconstruction may include trajectory and shooting, glass fracture, and accident reconstruction. The final step is to analyze experimental data and draw a conclusion. The experimental data are analyzed to see if the hypothesis is true or false. Additionally, the interpretation of physical evidence analysis, witness and confession statements, and investigative information should also be considered.

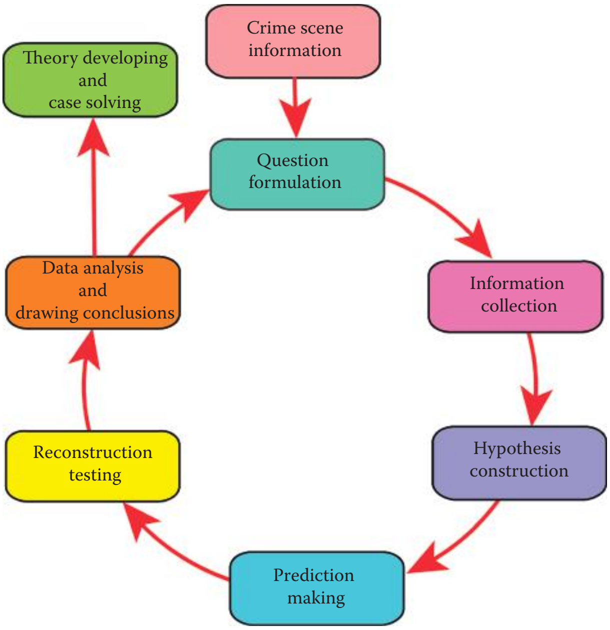
Crime scene information
Theory developing and case solving
Question formulation
Information collection
Data analysis and drawing conclusions
Hypothesis construction
Reconstruction testing
Prediction making

Figure 1.45 The scientific process of crime scene reconstruction. (© Richard C. Li.)

If the results of the experiment are consistent with the hypothesis, a theory can be developed that is intended to provide valuable information to the investigation and future prosecution of a case. Sometimes, forensic scientists may find that the hypothesis is inconsistent with the test results. In that case, an alternative hypothesis needs to be constructed to initiate another reconstruction process. Figure 1.45 illustrates the scientific process of crime scene reconstruction.

Bibliography

Ackermann, K., K.N. Ballantyne, and M. Kayser, Estimating trace deposition time with circadian biomarkers: A prospective and versatile tool for crime scene reconstruction. *Int J Legal Med*, 2010, **124**(5): 387–395.

Adderley, R. and J.W. Bond, The effects of deprivation on the time spent examining crime scenes and the recovery of DNA and fingerprints. *J Forensic Sci*, 2008, **53**(1): 178–182.

Barash, M., A. Reshef, and P. Brauner, The use of adhesive tape for recovery of DNA from crime scene items. *J Forensic Sci*, 2010, **55**(4): 1058–1064.

Benschop, C.C., et al., Post-coital vaginal sampling with nylon flocked swabs improves DNA typing. *Forensic Sci Int Genet*, 2010, **4**(2): 115–121.

Bond, J.W. and C. Hammond, The value of DNA material recovered from crime scenes. *J Forensic Sci*, 2008, **53**(4): 797–801.

Brownlow, R.J., K.E. Dagnall, and C.E. Ames, A comparison of DNA collection and retrieval from two swab types (cotton and nylon flocked swab) when processed using three QIAGEN extraction methods. *J Forensic Sci*, 2012, **57**(3): 713–717.

Buck, U., et al., Accident or homicide—Virtual crime scene reconstruction using 3D methods. *Forensic Sci Int*, 2013, **225**(1–3): 75–84.

Byard, R.W., et al., “Murder-suicide” or “murder-accident”? Difficulties with the analysis of cases. *J Forensic Sci*, 2010, **55**(5): 1375–1377.

Castello, A., F. Frances, and F. Verdu, Solving underwater crimes: Development of latent prints made on submerged objects. *Sci Justice*, 2013, **53**(3): 328–331.

Chen, T., et al., A rapid wire-based sampling method for DNA profiling. *J Forensic Sci*, 2012, **57**(2): 472–477.

Clarke, M., Alleged rape: An appeal case. *J Forensic Leg Med*, 2008, **15**(1): 32–36.

Courts, C., B. Madea, and C. Schyma, Persistence of biological traces in gun barrels—An approach to an experimental model. *Int J Legal Med*, 2011, **126**(3): 391–397.

Crispino, F., Nature and place of crime scene management within forensic sciences. *Sci Justice*, 2008, **48**(1): 24–28.

Durnal, E.W., Crime scene investigation (as seen on TV). *Forensic Sci Int*, 2010, **199**(1–3): 1–5.

Edelman, G., I. Alberink, and B. Hoogeboom, Comparison of the performance of two methods for height estimation. *J Forensic Sci*, 2010, **55**(2): 358–365.

Eldredge, K., E. Huggins, and L.C. Pugh, Alternate light sources in sexual assault examinations: An evidence-based practice project. *J Forensic Nurs*, 2012, **8**(1): 39–44.

Fisher, B., *Techniques of Crime Scene Investigation*, 7th edn., 2004. Boca Raton, FL: CRC Press.

Gokdogan, M.R. and J. Bafra, Development of a sexual assault evidence collection kit—The need for standardization in Turkey. *Nurse Educ Today*, 2010, **30**(4): 285–290.

Gosnell, J., Case three: Collection of evidence in a murder investigation. *Clin Lab Sci*, 2008, **21**(2): 124–125.

Gunasekera, R.S., A.B. Brown, and E.H. Costas, Tales from the grave: Opposing autopsy reports from a body exhumed. *J Forensic Leg Med*, 2012, **19**(5): 297–301.

Hammer, U. and A. Buttner, Distinction between forensic evidence and post-mortem changes of the skin. *Forensic Sci Med Pathol*, 2012, **8**(3): 330–333.

Hollmann, T., R.W. Byard, and M. Tsokos, The processing of skeletonized human remains found in Berlin, Germany. *J Forensic Leg Med*, 2008, **15**(7): 420–425.

Hornor, G., et al., Pediatric sexual assault nurse examiner care: Trace forensic evidence, ano-genital injury, and judicial outcomes. *J Forensic Nurs*, 2012, **8**(3): 105–111.

Hulse-Smith, L. and M. Illes, A blind trial evaluation of a crime scene methodology for deducing impact velocity and droplet size from circular bloodstains. *J Forensic Sci*, 2007, **52**(1): 65–69.

Ingemann-Hansen, O. and A.V. Charles, Forensic medical examination of adolescent and adult victims of sexual violence. *Best Pract Res Clin Obstet Gynaecol*, 2013, **27**(1): 91–102.

Ingemann-Hansen, O., et al., Legal aspects of sexual violence—Does forensic evidence make a difference? *Forensic Sci Int*, 2008, **180**(2–3): 98–104.

Jina, R., et al., Recovering of DNA evidence after rape. *S Afr Med J*, 2011, **101**(10): 758–759.

Kaliszan, M., et al., Striated abrasions from a knife with non-serrated blade—Identification of the instrument of crime on the basis of an experiment with material evidence. *Int J Legal Med*, 2011, **125**(5): 745–748.

Komar, D.A., S. Lathrop, and S. American, The use of material culture to establish the ethnic identity of victims in genocide investigations: A validation study from the American Southwest. *J Forensic Sci*, 2008, **53**(5): 1035–1039.

Larkin, B.A.J., et al., Crime scene investigation III: Exploring the effects of drugs of abuse and neurotransmitters on bloodstain pattern analysis. *Anal Methods*, 2012, **4**(3): 721–729.

Lee, H.C. and C. Ladd, Preservation and collection of biological evidence. *Croat Med J*, 2001, **42**(3): 225–228.

Lee, H.C. and E.M. Pagliaro, Forensic evidence and crime scene investigation. *J Forensic Invest*, 2013, **1**(1): 1–5.

Lee, H.C., T. Palmbach, and M.T. Miller, *Henry Lee's Crime Scene Handbook*, 2001. San Diego: Academic Press.

Lee, H.C., et al., Guidelines for the collection and preservation of DNA evidence. *J Forensic Ident*, 1991, **41**(5): 13.

Ma, M., H. Zheng, and H. Lallie, Virtual reality and 3D animation in forensic visualization. *J Forensic Sci*, 2010, **55**(5): 1227–1231.

Marchant, B. and C. Tague, Developing fingerprints in blood: A comparison of several chemical techniques. *J Forensic Ident*, 2007, **57**(1): 76–93.

Matte, M., et al., Prevalence and persistence of foreign DNA beneath fingernails. *Forensic Sci Int Genet*, 2012, **6**(2): 236–243.

Mennell, J., The future of forensic and crime scene science. Part II. A UK perspective on forensic science education. *Forensic Sci Int*, 2006, **157**(Suppl 1): S13–S20.

Morgan, J.A., Comparison of cervical os versus vaginal evidentiary findings during sexual assault exam. *J Emerg Nurs*, 2008, **34**(2): 102–105.

Mulligan, C.M., S.R. Kaufman, and L. Quarino, The utility of polyester and cotton as swabbing substrates for the removal of cellular material from surfaces. *J Forensic Sci*, 2011, **56**(2): 485–490.

National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002. Washington, DC: US Department of Justice, Office of Justice Programs, Washington, DC, 2002.

Newton, M., The forensic aspects of sexual violence. *Best Pract Res Clin Obstet Gynaecol*, 2013, **27**(1): 77–90.

Nikolic, S. and V. Zivkovic, A healed bony puzzle: An old gunshot wound to the head. *Forensic Sci Med Pathol*, 2013, **9**(1): 112–116.

Nunn, S., Touch DNA collection versus firearm fingerprinting: Comparing evidence production and identification outcomes. *J Forensic Sci*, 2013, **58**(3): 601–608.

Oliva, A., et al., State of the art in forensic investigation of sudden cardiac death. *Am J Forensic Med Pathol*, 2011, **32**(1): 1–16.

Oliver, W.R. and L. Leone, Digital UV/IR photography for tattoo evaluation in mummified remains. *J Forensic Sci*, 2012, **57**(4): 1134–1136.

Osterkamp, T., K9 water searches: Scent and scent transport considerations. *J Forensic Sci*, 2011, **56**(4): 907–912.

Porta, D., et al., The importance of an anthropological scene of crime investigation in the case of burnt remains in vehicles: 3 case studies. *Am J Forensic Med Pathol*, 2013, **34**(3): 195–200.

Pringle, J.K., J.P. Cassella, and J.R. Jervis, Preliminary soilwater conductivity analysis to date clandestine burials of homicide victims. *Forensic Sci Int*, 2010, **198**(1–3): 126–133.

Ribaux, O., et al., Intelligence-led crime scene processing. Part I: Forensic intelligence. *Forensic Sci Int*, 2010, **195**(1–3): 10–16.

Ribaux, O., et al., Intelligence-led crime scene processing. Part II: Intelligence and crime scene examination. *Forensic Sci Int*, 2010, **199**(1–3): 63–71.

Rios, L., J.I. Ovejero, and J.P. Prieto, Identification process in mass graves from the Spanish Civil War I. *Forensic Sci Int*, 2010, **199**(1–3): e27–e36.

Rutty, G.N., A. Hopwood, and V. Tucker, The effectiveness of protective clothing in the reduction of potential DNA contamination of the scene of crime. *Int J Legal Med*, 2003, **117**(3): 170–174.

Sakelliadis, E.I., C.A. Spiliopoulou, and S.A. Papadodima, Forensic investigation of child victim with sexual abuse. *Indian Pediatr*, 2009, **46**(2): 144–151.

Sansoni, G., et al., Scene-of-crime analysis by a 3-dimensional optical digitizer: A useful perspective for forensic science. *Am J Forensic Med Pathol*, 2011, **32**(3): 280–286.

Schmidt, A., Crime scene investigation approach to sudden cardiac death. *J Am Coll Cardiol*, 2013, **62**(7): 630–631.

Schyma, C., B. Madea, and C. Courts, Persistence of biological traces in gun barrels after fatal contact shots. *Forensic Sci Int Genet*, 2013, **7**(1): 22–27.

Sharma, L., V.P. Khanagwal, and P.K. Paliwal, Homicidal hanging. *Legal Med*, 2011, **13**(5): 259–261.

Shaw, J. and R. Campbell, Predicting sexual assault kit submission among adolescent rape cases treated in forensic nurse examiner programs. *J Interpers Violence*, 2013, **28**(18): 3400–3417.

Shuttlewood, A.C., J.W. Bond, and L.L. Smith, The relationship between deprivation and forensic material recovered from stolen vehicles: Is it affected by vehicle condition and tidiness? *J Forensic Sci*, 2011, **56**(2): 510–513.

Smith, P.A., et al., Measuring team skills in crime scene investigation: Exploring ad hoc teams. *Ergonomics*, 2008, **51**(10): 1463–1488.

Solla, M., et al., Experimental forensic scenes for the characterization of ground-penetrating radar wave response. *Forensic Sci Int*, 2012, **220**(1–3): 50–58.

Taroni, F., et al., Whose DNA is this? How relevant a question? (a note for forensic scientists). *Forensic Sci Int Genet*, 2013, **7**(4): 467–470.

Thackeray, J.D., et al., Forensic evidence collection and DNA identification in acute child sexual assault. *Pediatrics*, 2011, **128**(2): 227–232.

Trombka, J.I., et al., Crime scene investigations using portable, non-destructive space exploration technology. *Forensic Sci Int*, 2002, **129**(1): 1–9.

Uzun, I., et al., Identification procedures as a part of death investigation in Turkey. *Am J Forensic Med Pathol*, 2012, **33**(1): 1–3.

Verdon, T.J., R.J. Mitchell, and R.A. van Oorschot, Swabs as DNA collection devices for sampling different biological materials from different substrates. *J Forensic Sci*, 2014, **59**(4): 1080–1089.

Wahlsten, P., V. Koiranen, and P. Saukko, Survey of medico-legal investigation of homicides in the city of Turku, Finland. *J Forensic Leg Med*, 2007, **14**(5): 243–252.

Walter, B.S. and J.J. Schultz, Mapping simulated scenes with skeletal remains using differential GPS in open environments: An assessment of accuracy and practicality. *Forensic Sci Int*, 2013, **228**(1–3): e33–e46.

Westen, A.A., R.R. Gerretsen, and G.J. Maat, Femur, rib, and tooth sample collection for DNA analysis in disaster victim identification (DVI): A method to minimize contamination risk. *Forensic Sci Med Pathol*, 2008, **4**(1): 15–21.

Wiegand, P., et al., Transfer of biological stains from different surfaces. *Int J Legal Med*, 2011, **125**(5): 727–731.

Winskog, C., Underwater disaster victim identification: The process and the problems. *Forensic Sci Med Pathol*, 2012, **8**(2): 174–178.

Yoon, D.Y., Newsmaker interview: Yoon Duk Yong. Crime scene investigation: The sinking of the Cheonan. Interview by Dennis Normile. *Science*, 2010, **328**(5984): 1336–1337.

Zeren, C., et al., A case of a serial sexual offender: The first ever report from Turkey. *J Forensic Leg Med*, 2012, **19**(7): 428–430.

2

Crime Scene Bloodstain Pattern Analysis

Bloodstain pattern analysis is the application of scientific knowledge to the examination and the interpretation of the morphology, the sequence, and the distribution of bloodstains associated with a crime. These analyses may determine the sequence of events; the approximate blood source locations; the positioning of the victim; and the position, the intensity, and the number of impacts applied to the blood source. They also can assist in the determination of the manner of death and can distinguish between accidents, homicides, and suicides. Bloodstain pattern analysis provides critical information for crime scene reconstructions in violent crime investigations (Chapter 1).

2.1 Basic Biological Properties of Human Blood

Blood is a bodily fluid circulating within the body. An average adult has a blood volume of approximately 8% of his or her body weight. Blood consists of a cellular portion as well as a liquid portion known as plasma (Chapter 12). The cellular portion consists of blood cells and platelets. The plasma is mostly composed of water and other substances such as proteins, inorganic salts, and other organic substances. The mass density of blood is only slightly greater than that of water. Blood can form clots (or *thrombi*) that are the result of blood coagulation (Chapter 16). Coagulation begins after an injury occurs, stopping blood loss from a damaged vessel. The normal coagulation time for 1 mL of venous blood in a glass tube is 5–15 min. The coagulation time can be affected by many factors such as blood volume and mechanical disturbance.

2.2 Formation of Bloodstains

The formation of a blood droplet is a complex event that is influenced by viscosity, surface tension, cohesion force, and gravity. Blood is viscous, and blood viscosity is a measure of the blood's resistance to flow. The viscosity of blood is approximately five times greater than that of water. During the formation of a drop of blood, blood leaks out from a blood source. The surface tension of the blood causes it to hang from the opening of a blood source and to form a pendant drop of blood (Figure 2.1). The molecules of a blood drop are held together by the cohesion force to maintain the shape of a blood drop (Figure 2.2). Surface tension causes liquids to minimize their surface. As a result, the formed blood drop is spherical. As the volume of the drop gradually increases and exceeds a certain size, it detaches itself and falls. The falling drop is also held

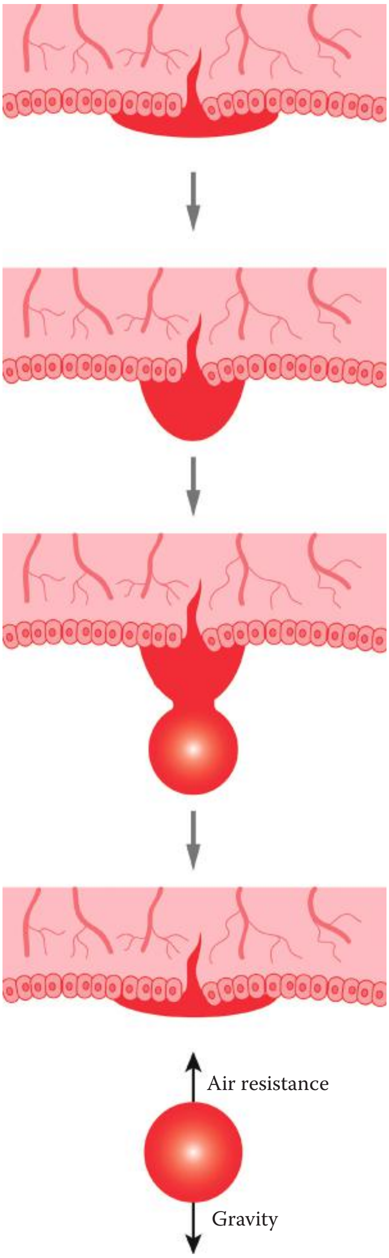

Figure 2.1 Forming a blood drop from a blood source. The blood that leaks out of the blood source forms a pendant drop of blood. As the volume of the pendant drop increases, the drop stretches in a downward direction. Eventually, the drop detaches and falls. The falling drop is largely influenced by the force of gravity and air resistance. (© Richard C. Li.)

together by surface tension. A falling blood drop is influenced by the downward force of gravity acting on the drop and the air resistance that acts in the opposite direction as the drop is in motion (Figure 2.1).

When a bloodstain lands on a surface, the shape and the size of the bloodstain is affected by the texture of the target surface. Bloodstains that land on porous or rough surfaces usually

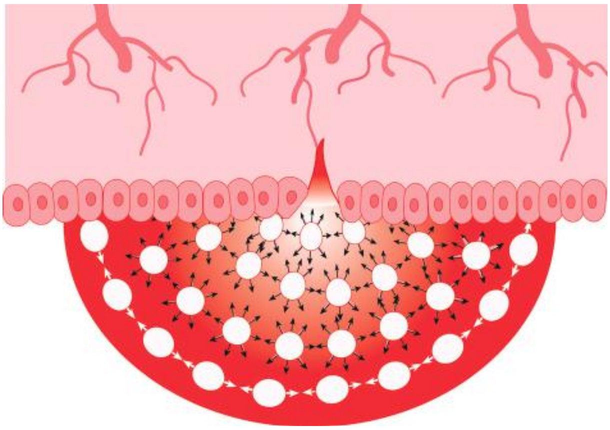

Figure 2.2 The particles of blood are attracted to each other by cohesive forces that are responsible for surface tension. As a result, a formed blood drop is spherical in order to minimize its surface area. Black arrow, cohesive force; white arrow, surface tension. (© Richard C. Li.)

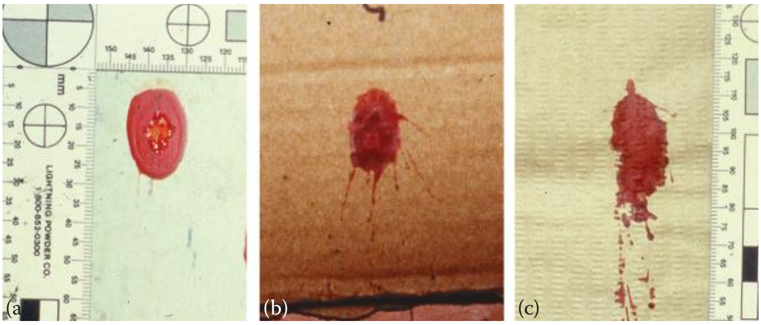

Figure 2.3 The respective morphologies of falling blood drops that land on surfaces with different textures at a 30° angle. (a) Tile, (b) cardboard, and (c) paper towel. (© Richard C. Li.)

have more distortion around the edges of the stains than those that land on smooth surfaces. A comparison of blood dropped onto different textures of target surfaces is shown in Figure 2.3.

2.3 Chemical Enhancement and Documentation of Bloodstain Evidence

Many chemical reagents react with blood to exhibit a color, a chemiluminescent light, or a fluorescent light (Chapter 12). These tests are extremely sensitive and thus are used as chemical enhancement reagents for detecting bloodstains. For bloodstain pattern analysis, the enhancement reagent is primarily used for detecting latent bloodstains such as diluted bloodstains that are visible on enhancement. A commonly used chemical reagent is luminol, which can be used for locating bloodstains at the scene. Other reagents such as phenolphthalein, leucomalachite green, and tetramethylbenzidine are not often used as enhancement reagents but rather as presumptive tests for blood. The positive reactions of all these reagents indicate the presence of blood.

Documenting bloodstain patterns at the scene is a major task of the investigation. Documenting bloodstain evidence can be done using a combination of photography, note-taking, and sketching. The general principle of crime scene documentation is described in Chapter 1. In bloodstain pattern analysis, special attention must obviously be given to bloodstains. The photographic documentation of bloodstains may be performed by multiple means, including film and digital photography, as well as videotaping. Photographs should be taken with an overall view followed by a medium-range and a close-up view of the bloodstain patterns. A scale of measurement must be included in the photograph, which is critical for bloodstain analysis. To avoid any distortion, the photographs should be taken with the camera lens parallel to the target surface where the bloodstains are located. An overall photograph provides an overall view of the scene including the bloodstain evidence (Figure 2.4a). A midrange photograph provides more details of the bloodstain pattern compared with that of the overall photograph (Figure 2.4b). Single bloodstains should be visible in midrange images. A close-up photograph, usually taken with a macro lens, provides a detailed image of single bloodstains, which is useful for spatter pattern analysis (Figure 2.4c).

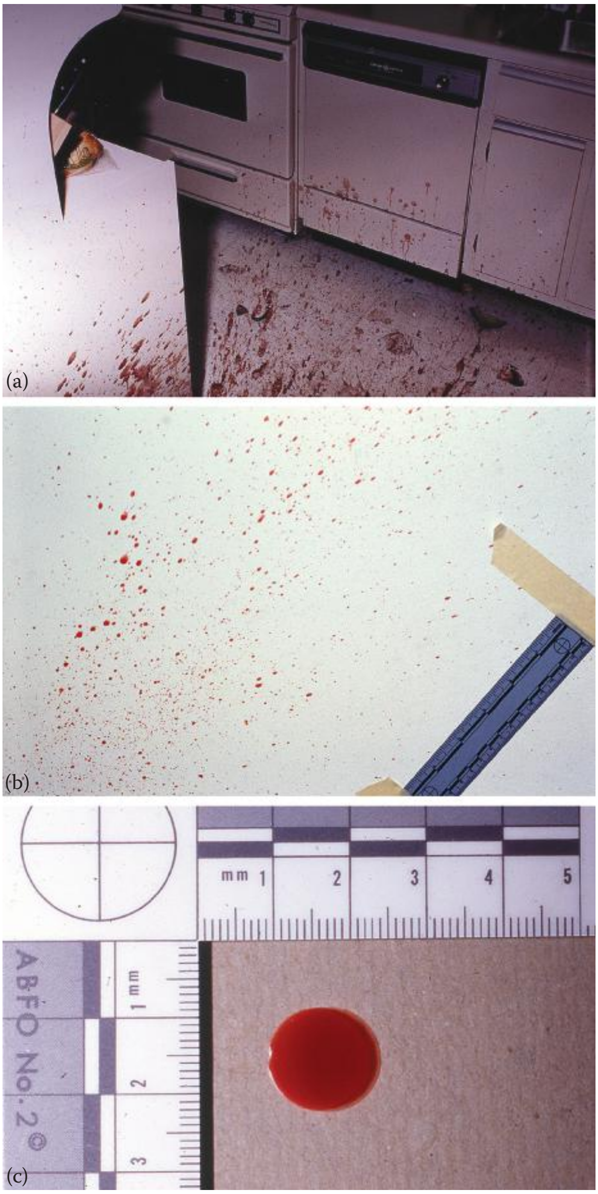

Figure 2.4 Crime scene photographic documentation. (a) Overall, (b) midrange, and (c) close-up photographs. (© Richard C. Li.)

2.4 Analyzing Spatter Stains

A *spatter stain*, based on the recommended terminology of the Scientific Working Group on Bloodstain Pattern Analysis (SWGSTAIN), is "a bloodstain resulting from a blood drop dispersed through the air due to an external force applied to a source of liquid blood." The patterns of spatter stains, including the shape and the size of the stains, are affected by the direction and the angle of impact (discussed in detail in Section 2.5.3) of the spatter stains that are projected. This information can be obtained from an analysis of the patterns of the spatter stains. Thus, it is possible to determine the area of origin (discussed in detail in Section 2.4.4) where an external force was directly applied to the blood source.

2.4.1 Velocity of Blood Droplets

The sizes of bloodstains are affected by the external force that is directly applied on a blood source. Increasing the energy of the external force will reduce the surface tension, thus decreasing the size of the droplets. Since these travelling blood droplets are driven by the energy derived from the external force, the higher the energy, the higher the velocities of the droplets. Bloodstains can be divided into three categories based on different travelling speeds. Low-velocity impact spatter is formed when a blood droplet is travelling at <1.5 m/s. The resulting stains are usually >4 mm in diameter (Figure 2.5a). As the travelling speed of blood droplets increases, the size of the spatter stain decreases. Medium-velocity impact spatter is formed when a blood source is subjected to a force associated with beatings or stabbings. The resulting stains range from 1 to 4 mm in diameter (Figure 2.5b). High-velocity impact spatter is formed when a blood source is subjected to a force associated with shooting using firearms. The resulting stains are usually <1 mm in diameter.

2.4.2 Determining the Directionality of the Stains

In this analysis, the effects of the directionality of the spatter stains projected are examined. SWGSTAIN defines the directionality to be "the characteristic of a bloodstain that indicates the direction blood was moving at the time of deposition." This analysis is applicable when the blood source is projected onto a surface at an angle of between 0° and 90°. Under this condition, the resulting spatter stain is an elongated ellipse (Figure 2.6), which is known as the *parent stain*. Additionally, *satellite stains* in the vicinity of the parent stain can be observed. As defined by SWGSTAIN, a satellite stain is "a smaller bloodstain that originated during the formation of the parent stain as a result of blood impacting a surface." More importantly, a spine is observed, which is the pointed edge away from the parent stain. When such a pattern is observed, the pointed end of the spine always points toward the direction of travel of the bloodstains.

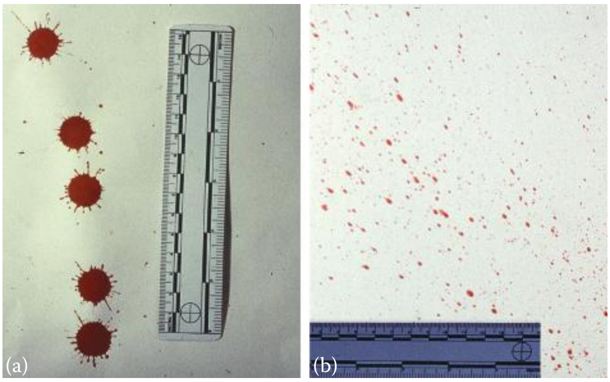

Figure 2.5 Bloodstains can be categorized based on their travelling velocities. (a) An example of a low-velocity impact spatter stain and (b) an example of a medium-velocity impact spatter stain. (© Richard C. Li.)

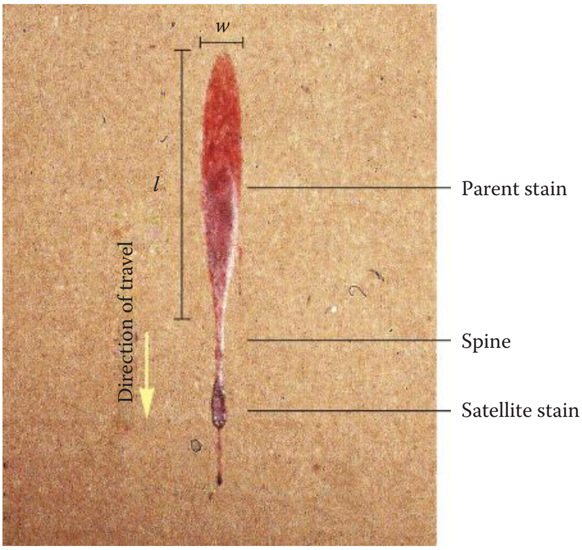

Figure 2.6 The morphology and directionality of a blood spatter stain. The arrow indicates the direction of travel. (© Richard C. Li.)

2.4.3 Determining Angles of Impact

SWGSTAIN defines the *angles of impact* to be "the acute angle (alpha), relative to the plane of a target at which a blood drop strikes the target." The shapes of the spatter stains are affected by the angle of impact. When a blood drop lands on a surface at a perpendicular angle (90°), a circular parent stain is formed (Figure 2.7), where the length and the width of the stain are equal. When a blood drop is projected onto a surface at an angle of between 0° and 90°, the stain is elongated. As the impact angle decreases, the shape of the spatter stain is more elongated (Figure 2.7), in which the length of the stain is greater than the width. It is observed that the ratio of the width and the length of the parent stain is proportional to the sine of the impact angle, which is summarized in the following trigonometric equation:

$$\sin \alpha = \frac{w}{l}$$

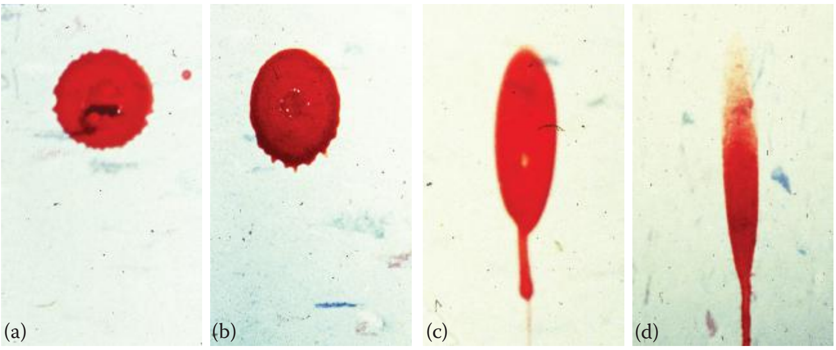

Figure 2.7 The effects of the impact angle on the shapes of blood spatter stains. Spatter stains are projected onto the surface of a ceramic tile at: (a) 90°, (b) 50°, (c) 20°, and (d) 10°. (© Richard C. Li.)

In this equation, α is the angle of impact, *l* is the length of the parent stain (major axis), and *w* is the width of the parent stain (minor axis). Thus, the angle of impact can be determined based on the relationship between the length and the width of the stain (Figure 2.8). Obviously, the measurement of the stain's axes is critical to the accuracy of the calculation of the angle of impact. To produce accurate and reproducible measurements, bloodstain pattern analysis software can be used, which superimposes a scaled close-up image of an individual bloodstain and calculates the angle of impact.

2.4.4 Determining Area of Origin

SWGSTAIN defines the *area of origin* to be "the three-dimensional location from which spatter originated." Using simple trigonometry, the area of origin can be determined based on the measurements from multiple elongated spatter stains (Figure 2.9). This can be accomplished by using the string method or the tangent method.

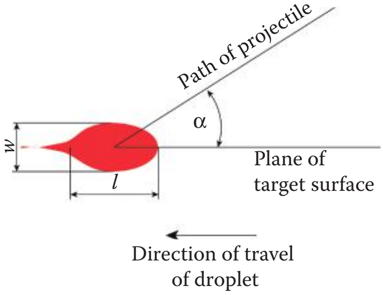

Figure 2.8 Impact angle. The angle between the path of a projectile and the plane of the target surface is shown. α, the impact angle; *l*, the length of the parent stain; and *w*, the width of the parent stain. (© Richard C. Li.)

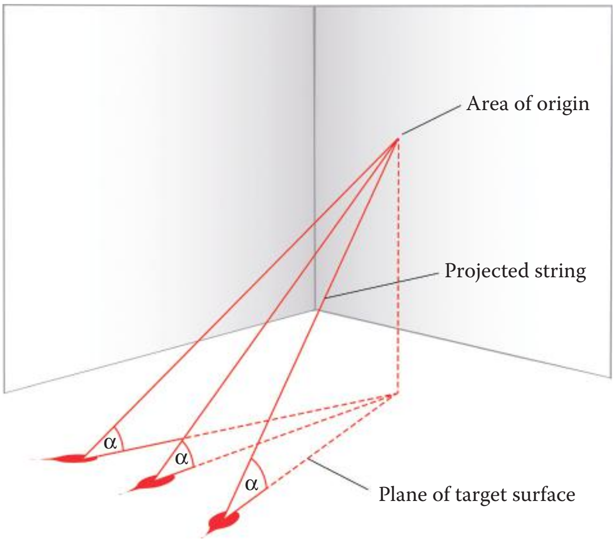

Figure 2.9 Area of origin. The area of origin is determined using the string method. Only three representative bloodstains are shown. α, impact angle. (© Richard C. Li.)

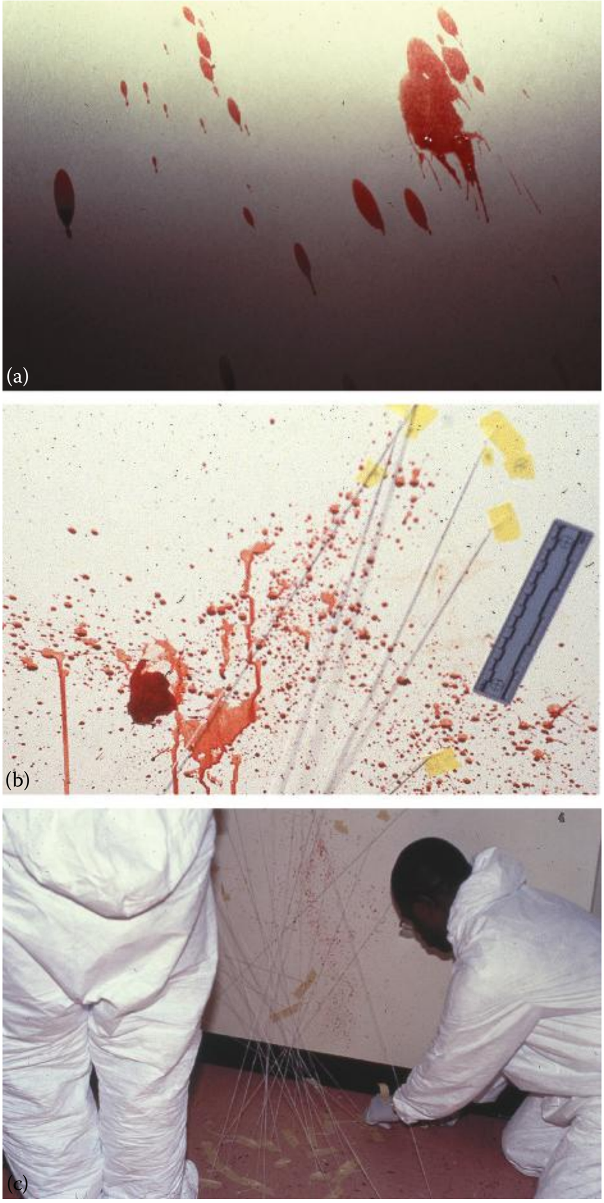

Figure 2.10 Determining the area of origin using the string method. (a) Selecting elongated spatter stains, (b) connecting strings, and (c) setting the path of the strings. (© Richard C. Li.)

In the string method, multiple (approximately two dozen) well-formed, elongated spatter stains are selected for analysis (Figure 2.10a). For each stain, the angle of impact is calculated. A piece of string is then connected between the stain and a surface with one end of the string precisely attached to the spatter stain (Figure 2.10b). The path of the string, indicating the trajectory of the stain, is set using a protractor based on the calculated angle of impact (Figures 2.10c and 2.11). This process is repeated until all the stains that have been selected are processed. For

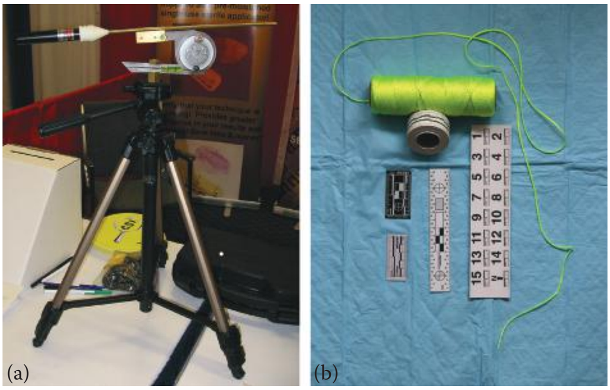

Figure 2.11 Tools for finding the area of origin. (a) Laser trajectory pointer and (b) string and scales. (© Richard C. Li.)

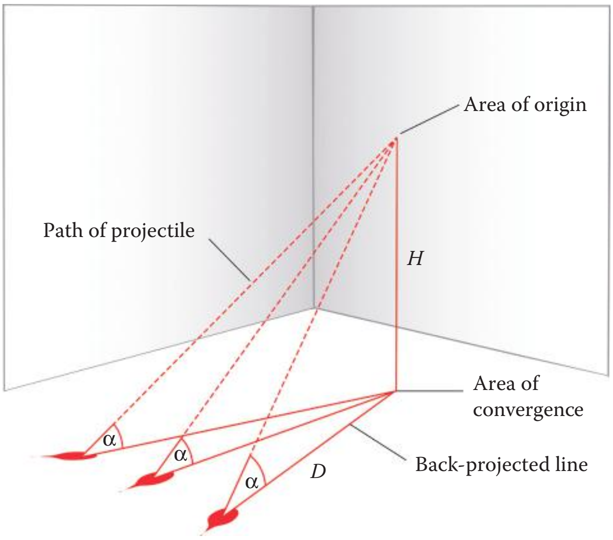

Figure 2.12 Determining the area of origin using the tangent method. Only three representative bloodstains are shown. α, the angle of impact; *H*, the height of the area of origin; and *D*, the distance from the spatter stain to the area of convergence. (© Richard C. Li.)

a spatter pattern generated from a single impact event, the strings converge. The area where the strings meet is the area of origin.

In the tangent method, the directionality of a single stain is determined first. A line is then back projected through the major axis of the bloodstain. For a single impact event, approximately two dozen stains are processed to determine the *area of convergence*. Based on SWGSTAIN's definition, the area of convergence is "the area containing the intersections generated by lines drawn through the long axes of individual stains that indicates in two dimensions the location of the blood source." Next, the angle of impact of each stain is calculated. The distance from the bloodstain to the area of convergence is measured (Figure 2.12). The height of the area of origin is calculated using the tangent function as shown:

$$H = D \cdot \tan \alpha$$

In the equation, α is the angle of impact, *H* is the height of the area of origin, and *D* is the distance from the spatter stain to the area of convergence.

2.5 Types of Bloodstain Patterns

Bloodstain patterns can be classified into three basic categories: passive, transfer, and projected bloodstains.

2.5.1 Passive Bloodstains

A *passive bloodstain* is formed due to bleeding from wounds, and the blood is deposited on a surface by the influence of the force of gravity alone. For example, a *drip stain* is formed when a falling drop of blood from an exposed wound or a blood-bearing object lands on a surface. If a blood source is moving, a *drip trail* is formed. A *drip pattern*, which is distinct from a drip stain, is formed when a liquid drips into another liquid, where one or both of the liquids are blood (Figure 2.13). As a result, secondary spatter stains are generated. As the dropping distance of the blood increases, the number of secondary spatter stains usually increases, and the size of these stains decreases. An approximate estimation of the dropping distance is possible. A *splash pattern* is formed when a volume of blood spills onto a surface (Figure 2.14). Splash patterns usually

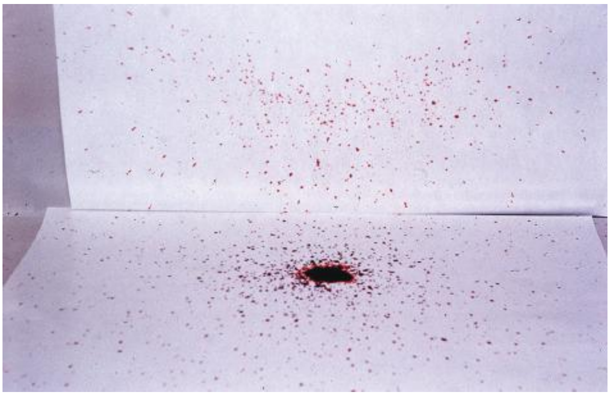

Figure 2.13 A drip pattern. The secondary spatter stains are shown. (© Richard C. Li.)

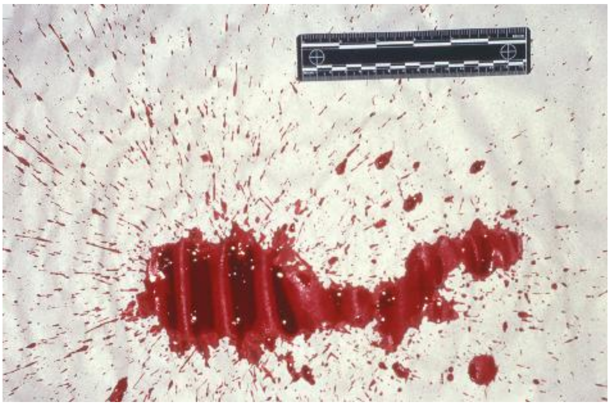

Figure 2.14 A splash pattern. Peripheral, elongated bloodstains are shown. (© Richard C. Li.)

have large stains surrounded by numerous, peripheral, elongated bloodstains. A *flow pattern* is caused by the movement of a large volume of blood on a surface either due to gravity or to the movement of the target such as a victim or postmortem disturbance. A *pool* is a bloodstain resulting from the accumulation of liquid blood on a surface (Figure 2.15). Sometimes, air bubbles in the blood may cause a *bubble ring* pattern (Figure 2.15). If blood is coagulated, gelatinous *blood clots* can be observed. Additionally, a *serum stain*, which consists of the liquid portion of the blood after a clot is formed, may also be present.

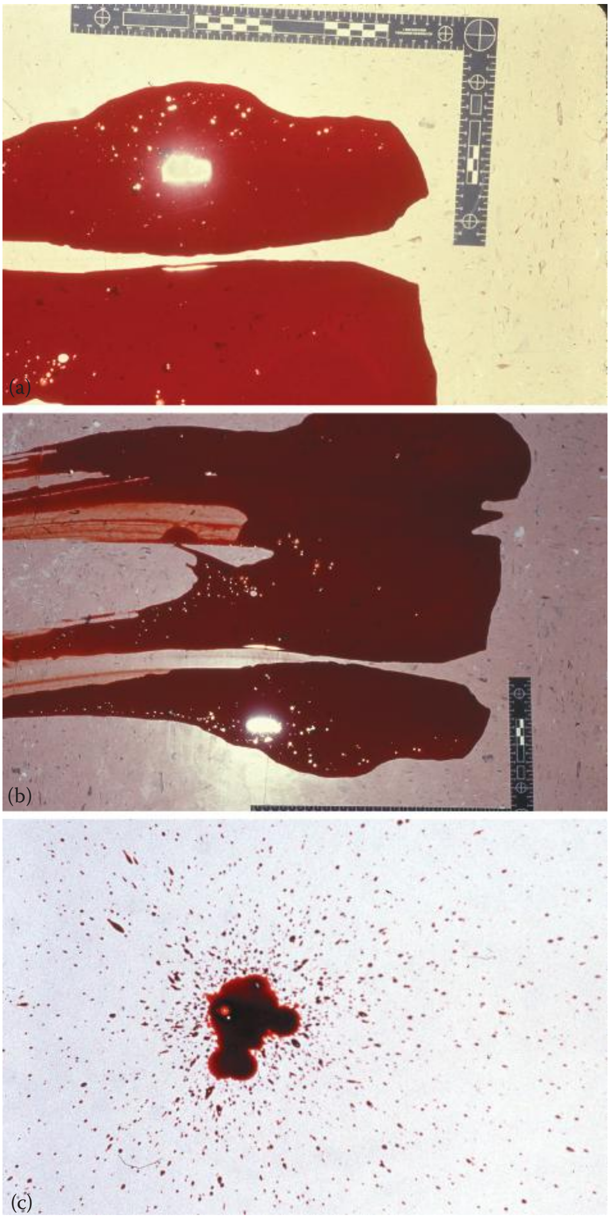

Figure 2.15 Pool and bubble ring patterns. (a) A pool pattern, (b) a disturbed pool pattern, and (c) a splash pattern with a bubble ring. Bubble rings are also present in (a) and (b). (© Richard C. Li.)

2.5.2 Transfer Bloodstains

A *transfer bloodstain*, based on SWGSTAIN, is "a bloodstain resulting from contact between a blood-bearing surface and another surface." For example, a *swipe pattern* is "a bloodstain pattern resulting from the transfer of blood from a blood-bearing surface onto another surface, with characteristics that indicate relative motion between the two surfaces." For example, bloody impressions can provide information about the shape, the size, and the pattern of the objects such as finger ridges, hands, and shoe soles. Examples of hand and shoe swipe patterns are shown in Figures 2.16 and 2.17. A *wipe pattern* is "an altered bloodstain pattern resulting from an object moving through a preexisting wet bloodstain." Examples of wipe patterns are shown in Figures 2.18 and 2.19. A *perimeter* stain, a type of wipe pattern, is a bloodstain that is disturbed before it is dried but it maintains the peripheral characteristics of the original stain (Figure 2.20). Perimeter stain patterns can be useful for the estimation of sequential events of acts. The pattern can also be used to estimate a time frame between the time of bleeding and the subsequent act. However, the drying time of a blood drop varies based on the surrounding conditions. Therefore, it is necessary to carry out a crime scene reconstruction under similar conditions to those of the scene to make such estimations.

2.5.3 Projected Bloodstains

A *projected bloodstain* is formed when a volume of blood is deposited on a surface under a pressure or a force that is greater than the force of gravity. For example, an *impact pattern* is

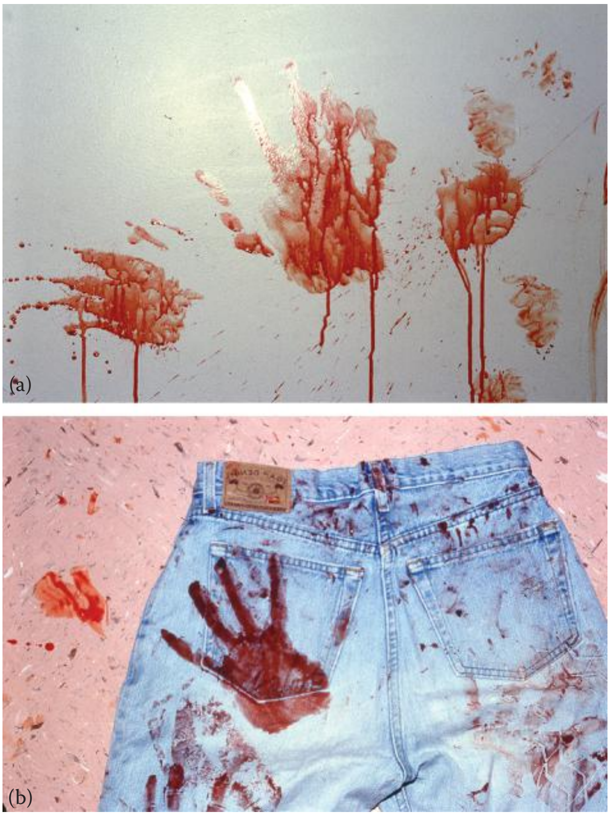

Figure 2.16 Bloody impressions. Bloody handprints are present on (a) a wall and (b) fabric. (© Richard C. Li.)

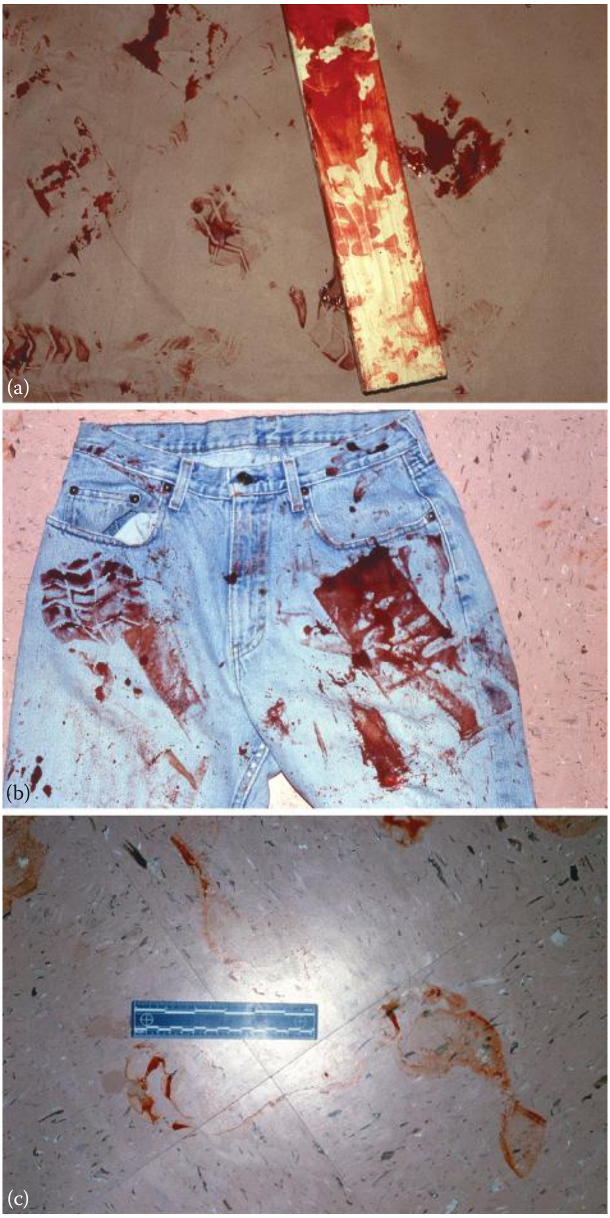

Figure 2.17 Bloody impressions. Bloody shoe prints on (a) paper and (b) fabric; and (c) bloody footprints on tile. (© Richard C. Li.)

formed when an object strikes liquid blood (Figure 2.21). A *cast-off pattern* is formed when blood drops are released from a moving blood-bearing object (Figure 2.22). Some spatter patterns are often associated with a wound penetrated by a projectile (Figure 2.23). A *forward spatter* is formed when blood drops travel from an exit wound in the same direction as a projectile, while a *back spatter* is formed when blood drops travel from an entry wound in the opposite direction of a projectile. Sometimes, internal bleeding caused by an injury may block the airway. An *expiration pattern* is formed when blood is forced by airflow through the

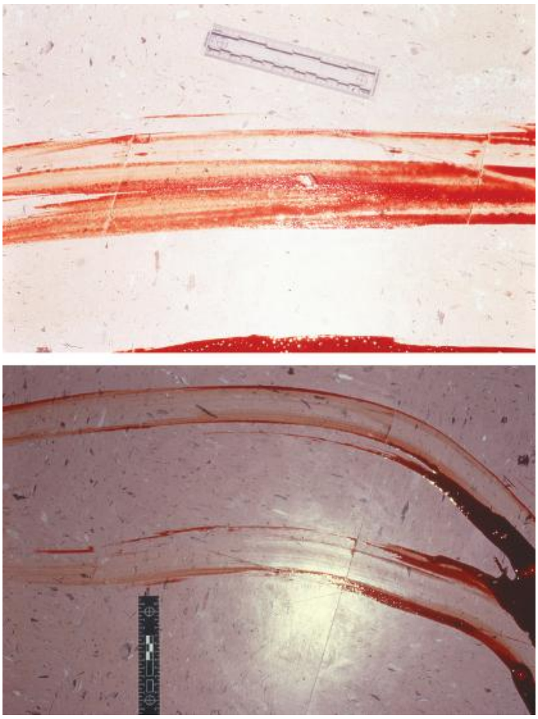

Figure 2.18 Wipe patterns. (© Richard C. Li.)

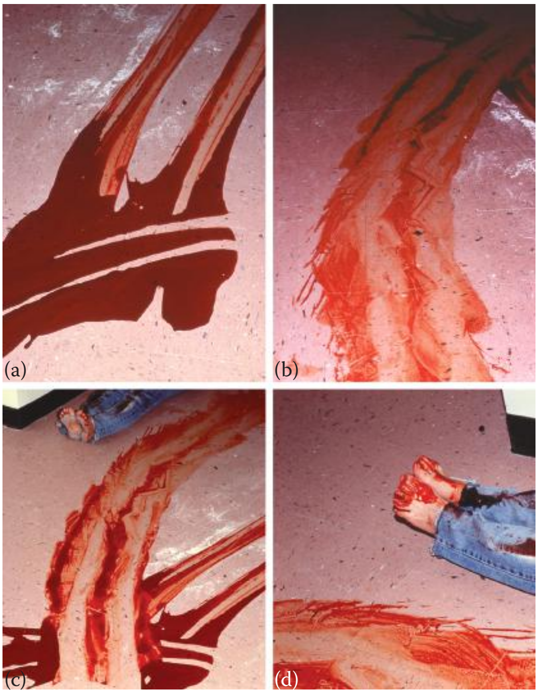

Figure 2.19 A wipe pattern caused by dragging a body through a pool of blood. (a) A pool of blood. Sections of the wipe pattern caused by dragging are shown in (b), (c), and (d). (© Richard C. Li.)

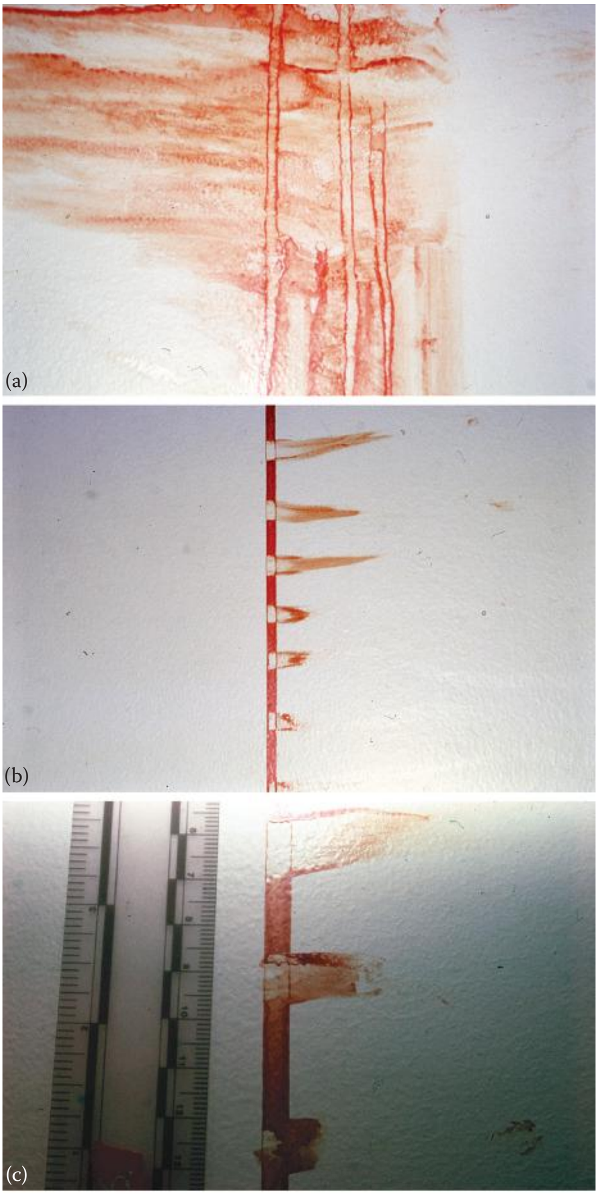

Figure 2.20 Perimeter stains. (a) Peripheral characteristics of the original stains are shown. Perimeter stains were created at different periods of time after the original stain was formed: (b) midrange view and (c) close-up view. (© Richard C. Li.)

trachea and out of the nose or mouth (Figure 2.24). An *arterial spurt pattern* is associated with wounds damaging arterial blood vessels where bloodstains are driven by arterial pressure. Although the shape of arterial patterns varies, these patterns usually have a series of large spurts with fluctuations corresponding to the systolic and the diastolic blood pressures. At a crime scene, if the projectile of bloodstains is blocked by an object, a *void pattern* is formed, which exhibits an area where there is an absence of blood surrounded by continuously distributed bloodstains.

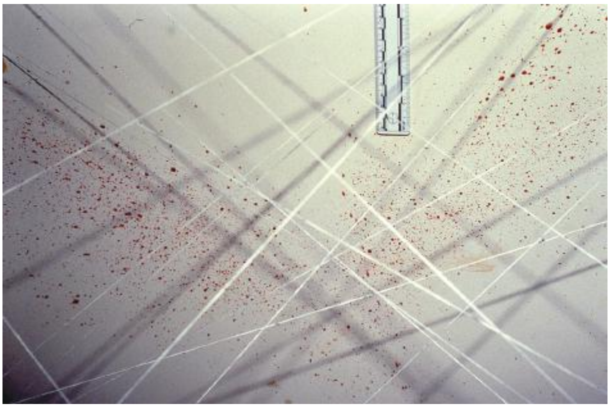

Figure 2.21 An impact bloodstain pattern as a result of using blunt force. (© Richard C. Li.)

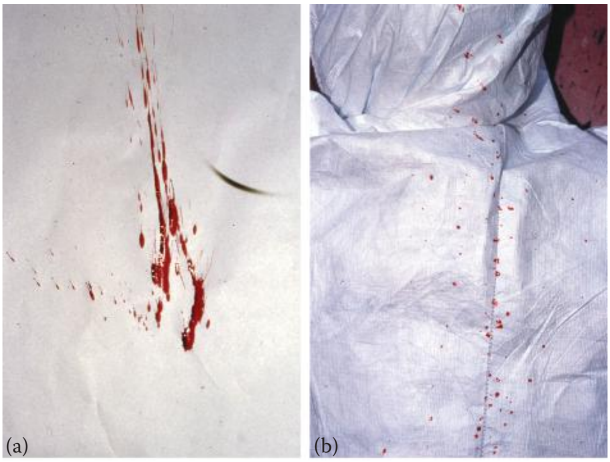

Figure 2.22 Cast-off patterns. Spatter stains are projected onto (a) a covered wall and (b) a lab coat. (© Richard C. Li.)

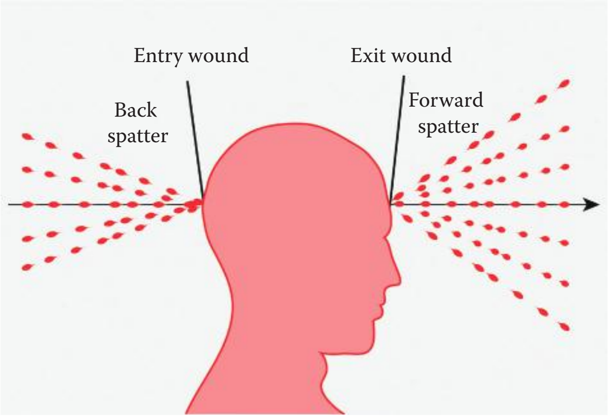

Entry wound
Exit wound
Back spatter
Forward spatter

Figure 2.23 Forward and back spatter patterns. Arrow, the direction of a projectile. (© Richard C. Li.)

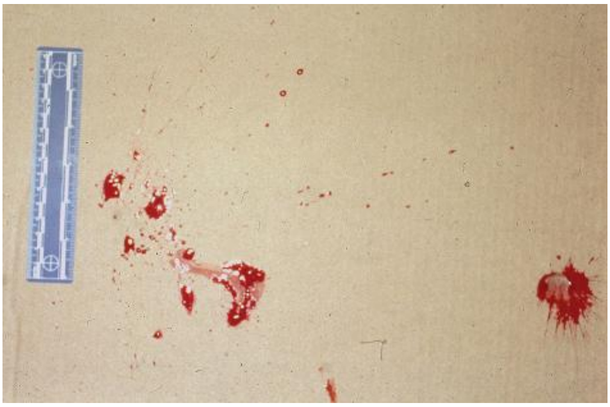

Figure 2.24 An expiration pattern. (© Richard C. Li.)

Bibliography

Adam, C.D., Fundamental studies of bloodstain formation and characteristics. *Forensic Sci Int*, 2012, **219**(1–3): 76–87.

Adelson, L., The coroner of Elsinore. *New Engl J Med*, 1960, **262**(5): 229–234.

Attinger, D., et al., Fluid dynamics topics in bloodstain pattern analysis: Comparative review and research opportunities. *Forensic Sci Int*, 2013, **231**(1–3): 375–396.

Behrooz, N., L. Hulse-Smith, and S. Chandra, An evaluation of the underlying mechanisms of bloodstain pattern analysis error. *J Forensic Sci*, 2011, **56**(5): 1136–1142.

Benecke, M. and L. Barksdale, Distinction of bloodstain patterns from fly artifacts. *Forensic Sci Int*, 2003, **137**(2–3): 152–159.

Buck, U., et al., 3D bloodstain pattern analysis: Ballistic reconstruction of the trajectories of blood drops and determination of the centres of origin of the bloodstains. *Forensic Sci Int*, 2011, **206**(1–3): 22–28.

Byard, R.W., et al., Blood stain pattern interpretation in cases of fatal haemorrhage from ruptured varicose veins. *J Forensic Leg Med*, 2007, **14**(3): 155–158.

Camana, F., Determining the area of convergence in bloodstain pattern analysis: A probabilistic approach. *Forensic Sci Int*, 2013, **231**(1–3): 131–136.

Connolly, C., M. Illes, and J. Fraser, Affect of impact angle variations on area of origin determination in bloodstain pattern analysis. *Forensic Sci Int*, 2012, **223**(1–3): 233–240.

Davidson, P.L., et al., Physical components of soft-tissue ballistic wounding and their involvement in the generation of blood backspatter. *J Forensic Sci*, 2012, **57**(5): 1339–1342.

de Bruin, K.G., R.D. Stoel, and J.C. Limborgh, Improving the point of origin determination in bloodstain pattern analysis. *J Forensic Sci*, 2011, **56**(6): 1476–1482.

de Castro, T., et al., Interpreting the formation of bloodstains on selected apparel fabrics. *Int J Legal Med*, 2013, **127**(1): 251–258.

Denison, D., et al., Forensic implications of respiratory derived blood spatter distributions. *Forensic Sci Int*, 2011, **204**(1–3): 144–155.

Donaldson, A.E., et al., Using oral microbial DNA analysis to identify expirated bloodspatter. *Int J Legal Med*, 2010, **124**(6): 569–576.

Donaldson, A.E., et al., Characterising the dynamics of expirated bloodstain pattern formation using high-speed digital video imaging. *Int J Legal Med*, 2011, **125**(6): 757–762.

Emes, A., Expirated blood—A review. *J Can Soc Forensic Sci*, 2001, **34**(4): 197–203.

Finnis, J., J. Lewis, and A. Davidson, Comparison of methods for visualizing blood on dark surfaces. *Sci Justice*, 2013, **53**(2): 178–186.

Fujikawa, A., et al., Changes in the morphology and presumptive chemistry of impact and pooled bloodstain patterns by *Lucilia sericata* (Meigen) (Diptera: Calliphoridae). *J Forensic Sci*, 2011, **56**(5): 1315–1318.

Illes, M. and M. Boue, Robust estimation for area of origin in bloodstain pattern analysis via directional analysis. *Forensic Sci Int*, 2013, **226**(1–3): 223–229.

Karger, B., et al., Bloodstain pattern analysis—Casework experience. *Forensic Sci Int*, 2008, **181**(1–3): 15–20.

Kettner, M., F. Ramsthaler, and A. Schnabel, "Bubbles"—A spot diagnosis. *J Forensic Sci*, 2010, **55**(3): 842–844.

Kunz, S.N., H. Brandtner, and H. Meyer, Unusual blood spatter patterns on the firearm and hand: A backspatter analysis to reconstruct the position and orientation of a firearm. *Forensic Sci Int*, 2013, **228**(1–3): e54–e57.

Larkin, B.A. and C.E. Banks, Preliminary study on the effect of heated surfaces upon bloodstain pattern analysis. *J Forensic Sci*, 2013, **58**(5): 1289–1296.

Larkin, B.A.J., et al., Crime scene investigation III: Exploring the effects of drugs of abuse and neurotransmitters on bloodstain pattern analysis. *Anal Methods*, 2012, **4**(3): 721–729.

Lee, W.C., et al., Statistical evaluation of alternative light sources for bloodstain photography. *J Forensic Sci*, 2013, **58**(3): 658–663.

Liesegang, J., Bloodstain pattern analysis—Blood source location. *J Can Soc Forensic Sci*, 2004, **37**(4): 215–222.

Maclean, B., K. Powley, and D. Dahlstrom, A case study illustrating another logical explanation for high velocity impact spatter. *J Can Soc Forensic Sci*, 2001, **34**(4): 191–195.

Makovicky, P., et al., The use of trigonometry in bloodstain analysis. *Soud Lek*, 2013, **58**(2): 20–25.

Maloney, A., et al., One-sided impact spatter and area-of-origin calculations. *J Forensic Ident*, 2011, **61**(2): 123–135.

Nikolic, S. and V. Zivkovic, Bloodstain pattern in the form of gushing in a case of fatal exsanguination due to ruptured varicose vein. *Med Sci Law*, 2011, **51**(1): 61–62.

Peschel, O., et al., Blood stain pattern analysis. *Forensic Sci Med Pathol*, 2011, **7**(3): 257–270.

Pizzola, P.A., et al., Commentary on "3D bloodstain pattern analysis: Ballistic reconstruction of the trajectories of blood drops and determination of the centres of origin of the bloodstains" by Buck et al. [Forensic Sci. Int. 206 (2011) 22–28]. *Forensic Sci Int*, 2012, **220**(1–3): e39–e40; author reply e41.

Power, D.A., et al., PCR-based detection of salivary bacteria as a marker of expirated blood. *Sci Justice*, 2010, **50**(2): 59–63.

Ramsthaler, F., et al., Drying properties of bloodstains on common indoor surfaces. *Int J Legal Med*, 2012, **126**(5): 739–746.

Randall, B., Blood and tissue spatter associated with chainsaw dismemberment. *J Forensic Sci*, 2009, **54**(6): 1310–1314.

Raymond, M.A., E.R. Smith, and J. Liesegang, Oscillating blood droplets—Implications for crime scene reconstruction. *Sci Justice*, 1996, **36**(3): 161–171.

Raymond, M.A., E.R. Smith, and J. Liesegang, The physical properties of blood—Forensic considerations. *Sci Justice*, 1996, **36**(3): 153–160.

Sauvageau, A., et al., Bloodstain pattern analysis in a case of fatal varicose vein rupture. *Am J Forensic Med Pathol*, 2007, **28**(1): 35–37.

Stephens, B.G. and T.B. Allen, Back spatter of blood from gunshot wounds. Observations and experimental simulation. *J Forensic Sci*, 1983, **28**(2): 437–439.

Striman, B., et al., Alteration of expirated bloodstain patterns by *Calliphora vicina* and *Lucilia sericata* (Diptera: Calliphoridae) through ingestion and deposition of artifacts. *J Forensic Sci*, 2011, **56**(Suppl 1): S123–S127.

Taylor, M.C., et al., The effect of firearm muzzle gases on the backspatter of blood. *Int J Legal Med*, 2011, **125**(5): 617–628.

Thanakiatkrai, P., A. Yaodam, and T. Kitpipit, Age estimation of bloodstains using smartphones and digital image analysis. *Forensic Sci Int*, 2013, **233**(1–3): 288–297.

Trombka, J.I., et al., Crime scene investigations using portable, non-destructive space exploration technology. *Forensic Sci Int*, 2002, **129**(1): 1–9.

White, B., Bloodstain patterns on fabrics: The effect of drop volume, dropping height and impact angle. *J Can Soc Forensic Sci*, 1986, **19**(1): 3–36.

Yen, K., et al., Blood-spatter patterns: Hands hold clues for the forensic reconstruction of the sequence of events. *Am J Forensic Med Pathol*, 2003, **24**(2): 132–140.

Yi, R. and Y. Wang, The experimental observation of secondary spatter bloodstain in morphology. *Chin J Forensic Med*, 2011, **26**(1): 26–29.

3

Forensic Biology

A Subdiscipline of Forensic Science

Forensic laboratories provide scientific analysis, evidence evaluation, and consultations to various criminal justice agencies for the investigation of criminal cases. Additionally, forensic laboratories provide expert testimony related to the resolution of criminal cases to the courts.

3.1 Common Disciplines of Forensic Laboratory Services

Many of the disciplines of the forensic laboratory services are commonly practiced in various municipal, county, state, and federal forensic laboratories in the United States. Forensic biology is a subdiscipline of forensic science. A full range of forensic laboratory services, known as "full service," usually includes: crime scene investigation (Figure 3.1); latent print examination (Figure 3.2); forensic biology (Figure 3.3); controlled substance analysis (Figure 3.4); postmortem toxicology (Figure 3.5); questioned document examination (Figure 3.6); firearm, toolmark, and other impression evidence examination (Figures 3.7 through 3.10); explosive and fire debris examination (Figure 3.11); and transfer (trace) evidence examination (Figures 3.12 and 3.13). Table 3.1 describes the services that are normally provided by a forensic laboratory with their respective analyses.

3.2 Laboratory Analysis of Biological Evidence

Forensic biology uses scientifically accepted protocols to analyze biological evidence. Laboratory analysis (Figures 3.14 and 3.15) utilizes scientific techniques for the examination of evidence, the reconstruction of a crime scene, the identification of biological fluids, and the comparison of individual characteristics of biological evidence.

3.2.1 Identification of Biological Evidence

The identification of biological evidence is the first step that is performed before further analyses are carried out. This includes the identification of biological fluids such as blood, saliva, and semen; this process is discussed in more detail in subsequent chapters (Chapters 12 through 14). The *identification* is based on a comparison of *class characteristics*—a set of characteristics that allows a sample to be placed in a category with similar materials. By comparing the

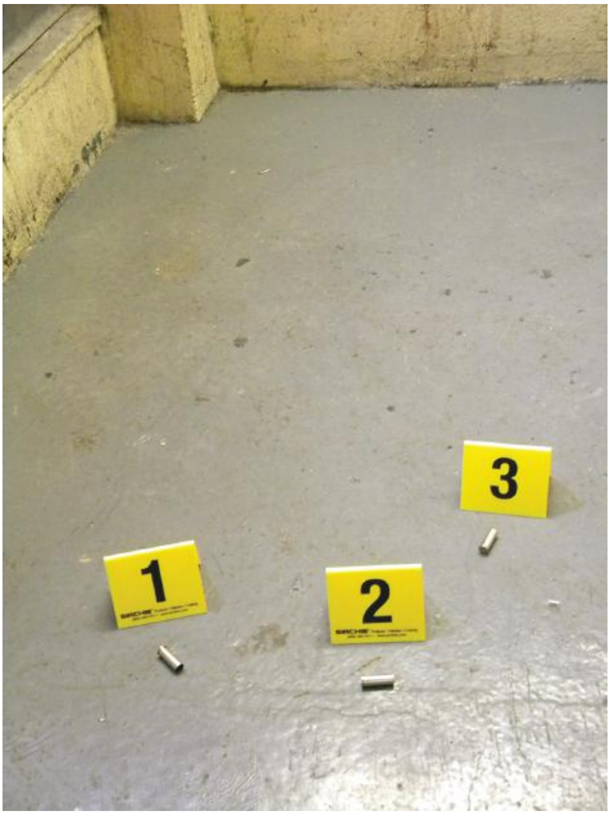

Figure 3.1 Crime scene investigation. Recovery of fired casings at the scene can aid in determining the position of the shooter. (© Richard C. Li.)

class characteristics of a sample with known standards of its class, biological samples can be identified.

3.2.2 Comparison of Individual Characteristics of Biological Evidence

Individual characteristics refer to the unique characteristics of both the evidence and a reference sample such as fingerprints, which share a common origin to a high degree of certainty. An example of biological evidence possessing individual characteristics is DNA polymorphisms. In the case of biological evidence, current forensic DNA profiling can compare individual characteristics of DNA evidence with a known reference sample. It is possible to determine that a biological stain originated from a particular individual, which is useful for human identification. The examination of individual characteristics of evidence can also exclude the possibility of a common origin. The specific methods utilized for the individualization of evidence are also discussed in subsequent chapters (Chapters 19 through 23).

3.2.3 Reporting Results and Expert Testimony

After the analysis of evidence is completed, a report is prepared based on the results of the analysis, which may include sections discussing the specific evidence analyzed, the method of analysis used, the results obtained, and the conclusions drawn. In the case of DNA evidence, the strength of the conclusion is usually evaluated via statistical computations (Chapter 25). A forensic scientist often serves as an expert witness whose testimony

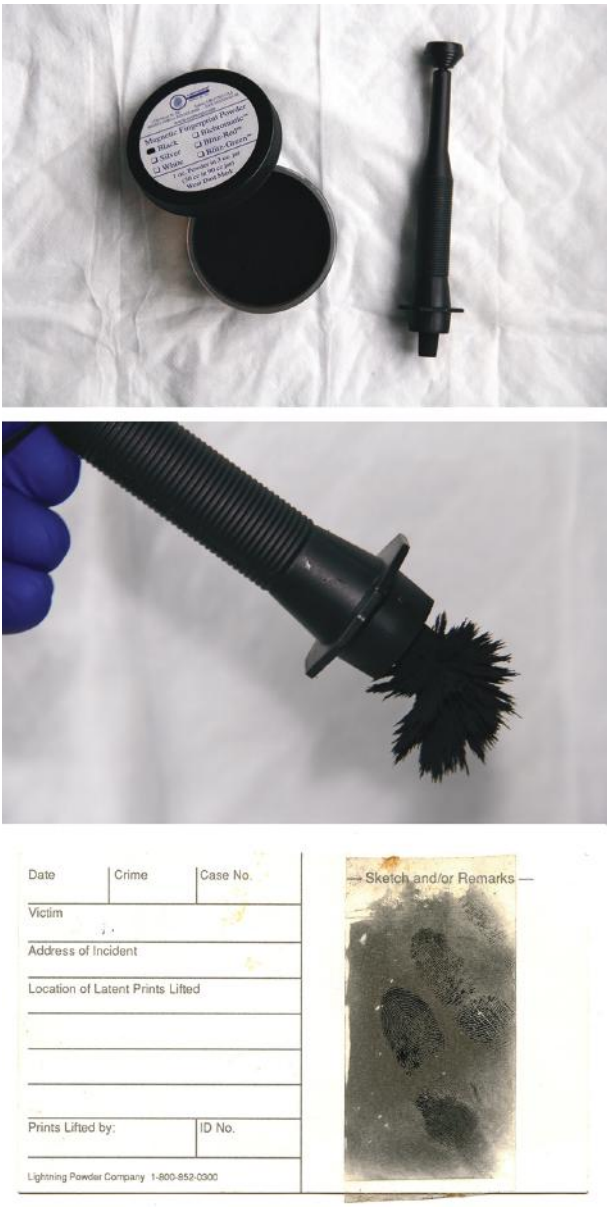
Magnetic Fingerprint Powder
☑ Black
☐ Silver
☐ White
☐ Bichromatic
☐ Blitz-Red
☐ Blitz-Green
Date
Crime
Case No.
Victim
Address of Incident
Location of Latent Prints Lifted
Prints Lifted by:
ID No.
Lightning Powder Company 1-800-852-0300
Sketch and/or Remarks

Figure 3.2 Developing latent fingerprints by dusting methods. Using magnetic fingerprint powder (top) that is held by a magnetic applicator (middle). Fingerprints (bottom) dusted by magnetic fingerprint powder, tape lifted and preserved on a fingerprint card. (© Richard C. Li.)

provides professional opinions about the evidence analyzed. Based on the federal rules of evidence, an expert witness is qualified based on his or her knowledge, skill, experience, training, or education, and may give an opinion to the court that is relevant to the analyses conducted. However, forensic scientists must also communicate their findings to attorneys, judges, and members of a jury. This requires the translation of technical information into layman's terms.

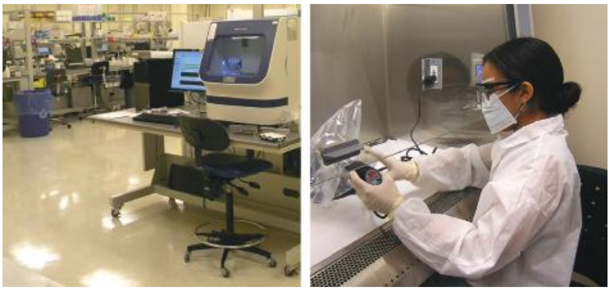

Figure 3.3 A section of a forensic biology laboratory showing an automated electrophoresis instrument used for forensic DNA profiling (left). Processing biological evidence in a biosafety cabinet (right). (© Richard C. Li.)

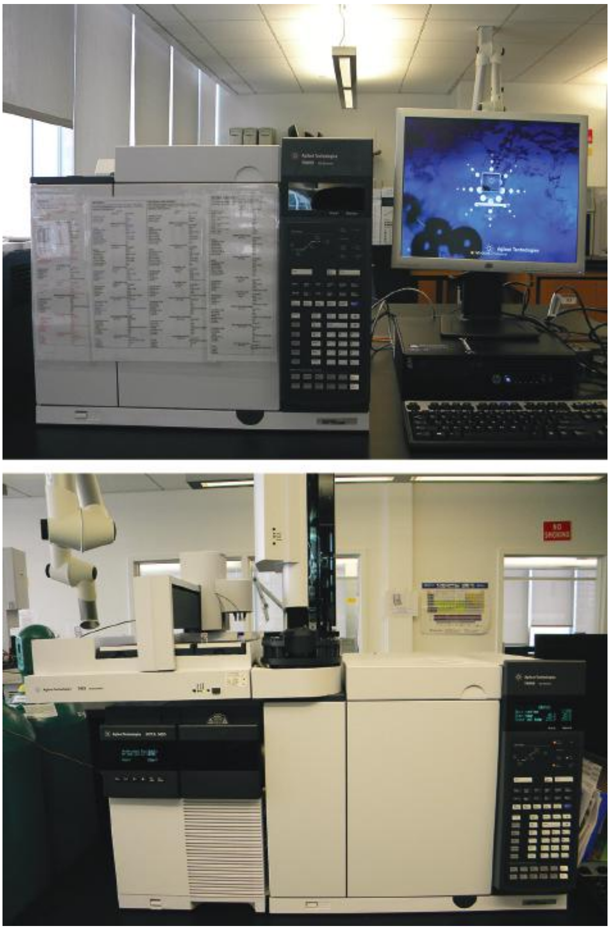

Figure 3.4 Gas chromatograph (top) and gas chromatograph-mass spectrometer (bottom) used for controlled substance analysis. (© Richard C. Li.)

Figure 3.5 Tissue samples (left) and gas chromatograph-mass spectrometer (right) for forensic postmortem toxicological analysis. (© Richard C. Li.)

Figure 3.6 A digital imaging system, using multiple illumination sources ranging from ultraviolet to infrared wavelengths, for examining altered and counterfeit documents (left). An electrostatic imaging system for detecting indented writing on questioned documents (right). This device generates an electrostatic image of indented writing, which is then visualized using charge-sensitive toners. (© Richard C. Li.)

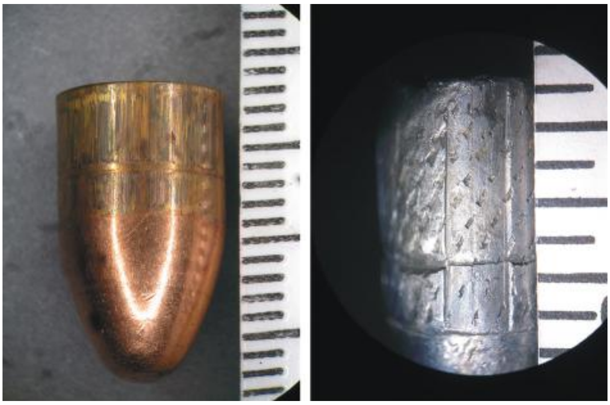

Figure 3.7 Striation marks on fired bullets can be analyzed to match a bullet to a gun. (Courtesy of P. Diaczuk.)

3.3 Forensic Science Services Related to Forensic Biology

A number of specialized forensic science services beyond those provided by forensic laboratories are routinely available to law enforcement agencies. For example, forensic services related to biological evidence and those involving more specialized analysis are available. These services are important aids to a criminal investigation and require the expertise of individuals who have highly specialized skills.

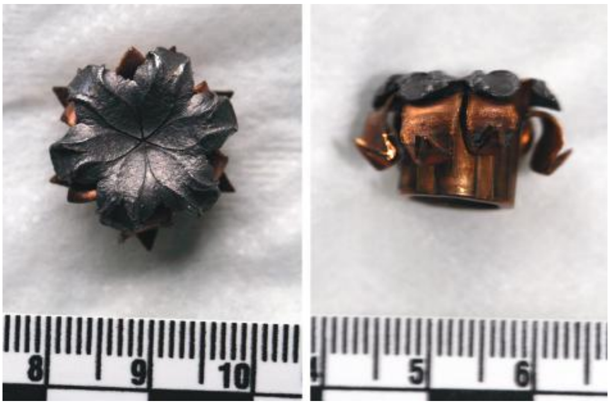

Figure 3.8 Fired hollowpoint bullets (.45 caliber ACP Winchester). Striation marks are visible on the side view of a bullet (right). (© Richard C. Li.)

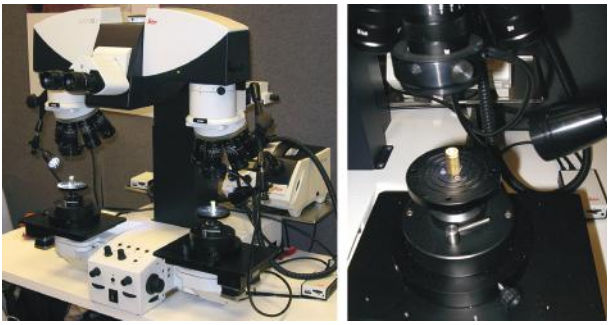

Figure 3.9 A comparison microscope is used for the simultaneous comparison of two items of firearm evidence side by side. (© Richard C. Li.)

3.3.1 Forensic Pathology

When a death is deemed suspicious or unexplained, medical examiners frequently perform autopsies to determine the exact cause (Figures 3.16 and 3.17). The manner of death is classified into one of five categories based on the circumstances: natural, homicide, suicide, accident, or undetermined. Additionally, a medical examiner participating in a criminal investigation is often responsible for estimating the time of death.

3.3.2 Forensic Anthropology

Forensic anthropology is the identification and the examination of human skeletal remains (Figures 3.18 and 3.19). Skeletal remains can reveal a number of individual characteristics that can be useful in attempting to identify an individual. An examination of bones may reveal an individual's origin, sex, approximate age, race, and the presence of a skeletal injury. A forensic anthropologist may also assist in creating facial reconstructions to aid in the identification of skeletal remains or may be called on to help collect and organize bone fragments in the course of identifying victims of mass disasters such as plane crashes as well as victims in mass graves discovered after wars or genocides.

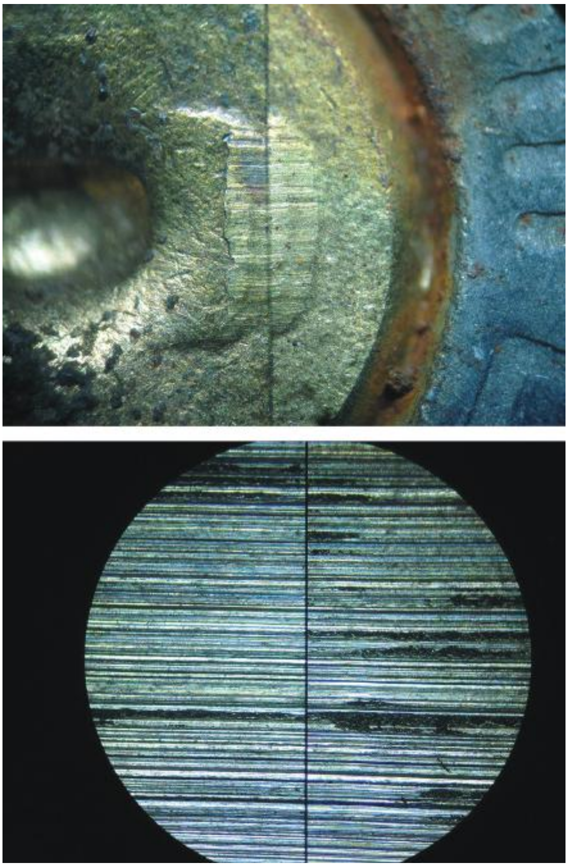

Figure 3.10 Comparing the striations between evidence and reference samples using a comparison microscope: casings (top) and bullets (bottom). (Courtesy of P. Diaczuk.)

Figure 3.11 Scanning electron microscope used in the analysis of gunshot residue and explosives. (© Richard C. Li.)

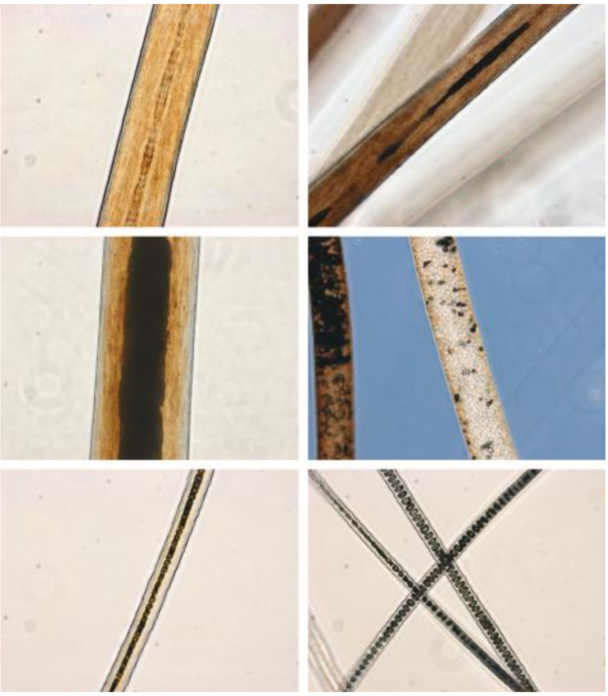

Figure 3.12 Trace evidence such as hairs can be transferred during the acts of a violent crime. The analysis and comparisons of these types of trace evidence can potentially establish a link between a suspect or a victim and a crime scene. Hairs from human (top), horse (middle left), deer (middle right), dog (bottom left), and cat (bottom right). (© Richard C. Li.)

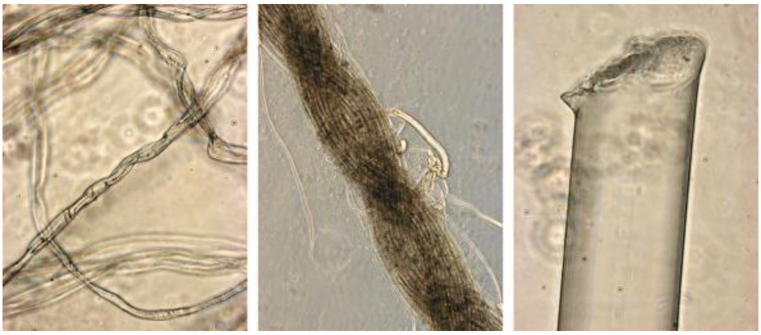

Figure 3.13 Trace evidence: fibers. Cotton (left), nylon (middle), and polyester (right). (© Richard C. Li.)

3.3.3 Forensic Entomology

The study of insects in relation to a criminal investigation is known as forensic entomology. This forensic discipline is valuable for estimating the time of death when the circumstances surrounding the crime are otherwise unknown. The stages of development of certain insect species present in or on a body can be identified and allow a forensic entomologist to approximate how long the body was left exposed (Figure 3.20).

Table 3.1 Common Services Provided by US Forensic Laboratories

Service	Function	Method
Crime scene investigation	Evidence recognition, documentation, collection, and preservation	Crime scene responses and related endeavors are diverse and vary with case and type of evidence
Latent print examination	Analysis of friction ridge detail in fingerprints Activities include visualization, recording, comparison, storage, and recovery of latent prints	Alternate light sources, physical (powder) and chemical enhancements Direct lifts, photography, and digital imaging Use of an Automated Fingerprint Identification System (AFIS) database
Forensic biology	Identification of biological fluids (blood, semen, and saliva) DNA profiling for individualization	Serological and biochemical methods Polymerase chain reaction (PCR)-based methods Automated electrophoresis platforms Use of Combined DNA Index System (CODIS)
Controlled substance analysis	Identification and quantification of drugs present in submitted evidence	Microscopic, chemical, chromatographic, and spectroscopic methodologies Gas chromatography–mass spectrometry or infrared spectrophotometry
Postmortem toxicology	Determination of concentrations of substances and their metabolites in biological fluids or tissues	Immunoassays and chemical methods Confirmatory techniques such as gas and liquid chromatography–mass spectrometry
Questioned document examination	Investigation of forgeries, tracings, disguised handwritings, computer manipulation of images, and recovery of altered documents Analysis of papers, inks, toners, word processors, typewriters, copiers, and printers	Macroscopic and microscopic comparisons Chromatographic and spectroscopic methods
Firearm and toolmark examination	Identification of firearms, tools, and other implements (expertise achieved predominantly through experience)	Microscopic comparisons of questioned and authenticated impressions Comparison of striae on recovered bullets Use of National Integrated Ballistics Information Network (NIBIN)
Explosive and fire debris examination	Identification, recovery, and detection of bulk explosives, residues, debris, and accelerants	Microscopic, spectroscopic, and chromatographic methods Gas chromatography–mass spectrometry may be needed to adequately characterize sample
Trace evidence examination	Analysis of transferred evidence such as hairs, fibers, soil, paints, and glass	Microscopic analysis of evidence with gas chromatograph-mass spectrometers, FTIR microscopes, scanning electron microscopes, basic and advanced microscopy, and capillary electrophoresis

Source: Adapted from US Department of Justice, Office of Justice Programs, National Institute of Justice (NIOJ), *Forensic Science: Review of Status and Needs*, 1999, US Department of Justice, Washington, DC.

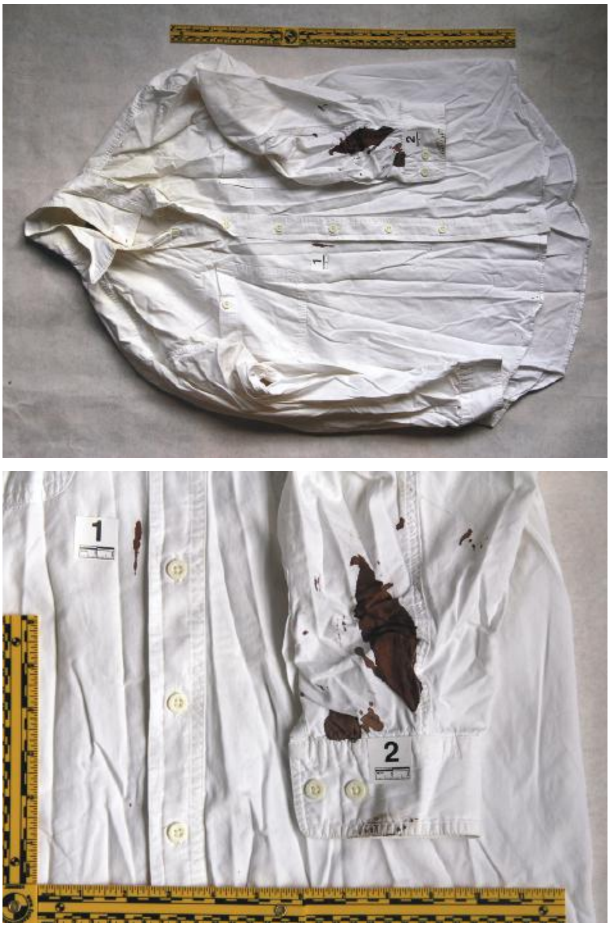

Figure 3.14 Photographic documentation of a bloodstained shirt after visual examination. (© Richard C. Li.)

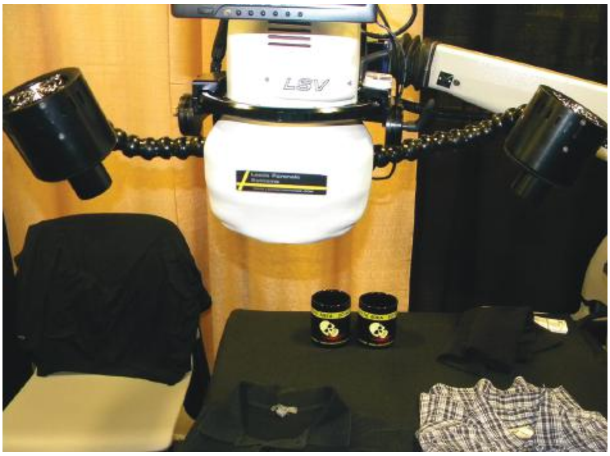

Figure 3.15 A multiwavelength viewing and imaging device used for the examination of various types of evidence, including bodily fluid stains. (© Richard C. Li.)

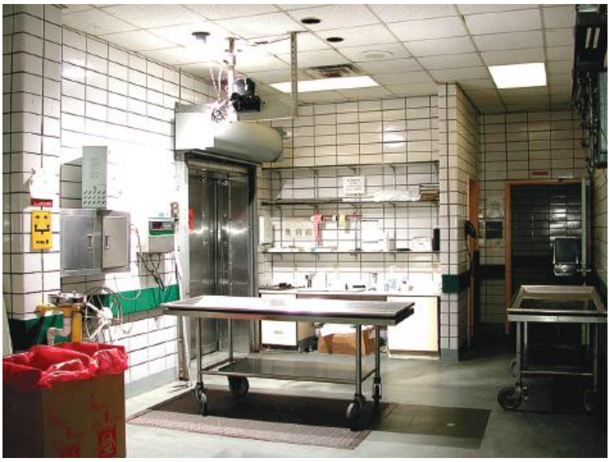

Figure 3.16 View of a forensic pathology facility. (Courtesy of G. Ledwell.)

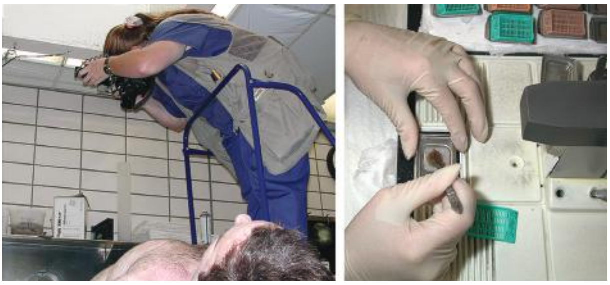

Figure 3.17 Photographic documentation prior to autopsy (left). Preparing specimens for histological sections for forensic pathological examination (right). (Courtesy of G. Ledwell.)

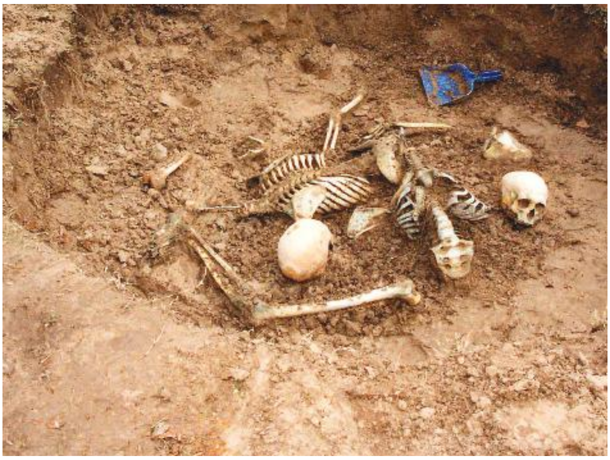

Figure 3.18 Buried human skeletons recovered by forensic anthropologists. (Courtesy of H. Brewster.)

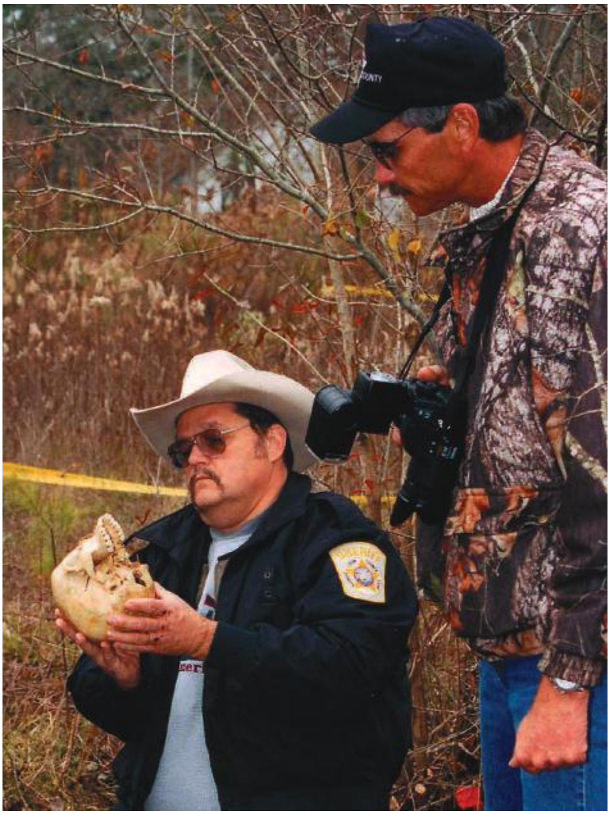

Figure 3.19 Human skull recovered at the scene. (Courtesy of H. Brewster.)

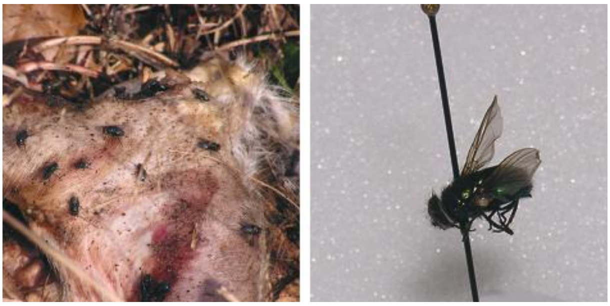

Figure 3.20 Insects found on a dead animal (left). Blow fly specimen. This insect is commonly encountered at crime scenes (right). (Courtesy of K. Wendler.)

3.3.4 Forensic Odontology

Practitioners of forensic odontology participate in the identification of victims whose bodies are left in an unrecognizable state. The characteristics of teeth, their alignment, and the overall structure of the mouth provide evidence that can identify a specific person. Dental records such as x-rays and dental casts allow a forensic odontologist to compare a set of dental remains with an alleged victim. Another application of forensic odontology in a criminal investigation is bite mark analysis (Figure 3.21). A forensic odontologist can analyze the marks left on a victim and compare them with the tooth structures of a suspect to make a comparison.

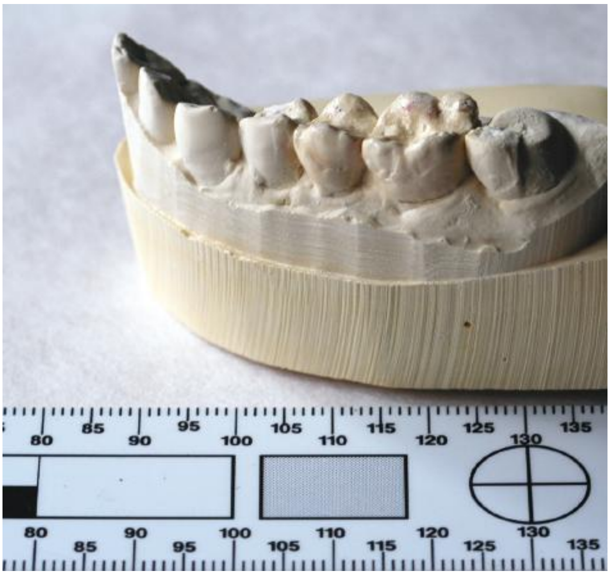
80
85
90
95
100
105
110
115
120
125
130
135

Figure 3.21 A dental cast can be utilized for identifying bite marks. (© Richard C. Li.)

3.4 Brief History of the Development of Forensic Biology

The developmental history of modern forensic biology spans three stages: (1) antigen polymorphism, (2) protein polymorphism, and (3) DNA polymorphism. Figure 3.22 illustrates this history.

3.4.1 Antigen Polymorphism

The human ABO blood groups were discovered in 1900 by Karl Landsteiner in a study of the causes of blood transfusion reactions. Landsteiner's discovery made blood transfusions feasible, and he received the Nobel Prize in 1930 when he revealed the four groups of human blood cells designated A, B, AB, and O. By the 1960s, a dozen more blood group systems had been characterized, and at least 29 systems are currently known (Chapter 18).

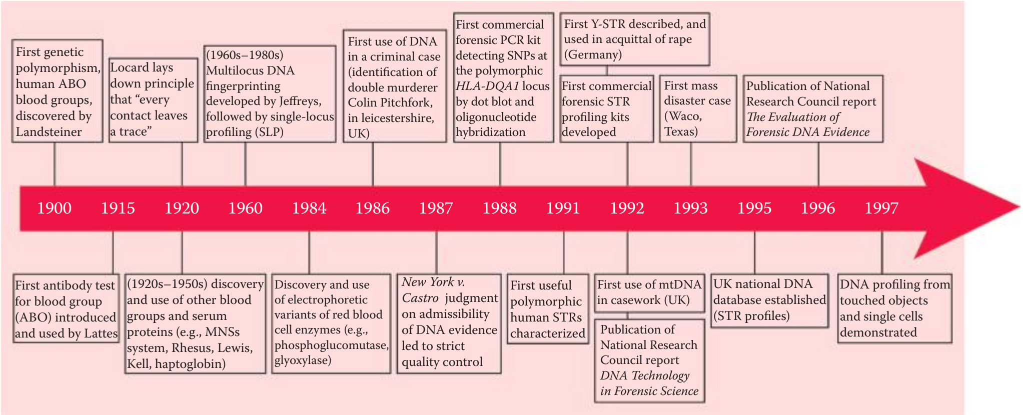
First genetic polymorphism, human ABO blood groups, discovered by Landsteiner
Locard lays down principle that "every contact leaves a trace"
(1960s–1980s) Multilocus DNA fingerprinting developed by Jeffreys, followed by single-locus profiling (SLP)
First use of DNA in a criminal case (identification of double murderer Colin Pitchfork, in leicestershire, UK)
First commercial forensic PCR kit detecting SNPs at the polymorphic *HLA-DQA1* locus by dot blot and oligonucleotide hybridization
First Y-STR described, and used in acquittal of rape (Germany)
First commercial forensic STR profiling kits developed
First mass disaster case (Waco, Texas)
Publication of National Research Council report *The Evaluation of Forensic DNA Evidence*
1900
1915
1920
1960
1984
1986
1987
1988
1991
1992
1993
1995
1996
1997
First antibody test for blood group (ABO) introduced and used by Lattes
(1920s–1950s) discovery and use of other blood groups and serum proteins (e.g., MNSs system, Rhesus, Lewis, Kell, haptoglobin)
Discovery and use of electrophoretic variants of red blood cell enzymes (e.g., phosphoglucomutase, glyoxylase)
New York v. Castro judgment on admissibility of DNA evidence led to strict quality control
First useful polymorphic human STRs characterized
First use of mtDNA in casework (UK)
Publication of National Research Council report *DNA Technology in Forensic Science*
UK national DNA database established (STR profiles)
DNA profiling from touched objects and single cells demonstrated

Figure 3.22 A brief history of the development of forensic biology. (From Jobling, M.A. and Gill, P., *Nat Rev Genet*, 5, 739–751, 2004. With permission.)

Subsequent studies found that the blood types in the ABO system were inherited, and the frequencies with which the four types appeared in specific human populations were found to differ. This led to the discovery of the first antigen polymorphic marker for use in human identification in forensic cases. In the past, forensic laboratories utilized blood group systems in a discipline known as forensic serology. While it is possible to exclude a suspect through the use of blood group typing, the evidence for the inclusion of a suspect is weak due to the high probability of a coincidental match between two unrelated persons.

3.4.2 Protein Polymorphism

Because of the limitations of antigen polymorphism, protein polymorphism was introduced for forensic identification (Chapter 18). Initially, a few polymorphisms in serum proteins and erythrocyte enzymes were reported. By the 1980s, however, approximately a hundred protein polymorphisms had been discovered. A few systems were commonly used in forensic laboratories, including the polymorphisms of erythrocyte enzymes, serum proteins, and hemoglobin. Blood groups and protein polymorphism analysis were combined in forensic investigations to lower the probability of a match between two unrelated individuals. However, more powerful methods were still sought.

3.4.3 DNA Polymorphism

The human genome contains all the necessary biological information for cellular and organ structure and function. It consists of the *nuclear genome* and the *mitochondrial genome* (Chapter 23 discusses the mitochondrial genome). The human nuclear genome, a set of 23 chromosomes, contains approximately 3 billion base pairs (bp). The Human Genome Project was initiated in 1990 to sequence the entire human nuclear genome. In 2003, 99% of the human genome, including the most important parts of the genome, was sequenced. Further analyses on the human genome sequences continue. The genome contains genes and intergenic noncoding sequences.

3.4.3.1 Genes and Related Sequences

Approximately 20,000–25,000 genes have been identified in the human genome, which encode the information for the synthesis of proteins. The functions of nearly half of these genes have been identified. Most encode the proteins that are responsible for the maintenance of the genome, the functioning of the cells, the immune response, and the structural proteins of cells.

Most human genes are discontinuous. The coding regions of genes are called *exons* and are separated by *introns*. During gene expression, the precursor messenger RNA transcript (pre-mRNA), consisting of both the exons and introns, is produced. The mRNA is a template for protein synthesis in which the sequence is based on a complementary strand of DNA. Through the process of splicing, the introns are removed and the exons are joined, producing the spliced mRNA form, which can be used for protein synthesis via the translation process. Other gene-related sequences include those responsible for gene transcription such as *promoter sequences*; those responsible for gene regulation such as *cis-regulatory sequences* (or enhancers); and *untranslated sequences*, which are transcribed but do not encode proteins. Figure 3.23 depicts the features of a representative human gene.

3.4.3.2 Intergenic Noncoding Sequences

More than 90% of the human genome sequence consists of intergenic noncoding sequences located between genes. The functions of these sequences are yet to be discovered. The intergenic noncoding sequences contain large quantities of various types of *repetitive DNA*, which falls into two categories: tandem repeats and genome-wide or interspersed repeats.

Tandem repeats are repeat units placed next to each other in an array. One type is called *satellite DNA* because of the observation of satellite bands containing DNA with tandem repeats

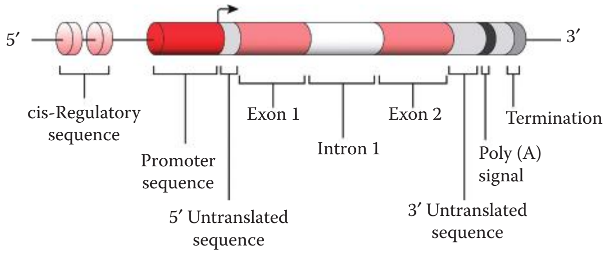
5′
3′
cis-Regulatory sequence
Promoter sequence
5′ Untranslated sequence
Exon 1
Intron 1
Exon 2
3′ Untranslated sequence
Poly (A) signal
Termination

Figure 3.23 Gene structure. Transcription, which can be regulated by the cis-regulatory sequence, is initiated at the transcription start site (arrow) near the promoter. The exons, noncoding introns, and the untranslated sequences are also shown. (© Richard C. Li.)

during density gradient centrifugation. Satellite DNA can be found at centromeres and telomeres consisting of regions composed of long stretches of tandem repeats. Minisatellites and microsatellites are two other types of shorter tandem repeats. *Minisatellites*, also known as *variable number tandem repeats* (VNTRs), form arrays of tandem repeats with a *repeat unit length* from several to hundreds of base pairs. In a *microsatellite*, also known as a *short tandem repeat* (STR) or a *simple sequence repeat* (SSR), the repeat unit length can be 2–6 bp long.

Mobile elements (*interspersed repeats*) are randomly located throughout the human genome (Figure 3.24). Two human types have been characterized: *DNA transposons* and *retrotransposons*. The mobile elements change their locations, a process called *transposition*, by which these sequences are inserted into a new site in the genome. The transposition of DNA transposons is through a "cut-and-paste" mechanism. During transposition, DNA transposons are excised from one site and inserted at a new site in the genome. In contrast, retrotransposons duplicate themselves during transposition and propagate throughout the genome, which is a copy-and-paste mechanism: a copy of the original retrotransposons is generated at the new site and the original copy is retained. Additionally, the transposition of retrotransposons requires an RNA intermediate, a process called *retrotransposition*. Retrotransposons have two subtypes: *long terminal repeat* (LTR) and *non-LTR retroposons*. The non-LTR retroposons can be further divided into two subtypes: *long interspersed elements* (LINEs) and *short interspersed elements* (SINEs). Alu elements are the most abundant type of human SINE. There are more than one million copies of Alu elements in the human genome (Figure 3.25). Some members of the Alu elements are

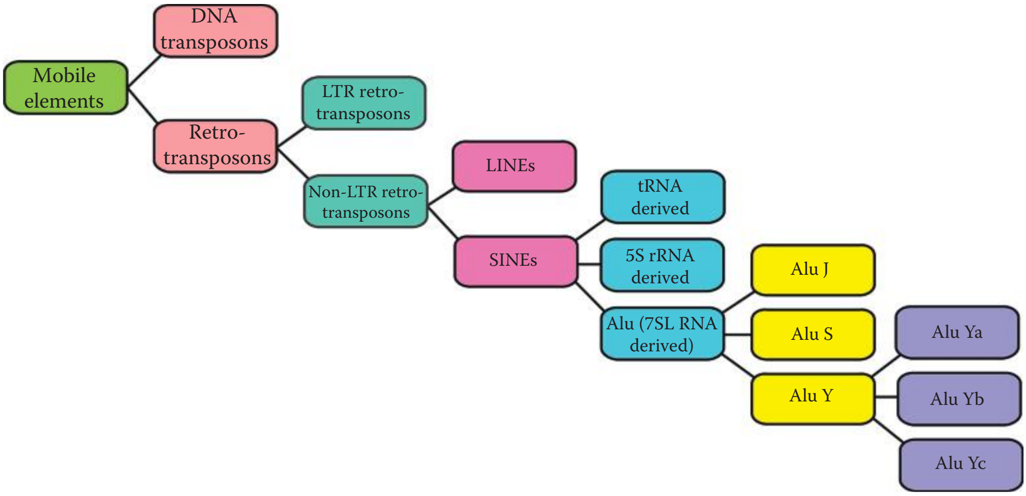
DNA transposons
Mobile elements
LTR retro-transposons
Retro-transposons
LINEs
Non-LTR retro-transposons
tRNA derived
SINEs
5S rRNA derived
Alu J
Alu (7SL RNA derived)
Alu S
Alu Ya
Alu Y
Alu Yb
Alu Yc

Figure 3.24 Mobile elements classification. (© Richard C. Li.)

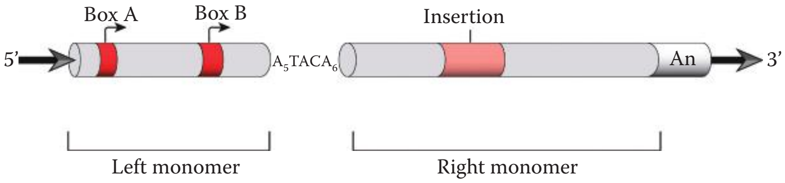

Figure 3.25 Structure of Alu elements. The name, Alu elements, was given since these elements usually contain DNA sequences cleaved by the restriction enzyme *Alu*I. The Alu element is a dimeric structure including two regions separated by an A-rich sequence. The 5′ region has boxes A and B containing RNA polymerase III promoter sequences. The 3′ region contains a 31-nucleotide insertion. There is a short polyadenylation tail located at the 3′ end of the Alu element (terminal A-stretch). Full-length Alu elements are approximately 300 bp long. (© Richard C. Li.)

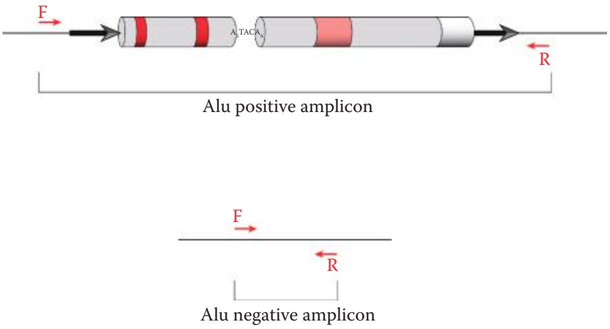

Figure 3.26 Detecting Alu element insertion polymorphism for human identification. The Alu element insertion site is amplified using the polymerase chain reaction assay. The presence and absence of the Alu element at the site can be detected based on the length of the DNA fragment analyzed. F, forward primer; R, reverse primer. (© Richard C. Li.)

polymorphic for the presence or absence of insertion that can potentially be used for forensic human identification (Figure 3.26).

3.4.3.3 Human DNA Polymorphic Markers

Most individual human genome sequences are very similar. However, variations in sequences do occur. The differences between individual genomes that occur at the DNA level are called *DNA polymorphisms*. In particular, a DNA polymorphism with alternative forms of a chromosomal locus that differ in nucleotide sequence is known as a *sequence polymorphism*. A DNA polymorphism that differs in the numbers of tandem repeat units is known as a *length polymorphism*. A DNA polymorphism can occur anywhere in a genome including genes and other chromosomal locations. Many DNA polymorphisms are useful for genetic mapping studies and hence are called *DNA markers*. DNA polymorphisms form the basis of forensic DNA profiling. The focus of this text is on the human genome, but polymorphisms also occur in the genomes of other organisms.

Most DNA polymorphisms are single nucleotide polymorphisms (SNPs) involving a single-base-pair change or a point mutation. Over one million SNPs have been identified. SNPs arise by spontaneous mutation. Most SNPs occur in noncoding regions of the genome, although some appear in coding regions as well. Other forms of DNA polymorphisms are tandem repeats such as STRs and VNTRs. Although their biological functions are unknown, STRs and VNTRs are very useful for forensic DNA analysis. Many are highly polymorphic, and the number of repeat units varies greatly among different individuals of a population. It is unlikely that two unrelated individuals will have exactly the same combination of STR or VNTR polymorphisms if sufficient markers are examined. Thus, a resulting genetic profile can be used for human identification.

Alternative forms of DNA polymorphisms are called *alleles*. The same allele present in both homologous chromosomes is referred to as *homozygous*. Two different alleles present in homologous chromosomes are referred to as *heterozygous*. A combination of alleles at a given locus is a *genotype*. In forensic analysis, the genotype for a panel of analyzed loci is called the *DNA profile*.

3.4.3.4 Forensic DNA Polymorphism Profiling

In 1984, Sir Alec Jeffreys (Figure 3.27) developed a DNA profiling technique using a VNTR technique involving multilocus profiling and later followed by single-locus profiling (Chapter 19). This technique led to the solving of a double murder that had been committed in Leicestershire (United Kingdom) in the 1980s. The case was the first to apply DNA evidence to a criminal investigation. During the investigation, DNA profiling not only identified the true perpetrator but it also excluded an innocent suspect. This case demonstrated DNA profiling's great potential in forensic investigations.

DNA profiling offered a number of advantages compared with earlier systems. The most important is the ability of the technique to reveal far greater individual variability in DNA than can be revealed by antigen and protein polymorphic markers. The probability of two unrelated individuals having the same DNA profile is very low. The great variability of DNA polymorphisms has made it possible to offer strong support for concluding that DNA from a suspect and from a crime scene originated from the same source. This technique was subsequently implemented in forensic laboratories worldwide.

In the mid-1980s, Kary Mullis and his coworkers developed the polymerase chain reaction (PCR) technique, which amplifies a small quantity of DNA. Mullis's invention had a powerful impact on molecular biology and earned him a Nobel Prize in 1993. Since the introduction of PCR, new techniques have been developed for forensic DNA testing purposes.

Figure 3.27 Sir Alec Jeffreys. (© Richard C. Li.)

The application of PCR-based assays makes forensic DNA analysis possible when only minute quantities of DNA can be recovered from a crime scene, for example, from hairs and cigarette butts. These assays have greatly increased the sensitivity of forensic DNA testing. The first forensic application of a PCR-based assay utilizing SNPs at the *HLA-DQA1* locus (formerly called DQα) was developed in 1986 (Chapter 22). One major disadvantage of the assay was the high probability (approximately 1 in 4000) of a match between two unrelated persons. Amplified fragment length polymorphism (AFLP) at the D1S80 locus has also been implemented in forensic laboratories. The D1S80 locus is a small-size VNTR marker that can be amplified by PCR. The *HLA-DQA1* and AFLP assays were used for some years until the introduction of STR assays.

In the late 1990s, forensic laboratories started utilizing STR loci. STRs have a number of advantages compared with VNTRs. For example, STRs can be amplified by PCR because of their smaller size, which greatly increases the sensitivity of the assay. Furthermore, STR markers are as highly variable as VNTRs. With the application of multiple STR loci, the probability of a match between two unrelated persons becomes extremely low. As a result of DNA testing, perpetrators have been identified and wrongly convicted innocent people have been exonerated (Chapter 20).

In 1995, the United Kingdom established the first national DNA database for criminal investigations. By the end of 1998, several other countries, including the United States, had created their own national DNA databases. The United States has selected 13 STR loci for the Combined DNA Index System (CODIS). These national DNA databases play important roles in solving criminal cases.

Another technique known as mitochondrial DNA (mtDNA) profiling has also been used for forensic testing. mtDNA is maternally inherited genetic material and is therefore particularly useful for human identification. Each cell contains multiple copies of mtDNA. Thus, mtDNA typing can be useful for analysis when nuclear DNA is severely degraded or present in very limited amounts, such as in cases involving decomposed human remains. Alternatively, polymorphic markers at the Y chromosome have also been utilized for forensic DNA testing. Y chromosomal markers are paternally inherited so they can be used for paternity testing. These markers are also very useful in analyzing DNA from multiple contributors in sexual assault cases.

Bibliography

Aleksander, A., Forensic expertise and judicial practice: Evidence or proof? *J Eval Clin Pract*, 2012, **18**(6): 1147–1150.

Alvarez-Cubero, M.J., et al., Genetic identification of missing persons: DNA analysis of human remains and compromised samples. *Pathobiology*, 2012, **79**(5): 228–238.

Andelinovic, S., et al., Twelve-year experience in identification of skeletal remains from mass graves. *Croat Med J*, 2005, **46**(4): 530–539.

Apostolov, A., S. Hristov, and E. Angelova, DNA identification of biological traces and interpretation in a sexual assault case. *Am J Forensic Med Pathol*, 2009, **30**(1): 57–60.

Aquila, I., et al., The role of forensic botany in crime scene investigation: Case report and review of literature. *J Forensic Sci*, 2014, **59**(3): 820–824.

Asari, M., et al., Multiplex PCR-based Alu insertion polymorphisms genotyping for identifying individuals of Japanese ethnicity. *Genomics*, 2012, **99**(4): 227–232.

Auvdel, M.J., Comparison of laser and ultraviolet techniques used in the detection of body secretions. *J Forensic Sci*, 1987, **32**(2): 326–345.

Baraybar, J.P., When DNA is not available, can we still identify people? Recommendations for best practice. *J Forensic Sci*, 2008, **53**(3): 533–540.

Bauer, C.M., et al., Molecular genetic investigations on Austria's patron saint Leopold III. *Forensic Sci Int Genet*, 2013, **7**(2): 313–315.

Bogdanowicz, W., et al., Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus. *Proc Natl Acad Sci U S A*, 2009, **106**(30): 12279–12282.

Boljuncic, J., DNA analysis of early mediaeval individuals from the Zvonimirovo burial site in Northern Croatia: Investigation of kinship relationships by using multiplex system amplification for short tandem repeat loci. *Croat Med J*, 2007, **48**(4): 536–546.

Bond, J.W., Value of DNA evidence in detecting crime. *J Forensic Sci*, 2007, **52**(1): 128–136.

Boric, I., J. Ljubkovic, and D. Sutlovic, Discovering the 60 years old secret: Identification of the World War II mass grave victims from the island of Daksa near Dubrovnik, Croatia. *Croat Med J*, 2011, **52**(3): 327–335.

Brettell, T.A., et al., Forensic science. *Anal Chem*, 1999, **71**(12): 235R–255R.

Bruce-Chwatt, R.M., A brief history of forensic odontology since 1775. *J Forensic Leg Med*, 2010, **17**(3): 127–130.

Budowle, B. and A. van Daal, Extracting evidence from forensic DNA analyses: Future molecular biology directions. *Biotechniques*, 2009, **46**(5): 339–340, 342–350.

Budowle, B., et al., Role of law enforcement response and microbial forensics in investigation of bioterrorism. *Croat Med J*, 2007, **48**(4): 437–449.

Burg, A., R. Kahn, and K. Welch, DNA testing of sexual assault evidence: The laboratory perspective. *J Forensic Nurs*, 2011, **7**(3): 145–152.

Byard, R.W., K. Both, and E. Simpson, The identification of submerged skeletonized remains. *Am J Forensic Med Pathol*, 2008, **29**(1): 69–71.

Caenazzo, L., P. Tozzo, and D. Rodriguez, Ethical issues in DNA identification of human biological material from mass disasters. *Prehosp Disaster Med*, 2013, **28**(4): 393–396.

Caramelli, D., et al., Genetic analysis of the skeletal remains attributed to Francesco Petrarca. *Forensic Sci Int*, 2007, **173**(1): 36–40.

Carey, L. and L. Mitnik, Trends in DNA forensic analysis. *Electrophoresis*, 2002, **23**(10): 1386–1397.

Cattaneo, C., et al., Unidentified bodies and human remains: An Italian glimpse through a European problem. *Forensic Sci Int*, 2010, **195**(1–3): 167, e1–e6.

Cavard, S., et al., Forensic and police identification of “X” bodies. A 6-years French experience. *Forensic Sci Int*, 2011, **204**(1–3): 139–143.

Charlier, P., et al., Genetic comparison of the head of Henri IV and the presumptive blood from Louis XVI (both Kings of France). *Forensic Sci Int*, 2013, **226**(1–3): 38–40.

Claes, P., et al., Computerized craniofacial reconstruction: Conceptual framework and review. *Forensic Sci Int*, 2010, **201**(1–3): 138–145.

Coble, M.D., et al., Mystery solved: The identification of the two missing Romanov children using DNA analysis. *PLoS One*, 2009, **4**(3): e4838.

Connery, S.A., Three decade old cold case murder solved with evidence from a sexual assault kit. *J Forensic Leg Med*, 2013, **20**(4): 355–356.

Cordner, S.M., N. Woodford, and R. Bassed, Forensic aspects of the 2009 Victorian Bushfires Disaster. *Forensic Sci Int*, 2011, **205**(1–3): 2–7.

Courts, C. and B. Madea, Genetics of the sudden infant death syndrome. *Forensic Sci Int*, 2010, **203**(1–3): 25–33.

Courts, C. and B. Madea, Full STR profile of a 67-year-old bone found in a fresh water lake. *J Forensic Sci*, 2011, **56**(Suppl 1): S172–S175.

Craft, K.J., J.D. Owens, and M.V. Ashley, Application of plant DNA markers in forensic botany: Genetic comparison of *Quercus* evidence leaves to crime scene trees using microsatellites. *Forensic Sci Int*, 2007, **165**(1): 64–70.

Davis, G.G., Forensic toxicology. *Clin Lab Sci*, 2012, **25**(2): 120–124.

Davoren, J., et al., Highly effective DNA extraction method for nuclear short tandem repeat testing of skeletal remains from mass graves. *Croat Med J*, 2007, **48**(4): 478–485.

De Ungria, M.C., et al., Forensic DNA evidence and the death penalty in the Philippines. *Forensic Sci Int Genet*, 2008, **2**(4): 329–332.

Definis Gojanovic, M. and D. Sutlovic, Skeletal remains from World War II mass grave: From discovery to identification. *Croat Med J*, 2007, **48**(4): 520–527.

Dekeirsschieter, J., et al., Forensic entomology investigations from Doctor Marcel Leclercq (1924–2008): A review of cases from 1969 to 2005. *J Med Entomol*, 2013, **50**(5): 935–954.

Deng, Y.J., et al., Preliminary DNA identification for the tsunami victims in Thailand. *Genomics Proteom Bioinform*, 2005, **3**(3): 143–157.

Dirkmaat, D.C., et al., New perspectives in forensic anthropology. *Am J Phys Anthropol*, 2008, **137**(Suppl 47): 33–52.

Dissing, J., et al., The last Viking king: A royal maternity case solved by ancient DNA analysis. *Forensic Sci Int*, 2007, **166**(1): 21–27.

Djuric, M., et al., Identification of victims from two mass-graves in Serbia: A critical evaluation of classical markers of identity. *Forensic Sci Int*, 2007, **172**(2–3): 125–129.

Dobberstein, R.C., et al., Degradation of biomolecules in artificially and naturally aged teeth: Implications for age estimation based on aspartic acid racemization and DNA analysis. *Forensic Sci Int*, 2008, **179**(2–3): 181–191.

Dolan, S.M., et al., The emerging role of genetics professionals in forensic kinship DNA identification after a mass fatality: Lessons learned from Hurricane Katrina volunteers. *Genet Med*, 2009, **11**(6): 414–417.

Dzijan, S., et al., Evaluation of the reliability of DNA typing in the process of identification of war victims in Croatia. *J Forensic Sci*, 2009, **54**(3): 608–609.

Enserink, M., Can this DNA sleuth help catch criminals? *Science*, 2011, **331**(6019): 838–840.

Ferri, G., et al., Forensic botany: Species identification of botanical trace evidence using a multigene barcoding approach. *Int J Legal Med*, 2009, **123**(5): 395–401.

Fondevila, M., et al., Case report: Identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur. *Forensic Sci Int Genet*, 2008, **2**(3): 212–218.

Foran, D.R., et al., The conviction of Dr. Crippen: New forensic findings in a century-old murder. *J Forensic Sci*, 2011, **56**(1): 233–240.

Foster, E.A., et al., Jefferson fathered slave's last child. *Nature*, 1998, **396**(6706): 27–28.

Ge, J., B. Budowle, and R. Chakraborty, Choosing relatives for DNA identification of missing persons. *J Forensic Sci*, 2011, **56**(Suppl 1): S23–S28.

Geserick, G. and I. Wirth, Genetic kinship investigation from blood groups to DNA markers. *Transfus Med Hemother*, 2012, **39**(3): 163–175.

Gewin, V., Forensics: The call of the crime lab. *Nature*, 2011, **473**(7347): 409–411.

Gill, P., Role of short tandem repeat DNA in forensic casework in the UK—Past, present, and future perspectives. *Biotechniques*, 2002, **32**(2): 366–368, 370, 372, passim.

Gill, P., A.J. Jeffreys, and D.J. Werrett, Forensic application of DNA "fingerprints". *Nature*, 1985, **318**(6046): 577–579.

Gill, P., et al., Identification of the remains of the Romanov family by DNA analysis. *Nat Genet*, 1994, **6**(2): 130–135.

Gingerich, O., The Copernicus grave mystery. *Proc Natl Acad Sci U S A*, 2009, **106**(30): 12215–12216.

Gonzalez-Andrade, F., et al., Two fathers for the same child: A deficient paternity case of false inclusion with autosomic STRs. *Forensic Sci Int Genet*, 2009, **3**(2): 138–140.

Graham, E.A., DNA reviews: Low level DNA profiling. *Forensic Sci Med Pathol*, 2008, **4**(2): 129–131.

Grellner, W. and B. Madea, Demands on scientific studies: Vitality of wounds and wound age estimation. *Forensic Sci Int*, 2007, **165**(2–3): 150–154.

Hampikian, G., E. West, and O. Akselrod, The genetics of innocence: Analysis of 194 U.S. DNA exonerations. *Annu Rev Genomics Hum Genet*, 2011, **12**: 97–120.

Hartman, D., et al., The contribution of DNA to the disaster victim identification (DVI) effort. *Forensic Sci Int*, 2011, **205**(1–3): 52–58.

Hartman, D., et al., The importance of Guthrie cards and other medical samples for the direct matching of disaster victims using DNA profiling. *Forensic Sci Int*, 2011, **205**(1–3): 59–63.

Hatsch, D., et al., A rape case solved by mitochondrial DNA mixture analysis. *J Forensic Sci*, 2007, **52**(4): 891–894.

Hinchliffe, J., Forensic odontology, part 4. Human bite marks. *Br Dent J*, 2011, **210**(8): 363–368.

Holland, M.M., et al., Mitochondrial DNA sequence analysis of human skeletal remains: Identification of remains from the Vietnam War. *J Forensic Sci*, 1993, **38**(3): 542–553.

Holland, M.M., et al., Development of a quality, high throughput DNA analysis procedure for skeletal samples to assist with the identification of victims from the World Trade Center attacks. *Croat Med J*, 2003, **44**(3): 264–272.

Holobinko, A., Forensic human identification in the United States and Canada: A review of the law, admissible techniques, and the legal implications of their application in forensic cases. *Forensic Sci Int*, 2012, **222**(1–3): 394, e1–e13.

Honda, K., The Ashikaga case of Japan—Y-STR testing used as the exculpatory evidence to free a convicted felon after 17.5 years in prison. *Forensic Sci Int Genet*, 2013, **7**(1): e1–e2.

Horsman-Hall, K.M., et al., Development of STR profiles from firearms and fired cartridge cases. *Forensic Sci Int Genet*, 2009, **3**(4): 242–250.

Howard, C., et al., A *Cannabis sativa* STR genotype database for Australian seizures: Forensic applications and limitations. *J Forensic Sci*, 2009, **54**(3): 556–563.

Howlett, R., DNA forensics and the FBI. *Nature*, 1989, **341**(6239): 182–183.

Hughes, V.K. and N.E. Langlois, Use of reflectance spectrophotometry and colorimetry in a general linear model for the determination of the age of bruises. *Forensic Sci Med Pathol*, 2010, **6**(4): 275–281.

Irwin, J.A., et al., DNA identification of "Earthquake McGoon" 50 years postmortem. *J Forensic Sci*, 2007, **52**(5): 1115–1118.

Ishiko, A., et al., Experimental studies on identification of the driver based on STR analysis. *Leg Med (Tokyo)*, 2008, **10**(3): 115–118.

Janisch, S., et al., Analysis of clinical forensic examination reports on sexual assault. *Int J Legal Med*, 2010, **124**(3): 227–235.

Jeffreys, A.J., Genetic fingerprinting. *Nat Med*, 2005, **11**(10): 1035–1039.

Jewkes, R., et al., Medico-legal findings, legal case progression, and outcomes in South African rape cases: Retrospective review. *PLoS Med*, 2009, **6**(10): e1000164.

Jobling, M.A. and P. Gill, Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*, 2004, **5**(10): 739–751.

Just, R.S., et al., Titanic's unknown child: The critical role of the mitochondrial DNA coding region in a re-identification effort. *Forensic Sci Int Genet*, 2011, **5**(3): 231–235.

Karlsson, A.O., et al., DNA-testing for immigration cases: The risk of erroneous conclusions. *Forensic Sci Int*, 2007, **172**(2–3): 144–149.

Kayser, M. and P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: Motivations, scientific challenges, and ethical considerations. *Forensic Sci Int Genet*, 2009, **3**(3): 154–161.

Kim, N.Y., et al., A genetic investigation of Korean mummies from the Joseon Dynasty. *Mol Biol Rep*, 2011, **38**(1): 115–121.

Kjellstrom, A., et al., An analysis of the alleged skeletal remains of Carin Goring. *PLoS One*, 2012, **7**(12): e44366.

Krishan, K., Individualizing characteristics of footprints in Gujjars of North India—Forensic aspects. *Forensic Sci Int*, 2007, **169**(2–3): 137–144.

Kupiec, T. and W. Branicki, Genetic examination of the putative skull of Jan Kochanowski reveals its female sex. *Croat Med J*, 2011, **52**(3): 403–409.

Lalueza-Fox, C., et al., Genetic analysis of the presumptive blood from Louis XVI, King of France. *Forensic Sci Int Genet*, 2011, **5**(5): 459–463.

Leclair, B., et al., Bioinformatics and human identification in mass fatality incidents: The World Trade Center disaster. *J Forensic Sci*, 2007, **52**(4): 806–819.

Lee, H.C., et al., DNA typing in forensic science. I. Theory and background. *Am J Forensic Med Pathol*, 1994, **15**(4): 269–282.

Lee, H.Y., et al., Genetic characterization and assessment of authenticity of ancient Korean skeletal remains. *Hum Biol*, 2008, **80**(3): 239–250.

Lee, H.Y., et al., DNA typing for the identification of old skeletal remains from Korean War victims. *J Forensic Sci*, 2010, **55**(6): 1422–1429.

Lehman, D.C., Introduction to forensic science. *Clin Lab Sci*, 2012, **25**(2): 107–108.

Li, X., et al., Mitochondrial DNA and STR analyses for human DNA from maggots crop contents: A forensic entomology case from central-southern China. *Trop Biomed*, 2011, **28**(2): 333–338.

Loreille, O.M., et al., Integrated DNA and fingerprint analyses in the identification of 60-year-old mummified human remains discovered in an Alaskan glacier. *J Forensic Sci*, 2010, **55**(3): 813–818.

Luna, A., Is postmortem biochemistry really useful? Why is it not widely used in forensic pathology? *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S27–S30.

Madea, B., et al., Molecular pathology in forensic medicine—Introduction. *Forensic Sci Int*, 2010, **203**(1–3): 3–14.

Maiquilla, S.M., et al., Y-STR DNA analysis of 154 female child sexual assault cases in the Philippines. *Int J Legal Med*, 2011, **125**(6): 817–824.

Manasatienkij, C. and C. Ra-ngabpai, Clinical application of forensic DNA analysis: A literature review. *J Med Assoc Thai*, 2012, **95**(10): 1357–1363.

Manhart, J., A. Bittorf, and A. Buttner, Disaster victim identification-experiences of the "Autobahn A19" disaster. *Forensic Sci Med Pathol*, 2012, **8**(2): 118–124.

Manjunath, B.C., et al., DNA profiling and forensic dentistry—A review of the recent concepts and trends. *J Forensic Leg Med*, 2011, **18**(5): 191–197.

Marjanovic, D., et al., DNA identification of skeletal remains from the World War II mass graves uncovered in Slovenia. *Croat Med J*, 2007, **48**(4): 513–519.

Marjanovic, D., et al., Identification of skeletal remains of Communist Armed Forces victims during and after World War II: Combined Y-chromosome (STR) and MiniSTR approach. *Croat Med J*, 2009, **50**(3): 296–304.

Matheson, C.D., et al., Molecular exploration of the first-century Tomb of the Shroud in Akeldama, Jerusalem. *PLoS One*, 2009, **4**(12): e8319.

McDonald, J. and D.C. Lehman, Forensic DNA analysis. *Clin Lab Sci*, 2012, **25**(2): 109–113.

Meissner, C. and S. Ritz-Timme, Molecular pathology and age estimation. *Forensic Sci Int*, 2010, **203**(1–3): 34–43.

Melchior, L., et al., Evidence of authentic DNA from Danish Viking Age skeletons untouched by humans for 1000 years. *PLoS One*, 2008, **3**(5): e2214.

Milde-Kellers, A., et al., An illicit love affair during the Third Reich: Who is my grandfather? *J Forensic Sci*, 2008, **53**(2): 377–379.

Monckton, D.G. and A.J. Jeffreys, DNA profiling. *Curr Opin Biotechnol*, 1993, **4**(6): 660–664.

Montelius, K. and B. Lindblom, DNA analysis in disaster victim identification. *Forensic Sci Med Pathol*, 2012, **8**(2): 140–147.

Mundorff, A.Z., E.J. Bartelink, and E. Mar-Cash, DNA preservation in skeletal elements from the World Trade Center disaster: Recommendations for mass fatality management. *J Forensic Sci*, 2009, **54**(4): 739–745.

Nakamura, Y., DNA variations in human and medical genetics: 25 years of my experience. *J Hum Genet*, 2009, **54**(1): 1–8.

National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002. Washington, DC: US Department of Justice, Office of Justice Programs.

Neuhuber, F., et al., An unusual case of identification by DNA analysis of siblings. *Forensic Sci Int Genet*, 2012, **6**(1): 121–123.

Njoroge, S.K., et al., Microchip electrophoresis of Alu elements for gender determination and inference of human ethnic origin. *Electrophoresis*, 2010, **31**(6): 981–990.

Nunez, C., et al., Genetic analysis of 7 medieval skeletons from the Aragonese Pyrenees. *Croat Med J*, 2011, **52**(3): 336–343.

Nuzzolese, E. and G. Di Vella, Future project concerning mass disaster management: A forensic odontology prospectus. *Int Dent J*, 2007, **57**(4): 261–266.

Onoja, A.M., Paternity testing. *Niger J Med*, 2011, **20**(4): 406–408.

Opdal, S.H., et al., Mitochondrial tRNA genes and flanking regions in sudden infant death syndrome. *Acta Paediatr*, 2007, **96**(2): 211–214.

Page, M., J. Taylor, and M. Blenkin, Expert interpretation of bitemark injuries—A contemporary qualitative study. *J Forensic Sci*, 2013, **58**(3): 664–672.

Palo, J.U., et al., Repatriation and identification of the Finnish World War II soldiers. *Croat Med J*, 2007, **48**(4): 528–535.

Park, D.K., et al., The role of forensic anthropology in the examination of the Daegu subway disaster (2003, Korea). *J Forensic Sci*, 2009, **54**(3): 513–518.

Parson, W., et al., Unravelling the mystery of Nanga Parbat. *Int J Legal Med*, 2007, **121**(4): 309–310.

Piccinini, A., et al., World War One Italian and Austrian soldier identification project: DNA results of the first case. *Forensic Sci Int Genet*, 2010, **4**(5): 329–333.

Pietrangeli, I., et al., Forensic DNA challenges: Replacing numbers with names of Fosse Ardeatine's victims. *J Forensic Sci*, 2009, **54**(4): 905–908.

Pinckard, J.K., Memorial Eckert paper for 2007 forensic DNA analysis for the medical examiner. *Am J Forensic Med Pathol*, 2008, **29**(4): 375–381.

Pollak, S., Medical criminalistics. *Forensic Sci Int*, 2007, **165**(2–3): 144–149.

Pomara, C., et al., A medieval murder. *Am J Forensic Med Pathol*, 2008, **29**(1): 72–74.

Prainsack, B. and M. Kitzberger, DNA behind bars: Other ways of knowing forensic DNA technologies. *Soc Stud Sci*, 2009, **39**(1): 51–79.

Pretty, I.A., Forensic dentistry: 1. Identification of human remains. *Dent Update*, 2007, **34**(10): 621–622, 624–626, 629–630, passim.

Pretty, I.A., Forensic dentistry: 2. Bitemarks and bite injuries. *Dent Update*, 2008, **35**(1): 48–50, 53–54, 57–58 passim.

Primorac, D. and M.S. Schanfield, Application of forensic DNA testing in the legal system. *Croat Med J*, 2000, **41**(1): 32–46.

Prottas, J.M. and A.A. Noble, Use of forensic DNA evidence in prosecutors' offices. *J Law Med Ethics*, 2007, **35**(2): 310–315.

Puentes, K., et al., Three-dimensional reconstitution of bullet trajectory in gunshot wounds: A case report. *J Forensic Leg Med*, 2009, **16**(7): 407–410.

Raymond, J.J., et al., Trace DNA analysis: Do you know what your neighbour is doing? A multi-jurisdictional survey. *Forensic Sci Int Genet*, 2008, **2**(1): 19–28.

Rios, L., et al., Identification process in mass graves from the Spanish Civil War II. *Forensic Sci Int*, 2012, **219**(1–3): e4–e9.

Roewer, L., Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, 2009, **5**(2): 77–84.

Rogaev, E.I., et al., Genomic identification in the historical case of the Nicholas II royal family. *Proc Natl Acad Sci U S A*, 2009, **106**(13): 5258–5263.

Roper, S.M. and O.L. Tatum, Forensic aspects of DNA-based human identity testing. *J Forensic Nurs*, 2008, **4**(4): 150–156.

Ruffell, A. and A. Sandiford, Maximising trace soil evidence: An improved recovery method developed during investigation of a $26 million bank robbery. *Forensic Sci Int*, 2011, **209**(1–3): e1–e7.

Rutkowska, J., et al., Donor DNA is detected in recipient blood for years after kidney transplantation using sensitive forensic medicine methods. *Ann Transplant*, 2007, **12**(3): 12–14.

Saigusa, K., et al., Practical applications of molecular biological species identification of forensically important flies. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S344–S347.

Sato, Y., et al., Multiplex STR typing of aortic tissues from unidentified cadavers. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S455–S457.

Schou, M.P. and P.J. Knudsen, The Danish disaster victim identification effort in the Thai tsunami: Organisation and results. *Forensic Sci Med Pathol*, 2012, **8**(2): 125–130.

Schwark, T., A. Heinrich, and N. von Wurmb-Schwark, Genetic identification of highly putrefied bodies using DNA from soft tissues. *Int J Legal Med*, 2011, **125**(6): 891–894.

Schwark, T., et al., Reliable genetic identification of burnt human remains. *Forensic Sci Int Genet*, 2011, **5**(5): 393–399.

Scott, R. and C. Skellern, DNA evidence in jury trials: The "CSI effect". *J Law Med*, 2010, **18**(2): 239–262.

Seo, Y., et al., STR and mitochondrial DNA SNP typing of a bone marrow transplant recipient after death in a fire. *Leg Med (Tokyo)*, 2012, **14**(6): 331–335.

Skinner, M., D. Alempijevic, and A. Stanojevic, In the absence of dental records, do we need forensic odontologists at mass grave sites? *Forensic Sci Int*, 2010, **201**(1–3): 22–26.

Skopp, G., Postmortem toxicology. *Forensic Sci Med Pathol*, 2010, **6**(4): 314–325.

Speller, C.F., et al., Personal identification of cold case remains through combined contribution from anthropological, mtDNA, and bomb-pulse dating analyses. *J Forensic Sci*, 2012, **57**(5): 1354–1360.

Stadlbauer, C., et al., History of individuals of the 18th/19th centuries stored in bones, teeth, and hair analyzed by LA-ICP-MS—A step in attempts to confirm the authenticity of Mozart's skull. *Anal Bioanal Chem*, 2007, **388**(3): 593–602.

Stene, L.E., K. Ormstad, and B. Schei, Implementation of medical examination and forensic analyses in the investigation of sexual assaults against adult women: A retrospective study of police files and medical journals. *Forensic Sci Int*, 2010, **199**(1–3): 79–84.

Sudoyo, H., et al., DNA analysis in perpetrator identification of terrorism-related disaster: Suicide bombing of the Australian Embassy in Jakarta 2004. *Forensic Sci Int Genet*, 2008, **2**(3): 231–237.

Suzuki, S., et al., Guilty by his fibers: Suspect confession versus textile fibers reconstructed simulation. *Forensic Sci Int*, 2009, **189**(1–3): e27–e32.

Swift, A.M., Case three: Ethics of coercion. *Clin Lab Sci*, 2008, **21**(2): 122–123.

Thompson, R., S. Zoppis, and B. McCord, An overview of DNA typing methods for human identification: Past, present, and future. *Methods Mol Biol*, 2012, **830**: 3–16.

Tokdemir, M., et al., Forensic value of gunpowder tattooing in identification of multiple entrance wounds from one bullet. *Leg Med (Tokyo)*, 2007, **9**(3): 147–150.

Tokutomi, T., et al., Identification using DNA from skin contact: Case reports. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S576–S577.

Tracey, M., Short tandem repeat-based identification of individuals and parents. *Croat Med J*, 2001, **42**(3): 6.
Travis, J., Forensic science. Scientists decry isotope, DNA testing of "nationality". *Science*, 2009, **326**(5949): 30–31.
Ubelaker, D.H., The forensic evaluation of burned skeletal remains: A synthesis. *Forensic Sci Int*, 2009, **183**(1–3): 1–5.
US Department of Justice, Office of Justice Programs, National Institute of Justice (NIOJ), *Forensic Science: Review of Status and Needs*, 1999. Washington, DC: US Department of Justice.
US Department of Justice, Office of Justice Programs, National Institute of Justice (NIOJ), *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group*, 2000. Washington, DC: US Department of Justice.
Vanek, D., L. Saskova, and H. Koch, Kinship and Y-chromosome analysis of 7th century human remains: Novel DNA extraction and typing procedure for ancient material. *Croat Med J*, 2009, **50**(3): 286–295.
Varsha, DNA fingerprinting in the criminal justice system: An overview. *DNA Cell Biol*, 2006, **25**(3): 181–188.
Vaz, M. and F.S. Benfica, The experience of the forensic anthropology service of the medical examiner's office in Porto Alegre, Brazil. *Forensic Sci Int*, 2008, **179**(2–3): e45–e49.
Verdiani, S., et al., An unusual observation of tetragametic chimerism: Forensic aspects. *Int J Legal Med*, 2009, **123**(5): 431–435.
von Wurmb-Schwark, N. and T. Schwark, Genetic determination of sibship and twin zygosity in a case of an alleged double infant homicide. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S510–S511.
von Wurmb-Schwark, N., et al., A new multiplex-PCR comprising autosomal and y-specific STRs and mitochondrial DNA to analyze highly degraded material. *Forensic Sci Int Genet*, 2009, **3**(2): 96–103.
Walsh, S.J., Recent advances in forensic genetics. *Expert Rev Mol Diagn*, 2004, **4**(1): 31–40.
Weedn, V.W. and H.J. Baum, DNA identification in mass fatality incidents. *Am J Forensic Med Pathol*, 2011, **32**(4): 393–397.
Wenk, R.E., Testing for parentage and kinship. *Curr Opin Hematol*, 2004, **11**(5): 357–361.
Williamson, R. and R. Duncan, DNA testing for all. *Nature*, 2002, **418**(6898): 585–586.
Willott, G.M., The role of the forensic biologist in cases of sexual assault. *J Forensic Sci Soc*, 1975, **15**(4): 269–276.
Wilms, H.R., et al., Evaluation of autopsy and police reports in the investigation of sudden unexplained death in the young. *Forensic Sci Med Pathol*, 2012, **8**(4): 380–389.
Wyman, J.F., Principles and procedures in forensic toxicology. *Clin Lab Med*, 2012, **32**(3): 493–507.
Zaya, D.N. and M.V. Ashley, Plant genetics for forensic applications. *Methods Mol Biol*, 2012, **862**: 35–52.
Zietkiewicz, E., et al., Current genetic methodologies in the identification of disaster victims and in forensic analysis. *J Appl Genet*, 2012, **53**(1): 41–60.
Zupanic Pajnic, I., B. Gornjak Pogorelc, and J. Balazic, Molecular genetic identification of skeletal remains from the Second World War Konfin I mass grave in Slovenia. *Int J Legal Med*, 2010, **124**(4): 307–317.

4

Sources of Biological Evidence

Biological evidence analysis is one of the standard forensic examinations in the investigation of a wide variety of crimes. In particular, DNA evidence facilitates investigators' efforts to link offenders to crime scenes by matching DNA profiles. DNA evidence can also be used to eliminate suspects. The DNA evidence that is routinely encountered at crime scenes can often be categorized into several groups or types. Table 4.1 lists sources of DNA that are frequently found on personal items. Figures 4.1 and 4.2 illustrate representative types of evidence that are processed and their success rates.

4.1 Bodily Fluids

Bodily fluids and their stains are useful biological evidence for forensic, serological, and DNA analysis and may be useful in solving crimes. The most common bodily fluids in forensic analysis are blood, seminal fluid, and saliva. Blood evidence, such as peripheral blood that circulates through the heart, arteries, veins, and capillaries, is one of the most common types of biological evidence that is found at crime scenes (Chapter 12). The fluid portion of blood is called *plasma*, a subcompartment of extracellular fluid, which is the bodily fluid outside cells. Blood contains various suspended blood cells. The cellular portion of the blood consists of *erythrocytes* (also known as red blood cells), *leucocytes* (also known as white blood cells), and *platelets*. Because mature human erythrocytes and platelets do not have nuclei, they are not useful sources of nuclear DNA. For forensic DNA profiling, the nuclear DNA in blood samples (Figure 4.3) is primarily isolated from leucocytes, which are nucleated. Besides peripheral blood, menstrual blood can be analyzed to investigate the possibility of the occurrence of a sexual assault crime (Chapter 16).

Other bodily fluids are *transcellular fluids*. These fluids are considered to be a subcompartment of the extracellular fluid that is contained within epithelial-lined extracellular spaces. For example, seminal fluid (Chapter 14) and saliva (Chapter 15) stains as well as vaginal secretions are analyzed for the investigation of sexual assault crimes. Sometimes, urine stains and fecal materials (Chapter 16) are related to assault crimes as well. Sweat, which is secreted from the eccrine and apocrine sweat glands in the skin (Chapter 17), and cerumen, also known as earwax (a waxy substance secreted in the ear canal), can be potentially used for human identification. Fluids present in vomitus (Chapter 17) can be potentially important for forensic investigations of violent crimes.

4.1.1 Extracellular Nucleic Acids

Blood plasma and other various bodily fluids usually contain small amounts of nucleic acids (DNA or RNA) known as extracellular nucleic acids. The nucleic acids circulating in plasma

Table 4.1 Common Items of Evidence

Evidence	Possible DNA Location	Source of DNA
Baseball bat	Handle	Skin cells, sweat, blood, tissue
Hat, bandana, mask	Inside surfaces	Sweat, hair, skin cells, dandruff, saliva
Eyeglasses	Nose, ear piece, lens	Sweat, skin cells
Facial tissue, cotton swab	Surface	Mucus, blood, sweat, semen, earwax
Dirty laundry	Surface	Blood, sweat, semen, saliva
Toothpick	Surface	Saliva
Used cigarette	Butt (filter area)	Saliva
Used stamp, envelope seal	Moistened area	Saliva
Tape or ligature	Inside or outside surface	Skin cells, sweat, saliva
Bottle, can, glass	Mouthpiece, rim, outer surface	Saliva, sweat, skin cells
Used condom	Inside surface, outside surface	Semen, vaginal cells, rectal cells
Bed linen	Surface	Sweat, hair, semen, saliva, blood
Through-and-through bullet	Outside surface	Blood, tissue
Bite mark	Skin surface	Saliva
Fingernail, partial fingernail	Scrapings	Blood, sweat, tissue, skin cells

Source: National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002, US Department of Justice, Office of Justice Programs, Washington, DC.

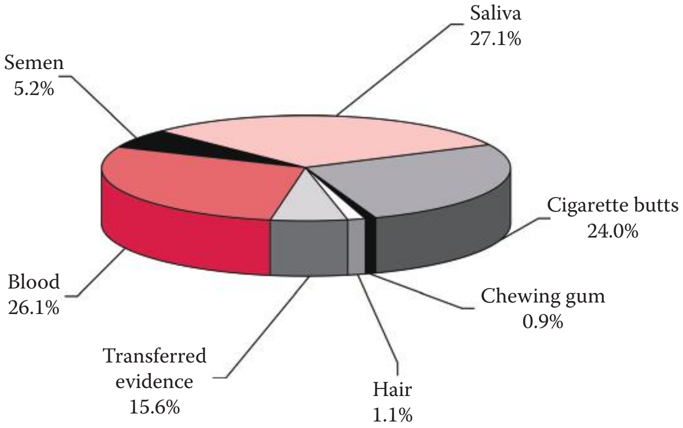

Figure 4.1 Representative types of evidence samples. Data compiled from the third quarter 2005 (July–September) results for all police forces in England and Wales. (Adapted from Bond, J.W., *J Forensic Sci*, 52, 128–136, 2007.)

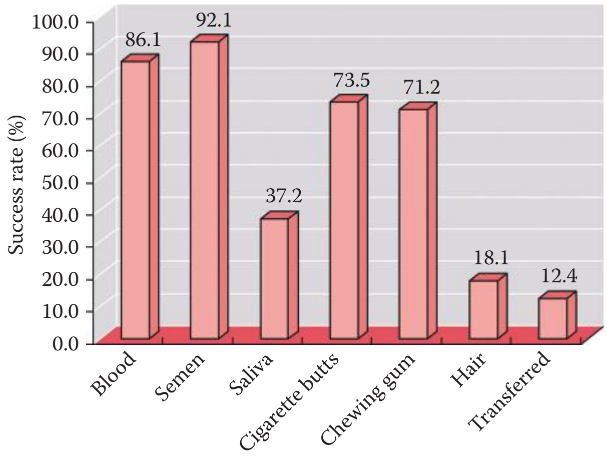

Figure 4.2 Success rate of obtaining suitable profiles from processed samples for submitting to DNA databases. Data are from the third-quarter 2005 (July–September) results for all police forces in England and Wales. (Adapted from Bond, J.W., *J Forensic Sci*, 52, 128–136, 2007.)

Figure 4.3 Biosafety cabinet for the extraction of DNA from biological evidence (left). Blood samples to be processed in a biosafety cabinet for DNA isolation (right). (© Richard C. Li.)

are referred to as *circulating* nucleic acids; the nucleic acids that are found in other body fluids, such as saliva and urine, are referred to as *cell-free* nucleic acids. Extracellular nucleic acids can remain soluble or form complexes with proteins and lipids.

The potential sources of extracellular nucleic acids are *extracellular vesicles* (EVs). EVs are endogenous vesicular structures, containing proteins and nucleic acids that are secreted by most eukaryotic cells (Figure 4.4). There are many types of EVs, including exosomes and microvesicles, which can be detected in bodily fluids. *Exosomes* are one potential source of extracellular nucleic acids. Exosomes are derived from multivesicular bodies (MVBs), which are intracellular organelles of the endocytic pathway (Figure 4.4a). MVBs fuse with the plasma membrane. As a result, the vesicles are released into the extracellular compartment as exosomes. *Microvesicles*, also called shed vesicles or ectosomes, are another possibility. Microvesicles shed from the plasma membrane and thus carry along membrane and cytosolic materials including nucleic acids (Figure 4.4b). *Apoptotic bodies* are a special type of microvesicle that is formed in apoptotic cells. During apoptosis, a process of programmed cell death, cells shrink and eventually form apoptotic bodies (Figure 4.5). Apoptotic bodies contain fragmented DNA by nucleolytic degradation that resembles similar characteristics to those observed in extracellular nucleic acids. However, messenger RNA (mRNA) within apoptotic bodies is protected from RNase degradation. These

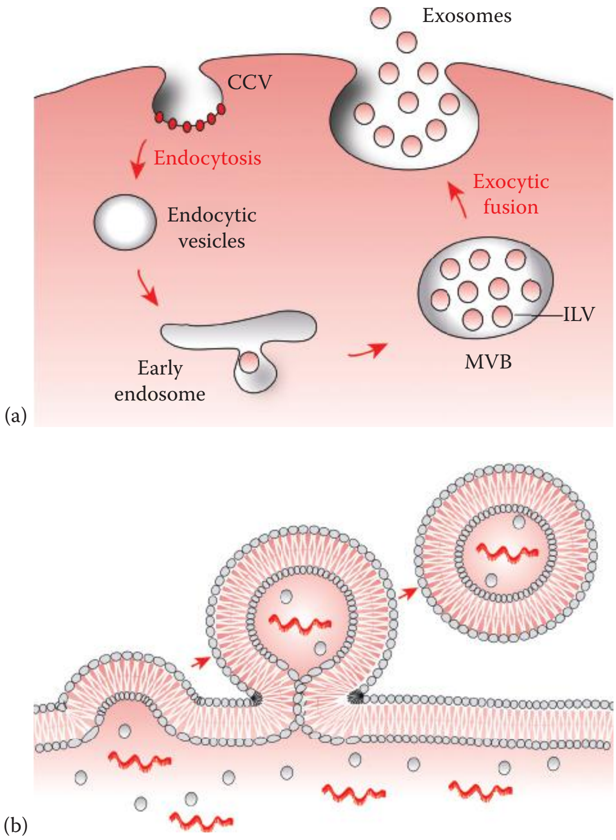

Figure 4.4 Exosomes and microvesicles released from healthy cells. (a) Exosomes are derived from the endocytic pathway. In the endocytic pathway, clathrin-coated vesicles (CCV) are formed at the plasma membrane. The endocytic vesicles are then formed through endocytosis and are transported to early endosomes. Multivesicular bodies (MVBs) containing intraluminal vesicles (ILVs) are then developed from early endosomes. Through the exocytic fusion of the MVB membrane with the plasma membrane of the cell, exosomes, containing cellular components including nucleic acids, are released to the extracellular space. (b) Microvesicles are derived from the budding of vesicles from the plasma membrane. Microvesicles are then shed from the plasma membrane carrying along cellular components including nucleic acids that are present in the cytoplasm. (© Richard C. Li.)

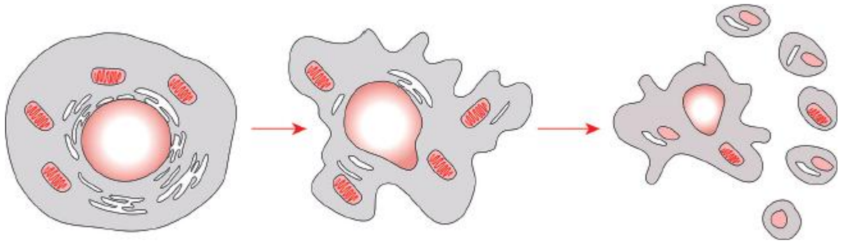

Figure 4.5 Apoptotic bodies released during apoptosis. During apoptosis, the preapoptotic cell (left) undergoes morphological changes. The early apoptotic cell (middle) shows deformation: the cell shrinks, the membrane blebs, and chromatin condense. The late apoptotic cells (right) are fragmented, releasing apoptotic bodies, which contain cytosol and organelles, and nuclear fragments. In addition, proteolytic enzymes are activated that cleave the genomic DNA into fragments. (© Richard C. Li.)

extracellular nucleic acids are potentially useful sources for the forensic identification of bodily fluids and for forensic DNA profiling.

4.2 Cells

4.2.1 Cell Surface Markers

Many of the forensic markers that are analyzed in forensic laboratories are from cells, the building blocks of the human body. All cells have membranes, also known as plasma membranes, which constitute their outer boundaries (Figure 4.6). The functions of cell membranes include exchanges with the environment, signal transduction, and structural support. The cell membrane is a phospholipid bilayer containing lipids, proteins, and carbohydrates. Membrane proteins can act as enzymes, receptors, or ion channels. Many cells also have carbohydrate-rich molecules, including proteoglycans, glycoproteins, and glycolipids, on their membrane surfaces (Figure 4.6). Many of these molecules act as cell surface antigens (Chapter 18).

4.2.2 Nucleated Cells

Probative biological evidence usually contains nucleated cells. The *nucleus* (Figure 4.7) is surrounded by a nuclear envelope and contains chromosomes and a nucleolus. A nucleolus is a dense, non-membrane-bound structure due to its high RNA content. The function of a nucleolus is to transcribe ribosomal RNA and to form ribosomes.

There are two types of cells in the body: sex cells (sperm and oocytes) and somatic cells (all other types). Spermatozoa and ova, which are formed by germ cells, are called *gametes*. In humans, each gamete is haploid, containing 22 *autosomes* (chromosomes other than sex chromosomes) plus one *sex chromosome*. In ova, the sex chromosome is always an X, while in spermatozoa it may be an X or a Y. After fertilization, the zygote, which is the fertilized egg cell formed when two gamete cells are joined, becomes diploid as a result of the fusion of the haploid spermatozoon and ovum. Most of the other cells of the body, known as *somatic cells*, are diploid. This means that they have two copies of each autosome plus two sex chromosomes, XX for females or XY for males. This results in a total of 46 chromosomes per diploid cell. The two chromosomes of a pair in a diploid cell are *homologous chromosomes*. One of the homologous chromosomes is inherited from the spermatozoon and the other from the ovum.

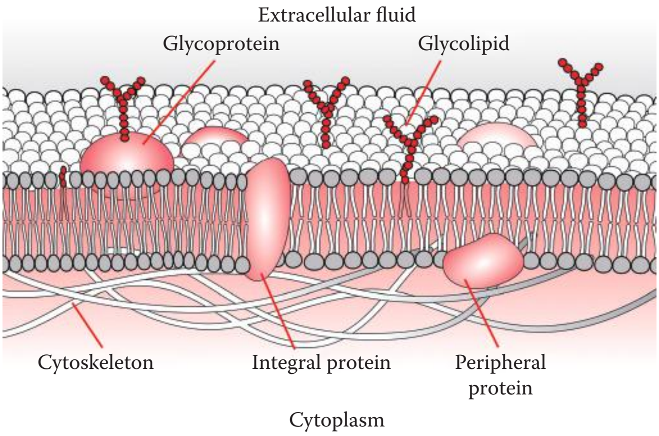

Figure 4.6 Cell membrane and carbohydrate-containing glycoproteins and glycolipids. (© Richard C. Li.)

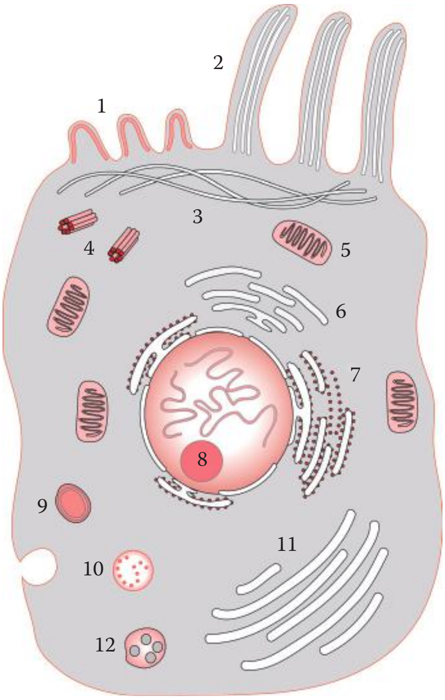

Figure 4.7 A cross-sectional view of a cell. 1, Microvilli; 2, cilia; 3, cytoskeleton; 4, centrioles; 5, mitochondrium; 6, smooth endoplasmic reticulum; 7, rough endoplasmic reticulum; 8, nucleolus; 9, peroxisome; 10, vesicle; 11, Golgi apparatus; and 12, lysosome. (© Richard C. Li.)

Although most somatic cells are diploid, exceptions exist. Some differentiated cells such as red blood cells and platelets have no nuclei and are designated *nulliploid*. A few other cells have more than two sets of chromosomes as a result of DNA replication without cell division and are referred to as *polyploid*. For example, the regenerating cells of the liver and other tissues are naturally tetraploid, while the giant megakaryocytes of the bone marrow may contain 8, 16, or even 32 copies of chromosomes.

The nuclear *chromosomes* of humans consist of complexes of DNA, histone proteins, and nonhistone chromosomal proteins. Each chromosome consists of one linear, double-stranded DNA molecule. The large amounts of DNA present in the human chromosome are compacted by their association with histones into nucleosomes and even further compacted by higher levels of folding of the nucleosomes into *chromatin* fibers. Each chromosome contains a large number of looped domains of chromatin fibers attached to a protein scaffold.

The degree of DNA packing varies throughout the cell cycle. During the metaphase of mitosis and meiosis of the cell cycle, chromatin is the most condensed. Two forms of chromatin have been defined on the basis of their chromosome-staining properties. *Euchromatin* regions are areas of chromosomes that undergo normal chromosome condensation and decondensation during the cell cycle. The intensity of staining of euchromatin is darkest in the metaphase and lightest in the synthesis (S) phase. Euchromatic regions account for most of the genome and lack repetitive DNA. Usually, the genes within the euchromatin can be expressed. *Heterochromatin* comprises the chromosomal regions that usually remain condensed throughout the cell cycle. It contains repetitive DNA and can be found at centromeres, much of the Y chromosome long arm, and the short arms of the acrocentric chromosomes (chromosomes with centromeres near one end). Genes within heterochromatic DNA are usually inactive.

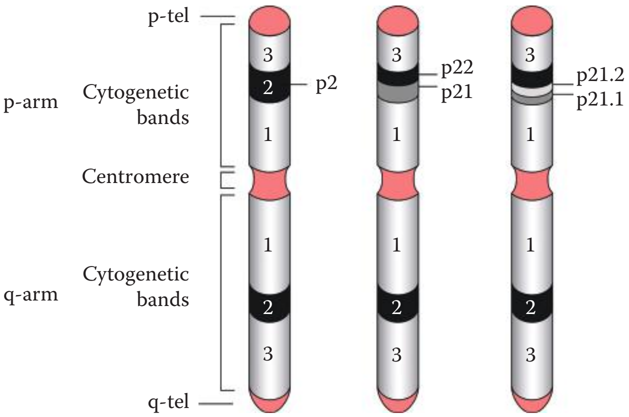

Figure 4.8 The ideogram of a human chromosome and cytogenetic banding nomenclature. The cytogenetic banding pattern is shown after chemical staining such as G-banding, which treats chromosomes with Giemsa dye. The short arm is designated p; the long arm is q. The centromere, p telomere (p-tel), and q telomere (q-tel) are also shown. G-bands at increasing resolution (from left to right) are shown. (© Richard C. Li.)

Each human chromosome has a short arm, designated p (for petit), and a long arm, designated q (for queue), separated by a centromere (Figure 4.8). *Centromeres* are the DNA sequences that are found near the points of attachment of mitotic or meiotic spindle fibers. The centromere region of each chromosome is responsible for accurately segregating the replicated chromosomes to daughter cells during cell divisions. The ends of the chromosome are called *telomeres* and they help stabilize the chromosome and play a role in the replication of DNA in the chromosome.

Chemical staining of metaphase chromosomes results in an alternating dark and light banding pattern (cytogenetic banding) that can be seen under a microscope. Each chromosome arm is divided into regions based on the cytogenic bands. This process is known as *cytogenetic mapping*. The cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, and so on, counting from the centromere out toward the telomeres. At higher resolutions, subbands can be observed. For example, the cytogenetic map location of a gene termed *AMELY* (amelogenin, Y-linked; Chapter 21) is Yp11.2, which indicates its location on chromosome Y, p arm, band 11, subband 2. The visually distinct banding pattern gives each chromosome a unique appearance. Recently, the cytogenic map has been integrated with the human genome sequence to allow the determination of the positions of cytogenetic bands within the DNA sequence.

Chromosomes can be identified on the basis of the size and the positions of the centromeres and cytogenetic banding patterns. The chromosome constitution is described as a *karyotype* and can be displayed as a *karyogram*, which includes the total number of chromosomes and the sex chromosome composition (Figure 4.9). Chromosomes are numbered in order of their size, with chromosome 1 as the largest (except chromosome 21 is smaller than 22). In the cases of chromosomal abnormality, the karyotype can also reflect the type of abnormality and allow visualization of the affected chromosome bands.

4.2.3 Mitochondria and Other Organelles

The cytoplasm contains the cytosol fluid in which organelles are suspended (Figure 4.7). Multiple copies of *mitochondria* are located within the cytoplasm. Mitochondria are surrounded by phospholipid membranes that separate them from the cytosol. The mitochondria are responsible for energy production through aerobic metabolism by producing molecules containing high-energy bonds, such as adenosine triphosphate (ATP). Mitochondria have their own genome, which can be analyzed for human identification (Chapter 23).

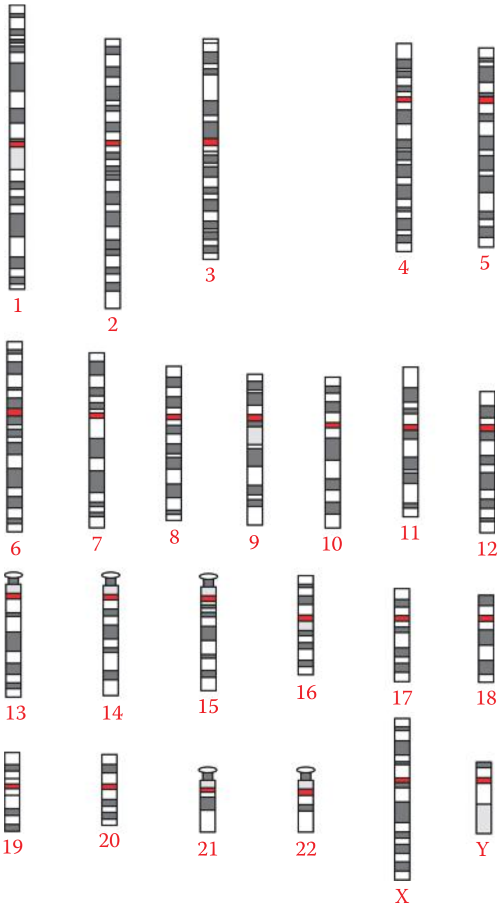

Figure 4.9 Human karyogram. The chromosomes are numbered. Cytogenic patterns show alternating dark and light bands (From www.genome.gov). Centromeres (in red) are shown.

Two types of *endoplasmic reticulum* (ER) can exist within the cytoplasm: the smooth ER (SER) is involved in lipid synthesis; the rough ER (RER) contains ribosomes on its outer surface and forms transport vesicles. The *Golgi apparatus* is responsible for the production of secretory vesicles and new membrane components, and also for the packaging of *lysosomes* (vesicles containing digestive enzymes for the degradation of injured cells). *Peroxisomes* carry enzymes that neutralize potentially harmful free radicals. Other organelles found within the cytoplasm of eukaryotic cells include the cytoskeleton, microvilli, centrioles, and cilia.

4.2.4 Cytosol

4.2.4.1 Messenger RNAs

The chromosomal DNA contains genes that encode for specific proteins. The genetic code is read as an array of triplet codes, a sequence of three bases that specifies the identity of a single amino acid. As gene expression is activated, transcription occurs in which precursor mRNA (pre-mRNA) is produced from a DNA template. After transcription, the pre-mRNA is capped, polyadenylated, and spliced to form matured mRNA. Only the matured mRNA is transported from the nucleus to the cytoplasm. Tissue-specific mRNA can be potentially used for the identification of biological evidence (Chapter 11). The proteins are synthesized in a process known as translation, in which amino acids are assembled based on the codons derived from the triplet code of the DNA contained in the sequence of the mRNA strand. Various components including the ribosomal complex are involved in translation. The cytosol also contains many proteins that can be used for the identification of bodily fluids (Chapters 11 and 12).

4.2.4.2 MicroRNAs

MicroRNAs (miRNAs) are short RNA molecules that are 21–23 nucleotides in length. In eukaryotic organisms, miRNAs function as negative regulators of gene expression. They play roles in development and cell differentiation. Additionally, the altered expression of miRNAs can be detected in many human diseases. The biological function and the potential application of miRNAs in forensic biology are discussed in Chapter 11.

It is estimated that the human genome may encode approximately 1000 miRNAs. Based on their genomic location and structure, miRNAs can be characterized into three types: intergenic, intronic, and exonic (Figure 4.10). *Intergenic* miRNA genes are distinct transcriptional units that are found in genomic regions. In humans, most of the intergenic miRNA genes have a transcription start site and a polyadenylation site. *Intronic* miRNAs reside within the introns of protein-coding and noncoding genes. The orientation of the intronic miRNAs can be the same as that of the sense (coding) strand of a host gene or that of the antisense strand (complementary to the sense strand) of a host gene. Sense intronic miRNAs are transcribed from the same promoter as their host genes. The antisense intronic miRNAs are transcribed from their own promoters. *Exonic* miRNAs are rare in eukaryotic genomes and reside in genomic regions that overlap with an exon and an intron of a pseudogene, which is a noncoding gene or is no longer transcribed. These miRNAs are also transcribed from their host gene promoter. miRNAs can also be monocistronic or polycistronic. A *monocistronic* miRNA has a single transcriptional unit with its own promoter. In *polycistronic* miRNAs, several miRNAs reside as a cluster of transcriptional units with a shared promoter.

The biogenesis of miRNAs begins in the nucleus. miRNAs are transcribed by RNA polymerase II. Nascent transcripts, referred to as primary transcripts (pri-miRNAs), can be several hundreds to thousands of nucleotides in length. The pri-miRNAs have a hairpin secondary

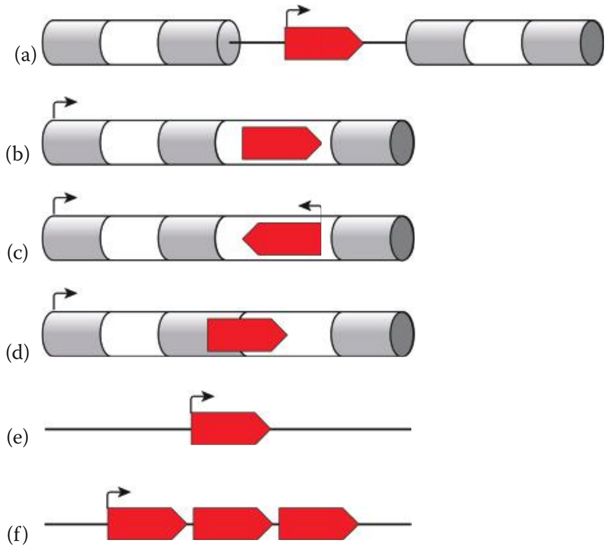

Figure 4.10 The genomic location and the structure of miRNAs. (a) An intergenic miRNA gene, (b) a sense intronic miRNA gene, (c) an antisense intronic miRNA gene, (d) an exonic miRNA gene, (e) a monocistronic miRNA gene, and (f) polycistronic miRNA genes. Exons (gray), introns (white), and miRNA genes (red) are shown. (© Richard C. Li.)

structure that is approximately 70 nucleotides long with imperfect base pairing in the stem. miRNA processing is initiated in the nucleus (Figure 4.11). The hairpin region of a pri-miRNA is then cleaved from the pri-miRNA by the nuclear RNase III endonuclease, Drosha. As a result, a pre-miRNA is formed. The pre-miRNA is transported from the nucleus to the cytoplasm by a nuclear transporter, Exportin 5. In the cytoplasm, a cytoplasmic RNAse III–like endonuclease, Dicer, cleaves the pre-miRNA to generate a double-stranded miRNA that is approximately 21–23 nucleotides in length. The mature miRNA strand is bound to the Argonaut protein to assemble the RNA-induced silencing complex (RISC). The complementary RNA strand is degraded.

A specific miRNA is designated with a prefix. The prefix for a mature miRNA is a capitalized "miR," while the prefix for a pre-miRNA is an uncapitalized "mir." The prefix "miR" is followed by a dash and a number (e.g., miR-135). Experimentally confirmed miRNAs are sequentially numbered and are deposited in the miRBase, which is a database that archives miRNA sequences and annotations.

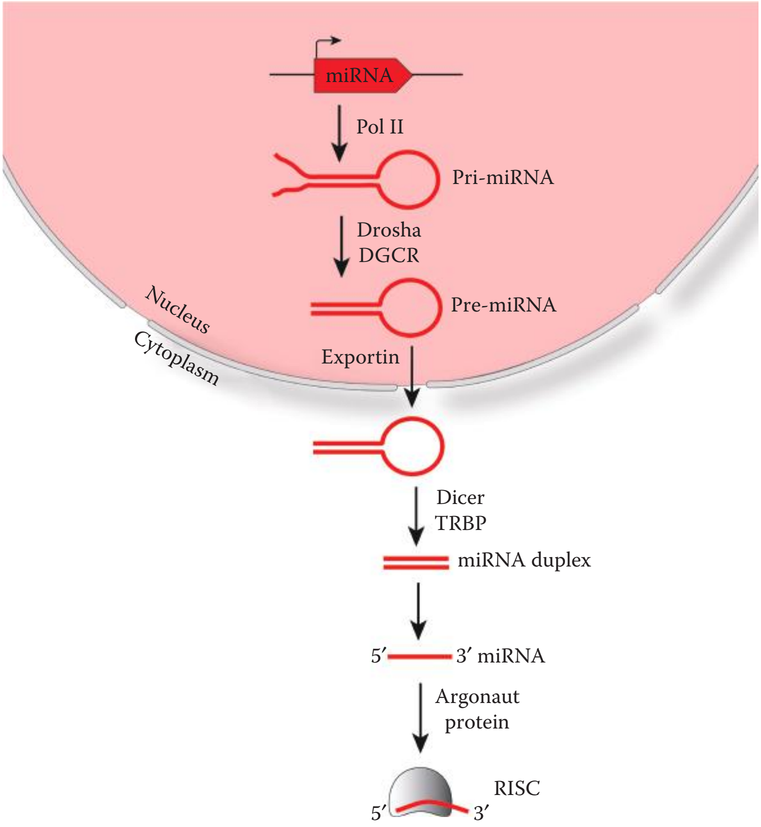

Figure 4.11 The biogenesis of miRNAs. In the nucleus, the pri-miRNA (primary transcript) is processed by Drosha (nuclear RNase for double-stranded RNA) and DGCR (nuclear double-stranded RNA-binding protein), forming the pre-miRNA. The pre-miRNA is transported from the nucleus to the cytoplasm by the exportin. In the cytoplasm, the pre-miRNA is further processed by Dicer (cytoplasmic RNase for double-stranded RNA) and TRBP (cytoplasmic double-stranded RNA-binding protein). The mature miRNA strand is bound to the Argonaut protein to assemble the RISC (RNA-induced silencing complex). (© Richard C. Li.)

4.3 Tissues

4.3.1 Skin

4.3.1.1 Biology of Skin

The biological evidence related to skin is important in forensic investigations. For example, fingerprints are ridge skin impressions that are usually collected at crime scenes, and shed skin tissue is a source of DNA for human identification. The skin covers the entire body surface. Additionally, the skin contains specialized structures that include sebaceous and sweat glands, hair follicles, and nails. The thickness of skin varies throughout the body. The skin of the dorsal area of the body is usually thicker than that of the ventral area of the body. The skin consists of different layers (Figure 4.12). The *epidermis* is the outer layer of the skin. The epidermis also contains melanocytes that produce the skin pigment melanin. The *dermis* is the middle layer of the skin. It is filled with fibrous collagen proteins secreted by fibroblasts and contains hair follicles and sweat glands. Additionally, it contains blood, lymph vessels, and nerves. The *subcutaneous layer* is the deepest layer of the skin. The subcutaneous layer consists of collagen networks and adipose tissue to prevent loss of heat.

The epidermis is a multilayered tissue that includes a number of morphologically distinct zones: the basal, the spinous, the granular, and the cornified layers (Figure 4.12). The epidermis renews continually through the proliferation and differentiation of keratinocytes. The *basal layer* contains newly formed keratinocytes that are proliferative. Epidermal differentiation begins with the migration of the keratinocytes from the basal layer toward the outer layer of skin. Once the migrating keratinocytes reach the *spinous* and *granular* layers, the keratinocytes become nonproliferating and partially differentiated. As the cells reach the *cornified* layer, these cells are filled with keratin filaments and are differentiated into corneocytes, which are dead, and terminally differentiated keratinocytes. In the course of differentiation, the cells are flattened, and all organelles including the nucleus are lost. The corneocytes are then shed from the skin surface.

4.3.1.2 Skin as Source of DNA Evidence

Evidence from skin contact, also referred to as *touched evidence*, can be collected and used for forensic DNA analysis. One example of this evidence is shed skin cells that are found on worn clothing, which is frequently encountered in crime scene investigations. For instance, a perpetrator's shed skin cells that are deposited on worn clothing are potential evidence to be collected. After collection, a DNA profile can be obtained from shed skin cells, providing the forensic evidence or "lead" that is required for the criminal investigation. Touched evidence becomes more

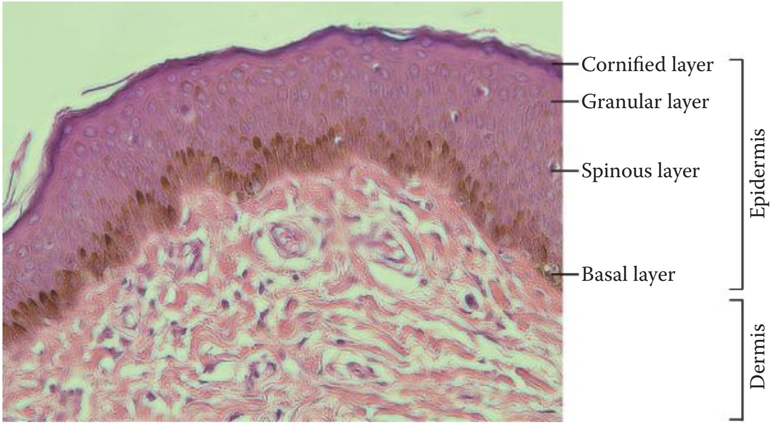

Figure 4.12 A sectional view of the skin. Multilayered epidermis tissues are shown. (© Richard C. Li.)

useful and important when no other type of evidence (e.g., a fingerprint, bloodstain, or seminal stain) can be collected at the crime scene, especially for the investigation of certain criminal cases such as cold cases and property crimes such as theft and burglary (Figures 4.13 through 4.15). Touched evidence usually contains minute quantities of nuclear DNA. DNA recovered from touched evidence is referred to as *transfer DNA*. However, the source of DNA transferred through physical contact is not well understood. Although the shed skin cells from touched evidence lack nuclei, these cells may contain DNA remnants from partial degradation during the differentiation process. These can be a possible DNA source. Additionally, the sweat glands in the skin produce sweat. When the skin makes contact with an item, a residue of sweat is left on the surface of the item. It is known that sweat contains cell-free DNA. Thus, cell-free DNA from sweat is another possible source of DNA. Furthermore, a small number of nucleated cells can be observed in touched evidence. These cells possibly originated from sweat glands and ducts.

4.3.2 Hair

Hairs, including scalp and pubic hairs, frequently constitute biological evidence that is found at crime scenes, and their identification can be of great forensic importance. Formerly, the

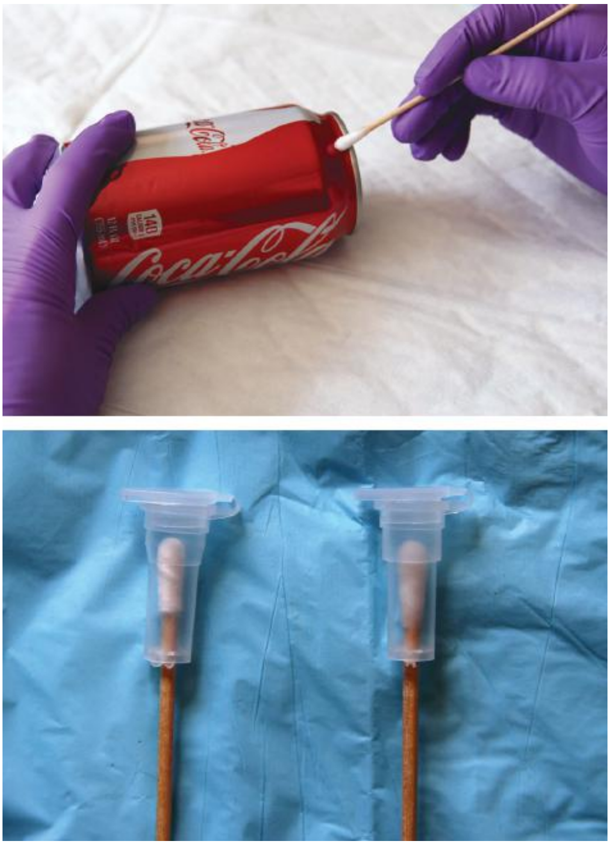

Figure 4.13 Evidence collection using swabbing. The evidence from skin contact for forensic DNA analysis is usually collected by swabbing. More DNA can be recovered when evidence is collected with a double swab method than with a single swab. The double swab method involves applying a moistened cotton swab followed by a second dry cotton swab onto the same target surface of evidence. The target surface is swabbed using a moistened swab first. The moisture left by the first swab is absorbed by the second dry swab. Both swabs can be pooled for DNA extraction. (© Richard C. Li.)

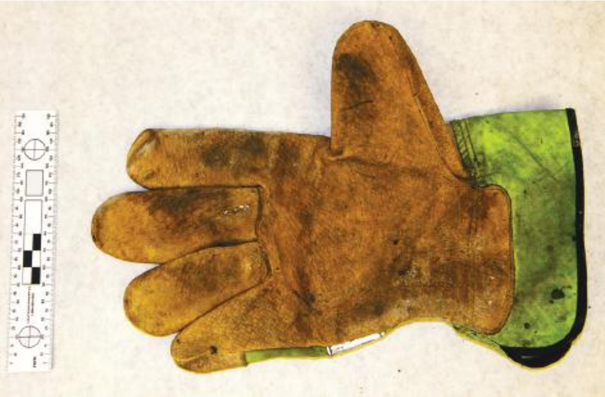

Figure 4.14 A worn glove as a source of DNA evidence. Shed skin cells on a worn glove can potentially generate DNA profiles. (© Richard C. Li.)

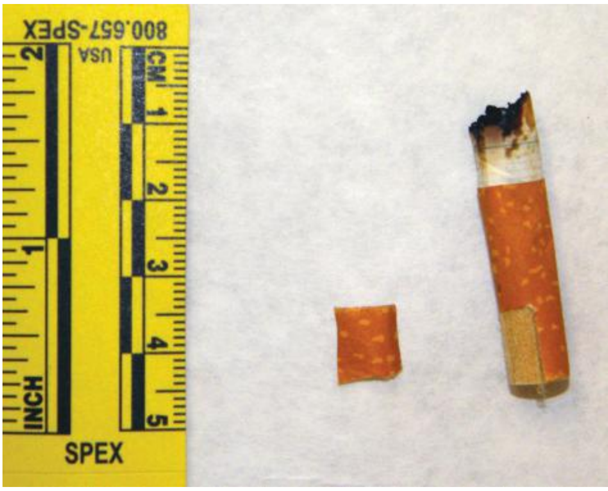

Figure 4.15 A cigarette butt as a source of DNA evidence. Shed cells left behind on a cigarette butt can be a source of DNA. A portion (~1 cm^2) of the filter paper of a smoked cigarette butt can be cut for isolating DNA. (© Richard C. Li.)

principal methods that were utilized in forensic hair analysis were limited to morphological analysis and comparisons. Since then, protein polymorphisms provide some potential for identifying individuals from single hairs. However, human hairs contain DNA and as a DNA source they may be used for forensic analysis. The development of the polymerase chain reaction (PCR) amplification technique made it possible to analyze very small quantities of DNA in hair, and the use of hair as evidence of identification has become more significant.

4.3.2.1 Biology of Hair

The human hair shaft is a keratinized cylindrical structure (Figure 4.16). The center or core of the hair is called the *medulla*, which is present in the majority of hairs. The medulla is surrounded by a *cortex*, which is the outer layer of the hair shaft. The *cuticle* consists of overlapping layers of flattened keratinized cells that protect the hair. Hairs are produced in *hair follicles* (Figures 4.17 and 4.18). Each hair follicle is located deep in the dermis (a skin layer beneath the epidermis) and opens onto the surface of the epidermis (the outer layer of the skin). The hair follicle is composed

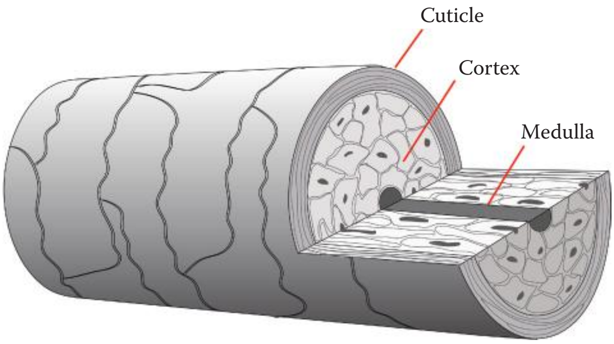

Figure 4.16 A sectional view of a hair shaft. (© Richard C. Li.)

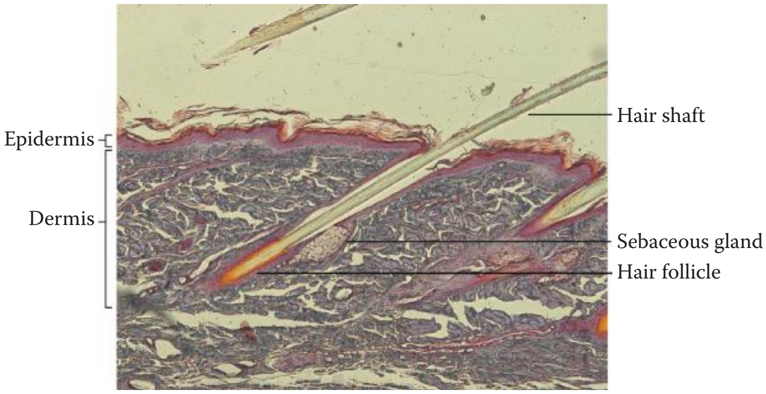

Figure 4.17 Longitudinal section view of a scalp hair follicle with accessory structures. (© Richard C. Li.)

of the *bulge* and the *bulb* regions. The bulge exports the stem cells that migrate down and give rise to bulb cells. The bulge also produces the stem cells that migrate up to form skin cells. The matrix cells, which generate the hair shaft cells, are located at the lower portion of the bulb. The *dermal papilla* is situated at the base of the bulb and contains cells that regulate hair growth, which are nourished by blood vessels and nerves. The growing hair shaft is surrounded by two concentric layers of cells, which are referred to as the inner root sheath and the outer root sheath, respectively. The entire hair follicle is surrounded by a connective tissue sheath.

Human scalp hairs grow for a few years and shed according to the hair follicle cycle (Figures 4.19 through 4.21). A human scalp hair can grow at its highest rate of approximately 1 mm per day. The growing phase of hair is called the *anagen* phase. The matrix cells undergo rapid proliferation and eventually become differentiated cells such as hair shaft cells. As a hair grows, it is pushed toward the surface of the skin and becomes longer. During the migration upward, keratinization occurs as cells are filled with fibril keratin proteins. In these keratinizing cells, nuclei are absent. However, mitochondrial remnants can be observed. By the time a hair approaches the skin surface, cell death occurs at the medulla, cortex, and cuticle. At the end of the anagen phase, the matrix cells enter the *catagen* phase and undergo cell death, thus leading to the regression of the bulb. Hair follicles then enter the *telogen* phase: the stage of rest. When another cycle begins, the follicle produces a new bulb and the telogenic hair, also known as the club hair, is pushed to the surface and shed. On average, an adult loses approximately dozens of hairs daily.

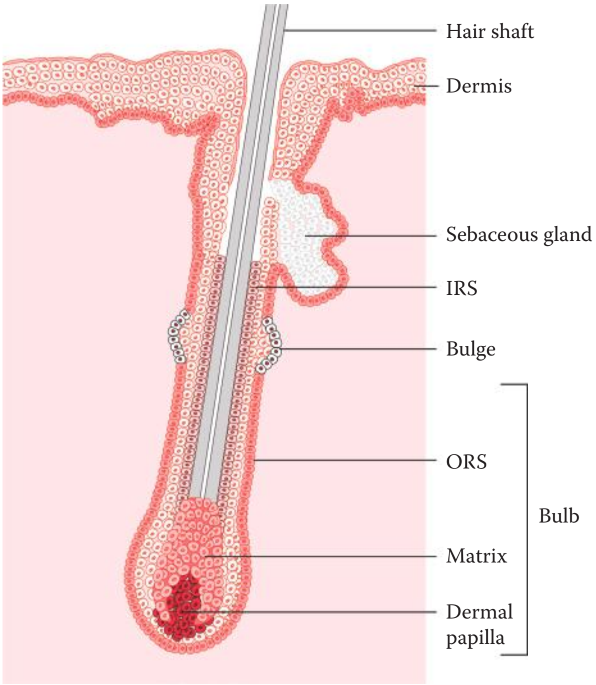

Figure 4.18 Structure of an anagen-phase hair follicle. The diagram shows the hair follicle structure including the dermal papilla, sebaceous gland, bulge, bulb, and hair shaft. Concentric layers of the outer root sheath (ORS), the inner root sheath (IRS), and the hair shaft are also shown. (© Richard C. Li)

4.3.2.2 Hair as Source of DNA Evidence

The isolation of DNA from intact hair roots is routinely used in nuclear DNA analysis. In comparison, the quantities of DNA in telogen hair roots are considerably less than the DNA that is found in the roots of anagen hairs. Nuclear DNA analysis is usually accomplished by using freshly plucked hair roots (Figures 4.22 and 4.23) because cells at the root region may contain nuclear DNA. Unfortunately, most human hairs recovered from crime scenes are shed naturally (in the telogen phase) and contain little nuclear DNA (Figure 4.24). Thus, multiple telogen hairs with roots are necessary to isolate enough nuclear DNA. However, shed hairs that are found at crime scenes may be derived from different individuals. Therefore, the ability to perform a forensic DNA analysis of a single shed hair would be highly desirable. Nuclear DNA isolation from hair shafts is still far less reliable because hair shafts contain very low amounts of nuclear DNA. In addition, variations in the amounts of DNA isolated from hair shafts are observed in a comparison between different hairs from the same head and hairs from different individuals.

A hair follicle cell contains multiple copies of mitochondria. As a result, mitochondrial DNA (mtDNA) can be successfully isolated from hair roots. Additionally, mtDNA is embedded in the keratin matrix of hair shaft cells, which protects the mtDNA molecules from degradation. Thus, mtDNA can also be isolated from hair shafts. A sequence polymorphism analysis of mtDNA from hair can be carried out. mtDNA is maternally inherited, which is useful to identify maternal relatives but cannot be used to perform paternity testing. Additionally, the mtDNA profiling results are not as discriminating as nuclear DNA profiling. Furthermore, mtDNA analysis is time-consuming. Therefore, the typing of nuclear DNA from hair would be preferable for forensic DNA analysis.

Sometimes, a mixture of more than one mtDNA sequence in the same individual is observed. This heterogeneous pool of mtDNA molecules is referred to as *heteroplasmy*. In hair, it is

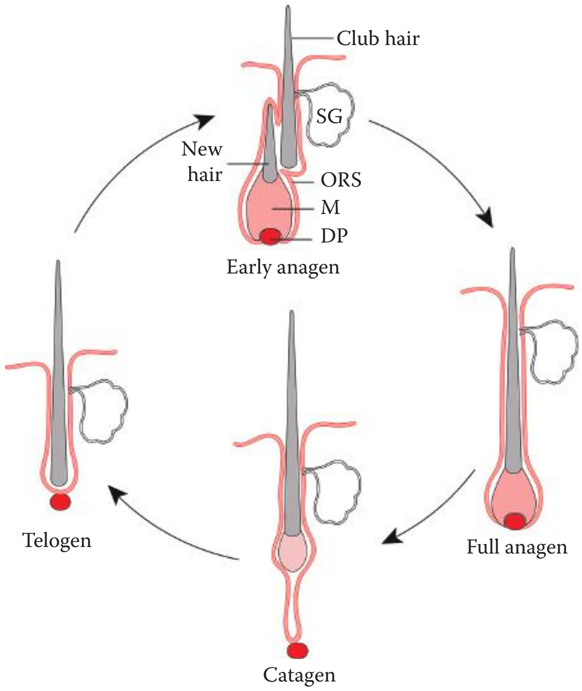

Figure 4.19 Diagrammatic representation of the scalp hair cycle. The morphological characteristics of a hair follicle are shown in three distinct phases encompassing the entire hair cycle: anagen (including early and full anagen), catagen, and telogen. DP, dermal papilla; M, matrix; ORS, outer root sheath; SG, sebaceous gland. (© Richard C. Li.)

believed that it is due to a mixture of mtDNA molecules from keratinocyte and melanocyte-derived mitochondria (Figure 4.25). Hair follicle melanocytes are formed at the beginning of each hair cycle and die at the end of the cycle. Melanocytes are located in the bulb hair follicle. Hair melanocytes play roles in hair pigmentation, which determine hair color. Melanin, which is produced by the melanocytes, is contained in an organelle called the *melanosome*. The melanosomes are transferred to neighboring keratinocytes though *dendritic processes*. In addition to melanosomes, the melanocyte mitochondria can also be transferred to keratinocytes. Thus, the keratinized cells in the hair shaft may carry more than one type of mitochondria, one from the keratinocytes and the other from the melanocytes. As a result, heteroplasmy of a mixture of different mtDNA molecules, with different DNA sequences, can occur in hairs.

4.3.3 Bone

4.3.3.1 Biology of Bone

The bodies of human remains begin to decompose shortly after death. The rate of decomposition of human remains varies greatly with environmental conditions such as climate, bacterial growth, and the presence of insects and other animal scavengers. However, soft tissues may be lost first while more stable bone tissues may remain. Identifying human skeletal remains can be applied in a variety of cases including mass fatality incidents, missing persons, fires, explosions, and violent crime cases involving skeletal remains.

An adult human skeleton consists of 206 bones (Figures 4.26 and 4.27). The shaft of a long bone, such as an arm or a leg bone, consists largely of an outer layer of *cortical* (or compact) bone, which is solid and strong. The shaft of a long bone forms a *marrow cavity*, which is filled with a specialized type of connective tissue called bone marrow. The portion at each end of a

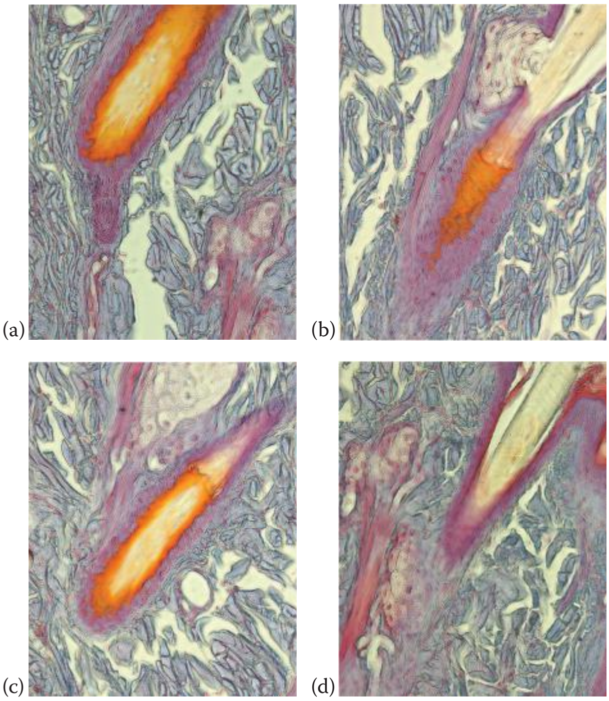

Figure 4.20 Longitudinal section view of a scalp hair follicle during the hair cycle. (a) Early anagen, (b) full anagen, (c) catagen, and (d) telogen. (© Richard C. Li.)

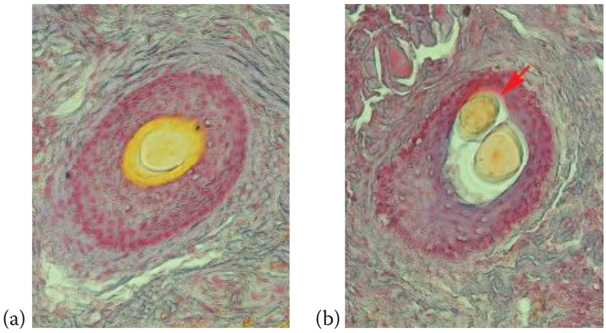

Figure 4.21 Cross-sectional view of hair follicles. (a) Early anagen and (b) full anagen. Note that the new hair (arrow) and the club hair are shown. (© Richard C. Li.)

long bone is called the *epiphysis*, which is composed largely of *cancellous* (or spongy) bone, and can bear the force of compression. A flat bone can have primarily either cortical or cancellous bone. For instance, a rib consists of primarily cancellous bone surrounded by a thin layer of cortical bone. A skull bone usually consists of largely cortical bone.

Bone, which is a connective tissue, contains a matrix and cells. The bone matrix consists of an inorganic and an organic matrix. Calcium and phosphorus are the major components of the inorganic matrix, which consists mainly of hydroxyapatite crystals, $Ca_{10}(PO_4)_6(OH)_2$. The organic matrix consists of collagens, primarily type I collagen, which are insoluble fibrous proteins. With the deposition of calcium hydroxyapatite crystals around the collagen fibrils, bone becomes a weight-bearing hard tissue.

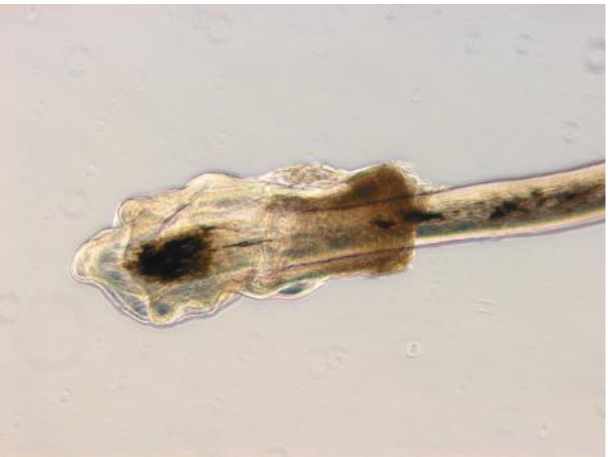

Figure 4.22 Hair root of a pulled hair with visible soft tissue attached. (© Richard C. Li.)

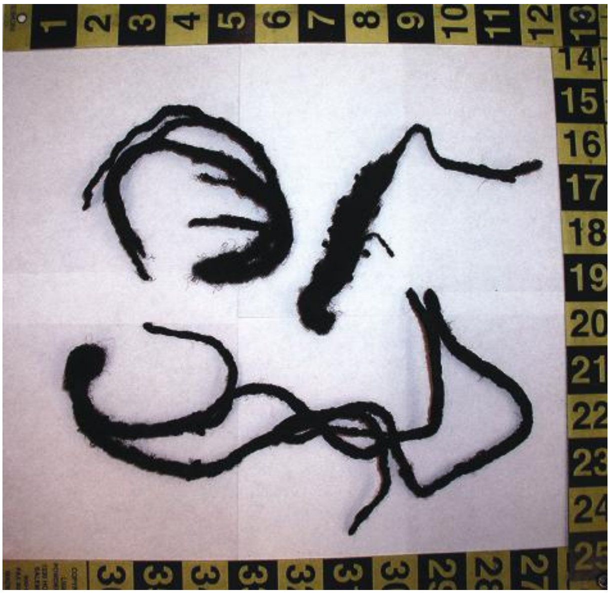

Figure 4.23 Pulled dreadlocks recovered from a crime scene. (© Richard C. Li.)

Developing bones contain small numbers of *osteoprogenitor* cells. These cells can divide to produce cells that differentiate into *osteoblasts*. Osteoblast cells regulate the calcification of the bone matrix. Osteoblasts that are embedded in the bone matrix are termed *osteocytes* and are the most abundant cells in bone. Osteocytes play a role in maintaining the surrounding matrix and repairing damaged bone. Another type of cell that can be found in bone tissues is *osteoclasts*. These cells are giant cells containing 50 or more nuclei and are responsible for dissolving and recycling the bone matrix.

4.3.3.2 Bone as Source of DNA Evidence

A number of methods are used to identify human remains, for example: the identification of facial characteristics; the recognition of individualizing scars, marks, and other special body

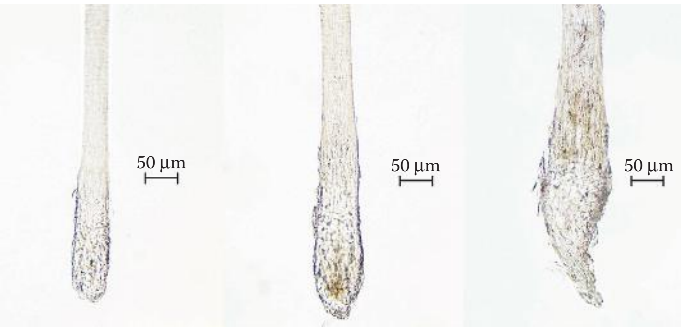

Figure 4.24 Telogen hair roots. From left to right: telogen club root with no visible soft tissue; telogen hair root with some visible soft tissue; telogen hair root with visible soft tissue. (From Bourguignon, L. et al., *Forensic Sci Int Genet*, 3, 27–31, 2008. With permission.)

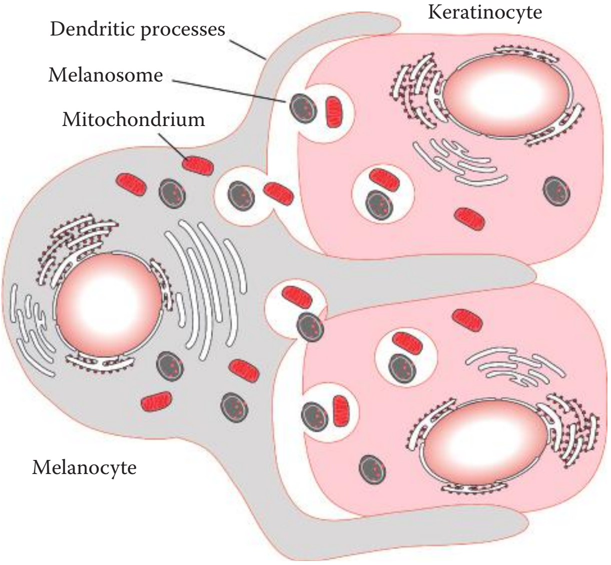

Figure 4.25 Scheme diagram of the melanosome transportation. Melanosomes are released into the extracellular space from the melanocyte dendrites through exocytosis and subsequent endocytosis by keratinocytes. During the process, mitochondria originating from the melanocytes can be potentially transported to the keratinocytes. (© Richard C. Li.)

features; the matching of dentition with premortem dental x-rays; and the comparison of fingerprints. In some situations, these methods cannot be used because of the extensive decomposition of the remains. The mass fatality terrorist attack on the World Trade Center in 2001 (Figure 4.28) serves as an example of a situation where common identification techniques may not be useful. Large quantities of compromised human skeletal fragments were recovered at the fatality site. In these cases, DNA typing is a powerful tool for identifying human remains.

Most DNA in cortical bone is located in the osteocytes. It has been estimated that there are approximately 20,000 osteocytes per cubic millimeter of calcified bone matrix. As a result, microgram quantities of DNA can potentially be obtained from a gram of bone. Thus, compact bone tissue should contain sufficient amounts of nuclear DNA for analysis. However, the skeletal fragments recovered from burial sites are often subjected to decomposition (Figures 4.29 and 4.30). During the decomposition process, both nuclear and mtDNA can be degraded.

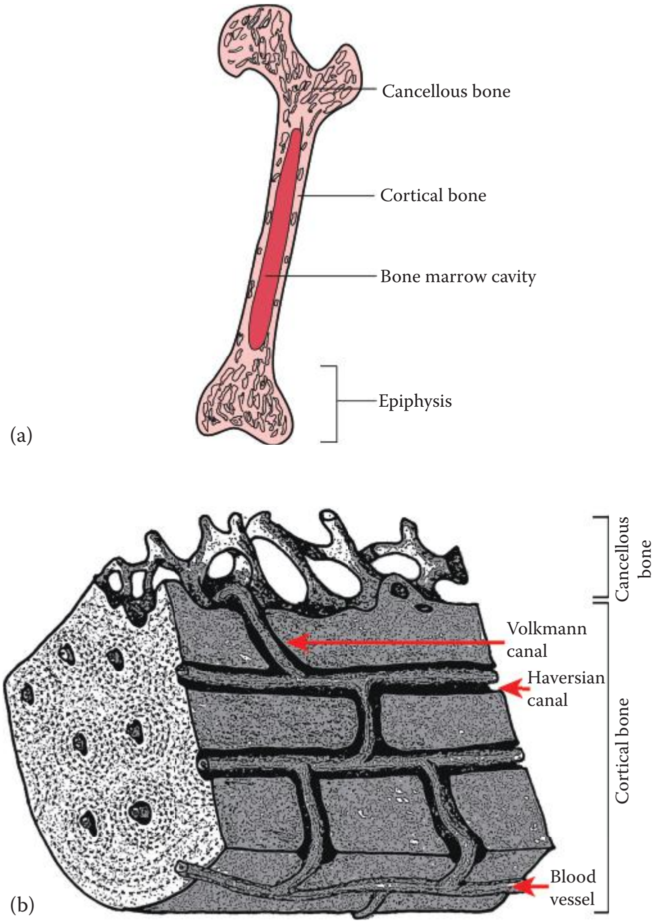

Figure 4.26 Bone structure. An adult human consists of 206 bones. A long bone, such as an arm or a leg bone, consists of an outer cylinder of cortical bone surrounding a marrow cavity. Each end of a long bone is called the epiphysis, which is composed largely of cancellous bone. Flat bones have variable structures; for example, the skull consists mainly of cortical bone whereas the spine consists mainly of cancellous bone. (a) Diagram of a long bone. (b) A bone fragment. Cortical and cancellous bones are shown. Blood vessels can be found in the Volkmann and Haversian canals. (© Richard C. Li.)

Additionally, burial conditions with high humidity and temperature promote the degradation of DNA. Thus, the identification of partial DNA profiles or the complete failure to obtain DNA profiles can occur after samples from decomposed remains are analyzed.

Processing bone samples for DNA extraction is a time-consuming task (Figure 4.31). Due to the potential for commingled remains and contaminants that interfere with forensic analysis, a bone sample initially must be cleaned prior to isolating DNA. The outer surfaces of bone fragments are usually removed by using a mechanical method such as sanding. However, to avoid cross-contamination of samples, the bone dust that is generated by sanding must be removed. Additionally, special protective equipment and safety procedures are necessary to protect analysts from exposure to blood-borne pathogens.

To obtain adequate quality and quantity of DNA from a bone sample, a high-yield DNA extraction method should be selected. The bone samples can be ground to powder to aid in DNA extraction (Chapter 5). The osteocytes containing DNA are embedded in a calcified bone matrix, which is a barrier for extracting DNA from the osteocytes. The bone matrix must be

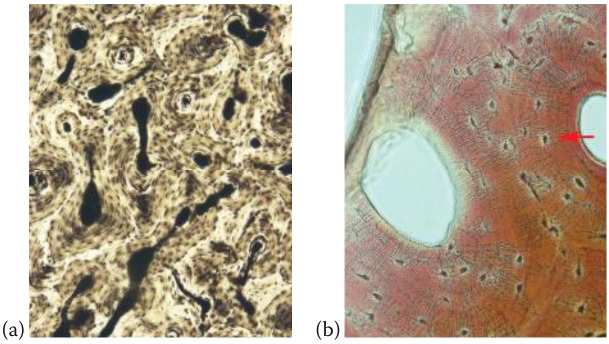

Figure 4.27 Cross-sectional view of cortical bone. (a) The functional unit of cortical bones is a cylindrical structure known as osteon. Haversian canals are shown in the center of the osteon. (b) Detailed view of an osteon. Osteocytes (arrows) are shown within osteons. (© Richard C. Li.)

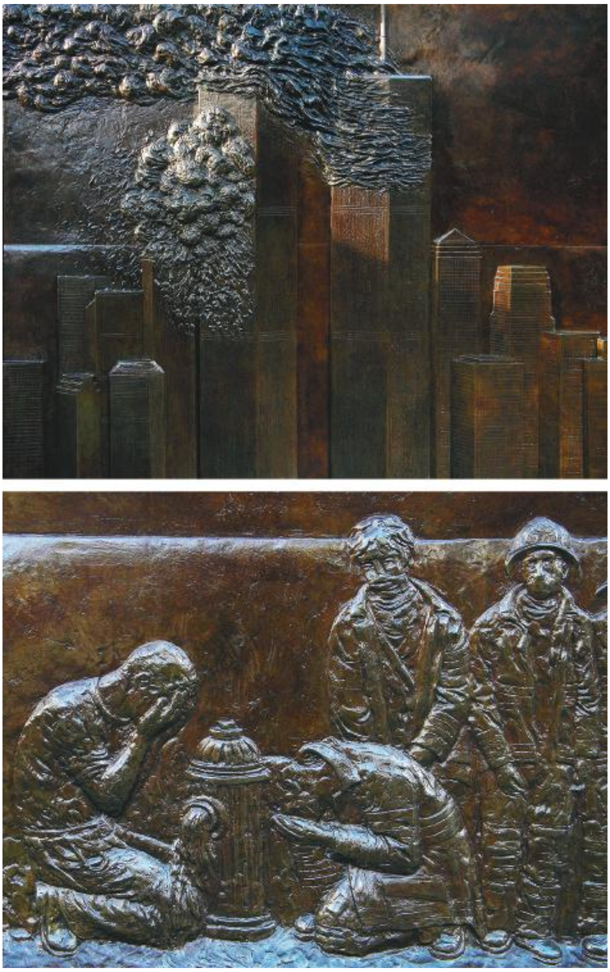

Figure 4.28 Sections of the Fire Department of New York City (FDNY) Memorial Wall. Memorial to the Fallen Firefighters of 9/11, at FDNY Engine Co. 10 on Liberty Street, New York, by Rambusch Studios. (© Richard C. Li.)

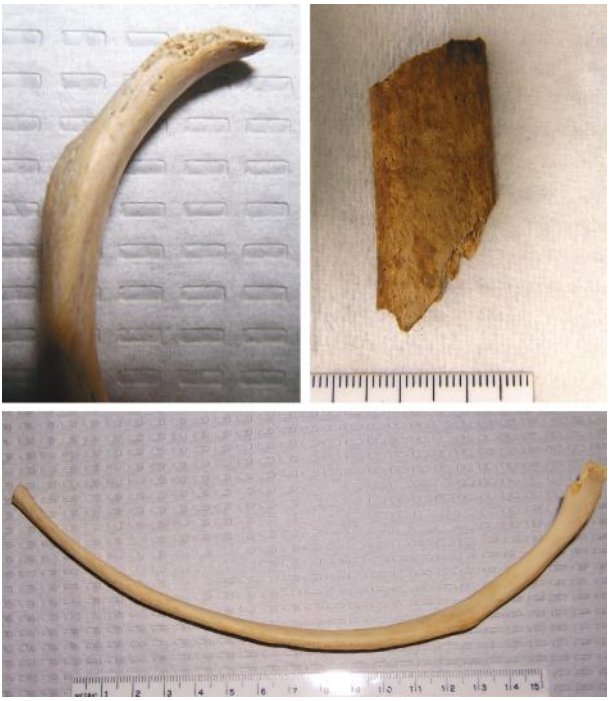

Figure 4.29 Human rib bone fragments recovered from burial site. (© Richard C. Li.)

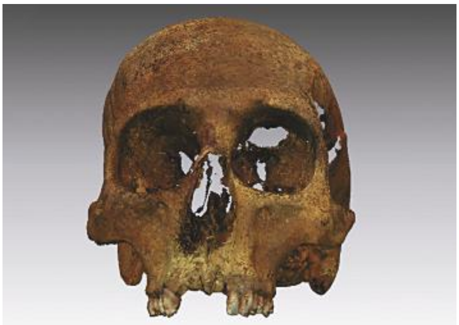

Figure 4.30 Skeletal remains, exhumed from Frombork Cathedral in Poland, are thought to be those of astronomer Nicolaus Copernicus (1473–1543). (From Bogdanowicz, W., et al., *Proc Natl Acad Sci U S A*, 106, 12279–12282, 2009. With permission.)

removed to improve the yield of DNA. A decalcification method can be utilized to dissolve calcium ions to soften the bone tissue. Additionally, the application of proteinase can be used to digest the matrix proteins, thus increasing the yield of DNA that is harvested from osteocytes.

4.3.4 Teeth

4.3.4.1 Biology of Teeth

During embryonic development, two sets of teeth begin to form. The first to appear are the *deciduous teeth* or *primary teeth*. Most children have 20 deciduous teeth, which are later replaced by 32 teeth known as the *secondary dentition* or *permanent dentition* (Figure 4.32).

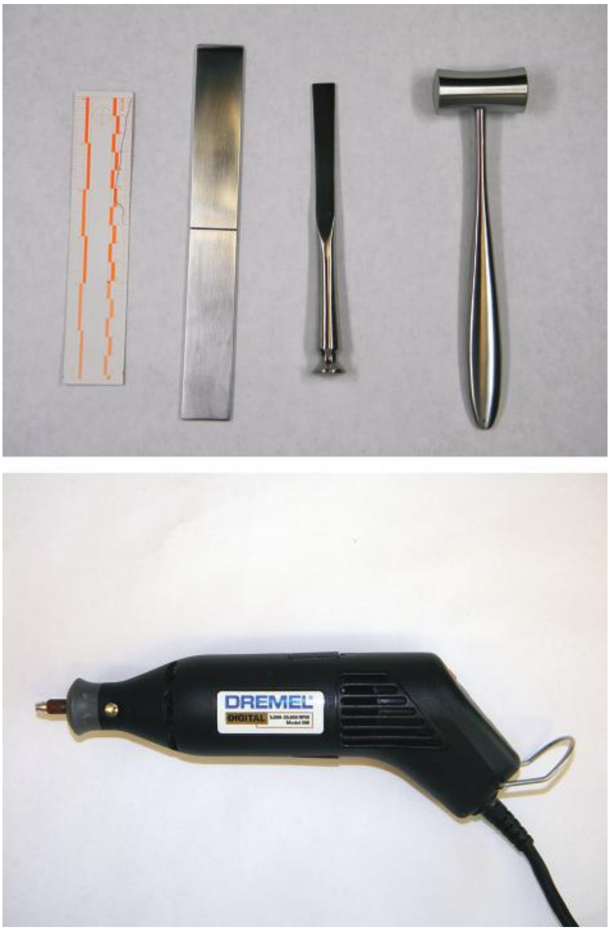

Figure 4.31 Tools for cutting bone samples. Osteotomes and a mallet (top), and a rotary device (bottom). (© Richard C. Li.)

The bulk of each tooth consists of a calcified connective tissue called *dentin* (Figure 4.33). The dentin of the crown is covered by a layer of *enamel*. The rest of the dentin is referred to as radicular dentin and is covered by a layer of *cementum*, which separates the tooth from the surrounding jawbone.

Similar to bone, tooth tissue contains a matrix. The inorganic matrix of tooth contains hydroxyapatite, a calcium phosphate in a crystalline form. The organic matrices of dentin and cementum are primarily collagens. In enamel, amelogenin is the major protein of the organic matrix. Other proteins include ameloblastin and enamelin, which are also the components of the enamel organic matrix.

The interior chamber within the tooth surrounded by dentin is known as the pulp cavity. The *dental pulp*, found within the pulp cavity, is the connective tissue made up of nerve fibers, blood vessels, and various cells. The blood vessels and nerves in the pulp cavity are innervated through the root canal, a narrow tunnel located at the root of the tooth. Incisor and cuspid teeth have single roots. Bicuspids have one or two roots. Molars typically have three or more roots.

The columnar cell bodies of *odontoblasts* are located along the peripheral dental pulp. A single *odontoblast process*, arising from each cell body of the odontoblasts, projects into the dentinal tubule (Figure 4.34). Odontoblasts play important roles in the formation of dentin. Odontoblasts secrete collagens and ground substances that are the components of the

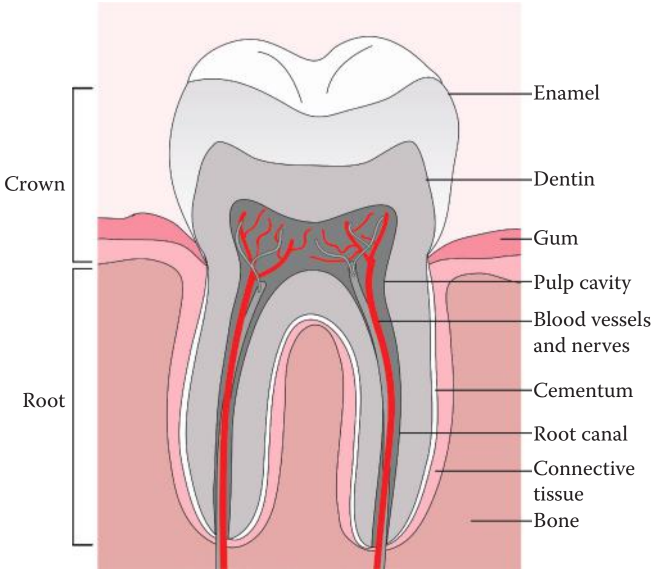
Crown
Root
Enamel
Dentin
Gum
Pulp cavity
Blood vessels and nerves
Cementum
Root canal
Connective tissue
Bone

Figure 4.32 A sectional view of an adult tooth. (© Richard C. Li.)

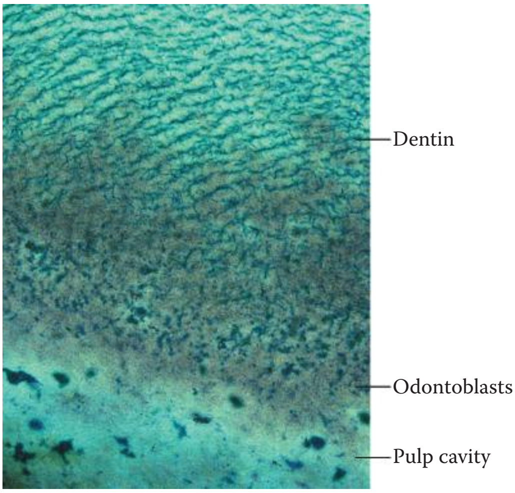
Dentin
Odontoblasts
Pulp cavity

Figure 4.33 Cross-sectional view of a tooth. (© Richard C. Li.)

dentinal matrix. Additionally, odontoblasts regulate the calcification of the matrix of dentin. *Cementoblasts* are cells that play roles in forming the cementum. Cementoblasts secrete collagens and ground substances to form the extracellular matrix of the cementum. Through the process of forming cementum, cementoblasts become trapped in the extracellular matrix. The cementoblasts embedded in the cementum are referred to as *cementocytes*. *Ameloblasts* are cells that play a role in producing enamel and are subsequently lost during tooth eruption.

4.3.4.2 Teeth as Source of DNA Evidence

The characteristics of teeth, their alignment, and the overall structure of the mouth provide information for identifying a person. The use of dental records such as x-rays and dental casts can allow dental remains to be connected to a victim. Particularly in circumstances such as

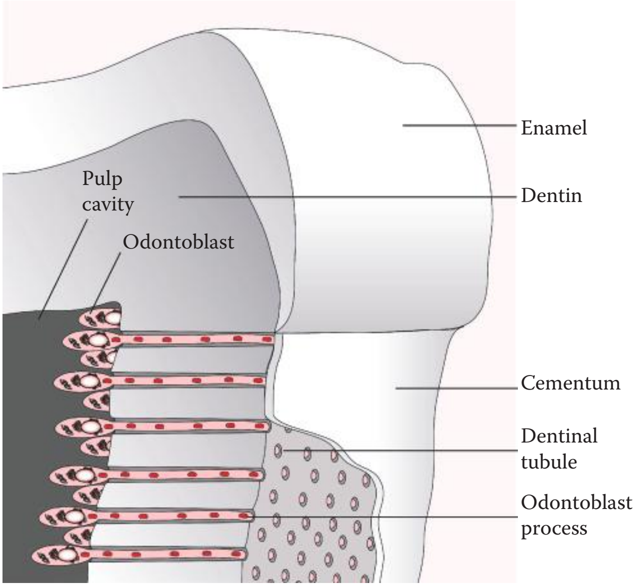

Figure 4.34 Odontoblast processes. See Section 4.3.4.1. (© Richard C. Li.)

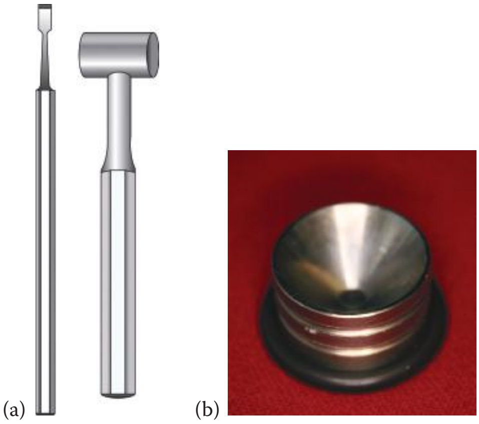

Figure 4.35 Tools for dissecting teeth. (© Richard C. Li.) (a) A dental chisel and a mallet. (© Richard C. Li.) (b) An amalgam well. (Courtesy of Dr. Ken Hermsen.)

decomposition, an odontological comparison is possible since the dental evidence often remains intact. When no antemortem dental record is available for comparison, forensic DNA testing can be carried out for postmortem human identification. The mineralized dental structure protects DNA from degradation in cases where it may be degraded in other tissues of the body. Thus, teeth are an excellent source of DNA for forensic DNA analysis under such conditions.

Dental pulp tissue contains various cells and is a suitable source of DNA. However, when tooth evidence has been exposed to high temperature and humid environments, decomposition of the pulp tissue can occur. Thus, cementoblasts within the cementum, containing both nuclei and mitochondria, can then be utilized as a source of DNA. Additionally, odontoblast processes within dentin, containing mitochondria, can also be used. Several different methods are used to obtain dental tissues for DNA isolation (Figure 4.35). A vertical section is cut along the longitudinal axis of the tooth, which allows the dissection of the pulp, dentin, and cementum tissues

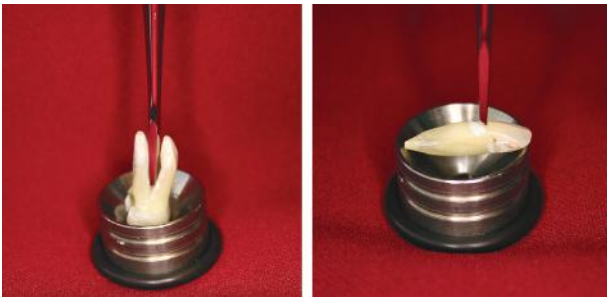

Figure 4.36 Dissecting teeth for DNA isolation. Vertical cutting and dissecting of the pulp, dentin, and cementum tissues (left) and horizontal cutting and dissecting of the root portion (right). (Courtesy of Dr. Ken Hermsen.)

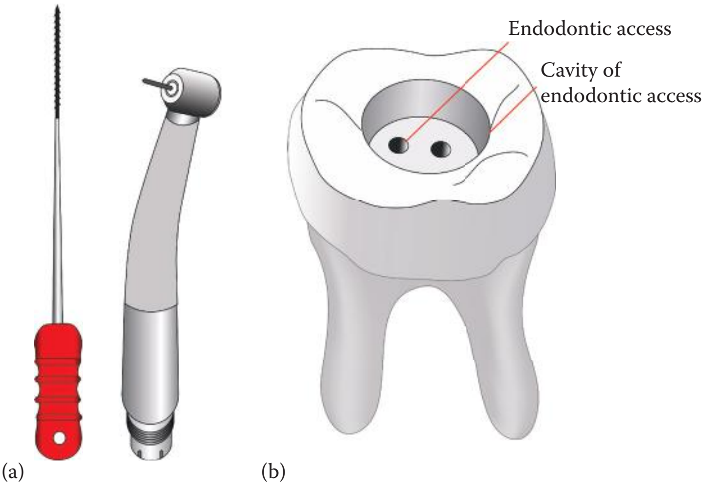

Figure 4.37 Extraction of pulp tissue using endodontic access procedures: trepanning the occlusal surface of a tooth using a dental bur mounted on a turbine (a), creating a cavity for endodontic access (b), and extracting pulp tissue using a nerve broach (a). (© Richard C. Li.)

(Figure 4.36). A horizontal section is cut through the cementum–enamel junction of the tooth (Figure 4.36). The root portion can be pulverized to a fine powder for DNA isolation. The crown can be preserved for forensic odontological comparisons if needed. Additionally, the extraction of pulp tissue can be carried out by standard endodontic access (Figure 4.37). For calcified tissues such as dentin and cementum, a decalcification step is needed to soften the dental matrix, which facilitates DNA isolation.

Bibliography

Abaz, J., et al., Comparison of the variables affecting the recovery of DNA from common drinking containers. *Forensic Sci Int*, 2002, **126**(3): 233–240.

Aditya, S., C.N. Bhattacharyya, and K. Chaudhuri, Generating STR profile from "Touch DNA". *J Forensic Leg Med*, 2011, **18**(7): 295–298.

Alakoc, Y.D. and P.S. Aka, "Orthograde entrance technique" to recover DNA from ancient teeth preserving the physical structure. *Forensic Sci Int*, 2009, **188**(1–3): 96–98.

Bibliography

Allouche, M., et al., Genetic identification of decomposed cadavers using nails as DNA source. *Forensic Sci Int Genet*, 2008, **3**(1): 46–49.

Alvarez Garcia, A., et al., Effect of environmental factors on PCR-DNA analysis from dental pulp. *Int J Legal Med*, 1996, **109**(3): 125–129.

Amory, S., et al., STR typing of ancient DNA extracted from hair shafts of Siberian mummies. *Forensic Sci Int*, 2007, **166**(2–3): 218–229.

Anderson, T.D., et al., A validation study for the extraction and analysis of DNA from human nail material and its application to forensic casework. *J Forensic Sci*, 1999, **44**(5): 1053–1056.

Anzai-Kanto, E., et al., DNA extraction from human saliva deposited on skin and its use in forensic identification procedures. *Pesqui Odontol Bras*, 2005, **19**(3): 216–222.

Barbaro, A., P. Cormaci, and A. Barbaro, Detection of STRs from body fluid collected on IsoCode paper-based devices. *Forensic Sci Int*, 2004, **146**(Suppl): S127–S128.

Barbaro, A., P. Cormaci, and A. Barbaro, DNA analysis from mixed biological materials. *Forensic Sci Int*, 2004, **146**(Suppl): S123–S125.

Barbaro, A., et al., Anonymous letters? DNA and fingerprints technologies combined to solve a case. *Forensic Sci Int*, 2004, **146**(Suppl): S133–S134.

Berger, B., et al., Chimerism in DNA of buccal swabs from recipients after allogeneic hematopoietic stem cell transplantations: Implications for forensic DNA testing. *Int J Legal Med*, 2013, **127**(1): 49–54.

Bogdanowicz, W., et al., Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus. *Proc Natl Acad Sci U S A*, 2009, **106**(30), 12279–12282.

Bond, J.W., Value of DNA evidence in detecting crime. *J Forensic Sci*, 2007, **52**(1): 128–136.

Bourguignon, L., et al., A fluorescent microscopy-screening test for efficient STR-typing of telogen hair roots. *Forensic Sci Int Genet*, 2008, **3**(1): 27–31.

Bright, J.A., et al., The effect of cleaning agents on the ability to obtain DNA profiles using the Identifiler™ and PowerPlex® Y Multiplex Kits. *J Forensic Sci*, 2011, **56**(1): 181–185.

Bruck, S., et al., Single cells for forensic DNA analysis—From evidence material to test tube. *J Forensic Sci*, 2011, **56**(1): 176–180.

Burger, M.F., E.Y. Song, and J.W. Schumm, Buccal DNA samples for DNA typing: New collection and processing methods. *Biotechniques*, 2005, **39**(2): 257–261.

Castella, V., M. Lesta Mdel, and P. Mangin, One person with two DNA profiles: A(nother) case of mosaicism or chimerism. *Int J Legal Med*, 2009, **123**(5): 427–430.

Castella, V., et al., Successful DNA typing of ultrafiltered urines used to detect EPO doping. *Forensic Sci Int Genet*, 2007, **1**(3–4): 281–282.

Chiou, F.S., et al., Extraction of human DNA for PCR from chewed residues of betel quid using a novel "PVP/CTAB" method. *J Forensic Sci*, 2001, **46**(5): 1174–1179.

Cina, M.S., et al., Isolation and identification of male and female DNA on a postcoital condom. *Arch Pathol Lab Med*, 2000, **124**(7): 1083–1086.

Cook, O. and L. Dixon, The prevalence of mixed DNA profiles in fingernail samples taken from individuals in the general population. *Forensic Sci Int Genet*, 2007, **1**(1): 62–68.

Corte-Real, A., et al., The tooth for molecular analysis and identification: A forensic approach. *J Forensic Odontostomatol*, 2012, **30**(1): 22–28.

Courts, C. and B. Madea, Micro-RNA—A potential for forensic science? *Forensic Sci Int*, 2010, **203**(1–3): 106–111.

Courts, C., B. Madea, and C. Schyma, Persistence of biological traces in gun barrels—An approach to an experimental model. *Int J Legal Med*, 2012, **126**(3): 391–397.

Daly, D.J., C. Murphy, and S.D. McDermott, The transfer of touch DNA from hands to glass, fabric and wood. *Forensic Sci Int Genet*, 2012, **6**(1): 41–46.

de Lourdes Chavez-Briones, M., et al., Identification of human remains by DNA analysis of the gastrointestinal contents of fly larvae. *J Forensic Sci*, 2013, **58**(1): 248–250.

Delabarde, T., et al., The potential of forensic analysis on human bones found in riverine environment. *Forensic Sci Int*, 2013, **228**(1–3): e1–e5.

Dimo-Simonin, N., F. Grange, and C. Brandt-Casadevall, PCR-based forensic testing of DNA from stained cytological smears. *J Forensic Sci*, 1997, **42**(3): 506–509.

Dissing, J., A. Sondervang, and S. Lund, Exploring the limits for the survival of DNA in blood stains. *J Forensic Leg Med*, 2010, **17**(7): 392–396.

Dowlman, E.A., et al., The prevalence of mixed DNA profiles on fingernail swabs. *Sci Justice*, 2010, **50**(2): 64–71.

Durdle, A., R.J. Mitchell, and R.A. van Oorschot, The human DNA content in artifacts deposited by the blowfly *Lucilia cuprina* fed human blood, semen and saliva. *Forensic Sci Int*, 2013, **233**(1–3): 212–219.

Edson, S.M. and A.F. Christensen, Field contamination of skeletonized human remains with exogenous DNA. *J Forensic Sci*, 2013, **58**(1): 206–209.

Edson, J., et al., A quantitative assessment of a reliable screening technique for the STR analysis of telogen hair roots. *Forensic Sci Int Genet*, 2013, **7**(1): 180–188.

Ellis, M.A., et al., An evaluation of DNA yield, DNA quality and bite registration from a dental impression wafer. *J Am Dent Assoc*, 2007, **138**(9): 1234–1240; quiz 1267.

Farmen, R.K., et al., Assessing the presence of female DNA on post-coital penile swabs: Relevance to the investigation of sexual assault. *J Forensic Leg Med*, 2012, **19**(7): 386–389.

Flanagan, N. and C. McAlister, The transfer and persistence of DNA under the fingernails following digital penetration of the vagina. *Forensic Sci Int Genet*, 2011, **5**(5): 479–483.

Foran, D.R., M.E. Gehring, and S.E. Stallworth, The recovery and analysis of mitochondrial DNA from exploded pipe bombs. *J Forensic Sci*, 2009, **54**(1): 90–94.

Fridez, F. and R. Coquoz, PCR DNA typing of stamps: Evaluation of the DNA extraction. *Forensic Sci Int*, 1996, **78**(2): 103–110.

Goray, M., R.J. Mitchell, and R.A. van Oorschot, Investigation of secondary DNA transfer of skin cells under controlled test conditions. *Leg Med (Tokyo)*, 2010, **12**(3): 117–120.

Goray, M., J.R. Mitchell, and R.A. van Oorschot, Evaluation of multiple transfer of DNA using mock case scenarios. *Leg Med (Tokyo)*, 2012, **14**(1): 40–46.

Goray, M., R.A. van Oorschot, and J.R. Mitchell, DNA transfer within forensic exhibit packaging: Potential for DNA loss and relocation. *Forensic Sci Int Genet*, 2012, **6**(2): 158–166.

Goray, M., et al., Secondary DNA transfer of biological substances under varying test conditions. *Forensic Sci Int Genet*, 2010, **4**(2): 62–67.

Graham, E.A. and G.N. Rutty, Investigation into "normal" background DNA on adult necks: Implications for DNA profiling of manual strangulation victims. *J Forensic Sci*, 2008, **53**(5): 1074–1082.

Graham, E.A., E.E. Turk, and G.N. Rutty, Room temperature DNA preservation of soft tissue for rapid DNA extraction: An addition to the disaster victim identification investigators toolkit? *Forensic Sci Int Genet*, 2008, **2**(1): 29–34.

Grskovic, B., et al., Effect of ultraviolet C radiation on biological samples. *Croat Med J*, 2013, **54**(3): 263–271.

Grubwieser, P., et al., Systematic study on STR profiling on blood and saliva traces after visualization of fingerprint marks. *J Forensic Sci*, 2003, **48**(4): 733–741.

Gurvitz, A., L.Y. Lai, and B.A. Neilan, Exploiting biological materials in forensic science. *Australas Biotechnol*, 1994, **4**(2): 88–91.

Hellmann, A., et al., STR typing of human telogen hairs—A new approach. *Int J Legal Med*, 2001, **114**(4–5): 269–273.

Herber, B. and K. Herold, DNA typing of human dandruff. *J Forensic Sci*, 1998, **43**(3): 648–656.

Higgins, D. and J.J. Austin, Teeth as a source of DNA for forensic identification of human remains: A review. *Sci Justice*, 2013, **53**(4): 433–441.

Higuchi, R., et al., DNA typing from single hairs. *Nature*, 1988, **332**(6164): 543–546.

Hochmeister, M.N., PCR analysis of DNA from fresh and decomposed bodies and skeletal remains in medicolegal death investigations. *Methods Mol Biol*, 1998, **98**: 19–26.

Hochmeister, M.N., O. Rudin, and E. Ambach, PCR analysis from cigarette butts, postage stamps, envelope sealing flaps, and other saliva-stained material. *Methods Mol Biol*, 1998, **98**: 27–32.

Hochmeister, M.N., et al., PCR-based typing of DNA extracted from cigarette butts. *Int J Legal Med*, 1991, **104**(4): 229–233.

Hoile, R., et al., Gamma irradiation as a biological decontaminant and its effect on common fingermark detection techniques and DNA profiling. *J Forensic Sci*, 2010, **55**(1): 171–177.

Hopkins, B., et al., The use of minisatellite variant repeat-polymerase chain reaction (MVR-PCR) to determine the source of saliva on a used postage stamp. *J Forensic Sci*, 1994, **39**(2): 526–531.

Hopwood, A.J., A. Mannucci, and K.M. Sullivan, DNA typing from human faeces. *Int J Legal Med*, 1996, **108**(5): 237–243.

Hopwood, A., et al., Rapid quantification of DNA samples extracted from buccal scrapes prior to DNA profiling. *Biotechniques*, 1997, **23**(1): 18–20.

Johnson, D.J., A.C. Calderaro, and K.A. Roberts, Variation in nuclear DNA concentrations during urination. *J Forensic Sci*, 2007, **52**(1): 110–113.

Kaarstad, K., et al., The detection of female DNA from the penis in sexual assault cases. *J Forensic Leg Med*, 2007, **14**(3): 159–160.

Kaiser, C., et al., Molecular study of time dependent changes in DNA stability in soil buried skeletal residues. *Forensic Sci Int*, 2008, **177**(1): 32–36.

Kamodyova, N., et al., Prevalence and persistence of male DNA identified in mixed saliva samples after intense kissing. *Forensic Sci Int Genet*, 2013, **7**(1): 124–128.

Kamphausen, T., et al., Good shedder or bad shedder—The influence of skin diseases on forensic DNA analysis from epithelial abrasions. *Int J Legal Med*, 2012, **126**(1): 179–183.

Kenna, J., et al., The recovery and persistence of salivary DNA on human skin. *J Forensic Sci*, 2011, **56**(1): 170–175.

Kester, K.M., et al., Recovery of environmental human DNA by insects. *J Forensic Sci*, 2010, **55**(6): 1543–1551.

Kim, M., et al., Identification and long term stability of DNA captured on a dental impression wafer. *Pediatr Dent*, 2012, **34**(5): 373–377.

Kita, T., et al., Morphological study of fragmented DNA on touched objects. *Forensic Sci Int Genet*, 2008, **3**(1): 32–36.

Kitayama, T., et al., Evaluation of a new experimental kit for the extraction of DNA from bones and teeth using a non-powder method. *Leg Med (Tokyo)*, 2010, **12**(2): 84–89.

Kline, M.C., et al., Polymerase chain reaction amplification of DNA from aged blood stains: Quantitative evaluation of the "suitability for purpose" of four filter papers as archival media. *Anal Chem*, 2002, **74**(8): 1863–1869.

Kohnemann, S., et al., qPCR and mtDNA SNP analysis of experimentally degraded hair samples and its application in forensic casework. *Int J Legal Med*, 2010, **124**(4): 337–342.

Lee, E.J., et al., The effects of different maceration techniques on nuclear DNA amplification using human bone. *J Forensic Sci*, 2010, **55**(4): 1032–1038.

Li, R.C. and H.A. Harris, Using hydrophilic adhesive tape for collection of evidence for forensic DNA analysis. *J Forensic Sci*, 2003, **48**(6): 1318–1321.

Lijnen, I. and G. Willems, DNA research in forensic dentistry. *Methods Find Exp Clin Pharmacol*, 2001, **23**(9): 511–517.

Linacre, A., et al., Generation of DNA profiles from fabrics without DNA extraction. *Forensic Sci Int Genet*, 2010, **4**(2): 137–141.

Linch, C.A., The ultrastructure of tissue attached to telogen hair roots. *J Forensic Sci*, 2008, **53**(6): 1363–1366.

Linch, C.A., Degeneration of nuclei and mitochondria in human hairs. *J Forensic Sci*, 2009, **54**(2): 346–349.

Linch, C.A., et al., Specific melanin content in human hairs and mitochondrial DNA typing success. *Am J Forensic Med Pathol*, 2009, **30**(2): 162–166.

Lindahl, T., Instability and decay of the primary structure of DNA. *Nature*, 1993, **362**(6422): 709–715.

Lorente, M., et al., Dandruff as a potential source of DNA in forensic casework. *J Forensic Sci*, 1998, **43**(4): 901–902.

Maguire, S., et al., Retrieval of DNA from the faces of children aged 0–5 years: A technical note. *J Forensic Nurs*, 2008, **4**(1): 40–44.

Malsom, S., et al., The prevalence of mixed DNA profiles in fingernail samples taken from couples who cohabit using autosomal and Y-STRs. *Forensic Sci Int Genet*, 2009, **3**(2): 57–62.

Marrone, A. and J. Ballantyne, Changes in dry state hemoglobin over time do not increase the potential for oxidative DNA damage in dried blood. *PLoS One*, 2009, **4**(4): e5110.

Mayntz-Press, K.A., et al., Y-STR profiling in extended interval (> or = 3 days) postcoital cervicovaginal samples. *J Forensic Sci*, 2008, **53**(2): 342–348.

McCartney, C., LCN DNA: Proof beyond reasonable doubt? *Nat Rev Genet*, 2008, **9**(5): 325.

Medintz, I., et al., Restriction fragment length polymorphism and polymerase chain reaction-HLA DQ alpha analysis of casework urine specimens. *J Forensic Sci*, 1994, **39**(6): 1372–1380.

Misner, L.M., et al., The correlation between skeletal weathering and DNA quality and quantity. *J Forensic Sci*, 2009, **54**(4): 822–828.

National Institute of Justice. Using DNA to solve cold cases, Special Report, 2002, US Department of Justice, Office of Justice Programs, Washington, DC.

Ng, L.K., et al., Optimization of recovery of human DNA from envelope flaps using DNA IQ System for STR genotyping. *Forensic Sci Int Genet*, 2007, **1**(3–4): 283–286.

Niemcunowicz-Janica, A., et al., Typeability of PowerPlex Y (Promega) profiles in selected tissue samples incubated in various environments. *Arch Med Sadowej Kryminol*, 2007, **57**(4): 385–388.

Niemcunowicz-Janica, A., et al., Detectability of SGM Plus profiles in heart and lungs tissue samples incubated in different environments. *Leg Med (Tokyo)*, 2008, **10**(1): 35–38.

Nurit, B., et al., Evaluating the prevalence of DNA mixtures found in fingernail samples from victims and suspects in homicide cases. *Forensic Sci Int Genet*, 2011, **5**(5): 532–537.

O'Driscoll, L., Extracellular nucleic acids and their potential as diagnostic, prognostic and predictive biomarkers. *Anticancer Res*, 2007, **27**(3A): 1257–1265.

Offele, D., et al., Soft tissue removal by maceration and feeding of *Dermestes* sp.: Impact on morphological and biomolecular analyses of dental tissues in forensic medicine. *Int J Legal Med*, 2007, **121**(5): 341–348.

Ohira, H., et al., Effective appropriate use of dental remains and forensic DNA testing for personal identity confirmation. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S560–S562.

Opel, K.L., et al., Evaluation and quantification of nuclear DNA from human telogen hairs. *J Forensic Sci*, 2008, **53**(4): 853–857.

Paneto, G.G., et al., Heteroplasmy in hair: Differences among hair and blood from the same individuals are still a matter of debate. *Forensic Sci Int*, 2007, **173**(2–3): 117–121.

Pelotti, S., et al., Cancerous tissues in forensic genetic analysis. *Genet Test*, 2007, **11**(4): 397–400.

Pfeiffer, H., et al., Influence of soil storage and exposure period on DNA recovery from teeth. *Int J Legal Med*, 1999, **112**(2): 142–144.

Phipps, M. and S. Petricevic, The tendency of individuals to transfer DNA to handled items. *Forensic Sci Int*, 2007, **168**(2–3): 162–168.

Pinchi, V., et al., Techniques of dental DNA extraction: Some operative experiences. *Forensic Sci Int*, 2011, **204**(1–3): 111–114.

Prado, V.F., et al., Extraction of DNA from human skeletal remains: Practical applications in forensic sciences. *Genet Anal*, 1997, **14**(2): 41–44.

Quinones, I. and B. Daniel, Cell free DNA as a component of forensic evidence recovered from touched surfaces. *Forensic Sci Int Genet*, 2012, **6**(1): 26–30.

Raimann, P.E., et al., Procedures to recover DNA from pre-molar and molar teeth of decomposed cadavers with different post-mortem intervals. *Arch Oral Biol*, 2012, **57**(11): 1459–1466.

Ramasamy, S., A. Houspian, and F. Knott, Recovery of DNA and fingermarks following deployment of render-safe tools for vehicle-borne improvised explosive devices (VBIED). *Forensic Sci Int*, 2011, **210**(1–3): 182–187.

Remualdo, V.R. and R.N. Oliveira, Analysis of mitochondrial DNA from the teeth of a cadaver maintained in formaldehyde. *Am J Forensic Med Pathol*, 2007, **28**(2): 145–146.

Riemer, L.B., D. Fairley, and D. Sweet, DNA collection from used toothbrushes as a means to decedent identification. *Am J Forensic Med Pathol*, 2012, **33**(4): 354–356.

Roberts, K.A. and C. Calloway, Mitochondrial DNA amplification success rate as a function of hair morphology. *J Forensic Sci*, 2007, **52**(1): 40–47.

Roeper, A., W. Reichert, and R. Mattern, The Achilles tendon as a DNA source for STR typing of highly decayed corpses. *Forensic Sci Int*, 2007, **173**(2–3): 103–106.

Rubio, L., et al., Study of short- and long-term storage of teeth and its influence on DNA. *J Forensic Sci*, 2009, **54**(6): 1411–1413.

Rubio, L., et al., Time-dependent changes in DNA stability in decomposing teeth over 18 months. *Acta Odontol Scand*, 2013, **71**(3–4): 638–643.

Salvador, J.M. and M.C. De Ungria, Isolation of DNA from saliva of betel quid chewers using treated cards. *J Forensic Sci*, 2003, **48**(4): 794–797.

Schneider, H., et al., Hot flakes in cold cases. *Int J Legal Med*, 2011, **125**(4): 543–548.

Schwark, T., et al., Phantoms in the mortuary—DNA transfer during autopsies. *Forensic Sci Int*, 2012, **216**(1–3): 121–126.

Sewell, J., et al., Recovery of DNA and fingerprints from touched documents. *Forensic Sci Int Genet*, 2008, **2**(4): 281–285.

Shaw, K., et al., Comparison of the effects of sterilisation techniques on subsequent DNA profiling. *Int J Legal Med*, 2007, **122**(1): 29–33.

Shintani-Ishida, K., et al., Usefulness of blood vessels as a DNA source for PCR-based genotyping based on two cases of corpse dismemberment. *Leg Med (Tokyo)*, 2010, **12**(1): 8–12.

Simons, J.L. and S.K. Vintiner, Effects of histological staining on the analysis of human DNA from archived slides. *J Forensic Sci*, 2011, **56**(Suppl 1): S223–S228.

Soares-Vieira, J.A., et al., Y-STRs in forensic medicine: DNA analysis in semen samples of azoospermic individuals. *J Forensic Sci*, 2007, **52**(3): 664–670.

Sosa, C., et al., Nuclear DNA typing from ancient teeth. *Am J Forensic Med Pathol*, 2012, **33**(3): 211–214.

Sosa, C., et al., Association between ancient bone preservation and DNA yield: A multidisciplinary approach. *Am J Phys Anthropol*, 2013, **151**(1): 102–109.

Springer, E.E., T.L. Laber, and B.B. Randall, The examination of vaginally inserted plastic tampon applicators for genetic markers and evidence of prior sexual intercourse. *J Forensic Sci*, 1988, **33**(5): 1139–1145.

Steadman, D.W., et al., The effects of chemical and heat maceration techniques on the recovery of nuclear and mitochondrial DNA from bone. *J Forensic Sci*, 2006, **51**(1): 11–17.

Szabo, S., et al., *In situ* labeling of DNA reveals interindividual variation in nuclear DNA breakdown in hair and may be useful to predict success of forensic genotyping of hair. *Int J Legal Med*, 2012, **126**(1): 63–70.

Taguchi, M., et al., DNA identification of formalin-fixed organs is affected by fixation time and type of fixatives: Using the AmpF l STR(R) Identifiler(R) PCR Amplification Kit. *Med Sci Law*, 2012, **52**(1): 12–16.

Tanaka, M., et al., Usefulness of a toothbrush as a source of evidential DNA for typing. *J Forensic Sci*, 2000, **45**(3): 674–676.

Tang, D.L., et al., Multiplex fluorescent PCR for noninvasive prenatal detection of fetal-derived paternally inherited diseases using circulatory fetal DNA in maternal plasma. *Eur J Obstet Gynecol Reprod Biol*, 2009, **144**(1): 35–39.

Tate, C.M., et al., Evaluation of circular DNA substrates for whole genome amplification prior to forensic analysis. *Forensic Sci Int Genet*, 2012, **6**(2): 185–190.

Theodore Harcke, H., et al., Forensic imaging-guided recovery of nuclear DNA from the spinal cord. *J Forensic Sci*, 2009, **54**(5): 1123–1126.

Tilotta, F., et al., A comparative study of two methods of dental pulp extraction for genetic fingerprinting. *Forensic Sci Int*, 2010, **202**(1–3): e39–e43.

Toothman, M.H., et al., Characterization of human DNA in environmental samples. *Forensic Sci Int*, 2008, **178**(1): 7–15.

Tsuchimochi, T., et al., Chelating resin-based extraction of DNA from dental pulp and sex determination from incinerated teeth with Y-chromosomal alphoid repeat and short tandem repeats. *Am J Forensic Med Pathol*, 2002, **23**(3): 268–271.

van Oorschot, R.A., et al., Beware of the possibility of fingerprinting techniques transferring DNA. *J Forensic Sci*, 2005, **50**(6): 1417–1422.

Vandewoestyne, M., et al., Presence and potential of cell free DNA in different types of forensic samples. *Forensic Sci Int Genet*, 2013, **7**(2): 316–320.

von Wurmb-Schwark, N., et al., Fast and simple DNA extraction from saliva and sperm cells obtained from the skin or isolated from swabs. *Leg Med (Tokyo)*, 2006, **8**(3): 177–181.

Warshauer, D.H., et al., An evaluation of the transfer of saliva-derived DNA. *Int J Legal Med*, 2012, **126**(6): 851–861.

Webb, L.G., S.E. Egan, and G.R. Turbett, Recovery of DNA for forensic analysis from lip cosmetics. *J Forensic Sci*, 2001, **46**(6): 1474–1479.

Wiegand, P., T. Bajanowski, and B. Brinkmann, DNA typing of debris from fingernails. *Int J Legal Med*, 1993, **106**(2): 81–83.

Yasuda, T., et al., A simple method of DNA extraction and STR typing from urine samples using a commercially available DNA/RNA extraction kit. *J Forensic Sci*, 2003, **48**(1): 108–110.

SECTION II

Basic Techniques in Forensic Biology

5

Nucleic Acid Extraction

5.1 Basic Principles of DNA Extraction

Deoxyribonucleic acid (DNA) is a linear polynucleotide consisting of four types of monomeric nucleotides. Each nucleotide contains three components: a deoxyribose, a nitrogenous base, and a phosphate group (Figure 5.1). The four bases for DNA are adenine (A), cytosine (C), guanine (G), and thymine (T) (Figure 5.2). The deoxyribose is attached to the nitrogen of a base. The phosphate group is attached to the deoxyribose. In a *polynucleotide*, individual nucleotides are linked by phosphodiester bonds (Figure 5.3).

Interestingly, the first attempt to isolate DNA from humans, which involved the use of human leucocytes, was accomplished by the Swiss physician Friedrich Miescher in 1869. This study led to the discovery of the DNA molecule, which he referred to as "nuclein." Over the years, various methods for isolating DNA from blood samples were developed. However, many protocols still use the basic principles Miescher developed more than 100 years ago.

This chapter introduces basic techniques for DNA extraction used in forensic laboratories (Table 5.1). The commonly used methods isolate total cellular DNA, which is suitable for most forensic DNA analyses. However, DNA extraction procedures may vary according to the type of biological evidence, which can include cell types, substrates, the quantity of biological evidence collected, and the type of test that is being performed. Specific DNA extraction procedures for a sample of interest can be found in the literature. The extraction method of choice yields an optimal quantity, quality, and purity of DNA to satisfy forensic DNA testing needs. A sufficient quantity of DNA ensures the generation of a complete DNA profile. Poor quality of DNA, such as fragmentation due to DNA degradation, may result in a partial DNA profile or a failure to obtain a profile. Poor purity of DNA may cause interference during subsequent DNA testing. For example, DNA polymerase inhibitors interfere with DNA amplification (Chapter 7). Additional criteria for selecting proper DNA extraction methods include adaptability to automation, throughput potential, simplicity, the reduction of contamination risks, and cost-effectiveness. The most common DNA extraction protocols include basic components discussed below.

5.1.1 Cell and Tissue Disruption

In most DNA extraction protocols, enzymatic digestions, such as those with proteinase K, are used for cell and tissue disruption. The disruption process can also be carried out by boiling and by using alkali treatment and mechanical methods (Figures 5.4 and 5.5). Materials such as bone and teeth can be frozen in liquid nitrogen and then ground to a fine powder with a mortar or cryogenic grinder such as the SPEX CertiPrep® freezer mill (Figure 5.6). In such specimens, cells containing DNA

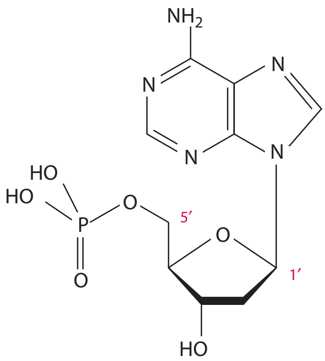

Figure 5.1 Structure of a nucleotide. Each nucleotide in a DNA polymer is made up of three components: a deoxyribose, a nitrogenous base (adenine is shown as an example), and a phosphate group. The 1′ carbon of the deoxyribose is attached to the nitrogen of a nitrogenous base. A phosphate group is attached to the 5′ carbon of the deoxyribose.

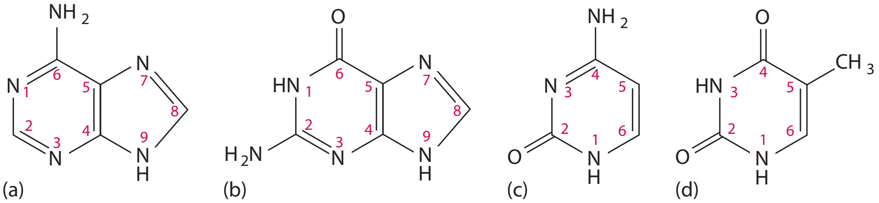

Figure 5.2 Chemical structures of nitrogenous bases: (a) adenine, (b) guanine, (c) cytosine, and (d) thymine.

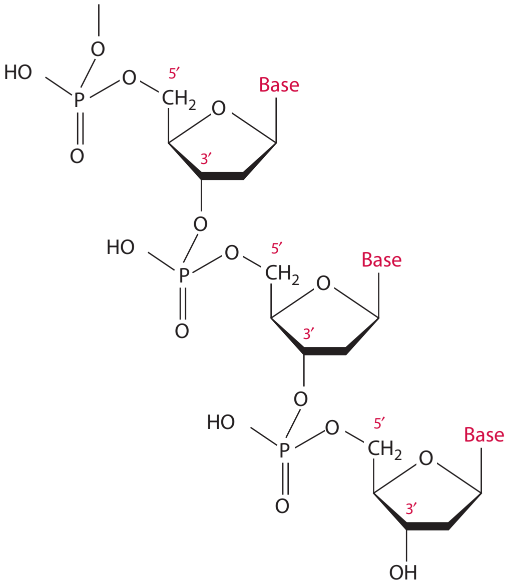

Figure 5.3 DNA polynucleotide chain. Individual nucleotides are linked by phosphodiester bonds between their 5′ and 3′ carbons.

Table 5.1 Three Basic DNA Extraction Methods

Method	Forensic Application	Purity	DNA Strand	DNA Size	Throughput	Note
Solvent-based	RFLP- and PCR-based assay	High	Double stranded	Large	Time-consuming; difficult to adapt for automation	Uses toxic solvent; multiple transfer steps between tubes
Boiling	PCR-based assay	Low	Single stranded	Small	~30 min per sample	No transfer required
Silica-based	RFLP- and PCR-based assay	High	Double stranded	Large	~1 h per sample; amenable to automation	Minimum transfer required

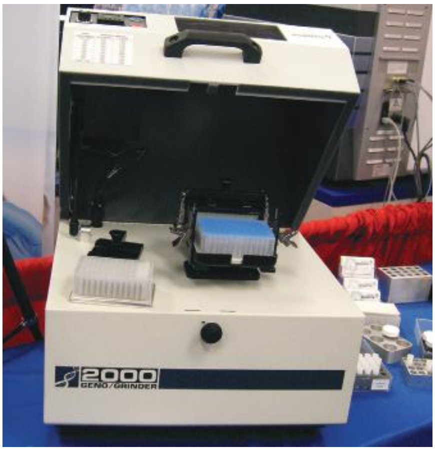

Figure 5.4 An automated mechanical disruption device for efficient disruption of tissues. It can be used for high-throughput applications involving sample preparation for the isolation of nucleic acids. (© Richard C. Li.)

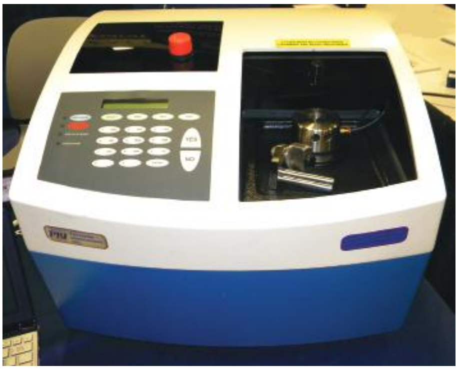

Figure 5.5 Tissue disruption using a pressure-generating instrument also known as a barocycler. During the process, a barocycler is utilized to apply alternating cycles of high and low pressures onto specimens placed in single-use processing containers. The technique is known as pressure-cycling technology (PCT). (© Richard C. Li.)

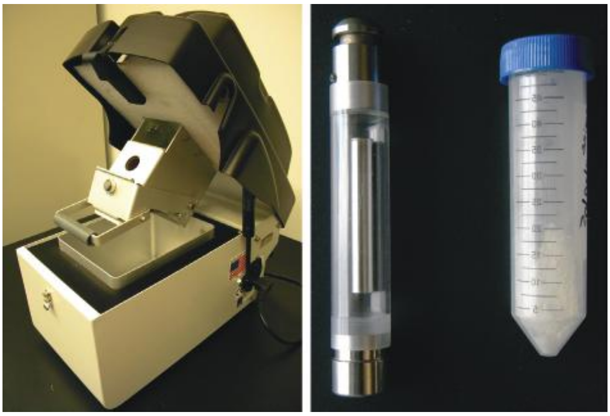

Figure 5.6 Tissue disruption using cryogenic grinding. A cryogenic mill (left), also known as freezer mill, can be used for the pulverization of hard tissue such as bone samples into fine powder prior to DNA extraction. During a grinding cycle, an impactor (right) moves back and forth, under a magnetic field, inside a grinding vial, grinding the sample to a fine powder. The grinding process is carried out at low temperatures (approximately –196°C), using liquid nitrogen to protect temperature-sensitive analytes such as DNA. (© Richard C. Li.)

are embedded in a calcified matrix that is responsible for the rigid structure of bone and teeth. In order to achieve a high yield of DNA extraction, it is necessary to remove the calcified matrix. A decalcification process removes calcium ions from the matrix, thus making the specimen suitable for DNA extraction. The most commonly used decalcifying agents for DNA extraction purposes are chelating agents such as ethylenediaminetetraacetic acid (EDTA), which sequesters the calcium ions. Decalcification is a lengthy procedure. Bone or teeth powder is usually treated with the decalcifying agent overnight or for a longer period of time, depending on the size and source of the sample.

5.1.2 Lysis of Cellular and Organelle Membranes

During or after tissue disruption, membranes—including those of cells, nuclei, and mitochondria—are lysed in order to release DNAs, including nuclear and mitochondrial DNA. The lysis can be carried out using salts and chaotropic agents (Section 5.2.3) such as guanidinium salts and detergents such as sarkosyl and sodium dodecyl sulfate (SDS). These substances can destroy membranes, denature proteins, and dissociate proteins such as histones from DNA. The lysis procedure is usually carried out in a buffer, such as Tris, in order to maintain a pH where endogenous *deoxyribonucleases* (DNases) remain inactive. DNases are a type of nuclease that catalyzes the cleavage of phosphodiester bonds of DNA. Endogenous DNases are located in cytoplasmic lysosomes and play a role in degrading the DNA of invading viruses. When cells are lysed, the DNases that are also released can degrade the extracted DNA. Chelating agents such as EDTA or Chelex® (Section 5.2.2) can therefore be used to chelate the divalent cations that are the cofactors of DNases, in order to inhibit DNase activities. Furthermore, reducing agents such as mercaptoethanol or dithiothreitol (DTT) can be used to inhibit the oxidization processes that can damage DNA.

5.1.3 Removal of Proteins and Cytoplasmic Constituents

After lysis, cytoplasmic constituents, such as proteins and liquids that interfere with DNA extraction, are removed. Proteins and lipids are usually removed by one or more rounds of extraction with organic solvents such as phenol–chloroform mixtures (Section 5.2.1). Another strategy to remove cytoplasmic constituents is to utilize the reversible binding of DNA to a solid material such as silica, which selectively binds DNA in chaotropic salt solutions. The proteins and cytoplasmic constituents can then be removed through washing steps.

5.1.4 Storage of DNA Solutions

Purified high-molecular-weight DNA is usually stored in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). EDTA is usually included in storage solutions to chelate divalent cations and thereby inhibit DNases. Such a DNA solution may be stored at 4°C or −20°C. For long-term storage, −80°C is recommended. Frequent freezing or thawing cycles should be avoided because temperature fluctuations may cause breaks of single- and double-stranded DNA. Additionally, DNA containing impurities such as extracts generated via the Chelex method are much less stable. The presence of heavy metals, such as cadmium and cobalt, can cause breakage of phosphodiester bonds in the molecules. Additionally, unsuccessful removal of exogenous or intracellular free radicals, such as a hydroxyl radical (HO·), which have unpaired valence electrons, can cause DNA damage such as strand breaks. Furthermore, contamination by DNases may lead to subsequent degradation of DNA.

5.1.5 Contamination

Contamination is usually caused by the introduction of exogenous DNA to an evidence sample. It can occur between samples, between an individual and a sample, between other organisms and a sample, or between amplified DNA and a sample. Certain procedures can be utilized to prevent the occurrence of contamination. For example, evidence and reference samples should be processed separately in different rooms to avoid sample-to-sample contamination. In situations where space is limited, the evidence sample should be processed before the reference sample. Additionally, evidence samples should be processed and extracted for isolating DNA in separate areas (or at different times) from DNA amplification areas.

Solutions and test tubes used for extraction should be DNA-free, and aerosol-resistant pipet tips should be used during the extraction process. Additionally, the levels of contamination should be monitored by using *extraction reagent blanks* (having reagents but no samples), which monitor contamination from the extraction.

5.2 Methods of DNA Extraction

5.2.1 Extraction with Phenol–Chloroform

This method is also called organic extraction. Major steps include the following:

5.2.1.1 Cell Lysis and Protein Digestion

These steps can be achieved by digestion with proteolytic enzymes such as proteinase K before extraction with organic solvents.

5.2.1.2 Extraction with Organic Solvents

The removal of proteins is carried out by extracting aqueous solutions containing DNA with a mixture of phenol:chloroform:isoamyl alcohol (25:24:1). Phenol is used to extract the proteins from the aqueous solution. Although phenol has a slightly higher density than water, it is sometimes difficult to separate it from the aqueous phase. Therefore, chloroform is utilized as it has a higher density than phenol. As a result, the phenol–chloroform mixture forms the organic phase at the bottom of the tube and is easily separated from the aqueous phase. Isoamyl alcohol is often added to the phenol–chloroform mixture to reduce foaming. During partition, DNA is solubilized in the aqueous phase, while lipids are solubilized in the organic phase. Proteins are located at the interface between the two phases (Figure 5.7).

5.2.1.3 Concentrating DNA

Two common methods for concentrating DNA are ethanol precipitation and ultrafiltration. In the first method, the DNA is precipitated from the aqueous solution with ethanol and salts. Ethanol depletes the hydration shell of DNA, thus exposing its negatively charged phosphate groups. The

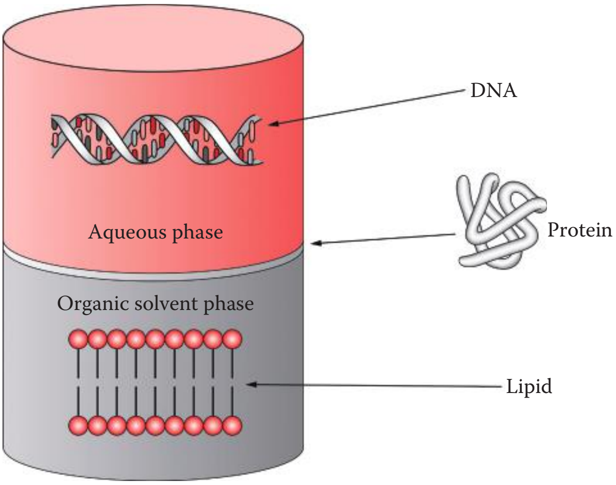
DNA
Protein
Aqueous phase
Organic solvent phase
Lipid

Figure 5.7 DNA extraction using organic solvent. The DNA is contained in the aqueous phase, while cellular materials such as lipids are contained in the organic-solvent phase. Proteins remain in the barrier between the two phases. (© Richard C. Li.)

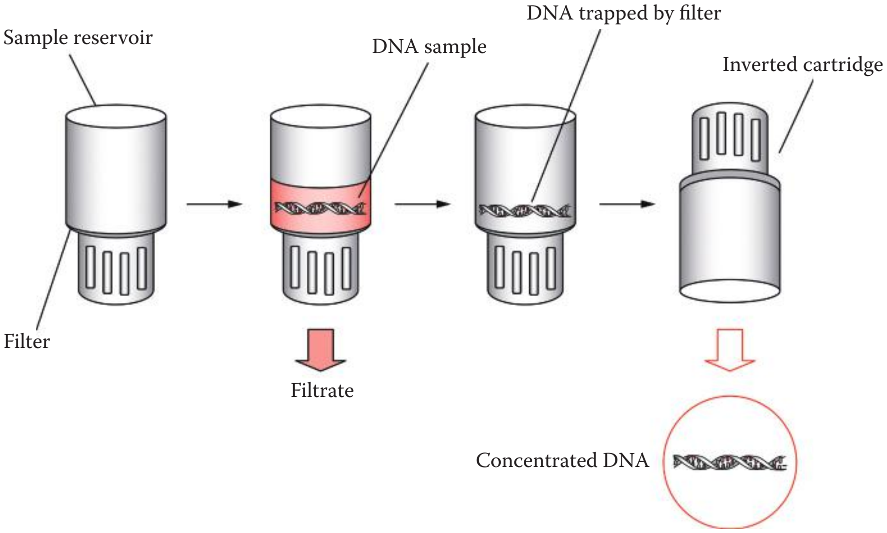
Sample reservoir
DNA sample
DNA trapped by filter
Inverted cartridge
Filter
Filtrate
Concentrated DNA

Figure 5.8 Concentrating DNA solutions using filtration devices. DNA samples are loaded into the reservoir. The liquid is filtered by centrifugation and the DNA becomes trapped in the membrane. The cartridge is then inverted to recover the trapped DNA by centrifugation. (© Richard C. Li.)

precipitation can only occur if sufficient quantities of cations are present in the solution. The most commonly used cations, such as ammonium, lithium, and sodium, neutralize the charges on the phosphate residues of DNA, forming a precipitate. Ultrafiltration is an alternative to ethanol precipitation for concentrating DNA solutions. The Microcon® and Amicon® are centrifugal ultrafiltration devices that can concentrate DNA samples (Figure 5.8). A proper Microcon® unit can be selected with a nucleotide cutoff equal to or smaller than the molecular weight of the DNA fragment of interest. Usually, the cutoff size is 100 kDa for forensic DNA samples.

Phenol–chloroform extraction yields large, double-stranded DNA and can be used for either restriction fragment length polymorphism (RFLP)-based or polymerase chain reaction

(PCR)-based analysis. However, the organic extraction method is time-consuming, involves the use of hazardous reagents, and requires the transfer of samples among tubes.

5.2.2 Extraction by Boiling Lysis and Chelation

This technique, also called the Chelex® extraction method, was introduced in the early 1990s to forensic DNA laboratories. It usually includes the following steps:

5.2.2.1 Washing

This step removes contaminants and inhibitors that may interfere with DNA amplification. For example, heme compounds found in blood samples should be removed from blood samples because it inhibits DNA amplification.

5.2.2.2 Boiling

Cells are suspended in a solution and incubated at 56°C, where DNases are not active, for 20 min. This preboiling step softens cell membranes and separates clumps of cells from each other. The cells are then lysed by heating to boiling temperature in order to break open the membranes and to release the DNA. Additionally, the lysis of cells releases all of their cellular constituents, including DNases. The DNase degradation of the extracted DNA can be blocked by applying a chelating resin (Chelex® 100) during the extraction process. Chelex® 100 is an ion-exchange resin composed of styrene divinylbenzene copolymers. The paired iminodiacetate ion groups in Chelex® 100 act as chelators by binding to divalent metal ions such as magnesium. Magnesium is a cofactor of endogenous DNases. Thus, sequestering magnesium in the solution using Chelex® 100 protects DNA from degradation by DNases (Figure 5.9).

5.2.2.3 Centrifugation

Brief centrifugation is performed to pull the Chelex® 100 resin and cellular debris to the bottom of the tube. The supernatant is used for DNA analysis. Carrying the Chelex® 100 resin over into the DNA amplification solutions should be avoided because the resin chelates magnesium, which is a necessary cofactor for DNA polymerases used for amplification.

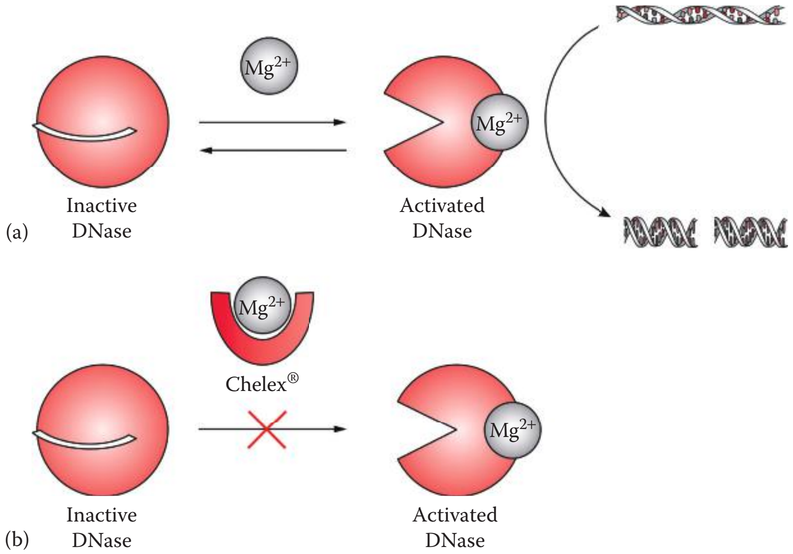

Figure 5.9 Utilizing Chelex® for preparing DNA samples. (a) Cations such as Mg^{2+} are required for the activity of endogenous DNase, which degrades DNA. (b) Chelex® prevents DNA degradation from endogenous DNase by sequestering the Mg^{2+}. (© Richard C. Li.)

This method is simple and rapid and uses only a single tube for extraction, thus reducing the risks of contamination and sample mix-ups. However, the heating step of this method disrupts and denatures proteins and also affects the chromosomal DNA. The resulting DNA extracted from the solution is fragmented single-stranded DNA. Thus, the DNA extracted is not suitable for RFLP analysis because RFLP requires double-stranded DNA samples. The DNA obtained by lysis and chelation can only be used for PCR-based DNA analysis.

5.2.3 Silica-Based Extraction

The method of adsorbing DNA molecules to solid silica surfaces is used for extraction (Figures 5.10 and 5.11). The method is based on the phenomenon that DNA is reversibly adsorbed to silica—silicon dioxide (SiO_2)—in the presence of high concentrations of chaotropic salts. Chaotropic salts can disrupt hydrogen bonding, affecting the three-dimensional structures of macromolecules. These salts are used to denature proteins. Additionally, chaotropic salts can facilitate the adsorption of DNA to silica.

In aqueous solutions, the hydration shells of nucleic acids shield the negative charges of phosphates at the phosphodiester backbone of nucleic acids. As a result, nucleic acids are usually hydrophilic in aqueous solutions. In the presence of chaotropic salts, nucleic acids become hydrophobic. The dehydrating effect caused by chaotropic salts allows the phosphate residues to become available for adsorption to the silica surface. Common chaotropic salts utilized for DNA extraction include guanidinium salts such as guanidinium thiocyanate (GuSCN) and guanidinium hydrochloride (GuHCl). GuSCN is a more potent chaotropic salt and also facilitates cell lysis and DNA adsorption. This technique usually includes the following steps:

5.2.3.1 Cell Lysis and Protein Digestion

This is carried out by proteinase K digestion. The cell membranes are broken open, and DNA is released.

5.2.3.2 DNA Adsorption onto Silica

This step utilizes silica as the stationary phase in a membrane configuration to which the DNA in the cell lysate binds. Adsorption of the DNA to the silica occurs in the presence of high concentrations of chaotropic agents (some protocols adjust pH conditions to enhance adsorption). Under these conditions, cellular materials and other contaminants that can inhibit DNA amplification reactions are not retained on the silica membrane. The adsorbed DNA is largely double stranded.

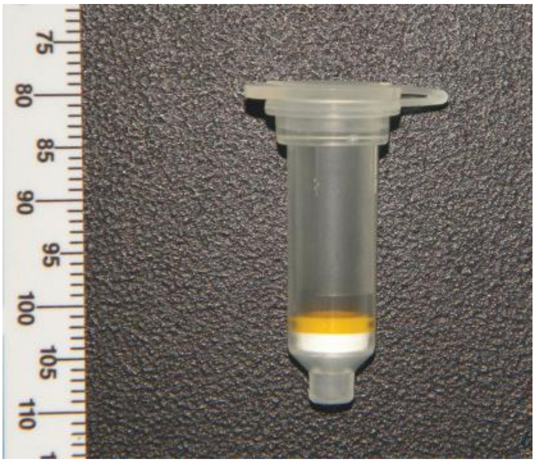

Figure 5.10 A silica-membrane spin column. (© Richard C. Li.)

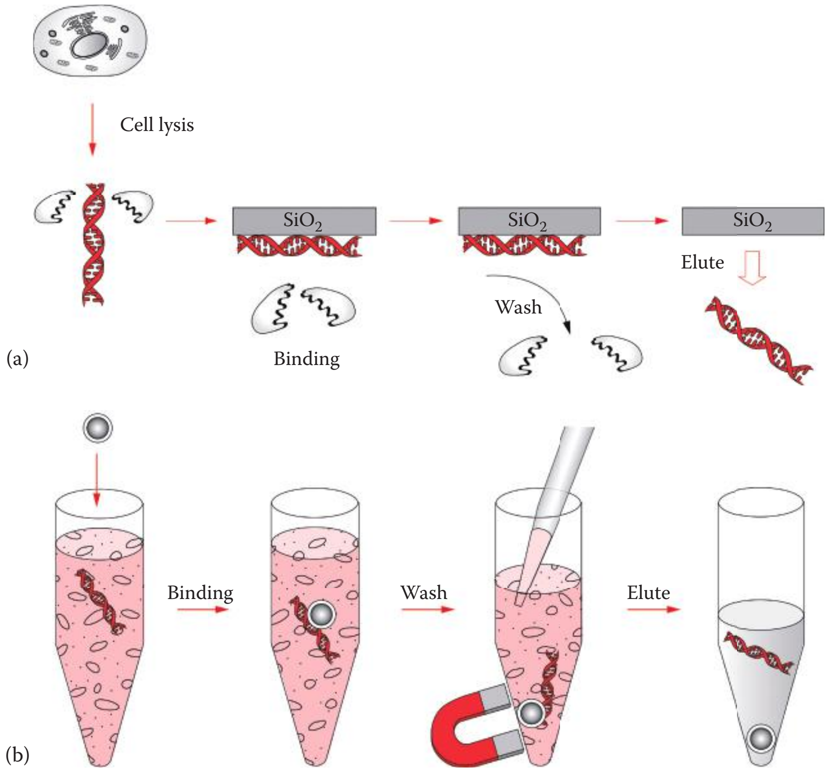

Figure 5.11 Silica-based DNA extractions. (a) Cells are lysed in the presence of proteinase. The DNA then binds to the silica matrix. A washing step removes unbound cellular materials and salts from the matrix. The purified DNA is then eluted for use in downstream applications. (b) Using silica-coated paramagnetic particles for DNA extraction. The particle is added into a lysate, and then DNA binds to the silica surface of the particle. The particle is then captured by a magnetic field. Afterwards, the supernatant is removed, and cellular materials are washed away. DNA elution follows. See Section 5.2.3. (© Richard C. Li.)

5.2.3.3 Washing

This step removes chaotropic agents and other contaminants. An ethanol-based wash solution is used. This wash solution does not remove DNA from the silica. The chaotropic agents and contaminants that are present in the solution can be removed using the ethanol-based wash solution.

5.2.3.4 Elution of DNA

The adsorbed DNA can be eluted by rehydration with aqueous low-salt solutions. The eluted DNA is double stranded and can be used for a wide variety of applications.

Silica-based extraction methods yield high-quality DNA. Silica membrane devices can also be adapted for automation; for example, this can be done by using 96-well silica membrane plates and a variety of robotic platforms. Another type of device utilizes silica-coated paramagnetic particles that adsorb DNA in the solution (Figure 5.11b). A magnet is used for particle capture instead of centrifugation or vacuum filtration. The magnetic particles can be resuspended during the wash steps, and the solution containing contaminants and cellular materials is then discarded. DNA is eluted after washing. This device can also be adapted to automated, high-throughput methods.

Over the years, high-throughput silica-based procedures have been developed to process large numbers of samples in parallel. Some of these methods are adapted for automated DNA extraction platforms (Figures 5.12 and 5.13).

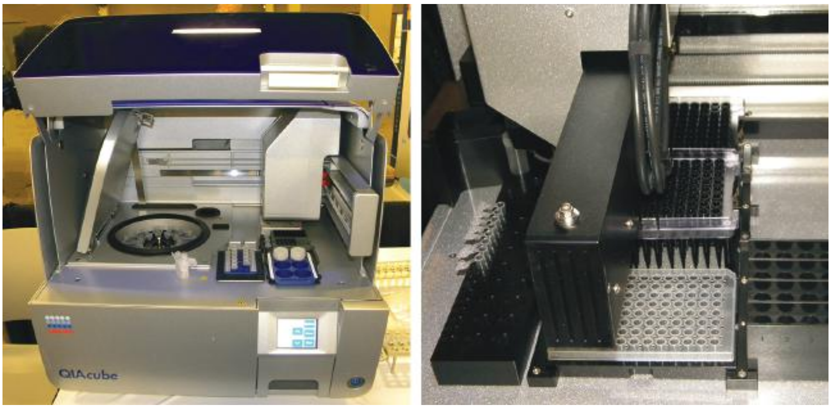

Figure 5.12 Automated bench-top DNA purification systems that enable the isolation of genomic DNA from a wide variety of forensic samples. A low-throughput sample-preparation system that can process up to 12 samples (left) and a low-to-medium throughput sample-preparation system (right). (© Richard C. Li.)

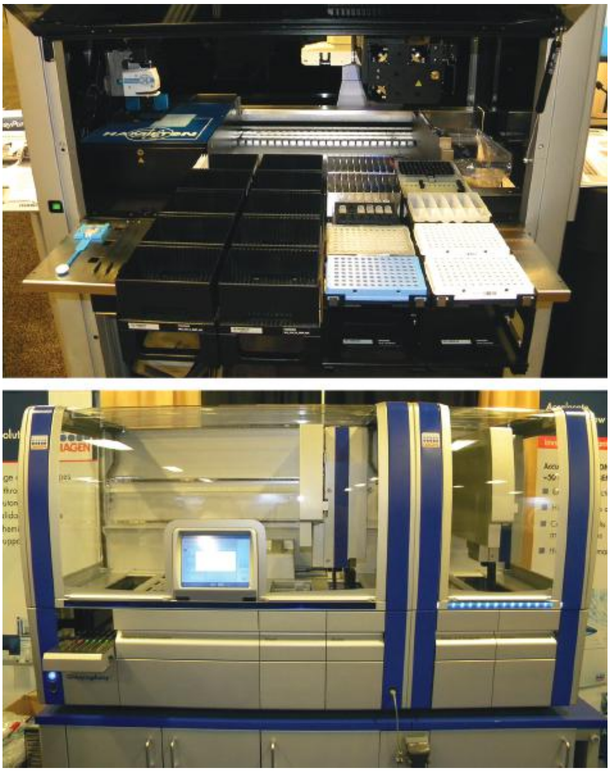

Figure 5.13 Integrated platforms enabling automated DNA sample preparation and liquid handling for subsequent assay setup. These systems utilize silica-coated bead chemistries and can achieve moderate to high throughput of processing (1–96 samples) with bar coding for sample tracking. A Hamilton (top) and a Qiagen system (bottom). (© Richard C. Li.)

5.2.4 Differential Extraction

This method is very useful for the extraction of DNA from biological evidence derived from sexual assault cases, such as vaginal swabs and bodily fluid stains. These types of evidence often contain mixtures of spermatozoa from a male contributor and nonsperm cells such as epithelial cells from a female victim. Mixtures of individual DNA profiles can complicate data interpretation.

This method selectively lyses the nonsperm and spermatozoa in separate steps based on the differences in cell-membrane properties of spermatozoa and other types of cells. Thus, the DNA from spermatozoa and nonsperm cell fractions can be sequentially isolated.

First, the differential extraction procedure involves preferentially lysing the nonsperm cells with proteolytic degradation using proteinase. Sperm plasma membrane contains proteins cross-linked by disulfide bonds. The membrane exhibits a much higher mechanical stability than nonsperm cells and is thus resistant to proteolytic degradation. The nonsperm DNA is released into the supernatant and the liquid containing it (the nonsperm fraction) is extracted, yielding a fraction that predominantly contains DNA from nonsperm cells.

To lyse the sperm cells, it is necessary to cleave the disulfide bonds in addition to proteolytic digestion. The application of DTT, a reducing agent, is an approach that can be used for cleavage. In the presence of DTT and proteinase K, the sperm plasma membrane is then lysed (Figure 5.14). Subsequently, DNA from the sperm cells can be extracted (Figure 5.15).

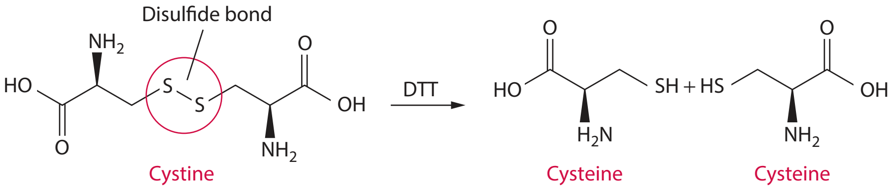

Figure 5.14 DTT reaction. The breaking of disulfide bonds in cystine residues is carried out by adding a reducing agent such as DTT.

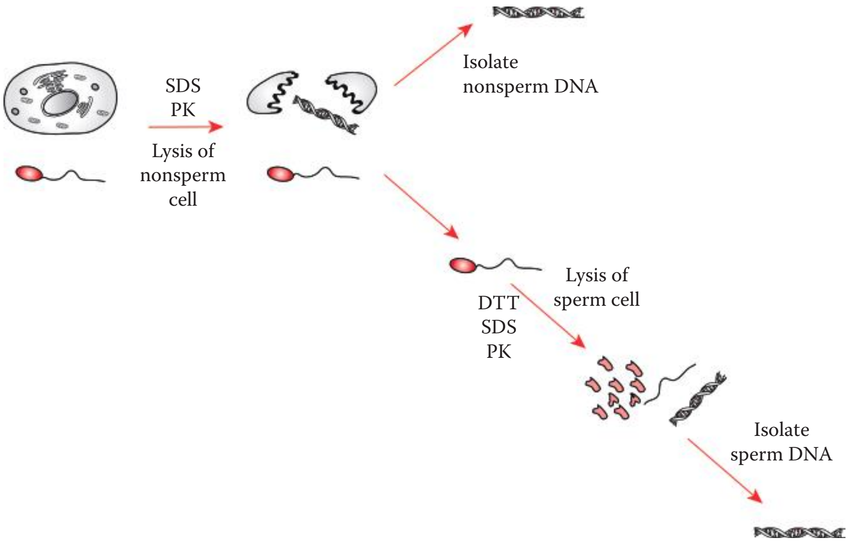

Figure 5.15 Differential extraction process used to separate sperm cells from nonsperm cells. Nonsperm cells are lysed in the presence of SDS and proteinase K (PK); sperm cells are resistant to such conditions. The nonsperm cell DNA is extracted. The sperm cells are then lysed separately in the presence of SDS and proteinase K plus DTT to extract the sperm DNA. (© Richard C. Li.)

This process isolates sperm and nonsperm cell DNA separately for obtaining DNA profiles from male and female contributors, respectively. However, the non-sperm-cell DNA and sperm-cell DNA may not be completely separated from one another; this can happen, for example, if the sperm cells have already lysed due to poor sample conditions. Some sperm DNA may be present in the non-sperm-cell fraction. Additionally, if a mixture has an abundance of nonsperm cells and fewer sperm cells, non-sperm-cell DNA may be detected in the sperm fraction. Thus, new methods that can overcome these problems are highly desired.

5.3 Essential Features of RNA

Like DNA, *ribonucleic acid* (RNA) is a linear molecule containing four types of nucleotides linked by phosphodiester bonds (Figures 5.16 and 5.17). However, certain properties of RNA differ from those of DNA. Unlike DNA containing deoxyribose (see Figure 5.18), the sugar residue

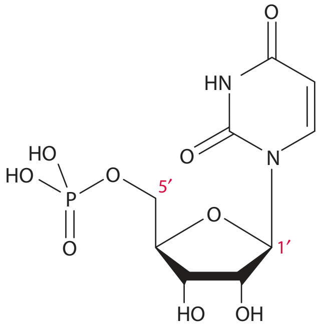

Figure 5.16 The structure of ribonucleotide. Each ribonucleotide in a RNA polymer consists of three components: a ribose, a nitrogenous base (uracil is shown as an example), and a phosphate group. The 1′ carbon of the ribose is attached to the nitrogen of the nitrogenous base. A phosphate group is attached to the 5′ carbon of the ribose.

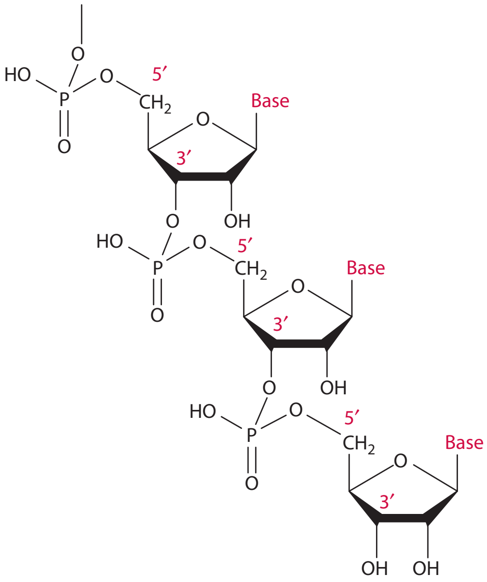

Figure 5.17 Chemical structure of RNA (polyribonucleotide) chain.

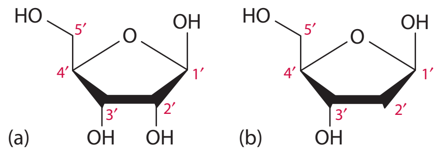

Figure 5.18 Comparison of (a) ribose and (b) deoxyribose. The hydroxyl group attached to the 2′ carbon of ribose is replaced by a hydrogen group in deoxyribose.

of RNA is a ribose, which has a hydroxyl (OH) group at the 2′ carbon position. Therefore, RNA has a lower pKa, which means that it is more acidic than DNA. RNA contains uracil (U) in the place of thymine in DNA (Figure 5.19). RNA is typically found in cells as a single-stranded molecule, while DNA is double stranded. Polyribonucleotides can form complementary helices with DNA strands by base pairing (with the exception of uracil pairs with adenine) (see Figure 5.20).

Total RNA contains all the RNA of the cells, including RNAs involved in protein synthesis and posttranscriptional modification, as well as regulatory RNAs. The RNA that carries codes from a DNA template is called messenger RNA (mRNA) and usually contains a cap and polyadenine tail at the 5′ and 3′ ends of the molecule, respectively (Figure 5.21).

The stability of RNA is an issue for mRNA-based forensic analysis. Single-stranded RNA is chemically less stable than DNA. The 2′ OH group of RNA can react with its phosphodiester backbone, potentially causing the nonenzymatic hydrolysis of an RNA molecule. Several factors, such as moisture, UV light, high temperature, and extreme pH, can facilitate the nonenzymatic

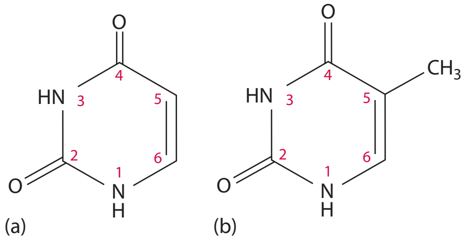

Figure 5.19 Comparison of (a) uracil and (b) thymine. Thymine has a methyl group attached at the 5-carbon position.

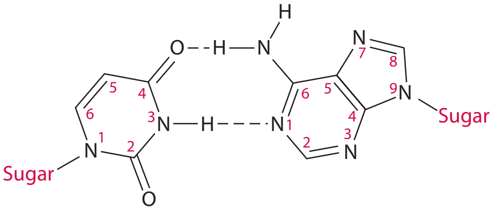

Figure 5.20 Base pairing of uracil (left) and adenine (right).

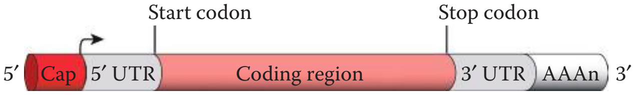

Figure 5.21 Structure of mRNA. Eukaryotic mRNA is modified at the 5′ end (cap) and 3′ end (polyadenine tail) needed for protein synthesis. Start and stop codons are required for the initiation and termination of protein synthesis. (© Richard C. Li.)

hydrolysis of RNA phosphodiesters, which leads to the degradation of RNA. Moreover, endogenous *ribonucleases* (RNases), present within cells, represent the major factor causing RNA degradation. Environmental microorganisms are also a common source of RNase contamination. RNases are very stable and usually do not require cofactors to carry out enzymatic reactions. Small amounts of RNases are sufficient to degrade RNA. RNA can be protected by adding RNase inhibitors such as diethylpyrocarbonate. Additionally, using RNase-free laboratory supplies and reagents and wearing disposable gloves while handling samples and reagents can reduce the risks of RNase contamination.

5.4 Methods of RNA Extraction

5.4.1 RNA–DNA Coextraction

For most forensic applications, total RNA is usually isolated. The procedure for isolating total RNA is simpler than that for mRNA. The total RNA isolated usually contains a sufficient amount of mRNA for subsequent reverse transcriptase PCR (RT-PCR; see Chapter 7). However, if the quantity of a target mRNA is very low, the procedures specifically for isolating mRNA should be used. There are a variety of protocols for isolating total RNA. However, RNA–DNA coextraction methods allow for the simultaneous extraction of high-quality DNA and RNA for forensic DNA analysis and bodily-fluid identification (Chapter 11), respectively, without the consumption of additional samples when extracted separately.

Biological samples are first lysed in lysis buffer. The lysis buffer usually contains chaotropic salts that can facilitate the lysis of cells and denature proteins, thus inactivating endogenous RNases to protect RNA. The lysate is then passed through a silica membrane column. A high-salt environment in the lysate allows selective binding of silica to genomic DNA over RNA. The column is washed, and purified DNA is then eluted.

The extraction of RNA is achieved by utilizing a high concentration of chaotropic salts that are already present in the lysate, along with ethanol, in order to decrease the hydrophilic property of RNA and increase its affinity for silica. After ethanol is added to the flow-through that passed the first column, the sample is then applied to a second silica column, where total RNA molecules, longer than 200 nucleotides, bind to the silica membrane. RNA including mRNA is then eluted (Figure 5.22). This method does not isolate small RNAs such as miRNA, rRNA, and tRNA, which comprise 15%–20% of total RNA.

5.4.2 miRNA Extraction

miRNAs (Chapter 4) are low-molecular-weight RNAs ranging between 15 and 30 nucleotides in length. Conventional methods routinely used for extracting total RNA do not effectively recover small RNAs; thus, they are not suitable for isolating miRNA. One approach to extracting small RNA molecules including miRNA involves two steps: organic-solvent extraction to isolate total RNA and solid-phase extraction to enrich small RNA.

To begin the process, tissues are disrupted and cells are lysed. The lysis reagents also inactivate RNases to protect RNA. The lysate is then extracted with an organic solvent, such as phenol and chloroform, with a high concentration of chaotropic salts such as GuSCN. The partitioning of DNA and RNA between the organic phase and the aqueous phase is determined by the pH during the organic extraction. At an acidic pH, DNA partitions to the organic phase. The pKa of RNA is usually lower than DNA; thus, RNA remains hydrophilic and is retained in the aqueous phase under this condition. This step removes most of the DNA, proteins, and other cellular components from the lysate into the organic phase.

The second step, silica-based extraction, further purifies and enriches small RNAs. It is achieved by utilizing the high concentration of chaotropic salt that is already present in the lysate, along with ethanol. The size-fractioning of RNAs is achieved using different ethanol

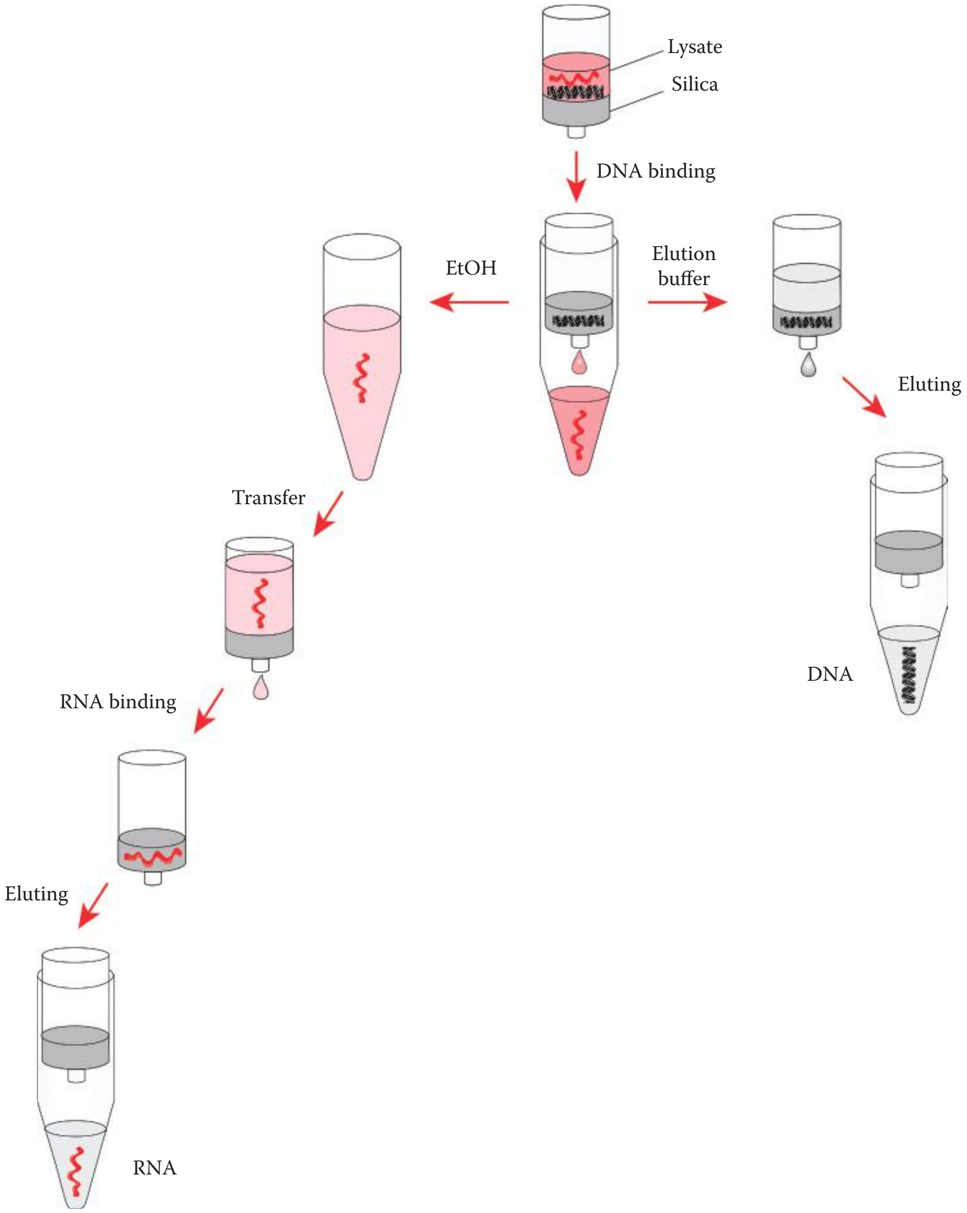
Lysate
Silica
DNA binding
EtOH
Elution buffer
Eluting
Transfer
DNA
RNA binding
Eluting
RNA

Figure 5.22 RNA and DNA coextraction using a silica-based method. See Section 5.4.1. (© Richard C. Li.)

concentrations. At a low ethanol concentration, the solvent-extracted lysate is passed through a silica membrane filter. Large RNAs are retained on the silica membrane, while small RNAs are not and are collected in the filtrate. The filtrate is passed through a second silica membrane filter. At a high ethanol concentration, the small RNAs are retained on the silica membrane. The small RNAs are then eluted in a low ionic strength solution (Figure 5.23). RNA less than 200 nucleotides, including miRNA, can be obtained.

After RNA is extracted, the small amounts of RNA can be quantified using a quantitative RT-PCR (Chapter 7) or fluorescent intercalating dye assay (Chapter 6). Purified RNA can be

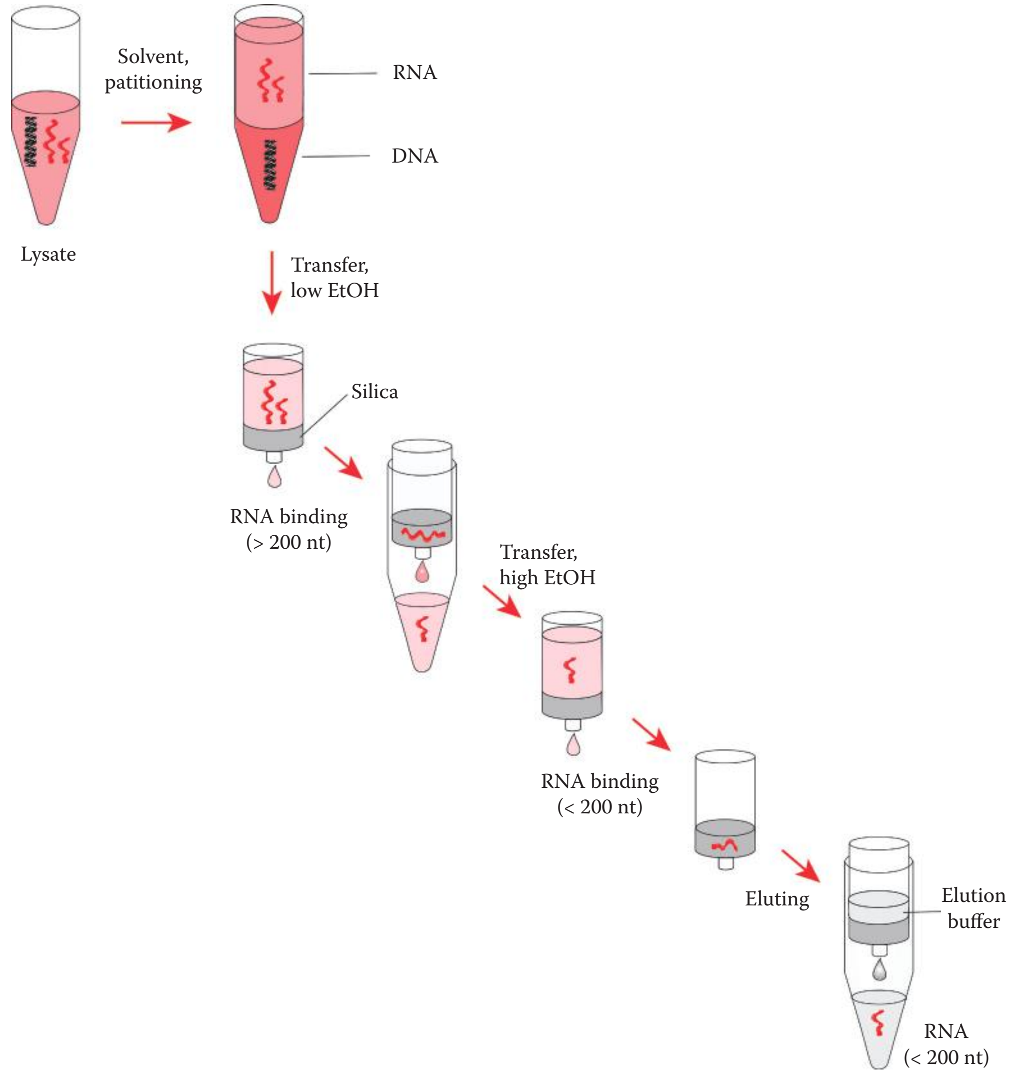
Solvent,
patitioning
RNA
DNA
Lysate
Transfer,
low EtOH
Silica
RNA binding
(> 200 nt)
Transfer,
high EtOH
RNA binding
(< 200 nt)
Eluting
Elution
buffer
RNA
(< 200 nt)

Figure 5.23 Schematic illustration of an miRNA extraction method. See Section 5.4.2. (© Richard C. Li.)

stored at −80°C in RNase-free water or Tris-EDTA buffer (pH 7). Single-use aliquots are preferred when possible to avoid multiple freeze–thaw cycles.

The integrity of extracted RNA is a great concern for RNA-based forensic analysis. RNA quality is traditionally assessed using the 28S:18S rRNA ratio. There are some conditions where this method has been shown to be inconsistent; that is, in which the 28S:18S rRNA ratio does not reflect the true level of RNA degradation. Recently, the *RNA integrity number* (RIN) has provided an effective method for determining RNA quality. RIN is a software algorithm that takes the electrophoretic RNA measurements into account in order to assign integrity values to RNA samples. RIN ranges from 1 to 10, with 10 being the most intact. Thus, the RIN method facilitates the assessment of the integrity of RNA samples. As reflected by RIN, RNA degradation has a negative influence on the reproducibility of the results of RNA-based forensic analysis.

Bibliography

Alvarez, M., J. Juusola, and J. Ballantyne, An mRNA and DNA co-isolation method for forensic casework samples. *Anal Biochem*, 2004, **335**(2): 289–298.

Amory, S., et al., Automatable full demineralization DNA extraction procedure from degraded skeletal remains. *Forensic Sci Int Genet*, 2012, **6**(3): 398–406.

Ananian, V., et al., Tumoural specimens for forensic purposes: Comparison of genetic alterations in frozen and formalin-fixed paraffin-embedded tissues. *Int J Legal Med*, 2011, **125**(3): 327–332.

Anzai-Kanto, E., et al., DNA extraction from human saliva deposited on skin and its use in forensic identification procedures. *Pesqui Odontol Bras*, 2005, **19**(3): 216–222.

Archer, E., et al., Validation of a dual cycle ethylene oxide treatment technique to remove DNA from consumables used in forensic laboratories. *Forensic Sci Int Genet*, 2010, **4**(4): 239–243.

Baker, L.E., W.F. McCormick, and K.J. Matteson, A silica-based mitochondrial DNA extraction method applied to forensic hair shafts and teeth. *J Forensic Sci*, 2001, **46**(1): 126–130.

Benoit, J.N., et al., An alternative procedure for extraction of DNA from ancient and weathered bone fragments. *Med Sci Law*, 2013, **53**(2): 100–106.

Bowden, A., R. Fleming, and S. Harbison, A method for DNA and RNA co-extraction for use on forensic samples using the Promega DNA IQ system. *Forensic Sci Int Genet*, 2011, **5**(1): 64–68.

Brevnov, M.G., et al., Developmental validation of the PrepFiler Forensic DNA Extraction Kit for extraction of genomic DNA from biological samples. *J Forensic Sci*, 2009, **54**(3): 599–607.

Castella, V., et al., Forensic evaluation of the QIAshredder/QIAamp DNA extraction procedure. *Forensic Sci Int*, 2006, **156**(1): 70–73.

Cattaneo, C., et al., A simple method for extracting DNA from old skeletal material. *Forensic Sci Int*, 1995, **74**(3): 167–174.

Crouse, C.A., J.D. Ban, and J.K. D'Alessio, Extraction of DNA from forensic-type sexual assault specimens using simple, rapid sonication procedures. *Biotechniques*, 1993, **15**(4): 641–644, 646, 648.

Davis, C.P., et al., Extraction platform evaluations: A comparison of AutoMate Express, EZ1(R) Advanced XL, and Maxwell(R) 16 Bench-top DNA extraction systems. *Leg Med (Tokyo)*, 2012, **14**(1): 36–39.

Di Nunno, N., et al., DNA extraction: An anthropologic aspect of bone remains from sixth- to seventh-century ad bone remains. *Am J Forensic Med Pathol*, 2007, **28**(4): 333–341.

Diegoli, T.M., et al., An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA. *Forensic Sci Int Genet*, 2012, **6**(4): 498–503.

Dieltjes, P., et al., A sensitive method to extract DNA from biological traces present on ammunition for the purpose of genetic profiling. *Int J Legal Med*, 2011, **125**(4): 597–602.

Duarte, G.R., et al., Characterization of dynamic solid phase DNA extraction from blood with magnetically controlled silica beads. *Analyst*, 2010, **135**(3): 531–537.

Duval, K., et al., Optimized manual and automated recovery of amplifiable DNA from tissues preserved in buffered formalin and alcohol-based fixative. *Forensic Sci Int Genet*, 2010, **4**(2): 80–88.

Edson, S.M. and A.F. Christensen, Field contamination of skeletonized human remains with exogenous DNA. *J Forensic Sci*, 2013, **58**(1): 206–209.

Elliott, K., et al., Use of laser microdissection greatly improves the recovery of DNA from sperm on microscope slides. *Forensic Sci Int*, 2003, **137**(1): 28–36.

Eminovic, I., et al., A simple method of DNA extraction in solving difficult criminal cases. *Med Arh*, 2005, **59**(1): 57–58.

Farrugia, A., C. Keyser, and B. Ludes, Efficiency evaluation of a DNA extraction and purification protocol on archival formalin-fixed and paraffin-embedded tissue. *Forensic Sci Int*, 2010, **194**(1–3): e25–e28.

Fattorini, P., et al., CE analysis and molecular characterisation of depurinated DNA samples. *Electrophoresis*, 2011, **32**(21): 3042–3052.

Fattorini, P., et al., Assessment of DNA damage by micellar electrokinetic chromatography. *Methods Mol Biol*, 2013, **984**: 341–351.

Fracasso, T., et al., Ultrasound-accelerated formalin fixation improves the preservation of nucleic acids extraction in histological sections. *Int J Legal Med*, 2009, **123**(6): 521–525.

Fregeau, C.J. and A. De Moors, Competition for DNA binding sites using Promega DNA IQ paramagnetic beads. *Forensic Sci Int Genet*, 2012, **6**(5): 511–522.

Fregeau, C.J., C.M. Lett, and R.M. Fourney, Validation of a DNA IQ-based extraction method for TECAN robotic liquid handling workstations for processing casework. *Forensic Sci Int Genet*, 2010, **4**(5): 292–304.

Fregeau, C.J., et al., Automated processing of forensic casework samples using robotic workstations equipped with nondisposable tips: Contamination prevention. *J Forensic Sci*, 2008, **53**(3): 632–651.

Fridez, F. and R. Coquoz, PCR DNA typing of stamps: Evaluation of the DNA extraction. *Forensic Sci Int*, 1996, **78**(2): 103–110.

Frippiat, C., et al., Evaluation of novel forensic DNA storage methodologies. *Forensic Sci Int Genet*, 2011, **5**(5): 386–392.

Fujita, Y. and S. Kubo, Application of FTA technology to extraction of sperm DNA from mixed body fluids containing semen. *Leg Med (Tokyo)*, 2006, **8**(1): 43–47.

Garvin, A.M. and A. Fritsch, Purifying and concentrating genomic DNA from mock forensic samples using Millipore Amicon filters. *J Forensic Sci*, 2013, **58**(Suppl 1): S173–S175.

Garvin, A.M., et al., DNA preparation from sexual assault cases by selective degradation of contaminating DNA from the victim. *J Forensic Sci*, 2009, **54**(6): 1297–1303.

Ginestra, E., et al., DNA extraction from blood determination membrane card test. *Forensic Sci Int*, 2004, **146**(Suppl): S145–S146.

Greenspoon, S.A., et al., QIAamp spin columns as a method of DNA isolation for forensic casework. *J Forensic Sci*, 1998, **43**(5): 1024–1030.

Greenspoon, S.A., et al., Application of the BioMek 2000 Laboratory Automation Workstation and the DNA IQ System to the extraction of forensic casework samples. *J Forensic Sci*, 2004, **49**(1): 29–39.

Grubb, J.C., et al., Implementation and validation of the Teleshake unit for DNA IQ robotic extraction and development of a large volume DNA IQ method. *J Forensic Sci*, 2010, **55**(3): 706–714.

Hagan, K.A., et al., Chitosan-coated silica as a solid phase for RNA purification in a microfluidic device. *Anal Chem*, 2009, **81**(13): 5249–5256.

Hanselle, T., et al., Isolation of genomic DNA from buccal swabs for forensic analysis, using fully automated silica-membrane purification technology. *Leg Med (Tokyo)*, 2003, **5**(Suppl 1): S145–S149.

Hedman, J., et al., A fast analysis system for forensic DNA reference samples. *Forensic Sci Int Genet*, 2008, **2**(3): 184–189.

Heinrich, M., et al., Successful RNA extraction from various human postmortem tissues. *Int J Legal Med*, 2007, **121**(2): 136–142.

Hennekens, C.M., et al., The effects of differential extraction conditions on the premature lysis of spermatozoa. *J Forensic Sci*, 2013, **58**(3): 744–752.

Ho, T.T. and R. Roy, Generating DNA profiles from immunochromatographic cards using LCN methodology. *Forensic Sci Int Genet*, 2011, **5**(3): 210–215.

Hochmeister, M.N., et al., PCR-based typing of DNA extracted from cigarette butts. *Int J Legal Med*, 1991, **104**(4): 229–233.

Hoff-Olsen, P., et al., Extraction of DNA from decomposed human tissue. An evaluation of five extraction methods for short tandem repeat typing. *Forensic Sci Int*, 1999, **105**(3): 171–183.

Howlett, S.E., et al., Evaluation of DNAstable for DNA storage at ambient temperature. *Forensic Sci Int Genet*, 2014, **8**(1): 170–178.

Hudlow, W.R. and M.R. Buoncristiani, Development of a rapid, 96-well alkaline based differential DNA extraction method for sexual assault evidence. *Forensic Sci Int Genet*, 2012, **6**(1): 1–16.

Hudlow, W.R., et al., The NucleoSpin(R) DNA Clean-up XS kit for the concentration and purification of genomic DNA extracts: An alternative to microdialysis filtration. *Forensic Sci Int Genet*, 2011, **5**(3): 226–230.

Hulme, P., J. Lewis, and G. Davidson, Sperm elution: An improved two phase recovery method for sexual assault samples. *Sci Justice*, 2013, **53**(1): 28–33.

Iwasa, M., et al., Y-chromosomal short tandem repeats haplotyping from vaginal swabs using a chelating resin-based DNA extraction method and a dual-round polymerase chain reaction. *Am J Forensic Med Pathol*, 2003, **24**(3): 303–305.

Jakubowska, J., A. Maciejewska, and R. Pawlowski, Comparison of three methods of DNA extraction from human bones with different degrees of degradation. *Int J Legal Med*, 2012, **126**(1): 173–178.

Karija Vlahovic, M. and M. Kubat, DNA extraction method from bones using Maxwell(R) 16. *Leg Med (Tokyo)*, 2012, **14**(5): 272–275.

Kim, K., et al., Technical note: Improved ancient DNA purification for PCR using ion-exchange columns. *Am J Phys Anthropol*, 2008, **136**(1): 114–121.

Kochl, S., H. Niederstatter, and W. Parson, DNA extraction and quantitation of forensic samples using the phenol–chloroform method and real-time PCR. *Methods Mol Biol*, 2005, **297**: 13–30.

Kopka, J., et al., New optimized DNA extraction protocol for fingerprints deposited on a special self-adhesive security seal and other latent samples used for human identification. *J Forensic Sci*, 2011, **56**(5): 1235–1240.

Lee, E.J., et al., The effects of different maceration techniques on nuclear DNA amplification using human bone. *J Forensic Sci*, 2010, **55**(4): 1032–1038.

Lee, H.Y., et al., Simple and highly effective DNA extraction methods from old skeletal remains using silica columns. *Forensic Sci Int Genet*, 2010, **4**(5): 275–280.

Lee, S.B., et al., Assessing a novel room temperature DNA storage medium for forensic biological samples. *Forensic Sci Int Genet*, 2012, **6**(1): 31–40.

Li, R. and L. Liriano, A bone sample cleaning method using trypsin for the isolation of DNA. *Leg Med (Tokyo)*, 2011, **13**(6): 304–308.

Li, R. and S. Klempner, The effect of an enzymatic bone processing method on short tandem repeat profiling of challenged bone specimens. *Leg Med (Tokyo)*, 2013, **15**(4): 171–176.

Li, R., et al., Developing a simple method to process bone samples prior to DNA isolation. *Leg Med (Tokyo)*, 2009, **11**(2): 76–79.

Li, C.X., et al., New cell separation technique for the isolation and analysis of cells from biological mixtures in forensic caseworks. *Croat Med J*, 2011, **52**(3): 293–298.

Liu, J.Y., et al., AutoMate Express forensic DNA extraction system for the extraction of genomic DNA from biological samples. *J Forensic Sci*, 2012, **57**(4): 1022–1030.

Loreille, O.M., et al., High efficiency DNA extraction from bone by total demineralization. *Forensic Sci Int Genet*, 2007, **1**(2): 191–195.

Lounsbury, J.A., et al., An enzyme-based DNA preparation method for application to forensic biological samples and degraded stains. *Forensic Sci Int Genet*, 2012, **6**(5): 607–615.

Ma, H.W., J. Cheng, and B. Caddy, Extraction of high quality genomic DNA from microsamples of human blood. *J Forensic Sci Soc*, 1994, **34**(4): 231–235.

McMichael, G.L., et al., Comparison of DNA extraction methods from small samples of newborn screening cards suitable for retrospective perinatal viral research. *J Biomol Tech*, 2011, **22**(1): 5–9.

Meredith, M., et al., Development of a one-tube extraction and amplification method for DNA analysis of sperm and epithelial cells recovered from forensic samples by laser microdissection. *Forensic Sci Int Genet*, 2012, **6**(1): 91–96.

Montpetit, S.A., I.T. Fitch, and P.T. O'Donnell, A simple automated instrument for DNA extraction in forensic casework. *J Forensic Sci*, 2005, **50**(3): 555–563.

Nagy, M., et al., Optimization and validation of a fully automated silica-coated magnetic beads purification technology in forensics. *Forensic Sci Int*, 2005, **152**(1): 13–22.

Norris, J.V., et al., Acoustic differential extraction for forensic analysis of sexual assault evidence. *Anal Chem*, 2009, **81**(15): 6089–6095.

Parker, C., E. Hanson, and J. Ballantyne, Optimization of dried stain co-extraction methods for efficient recovery of high quality DNA and RNA for forensic analysis. *Forensic Sci Int: Genet Suppl Series*, 2011, **3**(1): e309–e310.

Phillips, K., N. McCallum, and L. Welch, A comparison of methods for forensic DNA extraction: Chelex-100(R) and the QIAGEN DNA Investigator Kit (manual and automated). *Forensic Sci Int Genet*, 2012, **6**(2): 282–285.

Piglionica, M., et al., Extraction of DNA from bones in cases where expectations for success are low. *Am J Forensic Med Pathol*, 2012, **33**(4): 322–327.

Raimann, P.E., et al., Procedures to recover DNA from pre-molar and molar teeth of decomposed cadavers with different post-mortem intervals. *Arch Oral Biol*, 2012, **57**(11): 1459–1466.

Reedy, C.R., et al., Volume reduction solid phase extraction of DNA from dilute, large-volume biological samples. *Forensic Sci Int Genet*, 2010, **4**(3): 206–212.

Rerkamnuaychoke, B., et al., Comparison of DNA extraction from blood stain and decomposed muscle in STR polymorphism analysis. *J Med Assoc Thai*, 2000, **83**(Suppl 1): S82–S88.

Reshef, A., et al., STR typing of formalin-fixed paraffin embedded (FFPE) aborted foetal tissue in criminal paternity cases. *Sci Justice*, 2011, **51**(1): 19–23.

Rothe, J., L. Roewer, and M. Nagy, Individual specific extraction of DNA from male mixtures—First evaluation studies. *Forensic Sci Int Genet*, 2011, **5**(2): 117–121.

Rucinski, C., et al., Comparison of two methods for isolating DNA from human skeletal remains for STR analysis. *J Forensic Sci*, 2012, **57**(3): 706–712.

Scherczinger, C.A., et al., DNA extraction from liquid blood using QIAamp. *J Forensic Sci*, 1997, **42**(5): 893–896.

Seo, S.B., et al., Technical note: Efficiency of total demineralization and ion-exchange column for DNA extraction from bone. *Am J Phys Anthropol*, 2010, **141**(1): 158–162.

Shalhoub, R., et al., The recovery of latent fingermarks and DNA using a silicone-based casting material. *Forensic Sci Int*, 2008, **178**(2–3): 199–203.

Shan, Z., et al., PCR-ready human DNA extraction from urine samples using magnetic nanoparticles. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2012, **881–882**: 63–68.

Shaw, K., et al., Comparison of the effects of sterilisation techniques on subsequent DNA profiling. *Int J Legal Med*, 2008, **122**(1): 29–33.

Sinclair, K. and V.M. McKechnie, DNA extraction from stamps and envelope flaps using QIAamp and QIAshredder. *J Forensic Sci*, 2000, **45**(1): 229–230.

Stangegaard, M., et al., Automated extraction of DNA from blood and PCR setup using a Tecan Freedom EVO liquid handler for forensic genetic STR typing of reference samples. *J Lab Autom*, 2011, **16**(2): 134–140.

Stangegaard, M., et al., Biomek 3000: The workhorse in an automated accredited forensic genetic laboratory. *J Lab Autom*, 2012, **17**(5): 378–386.

Stangegaard, M., et al., Automated extraction of DNA from biological stains on fabric from crime cases. A comparison of a manual and three automated methods. *Forensic Sci Int Genet*, 2013, **7**(3): 384–388.

Stangegaard, M., et al., Evaluation of four automated protocols for extraction of DNA from FTA cards. *J Lab Autom*, 2013, **18**(5): 404–410.

Startari, L., et al., Comparison of extractable DNA from bone following six-month exposure to outdoor conditions, garden loam, mold contamination or room storage. *Med Sci Law*, 2013, **53**(1): 29–32.

Sweet, D., et al., Increasing DNA extraction yield from saliva stains with a modified Chelex method. *Forensic Sci Int*, 1996, **83**(3): 167–177.

Tack, L.C., M. Thomas, and K. Reich, Automated forensic DNA purification optimized for FTA card punches and identifiler STR-based PCR analysis. *Clin Lab Med*, 2007, **27**(1): 183–191.

Tsuchimochi, T., et al., Chelating resin-based extraction of DNA from dental pulp and sex determination from incinerated teeth with Y-chromosomal alphoid repeat and short tandem repeats. *Am J Forensic Med Pathol*, 2002, **23**(3): 268–271.

Vandenberg, N., R.A. van Oorschot, and R.J. Mitchell, An evaluation of selected DNA extraction strategies for short tandem repeat typing. *Electrophoresis*, 1997, **18**(9): 1624–1626.

Vandewoestyne, M., et al., Evaluation of three DNA extraction protocols for forensic STR typing after laser capture microdissection. *Forensic Sci Int Genet*, 2012, **6**(2): 258–262.

von Wurmb-Schwark, N., et al., The impact of DNA contamination of bone samples in forensic case analysis and anthropological research. *Leg Med (Tokyo)*, 2008, **10**(3): 125–130.

Walsh, P.S., D.A. Metzger, and R. Higuchi, Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*, 1991, **10**(4): 506–513.

Wilkinson, D.A., et al., The fate of the chemical warfare agent during DNA extraction. *J Forensic Sci*, 2007, **52**(6): 1272–1283.

Winskog, C., et al., The use of commercial alcohol products to sterilize bones prior to DNA sampling. *Forensic Sci Med Pathol*, 2010, **6**(2): 127–129.

Witt, S., et al., Establishing a novel automated magnetic bead-based method for the extraction of DNA from a variety of forensic samples. *Forensic Sci Int Genet*, 2012, **6**(5): 539–547.

Wolfe, K.A., et al., Toward a microchip-based solid-phase extraction method for isolation of nucleic acids. *Electrophoresis*, 2002, **23**(5): 727–733.

Wolfgramm Ede, V., et al., Simplified buccal DNA extraction with FTA Elute Cards. *Forensic Sci Int Genet*, 2009, **3**(2): 125–127.

Ye, J., et al., A simple and efficient method for extracting DNA from old and burned bone. *J Forensic Sci*, 2004, **49**(4): 754–759.

Yoshida-Yamamoto, S., et al., Efficient DNA extraction from nail clippings using the protease solution from Cucumis melo. *Mol Biotechnol*, 2010, **46**(1): 41–48.

Zamir, A., Y. Cohen, and M. Azoury, DNA profiling from heroin street dose packages. *J Forensic Sci*, 2007, **52**(2): 389–392.

Zehner, R., "Foreign" DNA in tissue adherent to compact bone from tsunami victims. *Forensic Sci Int Genet*, 2007, **1**(2): 218–222.

Zhang, H., et al., Solid-phase based on-chip DNA purification through a valve-free stepwise injection of multiple reagents employing centrifugal force combined with a hydrophobic capillary barrier pressure. *Analyst*, 2013, **138**(6): 1750–1757.

Zhang, S.H., et al., Genotyping of urinary samples stored with EDTA for forensic applications. *Genet Mol Res*, 2012, **11**(3): 3007–3012.

6

DNA Quantitation

Determining the amount of DNA in a sample is essential for polymerase chain reaction (PCR)-based DNA testing (Chapter 7). Because PCR-based DNA testing is very sensitive, a narrow concentration range is required for amplification to be successful. If too much DNA template is used in PCR, the resulting artifacts may interfere with data analysis and interpretation. On the other hand, very low amounts of DNA template may result in a partial DNA profile or failure to attain a profile. Samples with poor purity may contain PCR inhibitors that lead to a failure of DNA amplification. Thus, a test that can measure the quality and quantity of the DNA template of a sample is desirable.

Additionally, forensic samples containing DNA are often mixtures that may include nonhuman DNA. For instance, microbial DNA may be present. Thus, it is necessary to use human-specific DNA quantitation methods to selectively determine the amount of human DNA present. In the United States, quality-assurance guidelines require the use of a quantitation method that estimates the amount of human nuclear DNA in crime-scene evidence samples.

In this chapter, the slot blot, intercalating dye, and quantitative PCR methods are introduced. The quantitative PCR method is the most sensitive of the three methods. It is the only method that can detect PCR inhibitors. Both the slot blot and the quantitative PCR methods can detect human and nonhuman primate DNA. Current technology cannot distinguish between human and nonhuman primate DNA.

6.1 Slot Blot Assay

Historically, the slot blot assay was used to detect human genomic DNA in a sample. Methodically, the slot blot assay works based on the following principles. Prior to the quantitation, an alkaline solution is added to the genomic DNA sample, which denatures DNA. Generating single-stranded DNA is necessary for DNA to be cross-linked onto a nitrocellulose membrane. The DNA sample is then spotted, using a slot blot device, onto a nitrocellulose membrane (Figure 6.1). The single-stranded DNA is then immobilized onto a nylon membrane. The targeted sequence is revealed by hybridization with a labeled 40-nucleotide probe complementary to a primate-specific α-satellite DNA sequence at the D17Z1 locus (Figure 6.2). In humans, the α-satellite DNA sequences are highly repetitive sequences located near the centromeres of chromosomes. These sequences are usually distinct for each chromosome.

Three detecting schemes of the slot blot assay have been developed. Initially, the D17Z1 probe was labeled with radioisotopes that could be visualized by exposing the slot blot membrane to x-ray film. The hazardous radioisotope detection method was then replaced by alkaline phosphatase-labeled and biotinylated probes. The alkaline phosphatase-labeled probe can be coupled with chemiluminescent detection (Lumi-Phos Plus kit, Lumigen, Inc.). The biotinylated probe

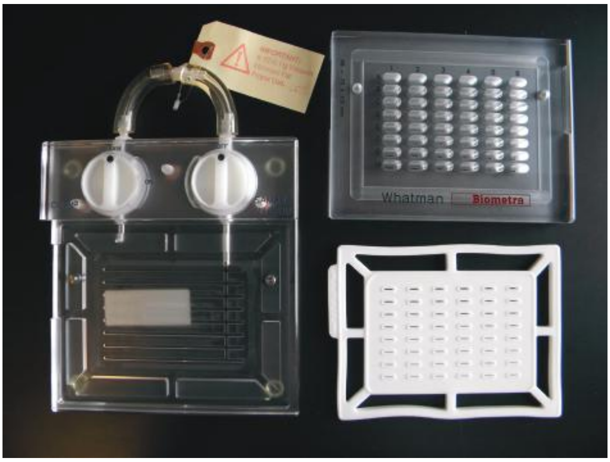

Figure 6.1 A slot blot device. (© Richard C. Li.)

(QuantiBlot Human DNA Quantitation Kit, Applied Biosystems) can be coupled with either colorimetric or chemiluminescent detection.

In colorimetric detection, the biotin moiety of the probe is bound to streptavidin. The streptavidin is conjugated with horseradish peroxidase, which catalyzes the oxidation reaction of tetramethylbenzidine (TMB), a substrate, forming a blue precipitate. With the chemiluminescent detection method, the horseradish peroxidase catalyzes the oxidation reaction of a substrate such as luminol, emitting photons that can be detected by exposure to x-ray film. The sensitivity of chemiluminescent detection is slightly higher than that of the colorimetric detection. The detection mechanisms will be discussed in Chapter 9.

The detected signal intensity is proportional to the concentration of the DNA sample in question. Quantitative measurements can be made by comparing an unknown sample to a set of standards with known DNA concentrations (Figure 6.3).

The assay typically quantifies DNA over the range of 150 pg–10 ng. However, the quantitation results are manually read, and conclusions are based on subjective judgments. Additionally, the D17Z1 sequences of humans and other primates share homology. The probe cannot distinguish between human and other primate genomic DNA. This is not a great concern because cases involving nonhuman primates are rare. Nevertheless, the probe does not cross-react with all other species. This assay has been replaced by recently developed quantitative PCR assays (see Section 6.3).

6.2 Fluorescent Intercalating Dye Assay

Small quantities of DNA can also be quantified by using a fluorescent intercalating dye method. Intercalating dyes, usually planar molecules, can slide themselves in between base pairs of DNA without breaking the DNA double helix. The Quant-iT™ PicoGreen® dsDNA reagent (Invitrogen) is a fluorescent intercalating dye that stains double-stranded DNA (dsDNA) for quantitation in a sample (Figure 6.4a). The detection limit of this method is approximately 250 pg. Intercalating dyes, not specific to human DNA, bind to all DNA molecules. Therefore, fluorescent intercalating dye assay can be utilized for the quantitation of known reference samples. For instance, DNA database samples from known sources can be quantified using this method. The assay has also been adapted for automation and is, thus, a high-throughput method.

DNA samples are simply added to a solution containing the fluorescent intercalating dye. The fluorescence, proportional to the quantities of DNA (Figure 6.4b), is measured using a standard spectrofluorometer with excitation and emission wavelengths of the light source. A standard

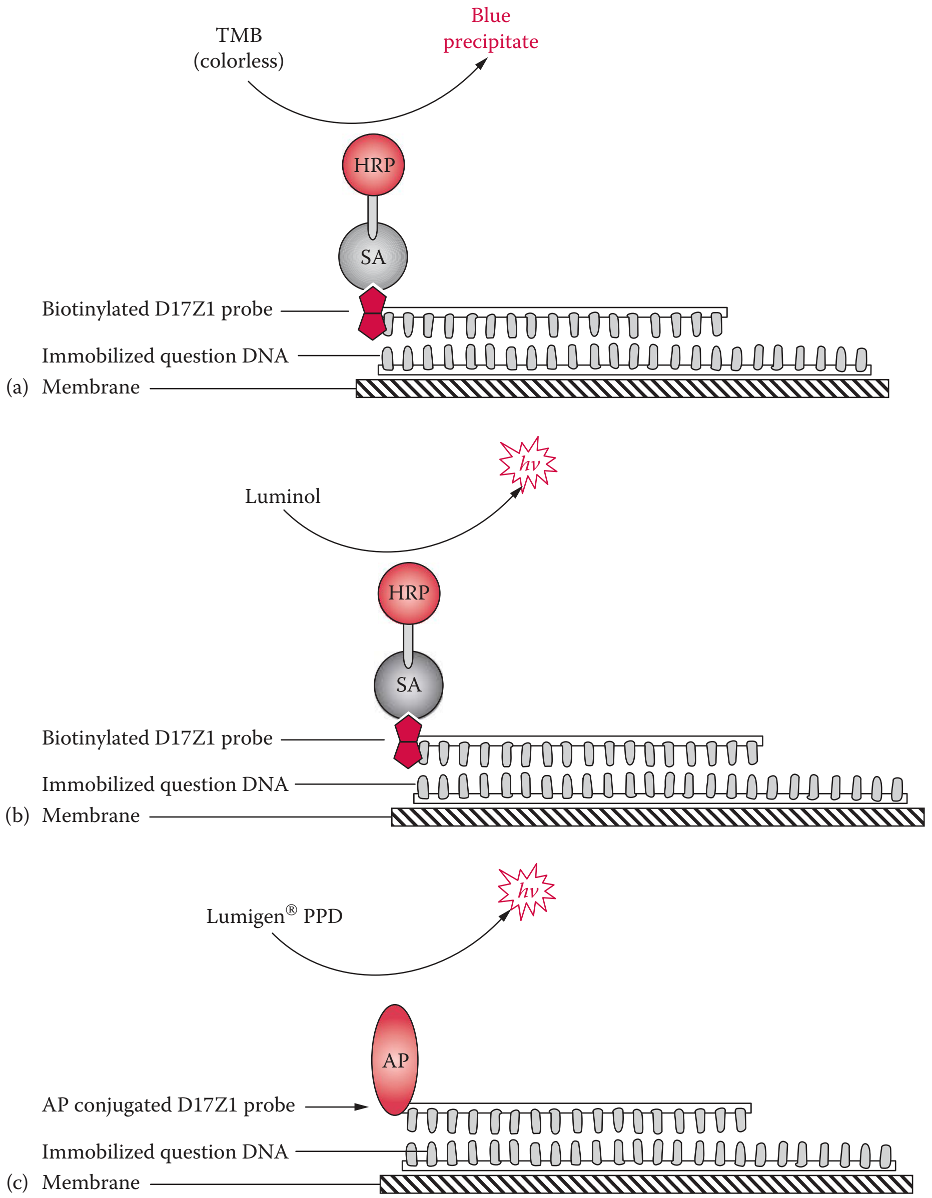
TMB
(colorless)
Blue
precipitate
HRP
SA
Biotinylated D17Z1 probe
Immobilized question DNA
(a) Membrane
Luminol
hv
HRP
SA
Biotinylated D17Z1 probe
Immobilized question DNA
(b) Membrane
Lumigen® PPD
hv
AP
AP conjugated D17Z1 probe
Immobilized question DNA
(c) Membrane

Figure 6.2 Slot blot assay. Questioned DNA is immobilized onto a solid-phase membrane then hybridized with a biotinylated D17Z1 probe. The detection of the hybridization is carried out by (a) streptavidin (SA) and horseradish peroxidase (HRP) conjugate, and a colorimetric reaction is catalyzed by HRP using tetramethylbenzidine (TMB) as a substrate; (b) SA and HRP conjugate, and a chemiluminescent reaction is catalyzed by HRP using Luminol as a substrate; (c) immobilized DNA is hybridized with an alkaline phosphatase (AP)-labeled D17Z1 probe. The detection of the hybridization is carried out by a chemiluminescent reaction catalyzed by AP using Lumigen® PPD as a substrate. (© Richard C. Li.)

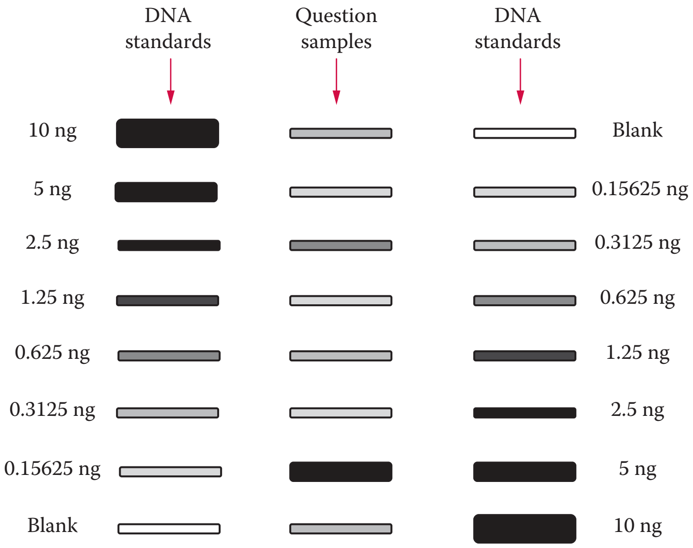
DNA standards
Question samples
DNA standards
10 ng
5 ng
2.5 ng
1.25 ng
0.625 ng
0.3125 ng
0.15625 ng
Blank
Blank
0.15625 ng
0.3125 ng
0.625 ng
1.25 ng
2.5 ng
5 ng
10 ng

Figure 6.3 Human DNA quantitation using the slot blot assay. Standards with known amounts of human DNA are applied, and unknown samples and a set of standards are compared. The quantities in the unknown samples are estimated by visual comparison to the standards. (© Richard C. Li.)

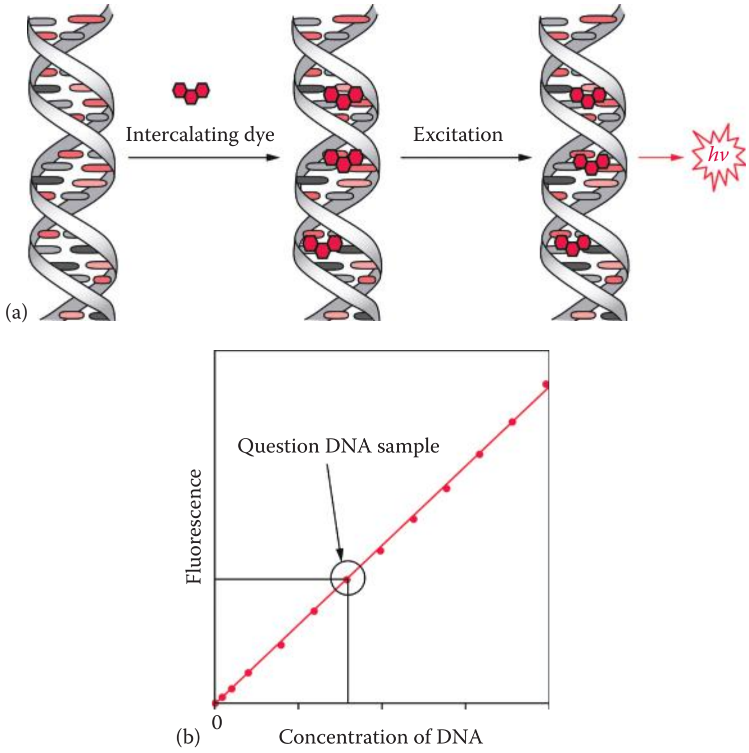
Intercalating dye
Excitation
hv
(a)
Question DNA sample
Fluorescence
0
Concentration of DNA
(b)

Figure 6.4 DNA quantitation using intercalating dye. (a) Fluorescent dye intercalates into DNA. The fluorescence can be measured upon applying an excitation light source. (b) A standard curve can be constructed using known amounts of DNA standards. The amount of questioned DNA can be determined by comparing the standard curve. (© Richard C. Li.)

curve is first created using samples containing known amounts of DNA. The assay is then performed for the unknown samples and the quantities of DNA in the samples are determined by comparing the results to the standard curve.

6.3 Quantitative PCR Assay

Based on the principle of PCR amplification (Chapter 7), the amount of PCR product amplified correlates with the initial concentration of DNA templates. Thus, the DNA concentration of a sample can be determined. There are two types of quantitative PCR methods. *End-point PCR* methods measure the quantity of amplified product at the end of PCR. Usually, the fluorescence is emitted by the dyes that intercalate into the double-stranded DNA. The quantity of amplified DNA is measured from the amount of fluorescence emitted from dyes such as SYBR (Figures 6.5 and 9.1b).

Real-time PCR methods can quantify the amplified DNA during the exponential phase of PCR (Chapter 7). The quantitation result is not affected to a significant extent by slight variations in PCR conditions. Thus, the precision of the quantitation of target sequences is improved with this method.

6.3.1 Real-Time Quantitative PCR

Real-time quantitative PCR (qPCR) was developed in the early 1990s, and it analyzes the amplification of a target sequence at each cycle of PCR. A fluorescent reporter is used to monitor the accumulation of amplified products during PCR. The fluorescence signals of the reporter molecule increase as amplified products accumulate with each cycle of PCR. qPCR is commonly used because of the following advantages:

- Better objectivity than the QuantiBlot method
- Increased sensitivity with a large dynamic range (30 pg–100 ng)
- More accurate measurements of small quantities of DNA in samples
- Fewer laboratory manipulations; amenable to automation
- Ability to detect PCR inhibitors (Section 6.3.1.1)

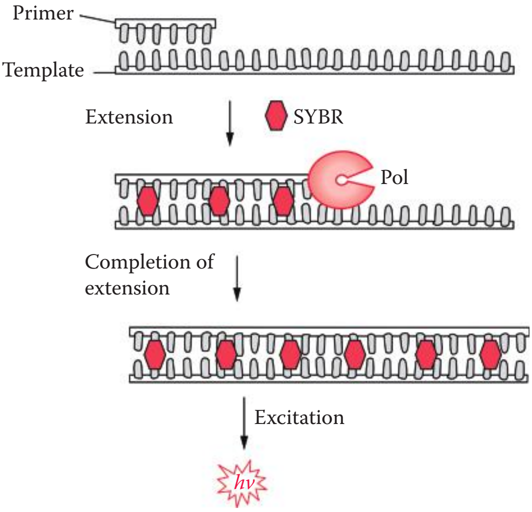

Figure 6.5 End-point PCR using SYBR Green detection. During the extension phase of PCR, in which DNA synthesis occurs, the dye binds to the double-stranded amplicons. Upon excitation, the emission intensity of the dye can be measured. Pol represents Taq polymerase. (© Richard C. Li.)

The technique is amenable to multiplexing to detect more than one type of DNA target sequence in a single reaction. Commercial qPCR kits for human DNA and Y-chromosome DNA quantitation are available. Additionally, the qPCR method for mtDNA quantitation is possible. qPCR uses commercially available fluorescence-detecting thermocyclers to amplify specific DNA sequences and measure their concentrations simultaneously. The fluorescent reporter can be a nonspecific intercalating double-stranded DNA-binding dye or a sequence-specific fluorescently labeled oligonucleotide probe (Chapter 9). The target sequences are amplified and detected by the same instrument, and the reporter fluorescence is monitored externally. Thus the reaction tubes do not need to be opened. This minimizes aerosol contamination and reduces the risk of false-positive results. A widely used qPCR probe technique is the TaqMan method (Applied Biosystems).

6.3.1.1 TaqMan Method

This method utilizes the 5′ exonuclease activity of *Taq* polymerase to cleave the probe during PCR (also known as the 5′ exonuclease assay). The probe is designed to anneal to the target sequence between the upstream and downstream primers and is added to the PCR mixture together with primers (Figure 6.6). The probe T_m (melting temperature; see Chapter 7) should be higher than the amplification primer T_m. A minor groove binder (MGB), such as dihydrocyclopyrroloindole tripeptide, is often linked at the 3′ end of the probe (Figure 6.6). A conjugated MGB binds to the minor groove of a B-form DNA helix (Figure 6.7), which is stabilized by van

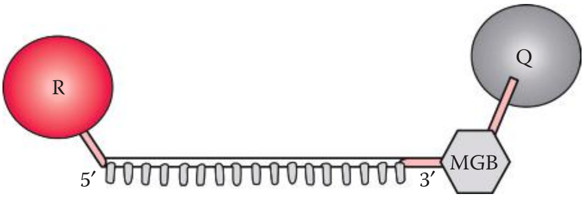

Figure 6.6 TaqMan probe. Each probe is labeled with a reporter dye (R) on the 5′ end and a fluorescence quencher (Q) on the 3′ end. MGB represents a minor groove binder. (© Richard C. Li.)

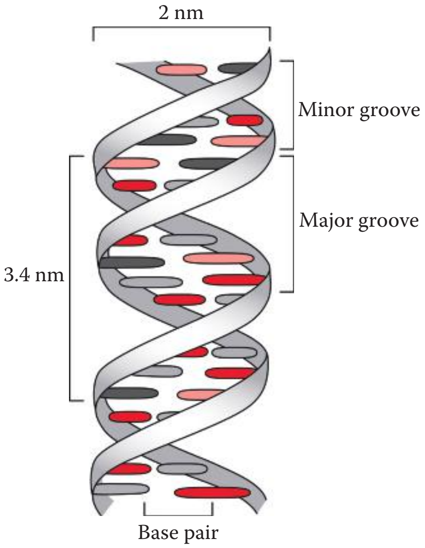

Figure 6.7 DNA double helix. The double-helical structure of B-DNA, the most common form of DNA, is shown. With a helical diameter of 2 nm, each turn of the helix takes 3.4 nm, which corresponds to 10 base pairs per turn. The major and minor grooves are shown. (© Richard C. Li.)

der Waals forces. As a result, conjugating an MGB into a probe increases the T_m values, allowing for the use of shorter probes. As a result, having the probe T_m higher than that of the primers ensures that the probe is fully hybridized during primer extension. The oligonucleotide probe is labeled with both a *reporter* fluorescent dye, usually 6-carboxyfluorescein (6-FAM) or tetrachlorofluorescein (TET) at the 5′ end; and a nonfluorescent *quencher* moiety, such as tetramethylrhodamine (TAMRA), usually at the 3′ end or any thymine position. While the probe is intact, the quencher greatly reduces the fluorescence emitted by the reporter via fluorescent resonance energy transfers (FRET). FRET is a distance-dependent interaction between two molecules in which the excitation energy is transferred from a photon donor molecule (reporter) to an acceptor molecule (quencher) without emission of a photon. During the extension phase of the PCR cycle, the 5′ exonuclease activity of Taq polymerase cleaves the reporter dye from the probe. Because the reporter dye is no longer in close proximity to the quencher, the FRET is disrupted and the probe begins to fluoresce (Figure 6.8). The intensity of fluorescence can be measured (Chapter 9) and is proportional to the amount of target DNA synthesized during the PCR.

In this assay, the rate of accumulation of amplified DNA over the entire course of a PCR is generated. The greater the initial concentration of target templates in a sample, the fewer cycles required to reach a particular quantity of amplified product. The initial concentration of target templates can be expressed using the *cycle threshold* (C_T). C_T is defined as the number of PCR cycles required for the fluorescent signal to cross a threshold of amplification where the signal exceeds background level or baseline noise. A plot of C_T against the $\log_{10}$ of the initial concentration of a set of DNA standards yields a straight line as a standard curve (Figure 6.9). The target sequences in an unknown sample can be quantified by comparing to the standard curve. Additionally, qPCR has the ability to detect PCR inhibitors (Chapter 7) that interact with DNA or with DNA polymerases. The presence of PCR inhibitors in the DNA extracts can be measured by monitoring the amplification of the internal positive control (IPC). Most human DNA quantitation kits contain a known amount of exogenous DNA as IPC that can be fortified to the sample and amplified. Monitoring the amplification of IPC enables the detection of PCR failure due to inhibition when the IPC's C_T value is higher than that of an uninhibited PCR reaction.

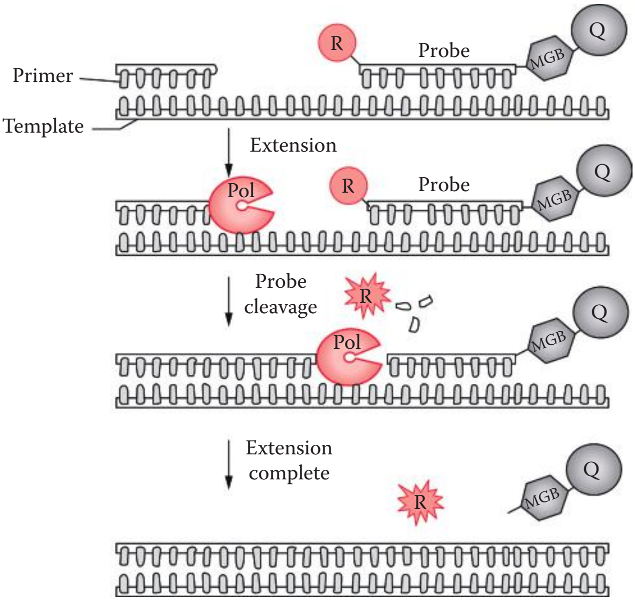

Figure 6.8 Real-time PCR with TaqMan detection. The TaqMan probe is shown. During extension, the 5′ nuclease activity of Taq polymerase (Pol) cleaves the probe. Reporter dye is released during each cycle of PCR. (© Richard C. Li.)

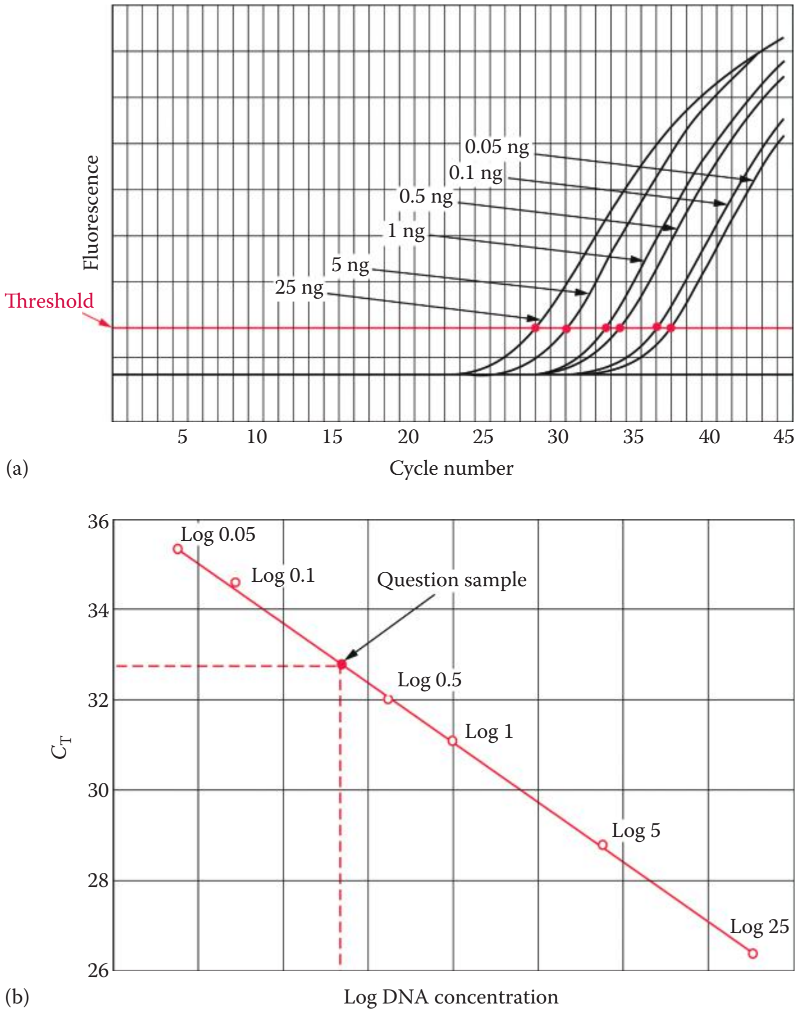

Figure 6.9 Real-time quantitative PCR. (a) Amplification curves for a dilution series of standards with known quantities of DNA. C_T is the cycle threshold at which the amplification curve crosses the threshold, as indicated by the red line. (b) A standard curve based on data obtained from the amplification curves. The quantity of DNA in a questioned sample can be determined from the standard curve. (© Richard C. Li.)

Bibliography

Ahn, S.J., J. Costa, and J.R. Emanuel, PicoGreen quantitation of DNA: Effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res*, 1996, **24**(13): 2623–2625.

Alonso, A. and P. Martin, A real-time PCR protocol to determine the number of amelogenin (X-Y) gene copies from forensic DNA samples. *Methods Mol Biol*, 2005, **297**: 31–44.

Alonso, A., et al., Specific quantification of human genomes from low copy number DNA samples in forensic and ancient DNA studies. *Croat Med J*, 2003, **44**(3): 273–280.

Alonso, A., et al., Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies. *Forensic Sci Int*, 2004, **139**(2–3): 141–149.

Andreasson, H. and M. Allen, Rapid quantification and sex determination of forensic evidence materials. *J Forensic Sci*, 2003, **48**(6): 1280–1287.

Andreasson, H., U. Gyllensten, and M. Allen, Real-time DNA quantification of nuclear and mitochondrial DNA in forensic analysis. *Biotechniques*, 2002, **33**(2): 402–404, 407–411.

Andreasson, H., et al., Nuclear and mitochondrial DNA quantification of various forensic materials. *Forensic Sci Int*, 2006, **164**(1): 56–64.

Budowle, B., et al., Simple protocols for typing forensic biological evidence: Chemiluminescent detection for human DNA quantitation and restriction fragment length polymorphism (RFLP) analyses and manual typing of polymerase chain reaction (PCR) amplified polymorphisms. *Electrophoresis*, 1995, **16**(9): 1559–1567.

Budowle, B., et al., The presumptive reagent fluorescein for detection of dilute bloodstains and subsequent STR typing of recovered DNA. *J Forensic Sci*, 2000, **45**(5): 1090–1092.

Budowle, B., et al., Using a CCD camera imaging system as a recording device to quantify human DNA by slot blot hybridization. *Biotechniques*, 2001, **30**(3): 680–685.

Del Rio, S.A., M.A. Marino, and P. Belgrader, Reusing the same blood-stained punch for sequential DNA amplifications and typing. *Biotechniques*, 1996, **20**(6): 970–972, 974.

Duewer, D.L., et al., NIST mixed stain studies #1 and #2: Interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples. *J Forensic Sci*, 2001, **46**(5): 1199–1210.

Fox, J.C., C.A. Cave, and J.W. Schumm, Development, characterization, and validation of a sensitive primate-specific quantification assay for forensic analysis. *Biotechniques*, 2003, **34**(2): 314–318, 320, 322.

Green, R.L., et al., Developmental validation of the quantifiler real-time PCR kits for the quantification of human nuclear DNA samples. *J Forensic Sci*, 2005, **50**(4): 809–825.

Hayn, S., et al., Evaluation of an automated liquid hybridization method for DNA quantitation. *J Forensic Sci*, 2004, **49**(1): 87–91.

Hewakapuge, S., et al., Investigation of telomere lengths measurement by quantitative real-time PCR to predict age. *Leg Med* (Tokyo), 2008, **10**(5): 236–242.

Hudlow, W.R., et al., A quadruplex real-time qPCR assay for the simultaneous assessment of total human DNA, human male DNA, DNA degradation and the presence of PCR inhibitors in forensic samples: A diagnostic tool for STR typing. *Forensic Sci Int Genet*, 2008, **2**(2): 108–125.

Kihlgren, A., A. Beckman, and S. Holgersson, Using D3S1358 for quantification of DNA amenable to PCR and for genotype screening. *Prog Forensic Genet*, 1998, **7**: 3.

Kline, M.C., et al., NIST Mixed Stain Study 3: DNA quantitation accuracy and its influence on short tandem repeat multiplex signal intensity. *Anal Chem*, 2003, **75**(10): 2463–2469.

Koukoulas, I., et al., Quantifiler observations of relevance to forensic casework. *J Forensic Sci*, 2008, **53**(1): 135–141.

Krenke, B.E., et al., Developmental validation of a real-time PCR assay for the simultaneous quantification of total human and male DNA. *Forensic Sci Int Genet*, 2008, **3**(1): 14–21.

LaSalle, H.E., G. Duncan, and B. McCord, An analysis of single and multi-copy methods for DNA quantitation by real-time polymerase chain reaction. *Forensic Sci Int Genet*, 2011, **5**(3): 185–193.

Mandrekar, M.N., et al., Development of a human DNA quantitation system. *Croat Med J*, 2001, **42**(3): 336–339.

Morrison, T.B., J.J. Weis, and C.T. Wittwer, Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques*, 1998, **24**(6): 954–958, 960, 962.

Myers, J.R., Validation of a DNA quantitation method on the Biomek® 3000. *J Forensic Sci*, 2010, **55**(6): 1570–1575.

Nicklas, J.A. and E. Buel, Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples. *J Forensic Sci*, 2003, **48**(5): 936–944.

Nicklas, J.A. and E. Buel, Quantification of DNA in forensic samples. *Anal Bioanal Chem*, 2003, **376**(8): 1160–1167.

Nicklas, J.A., T. Noreault-Conti, and E. Buel, Development of a real-time method to detect DNA degradation in forensic samples. *J Forensic Sci*, 2012, **57**(2): 466–471.

Niederstatter, H., et al., Characterization of mtDNA SNP typing and mixture ratio assessment with simultaneous real-time PCR quantification of both allelic states. *Int J Legal Med*, 2006, **120**(1): 18–23.

Niederstatter, H., et al., A modular real-time PCR concept for determining the quantity and quality of human nuclear and mitochondrial DNA. *Forensic Sci Int Genet*, 2007, **1**(1): 29–34.

Nielsen, K., et al., Comparison of five DNA quantification methods. *Forensic Sci Int Genet*, 2008, **2**(3): 226–230.

Omelia, E.J., M.L. Uchimoto, and G. Williams, Quantitative PCR analysis of blood- and saliva-specific microRNA markers following solid-phase DNA extraction. *Anal Biochem*, 2013, **435**(2): 120–122.

Ong, Y.L. and A. Irvine, Quantitative real-time PCR: A critique of method and practical considerations. *Hematology*, 2002, **7**(1): 59–67.

Puch-Solis, R., et al., Practical determination of the low template DNA threshold. *Forensic Sci Int Genet*, 2011, **5**(5): 422–427.

Richard, M.L., R.H. Frappier, and J.C. Newman, Developmental validation of a real-time quantitative PCR assay for automated quantification of human DNA. *J Forensic Sci*, 2003, **48**(5): 1041–1046.

Shewale, J.G., et al., Human genomic DNA quantitation system, H-Quant: Development and validation for use in forensic casework. *J Forensic Sci*, 2007, **52**(2): 364–370.

Sifis, M.E., K. Both, and L.A. Burgoyne, A more sensitive method for the quantitation of genomic DNA by Alu amplification. *J Forensic Sci*, 2002, **47**(3): 589–592.

Swango, K.L., et al., A quantitative PCR assay for the assessment of DNA degradation in forensic samples. *Forensic Sci Int*, 2006, **158**(1): 14–26.

Swango, K.L., et al., Developmental validation of a multiplex qPCR assay for assessing the quantity and quality of nuclear DNA in forensic samples. *Forensic Sci Int*, 2007, **170**(1): 35–45.

Tak, Y.K., et al., Highly sensitive polymerase chain reaction-free quantum dot-based quantification of forensic genomic DNA. *Anal Chim Acta*, 2012, **721**: 85–91.

Thomas, J.T., et al., Qiagen's Investigator Quantiplex Kit as a predictor of STR amplification success from low-yield DNA samples. *J Forensic Sci*, 2013, **58**(5): 1306–1309.

Timken, M.D., et al., A duplex real-time qPCR assay for the quantification of human nuclear and mitochondrial DNA in forensic samples: Implications for quantifying DNA in degraded samples. *J Forensic Sci*, 2005, **50**(5): 1044–1060.

Tozzo, P., et al., Discrimination between human and animal DNA: Application of a duplex polymerase chain reaction to forensic identification. *Am J Forensic Med Pathol*, 2011, **32**(2): 180–182.

Walker, J.A., et al., Multiplex polymerase chain reaction for simultaneous quantitation of human nuclear, mitochondrial, and male Y-chromosome DNA: Application in human identification. *Anal Biochem*, 2005, **337**(1): 89–97.

Walsh, P.S., J. Varlaro, and R. Reynolds, A rapid chemiluminescent method for quantitation of human DNA. *Nucleic Acids Res*, 1992, **20**(19): 5061–5065.

Waye, J.S., et al., A simple and sensitive method for quantifying human genomic DNA in forensic specimen extracts. *Biotechniques*, 1989, **7**(8): 852–855.

Westring, C.G., et al., Validation of reduced-scale reactions for the Quantifiler Human DNA kit. *J Forensic Sci*, 2007, **52**(5): 1035–1043.

Wilson, J., et al., Molecular assay for screening and quantifying DNA in biological evidence: The modified Q-TAT assay. *J Forensic Sci*, 2010, **55**(4): 1050–1057.

7

Amplification by Polymerase Chain Reaction

7.1 Denaturation and Renaturation of DNA

The DNA double helix is stabilized by chemical interactions. Base pairing of the two strands involves the formation of hydrogen bonds that provide weak electrostatic attractions between electronegative atoms. An adenine always pairs with a thymine (two hydrogen bonds), and a cytosine pairs with a guanine (three hydrogen bonds) (see Figure 7.1). Base stacking involves hydrophobic interactions between adjacent base pairs and provides stability to the double helix.

Double-stranded DNA is maintained by hydrogen bonding between the bases of complementary pairs. *Denaturation* occurs when the hydrogen bonds of DNA are disrupted, and the strands are separated. A *melting curve* can be obtained from measuring DNA denaturation by slowly heating a solution of DNA. As shown in Figure 7.2, an increasing temperature increases the percentage of the DNA that is denatured. The temperature at which 50% of DNA strands are denatured is defined as the *melting temperature* (T_m). The value of T_m is affected by the salt concentration of the solution, but can also be affected by nucleotide content, high pH, and length of the molecule.

Nucleotide content affects the value of T_m because GC pairs are joined by three hydrogen bonds while the AT pairs are joined by only two. Increasing the GC content of a DNA molecule increases the T_m. Excessively high pH causes the hydrogen bonds to break and the paired strands to separate. Finally, the length of the molecule also affects the T_m simply because a longer molecule of DNA requires more energy to break more bonds than a shorter molecule.

The single strands in a solution of denatured DNA can, under certain conditions, reanneal into double-stranded DNA. The process is called *renaturation* and two requirements must be met for it to occur. First, sufficient amounts of charged molecules, such as salts, must be present in the solution to neutralize the negative charges of the phosphate groups in DNA. This prevents the complementary strands from repelling each other. Additionally, the temperature must be high enough to disrupt hydrogen bonds that formed randomly between the bases of DNA strands. However, excessively elevated temperatures can disrupt the base pairs between the complementary DNA strands.

7.2 Basic Principles of Polymerase Chain Reaction

The *polymerase chain reaction* (*PCR*) allows the exponential amplification of specific sequences of DNA to yield sufficient amplified products, also known as *amplicons*, for various downstream

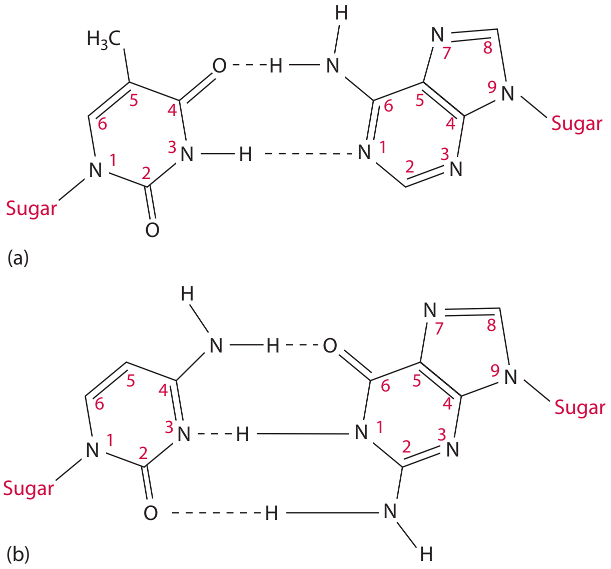

Figure 7.1 Base pairing between two DNA strands. (a) An adenine (right) pairs with a thymine (left). (b) A cytosine (left) pairs with a guanine (right).

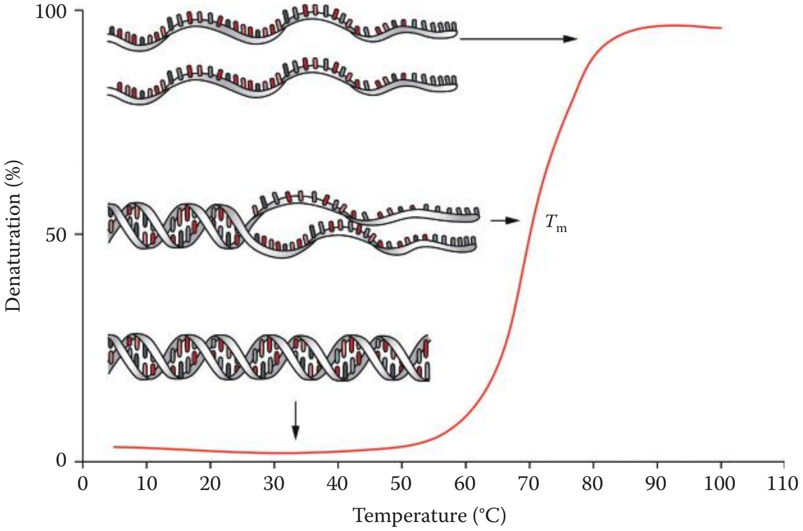

Figure 7.2 Melting curve of DNA. The degree of DNA denaturation is increased by increasing the temperature. T_m and possible shapes of a DNA molecule are shown. (© Richard C. Li.)

applications. The technique is highly sensitive and can amplify very small quantities of DNA. Therefore, it can be utilized for the analysis of samples of limited quantity. PCR-based assays are rapid and robust. Thus, PCR forms the basis of many forensic DNA assays such as DNA quantitation (Chapter 6), short tandem repeat (STR) profiling (Chapter 20), and mitochondrial DNA (mtDNA) sequencing (Chapter 23).

The concept of synthesizing DNA by a cycling process was first proposed in the early 1970s. In the mid-1980s, PCR technology was finally developed by Kary Mullis and his coworkers (Cetus Corporation) to amplify the β-globin gene for the diagnosis of sickle-cell anemia. In the late 1980s, a

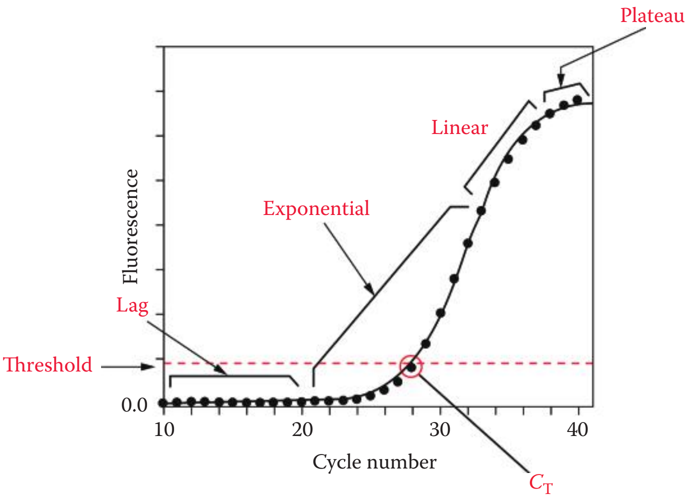

Figure 7.3 PCR amplification curve. A fluorescence signal results from the accumulation of amplicon. The location of the signal threshold is indicated as a dashed line. The C_T value for the curve is the cycle at which the amplification curve crosses the threshold. (© Richard C. Li.)

thermostable polymerase from *Thermus aquaticus* was utilized for PCR. This step greatly increased efficiency and allowed the process to be automated. The result was a powerful impact on molecular biology. In 1993, Mullis was awarded the Nobel Prize for the invention of PCR technology.

In the early 1990s, the technique for the simultaneous amplification and detection of the accumulation of amplicons at each PCR cycle was developed, and the concept of *real-time PCR* was born. This process allows the monitoring of amplicon production at each cycle of the PCR process. The fundamental processes were studied by characterizing the amplification kinetics of PCR using a graph that plotted the amount of amplicon yield at each cycle versus the cycle number (Figure 7.3).

An S-shaped amplification curve is obtained and divided into an exponential phase, a linear phase, and a plateau. During the exponential phase, the amplicon accumulates exponentially. It was revealed that the amplicon accumulation during PCR was correlated to the starting copy number of DNA template. Thus, the amount of amplicon produced during the exponential amplification phase can be used to determine the amount of starting material. This relationship can be further examined using a plot of cycle numbers versus a log scale of the serial dilution of the starting concentration of DNA template, which results in a linear relationship (see Chapter 6). It demonstrates that fewer cycles are needed if larger quantities of starting DNA template are present. The slope of this linear curve (Figure 6.9b) is known as the *amplification efficiency*. The exponential phase continues until one or more of the components (Section 7.3) in the reaction become limited. At this point, the amplification efficiency decreases, the amplicon no longer accumulates exponentially, and PCR enters the linear phase of the curve. At the plateau phase, no more amplicon is accumulated due to the exhaustion of reagents and polymerase.

7.3 Essential PCR Components

A PCR reaction requires thermostable DNA polymerases, primers, and other components, as described below (Figure 7.4).

7.3.1 Thermostable DNA Polymerases

A wide variety of DNA polymerases are available. They vary in fidelity, efficiency, and ability to synthesize longer DNA fragments. Nonetheless, *Taq* polymerase is the most commonly used

HAMILTON
microLAB
NIMBUS
NIMBUS PCR
Automated Assay Setup

Figure 7.4 A liquid handling workstation for automated PCR assay setup. (© Richard C. Li.)

enzyme for routine PCR applications (0.5–5 units per reaction). Currently, AmpliTaq Gold™ DNA polymerase (Applied Biosystems) is the most common DNA polymerase for forensic applications.

PCR reactions are usually set up at room temperature. Nonspecific annealing between primers and template DNA can occur, resulting in the formation of nonspecific amplicons. Additionally, annealing between the primers can occur to form primer dimers. Nonspecific annealing interferes with PCR amplification by reducing the amplification efficiency of the specific sequences of interest.

Such interference can be minimized by a *hot-start PCR* approach. The AmpliTaq Gold™ DNA polymerase, a modified enzyme, remains in an inactive form until activated with a pH below 7 prior to the PCR cycling in which the inhibitory motif is inactivated. The pH of the buffer system used in the PCR reaction is temperature sensitive; increasing the temperature decreases the pH of the solution. Thus, the activation of the enzyme can actually be carried out by a heating step at 95°C prior to the start of the cycling. During the heating process, the DNA strands also denature, which can prevent the formation of nonspecific PCR products.

7.3.2 PCR Primers

PCR primers are the oligonucleotides that are complementary to the sequences that flank the target region of the template. A pair (forward and reverse) of primers (typically 0.1–1 μM) is required. Properly designed primers are critical to the success of a PCR reaction. Computer software such as Primer3 is available to assist and optimize the designing of primers.

A primer must be specific to the target sequence; otherwise, nonspecific products that might interfere with the proper interpretation of a DNA profile might be produced. The primers within a pair should have similar T_m values. The estimated T_m values of a primer pair should not differ by more than 5°C. The T_m of an oligonucleotide primer can be predicted and calculated using the following equation:

$$T_m = 81.5°\text{C} + 16.6\left(\log_{10}[K^+]\right) + 0.41\left(\%\ [\text{G}+\text{C}]\right) - \left(\frac{675}{n}\right)$$

where:

$[K^+]$ = concentration of the potassium ion
$[\text{G}+\text{C}]$ = GC content (%) of the oligonucleotide
n = number of bases in the oligonucleotide

This equation shows that the T_m can be affected by primer base composition (GC content) and primer length. The GC content of an oligonucleotide primer should be 40%–60%, and the length of an oligonucleotide primer should be 15–25 base pairs, although longer primers can be used.

A primer should not contain self-complementary sequences that may form hairpin structures interfering with the annealing of primers and the template. Additionally, the primers in a pair should not share similar sequences, to avoid the primers from annealing to each other. These annealed primers may then be amplified during PCR, creating products known as *primer dimers*, which compete with the target DNA template for PCR components.

Multiplex PCR is the simultaneous amplification of more than one region of a DNA template in a single reaction to achieve high-throughput analysis. Multiplex PCR consists of multiple primer sets in a single reaction to produce amplicons of multiple target DNA regions. The primers should be designed to yield proper sizes of amplicons to be resolvable in downstream separation and detection procedures such as electrophoresis. To prevent preferential amplification of one target sequence over another, the annealing temperatures should be similar among multiplex PCR primer pairs. Additionally, the primers should lead to similar amplification efficiencies among the loci to be tested. Forensic applications of multiplex PCR, such as autosomal and Y-chromosomal short tandem repeat analysis, are discussed in Chapters 20 and 21.

7.3.3 Other Components

Essential components include template DNA with target sequences in either linear form (nuclear genomic DNA) or circular form (mitochondrial DNA). Both single- and double-stranded DNA can be used as a template for PCR. Typically, 1–2.5 ng of template DNA is utilized for forensic applications using PCR.

Deoxynucleoside triphosphates (dNTPs) are the substrates for DNA synthesis. A PCR assay usually contains equal molar amounts (typically, 200 μM) of dATP, dCTP, dGTP, and dTTP.

Divalent cations, such as Mg^{2+}, are required for the enzymatic activity of DNA polymerases. A PCR assay usually contains 1.5–2.5 mM Mg^{2+}. Monovalent cations, such as K^+ (50 mM), are usually recommended, and a buffer is often utilized to maintain pH between 8.3 and 8.8 at room temperature.

Controls should be used to monitor the effectiveness of PCR amplification. A *positive control* shows that PCR components such as reagents and PCR cycle parameters are working properly during a PCR. A standard DNA template should be used as a positive control and amplified with the same PCR components used on the rest of the samples. The amplification negative control and extraction reagent blank are discussed in Section 7.5.4.

7.4 Cycle Parameters

PCR cycling protocols may vary according to the type of analysis. Figure 7.5 shows representative PCR cycling conditions commonly used by forensic DNA laboratories. PCR utilizes a number of cycles for the replication of a specific region of a DNA template. During each cycle, a copy of the target DNA sequence is synthesized. A PCR cycle consists of three elements: denaturation, annealing, and extension. Precise and accurate temperatures at denaturation, annealing, and extension are critical to achieve a successful amplification. At the beginning of each cycle, the two complementary DNA template strands are separated at high temperatures (94°C–95°C) in a process called *denaturation*. The temperature is then decreased to allow *annealing* between the oligonucleotide primers and the template. The temperature for annealing is usually 3°C–5°C lower than the T_m of the oligonucleotide primer.

The annealing temperature is critical. If it is too high, a very low quantity of amplicon is yielded because of the failure of annealing between the primer and the template. If the annealing temperature is too low, nonspecific amplification can occur. Next, optimal temperature for DNA polymerase is reached, thus allowing for DNA replication (*extension*). By the end of each cycle, the copy number of the amplicon is nearly doubled (Figure 7.6).

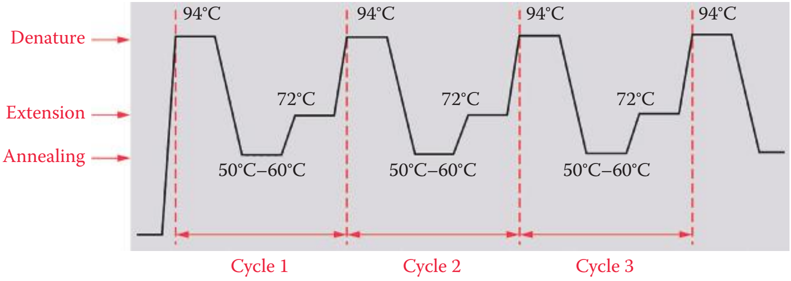

Figure 7.5 Temperature parameters during thermal cycling of the PCR process. The first three cycles are shown. (© Richard C. Li.)

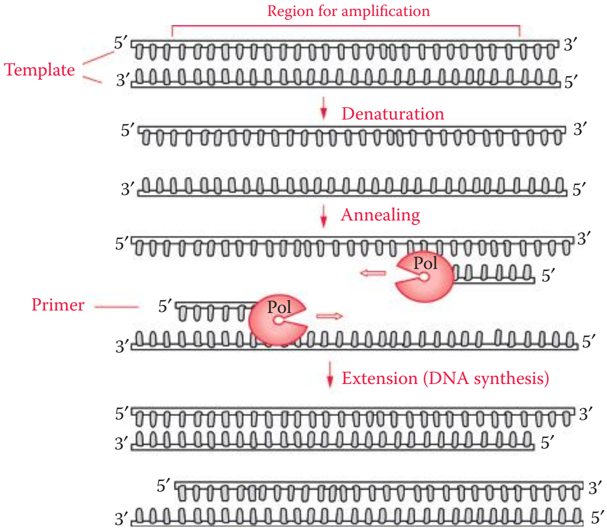

Figure 7.6 First cycle of PCR. The amplified region is determined by the positions of the primers. The direction (5′ to 3′) of DNA synthesis is indicated by arrows. Pol represents *Taq* polymerase. (© Richard C. Li.)

The number of cycles needed for PCR depends primarily on the number of copies of starting DNA template. The relationship can be expressed as the following equation:

$$N_x = N_0\left(1+E\right)^x$$

where:

X = number of PCR cycles
N_x = copy number of the amplicon after x cycles of PCR
N_0 = initial copy number of the template
E = amplification efficiency of the *Taq* polymerase

For example, in a 28-cycle PCR amplification, the DNA template can theoretically be amplified by a factor of approximately 10^8. If the cycle number is increased to 34, a factor of 10^{10} can theoretically be reached.

PCR amplification can be carried out using an instrument known as a *thermal cycler* (Figure 7.7). Various types of thermal cyclers differ in the number of samples they can process, the sizes of the sample tubes, and temperature control features.

Figure 7.7 Thermal cyclers. A thermal cycler can provide rapid temperature changes as desired to carry out PCR. A PCR thermal cycler (left) and a real-time PCR instrument (right). (© Richard C. Li.)

7.5 Factors Affecting PCR

7.5.1 Template Quality

It is important to prevent degradation of the DNA during the collection and processing of evidence. Degradation causes DNA to break into smaller fragments. If the damage occurs at a region to be amplified, the result can be failure in PCR amplification. In a degraded sample, the longer the amplicon length, the higher the risk of failure in PCR. Low copy number (LCN) of DNA template is often encountered in forensic samples. When amplifying very low levels (approximately 100 pg of DNA) of a template, the following phenomenon is often observed: one of the two alleles fails to be detected from a heterozygote and can falsely be identified as a homozygote. This phenomenon is also known as the *stochastic effect* in which the two alleles in a heterozygous individual are unequally detected at a low level of starting DNA template. Approaches such as increasing the cycle number, from 28 to 34 (Section 7.4), have been introduced to address the LCN problem.

7.5.2 Inhibitors

Inhibitors, if present, can interact with the DNA template or polymerase, causing PCR amplification failure. The presence of PCR inhibitors can be detected using an internal positive control (Section 6.3.1.1). A number of PCR inhibitors commonly encountered in evidence samples include heme molecules from blood, indigo dyes from fabrics, and melanin from hair samples. Thus, it is important to remove PCR inhibitors during DNA extraction. If PCR inhibitors are not eliminated during the extraction process, additional procedures such as the use of centrifugal filtration devices can be used. Centrifugal filtration devices can separate molecules by size. After the centrifugation step, small molecular weight inhibitors are filtered by passing through the membrane and are discarded. Alternatively, increasing the amount of DNA polymerase or adding bovine serum albumin (BSA) in the reaction can overcome the inhibition effects.

7.5.3 Contamination

PCR is a highly sensitive method; therefore, procedures that minimize the risk of contamination are necessary. To prevent contamination, pre- and post-PCR samples should be processed in separate areas or at different times. Additionally, reagents, supplies, and equipment used for pre- and post-PCR steps should be separated as well. Protective gear should include laboratory coats and disposable gloves. Facial masks and hair caps may be used if necessary. Aerosol-resistant pipet tips and DNA-free solutions and test tubes should also be used.

The levels of contamination must be monitored using controls. Extraction *reagent blanks*, which contain all extraction reagents but no sample, monitor contamination from extraction to PCR. Contamination detected in an extraction reagent blank but not in an amplification-negative control indicates that the reagents used for extraction are contaminated. Amplification-*negative* *controls*, which contain all PCR reagents and no DNA template, monitor contamination of the amplification. Contamination observed in an amplification-negative control but not in an extraction reagent blank indicates that the contamination occurred during the amplification step. Contamination observed in both an amplification reagent blank and an extraction-negative control indicates that amplification reagents are contaminated. A collection of DNA profiles of each member of a laboratory should be readily available for comparisons. Sources of laboratory contamination can be identified by comparing results with an analyst's DNA profile.

7.6 Reverse Transcriptase PCR for RNA-Based Assays

The pathway for the flow of genetic information is called the *central dogma*—a term coined by Francis Crick in 1956. According to the central dogma, parental DNA serves as the template for *DNA replication*. With RNA synthesis or *transcription*, the process is carried out using the DNA as a template. Conversely, RNA chains can be used as templates for the synthesis of a DNA strand of complementary sequence, in which the end product is referred to as *complementary DNA (cDNA).* Protein synthesis, also known as *translation*, is directed by an RNA template (Figure 7.8).

The flow of genetic information from RNA to DNA is referred to as *reverse transcription*. It was discovered independently by David Baltimore and Howard Temin in 1970, and they shared the Nobel Prize for their work. Reverse transcription is carried out by a *reverse transcriptase* that forms the basis of reverse transcriptase PCR as described below.

Reverse transcriptase PCR (RT-PCR) is highly sensitive and can be used to detect very small quantities of mRNA. It can be utilized to measure levels of gene expression even when the RNA of interest is expressed at very low levels. Detecting mRNAs of tissue-specific genes can be utilized for bodily fluid identification in forensic investigations (Chapter 11). During an RT-PCR process, a single-stranded cDNA is synthesized from a template mRNA using reverse transcription. The cDNA is then amplified by PCR for detection and analysis.

7.6.1 Reverse Transcription

The synthesis of single-stranded cDNA from an mRNA template is catalyzed by reverse transcriptase. Reverse transcriptases share many features in structure and function with DNA polymerases. The catalytic function of cDNA synthesis requires a primer that anneals to a complementary mRNA template. The primer can be either RNA or DNA; however, DNA primers are more efficient than RNA primers. During the elongation of the primer, reverse transcriptase incorporates the corresponding deoxyribonucleotide triphosphate according to the rules of base pairing with the RNA template. The RNA template is then degraded by an intrinsic RNase H activity of reverse transcriptase during the reverse transcription reaction. The retroviral RNase H is a domain of the viral reverse transcriptase enzyme. It is a nonspecific endonuclease that cleaves RNA.

Several reverse transcriptases derived from retroviruses can be used to generate cDNA from an RNA template. The most common reverse transcriptases used for cDNA synthesis are

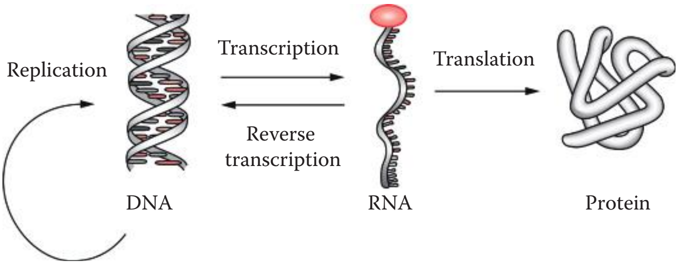

Figure 7.8 Pathway for the flow of genetic information. (© Richard C. Li.)

encoded by the *pol* gene from the avian myeloblastosis virus (AMV) and the Moloney strain of the murine leukemia virus (MMLV). These enzymes are genetically engineered reverse transcriptases but lack RNase H activity. Genetically engineered reverse transcriptases that are stable at higher temperatures (up to 60°C) are also produced. Increased thermal stability of reverse transcriptase allows the reverse transcription to be carried out at a higher than ambient temperature, which eliminates RNA secondary structures and improves the specificity of the reaction and the yields of the synthesis of full-length cDNA.

7.6.2 Oligodeoxynucleotide Priming

Oligodeoxynucleotide primers are an essential component for a reverse transcriptase reaction. Different priming strategies can be utilized to synthesize cDNA from a particular target mRNA or from all mRNA in a sample. *Gene-specific primers* are designed to hybridize to a particular mRNA sequence for the conversion of a specific gene sequence into cDNA. *Universal primers* can hybridize to any mRNA sequence in a sample to convert all mRNAs to cDNA. Two types of universal primers can be used: oligo (dT) and random hexamer primers. An *oligo (dT) primer* can hybridize to the 3′ termini poly (A) tails of eukaryotic mRNAs. Reverse transcriptases with oligo (dT) only synthesize cDNA from transcribed genes. *Random hexamer primers* are another type of universal primer. Random hexamers are nonspecific primers that can hybridize, at multiple sites, to any RNA sequence including non-mRNA templates such as ribosomal RNA.

7.6.3 Reverse Transcriptase PCR

During an RT-PCR process, a single-stranded cDNA is synthesized from a template mRNA using reverse transcription. The cDNA is then amplified by PCR with a pair of oligonucleotide primers corresponding to a specific sequence in the cDNA (Figure 7.9).

Two strategies of RT-PCR exist: a one-step and a two-step RT-PCR (Figure 7.10). One-step RT-PCR combines the reverse transcription reaction and PCR in a single tube. Only gene-specific

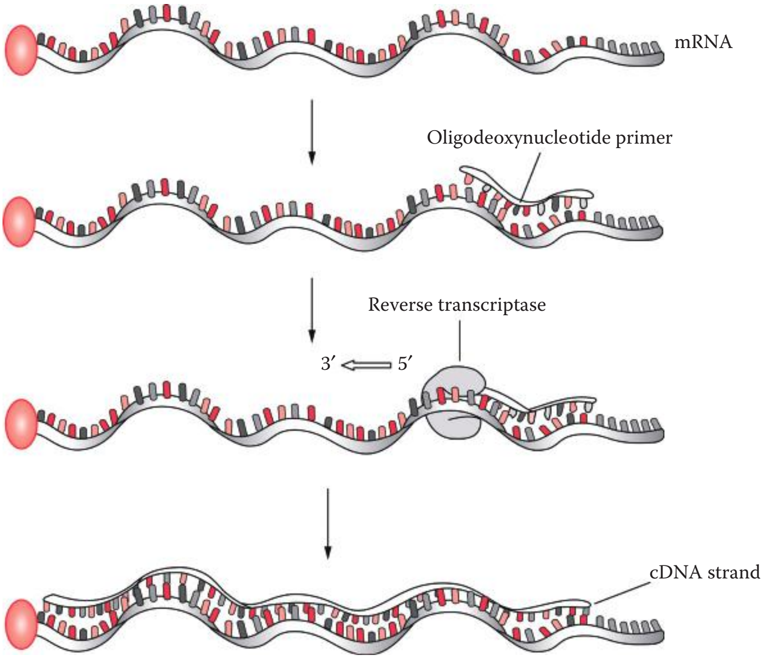

Figure 7.9 cDNA synthesis. The synthesis of a DNA strand transcribed from mRNA can be carried out using a primer and reverse transcriptase. A gene-specific primer that amplifies a specific target sequence is shown. (© Richard C. Li.)

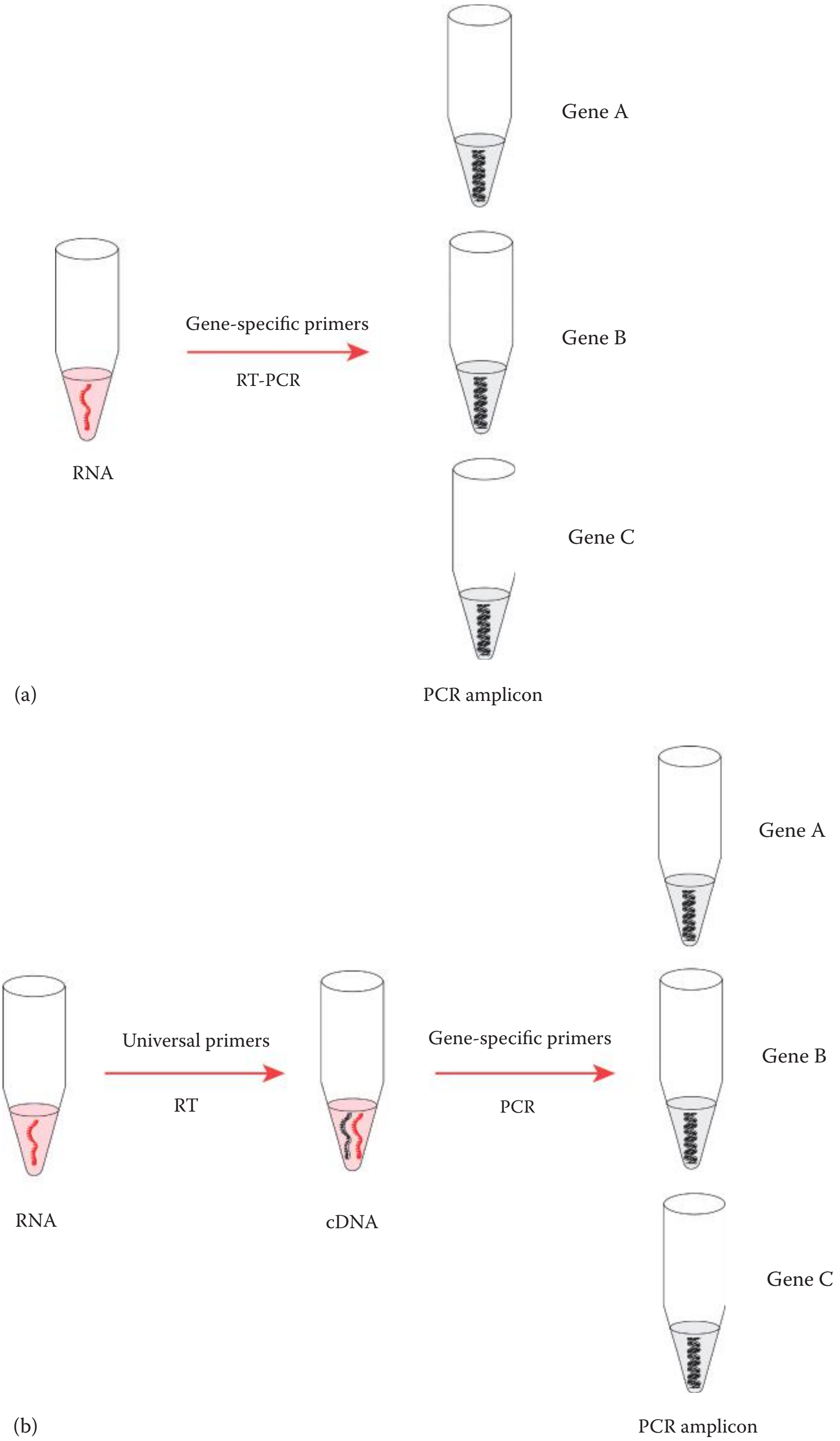
Gene A
Gene-specific primers
RT-PCR
Gene B
RNA
Gene C
(a)
PCR amplicon
Gene A
Universal primers
RT
Gene-specific primers
PCR
Gene B
RNA
cDNA
Gene C
(b)
PCR amplicon

Figure 7.10 RT-PCR strategies. (a) One-step RT-PCR and (b) two-step RT-PCR. (© Richard C. Li.)

primers can be used for the reverse transcription reaction and PCR. RNAs from either total RNA or mRNA can be used. This approach simplifies the reaction setup as it is useful for processing large numbers of samples. The tubes are not opened between reverse transcription and PCR, thus minimizing the risk of pipetting errors and carryover contamination.

During a two-step PCR, the reverse transcription reaction and PCR are carried out in separate tubes. Following the reverse transcription reaction, the cDNA is transferred to a separate tube for the PCR amplification. Either oligo (dT) or random hexamer primers can be used for the reverse transcription reaction. Total RNA or polyadenylated RNA can be used for the reverse transcriptase reaction. This allows for the ability to convert all the messages in an RNA sample into cDNA. The two-step RT-PCR is useful for analyzing multiple mRNAs from a single sample.

Two types of PCR methods can be utilized for analyzing amplified products. The end-point PCR method measures the amount of amplified product synthesized during PCR at the end of the PCR amplification. The detection of the amplified product indicates the presence of the mRNA of interest. With the real-time PCR method, the amplified product is quantified during the exponential phase of PCR. The hot-start PCR strategy is typically used to increase sensitivity, specificity, and yield. Usually, the hot-start strategy is carried out at a high temperature, for example 94°C, prior to PCR cycling. Under this condition, *Taq* DNA polymerase is activated. Additionally, the RNA-cDNA hybrid is denatured, and reverse transcriptase is inactivated at 94°C. Multiplex RT-PCR assays have been developed that can detect multiple bodily fluid mRNAs from single or mixed stains (Chapter 11). The commonly used RT-PCR for forensic identification of bodily fluids is the two-step RT-PCR. The end-point PCR is usually utilized with the forward primers labeled with florescent dyes at the 5′ end. The amplified products are separated and detected in a standard capillary electrophoresis instrument.

Bibliography

Aboud, M., H.H. Oh, and B. McCord, Rapid direct PCR for forensic genotyping in under 25 min. *Electrophoresis*, 2013, **34**(11): 1539–1547.

Afonina, I.A., et al., Minor groove binder-conjugated DNA probes for quantitative DNA detection by hybridization-triggered fluorescence. *Biotechniques*, 2002, **32**(4): 940–944, 946–949.

Alaeddini, R., Forensic implications of PCR inhibition—A review. *Forensic Sci Int Genet*, 2012, **6**(3): 297–305.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. *Forensic Sci Int*, 2007, **166**(1): 35–41.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Increased amplification success from forensic samples with locked nucleic acids. *Forensic Sci Int Genet*, 2011, **5**(4): 276–280.

Ballantyne, K.N., R.A. van Oorschot, and R.J. Mitchell, Locked nucleic acids in PCR primers increase sensitivity and performance. *Genomics*, 2008, **91**(3): 301–305.

Ballantyne, K.N., et al., Decreasing amplification bias associated with multiple displacement amplification and short tandem repeat genotyping. *Anal Biochem*, 2007, **368**(2): 222–229.

Bassam, B.J., et al., Nucleic acid sequence detection systems: Revolutionary automation for monitoring and reporting PCR products. *Aust Biotechnol*, 1996, **6**: 285–294.

Bekaert, B., et al., Automating a combined composite-consensus method to generate DNA profiles from low and high template mixture samples. *Forensic Sci Int Genet*, 2012, **6**(5): 588–593.

Birch, D.E., et al., Simplified hot start PCR. *Nature*, 1996, **318**: 2.

Bloch, W., A biochemical perspective of the polymerase chain reaction. *Biochemistry*, 1991, **30**(11): 2735–2747.

Bright, J.A., et al., The effect of cleaning agents on the ability to obtain DNA profiles using the Identifiler and PowerPlex® Y multiplex kits. *J Forensic Sci*, 2011, **56**(1): 181–185.

Brotherton, P., et al., Preferential access to genetic information from endogenous hominin ancient DNA and accurate quantitative SNP-typing via SPEX. *Nucleic Acids Res*, 2010, **38**(2): e7.

Budowle, B., A.J. Eisenberg, and A. van Daal, Validity of low copy number typing and applications to forensic science. *Croat Med J*, 2009, **50**(3): 207–217.

Chung, D.T., et al., A study on the effects of degradation and template concentration on the amplification efficiency of the STR Miniplex primer sets. *J Forensic Sci*, 2004, **49**(4): 733–740.

Date-Chong, M., et al., An examination of the utility of a nuclear DNA/mitochondrial DNA duplex qPCR assay to assess surface decontamination of hair. *Forensic Sci Int Genet*, 2013, **7**(3): 392–396.

Davis, C.P., et al., Multiplex short tandem repeat amplification of low template DNA samples with the addition of proofreading enzymes. *J Forensic Sci*, 2011, **56**(3): 726–732.

Deguilloux, M.F., et al., Analysis of ancient human DNA and primer contamination: One step backward one step forward. *Forensic Sci Int*, 2011, **210**(1–3): 102–109.

Dieffenbach, C.W., T.M. Lowe, and G.S. Dveksler, General concepts for PCR primer design. *PCR Methods Appl*, 1993, **3**(3): S30–S37.

Eckert, K.A. and T.A. Kunkel, DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl*, 1991, **1**(1): 17–24.

Eckert, K.A. and T.A. Kunkel, High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Nucleic Acids Res*, 1990, **18**(13): 3739–3744.

Edwards, M.C. and R.A. Gibbs, Multiplex PCR: Advantages, development, and applications. *PCR Methods Appl*, 1994, **3**(4): S65–S75.

Eilert, K.D. and D.R. Foran, Polymerase resistance to polymerase chain reaction inhibitors in bone*. *J Forensic Sci*, 2009, **54**(5): 1001–1007.

Endicott, P., et al., Genotyping human ancient mtDNA control and coding region polymorphisms with a multiplexed Single-Base-Extension assay: The singular maternal history of the Tyrolean Iceman. *BMC Genet*, 2009, **10**: 29.

Erlich, H.A., D. Gelfand, and J.J. Sninsky, Recent advances in the polymerase chain reaction. *Science*, 1991, **252**(5013): 1643–1651.

Estes, M.D., et al., Optimization of multiplexed PCR on an integrated microfluidic forensic platform for rapid DNA analysis. *Analyst*, 2012, **137**(23): 5510–5519.

Forster, L., J. Thomson, and S. Kutranov, Direct comparison of post-28-cycle PCR purification and modified capillary electrophoresis methods with the 34-cycle "low copy number" (LCN) method for analysis of trace forensic DNA samples. *Forensic Sci Int Genet*, 2008, **2**(4): 318–328.

Funes-Huacca, M.E., et al., A comparison of the effects of PCR inhibition in quantitative PCR and forensic STR analysis. *Electrophoresis*, 2011, **32**(9): 1084–1089.

Gasparini, P., et al., Amplification of DNA from epithelial cells in urine. *N Engl J Med*, 1989, **320**(12): 809.

Gavazaj, F.Q., et al., Optimization of DNA concentration to amplify short tandem repeats of human genomic DNA. *Bosn J Basic Med Sci*, 2012, **12**(4): 236–239.

Gefrides, L.A., et al., UV irradiation and autoclave treatment for elimination of contaminating DNA from laboratory consumables. *Forensic Sci Int Genet*, 2010, **4**(2): 89–94.

Ghasemi, A., et al., Design of a biological method for rapid detection of presence of PCR inhibitors in aged bone DNA. *Clin Lab*, 2012, **58**(7–8): 681–686.

Giardina, E., et al., Whole genome amplification and real-time PCR in forensic casework. *BMC Genom*, 2009, **10**: 159.

Giese, H., et al., Fast multiplexed polymerase chain reaction for conventional and microfluidic short tandem repeat analysis. *J Forensic Sci*, 2009, **54**(6): 1287–1296.

Gill, P., et al., An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Sci Int*, 2000, **112**(1): 17–40.

Grgicak, C.M., Z.M. Urban, and R.W. Cotton, Investigation of reproducibility and error associated with qPCR methods using Quantifiler® Duo DNA quantification kit. *J Forensic Sci*, 2010, **55**(5): 1331–1339.

Gyllensten, U.B. and H.A. Erlich, Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. *Proc Natl Acad Sci U S A*, 1988, **85**(20): 7652–7656.

Haas, C., E. Hanson, and J. Ballantyne, Capillary electrophoresis of a multiplex reverse transcription-polymerase chain reaction to target messenger RNA markers for body fluid identification. *Methods Mol Biol*, 2012, **830**: 169–183.

Haff, L., et al., A high-performance system for automation of the polymerase chain reaction. *Biotechniques*, 1991, **10**(1): 102–103, 106–112.

Hedman, J., et al., Improved forensic DNA analysis through the use of alternative DNA polymerases and statistical modeling of DNA profiles. *Biotechniques*, 2009, **47**(5): 951–958.

Henegariu, O., et al., Multiplex PCR: Critical parameters and step-by-step protocol. *Biotechniques*, 1997, **23**(3): 504–511.

Herrmann, D., et al., Microarray-based STR genotyping using RecA-mediated ligation. *Nucleic Acids Res*, 2010, **38**(17): e172.

Higuchi, R., et al., Kinetic PCR analysis: Real-time monitoring of DNA amplification reactions. *Biotechnology* (N Y), 1993, **11**(9): 1026–1030.

Higuchi, R., et al., Simultaneous amplification and detection of specific DNA sequences. *Biotechnology* (N Y), 1992, **10**(4): 413–417.

Hochmeister, M.N., et al., Confirmation of the identity of human skeletal remains using multiplex PCR amplification and typing kits. *J Forensic Sci*, 1995, **40**(4): 701–705.

Holland, P.M., et al., Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci U S A*, 1991, **88**(16): 7276–7280.

Hughes-Stamm, S.R., K.J. Ashton, and A. van Daal, Assessment of DNA degradation and the genotyping success of highly degraded samples. *Int J Legal Med*, 2011, **125**(3): 341–348.

Irwin, J.A., et al., Characterization of a modified amplification approach for improved STR recovery from severely degraded skeletal elements. *Forensic Sci Int Genet*, 2012, **6**(5): 578–587.

Isacsson, J., et al., Rapid and specific detection of PCR products using light-up probes. *Mol Cell Probes*, 2000, **14**(5): 321–328.

Ju, J., et al., Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc Natl Acad Sci U S A*, 1995, **92**(10): 4347–4351.

Kaminiwa, J., et al., Vanadium accelerates polymerase chain reaction and expands the applicability of forensic DNA testing. *J Forensic Leg Med*, 2013, **20**(4): 326–333.

Kavlick, M.F., et al., Quantification of human mitochondrial DNA using synthesized DNA standards. *J Forensic Sci*, 2011, **56**(6): 1457–1463.

Kermekchiev, M.B., et al., Mutants of *Taq* DNA polymerase resistant to PCR inhibitors allow DNA amplification from whole blood and crude soil samples. *Nucleic Acids Res*, 2009, **37**(5): e40.

Kimpton, C., et al., Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. *Int J Legal Med*, 1994, **106**(6): 302–311.

Kitayama, T., et al., Estimation of the detection rate in STR analysis by determining the DNA degradation ratio using quantitative PCR. *Leg Med (Tokyo)*, 2013, **15**(1): 1–6.

Koon, H.E., et al., Diagnosing post-mortem treatments which inhibit DNA amplification from US MIAs buried at the Punchbowl. *Forensic Sci Int*, 2008, **178**(2–3): 171–177.

Koppelkamm, A., et al., Validation of adequate endogenous reference genes for the normalization of qPCR gene expression data in human post mortem tissue. *Int J Legal Med*, 2010, **124**(5): 371–380.

Krawczak, M., et al., Polymerase chain reaction: Replication errors and reliability of gene diagnosis. *Nucleic Acids Res*, 1989, **17**(6): 2197–2201.

Kunkel, T.A., DNA replication fidelity. *J Biol Chem*, 1992, **267**(26): 18251–18254.

Kutyavin, I.V., et al., 3′-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. *Nucleic Acids Res*, 2000, **28**(2): 655–661.

Kwok, S. and R. Higuchi, Avoiding false positives with PCR. *Nature*, 1989, **339**(6221): 237–238.

Laurin, N. and C. Fregeau, Optimization and validation of a fast amplification protocol for AmpFlSTR® Profiler Plus® for rapid forensic human identification. *Forensic Sci Int Genet*, 2012, **6**(1): 47–57.

Lee, J.C., et al., Evaluating the performance of whole genome amplification for use in low template DNA typing. *Med Sci Law*, 2012, **52**(4): 223–228.

Maciejewska, A., J. Jakubowska, and R. Pawlowski, Whole genome amplification of degraded and nondegraded DNA for forensic purposes. *Int J Legal Med*, 2013, **127**(2): 309–319.

Markoulatos, P., N. Siafakas, and M. Moncany, Multiplex polymerase chain reaction: A practical approach. *J Clin Lab Anal*, 2002, **16**(1): 47–51.

Marshall, P.L., et al., Pressure cycling technology (PCT) reduces effects of inhibitors of the PCR. *Int J Legal Med*, 2013, **127**(2): 321–333.

McCloskey, M.L., et al., Encoding PCR products with batch-stamps and barcodes. *Biochem Genet*, 2007, **45**(11–12): 761–767.

Mitsuhashi, M., Technical report: Part 2. Basic requirements for designing optimal PCR primers. *J Clin Lab Anal*, 1996, **10**(5): 285–293.

Moretti, T., B. Koons, and B. Budowle, Enhancement of PCR amplification yield and specificity using AmpliTaq Gold DNA polymerase. *Biotechniques*, 1998, **25**(4): 716–722.

Morling, N., PCR in forensic genetics. *Biochem Soc Trans*, 2009, **37**(Pt 2): 438–440.

Myers, B.A., J.L. King, and B. Budowle, Evaluation and comparative analysis of direct amplification of STRs using PowerPlex® 18D and Identifiler® Direct systems. *Forensic Sci Int Genet*, 2012, **6**(5): 640–645.

Nazarenko, I.A., S.K. Bhatnagar, and R.J. Hohman, A closed tube format for amplification and detection of DNA based on energy transfer. *Nucleic Acids Res*, 1997, **25**(12): 2516–2521.

Nguyen, Q., et al., STR melting curve analysis as a genetic screening tool for crime scene samples. *J Forensic Sci*, 2012, **57**(4): 887–899.

Nicklas, J.A., T. Noreault-Conti, and E. Buel, Development of a fast, simple profiling method for sample screening using high resolution melting (HRM) of STRs. *J Forensic Sci*, 2012, **57**(2): 478–488.

Nogami, H., et al., Rapid and simple sex determination method from dental pulp by loop-mediated isothermal amplification. *Forensic Sci Int Genet*, 2008, **2**(4): 349–353.

Opel, K.L., D. Chung, and B.R. McCord, A study of PCR inhibition mechanisms using real time PCR. *J Forensic Sci*, 2010, **55**(1): 25–33.

Park, S.J., et al., Direct STR amplification from whole blood and blood- or saliva-spotted FTA without DNA purification. *J Forensic Sci*, 2008, **53**(2): 335–341.

Reynolds, R., G. Sensabaugh, and E. Blake, Analysis of genetic markers in forensic DNA samples using the polymerase chain reaction. *Anal Chem*, 1991, **63**(1): 2–15.

Roeder, A.D., et al., Maximizing DNA profiling success from sub-optimal quantities of DNA: A staged approach. *Forensic Sci Int Genet*, 2009, **3**(2): 128–137.

Rothe, J., N.E. Watkins Jr., and M. Nagy, New prediction model for probe specificity in an allele-specific extension reaction for haplotype-specific extraction (HSE) of Y chromosome mixtures. *PLoS One*, 2012, **7**(9): e45955.

Saiki, R.K., et al., Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 1985, **230**(4732): 1350–1354.

Saiki, R.K., et al., Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 1988, **239**(4839): 487–491.

Sanchez, J.J., et al., Multiplex PCR and minisequencing of SNPs—A model with 35 Y chromosome SNPs. *Forensic Sci Int*, 2003, **137**(1): 74–84.

Schoske, R., et al., Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci. *Anal Bioanal Chem*, 2003, **375**(3): 333–343.

Seo, S.B., et al., Effects of humic acid on DNA quantification with Quantifiler® Human DNA Quantification kit and short tandem repeat amplification efficiency. *Int J Legal Med*, 2012, **126**(6): 961–968.

Shuber, A.P., V.J. Grondin, and K.W. Klinger, A simplified procedure for developing multiplex PCRs. *Genome Res*, 1995, **5**(5): 488–493.

Smith, P.J. and J. Ballantyne, Simplified low-copy-number DNA analysis by post-PCR purification. *J Forensic Sci*, 2007, **52**(4): 820–829.

Solinas, A., N. Thelwell, and T. Brown, Intramolecular TaqMan probes for genetic analysis. *Chem Commun (Camb)*, 2002, (19): 2272–2273.

Steadman, S.A., et al., Recovery and STR amplification of DNA from RFLP membranes. *J Forensic Sci*, 2008, **53**(2): 349–358.

Sutlovic, D., et al., Interaction of humic acids with human DNA: Proposed mechanisms and kinetics. *Electrophoresis*, 2008, **29**(7): 1467–1472.

Swaran, Y.C. and L. Welch, A comparison between direct PCR and extraction to generate DNA profiles from samples retrieved from various substrates. *Forensic Sci Int Genet*, 2012, **6**(3): 407–412.

Tindall, K.R. and T.A. Kunkel, Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Biochemistry*, 1988, **27**(16): 6008–6013.

Tokutomi, T., et al., Real-time PCR method for identification of Asian populations in forensic casework. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S106–S108.

Turrina, S., et al., STR typing of archival Bouin's fluid-fixed paraffin-embedded tissue using new sensitive redesigned primers for three STR loci (CSF1P0, D8S1179 and D13S317). *J Forensic Leg Med*, 2008, **15**(1): 27–31.

Verheij, S., J. Harteveld, and T. Sijen, A protocol for direct and rapid multiplex PCR amplification on forensically relevant samples. *Forensic Sci Int Genet*, 2012, **6**(2): 167–175.

Wallin, J.M., et al., Constructing universal multiplex PCR systems for comparative genotyping. *J Forensic Sci*, 2002, **47**(1): 52–65.

Walsh, P.S., H.A. Erlich, and R. Higuchi, Preferential PCR amplification of alleles: Mechanisms and solutions. *PCR Methods Appl*, 1992, **1**(4): 241–250.

Wang, J. and B. McCord, The application of magnetic bead hybridization for the recovery and STR amplification of degraded and inhibited forensic DNA. *Electrophoresis*, 2011, **32**(13): 1631–1638.

Weiler, N.E., A.S. Matai, and T. Sijen, Extended PCR conditions to reduce drop-out frequencies in low template STR typing including unequal mixtures. *Forensic Sci Int Genet*, 2012, **6**(1): 102–107.

Westwood, S.A. and D.J. Werrett, An evaluation of the polymerase chain reaction method for forensic applications. *Forensic Sci Int*, 1990, **45**(3): 201–215.

Weusten, J. and J. Herbergs, A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications. *Forensic Sci Int Genet*, 2012, **6**(1): 17–25.

Whitcombe, D., et al., A homogeneous fluorescence assay for PCR amplicons: Its application to real-time, single-tube genotyping. *Clin Chem*, 1998, **44**(5): 918–923.

Wong, H.Y., E.S. Lim, and W.F. Tan-Siew, Amplification volume reduction on DNA database samples using FTA Classic Cards. *Forensic Sci Int Genet*, 2012, **6**(2): 176–179.

Yu, P.H. and M.M. Wallace, Effect of 1,2-indanedione on PCR-STR typing of fingerprints deposited on thermal and carbonless paper. *Forensic Sci Int*, 2007, **168**(2–3): 112–118.

8

DNA Electrophoresis

It is necessary to separate various sizes of DNA fragments so that the DNA fragment in question can be identified and analyzed. This can be achieved by electrophoresis, a process in which fragments are separated based on the migration of charged macromolecules in an electric field.

8.1 Basic Principles

DNA is a negatively charged molecule in an aqueous environment, with the phosphate groups of DNA nucleotides carrying negative charges. DNA molecules migrate from the negative electrode (cathode) toward the positive electrode (anode) in an electric field during electrophoresis. The *electrical potential*, a measure of the work required to move a charged molecule in an electric field, is the force responsible for moving the charged macromolecules during electrophoresis.

The *electrophoretic mobility* of macromolecules is primarily determined by their *charge-to-mass ratios* and their shapes. However, because the phosphate group of every DNA nucleotide carries a negative charge, the charge-to-mass ratio of DNA molecules is almost the same even if the length of the DNA fragment varies. Additionally, forensic DNA testing usually analyzes linear DNA molecules such as double-stranded linear DNA for variable number tandem repeat (VNTR) analysis (Chapter 19) and single-stranded DNA for short tandem repeat (STR) analysis (Chapter 20). The shapes of linear DNA molecules are similar. Therefore, the electrophoretic separation of different-sized DNA fragments, through a series of pores of the supporting matrix in which the fragments travel, is based more on their sizes than their shapes.

It is obviously easier for smaller molecules to migrate through the pores of a supporting matrix than larger molecules; this is why smaller molecules migrate faster through the matrix. Hence, the electrophoretic mobility increases as the size of the DNA molecule decreases. Conversely, larger DNA molecules migrate much more slowly because they experience more friction and collisions as they travel through the net of pores in a matrix. Therefore, the separation of DNA molecules with different sizes can be accomplished. Figure 8.1 depicts the models for DNA electrophoresis.

8.2 Supporting Matrices

Although the actual separation occurs in an aqueous phase, most variants of electrophoresis use a physical support material also called a *matrix*. As discussed in Section 8.1, the matrix can be used as a molecular sieve for the separation of DNA molecules, and it also reduces diffusion and convection during electrophoresis. Agarose and polyacrylamide are commonly used electrophoresis matrices because of their good reproducibility, reliability, and versatility.

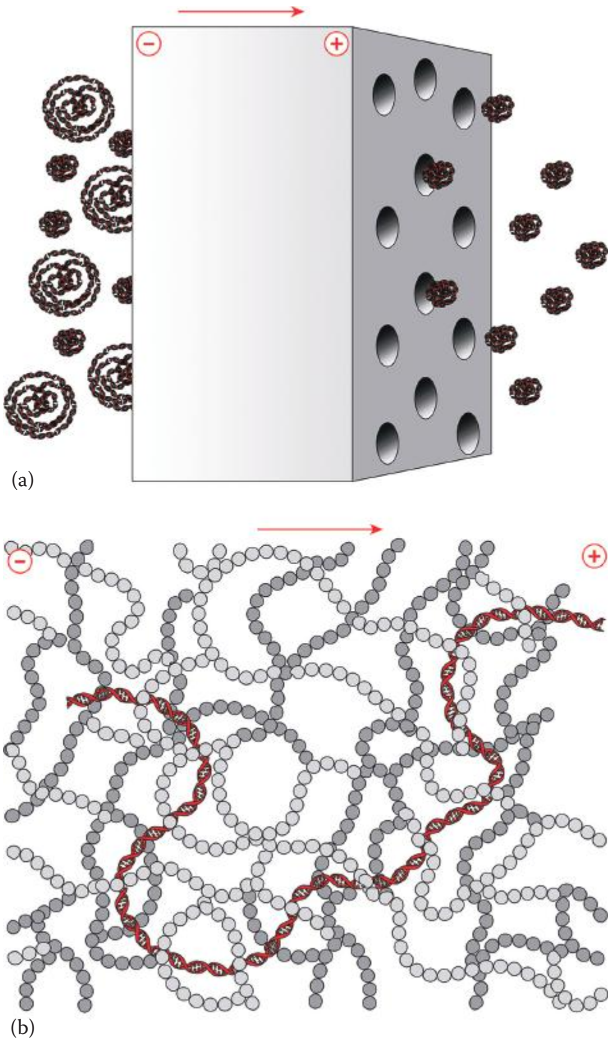

Figure 8.1 Schematic illustrations showing some of the nucleic acid separation mechanisms. (a) Ogston sieving model for the separation of DNA fragments through electrophoresis. Based on this model, the matrix used for gel electrophoresis consists of randomly distributed pores. DNA molecules are considered to behave as globular objects. During the migration, smaller DNA molecules migrate faster than their larger counterparts through the matrix. However, this model does not apply to DNA molecules with a radius much larger than the pore size of the matrix. (b) A reptation model presents an alternative mechanism of gel electrophoretic separation of DNA molecules. Based on this model, DNA molecules are flexible and can migrate through the pores of the gel matrix through a reptation process that is driven by an applied electric field. This model can also apply to capillary electrophoresis separation of DNA molecules in entangled polymer solutions. +, anode; −, cathode; arrow, direction of migration. (© Richard C. Li.)

8.2.1 Agarose

Agarose is a linear polymer composed of alternating residues of D- and L-galactose (Figure 8.2). A gelatinized agarose forms a three-dimensional sieve with pores from 50 to 200 nm in diameter (Figure 8.3). DNA fragments ranging from 50 to 20,000 base pairs in size are best resolved in agarose gels.

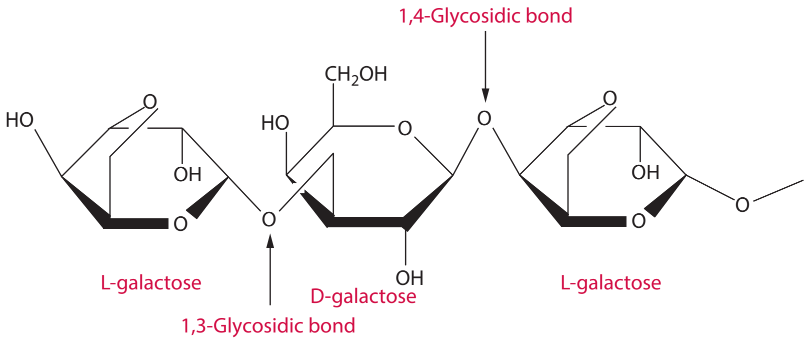

Figure 8.2 Chemical components of agarose, a linear polysaccharide composed of alternating units of D- and L-galactose. Often, the L-galactose residue has an anhydro bridge between the 3 and 6 positions and is called 3,6-anhydro-L-galactose.

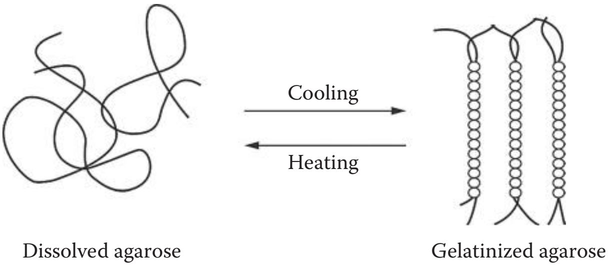

Figure 8.3 Agarose chains. Gelatinized agarose consists of a three-dimensional network in which chains of agarose form helical fibers that aggregate into supercoiled structures. (© Richard C. Li.)

The electrophoretic mobility of double-stranded DNA through agarose gel matrices is inversely proportional to the $\log_{10}$ of the size of DNA fragments ranging from 50 to 20,000 base pairs. The electrophoretic mobility of a linear DNA fragment is also inversely proportional to the concentration of agarose in the gel. A low concentration of agarose in a gel forms larger pores, allowing the separation of larger DNA fragments.

8.2.2 Polyacrylamide

A polyacrylamide gel matrix is very effective for the separation of smaller fragments of DNA (5–500 base pairs). Additionally, a single nucleotide difference in the length of a DNA fragment can be resolved with this type of gel. Polyacrylamide produces much smaller pore sizes than agarose gels and thus has a much higher resolving power than agarose gels for low-molecular-weight DNA molecules. A polyacrylamide gel matrix is formed by polymerization and cross-linking reactions.

8.2.2.1 Polymerization Reaction

Long linear chains of polyacrylamide are polymerized from acrylamide monomers (Figure 8.4). This polymerization reaction is initiated in the presence of free radicals that are generated from the reduction of ammonium persulfate (APS) by *N*,*N*,*N*′,*N*′-tetramethylethylene diamine (TEMED). See Figure 8.5.

8.2.2.2 Cross-Linking Reaction

As shown in Figure 8.5, three-dimensionally cross-linked polyacrylamide chains can be formed with the use of cross-linking agents, such as *N*,*N*′-methylenebisacrylamide (BIS; Figure 8.4).

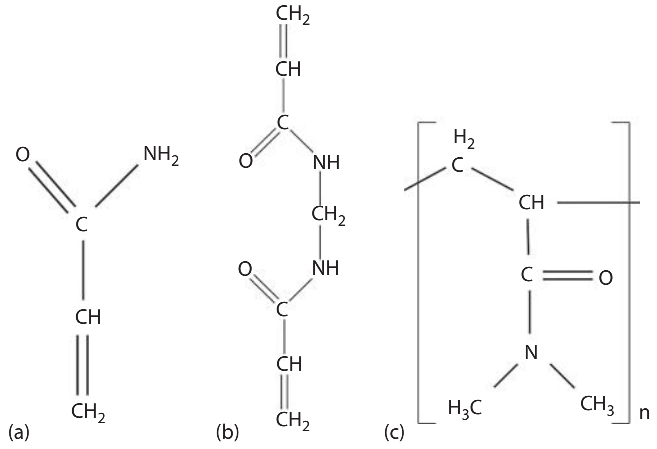
(a)
(b)
(c)
O
NH2
C
CH
CH2
CH2
CH
C
O
NH
CH2
NH
O
C
CH
CH2
H2
C
CH
C
O
N
H3C
CH3
n

Figure 8.4 Chemical structures of (a) acrylamide, (b) *N*,*N*′-methylenebisacrylamide (BIS), and (c) polydimethylacrylamide (a linear polymer for capillary electrophoresis).

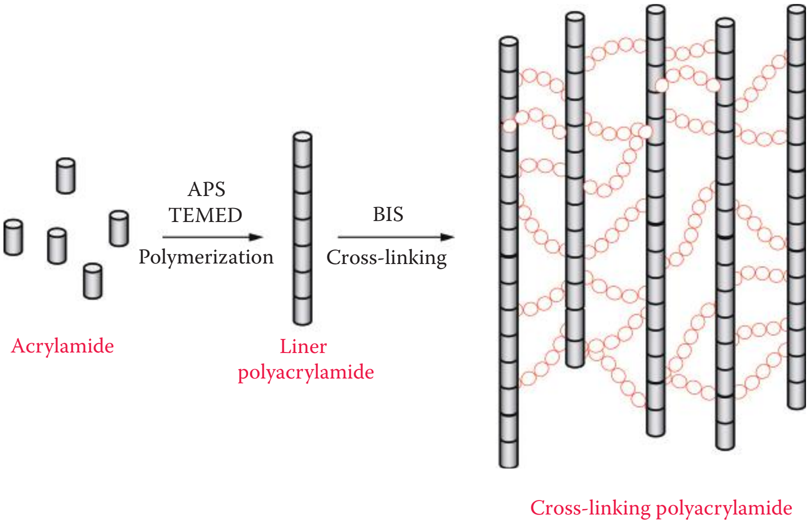
APS
TEMED
Polymerization
BIS
Cross-linking
Acrylamide
Liner
polyacrylamide
Cross-linking polyacrylamide

Figure 8.5 Formation of polyacrylamide gel. Polymerization is followed by cross-linking in which the pore size of the gel is determined by its degree of polymerization and cross-linking. (© Richard C. Li.)

The porosity of the resulting gel is determined by the lengths of these chains and the degree of cross-linking between them. Therefore, the sizes of the pores that are formed can be adjusted by altering the concentrations of the acrylamide monomer and the cross-linking reagent.

Polyacrylamide can also be used in a capillary electrophoresis matrix. It is very difficult to insert a cross-linked polyacrylamide matrix into a capillary. Therefore, a solution of a linear polymer (non-cross-linked) of polydimethylacrylamide (Figure 8.4) is used as a matrix for capillary electrophoresis. For instance, POP-4 (4% un-cross-linked polydimethylacrylamide polymer) is used for fragment analysis, such as forensic STR analysis (Chapter 18), in which fluorescently labeled DNA amplicons are separated using electrophoresis and are identified based on their sizes by comparing them with a size standard. POP-6 (6% un-cross-linked polydimethylacrylamide polymer) is used for sequencing, such as mitochondrial DNA (mtDNA) sequencing (Chapter 21). These un-cross-linked polymers are commercially available (Applied Biosystems).

8.2.2.3 Denaturing Polyacrylamide Electrophoresis

Polyacrylamide-based electrophoresis can be carried out with both double- and single-stranded DNA. Electrophoresis that is performed under the conditions of single-stranded DNA is called *denaturing electrophoresis*. Denatured DNA migrates through the gel as linear molecules at a rate independent of the base composition and sequence. With short single-stranded fragments, molecules differing in size by a single nucleotide can be separated. This extraordinary sensitivity to size is extremely useful for the separation of small DNA fragments that are used for DNA sequencing and STR analysis in forensic applications. The denaturing condition can be best achieved by adding chemicals such as urea or formamide to the matrix or by increasing the temperature or pH during electrophoresis.

8.3 Apparatus and Forensic Applications

Agarose electrophoresis is carried out in a slab gel. Agarose gel can be prepared in a variety of shapes and sizes and can be run in different configurations. The most common gel for forensic use is the horizontal slab. Polyacrylamide can be used either in a slab gel or in a capillary electrophoretic apparatus.

8.3.1 Slab Gel Electrophoresis

8.3.1.1 Agarose Gel Electrophoresis

One forensic application of agarose gel electrophoresis is restriction fragment length polymorphism (RFLP) analysis of VNTR loci (see Chapter 19). An agarose gel is used to separate the DNA fragments by size ranging from 500 to 20,000 base pairs of commonly used VNTR loci for forensic testing. This type of electrophoresis is done under nondenatured conditions (double-stranded DNA).

Figure 8.6 shows an example of the apparatus used for gel electrophoresis of DNA. A slab (3–4 mm) of agarose gel is prepared by allowing liquid agarose to gelatinize in a cast. The gel is submerged in a buffer tank filled with an electrophoretic buffer with proper ionic strength, which is necessary to achieve efficient electrical conductance. The gel contains small wells for loading samples. The DNA samples are mixed with a gel loading buffer prior to loading them into the gel. The gel loading buffer contains dyes that add color to the sample to facilitate the process of loading. Dyes such as bromophenol blue and xylene cyanol FF migrate toward the anode during electrophoresis. In most forensic applications, these dyes migrate faster than DNA fragments. The dyes are visible and can thus be used to track the progression of electrophoresis. The electrophoresis can be stopped as the dye front reaches the bottom of the gel (anode side).

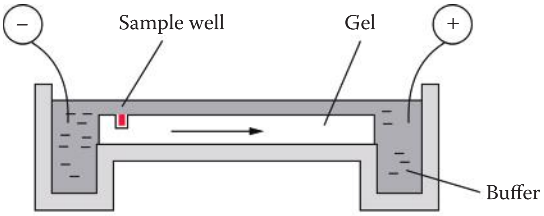

Figure 8.6 Horizontal agarose slab gel apparatus. The gel can be prepared by heating an agarose suspension to dissolve it. The agarose solution can then be poured into a gel cast and allowed to cool until gelatinized. An electrophoresis buffer is poured over the gel to submerge it. The samples, usually mixed with loading buffer containing dye to allow visualization, are loaded into the wells of the submerged gel using a pipet. The electrodes are indicated. (© Richard C. Li.)

The samples are loaded into the sample wells situated proximal to the cathode. An electric field is applied and the negatively charged DNA molecules migrate through the agarose toward the anode. Slab gel electrophoresis is capable of running multiple samples simultaneously. The sample is made visible by using a fluorescent intercalating reagent that can make DNA fluoresce under ultraviolet (UV) light (Chapter 9), either using the fluorescent intercalating reagent incorporated into the gel or staining the gel after electrophoresis. The separated DNA samples appear as bands and the sizes of the DNA fragments can be estimated by comparing them with the sizes of standards run concurrently.

8.3.1.2 Polyacrylamide Gel Electrophoresis

The forensic application of this apparatus is the separation of STR fragments and DNA sequencing reaction products of mtDNA. The sizes of DNA fragments that can be separated range from 100 to 500 base pairs—much smaller than what can be separated efficiently with agarose gels. Single-nucleotide resolution to distinguish similarly sized fragments can be achieved with this technique under denatured conditions with only single-stranded DNA. Polyacrylamide thin slab gels (0.75–1.5 mm) are usually used. Samples are loaded into wells flanked by two pieces of glass. Therefore, polyacrylamide gels are usually run in a vertical configuration (Figure 8.7). The detection of DNA bands in polyacrylamide gels will be described in Chapter 9. The sizes of DNA fragments can be calculated by including an internal size standard. As with agarose gel, polyacrylamide gel electrophoresis is capable of running multiple samples simultaneously and high throughput (a measurement of the rate that a sample is processed by a given analysis) can be achieved. However, caution should be taken to ensure that cross-contamination does not occur from the spilling of samples from adjacent wells. Additionally, polyacrylamide gels are more difficult to prepare than agarose gels.

8.3.2 Capillary Electrophoresis

Capillary electrophoresis (Figures 8.8 through 8.11) is a newer method than the slab gel method, and can be utilized to separate charged macromolecules, such as DNA, RNA, polysaccharides, and proteins.

One essential component of the capillary electrophoresis instrument is the capillary, a thin hollow tube made of fused silica (which typically has a diameter between 50 and 100 µm and a length between 10 and 50 cm). The capillary contains a translucent detection window for the instrument to detect signals from the labeled DNA fragments during electrophoresis.

Linear polydimethylacrylamide is used as the matrix. Capillary electrophoresis is conducted under denatured conditions for forensic applications such as STR analysis and mtDNA sequencing analysis. The denatured condition is achieved by including urea in the electrophoresis matrix and formamide during sample preparation. The injection of samples into the capillary is performed by an autosampler using an *electrokinetic* mechanism (an injection based on the charge of molecules). Only small quantities of sample are required for each injection, and any remaining sample can be saved in case an analysis needs to be repeated.

During electrophoresis, the capillary is connected to buffer reservoirs that are connected to electrodes. The efficient heat dissipation property of thin capillaries allows the separation to be performed at higher voltages, as the electric field (typically 200–300 V/cm) is much higher than that of a slab gel platform. Thus, the separation in capillary electrophoresis is rapid. During electrophoresis, DNA fragments migrate through the capillaries toward the anode. Similar to gel electrophoresis, DNA fragments are separated according to their sizes, with the shorter fragments moving faster than the longer fragments. During capillary electrophoresis, linear polymers in the solution can act as obstacles to the migrating DNA fragments. Thus, electrophoretic separation models (Figure 8.1) developed for gels are applicable for capillary electrophoresis. Currently, capillary electrophoresis can separate DNA fragments of up to 1000 nucleotides with single-nucleotide resolution.

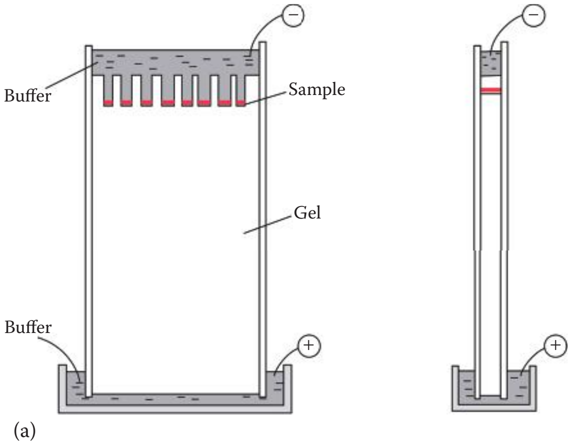

Figure 8.7 Vertical slab polyacrylamide gel. (a) Front and side view of vertical slab polyacrylamide gel apparatus. (b) Automated gel electrophoresis instrument, ABI PRISM 377® Genetic Analyzer, which was used in forensic laboratories but is now discontinued. (© Richard C. Li.)

Capillary electrophoresis instruments are equipped with detection systems utilizing laser excitation sources that excite the fluorescently labeled DNA fragments. They also include fluorescence detectors that record the signals emitted from the labeled DNA fragments (Chapter 9). In capillary electrophoresis, samples can only be analyzed sequentially so the throughput is much more limited than with slab gels. This disadvantage can be overcome by utilizing capillary array systems that can run up to 16 capillaries at one time (Figure 8.10).

8.3.3 Microfluidic Devices

Forensic DNA analysis includes the extraction of DNA from a sample, DNA quantification, polymerase chain reaction (PCR) amplification, capillary electrophoresis, data collection, and genotyping. Various microfluidic devices have been developed for forensic DNA analysis. Microfluidic devices control the movement of samples and reagents in a small, geometrically constrained environment and carry out biochemical reactions and analysis. These devices are made using microfabrication technology. This technology fabricates miniature structures historically used for integrated circuit fabrication in semiconductor manufacturing. The microfluidic

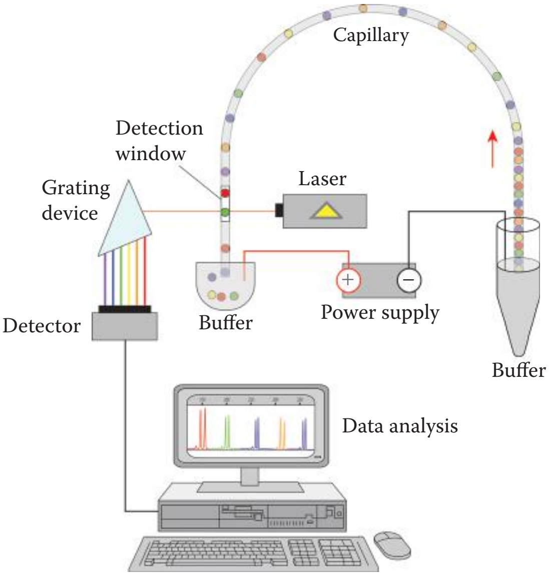

Figure 8.8 Essential components of a capillary electrophoresis system. The system includes a capillary, buffer reservoirs, two electrodes, a laser excitation source, a fluorescence detector, and an autosampler that holds the sample. Sample injection, electrophoresis, and data collection are automated and controlled by a computer. (© Richard C. Li.)

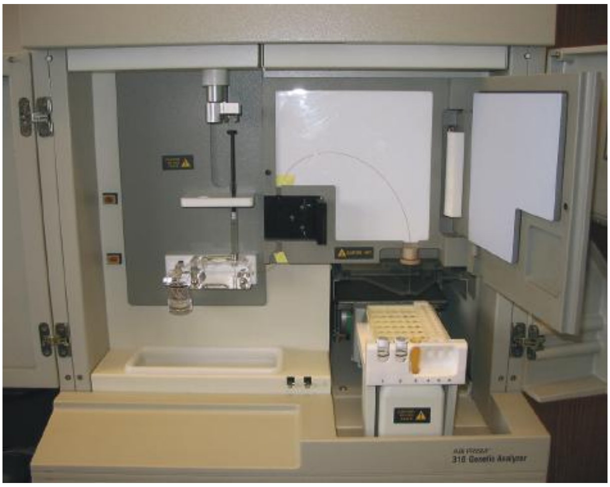

Figure 8.9 A single-capillary electrophoresis instrument, ABI PRISM 310® Genetic Analyzer. This was used in forensic laboratories but is now discontinued. (© Richard C. Li.)

devices can be classified as modular and integrated devices. Overall, the microfluidic device has many advantages compared with conventional techniques. The reaction volume required in microfluidic devices is usually in the nanoliter range, which decreases reagent and sample consumption. Additionally, due to the high surface area-to-volume ratio of the system, the efficiency of the system is greatly improved. Lastly, it can be automated, which makes the device a potential new platform for forensic DNA analysis.

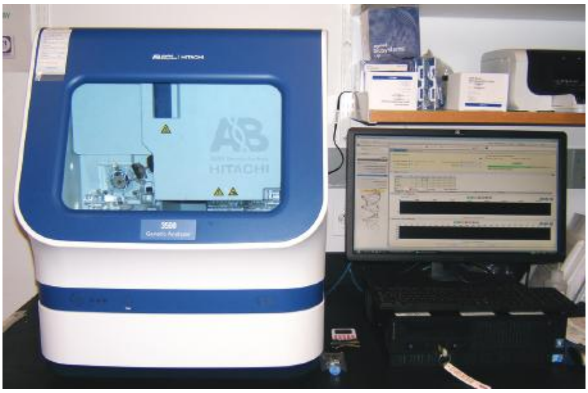

Figure 8.10 Photo of a multicapillary electrophoresis instrument. ABI PRISM 3500® Genetic Analyzer. (© Richard C. Li.)

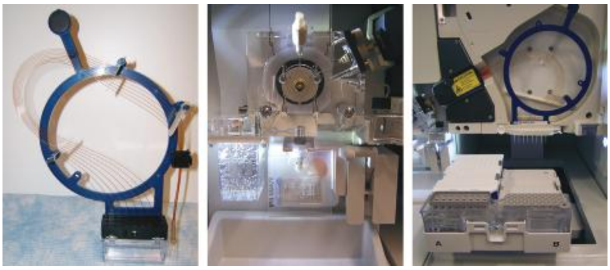

Figure 8.11 Components of the ABI PRISM 3500® Genetic Analyzer. A capillary array (left), electrophoretic buffer chambers (middle), and an autosampler (right). (© Richard C. Li.)

8.3.3.1 Modular Microfluidic Devices

The modular devices are separate devices for DNA extraction, quantitation, amplification, electrophoresis, and so on. Each processing step is carried out on different devices. The modular design is flexible, allowing laboratories to choose the best device suited to the procedure for each step. For example, an electrophoresis microfluidic device has been developed (Figure 8.12). The electrophoretic assay is carried out on a modular microfluidic chip format using the same basic principle of capillary electrophoresis. The single-use chip contains wells for samples, a sieving polymer matrix for electrophoresis, and size standards. The wells are connected through microchannels. When the interconnected wells and channels are filled, the chip becomes an integrated electrical circuit. Each well is then connected to an electrode with an independent power supply. DNA molecules, carrying negative charges, are

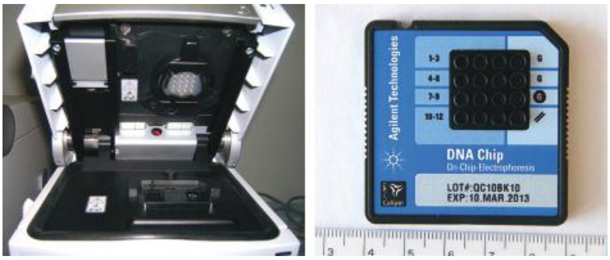

Figure 8.12 A modular instrument for gel electrophoresis. Left: Agilent 2100 Bioanalyzers contain 16-pin electrodes that fit into the wells of a chip. Right: The chip can accommodate sample wells, gel wells, and a well for a size standard. (© Richard C. Li.)

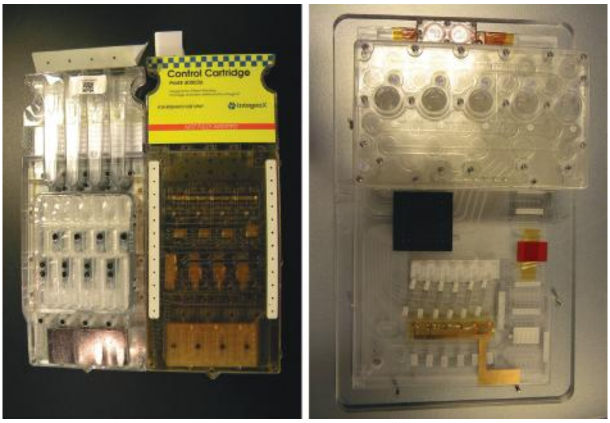

Figure 8.13 Integrated instruments for rapid DNA analysis. Cartridges developed by IntegeneX (left) and Zygem (right). (© Richard C. Li.)

electrophoretically separated based on their size. Smaller molecular weight fragments migrate faster than larger fragments. The polymer matrix contains fluorescent dye molecules that intercalate with DNA in a sample, which can be detected by laser-induced fluorescence. The duration of electrophoresis is approximately 30 min. The sensitivity is approximately 0.1 ng/μL; thus, sample consumption is small. The device provides the sizing and quantitation information of the nucleic acid fragment. In addition to DNA, the device can also be used for the analysis of RNA and protein. In the forensic laboratory, it has been used for the quantitation of amplified mtDNA product to monitor successful amplification and to normalize the quantities of DNA template for cycle sequencing.

8.3.3.2 Integrated Microfluidic Devices

It is possible to integrate several different modular devices into a single device (Figure 8.13). These devices are also known as "micro-total analysis" systems. An integrated system is a fully automated process. It also reduces the risk of contamination from laboratory sources. Full integration of all the constituent steps required for forensic STR DNA analysis has been achieved by a *rapid DNA* instrument. This instrument integrates various microfluidic devices including DNA extraction, PCR amplification, electrophoresis, detection, and genotyping steps into a single process on a cartridge. The cartridge, made from a variety of materials such as polycarbonate, incorporates pumps and valves controlling fluidic movement for the delivery of samples and prefilled reagents to the chambers for processing. DNA extraction usually utilizes silica-coated magnetic particles that can be collected by a magnet for DNA extraction. Microquantitative PCR utilizes the Taqman real-time quantitative PCR technique. During amplification, multiplex PCR is accomplished using commercial STR kits. Additionally, the cartridge also integrates a microcapillary electrophoresis chip for the separation of amplified products. The fluorescently labeled DNA fragments are detected based on laser-induced fluorescence. Data analysis and genotyping are then carried out to produce a DNA profile, which is compatible with the CODIS DNA database (Chapter 24). Rapid DNA instruments are fully automated platforms that are designed to process a single-source sample such as a buccal swab to generate a DNA profile and database search in less than 2 h. The goal is for these instruments to be deployed in field tests at the scene or at the booking stations operated by trained law enforcement agents.

8.4 Estimation of DNA Size

8.4.1 Relative Mobility

The *relative mobility* (R_f) of a DNA molecule during electrophoresis can be calculated as the distance of band migration divided by the distance of tracking dye migration. The DNA

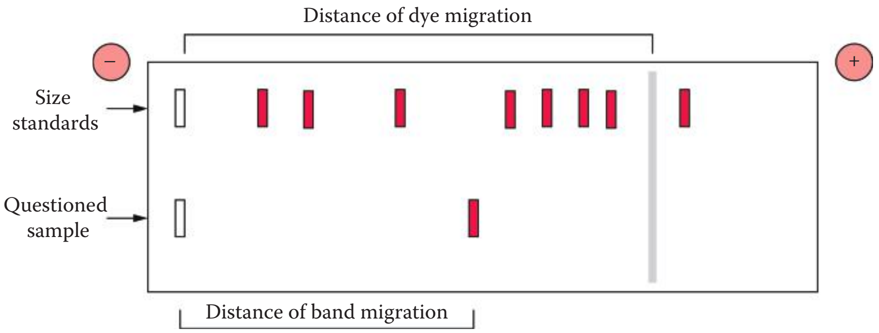

Figure 8.14 Measurement of the electromobility of DNA. Questioned DNA samples are analyzed concurrently with a set of standard DNA fragments of known size. The distances of DNA migration can thus be measured. The samples are usually mixed with a gel loading buffer prior to loading samples to the gel. The buffer contains dyes that add color to the sample to facilitate the process of loading. Dyes such as bromophenol blue and xylene cyanol FF migrate toward the positive electrode during electrophoresis. They can be used for tracking purposes as well. (© Richard C. Li.)

band migration is the distance from the sample origin to the center of the band. The tracking dye migration distance extends from the sample origin to the center of the dye band (Figure 8.14). To estimate the size of the DNA, standards containing DNA of known size and questioned samples are run on the same gel at the same time. The standards can be used to estimate the size of an unknown DNA molecule. A plot of $\log_{10}$ base pair of the standards versus R_f for a given gel can be constructed. A linear relationship, over a size range, between $\log_{10}$ base pair of the DNA molecule and the R_f can be observed. The R_f of the test sample is interpolated on the plot from which the size of an unknown DNA molecule can be determined (Figure 8.15).

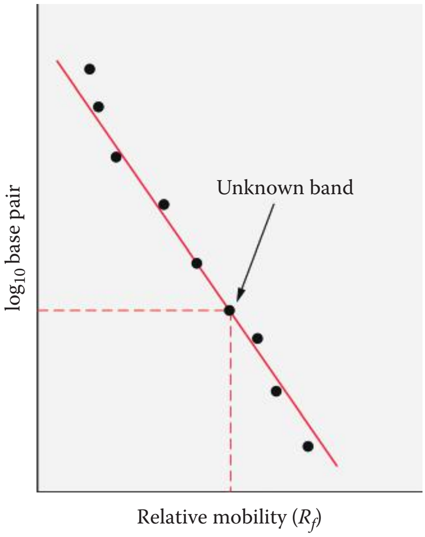

Figure 8.15 Estimation of the size of a DNA fragment. A plot of size standards versus relative migration during electrophoresis is shown. The size of the questioned DNA fragment can be estimated using this plot. (© Richard C. Li.)

8.4.2 Local Southern Method

The size of a DNA fragment is determined by an internal size standard that is a set of synthetic fragments with known molecular weight. The standard is labeled with a different colored dye so that it can be spectrally distinguished from DNA fragments of an unknown size (Figure 8.16). The sample including the internal size standard is then mixed in with DNA samples and is analyzed by electrophoresis. To determine the sizes of DNA fragments, a standard curve using the internal size standards must be established based on the reciprocal relationship between the electrophoretic mobility and the sizes of DNA fragments. However, this relationship is not exactly linear; instead, it appears to be sigmoidal (Figure 8.17). Therefore, the *Local Southern method*, described by Sir Edwin Southern, is used to generate standard curves for determining the sizes of DNA fragments. The equation is as follows:

$$L = \left[\frac{C}{M - M_0}\right] + L_0$$

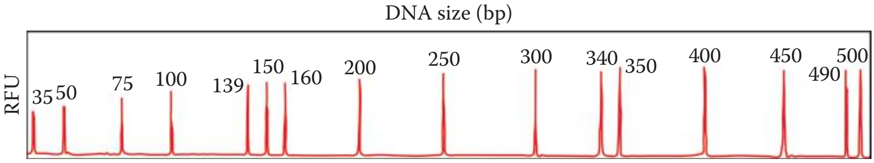

Figure 8.16 Electropherogram of a synthetic molecular weight size standard, GeneScan™ 500 size standard (Applied Biosystems). RFU, relative fluorescence units. (© Richard C. Li.)

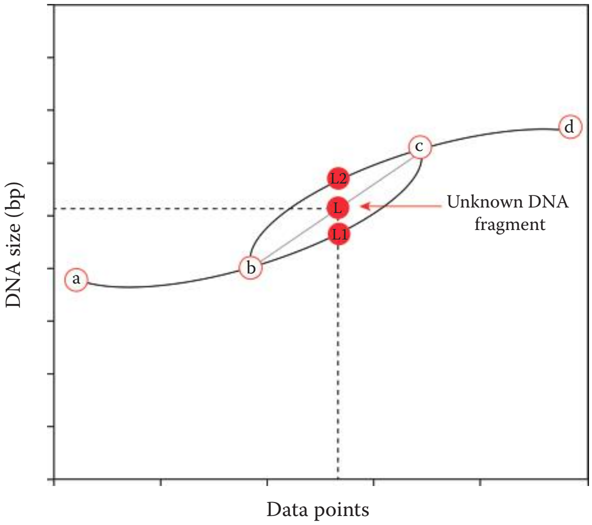

Figure 8.17 Estimation of unknown DNA fragment sizes using the Local Southern method. The migration of ssDNA fragments (20–1200 base pairs) is not linear over the entire fragment size range and is somewhat sigmoidal. To determine the size of a DNA fragment in question, two curves are generated. The first curve is generated using three standard points: two points (a and b) that are smaller and one point (c) that is larger than the questioned fragment. As a result, a fragment size (L_1) is determined. The second curve is then generated using an additional three standard points: one point (b) that is smaller and two points (c and d) that are larger than the questioned fragment. Thus, a fragment size, L_2, is determined. The two values (L_1 and L_2) are averaged to determine the size (L) of the questioned fragment. (Adapted from Southern, E.M., *Anal Biochem*, 100, 319–323, 1979.)

In the equation, L is the size of the unknown fragment; M is the mobility of the fragment; and C, M_0, and L_0 are constants that are obtained from the fragments of known size. To determine the size of an unknown fragment, two curves are generated based on the equation. Each curve utilizes three data points. The first curve utilizes two size standards that are smaller and one size standard that is greater than that of the unknown fragment. As a result, size L_1 is obtained. The second curve utilizes a size standard that is smaller and two size standards that are greater than that of the unknown fragment. Thus, size L_2 is obtained. Finally, the average of L_1 and L_2 is defined as the size of the unknown fragment. In this method, only the size standards that are in the proximity of the unknown fragment are analyzed. However, the accuracy of the standard curve relies on the proper separation of the standards during electrophoresis.

Bibliography

Aboud, M.J., M. Gassmann, and B.R. McCord, The development of mini pentameric STR loci for rapid analysis of forensic DNA samples on a microfluidic system. *Electrophoresis*, 2010, **31**(15): 2672–2679.

Biega, L.A. and B.W. Duceman, Substitution of H_2O for formamide in the sample preparation protocol for STR analysis using the capillary electrophoresis system: The effects on precision, resolution, and capillary life. *J Forensic Sci*, 1999, **44**: 3.

Bienvenue, J.M., et al., An integrated microfluidic device for DNA purification and PCR amplification of STR fragments. *Forensic Sci Int Genet*, 2010, **4**(3): 178–186.

Buel, E., M.B. Schwartz, and M.J. LaFountain, Capillary electrophoresis STR analysis: Comparison to gel-based systems. *J Forensic Sci*, 1998, **43**(1): 164–170.

Butler, J.M., C.M. Ruitberg, and P.M. Vallone, Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. *Fresenius J Anal Chem*, 2001, **369**(3–4): 200–205.

Butler, J.M., et al., Rapid analysis of the short tandem repeat HUMTH01 by capillary electrophoresis. *Biotechniques*, 1994, **17**(6): 1062–1064, 1066, 1068, passim.

Butler, J.M., et al., Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, 2004, **25**(10–11): 1397–1412.

Chen, Y., et al., Sample stacking capillary electrophoretic microdevice for highly sensitive mini Y short tandem repeat genotyping. *Electrophoresis*, 2010, **31**(17): 2974–2980.

Choi, J.Y. and T.S. Seo, An integrated microdevice for high-performance short tandem repeat genotyping. *Biotechnol J*, 2009, **4**(11): 1530–1541.

Fernandez, C. and A. Alonso, Microchip capillary electrophoresis protocol to evaluate quality and quantity of mtDNA amplified fragments for DNA sequencing in forensic genetics. *Methods Mol Biol*, 2012, **830**: 367–379.

Frazier, R.R., et al., Validation of the applied biosystems Prism 377 automated sequencer for the forensic short tandem repeat analysis. *Electrophoresis*, 1996, **17**(10): 1550–1552.

French, D.J., et al., Interrogation of short tandem repeats using fluorescent probes and melting curve analysis: A step towards rapid DNA identity screening. *Forensic Sci Int Genet*, 2008, **2**(4): 333–339.

Gale, N., et al., Rapid typing of STRs in the human genome by HyBeacon melting. *Org Biomol Chem*, 2008, **6**(24): 4553–4559.

Gill, P., P. Koumi, and H. Allen, Sizing short tandem repeat alleles in capillary array gel electrophoresis instruments. *Electrophoresis*, 2001, **22**(13): 2670–2678.

Greenspoon, S.A., et al., A forensic laboratory tests the Berkeley microfabricated capillary array electrophoresis device. *J Forensic Sci*, 2008, **53**(4): 828–837.

Hagan, K.A., et al., A valveless microfluidic device for integrated solid phase extraction and polymerase chain reaction for short tandem repeat (STR) analysis. *Analyst*, 2011, **136**(9): 1928–1937.

Horsman, K.M., et al., Forensic DNA analysis on microfluidic devices: A review. *J Forensic Sci*, 2007, **52**(4): 784–799.

Hurth, C., et al., An automated instrument for human STR identification: Design, characterization, and experimental validation. *Electrophoresis*, 2010, **31**(21): 3510–3517.

Isenberg, A.R., et al., Analysis of two multiplexed short tandem repeat systems using capillary electrophoresis with multiwavelength fluorescence detection. *Electrophoresis*, 1998, **19**(1): 94–100.

Issaq, H.J., K.C. Chan, and G.M. Muschik, The effect of column length, applied voltage, gel type, and concentration on the capillary electrophoresis separation of DNA fragments and polymerase chain reaction products. *Electrophoresis*, 1997, **18**(7): 1153–1158.

Koumi, P., et al., Evaluation and validation of the ABI 3700, ABI 3100, and the MegaBACE 1000 capillary array electrophoresis instruments for use with short tandem repeat microsatellite typing in a forensic environment. *Electrophoresis*, 2004, **25**(14): 2227–2241.

Lazaruk, K., et al., Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. *Electrophoresis*, 1998, **19**(1): 86–93.

Liu, P., et al., Integrated portable polymerase chain reaction-capillary electrophoresis microsystem for rapid forensic short tandem repeat typing. *Anal Chem*, 2007, **79**(5): 1881–1889.

Liu, P., et al., Real-time forensic DNA analysis at a crime scene using a portable microchip analyzer. *Forensic Sci Int Genet*, 2008, **2**(4): 301–309.

Liu, P. and R.A. Mathies, Integrated microfluidic systems for high-performance genetic analysis. *Trends Biotechnol*, 2009, **27**(10): 572–581.

Liu, P., et al., Integrated DNA purification, PCR, sample cleanup, and capillary electrophoresis microchip for forensic human identification. *Lab Chip*, 2011, **11**(6): 1041–1048.

Liu, P., et al., Integrated sample cleanup and capillary array electrophoresis microchip for forensic short tandem repeat analysis. *Forensic Sci Int Genet*, 2011, **5**(5): 484–492.

Liu, P., et al., Integrated sample cleanup and microchip capillary array electrophoresis for high-performance forensic STR profiling. *Methods Mol Biol*, 2012, **830**: 351–365.

Lounsbury, J.A., et al., From sample to PCR product in under 45 minutes: A polymeric integrated microdevice for clinical and forensic DNA analysis. *Lab Chip*, 2013, **13**(7): 1384–1393.

Madabhushi, R.S., Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis*, 1998, **19**(2): 224–230.

Mansfield, E.S., et al., Sensitivity, reproducibility, and accuracy in short tandem repeat genotyping using capillary array electrophoresis. *Genome Res*, 1996, **6**(9): 893–903.

Mansfield, E.S., et al., Analysis of multiplexed short tandem repeat (STR) systems using capillary array electrophoresis. *Electrophoresis*, 1998, **19**(1): 101–107.

McLaren, R.S., et al., Post-injection hybridization of complementary DNA strands on capillary electrophoresis platforms: A novel solution for dsDNA artifacts. *Forensic Sci Int Genet*, 2008, **2**(4): 257–273.

Montesino, M. and L. Prieto, Capillary electrophoresis of big-dye terminator sequencing reactions for human mtDNA control region haplotyping in the identification of human remains. *Methods Mol Biol*, 2012, **830**: 267–281.

Moretti, T.R., et al., Validation of STR typing by capillary electrophoresis. *J Forensic Sci*, 2001, **46**(3): 661–676.

Nai, Y.H., S.M. Powell, and M.C. Breadmore, Capillary electrophoretic system of ribonucleic acid molecules. *J Chromatogr A*, 2012, **1267**: 2–9.

Njoroge, S.K., et al., Integrated microfluidic systems for DNA analysis. *Top Curr Chem*, 2011, **304**: 203–260.

Pascali, J.P., F. Bortolotti, and F. Tagliaro, Recent advances in the application of CE to forensic sciences, an update over years 2009–2011. *Electrophoresis*, 2012, **33**(1): 117–126.

Pereira, J., et al., MtDNA typing of single-sperm cells isolated by micromanipulation. *Forensic Sci Int Genet*, 2012, **6**(2): 228–235.

Rosenblum, B.B., et al., Improved single-strand DNA sizing accuracy in capillary electrophoresis. *Nucleic Acids Res*, 1997, **25**(19): 3925–3929.

Roy, R., et al., Producing STR locus patterns from bloodstains and other forensic samples using an infrared fluorescent automated DNA sequencer. *J Forensic Sci*, 1996, **41**(3): 418–424.

Schneider, C., et al., Low copy number DNA profiling from isolated sperm using the aureka® micromanipulation system. *Forensic Sci Int Genet*, 2012, **6**(4): 461–465.

Sgueglia, J.B., S. Geiger, and J. Davis, Precision studies using the ABI prism 3100 genetic analyzer for forensic DNA analysis. *Anal Bioanal Chem*, 2003, **376**(8): 1247–1254.

Siles, B.A., et al., The use of a new gel matrix for the separation of DNA fragments: A comparison study between slab gel electrophoresis and capillary electrophoresis. *Appl Theor Electrophor*, 1996, **6**(1): 15–22.

Sinville, R. and S.A. Soper, High resolution DNA separations using microchip electrophoresis. *J Sep Sci*, 2007, **30**(11): 1714–1728.

Southern, E.M., Measurement of DNA length by gel electrophoresis. *Anal Biochem*, 1979, **100**, 319–323.

Tagliaro, F. and F. Bortolotti, Recent advances in the applications of CE to forensic sciences (2005–2007). *Electrophoresis*, 2008, **29**(1): 260–268.

Tereba, A., K.A. Micka, and J.W. Schumm, Reuse of denaturing polyacrylamide gels for short tandem repeat analysis. *Biotechniques*, 1998, **25**(5): 892–897.

Villablanca, A., et al., Suspension bead array branch migration displacement assay for rapid STR analysis. *Electrophoresis*, 2008, **29**(19): 4109–4114.

Wang, D.Y., et al., Identification and secondary structure analysis of a region affecting electrophoretic mobility of the STR locus SE33. *Forensic Sci Int Genet*, 2012, **6**(3): 310–316.

Wang, Y., et al., Rapid sizing of short tandem repeat alleles using capillary array electrophoresis and energy-transfer fluorescent primers. *Anal Chem*, 1995, **67**(7): 1197–1203.

Wenz, H., et al., High-precision genotyping by denaturing capillary electrophoresis. *Genome Res*, 1998, **8**(1): 69–80.

Westen, A.A., et al., Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing. *J Forensic Sci*, 2009, **54**(3): 591–598.

Yang, J., et al., An integratable microfluidic cartridge for forensic swab samples lysis. *Forensic Sci Int Genet*, 2014, **8**(1): 147–158.

9

Detection Methods

A variety of techniques are available for the detection of DNA fragments in forensic DNA analysis. For example, the direct detection of DNA fragments in a gel can be achieved via staining. The detection of DNA probes in a hybridization-based assay can be performed with radioisotopes, colorimetric assays, and chemiluminescence labeling. For polymerase chain reaction (PCR)-based assays, DNA primers can be labeled directly with fluorescent dyes.

9.1 Direct Detection of DNA in Gels

This section describes two simple and rapid detection methods used for detecting DNA: staining agarose gels with fluorescent intercalating dyes and staining denatured polyacrylamide gels with silver.

9.1.1 Fluorescent Intercalating Dye Staining

The location of DNA in an agarose gel can be determined directly by staining with low concentrations of fluorescent intercalating dyes (Chapter 6) such as ethidium bromide (Figures 9.1 through 9.3). These techniques allow the detection of DNA bands as small as 10 ng in agarose gels. Staining of DNA in agarose gels can be achieved by including an intercalating dye in the gel or staining the gel after electrophoresis in a dye-containing solution, followed by a washing step known as de-staining to reduce nonspecific staining. Historically, ethidium bromide-stained gels were photographed with a standard ultraviolet transilluminator at approximately 300 nm using a camera with an orange filter, whereas gels can currently be documented using digital techniques. When examining a DNA-containing gel under a UV lamp, the eyes and the skin should be protected from UV exposure. Ethidium bromide is a mutagen and a potential carcinogen. Ethidium bromide should be handled according to the Material Safety Data Sheet and safety protection wear should be used while handling the chemical. Ethidium bromide waste is usually disposed of as hazardous waste and is handled in accordance with laboratory guidelines. Alternatives to ethidium bromide are available, for example, fluorescent dyes such as SYBR® stains (Molecular Probes; Figures 9.2 and 9.3). Some of these alternatives are less mutagenic and have better performance than ethidium bromide.

9.1.2 Silver Staining

Electrophoretically separated DNA fragments can also be detected with silver nitrate staining. Silver staining of polyacrylamide gels has been used for the amplified fragment length polymorphism (AFLP; see Chapter 19) method of variable number tandem repeat (VNTR; see

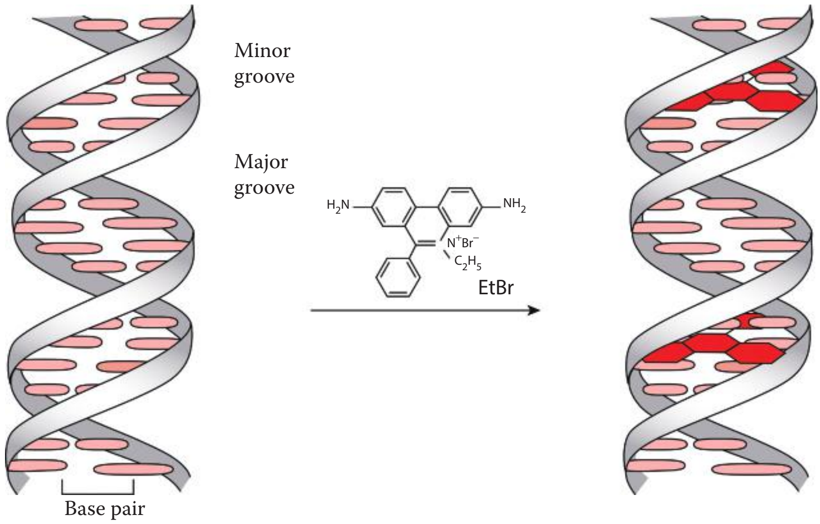

Figure 9.1 Schematic illustration showing binding of an intercalating agent, ethidium bromide (EtBr) to DNA. It intercalates into the minor groove of DNA. (© Richard C. Li.)

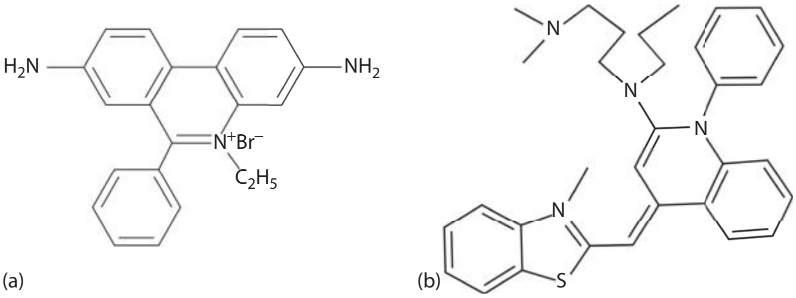

Figure 9.2 Intercalating agents used as nucleic acid stains. (a) Chemical structure of ethidium bromide. Ethidium bromide has UV absorbance. The emission maximum of the DNA–dye complex in aqueous solution is approximately 590 nm. (b) Chemical structure of SYBR Green 1. The stain is a cyanine dye that binds to DNA and is used as a nucleic acid stain. The excitation and emission maxima of the DNA–dye complex are 494 and 521 nm, respectively. The stain preferentially binds to double-stranded DNA rather than single-stranded DNA and RNA.

Chapter 19) profiling. The sensitivity of silver staining is approximately 100 times higher than that obtained with ethidium bromide, and silver staining is less hazardous than ethidium bromide detection. Also, the developing chemicals are readily available at low cost.

Silver staining involves processing a gel followed by exposure to a series of chemicals. First, the gel is submerged in a silver nitrate solution. Silver ions are positively charged, and DNA is negatively charged. Therefore, silver ions bind to the DNA and are subsequently reduced using formaldehyde to form a deposit of metallic silver on the DNA in the gel (Figure 9.4). A photograph of the gel with images of the silver-stained DNA strands is kept as a permanent record. Alternatively, the gels may be sealed and preserved for record purposes. One disadvantage of this method is that silver stains RNA and proteins along with DNA. The presence of restriction enzymes and polymerase should therefore be minimized. Additionally, bands from both complementary DNA strands may be detected in a denatured polyacrylamide gel, which leads to a two-band pattern.

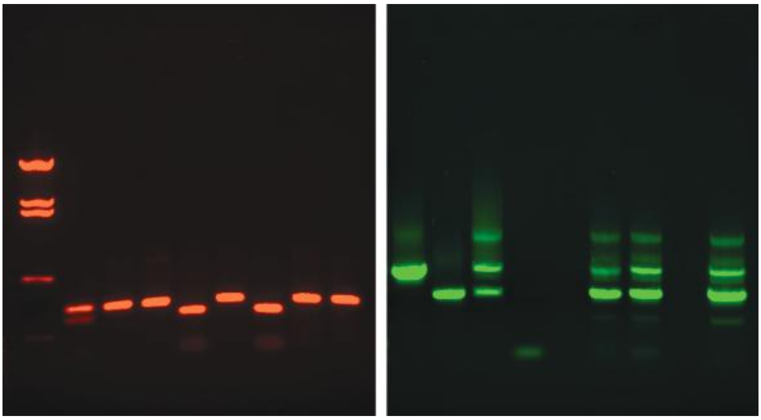

Figure 9.3 A DNA–dye complex emitting fluorescence on UV-light exposure. Ethidium bromide–containing agarose gel (left). SYBR Green 1–containing agarose gel (right). (© Richard C. Li.)

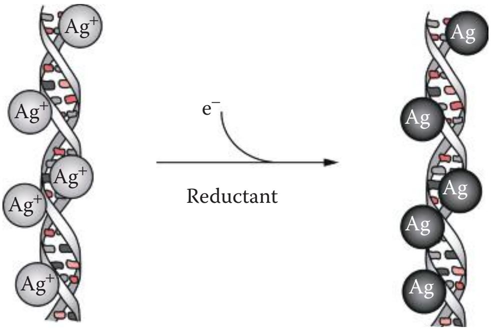

Figure 9.4 Silver staining of DNA. Silver (Ag$^+$) ions bind to DNA and are reduced to metallic silver (Ag) to form dark particles. (© Richard C. Li.)

9.2 Detection of DNA Probes in Hybridization-Based Assays

9.2.1 Radioisotope Labeled Probes

Radioisotope probe labeling was used for early versions of VNTR testing and DNA quantitation. Labeling can be accomplished in several ways (Figure 9.5). For example, nick translation incorporates labeled deoxyribonucleotides (dNTPs) into double-stranded DNA. DNase I is used to introduce single-strand nicks within the DNA fragment to be labeled. Next, DNA Polymerase I recognizes the nicks and replaces the preexisting nucleotides with new strands containing labeled dNTPs, resulting in the generation of ^{32}P-labeled double-stranded DNA molecules. ^{32}P, with a half-life of approximately 14 days, is the most common radioisotope used in this technique. Nick translation can utilize any dNTP labeled with ^{32}P.

Prior to hybridization, the probe is denatured into single-stranded fragments by boiling for a few minutes followed by rapid cooling on ice. After the hybridization process, these probes can be visualized by exposing the DNA-containing membrane to an x-ray film. The radioactive object is commonly placed in an x-ray cassette. The energy released from the decay of the radioisotopes is absorbed by the silver halide grains in the film emulsion and forms a latent image. A chemical development process amplifies the latent image and renders it visible on film (Figure 9.6). Because most ^{32}P emissions pass through the thin film emulsion without

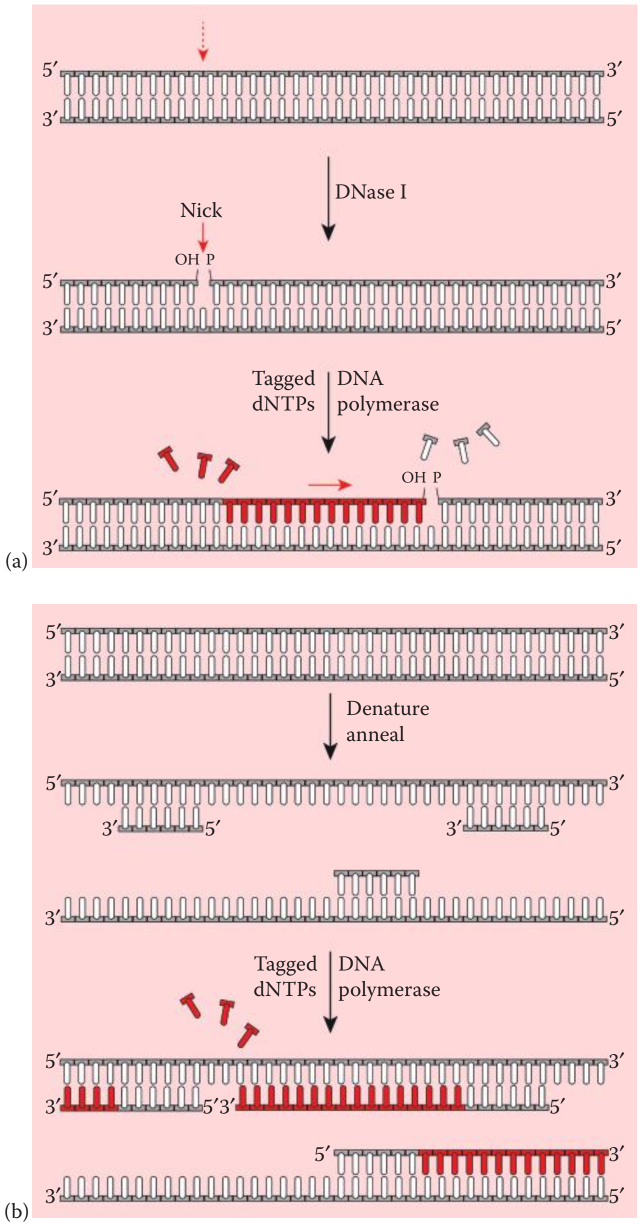
5′
3′
3′
5′
DNase I
Nick
OH P
5′
3′
3′
5′
Tagged dNTPs
DNA polymerase
OH P
5′
3′
3′
5′
(a)
5′
3′
3′
5′
Denature anneal
5′
3′
3′ 5′
3′ 5′
3′
5′
Tagged dNTPs
DNA polymerase
5′
3′
3′ 5′ 3′ 5′
5′ 3′
3′
5′
(b)

Figure 9.5 DNA probe labeling using dNTPs. (a) Nick translation method. A DNase is utilized to create nicks along a DNA strand. Tagged nucleotides (one or more than one of the dNTPs) are then incorporated into the DNA strand using a DNA polymerase. DNA synthesis extends from the 5′ to 3′ end and the original strand is degraded. (b) Random primer labeling method. A template DNA is denatured and separated into single strands. A random mixture of hexameric deoxynucleotide fragments, as primers, is then annealed to the template strand. The tagged nucleotides are then incorporated by the Klenow fragment of the DNA polymerase. (© Richard C. Li.)

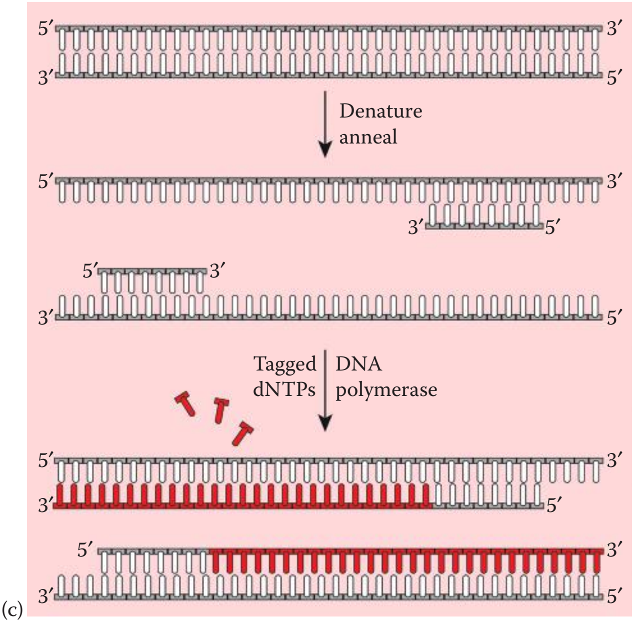

Figure 9.5 (Continued) (c) PCR method. DNA templates are denatured and then annealed to primers flanking the region of interest. During the DNA synthesis phase of a PCR cycle (see Chapter 7), tagged nucleotides are incorporated into the DNA strand using a Taq DNA polymerase, producing new double-stranded DNA. The process is repeated for many cycles. Only the first cycle is shown. (© Richard C. Li.)

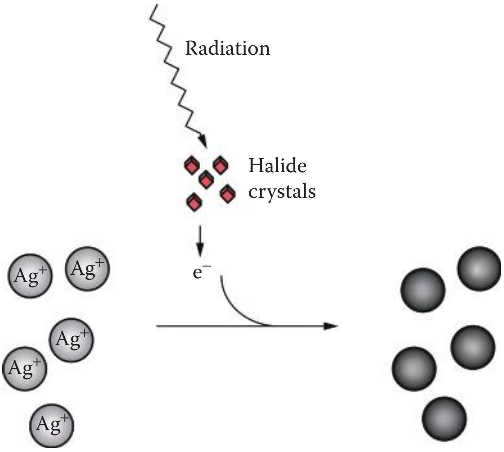

Figure 9.6 Autoradiography. Exposure to radiation causes halide crystals to release electrons, thus reducing silver (Ag^+) ions to metallic silver (Ag). (© Richard C. Li.)

contributing to the final image, the detection process may require long exposure times. Signal intensity can be enhanced, however, by using intensifying screens at low temperatures. The screens emit photons upon receiving radioactive β-particles, thus further enhancing signals. Additionally, highly sensitive x-ray film can be used. One disadvantage of using ^{32}P is that it is a safety hazard. Additionally, autoradiography is a lengthy process. Therefore, nonradioisotopic detection methods have become popular alternatives.

9.2.2 Enzyme-Conjugated Probe with Chemiluminescence Reporting System

The use of alkaline phosphatase (AP)-conjugated probes with chemiluminescent substrates comprises a highly sensitive nonradioisotopic detection system.

Alkaline phosphatase can cleave the phosphate groups from a variety of substrate molecules. Its enzymatic activity can be measured using dioxetane-based chemiluminescent substrates such as Lumigen® PPD (Figure 9.7). The Lumi-Phos Plus kit of Lumigen Inc. contains this substrate and can serve as a detection system for slot blot assays for DNA quantitation (Chapter 6) and RFLP assays for VNTR profiling (Chapter 19). AP catalyzes the cleavage of the phosphate ester of Lumigen® PPD, resulting in the release of a photon (Figure 9.8). The Lumigen® PPD substrate yields a long-lasting light emission that can be detected by exposure to x-ray film. This system provides a highly sensitive chemiluminescent detection method for AP-conjugated DNA probes in solution or on a solid matrix such as a membrane (Table 9.1).

9.2.3 Biotinylation of DNA with Colorimetric Reporting Systems

9.2.3.1 Biotin

Biotin, also known as vitamin H, is a water-soluble molecule found in egg yolk (Figure 9.9). It can be incorporated onto oligonucleotide probes without interfering with the ability of probes to hybridize because of its small size (molecular weight: 244.31 u). Signals from a biotinylated probe can be detected with an enzyme-conjugated avidin system. Two steps are required to detect biotin-labeled probes. First, an avidin conjugate consisting of a reporter enzyme is added. Then, the reporter enzyme is assayed with substrates.

9.2.3.2 Enzyme-Conjugated Avidin

Avidin is a glycoprotein found in egg white; it binds to biotin with extremely high affinity. Thus, a biotin–avidin complex is very stable. However, avidin detection has a high background due to nonspecific binding. The nonspecific binding can be reduced by replacing avidin with its

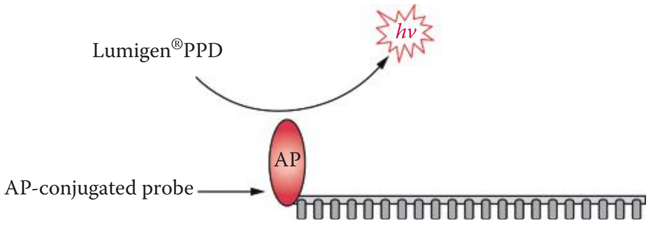

Figure 9.7 Detection system using AP-conjugated probe. Chemiluminescence is generated by using the Lumigen® PPD as an AP substrate. (© Richard C. Li.)

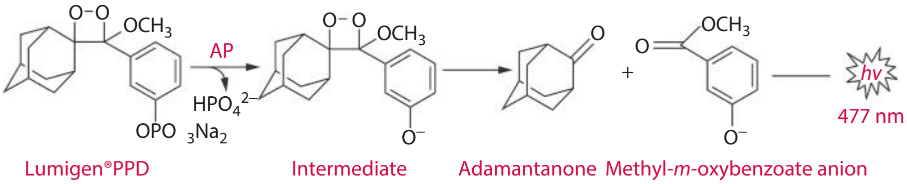

Figure 9.8 Lumi-Phos Plus contains Lumigen® PPD (4-methoxy-4-[3-phosphatephenyl]spiro[1,2-dioxetane-3,2′-adamantane], disodium salt). AP catalyzes the removal of the phosphate group of Lumigen® PPD and generates a chemiluminescent intermediate that is subsequently broken down to the excited state of methyl-*m*-oxybenzoate. The decay of the excited state end-product releases a photon.

Table 9.1 Forensic Applications of Enzyme Reporting Systems for Detecting DNA

Reporter Enzyme	Labeling	Detection Mechanism	Substrate	Forensic Application
AP	AP-conjugated probe	Chemiluminescent	Lumigen®PPD	RFLP; blot assay for DNA quantitation
Horseradish peroxidase (HRP)	Biotinylated probe, recognized by streptavidin conjugated with HRP	Colorimetric/chemiluminescent	TMB/luminol	Blot assay for DNA quantitation
	Biotinylated primer, recognized by streptavidin conjugated with HRP	Colorimetric	TMB	Reverse blot assay for DQA1 typing and for mtDNA typing

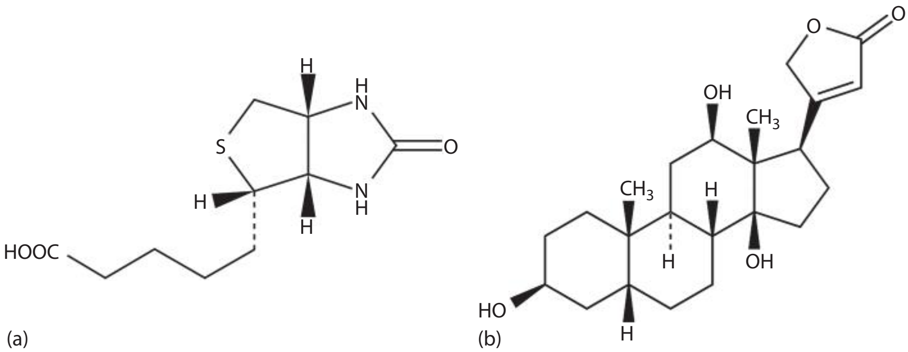

Figure 9.9 Nonradioactive tags for labeling and detecting nucleic acids. (a) Chemical structure of biotin, also known as vitamin H. Molecular formula: $C_{10}H_{16}N_2O_3S$. Molecular weight: 244.31. (b) Chemical structure of digoxigenin. Molecular formula: $C_{23}H_{34}O_5$. Molecular weight: 390.51.

streptavidin counterpart from *Streptomyces avidinii*. To detect binding, an enzyme-conjugated streptavidin such as horseradish peroxidase (HRP)-conjugated streptavidin can be used. HRP isolated from horseradish roots contains heme residues that catalyze the oxidation reactions of substrates (Chapter 12).

9.2.3.3 Reporter Enzyme Assay

HRP can be assayed with colorimetric, chemiluminescent, or fluorogenic substrates. One common colorimetric substrate for forensic DNA testing is 3,3′,5,5′-tetramethylbenzidine (TMB), which is oxidized by the peroxidase to form an insoluble precipitate of intense blue at an acidic pH (Figure 9.10). Because the colored precipitate is difficult to remove from membranes, TMB is not suitable if reprobing for RFLP analysis is required. This technique has been used for forensic DNA testing such as slot blot assays for DNA quantitation (Table 9.1). A chemiluminescent substrate such as a luminol-based reagent can also be utilized with

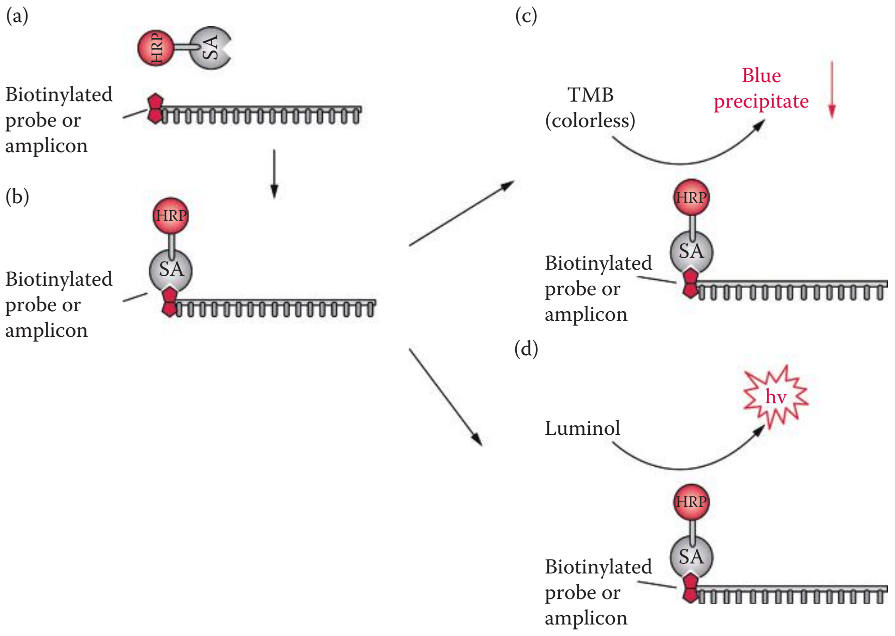
(a)
HRP
SA
Biotinylated probe or amplicon
(b)
HRP
SA
Biotinylated probe or amplicon
(c)
TMB (colorless)
Blue precipitate
HRP
SA
Biotinylated probe or amplicon
(d)
Luminol
hv
HRP
SA
Biotinylated probe or amplicon

Figure 9.10 Detection system using biotinylated DNA probes with colorimetric and chemiluminescent reactions. (a) Biotinylated probe is incubated with a streptavidin (SA) and horseradish peroxidase (HRP) conjugate complex. (b) Biotin is recognized by the complex. Reporter enzyme assays can be carried out using either a colorimetric reaction with a TMB substrate (c) or a chemiluminescent reaction using luminol analogs (d). (© Richard C. Li.)

HRP. The peroxidase catalyzes the oxidation of luminol to form a chemiluminescent product (Figure 9.10).

9.3 Detection Methods for PCR-Based Assays

9.3.1 Fluorescence Labeling

9.3.1.1 Fluorescent Dyes

The advantages of fluorescence detection methods include a higher sensitivity and broader dynamic range than comparable colorimetric detection methods. Furthermore, they have the capacity for simultaneous analysis of complex samples such as multiplex PCR products with different fluorescent labels (Figure 9.11), allowing the distinction of various amplicons. Commonly used fluorescent dyes in DNA labeling emit fluorescence in the range of 400–600 nm (Figures 9.12 and 9.13).

9.3.1.2 Labeling Methods

Fluorescent dye labeling can be incorporated into a DNA fragment using a 5′-end fluorescently labeled oligonucleotide primer (Figures 9.14 and 9.15). The dye-labeled primer method is usually used for STR profiling (see Chapter 20) in which only one primer from each primer pair is labeled; therefore, only one strand can be detected. The two-band pattern observed with silver staining does not appear with this method. Additionally, dye-labeled primers allow multiplex PCR amplifications in the same tube.

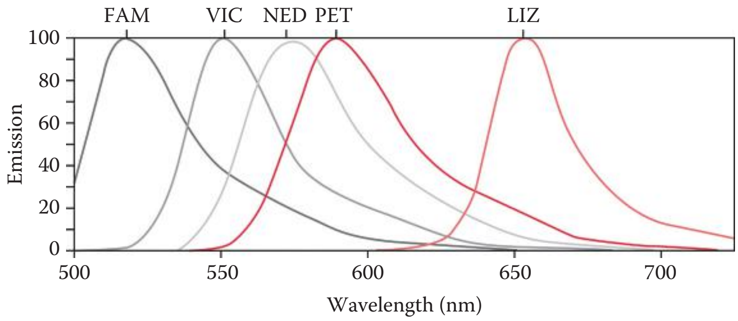

Figure 9.11 Emission spectra of common fluorescent dyes used for forensic DNA analysis. (© Richard C. Li.)

Color	Fluorescent dye	Globalfiler	AmpFlSTR Yfiler	PowerPlex fusion system	PowerPlex Y23 system
Purple dye	SID	☑			
Blue dyes	FAM	☑	☑		
	Fluorescein			☑	☑
Green dyes	JOE			☑	☑
	VIC	☑	☑		
Yellow dyes	NED	☑	☑		
	TMR (TAMRA)			☑	☑
Orange dyes	CC5			☑	☑
	LIZ	☑	☑		
Red dyes	PET		☑		
	ROX (CXR)			☑	☑
	TAZ	☑			

Figure 9.12 Examples of fluorescent dyes and their applications in STR kits. *Note*: Globalfiler and AmpFlSTR Yfiler are manufactured by Applied Biosystems. PowerPlex Fusion system and PowerPlex® Y23 System are manufactured by Promega.

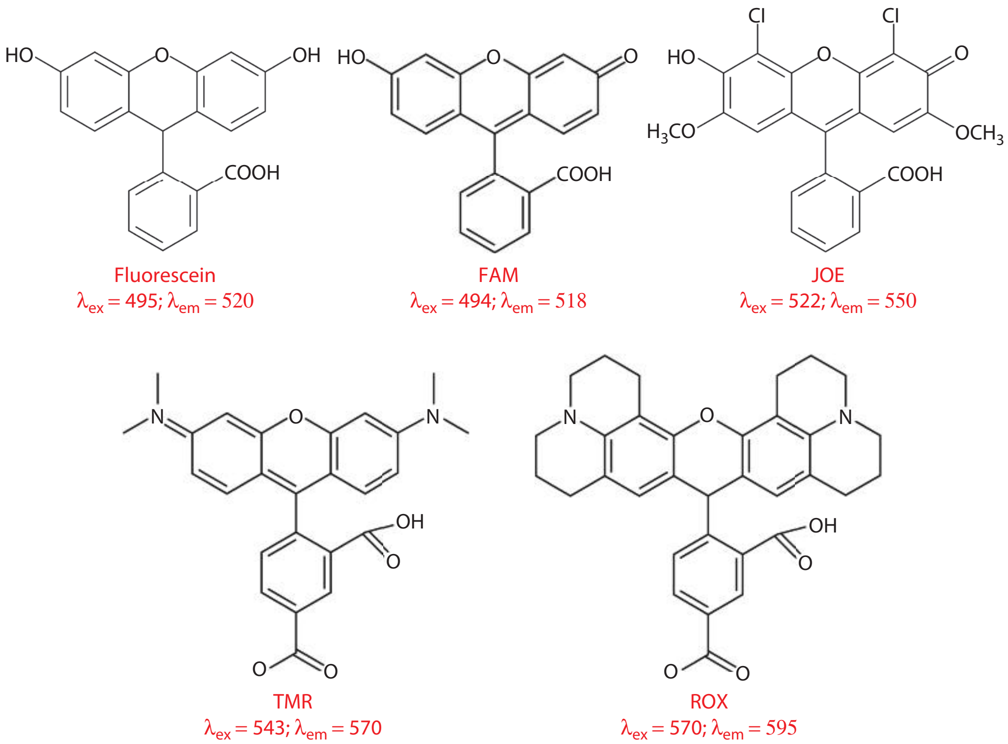
HO
O
OH
COOH
Fluorescein
$\lambda_{ex} = 495; \lambda_{em} = 520$
HO
O
O
COOH
FAM
$\lambda_{ex} = 494; \lambda_{em} = 518$
Cl
Cl
HO
O
O
H_3CO
OCH_3
COOH
JOE
$\lambda_{ex} = 522; \lambda_{em} = 550$
N
O
N
OH
O
O
O
TMR
$\lambda_{ex} = 543; \lambda_{em} = 570$
N
O
N
OH
O
O
O
ROX
$\lambda_{ex} = 570; \lambda_{em} = 595$

Figure 9.13 Chemical structures of representative fluorescent dyes. Fluorescein, 3′,6′-dihydroxyspiro[2-benzofuran-3,9′-xanthene]-1-one; FAM, 5(6)-carboxyfluorescein; JOE, 6-carboxy-4′,5′-dichloro-2′,7′-dimethoxy-fluorescein; TMR, 5-carboxytetramethylrhodamine; ROX, 5(6)-carboxy-X-rhodamine; λ_{ex}, peak excitation wavelength (nm); λ_{em}, peak emission wavelength (nm).

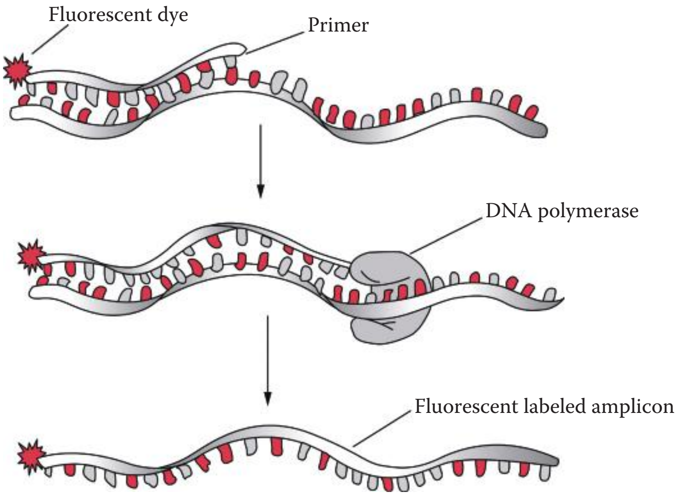
Fluorescent dye
Primer
DNA polymerase
Fluorescent labeled amplicon

Figure 9.14 A fluorescent dye–labeled primer can be used for the amplification of DNA. The dye is conjugated at the 5′ end of the primer. The amplified product is fluorescently labeled. (© Richard C. Li.)

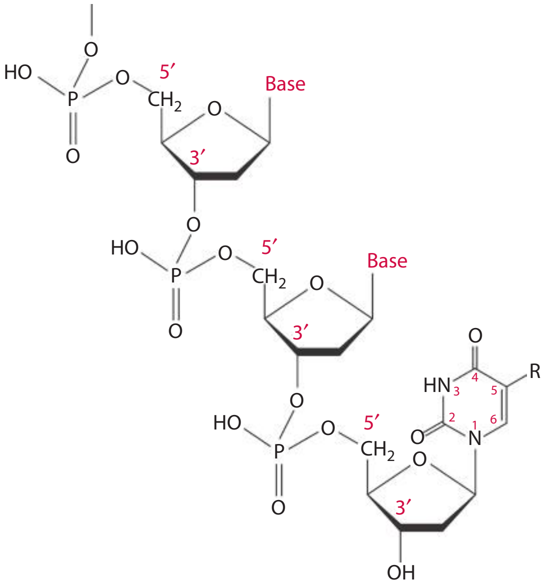

Figure 9.15 Direct labeling of nucleic acids with fluorescent dye. Labeled cytosine is shown as an example. R, fluorescent dye.

Alternatively, fluorescent dye labeling of DNA fragments can be carried out by incorporating fluorescently labeled dideoxynucleotides (ddNTPs) in the PCR product. This labeling method is usually used in DNA sequencing such as mtDNA sequence profiling (Chapter 23).

9.3.1.3 Fluorophore Detection

The *fluorophore* is a component of a florescent dye molecule that causes the molecule to be fluorescent. First, a laser strikes a fluorophore covalently linked to the end of a DNA fragment. An electron of the fluorophore is then excited, rising to an excited state from a ground state. The excited electron then descends to the ground state and releases a photon (Figure 9.16). The emitted photon has a longer wavelength (emission spectrum) than that of the excitation photon (excitation spectrum). The wavelengths of excitation and emission spectra (Figure 9.17) are largely dependent on the chemical structure of the fluorophore.

Lasers are commonly used as *excitation sources* because laser light emissions have high intensity and are monochromatic (single wavelength). The argon ion gas laser is frequently used in applications such as fluorescence-labeled STR and mtDNA sequence analysis because the excitation wavelength of commonly used fluorescent dyes matches the wavelength of the argon laser.

Optical filters are used to filter out undesired light and to allow only one particular wavelength to pass through. An essential optical filter consists of three components: an excitation filter, a dichroic beam splitter, and an emission filter. The excitation filter selectively transmits light from an excitation source. The light is then directed by the dichroic beam splitter to DNA molecules labeled with fluorescent dyes. The light emitted from a fluorophore is also transmitted by the dichroic beam splitter toward the detector. The emission filter selectively blocks undesired light, thus transmitting a specific wavelength of the emitted fluorescence. The light intensity emitted from a fluorophore is detected using a photosensitive device such as a charge-coupled device (CCD). The signal from the fluorophore is collected and converted to an electronic signal expressed in an arbitrary unit such as a *relative fluorescence unit* (RFU). Signals from multiple fluorophores in the same sample can be recorded separately using optical filters and a mathematical *matrix* (fluorophore separation algorithm). The function of a matrix is to subtract the

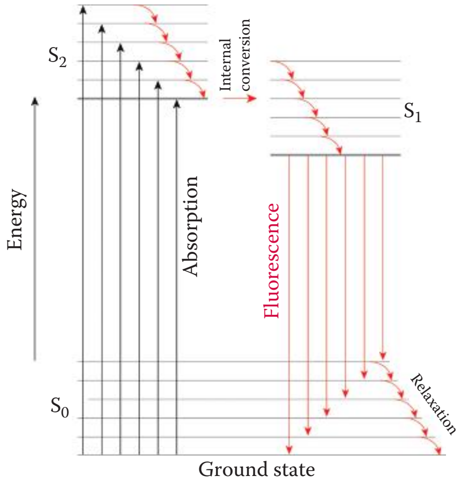

Figure 9.16 A Jablonski diagram illustrating the absorption of a single photon and subsequent emission of fluorescence. During an excitation stage, a single photon with the appropriate energy excites a molecule to the first excited energy level (S_2). In fluorescence reactions, the energy is supplied by an external light source such as a laser, while in chemiluminescence reactions, the energy is derived from a chemical reaction. The excited molecule is not stable and relaxes to its lowest vibrational level in the excited state (S_1). The molecule then returns to the ground state (S_0), emitting a photon. (© Richard C. Li.)

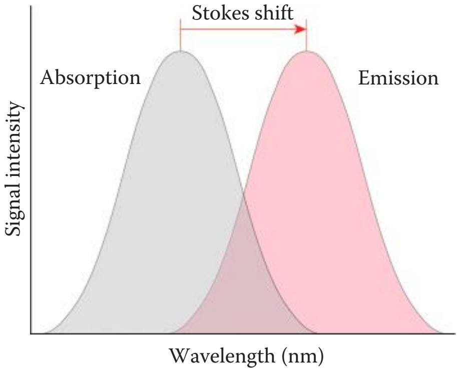

Figure 9.17 A fluorescence excitation spectrum. In fluorescence reactions, the energy of the photon emitted is lower than that of the excitation photon. As a result, the wavelength of an emission photon is longer than that of an excitation photon. The difference in wavelength is known as the Stokes shift. (© Richard C. Li.)

color overlaps of the various fluorescent dyes. The matrix can be established by the calibration of fluorescently labeled standards.

Bibliography

Bassam, B.J., G. Caetano-Anolles, and P.M. Gresshoff, Fast and sensitive silver staining of DNA in polyacrylamide gels. *Anal Biochem*, 1991, **196**(1): 80–83.

Bregu, J., et al., Analytical thresholds and sensitivity: Establishing RFU thresholds for forensic DNA analysis. *J Forensic Sci*, 2013, **58**(1): 120–129.

Bronstein, I. and P. McGrath, Chemiluminescence lights up. *Nature*, 1989, **338**(6216): 599–600.

Bronstein, I., et al., Rapid and sensitive detection of DNA in Southern blots with chemiluminescence. *Biotechniques*, 1990, **8**(3): 310–314.

Brunk, C.F. and L. Simpson, Comparison of various ultraviolet sources for fluorescent detection of ethidium bromide-DNA complexes in polyacrylamide gels. *Anal Biochem*, 1977, **82**(2): 455–462.

Budowle, B., et al., Simple protocols for typing forensic biological evidence: Chemiluminescent detection for human DNA quantitation and restriction fragment length polymorphism (RFLP) analyses and manual typing of polymerase chain reaction (PCR) amplified polymorphisms. *Electrophoresis*, 1995, **16**(9): 1559–1567.

Buel, E. and M. Schwartz, The use of DAPI as a replacement for ethidium bromide in forensic DNA analysis. *J Forensic Sci*, 1995, **40**(2): 275–278.

Cardullo, R.A., et al., Detection of nucleic acid hybridization by nonradiative fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A*, 1988, **85**(23): 8790–8794.

Decorte, R. and J.J. Cassiman, Evaluation of the ALF DNA sequencer for high-speed sizing of short tandem repeat alleles. *Electrophoresis*, 1996, **17**(10): 1542–1549.

Feinberg, A.P. and B. Vogelstein, A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem*, 1983, **132**(1): 6–13.

French, B.T., H.M. Maul, and G.G. Maul, Screening cDNA expression libraries with monoclonal and polyclonal antibodies using an amplified biotin-avidin-peroxidase technique. *Anal Biochem*, 1986, **156**(2): 417–423.

Gillespie, P.G. and A.J. Hudspeth, Chemiluminescence detection of proteins from single cells. *Proc Natl Acad Sci U S A*, 1991, **88**(6): 2563–2567.

Giusti, A.M. and B. Budowle, A chemiluminescence-based detection system for human DNA quantitation and restriction fragment length polymorphism (RFLP) analysis. *Appl Theor Electrophor*, 1995, **5**(2): 89–98.

Green, N.M., Avidin. *Adv Protein Chem*, 1975, **29**: 85–133.

Green, N.M. and E.J. Toms, The properties of subunits of avidin coupled to sepharose. *Biochem J*, 1973, **133**(4): 687–700.

Halpern, M.D. and J. Ballantyne, An STR melt curve genotyping assay for forensic analysis employing an intercalating dye probe FRET. *J Forensic Sci*, 2011, **56**(1): 36–45.

Johnson, E.D. and T.M. Kotowski, Chemiluminescent detection of RFLP patterns in forensic DNA analysis. *J Forensic Sci*, 1996, **41**(4): 569–578.

Laskey, R.A. and A.D. Mills, Enhanced autoradiographic detection of 32P and 125I using intensifying screens and hypersensitized film. *FEBS Lett*, 1977, **82**(2): 314–316.

Lee, L.G., et al., Seven-color, homogeneous detection of six PCR products. *Biotechniques*, 1999, **27**(2): 342–349.

LePecq, J.B. and C. Paoletti, A fluorescent complex between ethidium bromide and nucleic acids. Physical-chemical characterization. *J Mol Biol*, 1967, **27**(1): 87–106.

Mansfield, E.S. and M.N. Kronick, Alternative labeling techniques for automated fluorescence based analysis of PCR products. *Biotechniques*, 1993, **15**(2): 274–279.

Mayrand, P.E., et al., The use of fluorescence detection and internal lane standards to size PCR products automatically. *Appl Theor Electrophor*, 1992, **3**(1): 1–11.

McKimm-Breschkin, J.L., The use of tetramethylbenzidine for solid phase immunoassays. *J Immunol Methods*, 1990, **135**(1–2): 277–280.

Merril, C.R., Silver staining of proteins and DNA. *Nature*, 1990, **343**(6260): 779–780.

Mitchell, L.G., A. Bodenteich, and C.R. Merril, Use of silver staining to detect nucleic acids. *Methods Mol Biol*, 1994, **31**: 197–203.

Morin, P.A. and D.G. Smith, Nonradioactive detection of hypervariable simple sequence repeats in short polyacrylamide gels. *Biotechniques*, 1995, **19**(2): 223–228.

Renz, M. and C. Kurz, A colorimetric method for DNA hybridization. *Nucleic Acids Res*, 1984, **12**(8): 3435–3444.

Roberts, I.M., et al., A comparison of the sensitivity and specificity of enzyme immunoassays and time-resolved fluoroimmunoassay. *J Immunol Methods*, 1991, **143**(1): 49–56.

Selvin, P.R., Fluorescence resonance energy transfer. *Methods Enzymol*, 1995, **246**: 300–334.

Southern, E.M., Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, 1975, **98**(3): 503–517.

Stone, T. and I. Durrant, Enhanced chemiluminescence for the detection of membrane-bound nucleic acid sequences: Advantages of the Amersham system. *Genet Anal Tech Appl*, 1991, **8**(8): 230–237.

Tizard, R., et al., Imaging of DNA sequences with chemiluminescence. *Proc Natl Acad Sci U S A*, 1990, **87**(12): 4514–4518.

Tuma, R.S., et al., Characterization of SYBR Gold nucleic acid gel stain: A dye optimized for use with 300-nm ultraviolet transilluminators. *Anal Biochem*, 1999, **268**(2): 278–288.

Waring, M.J., Complex formation between ethidium bromide and nucleic acids. *J Mol Biol*, 1965, **13**(1): 269–282.

Watkins, T.I. and G. Woolfe, Effect of changing the quaternizing group on the trypanocidal activity of dimidium bromide. *Nature*, 1952, **169**(4299): 506–507.

Yeung, S.H., et al., Fluorescence energy transfer-labeled primers for high-performance forensic DNA profiling. *Electrophoresis*, 2008, **29**(11): 2251–2259.

10

Serology Concepts

10.1 Serological Reagents

10.1.1 Immunogens and Antigens

A foreign substance that is capable of eliciting antibody formation when introduced into a host is called an *immunogen*. Natural immunogens are usually macromolecules such as proteins and polysaccharides. Other molecular structures can also act as immunogens, for example, glycolipids (such as A, B, and O blood group antigens) and glycoproteins (such as Rh and Lewis antigens). However, they must be foreign to their hosts. The molecular structure of an immunogen, usually a small portion recognized by an antibody, is called an *epitope* or determinant site. An immunogen usually consists of multiple epitopes and is thus considered *multivalent* (Figure 10.1). Each epitope can elicit the production of its own corresponding antibody.

An *antigen* is a foreign substance that is capable of reacting with an antibody. All immunogens can be considered antigens, but not all antigens can elicit antibody formation. *Hapten* is one example of a substance that is antigenic but not immunogenic. Haptens are chemical compounds that are too small to elicit antibody production when they are introduced to a host animal. However, a hapten can be coupled to a *carrier*, usually a macromolecule, to produce antibodies. A hapten-conjugated carrier can become immunogenic to elicit the formation of an antibody specific to the hapten. The resulting antibody can bind to free haptens. Certain controlled substances such as cocaine and amphetamines are haptens and can be detected through corresponding antibodies for forensic toxicological analysis.

10.1.2 Antibodies

Antibodies, also known as *immunoglobulins*, are capable of binding specifically to antigens and are designated with an Ig prefix. The five major classes of immunoglobulins are designated IgG, IgA, IgM, IgD, and IgE. Additionally, IgG immunoglobulins can be further divided into four subclasses (IgG1–IgG4) and IgA immunoglobulins can be further divided into two subclasses (IgA1 and IgA2). Thus, there are a total of nine immunoglobulin isotypes in humans. *Isotypes* are the immunoglobulins that differ based on the molecular variations in the constant domains of the heavy and light chains (Figure 10.2). IgD, IgE, and IgG are usually monomers. IgM can be a membrane-bound monomer or a cross-linked pentamer (secreted form). IgA can be a monomer, dimer, or trimer. In an immune response, an initial exposure to an immunogen elicits a primary response, producing IgM immunoglobulins. Further exposure to the immunogen can elicit a secondary response, producing IgG, IgA, IgE, and IgD immunoglobulins. IgG is the most abundant immunoglobulin in serum. A majority of serology tests are based on the IgG immunoglobulins.

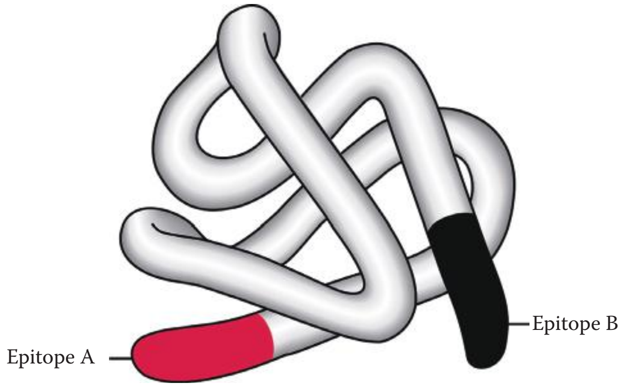

Figure 10.1 Multivalent immunogen. A protein with two different epitopes is shown. (© Richard C. Li)

Immunoglobulins have many similarities in their molecular structures. Figure 10.2 illustrates a diagram of the IgG molecule. The structures of immunoglobulins were first revealed by Gerald Edelman and Rodney Porter who shared a Nobel Prize in 1972. Immunoglobulins are composed of four polypeptide chains: two heavy (H) chains and two light (L) chains. The polypeptide chains are linked by disulfide bonds into a Y-shaped complex. The H chain can be divided into fragment antigen-binding (*Fab*) and fragment crystallizable (*Fc*) fragments. The L chain consists of a Fab fragment only. A typical antibody has two identical antigen-binding sites and is thus considered *bivalent*. The antigen-binding activity is located within the Fab fragments. In particular, the N-terminal ends of the L and H chains together form antigen-binding sites. At the amino acid sequence level, both H and L chains have *variable and constant domains* (Figure 10.2). The variable domains are located at the N-terminal ends of the immunoglobulins. Additionally, three small *hypervariable regions* are located within the variable domain of each chain.

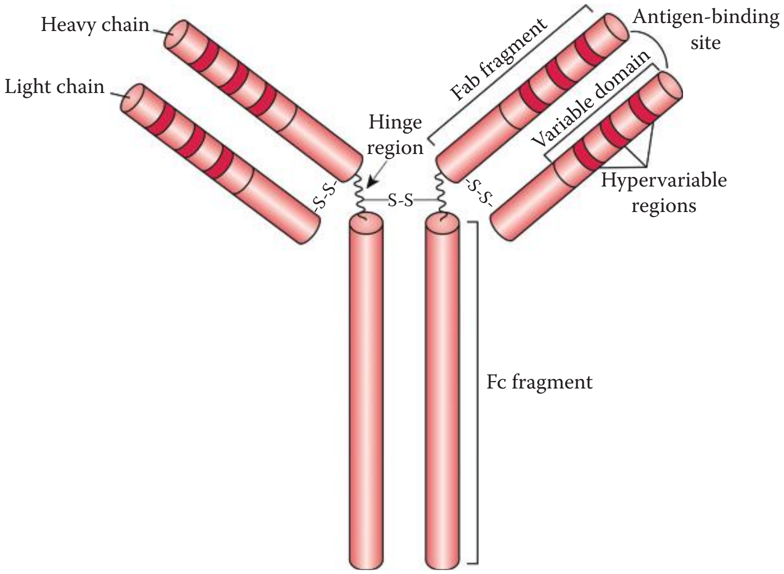

Figure 10.2 Immunoglobulin IgG structure. Immunoglobulin IgG is composed of light and heavy chains that contain variable domains. The remaining portions of chains form the constant domains. The variable domains of both light and heavy chains contain three hypervariable regions that form the antigen-binding site of the immunoglobulin. (© Richard C. Li.)

The diversity in the amino acid sequences of these hypervariable regions determines the specificity of the antigen-binding sites. The hinge regions provide flexibility to the antibody molecule and are important for the efficiency of the binding and cross-linking reactions. The basal portion of the H chain consists of *Fc* fragment.

The binding affinity and specificity of antibodies make them useful reagents for serological testing. Two types of antibodies are commonly used: polyclonal and monoclonal antibodies.

10.1.2.1 Polyclonal Antibodies

To produce an antibody, an immunogen is usually introduced into a host animal. A multivalent immunogen is capable of eliciting a mixture of antibodies with diverse specificities for the immunogen. As a result, a polyclonal antibody is produced by different B lymphocyte clones in response to the different epitopes of the immunogen. Antibodies can be circulating (in blood or other bodily fluids) or tissue bound (in cell surface antibodies). Circulating immunoglobulins are referred to as humoral antibodies.

The blood from an immune host is drawn and allowed to clot, resulting in the formation of a solid consisting largely of blood cells and a liquid portion known as *serum* containing antibodies (Figure 10.3). Such a preparation of humoral antibodies is also called a *polyclonal antiserum.* Depending on the type of animal used, the antibodies produced are classified as avian (B), rabbit (R), or horse (H) type. The characteristics of polyclonal antibodies may vary if they are produced from different individual host animals of the same species. Variations in reactions among different sources of antibodies should be monitored by quality-control procedures.

10.1.2.2 Monoclonal Antibodies

To produce a monoclonal antibody, spleen cells are harvested from a host animal, such as a mouse, inoculated with an immunogen (Figure 10.4). Next, the plasma cells of the spleen, which produce antibodies, are fused with myeloma cells. Since only a small population of cells fuse, a selection step is needed to allow only fused cells to grow (Figure 10.4). The fused cells, called *hybridoma* cells, are immortal (proliferate indefinitely) in cell cultures. Pools of hybridoma cells are diluted into single clones and are allowed to proliferate. The clones are then screened for the specific antibody of interest.

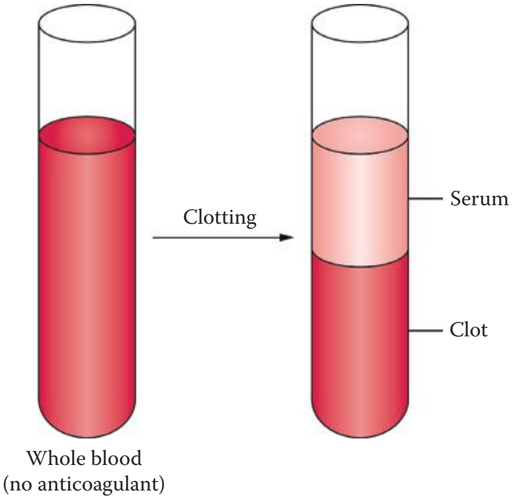

Figure 10.3 Serum component of blood. The blood of an immunized animal is collected in the absence of an anticoagulant and is allowed to clot. The resulting liquid portion of the blood is serum. (© Richard C. Li.)

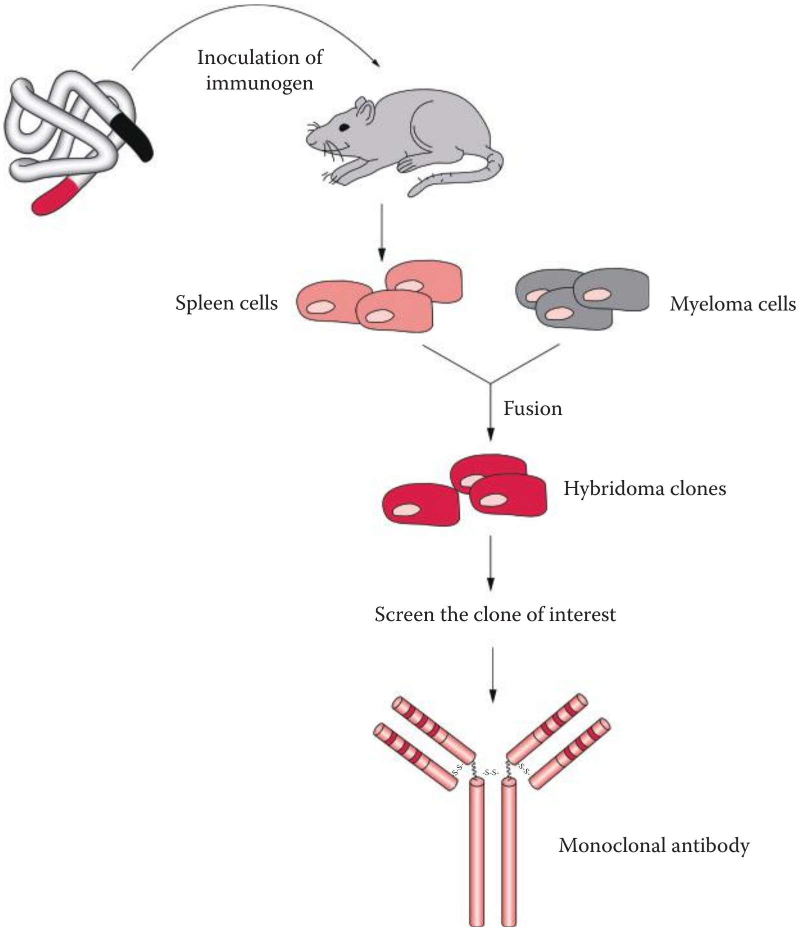

Figure 10.4 Preparation of a monoclonal antibody. A mouse is immunized with an immunogen, and its spleen cells are fused with myeloma cells to generate hybridoma cells. The clone that synthesizes and secretes the monoclonal antibody of interest can then be identified. (© Richard C. Li.)

The desired hybridoma clone can be maintained indefinitely, and it produces a monoclonal antibody that reacts with a single epitope. The hybridoma-derived monoclonal antibodies are specific and homogenous, and can be obtained in unlimited quantities. Monoclonal antibodies have been utilized in many serological assays, as discussed in Chapter 11. However, they have certain limitations in serology assays. For instance, monoclonal antibodies react with only a single epitope of a multivalent antigen and, therefore, cannot form cross-linked networks in precipitation assays (see Section 10.3.2.1).

10.1.2.3 Antiglobulins

Immunoglobulins are proteins that can also function as immunogens. If a purified foreign immunoglobulin or a fragment of a foreign immunoglobulin is introduced into a host, the antibodies produced are known as *antiglobulins*. Antiglobulins that are specific to a particular isotype can be produced in laboratories. In addition to specific antiglobulins, it is possible to produce nonspecific antiglobulins, which recognize an epitope that is common to all isotypes of an immunoglobulin class, such as the Fc portion of the heavy chain of all subclasses of human IgG. Antiglobulins are important reagents in many serological tests.

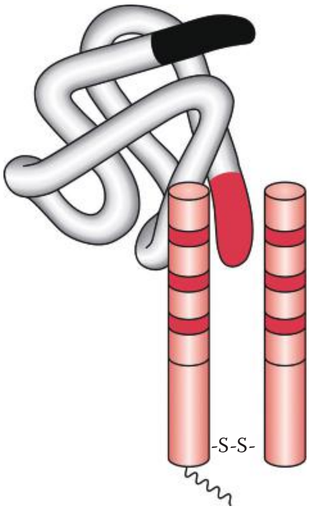

Figure 10.5 Affinity is a measure of the interaction between a single epitope on an antigen and a single binding site on an antibody. (© Richard C. Li.)

10.2 Strength of Antigen–Antibody Binding

The binding of an antigen to its specific antibody is mediated by the interaction between the epitope of an antigen and the binding site of its antibody. Noncovalent bonds can be formed during antigen–antibody binding. Various forces act cooperatively during antigen–antibody binding. These include hydrogen bonding, hydrophobic interactions that exclude water molecules from the area of contact, and van der Waals forces arising from the asymmetric distribution of the charges of electrons. The binding process occurs rapidly and the formation of the antigen–antibody complex is reversible. Such binding occurs at short distances when the antigen and antibody are in close proximity. Additionally, the strongest binding occurs only if the shape of the epitope fits the binding site of the antibody. The strength of the interaction between the antigen and the antibody depends on two characteristics, designated affinity and avidity.

Affinity is the energy of the interaction of a single epitope on an antigen and a single binding site on a corresponding antibody (Figure 10.5). The strength of the interaction depends on the specificity of the antibody for the antigen. Nevertheless, antibodies can bind with lower strength to antigens that are structurally similar to the immunogen. Such binding is known as *cross-reaction*.

Avidity is the overall strength of the binding of an antibody and an antigen (Figure 10.6). Since an antigen is usually considered multivalent and an antibody is bivalent, the avidity reflects the combined synergistic strength of the binding of all the binding sites of antigens and antibodies rather than the sum of individual affinities. It also reflects the overall stability of an antigen–antibody complex that is essential for many serological assays.

10.3 Antigen–Antibody Binding Reactions

The binding of an antigen to an antibody is an equilibrium reaction consisting of three types of reactions. The primary and secondary reactions form the basis for many forensic serological assays and will be discussed in the following subsections. The third type is called the tertiary reaction. It is used to measure *in vivo* immune responses such as inflammation and phagocytosis. Because most forensic serology tests are *in vitro* assays, the tertiary reaction is not commonly utilized in forensic serology testing and will not be discussed here.

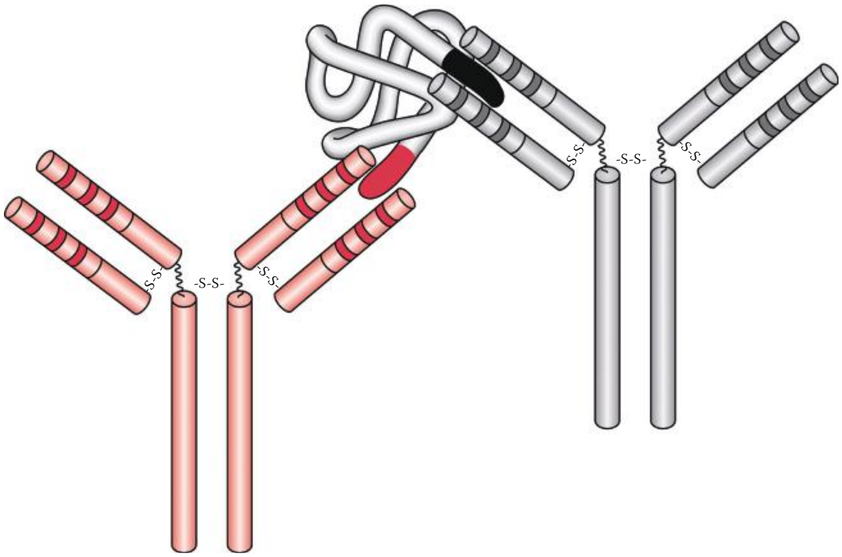

Figure 10.6 Avidity is a measure of the overall strength of the binding between antigens and antibodies. (© Richard C. Li.)

10.3.1 Primary Reactions

A primary reaction is the initial binding of a single epitope of an antigen (Ag) and a single binding site of an antibody (Ab) to form an antigen–antibody complex (Figure 10.7). This rapid and reversible binding reaction can be expressed as

$$\mathrm{Ag} + \mathrm{Ab} \rightleftarrows \mathrm{AgAb}$$

At equilibrium, the strength of the interaction can be expressed as the affinity constant (K_a) that reflects the affinity of binding, where:

$$K_\mathrm{a} = \frac{[\mathrm{AgAb}]}{[\mathrm{Ag}][\mathrm{Ab}]}$$

The square brackets indicate the concentration of each component at equilibrium. K_a is the reciprocal of the concentration of free epitopes when half the antibody-binding sites are occupied. Thus, a higher K_a corresponds to a stronger binding interaction.

Techniques such as enzyme immunoassays, immunofluorescence assays, radioimmunoassays, and dye-labeled immunochromatography can measure the concentrations of antigen–antibody complexes formed by primary reactions (Chapter 11). These techniques are the most sensitive for detecting the presence of an antigen and an antibody in a sample. Additionally, many forensic serology assays are based on the detection of primary reactions and will be discussed in Chapters 12 through 17.

10.3.2 Secondary Reactions

The primary reaction between an antigen and an antibody is often followed by a secondary reaction. The three types of secondary reactions are precipitation, agglutination, and complement fixation. The techniques that detect secondary reactions are usually less sensitive but easier to perform than primary reaction assays. The precipitation and agglutination reactions form the basis for many serologic assays performed in forensic laboratories. These reactions will be discussed in the following subsections in detail. The third type of reaction is called *complement fixation*. If an antigen is located on a cell surface, the binding of the antigen and the antibody

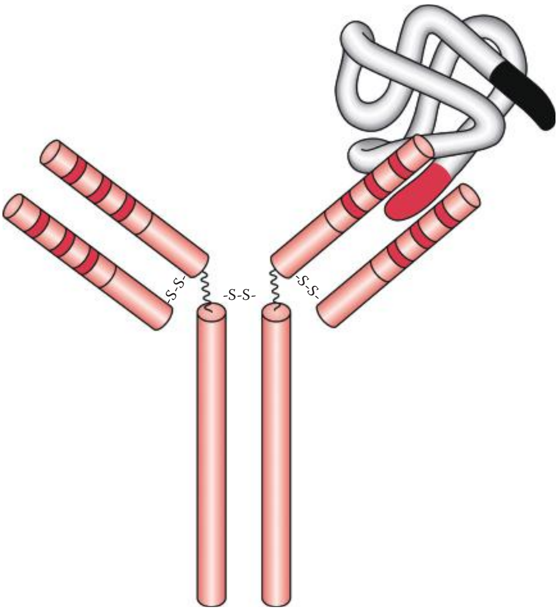

Figure 10.7 Primary reaction. Initial binding forms an antigen–antibody complex. (© Richard C. Li.)

may activate the classical complement pathway and lead to cell lysis, also known as a complement fixation reaction. The detection of this type of reaction is not commonly used in forensic serology.

10.3.2.1 Precipitation

If a soluble antigen is mixed and incubated with its antibody, the antigen–antibody complexes can form cross-linked complexes at the optimal ratio of antigen-to-antibody concentration. The cross-linked complex is insoluble and eventually forms a precipitate that settles to the bottom of a test tube. Antibodies that produce such precipitation are also called *precipitins*.

This precipitation reaction can be characterized by examining the effect of varying the relative ratio of antigen and antibody. If an increasing amount of soluble antigen is mixed with a constant amount of antibody, the amount of precipitate formed can be plotted. A precipitin curve (Figure 10.8)

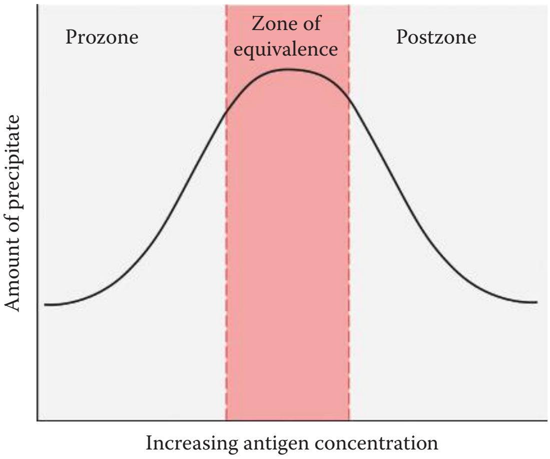

Figure 10.8 Precipitin curve. (© Richard C. Li.)

illustrates the results observed when antigens and antibodies are mixed in various concentration ratios. The curve can be divided into three zones known as the prozone, the zone of equivalence, and the postzone.

10.3.2.1.1 Prozone

At this zone, the ratio of antigen–antibody concentration is low. In other words, the antibody is in excess. Each antigen molecule is rapidly saturated with antibody, thus preventing cross-linking (Figure 10.9a). No precipitate is formed at the prozone stage.

10.3.2.1.2 Zone of Equivalence

As the concentrations of antigen increase, the amount of precipitate increases until it reaches a maximum. The amount of precipitation depends on the relative proportions of antigens and antibodies present. The maximum precipitation occurs in what is called the zone of equivalence. In the zone of equivalence, the ratio of antibody to antigen concentration is optimal and

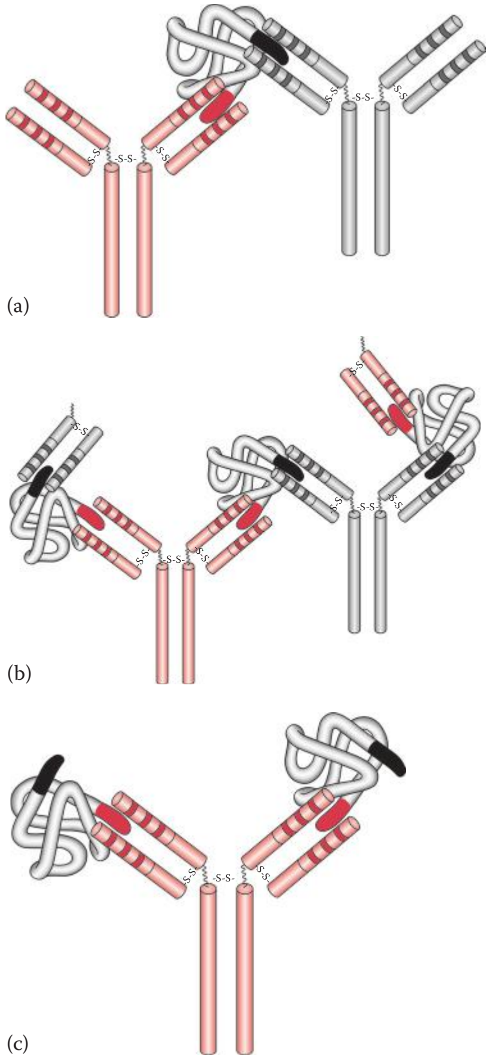

Figure 10.9 Antigen–antibody binding in (a) prozone, (b) zone of equivalence, and (c) postzone. (© Richard C. Li.)

precipitation occurs as a result of forming cross-linked networks (Figure 10.9b). Precipitation assays are usually carried out under the condition of the zone of equivalence, forming a sufficient quantity of precipitation to be detected.

10.3.2.1.3 Postzone

With the addition of more antigens, the ratio of antigen–antibody concentrations is high. In other words, the antigen is in excess. The amount of precipitate decreases and eventually diminishes. Each antibody molecule is saturated with antigen molecules (Figure 10.9c). Cross-linkage cannot form, and precipitation does not occur.

10.3.2.2 Agglutination

As discussed in the previous subsection, precipitation reactions involve soluble antigens. If the antigens are located on the surfaces of cells or carriers (carrier cells such as sheep erythrocytes,

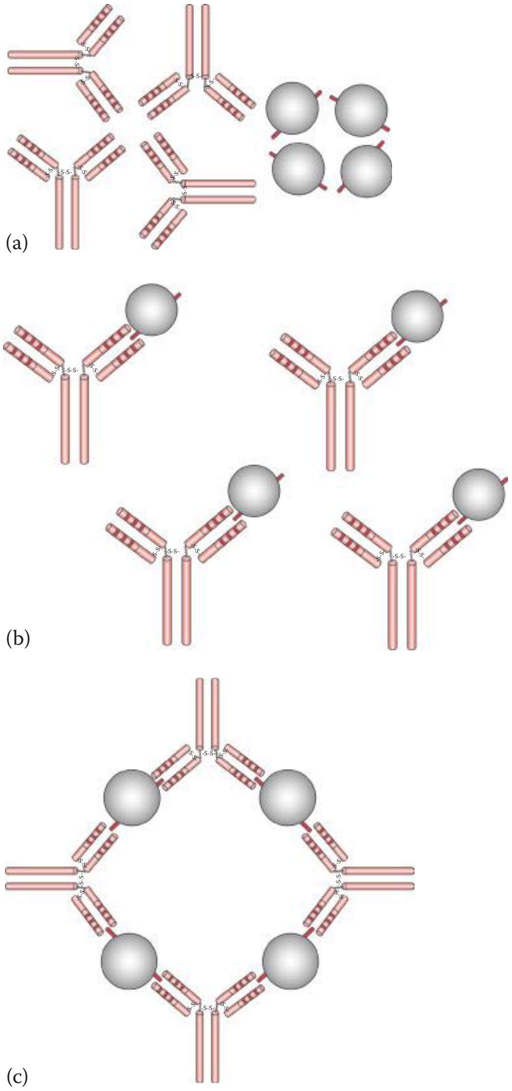

Figure 10.10 Agglutination reaction: (a) antigens are mixed with antibodies, (b) an antigen–antibody complex is formed during initial binding, and (c) the lattice is formed. (© Richard C. Li.)

bacteria, or latex particles), antibodies can bind to the surface antigens and can form cross-links among cells or carriers, causing them to aggregate. This aggregation is referred to as *agglutination*. In agglutination, a visible clumping can be observed as an indicator of the reaction of the antigen and the antibody. If the antigen is located on an erythrocyte, the agglutination reaction is designated *hemagglutination*. Agglutination is a two-step process that includes initial binding and lattice formation (Figure 10.10a).

10.3.2.2.1 Initial Binding

The first step of the reaction involves antigen–antibody binding at a single epitope on the cell surface (Figure 10.10b). This initial binding is rapid and reversible.

10.3.2.2.2 Lattice Formation

The second step involves the formation of a cross-linked network resulting in visible aggregates that constitute a lattice (Figure 10.10c). This involves an antibody binding to multiple epitopes because each antibody has two binding sites and antigens are multivalent. Lattice formation is a much slower process than the initial binding step. The cross-linking of cells requires physical contact. Additionally, an antibody must bind to epitopes on two different cells. The ability to cross-link cells depends on the nature of the antibody. Antibodies that produce such reactions are often called *agglutinins*.

Additionally, a *complete antibody* is capable of carrying out both primary and secondary interactions that result in agglutination. An antibody that can carry out initial binding but fails to form agglutination is called an *incomplete antibody*. This type of antibody is believed to have only one active antigen-binding site and is thus not capable of agglutination. It is potentially caused by the presence of steric obstruction due to the conformation of the incomplete antibody molecule, preventing the binding of antigens at the second binding site. However, other incomplete antibodies have two active sites but cannot bridge the distance between cells, thus failing to form lattices. Certain antibodies such as IgG are small and lack flexibility at the hinge region, and this may prevent agglutination. In contrast, the large IgM antibodies produce agglutination much more easily than IgG. Agglutination reactions have a wide variety of applications in the detection of antigens and antibodies. Such assays have high degrees of sensitivity and have been used for many years in forensic serology.

Bibliography

Baxter, S.J. and B. Rees, The use of anti-human haemoglobin in forensic serology. *Med Sci Law*, 1974, **14**(3): 159–162.

Dodd, B.E., Some recent advances in forensic serology. *Med Sci Law*, 1972, **12**(3): 195–199.

Fletcher, S. and M.J. Davie, Monoclonal antibody: A major new development in immunology. *J Forensic Sci Soc*, 1980, **20**(3): 163–167.

Giusti, G.V., Leone Lattes: Italy's pioneer in forensic serology. *Am J Forensic Med Pathol*, 1982, **3**(1): 79–81.

Goldsby, R., Kindt, T, and Osborne, B., *Immunology*, 6th edn. 2007, New York: W.H. Freeman and Company.

Lee, H.C., et al., Bits and pieces: Serology, criminalistics, anthropology and odontology. *J Forensic Sci Soc*, 1991, **31**(2): 293–296.

Patzelt, D., History of forensic serology and molecular genetics in the sphere of activity of the German Society for Forensic Medicine. *Forensic Sci Int*, 2004, **144**(2–3): 185–191.

Tietz, N. (ed.), *Fundamentals of Clinical Chemistry*, 1987. Philadelphia, PA: W.B. Saunders Company.

11

Serology Techniques

Past, Current, and Future

11.1 Introduction to Forensic Serology

11.1.1 The Scope of Forensic Serology

Forensic serology is the component of forensic biology that deals with the examination and identification of biological evidence. In particular, it focuses on determining the presence and identification of various bodily fluids such as blood, semen, and saliva in a questioned sample.

Bodily fluid stains are commonly associated with violent criminal cases. For instance, the identification of blood evidence (Chapter 12) is often necessary for investigating cases involving homicide, aggravated assault, sexual assault, and burglary. Proving the presence of blood has probative value or may corroborate allegations of violent acts. This evidence can also be used for investigative purposes. For example, the forensic DNA analysis of evidence such as a victim's blood on a suspect's weapon can establish a link between a victim and a suspect. The identification of semen (Chapter 14) and saliva (Chapter 15) is important for the investigation of a sexual assault case. For instance, a stain from a victim's clothing would be processed with forensic DNA analysis in a sexual assault case so that the DNA profile of the stain could be compared with that of an alleged suspect. Matching DNA profiles prove that the suspect's DNA was found on the victim's clothing, establishing a link between the suspect and the victim. Additionally, the identification of a biological stain as a semen stain through forensic serology testing proves that semen was found on the clothing taken from the victim and that the suspect is the source of the DNA from the semen stain. This evidence can then be used in court to support allegations that a sexual act occurred. Likewise, the presence of a suspect's saliva stains on a victim's genital area may corroborate an alleged oral copulation. The identification of bodily fluids other than blood, semen, and saliva will be discussed in Chapters 16 and 17.

11.1.2 Class Characteristics and Individual Characteristics of Biological Evidence

The identification of an unknown fluid sample is based on a comparison of the class characteristics (Section 3.2.1) of a sample with known standards of its class. Forensic identification typically involves bodily fluids, such as blood, semen, and saliva. If the presence of the bodily fluid is confirmed, the individual characteristics (Section 3.2.2) of the biological evidence are then determined to find out whether or not a bodily fluid sample has come from a particular individual. Today, this analysis can be achieved through forensic DNA analysis. The most commonly utilized forensic DNA analysis is the short tandem repeat (STR) analysis used for human identification. Additionally, Y-chromosomal STR analysis is often utilized for the investigation of sexual assault crimes. Furthermore, mitochondrial DNA analysis is used for the identification of human remains. Therefore, establishing the probative value of a sample requires both the identification of its class characteristics and the individualization of its contributor. Therefore, the identification of bodily fluids cannot be replaced by forensic DNA analysis.

11.1.3 Presumptive and Confirmatory Assays

The identification of bodily fluids can be carried out using presumptive and confirmatory assays to identify the type of bodily fluid in question. The advantages of *presumptive assays* are that these assays are sensitive, rapid, and simple. A positive reaction of a presumptive assay indicates the possibility of the presence of the bodily fluid in question. However, presumptive assays are not very specific. Therefore, they should not be considered conclusive for the presence of a type of bodily fluid. In contrast, a negative assay indicates that the questioned bodily fluid is absent. Thus, presumptive assays can be used as a screening method and for narrowing down bodily fluid stains prior to other types of analyses, such as forensic DNA testing. Moreover, these assays can be used as a search method to locate bodily fluid stains at the crime scene. Additional assays, such as confirmatory assays, should be conducted afterward if necessary.

Confirmatory assays are more specific for the bodily fluid in question. These assays are utilized to identify bodily fluids with higher certainty than presumptive assays. For example, bloodstains are commonly associated with criminal investigations. A reddish-brown stain that has been identified through visual examination is usually tested by using presumptive assays. If the result of the presumptive assay on the alleged bloodstain is positive, the stain is then further analyzed by forensic DNA analysis. This approach *indicates* the presence of blood. Confirmatory assays are performed when a sample has to be *identified* as blood (Figure 11.1). Additionally, the human or animal origin of blood evidence can be determined if necessary and the techniques to do so will be discussed in Chapter 13.

11.1.4 Primary and Secondary Binding Assays

Traditionally, the detection and measurement of the antigen–antibody binding reactions serve as the bases of forensic serology. These assays fall into two categories: primary and secondary binding assays. Recall that the primary binding assays involve the initial binding between a single epitope of an antigen and a single binding site of an antibody (see Section 10.3.1). Therefore, they are very sensitive. Although the secondary binding assays are less sensitive than the primary binding assays, they are easier to perform than primary assays. The secondary assays consist of precipitation-based and agglutination-based assays. Precipitation-based assays have been used for species identification. Agglutination-based assays are more sensitive

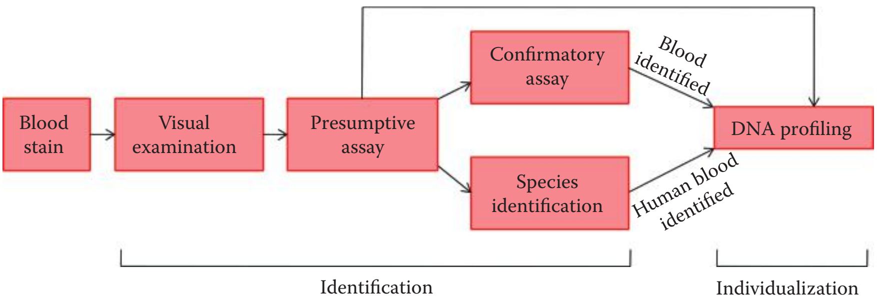

Figure 11.1 An example of work flow for processing blood samples. (© Richard C. Li.)

than precipitation-based assays. Agglutination-based assays, which detect antigens located on the surface of cells or carriers, are normally applied to blood group typing. Recently, emerging techniques such as RNA, proteomic, and DNA methylation assays, and fluorescence as well as Raman spectroscopy can potentially be used for the identification of bodily fluids.

11.2 Primary Binding Assays

11.2.1 Enzyme-Linked Immunosorbent Assay (ELISA)

An enzyme-linked immunosorbent assay (ELISA) is an immunoenzyme assay that can be used to detect and measure the antibody or antigen in question. The most common ELISA that is used in forensic serology is the antibody-sandwich ELISA (Figure 11.2). It is utilized to detect the prostate-specific antigen (PSA) to identify seminal stains and amylase for the identification of saliva (Chapters 13 and 14).

An antibody (usually monoclonal) coating is formed by nonspecific adsorption onto a solid phase such as the wells of a polystyrene plate. A sample containing the antigen to be tested is then added and binds to the solid-phase antibody. Subsequently, a second antibody is added to form an antibody–antigen–antibody sandwich complex. The second antibody binds to different epitopes of the antigen. Next, an enzyme-labeled antiglobulin (Chapter 10) is added to bind the sandwich. Subsequently, the excess enzyme-labeled antiglobulin is removed by washing and the bound antiglobulin can be detected.

A number of enzymes such as alkaline phosphatase and horseradish peroxidase have been used as reporting enzymes (Chapter 9) to label the antiglobulin for ELISA. The enzyme catalyzes the substrate and produces colorimetric or fluorometric signals. The intensity of the signals can

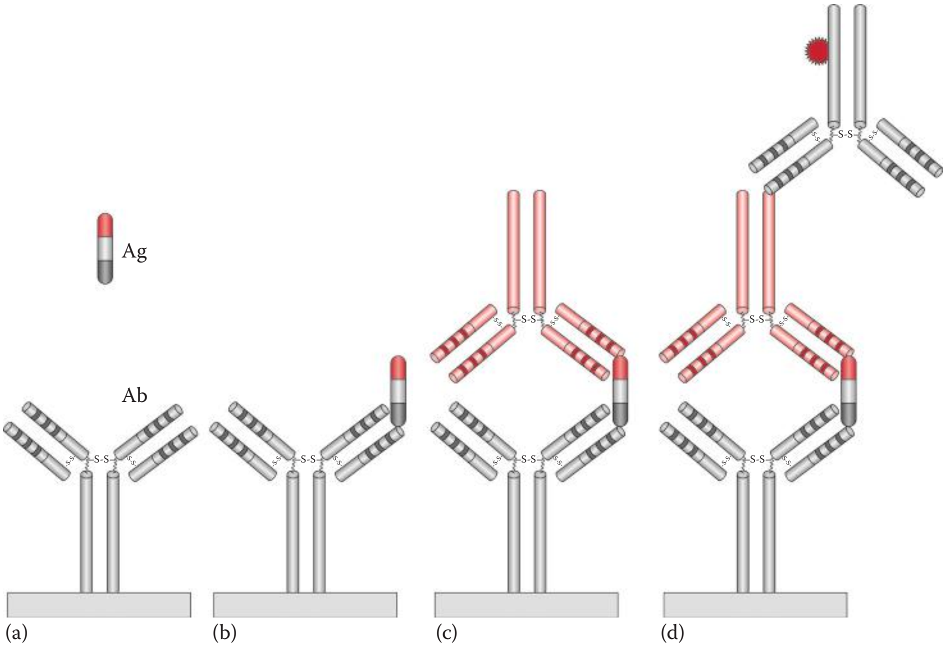

Figure 11.2 ELISA assay. (a) Sample containing Ag (antigen) is applied to a sample well where Ab (antibody) is immobilized. (b) Ag binds to an immobilized Ab to form an Ag–Ab complex. (c) A second Ab to a different epitope is added to form an Ab–Ag–Ab sandwich. (d) Labeled antiglobulin binds to the sandwich. The bound antiglobulin can be detected by various reporting schemes. (© Richard C. Li.)

be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of antigen can be quantified by comparing the standard with known concentrations. Recall that antiglobulin can recognize an epitope that is common to all isotypes within an antibody class (Section 10.1.2.3); the enzyme-labeled antiglobulin is a universal reagent regardless of the antibody used, as long as they are of the same class.

Alternatively, antibodies in a sample can also be detected and may be quantified by an ELISA system in which the antigen is bound to a solid phase instead of the antibody. After an antibody in a sample binds to the solid-phase antigen, an enzyme-labeled antiglobulin is added to bind to the bound antibody. The bound antiglobulin can be detected and measured by the addition of an enzyme substrate. The enzymatic catalytic reaction is similar to the antibody sandwich ELISA procedure previously described.

11.2.2 Immunochromatographic Assays

Figure 11.3 depicts a test using an immunochromatographic membrane device. A dye-labeled monoclonal antibody is contained in a sample well and a polyclonal antibody for the antigen (or a second monoclonal antibody to a different epitope of the antigen) is immobilized onto a test zone of a nitrocellulose membrane. An antiglobulin that recognizes the antibody is immobilized onto a control zone.

The assay is carried out by loading a sample into the sample well. The antigen in the sample binds to the dye-labeled antibody already in the sample well to form an antigen–antibody complex. The complex then diffuses across the nitrocellulose membrane until it reaches the test zone. The antibody immobilized at the test zone traps the antigen–antibody complex to form an antibody–antigen–antibody sandwich. The presence of the antigen in the sample results in a colored vertical line at the test zone. (Figure 11.4)

The immunochromatographic device also utilizes a control zone to ensure that the device works properly and that the sample has diffused completely along the test strip. Unbound

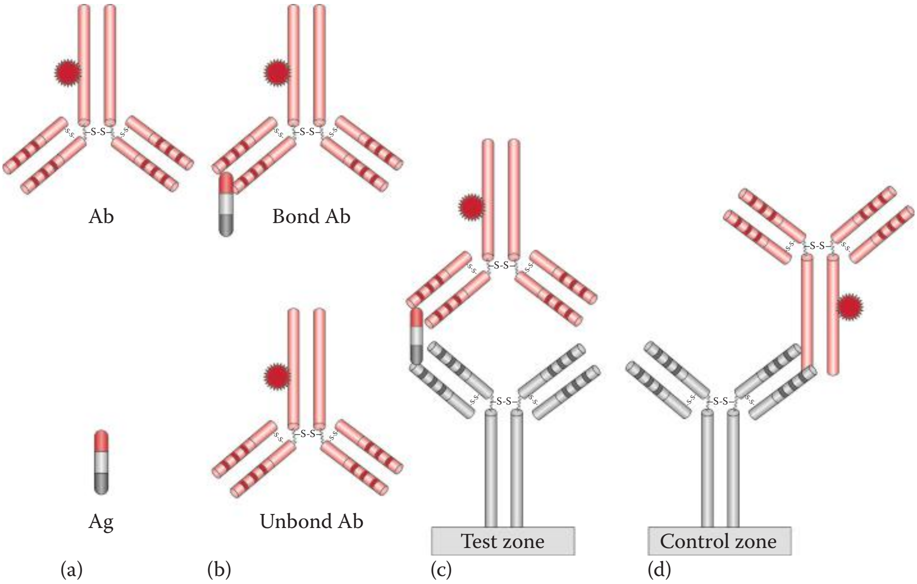

Figure 11.3 Immunochromatographic assay. (a) Sample containing Ag (antigen) is loaded in a sample well. (b) Ag binds to a labeled Ab (antibody) to form a labeled Ab–Ag complex. (c) At the test zone, the labeled Ab–Ag complex binds to an immobilized Ab to form a labeled Ab–Ag–Ab sandwich. (d) At the control zone, a labeled Ab binds to an immobilized antiglobulin and is captured at the control zone. (© Richard C. Li.)

monoclonal antibodies diffuse across the membrane until they reach the control zone where they are trapped by the immobilized antiglobulin. This antibody–antiglobulin complex at the control zone also results in a colored vertical line (Figure 11.4). The test is considered valid only if the line in the control zone is observed. The presence of an antigen results in a line at both the test zone and the control zone, while the absence of an antigen results in a line in the control zone only (Figure 11.4). This method is rapid and simple and thus can be used as a screening test in laboratories and as a field test at crime scenes to identify semen, saliva, and species (Table 11.1).

However, a false-negative result may be obtained if a sample contains a very high concentration of an antigen (Figure 11.5). Under this condition, the dye-labeled antibody will become saturated with the antigen to form antigen–antibody complexes. The unbound antigen will diffuse along with the

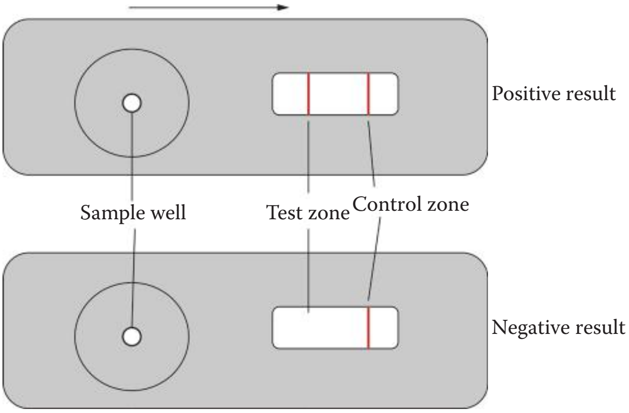

Figure 11.4 Immunochromatographic device. Positive and negative results are shown. (© Richard C. Li.)

Table 11.1 Common Immunochromatographic Assays for Forensic Applications

Assay	Antigen	Labeled Antibody	Immobilized Antibody	Forensic Application
ABAcard® HemaTrace® (Abacus Diagnostics)	Hemoglobin (Hb)	Monoclonal antihuman Hb antibody	Polyclonal antihuman Hb antibody	Blood and species identification
RSID™-Blood (Independent Forensics)	Glycophorin A (GPA)	Monoclonal antihuman GPA antibody	Monoclonal antihuman GPA antibody	Blood and species identification
RSID™-Saliva (Independent Forensics)	Human salivary α-amylase (HAS)	Monoclonal antihuman HAS antibody	Monoclonal antihuman HAS antibody[a]	Saliva identification
One-Step ABAcard PSA® (Abacus Diagnostics)	Prostate-specific antigen (PSA)	Monoclonal antihuman PSA antibody	Polyclonal antihuman PSA antibody	Semen identification
RSID™-Semen (Independent Forensics)	Semenogelin (Sg)	Monoclonal antihuman Sg antibody	Monoclonal antihuman Sg antibody[a]	Semen identification

[a] The epitope recognized by the immobilized antibody is different from that of the labeled antibody.

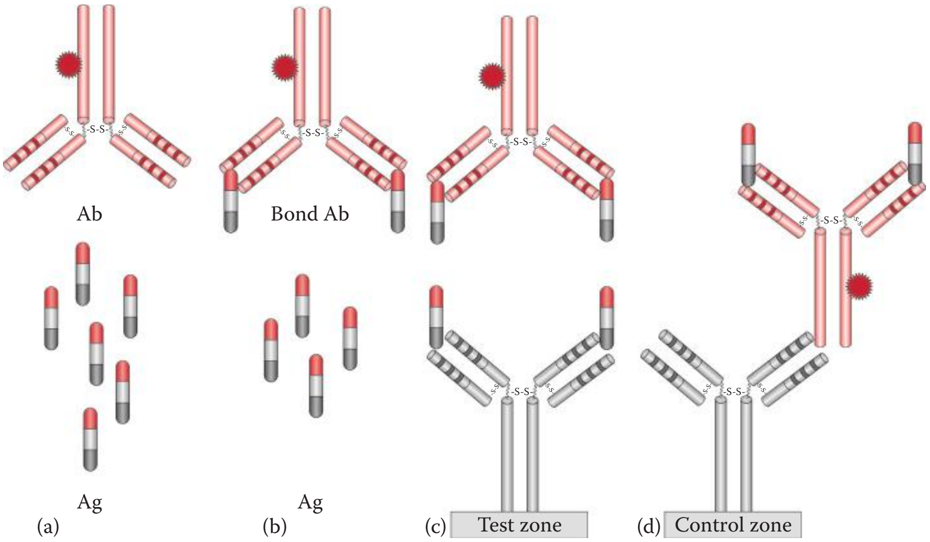

Figure 11.5 High-dose hook effect of immunochromatographic assay. (a) Ag (antigen) in sample is loaded in the sample well. (b) Ag binds to a labeled Ab (antibody) to form a labeled Ag–Ab complex. (c) At the test zone, free Ag binds to an immobilized Ab to form an unlabeled Ab–Ag complex and prevents the formation of a labeled Ab–Ag–Ab sandwich. (d) At the control zone, a labeled Ab binds to immobilized antiglobulin and is captured at the control zone. (© Richard C. Li.)

antigen–antibody complex toward the test zone. At the test zone, the unbound antigen will compete with the antigen–antibody complexes for the antibody immobilized at the test zone. Since the antigen is in excess, the unbound antigen binds to the immobilized antibody, preventing the formation of the antibody–antigen–antibody sandwich. The resulting reading appears as a negative result. This artifact is known as the *high-dose hook* effect. To prevent the high-dose hook effect, a smaller volume of sample can be applied, or the sample can be diluted to reduce the amount of antigen applied.

11.3 Secondary Binding Assays

11.3.1 Precipitation-Based Assays

These techniques are based on the precipitation reaction and are used primarily for species identification in forensic laboratories.

11.3.1.1 Immunodiffusion

Immunodiffusion is a passive method in which an antigen or an antibody or both are allowed to diffuse and therefore a gradient, from low to high concentration, is established for an antigen or an antibody or both. As a result, precipitation occurs due to the interaction between an antigen and an antibody. The assay can be carried out in a liquid or a semisolid medium such as agarose gel. The semisolid medium can stabilize the diffusion process and reduce interference such as convection, which is the movement of molecules within liquids. The two types are single immunodiffusion and double immunodiffusion.

11.3.1.1.1 Single Immunodiffusion

A concentration gradient is established for either an antigen or an antibody. One serology technique based on this principle is radial immunodiffusion—a single diffusion method in which a concentration gradient is established for an antigen. The antibody is uniformly distributed in the gel matrix (Figure 11.6). The antigen is loaded into a sample well and allowed to diffuse from the well into the gel until a precipitation reaction occurs. The precipitate ring around the well is observed; the area within the ring of precipitate is proportional to the amount of antigen loaded in the well. Standards using known concentrations of antigen can be included in the same assay along with the samples, and a standard curve can be plotted. The amounts of antigen in samples can then be quantified by comparing the results with the standard curve.

11.3.1.1.2 Double Immunodiffusion

The second type of assay is double diffusion in which a concentration gradient is established for both an antigen and an antibody. The most common examples are the ring assay and the Ouchterlony assay. The *ring assay* can be performed in a test tube or capillary tube (Figure 11.7). An antiserum, a denser phase, is placed in a small tube. An antigen solution is carefully layered on top of an antibody solution without mixing. Both the antigen and the antibody will diffuse toward each other. In a positive reaction, a ring of precipitate can be observed at the interface of the two solutions. A negative reaction is indicated by a lack of precipitation. The assay requires positive and negative controls along with questioned samples.

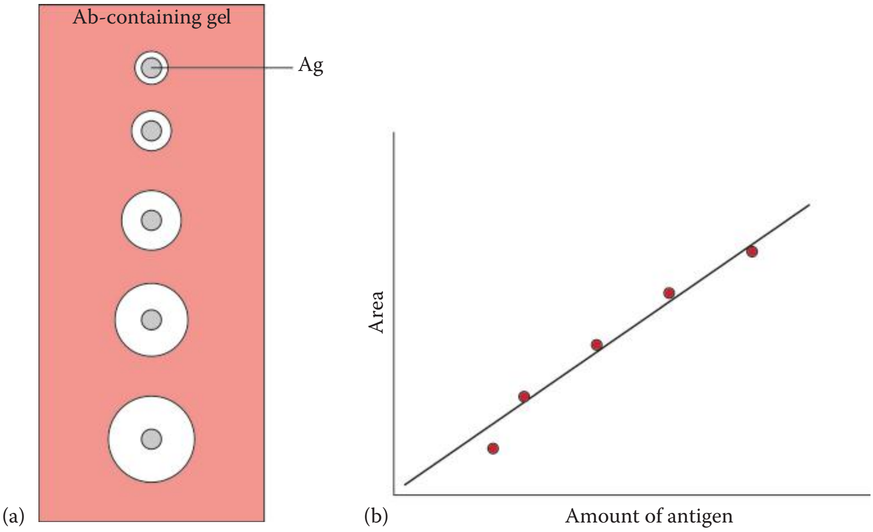

Figure 11.6 Radial immunodiffusion assay. (a) Immunodiffusion of antigens from a well into an antibody-containing gel. (b) Standard curve based on results of standards with known amounts of antigens. (© Richard C. Li.)

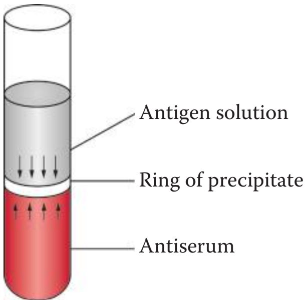

Figure 11.7 Ring assay. (© Richard C. Li.)

The *Ouchterlony assay* is named after the Swedish immunologist, Örjan Ouchterlony, who developed it. The assay can be performed in an agarose gel supported by a glass slide or polyester film (Figure 11.8). Wells are created by punching holes in the gel layer at desired locations. Often, a pattern with six wells surrounding a center well is used. The antibody is loaded in the central well while the questioned samples and the controls are loaded in the surrounding wells. The double diffusion of the antigen and the antibody from the wells is allowed to occur during incubation. If the reaction is positive, a precipitate line between wells can be observed at the end of incubation. The precipitate can be stained, enhancing visibility to aid observation. A single assay can compare more than one antigen to determine whether the antigens in question react the same way or differently with the antibody (see Chapter 7). This method is sometimes used to determine whether samples have come from the same or different origins.

11.3.1.2 Immunoelectrophoretic Methods

Diffusion techniques can be combined with electrophoresis to enhance test results. Electrophoresis separates molecules according to the differences in their electrophoretic mobility.

11.3.1.2.1 Immunoelectrophoresis

Immunoelectrophoresis (IEP) is a two-step procedure that can analyze a wide range of antigens. This technique uses electrophoresis to separate the antigen mixture prior to immunodiffusion (Figure 11.9). In the first step, the antigens in a sample are separated using agarose gel electrophoresis.

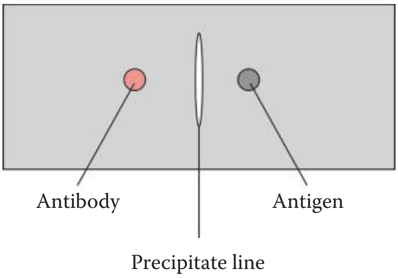

Figure 11.8 Ouchterlony assay. (© Richard C. Li.)

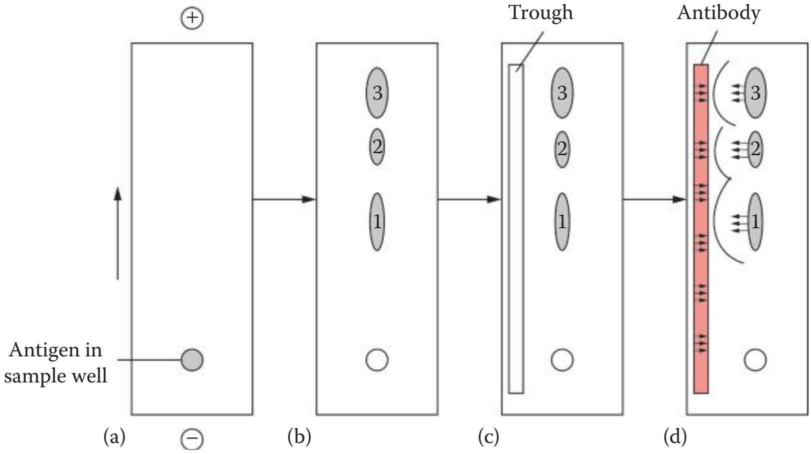

Figure 11.9 IEP assay. (a) Electrophoresis of antigens is carried out. (b) Various antigens are separated after electrophoresis. (c) Trough is cut. (d) Antibody is applied to the trough allowing diffusion to occur and forming precipitate lines. +, anode; −, cathode. (© Richard C. Li.)

Then, a trough is cut in the gel parallel to the array of antigens separated by electrophoresis. In the second step, an antibody is loaded in the trough and the gel is incubated for double diffusion. An arc-shaped precipitate line can be observed for a positive reaction. Multiple precipitate lines can occur if more than one antigen reacts with the antibody. The shapes, intensities, and locations of the precipitate lines of a known control and a questioned sample can be compared.

11.3.1.2.2 Crossed Immunoelectrophoresis

Crossed immunoelectrophoresis (CRIE), also known as two-dimensional IEP, is a modification of IEP. The first step utilizes electrophoresis to separate antigens contained in a sample (Figure 11.10). A strip of gel containing separated antigens is cut for the second round of electrophoresis. The gel including the gel strip is turned at a 90° angle and is further separated by a second-dimension electrophoresis. This drives the antigens from the gel strip into an agarose gel that contains uniformly distributed antibodies. Following the second-dimension electrophoresis, an arc-shaped precipitate line is formed. The area of the arc can be measured, and a sample can be identified by comparison with a known standard. This technique is more sensitive than IEP.

11.3.1.2.3 Rocket Immunoelectrophoresis

An antibody-containing agarose gel is used. The antigen is loaded into the well (Figure 11.11). Electrophoresis then drives the antigen from the well into the agarose gel. In a positive reaction,

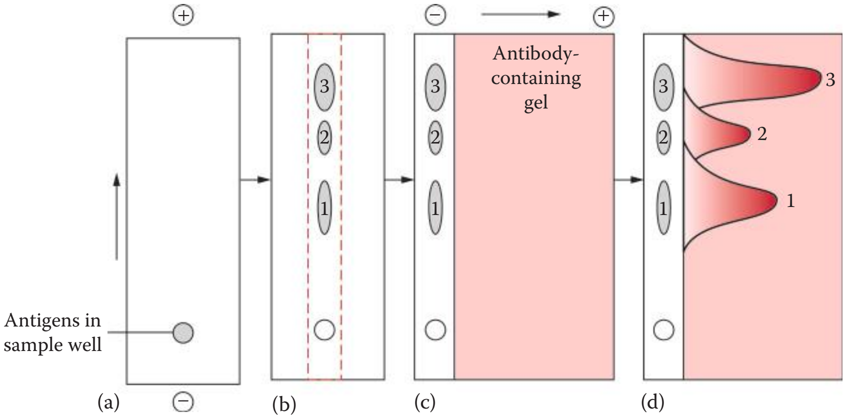

Figure 11.10 CRIE assay. (a) Electrophoresis of antigens is carried out. (b) Various antigens are separated after electrophoresis. A strip of gel containing separated antigens is excised and then used for second-dimension electrophoresis. (c) As shown, second-dimension electrophoresis is carried out to drive the antigens into an antibody-containing gel; (d) arc-shaped precipitate lines are formed as a result. (© Richard C. Li.)

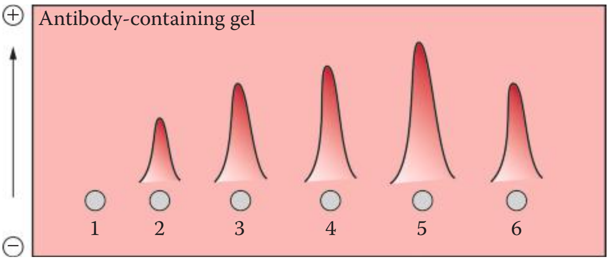

Figure 11.11 Rocket immunoelectrophoresis assay. Antigens are driven from the wells into an antibody-containing gel forming precipitate lines. (© Richard C. Li.)

a rocket-shaped precipitate line can be observed. The height of the rocket is in proportion to the amount of antigen in the sample. Quantitation can be achieved by comparing standards and the sample in the same gel.

11.3.1.2.4 Crossed-Over Immunoelectrophoresis

This technique is also known as counterimmunoelectrophoresis (CIE). Two arrays of opposing wells are created by punching holes in the agarose gel (Figure 11.12). The antibody and samples are loaded in opposing wells arranged by pairs. Electrophoresis is used to drive the antigen and the antibody toward each other. The wells containing the antibody should be proximal to the anode and the wells containing the samples should be proximal to the cathode. During gel electrophoresis, the antigen, which is usually negatively charged, migrates toward the anode. The antibody migrates in the opposite direction as a result of electroendosmosis—a phenomenon in which the movement of molecules is caused by fluid flow. A precipitate line is formed between the opposing wells if the antigen reacts with its specific antibody.

11.3.2 Agglutination-Based Assays

Agglutination reactions can be used as forensic serological assays such as for blood group typing (Chapter 18) and menstrual blood identification (Chapter 16). Agglutination assays are qualitative, indicating the absence or presence of antigens or antibodies. Semiquantitative results can be obtained by titration (diluting the antigen or antibody). Many types of agglutination reactions are available; only those used in forensic serology are discussed next.

Direct agglutination assays involve reactions in which an antibody interacts with antigens originally located on cell surfaces. In a *hemagglutination* reaction (Figure 11.13a), an antibody binds to the antigens located on erythrocytes. This method is used for the identification of blood types, for example, the testing of erythrocytes for ABO blood group typing. The assay can be carried out on a glass slide and the agglutination, as indicated by forming cell clumps, can be observed under a microscope (Chapter 18). The assay can also be carried out in a test tube. In a positive reaction, agglutinated cells are formed on the bottom of the test tube. Test tubes can also be centrifuged and then swirled to determine whether the cell clumps can be resuspended. Agglutinated cells cannot be resuspended. In a negative reaction, unagglutinated cells can be resuspended.

In *agglutination inhibition assays*, the presence of an antigen in question is indirectly detected (Figure 11.13b). If a known antigen is added to a mixture consisting of antigen-containing cells and antibodies, the added antigen will compete with the antigen located on the cell surfaces for antibody binding and inhibit the agglutination reaction. Another indirect agglutination assay is the absorption–elution assay (Chapter 18), which can be used for ABO blood group typing.

Passive agglutination assays are different from direct agglutination assays, in which the antigen is coated on the surface of carrier cells such as tannic acid–treated sheep erythrocytes. Tannic acid is considered to be a fixative. Treating with tannic acid can stabilize the carrier cells

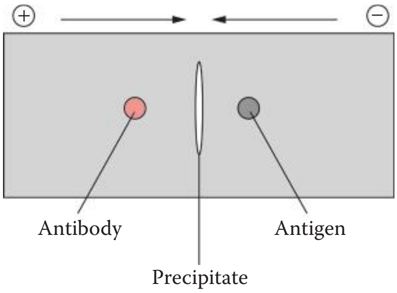

Figure 11.12 Crossed-over immunoelectrophoresis assay. (© Richard C. Li.)

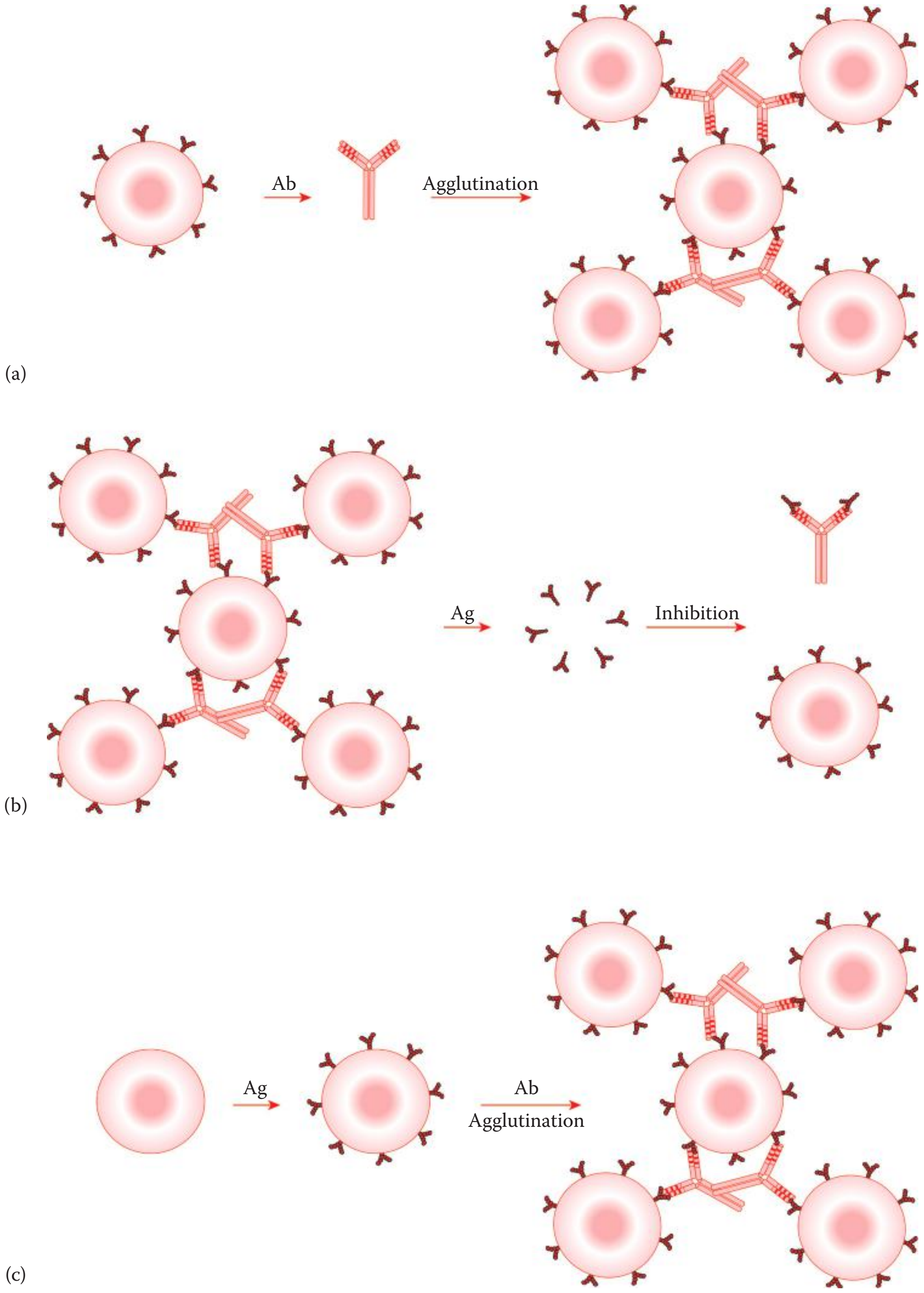
Ab
Agglutination
(a)
Ag
Inhibition
(b)
Ag
Ab
Agglutination
(c)

Figure 11.13 Agglutination-based assays. (a) Direct agglutination assay and (b) agglutination inhibition assay. Hemagglutination reactions are shown; antigens are located on erythrocytes. (c) Passive agglutination assay. (© Richard C. Li.)

for a subsequent coating of antigens. The carrier cells are then incubated with samples containing antigens. A coating of antigens can be formed on the surfaces of carriers. Such antigen-coated carrier cells can then be agglutinated using an antibody that is specific to the absorbed antigens (Figure 11.13c).

11.4 DNA Methylation Assays for Bodily Fluid Identification

In the eukaryotic genome, methylation occurs at the cytosine residues commonly in the CpG dinucleotide sequences of both DNA strands. The "p" in CpG refers to the phosphodiester bond between cytosine and guanine. Cytosine methylation is carried out *in vivo* by methyltransferase (DNMT). DNMT catalyzes the transfer of a methyl group (CH_3) from S-adenosylmethionine (SAM) to the C^5 position of cytosine (Figure 11.14). As a result, cytosine is converted to 5-methylcytosine. Cytosine methylation is observed throughout the eukaryotic genomes in both the coding and intergenic regions located between genes. However, cytosine methylation is rare at CpG islands, which are stretches of DNA containing a high frequency of CpG dinucleotides in the regulatory regions of genes (Figure 11.15). It is implicated that cytosine methylation alters the chromatin structure and causes chromatin condensation. As a result, gene expression is usually inhibited by cytosine methylation.

Genomic loci, known as *tissue-specific differentially methylated regions* (TDMRs), have been identified. TDMRs show differential DNA methylation patterns between different tissues. For example, semen-specific TDMRs are consistently hypomethylated in spermatozoa, which can be used to identify semen samples. Other TDMRs display varying degrees of methylation in different types of bodily fluids. This suggests that differential DNA methylation assays may be potentially useful for bodily fluid identification.

The detection of DNA methylation can be carried out by a number of methods. *Methylation-sensitive restriction enzyme digestion polymerase chain reaction* (MSRE-PCR) can be used for the

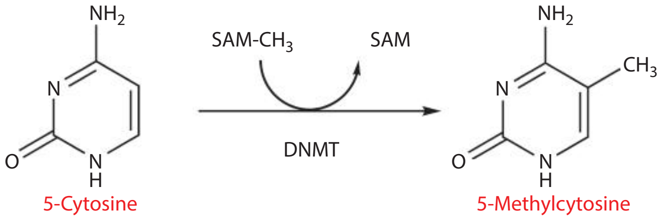

Figure 11.14 Eukaryotic cytosine methylation catalyzed by methyltransferase.

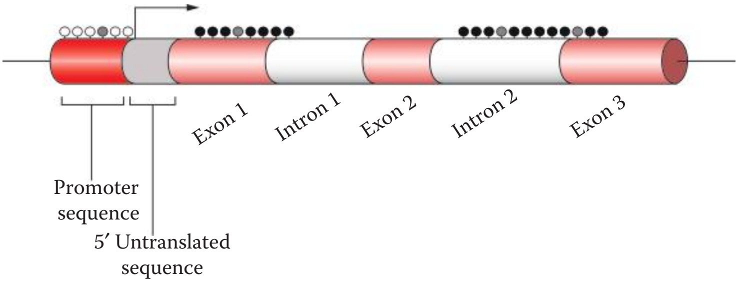

Figure 11.15 Methylation at the cytosine residues in CpG dinucleotide sequences. Note that the CpG dinucleotide sequences are usually unmethylated in the promoter regions. Black circle: methylated. Clear circle: unmethylated. Gray circle: partial methylated. (© Richard C. Li.)

rapid detection of DNA methylation. This method utilizes the *methylation-sensitive restriction enzyme* (MSRE). MSREs cleave DNA at unmethylated restriction sites, but they are not able to cleave once a cytosine residue is methylated. Thus, the methylated DNA remains intact. After PCR amplification, the methylated DNA is amplified and the amplicon is detected. In contrast, the unmethylated DNA is cleaved by the MSRE and cannot be amplified (Figure 11.16). *Bisulfite sequencing* is another widely used method. The genomic DNA is treated with sodium bisulfite, which catalyzes the hydrolytic deamination of unmethylated cytosines (Figure 11.17). As a result, unmethylated cytosines are converted to uracils, while methylated cytosines are resistant to conversion and remain unchanged. Next, uracils are replaced by thymines during PCR. After sequencing, the methylation sites can be deduced at a single-nucleotide resolution by comparing them with the reference sequence (Figure 11.18). *Methyl-DNA immunoprecipitation* (MeDIP) is a useful method for isolating methylated DNA. In this method, DNA is first sheared into fragments. The fragmented DNA is then treated with an antimethylcytosine antibody that binds to methylated cytosines. Methylated DNA fragments can then be isolated by immunoprecipitation for further analysis (Figure 11.19).

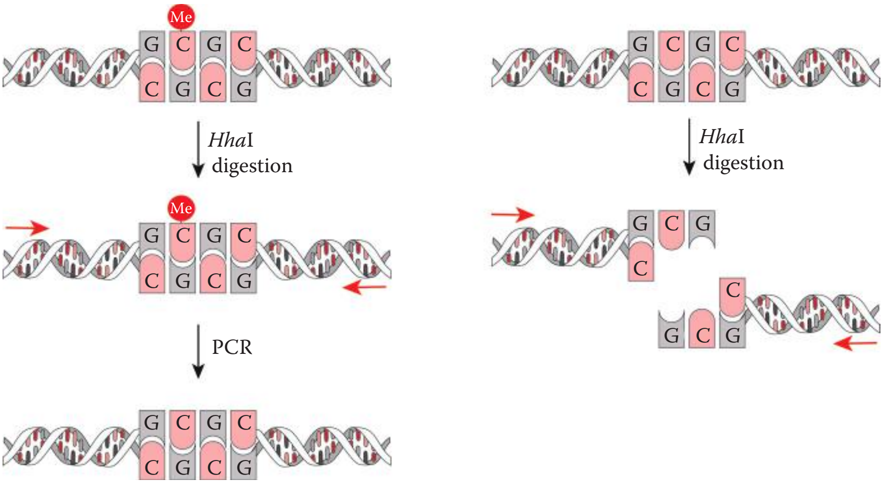

Figure 11.16 Methylation-sensitive restriction enzyme digestion PCR (MSRE-PCR) for the detection of DNA methylation. This method utilizes the methylation-sensitive restriction enzyme (MSRE) such as *Hha*I. *Hha*I cleaves DNA at the unmethylated sequence GCGC but is not able to cleave once a cytosine residue is methylated (GCmGC). The *Hha*I-treated DNA is then PCR amplified. The *Hha*I uncleaved DNA results in an amplicon, indicating the presence of a methylated cytosine in the sequence. The *Hha*I cleaved DNA cannot be amplified, indicating the absence of methylated cytosine in the sequence. Me, 5-methylcytosine site; red arrow, PCR primers. (Adapted from Frumkin, D. et al., *Forensic Sci Int Genet*, 5, 517–524, 2011.)

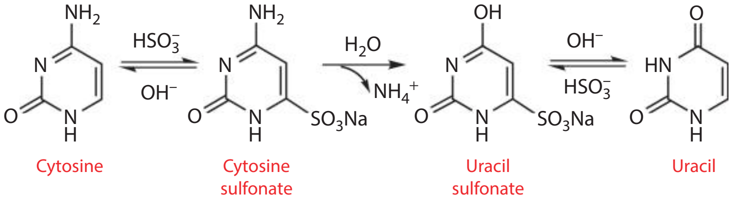

Figure 11.17 Bisulfite conversion of 5-methylcytosine.

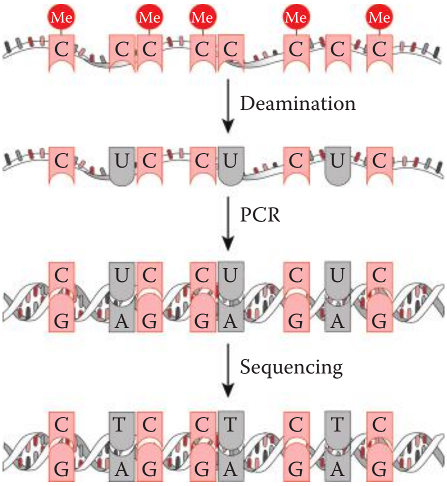

Figure 11.18 Schematic illustration of bisulfite sequencing of cytosine methylation. In a deamination reaction, unmethylated cytosines are converted to uracil. PCR and DNA sequencing are then carried out. The retention of C indicates that the site was methylated; the conversion of C to T indicates that the site was unmethylated in the DNA sample. Me, 5-methylcytosines site. (© Richard C. Li)

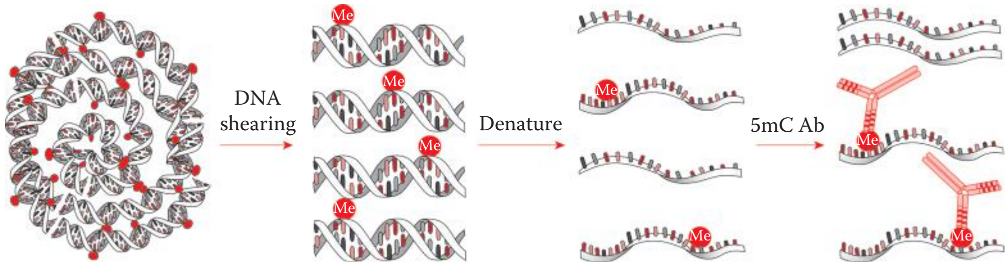

Figure 11.19 Methylated DNA immunoprecipitation (MeDIP) for the detection of methylated DNA. Genomic DNA is randomly and mechanically sheared into fragments. Fragments are then denatured to single-stranded DNA. The fragments containing the methylated region can be immunoprecipitated using an antibody that specifically recognizes 5-methylcytosine. The methylated DNA sequence in the immunoprecipitated fragments can then be determined by DNA sequencing and methylation analysis. Me, 5-methylcytosine site; 5mC Ab, 5-methylcytosine antibody. (© Richard C. Li.)

11.5 Forensic Applications of RNA-Based Assays and RNA Profiling

11.5.1 Messenger RNA-Based Assays

mRNA-based assays have been developed to identify bodily fluids for forensic investigation. In forensic pathological investigation, they can potentially be used for wound age estimation and the age of biological stains. The assays are based on the expression of certain genes in certain cell or tissue types. Candidate tissue-specific genes that may be useful for the identification of bodily fluids have been identified. Thus, the techniques that are used in the identification of bodily fluids are based on the detection of specific types of mRNA that are expressed exclusively in certain cells. The *tissue-specific genes* that are utilized for bodily fluid identification are summarized in Table 11.2. Additionally, *reference genes* that are constitutively expressed housekeeping genes are utilized as internal controls. The amount of mRNA can be assessed by normalizing the target gene to the expression level of the reference genes. Compared with conventional assays that are used for bodily fluid identification, the mRNA-based assay has higher specificity and is amenable to automation. However, one limitation is that the mRNA is unstable due to degradation by

Table 11.2 Representative Markers of mRNA-Based Assays for Bodily Fluid Identification

Bodily Fluid	Gene Symbol	Description
Blood	*ALAS2*	Aminolevulinatesynthase 2
	AMICA1	Adhesion molecule, interacts with CXADR antigen 1
	ANK1	Ankyrin1
	CD3G	CD3 gammamolecule
	CD93	CD93 molecule
	HBA1	Alpha 1 hemoglobin
	HBB	Beta hemoglobin
	PBGD	Porphobilinogen deaminase
	SPTB	Beta spectrin
Saliva	*HTN3*	Histatin 3
	MUC7	Mucin 7
	STATH	Statherin
Semen	*KLK3*	Kallikrein 3 (prostate-specific antigen)
	PRM1	Protamine 1
	PRM2	Protamine 2
	SEMG1	Semenogelin 1
	TGM4	Transglutaminase 4
Vaginal secretions	*CYP2B7P1*	Cytochrome P450, family 2, subfamily B, polypeptide 7 pseudogene 1
	DKK4	Dickkopf homolog 4
	FUT6	Fucosyltransferase 6
	HBD1	Beta defensin 1
	IL19	Interleukin 19
	MUC4	Mucin 4
	MYOZ1	Myozenin 1
	SFTA2	Surfactant associated 2
Menstrual blood	*MMP7*	Matrix metalloproteinase 7
	MMP11	Matrix metalloproteinase 11
Skin	*CDSN*	Corneodesmosin
	LOR	Loricrin

(*continued*)

Table 11.2 (Continued) Representative Markers of mRNA-Based Assays for Bodily Fluid Identification

Bodily Fluid	Gene Symbol	Description
Reference gene	*18S rRNA*	18S ribosomal RNA
	ACTB	Beta actin
	G6PDH	Hexose-6-phosphate dehydrogenase
	RPS15	Ribosomal protein S15

Sources: Adapted from Bauer, M. and Patzelt, D., *J Forensic Sci*, 47, 1278–1282, 2002; Hanson, E.K. and Ballantyne, J., *Sci Justice*, 53, 14–22, 2013; Lindenbergh, A., Maaskant, P., and Sijen, T., *Forensic Sci Int Genet*, 7, 159–166, 2013; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

endogenous ribonucleases. Additionally, bodily fluid stains collected from crime scenes are often exposed to ultraviolet (UV) light, moisture, and high temperature, which can promote mRNA degradation. Nevertheless, the successful detection of mRNA from aged samples is possible.

11.5.2 MicroRNA-Based Assays

The biological function of microRNAs (miRNAs) is to regulate gene expression. The mature miRNA strand associated with the RNA-induced silencing complex (RISC) binds to its target mRNA. In animal cells, the mature miRNA forms a base pairing with its complementary sequence of the *miRNA responsive element* (MRE) within the 3′ untranslated regions (3′UTRs) of the target mRNA (Figure 11.20). A single miRNA can bind different mRNA transcripts encoded by multiple genes. Animal miRNAs usually form a base pairing with their target mRNAs through partial complementary sequences, which lead to translation repression to inhibit protein synthesis. If the base pairing is exactly complementary to its target mRNA sequence, then the cleavage and the degradation of the target mRNA occurs. As a result, the synthesis of the protein encoded by the mRNA is reduced. Thus, miRNAs play important roles in cellular function.

miRNAs have potential applications in forensic identification. Recent studies have revealed that miRNAs can be detected from various bodily fluids such as blood, saliva, semen, vaginal secretions, and menstrual blood. The type of miRNA expressed is not exclusively tissue- or cell-type specific. However, the expression patterns of specific miRNAs are unique for various bodily fluids. Thus, a bodily fluid in question can potentially be identified using multiple differentially expressed miRNAs as markers (Table 11.3). miRNA markers for bodily fluid identification have their advantages compared with those of mRNA. Mature miRNAs are much

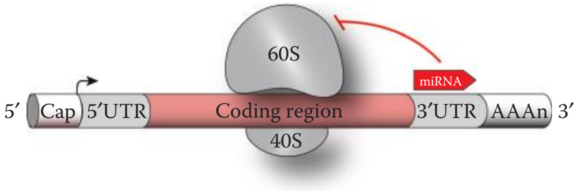

Figure 11.20 Schematic illustration of the roles of miRNA in translational inhibition. In animal cells, miRNAs bind to 3′UTR of their target mRNA with partially complementary Watson–Crick base pairing. As a result, the binding of miRNA to the target mRNA inhibits translation of the target mRNA. Eukaryotic 40S and 60S ribosomal subunits are shown. (© Richard C. Li.)

Table 11.3 Differentially Expressed miRNA Markers for Bodily Fluid Identification

Biological Fluid	miRNA Set
Blood	miR16; miR451
	miR16; miR486
	miR20a; miR106a; miR144; miR185
	miR126; miR150; miR451
Semen	miR10b; miR135b
	miR10a; miR135a; miR943; miR507; miR891a
Saliva	miR205; miR658
	miR200c; miR203; miR205
Vaginal secretions	miR124a; miR372
Menstrual blood	miR214
	miR412; miR451
Reference gene	*RNU6-2* (U6 small nuclear 2 RNA)

Sources: Adapted from Courts, C. and Madea, B., *J Forensic Sci*, 56, 1464–1470, 2011; Hanson, E.K., Lubenow, H., and Ballantyne, J., *Anal Biochem*, 387, 303–314, 2009; Wang, Z. et al., *Forensic Sci Int Genet*, 7, 116–123, 2013; Zubakov, D. et al., *Int J Legal Med*, 124, 217–226, 2010.

smaller, approximately 21–23 nucleotides in length, than mRNA. As a result, miRNAs are less susceptible to degradation and more stable than mRNAs. Additionally, the sensitivity of the miRNA-based method is much higher than that of the mRNA-based methods.

11.6 Proteomic Approaches Using Mass Spectrometry for Bodily Fluid Identification

The forensic application of proteomic approaches is to identify biomarker proteins derived from bodily fluids to determine the type of bodily fluid in question for forensic investigations. Mass spectrometry (MS) is a highly sensitive and rapid technique for protein identification from complex biological samples.

11.6.1 Mass Spectrometric Instrumentation for Protein Analysis

A mass spectrometer is an analytical technique to identify a molecule by measuring the ratio of the mass (m) to the charge (z) of a charged molecule (Figure 11.21). A typical mass spectrometer instrument analyzes ionized molecules while in its gas phase. A mass spectrometer consists of an *ion source* that converts analytes into gaseous phase ions, through a process known as *ionization*, by gaining a positive or a negative charge from a neutral species. The ions are then introduced into the *mass analyzer* of a mass spectrometer, where ions are accelerated under electric fields and separated based on their m/z ratio. The separated ions are detected by an *ion detector* that records the number of ions at each m/z value. The result of molecular ionization,

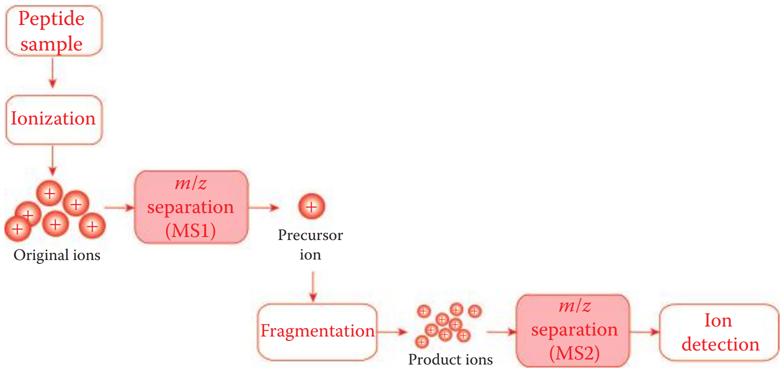

Figure 11.21 Diagram of tandem mass spectrometry. For the initial mass analysis of thermolabile analytes including peptides and proteins, soft ionization techniques such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) can be used to produce original ions. MALDI utilizes a laser beam to trigger ionization and facilitate the vaporization of the analyte. In the ESI method, a fine mist of droplets is formed in the presence of a high electric field, producing charged ions. In the first round of mass analysis, ions are separated based on their *m/z* ratio. A particular ion, called a precursor ion, is then selected for fragmentation, generating product ions. The fragmentation can be carried out using collision techniques such as collision-induced dissociation (CID) by multiple collisions with rare gas atoms. Additionally, a new fragmentation technique, electron-capture dissociation (ECD), can be used in which the fragmentation is induced through the capture of a thermal electron by a protonated peptide cation. The product ions are analyzed by the second round of mass analysis and detected by an ion detector. MS, mass analyzer. (© Richard C. Li.)

ion separation, and ion detection is a spectrum that can be used to determine the mass (or molecular weight) and potentially the structure of the molecule.

Furthermore, biomarker proteins can be identified by *peptide sequencing* using the tandem mass spectrometry (MS/MS) mode of operation in which a mass spectrometer uses two or more mass analyzers. After an initial mass analysis, individual peptide ions in the mass spectrum can be selected. The selected ions are known as *precursor ions*. A precursor ion is then subjected to a second round of fragmentation through collision and is broken into smaller ions known as *product ions*. The resulting product ions are further analyzed by MS. Thus, the corresponding protein can be identified through searching a protein database, based on the mass of a specific peptide obtained.

11.6.2 Analysis Strategies for Protein Identification

To date, various MS-based protein identification strategies have been developed, which can be divided into two categories: "top-down" and "bottom-up" strategies (Figure 11.22). In the *top-down* strategy, intact proteins in a complex mixture are fractionated and separated into less complex protein mixtures or single proteins. Intact proteins are then analyzed using MS. A liquid chromatography–mass spectrometry (LC–MS) system can be used for such an application. The advantage of this approach is that an entire protein is analyzed for identification purposes. However, the throughput and efficiency of this approach is still a major challenge for large-scale sample analysis in forensic applications.

The *bottom-up* strategy is commonly used for high-throughput analysis of small peptides derived from highly complex samples. Current MS usually cannot resolve the mass measurement of large proteins; rather, MS is ideal for the analysis of small peptides. One approach is the *sort-then-break* approach in which proteins in a complex mixture are first isolated and then enzymatically cleaved. The target proteins are isolated from their complex biological source using 1-D or 2-D gel electrophoresis (Chapter 18). The isolated proteins are then enzymatically

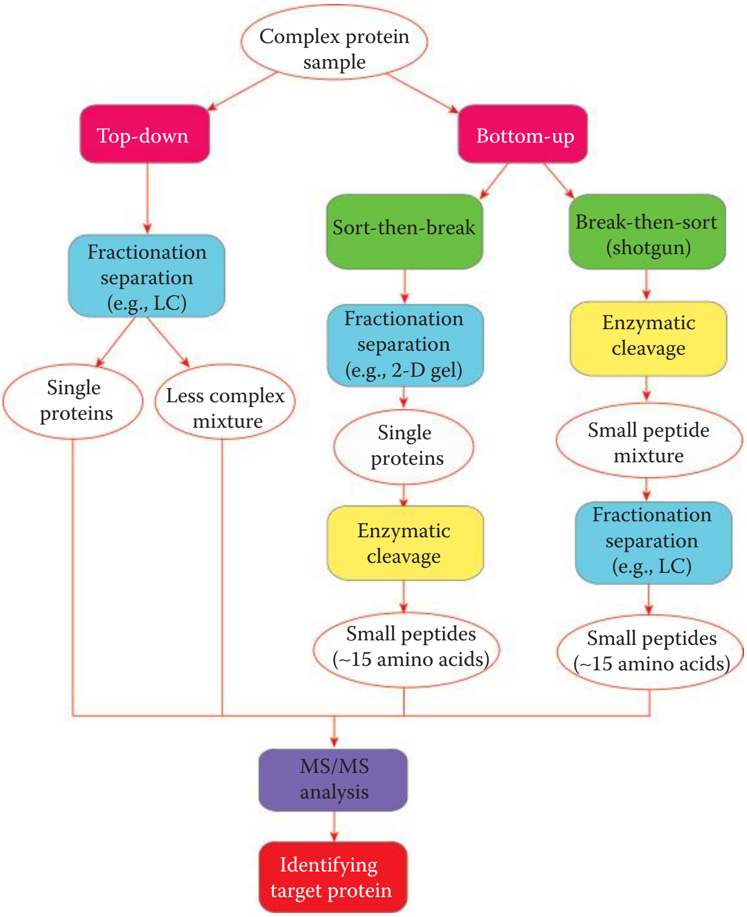
Complex protein sample
Top-down
Bottom-up
Fractionation separation (e.g., LC)
Single proteins
Less complex mixture
Sort-then-break
Break-then-sort (shotgun)
Fractionation separation (e.g., 2-D gel)
Single proteins
Enzymatic cleavage
Small peptides (~15 amino acids)
Enzymatic cleavage
Small peptide mixture
Fractionation separation (e.g., LC)
Small peptides (~15 amino acids)
MS/MS analysis
Identifying target protein

Figure 11.22 Top-down and bottom-up strategies for MS-based protein identification. LC, liquid chromatography; MS/MS, tandem mass spectrometry. (© Richard C. Li.)

cleaved into small peptides, approximately 15 amino acids in length, using proteases such as trypsin. These peptides are then analyzed by peptide sequencing using tandem MS to identify the target protein.

Break-then-sort, also known as the *shotgun* approach, is another approach of the bottom-up strategy. In this approach, proteins are cleaved first and the resulting peptide mixture is then separated by LC and analyzed by tandem MS (usually coupled to the LC). In a complex mixture, a peptide mass measurement alone is not sufficient for the identification of a target protein. Database searches are carried out utilizing software that compares the observed MS peptide spectra against all candidate spectra of the proteins that are present in a protein sequence (or translated from genomic DNA sequences) database for possible matches. Thus, a biomarker protein can be identified.

11.7 Microbial DNA Analysis for Bodily Fluid Identification

Human bodies harbor a large number of microbes. It is estimated that there are over 10 times more microbial cells than human cells in and on the human body, in the nose, mouth, throat, intestine, vagina, and skin. This microbial community is referred to as the human *microbiota*.

It is known that the microbiota has an important influence on human health. A decade ago, the concept of the *microbiome* was proposed by Nobel laureate Joshua Lederberg to identify and characterize the microbiota, their genomes, and their environmental interactions in a defined community. Traditionally, the studies of microbial genomes rely on cultivated individual species. The disadvantage of cultivation-based methods, however, is that large numbers of microbial species cannot be successfully isolated because many microbial species require specific growth conditions. Based on the advances in DNA sequencing technologies, a new approach, known as *metagenomics*, has been developed. Instead of studying the genome of an individual cultivated microbial species, metagenomics allows the study of the microbial genome of complete microbial communities harvested directly from natural environments. Metagenomics has been a powerful tool to analyze the human microbiome. Recently, human microbiota have been surveyed, revealing that microbial diversity is greatest in the teeth and the intestines, intermediate on the skin and the inside surface of the cheek, and lowest in the vagina (Table 11.4). More interestingly, each habitat is characterized by a small number of highly predominant bacterial taxa. The predominant taxa of each habitat can potentially be utilized for the forensic analysis of biological evidence. Bodily fluids identification is an important task in the forensic analysis of biological evidence. The detection of predominant taxa of bacteria can aid in the forensic identification of a particular type of bodily fluid (Table 11.4).

Table 11.4 Representative Microbiota that are Potentially Can Be Used for Forensic Identification

Body Habitats	Estimated Number of Microbial Species	Predominent Taxa	Potential Forensic Applications	DNA Marker	Further Reading
Cheek	800	*Streptococcus* spp.	Saliva identification	*gtf*	Nakanishi et al. (2011)
			Expirated blood	*gtf*	Donaldson et al. (2010)
				16S rRNA	Power et al. (2010)
Vagina	300	*Lactobacillus* spp.	Vaginal secretion identification	16S rRNA	Akutsu et al. (2012)
				16S–23S ISR	Fleming and Harbison (2010)
					Giampaoli et al. (2012)
Gastrointestinal tract	4000	*Bacteroides* spp.	Fecal matter identification	*rpoB*	Nakanishi et al. (2013)
Skin	1000	*Staphylococcus* spp.	Touched evidence identification	16S rRNA	Fierer et al. (2010)

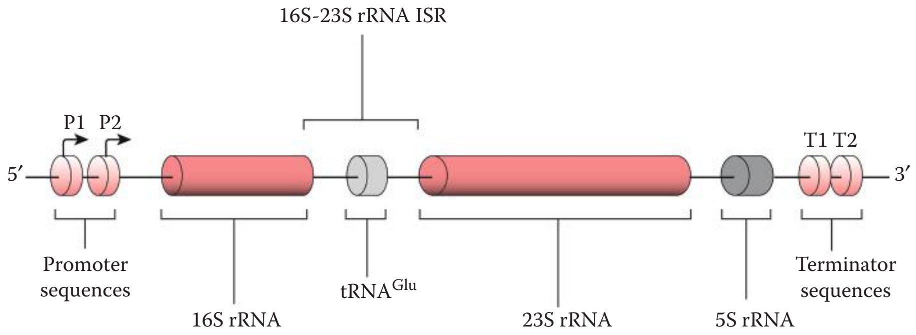

Figure 11.23 Schematic diagram of an *Escherichia coli* rRNA operon. *E. coli* has seven rRNA operons. A typical rRNA operon is shown. The rRNA operon encodes for three ribosomal RNA (rRNA) genes: the 5S and 23S genes code for the large ribosomal subunit, while the 16S gene encodes for the small ribosomal subunit. *E. coli* encodes one tRNA gene in the rRNA operon. The intergenic spacer regions (ISRs) and the promoter and terminator sequences are also shown. The rRNA operon is transcribed into a single precursor RNA transcript that is cleaved into separate rRNA and tRNA transcripts. The diagram is not to scale to the physical map of rRNA operons. (© Richard C. Li.)

The detection of bacteria can be achieved by utilizing DNA markers. The most commonly used DNA markers are the essential genes present in all bacterial species. These markers have highly variable regions that can be used to distinguish various bacterial species. The DNA sequences within these variable regions can also suggest taxonomic relationships between different taxa. Additionally, DNA markers contain conserved regions that are necessary for use as primer-binding sites for PCR-based detection.

For example, the rRNA genes (Figure 11.23), encoding for the ribosomal RNA, are conserved in all living organisms including bacteria. The bacterial rRNA operon contains three rRNA genes: the 5S and 23S rRNA genes encode the RNA components of the large subunit of the ribosome, while the 16S rRNA gene encodes the RNA component of the small subunit of the ribosome. The 16S rRNA gene is the most commonly used marker for the taxonomic identification of bacteria. The bacterial 16S rRNA genes are similar in length, approximately 1.5 kb. They contain highly conserved regions that can be used as primer-binding sites for the amplification of the adjacent DNA regions. Additionally, they contain variable regions that allow for taxonomic identification. In forensic applications, they are used as a marker for the identification of vaginal bacteria as an indicator of vaginal secretions (Chapter 16). However, the number of variable regions in the 16S rRNA genes of some bacteria species is limited. Furthermore, closely related species have high levels of sequence similarity to the 16S rRNA genes. Thus, the 16S rRNA gene marker is not adequate for distinguishing closely related species.

DNA markers of bacteria identification can also be located in noncoding regions of DNA. The *intergenic spacer region* (ISR) between the 16S rRNA and 23S rRNA genes in the rRNA operon is another commonly used marker. The 16S–23S rRNA ISR contains both length and sequence variations between species. These variations are partially caused by the number and type of tRNA genes that this region of operon contains. For example, most gram-negative bacteria contain both a copy of the tRNA^{ala} and the tRNA^{ile} genes, while some contain only a copy of the tRNA^{glu} gene. In contrast, most gram-positive bacteria have no tRNA gene; some contain either a copy of the tRNA^{ala} gene or the tRNA^{ile} gene or both. Thus, heterogeneity in the length and in the sequence of the 16S–23S rRNA ISR allows bacteria to be identified at the species level.

In addition to rRNA genes, many other genes are utilized as DNA markers for bacterial species identification. For example, the *rpoB* gene encodes the β subunit of bacterial RNA polymerase and is used as a marker for the forensic identification of fecal matter (Chapter 17). The sequence variation of the *rpoB* gene allows the distinguishing of species when they are not distinguishable, using 16S rRNA gene sequences. Several oral streptococcal species that produce extracellular polysaccharides contain the *gtf* gene, which encodes glucosyltransferase. Thus, the *gtf* gene is a useful marker for oral streptococcal species identification as a potential indicator of saliva in forensic investigations. However, this marker cannot be used to identify bacteria other than *streptococcus*. Nevertheless, combining several DNA markers can be used for the identification of bacteria species.

11.8 Nondestructive Assays for the Identification of Bodily Fluids

Most of the bodily fluid identification techniques mentioned previously consume a portion of the evidence. When the amount of sample evidence is very limited, such destructive assays may consume the evidence that is also needed for subsequent forensic DNA analysis. Moreover, most of these tests are used for a single type of bodily fluid. Tests for multiple bodily fluids in question will consume additional amounts of the evidence. Therefore, nondestructive identification techniques are highly desired. Additionally, it is important to develop an assay that can be utilized to test multiple types of bodily fluids. Furthermore, a portable device that can identify bodily fluids at a crime scene is very useful. It provides investigators with test results immediately to aid the crime scene investigations. Two techniques have been developed that can potentially be used for nondestructive bodily fluid identification: fluorescence spectroscopy and Raman spectroscopy.

Fluorescence spectroscopy is one example of a nondestructive technique. Fluorescence is the emission of light by a fluorophore. A fluorophore is a moiety in a molecule that fluoresces on absorbing the energy from an excitation light source or radiation. The emitted light usually has a longer wavelength and lower energy than the absorbed radiation. Constituents present in bodily fluids, such as nucleic acids, proteins, lipids, and metabolites, can exhibit fluorescence. The unique composition of a bodily fluid emits characteristic emission spectra, thus making it identifiable. Using multiple wavelengths of an excitation light source, fluorescent emissions at a wide range of wavelengths can be detected, which allows for the identification of various bodily fluids. This technique is sensitive and rapid. It does not utilize chemical reagents and the detection does not require physical contact with a sample. After identification, the same biological sample can be used for forensic DNA analysis. Additionally, portable fluorescence instruments can potentially be used at a crime scene for bodily fluid identification. However, it is not clear if exposure to an excitation light source can damage DNA evidence.

Raman spectroscopy is another example of a nondestructive technique. Raman spectroscopy utilizes a near-infrared excitation light source and measures the scattering of laser light caused by the vibrating molecules of a sample. A typical Raman spectrum provides information about the molecular structure and the "signature" based on the properties of the constituents of a sample. Thus, it is possible to obtain unique Raman spectra for a particular type of bodily fluid. Raman spectroscopy is highly sensitive. The measurements can be carried out on very small amounts of sample, ranging from a few picoliters to femtoliters. This technique does not require any chemical reagent. Additionally, it does not consume the sample, which can be used for subsequent DNA analysis. Raman spectroscopy has been utilized for various forensic purposes including the identification of drugs, trace evidence such as fibers, and questioned document evidence such as inks and paints. It can potentially be used for the nondestructive identification of various bodily fluids. A portable Raman spectrometer will allow bodily fluids to be identified at crime scenes.

Bibliography

Aaspõllu, A., et al., Can microbes on skin help linking persons and crimes? *Forensic Sci Int Genet Suppl Series*, 2011, **3**(1): e269–e270.

Akutsu, T., et al., Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. *Leg Med*, 2012, **14**(3): 160–162.

Allen, S.M., An enzyme linked immunosorbent assay (ELISA) for detection of seminal fluid using a monoclonal antibody to prostatic acid phosphatase. *J Immunoassay*, 1995, **16**(3): 297–308.

Alvarez, M., J. Juusola, and J. Ballantyne, An mRNA and DNA co-isolation method for forensic casework samples. *Anal Biochem*, 2004, **335**(2): 289–298.

Amin, K.M., et al., The exon-intron organization of the human erythroid beta-spectrin gene. *Genomics*, 1993, **18**(1): 118–125.

An, J.H., et al., Body fluid identification in forensics. *BMB Rep*, 2012, **45**(10): 545–553.

An, J.H., et al., DNA methylation-specific multiplex assays for body fluid identification. *Int J Legal Med*, 2013, **127**(1): 35–43.

Anderson, J.R., The agglutination of sensitised red cells by antibody to serum, with special reference to nonspecific reactions. *Br J Exp Pathol*, 1952, **33**(5): 468–483.

Bauer, M., RNA in forensic science. *Forensic Sci Int Genet*, 2007, **1**(1): 69–74.

Bauer, M. and D. Patzelt, Evaluation of mRNA markers for the identification of menstrual blood. *J Forensic Sci*, 2002, **47**(6): 1278–1282.

Bauer, M. and D. Patzelt, Protamine mRNA as molecular marker for spermatozoa in semen stains. *Int J Legal Med*, 2003, **117**(3): 175–179.

Bauer, M., A. Kraus, and D. Patzelt, Detection of epithelial cells in dried blood stains by reverse transcriptase-polymerase chain reaction. *J Forensic Sci*, 1999, **44**(6): 1232–1236.

Benschop, C.C.G., et al., Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? *Int J Legal Med*, 2012, **126**(2): 303–310.

Berti, A., et al., Expression of seminal vesicle-specific antigen in serum of lung tumor patients. *J Forensic Sci*, 2005, **50**(5): 1114–1115.

Black, C.M., et al., Multicenter study of nucleic acid amplification tests for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in children being evaluated for sexual abuse. *Pediatr Infect Dis J*, 2009, **28**(7): 608–613.

Bocklandt, S., et al., Epigenetic predictor of age. *PLoS One*, 2011, **6**(6): e14821.

Bowden, A., R. Fleming, and S. Harbison, A method for DNA and RNA co-extraction for use on forensic samples using the Promega DNA IQ™ system. *Forensic Sci Int Genet*, 2011, **5**(1): 64–68.

Brenig, B., J. Beck, and E. Schutz, Shotgun metagenomics of biological stains using ultra-deep DNA sequencing. *Forensic Sci Int Genet*, 2010, **4**(4): 228–231.

Brzezinski, J.L. and D.L. Craft, Characterization of microorganisms isolated from counterfeit toothpaste. *J Forensic Sci*, 2012, **57**(5): 1365–1367.

Chu, Z.L., et al., Erythroid-specific processing of human beta spectrin I pre-mRNA. *Blood*, 1994, **84**(6): 1992–1999.

Counsil, T.I. and J.L. McKillip, Forensic blood evidence analysis using RNA targets and novel molecular tools. *Biologia*, 2010, **65**(2): 175–182.

Courts, C. and B. Madea, Micro-RNA—A potential for forensic science? *Forensic Sci Int*, 2010, **203**(1–3): 106–111.

Courts, C. and B. Madea, Specific micro-RNA signatures for the detection of saliva and blood in forensic body-fluid identification. *J Forensic Sci*, 2011, **56**(6): 1464–1470.

Courts, C. and B. Madea, [Ribonucleic acid—Importance in forensic molecular biology]. *Rechtsmedizin*, 2012, **22**(2): 135–144.

Courts, C., M. Grabmuller, and B. Madea, Dysregulation of heart and brain specific micro-RNA in sudden infant death syndrome. *Forensic Sci Int*, 2013, **228**(1–3): 70–74.

Culliford, B.J., Precipitin reactions in forensic problems. A new method for precipitin reactions on forensic blood, semen and saliva stains. *Nature*, 1964, **201**: 1092–1093.

Cummings, C.A. and D.A. Relman, Genomics and microbiology. Microbial forensics— "cross-examining pathogens". *Science*, 2002, **296**(5575): 1976–1979.

Donaldson, A.E., et al., Using oral microbial DNA analysis to identify expirated bloodspatter. *Int J Legal Med*, 2010, **124**(6): 569–576.

Elkins, K.M., Rapid presumptive "fingerprinting" of body fluids and materials by ATR FT-IR spectroscopy. *J Forensic Sci*, 2011, **56**(6): 1580–1587.

Ferguson, L.S., et al., Direct detection of peptides and small proteins in fingermarks and determination of sex by MALDI mass spectrometry profiling. *Analyst*, 2012, **137**(20): 4686–4692.

Fierer, N., et al., Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A*, 2010, **107**(14): 6477–6481.

Fleming, R.I. and S. Harbison, The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. *Forensic Sci Int Genet*, 2010, **4**(4): 244–256.

Fleming, R.I. and S. Harbison, The use of bacteria for the identification of vaginal secretions. *Forensic Sci Int Genet*, 2010, **4**(5): 311–315.

Fordyce, S.L., et al., Long-term RNA persistence in postmortem contexts. *Investig Genet*, 2013, **4**(1): 7.

Frascione, N., et al., Detection and identification of body fluid stains using antibody-nanoparticle conjugates. *Analyst*, 2012, **137**(2): 508–512.

Frumkin, D., et al., Authentication of forensic DNA samples. *Forensic Sci Int Genet*, 2010, **4**(2): 95–103.

Frumkin, D., et al., DNA methylation-based forensic tissue identification. *Forensic Sci Int Genet*, 2011, **5**(5): 517–524.

Gao, L.L., et al., Application and progress of RNA in forensic science. *J Forensic Med*, 2011, **27**(6): 455–459.

Gauvin, J., et al., Forensic pregnancy diagnostics with placental mRNA markers. *Int J Legal Med*, 2010, **124**(1): 13–17.

Gelmini, S., et al., Real-time quantitative reverse transcriptase-polymerase chain reaction (RT-PCR) for the measurement of prostate-specific antigen mRNA in the peripheral blood of patients with prostate carcinoma using the taqman detection system. *Clin Chem Lab Med*, 2001, **39**(5): 385–391.

Giampaoli, S., et al., Molecular identification of vaginal fluid by microbial signature. *Forensic Sci Int Genet*, 2012, **6**(5): 559–564.

Gipson, I.K., et al., Mucin genes expressed by human female reproductive tract epithelia. *Biol Reprod*, 1997, **56**(4): 999–1011.

Gipson, I.K., et al., MUC4 and MUC5B transcripts are the prevalent mucin messenger ribonucleic acids of the human endocervix. *Biol Reprod*, 1999, **60**(1): 58–64.

Goga, H., Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners. *Int J Legal Med*, 2012, **126**(5): 815–823.

Grskovic, B., et al., DNA methylation: The future of crime scene investigation? *Mol Biol Rep*, 2013, **40**(7): 4349–4360.

Gubin, A.N. and J.L. Miller, Human erythroid porphobilinogen deaminase exists in 2 splice variants. *Blood*, 2001, **97**(3): 815–817.

Gunn, A. and S.J. Pitt, Microbes as forensic indicators. *Trop Biomed*, 2012, **29**(3): 311–330.

Haas, C., E. Hanson, and J. Ballantyne, Capillary electrophoresis of a multiplex reverse transcription-polymerase chain reaction to target messenger RNA markers for body fluid identification. *Methods Mol Biol*, 2012, **830**: 169–183.

Haas, C., et al., mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR. *Forensic Sci Int Genet*, 2009, **3**(2): 80–88.

Haas, C., et al., mRNA profiling for the identification of sperm and seminal plasma. *Forensic Sci Int Genet Suppl Series*, 2009, **2**(1): 534–535.

Haas, C., et al., Collaborative EDNAP exercises on messenger RNA/DNA co-analysis for body fluid identification (blood, saliva, semen) and STR profiling. *Forensic Sci Int Genet Suppl Series*, 2011, **3**(1): e5–e6.

Haas, C., et al., MRNA profiling for the identification of blood—Results of a collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2011, **5**(1): 21–26.

Haas, C., et al., Selection of highly specific and sensitive mRNA biomarkers for the identification of blood. *Forensic Sci Int Genet*, 2011, **5**(5): 449–458.

Haas, C., et al., RNA/DNA co-analysis from blood stains—Results of a second collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2012, **6**(1): 70–80.

Haas, C., et al., RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: Results of a fourth and fifth collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2014, **8**(1): 203–212.

Haas, C., et al., RNA/DNA co-analysis from human saliva and semen stains—Results of a third collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2013, **7**(2): 230–239.

Hampson, C., J. Louhelainen, and S. McColl, An RNA expression method for aging forensic hair samples. *J Forensic Sci*, 2011, **56**(2): 359–365.

Hanson, E.K. and J. Ballantyne, Circulating microRNA for the identification of forensically relevant body fluids. *Methods Mol Biol*, 2013, **1024**: 221–234.

Hanson, E.K. and J. Ballantyne, Highly specific mRNA biomarkers for the identification of vaginal secretions in sexual assault investigations. *Sci Justice*, 2013, **53**(1): 14–22.

Hanson, E.K., H. Lubenow, and J. Ballantyne, Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Anal Biochem*, 2009, **387**(2): 303–314.

Hanson, E., et al., Identification of skin in touch/contact forensic samples by messenger RNA profiling. *Forensic Sci Int Genet Suppl Series*, 2011, **3**(1): e305–e306.

Hanson, E., et al., Specific and sensitive mRNA biomarkers for the identification of skin in 'touch DNA' evidence. *Forensic Sci Int Genet*, 2012, **6**(5): 548–558.

Harteveld, J., A. Lindenbergh, and T. Sijen, RNA cell typing and DNA profiling of mixed samples: Can cell types and donors be associated? *Sci Justice*, 2013, **53**(3): 261–269.

Heinrich, M., et al., Real-time PCR detection of five different "endogenous control gene" transcripts in forensic autopsy material. *Forensic Sci Int Genet*, 2007, **1**(2): 163–169.

Hochmeister, M.N., et al., Evaluation of prostate-specific antigen (PSA) membrane test assays for the forensic identification of seminal fluid. *J Forensic Sci*, 1999, **44**(5): 1057–1060.

Hochmeister, M.N., et al., Validation studies of an immunochromatographic 1-step test for the forensic identification of human blood. *J Forensic Sci*, 1999, **44**(3): 597–602.

Hsu, L., et al., Amplification of oral streptococcal DNA from human incisors and bite marks. *Curr Microbiol*, 2012, **65**(2): 207–211.

Iida, R., T. Yasuda, and K. Kishi, Identification of novel fibronectin fragments detected specifically in juvenile urine. *FEBS J*, 2007, **274**(15): 3939–3947.

Jakubowska, J., A. Maciejewska, and R. Pawłowski, mRNA profiling in identification of biological fluids in forensic genetics. *Z Zagadnien Nauk Sadowych*, 2011, **87**: 204–215.

Jakubowska, J., et al., mRNA profiling for vaginal fluid and menstrual blood identification. *Forensic Sci Int Genet*, 2013, **7**(2): 272–278.

Johnston, S., J. Newman, and R. Frappier, Validation study of the abacus diagnostics ABAcard HemaTrace Membrane Test for the forensic identification of human blood. *Can Soc Forensic Sci J*, 2003, **36**(3): 173–183.

Juusola, J. and J. Ballantyne, Messenger RNA profiling: A prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int*, 2003, **135**(2): 85–96.

Juusola, J. and J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**(1): 1–12.

Juusola, J. and J. Ballantyne, mRNA profiling for body fluid identification by multiplex quantitative RT-PCR. *J Forensic Sci*, 2007, **52**(6): 1252–1262.

Kakizaki, E., et al., Bioluminescent bacteria have potential as a marker of drowning in seawater: Two immersed cadavers retrieved near estuaries. *Leg Med*, 2009, **11**(2): 91–96.

Kakizaki, E., et al., Detection of marine and freshwater bacterioplankton in immersed victims: Post-mortem bacterial invasion does not readily occur. *Forensic Sci Int*, 2011, **211**(1–3): 9–18.

Kakizaki, E., et al., *In vitro* study of possible microbial indicators for drowning: Salinity and types of bacterioplankton proliferating in blood. *Forensic Sci Int*, 2011, **204**(1–3): 80–87.

Kakizaki, E., et al., Detection of diverse aquatic microbes in blood and organs of drowning victims: First metagenomic approach using high-throughput 454-pyrosequencing. *Forensic Sci Int*, 2012, **220**(1–3): 135–146.

Kohlmeier, F. and P.M. Schneider, Successful mRNA profiling of 23 years old blood stains. *Forensic Sci Int Genet*, 2012, **6**(2): 274–276.

Koppelkamm, A., et al., Validation of adequate endogenous reference genes for the normalisation of qPCR gene expression data in human post mortem tissue. *Int J Legal Med*, 2010, **124**(5): 371–380.

Koppelkamm, A., et al., RNA integrity in post-mortem samples: Influencing parameters and implications on RT-qPCR assays. *Int J Legal Med*, 2011, **125**(4): 573–580.

Kumagai, A., N. Nakayashiki, and Y. Aoki, Analysis of age-related carbonylation of human vitreous humor proteins as a tool for forensic diagnosis. *Leg Med (Tokyo)*, 2007, **9**(4): 175–180.

Lee, H.Y., et al., Potential forensic application of DNA methylation profiling to body fluid identification. *Int J Legal Med*, 2012, **126**(1): 55–62.

Lee, H.Y., et al., Potential forensic application of DNA methylation profiling to body fluid identification. *Int J Legal Med*, 2012, **126**(1): 55–62.

Lehman, D.C., Forensic microbiology. *Clin Lab Sci*, 2012, **25**(2): 114–119.

Li, W.C., et al., Estimation of postmortem interval using microRNA and 18S rRNA degradation in rat cardiac muscle. *J Forensic Med*, 2010, **26**(6): 413–417.

Li, C., et al., Differences of DNA methylation profiles between monozygotic twins' blood samples. *Mol Biol Rep*, 2013, **40**(9): 5275–5280.

Lincoln, P.J. and B.E. Dodd, The use of low ionic strength solution (LISS) in elution experiments and in combination with papain-treated cells for the titration of various antibodies, including eluted antibody. *Vox Sang*, 1978, **34**(4): 221–226.

Lindenbergh, A., et al., A multiplex (m)RNA-profiling system for the forensic identification of body fluids and contact traces. *Forensic Sci Int Genet*, 2012, **6**(5): 565–577.

Lindenbergh, A., P. Maaskant, and T. Sijen, Implementation of RNA profiling in forensic casework. *Forensic Sci Int Genet*, 2013, **7**(1): 159–166.

Lucci, A., et al., A promising microbiological test for the diagnosis of drowning. *Forensic Sci Int*, 2008, **182**(1–3): 20–26.

Madi, T., et al., The determination of tissue-specific DNA methylation patterns in forensic biofluids using bisulfite modification and pyrosequencing. *Electrophoresis*, 2012, **33**(12): 1736–1745.

Moreno, L.I., et al., Determination of an effective housekeeping gene for the quantification of mRNA for forensic applications. *J Forensic Sci*, 2012, **57**(4): 1051–1058.

Nagasawa, S., et al., Detection of *Helicobacter pylori* (*H. pylori*) DNA in digestive systems from cadavers by real-time PCR. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S458–S459.

Nakanishi, H., et al., A novel method for the identification of saliva by detecting oral streptococci using PCR. *Forensic Sci Int*, 2009, **183**(1–3): 20–23.

Nakanishi, H., et al., A simple identification method of saliva by detecting *Streptococcus salivarius* using loop-mediated isothermal amplification. *J Forensic Sci*, 2011, **56**(Suppl. 1): S158–S167.

Nakanishi, H., et al., Identification of feces by detection of *Bacteroides* genes. *Forensic Sci Int Genet*, 2013, **7**(1): 176–179.

Nakatome, M., et al., Methylation analysis of circadian clock gene promoters in forensic autopsy specimens. *Leg Med (Tokyo)*, 2011, **13**(4): 205–209.

Nakayashiki, N., et al., Analysis of the methylation profiles in imprinted genes applicable to parental allele discrimination. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S471–S472.

Nakayashiki, N., et al., Investigation of the methylation status around parent-of-origin detectable SNPs in imprinted genes. *Forensic Sci Int Genet*, 2009, **3**(4): 227–232.

Nussbaumer, C., E. Gharehbaghi-Schnell, and I. Korschineck, Messenger RNA profiling: A novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**(2–3): 181–186.

Oakley, C.L. and A.J. Fulthorpe, Antigenic analysis by diffusion. *J Pathol Bacteriol*, 1953, **65**(1): 49–60.

Odriozola, A., et al., miRNA analysis in vitreous humor to determine the time of death: A proof-of-concept pilot study. *Int J Legal Med*, 2013, **127**(3): 573–578.

Omelia, E.J., M.L. Uchimoto, and G. Williams, Quantitative PCR analysis of blood- and saliva-specific microRNA markers following solid-phase DNA extraction. *Anal Biochem*, 2013, **435**(2): 120–122.

Ota, M., et al., Restriction enzyme analysis of PCR products. *Methods Mol Biol*, 2009, **578**: 405–414.

Ouchterlony, O., Antigen-antibody reactions in gels. *Acta Pathol Microbiol Scand*, 1949, **26**: 507.

Pang, B.C. and B.K. Cheung, Identification of human semenogelin in membrane strip test as an alternative method for the detection of semen. *Forensic Sci Int*, 2007, **169**(1): 27–31.

Park, J.L., et al., Forensic body fluid identification by analysis of multiple RNA markers using nanoString technology. *Genomics Inform*, 2013, **11**(4): 277–281.

Park, S.M., et al., Genome-wide mRNA profiling and multiplex quantitative RT-PCR for forensic body fluid identification. *Forensic Sci Int Genet*, 2013, **7**(1): 143–150.

Parker, C., E. Hanson, and J. Ballantyne, Optimization of dried stain co-extraction methods for efficient recovery of high quality DNA and RNA for forensic analysis. *Forensic Sci Int Genet Suppl Series*, 2011, **3**(1): e309–e310.

Partemi, S., et al., Analysis of mRNA from human heart tissue and putative applications in forensic molecular pathology. *Forensic Sci Int*, 2010, **203**(1–3): 99–105.

Petrie, G.F., A specific precipitin reaction associated with with growth on agar plates on *meningococcus*, *pneumococcus*, and *B. dysenteriae* (shiga). *Br J Exp Pathol*, 1932, **13**: 380.

Phang, T.W., et al., Amplification of cDNA via RT-PCR using RNA extracted from postmortem tissues. *J Forensic Sci*, 1994, **39**(5): 1275–1279.

Power, D.A., et al., PCR-based detection of salivary bacteria as a marker of expirated blood. *Sci Justice*, 2010, **50**(2): 59–63.

Qi, B., L. Kong, and Y. Lu, Gender-related difference in bloodstain RNA ratio stored under uncontrolled room conditions for 28 days. *J Forensic Legal Med*, 2013, **20**(4): 321–325.

Qin, J., et al., Application situation and prospect of mRNA in forensic medicine. *Chin J Forensic Med*, 2013, **28**(1): 30–33.

Quarino, L., et al., An ELISA method for the identification of salivary amylase. *J Forensic Sci*, 2005, **50**(4): 873–876.

Richard, M.L., et al., Evaluation of mRNA marker specificity for the identification of five human body fluids by capillary electrophoresis. *Forensic Sci Int Genet*, 2012, **6**(4): 452–460.

Roeder, A.D. and C. Haas, mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification. *Int J Legal Med*, 2013, **127**(4): 707–721.

Sabatini, L.M., Y.Z. He, and E.A. Azen, Structure and sequence determination of the gene encoding human salivary statherin. *Gene*, 1990, **89**(2): 245–251.

Saeed, M., R.M. Berlin, and T.D. Cruz, Exploring the utility of genetic markers for predicting biological age. *Leg Med (Tokyo)*, 2012, **14**(6): 279–285.

Sakurada, K., et al., Evaluation of mRNA-based approach for identification of saliva and semen. *Leg Med*, 2009, **11**(3): 125–128.

Sakurada, K., et al., Identification of nasal blood by real-time RT-PCR. *Leg Med (Tokyo)*, 2012, **14**(4): 201–204.

Salamonsen, L.A. and D.E. Woolley, Matrix metalloproteinases in normal menstruation. *Hum Reprod*, 1996, **11**(Suppl 2): 124–133.

Salamonsen, L.A. and D.E. Woolley, Menstruation: Induction by matrix metalloproteinases and inflammatory cells. *J Reprod Immunol*, 1999, **44**(1–2): 1–27.

Sampaio-Silva, F., et al., Profiling of RNA degradation for estimation of post mortem interval. *PLoS One*, 2013, **8**(2): e56507.

Sato, I., et al., Rapid detection of semenogelin by one-step immunochromatographic assay for semen identification. *J Immunol Methods*, 2004, **287**(1–2): 137–145.

Schulz, M.M., et al., A new approach to the investigation of sexual offenses—Cytoskeleton analysis reveals the origin of cells found on forensic swabs. *J Forensic Sci*, 2010, **55**(2): 492–498.

Seidl, S., R. Hausmann, and P. Betz, Comparison of laser and mercury-arc lamp for the detection of body fluids on different substrates. *Int J Legal Med*, 2008, **122**(3): 241–244.

Setzer, M., J. Juusola, and J. Ballantyne, Recovery and stability of RNA in vaginal swabs and blood, semen, and saliva stains. *J Forensic Sci*, 2008, **53**(2): 296–305.

Sikirzhytski, V., A. Sikirzhytskaya, and I.K. Lednev, Multidimensional Raman spectroscopic signatures as a tool for forensic identification of body fluid traces: A review. *Appl Spectrosc*, 2011, **65**(11): 1223–1232.

Sikirzhytski, V., K. Virkler, and I.K. Lednev, Discriminant analysis of Raman spectra for body fluid identification for forensic purposes. *Sensors (Basel)*, 2010, **10**(4): 2869–2884.

Sikirzhytskaya, A., V. Sikirzhytski, and I.K. Lednev, Raman spectroscopic signature of vaginal fluid and its potential application in forensic body fluid identification. *Forensic Sci Int*, 2012, **216**(1–3): 44–48.

Sikirzhytski, V., A. Sikirzhytskaya, and I.K. Lednev, Advanced statistical analysis of Raman spectroscopic data for the identification of body fluid traces: Semen and blood mixtures. *Forensic Sci Int*, 2012, **222**(1–3): 259–265.

Sikirzhytski, V., A. Sikirzhytskaya, and I.K. Lednev, Multidimensional Raman spectroscopic signature of sweat and its potential application to forensic body fluid identification. *Anal Chim Acta*, 2012, **718**: 78–83.

Spencer, S.E., et al., Matrix-assisted laser desorption/ionization-time of flight-mass spectrometry profiling of trace constituents of condom lubricants in the presence of biological fluids. *Forensic Sci Int*, 2011, **207**(1–3): 19–26.

Steger, K., et al., Expression of protamine-1 and -2 mRNA during human spermiogenesis. *Mol Hum Reprod*, 2000, **6**(3): 219–225.

Suto, M., et al., PCR detection of bacterial genes provides evidence of death by drowning. *Leg Med*, 2009, **11**(Suppl. 1): S354–S356.

Sweet, G.H. and J.W. Elvins, Studies by crossed electroimmunodiffusion on the individuality and sexual origin of bloodstains. *J Forensic Sci*, 1976, **21**(3): 498–509.

Tims, S., et al., Microbial DNA fingerprinting of human fingerprints: Dynamic colonization of fingertip microflora challenges human host inferences for forensic purposes. *Int J Legal Med*, 2010, **124**(5): 477–481.

Tims, S., et al., Microbial DNA fingerprinting of human fingerprints: Dynamic colonization of fingertip microflora challenges human host inferences for forensic purposes. *Int J Legal Med*, 2010, **124**(5): 477–481.

Tran, T.N., et al., Classification of ancient mammal individuals using dental pulp MALDI-TOF MS peptide profiling. *PLoS One*, 2011, **6**(2): e17319.

Tran-Hung, L., et al., A new method to extract dental pulp DNA: Application to universal detection of bacteria. *PLoS One*, 2007, **2**(10): e1062.

Uchimoto, M.L., et al., Considering the effect of stem-loop reverse transcription and real-time PCR analysis of blood and saliva specific microRNA markers upon mixed body fluid stains. *Forensic Sci Int Genet*, 2013, **7**(4): 418–421.

Uchiyama, T., et al., A new molecular approach to help conclude drowning as a cause of death: Simultaneous detection of eight bacterioplankton species using real-time PCR assays with TaqMan probes. *Forensic Sci Int*, 2012, **222**(1–3): 11–26.

Valore, E.V., et al., Human beta-defensin-1: An antimicrobial peptide of urogenital tissues. *J Clin Invest*, 1998, **101**(8): 1633–1642.

van der Meer, D., M.L. Uchimoto, and G. Williams, Simultaneous analysis of micro-RNA and DNA for determining the body fluid origin of DNA profiles. *J Forensic Sci*, 2013, **58**(4): 967–971.

Van Steendam, K., et al., Mass spectrometry-based proteomics as a tool to identify biological matrices in forensic science. *Int J Legal Med*, 2013, **127**(2): 287–298.

Vennemann, M. and A. Koppelkamm, mRNA profiling in forensic genetics I: Possibilities and limitations. *Forensic Sci Int*, 2010, **203**(1–3): 71–75.

Vennemann, M. and A. Koppelkamm, Postmortem mRNA profiling II: Practical considerations. *Forensic Sci Int*, 2010, **203**(1–3): 76–82.

Virkler, K. and I.K. Lednev, Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene. *Forensic Sci Int*, 2009, **188**(1–3): 1–17.

Virkler, K. and I.K. Lednev, Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis. *Anal Chem*, 2009, **81**(18): 7773–7777.

Virkler, K. and I.K. Lednev, Raman spectroscopic signature of semen and its potential application to forensic body fluid identification. *Forensic Sci Int*, 2009, **193**(1–3): 56–62.

Virkler, K. and I.K. Lednev, Raman spectroscopy offers great potential for the nondestructive confirmatory identification of body fluids. *Forensic Sci Int*, 2008, **181**(1–3): e1–e5.

Virkler, K. and I.K. Lednev, Forensic body fluid identification: The Raman spectroscopic signature of saliva. *Analyst*, 2010, **135**(3): 512–517.

Virkler, K. and I.K. Lednev, Raman spectroscopic signature of blood and its potential application to forensic body fluid identification. *Anal Bioanal Chem*, 2010, **396**(1): 525–534.

Visser, M., et al., mRNA-based skin identification for forensic applications. *Int J Legal Med*, 2011, **125**(2): 253–263.

Wang, Z., et al., A model for data analysis of microRNA expression in forensic body fluid identification. *Forensic Sci Int Genet*, 2012, **6**(3): 419–423.

Wang, Z., et al., Screening and confirmation of microRNA markers for forensic body fluid identification. *Forensic Sci Int Genet*, 2013, **7**(1): 116–123.

Wasserstrom, A., et al., Demonstration of DSI-semen—A novel DNA methylation-based forensic semen identification assay. *Forensic Sci Int Genet*, 2013, **7**(1): 136–142.

Wiener, A.S., M.A. Hyman, and L. Handman, A new serological test (inhibition test) for human serum globulin. *Proc Soc Exp Biol Med*, 1949, **71**: 96.

Wilkinson, D.A., et al., The effects of aerosolized bacteria on fingerprint impression evidence. *J Forensic Identif*, 2009, **59**(1): 65–79.

Xu, H., et al., Bisulfite genomic sequencing of DNA from dried blood spot microvolume samples. *Forensic Sci Int Genet*, 2012, **6**(3): 306–309.

Yoshida, K., et al., Quantification of seminal plasma motility inhibitor/semenogelin in human seminal plasma. *J Androl*, 2003, **24**(6): 878–884.

Young, S.T., et al., Estimating postmortem interval using RNA degradation and morphological changes in tooth pulp. *Forensic Sci Int*, 2013, **229**(1–3): 163.e1–e6.

Zhao, D., et al., Postmortem mRNA quantification for investigation of infantile death: A comparison with adult cases. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S286–S289.

Zhao, D., et al., Postmortem quantitative mRNA analyses of death investigation in forensic pathology: An overview and prospects. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S43–S45.

Zhao, G., et al., Study on the application of parent-of-origin specific DNA methylation markers to forensic genetics. *Forensic Sci Int*, 2005, **154**(2–3): 122–127.

Zhao, G.S. and Q.E. Yang, Applications of DNA methylation markers in forensic medicine. *Fa Yi Xue Za Zhi*, 2005, **21**(1): 61–64.

Zhao, S.M. and C.T. Li, Perspective of DNA methylation in forensic genetics and new progress of its detection methods. *J Forensic Med*, 2009, **25**(4): 290–295.

Zhao, S.M., et al., Differences of DNA methylation profiles in monozygotic twins' blood samples. *J Forensic Med*, 2011, **27**(4): 260–264.

Zubakov, D., et al., MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. *Int J Legal Med*, 2010, **124**(3): 217–226.

Zubakov, D., et al., New markers for old stains: Stable mRNA markers for blood and saliva identification from up to 16-year-old stains. *Int J Legal Med*, 2009, **123**(1): 71–74.

Zubakov, D., et al., Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples. *Int J Legal Med*, 2008, **122**(2): 135–142.

SECTION III

Identification of Biological Evidence

12

Identification of Blood

12.1 Biological Properties

Blood constitutes about 8% of the human body weight of a healthy individual. *Plasma* is the fluid portion of the blood. The cellular portion of the blood consists of *red blood cells*, *white blood cells*, and *platelets*, all of which are suspended in the plasma (Figures 12.1 and 12.2).

12.1.1 Red Blood Cells

These cells are also called *erythrocytes*. Their life span in humans is approximately 3–4 months. Additionally, mature human erythrocytes do not have nuclei, and therefore lack nuclear DNA. Erythrocytes consist of hemoglobin—proteins that are responsible for the transportation of oxygen. Most adult human hemoglobin consists of four polypeptide chains, two α chains and two β chains. Thus, adult hemoglobin is designated as $\alpha_2\beta_2$. Other forms of hemoglobin will be discussed in Chapter 18. Under normal physiological conditions, each hemoglobin subunit contains a heme moiety that binds to oxygen (Figure 12.3). A heme molecule consists of an organic component known as protoporphyrin IX and a *ferrous* (Fe^{2+}) iron ion (Figure 12.4). A heme molecule is also known as *ferroprotoporphyrin*. The ferrous ion of heme forms four bonds with the nitrogens of protoporphyrin IX, along with a fifth bond with a hemoglobin chain and a sixth bond with a molecule of O_2. Other chemicals such as carbon monoxide and cyanide also bind to the ferrous iron of the heme molecule and can cause chemical asphyxia. Heme groups are also present in the blood of various animals and in other proteins such as myoglobin in muscles and neuroglobin in the brain.

12.1.2 White Blood Cells

Also called leucocytes, white blood cells are subdivided into three types: granulocytes, lymphocytes, and monocytes. White blood cells are involved in defending the body against infection. They have nuclei and thus represent the main sources of nuclear DNA from the blood.

12.1.3 Platelets

These cells are also known as thrombocytes, and they play a role in blood clotting (Chapter 16). Platelets aggregate at sites of vascular and blood vessel injury. Like erythrocytes, they lack nuclei.

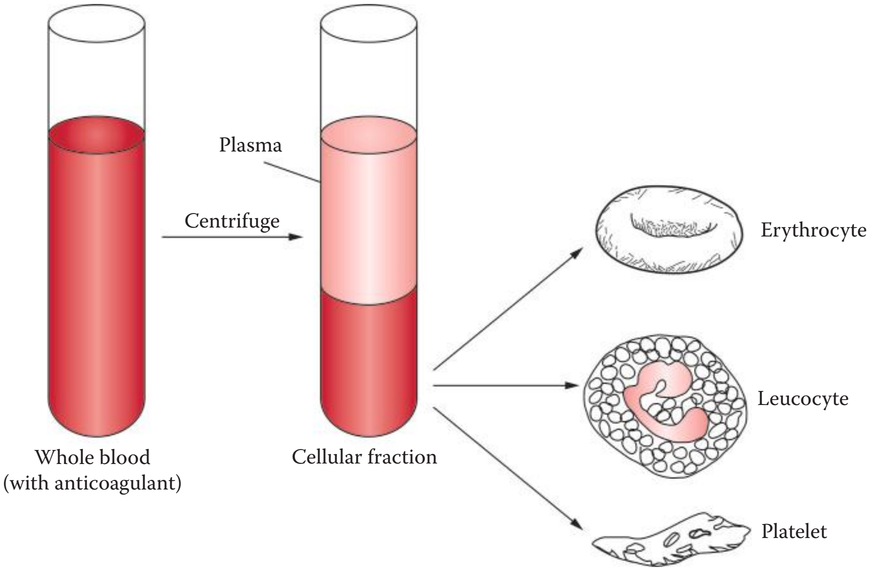

Figure 12.1 Basic composition of blood. Blood can be separated into two phases in the presence of an anticoagulant. The liquid portion called plasma accounts for approximately 55% of blood volume. The cellular elements include erythrocytes, leucocytes, and platelets. (© Richard C. Li.)

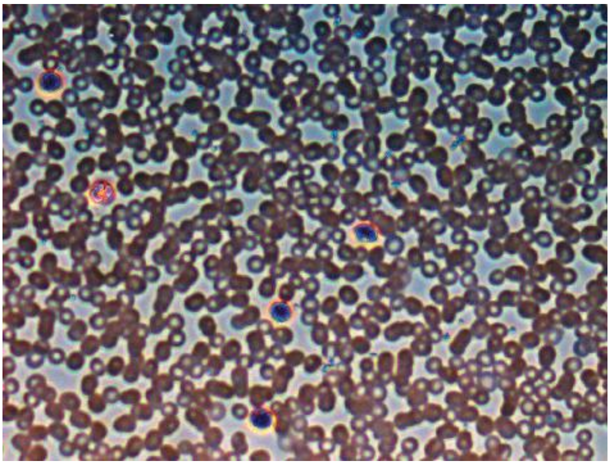

Figure 12.2 Blood cells. (© Richard C. Li.)

12.2 Presumptive Assays for Identification

12.2.1 Mechanisms of Presumptive Assays

Presumptive blood assays are designed to detect traces of blood. These assays are based on the basic principle of the oxidation–reduction reaction catalyzed by the heme moiety of the hemoglobin. As a result, colorless substrates catalyzed by heme undergo an oxidation reaction, causing either chemiluminescence, fluorescence, or a change of color. These assays are very sensitive and can detect blood in samples with 10^{-5}–10^{-6}-fold dilutions. A positive reaction indicates the possible presence of blood. Additionally, most of these assays do not interfere with forensic DNA analysis.

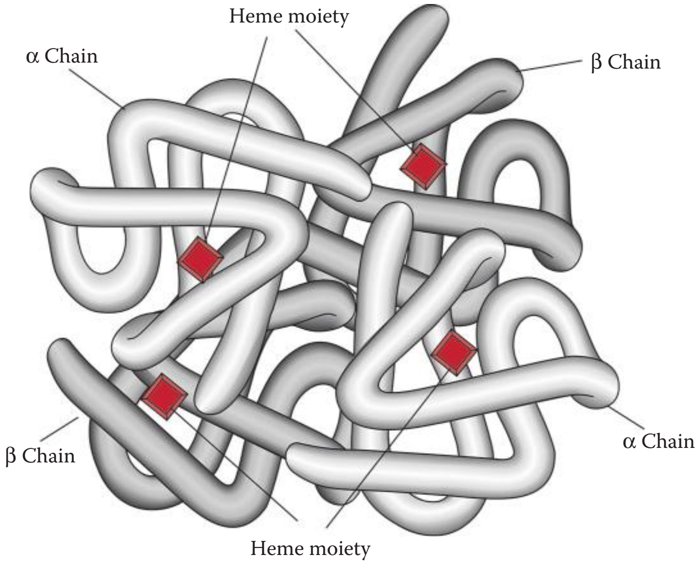
Heme moiety
α Chain
β Chain
β Chain
α Chain
Heme moiety

Figure 12.3 Human adult hemoglobin. Four subunits, two α and two β chains, of human adult hemoglobin are shown. Each hemoglobin subunit contains a heme moiety. (© Richard C. Li.)

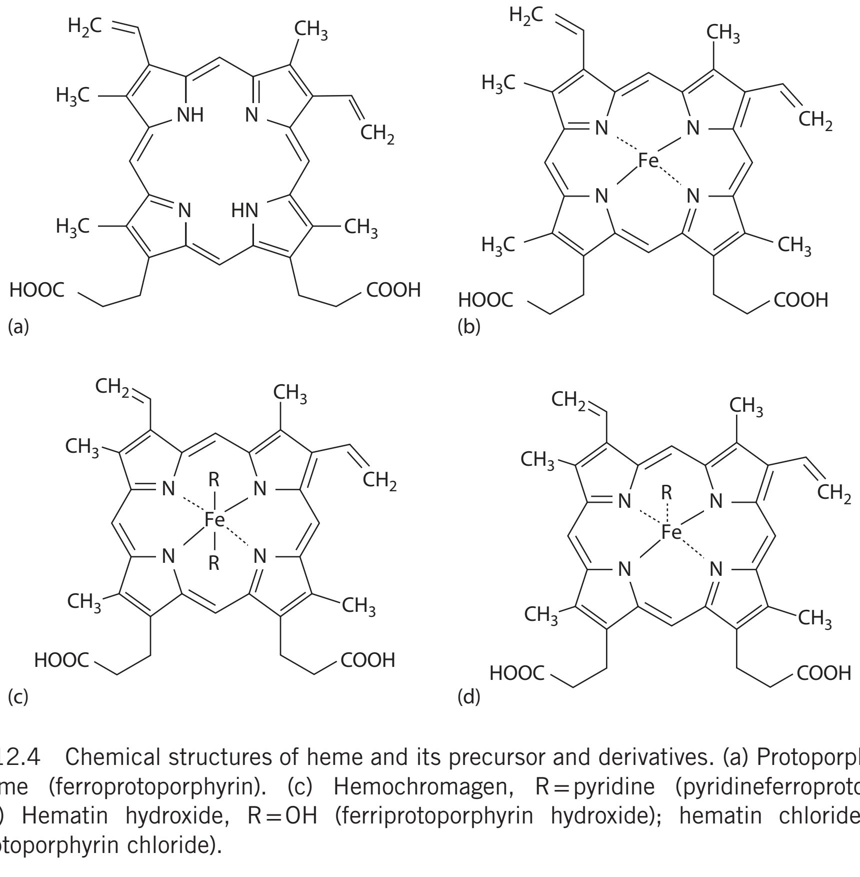
H2C
CH3
H3C
NH
N
CH2
N
HN
H3C
CH3
HOOC
COOH
(a)
H2C
CH3
H3C
N
N
CH2
Fe
N
N
H3C
CH3
HOOC
COOH
(b)
CH2
CH3
CH3
N
R
N
CH2
Fe
N
R
N
CH3
CH3
HOOC
COOH
(c)
CH2
CH3
CH3
N
R
N
CH2
Fe
N
N
CH3
CH3
HOOC
COOH
(d)

Figure 12.4 Chemical structures of heme and its precursor and derivatives. (a) Protoporphyrin IX. (b) Heme (ferroprotoporphyrin). (c) Hemochromagen, R = pyridine (pyridineferroprotoporphyrin). (d) Hematin hydroxide, R = OH (ferriprotoporphyrin hydroxide); hematin chloride, R = Cl (ferriprotoporphyrin chloride).

12.2.1.1 Oxidation–Reduction Reactions

An oxidation–reduction reaction involves a change in the oxidation state of a molecule. Specifically, the *oxidation* of a molecule means that the molecule has lost electrons, and the *reduction* of a molecule means that the molecule has gained electrons. Chemicals that can be reduced and therefore gain electrons from other molecules are called *oxidants*. In contrast, *reductants* are chemicals that can be oxidized and therefore lose electrons to other molecules. In biochemical reactions, oxidation often coincides with a loss of hydrogen. Figure 12.5 depicts an example of an oxidation–reduction reaction for blood identification. In presumptive assays, heme is utilized as the catalyst, and hydrogen peroxide is utilized as the oxidant for the reaction. In the presence of heme, a colorless substrate is oxidized, yielding a product with color, chemiluminescence, or fluorescence.

12.2.2 Colorimetric Assays

Many procedures are available for detecting heme in blood through color reactions. The most common agents are phenolphthalin, leucomalachite green, and benzidine derivatives. The color reactions produced by these assays can be observed immediately with the naked eye.

12.2.2.1 Phenolphthalin Assay

Phenolphthalein, a member of a class of indicators and dyes, is used in titrations of mineral and organic acids as well as most alkalis. The phenolphthalin assay for blood identification is also known as the Kastle–Meyer test. Kastle published a study in 1901, presenting the results of a reaction in which phenolphthalin, a colorless compound, is catalyzed by heme with hydrogen peroxide as the oxidant (Figures 12.6 and 12.7). The oxidized derivative is phenolphthalein, which appears pink under alkaline conditions.

12.2.2.2 Leucomalachite Green (LMG) Assay

Malachite green is a triphenylmethane dye. The leuco base form of malachite green is colorless and can be oxidized by the catalysis of heme to produce a green color. The reaction is carried out under acid conditions with hydrogen peroxide as the oxidant (Figures 12.8 and 12.9).

12.2.2.3 Benzidine and Derivatives

Historically, *benzidine* was used as an intermediate in dye manufacturing (Figure 12.10). Subsequently, it was used as a presumptive assay for the presence of blood after the discovery that the oxidation of benzidine can be catalyzed by heme to produce a blue to dark blue color (carried out in an acid solution). Since the blue color may eventually turn brown, the reaction must be read

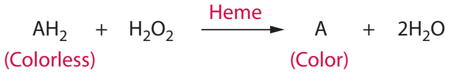

$$\underset{\text{(Colorless)}}{AH_2} + H_2O_2 \xrightarrow{\text{Heme}} \underset{\text{(Color)}}{A} + 2H_2O$$

Figure 12.5 Oxidation–reduction reaction as the basis for presumptive assays for blood identification. AH_2, substrate; A, oxidized substrate.

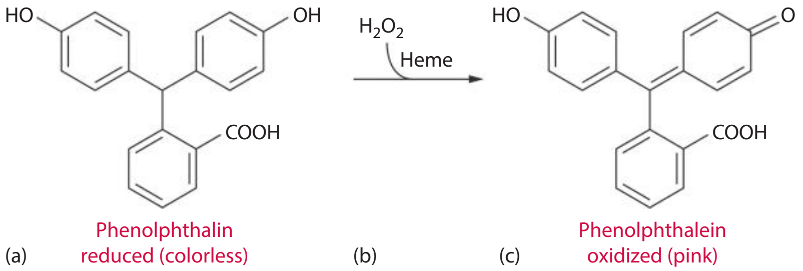

Figure 12.6 Chemical reaction of phenolphthalin assay.

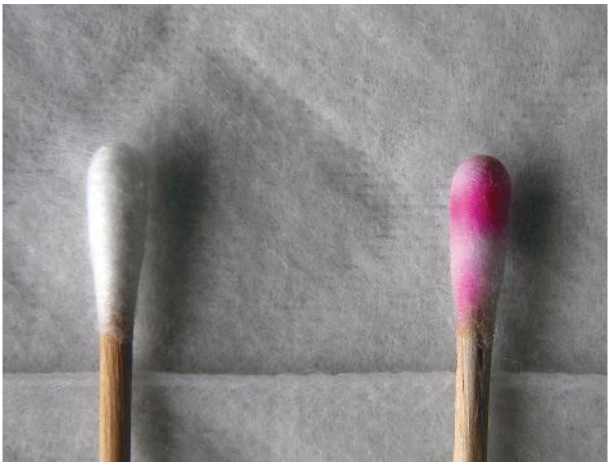

Figure 12.7 Photograph of phenolphthalin assay results. Negative (left) and positive (right) reaction. (© Richard C. Li.)

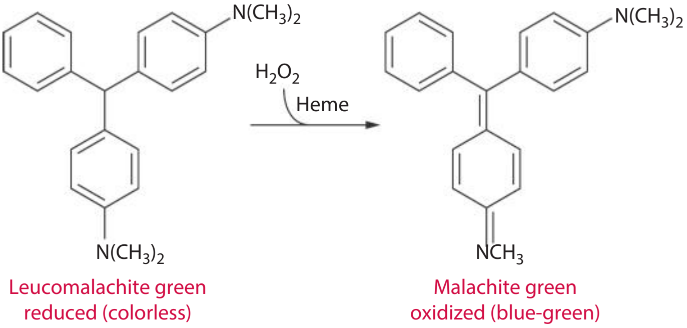
N(CH3)2
H2O2
Heme
N(CH3)2
N(CH3)2
NCH3
Leucomalachite green
reduced (colorless)
Malachite green
oxidized (blue-green)

Figure 12.8 Chemical reaction of leucomalachite green assay.

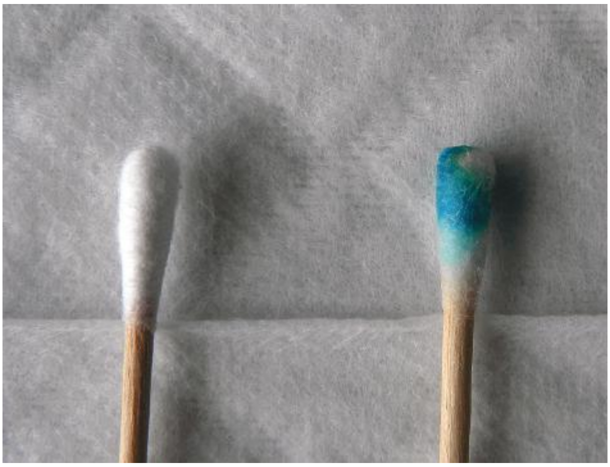

Figure 12.9 Photograph of leucomalachite green assay results. Negative (left) and positive (right) reaction. (© Richard C. Li.)

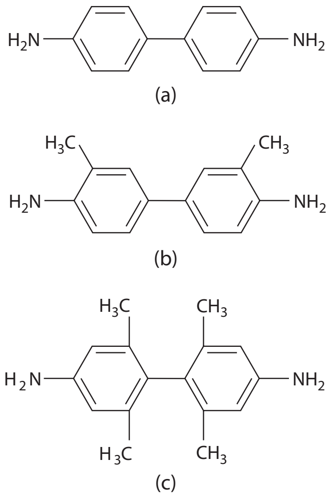

Figure 12.10 Chemical structures of benzidine and derivatives: (a) benzidine; (b) orthotolidine; (c) tetramethylbenzidine.

immediately. Benzidine was found to be a carcinogen and is therefore no longer used for forensic testing. *Orthotolidine* is a dimethyl derivative of benzidine. Its oxidation reaction can be catalyzed by heme to produce a blue color reaction under acidic conditions (Figure 12.11). Orthotolidine is also considered a potential carcinogen based on animal studies, and for this reason it has been replaced by tetramethylbenzidine. *Tetramethylbenzidine* (TMB) is a tetramethyl derivative of benzidine. The oxidation of TMB can be catalyzed by heme to produce a green to blue-green color under acidic conditions. TMB continues to be used. The Hemastix® assay kit (Miles Laboratories) is a TMB-based assay that utilizes a TMB-containing strip device. A test is carried out by applying a moistened sample to a Hemastix® strip. The appearance of a green or a blue-green color indicates the presence of blood.

12.2.3 Chemiluminescence and Fluorescence Assays

Other organic compounds whose oxidation products have chemiluminescent or fluorescent properties are utilized for testing. In the *chemiluminescence assay*, light is emitted as a product of a chemical reaction. In this category, luminol produces chemiluminescence when blood is present. In contrast, a *fluorescence assay* requires the exposure of an oxidized product, such as fluorescin, to a particular wavelength of an excitation light source. The fluorescence is then emitted at longer wavelengths than that of the excitation light source.

One advantage of chemiluminescent and fluorescent reagents is that they can be sprayed over large areas where latent bloodstains are potentially located. A positive reaction identifies blood and also reveals the patterns of bloody impressions such as footprints and fingerprints. These methods are very sensitive and can pinpoint the locations of even small traces of blood. Additionally, they are useful for detecting blood at crime scenes that have been cleaned and show no visible staining. One disadvantage of chemiluminescent and fluorescent reagents is that precautions must be taken if stains are very small or have been washed. The spraying of

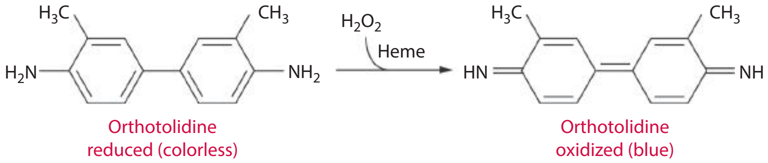
H₂O₂
Heme
Orthotolidine reduced (colorless)
Orthotolidine oxidized (blue)

Figure 12.11 Chemical reaction of orthotolidine assay.

presumptive assay reagents may further dilute a sample and thus lead to difficulty in isolating sufficient amounts of DNA for forensic DNA analysis.

12.2.3.1 Luminol (3-Aminophthalhydrazide)

Luminol is usually utilized as a chemiluminescent reagent. The oxidation reaction of luminol catalyzed by heme produces light in the presence of an oxidant (Figures 12.12 through 12.14). The light emitted from a positive reaction can only be observed in the dark, which limits the

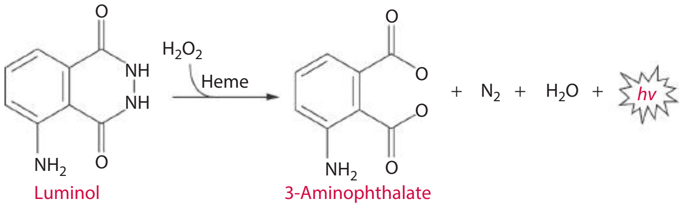

Figure 12.12 Chemical reaction of luminol assay.

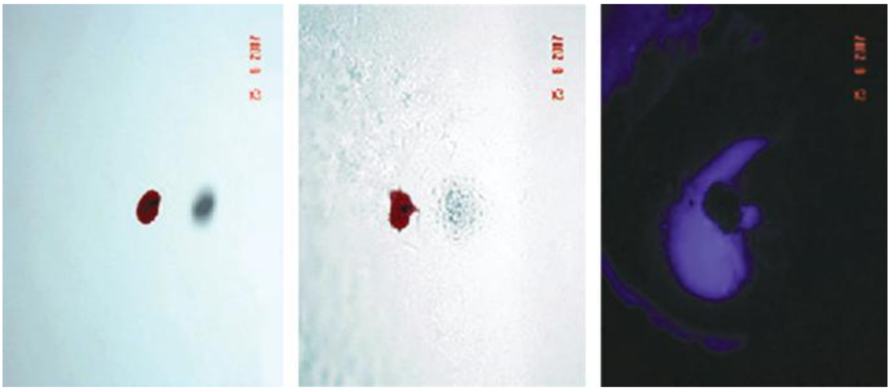

Figure 12.13 Results obtained from testing a drop of blood on a glass plate with luminol. From left to right: drop of blood; drop of blood after spraying with luminol; drop of blood after spraying with luminol and viewing in the dark. A blue chemiluminescence indicates the presence of blood. (From Bergervoet, P.W., et al., *J Hosp Infect*, 68, 329–333, 2008. With permission.)

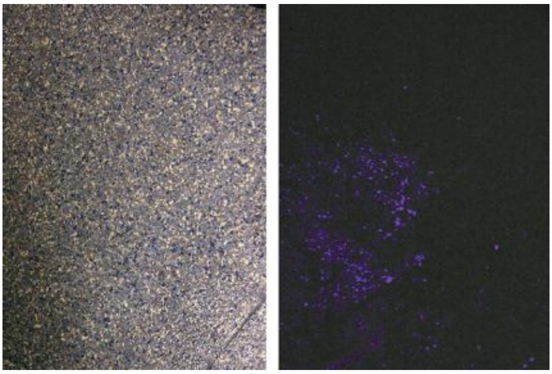

Figure 12.14 Detecting latent bloodstains on a floor with luminol. Left: with the light on; right: immediately after the light is switched off. A blue chemiluminescence indicates the presence of blood traces. (From Bergervoet, P.W., et al., *J Hosp Infect*, 68, 329–333, 2008. With permission.)

applications of luminol. The photodocumentation of a luminol-enhanced pattern should be done immediately before it fades away.

12.2.3.2 Fluorescin

Fluorescin is another reagent that is used to test for the presence of bloodstains at a crime scene (Figures 12.15 and 12.16). When oxidized and catalyzed by heme, fluorescin demonstrates fluorescent properties. Usually, fluorescin-sprayed stains are exposed to light in the range of 425–485 nm using an alternate light source device. In a positive reaction, the oxidized fluorescin emits an intense yellowish-green fluorescent light, which indicates the presence of a bloodstain. The light emitted from fluorescin-sprayed stains lasts longer than that of luminol.

12.2.4 Factors Affecting Presumptive Assay Results

The catalytic assays discussed in the previous sections are not specific to blood only, which can possibly lead to the observation of false-positive or false-negative results (Figure 12.17).

12.2.4.1 Oxidants

Chemicals that are strong oxidants may cause a false-positive reaction. Such chemicals can catalyze the oxidation reaction even in the absence of heme and result in a false-positive reaction. Certain metal salts, such as copper and nickel salts, household bleaches and cleaners that contain hypochlorite ions, and hair-coloring products that contain hydrogen peroxide, work as oxidants. To address this problem, a two-step catalytic assay should be performed. The substrate

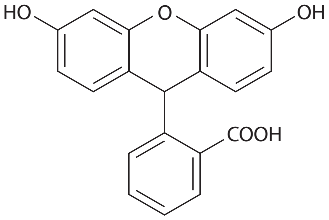

Figure 12.15 Chemical structure of fluorescein.

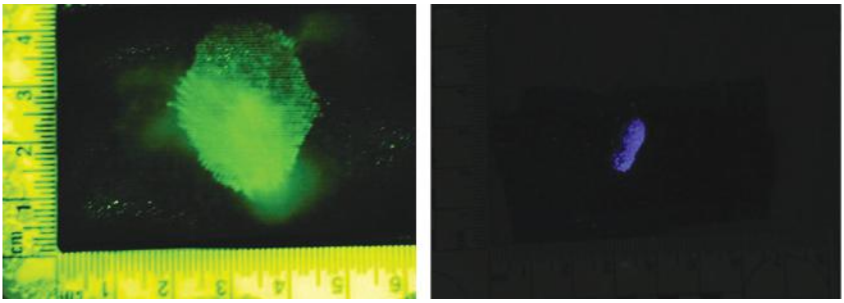

Figure 12.16 Detecting diluted bloodstains on black cotton fabric with fluorescin (left) and Bluestar (right). (From Finnis, J., Lewis, J., and Davidson, A., *Sci Justice*, 53, 178–186, 2013. With permission.)

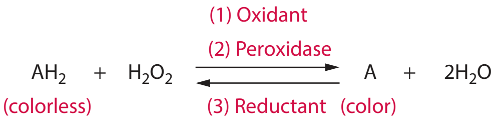

$$\underset{\text{(colorless)}}{AH_2} + H_2O_2 \underset{\text{(3) Reductant}}{\overset{\text{(1) Oxidant}}{\underset{}{\overset{\text{(2) Peroxidase}}{\rightleftarrows}}}} \underset{\text{(color)}}{A} + 2H_2O$$

Figure 12.17 Factors affecting presumptive assay results. Strong oxidant and peroxidase may cause false-positive results; reductant may cause false-negative results.

is applied first to the sample in question. A color change occurring before the addition of hydrogen peroxide indicates a false-positive result due to a possible oxidant in the sample. If a color change is observed after the addition of hydrogen peroxide, the result is a true positive.

12.2.4.2 Plant Peroxidases

Many types of plants such as horseradish contain peroxidases. Plant peroxidases may also catalyze oxidation reactions and lead to false-positive results. However, plant peroxidases are usually heat sensitive and may be inactivated by high temperatures. Because the heme molecule is relatively stable at high temperatures, samples can be retested after heating. This will inactivate any plant peroxidases in a sample.

12.2.4.3 Reductants

Although not common, a false-negative result can occur when a strong reductant is present in a sample. Strong reductants such as certain metal ions including lithium and zinc may inhibit the oxidation reaction.

12.3 Confirmatory Assays for Identification

12.3.1 Microcrystal Assays

Microcrystal assays apply chemicals to treat bloodstains, forming crystals of heme molecules. The morphologies of the resulting crystals are distinctive for heme and can be compared with a known standard using a microscope. A positive microcrystal assay strongly indicates the presence of blood. However, confirmatory assays are usually not as sensitive as presumptive assays. Additionally, these assays cannot distinguish between human and animal blood.

12.3.1.1 Hemochromagen Crystal Assay

Hemochromagens are heme derivatives in which the ferrous iron of the heme forms two bonds with nitrogenous bases (Figure 12.4). The method for forming hemochromagen crystals was documented in 1864. Since then, various modifications have been reported. The *Takayama crystal assay*, published in 1912, has been the method preferred by many forensic laboratories. A bloodstain is treated with pyridine and glucose (a reducing sugar that is capable of reducing ferric ion) under alkaline conditions to form crystals of pyridine ferroprotoporphyrin (Figure 12.18).

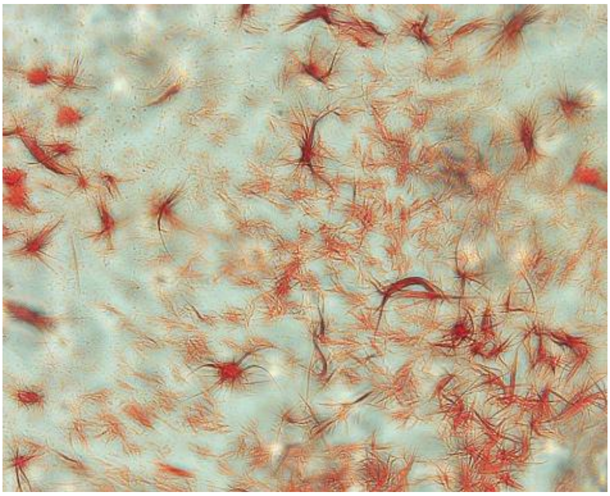

Figure 12.18 Microcrystal assays using the Takayama method. (© Richard C. Li.)

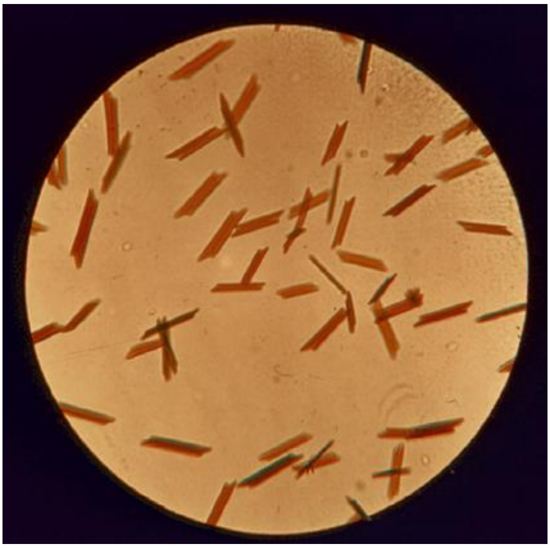

Figure 12.19 Microcrystal assays using the Teichmann method. (From James, S.H., Nordby, J.J., and Bell, S., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 4th edn., CRC Press, Boca Raton, 2014. With permission.)

12.3.1.2 Hematin Crystal Assay

This assay is also known as the *Teichmann crystal assay*. In 1853, Teichmann documented a method of forming crystals of blood specimens. When blood specimens are treated with glacial acetic acid and salts, and subsequently heated, hematin chloride (ferriprotoporphyrin chloride), a prismatic brown-colored crystal, is formed (Figure 12.19). Hematin (Figure 12.4) is a heme derivative; its iron is in the *ferric* (Fe^{+3}) state. This hematin assay has a similar sensitivity and specificity as hemochromagen assays. The hematin assay has the advantage of being more reliable than hemochromagen assays for aged blood samples.

12.3.2 Other Assays

Additional techniques may be used to confirm the presence of hemoglobin. For example, *chromatographic* and *electrophoretic methods* can identify hemoglobin by its mobility characteristics. *Spectrophotometric methods* for identifying hemoglobin are based on measurements of the

Table 12.1 Application of RT-PCR Assay for Blood Identification

Gene Symbol	Gene Product	Description	Further Reading
HBA1	Hemoglobin α1	Hemoglobin α1 chain (abundant in erythrocytes)	Waye and Chui (2001)
PBGD[a]	Porphobilinogen deaminase	Erythrocyte-specific isoenzyme of heme biosynthesis pathway	Gubin and Miller (2001)
SPTB	β-Spectrin	Subunit of major protein of erythrocyte membrane skeleton	Amin et al. (1993)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

[a] Also known as hydroxymethylbilane synthase (HMBS).

characteristic light spectra, with peak absorbance at 400–425 nm, absorbed by hemoglobin and its derivatives. Finally, *immunological methods* utilize antihuman hemoglobin antibodies. This antibody can be used to detect human hemoglobin and thus indicate the presence of human blood (see Chapter 13).

Recently, *RNA-based assays* have been developed to identify blood. These assays are based on the fact that certain genes are specifically expressed in certain cell types (Chapter 11). Thus, the techniques used in the identification of blood are based on the detection of specific types of messenger RNA (mRNA) that are expressed exclusively in erythrocytes. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect the gene expression levels of mRNAs for blood identification. Table 12.1 summarizes the tissue-specific genes utilized for blood identification. Compared with conventional assays that are used for blood identification, the RNA-based assays have higher specificity and are amenable to automation. However, one limitation is that RNA is unstable due to degradation by endogenous and environmentally born ribonucleases.

Bibliography

Amin, K.M., et al., The exon-intron organization of the human erythroid beta-spectrin gene. *Genomics*, 1993, **18**(1): 118–125.

Andersen, J. and S. Bramble, The effects of fingermark enhancement light sources on subsequent PCR-STR DNA analysis of fresh bloodstains. *J Forensic Sci*, 1997, **42**(2): 303–306.

Anderson, S.E., G.R. Hobbs, and C.P. Bishop, Multivariate analysis for estimating the age of a bloodstain. *J Forensic Sci*, 2011, **56**(1): 186–193.

Anderson, S., et al., A method for determining the age of a bloodstain. *Forensic Sci Int*, 2005, **148**(1): 37–45.

Andrasko, J., The estimation of age of bloodstains by HPLC analysis. *J Forensic Sci*, 1997, **42**(4): 601–607.

Anslinger, K., et al., Ninhydrin treatment as a screening method for the suitability of swabs taken from contact stains for DNA analysis. *Int J Legal Med*, 2004, **118**(2): 122–124.

Arany, S. and S. Ohtani, Age estimation of bloodstains: A preliminary report based on aspartic acid racemization rate. *Forensic Sci Int*, 2011, **212**(1–3): e36–e39.

Barnett, P.D., et al., Discussion of "Effects of presumptive test reagents on the ability to obtain restriction fragment length polymorphism (RFLP) patterns from human blood and semen stains". *J Forensic Sci*, 1992, **37**(2): 369–370.

Bauer, M., S. Polzin, and D. Patzelt, Quantification of RNA degradation by semi-quantitative duplex and competitive RT-PCR: A possible indicator of the age of bloodstains? *Forensic Sci Int*, 2003, **138**(1–3): 94–103.

Bergervoet, P.W., et al., Application of the forensic Luminol for blood in infection control. *J Hosp Infect*, 2008, **68**: 329–333.

Botoniic-Sehic, E., et al., Forensic application of near-infrared spectroscopy: Aging of bloodstains. *Spectroscopy* (Santa Monica), 2009, **24**(2): 42–48.

Boyd, S., et al., Highly sensitive detection of blood by surface enhanced Raman scattering. *J Forensic Sci*, 2013, **58**(3): 753–756.

Bremmer, R.H., et al., Remote spectroscopic identification of bloodstains. *J Forensic Sci*, 2011, **56**(6): 1471–1475.

Bremmer, R.H., et al., Forensic quest for age determination of bloodstains. *Forensic Sci Int*, 2012, **216**(1–3): 1–11.

Budowle, B., et al., The presumptive reagent fluorescein for detection of dilute bloodstains and subsequent STR typing of recovered DNA. *J Forensic Sci*, 2000, **45**(5): 1090–1092.

Caldwell, J.P., W. Henderson, and N.D. Kim, ABTS: A safe alternative to DAB for the enhancement of blood fingerprints. *J Forensic Sci*, 2000, **45**(4): 785–794.

Cheeseman, R., Direct sensitivity comparison to the fluorescein and luminol bloodstain enhancement techniques. *J Forensic Ident*, 1999, **49**(3): 8.

Cheeseman, R. and L.A. DiMeo, Fluorescein as a field-worthy latent bloodstain detection system. *J Forensic Ident*, 1995, **45**(6): 1.

Cheeseman, R. and R. Tomboc, Fluorescein technique performance study on bloody foot trails. *J Forensic Ident*, 2001, **51**(1): 12.

Courts, C. and B. Madea, Specific micro-RNA signatures for the detection of saliva and blood in forensic body-fluid identification. *J Forensic Sci*, 2011, **56**(6): 1464–1470.

Cox, M., Effect of fabric washing on the presumptive identification of bloodstains. *J Forensic Sci*, 1990, **35**(6): 1335–1341.

Cox, M., A study of the sensitivity and specificity of four presumptive tests for blood. *J Forensic Sci*, 1991, **36**(5): 1503–1511.

Creamer, J.I., et al., A comprehensive experimental study of industrial, domestic and environmental interferences with the forensic luminol test for blood. *Luminescence*, 2003, **18**(4): 193–198.

Creamer, J.I., et al., Attempted cleaning of bloodstains and its effect on the forensic luminol test. *Luminescence*, 2005, **20**(6): 411–413.

DeForest, P., R. Gaensslen, and H.C. Lee, *Forensic Science: An Introduction to Criminalistics*, 1983. New York: McGraw-Hill Book Company.

Della Manna, A. and S. Montpetit, A novel approach to obtaining reliable PCR results from luminol treated bloodstains. *J Forensic Sci*, 2000, **45**(4): 886–890.

Dixon, T.R., et al., A scanning electron microscope study of dried blood. *J Forensic Sci*, 1976, **21**(4): 797–803.

Dorward, D.W., Detection and quantitation of heme-containing proteins by chemiluminescence. *Anal Biochem*, 1993, **209**(2): 219–223.

Edelman, G., et al., Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy. *Forensic Sci Int*, 2012, **220**(1–3): 239–244.

Farrugia, K.J., et al., Chemical enhancement of footwear impressions in blood on fabric—Part 3: Amino acid staining. *Sci Justice*, 2013, **53**(1): 8–13.

Finnis, J., J. Lewis, and A. Davidson, Comparison of methods for visualizing blood on dark surfaces. *Sci Justice*, 2013, **53**(2): 178–186.

Frascione, N., V. Pinto, and B. Daniel, Development of a biosensor for human blood: New routes to body fluid identification. *Anal Bioanal Chem*, 2012, **404**(1): 23–28.

Fregeau, C.J., O. Germain, and R.M. Fourney, Fingerprint enhancement revisited and the effects of blood enhancement chemicals on subsequent profiler plus fluorescent short tandem repeat DNA analysis of fresh and aged bloody fingerprints. *J Forensic Sci*, 2000, **45**(2): 354–380.

Fujita, Y., et al., Estimation of the age of human bloodstains by electron paramagnetic resonance spectroscopy: Long-term controlled experiment on the effects of environmental factors. *Forensic Sci Int*, 2005, **152**(1): 39–43.

Gaensslen, R.E., *Sourcebook in Forensic Serology, Immunology, and Biochemistry*, 1983. Washington, DC: US Government Printing Office.

Garner, D.D., et al., An evaluation of tetramethylbenzidine as a presumptive test for blood. *J Forensic Sci*, 1976, **21**(4): 816–821.

Gimeno, F.E., Fill flash color photography to photograph luminol blood stain patterns. *J Forensic Ident*, 1989, **39**(5): 305–306.

Gimeno, F.E. and G.E. Rini, Fill flash luminescence to photograph luminol blood stain patterns. *J Forensic Ident*, 1989, **39**(3): 1.

Grafit, A., A. Cohen, and Y. Cohen, Estimation of original volume of dry bloodstains using spectrophotometric method. *J Forensic Ident*, 2012, **62**(4): 305–314.

Greenfield, A., M. Sloan, and R. Spaulding, Identification of blood and body fluids, in S. James, J. Nordby and S. Bell (eds), *Forensic Science: An Introduction to Scientific and Investigative Techniques*, pp. 205–228, 2014. Boca Raton, FL: CRC Press.

Grispino, R.R.J., The effect of luminol on the serological analysis of dried human bloodstains. *Crime Lab Digest*, 1990, **17**(1): 13–23.

Gross, A.M., K.A. Harris, and G.L. Kaldun, The effect of luminol on presumptive tests and DNA analysis using the polymerase chain reaction. *J Forensic Sci*, 1999, **44**(4): 837–840.

Gubin, A.N. and J.L. Miller, Human erythroid porphobilinogen deaminase exists in 2 splice variants. *Blood*, 2001, **97**(3): 815–817.

Guo, K., S. Achilefu, and M.Y. Berezin, Dating bloodstains with fluorescence lifetime measurements. *Chemistry*, 2012, **18**(5): 1303–1305.

Haas, C., et al., MRNA profiling for the identification of blood—Results of a collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2011, **5**(1): 21–26.

Haas, C., et al., Selection of highly specific and sensitive mRNA biomarkers for the identification of blood. *Forensic Sci Int Genet*, 2011, **5**(5): 449–458.

Haas, C., et al., RNA/DNA co-analysis from blood stains—Results of a second collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2012, **6**(1): 70–80.

Hanson, E., A. Albornoz, and J. Ballantyne, Validation of the hemoglobin (Hb) hypsochromic shift assay for determination of the time since deposition (TSD) of dried bloodstains. *Forensic Sci Int Genet Suppl Series*, 2011, **3**(1): e307–e308.

Hanson, E.K. and J. Ballantyne, A blue spectral shift of the hemoglobin soret band correlates with the age (time since deposition) of dried bloodstains. *PloS ONE*, 2010, **5**(9): e12830.

Hatch, A.L., A modified reagent for the confirmation of blood. *J Forensic Sci*, 1993, **38**(6): 1502–1506.

Higaki, R.S. and W.M.S. Philp, A study for the sensitivity, stability, and specificity of phenolphthalein as an indicator test for blood. *Can Soc Forensic Sci*, 1976, **9**(3): 97–102.

Hochmeister, M.N., B. Budowle, and F.S. Baechtel, Effects of presumptive test reagents on the ability to obtain restriction fragment length polymorphism (RFLP) patterns from human blood and semen stains. *J Forensic Sci*, 1991, **36**(3): 656–661.

Inoue, H., et al., Identification of fetal hemoglobin and simultaneous estimation of bloodstain age by high-performance liquid chromatography. *Int J Legal Med*, 1991, **104**(3): 127–131.

Inoue, H., et al., A new marker for estimation of bloodstain age by high performance liquid chromatography. *Forensic Sci Int*, 1992, **57**(1): 17–27.

Introna, F. Jr., G. Di Vella, and C.P. Campobasso, Determination of postmortem interval from old skeletal remains by image analysis of luminol test results. *J Forensic Sci*, 1999, **44**(3): 535–538.

James, S.H. and Nordby, J.J., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 2nd edn., 2005. Boca Raton, FL: CRC Press.

Juusola, J. and Ballantyne, J., Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**: 1–12.

Kashyap, V.K., A simple immunosorbent assay for detection of human blood. *J Immunoassay*, 1989, **10**(4): 315–324.

Kent, E.J., D.A. Elliot, and G.M. Miskelly, Inhibition of bleach-induced luminol chemiluminescence. *J Forensic Sci*, 2003, **48**(1): 64–67.

Kohlmeier, F. and P.M. Schneider, Successful mRNA profiling of 23 years old blood stains. *Forensic Sci Int Genet*, 2012, **6**(2): 274–276.

Laux, D.L., Effects of luminol on the subsequent analysis of bloodstains. *J Forensic Sci*, 1991, **36**(5): 1512–1520.

Laux, D.L., The detection of blood using luminol, in S.H. James, P.E. Kish, and T.P. Sutton (eds), *Principles of Bloodstain Pattern Analysis: Theory and Practice*, pp. 369–390, 2005. Boca Raton, FL: CRC Press.

Lee, H.C., Identification and grouping of bloodstains, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 267–337, 1982. Englewood Cliffs, NJ: Prentice Hall.

Lee, H.C., T. Palmbach, and M.T. Miller, *Henry Lee's Crime Scene Handbook*, 2001. San Diego: Academic Press.

Lytle, L.T. and D.G. Hedgecock, Chemiluminescence in the visualization of forensic bloodstains. *J Forensic Sci*, 1978, **23**(3): 550–562.

Matsuoka, T., T. Taguchi, and J. Okuda, Estimation of bloodstain age by rapid determinations of oxyhemoglobin by use of oxygen electrode and total hemoglobin. *Biol Pharm Bull*, 1995, **18**(8): 1031–1035.

Miyaishi, S., et al., Discrimination between postmortem and antemortem blood by a dot-ELISA for human myoglobin. *Nihon Hoigaku Zasshi*, 1994, **48**(6): 433–438.

Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., Messenger RNA profiling: a novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**: 181–186.

Omelia, E.J., M.L. Uchimoto, and G. Williams, Quantitative PCR analysis of blood- and saliva-specific microRNA markers following solid-phase DNA extraction. *Anal Biochem*, 2013, **435**(2): 120–122.

Ponce, A.C. and F.A. Verdu Pascual, Critical revision of presumptive tests for bloodstains. *Forensic Sci Commun*, 1999, **1**(2): 1–15.

Quickenden, T.I. and J.I. Creamer, A study of common interferences with the forensic luminol test for blood. *Luminescence*, 2001, **16**(4): 295–298.

Quickenden, T.I. and P.D. Cooper, Increasing the specificity of the forensic luminol test for blood. *Luminescence*, 2001, **16**(3): 251–253.

Quickenden, T.I., C.P. Ennis, and J.I. Creamer, The forensic use of luminol chemiluminescence to detect traces of blood inside motor vehicles. *Luminescence*, 2004, **19**(5): 271–277.

Richards, A. and R. Leintz, Forensic reflected ultraviolet imaging. *J Forensic Ident*, 2013, **63**(1): 46–69.

Santos, V.R.D., W.X. Paula, and E. Kalapothakis, Influence of the luminol chemiluminescence reaction on the confirmatory tests for the detection and characterization of bloodstains in forensic analysis. *Forensic Sci Int Genet Suppl Series*, 2009, **2**(1): 196–197.

Shaler, R.C., Modern forensic biology, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 525–614, 2002. Upper Saddle River, NJ: Pearson Education.

Shipp, E., et al., Effects of argon laser light, alternate source light, and cyanoacrylate fuming on DNA typing of human bloodstains. *J Forensic Sci*, 1993, **38**(1): 184–191.

Spalding, R., The identification and characterization of blood and bloodstains, in S. James and J. Nordby (eds), *Forensic Science: An Introduction to Scientific and Investigative Techniques*, pp. 237–260, 2005. Boca Raton, FL: CRC Press.

Stoilovic, M., Detection of semen and blood stains using polilight as a light source. *Forensic Sci Int*, 1991, **51**(2): 289–296.

Strasser, S., et al., Age determination of blood spots in forensic medicine by force spectroscopy. *Forensic Sci Int*, 2007, **170**(1): 8–14.

Sutton, T.P., Presumptive blood testing, in S.H. James (ed.), *Scientific and Legal Applications of Bloodstain Pattern Interpretation*, pp. 48–70, 1999. Boca Raton, FL: CRC Press.

Suwa, N., et al., Human blood identification using the genome profiling method. *Leg Med*, 2012, **14**(3): 121–125.

Thomas, P. and K. Farrugia, An investigation into the enhancement of fingermarks in blood on paper with genipin and lawsone. *Sci Justice*, 2013, **53**(3): 315–320.

Tobe, S.S., N. Watson, and N.N. Daeid, Evaluation of six presumptive tests for blood, their specificity, sensitivity, and effect on high molecular-weight DNA. *J Forensic Sci*, 2007, **52**(1): 102–109.

Tsutsumi, H. and Y. Katsumata, Forensic study on stains of blood and saliva in a chimpanzee bite case. *Forensic Sci Int*, 1993, **61**(2–3): 101–110.

Tsutsumi, A., Y. Yamamoto, and H. Ishizu, Determination of the age of bloodstains by enzyme activities in blood cells. *Nihon Hoigaku Zasshi*, 1983, **37**(6): 770–776.

Vandenberg, N. and R.A. van Oorschot, The use of Polilight in the detection of seminal fluid, saliva, and bloodstains and comparison with conventional chemical-based screening tests. *J Forensic Sci*, 2006, **51**(2): 361–370.

Virkler, K. and I.K. Lednev, Raman spectroscopic signature of blood and its potential application to forensic body fluid identification. *Anal Bioanal Chem*, 2010, **396**(1): 525–534.

Waye, J.S. and D.H. Chui, The alpha-globin gene cluster: Genetics and disorders. *Clin Invest Med*, 2001, **24**(2): 103–109.

Webb, J.L., J.I. Creamer, and T.I. Quickenden, A comparison of the presumptive luminol test for blood with four non-chemiluminescent forensic techniques. *Luminescence*, 2006, **21**(4): 214–220.

Yanagida, J., M. Hara, and H. Nakamura, Estimation of age of dried bloodstains by a surface absorption spectrophotometric method. *J Saitama Med School*, 1978, **5**(3): 221–225.

Yanagida, J., et al., Colorimetric study on the color changes of ageing bloodstains. *J Saitama Med School*, 1981, **8**(1–2): 15–20.

Zubakov, D., et al., Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples. *Int J Legal Med*, 2008, **122**(2): 135–142.

Zubakov, D., et al., New markers for old stains: Stable mRNA markers for blood and saliva identification from up to 16-year-old stains. *Int J Legal Med*, 2009, **123**(1): 71–74.

Zweidinger, R.A., L.T. Lytle, and C.G. Pitt, Photography of bloodstains visualized by luminol. *J Forensic Sci*, 1973, **18**(4): 296–302.

13

Species Identification

Chapter 12 discussed the principles of the identification of blood. If a stain is identified as blood, the evidence can be tested to determine whether the blood is of human origin. If the bloodstain is nonhuman, further analysis is usually not necessary.

Before forensic DNA techniques were implemented, species identification was largely determined by serological methods. Currently, most forensic laboratories perform DNA quantitation prior to DNA profile analysis. The quantitation method specifically detects higher-primate DNA. The presence of DNA measured by the quantitation assays concurrently identifies a sample as being of human origin (since crimes involving primate blood are extremely rare). Thus, species identification is usually not performed in forensic laboratories.

Nevertheless, species identification assays can be useful for screening to exclude or eliminate nonhuman samples unrelated to an investigation. Thus, it is practical for small laboratories to eliminate unnecessary analyses due to time and budget constraints. Additionally, species identification kits such as immunochromatographic devices allow field testing by crime scene investigators.

In cases involving the killing, trading, and possession of products derived from species that are protected from illegal hunting, it may be necessary to identify the animal species prior to further analysis. Species identification using DNA analysis can be performed using commonly used loci at the mitochondrial cytochrome b gene (*Cytb*), the cytochrome c oxidase I gene (*COI*), and the D-loop region. This type of identification is usually within the scope of wildlife forensic science and is thus not discussed here.

13.1 General Considerations

Most assays for species identification are based on serological techniques, including primary and secondary binding assays. The most common primary binding assays are immunochromatographic assays. The most commonly used secondary binding assays are precipitation-based assays that rely on the binding of an antigen to an antibody, causing the formation of visible precipitation. These precipitation-based assays include ring assays, Ouchterlony assays, and crossed-over immunoelectrophoresis. These assays utilize antihuman and antianimal antibodies to identify human and animal species, respectively.

13.1.1 Types of Antibodies

An antihuman antibody that is used in the identification of human samples can be made by introducing human serum into a host animal, which then produces specific antibodies against the human serum proteins. Antibodies produced from different species of host animals may produce variations in the characteristics of reactions. Since albumin is the most abundant protein in human serum, the antihuman antibody that is produced reacts strongly with human albumin. Albumin is a protein that plays important roles in the maintenance of the vascular circulating fluid and the transportation of various substances such as nutrients, hormones, and metabolic products. Blood is drawn from the host animal and the serum portion is collected. The collected serum is a polyclonal antihuman antiserum containing a mixture of antibodies against various human serum proteins. Likewise, an antibody against animal serum proteins can also be made to identify animal species of interest.

Other antibodies such as antihuman hemoglobin (Hb) antibodies can also be used to identify the human origin of a sample. Hb is an oxygen-transport protein that is found in erythrocytes (Chapter 12). Purified Hb can be used to generate monoclonal and polyclonal antihuman Hb antibodies. Likewise, antibodies recognizing glycophorin A (GPA), a human erythrocyte membrane antigen (see Section 13.2.1.2), can also be produced in a similar manner.

13.1.2 Titration of Antibodies

Recall that the ratio of antigen to antibody is critical for the success of a secondary reaction (Chapter 10). An extreme excess of antigen or antibody concentrations can inhibit secondary reactions. The prozone and postzone phenomena must be considered, and the concentrations of antigen and antibody must be carefully determined for forensic serology assays. For instance, in the prozone situation, a false-negative reaction may occur due to the presence of a high concentration of antibody.

Quality-control procedures can be used to estimate the amount of a specific antibody that is present, often via titration (Figure 13.1). To titrate an antiserum, a series of dilutions are made and each dilution is then tested for activity using precipitation or agglutination methods. The reciprocal of the highest dilution giving a positive reaction is known as the *titer*. This reflects the amount of antibody in the antiserum. Additionally, the polyclonal antiserum is a mixture of antibodies; thus, the reaction of the antiserum may vary from animal to animal (of the same species). Each lot or batch of antiserum must be validated by titration.

13.1.3 Antibody Specificity

In addition to the titer, the specificity of the antihuman antibody must be tested. Most antihuman antibodies usually have cross-reactivity with higher primates. This is not a great

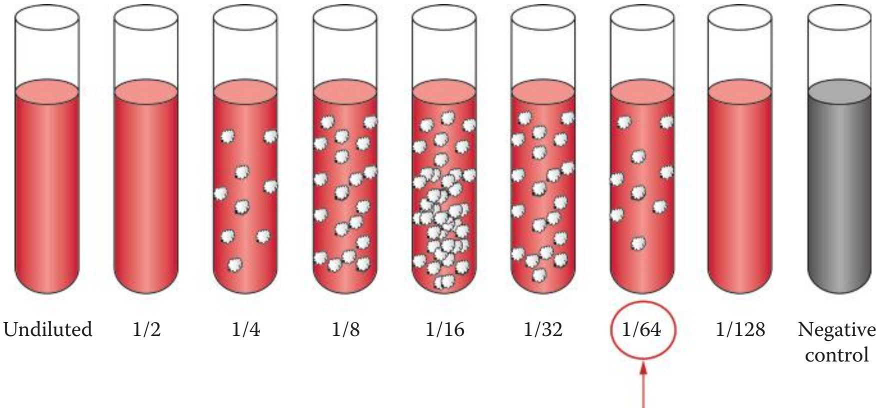

Figure 13.1 Titration of antibodies. Serum is serially diluted and a constant amount of antigen is applied to each tube. The mixture is incubated, allowing agglutination to occur. The reciprocal of the highest dilution giving a positive agglutination reaction is 64 (the titer). (© Richard C. Li.)

concern because crimes involving nonhuman primates are very rare. Nevertheless, the antihuman antibody must not cross-react with other commonly encountered animals. Antisera and positive control samples must be validated for cross-reactivity. Tissue specificity must also be validated. The antiserum against human serum is usually reactive with other human biological fluids such as semen and saliva.

13.1.4 Optimal Conditions for Antigen–Antibody Binding

A number of factors can affect antigen–antibody binding. For example, increasing ionic strength can inhibit the binding of an antigen and an antibody. Stronger inhibition is usually observed for ions with large ionic radii and small radii of hydration. It is believed that the lower degree of hydration permits interactions of ions and the antibody-binding site, leading to inhibition. A proper buffer system must be selected in serological assays to ensure reliable results. The introduction of polymers can facilitate precipitation in secondary binding reactions because the presence of a polymer in a solution decreases the solubility of proteins. Linear hydrophilic polymers with high molecular weights (e.g., polyethylene glycol) are preferred. Additional factors such as temperature and pH can also affect antigen–antibody binding.

13.2 Assays

Samples can be prepared by cutting out a portion of a stain or scraping stains from a surface. A sample is usually extracted with a small volume of saline or buffer. The extracted sample can be tested using the assays described in the following subsections. Controls should be included, for example, by using a known human serum as a positive control and an extraction blank as a negative control.

13.2.1 Immunochromatographic Assays

Immunochromatographic assays are rapid, specific, and sensitive and can be used in both laboratory and field tests for species identification. Two types of assays are discussed, including those based on the detection of human erythrocyte proteins. Chapter 11 discusses the principle of immunochromatographic assays in more detail.

13.2.1.1 Identification of Human Hemoglobin Protein

Commercially produced immunochromatographic kits such as the Hexagon OBTI (Human Gesellschaft für Biochemica und Diagnostica mbH, Wiesbaden) and the ABAcard HemaTrace® (Abacus Diagnostics, California) are available. They utilize the antibody–antigen–antibody sandwich method by using antibodies that recognize human Hb. The ABAcard HemaTrace assay utilizes a labeled monoclonal antihuman Hb antibody contained in a sample well, and a polyclonal antihuman Hb antibody immobilized at a test zone of a nitrocellulose membrane. Additionally, an antiglobulin that recognizes the antibody is immobilized onto a control zone (Figure 13.2).

A sample can be prepared by cutting a small portion (2 mm diameter) of a stain or a swab. Each sample is extracted for 5 min in 2 mL of extraction buffer. A longer extraction time may be used for older stains. The samples are loaded into the sample well, and the antigen in the sample binds to the labeled antibody in the well to form an antigen–antibody complex, which then diffuses across the nitrocellulose membrane. At the test zone, the solid-phase antihuman Hb antibody binds to the antigen–antibody complex to form a labeled antibody–antigen–antibody sandwich.

The ABAcard HemaTrace® uses a pink dye that is visualized in a positive result as a pink horizontal line at the test zone (Figure 13.3). In the control zone, unbound labeled antihuman

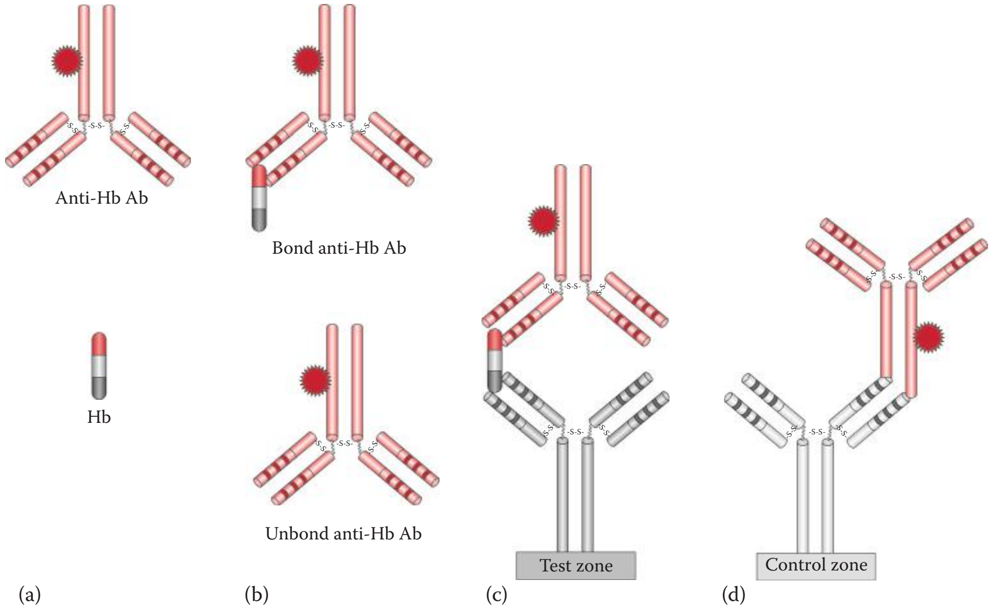

Figure 13.2 Immunochromatographic assays for the identification of Hb in human blood. (a) In a sample well, Hb in a blood sample is mixed with a labeled anti-Hb Ab. (b) The Hb binds to the labeled anti-Hb Ab to form a labeled Ab–Hb complex. The complex diffuses toward the test zone. (c) At the test zone, the labeled Ab–Hb complex binds to an immobilized anti-Hb Ab to form a labeled Ab–Hb–Ab sandwich. (d) At the control zone, the labeled anti-Hb Ab binds to an immobilized antiglobulin and is captured. Ab and Hb represent the antibody and hemoglobin, respectively. (© Richard C. Li.)

Hb antibody binds to the solid-phase antiglobulin. This antibody–antiglobulin complex at the control zone also produces a pink horizontal line. The test is considered valid only if the line in the control zone is observed. The presence of human Hb results in a pink line at both the test and control zones. The absence of human Hb results in a pink line in the control zone only. A positive result can appear in less than a minute.

Validation studies have revealed that the sensitivity of the ABAcard HemaTrace® can be as low as 0.07 μg/mL of Hb. The normal blood Hb concentration is 14–18 and 12–16 g/dL among males and females, respectively. This assay is more sensitive than the Kastle–Meyer assay (Chapter 12). Additionally, the assay is responsive to aged stains and degraded materials. Specificity studies have shown that it is specific for blood of higher primates, including humans. However, it is also responsive to seminal stains, and oral, vaginal, anal, and rectal swabs. It is believed that these biological fluids contain very low amounts of Hb, which can still be detected by highly sensitive assays. However, if the concentration of blood is too high, a false negative can result due to the high-dose hook effect described in Chapter 11.

13.2.1.2 Identification of Human Glycophorin A Protein

Commercially produced immunochromatographic kits such as RSID™-Blood (Independent Forensics, Hillside, IL) use antibodies that recognize human GPA (Figure 13.4). A labeled monoclonal antihuman GPA antibody is contained in a sample well, and a second monoclonal antihuman GPA antibody, to a different epitope of GPA, is immobilized onto a test zone of the membrane. An antiglobulin that recognizes the antibody is immobilized onto a control zone (Figure 13.5).

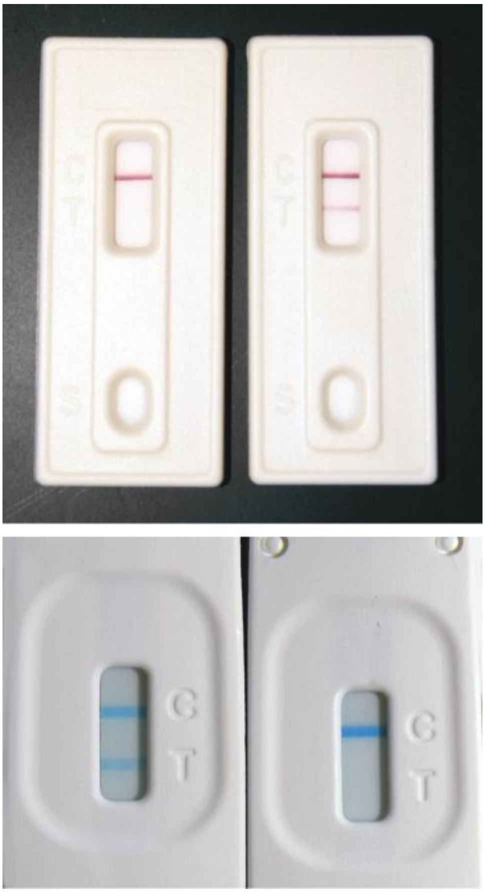

Figure 13.3 Human blood identification using immunochromatic devices. Top: In an assay using an ABAcard HemaTrace device, the negative (left) and positive (right) results are shown. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. The sample well is labeled "S." (© Richard C. Li.) Bottom: The positive (left) and negative (right) results are shown using a Hexagon-OBTI device. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. (From Ramsthaler, F., et al., Postmortem interval of skeletal remains through the detection of intraosseal hemin traces. A comparison of UV-fluorescence, luminol, Hexagon-OBTI(R), and Combur(R) tests. *Forensic Sci Int*, 209, 59–63, 2011. With permission.)

The sample can be collected by cutting out a small portion of a stain or a swab. The sample is then extracted overnight in an extraction buffer. The extract is removed and mixed with a running buffer. The assay is carried out by loading the extracted sample into the sample well. Again, the presence of GPA results in a pink line at both the test zone and the control zone, while the absence of GPA results in a pink line in the control zone only. The test is considered valid only if the line in the control zone is observed. A result can be read after 10 min.

Validation studies revealed that the sensitivity of the RSID kit can be as low as 100 nL of human blood. Species specificity studies showed no cross-reactivity with various animal species, including nonhuman primates. Biological fluid specificity studies revealed that the kit is not responsive to other human biological fluids such as semen, saliva, urine, milk, and amniotic and vaginal fluid. No high-dose hook effects were observed in samples containing up to 5 μL of blood.

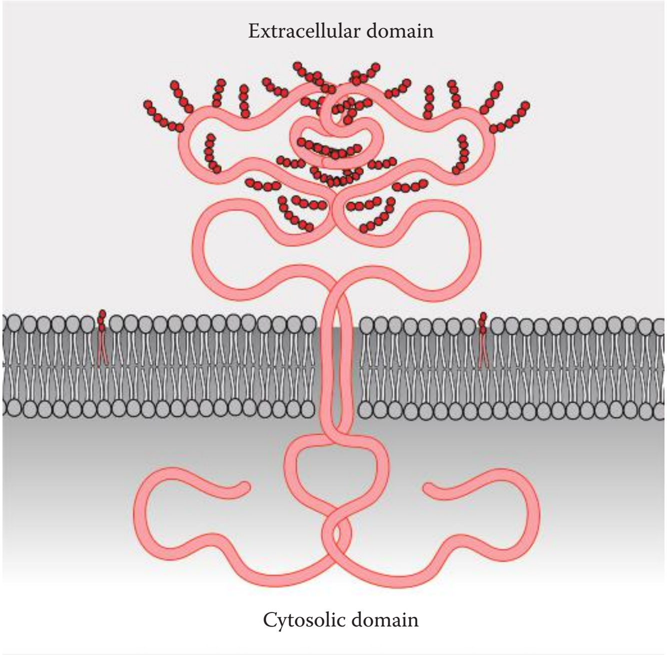
Extracellular domain
Cytosolic domain

Figure 13.4 Diagram of the structure of glycophorin (GPA) protein. GPA is a transmembrane protein on the human erythrocyte membrane. A GPA dimer is shown. The extracellular domain of the GPA is glycosylated with carbohydrate side chains (red). Various GPA epitopes are antigenic determinants of several blood group systems. (© Richard C. Li.)

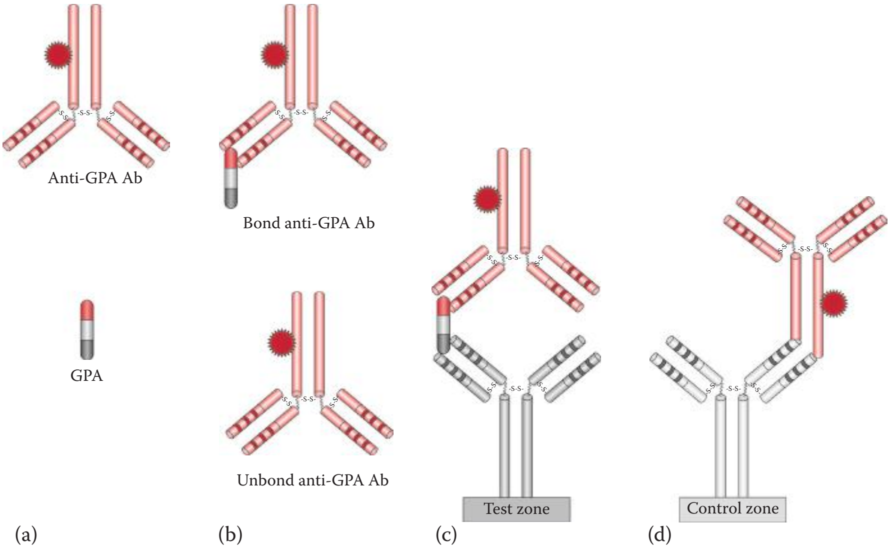
Anti-GPA Ab
GPA
Bond anti-GPA Ab
Unbond anti-GPA Ab
Test zone
Control zone
-S-S-
(a)
(b)
(c)
(d)

Figure 13.5 Immunochromatographic assays for the identification of GPA in human blood. (a) In a sample well, GPA in a blood sample is mixed with a labeled anti-GPA Ab. (b) The GPA binds to the labeled anti-GPA Ab to form a labeled Ab–GPA complex. (c) At the test zone, the labeled Ab–GPA complex binds to an immobilized anti-GPA Ab to form a labeled Ab–GPA–Ab sandwich. (d) At the control zone, the labeled anti-GPA Ab binds to an immobilized antiglobulin and is captured. Ab and GPA represent the antibody and GPA protein, respectively. (© Richard C. Li.)

13.2.2 Double Immunodiffusion Assays

13.2.2.1 Ring Assay

Chapter 11 discussed the basic principle of ring assay in detail. In this double immunodiffusion assay, an antihuman antibody reagent is placed at the bottom of a test tube and a bloodstain extract is placed on top of the bottom layer, as illustrated in Figure 13.6. The procedure is described in Box 13.1. In a positive reaction, a white precipitate between the two layers observed after several minutes indicates that a sample is of human origin. If the bloodstain extract is not human, no precipitation should appear.

13.2.2.2 Ouchterlony Assay

The basic principle of this double immunodiffusion assay has also been discussed in Chapter 11. The procedure of the Ouchterlony assay is described in Box 13.2. In a positive reaction, a line of precipitate will form between each antigen well and antibody well. This assay can also determine the similarity of the antigens (Figure 13.7). During the diffusion process, different antigen–antibody complexes migrate at different rates. Consequently, a separate line of precipitate will appear in the gel for each antigen–antibody complex. In an assay in which two antigens are loaded in adjacent wells and an antibody in the third well, the following results can be observed:

- If the two antigens are identical, the two lines will become fused. This phenomenon is referred to as *identity*.
- If the two antigens are totally unrelated, the lines will cross each other but not fuse; this is known as *nonidentity*.
- If the two antigens are related (share a common epitope) but are not identical, the lines will merge with spur formation. The spurs are continuations of the line formed by the antigen due to its unique epitope. This phenomenon is known as *partial identity*.

Thus, in a species test, a positive result is noted when the precipitate lines for the positive controls and the samples fuse. No spur formation should be observed.

13.2.3 Crossed-Over Electrophoresis

This method is a combination of immunodiffusion and electrophoresis (also see Chapter 11). The procedure for the assay is described in Box 13.3. With this technique, a sharp precipitate band is visualized in a positive reaction (Figure 13.8). However, false-negative results can occur due to the postzone phenomenon, in which excess antigen may inhibit precipitation. In this

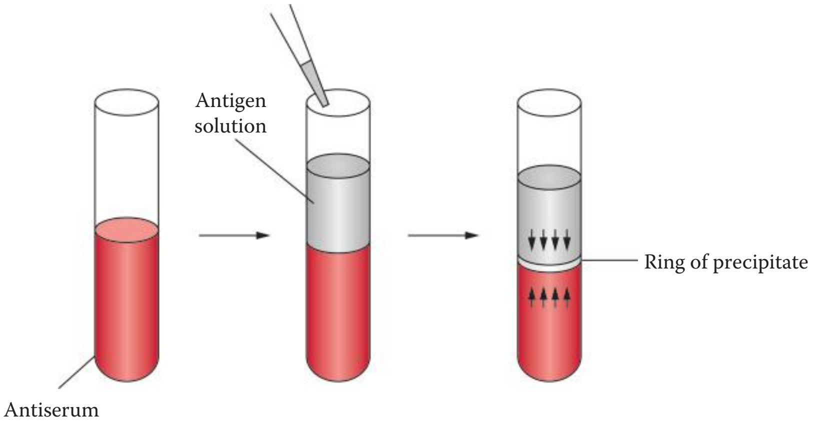

Figure 13.6 Ring assay. The antigen solution is carefully applied over the antiserum solution. Incubation may be necessary in order to form a ring of precipitate. (© Richard C. Li.)

BOX 13.1 RING ASSAY PROCEDURE

Sample preparation and extraction

- Extract a portion of a stain with saline at 4°C overnight.

Controls

- Include a positive control (known human serum sample) and a negative control (extraction blank).

Loading of antibody and samples

- Spin the antihuman antibody in a microfuge and transfer the supernatant into test tubes or capillary tubes (depending on the volume of the stain and the antiserum extracted).
- Place the sample carefully over the top of the antiserum solution, which is usually denser than the sample.

Immunodiffusion reaction

- Carry out the reaction at room temperature.
- In a positive reaction, white precipitate between the two layers can be observed after several minutes. This indicates that the sample is of human origin. No precipitate is formed if a bloodstain extract is from a nonhuman origin.

BOX 13.2 OUCHTERLONY ASSAY PROCEDURE

Sample preparation and extraction

- Cut out a small portion (approximately 5×5 mm) of a stain or a portion of a swab.
- Extract at room temperature in 100 μL of water for 30 min. The extract can be diluted if necessary. Alternatively, a very small piece of the stain or swab can be inserted directly into the well.

Controls

- Positive (known serum)
- Negative (extraction blank)
- Substrate controls (extraction of substrate from unstained area) if applicable

Agarose gel preparation

Heat a suspension of agarose (4%) until liquefied. Cool the solution in a water bath at 55°C. Pour the agarose onto a piece of glass slide and let the gel solidify to a thickness of about 2–3 mm. Alternatively, a polyester support film such as GelBond (Cambrex, New Zealand) can be used as a gel support. The agarose should be poured onto the hydrophilic side of a piece of GelBond film (6×9 cm). Punch wells consisting of a central well surrounded by four wells using a template.

Loading antibodies and samples

- Apply antihuman antibody to the central well. Apply the positive control to one of the surrounding wells.
- Apply the sample(s) in question next to a positive control.
- Apply negative and substrate controls to the remaining wells; only one negative control is needed per gel.

Immunodiffusion reaction

Incubate the plate overnight in a moisture chamber at 37°C.

Staining

Soak the gel overnight in saline solution and then soak it in deionized water for 10 min. Repeat once. Dry the gel between paper towels with a weight on top for 30 min. Dry in an oven for 30 min. Stain the gel with Coomassie blue. Stained precipitate bands appear blue.

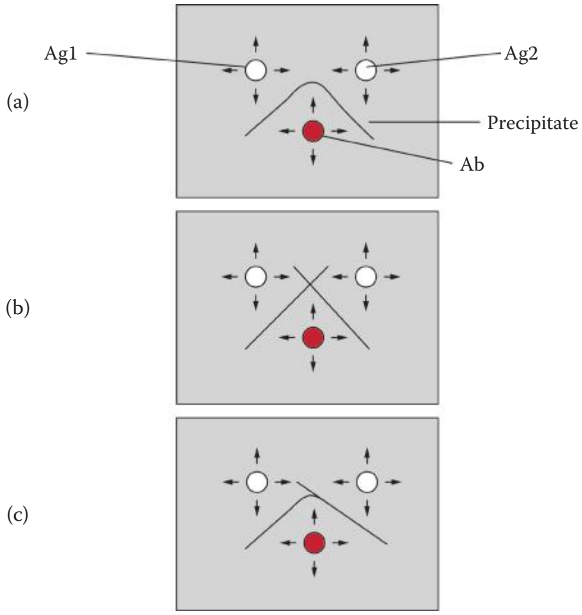

Figure 13.7 Results of Ouchterlony assay. Ab indicates an antibody. Ag1 and Ag2 are the antigen samples in question. (a) Identity. The two antigens are identical (fused line). (b) Nonidentity. The two antigens are unrelated (spur). (c) Partial identity. The two antigens are related but not identical (fused line with spur). (© Richard C. Li.)

situation, the sample can be diluted and the assay can be repeated. False-negative results can also occur due to simple mistakes made during electrophoresis:

- Electrophoresis is carried out in the opposite direction, which results in samples running off the gel.
- Electrophoresis is carried out using an incorrect buffer system, affecting antigen–antibody binding. The amount of current applied during the electrophoresis is too strong and generates heat and denatures proteins.

BOX 13.3 CROSSED-OVER ELECTROPHORESIS PROCEDURE

Sample preparation and extraction

- Cut out a small portion (~5 × 5 mm) of a stain or a portion of a swab.
- Extract at room temperature in 100 μL of water for 30 min. The extract can be diluted if necessary. Alternatively, a very small piece of the stain or swab can be inserted directly into the well.

Controls

- Positive (known serum)
- Negative (extraction blank)
- Substrate (extraction of substrate from unstained area) if applicable

Agarose gel preparation

Heat a suspension of agarose (4%) until liquefied. Cool the solution in a water bath at 55°C. Pour the agarose onto a piece of glass slide and let it solidify. Alternatively, a polyester support film such as GelBond can be used as a gel support. The agarose should be poured onto the hydrophilic side of a piece of GelBond (6 × 9 cm). Punch small wells (about 1–2 mm) in rows using a template.

Loading antibodies and samples

- Apply antihuman antibody in one row of wells.
- Apply samples in the other row of wells. Apply the positive, negative, and substrate controls.

Electrophoresis

Submerge the agarose gel in an electrophoresis tank in proper orientation. The wells containing antihuman antibody should be closest to the anode (positive electrode) and the wells containing samples should be closest to the cathode (negative electrode). During electrophoresis, the antibody in the antiserum should migrate toward the cathode while the antigen migrates toward the anode. Electrophoresis is carried out at 10 V/cm for 20 min.

Staining

Soak the gel overnight in a saline solution and then soak it in deionized water for 10 min. Repeat once. Dry the gel between paper towels with a weight on top for 30 min. Dry in an oven for 30 min. Stain the gel with Coomassie blue. Stained precipitate bands appear blue.

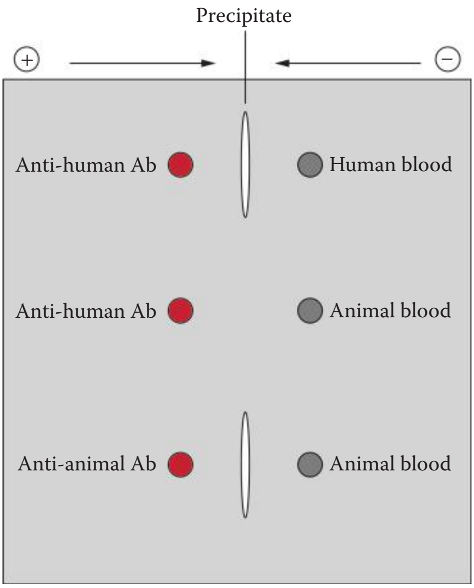

Figure 13.8 Results of crossed-over electrophoresis. A precipitate line is formed between a human blood sample and an antihuman antibody. No precipitate line is formed when the antihuman antibody is tested for an animal blood sample. A precipitate line is formed between the animal blood sample and the antianimal antibody. (© Richard C. Li.)

Bibliography

Allison, A.C. and J.A. Morton, Species specificity in the inhibition of antiglobulin sera: A technique for the identification of human and animal bloods. *J Clin Pathol*, 1953, **6**(4): 314–319.

Berger, B., et al., Validation of two canine STR multiplex-assays following the ISFG recommendations for non-human DNA analysis. *Forensic Sci Int Genet*, 2014, **8**(1): 90–100.

Bhatia, R.Y., Specificity of some plant lectins in the differentiation of animal blood. *Forensic Sci*, 1974, **4**(1): 47–52.

Coomber, N., et al., Validation of a short tandem repeat multiplex typing system for genetic individualization of domestic cat samples. *Croat Med J*, 2007, **48**(4): 547–555.

Dawnay, N., et al., Validation of the barcoding gene *COI* for use in forensic genetic species identification. *Forensic Sci Int*, 2007, **173**(1): 1–6.

DeForest, P., R. Gaensslen, and H.C. Lee, *Forensic Science: An Introduction to Criminalistics*, 1983. New York: McGraw-Hill.

Dorrill, M. and P.H. Whitehead, The species identification of very old human blood-stains. *Forensic Sci Int*, 1979, **13**(2): 111–116.

El-Sayed, Y.S., et al., Using species-specific repeat and PCR-RFLP in typing of DNA derived from blood of human and animal species. *Forensic Sci Med Pathol*, 2010, **6**(3): 158–164.

Grobbelaar, B.G., D. Skinner, and H.N. van de Gertenbach, The anti-human globulin inhibition test. In the identification of human blood stains. *J Forensic Med*, 1970, **17**(3): 103–111.

Hochmeister, M.N., et al., Validation studies of an immunochromatographic 1-step test for the forensic identification of human blood. *J Forensic Sci*, 1999, **44**(3): 597–602.

Imaizumi, K., et al., Development of species identification tests targeting the 16S ribosomal RNA coding region in mitochondrial DNA. *Int J Legal Med*, 2007, **121**(3): 184–191.

Johnston, S., J. Newman, and R. Frappier, Validation study of the Abacus Diagnostics ABAcard® HemaTrace® membrane test for the forensic identification of human blood. *Can Soc Forensic Sci*, 2003, **36**(3): 173–183.

Juusola, J. and J. Ballantyne, Messenger RNA profiling: A prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int*, 2003, **135**(2): 85–96.

Kanthaswamy, S., et al., Quantitative real-time PCR (qPCR) assay for human-dog-cat species identification and nuclear DNA quantification. *Forensic Sci Int Genet*, 2012, **6**(2): 290–295.

Karlsson, A.O. and G. Holmlund, Identification of mammal species using species-specific DNA pyrosequencing. *Forensic Sci Int*, 2007, **173**(1): 16–20.

Kun, T., et al., Developmental validation of Mini-DogFiler for degraded canine DNA. *Forensic Sci Int Genet*, 2013, **7**(1): 151–158.

Lawton, M.E. and J.G. Sutton, Species identification of deer blood by isoelectric focusing. *J Forensic Sci Soc*, 1982, **22**(4): 361–366.

Lee, H.C., Identification and grouping of bloodstains, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 267–337, 1982. Englewood Cliffs, NJ: Prentice Hall.

Lee, H.C. and P.R. De Forest, A precipitin-inhibition test on denatured bloodstains for the determination of human origin. *J Forensic Sci*, 1976, **21**(4): 804–810.

Lowenstein, J.M., et al., Identification of animal species by protein radioimmunoassay of bone fragments and bloodstained stone tools. *Forensic Sci Int*, 2006, **159**(2–3): 182–188.

Nakaki, S., et al., Study of animal species (human, dog and cat) identification using a multiplex single-base primer extension reaction in the cytochrome b gene. *Forensic Sci Int*, 2007, **173**(2–3): 97–102.

Ogden, R., Forensic science, genetics and wildlife biology: Getting the right mix for a wildlife DNA forensics lab. *Forensic Sci Med Pathol*, 2010, **6**(3): 172–179.

Parson, W., et al., Species identification by means of the cytochrome b gene. *Int J Legal Med*, 2000, **114**(1–2): 23–28.

Ramsthaler, F., et al., Postmortem interval of skeletal remains through the detection of intraosseal hemin traces. A comparison of UV-fluorescence, luminol, Hexagon-OBTI(R), and Combur(R) tests. *Forensic Sci Int*, 2011, **209**: 59–63.

Sato, I., et al., Forensic hair analysis to identify animal species on a case of pet animal abuse. *Int J Legal Med*, 2010, **124**(3): 249–256.

Shaler, R.C., Modern forensic biology, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 525–614, 2002. Upper Saddle River, NJ: Pearson Education.

Sivaram, S., et al., Differentiation between stains of human blood and blood of monkey. *Forensic Sci*, 1975, **6**(3): 145–152.

Spear, T.F. and S.A. Binkley, The HemeSelect test: A simple and sensitive forensic species test. *J Forensic Sci Soc*, 1994, **34**(1): 41–46.

Tabata, N. and M. Morita, Immunohistochemical demonstration of bleeding in decomposed bodies by using anti-glycophorin A monoclonal antibody. *Forensic Sci Int*, 1997, **87**(1): 1–8.

Tanaka, Y., et al., Enzyme-linked immunospot assay for detecting cells secreting antibodies against human blood group A epitopes. *Transplant Proc*, 2003, **35**(1): 555–556.

Tobe, S. and A. Linacre, Species identification of human and deer from mixed biological material. *Forensic Sci Int*, 2007, **169**(2–3): 278–279.

Tobe, S.S., et al., Recovery of human DNA profiles from poached deer remains part 2: Improved recovery protocol without the need for LCN analysis. *Sci Justice*, 2013, **53**(1): 23–27.

Tsutsumi, H. and Y. Katsumata, Forensic study on stains of blood and saliva in a chimpanzee bite case. *Forensic Sci Int*, 1993, **61**(2–3): 101–110.

van Asch, B. and A. Amorim, Capillary electrophoresis analysis of a 9-plex STR assay for canine genotyping. *Methods Mol Biol*, 2012, **830**: 231–240.

van Asch, B., et al., A new autosomal STR nineplex for canine identification and parentage testing. *Electrophoresis*, 2009, **30**(2): 417–423.

van Asch, B., et al., Forensic analysis of dog (*Canis lupus familiaris*) mitochondrial DNA sequences: An inter-laboratory study of the GEP-ISFG working group. *Forensic Sci Int Genet*, 2009, **4**(1): 49–54.

Virkler, K. and I.K. Lednev, Blood species identification for forensic purposes using Raman spectroscopy combined with advanced statistical analysis. *Anal Chem*, 2009, **81**(18): 7773–7777.

Whitehead, P.H. and A. Brech, A micro-technique involving species identification and ABO grouping on the same fragment of blood. *J Forensic Sci Soc*, 1974, **14**(2): 109–110.

Wictum, E., et al., Developmental validation of DogFiler, a novel multiplex for canine DNA profiling in forensic casework. *Forensic Sci Int Genet*, 2013, **7**(1): 82–91.

Wilson-Wilde, L., et al., Current issues in species identification for forensic science and the validity of using the cytochrome oxidase I (COI) gene. *Forensic Sci Med Pathol*, 2010, **6**(3): 233–241.

14

Identification of Semen

14.1 Biological Characteristics

A typical ejaculation releases 2–5 mL of semen, which contains seminal fluid and sperm cells (*spermatozoa*). A normal sperm count ranges from 10^7 to 10^8 spermatozoa per milliliter of semen. The spermatozoa are formed from spermatogonia in the seminiferous tubules of the testes. This process of generating spermatozoa is referred to as spermatogenesis (Figure 14.1). The spermatozoa are then transported and stored in the tubular network of the epididymis where they undergo functional maturation (spermatogenesis and maturation take approximately 3 months). The epididymis joins the ductus deferens, which transports matured sperm from the epididymis to the ejaculatory duct. From there, spermatozoa follow the ejaculatory ducts into the prostatic urethra where they are joined with secretions from the prostate. Figure 14.2 illustrates the anatomy of the male reproductive system.

Seminal fluid is a complex mixture of glandular secretions. A typical sample of seminal fluid contains the combined secretions of several accessory glands. Seminal vesicle fluid accounts for approximately 60% of the ejaculate. Various proteins secreted by the seminal vesicles play a role in the coagulation of the ejaculate. Additionally, seminal vesicle fluid contains flavin, which causes semen to fluoresce under ultraviolet light, often utilized when searching for semen-stain evidence.

Prostatic fluid secretions account for approximately 30% of the ejaculate. The components of this fluid are complex as well. This portion of semen contains high concentrations of *acid phosphatase* (AP) and *prostate-specific antigen* (PSA). Both are useful markers for the identification of semen in forensic laboratories. The epididymis and the bulbourethral secretions each account for approximately 5% of the ejaculate.

A *vasectomy* is the surgical removal of a bilateral segment of the ductus deferens. The surgery prevents spermatozoa from reaching the distal portions of the male reproductive tract. However, a vasectomized male can still produce ejaculate that contains only seminal vesicle fluid, prostatic fluid, and bulbourethral fluid. The condition by which males have abnormally low sperm counts is known as *oligospermia*. *Azoospermia* is a condition that causes males to produce no spermatozoa. However, the secretion of seminal fluid is not affected in males who have these conditions. DNA derived from epithelial cells can be isolated from the seminal fluids of these individuals.

14.1.1 Spermatozoa

A human spermatozoon has three morphologically distinct structures: the head, the middle piece, and the tail (Figure 14.3). The head contains a nucleus with densely packed chromosomes. At the tip of the head is the *acrosomal cap*, which is a membranous compartment

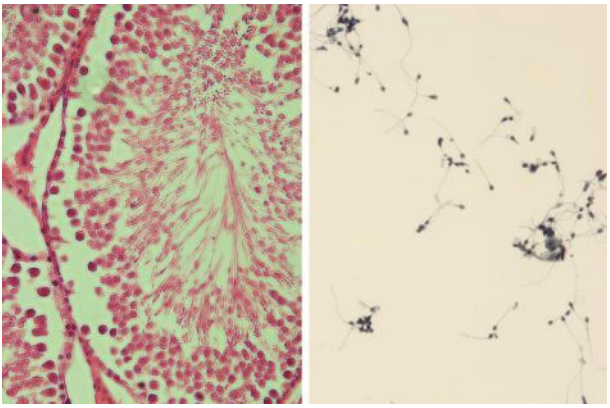

Figure 14.1 Spermatogenesis. In the seminiferous tubules (left) of the testes, spermatogonia (located at peripheral area of the seminiferous tubules) are differentiated to spermatids (located at the center of the seminiferous tubules). The spermatids are eventually differentiated to matured spermatozoa (right). (© Richard C. Li.)

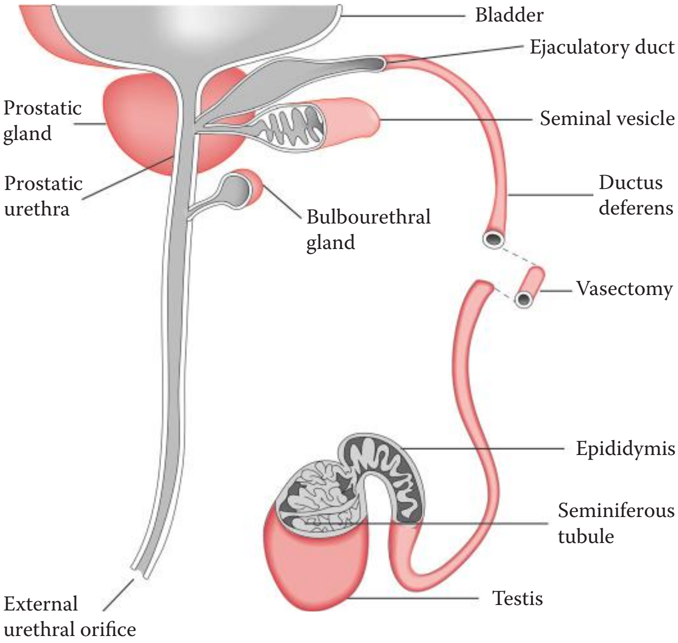
Bladder
Ejaculatory duct
Prostatic gland
Seminal vesicle
Prostatic urethra
Ductus deferens
Bulbourethral gland
Vasectomy
Epididymis
Seminiferous tubule
Testis
External urethral orifice

Figure 14.2 Male reproductive system and accessory glands (unilateral view). (© Richard C. Li.)

containing enzymes essential for fertilization. The head is attached to the middle piece through a short neck where the mitochondria that provide the energy for moving the tail are located. The tail or flagellum is responsible for spermatozoon motility. In contrast to other cell types, a mature spermatozoon lacks various intracellular organelles such as an endoplasmic reticulum, Golgi apparatus, lysosomes, and peroxisomes. In a normal male, at least 60% of spermatozoa have normal morphology, so morphological abnormalities can often be observed.

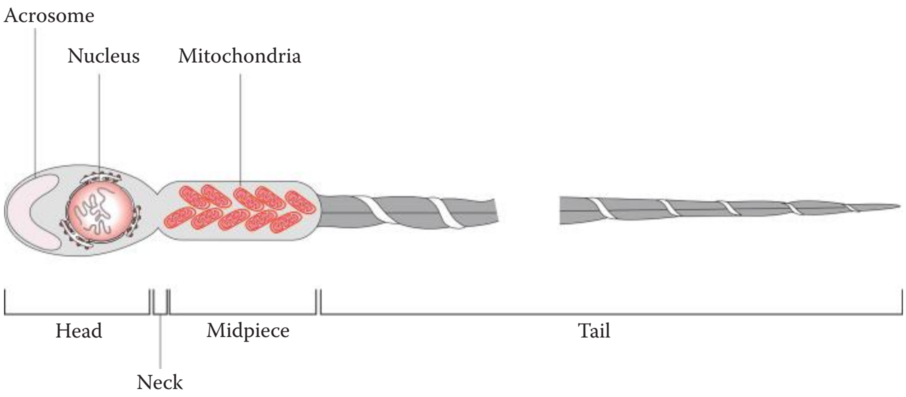

Figure 14.3 Structure of spermatozoon. (© Richard C. Li.)

14.1.2 Acid Phosphatase

Acid phosphatase (AP) consists of a group of phosphatases with optimal activity in an acidic pH environment. The greatest forensic importance of AP is that the prostate-derived AP contributes most of the AP activity present in semen. AP levels in semen are not affected by vasectomies. AP isoenzymes are also found in other tissues (Section 16.2).

The half-life of AP activity at 37°C is 6 months. However, the half-life is decreased if a sample is stored in a wet environment. AP activity can be detected from dry seminal stains stored at −20°C up to 1 year. Low levels of prostatic AP are present in the sera of healthy males. Elevated levels of prostatic AP found in serum are useful in diagnosing and monitoring prostate carcinoma. Many AP tests utilized in clinical testing may be used to identify semen for forensic applications.

14.1.3 Prostate-Specific Antigen

Prostate-specific antigen (PSA) is a major protein present in seminal fluid at concentrations of 0.5–2.0 mg/mL. PSA is produced in the prostate epithelium and secreted into the semen. PSA can also be found in the paraurethral glands, perianal glands, apocrine sweat glands, and mammary glands. Thus small quantities can be detected in urine, fecal material, sweat, and milk. PSA can also be found at much lower levels in the bloodstream. An elevated plasma PSA is present in prostate cancer patients, and it is widely used as a screening test for this disease. PSA is also elevated in cases of benign prostatic hyperplasia and prostatitis. The synthesis of PSA is stimulated by androgen, a steroid hormone.

PSA is a protein that has a molecular weight of 30 kDa and is thus also known as *P30*. It is responsible for hydrolyzing *semenogelin* (Sg), which mediates gel formation in semen (Section 14.1.4). PSA is a member of the tissue *kallikrein* (serine protease) family and is encoded by the *KLK3* locus located on chromosome 19. In addition to PSA, other tissue kallikreins encoded by *KLK2* and *KLK4* loci are expressed in the prostate. The half-life for PSA in a dried semen stain is about 3 years at room temperature. The half-life is greatly reduced when a sample is stored in wet conditions.

14.1.4 Seminal Vesicle–Specific Antigen

Human *seminal vesicle–specific antigen* (SVSA) includes two major types, semenogelin I (SgI) and semenogelin II (SgII), and constitutes the major seminal vesicle–secreted protein in semen. On ejaculation, SVSA forms a coagulum that is liquefied after a few minutes due to the degradation of SVSA by PSA. In humans, both SgI and SgII are present in a number of tissues of the male reproductive system, including the seminal vesicles, ductus deferens, prostate, and epididymis. They are also present in several other tissues such as skeletal muscle, kidney, colon, and trachea. They have also been found in the sera of lung cancer patients.

The use of Sg as a marker for semen identification instead of PSA presents certain advantages. The concentration of Sg in seminal fluid is much higher than that of PSA, and this is beneficial for the sensitivity of detection. Sg is present in seminal fluid and absent in urine, milk, and sweat, where PSA can be found. Although Sg compounds are present in skeletal muscle, kidney, and colon, this is not a great concern because these tissue samples are not routinely collected for semen detection in sexual assault cases.

14.2 Analytical Techniques for Identifying Semen

The location of semen stains is usually carried out through visual examination. Particularly, the application of alternate light sources (ALSs) can facilitate searches for semen stains. The presumptive identification of semen is largely based on the detection of the presence of prostatic AP activity in a sample. However, most presumptive assays cannot completely distinguish prostatic AP from nonprostatic AP. Confirmatory assays for the identification of semen are available, including the microscopic examination of spermatozoa, the identification of PSA and SVSA, and the RNA-based assay.

14.2.1 Presumptive Assays

14.2.1.1 Lighting Techniques for Visual Examination of Semen Stains

Lighting techniques can be used to aid in searching for semen stains. A dried semen stain fluoresces under certain light sources such as ALSs or argon lasers. ALSs are most commonly utilized for the visual examination of semen stains (Chapter 1; Figures 14.4 and 14.5). Excitation wavelengths between 450 and 495 nm can be used, allowing for the visualization of fluorescence with orange goggles. However, this approach is not specific for semen. Other bodily fluid stains, such as saliva and urine stains, can also fluoresce with less intensity. Additionally, the intensity of the fluorescence can be affected by different colors of substrates, and the material, such as clothing, where semen stains have been deposited.

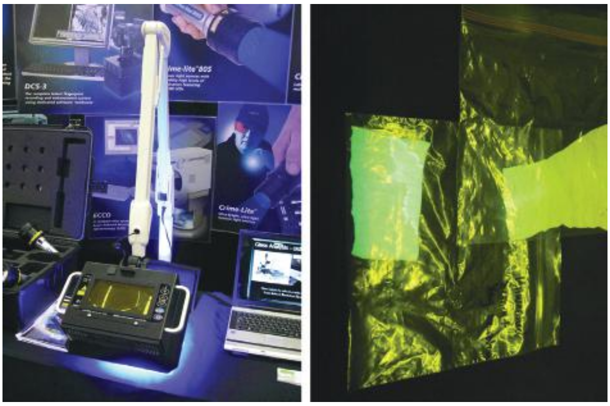

Figure 14.4 A tabletop ALS device (left) for the detection of semen stains (right). (© Richard C. Li.)

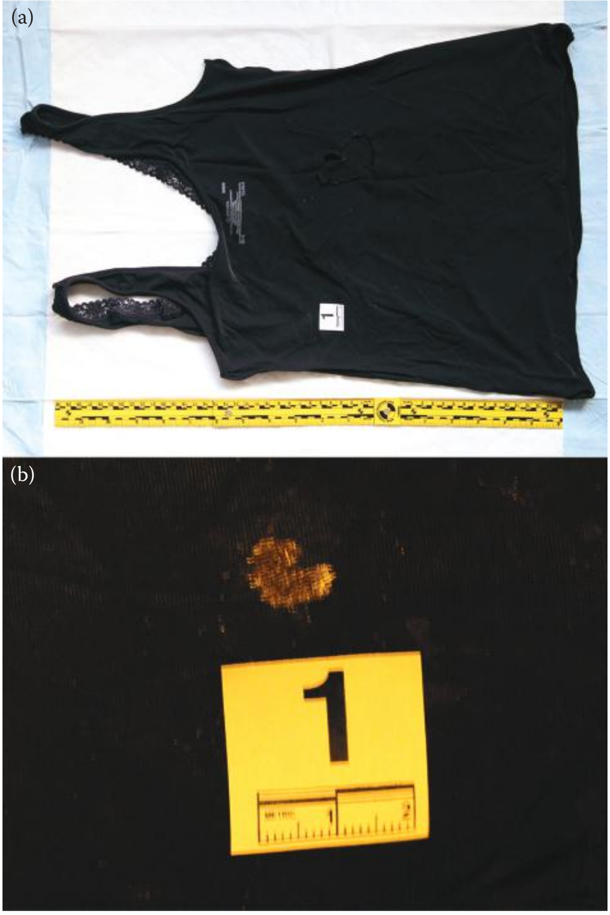

Figure 14.5 Examining a garment for semen stains. (a) A potential semen stain (labeled) is found on a garment, and (b) the stain fluoresces when irradiated with an ALS device. (© Richard C. Li.)

14.2.1.2 Acid Phosphatase Techniques

14.2.1.2.1 Colorimetric Assays

Colorimetric assays can be used for the presumptive identification of semen. The AP contained in semen can hydrolyze a variety of phosphate esters. It catalyzes the removal of the phosphate group from a substrate (Figure 14.6). Subsequently, an insoluble colored precipitate at sites of acid phosphatase activity is formed with a stabilized diazonium salt (usually in the form of zinc double salts).

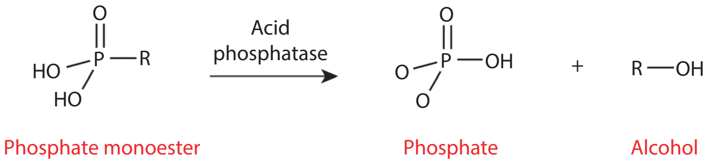

Figure 14.6 Reaction catalyzed by acid phosphatases (EC 3.1.3.2). The optimal pH of the reaction is usually under pH 7.

However, interference during a test by nonprostatic AP isoenzymes (multiple forms of AP), such as contamination by AP commonly present in vaginal secretions (Chapter 16), can create problems in specimens collected from victims. Thus, it is desirable to be able to increase the specificity of the assay for prostatic AP. One solution is the application of substrates that are hydrolyzed rapidly by the prostatic enzyme and at a slower rate by the other forms of AP isoenzymes. For example, α-naphthylphosphate and thymolphthalein monophosphate are more specific to prostatic AP than phenyl phosphate and 4-nitrophenyl phosphate (Figure 14.7). The most common method for forensic applications is the use of α-naphthyl phosphate as a substrate. In the presence of AP, α-naphthylphosphate is hydrolyzed to phosphate and α-naphthol. Subsequently, the Fast Blue B, a stabilized diazonium salt, is added to carry out an azo coupling reaction, producing a purple azo dye (Figures 14.8 through 14.10).

Prostatic AP is water soluble. Thus, a moistened cotton swab or piece of filter paper can be used to transfer a small amount of sample from a stain by briefly pressing onto the questioned stain area. The α-naphthylphosphate reagent is added to the swab or filter paper followed by the addition of Fast Blue B reagent. If a purple coloration develops within 1 min, the test is considered a positive indication for semen. Color that develops after more than 1 min may arise from the activity of nonprostatic AP.

Additionally, the prostatic enzyme is strongly inhibited by dextrorotatory tartrate ions. Thus, these inhibitors, particularly tartrate, allow a distinction to be made between prostatic AP and other AP isoenzymes. Prostate and vaginal acid phosphatase can also be distinguished by using gel electrophoresis (Chapter 16).

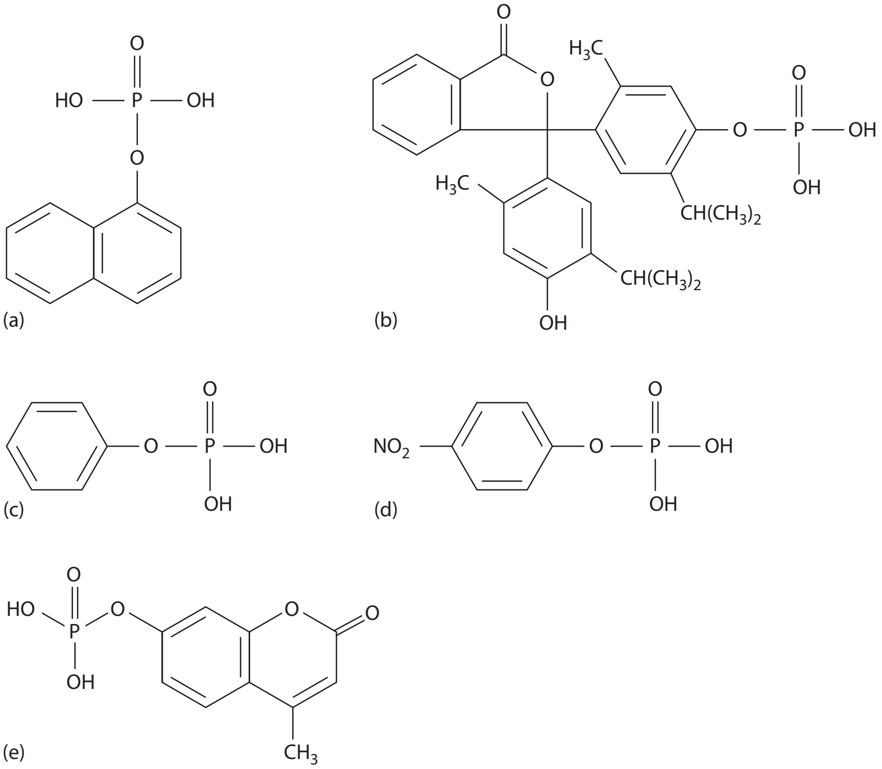

Figure 14.7 Chemical structures of acid phosphatase substrates. (a) α-Naphthyl phosphate, (b) thymolphthalein monophosphate, (c) phenyl phosphate, (d) 4-nitrophenyl phosphate, and (e) MUP.

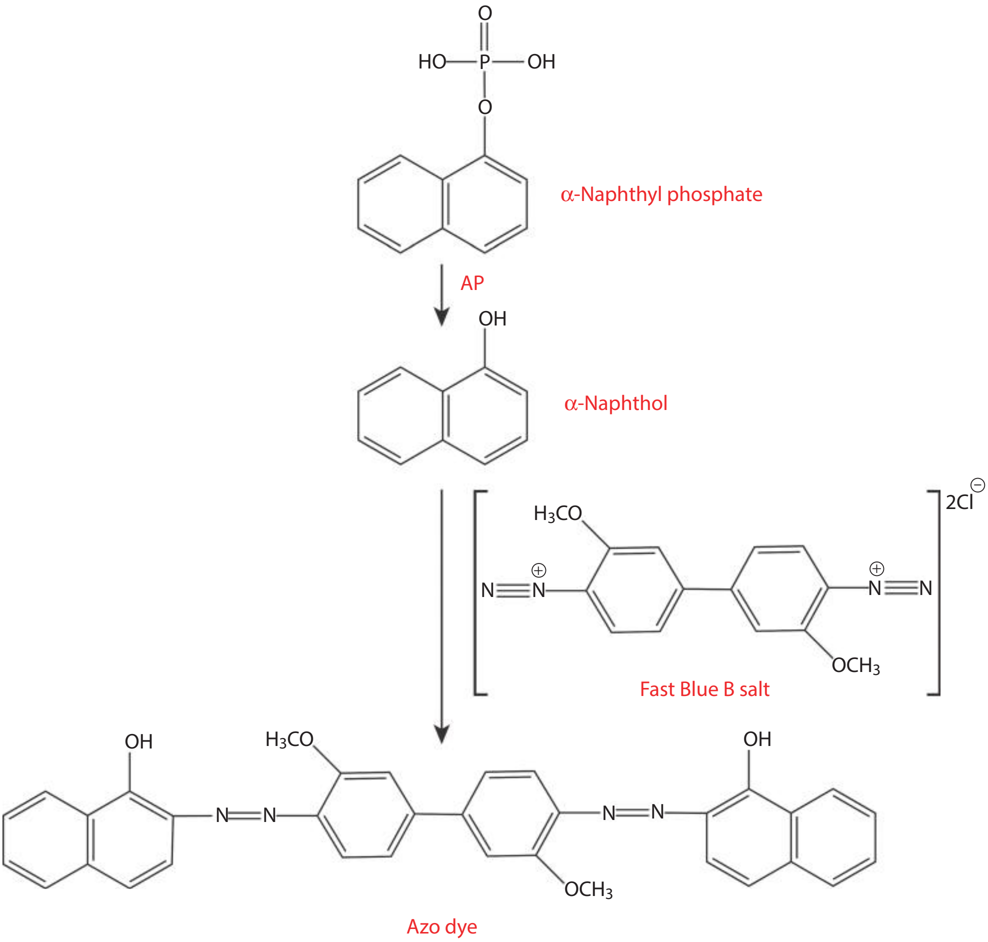

Figure 14.8 A colorimetric acid-phosphatase assay. In this assay, α-naphthylphosphate is hydrolyzed by acid phosphatase to phosphate and α-naphthol. The α-naphthol is subsequently converted into a purple azo dye with a diazonium salt such as Fast Blue B salt. AP, acid phosphatase.

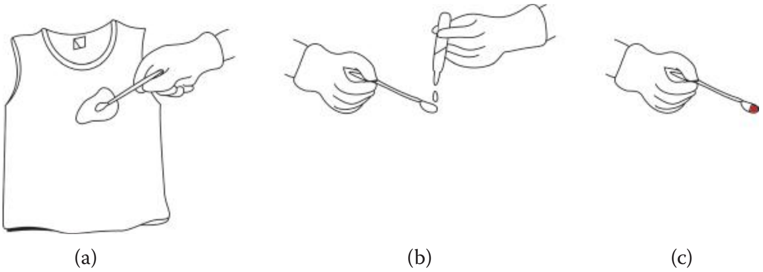

Figure 14.9 AP colorimetric assay. (a) A small amount of sample from the stain is transferred using a moistened cotton swab. (b) The substrate reagent is applied, followed by adding Fast Blue B reagent. (c) Purple coloration indicates a positive reaction. (© Richard C. Li.)

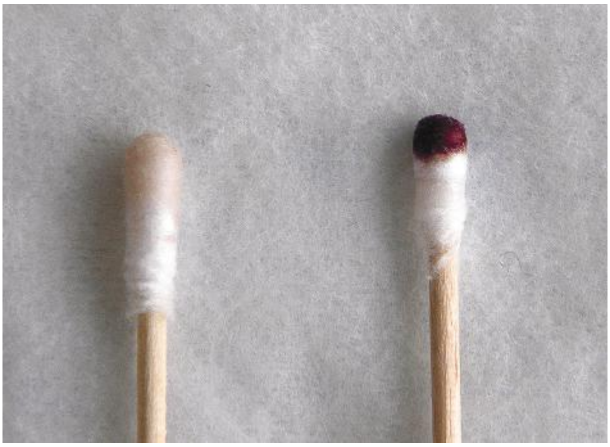

Figure 14.10 Photo of a colorimetric acid-phosphatase assay using α-naphthylphosphate as a substrate. (© Richard C. Li.)

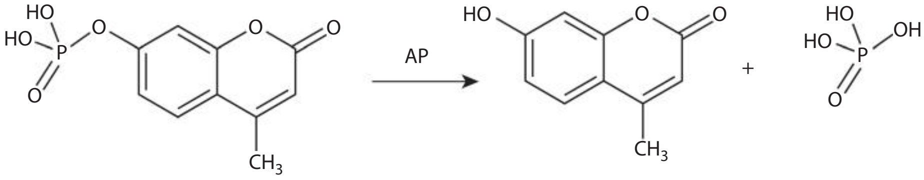

MUP

MU

Figure 14.11 The principle of MUP assay for detecting acid phosphatase activity. In the presence of acid phosphatase, 4-methylumbelliferone phosphate (MUP) is hydrolyzed, forming phosphate and 4-methylumbelliferone (MU), which fluoresces. AP, acid phosphatase.

14.2.1.2.2 Fluorometric Assays

Fluorometric methods are more sensitive than the colorimetric detection of AP and are used for semen stain mapping. AP catalyzes the removal of the phosphate residue on a 4-methylumbelliferone phosphate (MUP) substrate (Figure 14.11), a reaction that generates fluorescence under ultraviolet light. A piece of moistened filter paper, marked for proper orientation and identification, is used for transferring the prostatic AP. The evidence to be tested, a garment for example, is covered by the filter paper. Gloved hands are used to press the filter paper onto the stained area, ensuring that the evidence is in close contact with the paper. The filter paper is lifted from the evidence and examined in a dark room using long-wave ultraviolet light to detect any background fluorescence, which is then marked on the paper. The paper can then be sprayed with MUP reagent in a fume hood. The AP reaction on the paper can be visualized immediately. Areas where semen is present can be visualized as fluorescent areas on the filter paper (Figure 14.12).

14.2.2 Confirmatory Assays

14.2.2.1 Microscopic Examination of Spermatozoa

The cells from a questioned stain on an absorbent material can be transferred to a microscope slide by extracting a small portion of a stain with water, followed by gentle vortexing. The suspension is then transferred to a slide and evaporated at room temperature or fixed with low heat. Alternatively, it can be transferred by dampening the stain with water and rubbing or rolling it onto a microscope slide.

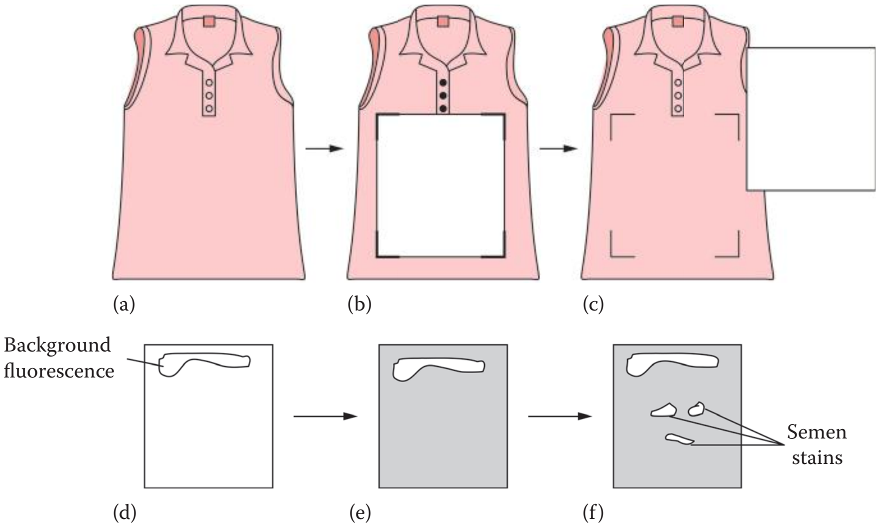

Figure 14.12 Fluorometric assay of acid phosphatase for locating semen stains. Evidence item (a) is closely covered by a piece of moistened filter paper (b) to allow transfer of a small amount of stain. The orientation of the paper is marked (c). The paper is lifted. The background fluorescence is marked under UV light (d). The filter paper is treated with MUP (e). The presence of fluorescence under ultraviolet light indicates semen stains (f). (© Richard C. Li.)

Microscopic identification of spermatozoa provides the proof of a seminal stain. Histological staining can facilitate microscopic examination. The most common staining technique is the *Christmas tree stain* (Figure 14.13). The red component known as Nuclear Fast Red (NFR) is a dye used for staining the nuclei of spermatozoa in the presence of aluminum ions. The green component, picroindigocarmine (PIC), stains the neck and tail portions of the sperm. The acrosomal cap and the nucleus stain pink-red, and the sperm tails and the midpiece stain blue-green. Epithelial cells, if present in the sample, appear blue-green and have red nuclei. Additionally, fluorescent detection utilizing SPERM HY-LITER Fluorescent Staining Kit can facilitate the identification of spermatozoa.

Laser capture microdissection (LCM) has been shown to be an effective technique for separating spermatozoa from nonsperm cells (i.e., epithelial cells from the victim) on a glass slide (Figure 14.14). This technique involves using a thin layer of a thermosensitive polymer that is placed on the surface of an LCM cap. Once spermatozoa are identified on the slide under a microscope, a polymer-containing LCM cap is placed over the spermatozoa on the slide. An infrared laser melts the polymer and causes it to adhere only to the targeted spermatozoa. The spermatozoa are then lifted off the slide. This allows spermatozoa to be separated and placed into snap-cap tubes for forensic DNA analysis.

14.2.2.2 Identification of Prostate-Specific Antigen

Over the years, a number of methods have been utilized to detect PSA: immunodiffusion, immunoelectrophoresis, enzyme-linked immunosorbent assay (ELISA), and immunochromatographic assays. ELISA and immunochromatographic assays have been found to be the most sensitive methods (Chapter 11).

14.2.2.2.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits such as the PSA-check-1 (VED-LAB, Alencon), Seratec® PSA Semiquant (Seratec Diagnostica, Göttingen), and One Step ABAcard

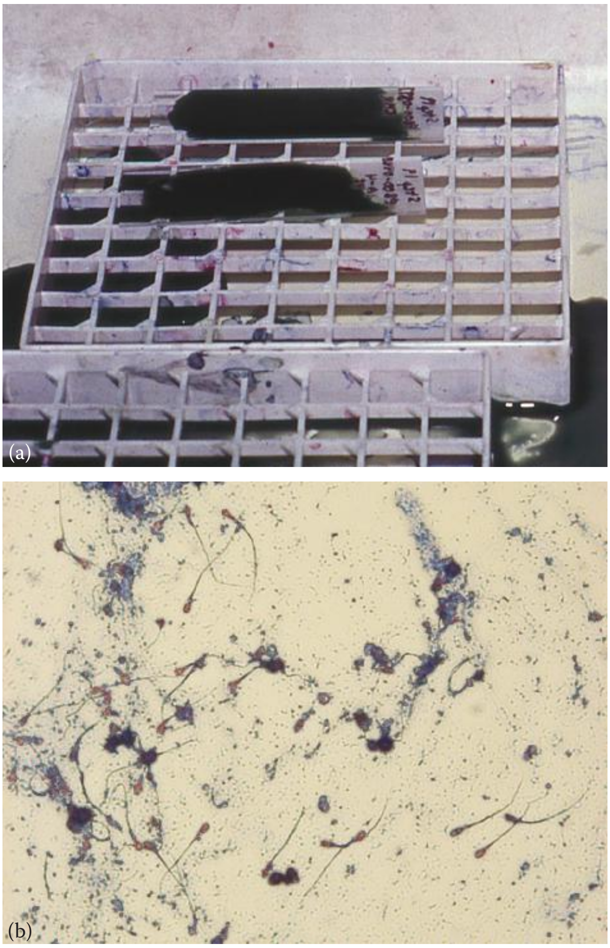

Figure 14.13 Human spermatozoa stained with Christmas tree stain. (a) Staining spermatozoa on a microscope slide and (b) stained spermatozoa. (© Richard C. Li.)

Figure 14.14 A microscopic device for fluorescent detection utilizing SPERM HY-LITER Fluorescent Staining Kit (left) and a laser-capture microdissection device for separating sperm cells from other types of cells (right). (© Richard C. Li.)

PSA® (Abacus Diagnostics, California) are available. These devices utilize antihuman PSA antibodies. In the ABAcard PSA® assay, a labeled monoclonal antihuman PSA antibody is contained in a sample well, a polyclonal antihuman PSA antibody is immobilized on a test zone of a nitrocellulose membrane, and an antiglobulin that recognizes the antibody is immobilized on a control zone (Figures 14.15 and 14.16).

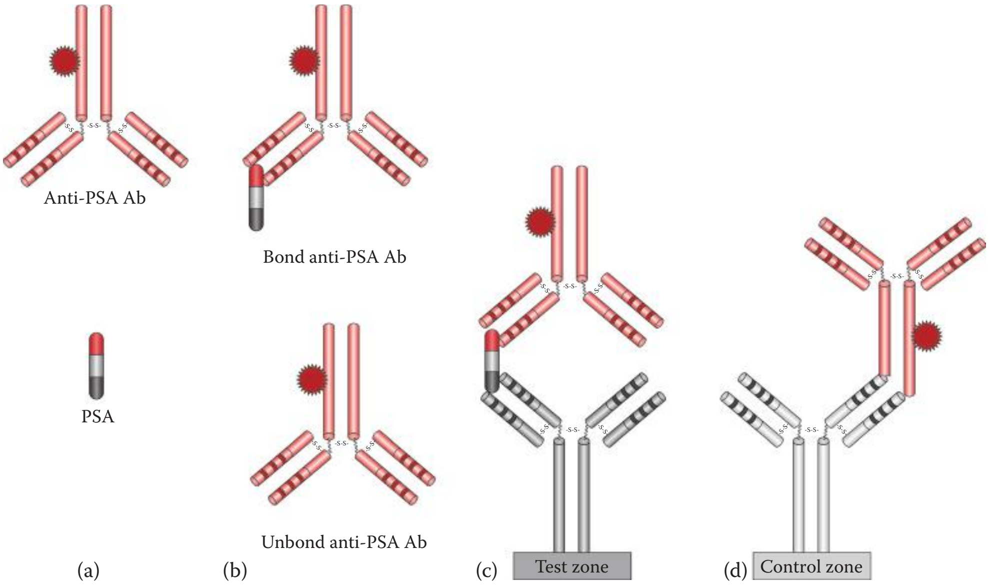

Figure 14.15 Immunochromatographic assays for identification of PSA in semen. (a) In a sample well, PSA in a semen sample is mixed with labeled anti-PSA Ab. (b) The PSA binds to the labeled anti-PSA Ab to form a labeled Ab–PSA complex. (c) At the test zone, the labeled complex binds to an immobilized anti-PSA Ab to form a labeled Ab–PSA–Ab sandwich. (d) At the control zone, the labeled anti-PSA Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab and PSA represent antibody and prostate-specific antigen, respectively. (© Richard C. Li.)

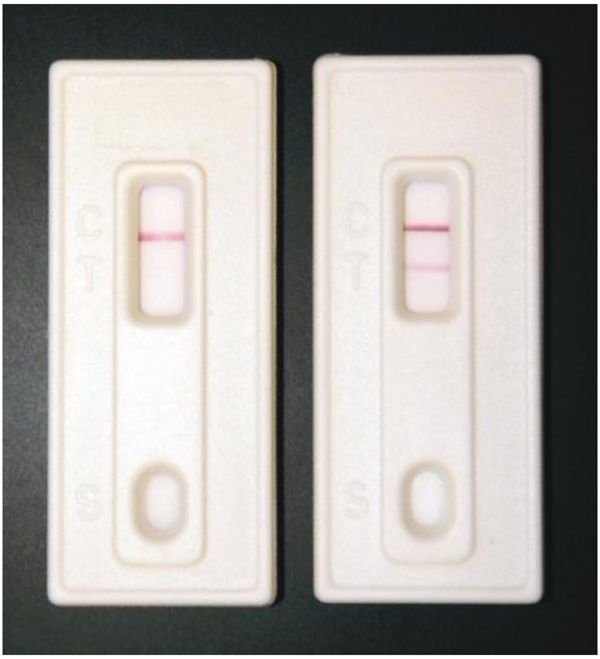

Figure 14.16 Semen identification using an immunochromatic device (ABAcard PSA). The negative (left) and positive (right) results are shown. The "C" band indicates that the test is valid. The "T" band indicates the presence of human blood. The sample well is labeled as "S". (© Richard C. Li.)

The assay is carried out by loading an extracted sample into the sample well. The antigen in the sample binds to the labeled antibody in the sample well to form an antigen–antibody complex. The complex then diffuses across the nitrocellulose membrane. At the test zone, the immobilized antihuman PSA antibody binds with the antigen–antibody complex to form an antibody–antigen–antibody sandwich. The ABAcard PSA® uses a pink dye that allows for the visualization of a positive test with a pink line at the test zone. In the control zone, unbound labeled antihuman PSA antibody binds to the immobilized antiglobulin. This antibody–antiglobulin complex at the control zone also results in a pink line. The test is considered valid only if the line in the control zone is observed.

The presence of human PSA results in a pink line at both the test and control zones. The absence of human PSA produces a pink line in the control zone only. A positive result can appear within 1 min; a negative result is read after 10 min. However, the high-dose hook effect, an artifact that may cause false-negative results (Chapter 10), occurs when high quantities of seminal fluid are tested.

14.2.2.2.2 ELISA

The ELISA method can be used to detect PSA with anti-PSA antibodies. The most common method used in forensic serology is antibody sandwich ELISA, in which an antibody–antigen–antibody sandwich complex is formed (Figure 14.17). The intensity of the signal can be detected and is proportional to the amount of bound antigen. The amount of PSA can also be quantified by comparing a standard with known concentrations. Although this method is specific and highly sensitive, it is time-consuming. Chapter 11 discusses the principle of ELISA in further detail.

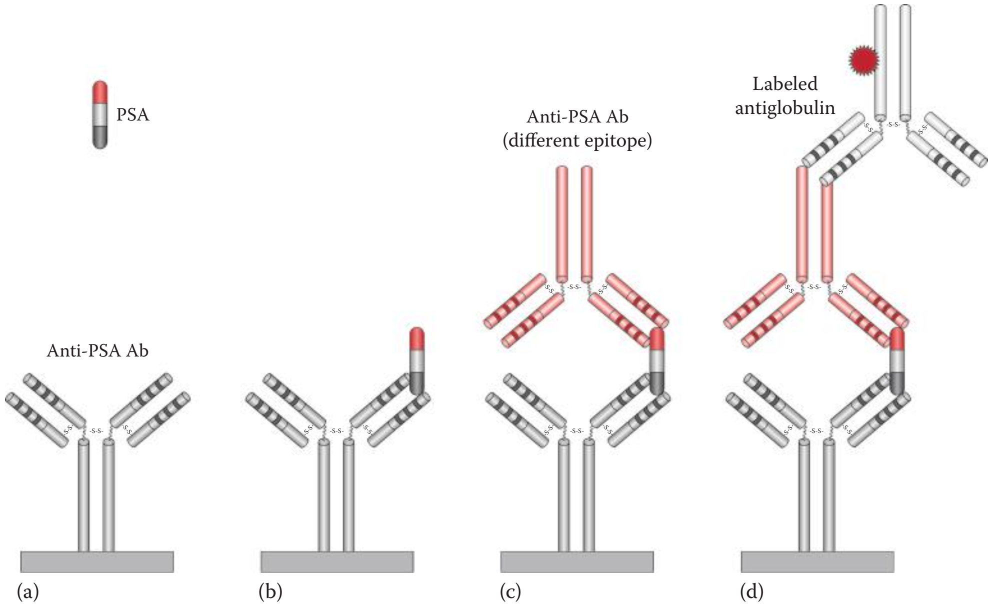

Figure 14.17 Use of ELISA for identification of PSA in semen. (a) Sample containing PSA is applied to polystyrene tubes in which anti-PSA Ab is immobilized. (b) The PSA binds to immobilized Ab to form a PSA–Ab complex. (c) A second anti-PSA Ab, specific for a different epitope of PSA, is added to form an Ab–PSA–Ab sandwich. (d) A labeled antiglobulin then binds to the Ab–PSA–Ab sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab and PSA represent antibody- and prostate-specific antigen, respectively. (© Richard C. Li.)

14.2.2.3 Identification of Seminal Vesicle–Specific Antigen

14.2.2.3.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits include the RSID®-Semen test (Independent Forensics, Hillside, IL) and the Nanotrap Sg. In the RSID®-Semen assay, a labeled monoclonal anti-Sg antibody is contained in a sample well, and a second monoclonal anti-Sg antibody, to a different epitope of Sg, is immobilized on the test zone of the membrane. An antiglobulin that recognizes the antibody is immobilized on a control zone (Figure 14.18).

The sample can be prepared by cutting a small portion of a stain or a swab and is extracted for 1–2 h in an extraction buffer (200–300 μL). Approximately 10% of the extract is removed and mixed with the running buffer. The assay is carried out by loading an extracted sample into the sample well. The antigen in the sample binds to the labeled anti-Sg antibody in the sample well to form a labeled antibody–antigen complex that then diffuses across the membrane. At the test zone, the solid-phase anti-Sg antibody binds with the labeled complex to form a labeled antibody–antigen–antibody sandwich. The antigen in the sample produces a pink line at the test zone. In the control zone, unbound labeled anti-Sg antibody binds to the solid-phase antiglobulin. This labeled antibody–antiglobulin complex at the control zone also results in a pink line. The presence of Sg generates a pink line at both the test and control zones. The absence of Sg results in a pink line in the control zone only. Results may be read after 10 min.

Validation studies have revealed that the sensitivity of the RSID-Semen kit for detecting seminal fluid can be as low as a 5×10^4-fold dilution. Species specificity studies have shown no cross-reactivity with various animal species including ruminants and small mammals. Bodily fluid specificity studies have also shown that the assay is not responsive to human blood, saliva,

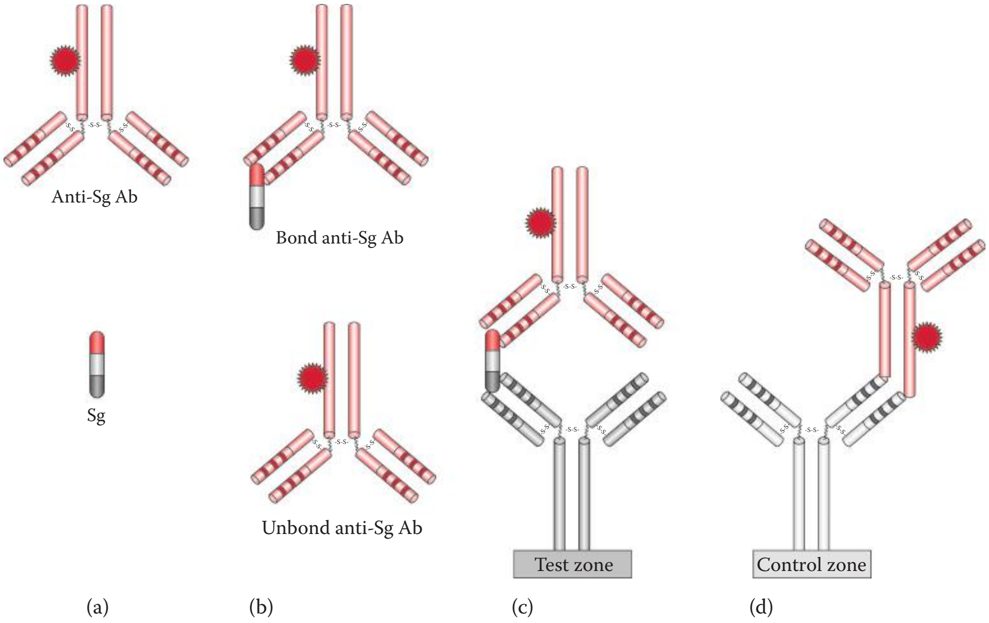

Figure 14.18 Immunochromatographic assays for identification of semenogelin (Sg) in semen. (a) In a sample well, Sg in a semen sample is mixed with labeled anti-Sg Ab. (b) Sg binds to the labeled anti-Sg Ab to form a labeled Ab–Sg complex. (c) At the test zone, the labeled complex binds to an immobilized anti-Sg Ab to form a labeled Ab–Sg–Ab sandwich. (d) At the control zone, the labeled anti-Sg Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab represents antibody. (© Richard C. Li.)

urine, sweat, fecal matter, milk, or vaginal secretions. The assay results are not affected by condom lubricants or spermicides such as nonoxynol-9 and menfegol. However, the high-dose hook effect occurs when more than 3 μL of seminal fluid is tested.

14.2.2.3.2 ELISA

Identification of Sg for semen detection has also been carried out with ELISA. Anti-Sg antibodies are utilized. An antibody–antigen–antibody sandwich complex is formed (Figure 14.19). The intensity of the colorimetric or fluorometric signals can be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of Sg can be quantified by comparing a standard with known concentrations.

14.2.2.4 RNA-Based Assays

RNA-based assays (Chapter 11) have been developed to identify semen. The assays are based on the expression of certain genes in certain cell or tissue types. Thus, the techniques used in the identification of semen are based on the detection of specific types of mRNA expressed exclusively in spermatozoa and in certain cells of male accessory glands. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect gene expression levels of mRNAs for semen identification. Table 14.1 lists the tissue-specific genes utilized for semen identification. Compared to conventional assays used for semen identification, the RNA-based assay has higher specificity and is amenable to automation. However, one limitation is that the RNA is unstable due to degradation by endogenous ribonucleases.

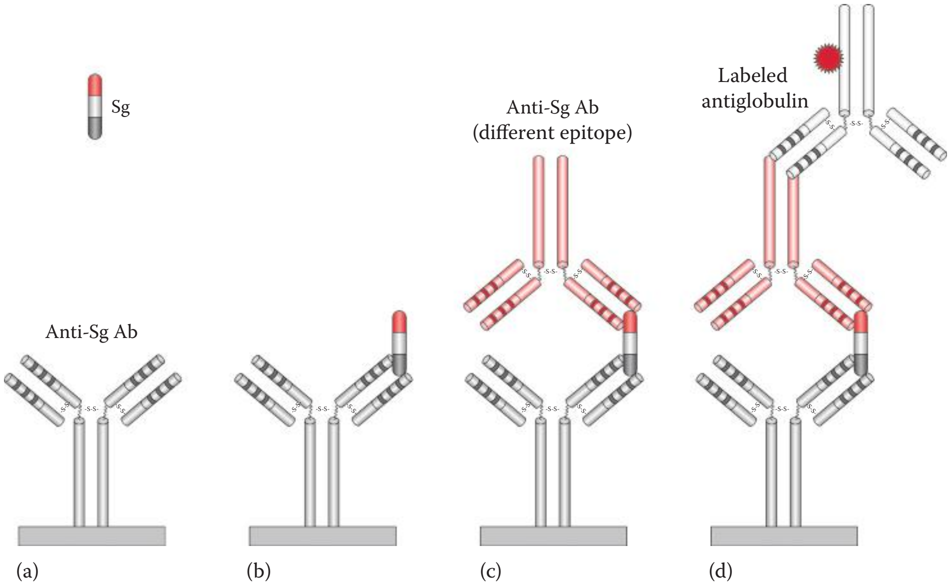

Figure 14.19 Use of ELISA for identification of semenogelin (Sg) in semen. (a) Sample containing Sg is applied to polystyrene tubes in which anti-Sg Ab is immobilized. (b) Sg binds to immobilized Ab to form a Sg–Ab complex. (c) A second anti-Sg Ab, specific for a different epitope of Sg, is added to form an Ab–Sg–Ab sandwich. (d) A labeled antiglobulin then binds to the Ab–Sg–Ab sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab represents antibody. (© Richard C. Li.)

Table 14.1 Application of RT-PCR for Semen Identification

Gene Symbol	Gene Product	Description	Further Reading
KLK3	Kallikrein 3	Also called prostate-specific antigen (PSA)	Gelmini et al. (2001)
PRM1	Protamine 1	DNA-binding proteins involved in condensation of sperm chromatin	Steger et al. (2000)
PRM2	Protamine 2	DNA-binding proteins involved in condensation of sperm chromatin	Steger et al. (2000)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

Bibliography

Ablett, P.J., The identification of the precise conditions for seminal acid phosphatase (SAP) and vaginal acid phosphatase (VAP) separation by isoelectric focusing patterns. *J Forensic Sci Soc*, 1983, **23**(3): 255–256.

Allard, J.E., The collection of data from findings in cases of sexual assault and the significance of spermatozoa on vaginal, anal and oral swabs. *Sci Justice*, 1997, **37**(2): 99–108.

Allard, J.E., et al., A comparison of methods used in the UK and Ireland for the extraction and detection of semen on swabs and cloth samples. *Sci Justice*, 2007, **47**(4): 160–167.

Allen, S.M., An enzyme linked immunosorbent assay (ELISA) for detection of seminal fluid using a monoclonal antibody to prostatic acid phosphatase. *J Immunoassay*, 1995, **16**(3): 297–308.

Allery, J.P., et al., Cytological detection of spermatozoa: Comparison of three staining methods. *J Forensic Sci*, 2001, **46**(2): 349–351.

Anoruo, B., et al., Isolating cells from non-sperm cellular mixtures using the PALM microlaser micro dissection system. *Forensic Sci Int*, 2007, **173**(2–3): 93–96.

Anslinger, K., et al., Sex-specific fluorescent labelling of cells for laser microdissection and DNA profiling. *Int J Legal Med*, 2007, **121**(1): 54–56.

Astrup, B.S., et al., Detection of spermatozoa following consensual sexual intercourse. *Forensic Sci Int*, 2012, **221**(1–3): 137–141.

Auvdel, M.J., Comparison of laser and high-intensity quartz arc tubes in the detection of body secretions. *J Forensic Sci*, 1988, **33**(4): 929–945.

Baechtel, S., The identification and individualization of semen stains, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 369–374, 1988. Englewood Cliffs, NJ: Prentice Hall.

Bauer, M. and D. Patzelt, Protamine mRNA as molecular marker for spermatozoa in semen stains. *Int J Legal Med*, 2003, **117**(3): 175–179.

Baxter, S.J., Immunological identification of human semen. *Med Sci Law*, 1973, **13**(3): 155–165.

Berti, A., et al., Expression of seminal vesicle–specific antigen in serum of lung tumor patients. *J Forensic Sci*, 2005, **50**(5): 1114–1145.

Bitner, S.E., False positives observed on the Seratec(R) PSA SemiQuant cassette test with condom lubricants. *J Forensic Sci*, 2012, **57**(6): 1545–1548.

Blake, E.T. and G.F. Sensabaugh, Genetic markers in human semen: A review. *J Forensic Sci*, 1976, **21**(4): 785–796.

Chapman, R.L., N.M. Brown, and S.M. Keating, The isolation of spermatozoa from sexual assault swabs using proteinase K. *J Forensic Sci Soc*, 1989, **29**(3): 207–212.

Chen, J.T. and G.L. Hortin, Interferences with semen detection by an immunoassay for a seminal vesicle-specific antigen. *J Forensic Sci*, 2000, **45**(1): 234–235.

Christoffersen, S., Immunohistochemical staining of human sperm cells in smears from sexual assault cases. *Int J Legal Med*, 2011, **125**(6): 887–890.

Clery, J.M., Stability of prostate specific antigen (PSA), and subsequent Y-STR typing, of *Lucilia* (*Phaenicia*) *sericata* (Meigen) (Diptera: Calliphoridae) maggots reared from a simulated postmortem sexual assault. *Forensic Sci Int*, 2001, **120**(1–2): 72–76.

Collins, K.A., et al., Identification of sperm and non-sperm male cells in cervicovaginal smears using fluorescence *in situ* hybridization: Applications in alleged sexual assault cases. *J Forensic Sci*, 1994, **39**(6): 1347–1355.

Cortner, G.V. and A.J. Boudreau, Phase contrast microscopy versus differential interference contrast microscopy as applicable to the observation of spermatozoa. *J Forensic Sci*, 1978, **23**(4): 830–832.

Dahlke, M.B., et al., Identification of semen in 500 patients seen because of rape. *Am J Clin Pathol*, 1977, **68**(6): 740–746.

Davidson, G. and T.B. Jalowiecki, Acid phosphatase screening—Wetting test paper or wetting fabric and test paper? *Sci Justice*, 2012, **52**(2): 106–111.

Davies, A., Evaluation of results from tests performed on vaginal, anal and oral swabs received in casework. *J Forensic Sci Soc*, 1978, **17**(2–3): 127–133.

Davies, A. and E. Wilson, The persistence of seminal constituents in the human vagina. *Forensic Sci*, 1974, **3**(1): 45–55.

De Moors, A., et al., Sperm Hy-Liter: An effective tool for the detection of spermatozoa in sexual assault exhibits. *Forensic Sci Int Genet*, 2013, **7**(3): 367–379.

Duenhoelter, J.H., et al., Detection of seminal fluid constituents after alleged sexual assault. *J Forensic Sci*, 1978, **23**(4): 824–829.

Elliott, K., et al., Use of laser microdissection greatly improves the recovery of DNA from sperm on microscope slides. *Forensic Sci Int*, 2003, **137**(1): 28–36.

Engelbertz, F., et al., Longevity of spermatozoa in the post-ejaculatory urine of fertile men. *Forensic Sci Int*, 2010, **194**(1–3): 15–19.

Enos, W.F. and J.C. Beyer, Prostatic acid phosphatase, aspermia, and alcoholism in rape cases. *J Forensic Sci*, 1980, **25**(2): 353–356.

Enos, W.F. and J.C. Beyer, Spermatozoa in the anal canal and rectum and in the oral cavity of female rape victims. *J Forensic Sci*, 1978, **23**(1): 231–233.

Eungprabhanth, V., Finding of the spermatozoa in the vagina related to elapsed time of coitus. *Z Rechtsmed*, 1974, **74**(4): 301–304.

Evers, H., et al., Investigative strategy for the forensic detection of sperm traces. *Forensic Sci Med Pathol*, 2009, **5**(3): 182–188.

Fraysier, H.D., A rapid screening technique for the detection of spermatozoa. *J Forensic Sci*, 1987, **32**(2): 527–530.

Gaensslen, R.E., *Sourcebook in Forensic Serology, Immunology, and Biochemistry*, 1983. Washington, DC: US Government Printing Office.

Gelmini, S., et al., Real-Time quantitative reverse transcriptase-polymerase chain reaction (RT-PCR) for the measurement of prostate-specific antigen mRNA in the peripheral blood of patients with prostate carcinoma using the taqman detection system. *Clin Chem Lab Med*, 2001, **39**(5): 385–391.

Greenfield, A. and Sloan, M., Identification of biological fluids and stains, in James, S.H., Nordby, J.J., and Bell, S. (eds), *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 2009. Boca Raton, FL: Taylor & Francis.

Haas, C., et al., mRNA profiling for the identification of sperm and seminal plasma. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 534–535.

Haas, C., et al., RNA/DNA co-analysis from human saliva and semen stains—Results of a third collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2013, **7**(2): 230–239.

Hardinge, P., et al., Optimisation of choline testing using Florence Iodine reagent, including comparative sensitivity and specificity with PSA and AP tests. *Sci Justice*, 2013, **53**(1): 34–40.

Healy, D.A., et al., Biosensor developments: Application to prostate-specific antigen detection. *Trends Biotechnol*, 2007, **25**(3): 125–131.

Herr, J.C. and M.P. Woodward, An enzyme-linked immunosorbent assay (ELISA) for human semen identification based on a biotinylated monoclonal antibody to a seminal vesicle–specific antigen. *J Forensic Sci*, 1987, **32**(2): 346–356.

Herr, J.C., et al., Characterization of a monoclonal antibody to a conserved epitope on human seminal vesicle–specific peptides: A novel probe/marker system for semen identification. *Biol Reprod*, 1986, **35**(3): 773–784.

Hochmeister, M.N., et al., High levels of alpha-amylase in seminal fluid may represent a simple artifact in the collection process. *J Forensic Sci*, 1997, **42**(3): 535–536.

Hochmeister, M.N., et al., Evaluation of prostate-specific antigen (PSA) membrane test assays for the forensic identification of seminal fluid. *J Forensic Sci*, 1999, **44**(5): 1057–1060.

Hueske, E.E., Techniques for extraction of spermatozoa from stained clothing: A critical review. *J Forensic Sci*, 1977, **22**(3): 596–598.

Ishiyama, I., Rapid histological examination of trace evidence by means of cellophane tape. *J Forensic Sci*, 1981, **26**(3): 570–575.

Iwasaki, M., et al., A demonstration of spermatozoa on vaginal swabs after complete destruction of the vaginal cell deposits. *J Forensic Sci*, 1989, **34**(3): 659–664.

Jones, E., The identification of semen and other body fluids, in R. Saferstein (ed.), *Forensic Science Handbook*, 2005. New Jersey: Pearson Prentice Hall.

Juusola, J. and J. Ballantyne, Messenger RNA profiling: A prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int*, 2003, **135**(2): 85–96.

Juusola, J. and J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**(1): 1–12.

Keil, W., J. Bachus, and H.D. Troger, Evaluation of MHS-5 in detecting seminal fluid in vaginal swabs. *Int J Legal Med*, 1996, **108**(4): 186–190.

Khaldi, N., et al., Evaluation of three rapid detection methods for the forensic identification of seminal fluid in rape cases. *J Forensic Sci*, 2004, **49**(4): 749–753.

Laffan, A., et al., Evaluation of semen presumptive tests for use at crime scenes. *Med Sci Law*, 2011, **51**(1): 11–17.

LaRue, B.L., J.L. King, and B. Budowle, A validation study of the Nucleix DSI-Semen kit—A methylation-based assay for semen identification. *Int J Legal Med*, 2013, **127**(2): 299–308.

Laux, D.L. and J.P. Barnhart, Validation of the Seratec(R) SeraQuant for the quantitation of prostate-specific antigen levels on immunochromatographic membranes. *J Forensic Sci*, 2011, **56**(6): 1574–1579.

Levine, B., et al., Use of prostate specific antigen in the identification of semen in postmortem cases. *Am J Forensic Med Pathol*, 2004, **25**(4): 288–290.

Lewis, J., et al., The fallacy of the two-minute acid phosphatase cut off. *Sci Justice*, 2012, **52**(2): 76–80.

Liedtke, R.J. and J.D. Batjer, Measurement of prostate-specific antigen by radioimmunoassay. *Clin Chem*, 1984, **30**(5): 649–652.

Lilja, H., A kallikrein-like serine protease in prostatic fluid cleaves the predominant seminal vesicle protein. *J Clin Invest*, 1985, **76**(5): 1899–1903.

Lin, M.F., et al., Fundamental biochemical and immunological aspects of prostatic acid phosphatase. *Prostate*, 1980, **1**(4): 415–425.

Linde, H.G. and K.E. Molnar, The simultaneous identification of seminal acid phosphatase and phosphoglucomutase by starch gel electrophoresis. *J Forensic Sci*, 1980, **25**(1): 113–117.

Lunetta, P. and H. Sippel, Positive prostate-specific antigen (PSA) reaction in post-mortem rectal swabs: A cautionary note. *J Forensic Leg Med*, 2009, **16**(7): 397–399.

Maher, J., et al., Evaluation of the BioSign PSA membrane test for the identification of semen stains in forensic casework. *N Z Med J*, 2002, **115**(1147): 48–49.

Masibay, A.S. and N.T. Lappas, The detection of protein p30 in seminal stains by means of thin-layer immunoassay. *J Forensic Sci*, 1984, **29**(4): 1173–1177.

Mauck, C.K., G.F. Doncel, and G. Biomarkers of semen exposure clinical working, biomarkers of semen in the vagina: Applications in clinical trials of contraception and prevention of sexually transmitted pathogens including HIV. *Contraception*, 2007, **75**(6): 407–419.

McAlister, C., The use of fluorescence *in situ* hybridisation and laser microdissection to identify and isolate male cells in an azoospermic sexual assault case. *Forensic Sci Int Genet*, 2011, **5**(1): 69–73.

McGee, R.S. and J.C. Herr, Human seminal vesicle–specific antigen is a substrate for prostate-specific antigen (or P-30). *Biol Reprod*, 1988, **39**(2): 499–510.

McWilliams, S. and B. Gartside, Identification of prostate-specific antigen and spermatozoa from a mixture of semen and simulated gastric juice. *J Forensic Sci*, 2009, **54**(3): 610–611.

Miller, K.W., et al., Developmental validation of the SPERM HY-LITER kit for the identification of human spermatozoa in forensic samples. *J Forensic Sci*, 2011, **56**(4): 853–865.

Mitchell, A., Validation study of KPICS SpermFinder™ by NicheVision Forensics, LLC for the identification of human spermatozoa. *J Forensic Sci*, 2012, **57**(4): 1042–1050.

Montagna, C.P., The recovery of seminal components and DNA from the vagina of a homicide victim 34 days postmortem. *J Forensic Sci*, 1996, **41**(4): 700–702.

Murray, C., C. McAlister, and K. Elliott, Identification and isolation of male cells using fluorescence *in situ* hybridisation and laser microdissection, for use in the investigation of sexual assault. *Forensic Sci Int Genet*, 2007, **1**(3–4): 247–252.

Nettikadan, S., et al., Detection and quantification of protein biomarkers from fewer than 10 cells. *Mol Cell Proteomics*, 2006, **5**(5): 895–901.

Nishi, K., et al., Utilization of lectin-histochemistry in forensic neuropathology: Lectin staining provides useful information for postmortem diagnosis in forensic neuropathology. *Leg Med* (Tokyo), 2003, **5**(3): 117–131.

Norris, J.V., et al., Expedited, chemically enhanced sperm cell recovery from cotton swabs for rape kit analysis. *J Forensic Sci*, 2007, **52**(4): 800–805.

Nussbaumer, C., E. Gharehbaghi-Schnell, and I. Korschineck, Messenger RNA profiling: A novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**(2–3): 181–186.

Old, J., et al., Developmental validation of RSID-Semen: A lateral flow immunochromatographic strip test for the forensic detection of human semen. *J Forensic Sci*, 2012, **57**(2): 489–499.

Pang, B.C. and B.K. Cheung, Identification of human semenogelin in membrane strip test as an alternative method for the detection of semen. *Forensic Sci Int*, 2007, **169**(1): 27–31.

Peonim, V., et al., Comparable between rapid one step immunochromatographic assay and ELISA in the detection of prostate specific antigen in vaginal specimens of raped women. *J Med Assoc Thai*, 2007, **90**(12): 2624–2629.

Poyntz, F.M. and P.D. Martin, Comparison of p30 and acid phosphatase levels in post-coital vaginal swabs from donor and casework studies. *Forensic Sci Int*, 1984, **24**(1): 17–25.

Randall, B., Glycogenated squamous epithelial cells as a marker of foreign body penetration in sexual assault. *J Forensic Sci*, 1988, **33**(2): 511–514.

Rao, A.R., H.G. Motiwala, and O.M. Karim, The discovery of prostate-specific antigen. *BJU Int*, 2008, **101**(1): 5–10.

Redhead, P. and M.K. Brown, The acid phosphatase test two minute cut-off: An insufficient time to detect some semen stains. *Sci Justice*, 2013, **53**(2): 187–191.

Rodrigues, R.G., et al., Semenogelins are ectopically expressed in small cell lung carcinoma. *Clin Cancer Res*, 2001, **7**(4): 854–860.

Romero-Montoya, L., et al., Relationship of spermatoscopy, prostatic acid phosphatase activity and prostate-specific antigen (p30) assays with further DNA typing in forensic samples from rape cases. *Forensic Sci Int*, 2011, **206**(1–3): 111–118.

Roy, A.V., M.E. Brower, and J.E. Hayden, Sodium thymolphthalein monophosphate: A new acid phosphatase substrate with greater specificity for the prostatic enzyme in serum. *Clin Chem*, 1971, **17**(11): 1093–1102.

Sakurada, K., et al., Evaluation of mRNA-based approach for identification of saliva and semen. *Leg Med* (Tokyo), 2009, **11**(3): 125–128.

Sato, I., et al., A dot-blot-immunoassay for semen identification using a polyclonal antibody against semenogelin, a powerful seminal marker. *Forensic Sci Int*, 2001, **122**(1): 27–34.

Sato, I., et al., Use of the "SMITEST" PSA card to identify the presence of prostate-specific antigen in semen and male urine. *Forensic Sci Int*, 2002, **127**(1–2): 71–74.

Sato, I., et al., Rapid detection of semenogelin by one-step immunochromatographic assay for semen identification. *J Immunol Methods*, 2004, **287**(1–2): 137–145.

Sato, I., et al., Applicability of Nanotrap Sg as a semen detection kit before male-specific DNA profiling in sexual assaults. *Int J Legal Med*, 2007, **121**(4): 315–319.

Sato, I., et al., Urinary prostate-specific antigen is a noninvasive indicator of sexual development in male children. *J Androl*, 2007, **28**(1): 150–154; discussion 155–157.

Schiff, A.F., Reliability of the acid phosphatase test for the identification of seminal fluid. *J Forensic Sci*, 1978, **23**(4): 833–844.

Sensabaugh, G.F., Isolation and characterization of a semen-specific protein from human seminal plasma: A potential new marker for semen identification. *J Forensic Sci*, 1978, **23**(1): 106–115.

Sensabaugh, G.F., The quantitative acid phosphatase test. A statistical analysis of endogenous and postcoital acid phosphatase levels in the vagina. *J Forensic Sci*, 1979, **24**(2): 19.

Shaler, R.C. and P. Ryan, High acid phosphatase levels as a possible false indicator of the presence of seminal fluid. *Am J Forensic Med Pathol*, 1982, **3**(2): 161–163.

Simich, J.P., et al., Validation of the use of a commercially available kit for the identification of prostate specific antigen (PSA) in semen stains. *J Forensic Sci*, 1999, **44**(6): 1229–1231.

Sokoll, L.J. and D.W. Chan, Prostate-specific antigen. Its discovery and biochemical characteristics. *Urol Clin North Am*, 1997, **24**(2): 253–259.

Standefer, J.C. and E.W. Street, Postmortem stability of prostatic acid phosphatase. *J Forensic Sci*, 1977, **22**(1): 165–172.

Stefanidou, M., G. Alevisopoulos, and C. Spiliopoulou, Fundamental issues in forensic semen detection. *West Indian Med J*, 2010, **59**(3): 280–283.

Steger, K., et al., Expression of protamine-1 and -2 mRNA during human spermiogenesis. *Mol Hum Reprod*, 2000, **6**(3): 219–225.

Stoilovic, M., Detection of semen and blood stains using polilight as a light source. *Forensic Sci Int*, 1991, **51**(2): 289–296.

Stowell, L.I., L.E. Sharman, and K. Hamel, An enzyme-linked immunosorbent assay (ELISA) for prostate-specific antigen. *Forensic Sci Int*, 1991, **50**(1): 125–138.

Stubbings, N.A. and P.J. Newall, An evaluation of gamma-glutamyl transpeptidase (GGT) and p30 determinations for the identification of semen on postcoital vaginal swabs. *J Forensic Sci*, 1985, **30**(3): 604–614.

Talthip, J., et al., An autopsy report case of rape victim by the application of PSA test kit as a new innovation for sexual assault investigation in Thailand. *J Med Assoc Thai*, 2007, **90**(2): 348–351.

Vandewoestyne, M. and D. Deforce, Laser capture microdissection in forensic research: A review. *Int J Legal Med*, 2010, **124**(6): 513–521.

Vandewoestyne, M., et al., Automatic detection of spermatozoa for laser capture microdissection. *Int J Legal Med*, 2009, **123**(2): 169–175.

Vandewoestyne, M., et al., Suspension fluorescence *in situ* hybridization (S-FISH) combined with automatic detection and laser microdissection for STR profiling of male cells in male/female mixtures. *Int J Legal Med*, 2009, **123**(5): 441–447.

Vennemann, M., et al., Sensitivity and specificity of presumptive tests for blood, saliva and semen. *Forensic Sci Med Pathol*, 2014, **10**(1): 69–75.

Virkler, K. and I.K. Lednev, Raman spectroscopic signature of semen and its potential application to forensic body fluid identification. *Forensic Sci Int*, 2009, **193**(1–3): 56–62.

Ward, A.M., J.W. Catto, and F.C. Hamdy, Prostate specific antigen: Biology, biochemistry and available commercial assays. *Ann Clin Biochem*, 2001, **38**(Pt 6): 633–651.

Wasserstrom, A., et al., Demonstration of DSI-semen—A novel DNA methylation-based forensic semen identification assay. *Forensic Sci Int Genet*, 2013, **7**(1): 136–142.

Willott, G.M. and J.E. Allard, Spermatozoa—Their persistence after sexual intercourse. *Forensic Sci Int*, 1982, **19**(2): 135–154.

Willott, G.M. and M.A. Crosse, The detection of spermatozoa in the mouth. *J Forensic Sci Soc*, 1986, **26**(2): 125–128.

Yokota, M., et al., Evaluation of prostate-specific antigen (PSA) membrane test for forensic examination of semen. *Leg Med* (Tokyo), 2001, **3**(3): 171–176.

Yoshida, K., et al., Quantification of seminal plasma motility inhibitor/semenogelin in human seminal plasma. *J Androl*, 2003, **24**(6): 878–884.

Yoshida, M., et al., Examination of seminal stain by HPLC assay of phenolphthalein. *Leg Med* (Tokyo), 2009, **11**(Suppl 1): S357–S359.

Yu, H. and E.P. Diamandis, Prostate-specific antigen in milk of lactating women. *Clin Chem*, 1995, **41**(1): 54–58.

Zarghami, N., et al., Prostate-specific antigen in serum during the menstrual cycle. *Clin Chem*, 1997, **43**(10): 1862–1867.

15

Identification of Saliva

15.1 Biological Characteristics of Saliva

The human salivary glands produce 1.0–1.5 L of saliva daily. About 70% of saliva is produced from the submandibular salivary glands, 25% from the parotids, and 5% from the sublingual salivary glands (Figure 15.1). Although a continuous basal level of saliva secretion is maintained, a large amount of saliva is produced during eating. Saliva is largely water containing small quantities of electrolytes, proteins, antibodies, and enzymes. The digestion of carbohydrates in the diet begins in the oral cavity, where amylase in the saliva breaks down carbohydrates such as starch. Thus, detecting amylase indicates the presence of saliva.

15.1.1 Amylases

Amylases are enzymes that cleave polysaccharides such as starches, which are composed of D-glucose units connected by α1→4 linkages. Starches contain two types of glucose polymers: *amylose* and *amylopectin* (Figure 15.2). Amylose consists of long, linear chains of glucose residues connected by α1→4 linkages. Amylopectin is highly branched and consists of linear chains of glucose residues connected by α1→4 linkages with the branch points connected by α1→6 linkages. Both linear amylose and amylopectin can be hydrolyzed by amylase by cleaving the chains at alternate α1→4 linkages. Amylase cleaves off one maltose (two glucose units) at a time. However, the α1→6 linkages at the branch points are not cleaved by the amylase.

Two types of amylases are characterized. β-*Amylases* found in plant and bacterial sources cleave only at the terminal reducing end of a polysaccharide chain. The end of a chain with a free anomeric carbon (not involved in a glycosidic bond) is called the reducing end. Human α-*amylases* cleave at α1→4 linkages randomly along the polysaccharide chain.

Human α-amylases have two major isoenzymes (multiple forms that differ in their amino acid sequences). Human *salivary* α-*amylase* (HSA) is encoded by the *Amy1* locus, synthesized at the salivary glands and secreted into the oral cavity. Human *pancreatic* α-amylase (HPA), encoded by the *Amy2* locus, is synthesized by the pancreas and secreted into the duodenum through the pancreatic duct. The amino acid sequences of the HSA and HPA are highly homologous. Therefore, monoclonal antibodies against HSA also cross-react with HPA. However, HSA is inactivated by acids in the stomach, while most HPA is inactivated in the lower portions of the intestine, and some amylase activity remains in the feces.

Amylase activity is found in various bodily fluids including semen, tears, milk, perspiration, and vaginal secretions. Most amylase present in normal serum consists of HPA and HSA. The amylases are small molecules and can pass through the glomeruli of the kidney (Chapter 17).

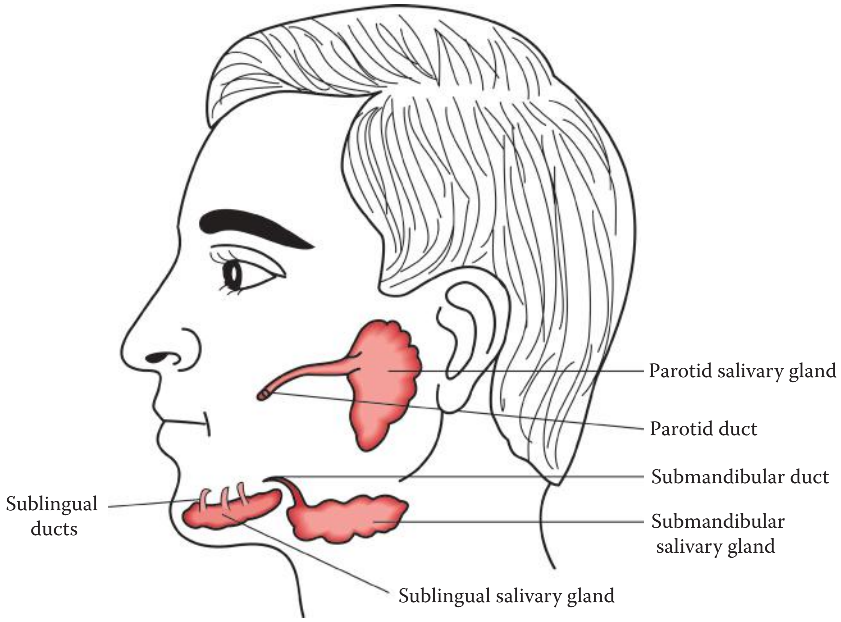
Parotid salivary gland
Parotid duct
Submandibular duct
Submandibular salivary gland
Sublingual ducts
Sublingual salivary gland

Figure 15.1 Human salivary glands. (© Richard C. Li.)

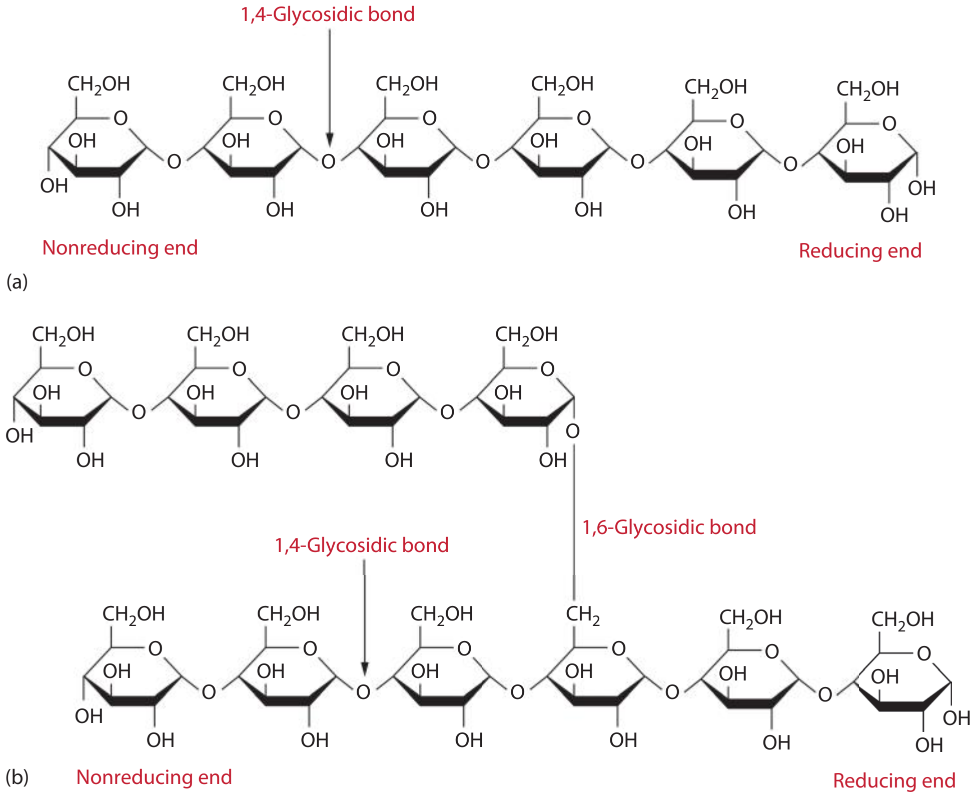
1,4-Glycosidic bond
Nonreducing end
Reducing end
(a)
1,6-Glycosidic bond
1,4-Glycosidic bond
(b)
Nonreducing end
Reducing end

Figure 15.2 Chemical structures of polysaccharides found in starch: (a) amylose and (b) amylopectin.

Thus, the amylase present in urine is derived from plasma. Amylase can be inactivated under boiling temperatures and strong acidic and alkaline conditions. Based on various studies, its stability varies from a few weeks to several months.

15.2 Analytical Techniques for Identification of Saliva

The use of alternate light sources facilitates the search for and the visual examination of saliva stains. The identification of saliva is largely based on detecting the presence of amylase in a sample. Two types of amylase assays can be utilized. The first type measures the enzymatic activity of total amylase. This type of assay cannot distinguish HSA from other amylases including HPA and nonhuman amylases, such as those from plants, animals, and microorganisms. The second type of assay, which includes direct detection of HSA proteins and RNA-based assays, is more confirmatory than enzymatic assays.

15.2.1 Presumptive Assays

15.2.1.1 Visual Examination

The lighting techniques used to search for semen stains can be utilized in searching for saliva stains. For example, a 470 nm excitation wavelength can be used with orange goggles to allow visualization of fluorescence. However, the fluorescence of a saliva stain is usually less intense than that of a seminal stain (Figure 15.3). Microscopic examination with proper histological staining can also be performed to identify the buccal epithelial cells, indicating the presence of a saliva stain (Figure 15.4).

15.2.1.2 Determination of Amylase Activity

15.2.1.2.1 Starch–Iodine Assay

Iodine (I_2) is used to test for the presence of starch. The amylose in starch reacts strongly with iodine to form a dark blue complex, while amylopectin develops a reddish-purple color. In the presence of amylases, starch is broken down to mono- or disaccharides. Consequently, such colors do not develop when iodine is added.

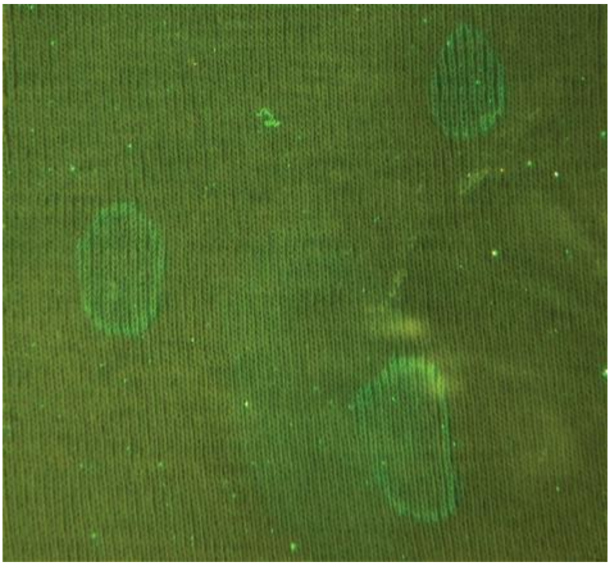

Figure 15.3 Examining saliva stains using an ALS technique. (© Richard C. Li.)

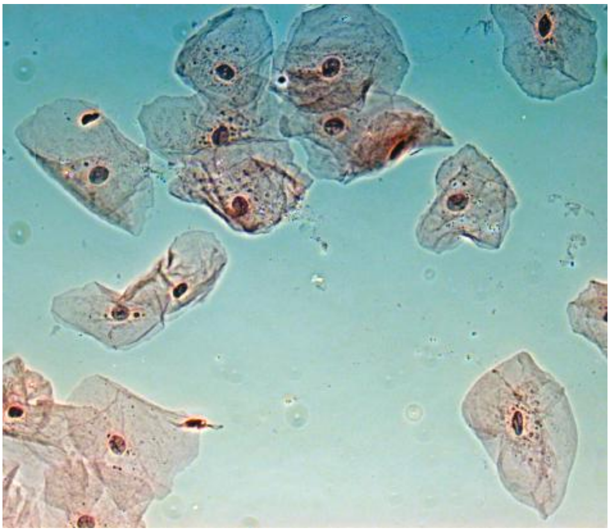

Figure 15.4 Buccal epithelial cells. (© Richard C. Li.)

A common configuration of the method is the radial diffusion assay (Figure 15.5). An agar gel containing starch is prepared. A sample well is created by punching a hole in the gel and an extract of the questioned sample is placed into the well. If amylase is present in the sample, it diffuses from the sample well and hydrolyzes the starch in the gel. The gel is then stained using an iodine solution. A clear area in the gel indicates amylase activity, and the size of the clear area is proportional to the amount of amylase in the sample. A linear standard curve

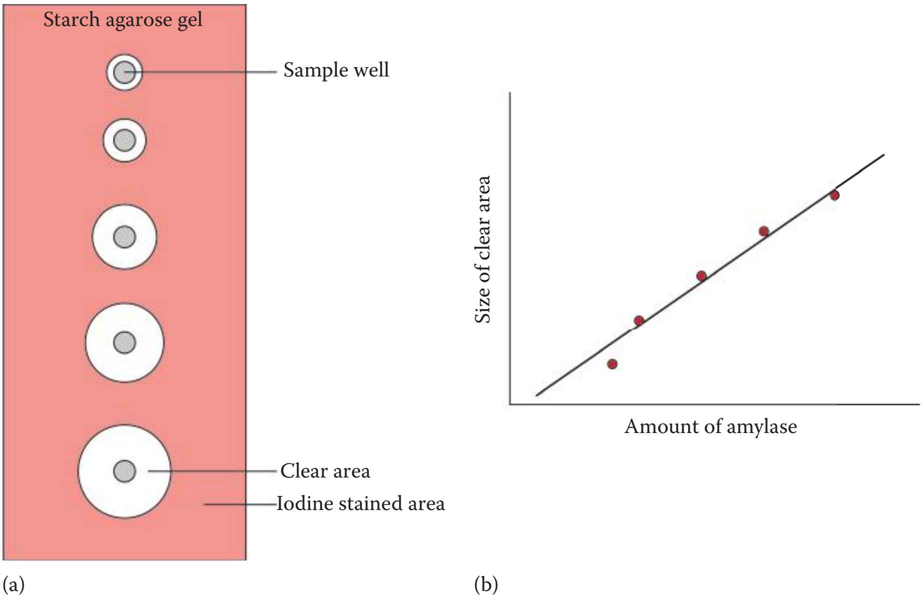

Figure 15.5 Radial diffusion assay for the identification of amylase. (a) Known amounts of amylase standards are applied to the well and allowed to diffuse. (b) The size of the clear area arising from amylase activity is plotted. The standard curve can be used to determine the amount of amylase in a questioned sample. (© Richard C. Li.)

(in log scale) can be prepared using a series of standard amylases with known concentrations. The amount of amylase can be quantified by comparing the results with the standard curve. However, this assay is not specific to HSA and can produce false-positive results. Moreover, the quantities of amylase can be used for determining the amount of sample needed for Y-STR analysis (Chapter 19), which now can be determined by using Y-chromosome-specific quantitative PCR assays (Chapter 6).

15.2.1.2.2 Colorimetric Assays

Dye-labeled amylase substrates such as dye-conjugated amylose or amylopectin are utilized. These substrates are not soluble in water. In the presence of amylase, the dye-containing moieties are cleaved and are soluble in water to produce a color. The degree of coloration, which can be measured colorimetrically by spectrophotometric methods, is proportional to the amount of amylase in the sample. Most of these assays are not HSA specific, although their specificity can be tested by using inhibitors that preferentially inhibit HSA, such as α-amylase inhibitors derived from the seeds of the wheat plant, *Triticum aestivum*. These amylase activity assays are considered presumptive, which means they are not conclusive for the presence of saliva in a sample.

While many substrates are available, Phadebas reagent (Pharmacia) is usually used in forensic laboratories. Produced in a tablet form, it is used to detect α-amylase in specimens for clinical diagnostic purposes. A small portion (approximately 3 mm^2) of a sample is cut and placed in a tube and incubated for 5 min at 37°C. One Phadebas tablet is added to each tube and mixed. Samples are then incubated for 15 min at 37°C, and the reaction is stopped at an alkaline pH by the addition of sodium hydroxide. The amylase substrate is an insoluble blue dye conjugated to starch. Amylase hydrolyzes the substrate to generate a blue color that can be measured at 620 nm using a spectrophotometer. The optical density of the supernatant is read and can be converted to amylase units by comparing to a standard curve.

The amylase assay can also be used for amylase mapping as a method of searching for possible saliva stains (Figures 15.6 and 15.7). These assays are based on the principle that amylase is water soluble and can be transferred from evidence to filter paper and then analyzed via colorimetric assay. This procedure is also referred to as a *press test*. The sensitivity of the method is similar to that of the test tube method.

The substrate can be prepared by evenly spraying the Phadebas reagent on a sheet of filter paper and allowing it to air-dry. The dried substrate-containing paper can be used immediately or stored until needed. To perform amylase mapping, a piece of paper is placed over the entire area to be tested (the item to be tested must be fairly flat to ensure good contact with the paper). The paper is dampened slightly by spraying with distilled water. An outline may be drawn on the paper to aid in locating stains. A piece of plastic wrap is placed on top to prevent the paper from drying during the assay, and a weight is applied to ensure close contact of substrate and evidence. The test is observed every minute for the first 10 min, and every 5 min thereafter up to 40 min, when a positive reaction should appear as a light blue area.

The SALIgAE® kit (Abacus Diagnostics), another commercially available colorimetric assay, has been validated for saliva identification (Figure 15.8). Its manufacturer also produces the SALIgAE spray kit, which can be used for amylase mapping.

15.2.2 Confirmatory Assays

15.2.2.1 Identification of Human Salivary α-Amylase

15.2.2.1.1 Immunochromatographic Assays

Commercially produced immunochromatographic kits include the RSID®-Saliva kit (Independent Forensics). A labeled monoclonal anti-HSA antibody is contained in a sample well. A second monoclonal anti-HSA antibody is immobilized onto a test zone of a membrane, and an antiglobulin that recognizes the antibody is immobilized onto a control zone (Figures 15.9 and 15.10).

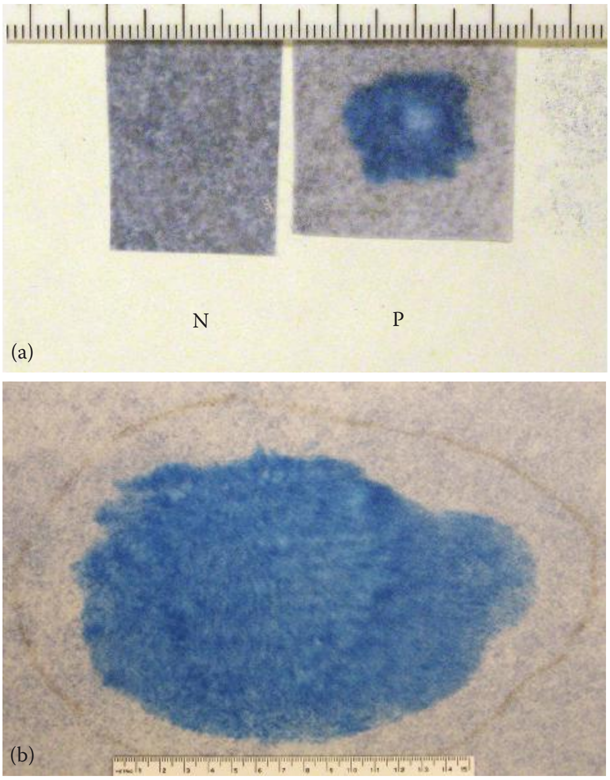

Figure 15.6 Amylase colorimetric assay using Phadebas reagent. (a) A spot test for saliva and (b) amylase mapping result showing a saliva-stained area. N, a negative result; P, a positive result. (© Richard C. Li.)

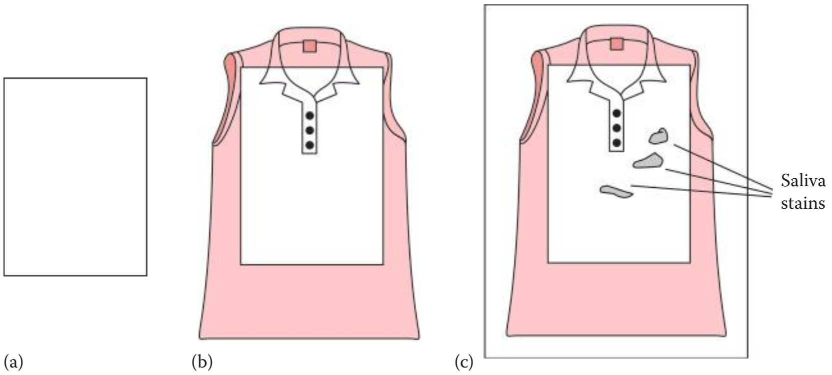

Figure 15.7 Amylase mapping for saliva stains. (a) Amylase substrate is sprayed on a sheet of filter paper. (b) Substrate-containing paper is placed over the area to be tested. The orientation of the filter paper is marked to aid in locating the stain. (c) The filter paper is dampened by spraying it with water, and plastic wrap is placed on top to prevent the paper from drying. Blue color indicates a positive reaction. (© Richard C. Li.)

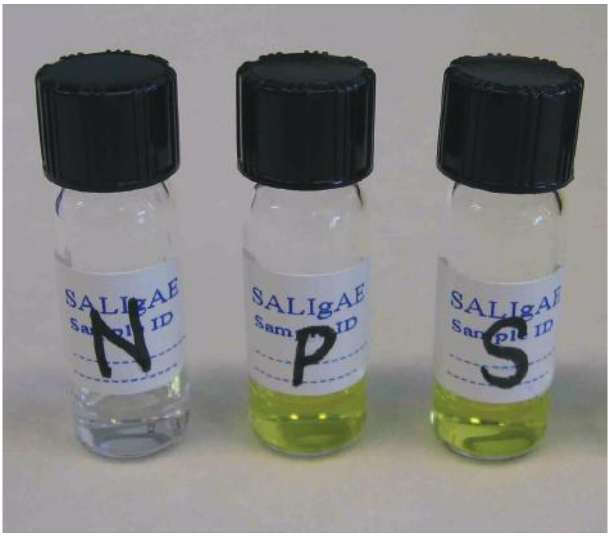

Figure 15.8 Saliva identification using SALIgAE reagent. P, a positive result; N, a negative result; S, a positive control. (© Richard C. Li.)

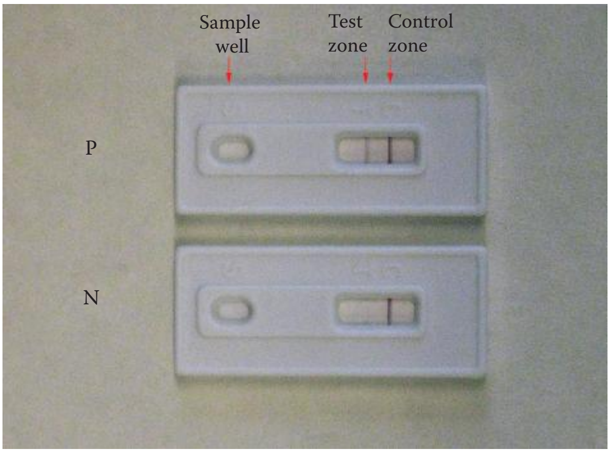

Figure 15.9 Identification of saliva using immunochromatographic assays. P, a positive result; N, a negative result. (© Richard C. Li.)

A sample can be prepared by cutting out a small portion of a stain or a swab. Each sample is extracted for 1–2 h in 200–300 μL of an extraction buffer. Approximately 10% of the extract is removed and mixed with a running buffer. The assay is carried out by loading an extracted sample into the sample well where the antigen in the sample binds to the labeled anti-HSA antibody in the well to form a labeled antibody–antigen complex. The complex then diffuses across the membrane to the test zone, where the solid-phase anti-HSA antibody binds with the labeled complex to form a labeled antibody–antigen–antibody sandwich.

The presence of antigen in the sample results in a pink line at the test zone. In the control zone, unbound labeled anti-HSA antibody binds to the solid-phase antiglobulin. The labeled antibody–antiglobulin complex at the control zone also results in a pink line. The test is

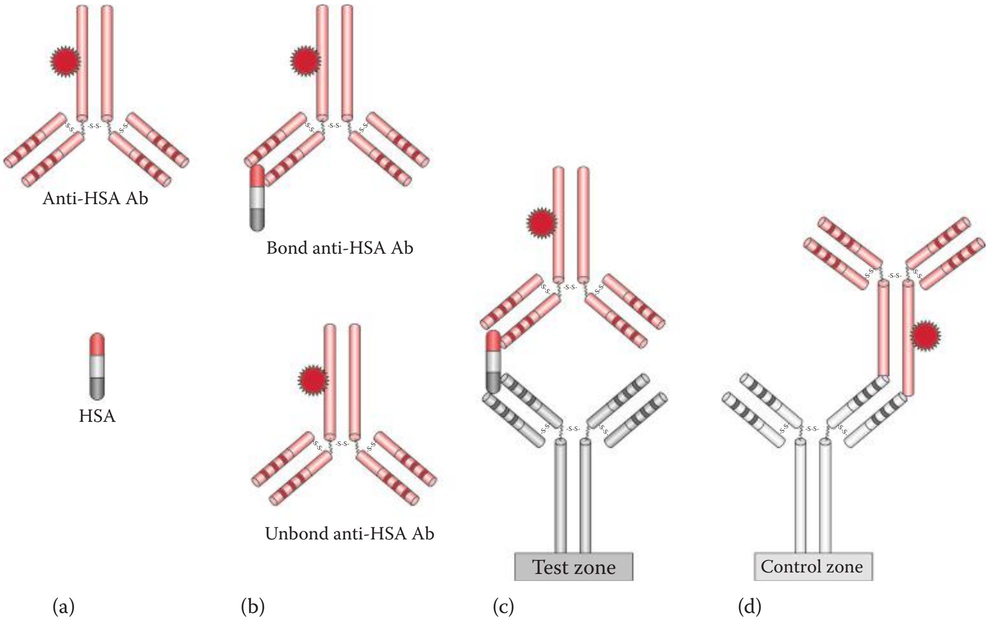

Figure 15.10 Immunochromatographic assay for the identification of HSA in saliva. (a) Sample containing amylases is loaded in a sample well. (b) Antigen binds to a labeled anti-HSA Ab to form a labeled Ab–HSA complex. (c) At the test zone, the labeled complex binds to an immobilized antihuman HSA Ab to form a labeled Ab–HSA–Ab sandwich. (d) At the control zone, the labeled anti-HSA Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, an antibody; HSA, human salivary α-amylase. (© Richard C. Li.)

considered valid only if the line in the control zone is observed. The presence of HSA results in a pink line at both the test zone and the control zone, while the absence of HSA results in a pink line in the control zone only. A result can be read after 10 min.

The sensitivity of the RSID®-Saliva kit can be as low as 1 μL of saliva. Additionally, the assay is responsive to samples extracted from saliva stains on both smooth and porous surfaces. In terms of species specificity, the kit has no cross-reactivity with various animal species, including monkeys (tamarin and callimico). As for bodily fluid specificity, it has also been shown that it is not responsive to human blood, semen, or urine. The high-dose hook effect, which creates an artifact that may cause false-negative results as described previously (Chapter 11), is not observed when up to 50 μL of saliva is tested. This method is rapid, specific, and sensitive and can be used in both laboratory and field analysis.

15.2.2.1.2 Enzyme-Linked Immunosorbent Assay (ELISA)

This method can be used to detect and to quantify a sample with the use of an anti-HSA antibody. The most common configuration in forensic serology is the antibody–antigen–antibody sandwich (Figure 15.11). ELISA utilizes reporting enzymes to produce colorimetric or fluorometric signals. The intensity of the signal can be detected spectrophotometrically and is proportional to the amount of bound antigen. The amount of HSA can be quantified by comparison with a standard of known concentration. This method is specific and highly sensitive in detecting HSA, but it is time-consuming. Chapter 11 discusses the ELISA principle in further detail.

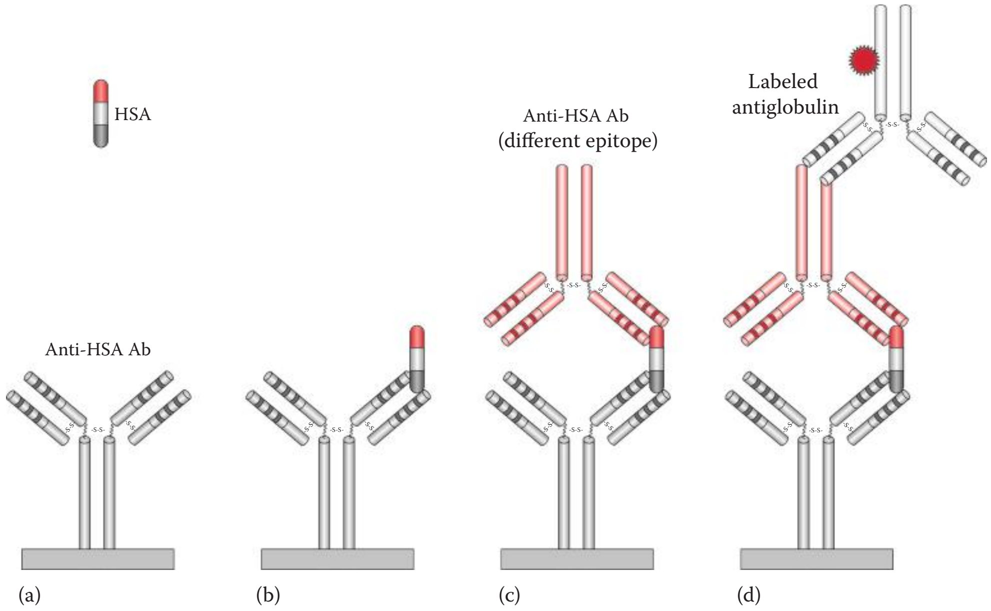

Figure 15.11 ELISA for identification of HSA in saliva. (a) Sample containing antigen is applied to polystyrene tubes where anti-HSA Ab is immobilized. (b) Antigen binds to immobilized Ab to form HSA–Ab complex. (c) Second anti-HSA Ab, specific for a different epitope of HSA, is added to form Ab–HSA–Ab sandwich. (d) Labeled antiglobulin then binds to the sandwich. The bound antiglobulin can be detected by various reporting schemes. Ab, an antibody; HSA, human salivary α-amylase. (© Richard C. Li.)

15.2.2.2 RNA-Based Assays

RNA-based assays (Chapter 11) have been developed recently for the identification of saliva. They are based on the expression of certain genes in certain cell or tissue types. Thus, the techniques used in the identification of saliva are based on the detection of specific types of mRNA expressed exclusively in certain cells in the oral cavity. These assays utilize reverse transcriptase polymerase chain reaction (RT-PCR; see Chapter 7) methods to detect gene expression levels of mRNAs for saliva identification. Table 15.1 summarizes the tissue-specific genes utilized for saliva identification. Compared to conventional assays used for saliva identification, RNA-based assays present higher specificity and are amenable to automation. However, one limitation is that the RNA is unstable because of degradation by endogenous ribonucleases.

Table 15.1 Application of RT-PCR Assay for Saliva Identification

Gene Symbol	Gene Product	Description	Further Reading
HTN3	Histatin 3	Histidine-rich protein involved in nonimmune host defense in oral cavity	Sabatini et al. (1993)
STATH	Statherin	Inhibitor of precipitation of calcium phosphate salts in oral cavity	Sabatini et al. (1990)

Source: Adapted from Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005.

Bibliography

Akutsu, T., et al., Applicability of ELISA detection of statherin for forensic identification of saliva. *Int J Legal Med*, 2010, **124**(5): 493–498.

Ali, M.M., et al., PCR applications in identification of saliva samples exposed to different conditions (streptococci detection based). *Pak J Biol Sci*, 2013, **16**(12): 575–579.

Aps, J.K. and L.C. Martens, Review: The physiology of saliva and transfer of drugs into saliva. *Forensic Sci Int*, 2005, **150**(2–3): 119–131.

Auvdel, M.J., Amylase levels in semen and saliva stains. *J Forensic Sci*, 1986, **31**(2): 426–431.

Auvdel, M.J., Comparison of laser and high-intensity quartz arc tubes in the detection of body secretions. *J Forensic Sci*, 1988, **33**(4): 929–945.

Barni, F., et al., Alpha-amylase kinetic test in bodily single and mixed stains. *J Forensic Sci*, 2006, **51**(6): 1389–1396.

Casey, D.G. and J. Price, The sensitivity and specificity of the RSID-saliva kit for the detection of human salivary amylase in the Forensic Science Laboratory, Dublin, Ireland. *Forensic Sci Int*, 2010, **194**(1–3): 67–71.

Courts, C. and B. Madea, Specific micro-RNA signatures for the detection of saliva and blood in forensic body-fluid identification. *J Forensic Sci*, 2011, **56**(6): 1464–1470.

Culliford, B.J., Precipitin reactions in forensic problems. A new method for precipitin reactions on forensic blood, semen and saliva stains. *Nature*, 1964, **201**: 1092–1093.

DeForest, P., R. Gaensslen, and H.C. Lee, *Forensic Science: An Introduction to Criminalistics*, 1983. New York: McGraw-Hill Book Company.

De Leo, D., et al., A sensitive and simple assay of saliva on stamps. *Z Rechtsmed*, 1985, **95**(1): 27–33.

Eckersall, P.D., et al., The production and evaluation of an antiserum for the detection of human saliva. *J Forensic Sci Soc*, 1981, **21**(4): 293–300.

Gaensslen, R.E., *Sourcebook in Forensic Serology, Immunology, and Biochemistry*, 1983. Washington, DC: US Government Printing Office.

Greenfield, A. and M. Sloan, Identification of biological fluids and stains, in S. James and J. Nordby (eds), *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 2nd edn., 2005. CRC Press.

Haas, C., et al., RNA/DNA co-analysis from human saliva and semen stains—Results of a third collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2013, **7**(2): 230–239.

Hedman, J., et al., Evaluation of amylase testing as a tool for saliva screening of crime scene trace swabs. *Forensic Sci Int Genet*, 2011, **5**(3): 194–198.

Hochmeister, M.N., et al., High levels of alpha-amylase in seminal fluid may represent a simple artifact in the collection process. *J Forensic Sci*, 1997, **42**(3): 535–536.

Hsu, L., et al., Amplification of oral streptococcal DNA from human incisors and bite marks. *Curr Microbiol*, 2012, **65**(2): 207–211.

Jones, E., The identification of semen and other body fluids, in R. Saferstein (ed.), *Forensic Science Handbook*, 2005. New Jersey: Pearson Prentice Hall.

Jones, E.L. Jr. and J.A. Leon, Lugol's test reexamined again: Buccal cells. *J Forensic Sci*, 2004, **49**(1): 64–67.

Juusola, J. and J. Ballantyne, Messenger RNA profiling: A prototype method to supplant conventional methods for body fluid identification. *Forensic Sci Int*, 2003, **135**(2): 85–96.

Juusola, J. and J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**(1): 1–12.

Keating, S.M. and D.F. Higgs, The detection of amylase on swabs from sexual assault cases. *J Forensic Sci Soc*, 1994, **34**(2): 89–93.

Kennedy, D.M., et al., Microbial analysis of bite marks by sequence comparison of streptococcal DNA. *PLoS One*, 2012, **7**(12): e51757.

Lee, H.C., Identification and grouping of bloodstains, in R. Saferstein (ed.), *Forensic Science Handbook*, 1982. Englewood Cliffs, NJ: Prentice Hall.

Martin, N.C., N.J. Clayson, and D.G. Scrimger, The sensitivity and specificity of red-starch paper for the detection of saliva. *Sci Justice*, 2006, **46**(2): 97–105.

Miller, D.W. and J.C. Hodges, Validation of Abacus SALIgAE® test for the forensic identification of saliva, 2005. Available from: http://www.dnalabsinternational.com/SalivaValidation.pdf. Retrieved April 21, 2007.

Myers, J.R. and W.K. Adkins, Comparison of modern techniques for saliva screening. *J Forensic Sci*, 2008, **53**(4): 862–867.

Nakanishi, H., et al., A novel method for the identification of saliva by detecting oral streptococci using PCR. *Forensic Sci Int*, 2009, **183**(1–3): 20–23.

Nakanishi, H., et al., A simple identification method of saliva by detecting Streptococcus salivarius using loop-mediated isothermal amplification. *J Forensic Sci*, 2011, **56**(Suppl 1): S158–S161.

Nussbaumer, C., E. Gharehbaghi-Schnell, and I. Korschineck, Messenger RNA profiling: A novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**(2–3): 181–186.

Old, J.B., et al., Developmental validation of RSID-saliva: A lateral flow immunochromatographic strip test for the forensic detection of saliva. *J Forensic Sci*, 2009, **54**(4): 866–873.

Omelia, E.J., M.L. Uchimoto, and G. Williams, Quantitative PCR analysis of blood- and saliva-specific microRNA markers following solid-phase DNA extraction. *Anal Biochem*, 2013, **435**(2): 120–122.

Pandeshwar, P. and R. Das, Role of oral fluids in DNA investigations. *J Forensic Leg Med*, 2014, **22C**: 45–50.

Pang, B.C. and B.K. Cheung, Applicability of two commercially available kits for forensic identification of saliva stains. *J Forensic Sci*, 2008, **53**(5): 1117–1122.

Power, D.A., et al., PCR-based detection of salivary bacteria as a marker of expirated blood. *Sci Justice*, 2010, **50**(2): 59–63.

Pretty, I.A., The barriers to achieving an evidence base for bitemark analysis. *Forensic Sci Int*, 2006, **159**(Suppl 1): S110–S120.

Quarino, L., et al., Differentiation of α-amylase from various sources: An approach using selective inhibitors. *J Forensic Sci Soc*, 1993, **33**(2): 87–94.

Quarino, L., et al., An ELISA method for the identification of salivary amylase. *J Forensic Sci*, 2005, **50**(4): 873–876.

Ricci, U., I. Carboni, and F. Torricelli, False-positive results with amylase testing of citrus fruits. *J Forensic Sci*, 2014, **59**(5):1410–1412.

Rushton, C., et al., The distribution and significance of amylase-containing stains on clothing. *J Forensic Sci Soc*, 1979, **19**(1): 53–58.

Sabatini, L.M., Y.Z. He, and E.A. Azen, Structure and sequence determination of the gene encoding human salivary statherin. *Gene*, 1990, **89**(2): 245–251.

Sabatini, L.M., T. Ota, and E.A. Azen, Nucleotide sequence analysis of the human salivary protein genes HIS1 and HIS2, and evolution of the STATH/HIS gene family. *Mol Biol Evol*, 1993, **10**(3): 497–511.

Sakurada, K., et al., Evaluation of mRNA-based approach for identification of saliva and semen. *Leg Med* (*Tokyo*), 2009, **11**(3): 125–128.

Searcy, R.L., et al., The interaction of human serum protein fractions with the starch–iodine complex. *Clin Chim Acta*, 1965, **12**(6): 631–638.

Shaler, R.C., Modern forensic biology, in R. Saferstein (ed.), *Forensic Science Handbook*, 2nd edn., 2002. Upper Saddle River, NJ: Pearson Education.

Soukos, N.S., et al., A rapid method to detect dried saliva stains swabbed from human skin using fluorescence spectroscopy. *Forensic Sci Int*, 2000, **114**(3): 133–138.

Sweet, D., et al., An improved method to recover saliva from human skin: The double swab technique. *J Forensic Sci*, 1997, **42**(2): 320–322.

Uchimoto, M.L., et al., Considering the effect of stem-loop reverse transcription and real-time PCR analysis of blood and saliva specific microRNA markers upon mixed body fluid stains. *Forensic Sci Int Genet*, 2013, **7**(4): 418–421.

Vandenberg, N. and R.A. van Oorschot, The use of Polilight in the detection of seminal fluid, saliva, and bloodstains and comparison with conventional chemical-based screening tests. *J Forensic Sci*, 2006, **51**(2): 361–370.

Vennemann, M., et al., Sensitivity and specificity of presumptive tests for blood, saliva and semen. *Forensic Sci Med Pathol*, 2014, **10**(1): 69–75.

Wawryk, J. and M. Odell, Fluorescent identification of biological and other stains on skin by the use of alternative light sources. *J Clin Forensic Med*, 2005, **12**(6): 296–301.

Whitehead, P.H. and A.E. Kipps, A test paper for detecting saliva stains. *J Forensic Sci Soc*, 1975, **15**(1): 39–42.

Whitehead, P.H. and A.E. Kipps, The significance of amylase in forensic investigations of body fluids. *Forensic Sci*, 1975, **6**(3): 137–144.

Willott, G.M., An improved test for the detection of salivary amylase in stains. *J Forensic Sci Soc*, 1974, **14**(4): 341–344.

Willott, G.M. and M. Griffiths, A new method for locating saliva stains: Spotty paper for spotting spit. *Forensic Sci Int*, 1980, **15**(1): 79–83.

Zimmermann, B.G., N.J. Park, and D.T. Wong, Genomic targets in saliva. *Ann N Y Acad Sci*, 2007, **1098**: 184–191.

Zubakov, D., et al., Stable RNA markers for identification of blood and saliva stains revealed from whole genome expression analysis of time-wise degraded samples. *Int J Legal Med*, 2008, **122**(2): 135–142.

Zubakov, D., et al., New markers for old stains: Stable mRNA markers for blood and saliva identification from up to 16-year-old stains. *Int J Legal Med*, 2009, **123**(1): 71–74.

16

Identification of Vaginal Secretions and Menstrual Blood

The identification of vaginal secretions and menstrual blood is important for the investigation of sexual assault cases. Such identification can help to corroborate allegations of sexual assault. For example, in a sexual assault investigation, a stain was observed after examining the suspect's clothing. Subsequently, forensic DNA analysis revealed that the DNA of the stain originated from the victim, establishing a link between the victim and the suspect. However, the defendant may assert that the victim's DNA originated from a sweat stain as a result of casual contact and deny any criminal act. If vaginal secretions were found in the stain, the evidence has probative value to corroborate an allegation of a sexual act. Additionally, postcoital drainage stains on clothing or bedding are often recovered at crime scenes. These stains usually consist of a mixture of semen and vaginal secretions. In such cases, the presence of vaginal secretions in these stains indicates the occurrence of sexual intercourse. Sometimes, vaginal secretions can be transferred onto a perpetrator during a sexual assault. For example, the presence of vaginal secretions on a suspect's genital area can indicate the occurrence of sexual intercourse. Furthermore, the presence of vaginal secretions on an object can corroborate an allegation of vaginal rape with a foreign object.

If a sexual assault victim is in menses when an assault occurred, blood evidence, such as the victim's bloodstains located on the suspect's clothing, may be recovered at the scene. The defense may argue that the bloodstains resulted from an injury and deny that any sexual act occurred. In this case, the identification of menstrual blood would corroborate an alleged rape. Therefore, it is necessary to distinguish between peripheral and menstrual blood in investigations of sexual assault where blood evidence is found at the scene.

16.1 Identification of Vaginal Stratified Squamous Epithelial Cells

A normal human vagina is covered by the *squamous mucosa*, which is composed of stratified squamous epithelial tissue (Figure 16.1a). Lying under the squamous mucosa is the *submucosa*, which contains an abundance of connective tissue and capillaries. Below the submucosa is the *muscularis*, which is made up of smooth muscle.

The squamous mucosa consists of multiple layers of cells (Figure 16.1b). At the *basal layer* of the squamous mucosa, basal cells are anchored to the basement membrane that separates the squamous mucosa from the submucosa. The basal cells are small in size with relatively large nuclei and are highly proliferative. As the cells migrate up from the basal layer to the *parabasal layer*, the cells undergo differentiation. At the *intermediate layer*, the cells are flattened and their nuclei are compressed. As the epithelial cells reach the apical layer, the *superficial layer*, the cells are fully differentiated with small and dense nuclei.

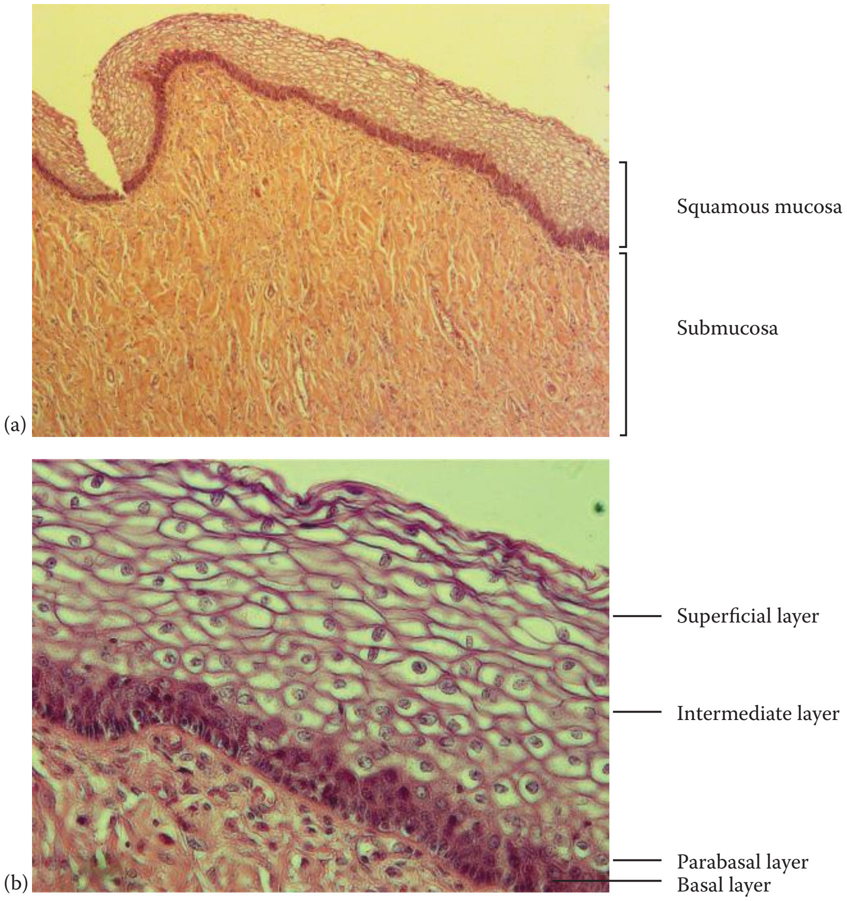

Figure 16.1 Histological structure of human vaginal tissues. (a) The squamous mucosa and the submucosa. (b) Multiple layers of cells in the squamous mucosa. (© Richard C. Li.)

Vaginal epithelial cells do not accumulate keratin, which is different from skin cells (Chapter 4). The intermediate and superficial layer cells contain abundant glycogen in their cytoplasm. The presence of glycogen in these cells is an indication of normal development and the differentiation of the vaginal epithelial cells. The apical surface of the vaginal squamous mucosa is usually covered by mucus that is secreted from glands that are located deep in the epithelium. The cells of the apical layers are eventually sloughed and are continuously replaced by the cells of deeper layers. When the lining of the vagina is swabbed during the collection of evidence, glycogenated cells with small numbers of parabasal cells are usually recovered (Figure 16.2). In addition to the vagina, glycogenated squamous epithelial cells are found in the linings of the oral cavity, pharynx, esophagus, anus, and the apex of the urethra. The differentiation of the vaginal epithelial cells requires estrogen. However, in premenarche and postmenopausal women, estrogen levels are very low; thus, the vaginal epithelial cells only differentiate to the parabasal cells. Primarily, parabasal cells are found in specimens from these individuals.

16.1.1 Lugol's Iodine Staining and Periodic Acid–Schiff Method

Lugol's iodine solution, named after the French physician Jean Lugol, is originally used as an antiseptic that is applied to skin or tissue to prevent infection. In forensic applications, it is utilized for the identification of glycogenated vaginal epithelial cells. The technique is based on the

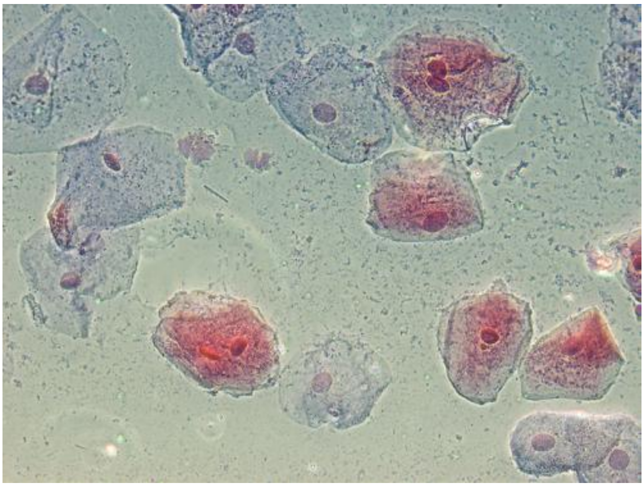

Figure 16.2 Vaginal stratified squamous epithelial cells. (© Richard C. Li.)

principle that iodine reacts with intracellular glycogen to exhibit a color. Glycogen is the principal storage form of glucose in animal and human cells and is found in the granules in the cytoplasm of the cells of many tissues. In addition to squamous epithelial cells, glycogen is also found in hepatocytes, which have the highest glycogen content, as well as muscle cells. Glycogen is a polysaccharide composed of D-glucose units. Similar to the amylopectin in plant cells (Chapter 15), glycogen is a branched polysaccharide consisting of linear and branched chains (Figure 16.3). The D-glucose residues of the linear chain are connected by α1→4 glycosidic bonds, while the branching points are connected by α1→6 glycosidic bonds. However, glycogen is more branched than

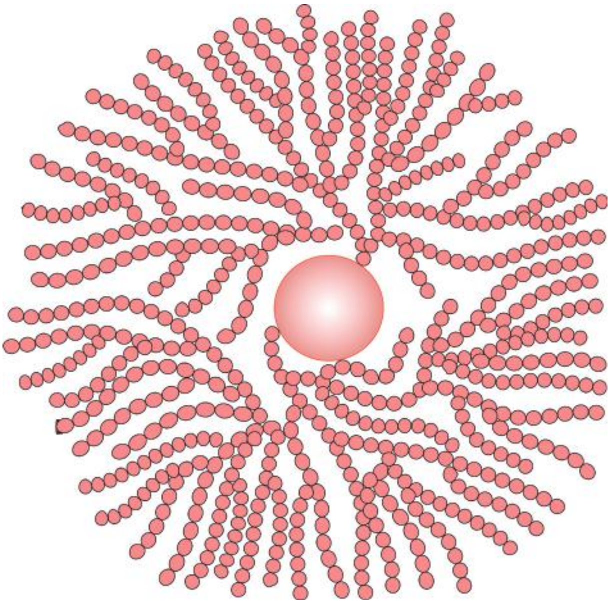

Figure 16.3 Structure of glycogen. Glycogen is a highly branched polysaccharide composed of glucose. Glycogen has a similar structure to amylopectin in starch (see Chapter 15). However, glycogen is more branched than starch. Additionally, glycogen contains a protein known as glycogenin at the center of its structure. (© Richard C. Li.)

amylopectin. Branch points occur approximately every 10 glucose residues in glycogen while they occur every 25 glucose residues in amylopectin. Lugol's stock solution is an aqueous solution of 5% iodine (I_2) and 10% potassium iodide (KI). Potassium iodide allows the iodine to be soluble in water through the formation of the triiodide ion. For staining the vaginal epithelial cells, 5% of the stock solution is usually used as the working solution. Iodine atoms fit into the helices of glycogen to form a dark brown glycogen–iodine complex (Figures 16.4 and 16.5). Vaginal epithelial cells can also be stained using the periodic acid–Schiff method. The cytoplasm of the vaginal epithelial cells is stained magenta, and the nucleus is stained purple.

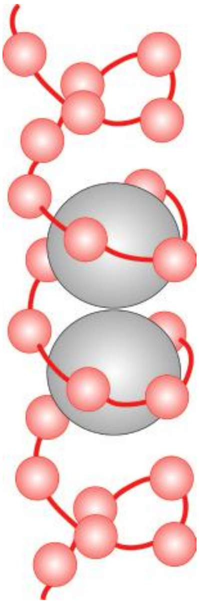

Figure 16.4 Diagram of a glycogen–iodine complex. In the complex, the glycogen chain forms a helix structure with six monosaccharide residues (red) per turn. Iodine molecules (gray) fit in the helix to form the glycogen–iodine complex, exhibiting a color. (© Richard C. Li.)

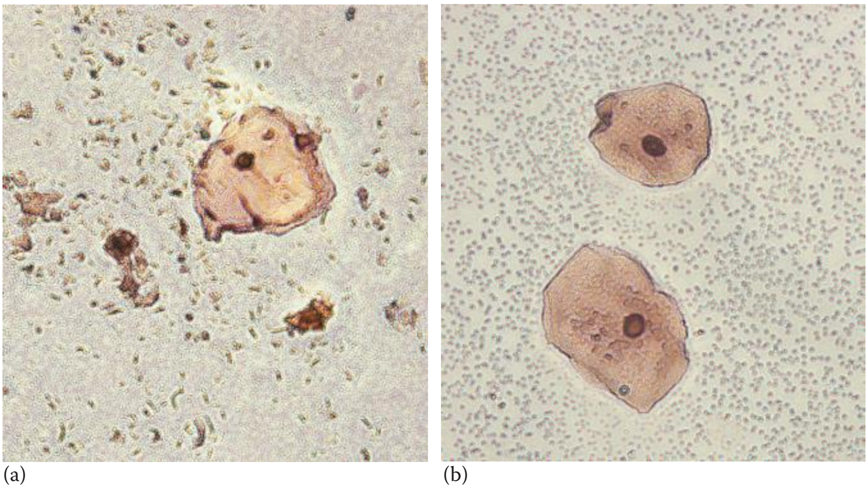

(a) (b)

Figure 16.5 Epithelial cells stained with Lugol's iodine solution. (a) Vaginal and (b) buccal epithelial cells. (© Richard C. Li.)

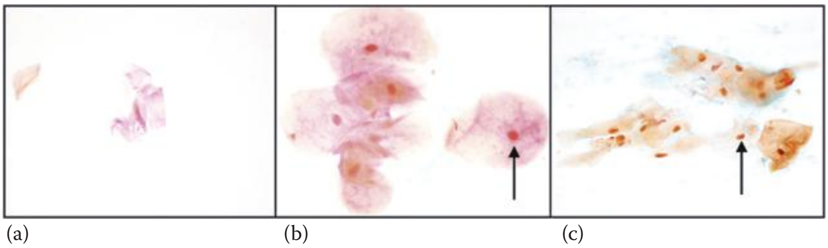
(a) (b) (c)

Figure 16.6 Staining of epithelial cells with Dane's method. The results obtained using methanol fixation followed by Dane's staining on skin (a), buccal (b), and vaginal epithelial (c) cells are shown. The arrows indicate red nuclei in buccal and orange nuclei in vaginal cells. (From French, C.E., et al., *Forensic Sci Int*, 178, 1–6, 2008. With permission.)

16.1.2 Dane's Staining Method

Evidence containing skin, buccal, and vaginal epithelial cells is often recovered in forensic investigations, particularly in sexual assault cases. Thus, it is necessary to distinguish between these cells. Skin, buccal, and vaginal epithelial cells belong to stratified squamous epithelium. Differentiated skin epithelial cells are keratinized and are classified as keratinizing squamous epithelial cells, while buccal and vaginal epithelial cells are nonkeratinizing squamous epithelial cells. Additionally, the skin epithelial cells lose nuclei and other cellular organelles during differentiation. In contrast, buccal and vaginal cells contain nuclei. Based on their morphology, it is possible to distinguish skin epithelial cells from buccal and vaginal epithelial cells. However, buccal and vaginal cells are morphologically indistinguishable from each other. Although Lugol's iodine solution and periodic acid–Schiff can stain vaginal epithelial cells, these stains cannot distinguish vaginal from other glycogenated epithelial cells in the oral mucosa and the urinary tract. Recently, Dane's staining method has been developed to distinguish all three types of cells. Skin cells are stained red and orange; buccal cells are stained predominantly orange-pink with red nuclei; and vaginal cells are stained bright orange with orange nuclei (Figure 16.6).

16.2 Identification of Vaginal Acid Phosphatase

Acid phosphatases are a group of enzymes that are capable of hydrolyzing a variety of small organic phosphomonoesters under acidic conditions. To date, at least five different acid phosphatase isoenzymes have been identified in human tissues: erythroid acid phosphatase (encoded by the *ACP1* gene), lysosomal acid phosphatase (encoded by the *ACP2* gene), prostate acid phosphatase (encoded by the *ACPP* gene, also known as *ACP3*), macrophage acid phosphatase (encoded by the *ACP5* gene), and testicular acid phosphatase (encoded by the *ACPT* gene). Human prostatic acid phosphatase is found in large quantities in seminal fluid and is used as a biomarker for semen identification (Chapter 14). The prostatic acid phosphatase is a homodimer containing two identical subunits with a molecular weight of 50 kDa. Small amounts of acid phosphatase can be detected in vaginal fluid, which is produced in normal cervical epithelial cells. However, the molecular characteristics of vaginal acid phosphatase are still not known. Historically, vaginal acid phosphatase has been used as a biomarker for the identification of vaginal secretions using acid phosphatase catalytic assays (Chapter 14). In sexual assault investigations, it is important to distinguish vaginal acid phosphatase from prostate acid phosphatase originating from semen exposure. These two enzymes have identical molecular weights, enzymatic specificities, and responses to the same inhibitors. Nevertheless, vaginal and prostate acid phosphatases can be distinguished using agarose electrophoresis. Based on their electrophoretic mobility, bands of vaginal and prostate acid phosphatases can be separated. The prostate acid phosphatase

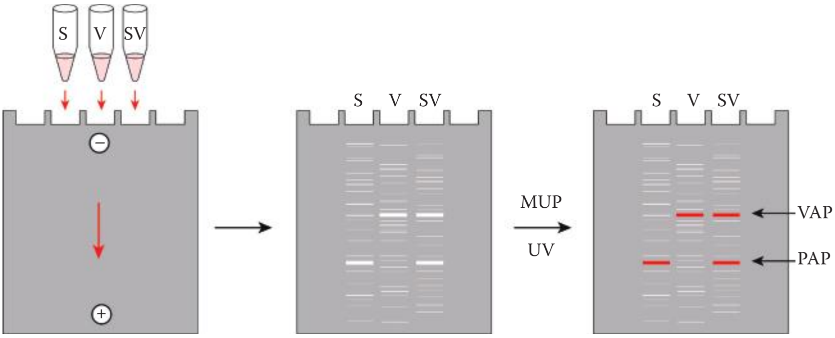

Figure 16.7 Distinguishing prostate and vaginal acid phosphatases using polyacrylamide gel electrophoresis. Bands of acid phosphatases are detected using 4-methylumbelliferyl phosphate (MUP). MUP is a substrate of acid phosphatases, resulting in a product that fluoresces in UV illumination (see Chapter 14). The prostate acid phosphatase has higher electrophoretic mobility toward an anode (a positively charged electrode) than vaginal acid phosphatase. S, semen; V, vaginal secretions; SV, mixture of semen and vaginal secretions; PAP, prostate acid phosphatase; VAP, vaginal acid phosphatase; MUP, 4-methylumbelliferyl phosphate; UV, ultraviolet illumination. (© Richard C. Li.)

has higher electrophoretic mobility toward an anode (a positively charged electrode) than the vaginal acid phosphatase (Figure 16.7). Thus, the presence of vaginal acid phosphatase can be determined.

16.3 Identification of Vaginal Bacteria

Lactobacillus can be found in the respiratory, the gastrointestinal, and the urogenital tract of healthy humans and animals. *Lactobacillus* taxa are the predominant bacteria in the vagina of women of reproductive age, and they play an important role in protecting the host against invasive pathogenic organisms. *Lactobacillus* consists of rod-shaped, nonmotile, and non-spore-forming gram-positive bacteria (Figure 16.8). Since these bacteria survive on carbohydrates, *Lactobacillus* bacteria produce lactic acid. As a result, a low pH environment is established in

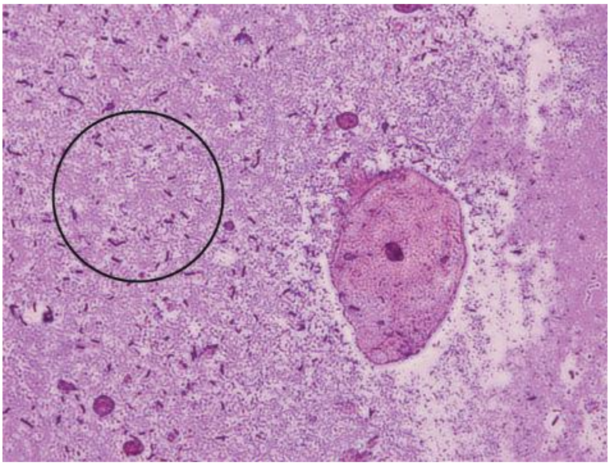

Figure 16.8 Vaginal *Lactobacillus* bacteria (circled). (© Richard C. Li.)

the lumen of the vagina, which restricts the growth of pathogenic organisms. Such vaginal lactic acid–producing bacteria were thought to be *Lactobacillus acidophilus*. In the 1980s, it was determined that *L. acidophilus* was not a single species, but actually a group of related species known as the *L. acidophilus* complex. The species of the complex can now be distinguished based on their DNA sequences. The most frequently occurring *Lactobacillus* species found in the vagina are *L. iners*, *L. crispatus*, *L. gasseri*, and *L. jensenii*. Among them, *L. iners* is the most common species of *Lactobacillus* in women. *Lactobacillus* can be identified based on the sequences of DNA markers such as the 16S rRNA gene and the intergenic spacer region between the 16S rRNA and 23S rRNA genes (Chapter 11). In forensic applications, *L. iners*, *L. crispatus*, *L. gasseri*, and *L. jensenii* can be detected in vaginal secretions. Thus, the identification of *Lactobacillus* taxa can potentially be utilized for the forensic identification of vaginal fluid. However, the presence of *Lactobacillus* taxa is not specific enough to vaginal fluid. Some studies show that *Lactobacillus* taxa are present in semen while others show that *Lactobacillus* can be found in female urine. The openings of the female urogenital system are in close proximity; thus, it is possible to have partially overlapping microbiota between the urine and vaginal secretions. Thus, the identification of multiple bacterial species, at least the four *Lactobacillus* species previously mentioned, should be carried out to distinguish these samples.

16.4 Outlook for Confirmatory Assays of Vaginal Secretions

The identification of the vaginal stratified squamous epithelial cells provides important probative evidence in forensic investigations. However, the existing methods described earlier can sometimes give false-negative or false-positive results. In some situations, these assays also cross-react with other types of bodily fluids. Thus, none of these assays is confirmatory. A useful identification method for vaginal secretions should be able to distinguish vaginal secretions from other bodily fluids and should be easy to perform. For example, nondestructive confirmatory identification methods such as fluorescence spectroscopy and Raman spectroscopy (Chapter 11) can potentially be useful for the identification of vaginal secretions. Recently, the analysis of tissue-specific gene expression has been utilized for the identification of vaginal secretions. Using the reverse transcription polymerase chain reaction (RT-PCR) technique (Chapter 7), the mRNAs of the tissue-specific genes of vaginal epithelial cells can be detected. For example, two commonly used markers for vaginal secretion identification are *MUC4* and *HBD1*. *MUC4* encodes a mucin protein that is a major component of vaginal mucus, and *HBD1*, the human β defensin 1, encodes a vaginal antimicrobial peptide. Both *MUC4* and *HBD1* are expressed in vaginal epithelial cells and are considered reliable markers of vaginal fluid. Additional mRNA and miRNA markers are described in Tables 16.1 and 16.2.

16.5 Menstruation

Menstruation is the periodic discharge of blood and the elimination of the degenerated lining of the endometrium from the uterus of nonpregnant women. From menarche to menopause, women may menstruate up to 400 times during their reproductive age. The uterus plays an important role in preparing the uterine endometrium for the possible implantation of a developing embryo. The linings of the uterus are composed of the myometrium and the endometrium. The *myometrium* consists of the muscle fibers of the uterus. The *endometrium* consists of the simple columnar epithelium and the stroma (Figure 16.9). The simple columnar epithelium is formed by single-layered elongated cells located at the apical surface of the endometrium. The stroma consists of connective tissues as well as spiral arteries. Spiral arteries are small arteries that ascend through the endometrium and form a coil-like structure, which supplies blood to the endometrium.

Table 16.1 Representative Markers of mRNA-Based Assays for Vaginal Secretions and Menstrual Blood Identification

Biological Fluid	Gene Symbol	Description
Vaginal secretions	*CYP2B7P1*	Cytochrome P450, family 2, subfamily B, polypeptide 7, pseudogene 1
	DKK4	Dickkopf homolog 4
	FUT6	Fucosyltransferase 6
	HBD1	β Defensin 1
	IL19	Interleukin 19
	MUC4	Mucin 4
	MYOZ1	Myozenin 1
	SFTA2	Surfactant associated 2
Menstrual blood	*MMP7*	Matrix metalloproteinase 7
	MMP11	Matrix metalloproteinase 11

Source: Adapted from Bauer, M. and Patzelt, D., *J Forensic Sci*, 47, 1278–1282, 2002; Hanson, E.K. and Ballantyne, J., *Sci Justice*, 53, 14–22, 2013; Juusola, J. and Ballantyne, J., *Forensic Sci Int*, 152, 1–12, 2005; Nussbaumer, C., Gharehbaghi-Schnell, E., and Korschineck, I., *Forensic Sci Int*, 157, 181–186, 2006.

Table 16.2 miRNA Markers for Vaginal Secretions and Menstrual Blood Identification

Bodily Fluid	miRNA Marker	Sequence
Vaginal secretions	miR124*a*	UAAGGCACGCGGUGAAUGCC
	miR372	AAAGUGCUGCGACAUUUGAGCGU
	miR617	AGACUUCCCAUUUGAAGGUGGC
	miR891a	UGCAACGAACCUGAGCCACUGA
Menstrual blood	miR214	UGCCUGUCUACACUUGCUGUGC
	miR412	ACUUCACCUGGUCCACUAGCCGU
	miR451	AAACCGUUACCAUUACUGAGUU

Source: Adapted from Hanson, E.K., Lubenow, H., and Ballantyne, J., *Anal Biochem*, 387, 303–314, 2009; Wang, Z., et al., *Forensic Sci Int Genet*, 7, 116–123, 2013.

16.5.1 Uterine Cycle

The endometrium can be divided into two zones: the functionalis and the basalis (Figure 16.10). The *functionalis* is the luminal part of the endometrium. It is the zone of cyclic changes in the endometrium and is shed during menstruation. The *basalis* is the basal part of the endometrium and is not shed during menstruation. This zone produces cells to regenerate the functionalis during the next cycle.

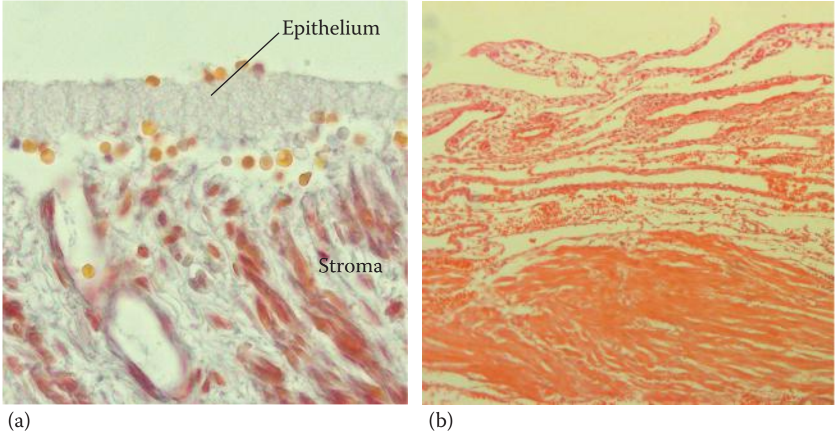
Epithelium
Stroma
(a)
(b)

Figure 16.9 Uterus endometrium. (a) The proliferative phase and (b) the menstrual phase. (© Richard C. Li.)

During the uterine cycle, repetitive physiological changes occur in the functionalis (Figure 16.10). The cycle is divided into three phases: the menstrual, proliferative, and secretory phases. The first day of menstrual bleeding is considered the onset of the menstrual phase. During the menstrual phase, the functionalis degenerates and is sloughed off from the uterine wall and bleeding occurs, known as menses. During the proliferative phase, the functionalis begins regeneration in which the spiral arteries are proliferated. During the secretory phase, the spiral arteries are further developed and coiled. In the absence of pregnancy, a decrease in the progesterone level leads to the constriction of the spiral arteries. As a result, the functionalis becomes ischemic (insufficient blood flow), leading to hypoxia (low levels of oxygen in tissue). In addition, the activation of an enzymatic degradation process causes the destruction and

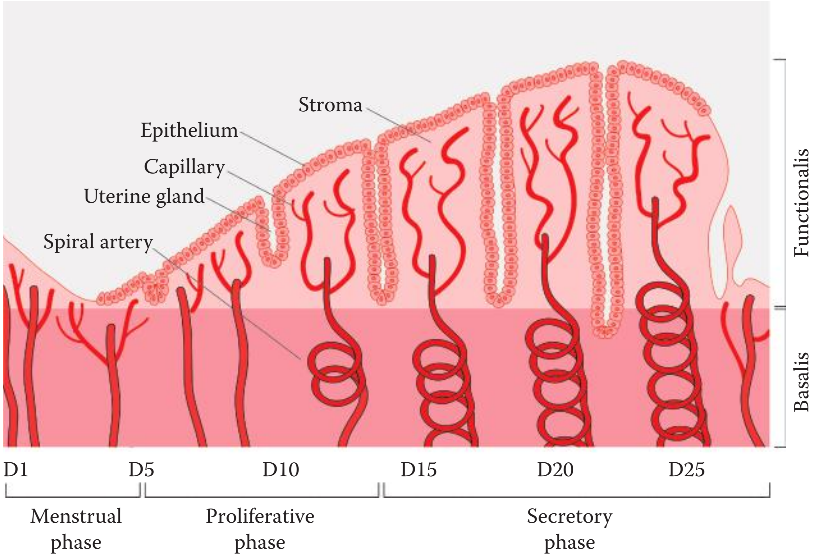
Stroma
Epithelium
Capillary
Uterine gland
Spiral artery
Functionalis
Basalis
D1
D5
D10
D15
D20
D25
Menstrual phase
Proliferative phase
Secretory phase

Figure 16.10 Changes in the functionalis of the uterine mucosa during a uterine cycle. D, day. (© Richard C. Li.)

shedding of the functionalis during menses. Menstrual fluid contains blood, the functionalis layer tissue, and mucus.

16.5.2 Uterine Endometrial Hemostasis

The cessation of menstrual bleeding is achieved by endometrial *hemostasis* that is initiated when injury occurs due to the shedding of the endometrium (Figure 16.11). Hemostasis begins with platelet activation and aggregation to form platelet plugs at the site of injury. Additionally, the blood coagulation cascade is activated to produce *thrombin*. Thrombin, a serine protease, converts soluble fibrinogen into fibrin. Fibrin, a protein involved in blood clotting, aggregates with the platelet plugs and leads to the cessation of bleeding by forming blood clots, known as *thrombi*. Under normal physiological conditions, uterine endometrial hemostasis is a balanced process between blood coagulation and clot dissolution to control blood loss and to prevent clot accumulation within the uterus. As a result, the balance of these two processes allows the removal of tissue fragments from the uterus cavity in order to reduce the risks of infection. Blood clots are prevented from accumulating during menstruation by forming low amounts of platelet plugs and synthesizing coagulation factor inhibitors that inhibit blood coagulation. Additionally, *fibrinolysis* is activated, during which thrombus is broken down by a protease known as *plasmin*. Plasmin cleaves fibrin, generating soluble degradation products. As a result, fibrinolysis can inhibit blood clot formation.

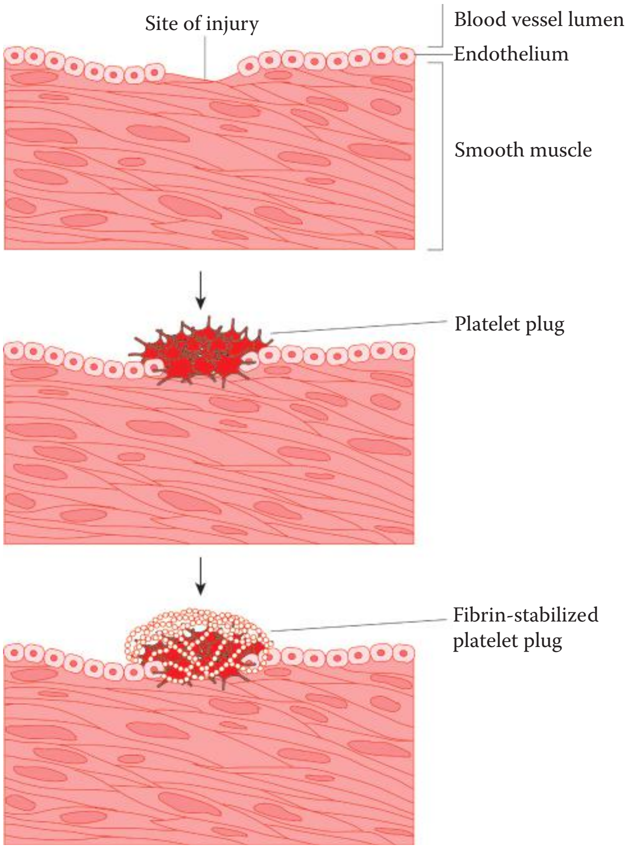

Figure 16.11 Diagram of hemostasis. Hemostasis occurs at the site of injury of a blood vessel. A platelet plug is formed as a result of the aggregation of platelets. The plug is further stabilized by the formation of a fibrin clot over the platelet plug. (© Richard C. Li.)

16.6 D-dimer Assay

During fibrinolysis, cross-linked fibrin is cleaved by plasmin, producing a degradation product known as *D-dimer* (Figure 16.12). Assays for D-dimer fragments have been utilized in the forensic identification of menstrual blood. A number of different formats of D-dimer assays can be used. In an enzyme-linked immunosorbent assay (ELISA; Chapter 11), antibodies bind to the D-dimer antigens on the solid phase. The D-dimer–antibody complex is subsequently analyzed using an antibody-based detection system. This method is highly sensitive, but is time-consuming. Latex agglutination assays are based on the interaction of antibodies and D-dimers that are located on carriers to form aggregates during the agglutination process (Chapter 11). However, the magnitude of the agglutination response is manually read and conclusions are based on subjective judgments. Immunochromatographic assays (Chapter 11) utilize monoclonal antibodies specific to D-dimers, which have been developed recently (Figures 16.13 and 16.14). This immunochromatographic assay is very specific, sensitive, and rapid, and can be completed within minutes. The immunochromatographic devices are portable, and thus can potentially be used at crime scenes. The D-dimer assays can positively identify menstrual blood samples. Although peripheral blood contains low levels of D-dimer, these assays do not show positive reactions with peripheral blood. Thus, menstrual blood can be distinguished from peripheral blood using D-dimer assays. However, postmortem blood also contains these D-dimers, which are detected by these assays. Although the detection of postmortem blood would complicate the interpretation of results, postmortem blood is not often encountered in sexual assault cases.

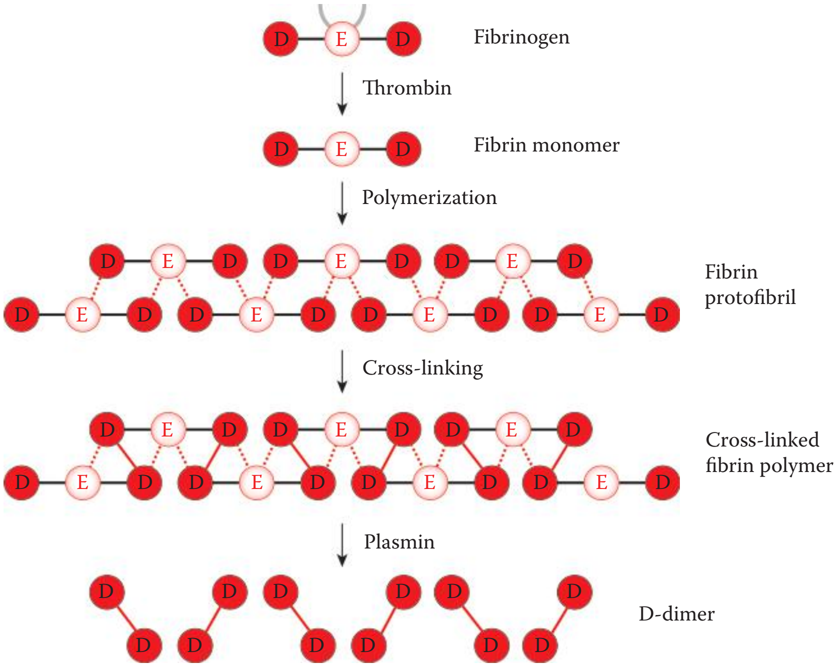

Figure 16.12 The formation and the degradation of a fibrin polymer. The activation of coagulation ultimately generates thrombin, which catalyzes the conversion of fibrinogen to fibrin by cleaving the fibrinopeptides (gray). A fibrin monomer contains an E domain and two D domains. The fibrin monomers are held together by noncovalent bonds (dotted red) between the D domains and E domain to form a fibrin polymer. Fibrin polymers are then covalently linked (solid red) to form a cross-linked fibrin polymer, which plays a role in forming clots. During fibrinolysis, plasmin cleaves the cross-linked fibrin at multiple sites giving rise to fibrin degradation products including D-dimer. (© Richard C. Li.)

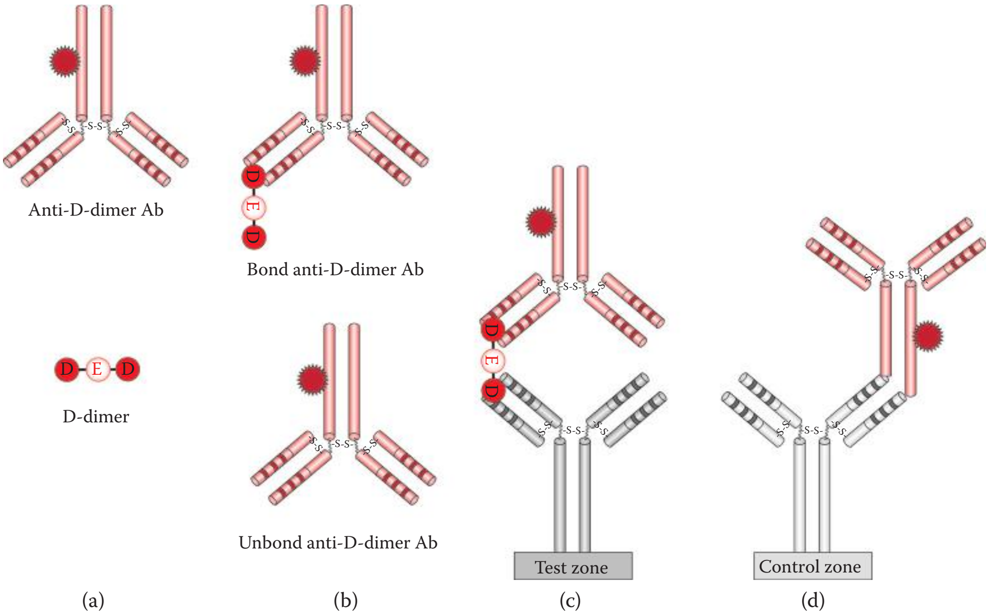
Anti-D-dimer Ab
Bond anti-D-dimer Ab
D-dimer
Unbond anti-D-dimer Ab
D
E
D
-S-S-
Test zone
Control zone
(a)
(b)
(c)
(d)

Figure 16.13 Immunochromatographic assays for the identification of D-dimer in menstrual blood. (a) In a sample well, D-dimer antigen in a menstrual blood sample is mixed with labeled anti-D-dimer Ab. (b) The D-dimer binds to the labeled anti-D-dimer Ab to form a labeled Ab–antigen complex. (c) At the test zone, the labeled complex binds to an immobilized anti-D-dimer Ab to form a labeled Ab–antigen–Ab sandwich. (d) At the control zone, the labeled anti-D-dimer Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, antibody. (© Richard C. Li.)

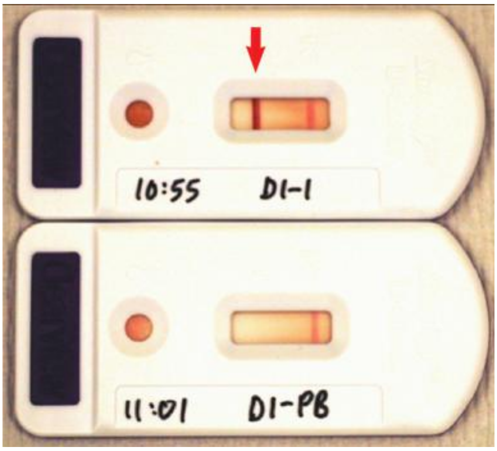
10:55
D1-1
11:01
D1-PB

Figure 16.14 Detecting D-dimer using immunochromatographic assays. The results obtained using the Clearview Simplify D-dimer devices (Alere, Cheshire, UK) on menstrual (top) and peripheral blood (bottom) are shown. A positive result (arrow) indicates the presence of D-dimer. (From Baker, D.J., Grimes, E.A., and Hopwood, A.J., *Forensic Sci Int*, 212, 210–214, 2011. With permission.)

16.7 Lactate Dehydrogenase Assay

Lactate dehydrogenase (LDH) is an enzyme that plays an important role in glycolysis. In humans, LDH catalyzes the reversible reduction of pyruvate into lactate when the amount of oxygen is limited. LDH is a tetrameric enzyme consisting of three different types of subunits. The A subunit, also known as the M subunit, is encoded by the *LDHA* gene and is primarily expressed in skeletal muscle. The B subunit, also known as the H subunit, is encoded by the *LDHB* gene and is primarily expressed in cardiac muscle. The C subunit, encoded by the *LDHC* gene, is expressed restrictively in the testes. LDHs are found in various human tissues. Five isoenzymes can be found in blood (Figure 16.15). LDH1 consists of four identical B subunits; LDH2 consists of one A and three B subunits; LDH3 consists of two A and two B subunits; LDH4 consists of three A and one B subunits; and LDH5 consists of four identical A subunits. The five isoenzymes can be separated using electrophoresis (Figure 16.16) and detected using a colorimetric assay (Figures 16.17 and 16.18). According to their electrophoretic mobility, five bands can be identified. LDH1 has the highest electrophoretic mobility (toward an anode that is a positively charged electrode) and LDH5 has the lowest electrophoretic mobility. In peripheral blood, LDH1, 2, and 3 are the predominant forms of the isoenzymes and LDH4 and 5 are the minor forms of the isoenzymes. In contrast, LDH4 and 5 are consistently the predominant isoenzymes in menstrual blood, while the amounts of LDH1, LDH2, and LDH3 vary. Thus, menstrual blood can be distinguished from peripheral blood. LDH was historically used as a marker for the forensic identification of menstrual blood.

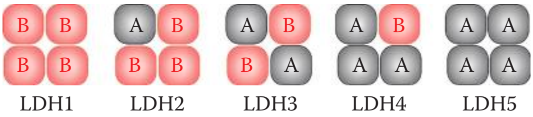

Figure 16.15 Isozymes of lactate dehydrogenase. Human lactate dehydrogenases are composed of four subunits. Five types of isozymes with their subunits are shown. A, LDH A subunit; B, LDH B subunit.

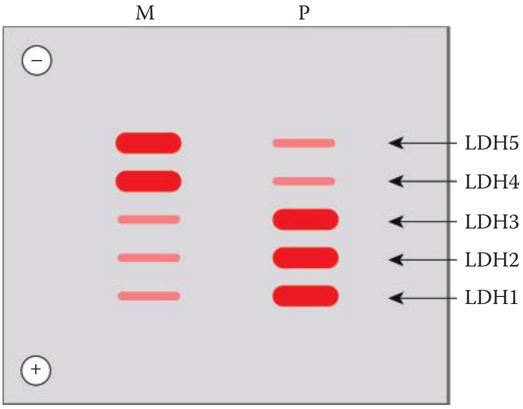

Figure 16.16 Electrophoretic separation of lactate dehydrogenase (LDH) isozymes. Extracted blood samples are loaded onto a cellulose acetate membrane. Electrophoresis is then carried out. Separated bands of LDH isozymes are detected using a colorimetric assay (see Figure 16.15). M, menstrual blood; P, peripheral blood. (© Richard C. Li.)

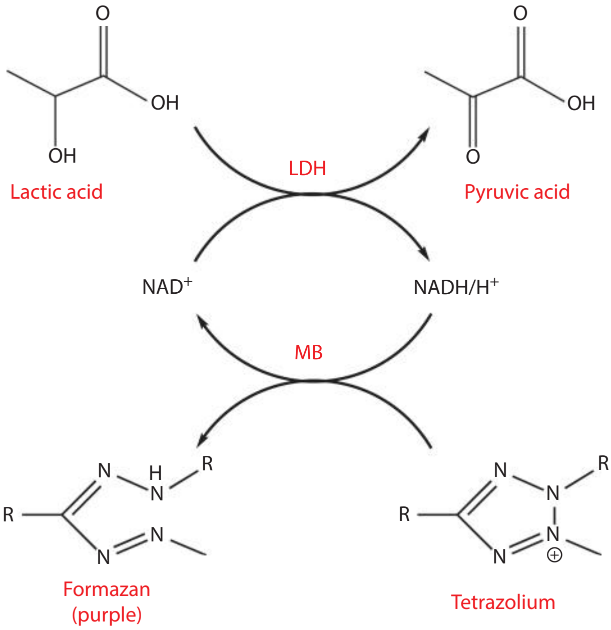

Figure 16.17 Detecting LDH activity using a colorimetric assay forming a colored dye. At the start of a cascade reaction, the LDH catalyzes the conversion of lactic acid to pyruvic acid. Nicotinamide adenine dinucleotide (NAD^+) is then reduced to $NADH/H^+$. Subsequently, Meldola's Blue (MB; 8-dimethylamino-2-benzophenoxazine), an electron carrier, transfers H/H^+ from $NADH/H^+$ to a tetrazolium salt. The pale yellow tetrazolium salt, nitroblue tetrazolium (NBT), is reduced to form a purple formazan dye.

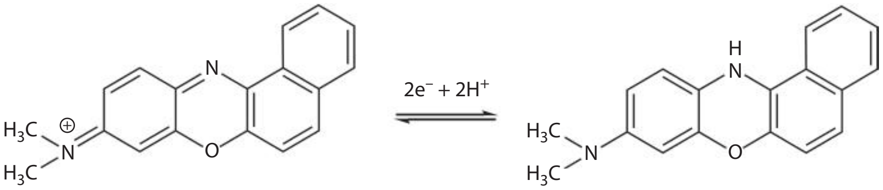

Figure 16.18 The reduction and oxidation of Meldola's Blue. Meldola's Blue is an electron carrier, which transfers hydrogen atoms (coupled to electron transfer).

16.8 RNA-Based Assays

Matrix metalloproteinase (MMP) genes are considered tissue-specific markers for human endometrium tissues. MMPs are zinc-dependent endopeptidases that degrade extracellular matrix components. Additionally, they cleave other proteins such as cytokines, chemokines, and growth factors. The *extracellular matrix* (ECM) is the extracellular space of tissue that is filled by macromolecules such as collagens, laminins, fibronectins, and proteoglycans. The ECM can be divided into two categories: the interstitial matrix and the basement membrane. The interstitial matrices are located in the intercellular spaces. The basement membranes are thin layers of macromolecule fibers that usually lie under the epithelium and the endothelium. Both the interstitial matrices and the basement membranes provide structural support to the cells. It is proposed that MMPs play an important role in the degradation of ECM, leading to the destruction of endometrium tissues during the uterine cycle. To date, a total of 23 MMPs, divided into five subgroups, have been found in humans. The most commonly used markers for the forensic identification of menstrual blood are *MMP7* and *MMP11* (Table 16.1).

Many MMPs are expressed in human endometrium throughout the uterine cycle. The patterns of the MMP gene expressions are correlated with their functions in endometrium tissue breakdown during menstruation. *MMP7* is predominantly expressed in epithelial cells, while *MMP11* is expressed in the stromal cells of the endometrium. Both *MMP7* and *MMP11* mRNA expressions are elevated at the menstrual phase and remain at high levels during the proliferative phase. It is also known that MMPs' mRNA may be elevated in postpartum, wound healing, and metastatic cancer conditions, which may potentially lead to a false-positive identification of menstrual blood.

In menstrual blood samples, among all of the MMP genes tested, *MMP11* is the most sensitive and specific marker for distinguishing menstrual blood from peripheral blood. Using the RT-PCR technique (Chapter 7), *MMP11* mRNA can be detected in menstrual blood from the first to the eighth day of menstruation but it is absent in peripheral blood and vaginal secretions. Therefore, *MMP11* can be used as a marker for the identification of menstrual blood. Likewise, *MMP7* is also a useful marker for menstrual blood identification. However, the *MMP7* mRNA level in menstrual blood is less than that of *MMP11*. Additional markers such as miRNA markers for menstrual blood identification are included in Table 16.2.

Bibliography

Ablett, P.J., The identification of the precise conditions for seminal acid phosphatase (SAP) and vaginal acid phosphatase (VAP) separation by isoelectric focusing patterns. *J Forensic Sci Soc*, 1983, **23**(3): 255–256.

Adams, E.G. and Wraxall, B.G., Phosphatases in body fluids: The differentiation of semen and vaginal secretion. *Forensic Sci*, 1974, **3**: 57–62.

Akutsu, T., et al., Evaluation of latex agglutination tests for fibrin-fibrinogen degradation products in the forensic identification of menstrual blood. *Leg Med (Tokyo)*, 2012, **14**(1): 51–54.

Akutsu, T., et al., Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. *Leg Med (Tokyo)*, 2012, **14**(3): 160–162.

Albrecht, K., et al., Immunohistochemical distribution of cyclic nucleotide phosphodiesterase (PDE) isoenzymes in the human vagina: A potential forensic value? *J Forensic Leg Med*, 2007, **14**(5): 270–274.

Alsawaf, K. and A.T. Tu, Isotachophoretic analysis of bloodstains: Differentiation of human, menstrual, bovine, and ovine bloods. *J Forensic Sci*, 1985, **30**(3): 922–930.

Baker, D.J., E.A. Grimes, and A.J. Hopwood, D-dimer assays for the identification of menstrual blood. *Forensic Sci Int*, 2011, **212**(1–3): 210–214.

Bauer, M. and D. Patzelt, Evaluation of mRNA markers for the identification of menstrual blood. *J Forensic Sci*, 2002, **47**(6): 1278–1282.

Bauer, M. and D. Patzelt, Identification of menstrual blood by real time RT-PCR: Technical improvements and the practical value of negative test results. *Forensic Sci Int*, 2008, **174**(1): 55–59.

Benschop, C.C., et al., Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? *Int J Legal Med*, 2012, **126**(2): 303–310.

Divall, G.B. and M. Ismail, Lactate dehydrogenase isozymes in vaginal swab extracts: A problem for the identification of menstrual blood. *Forensic Sci Int*, 1983, **21**(2): 139–147.

Ferri, G., et al., Successful identification of two years old menstrual bloodstain by using MMP-11 shorter amplicons. *J Forensic Sci*, 2004, **49**(6): 1387.

Fleming, R.I. and S. Harbison, The use of bacteria for the identification of vaginal secretions. *Forensic Sci Int Genet*, 2010, **4**(5): 311–315.

French, C.E., et al., A novel histological technique for distinguishing between epithelial cells in forensic casework. *Forensic Sci Int*, 2008, **178**(1): 1–6.

Fujita, Y., I. Tokunaga, and S.I. Kubo, Forensic identification of a vaginal fluid and saliva mixture through DNA analysis. *Acta Criminologiae et Medicinae Legalis Japonica*, 2003, **69**(2): 48–52.

Giampaoli, S., et al., Molecular identification of vaginal fluid by microbial signature. *Forensic Sci Int Genet*, 2012, **6**(5): 559–564.

Gray, D., N. Frascione, and B. Daniel, Development of an immunoassay for the differentiation of menstrual blood from peripheral blood. *Forensic Sci Int*, 2012, **220**(1–3): 12–18.

Hanson, E.K. and J. Ballantyne, Highly specific mRNA biomarkers for the identification of vaginal secretions in sexual assault investigations. *Sci Justice*, 2013, **53**(1): 14–22.

Hanson, E.K., Lubenow, H., and Ballantyne, J., Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Anal Biochem*, 2009, **387**: 303–314.

Hausmann, R. and B. Schellmann, Forensic value of the Lugol's staining method: Further studies on glycogenated epithelium in the male urinary tract. *Int J Legal Med*, 1994, **107**(3): 147–151.

Jakubowska, J., et al., mRNA profiling for vaginal fluid and menstrual blood identification. *Forensic Sci Int Genet*, 2013, **7**(2): 272–278.

Jones, E., The identification of semen and other body fluids, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 329–399, 2005. Upper Saddle River, NJ: Pearson Prentice Hall.

Juusola, J. and J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**(1): 1–12.

Kaur, G. and V.K. Sharma, The determination of esterase D in menstrual blood stains. *Adli Tip Dergisi*, 1988, **4**(1–2): 3–9.

Metropolitan Police Forensic Science Laboratory, *Biology Methods Manual*, p. 5, 1978. London: Metropolitan Police Forensic Science Laboratory.

Miyaishi, S., et al., Identification of menstrual blood by the simultaneous determination of FDP-D-dimer and myoglobin contents. *Nihon Hoigaku Zasshi*, 1996, **50**(6): 400–403.

Nussbaumer, C., E. Gharehbaghi-Schnell, and I. Korschineck, Messenger RNA profiling: A novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**(2–3): 181–186.

Omelia, E.J., M.L. Uchimoto, and G. Williams, Quantitative PCR analysis of blood- and saliva-specific microRNA markers following solid-phase DNA extraction. *Anal Biochem*, 2013, **435**(2): 120–122.

Peabody, A.J., R.M. Burgess, and R.E. Stockdale, Re-examination of the Lugol's iodine test, pp. 1–18, 1981. Aldermaston: Home Office Central Research Establishment.

Peri, S. and P.K. Chattopadhyay, Differentiation of menstrual blood from other blood stains. *J Forensic Med Toxicol*, 2000, **17**(1): 6–11.

Peri, S. and P.K. Chattopadhyay, Studies on menstrual bloodstains: For species origin and blood groups. *Int J Med Toxicol Leg Med*, 2000, **2**(2): 1–8.

Plac-Bobula, E. and B. Turowska, Identification of menstrual bloodstains by determination of lactate dehydrogenase isoenzymes. *Arch Med Sadowej Kryminol*, 1979, **29**(4): 263–267.

Randall, B., Glycogenated squamous epithelial cells as a marker of foreign body penetration in sexual assault. *J Forensic Sci*, 1988, **33**(2): 511–514.

Sakurada, K., et al., Expression of statherin mRNA and protein in nasal and vaginal secretions. *Leg Med (Tokyo)*, 2011, **13**(6): 309–313.

Sikirzhytskaya, A., V. Sikirzhytski, and I.K. Lednev, Raman spectroscopic signature of vaginal fluid and its potential application in forensic body fluid identification. *Forensic Sci Int*, 2012, **216**(1–3): 44–48.

Simons, J.L. and S.K. Vintiner, Efficacy of several candidate protein biomarkers in the differentiation of vaginal from buccal epithelial cells. *J Forensic Sci*, 2012, **57**(6): 1585–1590.

Stombaugh Jr, P.M. and J.J. Kearney, Factors affecting the use of lactate dehydrogenase as a means of bloodstain differentiation. *J Forensic Sci*, 1978, **23**(1): 94–105.

Takasaka, T., et al., Trials of the detection of semen and vaginal fluid RNA using the genome profiling method. *Leg Med (Tokyo)*, 2011, **13**(5): 265–267.

Thompson, J. and K. Belschner, Differentiation of various menstrual bloodstains using LDH. *Calif Depart Justice Tie-Line*, 1997, **21**: 34–39.

Wang, Y.X., et al., A new method of identifying the peripheral blood and the menstrual blood. *Fa Yi Xue Za Zhi*, 2012, **28**(5): 359–361.

Wang, Z., et al., Screening and confirmation of microRNA markers for forensic body fluid identification. *Forensic Sci Int Genet*, 2013, **7**: 116–123.

Whitehead, P.H. and G.B. Divall, Assay of "soluble fibrinogen" in bloodstain extracts as an aid to identification of menstrual blood in forensic science. Preliminary findings. *Clin Chem*, 1973, **19**(7): 762–765.

Yang, H., et al., Proteomic analysis of menstrual blood. *Mol Cell Proteomics*, 2012, **11**(10): 1024–1035.

Yao, Y.N., H.L. Lu, and S. Chen, Preparation of anti-human MMP-11 polyclonal antibody for forensic application. *Chin J Forensic Med*, 2007, **22**(5): 305–307, 310.

Yao, Y.N., et al., Detection of MMP-11 from menstrual blood using immunohistochemistry. *Fa Yi Xue Za Zhi*, 2008, **24**(1): 32–33.

Zhang, Y.Q., H.L. Lu, and Y.N. Yao, Detection of matrix metalloproteinase-11 in menstrual blood by enhanced chemiluminescence method. *Fa Yi Xue Za Zhi*, 2012, **28**(2): 109–111.

Zheng, J., H.L. Lu, and S. Chen, Menstrual blood identification and matrix metalloproteinase-11. *Fa Yi Xue Za Zhi*, 2006, **22**(1): S1–S3.

17

Identification of Urine, Sweat, Fecal Matter, and Vomitus

17.1 Identification of Urine

The identification of the presence of urine is useful for the forensic investigation of an alleged sexual assault and harassment involving urination. More importantly, the identification of urine stains can aid in the investigation of homicides involving either ligature or manual strangulation. In these incidents, strangulation victims often involuntarily excrete urine prior to death. The locations of urine stains at the crime scene provide useful information to determine the site where the violence has occurred.

17.1.1 Urine Formation

In humans, the urinary excretory system eliminates soluble toxic wastes that are cellular metabolic by-products. The urinary system consists of the kidney, the ureter, the urinary bladder, and the urethra (Figure 17.1). The formation of urine takes place in the kidneys and, in particular, in the nephrons. The *nephron* is the basic structural and functional unit of the kidney (Figures 17.2 and 17.3). Each nephron is composed of a glomerulus, a Bowman's capsule, and a renal tubule. The glomerulus is formed by a network of capillaries and is surrounded by a Bowman's capsule. *Filtration* is the first step in urine formation. As blood flows through the glomeruli, much of its fluid, except cells and large molecules, is filtered through the capillaries into the Bowman's capsule. The glomerular filtrate, the preliminary form of urine, consists of water, salts (largely sodium and potassium ions), glucose, and waste products such as urea. The filtrate is then passed through the renal tubule where *reabsorption* occurs. During the reabsorption process, water, glucose, nutrients, and ions such as sodium are reabsorbed back into the blood. The last process of urine formation is *secretion*. Secretion occurs at the distal and the collecting tubules of the nephron where ions (such as hydrogen and potassium ions), ammonia, and certain metabolites are secreted from the blood into the lumen of the renal tubule to be eliminated in the urine. The urine is drained from the kidneys through the ureters and is stored in the bladder before it is finally excreted through the urethra.

Urine is an aqueous solution consisting largely of water. Urea is the most abundant waste product in urine, resulting from the elimination of ammonia that is produced from the metabolic process of amino acids. The average urea concentration in human urine is approximately 9 g/L. Other major components are creatinine, uric acid, and ions such as phosphate, sulfate,

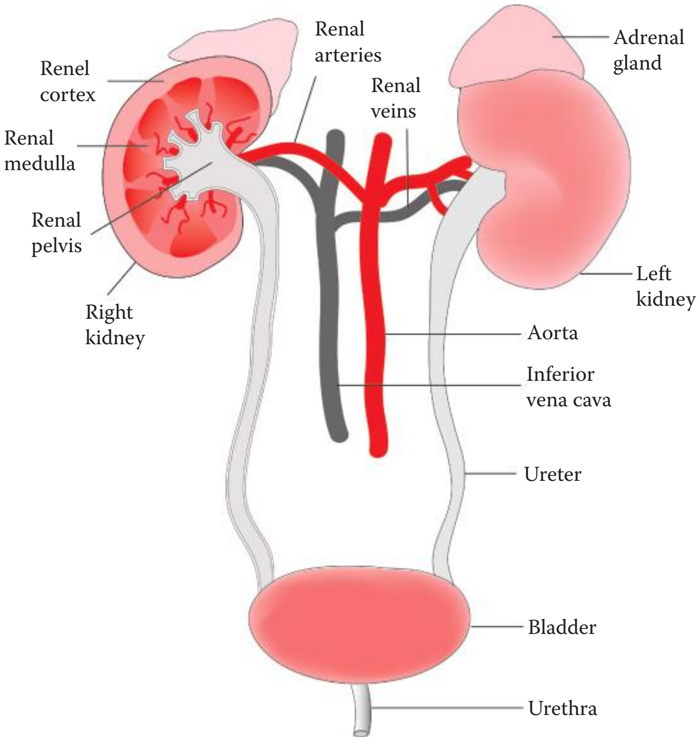

Figure 17.1 Human urinary system. A longitudinal view of the right kidney is shown. (© Richard C. Li.)

chloride, sodium, and potassium. Healthy individuals also excrete a small amount of protein. Urine has a characteristic yellow color. Subsequent to excretion from the body, urine acquires an odor that results from the ammonia that is released from the breakdown of urea by bacteria.

17.1.2 Presumptive Assays

In forensic identification, urine stains can be located by visual examination based on the characteristic yellow color of urine and the detection of the distinctive odor of urine stains. Under alternative light sources, urine stains emit a fluorescent light that facilitates the locating of urine stains in clothing and bedding. Chemical analysis can be carried out to detect the major inorganic anions in urine such as phosphate and sulfate as well as the major organic compounds in urine such as urea, creatinine, and uric acid. These assays are summarized in Table 17.1. However, these assays are not specific to urine. Other bodily fluids, such as sweat, also contain these chemical components.

17.1.2.1 The Identification of Urea

The *para-dimethylaminocinnamaldehyde* (DMAC) assay is simple and rapid and is the most commonly used presumptive assay for the forensic identification of urine stains (Figure 17.4). The DMAC assay can be performed using two different methods: the colorimetric and the fluorometric methods (Figure 17.5).

In the colorimetric method, a portion of a stain (~1 cm^2) is cut and is extracted with 1 mL of distilled water. The extraction is transferred onto a piece of filter paper and is allowed to dry. One drop of 0.1% DMAC solution is then added to the filter paper. DMAC reacts specifically with urea, if present, producing a pink-colored (or magenta-colored) product. DMAC does not react with creatinine, ammonia, or uric acid. The appearance of a pink color within 30 min after applying the DMAC reagent is considered a positive reaction. No color change within 30 min is

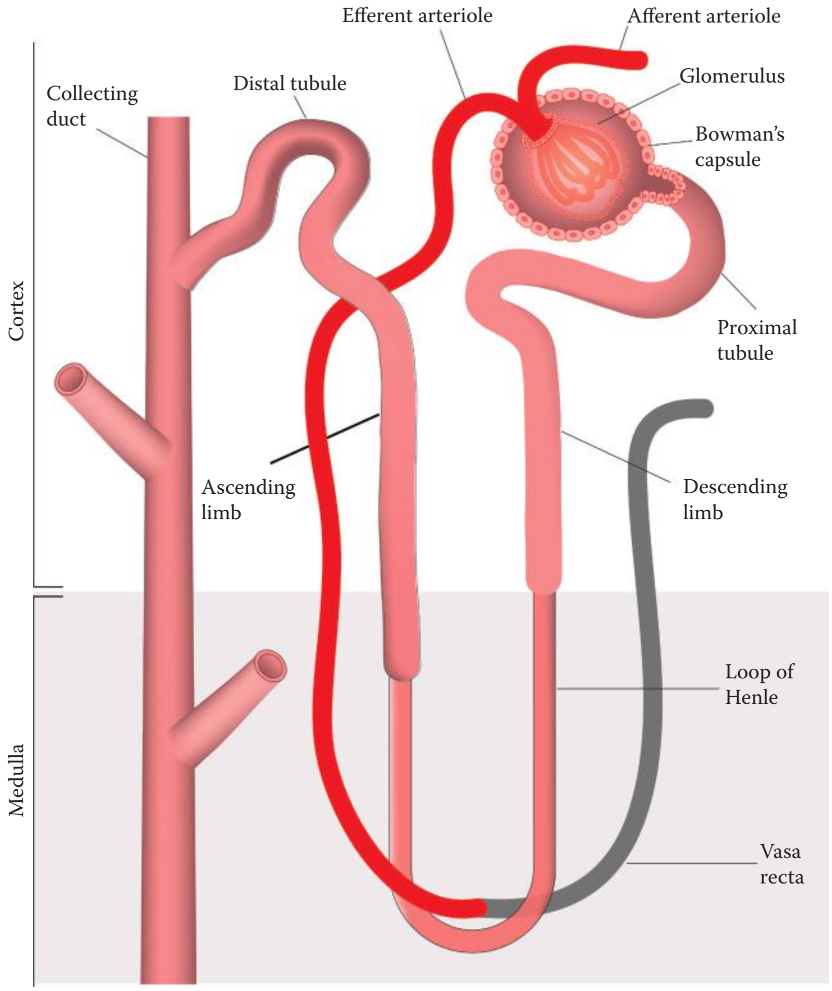
Efferent arteriole
Afferent arteriole
Collecting duct
Distal tubule
Glomerulus
Bowman's capsule
Proximal tubule
Cortex
Ascending limb
Descending limb
Loop of Henle
Medulla
Vasa recta

Figure 17.2 Diagram of the human nephron. (© Richard C. Li.)

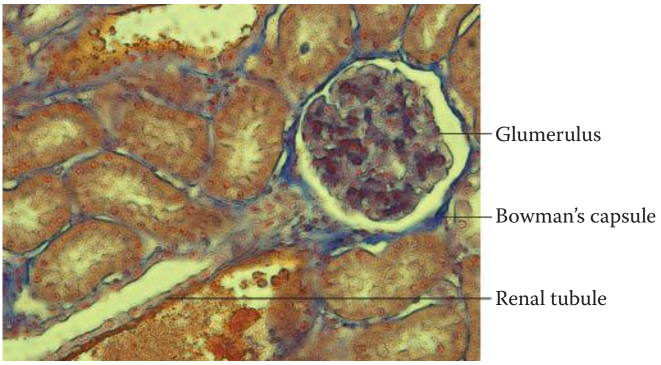
Glumerulus
Bowman's capsule
Renal tubule

Figure 17.3 Section view of the human nephron. (© Richard C. Li.)

Table 17.1 Assays for Identification of Urine and Sweat

Biological Fluid	Assays	Identification	Assay	Method	Further Reading
Urine	Presumptive assays	Locating stains	Visual examination	Alternate light source	Wawryk and Odell (2005)
		Urea	DMAC assay	Colorimetric method	Ong et al. (2012)
				Fluorometric method	Farrugia et al. (2012)
			Microscopic crystal assay	Microscopic examination of urea crystals and urea nitrate crystals converted from urea	Nickolls (1956)
				Microscopic examination of crystals using xanthydrol	Biles and Ziobro (1998)
			Urease assay	The enzyme catalyzes the breakdown of urea and releases ammonia and carbon dioxide. The ammonia is detected using an acid-base indicator, bromthymol blue, which exhibits a blue color	Bedrosian et al. (1984)
				The ammonia is detected using Mn and Ag nitrates, which exhibit a black color	Gaensslen (1983)
			Chromatography	Gas chromatograph–mass spectrometer, thin-layer chromatography	Lum (1991)
		Creatinine	Colorimetric assay	Jaffe's test, using picric acid to convert creatinine to a red creatinine picrate compound	Greenwald (1930)
				Using potassium ferricyanide to produce a blue nitroprusside compound	O'Leary et al. (1992)
			Chromatography	Gas chromatograph–mass spectrometer	Lum (1991)

		Uric acid	Uricase assay	Measuring the disappearance of the absorption of uric acid at 293 nm after treatment with uricase that catalyzes the oxidation of uric acid to 5-hydroxyisourate	Kageyama (1971)
	Confirmatory assays	Tamm–Horsfall glycoprotein	Immunological assays	RIA, crossover electrophoresis, Ouchterlony, ELISA, and immunochromatographic methods	Tsutsumi et al. (1988)
		17-ketosteroids	Mass spectrometer	Liquid chromatograph–mass spectrometer	Nakazono et al. (2002)
Sweat	Presumptive assays	Lactate, lactic acid, urea, and single amino acids	Various	SEM coupled with EDX and Raman microspectroscopy	Sikirzhytski et al. (2011)
	Confirmatory assays	Dermcidin	Immunological assay	ELISA	Sagawa et al. (2003)
			RNA-based assay	RT-PCR	Sakurada et al. (2010)

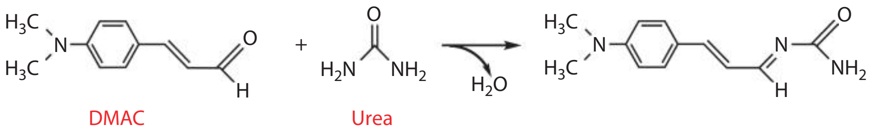

Figure 17.4 Chemical reaction of DMAC assay.

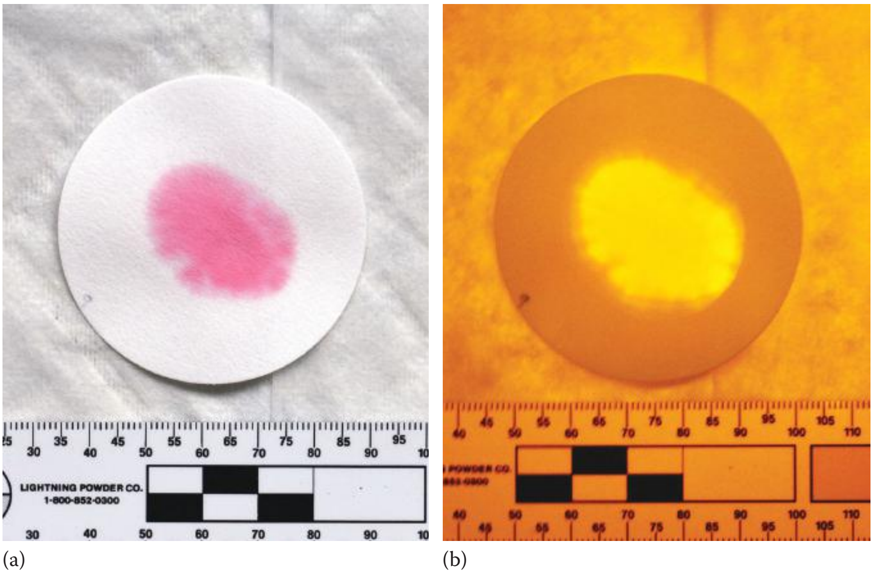
(a)
(b)

Figure 17.5 Positive reaction of DMAC assay of a urine stain. (a) Colorimetric and (b) fluorometric method. (© Richard C. Li.)

considered a negative reaction. However, this method is not specific to urine, as other bodily fluids such as saliva, semen, sweat, and vaginal secretions can also give positive reactions. A diluted DMAC solution, from 0.1% to 0.05%, can maintain appropriate sensitivity to urine stains and minimize false-positive reactions caused by the detection of low levels of urea that are present in other bodily fluids.

The fluorometric method is useful for locating urine stains on large pieces of evidence such as clothing and bedding (Figure 17.6). Additionally, this method can detect patterns of urine stains (Figure 17.7), which can be useful in crime scene reconstructions. In the fluorometric method, the evidence to be examined, such as a garment, is covered by a sheet of filter paper that has been preabsorbed with the DMAC solution. The evidence and the filter are then wrapped together in a sheet of aluminum foil and are left overnight in a press, ensuring that the evidence is in close contact with the filter paper. Alternatively, the layers can be heated for 30 s using an iron. Using a light source at 473–548 nm, the DMAC-treated urine stain fluoresces. The fluoresced stain is best observed with a 549 nm filter. However, colored fabrics interfere with the assay since dyes and pigments can inhibit the fluorescence.

In addition to the DMAC test, the identification of urea that is present in urine stains can be carried out by urease assays. Ureases catalyze the breakdown of urea, thus releasing ammonia and carbon dioxide (Figure 17.8). The ammonia is detected using an acid-base indicator, bromthymol blue, which exhibits a blue color. Alternatively, the ammonia can be detected by manganese and silver nitrates, which exhibit a black color. Additional assays such as microscopic crystal and chromatographic assays are summarized in Table 17.1.

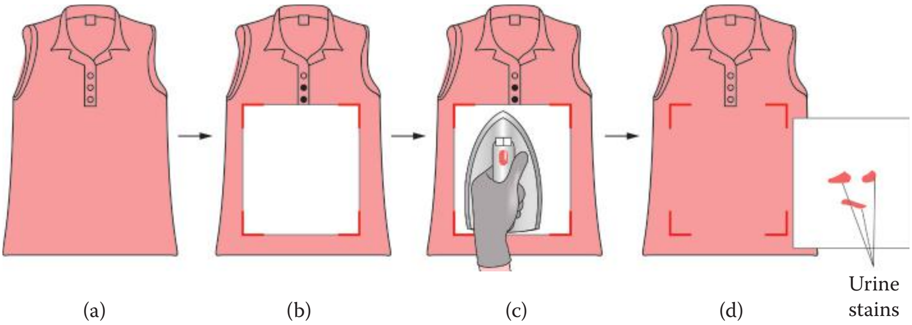

Figure 17.6 Fluorometric DMAC assay of urine stains. Evidence item (a) is closely covered by a piece of moistened filter paper containing DMAC. The orientation of the paper is marked (b). The layers are pressed to allow the transfer of a small amount of the stain onto the paper and are heated using an iron (c). The paper is then lifted. The DMAC-treated urine stain fluoresces under a light source at 473–548 nm (d). (© Richard C. Li.)

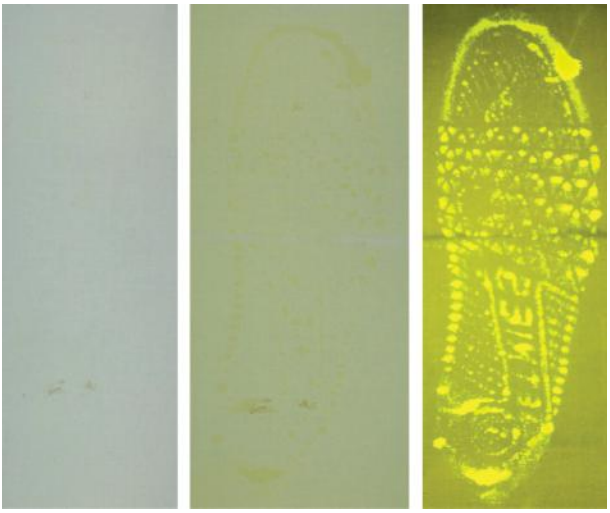

Figure 17.7 Fluorometric DMAC enhancement of a footwear impression in urine on white cotton fabric. From left to right: urine impression, DMAC-treated urine stains under white light, and DMAC-treated urine stains using an excitation light source. (From Farrugia, K.J., et al., *Forensic Sci Int*, 214, 67–81, 2012. With permission.)

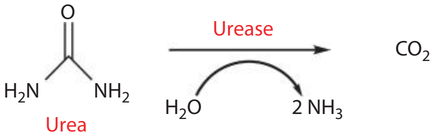

Figure 17.8 Biochemical reaction of urease.

17.1.2.2 Identification of Creatinine

Creatinine is produced during normal muscle cell metabolism (Figure 17.9). During this metabolism, phosphocreatine, an energy-storing molecule in muscle cells, breaks down to form creatine. Creatine is then metabolized to creatinine, which is released from the muscle cells into the blood. Serum creatinine is largely filtrated by the renal glomeruli. A small amount of creatinine is secreted by the renal distal tubules. The amount of creatinine excreted in urine is proportional to the muscle mass of an individual. However, creatinine is not only present in urine. It is also present in other bodily fluids such as blood and semen.

The creatinine present in urine can be detected using the Jaffe color test (Figure 17.10). In this test, picric acid is used to convert creatinine, under alkaline conditions, to form creatinine picrate, which is a bright red product. Additional tests for urine creatinine are summarized in Table 17.1. Recently, a Uritrace device (Abacus Diagnostics) has been made commercially available for the detection of creatinine.

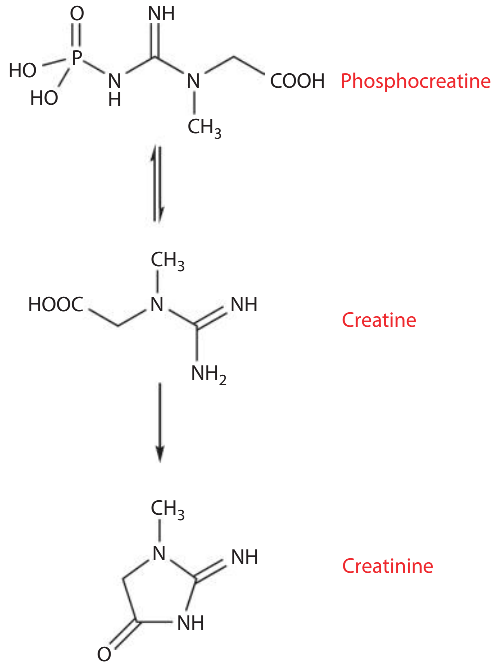

Figure 17.9 Schematic view of the formation of creatinine.

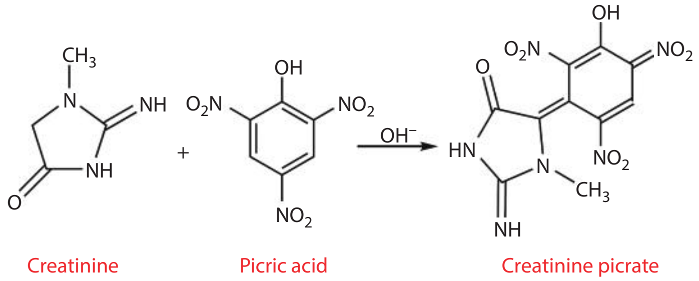

Figure 17.10 Chemical reaction of Jaffe's assay.

17.1.3 Confirmative Assays

17.1.3.1 Identification of Tamm–Horsfall Protein

Tamm–Horsfall protein (THP), also known as *uromodulin*, is the most abundant protein in urine, and accounts for 40% of the urine proteins. THP is exclusively synthesized in the epithelial cells of Henle's loop. THP is secreted from the apical plasma membrane of the epithelial cells into the lumen. Under physiological conditions, an adult excretes 20–100 mg of THP into urine daily. The biological function of THP is not fully understood. It is speculated that it prevents the body from contracting urinary tract infections and from forming renal stones.

THP is a urine-specific biomarker for forensic urine identification. It can be detected using an enzyme-linked immunosorbent assay (ELISA). More recently, it can be detected using an immunochromatographic assay, RSID-Urine, which utilizes a polyclonal rabbit antibody that is specific to THP (Figures 17.11 and 17.12). This test is rapid and simple and thus can be used as a screening test in laboratories and as a field test at crime scenes to identify urine. The detection limit of RSID-Urine for THP is 0.5 μL of urine. Although the sensitivity of RSID-Urine is lower than that of ELISA detection of THP, it is sufficient for forensic applications. These methods are specific to THP that is present only in urine while no detection is found in other bodily fluids such as plasma, saliva, semen, vaginal fluid, or sweat.

17.1.3.2 Identification of 17-Ketosteroids

In humans, urine contains derivatives of 17-ketosteroids such as androsterone, dehydroepiandrosterone (DHEA), and etiocholanolone. *Androsterone* is a steroid hormone with a weak potency of testosterone. *DHEA* is a metabolic intermediate in the biosynthesis of the gonadal steroids. DHEA also has a potential function as a steroid hormone. *Etiocholanolone* is a metabolite of testosterone. These compounds are present in urine as conjugates, in which 17-ketosteroids

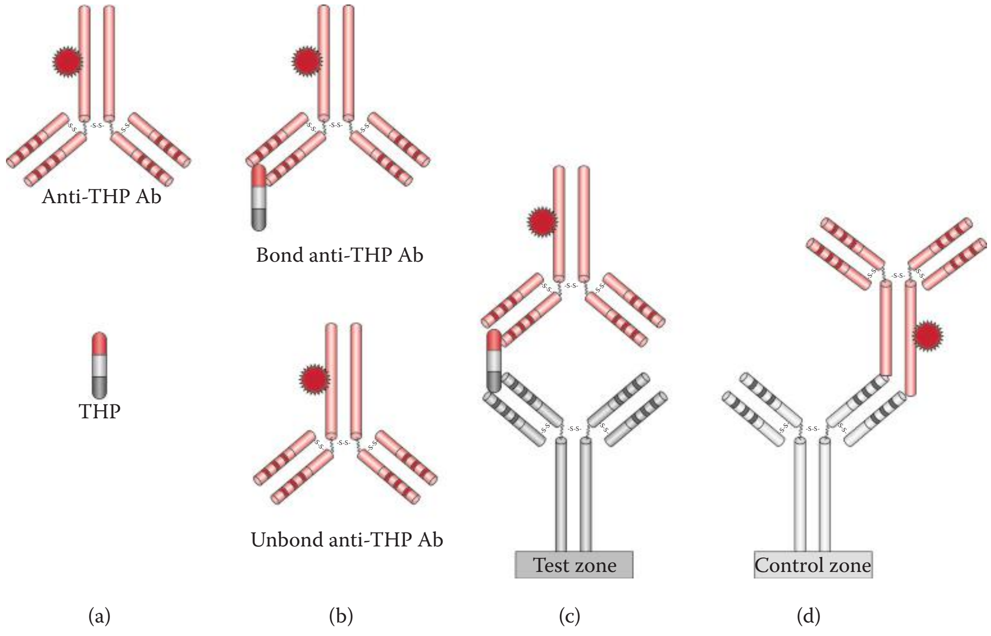

Figure 17.11 Immunochromatographic assays for the identification of THP in urine. (a) In a sample well, THP in a urine sample is mixed with labeled anti-THP Ab. (b) The THP binds to the labeled anti-THP Ab to form a labeled Ab–THP complex. (c) At the test zone, the labeled complex binds to an immobilized anti-THP Ab to form a labeled Ab–THP–Ab sandwich. (d) At the control zone, the labeled anti-THP Ab binds to an immobilized antiglobulin and is captured at the control zone. Ab, antibody; THP, Tamm–Horsfall protein antigen. (© Richard C. Li.)

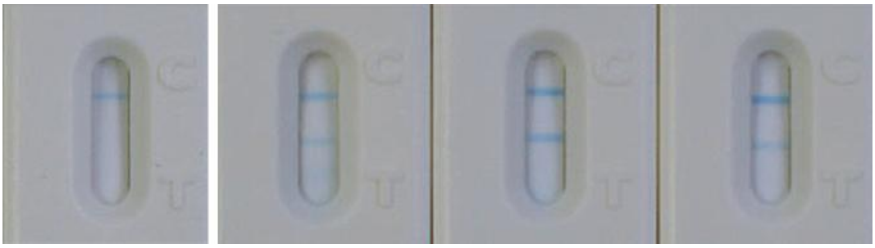

Figure 17.12 Results obtained using immunochromatographic assays (RSID™-Urine; Independent Forensics, Lombard, IL) for the identification of THP in urine. From left to right: negative control, undiluted, 1:10 diluted, and 1:100 diluted urine samples. A visible blue line at the test (T) zone and control (C) zone indicates a positive result. The blue line at the control zone only indicates a negative result. (From Akutsu, T., Watanabe, K., and Sakurada, K., *J Forensic Sci*, 57, 1570–1573, 2012. With permission.)

are modified in the liver through glucuronidation and sulfation. In glucuronidation, glucuronic acids are added to 17-ketosteroids while in sulfation, sulfates are transferred to 17-ketosteroids. As a result, these conjugates, containing charged moieties, are more water soluble than the nonconjugated 17-ketosteroids. Thus, they are excreted into urine to be eliminated from the body. The five major components of the 17-ketosteroid conjugates present in human urine are androsterone glucuronide, androsterone sulfate, DHEA sulfate, etiocholanolone glucuronide, and etiocholanolone sulfate (Figure 17.13). Thus, the analysis of 17-ketosteroid conjugates in urine stains is useful for the identification of human urine stains. These five 17-ketosteroid conjugates can be identified using liquid chromatography–mass spectrometry (LC-MS). However, some of these 17-ketosteroid conjugates are also detected in serum. Therefore, the presence of all five conjugated 17-ketosteroids in a sample is required to identify a urine stain. Additionally, the profiles of the 17-ketosteroid conjugates are human specific and are distinguishable between humans and animals.

17.2 Identification of Sweat

Sweat is the least common bodily fluid analyzed in forensic laboratories compared with others that have been mentioned in previous chapters. However, sweat identification is still useful for forensic investigations. For example, forensic DNA analysis allows the generation of DNA profiles from trace biological evidence such as fingerprints. Identifying sweat can be important for the analysis of these types of evidence.

17.2.1 Biology of Perspiration

Humans have two types of secretory sweat glands: the eccrine and the apocrine sweat glands (Figure 17.14). *Eccrine* sweat glands are distributed almost all over the body and are controlled by the sympathetic nervous system. Eccrine sweat glands play a role in regulating body temperature. When the body temperature rises, eccrine sweat glands secrete a watery sweat to the skin's surface where heat is carried away through the evaporation of the sweat to maintain normal body temperature. In humans, *apocrine* sweat glands, which are associated with hair follicles, are usually restricted to the underarm and genital areas and are controlled by emotional stress. Apocrine sweat glands are inactive until puberty. This type of sweat gland secretes an oily sweat that is odorous after being processed by skin bacteria. The majority of sweat evidence that is analyzed in forensic laboratories is sweat stains secreted from eccrine glands. Sweat contains water, minerals, lactate, and urea. Its biochemical composition varies

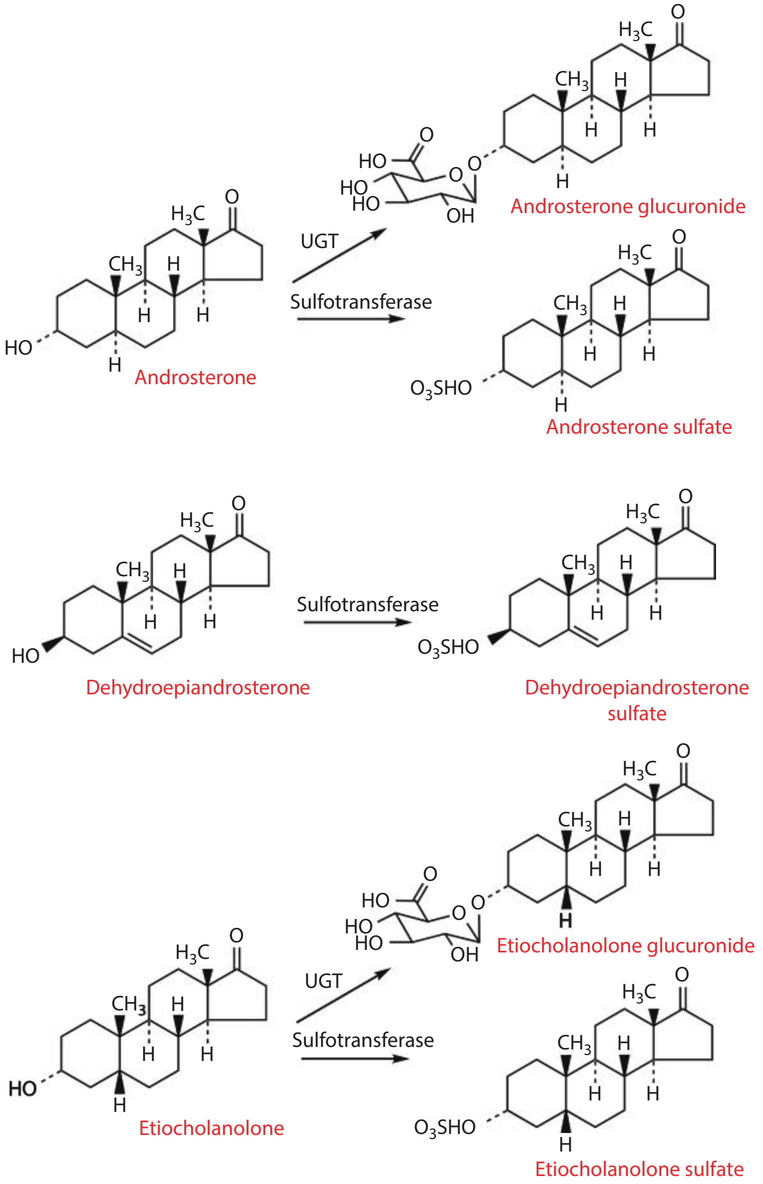
Androsterone glucuronide
UGT
Sulfotransferase
Androsterone
Androsterone sulfate
Dehydroepiandrosterone
Sulfotransferase
Dehydroepiandrosterone sulfate
Etiocholanolone glucuronide
UGT
Sulfotransferase
Etiocholanolone
Etiocholanolone sulfate

Figure 17.13 Structural formula of 17-ketosteroids and 17-ketosteroid conjugates present in human urine. UGT, UDP glucuronosyltransferase. (© Richard C. Li.)

among individuals and their physical activities. Sweat contains low levels of constituents that are also present in other bodily fluids such as urine. Thus, it has been considered a difficult bodily fluid to identify.

17.2.2 Sweat Identification Assays

Sweat evidence has been analyzed using presumptive assays (Table 17.1) such as elemental analysis using scanning electron microscopes coupled with energy dispersive x-ray spectroscopy in the detection of lactic acid. Since sweat contains many inorganic and organic compounds that are also present in other bodily fluids, these assays are not specific to sweat. Raman

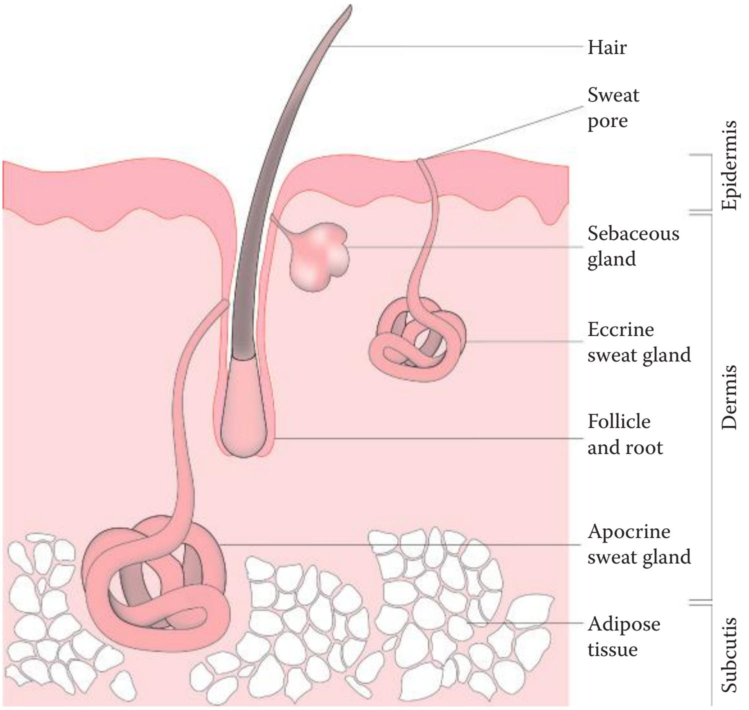

Figure 17.14 Diagram of the human eccrine and apocrine sweat glands. (© Richard C. Li.)

microspectroscopy is potentially useful for the identification of sweat for forensic purposes, which is based largely on the profiles of lactate, lactic acid, urea, and single amino acids in urine.

Recently, *dermcidin* has been identified as a potential biomarker of human sweat. Dermcidin belongs to a class of human antimicrobial peptides of the innate immune defense system and plays an important role in protecting epithelial barriers from infections. Dermcidin, specifically expressed in eccrine sweat glands (Figure 17.15), is secreted into the sweat and is transferred to the epidermal surface. In forensic applications, dermcidin can potentially be utilized as a biomarker for the confirmatory assay of sweat identification. The detection of dermcidin in sweat stains can be performed using ELISA assays utilizing antibodies specific to human dermcidin. This method is highly sensitive and is able to detect dermcidin in sweat samples that are diluted 10,000-fold. Dermcidin is encoded by the *DCD* gene. Its mRNA can be detected using reverse transcription polymerase chain reaction (RT-PCR) assays (Chapter 7) that can detect *DCD* mRNA in 10 µL of sweat sample. Dermcidin assays are also specific to sweat as dermcidin is not detected in other bodily fluids such as semen, saliva, and urine.

17.3 Identification of Fecal Matter

The examination of feces has been used in criminal investigations for over a century. Specifically in 1948, a burglary case was reported using fecal analysis to link the shoes of a suspect to a crime scene. One of the aspects of fecal analysis is to determine a common origin of the reference sample and the fecal evidence, thus potentially linking a suspect to a crime scene. Today, the individual characteristics of a fecal sample can be effectively determined using forensic DNA analysis of sloughed intestinal epithelial cells that are present in fecal matter. The identification of fecal matter is valuable in providing important information for a criminal investigation. For example, the presence of fecal matter may corroborate a sexual

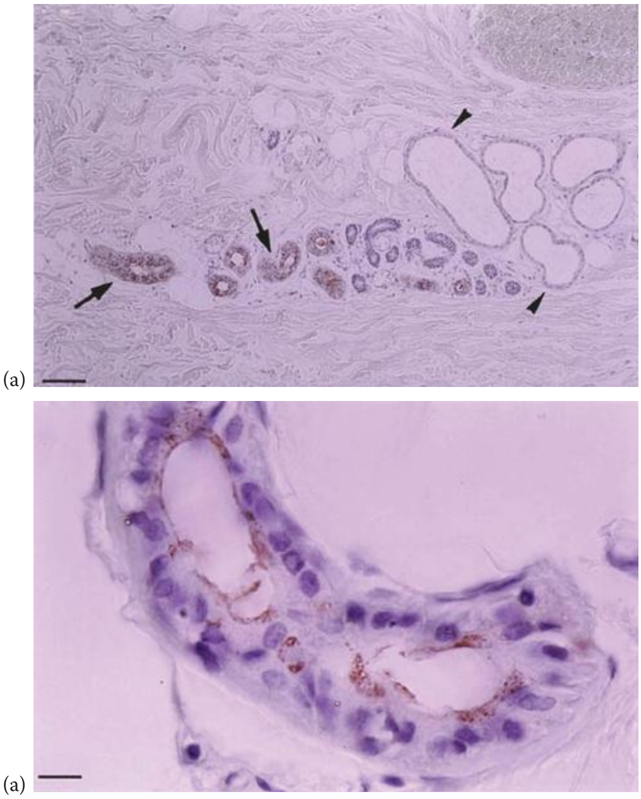

Figure 17.15 Localization of the dermcidin peptide in the sweat glands. Skin tissue sections are treated with a dermcidin antibody and stained. (a) Both the eccrine and apocrine sweat glands are shown. Only cells of the eccrine gland in the skin express dermcidin (brown stained). Arrows: eccrine sweat glands. Arrowheads: apocrine sweat glands. Scale bar: 100 µm. (b) Close-up view of the eccrine gland. The presence of the dermcidin peptide is observed (brown). Scale bar: 10 µm. (From Sagawa, K., et al., *Int J Legal Med*, 117, 90–95, 2003. With permission.)

assault involving sodomy, assault with fecal matter, vandalism, and burglary during which the perpetrator defecated at the scene.

17.3.1 Fecal Formation

Feces are a type of waste matter that is the direct result of food that has been processed by the digestive system (Figure 17.16). Human feces contain undigested foodstuffs, sloughed intestinal epithelial cells, intestinal bacteria, bile pigments, electrolytes, and water. Feces are formed in the intestines during the last phase of digestion. Feces first enter the colon in liquid form. Most of the nutrients are absorbed on the surface area of the small intestine. In the large intestine, water, sodium, and chloride are absorbed on the surface of the lumen. The remaining luminal contents are converted into feces. Food stays for approximately 2–6 h in the stomach. It takes an additional 3–5 h to travel through the small intestine and 12–24 h to travel through the large intestine.

17.3.2 Fecal Matter Identification Assays

A fecal analysis includes macroscopic and microscopic examination, chemical tests, and fecal bacterial identification.

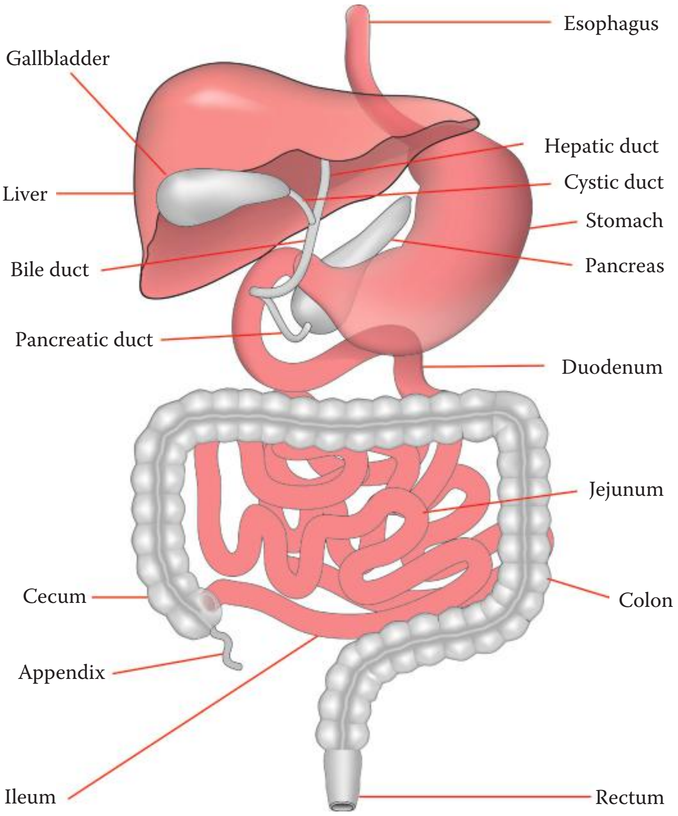

Figure 17.16 Diagram of the human digestive system. The esophagus, stomach, liver, gallbladder, pancreas, small intestine (duodenum, jejunum, and ileum), and large intestine (cecum, colon, appendix, and rectum) are shown. (© Richard C. Li.)

17.3.2.1 Macroscopic and Microscopic Examination

The color and odor of human feces are useful characteristics for fecal identification. The normal brown color of feces primarily results from the presence of urobilinoids, which are heme catabolic by-products. The characteristic odor of feces is caused by the metabolic by-products of the intestinal bacterial flora. Indole, skatole, and hydrogen sulfide are the compounds that are responsible for the odor of feces.

When fecal stains are analyzed, the microscopic examination of the feces can be performed on an aliquot of fecal suspension. The presence of characteristic undigested foodstuffs can indicate human feces. Fecal matter can be transferred from samples of clothing by scraping with a sterile stainless steel spatula. The fecal matter is then hydrated in 6% formalin solution for 24–48 h prior to microscopic examination. Undigested foodstuffs such as vegetable fragments and meat fibers are often present in human feces. Vegetable fragments (Figure 17.17) are often undigested vegetable dermal tissues that cover and protect the plant and fragments of vascular tissues that play roles in transporting water and nutrients throughout the plant. The types of vegetables can be identified by comparing the observed fragments with known plants. Meat fibers (Figure 17.18) are undigested animal skeletal muscle fibers. These fibers have characteristic striations, usually rectangular in shape, that are used for comparison with known animal tissues.

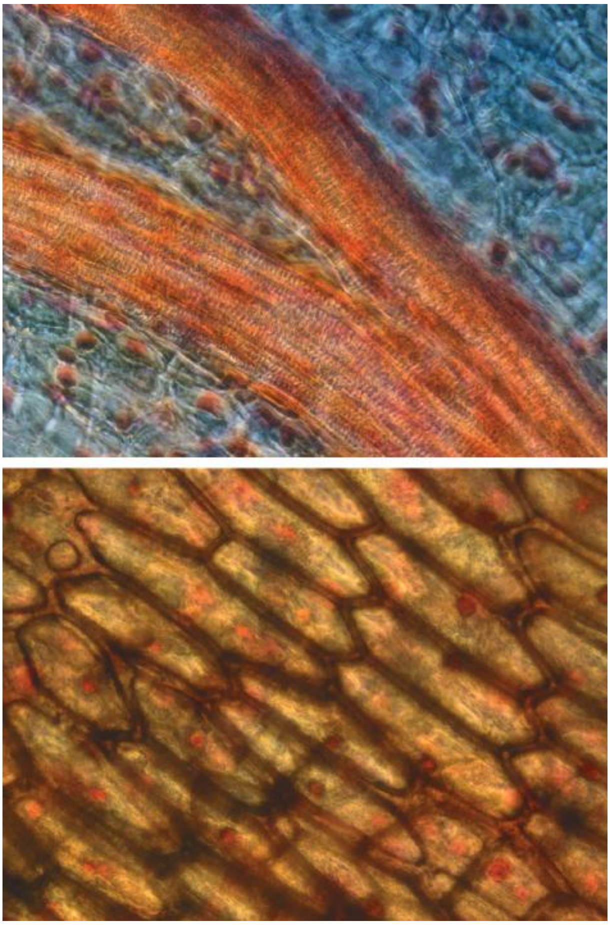

Figure 17.17 Scallion vascular tissue (top) and dermal tissue (bottom). (© Richard C. Li.)

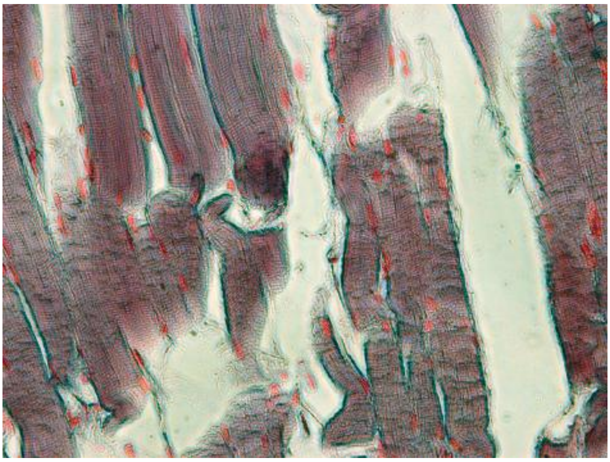

Figure 17.18 Cattle meat fibers. (© Richard C. Li.)

17.3.2.2 Urobilinoids Tests

The forensic analysis of fecal matter often involves the identification of fecal stains on swabs and clothing from which small amounts of sample may be available. Chemical tests can be useful for the analysis of fecal stains. By far the most common chemical test performed on feces is the detection of urobilinoids. *Urobilinoids*, including urobilin and stercobilin, are generated from the degradation of heme and are excreted into feces.

The average lifetime of erythrocytes is approximately 3–4 months. Erythrocytes are continuously undergoing hemolysis in which erythrocytes are naturally broken down and are usually processed in the reticuloendothelial system of the spleen. Hemoglobin is released daily during the hemolysis process and is degraded into heme, globin, and iron. Other sources of heme are derived from the degradation of erythrocyte precursors in bone marrow and other heme-containing proteins such as myoglobin and cytochromes. In the peripheral tissues, heme undergoes catabolism to form *bilirubin* (Figures 17.19 and 17.20). Bilirubin is further converted to *urobilinogen* in the intestine. A portion of urobilinogen is reduced to *stercobilinogen*. In the large intestine, the spontaneous oxidation of urobilinogen and stercobilinogen results in the formation of

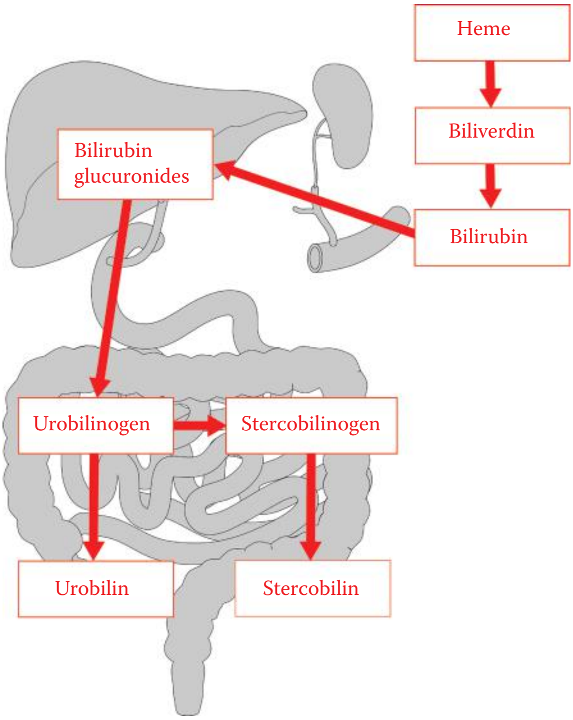

Figure 17.19 The formation of urobilinoids. Aged erythrocytes are disposed in the spleen, releasing hemoglobin that is broken down to heme. The heme is converted to biliverdin and is subsequently reduced to bilirubin. The bilirubin is then released into the bloodstream where it is bound to albumin, which cannot be filtrated at the glomeruli. The bilirubin is transported through the bloodstream to the liver where it is conjugated with glucuronic acid, forming water-soluble bilirubin monoglucuronide and diglucuronide. The bilirubin glucuronides are excreted into the bile and are subsequently excreted into the small intestine. In the intestines, the glucuronic acid of the conjugated bilirubin is removed. The unconjugated bilirubin is metabolized by intestinal bacteria, forming urobilinogen. A portion of the urobilinogen is further metabolized to stercobilinogen. The urobilinogen and the stercobilinogen are oxidized by intestinal bacteria, forming urobilin and stercobilin, respectively, which are excreted into the feces. (© Richard C. Li.)

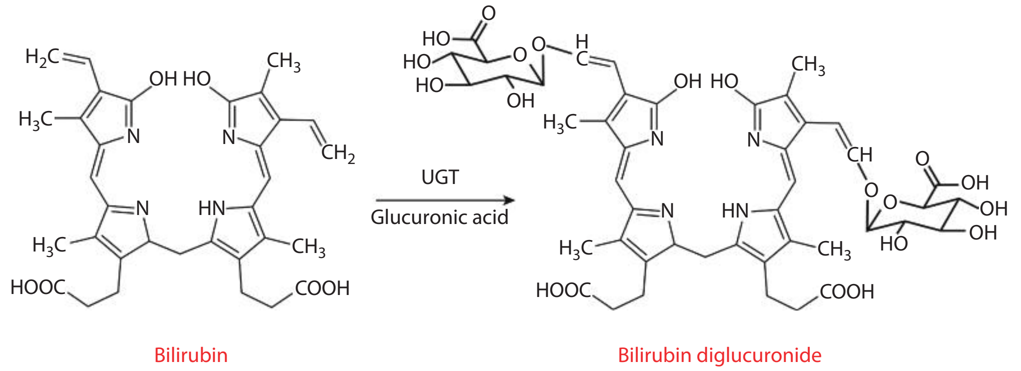

Figure 17.20 The formation of bilirubin glucuronides in the liver. UGT, UDP glucuronosyltransferases.

urobilin and stercobilin (Figure 17.21), respectively. These compounds are brown colored and are responsible for the characteristic color of feces.

The urobilinoids can be detected using the Schlesinger and Edelman tests. In the *Schlesinger* test, a sample is mixed with saturated zinc acetate in ethanol solution to form aurobilinoid–zinc chelation complex that emits a characteristic green fluorescence under ultraviolet light. The *Edelman* test is a variation of the Schlesinger test. A sample is treated with a mercuric salt solution to yield a pink-colored compound. Further treatment with a zinc salt produces fluorescence. However, less fluorescence is observed in the Edelman test than in the Schlesinger test. Inconclusive and inconsistent results are often obtained using these tests where fecal material sometimes gives no visible fluorescence. Additionally, the intensity of the fluorescence observed varies between samples. The reliability and selectivity of the tests can be increased using a spectrometric measurement of the fluorescence detection of fecal urobilinoids based on the principle of the Schlesinger test. A dry sample is treated with 1 mL of zinc acetate solution (1% zinc acetate methoxyethanol solution and 0.2% Tris). The suspension is then sonicated for 5 min, heated at 100°C for 10 min, cooled, and centrifuged. The presence of urobilinoids can be detected using excitation and emission maxima at 507 and 514 nm, respectively.

The disadvantages of the Schlesinger and the Edelman tests are their low species specificity as both tests cannot distinguish between human and other mammalian fecal materials. Moreover, under normal circumstances, up to 5% of urobilinogen is transported to the kidney and oxidized to urobilin in urine. Under some pathological conditions such as hepatic function disorders, the level of urobilin in the urine can be very high. Both the Schlesinger and Edelman tests respond to the urobilin of urine stains as well.

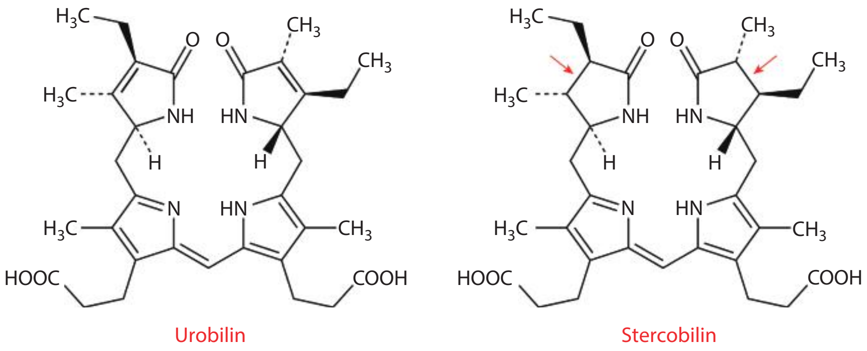

Figure 17.21 Urobilinoids: urobilin and stercobilin. Arrow, site of reduction.

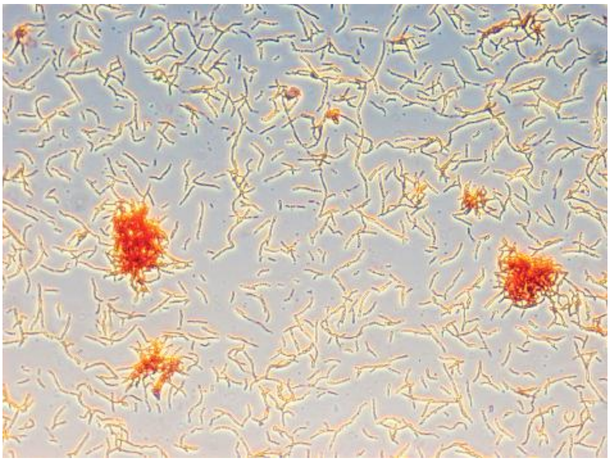

Figure 17.22 Intestinal bacteria. (© Richard C. Li.)

17.3.2.3 Fecal Bacterial Identification

The human intestinal microbiota contains more than 4000 bacterial species (Figure 17.22). Traditional methods for detecting fecal materials utilize the cultivation of fecal indicator bacteria such as *Escherichia coli* or *Enterococci* spp. However, these fecal indicator bacteria constitute only a small portion of the fecal microbiota, and thus may not be adequate for forensic identification purposes. *Bacteroides*, however, accounts for approximately 30% of fecal microbiota that are the predominant bacteria in human feces. *Bacteroides* can potentially be used for the forensic identification of feces. *Bacteroides* is a genus of rod-shaped, anaerobic gram-negative bacteria. These bacteria play a role in digesting complex carbohydrates and other substances that cannot be digested by human enzymes. The identification of *bacteroides* can be carried out by detecting specific DNA sequences of the *rpoB* gene, which encodes the β subunit of bacterial RNA polymerase (Chapter 11). The presence of the species-specific DNA sequence can be detected by RT-PCR utilizing primers that are specific to the target species but not to other species. Thus, only a targeted species can be amplified if it is present. Two fecal predominant bacteroides species, *B. uniformis* and *B. vulgatus*, can be detected in feces. *B. uniformis* is not detectable in blood, saliva, semen, urine, vaginal fluids, or on skin surfaces. Therefore, *B. uniformis* is considered as a specific indicator bacterium for forensic fecal identification. Sometimes, *B. vulgatus* can also be detected in vaginal fluid samples. Therefore, precaution should be taken in interpreting the results obtained using a *B. vulgatus* assay. Additionally, this method alone cannot discriminate between human and animal feces. Furthermore, fecal microbial populations can be affected by the host's diet. Individuals who consume saturated fats and proteins, which are abundant in Western diets, have predominantly *Bacteroides* species in their feces. However, individuals who consume a low-fat and carbohydrate-rich diet have predominantly *Prevotella* species, also a genus of gram-negative bacteria, in their feces.

17.4 Identification of Vomitus

17.4.1 Biology of Gastric Fluid

Gastric fluid can be found in stains derived from stomach wounds. Most often, gastric fluid is from vomitus found at a crime scene or as dried stains on clothing. Vomiting is the forceful expulsion of the contents of the stomach through the mouth (Figure 17.23). It is usually preceded by salivation, sweating, and the sensation of nausea. Vomiting usually begins with a deep

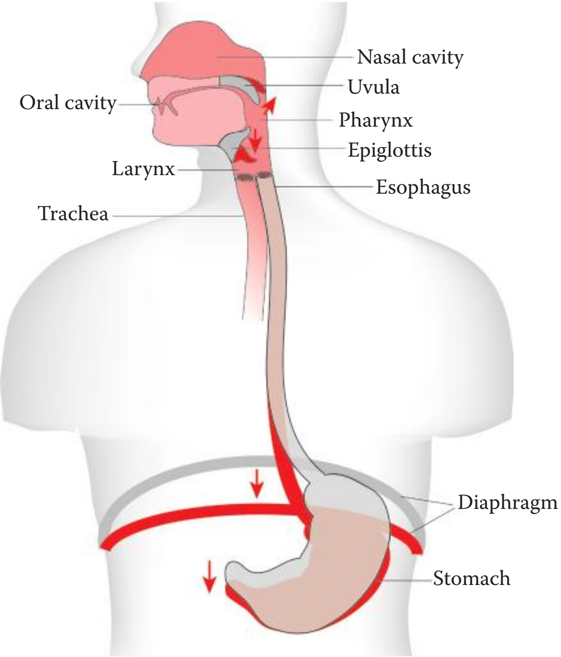

Figure 17.23 Act of vomiting. The contraction of the diaphragm and the stomach during the phase of vomiting is shown (in red). Additionally, the positions of the epiglottis and the uvula during vomiting are shown (in red). As a result, the stomach contents are expelled through the mouth. (© Richard C. Li.)

inhalation and the closure of the glottis. Subsequent contractions of the diaphragm and the abdominal muscles compress the stomach. The gastric contents are then forced upward through the relaxed sphincters and the esophagus, and are expelled through the mouth. Since the glottis is closed, vomitus usually does not enter the respiratory tract. Although the uvula is usually raised to close the nasal cavity, vomitus sometimes enters the nose. Vomiting can be caused by a wide variety of pathological conditions. Vomiting may be a specific response to acute intoxication in homicidal poisoning cases or it may be caused by trauma in a violent assault. Thus, the forensic identification of the gastric fluid can corroborate a criminal act.

The stomach stores ingested food until it can be emptied into the small intestines. When food enters the stomach, hydrochloric acid is secreted in large quantities, which facilitates the initial degradation of proteins. The stomach also secretes mucus that lubricates the gastric surface to protect the epithelium from acidic environments. Hormones such as gastrin, which are found in the gastric fluid, regulate acid secretion and gastric movement. A number of enzymes are secreted into the gastric fluid, including lipase, which plays a role in lipid hydrolysis, and gelatinase, which can hydrolyze gelatin. The stomach also secretes pepsinogens, which are enzyme precursors, into the gastric fluid. In the stomach, pepsinogens are activated by hydrochloric acid into pepsin, which is largely responsible for the digestion of proteins.

17.4.2 Vomitus Identification Assays

Vomitus is highly acidic and tends to be malodorous. The color of vomitus may be of forensic interest. Fresh blood in the vomit is usually bright red and suggests bleeding due to injuries, while dark red blood clots suggest bleeding in the stomach due to pathological conditions such as an ulcer. A microscopic examination (Section 17.3.2.1) can be performed to identify recently ingested food particles that are present in a sample. The presence of gastric fluid in vomitus can also be identified by the detection of pepsins secreted from the stomach. This identification test

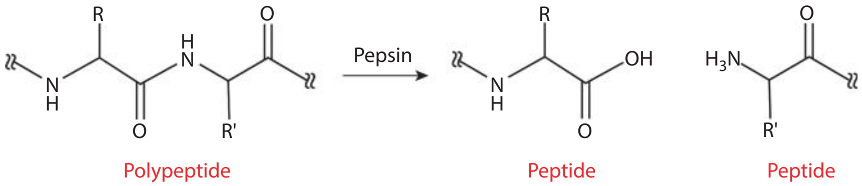

Figure 17.24 The cleavage of polypeptide by pepsin. Pepsin cleaves polypeptide at sites of aromatic amino acids. R and R′: tryptophan, phenylalanine, or tyrosine.

is based on the proteolytic activity of pepsins. Pepsins are endopeptidases that cleave primarily on peptide bonds in the middle of the protein. Aromatic amino acids such as tryptophan, phenylalanine, and tyrosine are the preferred targeted amino acids for the cleavage reaction by pepsins (Figure 17.24). In the pepsin-proteolytic assay, fibrin blue is used as a substrate for pepsin. Fibrin blue is an insoluble protein–dye complex that is colorless. In the presence of vomitus, pepsin cleaves fibrin blue and releases a chromophore that is soluble in water and exhibits a blue color. In the assay, a fibrin blue–containing agarose gel is utilized. The sample from vomitus is loaded onto the gel plate. After incubation, a blue ring around the sample can be observed as a result of the enzymatic reactivity of pepsin. The amounts of pepsin in a sample can be quantified. The results on the gel plate can be photographed and the dried gel plate can also be preserved as evidence. This method can determine the pepsin content of fresh and aged forensic samples. Bodily fluids other than vomitus do not show positive reactions with the use of this method. However, this method cannot distinguish the vomitus of humans from that of other vertebrates.

Bibliography

Akutsu, T., K. Watanabe, and K. Sakurada, Specificity, sensitivity, and operability of RSID-urine for forensic identification of urine: Comparison with ELISA for Tamm–Horsfall protein. *J Forensic Sci*, 2012, **57**(6): 1570–1573.

Akutsu, T., et al., Evaluation of Tamm–Horsfall protein and uroplakin III for forensic identification of urine. *J Forensic Sci*, 2010, **55**(3): 742–746.

Bedrosian, J.L., M.D. Stolorow, and M.A. Tahir, Development of a radial gel diffusion technique for the identification of urea in urine stains. *J Forensic Sci*, 1984, **29**(2): 601–606.

Biles, P.V. and G.C. Ziobro, Identification of the source of reagent variability in the xanthydrol/urea method. *J Assoc Official Anal Chem Int*, 1998, **81**(6): 1155–1161.

Bock, J.H., M.A. Lane, and D.O. Norris, *Identifying Plant Food Cells in Gastric Contents for Use in Forensic Investigations: A Laboratory Manual*, 1988. Washington, DC: US Department of Justice, National Institute of Justice.

Brunzel, N.A., *Fundamentals of Urine and Body Fluid Analysis*, 2nd edn., 2004. Philadelphia, PA: W.B. Saunder.

De Araújo, W.R., M.O. Salles, and T.R.L.C. Paixão, Development of an enzymeless electroanalytical method for the indirect detection of creatinine in urine samples. *Sensor Actuat B-Chem*, 2012, **173**: 847–851.

Farrugia, K.J., et al., Chemical enhancement of footwear impressions in urine on fabric. *Forensic Sci Int*, 2012, **214**(1–3): 67–81.

Farrugia, K.J., et al., A comparison of enhancement techniques for footwear impressions on dark and patterned fabrics. *J Forensic Sci*, 2013, **58**(6): 1472–1485.

Fujishiro, M., et al., Identification of human urinary stains by the uric acid/creatinine quotient and HPLC chromatogram. *J Showa Med Assoc*, 2008, **68**(3): 175–181.

Gaensslen, R.E., Identification of urine, in *Sourcebook in Forensic Serology, Immunology, and Biochemistry*, pp. 191–195, 1983. Washington, DC: US Government Printing Office.

Greenwald, I., The chemistry of Jaffe's reaction for creatinine: VI. A compound of picric acid with two molecules of creatinine. Its combinations with acid and alkali. *J Biol Chem*, 1930, **86**: 333–343.

Iida, R., T. Yasuda, and K. Kishi, Identification of novel fibronectin fragments detected specifically in juvenile urine. *Fed Eur Biochem Soc J*, 2007, **274**(15): 3939–3947.

Johnson, D.J., L.R. Martin, and K.A. Roberts, STR-typing of human DNA from human fecal matter using the QIAGEN QIAamp® stool mini kit. *J Forensic Sci*, 2005, **50**(4): 802–808.

Jones, E., The identification of semen and other body fluids, in R. Saferstein (ed.), *Forensic Science Handbook*, pp. 329–399, 2005. Upper Saddle River, NJ: Prentice Hall.

Juusola, J. and J. Ballantyne, Multiplex mRNA profiling for the identification of body fluids. *Forensic Sci Int*, 2005, **152**(1): 1–12.

Kageyama, N., A direct colorimetric determination of uric acid in serum and urine with uricase-catalase system. *Clin Chim Acta*, 1971, **31**(2): 421–426.

Knight, B.H., The significance of the postmortem discovery of gastric contents in the air passages. *Forensic Sci*, 1975, **6**(3): 229–234.

Lai, Y., C.H. Wang, and J. Qin, A study on STR typing of human sweat latent fingerprints developed by various methods. *Chin J Forensic Med*, 2007, **22**(6): 366–368.

Lee, H.C., et al., Enzyme assays for the identification of gastric fluid. *J Forensic Sci*, 1985, **30**(1): 97–102.

Lloyd, J.B. and N.T. Weston, A spectrometric study of the fluorescence detection of fecal urobilinoids. *J Forensic Sci*, 1982, **27**(2): 352–365.

Lum, P., Seven month old substituted urine sample. *Calif Depart Justice Tie-Line*, 1991, **16**: 79–82.

Metropolitan Police Forensic Science Laboratory, *Biology Methods and Manual*, p. 5, 1978. London: Metropolitan Police Forensic Science Laboratory.

Nakanishi, H., et al., Identification of feces by detection of bacteroides genes. *Forensic Sci Int Genet*, 2013, **7**(1): 176–179.

Nakazono, T., et al., Identification of human urine stains by HPLC analysis of 17-ketosteroid conjugates. *J Forensic Sci*, 2002, **47**(3): 568–572.

Nakazono, T., et al., Dual examinations for identification of urine as being of human origin and for DNA-typing from small stains of human urine. *J Forensic Sci*, 2008, **53**(2): 359–363.

Nickolls, L.C., Urine, in *The Scientific Investigation of Crime*, pp. 209–210, 1956. London: Butterworth.

Norris, D.O. and J.H. Bock, Use of fecal material to associate a suspect with a crime scene: Report of two cases. *J Forensic Sci*, 2000, **45**(1): 184–187.

Nussbaumer, C., E. Gharehbaghi-Schnell, and I. Korschineck, Messenger RNA profiling: A novel method for body fluid identification by real-time PCR. *Forensic Sci Int*, 2006, **157**(2–3): 181–186.

O'Leary, N., A. Pembroke, and P.F. Duggan, A simplified procedure for eliminating the negative interference of bilirubin in the Jaffe reaction for creatinine. *Clin Chem*, 1992, **38**(9): 1749–1751.

Ong, S.Y., et al., Forensic identification of urine using the DMAC test: A method validation study. *Sci Justice*, 2012, **52**(2): 90–95.

Poon, H.H.L., Identification of human urine by immunological techniques. *Can Soc Forensic Sci J*, 1984, **17**: 81–89.

Reddy, K. and E.J. Lowenstein, Forensics in dermatology: Part I. *J Am Acad Dermatol*, 2011, **64**(5): 801–808.

Rhodes, E.F. and J.I. Thornton, DNAC test for urine stains. *J Police Sci Admin*, 1976, **4**: 88–89.

Roy, R., Analysis of human fecal material for autosomal and Y chromosome STRs. *J Forensic Sci*, 2003, **48**(5): 1035–1040.

Sagawa, K., et al., Production and characterization of a monoclonal antibody for sweat-specific protein and its application for sweat identification. *Int J Legal Med*, 2003, **117**(2): 90–95.

Sakurada, K., et al., Detection of dermcidin for sweat identification by real-time RT-PCR and ELISA. *Forensic Sci Int*, 2010, **194**(1–3): 80–84.

Sato, I., Detection of α S1-casein in vomit from bottle-fed babies by enzyme-linked immunosorbent assay. *Int J Legal Med*, 1992, **105**(3): 127–131.

Sato, K., et al., Identification of human urinary stains by the quotient uric acid/urea nitrogen. *Forensic Sci Int*, 1990, **45**(1–2): 27–38.

Sikirzhytski, V., A. Sikirzhytskaya, and I.K. Lednev, Multidimensional Raman spectroscopic signatures as a tool for forensic identification of body fluid traces: A review. *Appl Spectrosc*, 2011, **65**(11): 1223–1232.

Sikirzhytski, V., A. Sikirzhytskaya, and I.K. Lednev, Multidimensional Raman spectroscopic signature of sweat and its potential application to forensic body fluid identification. *Anal Chim Acta*, 2012, **718**: 78–83.

Srch, M., Examination of vomit [in Czech]. *Ceskoslovenska Patologie*, 1975, **11**(3): 46–47.

Srch, M., Examination of vomits. *Vysetrovánizvratků*, 1975, **20**(3): 46–48.

Taylor, M.C. and J.S. Hunt, Forensic identification of human urine by radioimmunoassay for Tamm–Horsfall urinary glycoprotein. *J Forensic Sci Soc*, 1983, **23**(1): 67–72.

Tsutsumi, H., et al., Identification of human urinary stains by enzyme-linked immunosorbent assay for human uromucoid. *J Forensic Sci*, 1988, **33**(1): 237–243.

Virkler, K. and I.K. Lednev, Analysis of body fluids for forensic purposes: From laboratory testing to nondestructive rapid confirmatory identification at a crime scene. *Forensic Sci Int*, 2009, **188**(1–3): 1–17.

Wawryk, J. and M. Odell, Fluorescent identification of biological and other stains on skin by the use of alternative light sources. *J Clin Forensic Med*, 2005, **12**(6): 296–301.

Yamada, S., et al., Vomit identification by a pepsin assay using a fibrin blue-agarose gel plate. *Forensic Sci Int*, 1992, **52**(2): 215–221.

Yan, P., et al., DNA-based species identification for faecal samples: An application on the mammalian survey in Mountain Huangshan scenic spot. *Afr J Biotechnol*, 2011, **10**(57): 12134–12141.

Zhang, S.H., et al., Genotyping of urinary samples stored with EDTA for forensic applications. *Genet Mol Res*, 2012, **11**(3): 3007–3012.

SECTION IV

Individualization of Biological Evidence

18

Blood Group Typing and Protein Profiling

18.1 Blood Group Typing

18.1.1 Blood Groups

For the purposes of this text, blood groups are defined as antigen polymorphisms present on erythrocyte surfaces. Human erythrocyte surface membranes contain a variety of blood group antigens. Transfusion reactions occur when an incompatible type of blood is transfused into an individual, which can lead to severe symptoms or even death. Karl Landsteiner discovered the first blood group, known as the ABO system, in the early 1900s, while studying transfusion and transplantation. The discovery made blood transfusions feasible, and Landsteiner was awarded the Nobel Prize in 1930.

The International Society of Blood Transfusion currently recognizes 29 blood group systems, which include hundreds of antigen polymorphisms (Table 18.1). From the 1950s to the 1970s, the structures and biosynthesis pathways of many blood group antigens were determined. The genes for most of these blood group systems have been identified as well. The isolation of blood group genes has made it possible to understand the molecular mechanisms of the antigenic characteristics of the blood group systems.

The ABO system of antigens in human erythrocytes is the most commonly used blood group system for forensic applications. Forensic laboratories also use others, including the Rh, MNS, Kell, Duffy, and Kidd systems.

18.1.2 ABO Blood Group System

In the ABO blood group system, two types of antigens, designated A and B, give rise to four blood types:

- Type A individuals have the A antigen.
- Type B individuals have the B antigen.
- Type AB individuals have both A and B antigens.
- Type O individuals have neither A nor B antigens.

The antigens may be found in other bodily fluids as well as blood, such as amniotic fluid, saliva, and semen as well as many organs including the kidney, pancreas, liver, and lungs.

Table 18.1 Blood Group Systems

Number	Name	Symbol	Number of Antigens	Gene Names	Chromosomal Location
001	ABO	ABO	4	*ABO*	9q34.2
002	MNS	MNS	43	*GYPA, GYPB, GYPE*	4q31.21
003	P	P1	1		22q11.2-qter
004	Rh	RH	49	*RHD, RHCE*	1p36.11
005	Lutheran	LU	20	*LU*	19q13.32
006	Kell	KEL	25	*KEL*	7q34
007	Lewis	LE	6	*FUT3*	19p13.3
008	Duffy	FY	6	*FY*	1q23.2
009	Kidd	JK	3	*SLC41A1*	18q12.3
010	Diego	DI	21	*SLC4A1*	17q21.31
011	Yt	YT	2	*ACHE*	7q22.1
012	Xg	XG	2	*XG, MIC2*	Xp22.33, Yp11.3
013	Scianna	SC	5	*ERMAP*	1p34.2
014	Dombrock	DO	5	*DO*	12p12.3
015	Colton	CO	3	*AQP1*	7p14.3
016	Landsteiner-Wiener	LW	3	*ICAM4*	19p13.2
017	Chido/ Rodgers	CH/RG	9	*C4A, C4B*	6p21.3
018	H	H	1	*FUT1*	19q13.33
019	Kx	XK	1	*XK*	Xp21.1
020	Gerbich	GE	8	*GYPC*	2q14.3
021	Cromer	CROM	12	*DAF*	1q32.2
022	Knops	KN	8	*CR1*	1q32.2
023	Indian	IN	2	*CD44*	11p13
024	Ok	OK	1	*BSG*	19p13.3
025	Raph	RAPH	1	*CD151*	11p15.5
026	John Milton Hagen	JMH	1	*SEMA7A*	15q24.1
027	I	I	1	*GCNT2*	6p24.2
028	Globoside	GLOB	1	*B3GALT3*	3q26.1
029	Gill	GIL	1	*AQP3*	9p13.3

Source: Adapted from Daniels, G.L., et al., *Vox Sang*, 87, 304–316, 2004.

18.1.2.1 Biosynthesis of Antigens

All individuals generate the *O antigen*, also known as the *H antigen*. The O antigen is synthesized by *fucosyltransferase*, a fucose transferase encoded by the *FUT* genes, which adds a fucose on the end of a glycolipid (in erythrocytes) or glycoprotein (in tissues). An additional monosaccharide (Figure 18.1) is then transferred to the O antigen by a transferase encoded by the *ABO* locus. The specificity of this enzyme determines the ABO blood type (Figure 18.2):

- The *A* allele produces the A-transferase, which transfers N-acetylgalactosamine to the O antigen and thus synthesizes the A antigen.
- The *B* allele produces the B-transferase, which transfers galactose to the O antigen and thus synthesizes the B antigen.
- The *O* allele has a mutation (small deletion), which eliminates transferase activity, and no modification of the O antigen occurs.

As a result, the A and B antigens differ in their terminal sugar molecules. Subgroups of blood types A and B have been described. The most important are the A_1 and A_2 antigens. Both A_1 and A_2 (and A_1B and A_2B) cells react with anti-A antibodies. However, A_1 cells react more strongly than A_2 cells. The apparent difference between A_1 and A_2 is that each A_1 cell contains more copies of the A antigen than A_2 cells.

18.1.2.2 Molecular Basis of the ABO System

A- and B-transferases are encoded by a single gene, *ABO*, on chromosome 9. The *ABO* gene (approximately 20 kb) is organized into seven exons. Most of its coding regions are located in exons 6 and 7 of the *ABO* locus, including the domain responsible for catalytic activity (Figure 18.3). The gene products of the *A* and *B* alleles differ by four amino acid substitutions (Table 18.2). In particular, amino acid residues at positions 266 and 268 are more important in determining the enzymatic property of a transferase.

The A^1 allele and A^2 allele differ in a single nucleotide deletion upstream from the translation stop codon. The resulting reading-frame shift in the A^2 allele abolishes the stop codon, yielding a product with an extra 21–amino acid residue at the C-terminus.

Subgroups of blood types O have also been reported. The sequence of the O^1 allele has a deletion of a single nucleotide at exon 6. This nucleotide deletion leads to a reading-frame shift generating a truncated protein, which lacks the catalytic domain. While the O^{1y} allele also has a single nucleotide deletion, it differs from O^1 by nine nucleotides within the coding sequence. O^1 and O^{1y} have identical phenotypes. There is also an O^2 allele, which is inactivated by a substitution

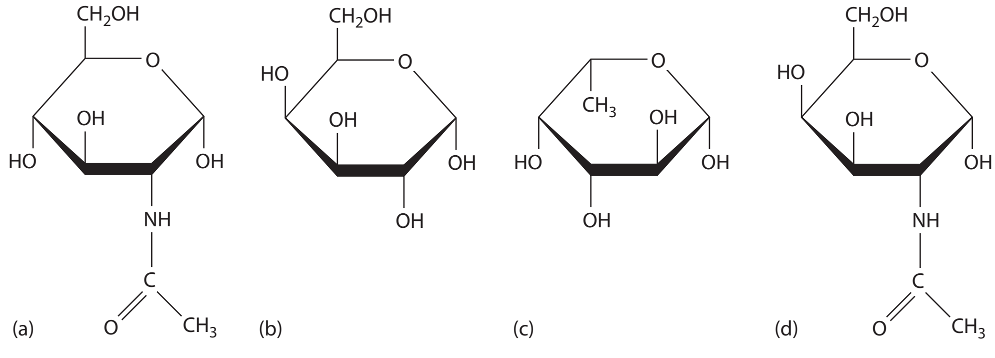

Figure 18.1 Chemical structures of (a) *N*-acetylglucosamine, (b) galactose, (c) fucose, and (d) *N*-acetylgalactosamine.

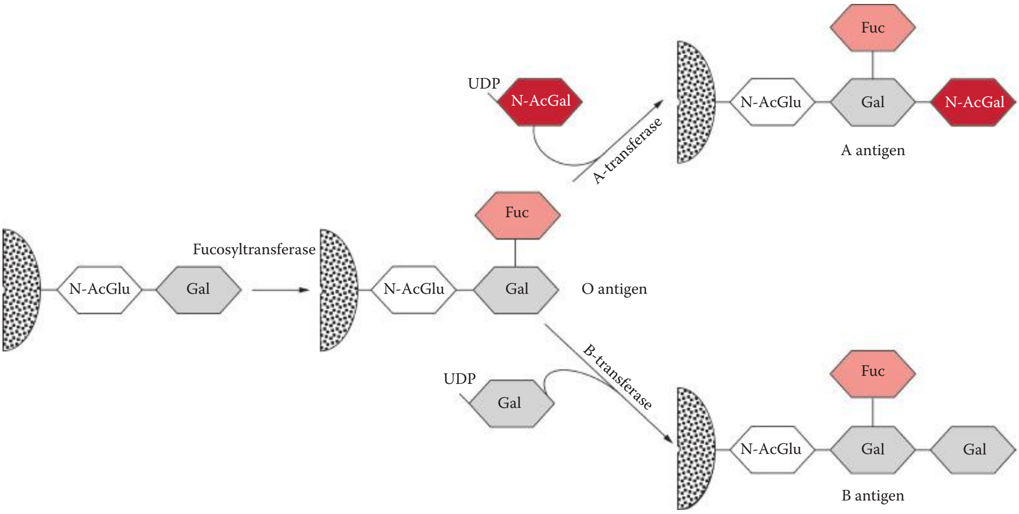
UDP
N-AcGal
A-transferase
Fuc
N-AcGlu
Gal
N-AcGal
A antigen
Fucosyltransferase
N-AcGlu
Gal
Fuc
N-AcGlu
Gal
O antigen
B-transferase
UDP
Gal
Fuc
N-AcGlu
Gal
Gal
B antigen

Figure 18.2 Biosynthesis of ABO antigens. O-antigen biosynthesis is catalyzed by fucosyltransferase. A-antigen biosynthesis is catalyzed by the A-transferase that transfers the N-acetylgalactosamine from the donor and uridine diphosphate (UDP)-*N*-acetylgalactosamine to the O antigen. B-antigen biosynthesis is catalyzed by the B-transferase that transfers the galactose from UDP-galactose to the O antigen. N-AcGlu, *N*-acetytglucosamine; Gal, galactose; Fuc, fucose; N-AcGal, *N*-acetylgalactosamine. (© Richard C. Li.)

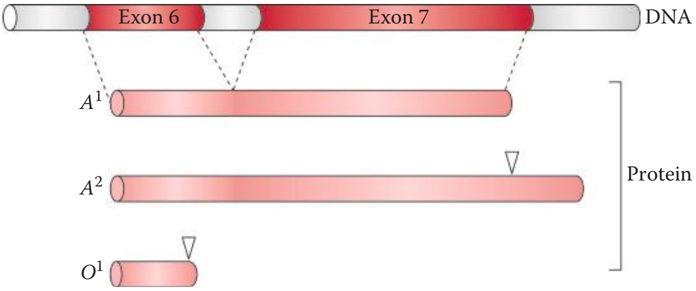
Exon 6
Exon 7
DNA
A1
A2
O1
Protein

Figure 18.3 Structure of ABO gene and variants. Exons 6 and 7 are shown. The deletion mutation in A^2 and O^1 variants is indicated by an inverted triangle and leads to their A_2 and O phenotypes, respectively. (© Richard C. Li.)

Table 18.2 Amino Acid Substitutions at Four Positions in Human ABO Variants A^1, B, and O^2

ABO Variant	Amino Acid Position			
	176	235	266	268
A^1	Arginine	Glycine	Leucine	Glycine
B	Glycine	Serine	Methionine	Alanine
O^2	Glycine	Glycine	Leucine	Arginine

mutation at glycine (position 268) by arginine. Additionally, a few dozen other rare *O* alleles, which yield inactive proteins, have also been documented.

18.1.2.3 Secretors

In addition to erythrocytes, individuals whose A, B, and O antigens can be found in other types of bodily fluids are referred to as *secretors*. Eighty percent of Caucasians are secretors. As described earlier, the O antigen is the substrate for the A- and B-transferase because the A- and B-transferase can only utilize a fucosylated substrate. The O antigen is synthesized by fucosylation of the terminal galactosyl residue catalyzed by the fucosyltransferase, which is encoded by *FUT* genes.

Chromosome 19 contains two homologous genes: *FUT1* and *FUT2*. *FUT1* is expressed in tissues of mesodermal origin (embryonic tissues that serve as precursors of hemopoietic tissues, muscle, the skeleton, and internal organs) and is responsible for the synthesis of the O antigen in erythrocytes. *FUT2* is expressed in tissues of endodermal origin (embryonic tissues that are precursors of the gut and other internal organs); it is responsible for the synthesis of the O antigen in secretions.

About 20% of Caucasian individuals (called *nonsecretors*) are homozygous for a nonsense mutation in *FUT2* at amino acid position 143, resulting in a truncated protein. Bodily fluids such as the semen of type A or B nonsecretors (who carry homozygous *FUT2* mutations) contain no A or B antigens despite containing active A- or B-transferases (Figure 18.4). This can be a problem in investigating sexual assault cases when the blood type of the seminal evidence needs to be determined. However, nonsecretors have O antigens on erythrocytes synthesized by *FUT1* and thus have A or B antigens in blood. Individuals carrying very rare homozygous *FUT1* mutations produce erythrocyte O-deficient phenotypes in which the erythrocytes express no O antigens and thus express neither A nor B antigens, regardless of *ABO* genotype. Individuals

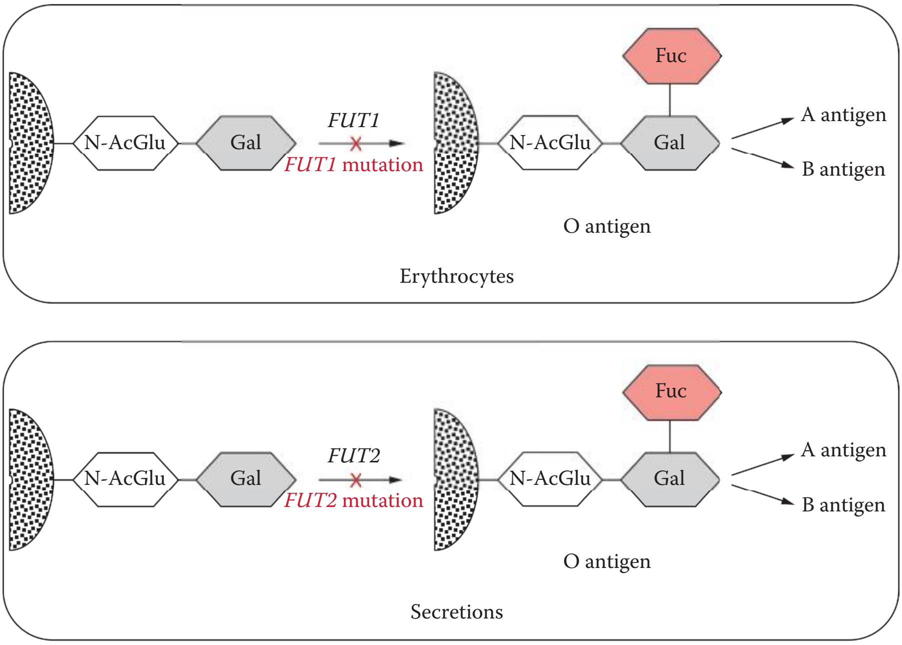

Figure 18.4 Tissue-specific O-antigen biosynthesis by *FUT1* and *FUT2* gene products. Tissue-specific O-antigen biosynthesis in erythrocytes is catalyzed by the *FUT1* gene product; in secretions, it is catalyzed by the *FUT2* gene product. The mutations abolishing the biosynthesis of O antigens are indicated. The *FUT2* mutation produces a nonsecretor phenotype. (© Richard C. Li.)

who carry both *FUT1* and *FUT2* mutations have no O antigens (nor A nor B antigens) in their erythrocytes and other bodily fluids and are known as Bombay (O_h) phenotypes.

18.1.2.4 Inheritance of A and B Antigens

A and *B* alleles are dominant. For *AO* and *BO* heterozygotes, the corresponding transferase synthesizes the A or B antigen. *A* and *B* alleles are codominant in *AB* heterozygotes because both transferase activities are expressed. The *OO* homozygote produces neither transferase activity and therefore lacks both antigens. The inheritance of *A* and *B* alleles obeys Mendelian principles (Chapter 25). For example, an individual with type B blood may have inherited a *B* allele from each parent or a *B* allele from one parent and an *O* allele from the other; thus, an individual whose phenotype is B may have the *BB* (homozygous) or *BO* (heterozygous) genotype. Conversely, if the blood types of the parents are known, the possible genotypes of their children can be determined. When both parents are type B (heterozygous), they may produce children with the genotype *BB* (B antigens from both parents), *BO* (B antigen from one parent, O from the other heterozygous parent), or *OO* (O antigens from parents who are both heterozygous). Thus, blood group typing can be used for paternity testing.

18.1.3 Forensic Applications of Blood Group Typing

The application and usefulness of blood typing in forensic identification are based on the ability to group individuals into four different types using the ABO blood system, allowing individuals to be identified. For example, if one crime scene blood sample is type B and a suspect has type A, the crime scene sample must have a different origin. However, if both the sample and the suspect are type A, the sample may have come from the same origin or from a different origin that happened to be type A.

Unfortunately, the probability that any two randomly chosen individuals have an identical blood type is very high. Approximately 42% of Caucasians have type A blood. The frequency of other blood types within the ABO system is shown in Figure 18.5. Multiple blood group systems were utilized to decrease the probability of a coincident match.

The A and B antigens are very stable and can be identified in dried blood even after many years. They can also be found in semen and other bodily fluids of secretors. Thus, in sexual assault cases, for example, the ABO type of a semen sample can be examined to identify a perpetrator.

18.1.4 Blood Group Typing Techniques

The most common assays used in forensic serology involve agglutination and include the Lattes crust and absorption–elution assays.

18.1.4.1 Lattes Crust Assay

In the early 1900s, Karl Landsteiner used his blood and blood obtained from his laboratory coworkers to test the effects of serum on erythrocytes. He discovered that naturally occurring

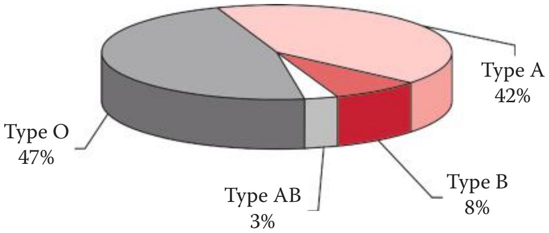

Figure 18.5 Frequency distributions of ABO types observed in American Caucasians. Different human populations may exhibit different frequencies of the four blood types. (© Richard C. Li.)

antibodies in serum caused agglutination of certain erythrocytes, and the agglutination patterns observed were designated A, B, and O. Each pattern indicated the presence or absence of a particular antigen on erythrocytes.

Shortly after birth, newborn infants develop antibodies against antigens that are not present in their own bodies. For example, type A individuals develop anti-B antibodies, type B individuals develop anti-A antibodies, type O individuals develop both types of antibodies, and type AB individuals do not develop anti-A or anti-B antibodies. When the plasma of a type A individual is mixed with type B cells, the anti-B antibodies from the type A individual cause the type B cells to agglutinate. This result forms the basis for blood group typing.

The Lattes crust assay relies on the principles of Landsteiner's experiments. It is an agglutination-based assay that utilizes the A, B, and O indicator cells to test the agglutination reaction with its corresponding naturally occurring serum antibodies in a questioned sample. The procedure for the Lattes crust assay is described in Box 18.1 and illustrated in Figure 18.6. Typical results are summarized in Table 18.3 and illustrated in Figure 18.7. Type A blood contains naturally occurring anti-B antibodies that agglutinate only with B cells. Likewise, type B blood agglutinates only with A cells, type O blood agglutinates with both A and B cells, and type AB blood does not agglutinate with any cells.

The Lattes crust assay is simple and rapid. However, one limitation is that the assay is not very sensitive and requires a large quantity of blood. Recall that successful agglutination reactions usually require intact cells. The agglutination assay of forensic samples is, therefore, difficult to carry out because blood cells lyse when they are dry. Therefore, this method is not reliable for testing old stains.

BOX 18.1 LATTES CRUST ASSAY PROCEDURE

1. Place small quantities of blood crust from a specimen on a microscopic slide and place a cover slide over the crusts. Prepare slides for A, B, and O cells separately.
2. Prepare cell suspensions with saline (0.85% NaCl in phosphate buffer, pH 7.4) for the A, B, and O cells separately.
3. Apply a few drops of the A-cell suspension and allow the cells to diffuse under the cover slip. Repeat this step for B cells and O cells.
4. Incubate the slides in a moisture chamber at room temperature for 2 h.
5. Examine results under a microscope.

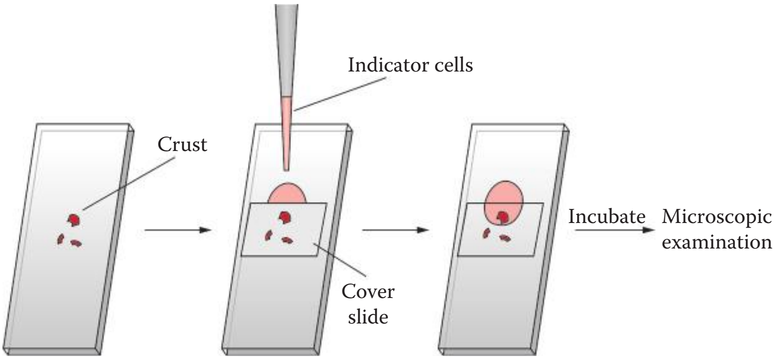

Figure 18.6 Lattes crust assay. (© Richard C. Li.)

Table 18.3 Representative Results of Lattes Crust Assay

Blood Type	Serum Antibody	Agglutination Reaction Observed
A	Anti-B	B cells
B	Anti-A	A cells
O	Anti-A, anti-B	A cells, B cells
AB	None	None

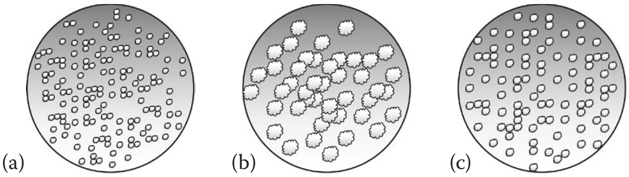

Figure 18.7 Diagram of Lattes crust assay results. (a) Indicator cells added before incubation. (b) Strong agglutination: large clumps are observed after incubation. (c) Negative agglutination: a cloudy background may be observed after incubation.

18.1.4.2 Absorption–Elution Assay

The absorption–elution assay is highly sensitive and can be used for testing dried bloodstains. This method indirectly detects the presence of antigens. The antigens are immobilized in a solid phase (Figure 18.8). At low temperatures, the antigens bind to their corresponding antibodies: anti-A antibodies, anti-B antibodies, or anti-O lectins. (The anti-O lectin is isolated from plants and reacts strongly with the O antigen present in type O blood, but has some cross-reaction with the A antigen). The excess unbound antibodies are removed by washing, and the bound antibodies are then eluted at higher temperatures (recall that antigen–antibody binding can be affected by temperature; Chapter 13). The eluted antibodies can then be identified by an agglutination assay using A, B, and O indicator cells.

Typical results of an absorption–elution assay are summarized in Table 18.4. The bloodstains containing the A antigen can bind to anti-A antibodies. The eluted anti-A antibody can form agglutination with A cells. Likewise, for type B blood, the eluted anti-B antibody can form agglutination with B cells; for type AB blood, the eluted antibodies can form agglutination with both A and B cells; and with type O blood, the eluted anti-O lectins can form agglutination with O cells.

18.2 Forensic Protein Profiling

Because of the limitations of blood group systems, inherited protein polymorphic markers have been utilized to decrease the chances of matches between two unrelated individuals. The amino acid sequences of many proteins vary in the human population. An estimated 20%–30% of the proteins in humans are polymorphic. Some of the variations in amino acid sequences affect the function of proteins, but many of them exert little or no effect on protein function. Thus, individuals can be divided into groups based on the types of protein polymorphisms. A combination of the blood group systems and protein polymorphic markers can be used for

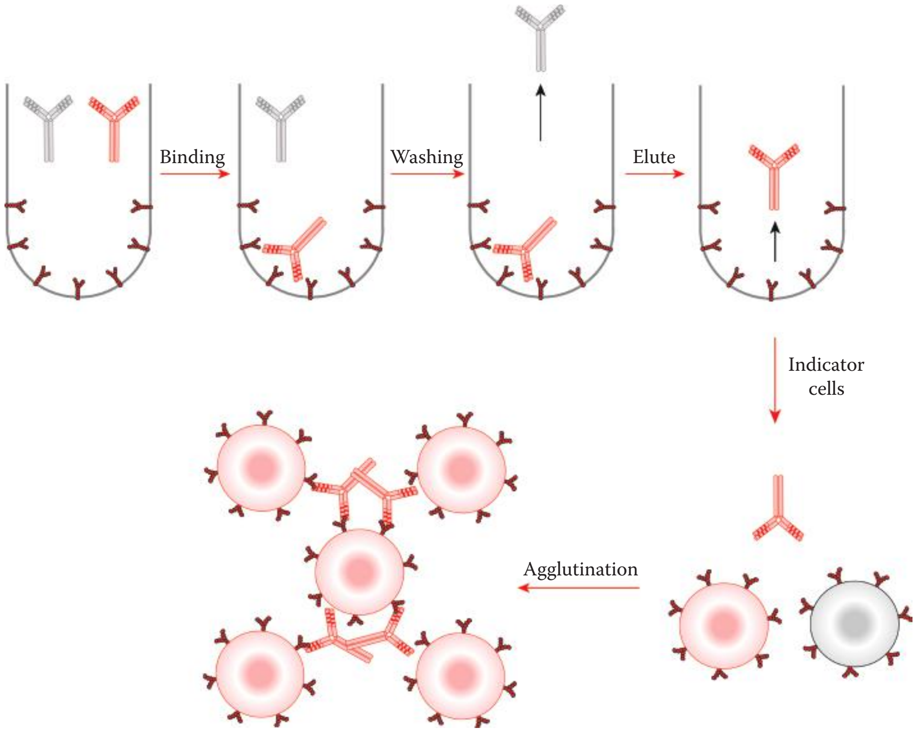

Figure 18.8 Absorption–elution assay. An antigen sample is immobilized on a solid-phase matrix. Antibodies (two different types are shown) are added. The antibody that is specific to the antigen binds. Unbound antibody is washed away. The bound antibody is then eluted. The eluted antibody is tested with indicator cells (two different types are shown). The indicator cell that is specific to the antibody eluted shows a positive agglutination reaction. (© Richard C. Li.)

Table 18.4 Representative Results of Absorption–Elution Assay

Blood Type (Stain)	Antibody Bound and Eluted	Agglutination Reaction Observed
A	Anti-A	A cells
B	Anti-B	B cells
O	Anti-O	O cells
AB	Anti-A, anti-B	A cells, B cells

criminal investigations and paternity testing. The probability that results for two unrelated persons would match is decreased to one in several hundred through use of the blood-typing and protein-profiling techniques.

18.2.1 Methods

Identification of protein polymorphisms is performed through electrophoretic separation based on the molecular weights (*Mr*) and charges of the protein variants.

18.2.1.1 Matrices Supporting Protein Electrophoresis

Electrophoresis of proteins is generally carried out in a support material, also called the matrix, to separate various macromolecules. The matrix also reduces the effects of diffusion and convection on the macromolecules. Historically, protein profiling for forensic application utilizes two types of matrices: papers such as cellulose acetate; and gels composed of starch, agar, agarose, or polyacrylamide. The first polymorphic protein marker, phosphoglucomutase, was characterized by starch-gel electrophoresis. However, agarose and polyacrylamide became more commonly used in electrophoresis due to good reproducibility and reliability (Table 18.5).

18.2.1.2 Separation by Molecular Weight

An electrophoretic method is frequently utilized to resolve various proteins based on their molecular weights. Native electrophoresis, also known as nondenaturing electrophoresis, can be used to isolate proteins for studying the functions of proteins. Biological activity of the protein can be retained for further analysis. However, some proteins are not well separated in electrophoresis in their native form. Thus, it may be necessary to denature the proteins in order for them to be resolved better during separation. This process is called *denaturing protein electrophoresis*. The following additives can be used:

18.2.1.2.1 Reducing Agents

It is common to include reducing agents such as mercaptoethanol (ME), dithiothreitol (DTT), or sodium mercaptoethane sulfonate (MESNA) to denature proteins. Reducing agents cleave the disulfide bonds of proteins. As a result, protein shape becomes unfolded and linear. These agents can be used during sample preparation and can also be added to the electrophoresis buffer.

18.2.1.2.2 Detergents

Detergents disrupt noncovalent interactions within the structures of native proteins. The procedure is generally performed with sodium dodecylsulfate (SDS), a strong anion detergent that binds to most proteins in amounts proportional to the molecular weight of the protein (approximately one molecule of SDS for two amino acids). The bound SDS contributes a large net negative charge on the protein, which masks any surface charges of the native protein. As a result, the charge-to-mass ratio of the protein becomes a constant. As with reducing agents, the various native conformations of proteins change to a more uniformly linear shape when SDS is bound.

Table 18.5 Properties of Matrices Supporting Protein Electrophoresis

Supporting Matrix	Pore Size	EEO	Reproducibility	Strength	Preparation	Toxicity
Cellulose acetate	Large	High	Poor	Good	Simple	Nontoxic
Starch	Large	High	Poor	Fragile	Simple	Nontoxic
Agar	Large	High	Poor	Fragile	Simple	Nontoxic
Agarose	Large	Low	Good	Fragile	Simple	Nontoxic
Polyacrylamide	Small	Very Low	Good	High; tolerates high electronic field	Complex	Toxic

Note: Electroendosmosis (EEO) occurs when fixed charges of the supporting matrix cause liquid flow toward the electrodes. A matrix with high EEO may affect the mobilities and separation performances of proteins during electrophoresis.

Electrophoretic mobility in the presence of SDS, therefore, becomes based on *M*r rather than both *M*r and the charge. Smaller proteins move through the pores of the gel matrix more rapidly than larger proteins. As a result, the larger the size of the protein, the smaller its electrophoretic mobility.

SDS gel electrophoresis can also be used to determine the *M*r of an unidentified protein based on its electrophoretic mobility on the gel. Standard marker proteins of known molecular weight are run on the same gel and allow the estimation of the *M*r of an unknown protein. A linear plot of log *M*r values of marker proteins versus relative migration during electrophoresis allows the molecular weight of the unknown protein to be determined from the graph.

18.2.1.3 Separation by Isoelectric Point

The *isoelectric focusing (IEF)* technique can be used to separate proteins according to their *isoelectric points (pI)*. The pI is the pH value at which the net electric charge of an amino acid is zero. All proteins are composed of amino acids, and each has its own characteristic pI at which its net electric charge is zero and does not migrate in an electric field.

In IEF electrophoresis, a pH gradient is created in a gel between the electrodes, and a protein sample is placed in a well on the gel. With an applied electric field, proteins enter the gel and migrate until they reach a pH equivalent to their pI values, at which they lose mobility (Figure 18.9). IEF, based on molecular charge, is capable of producing sharper bands than denaturing protein electrophoresis and thus has a higher resolving power. The technique can detect very low quantities of proteins in samples. A pH gradient in the gel is established by utilizing materials such as carrier ampholytes or immobilines that are dispersed in the gel. Carrier ampholytes are synthetic amphoteric compounds that contain multiple weak ionizable moieties acting as either acids or bases. To establish a pH gradient, a mixture of ampholytes with slightly different pIs is directly added to an IEF gel. The pH gradient is generated by applying an electric field on the ampholyte-containing gel. Under the electric field, the negatively charged ampholytes migrate toward the anode, and the positively charged ampholytes migrate to the cathode. As a result, a gradual pH gradient is created between the anodal end of the gel (acidic) and the cathodal side of the gel (basic). Immobilines are a series of modified acrylamide monomers that can be acidic or basic. The pH gradient of an IEF gel can also be established using a gradient-forming device that changes the proportion of the immobilines added to the gel matrix mixture as it is loaded into the gel-casting apparatus.

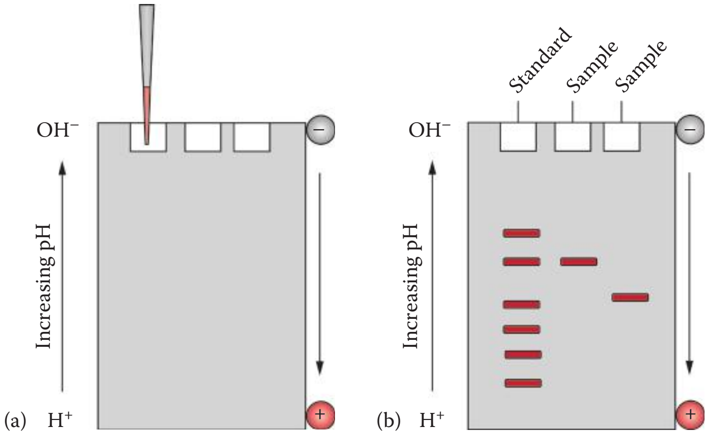

Figure 18.9 Isoelectric focusing. (a) A pH gradient is established by allowing low-molecular-weight organic acids and bases to distribute themselves in an electric field across a gel. A sample containing a protein mixture is loaded into a sample well. (b) During the electrophoresis, each protein migrates until it matches its pI. Proteins with different pIs are separated. (© Richard C. Li.)

18.2.2 Erythrocyte Protein Polymorphisms

18.2.2.1 Erythrocyte Isoenzymes

The human erythrocyte contains a number of *isoenzymes*, which are multiple forms of an enzyme that catalyze the same reaction but differ in their amino acid sequences. Individuals can be divided into groups on the basis of the different isoenzymes present in their erythrocytes. The isoenzyme type is also inherited according to Mendelian principles.

The polymorphism of erythrocyte phosphoglucomutase (PGM) was first described in the 1960s and was later successfully applied to the testing of bloodstains. PGM, an important metabolic enzyme, catalyzes the reversible conversion of glucose-1-phosphate and glucose-6-phosphate. The PGM found in erythrocytes is encoded at the *PGM1* locus at chromosome 1. The PGM encoded by *PGM1* can also be found in semen and thus can be utilized for the testing of semen samples in sexual assault cases. The protein polymorphisms of the PGM have two alleles, which result in three different phenotypes, depending on the combination of the two alleles. The success in the forensic application of PGM led to the similar use of many other erythrocyte isoenzyme polymorphisms. The most commonly used erythrocyte isoenzyme systems are listed in Table 18.6.

18.2.2.2 Hemoglobin

Recall that the use of hemoglobin (Hb) in screening and confirmatory blood tests was discussed in Chapter 12. Adult human Hb consists of two α chains and two β chains. Each polypeptide chain contains a heme group involved in oxygen binding. A very small portion of blood possesses a form of the human adult Hb consisting of two α chains and two δ chains.

More than 200 Hb variants have been identified and can be useful as markers for forensic applications. In particular, two types of human Hb variants are important in forensic testing: fetal Hb and sickle-cell Hb (Hb S). Hb S is the factor responsible for sickle-cell disease (Figure 18.10). Hb variants can be resolved using electrophoresis (Figure 18.11).

18.2.2.2.1 Fetal Hemoglobin

Humans have three forms of Hb during their development: embryonic, fetal, and adult Hb. In adults, the Hb tetramer consists of two identical α and two identical β chains. Embryonic erythrocytes contain Hb tetramers that are different from the adult form. Each embryonic Hb consists of two identical α-like chains and two identical β-like chains. The embryonic Hb is gradually replaced during pregnancy (approximately 3 months after conception) by fetal Hb, which comprises approximately 70% of the Hb in fetal blood. The fetal Hb has two identical α chains and two identical γ chains.

TABLE 18.6 Common Isoenzymes Used for Forensic Protein Profiling

Erythrocyte Isoenzyme	Protein Symbol	Number of Alleles
Phosphoglucomutase	PGM	2[a]
Erythrocyte acid phosphatase	ACP/EAP	3
Esterase D	ESD	2
Adenylate kinase	AK	2
Glyoxalase I	GLO	2
Adenosine deaminase	ADA	2

[a] Ten alleles can be observed using IEF electrophoresis.

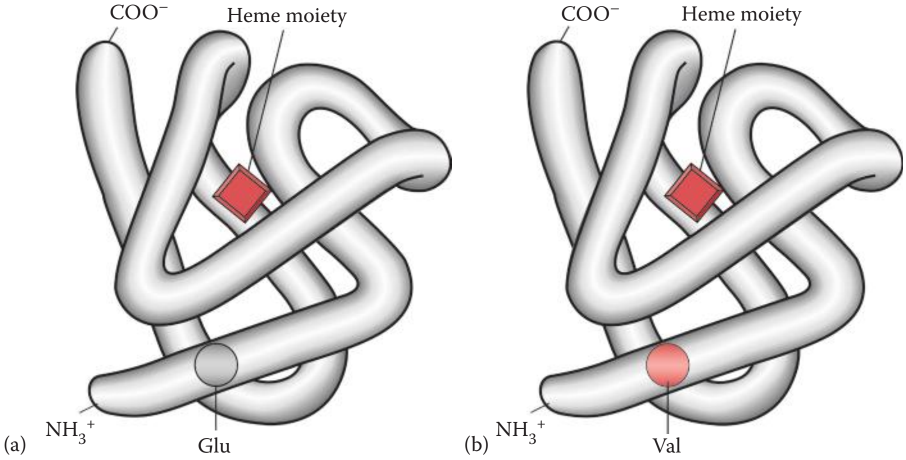

Figure 18.10 Normal and sickle-cell hemoglobin β chains. (a) Normal hemoglobin β chain contains a glutamic acid residue (Glu) at position 6 of the N-terminal of the protein. (b) At position 6 of the sickle-cell hemoglobin β chain, the glutamic acid residue is replaced by a valine (Val). (© Richard C. Li.)

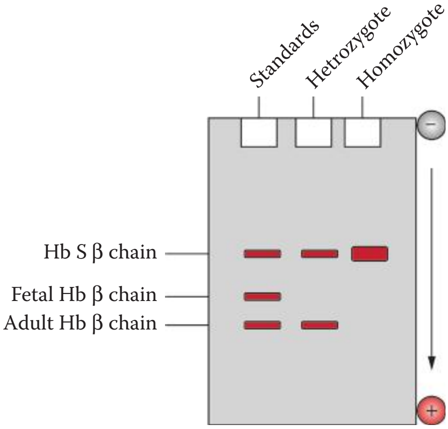

Figure 18.11 Hemoglobins resolved by isoelectric focusing electrophoresis. (© Richard C. Li.)

The embryonic and fetal Hbs have higher affinities to oxygen, required to provide the embryo and fetus with sufficient amounts of oxygen taken from maternal blood. Fetal Hb is replaced by adult Hb approximately 6 months after birth. These Hbs are encoded by their corresponding genes located at the globin gene clusters. The detection of fetal Hb in a bloodstain via electrophoresis can provide important evidence in cases of infanticide and concealed birth.

18.2.2.2.2 Hemoglobin S

Hb S polymorphism has forensic importance in identifying individuals. The Hb S polymorphism is observed in high frequencies among those of African heritage and some Hispanic populations. Such a protein polymorphic marker can provide investigational leads for the indication of the ethnic origin of a perpetrator. Hb S transports oxygen much less efficiently than normal Hb. Individuals who are homozygous for Hb S usually die early after suffering from sickle-cell anemia and related complications. However, a heterozygous individual (an individual with a copy of the wild-type Hb allele from one parent and a copy of the Hb S allele from the other) can survive. This condition is known as the sickle-cell trait.

In the 1950s, Vernon Ingram of Cambridge University discovered the molecular mechanism of the Hb S defect. His work revealed that the Hb S bears a mutation, which changes the glutamic acid in wild type to a valine at the sixth amino acid from the N-terminal end of the β chain. This substitution of amino acids causes a major change in the structure of the β chain, which in turn results in sickle-cell anemia.

18.2.3 Serum Protein Polymorphisms

The serum portion of blood consists of a large number of proteins. The work on serum proteins for forensic purposes started in the 1950s, when variations in serum proteins were found useful for distinguishing individuals. Over the years, a number of serum proteins were characterized and applied for forensic testing. Haptoglobin (Hp) was the most widely used of the polymorphic serum proteins in forensic biology (Figure 18.12). Haptoglobin is a protein that binds and transports Hb from the bloodstream to the liver for the recycling of the iron contained in the Hb.

Immunoglobulin (Chapter 10) accounts for approximately 15% of serum protein and has been found to be highly variable. Two immunoglobulin proteins are utilized for forensic application. The γ chain protein (G_m) is the heavy chain of immunoglobulin G and the κ chain protein (K_m) is one of two types of the light chain of all immunoglobulins. Table 18.7 lists common serum group systems. All exhibit genetic variations and can be detected in bloodstains. The variants of these proteins can be determined by electrophoresis or serological methods.

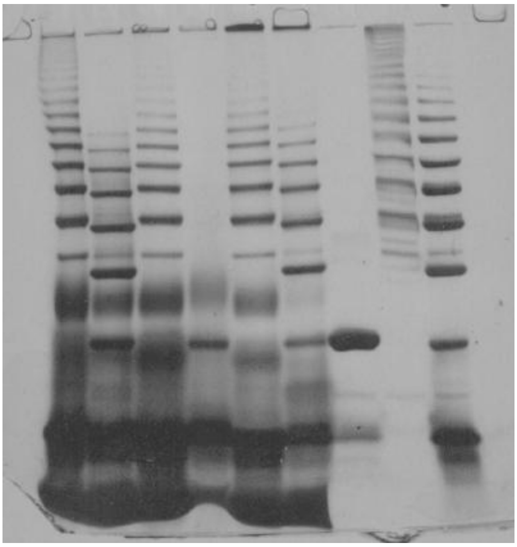

Figure 18.12 Polyacrylamide gel electrophoresis of haptoglobin proteins. From left to right: Hp2, Hp2-1, Hp2, Hp1, Hp2, Hp2-1, Hp1, Hp2, Hp2-1. An anode is at the bottom. (From James, S. and Nordby, J.J., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, CRC Press, Boca Raton, FL, 2005. With permission.)

TABLE 18.7 Serum Proteins Used for Forensic Protein Profiling

Serum Protein	Protein Symbol	Gene Symbol	Chromosomal Location	Number of Amino Acids	Number of Alleles
Haptoglobin	Hp	*HP*	16q22.1	328/387	2
Group-specific component	Gc	*GC*	4q12–13	458	2
Transferrin	Tf	*TF*	3q22.1	679	3
Protease inhibitor (α1-antitrypsin)	Pi	*PI*	14q32.1	394	Many

Source: Adapted from Yuasa, I. and Umetsu, K., *Leg Med*, 7, 251–254, 2005.

Bibliography

Blood Groups

Achermann, F.J., et al., Soluble type A substance in fresh-frozen plasma as a function of ABO and secretor genotypes and Lewis phenotype. *Transfus Apher Sci*, 2005, **32**(3): 255–262.

Aki, K., A. Izumi, and E. Hosoi, The evaluation of histo-blood group ABO typing by flow cytometric and PCR-amplification of specific alleles analyses and their application in clinical laboratories. *J Med Invest*, 2012, **59**(1–2): 143–151.

Allen, F.H. Jr. and R.E. Rosenfield, Review of Rh serology. Eight new antigens in nine years. *Haematologia* (Budap), 1972, **6**(1): 113–120.

Anstee, D.J. and M.J. Tanner, Separation of ABH, I, Ss antigenic activity from the MN-active sialoglycoprotein of the human erythrocyte membrane. *Vox Sang*, 1975, **29**(5): 378–389.

Asari, M., et al., A new method for human ABO genotyping using a universal reporter primer system. *J Forensic Sci*, 2010, **55**(6): 1576–1581.

Bargagna, M. and M. Pereira, A study of absorption–elution as a method of identification of Rhesus antigens in dried bloodstains. *J Forensic Sci Soc*, 1967, **7**(3): 123–130.

Bargagna, M., M. Sabelli, and C. Giacomelli, The detection of Rh antigens (D,C,c,E,e) on bloodstains by a micro-elution technique using low ionic strength solution (LISS) and papain-treated red cells. *Forensic Sci Int*, 1982, **19**(2): 197–203.

Bassler, G., Determination of the Lewis blood group substances in stains of forensically relevant body fluids. *Forensic Sci Int*, 1986, **30**(1): 29–35.

Boorman, K.E., B.E. Dodd, and P.J. Lincoln, *Blood Group Serology*, 5th edn., vol. 4, p. 495, 1977. London: Churchill Livingstone.

Brauner, P., DNA typing and blood transfusion. *J Forensic Sci*, 1996, **41**(5): 895–897.

Brewer, C.A., P.L. Cropp, and L.E. Sharman, A low ionic strength, hemagglutinating, autoanalyzer for rhesus typing of dried bloodstains. *J Forensic Sci*, 1976, **21**(4): 811–815.

Busuttil, A., et al., Assessment of Lewis blood group antigens and secretor status in autopsy specimens. *Forensic Sci Int*, 1993, **61**(2–3): 133–140.

Camp, F.R. Jr., Forensic serology in the United States. I. Blood grouping and blood transfusion—Historical aspects. *Am J Forensic Med Pathol*, 1980, **1**(1): 47–55.

Cartron, J.P., et al., "Weak A" phenotypes. Relationship between red cell agglutinability and antigen site density. *Immunology*, 1974, **27**(4): 723–727.

Cayzer, I. and P.H. Whitehead, The use of sensitized latex in the identification of human bloodstains. *J Forensic Sci Soc*, 1973, **13**(3): 179–181.

Chatterji, P.K., A simplified mixed agglutination technique for ABO grouping of dried bloodstains using cellulose acetate sheets. *J Forensic Sci Soc*, 1978, **17**(2–3): 143–144.

Chester, M.A. and M.L. Olsson, The ABO blood group gene: A locus of considerable genetic diversity. *Transfus Med Rev*, 2001, **15**(3): 177–200.

Daniels, G., *Human Blood Groups*, 2nd edn., p. 576, 2002. Oxford: Blackwell Science.

Daniels, G.L., et al., Blood group terminology 2004: From the International Society of Blood Transfusion committee on terminology for red cell surface antigens. *Vox Sang*, 2004, **87**(4): 304–316.

DeForest, P., R. Gaensslen, and H.C. Lee, Blood, in *Forensic Science*: *An Introduction to Criminalistics*, pp. 230–263, 1983. New York: McGraw-Hill.

De Soyza, K. and D.G. Garland, Studies and observations on Lewis grouping of body fluids and stains. *Forensic Sci Int*, 1988, **38**(1–2): 129–137.

Denomme, G.A., The structure and function of the molecules that carry human red blood cell and platelet antigens. *Transfus Med Rev*, 2004, **18**(3): 203–231.

Erskine, A.G., *The Principles and Practices of Blood Grouping*, 1973. St. Louis, MO: Mosby.

Fayrouz, I.N., N. Farida, and A.H. Irshad, Relation between fingerprints and different blood groups. *J Forensic Leg Med*, 2012, **19**(1): 18–21.

Ferri, G. and S. Pelotti, Multiplex ABO genotyping by minisequencing. *Methods Mol Biol*, 2009, **496**: 51–58.

Fiori, A. and P. Benciolini, The ABO grouping of stains from body fluids. *Z Rechtsmed*, 1972, **70**(4): 214–222.

Fiori, A., M. Marigo, and P. Benciolini, Modified absorption–elution method Siracusa for ABO and MN grouping of blood-stains. *J Forensic Sci*, 1963, **8**(3): 419–445 contd.

Gaensslen, R.E., et al., Evaluation of antisera for bloodstain grouping. II. Ss, Kell, Duffy, Kidd, and G_m/K_m. *J Forensic Sci*, 1985, **30**(3): 655–676.

Gardas, A. and J. Koscielak, Megaloglycolipids—Unusually complex glycosphingolipids of human erythrocyte membrane with A, B, H and I blood group specificity. *Fed Eur Biochem Soc Lett*, 1974, **42**(1): 101–104.

Grunbaum, B.W., Some new approaches to the individualization of fresh and dried bloodstains. *J Forensic Sci*, 1976, **21**(3): 488–509.

Haak, W., J. Burger, and K.W. Alt, ABO genotyping by PCR-RFLP and cloning and sequencing. *Anthropol Anz*, 2004, **62**(4): 397–410.

Hakomori, S. and A. Kabata, Blood group antigens, in *The Antigens*, 1974. New York: Academic Press.

Hamaguchi, H. and H. Cleve, Solubilization of human erythrocyte membrane glycoproteins and separation of the MN glycoprotein from a glycoprotein with I, S, and A activity. *Biochim Biophys Acta*, 1972, **278**(2): 271–280.

Harrington, J.J., R. Martin, and L. Kobilinsky, Detection of hemagglutinins in dried saliva stains and their potential use in blood typing. *J Forensic Sci*, 1988, **33**(3): 628–637.

Harrington, J.J., et al., Chemically sensitized erythrocytes for hemagglutination reactions. *J Forensic Sci*, 1990, **35**(5): 1115–1124.

Henry, S., et al., Molecular basis for erythrocyte Le(a+ b+) and salivary ABH partial-secretor phenotypes: Expression of a *FUT2* secretor allele with an A→T mutation at nucleotide 385 correlates with reduced alpha (1,2) fucosyltransferase activity. *Glycoconj J*, 1996, **13**(6): 985–993.

Hughes-Jones, N.C. and B. Gardner, The Kell system studied with radioactively-labelled anti-K. *Vox Sang*, 1971, **21**(2): 154–158.

Hughes-Jones, N.C., B. Gardner, and P.J. Lincoln, Observations of the number of available c,D, and E antigen sites on red cells. *Vox Sang*, 1971, **21**(3): 210–216.

Huh, J.Y., et al., A rapid long PCR-direct sequencing analysis for ABO genotyping. *Ann Clin Lab Sci*, 2011, **41**(4): 340–345.

Jaff, M.S., Higher frequency of secretor phenotype in O blood group—Its benefits in prevention and/or treatment of some diseases. *Int J Nanomed*, 2010, **5**: 901–905.

Jiang, X., et al., An integrated system of ABO typing and multiplex STR testing for forensic DNA analysis. *Forensic Sci Int Genet*, 2012, **6**(6): 785–797.

Kabat, E.A., *The Blood Group Substances*, p. 330, 1956. New York: Academic Press.

Kaneko, M., et al., Molecular characterization of a human monoclonal antibody to B antigen in ABO blood type. *Immunol Lett*, 2003, **86**(1): 45–51.

Karpoor, C., et al., Study of secretors and non-secretors in normal healthy population—Its forensic implication in human identification. *Indian J Forensic Med Tox*, 2010, **4**(1): 11–13.

Kelly, R.J., et al., Molecular basis for H blood group deficiency in Bombay (Oh) and para-Bombay individuals. *Proc Natl Acad Sci U S A*, 1994, **91**(13): 5843–5847.

Kelly, R.J., et al., Sequence and expression of a candidate for the human secretor blood group alpha (1,2) fucosyltransferase gene (*FUT2*). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J Biol Chem*, 1995, **270**(9): 4640–4649.

Kimura, H. and S. Matsuzawa, Lewis blood group determination in bloodstains by planimetric measurement of eluted monoclonal antibodies. *J Forensic Sci*, 1991, **36**(4): 999–1009.

Kimura, A., et al., ABO blood grouping of semen from mixed body fluids with monoclonal antibody to tissue-specific epitopes on seminal ABO blood group substance. *Int J Legal Med*, 1991, **104**(5): 255–258.

Kimura, A., et al., Blood group A glycosphingolipid accumulation in the hair of patients with alpha-*N*-acetylgalactosaminidase deficiency. *Life Sci*, 2005, **76**(16): 1817–1824.

Kind, S.S., Absorption–elution grouping of dried blood smears. *Nature*, 1960, **185**: 397–398.

Kind, S.S., Absorption–elution grouping of dried blood-stains on fabrics. *Nature*, 1960, **187**: 789–790.

Kind, S.S. and R.M. Cleevely, The fluorescent antibody technique. Its application to the detection of blood group antigens in stains. *J Forensic Med*, 1970, **17**(3): 121–129.

Kobilinsky, L. and J.J. Harrington, Detection and use of salivary hemagglutinins for forensic blood grouping. *J Forensic Sci*, 1988, **33**(2): 396–403.

Koda, Y., M. Soejima, and H. Kimura, The polymorphisms of fucosyltransferases. *Leg Med (Tokyo)*, 2001, **3**(1): 2–14.

Koda, Y., et al., Missense mutation of *FUT1* and deletion of *FUT2* are responsible for Indian Bombay phenotype of ABO blood group system. *Biochem Biophys Res Commun*, 1997, **238**(1): 21–25.

Korchagina, E.Y., et al., Design of the blood group AB glycotope. *Glycoconj J*, 2005, **22**(3): 127–133.

Ladd, C., et al., A PCR-based strategy for ABO genotype determination. *J Forensic Sci*, 1996, **41**(1): 134–137.

Lappas, N.T. and M.E. Fredenburg, The identification of human bloodstains by means of a micro-thin-layer immunoassay procedure. *J Forensic Sci*, 1981, **26**(3): 564–569.

Lattes, L., *The Individuality of Blood*, p. 413, 1932. London: Oxford University Press.

Lee, H.C., Identification and grouping of bloodstains, in R. Saferstein (ed.), *Forensic Science Handbook*, 1982. Englewood Cliffs, NJ: Prentice Hall.

Lee, H.C., et al., Genetic markers in human bone: II. Studies on ABO (and IGH) grouping. *J Forensic Sci*, 1991, **36**(3): 639–655.

Lee, H.Y., et al., Rapid direct PCR for ABO blood typing. *J Forensic Sci*, 2011, **56**(Suppl 1): S179–S182.

Lee, J.C., et al., ABO genotyping by mutagenically separated polymerase chain reaction. *Forensic Sci Int*, 1996, **82**(3): 227–232.

Lee, J.C., et al., ABO genotyping by single strand conformation polymorphism—Using CE. *Electrophoresis*, 2009, **30**(14): 2544–2548.

Lee, J.C., et al., ABO genotyping by capillary electrophoresis. *Methods Mol Biol*, 2013, **919**: 113–120.

Lee, S.H., et al., Rapid ABO genotyping using whole blood without DNA purification. *Korean J Lab Med*, 2009, **29**(3): 231–237.

Lincoln, P.J. and B.E. Dodd, The application of a micro-elution technique using anti-human globulin for the detection of the S, s, K, Fya, Fyb and Jka antigens in stains. *Med Sci Law*, 1975, **15**(2): 94–101.

Liu, Y.H., et al., Distribution of H type 1 and of H type 2 antigens of ABO blood group in different cells of human submandibular gland. *J Histochem Cytochem*, 1998, **46**(1): 69–76.

Maeda, K., et al., ABO genotyping by TaqMan assay and allele frequencies in a Japanese population. *Leg Med (Tokyo)*, 2013, **15**(2): 57–60.

Martin, P.D., A manual method for the detection of Rh antigens in dried bloodstains. *J Forensic Sci Soc*, 1978, **17**(2–3): 139–142.

McDowall, M.J., P.J. Lincoln, and B.E. Dodd, Increased sensitivity of tests for the detection of blood group antigens in stains using a low ionic strength medium. *Med Sci Law*, 1978, **18**(1): 16–23.

McDowall, M.J., P.J. Lincoln, and B.E. Dodd, Observations on the use of an autoanalyser and a manual technique for the detection of the red cell antigens C, D, E, c, K and S in bloodstains. *Forensic Sci*, 1978, **11**(2): 155–164.

Miller, C.H., et al., Measurement of von Willebrand factor activity: Relative effects of ABO blood type and race. *J Thromb Haemost*, 2003, **1**(10): 2191–2197.

Mollicone, R., A. Cailleau, and R. Oriol, Molecular genetics of H, Se, Lewis and other fucosyltransferase genes. *Transfus Clin Biol*, 1995, **2**(4): 235–242.

Moureau, P., Determination of blood groups in bloodstains, in F. Lundquist (ed.), *Methods of Forensic Science*, p. 137, 1963. New York: Interscience.

Muro, T., et al., Determination of ABO genotypes by real-time PCR using allele-specific primers. *Leg Med (Tokyo)*, 2012, **14**(1): 47–50.

Nakanishi, H., et al., Preparation of latex reagents combined with IgM and its F(ab′)2 fragment from commercial ABO blood grouping reagent. *Colloid Surf B Biointerf*, 2007, **54**(1): 114–117.

Nickolls, L.C. and M. Pereira, A study of modern method of grouping dried bloodstains. *Med Sci Law*, 1962, **2**: 172.

Ohmori, T., et al., Monoclonal antibodies against blood group A secretors and nonsecretors saliva. *Hybrid Hybridomics*, 2003, **22**(3): 183–186.

Okiura, T., et al., A-elute alleles of the ABO blood group system in Japanese. *Leg Med (Tokyo)*, 2003, **5**(Suppl 1): S207–S209.

Oriol, R., ABO, Hh, Lewis, and secretion: Serology, genetics, and tissue distribution, in J.-P. Cartron and P. Rouger (eds), *Blood Cell Chemistry*, pp. 36–73, 1995. New York: Plenum Press.

Oriol, R., J.J. Candelier, and R. Mollicone, Molecular genetics of H. *Vox Sang*, 2000, **78**(Suppl 2): 105–108.

Outteridge, R.A., Absorption–elution grouping of bloodstains: Modification and development. *Nature*, 1962, **194**: 385.

Outteridge, R.A., Absorption–elution method of grouping blood-stains. *Nature*, 1962, **195**: 818–819.

Outteridge, R.A., Recent advances in the grouping of dried blood and secretion stains. A.S. Curry (ed.), *Methods of Forensic Science*,, 1965. New York: Interscience.

Painter, T.J., W.M. Watkins, and W.T. Morgan, Serologically active fucose-containing oligosaccharides isolated from human blood-group A and B substances. *Nature*, 1965, **206**(984): 594–597.

Prokop, O. and G. Uhlenbruck, *Human Blood and Serum Groups*, 2nd edn., p. 891, 1969. New York: Wiley Interscience.

Reid, M.E. and C. Lomas-Francis, *The Blood Group Antigen Facts Book*, 2nd edn., 2004. London: Academic Press.

Reid, M.E. and N. Mohandas, Red blood cell blood group antigens: Structure and function. *Semin Hematol*, 2004, **41**(2): 93–117.

Rosenfield, R.E., F.H. Allen Jr., and P. Rubinstein, Genetic model for the Rh blood-group system. *Proc Natl Acad Sci U S A*, 1973, **70**(5): 1303–1307.

Ruan, L., H. Zhao, and Q. Li, Multicolor real-time PCR genotyping of ABO system using displacing probes. *J Forensic Sci*, 2010, **55**(1): 19–24.

Sasaki, M. and H. Shiono, ABO genotyping of suspects from sperm DNA isolated from postcoital samples in sex crimes. *J Forensic Sci*, 1996, **41**(2): 275–278.

Schleyer, F., Investigation of biological stains with regard to species origin, in F. Lundquist (ed.), *Methods of Forensic Science*, p. 291, 1962. New York: Interscience.

Shaler, R.C., A.M. Hagins, and C.E. Mortimer, MN determination in bloodstains–selective destruction of cross-reacting activity. *J Forensic Sci*, 1978, **23**(3): 570–576.

Shintani-Ishida, K., et al., A new method for ABO genotyping to avoid discrepancy between genetic and serological determinations. *Int J Legal Med*, 2008, **122**(1): 7–9.

Smeets, B., H. van de Voorde, and P. Hooft, ABO bloodgrouping on tooth material. *Forensic Sci Int*, 1991, **50**(2): 277–284.

Snyder, L.H., *Blood Groups*, p. 35, 1973. Minneapolis: Burgess.

Springer, G.F. and P.R. Desai, Human blood-group MN and precursor specificities: Structural and biological aspects. *Carbohydr Res*, 1975, **40**(1): 183–192.

Storry, J.R. and M.L. Olsson, Genetic basis of blood group diversity. *Br J Haematol*, 2004, **126**(6): 759–771.

Styles, W.M., B.E. Dodd, and R.R. Coombs, Identification of human bloodstains by means of the mixed antiglobulin reaction on separate cloth fibrils. *Med Sci Law*, 1963, **20**: 257–267.

Tsuji, A., et al., A familial case of ABO phenotype-genotype discrepancy. *Fukuoka Igaku Zasshi*, 2013, **104**(2): 40–45.

Watanabe, K., et al., A novel method for ABO genotyping using a DNA chip. *J Forensic Sci*, 2011, **56**(Suppl 1): S183–S187.

Watkins, W.M., Blood-group substances. *Science*, 1966, **152**(3719): 172–181.

Watkins, W.M., Commemoration of the centenary of the discovery of the ABO blood group system. *Transfu Med*, 2001, **11**: 239–351.

Wynbrandt, F. and W.J. Chisum, Determination of the ABO blood group in hair. *J Forensic Sci Soc*, 1971, **11**(3): 201–204.

Xingzhi, X., et al., ABO blood grouping on dental tissue. *J Forensic Sci*, 1993, **38**(4): 956–960.

Yamada, M., et al., Determination of ABO genotypes with DNA extracted from formalin-fixed, paraffin-embedded tissues. *Int J Legal Med*, 1994, **106**(6): 285–287.

Yamamoto, F., et al., Molecular genetic basis of the histo-blood group ABO system. *Nature*, 1990, **345**(6272): 229–233.

Yu, Q., et al., Congenital tetragametic blood chimerism explains a case of questionable paternity. *J Forensic Sci*, 2011, **56**(5): 1346–1348.

Protein Profiling

Allen, R.C., R.A. Harley, and R.C. Talamo, A new method for determination of alpha-1-antitrypsin phenotypes using isoelectric focusing on polyacrylamide gel slabs. *Am J Clin Pathol*, 1974, **62**(6): 732–739.

Bark, J.E., M.J. Harris, and M. Firth, Typing of the common phosphoglucomutase variants using isoelectric focusing—A new interpretation of the phosphoglucomutase system. *J Forensic Sci Soc*, 1976, **16**(2): 115–120.

Blake, N.M. and K. Omoto, Phosphoglucomutase types in the Asian-Pacific area: A critical review including new phenotypes. *Ann Hum Genet*, 1975, **38**(3): 251–273.

Budowle, B., A method for subtyping group-specific component in bloodstains. *Forensic Sci Int*, 1987, **33**(3): 187–196.

Budowle, B., A method to increase the volume of sample applied to isoelectric focusing gels. *Forensic Sci Int*, 1984, **24**(4): 273–277.

Budowle, B. and E. Scott, Transferrin subtyping of human bloodstains. *Forensic Sci Int*, 1985, **28**(3–4): 269–275.

Budowle, B. and P. Eberhardt, Ultrathin-layer polyacrylamide gel isoelectric focusing for the identification of hemoglobin variants. *Hemoglobin*, 1986, **10**(2): 161–172.

Budowle, B., S. Sundaram, and R.E. Wenk, Population data on the forensic genetic markers: Phosphoglucomutase-1, esterase D, erythrocyte acid phosphatase and glyoxylase I. *Forensic Sci Int*, 1985, **28**(2): 77–81.

Burdett, P.E. and P.H. Whitehead, The separation of the phenotypes of phosphoglucomutase, erythrocyte acid phosphatase, and some haemoglobin variants by isoelectric focusing. *Anal Biochem*, 1977, **77**(2): 419–428.

Burdett, P.E., Isoelectric focusing in agarose: Phosphoglucomutase (*PGM* locus 1) typing. *J Forensic Sci*, 1981, **26**(2): 405–409.

Carracedo, A. and L. Concheiro, The typing of alpha-1-antitrypsin in human bloodstains by isoelectric focusing. *Forensic Sci Int*, 1982, **19**(2): 181–184.

Carracedo, A., et al., A silver staining method for the detection of polymorphic proteins in minute bloodstains after isoelectric focusing. *Forensic Sci Int*, 1983, **23**(2–3): 241–248.

Constans, J. and M. Viau, Group-specific component: Evidence for two subtypes of the Gc1 gene. *Science*, 1977, **198**(4321): 1070–1071.

Constans, J., et al., Analysis of the Gc polymorphism in human populations by isoelectrofocusing on polyacrylamide gels. Demonstration of subtypes of the Gc allele and of additional Gc variants. *Hum Genet*, 1978, **41**(1): 53–60.

Cox, D.W., S. Smyth, and G. Billingsley, Three new rare variants of alpha-1-antitrypsin. *Hum Genet*, 1982, **61**(2): 123–126.

Divall, G.B., Studies on the use of isoelectric focusing as a method of phenotyping erythrocyte acid phosphatase. *Forensic Sci Int*, 1981, **18**(1): 67–78.

Divall, G.B., The esterase D polymorphism as revealed by isoelectric focusing in ultra-thin polyacrylamide gels. *Forensic Sci Int*, 1984, **26**(4): 255–267.

Divall, G.B. and M. Ismail, Studies and observations on the use of isoelectric focusing in ultra-thin polyacrylamide gels as a method of typing human red cell phosphoglucomutase. *Forensic Sci Int*, 1983, **22**(2–3): 253–263.

Dorrill, M. and P.H. Whitehead, The species identification of very old human blood-stains. *Forensic Sci Int*, 1979, **13**(2): 111–116.

Dykes, D.D. and H.F. Polesky, Review of isoelectric focusing for Gc, PGM1, Tf, and Pi subtypes: Population distributions. *Crit Rev Clin Lab Sci*, 1984, **20**(2): 115–151.

Frants, R.R. and A.W. Eriksson, Alpha-1-antitrypsin: Common subtypes of Pi M. *Hum Hered*, 1976, **26**(6): 435–440.

Gorg, A., W. Postel, and R. Westermeier, Ultrathin-layer isoelectric focusing in polyacrylamide gels on cellophane. *Anal Biochem*, 1978, **89**(1): 60–70.

Grunbaum, B.W. and P.L. Zajac, Rapid phenotyping of the group specific component by immunofixation on cellulose acetate. *J Forensic Sci*, 1977, **22**(3): 586–589.

Grunbaum, B.W. and P.L. Zajac, Phenotyping of erythrocyte acid phosphatase in fresh blood and in bloodstains on cellulose acetate. *J Forensic Sci*, 1978, **23**(1): 84–88.

Hoste, B., Group-specific component (Gc) and transferrin (Tf) subtypes ascertained by isoelectric focusing. A simple nonimmunological staining procedure for Gc. *Hum Genet*, 1979, **50**(1): 75–79.

Itoh, Y. and S. Matsuzawa, Detection of human hemoglobin A (HbA) and human hemoglobin F (HbF) in biological stains by microtiter latex agglutination-inhibition test. *Forensic Sci Int*, 1990, **47**(1): 79–89.

James, S. and Nordby, J.J., *Forensic Science: An Introduction to Scientific and Investigative Techniques*, 2005. Boca Raton, FL: CRC Press.

Jones, D.A., Blood samples: Probability of discrimination. *J Forensic Sci Soc*, 1972, **12**(2): 355–359.

Khalap, S. and G.B. Divall, Gm(5) grouping of dried bloodstains. *Med Sci Law*, 1979, **19**(2): 86–88.

Kido, A., et al., A stability study on Gc subtyping in bloodstains: Comparison by two different techniques. *Forensic Sci Int*, 1984, **26**(1): 39–43.

Kimura, H., et al., The typing of group-specific component (Gc protein) in human blood stains. *Forensic Sci Int*, 1983, **22**(1): 49–55.

Kipps, A.E., G_m and K_m typing in forensic science—A methods monograph. *J Forensic Sci Soc*, 1979, **19**(1): 27–47.

Kipps, A.E., V.E. Quarmby, and P.H. Whitehead, The detection of mixtures of blood and other body secretions in stains. *J Forensic Sci Soc*, 1978, **18**(3–4): 189–191.

Kueppers, F. and B. Harpel, Group-specific component (Gc) "subtypes" of Gc1 by isoelectric focusing in US blacks and whites. *Hum Hered*, 1979, **29**(4): 242–249.

Kuhnl, P. and W. Spielmann, A third common allele in the transferrin system, TfC3, detected by isoelectric focusing. *Hum Genet*, 1979, **50**(2): 193–198.

Kuhnl, P. and W. Spielmann, Transferrin: Evidence for two common subtypes of the TfC allele. *Hum Genet*, 1978, **43**(1): 91–95.

Kuhnl, P., U. Schmidtmann, and W. Spielmann, Evidence for two additional common alleles at the *PGM1* locus (phosphoglucomutase—E.C.: 2.7.5.1). A comparison by three different techniques. *Hum Genet*, 1977, **35**(2): 219–223.

Lamm, L.U., Family studies of red cell acid phosphatase types. Report of a family with the D variant. *Hum Hered*, 1970, **20**(3): 329–335.

Lincoln, P.J. and B.E. Dodd, An evaluation of factors affecting the elution of antibodies from bloodstains. *J Forensic Sci Soc*, 1973, **13**(1): 37–45.

Markert, C.L., The molecular basis for isozymes. *Ann N Y Acad Sci*, 1968, **151**(1): 14–40.

Miscicka, D., T. Dobosz, and S. Raszeja, Determination of phenotypes of phosphoglucomutase (PGM1) in bloodstains by cellulose acetate electrophoresis. *Z Rechtsmed*, 1977, **79**(4): 297–300.

Murch, R.S. and B. Budowle, Applications of isoelectric focusing in forensic serology. *J Forensic Sci*, 1986, **31**(3): 869–880.

Neilson, D.M., et al., Simultaneous electrophoresis of peptidase A, phosphoglucomutase, and adenylate kinase. *J Forensic Sci*, 1976, **21**(3): 510–513.

Olaisen, B., et al., The ESD polymorphism: Further studies of the *ESD2* and *ESD5* allele products. *Hum Genet*, 1981, **57**(4): 351–353.

Randall, T., W.A. Harland, and J.W. Thorpe, A method of phenotyping erythrocyte acid phosphatase by isoelectric focusing. *Med Sci Law*, 1980, **20**(1): 43–47.

Rees, B., T.J. Rothwell, and J. Bonnar, Correlation of phosphoglucomutase isoenzymes in blood and semen. *Lancet*, 1974, **1**(7861): 783.

Sensabaugh, G.F., Biochemical markers for individuality, in R. Saferstein (ed.), *Forensic Science Handbook*, 1982. Englewood Cliffs, NJ: Prentice Hall.

Shaler, R.C., Interpretation of Gm testing results: Two case histories. *J Forensic Sci*, 1982, **27**(1): 231–235.

Sonneborn, H.H., Comments on the determination of isoenzyme polymorphism (ADA, AK, 6-PGD, PGM) by cellulose acetate electrophoresis. *Humangenetik*, 1972, **17**(1): 49–55.

Sorensen, S.A., Agarose gel electrophoresis of the human red cell acid phosphatase. *Vox Sang*, 1974, **27**(6): 556–563.

Sorensen, S.A., Report and characterization of a new variant, EB, of human red cell acid phosphatase. *Am J Hum Genet*, 1975, **27**(1): 100–109.

Stedman, R., Human population frequencies in twelve blood grouping systems. *J Forensic Sci Soc*, 1972, **12**(2): 379–413.

Stolorow, M.D. and B.G. Wraxall, An efficient method to eliminate streaking in the electrophoretic analysis of haptoglobin in bloodstains. *J Forensic Sci*, 1979, **24**(4): 856–863.

Sutton, J.G. and R. Burgess, Genetic evidence for four common alleles at the phosphoglucomutase-1 locus (*PGM1*) detectable by isoelectric focusing. *Vox Sang*, 1978, **34**(2): 97–103.

Teige, B., et al., Forensic aspects of haptoglobin: Electrophoretic patterns of haptoglobin allotype products and an evaluation of typing procedure. *Electrophoresis*, 1988, **9**(8): 384–392.

Thogmartin, J.R., et al., Sickle cell trait-associated deaths: A case series with a review of the literature. *J Forensic Sci*, 2011, **56**(5): 1352–1360.

Turowska, B. and M. Bogusz, The rare silent gene Po of human red cell acid phosphatase in a second family in Poland. *Forensic Sci*, 1978, **11**(3): 175–176.

Twibell, J. and P.H. Whitehead, Enzyme typing of human hair roots. *J Forensic Sci*, 1978, **23**(2): 356–360.

Whitehead, P.H., L.A. King, and D.J. Werrett, New information from bloodstains. *Naturwissenschaften*, 1979, **66**(9): 446–451.

Whitehead, P.H., et al., The examination of bloodstains by Laurell electrophoresis (antigen-antibody crossed electrophoresis). *J Forensic Sci Soc*, 1970, **10**(2): 83–90.

Wraxall, B.G. and E.G. Emes, Erythrocyte acid phosphatase in bloodstains. *J Forensic Sci Soc*, 1976, **16**(2): 127–132.

Wrede, B., E. Koops, and B. Brinkmann, Determination of three enzyme polymorphisms in one electrophoretic step on horizontal polyacrylamide gels. *Humangenetik*, 1971, **13**(3): 250–252.

Yuasa, I. and K. Umetsu, Molecular aspects of biochemical markers. *Leg Med*, 2005, **7**: 251–254.

19

Variable Number Tandem Repeat Profiling

Tandem repeats are abundant in the human genome. *Minisatellites* were first defined as a class of tandem repeats in the 1980s. Some of these repeats share a GC-rich core sequence. Subsequently, tandem repeats with higher AT contents of core sequence have also been characterized. The minisatellites are also called *variable number tandem repeats* (*VNTRs*), as shown in Figure 19.1. The repeat unit length of a VNTR can range from several to hundreds of base pairs (bp). The tandem repeat arrays can be kilobases (kb, corresponding to 10^3 bp) long, and the numbers of tandem repeat units in some VNTR loci are highly variable, leading to variable lengths of DNA fragments. A genotype is defined by a particular number of tandem repeat units at a given locus.

Table 19.1 lists the common VNTR loci used for forensic testing. To achieve high discriminating power, the VNTR loci should not be linked, which means that they should be inherited independently of each other. For example, loci located on different chromosomes or far apart on the same chromosome can be used (Section 21.2). Many VNTR loci used for forensic applications are highly polymorphic, and as many as hundreds of different genotypes per locus can be observed among the population. The discriminating power of VNTR loci used for forensic testing can be measured by *population match probability* (P_m; Chapter 25). The lower the P_m, the less likely a match will occur between two randomly chosen individuals. A P_m of up to 10^{-12} can be achieved by testing several VNTR loci.

19.1 Restriction Fragment Length Polymorphism

VNTR profiling utilizes RFLP—the first historical method used in forensic DNA testing (Figure 19.2). It utilizes *restriction endonucleases* that recognize and cleave specific sites along the DNA sequence. Cleavage of a DNA sample with a particular restriction endonuclease results in a reproducible set of restriction fragments of various lengths. Appropriate restriction endonucleases should be selected so that the genomic DNA is cleaved at sites that flank the VNTR core repeat region. The resulting fragments are then separated according to their sizes by gel electrophoresis through a standard agarose gel (Chapter 9).

The DNA is then processed using the *Southern transfer and hybridization* technique. The DNA is denatured and transferred from the gel to a supporting matrix such as a nylon or nitrocellulose membrane. The DNA immobilized on the membrane is then hybridized with a labeled *probe*. Only bands of DNA that have complementary sequences to the probe are recognized by detection systems such as autoradiography (Section 19.1.4). Using the RFLP technique, the

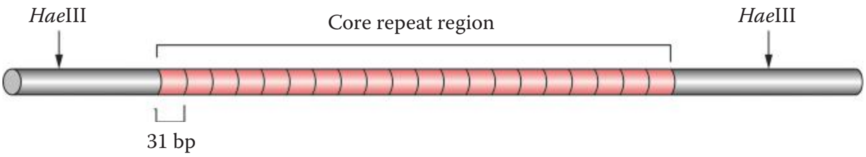

Figure 19.1 VNTR locus D2S44 (2q21.3–2q22). Each repeat unit consists of 31 bp. *Hae*III represents the *Hae*III restriction site.

Table 19.1 Common VNTR Loci

Locus	Chromosome Location	Repeat Unit Length (bp)	*Hae*III Fragment Size (kb)	Probe
D1S7	1	9	0.5–12	MS1
D2S44	2	31	0.7–8.5	yNH24
D4S139	4	31	2–12	pH30
D10S28	10	33	0.4–10	pTBQ7
D14S13	14	15	0.7–12	pCMM101
D16S85	16	17	0.2–5	3′HVR
D17S26	17	18	0.7–11	pEFD52
D17S79	17	38	0.5–3	V1

Source: Adapted from Budowle, B., et al., *DNA Typing Protocols: Molecular Biology and Forensic Analysis*, Eaton Publishing, Natick, MA, 2000; Office of Justice Programs, Future of forensic DNA testing: Predictions of the Research and Development Working Group. National Institute of Justice, US Department of Justice, 2000.

length variations among restriction sites can be detected. Most forensic applications focus on the length variations of VNTR regions located between two restriction sites.

In summary, the RFLP method includes several steps: (1) genomic DNA preparation, (2) restriction endonuclease digestion of the genomic DNA into fragments, (3) agarose gel electrophoretic separation of the DNA fragments according to size, (4) transfer of DNA fragments using Southern transfer, (5) hybridization with locus-specific probes, and (6) detection of locus-specific bands by autoradiography or chemiluminescence.

19.1.1 Restriction Endonuclease Digestion

Restriction endonucleases are enzymes that cleave the phosphodiester bond of DNA at or near specific recognition nucleotide sequences known as restriction sites. A restriction site usually is a short motif that is 4–8 bp in length. It often has a specific palindromic recognition sequence, that is, a segment of double-stranded DNA in which the nucleotide's sequence is identical with an inverted sequence in the complementary strand. Thus, double-stranded DNA is required to be cleaved by most restriction endonucleases. As a result, both sticky ends and blunt ends of restriction fragments can be generated after the cleavage (Figure 19.3).

To date, hundreds of restriction endonucleases have been described. They are traditionally classified into three types on the basis of subunit composition and enzymatic properties. Type II restriction endonucleases are most commonly used in molecular biology applications. Type II restriction endonucleases, requiring magnesium as a cofactor, usually cleave DNA at defined positions within their recognition sequences. The Enzyme Commission (EC) number of type II

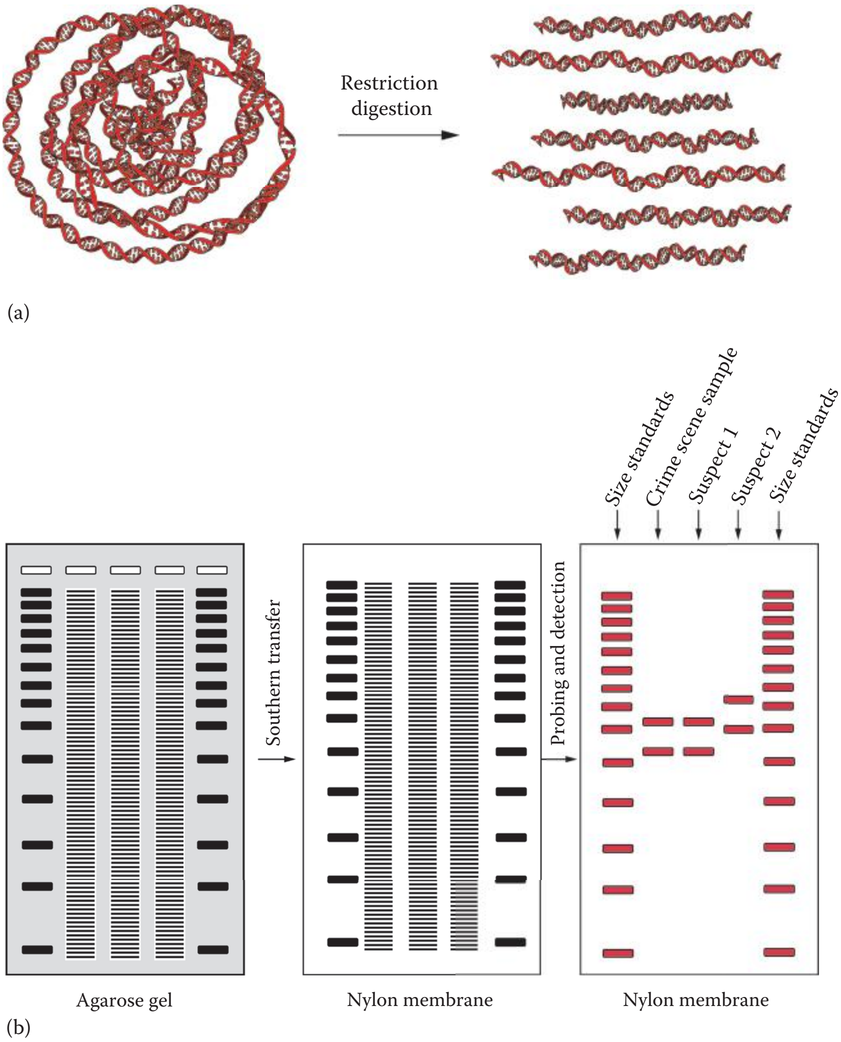
Restriction
digestion
(a)
Size standards
Crime scene sample
Suspect 1
Suspect 2
Size standards
Southern transfer
Probing and detection
Agarose gel
Nylon membrane
Nylon membrane
(b)

Figure 19.2 RFLP. (a) Restriction digestion generates restriction fragments with various lengths of genomic DNA. (b) Restriction fragments are separated by gel electrophoresis. DNA is transferred to a solid phase and probed. The signal is detected and the DNA fragment of interest can be observed. Band patterns of heterozygous loci of individuals are shown. (© Richard C. Li.)

restriction endonucleases is EC 3.1.21.4. The EC number is a numerical classification system of nomenclature based on the chemical reaction that is catalyzed by the enzyme. In contrast, type I (EC 3.1.21.3) and type III (EC 3.1.21.5) restriction endonucleases cleave at sites remote from their recognition site, which are not utilized for RFLP applications.

Restriction endonucleases are isolated from various bacteria. Each enzyme is named using a nomenclature system after the bacterium from which it was isolated. For example, in the restriction enzyme *Hae*III, H is from the genus name *Haemophilus*, ae is from the species name *aegypticus*, and III stands for the third endonuclease isolated from the *Haemophilus aegyptius* bacteria.

The restriction endonucleases play a role in protecting the bacteria from phage (bacterial virus) infections by using their endonucleases to destroy foreign DNA molecules. Bacterial DNA is usually methylated. In prokaryotes, the *Dam* methylase transfers a methyl group to the N6 position of the adenine in the sequence GmATC and the *Dcm* methylase transfers a methyl

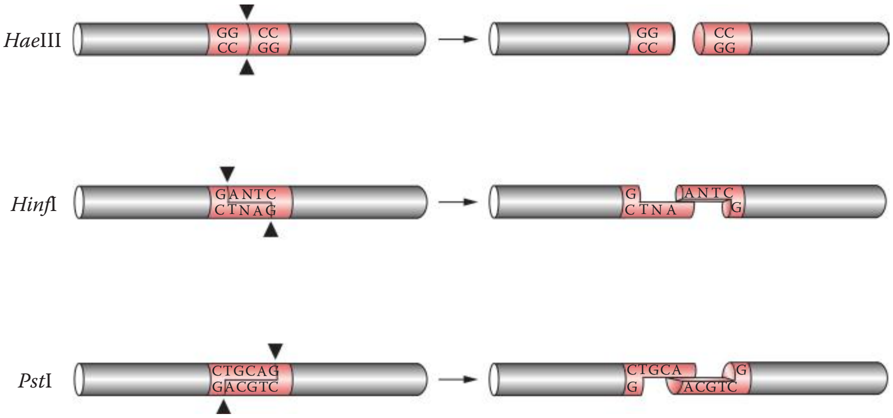

Figure 19.3 Restriction sites for *Hae*III, *Hinf*I, and *Pst*I. *Hae*III digestion produces a blunt end DNA fragment. *Hinf*I and *Pst*I digestions produce sticky ends. *N* represents any nucleotide. (© Richard C. Li.)

group to the C5 position of cytosine in the sequences CmCAGG and CmCTGG. The activities of restriction endonucleases can be influenced by DNA methylation. Many restriction endonucleases cannot cleave methylated DNA. Therefore, bacterial DNA is distinguishable from foreign DNA by the bacteria's restriction endonucleases. This phenomenon protects bacterial DNA from digestion by their own endonucleases. Note that because DNA methylation also occurs in the human genome (Chapter 11), it is important to choose the restriction endonucleases that are not affected by the methylation of human genomic DNA for RFLP analysis.

The type II restriction endonucleases were used in RFLP analysis for forensic DNA testing (Figure 19.3). In order to perform this analysis, the preferred restriction endonucleases for RFLP were those that cleave at the flanking regions of VNTR repeat units but not within the core repeat sequences of the VNTR. Most forensic laboratories used a single restriction endonuclease for a panel of VNTR loci because the DNA in evidence samples was often insufficient for performing multiple tests with different restriction endonucleases. For instance, *Hinf*I-based RFLP was commonly used in European forensic laboratories, and *Hae*III-based RFLP was used in North American and some European forensic laboratories. Other restriction endonuclease-based RFLPs such as *Pst*I were also used.

The use of these common restriction endonucleases allows the comparison of data of various laboratories. Several VNTR loci are suitable for RFLP analysis with these restriction endonucleases. For example, *Hae*III presents several advantages for forensic RFLP analysis. It recognizes a four-base sequence, 5′-GGCC-3′, and cleaves the DNA between the internal G and C residues of the recognition site (GG/CC). *Hinf*I recognizes a five-base restriction site, and *Pst*I recognizes a six-base restriction site. Hypothetically, four-base restriction sites are likely to occur more often than five- and six-base restriction sites in the human genome. Thus, *Hae*III restriction sites occur more frequently than *Hinf*I and *Pst*I sites. As a result, *Hae*III-cleaved DNA fragments are smaller than those of *Hinf*I and *Pst*I. The *Hae*III-generated VNTR allele sizes are easier to separate using conventional agarose gels, also called analytic gel electrophoresis. After electrophoresis, a smear of various sizes of DNA fragments can be observed. The analytic gel is then processed for Southern transfer. Moreover, the enzymatic activity of *Hae*III is not affected by the methylation of human genomic DNA. Its enzymatic activities also appear unaffected when a reaction proceeds under nonoptimal conditions. Additionally, low star activity is observed (Section 19.1.5.2.2).

19.1.2 Southern Transfer

Also known as *Southern blotting*, this technique was named after Sir Edwin Southern, who developed it in the United Kingdom in the mid-1970s. The method can be used to transfer

DNA from an agarose gel to a solid matrix so that it can be detected with a hybridization probe (Section 19.1.3). This method is still used today in many research laboratories. Prior to the transfer of DNA, the DNA in the gel must be denatured, under alkaline conditions such as treatment with sodium hydroxide, into single-stranded DNA. The single-stranded DNA in the gel is then transferred by capillary action to a solid matrix such as a piece of nylon membrane. The single-stranded DNA fragments transferred can be immobilized on a nylon membrane by an ultraviolet cross-linking process (Figure 19.4).

19.1.3 Hybridization with Probes

A hybridization probe of RFLP is a small segment of labeled DNA that is usually several hundred to a thousand bases in length containing the VNTR sequence. It is utilized to detect the presence of much longer target DNA sequences, in this case the VNTR sequences that are complementary to the nucleotide sequences of the probe. The probe is first denatured by heating or by exposure to alkaline conditions into single-stranded DNA. The hybridization process allows complementary pairing between the probe and the target sequence. Two types of probe techniques were developed for VNTR analysis: the multilocus probe and single-locus probe techniques.

19.1.3.1 Multilocus Probe Technique

The *multilocus probe* (*MLP*) technique can detect multiple VNTR loci simultaneously (Figure 19.5). Some VNTRs in the human genome share a short GC-rich core sequence of 10–15 bp. The MLP consists of this core sequence and hybridizes to multiple VNTRs that share these core sequences. As a result, the utilization of MLP produces a complex bar-code-like band pattern from alleles of multiple VNTR loci (Figure 19.6).

The MLP technique was pioneered by Sir Alec Jeffreys in 1984 at the University of Leicester in the United Kingdom and was called *DNA fingerprinting*. Because of its excellent discriminating power, the method was used for parentage testing in immigration disputes with great success.

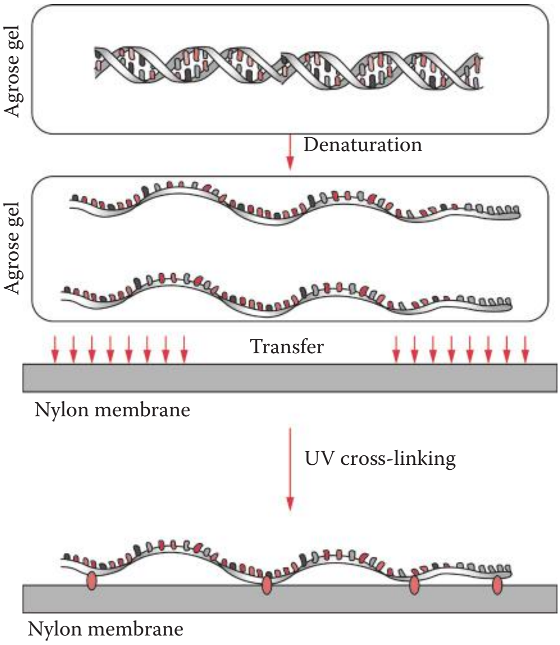

Figure 19.4 Southern blotting. DNA in agarose gel is denatured into single-stranded DNA and transferred to a solid-phase membrane where the single-stranded DNA is immobilized by ultraviolet cross-linking. (© Richard C. Li.)

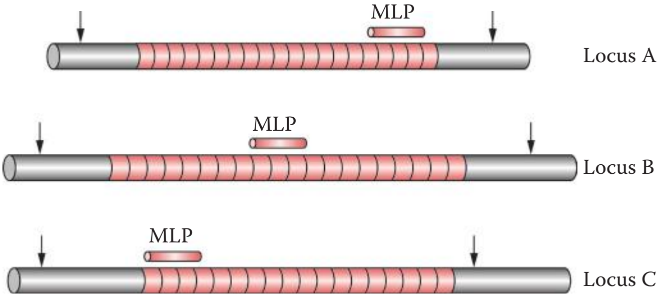

Figure 19.5 VNTR analysis using the MLP method. The technique can detect multiple VNTR loci simultaneously. Restriction sites are indicated by arrows. (© Richard C. Li.)

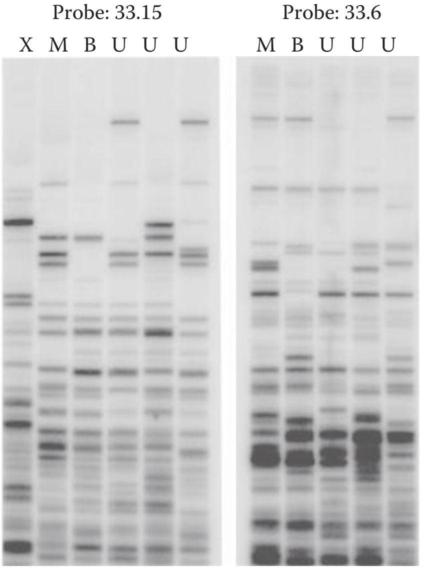

Figure 19.6 First application of DNA fingerprinting. The MLP method was used to analyze samples for an immigration case. M, mother; U, three undisputed children; B, male child in dispute; X, an unrelated individual. All bands in B can be traced back to M or U. (From Jeffreys, A.J., *Nat Med*, 11, 1035, 2005. With permission.)

However, one of the disadvantages of the MLP approach is that the interpretation of a mixed DNA sample from more than one individual is nearly impossible due to its complex DNA fingerprinting patterns. Therefore, MLP analysis was not widely utilized in forensic DNA laboratories.

19.1.3.2 Single-Locus Probe Technique

To resolve the disadvantages encountered in the MLP technique, probes that recognize the genomic DNA at the flanking regions of specific VNTR loci can be used. The probe only hybridizes to a single VNTR locus and the technique is called the *single-locus probe* (*SLP*), as depicted in Figure 19.7. SLP generates a simple pattern called a *DNA profile*, consisting of one band for a homozygote and two bands for a heterozygote per locus. In order to improve the discriminating power of the test, SLP analyses of different VNTR loci can be performed by using different probes sequentially with a single locus at one time. SLP can analyze mixed DNA samples from two or more contributors. The sizes of fragments can be estimated and converted into a numerical form suitable for databasing. Therefore, DNA profiles can be compared among different laboratories.

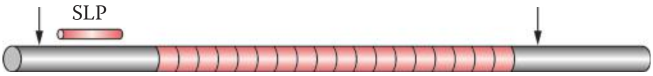

Figure 19.7 VNTR analysis using the SLP method. The technique can detect a single VNTR locus. Restriction sites are indicated by arrows. (© Richard C. Li.)

This technique led to the solving of a double murder case in Leicestershire in the 1980s. The case was the first to apply DNA evidence to a criminal investigation. DNA profiling identified the true perpetrator and also excluded an innocent suspect (Figure 19.8). In 1983 and 1986, two girls were raped and murdered. Crime scene evidence suggested that the two cases were committed by the same perpetrator. A young local man, Richard Buckland, was the suspect. However, the DNA evidence revealed that the semen samples from both crimes did not originate from Buckland. To solve the crimes, an investigation was carried out in which 5000 local men were asked to volunteer DNA samples for testing. Several months after the investigation, a witness tipped off police that a man named Pitchfork had paid someone for giving a blood sample as Pitchfork's. In 1987, Pitchfork was arrested. It was discovered that Pitchfork's DNA profile matched that of the crime scene evidence. He was sentenced to life imprisonment. This case demonstrated the great potential of DNA profiling in forensic investigations. Consequently, SLP became a common method in most forensic laboratories in the late 1980s–1990s.

19.1.4 Detection

To detect VNTR loci, a labeled SLP probe is hybridized to the target sequence of DNA, which has been immobilized on a solid matrix such as a piece of nylon membrane (see Southern

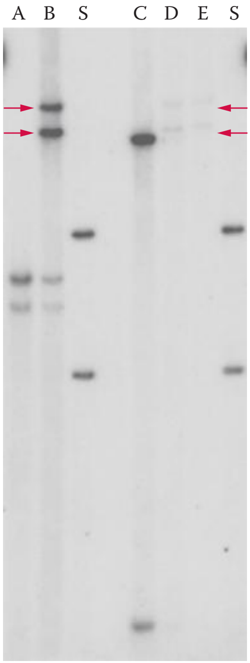

Figure 19.8 First application of DNA profiling in a criminal investigation using the SLP method. A, hair roots from the first victim; B, a mixture of semen and vaginal fluid from the first victim; C, blood from second victim; D, a vaginal swab from the second victim; E, a semen stain on clothing from the second victim; S, blood from the suspect. Alleles (arrows) are matched with the profiles of the two cases but not with the suspect profile. (From Jeffreys, A.J., *Nat Med*, 11, 1035, 2005. With permission.)

transfer, Section 19.1.2). Any unbound probes are washed away so that they do not interfere with the signal. Two types of detection systems are used for VNTR analysis. Radioisotope labeling, such as with a ^{32}P-labeled probe, can be used. The hybridized probe can be detected by exposing the membrane to a sheet of x-ray film to generate an autoradiograph. Alternatively, an enzyme-conjugated probe can also be used. Alkaline phosphatase (Section 9.2.2) is an example of an enzyme used in this type of probe. Its enzymatic activity can be detected with chemiluminescent substrates. A chemiluminescent signal can be detected by exposure to x-ray film as well.

In RFLP analysis, several loci are commonly analyzed sequentially using the same membrane. This approach has the advantage of not consuming additional DNA samples, which are often limited in forensic cases. When probing for multiple loci, multiple probes for each locus are sequentially hybridized and removed one at a time. Once the analysis of the first probe is completed, the probe is removed by a procedure called probe stripping, which is carried out under conditions such as high-temperature washing to denature the DNA strands (Chapter 17). The probe for the next locus to be analyzed is then hybridized to the same membrane, and the process is repeated for each probe to be tested.

Typically, a size standard is utilized on each gel. Band sizes can thus be estimated by comparison to these standards. However, the VNTR alleles that differ by only one or two repeat units are usually not distinguishable. For this reason, genotypes can be determined by bins but not discrete alleles. A *bin* is a range of DNA fragments that differ by only a few repeat units. A sample with a known VNTR genotype is also utilized on each gel as a positive control where historically a genomic DNA sample from cell line K562 (a human erythroleukemic cell line) was used for the positive control. DNA samples to be compared can be loaded side by side on the same gel. As a result, the patterns of VNTR fragments can be compared from sample to sample. The following possible conclusions can be made. If the VNTR fragments are at corresponding positions (profiles match), they are considered to be a match (inclusion). Chapter 25 evaluates and discusses the strengths of the results. If the profiles are different, the two DNA samples are considered to have come from different origins (exclusion).

19.1.5 Factors Affecting RFLP Results

The accuracy of VNTR profiling results can be affected by certain factors such as sample conditions, genetic mutations, and experimental artifacts appearing during the procedure. Consequently, these factors can impact data interpretation and are explained in the following sections.

19.1.5.1 DNA Degradation

RFLP analysis requires the genomic DNA to be intact. DNA degradation results in damage such as creating nicks and breaks in the strand. The more severe the degradation, the smaller the average size of the DNA fragments. When the average size of DNA fragments becomes too small, the allele may not be detected. Many VNTR tandem arrays can span several kilobases in length. In theory, large alleles are more likely to be affected by degradation than smaller alleles at a different locus.

A two-banded heterozygous profile can be observed as a one-banded homozygous profile if the larger band is not detected due to degradation. This artifact could lead to a false determination of exclusion. However, DNA degradation can be detected prior to conducting RFLP by the use of agarose gel electrophoresis, also known as a *yield gel*, used for evaluating the yield and integrity of the isolated genomic DNA. High-molecular-weight genomic DNA bands are usually observed for a typical genomic DNA sample. In contrast, a smear of low-molecular-weight DNA bands can be observed if DNA is degraded. The sizes of the DNA can be estimated by comparison to a size standard run on the same gel. Additionally, the yield of DNA can be estimated by comparing the intensity of the size standards with a known quantity of DNA.

19.1.5.2 Restriction Digestion–Related Artifacts

19.1.5.2.1 Partial Restriction Digestion

Complete restriction digestion should be achieved for RFLP analysis. If partial digestion occurs, the partially cleaved DNA strands are longer than the cleaved fragments (Figure 19.9). Thus, partial digestion results in a mixture of fragments with correct sizes and slightly larger fragments. Under these conditions, a larger uncleaved band, usually lower in intensity than the true bands, can be observed. The multibanded pattern due to the partial digestion can be observed at multiple loci analyzed in the same nylon membrane.

Detection of more than two bands in an RFLP profile may lead to a false interpretation and be incorrectly concluded to be a mixture. However, partial digestion can be detected after restriction digestion. DNA cleavage by restriction endonuclease digestion can be examined using agarose gel electrophoresis. A small portion of a sample can be analyzed. After separation using electrophoresis, a smear of various sizes of cleaved DNA fragments can usually be observed if restriction digestion is completed. Conversely, high-molecular-weight genomic DNA can still be observed from partially digested DNA samples. Additionally, comparisons can be made between the sample and uncleaved and completely cleaved standard samples of DNA. This quick assay is also called a *test gel*, used to determine if the DNA was cleaved to completion.

If partial digestion occurs, procedures such as additional purification of the DNA sample can be carried out. Additionally, optimal amounts of DNA, restriction enzymes, buffer, and proper incubation conditions should be used in achieving complete digestion.

19.1.5.2.2 Star Activity

Star activity refers to a deviation of the specificity of a cleavage site of a restriction endonuclease under certain conditions, such as a high concentration ratio of enzyme to DNA, the use of non-optimal buffers for restriction digestion, prolonged digestion time, the substitution of Mg^{2+} with other divalent cations, and the presence of organic solvents such as ethanol and concentrated glycerol. For instance, *Hae*III cleaves at the GGCC DNA sequence. When star activity occurs, the enzymatic specificity is reduced and cleaves at a sequence slightly different from GGCC. If the start site is presented at an internal location of a VNTR locus, the enzyme would cleave the GGCC sequences and additionally cleave at the internal star site (Figure 19.10). This would result in an additional band smaller than the true alleles, although the intensity of this band is usually not the same as other bands. Thus, a multiband pattern is observed. However, the star

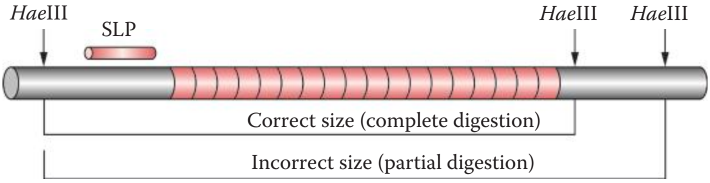
*Hae*III
SLP
*Hae*III
*Hae*III
Correct size (complete digestion)
Incorrect size (partial digestion)

Figure 19.9 Effects of partial restriction digestion on the RFLP profile. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites are indicated by arrows. (© Richard C. Li.)

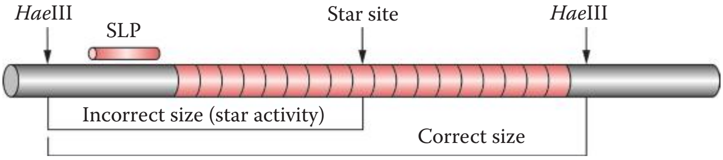

Figure 19.10 Star effects on RFLP profiles. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites and star sites are indicated by arrows. (© Richard C. Li.)

site presented externally to a VNTR locus cannot be detected by the probe and does not affect the profiling results. Star activity can usually be avoided by carrying out restriction digestion reactions under conditions recommended by their manufacturers.

19.1.5.2.3 Point Mutations

A *point mutation* is caused by the substitution, deletion, or insertion of a single nucleotide. Point mutations at a restriction site within flanking regions may abolish the site, and the result is a band slightly larger than the true allele (Figure 19.11). The point mutation may also be present internally in a VNTR sequence. If such a point mutation creates a restriction enzyme site, the enzyme will cleave at the regular site and at the mutation site and yield two smaller bands. If the created restriction site is located internal to the probe binding region, both bands will be detected for that allele. These rare mutations obey Mendelian inheritance.

19.1.5.3 Electrophoresis and Blotting Artifacts

19.1.5.3.1 Partial Stripping

If more than one VNTR locus is analyzed sequentially using the same membrane, a probe must be removed by the stripping process before the application of the next probe. Any probe remaining on the membrane due to partial stripping may generate additional bands when the next probe is analyzed. However, bands due to partial stripping are usually faint and have the same electrophoretic mobility as the previous autoradiograph.

19.1.5.3.2 Separation Resolution Limits and Band Shifting

Agarose gel electrophoresis cannot resolve restriction fragments that differ by one or a few repeat units, especially for high-molecular-weight fragments. These bands may not be

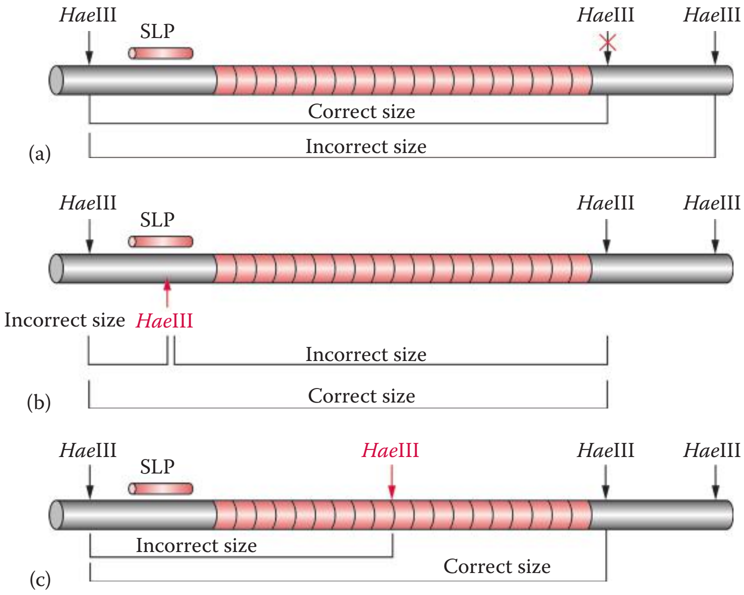

Figure 19.11 Effects of point mutations on the RFLP profile. Only the restriction fragments detectable by the probe are shown. *Hae*III restriction sites are indicated by arrows. (a) Point mutation (in red) abolishes the *Hae*III restriction site. (b) A point mutation (in red) creates an internal *Hae*III restriction site residing within the probe-binding region. (c) A point mutation (in red) creates an internal *Hae*III restriction site residing outside the probe-binding region. (© Richard C. Li.)

separated and will appear as a single band. This may lead to a false interpretation as a homozygous profile. Additionally, minor variations in the electrophoretic mobility of DNA fragments, known as band shifting, can cause two samples from the same individual to appear different.

19.1.5.3.3 Bands Running off Gel

The commonly used VNTR loci generate bands from hundreds of base pairs to 20 kb in length. The small bands have higher electrophoretic mobility and may run off the front edge of a gel during electrophoresis and fail to be detected. This phenomenon may also lead to a false interpretation as a homozygous profile. To prevent these DNA fragments from running off the gel into the buffer tank, a longer gel can be used. Alternatively, the electrophoresis can be stopped before the dye front, the furthest extent that dyes migrate, reaches the front edge of the gel.

19.2 Amplified Fragment Length Polymorphism

The RFLP analysis of VNTR profiling does not perform well for degraded or limited quantities of DNA from crime scene samples. For these reasons, an improved VNTR method was developed. Some VNTR loci have relatively short alleles (<1 kb). These loci are suitable for PCR amplification. This technique is called *amplified fragment length polymorphism* (*AFLP*). One locus, D1S80, was used by forensic DNA laboratories for AFLP analysis. Fragments in the range of 14–42 repeat units (16 bp per repeat) were amplified using the AFLP method (Figure 19.12). The amplified DNA fragments were commonly separated according to size using polyacrylamide gel electrophoresis and detected using a silver stain (Figure 19.13).

D1S80 loci are detected as discrete alleles and thus can be compared directly to an allelic ladder (a collection of common alleles used as a standard) on the same gel. This technique represented an improvement over the RFLP system. RFLP allele sizing cannot be performed with precision and the resolution limits of agarose gel electrophoresis are much lower than those of the polyacrylamide gels.

The AFLP technique requires less DNA than the RFLP method and performs better for degraded samples. The AFLP method at the D1S80 locus can be analyzed in a multiplex fashion with an amelogenin locus (Chapter 21). The amelogenin gene is used for forensic sex-typing applications. Typing the amelogenin gene enables the determination of the sex of the contributor of a biological sample.

Due to the wide variation in allele sizes at the D1S80 locus, preferential amplification may be observed. Under certain conditions, the larger alleles may not be as consistently amplified as the small alleles, which may cause lower signal intensity of the larger allele. Additionally, only one locus is analyzed in this system. Furthermore, the D1S80 locus contains two alleles that are very common in some populations. Thus, the discriminating power is reduced compared to RFLP. D1S80 was gradually replaced by multiplex STR systems in the late 1990s.

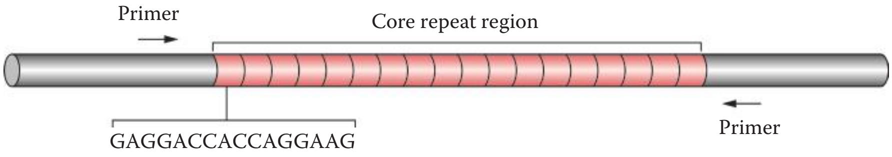

Figure 19.12 VNTR locus D1S80 (chromosome 1p). Each repeat unit is 16 bp long. PCR primers are indicated to amplify the core repeat region.

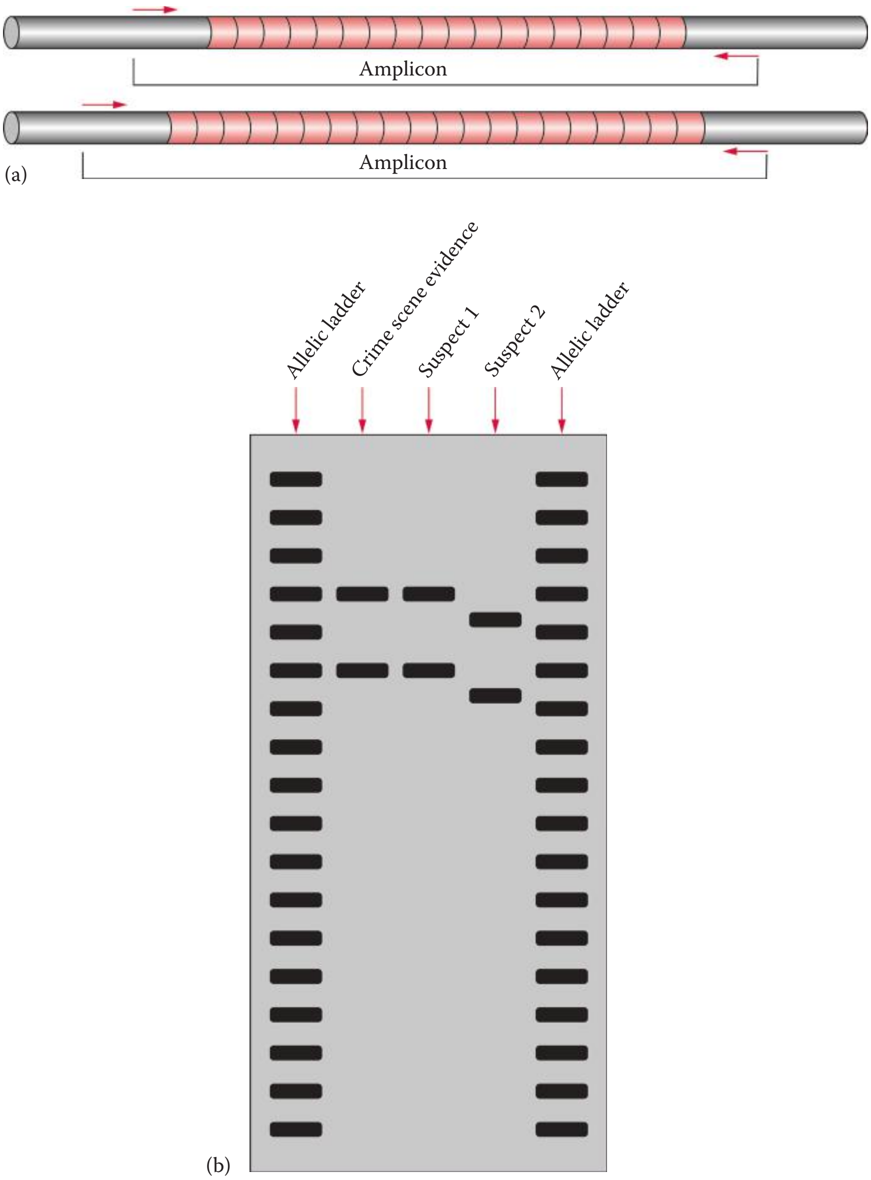
Amplicon
Amplicon
(a)
Allelic ladder
Crime scene evidence
Suspect 1
Suspect 2
Allelic ladder
(b)

Figure 19.13 AFLP analysis. (a) Heterozygous D1S80 loci amplified by PCR. PCR primers are indicated as *arrows*. (b) Silver-stained polyacrylamide gel showing D1S80 amplicons along with allelic ladders. (© Richard C. Li.)

Bibliography

RFLP

Balazs, I., et al., Human population genetic studies of five hypervariable DNA loci. *Am J Hum Genet*, 1989, **44**(2): 182–190.

Benzinger, E.A., et al., Time course and inhibitors of *Hae*III digestion in the forensic laboratory. *Appl Theor Electrophor*, 1995, **4**(4): 179–188.

Benzinger, E.A., et al., Products of partial digestion with *Hae*III. Part 1. Characterization, casework experience and confirmation of the theory of three-, four- and five-banded RFLP pattern origins using partial digestion. *J Forensic Sci*, 1997, **42**(5): 850–863.

Benzinger, E.A., et al., An illustrated guide to RFLP troubleshooting. *J Forensic Sci*, 1998, **43**(3): 665–679.

Bragg, T., et al., Isolation and mapping of a polymorphic DNA sequence (cTBQ7) on chromosome 10 [D10S28]. *Nucleic Acids Res*, 1988, **16**(23): 11395.

Budowle, B., et al., *Hae*III—A suitable restriction endonuclease for restriction fragment length polymorphism analysis of biological evidence samples. *J Forensic Sci*, 1990, **35**(3): 530–536.

Budowle, B., et al., Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. *Am J Hum Genet*, 1991, **48**(5): 841–855.

Budowle, B., et al., Simple protocols for typing forensic biological evidence: Chemiluminescent detection for human DNA quantitation and restriction fragment length polymorphism (RFLP) analyses and manual typing of polymerase chain reaction (PCR) amplified polymorphisms. *Electrophoresis*, 1995, **16**(9): 1559–1567.

Budowle, B., et al., The assessment of frequency estimates of *Hae*III-generated VNTR profiles in various reference databases. *J Forensic Sci*, 1994, **39**(2): 319–352.

Budowle, B., et al., *DNA Typing Protocols: Molecular Biology and Forensic Analysis*, 2000. Natick, MA: Eaton Publishing.

Duewer, D.L., K.L. Richie, and D.J. Reeder, RFLP band size standards: NIST standard reference material 2390. *J Forensic Sci*, 2000, **45**(5): 1093–1105.

Duewer, D.L., et al., Interlaboratory comparison of autoradiographic DNA profiling measurements. 4. Protocol effects. *Anal Chem*, 1997, **69**(10): 1882–1892.

Elder, J.K. and E.M. Southern, Measurement of DNA length by gel electrophoresis II: Comparison of methods for relating mobility to fragment length. *Anal Biochem*, 1983, **128**(1): 227–231.

Eriksen, B. and O. Svensmark, Comparison of two molecular weight markers used in DNA-profiling. *Int J Legal Med*, 1992, **105**(3): 145–148.

Gary, K.T., D.L. Duewer, and D.J. Reeder, Graphical tools for RFLP measurement quality assurance: Laboratory performance charts. *J Forensic Sci*, 1999, **44**(5): 978–982.

Gill, P. and D.J. Werrett, Exclusion of a man charged with murder by DNA fingerprinting. *Forensic Sci Int*, 1987, **35**(2–3): 145–148.

Giusti, A.M. and B. Budowle, Effect of storage conditions on restriction fragment length polymorphism (RFLP) analysis of deoxyribonucleic acid (DNA) bound to positively charged nylon membranes. *J Forensic Sci*, 1992, **37**(2): 597–603.

Hartmann, J.M., et al., The effect of sampling error and measurement error and its correlation on the estimation of multi-locus fixed-bin VNTR RFLP genotype probabilities. *J Forensic Sci*, 1997, **42**(2): 241–245.

Hau, P. and N. Watson, Sequencing and four-state minisatellite variant repeat mapping of the D1S7 locus (MS1) by fluorescence detection. *Electrophoresis*, 2000, **21**(8): 1478–1483.

Jeffreys, A.J., Genetic fingerprinting. *Nat Med*, 2005, **11**: 1035.

Jeffreys, A.J., V. Wilson, and S.L. Thein, Hypervariable "minisatellite" regions in human DNA. *Biotechnology*, 1985, **24**: 467–472.

Jeffreys, A.J., V. Wilson, and S.L. Thein, Individual-specific "fingerprints" of human DNA. *Nature*, 1985, **316**(6023): 76–79.

Jeffreys, A.J., M. Turner, and P. Debenham, The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework. *Am J Hum Genet*, 1991, **48**(5): 824–840.

Kanter, E., et al., Analysis of restriction fragment length polymorphisms in deoxyribonucleic acid (DNA) recovered from dried bloodstains. *J Forensic Sci*, 1986, **31**(2): 403–408.

Kasai, K., Y. Nakamura, and R. White, Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. *J Forensic Sci*, 1990, **35**(5): 1196–1200.

Laber, T.L., et al., Evaluation of four deoxyribonucleic acid (DNA) extraction protocols for DNA yield and variation in restriction fragment length polymorphism (RFLP) sizes under varying gel conditions. *J Forensic Sci*, 1992, **37**(2): 404–424.

Laber, T.L., et al., Validation studies on the forensic analysis of restriction fragment length polymorphism (RFLP) on LE agarose gels without ethidium bromide: Effects of contaminants, sunlight, and the electrophoresis of varying quantities of deoxyribonucleic acid (DNA). *J Forensic Sci*, 1994, **39**(3): 707–730.

Laber, T.L., et al., The evaluation and implementation of match criteria for forensic analysis of DNA. *J Forensic Sci*, 1995, **40**(6): 1058–1064.

Lander, E.S., DNA fingerprinting on trial. *Nature*, 1989, **339**(6225): 501–505.

Lewis, M.E., et al., Restriction fragment length polymorphism DNA analysis by the FBI laboratory protocol using a simple, convenient hardware system. *J Forensic Sci*, 1990, **35**(5): 1186–1190.

McNally, L., et al., The effects of environment and substrata on deoxyribonucleic acid (DNA): The use of casework samples from New York City. *J Forensic Sci*, 1989, **34**(5): 1070–1077.

Milner, E.C., et al., Isolation and mapping of a polymorphic DNA sequence pH30 on chromosome 4 [HGM provisional no. D4S139]. *Nucleic Acids Res*, 1989, **17**(10): 4002.

Nakamura, Y., et al., Isolation and mapping of a polymorphic DNA sequence pYNH24 on chromosome 2 (D2S44). *Nucleic Acids Res*, 1987, **15**(23): 10073.

Office of Justice Programs, Future of forensic DNA testing: Predictions of the Research and Development Working Group, 2000. National Institute of Justice, US Department of Justice.

Rankin, D.R., et al., Restriction fragment length polymorphism (RFLP) analysis on DNA from human compact bone. *J Forensic Sci*, 1996, **41**(1): 40–46.

Reed, K.C. and D.A. Mann, Rapid transfer of DNA from agarose gels to nylon membranes. *Nucleic Acids Res*, 1985, **13**(20): 7207–7221.

Richie, K.L., et al., Long PCR for VNTR analysis. *J Forensic Sci*, 1999, **44**(6): 1176–1185.

Rossi, U., International recommendations (as of March, 1988) on the application of methods involving DNA-polymorphisms in forensic haematology. *Haematologica*, 1989, **74**(2): 219–221.

Sharma, B.R., et al., A comparative study of genetic variation at five VNTR loci in three ethnic groups of Houston, Texas. *J Forensic Sci*, 1995, **40**(6): 933–942.

Stolorow, A.M., et al., Interlaboratory comparison of autoradiographic DNA profiling measurements. 3. Repeatability and reproducibility of restriction fragment length polymorphism band sizing, particularly bands of molecular size >10K base pairs. *Anal Chem*, 1996, **68**(11): 1941–1947.

Tahir, M.A., et al., Restriction fragment length polymorphism (RFLP) typing of DNA extracted from nasal secretions. *J Forensic Sci*, 1995, **40**(3): 459–463.

Tahir, M.A., et al., Deoxyribonucleic acid profiling by restriction fragment length polymorphism analysis—A compilation of validation studies. *Sci Justice*, 1996, **36**(3): 173–182.

Tully, G., K.M. Sullivan, and P. Gill, Analysis of 6 VNTR loci by "multiplex" PCR and automated fluorescent detection. *Hum Genet*, 1993, **92**(6): 554–562.

Wyman, A.R. and R. White, A highly polymorphic locus in human DNA. *Proc Natl Acad Sci U S A*, 1980, **77**(11): 6754–6758.

AFLP

Baechtel, F.S., K.W. Presley, and J.B. Smerick, D1S80 typing of DNA from simulated forensic specimens. *J Forensic Sci*, 1995, **40**(4): 536–545.

Baechtel, F.S., et al., Multigenerational amplification of a reference ladder for alleles at locus D1S80. *J Forensic Sci*, 1993, **38**(5): 1176–1182.

Budowle, B., et al., Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. *Am J Hum Genet*, 1991, **48**(1): 137–144.

Cosso, S. and R. Reynolds, Validation of the AmpliFLP D1S80 PCR amplification kit for forensic casework analysis according to TWGDAM guidelines. *J Forensic Sci*, 1995, **40**(3): 424–434.

Gross, A.M., G. Carmody, and R.A. Guerrieri, Validation studies for the genetic typing of the D1S80 locus for implementation into forensic casework. *J Forensic Sci*, 1997, **42**(6): 1140–1146.

Kasai, K., Y. Nakamura, and R. White, Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. *J Forensic Sci*, 1990, **35**(5): 1196–1200.

Kloosterman, A.D., B. Budowle, and P. Daselaar, PCR-amplification and detection of the human D1S80 VNTR locus. Amplification conditions, population genetics and application in forensic analysis. *Int J Legal Med*, 1993, **105**(5): 257–264.

Koseler, A., A. Atalay, and E.O. Atalay, Allele frequency of VNTR locus D1S80 observed in Denizli Province of Turkey. *Biochem Genet*, 2009, **47**(7–8): 540–546.

Latorra, D. and M.S. Schanfield, Analysis of human specificity in AFLP systems APOB, PAH, and D1S80. *Forensic Sci Int*, 1996, **83**(1): 15–25.

Nakamura, Y., et al., Isolation and mapping of a polymorphic DNA sequence (pMCT118) on chromosome 1p [D1S80]. *Nucleic Acids Res*, 1988, **16**(19): 9364.

Sajantila, A., et al., Amplification of reproducible allele markers for amplified fragment length polymorphism analysis. *Biotechniques*, 1992, **12**(1): 16, 18, 20–22.

Skinker, D.M., et al., DNA typing of azoospermic semen at the D1S80 locus. *J Forensic Sci*, 1997, **42**(4): 718–720.

Skowasch, K., P. Wiegand, and B. Brinkmann, pMCT 118 (D1S80): A new allelic ladder and an improved electrophoretic separation lead to the demonstration of 28 alleles. *Int J Legal Med*, 1992, **105**(3): 165–168.

Walsh, S.J. and C. Eckhoff, Australian Aboriginal population genetics at the D1S80 VNTR locus. *Ann Hum Biol*, 2007, **34**(5): 557–565.

Watanabe, G., et al., Nucleotide substitution in the 5′ flanking region of D1S80 locus. *Forensic Sci Int*, 1997, **89**(1–2): 75–80.

20

Autosomal Short Tandem Repeat Profiling

A *short tandem repeat* (STR) is a region of genomic DNA containing an array of short repeating sequences. STRs are also called *microsatellites* or *simple sequence repeats*. A STR repeat unit can be several base pairs (bp) in length. Arrays range from several to approximately a hundred *repeat units*, which are the component of repetition. The number of STR repeat units varies among individuals. The most commonly used STR loci are 100–500 bp in length, which are shorter than the smallest variable number tandem repeats (VNTRs) (approximately 1000 bp). Thus, STR loci have many advantages compared to VNTR loci:

- STR loci can be amplified by PCR.
- STR profiling can be carried out for degraded DNA samples.
- Preferential amplification is reduced at STR loci.
- The resolution of electrophoretic separation of STR fragments is superior.
- STR loci are suitable for multiplex amplification.

Additionally, STR profiling, as with VNTR profiling, is suitable for the interpretation of mixed DNA profiles from multiple individuals. Thus, STR loci are better candidates for forensic DNA testing than VNTR loci. This chapter will discuss autosomal STR profiling. Male-specific Y chromosomal STR will be discussed in Chapter 21.

20.1 Characteristics of STR Loci

More than 10^5 STRs exist in the human genome. Many STRs have been characterized and used in various types of studies such as genetic mapping and linkage analysis. Some STRs have been characterized specifically for forensic DNA profiling.

20.1.1 Core Repeat and Flanking Regions

The *core repeat region* of each STR locus contains tandemly repeated sequences. The designation of genotypes for human identification is based on the number of tandem repeat units at a STR locus, which varies among human individuals (Figure 20.1). The *flanking regions* surrounding the core repeat region are also needed for STR analysis. PCR primers complementary to these flanking regions are used, allowing the core repeat regions to be amplified.

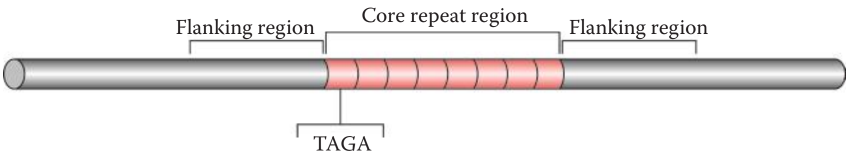

Figure 20.1 Core repeat and flanking regions of CSF1PO STR locus. It consists of eight repeating units of tetrameric nucleotides (TAGA); thus, it is designated as allele 8. (© Richard C. Li.)

20.1.2 Repeat Unit Length

Repeat unit length is the number of nucleotides in a single repeat unit. Dimeric, trimeric, tetrameric, pentameric, and hexameric repeat units appear in the human genome. For example, dimeric and trimeric repeats are very abundant, but they are not usually used for forensic applications. High frequencies of stutter peaks (Section 20.4.2.1) that interfere with genotype interpretation are observed when dimeric and trimeric repeats are amplified. On the contrary, only a few thousand pentameric repeats and a few hundred hexameric repeats exist in the human genome. The pentameric and hexameric repeats are very polymorphic. Only a few pentameric and hexameric repeats are used for forensic applications because they are less abundant in the human genome. The human genome has at least 10^4 tetrameric repeats representing approximately 9% of the total STRs. The STRs with tetrameric repeats are very polymorphic. When they are amplified by PCR, this category of STRs exhibits fewer frequencies of stutter peaks than STRs with dimeric and trimeric repeats. Therefore, the most commonly used STR loci for forensic DNA profiling are the STRs with tetrameric repeats.

20.1.3 Repeat Unit Sequences

STR loci compatible for forensic use can be divided into several classes based on their repeat unit sequences. Figure 20.2 shows representative examples of core repeat sequences. *Simple repeats* consist of tandem repeats with identical repeat unit sequences (Figure 20.2a). Allele designation is based on the number of repeat units in the core repeat region. For example, a D5S818 allele consisting of ten repeating units of the tetrameric nucleotide sequence AGAT is designated as allele 10. *Compound repeats* consist of more than one type of simple repeat (Figure 20.2b). *Complex repeats* contain several clusters of different tandem repeats with intervening sequences (Figure 20.2c).

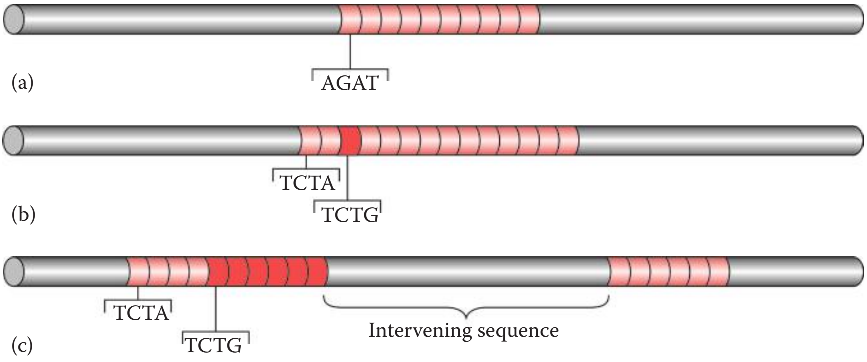

Figure 20.2 Examples of core repeat sequences. (a) A simple repeat in which D5S818 $[AGAT]_{10}$ is designated as allele 10, consisting of 10 repeating units of the tetrameric nucleotides, AGAT. (b) Compound repeats. Allele 14 of D8S1179 consists of two types of repeating units: $[TCTA]_2$, $[TCTG]_1$, and $[TCTA]_{11}$. (c) Complex repeats. Allele 24 of D21S11 contains several clusters of different tandem repeats, $[TCTA]_4$, $[TCTG]_6$, and $[TCTA]_6$, with a 43 bp intervening sequence: $[TCTA]_3$ TA $[TCTA]_3$ TCA $[TCTA]_2$ TCCATA. (© Richard C. Li.)

Nonconsensus alleles with partial repeat units also appear in the population. These nonconsensus alleles, also known as *microvariants*, differ from common alleles by one or more nucleotides. They are designated by the number of consensus repeats, followed by a decimal point and the number of nucleotides of the partial repeat, for example, the TH01 allele 9.3 is 1 nucleotide shorter than allele 10.

Another type of nonconsensus allele can result from a limitation of STR analysis. These alleles have the same number of tandem repeats as common alleles but contain different sequences. These microvariants cannot be distinguished by STR profiling because their length is identical to the lengths of common alleles.

20.2 STR Loci Commonly Used for Forensic DNA Profiling

In the early 1990s, STR loci were initially utilized for genetic studies and were later applied to forensic DNA profiling. The first STR multiplex system, known as the quadruplex, was developed by Forensic Science Services in the United Kingdom. It consisted of four STR loci (F13A1, FES, TH01, and VWA) with a *population match probability* (P_m) of 10^{-4} (Figure 20.3). P_m measures the discriminating power of an STR locus used for forensic DNA analysis. The lower the P_m (i.e., the higher the discriminating power), the less likely a match will occur between two randomly chosen profiles from different individuals (Chapter 25). In 1995, the first national DNA database was established in the United Kingdom. It contained six STR loci, also known as the *second-generation multiplex* (SGM), consisting of D8S1179, D18S51, D21S11, FGA, TH01, and VWA, with a P_m value of 10^{-7}. The SGM system also included the amelogenin locus (Section 21.3.1) for determining the sexes of DNA contributors. Subsequently, four additional loci were added to SGM with a P_m as low as 10^{-13} (SGM Plus).

To allow for international data exchange, the European DNA Profiling Group (EDNAP) recommended the use of TH01 and VWA loci for all participating European laboratories in 1996. In 1998, the *European Standard Set* (ESS) of loci was established and included TH01, VWA, FGA, and D21S11 for forensic use in Europe. Thereafter, D3S1358, D8S1179, and D18S51 loci were added to the ESS. Other loci are used as well, such as SE33, which is used in Germany's database. In 1998, the US Federal Bureau of Investigation established the *Combined DNA Index System* (CODIS). It contains 13 core STR loci plus the amelogenin sex-typing locus with a P_m

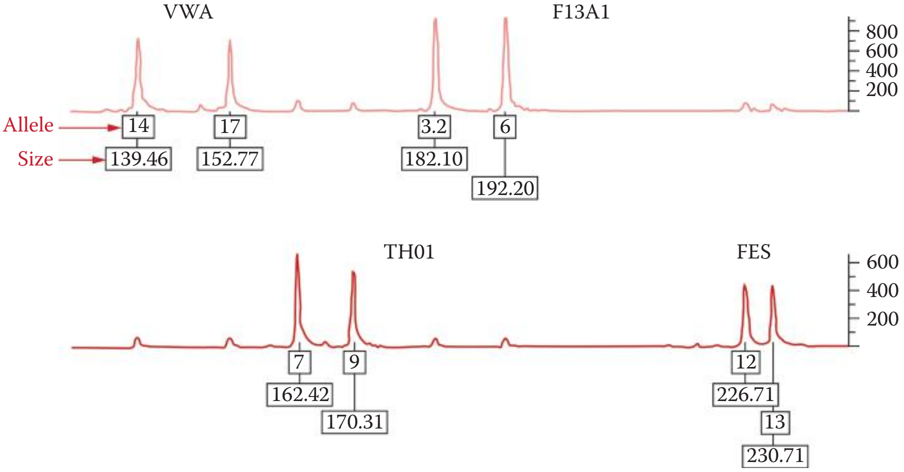

Figure 20.3 DNA profile obtained using the first STR multiplex system: the quadruplex. F13A1, FES, TH01, and VWA loci are shown. (© Richard C. Li.)

Table 20.1 Common STR Loci

Locus Symbol	Repeat Motif	Repeat Category	Cytogenetic Map Location on Chromosome	Distance from pter (Mb)	Structural Gene
CSFIPO	TAGA	Simple	5q33.1	149.4	Intron 6 of c-fms protooncogene
FGA	[CTTT] [CCTT]	Compound	4q31.3	155.5	Intron 3 of fibrinogen α chain gene
TH01	AATG	Simple	11p15.5	2.2	Intron 1 of tyrosine hydroxylase gene
TPOX	TGAA	Simple	2p25.3	1.5	Intron 10 of thyroid peroxidase gene
SE33/ ACTBP2	AAAG	Complex	6q14	89	5′ Flanking sequence of β-actin-related pseudogene 2 gene
VWA	[TCTG] [TCTA]	Compound	12p13.31	6.1	Intron 40 of von Willebrand factor gene
D1S1656	[TAGA] [TAGC]	Compound	1q42.2	230.9	Anonymous
D2S441	[TCTA] [TCAA]	Compound	2p14	68.2	Anonymous
D2S1338	[TGCC] [TTCC]	Compound	2q35	218.9	Anonymous
D3S1358	[TCTG] [TCTA]	Compound	3p21.31	45.6	Anonymous
D5S818	AGAT	Simple	5q23.2	123.1	Anonymous
D7S820	GATA	Simple	7q21.11	83.8	Anonymous
D8S1179	[TCTA] [TCTG]	Compound	8q24.13	125.9	Anonymous
D10S1248	GGAA	Simple	10q26	–[a]	Anonymous
D12S391	[AGAT] [AGAC]	Compound	12p13.2	12.5	Anonymous
D13S317	TATC	Simple	13q31.1	82.7	Anonymous
D16S539	GATA	Simple	16q24.1	86.4	Anonymous
D18S51	AGAA	Simple	18q21.33	60.9	Anonymous
D19S433	[AAGG] [TAGG]	Compound	19q12	30.4	Anonymous
D21S11	[TCTA] [TCTG]	Complex	21q21.1	20.6	Anonymous
D22S1045	ATT	Simple	22q12.3	37.5	Anonymous

Source: Ensembl Homo sapiens version 75.37 (GRCh37). Mb, megabase.

[a] Locus is not mapped to the assembly in the current Ensembl database.

of 10^{-15}. As the database grows rapidly, the chances of finding incidental matches among DNA profiles are increasing. Recently, additional loci have been added to CODIS and ESS core loci to reduce the likelihood of adventitious matches, as well as to facilitate international data sharing among law enforcement agencies and to improve the discrimination power of forensic STR analysis.

Commonly used STR loci characterized for forensic DNA profiling are summarized in Tables 20.1 and 20.2. To achieve low P_m in forensic STR profiling, desired STR loci should possess certain characteristics as described below. First, the alleles of STR loci selected should be highly variable among individuals. Second, if more than one locus is selected, the loci should not be linked to each other or inherited together (Section 21.2). The STR loci utilized are usually located at different chromosomes to ensure that they are not linked. However, loci that are far

TABLE 20.2 Core STR Loci

Locus	SGM	SGM Plus	ESS	ESS-Extended	CODIS	CODIS-Extended
Amel	☑	☑	☑	☑	☑	☑
CSFIPO					☑	☑
D1S1656				☑		☑
D2S441				☑		☑
D2S1338		☑				☑
D3S1358		☑	☑	☑	☑	☑
D5S818					☑	☑
D7S820					☑	☑
D8S1179	☑	☑	☑	☑	☑	☑
D10S1248				☑		☑
D12S391				☑		☑
D13S317					☑	☑
D16S539		☑			☑	☑
D18S51	☑	☑	☑	☑	☑	☑
D19S433		☑				☑
D21S11	☑	☑	☑	☑	☑	☑
D22S1045				☑		2
DYS391						☑
FGA	☑	☑	☑	☑	☑	☑
THO1	☑	☑	☑	☑	☑	☑
TPOX					☑	1
VWA	☑	☑	☑	☑	☑	☑
SE33						3

Source: Hares, D.R., *Forensic Sci Int Genet*, 6, e135, 2012; Gill, P., et al., *Forensic Sci Int*, 156, 242–244, 2006.

☑, minimum required loci; 1–3, recommended loci in ranked order of preference.

enough apart on the same chromosome can still be used (Figures 20.4 and 20.5), since they are not linked. Additionally, STR loci with fewer amplification artifacts such as stutter products are desired. Stutters can complicate the interpretation of profiles derived from a mixed DNA sample from more than one contributor. Moreover, STR loci with short amplicon (amplified product) lengths are preferred for multiplex STR analysis and the testing of degraded DNA samples.

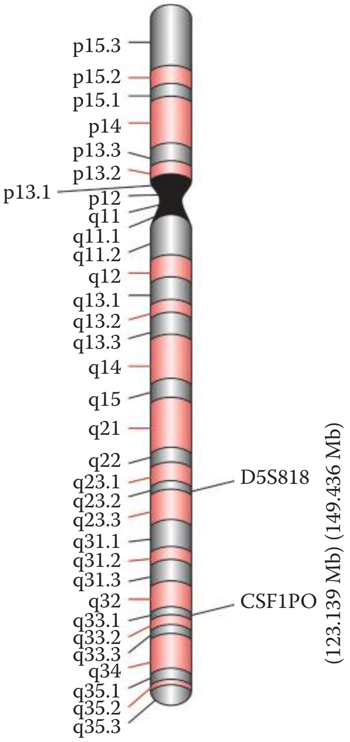
p15.3
p15.2
p15.1
p14
p13.3
p13.2
p13.1
p12
q11
q11.1
q11.2
q12
q13.1
q13.2
q13.3
q14
q15
q21
q22
q23.1
q23.2
q23.3
q31.1
q31.2
q31.3
q32
q33.1
q33.2
q33.3
q34
q35.1
q35.2
q35.3
D5S818
CSF1PO
(123.139 Mb) (149.436 Mb)

Figure 20.4 Cytogenetic map showing the locations of STR markers on chromosome 5. CSF1PO and D5S818 are separated by 26 Mb (megabases). (© Richard C. Li.)

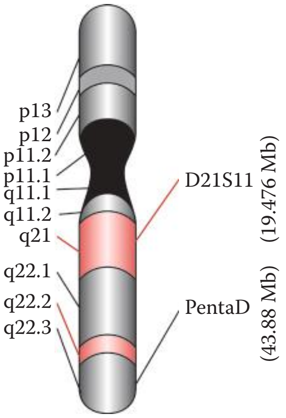
p13
p12
p11.2
p11.1
q11.1
q11.2
q21
q22.1
q22.2
q22.3
D21S11
PentaD
(19.476 Mb)
(43.88 Mb)

Figure 20.5 Cytogenetic map showing the locations of STR markers on chromosome 21. D21S11 and PentaD are separated by 24 Mb (megabases). (© Richard C. Li.)

20.3 Forensic STR Analysis

STR loci are amplified using fluorescent dye-labeled primers. A multiplex STR system utilizes multiple fluorescent dyes to label each amplicon. The amplicons are separated via electrophoresis. The different fluorescent dye colors are resolved by the detector, and the signals corresponding to each DNA fragment are identified using specialized computer software. The data collection process generates an *electropherogram* that shows a profile of peaks corresponding to each DNA fragment. The positions of these peaks represent the electrophoretic mobility of the DNA fragments. A small fragment, migrating faster than a large one, peaks earlier in the electropherogram than a longer fragment. The DNA fragments are sized by comparison to an internal size standard (Figure 20.6; Chapter 8). Figure 20.7 summarizes the work flow of a forensic STR analysis.

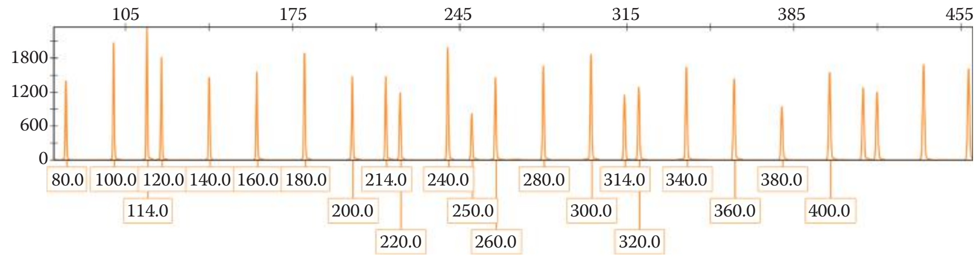

Figure 20.6 Electropherogram of GeneScan™ 500 size standard (Applied Biosystems). RFU represents relative fluorescence unit. (© Richard C. Li.)

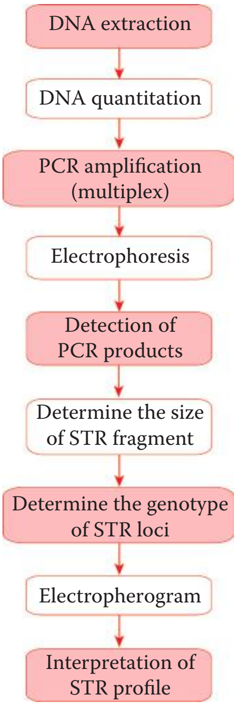
DNA extraction
DNA quantitation
PCR amplification (multiplex)
Electrophoresis
Detection of PCR products
Determine the size of STR fragment
Determine the genotype of STR loci
Electropherogram
Interpretation of STR profile

Figure 20.7 Capillary electrophoresis separation of amplified STR products. Fluorescent dye-labeled amplification products are separated and subsequently detected. Various fluorescent dye colors are resolved by the detector. The peaks corresponding to each DNA fragment are identified. (© Richard C. Li.)

The area or amplitude of the peak, expressed as relative fluorescence units (RFU) (Section 9.3.1.3), reflects the fluorescent signal intensity. The RFU value of peak height is proportional to the amount of DNA amplicons being analyzed. When the RFU value is very low, it is difficult to distinguish a signal from background noise. The manufacturer of commonly used instruments for forensic DNA analysis recommends 150 RFU as the threshold of detection. Peaks below 150 RFU should be interpreted with caution. Some forensic laboratories use lower thresholds, as low as 50 RFU, based on their own validation studies. In contrast, when the RFU value is too high, it saturates the sensitivity of an instrument as well as causing artifacts such as pull-up signals (Section 20.4.3.1). The maximum RFU allowed is usually 6000 RFU.

20.3.1 Determining the Genotypes of STR Fragments

As noted earlier, electropherograms are usually plotted as fluorescent signal intensity (in RFU) versus the sizes of the DNA fragments. The data in an electropherogram can then be converted into a genotype. The genotype for a specific STR locus is defined as the number of repeat units of the allele. STR genotype data generated from different laboratories can be compared easily and are suitable for databasing.

The genotype is determined by using an allelic ladder, which is important to achieve accurate genotype profiling. An allelic ladder is a collection of synthetic fragments corresponding to common alleles observed in the human population for a given set of STR loci (Figures 20.8 and 20.9). The ladders are compared to data obtained from an electropherogram of a questioned sample to determine the genotype. Thus, each allele in a ladder must be resolved properly in order to determine correct STR alleles for a sample. The sizes of DNA fragments of a sample are correlated to sizes of fragments for each allele in an allelic ladder in order to determine the allele designation (genotype) of a questioned sample (Figures 20.10 and 20.11). If a rare allele fails to match alleles within an allelic ladder, it is considered an *off-ladder allele*. If an off-ladder allele is present, the sample should be reanalyzed. The presence of a rare allele can be confirmed by repeating the electrophoresis process based on the characteristic electrophoretic mobility of the rare allele.

20.3.2 Interpretation of STR Profiling Results

General guidelines for the interpretation and the reporting of STR profile results were set by the Scientific Working Group on DNA Analysis Methods (SWGDAM) and the DNA Commission of the International Society of Forensic Genetics (ISFG). Typically, conclusions are categorized as inclusion, exclusion, or inconclusive result.

20.3.2.1 Inclusion (Match)

Peaks of compared STR loci, such as those between the profiles of suspect and crime scene evidence or victim and crime scene evidence, show identical genotypes. The strength of this conclusion can be evaluated via statistical analysis and is usually cited in the case report (Chapter 25).

20.3.2.2 Exclusion

The genotypes of two or more samples differ, and the profile of the sample is determined to be an exclusion, meaning that the profiles originated from different sources.

20.3.2.3 Inconclusive Result

The data do not support a conclusion of inclusion or exclusion. In other words, insufficient information is available to reach a conclusion.

20.4 Factors Affecting Genotyping Results

A number of genetics-, amplification-, and electrophoresis-related factors may affect the accuracy of genotypic profiles.

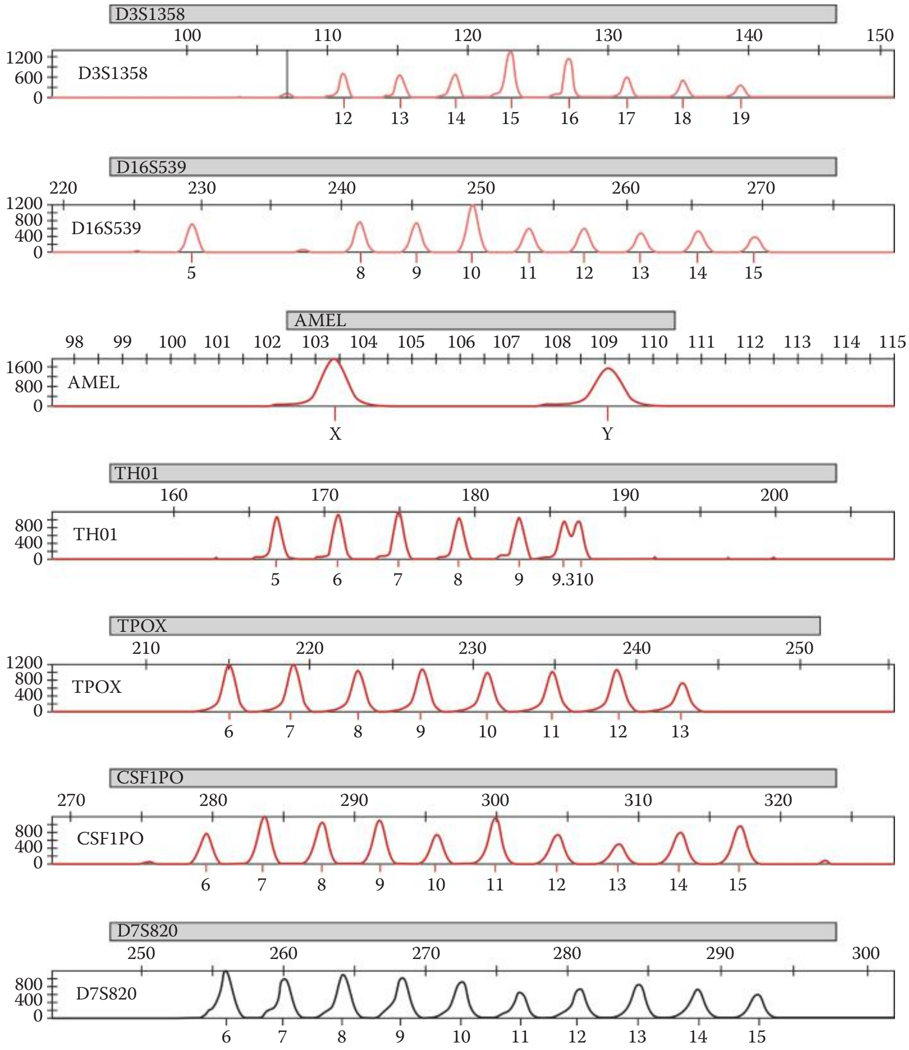

Figure 20.8 Electropherogram of the allelic ladder of the AmpFlSTR® COfiler® PCR Amplification Kit (Applied Biosystems). AMEL, CSF1PO, D3S1358, D7S820, D16S539, TH01, and TPOX loci are shown. (© Richard C. Li.)

20.4.1 Mutations

STR loci with low mutation frequencies are desired, in particular, for human identification in mass disasters and for missing person and paternity cases. However, STR mutations do occur, which can affect the profiling results.

20.4.1.1 Mutations at STR Core Repeat Regions

Mutations, usually resulting in a gain or a loss of a single repeat unit, are observed at STR loci. If a mutation occurs in the germ cells (cells that form gametes), the mutant allele will be transmitted to and be present in all cell types of the progeny. This type of inheritable mutation in germ cell lineage is called a *germ-line mutation*. The frequency of germ-line mutation can be measured by the *mutation rate*, expressed as the number of mutations per generation (germ-line transmission). The average mutation rate of commonly used STR loci is about 10^{-4} mutations per germ-line transmission. However, the mutation rate may vary among different STR loci.

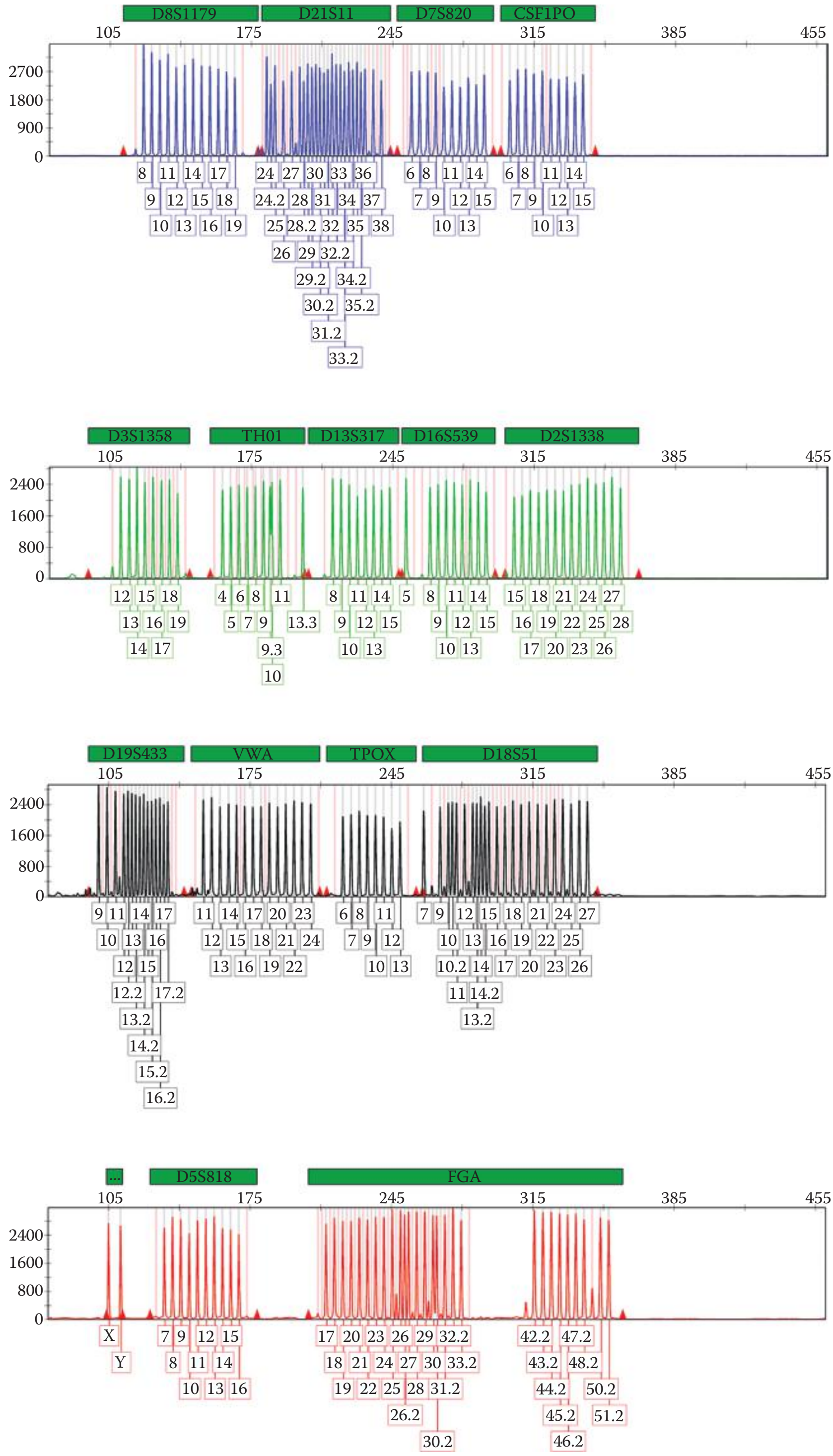

Figure 20.9 Electropherogram of the allelic ladder of the AmpFlSTR® Identifiler® Plus Kit (Applied Biosystems). AMEL, CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, TH01, TPOX, and VWA loci are shown. (© Richard C. Li.)

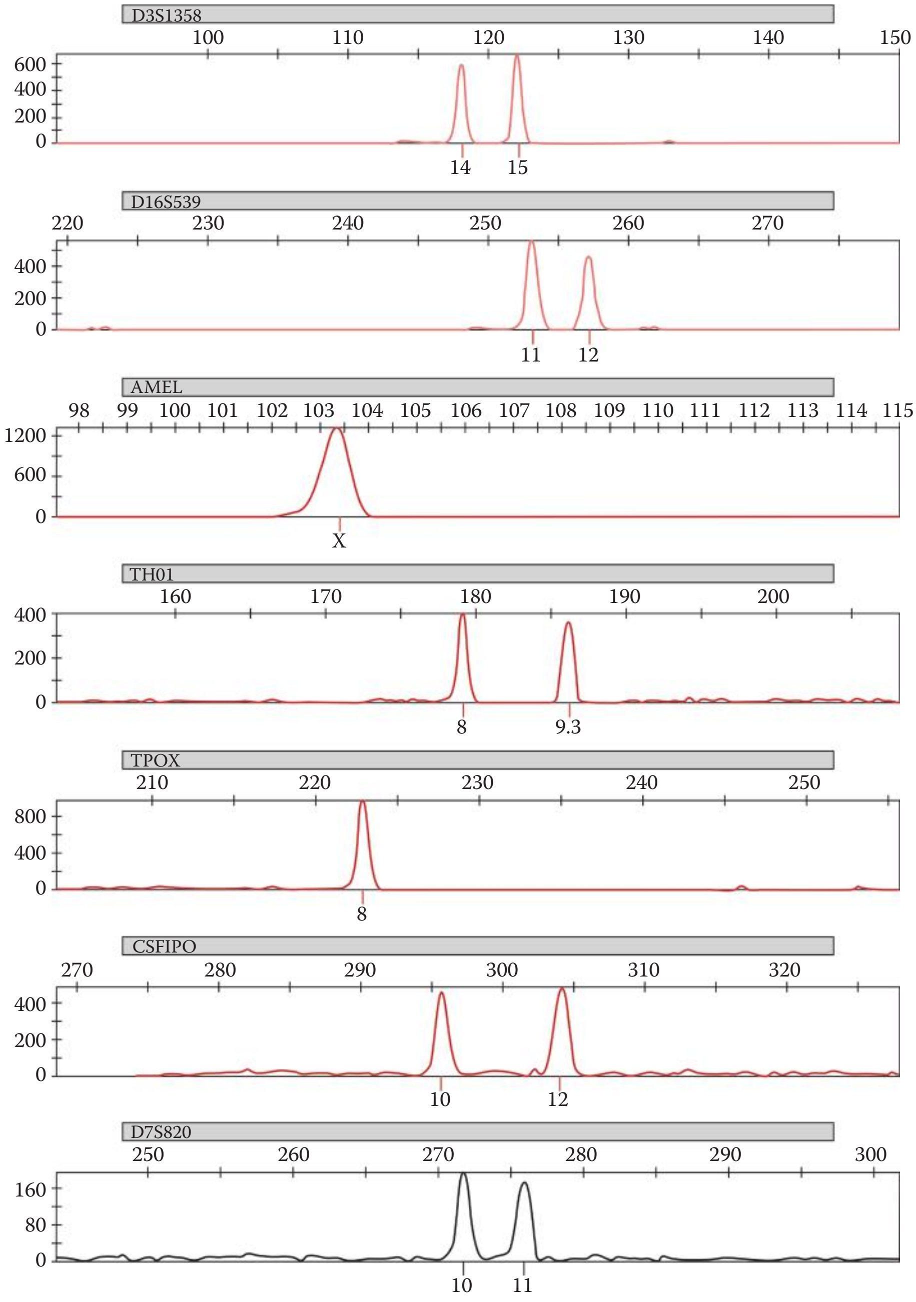

Figure 20.10 Individual DNA profile (Cofiler). The genotype of the DNA profile is shown. AMEL: X, X. CSF1PO: 10, 12. D3S1358: 14, 15. D16S539: 11, 12. D7S820:10, 11. TH01:8, 9.3. TPOX: 8, 8. (© Richard C. Li.)

In contrast, *somatic mutations* involve the mutation of only somatic cells. The germ cells are not affected, and thus a mutant allele is not transmitted to the progeny. A somatic mutation occurring at the core repeat region of an STR locus can be detected and compared to the wild-type allele. The ratio of the signal intensities of the wild-type and mutant alleles varies, depending on the number

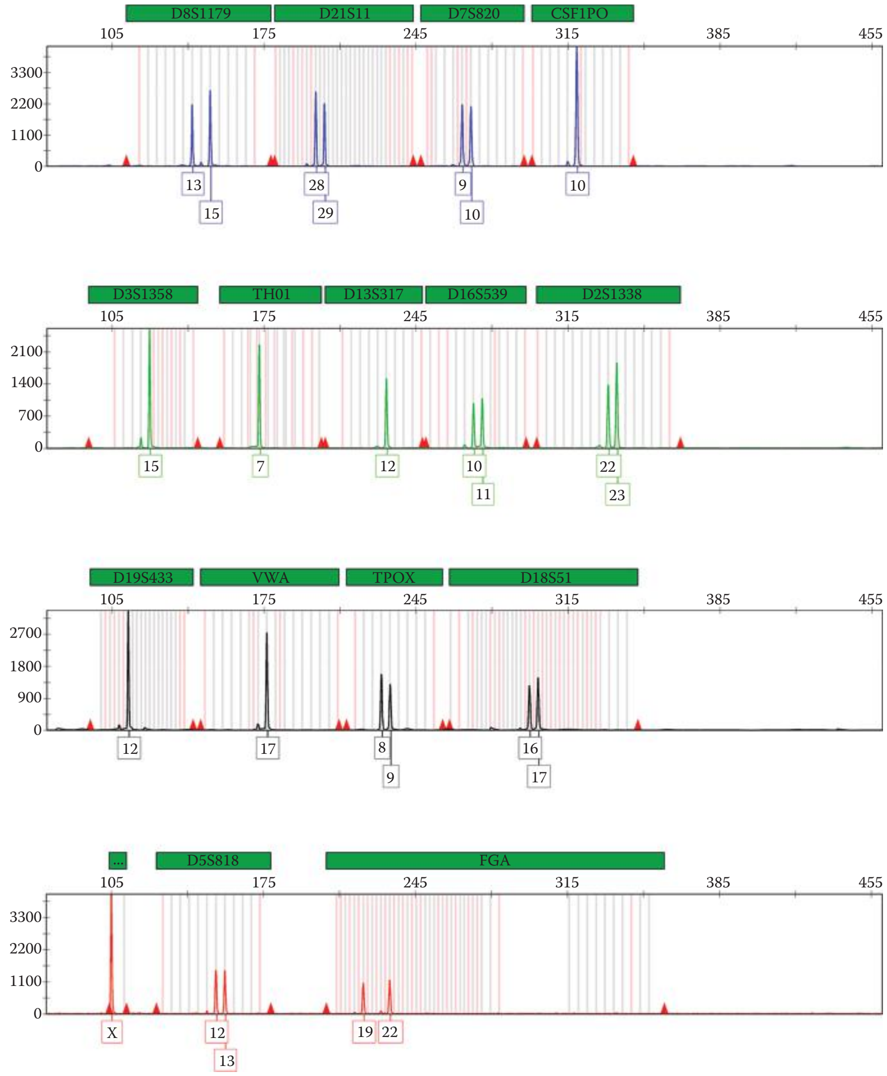

Figure 20.11 Individual DNA profile (Identifiler). The genotype of the DNA profile is shown. AMEL: X, X. CSF1PO: 10, 10. D2S1338: 22, 23. D3S1358: 15, 15. D5S818: 12, 13. D7S820: 9, 10. D8S1179: 13, 15. D13S317: 12, 12. D16S539: 10, 11. D18S51: 16, 17. D19S433: 12, 12. D21S11: 28, 29. FGA: 9, 22. TH01: 7, 7. TPOX: 8, 9. VWA: 17, 17. (© Richard C. Li.)

of mutation-carrying cells in the tissue. Somatic mutations are usually tissue specific. STR profiles from different tissues of the same individual can be compared if a somatic mutation is suspected.

20.4.1.2 Chromosomal and Gene Duplications

Duplicating one of the homologous chromosomes results in a total of three copies of a particular chromosome. This condition, called *trisomy*, is rare and often associated with genetic diseases such as Down's syndrome (chromosome 21 duplication). Duplications have also been observed in chromosomes 13, 18, and X. Other anomalies include duplication of a portion of a chromosome and a single or group of genes instead of an entire chromosome.

A duplication bearing a mutation within the STR core repeat region can affect the number of tandem repeat units. If the duplicated locus is mutated, a *triallelic* or three-peaks pattern can be detected at a single locus, but not at loci located at other chromosomes in a multiplex STR profile (Figure 20.12). The three alleles usually have equal signal intensity (peak amplitude or peak area). Triallelic patterns at STR loci commonly used for forensic DNA analysis have been documented. Many occur at the TPOX, FGA, and CSF1PO loci. If the duplicate locus is not mutated, only two alleles will be observed in a heterozygote. However, the ratio of the peak amplitude of the alleles will be 1:2 (one copy vs. two copies including the duplicate) at that particular locus. However, STR profiles at the loci located at other chromosomes are not affected.

20.4.1.3 Point Mutations

Point mutations involve the changing of a nucleotide sequence through nucleotide substitution, insertion, or deletion. Insertion or deletion mutations affect the lengths of the core repeat regions and the amplified flanking regions of STR loci and thus affect STR profiles. Nucleotide substitution mutations (except those residing within the primer-binding regions) do not affect the length of DNA and thus do not affect STR profiles.

However, mutations occurring at the primer-binding sequences of the flanking regions of STR loci may affect genotype results. If a mutation at a primer-binding sequence prevents the primer from annealing to the template, this leads to a complete failure of the amplification of the allele. This phenomenon is known as a *null allele* or silent allele (Figure 20.13). To overcome the consequences of a null allele, an alternative primer annealing to a flanking region away from the mutated sequence can be used. Additionally, a primer with the sequence that is complementary to the known mutation can also be used. If the mutation does not completely prevent the primer from annealing but reduces the efficiency of the amplification, the resulting signal intensity of the allele is usually decreased. This problem may be solved by modifying the condition of amplification for the mutant allele.

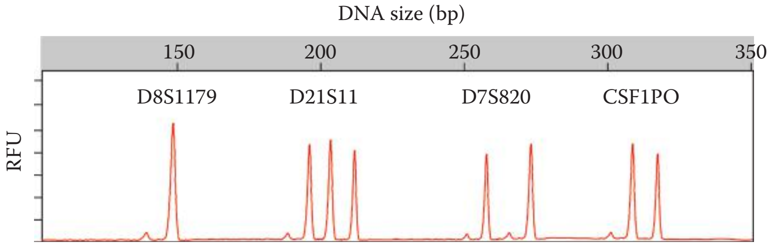

Figure 20.12 Triallele. In this example, the triallele is observed only at D21S11 and not at other STR loci. (© Richard C. Li.)

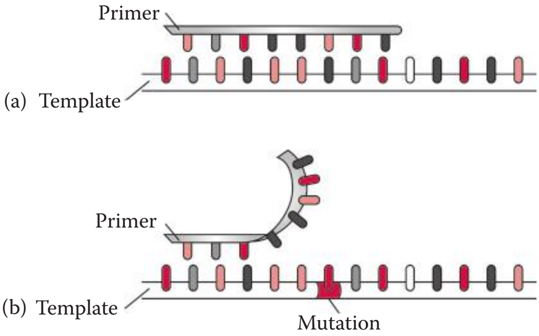

Figure 20.13 Null allele. An allele present in the sample failed to be amplified by one of the primer sets as a result of a rare mutation at the primer-binding sequence of the flanking region: (a) wild type and (b) mutation. (© Richard C. Li.)

20.4.2 Amplification Artifacts

20.4.2.1 Stuttering

A stutter is a minor allele peak, also known as a stutter peak, whose repeat units are shorter or longer than the parental allele peak (Figure 20.14). Less stuttering is observed with pentameric and hexameric repeat unit loci compared to shorter repeat unit loci. The loci that contain complex repeat sequences usually exhibit reduced stuttering. At a given STR locus, large alleles appear to yield more stutter than smaller alleles.

Commonly observed stutters are one repeat unit shorter than the parental allele. It is believed that stuttering is due to the slippage of polymerase, which may have occurred during amplification reactions (Figure 20.15). Stutters with repeat units longer than the parental allele peak can also be observed, but are very rare. The *stutter ratio* is defined as the area of the stutter peak divided by the area of the parental peak. The stutter ratio is usually less than 0.15. A ratio of greater than 0.15 should be interpreted with caution due to the potential presence of DNA from more than one contributor.

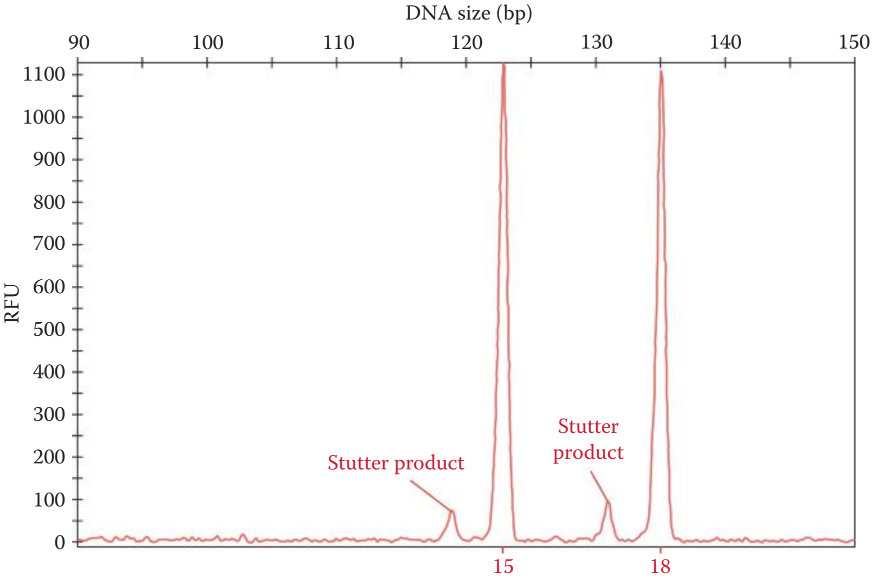

Figure 20.14 Stutter products. (© Richard C. Li.)

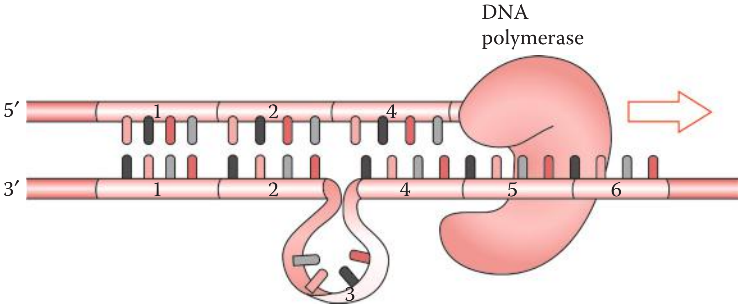

Figure 20.15 Proposed mechanism for stutter products. During the DNA synthesis step of PCR amplification, a DNA polymerase slips, and a region of the primer–template complex becomes unpaired, causing the template strand to form a loop. The consequence of this one-repeat loop is a shortened PCR product smaller than the template by a single repeat unit. (© Richard C. Li.)

20.4.2.2 Nontemplate Adenylation

During PCR amplification, DNA polymerase often adds an extra nucleotide, usually an adenosine, to the 3′-end of an amplicon. Such a phenomenon is referred to as a *nontemplate addition* resulting in an amplicon that is one base pair longer than the parental allele (designated the +A peak), as shown in Figure 20.16. As a result, an amplicon has both the –A amplicon, which corresponds to the size of the parental allele, and the +A amplicon, which represents the amplicon with the nontemplate addition. To simplify the analysis, the most commonly used multiplex STR kits utilize amplification conditions that favor the adenylation of amplicons. Thus, most amplicons in a sample contain an additional adenosine at the 3′ end (+A peak). However, partial nontemplate addition can occur when too much DNA template is utilized in PCR amplification. As a result, a mixture of –A and +A peaks is usually observed.

20.4.2.3 Heterozygote Imbalance

Heterozygote imbalance occurs when one of the alleles has greater peak area or amplitude than the other allele within the same locus in which the two alleles of a heterozygote are compared (Figure 20.17). It is believed that heterozygote imbalance may arise if the DNA sample contains unequal copies of DNA template of the two alleles for the heterozygote, or the two alleles of a heterozygote are unequally amplified, a condition known as *preferential amplification*. Preferential amplification usually refers to an event where a smaller allele is amplified more efficiently than larger ones. As a result, the presence of heterozygote imbalance interferes with the interpretation of samples with a DNA mixture derived from more than one contributor.

20.4.2.4 Allelic Dropout

Allelic dropout occurs when an allele, usually one of the heterozygote alleles, fails to be detected. To date, our understanding of what causes the dropout is very limited. The occurrence of allelic dropout can be the result of an extreme situation of preferential amplification or heterozygote imbalance. Additionally, certain mutations leading to amplification failure (Section 20.4.1.3) can cause allelic dropout.

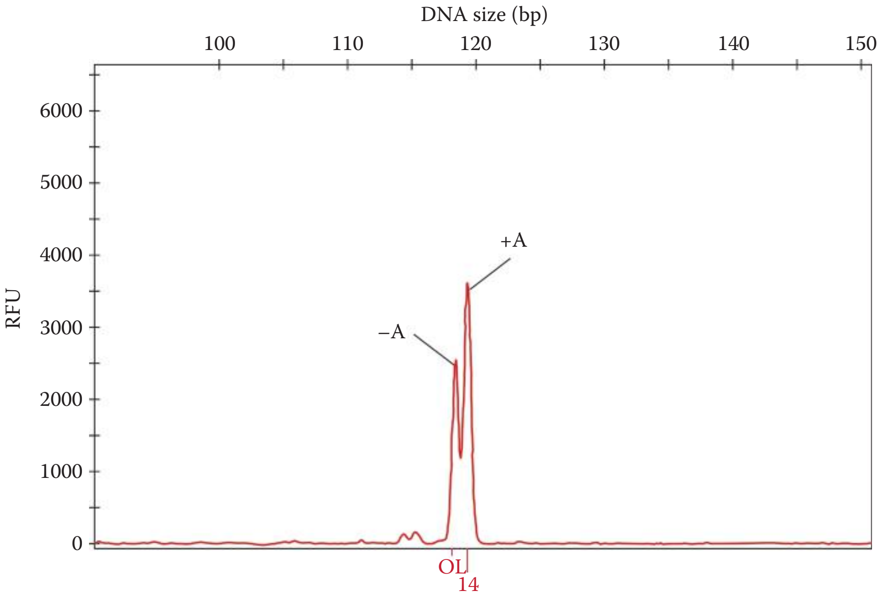

Figure 20.16 Nontemplate adenylation. OL represents the off-ladder allele. (© Richard C. Li.)

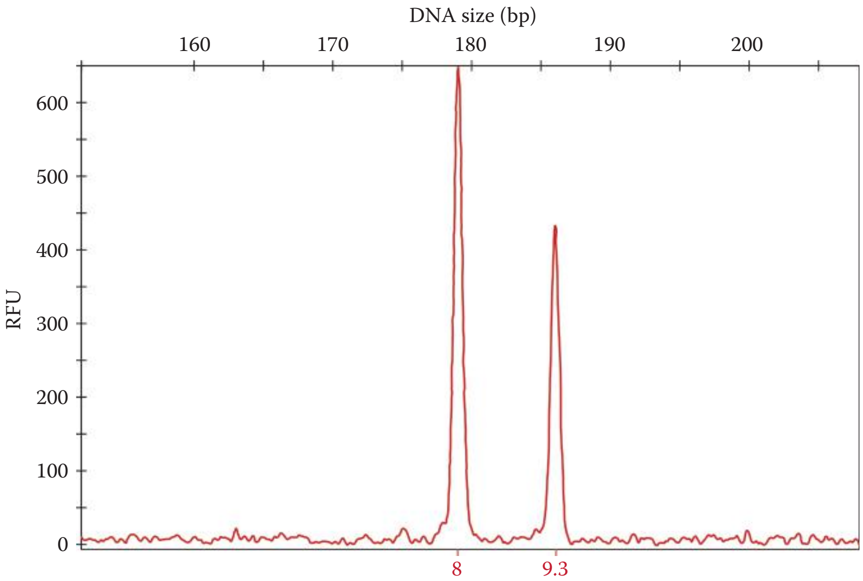

Figure 20.17 Heterozygote imbalance. The signal intensity of one allele is greater than that of the other allele within the same locus. (© Richard C. Li.)

20.4.3 Electrophoretic Artifacts

20.4.3.1 Pull-Up Peaks

A *pull-up peak* occurs when a minor peak of one color on an electropherogram is pulled up from a major allelic peak in another color (Figure 20.18) when the colors have overlapping spectra. For example, a green peak may pull up a yellow peak, or a blue peak may pull up a green peak. A pull-up peak may contribute to the inaccuracy of a profile if the position of a pull-up peak corresponds to the position of an allele. A pull-up peak often occurs when a sample is overloaded or a matrix file (a spectral calibration) is not updated. Thus, loading a proper amount of sample or installing an appropriate matrix file can prevent the occurrence of pull-up peaks.

20.4.3.2 Spikes

Spikes are sharp peaks, with similar signal intensities, that are present in all color panels of an electropherogram (Figure 20.19). Spikes are caused by air bubbles and urea crystals in the capillary of an electrophoretic platform. Voltage spikes can also contribute to spike peaks. The spikes are electrophoretic artifacts and are not reproducible. Thus, electrophoresis can be repeated to verify that the spikes occurred in a previous electrophoresis result.

20.5 Genotyping of Challenging Forensic Samples

20.5.1 Degraded DNA

Environmental exposure, such as high humidity and temperature, of biological evidence can lead to DNA degradation such as the breaking of DNA molecules into small fragments. The more severe the degradation, the more intensive the fragmentation. In forensic DNA analysis, the size range of STR amplicons is usually 100–500 bp in length. When a sample experiences some degradation, large alleles are less likely to be amplified than small alleles (Figures 20.20 and 20.21). As a result, the dropout of large alleles often occurs, leading to a partial DNA profile or even a failure in obtaining a DNA profile. To address this issue, the PCR primers can be

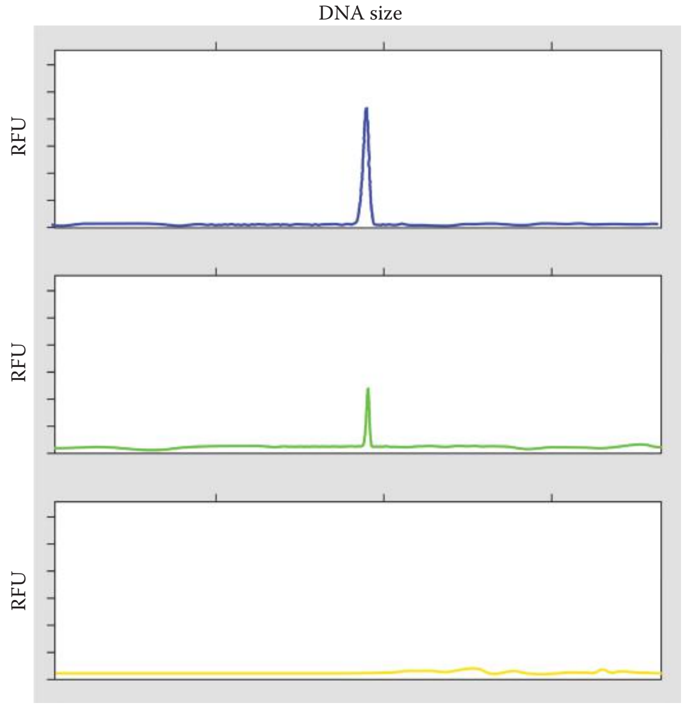

Figure 20.18 Pull-up peaks. The peaks with overlapping spectra observed in the top and middle panels are not observed in the bottom panel. (© Richard C. Li.)

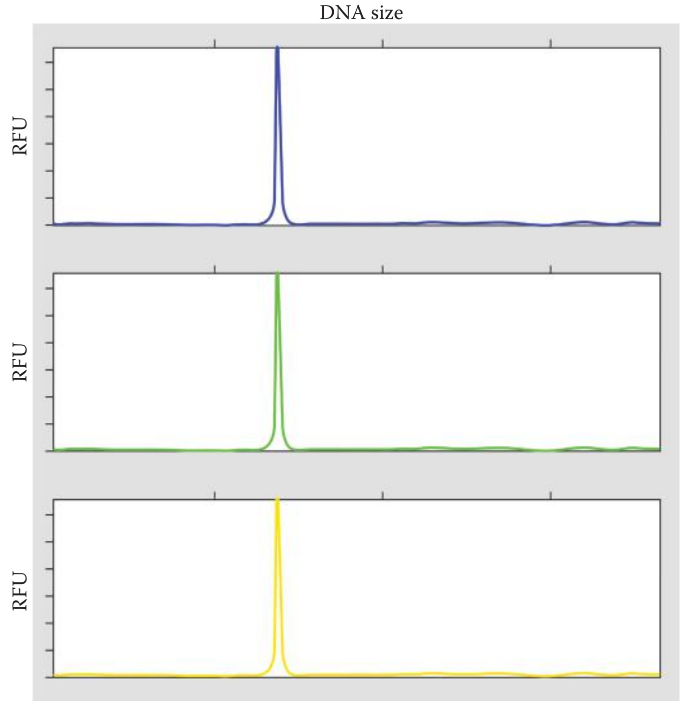

Figure 20.19 Spike peaks can be observed in various intensities. (© Richard C. Li.)

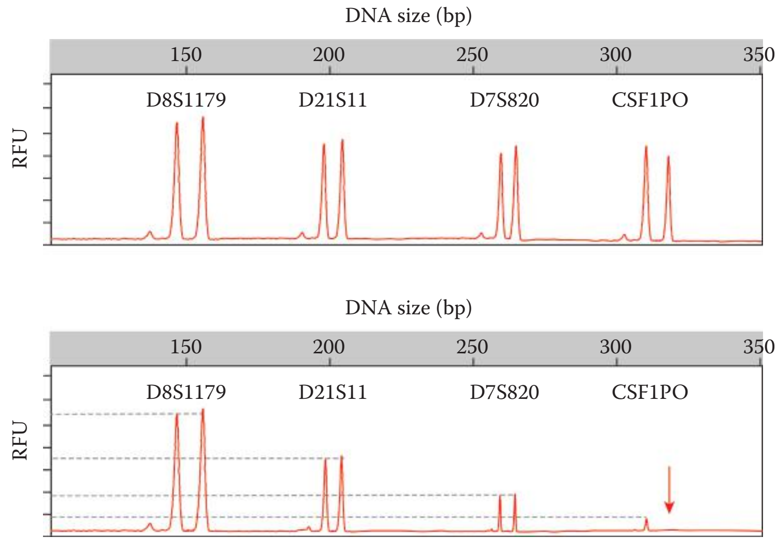

Figure 20.20 Diagram of STR profiles of samples: without DNA degradation (top) and with degradation (bottom). Arrow: allelic dropout. (© Richard C. Li.)

redesigned to anneal more proximally to the STR core repeat region than standard STR primers, yielding small amplicons also known as miniSTRs. Using the miniSTR strategy, more alleles can be detected in degraded DNA samples than using the standard STR primers.

20.5.2 Low Copy Number DNA Testing

Low copy number (LCN) DNA analysis involves the testing of very small amounts of DNA (<100 pg) in a sample. LCN DNA analysis is often needed for samples derived from evidence such as fingerprints and tools and weapons handled by perpetrators. STR analysis of extremely low levels of human DNA can be achieved by increasing the number of PCR cycles (e.g., increasing from 28 to 34 cycles) to improve the yield of amplicons, thus improving the sensitivity of the analysis.

However, this approach also increases the appearance of artifacts that can make interpretation difficult. For instance, the occurrence of allele dropout, heterozygote imbalance, and stuttering is frequently observed in LCN DNA analysis. Additionally, allele drop-in can arise from contamination. The phenomenon of allele drop-in is usually not reproducible. In an LCN DNA analysis, genotypes can be determined if identical alleles can be detected from two independent amplification reactions.

20.5.3 Mixtures

Samples of DNA from two or more contributors are commonly encountered in forensic cases such as sexual assaults in which the evidence recovered from a victim is mixed with a suspect's bodily fluids (Figure 20.22). The interpretation of DNA profiles of mixed stains is known as *mixture interpretation*, which is described below:

1. *To determine the presence of a mixture*: First, determine whether the source of the DNA in the sample came from one or more individuals by examining the number of alleles at multiple loci. The characteristics listed below usually indicate a mixture:
   a. Severe heterozygote imbalance.
   b. Stutter ratio above 0.15.
   c. Presence of three or more alleles per locus at multiple loci.

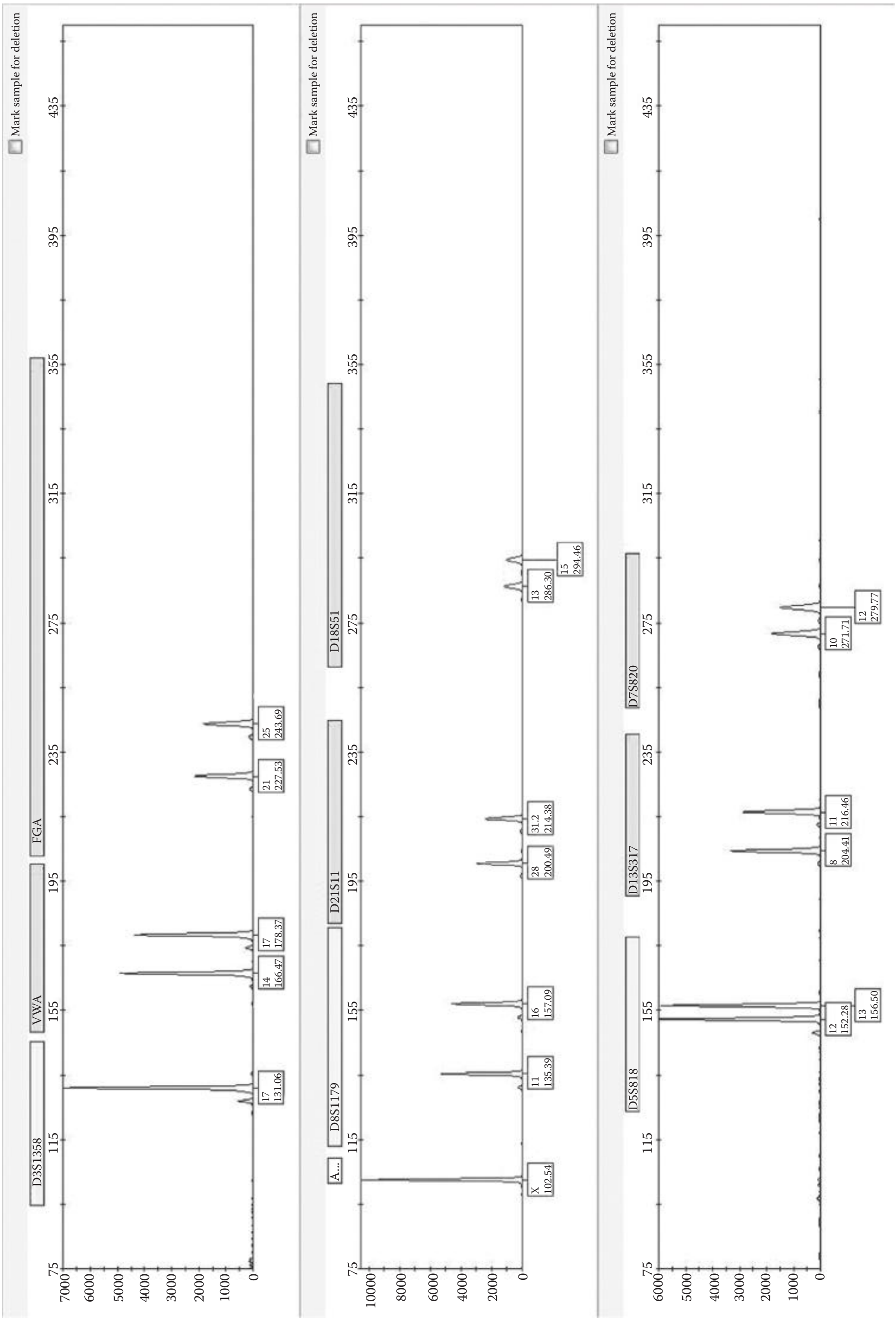

Figure 20.21 DNA profile (AmpFlSTR® Profiler Plus® PCR Amplification Kit; Applied Biosystems) of the degraded sample. (© Richard C. Li.)

Additionally, caution should be taken to distinguish the presence of a mixture and various artifacts such as stutters and nontemplate adenylation.

2. *To determine the genotypes of all alleles and to identify the number of contributors*: Note that the maximum number of alleles at any given locus is two per individual. In the case of homozygous or allele overlap, the number of alleles observed can be less than two per individual.
3. *To estimate the ratios of the contributions*: Determine the relative ratios of the contributions to the mixture made by each individual by comparing the peak areas or

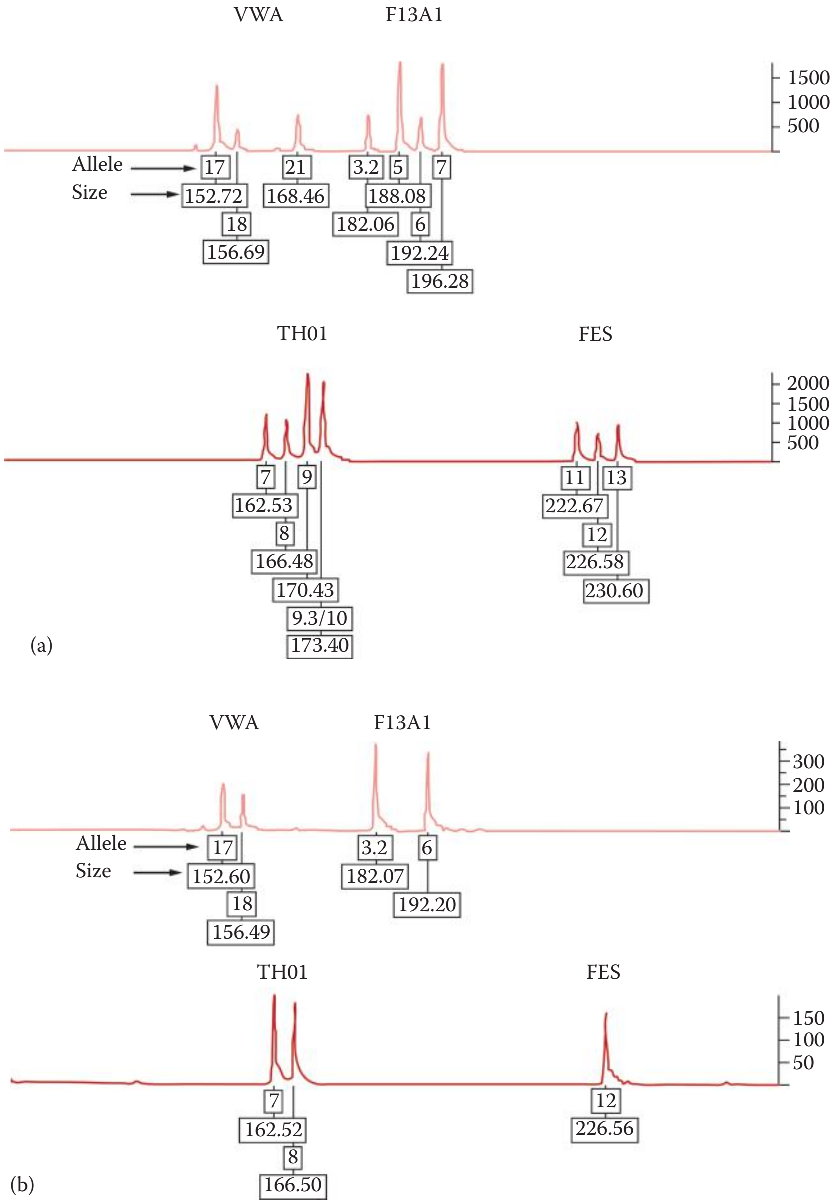

Figure 20.22 DNA profiles of mixed bodily fluids. (a) DNA profile of mixed stains from evidence, (b) DNA profile of the victim. (© Richard C. Li.)

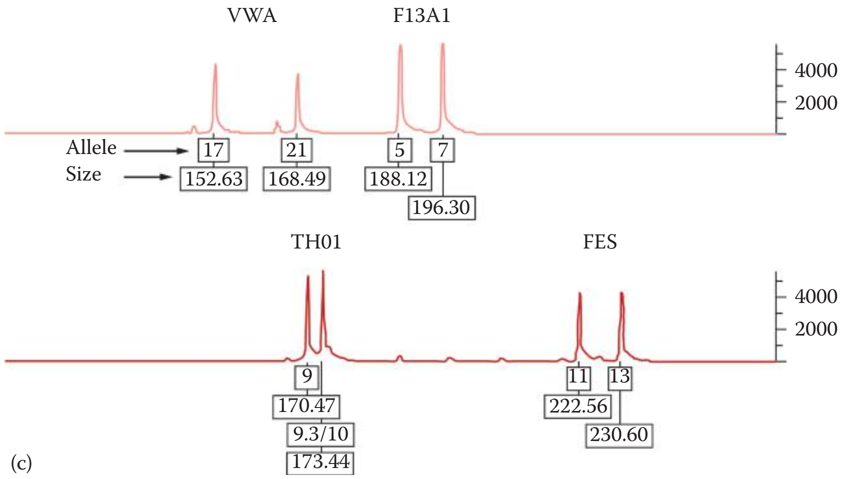

Figure 20.22 (Continued) (c) DNA profile of the suspect.

amplitudes. Amelogenin, a sex-typing marker, is useful in determining the genders of DNA contributors.

4. *To consider all possible genotype combinations*: This may be done by pair-wise comparisons to determine the allele combinations that belong to the minor contributor and those that belong to the major contributor.
5. *To compare reference samples*: The final step is to compare the genotype profiles with the genotypes of reference samples from a suspect and victim. If the DNA profile of the suspect's reference sample matches a major or minor component of the mixture, the suspect cannot be excluded as a contributor.

Detailed guidance for basic steps in mixture interpretation is provided in a number of published guidelines. For example, the Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories was recently established by the Scientific Working Group on DNA Analysis Methods (SWGDAM), which provides direction on DNA profiling and mixture interpretation. Additionally, guidance for the statistical evaluations of the mixture analysis results is also provided.

Bibliography

STR Profiling

Andersen, J., et al., Report on the third EDNAP collaborative STR exercise. *Forensic Sci Int*, 1996, **78**(2): 83–93.

Anker, R., T. Steinbrueck, and H. Donis-Keller, Tetranucleotide repeat polymorphism at the human thyroid peroxidase (hTPO) locus. *Hum Mol Genet*, 1992, **1**(2): 137.

Asamura, H., et al., MiniSTR multiplex systems based on non-CODIS loci for analysis of degraded DNA samples. *Forensic Sci Int*, 2007, **173**(1): 7–15.

Asamura, H., et al., MiniY-STR quadruplex systems with short amplicon lengths for analysis of degraded DNA samples. *Forensic Sci Int Genet*, 2007, **1**(1): 56–61.

Aznar, J.M., et al., I-DNASE21 system: Development and SWGDAM validation of a new STR 21-plex reaction. *Forensic Sci Int Genet*, 2014, **8**(1): 10–19.

Balding, D.J., Evaluation of mixed-source, low-template DNA profiles in forensic science. *Proc Natl Acad Sci U S A*, 2013, **110**(30): 12241–12246.

Ballantyne, J., E.K. Hanson, and M.W. Perlin, DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: Combining quantitative data for greater identification information. *Sci Justice*, 2013, **53**(2): 103–114.

Bar, W., et al., DNA recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *Int J Legal Med*, 1997, **110**(4): 175–176.

Barbaro, A., et al., Casework application of a stand-alone pentaplex assay of extended-ESS STRs. *Leg Med (Tokyo)*, 2013, **15**(4): 217–221.

Barber, M.D. and B.H. Parkin, Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179. *Int J Legal Med*, 1996, **109**(2): 62–65.

Barber, M.D., B.J. McKeown, and B.H. Parkin, Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus. *Int J Legal Med*, 1996, **108**(4): 180–185.

Becker, D., et al., New alleles and mutational events at 14 STR loci from different German populations. *Forensic Sci Int Genet*, 2007, **1**(3–4): 232–237.

Benschop, C., H. Haned, and T. Sijen, Consensus and pool profiles to assist in the analysis and interpretation of complex low template DNA mixtures. *Int J Legal Med*, 2013, **127**(1): 11–23.

Benschop, C.C., et al., Assessment of mock cases involving complex low template DNA mixtures: A descriptive study. *Forensic Sci Int Genet*, 2012, **6**(6): 697–707.

Betz, T., et al., "Paterniplex", a highly discriminative decaplex STR multiplex tailored for investigating special problems in paternity testing. *Electrophoresis*, 2007, **28**(21): 3868–3874.

Bill, M., et al., PENDULUM—A guideline-based approach to the interpretation of STR mixtures. *Forensic Sci Int*, 2005, **148**(2–3): 181–189.

Bornman, D.M., et al., Short-read, high-throughput sequencing technology for STR genotyping. *Biotechniques*, 2012, 10.2144/000113857: 1–6.

Brinkmann, B., E. Meyer, and A. Junge, Complex mutational events at the HumD21S11 locus. *Hum Genet*, 1996, **98**(1): 60–64.

Brookes, C., et al., Characterising stutter in forensic STR multiplexes. *Forensic Sci Int Genet*, 2012, **6**(1): 58–63.

Budimlija, Z.M. and T.A. Caragine, Interpretation guidelines for multilocus STR forensic profiles from low template DNA samples. *Methods Mol Biol*, 2012, **830**: 199–211.

Budowle, B., et al., Validation studies of the CTT STR multiplex system. *J Forensic Sci*, 1997, **42**(4): 701–707.

Budowle, B., et al., *DNA Typing Protocols: Molecular Biology and Forensic Analysis*, 2000. Natick, MA: Eaton Publishing.

Budowle, B., et al., STR primer concordance study. *Forensic Sci Int*, 2001, **124**(1): 47–54.

Buse, E.L., et al., Performance evaluation of two multiplexes used in fluorescent short tandem repeat DNA analysis. *J Forensic Sci*, 2003, **48**(2): 348–357.

Butler, J.M., Constructing STR multiplex assays. *Methods Mol Biol*, 2005, **297**: 53–66.

Butler, J.M., Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci*, 2006, **51**(2): 253–265.

Butler, J.M., Y. Shen, and B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA. *J Forensic Sci*, 2003, **48**(5): 1054–1064.

Caragine, T., et al., Validation of testing and interpretation protocols for low template DNA samples using AmpFlSTR Identifiler. *Croat Med J*, 2009, **50**(3): 250–267.

Castella, V., J. Gervaix, and D. Hall, DIP-STR: Highly sensitive markers for the analysis of unbalanced genomic mixtures. *Hum Mutat*, 2013, **34**(4): 644–654.

Chakraborty, R., et al., The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. *Electrophoresis*, 1999, **20**(8): 1682–1696.

Chen, P.Y., et al., Substitution mutation induced migration anomaly of a D10S2325 allele on capillary electrophoresis. *Int J Legal Med*, 2013, **127**(2): 363–368.

Clayton, T.M., et al., A genetic basis for anomalous band patterns encountered during DNA STR profiling. *J Forensic Sci*, 2004, **49**(6): 1207–1214.

Coble, M.D. and J.M. Butler, Characterization of new miniSTR loci to aid analysis of degraded DNA. *J Forensic Sci*, 2005, **50**(1): 43–53.

Coble, M.D., Capillary electrophoresis of miniSTR markers to genotype highly degraded DNA samples. *Methods Mol Biol*, 2012, **830**: 31–42.

Collins, J.R., et al., An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*, 2003, **82**(1): 10–19.

Collins, P.J., et al., Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: The AmpFlSTR Identifiler PCR amplification kit. *J Forensic Sci*, 2004, **49**(6): 1265–1277.

Cotton, E.A., et al., Validation of the AMPFlSTR SGM plus system for use in forensic casework. *Forensic Sci Int*, 2000, **112**(2–3): 151–161.

Cowell, R.G., Validation of an STR peak area model. *Forensic Sci Int Genet*, 2009, **3**(3): 193–199.

Cowell, R.G., S.L. Lauritzen, and J. Mortera, Identification and separation of DNA mixtures using peak area information. *Forensic Sci Int*, 2007, **166**(1): 28–34.

Cowen, S., et al., An investigation of the robustness of the consensus method of interpreting low-template DNA profiles. *Forensic Sci Int Genet*, 2011, **5**(5): 400–406.

Crouse, C.A. and J. Schumm, Investigation of species specificity using nine PCR-based human STR systems. *J Forensic Sci*, 1995, **40**(6): 952–956.

Cupples, C.M., et al., STR profiles from DNA samples with "undetected" or low quantifiler results. *J Forensic Sci*, 2009, **54**(1): 103–107.

Dauber, E.M., et al., Germline mutations of STR-alleles include multi-step mutations as defined by sequencing of repeat and flanking regions. *Forensic Sci Int Genet*, 2012, **6**(3): 381–386.

Davis, C., et al., Variants observed for STR locus SE33: A concordance study. *Forensic Sci Int Genet*, 2012, **6**(4): 494–497.

Debernardi, A., et al., One year variability of peak heights, heterozygous balance and inter-locus balance for the DNA positive control of AmpFlSTR© Identifiler© STR kit. *Forensic Sci Int Genet*, 2011, **5**(1): 43–49.

Dib, C., et al., A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, 1996, **380**(6570): 152–154.

Divne, A.M., H. Edlund, and M. Allen, Forensic analysis of autosomal STR markers using pyrosequencing. *Forensic Sci Int Genet*, 2010, **4**(2): 122–129.

Drabek, J., et al., Concordance study between Miniplex assays and a commercial STR typing kit. *J Forensic Sci*, 2004, **49**(4): 859–860.

Duewer, D.L., et al., NIST mixed stain study 3: Signal intensity balance in commercial short tandem repeat multiplexes. *Anal Chem*, 2004, **76**(23): 6928–6934.

Edwards, A., et al., DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet*, 1991, **49**(4): 746–756.

Edwards, A., et al., Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 1992, **12**(2): 241–253.

Ellegren, H., Microsatellites: Simple sequences with complex evolution. *Nat Rev Genet*, 2004, **5**(6): 435–445.

Ensenberger, M.G., et al., Developmental validation of the PowerPlex 16 HS system: An improved 16-locus fluorescent STR multiplex. *Forensic Sci Int Genet*, 2010, **4**(4): 257–264.

Ensenberger, M.G., et al., Developmental validation of the PowerPlex(®) 21 system. *Forensic Sci Int Genet*, 2014, **9**: 169–178.

Flores, S., et al., Internal validation of the GlobalFiler express PCR amplification kit for the direct amplification of reference DNA samples on a high-throughput automated workflow. *Forensic Sci Int Genet*, 2014, **10C**: 33–39.

Fordyce, S.L., et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform. *Biotechniques*, 2011, **51**(2): 127–133.

Fregeau, C.J. and R.M. Fourney, DNA typing with fluorescently tagged short tandem repeats: A sensitive and accurate approach to human identification. *Biotechniques*, 1993, **15**(1): 100–119.

Fregeau, C.J., K.L. Bowen, and R.M. Fourney, Validation of highly polymorphic fluorescent multiplex short tandem repeat systems using two generations of DNA sequencers. *J Forensic Sci*, 1999, **44**(1): 133–166.

Fregeau, C.J., et al., AmpFlSTR Profiler Plus short tandem repeat DNA analysis of casework samples, mixture samples, and nonhuman DNA samples amplified under reduced PCR volume conditions (25 microL). *J Forensic Sci*, 2003, **48**(5): 1014–1034.

Ge, J., A. Eisenberg, and B. Budowle, Developing criteria and data to determine best options for expanding the core CODIS loci. *Investig Genet*, 2012, **3**: 1.

Gibb, A.J., et al., Characterisation of forward stutter in the AmpFlSTR SGM Plus PCR. *Sci Justice*, 2009, **49**(1): 24–31.

Gil, A., et al., Linkage between HPRTB STR alleles and Lesch-Nyhan syndrome inside a family: Implications in forensic casework. *Forensic Sci Int Genet*, 2013, **7**(1): e5–e6.

Gilder, J.R., et al., Run-specific limits of detection and quantitation for STR-based DNA testing. *J Forensic Sci*, 2007, **52**(1): 97–101.

Gilder, J.R., et al., Magnitude-dependent variation in peak height balance at heterozygous STR loci. *Int J Legal Med*, 2011, **125**(1): 87–94.

Gill, P. and J. Buckleton, Commentary on: Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci 2009;54(4):810–821. *J Forensic Sci*, 2010, **55**(1): 265–268; author reply 269–272.

Gill, P., A. Kirkham, and J. Curran, LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Sci Int*, 2007, **166**(2–3): 128–138.

Gill, P., et al., The evolution of DNA databases—Recommendations for new European STR loci. *Forensic Sci Int*, 2006, **156**(2–3): 242–244.

Gill, P., et al., Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Sci Int Genet*, 2008, **2**(2): 91–103.

Gill, P., et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Sci Int Genet*, 2012, **6**(6): 679–688.

Gill, P., Role of short tandem repeat DNA in forensic casework in the UK—Past, present, and future perspectives. *Biotechniques*, 2002, **32**(2): 366–368, 370, 372, passim.

Goodwin, W. and C. Peel, Theoretical value of the recommended expanded European standard set of STR loci for the identification of human remains. *Med Sci Law*, 2012, **52**(3): 162–168.

Graydon, M., F. Cholette, and L.K. Ng, Inferring ethnicity using 15 autosomal STR loci—Comparisons among populations of similar and distinctly different physical traits. *Forensic Sci Int Genet*, 2009, **3**(4): 251–254.

Green, R.L., et al., Developmental validation of the AmpFlSTR® NGM SElect PCR amplification kit: A next-generation STR multiplex with the SE33 locus. *Forensic Sci Int Genet*, 2013, **7**(1): 41–51.

Greenspoon, S.A., et al., Validation of the PowerPlex 1.1 loci for use in human identification. *J Forensic Sci*, 2000, **45**(3): 677–683.

Greenspoon, S.A., et al., Validation and implementation of the PowerPlex 16 BIO System STR multiplex for forensic casework. *J Forensic Sci*, 2004, **49**(1): 71–80.

Griffiths, R.A., et al., New reference allelic ladders to improve allelic designation in a multiplex STR system. *Int J Legal Med*, 1998, **111**(5): 267–272.

Guo, F., et al., Development of a 24-locus multiplex system to incorporate the core loci in the combined DNA index system (CODIS) and the European standard set (ESS). *Forensic Sci Int Genet*, 2014, **8**(1): 44–54.

Hammond, H.A., et al., Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet*, 1994, **55**(1): 175–189.

Haned, H., et al., Estimating drop-out probabilities in forensic DNA samples: A simulation approach to evaluate different models. *Forensic Sci Int Genet*, 2011, **5**(5): 525–531.

Harder, M., et al., STR-typing of ancient skeletal remains: Which multiplex-PCR kit is the best? *Croat Med J*, 2012, **53**(5): 416–422.

Hares, D.R., Addendum to expanding the CODIS core loci in the United States. *Forensic Sci Int Genet*, 2012, **6**(5): e135.

Hares, D.R., Expanding the CODIS core loci in the United States. *Forensic Sci Int Genet*, 2012, **6**(1): e52–e54.

Heinrich, M., et al., Allelic drop-out in the STR system ACTBP2 (SE33) as a result of mutations in the primer binding region. *Int J Legal Med*, 2004, **118**(6): 361–363.

Hering, S., J. Edelmann, and J. Dressler, Sequence variations in the primer binding regions of the highly polymorphic STR system SE33. *Int J Legal Med*, 2002, **116**(6): 365–367.

Hering, S., et al., Complex variability of intron 40 of the von Willebrand factor (vWF) gene. *Int J Legal Med*, 2008, **122**(1): 67–71.

Hill, C.R., Capillary electrophoresis and 5-channel LIF detection of a 26plex autosomal STR assay for human identification. *Methods Mol Biol*, 2012, **830**: 17–29.

Hill, C.R., et al., Concordance study between the AmpFlSTR MiniFiler PCR amplification kit and conventional STR typing kits. *J Forensic Sci*, 2007, **52**(4): 870–873.
Hill, C.R., et al., Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. *J Forensic Sci*, 2008, **53**(1): 73–80.
Holland, M.M. and W. Parson, GeneMarker® HID: A reliable software tool for the analysis of forensic STR data. *J Forensic Sci*, 2011, **56**(1): 29–35.
Holt, C.L., et al., TWGDAM validation of AmpFlSTR PCR amplification kits for forensic DNA casework. *J Forensic Sci*, 2002, **47**(1): 66–96.
Huang, Y.M., et al., Assessment of application value of 19 autosomal short tandem repeat loci of GoldenEye 20A kit in forensic paternity testing. *Int J Legal Med*, 2013, **127**(3): 587–590.
Irwin, J.A., et al., Application of low copy number STR typing to the identification of aged, degraded skeletal remains. *J Forensic Sci*, 2007, **52**(6): 1322–1327.
Jamieson, A., et al., Two-, three-, and four-person mixtures in forensic casework: Difficulties and questions. *Croat Med J*, 2011, **52**(5): 653–654; author reply 654–656.
Jiang, X., et al., Development of a 20-locus fluorescent multiplex system as a valuable tool for national DNA database. *Forensic Sci Int Genet*, 2013, **7**(2): 279–289.
Jobling, M.A., In the name of the father: Surnames and genetics. *Trends Genet*, 2001, **17**(6): 353–357.
Junge, A., et al., Validation of the multiplex kit genRESMPX-2 for forensic casework analysis. *Int J Legal Med*, 2003, **117**(6): 317–325.
Kadash, K., et al., Validation study of the TrueAllele automated data review system. *J Forensic Sci*, 2004, **49**(4): 660–667.
Katsanis, S.H. and J.K. Wagner, Characterization of the standard and recommended CODIS markers. *J Forensic Sci*, 2013, **58**(Suppl 1): S169–S172.
Kimpton, C., A. Walton, and P. Gill, A further tetranucleotide repeat polymorphism in the vWF gene. *Hum Mol Genet*, 1992, **1**(4): 287.
Kimpton, C.P., et al., Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl*, 1993, **3**(1): 13–22.
Kimpton, C., et al., Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. *Int J Legal Med*, 1994, **106**(6): 302–311.
Kline, M.C., et al., STR sequence analysis for characterizing normal, variant, and null alleles. *Forensic Sci Int Genet*, 2011, **5**(4): 329–332.
Konarzewska, M., et al., Population data and sequence analysis of a “new” microsatellite locus HumHUU (D16S3433). *Forensic Sci Int Genet*, 2010, **4**(5): e143–e144.
Krenke, B.E., et al., Validation of a 16-locus fluorescent multiplex system. *J Forensic Sci*, 2002, **47**(4): 773–785.
Laberke, P.J., et al., Method to predict the chance of developing a male profile out of mixtures of male and female DNA. *Int J Legal Med*, 2012, **126**(1): 157–160.
Laird, R., P.M. Schneider, and S. Gaudieri, Forensic STRs as potential disease markers: A study of VWA and von Willebrand’s disease. *Forensic Sci Int Genet*, 2007, **1**(3–4): 253–261.
Lareu, M.V., et al., A highly variable STR at the D12S391 locus. *Int J Legal Med*, 1996, **109**(3): 134–138.
Laurin, N., A. DeMoors, and C. Fregeau, Performance of Identifiler Direct and PowerPlex 16 HS on the applied biosystems 3730 DNA analyzer for processing biological samples archived on FTA cards. *Forensic Sci Int Genet*, 2012, **6**(5): 621–629.
Leclair, B., et al., Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples. *J Forensic Sci*, 2004, **49**(5): 968–980.
Lederer, T., et al., Characterization of two unusual allele variants at the STR locus ACTBP2 (SE33). *Forensic Sci Med Pathol*, 2008, **4**(3): 164–166.
Leibelt, C., et al., Identification of a D8S1179 primer binding site mutation and the validation of a primer designed to recover null alleles. *Forensic Sci Int*, 2003, **133**(3): 220–227.
Levedakou, E.N., et al., Characterization and validation studies of powerPlex 2.1, a nine-locus short tandem repeat (STR) multiplex system and penta D monoplex. *J Forensic Sci*, 2002, **47**(4): 757–772.
Li, H., et al., Three tetranucleotide polymorphisms for loci: D3S1352; D3S1358; D3S1359. *Hum Mol Genet*, 1993, **2**(8): 1327.
Lohmueller, K.E., Graydon et al. provide no new evidence that forensic STR loci are functional. *Forensic Sci Int Genet*, 2010, **4**(4): 273–274.

Luce, C., et al., Validation of the AmpFlSTR MiniFiler PCR amplification kit for use in forensic casework. *J Forensic Sci*, 2009, **54**(5): 1046–1054.

Lucy, D., et al., The probability of achieving full allelic representation for LCN-STR profiling of haploid cells. *Sci Justice*, 2007, **47**(4): 168–171.

Lygo, J.E., et al., The validation of short tandem repeat (STR) loci for use in forensic casework. *Int J Legal Med*, 1994, **107**(2): 77–89.

Manabe, S., et al., Mixture interpretation: Experimental and simulated reevaluation of qualitative analysis. *Leg Med (Tokyo)*, 2013, **15**(2): 66–71.

Mardini, A.C., et al., Mutation rate estimates for 13 STR loci in a large population from Rio Grande do Sul, Southern Brazil. *Int J Legal Med*, 2013, **127**(1): 45–47.

Margolis-Nunno, H., et al., A new allele of the short tandem repeat (STR) locus, CSF1PO. *J Forensic Sci*, 2001, **46**(6): 1480–1483.

Martinez-Gonzalez, L.J., et al., Intentional mixed buccal cell reference sample in a paternity case. *J Forensic Sci*, 2007, **52**(2): 397–399.

Masibay, A., T.J. Mozer, and C. Sprecher, Promega corporation reveals primer sequences in its testing kits. *J Forensic Sci*, 2000, **45**(6): 1360–1362.

Mills, K.A., D. Even, and J.C. Murray, Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). *Hum Mol Genet*, 1992, **1**(9): 779.

Miozzo, M.C., et al., Characterization of the variant allele 9.2 of Penta D locus. *J Forensic Sci*, 2007, **52**(5): 1073–1076.

Mitchell, A.A., et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. *Forensic Sci Int Genet*, 2012, **6**(6): 749–761.

Moller, A., E. Meyer, and B. Brinkmann, Different types of structural variation in STRs: HumFES/FPS, HumVWA and HumD21S11. *Int J Legal Med*, 1994, **106**(6): 319–323.

Moretti, T.R., et al., Validation of short tandem repeats (STRs) for forensic usage: Performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. *J Forensic Sci*, 2001, **46**(3): 647–660.

Mulero, J.J., et al., Development and validation of the AmpFlSTR MiniFiler PCR amplification kit: A MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA. *J Forensic Sci*, 2008, **53**(4): 838–852.

Muller, K., et al., Q8—A short amplicon multiplex including the German DNA database systems. *Forensic Sci Int Genet*, 2007, **1**(2): 205–207.

Muller, K., et al., Validation of the short amplicon multiplex q8 including the German DNA database systems. *J Forensic Sci*, 2009, **54**(4): 862–865.

Muller, K., et al., Casework testing of the multiplex kits AmpFlSTR SEfiler Plus PCR amplification kit (AB), PowerPlex S5 system (Promega) and AmpFlSTR MiniFiler PCR amplification kit (AB). *Forensic Sci Int Genet*, 2010, **4**(3): 200–205.

Narkuti, V., et al., Microsatellite mutation in the maternally/paternally transmitted D18S51 locus: Two cases of allele mismatch in the child. *Clin Chim Acta*, 2007, **381**(2): 171–175.

Narkuti, V., et al., Mother-child double incompatibility at vWA and D5S818 loci in paternity testing. *Clin Chem Lab Med*, 2007, **45**(10): 1288–1291.

Narkuti, V., et al., Single and double incompatibility at vWA and D8S1179/D21S11 loci between mother and child: Implications in kinship analysis. *Clin Chim Acta*, 2008, **395**(1–2): 162–165.

Nelson, M.S., et al., Detection of a primer-binding site polymorphism for the STR locus D16S539 using the Powerplex 1.1 system and validation of a degenerate primer to correct for the polymorphism. *J Forensic Sci*, 2002, **47**(2): 345–349.

Odriozola, A., et al., Development and validation of I-DNA1: A 15-Loci multiplex system for identity testing. *Int J Legal Med*, 2011, **125**(5): 685–694.

Odriozola, A., et al., Development and validation for identity testing of I-DNADuo, a combination of I-DNA1 and a new multiplex system, I-DNA2. *Int J Legal Med*, 2012, **126**(1): 167–172.

Odriozola, A., et al., Recent advances and considerations for the future in forensic analysis of degraded DNA by autosomic miniSTR multiplex genotyping. *Recent Pat DNA Gene Seq*, 2011, **5**(2): 110–116.

Oh, C.S., et al., Autosomal short tandem repeat analysis of ancient DNA by coupled use of mini- and conventional STR kits. *J Forensic Sci*, 2012, **57**(3): 820–825.

Oostdik, K., et al., Developmental validation of the PowerPlex® 18D system, a rapid STR multiplex for analysis of reference samples. *Forensic Sci Int Genet*, 2013, **7**(1): 129–135.

Opel, K.L., et al., The application of miniplex primer sets in the analysis of degraded DNA from human skeletal remains. *J Forensic Sci*, 2006, **51**(2): 351–356.

Opel, K.L., et al., Developmental validation of reduced-size STR miniplex primer sets. *J Forensic Sci*, 2007, **52**(6): 1263–1271.

Park, H., et al., Detection of very large off-ladder alleles at the PentaE locus in a 15 locus autosomal STR database of 199 Korean individuals. *Forensic Sci Int Genet*, 2012, **6**(6): e189–e191.

Parsons, T.J., et al., Application of novel "mini-amplicon" STR multiplexes to high volume casework on degraded skeletal remains. *Forensic Sci Int Genet*, 2007, **1**(2): 175–179.

Pascali, V.L. and S. Merigioli, Joint Bayesian analysis of forensic mixtures. *Forensic Sci Int Genet*, 2012, **6**(6): 735–748.

Perez, J., et al., Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts. *Croat Med J*, 2011, **52**(3): 314–326.

Perlin, M.W., et al., Validating TrueAllele® DNA mixture interpretation. *J Forensic Sci*, 2011, **56**(6): 1430–1447.

Perlin, M.W., J.L. Belrose, and B.W. Duceman, New York State TrueAllele® casework validation study. *J Forensic Sci*, 2013, **58**(6): 1458–1466.

Perlin, M.W., et al., TrueAllele casework on Virginia DNA mixture evidence: Computer and manual interpretation in 72 reported criminal cases. *PLoS One*, 2014, **9**(3): e92837.

Petricevic, S., et al., Validation and development of interpretation guidelines for low copy number (LCN) DNA profiling in New Zealand using the AmpFlSTR SGM Plus multiplex. *Forensic Sci Int Genet*, 2010, **4**(5): 305–310.

Pfeifer, C.M., et al., Comparison of different interpretation strategies for low template DNA mixtures. *Forensic Sci Int Genet*, 2012, **6**(6): 716–722.

Phillips, C., et al., Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. *Electrophoresis*, 2013, **34**(8): 1151–1162.

Picanco, J.B., et al., Tri-allelic pattern at the TPOX locus: A familial study. *Gene*, 2014, **535**(2): 353–358.

Pitterl, F., et al., Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations. *Int J Legal Med*, 2010, **124**(6): 551–558.

Pitterl, F., et al., The next generation of DNA profiling—STR typing by multiplexed PCR—Ion-pair RP LC-ESI time-of-flight MS. *Electrophoresis*, 2008, **29**(23): 4739–4750.

Planz, J.V., et al., Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. *Forensic Sci Int Genet*, 2012, **6**(5): 594–606.

Poetsch, M., et al., Powerplex® ES versus Powerplex® S5—Casework testing of the new screening kit. *Forensic Sci Int Genet*, 2011, **5**(1): 57–63.

Poltl, R., C. Luckenbach, and H. Ritter, The short tandem repeat locus D3S1359. *Forensic Sci Int*, 1998, **95**(2): 163–168.

Poltl, R., et al., Typing of the short tandem repeat D8S347 locus with different fluorescence markers. *Electrophoresis*, 1997, **18**(15): 2871–2873.

Polymeropoulos, M.H., et al., Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH). *Nucleic Acids Res*, 1991, **19**(13): 3753.

Puch-Solis, R., et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Sci Int Genet*, 2013, **7**(5): 555–563.

Puers, C., et al., Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder. *Am J Hum Genet*, 1993, **53**(4): 953–958.

Reichenpfader, B., R. Zehner, and M. Klintschar, Characterization of a highly variable short tandem repeat polymorphism at the D2S1242 locus. *Electrophoresis*, 1999, **20**(3): 514–517.

Repnikova, E.A., et al., Characterization of copy number variation in genomic regions containing STR loci using array comparative genomic hybridization. *Forensic Sci Int Genet*, 2013, **7**(5): 475–481.

Ricci, U., et al., A single mutation in the FGA locus responsible for false homozygosities and discrepancies between commercial kits in an unusual paternity test case. *J Forensic Sci*, 2007, **52**(2): 393–396.

Rockenbauer, E., et al., Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. *Forensic Sci Int Genet*, 2014, **8**(1): 68–72.

Ruitberg, C.M., D.J. Reeder, and J.M. Butler, STRBase: A short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res*, 2001, **29**(1): 320–322.

Schilz, F., S. Hummel, and B. Herrmann, Design of a multiplex PCR for genotyping 16 short tandem repeats in degraded DNA samples. *Anthropol Anz*, 2004, **62**(4): 369–378.

Schneider, H.R., et al., ACTBP2-nomenclature recommendations of GEDNAP. *Int J Legal Med*, 1998, **111**(2): 97–100.

Schumm, J.W., et al., A 27-locus STR assay to meet all United States and European law enforcement agency standards. *J Forensic Sci*, 2013, **58**(6): 1584–1592.

Seidl, C., et al., Sequence analysis and population data of short tandem repeat polymorphisms at loci D8S639 and D11S488. *Int J Legal Med*, 1999, **112**(6): 355–359.

Sharma, V. and M. Litt, Tetranucleotide repeat polymorphism at the D21S11 locus. *Hum Mol Genet*, 1992, **1**(1): 67.

Slooten, K. and F. Ricciardi, Estimation of mutation probabilities for autosomal STR markers. *Forensic Sci Int Genet*, 2013, **7**(3): 337–344.

Smith, R.N., Accurate size comparison of short tandem repeat alleles amplified by PCR. *Biotechniques*, 1995, **18**(1): 122–128.

Sparkes, R., et al., The validation of a 7-locus multiplex STR test for use in forensic casework. (I). Mixtures, ageing, degradation and species studies. *Int J Legal Med*, 1996, **109**(4): 186–194.

Spathis, R. and J.K. Lum, An updated validation of Promega's PowerPlex 16 system: High throughput databasing under reduced PCR volume conditions on applied biosystem's 96 capillary 3730xl DNA analyzer. *J Forensic Sci*, 2008, **53**(6): 1353–1357.

Sun, H.Y., et al., A paternity case with mutations at three CODIS core STR loci. *Forensic Sci Int Genet*, 2012, **6**(1): e61–e62.

Takayama, T., et al., Identification of a rare mutation in a TH01 primer binding site. *Leg Med (Tokyo)*, 2007, **9**(6): 289–292.

Tomsey, C.S., et al., Comparison of PowerPlex 16, PowerPlex 1.1/2.1, and ABI AmpFiSTR Profiler Plus/COfiler for forensic use. *Croat Med J*, 2001, **42**(3): 239–243.

Tucker, V.C., A.J. Kirkham, and A.J. Hopwood, Forensic validation of the PowerPlex® ESI 16 STR multiplex and comparison of performance with AmpFlSTR® SGM Plus®. *Int J Legal Med*, 2012, **126**(3): 345–356.

Tucker, V.C., et al., Developmental validation of the PowerPlex® ESI 16 and PowerPlex® ESI 17 systems: STR multiplexes for the new European standard. *Forensic Sci Int Genet*, 2011, **5**(5): 436–448.

Tucker, V.C., et al., Developmental validation of the PowerPlex® ESX 16 and PowerPlex® ESX 17 systems. *Forensic Sci Int Genet*, 2012, **6**(1): 124–131.

Tvedebrink, T., et al., Identifying contributors of DNA mixtures by means of quantitative information of STR typing. *J Comput Biol*, 2012, **19**(7): 887–902.

Tvedebrink, T., et al., Performance of two 17 locus forensic identification STR kits-applied biosystems's AmpFlSTR® NGMSElect™ and Promega's PowerPlex® ESI17 kits. *Forensic Sci Int Genet*, 2012, **6**(5): 523–531.

Tvedebrink, T., et al., Statistical model for degraded DNA samples and adjusted probabilities for allelic dropout. *Forensic Sci Int Genet*, 2012, **6**(1): 97–101.

Urquhart, A., C.P. Kimpton, and P. Gill, Sequence variability of the tetranucleotide repeat of the human beta-actin related pseudogene H-beta-Ac-psi-2 (ACTBP2) locus. *Hum Genet*, 1993, **92**(6): 637–638.

Urquhart, A., et al., Variation in short tandem repeat sequences—A survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med*, 1994, **107**(1): 13–20.

Vallone, P.M., C.R. Hill, and J.M. Butler, Demonstration of rapid multiplex PCR amplification involving 16 genetic loci. *Forensic Sci Int Genet*, 2008, **3**(1): 42–45.

Van Neste, C., et al., Forensic STR analysis using massive parallel sequencing. *Forensic Sci Int Genet*, 2012, **6**(6): 810–818.

Vanderheyden, N., et al., Identification and sequence analysis of discordant phenotypes between AmpFlSTR SGM Plus™ and PowerPlex® 16. *Int J Legal Med*, 2007, **121**(4): 297–301.

Voskoboinik, L. and A. Darvasi, Forensic identification of an individual in complex DNA mixtures. *Forensic Sci Int Genet*, 2011, **5**(5): 428–435.

Walsh, P.S., H.A. Erlich, and R. Higuchi, Preferential PCR amplification of alleles: Mechanisms and solutions. *PCR Methods Appl*, 1992, **1**(4): 241–250.

Walsh, P.S., N.J. Fildes, and R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. *Nucleic Acids Res*, 1996, **24**(14): 2807–2812.

Wang, D.Y., et al., Development and validation of the AmpFlSTR® Identifiler® direct PCR amplification kit: A multiplex assay for the direct amplification of single-source samples. *J Forensic Sci*, 2011, **56**(4): 835–845.

Wang, D.Y., et al., Developmental validation of the AmpFlSTR® Identifiler® Plus PCR amplification kit: An established multiplex assay with improved performance. *J Forensic Sci*, 2012, **57**(2): 453–465.

Warshauer, D.H., et al., STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet*, 2013, **7**(4): 409–417.

Weber, J.L. and P.E. May, Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*, 1989, **44**(3): 388–396.

Welch, L.A., et al., European Network of Forensic Science Institutes (ENFSI): Evaluation of new commercial STR multiplexes that include the European standard set (ESS) of markers. *Forensic Sci Int Genet*, 2012, **6**(6): 819–826.

Westen, A.A., et al., Assessment of the stochastic threshold, back- and forward stutter filters and low template techniques for NGM. *Forensic Sci Int Genet*, 2012, **6**(6): 708–715.

Westen, A.A., et al., Combining results of forensic STR kits: HDplex validation including allelic association and linkage testing with NGM and Identifiler loci. *Int J Legal Med*, 2012, **126**(5): 781–789.

Wetton, J.H., et al., Analysis and interpretation of mixed profiles generated by 34 cycle SGM Plus(®) amplification. *Forensic Sci Int Genet*, 2011, **5**(5): 376–380.

Wiegand, P., et al., D18S535, D1S1656 and D10S2325: Three efficient short tandem repeats for forensic genetics. *Int J Legal Med*, 1999, **112**(6): 360–363.

Wiegand, P., et al., Forensic validation of the STR systems SE 33 and TC 11. *Int J Legal Med*, 1993, **105**(6): 315–320.

Yeung, S.H., et al., Rapid and high-throughput forensic short tandem repeat typing using a 96-lane microfabricated capillary array electrophoresis microdevice. *J Forensic Sci*, 2006, **51**(4): 740–747.

Yoshida, K., et al., Efficacy of extended kinship analyses utilizing commercial STR kit in establishing personal identification. *Leg Med (Tokyo)*, 2011, **13**(1): 12–15.

Yuasa, I., et al., A hypervariable STR polymorphism in the CFI gene: Mutation rate and no linkage disequilibrium with FGA. *Leg Med (Tokyo)*, 2013, **15**(3): 161–163.

Zhang, S., et al., A new multiplex assay of 17 autosomal STRs and amelogenin for forensic application. *PLoS One*, 2013, **8**(2): e57471.

Zhou, Y., et al., DNA profiling in blood, buccal swabs and hair follicles of patients after allogeneic peripheral blood stem cells transplantation. *Leg Med (Tokyo)*, 2011, **13**(1): 47–51.

Zupanic Pajnic, I., et al., Highly efficient nuclear DNA typing of the World War II skeletal remains using three new autosomal short tandem repeat amplification kits with the extended European standard set of loci. *Croat Med J*, 2012, **53**(1): 17–23.

Genetic Population Data

Abrahams, Z., et al., Allele frequencies of six non-CODIS miniSTR loci (D1S1627, D3S4529, D5S2500, D6S1017, D8S1115 and D9S2157) in three South African populations. *Forensic Sci Int Genet*, 2011, **5**(4): 354–355.

Abu Halima, M.S., L.P. Bernal, and F.A. Sharif, Genetic variation of 15 autosomal short tandem repeat (STR) loci in the Palestinian population of Gaza Strip. *Leg Med (Tokyo)*, 2009, **11**(4): 203–204.

Aguiar, V.R., et al., Updated Brazilian STR allele frequency data using over 100,000 individuals: An analysis of CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA loci. *Forensic Sci Int Genet*, 2012, **6**(4): 504–509.

Akhteruzzaman, S., et al., Forensic evaluation of 11 non-standard STR loci in Bangladeshi population. *Leg Med (Tokyo)*, 2013, **15**(2): 106–108.

Al-Enizi, M., et al., Population genetic analyses of 15 STR loci from seven forensically-relevant populations residing in the state of Kuwait. *Forensic Sci Int Genet*, 2013, **7**(4): e106–e107.

Alenizi, M., et al., STR data for the AmpFlSTR Identifiler loci in Kuwaiti population. *Leg Med (Tokyo)*, 2008, **10**(6): 321–325.

Alenizi, M.A., et al., Concordance between the AmpFlSTR MiniFiler and AmpFlSTR Identifiler PCR amplification kits in the Kuwaiti population. *J Forensic Sci*, 2009, **54**(2): 350–352.

Ali, I., et al., Allele frequencies of 15 STR loci using AmpF/STR Identifiler kit in the Maldivian population. *Forensic Sci Int Genet*, 2012, **6**(5): e136.

Andreassen, R., S. Jakobsen, and B. Mevaag, Norwegian population data for the 10 autosomal STR loci in the AmpFlSTR® SGM Plus™ system. *Forensic Sci Int*, 2007, **170**(1): 59–61.

Andreassen, R., et al., Icelandic population data for the STR loci in the AmpFlSTR SGM Plus system and the PowerPlex Y-system. *Forensic Sci Int Genet*, 2010, **4**(4): e101–e103.

Andreini, E., et al., Allele frequencies for nine STR loci (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820) in the Italian population. *Forensic Sci Int*, 2007, **168**(1): e13–e16.

Aranda, X.G., et al., Genetic composition of six miniSTR in a Brazilian Mulatto sample population. *J Forensic Leg Med*, 2011, **18**(4): 184–186.

Asamura, H., M. Ota, and H. Fukushima, Population data on 10 non-CODIS STR loci in Japanese population using a newly developed multiplex PCR system. *J Forensic Leg Med*, 2008, **15**(8): 519–523.

Bagdonavicius, A., et al., Western Australian sub-population data for the thirteen AmpFiSTR Profiler Plus and COfiler STR loci. *J Forensic Sci*, 2002, **47**(5): 1149–1153.

Barni, F., et al., Allele frequencies of 15 autosomal STR loci in the Iraq population with comparisons to other populations from the middle-eastern region. *Forensic Sci Int*, 2007, **167**(1): 87–92.

Bindu, G.H., R. Trivedi, and V.K. Kashyap, Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India. *Forensic Sci Int*, 2007, **170**(1): 76–85.

Borosky, A., L. Catelli, and C. Vullo, Analysis of 17 STR loci in different provinces of Argentina. *Forensic Sci Int Genet*, 2009, **3**(3): e93–e95.

Branham, A., R. Wenk, and F. Chiafari, Allele frequencies of fifteen STR loci in U.S. immigrants from Haiti compared with African Americans and Afro-Caribbeans. *Forensic Sci Int Genet*, 2012, **6**(1): e3–e4.

Bright, J.A., J.S. Buckleton, and C.E. McGovern, Allele frequencies for the four major sub-populations of New Zealand for the 15 Identifiler loci. *Forensic Sci Int Genet*, 2010, **4**(2): e65–e66.

Brisighelli, F., et al., Allele frequencies of fifteen STRs in a representative sample of the Italian population. *Forensic Sci Int Genet*, 2009, **3**(2): e29–e30.

Budowle, B., D.A. Defenbaugh, and K.M. Keys, Genetic variation at nine short tandem repeat loci in Chamorros and Filipinos from Guam. *Leg Med (Tokyo)*, 2000, **2**(1): 26–30.

Budowle, B., et al., Population genetic analyses of the NGM STR loci. *Int J Legal Med*, 2011, **125**(1): 101–109.

Cai, J.F., et al., Population genetics of two STR loci D2S1322 and D2S1356 in a Chinese Han population in Hohhot. *J Forensic Sci*, 2007, **52**(4): 1001.

Camacho, M.V., C. Benito, and A.M. Figueiras, Allelic frequencies of the 15 STR loci included in the AmpFlSTR Identifiler PCR amplification kit in an autochthonous sample from Spain. *Forensic Sci Int*, 2007, **173**(2–3): 241–245.

Cassar, M., C. Farrugia, and C. Vidal, Allele frequencies of 14 STR loci in the population of Malta. *Leg Med (Tokyo)*, 2008, **10**(3): 153–156.

Castro, E.A., et al., Genetic polymorphism and forensic parameters of nine short tandem repeat loci in Ngobe and Embera Amerindians of Panama. *Hum Biol*, 2007, **79**(5): 563–577.

Chang, Y.F., et al., Genetic polymorphism of 17 STR loci in Chinese population from Hunan province in Central South China. *Forensic Sci Int Genet*, 2012, **6**(5): e151–e153.

Chen, J.G., et al., Population genetic data of 15 autosomal STR loci in Uygur ethnic group of China. *Forensic Sci Int Genet*, 2012, **6**(6): e178–e179.

Chouery, E., et al., Population genetic data for 17 STR markers from Lebanon. *Leg Med (Tokyo)*, 2010, **12**(6): 324–326.

Chula, F.G., et al., 15 STR loci frequencies with mutation rates in the population from Rio Grande do Sul, Southern Brazil. *Forensic Sci Int Genet*, 2009, **3**(2): e35–e38.

Chung, U., et al., Population data of nine miniSTR loci in Koreans. *Forensic Sci Int*, 2007, **168**(2–3): e51–e53.

Clark, D., et al., STR data for the AmpFlSTR SGM Plus loci from two South Asian populations. *Leg Med (Tokyo)*, 2009, **11**(2): 97–100.

Cortellini, V., N. Cerri, and A. Verzeletti, Genetic variation at 5 new autosomal short tandem repeat markers (D10S1248, D22S1045, D2S441, D1S1656, D12S391) in a population-based sample from Maghreb region. *Croat Med J*, 2011, **52**(3): 368–371.

Cortellini, V., N. Cerri, and A. Verzeletti, Population data on 5 non-CODIS STR loci (D10S1248, D22S1045, D2S441, D1S1656, D12S391) in a population sample from Brescia county (Northern Italy). *Forensic Sci Int Genet*, 2011, **5**(3): e97–e98.

Coudray, C., et al., Allele frequencies of 15 short tandem repeats (STRs) in three Egyptian populations of different ethnic groups. *Forensic Sci Int*, 2007, **169**(2–3): 260–265.

Coudray, C., et al., Allele frequencies of 15 tetrameric short tandem repeats (STRs) in Andalusians from Huelva (Spain). *Forensic Sci Int*, 2007, **168**(2–3): e21–e24.

Coudray, C., et al., Population genetic data of 15 tetrameric short tandem repeats (STRs) in Berbers from Morocco. *Forensic Sci Int*, 2007, **167**(1): 81–86.

Curic, G., et al., Genetic parameters of five new European standard set STR loci (D10S1248, D22S1045, D2S441, D1S1656, D12S391) in the population of Eastern Croatia. *Croat Med J*, 2012, **53**(5): 409–415.

da Costa Francez, P.A., et al., Allelic frequencies and statistical data obtained from 12 codis STR loci in an admixed population of the Brazilian Amazon. *Genet Mol Biol*, 2011, **34**(1): 35–39.

de Assis Poiares, L., et al., Allele frequencies of 15 STRs in a representative sample of the Brazilian population. *Forensic Sci Int Genet*, 2010, **4**(2): e61–e63.

Demeter, S.J., et al., Genetic variation at 15 polymorphic, autosomal, short tandem repeat loci of two Hungarian populations in Transylvania, Romania. *Croat Med J*, 2010, **51**(6): 515–523.

Deng, Q., et al., Genetic relationships among four minorities in Guangxi revealed by analysis of 15 STRs. *J Genet Genomics*, 2007, **34**(12): 1072–1079.

Deng, Y., et al., Genetic polymorphisms of 15 STR loci of Chinese Dongxiang and Salar ethnic minority living in Qinghai Province of China. *Leg Med (Tokyo)*, 2007, **9**(1): 38–42.

Deng, Y.J., et al., Genetic polymorphism analysis of 15 STR loci in Chinese Hui ethnic group residing in Qinghai province of China. *Mol Biol Rep*, 2011, **38**(4): 2315–2322.

Deng, Y.J., et al., Polymorphic analysis of 21 new STR loci in Chinese Uigur group. *Forensic Sci Int Genet*, 2013, **7**(3): e97–e98.

Di Cristofaro, J., et al., Genetic data of 15 STR loci in five populations from Afghanistan. *Forensic Sci Int Genet*, 2012, **6**(1): e44–e45.

Dognaux, S., et al., Allele frequencies for the new European standard set (ESS) loci and D1S1677 in the Belgian population. *Forensic Sci Int Genet*, 2012, **6**(2): e75–e77.

dos Santos, L.L., et al., Allele distribution of six STR/miniSTR loci (CD4, FABP2, D12S391, D14S1434, D22S1045 and D10S1248) for forensic purposes in Southeastern Brazil. *Ann Hum Biol*, 2011, **38**(1): 110–113.

Du, B., et al., Allele frequencies for two STR loci D6S1274 and D17S1299 in Chinese and Thai populations. *J Forensic Sci*, 2007, **52**(1): 225.

Dubey, B., et al., Forensic STR profile of two endogamous populations of Madhya Pradesh, India. *Leg Med (Tokyo)*, 2009, **11**(1): 41–44.

Eaaswarkhanth, M., S. Roy, and I. Haque, Allele frequency distribution for 15 autosomal STR loci in two Muslim populations of Tamilnadu, India. *Leg Med (Tokyo)*, 2007, **9**(6): 332–335.

Eaaswarkhanth, M., T.S. Vasulu, and I. Haque, Genetic affinity between diverse ethnoreligious communities of Tamil Nadu, India: A microsatellite study. *Hum Biol*, 2008, **80**(6): 601–609.

Eaaswarkhanth, M., et al., Diverse genetic origin of Indian Muslims: Evidence from autosomal STR loci. *J Hum Genet*, 2009, **54**(6): 340–348.

Eaaswarkhanth, M., et al., Microsatellite diversity delineates genetic relationships of Shia and Sunni Muslim populations of Uttar Pradesh, India. *Hum Biol*, 2009, **81**(4): 427–445.

Eckhoff, C., S.J. Walsh, and J.S. Buckleton, Population data from sub-populations of the Northern Territory of Australia for 15 autosomal short tandem repeat (STR) loci. *Forensic Sci Int*, 2007, **171**(2–3): 237–249.

El Andari, A., et al., Population genetic data for 23 STR markers from Lebanon. *Forensic Sci Int Genet*, 2013, **7**(4): e108–e113.

El Ossmani, H., et al., Assessment of phylogenetic structure of Berber-speaking population of Azrou using 15 STRs of Identifiler kit. *Leg Med (Tokyo)*, 2010, **12**(1): 52–56.

Elmrghni, S., et al., Genetic data provided by 15 autosomal STR loci in the Libyan population living in Benghazi. *Forensic Sci Int Genet*, 2012, **6**(3): e93–e94.

Erkol, Z., et al., STR data for the AmpFlSTR identifier loci from an old settlement in Northwestern Turkey. *Forensic Sci Int*, 2007, **173**(2–3): 238–240.

Ferdous, A., et al., Forensic evaluation of STR data for the PowerPlex 16 system loci in a Bangladeshi population. *Leg Med (Tokyo)*, 2009, **11**(4): 198–199.

Forward, B.W., et al., AmpPFlSTR Identifiler STR allele frequencies in Tanzania, Africa. *J Forensic Sci*, 2008, **53**(1): 245.

Fridman, C., et al., Brazilian population profile of 15 STR markers. *Forensic Sci Int Genet*, 2008, **2**(2): e1–e4.

Fujihara, J., et al., Allele frequencies for nine STR loci in Ovambo population using AmpFlSTR Profiler Kit. *Forensic Sci Int*, 2007, **169**(1): e7–e9.

Gazi, N.N., et al., Genetic polymorphisms of 15 autosomal STR loci in three isolated tribal populations of Bangladesh. *Forensic Sci Int Genet*, 2010, **4**(4): 265–266.

Goetz, R., et al., Population data from the New South Wales Aboriginal Australian sub-population for the Profiler Plus autosomal short tandem repeat (STR) loci. *Forensic Sci Int*, 2008, **175**(2–3): 235–237.

Gomes, A.V., et al., 13 STR loci frequencies in the population from Paraiba, Northeast Brazil. *Forensic Sci Int*, 2007, **173**(2–3): 231–234.

Gomes, V., et al., Population data defined by 15 autosomal STR loci in Karamoja population (Uganda) using AmpFlSTR Identifiler kit. *Forensic Sci Int Genet*, 2009, **3**(2): e55–e58.

Gomez, M.A., et al., Allele frequencies of 15 STR loci in the population of the city of Quito, Ecuador. *J Forensic Sci*, 2008, **53**(2): 510–511.

Gorostiza, A., et al., Allele frequencies of the 15 AmpFlSTR Identifiler loci in the population of Metztitlan (Estado de Hidalgo), Mexico. *Forensic Sci Int*, 2007, **166**(2–3): 230–232.

Grubwieser, P., et al., Evaluation of an extended set of 15 candidate STR loci for paternity and kinship analysis in an Austrian population sample. *Int J Legal Med*, 2007, **121**(2): 85–89.

Gutierrez, C.C., et al., Population genetic data for 15 STR loci (PowerPlex 16 kit) in Nicaragua. *Forensic Sci Int Genet*, 2011, **5**(5): 563–564.

Hatzer-Grubwieser, P., et al., Allele frequencies and concordance study of 16 STR loci—Including the new European standard set (ESS) loci—In an Austrian population sample. *Forensic Sci Int Genet*, 2012, **6**(1): e50–e51.

Havas, D., et al., Population genetics of 15 AmpFlSTR Identifiler loci in Macedonians and Macedonian Romani (Gypsy). *Forensic Sci Int*, 2007, **173**(2–3): 220–224.

He, J. and F. Guo, Allele frequencies for fifteen autosomal STR loci in a Xibe population from Liaoning Province, Northeast China. *Forensic Sci Int Genet*, 2013, **7**(3): e80–e81.

He, J. and F. Guo, Genetic variation of fifteen autosomal STR loci in a Manchu population from Jilin Province, Northeast China. *Forensic Sci Int Genet*, 2013, **7**(2): e45–e46.

Herrera-Paz, E.F., et al., Allele frequencies distributions for 13 autosomal STR loci in 3 Black Carib (Garifuna) populations of the Honduran Caribbean coasts. *Forensic Sci Int Genet*, 2008, **3**(1): e5–e10.

Hill, C.R., et al., U.S. population data for 29 autosomal STR loci. *Forensic Sci Int Genet*, 2013, **7**(3): e82–e83.

Hong, S.B., et al., Korean population genetic data and concordance for the PowerPlex® ESX 17, AmpFlSTR Identifiler®, and PowerPlex® 16 systems. *Forensic Sci Int Genet*, 2013, **7**(3): e47–e51.

Hou, G., et al., Genetic distribution on 15 STR loci from a Han population of Shenyang region in northeast China. *Forensic Sci Int Genet*, 2013, **7**(3): e86–e87.

Hou, Y.P., et al., D20S161 data for three ethnic populations and forensic validation. *Int J Legal Med*, 1999, **112**(6): 400–402.

Huang, Q., et al., Genetic polymorphism of 15 STR loci in Chinese Han population from Shanghai municipality in East China. *Forensic Sci Int Genet*, 2013, **7**(2): e31–e34.

Huang, S., et al., Genetic variation analysis of 15 autosomal STR loci of AmpFlSTR Sinofiler PCR amplification kit in Henan (central China) Han population. *Leg Med (Tokyo)*, 2010, **12**(3): 160–161.

Hwa, H.L., et al., Fifteen non-CODIS autosomal short tandem repeat loci multiplex data from nine population groups living in Taiwan. *Int J Legal Med*, 2012, **126**(4): 671–675.

Hwa, H.L., et al., Fourteen non-CODIS autosomal short tandem repeat loci multiplex data from Taiwanese. *Int J Legal Med*, 2011, **125**(2): 219–226.

Illeperuma, R.J., et al., Genetic profile of 11 autosomal STR loci among the four major ethnic groups in Sri Lanka. *Forensic Sci Int Genet*, 2009, **3**(3): e105–e106.

Illeperuma, R.J., et al., Genetic variation at 11 autosomal STR loci in the aboriginal people, the Veddahs of Sri Lanka. *Forensic Sci Int Genet*, 2010, **4**(2): 142.

Jacewicz, R., et al., Population database on 15 autosomal STR loci in 1000 unrelated individuals from the Lodz region of Poland. *Forensic Sci Int Genet*, 2008, **2**(1): e1–e3.

Jedrzejczyk, M., et al., Genetic population studies on 15 NGM STR loci in central Poland population. *Forensic Sci Int Genet*, 2012, **6**(4): e119–e120.

Jeran, N., et al., Genetic diversity of 15 STR loci in a population of Montenegro. *Coll Antropol*, 2007, **31**(3): 847–852.

Jin, H.J., et al., Forensic and population genetic analyses of eighteen non-CODIS miniSTR loci in the Korean population. *J Forensic Leg Med*, 2013, **20**(8): 1093–1097.

Juarez-Cedillo, T., et al., Genetic admixture and diversity estimations in the Mexican Mestizo population from Mexico City using 15 STR polymorphic markers. *Forensic Sci Int Genet*, 2008, **2**(3): e37–e39.

Kalpana, D., et al., Pentaplex typing of new European standard set (ESS) STR loci in Indian population. *Forensic Sci Int Genet*, 2012, **6**(3): e86–e89.

Kang, L., et al., Genetic polymorphisms of 15 STR loci in two Tibetan populations from Tibet Changdu and Naqu, China. *Forensic Sci Int*, 2007, **169**(2–3): 239–243.

Kang, L.L., et al., Allele frequencies of 15 STR loci of Tibetan lived in Tibet Lassa. *Forensic Sci Int*, 2007, **168**(2–3): 236–240.

Kee, B.P., et al., Genetic data for 15 STR loci in a Kadazan-Dusun population from East Malaysia. *Genet Mol Res*, 2011, **10**(2): 739–743.

Khodjet-el-Khil, H., et al., Allele frequencies for 15 autosomal STR markers in the Libyan population. *Ann Hum Biol*, 2012, **39**(1): 80–83.

Kido, A., et al., Population data on the AmpFlSTR Identifiler loci in Africans and Europeans from South Africa. *Forensic Sci Int*, 2007, **168**(2–3): 232–235.

Korzebor, A., et al., Statistical analysis of six STR loci located in MHC region in Iranian population for pre-implantation genetic diagnosis. *Int J Immunogenet*, 2007, **34**(6): 441–443.

Kraaijenbrink, T., et al., Allele frequency distribution for 21 autosomal STR loci in Bhutan. *Forensic Sci Int*, 2007, **170**(1): 68–72.

Kraaijenbrink, T., et al., Allele frequency distribution for 21 autosomal STR loci in Nepal. *Forensic Sci Int*, 2007, **168**(2–3): 227–231.

Krithika, S., et al., Allele frequency distribution at 15 autosomal STR loci in Panggi, Komkar and Padam sub tribes of Adi, a Tibeto-Burman speaking population of Arunachal Pradesh, India. *Leg Med (Tokyo)*, 2007, **9**(4): 210–217.

Lee, M.H., et al., Genetic variation of three autosomal STR loci D21S1435, D21S1411, and D21S1412 in Korean population. *Mol Biol Rep*, 2010, **37**(1): 99–104.

Li, B., et al., Population genetic analyses of the STR loci of the AmpFlSTR NGM SElect kit for Han population in Fujian Province, China. *Int J Legal Med*, 2013, **127**(2): 345–346.

Li, C., et al., Genetic polymorphism of 17 STR loci for forensic use in Chinese population from Shanghai in East China. *Forensic Sci Int Genet*, 2009, **3**(4): e117–e118.

Li, S., et al., Allele frequencies of nine non-CODIS STR loci in Chinese Uyghur ethnic minority group. *Forensic Sci Int Genet*, 2012, **6**(1): e11–e12.

Liao, G., et al., Population genetic study of 15 STR loci in a Chinese population. *J Forensic Sci*, 2008, **53**(1): 252–253.

Liu, C., et al., Genetic polymorphisms of 9 non-CODIS short tandem repeat loci in two ethnic minority populations in Southern China. *Forensic Sci Int Genet*, 2013, **7**(4): e114–e115.

Liu, J., et al., Allele frequencies of 19 autosomal STR loci in Manchu population of China with phylogenetic structure among worldwide populations. *Gene*, 2013, **529**(2): 282–287.

Lu, D.J., Q.L. Liu, and H. Zhao, Genetic data of nine non-CODIS STRs in Chinese Han population from Guangdong Province, Southern China. *Int J Legal Med*, 2011, **125**(1): 133–137.

Manamperi, A., et al., STR polymorphisms in Sri Lanka: Evaluation of forensic utility in identification of individuals and parentage testing. *Ceylon Med J*, 2009, **54**(3): 85–89.

Manta, F.S., et al., Terena Amerindian group autosomal STR data: Comparison studies with other Brazilian populations. *Mol Biol Rep*, 2012, **39**(4): 4455–4459.

Marian, C., et al., STR data for the 15 AmpFlSTR identifiler loci in the Western Romanian population. *Forensic Sci Int*, 2007, **170**(1): 73–75.

Martin, P., et al., Allele frequencies of six miniSTR loci (D10S1248, D14S1434, D22S1045, D4S2364, D2S441 and D1S1677) in a Spanish population. *Forensic Sci Int*, 2007, **169**(2–3): 252–254.

Martinez, B., J.J. Builes, and L. Caraballo, Genetic data analysis of nine STRs in two Caribbean Colombian populations: Cesar and Guajira. *J Forensic Sci*, 2008, **53**(1): 254–255.

Martinez-Cortes, G., et al., Origin and genetic differentiation of three native Mexican groups (Purepechas, Triquis and Mayas): Contribution of CODIS-STRs to the history of human populations of Mesoamerica. *Ann Hum Biol*, 2010, **37**(6): 801–819.

Martins, J.A., et al., Genetic data of 15 autosomal STR loci: An analysis of the Araraquara population colonization (Sao Paulo, Brazil). *Mol Biol Rep*, 2011, **38**(8): 5397–5403.

Maruyama, S., et al., Population data on 15 STR loci using AmpFlSTR Identifiler kit in a Malay population living in and around Kuala Lumpur, Malaysia. *Leg Med (Tokyo)*, 2008, **10**(3): 160–162.

Mastana, S.S., et al., Genetic variation of 13 STR loci in the four endogamous tribal populations of Eastern India. *Forensic Sci Int*, 2007, **169**(2–3): 266–273.

Melo, M.M., et al., Genetic study of 15 STRs loci of Identifiler system in Angola population. *Forensic Sci Int Genet*, 2010, **4**(5): e153–e157.

Montelius, K., A.O. Karlsson, and G. Holmlund, STR data for the AmpFlSTR Identifiler loci from Swedish population in comparison to European, as well as with non-European population. *Forensic Sci Int Genet*, 2008, **2**(3): e49–e52.

Monterrosa, J.C., et al., Population genetic data for 16 STR loci (PowerPlex ESX-17 kit) in El Salvador. *Forensic Sci Int Genet*, 2012, **6**(5): e134.

Mornhinweg, E., et al., D3S1358: Sequence analysis and gene frequency in a German population. *Forensic Sci Int*, 1998, **95**(2): 173–178.

Munoz, A., et al., Allele frequencies of 15 STRs in the Calchaqui Valleys population (North-Western Argentina). *Forensic Sci Int Genet*, 2012, **6**(1): e58–e60.

Nascimento, E., E. Cerqueira, and L. Gusmao, Population database defined by 13 autosomal STR loci in a representative sample from Bahia, Northeast Brazil. *Forensic Sci Int Genet*, 2011, **5**(1): e38–e40.

Nasibov, E., et al., Allele frequencies of 15 STR loci in Azerbaijan population. *Forensic Sci Int Genet*, 2013, **7**(3): e99–e100.

Nie, S., et al., Genetic data of 15 STR loci in Chinese Yunnan Han population. *Forensic Sci Int Genet*, 2008, **3**(1): e1–e3.

Noor, S., S. Roy, and I. Haque, An autosomal STR database of Muslims: The largest minority community, Uttar Pradesh, India. *Forensic Sci Int Genet*, 2011, **5**(4): e117–e118.

Noor, S., et al., Allele frequency distribution for 15 autosomal STR loci in Afridi Pathan population of Uttar Pradesh, India. *Leg Med (Tokyo)*, 2009, **11**(6): 308–311.

Novokmet, N., et al., Forensic efficiency parameters for the 15 STR loci in the population of the island of Cres (Croatia). *Coll Antropol*, 2009, **33**(4): 1319–1322.

Ocampos, M., et al., 15 STR loci frequencies in the population from Santa Catarina, Southern Brazil. *Forensic Sci Int Genet*, 2009, **3**(4): e129–e131.

Omran, G.A., G.N. Rutty, and M.A. Jobling, Genetic variation of 15 autosomal STR loci in Upper (Southern) Egyptians. *Forensic Sci Int Genet*, 2009, **3**(2): e39–e44.

Ossmani, H.E., et al., Allele frequencies of 15 autosomal STR loci in the Southern Morocco population with phylogenetic structure among worldwide populations. *Leg Med (Tokyo)*, 2009, **11**(3): 155–158.

Ota, M., et al., Allele frequencies for 15 STR loci in Tibetan populations from Nepal. *Forensic Sci Int*, 2007, **169**(2–3): 234–238.

Park, J.H., et al., Genetic variation of 23 autosomal STR loci in Korean population. *Forensic Sci Int Genet*, 2013, **7**(3): e76–e77.

Parys-Proszek, A., et al., Genetic variation of 15 autosomal STR loci in a population sample from Poland. *Leg Med (Tokyo)*, 2010, **12**(5): 246–248.

Pepinski, W., et al., Polymorphism of 11 non-CODIS STRs in a population sample of Lithuanian minority residing in northeastern Poland. *Forensic Sci Int Genet*, 2011, **5**(1): e37.

Pepinski, W., et al., Polymorphism of 11 non-CODIS STRs in a population sample of religious minority of Old Believers residing in Northeastern Poland. *Adv Med Sci*, 2010, **55**(2): 328–332.

Petric, G., et al., Genetic variation at 15 autosomal STR loci in the Hungarian population of Vojvodina Province, Republic of Serbia. *Forensic Sci Int Genet*, 2012, **6**(6): e163–e165.

Phillips, C., et al., Analysis of global variability in 15 established and 5 new European standard set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Sci Int Genet*, 2011, **5**(3): 155–169.

Piatek, J., et al., Population genetics of 15 autosomal STR loci in the population of Pomorze Zachodnie (NW Poland). *Forensic Sci Int Genet*, 2008, **2**(3): e41–e43.

Pinto, L.M., et al., Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian populations. *Forensic Sci Int Genet*, 2014, **9**: e16–e17.

Poetsch, M. and N. von Wurmb-Schwark, Allele frequencies for the 16 short tandem repeats of the Powerplex ESX17 kit in a population from Turkey. *Int J Legal Med*, 2013, **127**(3): 591–592.

Poetsch, M., et al., First experiences using the new Powerplex® ESX17 and ESI17 kits in casework analysis and allele frequencies for two different regions in Germany. *Int J Legal Med*, 2011, **125**(5): 733–739.

Poiares Lde, A., et al., 15 STR loci frequencies in the population from Parana, Southern Brazil. *Forensic Sci Int Genet*, 2009, **4**(1): e23–e24.

Pontes, M.L. and M.F. Pinheiro, Population data of the AmpFISTR® NGM STR loci in a North of Portugal sample. *Forensic Sci Int Genet*, 2012, **6**(5): e127–e128.

Porras, L., et al., Genetic polymorphism of 15 STR loci in central Western Colombia. *Forensic Sci Int Genet*, 2008, **2**(1): e7–e8.

Projic, P., et al., Allele frequencies for 15 short tandem repeat loci in representative sample of Croatian population. *Croat Med J*, 2007, **48**(4): 473–477.

Raczek, E., Population data on the 11 STR loci in the Upper Silesia (Poland). *Forensic Sci Int*, 2007, **168**(1): 68–72.

Rak, S.A., et al., Population genetic data on 15 STR loci in the Hungarian population. *Forensic Sci Int Genet*, 2011, **5**(5): 543–544.

Rakha, A., et al., Genetic analysis of Kashmiri Muslim population living in Pakistan. *Leg Med (Tokyo)*, 2008, **10**(4): 216–219.

Rakha, A., et al., Population genetic data on 15 autosomal STRs in a Pakistani population sample. *Leg Med (Tokyo)*, 2009, **11**(6): 305–307.

Rangel-Villalobos, H., et al., Evaluation of forensic and anthropological potential of D9S1120 in Mestizos and Amerindian populations from Mexico. *Croat Med J*, 2012, **53**(5): 423–431.

Rangel-Villalobos, H., et al., Forensic parameters for 15 STRs in eight Amerindian populations from the north and west of Mexico. *Forensic Sci Int Genet*, 2013, **7**(3): e62–e65.

Rebala, K., et al., Belarusian population genetic database for 15 autosomal STR loci. *Forensic Sci Int*, 2007, **173**(2–3): 235–237.

Riccardi, L.N., et al., Genetic profiling of Bolivian population using 15 STR markers of forensic importance. *Leg Med (Tokyo)*, 2009, **11**(3): 149–151.

Rocchi, A., et al., Italian data of 23 STR loci amplified in a single multiplex reaction. *Forensic Sci Int Genet*, 2012, **6**(6): e157–e158.

Rodenbusch, R., et al., Allele frequencies of the five new generation forensic STR (D1S1656, D2S441, D10S1248, D12S391 and D22S1045) in the population from Rio Grande do Sul, Southern Brazil. *Forensic Sci Int Genet*, 2012, **6**(1): e55–e57.

Rodrigues, E.L., et al., Genetic data on 15 STR autosomal loci for a sample population of the Northern Region of the State of Rio de Janeiro, Brazil. *Forensic Sci Int Genet*, 2009, **4**(1): e25–e26.

Rodrigues, E.M., J. Palha Tde, and S.E. dos Santos, Allele frequencies data and statistic parameters for 13 STR loci in a population of the Brazilian Amazon Region. *Forensic Sci Int*, 2007, **168**(2–3): 244–247.

Rodriguez, A., et al., Population genetic data for 18 STR loci in Costa Rica. *Forensic Sci Int*, 2007, **168**(1): 85–88.

Roy, S., et al., Autosomal STR variations in three endogamous populations of West Bengal, India. *Leg Med (Tokyo)*, 2008, **10**(6): 326–332.

Rubi-Castellanos, R., et al., Genetic data of 15 autosomal STRs (Identifiler kit) of three Mexican Mestizo population samples from the States of Jalisco (West), Puebla (Center), and Yucatan (Southeast). *Forensic Sci Int Genet*, 2009, **3**(3): e71–e76.

Sanchez-Diz, P., et al., 16 STR data of a Greek population. *Forensic Sci Int Genet*, 2008, **2**(4): e71–e72.

Sanchez-Diz, P., et al., Population data on 15 autosomal STRs in a sample from Colombia. *Forensic Sci Int Genet*, 2009, **3**(3): e81–e82.

Santos, C., et al., Genetic structure of the Azores Islands: A study using 15 autosomal short tandem repeat loci. *Coll Antropol*, 2009, **33**(4): 991–999.

Schlebusch, C.M., H. Soodyall, and M. Jakobsson, Genetic variation of 15 autosomal STR loci in various populations from southern Africa. *Forensic Sci Int Genet*, 2012, **6**(1): e20–e21.

Shi, M., et al., Genetic polymorphism of 14 non-CODIS STR loci for forensic use in Southeast China population. *Forensic Sci Int*, 2008, **174**(1): 77–80.

Shotivaranon, J., et al., DNA database of populations from different parts in the Kingdom of Thailand. *Forensic Sci Int Genet*, 2009, **4**(1): e37–e38.

Silva, B.C., et al., Genetic diversity and statistical parameters of 15 autosomal STR loci in the Pomeranian subpopulation of Espirito Santo State, Brazil. *Mol Biol Rep*, 2011, **38**(5): 3013–3016.

Silva, M.B., et al., Allele frequencies of fifteen STR loci in a population from Central Brazil. *Forensic Sci Int Genet*, 2010, **4**(5): e151–e152.

Simkova, H., et al., Allele frequency data for 17 short tandem repeats in a Czech population sample. *Forensic Sci Int Genet*, 2009, **4**(1): e15–e17.

Smith, B.G., et al., Population data for 15 STR loci (Identifiler kit) in a Filipino population. *Leg Med (Tokyo)*, 2009, **11**(3): 159–161.

Soltyszewski, I., et al., Analysis of forensically used autosomal short tandem repeat markers in Polish and neighboring populations. *Forensic Sci Int Genet*, 2008, **2**(3): 205–211.

Sotak, M., et al., Genetic variation analysis of 15 autosomal STR loci in Eastern Slovak Caucasian and Romany (Gypsy) population. *Forensic Sci Int Genet*, 2008, **3**(1): e21–e25.

Sotak, M., et al., Population database of 17 autosomal STR loci from the four predominant Eastern Slovakia regions. *Forensic Sci Int Genet*, 2011, **5**(3): 262–263.

Stanciu, F., O.R. Popescu, and I.M. Stoian, Allele frequencies of 15 STR loci in Moldavia region (NE Romania). *Forensic Sci Int Genet*, 2009, **4**(1): e39–e40.

Stanciu, F., O.R. Popescu, and I.M. Stoian, STR data for the AmpFlSTR Identifiler from Dobruja region (SE Romania). *Forensic Sci Int Genet*, 2009, **3**(2): 146–147.

Stanciu, F., I.M. Stoian, and O.R. Popescu, Comprehensive STR data for the AmpFlSTR Identifiler from Transylvania (NW Romania). *Leg Med (Tokyo)*, 2009, **11**(1): 48–50.

Stanciu, F., I.M. Stoian, and O.R. Popescu, Population data for 15 short tandem repeat loci from Wallachia Region, South Romania. *Croat Med J*, 2009, **50**(3): 321–325.

Stanciu, F., et al., Genetic parameters and allele frequencies of five new European standard set STR loci (D10S1248, D22S1045, D2S441, D1S1656, D12S391) in the population of Romania. *Croat Med J*, 2013, **54**(3): 232–237.

Steinlechner, M., et al., Population genetics of ten STR loci (AmpFlSTR SGM plus) in Austria. *Int J Legal Med*, 2001, **114**(4–5): 288–290.

Stepanov, V.A., et al., Genetic variability of 15 autosomal STR loci in Russian populations. *Leg Med (Tokyo)*, 2010, **12**(5): 256–258.

Suadi, Z., et al., STR data for the AmpFlSTR Identifiler loci from the three main ethnic indigenous population groups (Iban, Bidayuh, and Melanau) in Sarawak, Malaysia. *J Forensic Sci*, 2007, **52**(1): 231–234.

Talledo, M., et al., Comparative allele distribution at 16 STR loci between the Andean and coastal population from Peru. *Forensic Sci Int Genet*, 2010, **4**(4): e109–e117.

Tamaki, K., et al., Evaluation of tetranucleotide repeat locus D7S809 (wg1g9) in the Japanese population. *Forensic Sci Int*, 1996, **81**(2–3): 133–140.

Tang, J., et al., Genetic analyzing of 15 STR loci in a Han population of Jinan (Northern China). *Leg Med (Tokyo)*, 2009, **11**(3): 144–146.

Taylor, D.A., J.M. Henry, and S.J. Walsh, South Australian Aboriginal sub-population data for the nine AmpFlSTR Profiler Plus short tandem repeat (STR) loci. *Forensic Sci Int Genet*, 2008, **2**(2): e27–e30.

Teng, Y., et al., Genetic variation of new 21 autosomal short tandem repeat loci in a Chinese Salar ethnic group. *Mol Biol Rep*, 2012, **39**(2): 1465–1470.

Thangaraj, K., et al., Autosomal STR data on the enigmatic Andaman Islanders. *Forensic Sci Int*, 2007, **169**(2–3): 247–251.

Tillmar, A.O., G. Backstrom, and K. Montelius, Genetic variation of 15 autosomal STR loci in a Somali population. *Forensic Sci Int Genet*, 2009, **4**(1): e19–e20.

Tong da, Y., et al., Polymorphism analysis and evaluation of nine non-CODIS STR loci in the Han population of Southern China. *Ann Hum Biol*, 2010, **37**(6): 820–826.

Tong, D., et al., Polymorphism analysis of 15 STR loci in a large sample of the Han population in Southern China. *Forensic Sci Int Genet*, 2009, **4**(1): e27–e29.

Tong, D., et al., Polymorphism analysis and evaluation of 19 STR loci in the Han population of Southern China. *Ann Hum Biol*, 2013, **40**(2): 191–196.

Toscanini, U., et al., Testing for genetic structure in different urban Argentinian populations. *Forensic Sci Int*, 2007, **165**(1): 35–40.

Turrina, S., et al., Evaluation of genetic parameters of 22 autosomal STR loci (PowerPlex® Fusion System) in a population sample from Northern Italy. *Int J Legal Med*, 2014, **128**(2): 281–283.

Uchihi, R., et al., Population data of six STR loci using hexaplex PCR amplification and typing system, "Midi-6" in four regional populations in Japan. *Forensic Sci Int*, 2007, **169**(2–3): 255–259.

Untoro, E., et al., Allele frequency of CODIS 13 in Indonesian population. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S203–S205.

Vargas-Diaz, R., et al., A genetic study of the Identifiler(™)System 15 STR loci in the general population of Nicaragua, Central America. *Leg Med (Tokyo)*, 2011, **13**(4): 213–214.

Vergara, I.A., et al., Autosomal STR allele frequencies for the CODIS system from a large random population sample in Chile. *Forensic Sci Int Genet*, 2012, **6**(3): e83–e85.

Veselinovic, I.S., et al., Serbian population data at the CODIS core STR loci. *J Forensic Sci*, 2007, **52**(6): 1426–1427.

Vieira, T.C., et al., Allelic frequencies and statistical data obtained from 15 STR loci in a population of the Goias State. *Genet Mol Res*, 2013, **12**(1): 23–27.

Vullo, C., et al., Frequency data for 12 mini STR loci in Argentina. *Forensic Sci Int Genet*, 2010, **4**(3): e79–e81.

Walsh, S.J. and J.S. Buckleton, Autosomal microsatellite allele frequencies for 15 regionally defined Aboriginal Australian population datasets. *Forensic Sci Int*, 2007, **168**(2–3): e29–e42.

Walsh, S.J. and J.S. Buckleton, Autosomal microsatellite allele frequencies for a nationwide dataset from the Australian Caucasian sub-population. *Forensic Sci Int*, 2007, **168**(2–3): e47–e50.

Wang, H.D., et al., Allelic frequency distributions of 21 non-combined DNA index system STR loci in a Russian ethnic minority group from Inner Mongolia, China. *J Zhejiang Univ Sci B*, 2013, **14**(6): 533–540.

Wang, J., et al., Allele frequencies of nine non-CODIS STR loci in Western Chinese Han population. *Forensic Sci Int Genet*, 2013, **7**(3): e88–e89.

Wang, R., et al., Genetic distribution on 15 STR loci from a population of Southern Liaoning in Northeast of China. *Forensic Sci Int Genet*, 2008, **2**(2): e25–e26.

Wei, Y., et al., Genetic diversity of five STR loci in Southeastern China. *J Forensic Sci*, 2007, **52**(4): 999–1000.

Wu, Y.M., et al., Genetic polymorphisms of 15 STR loci in Chinese Han population living in Xi'an city of Shaanxi Province. *Forensic Sci Int Genet*, 2008, **2**(2): e15–e18.

Xing, J., et al., Genetic polymorphism of 15 STR loci in a Manchu population in Northeast China. *Forensic Sci Int Genet*, 2011, **5**(3): e93–e95.

Xu, L.N., S.P. Hu, and G.Y. Feng, STR polymorphisms of the Henan population and investigation of the Central Plains Han origin of Chaoshanese. *Biochem Genet*, 2009, **47**(7–8): 569–581.

Yan, J., et al., Genetic analysis of 15 STR loci on Chinese Tibetan in Qinghai Province. *Forensic Sci Int*, 2007, **169**(1): e3–e6.

Yan, J., et al., Genetic polymorphism of two novel STR loci AY639920 and AY639923 in a Chinese population. *J Forensic Sci*, 2007, **52**(3): 754.

Yong, R.Y., et al., Allele frequencies of six miniSTR loci of three ethnic populations in Singapore. *Forensic Sci Int*, 2007, **166**(2–3): 240–243.

Yoo, S.Y., et al., A large population genetic study of 15 autosomal short tandem repeat loci for establishment of Korean DNA profile database. *Mol Cells*, 2011, **32**(1): 15–19.

Yuan, L., et al., Population data of 21 non-CODIS STR loci in Han population of Northern China. *Int J Legal Med*, 2012, **126**(4): 659–664.

Zeng, Z., et al., Genetic data of six STR loci of Han population in Henan Province (Central China). *J Forensic Sci*, 2008, **53**(2): 515–516.

Zeng, Z.S., et al., Population genetic data of 15 STR loci in Han population of Henan province (central China). *Leg Med (Tokyo)*, 2007, **9**(1): 30–32.

Zha, L., et al., Genetic polymorphism of 21 non-CODIS STR loci in the Chinese Mongolian ethnic minority. *Forensic Sci Int Genet*, 2014, **9**: e32–e33.

Zhang, H., et al., Analysis of 15 STR loci in Chinese population from Sichuan in West China. *Forensic Sci Int*, 2007, **171**(2–3): 222–225.

Zhang, H.B., et al., Nine polymorphic STR loci in the HLA region in the Shaanxi Han population of China. *Genet Mol Res*, 2012, **11**(3): 2534–2538.

Zhang, S., et al., Genetic polymorphisms in 12 autosomal STRs in a Shanghai Han population from China. *Electrophoresis*, 2013, **34**(4): 613–617.

Zheng, X., et al., Genetic data of 9 STR loci from Henan Province (central China). *Forensic Sci Int*, 2007, **169**(2–3): 244–246.

Zhivotovsky, L.A., et al., A comprehensive population survey on the distribution of STR frequencies in Belarus. *Forensic Sci Int*, 2007, **172**(2–3): 156–160.

Zhivotovsky, L.A., et al., A reference database on STR allele frequencies in the Belarus population developed from paternity cases. *Forensic Sci Int Genet*, 2009, **3**(3): e107–e109.

Zhivotovsky, L.A., et al., An STR database on the Volga-Ural population. *Forensic Sci Int Genet*, 2009, **3**(4): e133–e136.

Zhivotovsky, L.A., et al., Developing STR databases on structured populations: The native South Siberian population versus the Russian population. *Forensic Sci Int Genet*, 2009, **3**(4): e111–e116.

Zhou, A., et al., Population genetic data of the NGM SElect STR loci in Chinese Han population from Zhejiang region, China. *Int J Legal Med*, 2013, **127**(2): 377–378.

Zhu, B., et al., Population genetic analysis of 15 STR loci of Chinese Tu ethnic minority group. *Forensic Sci Int*, 2008, **174**(2–3): 255–258.

Zhu, B.F., et al., Genetic diversities of 21 non-CODIS autosomal STRs of a Chinese Tibetan ethnic minority group in Lhasa. *Int J Legal Med*, 2011, **125**(4): 581–585.

Zhu, B.F., et al., Population data of 15 STR loci of Chinese Yi ethnic minority group. *Leg Med (Tokyo)*, 2008, **10**(4): 220–224.

Zhu, B.F., et al., Population genetic analysis of 15 autosomal STR loci in the Russian population of Northeastern Inner-Mongolia, China. *Mol Biol Rep*, 2010, **37**(8): 3889–3895.

Zhu, J., et al., Genetic polymorphisms of 15 STR in Chinese Salar ethnic minority group. *Forensic Sci Int*, 2007, **173**(2–3): 210–213.

Zhu, Y., et al., Genetic analysis of 15 STR loci in the population of Zhejiang Province (Southeast China). *Forensic Sci Int Genet*, 2009, **3**(4): e139–e140.

21

Sex Chromosome Haplotyping and Gender Identification

21.1 Y Chromosome Haplotyping

21.1.1 Human Y Chromosome Genome

The Y chromosome is inherited from the father and is passed on to all male offspring; this is known as *patrilineage* (Figure 21.1). Thus, the Y chromosome is unique to males. The chromosome encodes dozens of genes required for male-specific functions, including sex determination and spermatogenesis. The human Y chromosome genome contains approximately 59 million base pairs (bp) and likely contains 50–60 genes. The Y chromosome can be divided into two regions: the pseudoautosomal region and the male-specific Y region (Figure 21.2).

21.1.1.1 Pseudoautosomal Region

The *pseudoautosomal regions (PARs)* are homologous nucleotide sequences that are present on the X and Y chromosomes. There are two PARs on each X and Y chromosome: PAR1 and PAR2. In the Y chromosome, PAR1 is located on the terminal region of the short arm. The PAR1 comprises 2.6 Mb. Twenty-four genes have been identified within the PAR1. The PAR2 of the Y chromosome is located at the tip of the long arm. The PAR2 comprises 320 kb, with only four genes identified so far.

The PARs play an important role in proper segregation of the X and Y chromosomes during meiosis. During meiosis in males, the PARs allow the Y chromosome to pair with the X chromosome. Crossing over and recombination (Chapter 25) within the PARs of the X and Y chromosomes can occur. As a result, males can inherit an allele originally present on the PARs of the X chromosome, and females can inherit an allele originally present on the PARs of the Y chromosome. In particular, PAR1 plays a major role in X–Y chromosome pairing. The deletion of PAR1 results in failure of pairing and leads to male infertility. PAR2 is much shorter than PAR1, and thus shows a lower frequency of pairing than PAR1. The deletion of PAR2 exhibits less severe phenotypes than that of PAR1.

21.1.1.2 Male-Specific Y Region

The remainder of the Y chromosome is known as the *male-specific Y (MSY)* region. It was previously called the *nonrecombining Y (NRY)* region. The bulk of this region does not participate in homologous recombination. However, certain sections involve intrachromosomal gene conversion, which is the nonreciprocal transfer of genomic DNA between a pair of repeated genes on a single chromosome; in this case, the Y chromosome. About 40 megabases (Mb) within the

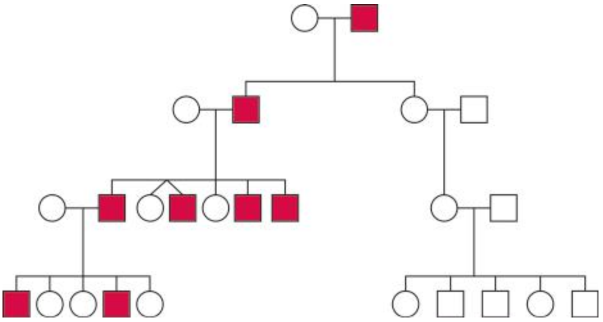

Figure 21.1 Human family pedigree showing inheritance of the Y chromosome. Females and males are denoted by circles and squares, respectively. Red symbols indicate individuals who inherited the same Y chromosome.

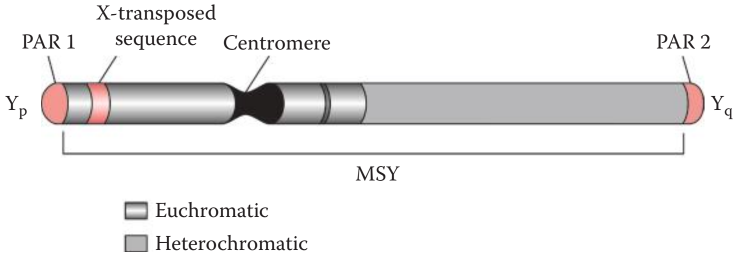

Figure 21.2 Human Y chromosome structure. PAR, the pseudoautosomal region; MSY, male-specific Y region; Y_p, the short arm of the Y chromosome; Y_q, the long arm of the Y chromosome. (© Richard C. Li.)

MSY region are heterochromatic (highly repetitive sequences) including the centromeric region and the bulk of the distal long arm. The euchromatic region, which comprises transcriptional active genes, is approximately 23 Mb. Certain sections of the euchromatic region share some homology with the X chromosome. For instance, X-transposed sequences of the Y chromosome are 99% identical to sequences within Xq21 (a band in the long arm of the X chromosome). The X-transposed sequences are the sequences, a total of approximately 3.4 Mb in length, that were transposed from the X chromosome to the Y chromosome several million years ago. Additionally, dozens of genes located in the euchromatic region share 60%–96% homology with their X chromosome counterparts. The regions sharing homology with the X chromosome should be avoided when selecting Y chromosome–specific markers for forensic DNA profiling.

21.1.1.3 Polymorphic Sequences

The Y chromosome contains an abundance of polymorphic markers. DYF155S1, also known as MSY1, is the first characterized variable number of tandem repeats (VNTR) or minisatellite at the human Y chromosome. It consists of an array of AT-rich repeats at 25 bp per unit repeat. The DYF155S1 locus is highly polymorphic. The unit repeat sequence varies through base substitution. At least five different variant types of unit repeat sequences exist. Additionally, the numbers of these units also vary, ranging from 50 to 115 repeats. Thus, the length of alleles range from approximately 1200 to 2800 bp. DYF155S1 is of considerable interest as a potential marker in forensic testing. However, the analysis method is labor intensive, making it difficult to implement in forensic casework, such as the investigation of sex crimes. Moreover, many single nucleotide polymorphisms (SNPs) and mobile elements exist at the Y chromosome. However, the discrimination power of SNPs and mobile elements is considerably less than that of Y chromosome short tandem repeats (Y-STRs). To date, Y-STRs are usually used for Y chromosome DNA testing due to the high-throughput analysis and good discrimination power (Section 21.1.2).

21.1.2 Y-STR

Y chromosome loci are very important for forensic DNA profiling, and this chapter will discuss such applications. For instance, the Y-STRs used in forensic DNA testing are male specific (for humans and certain higher primates) and are thus useful in the investigation of sexual assault cases involving male suspects. The evidence gathered in such cases usually consists of a mixture of high levels of female DNA and low levels of male DNA. The Y chromosome–specific loci can be examined without interference from large amounts of female DNA; differential extraction of sperm and nonsperm cells may not be needed. Furthermore, the Y-STR system is useful for determining the numbers of unrelated male perpetrators in sexual assault cases. The Y-STR loci used for forensic applications are located in the nonrecombining section of the Y chromosome so that patrilineage can be established. The technique can be used for paternity testing and the identification of missing persons. Finally, data interpretation can be simplified by the use of a single allele per Y-STR locus profile at most loci. Reference databases are available for estimating Y-STR haplotype frequencies among various human populations for statistical analysis of profiling results.

The major disadvantage of Y-STR loci is that their discriminating power is lowerthan that of autosomal loci. Because Y chromosome loci are linked, the product rule for statistical calculations for profile probability does not apply. Chapter 25 discusses the statistical evaluation of the strength of the matches. Also, the current Y-STR profiling cannot distinguish individuals with the same patrilineage.

More than 400 STR loci have been identified in the Y chromosome genome. The precise locations of these loci have been sequentially mapped using human genome sequencing data. Most Y-STR loci, approximately 60% of the 400 identified, are located on the long arm of the chromosome; about 22% are located on the short arm and a few are found in the centromeric region (Figure 21.3). Y-STRs in the telomeric region have yet to be identified. Only about 5% of Y-STRs are located within 5′ untranslated or intron regions of protein-coding genes. The

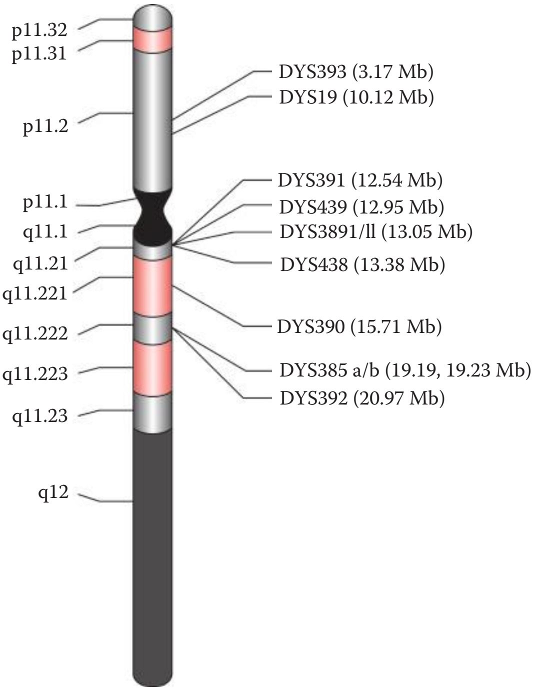

Figure 21.3 Human cytogenetic map of the Y chromosome. The Y-STRs and positions are shown (Mb = megabase). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

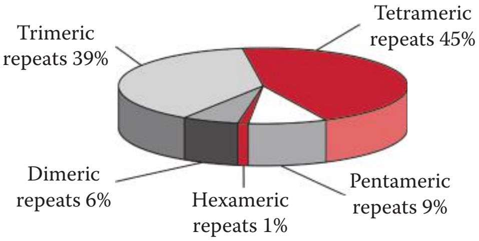

Figure 21.4 Human Y-STRs with different repeat unit length. About 400 Y-STRs have been identified and categorized according to repeat unit length. (Adapted from Hanson, E.K. and Ballantyne, J., *Leg Med (Tokyo)*, 8, 110, 2006.)

repeat unit lengths of identified Y-STRs have been analyzed. Among the 400 Y-STRs, 6% are dimeric repeats, 39% are trimeric, 45% are tetrameric, 9% are pentameric, and 1% are hexameric (Figure 21.4).

Fewer than half of the Y-STRs have been characterized. Some loci are polymorphic and are useful for forensic applications and developing new Y-STR multiplex systems. Since homologous recombination does not occur on the majority of the Y chromosome, alleles of Y-STR loci are linked (inherited together); they are referred to as *haplotypes*. As a result, the discrimination power of Y-STRs is much lower than that of autosomal STRs. The most commonly used Y-STR loci for forensic testing are described below.

21.1.2.1 Core Y-STR Loci

In 1997, the *European minimal haplotype* locus set, also known as the minimal haplotype loci, was recommended by the International Y-STR User Group for forensic applications (Table 21.1; Figure 21.5). This haplotype set includes a core set of nine Y-STR loci: DYS19, DYS385a and b, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. In 2003, two additional loci were recommended by the Scientific Working Group for DNA Analysis Methods (SWGDAM) for forensic DNA analysis. The SWGDAM loci include the European minimal haplotype loci plus two additional loci: DYS438 and DYS439 (Table 21.1).

The application Y-STR for forensic casework has been expanded with additional Y-STR markers. The use of Y-STR loci has been facilitated by commercially available PCR amplification

Table 21.1 Common Y-STR Core Loci

Locus	EMH	US Haplotype Loci	Repeat Motif
DYS19	☑	☑	TAGA
DYS385a/b	☑	☑	GAAA
DYS389 I	☑	☑	TCTA
DYS389 II	☑	☑	[TCTG] [TCTA]
DYS390	☑	☑	[TCTG] [TCTA]
DYS391	☑	☑	TCTA
DYS392	☑	☑	TAT
DYS393	☑	☑	AGAT
DYS438		☑	TTTTC
DYS439		☑	GATA

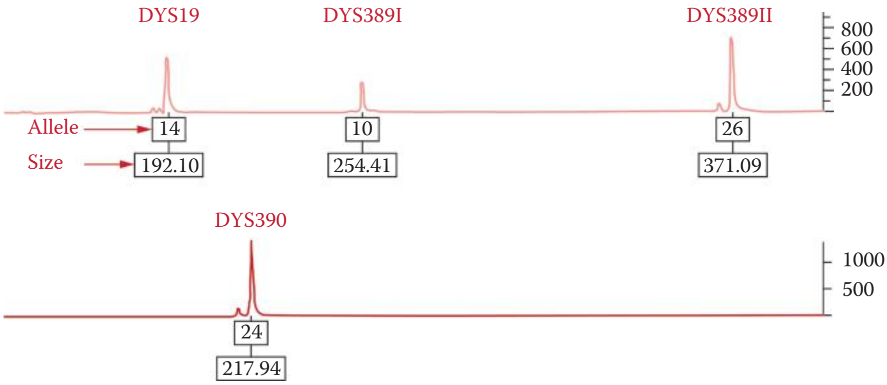

Figure 21.5 Electropherogram of Y-STR profile using an early generation of multiplex of four loci. The genotype of the DNA profile is shown (DYS19: 14. DYS389I: 10. DYS389II: 26. DYS390: 24). (© Richard C. Li.)

kits in multiplex Y-STR systems. To improve discriminating power, multiplex systems including new Y-STR loci are desired. Many new Y-STR loci are being characterized for developing new multiplex systems. Commercially available kits with more Y-STR loci are now available and have been validated for forensic use (Table 21.2; Figures 21.6 and 21.7). Some of the additional Y-STR loci are highly discriminating, allowing for further distinction between unrelated male individuals. Additionally, Y-STR loci, such as DYS570 and DYS576, with high mutation rates (see rapidly mutating Y-STR loci; Section 21.1.2.3) are included, which are useful for the discrimination of related individuals.

21.1.2.2 Multilocal Y-STR Loci

DYS385 and DYS389 are *multilocal Y-STR loci* (*MLL*). The MLL designation refers to the presence of a particular STR at more than one site on the Y chromosome due to duplication. To date, about 50 such MLL Y-STRs have been identified. Further MLL subdivisions are designated bilocal, trilocal, and so on. DYS385 and DYS389 are bilocal.

The DYS385 locus has two inverted duplicated clusters and is separated by a 4×10^4 bp interstitial region (Figure 21.8). It can be amplified by a single set of primers. One allele is observed if the duplicates are the same length. If the duplicated clusters have different lengths, they can generate two different alleles when amplified. The smaller allele is designated “a” and the larger allele is designated “b.” Moreover, the DYS389 locus has two duplicated clusters with the same orientation (Figure 21.9). In a single set of PCR primers, there are two binding sites for the same forward primer at each 5′ flanking sequence of the core repeat region of DYS389. These binding sites between DYS389I and DYS389II are about 120 bp apart. Therefore, two amplicons are produced. DYS389I is designated for the smaller allele, and DYS389II is designated for the larger allele.

21.1.2.3 Rapidly Mutating Y-STR

Current Y-STR loci implemented in forensic casework have adequate resolution of males from different patrilineages. However, most Y-STR sets have limited abilities to differentiate related males who belong to the same patrilineage. Thus, current forensic Y-STR profiling is not able to exclude patrilineal relatives of the suspect. Recently, rapidly mutating Y-STR (RM Y-STR) loci have been identified and characterized. The mutation rates of RM Y-STR are above 10^{-2}, which is considerably higher than the average mutation rates of Y-STRs. The average mutation rate for the majority of the Y-STRs characterized, including the core Y-STR loci currently used in forensic casework, is approximately 10^{-4}–10^{-3} per generation. Due to the high mutation rates, RM Y-STRs can improve patrilineage resolution. For use in forensic casework, the RM Y-STR loci have potential abilities to differentiate both paternally related and unrelated males.

Table 21.2 Y-STR Loci Covered by Representative Commercial Kits

	Loci	Yfiler[a]	PowerPlexY23[b]
Core loci	DYS19	☑	☑
	DYS385a/b	☑	☑
	DYS389 I	☑	☑
	DYS389 II	☑	☑
	DYS390	☑	☑
	DYS391	☑	☑
	DYS392	☑	☑
	DYS393	☑	☑
	DYS438	☑	☑
	DYS439	☑	☑
Additional loci	DYS437	☑	☑
	DYS448	☑	☑
	DYS456	☑	☑
	DYS458	☑	☑
	DYS481		☑
	DYS533		☑
	DYS549		☑
	DYS570		☑
	DYS576		☑
	DYS635	☑	☑
	DYS643		☑
	Y-GATA-H4	☑	☑

[a] AmpFlSTR® Yfiler® PCR Amplification Kit (Applied Biosystems).

[b] PowerPlex® Y23 System (Promega).

21.2 X Chromosome Haplotyping

X-chromosomal STR (X-STR) profiling is a useful tool in kinship testing in forensic investigations. For example, males under usual circumstances have only one X chromosome; thus males are hemizygous for the X-STR loci on the X chromosome. Homologous recombination between the X and Y chromosomes is restricted to the homologous PARs (Figure 21.2). The paternal X chromosome is inherited by daughters as haplotypes. Thus, father–daughter kinship is easier to determine using X-STRs than autosomal STRs. In females, there are two copies of the X chromosomes. Homologous recombination can occur between two X chromosomes in the mother–child transmission. Although it is possible to determine mother–child kinship using X-STR analysis, the level of certainty is less than that of father–daughter kinship analysis. Nevertheless, sharing of rare X-STR haplotypes can strengthen an indication of kinship. The most useful

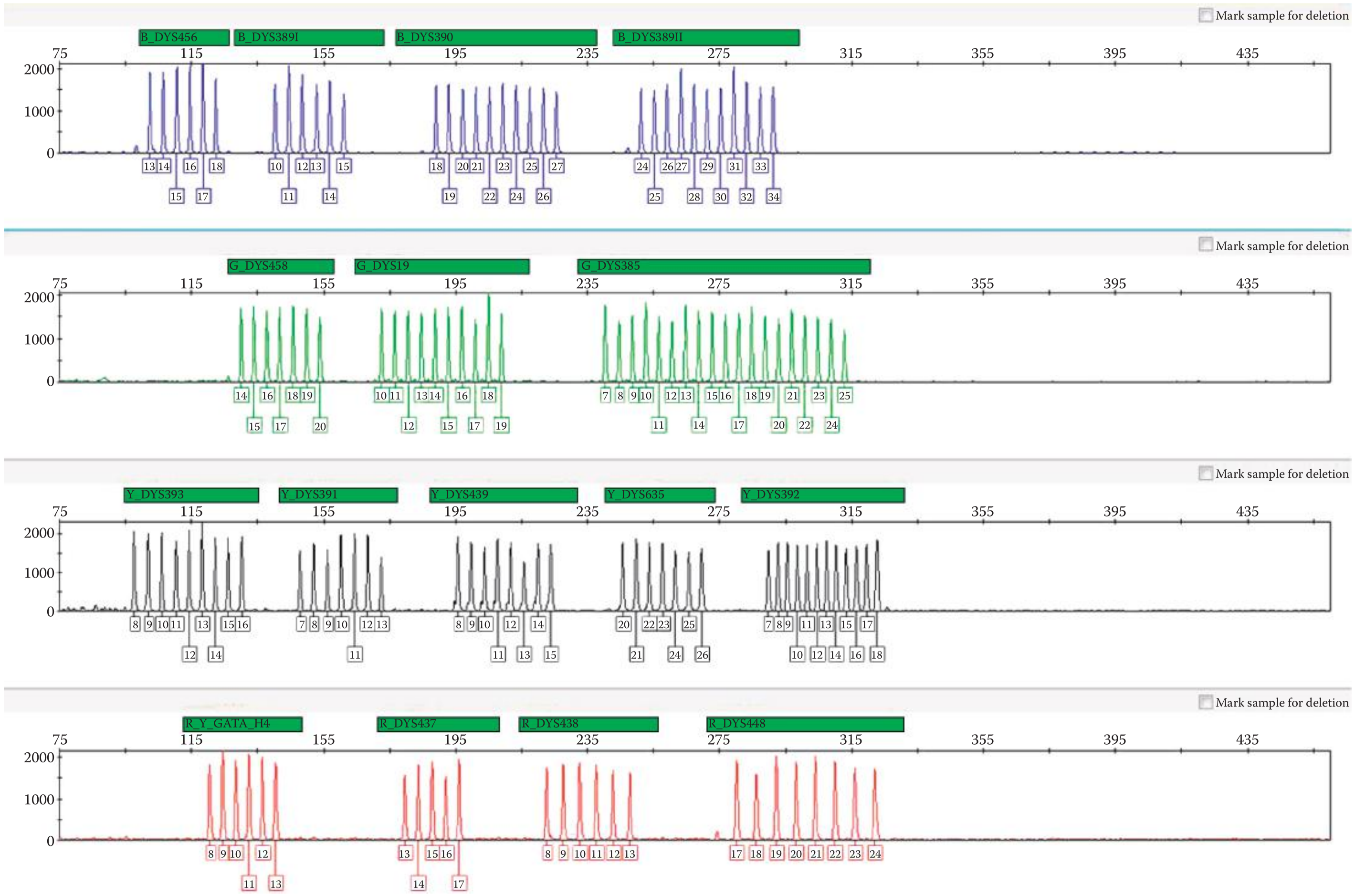

Figure 21.6 Electropherogram of the allelic ladder of AmpFISTR® Yfiler® PCR Amplification Kit (Applied Biosystems). DYS19, DYS385 a/b, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y-GATA-H4 loci are shown. (© Richard C. Li.)

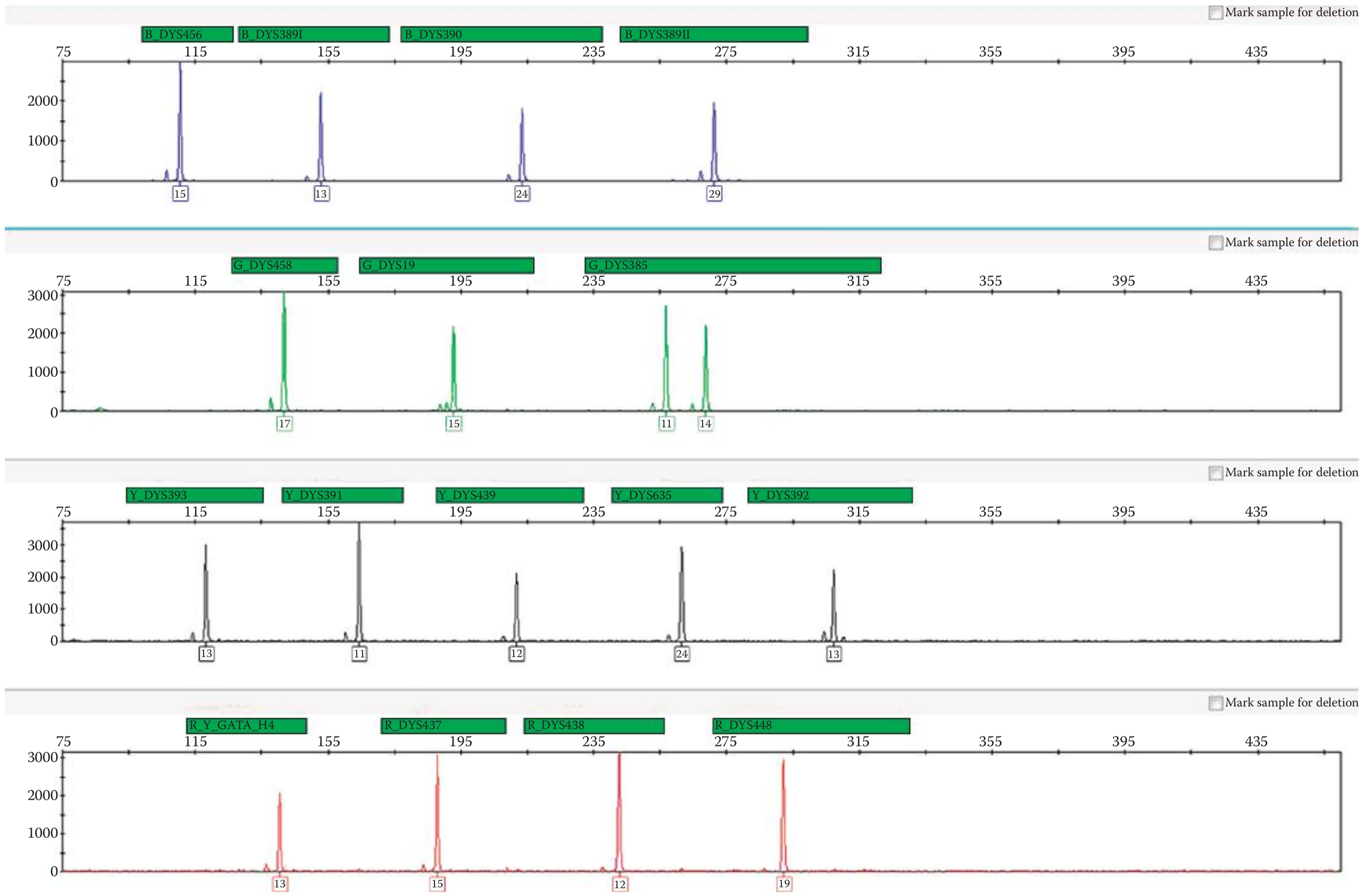

Figure 21.7 Individual Y-STR profile (AmpFlSTR® Yfiler® PCR Amplification Kit). The genotype of the DNA profile is shown. DYS19: 15. DYS385 a/b: 11, 14. DYS389 I: 13. DYS389 II: 29. DYS390: 24. DYS391: 11. DYS392: 13. DYS393: 13. DYS437: 15. DYS438: 12. DYS439: 12. DYS448: 19. DYS456: 15. DYS458: 17. DYS635: 24.Y-GATA-H4: 13. (© Richard C. Li.)

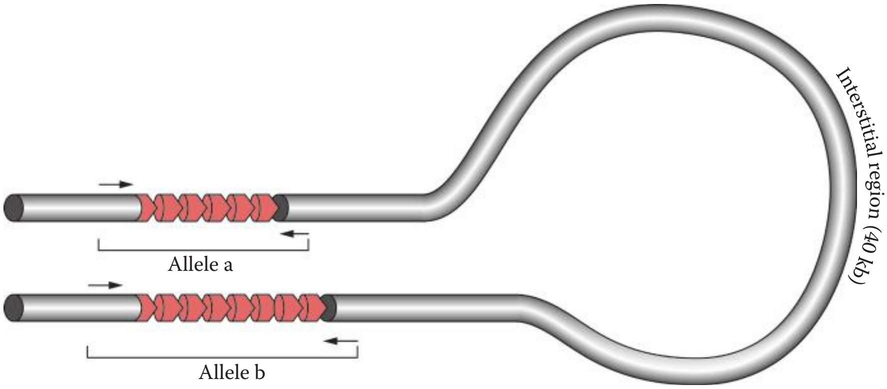

Figure 21.8 MLL Y-STR locus DYS385. At the DYS385 locus, note the two inverted duplicated regions of the Y chromosome with an interstitial region of 40 kilobases (kb). These inverted regions can be amplified with a single pair of primers (indicated by arrows) in one PCR reaction. The allele designations are described in the text. (© Richard C. Li.)

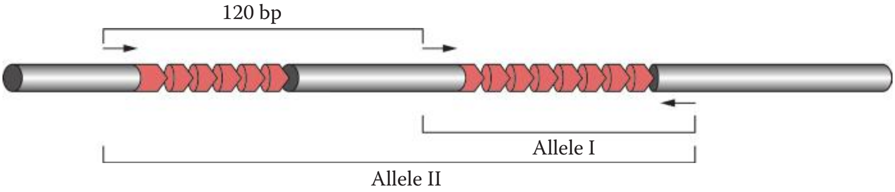

Figure 21.9 MLL Y-STR locus DYS389. At the DYS389 locus, note the two duplicated regions of the Y chromosome with the same orientation. These duplicated regions can be amplified with a single pair of primers (indicated by arrows) in one PCR reaction. The allele designations are described in the text. (© Richard C. Li.)

application of X-STR profiling, however, is in situations where autosomal STR profiling fails to determine kinship with high levels of certainty, and where Y-chromosomal or mitochondrial DNA profiling results are inconclusive.

Many X-STR loci have been identified spanning the entire human X chromosome. Increasing the number of X-STR loci to be analyzed in the inherited region would increase the degree of certainty in determining kinship. However, it is necessary to take into consideration the linkage between X-STR loci on the X chromosome. Based on Gregor Mendel's first law, the different alleles of two loci segregate independently. This is definitely true for loci localized on different chromosomes. The segregation can also be observed for loci located on the same chromosome, including the X chromosome, due to homologous recombination and meiotic crossing-over. Generally speaking, the farther apart the two loci are on the chromosome in physical distance, the more likely those loci are to segregate independently. When two loci are very close to each other, the alleles of the two loci are inherited together as a haplotype. Therefore, including additional closely linked STR loci does not necessarily improve the discrimination power in kinship analysis.

Homologous recombination events at different regions of the genome may occur with different probabilities. Thus, the tightness of linkage between two X-STR loci cannot be accurately measured based on the physical distance. In order to determine how closely two different loci are linked, the *recombination fraction* (or recombination frequency) can be used, which is the percentage of recombinants resulting from chromosomal crossover between two loci during meiosis among all the offspring. The recombination fraction reflects how often the loci are inherited together or how often they are separated. The recombination fraction ranges from 0% to 50%. The minimum recombination fraction is 0, indicating perfect linkage (no recombinants). The maximum is 50%, meaning complete independence of the two loci (they are actually located on different chromosomes or on the same chromosome with great distance between them).

Multiplex X-STR assays have also been developed. Commonly used X-STR loci can be grouped into linkage groups. Each linkage group is presumably transmitted independently from one another. These loosely linked loci can be treated as if they were located on different chromosomes. The distribution of alleles at these X-STR loci should be in accordance with the Hardy–Weinberg principle (Chapter 25). Within each linkage group, the occurrence of recombination is negligible. Thus, loci within groups are generally considered as closely linked and are thus treated as a haplotype.

21.3 Sex Typing for Gender Identification

Sex typing of a biological sample is useful in forensic investigation. For example, the application of sex typing facilitates the identification of the victim and the offender's DNA evidence in sexual assault cases and the remains of victims in mass disasters or missing persons investigations. One commonly used sex-typing marker is the amelogenin (*AMEL*) locus. The use of multiplex PCR systems with an additional amelogenin marker, a non-STR marker, leads to potential gender determination.

21.3.1 Amelogenin Locus

This region encodes extracellular matrix proteins involved in tooth enamel formation (Table 21.3). Mutations in the *AMEL* gene can lead to an enamel defect known as amelogenesis imperfecta. Amelogenesis imperfecta is a disorder that causes abnormal formation of tooth enamel in both primary and permanent teeth. The formed tooth enamel is soft and thin; therefore it is easily damaged. The *AMEL* locus has two homologous genes: *AMELX* (Xp22.2) is located on the human X chromosome (Figure 21.10) and *AMELY* (Yp11.2) is located on the human Y chromosome (Figure 21.11). Although the genes constitute a homologous pair, they differ in size and sequence. The coding sequence of the *AMELX* gene has a size of 2872 bp and the *AMELY* gene has a size of 3272 bp in length.

The most commonly used sex-typing method at the *AMEL* locus is the detection of a 6 bp deletion at intron 3 of *AMELX* (Figures 21.12 and 21.13). This deletion is not present in *AMELY*. Sex typing can be performed using primers specifically amplifying the region of the *AMEL* locus. Primer sets were developed to amplify both alleles in a single PCR. The two most commonly

Table 21.3 Sex-Typing Makers and Homologous Genes on the Human X and Y Chromosomes

Sex-Typing Markers					
X Chromosome			Y Chromosome		
Gene Symbol	Chromosomal Location	Distance from Xpter (Mb)	Gene Symbol	Chromosomal Location	Distance from Ypter (Mb)
AMELX	Xp22.2	11.3	*AMELY*	Yp11.2	6.7
DXYS156	Xq21.31	88.9	DXYS156	Yp11.2	3.2
SOX3	Xq27.1	139.6	*SRY*	Yp11.31	2.7
STS	Xp22.31	7.1–7.3	*STSP1*	Yq11.221	17.7
TSPYL2	Xp11.2	53.1	*TSPY1*	Yp11.2	9.2–9.3

Source: Ensembl Homo sapiens version 75.37 (GRCh37). Mb, megabase.

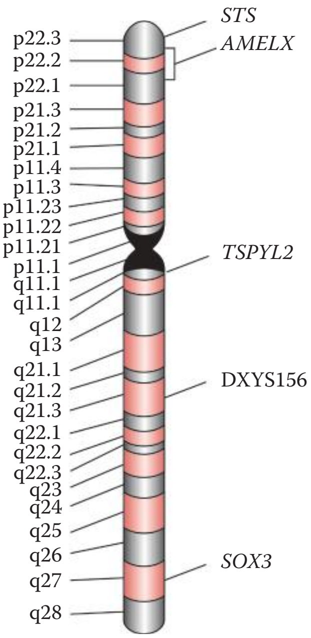
p22.3
p22.2
p22.1
p21.3
p21.2
p21.1
p11.4
p11.3
p11.23
p11.22
p11.21
p11.1
q11.1
q11.1
q12
q13
q21.1
q21.2
q21.3
q22.1
q22.2
q22.3
q23
q24
q25
q26
q27
q28
STS
AMELX
TSPYL2
DXYS156
SOX3

Figure 21.10 Cytogenetic map of the human X chromosome and sex-typing loci shown with physical positions (Mb = megabases). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

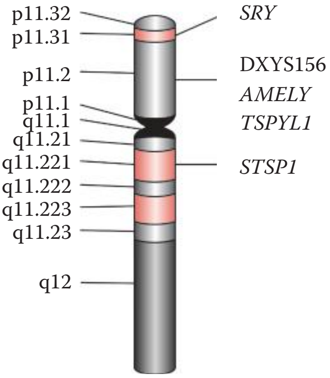
p11.32
p11.31
p11.2
p11.1
q11.1
q11.21
q11.221
q11.222
q11.223
q11.23
q12
SRY
DXYS156
AMELY
TSPYL1
STSP1

Figure 21.11 Cytogenetic map of the human Y chromosome and sex-typing loci shown with physical positions (Mb = megabases). Cytogenetic patterns with alternating dark and light bands are shown. (© Richard C. Li.)

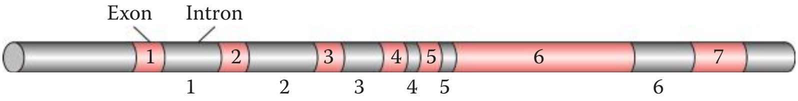
Exon
Intron
1
2
3
4
5
6
7
1
2
3
4
5
6

Figure 21.12 Structure of human *AMELY* gene. Exons 1 through 7 and introns 1 through 6 are numbered. (© Richard C. Li.)

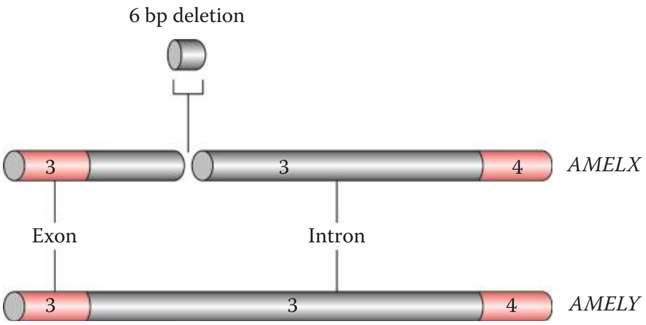

Figure 21.13 Sex typing using *AMEL* markers. A 6-bp deletion in intron 3 is present in *AMELX* but not in *AMELY* and can be resolved using electrophoresis as described in the text. (© Richard C. Li.)

used amelogenin primer sets generate amplicons of 106 and 112 bp or 212 and 218 bp for *AMELX* and *AMELY* loci, respectively. The amplicons generated from *AMELX* and *AMELY* are separated by electrophoresis. The observation of the *AMELX* fragment alone indicates a female, whereas the observation of both *AMELX* and *AMELY* fragments indicates a male (Figure 20.8). Nevertheless, DNA from primates and some rudiments can be amplified as well, but the amplicon sizes vary. The *AMEL* locus has been coamplified with other markers to provide a combined sex and identity test. Such combined tests have been used in historical D1S80 AFLP profiling (Chapter 19) and various current STR multiplex analyses.

Cases of *AMELY* null mutations have been reported where only the *AMELX* fragment is detected in *AMELY* null males. Many of them are phenotypically normal but present the *AMELX* sex types of females. Various interstitial deletions at the Y chromosome short arm and point mutations within the primer-binding sites have been identified as the possible cause of some *AMELY* null sex typing. The frequency of *AMELY* null males is rare, but is higher in Sri Lanka and India. However, the mutations may lead to incorrect identification of sex if the mutations occurred in the DNA evidence of criminal investigations such as sexual assaults or in the identification of human remains in mass disasters. In the case of a female individual, an *AMELX* mutation, which is rare, usually affects one copy of *AMELX*; however, the wild-type copy of the *AMELX* is detected. Thus, the *AMELX* mutation is less problematic in females than the *AMELY* mutation in sex typing in males.

21.3.2 Other Loci

To solve the incorrect identification problem with null *AMEL* mutations, additional candidate genetic markers are used in combination with *AMEL* analysis (Table 21.3). One useful marker is the sex-determining region Y (*SRY*) gene located on the Yp11.31 of the Y chromosome. *SRY* encodes a transcription factor that plays a role in the regulation of sex determination toward male development. The *SRY* protein contains a DNA-binding domain known as the HMG box. *SOX3*, an *SRY*-related HMG box-containing gene, has been identified at Xq27.1 of the X chromosome, which shares sequence homology with *SRY*. *SRY*-specific amplification is achieved using *SRY*-specific primers.

Another marker is the *TSPY* locus located on Yp11.2 of the Y chromosome. The *TSPY* locus encodes the testis-specific protein Y-encoded gene that is only expressed in the testis and may play a role in spermatogenesis. The *TSPY* structural gene is approximately 20 kb in length, consisting of 6 exons and 5 introns. The *TSPY* gene has multiple copies. A single Y chromosome contains several dozen copies of *TSPY* genes that are organized as a tandem array. The *TSPY* gene cluster shows copy number variation (CNV) between individuals in a population ranging from 23 to 64 copies. *DYS14* is a marker utilized for the characterization of the *TSPY* locus. The

DYS14 locus includes the partial sequence of the first exon and intron of the *TSPY* gene and shares approximately 98% sequence homology with other members of the *TSPY* gene cluster.

A *TSPY*-like (*TSPYL*) gene has been identified on the short arm of the X chromosome and is designated as *TSPYL2*. *TSPYL2* is a single-copy gene per X chromosome and is 6.3 kb in length, consisting of 7 exons and 6 introns. *TSPYL2* shares homology with *TSPY*. Additionally, several *TSPYL* genes have been identified on the autosomes. Primers are designed to amplify *TSPY* only and not amplify *TSPYL2* and other *TSPYL* gene sequences. The sensitivity of detecting *TSPY*, due to multiple copies, can be considerably higher than any single copy genetic maker such as *SRY*.

Furthermore, the DXYS156 locus, located at the pseudoautosomal region of both X and Y chromosomes, is another candidate marker used for sex typing. DXYS156 is a polymorphic pentanucleotide STR. The DXYS156 Y alleles have an additional adenine insertion in the repeat units of STR. Thus, the DXYS156 Y alleles can be distinguished from the DXYS156 X alleles because of the insertion. Additionally, the DXYS156 alleles are ethnically distributed. The DXYS156 profiling may potentially indicate the ethnic origin of an individual as an investigative lead.

Moreover, the steroid sulfatase (*STS*) gene encodes an enzyme that catalyzes the conversion of sulfated steroid precursors to biologically active steroids such as estrogens and androgens. The *STS* gene is located on the distal part of the short arm of the X chromosome within Xp22.31. The *STS* gene is 146 kb in length, consisting of 10 exons and 9 introns. The size of exons is between 120 and 160 bp long. However, the sizes of introns vary greatly, ranging from 102 bp to 35 kb. The entire coding sequence of the *STS* gene is 1542 bp in length. Deletions and point mutations of the *STS* gene have been associated with X-linked ichthyosis, a skin disease that affects males. There is an *STS* pseudogene, 100 kb in length, on the long arm of the Y chromosome known as *STSP1*. Although *STSP1* shares some sequence homologies with the *STS* gene, the pseudogene does not encode a functional gene. Primer sets are designed to amplify the target sequences within the first intron (35 kb in length) of the *STS* gene and its homologous *STSP1* pseudogene sequence. Although the homology of these target sequences of the *STS* gene and *STSP1* pseudogene is approximately 80%, amplicons with two different sizes are produced to identify *STS* and *STSP1* alleles for sex typing.

Bibliography

Y Chromosome Haplotyping

Alves, C., et al., Evaluating the informative power of Y-STRs: A comparative study using European and new African haplotype data. *Forensic Sci Int*, 2003, **134**(2–3): 126–133.

Ambers, A., et al., Autosomal and Y-STR analysis of degraded DNA from the 120-year-old skeletal remains of Ezekiel Harper. *Forensic Sci Int Genet*, 2014, **9**: 33–41.

Andersen, M.M., et al., Estimating stutter rates for Y-STR alleles. *Forensic Sci Int Genet Suppl*, 2011, **3**(1): e192–e193.

Andersen, M.M., et al., Estimating Y-STR allelic drop-out rates and adjusting for interlocus balances. *Forensic Sci Int Genet*, 2013, **7**(3): 327–336.

Asamura, H., et al., Evaluation of miniY-STR multiplex PCR systems for extended 16 Y-STR loci. *Int J Legal Med*, 2008, **122**(1): 43–49.

Ayadi, I., et al., Combining autosomal and Y-chromosomal short tandem repeat data in paternity testing with male child: Methods and application. *J Forensic Sci*, 2007, **52**(5): 1068–1072.

Ayub, Q., et al., Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res*, 2000, **28**(2): e8.

Balaresque, P., et al., Genomic complexity of the Y-STR DYS19: Inversions, deletions and founder lineages carrying duplications. *Int J Legal Med*, 2009, **123**(1): 15–23.

Ballantyne, K.N., et al., A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Sci Int Genet*, 2012, **6**(2): 208–218.

Ballantyne, K.N., et al., Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *Am J Hum Genet*, 2010, **87**(3): 341–353.

Ballard, D.J., et al., A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs. *Forensic Sci Int*, 2005, **155**(1): 65–70.

Beleza, S., et al., Extending STR markers in Y chromosome haplotypes. *Int J Legal Med*, 2003, **117**(1): 27–33.

Bosch, E., et al., High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic Sci Int*, 2002, **125**(1): 42–51.

Bouakaze, C., et al., First successful assay of Y-SNP typing by SNaPshot minisequencing on ancient DNA. *Int J Legal Med*, 2007, **121**(6): 493–499.

Brion, M., et al., Hierarchical analysis of 30 Y-chromosome SNPs in European populations. *Int J Legal Med*, 2005, **119**(1): 10–15.

Burgarella, C. and M. Navascues, Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. *Eur J Hum Genet*, 2011, **19**(1): 70–75.

Butler, J.M., et al., A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci Int*, 2002, **129**(1): 10–24.

Caglia, A., et al., Increased forensic efficiency of a STR-based Y-specific haplotype by addition of the highly polymorphic DYS385 locus. *Int J Legal Med*, 1998, **111**(3): 142–146.

Capelli, C., et al., Phylogenetic evidence for multiple independent duplication events at the DYS19 locus. *Forensic Sci Int Genet*, 2007, **1**(3–4): 287–290.

Carboni, I. and U. Ricci, Unexpected patterns in Y-STR analyses and implications for profile identification. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 55–56.

Carracedo, A., et al., Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format. *Forensic Sci Int*, 2001, **119**(1): 28–41.

Cerri, N., et al., Mixed stains from sexual assault cases: Autosomal or Y-chromosome short tandem repeats? *Croat Med J*, 2003, **44**(3): 289–292.

Corach, D., et al., Routine Y-STR typing in forensic casework. *Forensic Sci Int*, 2001, **118**(2–3): 131–135.

D'Amato, M.E., V.B. Bajic, and S. Davison, Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. *Forensic Sci Int Genet*, 2011, **5**(2): 122–125.

Davis, C., et al., Prototype PowerPlex® Y23 system: A concordance study. *Forensic Sci Int Genet*, 2013, **7**(1): 204–208.

de Knijff, P., et al., Chromosome Y microsatellites: Population genetic and evolutionary aspects. *Int J Legal Med*, 1997, **110**(3): 134–149.

Decker, A.E., et al., The impact of additional Y-STR loci on resolving common haplotypes and closely related individuals. *Forensic Sci Int Genet*, 2007, **1**(2): 215–217.

Decker, A.E., et al., Analysis of mutations in father-son pairs with 17 Y-STR loci. *Forensic Sci Int Genet*, 2008, **2**(3): e31–e35.

Dekairelle, A.F. and B. Hoste, Application of a Y-STR-pentaplex PCR (DYS19, DYS389I and II, DYS390 and DYS393) to sexual assault cases. *Forensic Sci Int*, 2001, **118**(2–3): 122–125.

Delfin, F.C., et al., Y-STR analysis for detection and objective confirmation of child sexual abuse. *Int J Legal Med*, 2005, **119**(3): 158–163.

Deng, Z.H., et al., Application of 17 Y-chromosome specific STR loci in paternity testing. *Zhongguo Shi Yan Xue Ye Xue Za Zhi*, 2008, **16**(3): 699–703.

Djelloul, S. and V. Sarafian, Validation of a 17-locus Y-STR multiplex system. *Forensic Sci Int Genet Suppl*, 2008, **1**(1): 198–199.

Dupuy, B.M., et al., Y-chromosomal microsatellite mutation rates: Differences in mutation rate between and within loci. *Hum Mutat*, 2004, **23**(2): 117–124.

Edlund, H. and M. Allen, Y chromosomal STR analysis using pyrosequencing technology. *Forensic Sci Int Genet*, 2009, **3**(2): 119–124.

Ge, J., B. Budowle, and R. Chakraborty, Interpreting Y chromosome STR haplotype mixture. *Leg Med*, 2010, **12**(3): 137–143.

Ge, J., et al., Mutation rates at Y chromosome short tandem repeats in Texas populations. *Forensic Sci Int Genet*, 2009, **3**(3): 179–184.

Ge, J., et al., US forensic Y-chromosome short tandem repeats database. *Leg Med (Tokyo)*, 2010, **12**(6): 289–295.

Geppert, M., J. Edelmann, and R. Lessig, The Y-chromosomal STRs DYS481, DYS570, DYS576 and DYS643. *Leg Med*, 2009, **11**(Suppl 1): S109–S110.

Geppert, M., et al., Hierarchical Y-SNP assay to study the hidden diversity and phylogenetic relationship of native populations in South America. *Forensic Sci Int Genet*, 2011, **5**(2): 100–104.

Gill, P., et al., DNA commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Int J Legal Med*, 2001, **114**(6): 305–309.

Goedbloed, M., et al., Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFSTR® Yfiler® PCR amplification kit. *Int J Legal Med*, 2009, **123**(6): 471–482.

Gonzalez-Neira, A., et al., Sequence structure of 12 novel Y chromosome microsatellites and PCR amplification strategies. *Forensic Sci Int*, 2001, **122**(1): 19–26.

Gross, A.M., et al., Internal validation of the AmpFlSTR Yfiler amplification kit for use in forensic casework. *J Forensic Sci*, 2008, **53**(1): 125–134.

Gusmao, L., C. Alves, and A. Amorim, Molecular characteristics of four human Y-specific microsatellites (DYS434, DYD437, DYS438, DYS439) for population and forensic studies. *Ann Hum Genet*, 2001, **65**(Pt 3): 285–291.

Gusmao, L., et al., Robustness of the Y STRs DYS19, DYS389 I and II, DYS390 and DYS393: Optimization of a PCR pentaplex. *Forensic Sci Int*, 1999, **106**(3): 163–172.

Gusmao, L., et al., Alternative primers for DYS391 typing: Advantages of their application to forensic genetics. *Forensic Sci Int*, 2000, **112**(1): 49–57.

Hall, A. and J. Ballantyne, Novel Y-STR typing strategies reveal the genetic profile of the semen donor in extended interval post-coital cervicovaginal samples. *Forensic Sci Int*, 2003, **136**(1–3): 58–72.

Hall, T.A., et al., Human Y-STR profiling using fully automated electrospray ionization time of flight mass spectrometry. *Forensic Sci Int Genet Suppl*, 2011, **3**(1): e451–e452.

Hammer, M.F., et al., Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol*, 2001, **18**(7): 1189–1203.

Hammer, M.F., et al., Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int*, 2006, **164**(1): 45–55.

Hanson, E., et al., Performance evaluation and optimization of multiplex PCRs for the highly discriminating OSU 10-locus set Y-STRs. *J Forensic Sci*, 2012, **57**(1): 52–59.

Hanson, E.K. and J. Ballantyne, Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Leg Med (Tokyo)*, 2006, **8**(2): 110–120.

Hanson, E.K. and J. Ballantyne, An ultra-high discrimination Y chromosome short tandem repeat multiplex DNA typing system. *PLoS One*, 2007, **2**(8): e688.

Hartzell, B., K. Graham, and B. McCord, Response of short tandem repeat systems to temperature and sizing methods. *Forensic Sci Int*, 2003, **133**(3): 228–234.

He, M., et al., Geographical affinities of the HapMap samples. *PLoS One*, 2009, **4**(3): e4684.

Hedman, M., et al., Dissecting the Finnish male uniformity: The value of additional Y-STR loci. *Forensic Sci Int Genet*, 2011, **5**(3): 199–201.

Henke, J., et al., Application of Y-chromosomal STR haplotypes to forensic genetics. *Croat Med J*, 2001, **42**(3): 292–297.

Heyer, E., et al., Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet*, 1997, **6**(5): 799–803.

Holtkemper, U., et al., Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques. *Hum Mol Genet*, 2001, **10**(6): 629–633.

Honda, K., The Ashikaga case of Japan—Y-STR testing used as the exculpatory evidence to free a convicted felon after 17.5 years in prison. *Forensic Sci Int Genet*, 2013, **7**(1): e1–e2.

Huang, D., et al., Y-haplotype screening of local patrilineages followed by autosomal STR typing can detect likely perpetrators in some populations. *J Forensic Sci*, 2011, **56**(5): 1340–1342.

Huang, Y.M., et al., Assessment of application value of 19 autosomal short tandem repeat loci of GoldenEye™ 20A kit in forensic paternity testing. *Int J Legal Med*, 2013, **127**(3): 587–590.

Hurles, M.E. and M.A. Jobling, A singular chromosome. *Nat Genet*, 2003, **34**(3): 246–247.

Jobling, M.A., Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Sci Int*, 2001, **118**(2–3): 158–162.

Jobling, M.A., A. Pandya, and C. Tyler-Smith, The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med*, 1997, **110**(3): 118–124.

Jobling, M.A. and C. Tyler-Smith, Fathers and sons: The Y chromosome and human evolution. *Trends Genet*, 1995, **11**(11): 449–456.

Jobling, M.A. and C. Tyler-Smith, New uses for new haplotypes the human Y chromosome, disease and selection. *Trends Genet*, 2000, **16**(8): 356–362.

Jobling, M.A. and C. Tyler-Smith, The human Y chromosome: An evolutionary marker comes of age. *Nat Rev Genet*, 2003, **4**(8): 598–612.

Jobling, M.A., et al., Y-chromosome-specific microsatellite mutation rates re-examined using a minisatellite, MSY1. *Hum Mol Genet*, 1999, **8**(11): 2117–2120.

Kayser, M. and A. Sajantila, Mutations at Y-STR loci: Implications for paternity testing and forensic analysis. *Forensic Sci Int*, 2001, **118**(2–3): 116–121.

Kayser, M., et al., Evaluation of Y-chromosomal STRs: A multicenter study. *Int J Legal Med*, 1997, **110**(3): 125–133, 141–149.

Kayser, M., et al., A comprehensive survey of human Y-chromosomal microsatellites. *Am J Hum Genet*, 2004, **74**(6): 1183–1197.

Kayser, M., et al., Characteristics and frequency of germ-line mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet*, 2000, **66**(5): 1580–1588.

Krawczak, M., Forensic evaluation of Y-STR haplotype matches: A comment. *Forensic Sci Int*, 2001, **118**(2–3): 114–115.

Krenke, B.E., et al., Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Sci Int*, 2005, **148**(1): 1–14.

Lagoa, A.M., T. Magalhães, and M.F. Pinheiro, Autosomic STR, Y-STR and miniSTR markers evaluation for genetic analysis of fingerprints. *Forensic Sci Int Genet Suppl*, 2008, **1**(1): 435–436.

Lahn, B.T., N.M. Pearson, and K. Jegalian, The human Y chromosome, in the light of evolution. *Nat Rev Genet*, 2001, **2**(3): 207–216.

Laouina, A., et al., Mutation rate at 17 Y-STR loci in "father/son" pairs from Moroccan population. *Leg Med (Tokyo)*, 2013, **15**(5): 269–271.

Lareu, M., et al., The use of the LightCycler for the detection of Y chromosome SNPs. *Forensic Sci Int*, 2001, **118**(2–3): 163–168.

Lim, S.K., et al., Variation of 52 new Y-STR loci in the Y chromosome consortium worldwide panel of 76 diverse individuals. *Int J Legal Med*, 2007, **121**(2): 124–127.

Ma, Y., et al., A Y-chromosomal haplotype with two short tandem repeat mutations. *J Forensic Sci*, 2012, **57**(6): 1630–1633.

Malsom, S., et al., The prevalence of mixed DNA profiles in fingernail samples taken from couples who co-habit using autosomal and Y-STRs. *Forensic Sci Int Genet*, 2009, **3**(2): 57–62.

Marjanović, D., et al., Identification of skeletal remains of communist armed forces victims during and after World War II: Combined Y-chromosome short tandem repeat (STR) and miniSTR approach. *Croat Med J*, 2009, **50**(3): 296–304.

Maybruck, J.L., et al., A comparative analysis of two different sets of Y-chromosome short tandem repeats (Y-STRs) on a common population panel. *Forensic Sci Int Genet*, 2009, **4**(1): 11–20.

Mayntz-Press, K.A. and J. Ballantyne, Performance characteristics of commercial Y-STR multiplex systems. *J Forensic Sci*, 2007, **52**(5): 1025–1034.

Mayntz-Press, K.A., et al., Y-STR profiling in extended interval (≥3 days) postcoital cervicovaginal samples. *J Forensic Sci*, 2008, **53**(2): 342–348.

Mulero, J.J., C.W. Chang, and L.K. Hennessy, Characterization of the N + 3 stutter product in the trinucleotide repeat locus DYS392. *J Forensic Sci*, 2006, **51**(5): 1069–1073.

Naitoh, S., et al., Assignment of Y-chromosomal SNPs found in Japanese population to Y-chromosomal haplogroup tree. *J Hum Genet*, 2013, **58**(4): 195–201.

Nebel, A., et al., Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392. *Eur J Hum Genet*, 2001, **9**(1): 22–26.

Niederstatter, H., et al., Separate analysis of DYS385a and b versus conventional DYS385 typing: Is there forensic relevance? *Int J Legal Med*, 2005, **119**(1): 1–9.

Olofsson, J., et al., Evaluation of Y-STR analyses of sperm cell negative vaginal samples. *Forensic Sci Int Genet Suppl*, 2011, **3**(1): e141–e142.

Onofri, V., L. Buscemi, and A. Tagliabracci, Evaluating Y-chromosome STRs mutation rates: A collaborative study of the Ge.F.I.-ISFG Italian group. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 419–420.

Oz, C., et al., A Y-chromosome STR marker should be added to commercial multiplex STR kits. *J Forensic Sci*, 2008, **53**(4): 858–861.

Paracchini, S., et al., Hierarchical high-throughput SNP genotyping of the human Y chromosome using MALDI-TOF mass spectrometry. *Nucleic Acids Res*, 2002, **30**(6): e27.

Park, M.J., et al., Characterization of deletions in the DYS385 flanking region and null alleles associated with AZFc microdeletions in Koreans. *J Forensic Sci*, 2008, **53**(2): 331–334.

Park, M.J., et al., Y-SNP miniplexes for East Asian Y-chromosomal haplogroup determination in degraded DNA. *Forensic Sci Int Genet*, 2013, **7**(1): 75–81.

Park, M.J., et al., Y-STR analysis of degraded DNA using reduced-size amplicons. *Int J Legal Med*, 2007, **121**(2): 152–157.

Parson, W., et al., When autosomal short tandem repeats fail: Optimized primer and reaction design for Y-chromosome short tandem repeat analysis in forensic casework. *Croat Med J*, 2001, **42**(3): 285–287.

Parson, W., et al., Improved specificity of Y-STR typing in DNA mixture samples. *Int J Legal Med*, 2003, **117**(2): 109–114.

Parson, W., et al., Y-STR analysis on DNA mixture samples—Results of a collaborative project of the ENFSI DNA working group. *Forensic Sci Int Genet*, 2008, **2**(3): 238–242.

Pascali, V.L., M. Dobosz, and B. Brinkmann, Coordinating Y-chromosomal STR research for the courts. *Int J Legal Med*, 1999, **112**(1): 1.

Prinz, M. and M. Sansone, Y chromosome-specific short tandem repeats in forensic casework. *Croat Med J*, 2001, **42**(3): 288–291.

Prinz, M., et al., Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Sci Int*, 1997, **85**(3): 209–218.

Prinz, M., et al., Validation and casework application of a Y chromosome specific STR multiplex. *Forensic Sci Int*, 2001, **120**(3): 177–188.

Raina, A., et al., Misinterpretation of results in medico-legal cases due to microdeletion in the Y-chromosome. *Mol Cell Probes*, 2010, **24**(6): 418–420.

Ravid-Amir, O. and S. Rosset, Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. *Bioinformatics*, 2010, **26**(18): i440–i445.

Redd, A.J., S.L. Clifford, and M. Stoneking, Multiplex DNA typing of short-tandem-repeat loci on the Y chromosome. *Biol Chem*, 1997, **378**(8): 923–927.

Redd, A.J., et al., Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int*, 2002, **130**(2–3): 97–111.

Rodig, H., et al., Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. *Forensic Sci Int*, 2008, **174**(2–3): 182–188.

Roewer, L., Y chromosome STR typing in crime casework. *Forensic Sci Med Pathol*, 2009, **5**(2): 77–84.

Roewer, L. and M. Geppert, Interpretation guidelines of a standard Y-chromosome STR 17-plex PCR-CE assay for crime casework. *Methods Mol Biol*, 2012, **830**: 43–56.

Roewer, L., et al., A new method for the evaluation of matches in non-recombining genomes: Application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci Int*, 2000, **114**(1): 31–43.

Roewer, L., et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int*, 2001, **118**(2–3): 106–113.

Rolf, B., et al., Paternity testing using Y-STR haplotypes: Assigning a probability for paternity in cases of mutations. *Int J Legal Med*, 2001, **115**(1): 12–15.

Rozen, S., et al., Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, 2003, **423**(6942): 873–876.

Sanchez, J.J., et al., Multiplex PCR and minisequencing of SNPs—A model with 35 Y chromosome SNPs. *Forensic Sci Int*, 2003, **137**(1): 74–84.

Sanchez-Diz, P., et al., Results of the GEP-ISFG collaborative study on two Y-STRs tetraplexes: GEPY I (DYS461, GATA C4, DYS437 and DYS438) and GEPY II (DYS460, GATA A10, GATA H4 and DYS439). *Forensic Sci Int*, 2003, **135**(2): 158–162.

Santos, F.R., N.O. Bianchi, and S.D. Pena, Worldwide distribution of human Y-chromosome haplotypes. *Genome Res*, 1996, **6**(7): 601–611.

Santos, F.R., S.D. Pena, and J.T. Epplen, Genetic and population study of a Y-linked tetranucleotide repeat DNA polymorphism with a simple non-isotopic technique. *Hum Genet*, 1993, **90**(6): 655–656.

Schneider, P.M., et al., Results of collaborative study regarding the standardization of the Y-linked STR system DYS385 by the European DNA Profiling (EDNAP) group. *Forensic Sci Int*, 1999, **102**(2–3): 159–165.

Schoske, R., et al., High-throughput Y-STR typing of U.S. populations with 27 regions of the Y chromosome using two multiplex PCR assays. *Forensic Sci Int*, 2004, **139**(2–3): 107–121.

Schoske, R., et al., Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci. *Anal Bioanal Chem*, 2003, **375**(3): 333–343.

Seo, Y., et al., A method for genotyping Y chromosome-linked DYS385a and DYS385b loci. *Leg Med (Tokyo)*, 2003, **5**(4): 228–232.

Sibille, I., et al., Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci Int*, 2002, **125**(2–3): 212–216.

Sinha, S.K., et al., Development and validation of a multiplexed Y-chromosome STR genotyping system, Y-PLEX 6, for forensic casework. *J Forensic Sci*, 2003, **48**(1): 93–103.

Sinha, S.K., et al., Development and validation of the Y-PLEX 5, a Y-chromosome STR genotyping system, for forensic casework. *J Forensic Sci*, 2003, **48**(5): 985–1000.

Skaletsky, H., et al., The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 2003, **423**(6942): 825–837.

Sturk, K.A., et al., Evaluation of modified Yfiler™ amplification strategy for compromised samples. *Croat Med J*, 2009, **50**(3): 228–238.

Szibor, R., M. Kayser, and L. Roewer, Identification of the human Y-chromosomal microsatellite locus DYS19 from degraded DNA. *Am J Forensic Med Pathol*, 2000, **21**(3): 252–254.

Thompson, J.M., et al., Developmental validation of the PowerPlex® Y23 System: A single multiplex Y-STR analysis system for casework and database samples. *Forensic Sci Int Genet*, 2013, **7**(2): 240–250.

Tilford, C.A., et al., A physical map of the human Y chromosome. *Nature*, 2001, **409**(6822): 943–945.

Tsuji, A., et al., Personal identification using Y-chromosomal short tandem repeats from bodily fluids mixed with semen. *Am J Forensic Med Pathol*, 2001, **22**(3): 288–291.

Tsukada, K., et al., DNA typing using AmpFlSTR Y-filer for very long-term stored specimens. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S466–S467.

Tsutsumi, H., et al., A case of personal identification due to detection of rare DNA types from seminal stain. *J Oral Sci*, 2009, **51**(4): 645–650.

Underhill, P.A., et al., Y chromosome sequence variation and the history of human populations. *Nat Genet*, 2000, **26**(3): 358–361.

Underhill, P.A., et al., The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet*, 2001, **65**(Pt 1): 43–62.

Vallone, P.M. and J.M. Butler, Y-SNP typing of U.S. African American and Caucasian samples using allele-specific hybridization and primer extension. *J Forensic Sci*, 2004, **49**(4): 723–732.

Valverde, L., et al., Improving the analysis of Y-SNP haplogroups by a single highly informative 16 SNP multiplex PCR-minisequencing assay. *Electrophoresis*, 2013, **34**(4): 605–612.

van Oven, M., A. Ralf, and M. Kayser, An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups. *Int J Legal Med*, 2011, **125**(6): 879–885.

Vermeulen, M., et al., Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Sci Int Genet*, 2009, **3**(4): 205–213.

Wei, W., et al., Exploring of new Y-chromosome SNP loci using Pyrosequencing and the SNaPshot methods. *Int J Legal Med*, 2012, **126**(6): 825–833.

White, P.S., et al., New, male-specific microsatellite markers from the human Y chromosome. *Genomics*, 1999, **57**(3): 433–437.

Willuweit, S., L. Roewer, and International Forensic Y Chromosome User Group, Y chromosome haplotype reference database (YHRD): Update. *Forensic Sci Int Genet*, 2007, **1**(2): 83–87.

Zhivotovsky, L.A., et al., The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet*, 2004, **74**(1): 50–61.

Zuccarelli, G., et al., Rapid screening for Native American mitochondrial and Y-chromosome haplogroups detection in routine DNA analysis. *Forensic Sci Int Genet*, 2011, **5**(2): 105–108.

Genetic Population Data of Y Chromosome Haplotyping

Aboukhalid, R., et al., Haplotype frequencies for 17 Y-STR loci (AmpFlSTRY-filer) in a Moroccan population sample. *Forensic Sci Int Genet*, 2010, **4**(3): e73–e74.

Achakzai, N.M., et al., Y-chromosomal STR analysis in the Pashtun population of Southern Afghanistan. *Forensic Sci Int Genet*, 2012, **6**(4): e103–e105.

Alakoc, Y.D., et al., Y-chromosome and autosomal STR diversity in four proximate settlements in Central Anatolia. *Forensic Sci Int Genet*, 2010, **4**(5): e135–e137.

Alam, S., et al., Haplotype diversity of 17 Y-chromosomal STR loci in the Bangladeshi population. *Forensic Sci Int Genet*, 2010, **4**(2): e59–e60.

Ali, N., et al., Announcement of population data: Genetic data for 17 Y-STR AmpFlSTR® Yfiler™ markers from an immigrant Pakistani population in the UK (British Pakistanis). *Forensic Sci Int Genet*, 2013, **7**(2): e40–e42.

Alshamali, F., et al., Local population structure in Arabian Peninsula revealed by Y-STR diversity. *Hum Hered*, 2009, **68**(1): 45–54.

Ambrosio, B., et al., Y-STR genetic diversity in autochthonous Andalusians from Huelva and Granada provinces (Spain). *Forensic Sci Int Genet*, 2012, **6**(2): e66–e71.

Andersen, M.M., P.S. Eriksen, and N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *J Theor Biol*, 2013, **329**: 39–51.

Andersen, M.M., et al., Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Sci Int Genet*, 2013, **7**(2): 264–271.

Baeza, C., et al., Population data for 15 Y-chromosome STRs in a population sample from Quito (Ecuador). *Forensic Sci Int*, 2007, **173**(2–3): 214–219.

Bai, R., et al., Haplotype diversity of 17 Y-STR loci in a Chinese Han population sample from Shanxi Province, Northern China. *Forensic Sci Int Genet*, 2013, **7**(1): 214–216.

Balamurugan, K., et al., Y chromosome STR allelic and haplotype diversity in five ethnic Tamil populations from Tamil Nadu, India. *Leg Med (Tokyo)*, 2010, **12**(5): 265–269.

Bembea, M., et al., Y-chromosome STR haplotype diversity in three ethnically isolated population from North-Western Romania. *Forensic Sci Int Genet*, 2011, **5**(3): e99–e100.

Bento, A.M., et al., Distribution of Y-chromosomal haplotypes in the Central Portuguese population using 17-STRs. *Forensic Sci Int Genet*, 2009, **4**(1): e35–e36.

Bing, L., et al., Population genetics for 17 Y-STR loci(AmpFISTR® Y-filer™) in Luzhou Han ethnic group. *Forensic Sci Int Genet*, 2013, **7**(2): e23–e26.

Borjas, L., et al., Usefulness of 12 Y-STRs for forensic genetics evaluation in two populations from Venezuela. *Leg Med (Tokyo)*, 2008, **10**(2): 107–112.

Brisighelli, F., et al., Patterns of Y-STR variation in Italy. *Forensic Sci Int Genet*, 2012, **6**(6): 834–839.

Budowle, B., et al., Texas population substructure and its impact on estimating the rarity of Y STR haplotypes from DNA evidence. *J Forensic Sci*, 2009, **54**(5): 1016–1021.

Budowle, B., et al., The effects of Asian population substructure on Y STR forensic analyses. *Leg Med (Tokyo)*, 2009, **11**(2): 64–69.

Builes, J.J., et al., Y chromosome STR haplotypes in the Caribbean city of Cartagena (Colombia). *Forensic Sci Int*, 2007, **167**(1): 62–69.

Caine, L.M., M.M. de Pancorbo, and F. Pinheiro, Y-chromosomal STR haplotype diversity in males from Santa Catarina, Brazil. *J Forensic Leg Med*, 2010, **17**(2): 92–95.

Carvalho, M., et al., Paternal and maternal lineages in Guinea-Bissau population. *Forensic Sci Int Genet*, 2011, **5**(2): 114–116.

Chang, Y.M., et al., Haplotype diversity of 16 Y-chromosomal STRs in three main ethnic populations (Malays, Chinese and Indians) in Malaysia. *Forensic Sci Int*, 2007, **167**(1): 70–76.

Cloete, K., et al., Analysis of seventeen Y-chromosome STR loci in the Cape Muslim population of South Africa. *Leg Med (Tokyo)*, 2010, **12**(1): 42–45.

Coble, M.D., C.R. Hill, and J.M. Butler, Haplotype data for 23 Y-chromosome markers in four U.S. population groups. *Forensic Sci Int Genet*, 2013, **7**(3): e66–e68.

Cortellini, V., et al., Y-chromosome polymorphisms and ethnic group—A combined STR and SNP approach in a population sample from northern Italy. *Croat Med J*, 2013, **54**(3): 279–285.

D'Amato, M.E., M. Benjeddou, and S. Davison, Evaluation of 21 Y-STRs for population and forensic studies. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 446–447.

Davis, C., et al., Y-STR loci diversity in native Alaskan populations. *Int J Legal Med*, 2011, **125**(4): 559–563.

Diaz-Lacava, A., et al., Geostatistical inference of main Y-STR-haplotype groups in Europe. *Forensic Sci Int Genet*, 2011, **5**(2): 91–94.

Diaz-Lacava, A., et al., Spatial assessment of Argentinean genetic admixture with geographical information systems. *Forensic Sci Int Genet*, 2011, **5**(4): 297–302.

Ehler, E., R. Marvan, and D. Vanek, Evaluation of 14 Y-chromosomal short tandem repeat haplotype with focus on DYS449, DYS456, and DYS458: Czech population sample. *Croat Med J*, 2010, **51**(1): 54–60.

Elmrghni, S., et al., Population genetic data for 17 Y STR markers from Benghazi (East Libya). *Forensic Sci Int Genet*, 2012, **6**(2): 224–227.

Ferri, G., et al., Y-STR variation in Albanian populations: Implications on the match probabilities and the genetic legacy of the minority claiming an Egyptian descent. *Int J Legal Med*, 2010, **124**(5): 363–370.

Frank, W.E., H.C. Ralph, and M.A. Tahir, Y chromosome STR haplotypes and allele frequencies in a southern Indian male population. *J Forensic Sci*, 2008, **53**(1): 248–251.

Gayden, T., et al., Y-chromosomal microsatellite diversity in three culturally defined regions of historical Tibet. *Forensic Sci Int Genet*, 2012, **6**(4): 437–446.

Gois, C.C., et al., Genetic population data of 12 STR loci of the PowerPlex Y system in the state of Sao Paulo population (Southeast of Brazil). *Forensic Sci Int*, 2008, **174**(1): 81–86.

Gonzalez-Andrade, F., et al., Y-chromosome STR haplotypes in three different population groups from Ecuador (South America). *J Forensic Sci*, 2008, **53**(2): 512–514.

Gonzalez-Andrade, F., et al., Y-STR variation among ethnic groups from Ecuador: Mestizos, Kichwas, Afro-Ecuadorians and Waoranis. *Forensic Sci Int Genet*, 2009, **3**(3): e83–e91.

Grskovic, B., et al., Population genetic analysis of haplotypes based on 17 short tandem repeat loci on Y chromosome in population sample from eastern Croatia. *Croat Med J*, 2010, **51**(3): 202–208.

Grskovic, B., et al., Genetic polymorphisms of 17 short tandem repeat loci on Y chromosome in central Croatian population. *Forensic Sci Med Pathol*, 2011, **7**(2): 155–161.

Grskovic, B., et al., Population data for 17 short tandem repeat loci on Y chromosome in northern Croatia. *Mol Biol Rep*, 2011, **38**(3): 2203–2209.

Guo, H., et al., Genetic polymorphisms for 17 Y-chromosomal STRs haplotypes in Chinese Hui population. *Leg Med (Tokyo)*, 2008, **10**(3): 163–169.

Haliti, N., et al., Evaluation of population variation at 17 autosomal STR and 16 Y-STR haplotype loci in Croatians. *Forensic Sci Int Genet*, 2009, **3**(4): e137–e138.

Hallenberg, C., et al., Y-chromosome STR haplotypes in males from Greenland. *Forensic Sci Int Genet*, 2009, **3**(4): e145–e146.

Hara, M., et al., Genetic data for 16 Y-chromosomal STR loci in Japanese. *Leg Med (Tokyo)*, 2007, **9**(3): 161–170.

Hashiyada, M., et al., Population genetics of 17 Y-chromosomal STR loci in Japanese. *Forensic Sci Int Genet*, 2008, **2**(4): e69–e70.

He, J. and F. Guo, Population genetics of 17 Y-STR loci in Chinese Manchu population from Liaoning Province, northeast China. *Forensic Sci Int Genet*, 2013, **7**(3): e84–e85.

Hedman, M., et al., Dissecting the Finnish male uniformity: The value of additional Y-STR loci. *Forensic Sci Int Genet*, 2011, **5**(3): 199–201.

Huang, T.Y., et al., Polymorphism of 17 Y-STR loci in Taiwan population. *Forensic Sci Int*, 2008, **174**(2–3): 249–254.

Hwa, H.L., et al., Seventeen Y-chromosomal short tandem repeat haplotypes in seven groups of population living in Taiwan. *Int J Legal Med*, 2010, **124**(4): 295–300.

Hwang, J.H., et al., Haplotypes for 12 Y-chromosomal STR loci in a Korean population (the central region). *Forensic Sci Int*, 2007, **168**(1): 73–84.

Illeperuma, R.J., et al., Haplotype data for 12 Y-chromosome STR loci of Sri Lankans. *Forensic Sci Int Genet*, 2010, **4**(4): e119–e120.

Jakovski, Z., et al., Genetic data for 17 Y-chromosomal STR loci in Macedonians in the Republic of Macedonia. *Forensic Sci Int Genet*, 2011, **5**(4): e108–e111.

Julieta Avila, S., I. Briceno, and A. Gomez, Genetic population analysis of 17 Y-chromosomal STRs in three states (Valle del Cauca, Cauca and Narino) from Southwestern Colombia. *J Forensic Leg Med*, 2009, **16**(4): 204–211.

Kang, L.L., K. Liu, and Y. Ma, Y chromosome STR haplotypes of Tibetan Living Tibet Lassa. *Forensic Sci Int*, 2007, **172**(1): 79–83.

Katsaloulis, P., et al., Genetic population study of 11 Y chromosome STR loci in Greece. *Forensic Sci Int Genet*, 2013, **7**(3): e56–e58.

Kayser, M., et al., Relating two deep-rooted pedigrees from central Germany by high-resolution Y-STR haplotyping. *Forensic Sci Int Genet*, 2007, **1**(2): 125–128.

Kim, S.H., et al., Genetic polymorphisms of 16 Y chromosomal STR loci in Korean population. *Forensic Sci Int Genet*, 2008, **2**(2): e9–e10.

Kim, S.H., et al., Population genetics and mutational events at 6 Y-STRs in Korean population. *Forensic Sci Int Genet*, 2009, **3**(2): e53–e54.

Kim, S.H., et al., Y chromosome homogeneity in the Korean population. *Int J Legal Med*, 2010, **124**(6): 653–657.

Kim, S.H., et al., Forensic genetic data of 6 Y-STR loci: An expanded Korean population database. *Forensic Sci Int Genet*, 2012, **6**(1): e35–e36.

Kovacevic, L., et al., Haplotype data for 23 Y-chromosome markers in a reference sample from Bosnia and Herzegovina. *Croat Med J*, 2013, **54**(3): 286–290.

Kovatsi, L., J.L. Saunier, and J.A. Irwin, Population genetics of Y-chromosome STRs in a population of Northern Greeks. *Forensic Sci Int Genet*, 2009, **4**(1): e21–e22.

Kumagai, R., et al., Haplotype analysis of 17 Y-STR loci in a Japanese population. *Forensic Sci Int*, 2007, **172**(1): 72–78.

Lacau, H., et al., Y-STR profiling in two Afghanistan populations. *Leg Med (Tokyo)*, 2011, **13**(2): 103–108.

Laouina, A., et al., Allele frequencies and population data for 17 Y-STR loci (The AmpFlSTR® Y-filer™) in Casablanca resident population. *Forensic Sci Int Genet*, 2011, **5**(1): e1–e3.

Larmuseau, M.H., et al., Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. *Forensic Sci Int Genet*, 2011, **5**(2): 95–99.

Leat, N., et al., Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic Sci Int*, 2007, **168**(2–3): 154–161.

Lecerf, M., et al., Allele frequencies and haplotypes of eight Y-short tandem repeats in Bantu population living in Central Africa. *Forensic Sci Int*, 2007, **171**(2–3): 212–215.

Lee, H.Y., et al., Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs. *Int J Legal Med*, 2007, **121**(2): 128–135.

Lee, J., et al., Y chromosomal STRs haplotypes in two populations from Bolivia. *Leg Med (Tokyo)*, 2007, **9**(1): 43–47.

Lim, S.K., et al., Variation of 52 new Y-STR loci in the Y chromosome consortium worldwide panel of 76 diverse individuals. *Int J Legal Med*, 2007, **121**(2): 124–127.

Ljubkovic, J., et al., Y-chromosomal short tandem repeat haplotypes in southern Croatian male population defined by 17 loci. *Croat Med J*, 2008, **49**(2): 201–206.

Lowery, R.K., et al., Sub-population structure evident in forensic Y-STR profiles from Armenian geographical groups. *Leg Med (Tokyo)*, 2013, **15**(2): 85–90.

Maybruck, J.L., et al., A comparative analysis of two different sets of Y-chromosome short tandem repeats (Y-STRs) on a common population panel. *Forensic Sci Int Genet*, 2009, **4**(1): 11–20.

Melo, M.M., et al., Y-STR haplotypes in three ethnic linguistic groups of Angola population. *Forensic Sci Int Genet*, 2011, **5**(3): e83–e88.

Mendes-Junior, C.T., et al., Y-chromosome STR haplotypes in a sample from Sao Paulo State, southeastern Brazil. *J Forensic Sci*, 2007, **52**(2): 495–497.

Mertens, G., et al., Twelve-locus Y-STR haplotypes in the Flemish population. *J Forensic Sci*, 2007, **52**(3): 755–757.

Mielnik-Sikorska, M., et al., Genetic data from Y chromosome STR and SNP loci in Ukrainian population. *Forensic Sci Int Genet*, 2013, **7**(1): 200–203.

Mizuno, N., et al., 16 Y chromosomal STR haplotypes in Japanese. *Forensic Sci Int*, 2008, **174**(1): 71–76.

Mizuno, N., et al., A forensic method for the simultaneous analysis of biallelic markers identifying Y chromosome haplogroups inferred as having originated in Asia and the Japanese archipelago. *Forensic Sci Int Genet*, 2010, **4**(2): 73–79.

Mršić, G., et al., Croatian national reference Y-STR haplotype database. *Mol Biol Rep*, 2012, **39**(7): 7727–7741.

Muro, T., et al., Simultaneous determination of seven informative Y chromosome SNPs to differentiate East Asian, European, and African populations. *Leg Med (Tokyo)*, 2011, **13**(3): 134–141.

Nagy, M., et al., Searching for the origin of Romanies: Slovakian Romani, Jats of Haryana and Jat Sikhs Y-STR data in comparison with different Romani populations. *Forensic Sci Int*, 2007, **169**(1): 19–26.

Nonaka, I., K. Minaguchi, and N. Takezaki, Y-chromosomal binary haplogroups in the Japanese population and their relationship to 16 Y-STR polymorphisms. *Ann Hum Genet*, 2007, **71**(Pt 4): 480–495.

Nunez, C., et al., Y chromosome haplogroup diversity in a Mestizo population of Nicaragua. *Forensic Sci Int Genet*, 2012, **6**(6): e192–e195.

Onofri, V., et al., Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. *Int J Legal Med*, 2007, **121**(3): 234–237.

Padilla-Gutierrez, J.R., et al., Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, Mexico. *Leg Med (Tokyo)*, 2008, **10**(6): 319–320.

Palet, L., et al., Y-STR genetic diversity in Moroccans from the Figuig oasis. *Forensic Sci Int Genet*, 2010, **4**(5): e139–e141.

Palha Tde, J., E.M. Rodrigues, and S.E. Dos Santos, Y-chromosomal STR haplotypes in a population from the Amazon region, Brazil. *Forensic Sci Int*, 2007, **166**(2–3): 233–239.

Palha, T., et al., Fourteen short tandem repeat loci Y chromosome haplotypes: Genetic analysis in populations from northern Brazil. *Forensic Sci Int Genet*, 2012, **6**(3): 413–418.

Palha, T.J., E.M. Rodrigues, and S.E. dos Santos, Y-STR haplotypes of Native American populations from the Brazilian Amazon region. *Forensic Sci Int Genet*, 2010, **4**(5): e121–e123.

Palo, J.U., et al., High degree of Y-chromosomal divergence within Finland—Forensic aspects. *Forensic Sci Int Genet*, 2007, **1**(2): 120–124.

Palo, J.U., et al., The effect of number of loci on geographical structuring and forensic applicability of Y-STR data in Finland. *Int J Legal Med*, 2008, **122**(6): 449–456.

Park, S.W., et al., Development of a Y-STR 12-plex PCR system and haplotype analysis in a Korean population. *J Genet*, 2009, **88**(3): 353–358.

Parkin, E.J., et al., Diversity of 26-locus Y-STR haplotypes in a Nepalese population sample: Isolation and drift in the Himalayas. *Forensic Sci Int*, 2007, **166**(2–3): 176–181.

Parvathy, S.N., A. Geetha, and C. Jagannath, Haplotype analysis of the polymorphic 17 YSTR markers in Kerala nontribal populations. *Mol Biol Rep*, 2012, **39**(6): 7049–7059.

Petrejcikova, E., et al., Allele frequencies and population data for 11 Y-chromosome STRs in samples from Eastern Slovakia. *Forensic Sci Int Genet*, 2011, **5**(3): e53–e62.

Piglionica, M., et al., Population data for 17 Y-chromosome STRs in a sample from Apulia (Southern Italy). *Forensic Sci Int Genet*, 2013, **7**(1): e3–e4.

Ploski, R., et al., Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis. *Hum Genet*, 2002, **110**(6): 592–600.

Pokupcic, K., et al., Y-STR genetic diversity of Croatian (Bayash) Roma. *Forensic Sci Int Genet*, 2008, **2**(2): e11–e13.

Pontes, M.L., et al., Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR® Y-filer™) in a Northern Portuguese population sample. *Forensic Sci Int*, 2007, **170**(1): 62–67.

Rafiee, M.R., et al., Analysis of Y-chromosomal short tandem repeat (STR) polymorphism in an Iranian Sadat population. *Genetika*, 2009, **45**(8): 1105–1109.

Rapone, C., et al., Y chromosome haplotypes in central-south Italy: Implication for reference database. *Forensic Sci Int*, 2007, **172**(1): 67–71.

Rebala, K., et al., Forensic analysis of polymorphism and regional stratification of Y-chromosomal microsatellites in Belarus. *Forensic Sci Int Genet*, 2011, **5**(1): e17–e20.

Rebala, K., et al., Y-STR variation among Slavs: Evidence for the Slavic homeland in the middle Dnieper basin. *J Hum Genet*, 2007, **52**(5): 406–414.

Roewer, L., et al., Y-chromosomal STR haplotypes in Kalmyk population samples. *Forensic Sci Int*, 2007, **173**(2–3): 204–209.

Roewer, L., et al., Analysis of Y chromosome STR haplotypes in the European part of Russia reveals high diversities but non-significant genetic distances between populations. *Int J Legal Med*, 2008, **122**(3): 219–223.

Roewer, L., et al., A Y-STR database of Iranian and Azerbaijanian minority populations. *Forensic Sci Int Genet*, 2009, **4**(1): e53–e55.

Romero, R.E., et al., A Colombian Caribbean population study of 16 Y-chromosome STR loci. *Forensic Sci Int Genet*, 2008, **2**(2): e5–e8.

Salvador, J.M., K.A. Tabbada, and M.C. De Ungria, Population data of 10 Y-chromosomal STR loci in Cebu province, Central Visayas (Philippines). *J Forensic Sci*, 2008, **53**(1): 256–258.

Sanchez, C., et al., Haplotype frequencies of 16 Y-chromosome STR loci in the Barcelona metropolitan area population using Y-Filer kit. *Forensic Sci Int*, 2007, **172**(2–3): 211–217.

Schwengber, S.P., et al., Population data of 17 Y-STR loci from Rio Grande do Sul state (South Brazil). *Forensic Sci Int Genet*, 2009, **4**(1): e31–e33.

Seong, K.M., et al., Population genetic polymorphisms of 17 Y-chromosomal STR loci in South Koreans. *Forensic Sci Int Genet*, 2011, **5**(5): e122–e123.

Serin, A., et al., Haplotype frequencies of 17 Y-chromosomal short tandem repeat loci from the Cukurova region of Turkey. *Croat Med J*, 2011, **52**(6): 703–708.

Shi, M., et al., Southwest China Han population data for nine Y-STR loci by multiplex polymerase chain reaction. *J Forensic Sci*, 2007, **52**(1): 228–230.

Shi, M.S., et al., Haplotypes of 20 Y-chromosomal STRs in a population sample from southeast China (Chaoshan area). *Int J Legal Med*, 2007, **121**(6): 455–462.

Shi, M., et al., Population genetics for Y-chromosomal STRs haplotypes of Chinese Tujia ethnic group. *Forensic Sci Int Genet*, 2008, **2**(4): e65–e68.

Shi, M., et al., Haplotype diversity of 22 Y-chromosomal STRs in a southeast China population sample (Chaoshan area). *Forensic Sci Int Genet*, 2009, **3**(2): e45–e47.

Shi, M., et al., Population genetics for Y-chromosomal STRs haplotypes of Chinese Xibe ethnic group. *Forensic Sci Int Genet*, 2011, **5**(5): e119–e121.

Sims, L.M., D. Garvey, and J. Ballantyne, Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs. *Hum Mutat*, 2007, **28**(1): 97.

Siriboonpiputtana, T., et al., Y-chromosomal STR haplotypes in central Thai population. *Forensic Sci Int Genet*, 2010, **4**(3): e71–e72.

Soltyszewski, I., et al., Y-chromosomal haplotypes for the AmpFlSTR Yfiler PCR amplification kit in a population sample from Central Poland. *Forensic Sci Int*, 2007, **168**(1): 61–67.

Stanciu, F., et al., Population data for Y-chromosome haplotypes defined by 17 STRs in South-East Romania. *Leg Med (Tokyo)*, 2010, **12**(5): 259–264.

Stevanovic, M., et al., Human Y-specific STR haplotypes in population of Serbia and Montenegro. *Forensic Sci Int*, 2007, **171**(2–3): 216–221.

Taylor, D., et al., An investigation of admixture in an Australian Aboriginal Y-chromosome STR database. *Forensic Sci Int Genet*, 2012, **6**(5): 532–538.

Taylor, D.A. and J.M. Henry, Haplotype data for 16 Y-chromosome STR loci in Aboriginal and Caucasian populations in South Australia. *Forensic Sci Int Genet*, 2012, **6**(6): e187–e188.

Thangaraj, K., et al., Y-chromosomal STR haplotypes in two endogamous tribal populations of Karnataka, India. *J Forensic Sci*, 2007, **52**(3): 751–753.

Theves, C., et al., Population genetics of 17 Y-chromosomal STR loci in Yakutia. *Forensic Sci Int Genet*, 2010, **4**(5): e129–e130.

Tian-Xiao, Z., Y. Li, and L. Sheng-Bin, Y-STR haplotypes and the genetic structure from eight Chinese ethnic populations. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S198–S200.

Triki-Fendri, S., et al., Population genetics of 17 Y-STR markers in West Libya (Tripoli region). *Forensic Sci Int Genet*, 2013, **7**(3): e59–e61.

Trynova, E.G., et al., Presentation of 17 Y-chromosomal STRs in the population of the Sverdlovsk region. *Forensic Sci Int Genet*, 2011, **5**(3): e101–e104.

Valverde, L., et al., 17 Y-STR haplotype data for a population sample of residents in the Basque Country. *Forensic Sci Int Genet*, 2012, **6**(4): e109–e111.

Valverde, L., et al., Y-STR variation in the Basque diaspora in the Western USA: Evolutionary and forensic perspectives. *Int J Legal Med*, 2012, **126**(2): 293–298.

Veselinovic, I.S., et al., Allele frequencies and population data for 17 Y-chromosome STR loci in a Serbian population sample from Vojvodina province. *Forensic Sci Int*, 2008, **176**(2–3): e23–e28.

Volgyi, A., et al., Hungarian population data for 11 Y-STR and 49 Y-SNP markers. *Forensic Sci Int Genet*, 2009, **3**(2): e27–e28.

Weng, W., et al., Mutation rates at 16 Y-chromosome STRs in the South China Han population. *Int J Legal Med*, 2013, **127**(2): 369–372.

Willuweit, S., et al., Y-STR frequency surveying method: A critical reappraisal. *Forensic Sci Int Genet*, 2011, **5**(2): 84–90.

Wolanska-Nowak, P., et al., A population data for 17 Y-chromosome STR loci in South Poland population sample—Some DYS458.2 variants uncovered and sequenced. *Forensic Sci Int Genet*, 2009, **4**(1): e43–e44.

Wolfgramm Ede, V., et al., Genetic analysis of 15 autosomal and 12 Y-STR loci in the Espirito Santo State population, Brazil. *Forensic Sci Int Genet*, 2011, **5**(3): e41–e43.

Wozniak, M., et al., Continuity of Y chromosome haplotypes in the population of Southern Poland before and after the Second World War. *Forensic Sci Int Genet*, 2007, **1**(2): 134–140.

Wu, F.C., et al., Y-chromosomal STRs haplotypes in the Taiwanese Paiwan population. *Int J Legal Med*, 2011, **125**(1): 39–43.

Wu, W., et al., Population genetics of 17 Y-STR loci in a large Chinese Han population from Zhejiang Province, Eastern China. *Forensic Sci Int Genet*, 2011, **5**(1): e11–e13.

Yadav, B., A. Raina, and T.D. Dogra, Genetic polymorphisms for 17 Y-chromosomal STR haplotypes in Jammu and Kashmir Saraswat Brahmin population. *Leg Med (Tokyo)*, 2010, **12**(5): 249–255.

Yadav, B., A. Raina, and T.D. Dogra, Haplotype diversity of 17 Y-chromosomal STRs in Saraswat Brahmin community of North India. *Forensic Sci Int Genet*, 2011, **5**(3): e63–e70.

Yanmei, Y., et al., Genetic polymorphism of 11 Y-chromosomal STR loci in Yunnan Han Chinese. *Forensic Sci Int Genet*, 2010, **4**(2): e67–e69.

Yoshida, Y. and S. Kubo, Y-SNP and Y-STR analysis in a Japanese population. *Leg Med (Tokyo)*, 2008, **10**(5): 243–252.

Zahra, N., et al., The analysis of UAE populations using AmpFlSTR® Y Filer™: Identification of novel and null alleles. *Forensic Sci Int Genet Suppl*, 2008, **1**(1): 255–256.

Zastera, J., et al., Assembly of a large Y-STR haplotype database for the Czech population and investigation of its substructure. *Forensic Sci Int Genet*, 2010, **4**(3): e75–e78.

Zhang, D., et al., Haplotypes of six miniY-STR loci in the Han population from Sichuan province and the Zhuang population in Guangxi Zhuang autonomous region. *Forensic Sci Int Genet*, 2009, **3**(2): e49–e51.

Zhang, G.Q., et al., Structure and polymorphism of 16 novel Y-STRs in Chinese Han population. *Genet Mol Res*, 2012, **11**(4): 4487–4500.

Zhang, H., et al., Haplotype of 12 Y-STR loci of the PowerPlex Y-system in Sichuan Han ethnic group in west China. *Forensic Sci Int*, 2008, **175**(2–3): 244–249.

Zhang, Y. and J. Lee, Genetic polymorphism of Y-chromosomal STR loci in South Korean population. *Forensic Sci Int*, 2007, **173**(2–3): 225–230.

Zhang, Y., et al., Population genetics for Y-chromosomal STRs haplotypes of Chinese Korean ethnic group in northeastern China. *Forensic Sci Int*, 2007, **173**(2–3): 197–203.

Zhang, Y., et al., Allele frequencies of 12 Y-chromosomal STRs in Chinese Tuvans in the Altay region. *Forensic Sci Int Genet*, 2013, **7**(1): e7–e8.

Zhang, Y.J., et al., Allele frequencies of DYS19, DYS385, DYS388, DYS389 (I and II), DYS390, DYS391, DYS392, and DYS393 STR loci in a Korean population. *J Forensic Sci*, 2007, **52**(6): 1424–1425.

Zhang, Y.L., et al., Genetic diversity of Y-chromosome microsatellites in the Fujian Han and the Sichuan Han populations of China. *Anthropol Anz*, 2007, **65**(1): 1–14.

Zhu, B., et al., Genetic analysis of 17 Y-chromosomal STRs haplotypes of Chinese Tibetan ethnic group residing in Qinghai province of China. *Forensic Sci Int*, 2008, **175**(2–3): 238–243.

Zhu, B., et al., Population genetic polymorphisms for 17 Y-chromosomal STRs haplotypes of Chinese Salar ethnic minority group. *Leg Med (Tokyo)*, 2007, **9**(4): 203–209.

X Chromosome Haplotyping

Castaneda, M., et al., Haplotypic blocks of X-linked STRs for forensic cases: Study of recombination and mutation rates. *J Forensic Sci*, 2012, **57**(1): 192–195.

Chen, D.P., et al., Use of X-linked short tandem repeats loci to confirm mutations in parentage caseworks. *Clin Chim Acta*, 2009, **408**(1–2): 29–33.

Diegoli, T.M. and M.D. Coble, Development and characterization of two mini-X chromosomal short tandem repeat multiplexes. *Forensic Sci Int Genet*, 2011, **5**(5): 415–421.

Edelmann, J., et al., Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28. *Forensic Sci Int Genet*, 2008, **2**(1): 41–46.

Edelmann, J., et al., Validation of six closely linked STRs located in the chromosome X centromere region. *Int J Legal Med*, 2010, **124**(1): 83–87.

Freitas, N.S., et al., X-linked insertion/deletion polymorphisms: Forensic applications of a 33-markers panel. *Int J Legal Med*, 2010, **124**(6): 589–593.

Gusmao, L., et al., Capillary electrophoresis of an X-chromosome STR decaplex for kinship deficiency cases. *Methods Mol Biol*, 2012, **830**: 57–71.

Hering, S., et al., X chromosomal recombination—A family study analysing 39 STR markers in German three-generation pedigrees. *Int J Legal Med*, 2010, **124**(5): 483–491.

Li, C., et al., Development of 11 X-STR loci typing system and genetic analysis in Tibetan and Northern Han populations from China. *Int J Legal Med*, 2011, **125**(5): 753–756.

Li, L., et al., Analysis of 14 highly informative SNP markers on X chromosome by TaqMan SNP genotyping assay. *Forensic Sci Int Genet*, 2010, **4**(5): e145–e148.

Liu, Q.L., D.J. Lu, and H. Zhao, The establishment of a nine X-chromosome short tandem repeat loci multiplex PCR and detection of the polymorphism of these loci. *Chin J Med Genet*, 2009, **26**(6): 664–669.

Liu, Q.L., et al., Development and population study of the 12 X-STR loci multiplexes PCR systems. *Int J Legal Med*, 2012, **126**(4): 665–670.

Liu, Q.L., et al., Development of multiplex PCR system with 15 X-STR loci and genetic analysis in three nationality populations from China. *Electrophoresis*, 2012, **33**(8): 1299–1305.

Nothnagel, M., et al., Collaborative genetic mapping of 12 forensic short tandem repeat (STR) loci on the human X chromosome. *Forensic Sci Int Genet*, 2012, **6**(6): 778–784.

Penna, L.S., et al., Development of two multiplex PCR systems for the analysis of 14 X-chromosomal STR loci in a southern Brazilian population sample. *Int J Legal Med*, 2012, **126**(2): 327–330.

Pepinski, W., et al., X-chromosomal polymorphism data for the ethnic minority of Polish Tatars and the religious minority of Old Believers residing in northeastern Poland. *Forensic Sci Int Genet*, 2007, **1**(2): 212–214.

Pereira, R., et al., A method for the analysis of 32 X chromosome insertion deletion polymorphisms in a single PCR. *Int J Legal Med*, 2012, **126**(1): 97–105.

Turrina, S., et al., Development and forensic validation of a new multiplex PCR assay with 12 X-chromosomal short tandem repeats. *Forensic Sci Int Genet*, 2007, **1**(2): 201–204.

Zarrabeitia, M.T., et al., Analysis of 10 X-linked tetranucleotide markers in mixed and isolated populations. *Forensic Sci Int Genet*, 2009, **3**(2): 63–66.

Genetic Population Data of X Chromosome Haplotyping

Aler, M., et al., Genetic data of 10 X-STRs in a Spanish population sample. *Forensic Sci Int*, 2007, **173**(2–3): 193–196.

Barbaro, A., et al., Distribution of 8 X-chromosomal STR loci in an Italian population sample (Calabria). *Forensic Sci Int Genet*, 2012, **6**(6): e174–e175.

Becker, D., et al., Population genetic evaluation of eight X-chromosomal short tandem repeat loci using Mentype Argus X-8 PCR amplification kit. *Forensic Sci Int Genet*, 2008, **2**(1): 69–74.

Caine, L.M., et al., Genetic data of a Brazilian population sample (Santa Catarina) using an X-STR decaplex. *J Forensic Leg Med*, 2010, **17**(5): 272–274.

Fracasso, T., et al., An X-STR meiosis study in Kurds and Germans: Allele frequencies and mutation rates. *Int J Legal Med*, 2008, **122**(4): 353–356.

Gao, S., et al., Allele frequencies for 10 X-STR loci in Nu population of Yunnan, China. *Leg Med (Tokyo)*, 2007, **9**(5): 284–286.

Gomes, I., et al., Analysis of 10 X-STRs in three African populations. *Forensic Sci Int Genet*, 2007, **1**(2): 208–211.

Gomes, I., et al., The Karimojong from Uganda: Genetic characterization using an X-STR decaplex system. *Forensic Sci Int Genet*, 2009, **3**(4): e127–e128.

Grskovic, B., et al., Analysis of 8 X-chromosomal markers in the population of central Croatia. *Croat Med J*, 2013, **54**(3): 238–247.

Hedman, M., J.U. Palo, and A. Sajantila, X-STR diversity patterns in the Finnish and the Somali population. *Forensic Sci Int Genet*, 2009, **3**(3): 173–178.

Horvath, G., et al., A genetic study of 12 X-STR loci in the Hungarian population. *Forensic Sci Int Genet*, 2012, **6**(1): e46–e47.

Hou, Q.F., B. Yu, and S.B. Li, Genetic polymorphisms of nine X-STR loci in four population groups from Inner Mongolia, China. *Genom Proteom Bioinformat*, 2007, **5**(1): 59–65.

Hwa, H.L., et al., Genetic analysis of eight population groups living in Taiwan using a 13 X-chromosomal STR loci multiplex system. *Int J Legal Med*, 2011, **125**(1): 33–37.

Hwa, H.L., et al., Thirteen X-chromosomal short tandem repeat loci multiplex data from Taiwanese. *Int J Legal Med*, 2009, **123**(3): 263–269.

Illescas, M.J., et al., Population genetic data for 10 X-STR loci in autochthonous Basques from Navarre (Spain). *Forensic Sci Int Genet*, 2012, **6**(5): e146–e148.

Inturri, S., et al., Linkage and linkage disequilibrium analysis of X-STRs in Italian families. *Forensic Sci Int Genet*, 2011, **5**(2): 152–154.

Laouina, A., et al., Genetic data of X-chromosomal STRs in a Moroccan population sample (Casablanca) using an X-STR decaplex. *Forensic Sci Int Genet*, 2013, **7**(3): e90–e92.

Li, C., et al., Genetic analysis of the 11 X-STR loci in Uigur population from China. *Forensic Sci Int Genet*, 2012, **6**(5): e139–e140.

Li, H., et al., A multiplex PCR for 4 X chromosome STR markers and population data from Beijing Han ethnic group. *Leg Med (Tokyo)*, 2009, **11**(5): 248–250.

Lim, E.J., et al., Genetic polymorphism and haplotype analysis of 4 tightly linked X-STR duos in Koreans. *Croat Med J*, 2009, **50**(3): 305–312.

Liu, Q.L., et al., Development of the nine X-STR loci typing system and genetic analysis in three nationality populations from China. *Int J Legal Med*, 2011, **125**(1): 51–58.

Luczak, S., et al., Diversity of 15 human X chromosome microsatellite loci in Polish population. *Forensic Sci Int Genet*, 2011, **5**(3): e71–e77.

Luo, H.B., et al., Characteristics of eight X-STR loci for forensic purposes in the Chinese population. *Int J Legal Med*, 2011, **125**(1): 127–131.

Martins, J.A., et al., X-chromosome genetic variation in Sao Paulo State (Brazil) population. *Ann Hum Biol*, 2010, **37**(4): 598–603.

Mukerjee, S., et al., Genetic variation of 10 X chromosomal STR loci in Indian population. *Int J Legal Med*, 2010, **124**(4): 327–330.

Mukerjee, S., et al., Population genetic data for 11 X-STR loci in eleven populations of India. *Leg Med (Tokyo)*, 2012, **14**(3): 163–165.

Nadeem, A., et al., Development of pentaplex PCR and genetic analysis of X chromosomal STRs in Punjabi population of Pakistan. *Mol Biol Rep*, 2009, **36**(7): 1671–1675.

Nakamura, Y. and K. Minaguchi, Sixteen X-chromosomal STRs in two octaplex PCRs in Japanese population and development of 15-locus multiplex PCR system. *Int J Legal Med*, 2010, **124**(5): 405–414.

Nakamura, Y., et al., Multiplex PCR for 18 X-chromosomal STRs in Japanese population. *Leg Med (Tokyo)*, 2013, **15**(3): 164–170.

Nishi, T., et al., Application of a novel multiplex polymerase chain reaction system for 12 X-chromosomal short tandem repeats to a Japanese population study. *Leg Med (Tokyo)*, 2013, **15**(1): 43–46.

Pereira, V., et al., Study of 25 X-chromosome SNPs in the Portuguese. *Forensic Sci Int Genet*, 2011, **5**(4): 336–338.

Poetsch, M., et al., Allele frequencies of 11 X-chromosomal loci in a population sample from Ghana. *Int J Legal Med*, 2009, **123**(1): 81–83.

Poetsch, M., et al., Allele frequencies of 11 X-chromosomal loci of two population samples from Africa. *Int J Legal Med*, 2011, **125**(2): 307–314.

Ribeiro Rodrigues, E.M., et al., A multiplex PCR for 11 X chromosome STR markers and population data from a Brazilian Amazon region. *Forensic Sci Int Genet*, 2008, **2**(2): 154–158.

Ribeiro-Rodrigues, E.M., et al., Extensive survey of 12 X-STRs reveals genetic heterogeneity among Brazilian populations. *Int J Legal Med*, 2011, **125**(3): 445–452.

Samejima, M., Y. Nakamura, and K. Minaguchi, Population genetic study of six closely linked groups of X-STRs in a Japanese population. *Int J Legal Med*, 2011, **125**(6): 895–900.

Samejima, M., et al., Genetic study of 12 X-STRs in Malay population living in and around Kuala Lumpur using Investigator Argus X-12 kit. *Int J Legal Med*, 2012, **126**(4): 677–683.

Sim, J.E., et al., Population genetic study of four closely-linked X-STR trios in Koreans. *Mol Biol Rep*, 2010, **37**(1): 333–337.

Sun, R., et al., Genetic polymorphisms of 10 X-STR among four ethnic populations in northwest of China. *Mol Biol Rep*, 2012, **39**(4): 4077–4081.

Tariq, M.A., et al., Allele frequency distribution of 13 X-chromosomal STR loci in Pakistani population. *Int J Legal Med*, 2008, **122**(6): 525–528.

Tie, J., S. Uchigasaki, and S. Oshida, Genetic polymorphisms of eight X-chromosomal STR loci in the population of Japanese. *Forensic Sci Int Genet*, 2010, **4**(4): e105–e108.

Tillmar, A.O., et al., Analysis of linkage and linkage disequilibrium for eight X-STR markers. *Forensic Sci Int Genet*, 2008, **3**(1): 37–41.

Tomas, C., et al., X-chromosome SNP analyses in 11 human Mediterranean populations show a high overall genetic homogeneity except in north-west Africans (Moroccans). *BMC Evol Biol*, 2008, **8**: 75.

Uchigasaki, S., J. Tie, and D. Takahashi, Genetic analysis of twelve X-chromosomal STRs in Japanese and Chinese populations. *Mol Biol Rep*, 2013, **40**(4): 3193–3196.

Wu, W., et al., Allele frequencies of seven X-linked STR loci in Chinese Han population from Zhejiang Province. *Forensic Sci Int Genet*, 2009, **4**(1): e41–e42.

Xu, J., et al., Population data of 12 X-chromosomal STR loci in Chinese Han samples from Hebei Province. *Forensic Sci Int Genet*, 2013, **7**(2): e43–e44.

Zalan, A., et al., Hungarian population data of eight X-linked markers in four linkage groups. *Forensic Sci Int*, 2008, **175**(1): 73–78.

Zeng, X., A. Rakha, and S. Li, Genetic polymorphisms of 10 X-chromosome STR loci in Chinese Daur ethnic minority group. *Leg Med (Tokyo)*, 2009, **11**(3): 152–154.

Zeng, X.P., et al., Genetic polymorphisms of twelve X-chromosomal STR loci in Chinese Han population from Guangdong Province. *Forensic Sci Int Genet*, 2011, **5**(4): e114–e116.

Zhang, S., et al., Genetic polymorphisms of 12 X-STR for forensic purposes in Shanghai Han population from China. *Mol Biol Rep*, 2012, **39**(5): 5705–5707.

Zhang, S.H., et al., Genetic polymorphism of eight X-linked STRs of Mentype® Argus X-8 kit in Chinese population from Shanghai. *Forensic Sci Int Genet*, 2011, **5**(1): e21–e24.

Sex Typing for Gender Identification

Blagodatskikh, E.G., et al., Sex determination in biological specimens using the DYS14 marker. *Mol Biol*, 2010, **44**(4): 568–570.

Buel, E., G. Wang, and M. Schwartz, PCR amplification of animal DNA with human X-Y amelogenin primers used in gender determination. *J Forensic Sci*, 1995, **40**(4): 641–644.

Cadenas, A.M., et al., Male amelogenin dropouts: Phylogenetic context, origins and implications. *Forensic Sci Int*, 2007, **166**(2–3): 155–163.

Chang, Y.M., L.A. Burgoyne, and K. Both, Higher failures of amelogenin sex test in an Indian population group. *J Forensic Sci*, 2003, **48**(6): 1309–1313.

Chang, Y.M., et al., A distinct Y-STR haplotype for amelogenin negative males characterized by a large Y(p)11.2 (DYS458-MSY1-AMEL-Y) deletion. *Forensic Sci Int*, 2007, **166**(2–3): 115–120.

Davis, C., et al., A case of amelogenin Y-null: A simple primer binding site mutation or unusual genetic anomaly? *Leg Med (Tokyo)*, 2012, **14**(6): 320–323.

Esteve Codina, A., H. Niederstatter, and W. Parson, "GenderPlex" a PCR multiplex for reliable gender determination of degraded human DNA samples and complex gender constellations. *Int J Legal Med*, 2009, **123**(6): 459–464.

Fincham, A.G., et al., Human developing enamel proteins exhibit a sex-linked dimorphism. *Calcif Tissue Int*, 1991, **48**(4): 288–290.

Frances, F., et al., Amelogenin test: From forensics to quality control in clinical and biochemical genomics. *Clin Chim Acta*, 2007, **386**(1–2): 53–56.

Gibbon, V., et al., Novel methods of molecular sex identification from skeletal tissue using the amelogenin gene. *Forensic Sci Int Genet*, 2009, **3**(2): 74–79.

Graham, E.A.M., Sex determination. *Forensic Sci Med Pathol*, 2006, **2**(4): 283–286.

Haas-Rochholz, H. and G. Weiler, Additional primer sets for an amelogenin gene PCR-based DNA-sex test. *Int J Legal Med*, 1997, **110**(6): 312–315.

Jacot, T.A., et al., TSPY4 is a novel sperm-specific biomarker of semen exposure in human cervicovaginal fluids; potential use in HIV prevention and contraception studies. *Contraception*, 2013, **88**(3): 387–395.

Kashyap, V.K., et al., Deletions in the Y-derived amelogenin gene fragment in the Indian population. *BMC Med Genet*, 2006, **7**: 37.

Kastelic, V., B. Budowle, and K. Drobnic, Validation of SRY marker for forensic casework analysis. *J Forensic Sci*, 2009, **54**(3): 551–555.

Kim, K.Y., et al., A real-time PCR-based amelogenin Y allele dropout assessment model in gender typing of degraded DNA samples. *Int J Legal Med*, 2013, **127**(1): 55–61.

Kumagai, R., et al., DNA analysis of family members with deletion in Yp11.2 region containing amelogenin locus. *Leg Med (Tokyo)*, 2008, **10**(1): 39–42.

Lattanzi, W., et al., A large interstitial deletion encompassing the amelogenin gene on the short arm of the Y chromosome. *Hum Genet*, 2005, **116**(5): 395–401.

Lau, E.C., H.C. Slavkin, and M.L. Snead, Analysis of human enamel genes: Insights into genetic disorders of enamel. *Cleft Palate J*, 1990, **27**(2): 121–130.

Lau, E.C., et al., Human and mouse amelogenin gene loci are on the sex chromosomes. *Genomics*, 1989, **4**(2): 162–168.

Luptáková, L., et al., Sex determination of early medieval individuals through nested PCR using a new primer set in the SRY gene. *Forensic Sci Int*, 2011, **207**(1–3): 1–5.

Ma, Y., et al., Y chromosome interstitial deletion induced Y-STR allele dropout in AMELY-negative individuals. *Int J Legal Med*, 2012, **126**(5): 713–724.

Maciejewska, A. and R. Pawlowski, A rare mutation in the primer binding region of the amelogenin X homologue gene. *Forensic Sci Int Genet*, 2009, **3**(4): 265–267.

Mannucci, A., et al., Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin. *Int J Legal Med*, 1994, **106**(4): 190–193.

Mitchell, R.J., et al., An investigation of sequence deletions of amelogenin (AMELY), a Y-chromosome locus commonly used for gender determination. *Ann Hum Biol*, 2006, **33**(2): 227–240.

Morikawaa, T., Y. Yamamoto, and S. Miyaish, A new method for sex determination based on detection of SRY, STS and amelogenin gene regions with simultaneous amplification of their homologous sequences by a multiplex PCR. *Acta Medica Okayama*, 2011, **65**(2): 113–122.

Mukerjee, S., et al., Differential pattern of genetic variability at the DXYS156 locus on homologous regions of X and Y chromosomes in Indian population and its forensic implications. *Int J Legal Med*, 2013, **127**(1): 1–6.

Nakahori, Y., O. Takenaka, and Y. Nakagome, A human X-Y homologous region encodes "amelogenin". *Genomics*, 1991, **9**(2): 264–269.

Ou, X., et al., Null alleles of the X and Y chromosomal amelogenin gene in a Chinese population. *Int J Legal Med*, 2012, **126**(4): 513–518.

Oz, C., et al., A Y-chromosome STR marker should be added to commercial multiplex STR kits. *J Forensic Sci*, 2008, **53**(4): 858–861.

Parton, J., et al., Sex identification of ancient DNA samples using a microfluidic device. *J Archaeol Sci*, 2013, **40**(1): 705–711.

Quincey, D., et al., Difficulties of sex determination from forensic bone degraded DNA: A comparison of three methods. *Sci Justice*, 2013, **53**(3): 253–260.

Roffey, P.E., C.I. Eckhoff, and J.L. Kuhl, A rare mutation in the amelogenin gene and its potential investigative ramifications. *J Forensic Sci*, 2000, **45**(5): 1016–1019.

Sire, J.Y., S. Delgado, and M. Girondot, The amelogenin story: Origin and evolution. *Eur J Oral Sci*, 2006, **114**(Suppl 1): 64–77; discussion 93–95, 379–380.

Snead, M.L., et al., Of mice and men: Anatomy of the amelogenin gene. *Connect Tissue Res*, 1989, **22**(1–4): 101–109.

Steinlechner, M., et al., Rare failures in the amelogenin sex test. *Int J Legal Med*, 2002, **116**(2): 117–120.

Takayama, T., et al., Determination of deleted regions from Yp11.2 of an amelogenin negative male. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S578–S580.

Thangaraj, K., A.G. Reddy, and L. Singh, Is the amelogenin gene reliable for gender identification in forensic casework and prenatal diagnosis? *Int J Legal Med*, 2002, **116**(2): 121–123.

Tomas, C., et al., Typing of 48 autosomal SNPs and amelogenin with GenPlex SNP genotyping system in forensic genetics. *Forensic Sci Int Genet*, 2008, **3**(1): 1–6.

Tozzo, P., et al., Deletion of amelogenin Y-locus in forensics: Literature revision and description of a novel method for sex confirmation. *J Forensic Leg Med*, 2013, **20**(5): 387–391.

Turrina, S., G. Filippini, and D. De Leo, Evaluation of deleted region from Yp11.2 of two amelogenin negative related males. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 240–241.

Umeno, M., et al., A rapid and simple system of detecting deletions on the Y chromosome related with male infertility using multiplex PCR. *J Med Invest*, 2006, **53**(1–2): 147–152.

von Wurmb-Schwark, N., H. Bosinski, and S. Ritz-Timme, What do the X and Y chromosomes tell us about sex and gender in forensic case analysis? *J Forensic Leg Med*, 2007, **14**(1): 27–30.

22

Single Nucleotide Polymorphism Profiling

22.1 Basic Characteristics of SNPs

Human genomes contain sequence polymorphisms, which are the variations in nucleotide sequences among individuals. One type of sequence polymorphism is called *single-nucleotide polymorphism* (*SNP*). An SNP constitutes a single-base-pair change originating from a spontaneous mutation that can be a base substitution, insertion, or deletion at a single site. An estimated 10 million SNPs may exist in the human genome, and among them, over a million SNPs have been identified. Thus, SNPs are the most frequent form of DNA polymorphism observed in humans. Most SNPs appear in noncoding regions and some SNPs are found in coding regions of genes (Figure 22.1).

The vast majority of SNPs are biallelic, although very rare triallelic and tetraallelic SNPs also exist. If an SNP originating from a spontaneous mutation occurs in the germ line, it can be inherited by offspring and can spread in the human population. As a result, both the wild type (the typical form of an allele occurring in nature) and a mutant allele are produced. This type of SNP is referred to as a biallelic SNP. Subsequently, a third mutation, a very rare event, may occur at the same nucleotide site. This SNP has a wild type and two different mutant alleles, which results in a triallelic SNP. It is also possible that a triallelic SNP is created by introducing two of the mutations simultaneously at a single site, which is also a very rare event.

Generally speaking, there are a number of advantages to utilizing SNP loci as markers for forensic applications. First, SNPs are abundant within the human genome; therefore, a sufficient number of SNP loci that are suitable for human identification can be selected. Second, in SNP analysis, amplified fragments (amplicons) are usually 50–100 base pairs in length and are smaller than that in short tandem repeat (STR) analysis (Chapter 20). Therefore, SNP profiling can be more useful than STR for analyzing degraded DNA samples in which genomic DNA is fragmented (Section 20.5.1). Third, SNP loci have lower mutation rates than STRs and are, therefore, useful for specialized human identification such as paternity testing. Lastly, it is possible to achieve high-throughput SNP analysis by utilizing multiplex SNP assays, which are amendable for automation.

The application of SNPs for forensic DNA analysis also has some disadvantages, in that SNP loci are not as polymorphic as STR loci. It is estimated that the analysis of 50–60 SNP loci is needed to achieve a similar level of population match probability (P_m—the lower the P_m, the less likely a match occurs between two randomly chosen individuals; see Chapter 25) using 13

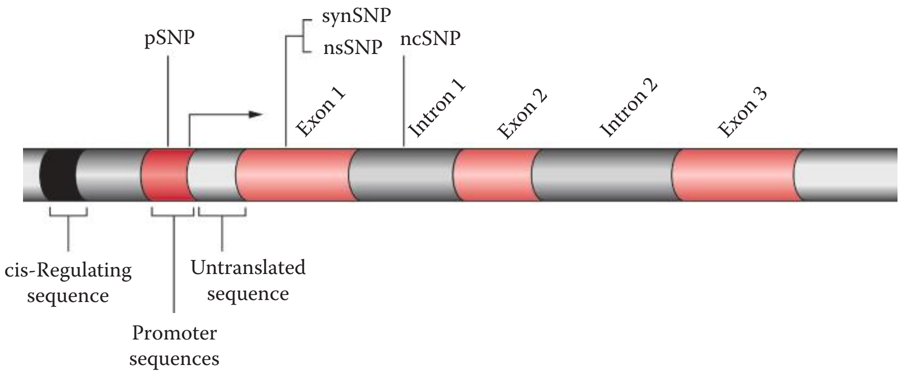

Figure 22.1 SNPs fall into several classes. Most reside in the noncoding regions of DNA and are designated as noncoding SNPs (ncSNPs). A subset of ncSNPs can also be found in introns. SNPs residing in exons are further divided into two types: The synonymous type (synSNP) is an exonic SNP that does not change the amino acid composition of the encoded polypeptide. Conversely, a nonsynonymous type (nsSNP) changes the encoded amino acid. SNPs in the promoter regions of the genome are known as promoter SNPs (pSNPs). Arrow: transcription start site. (© Richard C. Li.)

STR loci in CODIS. Moreover, most SNPs are biallelic; therefore, it is difficult to resolve mixed DNA profiles when a shared common allele exists in a mixture of more than one individual. Furthermore, most DNA databases contain STR profiles instead of SNP profiles, which means that it is not possible to carry out a database search for a matching SNP profile in cases where no suspect has yet been identified.

22.2 Forensic Applications of SNP Profiling

22.2.1 *HLA-DQA1* Locus

The first use of SNP-based profiling for forensic application involved sequence polymorphisms at the *HLA-DQA1* locus (formerly called the DQα locus). The *HLA-DQA1* gene is a member of the human leukocyte antigen (HLA) family, which contains a large number of genes involved in the immune response in humans. The *HLA-DQA1* locus is located within human HLA gene clusters on chromosome 6. The region tested for forensic use is highly polymorphic and is located at the second exon of the gene (Table 22.1).

22.2.1.1 DQα AmpliType and Polymarker Assays

The SNP profiling of the *HLA-DQA1* locus was carried out using the DQα AmpliType® kit. It was the first commercial kit, developed in the late 1980s by Cetus Corporation in Emeryville, CA. The *HLA-DQA1* panel can distinguish the following: alleles 1 (subtyped as 1.1, 1.2, or 1.3), 2, 3, and 4 (subtyped as 4.1 and 4.2/4.3, in which the 4.2 and 4.3 alleles are combined and cannot be distinguished). Therefore, a total of 28 possible genotypes from combinations of these alleles can be distinguished. Although the P_m of this SNP profiling is high (approximately 5×10^{-2}), it is useful as a preliminary test to quickly exclude suspects.

In addition to the *HLA-DQA1* locus, five additional loci—*LDLR, GYPA, HBGG,* D7S8, and *GC*—were utilized for forensic application in 1993 (Table 22.1). These loci were included in the second generation of the kit known as the AmpliType® PM PCR amplification and typing kit (also known as Polymarker), manufactured by Perkin-Elmer (Norwalk, CT). It consisted of one panel for the testing of *HLA-DQAl* and another panel for an additional five loci (Figure 22.2). Among these five additional loci, *LDLR, GYPA*, and D7S8 each have two alleles (designated

Table 22.1 Chromosomal Locations of SNP Loci Used in AmpliType® PM PCR Amplification and Typing Kit

Locus	Gene Product	Chromosome Location	Number of Alleles	Further Reading
HLA-DQA1	HLA-DQA1	6p21.3	7	Gyllensten and Erlich (1988)
LDLR	Low-density lipoprotein receptor	19p 13.1-13.3	2	Yamamoto et al. (1984)
GYPA	Glycophorin A	4q28-31	2	Siebert and Fukuda (1987)
HBGG	Hemoglobin G gammaglobin	11 p 15.5	3	Slightom et al. (1980)
D7S8	Anonymous	7q22-31.1	2	Horn et al. (1990)
GC	Group-specific component	4q11-13	3	Yang et al. (1985)

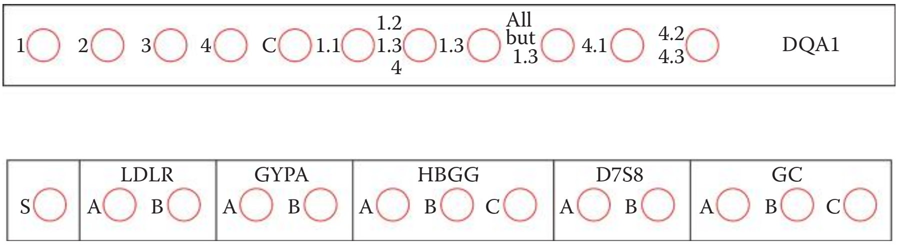

Figure 22.2 Panels of immobilized probes in the Polymarker kit. Top: *HLA-DQA1*. Bottom: additional five loci (*LDLR, GYPA, HBGG*, D7S8, and *GC*). C and S represent threshold control dots. (© Richard C. Li.)

A and B), while *HBGG* and *GC* each have three alleles (A, B, and C). As a result, the P_m of the Polymarker panels decreased to 10^{-4}.

The DNA profiles of the Polymarker system have been accepted in US courts. The Polymarker system has a number of advantages, particularly compared to variable number tandem repeat analysis (VNTR; see Chapter 19). First, the Polymarker system is a PCR-based method and is capable of analyzing a small amount of DNA sample (approximately 2 ng per analysis). Therefore, it is more sensitive than the method used in VNTR analysis. Second, the Polymarker system is an SNP assay with short amplicon sizes and, therefore, can analyze degraded DNA samples, which is not possible in VNTR analysis. Third, the Polymarker system is more rapid and less laborious than VNTR analysis. Lastly, the amplicon sizes of alleles at a given locus have identical lengths and, therefore, this assay does not exhibit preferential amplification as in amplified fragment length polymorphism (Chapter 19). However, the Polymarker system has its limitations compared to VNTR analysis. For example, compared to the VNTR analysis, the P_m of the Polymarker system is high, resulting in poor discrimination power in the comparison of two DNA profiles. Therefore, the Polymarker system is merely useful for excluding a suspect. Moreover, the SNP loci utilized in the Polymarker system are less polymorphic than VNTR loci. The limited number of alleles per SNP locus makes identifying the components of mixtures more difficult than with VNTR analysis in situations when contributors of a mixture sample share common alleles. For these reasons, the Polymarker system was replaced by STR profiling in the late 1990s.

22.2.1.2 Allele-Specific Oligonucleotide Hybridization

Both DQα AmpliType® and Polymarker kits utilize the *allele-specific oligonucleotide (ASO) hybridization* technique. This technique analyzes single-nucleotide variations, such as SNPs, at a given locus. It is based on the principle that ASO probes, usually 14–17 bases in length, hybridize to their complementary DNA sequences to distinguish known polymorphic alleles. ASO probes for multiple alleles at several loci can be arranged on the same panel to establish the presence or absence of specific alleles in PCR-amplified fragments of a DNA sample (Figure 22.3). Thus, the genotypes can be determined.

In the DQα AmpliType® and Polymarker kits, the oligonucleotides representing different alleles are immobilized to a solid matrix consisting of nylon membrane strips. Each immobilized probe at a particular site on the membrane is utilized to detect corresponding SNPs. Since the probe rather than the target DNA (as with regular blot format; see Chapter 9) is immobilized to the solid phase, the configuration used here is known as a *reverse blot* format.

The regions of the DNA in question are amplified by PCR. One of each pair of the primers is conjugated with biotin at the 5′ end (Figure 22.4). Thus, the amplicons are biotinylated for purposes of detection. This kit established multiplexing of a six-locus system, allowing simultaneous amplification of the *HLA-DQA1* locus along with *LDLR, GYPA, HBGG*, D7S8, and *GC* in

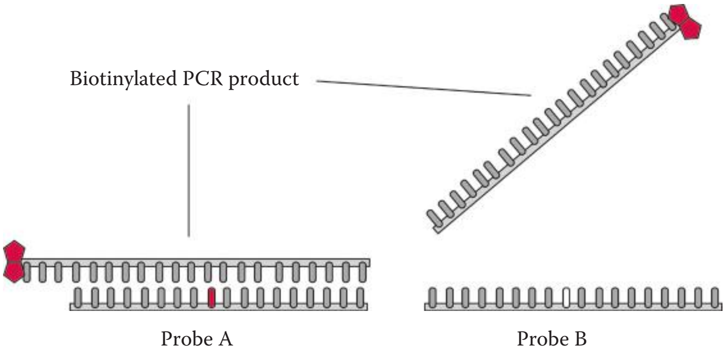

Figure 22.3 Hybridization with ASO. Two probes are incubated with the target DNA containing the SNP. Only the perfectly matched probe–target DNA can hybridize under optimal conditions.

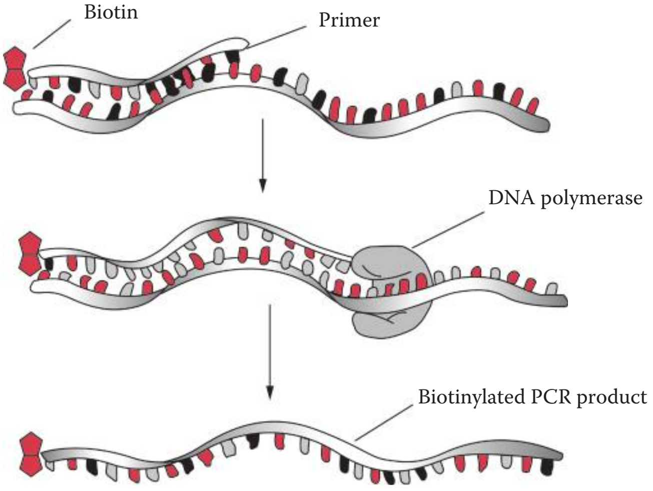

Figure 22.4 Amplification of DNA using biotinylated primers. The biotin is conjugated at the 5′ end of the primer. The amplified products are biotinylated. (© Richard C. Li.)

a single reaction for each sample. Following the denaturation of the PCR product to separate the two complementary DNA strands of the amplicons, the biotinylated strands are hybridized to the immobilized probes. Hybridization and washing conditions are established to ensure proper hybridization of the ASO probe and its target sequence. Unbound amplicons are washed away.

The presence of a PCR product bound to a specific probe can be detected by a colorimetric detection system (Figure 22.5). Since the amplicons are biotinylated, a horseradish peroxidase–conjugated streptavidin complex is utilized to bind to the biotinylated amplicons. The horseradish peroxidase then catalyzes the oxidation reaction of the colorless substrate, tetramethyl benzidine (TMB), into a blue precipitate at a designated location on the membrane strip, allowing the genotype of the sample to be determined (Figures 22.6 and 22.7).

The kit uses a threshold control (C dot in the *HLA-DQA1* panel and S dot in the panel of the other five loci) to distinguish between signal and background noise and to determine whether a

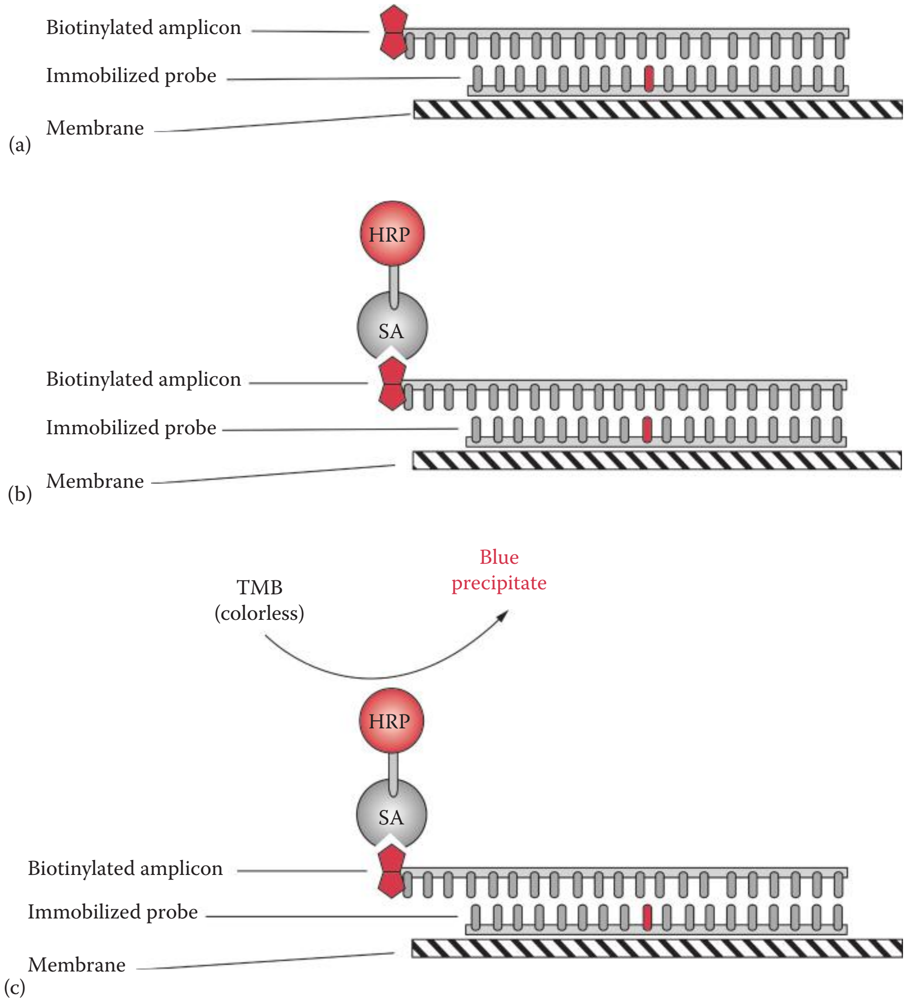
Biotinylated amplicon
Immobilized probe
Membrane
(a)
HRP
SA
Biotinylated amplicon
Immobilized probe
Membrane
(b)
TMB
(colorless)
Blue
precipitate
HRP
SA
Biotinylated amplicon
Immobilized probe
Membrane
(c)

Figure 22.5 Reverse blot assay. (a) The probe is immobilized onto a solid-phase membrane and hybridized with a biotinylated PCR product having the target sequence. (b) The detection of hybridization is carried out by streptavidin (SA) and horseradish peroxidase (HRP) conjugate. (c) Colorimetric reaction is catalyzed by HRP using TMB as a substrate. (© Richard C. Li.)

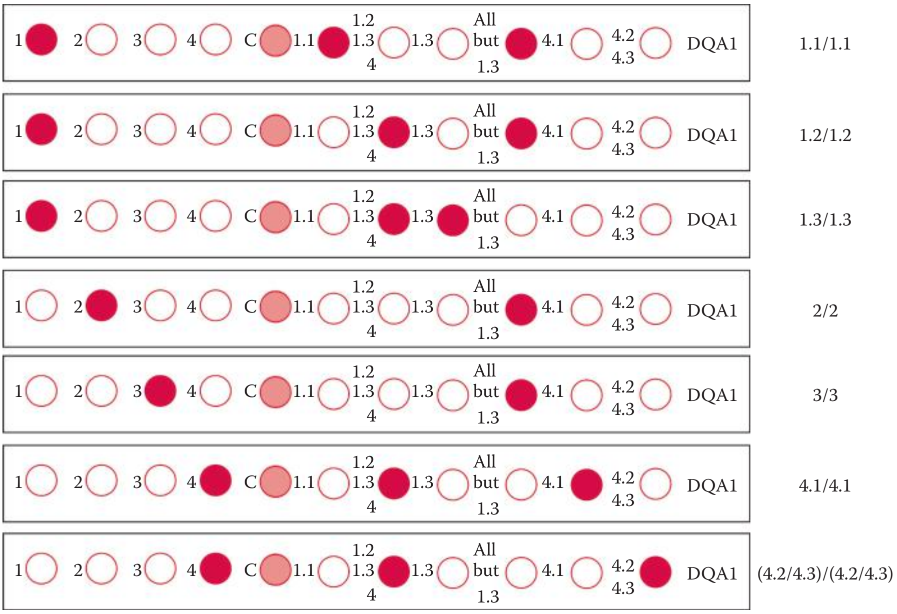
1
2
3
4
C
1.1
1.2 1.3 4
1.3
All but 1.3
4.1
4.2 4.3
DQA1
1.1/1.1
1.2/1.2
1.3/1.3
2/2
3/3
4.1/4.1
(4.2/4.3)/(4.2/4.3)

Figure 22.6 DNA typing of homozygous alleles at the *HLA-DQA1* locus using the reverse blot assay (Polymarker kit). The genotype of each sample is noted on the right. (© Richard C. Li.)

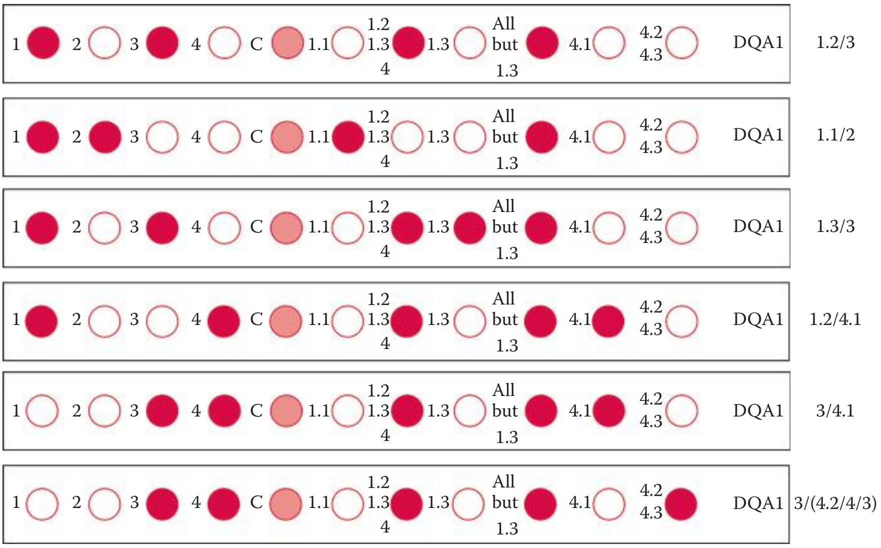
1
2
3
4
C
1.1
1.2 1.3 4
1.3
All but 1.3
4.1
4.2 4.3
DQA1
1.2/3
1.1/2
1.3/3
1.2/4.1
3/4.1
3/(4.2/4/3)

Figure 22.7 DNA typing of heterozygous alleles at the *HLA-DQA1* locus using the reverse blot assay (Polymarker kit). The representative genotypes of the samples are noted on the right. (© Richard C. Li.)

sufficient amount of DNA has been amplified to detect all the alleles present in a sample. If the signal intensities of allele dots are greater than or equal to the threshold control, the alleles are considered true. If the allele dots are less intense than the threshold control, they are considered inconclusive for the determination of full genotypes because an allele may not have been detected due to a low level of DNA template.

22.2.2 Current and Potential Applications of SNP Analysis

22.2.2.1 Application of SNP Analysis for Forensic Identification

The applications of SNP analysis for forensic human identification are summarized in Table 22.2. Autosomal SNP panels can be used for the most common types of forensic testing, including analysis of degraded DNA samples. The SNP panels for mitochondrial DNA (mtDNA) profiling are useful for identifying human remains by comparing the DNA profiles in question to those of potential relatives. This approach may serve as an alternative method to the DNA sequencing method of mtDNA analysis, which is time-consuming and laborious. SNP loci on the Y chromosome are also potentially useful markers for paternity testing because of low mutation rates. SNP loci such as ancestry informative markers (AIM) can be used to determine the ethnic origins of questioned samples to generate leads for investigations.

Table 22.2 Examples of Forensic Applications of SNP Profiling

Testing	Candidate SNP Loci	Application	Further Reading
Identity	Autosomal SNPs	Human identification	Kidd et al. (2006); Sanchez et al. (2006)
		Human identification via degraded DNA	Budowle (2004)
	mtDNA SNPs	Human identification	Grignani et al. (2006)
	Y chromosome SNPs	Paternity testing	Hammer et al. (2006)
Biogeographical origin	Ancestry informative markers (AIMs)	Ethnic group identification	Frudakis et al. (2003); Shriver and Kittles (2004)
Physical characteristics	*MC1R* (melanocortin 1 receptor gene)	Hair color identification (investigative lead)	Grimes et al. (2001)
	P (gene has role in pigmentation)	Eye color identification (investigative lead)	Rebbeck et al. (2002)
Pathology	*KCNH2* (cardiac potassium channel gene)	Determining cause of sudden death from cardiac arrhythmia long QT syndrome	Lunetta et al. (2002)
	SCN5A (encodes cardiac sodium channel gene)	Determining cause of sudden death from cardiac arrhythmia long QT syndrome	Burke et al. (2005)
Toxicology	*CYP2C19*, *CYP2D6*, *CYP3A4*, *CYP2E1* (drug metabolizing enzyme genes)	Investigation of drug overdose (including death)	Kupiec et al. (2006)

22.2.2.2 Potential Applications of SNPs for Phenotyping

One potential application of SNP analysis is in determining phenotypic information, also known as *phenotyping* (Table 22.2). The relevant SNP loci are usually *nonsynonymous SNPs* (*nsSNPs*); they reside in the exon and change the encoded amino acid, which leads to an altered phenotype. Phenotyping of a questioned sample can reveal physical characteristics of an individual, such as hair and eye color, to provide leads for investigations. A number of SNPs residing within the melanocortin 1 receptor gene (*MC1R*) are associated with red hair, fair skin, and freckles, while SNPs residing within the *P* gene, which play a role in pigmentation, are associated with eye color variations.

Phenotyping can also be employed in the area of forensic pathology. Cardiac arrhythmia long QT syndrome (LQTS) can cause sudden death. A number of LQTS-associated SNPs—for example, SNPs in *KCNH2* and *SCN5A* genes—have been shown to correlate to such deaths. Thus, these SNPs are potentially useful for investigating the causes of death. Finally, phenotyping also has applications in forensic toxicology. A number of SNPs in genes, such as *CYP2D6*, that are responsible for metabolizing drugs can serve as potential markers for postmortem investigations of drug-overdose cases.

22.3 SNP Techniques

Over the years, various techniques of SNP analysis have been developed and can be divided into several groups based on the mechanisms used: allele-specific hybridization, primer extension, oligonucleotide ligation, and invasive cleavage. In *allele-specific hybridization*, allele discrimination is based on an optimal condition allowing only the perfectly matched probe–target hybridization to form. *Primer extension* methods are based on the ability of DNA polymerase to incorporate specific deoxynucleotides (dNTPs) complementary to the sequence of the template DNA. *Allele-specific oligonucleotide ligation* is based on the condition that only the allelic probe perfectly matched to the target is ligated. In the *invasive cleavage* method, allelic discrimination is based on DNA sequence-specific cleavage by endonucleases. A number of detection methods can be utilized in SNP analysis, such as the measurements of fluorescence, luminescence, and molecular mass. Most assays are carried out in solutions or on solid matrices such as glass slides, chips, or beads. Table 22.3 summarizes the representative assays for SNP typing.

For decades, Sanger sequencing, using chain-termination chemistry, has been the standard method for DNA sequencing (Chapter 23). In recent years, *next-generation sequencing* (NGS), a rapidly developing technology, has had a profound impact on biology. Compared to Sanger sequencing, NGS is advantageous in that it can achieve substantially higher throughput, and at lower cost, than the Sanger method. For example, a human genome can now be sequenced in several days using NGS technologies. Although NGS is not yet widely utilized for forensic applications, it has great potential for forensic DNA analysis, particularly in SNP analysis.

22.3.1 Next-Generation Sequencing Technologies

NGS technologies have two categories of application: *de novo* sequencing and resequencing. In *de novo* genome sequencing, uncharacterized genomes or characterized genomes with substantial structural variations are sequenced. Sequence reads are assembled without any reference sequence. In *resequencing* applications, characterized genomes are sequenced. Sequence reads are assembled against an existing reference sequence to identify sequence polymorphisms. Thus, resequencing can potentially be used for forensic applications to detect polymorphisms associated with human identification. *Target resequencing* is a useful method of resequencing that can be utilized for forensic applications. Prior to sequencing, the genomic regions of interest from a DNA sample are selectively isolated through a method known as *enrichment*. Several target-enrichment strategies have been developed. PCR is the most widely used enrichment method.

Table 22.3 Techniques for SNP Assay

Basis of Technique	Representative Assay	Detection	Format	Notes	Further Reading
Allele-specific hybridization	Reverse Blot	Colorimetry	Membrane based	Useful for screen assay	Saiki et al. (1989)
	LightCycler® (Roche)	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	Lareu et al. (2001)
	TaqMan® (Applied Biosystems)	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	De la Vega et al. (2005)
	Molecular beacons	Fluorescence	Solution based	Applicable for automation; limited multiplexing capability	Tyagi et al. (1998)
	GeneChip® (Affymetrix)	Fluorescence	Chip based	Designed for genotyping large number of SNPs; exceeds forensic needs	Mei et al. (2000)
Primer extension	SNaPshot™ (Applied Biosystems)	Fluorescence	Solution based	Applicable for capillary electrophoresis and multiplexing	Sanchez et al. (2006)
	PinPoint (Applied Biosystems)	Mass spectrometry	Solution based	Multiplex capability lower than SNaPshot™	Haff and Smirnov (1997)
	Arrayed primer extension (APEX)	Fluorescence	Chip based	Low reproducibility	Pastinen et al. (1997)
	Pyrosequencing	Fluorescence	Solution based	May be applicable for analysis of mixed profiles; not applicable for automation; limited multiplexing capability	Ronaghi et al. (1998)
Allele-specific oligonucleotide ligation	SNPstream® UHT (Orchid-Gene Screen)	Fluorescence	Chip based	Developed for mass disaster cases	Bell et al. (2002)
	SNPlex™ (Applied Biosystems)	Fluorescence	Solution based	Applicable for multiplexing; requires higher amount of target DNA template than other PCR-based methods	De la Vega et al. (2005)
Invasive cleavage	Invader® (Third Wave Technology)	Fluorescence	Solution based	PCR not required; large amount of target DNA required	Lyamichev et al. (1999)

Source: Adapted from Sobrino, B., M. Brion, and A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int, 2005. 154(2-3): p. 181–94.; Budowle, B., SNP typing strategies. Forensic Sci Int, 2004. 146 Suppl: p. S139–42.

The PCR-based approach is highly effective in targeting genomic regions that are small in size. Nevertheless, a typical NGS involves several major steps, including sample, library, and template DNA preparation; template amplification; sequencing and detection; and base calling; as well as data analysis (Figure 22.8).

22.3.2 DNA Samples, Sequencing Library, and Template Preparation

For most sequencing applications, micrograms of purified DNA are needed. This requirement is still a challenge for forensic casework applications where nanograms of DNA are often obtained. Human autosomal genome sequencing usually requires converting a DNA sample into a sequencing library. Two types of sequencing libraries are usually used: mate-pair libraries and fragment libraries. A mate-pair library is often used in *de novo* sequencing applications, while the fragment libraries can be used for resequencing and forensic applications. To construct a fragment library, a DNA sample is fragmented using mechanical methods such as sonication, nebulization, or shearing; endonuclease digestion; or a transposon-based method. Subsequently, sequencing adapters containing primer-binding sites for universal PCR primers are ligated to both ends of the DNA fragments. In resequencing applications, multiplex sequencing of pooled samples is often carried out, which can improve the efficiency and reduce the costs of sequencing. For multiplex sequencing, an index tag containing a bar code can be ligated to each DNA fragment to allow it to be identified after sequencing. Indexing can minimize the risk of sample mix-ups and contaminations during the sample preparation.

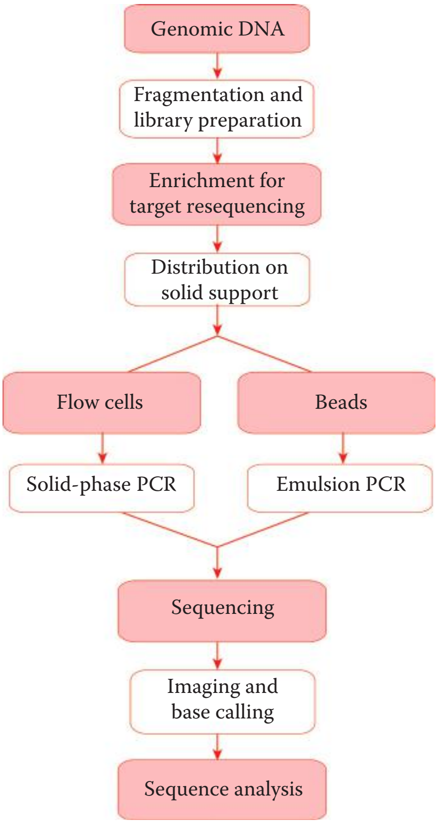
Genomic DNA
Fragmentation and library preparation
Enrichment for target resequencing
Distribution on solid support
Flow cells
Beads
Solid-phase PCR
Emulsion PCR
Sequencing
Imaging and base calling
Sequence analysis

Figure 22.8 NGS work flow for potential forensic applications. (© Richard C. Li.)

Most imaging systems of sequencing platforms are not equipped to detect single-molecule fluorescence. Thus, DNA templates are amplified using either emulsion or solid-phase PCR. In the *emulsion PCR*, single-stranded DNA templates are bound to beads. DNA-bound beads are suspended in a water-in-oil emulsion, where each bead is placed in a single aqueous droplet. DNA fragments are amplified using PCR. As a result, each bead is coated with thousands of copies of the same template sequence. Subsequently, beads carrying amplified DNA are deposited on a glass slide or into individual fiber-optic wells for sequencing. In *solid-phase PCR*, both forward and reverse primers are attached to a slide (Figure 22.9). DNA templates are hybridized to immobilized PCR primers. The extended primer bends and hybridizes to a second immobilized primer, forming a bridge. During the "bridge amplification," the single-stranded template is amplified to form clusters. Solid-phase amplification can produce millions of spatially separated template clusters with free ends, allowing universal sequencing primers to hybridize for sequencing.

22.3.3 NGS Chemistry

The amplified template fragments are then sequenced. At present, there are two major categories of NGS chemistry that are routinely used: *sequencing by synthesis* and *sequencing by ligation*. The *pyrosequencing* technology is one example of NGS using the sequencing-by-synthesis chemistry. During the pyrosequencing, each nucleotide substrate is introduced one at a time. Only the correct nucleotide corresponding to the template is incorporated, and a pyrophosphate is released (Figure 22.10a). Pyrophosphate is then converted to ATP, catalyzed by sulfurylase (Figure 22.10b). ATP is then utilized by luciferase to convert luciferin to oxyluciferin, and the reaction emits light that is detected by a charge-coupled device camera (Figure 22.10c). This technology allows the generation of sequences with long read lengths. *Cyclic reversible termination* is another example of NGS using sequencing-by-synthesis chemistry. In this method, chain terminators (fluorescent-labeled dideoxynucleotides) are used to extend a primer sequence complementary to the template DNA. Each of four nucleotides is labeled with a different fluorophore. After incorporation and fluorescence detection, the terminating and fluorescent moieties are cleaved and removed (Figure 22.11). As a result, the next sequencing cycle is carried out. This process is repeated until the sequence is completed. This is the most commonly used method throughout the field of NGS.

In the sequencing-by-ligation chemistry, probes (eight nucleotides in length) are utilized for ligation reactions. Each probe contains five specific nucleotides that are complementary to the template and three nucleotides that are universal. Probes containing all possible combinations of the first five nucleotides are utilized. Each probe is fluorescently labeled according to the first two bases of the probe so that it can be identified. During the first round of sequencing reaction, only the probe that is complementary to the template sequence can hybridize. The probe is then ligated to the 3′ end of the primer. After detection, three universal nucleotides, including fluorophore, are cleaved. The ligation process is repeated several times using a new set of probes each time. The newly made complementary DNA strand is then stripped off. In the next round of sequencing, new primers that are one nucleotide shorter than the previous primer are utilized, and the ligation reaction is repeated. In this round of sequencing, different nucleotide positions are read. This process is repeated for several rounds with different primers, until all nucleotides in the template have been sequenced twice. As a result, this sequencing method has a low error rate.

22.3.4 NGS Coverage

In NGS, sequencing reads are usually not distributed evenly over the genomic regions of interest. As a result, some nucleotides will be covered by fewer or more reads than the average. Therefore, multiple reads for each nucleotide are necessary to obtain a reliable sequence. The average number of times that each nucleotide in the genomic regions of interest is sequenced is known as the

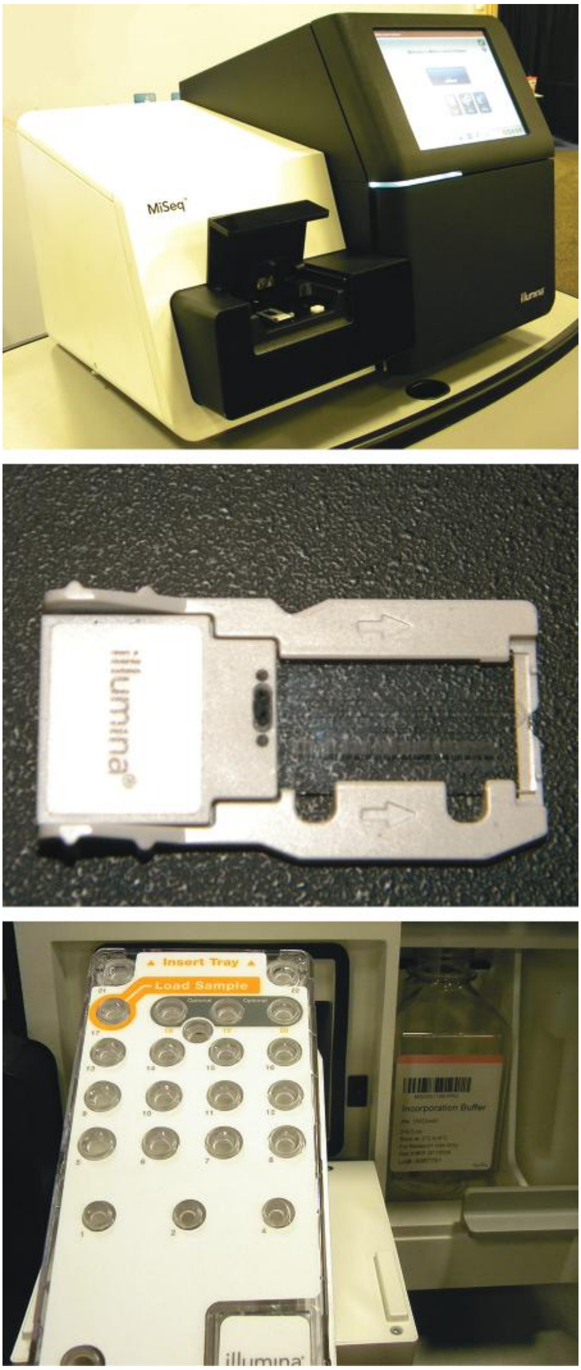
MiSeq
illumina
illumina
Insert Tray
Load Sample
Incorporation Buffer
illumina

Figure 22.9 A photo of an NGS platform. A MiSeq (Illumna) desktop sequencer (top), a flow-cell device (middle) where DNA templates are distributed for PCR, and a reagent cartridge (bottom) are shown. (© Richard C. Li.)

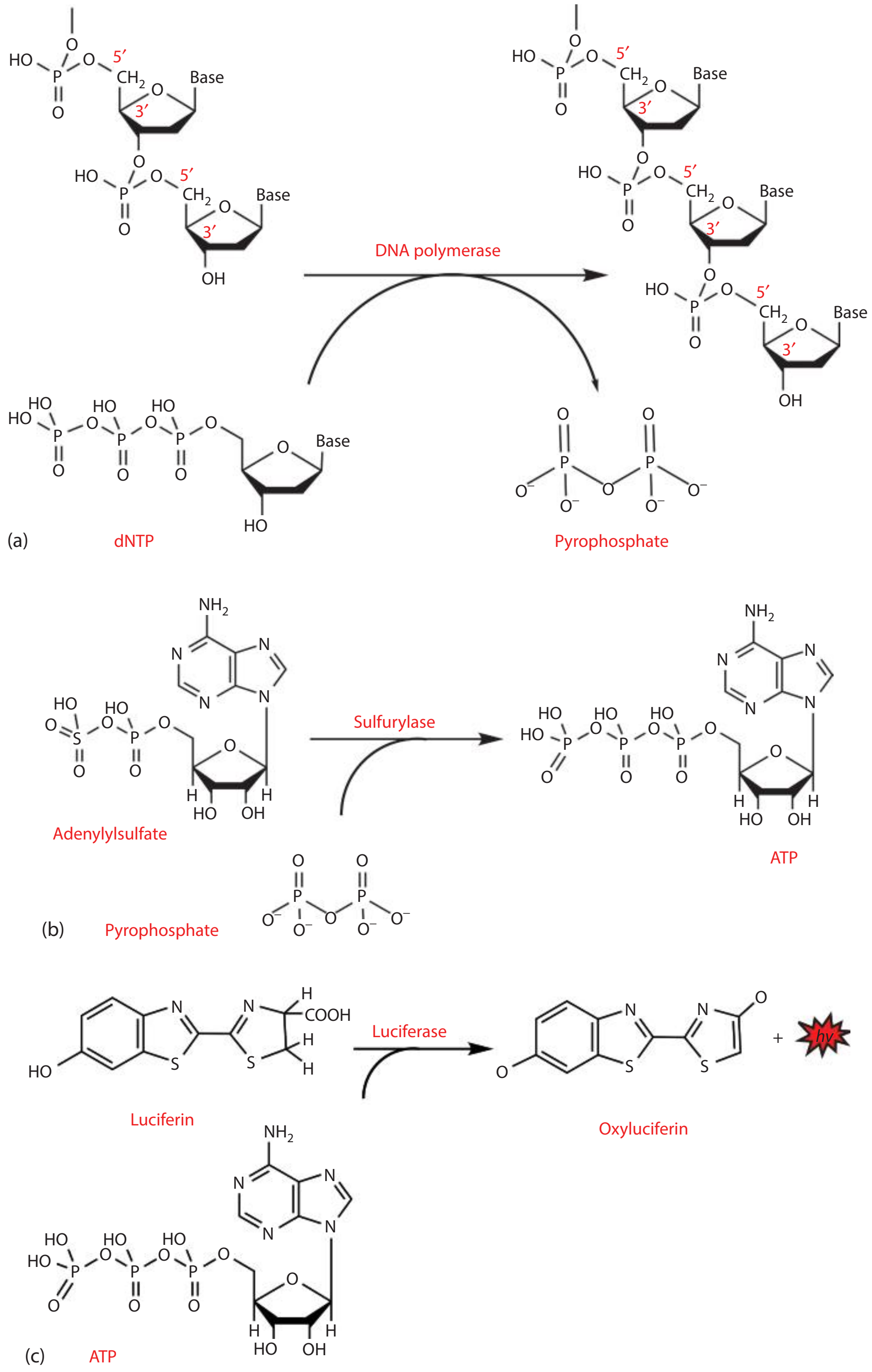
(a)
dNTP
DNA polymerase
Pyrophosphate
Base
5′
3′
OH
(b)
Adenylylsulfate
Sulfurylase
ATP
Pyrophosphate
(c)
Luciferin
Luciferase
Oxyluciferin
hν
ATP

Figure 22.10 Pyrosequencing chemistry. (a) During the pyrosequencing, a nucleotide substrate is incorporated, releasing a pyrophosphate. (b) Pyrophosphate is then converted to ATP. (c) ATP is utilized by luciferase, converting luciferin to oxyluciferin, which emits light.

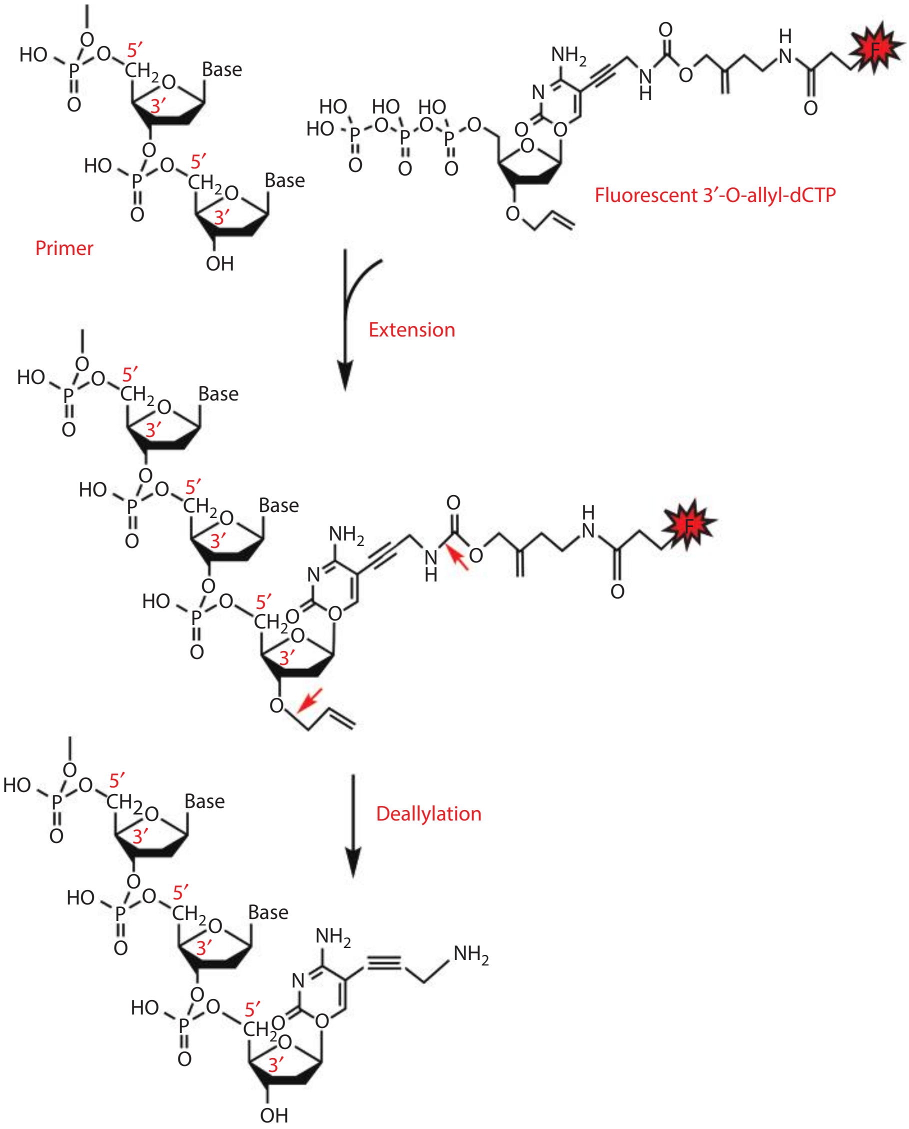
Primer
Fluorescent 3′-O-allyl-dCTP
Extension
Deallylation
Base
5′
3′
CH2
HO
OH
NH2
F

Figure 22.11 Cyclic reversible termination. A single sequencing cycle is shown. The incorporation of a C is illustrated as an example.

coverage or the *sequencing depth*. The level of coverage depends on the types of applications. For example, for human genome SNP analysis, the level of coverage is from 10 to 30. The coverage can be calculated using the following equation:

$$C = \frac{LN}{G}$$

where:

C = coverage
G = haploid genome length
L = read length
N = number of reads

Currently, NGS has not been widely used in forensic applications. One disadvantage of NGS is that the error rate of NGS is higher than that of Sanger sequencing. Although the error rate of NGS is not yet acceptable for forensic casework analysis, it can potentially be overcome by increasing the sequencing coverage.

Bibliography

Ackerman, M.J., D.J. Tester, and D.J. Driscoll, Molecular autopsy of sudden unexplained death in the young. *Am J Forensic Med Pathol*, 2001, **22**(2): 105–111.

Ali, M.E., et al., HLA-A, -B and -DRB1 allele frequencies in the Bangladeshi population. *Tissue Antigens*, 2008, **72**(2): 115–119.

Alkhayat, A., F. Alshamali, and B. Budowle, Population data on the PCR-based loci LDLR GYPA, HBGG, D7S8, Gc, HLA-DQA1, and D1S80 from Arabs from Dubai. *Forensic Sci Int*, 1996, **81**(1): 29–34.

Allah, R., L. Yang, and S.B. Li, SNPs and forensic DNA typing. *Fa Yi Xue Za Zhi*, 2007, **23**(5): 373–379.

Allwood, J.S. and S. Harbison, SNP model development for the prediction of eye colour in New Zealand. *Forensic Sci Int Genet*, 2013, **7**(4): 444–452.

Altshuler, D., et al., An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 2000, **407**(6803): 513–516.

Amigo, J., et al., The SNPforID browser: An online tool for query and display of frequency data from the SNPforID project. *Int J Legal Med*, 2008, **122**(5): 435–440.

Arnestad, M., et al., Prevalence of long-QT syndrome gene variants in sudden infant death syndrome. *Circulation*, 2007, **115**(3): 361–367.

Babol-Pokora, K. and J. Berent, SNP-minisequencing as an excellent tool for analysing degraded DNA recovered from archival tissues. *Acta Biochim Pol*, 2008, **55**(4): 815–819.

Bandelt, H.J. and A. Salas, Current next generation sequencing technology may not meet forensic standards. *Forensic Sci Int Genet*, 2012, **6**(1): 143–145.

Barbaro, A., et al., Genetic variability of the SNPforID 52-plex identification SNP panel in Italian population samples. *Forensic Sci Int Genet*, 2012, **6**(6): e185–e186.

Bell, B., et al., Distribution of types for six PCR-based loci; LDLR, GYPA, HBGG, D7S8, GC and HLA-DQA1 in central Pyrenees and Teruel (Spain). *J Forensic Sci*, 1997, **42**(3): 510–513.

Bell, P.A., et al., SNP stream UHT: Ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Bio Techniques*, 2002, **74**(Suppl 70-2): 76–77.

Birngruber, C., F. Ramsthaler, and M.A. Verhoff, The color(s) of human hair—Forensic hair analysis with SpectraCube. *Forensic Sci Int*, 2009, **185**(1–3): e19–e23.

Blake, E., et al., Polymerase chain reaction (PCR) amplification and human leukocyte antigen (HLA)-DQ alpha oligonucleotide typing on biological evidence samples: Casework experience. *J Forensic Sci*, 1992, **37**(3): 700–726.

Blanco-Verea, A., et al., Forensic validation and implementation of Y-chromosome SNP multiplexes. *Forensic Sci Int Genet Suppl*, 2008, **1**(1): 181–183.

Borsting, C. and N. Morling, Mutations and/or close relatives? Six case work examples where 49 autosomal SNPs were used as supplementary markers. *Forensic Sci Int Genet*, 2011, **5**(3): 236–241.

Borsting, C. and N. Morling, Reinvestigations of six unusual paternity cases by typing of autosomal single-nucleotide polymorphisms. *Transfusion*, 2012, **52**(2): 425–430.

Borsting, C., C. Tomas, and N. Morling, SNP typing of the reference materials SRM 2391b 1-10, K562, XY1, XX74, and 007 with the SNPforID multiplex. *Forensic Sci Int Genet*, 2011, **5**(3): e81–e82.

Borsting, C., C. Tomas, and N. Morling, Typing of 49 autosomal SNPs by single base extension and capillary electrophoresis for forensic genetic testing. *Methods Mol Biol*, 2012, **830**: 87–107.

Borsting, C., E. Rockenbauer, and N. Morling, Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Sci Int Genet*, 2009, **4**(1): 34–42.

Borsting, C., H.S. Mogensen, and N. Morling, Forensic genetic SNP typing of low-template DNA and highly degraded DNA from crime case samples. *Forensic Sci Int Genet*, 2013, **7**(3): 345–352.

Borsting, C., et al., Performance of the SNPforID 52 SNP-plex assay in paternity testing. *Forensic Sci Int Genet*, 2008, **2**(4): 292–300.

Bouakaze, C., et al., Pigment phenotype and biogeographical ancestry from ancient skeletal remains: Inferences from multiplexed autosomal SNP analysis. *Int J Legal Med*, 2009, **123**(4): 315–325.

Branicki, W., U. Brudnik, and A. Wojas-Pelc, Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype. *Ann Hum Genet*, 2009, **73**(2): 160–170.

Branicki, W., et al., Determination of phenotype associated SNPs in the MC1R gene. *J Forensic Sci*, 2007, **52**(2): 349–354.

Branicki, W., et al., Model-based prediction of human hair color using DNA variants. *Hum Genet*, 2011, **129**(4): 443–454.

Brenner, C.H. and B.S. Weir, Issues and strategies in the DNA identification of World Trade Center victims. *Theor Popul Biol*, 2003, **63**(3): 173–178.

Brito, R.M., et al., South Portuguese population data on the loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8 and Gc. *J Forensic Sci*, 1998, **43**(5): 1031–1036.

Brown, R.J., et al., Distribution of the HLA-DQA1 and polymarker alleles in the Basque population of Spain. *Forensic Sci Int*, 2000, **108**(2): 145–151.

Budowle, B. and A. van Daal, Forensically relevant SNP classes. *Biotechniques*, 2008, **44**(5): 603–608, 610.

Budowle, B., F.S. Baechtel, and R. Fejeran, Polymarker, HLA-DQA1, and D1S80 allele frequency data in Chamorro and Filipino populations from Guam. *J Forensic Sci*, 1998, **43**(6): 1195–1198.

Budowle, B., SNP typing strategies. *Forensic Sci Int*, 2004, **146**(Suppl): S139–S142.

Budowle, B., et al., PCR-based analysis: Allele-specific oligonucleotide assays, in *DNA Typing Protocols: Molecular Biology and Forensic Analysis*, pp. 95–118, 2000. Eaton: Natick, MA.

Budowle, B., et al., Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci), and HLA-DQ alpha using a multiplex amplification and typing procedure. *J Forensic Sci*, 1995, **40**(1): 45–54.

Bunten, H., et al., OPRM1 and CYP2B6 gene variants as risk factors in methadone-related deaths. *Clin Pharmacol Ther*, 2010, **88**(3): 383–389.

Burke, A., et al., Role of SCN5A Y1102 polymorphism in sudden cardiac death in blacks. *Circulation*, 2005, **112**(6): 798–802.

Caratti, S., et al., Subtyping of Y-chromosomal haplogroup E-M78 (E1b1b1a) by SNP assay and its forensic application. *Int J Legal Med*, 2009, **123**(4): 357–360.

Chen, X., K.J. Livak, and P.Y. Kwok, A homogeneous, ligase-mediated DNA diagnostic test. *Genome Res*, 1998, **8**(5): 549–556.

Comey, C.T. and B. Budowle, Validation studies on the analysis of the HLA DQ alpha locus using the polymerase chain reaction. *J Forensic Sci*, 1991, **36**(6): 1633–1648.

Comey, C.T., et al., PCR amplification and typing of the HLA DQ alpha gene in forensic samples. *J Forensic Sci*, 1993, **38**(2): 239–249.

Consortium, I.H., A haplotype map of the human genome. *Nature*, 2005, **437**(7063): 1299–1320.

Cowland, J.B., H.O. Madsen, and N. Morling, HLA-DQA1 typing in Danes by two polymerase chain reaction (PCR) based methods. *Forensic Sci Int*, 1995, **73**(1): 1–13.

De la Vega, F.M., et al., Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex senotyping system. *Mutat Res*, 2005, **573**(1–2): 111–135.

de Paula Careta, F. and G.G. Paneto, Recent patents on high-throughput single nucleotide polymorphism (SNP) genotyping methods. *Recent Pat DNA Gene Seq*, 2012, **6**(2): 122–126.

Dimo-Simonin, N. and C. Brandt-Casadevall, Evaluation and usefulness of reverse dot blot DNA-PolyMarker typing in forensic case work. *Forensic Sci Int*, 1996, **81**(1): 61–72.

Dixon, L.A., et al., Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci Int*, 2005, **154**(1): 62–77.

Doktycz, M.J., et al., Analysis of polymerase chain reaction-amplified DNA products by mass spectrometry using matrix-assisted laser desorption and electrospray: Current status. *Anal Biochem*, 1995, **230**(2): 205–214.

Drobnic, K., et al., Typing of 49 autosomal SNPs by SNaPshot in the Slovenian population. *Forensic Sci Int Genet*, 2010, **4**(5): e125–e127.

Duffy, D.L., et al., Interactive effects of MC1R and OCA2 on melanoma risk phenotypes. *Hum Mol Genet*, 2004, **13**(4): 447–461.

Edelmann, J., et al., Long QT syndrome mutation detection by SNaPshot technique. *Int J Legal Med*, 2012, **126**(6): 969–973.

Eiberg, H. and J. Mohr, Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. *Eur J Hum Genet*, 1996, **4**(4): 237–241.

Erlich, H.A., E.L. Sheldon, and G. Horn, HLA typing using DNA probes. *Biotechnology*, 1986, **4**: 7.

Erlich, H.A., HlA-DQ alpha typing of forensic specimens. *Forensic Sci Int*, 1992, **53**(2): 227–228.

Erlich, H.A., et al., Reliability of the HLA-DQ alpha PCR-based oligonucleotide typing system. *J Forensic Sci*, 1990, **35**(5): 1017–1019.

Fan, J.B., et al., Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res*, 2000, **10**(6): 853–860.

Fei, Z., T. Ono, and L.M. Smith, MALDI-TOF mass spectrometric typing of single nucleotide polymorphisms with mass-tagged ddNTPs. *Nucleic Acids Res*, 1998, **26**(11): 2827–2828.

Fildes, N. and R. Reynolds, Consistency and reproducibility of AmpliType PM results between seven laboratories: Field trial results. *J Forensic Sci*, 1995, **40**(2): 279–286.

Flanagan, N., et al., Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Hum Mol Genet*, 2000, **9**(17): 2531–2537.

Fondevila, M., et al., Forensic performance of two insertion-deletion marker assays. *Int J Legal Med*, 2012, **126**(5): 725–737.

Fondevila, M., et al., Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet*, 2013, **7**(1): 63–74.

Freire-Aradas, A., et al., A new SNP assay for identification of highly degraded human DNA. *Forensic Sci Int Genet*, 2012, **6**(3): 341–349.

Friis, S.L., et al., Typing of 30 insertion/deletions in Danes using the first commercial indel kit—Mentype® DIPplex. *Forensic Sci Int Genet*, 2012, **6**(2): e72–e74.

Frudakis, T., et al., Sequences associated with human iris pigmentation. *Genetics*, 2003, **165**(4): 2071–2083.

Gene, M., et al., Population study of Aymara Amerindians for the PCR-DNA polymorphisms HUMTH01, HUMVWA31A, D3S1358, D8S1179, D18S51, D19S253, YNZ22 and HLA-DQ alpha. *Int J Legal Med*, 2000, **113**(2): 126–128.

Geppert, M. and L. Roewer, SNaPshot® minisequencing analysis of multiple ancestry-informative Y-SNPs using capillary electrophoresis. *Methods Mol Biol*, 2012, **830**: 127–140.

Giardina, E., et al., Frequency assessment of SNPs for forensic identification in different populations. *Forensic Sci Int Genet*, 2007, **1**(3–4): e1–e3.

Gill, P., et al., An assessment of whether SNPs will replace STRs in national DNA databases—Joint considerations of the DNA working group of the European Network of Forensic Science Institutes (ENFSI) and the Scientific Working Group on DNA Analysis Methods (SWGDAM). *Sci Justice*, 2004, **44**(1): 51–53.

Gino, S., et al., LDLR, GYPA, HBGG, D7S8 and GC allele and genotype frequencies in the northwest Italian population. *J Forensic Sci*, 1999, **44**(1): 171–174.

Giroti, R. and V.K. Kashyap, Detection of the source of mislabeled biopsy tissue paraffin block and histopathological section on glass slide. *Diagn Mol Pathol*, 1998, **7**(6): 331–334.

Giroti, R.I., R. Biswas, and K. Mukherjee, Restriction fragment length polymorphism and polymerase chain reaction: HLA-DQA1 and polymarker analysis of blood samples from transfusion recipients. *Am J Clin Pathol*, 2002, **118**(3): 382–387.

Gra, O., et al., Microarray-based detection of CYP1A1, CYP2C9, CYP2C19, CYP2D6, GSTT1, GSTM1, MTHFR, MTRR, NQO1, NAT2, HLA-DQA1, and AB0 allele frequencies in native Russians. *Genet Test Mol Biomarkers*, 2010, **14**(3): 329–342.

Graf, J., R. Hodgson, and A. van Daal, Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum Mutat*, 2005, **25**(3): 278–284.

Graham, E.A., DNA reviews: Predicting phenotype. *Forensic Sci Med Pathol*, 2008, **4**(3): 196–199.

Grignani, P., et al., Subtyping mtDNA haplogroup H by SNaPshot minisequencing and its application in forensic individual identification. *Int J Legal Med*, 2006, **120**(3): 151–156.

Grimes, E.A., et al., Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Sci Int*, 2001, **122**(2–3): 124–129.

Gross, A.M. and R.A. Guerrieri, HLA DQA1 and Polymarker validations for forensic casework: Standard specimens, reproducibility, and mixed specimens. *J Forensic Sci*, 1996, **41**(6): 1022–1026.

Grossman, P.D., et al., High-density multiplex detection of nucleic acid sequences: Oligonucleotide ligation assay and sequence-coded separation. *Nucleic Acids Res*, 1994, **22**(21): 4527–4534.

Gyllensten, U.B. and H.A. Erlich, Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. *Proc Natl Acad Sci U S A*, 1988, **85**(20): 7652–7656.

Haff, L.A. and I.P. Smirnov, Multiplex genotyping of PCR products with MassTag-labeled primers. *Nucleic Acids Res*, 1997, **25**(18): 3749–3750.

Hall, J.G., et al., Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction. *Proc Natl Acad Sci U S A*, 2000, **97**(15): 8272–8277.

Hallenberg, C. and N. Morling, A report of the 1997, 1998 and 1999 Paternity Testing Workshops of the English Speaking Working Group of the International Society for Forensic Genetics. *Forensic Sci Int*, 2001, **116**(1): 23–33.

Hammer, M.F., et al., Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic Sci Int*, 2006, **164**(1): 45–55.

Hardenbol, P., et al., Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, 2003, **21**(6): 673–678.

Hart, K.L., et al., Improved eye- and skin-color prediction based on 8 SNPs. *Croat Med J*, 2013, **54**(3): 248–256.

Hashiyada, M., et al., Development of a spreadsheet for SNPs typing using Microsoft EXCEL. *Leg Med*, 2009, **11**(Suppl 1): S453–S454.

Heinrich, M., et al., Reduced-volume and low-volume typing of Y-chromosomal SNPs to obtain Finnish Y-chromosomal compound haplotypes. *Int J Legal Med*, 2009, **123**(5): 413–418.

Helmuth, R., et al., HLA-DQ alpha allele and genotype frequencies in various human populations, determined by using enzymatic amplification and oligonucleotide probes. *Am J Hum Genet*, 1990, **47**(3): 515–523.

Herrin, G., N. Fildes, and R. Reynolds, Evaluation of the AmpliType PM DNA test system on forensic case samples. *J Forensic Sci*, 1994, **39**: 9.

Hill, C.R., J.M. Butler, and P.M. Vallone, A 26plex autosomal STR assay to aid human identity testing*. *J Forensic Sci*, 2009, **54**(5): 1008–1015.

Hirschhorn, J.N., et al., SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. *Proc Natl Acad Sci U S A*, 2000, **97**(22): 12164–12169.

Hochmeister, M.N., et al., A method for the purification and recovery of genomic DNA from an HLA DQA1 amplification product and its subsequent amplification and typing with the AmpliType PM PCR amplification and typing kit. *J Forensic Sci*, 1995, **40**(4): 649–733.

Hochmeister, M.N., et al., Swiss population data on the loci HLA-DQ alpha, LDLR, GYPA, HBGG, D7S8, Gc and D1S80. *Forensic Sci Int*, 1994, **67**(3): 175–184.

Horn, G.T., et al., Characterization and rapid diagnostic analysis of DNA polymorphisms closely linked to the cystic fibrosis locus. *Clin Chem*, 1990, **36**(9): 1614–1619.

Hromadnikova, I., et al., Analysis of paternal alleles in nucleated red blood cells enriched from maternal blood. *Folia Biol*, 2001, **47**(1): 36–39.

Hsu, T.M., et al., Genotyping single-nucleotide polymorphisms by the invader assay with dual-color fluorescence polarization detection. *Clin Chem*, 2001, **47**(8): 1373–1377.

Huang, N.E. and B. Budowle, Chinese population data on the PCR-based loci HLA-DQ alpha, low-density-lipoprotein receptor, glycophorin A, hemoglobin gamma G, D7S8, and group-specific component. *Hum Hered*, 1995, **45**(1): 34–40.

Huckenbeck, W., et al., German population data on the loci low-density-lipoprotein receptor, glycophorin A, hemoglobin gamma G, D7S8 and group-specific component. *Anthropol Anz*, 1995, **53**(3): 193–198.

Inagaki, S., et al., A new 39-plex analysis method for SNPs including 15 blood group loci. *Forensic Sci Int*, 2004, **144**(1): 45–57.

Jankowski, L.B., et al., New Jersey Caucasian, African American, and Hispanic population data on the PCR-based loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, and Gc. *J Forensic Sci*, 1998, **43**(5): 1037–1040.

Johansen, P., et al., Evaluation of the iPLEX® sample ID Plus Panel designed for the Sequenom MassARRAY® system. A SNP typing assay developed for human identification and sample tracking based on the SNPforID panel. *Forensic Sci Int Genet*, 2013, **7**(5): 482–487.

Jung, J.M., et al., Extraction strategy for obtaining DNA from bloodstains for PCR amplification and typing of the HLA-DQ alpha gene. *Int J Legal Med*, 1991, **104**(3): 145–148.

Kastelic, V. and K. Drobnic, A single-nucleotide polymorphism (SNP) multiplex system: The association of five SNPs with human eye and hair color in the Slovenian population and comparison using a Bayesian network and logistic regression model. *Croat Med J*, 2012, **53**(5): 401–408.

Kastelic, V., et al., Prediction of eye color in the Slovenian population using the IrisPlex SNPs. *Croat Med J*, 2013, **54**(4): 381–386.

Keating, B., et al., First all-in-one diagnostic tool for DNA intelligence: Genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int J Legal Med*, 2013, **127**(3): 559–572.

Kersbergen, P., et al., Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. *BMC Genet*, 2009, **10**: 69.

Khodjet-el-Khil, H., et al., Genetic analysis of the SNPforID 34-plex ancestry informative SNP panel in Tunisian and Libyan populations. *Forensic Sci Int Genet*, 2011, **5**(3): e45–e47.

Kidd, K.K., et al., Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int*, 2006, **164**(1): 20–32.

Kiesler, K.M. and P.M. Vallone, Allele frequencies for 40 autosomal SNP loci typed for US population samples using electrospray ionization mass spectrometry. *Croat Med J*, 2013, **54**(3): 225–231.

Kim, J.J., et al., Development of SNP-based human identification system. *Int J Legal Med*, 2010, **124**(2): 125–131.

Kimura, R., et al., Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Mol Biol Evol*, 2008, **25**(8): 1750–1761.

Kingback, M., et al., Influence of CYP2D6 genotype on the disposition of the enantiomers of venlafaxine and its major metabolites in postmortem femoral blood. *Forensic Sci Int*, 2012, **214**(1–3): 124–134.

Kitano, T., et al., Allele frequencies of a SNP and a 27-bp deletion that are the determinant of earwax type in the ABCC11 gene. *Leg Med*, 2008, **10**(2): 113–114.

Kling, D., et al., DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects. *Forensic Sci Int Genet*, 2012, **6**(3): 322–329.

Kostrikis, L.G., et al., Spectral genotyping of human alleles. *Science*, 1998, **279**(5354): 1228–1229.

Krawczak, M., Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, 1999, **20**(8): 1676–1681.

Krjutskov, K., et al., Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on microarrays. *Nucleic Acids Res*, 2008, **36**(12): e75.

Krjutskov, K., et al., Evaluation of the 124-plex SNP typing microarray for forensic testing. *Forensic Sci Int Genet*, 2009, **4**(1): 43–48.

Kubo, S., et al., Personal identification from skeletal remain by D1S80, HLA DQA1, TH01 and polymarker analysis. *J Med Invest*, 2002, **49**(1–2): 83–86.

Kupiec, T.C., V. Raj, and N. Vu, Pharmacogenomics for the forensic toxicologist. *J Anal Toxicol*, 2006, **30**(2): 65–72.

Kuppuswamy, M.N., et al., Single nucleotide primer extension to detect genetic diseases: Experimental application to hemophilia B (factor IX) and cystic fibrosis genes. *Proc Natl Acad Sci U S A*, 1991, **88**(4): 1143–1147.

Lai, E., Application of SNP technologies in medicine: Lessons learned and future challenges. *Genome Res*, 2001, **11**(6): 927–929.

Lamason, R.L., et al., SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 2005, **310**(5755): 1782–1786.

Landegren, U., et al., A ligase-mediated gene detection technique. *Science*, 1988, **241**(4869): 1077–1080.

Lareu, M., et al., The use of the LightCycler for the detection of Y chromosome SNPs. *Forensic Sci Int*, 2001, **118**(2–3): 163–168.

Larsen, M.K., et al., Molecular autopsy in young sudden cardiac death victims with suspected cardiomyopathy. *Forensic Sci Int*, 2012, **219**(1–3): 33–38.

Larsen, M.K., et al., Postmortem genetic testing of the ryanodine receptor 2 (RYR2) gene in a cohort of sudden unexplained death cases. *Int J Legal Med*, 2013, **127**(1): 139–144.

LaRue, B.L., et al., A validation study of the Qiagen investigator DIPplex® kit; an INDEL-based assay for human identification. *Int J Legal Med*, 2012, **126**(4): 533–540.

Levo, A., et al., Post-mortem SNP analysis of CYP2D6 gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood. *Forensic Sci Int*, 2003, **135**(1): 9–15.

Li, C., et al., Genetic polymorphism of 29 highly informative InDel markers for forensic use in the Chinese Han population. *Forensic Sci Int Genet*, 2011, **5**(1): e27–e30.

Li, C., et al., Selection of 29 highly informative InDel markers for human identification and paternity analysis in Chinese Han population by the SNPlex genotyping system. *Mol Biol Rep*, 2012, **39**(3): 3143–3152.

Lou, C., et al., A SNaPshot assay for genotyping 44 individual identification single nucleotide polymorphisms. *Electrophoresis*, 2011, **32**(3–4): 368–378.

Lunetta, P., et al., Death in bathtub revisited with molecular genetics: A victim with suicidal traits and a LQTS gene mutation. *Forensic Sci Int*, 2002, **130**(2–3): 122–124.

Lyamichev, V., et al., Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat Biotechnol*, 1999, **17**(3): 292–296.

Manta, F., et al., Indel markers: Genetic diversity of 38 polymorphisms in Brazilian populations and application in a paternity investigation with post mortem material. *Forensic Sci Int Genet*, 2012, **6**(5): 658–661.

Martin, P., et al., Population genetic data of 30 autosomal indels in central Spain and the Basque Country populations. *Forensic Sci Int Genet*, 2013, **7**(2): e27–e30.

Martinez-Cadenas, C., et al., Gender is a major factor explaining discrepancies in eye colour prediction based on HERC2/OCA2 genotype and the IrisPlex model. *Forensic Sci Int Genet*, 2013, **7**(4): 453–460.

Medintz, I., et al., HLA-DQA1 and polymarker allele frequencies in two New York City Jewish populations. *J Forensic Sci*, 1997, **42**(5): 919–922.

Mei, R., et al., Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res*, 2000, **10**(8): 1126–1137.

Mein, C.A., et al., Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res*, 2000, **10**(3): 330–343.

Mengel-From, J., et al., Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet*, 2010, **4**(5): 323–328.

Mikkelsen, M., et al., Frequencies of 33 coding region mitochondrial SNPs in a Danish and a Turkish population. *Forensic Sci Int Genet*, 2011, **5**(5): 559–560.

Muddiman, D.C., et al., Length and base composition of PCR-amplified nucleic acids using mass measurements from electrospray ionization mass spectrometry. *Anal Chem*, 1997, **69**(8): 1543–1549.

Musgrave-Brown, E., et al., Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet*, 2007, **1**(2): 186–190.

Myakishev, M.V., et al., High-throughput SNP genotyping by allele-specific PCR with universal energy-transfer-labeled primers. *Genome Res*, 2001, **11**(1): 163–169.

Nakahara, H., et al., Automated SNPs typing system based on the Invader assay. *Leg Med*, 2009, **11**(Suppl 1): S111–S114.

Nakajima, T., et al., Evaluation of 7 DNA markers (D1S80, HLA-DQ alpha, LDLR, GYPA, HBGG, D7S8 and GC) in a Japanese population. *Int J Legal Med*, 1996, **109**(1): 47–48.

Neitzke-Montinelli, V., et al., Polymorphisms upstream of the melanocortin-1 receptor coding region are associated with human pigmentation variation in a Brazilian population. *Am J Hum Biol*, 2012, **24**(6): 853–855.

Neuvonen, A.M., et al., Discrimination power of investigator DIPplex loci in Finnish and Somali populations. *Forensic Sci Int Genet*, 2012, **6**(4): e99–e102.

Nicklas, J.A. and E. Buel, A real-time multiplex SNP melting assay to discriminate individuals. *J Forensic Sci*, 2008, **53**(6): 1316–1324.

Odriozola, A., et al., SNPSTR rs59186128_D7S820 polymorphism distribution in European Caucasoid, Hispanic, and Afro-American populations. *Int J Legal Med*, 2009, **123**(6): 527–533.

Osawa, M., et al., SNP association and sequence analysis of the NOS1AP gene in SIDS. *Leg Med*, 2009, **11**(Suppl 1): S307–S308.

Pai, C.Y., et al., Flow chart HLA-DQA1 genotyping and its application to a forensic case. *J Forensic Sci*, 1995, **40**(2): 228–235.

Pakstis, A.J., et al., Candidate SNPs for a universal individual identification panel. *Hum Genet*, 2007, **121**(3–4): 305–317.

Parson, W., et al., Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). *Forensic Sci Int Genet*, 2013, **7**(5): 543–549.

Pastinen, T., et al., Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res*, 1997, **7**(6): 606–614.

Pastinen, T., et al., A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res*, 2000, **10**(7): 1031–1042.

Pereira, R. and L. Gusmao, Capillary electrophoresis of 38 noncoding biallelic mini-Indels for degraded samples and as complementary tool in paternity testing. *Methods Mol Biol*, 2012, **830**: 141–157.

Pereira, R., et al., A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis*, 2009, **30**(21): 3682–3690.

Phillips, C., SNP databases. *Methods Mol Biol*, 2009, **578**: 43–71.

Phillips, C., M. Fondevila, and M.V. Lareau, A 34-plex autosomal SNP single base extension assay for ancestry investigations. *Methods Mol Biol*, 2012, **830**: 109–126.

Phillips, C., et al., Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Sci Int Genet*, 2007, **1**(2): 180–185.

Phillips, C., et al., Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, 2007, **1**(3–4): 273–280.

Phillips, C., et al., Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. *Forensic Sci Int Genet*, 2008, **2**(3): 198–204.

Phillips, C., et al., Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet*, 2013, **7**(3): 359–366.

Pilgrim, J.L., et al., Characterization of single nucleotide polymorphisms of cytochrome p450 in an Australian deceased sample. *Curr Drug Metab*, 2012, **13**(5): 679–692.

Pimenta, J.R. and S.D. Pena, Efficient human paternity testing with a panel of 40 short insertion-deletion polymorphisms. *Genet Mol Res*, 2010, **9**(1): 601–607.

Pinto, N., et al., Assessing paternities with inconclusive STR results: The suitability of bi-allelic markers. *Forensic Sci Int Genet*, 2013, **7**(1): 16–21.

Podini, D. and P.M. Vallone, SNP genotyping using multiplex single base primer extension assays. *Methods Mol Biol*, 2009, **578**: 379–391.

Poetsch, M., et al., Prediction of people's origin from degraded DNA—Presentation of SNP assays and calculation of probability. *Int J Legal Med*, 2013, **127**(2): 347–357.

Pomeroy, R., et al., A low-cost, high-throughput, automated single nucleotide polymorphism assay for forensic human DNA applications. *Anal Biochem*, 2009, **395**(1): 61–67.

Porras, L., et al., Genetic variability of the SNPforID 52-plex identification-SNP panel in Central West Colombia. *Forensic Sci Int Genet*, 2009, **4**(1): e9–e10.

Pulker, H., et al., Finding genes that underlie physical traits of forensic interest using genetic tools. *Forensic Sci Int Genet*, 2007, **1**(2): 100–104.

Rebbeck, T.R., et al., P gene as an inherited biomarker of human eye color. *Cancer Epidemiol Biomarkers Prev*, 2002, **11**(8): 782–784.

Rees, J.L., Genetics of hair and skin color. *Annu Rev Genet*, 2003, **37**: 67–90.

Reich, D.E., et al., Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet*, 2002, **32**(1): 135–142.

Riccardi, L.N., et al., Development of a tetraplex PCR assay for CYP2D6 genotyping in degraded DNA samples. *J Forensic Sci*, 2013, **59**(3): 690–695.

Rockenbauer, E., et al., Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. *Forensic Sci Int Genet*, 2014, **8**(1): 68–72.

Rodriguez-Calvo, M.S., et al., Population data on the loci LDLR, GYPA, HBGG, D7S8, and GC in three southwest European populations. *J Forensic Sci*, 1996, **41**(2): 291–296.

Rolf, B., N. Bulander, and P. Wiegand, Insertion-/deletion polymorphisms close to the repeat region of STR loci can cause discordant genotypes with different STR kits. *Forensic Sci Int Genet*, 2011, **5**(4): 339–341.

Romanini, C., et al., Typing short amplicon binary polymorphisms: Supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains. *Forensic Sci Int Genet*, 2012, **6**(4): 469–476.

Romero, R.L., et al., The applicability of formalin-fixed and formalin fixed paraffin embedded tissues in forensic DNA analysis. *J Forensic Sci*, 1997, **42**(4): 708–714.

Ronaghi, M., M. Uhlen, and P. Nyren, A sequencing method based on real-time pyrophosphate. *Science*, 1998, **281**(5375): 363, 365.

Ronaghi, M., et al., Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, 1996, **242**(1): 84–89.

Ross, P., et al., High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat Biotechnol*, 1998, **16**(13): 1347–1351.

Ruiz, Y., et al., Analysis of the SNPforID 52-plex markers in four Native American populations from Venezuela. *Forensic Sci Int Genet*, 2012, **6**(5): e142–e145.

Ruiz, Y., et al., Further development of forensic eye color predictive tests. *Forensic Sci Int Genet*, 2013, **7**(1): 28–40.

Sachidanandam, R., et al., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 2001, **409**(6822): 928–933.

Saiki, R.K., et al., Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature*, 1986, **324**(6093): 163–166.

Saiki, R.K., et al., Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc Natl Acad Sci U S A*, 1989, **86**(16): 6230–6234.

Sanchez, J.J., et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*, 2006, **27**(9): 1713–1724.

Santos, C., et al., A study of East Timor variability using the SNPforID 52-plex SNP panel. *Forensic Sci Int Genet*, 2011, **5**(1): e25–e26.

Sato, Y., et al., HLA typing of aortic tissues from unidentified bodies using hot start polymerase chain reaction-sequence specific primers. *Leg Med (Tokyo)*, 2003, **5**(Suppl 1): S191–S193.

Sauer, S., et al., A novel procedure for efficient genotyping of single nucleotide polymorphisms. *Nucleic Acids Res*, 2000, **28**(5): E13.

Schmid, D., B. Bayer, and K. Anslinger, Comparison of telogen hair analyses: GenRES MPX-2SP kit versus genRES MPX-SP1 and genRES MPX-SP2 kits. *Forensic Sci Int Genet*, 2008, **3**(1): 22–26.

Schneider, P.M. and C. Rittner, Experience with the PCR-based HLA-DQ alpha DNA typing system in routine forensic casework. *Int J Legal Med*, 1993, **105**(5): 295–299.

Scholl, S., et al., Navajo, Pueblo, and Sioux population data on the loci HLA-DQA1, LDLR, GYPA, HBGG, D7S8, Gc, and D1S80. *J Forensic Sci*, 1996, **41**(1): 47–51.

Senge, T., et al., STRs, mini STRs and SNPs—A comparative study for typing degraded DNA. *Leg Med*, 2011, **13**(2): 68–74.

Sharafi Farzad, M., et al., Analysis of 49 autosomal SNPs in three ethnic groups from Iran: Persians, Lurs and Kurds. *Forensic Sci Int Genet*, 2013, **7**(4): 471–473.

Shen, M., Y. Shi, and P. Xiang, CYP3A4 and CYP2C19 genetic polymorphisms and zolpidem metabolism in the Chinese Han population: A pilot study. *Forensic Sci Int*, 2013, **227**(1–3): 77–81.

Shi, M., et al., 6 Y-SNP typing of China and Korean samples using primer extension and DHPLC. *J Forensic Sci*, 2007, **52**(1): 235–236.

Shi, Y., et al., Analysis of 50 SNPs in CYP2D6, CYP2C19, CYP2C9, CYP3A4 and CYP1A2 by MALDI-TOF mass spectrometry in Chinese Han population. *Forensic Sci Int*, 2011, **207**(1–3): 183–187.

Shriver, M.D. and R.A. Kittles, Genetic ancestry and the search for personalized genetic histories. *Nat Rev Genet*, 2004, **5**(8): 611–618.

Siebert, P.D. and M. Fukuda, Molecular cloning of a human glycophorin B cDNA: Nucleotide sequence and genomic relationship to glycophorin A. *Proc Natl Acad Sci U S A*, 1987, **84**(19): 6735–6739.

Slightom, J.L., A.E. Blechl, and O. Smithies, Human fetal G gamma- and A gamma-globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell*, 1980, **21**(3): 627–638.

Soares-Vieira, J.A., et al., Brazilian population data on the polymerase chain reaction-based loci LDLR, GYPA, HBGG, D7S8, and Gc. *Am J Forensic Med Pathol*, 2003, **24**(3): 283–287.

Sobrino, B., M. Brion, and A. Carracedo, SNPs in forensic genetics: A review on SNP typing methodologies. *Forensic Sci Int*, 2005, **154**(2–3): 181–194.

Sokolov, B.P., Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Res*, 1990, **18**(12): 3671.

Spichenok, O., et al., Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Sci Int Genet*, 2011, **5**(5): 472–478.

Spinella, A., et al., Italian population allele and genotype frequencies for the AmpliType PM and the HLA-DQ-alpha loci. *J Forensic Sci*, 1997, **42**(3): 514–518.

Syvanen, A.C., et al., A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics*, 1990, **8**(4): 684–692.

Tahir, M.A., et al., DNA typing of samples for polymarker, DQA1, and nine STR loci from a human body exhumed after 27 years. *J Forensic Sci*, 2000, **45**(4): 902–907.

Takada, Y., et al., Application of SNPs in forensic casework: Identification of pathological and autoptical specimens due to sample mix-up. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S196–S197.

Taylor, M.S., A. Challed-Spong, and E.A. Johnson, Co-amplification of the amelogenin and HLA DQ alpha genes: Optimization and validation. *J Forensic Sci*, 1997, **42**(1): 130–136.

Templeton, J.E., et al., DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. *Investig Genet*, 2013, **4**(1): 26.

Tomas, C., C. Borsting, and N. Morling, A 48-plex autosomal SNP GenPlex assay for human individualization and relationship testing. *Methods Mol Biol*, 2012, **830**: 73–85.

Tomas, C., et al., Typing of 48 autosomal SNPs and amelogenin with GenPlex SNP genotyping system in forensic genetics. *Forensic Sci Int Genet*, 2008, **3**(1): 1–6.

Tomas, C., et al., Autosomal SNP typing of forensic samples with the GenPlex HID system: Results of a collaborative study. *Forensic Sci Int Genet*, 2011, **5**(5): 369–375.

Tomas, C., et al., Analysis of 49 autosomal SNPs in an Iraqi population. *Forensic Sci Int Genet*, 2013, **7**(1): 198–199.

Tsongalis, G.J. and M.M. Berman, Application of forensic identity testing in a clinical setting. Specimen identification. *Diagn Mol Pathol*, 1997, **6**(2): 111–114.

Tyagi, S. and F.R. Kramer, Molecular beacons: Probes that fluoresce upon hybridization. *Nat Biotechnol*, 1996, **14**(3): 303–308.

Tyagi, S., D.P. Bratu, and F.R. Kramer, Multicolor molecular beacons for allele discrimination. *Nat Biotechnol*, 1998, **16**(1): 49–53.

Valenzuela, R.K., et al., Predicting phenotype from genotype: Normal pigmentation. *J Forensic Sci*, 2010, **55**(2): 315–322.

Van Geystelen, A., et al., *In silico* detection of phylogenetic informative Y-chromosomal single nucleotide polymorphisms from whole genome sequencing data. *Electrophoresis*, 2014 (in press) 10.1002/elps.201300459.

Van Neste, C., et al., Forensic STR analysis using massive parallel sequencing. *Forensic Sci Int Genet*, 2012, **6**(6): 810–818.

Van Neste, C., et al., My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci Int Genet*, 2014, **9**: 1–8.

Vu, N.T., A.K. Chaturvedi, and D.V. Canfield, Genotyping for DQA1 and PM loci in urine using PCR-based amplification: Effects of sample volume, storage temperature, preservatives, and aging on DNA extraction and typing. *Forensic Sci Int*, 1999, **102**(1): 23–34.

Vural, B., et al., Turkish population data on the HLA-DO alpha, LDLR, GYPA, HBGG, D7s8, and GC loci. *Int J Legal Med*, 1998, **111**(1): 43–45.

Walsh, S., et al., Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet*, 2011, **5**(5): 464–471.

Walsh, S., et al., IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet*, 2011, **5**(3): 170–180.

Walsh, S., et al., DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet*, 2012, **6**(3): 330–340.

Walsh, S., et al., The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*, 2013, **7**(1): 98–115.

Watanabe, Y., et al., Japanese population DNA typing data for the loci LDLR, GYPA, HBGG, D7S8, and GC. *J Forensic Sci*, 1997, **42**(5): 911–913.

Weber-Lehmann, J., et al., Finding the needle in the haystack: Differentiating “identical” twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Sci Int Genet*, 2014, **9**: 42–46.

Wei, Y.L., et al., Forensic identification using a multiplex assay of 47 SNPs. *J Forensic Sci*, 2012, **57**(6): 1448–1456.

Westen, A.A., et al., Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. *Forensic Sci Int Genet*, 2009, **3**(4): 233–241.

Wilson, R.B., et al., Guidelines for internal validation of the HLA-DQ alpha DNA typing system. *Forensic Sci Int*, 1994, **66**(1): 9–22.

Woo, K.M. and B. Budowle, Korean population data on the PCR-based loci LDLR, GYPA, HBGG, D7S8, Gc, HLA-DQA1, and D1S80. *J Forensic Sci*, 1995, **40**(4): 645–648.

Yamamoto, T., et al., The human LDL receptor: A cysteine-rich protein with multiple Alu sequences in its mRNA. *Cell*, 1984, **39**(1): 27–38.

Yang, F., et al., Human group-specific component (Gc) is a member of the albumin family. *Proc Natl Acad Sci U S A*, 1985, **82**(23): 7994–7998.

Yang, H.C., et al., Integrative analysis of single nucleotide polymorphisms and gene expression efficiently distinguishes samples from closely related ethnic populations. *BMC Genomics*, 2012, **13**: 346.

Zaumsegel, D., M.A. Rothschild, and P.M. Schneider, A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. *Forensic Sci Int Genet*, 2013, **7**(2): 305–312.

Zeng, Z., et al., Genome-wide screen for individual identification SNPs (IISNPs) and the confirmation of six of them in Han Chinese with pyrosequencing technology. *J Forensic Sci*, 2010, **55**(4): 901–907.
Zeng, Z., et al., Evaluation of 96 SNPs in 14 populations for worldwide individual identification. *J Forensic Sci*, 2012, **57**(4): 1031–1035.
Zha, L., et al., Exploring of tri-allelic SNPs using pyrosequencing and the SNaPshot methods for forensic application. *Electrophoresis*, 2012, **33**(5): 841–848.
Zhao, S.M., S.H. Zhang, and C.T. Li, InDel-typer30: A multiplex PCR system for DNA identification among five Chinese populations. *Fa Yi Xue Za Zhi*, 2010, **26**(5): 343–348+356.
Zidkova, A., et al., Application of the new insertion-deletion polymorphism kit for forensic identification and parentage testing on the Czech population. *Int J Legal Med*, 2013, **127**(1): 7–10.

23

Mitochondrial DNA Profiling

Forensic mitochondrial DNA (mtDNA) analysis is an important tool for human identification and is especially useful for identifying victims such as missing persons and individuals in mass fatality cases. Because mtDNA is maternally inherited, the mtDNA profiles of these individuals can be compared to those of their maternal relatives, and thus these individuals can be identified. Additionally, cells contain a much higher copy number of the mtDNA genome than that of the nuclear genome. Therefore, mtDNA testing is frequently used to analyze evidence samples, such as hair shafts, that contain low amounts of nuclear DNA. Furthermore, buried bones and decomposed tissues, in which nuclear DNA may be degraded, can be tested with mtDNA analysis.

23.1 Human Mitochondrial Genome

Mitochondria are cellular organelles that serve as the energy-generating components of cells (Figure 23.1). Each cell contains hundreds of mitochondria, which have their own extrachromosomal genomes separated from the nuclear genomes. Although each human mitochondrion contains multiple copies of the mtDNA genome, the exact copy number varies for each cell. However, it is estimated that hundreds of copies of the mtDNA genome exist in most cells.

23.1.1 Genetic Contents of Mitochondrial Organelle Genomes

The human mitochondrial genome was first sequenced by Fred Sanger's laboratory at Cambridge University and was published in 1981; it is known as the *Cambridge reference sequence* (CRS). The sequence was largely derived from a placental sample from an individual of European descent and also partially from HeLa cells (a cell line derived from cervical cancer cells), as well as from bovine cells. It was later discovered, by resequencing the original mtDNA sample, that the CRS contains substitution errors at 10 nucleotide positions. The *revised Cambridge reference sequence* (rCRS) was published in 1999 and presented corrections to these substitution errors. Additionally, CRS contains a cytosine dimer at nucleotide positions 3106 and 3107, which is in fact a single cytosine nucleotide. This error is not corrected in the rCRS in order to retain the original nucleotide numbering system of CRS and thus to avoid inconsistency with the previous literature. The human mitochondrial DNA genome is circular with no beginning and end, which can make sequence comparison a potential problem. Therefore, the rCRS, with its nucleotide numbering system, is used as a reference when aligning with other mitochondrial sequences for comparison purposes.

Organelle genomes are usually much smaller than their nuclear counterparts. The human mtDNA genome consists of 16,569 base pairs (bp) containing 37 genes (Figure 23.2). Thirteen of

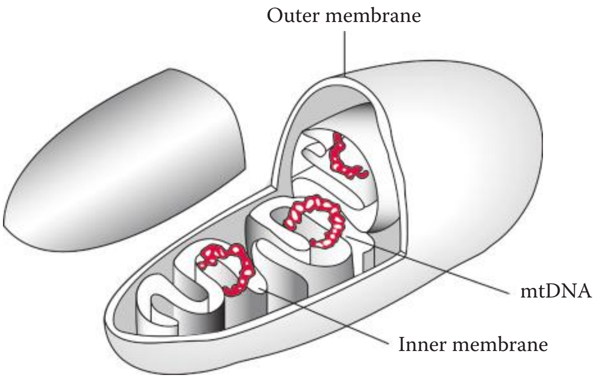
Outer membrane
mtDNA
Inner membrane

Figure 23.1 Mitochondrion. mtDNA, mitochondrial DNA. (© Richard C. Li.)

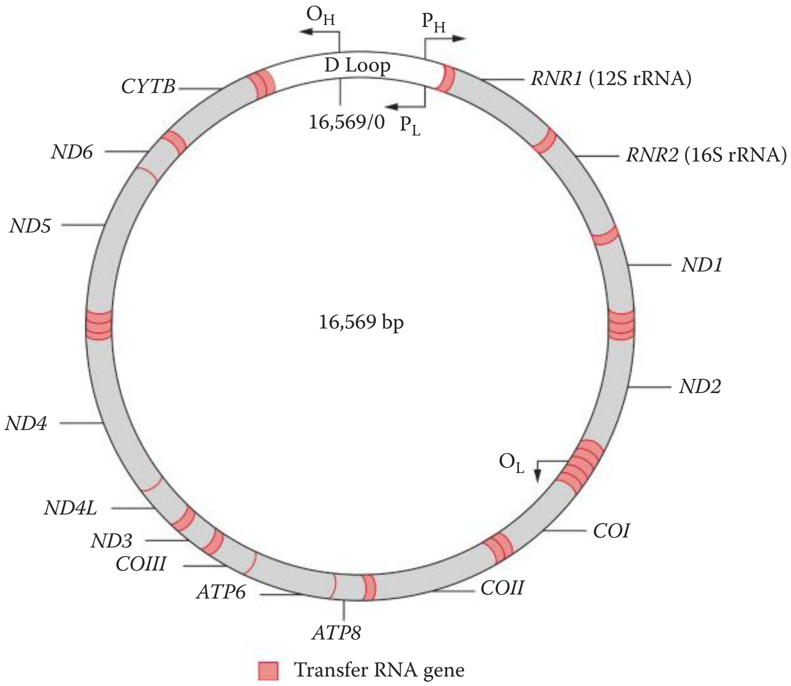
O_H
P_H
D Loop
16,569/0
P_L
CYTB
RNR1 (12S rRNA)
ND6
RNR2 (16S rRNA)
ND5
ND1
16,569 bp
ND2
ND4
O_L
ND4L
COI
ND3
COIII
COII
ATP6
ATP8
🟥 Transfer RNA gene

Figure 23.2 Human circular mitochondrial genome. The transcription direction for the H (heavy) and L (light) strands are indicated by arrows (P_H, P_L, respectively). The origins of replication are labeled O_H for the heavy strand and O_L for the light strand. The mitochondrial DNA genome encodes genes. ND, NADH coenzyme Q oxidoreductase complex; CO, cytochrome c oxidase complex; CYTB, cytochrome b; ATP, ATP synthase; rRNA, ribosomal RNA. Transfer RNA genes are shown as indicated. (© Richard C. Li.)

these genes code for proteins involved in the respiratory complex, a main energy-generating component in mitochondria. The other 24 specify noncoding RNA molecules required for the expression of the mitochondrial genome. The genes in the human mitochondrial genome are much more closely packed than in the nuclear genome and contain no introns. A *control region*, also known as a *displacement loop* (*D loop*), contains the origin of replication for one of the mtDNA strands but does not code for any gene products (Figure 23.2). An asymmetric distribution of

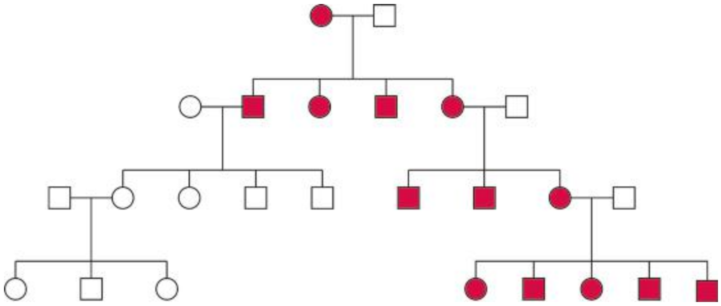

Figure 23.3 Pedigree of a human family showing inheritance of mtDNA. Females and males are denoted by circles and squares, respectively. Red symbols indicate individuals who inherited the same mtDNA.

nucleotides gives rise to light (L) and heavy (H) strands. The H strand contains a greater number of guanine nucleotides and a higher molecular weight in comparison to the L strand.

23.1.2 Maternal Inheritance of mtDNA

Maternal inheritance is typically observed for the mtDNA genome (Figure 23.3), which is inherited differently from nuclear genes. The inheritance of the mtDNA genome does not obey the rules of Mendelian inheritance and is thus called non-Mendelian inheritance.

The mitochondria of a spermatozoon are located in the midpiece (Chapter 14). At conception, only the head portion of a spermatozoon (containing a nucleus but no mitochondria) enters the egg. The fertilized egg contains the maternal mitochondria, which are transmitted to the progeny. Occasionally, the paternal mitochondria can enter the cell. However, paternal mitochondria in the spermatozoon that enter the egg are usually destroyed by the egg cell after fertilization (Figure 23.4). Therefore, the coinheritance of maternal and paternal mtDNA in a single individual is extremely rare in humans. The mtDNA sequence is identical for relatives within the same maternal lineage, a property that is useful when identifying individuals by comparing their mtDNA with that of maternal relatives.

Homologous DNA recombination (Chapter 25) has not been observed in the mtDNA genome. Thus, an mtDNA profile, also referred to as the *mitotype*, is considered a haplotype treated as a single locus. The mitochondrial genome has a higher mutation rate (up to 10 times higher) than its nuclear counterpart. The presence of mutations can be problematic in victim identification when comparing the mtDNA profiles of a victim with the relatives of the victim.

23.2 mtDNA Polymorphic Regions

23.2.1 Hypervariable Regions

The most polymorphic region of mtDNA is located within the D-loop (Figure 23.5). The three *hypervariable regions* in the D-loop are designated hypervariable region I (HV1: 16,024–16,365; 342 bp), hypervariable region II (HV2: 73–340; 268 bp), and hypervariable region III (HV3: 438–574; 137 bp). The most common polymorphic regions of the human mtDNA genome analyzed for forensic purposes are HV1 and HV2.

23.2.2 Heteroplasmy

Heteroplasmy occurs when an individual carries more than one mtDNA haplotype. Heteroplasmy may be observed with one kind of tissue and may be absent in other kinds of tissues; for example,

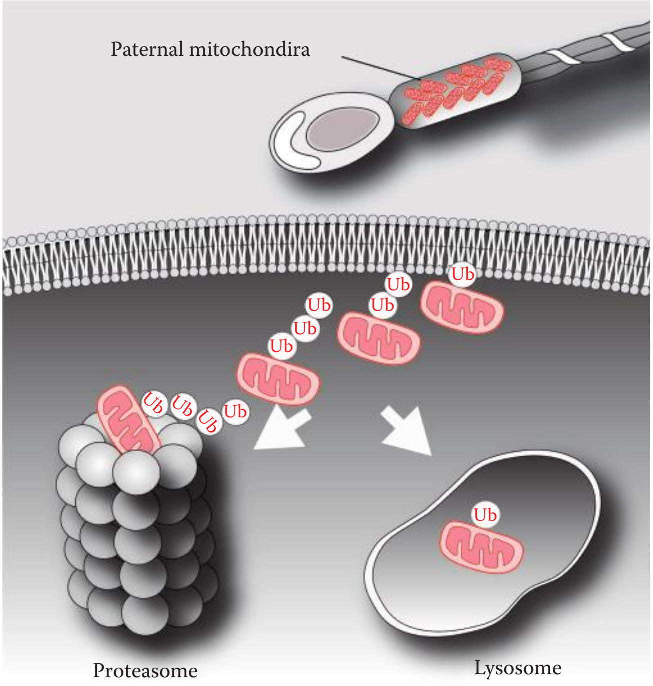

Figure 23.4 A model of uniparental mtDNA inheritance in humans. Paternal mitochondria occasionally enter the egg cell. Paternal mitochondria inside fertilized eggs are tagged by ubiquitin protein (Ub). It is proposed that the ubiquitination of sperm mitochondria leads to the degradation of paternal mitochondria in fertilized eggs. Tagged Ub can be recognized by proteasomes and lysosomes, which are cellular degradation machineries. A polyubiquitin chain with at least four ubiquitin units is needed to be recognized by the proteasomes. Ubiquitin with less than four ubiquitin units can be processed by the lysosomes. (© Richard C. Li.)

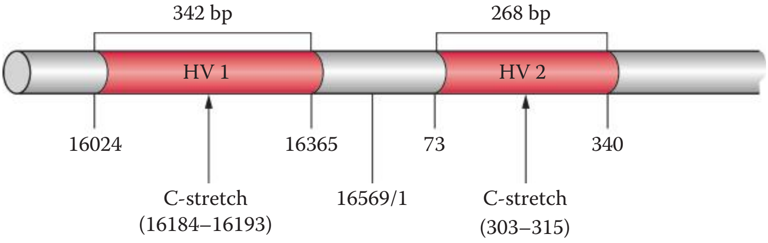

Figure 23.5 Hypervariable regions of the D-loop in mtDNA (with nucleotide positions). (© Richard C. Li.)

mtDNA heteroplasmy is commonly observed in hair samples (Chapter 4). Additionally, an individual may exhibit one mitotype in one tissue and a different mitotype in another. Thus, it is necessary to obtain and process additional samples to confirm the heteroplasmy when it is observed in a questioned sample but not in a known sample or vice versa. The two types of heteroplasmies are sequence and length heteroplasmy.

Sequence heteroplasmy is defined as the presence of two different nucleotides at a single position shown as overlapping peaks in a sequence electropherogram (Figure 23.6). Heteroplasmy usually occurs at one position, but on rare occasions it can be observed at more than one position. Hot spots for heteroplasmy have been documented at both HV1 and HV2 regions. Heteroplasmy may complicate the interpretation of mtDNA results, but its presence can also

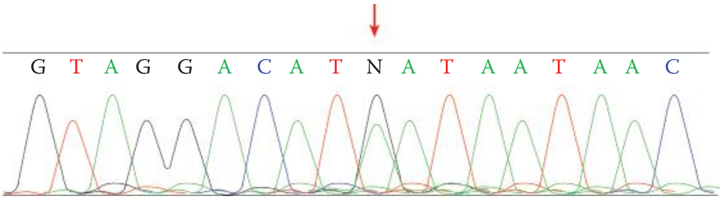

Figure 23.6 Electropherogram showing mtDNA sequence heteroplasmy at position 234R (A/G) as indicated by an arrow. N, unresolved sequence. (© Richard C. Li.)

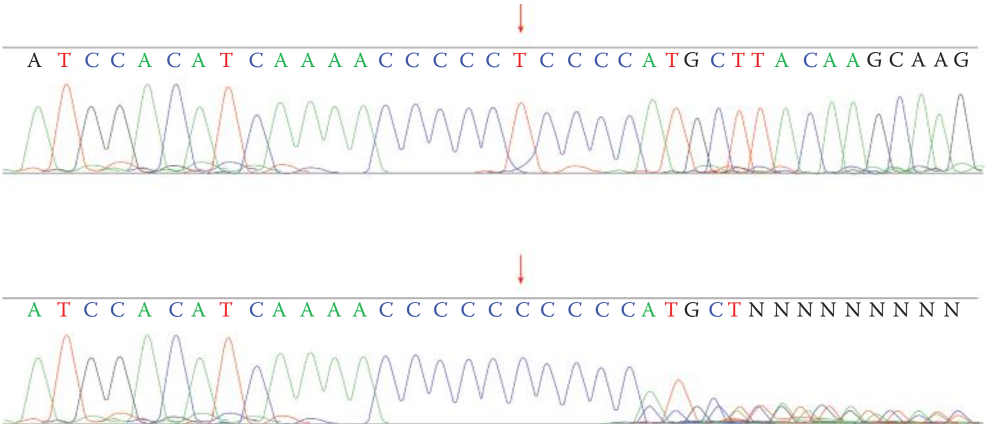

Figure 23.7 Length heteroplasmy. Electropherogram showing mtDNA length heteroplasmy at the C stretch of the HV1 region where position 16,189 is a T (top) and a C (bottom) as indicated by an arrow. N, unresolved sequence. (© Richard C. Li.)

improve the strength of a match. Both HV1 and HV2 of the human mtDNA D-loop region contain homopolymeric cytosine sequences known as *C stretches*. The HVI region contains a C stretch between nucleotide positions 16,184 and 16,193, interrupted by a thymine at position 16,189. If a base transition from T to C occurs at position 16,189 (a variant present in approximately 20% of the population), it results in an uninterrupted C stretch. A similar C stretch resides between positions 303 and 315 of the HV2 region.

Length heteroplasmies are often observed at the uninterrupted C stretches in sequencing, in which sequencing products with various lengths of polymeric cytosine residues are present. As a result, sequences downstream from the C stretch cannot be resolved (Figure 23.7). It is not clear whether the length heteroplasmy is due to replication slippage at the C stretches or results from the presence of a mixture of length variants in the cells. If length heteroplasmy occurs, alternative sequencing primers that anneal at the downstream of C stretches can be used to obtain the downstream sequences of the C stretches.

23.3 Forensic mtDNA Testing

23.3.1 General Considerations

mtDNA analysis is often used on samples derived from skeletal or decomposed remains. The surface of the sample should be cleaned to remove any adhering debris or contaminants. Bones and teeth are pulverized to facilitate extraction of the mtDNA (Chapter 5). Duplicate extractions (e.g., two sections of a single hair) are recommended if sufficient sample material is available.

mtDNA is extracted using a similar method to nuclear DNA (nuclear DNA is coextracted with mtDNA). The amount of mtDNA can therefore be estimated from the quantity of nuclear DNA obtained. Alternatively, mtDNA-specific quantization methods using real-time PCR (Chapter 6) can also be used to directly obtain measurements of mtDNA extracted.

For mtDNA sequencing, the analysis of both strands of the mtDNA in a given region must be performed to ensure accuracy. Due to the high sensitivity of mtDNA analysis, it is essential to minimize risks of contamination during the procedure. Contamination must be strictly monitored using proper controls such as extraction reagent blanks (Section 7.5.3) and amplification negative controls (samples containing all reagents except DNA template).

Finally, a positive control must also be used to monitor the success of the analysis. It should be introduced at the amplification step and remain through the sequencing process. A positive control consists of a DNA template of known sequence, such as DNA purified from the HL60 cell line.

23.3.2 mtDNA Screen Assay

One example of the assay for screening mtDNA variations is the allele-specific oligonucleotide (ASO) assay. It allows initial screening of mtDNA sequence polymorphisms and has the potential to reduce the number of samples required for mtDNA sequencing. This method is also useful for excluding or eliminating suspects from a case. However, HV1 and HV2 sequencing should be performed to obtain complete sequence information for the targeted HV regions to confirm a match.

The commercial Linear Array™ mtDNA HV1/HV2 region sequence typing kit (Roche Applied Sciences, Indianapolis) utilizes reverse ASO configuration with a panel of immobilized ASO probes that detect common polymorphic sites (Figures 23.8 and 23.9). The mtDNA is amplified at both HV1 (444 bp amplicon) and HV2 (415 bp amplicon) regions and the forward primers are biotin labeled at the 5′ ends of the oligonucleotides. Thus, the amplified PCR product (amplicon) is biotinylated. A horseradish peroxidase–conjugated streptavidin complex then binds to the biotinylated amplicon (see Chapter 9). Finally, colorimetric detection is carried out with tetramethylbenzidine (TMB) as the substrate to produce a colored precipitate at the designated location. The typing kit detects sequence variations in 19 positions within the HV1 and HV2 regions.

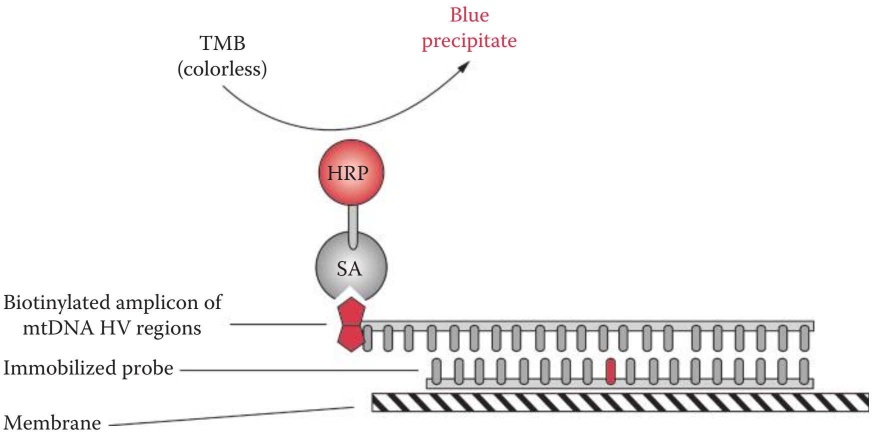

Figure 23.8 Reverse blot assay employed in mtDNA screen. A probe is immobilized onto a solid-phase membrane, and then hybridized with a biotinylated amplicon of the mtDNA HV sequences. Hybridization is detected by a streptavidin (SA) and horseradish peroxidase (HRP) conjugate. A colorimetric reaction is catalyzed by HRP using tetramethylbenzidine (TMB) as a substrate. (© Richard C. Li.)

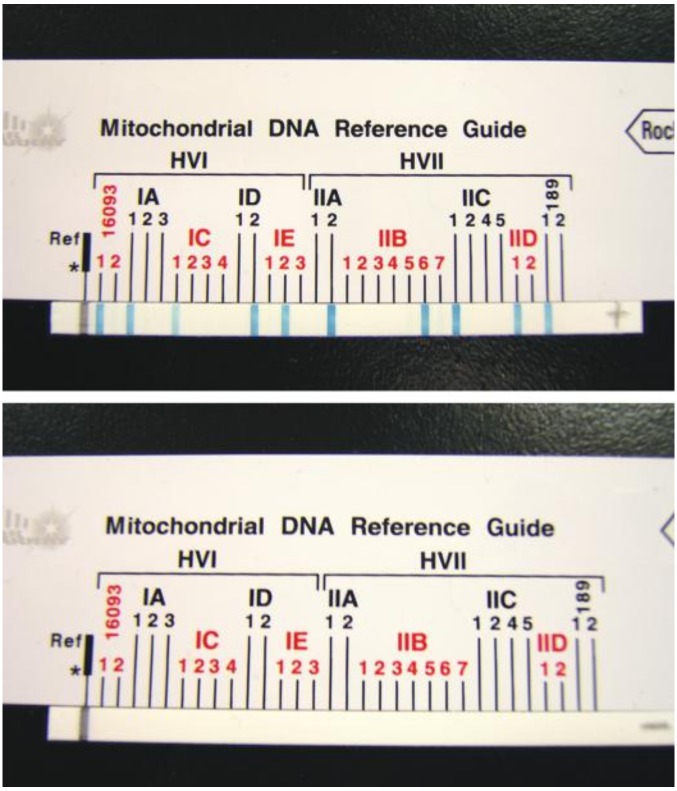

Figure 23.9 Linear array mtDNA assay results (top) and negative control (bottom). (© Richard C. Li.)

23.3.3 mtDNA Sequencing

To sequence a specific region of mtDNA, a combination of PCR amplification and DNA sequencing techniques is utilized to reduce the time and labor needed to obtain DNA sequences from genomic DNA templates. mtDNA sequencing usually consists of (1) PCR amplification, (2) DNA sequencing reactions, (3) separation using electrophoresis, and (4) data collection and sequence analysis (Figure 23.10). Chapter 25 describes the evaluation of the strength of the results via statistical analysis.

23.3.3.1 PCR Amplification

The extracted DNA samples must be amplified to yield sufficient quantities of template for sequencing reactions. PCR amplification of all or a part of the D-loop region can be carried out with various primer sets. If a sample contains intact mtDNA, the HV1 and HV2 regions can be amplified as two amplicons, each of about 350–400 bp in length. If an mtDNA sample is fragmented due to degradation, the hypervariable regions can be amplified as smaller amplicons. PCR amplification of mtDNA is usually done in 34–38 cycles. Protocols for highly degraded DNA specimens sometimes require 42 cycles. The use of higher PCR cycle numbers can improve the yield of the amplicons.

Following mtDNA amplification, a purification step is necessary to remove excess primers and deoxynucleotide triphosphates (dNTPs). This step can be performed using filtration devices such as a Microcon® to remove small molecules from the sample or using nuclease digestion with shrimp alkaline phosphatase or exonuclease I to degrade remaining primers and dNTPs. The concentration of the amplicons is important for an optimal sequencing reaction in the next phase of mtDNA sequencing. The quality and quantity of the mtDNA amplicon must be evaluated to confirm the presence or absence of amplicons and their concentrations. This can be done

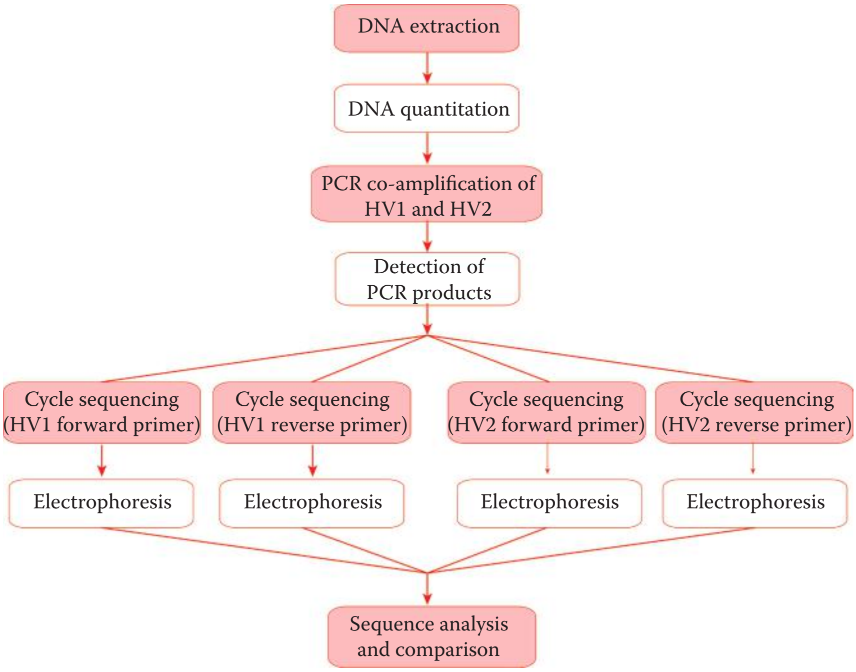
DNA extraction
DNA quantitation
PCR co-amplification of HV1 and HV2
Detection of PCR products
Cycle sequencing (HV1 forward primer)
Cycle sequencing (HV1 reverse primer)
Cycle sequencing (HV2 forward primer)
Cycle sequencing (HV2 reverse primer)
Electrophoresis
Electrophoresis
Electrophoresis
Electrophoresis
Sequence analysis and comparison

Figure 23.10 Sanger sequencing work flow for the mtDNA HV regions. (© Richard C. Li.)

using an agarose yield gel to visualize the amplicons of the sample or via capillary electrophoresis using a modular microfluidic chip device (Figure 8.12), a more informative method, for quantifying amplicons.

23.3.3.2 DNA Sequencing Reactions

The best-known DNA sequencing techniques are the *chain-termination* method and the *chemical degradation* method developed by Sanger and Gilbert (who shared the Nobel Prize for their work), respectively, in 1977. Over the years, the chain-termination method has become more common because it is suitable for automation and does not require the toxic chemicals necessary for the chemical degradation method.

23.3.3.2.1 Chain-Termination or Sanger Method

Using the chain-termination method, an oligonucleotide primer that can anneal to a single-stranded DNA template is utilized. A sequencing reaction contains DNA polymerase and the four dNTPs in order to carry out extension. The reaction also contains small quantities of dideoxynucleotide triphosphates (ddNTPs, Figure 23.11). Thus, a sequencing reaction involves a combination of extension and termination of the chain (Figure 23.12). If a ddNTP molecule is incorporated into a growing DNA chain, the absence of a 3′ OH group in the ddNTP molecule prevents the formation of a phosphodiester bond and thus disrupts the extension of the oligonucleotide chain.

The ratio of ddNTPs to dNTPs has been optimized to result in a collection of DNA fragments varying in length by one nucleotide from the primer length to the full length of the sequencing reaction product. As a result, the products of the sequencing reaction consist of a pool of various lengths of oligonucleotide chains terminated by ddNTPs. By using the four different ddNTPs, populations of DNA fragments are generated that terminate at positions occupied by every A, C, G, or T in the template strand.

The sequencing product of chain termination can be labeled with the *dye-terminator* system, in which the terminator is labeled, or with the *dye-primer* system, in which the primer is labeled.

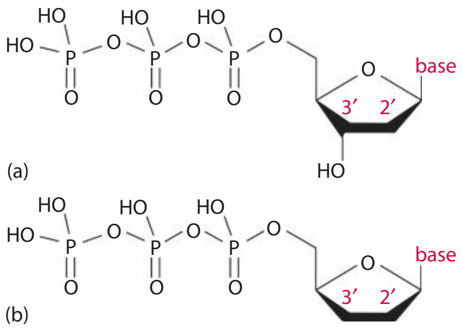
HO
HO
P
O
O
HO
P
O
O
HO
P
O
O
O
base
3′
2′
HO
(a)
(b)

Figure 23.11 Chemical structures of dNTP and ddNTP: (a) dNTP; (b) ddNTP. Both hydroxyl groups attached to the 2′ and 3′ carbons of ribose are replaced by hydrogens.

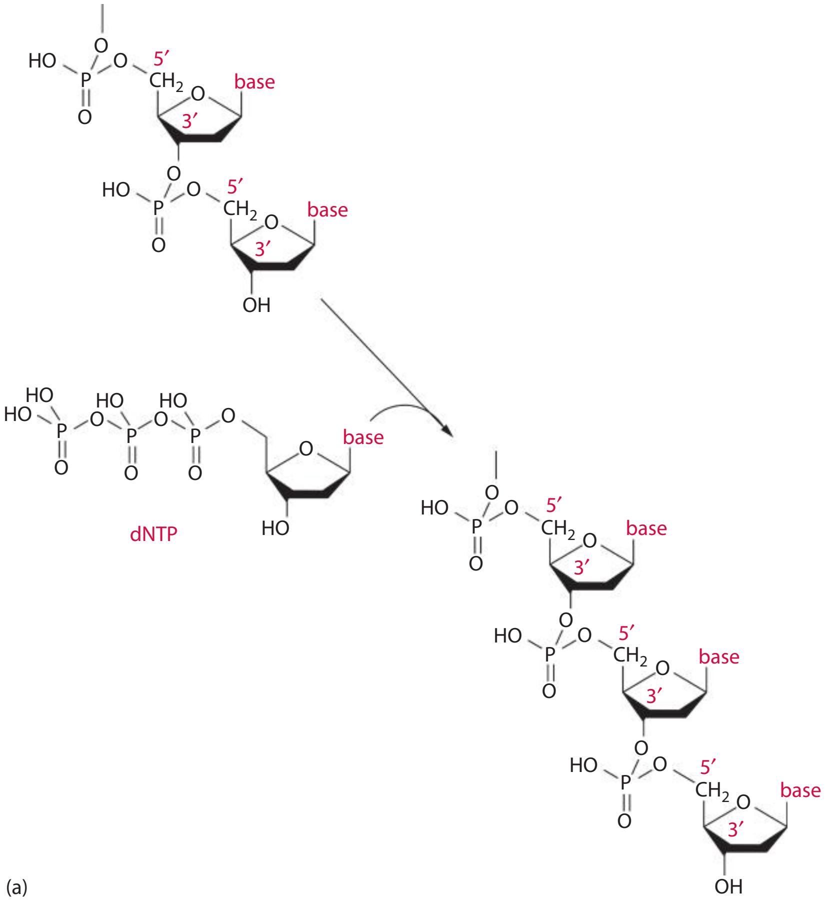
HO
P
O
O
O
5′
CH_2
base
3′
OH
dNTP
HO
(a)

Figure 23.12 Competition between extension and termination. (a) Growing chain in extension.

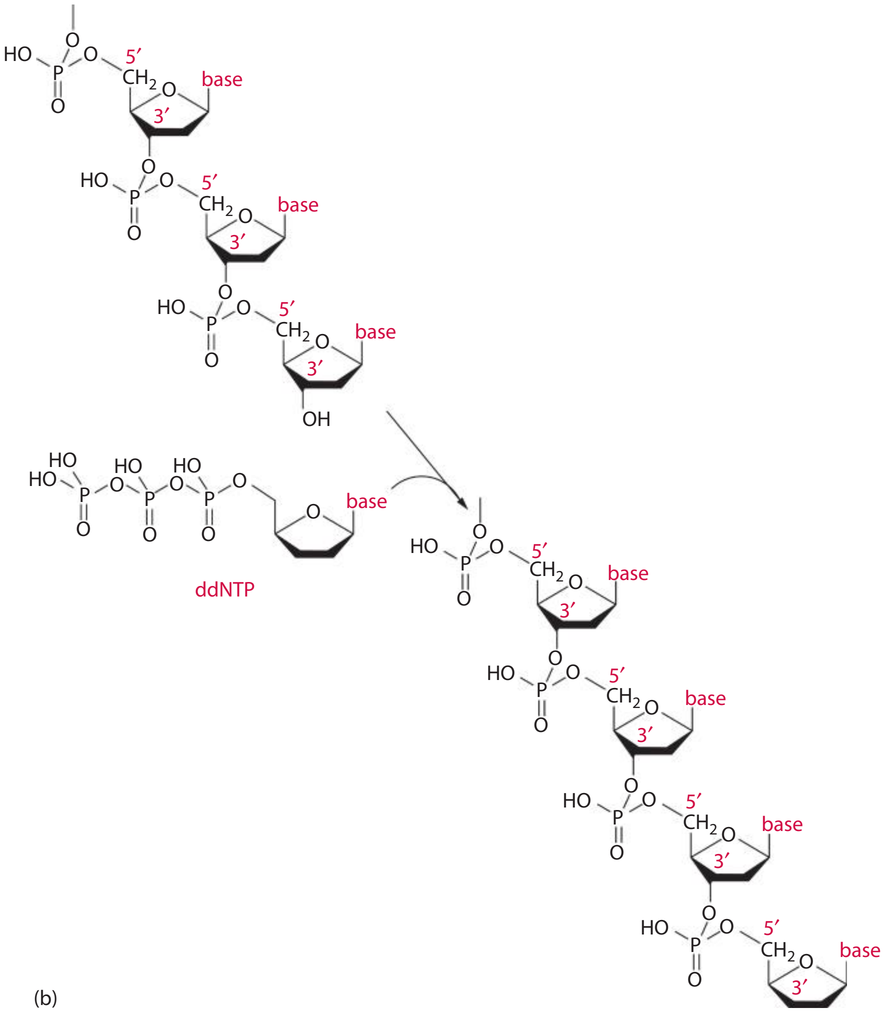

Figure 23.12 (Continued) (b) DNA strand with labeled ddNTP incorporated into a growing DNA chain that interrupts the incorporation of new nucleotides. The ddNTP usually is labeled with a fluorescent dye used for detection.

The dye-terminator system is commonly used for mtDNA sequencing in forensic laboratories. With the dye-terminator system, the ddNTPs are labeled with four different fluorescent dyes, each with a distinct spectrum. Thus, the sequencing with all four ddNTPs can be carried out in a single reaction. The labeled products of sequencing reactions are then resolved during electrophoresis and the sequencing data can be collected (Section 23.3.3.).

23.3.3.2.2 Cycle Sequencing

The chain-termination reaction is carried out using a *cycle sequencing* technique commonly used in forensic laboratories for mtDNA sequencing. Cycle sequencing, developed in the late 1980s, utilizes thermal cycling to generate a single-stranded template for chain-termination sequencing reactions. The application of thermal cycling in a sequencing reaction greatly increases the signal intensity and thus the sensitivity of the sequencing.

The sequencing reactions are carried out with multiple rounds of thermal cycling. Each cycle consists of three steps: denaturation of the double-stranded DNA template, annealing of a sequencing primer to its target sequence, and the extension of the annealed primer by DNA polymerase. Cycle sequencing utilizes only a single primer per reaction. During the extension of cycle sequencing, the extension of the strand is terminated with the incorporation of a ddNTP (Figure 23.13). The resulting partially double-stranded hybrid, consisting of the full-length template strand and its complementary chain-terminated product, is denatured during the first step of the next cycle, thereby liberating the template strand for another round of annealing, extension, and termination.

23.3.3.3 Electrophoresis, Sequence Analysis, and Mitotype Designations

The cycle sequencing products can be separated using electrophoresis in a 4% polyacrylamide denatured gel or a POP-6 polymer (Applied Biosystems) as the matrix for capillary electrophoresis (Figure 23.14). Following data collection, sequence data analysis can be performed with the Sequencher™ software (Gene Codes Corporation, Ann Arbor). Figure 23.15 shows sequence data.

Sequencing of a region of the mtDNA genome should be performed twice. Additionally, both strands of a region of the mtDNA genome should be sequenced to reduce ambiguities in sequence determination. The sequences of evidence samples and reference samples such as that of the victim or suspect can be compared. The nomenclature used in reporting should be compatible with International Union of Pure and Applied Chemistry (IUPAC) codes.

23.3.3.3.1 Reporting Format

The rCRS is used as a reference standard to facilitate the designation of mitotypes. For reporting purposes, sequence differences relative to the rCRS are listed in data format. When a difference between an individual's sequence and that of the rCRS sequence is observed, only the position (designated by a number) and the nucleotide differing from the reference standard are recorded. In this format, nucleotides identical to the rCRS are not listed. For example, at position 228 (HV2), the rCRS has a G. If a mitotype carries an A at position 228, the individual's mtDNA sequence is described as 228A. If an unresolved sequence ambiguity is observed at a position,

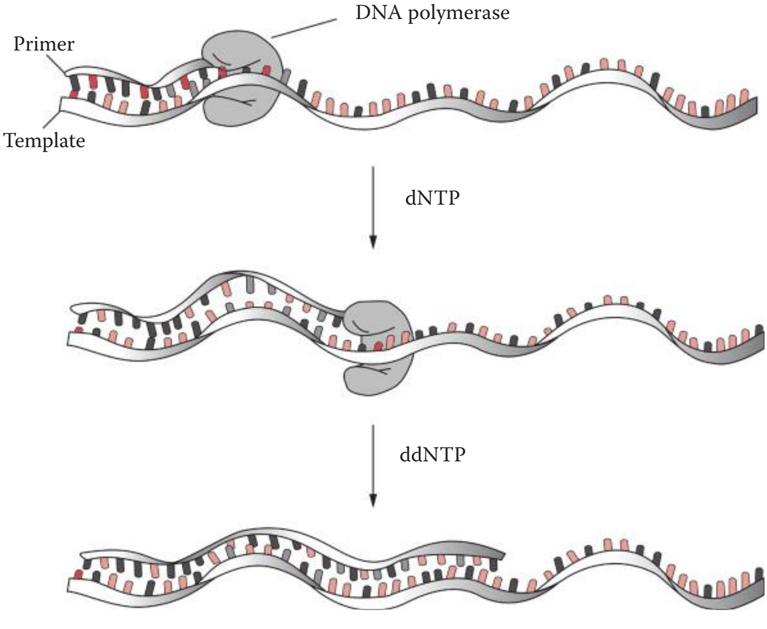

Figure 23.13 Cycle sequencing reaction. The reaction requires DNA polymerase, a template, and a primer. During DNA synthesis, the dNTP is incorporated by a new phosphodiester bond with the primer. The incorporation of ddNTP blocks further DNA synthesis of the growing chain. (© Richard C. Li.)

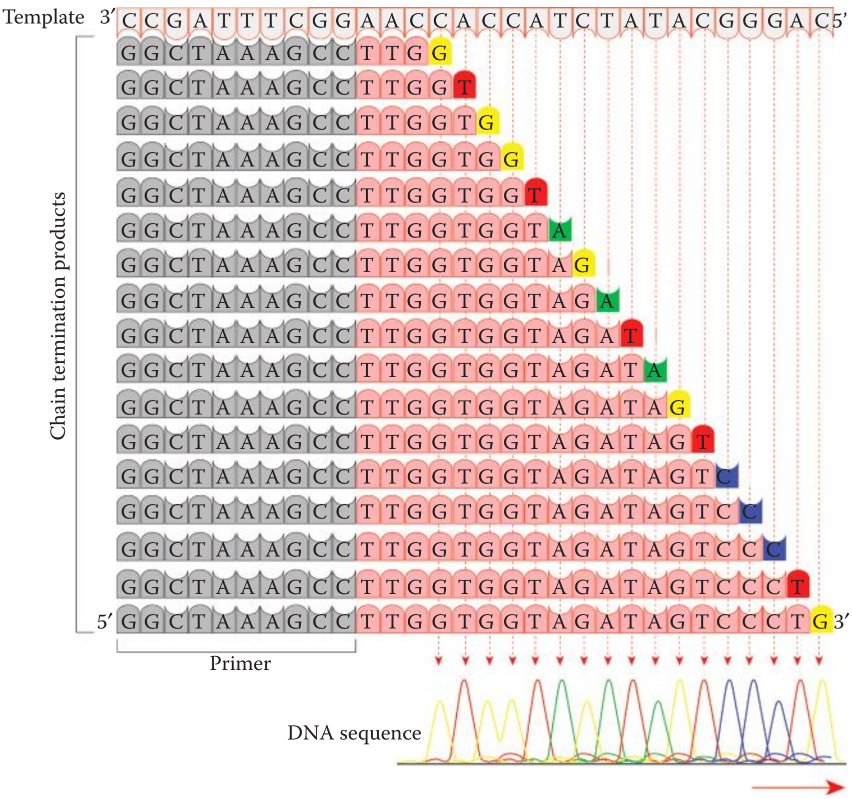
Template 3′ C C G A T T T C G G A A C C A C C A T C T A T A C G G A C 5′
Chain termination products
5′ G G C T A A A G C C T T G G T G G T A G A T A G T C C C T G 3′
Primer
DNA sequence

Figure 23.14 Diagram of the Sanger sequencing products. The DNA chain lengths are determined by the addition of ddNTP at different positions. The fluorescent dye-labeled cycle-sequencing products are separated using capillary electrophoresis. The fluorescent dyes are resolved by a detector, and the peaks corresponding to each DNA fragment are identified. (© Richard C. Li.)

the base number for the position is listed followed by an N. For example, 228N means an unresolved sequence ambiguity was observed at position 228.

23.3.3.3.2 Insertions

The insertion site is described by noting the position (at 5′ to the insertion) followed by a decimal point and a number. The number indicates the order of the insertions (e.g., 1 indicates the first insertion, 2 indicates the second, etc.). The base calling following the number indicates the inserted nucleotide (e.g., 524.1A, 524.2C).

23.3.3.3.3 Deletions

The deletion site designation is followed by the letter *d*. For example, a deletion at position 16,296 is recorded as 16,296d.

23.3.3.3.4 Heteroplasmic Sites

The IUPAC codes for base calling can be applied to heteroplasmic sites. For example, an A/G heteroplasmy can be designated as R, and a C/T heteroplasmy can be designated as Y.

23.3.4 Interpretation of mtDNA Profiling Results

Interpretation guidelines are used when comparing sequencing results between evidence and reference samples. General guidelines were set forth by the Scientific Working Group on DNA

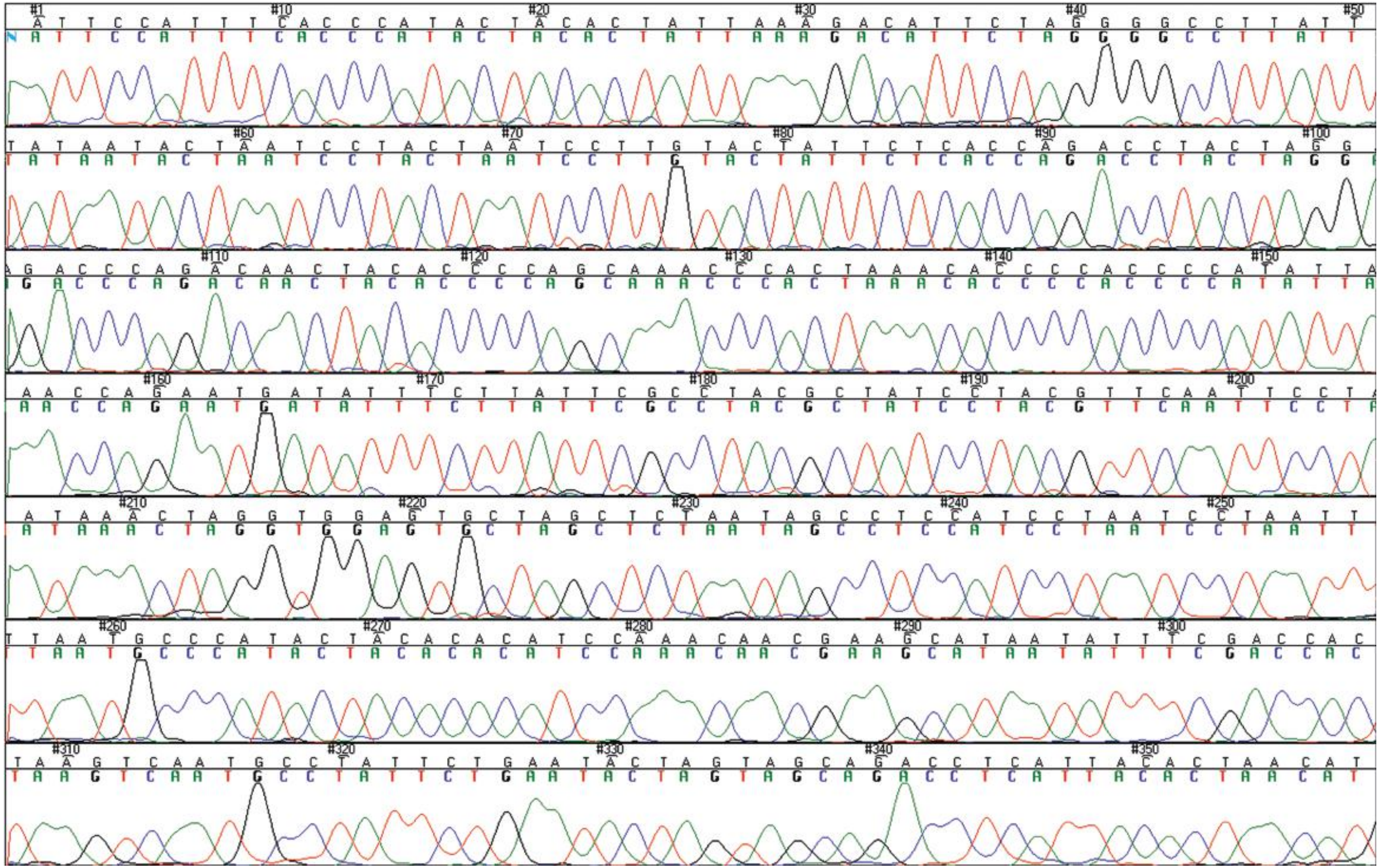

Figure 23.15 Electropherogram representing a DNA sequence presented as peaks with corresponding colors for each of the four bases. The bases associated with each peak are noted. The sequence is read from left to right (5′ to 3′). (© Richard C. Li.)

Analysis Methods (SWGDAM) and the DNA Commission of the International Society of Forensic Genetics (ISFG). In reporting mtDNA profiling results, the most common categories of conclusions are the following: cannot exclude, exclusion, and inconclusive result.

23.3.4.1 Exclusion

If the sequences of questioned and known samples are different, then the samples can be excluded as originating from the same source. It should be taken into account that higher mutation rates are found with the mtDNA genome than are found with the nuclear genome. The SWGDAM's guidelines define that the conclusion of exclusion can be made if there are two or more nucleotide differences between the questioned and known samples. Additionally, mutations seem to be more common in certain tissues. For this reason, the sources of the tissues analyzed should be taken into consideration.

23.3.4.2 Cannot Exclude

If the sequences are the same, the reference sample and evidence cannot be excluded as arising from the same source. When an mtDNA profile cannot be excluded, it is desirable to evaluate the weight of the evidence. In cases where the same heteroplasmy is observed in both questioned and known samples, its presence increases the strength of the evidence. However, if heteroplasmy is observed in a questioned sample but not in a known sample or vice versa, a maternal lineage still cannot be excluded.

23.3.4.3 Inconclusive Result

If the questioned and known samples differ by a single nucleotide, and no heteroplasmy is present, the results are considered to be inconclusive.

Bibliography

Mitochondrial DNA Profiling

Adachi, N., K. Umetsu, and H. Shojo, Forensic strategy to ensure the quality of sequencing data of mitochondrial DNA in highly degraded samples. *Leg Med (Tokyo)*, 2014, **16**(1): 52–55.

Anderson, S., et al., Sequence and organization of the human mitochondrial genome. *Nature*, 1981, **290**(5806): 457–465.

Andreasson, H., et al., Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. *Forensic Sci Int Genet*, 2007, **1**(1): 35–43.

Andrews, R.M., et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 1999, **23**(2): 147.

Anjum, G.M., et al., Pyrosequencing-based strategy for a successful SNP detection in two hypervariable regions: HV-I/HV-II of the human mitochondrial displacement loop. *Electrophoresis*, 2010, **31**(2): 309–314.

Arnestad, M., et al., A mitochondrial DNA polymorphism associated with cardiac arrhythmia investigated in sudden infant death syndrome. *Acta Paediatr*, 2007, **96**(2): 206–210.

Asari, M., et al., Differences in tissue distribution of HV2 length heteroplasmy in mitochondrial DNA between mothers and children. *Forensic Sci Int*, 2008, **175**(2–3): 155–159.

Bandelt, H.J. and W. Parson, Consistent treatment of length variants in the human mtDNA control region: A reappraisal. *Int J Legal Med*, 2008, **122**(1): 11–21.

Bandelt, H.J., M. van Oven, and A. Salas, Haplogrouping mitochondrial DNA sequences in legal medicine/forensic genetics. *Int J Legal Med*, 2012, **126**(6): 901–916.

Bendall, K.E. and B.C. Sykes, Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet*, 1995, **57**(2): 248–256.

Bendall, K.E., V.A. Macaulay, and B.C. Sykes, Variable levels of a heteroplasmic point mutation in individual hair roots. *Am J Hum Genet*, 1997, **61**(6): 1303–1308.

Bendall, K.E., et al., Heteroplasmic point mutations in the human mtDNA control region. *Am J Hum Genet*, 1996, **59**(6): 1276–1287.

Berger, C. and W. Parson, Mini-midi-mito: Adapting the amplification and sequencing strategy of mtDNA to the degradation state of crime scene samples. *Forensic Sci Int Genet*, 2009, **3**(3): 149–153.

Berger, C., et al., Evaluating sequence-derived mtDNA length heteroplasmy by amplicon size analysis. *Forensic Sci Int Genet*, 2011, **5**(2): 142–145.

Bintz, B.J., G.B. Dixon, and M.R. Wilson, Simultaneous detection of human mitochondrial DNA and nuclear-inserted mitochondrial-origin sequences (NumtS) using forensic mtDNA amplification strategies and pyrosequencing technology. *J Forensic Sci*, 2014, **59**(4): 1064–1073.

Brandon, M.C., et al., MITOMASTER: A bioinformatics tool for the analysis of mitochondrial DNA sequences. *Hum Mutat*, 2009, **30**(1): 1–6.

Brandstatter, A., et al., Dissection of mitochondrial superhaplogroup H using coding region SNPs. *Electrophoresis*, 2006, **27**(13): 2541–2550.

Buckleton, J.S., J.M. Curran, and P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Sci Int Genet*, 2007, **1**(1): 20–28.

Budowle, B., et al., Mitochondrial DNA regions HVI and HVII population data. *Forensic Sci Int*, 1999, **103**(1): 23–35.

Budowle, B., et al., Forensics and mitochondrial DNA: Applications, debates, and foundations. *Annu Rev Genom Hum Genet*, 2003, **4**: 119–141.

Budowle, B., et al., Addressing the use of phylogenetics for identification of sequences in error in the SWGDAM mitochondrial DNA database. *J Forensic Sci*, 2004, **49**(6): 1256–1261.

Budowle, B., et al., Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *J Forensic Sci*, 2009, **54**(4): 810–821.

Budowle, B., et al., Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region. *J Forensic Sci*, 2010, **55**(5): 1190–1195.

Butler, J.M. and B.C. Levin, Forensic applications of mitochondrial DNA. *Trends Biotechnol*, 1998, **16**(4): 158–162.

Calloway, C.D., et al., The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. *Am J Hum Genet*, 2000, **66**(4): 1384–1397.

Carracedo, A., et al., DNA commission of the international society for forensic genetics: Guidelines for mitochondrial DNA typing. *Forensic Sci Int*, 2000, **110**(2): 79–85.

Cassandrini, D., et al., A new method for analysis of mitochondrial DNA point mutations and assess levels of heteroplasmy. *Biochem Biophys Res Commun*, 2006, **342**(2): 387–393.

Cerezo, M., et al., Applications of MALDI-TOF MS to large-scale human mtDNA population-based studies. *Electrophoresis*, 2009, **30**(21): 3665–3673.

Chemale, G., et al., Development and validation of a D-loop mtDNA SNP assay for the screening of specimens in forensic casework. *Forensic Sci Int Genet*, 2013, **7**(3): 353–358.

Clayton, D.A., Structure and function of the mitochondrial genome. *J Inherit Metab Dis*, 1992, **15**(4): 439–447.

Comas, D., S. Paabo, and J. Bertranpetit, Heteroplasmy in the control region of human mitochondrial DNA. *Genome Res*, 1995, **5**(1): 89–90.

Crespillo, M., et al., Results of the 2003–2004 GEP-ISFG collaborative study on mitochondrial DNA: Focus on the mtDNA profile of a mixed semen-saliva stain. *Forensic Sci Int*, 2006, **160**(2–3): 157–167.

Curran, J.M., A MCMC method for resolving two person mixtures. *Sci Justice*, 2008, **48**(4): 168–177.

Curtis, P.C., et al., Optimization of primer-specific filter metrics for the assessment of mitochondrial DNA sequence data. *Mitochondrial DNA*, 2010, **21**(6): 191–197.

D'Eustachio, P., High levels of mitochondrial DNA heteroplasmy in human hairs by Budowle et al. *Forensic Sci Int*, 2002, **130**(1): 63–67; author reply 68–70.

Danielson, P.B., et al., Resolving mtDNA mixtures by denaturing high-performance liquid chromatography and linkage phase determination. *Forensic Sci Int Genet*, 2007, **1**(2): 148–153.

Danielson, P.B., et al., Separating human DNA mixtures using denaturing high-performance liquid chromatography. *Expert Rev Mol Diagn*, 2005, **5**(1): 53–63.

Desmyter, S. and B. Hoste, Influence of the electrophoresis-resequencing method on the forensic mtDNA profiling quality. *Forensic Sci Int Genet*, 2007, **1**(2): 199–200.

Eduardoff, M., et al., Mass spectrometric base composition profiling: Implications for forensic mtDNA databasing. *Forensic Sci Int Genet*, 2013, **7**(6): 587–592.

Eichmann, C. and W. Parson, "Mitominis": Multiplex PCR analysis of reduced size amplicons for compound sequence analysis of the entire mtDNA control region in highly degraded samples. *Int J Legal Med*, 2008, **122**(5): 385–388.

Fan, L. and Y.G. Yao, An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion*, 2013, **13**(4): 360–363.

Fattorini, P., et al., Estimating the integrity of aged DNA samples by CE. *Electrophoresis*, 2009, **30**(22): 3986–3995.

Fendt, L., et al., Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. *BMC Genom*, 2009, **10**: 139.

Fernández, C. and A. Alonso, Microchip capillary electrophoresis protocol to evaluate quality and quantity of mtDNA amplified fragments for DNA sequencing in forensic genetics. *Methods Mol Biol*, 2012, **830**: 367–379.

Forster, L., et al., Evaluating length heteroplasmy in the human mitochondrial DNA control region. *Int J Legal Med*, 2010, **124**(2): 133–142.

Gabriel, M.N., et al., Improved mtDNA sequence analysis of forensic remains using a "mini-primer set" amplification strategy. *J Forensic Sci*, 2001, **46**(2): 247–253.

Gidlof, O., et al., Complete discrimination of six individuals based on high-resolution melting of hypervariable regions I and II of the mitochondrial genome. *Biotechniques*, 2009, **47**(2): 671–672, 674, 676, passim.

Gill, P., et al., National recommendations of the technical UK DNA working group on mixture interpretation for the NDNAD and for court going purposes. *Forensic Sci Int Genet*, 2008, **2**(1): 76–82.

Grignani, P., et al., Subtyping mtDNA haplogroup H by SNaPshot minisequencing and its application in forensic individual identification. *Int J Legal Med*, 2006, **120**(3): 151–156.

Grignani, P., et al., A mini-primer set covering the mtDNA hypervariable regions for the genetic typing of old skeletal remains. *Forensic Sci Int Genet Suppl*, 2009, **2**(1): 265–266.

Grignani, P., et al., Multiplex mtDNA coding region SNP assays for molecular dissection of haplogroups U/K and J/T. *Forensic Sci Int Genet*, 2009, **4**(1): 21–25.

Grzybowski, T. and U. Rogalla, Mitochondria in anthropology and forensic medicine. *Adv Exp Med Biol*, 2012, **942**: 441–453.

Hall, T.A., et al., Base composition profiling of human mitochondrial DNA using polymerase chain reaction and direct automated electrospray ionization mass spectrometry. *Anal Chem*, 2009, **81**(18): 7515–7526.

Hartmann, A., et al., Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum Mutat*, 2009, **30**(1): 115–122.

Hauswirth, W.W. and D.A. Clayton, Length heterogeneity of a conserved displacement-loop sequence in human mitochondrial DNA. *Nucleic Acids Res*, 1985, **13**(22): 8093–8104.

He, Y., et al., Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature*, 2010, **464**(7288): 610–614.

Holland, M.M., M.R. McQuillan, and K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. *Croat Med J*, 2011, **52**(3): 299–313.

Holland, M.M., Molecular analysis of the human mitochondrial DNA control region for forensic identity testing. *Curr Protoc Hum Genet*, 2012, (Suppl 74) 10.1002/0471142905.hg1407s74.

Howard, R., et al., Comparative analysis of human mitochondrial DNA from World War I bone samples by DNA sequencing and ESI-TOF mass spectrometry. *Forensic Sci Int Genet*, 2013, **7**(1): 1–9.

Hwan, Y.L., et al., A modified mini-primer set for analyzing mitochondrial DNA control region sequences from highly degraded forensic samples. *BioTechniques*, 2008, **44**(4): 555–558.

Irwin, J.A., et al., mtGenome reference population databases and the future of forensic mtDNA analysis. *Forensic Sci Int Genet*, 2011, **5**(3): 222–225.

Jarman, P.G., S.L. Fentress, and D.E. Katz, Mitochondrial DNA validation in a state laboratory. *J Forensic Sci*, 2009, **54**(1): 95–102.

Just, R.S., et al., Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. *Forensic Sci Int Genet*, 2008, **2**(3): e45–e48.

Just, R.S., et al., The use of mitochondrial DNA single nucleotide polymorphisms to assist in the resolution of three challenging forensic cases. *J Forensic Sci*, 2009, **54**(4): 887–891.

Just, R.S., et al., Development of forensic-quality full mtGenome haplotypes: Success rates with low template specimens. *Forensic Sci Int Genet*, 2014, **10**: 73–79.

Kim, N.Y., et al., Modified midi- and mini-multiplex PCR systems for mitochondrial DNA control region sequence analysis in degraded samples. *J Forensic Sci*, 2013, **58**(3): 738–743.

Kohnemann, S. and H. Pfeiffer, Application of mtDNA SNP analysis in forensic casework. *Forensic Sci Int Genet*, 2011, **5**(3): 216–221.

Kohnemann, S., C. Hohoff, and H. Pfeiffer, An economical mtDNA SNP assay detecting different mitochondrial haplogroups in identical HVR 1 samples of Caucasian ancestry. *Mitochondrion*, 2009, **9**(5): 370–375.

Kohnemann, S., et al., A rapid mtDNA assay of 22 SNPs in one multiplex reaction increases the power of forensic testing in European Caucasians. *Int J Legal Med*, 2008, **122**(6): 517–523.

Kristinsson, R., S.E. Lewis, and P.B. Danielson, Comparative analysis of the HV1 and HV2 regions of human mitochondrial DNA by denaturing high-performance liquid chromatography. *J Forensic Sci*, 2009, **54**(1): 28–36.

Kurelac, I., et al., Searching for a needle in the haystack: Comparing six methods to evaluate heteroplasmy in difficult sequence context. *Biotechnol Adv*, 2012, **30**(1): 363–371.

Lacan, M., et al., Detection of the A189G mtDNA heteroplasmic mutation in relation to age in modern and ancient bones. *Int J Legal Med*, 2009, **123**(2): 161–167.

Lacan, M., et al., Detection of age-related duplications in mtDNA from human muscles and bones. *Int J Legal Med*, 2011, **125**(2): 293–300.

Lee, H.Y., et al., A modified mini-primer set for analyzing mitochondrial DNA control region sequences from highly degraded forensic samples. *Biotechniques*, 2008, **44**(4): 555–556, 558.

Lee, H.Y., et al., DNA typing for the identification of old skeletal remains from Korean War victims. *J Forensic Sci*, 2010, **55**(6): 1422–1429.

Lee, H.Y., et al., A one step multiplex PCR assay for rapid screening of East Asian mtDNA haplogroups on forensic samples. *Leg Med (Tokyo)*, 2013, **15**(1): 50–54.

Linch, C.A., D.A. Whiting, and M.M. Holland, Human hair histogenesis for the mitochondrial DNA forensic scientist. *J Forensic Sci*, 2001, **46**(4): 844–853.

Loreille, O.M. and J.A. Irwin, Capillary electrophoresis of human mtDNA control region sequences from highly degraded samples using short mtDNA amplicons. *Methods Mol Biol*, 2012, **830**: 283–299.

Mabuchi, T., et al., Typing the 1.1 kb control region of human mitochondrial DNA in Japanese individuals. *J Forensic Sci*, 2007, **52**(2): 355–363.

Malyarchuk, B.A., Improving the reconstructed sapiens reference sequence of mitochondrial DNA. *Forensic Sci Int Genet*, 2013, **7**(3): e74–e75.

Maricic, T., M. Whitten, and S. Paabo, Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*, 2010, **5**(11): e14004.

Marquez, M.C., Interpretation guidelines of mtDNA control region sequence electropherograms in forensic genetics. *Methods Mol Biol*, 2012, **830**: 301–319.

Melton, T., et al., Forensic mitochondrial DNA analysis of 691 casework hairs. *J Forensic Sci*, 2005, **50**(1): 73–80.

Miller, K.W. and B. Budowle, A compendium of human mitochondrial DNA control region: Development of an international standard forensic database. *Croat Med J*, 2001, **42**(3): 315–327.

Montesino, M. and L. Prieto, Capillary electrophoresis of big-dye terminator sequencing reactions for human mtDNA control region haplotyping in the identification of human remains. *Methods Mol Biol*, 2012, **830**: 267–281.

Montesino, M., et al., Analysis of body fluid mixtures by mtDNA sequencing: An inter-laboratory study of the GEP-ISFG working group. *Forensic Sci Int*, 2007, **168**(1): 42–56.

Moore, C.A., J. Gudikote, and G.C. Van Tuyle, Mitochondrial DNA rearrangements, including partial duplications, occur in young and old rat tissues. *Mutat Res*, 1998, **421**(2): 205–217.

Morling, N., et al., Interpretation of DNA mixtures—European consensus on principles. *Forensic Sci Int Genet*, 2007, **1**(3–4): 291–292.

Mosquera-Miguel, A., et al., Testing the performance of mtSNP minisequencing in forensic samples. *Forensic Sci Int Genet*, 2009, **3**(4): 261–264.

Nakahara, H., et al., Heteroplasmies detected in an amplified mitochondrial DNA control region from a small amount of template. *J Forensic Sci*, 2008, **53**(2): 306–311.

Nara, A., et al., Sequence analysis for HV I region of mitochondrial DNA using WGA (whole genome amplification) method. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S115–S118.

Naue, J., et al., Factors affecting the detection and quantification of mitochondrial point heteroplasmy using Sanger sequencing and SNaPshot minisequencing. *Int J Legal Med*, 2011, **125**(3): 427–436.

Nelson, K. and T. Melton, Forensic mitochondrial DNA analysis of 116 casework skeletal samples. *J Forensic Sci*, 2007, **52**(3): 557–561.

Nelson, T.M., et al., Development of a multiplex single base extension assay for mitochondrial DNA haplogroup typing. *Croat Med J*, 2007, **48**(4): 460–472.

Nicklas, J.A., T. Noreault-Conti, and E. Buel, Development of a real-time method to detect DNA degradation in forensic samples. *J Forensic Sci*, 2012, **57**(2): 466–471.

Nilsson, M., et al., Evaluation of mitochondrial DNA coding region assays for increased discrimination in forensic analysis. *Forensic Sci Int Genet*, 2008, **2**(1): 1–8.

Oberacher, H., H. Niederstatter, and W. Parson, Liquid chromatography-electrospray ionization mass spectrometry for simultaneous detection of mtDNA length and nucleotide polymorphisms. *Int J Legal Med*, 2007, **121**(1): 57–67.

Oberacher, H., et al., Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry. *Hum Mutat*, 2008, **29**(3): 427–432.

Pakendorf, B. and M. Stoneking, Mitochondrial DNA and human evolution. *Annu Rev Genom Hum Genet*, 2005, **6**: 165–183.

Paneto, G.G., et al., Heteroplasmy in hair: Study of mitochondrial DNA third hypervariable region in hair and blood samples. *J Forensic Sci*, 2010, **55**(3): 715–718.

Parson, W. and A. Dur, EMPOP—A forensic mtDNA database. *Forensic Sci Int Genet*, 2007, **1**(2): 88–92.

Parsons, T.J. and M.D. Coble, Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croat Med J*, 2001, **42**(3): 304–309.

Polanskey, D., et al., Comparison of mitotyper rules and phylogenetic-based mtDNA nomenclature systems. *J Forensic Sci*, 2010, **55**(5): 1184–1189.

Prieto, L., et al., 2006 GEP-ISFG collaborative exercise on mtDNA: Reflections about interpretation, artefacts, and DNA mixtures. *Forensic Sci Int Genet*, 2008, **2**(2): 126–133.

Prieto, L., et al., GHEP-ISFG proficiency test 2011: Paper challenge on evaluation of mitochondrial DNA results. *Forensic Sci Int Genet*, 2013, **7**(1): 10–15.

Quintans, B., et al., Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. *Forensic Sci Int*, 2004, **140**(2–3): 251–257.

Roberts, K.A. and C. Calloway, Characterization of mitochondrial DNA sequence heteroplasmy in blood tissue and hair as a function of hair morphology. *J Forensic Sci*, 2011, **56**(1): 46–60.

Rock, A., et al., SAM: String-based sequence search algorithm for mitochondrial DNA database queries. *Forensic Sci Int Genet*, 2011, **5**(2): 126–132.

Salas, A., et al., Phylogeographic investigations: The role of trees in forensic genetics. *Forensic Sci Int*, 2007, **168**(1): 1–13.

Salas, A., et al., A cautionary note on switching mitochondrial DNA reference sequences in forensic genetics. *Forensic Sci Int Genet*, 2012, **6**(6): e182–e184.

Schneider, P.M., et al., The German stain commission: Recommendations for the interpretation of mixed stains. *Int J Legal Med*, 2009, **123**(1): 1–5.

Seo, S.B., et al., Alterations of length heteroplasmy in mitochondrial DNA under various amplification conditions. *J Forensic Sci*, 2010, **55**(3): 719–722.

Serizawa, Y., et al., Detection of mitochondrial DNA polymorphisms from human hair shafts and formalin fixed tissue using whole genome amplification. *Int Med J*, 2008, **15**(3): 163–167.

Sosa, C., et al., A preliminary study on the incidence of heteroplasmy in mitochondrial DNA from vitreous humour. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S460–S462.

Stringer, P., et al., Interpretation of DNA mixtures—Australian and New Zealand consensus on principles. *Forensic Sci Int Genet*, 2009, **3**(2): 144–145.

Sutovsky, P., et al., Ubiquitin tag for sperm mitochondria. *Nature*, 1999, **402**: 371–372.

Tobe, S.S. and A.M. Linacre, A technique for the quantification of human and non-human mammalian mitochondrial DNA copy number in forensic and other mixtures. *Forensic Sci Int Genet*, 2008, **2**(4): 249–256.

Tully, G., et al., Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forensic Sci Int*, 2001, **124**(1): 83–91.

Tully, G., et al., Results of a collaborative study of the EDNAP group regarding mitochondrial DNA heteroplasmy and segregation in hair shafts. *Forensic Sci Int*, 2004, **140**(1): 1–11.

Tully, L.A. and B.C. Levin, Human mitochondrial genetics. *Biotechnol Genet Eng Rev*, 2000, **17**: 147–177.

Turchi, C., et al., Italian mitochondrial DNA database: Results of a collaborative exercise and proficiency testing. *Int J Legal Med*, 2008, **122**(3): 199–204.

Umetsu, K. and I. Yuasa, Recent progress in mitochondrial DNA analysis. *Leg Med (Tokyo)*, 2005, **7**(4): 259–262.

Vallone, P.M., Capillary electrophoresis of an 11-plex mtDNA coding region SNP single base extension assay for discrimination of the most common Caucasian HV1/HV2 mitotype. *Methods Mol Biol*, 2012, **830**: 159–167.

Vallone, P.M., J.P. Jakupciak, and M.D. Coble, Forensic application of the Affymetrix human mitochondrial resequencing array. *Forensic Sci Int Genet*, 2007, **1**(2): 196–198.

van Oven, M. and M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*, 2009, **30**(2): E386–E394.

Warshauer, D.H., et al., Validation of the PLEX-ID™ mass spectrometry mitochondrial DNA assay. *Int J Legal Med*, 2013, **127**(2): 277–286.

Wilson, M.R., et al., Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. *Biotechniques*, 1995, **18**(4): 662–669.

Wilson, M.R., et al., Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. *Forensic Sci Int*, 2002, **129**(1): 35–42.

Wilson, M.R., et al., Validation of mitochondrial DNA sequencing for forensic casework analysis. *Int J Legal Med*, 1995, **108**(2): 68–74.

Wu, J., et al., Multiplex mutagenically separated polymerase chain reaction assay for rapid detection of human mitochondrial DNA variations in coding area. *Croat Med J*, 2008, **49**(1): 32–38.

Xiu-Cheng Fan, A., et al., A rapid and accurate approach to identify single nucleotide polymorphisms of mitochondrial DNA using MALDI-TOF mass spectrometry. *Clin Chem Lab Med*, 2008, **46**(3): 299–305.

Genetic Population Data

Afonso Costa, H., et al., Mitochondrial DNA sequence analysis of a native Bolivian population. *J Forensic Leg Med*, 2010, **17**(5): 247–253.

Alshamali, F., et al., Mitochondrial DNA control region variation in Dubai, United Arab Emirates. *Forensic Sci Int Genet*, 2008, **2**(1): e9–e10.

Alvarez, J.C., et al., Characterization of human control region sequences for Spanish individuals in a forensic mtDNA data set. *Leg Med (Tokyo)*, 2007, **9**(6): 293–304.

Alvarez-Iglesias, V., et al., Coding region mitochondrial DNA SNPs: Targeting East Asian and native American haplogroups. *Forensic Sci Int Genet*, 2007, **1**(1): 44–55.

Al-Zahery, N., et al., Characterization of mitochondrial DNA control region lineages in Iraq. *Int J Legal Med*, 2013, **127**(2): 373–375.

Asari, M., et al., Utility of haplogroup determination for forensic mtDNA analysis in the Japanese population. *Leg Med (Tokyo)*, 2007, **9**(5): 237–240.

Ballantyne, K.N., et al., MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. *Forensic Sci Int Genet*, 2012, **6**(4): 425–436.

Barbosa, A.B., et al., Mitochondrial DNA control region polymorphism in the population of Alagoas state, north-eastern Brazil. *J Forensic Sci*, 2008, **53**(1): 142–146.

Bobillo, M.C., et al., Amerindian mitochondrial DNA haplogroups predominate in the population of Argentina: Towards a first nationwide forensic mitochondrial DNA sequence database. *Int J Legal Med*, 2010, **124**(4): 263–268.

Bodner, M., et al., Southeast Asian diversity: First insights into the complex mtDNA structure of Laos. *BMC Evol Biol*, 2011, **11**: 49.

Brandstatter, A., et al., Generating population data for the EMPOP database—An overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. *Forensic Sci Int*, 2007, **166**(2–3): 164–175.

Brandstatter, A., et al., Mitochondrial DNA control region variation in Ashkenazi Jews from Hungary. *Forensic Sci Int Genet*, 2008, **2**(1): e4–e6.

Cardoso, S., et al., Mitochondrial DNA control region variation in an autochthonous Basque population sample from the Basque Country. *Forensic Sci Int Genet*, 2012, **6**(4): e106–e108.

Cardoso, S., et al., Mitochondrial DNA control region data reveal high prevalence of nNative American lineages in Jujuy province, NW Argentina. *Forensic Sci Int Genet*, 2013, **7**(3): e52–e55.

Cardoso, S., et al., Variability of the entire mitochondrial DNA control region in a human isolate from the Pas Valley (northern Spain). *J Forensic Sci*, 2010, **55**(5): 1196–1201.

Castro de Guerra, D., et al., Sequence variation of mitochondrial DNA control region in North Central Venezuela. *Forensic Sci Int Genet*, 2012, **6**(5): e131–e133.

Chen, F., et al., Analysis of mitochondrial DNA polymorphisms in Guangdong Han Chinese. *Forensic Sci Int Genet*, 2008, **2**(2): 150–153.

Chen, F., et al., Genetic polymorphism of mitochondrial DNA HVS-I and HVS-II of Chinese Tu ethnic minority group. *J Genet Genom*, 2008, **35**(4): 225–232.

Chen, F., et al., Sequence-length variation of mtDNA HVS-I C-stretch in Chinese ethnic groups. *J Zhejiang Univ Sci B*, 2009, **10**(10): 711–720.

Diegoli, T.M., et al., Mitochondrial control region sequences from an African American population sample. *Forensic Sci Int Genet*, 2009, **4**(1): e45–e52.

Egyed, B., et al., Mitochondrial control region sequence variations in the Hungarian population: Analysis of population samples from Hungary and from Transylvania (Romania). *Forensic Sci Int Genet*, 2007, **1**(2): 158–162.

Fendt, L., et al., Mitochondrial DNA control region data from indigenous Angolan Khoe-San lineages. *Forensic Sci Int Genet*, 2012, **6**(5): 662–663.

Fendt, L., et al., MtDNA diversity of Ghana: A forensic and phylogeographic view. *Forensic Sci Int Genet*, 2012, **6**(2): 244–249.

Grzybowski, T., et al., Complex interactions of the eastern and western Slavic populations with other European groups as revealed by mitochondrial DNA analysis. *Forensic Sci Int Genet*, 2007, **1**(2): 141–147.

Hedman, M., et al., Finnish mitochondrial DNA HVS-I and HVS-II population data. *Forensic Sci Int*, 2007, **172**(2–3): 171–178.

Huang, D., et al., Typing of 24 mtDNA SNPs in a Chinese population using SNaPshot minisequencing. *J Huazhong Univ Sci Technol Med Sci*, 2010, **30**(3): 291–298.

Irwin, J., et al., Hungarian mtDNA population databases from Budapest and the Baranya county Roma. *Int J Legal Med*, 2007, **121**(5): 377–383.

Irwin, J., et al., Mitochondrial control region sequences from northern Greece and Greek Cypriots. *Int J Legal Med*, 2008, **122**(1): 87–89.

Irwin, J.A., et al., Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. *Forensic Sci Int Genet*, 2007, **1**(2): 154–157.

Irwin, J.A., et al., Mitochondrial DNA control region variation in a population sample from Hong Kong, China. *Forensic Sci Int Genet*, 2009, **3**(4): e119–e125.

Irwin, J.A., et al., The mtDNA composition of Uzbekistan: A microcosm of Central Asian patterns. *Int J Legal Med*, 2010, **124**(3): 195–204.

Kato, H., et al., Molecular analysis of mitochondrial hypervariable region 1 in 394 Japanese individuals. *Leg Med (Tokyo)*, 2009, **11**(Suppl 1): S443–S445.

Lander, N., et al., Haplotype diversity in human mitochondrial DNA hypervariable regions I-III in the city of Caracas (Venezuela). *Forensic Sci Int Genet*, 2008, **2**(4): e61–e64.

Lehocky, I., et al., A database of mitochondrial DNA hypervariable regions I and II sequences of individuals from Slovakia. *Forensic Sci Int Genet*, 2008, **2**(4): e53–e59.

Liu, C., et al., Mitochondrial DNA polymorphisms in Gelao ethnic group residing in Southwest China. *Forensic Sci Int Genet*, 2011, **5**(1): e4–e10.

Maruyama, S., et al., MtDNA control region sequence polymorphisms and phylogenetic analysis of Malay population living in or around Kuala Lumpur in Malaysia. *Int J Legal Med*, 2010, **124**(2): 165–170.

Maruyama, S., et al., Analysis of human mitochondrial DNA polymorphisms in the Japanese population. *Biochem Genet*, 2013, **51**(1–2): 33–70.

Mikkelsen, M., et al., Mitochondrial DNA HV1 and HV2 variation in Danes. *Forensic Sci Int Genet*, 2010, **4**(4): e87–e88.

Mikkelsen, M., et al., Forensic and phylogeographic characterisation of mtDNA lineages from Somalia. *Int J Legal Med*, 2012, **126**(4): 573–579.

Nohira, C., S. Maruyama, and K. Minaguchi, Phylogenetic classification of Japanese mtDNA assisted by complete mitochondrial DNA sequences. *Int J Legal Med*, 2010, **124**(1): 7–12.

Nur Haslindawaty, A.R., et al., Sequence polymorphisms of mtDNA HV1, HV2, and HV3 regions in the Malay population of Peninsular Malaysia. *Int J Legal Med*, 2010, **124**(5): 415–426.

Palencia, L., et al., Mitochondrial DNA diversity in a population from Santa Catarina (Brazil): Predominance of the European input. *Int J Legal Med*, 2010, **124**(4): 331–336.

Parson, W., et al., Identification of West Eurasian mitochondrial haplogroups by mtDNA SNP screening: Results of the 2006–2007 EDNAP collaborative exercise. *Forensic Sci Int Genet*, 2008, **2**(1): 61–68.

Powell, G.T., et al., The population history of the Xibe in northern China: A comparison of autosomal, mtDNA and Y-chromosomal analyses of migration and gene flow. *Forensic Sci Int Genet*, 2007, **1**(2): 115–119.

Rakha, A., et al., Forensic and genetic characterization of mtDNA from Pathans of Pakistan. *Int J Legal Med*, 2011, **125**(6): 841–848.

Ribeiro-dos-Santos, A.K., et al., Nucleotide variability of HV-I in Afro-descendents populations of the Brazilian Amazon region. *Forensic Sci Int*, 2007, **167**(1): 77–80.

Saunier, J.L., et al., Mitochondrial control region sequences from a U.S. "Hispanic" population sample. *Forensic Sci Int Genet*, 2008, **2**(2): e19–e23.

Saunier, J.L., et al., Mitochondrial control region sequences from an Egyptian population sample. *Forensic Sci Int Genet*, 2009, **3**(3): e97–e103.

Scheible, M., et al., Mitochondrial DNA control region variation in a Kuwaiti population sample. *Forensic Sci Int Genet*, 2011, **5**(4): e112–e113.

Szibor, R., et al., Mitochondrial D-loop $(CA)_n$ repeat length heteroplasmy: Frequency in a German population sample and inheritance studies in two pedigrees. *Int J Legal Med*, 2007, **121**(3): 207–213.

Tetzlaff, S., et al., Mitochondrial DNA population data of HVS-I and HVS-II sequences from a northeast German sample. *Forensic Sci Int*, 2007, **172**(2–3): 218–224.

Tillmar, A.O., et al., Homogeneity in mitochondrial DNA control region sequences in Swedish subpopulations. *Int J Legal Med*, 2010, **124**(2): 91–98.

Turchi, C., et al., Polymorphisms of mtDNA control region in Tunisian and Moroccan populations: An enrichment of forensic mtDNA databases with Northern Africa data. *Forensic Sci Int Genet*, 2009, **3**(3): 166–172.

Wilson, J.L., et al., Forensic analysis of mtDNA haplotypes from two rural communities in Haiti reflects their population history. *J Forensic Sci*, 2012, **57**(6): 1457–1466.

Wong, H.Y., et al., Sequence polymorphism of the mitochondrial DNA hypervariable regions I and II in 205 Singapore Malays. *Leg Med (Tokyo)*, 2007, **9**(1): 33–37.

Yan, J., et al., Allele frequencies of mitochondrial D-loop $(CA)_n$ repeat polymorphism in six Chinese ethnic groups. *Leg Med (Tokyo)*, 2007, **9**(6): 330–331.

Yang, Y., et al., A new strategy for the discrimination of mitochondrial DNA haplogroups in Han population. *J Forensic Sci*, 2011, **56**(3): 586–590.

Zgonjanin, D., et al., Sequence polymorphism of the mitochondrial DNA control region in the population of Vojvodina Province, Serbia. *Leg Med (Tokyo)*, 2010, **12**(2): 104–107.

Zimmermann, B., et al., Mitochondrial DNA control region population data from Macedonia. *Forensic Sci Int Genet*, 2007, **1**(3–4): e4–e9.

Zimmermann, B., et al., Forensic and phylogeographic characterization of mtDNA lineages from northern Thailand (Chiang Mai). *Int J Legal Med*, 2009, **123**(6): 495–501.

SECTION V

Forensic Issues

24

Forensic DNA Databases

Tools for Crime Investigations

24.1 Brief History of Forensic DNA Databases

Forensic DNA databases are networks for exchanging information among law enforcement agencies to assist in solving crimes. For example, forensic DNA databases allow forensic laboratories to search DNA profiles against the databases to identify criminals. In 1995, the United Kingdom established the world's first national DNA database, NDNAD, in England and Wales. Scotland and Northern Ireland have their own databases but also submit their profiles to NDNAD. NDNAD demonstrated initial success in solving crimes. Three years later, the United States introduced its national *Combined DNA Index System* or CODIS. By the end of 1998, other countries (such as Austria, Germany, the Netherlands, New Zealand, and Slovenia) had also introduced national DNA databases. Table 24.1 describes some of these national DNA databases. This chapter will focus on CODIS.

24.2 Infrastructure of CODIS

In the United States, a pilot project of DNA databasing was initiated by the Federal Bureau of Investigation (FBI) and 14 participating state and local laboratories. Subsequently in 1994, the Congressional DNA Identification Act authorized the FBI to establish a national DNA database including "DNA identification profiles of persons convicted of crimes, and analyses of DNA samples recovered from crime scenes and from unidentified human remains." By 1997, 13 STR loci were selected and in 1998 were implemented as the core loci for the national database, known today as CODIS. All 50 states, the District of Columbia, the federal laboratories, the US Army Criminal Investigation Laboratory, and Puerto Rico contribute to CODIS.

CODIS has three hierarchical levels: the Local DNA Index System (LDIS), the State DNA Index System (SDIS), and the National DNA Index System (NDIS) (Figure 24.1). The LDIS is maintained at crime laboratories operated by police departments, sheriff's offices, and local agencies. All forensic DNA profiles originating at the local level are stored in the LDIS and are transmitted to the SDIS and NDIS. Each state maintains a SDIS, which is typically operated by a designated state laboratory. An SDIS also stores the DNA profiles generated from state laboratories. The quality assurance standards for a qualified laboratory were set up by the DNA Advisory Board in 1998. The periodic revision of the standards is now carried out by the

Table 24.1 Characteristics of National DNA Databases

Country	Year Established	Suspect Entry Criteria	Convicted Offender Entry Criteria	Removal Criteria
United Kingdom	1995	Any recordable offense that leads to imprisonment	Entered as suspect	Never removed
New Zealand	1996	No suspects entered	Relevant offense (≥7 years in prison)	Never removed unless conviction quashed
Austria	1997	Any recordable offense that leads to imprisonment	Entered as suspect	Only after acquittal
The Netherlands	1997	No suspects entered except when suspect's DNA is tested for case	Offense leading to >4 years in prison	20–30 years after conviction
Germany	1998	Offense leading to >1 year in prison	After court decision	After acquittal or 5–10 years after conviction if prognosis good
Slovenia	1998	Any recordable offense that leads to imprisonment	Entered as suspect	Depends on severity of crime
United States	1998	No suspects entered; under revision	Depends on state law	Depends on state law
Finland	1999	Offense leading to >1 year in prison	Entered as suspect	Only after acquittal
Sweden	2000	No suspects entered	Offense leading to >2 years in prison	10 years after release from prison
Switzerland	2000	Any recordable offense that leads to imprisonment	Entered as suspect	After acquittal or 5–30 years after conviction
France	2001	No suspects entered	Sexual assaults and serious crimes	40 years after conviction

Source: Adapted from Jobling, M.A. and Gill, P., *Nat Rev Genet*, 5, 739–751, 2004.

Scientific Working Group of DNA Analysis and Methods (SWGDAM). The SDIS enables local laboratories within that state to compare DNA profiles. SDIS also serves as the communication path through which the LDIS and NDIS are able to exchange messages. Communication is mediated using a secured network with encryption. Only authorized personnel, approved by the FBI, have access to CODIS network terminals and servers. NDIS is the highest level of the CODIS infrastructure, which contains DNA profiles contributed by participating local and state

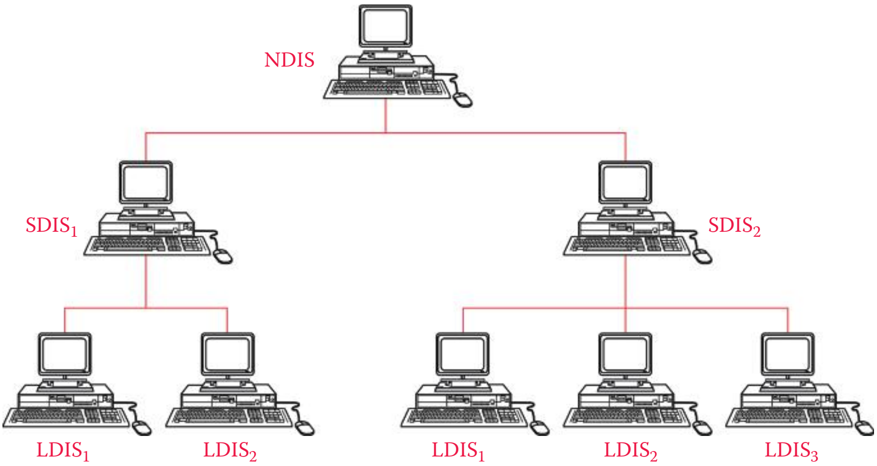

Figure 24.1 CODIS infrastructure. All DNA profiles originate at LDIS, and then enter SDIS and NDIS. SDIS allows laboratories within states to exchange DNA profiles. NDIS is the highest level of the infrastructure. It allows the participating laboratories to exchange DNA profiles on a national level.

laboratories. Additionally, NDIS stores DNA profiles generated by federal laboratories, such as those generated by the FBI and the US Army Criminal Investigation Laboratory. NDIS enables qualified state laboratories to compare DNA profiles. Searches of DNA profiles can be conducted at the national level. All DNA profiles submitted to NDIS are automatically searched weekly against the DNA profiles from other states. NDIS is administered by the FBI, which provides software, training, and support for all participating laboratories. Many law enforcement laboratories from other countries utilize the CODIS software for their own databases. However, laboratories in foreign countries do not have any access to the CODIS system. Nevertheless, a search through the CODIS database may be requested either from the FBI or International Criminal Police Organization (Interpol). Conversely, a search of an international DNA database can be arranged through Interpol.

24.3 Indexes of CODIS

The DNA profiles entered in CODIS are organized into categories known as indexes (Figure 24.2). The Convicted Offender Index contains DNA profiles of individuals convicted of crimes. The Arrestee Index contains DNA profiles of arrested individuals. It varies in SDIS databases based on each state's law permitting the collection of DNA samples from arrestees. The Forensic Index contains DNA profiles, also known as forensic profiles, derived from crime scene evidence, potentially originating from perpetrators but not including suspects.

Additionally, the FBI also established the National Missing Person DNA Database (NMPDD) Program for the identification of missing and unidentified persons at the national level. The DNA profiles entered into the NMPDD are categorized into three indexes that can be searched against each other. The Missing Person Index and the Unidentified Human Remains Index contain DNA profiles from missing persons and unidentified human remains, respectively. The Biological Relatives of the Missing Person Index contains DNA profiles voluntarily contributed from relatives. This index may also store patrilineal or matrilineal DNA profiles from the relative such as a biological father, mother, or child of the missing person to assist investigations. Additionally, a pedigree chart (a diagram showing the relationship between the missing person and relatives) can be created.

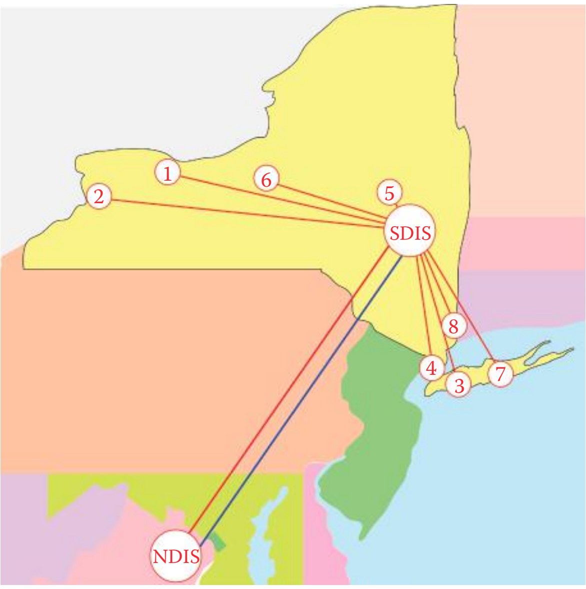

Figure 24.2 Example of an SDIS maintaining the Forensic Index (red) and Convicted Offender Index (blue). In this example, the New York State Police Forensic Investigation Center serves as the SDIS laboratory for New York State. LDIS sites: 1. Erie County Department of Central Police Services Forensic Laboratory, 2. Monroe County Public Safety Laboratory, 3. Nassau County Office of the Medical Examiner Division of Forensic Services, 4. New York City Office of the Chief Medical Examiner Department of Forensic Biology, 5. New York State Police Forensic Investigation Center, 6. Onondaga County Center for Forensic Sciences, 7. Suffolk County Crime Laboratory, and 8. Westchester County Department of Laboratories & Research Division of Forensic Sciences. (© Richard C. Li.)

In 2010, the FBI established a Rapid DNA Program Office for the purpose of developing rapid DNA technology. Rapid DNA technology is a fully automated process of performing STR analysis within 1–2 h to generate a CODIS core STR profile from a reference sample such as a buccal swab (Figure 24.3). The Rapid DNA Index System (RDIS) is the proposed index of NDIS. It shall be an integrated system capable of applying rapid DNA technology and carrying out database searches from police custody or booking units by trained police officers. The entire process, including obtaining any match from a database search, shall take less than 2 h.

The forensic DNA analysis of a reference sample, taken from an individual, takes 2–3 days if processed immediately in a forensic laboratory. The entire process includes the extraction of DNA from a sample, DNA quantification, PCR amplification, electrophoresis, and data collection and analysis. After being transported to a laboratory for processing, samples are usually stored and batched in laboratories prior to analysis. A typical turnaround time for analysis of such a sample is 1–3 months. While the sample is processed, the suspect is often released from custody. Perpetrators released on police bail may commit another crime. Thus, it is desirable to develop new technology that is capable of completing the forensic DNA analysis and database search while the suspect remains in custody. This technology would also facilitate rapid exclusion of a suspect, thus redirecting the investigation.

Rapid DNA instruments are portable, compact instruments designed to be deployed into field testing. These instruments are fully automated for processing reference samples in order to generate a DNA profile in less than 2 h. Several versions of the instruments have recently been developed by manufacturers and some of them are currently commercially available. In addition to rapid DNA instruments, rapid DNA profiling may be achieved through alternative processes. First, it can be achieved through rapid services that offer a quick turnaround time (<2 h). Second,

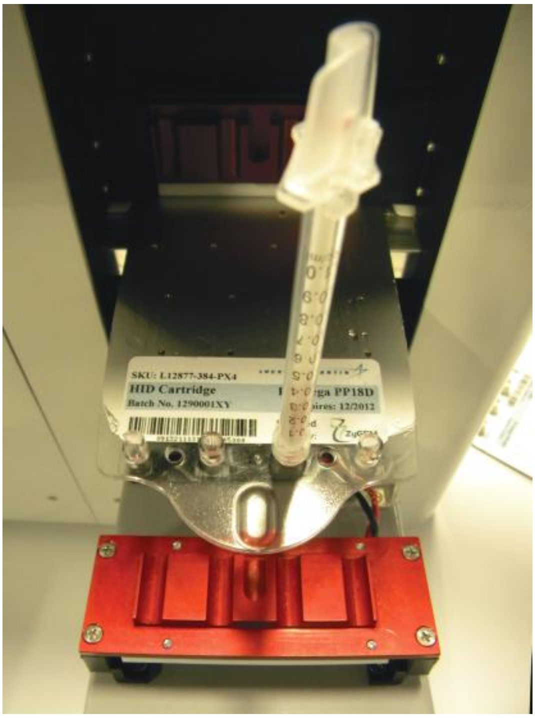

Figure 24.3 A rapid DNA instrument. A compact instrument that is designed to process swab samples and to produce DNA profiles within 2 h in the field. (© Richard C. Li.)

rapid DNA techniques can be applied using standard laboratory equipment but implementing specialized protocols to generate CODIS-compatible profiles in less than 2 h.

The primary goal of the rapid DNA initiative is to produce a CODIS-compatible DNA profile of a sample taken from an arrestee in police custody and to search DNA databases during the booking process in less than 2 h. This technology could be used, for example, to quickly identify other unsolved crimes or eliminate suspects, which can aid in investigative decision making. It can also potentially be useful for other investigations, such as identifying human remains in mass disasters; identifying detainees in counterterrorism applications; and, for immigration and border agents, confirming and verifying individual identifications or family relationships.

24.4 Database Entries

Currently, over 190 public law enforcement laboratories participate in CODIS across the United States. As of April 2013, CODIS contains the DNA profiles of more than 10 million convicted felons, over 1 million arrestees, and half a million crime scene samples. The proportion of the population represented in the database is approximately 3%. The NDIS is one of the largest DNA databases containing DNA profiles in terms of absolute numbers. Each CODIS entry consists of the DNA profile of the sample and the specimen identification number, as well as the information of the laboratory submitting the DNA profile and the laboratory analyst that performed the DNA analysis. The entry does not include case information or the personal information of the offenders or arrestees. Access to CODIS profiles is restricted to criminal justice agencies for law enforcement identification purposes.

The SDIS database retains samples collected from convicted offenders after DNA profiles are obtained. State policies vary in the retention of samples and DNA profiles in situations of dismissed charges, acquittal, or no charges for arrestees. Some states, such as Virginia, destroy the DNA sample and expunge the DNA profile from the database. Other states, such as California, require a petition to destroy the sample and expunge the DNA profile from the database. There are two major reasons for sample retention. First, the sample is needed for confirmatory processes for the purpose of quality assurance and control. In confirmatory processes, the DNA analysis of the original sample is repeated and compared against the prior analysis. Second, the retention of samples allows for possible retesting if new technology becomes available. It also allows retesting of a sample for purposes of updating with expanded loci. Considerable debate has surrounded the retention of samples. It can be argued from the opposite perspective that the database could reveal private genetic information that could then be misused. The objection is based on concern for the protection of privacy rights. If a sample were made available to an unauthorized person, confidential information could be disclosed.

24.5 Database Expansion

Currently, all 50 states have authorized the collection of samples from convicted felons for DNA databasing. Over the years, the demand for the utilization of databases has increased sharply (Figure 24.4). More jurisdictions are incorporating more felonies into the lists of crimes that require DNA profiles, and some jurisdictions plan to include all felonies in such databases. Additionally, the database system is projected to include the profiles of minor criminals, because statistics show that most offenders found guilty of serious crimes were previously convicted for minor crimes. Broadening the size of the database and including samples from more types of crimes could lead to the assumption that the number of crimes solved would also increase. Although state laws vary, each state has its own statute governing the entry criteria of the database samples. More and more states are authorizing the collection of additional types of DNA samples, including individuals convicted of misdemeanor crimes and adult felony arrestees who have not yet been convicted for the offense. It is known that offenders tend to commit multiple

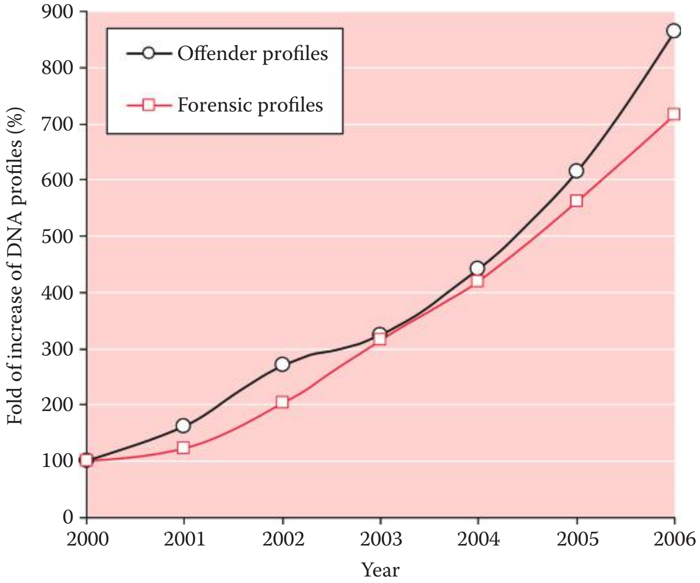

Figure 24.4 Growth of the CODIS database. In 2000, 460,365 offender profiles and 22,484 forensic profiles were entered.

crimes. Thus, including DNA profiles of arrestees in the database is beneficial in catching serial offenders and potentially preventing future crimes. Virginia is the first state to collect DNA from violent-felony arrestees, among other crimes. Over the years, more than half of states in the United States have amended their law to include all felony arrestees or a subset of arrestees for felonies that involve violence or sexual assault.

In 2009, Alonzo King was arrested in Wicomico County, Maryland, on violent assault charges. Maryland's DNA Collection Act authorizes law enforcement agencies to collect DNA samples from an individual who is arrested for violent crimes. During the booking process, under state law, King's buccal swabs were collected for forensic DNA identification. When King's DNA profile was entered into Maryland's SDIS, it matched to a crime scene DNA profile from an unsolved rape case in 2003. Based on the DNA evidence, King was indicted for first-degree rape by a grand jury. The defendant filed a motion to suppress the DNA evidence. King argued that using his DNA to investigate his connection to the 2003 rape was an unreasonable search and seizure under the Fourth Amendment, since the police had no reason to suspect his involvement in the 2003 rape. During the trial at Wicomico County Circuit Court, King's motion was denied. He was convicted of rape and sentenced to life in prison. The Court of Appeals of Maryland, the state's highest court, upheld that the DNA Collection Act was constitutional. The police had probable cause to arrest King on the assault charge, and thus collecting King's DNA sample was a reasonable search. However, the court overturned the low court and reversed King's conviction. It stated that the DNA Collection Act was inappropriately applied to King in this case. Thus, investigating King without probable cause for the unrelated rape was unconstitutional. To determine whether the Fourth Amendment permits states to collect and analyze the DNA of arrestees, a divided Supreme Court ruled 5–4 and reversed the Court of Appeals. The Supreme Court's judgment ruled that Maryland has the right to collect DNA evidence from individuals arrested for serious crimes. The ruling reflects the Supreme Court's view on the balance between the interest of criminal justice systems in solving violent crimes and an individual's interest in the Fourth Amendment, which protects them from unreasonable searches.

Some jurisdictions, such as the United Kingdom, allow DNA samples to be taken from individuals suspected of committing recordable offenses that may lead to prison sentences. In contrast, DNA profiles from suspects are not eligible for entry into CODIS. In some jurisdictions, it has been suggested that databases include many more offenders and suspects, as well as the general public. One advantage of including an entire population in a database is the ability to identify missing, kidnapped, and abducted individuals in addition to victims of major accidents and mass fatalities. Nevertheless, debates concerning the need to balance the benefits and dangers of developing a broader database will inevitably continue.

24.6 DNA Profiles

Currently, the CODIS software supports the storage and search of DNA profiles of short tandem repeat (STR), Y chromosome STR (Y-STR), and mitochondrial DNA (mtDNA). Y-STR and mtDNA profiles may only be searched within NMPDD-related indexes. The CODIS software no longer supports searches of DNA profiles generated by restriction fragment length polymorphism (RFLP) analysis. The 13 core CODIS STR loci are CSF1PO, FGA, THO1, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, and D21S11. DNA profiles are entered into one of the indexes, such as convicted offender, arrestee, forensic, unidentified human remains, missing person, or a relative of a missing person. There is a minimum loci requirement for each DNA profile entering CODIS: 13 core CODIS loci are required for a DNA profile entering the Convicted Offender Index and the Arrestee Index; 13 core CODIS loci and amelogenin, a sex-typing marker, are required for a DNA profile entering the Relatives of Missing Person Index; at least 10 CODIS loci are required (all 13 core loci must be tested) for the Forensic Index; and at least 8 loci and amelogenin are required (all 13 core loci must be tested) for the Missing Person and Unidentified Human Remains Index. The DNA profiles must be generated by an accredited and audited laboratory in accordance with the FBI's Quality Assurance Standards using approved commercially available kits. As the database is growing rapidly, the chances of finding incidental matches among DNA profiles is increasing. To increase discrimination power, it has recently been proposed that CODIS core loci be expanded to include 20 or more markers (Chapter 20). Additionally, the expanded core loci share additional loci compatible with international standards for forensic DNA analysis, which facilitates the exchange of information with international law enforcement agencies.

24.7 Routine Database Searches for Forensic Investigations

The ultimate goal of the database utilization is to provide investigative leads for law enforcement in solving crimes, particularly in cases where no suspect has yet been identified. Currently, DNA profiles uploaded to NDIS are automatically searched once a week (Figure 24.5). A hit is a match made from the information provided by comparing a target DNA profile against the DNA profiles contained in the database. There are two types of CODIS hits: an *offender hit* provides the identity of a potential suspect of a crime, while a *forensic hit* reveals the linkage between two or more crime scenes. Once a hit is identified, the match is verified by the laboratories that originally processed the evidence. A verified CODIS hit can be utilized as probable cause to allow law enforcement to obtain a court order to collect a DNA sample from a suspect. Collecting DNA with a warrant ensures admissibility in court. *Investigation-aided* cases are those assisted by CODIS hits, including case-to-case matches as well as case-to-offender matches. The number of investigations aided is a useful measure of the successful application of the database (Figure 24.6). As of April 2013, over 207,800 hits had been made with CODIS assistance in more than 199,200 investigations.

24.7.1 Case-to-Offender Searches

Matches of profiles from the Forensic Index and the Offender Index reveal the identities of perpetrators of crimes. For example, in 1998, a Florida man, Leon Dundas, became a suspect in a rape case but refused to provide his DNA reference sample for testing. Dundas was killed a year later in an illegal drug deal. Thus, a postmortem sample of Dundas was obtained and his DNA profile was compared with the Forensic Index of CODIS. It was discovered that Dundas's profile

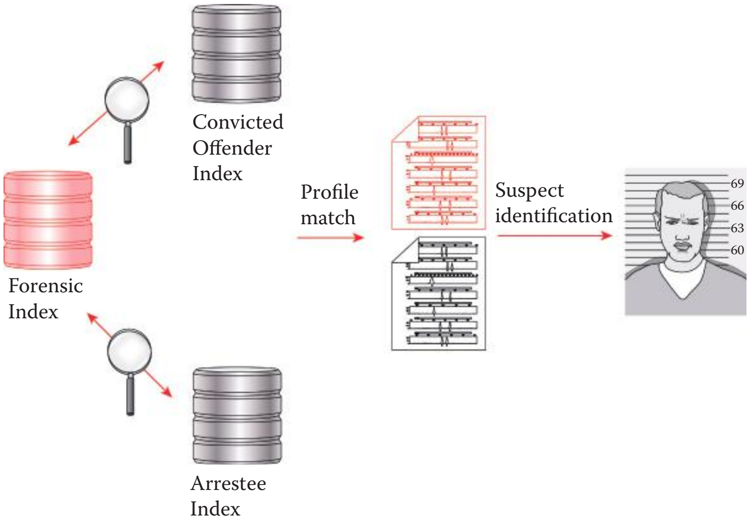

Figure 24.5 Example of a weekly routine CODIS search. (© Richard C. Li.)

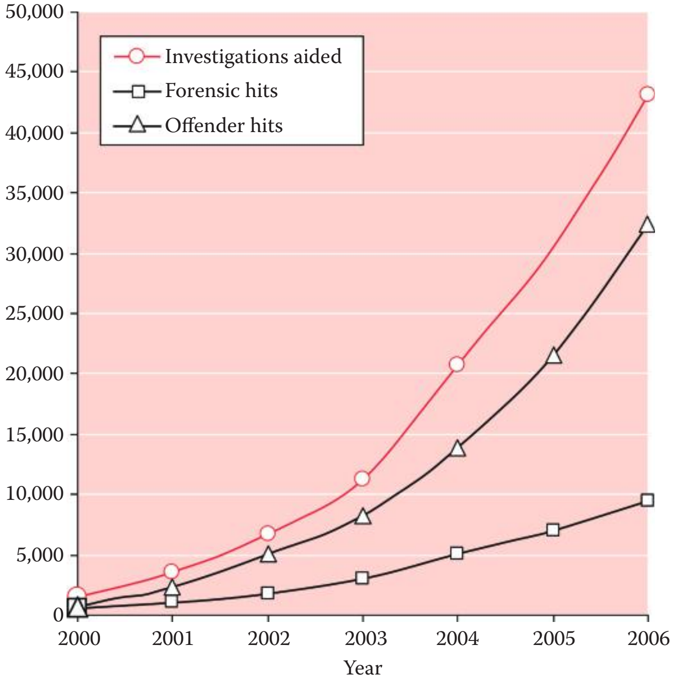

Figure 24.6 Number of hits and investigations aided by the CODIS database.

matched the DNA sample found in the rape case. Additionally, his DNA profile was linked to several other unsolved rapes in Jacksonville and Washington, DC. Such cases can now be solved by utilizing DNA databases.

24.7.2 Case-to-Case Searches

A target DNA profile from a crime scene is also searched against the profiles stored in the Forensic Index of the database. Matches of profiles among the target profile and the profile in the Forensic Index can link separate crime scenes and aid in identifying serial offenders. This helps law enforcement agencies in multiple jurisdictions to coordinate their investigations and share leads. For example, in 1996, two young girls were abducted from bus stops in St. Louis. Both girls were raped and DNA samples were collected. Both DNA profiles pointed to the same perpetrator. In 1999, the St. Louis police decided to reanalyze the samples using new STR technology through the CODIS database. The database found a match to a different rape case, to which the perpetrator, Dominic Moore, had already confessed, thus identifying him as the perpetrator of the 1996 rapes.

24.7.3 Search Stringency and Partial Matches

Database searches are carried out using the CODIS software with three stringency levels, which allow the search of complex forensic profiles against offender profiles (Figure 24.7). A high-stringency match requires an exact match in which all alleles are matched at each locus between the target and candidate profiles. A moderate-stringency match, as defined by the FBI, is a candidate match "between two single source profiles having at each locus all of the alleles of one sample represented in the other sample." In a moderate-stringency match, allelic drop-outs in a target or candidate profile are allowed, possibly resulting in a partial match at some loci. As a result, high-stringency matches are automatically included in a moderate-stringency

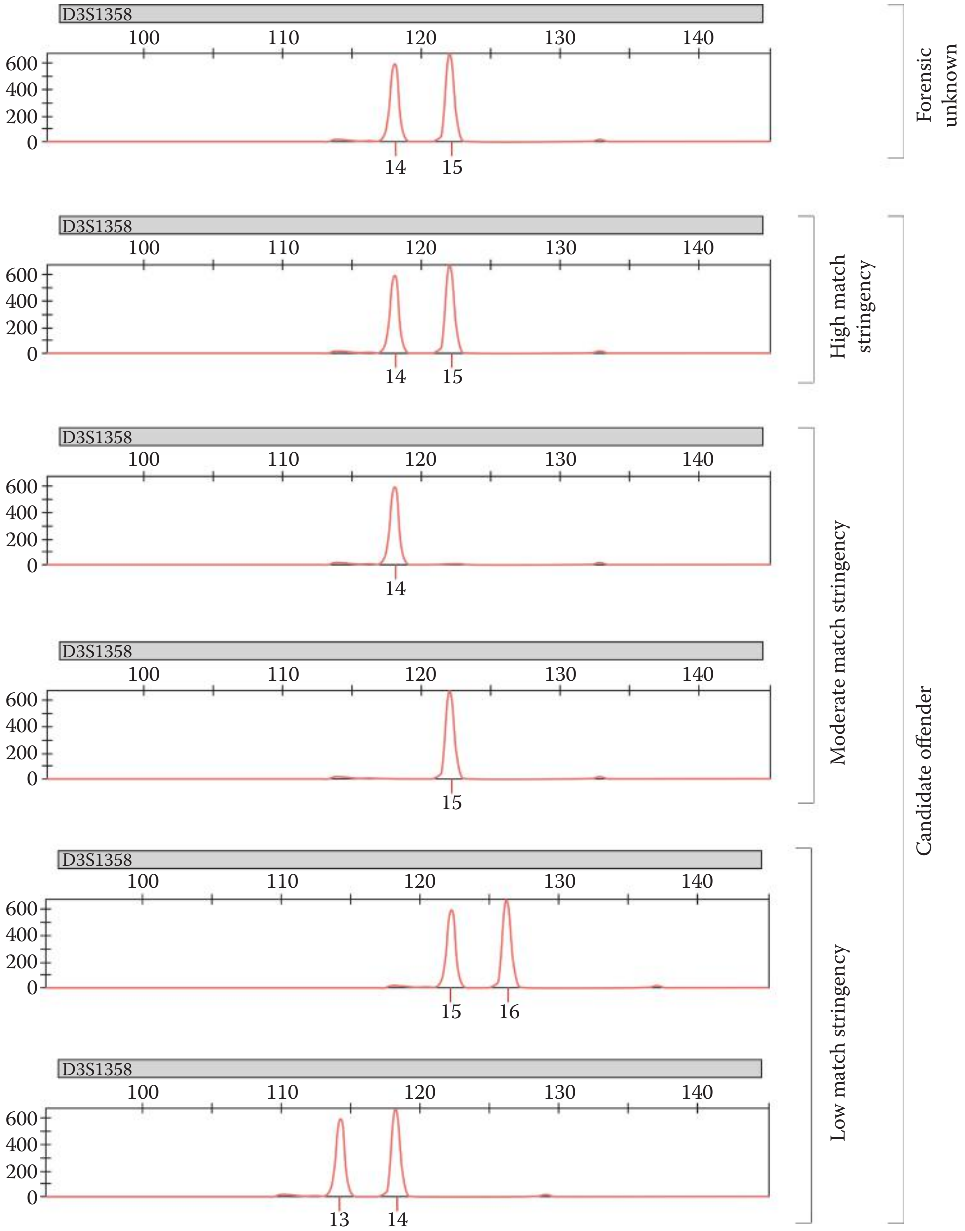

Figure 24.7 Diagram of match stringencies. A heterozygote (allele 14 and 15) at locus D3S1358 is utilized as the genotype for the forensic unknown profile. (© Richard C. Li.)

search. In a low-stringency search, both mismatches and allelic dropouts are allowed. Presently, NDIS searches are only carried out at moderate stringency. During forensic DNA analysis, DNA degradation may prevent full DNA profiles from being processed, producing only partial profiles. Additionally, mixture profiles derived from forensic samples containing DNA contributed by more than one individual may be encountered. Furthermore, due to mutations, null alleles may occur in the profiles produced with some primer sets but not other primers. Therefore, searching at moderate stringency allows the detection of matches under the situations described above.

Such moderate-stringency searches between the Forensic Index and Offender Index may occasionally generate partial match profiles. Since a partial match is not an exact match of the two profiles, further investigation is needed. For instance, additional Y-STR and mtDNA analysis is needed to eliminate unrelated individuals.

24.8 Familial Searches

Familial search, initiated in the United Kingdom, is a new method of applying databases in criminal investigations. It is known that, in the United States, nearly half of prison inmates have close relatives who have also been incarcerated. Familial search is based on the assumption that close relatives share more alleles of DNA profiles than unrelated individuals. Thus, databases may be utilized to identify perpetrators by finding a close relative, if the close relative has been convicted of a crime and is listed in the database. Familial search is an intentional search of a target crime scene profile against an offender database to obtain a list of candidate profiles that are similar to the target profile. This list may include the profile of a close relative of the perpetrator, who is the source of crime scene evidence. These matches most frequently involve siblings, parents, or children. The investigative leads produced by familial searches allow law enforcement to conduct further investigations to identify the perpetrator.

The first familial search leading to a successful prosecution was conducted in Surrey, England. In 2003, a truck driver was killed after a brick was thrown through his windshield from a bridge. The perpetrator's DNA profile was obtained from the brick. A search of the UK's national DNA database revealed no match for the perpetrator. Next, a familial search of the database was conducted. The system identified a close relative that led police to identify the perpetrator, Craig Harman, who was then convicted of manslaughter.

24.8.1 Legal and Ethical Issues of Familial Search

The use of forensic databases involves a balance of individual civil rights and the interests of the criminal justice systems. Many concerns have been raised, including the potential for these searches to violate the privacy of unrelated people whose genetic profiles happen to resemble those of individuals included in the databases. In the United States, the Fourth Amendment protects against unreasonable searches and seizures. The permissibility of familial searching under the Fourth Amendment is yet to be addressed by courts. The collection of biological material for the initial creation of a profile for law enforcement purposes is subject to the Fourth Amendment implications. In terms of familial searches, some legal experts argue that the biological materials are not collected directly from the individuals for a familial search, and that these individuals may thus be protected under the Fourth Amendment. Others argue, however, that the Fourth Amendment protects the initial creation of the profile, including the sample collection, forensic DNA analysis, and the databasing. They argue that the Fourth Amendment may not protect subsequent investigations of DNA profiles during familial searches. In addition, the CODIS database consists of a high percentage of profiles from individuals of racial minority groups, including African Americans. Familial searches can disproportionately focus on a specific racial group.

Familial searches are rare in the United States. Two jurisdictions, Maryland and the District of Columbia, have laws prohibiting the use of familial searches. Familial searches are not conducted at the NDIS. While familial searches are now being performed in several jurisdictions in the United States, policies on familial searches vary among jurisdictions. The major issues relate to criteria for privacy, information release, search approval, and the types of crimes eligible for familial searches. Currently, California, Colorado, Texas, and Virginia have state legislation permitting familial searches. Familial searches are initiated when a specific suspect is not known, and the cause for the search needs to be justified. Familial searches are usually conducted for crimes that pose a substantial threat to public safety, typically those cases involving the most serious offenses. Additionally, familial searches are only conducted after a routine search of a DNA profile has yielded no match in the database.

In 2003, for example, a crime scene DNA sample of a closed rape and murder case, committed decades previously, was reanalyzed using forensic DNA techniques. The crime scene DNA profile was compared to the profile of a man, Darryl Hunt, who was then imprisoned for the crime. Surprisingly, the crime scene DNA profile did not match either that of Hunt or those of convicted felons in the database. However, the database search revealed a close relative of the true perpetrator and thus led law enforcement to identify Willard Brown as the perpetrator. Brown was sentenced to life imprisonment. Additionally, Hunt was exonerated after 18 years in prison.

24.8.2 Familial Search Strategies

The familial search is usually carried out using specially designed software. Although the CODIS software is not designed for familial search, it can be used for familial search through a low-stringency search, which may result in a list of candidate profiles including close relatives of individuals, such as parent–offspring or full-sibling relatives. In a large offender DNA database, similar DNA profiles from unrelated people are often observed due to shared alleles. Therefore, the candidate profile list may also include unrelated individuals whose DNA profiles are similar to the target profile. As a result, a familial search can provide a list of potential candidates consisting of hundreds of profiles, which would be too labor intensive to pursue through further investigations. Several methods can be used to determine a cutoff analytical threshold in order to limit a pool of candidate profiles and exclude unrelated individuals from familial searches (Figure 24.8).

24.8.2.1 Identity-by-State and Kinship Index Method

The *identity-by-state* (IBS) method compares the number of shared alleles and loci between a target forensic profile and the offender profiles in a database but does not take into account allele frequencies. The analytical threshold for a familial search is determined by a preset number of shared alleles or loci in order to prompt further investigation. For example, some states use 15 shared alleles as the analytical threshold to be considered as a candidate, while other states require at least one shared allele for each locus. Additionally, this method can prioritize a pool of candidates based on the highest to the lowest number of shared alleles for investigation.

The *Kinship Index* (KI) method is a likelihood ratio–based method that evaluates the familial match by comparing the probability that two DNA profiles are from related individuals to the probability that they are unrelated. The KI method analyzes the allele frequency data, including all CODIS core loci, to calculate the Combined Kinship Index. The KI may vary based on the allele frequency data across the population. Thus, the accuracy of the KI method relies on the relevance of the population data set analyzed. False inclusions or exclusions may occur if nonrelevant population data are utilized. The KI method also allows the generation of a ranked candidate list according to the probability that the individuals are related. Generally speaking, the accuracy of the KI method is higher than that of the IBS method. However, using both IBS and KI methods is better than using a single method alone for familial search.

24.8.2.2 Focusing on Rare Alleles

During a familial search, target DNA profiles may contain rare alleles with low allelic frequencies. The chance of sharing rare alleles for two close relatives is higher than for two unrelated individuals. If a target profile carries a rare allele, it can be used to narrow the pool of candidates.

For example, Jeffrey Gafoor, then 23, had been living in a suburban neighborhood in Cardiff, Wales. He had a reputation for being a loner. Besides working in his family-owned shop, he spent most of his time at home. On February 13, 1988, Gafoor entered 7 James Street in the Butetown area of Cardiff. The first-floor unit was temporarily occupied by Lynette White, a 20-year-old

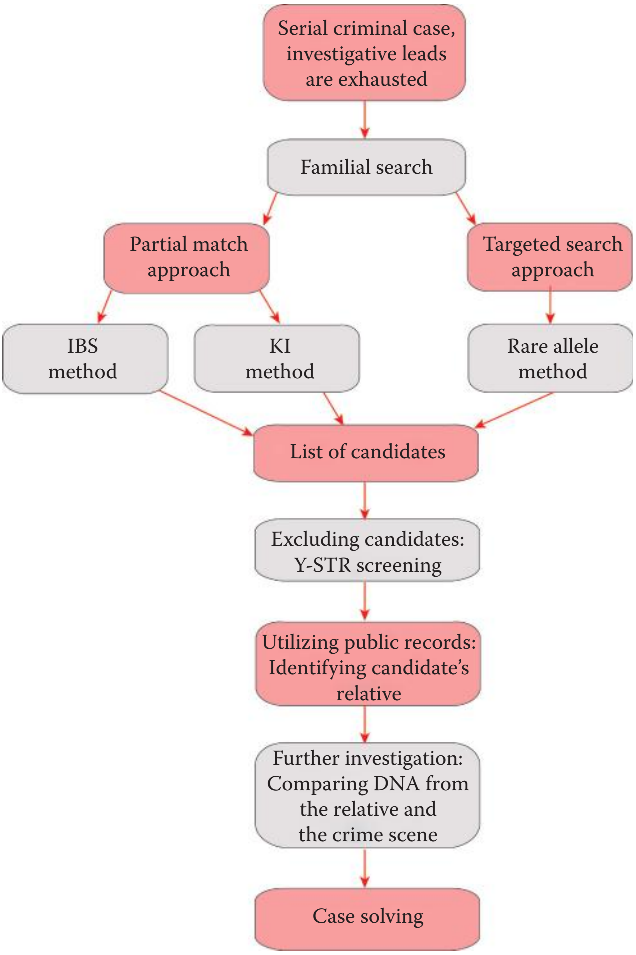
Serial criminal case, investigative leads are exhausted
Familial search
Partial match approach
Targeted search approach
IBS method
KI method
Rare allele method
List of candidates
Excluding candidates: Y-STR screening
Utilizing public records: Identifying candidate's relative
Further investigation: Comparing DNA from the relative and the crime scene
Case solving

Figure 24.8 Example of a familial search workflow. (© Richard C. Li.)

woman who was working as a prostitute in the area. Gafoor went there for the purpose of receiving sexual services. He left the scene in the early hours of Valentine's Day. On Valentine's night, White's body was discovered. She suffered from more than 50 stab, cut, and slash wounds as well as defensive wounds on her hands. The body was dragged to the corner of the room adjacent to the bed, the only piece of furniture in the room.

Gafoor left blood and semen evidence at the scene. He bled during the murder and deposited bloodstains at the scene. Additionally, he left seminal stains at the scene; the semen had no measurable level of sperm and most likely came from an individual who had a medical condition known as azoospermia (Chapter 14). After the murder, Gafoor lived at the same place as he did beforehand and kept the same job for many years.

Two years later, three local men, known as the "Cardiff Three," were convicted of White's murder and sentenced to life in prison. In 1992, however, their convictions were reversed by the court of appeal. Subsequently, the crime scene evidence was tested using new techniques in forensic DNA analysis that were not available in 1988. A DNA profile was obtained from the crime scene. A familial search revealed a similar profile from a 14-year-old boy in the UK National DNA Database. In particular, there was a rare allele match between the crime scene and the candidate DNA profiles. This search led to the further investigation of Gafoor, who was the uncle of the boy. Gafoor was charged with murder.

24.8.2.3 Excluding Candidates through Y-STR Screening

The vast majority of DNA profiles in CODIS come from males. Patrilineal markers, such as Y-STR loci, are successfully utilized to screen the candidates to verify the relationship between two individuals. This analysis helps to identify first-order relatives as well as paternal half-siblings. STR loci beyond the CODIS core loci and maternal lineage makers, such as mitochondrial DNA typing systems, can be used to narrow the pool of candidates and eliminate coincidental partial matches. One or more of these methods can be incorporated to evaluate highly ranked candidates.

For example, Lonnie Franklin lived in a southern Los Angeles neighborhood. Throughout the 1980s, Franklin worked for the city as a maintenance worker and a sanitation truck operator. His neighbors recalled that he often brought prostitutes to a camper parked in the backyard of his house. He also took photos of nude women, which he kept in his garage. Nevertheless, to many of his neighbors, Franklin seemed to be a friendly person who often chatted with them.

Since 1985, the bodies of many women, most of which were prostitutes, had been dumped in an alley running along Western Avenue in a southern Los Angeles neighborhood. The victims had been shot and some had been strangled after sexual contact. DNA and ballistics analysis revealed that at least 10 of the murders had been committed by the same perpetrator. The perpetrator was known to the general public as the “Grim Sleeper” because he had taken a 14-year break during the period in which these 10 murders had taken place. The crime scene DNA profile did not match any profile in the database.

In 2008, the first familial search of the “Grim Sleeper” case was conducted by the California Department of Justice. California state law allows familial searches for high-profile cases if all other leads have been exhausted. However, all candidate DNA profiles were excluded through Y-STR screening. As a result, no candidate was identified as a possible relative of the “Grim Sleeper.” In 2010, a second search targeted partial matches that shared at least 15 alleles. It produced a list of 100 candidate profiles ranked by the likelihood that they were related. Y-STR was used as a screen tool to identify possible patrilineal relatives. This time, the search generated a match to a new DNA profile of a felon, which had recently been entered into California’s SDIS, in 2009. The search suggested that the candidate was a relative of the source of the DNA from the crime scenes. This search led to the identification of Franklin as a suspect in the “Grim Sleeper” murders. Police detectives conducted surveillance and collected a discarded slice of pizza and a cup used by Franklin at a local restaurant for DNA identification. The DNA profile was identical to the crime scene DNA profiles. Franklin was arrested and was charged with ten counts of murder and one of attempted murder.

Bibliography

Alonso, A., et al., Challenges of DNA profiling in mass disaster investigations. *Croat Med J*, 2005, **46**(4): 540–548.

Aoki, Y., et al., Comparison of the likelihood ratio and identity-by-state scoring methods for analyzing sib-pair test cases: A study using computer simulation. *Tohoku J Exp Med*, 2001, **194**(4): 241–250.

Asplen, C. and S.A. Lane, International perspectives on forensic DNA databases. *Forensic Sci Int*, 2004, **146**(Suppl): S119–S121.

Baker, L.E. and E.J. Baker, Reuniting families: An online database to aid in the identification of undocumented immigrant remains. *J Forensic Sci*, 2008, **53**(1): 50–53.

Balding, D.J., The DNA database search controversy. *Biometrics*, 2002, **58**(1): 241–244.

Balding, D.J., et al., Decision-making in familial database searching: KI alone or not alone? *Forensic Sci Int Genet*, 2013, **7**(1): 52–54.

Barash, M., et al., A search for obligatory paternal alleles in a DNA database to find an alleged rapist in a fatherless paternity case. *J Forensic Sci*, 2012, **57**(4): 1098–1101.

Barber, M.D. and B.H. Parkin, Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179. *Int J Legal Med*, 1996, **109**(2): 62–65.

Benschop, C.C., et al., Low template STR typing: Effect of replicate number and consensus method on genotyping reliability and DNA database search results. *Forensic Sci Int Genet*, 2011, **5**(4): 316–328.

Bereano, P.L., DNA identification systems: Social policy and civil liberties concerns. *J Int Bioethique*, 1990, **1**(3): 146–155.

Bieber, F.R. and D. Lazer, Guilt by association: Should the law be able to use one person's DNA to carry out surveillance on their family? Not without a public debate. *New Sci*, 2004, **184**(2470): 20.

Bieber, F.R., Turning base hits into earned runs: Improving the effectiveness of forensic DNA data bank programs. *J Law Med Ethics*, 2006, **34**(2): 222–233.

Bingham, R., A database of the innocent? *Splice Life*, 2001, **7**(2–3): 8–9.

Birmingham, K., UK leads on science and technology bioethics. *Nat Med*, 1998, **4**(6): 651–652.

Bradford, L., et al., Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0. *Forensic Sci Int Genet*, 2011, **5**(4): 291–296.

Buckleton, J., J.A. Bright, and S.J. Walsh, Database crime to crime match rate calculation. *Forensic Sci Int Genet*, 2009, **3**(3): 200–201.

Budowle, B., et al., CODIS and PCR-based short tandem repeat loci: Law enforcement tools, in *Second European Symposium on Human Identification*, 1998. Innsbruck, Austria: Promega Corporation.

Burk, D.L., DNA identification testing: Assessing the threat to privacy. *Univ Toledo Law Rev*, 1992, **24**(1): 87–102.

Burk, D.L. and J.A. Hess, Genetic privacy: Constitutional considerations in forensic DNA testing. *Geoge Mason Univ Civ Rights Law J*, 1994, **5**(1): 1–53.

Cash, H.D., J.W. Hoyle, and A.J. Sutton, Development under extreme conditions: Forensic bioinformatics in the wake of the World Trade Center disaster. *Pac Symp Biocomput*, 2003, **8**: 638–653.

Chow-White, P.A. and T. Duster, Do health and forensic DNA databases increase racial disparities? *PLoS Med*, 2011, **8**(10): e1001100.

Chung, Y.K. and W.K. Fung, Identifying contributors of two-person DNA mixtures by familial database search. *Int J Legal Med*, 2013, **127**(1): 25–33.

Chung, Y.K., Y.Q. Hu, and W.K. Fung, Evaluation of DNA mixtures from database search. *Biometrics*, 2010, **66**(1): 233–238.

Corte-Real, F., Forensic DNA databases. *Forensic Sci Int*, 2004, **146**(Suppl): S143–S144.

Curran, J.M. and J.S. Buckleton, Effectiveness of familial searches. *Sci Justice*, 2008, **48**(4): 164–167.

Dawid, A.P., Comment on Stockmarr's "Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search". *Biometrics*, 2001, **57**(3): 976–980.

Duster, T., Explaining differential trust of DNA forensic technology: Grounded assessment or inexplicable paranoia? *J Law Med Ethics*, 2006, **34**(2): 293–300.

Evett, I.W., J. Scranage, and R. Pinchin, Efficient retrieval from DNA databases: Based on the second European DNA Profiling Group collaborative experiment. *Forensic Sci Int*, 1992, **53**(1): 45–50.

Evison, M., DNA database could end problem of identity fraud. *Nature*, 2002, **420**(6914): 359.

Gabriel, M., C. Boland, and C. Holt, Beyond the cold hit: Measuring the impact of the national DNA data bank on public safety at the city and county level. *J Law Med Ethics*, 2010, **38**(2): 396–411.

Gamero, J.J., et al., A study of Spanish attitudes regarding the custody and use of forensic DNA databases. *Forensic Sci Int Genet*, 2008, **2**(2): 138–149.

Ge, J., et al., Comparisons of familial DNA database searching strategies. *J Forensic Sci*, 2011, **56**(6): 1448–1456.

Gershaw, C.J., et al., Forensic utilization of familial searches in DNA databases. *Forensic Sci Int Genet*, 2011, **5**(1): 16–20.

Gill, P., et al., Databases, quality control and interpretation of DNA profiling in the hHome Office fForensic sScience sService. *Electrophoresis*, 1991, **12**(2–3): 204–209.

Gill, P., et al., A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations. *Forensic Sci Int*, 2003, **131**(2–3): 184–196.

Gill, P., et al., The evolution of DNA databases—Recommendations for new European STR loci. *Forensic Sci Int*, 2006, **156**(2–3): 242–244.

Gosline, A., Will DNA profiling fuel prejudice? *New Sci*, 2005, **186**(2494): 12–13.

Guillen Vazquez, M., C. Pestoni, and A. Carracedo, DNA databases for criminal investigation purposes: Technical aspects and ethical–legal problems. *Law Hum Genome Rev*, 1998, (8): 137–158.

Guillen, M., et al., Ethical-legal problems of DNA databases in criminal investigation. *J Med Ethics*, 2000, **26**(4): 266–271.

Haimes, E., Social and ethical issues in the use of familial searching in forensic investigations: Insights from family and kinship studies. *J Law Med Ethics*, 2006, **34**(2): 263–276.

Harbison, S.A., J.F. Hamilton, and S.J. Walsh, The New Zealand DNA databank: Its development and significance as a crime solving tool. *Sci Justice*, 2001, **41**(1): 33–37.

Harding, H.W. and R. Swanson, DNA database size. *J Forensic Sci*, 1998, **43**(1): 248–249.

Hedman, J., et al., A fast analysis system for forensic DNA reference samples. *Forensic Sci Int Genet*, 2008, **2**(3): 184–189.

Heninger, M. and R. Hanzlick, The Fulton County Medical Examiner's experience with the Federal Bureau of iInvestigation National Missing Person DNA Database Program, 2004–2007. *Am J Forensic Med Pathol*, 2011, **32**(1): 78–82.

Hepple, B., Forensic databases: Implications of the cases of S and Marper. *Med Sci Law*, 2009, **49**(2): 77–87.

Herkenham, M.D., Retention of offender DNA samples necessary to ensure and monitor quality of forensic DNA efforts: Appropriate safeguards exist to protect the DNA samples from misuse. *J Law Med Ethics*, 2006, **34**(2): 380–384.

Hibbert, M., DNA databanks: Law enforcement's greatest surveillance tool? *Wake Forest Law Rev*, 1999, **34**(3): 767–825.

Hicks, T., et al., Use of DNA profiles for investigation using a simulated national DNA database: Part I. Partial SGM Plus profiles. *Forensic Sci Int Genet*, 2010, **4**(4): 232–238.

Hicks, T., et al., Use of DNA profiles for investigation using a simulated national DNA database: Part II. Statistical and ethical considerations on familial searching. *Forensic Sci Int Genet*, 2010, **4**(5): 316–322.

Hoyle, R., Forensics. The FBI's national DNA database. *Nat Biotechnol*, 1998, **16**(11): 987.

Huang, D., et al., Y-haplotype screening of local patrilineages followed by autosomal STR typing can detect likely perpetrators in some populations. *J Forensic Sci*, 2011, **56**(5): 1340–1342.

Jobling, M.A. and Gill, P., Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*, 2004, **5**: 739–751.

Karjala, D.S., A legal research agenda for the human genome initiative. *Jurimetrics*, 1992, **32**(2): 121–222.

Kim, J., et al., Policy implications for familial searching. *Investig Genet*, 2011, **2**: 22.

Kimel, C.W., DNA profiles, computer searches, and the Fourth Amendment. *Duke Law J*, 2013, **62**(4): 933–973.

Kirby, R., The many dangers of relying on a DNA database. *Nature*, 2002, **419**(6904): 247.

Lakra, R. and F. Dimitriadis, Forensic DNA data banks: Considerations for the health sector. *Health Law Can*, 2002, **22**(3): 57–70.

Lehrman, S., Prisoner's DNA database ruled unlawful. *Nature*, 1998, **394**(6696): 818.

Lehrman, S., Partial to crime: Families become suspects as rules on DNA matches relax. *Sci Am*, 2006, **295**(6): 28–29.

Levitt, M., Forensic databases: Benefits and ethical and social costs. *Br Med Bull*, 2007, **83**: 235–248.

Linacre, A., The UK National DNA Database. *Lancet*, 2003, **361**(9372): 1841–1842.

Love, S.H., Allowing new technology to erode constitutional protections: A Fourth Amendment challenge to non-consensual DNA testing of prisoners. *Villanova Law Rev*, 1993, **38**(5): 1617–1660.

Machado, H., F. Santos, and S. Silva, Prisoners' expectations of the national forensic DNA database: Surveillance and reconfiguration of individual rights. *Forensic Sci Int*, 2011, **210**(1–3): 139–143.

Maguire, C.N., et al., Familial searching: A specialist forensic DNA profiling service utilising the national DNA database to identify unknown offenders via their relatives—The UK experience. *Forensic Sci Int Genet*, 2014, **8**(1): 1–9.

Mansel, C. and S. Davies, Minors or suspects? A discussion of the legal and ethical issues surrounding the indefinite storage of DNA collected from children aged 10–18 years on the national DNA database in England and Wales. *Med Sci Law*, 2012, **52**(4): 187–192.

Marjanovic, D., et al., Forensic DNA databases in Western Balkan region: Retrospectives, perspectives, and initiatives. *Croat Med J*, 2011, **52**(3): 235–244.

Martin, P.D., H. Schmitter, and P.M. Schneider, A brief history of the formation of DNA databases in forensic science within Europe. *Forensic Sci Int*, 2001, **119**(2): 225–231.

McEwen, J.E. and P.R. Reilly, A review of state legislation on DNA forensic data banking. *Am J Hum Genet*, 1994, **54**(6): 941–958.

McEwen, J.E., Forensic DNA data banking by state crime laboratories. *Am J Hum Genet*, 1995, **56**(6): 1487–1492.

McNamara, J.J. and R.J. Morton, Frequency of serial sexual homicide victimization in Virginia for a 10-year period. *J Forensic Sci*, 2004, **49**(3): 529–533.

Meester, R. and M. Sjerps, The evidential value in the DNA database search controversy and the two-stain problem. *Biometrics*, 2003, **59**(3): 727–732.

Melson, K.E., President's editorial—The journey to justice. *J Forensic Sci*, 2003, **48**(4): 705–707.

Mohanty, N.K. and S. Haldar, DNA test—A forensic boon. *J Indian Med Assoc*, 2006, **104**(2): 90–92, 94.

Murphy, E., The government wants your DNA. *Sci Am*, 2013, **308**(3): 72–77.

Myers, S.P., et al., Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach. *Forensic Sci Int Genet*, 2011, **5**(5): 493–500.

Nerko, C.J., Assessing Fourth Amendment challenges to DNA extraction statutes after Samson v. California. *Fordham Law Rev*, 2008, **77**(2): 917–949.

Nishimi, R.Y., et al., Genetic witness: Forensic uses of DNA tests. *J Int Bioethique*, 1991, **2**(1): 29–32.

Oz, C., M. Amiel, and A. Zamir, The Israel police DNA database: Recognizing the capability of databases in providing investigative leads. *Forensic Sci Int Genet*, 2012, **6**(1): e8–e10.

Parson, W. and M. Steinlechner, Efficient DNA database laboratory strategy for high through-put STR typing of reference samples. *Forensic Sci Int*, 2001, **122**(1): 1–6.

Pascali, V.L., G. Lago, and M. Dobosz, The dark side of the UK National DNA Database. *Lancet*, 2003, **362**(9386): 834.

Patyn, A. and K. Dierickx, Forensic DNA databases: Genetic testing as a societal choice. *J Med Ethics*, 2010, **36**(5): 319–320.

Peerenboom, E., Central criminal DNA database created in Germany. *Nat Biotechnol*, 1998, **16**(6): 510–511.

Pham-Hoai, E., F. Crispino, and G. Hampikian, The first successful use of a low stringency familial match in a French criminal investigation. *J Forensic Sci*, 2014, **59**(3): 816–819.

Raab, S., Cuomo seeks genetic data of offenders. *N Y Times*, May 10, 1992.

Reid, T.M., et al., Use of sibling pairs to determine the familial searching efficiency of forensic databases. *Forensic Sci Int Genet*, 2008, **2**(4): 340–342.

Reilly, P., Rights, privacy, and genetic screening. *Yale J Biol Med*, 1991, **64**(1): 43–45.

Rohlfs, R.V., S.M. Fullerton, and B.S. Weir, Familial identification: Population structure and relationship distinguishability. *PLoS Genet*, 2012, **8**(2): e1002469.

Rohlfs, R.V., et al., The influence of relatives on the efficiency and error rate of familial searching. *PLoS One*, 2013, **8**(8): e70495.

Rossy, Q., et al., Integrating forensic information in a crime intelligence database. *Forensic Sci Int*, 2013, **230**(1–3): 137–146.

Sankar, P., The proliferation and risks of government DNA databases. *Am J Public Health*, 1997, **87**(3): 336–337.

Scheck, B., DNA data banking: A cautionary tale. *Am J Hum Genet*, 1994, **54**(6): 931–933.

Schneider, P.M. and P.D. Martin, Criminal DNA databases: The European situation. *Forensic Sci Int*, 2001, **119**(2): 232–238.

Scott, G., M.F. Marshall, and C. Tindall, Of doctors and detectives. *N Y Times Web*, November 20, 2001, F7.

Shapiro, E.D. and M.L. Weinberg, DNA data banking: The dangerous erosion of privacy. *Clevel State Law Rev*, 1990, **38**(3): 455–486.

Shriver, M., T. Frudakis, and B. Budowle, Getting the science and the ethics right in forensic genetics. *Nat Genet*, 2005, **37**(5): 449–450; author reply 450–451.

Simoncelli, T., Dangerous excursions: The case against expanding forensic DNA databases to innocent persons. *J Law Med Ethics*, 2006, **34**(2): 390–397.

Sjerps, M. and R. Meester, Selection effects and database screening in forensic science. *Forensic Sci Int*, 2009, **192**(1–3): 56–61.

Smith, J.C., The precarious implications of DNA profiling. *Univ Pittsbg Law Rev*, 1994, **55**(3): 865–888.

Song, Y.S., et al., Average probability that a "cold hit" in a DNA database search results in an erroneous attribution. *J Forensic Sci*, 2009, **54**(1): 22–27.

Struse, H.M. and I.D. Montoya, Health services implications of DNA testing. *Clin Lab Sci*, 2001, **14**(4): 247–251.

Sullivan, K., et al., New developments and challenges in the use of the UK DNA database: Addressing the issue of contaminated consumables. *Forensic Sci Int*, 2004, **146**(Suppl): S175–S176.

Swinbanks, D., UK plans library of DNA profiles. *Nature*, 1993, **365**(6447): 596.

Taborda, J.G. and J. Arboleda-Florez, Forensic medicine in the next century: Some ethical challenges. *Int J Offender Ther Comp Criminol*, 1999, **43**(2): 188–201.

Taylor, D., et al., An investigation of admixture in an Australian Aboriginal Y-chromosome STR database. *Forensic Sci Int Genet*, 2012, **6**(5): 532–538.

Terry, S.F. and P.F. Terry, A consumer perspective on forensic DNA banking. *J Law Med Ethics*, 2006, **34**(2): 408–414.

Tvedebrink, T., et al., Analysis of matches and partial-matches in a Danish STR data set. *Forensic Sci Int Genet*, 2012, **6**(3): 387–392.

Van Camp, N. and K. Dierickx, The retention of forensic DNA samples: A socio-ethical evaluation of current practices in the EU. *J Med Ethics*, 2008, **34**(8): 606–610.

Voultsos, P., et al., Launching the Greek forensic DNA database. The legal framework and arising ethical issues. *Forensic Sci Int Genet*, 2011, **5**(5): 407–410.

Walsh, S.J., J.M. Curran, and J.S. Buckleton, Modeling forensic DNA database performance. *J Forensic Sci*, 2010, **55**(5): 1174–1183.

Walsh, S.J., et al., The collation of forensic DNA case data into a multi-dimensional intelligence database. *Sci Justice*, 2002, **42**(4): 205–214.

Werrett, D.J., The national DNA database. *Forensic Sci Int*, 1997, **88**(1): 33–42.

Williams, R. and P. Johnson, "Wonderment and dread": Representations of DNA in ethical disputes about forensic DNA databases. *New Genet Soc*, 2004, **23**(2): 205–223.

Williams, R. and P. Johnson, Inclusiveness, effectiveness and intrusiveness: Issues in the developing uses of DNA profiling in support of criminal investigations. *J Law Med Ethics*, 2005, **33**(3): 545–558.

Zamir, A., et al., The Israel DNA database—The establishment of a rapid, semi-automated analysis system. *Forensic Sci Int Genet*, 2012, **6**(2): 286–289.

Zimmermann, B., et al., Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection. *Forensic Sci Int Genet*, 2011, **5**(2): 133–137.

25

Evaluation of the Strength of Forensic DNA Profiling Results

25.1 A Review of Basic Principles of Genetics

25.1.1 Mendelian Genetics

Mendel's first law is the principle of segregation of alleles. Each pair of alleles segregates from others in the formation of gametes (mature reproductive cells such as spermatozoa and oocytes). One-half of the gamete carries one allele, and the other half carries the other allele.

Mendel's second law is the principle of independent assortment of alleles. The segregation of each pair of alleles is independent of the segregation of other pairs during the formation of gametes.

Gametes are formed during a process known as *meiosis*, in which cells with haploid chromosome numbers (23 in humans) are produced by the division of cells with diploid chromosome numbers (46 in humans). A fertilized human egg thus contains a diploid number of chromosomes. A diploid is composed of 22 pairs of autosomes and 2 sex chromosomes (XX in females and XY in males).

Based on Mendelian principles, genes on different chromosomes assort independently of one another in gamete production. Genes residing very closely together on the same chromosome are usually inherited together. Thus, they do not assort independently and are called *linked genes*. Genes distant from each other on the same chromosome are usually inherited separately. This results from an exchange of segments between a pair of *homologous* chromosomes when the chromosomes are paired during the early phases of meiotic division. These types of gene-exchange events on homologous chromosomes are collectively called *crossing over*, which results in the *recombination* of genes in a pair of chromosomes (Figure 25.1).

The Mendelian inheritances of genes can often be measured using *probabilities*. A probability is the ratio of the number of actual occurrences of an event to the number of possible occurrences. Additionally, the probability of two independent events occurring simultaneously is the product of each of their individual probabilities; this is known as the *product rule of probability*.

Mendelian principles apply to the inheritance of loci of the autosomal nuclear DNA genome commonly used for forensic testing. The Y-chromosomal genome is inherited paternally, which does not obey the rules for Mendelian principles. Mitochondria contain their own mitochondrial genomes and are inherited maternally. This maternal inheritance

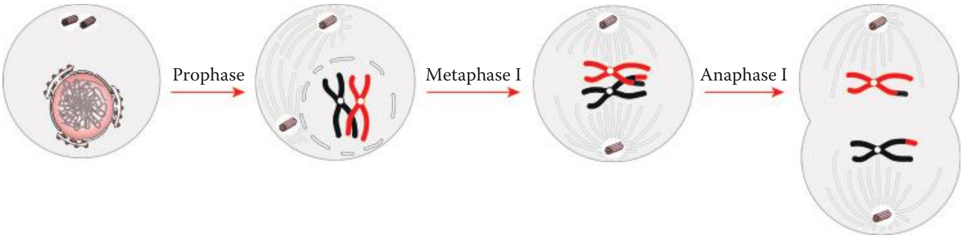

Figure 25.1 Crossing over hypothesis. Chromosomes are replicated prior to the first meiotic division. Each duplicate is called a chromatid, which forms into tetrads in the prophase. A crossover event occurs between the maternal and the paternal chromatids. Each chromatid breaks at the point of the cross and fuses with a portion of its counterpart. Chromosomes are associated with the microtubules during metaphase I and are pulled toward the spindle poles during anaphase I. The maternal and paternal homologs of one chromosome are shown. (© Richard C. Li.)

of mitochondrial genes also does not obey the rules for Mendelian principles. The inheritance of Y-chromosomal and mitochondrial genomes is referred to as non-Mendelian inheritance.

25.1.2 Population Genetics

Population genetics studies the causes of patterns of genetic variations within and among populations.

25.1.2.1 Allele Frequency

Allele frequency (p) can be calculated directly by counting the number of alleles of one type at a given locus and dividing it by the total number of alleles at that locus in a sampled population. This is called the *gene counting* method.

25.1.2.2 Genotype Frequency

Genotype frequency (P) observed at a given locus can be calculated by dividing the number of individuals with one genotype by the total number of individuals in a sampled population. Each genotype at the locus can be calculated separately. The summation of all genotype frequencies at that locus should equal 1.

25.1.2.3 Heterozygosity

Heterozygosity is the proportion of alleles, at a given locus, that are heterozygous and is calculated as

$$h = 1 - \sum_{i} p^2 \tag{25.1}$$

where:

h = heterozygosity
p = allelic frequency of the locus for homozygotes

The amount of heterozygosity at a locus in a sampled population is a measure of genetic variation. The higher the heterozygosity, the more variation there is at a given locus.

25.1.2.4 Hardy–Weinberg Principle

The Hardy–Weinberg (HW) principle, independently discovered by two scholars in the early 1900s, allows predictions of genotype frequencies to be made based on allelic frequencies. However, certain conditions must be met. The population must be large; mate randomly; and lack mutation, migration, and natural selection. If these conditions of the HW principle are met, the population will be in *equilibrium*, and the following results are expected:

1. The frequencies of the alleles will not change from one generation to the next.
2. Genotype frequencies can be predicted by the allelic frequencies (p^2 and q^2 for genotype frequencies of homozygotes and $2pq$ for genotype frequencies of heterozygotes). The sum of the genotype frequencies should equal 1.

$$p^2 + 2pq + q^2 = 1 \tag{25.2}$$

If the observed genotype proportions are different from those expected, one or more of the HW assumptions have been violated.

25.1.2.5 Testing for HW Proportions of Population Databases

To determine whether the genotypes of a population in question obey the HW principle, a population database can be constructed. Samples (usually 100–200 samples for STR loci) are collected and analyzed at the loci of interest. Allelic frequencies are obtained by using the gene counting method. Table 25.1 shows the allelic frequencies of CODIS loci from a population database. The *observed genotype frequencies* at a given locus, as described earlier in Section 25.1.2.2, are calculated by dividing the number of individuals with one genotype by the total number of individuals in the population sampled. The *expected genotype frequencies* are calculated using p^2, $2pq$, and q^2.

The observed and expected genotype frequencies are then compared using a chi-square test. The significance of the differences between observed and expected genotype frequencies can then be determined. The chi-square is calculated using the following formula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{25.3}$$

where:

O_i = ith observed genotype frequency
E_i = ith expected genotype frequency
n = total number of genotypes

Chi-square (χ^2) is calculated as the sum of all genotypes of a given locus.

The chi-square value and the degrees of freedom (the number of genotypes minus the numbers of alleles) are then used to obtain a p value (not to be confused with the allelic frequency designated p) from a table of p values, and such tables can be found at the backs of most statistics textbooks. If the p value exceeds 0.05 (5% significance level), the deviation of the expected genotype frequencies from the observed genotype frequencies is not considered statistically significant. Thus, the null hypothesis that the observed genotype frequencies fit the expected genotype frequencies predicted by the HW principle is not rejected if the p value is greater than 0.05.

Table 25.1 Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
D3S1358	(*N*=210)	(*N*=203)	(*N*=209)
<12	0.476	0.000	0.000
12	0.238	0.000	0.000
13	1.190	0.246	0.239
14	12.143	14.039	7.895
15	29.048	24.631	42.584
15.2	0.000	0.000	0.000
16	30.714	23.153	26.555
17	20.000	21.182	12.679
18	5.476	16.256	8.373
19	0.476	0.493	1.435
>19	0.238	0.000	0.239
VWA	(*N*=180)	(*N*=196)	(*N*=203)
11	0.278	0.000	0.246
13	0.556	0.510	0.000
14	6.667	10.204	6.158
15	23.611	11.224	7.635
16	26.944	20.153	35.961
17	18.333	26.276	22.167
18	13.611	22.194	19.458
19	7.222	8.418	7.143
20	2.778	1.020	1.232
21	0.000	0.000	0.000
FGA	(*N* = 180)	(*N* = 196)	(*N* = 203)
<18	0.278	0.000	0.000
18	0.833	3.061	0.246
18.2	0.833	0.000	0.000
19	5.278	5.612	7.882
19.2	0.278	0.000	0.000
20	7.222	14.541	7.143
20.2	0.000	0.255	0.246
21	12.500	17.347	13.054

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
21.2	0.000	0.000	0.246
22	22.500	18.878	17.734
22.2	0.556	1.020	0.493
22.3	0.000	0.000	0.000
23	12.500	15.816	14.039
23.2	0.000	0.000	0.739
24	18.611	13.776	12.562
24.2	0.000	0.000	0.000
24.3	0.000	0.000	0.000
25	10.000	6.888	13.793
26	3.611	1.786	8.374
27	2.222	1.020	3.202
28	1.667	0.000	0.246
29	0.556	0.000	0.000
30	0.278	0.000	0.000
>30	0.278	0.000	0.000
D8S1179	($N = 180$)	($N = 196$)	($N = 203$)
<9	0.278	1.786	0.246
9	0.556	1.020	0.246
10	2.500	10.204	9.360
11	3.611	5.867	6.158
12	10.833	14.541	12.069
13	22.222	33.929	32.512
14	33.333	20.153	24.631
15	21.389	10.969	11.576
16	4.444	1.276	2.463
17	0.833	0.255	0.739
18	0.000	0.000	0.000
D21S11	($N = 179$)	($N = 196$)	($N = 203$)
24.2	0.279	0.510	0.246
24.3	0.000	0.000	0.000

(*continued*)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
26	0.279	0.000	0.000
27	6.145	4.592	0.985
28	21.508	16.582	6.897
29	18.994	18.112	20.443
29.2	0.279	0.000	0.246
30	17.877	23.214	33.005
30.2	0.838	3.827	3.202
30.3	0.000	0.000	0.000
31	9.218	7.143	6.897
31.2	7.542	9.949	8.621
32	0.838	1.531	1.232
32.1	0.000	0.000	0.000
32.2	6.983	11.224	13.547
33	0.838	0.000	0.000
33.2	3.352	3.061	4.187
34	0.838	0.000	0.000
34.2	0.279	0.000	0.493
35	2.793	0.000	0.000
35.2	0.000	0.255	0.000
36	0.559	0.000	0.000
>36	0.559	0.000	0.000
D18S51	($N = 180$)	($N = 196$)	($N = 203$)
<11	0.556	1.276	0.493
11	0.556	1.276	1.232
12	5.833	12.755	10.591
13	5.556	12.245	16.995
13.2	0.556	0.000	0.000
14	6.389	17.347	16.995
14.2	0.000	0.000	0.000
15	16.667	12.755	13.793
15.2	0.000	0.000	0.000
16	18.889	10.714	11.576

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
17	16.389	15.561	13.793
18	13.056	9.184	5.172
19	7.778	3.571	3.695
20	5.556	2.551	1.724
21	1.111	0.510	1.970
21.2	0.000	0.000	0.000
22	0.556	0.255	0.739
>22	0.556	0.000	1.232
D5S818	(*N* = 180)	(*N* = 195)	(*N* = 203)
7	0.278	0.000	6.158
8	5.000	0.000	0.246
9	1.389	3.077	5.419
10	6.389	4.872	6.650
11	26.111	41.026	42.118
12	35.556	35.385	29.064
13	24.444	14.615	9.606
14	0.556	0.769	0.493
15	0.000	0.256	0.246
>15	0.278	0.000	0.000
D13S317	(*N* = 179)	(*N* = 196)	(*N* = 203)
7	0.000	0.000	0.000
8	3.631	9.949	6.650
9	2.793	7.653	21.921
10	5.028	5.102	10.099
11	23.743	31.888	20.197
12	48.324	30.867	21.675
13	12.570	10.969	13.793
14	3.631	3.571	5.665
15	0.279	0.000	0.000
D7S820	(*N* = 210)	(*N* = 203)	(*N* = 209)
6	0.000	0.246	0.239

(*continued*)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
7	0.714	1.724	2.153
8	17.381	16.256	9.809
9	15.714	14.778	4.785
10	32.381	29.064	30.622
10.1	0.000	0.000	0.000
11	22.381	20.197	28.947
11.3	0.000	0.000	0.000
12	9.048	14.039	19.139
13	1.905	2.956	3.828
14	0.476	0.739	0.478
CSF1PO	($N = 210$)	($N = 203$)	($N = 209$)
6	0.000	0.000	0.000
7	4.286	0.246	0.239
8	8.571	0.493	0.000
9	3.333	1.970	0.718
10	27.143	25.369	25.359
10.3	0.000	0.246	0.000
11	20.476	30.049	26.555
12	30.000	32.512	39.234
13	5.476	7.143	6.459
14	0.714	1.478	0.957
15	0.000	0.493	0.478
TPOX	($N = 209$)	($N = 203$)	($N = 209$)
6	8.612	0.000	0.478
7	2.153	0.246	0.478
8	36.842	54.433	55.502
9	18.182	12.315	3.349
10	9.330	3.695	3.349
11	22.488	25.369	27.273
12	2.392	3.941	9.330
13	0.000	0.000	0.239
TH01	($N = 210$)	($N = 203$)	($N = 209$)

Table 25.1 (Continued) Allelic Frequencies of 13 CODIS STR Loci

Allele	Allelic Frequency (%)		
	African American	Caucasian	Hispanic
5	0.000	0.000	0.239
6	10.952	22.660	23.206
7	44.048	17.241	33.732
8	18.571	12.562	8.134
8.3	0.000	0.246	0.000
9	14.524	16.502	10.287
9.3	10.476	30.542	24.163
10	1.429	0.246	0.239
D16S539	(*N* = 209)	(*N* = 202)	(*N* = 208)
8	3.589	1.980	1.683
9	19.856	10.396	7.933
10	11.005	6.683	17.308
11	29.426	27.228	31.490
12	18.660	33.911	28.606
13	16.507	16.337	10.337
14	0.957	3.218	2.404
15	0.000	0.248	0.240

Source: Budowle, B., et al., *J Forensic Sci*, 44, 1277–1286, 1999. With permission.

25.1.2.6 Probability of Match

The discriminating power of genetic loci used above can be measured by *population match probability* (P_m). P_m is defined as the probability of having a matching genotype between two randomly chosen individuals. The lower the P_m, the less likely a match between two randomly chosen individuals will occur. This is calculated as follows:

$$P_m = \sum_i \left(p^2\right)^2 + \sum_j \left(2pq\right)^2 \qquad (25.4)$$

where:

p and q = the frequencies of two different alleles

P_m can also be used to compare the discriminating powers of different loci

Tables 25.2 through 25.4 show P_m values of loci commonly used for forensic applications, including SNP, VNTR, and STR.

Table 25.2 Heterozygosity and P_m Values of Six SNP Loci

Locus	Allele	Heterozygosity	P_m
DQA1	7	0.828	0.053
LDLR	2	0.493	0.379
GYPA	2	0.498	0.376
HBGG	3	0.508	0.360
D7S8	2	0.476	0.388
GC	3	0.592	0.235
Average		0.566	
Product			2.5×10^{-4} (1 in 4000)

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group*, 2000.

Table 25.3 Heterozygosity and P_m Values of Six VNTR Loci

Locus	Bins	Heterozygosity	P_m
D1S7	28	0.945	0.0058
D2S44	26	0.926	0.0103
D4S139	19	0.899	0.0184
D10S28	24	0.943	0.0063
D14S13	30	0.899	0.0172
D17S79	19	0.799	0.0700
Average		0.902	
Product			8.26×10^{-12} (1 in 1.2×10^{11})

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group,* 2000.

25.2 Statistical Analysis of DNA Profiling Results

It is desirable to evaluate the strength of DNA profiling results, particularly if two DNA profiles match. A DNA profile from crime scene evidence and a profile from a suspect may be the same for two reasons: (1) the crime scene sample may have come from the suspect or (2) the suspect happens to have the same profile as the individual who left the evidence found at the crime scene. The significance of a match between DNA profiles can be evaluated by using statistical calculations that determine the rarity of a specific DNA profile in a relevant population. The statistical evaluation of the significance can be included in a case report

Table 25.4 Heterozygosity and P_m Values for CODIS Loci

Locus	Allele	Caucasian American		African American	
		Heterozygosity	P_m	Heterozygosity	P_m
CSF1PO	11	0.734	0.112	0.781	0.081
TPOX	7	0.621	0.195	0.763	0.090
TH01	7	0.783	0.081	0.727	0.109
VWA	10	0.811	0.062	0.809	0.063
D16S539	8	0.767	0.089	0.798	0.070
D7S820	11	0.806	0.065	0.782	0.080
D13S317	8	0.771	0.085	0.688	0.136
D5S818	10	0.682	0.158	0.739	0.112
FGA	19	0.860	0.036	0.863	0.033
D3S1358	10	0.795	0.075	0.763	0.094
D8S1179	10	0.780	0.067	0.778	0.082
D18S51	15	0.876	0.028	0.873	0.029
D2S11	20	0.853	0.039	0.861	0.034
Average		0.7812		0.7866	
Product			1.738×10^{-15} (1 in 5.753×10^{14})		1.092×10^{-15} (1 in 9.161×10^{14})

Source: Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group,* 2000.

(Figure 25.2). Guidelines for DNA profile interpretation such as those issued by the National Research Council, the DNA Advisory Board, and the European DNA Profiling Group can be consulted.

25.2.1 Genotypes

The approaches to performing statistical analysis are (1) calculation of profile probability and (2) use of the likelihood ratio method. Although profile probability is the most commonly used method because of its simplicity, both approaches lead to the same conclusion.

25.2.1.1 Profile Probability

Profile probability can be calculated based on the following steps:

1. The locus genotype frequency can be calculated as follows; p and q are the allelic frequencies observed in the database for a given allele:

   Locus genotype frequency for homozygotes:

$$P_i = p^2 \tag{25.5}$$

Locus genotype frequency for heterozygotes:

$$P_j = 2pq \tag{25.6}$$

2. Profile probability can then be calculated based on the product rule by multiplying all the locus genotype frequencies calculated as above.

The lower the profile probability is, the less likely that an individual chosen at random will have a coincident match with the DNA profile of the evidence sample. Table 25.5 shows calculations of profile probability from a DNA profile.

25.2.1.1.1 Structured Populations

The above calculation of profile probability is based on the assumption that a randomly selected individual is unrelated to a perpetrator. However, it is likely that the individual and the perpetrator are from the same subpopulation (groups within a population) and are thus not completely independent. Mating is more likely to occur within subpopulations than between subpopulations. As a result, the proportion of homozygotes increases and the proportion of heterozygotes decreases in a subpopulation because individuals in a subpopulation appear to be related.

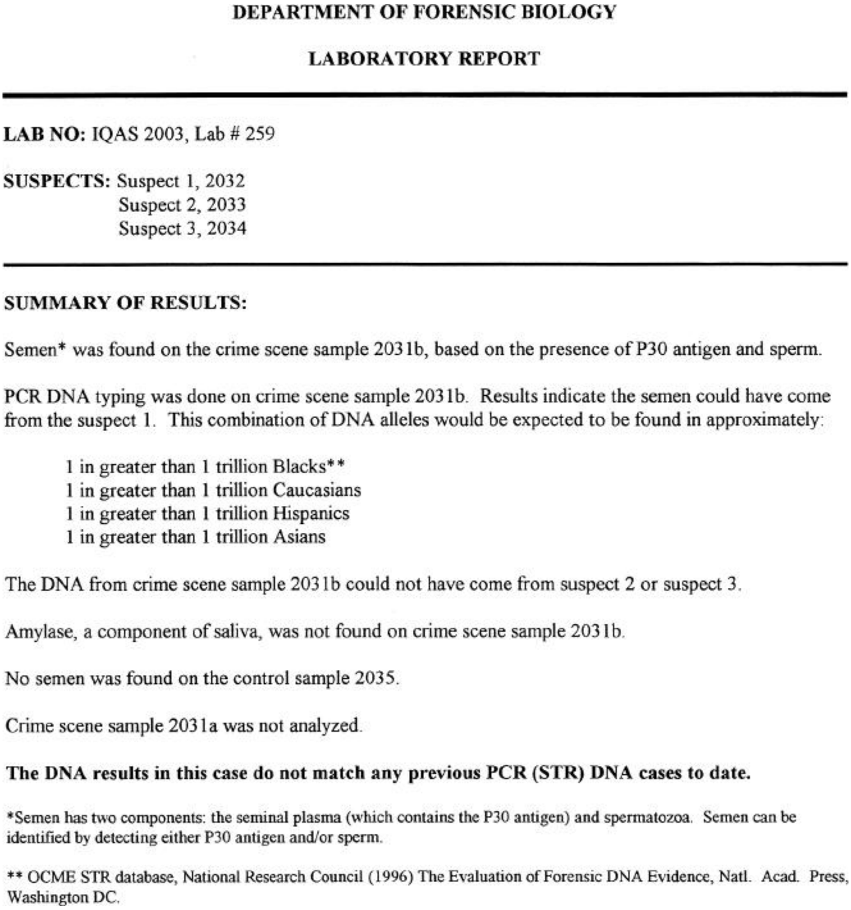

DEPARTMENT OF FORENSIC BIOLOGY

LABORATORY REPORT

LAB NO: IQAS 2003, Lab # 259

SUSPECTS: Suspect 1, 2032
Suspect 2, 2033
Suspect 3, 2034

SUMMARY OF RESULTS:

Semen* was found on the crime scene sample 2031b, based on the presence of P30 antigen and sperm.

PCR DNA typing was done on crime scene sample 2031b. Results indicate the semen could have come from the suspect 1. This combination of DNA alleles would be expected to be found in approximately:

1 in greater than 1 trillion Blacks**
1 in greater than 1 trillion Caucasians
1 in greater than 1 trillion Hispanics
1 in greater than 1 trillion Asians

The DNA from crime scene sample 2031b could not have come from suspect 2 or suspect 3.

Amylase, a component of saliva, was not found on crime scene sample 2031b.

No semen was found on the control sample 2035.

Crime scene sample 2031a was not analyzed.

The DNA results in this case do not match any previous PCR (STR) DNA cases to date.

*Semen has two components: the seminal plasma (which contains the P30 antigen) and spermatozoa. Semen can be identified by detecting either P30 antigen and/or sperm.

** OCME STR database, National Research Council (1996) The Evaluation of Forensic DNA Evidence, Natl. Acad. Press, Washington DC.

Figure 25.2 An example of a laboratory report. (© Richard C. Li.)

Table 25.5 Calculation of Profile Probability of STR Profile with 13 CODIS Loci

Locus	Profile	Allelic Frequency[a]	Formula	Locus Genotype Frequency
CSF1PO	10	0.25369	$2pq$	0.165
	12	0.32512		
D3S1358	14	0.14039	$2pq$	0.0691
	15	0.24631		
D5S818	11	0.41026	p^2	0.168
	11	0.41026		
D7S820	10	0.29064	$2pq$	0.117
	11	0.20197		
D8S1179	13	0.33929	p^2	0.115
	13	0.33929		
D13S317	11	0.31888	p^2	0.102
	11	0.31888		
D16S539	11	0.27228	$2pq$	0.185
	12	0.33911		
D18S51	15	0.12755	$2pq$	0.00911
	19	0.035710		
D21S11	30	0.23214	p^2	0.0539
	30	0.23214		
FGA	23	0.15816	$2pq$	0.0436
	24	0.13776		
TH01	8	0.12562	$2pq$	0.0767
	9.3	0.30542		
TPOX	8	0.54433	p^2	0.296
	8	0.54433		
VWA	17	0.26276	$2pq$	0.117
	18	0.22194		
				Profile probability $= 2.76 \times 10^{-14}$

[a] See Table 25.1.

The effect of population structure should be considered and an appropriate correction should be made to estimate profile probabilities. The correction can be made by using the factor θ. Thus, the locus genotype frequency can be calculated as follows:

Locus genotype frequency for homozygotes:

$$P_i = \frac{\left[2\theta + (1-\theta)p\right]\left[3\theta + (1-\theta)p\right]}{(1+\theta)(1+2\theta)} \tag{25.7}$$

Locus genotype frequency for heterozygotes:

$$P_j = \frac{2\left[\theta + (1-\theta)p\right]\left[\theta + (1-\theta)q\right]}{(1+\theta)(1+2\theta)} \tag{25.8}$$

The θ value is 0.01 for the majority US population and 0.03 for the Native American population. Table 25.6 shows the calculation of profile probability with subpopulation correction using $\theta = 0.01$. The profile probability is approximately three times higher than the value without the correction (Table 25.5). Additional corrections can be calculated for relatives, mixed stains, or database searches using formulas provided by the National Research Council's guidelines.

25.2.1.2 Likelihood Ratio

The *likelihood ratio* (LR) method is an alternative for evaluating the strength of a match. The method allows the calculation of the probability of the DNA profile under two hypothesis:

Hypothesis 1 (H_1) —The evidence and suspect profiles originated from the same source.

Hypothesis 2 (H_2) —The evidence and suspect profiles did not originate from the same source (i.e., the suspect happens to have the same profile as that of the individual who left the evidence).

The LR is the probability of hypothesis H_1 divided by the probability of hypothesis H_2. Where Pr_1 is the probability under hypothesis H_1, and Pr_2 is the probability under hypothesis H_2, this can be expressed as:

$$\mathrm{LR} = \frac{\mathrm{Pr}_1}{\mathrm{Pr}_2} \tag{25.9}$$

The greater the numerator (Pr_1), the greater the likelihood ratio becomes. The result favors hypothesis H_1 (the evidence and suspect profile originated from the same source). Pr_1 is equal to 1 (100%) when a match occurs, and Pr_2 is equal to the profile probability. A LR of 1000 indicates that the evidence is 1000 times as probable if the evidence and suspect profiles originated from the same source.

25.2.2 Haplotypes

The term *haplotype* was first used to describe very closely linked polymorphic loci. During meiosis, alleles at neighboring loci cosegregate (both alleles segregate as a single allele) because of the close linkage of loci. The term also applies to genetic regions within which recombination events are very rare, that is, within mitochondrial and Y-chromosomal DNA. The entire mitochondrial DNA (mtDNA) sequence can be considered to be a single locus or haplotype because of the absence of recombination. Likewise, Y chromosome loci can also be considered haplotypes.

Where recombination is very rare, certain allelic combinations occur in populations much more frequently than would be expected. This phenomenon is called *linkage disequilibrium*. As a result, the HW principle cannot be applied. The two methods for evaluating the strength of

Table 25.6 Calculation of Profile Probability of STR with 13 CODIS Loci and Correction Factor ($\theta=0.01$)

Locus	Profile	Allelic Frequency[a]	Formula	Locus Genotype Frequency
CSF1PO	10	0.25369	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.168
	12	0.32512		
D3S1358	14	0.14039	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0734
	15	0.24631		
D5S818	11	0.41026	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.180
	11	0.41026		
D7S820	10	0.29064	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.121
	11	0.20197		
D8S1179	13	0.33929	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.126
	13	0.33929		
D13S317	11	0.31888	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.113
	11	0.31888		
D16S539	11	0.27228	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.187
	12	0.33911		
D18S51	15	0.12755	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.012
	19	0.03571		
D21S11	30	0.23214	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.0630
	30	0.23214		
FGA	23	0.15816	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0473
	24	0.13776		
TH01	8	0.12562	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.0815
	9.3	0.30542		
TPOX	8	0.54433	$\frac{[2\theta+(1-\theta)p][3\theta+(1-\theta)p]}{(1+\theta)(1+2\theta)}$	0.309
	8	0.54433		
VWA	17	0.26276	$\frac{2[\theta+(1-\theta)p][\theta+(1-\theta)q]}{(1+\theta)(1+2\theta)}$	0.120
	18	0.22194		
				Profile probability $=7.8\times10^{-14}$

[a] See Table 23.1.

a match between haplotypes are (1) mitotype frequency and (2) likelihood ratios. The current most common approach for interpreting mtDNA profiles is mitotype frequency carried out with the gene counting method, that is, the calculation of the number of occurrences of a particular sequence or haplotype. The interpretation of Y chromosome profiles is similar to the interpretation of mtDNA profiles. The estimation of the frequency of a mitotype can be calculated as shown below.

25.2.2.1 Mitotypes Observed in Database

If a mitotype is observed at least once in a database, Equation 25.10 can be used. P_{mt} is the mitotype frequency, x is the number of observations of the haplotype, and n is the size of the database (the number of mitotype entries):

$$P_{\mathrm{mt}} = \frac{x+2}{n+2} \tag{25.10}$$

Any sampling error may be addressed by a confidence interval:

$$P_{\mathrm{mt}} \pm 1.96\sqrt{\frac{P_{\mathrm{mt}}\left(1-P_{\mathrm{mt}}\right)}{n}} \tag{25.11}$$

In this case, P_{mt} is the mitotype frequency, and n is the size of the database. The conservative upper bound of the frequency is usually quoted.

25.2.2.2 Mitotype Not Observed in Database

If a mitotype has not been observed in a database, Equation 25.12 can be used to calculate the mitotype frequency; α is 0.05 for a 95% confidence interval and n is the size of the database.

$$P_{\mathrm{mt}} = 1-\alpha^{1/n} \tag{25.12}$$

Bibliography

Aitken, C.G.G., *Statistics and the Evaluation of Evidence for Forensic Scientists*, p. 260, 1995. New York: Wiley.

Aktulga, H.M., et al., Identifying statistical dependence in genomic sequences via mutual information estimates. *EURASIP J Bioinform Syst Biol*, 2007, **2007**(1): 14741.

Allen, R.W., et al., Considerations for the interpretation of STR results in cases of questioned half-sibship. *Transfusion*, 2007, **47**(3): 515–519.

Amorim, A., A cautionary note on the evaluation of genetic evidence from uniparentally transmitted markers. *Forensic Sci Int Genet*, 2008, **2**(4): 376–378.

Andersen, M.M., P.S. Eriksen, and N. Morling, The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *J Theor Biol*, 2013, **329**: 39–51.

Andersen, M.M., et al., Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Sci Int Genet*, 2013, **7**(2): 264–271.

Balding, D.J., The prosecutor's fallacy and DNA evidence. *Crim Law Rev*, 1994, 711–721.

Balding, D.J., When can a DNA profile be regarded as unique? *Sci Justice*, 1999, **39**(4): 257–260.

Balding, D.J. and P. Donnelly, Inferring identify from DNA profile evidence. *Proc Natl Acad Sci U S A*, 1995, **92**(25): 11741–11745.

Balding, D.J. and P. Donnelly, Evaluating DNA profile evidence when the suspect is identified through a database search. *J Forensic Sci*, 1996, **41**(4): 603–607.

Balding, D.J. and R.A. Nichols, DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci Int*, 1994, **64**(2–3): 125–140.

Beecham, G.W. and B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence. *J Forensic Sci*, 2011, **56**(Suppl 1): S166–S171.

Behar, D.M., et al., A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am J Hum Genet*, 2012, **90**(4): 675–684.

Bodner, M., et al., Inspecting close maternal relatedness: Towards better mtDNA population samples in forensic databases. *Forensic Sci Int Genet*, 2011, **5**(2): 138–141.

Botstein, D., et al., Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 1980, **32**(3): 314–331.

Brenner, C.H., Difficulties in the estimation of ethnic affiliation. *Am J Hum Genet*, 1998, **62**(6): 1558–1560; author reply 1560–1561.

Brenner, C.H., Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Sci Int Genet*, 2010, **4**(5): 281–291.

Bright, J.A., J.M. Curran, and J.S. Buckleton, Relatedness calculations for linked loci incorporating subpopulation effects. *Forensic Sci Int Genet*, 2013, **7**(3): 380–383.

Buckleton, J., C. Triggs, and S. Walsh, *Forensic DNA Evidence Interpretation*, p. 534, 2005. New York: CRC Press.

Buckleton, J.S., M. Krawczak, and B.S. Weir, The interpretation of lineage markers in forensic DNA testing. *Forensic Sci Int Genet*, 2011, **5**(2): 78–83.

Buckleton, J.S., S. Walsh, and S.A. Harbison, The fallacy of independence testing and the use of the product rule. *Sci Justice*, 2001, **41**(2): 81–84.

Budowle, B., The effects of inbreeding on DNA profile frequency estimates using PCR-based loci. *Genetica*, 1995, **96**(1–2): 21–25.

Budowle, B. and T.R. Moretti, Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR based loci. *Forensic Sci Commun*, 1999, **1**(2). Retrieved from http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july1999/index.htm/budowle.htm.

Budowle, B., F.S. Baechtel, and R. Chakraborty, Partial matches in heterogeneous offender databases do not call into question the validity of random match probability calculations. *Int J Legal Med*, 2009, **123**(1): 59–63.

Budowle, B., et al., Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *J Forensic Sci*, 1999, **44**(6): 1277–1286.

Budowle, B., et al., Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci), and HLA-DQ alpha using a multiplex amplification and typing procedure. *J Forensic Sci*, 1995, **40**(1): 45–54.

Budowle, B., et al., Source attribution of a forensic DNA profile. *Forensic Sci Commun*, 2000, **2**(3). Retrieved from http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2000/index.htm/source.htm.

Budowle, B., et al., CODIS STR loci data from 41 sample populations. *J Forensic Sci*, 2001, **46**(3): 453–489.

Budowle, B., et al., Use of prior odds for missing persons identifications. *Investig Genet*, 2011, **2**(1): 15.

Butler, J.M., et al., Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *J Forensic Sci*, 2003, **48**(4): 908–911.

Carracedo, A., et al., New guidelines for the publication of genetic population data. *Forensic Sci Int Genet*, 2013, **7**(2): 217–220.

Chakraborty, R., Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample. *Genetics*, 1984, **108**(3): 719–731.

Chakraborty, R., Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Hum Biol*, 1992, **64**(2): 141–159.

Chakraborty, R. and D.N. Stivers, Paternity exclusion by DNA markers: Effects of paternal mutations. *J Forensic Sci*, 1996, **41**(4): 671–677.

Chakraborty, R., et al., The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. *Electrophoresis*, 1999, **20**(8): 1682–1696.

Clayton, D., On inferring presence of an individual in a mixture: A Bayesian approach. *Biostatistics*, 2010, **11**(4): 661–673.

Cockerton, S., K. McManus, and J. Buckleton, Interpreting lineage markers in view of subpopulation effects. *Forensic Sci Int Genet*, 2012, **6**(3): 393–397.

Collins, A. and N.E. Morton, Likelihood ratios for DNA identification. *Proc Natl Acad Sci U S A*, 1994, **91**(13): 5.

Condel, K. and M. Al Salih, A simple method for establishing concordance between short-tandem-repeat allele frequency databases. *Transfusion*, 2011, **51**(5): 986–992.

Crow, J.F., Hardy, Weinberg and language impediments. *Genetics*, 1999, **152**(3): 821–825.

Curran, J.M. and J. Buckleton, The appropriate use of subpopulation corrections for differences in endogamous communities. *Forensic Sci Int*, 2007, **168**(2–3): 106–111.

Curran, J.M. and J. Buckleton, Inclusion probabilities and dropout. *J Forensic Sci*, 2010, **55**(5): 1171–1173.

Curran, J.M. and J.S. Buckleton, An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations. *Forensic Sci Int Genet*, 2011, **5**(5): 512–516.

Curran, J.M., S.J. Walsh, and J. Buckleton, Empirical testing of estimated DNA frequencies. *Forensic Sci Int Genet*, 2007, **1**(3–4): 267–272.

Curran, J.M., et al., Interpreting DNA mixtures in structured populations. *J Forensic Sci*, 1999, **44**(5): 987–995.

Edwards, A., et al., Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics*, 1992, **12**(2): 241–253.

Egeland, T. and A. Salas, Estimating haplotype frequency and coverage of databases. *PLoS One*, 2008, **3**(12): e3988.

Egeland, T. and A. Salas, A statistical framework for the interpretation of mtDNA mixtures: Forensic and medical applications. *PLoS One*, 2011, **6**(10): e26723.

Egeland, T. and N. Sheehan, On identification problems requiring linked autosomal markers. *Forensic Sci Int Genet*, 2008, **2**(3): 219–225.

Evett, I. and B. Weir, *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, p. 278, 1992. Sunderland, MA: Sinauer Associates.

Evett, I.W. and P. Gill, A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. *Electrophoresis*, 1991, **12**(2–3): 226–230.

Evett, I.W., R. Pinchin, and C. Buffery, An investigation of the feasibility of inferring ethnic origin from DNA profiles. *J Forensic Sci Soc*, 1992, **32**(4): 301–306.

Fisher, R.A., Standard calculations for evaluating a blood-group system. *Heredity*, 1951, **5**(1): 95–102.

Foreman, L.A., A.F.M. Smith, and I. Evett, Bayesian validation of a quaduplex STR profiling system for identification purposes. *J Forensic Sci*, 1999, **44**: 378–486.

Foreman, L.A. and I.W. Evett, Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *Int J Legal Med*, 2001, **114**(3): 147–155.

Foreman, L.A., et al., Interpreting DNA evidence: A review. *Int Stat Rev*, 2003, **71**(3): 13.

Gaensslen, R.E., Journal policy on the publication of DNA population genetic data. *J Forensic Sci*, 1999, **44**(4): 4.

Ge, J. and B. Budowle, Kinship index variations among populations and thresholds for familial searching. *PLoS One*, 2012, **7**(5): e37474.

Ge, J., B. Budowle, and R. Chakraborty, Interpreting Y chromosome STR haplotype mixture. *Leg Med*, 2010, **12**(3): 137–143.

Ge, J., et al., Haplotype block: A new type of forensic DNA markers. *Int J Legal Med*, 2010, **124**(5): 353–361.

Ge, J., et al., Pedigree likelihood ratio for lineage markers. *Int J Legal Med*, 2011, **125**(4): 519–525.

Gill, P. and H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios. *Forensic Sci Int Genet*, 2013, **7**(2): 251–263.

Gill, P., et al., A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations. *Forensic Sci Int*, 2003, **131**(2–3): 184–196.

Gjertson, D.W., et al., ISFG: Recommendations on biostatistics in paternity testing. *Forensic Sci Int Genet*, 2007, **1**(3–4): 223–231.

Guo, S.W. and E.A. Thompson, Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics*, 1992, **48**(2): 361–372.

Haned, H., Forensim: An open-source initiative for the evaluation of statistical methods in forensic genetics. *Forensic Sci Int Genet*, 2011, **5**(4): 265–268.

Hardy, G.H., Mendelian proportions in a mixed population. *Science*, 1908, **17**: 49–50.

Hartman, D., et al., Examples of kinship analysis where Profiler Plus was not discriminatory enough for the identification of victims using DNA identification. *Forensic Sci Int*, 2011, **205**(1–3): 64–68.

Henderson, J.P., The use of DNA statistics in criminal trials. *Forensic Sci Int*, 2002, **128**(3): 183–186.

Hill, C.R., et al., Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex ESX 17 and ESI 17 Systems. *Forensic Sci Int Genet*, 2011, **5**(4): 269–275.

Holt, C.L., et al., Practical applications of genotypic surveys for forensic STR testing. *Forensic Sci Int*, 2000, **112**(2–3): 91–109.

Hong-Sheng, G., et al., HGD-Chn: The database of genome diversity and variation for Chinese populations. *Leg Med*, 2009, **11**(Suppl 1): S201–S202.

Hopwood, A.J., et al., Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Sci Justice*, 2012, **52**(3): 185–190.

Hosking, L., et al., Detection of genotyping errors by Hardy–Weinberg equilibrium testing. *Eur J Hum Genet*, 2004, **12**(5): 395–399.

Janica, J., et al., Database of genetic profiles at the Department of Forensic Medicine, Medical University of Bialystok. *Adv Med Sci*, 2008, **53**(1): 64–68.

Kling, D., T. Egeland, and A.O. Tillmar, FamLink—A user friendly software for linkage calculations in family genetics. *Forensic Sci Int Genet*, 2012, **6**(5): 616–620.

Lareu, M.V., et al., Analysis of a claimed distant relationship in a deficient pedigree using high density SNP data. *Forensic Sci Int Genet*, 2012, **6**(3): 350–353.

Lauc, G., et al., Empirical support for the reliability of DNA interpretation in Croatia. *Forensic Sci Int Genet*, 2008, **3**(1): 50–53.

Lee, J.C., et al., A novel strategy for sibship determination in trio sibling model. *Croat Med J*, 2012, **53**(4): 336–342.

Li, L., et al., Maternity exclusion with a very high autosomal STRs kinship index. *Int J Legal Med*, 2012, **126**(4): 645–648.

Lincoln, P. and A. Carracedo, Publication of population data of human polymorphisms. *Forensic Sci Int*, 2000, **110**(1): 3–5.

Lins, A.M., et al., Development and population study of an eight-locus short tandem repeat (STR) multiplex system. *J Forensic Sci*, 1998, **43**(6): 1168–1180.

Matsumura, S., et al., Kinship analysis using DNA typing from five skeletal remains with an unusual postmortem course. *Med Sci Law*, 2011, **51**(4): 240–243.

Montinaro, F., et al., Using forensic microsatellites to decipher the genetic structure of linguistic and geographic isolates: A survey in the eastern Italian Alps. *Forensic Sci Int Genet*, 2012, **6**(6): 827–833.

Morton, N.E., Genetic structure of forensic populations. *Proc Natl Acad Sci U S A*, 1992, **89**(7): 2556–2560.

Mueller, L.D., Can simple population genetic models reconcile partial match frequencies observed in large forensic databases? *J Genet*, 2008, **87**(2): 101–108.

Nei, M., Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 1978, **89**(3): 583–590.

Noris, G., et al., Mexican mestizo population sub-structure: Effects on genetic and forensic statistical parameters. *Mol Biol Rep*, 2012, **39**(12): 10139–10156.

Nothnagel, M., J. Schmidtke, and M. Krawczak, Potentials and limits of pairwise kinship analysis using autosomal short tandem repeat loci. *Int J Legal Med*, 2010, **124**(3): 205–215.

NRC (National Research Council), *The Evaluation of Forensic DNA Evidence*, 1996. Washington, DC: National Research Council Committee, National Academies Press.

Nunez, C., et al., Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers. *Am J Phys Anthropol*, 2010, **143**(4): 591–600.

O'Connor, K.L. and A.O. Tillmar, Effect of linkage between vWA and D12S391 in kinship analysis. *Forensic Sci Int Genet*, 2012, **6**(6): 840–844.

O'Connor, K.L., et al., Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples. *Forensic Sci Int Genet*, 2011, **5**(5): 538–540.

Office of Justice Programs, National Institute of Justice, US Department of Justice, *The Future of Forensic DNA Testing: Predictions of the Research and Development Working Group,* 2000.

Palo, J.U., et al., Genetic markers and population history: Finland revisited. *Eur J Hum Genet*, 2009, **17**(10): 1336–1346.

Parson, W. and H.J. Bandelt, Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci Int Genet*, 2007, **1**(1): 13–19.

Parson, W. and L. Roewer, Publication of population data of linearly inherited DNA markers in the *International Journal of Legal Medicine*. *Int J Legal Med*, 2010, **124**(5): 505–509.

Pereira, L., et al., PopAffiliator: Online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. *Int J Legal Med*, 2011, **125**(5): 629–636.

Pereira, R., et al., Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One*, 2012, **7**(1): e29684.

Phillips, C., et al., D9S1120, a simple STR with a common Native American-specific allele: Forensic optimization, locus characterization and allele frequency studies. *Forensic Sci Int Genet*, 2008, **3**(1): 7–13.

Phillips, C., et al., The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data. *Forensic Sci Int Genet*, 2012, **6**(3): 354–365.

Prieto, L., et al., The GHEP-EMPOP collaboration on mtDNA population data—A new resource for forensic casework. *Forensic Sci Int Genet*, 2011, **5**(2): 146–151.

Pu, C.E. and A. Linacre, Increasing the confidence in half-sibship determination based upon 15 STR loci. *J Forensic Leg Med*, 2008, **15**(6): 373–377.

Pu, C.E. and A. Linacre, Systematic evaluation of sensitivity and specificity of sibship determination by using 15 STR loci. *J Forensic Leg Med*, 2008, **15**(5): 329–334.

Puch-Solis, R., S. Pope, and I. Evett, Calculating likelihood ratios for a mixed DNA profile when a contribution from a genetic relative of a suspect is proposed. *Sci Justice*, 2010, **50**(4): 205–209.

Rajeevan, H., et al., Introducing the forensic research/reference on genetics knowledge base, FROG-kb. *Investig Genet*, 2012, **3**(1): 18.

Ravid-Amir, O. and S. Rosset, Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. *Bioinformatics*, 2010, **26**(18): i440–i445.

Raymond, M. and F. Rousset, Population genetics software for exact tests and ecumenicism. *J Hered*, 1995, **86**: 12.

Robertson, B. and G.A. Vigneaux, Expert evidence: Law, practice, and probability. *Oxford J Legal Stud*, 1992, **12**(1): 392–403.

Rodig, H., et al., Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. *Forensic Sci Int*, 2008, **174**(2–3): 182–188.

Roeder, K., DNA fingerprinting: A review of the controversy. *Stat Sci*, 1994, **9**(2): 222–278.

Rowold, D.J. and R.J. Herrera, On human STR sub-population structure. *Forensic Sci Int*, 2005, **151**(1): 59–69.

Royall, R., *Statistical Evidence: A Likelihood Paradigm*, 1997. New York: Chapman & Hall.

Sanchez, J.J., et al., Forensic typing of autosomal SNPs with a 29 SNP-multiplex—Results of a collaborative EDNAP exercise. *Forensic Sci Int Genet*, 2008, **2**(3): 176–183.

Sanchez-Diz, P., et al., Population and segregation data on 17 Y-STRs: Results of a GEP-ISFG collaborative study. *Int J Legal Med*, 2008, **122**(6): 529–533.

Santos, C., et al., Frequency and pattern of heteroplasmy in the control region of human mitochondrial DNA. *J Mol Evol*, 2008, **67**(2): 191–200.

Silva, N.M., et al., Human neutral genetic variation and forensic STR data. *PLoS One*, 2012, **7**(11): e49666.

Stockmarr, A., Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics*, 1999, **55**(3): 671–677.

Taroni, F., et al., Evaluation and presentation of forensic DNA evidence in European laboratories. *Sci Justice*, 2002, **42**(1): 21–28.

Taylor, C. and P. Colman, Forensics: Experts disagree on statistics from DNA trawls. *Nature*, 2010, **464**(7293): 1266–1267.

Thompson, W.C. and E.L. Schumann, Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law Hum Behav*, 1987, **11**(3): 11.

Toscanini, U., et al., Evaluating methods to correct for population stratification when estimating paternity indexes. *PLoS One*, 2012, **7**(11): e49832.

Tvedebrink, T., et al., Estimating the probability of allelic drop-out of STR alleles in forensic genetics. *Forensic Sci Int Genet*, 2009, **3**(4): 222–226.

van Oven, M., et al., Unexpected island effects at an extreme: Reduced Y chromosome and mitochondrial DNA diversity in Nias. *Mol Biol Evol*, 2011, **28**(4): 1349–1361.

Walsh, S.J., et al., Evidence in support of self-declaration as a sampling method for the formation of subpopulation DNA databases. *J Forensic Sci*, 2003, **48**(5): 1091–1093.

Walsh, S.J., et al., Use of subpopulation data in Australian forensic DNA casework. *Forensic Sci Int Genet*, 2007, **1**(3–4): 238–246.

Weir, B.S. and W.G. Hill, Estimating F-statistics. *Annu Rev Genet*, 2002, **36**: 721–750.

Wenk, R.E., Detection of genotype recycling fraud in U.S. immigrants. *J Forensic Sci*, 2011, **56**(Suppl 1): S243–S246.

Willuweit, S., et al., Y-STR frequency surveying method: A critical reappraisal. *Forensic Sci Int Genet*, 2011, **5**(2): 84–90.

26

Quality Assurance and Quality Control

Quality assurance (*QA*) for forensic services requires certain processes to ensure that a service will meet laboratory requirements for the integrity of testing. A QA program must include components that address:

- Continuing education, training, and certification of personnel
- Specification and calibration of equipment and reagents
- Documentation and validation of analytic methods
- Use of appropriate standards and controls
- Sample handling procedures
- Proficiency testing
- Data interpretation and reporting
- Audits (internal and external) and laboratory accreditation
- Corrective actions to address deficiencies and assessments for laboratory competence

Over the years, many guidelines for quality assurance in forensic DNA laboratories have been established. These guidelines will be introduced in this chapter.

Quality control (*QC*) for forensic services refers to the operational procedures necessary to meet quality requirements. QC procedures may include maintenance of calibration records for equipment and instruments as well as the testing of chemical reagents and supplies used in analysis to ensure reliable results.

26.1 US Quality Standards

DNA profiling methods were first used in criminal investigations in the 1980s. By the early 1990s, emerging forensic DNA techniques had undergone detailed reviews by the *National Research Council* (*NRC*) of the National Academy of Sciences. In 1992, the first published NRC report included recommendations in the areas of (1) technical considerations, (2) statistical interpretation, (3) laboratory standards, (4) data banks and privacy, (5) legal considerations, and (6) societal and ethical issues related to forensic DNA testing. The NRC report attempted to explain the basic scientific principles of forensic DNA technology and made suggestions for applications and improvements. However, the report received negative criticism from both the forensic and the legal communities.

In 1996, a second NRC committee was formed "to update and clarify discussion of the principles of population genetics and statistics as they apply to DNA evidence." Its report stated:

> The central question that the report addresses is this: What information can a forensic scientist, population geneticist, or statistician provide to assist a judge or jury in drawing inferences from the finding of a match? To answer this question, the committee reviewed the scientific literature and the legal cases and commentaries on DNA profiling, and it investigated the criticisms that have been voiced about population data, statistics, and laboratory error. Much has been learned since the last report. The technology for DNA profiling as well as the methods for estimating frequencies and related statistics have progressed to the point where the reliability and validity of properly collected and analyzed DNA data should not be in doubt. The new recommendations presented here should pave the way to more effective use of DNA evidence. (*The Evaluation of Forensic DNA Evidence* also known as the NRC II report)

The NRC II report consisted of (1) an introduction describing the 1992 report, changes made subsequent to that report, and the validity and application of DNA typing techniques; (2) assurance of high standards of laboratory performance; (3) population genetics issues; (4) statistical issues; and (5) DNA evidence in the legal system.

In 1995, The *DNA Advisory Board* (*DAB*) was formed as a result of the DNA Identification Act (1994) passed by Congress. The DAB served from 1995 to 2000 to develop guidelines for quality assurance in forensic laboratories. During that time, the DAB provided two sets of guidelines: Quality Assurance Standards for Forensic DNA Testing Laboratories (1998) and Quality Assurance Standards for Convicted Offender DNA Databasing Laboratories (1999). These standards describe the requirements to ensure quality and integrity of the data as well as competency of laboratories.

These standards were built on the previous standards set by the *Scientific Working Group on DNA Analysis Methods* (*SWGDAM*). After the DAB's assignment ended in 2000, the SWGDAM became responsible for providing guidelines to the US forensic community. The SWGDAM was established in 1988 by the FBI Laboratory. It was initially called the Technical Working Group on DNA Analysis Methods; the name changed in 1998. The SWGDAM comprises forensic scientists from DNA laboratories in the United States and Canada. Its purpose is to facilitate forensic DNA community discussions regarding necessary laboratory methods and to share protocols for forensic DNA testing. The FBI sponsors and hosts its meetings and plays an important role in its activities.

The SWGDAM established and revised several guidelines including the Guidelines for a Quality Assurance Program for DNA Analysis, published in 1989, 1991, and 1995 (the validation section was revised in 2003); Quality Assurance Standards for Forensic DNA Testing Laboratories (2011); Quality Assurance Standards for DNA Databasing Laboratories (2011); Validation Guidelines for DNA Analysis Methods (2012); and SWGDAM Training Guidelines, published in 2001 and revised in 2013.

The SWGDAM also formed subcommittees to provide guidelines in more specific areas of forensic DNA testing, for example, SWGDAM's Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories (2010), Y-chromosome Short Tandem Repeat (Y-STR) Interpretation Guidelines (2009), and Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories (2013). The SWGDAM also organized a number of interlaboratory and validation studies for new techniques.

26.2 International Quality Standards

In the early stages of forensic DNA testing, the *International Society for Forensic Genetics* (*ISFG*) recognized the potential of DNA testing for criminal investigations and made a number of recommendations related to the forensic application of DNA polymorphisms.

The ISFG provided various recommendations for forensic DNA testing of STR, mtDNA, and Y chromosome markers for the international community. It formed a working group called the *European DNA Profiling Group* (*EDNAP*) in 1991. The EDNAP has investigated systems for DNA profiling, has organized a number of collaborative exercises for the evaluation of new methods, and has published reports of its studies.

The *European Network of Forensic Science Institutes* (*ENFSI*) formed its DNA working group about a decade ago to address issues of quality and standards for forensic DNA testing. The *Interpol European Working Party on DNA Profiling* (*IEWPDP*) also makes recommendations for applying DNA evidence to criminal investigations in Europe. Based on the EDNAP exercises and recommendations by the ENFSI and IEWPDP, the *European Standard Set* (*ESS*) for autosomal STR core loci was established. The Standardization of DNA Profiling Techniques in the European Union (STADNAP) group has been working on the selection of forensic DNA profiling systems, methods for use among European countries, and the maintenance of European population databases.

26.3 Laboratory Accreditation

Accreditation is the process used to assess the qualification of a laboratory to meet established standards. During an accreditation process, the services and performance of a laboratory are evaluated, particularly in the areas of management, operations, personnel, procedures, equipment, physical plants, security, and personnel safety procedures. The accreditation process generally involves several components such as self-evaluation, the preparation of supporting documents, on-site inspection and reports, and accreditation review reports.

Accreditation in the United States is offered by the *Laboratory Accreditation Board* of the *American Society of Crime Laboratory Directors* (*ASCLD/LAB*) for forensic laboratories performing casework. The *American Association of Blood Banks* (*AABB*) provides accreditation for laboratories performing DNA parentage testing according to the AABB's standards.

The accreditation of a forensic laboratory is granted for 5 years. To remain in compliance, a laboratory must undergo audits to evaluate its operation according to established guidelines. The FBI has published the Quality Assurance Standards Audit for Forensic DNA Testing Laboratories (2011) and the Quality Assurance Standards Audit for DNA Databasing Laboratories (2011). The areas of operating protocols, instruments and equipment, and personnel training are evaluated based on guidelines. Problems identified during an audit must be documented and actions to resolve the problems must be addressed. Annual internal audits and external audits during alternate years are required under the guidelines.

26.4 Laboratory Validation

Validation is the process of confirming that a laboratory procedure is sufficiently robust, reliable, and reproducible. A robust method maintains successful performance and can cope with errors. A reliable method produces accurate results. A reproducible method achieves the same or very similar results each time a sample is tested.

Two types of laboratory validations, developmental and internal validations, are used for modifying methods for forensic DNA analysis. An *internal validation* is required when adopting a procedure for forensic applications. Based on the SWGDAM's Validation Guidelines for DNA Analysis Methods, the internal validation is "an accumulation of test data within the laboratory to demonstrate that established methods and procedures perform as expected in the laboratory." The *developmental validation* "is the acquisition of test data and determination of conditions and limitations of a new or novel DNA methodology for use on forensic, database, known or casework reference samples." For example, the characterization of a new genetic marker requires studies in the inheritance, the genomic location, the detection method, and the polymorphism of the genetic marker. However, developmental validation studies in the same subject but from fields other than forensic DNA analysis may be acceptable for forensic applications.

Additionally, the precision and the accuracy of the test are important criteria for developmental validation. Based on the SWGDAM, *precision* "characterizes the degree of mutual agreement among a series of individual measurements, values and/or results. Precision depends only on the distribution of random errors and does not relate to the true value or specified value." The measure of precision can be expressed as a standard deviation of test results. *Accuracy* is referred to as "the ability of a measuring instrument to give responses close to a true value," which can be assessed by a *performance check*, a quality assurance procedure monitoring the performance of instruments and equipment affecting the accuracy of forensic DNA testing.

26.5 Proficiency Testing

Proficiency testing is an important component of quality control and quality assurance. It evaluates a laboratory's performance of DNA analyses according to the laboratory's standard protocols. Proficiency testing also evaluates the quality of performance by individual analysts following laboratory protocols.

Proficiency tests of DNA analysts must be conducted every 6 months based on DNA Advisory Board Standards. The testing usually involves mock forensic case samples including questioned bodily fluid stains and reference samples. The test is assigned to an analyst for processing according to the laboratory procedures. A report must be prepared and is then reviewed. The proficiency test can be administered as either an open or a blind test. In the blind test, the analyst is not aware that he or she is being tested. Blind testing is considered a more effective means of evaluating performance.

The tests may be administered internally or by any of a number of external testing organizations. For example, Orchid Cellmark provides the *International Quality Assessment Scheme* (*IQAS*) DNA proficiency test, and the Collaborative Testing Services (CTS) Forensics Testing Program offers similar proficiency tests. These proficiency tests may include a selection of sample types (neat or mixture) for serological tests and/or DNA analysis (autosomal STR, Y-STR, and mitochondrial DNA). The College of American Pathologists (CAP) has a paternity testing proficiency program offering external proficiency tests. In Europe, the *German DNA Profiling Group* (*GEDNAP*) provides proficiency testing for participating European laboratories.

26.6 Certification

Certification is a voluntary process that recognizes the attainment of professional qualifications needed for practice in forensic services. Certification is not required, but is desired by some laboratories. In 2009, a report entitled Strengthening Forensic Science in the United States: A Path Forward was published by the Committee on Identifying the Needs of the Forensic Science Community at the National Academy of Sciences. The report recommends that certification should be mandatory for forensic science professionals, which includes written examinations, supervised practice, proficiency testing, and compliance to a code of ethics (Section 26.8). In the United States, the *American Board of Criminalistics* (*ABC*) offers three types of certification for forensic scientists. A diplomate must pass a general knowledge examination. ABC also requires a bachelor's degree in a natural science and 2 years of experience in a forensic laboratory. To obtain fellow status (higher than diplomate status), an applicant must have 2 years of experience in his or her specialty and must have met the diplomate requirements in addition to passing a written specialty examination and a proficiency test.

ABC has added a third certification that is completely separate from the other certification programs. The technical specialist certification for molecular biology was created to recognize the qualifications required for the analysis of biological materials through DNA profiling. An applicant must take a specialist examination containing questions from the general knowledge examination and a subset of questions from the forensic biology fellow examination. This certification also requires a bachelor's degree in a natural science and 2 years of experience along with successful completion of a proficiency examination within 12 months of taking the technical specialist certification examination.

26.7 Forensic DNA Analyst Qualifications

DAB's Quality Assurance Standards for Forensic DNA Testing Laboratories require that an examiner or analyst have "at a minimum a BA/BS degree or its equivalent degree in a biology, chemistry or forensic science-related area and must have successfully completed college course work (graduate or undergraduate level) covering the subject areas of biochemistry, genetics and molecular biology," as well as "course work and/or training in statistics and population genetics as it applies to forensic DNA analysis."

Additionally, SWGDAM Training Guidelines discuss the course work requirements in detail. At least nine cumulative semester hours are required for the course work covering the required subject areas. Particularly, the section in Fundamental Scientific Knowledge of the SWGDAM Training Guidelines defines required key elements to be covered in course work. The key elements also aid in evaluating the contents of the course work.

Biochemistry "refers to the nature of biologically important molecules in living systems, DNA replication and protein synthesis, and the quantitative and qualitative aspects of cellular metabolism," and may include but is not limited to:

- Structure and function of cellular components such as proteins, carbohydrates, lipids, nucleic acids, and other biomolecules
- Chemistry of enzyme-catalyzed reactions
- Metabolism
- DNA and RNA
- Protein synthesis
- Cell membrane transport
- Signal transduction

Genetics "refers to the study of inherited traits, genotype/phenotype relationships, and population/species differences in allele and genotype frequencies," and may include but is not limited to:

- Heredity
- Function of genes
- Gene expression
- Recombinant DNA
- Mitosis/meiosis

Molecular biology "covers theories, methods, and techniques used in the study and analysis of gene structure, organization, and function," and may include, but is not limited to:

- Interrelationship of DNA, RNA, and protein synthesis
- Central dogma

- Transcription, translation, replication
- Recombinant DNA techniques
- PCR
- Cloning

Course work that is similar in content and scope to those described above may be qualified as required course work. Compliance with the required course contents can be evaluated through transcripts, syllabi, and letters from the instructors.

DAB's Quality Assurance Standards for Forensic DNA Testing Laboratories also describes requirements for an analyst to meet prior to initiating independent casework analysis using DNA technology. First, an analyst must have at least 6 months of forensic DNA laboratory experience "including the successful analysis of a range of samples typically encountered in forensic casework." Second, an analyst must have "successfully completed a qualifying test before beginning independent casework responsibilities."

26.8 Code of Ethics of Forensic Scientists

Ethical codes for forensic scientists are used as guides for individuals making their decisions in distinguishing the difference between correctness and incorrectness. These codes are usually adopted by forensic organizations to regulate the profession. Failure to comply with a code of forensic practice can raise doubt in an individual's fitness for providing forensic services. Additionally, it may result in the expulsion of an individual from a forensic organization. A typical document of ethical codes for forensic scientists generally contains five sections: (1) ethics relating to the scientific method, (2) ethics relating to opinions and conclusions, (3) ethical aspects of court presentation, (4) ethics relating to the general practice of forensics, and (5) ethical responsibilities to the profession.

For the first section, the ethical application of scientific methods is discussed in detail. True scientific methods should be utilized to adequately analyse all the evidence. Such analyses should not be conducted by secret processes. Conclusions should be drawn from the analyses of evidence that appears representative, typical, or reliable.

Moreover, ethical standards relating to opinions and conclusions are very important to the forensic profession. Conclusions should always be drawn from the application of proven scientific methods. The purpose of experimental design and the interpretation of results is to reveal facts. During an analysis, experimental controls should always be utilized. If necessary, the results of analyses should be verified by repeating the analysis or using additional techniques. Explanations should be provided where inconclusive results are obtained. The opinion provided by a forensic scientist should be unbiased and should not be influenced by matters unrelated to the specific evidence in question. A forensic scientist should not choose the interpretation that is in favor of the side of his or her employer.

Pertaining to the courtroom, ethical standards relating to expert witness testimonies is a crucial component for a forensic scientist. An expert opinion is defined as a formal consideration of a subject within an individual's knowledge, skill, education, training, and experience. A forensic scientist should not extend an opinion beyond his or her competence. Appropriate terms should be used to represent the degrees of certainty of an expert opinion. A forensic scientist should not only present the evidence that supports the view of the side of his or her employer. When testifying as an expert witness, language that can be understood by lay jurors should be used to avoid misinterpretation.

Furthermore, in terms of the general practice of forensic science, a few things should be mentioned here. A forensic scientist in private practice should set a reasonable fee for the services provided, which should not be rendered on a contingency fee basis. When a different opinion is presented by another analyst in the reexamination of evidence, both analysts should resolve their contradiction before the trial. A forensic scientist may serve, in an advisory capacity, either the prosecutor or defense in the cross-examination of another expert.

Lastly, ethical responsibilities to the profession are an important component of ethical codes. Information regarding new developments or techniques of forensic analysis should be made available to other forensic scientists. Likewise, any information regarding methods in use that may appear unreliable should be brought to the attention of others. Individual forensic scientists should not seek publicity for the association of their name with specific cases, developments, publications, or organizations solely for the purpose of gaining personal prestige.

Bibliography

Andersen, J., et al., Report on the third EDNAP collaborative STR exercise. *Forensic Sci Int*, 1996, **78**(2): 83–93.

Balazic, J. and I. Zupanic, Quality control and quality assurance in DNA laboratories: Legal, civil and ethical aspects. *Forensic Sci Int*, 1999, **103**: S1–S5.

Bar, W., et al., DNA recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. International Society for Forensic Haemogenetics. *Int J Legal Med*, 1997, **110**(4): 175–176.

Buckleton, J., Validation issues around DNA typing of low level DNA. *Forensic Sci Int Genet*, 2009, **3**(4): 255–260.

Budowle, B., et al., A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *J Forensic Sci*, 2009, **54**(4): 798–809.

Carracedo, A., et al., DNA commission of the International Society for Forensic Genetics: Guidelines for mitochondrial DNA typing. *Forensic Sci Int*, 2000, **110**(2): 79–85.

Carracedo, A., et al., Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format. *Forensic Sci Int*, 2001, **119**(1): 28–41.

Cormier, K., L. Calandro, and D.J. Reeder, Evolution of the quality assurance documents for DNA laboratories. *Forensic Mag*, 2005, 2(1): 16–19.

Council, N.R., *NRC II: The Evaluation of Forensic DNA Evidence*, 1996. Washington, D.C.: National Academy Press.

Coyle, H.M., The importance of scientific evaluation of biological evidence—Data from eight years of case review. *Sci Justice*, 2012, **52**(4): 268–270.

DNA Advisory Board, Federal Bureau of Investigation, US Department of Justice, Quality assurance standards for forensic DNA testing laboratories. *Forensic Sci Commun*, 2000, **2**(3). Retrieved from http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2000/codis2a.htm.

Duewer, D.L., K.L. Richie, and D.J. Reeder, RFLP band size standards: NIST standard reference material 2390. *J Forensic Sci*, 2000, **45**(5): 1093–1105.

Duewer, D.L., et al., NIST mixed stain studies #1 and #2: Interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples. *J Forensic Sci*, 2001, **46**(5): 1199–1210.

Frank, W.E., et al., Validation of the AmpFlSTR Profiler Plus PCR amplification kit for use in forensic casework. *J Forensic Sci*, 2001, **46**(3): 642–646.

Gill, P., et al., Report of the European DNA profiling group (EDNAP)—Towards standardisation of short tandem repeat (STR) loci. *Forensic Sci Int*, 1994, **65**(1): 51–59.

Gill, P., et al., Report of the European DNA profiling group (EDNAP): An investigation of the complex STR loci D21S11 and HUMFIBRA (FGA). *Forensic Sci Int*, 1997, **86**(1–2): 25–33.

Gill, P., et al., Report of the European DNA profiling group (EDNAP)—An investigation of the hypervariable STR loci ACTBP2, APOAI1 and D11S554 and the compound loci D12S391 and D1S1656. *Forensic Sci Int*, 1998, **98**(3): 193–200.

Gill, P., et al., DNA commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Int J Legal Med*, 2001, **114**(6): 305–309.

Gill, P., et al., DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. *Forensic Sci Int*, 2006, **160**(2–3): 90–101.

Gill, P., et al., DNA commission of the International Iociety of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods. *Forensic Sci Int Genet*, 2012, **6**(6): 679–688.

Gusmao, L., et al., DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci Int*, 2006, **157**(2–3): 187–197.

Holt, C.L., et al., TWGDAM validation of AmpFlSTR PCR amplification kits for forensic DNA casework. *J Forensic Sci*, 2002, **47**(1): 66–96.

Houck, M., et al., The balanced scorecard: Sustainable performance assessment for forensic laboratories. *Sci Justice*, 2012, **52**(4): 209–216.

Jarman, P.G., S.L. Fentress, and D.E. Katz, Mitochondrial DNA validation in a state laboratory. *J Forensic Sci*, 2009, **54**(1): 95–102.

Kimpton, C., et al., Report on the second EDNAP collaborative STR exercise. European DNA profiling group. *Forensic Sci Int*, 1995, **71**(2): 137–152.

Kline, M.C., et al., Interlaboratory evaluation of short tandem repeat triplex CTT. *J Forensic Sci*, 1997, **42**(5): 897–906.

Kline, M.C., et al., NIST mixed stain study 3: DNA quantitation accuracy and its influence on short tandem repeat multiplex signal intensity. *Anal Chem*, 2003, **75**(10): 2463–2469.

Kline, M.C., et al., Production and certification of NIST standard reference material 2372 human DNA quantitation standard. *Anal Bioanal Chem*, 2009, **394**(4): 1183–1192.

LaFountain, M.J., et al., TWGDAM validation of the AmpFlSTR Profiler Plus and AmpFlSTR COfiler STR multiplex systems using capillary electrophoresis. *J Forensic Sci*, 2001, **46**(5): 1191–1198.

Lanning, K.A., et al., Scientific working group on materials analysis position on hair evidence. *J Forensic Sci*, 2009, **54**(5): 1198–1202.

Laux, D.L., Development of biological standards for the quality assurance of presumptive testing reagents. *Sci Justice*, 2011, **51**(3): 143–145.

Levin, B.C., et al., Comparison of the complete mtDNA genome sequences of human cell lines—HL-60 and GM10742A—From individuals with pro-myelocytic leukemia and leber hereditary optic neuropathy, respectively, and the inclusion of HL-60 in the NIST human mitochondrial DNA standard reference material—SRM 2392-I. *Mitochondrion*, 2003, **2**(6): 387–400.

Lincoln, P.J., DNA recommendations—Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. *Forensic Sci Int*, 1997, **87**(3): 181–184.

Marquez, M.C., Interpretation guidelines of mtDNA control region sequence electropherograms in forensic genetics. *Methods Mol Biol*, 2012, **830**: 301–319.

Micka, K.A., et al., TWGDAM validation of a nine-locus and a four-locus fluorescent STR multiplex system. *J Forensic Sci*, 1999, **44**(6): 1243–1257.

Morling, N., et al., Paternity testing commission of the International Society of Forensic Genetics: Recommendations on genetic investigations in paternity cases. *Forensic Sci Int*, 2002, **129**(3): 148–157.

Ong, B.B. and N. Milne, Quality assurance in forensic pathology. *Malays J Pathol*, 2009, **31**(1): 17–22.

Patterson, D. and R. Campbell, The problem of untested sexual assault kits: Why are some kits never submitted to a crime laboratory? *J Interpers Violence*, 2012, **27**(11): 2259–2275.

Peterson, J.L., et al., The feasibility of external blind DNA proficiency testing. I. Background and findings. *J Forensic Sci*, 2003, **48**(1): 21–31.

Peterson, J.L., et al., The feasibility of external blind DNA proficiency testing. II. Experience with actual blind tests. *J Forensic Sci*, 2003, **48**(1): 32–40.

Presley, L.A., The evolution of quality standards for forensic DNA analyses in the United States. *Profiles DNA*, 1999, 3(2): 10–11.

Prieto, L., et al., 2006 GEP-ISFG collaborative exercise on mtDNA: Reflections about interpretation, artefacts, and DNA mixtures. *Forensic Sci Int Genet*, 2008, **2**(2): 126–133.

Prieto, L., et al., GHEP-ISFG proficiency test 2011: Paper challenge on evaluation of mitochondrial DNA results. *Forensic Sci Int Genet*, 2013, **7**(1): 10–15.

Prinz, M., et al., DNA commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Sci Int Genet*, 2007, **1**(1): 3–12.

Rand, S., M. Schurenkamp, and B. Brinkmann, The GEDNAP (German DNA profiling group) blind trial concept. *Int J Legal Med*, 2002, **116**(4): 199–206.

Rand, S., et al., The GEDNAP blind trial concept part II. Trends and developments. *Int J Legal Med*, 2004, **118**(2): 83–89.

Reardon, S., Faulty forensic science under fire. *Nature*, 2014, **506**(7486): 13–14.

Reeder, D.J., Impact of DNA typing on standards and practice in the forensic community. *Arch Pathol Lab Med*, 1999, **123**(11): 1063–1065.

Roewer, L. and M. Geppert, Interpretation guidelines of a standard Y-chromosome STR 17-plex PCR-CE assay for crime casework. *Methods Mol Biol*, 2012, **830**: 43–56.

Schneider, P.M., Scientific standards for studies in forensic genetics. *Forensic Sci Int*, 2007, **165**(2–3): 238–243.

Schneider, P.M., et al., Results of collaborative study regarding the standardization of the Y-linked STR system DYS385 by the European DNA Profiling (EDNAP) group. *Forensic Sci Int*, 1999, **102**(2–3): 159–165.

Scott, C.L., Believing doesn't make it so: Forensic education and the search for truth. *J Am Acad Psychiatry Law*, 2013, **41**(1): 18–32.

SWGDAM (Scientific Working Group on DNA Analysis Methods), Revised validation guidelines. *Forensic Sci Commun*, 2004, **6**(3). Retrieved from http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2004/index.htm/standards/2004_03_standards02.htm.

Szibor, R., et al., Cell line DNA typing in forensic genetics—The necessity of reliable standards. *Forensic Sci Int*, 2003, **138**(1–3): 37–43.

Tully, G., et al., Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forensic Sci Int*, 2001, **124**(1): 83–91.

Tully, G., et al., Results of a collaborative study of the EDNAP group regarding mitochondrial DNA heteroplasmy and segregation in hair shafts. *Forensic Sci Int*, 2004, **140**(1): 1–11.

Wallin, J.M., et al., TWGDAM validation of the AmpFISTR blue PCR amplification kit for forensic casework analysis. *J Forensic Sci*, 1998, **43**(4): 854–870.

Welch, L.A., et al., European Network of Forensic Science Institutes (ENFSI): Evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers. *Forensic Sci Int Genet*, 2012, **6**(6): 819–826.

FORENSIC SCIENCE

FORENSIC BIOLOGY

SECOND EDITION

Over the last several years, new research and developments in analysis methods and practice have led to rapid advancements in forensic biology. Identifying critical points of knowledge and new methodological approaches in the field, ***Forensic Biology, Second Edition*** focuses on forensic serology and forensic DNA analysis. It provides students and professionals with a scientific grounding in biological evidence—both the techniques utilized to identify it and the methodology to analyze it.

This second edition:

- Introduces the language of forensic biology, enabling students to become comfortable with usage and terminology
- Provides clear explanations of the principles of forensic identification and analysis of biological evidence
- Explains forensic serology and DNA techniques used in the field and the laboratory
- Discusses the benefits and limitations that apply to various forensic biology techniques
- Includes schematic illustrations to clarify concepts
- Presents three new chapters (Chapters 2, 16, and 17) created for this edition
- Adds more than two hundred new full-color figures

Covering the full scope of forensic biology, the book uses an accessible, straightforward style designed to enhance students' education and training so they are prepared, both in the laboratory and in the field.

K14098

ISBN: 978-1-4398-8970-1
90000
9 781439 889701

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

www.crcpress.com

Third edition

KNIGHT'S
Forensic Pathology

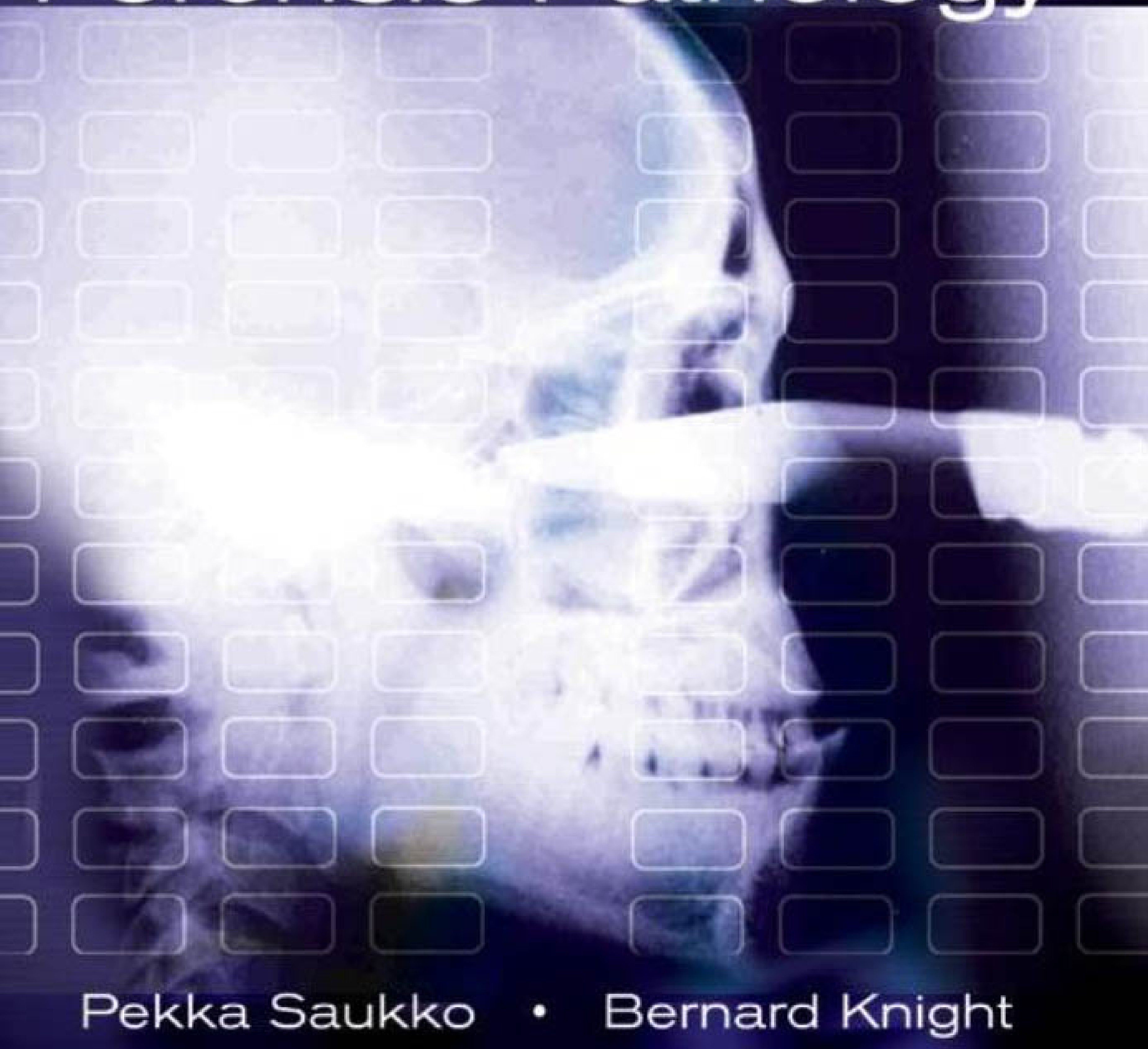

Pekka Saukko • Bernard Knight

KNIGHT'S
Forensic Pathology

Those who have dissected or inspected many bodies have at least learned to doubt, while those who are ignorant of anatomy and do not take the trouble to attend to it, are in no doubt at all.

Giovanni Morgagni 1682–1771
The Father of Morbid Anatomy

Taceant colloquia. Effugiat risus. Hic locus est
ubi mors gaudet succurrere vitae.
(Let conversation cease. Let laughter flee.
This is the place where death delights to help the living.)

Latin proverb

Seldom say never – seldom say always!

Forensic proverb

Third Edition

KNIGHT'S
Forensic Pathology

Pekka Saukko Dr.med.univ. (Vienna), Dr.Med.Sci., Specialist in Forensic Medicine, Dr.h.c.
Professor and Head, Department of Forensic Medicine, University of Turku, Finland

Bernard Knight CBE, MD, DM (Hon), BCh, MRCP, FRCPath, FHKCPath, DSc (Hon), LLD (Hon), PhD (Hon), DMJ (Path), Barrister of Gray's Inn
Emeritus Professor of Forensic Pathology, University of Wales College of Medicine, UK
Formerly Consultant Forensic Pathologist to the Home Office, UK

HODDER
ARNOLD
PART OF HACHETTE LIVRE UK

First published in Great Britain in 2004 by Hodder Arnold,
part of Hachette Livre UK
338 Euston Road, London NW1 3BH

http://www.hoddereducation.com

© 2004 Edward Arnold (Publishers) Ltd

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronically or mechanically, including photocopying, recording or any information storage or retrieval system, without either prior permission in writing from the publisher or a licence permitting restricted copying. In the United Kingdom such licences are issued by the Copyright Licensing Agency:
Saffron House, 6–10 Kirby Street London EC1N & 8TS.

Whilst the advice and information in this book are believed to be true and accurate at the date of going to press, neither the author[s] nor the publisher can accept any legal responsibility or liability for any errors or omissions that may be made. In particular (but without limiting the generality of the preceding disclaimer) every effort has been made to check drug dosages; however it is still possible that errors have been missed. Furthermore, dosage schedules are constantly being revised and new side-effects recognized. For these reasons the reader is strongly urged to consult the drug companies' printed instructions before administering any of the drugs recommended in this book.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978 0 340 76044 4

5 6 7 8 9 10

Commissioning Editor: Serena Bureau
Development Editor: Layla Vandenbergh
Project Editor: Anke Ueberberg
Production Controller: Deborah Smith
Cover Design: Stewart Larking

Typeset in 10/12.5 AGaramond by Charon Tec Ltd., A Macmillan Company
Printed and bound in India

What do you think about this book? Or any other Arnold title?
Please send your comments to feedback.arnold@hodder.co.uk

Contents

Preface to the third edition *vi*
Preface to the second edition *vii*
Preface to the first edition *viii*

1 The forensic autopsy 1
2 The pathophysiology of death 52
3 The establishment of identity of human remains 98
4 The pathology of wounds 136
5 Head and spinal injuries 174
6 Chest and abdominal injuries 222
7 Self-inflicted injury 235
8 Gunshot and explosion deaths 245
9 Transportation injuries 281
10 Abuse of human rights: deaths in custody 301
11 Burns and scalds 312
12 Electrical fatalities 326
13 Complications of injury 339
14 Suffocation and 'asphyxia' 352
15 Fatal pressure on the neck 368
16 Immersion deaths 395
17 Neglect, starvation and hypothermia 412
18 Deaths associated with sexual offences 421
19 Deaths associated with pregnancy 431
20 Infanticide and stillbirth 439
21 Sudden death in infancy 451
22 Fatal child abuse 461
23 Deaths associated with surgical procedures 480
24 Dysbarism and barotrauma 488
25 The pathology of sudden death 492
26 Forensic dentistry for the pathologist 527
27 Poisoning and the pathologist 541
28 Forensic aspects of alcohol 552
29 Carbon monoxide poisoning 560
30 Agrochemical poisoning 566
31 Poisoning by medicines 570
32 Death from narcotic and hallucinogenic drugs 577
33 Corrosive and metallic poisoning 585
34 Deaths from organic solvents 595

Appendix 1 Technical methods *600*
Appendix 2 Council of Europe Committee of Ministers Recommendation No. R (99) 3 *610*

Index *618*

Preface to the third edition

It is pleasant to record that the second edition of *Forensic Pathology* consolidated the international reputation of the original book as one of the foremost manuals of forensic pathology in the English language. This new third edition generally follows the same successful pattern, but there are some significant differences.

The editorship is being transferred to Professor Pekka Saukko and this volume is the product of the collaboration with the original author. As with previous editions, the text has undergone a complete revision to update where necessary and add new material where relevant. A major change is the extensive use of colour to replace original monochrome illustrations and in numerous new illustrations, including micrographs. More than 300 colour illustrations have been added, of which 175 are completely new from the archives of both the Cardiff and Finnish institutes, and others. Another major improvement has been in the reference material where both corrections and numerous additions have been made. 'Further reading' is intended for those readers who do not have access to electronic libraries and to offer an overview of the literature on a particular topic.

The new edition maintains the philosophy of evidence-based forensic pathology with emphasis on the avoidance of over-interpretation, which regrettably still leads to instances of miscarriage of justice.

It is hoped that the new edition of this well-established textbook, with its emphasis on practical procedures and common-sense evaluation of autopsy findings, will continue to be of assistance to forensic pathologists all over the world.

Pekka Saukko
Bernard Knight

Preface to the second edition

The most gratifying reception accorded in many parts of the world to the first edition of this book has led to the need for an updated version less than five years later. No radical changes have been made in the format, but like an airliner during its periodic major service, every part has been scrutinised and replaced where it has become outdated or defective. In addition, new material has been added where appropriate.

Forensic pathology, unlike its sister discipline forensic science, does not change quickly or have dramatic developments such as the current DNA revolution. Indeed, because its base is the interpetation of autopsy findings, forensic pathology still rests largely on the principles of morbid anatomy founded in the nineteenth century and earlier. However, this is not to say that it remains fossilised in the era of Virchow – new findings and techniques appear constantly, albeit at a measured pace compared with other disciplines. The most useful of these advances have been incorporated into this new edition.

In particular, the sections on child abuse, head injuries and traumatic subarachnoid haemorrhage have been amended or supplemented. Sixty new illustrations have been incorporated, and another area of attention has been the references and recommended reading, as there were some textual citations not listed in the first edition – and very many more have been added. However, the literature is now so vast, both in forensic journals and scattered profusely throughout other specialist publications, that it is futile to try to capture all of even the most seminal papers, which now need computerised and other modern library techniques for their retrieval.

One of the many reviewers of the previous edition complained that many of the references were old, but I make no apology for this. The value of a publication is in its content, not in its date, as Harvey's *De Motu Cordis* and Morgagni's *De Sedibus* clearly demonstrate! Some of the most valuable papers in forensic pathology were written decades or even a century ago; the critical writings of Moritz, Shapiro, Adelson, Helpern, Gonzales, Polson – right back to Taylor and Tardieu in the nineteenth century – are examples of careful observation and logical thinking which some modern pathologists, given to overinterpretation, would benefit from studying.

The critical attitude to every autopsy finding has been maintained in this new edition. A number of recent criminal trials and Appeals have emphasised the vital need for expert medical witnesses to be totally objective in their interpretation of physical findings. To quote from the new (1996) *Practice Guidelines* of the Policy Advisory Board in Forensic Pathology of the British Home Office: 'The evidence should be objective: speculation should be avoided. Unwarranted conclusions can never be defended. The role of the expert witness is not to provide the evidence which supports the case for the Crown nor for the defence, unless that opinion is objectively reached and has scientific validity.' Perhaps, in the interests of justice, an additional new forensic aphorism should be on the frontispiece: 'If you can't prove it, don't claim it!'

Bernard Knight
1996

Preface to the first edition

This is a textbook of forensic *pathology*, not forensic *medicine*. Though there is a considerable overlap, forensic medicine includes medical jurisprudence, the legal aspects of medical practice and many ethical matters, none of which are found – nor are intended to be found – in these pages.

Such topics have marked geographical limitations, as a result of legal, ethnic, cultural, and even religious variations from place to place. Even within the British Isles it is almost impossible to write a single book on medical jurisprudence that can do equal justice to the different legal systems of England and Wales, Scotland and Ireland.

The subject matter of this volume has no such geographical constraints, as it is solely concerned with the examination of the dead body for medicolegal purposes. Even in this limited sphere, police procedures and the habits of pathologists will vary considerably with country and with resources, but it is hoped that the routines, techniques and philosophy offered in this book will offer a guide to good practice that can then be modified according to local circumstances.

The contents are intended to lead the pathologist – and in some countries, the non-pathologist – through the procedures that are needed in the examination of a body found under obscure, suspicious, or frankly criminal circumstances.

In developed countries with sophisticated medicolegal systems, such autopsies will be performed only by forensic pathologists or by histopathologists with considerable experience, but in many areas of the world – especially the developing countries – lack of manpower and resources as well as considerations of distance and facilities mean that almost any doctor may be called upon to perform medicolegal examinations. For both of these classes of medical men and women this book aims to act as a guide and a source of reference.

The subject matter follows a fairly conventional pattern, but the treatment of each topic is designed to offer practical advice linked with a philosophical approach that leads the doctor to analyse and question the interpretations drawn from physical findings.

All too often, dogmatic opinions are derived from an unsound factual base, learned from lectures or textbooks that repeat previous dogma with little sense of critical evaluation. In some parts of the world forensic pathology is learned by rote from teachers who studied it themselves in the same fashion, and who have little or no practical experience in the hard schools of mortuary or witness-box. I hope that this book will at least stimulate trainee forensic pathologists to think twice, question and disagree – not least about some of the material in the following pages.

Some conventions have been discarded in the format of this book. The author feels that it is disruptive for ease of reading to have every line clogged with references to other authors, except where essential to the topic under discussion. Thus most references have been consolidated into groups, placed at the end of the appropriate sections in each chapter. In addition to actual references, there are many suggestions for further reading on the same topic. It is quite impossible to be comprehensive in offering relevant titles, as the amount published on forensic matters is now overwhelming, several hundred new titles appearing each month. A representative selection is offered, much of it quite recent – these papers themselves will then offer a relevant bibliography, so that the topic can be followed in almost geometric progression.

There is no discussion of forensic serology, as this is now a discipline of its own. Similarly, the toxicology is confined to a detailed consideration of the pathologist's role in obtaining samples and interpreting the results from the analytical laboratory. Though the most common poisons are discussed, this is purely from the autopsy aspect. Like serology, toxicology is now a vast subject that stands independently and it seems futile merely to scratch the surface in a text devoted to forensic morbid anatomy.

Another departure from the conventional text is that concise reminders of the forensic anatomy of important regions are given, adjacent to the topic under discussion. Many pathologists learned their anatomy a long time ago and it can be useful before an autopsy, or writing the report, or appearing in court, to have a quick refresher about the relations of the aortic arch to the second rib or the exact structure of the laryngeal cartilages. Illustrations in forensic textbooks are traditionally profuse and often spectacular. The only justification for a photograph is that it instructs, however, rather than entertains or shocks. This book uses photographs only where a relevant point is displayed, with a full caption that stands alone without having to refer to the text.

Many line drawings are used, as the author feels that they often get the message across better than a photograph, which of necessity contains irrelevant and perhaps distracting features. Similarly, many books are replete with anecdotal cases; these can be valuable, but often are more intriguing than useful. The space devoted to these can usually be put to better use and they are employed sparingly in this text.

Where the pronoun 'he' is used throughout this book, it is intended that the word 'she' is equally applicable, unless obviously inappropriate from the context.

Finally, an Appendix is offered with basic reference material on weights, sizes, temperatures, conversion tables and other useful facts. Histological and histochemical techniques relevant to forensic needs are given in detail.

Bernard Knight
1990

This page intentionally left blank

CHAPTER 1

The forensic autopsy

- Autopsy standards 1
- Types of autopsy 2
- The procedure for a forensic autopsy 2
- Examination of the scene of death 4
- Property, clothing and identification 8
- The use of the history of the case 9
- Precautions regarding potential infective conditions 10
- The autopsy: external examination 11
- The autopsy: internal examination 16
- Examination of organs 23
- Examination of the brain 28
- Ancillary investigations 29
- Autopsy radiology 31
- Forensic photography 32
- The autopsy report 33
- Post-mortem artefacts 36
- Exhumation 36
- The autopsy on the putrefied corpse 39
- Resuscitation artefacts at autopsy 40
- Mass disasters – the role of the pathologist 41
- The obscure autopsy 44
- References and further reading 47

Though 'necropsy' is semantically the most accurate description of the investigative dissection of a dead body, the word 'autopsy' is used so extensively that there is now no ambiguity about its meaning.

The term 'post-mortem examination' is a common alternative, especially in Britain, where its meaning is never in doubt. Unfortunately, it suffers from a lack of precision about the extent of the examination, for in some countries many bodies are disposed of after external examination without dissection.

The forensic investigation of deaths was introduced relatively early from the requirements of the judicial system. The earliest known forensic dissections took place in Italy, probably in the middle of the thirteenth century, at the University of Bologna. The first one was recorded by William of Saliceto, a surgeon and a teacher on the medical faculty there. He lived from 1210 to 1277 and in his book *Surgery* he mentioned a case examined in about 1275. The principles of the modern medico-legal investigation were developed based on the codes of the sixteenth century Europe: the Bamberg Code in 1507, the Caroline Code in 1532 and later the Theresian Code in 1769. The hospital or clinical autopsy became meaningful first after the introduction of modern concepts of pathogenesis of diseases by Carl von Rokitansky (1804–1878) and cellular pathology by Rudolf Virchow (1821–1902). Further development of medico-legal autopsy has been greatly influenced by the judicial system adopted, the main emphasis in most countries being in the detection and investigation of criminal and other unnatural or unexpected deaths.

AUTOPSY STANDARDS

Discussions about the quality of the autopsy started at the beginning of the nineteenth century. The Austrian decree of 1855 included, in its 134 paragraphs, very detailed instructions as to the performance of medico-legal autopsies. The idea of quality control and quality assurance, first introduced into industrial production of goods in the twentieth century, found its way into medicine later via laboratory medicine. Although regional and national differences in this respect are huge and the pace of progress varies, there is need for standards in specialist education, continuing professional development, harmonization of investigative procedures and quality systems leading perhaps eventually to some form of accreditation of individuals and institutions.

In addition to the national measures to create guidelines and to harmonize the medico-legal autopsy, there has been an increasing international interest in achieving harmonized and internationally recognized rules on the way autopsies should be carried out. In 1991, the General Assembly of the United Nations endorsed the Model Autopsy Protocol of the United Nations. The European Council of Legal Medicine (ECLM) document 'Harmonisation of the Performance of the Medico-Legal Autopsy' was adopted by the General Assembly in London 1995. The latter document served largely as basis for the Pan-European Council of Europe – *Recommendation No. R (99) 3 On the Harmonisation of Medico-Legal Autopsy Rules and Its Explanatory Memorandum* – which was adopted by the Committee of Ministers in February 1999 (see Appendix 2, p. 610). Although the document is a 'recommendation' by nature, it has legal implications, because Council of Europe member countries have to implement these principles into their national legislation.

TYPES OF AUTOPSY

Though medical conventions and legal systems vary considerably from country to country, there are generally two main types of autopsy:

- The **clinical** or **academic autopsy** is one in which the medical attendants, with the consent of relatives, seek to learn the extent of the disease for which they were treating the deceased patient. In most jurisdictions this type of autopsy should not be held to determine the nature of the fatal disease because, if this was unknown to the physicians, the death should have been reported for medico-legal investigation.
- The **medico-legal** or **forensic autopsy**, which is performed on the instructions of the legal authority responsible for the investigation of sudden, suspicious, obscure, unnatural, litigious or criminal deaths. This legal authority may be a coroner, a medical examiner, a procurator fiscal, a magistrate, a judge, or the police, the systems varying considerably from country to country.

In most systems the permission of the relatives is not required, as the object of the legal investigation would be frustrated if the objections of possibly guilty persons could prevent the autopsy.

In many jurisdictions the medico-legal autopsy is often further subdivided into:

- those held on apparently non-criminal deaths, such as accidents, suicides, deaths from sudden natural causes, or associated with medical and surgical treatment, industrial deaths, and so on
- the truly forensic autopsy held on suspicious or frankly criminal deaths, usually at the instigation of the police. These deaths comprise murder, manslaughter, infanticide and other categories that vary in different jurisdictions.

The type of pathologist that deals with these categories also varies from place to place but, as the systems are so diverse, there is little point in discussing the details. What is much more important is that whichever pathologist tackles each type of case, he or she should be trained and experienced in that particular field. Unfortunately, either from lack of staff and resources or because the system is deficient, medico-legal autopsies – especially of major criminal cases – are frequently performed by pathologists inexperienced in forensic procedures.

Another serious defect in many countries is the separation of those who practise forensic pathology from those who profess to teach it in universities. It is impossible to be a credible and convincing teacher unless one has continuing practical experience of the subject.

Even more unfortunate is the performance of medico-legal autopsies by doctors who have no training in pathology at all. Though lack of resources make this a widespread and inevitable practice in many developing countries, the same regrettable practice is seen in some parts of Europe, the Americas and elsewhere. This is to the detriment of the high standard of expertise that is vital for the support of law enforcement and the administration of justice. A poor opinion is often worse than no opinion at all, as in the latter case, the legal authorities will at least be aware of the deficiency in their evidence, rather than be misled by the often dogmatic inaccuracies of an inexperienced doctor.

THE PROCEDURE FOR A FORENSIC AUTOPSY

Though the actual performance of an autopsy is fairly uniform whatever the nature of the death, there are a number of associated matters that vary according to the circumstances. For example, the procedural precautions required in a murder are not necessary in a sudden natural death, and the dissection in a criminal abortion or fatal rape is different from that upon a drowned body.

There are, however, many facets of the autopsy that are common to every death. These will be discussed in turn here, though the full significance of each item is discussed in detail in other chapters.

The objectives of an autopsy

- To make a positive identification of the body and to assess the size, physique and nourishment.
- To determine the cause of death or, in the newborn, whether live birth occurred.
- To determine the mode of dying and time of death, where necessary and possible.
- To demonstrate all external and internal abnormalities, malformations and diseases.
- To detect, describe and measure any external and internal injuries.
- To obtain samples for analysis, microbiological and histological examination, and any other necessary investigations.
- To retain relevant organs and tissues as evidence.
- To obtain photographs and video for evidential and teaching use.
- To provide a full written report of the autopsy findings.
- To offer an expert interpretation of those findings.
- To restore the body to the best possible cosmetic condition before release to the relatives.

Preliminaries to an autopsy

Before the body is even approached, a number of preliminaries are necessary.

AUTHORIZATION AND CONSENT

A medico-legal autopsy is carried out at the behest of the appropriate authority. The pathologist must not begin his examination until he is satisfied that such authority has been issued in respect of that particular death. The means of delivering such authority will vary from place to place: it may be a written document, a verbal or telephoned message, or a tacit standing arrangement.

Where two official organizations are involved, it must be clear who has the premier right to order an autopsy. For example, in England and Wales, the police may request a pathologist to examine a body externally at the scene of death, but the right to order an autopsy is the sole prerogative of the coroner. Though in serious incidents he or she should take the advice of the Chief of Police regarding the choice of a pathologist, the final decision remains that of the coroner.

As stated earlier, in a medico-legal autopsy the relatives are not consulted for their consent to autopsy, as such permission cannot be withheld. As a matter of courtesy it is usual to inform them before the autopsy is carried out unless the circumstances are too urgent. In most cases, however, a relative will have attended to identify the body before the examination begins. Pathologists should be cautious about performing autopsies where the authorization or consent aspects are not clear. It sometimes happens that a 'consent document' for a hospital death has been obtained but the case has later been reported for medico-legal investigation. The pathologist should not then proceed under the 'consent' authority until the coroner or other official has decided whether to order his own autopsy (perhaps using another pathologist). If that official decides not to proceed, then usually the consent document becomes valid once more.

Permission for the retention of material from an autopsy (ranging from small fluid samples to the entire body, if necessary) is usually covered in a medico-legal case by the original authority issued for the examination. There is, however, considerable variation in the legal aspects of tissue and organ retention in different countries, and each pathologist must become fully conversant with local regulations.

In the legislation controlling the English coroner, it is not only permissible, but also obligatory, for the pathologist to retain any tissue that may assist in the further investigation of the death.

However, the coroner (or procurator fiscal in Scotland) has only a negative power to forbid the use of tissues or organs for purposes other than the investigation of the death, and so cannot give positive consent for the use of these for transplantation, teaching or research; such permission must be obtained from the relatives, if the coroner does not object, under the Human Tissue Act 1961 and the forthcoming Human Tissue Act 2005.

Who may be present at an autopsy

Once again, each nation or state has different legislation and regulations on the conduct of medico-legal examinations. Though it is pointless to try to survey these, it must be emphasized that every pathologist must learn and abide by the laws applicable to his or her position.

Often the relatives of the deceased person, or an accused person, must be informed of the place and time of the autopsy so that they may be represented by a lawyer or doctor acting on their behalf. In some jurisdictions, including Britain, there is nothing that specifically forbids the relatives attending in person and the authors have had several experiences of this nature. It is open to the pathologist to refuse to conduct the autopsy if he objects to their presence.

In many homicides, a second autopsy is performed by another pathologist acting on behalf of the defence lawyers representing the accused person. This usually takes place at a later date, after the accused has been charged and granted legal representation, but sometimes the second pathologist will attend the original autopsy. This is quite in order and is more valuable than a later exploration of extensively dissected and autolysed tissues.

The permission for such attendance is granted by the coroner, medical examiner, or equivalent legal authority. The first pathologist should grant his colleague every courtesy and facility in the expectation of similar behaviour when the roles are reversed.

Others entitled to be present naturally include the officials or deputies of the department ordering the autopsy – for example, a coroner, magistrate or judge. The police, including their technical teams, are also present if the death is criminal or suspicious. Whether other doctors and medical students are allowed depends upon the wishes of the official commissioning the examination. When the deceased has been under medical care before death, it is almost invariable practice to allow – and indeed encourage – the physician to be present, as he has the best knowledge of the medical history.

In a criminal or suspicious case, the pathologist should try to limit the number of those present to a minimum. Not only is there a greater risk of loss of confidentiality, but sheer physical numbers, especially of large policemen, can make the mortuary overcrowded. This hampers movement, causes distraction, and adds to the risk of infection and contamination, especially with the present concern about the various forms of hepatitis (B, C, D, E), tuberculosis and human immunodeficiency virus (HIV) infectivity. No one should be present merely as a casual observer, not even senior police officers not directly involved in the investigation. With the increased sophistication and complexity of forensic and police procedures, more and more people cram into the mortuary, cluttered with cameras, videos, recorders, scene-of-crime kits and so on, until there is hardly room for the pathologist to move.

Whoever is present at an autopsy on a criminal or suspicious death should be listed and named by the pathologist on the autopsy report.

EXAMINATION OF THE SCENE OF DEATH

In homicide, suspected homicide, and other suspicious or obscure cases, the pathologist should visit the scene of the death before the body is removed.

Once again, local practice varies but any doctor claiming to be a forensic pathologist should always make himself available to accompany the police to the locus of the death. This duty is often formalized and made part of a contract of service for those pathologists who are either full-time or substantially involved in assisting the police. In England and Wales, the 'Home Office Pathologists' are permanently on call for such visits and in many other jurisdictions, such as the medical examiner systems in the USA, and the European State and University Institutes of Forensic Medicine, there is usually a prearranged duty roster for attendance at scenes of death. The function of an experienced forensic pathologist at the scene of death is generally to assess the environment, the local circumstances, and the position and the condition of the body. In a large proportion of instances crime can rapidly be excluded in favour of accident, suicide or even natural causes. This is a most useful and cost-effective function, as a spurious murder investigation involving perhaps scores of police, many vehicles and other expensive public facilities, may be disbanded promptly when the pathologist recognizes innocent circumstances.

The pathologist should always have appropriate equipment ready to take to a scene investigation at a moment's notice. Further equipment may be carried if autopsies have to be carried out in places where good mortuary facilities are not available. Most forensic pathologists carry a 'murder bag' in their car and though every expert has his own choice of equipment, the following is a reasonable inventory:

- waterproof apron and rubber gloves
- thermometer, syringes and needles, sterile swabs
- autopsy dissection set, including handsaw
- cutting needles and twine for body closure
- swabs and containers for blood and body fluids
- formalin jars for histological samples
- plastic bags, envelopes, paper, spare pen and pencil
- printed body charts for recording external injuries
- hand lens, electric torch, mini-tape recorder
- camera, usually 35 mm single-lens reflex with electronic flash. The recent advent of compact digital cameras or digital video cameras with the facility to take still pictures has made instant reviewing possible.

The thermometer can be either a long chemical mercury type, reading from 0 to 50°C, or the more modern electronic digital variety with a probe carrying a thermocouple. The amount of equipment varies with the facilities likely to be available. In developed countries there are likely to be good mortuary facilities available in a hospital or municipal mortuary and the police forces will have extensive scenes-of-crime expertise with photography, specimen containers and so on. In developing countries and the more remote areas of other states, the pathologist may have to be virtually self-sufficient in respect of both crime investigation and the subsequent autopsy.

In addition to medical kit, the experienced forensic pathologist will always have appropriate clothing such as rubber boots and rain- or snow-wear ready to hand for any call.

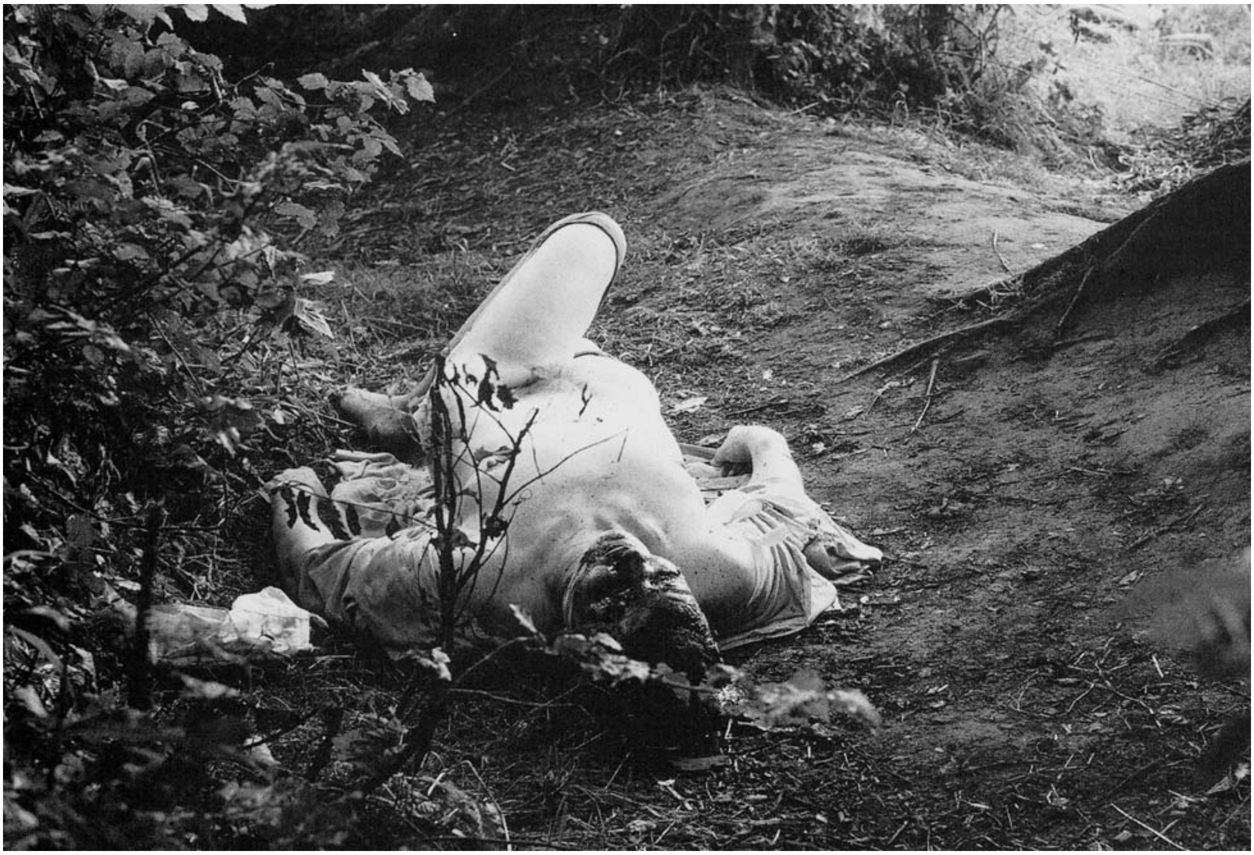

FIGURE 1.1 *The scene of a rape-homicide. The attending pathologist must record the posture, state of clothing and general relationship of the victim to the surroundings, but limit the examination at the scene to avoid any interference with the vital process of obtaining trace evidence. Rectal temperatures should not be taken until all swabs are completed, usually in the mortuary. Temperatures can be taken in the ear or deep nostrils, though with severe head injuries (as here) even this must be done with care.*

When at the scene of death, the pathologist's actions are dictated by the particular circumstances. Much will depend upon the availability or otherwise of police and forensic science assistance. In Britain, for example, several teams converge on a scene of crime, including photographers and video operators, and Scene of Crime Officers (SOCOs) whose function is to collect trace evidence. Scientists from the nearest forensic laboratory often attend with their police liaison officers, as well as fingerprint officers and, of course, the investigating officers from the Criminal Investigation Department.

Where no such backup is available, the pathologist must try to collect trace evidence himself, but he should remain within the limits of his own expertise.

The pathologist should accept the instructions of police officers in relation to the approach to the body so as to preserve the immediate environment as much as possible. Out-of-doors access is often limited to a single pathway marked by tapes, and in a building a track to the corpse is usually pointed out by the detective in charge.

The doctor should not touch anything unnecessarily and certainly not smoke or leave any object or debris of his or her own. Increasingly, those visiting the scene of a crime are given disposable overalls and overshoes to wear, so that fibres, hairs and so on from the visitor are not spuriously transferred to the scene.

If a forensic doctor has to act both as pathologist and clinical medico-legist or 'police surgeon' (that is, examining live victims or suspects), he or she must change clothing or wear a new set of protective garments to avoid transferring trace evidence, such as fibres or hairs, from victim to suspect.

The pathologist should observe a great deal, but do very little. He or she should note the position of the body in relation to nearby objects and establish the plan of the premises if indoors. A sketch or his own photograph is sometimes useful, and some pathologists use a Polaroid, digital or video camera for instant recording of the death scene.

Any obvious cause of death should be observed, and any blood pools or splashes noted in relation to the position of the corpse. The shape of such splashes should be observed,

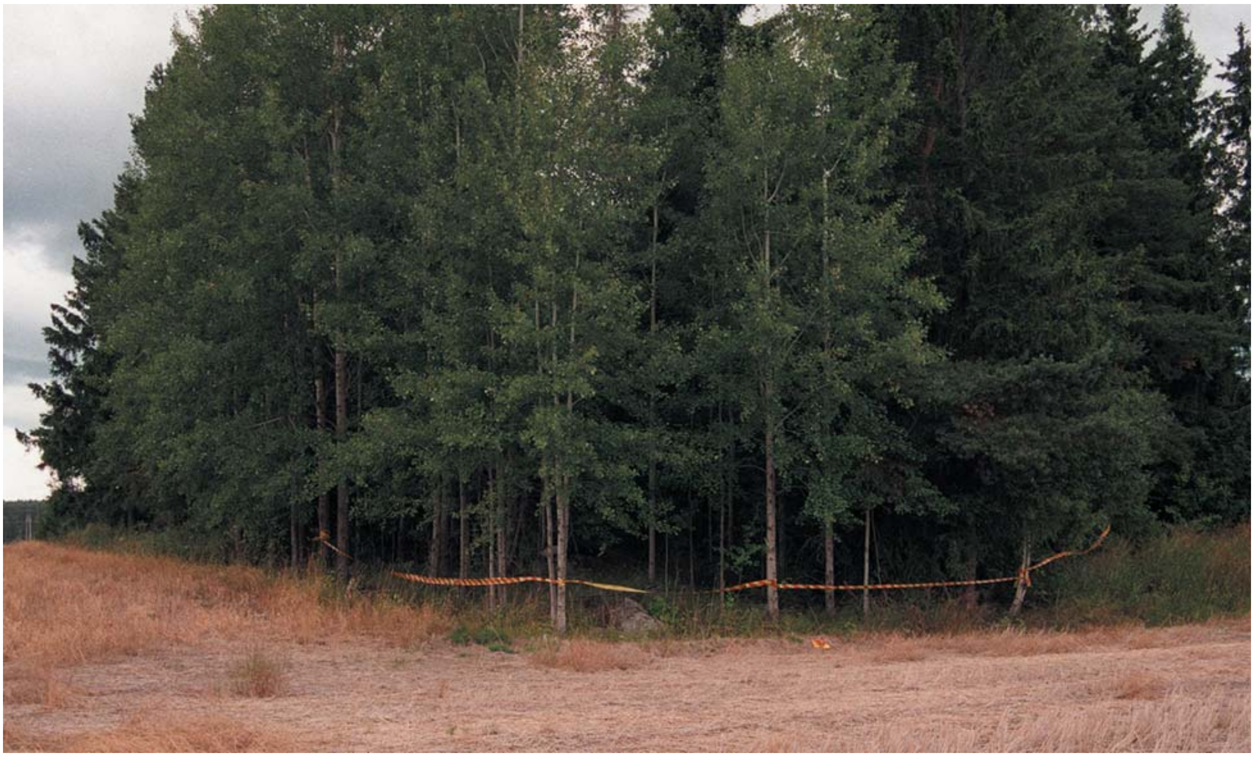

FIGURE 1.2 *Scene of a homicide-suicide where the locus of bodies should be widely marked off by tape to restrict access, and the actual body protected by a shelter from the weather and unauthorized onlookers.*

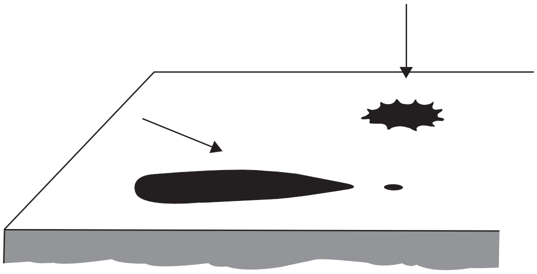

FIGURE 1.3 *At a scene of crime, the position of the body at the time of an assault can sometimes be clarified by observing the direction of blood splashing that may have been thrown off a weapon or blood-stained hands. Blood striking vertically causes round or crenellated circles, while blood hitting at an angle is lance-shaped with the narrow end facing the direction of travel. There is often a preceding drop of blood in advance of the major splash, forming an 'exclamation mark' stain.*

as blood striking perpendicularly to a surface leaves a circular mark, whilst that landing obliquely is pear-shaped, with the sharper end towards the direction of flight.

When photographs have been taken of the body in its original position, the pathologist then approaches it, after checking with the investigating officers that it is appropriate for him to do so. Close examination can be made and the skin felt to assess temperature. The eyes, neck and hands can be examined and where necessary, clothing gently moved aside to look at the throat or upper chest. Any relevant findings should be photographed by the police before further disturbance.

If forensic scientists or scene of crime officers require any samples at this stage, their wishes must be respected. They may want to 'tape' the body, that is, press adhesive tape across the skin and clothing to capture any loose hairs and fibres. The body can then be moved to look at the sides and undersurface, again with caution so that no weapon, other object or trace evidence is disturbed. No set routine can be advised, as each case has its own individual aspects. The prime object is to assess the whole scene while causing the least disturbance to the body and clothing before the scientific teams have finished their tasks.

Estimating the post-mortem interval at the scene

The last sentence above raises the matter of estimating the time since death. The problem is fully discussed in Chapter 3, but has relevance here in relation to the pathologist's actions at the scene of death.

The general warmth or coolness of the hands and face can be assessed by touch, and the degree of rigor mortis felt by gently testing the limbs. The ambient (environmental) temperature must be taken as soon as possible after the discovery of the body, preferably by police scene of crime officers who usually arrive at the locus before the pathologist. The ambient temperature should be taken as near to the body as possible, as microenvironments can exist, even inside buildings or rooms. Information should be sought as to how much disturbance of the ambient temperature might have occurred, such as opening doors and windows, or turning fires or central heating on or off, so that some idea of post-discovery distortions of temperature can be estimated later. The insertion of a thermometer into the rectum at this stage in the investigation, as advocated by some textbooks, is controversial.

At a scene of death, this usually means either pulling down trousers or pants, and otherwise disturbing clothing, often in cramped and ill-lit places, frequently out in the open. It also risks contaminating the rectum and perineum, by introducing seminal fluid from the anal margin into the rectum, making subsequent examination of that area (and taking swabs for semen) of reduced value. As so many violent crimes now have sexual or homosexual overtones, the practice of taking rectal temperatures at the scene should be performed only if the forensic scientists or police scene of crime officers are satisfied that trace evidence from the clothing, swabs from the vulva, vagina and anus, etc., can be obtained satisfactorily before rectal thermometry is performed.

In other words, a cost–benefit analysis must be made at the scene, to decide if the difficulties of taking a rectal temperature are worth the small potential advantage of an earlier measurement. In many cases, where the body has obviously been there long enough for the core temperature to have reached ambient – or where other circumstantial evidence has indicated that the time of death is known to a greater degree of accuracy that can be hoped for by thermometry – then nothing is lost by postponing the procedure until the body arrives at the mortuary for autopsy, which, in British practice, is usually directly after the body is moved from the scene.

If the autopsy is to be delayed for many hours owing to difficulties with transport or lack of facilities, then much more must be done at the scene and temperature measurements are justified.

An alternative is to use a place other than the rectum. The axilla and mouth give low readings, which cannot reliably be correlated with the deep temperature because of variable exposure to the air temperature. More useful is the auditory meatus or nostril, the thermometer or thermocouple probe being inserted as deeply as possible. Reliable, reproducible readings can be obtained from these sites, which have the great advantage of being easily accessible without moving clothing, as well as not being required for swabbing to investigate possible sexual assaults. The use of temperatures to attempt to calculate the post-mortem interval is discussed fully in Chapter 3.

When the pathologist has made the best examination possible in the circumstances, his next function is to ensure that the corpse is removed to the mortuary for autopsy with the least disturbance and loss of evidence. He should supervise the removal himself or at least delegate the duty to another person whom he knows is careful and competent. Each hand should be enclosed in a bag, secured at the wrist by adhesive tape or string. A similar bag should be placed over the head. The packaging medium may vary, but generally paper bags are recommended.

The body should be placed gently in a ‘body-bag’, which has a zip closure, or moved on to a large, new plastic sheet, at least 2 metres square. If a sheet is used, the edges should be wrapped over the body and secured with adhesive tape. The object of the exercise is to retain any loose objects, hairs and fibres that may be adhering to the body or the clothing. The sheet or bag is taken by the forensic laboratory after the body is removed in the mortuary so that they may screen it for trace evidence. The transport of the body is the responsibility of the police or other agency such as the coroner or his officer. The body in its plastic wrapping should be placed in a rigid fibreglass ‘shell’ or ordinary coffin, and taken by hearse, van or police transport to the chosen mortuary.

Physical damage during the removal should be avoided as much as possible, though in difficult or inaccessible sites this is easier said than done. In fires, the body may be seriously damaged before or during recovery, sometimes because its presence is not suspected in the smoke-filled, often waterlogged, debris of a conflagration. The author (BK) has experience of such a scene where firemen walked over two bodies for several hours before it was realized that they were buried under burnt furniture and other debris. Handling brittle, charred, bodies can easily cause the splits at joints that may mimic ante-mortem injuries (see Chapter 12).

In summary, the function of a pathologist at any scene of suspicious death is to observe the situation, to conserve any fragile evidence, to supervise the removal of the body and offer an opinion, based on experience, about the nature of death where this can reasonably be done. He is not there to act as a latter-day Sherlock Holmes, voicing unsubstantiated theories on non-medical matters, nor attempting to overinterpret the situation from the flimsiest of facts. The pathologist is part of a team of specialists, all experts in their own field, and it is as a member of such a cooperative, coordinated group that his best contributions can be made.

PROPERTY, CLOTHING AND IDENTIFICATION

Whether or not the pathologist has been to the scene of a death, he should take notice of the clothing and other property of the body upon which he is to carry out an autopsy. This applies not only to every criminal or suspicious death, but also to many traffic and industrial accidents, as well as to the victims of falls from a height, drowning and so on.

In many instances, there will be no opportunity to examine the clothed body in the mortuary: if death has occurred in hospital or in an accident department, the clothing may have been removed before transfer to the mortuary. The pathologist should make a permanent request that a body from any traffic accident, or other case where trauma was the provisional cause of death, should be brought to the mortuary without having the clothing removed by police or nursing staff – unless of course the victim was still alive on arrival at hospital, and undressing was performed before attempted resuscitation or treatment. As a second best, the removed clothing should accompany the body to the mortuary so that it can be examined, if necessary, for damage, stains and other evidence. Unfortunately, such clothing is frequently destroyed before it can be seen.

The mortuary staff should be trained to regard clothing and property as important items of evidence, and a system should be established to retain, identify and store these, both from the evidential aspects and for the safety of valuables. The contents of the pockets, documents, keys and other items all assist in identification. Though primarily the task of the police, the pathologist will sometimes have an interest in this aspect. The clothes themselves, the style, fabric, colour and labels all assist in identity.

In trauma deaths, the injuries on the body should be matched up with damage on the clothing. Tears, slashes, stab wounds and especially firearm wounds in the clothes must be compared with the position of external lesions on the body, making allowance for movement and displacement during life. Some self-inflicted wounds may be confirmed by the non-alignment of the clothing damage with the injuries (see Chapter 8).

Blood, seminal, vaginal and other body secretions may be found on the clothing, and though this is primarily the responsibility of the forensic science laboratory, the pathologist may be the first or only person to detect their presence. In firearm deaths, gunshot residues on the clothing may be vital evidence about the range of the discharge and the identity of the ammunition.

In traffic fatalities, tearing of the clothes, grease marks, road dirt, broken lamp or windscreen glass, and even metallic or paint fragments from the vehicle may all assist in reconstructing the event and in identifying the unknown vehicle in a ‘hit-and-run’ tragedy.

Other objects associated with the body that may be helpful include medicines, which may assist in determining the nature of the disease from which the deceased had suffered, for example, amyl nitrite or insulin. In some suicides, empty drug or poison containers may be with the body. Other helpful artefacts include such items as hearing aids, syringes, external pacemakers and inhalers. The clothing must be removed carefully and, especially in criminal or suspicious cases, the pathologist should supervise and assist the mortuary technician, especially as some technicians are not always aware of the importance of clothing in the reconstruction of events.

If the body is not bleeding or otherwise fouled, it is best to remove clothing in the usual way by pulling over the head and limbs, unless this might interfere with any injuries or soiling. If rigor is intense or if there is blood on the face or hands, it may be advisable to cut off some or all of the clothing. This should be done after consultation with the forensic scientists, if they are present, so that the cuts will be made where they will least interfere with later laboratory examination. In any event, cuts should avoid passing through pre-existing damage or staining of the garments. Each item of the clothing should be placed separately into a paper bag to allow them to ‘breathe’.

Identification of the body

The identification of an unknown body is a major forensic exercise, and is fully discussed in Chapter 4. Here, however, we are concerned with the formal recognition of the subject of an imminent autopsy. Before a pathologist makes any examination (and certainly before he begins any mutilating dissection), it is essential that he ensures that the corpse before him is indeed the correct person.

In every medico-legal case, some responsible person must have identified the body. This is usually a relative or close friend of the deceased person, who looks at the face of the body and verbally certifies to a police officer, mortuary attendant or doctor that it is indeed ‘John Smith’. Where the body is burned beyond recognition, mutilated or putrefied, attempts at identification must be made by showing the relative documents, or items such as clothing or jewellery.

In non-criminal cases, such as sudden deaths and most accidents and suicides, the continuity of identity is carried on by a label or tag attached to the body by the police officer, nursing staff or mortuary attendant, which carries the name, address, serial number and other relevant details. This label may be tied firmly to a toe, or a durable wrist or ankle bracelet may be used. Some mortuaries write the name on the leg with an indelible marker, but this can be smudged or obscured during autopsy.

The pathologist should always satisfy himself of the identity by comparing the documents authorizing the autopsy with the toe label or bracelet. The details should agree and, if they do not, he should not proceed with the autopsy until the discrepancy is cleared up, if necessary by calling back the police officer or even the relative to make absolutely sure of the identity.

Many mistakes have been made in the past, with adverse publicity, embarrassing enquiries and even legal consequences. Autopsies on the wrong person, incorrect causes of death, relatives attending the wrong funeral and even cremation of the wrong body are regularly reported as a result of laxity in identification. To avoid confusion, the body should always be labelled as soon as it arrives in the mortuary. Labels on shrouds become detached and those on refrigerator doors are unreliable, as other parties (such as night porters) may shift occupants around without the knowledge of the pathologist or regular mortuary staff.

Some large mortuaries use special means of identity. One known to the authors has an automatic camera fitted in the ceiling of the mortuary entrance. Every new body wheeled in is photographed on its trolley with a serial number prominently displayed on a board placed across the upper chest, so that the face appears with the record number to resolve any later discrepancy.

In serious cases such as homicide, the pathologist must not rely on second-hand evidence of identity such as a toe label. Before the autopsy, someone must verbally confirm to him that the body they are both then viewing is 'John Smith'. This person must be either a relative or the police officer to whom a relative has already formally identified the corpse.

The pathologist must record in his autopsy report the date, time and particulars of the person identifying the body to him, so that 'continuity of evidence' is ensured for legal purposes, which cannot successfully be challenged by the defence at a subsequent trial.

THE USE OF THE HISTORY OF THE CASE

As in clinical medicine, the history of the deceased 'patient' or victim is a vital and indispensable part of the investigation. The extent to which it should influence the pathologist in coming to a decision as to the cause of death, however, is more controversial, as is discussed below.

In medico-legal autopsies – compared with hospital 'clinical' cases – the history is often scanty, absent or misleading. If a person has been found dead, having had no previous medical attendant, there may be virtually no information at all. In criminal deaths, the person who knows most about the death may well be the perpetrator, who naturally may remain silent or give a distorted, misleading or totally false account of the circumstances. The forensic pathologist then often has to contend with a scanty or incorrect history.

Even when a story is given in good faith by relatives, the medical facts are often distorted because of incomplete knowledge and understanding. It may be further warped by transmission by a police officer or other non-medical person, so that only a garbled history reaches the pathologist. The latter should attempt to supplement deficiencies in information by further requests to the commissioning authority, such as the coroner or the police. Wherever possible, the pathologist should try to contact the doctor who may have treated the deceased in the recent past. In the urgency of forensic work, however, frequently at night or weekends, the autopsy may have to proceed without any additional history.

Though more detail may be obtained afterwards, the course of the autopsy might well have been different if such knowledge was available beforehand. For example, the author (BK) once performed an autopsy on a coal miner with no history other than a knowledge of his occupation and chronic chest disease. Autopsy revealed sufficient cardiopulmonary disease to account for death and the body was released for burial. It was not until the following day that the police in the rural area where the man had died, tardily produced the information that both an empty bottle of sleeping tablets and a suicide note had been found with the body. Paradoxically, the history is often less important in violent deaths where the wounds are self-evident, though a good description of the circumstances may assist in interpreting the direction of injuries, the nature of the weapon and other aspects.

In most autopsies the history directs the pathologist to the appropriate ancillary investigations, as should have happened in the case described above. Some pathologists have in the past advocated that the autopsy should be performed 'blind', so that the history does not prejudice the opinion of the pathologist. This is patently impracticable, for every autopsy would then have to be totally comprehensive, including such techniques as the removal of the spinal cord in every case, and all possible ancillary investigations, such as toxicology, microbiology, virology, radiology, diatoms, histology and so on, as there would be no means of knowing what was necessary and what was irrelevant. Apart from the intolerable burden of work, the expense of such an approach would be prohibitive, if applied to all autopsies.

A particular difficulty often arises in relation to bodies returned to a home country after death in a foreign state.

The problems vary greatly with the degree of both pathological and administrative sophistication of the foreign country. Where these are poor, then the problems of obtaining a good – or even any – history may be great, even after attempts at using international police or diplomatic channels.

Sometimes, a 'doctor-to-doctor' approach by telephone or fax may be the best way of obtaining information. Even a cause of death from the foreign country may be either unobtainable or so vague (such as 'heart failure') as to be useless. Previous autopsies may be incomplete, of poor quality, or even fraudulent, such as finding a stitched skin incision but no underlying dissection of organs. Several surveys of such problem autopsies have been published, with a plea for some kind of International Death Certificate (Green 1982; Leadbeatter 1991).

Difficulties arise when objective autopsy findings are scanty or even absent. The pathologist then has the dilemma of choosing between a cause of death based on a subjective knowledge of the alleged history (which may be wrong) or admitting that the cause of death is unascertained.

This problem, fully discussed by Leadbeatter and Knight (1987), is by no means uncommon. For example, sufferers from epilepsy or asthma are known to die suddenly and unexpectedly, an autopsy usually revealing no adequate morphological cause for their death. Another prime example is the sudden infant death syndrome, discussed in Chapter 21, where by definition there are no significant autopsy lesions and the history is usually essentially barren.

In hypothermia where the patient is warmed up in hospital but dies in a day or two, there may be nothing to find at autopsy, yet there is firm clinical history of a rectal temperature of 26°C. Moving to more forensic cases, a body may be recovered from a river after clothing and a suicide note have been found on the bank – yet at autopsy there may be no visible evidence of drowning whatsoever.

In these cases, should the pathologist – in the absence of objective evidence of hypothermia or drowning – decline to offer a cause of death and perhaps be accused of being obstructive or perverse? Or should he certify 'hypothermia' or 'drowning', and perhaps be challenged in court by a lawyer who claims that the pathologist relied only on hearsay and has no objective evidence to offer?

The best course is for the pathologist, if he feels that the history is sufficiently strong, to give the most likely and reasonable cause of death, but make it clear in his discussion in the autopsy report that this opinion is based on a consideration of the circumstances and is not a dogmatic statement of the kind that could be offered if the victim was shot through the brain. Where a summary of the history prefaces the autopsy report, the pathologist should be careful to indicate what is 'hearsay' as far as he is concerned. Thus he should not say 'The deceased was struck on the head by two men', but 'I was told that the deceased was struck on the head by two men'.

As is discussed later, the autopsy report should not be merely a bald recital of the anatomical findings, but should have a final commentary that includes a 'differential diagnosis' where the cause of death is not clear cut, as well as a justification for the eventual opinion or an admission that there is no way of deciding between alternative possibilities.

The main difficulty arises with the actual 'cause of death' phraseology, which will be entered into the legal and national records in the format required by the World Health Organization. This gives no opportunity to express preferences or give explanations but, on the 'best-guess' principle, the pathologist can do no more than enter his most reasoned choice or, with legitimate honesty, state that the cause is 'undetermined'.

PRECAUTIONS REGARDING POTENTIAL INFECTIVE CONDITIONS

Many forensic situations involve drug abusers and persons with promiscuous sexual behaviour, where the statistical risk of HIV and hepatitis infection is markedly greater than in the general autopsy population. This poses a risk to pathologists, mortuary staff, police and laboratory staff who may deal with post-autopsy samples. A great deal has been written about such risks but no consensus has yet been reached, though an excellent survey has been published by the Royal College of Pathologists (1995).

One school maintains that all autopsies should be carried out with total precautions against infective risks, so that it does not matter what case is handled. However, this is almost impossible to achieve in a busy coroner or medical examiner practice and does not solve the problem of possibly infected material being sent out to other laboratories.

A more common regime is to carry out pre-autopsy testing for HIV and hepatitis, using blood from a femoral needle puncture. The result can often be returned within hours, when a decision may be made as how to handle the autopsy – or even whether to abandon it, if the risk–benefit aspect is high. More usually, a positive result will result in the autopsy being carried out with special care, additional protective clothing, visors, masks and metal gloves, with restriction of access to observers, choice of more senior technicians and warnings sent to laboratories liable to handle samples.

In this respect, hepatitis is more of a risk than HIV infection. However, so far, about 100 health-care workers have acquired HIV infection from definite occupational exposure (Sidwell *et al.* 1999) and one pathologist is known to have become infected from autopsy work (Johnson *et al.* 1997). Also three morticians in the USA have possibly contracted occupation-related HIV positivity.

The time for which a corpse remains potentially contagious with HIV is variable. Infectious virus has been recovered from liquid blood held at room temperature for 2 months and virus in high concentrations has been found to remain viable for 3 weeks (Cao *et al.* 1993). Bankowski *et al.* (1992) found 51 per cent survival of virus in plasma and monocyte fractions from infected cadavers up to 21 hours post-mortem. Other series found survival in corpses from 18 hours to 11 days after death. Virus has been recovered from the spleen after 14 days post-mortem. Refrigeration seems to make little difference to viability. Douceron *et al.* (1993) cultured blood and effusions from refrigerated bodies and obtained viable virus up to 16 days post-mortem and concluded that there was no safe maximum time at which corpses ceased to be an infective risk. In industrialized countries, it has become standard to offer post-exposure prophylaxis after significant percutaneous exposure to blood or tissues of HIV patients.

Other infections, such as tuberculosis, hepatitis viruses, anthrax, plague, Creutzfeldt–Jakob, Marburg, Green Monkey diseases, etc., are usually the subject of specific health and safety regulations, which vary from country to country.

THE AUTOPSY: EXTERNAL EXAMINATION

In contrast to the 'clinical autopsy' performed to evaluate natural disease, the importance of the external examination is far greater in the forensic case, especially in deaths from trauma. In the latter, the medico-legal value of the external description may be paramount, as it is often from the outer evidence that inferences may be made about the nature of the weapon, the direction of attack and other vital aspects. Thus the forensic pathologist must spend all the time that is necessary in a careful evaluation of the body surface and not be too impatient to wield the knife mainly to seek material for histology, which is more justified in the purely clinical autopsy.

The routine for external examination will naturally vary according to the nature of the case but certain general principles apply. The following procedure is a useful baseline and may be adjusted according to personal preference. Variations in criminal cases to accommodate the needs of the investigating and scientific teams are mentioned later.

- After identification and removal of any clothing, the race and gender are noted. The apparent age is assessed in children by size and in adults by changes in skin and eyes, such as the loss of skin elasticity, senile hyperkeratosis, Campbell de Morgan spots, senile purpura and arcus senilis. Hair colour, tooth loss and arthritic changes are also obvious signs of ageing. The apparent age should be compared with the alleged age and enquiries made about any obvious discrepancy, in case it is the wrong body, an error which plagues most autopsy rooms from time to time.
- The body length is measured from heel to crown (in infants, more detailed measurements are described later). Ensure that the attendant does not take the 'undertaker's height' from toe to crown, as due to the plantar flexion of rigor, this can be a considerable number of centimetres more than the live standing height.

  It should also be appreciated that the post-mortem height may differ from the known living height by several centimetres. There are several opposing causes of variation, which do not necessarily cancel each other.

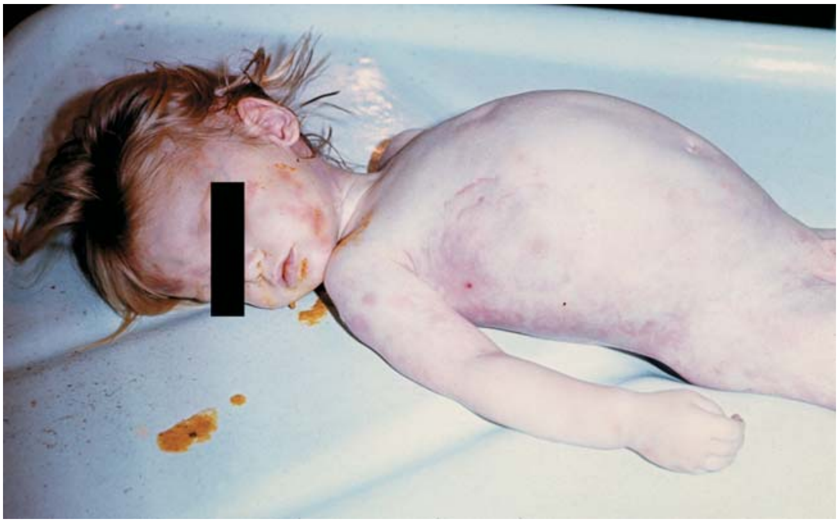

FIGURE 1.4 *External examination must note every feature; here there is obvious abdominal distension in a battered child with facial bruising. The intestine had been ruptured by a blow, and oxygen administered by the ambulance crew escaped to distend the peritoneal cavity.*

For example, muscle flaccidity allows joints to relax, unless rigor is present, but intervertebral discs appear to shrink, allowing shortening.

- The body weight in kilograms is measured if facilities are available; if not, it should be estimated. The weight of infants must always be measured. The general nutrition and physique is assessed in terms of obesity, leanness, dehydration, oedema, emaciation, and so on.
- The state of cleanliness, personal hygiene, hair and beard length, toenail and fingernail state, and urinary and faecal soiling is noted. Any parasitic infestation, such as fleas or lice, is combated before proceeding.
- The general skin colour is noted, especially hypostasis (discussed at length in Chapter 2). Congestion or cyanosis of the face, hands and feet is sought. Localized discoloration, especially unilateral in a limb, suggests arterial embolism or incipient gangrene. Pink or brownish pink patches over the large joints may indicate hypothermia (Chapter 17). Other abnormal colours include the brownish hue of methaemoglobinaemia in some poisonings, the bronze speckling of clostridial septicaemia and the dark red of cyanide that somewhat resembles the cherry-pink coloration of carboxyhaemoglobin. Naturally racial pigmentation will modify the ease with which abnormal skin coloration can be seen.
- Congenital deformities of any type are recorded, from talipes equinovarus to spina bifida, from a naevus to extra toes.
- Acquired external marks may be important for identification purposes or in relation to past injuries and disease. Tattoos, circumcision, amputations, surgical scars, old fracture deformities and scars of injuries, burns or suicidal attempts on the wrist and throat are noted. These are discussed further in Chapter 3. Increasingly, artefacts – both external and internal – arise from resuscitation attempts and must be carefully distinguished from original trauma. This emphasizes the importance of the history, to determine whether cardiopulmonary resuscitation was attempted by trained or untrained persons.
- The hands should be carefully examined for such signs as old and new injuries, defence wounds, bruised knuckles and electrical marks. The latter are often insignificant and difficult to see unless the rigor of flexed fingers is overcome, either by force or even cutting the flexor tendons at the wrist.
- Vomit, froth or blood may be present at the mouth and nostrils, and faeces and urine may have been voided. This must be correlated with the degree of post-mortem decomposition, which often leads to purging of fluids from orifices; most forensic pathologists have had the experience of being called by the police to the scene of ‘a fatal haemorrhage’, to discover only bloody fluid being purged by gases from a decomposing corpse.

Vaginal discharge or bleeding is noted and the ears examined for leakage of blood or cerebrospinal fluid. Post-mortem ejaculation of semen from the external

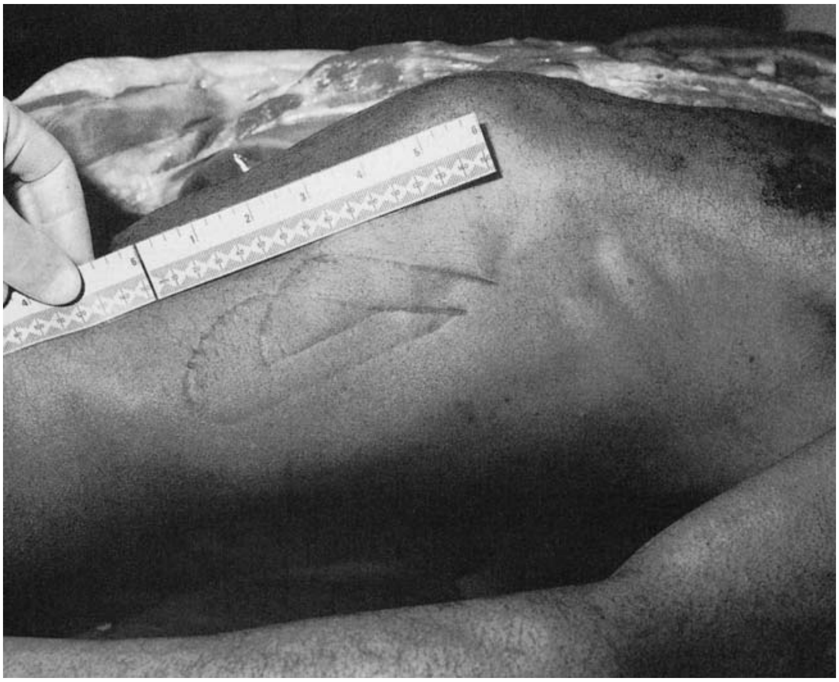

FIGURE 1.5 *Post-mortem artefacts on the body of a homicide victim. These are due to defibrillator paddles applied in a hospital casualty department.*

meatus is of no significance and can be seen in any type of death. It was shown by Mant (1984) not to be associated with sexual activity immediately before death. It is not particularly associated with asphyxial deaths, as sometimes stated.

- The degree of rigor mortis is assessed by flexing the arms and legs to test the resistance. The significance of this and hypostasis is discussed in Chapter 2.
- Recent injuries (other than scars already noted) are carefully examined. In fact this part of the autopsy is often the most significant element of forensic cases. Injuries may be conveniently recorded in the mortuary by marking them on printed body diagrams. The forms can be secured on a clipboard and the data later transcribed into a written description for the autopsy report. The original diagrams should be retained, as their production may be demanded in court if there is controversy about any particular injury.

All traumatic lesions should be clearly differentiated into abrasions, bruises, lacerations, incised wounds, burns and so on, according to the definitions given in Chapter 4. The shape and condition of the margins of each injury should be described where appropriate. It must be meticulously measured in terms of length, breadth, orientation to the axis of the body and the position with reference to surface anatomical landmarks. For example, a knife wound of the thorax might be described as follows, using ordinary language to explain any medical terms:

> A stab wound was present on the left upper chest, placed obliquely, with the inner end lower than the upper outer end. The wound was 20 mm in length, which extended to 22 mm when the edges were opposed. The maximum width at the centre was 4 mm. The centre of the wound was just below the line joining the nipples, being 6 cm from the midline, 7 cm from the left nipple and 18 cm below the centre of the left clavicle (collar bone). The wound was 132 cm above heel level and was elliptical in shape, with the inner lower end slightly more sharply cut than the rather blunt upper end. The wound was shelved in a downwards direction, with subcutaneous tissue visible along the inside of the upper edge.

This description, along with police photographs, will convey an excellent impression of the wound at any future date, including at the trial, which may be many months later. The position of the injury is related to obvious anatomical points and the height above the ground may have relevance if the stature of the victim and assailant, and the angle of attack, become legal issues. On the scalp, the occiput and the tip of the ear can be used as reference points, together with the vertex and the centre line of the head. On the face, the obvious landmarks are the eyebrows, nasion, tip of nose, lips, point of chin and angle of jaw. Where large areas of abrasion or bruising are present, similar measurements of lesion size are needed, but more general descriptions of position may be sufficient, such as ‘covering the external aspect of the left thigh’ or ‘extensive bruising 23 cm by 18 cm on the right side of the chest extending from axilla to costal margin’.

Where widespread burns are present, an estimate of the total area using the ‘Rule of Nine’ should be made, as described in Chapter 11.

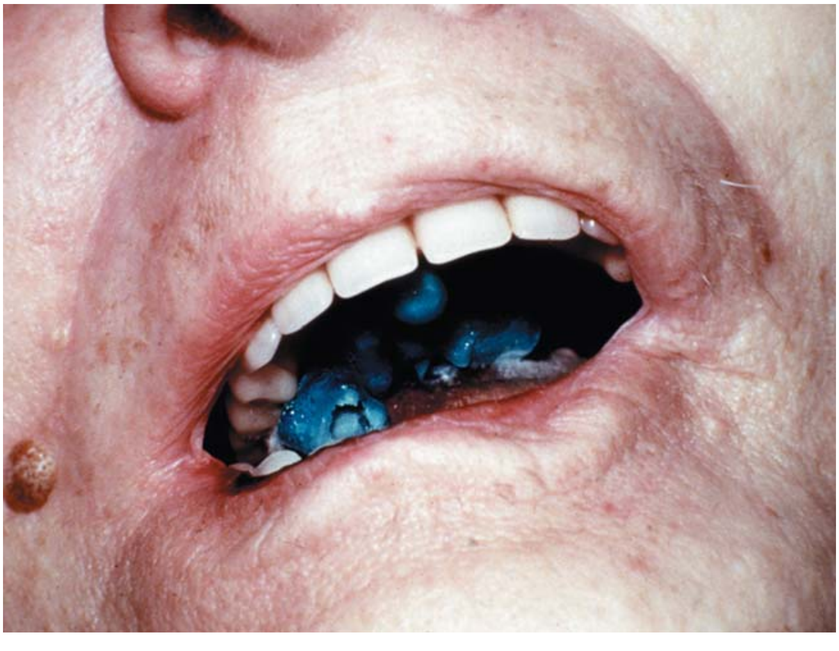

FIGURE 1.6 *Full external examination is vital, including the interior of the mouth. Here a large number of amylobarbitone capsules offer a presumptive cause of death, though analytical confirmation is still required.*

Other injuries, such as firearm wounds, are described in the same manner, being careful not to clean off any powder residues or other trace evidence before the forensic scientists have collected their samples.

With head injuries, the scalp is examined in its original condition first and any trace evidence collected. Then any clotted blood that frequently obscures the injuries can be gently removed, using a sponge and water. After this stage has been studied, it is usually necessary to shave off hair carefully around the wounds, so that the full extent of the lacerations and especially the state of their margins can be assessed and photographed. This shaving is best carried out with a scalpel fitted with a new blade, the blade being kept almost parallel with the surface to avoid making false cuts. The procedure should be carried out by the pathologist himself and not delegated to a mortuary technician. In this way any artefactual cuts can be recognized as such and not misinterpreted as part of the original wound.

- The eyes must be examined carefully, especially to detect petechial haemorrhages on the outside of the eyelids, conjunctivae and sclera. Though these do not necessarily indicate an asphyxial process, such petechiae require an explanation. Petechiae should also be sought behind the ears and in the skin of the face, especially around the mouth, chin and forehead. Care must be taken to differentiate these fine petechiae from the more coarse intradermal blood spots often seen across the shoulders or upper chest, which are caused by post-mortem hypostasis in a body with marked agonal venous congestion – especially when that surface of the body has been dependent. Haemorrhages in postural hypostasis are of no diagnostic significance, as true ante-mortem bleedings can never be differentiated from those that commonly develop after death.

The size of the pupils is rarely useful, as rigor in the iris musculature can produce any degree of constriction, which may be unequal on either side (see Chapter 2). False eyes, contact lenses, lens opacity and other defects are to be noted. It is not possible to evaluate the degree of visual acuity during life from any type of post-mortem examination – though obviously a

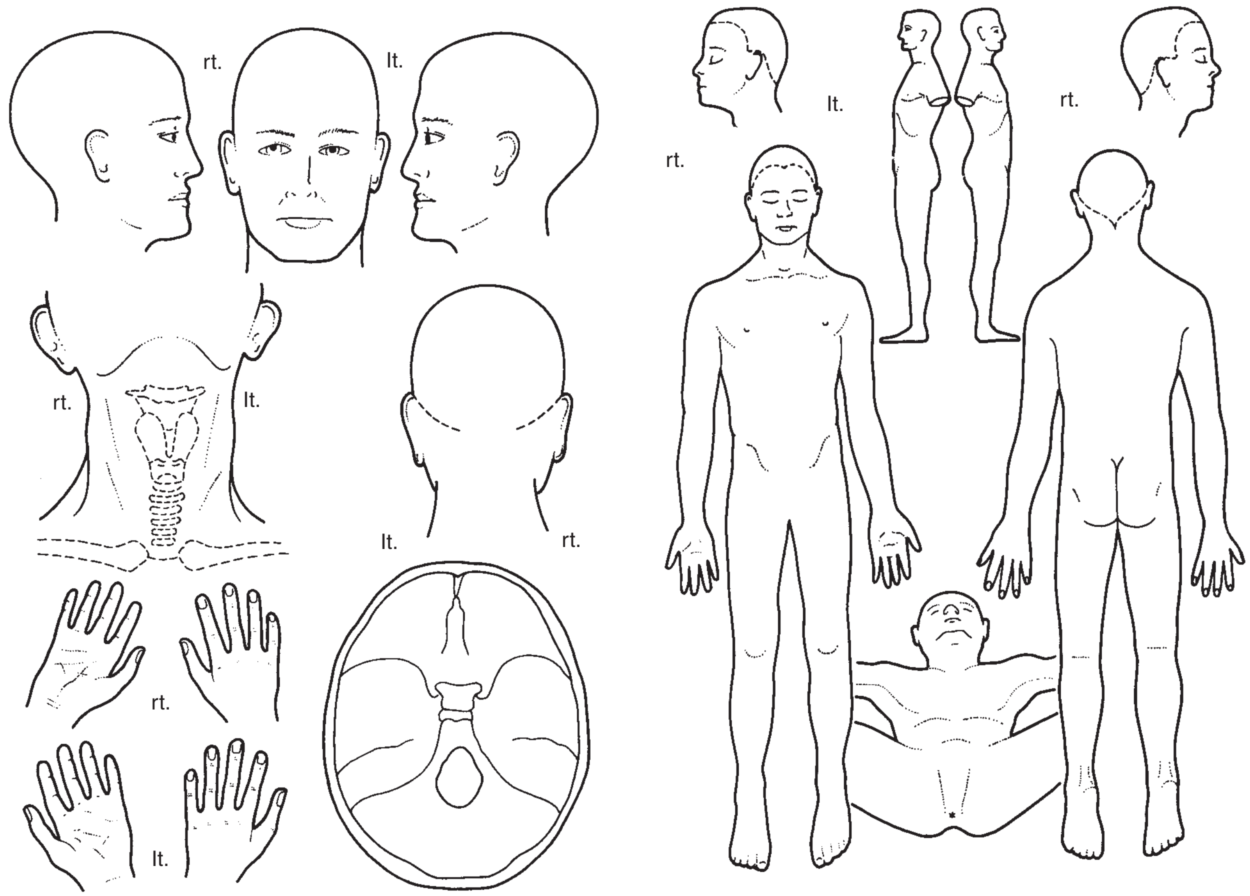

(a) (b)

FIGURE 1.7 *(a, b) Typical examples of the many anatomical diagrams available for the recording of external findings at autopsy.*

significant defect such as a cataract or a large vitreous haemorrhage is likely to affect vision.

- The mouth may reveal foreign bodies, drugs, damaged teeth, injured gums and lips (especially the ruptured frenulum of child abuse), and the bitten tongue of epilepsy or blows on the jaw. Dentures should be identified and removed before autopsy. Gastric contents in the mouth need not indicate ante-mortem regurgitation, as discussed in Chapter 14, but should be noted. Dried powder on the lips may suggest the recent taking of medicaments or poisons; corrosion of the mouth, lips and chin may be seen in irritant poisons. Bleeding from mouth, nostrils, or ears must be recorded, and later investigated as to source from internal examination. Frothy fluid, sometimes blood-tinged, may be seen issuing from the mouth or nostrils, or both, in drowning and from pulmonary oedema due to a variety of causes. Froth is sometimes pink or frankly blood-tinged. This has no particular significance unless gross, as in drowning, sudden infant death and other conditions with marked pulmonary oedema; rupture of small pulmonary or even pharyngeal vessels can add a little blood to colour the foam.

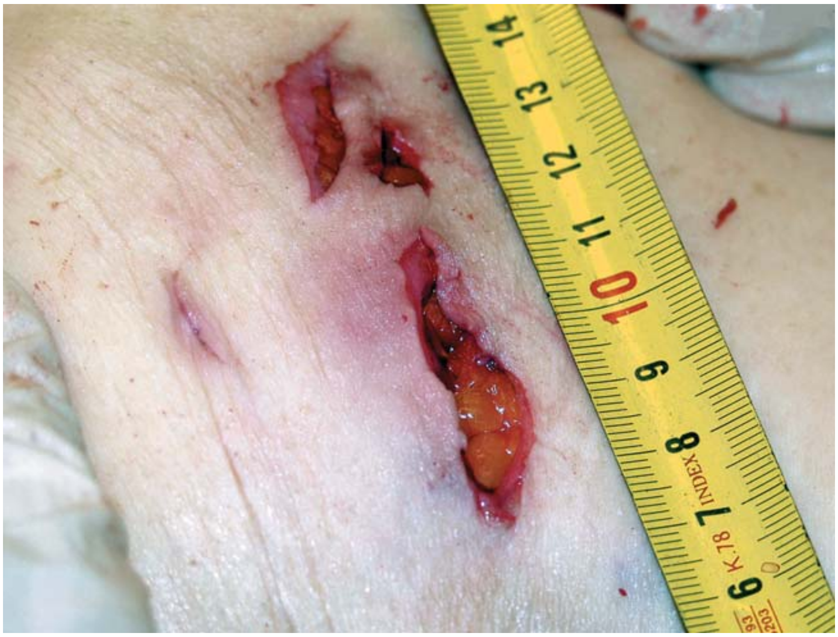

FIGURE 1.8 *Examination and photography of injuries; after full preliminary examination and forensic sampling, the area is cleaned, shaved if necessary, and photographed with a scale adjacent to the injuries.*

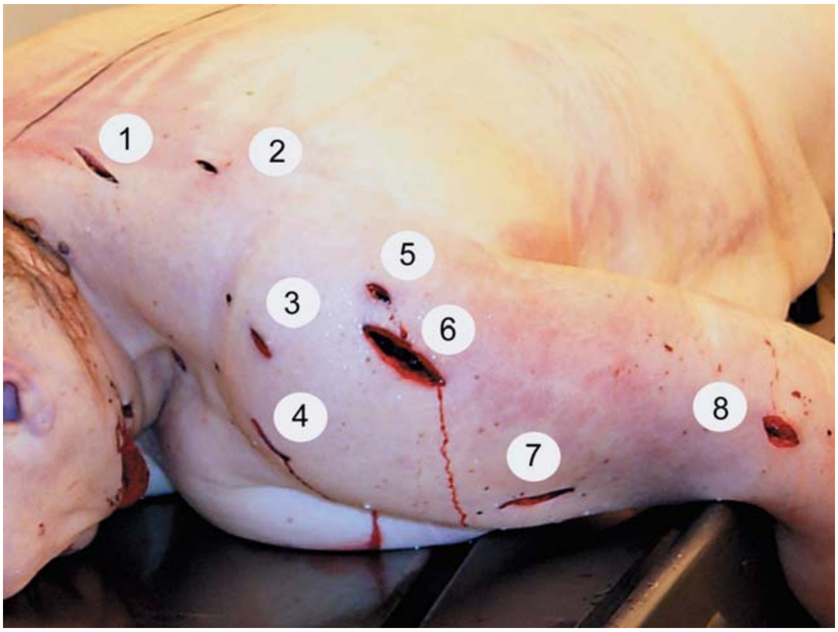

FIGURE 1.9 *Where there are numerous similar lesions, it assists the clarity of the autopsy report if numbers used during photography match those listed in the autopsy protocol.*

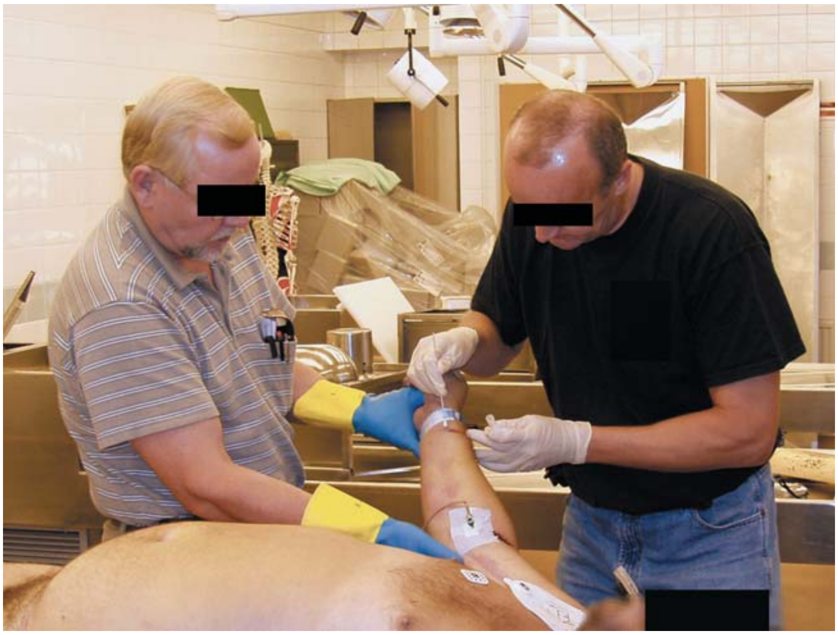

FIGURE 1.10 *Scene of crime officers recovering trace evidence from hands before a homicide autopsy. The pathologist must offer the rest of the team every facility to complete their work before he begins his dissection.*

- The external genitals require careful examination, as does the anus. The state of the latter can be misleading, as a widely open, patulous anus is often seen post-mortem, due to flaccidity of the sphincter. The inner mucosa is often visible through the orifice. This is also the case in infants and children, and a diagnosis of sexual abuse must not be assumed without other corroborative evidence such as fresh mucosal tears or swabs positive for semen.

  The 'funnel-shaped' anus, beloved of traditional textbooks as a sign of chronic homosexual activity, is such an extreme rarity that its true existence is doubtful. A funnel shape is a normal anatomical variant, the anus being set deeply between the buttocks, sometimes with a commissure of skin above it, making the approach seem even deeper. The alleged 'shiny, silvery hyperkeratinized skin of the habitual pederast' is also of little diagnostic value as scratching from chronic irritation, often associated with haemorrhoids, viral infections or threadworms, may lead to the same appearance. The only reliable criteria are fresh tears and old scars, as well as mucocutaneous eversion, though even these can arise from severe chronic constipation (see also Chapter 18).

  Examination of the vulva and vagina is made to exclude obvious injury and disease, unless the nature of the case suggests some sexual interference, when a far more detailed examination and a special autopsy technique is carried out (Chapter 18). Routine examination of the male genitals usually need only extend to general inspection of the penis, glans and scrotum, with palpation of the testes. Circumcision should be noted, as (rarely) it may assist in identification.

THE AUTOPSY: INTERNAL EXAMINATION

The dissection in a forensic case is basically similar to any other autopsy, with variations according to the nature of the death and the needs of the particular investigation, whether it be a criminal case, a civil dispute or an accident investigation.

There are a number of manuals devoted to the performance of an autopsy (see Ludwig 2002 or Knight 1996, for example) as well as one for the instruction of mortuary technicians (Knight 1984). Only an outline of the technique is offered here for those unaccustomed to examining the dead who may have autopsies thrust upon them by circumstances. In addition, special forensic procedures are described, though these are also discussed in each of the chapters devoted to the various types of injury and death. The autopsy on infants is dealt with in Chapters 20–22.

The usual incision is an almost straight line from laryngeal prominence to pubis, deviating to avoid the umbilicus. The upper end of this incision should not be prolonged above the larynx, as even a high-necked shroud will then fail to hide the subsequent suture line from relatives.

Another common method is to cut from behind each ear to a point above the manubrium and continue downwards in a ‘Y-shape’. This is often done in infants and wherever it is desired to avoid disfiguring the front of the neck. In the USA in particular, the Y-incision is favoured or even a deep ‘U-shape’ carried across the upper chest.

In strangulation, hanging, and any condition where the larynx might be damaged, the Y-incision is to be preferred, as the skin of the upper neck can then be dissected off the mandible and raised clear to give a wide approach to the neck structures. Here, the incision should not be made until the skull-cap and brain have been removed, to avoid the congestive artefactual haemorrhages in the neck structures, described by Prinsloo and Gordon (1951), which can be confused with true ante-mortem trauma. Gordon *et al.* (1988) have further suggested that the brain be removed first in all autopsies so that abnormal odours can be detected before being swamped by the smell from the opened abdominal cavity.

The Y-incision is also required when dissection of the face is necessary, to look for deep bruising or bony damage. In fact the whole face can be dissected off completely from the skull and replaced with little effect upon the cosmetic result, if careful dissection is employed. The Y-incision in the neck is made continuous with the transverse scalp incision, joining these behind each ear, so that the anterior neck and facial skin can be removed in entirety.

Returning to the autopsy in general, the primary incision should be shallow over the neck to avoid cutting underlying structures, especially the trachea. The thorax can be cut down to the sternum, but care must be taken over the abdomen, where a light cut is made, sufficient to incise only skin and fat. A small puncture should then be made in the peritoneum and a finger inserted to lift it away from the intestines. The knife is then used to cut outwards along the length of the abdomen, to avoid penetrating the intestines.

The incision for access to the skull is made from behind each ear, meeting over the crown of the head. It is well to keep this posterior to the actual vertex, again to make the stitching less obvious, especially where the hair is scanty or absent. With abundant hair – and especially in children – the hair should be wetted and then combed backwards and forwards from a transverse parting made in the line of the prospective incision. Hair is not then severed and can be combed back later to cover the stitching.

Exposing the body cavities

The skin, subcutaneous tissues and fat are flayed off laterally from the main incision, taking care not to let the edge or point of the knife come through the skin, especially in the neck area, where repair can be unsightly. The tissues are taken back to the lateral edge of the neck and to the outer third of the clavicles. Over the thorax, the tissues, including pectoral muscles, are flayed off to the midaxillary line in the upper part and even further posteriorly towards the costal margin.

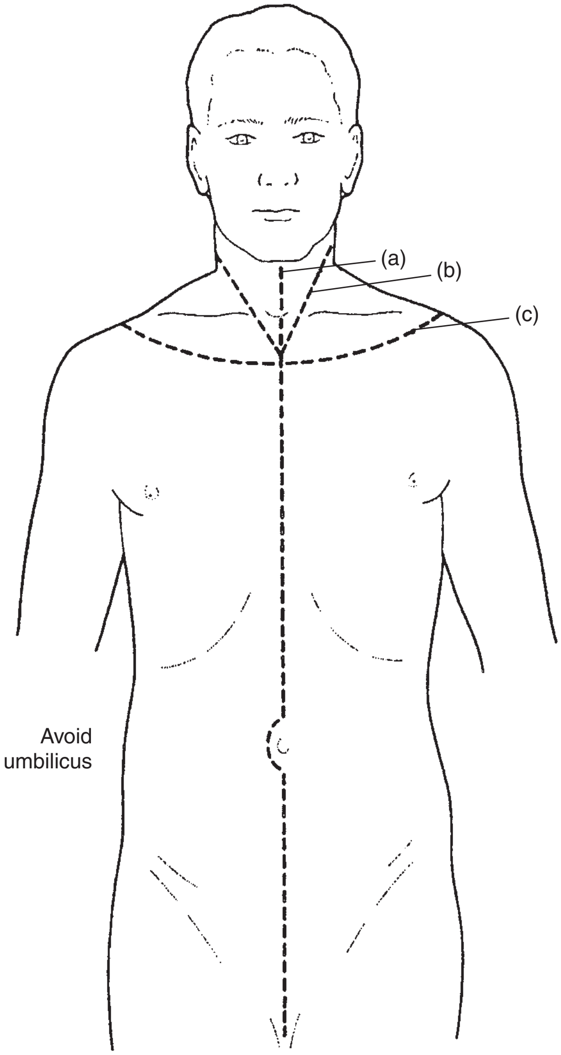

FIGURE 1.11 *Autopsy incisions: (a) standard midline, (b) ‘V-shape’ and (c) subclavicular.*

The anterior abdominal wall is similarly separated. This can be done either in two stages, first stripping back the skin and fat to expose the muscles – or the full thickness of skin, fat and muscle can be reflected together. The muscle must be cut from the costal margin and, if thick fat overlies, then some transverse relieving incisions can be made on the peritoneal surface of the lower abdominal wall, taking care not to come through to the skin.

Opening the thorax

The thorax is opened by first disarticulating both sternoclavicular joints. This is carried out by moving the shoulder tip with one hand, to identify the joint capsules. The point of a knife is then introduced vertically and cut laterally in a half-circle to separate the joints. If they are ankylosed, which often occurs in old age, then the clavicles can be cut through at the end of the next operation. This consists of the severing of the ribs and can be performed either with a handsaw or a rib shears. In children and many adults, the costal cartilages may be cut through with a knife, though this provides a rather narrow exposure of the chest contents. In infants, the soft cartilage can easily be divided with a scalpel: in older bodies, a stout knife should be kept for the purpose to save blunting the knife that is needed for organ dissection. Often the first rib has to be sawn through, even though the remainder are cut with a knife.

When a saw is used, the ribs are cut through lateral to the costochondral junctions from a point on the costal margin to the sternoclavicular joints or nearby. If a saw is used, it must be kept at a low angle to avoid the tip lacerating the underlying lung, especially if there are pleural adhesions. When the sternum and medial rib segments are free, the section is lifted and dissected away from the mediastinum, keeping the knife close to the bone to avoid cutting the pericardium. The sternal plate is examined for fractures or other lesions before it is put aside: damage caused by the considerable trauma of resuscitatory cardiac massage is frequently found at this point.

The whole of the thorax and abdomen is now open for inspection. The degree of inflation of the lungs should be assessed, noting complete or partial collapse, emphysema, overdistension and any asymmetry of inflation.

If a pneumothorax has been suspected beforehand, a post-mortem radiograph is the best confirmation. Alternatively, the chest wall can be punctured in the mid-axillary line after filling the reflected skin with water to observe if bubbles escape. This test is rarely successful and cannot succeed if there is a patent communication between the pleural cavity and the bronchial tree. If there is a marked tension pneumothorax, the hiss of escaping air may be heard when the tip of the knife penetrates the intercostal muscles and parietal pleura. The pleural cavities are inspected for adhesions, effusions, pus, blood, fibrin and even gastric contents.

Examining the abdomen

The abdomen is then inspected, though ascites, faeculent fluid, pus and blood may already have escaped on first opening the peritoneal cavity.

The omentum may show inflammation or fat necrosis. When moved aside, the loops of bowel are inspected for any abnormality, especially infarction, peritonitis and the distension of ileus. Beware of mistaking post-mortem hypostasis for the necrosis of mesenteric embolism or strangulated bowel. Though the dark colour may be similar, hypostasis usually has irregular segments when the gut is stretched out, whereas infarction occupies one continuous section – and if well established, the bowel wall will be lustreless and friable.

The bowel is gently moved aside to look at the posterior part of the abdomen: a retroperitoneal haemorrhage from a ruptured aorta may be visible, or an aneurysm itself.

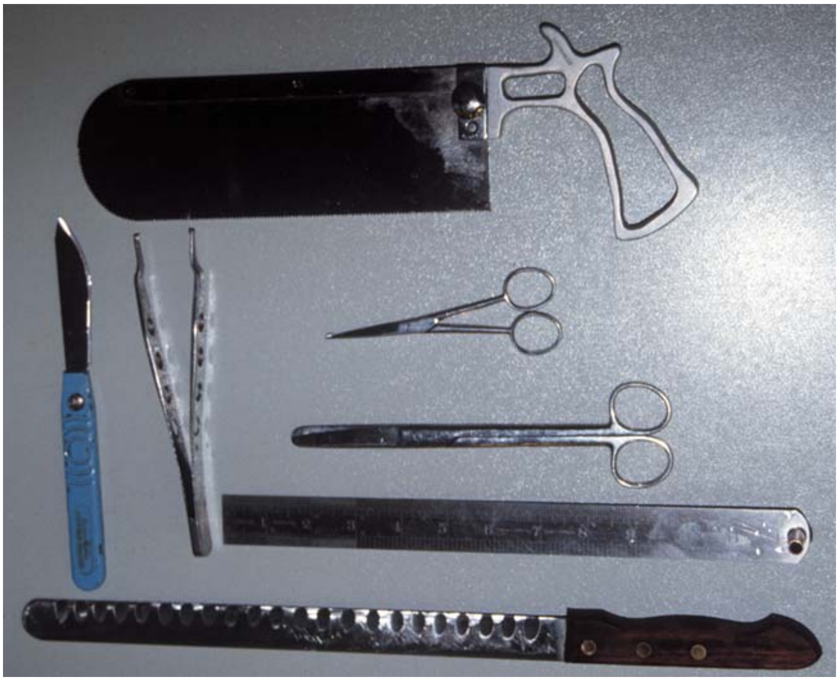

FIGURE 1.12 *The minimum instruments needed for performing an autopsy. Any further items may be useful but not essential.*

The collection of body fluid samples

In a large proportion of forensic autopsies and virtually every one with criminal connotations, samples of blood and other body fluids and tissues are needed for laboratory examination. In relation to sexual offences, the taking of swabs and other samples is described in Chapter 18. Where blood and other fluids are required for toxicological, biochemical, microbiological and serological investigations, they are usually collected at the early stages of the autopsy (see Chapter 27).

The site of collection depends on the nature of the test. When samples for toxicological analysis are required, considerable care should be employed in sampling; this is more fully discussed in Chapter 27.

It is not advisable to use visceral blood for sampling, especially for small molecule substances that can easily diffuse after death; these include alcohol and many pharmaceutical products. Though older textbooks advocate the use of heart blood as a convenient source of samples, it may be contaminated by post-mortem diffusion from the stomach and intestine. After death, the cellular barrier of mucosa and serous membranes breaks down, and substances in the stomach, intestine and air passages can migrate to other organs in the main thoracoabdominal cavity, causing a false rise of the true ante-mortem blood level.

Pounder (1985; see Chapter 27) found that a slurry containing alcohol and paracetamol, placed in the trachea after death to resemble gastric contents, gave rise to appreciable concentrations of those substances in blood samples taken from thoracic vessels, whereas femoral blood remained uncontaminated.

For substances such as carboxyhaemoglobin this does not matter, but alcohol is the prime example of a potential source of error if visceral blood is collected. The better choices for collection sites are:

- by needle and syringe puncture of the femoral vein before the autopsy dissection begins. This requires practice but, in adults, 20 ml can usually be aspirated without trouble. This is the method of choice if only an external examination is possible.
- from the subclavian or external iliac veins after the body has been eviscerated. By holding a small container under the cut end of a subclavian vein and raising the arm, blood can be collected directly. If the flow is slow, the arm can be massaged towards the shoulder. Similarly – and preferably if volume is required – by cutting across the iliac veins at the brim of the pelvis, a container can be dropped into the pelvis with its mouth under the vein and blood massaged into it by firm pressure moved up the inner thigh. As all these containers will be soiled externally, the blood should be decanted into a fresh tube for transmission to the laboratory.
- when the skin is dissected off the neck, the internal jugular vein is exposed, especially if a sternomastoid muscle is divided and pulled aside. When cut, a copious flow of blood is usually obtained, which can be collected directly into a container. The only disadvantage of this method is that if the blood wells

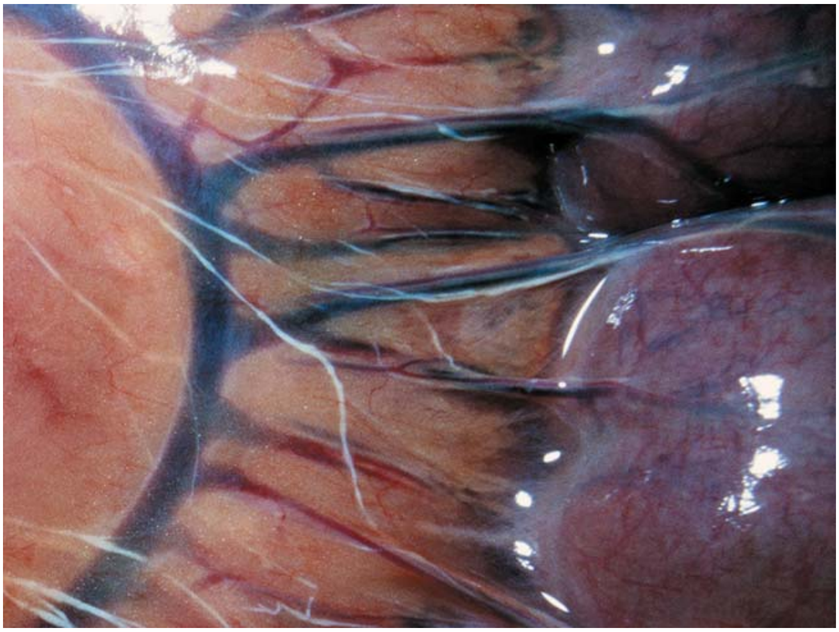

FIGURE 1.13 *White shining fat droplets (chylomicrons) in mesenterial lymphatic vessels suggesting postprandial state.*

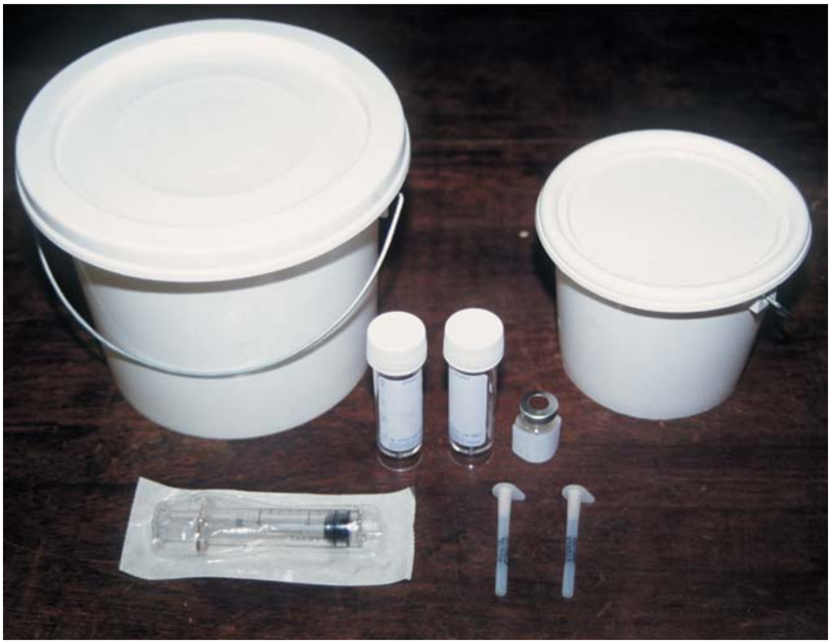

FIGURE 1.14 *Toxicology kit issued by the home office forensic science laboratories. Contained in a plastic bag are a large liver tub, a smaller stomach tub, two universal (30 ml) containers, one rubber-capped fluoride vial for blood alcohol, and a syringe and two needles.*

> up from the thoracic inlet via the superior vena cava, heart blood is likely to be admixed, with the possible errors mentioned earlier. If the blood is collected from the upper segment of the jugular, then blood from the head is collected: this flow can usually be stimulated by raising and lowering the head during collection.

Samples for serology, microbiology and for analysis for substances such as carboxyhaemoglobin, which are not absorbed from the gastrointestinal tract, can be collected from any blood vessel, but blood should never be scooped up from the general body cavity after evisceration, as this can be contaminated with any leakage from other structures, such as gastric or bowel contents, mucus, urine, pus or serous fluids.

Blood for microbiological culture has traditionally been taken from the heart but there is no particular merit in this as opposed to peripheral blood. If an infective endocarditis is suspected, it is best to open the heart later with a sterile scalpel and excise the mitral or aortic valve cusps or vegetations for direct culture. Otherwise, blood for culture for a suspected septicaemia is taken from a peripheral vein.

Urine can be collected by catheter before autopsy or even by suprapubic puncture with a syringe and long needle. However, it is usually obtained after the abdomen is opened, but before the organs are removed. If the bladder is full, the fundus is penetrated, and urine collected either by syringe or directly into a container. If almost empty and contracted, the fundus is gripped and pulled upwards so that it stretches, then is incised and the contents removed by syringe. Care should be taken not to contaminate the urine with blood.

The removal of vitreous humour and cerebrospinal fluid may be required for toxicology or for attempts at estimating the time since death by the potassium content (see Chapter 2). Vitreous humour must be aspirated with care if any reliable results are to be attained. A fine hypodermic needle attached to a 5 ml syringe is inserted into the outer canthus of the globe after pulling the eyelid aside. When released, the lid will cover the small puncture mark, hiding any sign of interference. The needle should be entered into the centre of the globe to avoid aspirating material near the retina, which has a markedly different chemical composition due to shreds of detached retina entering the aspirate. The fluid should be sucked off slowly and gently, and as much taken as possible to obtain a mixed fluid: both eyes should be used, as they often differ somewhat in their chemical composition. After withdrawing vitreous, the globes can be re-inflated with water, to improve the cosmetic appearance of the eyes.

Cerebrospinal fluid may be obtained in the same way as in living patients, by passing a needle into the theca between the lumbar spines. A baby can be held upright by an assistant in a fully flexed position; an adult must be pulled into flexion whilst lying sideways on the autopsy table. An alternative technique is to perform a cisternal puncture through the atlanto-occipital membrane.

As there is no pressure within the theca in a cadaver, the fluid must be actively aspirated: sometimes all attempts at obtaining fluid by external puncture fail. The only course then is to try to puncture the ventricles through the exposed brain surface when the skull is opened. Attempting to obtain clear cerebrospinal fluid from the interior of the skull after removal of the brain is generally useless: though blood-stained fluid can be centrifuged to clarity, its chemical composition is then unreliable.

Removal of the viscera

After the body cavities have been inspected, the organs are removed *en bloc* by a modified Rokitansky procedure, more accurately described as Letulle's method. First, the intestine is removed, as follows.

The omentum is lifted upwards to expose the coils of small intestine. The uppermost part of the jejunum is identified, where it passes retroperitoneally to join the termination of the duodenum. Here the mesentery is perforated with the knife and the gut cut through. If it is essential to retain the duodenal/gastric contents or the small intestinal contents, two string sutures may be passed through the hole and tied before cutting the gut between them, but little contents are lost if this is not done. The intestine is stripped out by cutting along the mesentery near the attachment with the bowel until the ileocaecal valve is reached.

The caecum is then mobilized medially using manual traction with minimum use of the knife, to avoid puncturing the lumen. When the hepatic flexure is reached, the omentum is pulled downwards to draw the transverse colon tense against the mesocolon, which is cut through, taking care not to open the adjacent stomach. The splenic flexure is then pulled medially and downwards, and the descending and sigmoid colon separated from the posterior abdominal wall. The upper rectum is cut across, though some pathologists merely lay the gut outside the body, leaving the sigmoid and rectum attached.

Removing the neck structures

To make the removal of the neck structures easier, a block 10–15 cm high should be placed under the shoulders of the cadaver. This allows the head to fall back and thus extends the neck. This should be done gently, as with all handling of the body, to avoid the well-known 'undertaker's fracture', which is a subluxation of the lower cervical spine due to tearing of the intervertebral disc at about C6–C7. This can be misinterpreted as an ante-mortem injury, especially if the other common artefact of haemorrhage over the anterior longitudinal ligament of the cervical spine is present (Prinsloo and Gordon 1951). The neck structures are then freed by passing a knife under the skin of the upper neck until it enters the floor of the mouth. The knife is then run around the inside of the mandible to free the tongue. The tissues at the back and sides of the pharynx are divided, and the tonsillar area cut through. Fingers are then passed up behind the mandibular symphysis to grasp the tongue, which is then drawn down, the remaining tissues behind the larynx being divided to release the neck structures. This should be done as far lateral as possible, so that the carotids can be removed with the laryngeal structures. It is now advisable to look into the pharynx and glottis before any further disturbance, to see if any obstruction, bleeding or other abnormality is present in the upper airway.

Removal of the thoracic contents

The subclavian bundles of vessels and nerves are divided by passing the knife from inside the thorax around the medial ends of the clavicles and first ribs to release the trachea and oesophagus. With gentle traction, the neck structures are held up and pulled caudally, whilst carefully clearing all attachments to the thoracic spine with the knife, taking care to keep the blade on the bone and not stray anteriorly to damage oesophagus or aorta.

Traction should be minimal and, as soon as the thorax is entered, the hand should move from the neck structures to place two fingers under the upper lobes of the lungs, lifting them and the mediastinum as the knife clears the midline structures down to the diaphragm. If the neck structures are pulled too hard by using them as a handle to drag out the thoracic viscera, they may be ripped off. In addition, the descending aorta may suffer transverse intimal tears from traction, which resemble the genuine 'ladder tears' seen in many traffic accidents (Chapter 9).

Pleural adhesions may prevent clean removal of the lungs. If there are only a few, they can be cut through. If the whole pleural cavity is obliterated by adhesions, they may be pulled away by making a cleavage plane with the hand and stripping them off. Sometimes (especially in industrial chest disease and old tuberculosis) the adhesions are dense, tough or even calcified. Removal of the lung may then be achieved by running a knife down the whole length of the parietal pleura over the inner anterior aspect of the ribs, and forcing a hand through the slit to form a cleavage plane to force the parietal pleura off the intercostal muscles and ribs.

Removal of the abdominal organs

When the chest organs are free, they are laid back in the thorax and the diaphragm incised. One hand should pull the liver and spleen medially, putting the left leaf of the diaphragm on the stretch, while the knife cuts it through laterally, near the costal margin. The cut curves posteriorly under the organs to reach the spine, where it must cut through the cruciate ligaments, then passes caudally behind the kidney, which is mobilized forwards.

The knife cut then curves up over the psoas muscle and ends at the brim of the pelvis. The same is done on the opposite side, the operator moving around the body if necessary. The chest organs are then lifted and gently pulled forwards to carry the abdominal viscera towards the feet. Any resistance is usually due to incompletely severed cruciate ligaments, which must be transected.

Eventually the organs will lay inverted across the pubis, secured only by the iliac vessels and ureters, which are cut through and the whole pluck of viscera taken away to the dissecting bench where running water and good illumination must be available.

Removal of the pelvic organs

The treatment of the pelvic contents depends on the type of case. Where the presumed cause of death is unrelated to pelvic lesions, then in men the bladder may be opened widely and the mucosa and trigone inspected before the prostate is incised for examination. The testes are pushed upwards through the inguinal canals, which are widened with the knife. In women, the ovaries are incised and the tubes examined from above before the uterus is sliced in the midline from fundus to cervix.

A more thorough examination of either gender may be made by enucleating the pelvic contents. The knife is passed circumferentially around the pelvic bowl after pulling the bladder away from the pubis. When the walls are free, the knife cuts through below the prostate and then through the lower rectum to allow the pelvic organs to be lifted out. In women the ovaries and tubes are mobilized forwards and the knife passed around the wall of the pelvic bowl, then in front of and below the bladder. The vault of the vagina and rectum is transected, freeing the whole contents. In cases where sexual interference or abortion is suspected, a special technique described in Chapter 19 is employed.

Removing the brain

Attention is then turned to the head. The scalp is incised across the posterior vertex from a point behind the ear to the corresponding place on the other side. Where a Y-incision is used on the neck, the limbs of the Y may be continued right across the scalp, especially if a face dissection is necessary.

The tissues are reflected forwards to the lower forehead and back to the occiput. The deep scalp tissues may peel off by traction, but often require touches of the knife to free them. Bruising is sought and, where head injuries are present or suspected, the scalp should be reflected right back to the nape of the neck, paying particular attention to the tissue behind and below each ear where injuries causing vertebrobasilar artery damage occur. Where there are facial injuries, the skin of the face may be peeled back from the jaw line and downwards from the forehead, restoration being excellent if care is taken not to perforate the facial skin during removal.

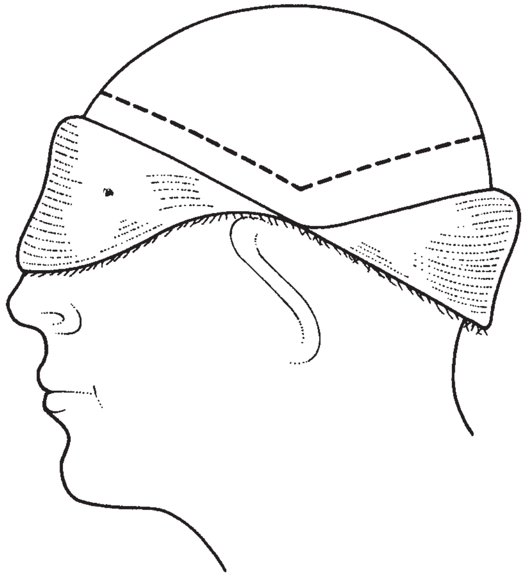

FIGURE 1.15 *Angled saw-cuts for removal of skull-cap to avoid slippage during reconstruction.*

The skull is sawn through, using either hand or power tools. The line of the cut should not be along a circumference, as it is then impossible to reconstitute the head without unsightly sliding of the calvarium. There should be an angled removal, with a horizontal cut from forehead to behind the ears joined by a second, which passes diagonally upwards at a shallow angle over the occipitoparietal area. Care must be taken not to place this posterior saw-cut too vertically (and thus anteriorly) on the skull, or the brain may be damaged by forcing its removal through too narrow an aperture.

The calvarium is then removed by leverage after complete cutting through. A mallet and chisel should not be used in forensic autopsies, even to ensure that the dura is kept intact. The risk of extending or even causing fractures by the use of excessive hammering is too great merely to justify an unmarked dural membrane. A cut dura is easily recognized as such by any competent pathologist. What is more important is to inspect the surface of the exposed dura and brain and assess any oedema, bleeding or inflammatory conditions that may be present. The skull-cap is carefully inspected for fractures and the dura peeled off the inside to study the inner skull surface.

To remove the brain, the dura is incised around the line of skull removal and two fingers slipped beneath each frontal lobe. With gentle traction the frontal lobes are lifted to expose the optic chiasma and anterior cranial nerves. The falx may have to be cut to free the brain, then a scalpel or blunt-pointed bistoury is passed along the floor of the skull to divide the cranial nerves, carotid arteries and pituitary stalk until the free edges of the tentorium are accessible. A cut is made along each side of the tentorium, following the line of the petrous temporal bones to the lateral wall of the skull. Continuing with traction on the brain, but being careful not to impact the upper surface against the posterior saw-cut, the knife severs the remaining posterior cranial nerves and then passes down into the foramen magnum to transect the spinal cord as far down as can be reached. The hand is now slid under the base of the brain, which is rotated backwards for removal, any attached dura being severed where necessary. The brain is taken into a scale pan and weighed before either fixation or dissection.

The floor of the skull is now examined and the basal dura stripped out with a strong forceps to reveal any basal fractures. Discarded dental forceps can be useful for this purpose. The venous sinuses are incised to search for thrombosis. Where appropriate – and always in infants – the petrous temporal bones are sawn, chiselled or cut with bone forceps to examine the middle and inner ears for infection.

Removal and examination of the spinal cord

It is not an invariable routine to remove the spinal cord at autopsy unless there are indications that some lesion may be present. Where there is the slightest possibility of damage to the vertebral column, its blood vessels or the contents of the spinal canal, however, there should never be any hesitation in extending the autopsy to include this area.

There are several methods of removing the cord and for full details, the texts of Ludwig (2002) or Knight (1983) should be consulted. Briefly, there are two main approaches to the spinal canal, the anterior and posterior.

In the anterior method, the vertebral bodies are removed after complete evisceration of the body, by sawing through the pedicles by a lateral cut down each side. The advantages are that the body need not be turned over on to its face and an extensive dorsal incision is avoided, which requires subsequent repair. The author finds this method more laborious, however, especially in the thoracic region where the heads of the ribs make the approach difficult.

The more usual posterior approach requires a midline incision from occiput to lumbar region, the paraspinal muscles being reflected along with subcutaneous tissues. Two parallel saw-cuts are then made down the length of the spine to divide the right and left laminae, and to give access to the spinal canal. This is best done with an electric oscillating saw, taking care not to cut so deeply that the spinal dura is penetrated. The strip of bone may be dissected off from below upwards to expose the spinal canal.

The cuts should be placed sufficiently lateral to allow the cord to be removed without difficulty. When the canal is exposed, the dura is examined for haemorrhage, infection or other abnormalities, then removed – still within its dural sheath – by transecting the nerve roots and dural attachments, and peeling it out progressively from below upwards. The dura is then carefully opened with forceps and scissors to examine the cord itself. It can be fixed in formalin, as with the brain, before cutting, or dissected immediately and samples taken for histology. Crushing, infarction, infection, haemorrhage and degeneration are the main lesions in a forensic context. The empty spinal canal must be carefully examined for disc protrusions, tumours, fractures, haemorrhage dislocations and vertebral collapse.

Where in any autopsy spinal damage is suspected, a good preliminary test is to slide the hands under the back of the eviscerated body on the autopsy table and lift the dorso-lumbar spine upwards, whilst watching the interior vertebral bodies. If a fracture or dislocation is present, abnormally acute angulation will be seen, instead of smooth bending. The cervical spine can be tested by manual manipulation. If suspicious angulation is seen, a slice can be taken along with the anterior spine, through the vertebral bodies and discs, with an electric or handsaw. This will reveal the interior of the spine and exhibit any crushing, haemorrhage, or torn disc spaces: if one of these is found, the cord must always be removed.

EXAMINATION OF ORGANS

Examination of viscera

The thoracic and abdominal viscera are laid on a cutting bench at a convenient height and under good illumination. Ample washing water should be available from a flexible pipe, to flush the tissues as dissection proceeds. Some pathologists maintain that this should not be done, as the water can have an effect on the quality of subsequent histological sections, but this has recently been disproved (Cotton and Stephenson 1988). In any case, the vastly inferior naked-eye examination that results if blood is not removed at frequent intervals greatly outweighs any unsubstantiated objections about the more exquisite details of cell structure, especially as in most forensic autopsies the gross appearances are usually far more important.

The viscera should be laid so that the tongue faces the pathologist, with the aorta upwards. The same sequence of examination should be carried out whatever the nature of the case, so that a fixed routine will ensure that nothing is left undone.

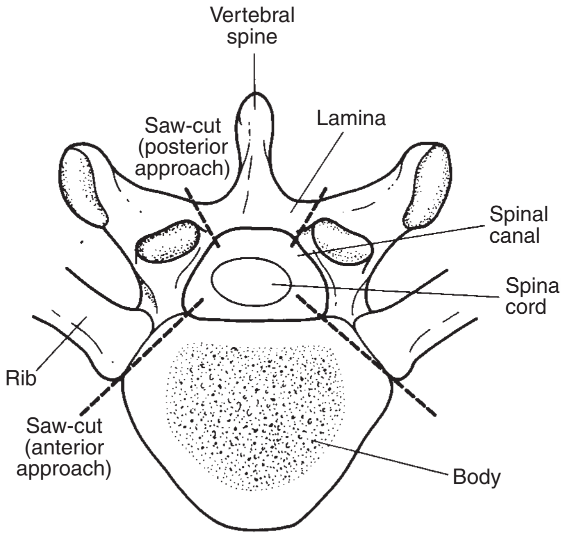

FIGURE 1.16 *The two approaches to the spinal canal for the removal of the spinal cord at autopsy.*

The neck structures

The tongue is examined for disease and injuries, including bites suggesting blows in the jaw or epilepsy. The tongue should be sliced to detect deep haemorrhages sometimes seen in strangulation. Such haemorrhage is seen mostly at the sides and centre of the mid-part of the tongue. Gross congestion, which may be due to either pressure on the neck or to other congestive modes of death, is usually in the posterior part of the tongue. The tonsils and pharyngeal walls are inspected.

The glottis is examined for mechanical or infective obstruction, and the hyoid and thyroid horns palpated for fractures. The oesophagus is opened with large (20 cm) blunt-nosed scissors, which along with a very sharp 10–15 cm bladed knife and a long-bladed 'brain knife', are the most useful tools for performing an autopsy.

The carotid arteries on each side are opened, including the bifurcations and sinuses. If necessary, the upper portions of the carotids are explored in the body itself and followed to the base of the skull. If thrombosis is suspected, the intracranial part should be examined in the cavernous sinus.

Returning to the neck structures, the thyroid should be sliced and inspected, then the oesophagus opened almost to the cardia of the stomach and any suspect material such as capsules, tablets or powder retained for analysis.

The scissors are then passed down the posterior line of the larynx and trachea to the carina. If pressure on the neck of any type, such as strangulation, is suspected, then special examination should be made, as described in Chapter 14.

The trachea and main bronchi should be inspected for disease and obstruction. Gastric contents are often found, but the significance of this is discussed in Chapter 13 – it should not be assumed that ante-mortem aspiration has occurred merely from the presence of gastric contents in the air passages.

The lungs

The lungs are then removed, after careful examination of their external surfaces for patchy collapse, emphysema, petechiae and so on. Almost every autopsy will reveal a few petechiae, especially around the hilum and in the interlobar fissures. Their significance is also discussed in Chapter 13.

The lungs are removed from the thoracic pluck by passing a long-bladed knife (such as a brain knife) under the hilum with the blunt edge upwards. The knife is settled in the correct position before turning the sharp edge upwards to cut through the hilum. Before doing this it may be necessary to remove adhesions over the diaphragm and to cut through the pulmonary ligament, a thin sheet of tissue that ties the inferior medial edge of the lower lobe to the mediastinum.

As the hilum is being cut, the pathologist must notice if any embolism is visible within the pulmonary arteries. It has happened that such an embolism has slid out and been washed unnoticed down the sink. Some pathologists insist on opening the main pulmonary trunk and even right ventricle before removing the lungs, to seek a saddle embolus. This is not necessary, as any large embolus will be readily visible on examining the heart and the lungs in the usual sequence.

Both lungs are taken off and the hilum inspected before being laid aside for cutting. The lung should be weighed before cutting, as appreciable oedema fluid can run away during dissection. Then each is laid with the hilum down on the dissecting board, the opportunity being taken during handling to evaluate weight and oedema, as well as emphysema.

The lung is held on the upper surface by the left hand of the operator (or by an interposed sponge) and the organ cut across in the sagittal plane from apex to base with the large brain knife, held parallel to the board. This produces an anteroposterior slice, the lower medial part carrying the hilum. The cut surfaces can now be opened like a book and the surface examined for oedema, tumour, pneumonias, infarction, trauma and so on. The smaller bronchi must be inspected for such signs as mucosal thickening, infection and blockage. The smaller pulmonary arteries may reveal thrombosis or embolism that was not visible in the larger vessels.

INFLATING THE LUNGS WITH FORMALIN

In some medico-legal autopsies, especially in industrial lung disease such as pneumoconiosis or asbestosis, one or both lungs need to be inflated with formalin for fixation before cutting. This preserves the shape and histology in excellent condition, but delays examination for at least several days. It is carried out by holding or tying a cannula into the bronchus while 10 per cent formol saline is perfused through a tube from a reservoir held about 1 metre above the lung. The lung is then left in a bath of formalin covered with a formalin-soaked cloth to prevent drying.

The heart and great vessels

There are almost as many ways of examining the heart as there are pathologists, and each operator must decide upon the method that appeals most. In this summary there is no space for discussion of post-mortem angiography, which is moving from the field of research and special interest into routine use. A common and practical routine for examining the heart is described here.

First, the organ mass minus the lungs is rotated so that the lower end now faces the pathologist. The scissors are passed into the cut end of a common iliac artery and passed right up to the aortic arch and around to a few centimetres above the aortic valve, staying outside the reflection of the pericardium. The interior of the aorta is studied, especially for the degree of atheroma and for any aneurysms or trauma. The inferior vena cava is opened from its lower end into the liver. The organ pluck is then turned over so that the heart is uppermost. The pericardium is inspected externally for fluid and blood tamponade, then opened widely with scissors. The heart is delivered through the incision and inspected externally for pericarditis, adhesions, discoloration of an underlying infarct and cardiac aneurysms, for example. In a child, the thymus would be inspected and dissected off at this stage.

The heart is then removed by holding it up with the left hand so that its attachment is tensed against the other organs. A long knife, such as a brain knife, is then passed horizontally across at the reflection of the pericardium, cutting through the root of the aorta and other great vessels just above the atria.

The now-detached heart is washed externally and placed in the anatomical position on the dissecting board, with the apex facing the operator and the anterior surface upwards. It should not be weighed until all the contained blood and clot is removed. The general size, shape and ventricular preponderance should be noted. Any dilatation or thickening of the pulmonary conus should be noted as an index of right ventricular hypertrophy, especially if striae of transverse muscle fibres are seen crossing the conus.

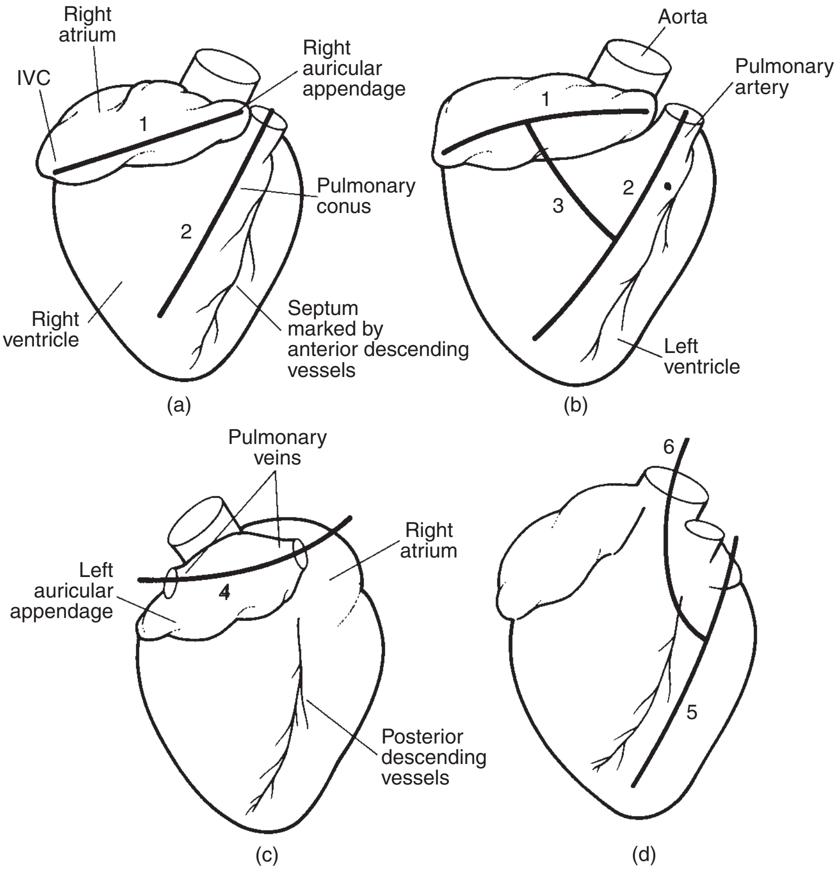

FIGURE 1.17 *Incisions for opening the heart at autopsy. (a) The right atrium is slit with an incision (1) to join the vena cava to the appendage; a cut parallel to the interventricular septum is made on the anterior wall of the right ventricle (2) passing up through the pulmonary conus. (b) These cuts are joined through the tricuspid valve (3). (c) The heart is reversed and the left atrium opened by a cut (4) joining pulmonary veins. (d) On the anterior wall a cut (5) is made parallel to the septum through the mitral valve and joined by (6) which passes through the aortic valve.*

The right atrium is then opened by introducing the scissors into the inferior vena cava and cutting across to the atrial appendage. The interior of the atrium is examined and the septum and tricuspid valve inspected.

The interventricular septum is then identified externally by the vessels running down the outside. With a knife a cut is made about 15 mm to the right of and parallel to the septum, over the right ventricle. This should be deep enough to enter the lumen, but not enough to cut the posterior wall. The scissors are now introduced into the cut and run up through the pulmonary conus and into the pulmonary artery until they meet the transected end. They are also extended downwards to the apex of the ventricle. The scissors are now put in midway down this linear cut and passed outwards at right angles, guided by the fingers of the left hand passed into the tricuspid valve from the opened atrium. The whole of the right side of heart is now open and displayed. It should be washed out and the endocardium and valve examined.

A similar routine is now employed on the left side. The scissors are introduced into a pulmonary vein and passed horizontally across to an opposite vein, thus opening the atrium. Fingers are introduced down through the mitral valve to estimate its size and detect any stenosis.

The heart is then restored to the anatomical position and a cut made again parallel to the septum, but on the left side, going deeper as the ventricle is thicker. Guided by the fingers still in the mitral valve, the cut is extended upwards through that valve and out at the top of the atrium.

A finger is now passed up the outflow tract to the aortic valve to estimate its size. Then the scissors are passed up at the side of the mitral valve, and the aortic valve and aortic stump opened. The whole heart is now open and can be washed out and weighed.

Various estimates of the normal heart weight exist and vary considerably (see the Appendix 1). It has been related to gender and body weight, though this is not altogether sound, as a fat person of moderate stature does not have a heart weight comparable with a large muscular person unless there is associated hypertension. This controversial matter cannot be pursued here, but as a rule of thumb the author (BK) accepts up to 380 g as normal in an adult man of average build.

After weighing, the endocardium and valves are examined, then the coronary arteries. Once again, controversy exists about methods of opening the coronary vessels, but the weight of opinion now lies almost universally with those who cut serial interrupted cross-sections with a knife, rather than open them longitudinally with small scissors. The disadvantages of lengthwise opening are that percentage assessment of stenosis or the recognition of total occlusion cannot be made once the vessel is flapped open, as restoring the cut edges can never reproduce the original conditions. In addition, the tip of the scissors may dislodge thrombus or an intimal flap. Cutting across the vessels allows an estimate of the percentage stenosis. It is admitted that this may not be the size of the lumen during life when normal blood pressure is operating, but the same disadvantage applies to longitudinal opening. An estimate of luminal size is relatively constant, however, in that the collapsed abnormal can still be compared with the collapsed normal.

The coronaries are therefore cut across at frequent intervals. Before the first cut, the ostia are examined for congenital variations (which are frequent) and for obstruction. The left coronary artery is then cut across from the epicardial surface, starting as close to the ostium as possible, as occlusion and severe stenosis can occur very near the origin. Serial cuts are then made at intervals of not more than 3 mm, first into the common trunk, then following the left circumflex laterally until the vessel becomes too small, usually when it dips down from the epicardium to become intramuscular. The anterior descending branch is then followed down the front of the septum almost to the apex.

Turning to the right coronary artery, the proximal segment is cut back from the point where the right ventricle was opened, transecting the artery in its mid-part. The cuts are made back to the aorta, then the distal segment is followed laterally until it becomes the posterior descending branch. During this process, the dominance of right versus left vessels is noted.

Difficulty arises where severe calcification exists, as the knife either fails to cut through the artery or shatters it because of the excessive force required. The lumen is crushed and the percentage stenosis is difficult or impossible to assess. Scissors may be used to exert more force than a knife, but the only real solution is to decalcify the vessels. Except for research purposes, it must be admitted that the days or even weeks of delay occasioned by decalcification provide a formidable deterrent for the busy coroner's pathologist unless the issues involved are important. Post-mortem angiography is the other alternative, though where autopsy caseloads are high, this can be difficult to arrange.

Once the coronary arteries have been examined, the myocardium can be studied more closely. One useful technique is the intramural or 'sandwich' cut through the thickness of the left ventricle. This is easier when there is some ventricular hypertrophy, but can be carried out in any heart. The heart is placed open on the cutting board, with the endocardium downwards. A long knife such as a brain knife is passed carefully into the cut edge of the left ventricle and sliced right through the muscle, keeping equidistant between endocardium and epicardium. The myocardium can then be opened out like a book, showing the interior with any infarcts or fibrotic plaques. If histological blocks through the entire wall are needed it is best to take these before slicing the heart as described, otherwise a full-thickness block will be hard to obtain. Some pathologists make a series of transverse heart slices about 8–10 mm thick, starting at the apex, before opening the heart. This displays the distal myocardium wall, but cannot expose the proximal part, as examination of the coronary arteries would then be compromised.

Examining the abdominal organs

The remaining organs are laid on the dissecting surface in the anatomical position, with the liver and stomach away from the pathologist, and the anterior surface upwards.

The stomach is opened and, if there is any question of requiring the contents for analysis, a suitable container should be ready. The stomach is washed externally with a stream of water and a small cut made in the greater curvature. If the contents are to be saved, this can best be done with the edge of the stomach projecting over the raised cutting board or over the edge of the sink, so that the container can be held underneath. The contents are allowed to run into the container, then the greater curvature is opened widely with scissors, and any remaining contents drained or scraped out. If powder or other substance is adherent to the mucosa, this can be scraped off and added to the container, or kept in a separate tube for analysis.

Some laboratories require the stomach wall to be retained for analysis as well as the contents. This can be done after the organ has been fully opened and the lining examined. If the contents are not required – or when they have been collected – the organ is opened up from cardia to pylorus along the greater curvature. The lining is washed and inspected. The scissors are now passed through the pylorus around the duodenum until they meet the point where the gut was detached earlier. The gall bladder may be squeezed to demonstrate the patency of the bile ducts. In some few cases, bile may be required for analysis, as in morphine or chlorpromazine overdose.

The adrenals are examined next, the right one being on top of its kidney, the left being on the medial side. If the right kidney is taken in the left hand and lifted against the weight of the liver, a single cut into the stretched tissues between liver and kidney will transect the adrenal. The left is buried in the tissue between pancreas, spleen and kidney. The amount of cortical lipoid and the absence of bleeding or other abnormality is noted in each gland.

The spleen is removed by cutting through its pedicle and is sliced after weighing. The pancreas lies under the stomach and should be cut lengthwise from the curve of the duodenum to its tail lying against the splenic hilum.

The kidneys are exposed by incising their capsules, often after a thick layer of perirenal fat is traversed. The kidneys can usually be peeled out of their capsules unless these are adherent.

The renal vessels were examined when the aorta was opened, but can be re-examined at this stage and the ureters inspected. The organs can be detached at their hilum and weighed, then cut lengthwise to inspect the interior. The width of the cortex is important, being about 1 cm in a

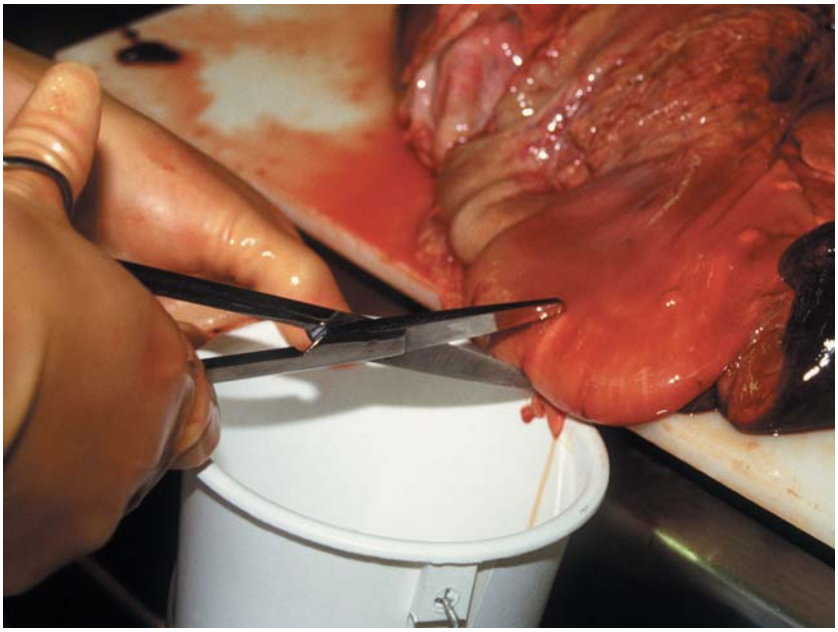

FIGURE 1.18 *Collecting stomach contents at autopsy. The exterior of the stomach is washed free from blood and the viscera moved to overhang the dissecting board. A chemically clean receptacle is held beneath the greater curvature while the latter is opened with scissors and the contents allowed to drain out. The stomach is then opened, the mucosa inspected and any adherent material scraped into the container. The stomach wall should then be dissected off and added to the contents if the laboratory is able to deal with tissue analysis.*

healthy subject. The granularity of the surface and the clarity of the corticomedullary junctions is assessed, as well as the size of the renal pelvis.

The small intestine may have been examined at the earlier stage of removal, but can be inspected now. It is not usual to open the whole length of the gut in a forensic autopsy unless there is any particular indication to do so, though ideally this should always be carried out.

EXAMINATION OF THE BRAIN

After weighing, a decision has to be made whether to examine the brain immediately – the so-called 'wet-cutting' – or to suspend it in formalin until fixed, a process which takes at least several weeks. The advantage of fixation is, of course, that the firmness of the tissue allows thinner and more accurate knife-cut sections to be taken, as well as better histological preservation. Where neurological issues are involved, either traumatic or from natural disease, it is almost mandatory for the brain to be fixed before cutting. Even the impatience of the investigative authorities can usually be overcome if the advantages of a higher standard of opinion are explained.

The technique of brain fixation is well known, but to summarize briefly, the brain is suspended in a container of 10 per cent buffered formalin, the volume being at least 5 litres and preferably 8. The brain is removed with the dura leaving the parasagittal bridging veins and falx intact and suspending the brain in an upright position by the falx. An alternative method is to pass a thread or metal paperclip under the basilar artery and tie it to a support across the mouth of the container, so that the vertex is clear of the bottom.

In the majority of autopsies there is no real need for fixation if no cerebral lesions are either expected or apparent on external examination of the brain. Here, 'wet-cutting' is sufficient, though if any unexpected lesions are found, the process can be stopped and the slices of brain fixed by placing them in a large volume of formalin, on cotton wool pads, to prevent distortion.

One lesion is better examined in the unfixed state, though the brain may be suspended later; this is subarachnoid haemorrhage. It is easier to wash away fresh, unfixed blood with a stream of water and blunt dissection than the hardened blood that develops during formalin fixation. Further details are given in Chapters 5 and 25.

Whether the brain is examined 'wet' or fixed, the sequence is the same. The weight is first considered, the normal for a young male adult being between about 1300 and 1450 g, the female equivalent being around 100 g less. More details are given in Appendix 1. It should be noted that formalin fixation adds about 8 per cent to the original weight.

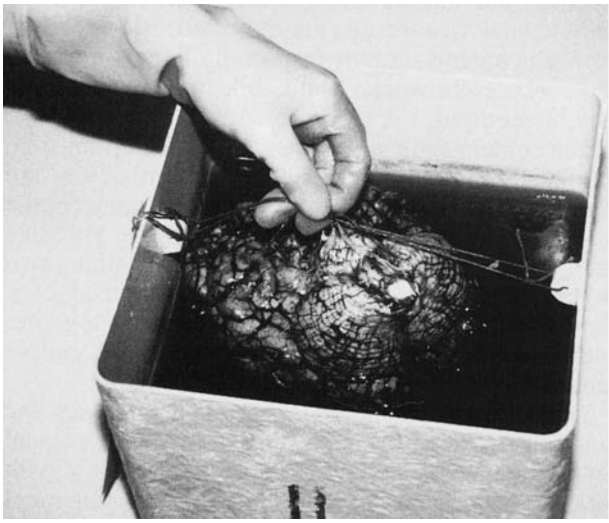

FIGURE 1.19 *A brain suspended in formalin for fixation before cutting. The tank is specially made of fibreglass, being cubical for stacking with a lid (not shown). There are lugs moulded into the sides to hold the suspensory strings, which support the brain by means of a paperclip hooked under the basilar artery. There should be sufficient fluid to allow the brain to float clear of the bottom of the receptacle.*

The brain is first examined for surface abnormalities, which in forensic practice usually means haemorrhage. This is dealt with in Chapter 5, but suffice to say here that meningeal bleeding is one of the most important lesions in forensic pathology, be it extradural, subdural or subarachnoid. This means that examination of the cerebral vessels, especially the arteries of the circle of Willis and the vertebral vessels, is vital – especially in the search for berry aneurysms. The general symmetry of the brain is then noted as well as any depression of the cortex from skull or meningeal masses. An estimate of cerebral oedema is made, partly from the weight, but mainly from flattening of the gyri, filling of the sulci and evidence of hippocampal herniation through the tentorial aperture. In lesser degree this may be seen as grooving of one or both unci, though a normal slight anatomical groove is often present. True uncal herniation is marked and often discoloured as a result of incipient infarction. Similarly, herniation or coning of the cerebellar tonsils through the foramen magnum must be distinguished from the common anatomical pouting of many tonsils: true pressure coning is often discoloured by local infarction. After careful inspection of the basal vessels and the exterior of the brain, and palpation for any fluctuant masses under the cortex such as internal haemorrhage, abscesses or cystic tumours, the organ is cut.

The first cut should be made through the cerebral peduncles using a long, broad-bladed knife to separate the cerebrum from the brainstem and cerebellum. The cerebellum is then held in one hand with the cut stem upwards.

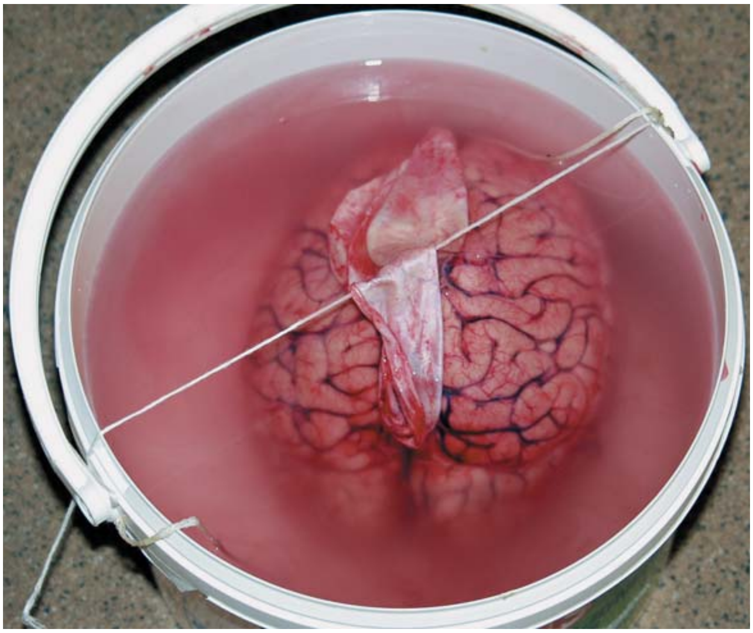

FIGURE 1.20 *An alternative method of suspension is to leave the falx intact and use it to suspend the brain base down in formalin.*

This is examined to assess the substantia nigra and aqueduct, as well as to note any primary or secondary haemorrhage, the latter often being due to raised intracranial pressure. The cerebellum and pons are then cut down vertically and opened like a book to display the fourth ventricle, dentate nuclei and the interior of the cerebellum. The lower pons and medulla can be further sectioned transversely or longitudinally.

The cerebral hemispheres are then placed base down on the cutting block and serial sections made in the coronal plane from the frontal lobes back to the occiput. The sections should be about 1 cm thick and each should be slid into a sequential place in rows along the cutting board so that orientation is preserved. Cuts should be made with careful but bold sweeps of the knife as erratic sawing motions will leave an irregular surface on each section, which obscures a good view. Cutting a fresh brain is less satisfactory than cutting a firmer fixed brain, especially if there is any post-mortem autolysis or softening as in a dead ischaemic brain after mechanical ventilation.

ANCILLARY INVESTIGATIONS

A wide range of samples may need to be taken either before, during or after the gross examination is completed. The nature of such ancillary investigations naturally depends upon the nature of the death, the history and the interests of the pathologist.

Microbiology

Though more common in clinical autopsies than forensic work, culture samples for ‘bacteriology’, virology and (rarely) fungi may be needed. Either plain swabs or swabs immersed in a transport medium can be employed for sampling a wide variety of sites at autopsy. Alternatively, tissue samples may be collected in sterile containers and this is the usual method for virological culture of lung and brain, for example. Blood cultures may be desired, and it is best to take blood with a sterile needle and syringe from a large vessel, such as the femoral vein, before starting the autopsy with its attendant inevitably widespread contamination with putrefactive organisms. Alternatively, blood can be taken from a freshly opened heart chamber using sterile instruments.

In all autopsy investigations, unless cultures are performed soon after death, there is often widespread contamination by organisms that travel rapidly through the dead tissues, especially from the gastrointestinal tract. It requires the expertise of an experienced microbiologist to advise on what is significant in the subsequent growth in the laboratory.

Toxicology

This has already been mentioned and is further discussed in the later chapters of this book but, as far as the collection of specimens is concerned, it may be repeated here that sampling is extremely important if reliable analytical results are expected. Blood, urine, stomach contents, organs (especially liver), intestinal contents, cerebrospinal fluid, bile and ocular fluid may be required.

The containers in which specimens are collected must be chemically clean to a very high standard. They are often supplied by the laboratory that will carry out the analyses. For example, the Home Office Forensic Science Service in Britain supply a kit containing two large plastic tubs with lids, several universal containers, fluoride tube, syringes and needles, together with instructions as to type and quantity of sample needed for various tests.

The pathologist should submit a form with the samples indicating the analyses required, the personal details of the deceased person, a brief history with details of suspected toxic substances, and a statement as to whether the subject was known to suffer from any infective condition, including hepatitis or HIV infection.

Histology

In most autopsies and inevitably in all criminal or litigious cases, the pathologist will need to carry out a histological examination on a range of tissues, even if only to exclude the possibility of some occult natural disease. The Royal College of Pathologists as well as the 1999 Council of Europe recommendation on the harmonization of medico-legal autopsy recommend that histology be taken from every autopsy. This advice should be followed wherever possible, even though in some medico-legal autopsies, the cost of such techniques may provide difficulty if the coroner or other commissioning authority declines to fund the procedure. It is customary to retain samples of liver, spleen, kidney, heart, lung, thyroid, adrenal, pancreas, muscle and brain as a minimum. Where there are indications to examine other parts of the body, these are retained in addition to the routine tissues. The tissues are either sampled by taking relatively large pieces at autopsy (which are later trimmed down to size) or by cutting blocks of a standard size (such as 20 × 12 × 5 mm) at the time of autopsy.

The tissue is placed in a large volume of buffered formol–saline and allowed to fix for at least several days before processing. The volume of fixative should be six times the total volume of tissue: it is all too common to see a mass of tissue squeezed into a small container, barely covered with formalin, half-fixed and even semi-dried. There are many specialized procedures also in use at present, such as histochemistry, fluorescent microscopy and immunohistochemical examination; these are used especially in the pathology of sudden death where evidence of early myocardial infarction is sought. This is discussed in the chapter on sudden natural death (Chapter 25).

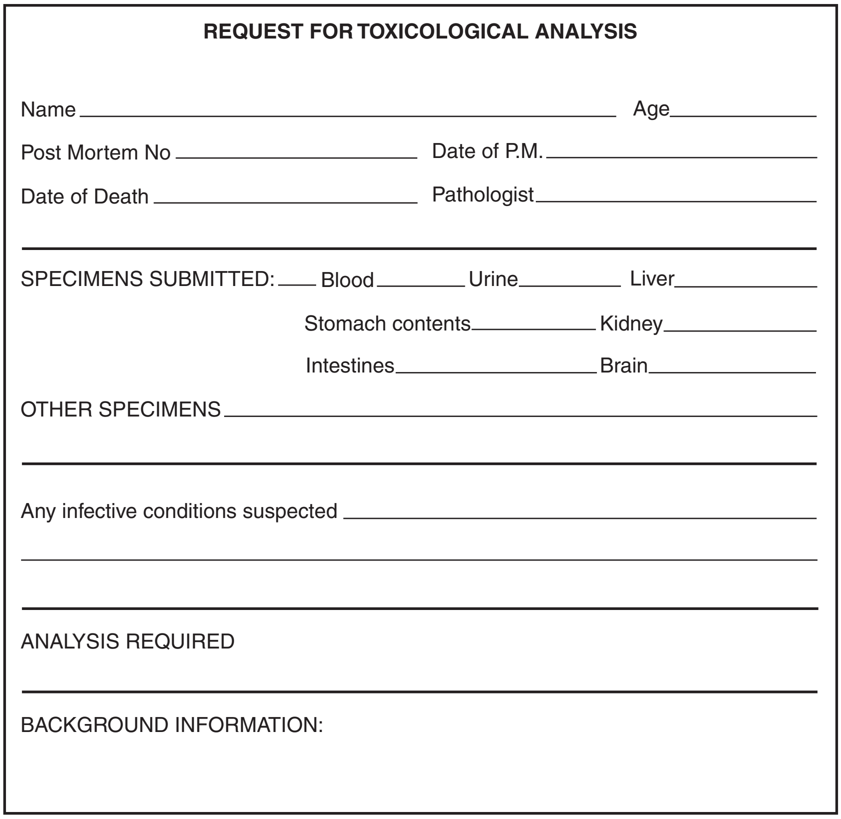

REQUEST FOR TOXICOLOGICAL ANALYSIS

Name ______________________ Age __________

Post Mortem No ______________ Date of P.M. ______________

Date of Death ______________ Pathologist ______________

SPECIMENS SUBMITTED: ___ Blood ______ Urine ______ Liver ______

Stomach contents ______ Kidney ______

Intestines ______ Brain ______

OTHER SPECIMENS ______________________

Any infective conditions suspected ______________________

ANALYSIS REQUIRED

BACKGROUND INFORMATION:

FIGURE 1.21 *Sample request form for toxicological analysis.*

AUTOPSY RADIOLOGY

The details of radiological findings are discussed in each appropriate chapter, particularly in relation to child abuse, gunshot wounds, identification and dentistry. In general, radiology is similar to photography in the autopsy room, though it requires more bulky apparatus, and someone to operate the equipment and process the films.

The quality of radiographic assistance available will vary widely from none to the most sophisticated techniques, even including computed tomography and even, in very few centres, magnetic resonance imaging (MRI).

The main difference is usually whether the autopsy is carried out in the mortuary of a well-equipped hospital where there is radiographic apparatus, radiographers to use it and radiologists to read the films – or whether the autopsy is performed in some remote public mortuary or in makeshift premises far from clinical facilities.

In many developing countries, radiography may be scarce for living patients let alone corpses, and the same standard of assistance for the pathologist cannot be expected. In larger, more affluent countries, remoteness may be the problem; here mobile equipment is available and sometimes used by forensic pathologists. There are small, portable X-ray kits that can be carried in two suitcases, and can function from an ordinary domestic power supply or from a portable petrol generator. In many instances a mortuary will have a small portable machine, sometimes one discarded from clinical use. This can be wheeled into position when needed, and films and technical assistance obtained from the nearest hospital, which is often adjacent.

For isolated organs and tissues, cabinet-type radiographic equipment is available, which can be used without the assistance of trained radiographers.

Where there is a large forensic institute, then full radiological facilities with the staff and equipment to operate and process plates will be available at all times. Where research projects involving radiology, such as post-mortem coronary angiography, are being carried out, there will usually also be ample facilities for routine use.

The stage at which radiology is employed in an autopsy will vary according to the individual circumstances, but is likely to be after the external examination is complete, but before dissection begins. Straight radiographs for bony injury are not often required, as the skeleton can be inspected directly by dissection in major trauma. The exception is child abuse, where, as fully described in Chapter 22, a full skeletal survey is needed before autopsy. Indeed, many forensic and paediatric pathologists would consider radiology essential in all infants who did not die of some obvious disease.

Suspected air embolism, pneumothorax, barotrauma, gunshot and explosive deaths should have radiological examination before the autopsy, and when a traumatic subarachnoid haemorrhage is suspected, vertebral artery angiography might be necessary, as described in Chapter 5.

Mutilated remains, especially those from mass disasters, may need to be X-rayed, as may victims from fires where the external damage makes dissection difficult. Where bombs or explosive devices are involved, it is essential to have radiographs to detect any parts of the mechanism that are embedded in the tissues.

Though the films are usually taken before the autopsy begins, some lesions may be better demonstrated on isolated

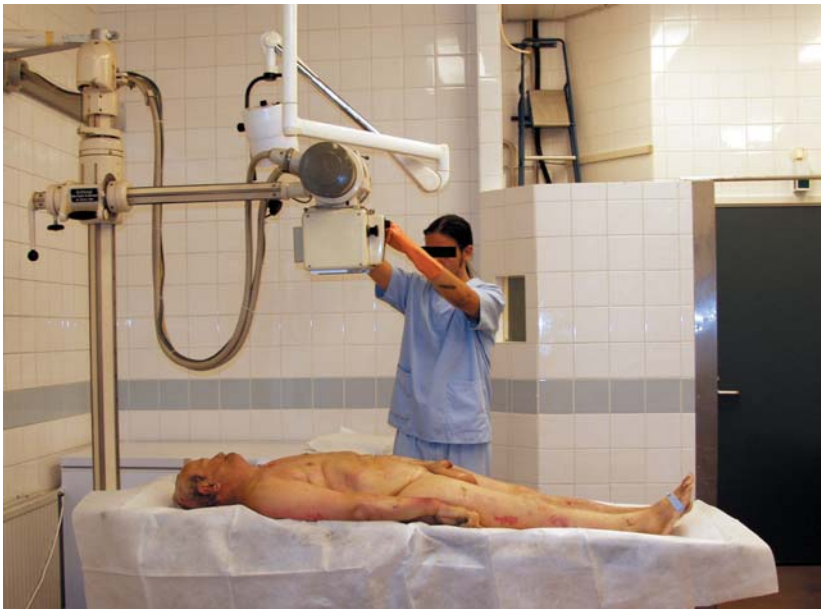

FIGURE 1.22 *Radiography is essential before some autopsies. Here it is being used to X-ray a gunshot wound of the chest.*

organs or structures. In child abuse, the callus of old fractures of posterior ribs may be better visualized on X-ray if the chest cage is dissected out, and radiographs taken without the soft tissues and the obscuring structures of sternum and anterior ribs. Similarly, a block of upper cervical spine may better reveal fractures of transverse processes carrying the vertebral artery, or a larynx may show fractures of the hyoid or thyroid cornuae more clearly when they are X-rayed outside the body.

FORENSIC PHOTOGRAPHY

Many forensic pathologists take their own photographs of both scenes of death and of autopsy appearances. Others rely on professional photographers such as the police and hospital medical photographers.

The standard of expertise amongst some doctor photographers is excellent, and their pictures grace many lectures and textbooks. It is to the more inexpert camera operators that these few general tips are offered.

The type of camera most favoured is the 35 mm single-lens reflex, of which there is now a vast range available, in all grades of sophistication and price. Some means of exact focusing is essential (either automatic or split-screen, for example) as picture sharpness is of prime importance.

The type of lens is a matter of personal choice, as some will prefer interchangeable lenses of various focal lengths. The standard 50 mm is most useful, but for scenes of crime or taking a full-length shot of a body in cramped conditions, a wide-angle lens of 28 or 30 mm is needed. A longer focal length of up to about 80 mm can be useful for close-up pictures of small lesions, but telephoto lenses of 100–200 mm are not required, though extension tubes, some with lenses, can be obtained for macrophotography.

Many pathologists, including the authors, prefer to combine lenses into a single variable-focus ‘zoom’ lens of 28–80 mm range. This saves time spent in changing lenses, and the image resolution of good-quality equipment is virtually indistinguishable from fixed focal length lenses.

Illumination is usually by electronic flash now often an integral part of the camera body, and the use of automatic thyristor control means that no complicated calculations about range are needed. For very close-range work, a flash attached to the camera may be unsatisfactory, so an extension cable is a useful and cheap accessory to keep the flashgun at a distance. Alternatively, the flash can be ‘bounced’ off the ceiling in the autopsy room or a ‘ring-flash’ around the lens used to avoid camera shadow.

Some will prefer tungsten light rather than flash, though this is more cumbersome unless a fixed station is installed in the autopsy room for taking photographs of organs. It is impracticable to use floodlights at the autopsy table or at a scene.

The type of film depends on the nature of the illumination. Speeds of 100 or 200 ASA are more than adequate for most flash work, though 400 ASA is now frequently used. Some keen medical photographers carry a separate camera body loaded with high-speed 1000 ASA film for special circumstances.

Ultraviolet (UV) and infrared sensitive film has been used to demonstrate surface lesions that are not visible to the human eye. It has been claimed that occult bruising can be revealed by UV photography, such as in child abuse, but care must be taken and experience gained to eliminate artefactual false positives.

Most doctors will want transparencies for projection at lectures, which can be made into prints, either colour or monochrome. If prints are specifically wanted for black and white publication in books or journals, it is better to take them originally on monochrome film, as colour negatives do not provide quite the same quality reproduction in black and white.

Some pathologists use Polaroid cameras either alone or together with conventional film. The advantage of instant prints is that a record of the scene of death can be obtained before the autopsy is carried out, and any findings checked back against the original surroundings. For example, if some linear mark is seen on the leg at autopsy, the pathologist can immediately refer to his Polaroid prints to check whether that leg had been resting on some object. Again, where multiple injuries exist, their number, position and size can be checked when writing the autopsy report, without waiting days or even longer for the police album to arrive. Correlation of wounds with blood splashes may also be useful. Such Polaroid pictures can be duplicated and even turned into transparencies, but it is preferable to have a conventional camera for these purposes.

Recent advances in electronics are revolutionizing photography, and have wide application and potential for forensic and autopsy work. Electronic cameras of varying grades of sophistication, resolution and cost, can now instantly store images on a small internal ‘memory stick’ or CD-ROM, which can then be viewed immediately on a computer VDU or printed on a colour laser printer. These images can be manipulated in many ways to enlarge sections or correct colour balance, sent via modems to distant locations, or incorporated within textual material and reports. At present, good resolution can still be obtained more cheaply with silver-based film stock but this situation is changing in favour of the digital photography. Alternatively, photographs and slides can be electronically scanned within minutes, the optical image being digitized into electronic storage for similar manipulation.

All electronic images can also be stored on CD-ROM or DVD equipment, which offers huge storage and retrieval capacities for record-keeping and educational display purposes. In addition, single frames from video camera recordings can be captured as still pictures and electronically stored.

In relation to the actual photography, a marked improvement in the pictorial quality can be achieved with a little care in composition. At scenes of death the viewpoint is often limited and the surroundings have to be photographed as they exist, without modification. The camera position can, however, usually be chosen to cut out as much extraneous background as possible. In the autopsy room, considerable improvement can be made by using the optimum viewpoint and by modifying the background. Many photographs are spoiled by having irrelevant and distracting objects in the background such as observers, buckets, boots and the other extraneous paraphernalia of the autopsy room. The picture frame should be filled as completely as possible with the object under display. A close shot should be taken to limit the irrelevant margins though, where necessary, anatomical landmarks should be included to orientate the viewer. The camera should be at right angles to the lesion being photographed, whenever possible. Tangential shots may foreshorten the required feature and, where size needs to be displayed, may distort both shape and length because of the foreshortening. A ruler or special white adhesive tape with centimetre markings should be placed very near the lesion or wound to provide a size reference.

Where a feature is obscured or insignificant, it may be pointed out by a probe or finger held in the appropriate position. Where part of the picture area consists of the metal or porcelain autopsy table or dissection bench, allowance should be made in exposure, as these very reflective areas may give a false reading in the light meter of the camera and thyristor of the flashgun.

If a full length or half shot of a body is required, an untidy background may be avoided by turning a mobile or rotatable table so that a blank wall forms the backdrop. If this is not possible, assistants can hold up a sheet behind the table to screen the distant confusion.

When photographing viscera, the camera should be near enough for the frame to be almost completely filled by the required object. The shot should be vertical to the lesion and it is often necessary either to place the dissecting board on the floor, or for the photographer to stand on a stool or some elevation to gain the required height.

Where isolated organs are being photographed, they should be placed on a green or blue cloth, such as a discarded operating gown. White can be used, though it may affect the exposure meter if much is visible around the periphery. The organ should be placed on the cloth in one movement and not moved thereafter, otherwise a wet dark stain will obtrude on the green or blue background. The organ should not be oozing blood onto its surface or onto the background. It should be dabbed with a dry cloth or sponge just before the photograph is taken to remove shiny wet highlights.

Ideally a special stage for organs should be used, such as a glass-topped table with a coloured (usually green) background set sufficiently far below the glass to be out of focus. Tungsten lights can be used to advantage on such a fixed stage.

THE AUTOPSY REPORT

Equally important as the autopsy itself is the report that the pathologist provides for whoever commissioned the examination. An autopsy is of little value if the findings and opinion of the forensic pathologist are not communicated in the most lucid and helpful way. The report is an integral part of the procedure and should receive as much attention as any physical procedure in the autopsy room. Unfortunately, some pathologists treat the process of making a report in a somewhat cavalier manner, which diminishes the expertise that they may otherwise possess.

The autopsy report is a permanent record of the findings and is especially vital for medico-legal purposes, when every word may be dissected in a court of law months or even years afterwards, and when all recollection of the examination has been driven from the mind of the pathologist by hundreds of subsequent autopsies. In a clinical autopsy in a hospital, the dissection may be demonstrated and discussed at the time with the interested physician. However, the report of a forensic autopsy becomes a legal document of possibly vital significance, and every effort must be made at the time to make it as comprehensive and useful as possible.

The form of the autopsy report

Reports fall into two main types described below, and local practice and indeed legislation may determine which is used, irrespective of the wishes of the pathologist. The choice is, however, often dictated by the nature of the case.

- A free-style ‘essay’, which usually adheres to a conventional sequence, but leaves the pathologist free to expand on various aspects according to his estimate of their importance. This type is usually used in criminal deaths and cases in which litigation is likely. It has the advantage that any part of the autopsy can be expanded without constraint; also the form of the report can be turned into a legal statement or a deposition for the court with little alteration.

- A printed proforma, in which the various sections of the examination and organ systems are already set out by title, leaving blank spaces for the insertion of the findings. The advantage includes the fact that this ‘shopping list’ acts as an *aide-mémoire* to those pathologists who do not conduct large numbers of autopsies – and also the non-medical recipient is more able to follow the set pattern of the report. One disadvantage is that the spacing prevents flexibility of description unless the proforma is large, when much of the space may be left blank. Also there is rarely enough space at the end of the form for an expansive discussion and opinion about the preceding factual findings. This type of form is commonly used for non-litigious autopsies – for example, the usual coroner’s cases of sudden death and suicide.

Concentrating upon the more serious cases, the report, whatever the format, must contain certain information in some logical order. Consecutive numeration, computer codings and other administrative aspects are naturally conditioned by local practice, but the following matters must be catered for in all autopsy reports, though not necessarily in this sequence:

- Full personal details of the deceased subject, unless unidentified. This includes the name, gender, age, occupation and address.
- The place, date and time of the autopsy.
- The name, qualifications and status of the pathologist.
- Persons present at the examination.
- Usually, the authority commissioning the autopsy.
- A record of who identified the body.
- The name and address of the deceased subject’s regular (or last) medical attendant.
- The date and time of death, where known.
- The history and circumstances of the death. The inclusion of this on the actual autopsy protocol may not be permitted in some jurisdictions as it is hearsay evidence, but unless expressly forbidden it should be included, as it remains a record for the pathologist’s own files. It also justifies his eventual cause of death in those cases where the morphological findings are scanty or even absent, as his conclusions will be strongly influenced by his pre-knowledge of the mode of death. When the autopsy report is converted to a statement or deposition for legal use, this history may be omitted by those legal authorities responsible for transcribing the document.
- External examination.
- Internal examination.
- A list of specimens and samples retained for further examination. Those handed to other agencies, such as the forensic science laboratory, should be formally identified by means of serial numbers and the name of the person to whom they were handed.
- The results of further examinations such as histology, microbiology, toxicology and serology. When the main report is issued soon after the autopsy, these will not yet be available and a supplementary report will be necessary.
- A summary of the lesions displayed by the autopsy (often coded for departmental computer retrieval).
- Discussion of the findings, if necessary in the light of the known history.
- An opinion as to the definite or most likely sequence of events leading to the death.
- A formal cause of death, in the format recommended by the World Health Organization, suitable for the completion of a death certificate.
- The signature of the pathologist.

The ‘external examination’ should record those details described earlier in the chapter, the major items being:

- The height, weight and apparent state of nutrition.
- The presence of natural disease such as oedema, abdominal swelling, cutaneous disease, senile changes, etc.
- Identifying features such as skin colour, tattoos, scars, congenital or acquired deformities, dentures, eye colour and hair colour. When identity is an issue, naturally this section will be greatly expanded.
- The presence of rigor, hypostasis, decomposition and abnormal skin coloration. Body and ambient temperature should be recorded where appropriate, with calculations concerning the estimated range of times since death, though this aspect may well be deferred until the final ‘Summary and Conclusions’.
- The condition of the eyes, including petechiae, arcus senilis, pupil size, and the condition of iris and lens.
- Condition of mouth and lips, including injuries, teeth and presence of foreign material.
- Condition of external genitals and anus.
- Listing and description of all external injuries, recent and old.

The internal examination records all abnormalities, usually in a conventional sequence such as:

- Cardiovascular system: heart weight, any dilatation, ventricular preponderance, congenital defects, the pericardium, epicardium, endocardium, valves, coronary arteries, myocardium, aorta, other great vessels and peripheral vessels.

- Respiratory system: external nares, glottis, larynx, trachea, bronchi, pleural cavities, pleura, lungs (including weight) and pulmonary arteries.
- Gastrointestinal system: mouth, pharynx, oesophagus, peritoneal cavity, omentum, stomach, duodenum, small and large intestine, liver (weight), pancreas, gall bladder and rectum.
- Endocrine system: pituitary, thyroid, thymus and adrenals.
- Reticuloendothelial system: spleen (weight) and lymph nodes.
- Genitourinary system: kidneys (weight), ureters, bladder, prostate, uterus, ovaries and testes.
- Musculoskeletal system: skull, spine, remaining skeleton and musculature where necessary.
- Central nervous system: scalp, skull, meninges, cerebral vessels, brain (weight), middle ears, venous sinuses and spinal cord (when examined).

The timing of the report

As with the format, there are two schools of practice in this respect. One advocates the issue of as full a report as possible on the gross findings as soon as the autopsy is completed, usually within a day or two. Obviously this can only be a preliminary, provisional report, as it may have to be modified (sometimes radically) by the results of ancillary investigations that may take days or weeks to return. Virological cultures, for instance, may take up to 6 weeks before a growth can be reported.

In a large proportion of forensic cases, however, especially those due to violence, the gross findings are unlikely to be substantially amended by ancillary investigations, though the possibility must always be left open until data have been gathered in.

The other philosophy will delay any report (except perhaps a provisional oral opinion) until everything is to hand, when a single final document is provided.

Whichever course is adopted, one aspect is vital to both. The descriptive facts must be recorded at or immediately after the completion of the autopsy. It is vital that no significant interval – certainly no more than few hours – be allowed between the physical performance of the examination and the setting down of the objective findings. The words ‘setting down’ are chosen carefully as the report may be handwritten or typed, or may be dictated to a secretary or into a tape recorder or other audio system. What is vital is that it is not consigned to the pathologists’ memory for a few days, even reinforced by notes written on scraps of paper.

In Britain and many other jurisdictions, the court may demand to see any contemporaneous notes and even tapes from a tape recorder. Any diagrams, notes or rough drafts must be preserved for production on demand of the judge, coroner or advocates in court. Where the report is dictated (either to a secretary or into a tape recorder), then the first typed (or word-processed) draft constitutes the ‘original report’, along with any contemporaneous notes including body sketches.

The use of printed body sketches can be a very useful aid in the autopsy room. Many versions exist from the simple front and back views of the whole body used in clinical neurology to multiple sketches portraying every possible view of the body surface. Separate diagrams are available for the male and female perineum, and for the different body proportions of infants. Used on a clipboard these diagrams can be most useful, especially where there are multiple injuries, or large areas of burns or abrasions. Each lesion can be drawn in, with a measurement noted alongside each and distances from anatomical landmarks recorded. The data from the sketches can be transposed to written form at the end of the examination.

Discussion and conclusions in an autopsy report

Some pathologists, usually those not normally concerned with criminal and litigious cases, claim that an autopsy report should be a bare recitation of the physical findings, with no discussion or interpretation of the significance of those findings. In the authors’ opinion, this is an abdication of the pathologist’s responsibility because, especially in criminal deaths, it is these conclusions that are of most interest and use to the investigating officers, lawyers and courts.

After the detailed description of the external and internal appearances, a short resumé should be offered of the major positive findings and their relationship to the cause of death. In many cases this will be obvious, as in a gunshot wound of the head. Matters such as the probable type of weapon, the range, the direction and the likely rapidity of death, however, should also be discussed.

When the findings are less clear cut, or are multiple, then the alternatives should be discussed, giving a differential diagnosis of the cause of death and detailing the possible sequence of events. If it is possible, a ranking order of probability of the various alternatives can be offered. Time of death and the limitations of accuracy in this particular case should be set out when the issue is relevant to the investigation. What is really required is as full an interpretation as possible, without venturing into the undesirable fields of unwarranted speculation or Sherlock Holmes style of over-interpretation, which was the bane of forensic pathology in former years and is still practised too much even today, to the detriment of the good reputation of the speciality.

POST-MORTEM ARTEFACTS

Forensic pathology can only be learned by experience but no account would be complete without drawing attention to common artefacts found at autopsy, which can mislead the pathologist with insufficient forensic experience and even lead to a miscarriage of justice.

Some of the 'classic' mistakes were described many years ago by Shapiro (1954) and Moritz (1942), but each generation of pathologists discovers them a new – or what is worse, fails to discover them. Most artefacts are described in the various chapters dealing with specific lesions, but a reminder of some of the most important is given here:

- The pancreas is one of the first organs to undergo autolysis, because of the proteolytic enzymes within it. The autolysed tissue is often haemorrhagic and can easily be mistaken for acute pancreatitis, though histology will rapidly resolve the problem.
- Patches of haemorrhage, sometimes quite large and confluent, can occur in the tissues behind the oesophagus in the neck. These lie on the anterior surface of the cervical vertebrae and are caused by distension and leakage from the venous plexuses that lie in this area. They were described well by Prinsloo and Gordon (1951) and are sometimes known by this name. Their importance lies in confusion with deep neck bleeding in strangulation (and sometimes with spurious neck fractures), which is why the skull should be opened before the neck in any suspected strangulation or hanging, to release the pressure in the neck veins before handling the tissues.
- Autolytic rupture of the stomach can occur post-mortem in both child and adult, described by John Hunter in the eighteenth century. This so-called 'gastromalacia' appears as a slimy brownish black disintegration of the fundus with release of the stomach contents into the peritoneal cavity. Sometimes, the left leaf of the diaphragm is also perforated through a ragged fenestration, with escape of gastric contents into the chest.
- Heat fractures of the bones, either skull plates or long bones, may be seen in victims of severe fires, but are not evidence of ante-mortem violence. Also in conflagrations, the 'heat haematoma' within the burned skull can resemble an extradural haemorrhage of ante-mortem origin. The site is often at the vertex or occiput; however, unlike the usual parietal haemorrhage, there is no fracture line crossing the middle meningeal artery, the usual cause of a true extradural bleed. The frothy brown appearance of the false clot, together with heating effects in the adjacent brain, should indicate the true diagnosis. Shrinkage of the dura due to heat may cause it to split, with herniation of the brain tissue into the extradural space. Severe burns of the body surface may lead to heat contractures of the limbs with tears over joints such as the elbow. These must not be confused with ante-mortem lacerations or incised wounds.
- The bloating, discoloration and blistering of a putrefying body must not be misinterpreted as disease on injury. Blisters are quite unlike those of burns and dark blackish areas of discoloration must be distinguished from bruising. The latter can be difficult and incision into the tissues to seek blood in the dermis is advised. Histological sections may help but, where a body is badly decomposed, special stains for blood traces in the histological sections may assist, such as alpha-glycophorin to detect red cell envelopes. However, often it is quite impossible to differentiate the discoloration of decomposition from true bruising.
- Blood or bloody fluid issuing from the mouth may be due to putrefaction, even if the body surface is not overtly decomposed. If the lungs and air passages are discoloured and filled with sanguineous liquid, then this must be taken to be cause of the purging from the mouth and nostrils.
- Dark red discoloration of the posterior part of the myocardium is usually due to post-mortem gravitational hypostasis, not early infarction. Similarly, segmental patches of dark red or purple discoloration of the intestine is hypostasis, not infarction. The latter tends to be a single continuous length, the serosa being dull and the gut wall friable.
- Large petechiae or ecchymoses, sometimes with raised blood blisters, are often seen in the dependent skin of persons who have died a congestive death or where the upper part of the body has been hanging down after death. The usual place where these are seen is over the front of the upper chest and across the back of the shoulders, though in dependent heads the face may be shot with haemorrhages.
- Resuscitation artefacts are of increasing importance to the forensic pathologist and are discussed later.

EXHUMATION

Exhumation is the retrieval of a previously buried body for post-mortem examination. This is usually followed by a first autopsy or a re-autopsy following new information. The term 'exhumation' is usually applied to the removal of a body buried in a legitimate fashion in a cemetery or graveyard ('inhumation'), rather than the recovery of an uncoffined, clandestinely buried victim of a suspicious death. The latter is really a true 'scene of crime' and the pathologist should treat it as such.

Exhumations are required for one of the following reasons:

- Where all or part of a graveyard has to be moved for some development of the ground. Often no special examination of each body is made unless there is some historical or anthropological interest.
- Where some civil legal matter needs to be investigated, such as personal injuries for insurance or civil litigation for negligence – usually after a road, industrial or other accident.
- Where new information or substantiated allegations arise to suggest that a death was due to criminal action, either from injury or poison.
- In ancient or historical circumstances to investigate either the individual or a series of individuals for academic interest. A number of such investigations have been carried out on medieval and later inhumations in England, to study disease patterns and nutritional states in old populations – though many of these have been from dry vaults, rather than earth burials.

The legal procedures authorizing an exhumation do not concern us here, as they vary greatly from country to country. In all jurisdictions, however, there must be strict safeguards to identify the grave and the coffin, so that no mistake can be made. The grave must be positively identified by the cemetery authorities by reference to plans and records: an official must personally point out the grave to be opened.

Traditionally, exhumations are carried out at dawn but, except to avoid spectators and publicity, there is no real need to stumble about in a dark cemetery before first light. A better plan is for the grave to be dug down to just above coffin level by a mechanical digger or workmen on the previous day so that, the following morning, the police, coroner, pathologist and others can arrive in time to see the final exposure of the coffin. The coffin nameplate must be cleaned and read to confirm the identity, and, if possible, the funeral director who carried out the original burial should be present to identify the coffin and the plate.

If there is a suspicion of poisoning, samples of earth should have been taken from the surface of the grave, from other parts of the cemetery and from immediately above the coffin. When the coffin has been removed, further samples should be obtained from the sides and beneath the coffin, but these matters will usually be attended to by forensic scientists.

When the coffin is lifted to the graveside it is as well for the lid to be loosened a little by slackening the holding screws or prising the lid loose. This allows foul gases to escape into the open air, rather than in the mortuary. The coffin is then transported to the mortuary; if it is in a bad state of decay, it may have to be supported on a rigid base such as a trestle, or placed inside an extra large coffin or fibreglass shell. Excess earth and mud should be removed before transportation to avoid excessive fouling of the autopsy room. Where criminal action is suspected or has been alleged, a full photographic record must be kept (usually by the police) of every stage from identification of the grave to the findings during the autopsy.

At the mortuary, the coffin lid should be completely removed and the contents again identified, if possible by the funeral director who originally buried the body. He can again confirm the coffin plate, but also identify the internal coffin fittings, such as fabrics and shroud. When the body has not been buried for too long, he may be able to identify the features of the corpse from personal knowledge.

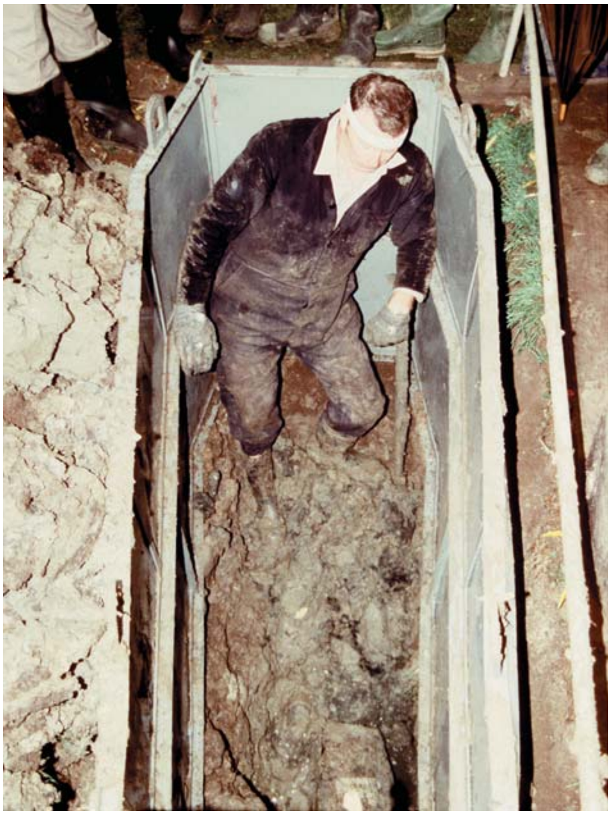

FIGURE 1.23 *Exhumation: a police officer making the final approach to the coffin. A metal cassion is lowered to protect the gravesides. The grave must be identified by the cemetery superintendent.*

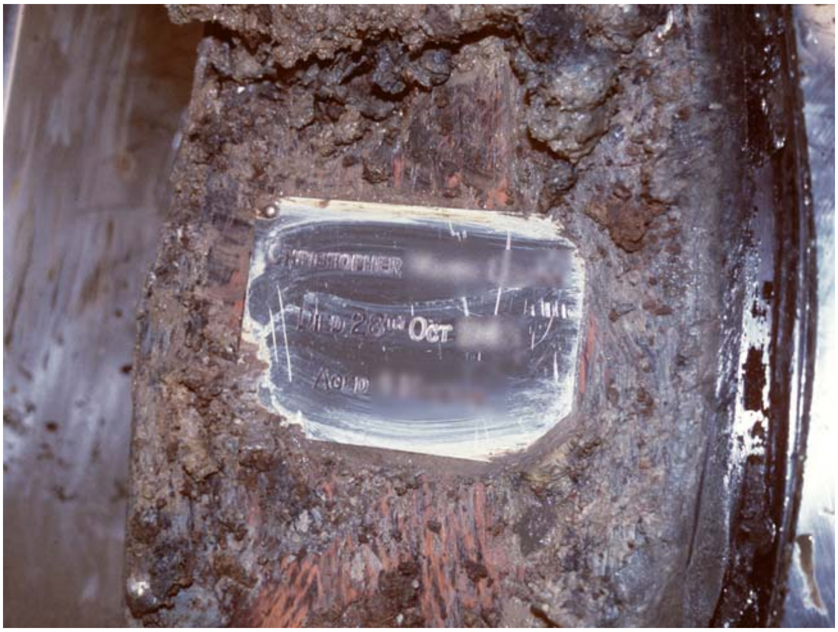

FIGURE 1.24 *Exhumation: when the coffin is exposed it must be identified by the funeral director and the coffin plate checked. If death from poisoning is suspected, samples of earth from above, below and the sides of the coffin as well as from a distant part of the cemetery must be taken, together with a sample of any grave water.*

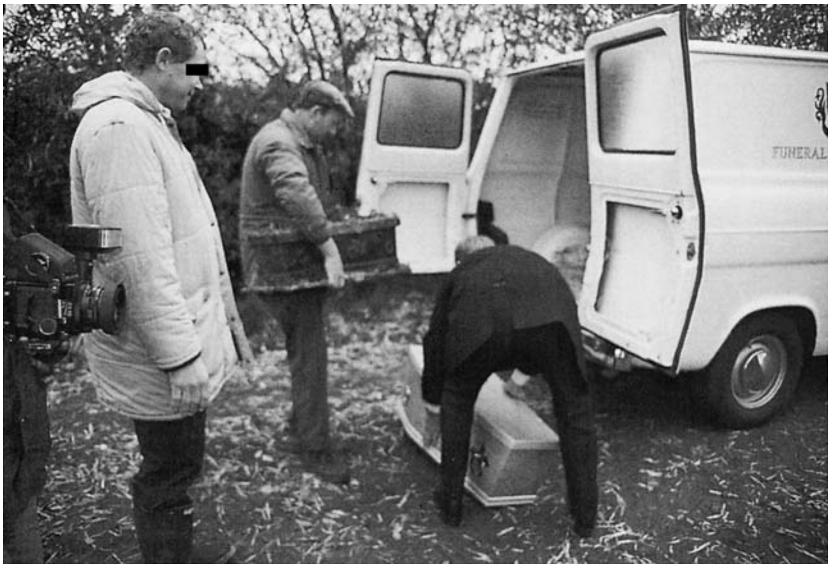

FIGURE 1.25 *Following exhumation, the body is transported to a mortuary for full autopsy. The former practice of autopsy in inadequate premises near the cemetery is no longer justified. Here the original rotted infant coffin is placed within a new coffin for transit.*

If poisoning is suspected, samples of the shroud, coffin trimming and any loose material such as packing or fluid should be retained for analysis. The body is then removed, undressed and a full autopsy carried out, as far as the condition of the body allows. Putrefaction, adipocere and mummification complicate the examination – sometimes, all three may be present in the same body.

The pathologist is sometimes asked by the authorities or by lawyers whether a proposed exhumation is worth carrying out, because of doubts about the usefulness of the result. Certainly the balance between the potential advantages must be weighed against the cost, publicity and distress to relatives that might be caused. In general, however, it is surprising how much information may be gained even when the body has been buried for many months or even a few years. Much will depend on the actual environment of the grave: a gravel or sandy soil, especially in an elevated position, will allow a body to remain in a much better state of preservation than one in the waterlogged loam of a valley. The author (BK) has seen burials only 20 years old with empty coffins containing silt, but no soft tissue or even bones (due to a constantly rising and falling water table) in an area of acidic peat.

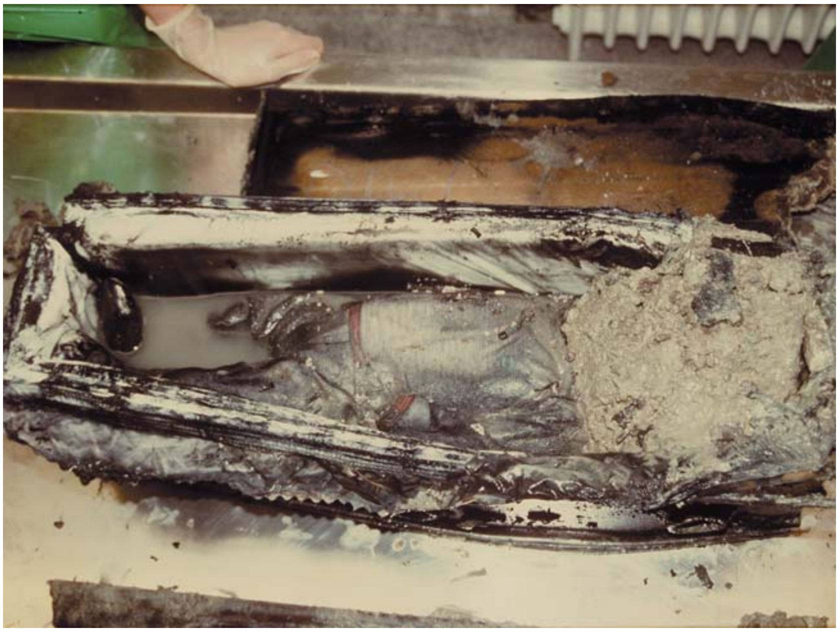

FIGURE 1.26 *When the coffin is first opened in the mortuary, the coffin fitting and fabrics should again be identified by the funeral director, as well as the body and its clothes, if they can be recollected. This infant has been buried for 3 years. The upper part of the coffin has leaked, allowing liquid mud to cover and substantially destroy the head.*

Even negative information gained at exhumation, such as the absence of alleged or suspected fractures, may be of considerable legal value. Some poisons, especially heavy metals, may persist for many years in a buried body and be detectable at exhumation. Even some organic chemicals may survive for a long time; barbiturates have been found after 7 years in a buried body. In all cases of suspected poisoning, it is vital that ample control samples are taken from the grave and its surroundings, to avoid the later accusation that any abnormal substances found were environmental artefacts or were unassociated with the corpse.

THE AUTOPSY ON THE PUTREFIED CORPSE

In forensic work, decomposed bodies are commonplace, especially in warm climates. Though the value of an autopsy is progressively reduced as the state of putrefaction advances, no short cuts should be taken by the pathologist merely because of the unpleasant nature of the examination. However bad the condition of the corpse, every effort should be made to carry out the autopsy as near to the usual routine as possible. It is often surprising – as with an exhumation – how much information can be gained. The interior of the body is often far better preserved than the outward appearances would suggest, so a policy of defeatism that leads to a skimped examination should never be adopted.

Externally, putrefaction hides bruising to a variable degree, the greenish black coloration of the skin masking the usual features of contusion. Abrasions, lacerations, incised wounds and gunshot wounds may, however, survive severe degrees of decomposition. Loss of bloody fluid from the mouth and nostrils (the so-called 'purging') is often mistaken by public, police and even some doctors as evidence of haemorrhage, but loss of serous, bloody or frothy fluid from any body orifice is common in the advanced stages of putrefaction.

Peeling and slippage of skin may hide some abrasions, though they may be visible when the desquamated epidermis is removed and the underlying skin examined. Marks around the neck from the tissues swelling with gas and becoming embedded in a collar have been mistaken for strangulation.

Where maggot or other insect infestation is present, some may be taken for expert entomological examination to help in assessing the post-mortem interval, as described in Chapter 2. As stated above, the external examination should be conducted as near as possible to the routine for a fresh body, and the back and perineum not neglected because of physical difficulties in handling the body.

Identity may be a problem when the facial features are too bloated for visual recognition. The usual procedures described in the next chapter may be employed where necessary. Fingerprints may be required by the police to assist in identity, but decomposition may seriously destroy the fingertips. These may become swollen and desquamate, or may shrivel and become leathery. Several methods of restoring the amputated tips have been described: some recommend a simple method of immersing the tips in 20 per cent acetic acid for 28–48 hours, when the shrivelling will swell to normal size. Others recommend immersion in glycerine.

Internally, much will depend on the state of decay. The abdominal and chest organs may be in a better state of preservation than the exterior. The subcutaneous tissues may be distended and crepitant with gas, as may be the swollen abdomen. Judicious penetration of the peritoneum with the tip of knife may be needed to release the pressure of gas. The technique of lighting the escaping methane with a burning newspaper may be spectacular and sometimes near explosive, but does little to reduce the smell and is not to be recommended.

Examination of the organs follows the usual pattern, modified according to state of putrefaction. The heart may be limp and discoloured, with haemolysis staining of the endocardium and vessels. The coronary arteries are often very well preserved, especially if atheromatous or calcified, or both. Ante-mortem thrombi may persist even after the muscle is semiglutinous. The larynx may be discoloured, but the hyoid and thyroid horns can be examined for fractures and may need to be X-rayed. It may be difficult to detect ante-mortem bleeding at the fracture sites. Fractures elsewhere in the skeleton naturally persist and may require radiography for their detection, as may foreign objects such as bullets.

The brain often decomposes early, and all too often is merely a pinkish-grey paste within the dura. Gross lesions like a large meningeal or intracranial haemorrhage may survive for examination, but the trauma of removing the calvarium and attached dura may seriously damage a semi-fluid brain. In Belgium, where exhumations are common due to the very low primary autopsy rate, a technique was developed in the University of Gent whereby the head of a decomposed body was removed from the body and deep frozen until solid. The head was then cut across in the coronal plane with a band-saw, leaving the brain in two halves within the cranium. These were then immersed in a large volume of formalin until fixed, when they could be removed in a relatively solid state for examination.

Internally, the detection of true subcutaneous bruising may be very difficult due to the discoloration of putrefaction. Histology is often highly unsatisfactory because of cellular lysis and degeneration; stains for haemoglobin may help in revealing focal collections of lysed blood, but are often diffusely positive even in control areas. Claims have been made for staining for glycophorin A to detect red cell envelopes, as opposed to diffuse haemolysis (Kibayashi *et al.* 1993).

RESUSCITATION ARTEFACTS AT AUTOPSY

In recent years, the advent of effective, but often aggressive and invasive, resuscitation procedures has made the task of the pathologist more difficult. At autopsy, injuries and abnormalities are now often found that could be due to terminal and even post-mortem resuscitatory measures. Where the pathologist is made aware of these, he can often exclude the damage being due to non-resuscitatory injury, though even when he is informed of the procedures they can still mask or mimic more sinister trauma. When he is not made aware – or such information cannot be made available – then the difficulties of interpretation are compounded.

There are now many publications describing resuscitation artefacts, including the review of Leadbeatter and Knight (1988). The following are the main categories of damage, for which the pathologist must always be alert:

- Bruising of the anterior chest wall, haemorrhage into the subcutaneous tissues and pectoral muscles, fractures of the sternum, fractures of the ribs, haemothorax, bruised lung, lacerated lung, pericardial haemorrhage, and even fractured dorsal spine, following energetic external cardiopulmonary resuscitation (CPR). Thoracic cage fractures are rare in children, however, because of the pliability of the ribs and costal cartilages, but it cannot be denied that they sometimes occur. The differentiation is naturally of vital interest where child abuse is being alleged.

  Internally all types of damage to the heart may occur, including ruptured atria and even ventricles, septal rupture and valve damage. The great vessels can suffer severe trauma, as described in the now extensive publications on resuscitation artefacts. Fat and bone marrow emboli in the pulmonary vessels have also been reported after cardiac massage (Mason 1993).

  Petechiae in the eyes and intra-ocular haemorrhages can occur after CPR, as well as after violent sneezing or coughing; they are well known to occur during whooping cough.
- Bruising of the face and neck, finger marks and nail marks on the face and neck, and damage to the lips and inner gums from mouth-to-mouth resuscitation, when the face and neck have been gripped by hands. Damage to lips, gums, teeth and pharynx can occur from the introduction of an artificial airway or endotracheal tube, especially in difficult, hurried emergency situations.

  Injuries to the larynx, even including fracture of the hyoid and thyroid cornuae, can occasionally occur from these procedures, which are difficult to distinguish from manual strangulation if the circumstances are obscure.
- Puncture marks for venepuncture may be confused with injection marks in drug dependence. The introduction of intravenous cannulae into veins in the neck may cause large haematomata and more diffuse bleeding into the tissues alongside the larynx.

Similar bruising may occur around puncture sites in the arms and groin. Intracardiac injections leave marks on the chest wall and may lead to a slight haemopericardium. The effects of injected noradrenaline and electrical defibrillation on the histological appearance of the myocardium are well recorded, with contraction bands being the most obvious artefacts which can be mistaken for pre-existing myocardial ischaemia.

- Damage to the mouth, palate, pharynx and larynx can occur from attempts to introduce a laryngoscope or airway. Even fracture of the mandible has been caused in this way. In infants, even digital clearance of the pharynx can cause mucosal damage. Damage to the pharyngeal mucosa may cause bleeding, which can seem sinister to police or relatives; this may be mixed with the fluid of pulmonary oedema to produce copious pink, bloody froth, seen in a number of cases, including sudden infant death syndrome.
- Electric defibrillator pads make marks on the chest, though these are usually easy to interpret, except where there is an unusual shape. Defibrillators and injected β-adrenergic catecholamines such as noradrenaline can, however, cause widespread histological damage to the myocardium, consisting of coagulation necrosis and contraction bands, which can be confused with infarction or electrocution. When both defibrillation and catecholamines have been used to resuscitate, the myocardial changes are even more marked (Karch 1987).
- During the Heimlich manoeuvre to clear an airway obstruction, rupture of the oesophagus, stomach and intestines have been reported. The oesophagus can be perforated by an incorrectly inserted airway. In the abdomen, external cardiac massage may cause ruptured stomach, ruptured liver, and damage to spleen and pancreas.
- Gastric contents in the air passages may have reached there by spontaneous agonal regurgitation or by pumping the chest and upper abdomen during resuscitation attempts. This makes the finding of vomit in the larynx and trachea of even less significance as a cause of death, as discussed in Chapter 14.
- The administration of oxygen by mask or tube may cause damage, as can overenergetic mouth-to-mouth resuscitation. Ruptures of oesophagus and lung have occurred, and other types of barotrauma include ruptured stomach and intestine. Where a pre-existing gut lesion exists, the administered gas may escape into the abdomen. The diagnosis of a pre-existing pneumothorax may be impossible where forced ventilation has been administered.
- In the central nervous system, subarachnoid haemorrhage has been described after external cardiac massage, and from hyperextension of the neck to pass an airway or perform mouth-to-mouth resuscitation. The forced posture can tear the vertebral arteries, making another source of confusion for the forensic pathologist in this difficult area of potential trauma to the neck (Chapter 5).
- Myocardial and pulmonary bone marrow embolism has been reported following cardiac massage (Dziecol *et al.* 1992).
- Retinal haemorrhages, classically a sign of raised intracranial pressure and of head injury, have also been described in whooping cough and after CPR.
- Miscellaneous artefacts of which every pathologist should be aware include: skin maceration from a body lying in urine, kerosene, etc., post-mortem burns from adjacent radiant heat or hot-water bottles; and petechiae and larger haemorrhages in the face from postural hypostasis.

MASS DISASTERS – THE ROLE OF THE PATHOLOGIST

With the exception of the more dramatic murders, the activity which focuses most public attention on the work of forensic pathologists is the mass disaster. Unfortunately, such tragedies are becoming more common with the increase in terrorism, the expansion of travel facilities and the larger size of passenger aircraft. Recent years have seen a tragically frequent need for the expertise of mass disaster teams, following such events as the Zeebrugge and Estonian ferry capsizes, football tragedies like Ibrox Park, Heysel, Moscow and Hillsborough, crush tragedies at Mecca, and numerous massive aircraft crashes, such as Tenerife, Air India, Japan, Lockerbie and most recently the terror attacks in the USA. Mass disaster management is a discipline in itself and only the briefest summary can be offered here, together with useful references at the end of the section.

Forward planning

Many pathologists will thankfully spend their entire careers without having to participate in a major mass disaster; a commonly accepted definition is the death of more than 12 victims in a single event. No one can predict when such a tragedy might occur, however, as typified by the 1988

Pan Am sabotage, when almost 300 bodies fell out of the sky on a quiet and unsuspecting Scottish village. It is thus essential that every forensic institute, department and individual pathologist should make some forward provision for such an eventuality. In Britain, the Royal College of Pathologists (1990) have published a useful booklet on the role of the pathologist in these disasters.

In most advanced countries each region now has a Mass Casualty Plan covering medical and hospital services, fire service and police. This is often very detailed, covering every aspect of transfusion, drugs, casualty transport, emergency surgery and anaesthesia. The plans are clinically orientated, but often completely ignore provision for the dead – or they have some cursory statement at the end such as ‘Mortuary accommodation will be provided’ with obviously no thought given to how several hundred corpses are to be accommodated and examined. Though clinical planning is, of course, vital – especially in rail disasters, multiple motorway crashes and urban bomb outrages – it has to be appreciated that, in many air crashes, there are few if any survivors, and that all the clinical planning may be redundant, leaving a massive and unprepared crisis in relation to the dead.

It is therefore essential that forensic pathologists should ensure that, in the area for which they are responsible, there is cooperative preplanning that includes adequate provision for collection, accommodation, examination and disposal of large numbers of dead victims. Naturally the pathologist is usually in no position to do this alone, but he is often the person with the most foresight and professional knowledge to act as the stimulus and catalyst between the major agencies responsible for overall planning. These are usually the police and the local health administration. When a Mass Casualty Plan does exist, but has serious omissions in respect of dealing with the dead, or where no such plan exists, the forensic pathologist should energetically stimulate the responsible authorities into making a comprehensive plan. This entails a series of meetings to identify danger areas in the region (such as airports, motorways, railways and military installations), and to discuss potential buildings for temporary mortuaries, the provision of materials such as markers, plastic bags and labels, which may be needed in large quantities at a few hours’ notice. Communications between such groups as police, pathologists, mortuary and laboratory technicians, radiographers, and dentists also need to be established in advance. The expertise of international funeral directors with experience in mass disaster fatalities is invaluable.

The objects of pathological investigation in mass disasters are:

- to retrieve and reconstruct bodies and fragmented bodies decently
- to establish personal identity
- to conduct autopsies on some or all of those bodies
- to establish the cause of death in some or all, especially air crew and drivers, and to assist in reconstructing the cause of the disaster
- to obtain material for toxicological analysis (especially alcohol and carbon monoxide) where appropriate
- to seek evidence of the cause of the disaster from autopsy examination, such as bomb or detonator fragments that may be embedded in the bodies.

Outline of necessities in mass disaster planning

PROVISION OF PATHOLOGISTS AND OTHER STAFF

Depending upon the scale of the tragedy, it may be necessary to recruit other pathologists to assist. In large cities, there may be sufficient persons with expert forensic knowledge, but elsewhere, hospital pathologists may be able to assist; forensic pathologists from a distance may also volunteer. It has to be remembered that a really large disaster may need many days or even weeks of work, and therefore personnel may be unavailable for the whole period. In addition, whatever the degree of willingness and unselfish devotion offered by doctors and all other staff, the physical and especially psychological stresses of this harrowing work mean that strictly limited periods of work should be imposed. Apart from the deleterious effect upon the doctor, the standard of work declines dramatically with fatigue and therefore it is in the interests of the investigation – as well as of the pathologists – that sufficient staff be recruited. This may, however, be easier said than done. One person must be in overall charge of the investigation. Usually a senior police officer has the ultimate responsibility, but the medical aspects should be firmly under the control of a senior pathologist, though he must delegate extensively to avoid being swamped by less important tasks and thereby being rendered totally inefficient. Again, all these aspects are modified by the scale of the disaster. Proper facilities for meals, rest and washing must be established, and again forward planning is essential for these mundane, but vital, features. Forensic dental and radiological expertise will be run independently by their own specialists, but there must still be a nominal chief in the form of the senior pathologist who acts as overall coordinator and arbiter of medical matters.

When a disaster, almost always an aircraft crash, occurs in some remote place or abroad where there are no satisfactory forensic or pathological services, it is the usual practice for a team from the country of origin of the aircraft – or a volunteer team arranged at governmental level – to fly to the scene and provide expertise to the local authorities.

In many advanced countries, especially Britain, the armed services have permanent aviation pathologists who deal with all service aircraft crashes anywhere in the world and who are often available to deal with, or attend as advisers, any civil air disaster.

PROVISION OF MORTUARY FACILITIES

Most hospital mortuaries and even larger public mortuaries have limited storage capacity for bodies, and even smaller provision for performing simultaneous autopsies. It must also be appreciated that the normal business of a mortuary has to continue even through a mass disaster, so the massive increase in work is an addition to, not a substitution for, the normal handling facilities. Where more than about ten bodies are involved, most mortuaries cannot cope with these problems, especially when it has to be remembered that some of the corpses may have to be retained for considerably longer periods than usual if identification cannot be made rapidly. Thus external facilities will have to be provided in some temporary accommodation. When the disaster is a long distance from the regular mortuary, transport and other logistic considerations may make it imperative to store and examine the bodies nearer to the crash site. Again forward planning is essential to identify hangars, storehouses, empty factories, halls and other buildings near potential danger spots such as airports. Whenever possible, all bodies should be taken to one site, as the separation of identification sites is a recipe for inconvenience, delay and mistakes. It is sometimes necessary, especially in remote regions, to set up tented mortuaries, but this is a last resort as facilities cannot be laid on with the same degree of efficiency. Where the crash is totally remote, it is usual for the authorities or the armed services to bring out the dead by helicopter or other transport to an urban site. A prime example was the Mount Erebus crash in Antarctica, where the dead were flown back to New Zealand. If a warehouse, or factory or building of similar size, can be used, certain minimum facilities are needed. Good electric lighting, portable lights for close inspection, and power points for radiographic equipment and electric instruments are required. Adequate piped water, and washing and toilet facilities are essential. If any of these is deficient, then portable generators and water tankers must be supplied by the armed services or the police. Telephone and, if possible, telex and fax facilities should be available for the input of identifying data.

In hot climates, or in the summer in temperate climates, body refrigeration is vital, not only for the decent preservation of the dead but for retention of tissues awaiting identification. If a disaster is too large to be handled in the usual mortuary, then some form of cooling is required. The renting of refrigerated trucks used for the transport of foodstuffs is the usual answer and, again, preplanning is necessary to discover sources of such facilities. Sometimes portable air-conditioning units can be installed in part of the mortuary, if this is sufficient to cool the storage area to an acceptable level. In disasters where there are both living and dead victims, the mortuary should be sited away from the clinical facilities or screened in some way to avoid the distressing sight of bodies arriving being visible to survivors or their relatives, and the press.

The temporary mortuary should be large enough for the expected load, and the area to be used for examinations and autopsies should not be too congested. No one should be allowed in who does not have direct business there, no matter of what eminence or rank. Security of admission should be tight, this being the responsibility of the police. Flooring should be waterproof and capable of being hosed down. It can be protected against blood, mud and burned fragments by covering with polythene from large rolls. Tables for examination can be wooden trestles, covered with polythene. These should be in rows one metre apart, with 2 metres between rows.

Retrieval of bodies

This is the task of the police or armed services, but must be carried out in a manner approved by the forensic identifying team. It is essential that every body is first certified as dead by a doctor at the scene. Often, volunteer casualty surgeons may be the first medical persons at the scene, whose primary duty is to rescue living survivors and to confirm death in the remainder.

Each body or fragment should be flagged with a sequential and unrepeatable serial number and marked on a grid plan, being photographed *in situ* wherever possible. It is then bagged and taken with its numbered label to the mortuary. Different teams will have different methods of dealing with the logistics of handling data and, increasingly, this is being performed on either microcomputers or on terminals linked to a central computer. The police have responsibility for these aspects, as they collect and record the clothing and personal belongings that play such an important part in personal identification.

Either the same serial number is used from the recovery stage or a different 'pathology' number is begun in the mortuary, which must be matched up with the previous serial number. Everything that has come from that body, including clothing, wallets, rings, teeth and jewellery, must carry the same number.

The pathologist and lay assistants undress the clothed bodies, and dictate a description of what is found as accompanying artefacts. The clothes are stored in strong paper

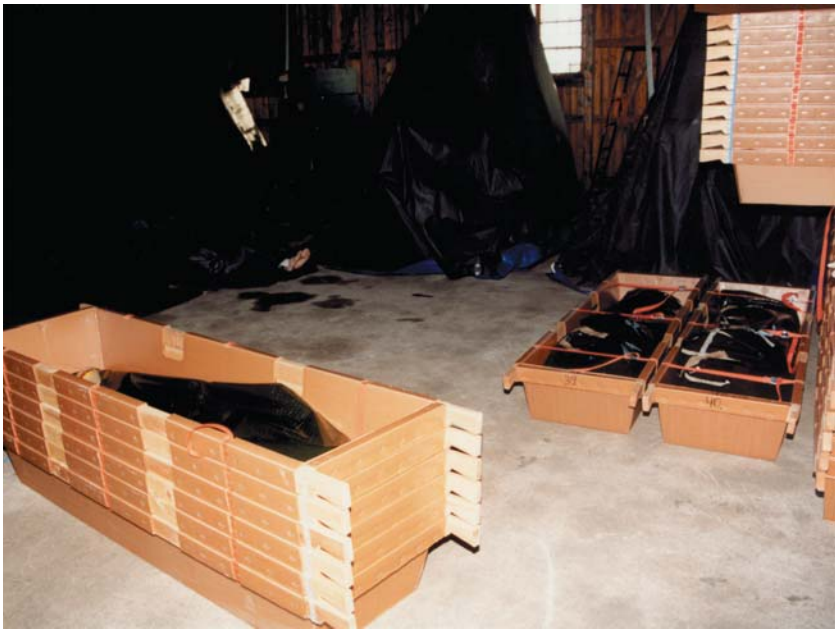

FIGURE 1.27 *Temporary coffins used to transport victims of the m/s Estonia disaster from the archipelago to the autopsy facility.*

bags to avoid the fungal growth that inevitably appears on damp fabric kept in plastic bags.

Many items will have been found loose at the site, but they must all be tagged and an attempt made to relate them to a nearby body, though this is obviously fraught with error. The objects found on a body are recorded on an inventory form and a separate form (usually the Interpol document) used for the purely medical and anatopathological aspects. The latter data are acquired by meticulous and systematic external examination, followed by an internal autopsy if this is to be carried out.

Whether or not an autopsy is to be performed on some or all of the victims will depend upon facilities, availability of pathologists, the legal direction and the system prevailing in the particular country. The wishes of the legal authority, such as coroner, judge, medical examiner or police, will be the deciding factor in determining how many victims are subjected to autopsy. The pathologist should exert his influence where there is reluctance to sanction any autopsies, emphasizing the benefit to the crash investigators of autopsy data on key victims. Pathology findings may often be of use – it has even been claimed that victims can be separated into smokers and non-smokers by finding 'smoker cells' in the lungs (Reiter and Risser 1994).

As mentioned, the aircrew or other persons in charge of a vehicle or train should always have a full examination and analysis of body fluids as part of the accident investigation. Full photography of the clothing and bodies should be taken, and then all physical features such as height, weight, gender, race, colouring, scars, tattoos and deformities must be recorded. The forensic dental examination is then carried out on those bodies where obvious identity cannot be established by non-medical or dental means.

Radiological examination will almost always be required to assist in identification from osteological features and dental aspects – and perhaps also to seek foreign bodies that may assist in accident reconstruction, such as metal fragments blown upwards into the thighs and buttock in a hold explosion in an aircraft bomb – or even parts of the bomb or detonator itself, more especially in land-based terrorist attacks.

Toxicology should be taken as extensively as possible, even from some of those bodies that are not to have an autopsy. The data acquired from this painstaking work is then forwarded to the police bureau who have been collecting personal data from the records and relatives of the victims, and efforts are made to match the two sets of data. This is increasingly being performed by computer, and many efforts have been made by international agencies such as Interpol to set up universally compatible systems so that data can be acquired and exchanged electronically via telephone systems and modems from any part of the world.

In 1990, the Royal College of Pathologists published a definitive guide for pathologists to the problem of mass disasters, which is strongly recommended as a standard set of procedures.

THE OBSCURE AUTOPSY

Several surveys in various countries have shown that where a physician offers a cause of death without the benefit of autopsy findings, the error rate is of the order of 25–50 per cent, even in deaths in hospital. Thus the value of an autopsy in improving the value of death certificates is undoubted, but it still has to be conceded that the autopsy is by no means infallible in revealing the true cause of death. Estimates of the frequency with which autopsies fail to deliver an adequate or even any cause of death vary from pathologist to pathologist, as well as with definitions of what type of case a is included in such series. For example, the sudden infant death syndrome (SIDS) could be considered to provide 'negative autopsies', in that by definition, no significant findings are discovered.

Even excluding SIDS, there is probably at least a 5 per cent 'failure rate' in autopsy series from large medical centres and forensic pathology departments. This rate will also vary according to the habits, personality and seniority of the pathologists involved. Contrary to what might be expected, a higher rate of negative conclusions will originate from older and more experienced pathologists than from juniors. The younger pathologist is often uneasy about failing to provide a cause of death, feeling that it reflects upon his ability, whereas the more grizzled doctor, enjoying the security of tenure and equality with – or even seniority over – his clinical and legal colleagues, is less inhibited in his admissions of ignorance when the cause of death remains obscure.

These 'obscure autopsies' are more common in the younger age group. Even apart from SIDS, many autopsies on infants, especially neonates, are less than satisfactory in terms of definite morphological lesions discoverable in the autopsy room or the laboratory. The deaths often have a biochemical or hypoxic basis and, though inferences can be made from the clinical history, even expert paediatric pathologists have to make do with minimal or even absent findings in a significant proportion of cases. In teenagers and young adults, up to the age of about 35 years, there is a higher proportion of negative autopsies than in the older group, which provides the vast majority of the autopsy workload. An example is the obscure syndrome seen in East Asia, where Thai construction workers in Singapore suddenly die with no demonstrable pathological cause, similar cases occurring in China, Japan and Hong Kong.

This preponderance in the young may carry an inherent fallacy, even though there may be an absolute increase in the incidence of obscure fatal causes. The young adult does not have the almost universal overlay of degenerative cardiovascular disease that is seen in the older group. Thus it is probable that the same occult disease processes occur in that older group as in the young, but as the latter do not have other lesions that can be grasped (usually quite legitimately) as a valid cause of death, then nothing presents itself to the pathologist. For example, on one autopsy table might be the body of a man of 22 who dropped dead shortly after taking part in a football match. There may be no adverse medical history, the gross examination revealed nothing abnormal and subsequently full histology with special stains, full toxicological screening, and microbiological and virological studies, were unrewarding. No cause of death can be extracted from these negative findings and the case must be recorded as 'unascertained' – or, as Professor Alan Usher of Sheffield points out, as 'unascertainable', if the pathologist is feeling particularly omnipotent!

On the next table in the autopsy room, however, may be a man of 60 who has been found dead with no available history. On examination he is found to have 60 per cent stenosis of the anterior descending branch of his left coronary artery, but no recent or old damage to the myocardium. Histology may well be performed, with no further information obtained. The cause of death is likely to be recorded as 'coronary artery disease' by most pathologists, yet he might well have died from the same obscure cause as the young man, but because he has just sufficient arterial degenerative disease to prove fatal, the latter is accepted as the most likely cause of death. This feeling of disillusion may be reinforced by the presence on the third autopsy table of another man of 60 who has been killed by a firearm, yet who has 80 per cent stenosis of all three coronary vessels, this obviously having played no part in the death.

Whatever the philosophical aspects of this problem, the practical difficulty remains of the pathologist's course of action when faced with negative findings in an autopsy.

At the end of the gross examination in the autopsy room, as the depressing realization increases that nothing significant has been found, further action will depend partly upon the facilities available to the pathologist, especially in relation to toxicological, biochemical, microbiological, virological and histological laboratory service.

Before proceeding further to review the dissection, attention must be paid to obtaining adequate samples for ancillary investigations. In many cases, especially where the history alerts the experienced pathologist that this might be a difficult case – and usually always in young people – samples of blood, urine and stomach contents will have been taken as routine during the removal and first dissection of the viscera.

If a blood sample has not already been taken, it should now be obtained, preferably from a peripheral vein such as the axillary or femoral to avoid contamination from the now empty trunk cavities. If urine was not retained from the bladder, a few drops may still be found in that opened organ by suction with a syringe. Stomach contents may well have been lost, but the liver can be retained for toxicological analysis. It might be useful to take vitreous humour if urine and stomach contents have been lost.

Some blood may be used for inoculation of blood culture bottles and swabs can still be taken for microbiological investigations. If a lung infection is possible, then either a few grams of lung tissue can be removed into a sterile jar, or swabs can be taken from peripheral bronchi or from the lung parenchyma itself. As the surface of all cut organs will have been contaminated during the first dissection, cuts should be made into the surface of the lung using a new sterile scalpel so that relatively fresh tissue is obtained for culture. At the same time, a cube of lung – or other organ if desired – should be taken into a sterile container for virus studies.

Before tissue for histology is taken, a full review of the more vital parts of the dissection should be reviewed. Even in younger victims the first area to be studied again is the coronary system, and in older obscure deaths it is essential to re-examine these vessels. Although the coronaries should have been transected at short intervals (not more than 3 mm apart) on the first examination, it is prudent to retrace each vessel carefully from the aortic origin distally, looking again at each cut segment and, where the interval appears too large, to make further intermediate cuts. In addition, it may be wise to join the segments longitudinally between cuts, in case a small isolated thrombus is lurking in the lumen. This should be done for common left trunk, anterior descending branch, left circumflex and right coronary artery.

This review of the coronary system sometimes pays dividends, especially in the middle-aged group that at first sight appears to have good coronary vessels. A tiny segment may be occluded or severely stenosed over a distance of only 2 or 3 mm, either by pure atheroma or by a subintimal haemorrhage, ruptured plaque or a localized thrombosis. If nothing is found, then after a quick review of the valves and great vessels, the myocardium can be sliced more extensively and a number of representative blocks taken for histology, to seek a myocarditis or some obscure cardiomyopathy.

The other organs should be examined again, the pulmonary arteries being scrutinized for pulmonary emboli in the smaller branches. This is rarely productive, but the author (BK) has on several occasions found multiple small emboli in more peripheral lung vessels, although whether these were a valid cause of death was debatable. The brain should be looked at once more, with special attention being paid to the basal arteries. These should be opened with coronary scissors and the dissection carried up into the middle cerebral arteries where they lie in the Sylvian fissures. It is also always worthwhile looking in the carotid arteries in the neck, though these should always be opened routinely at the first dissection from the aorta to a point just distal to the carotid sinuses; even the difficult area above this at the base of the skull has provided some surprises in the past. It is a place rarely examined, except by neuropathologists, but forensic pathologists should add it to their 'repertoire', as occasionally total thrombotic occlusion may be found, especially following some undisclosed neck injury.

When a complete review of the gross pathology has proved fruitless, then a full histological survey is required, especially of the myocardium. Special stains such as phosphotungstic acid-haematoxylin, dehydrogenase enzyme histochemistry (which requires unfixed, frozen sections), acridine-orange fluorescence stains and any other technique that the pathologist favours. Admittedly, where no cardiomegaly or coronary artery stenosis is present, the chances of finding an abnormality are slight, but an isolated myocarditis is an outside chance.

Toxicology may be difficult and expensive if there is no cause to suspect any particular drug or poison. A screen for unknown substances can be time consuming for the laboratory, which means a large expense to public funds. An alcohol estimation and a screen for acidic and basic substances, however, though not comprehensive, can at least exclude most common poisons. Microbiology and virology rarely point to a previously unsuspected fatal disease process unless the patient had clinical signs and symptoms before death.

When all the results are available, which may take several weeks (especially in the case of virus cultures), the case must be reviewed and an honest opinion offered as to whether any positive findings are sufficient to have caused death. In the experience of the authors, these ancillary investigations more often provide no help, than the times when they offer some assistance, but they must be carried out whenever possible in order to exclude such causes and to prevent allegations that the death was not investigated as fully as it should have been.

If at the end of the process no cause of death is apparent, then the appropriate investigating authority must be informed that no opinion can be offered in the present state of medical and scientific knowledge. It can be added, however, that the absence of injuries, evidence of poisoning, lethal infection or well-recognized natural disease is in itself significant negative evidence in that it confirms what the deceased did not die of, and the assumption is then that the balance of probabilities is that it was natural causes, rather than some unnatural external event.

There must be complete honesty on the part of the pathologist and a readiness to admit that the cause of death cannot be determined. The use of some meaningless euphemism such as 'heart failure' or 'cardiorespiratory arrest' is pointless and merely confuses the issue for non-medical persons such as police or magistrates. As mentioned above, there is a tendency on the part of some younger pathologists or histopathologists who rarely become involved in medico-legal cases to assume some unwarranted disease process or to attribute significance to insignificant findings. A **mode** of death is useless in lieu of a **cause** of death, so is the dangerous practice of using some agonal event such as 'inhalation of vomit' where there is no eyewitness confirmation of such an event taking place during life. It is also useless and positively dangerous to guess at some totally unprovable process, such as 'vagal inhibition', 'reflex cardiac arrest' or 'suffocation', just because these conditions are thought to leave no traces and therefore a death with no traces must be due to such a cause. This is illogical thinking and has led to several miscarriages of justice in the past, as well as to far more family anguish.

As the pathologist increases in experience and maturity, he or she is more ready to concede that he cannot find a cause of death, and this is far more satisfactory. In forensic work, this is not 'abrogating responsibilities' but being objective, sensible and just. There is no point in producing a speculative cause of death if that cannot be substantiated in later court evidence or legal statements. It is, of course, quite justifiable and indeed, part of the pathologist's responsibilities, to discuss the range of possible causes with the authorities responsible for investigating the death, but to be dogmatic about a single cause where the grounds for such a decision are tenuous does not help anyone and can lead to unfortunate consequences.

REFERENCES AND FURTHER READING

Adelson L. 1953. Possible neurological mechanisms for sudden death with minimal anatomical findings. *J Forensic Med* **1**:39–42.

Adelson L. 1974. *The pathology of homicide*. Thomas, Springfield.

Annegers JF, Coan SP. 1999. Sudep: overview of definitions and review of incidence data. *Seizure* **8**:347–52.

Asnaes S, Paaske F. 1980. Uncertainty of determining mode of death in medico-legal material without autopsy – a systematic autopsy study. *Forensic Sci Int* **15**:3–17.

Ball J, Desselberger U, Whitwell H. 1991. Long-lasting viability of HIV after patient's death. [letter] *Lancet* **6**:63.

Bankowski MJ, Landay AL, Staes B, *et al.* 1992. Postmortem recovery of human immunodeficiency virus type 1 from plasma and mononuclear cells. Implications for occupational exposure. *Arch Pathol Lab Med* **116**:1124–7.

Boukis D. 1986. Repeat autopsies on corpses from abroad. A futile effort? *Am J Forensic Med Pathol* **7**:216–18.

Bowen DA. 1966. The role of radiology and the identification of foreign bodies at postmortem examination. *J Forensic Sci Soc* **6**:28–34.

Braden G. 1975. Aircraft type crash injury investigation of a commuter train collision. *Aviat Space Environ Med* **46**:1157–60.

Brown T, Delaney R, Robinson W. 1952. Medical identification in the noronic disaster. *JAMA* **148**:621.

Buck N, Devlin HB, Lunn JN. 1988. *The report of a confidential enquiry into perioperative deaths. Perioperative deaths*, vol 39. Nuffield Provincial Hospitals Trust, London.

Cameron JM, Grant JH, Ruddick R. 1973. Ultra-violet photography in forensic medicine. *J Forensic Photogr* **2**:9–14.

Cameron HM, McGoogan E, Clarke J, Wilson BA. 1977. Trends in hospital necropsy rates. *Br Med J* **18**:1577–80.

Cameron HM, McGoogan E, Watson H. 1980. Necropsy: a yardstick for clinical diagnoses. *Br Med J* **11**:985–8.

Campling EA, Devlin HB, Lunn JN. 1990. *Gynaecology and the national confidential enquiry into perioperative deaths*. Royal College of Surgeons, London, pp. 466–7.

Camps FE. 1953. *Medical and scientific investigations in the Christie case*. Medical Publications, London.

Camps FE. 1972. Exhumation. *Criminologist* **7**:736.

Cao Y, Ngai H, Gu G, Ho DD. 1993. Decay of HIV-1 infectivity in whole blood, plasma and peripheral blood mononuclear cells. *AIDS* **7**:596–7.

Carpenter R, Dorn J, Hopkins B, *et al.* 1986. *Multiple death disaster response*. National Funeral Directors Association, Milwaukee, WI.

Clark DH. 1986. Dental identification problems in the Abu Dhabi air accident. *Am J Forensic Med Pathol* **7**:317–21.

Clark MA. 1981. The value of the hospital autopsy. Is it worth the cost? *Am J Forensic Med Pathol* **2**:231–7.

Clark MA. 1981. Applications of clinical laboratory tests to the autopsy: a practical guide for specimen collection. *Am J Forensic Med Pathol* **2**:75–81.

Claydon SM. 1993. The high risk autopsy. Recognition and protection. *Am J Forensic Med Pathol* **14**:253–6.

Copeland AR. 1986. Accidental non-commercial aircraft fatalities: the 7-year Metro-Dade county experience from 1977–1983. *Forensic Sci Int* **31**:13–20.

Cordner SM. 1985. The role of the second post-mortem examination. *Med Leg J* **53**:24–8.

Corwell WS. 1956. Radiography and photography in problems of identification. *Med Radiogr Photogr* **32**:34–40.

Cotton DW, Stephenson TJ. 1988. Impairment of autopsy histology by organ washing – a myth. *Med Sci Law* **28**:319–23.

Couch S, Noguchi TT, Wright J. 1981. Demonstrative evidence developed at the autopsy. *Am J Forensic Med Pathol* **2**:135–8.

de Craemer D. 1994. Postmortem viability of human immunodeficiency – implications for the teaching of anatomy. [letter] *N Engl J Med* **331**:1315.

Dodds JP, Nardone A, Mercey DE, *et al.* 2000. Increase in high risk sexual behaviour among homosexual men, London 1996–8:cross sectional, questionnaire study. *Br Med J* **320**:1510–11.

Donoghoe MC, Stimson GV, Dolan KA. 1989. Sexual behaviour of injecting drug users and associated risks of HIV infection for non-injecting sexual partners. *AIDS Care* **1**:51–8.

Douceron H, Deforges L, Gherardi R, *et al.* 1993. Long-lasting postmortem viability of human immunodeficiency virus: a potential risk in forensic medicine practice. *Forensic Sci Int* **60**:61–6.

Drake W, Lukash L. 1978. Reconstruction of mutilated victims for identification. *J Forensic Sci* **23**:218–30.

Durigon H. 1988. *Pathologie medico-legale*. Masson, Paris.

Durigon M, Ceccaldi PF, Renard J. 1983. Catastrophes et identifications: role des differentes structures. [Role of foreign physicians in the identification of catastrophy victims.] *Med-Armees* **3**:2524.

Dzieciol J, Kemona A, Gorska M, *et al.* 1992. Widespread myocardial and pulmonary bone marrow embolism following cardiac massage. *Forensic Sci Int* **56**:195–9.

Eckert WG. 1980. Catastrophes et morts collectives. Recent american experiences in mass deaths. *Am J Forensic Med Pathol* **1**:77–9.

Eckert WG. 1982. Fatal commercial air transport crashes, 1924–1981. Review of history and information on fatal crashes. *Am J Forensic Med Pathol* **3**:49–56.

Eckert WG. 1983. The untapped resource; the forensic and general autopsy. [editorial] *Am J Forensic Med Pathol* **4**:293.

Eckert WG. 1985. The disinterred body: do it right the first time. *Am J Forensic Med Pathol* **6**:5–6.

Enbom M, Sheldon J, Lennette E, *et al.* 2000. Antibodies to human herpesvirus 8 latent and lytic antigens in blood donors and potential high-risk groups in Sweden: variable frequencies found in a multicenter serological study. *J Med Virol* **62**:498–504.

Eriksson A, Krantz KP, Lowenhielm CG, *et al.* 1984. A device which allows dictation of the autopsy report during the autopsy. *Am J Forensic Med Pathol* **5**:279–82.

Evans KT, Knight B. 1986. Forensic radiology (see Notes). *Br J Hosp Med* **36**:14–20.

Fairburn G, McDonald K. 2000. Increased high risk sexual behaviour in homosexual men. Clarity may have been lost through including too much information. *Br Med J* **321**:1532.

Fatteh A. 1978. Exhumation, autopsy, second autopsy. *Leg Med Annu* 151–63.

Fatteh AV, Mann GT. 1969. The role of radiology in forensic pathology. *Med Sci Law* **9**:27–30.

Federal Emergency Management Agency. 1987. *Medical and emergency preparedness for disaster management: preventive medicine and public health aspects*. NATO-JCMMG, Study #4. Washington, DC.

Gee DJ. 1975. Radiology in forensic pathology. *Radiography* **41**:109–14.

Geller SA. 1990. HIV and the autopsy. [editorial] *Am J Clin Pathol* **94**:487–9.

Geller SA. 1990. The autopsy in acquired immunodeficiency syndrome. How and why [see comments]. *Arch Pathol Lab Med* **114**:324–9.

Gerber SR, Adelson L, Johnson L. 1956. The role of photography in the coroners office. *Med Radiogr Photogr* **32**:15–18.

Ghanem I. 1988. Permission for performing an autopsy: the pitfalls under Islamic law. *Med Sci Law* **28**:241–2.

Gilliland MG, McDonough ET, Fossum RM, *et al.* 1986. Disaster planning for air crashes. A retrospective analysis of delta airlines flight 191. *Am J Forensic Med Pathol* **7**:308–16.

Goldberg D, Scott G, Weir M, *et al.* 1998. HIV infection among homosexual/bisexual males attending genitourinary clinics in scotland. *Sex Transm Infect* **74**:185–8.

Goldman L. 1984. Diagnostic advances v the value of the autopsy. 1912–1980. *Arch Pathol Lab Med* **108**:501–5.

Goldman L, Sayson R, Robbins S, *et al.* 1983. The value of the autopsy in three medical eras. *N Engl J Med* **308**:1000–5.

Gonzales M, Halpern MH, Umberger CJ, Vance M. 1954. *Legal medicine and toxicology*, 2nd edn, vol. 64. Appleton Century Crofts, New York.

Gopinath C, Thuring J, Zayed I. 1978. Isoprenaline-induced myocardial necrosis in dogs. *Br J Exp Pathol* **59**:148–57.

Gordon I, Shapiro HA, Berson SD. 1988. *Forensic medicine; a guide to principles*, 3rd edn. Churchill Livingstone, Edinburgh.

Green MA. 1982. Sudden and suspicious deaths outside the deceased's own country – time for an international protocol. *Forensic Sci Int* **20**:71–5.

Gresham GA, Turner AF. 1979. *Post mortem procedures*. Wolfe Medical Publications, London.

Hartveit F, Karwinski B, Giertsen JC. 1987. Changes in autopsy profile – 1975 and 1984. *J Pathol* **153**:91–8.

Health Services Advisory Committee. 1991. *Safe working and prevention of infection in clinical labotories*. HMSO, London.

Heath Services Advisory Committee. 1991. *Safe working and the prevention of infection in the mortuary and postmortem room*. HMSO, London.

Hill IR. 1986. Toxicological findings in fatal aircraft accidents in the United Kingdom. *Am J Forensic Med Pathol* **7**:322–6.

Hirsch CS. 1971. The format of the medicolegal autopsy protocol. *Am J Clin Pathol* **55**:407–9.

Hollander N. 1987. Air crash disaster planning. [letter] *Am J Forensic Med Pathol* **8**:183–4.

Hunt AC. 1967. Instant photography in forensic medicine. *Med Sci Law* **7**:216–17.

Hunter J. 1772. *On the digestion of the stomach after death*. Philosophical Transactions of the Royal Society of London, Chapter 62, pp. 447–52. Royal Society of London, London.

Janssen W. 1984. *Forensic histopathology*. Springer Verlag, Berlin.

Jay GW, Leestma JE. 1981. Sudden death in epilepsy. A comprehensive review of the literature and proposed mechanisms. *Acta Neurol Scand Suppl* **82**:1–66.

Johnson HRM. 1969. The incidence of unnatural deaths which have been presumed to be natural in coroners' autopsies. *Med Sci Law* **9**:102–7.

Johnson MD, Schaffner W, Atkinson J, *et al.* 1997. Autopsy risk and acquisition of human immunodeficiency virus infection: a case report and reappraisal. *Arch Pathol Lab Med* **121**:64–6.

Kalichman SC. 1999. Psychological and social correlates of high-risk sexual behaviour among men and women living with HIV/AIDS. *AIDS Care* **11**:415–27.

Karch SB. 1987. Resuscitation-induced myocardial necrosis. Catecholamines and defibrillation. *Am J Forensic Med Pathol* **8**:3–8.

Kenyon CD. 1988. Identification and repatriation following major air disasters. *Forensic Sci Int* **36**:223–30.

Kibayashi K, Hamada K, Honjyo K, *et al.* 1993. Differentiation between bruises and putrefactive discolorations of the skin by immunological analysis of glycophorin A. *Forensic Sci Int* **61**:111–17.

Knight B. 1980. The obscure autopsy. *Forensic Sci Int* **16**:237–40.

Knight B. 1980. *Discovering the human body*. Lippincott & Crowell, New York.

Knight B. 1982. *Lawyer's guide to forensic medicine*. Heinemann Medical Books, London.

Knight B. 1983. *The coroner's autopsy*. Churchill Livingstone, London.

Knight B. 1984. *The post-mortem technician's handbook*. Blackwell Scientific Publications, Oxford.

Knight B. 1985. *Pocket picture guide to forensic medicine*. Gower Medical Publishing, London.

Knight B. 1988. A model medico-legal system. *Forensic Sci Int* **39**:1–4.

Knight B. 1996. *Simpson's forensic medicine*, 11th edn. Edward Arnold, London.

Koelmeyer TD, Beer B, Mullins PR. 1982. A computer-based analysis of injuries sustained by victims of a major air disaster. *Am J Forensic Med Pathol* **3**:11–16.

Kuller L, Lilienfeld A, Fisher R. 1966. Sudden and unexpected death in young adults: an epidemiological study. *JAMA* **198**:248–52.

Kunnen M, Thomas F, Van de Velde E. 1966. Proceedings of the British association in forensic medicine. Semi-microradiography of the larynx on post-mortem material. *Med Sci Law* **6**:218–19.

Landefeld CS, Chren MM, Myers A, *et al.* 1988. Diagnostic yield of the autopsy in a university hospital and a community hospital. *N Engl J Med* **318**:1249–54.

Langan Y, Nashef L, Sander JW. 2000. Sudden unexpected death in epilepsy: a series of witnessed deaths. *J Neurol Neurosurg Psychiatry* **68**:211–13.

Lawrence RD. 1985. Inaccuracy of measuring wounds on autopsy photographs. *Am J Forensic Med Pathol* **6**:17–18.

Leadbeatter S. 1986. Semantics of death certification. *J R Coll Physicians Lond* **20**:129–32.

Leadbeatter S. 1991. Deaths of British nationals abroad – a 10-year survey. *Forensic Sci Int* **49**:103–11.

Leadbeatter S, Knight B. 1987. The history and the cause of death. *Med Sci Law* **27**:132–5.

Leadbeatter S, Knight B. 1988. Resuscitation artefact. *Med Sci Law* **28**:200–4.

Lechat MF. 1976. The epidemiology of disasters. *Proc R Soc Med* **69**:421–6.

Leestma JE, Kalelkar MB, Teas SS, *et al.* 1984. Sudden unexpected death associated with seizures: analysis of 66 cases. *Epilepsia* **25**:84–8.

Leestma JE, Hughes JR, Teas SS, *et al.* 1985. Sudden epilepsy deaths and the forensic pathologist. *Am J Forensic Med Pathol* **6**:215–18.

Leestma JE, Annegers JF, Brodie MJ, *et al.* 1997. Sudden unexplained death in epilepsy: observations from a large clinical development program. *Epilepsia* **38**:47–55.

Li L, Zhang X, Constantine NT, *et al.* 1993. Seroprevalence of parenterally transmitted viruses (HIV-1, HBV, HCV, and HTLV-I/II) in forensic autopsy cases. *J Forensic Sci* **38**:1075–83.

Ludwig J. 2002. *Handbook of autopsy practice*, 3rd edn. Humana Press, Totowa, New Jersey.

Macdonald N, Evans B. 2000. Increased high risk sexual behaviour in homosexual men. There is no evidence for a decreased incidence of HIV infection. *Br Med J* **321**:1531–2; discussion 1532.

Mant AK. 1984. *Taylor's principles and practice of medical jurisprudence*, 13th edn. Churchill Livingstone, London.

Mason J. 1993. Resuscitation artefacts, including bone marrow. *Forensic medicine: an illustrated reference*. Chapman & Hall, London, pp. 39–42.

Mason JK. 1962. *Aviation accident pathology: a study of fatalities*. Butterworth, London.

Mason JK. 1962. The pathology of unsuccessful escape in flight. *Med Sci Law* **2**:124–30.

McCurdy WC. 1987. Postmortem specimen collection. *Forensic Sci Int* **35**:61–5.

McGarry P. 1967. A quick, simple method of removal of the spinal cord. *Arch Pathol* **83**:333–5.

Moritz A. 1942. *The pathology of trauma*. Lea & Febiger, Philadelphia.

Moritz AR. 1956. Classical mistakes in forensic pathology. *American Journal of Clinical Pathology* **26**:1383–92.

Morley AR. 1988. Value of the autopsy. *Assoc Clin Pathol News* **Spring issue**.

Nolte KB, Dasgupta A. 1996. Prevention of occupational cyanide exposure in autopsy prosectors. *J Forensic Sci* **41**:146–7.

Nolte KB, Taylor DG, Richmond JY. 2002. Biosafety considerations for autopsy. *Am J Forensic Med Pathol* **23**:107–22.

Peacock SJ, Machin D, Duboulay CE, *et al.* 1988. The autopsy: a useful tool or an old relic? *J Pathol* **156**:9–14.

Polson CJ, Marshall TK. 1975. *The disposal of the dead*, 3rd edn. English University Press, London.

Polson CJ, Gee D, Knight B. 1985. *Essentials of forensic medicine*, 4th edn. Pergamon Press, Oxford.

Pounder DJ. 1985. Natural disaster potential and counterdisaster planning in Australia. *Am J Forensic Med Pathol* **6**:347–57.

Prinsloo I, Gordon I. 1951. Post-mortem dissection artefacts of the neck and their differentiation from ante-mortem bruises. *S Afr Med J* **25**:358–61.

Puschel K, Lieske K, Hashimoto Y, *et al.* 1987. HIV infection in forensic autopsy cases. *Forensic Sci Int* **34**:169–74.

Puschel K, Lockemann U, Schneider V, *et al.* 1992. HIV-1 prevalence among drug deaths in major cities of central and northern europe. *Forensic Sci Int* **57**:57–62.

Reiter C, Risser D. 1994. Experience with the cytological demonstration of smoker cells in the identification of disaster victims illustrated by the findings concerning the Lauda Air airliner crash near Bangkok. *Forensic Sci Int* **66**:23–31.

Royal College of Pathologists. 1990. *Major disasters, the pathologist's role*. Royal College of Pathologists, London.

Royal College of Pathologists. 1993. *Guidelines for postmortem reports*. Royal Collage of Pathologists, London.

Royal College of Pathologists. 1995. *HIV and the practice of pathology; report of a working party of the Royal College of Pathologists*. Royal College of Pathologists, London.

Rutherford WH, de Boer J. 1983. The definition and classification of disasters. *Injury* **15**:10–12.

Saukko P. 1984. [Autopsy in Scandinavia.] *Beitr Gerichtl Med* **42**:339–41.

Saukko P, Pollak S. 2000. Postmortem examination: procedures and standards. In: Siegel JA, Saukko PJ, Knupfer GC (eds), *Encyclopedia of forensic sciences*. Academic Press, San Diego/San Francisco/New York/Boston/London/Sydney/Tokyo, Vol. 3, pp. 1272–5.

Schmidt G, Kallieris D. 1982. Use of radiographs in the forensic autopsy. *Forensic Sci Int* **19**:263–70.

Schultz TC. 1987. Simple method for demonstrating coronary arteries at postmortem examination. *Am J Forensic Med Pathol* **8**:313–16.

Shapiro H. 1954. Medicolegal mythology; some popular forensic fallacies. *J Forensic Med* **1**:144–69.

Sidwell RU, Green JS, Novelli V. 1999. Management of occupational exposure to HIV – what actually happens. *Commun Dis Public Health* **2**:287–90.

Simpson CK. 1965. *Taylor's principles and practice of medical jurisprudence*, 12th edn. J. & A. Churchill, London.

Skinner M. 1987. Planning the archaeological recovery of evidence from recent mass graves. *Forensic Sci Int* **34**:267–87.

Solheim T, van den Bos A. 1982. International disaster identification report. Investigative and dental aspects. *Am J Forensic Med Pathol* **3**:63–7.

Stevanovic G, Tucakovic G, Dotlic R, *et al.* 1986. Correlation of clinical diagnoses with autopsy findings: a retrospective study of 2,145 consecutive autopsies. *Hum Pathol* **17**:1225–30.

Stevens PJ. 1962. The medical investigation of fatal aircraft accidents. *Med Sci Law* **2**:101–9.

Stevens PJ. 1968. The pathology of fatal public transport aviation accidents. *Med Sci Law* **8**:41–8.

Stevens PJ, Tarlton S. 1963. Identification of mass casualties experience in four civil air disasters. *Med Sci Law* **3**:154–68.

Stewart TD (ed.). 1970. *Personal identification in mass disasters*. Smithsonian Museum. National Museum of Natural History, Washington, DC.

Sturt RH. 1988. Ver Heyden de Lancey lecture on a medico-legal matter. The role of the coroner with special reference to major disasters. *Med Sci Law* **28**:275–85.

Svendsen E, Hill RB. 1987. Autopsy legislation and practice in various countries. *Arch Pathol Lab Med* **111**:846–50.

Tedeschi LG. 1980. Future of the autopsy: a redelineation. *Am J Forensic Med Pathol* **1**:1034.

Titford M, Cumberland G. 1986. Forensic histotechnology: autopsy tissues and public issues. *J Histotechnol* **9**:263.

Tough SC, Green FH, Paul JE, *et al.* 1996. Sudden death from asthma in 108 children and young adults. *J Asthma* **33**:179–88.

Underwood JC, Slater DN, Parsons MA. 1983. The needle necropsy. *Br Med J Clin Res Ed* **286**:1632–4.

van den Bos A. 1980. Mass identification: a multidisciplinary operation. The Dutch experience. *Am J Forensic Med Pathol* **1**:265–70.

Vanezis P, Sims BG, Grant JH. 1978. Medical and scientific investigations of an exhumation in unhallowed ground. *Med Sci Law* **18**:209–21.

Waldron H, Vickerstaffe L. 1977. *Intimations of quality: antemortem and postmortem diagnoses*. Nuffield Provincial Hospitals Trust, London.

Wren C, O'Sullivan JJ, Wright C. 2000. Sudden death in children and adolescents. *Heart* **83**:410–13.

Wright RK, Tate LG. 1980. Forensic pathology. Last stronghold of the autopsy. *Am J Forensic Med Pathol* **1**:57–60.

CHAPTER 2

The pathophysiology of death

- Types of death 52
- The indications of death 53
- Hypostasis 55
- Rigor mortis 60
- Putrefaction 64
- Formation of adipocere 70
- Mummification 72
- Post-mortem damage by predators 73
- The entomology of the dead and post-mortem interval 76
- Estimation of the time since death by body cooling 78
- Stomach emptying as a measure of time since a death 83
- The use of vitreous humour chemistry in timing death 88
- Post-mortem chemistry 89
- References and further reading 90

Most of the discussion concerning the definitions of death belongs to the field of legal medicine and even that of ethics, rather than forensic pathology. Notwithstanding this, a quick review of this complex problem is appropriate here, as when we come to describe the various post-mortem phenomena, some appreciation is necessary of the zero point from which such changes are timed.

TYPES OF DEATH

It is conventional to describe two types of death:

- somatic death, in which the person irreversibly loses its sentient personality, being unconscious, unable to be aware of (or to communicate with) its environment, and unable to appreciate any sensory stimuli or to initiate any voluntary movement. Reflex nervous activity may, however, persist, and circulatory and respiratory functions continue either spontaneously or with artificial support so that the tissues and cells of the body, other than those already damaged in the central nervous system, are alive and functioning.
- cellular death, in which the tissues and their constituent cells are dead – that is, they no longer function or have metabolic activity, primarily aerobic respiration.

Cellular death follows the ischaemia and anoxia inevitably consequent upon cardiorespiratory failure, but it is a process rather than an event, except in the exceptionally rare circumstance of almost instantaneous total bodily destruction, such as falling into molten metal or a nuclear explosion. Even fragmentation of a body by a bomb does not kill all cells instantly.

Different tissues die at different rates, the cerebral cortex being vulnerable to only a few minutes' anoxia, whereas connective tissues and even muscle survive for many hours, even days after the cessation of the circulation.

Brain death

Without trespassing too far into clinical medicine, it may be further noted that somatic death is virtually equated with brain death. When the higher levels of cerebral activity are selectively lost, either from a period of hypoxia, trauma, or toxic insult, the victim will exist in a 'vegetative state'. Here the survival of the brainstem ensures that spontaneous breathing will continue and therefore cardiac function is not compromised. The victim can remain in deep coma almost indefinitely – certainly for years – though debilitating complications such as postural skin necrosis, muscle contractures, and secondary chest infections may well shorten life. Such vegetative patients are not considered 'dead' by most standards, though there is a faction of the medical profession who claim that, because of the irreversible loss of their awareness of themselves and the world around them, they can no longer be considered 'human beings'.

As they are not on life-support machines, however, there is nothing short of euthanasia that can be done about their condition. Attitudes are changing, however, as evidenced by the case of Tony Bland, who declined into a persistent vegetative state, with functioning respiration, after hypoxia caused by crush asphyxia during the Sheffield Hillsborough football disaster. After several years in this persistent vegetative state, his family obtained the consent of the courts in 1993 to allow nutrition to be withdrawn and natural death to supervene.

It is another matter when brain death spreads below the tentorium. When the brainstem (specifically the midbrain, pons and upper medulla) suffers neuronal damage, the loss of the 'vital centres' that control respiration, and of the ascending reticular activating system that sustains consciousness, cause the victim not only to be irreversibly comatose, but also to be incapable of spontaneous breathing. Without medical intervention, hypoxic cardiac arrest inevitably follows within minutes and then the usual progression of 'cellular death' ensues. This invariably occurred in the past, before the advent of sophisticated devices that allowed the support of respiration after spontaneous function has ceased. Because of continued oxygenation, the heart continues to function, though it has been pointed out by Pallis (1983) that the vast majority of brainstem-dead patients suffer a cardiac arrest within 48–72 hours, even when adequately oxygenated.

This then is the state of 'brainstem death', and almost all doctors and most jurisdictions now accept that, once irreversible damage to the brainstem has been proved, the patient is 'dead' in the somatic sense. They are, however, not yet dead in the cellular sense and it is through this 'physiological window' that the spectacular advances in cadaver donor transplant surgery have been made.

MEDICO-LEGAL ASPECTS OF BRAIN DEATH

In relation to forensic pathology and criminal procedure, these matters have now a largely historical relevance in that several decades ago a number of fatal assaults were defended in court on the basis that the original head injuries led to the victims being placed on life-support machines that were later disconnected when brainstem death was diagnosed. The defence was that the assailant did not 'kill' the victim, because 'death' was caused by the action of the doctors switching off the machine. Several judicial decisions have disposed of this defence, however, as it was accepted, first, that death was diagnosed by the doctors before discontinuing artificial ventilation and, second, that the criminal act that initiated the train of events was not remote enough to avoid culpability for the death.

THE INDICATIONS OF DEATH

Here the signs of somatic death are considered, as those of cellular death are manifested by post-mortem changes, to be discussed in later sections. When cardiorespiratory arrest occurs, brain function ceases within seconds as a result of the collapse of cerebral blood pressure and consequent cortical ischaemia. Within minutes, this loss of brain function becomes irreversible. The actual number of minutes for which total anoxia will cause cortical damage is controversial; it was formerly held that 3 minutes was sufficient, but this time has been extended to 7 or 9 minutes. Even much longer times of total hypoxia, such as immersion under water, has been survived without brain damage. In these cases, an element of hypothermia is usual, which reduces the oxygen needs of the tissues. Thus 20 or even 40 minutes of hypoxia has been claimed as not leading to brain damage. All that can be said is that the period is very variable and that 3–4 minutes of cardiac arrest or failure of respiratory gas exchange carries the risk of cortical damage, even though three or four times this period may sometimes be survived without ill-effects.

- Unconsciousness and loss of all reflexes occurs, and there is no reaction to painful stimuli. Rarely, there may be post-mortem coordinated muscle group activity for up to one hour after death, possibly due to surviving cells in the spinal cord (Nokes *et al.* 1989).
- Muscular flaccidity occurs immediately upon failure of cerebral and cerebellar function. All muscle tone is lost, though the muscles are physically capable of contraction for many hours.

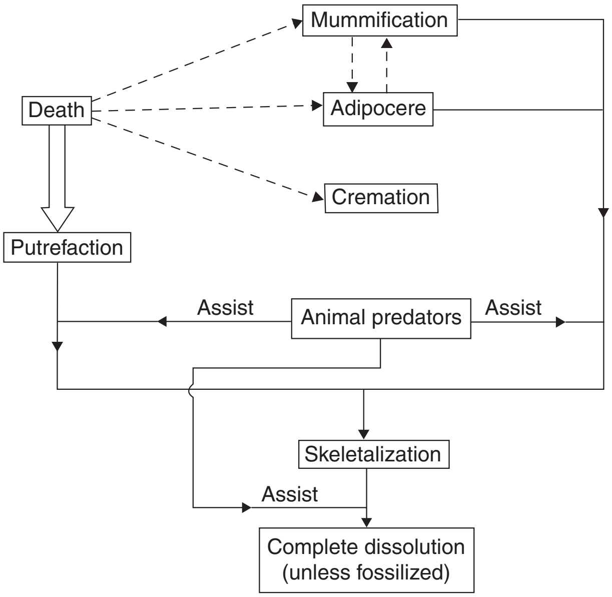

FIGURE 2.1 *The fate of a body after death.*

- Eye signs include loss of the corneal and light reflexes leading to insensitive corneas and fixed, unreactive pupils. Though the iris responds to chemical stimulation for hours after somatic death, the light reflex is lost as soon as the brainstem nuclei suffer ischaemic failure. The pupils usually assume a mid-dilated position, which is the relaxed neutral position of the pupillary muscle, though they may later alter as a result of rigor. There may be a marked difference in the degree of dilatation of each pupil, but this has no significance as a diagnostic sign either of a brain lesion or of drug intoxication. In conditions such as morphine poisoning, where the pupils may be contracted during life, death may allow this to persist or the pupils may dilate to the 'cadaveric position'.

In addition to irregular size, the pupils may lose their circular shape after death as a result of uneven relaxation. This is usually easy to differentiate from the more obvious irregularity caused by ante-mortem abnormality of the iris. The eye globe tension decreases rapidly, as it is dependent upon arterial pressure for its maintenance. The eyeball feels progressively softer within minutes and the cornea soon loses its normal glistening reflectivity because of laxity and failure of lachrymal moistening. Normal tension was measured by Nicati in 1894, who found that it halved at death, fell to as low as an eighth in 30 minutes, and was nil by 2 hours after death. The eyelids usually close, but this is commonly incomplete, the flaccid muscles failing to produce the full occlusion that occurs in voluntary closure. Where the sclera remains exposed, two yellow triangles of desiccated discoloration appear on each side of the cornea within a few hours, becoming brown and then sometimes almost black, giving rise to the name 'tache noire'.

When viewed with an ophthalmoscope, the retina provides one of the earliest positive signs of death. This is the well known 'trucking' of blood in the retinal vessels, when loss of blood pressure allows the blood to break up into segments, similar to trucks in a railway train. This phenomenon occurs all over the body, but only in the retina is it accessible to direct viewing.

The test is not easy to carry out, as the retina of a corpse seems far more difficult to visualize than in a living person. Many observers have described the agonal and early post-mortem appearances of the retina, though some of their accounts are rather contradictory and of little practical value, especially when they attempt to use them to estimate the time since death. Most observers – for example, Kervorkian (1956, 1961) and Salisbury and Melvin (1936) – confirm that the blood flow slows and becomes irregular in density before breaking up into segments. This was apparent within 15 minutes of death as assessed by other means. Kervorkian claimed that colour changes in the retina were closely related to the post-mortem interval. Tomlin believed that segmentation was more indicative of cerebral death than cessation of the circulation, and agreed with others in accepting that 'trucking' was of grave prognostic significance.

One of the best investigations into the phenomenon was by Wroblewski and Ellis (1970), who studied retinal and corneal changes at death in 300 patients. About a third of the total exhibited trucking, most of them within an hour of death. Part of the difficulty in

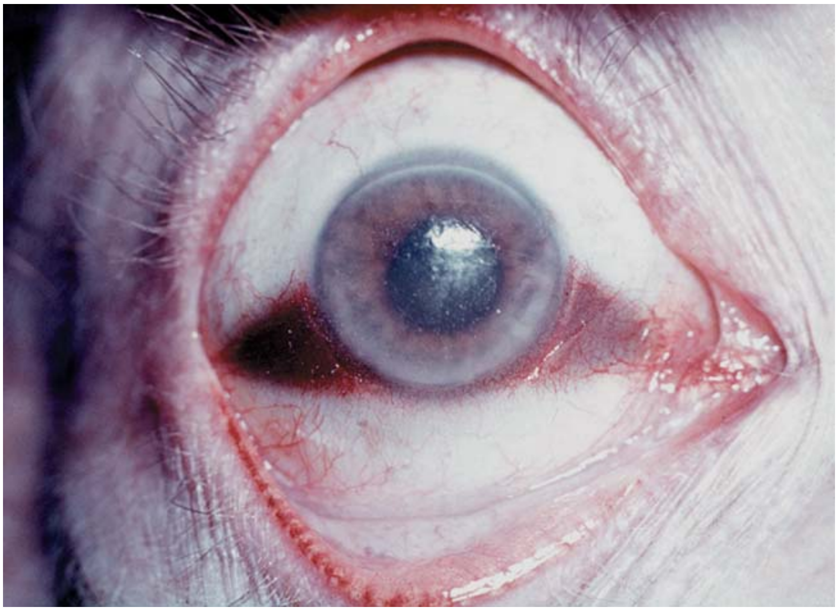

FIGURE 2.2 *Post-mortem change in the eye, the so-called 'tache noire'. These are brown areas of scleral drying caused by failure of the lids to close after death.*

examining the remainder was that clouding of the cornea occurred in 75 per cent of patients within 2 hours of death. They concluded that segmentation was a purely post-mortem change and any intravascular movement of blood, however irregular, was an indication for continued resuscitation.

- Cessation of the heart beat and respiratory movements was the primary marker of death until the advent of mechanical cardiorespiratory support as it led to immediate ischaemia and anoxia of all other tissues. Determination of cardiac arrest may be made by prolonged auscultation of the chest to exclude heart sounds though, as in life, a feeble heartbeat may be muffled by a thick chest wall. The electrocardiograph is unchallengeable in confirming cardiac arrest.

  Respiration is more difficult to confirm, especially in deep coma such as barbiturate poisoning, and prolonged listening with a stethoscope over the trachea or lung fields is necessary. All archaic procedures such as saucers of water on the chest, feathers before the nostrils and tourniquets around the fingers are only of historical interest.

The mode versus the cause of death

Confusion often arises, especially among students and younger doctors, about the distinction between the **mode** of death and its **cause**. This is particularly important in relation to the documentary certification of deaths, but the same confusion sometimes occurs among pathologists, especially those who are not habitually involved in medicolegal cases.

The mode of death refers to an abnormal physiological state that pertained at the time of death: for example, 'coma', 'congestive cardiac failure', 'cardiac arrest' and 'pulmonary oedema'. These offer no information as to the underlying pathological condition and should not be used as the definitive cause of death unless further qualified by the more fundamental aetiological process.

In most cases, the mode is unhelpful and immaterial in describing and understanding the cause of death, and some modal terms are quite useless, such as 'syncope' or 'cardiorespiratory failure'. Even 'bronchopneumonia' is such a common terminal event in numerous diseases that, used alone, it conveys no information about why the patient died. The British Registrar General has recently requested that it should be omitted from death certificates, the basic condition being preferable and sufficient for statistical purposes.

In addition to the mode and cause of death, here is also the **manner** of death, which is not really a medical decision. Manner refers to the circumstantial events, such as 'homicide, suicide, accident or natural cause' and is a legal or administrative categorization.

The recommendations of the World Health Organization, as given in its publication *The medical certification of death*, should be followed to improve both the comprehension and statistical accuracy of the cause of death. Even this booklet is inconsistent with the practice of certification in many advanced countries, in that it requests the doctor to enter the 'manner' of death (for example, homicide, accident or suicide) which, for example in the UK, is the prerogative of the legal authorities after full investigation (see Leadbeatter and Knight 1987).

POST-MORTEM CHANGES OF FORENSIC IMPORTANCE

To the forensic pathologist a number of post-mortem changes are of interest and potential usefulness, mainly in relation to the estimation of the post-mortem interval, possible interference with the body, and an indication of the cause of death.

HYPOSTASIS

Post-mortem hypostasis is known under a variety of older names, such as 'lucidity', 'staining' or 'cogitation', but the current title is most suitable as it indicates the cause. Hypostasis occurs when the circulation ceases, as arterial propulsion and venous return then fail to keep blood moving through the capillary bed, and the associated small afferent and efferent vessels.

Gravity then acts upon the now stagnant blood and pulls it down to the lowest accessible areas. The red cells are most

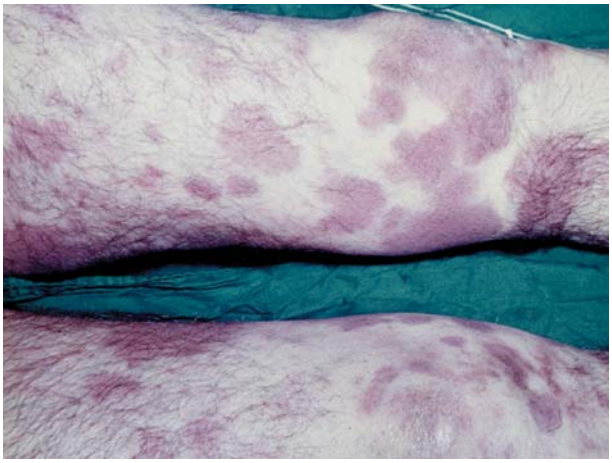

FIGURE 2.3 *Blotchy post-mortem hypostasis, forming in the early hours after death. The patchy disposition has no significance and this usually sinks down and becomes confluent in the most dependent areas within a few more hours.*

affected, sedimenting through the lax network, but plasma also drifts downwards to a lesser extent, causing an eventual post-mortem 'dependent oedema', which contributes to the skin blistering that is part of early post-mortem decay.

The arrival of erythrocytes in the lower areas is visible through the skin as a bluish red discoloration, 'hypostasis'. It often begins as blotchy patches on both lateral and dependent surfaces, but also on the upper surfaces of the legs, especially thighs. These soon coalesce and slide down to the lowest areas.

The distribution of hypostasis

The pattern of hypostasis depends on the posture of the body after death. It is most common when the deceased body is lying on its back, with the shoulders, buttocks and calves pressed against the supporting surface. This compresses the vascular channels in those areas, so that hypostasis is prevented from forming there, the skin remaining white. When the body lies for a sufficient time on the side or face, the hypostasis will distribute itself accordingly, again with white pressure areas at the zones of support.

If the body remains vertical after death, as in hanging, hypostasis will be most marked in the feet, legs and to lesser extent in the hands and distal part of the arms. In addition to pallor of the supporting areas, any local pressure can exclude hypostasis and produce a distinct pattern in contrast to the discoloured area. Examples include the irregular linear marks made by folds in rumpled bedlinen, the pattern of fabric from coarse cloth, the pressure of tight belts, brassière straps, pants' elastic and even socks. According to Bonte *et al.* (1986), when electrocution takes place in water (usually a bathtub), the hypostasis is sharply limited to a horizontal line corresponding to the water level.

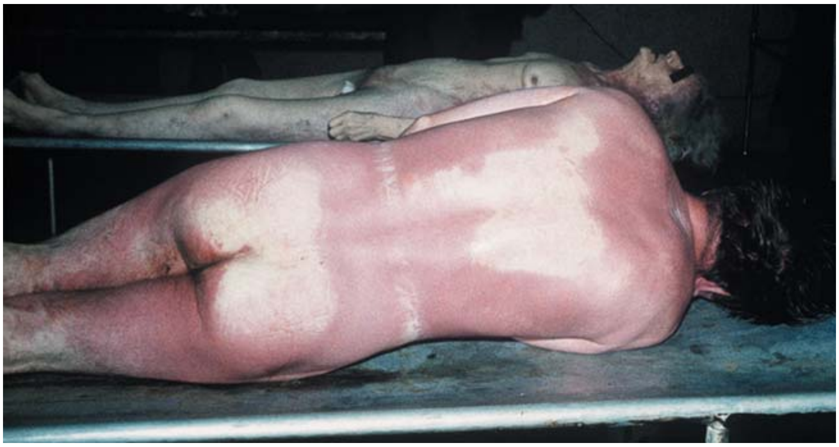

FIGURE 2.4 *Post-mortem hypostasis in the normal distribution. The pale areas are the result of pressure against a hard supporting surface.*

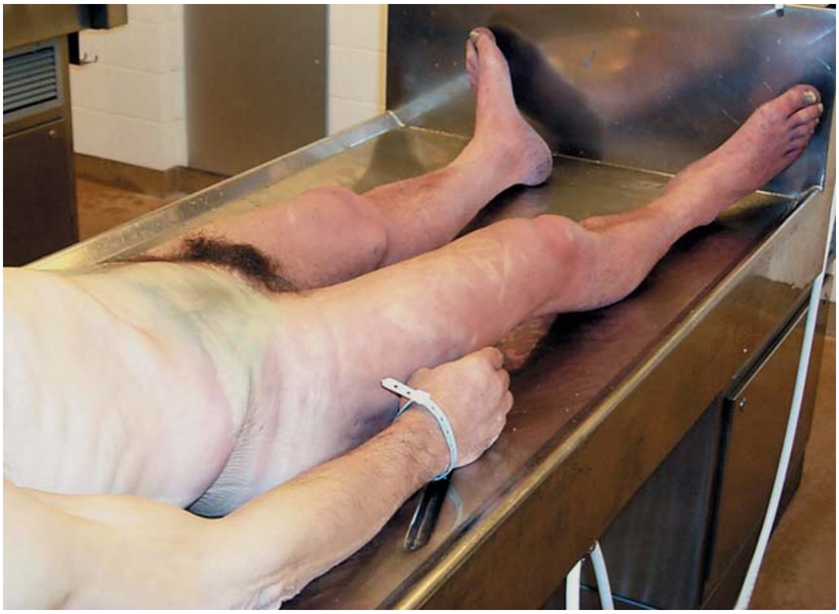

FIGURE 2.5 *Post-mortem hypostasis in a death from hanging. The discoloration of the skin is in the legs and hands, due to the vertical posture after death.*

The colour of hypostasis

The usual hue is a bluish red, but variation is wide. This depends partly on the state of oxygenation at death, those dying in a congested, hypoxic state having a darker tint as a result of reduced haemoglobin in the skin vessels. This is an unsure indicator of the mode of death, however, and no reliance can be placed on a cyanotic darkening of the hypostasis to indicate a hypoxic death in the sense of 'asphyxia'. Many natural deaths from coronary or other disease have markedly dark hypostasis. Often the colour of the hypostasis varies from area to area on the same body. Sometimes a rim of lighter colour may be seen along the margin of the lower darker area and sometimes there is a definite contrast between a bluish zone and a pink margin. This may appear and change as the post-mortem interval lengthens. Often the whole area of hypostasis is pink or bright red. When death has been due to hypothermia or exposure to cold in the agonal period, such as drowning, the colour may assist in confirming the cause of death; again this is relatively non-specific because bodies exposed to cold after death (especially in mortuary refrigeration) may turn pink after an initial stage of normal bluish-red tint.

The mechanism is not understood, but is obviously a result of oxyhaemoglobin forming at the expense of the darker reduced haemoglobin. This is understandable in hypothermia, where the reduced metabolism of the tissues fails to take up oxygen from the circulating blood, but its frequent formation in the post-mortem period is difficult to explain.

Several researchers have investigated the colour of hypostasis in relation to time since death. Schuller *et al.* (1987) noted an increasing paleness between 3 and 15 hours post-mortem, measuring this as a change in wavelength from 575 nm at 3 hours at an average rate of 2 nm per hour. Vanezis (1991) has used tristimulus colorimetry to study colour changes and secondary shifting of hypostasis, and claimed that there is a linear relationship between the fading colour and time during the first 2 hours after death, following which changes are unpredictable. Inoue *et al.* (1994) have also described measurement of hypostatic colour as a measure of time since death.

It may sometimes be noticed that originally bluish hypostasis becomes pink along the upper part of the horizontal margin, the lower parts remaining dark, so a quantitative change probably occurs, the haemoglobin being more easily re-oxygenated where the erythrocytes are packed less densely in the upper layers of hypostasis.

Other changes in the colour of hypostasis are more useful. The best known is the 'cherry-pink' of carboxyhaemoglobin, which is a unique colour and is often the first indication to the pathologist of carbon monoxide poisoning. Cyanide poisoning is said to have its own characteristic dark blue–pink hue, but it is really an index of the congested, cyanotic, mode of death and, if the pathologist was not already aware of the potential cause from the history – and perhaps the odour of cyanide – it is doubtful whether hypostasis would be a primary indication of the nature of the death. The hypostasis may be a brownish red in methaemoglobinaemia and may be various shades in aniline and chlorate poisoning. In deaths from septic abortions where *Clostridium perfringens* is the infecting agent, a pale bronze mottling may sometimes be seen on the skin, though this is not confined to the areas of hypostasis.

Skin haemorrhages, varying in size from small petechiae to large blotches and even palpable blood blisters may develop in areas of hypostasis. The most common place is the back of the shoulders and neck, though they may appear on the front of the chest, even on a body lying on its back. They are more common in cyanotic, congestive types of death and become more pronounced as the post-mortem interval lengthens. Their importance is in not being mistaken for the so-called signs of 'asphyxia'. They appear in the most gross form when a body dies or is left with the head downwards: the confluent petechiae and ecchymoses may be so marked that they virtually blacken the face and neck.

The timing and permanence of hypostasis

Too much has been claimed in the past for the usefulness of hypostasis as an indicator of the time of death and post-mortem disturbance of the body. The phenomenon appears at a variable time after death – indeed, it may not appear at all, especially in infants, old people or those with anaemia. It may be so faint as almost to escape detection.

Hypostasis can appear within half an hour of death or it may be delayed for many hours. Its variability is such that it is useless for any estimation of the time since death. It is claimed that hypostasis can sometimes be observed in the living if the heart action is failing or if venous return is impaired by the immobility of deep coma. The latter is certainly associated with skin blistering caused by dependent oedema.

Once hypostasis is established, there is controversy about its ability to undergo subsequent gravitational shift. If the body is moved into a different posture, the primary hypostasis may either:

- remain fixed
- move completely to the newly dependent zones or
- be partly fixed and partly relocated.

Thus if a corpse is found with the hypostasis in an obviously inappropriate distribution related to the present posture, it must have been moved after death. This fact may

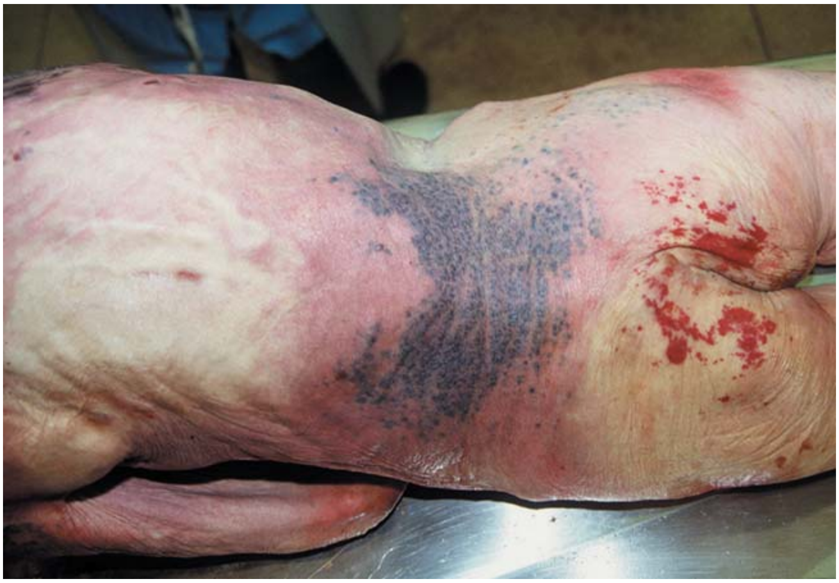

FIGURE 2.6 *Extensive confluent skin haemorrhages may occur within the dependent hypostasis. They worsen as the post-mortem interval lengthens and must not be taken to indicate so-called 'asphyxia'.*

have important significance in the investigation of criminal deaths, such as the return of the culprit or some other person to the scene. It may also migrate either partly or completely, however, thus negating its diagnostic usefulness. Older theories about 'fixation' of the staining after a certain time are not tenable, as there is no constant interval when this occurs.

Mallach (1964) analysed data including onset, confluence, maximum intensity, displacement and shifting, taken from publications between 1905 and 1963. These are summarized in Table 2.1, 'thumb pressure' meaning displacement by thumb pressure. Table 2.2 shows the variable opinions on hypostasis culled from standard textbooks.

A considerable amount of research has been devoted to the investigation of clotting of post-mortem blood, its subsequent lysis and inhibition of coagulation by fibrinolysins, but this cannot be translated into practice. Formerly, it was held that if a body had remained in its original posture for a certain minimum length of time, the blood would coagulate in the hypostatic areas, so that secondary shifting could not then occur if the body was moved. This is not true in the majority of instances, as there may be partial or complete secondary gravitation at any time – at least, until true staining of the tissues due to haemolysis begins as part of early decomposition, which is not until the second or third day in temperate conditions.

Suzutani *et al.* (1978) examined 430 bodies by pressure on hypostatic areas, finding that the colour could not be squeezed out in 30 per cent where death had occurred 6–12 hours previously. More than 50 per cent were fixed after 12–24 hours and no fading occurred in 70 per cent of those dead for more than 1–3 days. However, in a significant number, hypostasis was still mobile for at least 3 days.

Fechner *et al.* (1984) found no linear relationship between fixation and time of death and also observed variations in fixation according to storage temperatures.

TABLE 2.1 *Hypostasis related to the time of death (hpm) as derived from previous literature*

Stage	Mean	Standard deviation	Limits	
			Lower	Upper
Beginning	0.75	0.5	0.25	3
Confluence	2.50	1.0	1.00	4
Maximum	9.50	4.5	3.00	16
Thumb pressure	5.50	6.0	1.00	20
Complete shifting	3.75	1.0	2.00	6
Incomplete shifting	11.00	4.5	4.00	24

TABLE 2.2 *Time of onset of hypostasis as stated by previous authors*

Reference	Onset	Maximum
Adelson	30 min–4 h	8–12
Polson, Gee, Knight	30 min–2 h	6–12
Spitz and Fisher	2–4 h	8–12
Taylor (ed. Simpson)	0 h	12
Taylor (ed. Mant)	1 h	12
Gradwohl (ed. Camps)	20–30 min	6–12
Glaister, Brash	–	8–12
DiMaio	30 min–2 h	8–12
Sydney Smith	0 h	12
Mant	0 h	12
Gordon and Shapiro	'few' h	12

Hypostasis in other organs

Just as blood settles in dependent skin, so it does in other tissues and organs. The importance in forensic autopsy work is the differentiation of organ hypostasis from ante-mortem lesions. In the intestine, dependent loops of jejunum and ileum may be markedly discoloured and mislead the inexperienced pathologist into suspecting mesenteric infarction or strangulation. This hypostasis is discontinuous, however, revealing interrupted segments when the gut is laid out. Often loops in the pelvis are worst affected, because of their lower position.

The lungs almost always show a marked difference in colour from front to back, the anterior margins being pale and the posterior edges lying in the paravertebral gutters being dark blue. This is often accompanied by an obvious difference in fluid content, congestion and oedema being more marked posteriorly. The myocardium often shows a dark patch in the posterior wall of the left ventricle that must not be mistaken for early infarction.

One of the most important hypostatic artefacts is haemorrhage behind the oesophagus at the level of the larynx. This has many times been confused with the trauma of strangulation, but the true nature was best investigated by Prinsloo and Gordon (1951). Areas of congestion developing into haemorrhage appear in the loose tissues on the back of the oesophagus lying on the anterior longitudinal ligament of the cervical spine. They may be prominent, sufficient to produce actual haematomas. They arise in the veins and venous plexus on the front of the spine, and can be avoided in cases of strangulation and suspected strangulation by draining the neck of blood before starting the dissection. This can be achieved either by removing the brain first or by opening the chest and incising the great veins in the superior mediastinum.

Differentiation between hypostasis and bruising

This is rarely a problem in fresh bodies, but when decomposition begins the two conditions become blurred. In fresh material the appearance of hypostasis is of a regular, diffuse engorgement of the surface vessels, the colour varying between purple–red and bright pink. The density varies from place to place but there are no sudden changes in colour nor any sharply circumscribed areas as occurs in bruising. As mentioned above, in the early stages of hypostasis there may be mottled blue areas, which later coalesce. This mottling is unlikely to be confused with bruising by any but the most inexperienced pathologist. The position in dependent areas is also characteristic, giving a generally horizontal orientation. Bruises may be anywhere on the body, are often discoid or have an irregular margin, but rarely cover a large area with uniform density – and do not have a horizontal margin. Abrasions may be topographically associated with a bruise, but not with hypostasis. When there is difficulty in differentiating between the two – usually where there is racial pigmentation or really deep, cyanotic hypostasis that obscures possible bruises – the classic test is to incise the suspect area to see if the underlying blood is intravascular (hypostasis) or infiltrating the tissues outside the vessels (contusion).

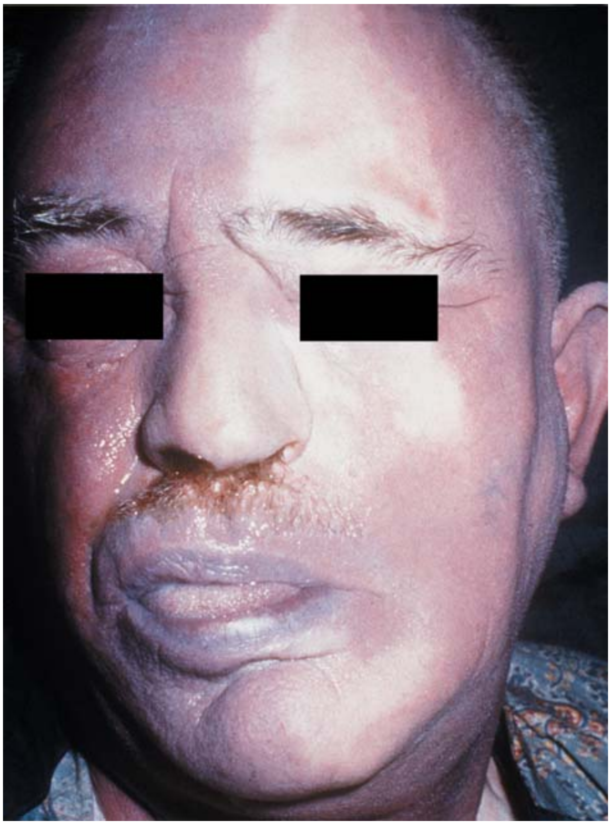

FIGURE 2.7 *White patches within the hypostasis on the face merely indicate pressure against the supporting surface after a face-down position. As in this case, they are usually post-mortem and are not indicators of suffocation.*

Fresh bruises may also be swollen and slightly raised above the surface. If a post-mortem pressure mark (such as from a belt or tight clothing) crosses an area of hypostasis, there will be a pale bloodless zone, but a bruise will not be affected. Hypostasis is in the most superficial layer of the dermis and any exuded blood can be wiped or washed away from the incised surface. A bruise is often deeper in the skin or underlying tissues and is fixed, being infiltrated through the tissues outside the ruptured vessels. An exception is intradermal bruising, but this is usually patterned or linear and rarely can be confused with hypostasis. Histological examination may be necessary finally to decide the matter.

When post-mortem autolysis has developed, the diffusion of haemolysed blood from vessels makes the differentiation between hypostasis and bruising difficult and eventually impossible. Even histological examination may be of little help when the tissues become markedly degenerate.

RIGOR MORTIS

Unlike hypostasis, the stiffening of the muscles after death has some relevance in determining the post-mortem interval. It has been known since antiquity that immediately after death there is general muscular flaccidity, usually followed by a period of partial or total rigidity, which in turn passes off as the signs of decomposition appear. The timing of this sequence of events is so variable, however, that it is a poor indicator of the time since death. The usual range of times when rigor appears can be summarized as follows.

- The flaccid period immediately after death is variable, but commonly extends to between 3 and 6 hours before stiffening is first detected, depending on environmental temperature and other factors. Rigor is first apparent in the smaller muscle groups, not because it begins there, but because the smaller joints, such as the jaw, are more easily immobilized. The sequence of spread of rigor is also variable but tends to affect the jaw, facial muscles and neck before being obvious in the wrists and ankles, then the knees, elbows and hips.

  The usual method of testing is by attempting to flex or extend the joints, though the whole muscle mass itself becomes hard, and finger pressure on the quadriceps or pectorals can also detect the change. The onset of rigor may be markedly accelerated or retarded by the factors mentioned below, so that stiffness may develop within half an hour of death (even excluding cadaveric spasm) or may be postponed almost indefinitely. Some subjects, usually infants, the cachectic and the aged, may never develop recognizable rigor mortis, mainly because of the feebleness of their musculature.
- Rigor spreads to involve the whole muscle mass, again within a variable period but in 'average' conditions it might be expected to reach a maximum within 6–12 hours. This state then remains constant until the muscle mass begins to undergo autolysis, which releases rigor gradually at a stage before overt post-mortem changes are visible externally, except perhaps for a commencing discoloration of the lower abdominal wall.

  The duration of full rigor may be 18–36 hours, until it begins to fade in roughly the same order of muscle groups as it appeared. However, there are many exceptions to this rather misleading generalization.

Factors affecting the timing of rigor mortis

As a chemical process, the speed of onset and the duration of rigor is modified by temperature. The colder the environment the slower the process and vice versa. In near freezing conditions rigor will be suspended almost indefinitely. The author (BK) has attended a scene outdoors in winter where the body was quite flaccid one week after death, rigor developing rapidly as soon as the corpse was brought into the comparative warmth of the mortuary. Conversely, hot weather or tropical conditions can speed up the whole cycle so that rigor appears within an hour or even less. Total stiffness develops rapidly, then fades during the first day as decomposition supervenes.

The other modifier of the speed of onset of rigor is physical activity shortly before death. As will be seen when the

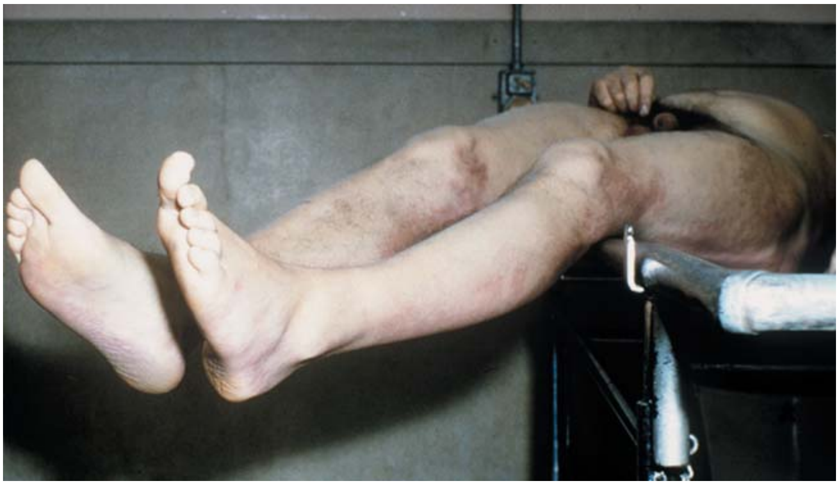

FIGURE 2.8 *Full rigor mortis 12 hours post-mortem.*

physicochemical basis of the phenomenon is discussed, the availability of glycogen and adenosine triphosphate in the muscle is a crucial element in rigor formation. Muscular exertion affects the interaction of these substances and hastens the onset of rigor. Cadaveric spasm, dealt with later, may well be an extreme variant of this accelerated rigor.

In view of the wide range of times at which the various stages of rigor appear and fade, it is a poor determinant of the time since death, compared with estimations of body temperature. Niderkorn's early work (1874) on 113 bodies showed a range of 2–13 hours for rigor to be complete, the main cluster being from 3 to 6 hours after death, a shorter period than would be accepted now. Mallach (1964) compiled a table from 150 years of published data, which suffered from gross variation in observer methodology (Table 2.3).

The following is a reasonable 'spot check' for use in average temperate conditions:

- If the body feels **warm** and is **flaccid**, it has been dead less than 3 hours.
- If the body feels **warm** and is **stiff**, it has been dead from 3 to 8 hours.
- If the body feels **cold** and is **stiff**, it has been dead from 8 to 36 hours.
- If the body feels **cold** and is **flaccid**, it has been dead more than 36 hours.

This crude estimate should never be used as a definitive statement in legal proceedings, as it is only meant as a rough guide 'on the spot'.

Rigor mortis in other tissues

Rigor occurs in all muscular tissues and organs, as well as the skeletal muscles. The iris is affected so that ante-mortem constriction or dilatation is modified. The rigor may be unequal in each eye, making the pupils unequal, confirming the fact that the post-mortem position is an unreliable indicator of toxic or neurological conditions during life.

In the heart, rigor causes the ventricles to contract, which may be mistaken by the inexperienced pathologist for left ventricular hypertrophy; this can be excluded by measuring the total weight, estimating the relative size of the left side, measuring the ventricular thickness (admittedly a rough guide) and – where the issue is important – by dissection and differential weighing of the two ventricles.

Rigor in the dartos muscle of the scrotum can compress the testes and epididymis which, together with the contraction of muscular fibres in the seminal vesicles and prostate, may lead to post-mortem extrusion of semen from the urethral meatus. This has been wrongly attributed to sexual activity and orgasm just before death in cases where a defence of provocation against a homosexual advance has been put forward. Mant showed that many corpses dying from a variety of causes have seminal fluid either at the meatus or in the penile urethra (Mant 1953, 1967; Mant and Furbank 1957). It is also said, without much foundation, that deaths from hanging and 'asphyxia' are commonly associated with post-mortem seminal emission.

Rigor in the erector pili muscles attached to the hair follicles can cause a pimpling or 'goose-flesh' appearance with elevation of the cutaneous hairs. This may have given rise to the persistent myth that the beard grows after death, though an additional explanation is that post-mortem desiccation and shrinkage of the skin allows the hair stubble to appear more prominent.

Studies by Krompecher and Bergerioux (1988) have shown that rigor also sets in more quickly in deaths from electrocution – and that it passes off more rapidly.

The biochemistry of rigor mortis

Though this has little direct forensic relevance apart from the association of the early onset of rigor with muscular

TABLE 2.3 *Time course of cadaveric rigidity as stated by previous literature*

		Hours post-mortem				
		Limits of 95.5 per cent probability (2s)		Variations		
Rigor phase	Mean with standard deviation(s)	Lower limit	Upper limit	Lower limit	Upper limit	Number of publications evaluated
Delay period	3 ± 2	–	7	< ½	7	26
Re-establishment possible	Up to 5	–	–	2	8	–
Complete rigidity	8 ± 1	6	10	2	20	28
Persistence	57 ± 14	29	85	24	96	27
Resolution	76 ± 32	12	140	24	192	27

exertion, the muscle chemistry of the phenomenon has been studied in great detail.

Szent-Gyorgi (1947) discovered that the essential contractile substances in muscle were the two proteins actin and myosin, arranged in interdigitating filaments. They form a loose physicochemical combination called ‘actomyosin’, which is physically shorter than the two substances uncombined (Hanson and Huxley 1955). If energy is supplied to the latter pair, the subsequent combination contracts. This energy is obtained by the splitting off of a phosphate complex from adenosine triphosphate (ATP) which then becomes adenosine diphosphate (ADP) (Erdos 1943). The free phosphate then engages in a phosphorylation reaction that converts glycogen to lactic acid, high energy being released in the process. Some is used to resynthesize the ATP from ADP, by the donation of a phosphate group from creatine phosphate; the remainder goes to activate the actin–myosin reaction.

In addition to supplying energy, ATP is responsible for the elasticity and plasticity of the muscle. The lactic acid is leached away back into the bloodstream and is returned to the liver for reconversion into glycogen. All these reactions are anaerobic and can continue after death, albeit in a distorted form.

In life there is a fairly constant concentration of ATP in the muscular tissues, there being a dynamic balance between utilization and resynthesis. At death, however, the ADP to ATP reaction ceases and the triphosphate is progressively diminished, with lactic acid accumulating. After a variable period, depending on temperature and the amount of ATP remaining, the actin and myosin become rigidly linked into a rigid, inextensible gel, with consequent stiffening of the muscle (see Bate-Smith and Bendall 1947; Forster 1964).

The resynthesis of ATP is dependent upon the supply of glycogen, which is depleted by vigorous activity before death; this explains the rapid onset of rigor in these circumstances. Normally there appears to be an initial period soon after death when the ATP level is maintained or even increased as a result of phosphate liberation by glycogenolysis.

Rigor is initiated when the ATP concentration falls to 85 per cent of normal, and the rigidity of the muscle is at maximum when the level declines to 15 per cent (see Bate-Smith and Bendall 1949).

Gross effects of rigor mortis

There has been some controversy over whether rigor only stiffens the muscles or actually shortens them. Sommer, as long ago as 1833, claimed that muscles contracted after death and the changes were actually known as ‘Sommer’s movements’. There have always been semi-apocryphal tales of corpses moving in the mortuary and there seems little doubt that, if a limb happens to be poised in unstable equilibrium on the edge of a tray or table, the developing tension may occasionally cause it to spring off as rigor supervenes. The author (BK) has seen athetoid writhing of the foot of a corpse, which lasted for 40 minutes, more than an hour after death.

There seems to be no doubt that some shortening does occur, but the noticeable effects are slight because both flexor and extensor muscle groups oppose each other across most limb joints. Bate-Smith and Bendall (1949) decided that shortening only occurred when there had been marked depletion of glycogen by activity before death, but Forster (1964) was of the opinion that, when a muscle was under some tension, it did shorten. His experiments showed that when the muscle was unloaded there was no change in length when rigor set in. Forster further showed that a high environmental temperature and poisons that increase muscle tone, such as parathion, lead to more shortening during rigor.

When fully established, rigor is ‘broken’ by forcible movements of the limbs or neck, then it will not return, a phenomenon utilized daily by mortuary staff and undertakers when preparing a body for a coffin. If rigor is still developing, it will continue in the new posture of the limbs after they have been stretched. ‘Breaking’ fully established rigor is an accurate description, as the rigid, inelastic fibres are physically ruptured – sometimes tearing the muscle insertions from the bone. Rarely, rigor can assist in showing that a body has been moved between death and discovery. If an arm or leg is found projecting into free space without support, in a posture that obviously could not have been maintained during primary post-mortem flaccidity, then it must have been rolled over or otherwise moved. In these cases, a simple restorative movement (after the scene has been fully examined) can usually indicate the original attitude quite simply.

Conversely, it can never be assumed that the posture of rigor in which a body is found was that which obtained at the time of death, as any amount of movement during the period of primary flaccidity will not be mirrored in the subsequent rigor.

Cadaveric spasm

This topic has received a disproportionate amount of notice in textbooks compared with its practical importance. No doubt this is because of its curiosity value rather than its usefulness. Cadaveric spasm is a rare form of virtually instantaneous rigor that develops at the time of death with no period of post-mortem flaccidity.

Krompecher (1994), a major authority on rigor mortis, is extremely sceptical about the existence of true cadaveric

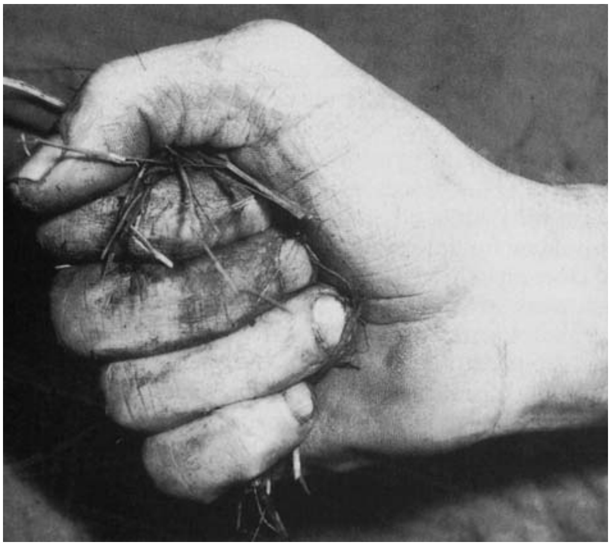

FIGURE 2.9 *Cadaveric spasm, an instantaneous form of rigor, in the victim of a fall into water. The victim was recovered within a short time (as can be seen from the absence of skin maceration) but had grass from the river bank firmly clutched in the hand.*

spasm, but many forensic pathologists claim to have seen such a phenomenon far too soon after death for it to be normal rigor mortis. The authors retain a healthy scepticism about the controversy, believing that most cases are misreported because of errors or uncertainties about the true time of death, so that early normal rigor could have supervened. However, one or two cases within BK's experience seem to be genuine enough to be remarkably early for true rigor.

It seems confined to those deaths that occur in the midst of intense physical and/or emotional activity, though how the latter can lead to instant post-mortem rigor is quite inexplicable. It presumably must be initiated by motor nerve action, but for some reason there is a failure of the normal relaxation. The phenomenon usually affects only one group of muscles, such as the flexors of one arm, rather than the whole body.

It was shown in the earlier section on the biochemistry of rigor that marked depletion of glycogen stores in the muscle by violent exertion immediately before death can hasten the onset of muscular rigidity. Most cases of cadaveric spasm occur in similar circumstances and it was said to be particularly common on the battlefield amongst soldiers slain in combat. In the civilian sphere it is most often seen in persons who fall into water or drop some distance down a precipitous slope such as a cliff. They may clutch at some nearby object, such as grass or shrubs, in an effort to break their fall and such material may be found held tightly in their fingers, even when the body is examined within a few minutes.

Another possibility, more common in detective fiction than in practice, is the gripping of a pistol with the finger still tightly flexed on the trigger, as evidence of true suicide rather than a 'planted' weapon in a homicide where an attempt has been made to simulate self-shooting. The chances of this actually being encountered by a pathologist are less than once in several professional lifetimes.

If found in the victim of drowning, or of a slide from a height, it has some value in confirming that the person was alive at the time of the fall, thus excluding the post-mortem disposal of an already dead body. Of course the body must be examined before ordinary rigor might be expected to have developed, or the presence of cadaveric spasm cannot then be assumed.

HEAT AND COLD STIFFENING

At extremes of temperature the muscles may undergo a false rigor. In extreme cold, well below zero, once the intrinsic body heat is lost, the muscles may harden because the body fluids may freeze solid – as in the commercial or domestic preservation of meat in a freezing cabinet. The temperature has to be below –5°C for cold stiffening to occur and is usually much lower. Part of the apparent stiffness is also due to solidification of the subcutaneous fat. When the body is warmed up, true rigor may supervene, though it may fail to appear as a result of intracellular damage caused by cell membranes being punctured by ice crystals.

Heat applied to the body also causes stiffness of the muscles, as the proteins of the tissues become denatured and coagulated as in cooking. The degree and depth of change depends on the intensity of the heat and the time for which it was applied. At autopsy the muscles may be shrivelled and desiccated, even carbonized on the surface. Beneath this there is a zone (which may be total) of brownish-pink 'cooked meat' and under that, if the process has not penetrated, normal red muscle. Marked shortening occurs, causing the well-known 'pugilistic attitude' of a burned body. This is because of the greater mass of flexor muscles compared with extensors, which forces the limbs into flexion and the spine into opisthotonus. These changes are purely post-mortem and are no indication of burning during life, as similar distortions occur during cremation.

POST-MORTEM DECOMPOSITION

Hypostasis and decomposition occur relatively soon after death when somatic death has occurred, but cellular death is incomplete. As discussed earlier, death is a process not an event and, while the cells of some tissues are still alive and even capable of movement (such as fibroblasts, leucocytes and muscle), others are dying or dead. The process of decomposition begin in some cells while others are still alive, and this overlap continues for several days in temperate climates.

Decomposition is a mixed process ranging from autolysis of individual cells by internal chemical breakdown to tissue autolysis from liberated enzymes, and external processes introduced by bacteria and fungi from the intestine and outer environment. Animal predators, from maggots to mammals, can be included in the range of destruction. Decomposition may differ from body to body, from environment to environment, and even from one part of the same corpse to another. Sometimes one portion of a corpse may show leathery, mummified preservation whilst the rest is in a state of liquefying putrefaction.

In addition, the time scale for decomposition may vary greatly in different circumstances and climates, and even in the same corpse: the head and arms may be skeletalized, whilst the legs and trunk, perhaps protected by clothing or other covering, may be moderately intact. All permutations may be found, making it even more difficult to estimate the probable time since death.

Within the general description of 'decomposition' there are several subclasses, which, as just stated, may merge into each other or may be uniform throughout the same cadaver. Most unembalmed bodies undergo putrefaction, in which the tissues become moist and gas-ridden, and eventually liquefy down to the skeleton.

Alternatives are dry decomposition, termed mummification, or a conversion to waxy substances, called adipocere; rarely, tannin occurs by the action of acid, protein-precipitating fluids in the anaerobic conditions of a peat bog. In some stillbirths an aseptic autolysis of a fetus dead *in utero* is known as maceration.

PUTREFACTION

The usual process of corruption of the dead body begins at a variable time after death, but in an average temperate climate may be expected to begin at about 3 days in the unrefrigerated corpse.

Even in temperate zones there can be a wide range of ambient temperatures, from below freezing to near blood heat. In the tropics, far higher temperatures are commonplace, but in high latitudes or elevations, deep-freeze conditions can keep decomposition at bay indefinitely, as in the case of modern discoveries of prehistoric mammoths and medieval Esquimaux. It is therefore futile to attempt to construct a timetable for the stages of decomposition, except to point out salient markers for an undisturbed body in an 'average' indoor environment of about 18°C in temperate climates. From this approximate baseline the pathologist must then extrapolate for local variations appropriate to his climatic and geographical conditions.

Sequence of putrefactive changes

Whatever the time scale, the general order of corruption is similar, though the degree of advancement may vary between different areas of even the same corpse.

Usually the first external naked-eye sign is discoloration of the lower abdominal wall, most often in the right iliac fossa where the bacteria-laden caecum lies fairly superficially. Direct spread of organisms from the bowel into the tissues of the abdominal wall breaks down haemoglobin into sulphaemoglobin and other pigmented substances. This discoloration spreads progressively over the abdomen, which in the later phase begins to become distended with gas.

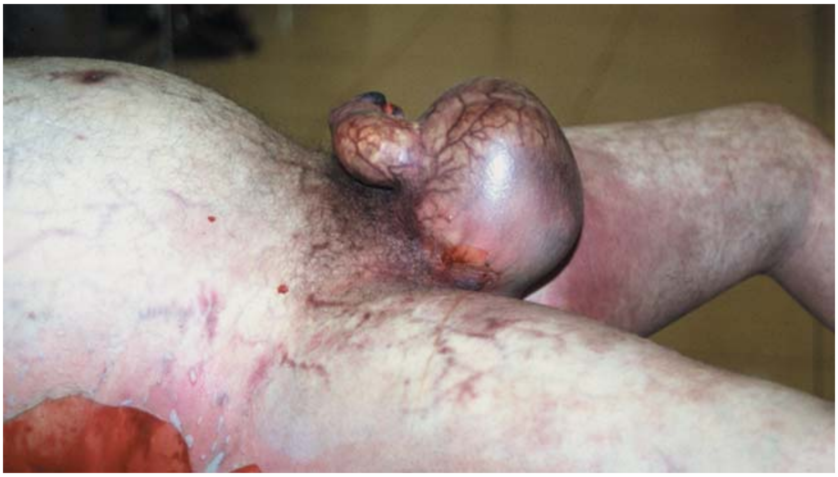

FIGURE 2.10 *Moderately early changes of decomposition, showing gaseous distension of scrotum and abdomen, and skin slippage and blistering in dependent areas. This was after about one week since death, but the changes vary greatly with environmental temperatures.*

At about this time, more generalized spread of bacteria begins to discolour the more moist tissues, which often comprise those dependent areas that show hypostatic coloration and oedema. The face and neck become reddish, and begin to swell. The putrefactive bacteria, which largely originate in the intestines and lungs, spread most easily in fluid so they tend to colonize the venous system, haemolysing the blood that stains the vessel walls and adjacent tissues. This gives rise to 'marbling', a branching outline of arborescent red, then greenish pattern in the skin, seen most clearly on the thighs, sides of the abdomen, and chest and shoulders. This stage may be about one week in the 'baseline' timetable, where the corpse is in air at 18–20°C.

At or even before the stage of marbling, skin blisters may appear, at first on the lower surfaces of trunk and thighs where hypostatic oedema has loaded the tissues with fluid. The upper epidermis becomes loosened, the so-called 'skin-slippage' giving rise to large, fragile sacs of clear, pink or red serous fluid. These may become so large that they are pendulous and soon burst, leaving areas of slimy, pink epidermis. Where this skin change occurs in hairy parts, such as scalp, axilla and pubis, the hair becomes detached and will slide off on slight pressure. Gas formation will now become marked, with increased tension in the abdomen.

The scrotum and penis may swell up to remarkable size and the neck and face will become grotesquely bloated, making visual identification difficult or impossible. The pressure may cause the eye globes and tongue to protrude.

Purging of urine and faeces may occur due to the intra-abdominal pressure and, occasionally, a uterine prolapse may be extruded. There are recorded instances of pregnant women having a macabre post-mortem 'delivery' of the fetus, from the same cause. Bloody fluid, which is tissue liquefaction stained by haemolysis, may leak from any orifice, especially the mouth, nostrils, rectum and vagina. By this stage, some 2–3 weeks may have elapsed since death.

To the inexperienced, this may suggest evidence of some lethal haemorrhage and many pathologists have been called to a scene of 'crime' because of the understandable concern of an investigating police officer that such bloody purging may be evidence of a violent mode of death. Pressure inside the chest due to gas formation in the abdomen may expel air, so that the fluids of decomposition in the trachea and bronchi may be blown into a bloody froth from the mouth and nostrils.

After several weeks, the reddish-green colour of the skin may deepen to a dark green or almost black, but there is marked variation in this aspect. Heavy maggot infestation will almost certainly have supervened except in winter conditions, and the destruction of skin by innumerable maggot holes and sinuses gives better access to other bacteria that may invade from the environment. The maggots secrete a proteolytic enzyme that speeds up the destruction of the tissues, in addition to the direct loss from the voracious appetites of these predators. Skin slippage causes shedding of the outer layers of the fingers and toes, making identification by fingerprints more difficult. Finger- and toenails last longer than the surrounding skin, but they too become loose and eventually fall out. Skin slippage may allow tattoos to become more visible and colourful for a time until the moist underlying dermis itself decomposes.

Internally, decomposition proceeds more slowly than at the surface. It is often quite surprising how valuable an

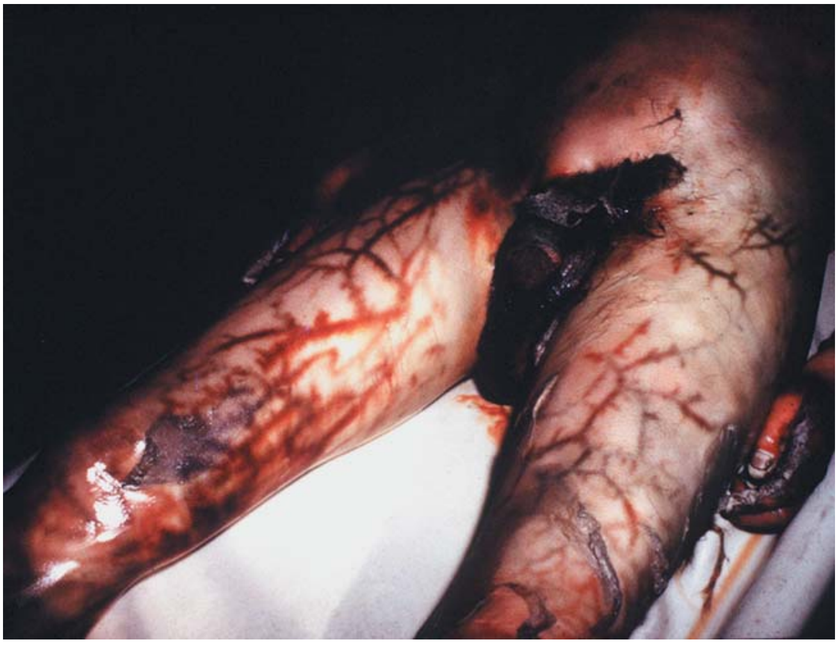

FIGURE 2.11 *Post-mortem decomposition of about 2 weeks' duration in water in a summer temperature. There is 'marbling' of the skin where breakdown products of haemoglobin have stained the venous channels.*

autopsy on a putrefying corpse can be, as the internal organs may be in far better condition than the exterior would suggest. Organs putrefy at markedly different rates. The lining of the intestine, the adrenal medulla and the pancreas autolyse within hours of death, yet the prostate and uterus may still be recognizable in a partially skeletalized body a year later. The brain soon becomes discoloured, being a soft pinkish-grey within a week and liquefying within a month. Meningeal haemorrhage and haematomata persist well, but apart from some tumours, non-haemorrhagic lesions in the brain substance vanish quite early.

The heart is moderately resistant, and examination of the coronary arteries for the degree of atheromatous stenosis may be well worthwhile for many months, though staining of the vessel walls by haemolysis makes recognition of mural thrombosis difficult. Complete occlusion by firm ante-mortem

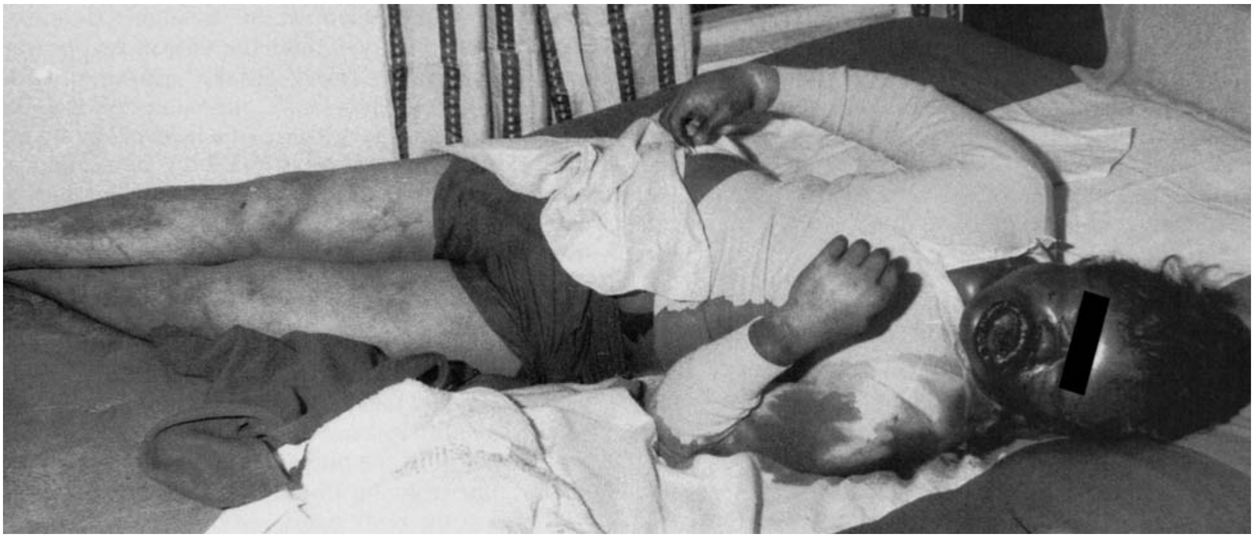

FIGURE 2.12 *Post-mortem decomposition showing putrefactive changes in the face, arms and trunk after one week in a warm room. The illustration shows the contrast sometimes seen between one part of the body and another, as the legs are hardly affected. The face, neck and hands are swollen with gas, and the clothing is stained by leaking skin blisters. The tongue is protruding because of gas pressure from the tissues below. Death was due to carbon monoxide poisoning caused by faulty installation of a room heater.*

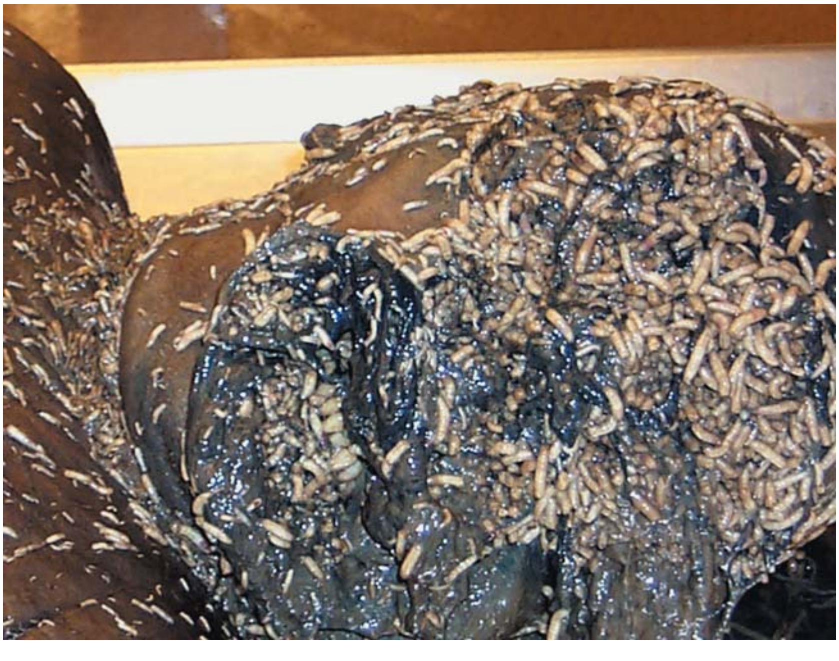

FIGURE 2.13 *Almost complete destruction of the facial soft tissues by maggots. Post-mortem time about 2 weeks in a centrally heated apartment.*

thrombus persists well. The heart may show a curious white granularity on the epicardial and endocardial surfaces, seen either during putrefaction or even on exhumation after a long interval. The nodules are a millimetre or less in size and are often called 'miliary plaques'. They were investigated by Gonzales *et al.* (1954) who found that microscopically they consisted of calcium and soapy material. The origin is obscure, but they seem to be confined to the heart and to be some degenerative product of the cardiac tissue.

In obese people, the body fat (especially perirenal, omental and mesenteric) may liquefy into a translucent, yellow fluid that fills the body cavities between the organs, and makes autopsy even more difficult and unpleasant.

Later putrefactive changes lead to the breakdown of the thoracic and abdominal walls, often hastened by the perforations of maggots and sometimes larger predators, such as rats and dogs. In some tropical countries, such as India and Sri Lanka, bodies left in the open are attacked by many animals, including monitor lizards, and in other parts of the world the indigenous fauna all contribute to the natural cycle that returns the body fats and proteins to the food chain.

After several months, the softer tissues and viscera progressively disintegrate, leaving the more solid organs, such as uterus, heart and prostate, together with the ligamentous and tendinous tissues attached to the skeleton. Often some areas of skin persist, especially where protected by clothing or under the body against the supporting surface. Much depends upon the environment, as a corpse in the open air will suffer far more from rain, wind and especially animal predators, compared with one in a locked room.

Eventually, the body will be reduced to a skeleton, but for some time, ligaments, cartilage and periosteal tags will survive. The season of year and the location will make a great difference to the time scale of skeletalization: a body dying outdoors in the late autumn will 'survive' longer through the cold winter and spring than one dying in the early summer. The effects of animal predators are profound in terms of removing soft tissues from the bones. In broad terms, a corpse outdoors in a temperate climate is likely to be converted to a skeleton carrying tendon tags within 12–18 months, and to a 'bare-bone' skeleton within 3 years; there are, of course, numerous exceptions, depending mainly on the local environment. In closed conditions indoors, a body may never skeletalize, often being converted to a dried, partly putrefied, partly mummified shell.

Decomposition in immersed bodies

The old rule-of-thumb that bodies decay twice as fast in air as in water is grossly inaccurate, but emphasizes the slower rate of decomposition of immersed corpses. Though in mortuary practice 'drowners' are generally regarded as being the major source of offensively rotten bodies, this is usually because discovery and recovery is far later than in deaths on dry land. Water, in fact, slows up putrefaction, mainly because of the lower ambient temperature, and protection from insect and small mammal predators.

The water also affects the usual processes of decay in that the epidermis becomes macerated and eventually detached, as described in the Chapter 16. Gas formation is the reason

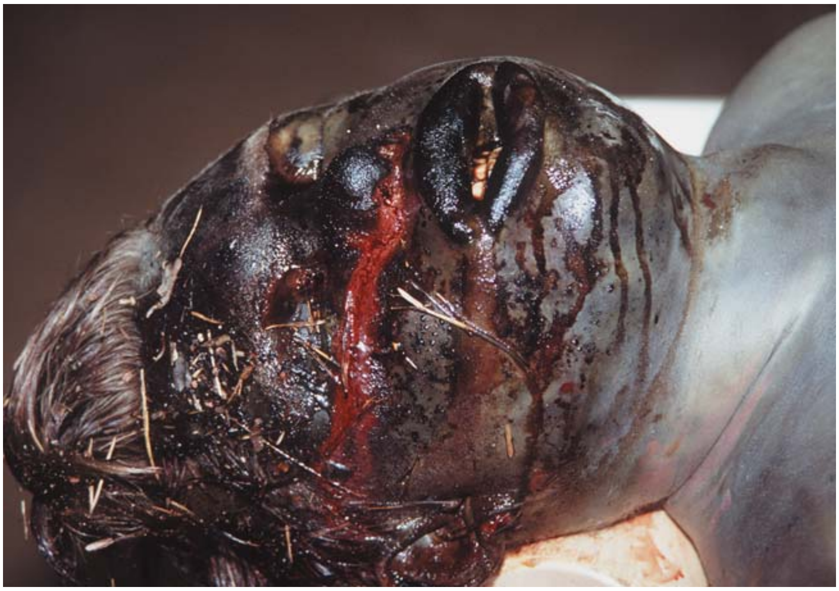

FIGURE 2.14 *Bloating of face and effusion of blood-tinged fluid from the nostrils and mouth. A body recovered from water, post-mortem time approximately 11 days.*

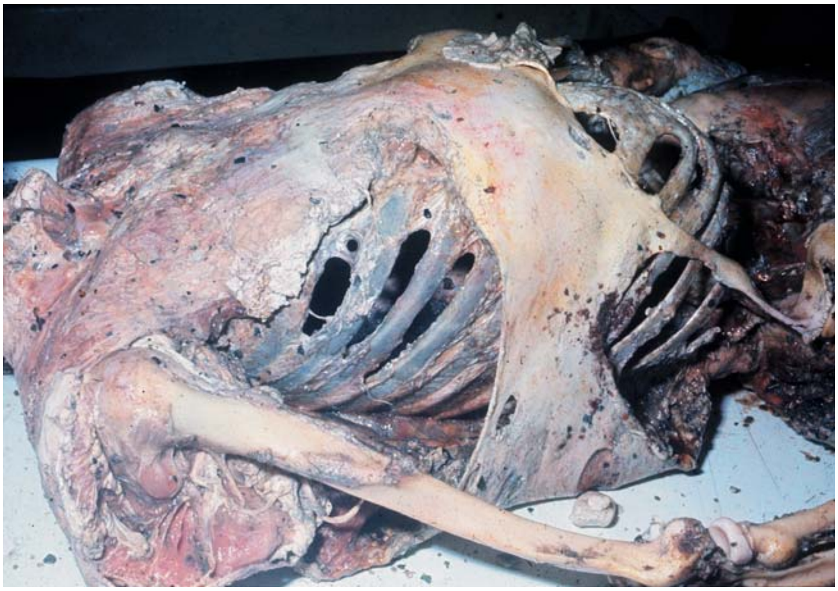

FIGURE 2.15 *Marine predation in a body after 3 months in the North Sea. The victim was from an oil rig and had floated on the surface in a life jacket. Much of the skin has been removed by crustaceans, and the arm muscles by larger fish who have cleaned out most of the body cavity.*

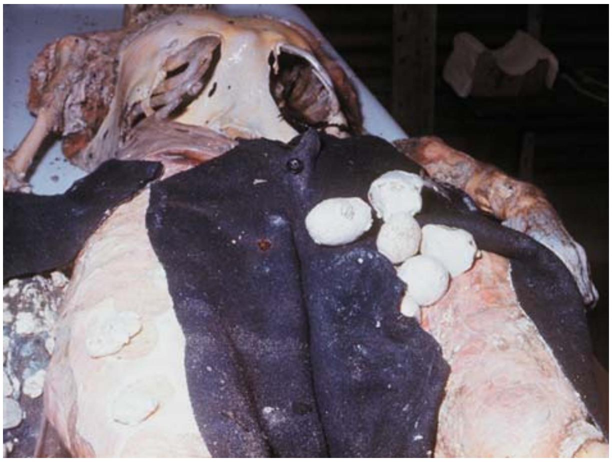

FIGURE 2.16 *Adipocere formation in a body after 3 months in the sea. Subcutaneous fat has been released from crustacean bites on the skin and has been converted into adipocere, which has been rolled by wave action within the clothing to form ovoid masses.*

for the inevitable flotation of an unweighted body, though the time of reappearance at the surface is extremely variable and certainly does not follow the speculative timetable of the older textbooks. The usual posture of a freely floating body is face down, as the head is relatively dense, and does not develop the early gas formation in the abdomen and thorax. This lower position favours fluid gravitation and hence more marked decomposition, so that the face is often badly putrefied in an immersed body, making visual recognition difficult or impossible at an early stage. As stated, temperature is the major determinant of the rate of putrefaction. Though it is usually claimed that the nature of the water in respect of sewage, for instance, is important, it is a minor factor as most of the micro-organisms responsible for decay come from the alimentary canal and respiratory passages of the body itself.

Decomposition in buried bodies

The rate of decay of bodies buried in earth is much slower than of those in either air or water. In fact the process of putrefaction may be arrested to a remarkable degree in certain conditions, allowing exhumations several years later to be of considerable value. In this respect the prospect of an exhumation should never be dismissed on the grounds that because of the lapse of time, it is bound to be worthless. It may turn out to be of little value, but this cannot be anticipated, and not infrequently the condition of the body may be surprisingly good.

The speed and extent of decay in interred corpses depends on a number of factors. If the body is buried soon after death, before the usual process of decay in air begins, putrefaction is less and may never proceed to the liquefying corruption usually inevitable on the surface. A lower temperature, exclusion of animal and insect predators, and lack of oxygen are important factors. Although most bacteria originate in the intestine, there is less access for secondary invaders and the restriction of oxygen inhibits aerobic organisms. If the body is rotting before burial then, although the process slows down, it will still severely damage the corpse, as enzymatic and bacterial growth have had initial encouragement from a higher ambient temperature and free access of air, thereby producing conditions in which secondary

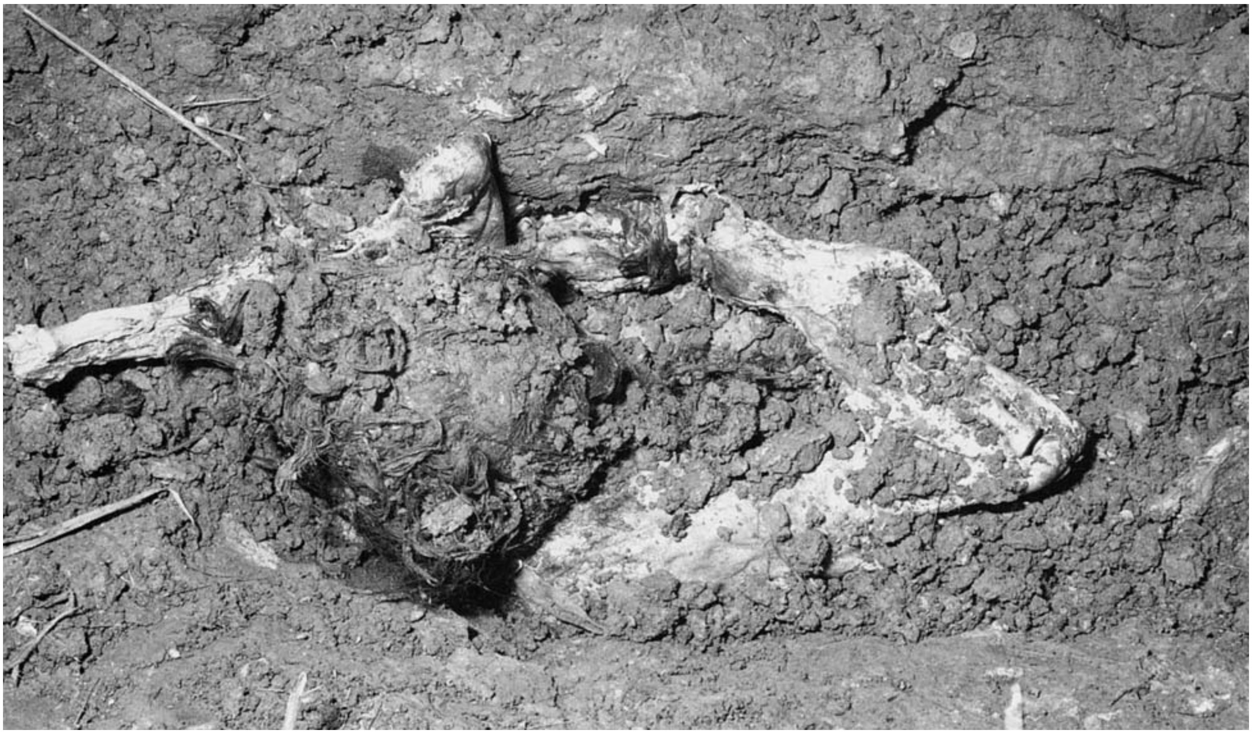

FIGURE 2.17 *Decomposition in a body buried in a shallow grave for 6 months (see Figure 1.2). The soil had collapsed into cavities around the corpse, partly due to the weight of tractors passing over it. This allowed access to flies and discovery by a dog. The body is partly putrefied, but has some adipocere.*

invaders (including anaerobes) can continue their work in a good culture medium that is already partly liquefied by the earlier stages.

Deep burial, as in the usual cemetery interment, preserves the corpse better than the shallow grave seen in some concealed homicides. The deep burial is colder (except in extremely cold weather), it excludes air better and, unless waterlogged, is not directly affected by rain.

The nature of the soil is not directly relevant except in its drainage and aeration properties. Heavy clay will exclude air and, if well above the water table, will exclude percolating surface water. By contrast, light sandy soil may allow access of both air and rainwater, but will drain more effectively. The soil factor is less important than the topography in which it lies – in a valley floor or below the water table, waterlogging is inevitable, but on a well-drained hillside the grave may remain relatively dry.

Another factor that aids the preservation of legitimately buried corpses is the coffin. Though modern coffins are often of wooden laminate or chipboard, which rapidly disintegrates when wet, any kind of coffin helps to exclude water and air for a time. A substantial, sound jointed coffin may last for years and the modern rarity of a sealed metal liner can keep a body in an excellent state of preservation for a long time.

A major factor that helps to slow decomposition is, of course, the absence of animal predation in burials. Again, if a corpse is buried before insect eggs are laid, the profound effect of maggot infestation is avoided. Rodents and larger mammals can only reach shallow burials, and in deep interments the coffin has to be breached before even the limited fauna of that stratum can gain access.

As in all putrefaction, it is unrealistic to try to construct a timetable for the stages of decay in burials. The permutations of factors mentioned above make it impossible, especially where adipocere formation takes place, as described below. The author (BK) has exhumed a chapel cemetery in the waterlogged peat of a Welsh valley and found not a trace of body, even skeleton, in graves with headstones confirming death only 20 years previously. By contrast, other bodies elsewhere were perfectly recognizable – and a standard autopsy could be carried out – one and a half years after death.

When bodies are buried in coffins in vaults, rather than in earth, then again there can be a variable rate of decay. Some bodies may develop adipocere, others may wholly or partly mummify.

There are many publications on this subject, mainly derived from the examination of historical sites or church crypts that have had to be cleared. The subject is of considerable interest but of limited forensic relevance.

FORMATION OF ADIPOCERE

An important and relatively common post-mortem change is the formation of adipocere, a waxy substance derived from the body fat. The name was given to it by Fourcroy in 1789, being derived from 'adipo' and 'cire', to indicate its affinity with both fat and wax. In most instances the change of adipocere is partial and irregular, though occasionally almost the whole body may be affected. In small amounts, adipocere is more common than usually thought in buried or concealed corpses. It is caused by hydrolysis and hydrogenation of adipose tissue, leading to the formation of a greasy or waxy substance if of recent origin. After months or years have passed, adipocere becomes brittle and chalky. The colour can vary from dead white, through pinkish, to a grey or greenish-grey. The substance itself is off-white, but staining with blood or products of decomposition can give it the red or greenish hues. The smell was accurately described by WED Evans (1962, 1963a, b) as being 'earthy, cheesy and ammoniacal'.

The chemistry of adipocere has been studied extensively. It contains palmitic, oleic and stearic fatty acids together with some glycerol, though the latter may have been leached out in older adipocere. These form a matrix for remnants of tissue fibres, nerves and muscles, which give some slight strength to the fats (see Mant and Furbank 1957). Crystals with radial markings can be found in the adipocere. Body fat at the time of death contains only about half of 1 per cent of fatty acids, but in adipocere, this may rise to 20 per cent within a month and more than 70 per cent in 3 months.

The formation of adipocere, as an alternative to total putrefaction, requires certain environmental conditions. A body left exposed in air, unless the conditions are conducive to mummification, will undergo moist putrefaction if the temperature remains above about 5–8°C. In burials, immersion in water, and in incarceration in vaults and crypts, adipocere often forms to a greater or lesser extent.

It is usually taught that moisture is necessary for the process, and it is undoubtedly a fact that most adipocere formation occurs in immersed bodies and those in wet graves and damp vaults. Numerous cases have been described, however, in which dry concealment also led to adipocere formation and here it must be assumed that the original internal body water was sufficient for the hydrolysis to proceed.

It is said that some warmth is necessary for adipocere formation, but the process seems to occur even in deep

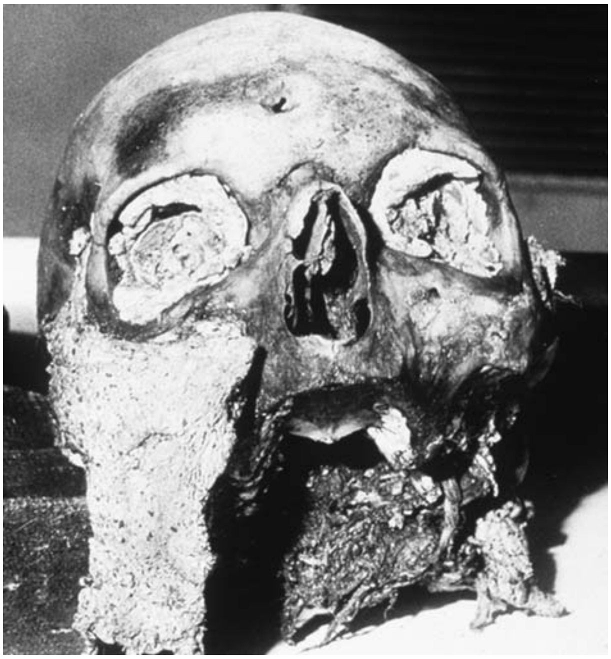

FIGURE 2.19 *Conversion of facial and orbital fat into adipocere after several months in water. There is also a penetrating wound on the forehead, obviously ante-mortem from the healing margins. The body was eventually identified by obtaining hospital radiographs of the person suspected of being the victim; these matched post-mortem radiographs in respect of the injury and frontal sinus outlines.*

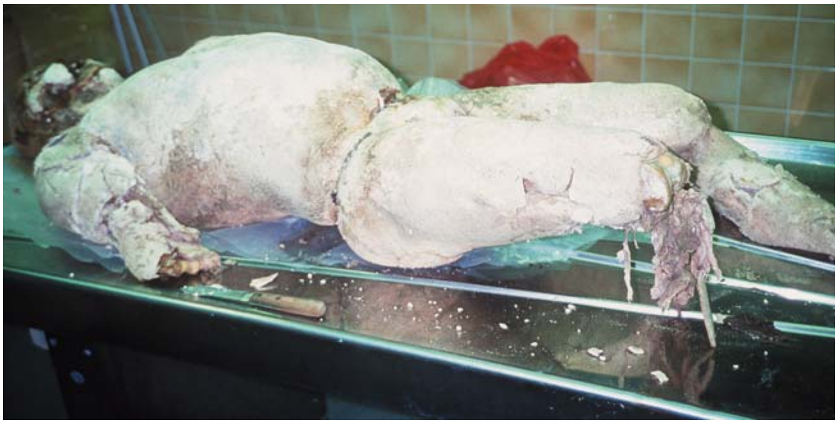

FIGURE 2.18 *Complete conversion of a body to adipocere. The body was recovered from a deep lake 8 years after disappearing in a boating accident during a heavy storm.*

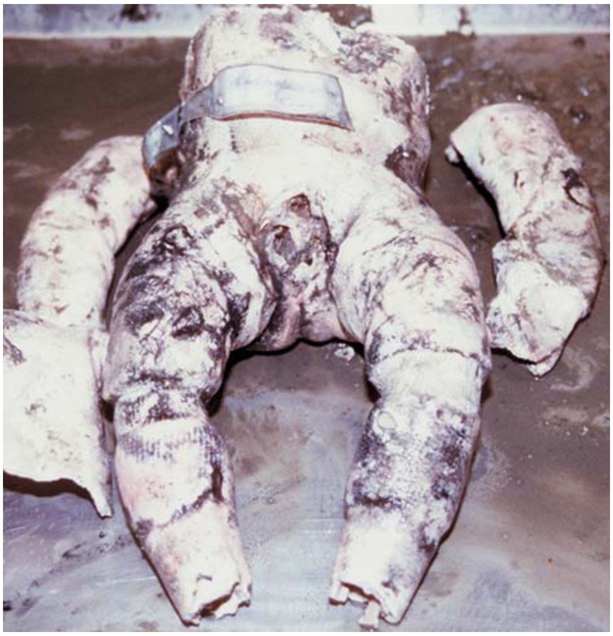

FIGURE 2.20 *Almost complete conversion of an infant's body to adipocere. The body was exhumed after 3 years' burial and was found to consist of a hollow shell of stiff adipocere with the skeletal elements loose within.*

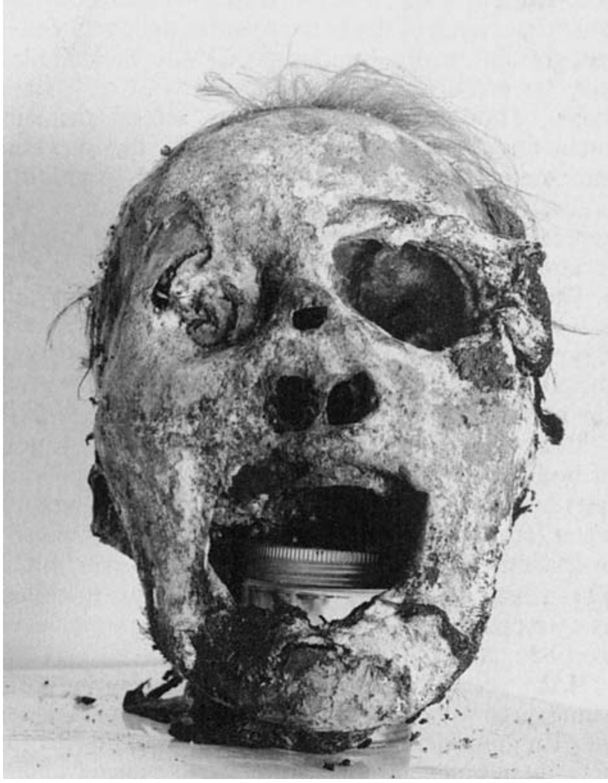

FIGURE 2.21 *Adipocere formation in body left in dry sewer tunnel for about 2 years. External water is not necessary for adipocere formation, as body fluids may be sufficient hydrolysis of fat.*

graves and in cold water. Early activity by anaerobes such as *Clostridium perfringens* assist in the reaction, as the bacteria produce lecithinase which facilitates hydrolysis and hydrogenation (see Mant 1967). Such bacteria need some warmth in the initial stages to reproduce and metabolize and Mant suggests that rapid cooling soon after death, such as by immersion in water, may actually retard the formation of adipocere.

The frequency with which adipocere forms may be gauged by the observations of those who have exhumed old cemeteries and vaults, as well as the victims of wartime slaughter (see Evans 1963a; Mant 1967; Polson *et al.* 1975; Evans 1963a). The latter found that over half the occupants of dry vaults had some adipocere. It was present in 63 per cent of women and 45 per cent of men.

The importance of adipocere formation

The process is more than a biochemical curiosity as, once it is formed, adipocere may persist for decades or even centuries. The usual dissolution of putrefaction is replaced by a permanent firm cast of the fatty tissues and, although distorted compared to the immediate post-mortem shape, it allows the form of the body and sometimes even of the facial features to be retained in recognizable form. Injuries, especially bullet holes, may be preserved in a remarkable fashion (Mant 1967). Though adipocere mainly affects the subcutaneous fat, it may also preserve the omental, mesenteric and perirenal adipose deposits; in addition, organs containing fat through pathological or degenerative processes may be preserved by adipocere forming in their parenchyma.

The time required for the production of adipocere has been a matter of controversy. The old authors claimed that 3–12 months were required, and Casper (1862) suggested 6 months, but it is obvious that the process can occur much faster than this.

There seems little reason to doubt that hydrolysis of body fat begins soon after death, as crystals of fatty acids have been found in infant bodies lying in water for only a week. This process is usually overtaken by liquefying putrefaction, however, which allows the fat, especially in obese persons, to collect as extracellular pools and pockets of triglycerides in the body cavities.

If the putrefactive process is slowed by burial or immersion, then hydrolysis and hydrogenation can outpace putrefaction. In addition, adipocere formation actually inhibits putrefaction, as the increasing acidity of tissues and the dehydration caused by the consumption of water in hydrolysis slow the growth and spread of the usual putrefactive organisms. The point at which adipocere becomes visible to the naked eye varies greatly, but it has been observed as early as 3 weeks, though 3 months is a more typical period. Adipocere is often mixed with the other forms of decomposition. Even within a coffin or other place of sequestration, there may be several different 'mini-environments'. One end of the body may be putrefied or skeletalized, whilst other parts may be mummified or in adipocere. Certain areas tend to develop adipocere, such as the cheeks, orbits, chest, abdominal wall and buttocks. Only rarely is the face preserved well enough by adipocere to be genuinely recognizable, as disintegration of the eye globes and shrinkage of the tissues around the nose and mouth obscure the most characteristic features.

MUMMIFICATION

The third type of long-term change after death is mummification, a drying of the tissues in place of liquefying putrefaction. Like the other modes of decomposition, this can be partial and can coexist with them in different areas of the same body. It is, however, more likely than the others to extend over the whole corpse.

Mummification can only occur in a dry environment, which is usually, but not exclusively, also a warm place. Mummification can occur in freezing conditions, partly because of the dryness of the air and partly because of the inhibition of bacterial growth. The most widely known forms of mummification are those in hot, desert zones. The process of artificial mummification practised for millennia in Egypt was an imitative process founded on the natural mummification of bodies in predynastic times. The essential requirement for mummification is a dry environment, preferably with a moving air current. If sterility of the tissues can be attained, as in a newborn baby, then putrefaction is held at bay whilst drying occurs. The appearances of mummification display desiccation and brittleness of the skin, which is stretched tightly across anatomical prominences such as the cheekbones, chin, costal margin and hips. The skin is discoloured (usually being brown), though secondary colonization by moulds may add patches of white, green or black.

The skin and underlying tissues are hard, making autopsy dissection difficult. The condition of the internal organs is variable, depending partly on the length of time since death. They may be partly dried, partly putrefied – and adipocere is not uncommon. In fact, slight adipocere formation is common in mummification, and perhaps the two are related in that the utilization of body water to hydrolyse fat in turn helps to dehydrate the tissues. The major factor, however, is evaporation from the surface in dry conditions.

Like adipocere, facial recognition may be possible in some instances, though the shrunken distortion and loss of eyes and full lips considerably alter the features. Mummification is likely to occur in temperate climates when the body is left undisturbed in a dry, warm place. These include closed rooms and cupboards, haylofts, attics and beneath floorboards. Because concealment favours mummification, a number of these bodies are hidden homicides – one of the best known being the 'Rhyl mummy' in North Wales, when a strangled woman was concealed in a cupboard for many years, whilst her pension continued to be claimed. Other instances are where a natural death has been concealed for some purpose. The author (BK) has also seen two deaths in hay barns, where the corpses were undisturbed for over a year.

Most experienced forensic pathologists are familiar with the mummified fetus or newborn infant, concealed in domestic circumstances, such as a house loft, where complete mummification has taken place. With small bodies such as these, evaporation is faster and more complete, and may extend

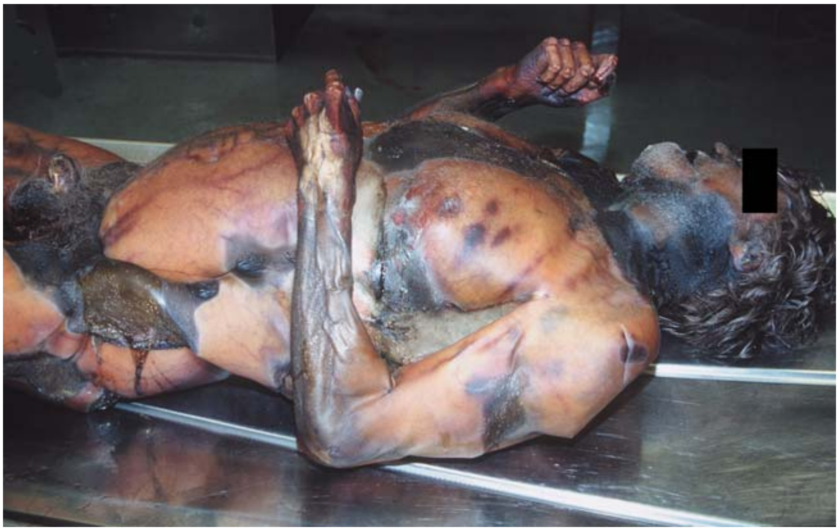

FIGURE 2.22 *Mummification in a man dead in a locked room for 10 weeks. The corpse is dry and leathery, with very little moist putrefaction.*

to all organs, especially as there would have been no invasion of putrefactive micro-organisms from the gut.

The timing of mummification is not well documented, as most mummified corpses have been concealed so well that discovery does not occur until long after the process has reached its maximum effect. It certainly takes some weeks and the early stages are often mixed with a degree of putrefactive change, especially in the internal organs.

After complete drying has taken place, the body may remain in that state for many years. Eventually, mould formation and physical deterioration progress, the dried tissues becoming split and powdery and gradually disintegrating. This process is usually hastened by animal predation. Even in the shelter of a barn or house, moths, beetles, mice and rats will wreak damage on the corpse. It will eventually skeletalize, though tough, leathery shreds of skin, tendon and ligament may persist for many years. Mummification allows major injuries to be preserved, though as in the putrefied body, the detection of bruises and abrasion may be difficult or impossible to differentiate from discoloration, artefacts and fungal damage. The stiff tissues can be softened for better examination and histology, by soaking in a 15 per cent solution of glycerine for several days (see Evans 1962).

POST-MORTEM DAMAGE BY PREDATORS

Animal predation is part of the natural food chain, which returns the proteins, fats and carbohydrates of dead bodies

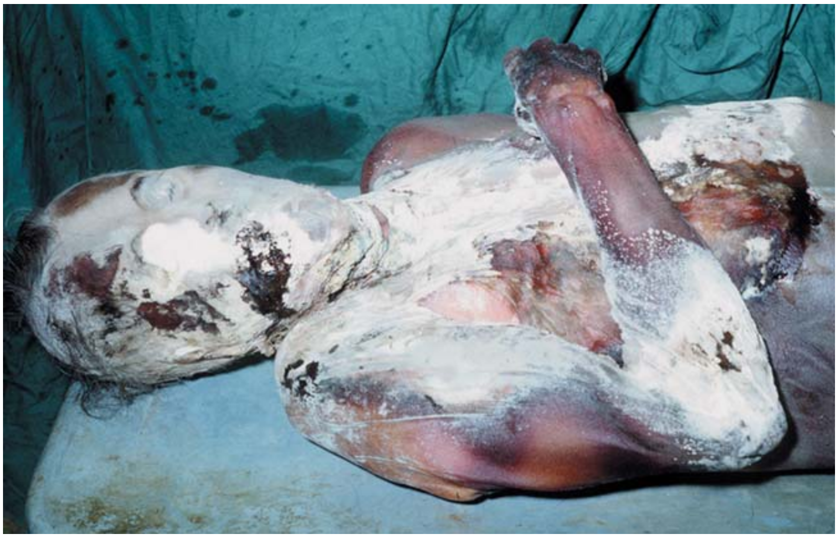

FIGURE 2.23 *The mummified body of an absconder from a mental hospital who was found a year later in a hay barn. The dry environment had inhibited wet decomposition.*

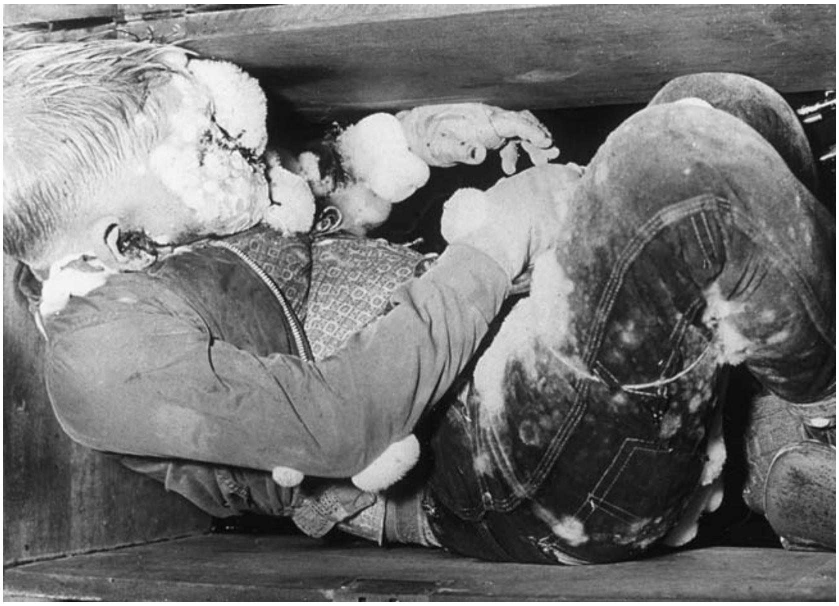

FIGURE 2.24 *Extensive mould formation and lack of putrefactive changes in a child who accidentally locked himself in a box and asphyxiated. The body was not found for 6 weeks.*

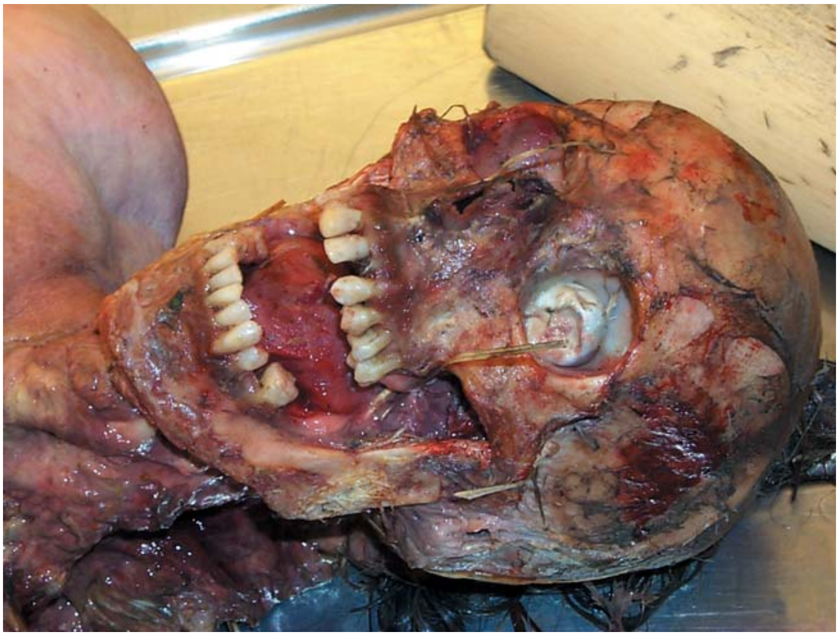

FIGURE 2.25 *Loss of all soft tissues of head and neck, in areas not covered by clothing, by post-mortem animal predation.*

to other animals, some of it passing back to the vegetable kingdom through soil nutrition. All types and sizes of animal are involved, the largest seen by the author (BK) being a 'near-miss' by a Malayan tiger, who dragged a sleeping soldier from his tent. Most animal damage is more mundane, the beasts varying from ants to foxes, and from bluebottles to monitor lizards. Foxes can drag parts of a body away to a distance of at least 2 miles.

The type of predation varies greatly with geography, season, and whether the dead body is indoors or out in the open. If it is lying in the countryside, large predators will cause prompt and severe damage, even complete destruction being possible in a short time. In temperate countries, foxes and dogs form the main agents, and the body may be dismembered and scattered over a wide area, especially if advancing putrefaction makes disarticulation easy. If corpses are in inland waters, damage from water rats and fish is possible, as well as attacks from dogs and foxes, if the body lies exposed on the bank of a river or a lake.

The type of damage from canine and rodent predators is usually obvious, as the local removal of large amounts of flesh is usually accompanied by evidence of teeth marks. Rats and cats leave a crenated edge to fairly clean-cut wounds, the post-mortem nature of which is obvious from the lack of bleeding or an inflamed marginal zone. Mice rarely attack bodies, but may help to remove the dried, crumbling tissue of a mummified corpse. The most active tissue removers are maggots, the larval stage of bluebottles (*Calliphora*) and flies (*Musca*). The use of the life cycle in timing death is dealt with elsewhere, but here we are concerned with their destructive effects on corpses. In temperate zones their activity is seasonal, but in hot countries their predatory work is ever present. The adult insects lay eggs on the fresh corpse (or even on a debilitated live victim), choosing wounds or moist areas such as the eyelids, lips, nostrils and genitalia. Once skin decomposition begins, the eggs can be deposited anywhere. The eggs hatch in a day or so, and several cycles of maggot develop, shedding their cases at intervals, depending on the species.

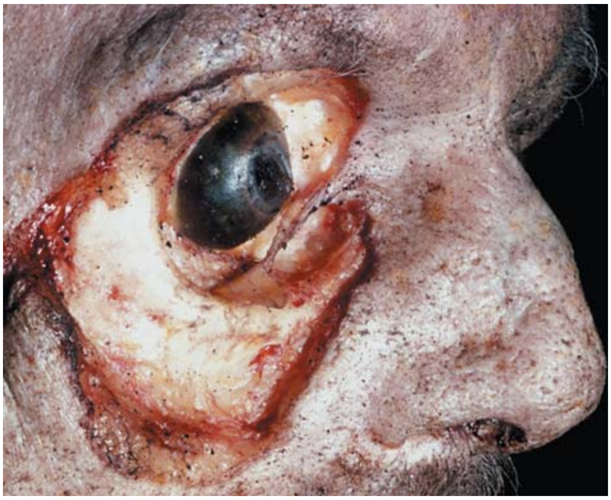

FIGURE 2.26 *Post-mortem rat bites of the orbit. Such injuries are – not unnaturally – often regarded with suspicion by the police. The complete absence of bleeding or reddening of the wound margins, as well as the unlikely shape and situation of the injury, make its post-mortem origin obvious.*

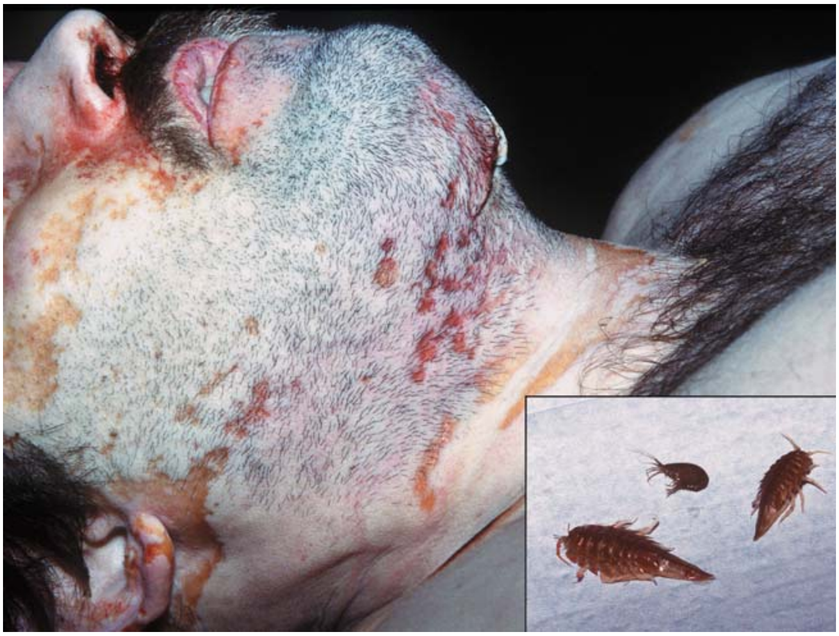

FIGURE 2.27 *A body recovered from the Baltic Sea with numerous superficial skin defects on the face and neck due to post-mortem predation, probably by the crustacean Saduria entomon, shown in the inserted picture with a smaller shrimp in the middle.*

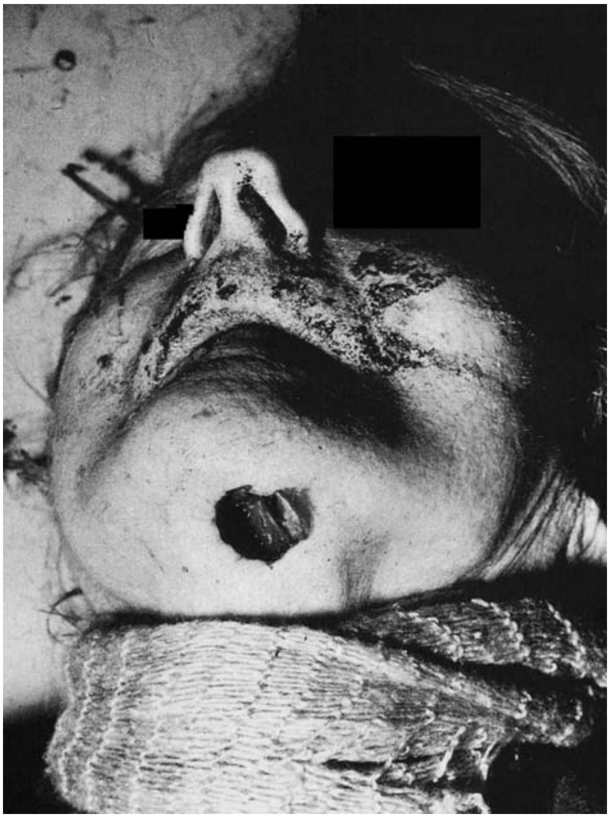

FIGURE 2.28 *A post-mortem animal bite that was mistaken for a criminal assault. The old lady was found locked in with a cat, having died from a gastric haemorrhage. The wound has not bled, the margins are not reddened and tooth crenations can be seen round its edge.*

The maggots are voracious and energetic, first exploring the natural passages, such as mouth and nostrils, then burrowing into the tissues. They secrete digestive fluids with proteolytic enzymes that help soften the tissues, burrowing beneath the skin, and making tunnels and sinuses that hasten putrefaction by admitting air and access to external micro-organisms. Successive waves of eggs are laid, producing new generations of maggots. Eventually, loss of tissue and drying make the host cadaver unattractive to egg-laying insects, and the later stages of decomposition become free from maggots.

Beetles and many other types of insect and arthropod join in the destruction, but one particular insect must be mentioned. This is the ant, which can attack the body soon after death before putrefaction begins. A common place for ant attack is around the eyelids, lips and on the knuckles. The lesions are superficial ulcers with scalloped, serpiginous margins. They can be mistaken for ante-mortem abrasions, but their position, margins and lack of bleeding or inflammatory changes usually make them easily recognizable. Shapiro, however, has documented cases where extensive linear ant lesions have resembled ligature abrasions around the neck (Shapiro *et al.* 1988).

Tropical ants can devour tissues to a considerable extent and, although stories of soldier ants reducing horses to skeletons within minutes are apocryphal, predation by large numbers of insects can be extensive.

In water, all manner of aquatic animals can mutilate immersed bodies, as well as land mammals if the corpse is in shallow water or grounded on a mudbank or the beach. Large fish can be totally destructive, especially in tropical waters, but even in temperate seas severe damage can occur. Some of the victims of the Air India aircraft disaster near Ireland in 1985 were ravaged by sharks. Large gouges from the wound edges, punctures from teeth and extensive tissue loss, even with fractures of large bones such as the femur, characterize shark attacks.

Smaller fish can wreak severe damage, the prime example being piranhas, though many species will devour both fresh and decomposing bodies. Crustaceans are also predatory and can remove circular areas of skin to expose the subcutaneous fat, which may detach inside the clothing to form masses of adipocere. Birds also cause damage, usually to the dead, though crows may mutilate live lambs. The habits of carrion birds, such as the vulture, are well known, but smaller birds will inflict injury on exposed bodies, especially when putrefaction softens the tissues.

All animal predation varies in appearance according to the size and shape of the teeth or jaws of the predator, but all such post-mortem injuries have features in common. There is no bleeding apart from the minute quantity actually in the vessels of the damaged part, and certainly no active haemorrhage into the wound margins. Naturally, there is no oedema or reddening of the edges as might be seen in an ante-mortem wound earlier than the perimortal period. Crenation of the edges is a useful guide where small rodents or other animals are involved, though large carnivores can make totally irregular tears in the tissue. Some animals, such as dogs and foxes, may leave punctured wounds adjacent to the damaged edge where sharp teeth have penetrated. On bones, the incisor teeth of rodents and larger carnivores can leave parallel gouges, which must not be mistaken for illicit human activities. In domestic surroundings, pets may inflict drastic post-mortem injuries if they are locked in with a dead person. Dogs and cats can cause injuries that puzzle those, such as the police, who are unaware of the possibility. Complete decapitation by Alsatian dogs has been seen by both of the authors, as well as genital injury.

THE ENTOMOLOGY OF THE DEAD AND POST-MORTEM INTERVAL

This is a highly specialized subject and, when the issue of time since death is important, such as a criminal investigation, it is essential that, wherever possible, the pathologist has the assistance of an entomologist with forensic experience. As in forensic toxicology and serology when serious medico-legal issues are at stake, there is no place in forensic entomology for the 'occasional expert' who dabbles in the subject when the opportunity arises. Having said this, it is helpful to know sufficient about the topic to recognize when expert help is required, and to understand what information and specimens need to be obtained.

The rationale of forensic entomology is that after death invasion of an unprotected body by sarcosaprophagous insects and other small fauna comes in successive waves. Different species of arthropods colonize the corpse at different periods after death. In addition, some species (including the most common blowflies) pass through complex life cycles that can be used to determine at least the minimum time since death by studying the stage of maturation of the insects. The science is inexact and is modified by a number of factors, both climatic and geographical, but in expert hands can sometimes yield useful information about the date of death at a period when other indicators have ceased to function. These methods have been used since the middle of the eighteenth century, but Megnin placed forensic entomology on a sound basis with his publication of *La faune des cadavres* in 1894. There are now a substantial number of publications on the subject, though most of them need specialist zoological knowledge for the interpretation of the species involved and their complicated maturation cycles. It is essential that the entomologist has the fullest possible information about the nature of the environment in which the body was found and, if possible, he should visit the scene himself. It is also vital to have accurate data on the weather, especially the ambient temperature of the area during the period in which the body had been lying there, as the maturation times of insects (especially blowflies) are markedly altered by climatic conditions.

The most common insect found on relatively fresh corpses is the blowfly, a colloquial name for a group of flying insects of the order **Diptera**, with almost worldwide distribution. They mainly comprise the bluebottles, the greenbottles and the housefly. There are numerous species, with variations in life cycles that are altered by climatic conditions. The bluebottle (*Calliphora*) is the most common, especially *Calliphora vicina* Robineau–Desvoidy (*Calliphora erythrocephala* Meigen), the most frequent invader of dead flesh. These are large, bristly flies about 6–14 mm long, with iridescent blue abdomens.

They do not fly in the dark and thus eggs are laid only in daylight. This means that a corpse found at night or in the morning with eggs upon it, almost certainly died the previous day. Bluebottles rarely fly in winter, but may do on fine days, though when the temperature is below 12°C they are unlikely to lay eggs. Rain is also a marked deterrent to egg-laying, except in the case of the **Sarcophagidae**. Bluebottles prefer fresh rather than decayed carrion and lay their eggs soon after death, which may be of medico-legal significance. Indeed, they may lay eggs on the living, especially when the victim (man or animal) is debilitated or wounded. A single bluebottle may lay up to 300–2000 eggs in clusters of

30–150. The eggs are laid on moist areas, such as the eyelids, canthi of the eyes, nostrils, lips, mouth, genitals and anus. If there are any open wounds or weeping abrasions, these may also be colonized. The eggs are yellow and banana-shaped, about l.7 mm in length.

The many tables and diagrams available for the succession of sarcosaprophagous insects on corpses must be read in the knowledge that the environment, especially the temperature, makes a marked difference to the time span of the larval and pupal stages. As Nuorteva *et al.* (1967) state:

> Crimes do not occur under experimental conditions and standardized food supplies. Flies, as medico-legal indicators, must therefore be used in conjunction with the records of meteorological conditions existing at the time subsequent to the presumed crime.

Thus the following examples, mainly taken from Glaister and Brash's (1937) account of the notorious Ruxton murder, must be read with this caution firmly in mind – and wherever possible, the pathologist must enlist the aid of an experienced entomologist.

Adult bluebottles will begin to lay eggs about 4–5 days after emergence from pupation. Eggs will not hatch at temperatures below 4°C, but they will develop at 6–7°C. The time from laying to hatching is 8–14 hours, depending on the temperature.

The first maggot stage (first 'instar') tries to penetrate the tissues or enter any nearby body cavity such as the mouth or wounds. Maggots have powerful proteolytic enzymes that aid dissolution of the tissues and facilitate penetration. The first larval instar persists for another 8–14 hours, then the outer skin is shed and the second larger instar feeds for another 2–3 days.

After a final moult, the third instar spends about 6 days on the body before leaving it to migrate some distance to hide in the ground or under some other cover (even a carpet) to pupate. The pupa is a brown leathery capsule in which the insect metamorphoses into the winged fly after about 12 days. The new bluebottle emerges and leaves the empty pupa case behind, ready to begin the cycle over again, which thus lasts about 18–24 days from egg to adult. The greenbottle (*Lucilla caesar*) and the sheep maggot fly (*Lucilla sericata*) have life histories that are similar to that of the bluebottle.

The common house fly (*Musca domestica*) is different in that, unlike the bluebottles, it prefers to lay its eggs on already decomposed flesh, though it is more attracted to garbage and manure than to cadavers. The eggs are much smaller and white rather than yellow. Hatching time is 8–12 hours, the first instar at about 36 hours, the second instar at 1–2 days and the third at about 3–4 days. Pupation lasts up to 7 days, less in warm weather. The whole cycle is thus about 14 days at about 22°C, though again it must be emphasized that variations in ambient temperature make considerable differences to the speed of maturation. A table by Busvine (1980) shows that the total cycle length varies from 9 days at 40°C to 32 days at 16°C.

Entomologists have constructed tables of variation in larval length according to age and temperature; expert knowledge and strict identification of the species is needed before definite opinions upon minimum times since cadaver death are offered. Some of the times quoted above were taken from the extensive entomological work performed on the Ruxton case in 1935, but do not correspond to the cycle periods published by Kamal in 1958. The total maturation times that he found in experimental conditions at 22°C are much longer, being 14–25 days (average 18) for *Calliphora vicina*.

In many of the published cases where *Calliphora* have assisted in the estimate of the time since death, it is an assessment of the age of larvae that is made from their size. Of course this can only be a minimum estimate, especially if mixed generations are present as a result of successive waves of flies laying eggs at different times. As already said, however, bluebottles do not favour decomposed tissue for oviposition, so egg-laying is likely to cease within a couple of days of death. If pupa cases are seen either on or near the body, then any larvae present may be from later waves – even second- or third-generation insects.

In addition to maggots on the body, other fauna in the soil beneath a corpse leave the area and a reduction in the number of species of normal soil insects may be a marker for the site of a body, even for some time after the latter has been removed. The minimum population of arthropods under a corpse is reached after about 2 months; it then revives, but with quite a different range of insects from the original inhabitants. The absence of eggs or maggots on a body may sometimes be of use, as it may indicate that a body has been in a sheltered place inaccessible to flies, even if it was later moved out. Where a body has been buried, even a shallow covering of soil may protect it from oviposition; where part of the body protrudes, this may well be the focus of maggot infestation.

Different types of insects invade the body after the moist, putrefaction stage that is associated with blowfly maggots has passed. At different seasons, different insects may be prevalent so that the finding of say, **Phoridae** and **Sphaerocerdiae** in Europe would suggest that the body had been there in the summer months. Larder beetles (*Dermestes*) may arrive 3–6 months after death; house moths may attack dried tissues such as mummies and can dispose of keratin such as hair and beard. Beetles (*Coleoptera*) are late arrivals and are usually found in dried or adipocerous corpses. These and many more are related more to season than time, so again the considered opinion of an expert is necessary. The absence of insects in certain climates may indicate that death took place in the winter months when no active colonization was occurring.

Collection of material for entomological study

To the practising forensic pathologist, the main concern in this context is the careful collection and dispatch of specimens to an entomologist with forensic experience.

First, the entomologist needs to know all available facts about the environment in which the body was found – preferably he should visit the locus himself if the matter is of legal importance. He should be told the temperature of the maggot mass and ambient temperature around the time the body was found. If no actual temperature measurements were taken by the pathologist at the scene – and even if they were – local records should be obtained from the nearest meteorological centre. Without this information the probable maturation rate cannot be estimated. The nature of the locus should be described, in terms of vegetation, trees and undergrowth if outdoors. Where maggots are present, some should be placed live in a tube and, if there is delay in transit to the entomology laboratory, a fragment of meat should be included for food. Any adult flies, other insects and eggs should also be collected without preservative.

In addition, some maggots, adults, pupae, empty pupa cases and eggs should be sent after fixation. They can be placed directly into 80 per cent alcohol, but preferably in three parts 80 per cent alcohol with one part glacial acetic acid. Even better is Pimple's fluid, which is six parts 35 per cent formalin, two parts glacial acetic acid, 15 parts 95 per cent alcohol and 30 parts water. Formol saline 10 per cent (as used for histological fixation) should not be used.

If there are different insects on the body, they should be placed in separate tubes – especially if they are alive – as some may devour the others before they reach the laboratory.

Whatever is sent should be labelled or numbered according to a key that records where on the body they were found. If outdoors, the soil beneath the body should be sampled and, where a full cycle has occurred, pupae or cases should be obtained by shallow trenching in radii around the body.

When the body has been recovered from water, any insects or crustaceans should be recovered. Body parasites may occasionally be useful. Fleas and lice can survive total immersion for some hours, but the advice of an entomologist should be sought if parasites are found on an immersed body to determine the minimum survival time of that particular species under the existing circumstances.

DETERMINATION OF TIME SINCE DEATH

ESTIMATION OF THE TIME SINCE DEATH BY BODY COOLING

Though the study of the rate of cooling of the dead human body is essentially an exercise in physiology, its potential use in determining the time since death have made it the most frequent topic for research in forensic medicine.

For more than one and a half centuries, papers have been published devoted to refining the problem further, which has obvious and important connotations in the investigation of criminal deaths. Unfortunately, the vast amount of labour in this direction has not been rewarded by comparable improvements in accuracy because of the permutations of factors that defeat exact calculation of the post-mortem interval. The history of this research is in itself extensive (Knight 1988) but a few cardinal points might be mentioned here to mark the various phases of investigation.

Though the fact that a dead body becomes progressively colder after death has naturally been known since earliest times, scientific measurements were first published in the nineteenth century. Dr John Davey in 1839 recounted experiments with dead soldiers in Malta and Britain, using a mercury thermometer. Though this pioneer made no practical contribution to the problem, some of his comments are remarkably pertinent across a gap of more than 150 years:

> Much judgement, however and nice discrimination may be requisite on the part of the medical man, in appreciating the circumstances likely to modify temperature, so as to enable him when called upon for his opinion (of the time of death), to give one which will be satisfactory to the legal officers – and to himself – on reflection.

These cautionary words are just as applicable nowadays when, unfortunately, some doctors offer a time of death with a confidence often inversely proportional to their experience. In 1863, Taylor and Wilkes wrote a classic paper which introduced many of the current concepts, such as the initial temperature plateau, the core temperature, heat gradient and the effect of insulation. Taylor was, of course, Alfred Swaine Taylor of Guy's Hospital, author of the textbook that remained the definitive work on forensic medicine for almost a century. Later in the nineteenth century, Rainy of Glasgow first applied mathematical concepts to the problem and produced a formula for calculating the time of death. He also pointed out that Newton's Law of Cooling did not apply to the human body. In 1887, Womack first used centigrade units, though his use of the now familiar temperature graphs was anticipated by Burman in 1880.

In this century, the most quoted papers were those by De Saram in Ceylon (De Saram *et al.* 1955; De Saram 1957), who published careful and detailed measurements of control cases obtained from executed prisoners. His data are still being studied and reworked. The names Fiddes and Patten are known to all those interested in this subject as, in 1958, they produced a paper that was a classic, as Taylor's had been in the previous century. Using repeated temperature measurements they devised a percentage cooling method and explored complex theoretical aspects, such as the 'infinite cylinder' model, for the human body.

Marshall and his collaborators dominated the publications in the 1960–1970 period, with papers that explored in depth the mathematical aspects and confirmed the double exponential or 'sigmoid' shape of the rectal cooling curve. In the last decade or so, many more papers and new techniques have been produced, with computer assistance increasingly used. Amongst these, Henssge and Madea in Germany have been predominant. Microwave and infrared thermography have been explored and the physics and mathematics of body cooling probed in such depth that often the scientist in the research group has difficulty in explaining the concepts to his medical collaborator. In spite of all this activity, practical methods of determining the time of death continue to lack accuracy. Though some recent publications offer firm advice and describe methods, they can do no more than provide a 'time bracket' of probability within which death is thought to have occurred.

Post-mortem cooling

Except where the environmental (ambient) temperature remains at or even above 37°C, the human body will cool after death. A uniform, homogeneous laboratory 'body' will cool according to Newton's Law of Cooling, which states that the rate of cooling is proportional to the difference in temperature between the body surface and its surroundings. Graphically represented, this will display the curve of a single exponential expression, not a straight line. A human body does not obey Newton's Law, though the size of the discrepancy varies according to several factors. When death occurs, heat transfer within the body through the circulation ceases. Metabolic heat production, occurring mainly in the muscles and liver, does not cease uniformly and some heat generation continues for a variable time. As soon as the supply of warmed blood ceases with cardiac arrest, the skin surface immediately begins to lose heat. The rate is variable because of clothing, posture and shielding against the supporting surface and, of course, the environmental temperature.

The centre or 'core' of the body cannot begin to cool until a 'temperature gradient' is set up by the cooling at the skin surface. As the tissues are poor heat conductors, this gradient takes a variable time to become established and therefore a thermometer placed near the core (usually in the rectum) will not register a fall for some time. This is the well-known 'plateau', which forms the upper flattened or slightly sloping part of the double exponential curve when

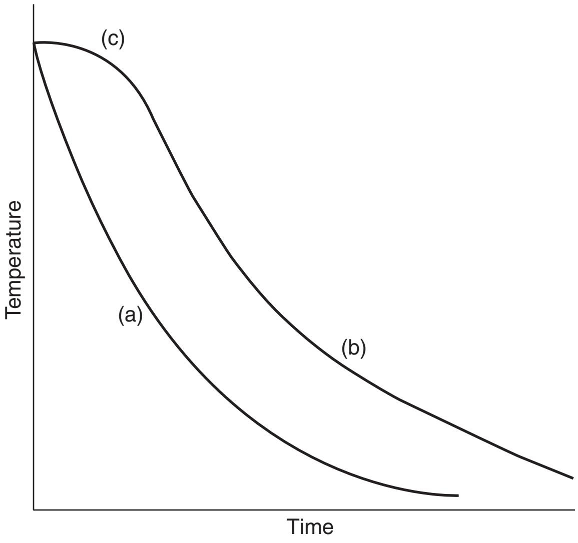

FIGURE 2.29 *Cooling curves: the Newtonian single exponential curve (a) does not occur in practice, except on the surface of the body. Because of the variable plateau (c), the true curve for deep core temperature (b) assumes a double exponential shape.*

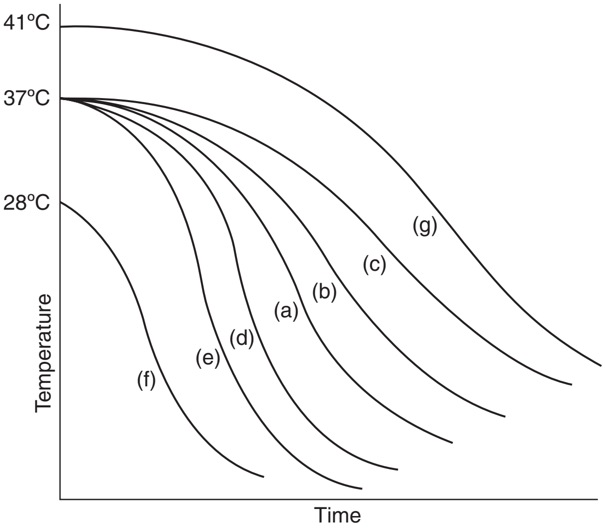

FIGURE 2.30 *Diagrammatic representation of some variables in body cooling curves: (a) average body, (b) obese body, (c) heavily clothed body, (d) thin body, (e) naked body, (f) hypothermic body and (g) febrile body.*

rectal temperatures are measured. If skin temperatures are used, as in some of the nineteenth century research, no plateau is found, and, using cranial temperatures taken through the nose, ear or skull, the plateau is less, as the core is nearer the surface.

The central part of the cooling curve approximates to Newtonian principles, being fairly straight or only a shallow curve. As the temperature differential between the body and environment approaches zero, the graph again flattens off into a lower shelf. Unlike the laboratory body, the human body rarely reaches the ambient temperature unless the latter is at or near freezing. This is probably because enzyme and bacterial action starts during early decomposition and, much as a compost heap warms up, the temperature may actually rise a few days after death.

The typical rectal cooling curve then, is a 'sigmoid' shape or 'double exponential' curve. The part that is of use in forensic medicine is the central section showing the steepest fall. Theoretically, if one assumes that the body temperature at the time of death was 37°C, then finding the point on this section corresponding to the measured rectal temperature should allow extrapolation back to the 37°C point which is at zero on the time scale, thus giving the post-mortem interval.

Unfortunately, a number of variables make this attractive proposition impossible to attain in practice. Not only are there variables, but the variables themselves often vary during the period before the body is discovered and examined. For example, the ambient temperature may change markedly – perhaps several times – and often this may not be known to the examiner. Even opening doors and allowing draughts to play on the body will have profound effects on the cooling curve, which cannot be detected or corrected retrospectively.

Factors affecting the cooling curve

INITIAL BODY TEMPERATURE

This cannot be assumed to be 37°C and in fact is incapable of ever being measured in retrospect. Not only is there a difference between the rectal, liver, brain, axillary, mouth and skin temperatures in the living person, but the absolute values vary slightly from person to person and from time to time, even in health.

If the oral temperature is taken to be 37°C, then that in the axilla will be several degrees lower and that in the rectum at least one degree higher. There is a diurnal variation of almost 1°C, the temperature being lowest between 0200 and 0600 and highest between 1600 and 1800. Strenuous exercise raises the temperature by up to 3°C, which persists for up to 30 minutes after returning to rest.

When there is illness or trauma, much wider variations occur, which again are impossible to detect retrospectively. In febrile illness from micro-organisms or parasitic infections, the temperature may be 4 or 5°C higher. In the forensic context, infected wounds or a septic abortion are obvious examples. Cerebral (especially pontine) haemorrhage may also cause hyperthermia, as may some drug reactions. It is traditionally stated in many textbooks that asphyxia and strangulation cause agonal hyperthermia, but there is very little hard evidence for this, apart from anecdotal opinions passed from one author to another. One explanation may be that a homicide victim struggling desperately for life against strangulation will generate muscular heat as in any violent exercise, irrespective of an asphyxial mode of death.

At the other end of the scale, hypothermia is common even in temperate winters. Many victims of criminal assault may be left exposed before death and their temperature may be as much as 10°C lower than normal.

As virtually all methods and formulae for calculating the time of death depend upon the body temperature being 37°C, even a slight variation can introduce a fundamental error.

THE BODY DIMENSIONS

The temperature gradient, which drives cooling, varies with the mass of the body and the surface area as well as with the conducting properties of the tissues. Some more complex calculations correct for mass:surface area ratio by means of nomograms (for example, those served by Henssge), but these can only be an approximation. The height and weight of the body must be known, which is often impossible at the scene of discovery. The amount of subcutaneous and abdominal fat will affect the insulating properties and hence the temperature gradient, but there is no way of assessing or correcting accurately for obesity. Oedema and dehydration both have a marked effect because of the high specific heat of water (see James and Knight 1965). In general, a thin person cools more quickly because of both the mass:surface area ratio and the lack of fatty insulation. Children have a larger surface area for a given body weight.

POSTURE

The loss of heat from the skin that drives the temperature gradient is affected by the access of air to the skin and the opportunity for radiation and convection. A body curled into a fetal position will expose much less surface than one in an extended, spread-eagled posture. Another factor is the amount of skin resting on the supporting surface and the nature of that surface. A body lying full length on its back will lose heat by conduction faster than one resting semi-prone, though radiation and convection may be facilitated.

A body on a metal mortuary tray will cool more quickly than one lying on straw.

CLOTHING AND COVERINGS

All too obvious is the effect of external insulation by clothing or other coverings, even a hat. Radiation is a minor pathway for heat loss because of the low biological temperatures involved, but convection and conduction are markedly reduced by coverings. Even more confusion may be added when coverings actually contribute heat, such as an electric blanket left on after death. A duvet or 'continental quilt' will markedly retard cooling and in fact will accelerate decomposition to a considerable degree. Wet clothing will accelerate cooling, compared with dry coverings, because of the uptake of heat for evaporation.

THE AMBIENT TEMPERATURE

This is, of course, the major factor in cooling and, as was said at the outset, a body will not cool after death if the temperature of the environment is higher than the nominal 37°C, in fact, it may warm up. This may be climatic and seasonal, as in large areas of the world, not necessarily in the tropics. It may also be due to local heating, usually in dwellings or other buildings. This includes radiant heat from fires left burning near the body after death, electric blankets, or death in a house or vehicle fire. Where a victim dies in a warm bath then the whole cooling process is grossly distorted, the opposite of when immersion takes place in the cold water of baths, rivers, lakes or the sea.

AIR MOVEMENT AND HUMIDITY

Most skin cooling takes place by convection and conduction with the adjacent air as the transporting medium. In still conditions, a layer of warm air clings to the skin, especially if clothed or hairy, blocking the temperature differential. Any air movement brings fresh cooler air into contact with the skin and encourages the gradient from the core. The humidity is a less active factor, but damp air conducts heat more readily than dry. A body in a small space will cool more slowly than one exposed to the open air, as transfer of heat to the small volume of air will reduce the temperature differential.

THE MEDIUM AROUND THE BODY

This is usually air, but when it is water or (rarely) another fluid, skin cooling is far more effective. A body immersed in water, especially the moving water of a river or the sea, will rapidly lose heat, as is all too familiar during life when fatal hypothermia can occur within minutes in a cold sea. It is commonly stated that cooling is less rapid in contaminated water (such as sewage) than in clean water, but this is hard to believe, given the same temperature for each type of medium. As mentioned above, death in the warm water of a bathtub reduces the cooling rate – and may even elevate the temperature, making any attempt at estimating the time of death futile.

HAEMORRHAGE

It is traditional to say that severe haemorrhage shortly before death causes more rapid cooling. As the estimation of the time of death is fraught with such inaccuracy, it is difficult to see how this opinion can have been derived. The volume of blood lost will reduce the mass of the body, but only in a minimal fashion. It may be that terminal bleeding may cause a shutdown of cutaneous circulation in an effort to maintain blood pressure and that this might encourage the early formation of the temperature gradient – but such vasoconstriction would relax immediately at death and play no part in post-mortem cooling.

Methods of measuring body temperature

Estimation of temperature by touching with the hand is a useful first manoeuvre when at the scene of a death. A hand placed on the forehead, face or exposed hand may give a first impression of whether death occurred recently or not. Even if these exposed areas are cold, feeling inside the clothing to touch the chest, abdomen or axilla may detect some heat, as may sliding a hand under the body where it is in contact with the supporting surface. These crude methods are combined with an estimate of rigor mortis to provide a preliminary screening of a recent, as opposed to a remote, time of death. Though conditions vary enormously, a body indoors will feel cold on exposed areas in 2–4 hours and in protected areas after some 6–8 hours. The traditional method of taking the post-mortem temperature is by placing a mercury thermometer in the rectum. This must be a chemical thermometer (not a clinical instrument) or thermocouple, ideally reading from 0 to 50°C. The tip must be inserted to at least 10 cm above the anus, the instrument preferably having most of its gradations still visible when *in situ*. It should be left in place for several minutes for the reading to stabilize before being recorded. Where possible, it should be left *in situ* for multiple readings at intervals, though in operational circumstances (especially in criminal deaths) this may be difficult to arrange. There is considerable controversy about when such measurements should be carried out. It is often recommended that a doctor at the scene of death should measure the rectal temperature at once but logistic difficulties exist. Many cases where estimation of the time of death

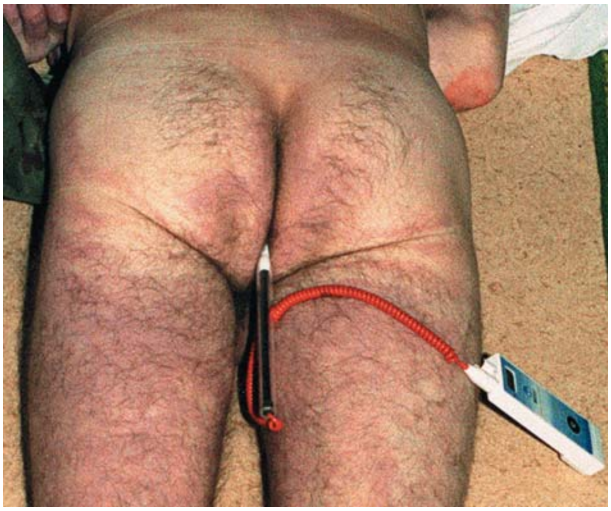

FIGURE 2.31 *Measuring the rectal temperature at the scene of a murder. This should be done only after forensic procedures such as rectal swabbing have been completed. Normally the temperature should be taken at the mortuary where removal of clothing can be carried out with full photographic and forensic science monitoring. In this case the body was unclothed at the scene and the circumstances did not warrant rectal swabbing.*

is important are criminal or suspicious deaths. These will often be associated with at least the possibility of sexual or homosexual assaults, and it may not be practicable to interfere with clothing in the perineal area or to contaminate the anal and vaginal region before full trace evidence procedures are carried out, usually by police scene of crime officers or forensic scientists, which may include adhesive taping of the clothing and skin, and the retention of underclothes for seminal stains, etc.

To wrestle with tight clothing, perhaps at night, in inclement weather and in confined spaces, and to try to introduce a thermometer into the rectum in these circumstances might ruin vital evidence of more importance than the admittedly uncertain calculations about time of death. However, many pathologists and police teams do carry out this manoeuvre as a routine, but each pathologist must decide in every individual case, whether it is preferable to wait until the body is taken to the mortuary where proper, controlled undressing is possible.

This applies where pathology and criminalistic expertise is readily available, but it is acknowledged that in less ideal circumstances, when no such expertise is likely to appear within a few hours (if at all), then some doctor at the scene should take a temperature. The procedure must be tailored to the individual circumstances but where there is a possibility of sexual interference, the rectum (and vagina) should be avoided until after full swabbing for semen and other examinations are completed. Alternatives are to use the axilla, deep nasal passage or external ear for the insertion of a thermometer.

The technique of introducing a mercury or 'rototherm' thermometer through a stab wound in the abdomen to measure liver temperature is never justified. It inevitably leads to blood contamination of the skin and clothing and also leads to intraperitoneal bleeding that might be confused with existing internal injuries.

More modern measuring devices include thermocouples, which register temperature accurately with minimum stabilization time. They may be part of a compact electronic instrument, which has a digital readout, or they may be connected to a computerized recorder that can analyse several other sites at regular intervals (see Morgan *et al.* 1988). Microwave thermography of brain and liver (see Al-Alousi *et al.* 1986, 1994) and infrared monitoring of skin temperature are at present research tools that may lead to practical devices in the future.

The use of thermometry in estimating the post-mortem interval

In spite of the great volume of research and publications already mentioned, accuracy in estimating the time since death from temperature remains elusive. The old rule-of-thumb was that the temperature fell at about 1.5°F/hour, something under 1°C/hour. Another rule-of-thumb was that the fall in °C from 37°C, plus three (to arbitrarily allow for the plateau), was equal to the time since death in hours. For example, if a rectal temperature was found to be 32°C, then 37 − 32 + 3 gave a post-mortem interval of 8 hours. The only confidence that one could place in these methods was that they were almost always wrong and that, if the answer happened to be correct, it was by chance rather than science! To hope for a linear fall was against all the principles of heat loss alluded to earlier. The first method took no account of the 'plateau' when rectal temperatures were used. This upper part of the double exponential curve is of variable length. If the body has been dead for some hours, serial temperature recordings fail to identify it, as the measurements begin on the steeper part of the curve, but where thermometry is begun soon after death, there is a variable flat area at the top of the sigmoid graph.

The length of the plateau has been discussed by several authors (see Shapiro 1965; Nokes *et al.*1985). Shapiro in particular has drawn attention to the invalidation of many formulae by the unknown length of the plateau. He claims that this can be as much as 4 hours 'and is possibly considerably longer'. Marshall and Hoare (1962) admit that the plateau may be as long as 5 hours. The plateau is the result of the lag in conducting heat from the core to the skin that is caused by the insulating properties of human tissue. This delays the establishment of the temperature gradient that must be in operation along the radius of the 'infinite cylinder', which is the model for the human body, before the core can begin to lose heat to the exterior. There may also be a small element of continuing heat production in the liver and musculature, as a number of investigators (including the first) noted that the body temperature could actually rise soon after death. As the length of the plateau is unknown in any particular case when there is a delay in starting measurements, most methods begin within this inbuilt inaccuracy.

At the lower end of the sigmoid curve, measurements again become almost useless, as when the temperature is within about 4°C of the environment, the rate of cooling becomes so slow that it provides no separation in time (see Fiddes and Patten 1958). Another empirical rule-of-thumb was devised by James and Knight (1965) to include variation in ambient temperature. The fall from 37°C was multiplied by a factor of 1, 1.25, 1.5, 1.75 or 2 for ambient temperatures of 0, 5, 10, 15 or 20°C, respectively, to give the time since death in hours. As another linear method, this was doomed to inaccuracy and was found to provide too short a post-mortem interval, though the authors emphasized that it was meant to be modified in each case by an appreciation of all other factors such as physique and clothing.

Nomogram method

At present one of the most useful practical guides is the nomogram published by Henssge. Based on conventional calculations such as those of De Saram and Marshall and backed up by a great volume of experimental data, Henssge has produced a method which can be carried out either by a simple computer program or by a nomogram. Adjustments are built in for body weight, ambient temperature, dry or wet clothing, still or moving air, or still or flowing water. The result is given within different ranges of error, with a 95 per cent probability of the true time of death falling within these ranges, which vary from 2.8 hours each side at the best estimate, down to 7 hours at worst. As with most methods, the difficulties arise in estimating the strength of the variable factors. It also cannot allow for variation of these factors over time, especially changes in ambient temperatures during the period before the body was examined. Henssge and Madea have recently refined their method to reduce the standard deviation of their nomograms so that the confidence range becomes smaller, by a whole battery of methods not dependent on temperature, such as electrical stimulation of muscle and the reaction of the eye pupils to drugs.

Another graphical method is that developed by Al-Alousi (1987) in Glasgow. In addition to his publications on microwave thermography, Al-Alousi has published a simple method based on rectal temperatures. He has constructed cooling curves for both naked and covered bodies, using the concept of 'temperature ratios'. The accompanying cooling curves may be used from a knowledge of the rectal temperature, ambient temperature and whether or not the body surface was covered. Again no correction is made for body size, posture or other variables, but the author claims that such factors made little difference to the standard errors, which he gives as shaded errors on each side of his graphs.

Multiple-site serial measurement methods

In an effort to reduce or eliminate the effect of the external variables when calculating the post-mortem interval, Morgan *et al.* (1988) have developed models using multiple sites in the body for temperature estimations. These are then taken serially at frequent intervals over a period using highly sensitive thermocouples and computerized data acquisition, and calculation, which is miniaturized sufficiently for use at scenes of crime.

The method also attempted to use the deep ear canal and nasal passages for temperature measurements, which avoids the problems of perineal contamination that occur when the rectum is used at scenes of death. The results were not encouraging, but recently French workers have revived the ear method, with claims of relative success. These methods have not been in routine use for sufficiently long to evaluate their operational accuracy.

STOMACH EMPTYING AS A MEASURE OF TIME SINCE A DEATH

With one exception, this controversial topic could be dismissed summarily as being quite irrelevant. For many years pathologists have argued over the reliability of the state of digestion of gastric contents as an indicator of the time between the last meal and death, the leading case in modern times being that of Truscott in Canada. There is now almost a consensus that with extremely circumscribed exceptions, the method is too uncertain to have much validity.

The original hypothesis was founded on the belief that food spent a fairly uniform time in the stomach before being released into the duodenum. In addition, it was claimed that its physical state was progressively altered by gastric juices and movements so that its appearance and volume was a measure of the time since it had been swallowed. Therefore, if the time of the last meal was known from circumstantial evidence, the time of death could be estimated.

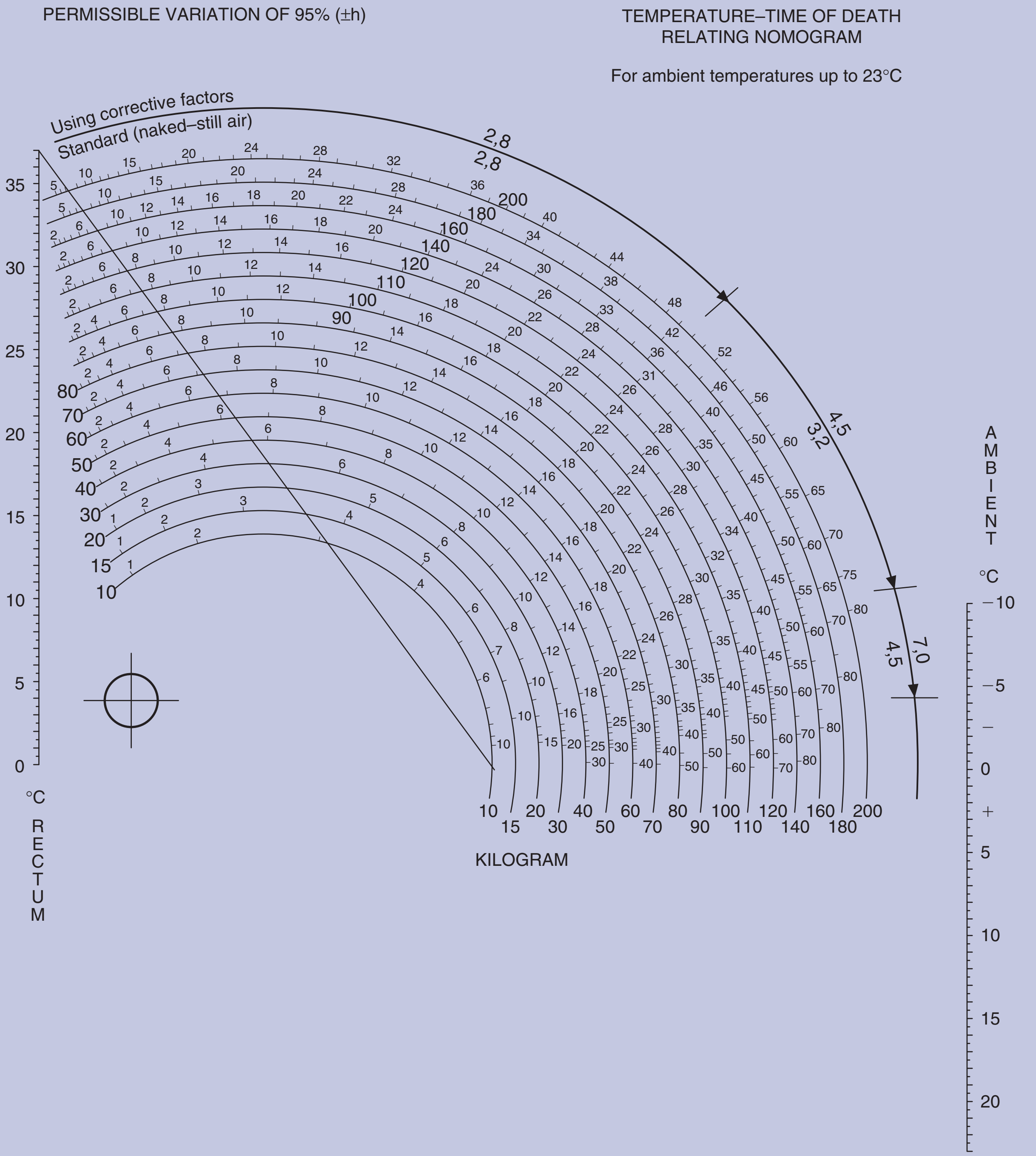

Figure 2.32 *Henssge's nomogram method for estimating the time since death from a single rectal temperature, where the environmental temperature is below 23°C. (From Henssge* et al. *2002.)*

The nomogram expresses the death time (t) by: $\frac{T_{\text{rectum}} - T_{\text{ambient}}}{37.2 - T_{\text{ambient}}} = 1.25\ \exp(Bt) - 0.25\ \exp(5Bt);\ B = -1.2815\ (\text{kg}^{-0.625}) + 0.0284$

The nomogram is related to the chosen standard, that is, naked body extended lying in still air. Cooling conditions differing from the chosen standard may be proportionally adjusted by corrective factors of the real body weight, giving the corrected body weight by which the death time is to be read off. Factors above 1.0 may correct thermal isolation conditions and factors below 1.0 may correct conditions accelerating the heat loss of a body.

How to read off the time of death

1. Connect the points of the scales by a straight line according to the rectal and ambient temperature. It crosses the diagonal of the nomogram at a special point.
2. Draw a second straight line going through the centre of the circle, below left of the nomogram and the intersection of the first line and the diagonal.

The second line crosses the semicircles, which represent the body weights. At the intersection of the semicircle of the body weight the time of death can be read off. The second line touches a segment of the outermost semicircle. Here can be seen the permissible variation of 95 per cent.

Example: temperature of the rectum 26.4°C; ambient temperature 12°C; body weight 90 kg. Result: time of death 16 ± 1.8 h. Statement: the death occurred within 13.2 and 18.8 (13 and 19) hours before the time of measurement (with a reliability of 95 per cent).

Note: if the values of the ambient temperature and/or the body weight (see 'corrective factors') are called into question, repeat the procedure with other values that might be possible. The range of death time can be seen in this way.

Empiric corrective factors of the body weight

Dry clothing/ covering	In air	Corrective factor	Wet through Clothing/covering wet body surface	In air/water
		0.35	Naked	Flowing water
		0.50	Naked	Still water
		0.70	Naked	Moving air
		0.70	1–2 thin layers clothing	Moving air
Naked	Moving	0.75		
1–2 thin layers clothing	Moving	0.90	2 or more thicker clothing	Moving air
Naked	Still	1.00		
1–2 thin layers clothing	Still	1.10	2 thicker layers clothing	Still air
2–3 thin layers clothing		1.20	More than 2 thicker layers clothing	Still air
1–2 thicker layers clothing	Moving or still	1.20		
3–4 thin layers clothing		1.30		
More thin/thicker layers clothing	Without influence	1.40		
		1.80		
Thick bedspread + clothing combined		2.40		

Note: for the selection of the corrective factor of any case, only the clothing or covering of the lower trunk is relevant.
Personal experience is needed; nevertheless, this is quickly achieved by consistent use of the method.

Requirements for use

- No strong radiation (for example, sun, heater, cooling system).
- No strong fever or general hypothermia.
- No uncertain* severe changes of the cooling conditions during the period between the time of death and examination (for example, the place of death must be the same as where the body was found).
- No high thermal conductivity of the surface beneath the body.[†]

* Known changes can be taken into account: a change of the ambient temperature can often be evaluation (for example, contact the weather station); use the mean ambient temperature of the period in question. Changes by the operations of the investigators (for example, take any cover off) since finding the body are negligible: take the conditions before into account!

[†] Measure the temperature of the surface beneath the body also. If there is a significant difference between the temperature of the air and the surface temperature, use the mean.

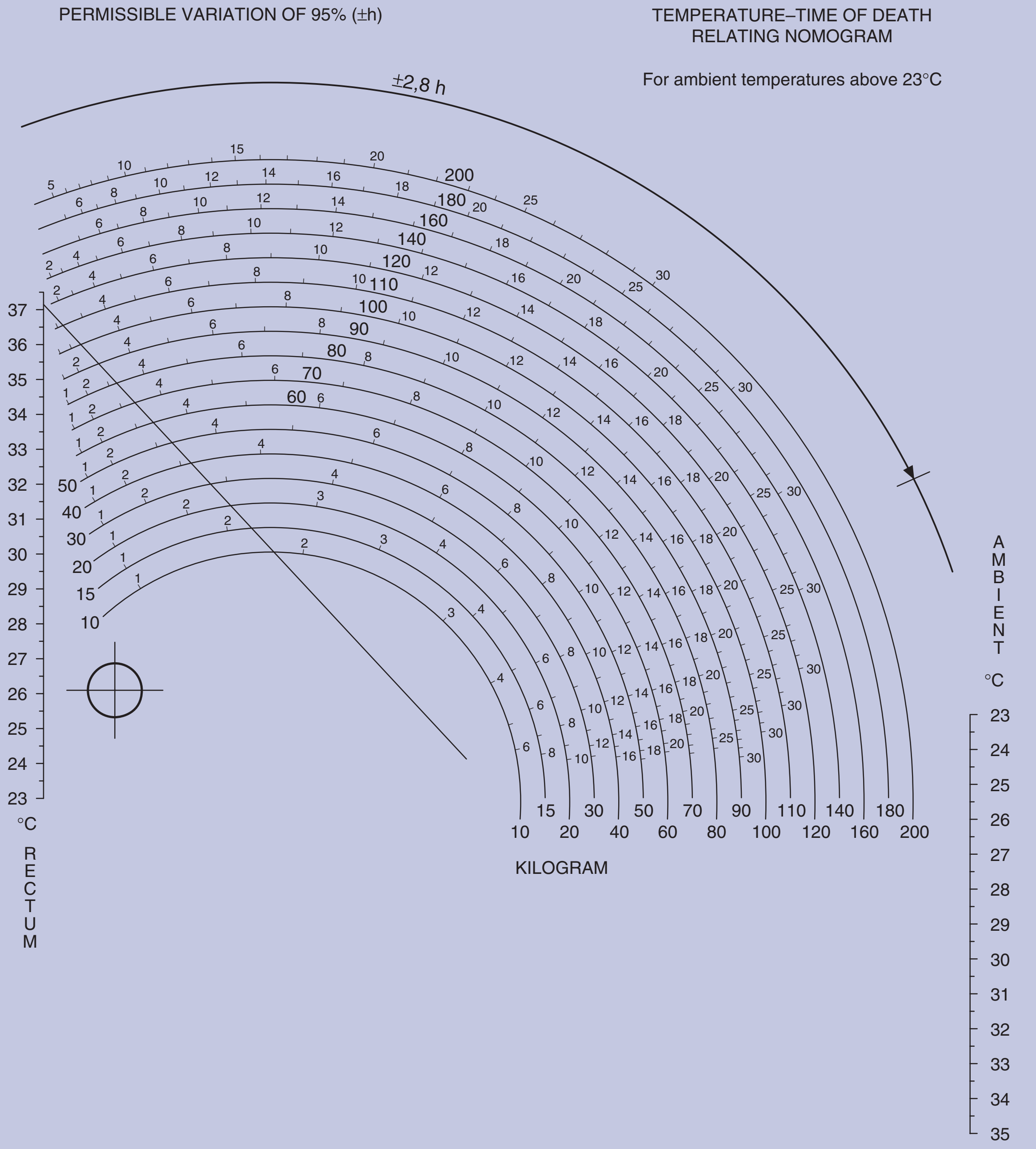

FIGURE 2.33 *Henssge's nomogram method for estimating the time since death from a single rectal temperature, where the environmental temperature is above 23°C. (From Henssge* et al. *2002.)*

The nomogram expresses the death time (t) by: $\dfrac{T_{\text{rectum}} - T_{\text{ambient}}}{37.2 - T_{\text{ambient}}} = 1.11\ \exp(Bt) - 0.11\ \exp(10Bt);\ B = -1.2815\ (\text{kg}^{-0.625}) + 0.0284$

The difference between this nomogram and that for ambient temperatures up to 23°C concerns only the relative length of the 'post-mortem plateau'. It is shorter in higher ambient temperatures (above 23°C) than in lower ones (that nomogram: up to 23°C). This nomogram for higher ambient temperatures was developed according to the data from De Saram, which were obtained in ambient temperatures between 27 and 32°C. These data show a relatively shorter length of the 'post-mortal plateau' than our own data of body cooling in lower ambient temperatures. Our own practical experience of the cooling of bodies in high ambient temperatures and therefore in using this nomogram, is small. It is recommended that the same permissible variation of 95 per cent as given in the nomogram for lower ambient temperatures is taken.

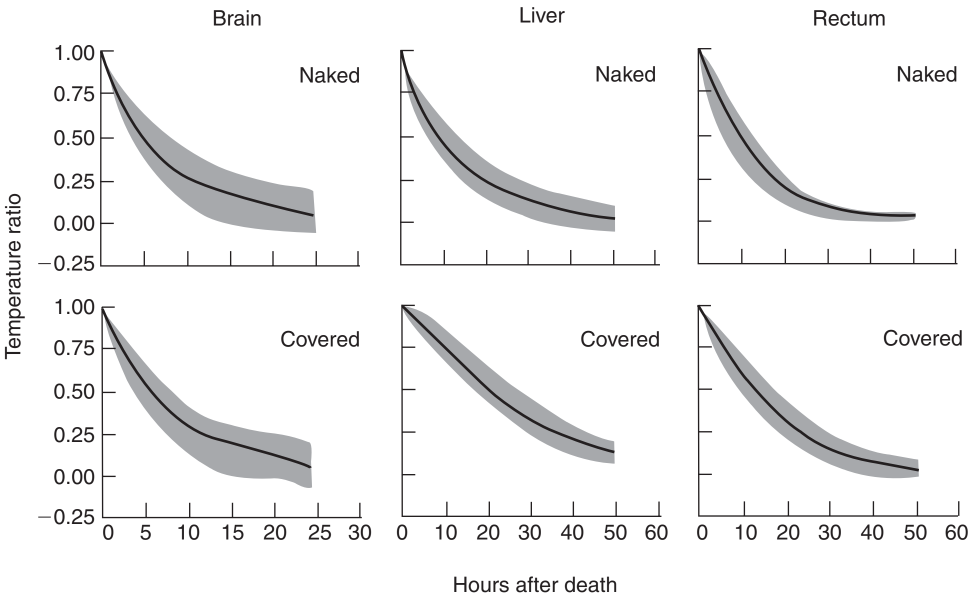

FIGURE 2.34 *Cooling curves for estimation of the post-mortem interval. In Al-Alousi's 'temperature ratio' method, the graphs above are used as follows. Assume that the body temperature (rectal) is 36.6°C at the time of death (T_b). Measure the environmental and rectal temperatures at the time of examination (T_r and T_e). Then the 'temperature ratio' is $T_r - T_e/T_b - T_e$. This ratio is then found on the vertical scale of the rectal graph, using either the graph for 'naked' or 'covered' as appropriate. The time since death is the point on the lower scale where a perpendicular falls from the point on the graph where a horizontal line meets it from the temperature ratio level. The shaded area indicates the limits of a 68 per cent probability at that particular point. For example, if the rectal temperature is 23.4°C and the air temperature 19°C, then the temperature ratio is 0.25. If the body is naked, then this gives 18 hours as the time since death, with a margin of plus 2.5 hours and minus 5 hours. (Reprinted with permission from Al-Alousi LM, Anderson RA. 1986. Microwave thermography in forensic medicine.* Police Surgeon, ***30****:30–42.)*

It was assumed that the physiological process of digestion of an 'average' meal lasted some 2–3 hours. This is based on the consumption of a test meal of gruel, however, hardly a representative example of a modern mixed diet. Moreover, the subjects of experimental work were healthy and presumably free from sudden stress during the experiments. More elaborate descriptions of digestion times of various foods have been drawn up, but they are of dubious value. As an example of the great variations offered, Modi (1957) gives 4–6 hours for a meat and vegetable meal and 6–7 hours for a farinaceous meal. Adelson (1974) stated that the stomach begins to empty within 10 minutes of swallowing, that a 'light' meal leaves the stomach by 2 hours, a 'medium' meal takes 3–4 hours and a large, heavy meal takes 4–6 hours. He noted that the head of an 'ordinary' meal usually reached the ileocaecal valve between 6 and 8 hours after ingestion. All these values, however, are subject to considerable variation.

More modern methods of investigation have used radioisotope techniques and have shown some interesting facts. When a solid meal is eaten and water drunk with it, the water leaves the stomach quickly irrespective of the nature or calorific content of the solid part. However, calorific liquids stay longer in the stomach (Brophy *et al.* 1986). The emptying rate also increases with the weight of the meal, as long as the calorific content remains constant. If the latter is increased, then the pyloric opening is delayed. Further work by Moore *et al.* (1984) again emphasized the marked variability of emptying times, even in normal circumstances.

The following factors frustrate the use of gastric emptying as a measure of time since death:

- Digestion may continue for some time after death.
- The physical nature of a meal has a profound effect on emptying time: the more fluid the consistency, the faster the emptying. Liquids entering an empty stomach pass through without any appreciable pause.
- The nature of the food modifies emptying time, notably fatty substances, which markedly delay the opening of the pylorus. Strong alcohol, such as spirits and liqueurs, also irritate the mucosa and tend to delay emptying.
- Importantly, any nervous or systemic shock or stress, mediated through the parasympathetic (vagus) system, can slow or stop gastric motility and digestive juice secretion as well as holding the pylorus firmly closed.

The intestine, as opposed to the stomach, may increase its motility with psychogenic upsets, as noted by De Saram in his autopsies on hanged criminals.

The state of digestion, as opposed to the volume and exit of food from the stomach, is almost impossible to assess. Variations in the type of food eaten make vast differences to the dissolution rate. Hard objects, such as nuts and seeds, resist digestion and may even be voided intact *per rectum*, whereas soft carbohydrates can liquefy almost immediately. The size and dispersion of particles is crucial, including the amount of chewing and admixed saliva. The efficiency of the teeth and gastric acid, enzymes and motility are also of prime importance, as is the amount of liquid either in the original food or drunk during the meal. Fat and oil in the meal slow up pyloric opening, and all these variables make it impossible to time the rate of digestion.

One of the most important factors in the forensic context is the effect of a physical or mental shock or stress during the digestion process. As stated, this can completely inhibit digestion, gastric motility and pyloric opening. The author (BK) recalls a victim of a traffic accident who lived in a coma for a week after a mortal head injury. At eventual autopsy, the large volume of gastric contents looked as fresh as if it had just been swallowed.

Even if one accepts an 'average' gastric transit time of an 'average' meal as being of the order of 2–3 hours, the assumption that death took place within this time can only be valid if the death was quite sudden and unexpected, with no stressful prodromal event. For example, if an unsuspecting person was suddenly shot or run down without warning and died almost immediately, then one can assume that the undisturbed physiological processes of digestion – albeit with its many variables – had taken place. If, however, a domestic dispute or a developing altercation culminated in a strangling or stabbing, the antecedent stresses would almost certainly affect gastric function and render invalid any interpretation of the condition of stomach contents at autopsy.

What is valid is the nature of the last meal, which may be helpful in establishing the time of death. If it was known to the investigators that the deceased person ate a certain type of meal at a certain time, whether it be chicken curry or green beans, the identification of such food in the stomach would be persuasive evidence that he died after taking that food and that it was his final meal, before any other substantial food of a different type was taken.

THE USE OF VITREOUS HUMOUR CHEMISTRY IN TIMING DEATH

This topic has been the subject of considerable research in forensic medicine over the last 30 years. It has never gained sufficient acceptance to become a routine tool, remaining a controversial procedure in spite of the large number of reports that now exists.

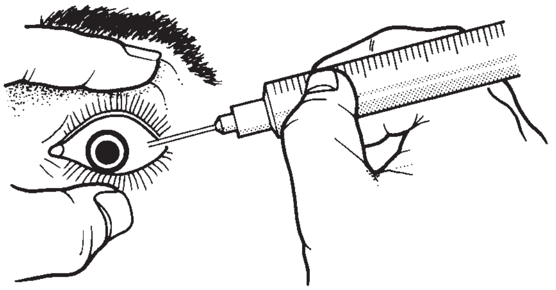

FIGURE 2.35 *Obtaining vitreous humour for analysis. The lids should be retracted and the needle introduced near the outer canthus, so that the hole will be covered when the lids are released. Fluid should be withdrawn slowly, keeping the tip of the needle in the centre of the globe to avoid dislodging the retina. Water can be reintroduced through the needle to restore the tension in the globe for cosmetic reasons.*

The most common chemical estimation performed on the vitreous fluid in the context of the post-mortem interval is that of potassium. There is a marked and progressive rise in the potassium concentration after death, the controversy revolving around whether this rise is simply linear or whether it is biphasic. The degree of confidence is also in dispute and the effect of variable factors is another contentious matter (see works by Coe, Sturner, Lange and Madea). First, the potassium values from either eye often differ, sometimes by a considerable amount. The sampling methods are then critical, as small or marginal samples vary greatly; if aspiration is forcible or from too near the retina, cellular fragments distort the values, because the potassium reaches the vitreous by leaching out from the retina. The effect of temperature changes post-mortem is also undoubtedly important. In addition, different analytical techniques used to estimate the potassium give different results, the older flame photometric methods producing a different range of values compared with modern selective-electrode procedures.

With all these variables, it is little wonder that different authors suggest widely different standard errors, varying from 1–7 hours to 26 hours either side of the true time of death. In addition, the errors seem much greater in persons dying of chronic illnesses, with nitrogen retention, compared with sudden deaths, presumably because of the premortal electrolyte disturbances in patients with metabolic disorders.

The rise in infants seems much faster than in adults – yet another variable – though in both adults and children the most potent influencing factor seems to be the post-mortem body temperature. The most popular calculation for estimating the time since death from vitreous potassium is that devised by Sturner, in which post-mortem interval =7.14 × potassium concentration (mEq/l) −39.1. This seems most satisfactory when the body has lain in an ambient temperature of not more than 10°C (50°F). The 95 per cent confidence limit during the first day is likely to be at least 4.7 hours either side of the true time and probably greater. Taking a sample from one eye at a later time than the first to attempt to calculate the rate of rise does not seem to add any accuracy to a single estimation method (see Adjutantis and Coutselenis 1972).

Hypoxanthine estimation in vitreous humour has also been used to attempt to time death, but more often as a marker of hypoxia. The results are not convincing, but there is an accumulating literature on the subject, more recent contributions being Madea *et al.* (1994) and also Rognum *et al.* (1991).

POST-MORTEM CHEMISTRY

The analysis of physiological chemical constituents of the body, as opposed to toxic substances, is often useful in investigating deaths from metabolic and biochemical disturbances.

Unfortunately, the concentration of many natural chemical substances in the dead body is rapidly distorted by post-mortem autolysis. Cell membranes become permeable to small molecules soon after the cells suffer ischaemic or anoxic damage, and agonal changes themselves may grossly alter the biochemical environment, even in the few minutes of the dying process. Thus both terminal and early post-mortem changes may render many of the analyses that are commonplace in clinical practice of little value – and indeed, quite misleading – when applied to body fluids obtained at autopsy. Some substances are more stable, however, and when results are carefully interpreted, considerable information can be obtained. Urea and creatinine are stable post-mortem, with little variation even up to 100 hours after death, so the diagnosis of ante-mortem nitrogen retention is quite reliable. The normal urea nitrogen range found in post-mortem serum is from 4.9 to 5.5 mmol/l, creatine being from about 70.7 to 212.2 μmol/l.

In the context of post-mortem chemistry the work of John Coe is best known in forensic pathology and his writings should be consulted for detailed information. Another very useful source is Madea, especially for vitreous fluid. The vitreous humour is much to be preferred to blood for post-mortem chemical analyses. Although still influenced by temperature changes, the vitreous is far less contaminated by body autolysis and is remote from the large organs and blood vessels of the abdominothoracic cavity.

There is a large body of literature on vitreous potassium, much of it centred around the controversy over the *in vivo* concentration of potassium, which can naturally rarely be measured directly in healthy persons, and the slope of the regression graph for relating potassium levels with post-mortem interval.

After death, intracellular potassium leaches from the retina through the now permeable cell membranes, into the vitreous body, naturally with an uneven distribution depending on the distance from the wall of eyeball (which is why all or a substantial proportion of the fluid should be withdrawn for analysis, to obtain a mean level). If fragments of retina are aspirated by the syringe, due to excessive suction, then a falsely elevated potassium measurement will be obtained.

Differences in the recommended regression gradient are observed between various workers: Madea states that 0.19 mmol/l/hour should be used, whereas Sturner claims 0.14 is preferable. The literature should be consulted for details (as in the 2002 book by Hennsge *et al.*); it is also pertinent to note that the methods of analysis can make a difference to the calculations. Generally speaking, the vitreous potassium method is of most use after the first 24–36 hours, when other methods have ceased to have application. Although the errors are great, some information can be derived for up to 100 hours post-mortem.

Caution must be used when interpreting results, as different analytical techniques provide different values. For instance, in relation to electrolytes, Coe and Apple (1985) state that flame photometric methods yield values about 5 mmol/l less for sodium, 7 mmol/l less for potassium and 10 mmol/l less for chloride, compared with more modern specific electrode methods. Electrolyte concentration differences between left and right vitreous humor samples were studied by Pounder *et al.* (1998) in 200 medico-legal autopsies using an ion-specific electrode system. Between-eye concentration differences of sodium and chloride were tolerable using this methodology, whereas differences in potassium, even in biochemically nonputrefied cases (potassium <15 mmol/l), were 0–2.34 mmol/l (0–21.8 per cent of mean) averaging 0.37 mmol/l (3.3 per cent), thus undermining the usefulness of vitreous potassium in estimation of time of death.

In relation to other vitreous electrolytes, the concentration of sodium and chlorides decrease after death, while potassium rises. The latter can be used as a check on the reliability of the others, as if potassium is <15 mmol/l, then the sodium and chloride concentrations may be acceptable. Chlorides decrease at less than 1 mmol/l/h and sodium by about 0.9 mmol/l/h, so the loss of this and sodium is insignificant in the first few hours, differing from potassium, which rises appreciably.

With the analysis of glucose in autopsy material, five different methods yielded five different results, presumably because interfering substances were included in varying amounts.

Returning to vitreous electrolytes, in mmol/l measured by flame photometry (in mmol/l), sodium >155, chloride > 135 and urea >40 is a reliable indication of ante-mortem dehydration. When sodium and chloride are normal but the urea exceeds 150, a diagnosis of uraemia is acceptable. These values can be distinguished from post-mortem decomposition, in which sodium is <130, chloride <105 and potassium >20.

In relation to glucose, a common problem is the autopsy diagnosis of uncontrolled diabetes and of hypoglycaemia. The vitreous glucose usually falls after death and can reach zero within a few hours. In 6000 analyses, Coe (1973) found that a vitreous glucose of more than 11.1 mmol/l was an invariable indicator of diabetes mellitus. The agonal or post-mortem rise in blood sugar is not reflected in the vitreous concentrations. In relation to hypoglycaemia, a vitreous glucose of less than 1.4 mmol/l was taken by Sturner *et al.* (1972) to be an indication of a low ante-mortem blood sugar, but others feel that whatever the concentration, no reliable interpretation can be made. In hypothermia there is also an elevated vitreous glucose, but never greater than 11.1 mmol/l.

REFERENCES AND FURTHER READING

Adelson L. 1974. *The pathology of homicide*. Thomas, Springfield.

Adelson L, Sunshine I, Rushforth NB, *et al.* 1963. Vitreous potassium concentration as an indicator of the post mortem interval. *J Forensic Sci* **8**:503–10.

Adjutantis G, Coutselinis A. 1972. Estimation of the time of death by potassium levels in the vitreous humour. *Forensic Sci* **1**:55–60.

Al-Alousi LM. 1987. *The post-mortem interval: a study of the postmortem cooling rate*. University of Glasgow, Glasgow.

Al-Alousi LM. 2002. A study of the shape of the post-mortem cooling curve in 117 forensic cases. *Forensic Sci Int* **125**:237–44.

Al-Alousi LM, Anderson RA. 1986. Microwave thermography in forensic medicine. *Police Surgeon* **30**:30–42.

Al-Alousi LM, Anderson RA, Land DV. 1994. A non-invasive method for postmortem temperature measurements using a microwave probe. *Forensic Sci Int* **64**:35–46.

Al-Alousi LM, Anderson RA, Worster DM, *et al.* 2002. Factors influencing the precision of estimating the postmortem interval using the triple-exponential formulae (TEF). Part I. A study of the effect of body variables and covering of the torso on the postmortem brain, liver and rectal cooling rates in 117 forensic cases. *Forensic Sci Int* **125**:223–30.

Ambrosi I, Carriero F. 1965. Hypostasis in the internal organs: histological aspects. *J Forensic Med* **12**:8–13.

Anonymous. 1963. Case of John David Potter (editorial). *J Forensic Med* **31**:195.

Balasooriya BA, St. Hill CA, Williams AR. 1984. The biochemistry of vitreous humour. A comparative study of the potassium, sodium and urate concentrations in the eyes at identical time intervals after death. *Forensic Sci Int* **26**:85–91.

Bate-Smith E, Bendall J. 1947. Rigor mortis and adenosine triphosphate. *J Physiol* **106**:177.

Bate-Smith E, Bendall J. 1949. Factors determining the time course of rigor mortis. *J Physiol* **110**:47.

Bendall J. 1960. *Post mortem changes in muscle*, vol. 3. Academic Press, London.

Benecke M. 1997. [Collecting insects, spiders and crustaceans for criminal forensic study.] *Arch Kriminol* **199**:167–76.

Betz P, Lignitz E, Eisenmenger W. 1995. The time-dependent appearance of black eyes. *Int J Legal Med* **108**:96–9.

Bohnert M, Weinmann W, Pollak S. 1999. Spectrophotometric evaluation of postmortem lividity. *Forensic Sci Int* **99**:149–58.

Bonte W, Sprung R, Huckenbeck W. 1986. [Problems in the evaluation of electrocution fatalities in the bathtub.] *Z Rechtsmed* **97**:7–19.

Borrman H, Du Chesne A, Brinkmann B. 1994. Medico-legal aspects of postmortem pink teeth. *Int J Legal Med* **106**:225–31.

Bray M. 1984. The effect of chilling, freezing, and rewarming on the postmortem chemistry of vitreous humor. *J Forensic Sci* **29**:404–11.

Brophy CM, Moore JG, Christian PE, *et al.* 1986. Variability of gastric emptying measurements in man employing standardized radiolabeled meals. *Dig Dis Sci* **31**:799–806.

Brown A, Marshall TK. 1974. Body temperature as a means of estimating the time of death. *Forensic Sci Int* **4**:125–33.

Brown A, Hicks B, Knight B, *et al.* 1985. Determination of time since death using the double exponential cooling model. *Med Sci Law* **25**:223–7.

Burton JF. 1974. Fallacies in the signs of death. *J Forensic Sci* **19**:529–34.

Busvine JR. 1980. *Insects and hygiene*, 3rd edn. Chapman & Hall, London.

Campobasso CP, Introna F. 2001. The forensic entomologist in the context of the forensic pathologist's role. *Forensic Sci Int* **120**:132–9.

Campobasso CP, Di Vella G, Introna F. 2001. Factors affecting decomposition and Diptera colonization. *Forensic Sci Int* **120**:18–27.

Camps FE. 1959. Establisment of time since death – a critical assessment. *J Forensic Sci* **4**:73–6.

Casper JL. 1862. *Handbook of forensic medicine*, 3rd edn. Engl transl. G W Balfour. New Sydenham Society, London.

Catts E, Goff M. 1992. Forensic entomology in criminal investigations. *Annu Rev Entomol* **37**:253–8.

Chowduri S, Chatterji PC, Banerjee SP. 1970. Histological study of liver tissue with regard to determining time of death. *J Indian Acad Forensic Sci* **9**:19–23.

Coe JI. 1969. Postmortem chemistries on human vitreous humor. *Am J Clin Pathol* **51**:741–50.

Coe JI. 1973. Further thoughts and observations on postmortem chemistry. *Forensic Sci Gaz* **5**:26.

Coe JI. 1974. Postmortem chemistry: practical considerations and a review of the literature. *J Forensic Sci* **19**:13–32.

Coe JI. 1974. Postmortem chemistries on blood with particular reference to urea nitrogen, electrolytes, and bilirubin. *J Forensic Sci* **19**:33–42.

Coe JI. 1977. Postmortem chemistry of blood, cerebrospinal fluid, and vitreous humor. *Leg Med Annu* **1976**:55–92.

Coe JI. 1989. Vitreous potassium as a measure of the postmortem interval: an historical review and critical evaluation. *Forensic Sci Int* **42**:201–13.

Coe JI. 1989. Postmortem chemistry of blood and vitreous humour in pediatric practice. In: Mason J (ed.), *Paediatric forensic medicine and pathology*. Chapman and Hall, London, pp. 191–203.

Coe JI. 1993. Postmortem chemistry update. Emphasis on forensic application. *Am J Forensic Med Pathol* **14**:91–117.

Coe JI, Apple FS. 1985. Variations in vitreous humor chemical values as a result of instrumentation. *J Forensic Sci* **30**:828–35.

Conference of Medical Royal Colleges and their Faculties in the United Kingdom. 1976. Diagnosis of brain death. *Br Med J* **2**:1187–8.

Contostavlos DL. 2001. Commentary on: Ely SF, Hirsch CS. Asphyxial deaths and petechiae. *J Forensic Sci* 2000; **45**(6):1274–7. *J Forensic Sci* **46**:1261.

De Letter EA, Piette MH. 1998. Can routinely combined analysis of glucose and lactate in vitreous humour be useful in current forensic practice? *Am J Forensic Med Pathol* **19**:335–42.

Department of Health and Social Security. 1983. *Code of practice on cadaveric organs for transplantation*. Her Majesty's Stationery Office, London.

Dervegie A. 1840. *Medicine-legale*. Baillière, Paris.

De Saram GS. 1957. Estimation of the time of death by medical criteria. *J Forensic Med* **4**:47–57.

De Saram GS, Webster G, Kathirgamatamby N. 1955. Postmortem temperature and the time of death. *J Crim Law Criminol Police Sci* **1**:562–77.

Devgun MS, Dunbar JA. 1986. Biochemical investigation of vitreus: applications in forensic medicine, especially in relation to alcohol. *Forensic Sci Int* **31**:27–34.

Eastham RD. 1971. *Biochemical values in clinical medicine*, 4th edn. John Wright, Bristol.

Easton AM, Smith KG. 1970. The entomology of the cadaver. *Med Sci Law* **10**:208–15.

Eliakis F. 1965. Determination of the hour of death by estimation of inorganic phosphorus content of the cerebro-spinal fluid. *Ann Med Leg* **45**:366–71.

Ely SF, Hirsch CS. 2000. Asphyxial deaths and petechiae: a review. *J Forensic Sci* **45**:1274–7.

Enticknap JB. 1960. Biochemical changes in cadaver sera in fatal acute heart attacks. *J Forensic Med* **7**:135–42.

Erdos T. 1943. Rigor, contracture and ATP. *Stud Ins Med Chem Univ Szeged* **3**:51–6.

Erzinclioglu YZ. 1983. The application of entomology to forensic medicine. *Med Sci Law* **23**:57–63.

Erzinclioglu YZ. 1989. Entomology, zoology and forensic science: the need for expansion. [editorial] *Forensic Sci Int* **43**:209–13.

Erzinclioglu YZ. 1989. Protocol for collecting entomological evidence. *Forensic Sci Int* **43**:211–13.

Evans WED. 1962. Some histological findings in spontaneously preserved bodies. *Med Sci Law* **2**:155–60.

Evans WED. 1963a. Adipocere formation in a relatively dry environment. *Med Sci Law* **3**:145–8.

Evans WED. 1963b. *The chemistry of death*. Thomas, Springfield.

Farmer JG, Benomran F, Watson AA, *et al.* 1985. Magnesium, potassium, sodium and calcium in post-mortem vitreous humour from humans. *Forensic Sci Int* **27**:1–13.

Fatteh A. 1966. *Estimation of time of death by chemical changes*. Commonwealth of Virginia, Office of Chief Medical Examiner, Virginia, pp. 163.

Fechner G, Koops E, Henssge C. 1984. [Cessation of livor in defined pressure conditions.] *Z Rechtsmed* **93**:283–7.

Fekete JF, Kerenyi NA. 1965. Postmortem blood sugar and blood urea nitrogen determinations. *Can Med Assoc Journal* **92**:970–74.

Fiddes F, Patten TA. 1958. Percentage method for representing the fall in body temperature after death. *J Forensic Med* **5**:2–15.

Findlay AB. 1976. Bone marrow changes in the post mortem interval. *J Forensic Sci Soc* **16**:213–18.

Fisher P. 1946. *Mechanism of death in the absence of postmortem evidence of disease or trauma. (seminar presentation)*, Department of Legal Medicine, University of Harvard, 22 October, 1946.

Foerch JS, Forman DT, Vye MV. 1979. Measurement of potassium in vitreous humor as an indicator of the postmortem interval. *Am J Clin Pathol* **72**:651–2.

Forman DT, Butts J. 1980. Electrolytes of the vitreous humor as a measure of the postmortem interval. *Clin Chem* **26**:1042–6.

Forster B. 1963. The plastic and elastic deformation of skeletal muscle in rigor mortis. *J Forensic Med* **10**:91–110.

Forster B. 1963. The contractile deformation of skeletal muscle in rigor mortis. *J Forensic Med* **10**:133–47.

Forster B. 1964. The plastic, elastic and contractile deformation of the heart muscle in rigor mortis. *J Forensic Med* **11**:148.

Gallois Montbrun FG, Barres DR, Durigon M. 1988. Postmortem interval estimation by biochemical determination in bird muscle. *Forensic Sci Int* **37**:189–92.

Gamero Lucas JJ, Romero JL, Ramos HM, *et al.* 1992. Precision of estimating time of death by vitreous potassium – comparison of various equations [see comments]. *Forensic Sci Int* **56**:137–45.

Gantner GE, Sturner W, Caffrey PR, *et al.* 1962. Ascorbic acid levels in the postmortem vitreous humor. *J Forensic Med* **9**:156–61.

Garrett G, Green MA, Murray LA. 1988. Technical method – rapid softening of adipocerous bodies. *Med Sci Law* **28**:98–9.

Glaister J, Brash J. 1937. *Medicolegal aspects of the Ruxton case*. Livingstone, Edinburgh.

Goff ML, Odom CB. 1987. Forensic entomology in the Hawaiian islands. Three case studies. *Am J Forensic Med Pathol* **8**:45–50.

Goff ML, Omori AI, Gunatilake K. 1988. Estimation of postmortem interval by arthropod succession. Three case studies from the Hawaiian Islands. *Am J Forensic Med Pathol* **9**:220–5.

Gonzales TA, Helpern M, Vance M. 1954. *Legal medicine and toxicology*, 2nd edn, vol. 64. Appleton Century Crofts, New York.

Gotouda H, Takatori T, Terazawa K, *et al.* 1988. The mechanism of experimental adipocere formation: hydration and dehydrogenation in microbial synthesis of hydroxy and oxo fatty acids. *Forensic Sci Int* **37**:249–57.

Grassberger M, Reiter C. 2002. Effect of temperature on development of the forensically important holarctic blow fly *Protophormia terraenovae* (Robineau-Desvoidy) (Diptera: Calliphoridae). *Forensic Sci Int* **128**:177–82.

Green MA, Wright JC. 1985. Postmortem interval estimation from body temperature data only. *Forensic Sci Int* **28**:35–46.

Green MA, Wright JC. 1985. The theoretical aspects of the time dependent Z equation as a means of postmortem interval estimation using body temperature data only. *Forensic Sci Int* **28**:53–62.

Haglund WD, Sperry K. 1993. The use of hydrogen peroxide to visualize tattoos obscured by decomposition and mummification. *J Forensic Sci* **38**:147–50.

Hanson J, Huxley H. 1955. The structural basis of contraction in striated muscle. *Symp Soc Exp Biol* **9**:228–64.

Hansson L, Uotila U, Lindfors R, *et al.* 1966. Potassium content of the vitreous body as an aid in determining the time of death. *J Forensic Sci* **11**:390–4.

Hellerich U, Bohnert M, Pollak S. 2001. [Hypostasis-induced changes in the breast area.] *Arch Kriminol* **207**:162–9.

Henry JB, Smith FA. 1980. Estimation of the postmortem interval by chemical means. *Am J Forensic Med Pathol* **1**:341–7.

Henssge C. 1988. Death time estimation in case work. I. The rectal temperature time of death nomogram. *Forensic Sci Int* **38**:209–36.

Henssge C, Madea B, Gallenkemper E. 1985. [Determination of the time of death – integration of various partial methods.] *Z Rechtsmed* **95**:185–96.

Henssge C, Madea B, Gallenkemper E. 1988. Death time estimation in case work. II. Integration of different methods. *Forensic Sci Int* **39**:77–87.

Henssge C, Knight B, Krompecher T, *et al.* 1995. *The estimation of the time since death in the early postmortem period*. Edward Arnold, London.

Henssge C, Althaus L, Bolt J, *et al.* 2000. Experiences with a compound method for estimating the time since death. II. Integration of non-temperature-based methods. *Int J Legal Med* **113**:320–31.

Henssge C, Knight B, Krompecher T, *et al.* 2002. *The estimation of the time since death in the early postmortem period*, 2nd edn. Arnold, London.

Henssge C, Wang H, Hoppe B. 2002. Light microscopical investigations on structural changes of skeletal muscle as artifacts after postmortem stimulation. *Forensic Sci Int* **125**:163–71.

Hill EV. 1941. Significance of dextrose and nondextrose reducing substances in postmortem blood. *Arch Pathol* **32**:452–6.

Hiraiwa K, Ohno Y, Kuroda F, *et al.* 1980. Estimation of postmortem interval from rectal temperature by use of computer. *Med Sci Law* **20**:115–25.

Hughes WM. 1965. Levels of potassium in the vitreous humour after death. *Med Sci Law* **5**:150–6.

Inoue M, Suyama A, Matuoka T, *et al.* 1994. Development of an instrument to measure postmortem lividity and its preliminary application to estimate the time since death. *Forensic Sci Int* **65**:185–93.

Introna F, Jr, Campobasso CP, Di Fazio A. 1998. Three case studies in forensic entomology from southern Italy. *J Forensic Sci* **43**:210–14.

Jaafar S, Nokes L. 1994. Examination of the eye as a means to determine the early postmortem period: a review of the literature. *Forensic Sci Int* **64**:185–9.

Jaffe F. 1962. Chemical postmortem changes in the intraocular fluid. *J Forensic Sci* **7**:231.

James RA, Hoadley PA, Sampson BG. 1997. Determination of postmortem interval by sampling vitreous humour. *Am J Forensic Med Pathol* **18**:158–62.

James WR, Knight B. 1965. Errors in estimating the time of death. *Med Sci Law* **5**:111–16.

Jit I, Sehgal S, Sahni D. 2001. An indian mummy: a case report. *Forensic Sci Int* **117**:57–63.

Johnston W, Villeneuve G. 1897. On the medicolegal applications of entomology. *Montreal Med* **26**:81–8.

Joseph AE, Schickele E. 1970. A general method for assessing factors controlling postmortem cooling. *J Forensic Sci* **15**:364–91.

Kamal AS. 1958. Comparative study of thirteen species of sacrosaprophagous Calliphoridae Sarcophagidae (Diptera). 1 Bionomics. *Ann Entomol Soc Am* **51**:261–71.

Kevorkian J. 1956. The fundus oculi and the determination of death. *Am J Pathol* **32**:1253–67.

Kevorkian J. 1961. *The eye in death*. Ciba Symposia, London.

Kevorkian J. 1961. The fundus oculi as a postmortem clock. *J Forensic Sci* **6**:261.

Knight B. 1968. Estimation of the time since death: a survey of practical methods. *J Forensic Sci Soc* **8**:91–6.

Knight B. 1979. The putrefied body. *Br Med J* **1**:1300–1.

Knight B. 1988. The evolution of methods for estimating the time of death from body temperature. *Forensic Sci Int* **36**:47–55.

Kominato Y, Harada S, Yamazaki K, *et al.* 1988. Estimation of postmortem interval based on the third component of complement (C3) cleavage. *J Forensic Sci* **33**:404–9.

Komura S, Oshiro S. 1977. Potassium levels in the aqueous and vitreous humor after death. *Tohoku J Exp Med* **122**:65–8.

Krompecher T. 1994. Experimental evaluation of rigor mortis. VIII. Estimation of time since death by repeated measurements of the intensity of rigor mortis on rats. *Forensic Sci Int* **68**:149–59.

Krompecher T, Bergerioux C. 1988. Experimental evaluation of rigor mortis. VII. Effect of ante- and post-mortem electrocution on the evolution of rigor mortis. *Forensic Sci Int* **38**:27–35.

Kuehn LA, Tikuisis P, Livingstone S, *et al.* 1979. Body cooling after death. *J Can Soc Forensic Sci* **12**:153–63.

Kuehn LA, Tikuisis P, Livingstone S, *et al.* 1980. Body cooling after death. *Aviat Space Environ Med* **51**:965–9.

Kulshrestha P, Chandra H. 1987. Time since death. An entomological study on corpses. *Am J Forensic Med Pathol* **8**:233–8.

Kulshrestha P, Satpathy DK. 2001. Use of beetles in forensic entomology. *Forensic Sci Int* **120**:15–7.

Lange N, Swearer S, Sturner WQ. 1994. Human postmortem interval estimation from vitreous potassium: an analysis of original data from six different studies. *Forensic Sci Int* **66**:159–74.

Langford AM, Pounder DJ. 1997. Possible markers for postmortem drug redistribution. *J Forensic Sci* **42**:88–92.

Langley RL. 1994. Fatal animal attacks in North Carolina over an 18-year period. *Am J Forensic Med Pathol* **15**:160–7.

Lasczkowski G, Riepert T, Rittner C. 1993. [Viewing the criminal site in unusual cadaver evidence: concealment, identification, determining time of death and reconstruction.] *Arch Kriminol* **192**:1–11.

Leadbeatter S, Knight B. 1987. The history and the cause of death. *Med Sci Law* **27**:132–5.

Leahy MS, Farber ER. 1967. Postmortem chemistry of human vitreous humor. *J Forensic Sci* **12**:214–22.

Leclercq M, Tinant-Dubois J. 1973. Entomologie et medicine legale. [Entomology and forensic medicine: dating of a death.] *Bull Med Legal Tox* **16**:251–5.

Lie JT. 1967. Changes of potassium concentration in the vitreous humor after death. *Am J Med Sci* **254**:136–43.

Lord WD, Burger JF. 1983. Collection and preservation of forensically important entomological materials. *J Forensic Sci* **28**:94–144.

Lorenzen GA, Lawson RL. 1971. A possible new approach for determining the postmortem interval by sperm motility. *J Crim Law Criminol* **62**:560–3.

Lothe F. 1964. The use of larval infestation in detemining time of death. *Med Sci Law* **4**:113–16.

Lucy D, Aykroyd R, Pollard M. 2001. Commentary on: Munoz JI, Suarez-Penaranda JM, Otero XL, Rodriguez-Calvo MS, Costas E, Miguens X, Concheiro L. A new perspective in the estimation of postmortem interval (PMI) based on vitreous. *J Forensic Sci* **46**:1527–8.

Lundquist F. 1956. Physical and chemical methods for estimating the time since death. *Acta Med Leg Soc* **9**:205–13.

Lyle H, Cleveland F. 1956. Determination of time since death by heat loss. *J Forensic Sci* **1**:11–24.

Madea B, Henssge C. 1990. Electrical excitability of skeletal muscle postmortem in casework. *Forensic Sci Int* **47**:207–27.

Madea B, Henssge C, Honig W, *et al.* 1989. References for determining the time of death by potassium in vitreous humor. *Forensic Sci Int* **40**:231–43.

Madea B, Herrmann N, Henssge C. 1990. Precision of estimating the time since death by vitreous potassium – comparison of two different equations. *Forensic Sci Int* **46**:277–84.

Madea B, Kaferstein H, Hermann N, *et al.* 1994. Hypoxanthine in vitreous humor and cerebrospinal fluid – a marker of postmortem interval and prolonged (vital) hypoxia? Remarks also on hypoxanthine in SIDS. *Forensic Sci Int* **65**:19–31.

Madea B, Kreuser C, Banaschak S. 2001. Postmortem biochemical examination of synovial fluid – a preliminary study. *Forensic Sci Int* **118**:29–35.

Mallach HJ. 1964. Zur Frage der Todeszeitbestimmung. [Definition of the time of death.] *Berlin Med* **18**:577–82.

Mant AK. 1953. Factors influencing changes after burial. In: Simpson K (ed.), *Modern trends in forensic medicine*. Butterworth, London.

Mant AK. 1967. Adipocere. In: Simpson K (ed.), *Modern trends in forensic medicine*, 2nd edn. Butterworth, London.

Mant AK, Furbank R. 1957. Adipocere – a review. *J Forensic Med* **4**:18–22.

Marshall TK. 1960. *The cooling of the body after death*. University of Leeds, Leeds.

Marshall TK. 1962. Estimating the time of death. *J Forensic Sci* **7**:210–21.

Marshall TK. 1965. Temperature methods of estimating the time of death. *Med Sci Law* **5**:224–32.

Marshall TK. 1969. The use of body temperature in estimating the time of death and its limitations. *Med Sci Law* **9**:178–82.

Marshall TK, Hoare F. 1962. Estimating the time of death – the rectal cooling after death and its mathematical representation. *J Forensic Sci* **7**:56–81.

Mason J, Klyne W, Lennox B. 1951. Potassium levels in the CSF after death. *J Clin Pathol* **4**:231.

Maxeiner H, Winklhofer A. 1999. [Eyelid petechiae and conjunctival hemorrhage after cardiopulmonary resuscitation] *Arch Kriminol* **204**:42–51.

Megnin P. 1894. *La faune des cadavres*. Gauthier-Villars, Paris.

Modi JP. 1957. *Medical jurisprudence and toxicology*. Tripathi, Bombay, India.

Mole RH. 1948. Fibrinolysin and the fluidity of blood postmortem. *J Pathol Bacteriol* **60**:413–19.

Moore JG, Christian PE, Brown JA, *et al.* 1984. Influence of meal weight and caloric content on gastric emptying of meals in man. *Dig Dis Sci* **29**:513–19.

Morgan C, Nokes LD, Williams JH, *et al.* 1988. Estimation of the post mortem period by multiple-site temperature measurements and the use of a new algorithm. *Forensic Sci Int* **39**:89–95.

Morovic-Buda KA. 1965. Experiences in the process of putrefaction in corpses buried in earth. *Med Sci Law* **5**:403.

Munoz JI, Suarez-Penaranda JM, Otero XL, *et al.* 2001. A new perspective in the estimation of postmortem interval (PMI) based on vitreous. *J Forensic Sci* **46**:209–14.

Munoz-Barus JI, Suarez-Penaranda J, Otero XL, *et al.* 2002. Improved estimation of postmortem interval based on differential behaviour of vitreous potassium and hypoxantine in death by hanging. *Forensic Sci Int* **125**:67–74.

Myskowiak JB, Doums C. 2002. Effects of refrigeration on the biometry and development of *Protophormia terraenovae* (Robineau-Desvoidy) (Diptera: Calliphoridae) and its consequences in estimating post-mortem interval in forensic investigations. *Forensic Sci Int* **125**:254–61.

Nanikawa R, Tawa N, Saito K. 1961. Chemical studies on adipocere formation: revaluation of the saponification theory. *Jap J Legal Med* **15**:258–64.

Nauman HN. 1959. Postmortem chemistry of the vitreous body in man. *Arch Ophthalmol* **62**:356–63.

Nicati S. 1894. Loss of eye tension as a sign of death. *Med Rec* **45**:4801.

Niderkorn PF. 1874. Rigor mortis. *Br Med J* **1**:303–4.

Nokes LD, Brown A, Knight B. 1983. A self-contained method for determining time since death from temperature measurements. *Med Sci Law* **23**:166–70.

Nokes LD, Hicks B, Knight BH. 1985. The post-mortem temperature plateau – fact or fiction? *Med Sci Law* **25**:263–4.

Nokes LD, Hicks B, Knight B. 1986. The use of trachea temperature as a means of determining the post-mortem period. *Med Sci Law* **26**:199–202.

Nokes LD, Barasi S, Knight BH. 1989. Case report: co-ordinated motor movement of a lower limb after death. *Med Sci Law* **29**:265.

Nuorteva P. 1974. Age determination of a blood stain in a decaying shirt by entomological means. *Forensic Sci* **3**:89–94.

Nuorteva P. 1977. *Sarcosaprophagous insects as forensic indicators*. In: Tedeschi CG, Eckert WG (eds), *Forensic medicine*. W. B. Saunders, Philadelphia, vol. II.

Nuorteva P, Isokoski M, Laiho K. 1967. Forensic entomology. *Ann Entomol Fenn* **33**:217–20.

Ohya I. 1994. [Some findings of the lung in medicolegal autopsy cases.] *Nippon Hoigaku Zasshi* **48**:379–94.

Osuna E, Garcia-Villora A, Perez-Carceles MD, *et al.* 1999. Vitreous humor fructosamine concentrations in the autopsy diagnosis of diabetes mellitus. *Int J Legal Med* **112**:275–9.

Osuna E, Garcia-Villora A, Perez-Carceles M, *et al.* 2001. Glucose and lactate in vitreous humor compared with the determination of fructosamine for the postmortem diagnosis of diabetes mellitus. *Am J Forensic Med Pathol* **22**:244–9.

Pallis C. 1983. *ABC of brain death*. British Medical Association, London.

Polson CJ, Marshall TK. 1975. *The disposal of the dead*, 3rd edn. English Universities Press, London.

Polson CJ, Gee D, Knight B. 1975. *The essentials of forensic medicine*, 4th edn. Pergamon Press, Oxford.

Pounder DJ, Carson DO, Johnston K, *et al.* 1998. Electrolyte concentration differences between left and right vitreous humor samples. *J Forensic Sci* **43**:604–7.

Pounder DJ, Stevenson RJ, Taylor KK. 1998. Alcoholic ketoacidosis at autopsy. *J Forensic Sci* **43**:812–16.

Prinsloo I, Gordon I. 1951. Post-mortem dissection artefacts of the neck and their differentiation from ante-mortem bruises. *S Afr Med J* **25**:358–61.

Ramu M, Robinson AE, Camps FF. 1969. The evaluation of post-mortem blood sugar levels and their correlation with the glycogen content of the liver. *Med Sci Law* **9**:23–6.

Rognum TO, Hauge S, Oyasaeter S, *et al.* 1991. A new biochemical method for estimation of postmortem time. *Forensic Sci Int* **51**:139–46.

Roll P, Rous F. 1991. Injuries by chicken bills: characteristic wound morphology. *Forensic Sci Int* **52**:25–30.

Rossi ML, Shahrom AW, Chapman RC, *et al.* 1994. Postmortem injuries by indoor pets. *Am J Forensic Med Pathol* **15**:105–9.

Salisbury CR, Melvin GS. 1936. Ophthalmoscopic signs of death. *Br Med J* **1**:1249–51.

Schleyer F. 1963. Determination of the time of death in the early post-mortem interval. In: Lundquist F (ed.), *Methods of forensic science*. Interscience Publishers, John Wiley and Sons, New York, pp. 253–93.

Schleyer F. 1975. Leichenveränderungen. Todeszeitbestimmung im früh-postmortalen Intervall. In: Müller B (ed.), *Gerichtliche Medizin*. Springer-Verlag, Berlin.

Schuller E, Pankratz H, Liebhardt E. 1987. [Colorimetry of livor mortis.] *Beitr Gerichtl Med* **45**:169–73.

Scott KS, Oliver JS. 2001. The use of vitreous humor as an alternative to whole blood for the analysis of benzodiazepines. *J Forensic Sci* **46**:694–7.

Sellier K. 1958. Determination of the time of death by extrapolation of the temperature decrease curve. *Acta Med Leg Soc* **2**:279–301.

Shapiro HA. 1950. Rigor mortis. *Br Med J* **2**:304.

Shapiro HA. 1954. Medico-legal mythology; the time of death. *J Forensic Med* **1**:144–69.

Shapiro HA. 1965. The post-mortem temperature plateau. *J Forensic Med* **12**:137–41.

Shapiro HA, Gordon I, Benson SD. 1988. *Forensic medicine – a guide to principles*, 3rd edn. Churchill Livingstone, Edinburgh.

Shattock FM. 1968. Injuries caused by wild animals. *Lancet* **2**:412–14.

Shikata I. 1958. On the mortality of the leucocyte and the time lapse after death. *Jap J Legal Med* **12**:227–31.

Simonsen J, Voigt J, Jeppesen N. 1977. Determination of the time of death by continuous post-mortem temperature measurements. *Med Sci Law* **17**:112–22.

Smith KGV. 1973. *Forensic entomology*. In: Smith KGV (ed.), *Insects and other arthropods of medical importance*. British Museum of Natural History, London.

Smith KGV. 1986. *A manual of forensic entomology*. British Museum of Natural History, London.

Stephens RJ, Richards RG. 1987. Vitreous humor chemistry: the use of potassium concentration for the prediction of the postmortem interval. *J Forensic Sci* **32**:503–9.

Sturner WQ. 1963. The vitreous humour: postmortem potassium changes. *Lancet* **1**:807–8.

Sturner WQ, Ganter G. 1964. The postmortem interval. A study of potassium in the vitreous humor. *Am J Clin Pathol* **42**:137–41.

Sturner WQ, Dowdey AB, Putnam RS, *et al.* 1972. Osmolality and other chemical determinations in postmortem human vitreous humor. *J Forensic Sci* **17**:387–93.

Suzuki S. 1987. Experimental studies on the presumption of the time after food intake from stomach contents. *Forensic Sci Int* **35**:83–117.

Suzutani T, Ishibashi H, Takatori T. 1978. [Studies on the estimation of the postmortem interval. 2. The postmortem lividity (author's transl).] *Hokkaido Igaku Zasshi* **52**:259–67.

Szent-Gyorgyi A. 1947. *Chemistry of muscular contraction*. Academic Press, New York.

Tagliaro F, Bortolotti F, Manetto G, *et al.* 2001. Potassium concentration differences in the vitreous humour from the two eyes revisited by microanalysis with capillary electrophoresis. *J Chromatogr A* **924**:493–8.

Takatori T, Ishiguro N, Tarao H, *et al.* 1986. Microbial production of hydroxy and oxo fatty acids by several microorganisms as a model of adipocere formation. *Forensic Sci Int* **32**:5–11.

Takatori T, Gotouda H, Terazawa K, *et al.* 1987. The mechanism of experimental adipocere formation: substrate specificity on microbial production of hydroxy and oxo fatty acids. *Forensic Sci Int* **35**:277–81.

Takeichi S, Wakasugi C, Shikata I. 1984. Fluidity of cadaveric blood after sudden death: part I. Postmortem fibrinolysis and plasma catecholamine level. *Am J Forensic Med Pathol* **5**:223–7.

Takeichi S, Tokunaga I, Hayakumo K, *et al.* 1986. Fluidity of cadaveric blood after sudden death: Part III. Acid-base balance and fibrinolysis. *Am J Forensic Med Pathol* **7**:35–8.

Taylor A, Wilkes D. 1863. On cooling of the human body after death. *Guy's Hosp Rep* 180–211.

Tomlin PJ. 1967. 'Railroading' in retinal vessels. *Br Med J* **3**:722–3.

Tough SC, Butt JC. 1993. A review of fatal bear maulings in alberta, canada. *Am J Forensic Med Pathol* **14**:22–7.

Trujillo O, Vanezis P, Cermignani M. 1996. Photometric assessment of skin colour and lightness using a tristimulus colorimeter: reliability of inter- and intra-investigator observations in healthy adult volunteers. *Forensic Sci Int* **81**:1–10.

Tsunerari S, Oho S, Sasaki S, *et al.* 1975. Histopathological studies on the changes of the postmortem cornea. *Jpn J Legal Med* **29**:298–303.

Vain A, Kauppila R, Humal LH, *et al.* 1992. Grading rigor mortis with myotonometry – a new possibility to estimate time of death. *Forensic Sci Int* **56**:147–50.

Valenzuela A. 1988. Postmortem diagnosis of diabetes mellitus. Quantitation of fructosamine and glycated hemoglobin. *Forensic Sci Int* **38**:203–8.

Van den Oever R. 1976. A review of the literature as to the present possibilities and limitations in estimating the time of death. *Med Sci Law* **16**:269–76.

Vanezis P. 1991. Assessing hypostasis by colorimetry. *Forensic Sci Int* **52**:1–3.

Vass AA, Barshick SA, Sega G, *et al.* 2002. Decomposition chemistry of human remains: a new methodology for determining the postmortem interval. *J Forensic Sci* **47**:542–53.

Webb PA, Terry HJ, Gee DJ. 1986. A method for time of death determination using ultrasound – a preliminary report. *J Forensic Sci Soc* **26**:393–9.

Wroblewski B, Ellis M. 1970. Eye changes after death. *Br J Surg* **57**:69–71.

CHAPTER 3

The establishment of identity of human remains

- Characteristics useful in identifying the dead 99
- Identification of skeletalized remains 106
- The determination of sex 108
- Estimation of stature from skeletal remains 114
- Estimating the subject's age from skeletal structures 116
- Determination of race from skeletal remains 122
- Personal identity from skeletal material 123
- The dating of skeletal remains 126
- Photosuperimposition techniques for identification 128
- Reconstruction of the facial appearance from the skull 128
- References and further reading 129

The identity of the dead is an essential part of post-mortem examination, for various reasons. These include:

- the ethical and humanitarian need to know which individual has died, especially for the information of surviving relatives
- to establish the fact of death in respect of that individual, for official, statistical and legal purposes
- to record the identity for administrative and ceremonial purposes in respect of burial or cremation
- to discharge legal claims and obligations in relation to property, estate and debts
- to prove claims for life insurance contracts, survivor's pensions, and other financial matters
- to allow legal investigations, inquests and other tribunals, such as those held by coroners, procurators fiscal, medical examiners, judges and accident enquiries to proceed with a firm knowledge of the identity of the decedent
- to facilitate police enquiries into overtly criminal or suspicious deaths, as the identity of the deceased person is a vital factor in initiating investigations.

The establishment of identity has many facets, some of which have no medical content, such as the recognition of clothing, documents and personal property. Some aspects, though primarily the responsibility of the police, have a medical content, including identification of fingerprints and personal property such as hearing aids, spectacles, medicines, trusses and surgical prostheses. False dentures have such central importance that they will be considered in the section on odontology.

Identification from anatomical and medical features comprises two major aspects:

- The establishment of certain broad groupings, such as sex, stature, race and age. These may be determined solely from the available bodily remains, though corroboration may be obtained from other evidence. An obvious example would be sex determination from clothing and jewellery, though even this may no longer be absolutely reliable.
- Comparison of the remains with ante-mortem information and records from those thought to be the victims, which restricts the method by making it dependent upon the existence or availability of such comparable material.

The establishment of identity may be required upon:

- **Intact fresh corpses.** Here visual recognition, directly or by photography, may be made. Hair colour, skin pigmentation, scars and tattoos can be examined without difficulty.
- **Decomposed corpses.** Many surface features may be partially or wholly lost, but more information can be obtained than from a skeleton. Direct measurement of body height, for example, may be possible, as well as

serological investigations and organ abnormalities, such as past surgical intervention, may be found.

- **Mutilated and dismembered corpses.** Depending on the degree of mutilation and the amount of tissue missing, identification may be hampered. If the remains are fresh, facts such as racial pigmentation may be determined, but direct measurement of stature may be impossible. Selective mutilation in some homicides may be deliberately directed at frustrating identification, such as the removal of teeth and finger-pads.
- **Skeletalized material.** If all soft tissues are absent, identity depends solely on osteological examination and measurements and the recognition of any pathological or anatomical abnormalities in bone.

CHARACTERISTICS USEFUL IN IDENTIFYING THE DEAD

Facial appearance

Even in perfectly fresh bodies, recognition may be difficult because of alterations in the features caused by death. It is a common occurrence in mortuary viewing rooms for a close relative, even a parent or spouse, to have doubts about – or even to deny or mistakenly agree to – the identity of the deceased person. Though distress and emotion play a part, changes in the features may be profound. Hypostasis, contact flattening, oedema, muscle flaccidity and pallor may all combine to distort the face. Recognition in the living is partly a dynamic process, aided by facial muscle tone and especially eye contact and movement, all of which are absent in the corpse.

Eye colour

In the fresh corpse, eye colour corresponds with the living state, but it quickly deteriorates. Loss of intraocular tension and clouding of the cornea develops progressively within a few hours, making the iris harder to observe. Collapse of the front of the globe occurs within a day or two and, with developing decomposition, all irises tend to darken to brown. It is unsafe to depend on eye colour as a criterion of identity later than a few days following death – or even much sooner where environmental conditions hasten decomposition.

Skin pigmentation

In undamaged, unputrefied bodies, the major ethnic differences in skin pigmentation are obvious, though the slight melanin increase of the misnamed 'yellow' races of Asia may be impossible to differentiate from Mediterranean or Middle Eastern races. The onset of deathly pallor and post-mortem hypostasis is more profound than slight variations between sunburnt White people, northern Asians or Semitic races. When putrefaction sets in, skin slippage progressively removes the pigmented layers and eventually pigmentation becomes unavailable as a marker of identity, though histologically melanin may still be visible in the basal layer of any surviving epidermis.

Burnt bodies may also lose pigmented skin, either by heat destruction or by deposition of soot and other combustion products on the surface, though it is rare for this to be so complete as to obscure all evidence of pigmentation.

Hair colour

The head, pubic and axillary hair is one of the most resistant identifying features, sometimes lasting millennia in favourable environments. The original colour may, however, alter after burial, becoming a brownish-red or 'foxy' colour within as short a time as 3 months. The colour may be obscured by dirt or staining and expert treatment (usually by a biologist or museum technologist) may be required to restore the original colour. The possibility of deliberate chemical coloration or bleaching before death may have to be considered and again expert chemical analysis in a forensic science laboratory may be needed to confirm or eliminate this if the hair colour conflicts with other positive evidence of identity.

Hair structure

This is more the province of the forensic biologist or anthropologist, but factors such as whether the hair ends have been cut or are naturally pointed may sometimes have importance in identity. Racial features exist, Negroid head hair being dark and having a spiral twist with a flattened, elliptical cross-section. Mongoloid hair is less pigmented and is straight with a cylindrical cross-section. Caucasian hair is round or ovoid in cross-section, but shows great variation in colour and morphology compared with the other two major ethnic groups. Though head hair in Caucasians is round to ovoid in section, eyebrow hair tends to be triangular and pubic hair flattened. Where a hair root is present, blood grouping and other serological criteria can be determined. Grouping of the shafts of hairs has been repeatedly attempted and success claimed by some workers. The cells of the hair root can give a DNA profile and mitochondrial DNA (mtDNA) has been detected even in a sample from a single hair shaft.

Microscopic examination of the hair can assist in determining the species, if not human. Human hair has broad groupings when viewed microscopically. However, the topic of individual identity from hair is now the province of forensic biologists, and has a specialism and literature of its own.

Tattoos

Deliberate ornamentation of the skin by introducing pigments under the epidermis has been practised for millennia and in all parts of the world. The word 'tattoo' comes from the Polynesian '*ta tau*', meaning 'to mark'. Some races such as the Ibans of Sarawak may be tattooed over much of their body surface, but many men and some women in most countries have localized tattoos, which can be of considerable assistance in identification. Many different pigments are used, as well as more unusual substances such as soot or gunpowder. The colour is pricked into the upper dermis with a sharp instrument, usually a needle used manually, or by an electric vibrator.

Once under the outer skin, the material persists for a long time. Colours such as blue, green or red may be scavenged by tissue cells and leached into the lymphatic system after a number of years or decades. Black pigments, usually carbon particles in the form of Indian ink, are so resistant as to be virtually lifelong, though some may be transported to regional lymph nodes. Thus an original pattern of different colours may fade differentially over years, the black areas standing out in strong relief.

The patterns are so diverse as to defy classification, though much has been written on the subject. From the point of view of identity, tattoos may be useful in both a general sense and for comparison with potential victims. The general applications require a knowledge of ethnic, national, cultural, religious and social practices, so that the tattoo can be assigned to some particular group of people. In Western society, the popularity of tattooing changes from time to time. Former statements that tattoos are seen mostly in the 'lower social classes' have little current validity, though it is obvious that bankers and parsons are less likely to sport tattoos than seamen and truck drivers. It is also now totally anachronistic to claim that most women with tattoos are prostitutes, though the nature and position of the tattoo is obviously relevant.

Certain tattoos have had particular significance such as a number, often '13', inscribed inside the lower lip of drug pushers or the bluebird on the extensor surface of the web of the thumb in homosexuals. The latter design is now so common as to be of little significance.

Tattoos are of most use when individual enough to be compared with known designs on a missing person. The design is often supplemented by actual names, though these are mostly forenames rather than family names. Designs are by no means unique and a professional tattooist may turn out hundreds of identical decorations, so names are a bonus when trying to achieve identity.

Initials may be used, but these can be misleading due to the use of nicknames and diminutives, such as 'B' for Bill, rather than William. In general, men have the names of girls on their tattoos, but initials are usually their own. Conversely, women with tattoos almost always display the names of sweethearts, rather than their own. Some tattoos may be numbers or unadorned names, the most tragic examples being those from Nazi concentration camps during the last war.

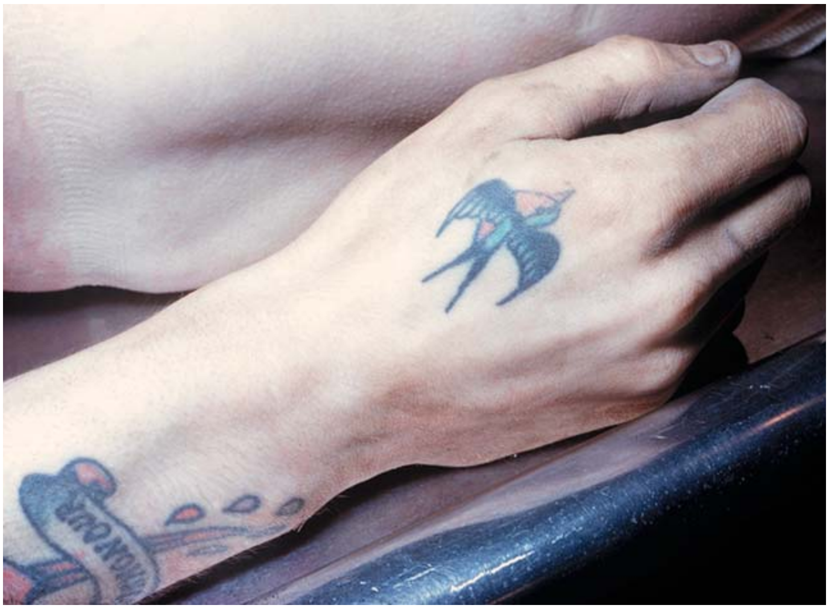

FIGURE 3.1 *Individualistic tattoos may assist the identification of unknown persons.*

The photography, distribution and matching of tattoos is police business, but the pathologist's duty is to make them accessible and clear to the police. Where decomposition has set in, tattoos may be obscured by wrinkled, peeling epidermis, but often this can be wiped off to reveal tattoos that are more vivid than on fresh skin. Once the tissues become green and slimy, then the pattern progressively vanishes, but early decomposition is no barrier to recognition. Where skin pigmentation may hide a design, skin slippage may actually enhance the clarity, especially of the weaker colours like blue, red and green. Attempts at the deliberate removal of tattoos are common, either from regret at having disfigured the body during youthful or drunken euphoria, or to remove evidence of identity. Many methods have been used, from surgical excision to scarification with sandpaper and from caustic soda to electrolysis. All methods depend upon damage to the epidermis and dermis, with consequent inflammation and scar formation. The tattoo can certainly be removed, but always by replacing it by a cicatrix to a greater or lesser extent, which will itself indicate that something pre-existed at that site.

It has been suggested (PA Edwards, personal communication) that the chemical nature of tattoo pigments might be used in identification. Although all manner of substances have been used to tattoo, most tattooists use black pigments containing carbon, green pigments containing potassium dichromate and red pigments containing mercuric chloride. Others use aniline-based dyes, so that microextraction and analysis in a forensic science laboratory might confirm or exclude the tattoo having been executed by a particular artist in the locality.

Finger, palm, foot and lip prints

The science of fingerprints, their classification, retrieval from records and methods of recording lie in the domain of police procedure, and, as with tattoos, the pathologist's role is merely to facilitate the taking of prints by police officers. Sometimes when a strange print is found at a scene of crime, the police may wish to take the fingerprints of doctors who visited the locus to eliminate them from their enquiries and there is obviously nothing objectionable about this. Indeed, some forensic pathologists who habitually assist the police leave their prints on file so that they are permanently available for exclusion. When called to a scene of death, the doctor should take care to avoid handling objects and furniture to reduce the need for such laborious screening of prints.

The doctor need know nothing about fingerprints except for his own interest, though it is as well to be prepared for the inevitable questions – no two fingerprints have ever been found to be identical and even identical twins have different prints. However, some recent claims point out that some prints are so similar that the criteria usually accepted as proof of congruence may not be sufficient. Police usually attend the mortuary to take full sets of fingerprints in the usual way by rolling ink on to the finger-pads and pressing the pulps against record cards.

The pathologist may help when there is strong rigor present, by either forcing the flexed fingers back or even by slitting the flexor tendons, though this has the danger of fouling the fingers with blood.

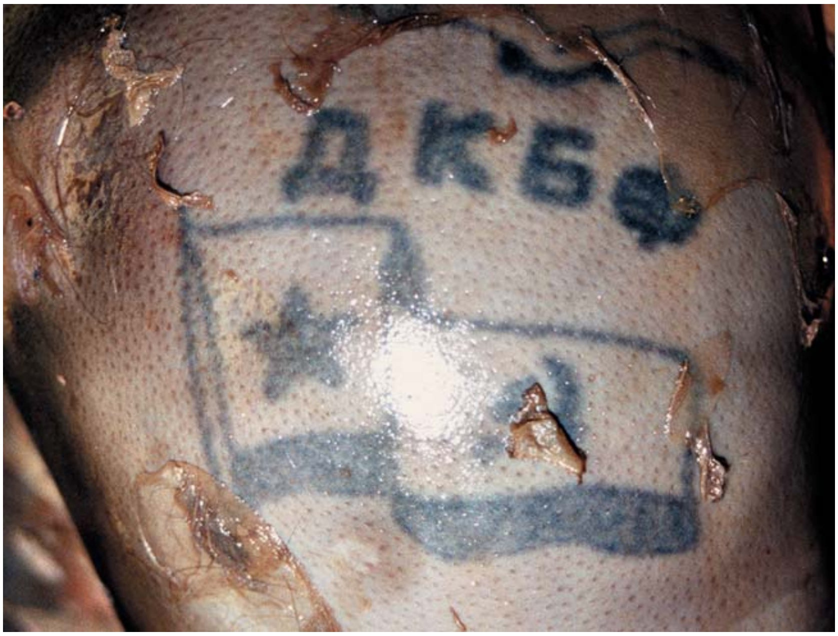

FIGURE 3.2 *Shoulder region of a decomposed body recovered from the sea showing a tattoo with Cyrillic letters (Д, К, Б, Ф) and a flag with a star and hammer and sickle suggesting a connection with the Russian Baltic Fleet.*

When the body is putrefied, the doctor can again assist by removing desquamated casts from the fingers for the police. These may need to be preserved or hardened up so that the decaying tissue does not lose further definition from the ridges that form the prints. The skin may be placed in formalin, alcohol or glycerine solution, especially where maceration in an immersed body has caused swelling and blurring of the epidermis. An alternative is immersion in 20 per cent aqueous acetic acid for 2–8 hours. Rarely, a criminal may attempt to obliterate his fingerprints by scarring the fingerpads. To be effective, this must be drastic enough to damage the underlying dermis and the consequent cicatrices will be more obvious than the original prints. Surgical grafting of skin from elsewhere on the body has been reported as a means of removing fingerprints, but again the result is self-evident proof of nefarious intent.

As with fingerprints, the skin pattern of palms, soles and even lips are said to be unique and have been used in identification, but again this is really police business. The pattern of veins on the back of hands is also said to be characteristic to one individual.

Identifying scars

These can be important in the identification of unknown bodies, even when some degree of putrefaction exists. The author (BK) used a cholecystectomy scar and scars of old surgical incisions on the legs to identify the body of a homicide victim submerged in a river for 6 weeks and many other cases are on record where scars provided similar vital information. Probably the best known instance is the Crippen case of 1910 in which much forensic controversy was generated over the distinction between a surgical scar and a skin crease on a piece of abdominal wall.

Scars on the skin may arise from any previous injury that has breached the epidermis, as superficial injuries to the upper layer of skin will heal without a trace. Where the dermis has been entered, healing occurs by organization of the plug of blood clot and/or granulation tissue, as described in Chapter 4. If the wound is narrow, such as that inflicted by a surgical instrument, sharp knife, razor or glass, then if the edges are kept together, especially by stitches or dressings, the resulting scar will be narrow and insignificant. Gaping or infection will widen the scar and obviously larger lacerations or burns will result in similarly greater scarring. In relation to identification, scars are of use only if those thought to be the victims are known to have similar scars, both in nature and position. The specificity of scars varies greatly, as many people may have appendicectomy scars, and many women have had hysterectomies or other gynaecological operations that leave lower abdominal scars. As always, scar evidence must be taken in conjunction with all other details of identity, but where a cicatrix is unusual or even unique in nature or position, its value is greatly enhanced.

Turning to traumatic scars other than surgical, the pathologist may be able to recognize the cause, which can assist in identity. Naturally the vast majority of non-surgical scars are the result of accidental trauma and where family, friends, photographs or medical records indicate the existence of specific scarred sites on the body, then confirmation of identity may be greatly assisted or confirmed.

Some scars are deliberately produced as part of some ethnic or religious tradition, such as the facial scars in some African tribes, as well as deformed or perforated ear lobes or lips. Regrettably, another class of scars is becoming more common in many parts of the world, due to physical torture; this is dealt with in Chapter 10. Linear scars, cicatrices of burns, keloid overgrowth (especially in negroid races), hyperpigmentation or depigmentation and other persistent lesions may assist in identity in the appropriate circumstances. Much depends on the accuracy of the information concerning the person thought to be the victim; relatives are notoriously vague and inaccurate about medical matters and frequently confuse different surgical operations, thus misleading the investigators about scars. Sometimes surgical scars may be almost invisible. The author (PS) once worked as surgical assistant in a private hospital where one of the surgeons used to perform appendicectomy (in non-acute cases) through a 1 cm wide incision, closing the wound, for cosmetic reasons, with subcutaneous catgut sutures. Most of these wounds healed with virtually no scar whatsoever, unless the person had a tendency to keloid formation. Wherever possible, old hospital records or the notes of a family physician should be obtained.

The old scars of wrist- or throat-slashing indicate previous attempts at suicide. Large areas of scarring on the trunk or limbs that have parallel grooves and marginal extensions suggest severe brush lacerations, probably from a traffic accident. Knife stab wounds may scar in an elliptical fashion, sometimes even with a blunt and a sharp end indicating the position of a single-edged blade. Bullet wound scars may persist – and in 1989, became a matter of political controversy over the identity of Nazi war prisoner Rudolf Hess, who was alleged to have been an imposter because he had no scar from an old rifle bullet wound through the chest. Burns may provide good evidence of identity, especially if they are widespread cicatrices that may have been quite noticeable during life. Many scars on the fronts of the lower legs may indicate the repeated stumblings of a chronic alcoholic, who frequently walks into or falls over furniture.

The age of scars

The age of a scar is very difficult to assess and, once it becomes mature, no change will occur for the rest of the person's life. When a linear wound is inflicted, be it a surgical operation or a knife slash, the edges become mechanically strong within a week, assuming that no infection or haematoma develops. The wound is brownish red at this stage and remains vascular for several months, depending upon its width. The pinkness of contained blood vessels gradually fades and a narrow surgical incision may be white by 4–6 months. The avascular collagen tends to shrink for a year or so, but is white–silver after a year and remains in this state indefinitely. These times are variable depending on the nature of the skin, its pigmentation and the part of the body involved. Where a wound crosses a skin crease that is frequently flexed or extended, it tends to gape and widen, the subsequent scar following suit.

Histology does not contribute much to the examination of scars, except to confirm that a skin mark actually is a scar. Stains for collagen or elastin may confirm the discontinuity in the dermis. Scars do not carry hair follicles, sweat glands or sebaceous glands, though an occasional accessory skin structure may be present as a result of the inclusion of a viable fragment of original skin in an irregular wound – or where surgical repair has detached or invaginated a piece of the margin.

Occupational stigmata

Formerly given prominence in most textbooks, distinctive physical marks of occupation have decreased in incidence and importance. They are still of use in certain areas, but local knowledge of occupations and industries is required. The Victorian days of pen callosities in clerks and shoemaker's kyphosis have gone, but many other occupations have their stigmata, though most are relatively nonspecific. The best known is the 'blue scar' of coal-miners, an involuntary tattoo caused by dust entering small lacerations on the hands and face. Some miners and quarrymen also have small facial scars due to rock splinters scattered during blasting. Steel workers and foundry men may have tiny burn scars on exposed areas from the spattering of hot metal. In more general terms, examination of the hands will indicate whether the deceased was a manual worker or involved in more sedentary occupations. Cuts, scars, callosities and hyperkeratosis of the hands are obvious indicators of rough work.

Internally, black, dust-laden lungs – with or without pneumoconiosis – points to mining, as coal trimmers no longer exist as dock workers. Thickened, pearly plaques on the parietal pleura and diaphragm are caused by exposure to asbestos, though in recent years the widespread use of this substance in so many industries makes it less helpful in identifying the occupation.

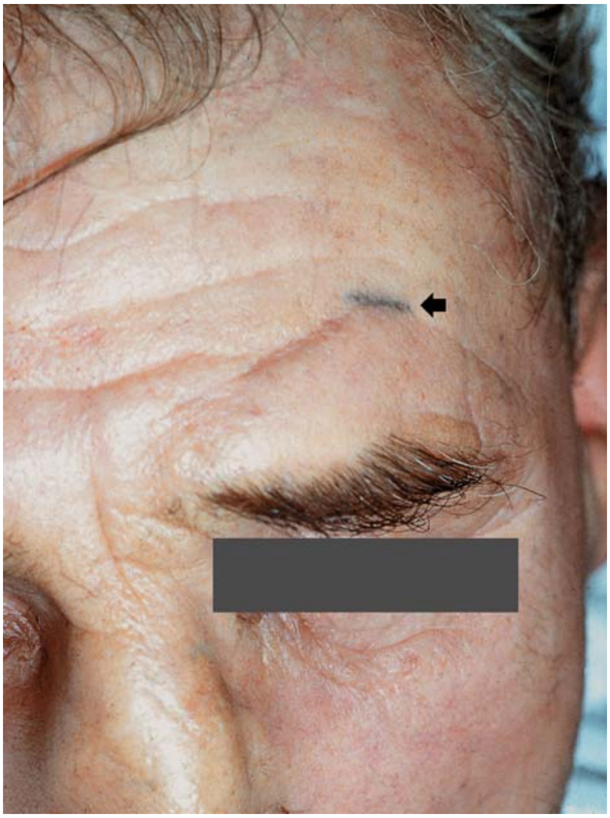

FIGURE 3.3 *Careful external examination may yield evidence that can assist identification. The small 'blue scar' on this forehead establishes that the deceased had been a coal-miner; coal dust enters small lacerations and remains visible after healing. Virtually every coal-worker has such marks on hands or face.*

The stature of an intact body

The stature of a corpse is not necessarily exactly that of the person measured during life. In the mortuary, care must be taken to measure the body from heel to crown, and not to accept the mortuary attendant's or funeral director's 'coffin measurement', which is made from toe to crown. Because of the usual plantar flexion of the foot, this latter height may be as much as 10 cm longer than the heel–crown length.

The height of a dead body may differ from the height during life in either direction, being slightly longer or shorter, though lengthening is much more common. In addition, the measured height may change a little at different periods after death. Because of the complete loss of muscle tone in the first stage of flaccidity, relaxation at large joints, such as the hip and knee, together with the lost tensioning effect of paraspinal muscles on intervertebral discs, the body may lengthen by up to 2–3 cm. Rigor mortis may then replace the muscle tone, shortening the body and tending to flex the legs a little. When rigor passes off and decomposition sets in, joints become lax. Loss of tension in the intervertebral discs tends to shorten the spinal column and hence the total height by a centimetre or so. According to Trotter and Gleser (1958, 1977), the average increase is about 2–5 cm. The errors in getting a really accurate post-mortem length include the difficulty of locating the measuring device exactly at the heel and the crown of the head, as well as fully extending the limbs, spine and neck of a body in rigor.

Determination of sex in non-skeletalized bodies

This is usually obvious and rarely presents problems. The issue of intersex and hermaphroditism is so rare as to be discounted in most forensic practice, except in some areas, such as South-east Asia, where deliberate transsexualism in men, aided by surgical intervention to feminize the genitalia and breasts occurs more frequently. Where this problem arises, expert advice must be sought. In general, the external genitalia remain recognizable until a late stage of putrefaction. In addition, breasts and general body shape, as well as the pattern of pubic hair, reveal the sex. Female pubic hair usually grows low on the abdomen, the upper margin being a horizontal or semicircular edge above the mons veneris. The male pattern rises more in the centre line, sometimes as far as the umbilicus. There are, however, always exceptions to these generalizations. The presence or absence of circumcision should be noted, as it may help in identification of certain ethnic and religious groups such as Jews and Muslims. Naturally it is of assistance here in eliminating only those where the prepuce is present, as many men of other religions are circumcised.

Clothing, hair length, hair style, hair coloration, earrings and other jewellery are no longer the reliable guide to sex that they once were. In certain cultures and religions there are obvious criteria that can assist in the determination of sex, but these are rarely needed as the more apparent anatomical differences can be used. Where putrefaction is advanced, examination of the internal pelvic organs may still reveal unequivocal evidence of sex. Internal viscera are often in better condition than the outside of the body. The uterus is the organ in the body most resistant to decomposition, though the prostate is also quite persistent.

Where almost total dissolution exists, then radiology of the pelvis or the criteria used in skeletal assessment can be used. Cytological examination of cell nuclei may indicate sex where only tiny fragments are available – such as tissue found on a road or rail vehicle where impact with a body is suspected. Usually it is the species rather than the sex that is the first consideration, and this is done by species-specific serology. When sex is to be determined, much depends on the state of preservation. When stained microscopic smears can be prepared or when tissue can be processed for histological sections, then sex-indicating nuclear chromatin may provide a definitive test. The 'drumsticks' of polymorphonuclear leucocytes, and the marginal chromatin masses or 'Barr bodies' in epithelial cells, such as the buccal mucosa, may indicate the sex, but in decomposed material this may be impossible.

The recent remarkable developments in cytogenetics and DNA analysis now give additional powerful tools in sexing and determining the human origin of sub-anatomical fragments, which will partly or wholly replace the older techniques mentioned above.

Other more recent techniques include counting the sex material within the nucleus stained with a fluorescent dye and viewed with ultraviolet light; this may be performed on neurones in a brain smear or in cells from the kidney.

Age of non-skeletalized bodies

Determining the personal age of a corpse is far harder than determining its sex. Ageing a living person can be difficult and a wide margin of error exists, especially as age advances, so the same exercise in a corpse can be expected to be even less exact. Where the body is relatively intact, the usual criteria known to the man in the street are applied to make a general estimate. Hair colour is of general applicability, though every one knows of someone whose hair is grey at 25 or someone who is dark haired at 70. The loss of elasticity in the skin, its thinness and hyperkeratosis and its red Campbell de Morgan spots are indicators of senility. These days teeth are less useful than in the past from the point of view of extractions, clearance, full dentures and oral hygiene, but the general impression remains. The objective determination of age from tooth attrition is a matter for a specialist forensic odontologist (see Chapter 26).

The eyes are useful, in that a grey or white ring around the pupil – the 'arcus senilis' – is rarely seen below 60 years. In infants and children, the height and weight may be compared with standard tables, but developmental defects, disease and malnourishment may introduce considerable errors. Internally, the evolution of degenerative changes such as arterial atherosclerosis and arthritis, may give a general impression, but personal variation is so great that it offers little more than a distinction between 'young' and 'old'. Where a distinction has to be made in a double death between two people of markedly different ages, however, then these generalities may be of considerable use. Up to the age of about 20–25, teeth, ossification centres and epiphyseal fusion are good indicators, but these are described in the section on skeletalized material.

Recent research, aimed mainly at skeletalized bodies but available for any corpse, is the determination of the racemization ratio of amino acids in teeth, as the preponderance of laevo- and dextro-compounds appears to be a function of personal age.

In addition, the Gustafson technique on teeth, described later, is also applicable to a non-skeletalized body, if ageing is important.

Racial and ethnic characteristics

Much of the identification of race depends on common knowledge and recognition of various ethnic traits, and is not primarily a medical problem. Everyone can recognize the deep pigmentation of a West African the epicanthic folds of the Mongoloid races, the red-haired Celt and the pale blond common in Scandinavia. There is a wide and almost continuous range of biological variability, however, not only because of extensive ethnic interbreeding, but also because of marked variations even within a coherent racial group.

In the intact or only partially decomposed body, the following factors may be used as pointers to ethnic classification, but the non-specificity of most of these must always be borne in mind:

- Clothing, ornaments and other associated objects. A vast catalogue can be assembled (such as the string around the wrist in Hindus), but it must be borne in mind that other races may adopt these through choice, marriage or deception.
- Hair texture, style and length and beards. Examples are the hair of the Sikh, the beard of the orthodox Jew, the crinkly hair or 'peppercorn' (clumped growth) of some Negroids, the multiplaiting of Rastafarian West Indians, the straight black hair of Asians and the ginger hair most common in Celtic people. With affected styles and convincing modern coloration and texture alteration, however, these features can be most misleading. Forensic science techniques may be needed to detect and identify bleaching and dyeing changes in hair.
- Physical artefacts include penile and vulval circumcision, nose and ear piercing, lip perforation, tribal facial scars, and earlobe distension in East Africans and some tribes in Sarawak. More bizarre forms, such as neck-stretching by rings, are now rare. Tattoos (mentioned elsewhere) may have an ethnic basis, as in the Ibans of Sarawak.
- Facial appearance is, of course, a major criterion with the high cheek bones and epicanthic folds of the Mongoloid peoples, and the prognathism and nasal differences in Negroid peoples. So many intermediate forms exist, however, such as the Slav and the Polynesian, that only the classic groupings can be identified with confidence.
- Body size gives rise to the same uncertainty, especially as racial differences have lessened with better nutrition in some groups. It is obvious, however, that Japanese are generally smaller than northern Europeans and that in Negroid bone structure the femora are proportionately longer and less curved than those of other races. Teeth are discussed later, but the concave upper central incisors of the Mongoloid races can be helpful on occasions, though they are not absolutely specific.

Finally, skin pigmentation can obviously differentiate the Negroid races from the other main groups, though again there is a continuous range of skin colour. To a lesser extent the slighter melanin content of the skin of Mongoloid races is recognizable, but merges into the variable tints of Semitic, Asian and Mediterranean peoples.

It may well be that, in the near future, genetic DNA studies will revolutionize this aspect of identification. They are also likely to do the same for personal identification.

Identification by DNA characteristics

One of the most revolutionary advances in identification in recent years is the so-called 'DNA profiling or fingerprinting'. This is a technique devised by Alec Jeffreys of Leicester University, in which virtually unique sequences of bases in the DNA strands of chromosomes are used to compare one blood or tissue sample with another, and to investigate genetic relationships. The details of the technique are complex and require highly specialized reagents and apparatus. The method needs highly trained forensic scientists for its application to medico-legal problems. All that the pathologist needs is a basic knowledge of the principles of the technique and how it can be applied in a forensic context.

The deoxyribonucleic acid (DNA) that forms genetic material is composed of two strands of sugar and phosphate molecules that are twisted into a double helix, which is bound together by links formed from adenine and thymine and cytosine and guanine. Each helical twist has ten such links, like rungs on a twisted ladder. A single DNA molecule may have millions of such links and the permutations of the bases in each adjacent link create the genes, the segments of the DNA that carry genetic information. A gene consists of a segment containing some hundreds or thousands of links, but separating these discrete genes are segments that seem to have no genetic function, being redundant. There is about a metre of

DNA in each human nucleus, but only 10 per cent is used for genetic coding, the rest being redundant. Of these silent segments (sometimes called 'stutters'), there may be 200–14 000 repeats of each identical sequence on each DNA strand.

During investigations into a specific gene (that for synthesizing myoglobin), Jeffreys found that the adjacent redundant segments varied greatly in base sequence, but that they were constant for a given individual and that they were transmitted from parents in a regular fashion. In other words, they formed a 'signature' that was virtually unique for every individual person, except identical (uni-ovular) twins, in whom they must be the same as they share the same DNA in the fertilized ovum. The chance of two persons sharing the same sequence is, according to Jeffreys, about one in a million billion. Even among siblings, the chances are only one in ten thousand billion.

To identify the redundant segment (often called the 'minisatellite' or 'hypervariable regions'), a minute amount of human tissue containing nuclear material is required. A small volume of blood containing leucocytes is sufficient, or a semen sample, pulp from a tooth or a hair with some sheath cells at the root. The DNA from the contained nuclear material is broken up into fragments by enzymes called 'restriction endonucleases' that cleave the DNA strands at predetermined spots. The fragments are then separated by gel electrophoresis, which spreads the fragments into bands, transferred to nylon or nitrocellulose sheets and subjected to a 'DNA probe'. This is a single strand of DNA labelled with radioactive phosphorus-32. This latches on to the separated minisatellite fragments and – by microradiography – the position of each band is rendered visible on an X-ray plate.

The end result is a series of parallel bars, similar in appearance to the 'bar code' printed on goods in a supermarket. However, more recently the bar pattern from multilocus probes has been supplemented or replaced by simpler patterns from single-locus probes, though the principles remain the same. From the presence of different bars in given positions, comparisons can be made with other samples, so this is a classic forensic technique of comparison. The test is useless without something to be compared with, but the blood on a weapon in the possession of a suspect can be matched with virtually complete confidence against the blood of the victim. Seminal fluid in the vagina of a victim of a murder/rape can be matched against the blood DNA pattern of a suspect – there is no need to match semen against semen, as all the DNA in a given person must, by definition, be identical. A few hair roots found on a blunt instrument in the possession of a suspect from a murder can be matched against the autopsy blood sample. In sexual crime cases, a great advantage of DNA testing over conventional blood-group secretor tests is that DNA can distinguish between mixed semen and vaginal fluids from a swab, which can confuse or negate blood-group techniques.

Another wide medico-legal use is in paternity testing where, unlike blood group, haptoglobin and enzyme tests, a positive identity of parents can be made, rather than a mere exclusion. Every bar in the 'bar-code' must have come from either mother or father, half the number from each. In testing, the coding is made from child, mother and father. The bars in the child's code are first matched with those in the mother's pattern. The remaining bars then must correspond with those from the father – if not, he cannot be the father.

SAMPLES REQUIRED FOR DNA TESTING

Although amplification techniques, such as polymerase chain reaction (PCR), etc., now allow minute samples to be tested, the larger the sample, the better the chance of a successful DNA test.

Post-mortem material is inferior to live blood and tissue for DNA fingerprinting if any significant post-mortem change has proceeded enough to break down nuclear chromatin. Plain blood samples should be taken, however, though leucocytes may have already disintegrated. Some 5 ml of blood is taken into an EDTA tube, which extracts metallic ions and not only prevents clotting, but inhibits enzymes in blood or micro-organisms, which may break down DNA during storage. The best material is said to be muscle or the spleen and, if decomposition is advancing, the bone marrow is recommended. At least 0.5 g of tissue should be cut from the parenchyma of an organ and placed in a small plastic tube with no fixative or preservative. This should be frozen at −20°C if there is likely to be the slightest delay in transmission to the laboratory.

In sexual offences associated with homicide, as much material as can be obtained from the vagina should be collected, either fluid by pipette or multiple swabs from vagina, rectum and mouth. These again should be frozen if there is more than a few hours delay in transit to the laboratory (see also Appendix 1).

IDENTIFICATION OF SKELETALIZED REMAINS

One of the classic problems of forensic pathology, the identification of a whole or partial skeleton, involves techniques and expertise that span a number of disciplines from anatomy to radiology, from archaeology to dentistry. It often overshadows the equally important need to identify intact or decomposed corpses, but the disproportionate amount of space occupied by skeletal identification in most textbooks can be partly justified by the fact that bones can survive for decades, centuries or millennia, so that the cumulative reservoir of material is vast. As was observed in relation to intact bodies, the procedure for identifying bony remains falls into two distinct sections:

- allotting the bones to general categories based on absolute criteria concerning species, race, sex, stature, age and date
- comparative studies, where the remains are matched against ante-mortem data derived from those people who might be potential victims.

The general categorization of skeletal remains

When objects thought to be skeletal remains are found, the following questions need to be asked and, where possible, answered:

- Are they bones?
- Are they human bones?
- What is the sex?
- What is the stature?
- What is the race?
- What is the age?
- How long have they been dead and/or concealed?
- What was the cause of death?

Are the remains actually bones?

This question is not as facetious as it may seem, as untrained persons may provide objects that falsely resemble bones. The author (BK) has been offered shaped stones, plastic and even hard wood, which were thought to be bones. The mistake is even more common when these objects are mixed with undoubted bones, usually animal in origin, which may be found buried or mingled with rubble.

Recognition is usually easy, by the shape, texture and especially weight of the objects. Greater difficulty may be experienced when simulated human 'bones' are found, such as the 'radius' brought to the author (BK). This originated from a medical student's anatomical skeleton and was made from plaster coated with a plastic polymer. Visually it was indistinguishable from a real bone, but the abnormal lightness indicated the true nature.

Are the bones of human origin?

The recognition of species is important, but deciding on a human origin is usually easy, unless marked fragmentation has occurred. Many animal bones are found by the general public and police who are not qualified to decide whether they are human. Extensive rebuilding programmes and the renovation and demolition of old sites reveal many caches of bones, but in the authors' experience the majority are of animal origin, often originating from butchers' debris.

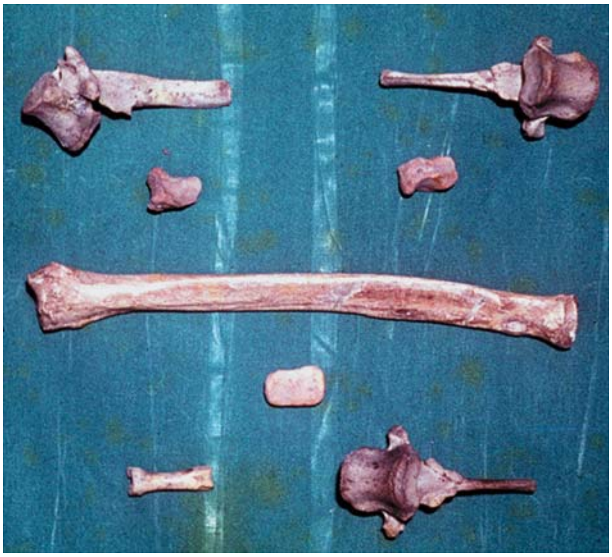

FIGURE 3.4 *All bones are not human – and not all 'bones' are bones. The smaller bones are animal, but the central 'radius' is from an ersatz student skeleton, made from plastic material, but thought by the police to be evidence of a criminal death.*

First, the size is assessed and many small, slender bones excluded on obvious grounds. Even a turkey thigh cannot be mistaken for a metatarsal, nor a beef rib for human. The gross anatomy is then studied, and even if the pathologist makes no pretensions to being a comparative anatomist, the majority of intact animal bones can be recognized from their lack of correspondence to any human bone. Difficulties arise with smaller bones from some animals, especially the hands and feet, where digits, metatarsals and metacarpals require careful study to distinguish them from the human. Several errors have been made in respect of bear paws, which closely resemble the human hand.

When bones are incomplete or fragmentary, the problems escalate rapidly. If the ends of longer bones are present, then their non-human shape may be more readily determined, but cylindrical segments of the central shaft have little in the way of distinguishing features, apart from size. Burnt bone fragments offer similar problems, added to which is the possibility of heat distortion and shrinkage.

The advice of an anatomist is needed in such cases, preferably one with forensic experience and enthusiasm for the project. Unfortunately, every anatomist is not necessarily an osteologist and it may be more effective for the bone to be examined by a pathologist who has both enthusiasm and the relevant reference books.

Histological examination may offer a species differentiation – or at least to exclude a human origin. The Haversian architecture is different in many animals, but the few specialist papers on this subject must be consulted for details.

Where the animal species is relevant to the enquiry, then a veterinary or comparative anatomist must be consulted. If the bones are too fragmentary to provide any anatomical data, then serological investigations must be attempted. These depend on species-specific proteins being extracted from the bone into solution where they can be tested against specific antisera prepared by immunizing animals (usually rabbits) against a range of animal proteins, usually derived from blood. Thus the test is essentially seeking plasma constituents within the bone, the recognition being carried out by techniques such as electrophoresis or gel diffusion. DNA can now identify human tissue, if not the alternative species.

The drawback of such serological tests is that they cannot be applied to bones that no longer have extractable proteins, and these include burnt or cremated bone and bone that has been dead for some years. The length of time for which identifiable protein persists is variable, but a negative result is usually to be expected after 10 years following death, though DNA techniques may be more sensitive.

THE DETERMINATION OF SEX

The accuracy of determination of the sex of skeletal remains varies with the age of the subject, the degree of fragmentation of the bones and biological variability (Table 3.1).

Particularly when studying the skull and pelvis, a subjective impression by the experienced observer defies complete analysis, yet objective measurement may be no more accurate: in other words, the processes comprise morphological traits versus morphometry. The determination of sex is statistically the most important criterion, as it immediately excludes approximately half the population whereas age, stature and race each provide points within a wide range of variables. Obvious sex differences do not become apparent until after puberty, usually in the 15–18-year period, though specialized measurements on the pelvis can indicate the sex even in fetal material. Sex and age are linked, especially where body size and weight are concerned. Similarly, race confuses sexing, for example, the size of the supraorbital ridges in a normal negroid female may exceed those in the average Caucasian male.

TABLE 3.1 *Pfitner's table of bodily dimensions in the female compared with the male (per cent of male dimensions). The common generalization that the female is 94 per cent of the male size varies in different areas of the body*

Stature	93.5	Arm length	91.5
Head breadth	98.0	Sitting height	94.5
Face breadth	94.0	Head circumference	96.0
Face height	90.0	Head height	96.0
Head length	95.5	Leg length	93.0

The accuracy of sexing is hard to estimate, as various loading factors exist. Krogman comments that he scored 100 per cent accuracy using the whole skeleton, 95 per cent on pelvis, 92 per cent on skull, 98 per cent on pelvis plus skull, 80 per cent on long bones and 98 per cent on long bones plus pelvis. He admitted, however, that, as most anatomy department material has a sex ratio of about 15:1 in favour of men, marked bias could be introduced by assigning all doubtful bones to the male category.

Stewart records that, for the whole skeleton, one can expect a 90–95 per cent success rate and for the skull alone only 80 per cent, but, if the mandible is also present, this rises to 90 per cent (Hrdlicka 1939). In general, adult female skeletal measurements are about 94 per cent that of the male of the same race, but different measurements may vary from 91 to 98 per cent.

The skull

The following features develop after puberty and are modified by senility, so are applicable only between about the 20th and the 55th year. Age as well as race has a profound effect.

- **General appearance.** The female skull is rounder and smoother than the rugged male.
- **Size.** Male skulls are larger, with an endocranial volume some 200 ml greater.
- **Muscle ridges** are more marked in male skulls, especially in occipital areas where larger muscles are

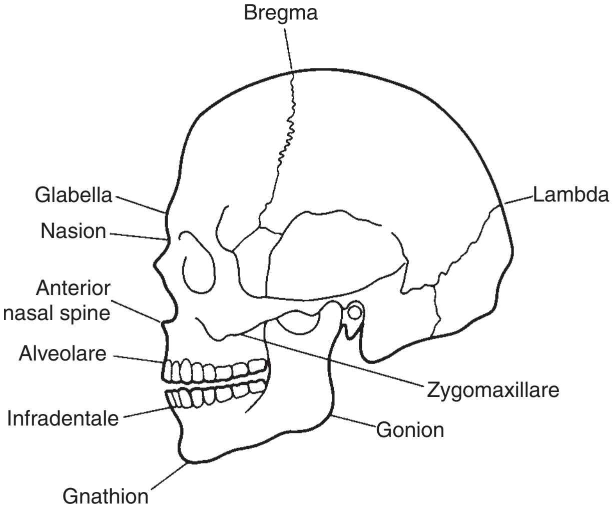

FIGURE 3.5 *Anatomical landmarks of the skull.*

attached to the nuchal crests and in temporal areas for larger masseter and temporalis muscles.

- **Supraorbital ridges** are more marked in male skulls and may be absent in female. The glabella (central forehead eminence above the nose) is small or absent in the female, and prominent in the male, though this is a poor discriminant.
- **Mastoid process.** This is larger in male skulls (see Hoshi).
- **Frontal and parietal eminences.** These are more prominent in female skulls, which resemble the shape in an infant more than in a male.
- **Palate.** This is larger and of a more regular U-shape in men. The smaller female palate tends to be parabolic.
- **Orbits** are set lower on the face in the male skull, with more square and less sharp edges (especially upper edge) than the female.
- **Nasal aperture.** This is higher and narrower in the male skull and has sharper edges. The nasal bones are larger and project further forward to meet at a more acute angle than in the female.
- **Forehead.** This is high and steep in the female skull, with a more rounded infantile contour than the male.
- **Teeth** are smaller in the female skull, molars usually having four cusps. The male often has a five-cusp lower first molar.
- **Zygomatic process.** The posterior ridge projects back beyond external auditory meatus in the male skull. The zygomatic arches bow outward more than the female, where they remain more medial.
- **Mandible.** This is large in the male skull, with a squarer symphysis region. Female jaws are more rounded and

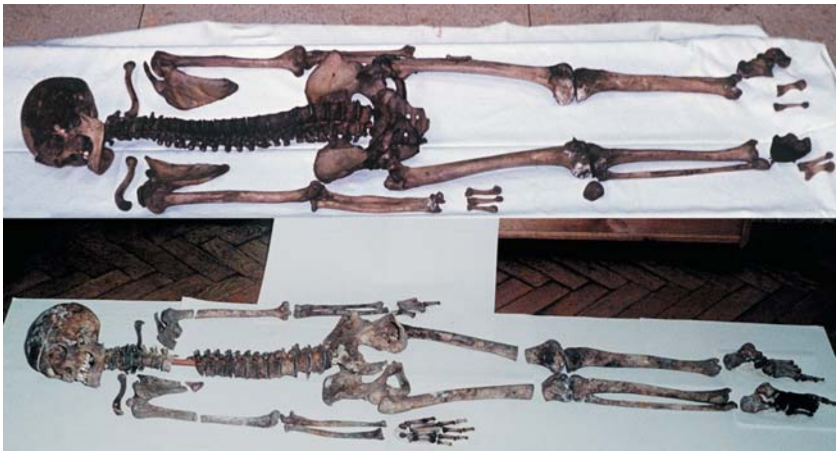

FIGURE 3.6 *Two recovered and reassembled skeletons. After cleaning, the bones should be laid out in correct anatomical position, to gain an approximate direct measure of the stature, allowing as well as possible for lost soft tissue. The upper skeleton, obviously male, shows the frequent loss of small hand and foot bones, but no injuries. The lower remains, from a hidden 40-year-old homicide, reveal that the body was sawn into six sections, there being cuts above the knees and through spine, scapula and humeri.*

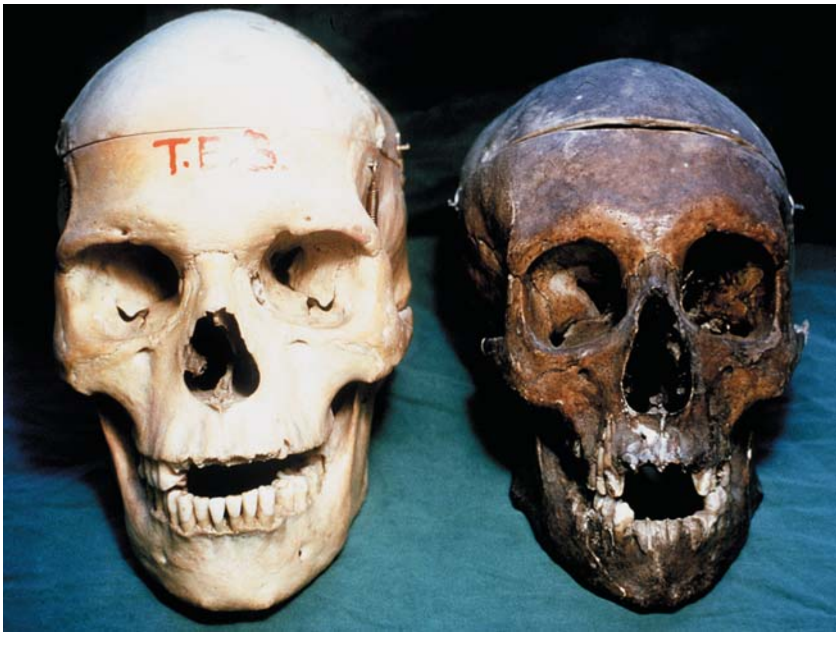

FIGURE 3.7 *Frontal view of male and female skull. The male (left) is larger, more massive and has heavier eyebrow ridges. The chin is squarer and the mandible heavier, The female has relatively larger height orbits, a smoother cranium, and fuller frontal and parietal eminences.*

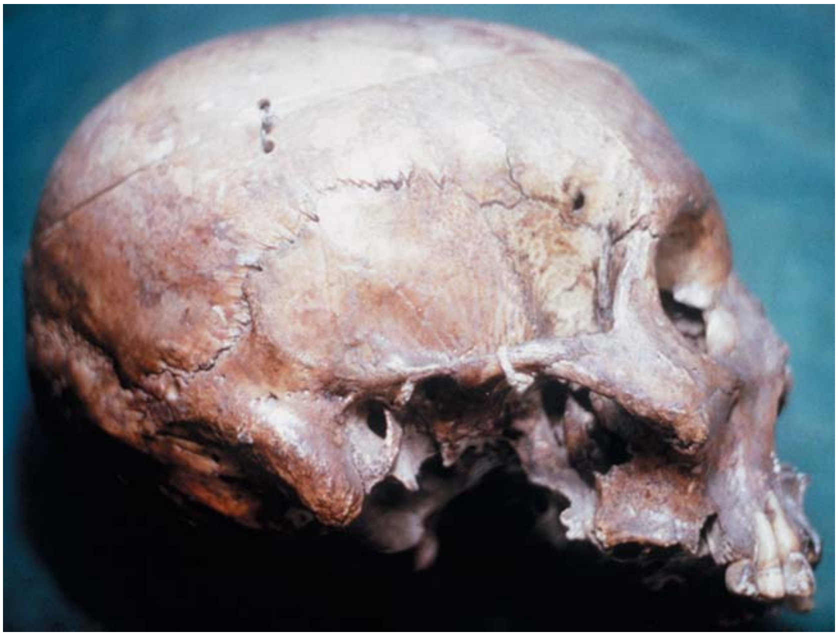

FIGURE 3.8 *A typical male skull, with a sloping forehead and prominent occipital ridges for large muscle attachment. Probably the best male feature is the large mastoid process; also, the zygomatic process extends well behind the auditory meatus.*

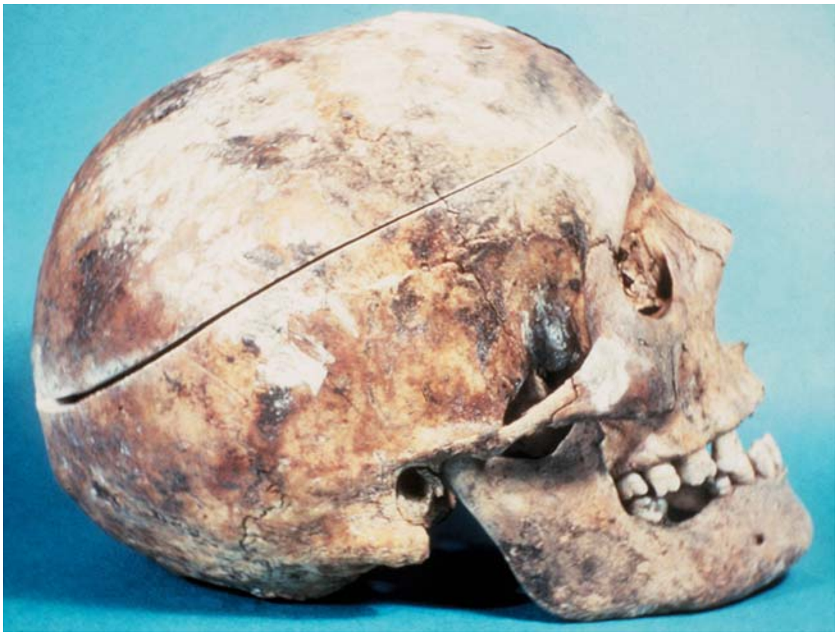

FIGURE 3.9 *A typical female skull, from a murder victim hidden for 40 years in a cave. The cranium is high, round and smooth with insignificant ridges for muscle attachment. The mastoid is tiny, one of the best discriminants. The post-zygomatic ridge does not continue behind the auditory meatus and there are no supraorbital ridges. The chin is round and smooth.*

project less at the anterior point. The vertical height at the symphysis is proportionately greater in the male. The angle formed by the body and the ramus is more upright in the male, being less than 125°. The condyles are larger in the male skull, as is the broader ascending ramus, and there is a more prominent coronoid process.

These sex variations represent the 'typical' White and, to a great extent, Asian skull. There is considerable overlap, especially in subjects from the Indian subcontinent, where osteological sex differences are much less marked. The criteria set out above exclude prepubertal and senile persons, and are less valid for those outside the 20–55 age group. For methods of sexing by discriminant function analysis, the work of Giles and Elliot (1963) should be consulted. In recent years craniometry has been applied to sexing the skull and multiple accurate measurements between discrete anatomical points have been used to produce discriminant function analysis. The papers of numerous authors must be consulted for details of this sophisticated and difficult technique that is accurate in 83–88 per cent of cases. The frequency of correct sexing is no greater than more subjective methods, but commands a greater degree of confidence level in respect to its accuracy.

Sex characteristics in the pelvis

The post-pubertal female pelvis is wider and shallower than the rather upright male girdle to allow for the passage of the fetus during childbirth. Though as always there is an overlap of the appearance and measurements in the least characteristic individuals, the usual variations are sufficient to allow sexing of the adult pelvis to be made with about a 95 per cent confidence rate (see work by Genovés). In the pelvis, unlike the skull, there are differences, albeit subtle, in immature (even fetal) pelves that allow sexual differentiation. The following features provide the most useful criteria. One must not be used in isolation, however; as many as possible must be assessed together. Sex features are independent of each other and one may even contradict the other in the same pelvis. It is often a first subjective impression by an experienced eye that determines the answer, though where the sex differences are slight, careful measurements may be needed to resolve the problem.

As in the skull, the male pelvis is more rugged because of the attachment of stronger muscles. It stands higher and more erect than the smoother, flatter, female pelvic girdle. The subpubic angle, measured at the medial intersection of two lines drawn along the best approximation of the lower border of both inferior pubic rami, approaches 90° in the female pelvis, but usually about 70° in the male.

This is often a subjective measurement, however and depends in turn upon the shape of the pubic bone itself. When the line of the inferior ramus is projected medially and intersected with a horizontal line laid across the upper border of the superior ramus, the reverse size of angle is seen, the male being wider than the female.

The body of the pubic bone, the block lateral to the symphysis, tends to be triangular in shape in the male, whilst the female pubis is more rectangular (Phenice 1969; Iscan and Derrick 1984). Certain sexual variations in the pubis have been used by Phenice. In the female these are:

- a bony ridge ('ventral arc') running down the ventral surface from the pubic crest
- a concavity of the lower margin of the inferior pubic ramus immediately lateral to the lower border of the symphysis
- a ridge of elevated bone on the medial aspect of the ischiopubic ramus, immediately lateral to the symphysis; in the male this area is broad and flat.

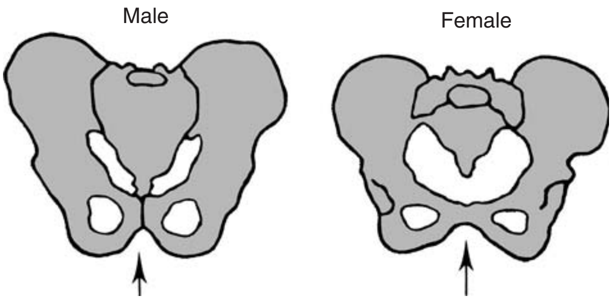

FIGURE 3.10 *Sex differences in the human pelvis. The most noticeable features are the narrower suprapubic angle and the higher iliac blades in the male pelvis.*

The 'ischiopubic index' devised by Washburn may be helpful, in which the pubic length (×100) is divided by the ischial length. The measurements must be carefully made, the pubic length being from the plane of the symphysis to the reference point in the acetabulum and the ischial length being from the same point to the most distal edge of the ischium. The reference point is the site of fusion of the three elements of the immature innominate bone, usually marked by a notch in the articular surface of the acetabulum (Schultz 1937). If the ischiopubic index (in White races) is less than 90, the pelvis is male; if over 95, it is female. The acetabulum is larger in the male, being an average of 52 mm in diameter, compared with 46 mm in the female. The male joint cup also faces more laterally than that of the female, which tends to look more forward. Naturally, acetabular size is related to that of the femoral head, which will be considered later. The greater sciatic notch is an important criterion, being deep and narrow in the male, wide and open in the female. The angle formed by the margins approaches closer to a right angle in the woman than it does in the man. Harrison (1968) and Hrdlicka (1939) both felt that the greater sciatic notch was one of the best discriminants for sex, the latter claiming a 75 per cent success rate using this criterion alone.

The obturator foramen is somewhat ovoid in the male, but triangular in the female. The pre-auricular sulcus, which marks the attachment of the anterior sacroiliac ligament, lies just lateral to the sacroiliac joint and is usually well marked in women but often absent in males. The pelvic inlet, looked at from above, is more circular in the female, the male being heart-shaped as a result of the protrusion of the sacrum into the posterior brim (Greulich and Thomas 1938). A number of other pelvic 'indices' have been devised by various authors (see, for example, the work of Turner, Greulich and Thomas, Caldwell and Molloy, Straus, and Derry).

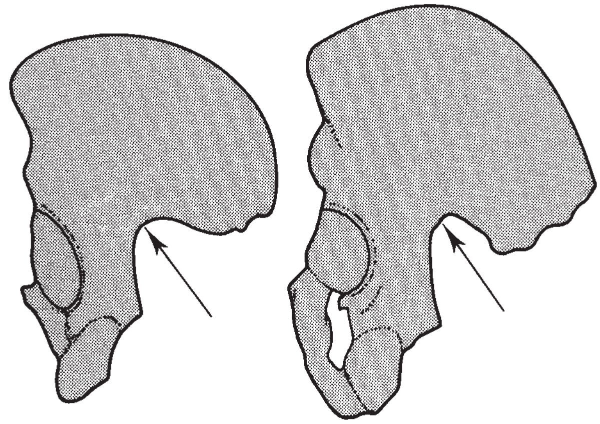

FIGURE 3.11 *The greater sciatic notch is narrower in the male innominate bone.*

Sex characteristics in the sacrum

The sacrum is functionally part of the pelvis and shares in its sex variations. The female sacrum is wide and has a shallow curve, again related to the larger pelvic canal for childbirth. It is shorter in the female and the curve is limited almost entirely to that distal part below the centre of the third sacral vertebra. The male sacrum may have more than five segments, which is rare in the female. The curve in the male is continuous down the whole bone and there may even be a slight forward projection of the coccyx. Fawcett (1938) compared the transverse diameter of the first sacral vertebra (CW) with that of the base of the sacrum (BW). The formula CW × 100/BW averaged 45 in the male and 40 in the female. Kimura (1982) has developed a ‘base wing index’, where the relative widths of the wing and base provide discriminant function coefficients for sex determination.

Sexing from the long bones

The femur is the most useful, its length and massiveness themselves being significant. There is, as usual, considerable overlap of all long-bone sex characteristics, but Brash’s series showed that the maximum (oblique) length in the male femur was around 459 mm, while that of the female was only 426 mm. Other figures from Pearson and Bell (1917) suggested mean values of 447 mm for men and 409 mm for women. Using the trochanteric oblique length, they suggested a range of 390–405 mm for women and 430–450 mm for men, though there was the usual overlap in the middle. Race and nutrition (which is related to the era and the place in which the samples were obtained) must be allowed for in such measurements.

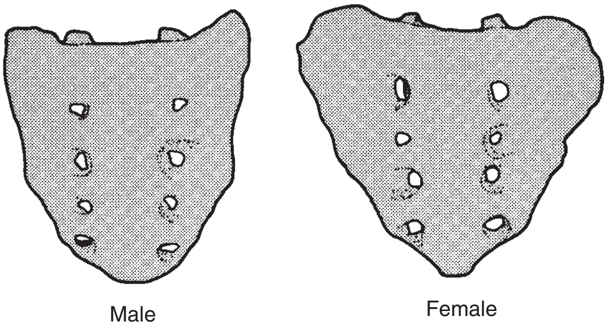

FIGURE 3.12 *The variation in shape of the male and female sacrum. The female is broader and more triangular.*

TABLE 3.2 *Dwight’s table for sexing from humeral and femoral head diameters (in mm)*

	Vertical humeral	Transverse humeral	Vertical femoral
Female	42.67	36.98	43.84
Male	48.76	44.66	49.68
Difference	6.09	5.68	5.84

TABLE 3.3 *Pearson and Bell’s table for the ‘mathematical sexing of the femur’. The area of uncertainty extends in increasing degrees of confidence outwards from the central column to the more definite limits on each side. Measurements are given in millimetres*

	Male	Male or female	Female
Vertical diameter of head	>45.5	43.5–41.5	<41.5
Popliteal length	>145	114–132	<106
Bicondylar width	>78	74–76	<72
Oblique trochanteric length	>450	405–430	<390

The size of the femoral heads is claimed to be a better discriminant of sex, the vertical diameter being said by Pearson and Bell to be greater than 45 mm in the male and less than 41 mm in the female, though again there is an overlap in the distribution curves around the 43 mm size (Table 3.2). Maltby (1917), however, measured 43–56 mm in the male and 37–46 mm in the female. Femoral head size is part of Pearson’s ‘mathematical sexing of the femur’, which incorporates several measurements (Table 3.3). Dwight (1904) studied the size of both femoral and humeral heads, claiming that they were more useful than bone length. Once again, discriminant function sexing has been carried out, using a number of measurements. Details should be sought in the work of Black and of Iscan and Miller-Shaivitz.

Another sex trait in the femur is the angle that the shaft makes with the vertical. Because the pelvis is relatively wider in the female, the shafts have to slope more to converge at knee level, so that the condyles at the lower end of

TABLE 3.4 *Long-bone lengths (mm) in White men and women of different stature*

Men							**Women**						
Humerus	**Radius**	**Ulna**	**Stature**	**Femur**	**Tibia**	**Fibula**	**Humerus**	**Radius**	**Ulna**	**Stature**	**Femur**	**Tibia**	**Fibula**
295	213	227	1530	392	319	318	263	193	203	1400	363	284	283
298	216	231	1552	398	324	323	266	195	206	1420	368	289	288
302	219	235	1571	404	330	328	270	197	209	1440	373	294	293
306	222	239	1590	410	335	333	273	199	212	1455	378	299	298
309	225	243	1605	416	340	338	276	201	215	1470	383	304	303
313	229	246	1625	422	346	344	279	203	217	1488	388	309	307
316	232	249	1634	428	351	349	282	205	219	1497	393	314	311
320	236	253	1644	434	357	353	285	207	222	1513	398	319	316
324	239	257	1654	440	362	358	289	209	225	1528	403	324	320
328	243	260	1666	446	368	363	292	211	228	1543	408	329	325
332	246	263	1677	453	373	368	297	214	231	1556	415	334	330
336	249	266	1686	460	378	373	302	218	235	1568	422	340	336
340	252	270	1697	467	383	378	307	222	239	1582	429	346	341
344	255	273	1716	475	389	383	313	226	243	1595	436	352	346
348	258	276	1730	482	394	388	318	230	247	1612	443	358	351
352	261	280	1755	490	400	393	324	234	251	1630	450	364	356
356	264	283	1767	497	405	398	329	238	254	1650	457	370	361
360	267	287	1785	504	410	403	334	242	258	1670	464	376	366
364	270	290	1812	512	415	408	339	246	261	1692	471	382	371
368	273	293	1830	519	420	413	344	250	264	1715	478	388	376

Modified by Krogman and Iscan (1986) from Hrdlicka (1939).

the femur sit horizontally on the tibial plateau. Thus when a female femur is seated on its condyles on a flat surface, the angle the shaft makes with that surface is of the order of 76°, while a male bone is more upright, the angle being around 80°. The angle of the neck on the shaft of the femur (the 'collodiaphyseal angle') was studied by Godycki (1957), the results suggesting that a bone with an angle of less than 40° had an 85 per cent chance of being male, while if the angle exceeded 50°, there was a 75 per cent chance of it being from a female (Table 3.4).

Most workers have worked with dry bone specimens; when methods using fresh bone are used, allowance must be made for articular cartilage where relevant. For example, the vertical diameter of a femoral head is 3 mm less in the dried specimen.

Sex determination from other bones

The sternum may be helpful in that the length of the manubrium in the female may equal or exceed half the length of the body, while the manubrium of the male is less than half the body length (Table 3.5). This was claimed in the nineteenth century by Hyrtl, but was denied by Krogman and by Dwight. The latter claimed that the ratio of manubrium: body was 52:100 in women and 49:100 in men, a poor discrimination. However, the method has been rehabilitated by Iordanidis (1961), who alleged a success rate of 80 per cent using the sternum alone. Stewart and McCormick (1983) used a radiographic technique and claimed total accuracy in stating that a sternal length of less than 121 mm must be female and over 173 mm must be male.

TABLE 3.5 *Sex determination from the sternum. Mean (range) lengths of manubrium and sternal body (mm)*

	Male	**Female**
Manubrial length	51.7 (41–73)	48.4 (39–61)
Body length	95.4 (74–122)	78.6 (59–95)
Combined length	147 (131–180)	127 (107–140)

Taken from Jit *et al.* (1980) North Indian population.

The scapula has been studied extensively, but mostly in relation to age. There are relatively poor sex variations in the vertical diameter of the glenoid cavity: the threshold, according to Dwight, is 36 mm, those smaller being female. Iordanidis made an extensive study of scapular measurements and concluded that scapular height was the best discriminant, the male usually being greater than 157 mm, the female less than 144 mm.

The humerus, radius and ulna yield little helpful sexing evidence, apart from overall size. The presence of a perforated olecranon fossa at the lower end of the humerus occurs more often in females and more commonly on the left side, there being a 3.7:1 ratio compared with males. Godycki studied this and other characteristics of the arm bones as sex determinants, but their value is poor.

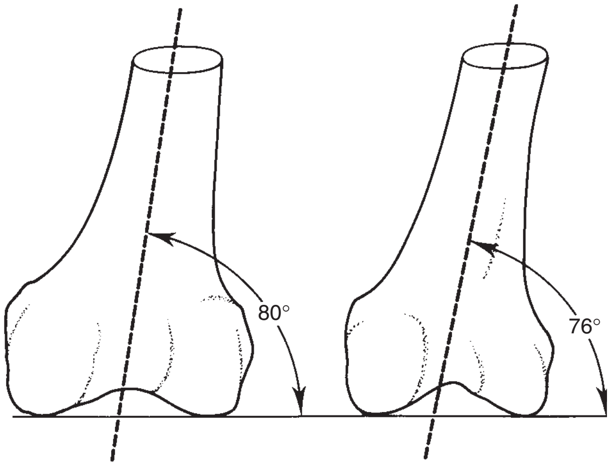

FIGURE 3.13 *Determination of sex of femur by angulation of the shaft. The condyles are set on a flat surface and the angle of the shaft measured. Because of the greater tilt caused by the relatively wider pelvis, the female angle is usually less than 80°.*

There are many reports on sexing from limb and girdle bones and the best approach seems to be a multiple assessment using the data for discriminant function analysis.

Evidence of pregnancy from the skeleton

More accurately, parturition causes some changes in the pelvis as a result of the local trauma of childbirth, which is reinforced by multiple pregnancies. These include 'pubic scars' from the tearing of tendon insertions and periosteum around the pubic bone. The dorsal pubic surface and the pre-auricular sulcus are best indicators, but most authorities agree that it is not possible to determine the number of births from osteological appearances. The papers of Angel and of Ullrich are most useful in this respect.

ESTIMATION OF STATURE FROM SKELETAL REMAINS

When a full skeleton is available, then obviously direct measurement of the correctly assembled bones will give the original height within a few centimetres. Allowance has to be made during assembly, however, for the loss of cartilage in joint spaces and especially for missing intervertebral discs. The accuracy achieved with direct measurement is not great, in that even with an intact body there can be a lengthening of up to 2.5 cm compared with the live height. The skeleton needs scalp and heel soft tissue thickness added – and with the obvious uncertainty about the cumulative allowances for disc thickness and joint cartilage, it is unrealistic to expect accuracy of less than about 4–8 cm compared with the living height.

When incomplete skeletal material is available, calculations have to be made on the basis of one or more bones. Where possible, all available bones should be used and a consensus of results assessed, though the accuracy derived from different bones varies and the result from – say, a femur – is more reliable than that from an ulna. The descending order of usefulness is: femur, tibia, humerus, radius.

Many formulae have been constructed for the determination of stature from the length of long bones. The following cautions must be appreciated before they are used:

- The tables were constructed from differing ethnic groups and at different times, so that racial and nutritional factors introduce significant variations. The table most appropriate to the bones under investigation should be used, though by definition the exact provenance of these bones may be unknown, even including their ethnic origins.
- There is a marked sex difference in stature estimation and ageing of the person also reduces stature relative to long bone length. Trotter and Gleser (1958) found in 855 bodies that there was a 1.2 cm loss of height for each two decades of age over the age of 30, a loss of height equivalent to about 0.6 mm a year after the fourth decade. The same investigation showed that the length of the cadaver was about 2.5 cm longer than the known height during life. The maximum stature (at least in American men) was found in 1951 to be reached at the age of 23 years, a change from earlier estimates in which the peak was found to occur between 18–21 years of age.
- The bone lengths must be measured in exactly the same way as that used by the author of the tables.
- Old dry bones are slightly shorter than recent material, even when one excludes the loss of articular cartilage. Telkkä (1950) stated that 2 mm must be deducted from the length of fresh bone before the calculation is made, as most tables are based on dried skeletons. These differences, however, are slight compared with the greater errors likely if incorrect measuring methods are employed.
- There are inbuilt errors, which are expressed in the formulae. These standard errors are given as 'plus or minus' a certain figure, which applies only when the length of the bone is near the mean of the usual range of length for that particular bone. When the bone is near the extremes of the range (that is from a very tall or very short person), then the standard deviation must be doubled in order to stay within the same range of confidence of accuracy. The same holds good if the tables used are not from the same ethnic group as those

TABLE 3.6 *The stature tables of Trotter and Gleser. The upper table represents their original 1952 figures, the lower their modification of 1977, as quoted by Krogman and Iscan (1986)*

Male Whites	**Male Negroes**
Stature = 63.05 + 1.31 (femur + fibula) ± 3.63 cm	Stature = 67.77 + 1.20 (femur + fibula) ± 3.63 cm
Stature = 67.09 + 1.26 (femur + tibia) ± 3.74 cm	Stature = 71.75 + 1.15 (femur + tibia) ± 3.68 cm
Stature = 75.50 + 2.60 fibula ± 3.86 cm	Stature = 72.22 + 2.10 femur ± 3.91 cm
Stature = 65.53 + 2.32 femur ± 3.94 cm	Stature = 85.36 + 2.19 tibia ± 3.96 cm
Stature = 81.93 + 2.42 tibia ± 4.00 cm	Stature = 80.07 + 2.34 fibula ± 4.02 cm
Stature = 67.97 + 1.82 (humerus + radius) ± 4.31 cm	Stature = 73.08 + 1.66 (humerus + radius) ± 4.18 cm
Stature = 66.98 + (humerus + ulna) ± 4.37 cm	Stature = 70.67 + 1.65 (humerus + ulna) ± 4.23 cm
Stature = 78.10 + 2.89 humerus ± 4.57	Stature = 75.48 + 2.88 humerus ± 4.23 cm
Stature = 79.42 + 3.79 radius ± 4.66	Stature = 85.43 + 3.32 radius ± 4.57 cm
Stature = 75.55 + 3.76 ulna ± 4.72	Stature = 82.77 + 3.20 ulna ± 4.74 cm
Female Whites	**Female Negroes**
Stature = 50.12 + 0.68 humerus + 1.17 femur + 1.15 tibia ± 3.51 cm	Stature = 56.33 + 0.44 humerus – 0.20 radius + 1.46 femur + 0.86 tibia ± 3.22 cm
Stature = 53.20 + 1.39 (femur + tibia) ± 3.55 cm	Stature = 58.54 + 1.53 femur + 0.96 tibia ± 3.23 cm
Stature = 53.07 + 1.48 femur + 1.28 tibia ± 3.55 cm	Stature = 59.72 + 1.26 (femur + tibia) ± 3.28 cm
Stature = 59.61 + 2.93 fibula ± 3.57 cm	Stature = 59.76 + 2.28 femur ± 3.41 cm
Stature = 61.53 + 2.90 tibia ± 3.66 cm	Stature = 62.80 + 1.08 humerus + 1.79 tibia ± 3.58 cm
Stature = 52.77 + 1.35 humerus + 1.95 tibia ± 3.67 cm	Stature = 72.65 + 2.45 tibia ± 3.70 cm
Stature = 54.10 + 2.47 femur ± 3.72 cm	Stature = 70.90 + 2.49 fibula ± 3.80 cm
Stature = 54.93 + 4.74 radius ± 4.24 cm	Stature = 64.67 + 3.08 humerus ± 4.25 cm
Stature = 57.76 + 4.27 ulna ± 4.20 cm	Stature = 75.38 + 3.31 ulna ± 4.83 cm
Stature = 57.97 + 3.36 humerus ± 4.45 cm	Stature = 94.51 + 2.75 radius ± 5.05 cm

White males	SE	**Black males**	SE
3.08 Humerus + 70.45	4.05	3.26 Humerus + 62.10	4.43
3.78 Radius + 79.01	4.32	3.42 Radius + 81.56	4.30
3.70 Ulna + 74.05	4.32	3.26 Ulna + 79.29	4.42
2.38 Femur + 61.41	3.27	2.11 Femur + 70.35	3.94
2.52 Tibia + 78.62	3.37	2.19 Tibia + 86.02	3.78
2.68 Fibula + 71.78	3.29	2.19 Fibula + 85.65	4.08
1.30 (Femur + Tibia) + 63.29	2.99	1.15 (Femur + Tibia) + 71.04	3.53
1.42 Femur + 1.24 Tibia + 59.88	2.99	0.66 Femur + 1.62 Tibia + 76.13	3.49
0.93 Humerus + 1.94 Tibia + 69.30	3.26	0.90 Humerus + 1.78 Tibia + 71.29	3.49
0.27 Humerus + 1.32 Femur+ 1.16 Tibia + 58.57	2.99	0.89 Humerus – 1.01 Radius + 0.38 Femur + 1.92 Tibia + 74.56	3.38
White females	SE	**Black females**	SE
3.36 Humerus + 57.97	4.45	3.08 Humerus + 64.67	4.25
4.74 Radius + 54.93	4.24	3.67 Radius + 71.79	4.59
4.27 Ulna + 57.76	4.30	3.31 Ulna + 75.38	4.83
2.47 Femur + 54.10	3.72	2.28 Femur + 59.76	3.41
2.90 Tibia + 61.53	3.66	2.45 Tibia + 72.65	3.70
2.93 Fibula + 59.61	3.57	2.49 Fibula + 70.90	3.80
1.39 (Femur + Tibia) + 53.20	3.55	1.26 (Femur + Tib) + 59.72	3.28
1.48 Femur + 1.28 Tibia + 53.07	3.55	1.53 Femur + 0.96 Tibia + 58.54	3.23
1.35 Humerus + 1.95 Tibia + 52.77	3.67	1.08 Humerus + 1.79 Tib	3.58
0.68 Humerus + 1.17 Femur + 1.15 Tibia + 50.122	3.51	0.44 Humerus – 0.20 Radius + 1.46 Femur + 0.86 Tibia + 56.33	3.22

All lengths are in centimetres, valid only for Americans between 18 and 30 years of age. Femur and tibia are maximum lengths. SE, standard error.

being estimated. Krogman indicated that a 95 per cent confidence rate applies only if twice the standard error is accepted each side of the favoured height – in an average man, this means a latitude of over 12 cm, which is too great to be of much use in identification.

Of the many available systems of calculation, those of Trotter and Gleser are most widely used, being based on Caucasian and Negroid Americans from the 1950s (Table 3.6). Older tables include those of Rollet whose measurements of French people were later reworked by Manouvrier and

Table 3.7 *Pearson's (1899) regression tables for calculating stature from dried long bones*

S = 81.306 + 1.880 Femur	S = 72.844 + 1.945 Femur
S = 70.651 + 2.894 Humerus	S = 71.475 + 2.754 Humerus
S = 78.664 + 3.378 Tibia	S = 74.774 + 2.352 Tibia
S = 85.925 + 3.271 Radius	S = 81.224 + 3.343 Radius
S = 71.272 + 1.159 (F + T)	S = 69.154 + 1.126 (F + T)
S = 71.441 + 1.220 F + 1.080 T	S = 69.561 + 1.117 F + 1.125 T
S = 66.855 + 1.730 (H + R)	S = 69.911 + 1.628 (H + R)
S = 69.788 + 2.769 H + 0.1958	S = 70.542 + 2.582 H + 0.2818
S = 68.397 + 1.030 F + 1.557 H	S = 67.435 + 1.339 F + 1.027 H
S = 67.049 + 0.913 F + 0.600 T	S = 67.467 + 0.782 F + 1.120 T

by Pearson and quoted in a review by Hrdlicka (1939). Pearson allowed a standard addition of 2.5 cm to compensate for the difference between dead and living height (Table 3.7). More modern tables include those of Dupertuis and Hadden (1951) and Trotter and Gleser (1958), from American subjects. The latter used casualties from World War II and later from the Korean War to construct their tables (Table 3.8). Other calculation systems include Breitinger (1937; Germans), Telkkä (Finns) (Table 3.9), Allbrook (1961; African Negroids and English), Shiati (1983; Chinese), Mendes-Correa (1932; Portuguese) and Stevenson (1929; Chinese). Modern critics of the accuracy of the estimation of stature from long bones include Wells (1959), who suggested that even the preferred method of Trotter and Gleser has less accuracy than is generally acknowledged.

Estimation of stature from bones other than the major limb bones is far more inaccurate. Measuring the spinal column (from the tip of the odontoid to the bottom of the fifth lumbar vertebra) has been used by Krogman and by Dwight; Fully and Pineau (1960) used the spine plus a long bone. Tibbetts (1981), in his more modern study on the spine, measured 23 individual vertebrae and produced a regression formula for both men and women. However, the results were far superior when long bones were used. Others have employed the clavicle (Jit and Singh 1956) and the scapula (Olivier 1969).

Table 3.9 *Telkkä's table for calculating stature of Finnish men and women*

Men	SE	**Women**	SE
169.4 + 2.8 (Humerus – 32.9)	5.0	156.8 + 2.7 (Humerus – 30.7)	3.9
169.4 + 3.4 (Radius – 22.7)	5.0	156.8 + 3.1 (Radius – 20.8)	4.5
169.4 + 3.2 (Ulna – 23.1)	5.2	156.8 + 3.3 (Ulna – 21.3)	4.4
169.4 + 2.1 (Femur – 45.5)	4.9	156.8 + 1.8 (Femur – 41.8)	4.0
169.4 + 2.1 (Tibia – 36.6)	4.6	156.8 + 1.9 (Tibia – 33.1)	4.6
169.4 + 2.5 (Fibula – 36.1)	4.4	156.8 + 2.3 (Fibula – 32.7)	4.5

The use of fragmentary bones and those immature bones from which the epiphyses have been lost have been studied, the survey by Krogman and Iscan being the best account of the techniques employed by researchers such as Steele and McKern (1969) and Muller (1935). The estimation of the heights of infants and fetuses is even more difficult when working with skeletal material, as substantial parts of these bones may be detached and missing because of non-fusion of epiphyses and non-appearance of ossification centres. Mehta and Singh (1972) have specifically addressed this problem, as well as others summarized by Krogman and Iscan (1986).

ESTIMATING THE SUBJECT'S AGE FROM SKELETAL STRUCTURES

The age of the subject at death is usually more important than the stature and rivals the sex as the most vital indication

Table 3.8 *Dupertuis and Hadden's (1951) tables for estimating stature from bones. The length of the bones (measured in centimetres according to the authors' criteria) is multiplied by the factor in the table and added to the constant in the right hand column to give the body length (cm)*

Men	**cm**	**Women**	**cm**
2.238 (Femur)	+ 69.089	2.317 (Femur)	+ 61.412
2.392 (Tibia)	+ 81.688	2.533 (Tibia)	+ 72.572
2.970 (Humerus)	+ 73.570	3.144 (Humerus)	+ 64.977
3.650 (Radius)	+ 80.405	3.876 (Radius)	+ 73.502
1.225 (Femur + Tibia)	+ 69.294	1.233 (Femur + Tibia)	+ 65.213
1.728 (Humerus + Radius)	+ 71.429	1.984 (Humerus + Radius)	+ 55.729
1.422 (Femur) + 1.062 (Tibia)	+ 66.544	1.657 (Femur) + 0.879 (Tibia)	+ 59.259
1.789 (Humerus) + 1.841 (Radius)	+ 66.400	2.164 (Humerus) + 1.525 (Radius)	+ 60.344
1.928 (Femur) 0.568 (Humerus)	+ 64.505	2.009 (Femur) + 0.566 (Humerus)	+ 57.600
1.442 (Femur) + 0.931 (Tibia) + 0.083 (Humerus) + 0.480 (Radius)	+ 56.006	1.544 (Femur) + 0.764 (Tibia) + 0.126 (Humerus) + 0.295 (Radius)	+ 57.495

of identity. The estimation of skeletal age falls into various groups, which have marked differences in both method and accuracy. In general, the greater the personal age the less the confidence quotient. There are many publications on the subject, many of them arising from archaeological rather than forensic interest, as the age structure of a skeletal population is of great interest to social anthropologists and historians.

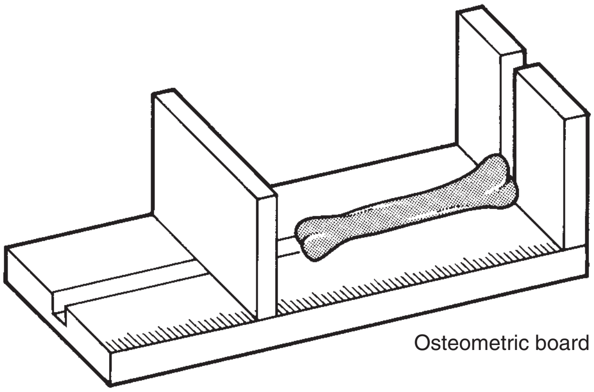

FIGURE 3.14 *Osteometric board for accurate measurement of bone length. There is a central slot in the headboard to accommodate the tibial head spines.*

The fetus and young infant

It is more usual to have to estimate fetal and neonatal age on the intact dead body, rather than the skeleton, as the immature bones are so easily dispersed, lost and destroyed compared to the more robust bones of older subjects. Both in archaeology and forensic pathology, however, such fetal and immature remains are sometimes present.

The major dating indices are the appearance of ossification centres in growing cartilage but, as stated above, these are rarely available in dried bones as the cartilage disintegrates within weeks or months and small ossification centres – both in diaphyses and epiphyses – rarely survive. The bones have to be examined with the cartilage still attached to offer much hope of ageing the remains. Radiology may provide more information than visual inspection.

Specialized books such as the classics by Fazekas and Kosa, Krogman and Iscan, and by Stewart and the numerous papers and monographs on the subject (a selection of which are listed at the end of this chapter) must be consulted where fetal issues are complex and important. Once again the help of interested anatomists and radiologists can be invaluable, especially as biological variation poses a constant trap for the inexperienced who slavishly follow printed tables without the knowledge to appreciate their limitations. Stewart has

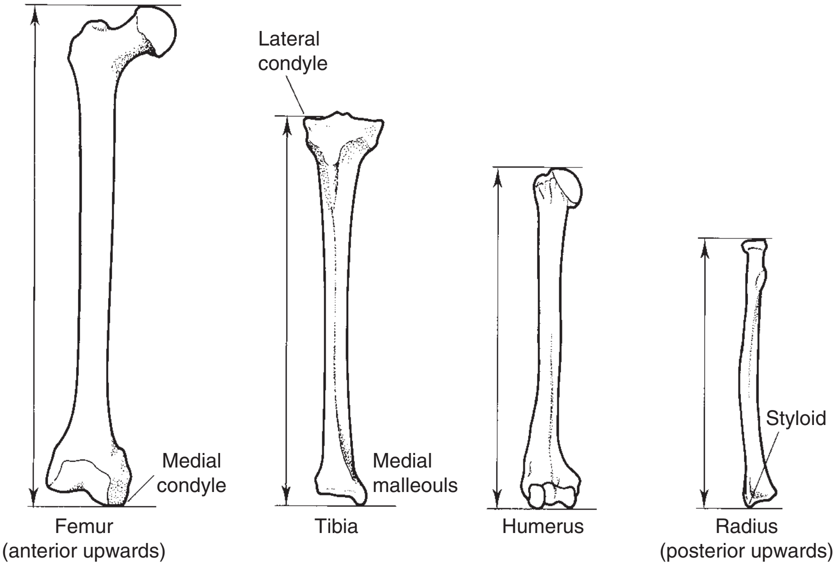

FIGURE 3.15 *Dimensions of dried bones for estimations of stature (Trotter and Gleser, Dupertuis and Hadden). The right side bone is used for preference.* ***Femur:*** *With the bone lying anterior surface upwards, the maximum length is measured from the medial condyle to the most proximal part of the head.* ***Tibia:*** *Maximum length between tip of medial malleolus and face of lateral condyle. The intercondylar eminence is not included.* ***Humerus:*** *Maximum overall length from the posterior margin of the trochlea to the upper edge of the head.* ***Radius:*** *From the tip of the styloid process to the head, lying with the posterior surface upwards. All bones should be measured either with an osteometric board or, if one is not available, on a flat bench with the maximum lengths taken between two vertical, parallel boards placed in contact with the bone ends. If the bones are not dry, but have articular cartilage in place, the following should be subtracted from the measured length before applying the formulae (Boyd and Trevor): radius and humerus 3 mm each, tibia 5 mm and femur 7 mm.*

produced a useful nomogram that relates fetal femoral length to crown–rump length and hence to approximate gestational age. Teeth are discussed in Chapter 26. Fazeka and Kosa's book is particularly useful to the pathologist (as bone lengths, etc. are directly related to fetal and infant age), as are the profuse illustrations.

Skeletal age in the child and young adult

There is no break in the methods used for the fetus and small infant and the older child and young adult. The appearance of ossification centres is complete by around 5 years and after this stage the fusion of epiphyses acts as a calendar up to the age of about 25, when the medial clavicular epiphysis is usually the last evidence of such fusion.

Lists and diagrams such as those shown here are the usual method of tracing the maturity of the skeleton. Several cautionary points must be appreciated and where possible the advice of a radiologist be obtained if there are important issues at stake. The paper by McKern and Stewart gives a valuable commentary on personal variation in skeletal age determination (Table 3.10).

- Maturity is not synonymous with calendar age. There are both sexual, racial, nutritional and other biological variations.
- Females are almost always in advance of males, and maturity tends to be accelerated in hotter climates, though the latter may be tempered by nutritional disadvantages.
- There is a marked range of closure dates in epiphyseal union, and the year suggested in some charts is merely the midpoint of that range.
- Union is a process, not an event. It may be different radiologically from gross inspection; it may also be slow in completion, so that it may be difficult to pinpoint a date that is needed to refer to published data. An example is the one that is usually the last to fuse, the medial end of the clavicle, which may slowly close during any period from about 18 to 31 years. Finally, as emphasized by Krogman and Iscan, multiple criteria of skeletal age should be used, dependence not being placed upon any single measurement.

Skeletal ageing in later years

The eruption of the third molar teeth and the fusion of the last epiphysis occurs approximately in the middle of the third decade. This is a watershed in skeletal dating, as the more objective markers of age are almost all on the younger side of this time. From around 25 years until old age, there are no dramatic events such as tooth eruption or the appearance of ossification centres. There are more subtle changes available for specialized interpretation, but the general decline through middle age to senility prevents assessment of age to within the nearest half-decade.

The major advances in this difficult period have been the use of the pubic symphysis and sternal ribs and the radiology of cancellous bone. However, dental technology may further refine the assessment of age.

Pubic symphysis and age

The opposing faces of the two pubic bones have been extensively studied by Todd (1920, 1921), by McKern and Stewart (1957) and by Gilbert and McKern (1973), who related the changing topography of the symphyseal faces to age. The technique is complex and requires bones that are free of cartilage, but not so eroded by drying and damage that the surface features are blurred.

ASSESSMENT OF AGE FROM THE MALE PUBIC SYMPHYSIS

As always, the details given in the original publications should be studied and practice gained by examining bones of known age. The method as originally devised by Todd has been radically modified by McKern and Stewart so that the technique is now the most useful in assessing the age of post-mature material. It must be noted that their method applies only to males and was not applied to female material until the work of Gilbert and McKern.

In summary, the face of the symphysis is analysed by reference to three 'components', each being scored on a scale of 6. In component I, the dorsal half of the face is assessed on a scale of 0–5; in component II, the ventral half is assessed; and in component III, the whole surface is considered in relation to different criteria to the preceding two stages.

Component I (the dorsal plateau) is scored as follows:

- 0: dorsal margin absent
- 1: slight margin formation appears in the middle third of the dorsal border
- 2: dorsal margin extends along whole dorsal border
- 3: grooves filled in with resorption of ridges to form an early plateau in the middle third of the dorsal demiface
- 4: te plateau, still with vestiges of billowing, extends over most of the dorsal demiface
- 5: billowing vanishes and the surface of the demiface becomes flat and slightly granular.

TABLE 3.10 *Time of appearance of major ossification centres*

(a) Healthy Caucasian boys

Birth
Calcaneus
Talus
Femur, distal
Tibia, proximal
Cuboid
Humerus, head

2 months
Capitate
Hamate
Lateral cuneiform

3 months
Femur, head
Capitulum
Tibia, distal

6 months
Fibula, distal

7 months
Humerus, greater tuberosity
Radius, distal

10 months
Triquetrum

11 months
Third finger–first phalanx
First toe–second phalanx

12 months
Second finger–first phalanx
Fourth finger–first phalanx
First finger–second phalanx

13 months
Third toe–first phalanx
Second metacarpal
Medial cuneiform

14 months
Fourth toe–first phalanx
Second toe–first phalanx
Fifth toe–second phalanx

15 months
Third metacarpal
Second toe–second phalanx
Fifth finger–first phalanx

16 months
Fourth toe second phalanx
Fourth metacarpal

18 months
Second finger–second phalanx
Third finger–second phalanx
Fourth finger–second phalanx
Fifth metacarpal

20 months
First toe–first phalanx
Middle cuneiform

21 months
Third finger–third phalanx
Fourth finger–third phalanx
Navicular of foot
Fifth toe–first phalanx

22 months
First metacarpal
First metatarsal

23 months
First finger–first phalanx

2 years
Fifth finger–second phalanx
Lunate

2 years, 2 months
Second metatarsal

2 years, 5 months
Second finger–third phalanx
Fifth finger–third phalanx

2 years, 11 months
Third metatarsal
Fibula, proximal

3 years, 1 month
Femur, great trochanter
Patella

3 years, 3 months
Fourth metatarsal

3 years, 4 months
Fifth toe–third phalanx

3 years, 7 months
Third toe–third phalanx
Fourth toe–third phalanx

3 years, 8 months
Fifth metatarsal
Second finger–third phalanx

3 years, 10 months
Radius, proximal

4 years, 2 months
Multangulate majus

4 years, 4 months
Navicular, hand

4 years, 8 months
Multangulate minus

5 years+
Humerus, medial epicond
Ulna, distal
Fifth toe–second phalanx

(b) Healthy Caucasian girls

Birth
Calcaneus
Talus
Femur, distal
Tibia, proximal
Cuboid
Humerus, head

2 months
Capitate
Hamate
Lateral cuneiform

3 months
Femur, head
Capitulum
Tibia, distal

4 months
Humerus, greater tuberosity

6 months
Fibula, distal
Radius, distal

7 months
First toe–second phalanx
Third finger–first phalanx
Fourth finger–first phalanx

8 months
Second finger–first phalanx
First finger–second phalanx
Third toe–first phalanx

9 months
Third toe–second phalanx
Fourth toe–first phalanx
Medial cuneiform

10 months
Second metacarpal
Second toe–second phalanx
Fourth toe–second phalanx
Third metacarpal
Second toe–first phalanx
Triquetrum

11 months
Fourth metacarpal
Fifth finger–first phalanx

12 months
Fourth finger–second phalanx
Third finger–second phalanx

13 months
Fifth metacarpal
Second finger–second phalanx

14 months
First metacarpal
First toe–first phalanx
Fifth finger–first phalanx
Third finger–third phalanx
Fourth finger–third phalanx
Navicular of foot
Middle cuneiform
First metatarsal

15 *months*
First finger–first phalanx
Fifth finger–second phalanx

17 months
Second finger–third phalanx
Fifth finger–third phalanx

19 months
Second metatarsal

21 months
Fifth toe–third phalanx

22 months
Third metatarsal

23 months
Patella

2 years
Lunate
Third toe–third phalanx
Fourth toe–third phalanx
Fibula, proximal
Femur, greater trochanter

2 years, 2 months
Second toe–third phalanx
Fourth metatarsal

2 years, 5 months
Fifth metatarsal

2 years, 8 months
Multangulate majus

2 years, 9 months
Humerus, medial epicon

3 years
Radius, proximal
Multangulate minus

3 years, 2 months
Navicular, hand

4 years, 6 months
Ulna, distal

5 years+
Fifth toe–second phalanx

After Francis and Werle (1939).

Component II (the ventral rampart) is scored as follows:

- 0: no ventral bevelling
- 1: ventral bevelling only at the superior end of the ventral border
- 2: bevelling extends along whole ventral border
- 3: ventral rampant starts as bony extensions from either or both extremities
- 4: extensive rampart, but still gaps, especially on the upper two thirds of the ventral border
- 5: complete ventral rampart.

Component III (symphyseal rim) is scored as follows:

- 0: no symphyseal rim
- 1: partial dorsal rim visible, usually at the upper end of the border; it is round, smooth and elevated above the surface
- 2: complete dorsal rim with beginnings of ventral rim, which starts at no particular site
- 3: complete circumferential symphyseal rim, enclosing a finely grained, slightly undulating surface
- 4: the rim begins to break down, no longer smooth but sharply defined. Some lipping on ventral edge
- 5: further breakdown of rim, especially along superior ventral edge. Rarefaction of symphyseal face. Disintegration and erratic ossification of ventral rim.

In practice, a three-figure score is made from evaluation of each component and added together. For example, component I = 3, component II = 4, component III = 1. Thus the score is 3 + 4 + 1 = 8 and by reference to the tables, the age suggested is between 24 and 28.

Since the work of Gilbert and McKern, a method has been offered to assess age in both males and females. Stewart has pointed out that, although Todd felt that the same criteria could be applied to females, there was risk of overestimation of age in women because pelvic trauma during parturition could deform the dorsal border of the symphysis in a manner likely to mimic age changes.

The other limitation of the method is that the upper limit is 50 years, so that a large segment of the range is not considered. Also the scoring is naturally subjective, depending on the experience and training of the observer. In spite of these drawbacks, the method remains the most useful ageing technique for post-mature males.

Gilbert and McKern established standards for females in 1973. The same three components are used, but the description of the symphyseal faces is different. Details must be sought in the original publications or in Krogman and Iscan's invaluable book. Gilbert concluded that, if the male criteria were used for females, the age assessment would be underestimated by about a decade, as the female pubis reaches full maturity about 10 years later than the male.

Several tests of the accuracy of this method have been made by Suchey (1979) and Meindl *et al.* (1985). There have been many reports on the subject, but it would seem that Todd's system remains the most accurate. Overall accuracy obtained by experienced observers lies in the range of 5–7 years around the true age.

The sternal rib methods

The costochondral junction has been studied by a number of physical anthropologists using different techniques. Michelson (1934) found that calcification in the first costal cartilage did not occur under 11 years and that after 16 years, males calcified much more quickly until 66, when the sexes again became uniform. Iscan and Loth (1986) made detailed studies of the shape of the rib end adjacent to the cartilage and constructed complex criteria for age-related changes. The original publications must be consulted for details as well as more recent developments such as those by Loth, Iscan and Sheuerman (1994).

Skull sutures and age

The use of skull suture fusion as an index of age has had a chequered history, beginning in the first century AD with a comment by Celsus. It is now generally discredited, except in the most broad terms.

It is common knowledge that most adults have at least part of their suture lines closed and that this tends to become more widespread as age increases. There are many exceptions and the rate of closure is not linear with time. This generality can be useful when skull fragments are found, as any visible fusion will at least indicate that the skull came from a mature individual, as it is unlikely below the age of 20 (Brothwell).

The claims of Dwight (1890), Parsons and Box (1905) and Todd and Lyon (1924, 1925) (who stated that sagittal fusion began at 22 and was complete by 35) for the use of fusion as an ageing tool, have been refuted by Singer (1953), Cobb (1952), McKern and Stewart (1957) and Genovés and Messmacher (1959). McKern and Stewart found that 25 per cent of 18-year-old males had begun closure of their sagittal sutures and that by 31–40 years, 90 per cent had some fusion. Yet many had no fusion at considerably older ages. Since these criticisms were made, much further work has been carried out, recorded by Krogman and Iscan's book in its 1986 edition. They concluded from the accumulated evidence of many publications that suture closure is not affected by sex, race or right/left differences. Only endocranial fusion must be studied, as that on the outer side is far more erratic. Even so, determination of age from suture closure is unsafe: they feel that in the 20–50 age range, it may be possible to place a skull within the correct decade only, older material being even more variable.

The basisphenoid synchondrosis cannot be included in this class, as its fusion is a relatively reliable indicator of a minimum age of about 20 years. The metopic suture, between the two halves of the frontal bone, usually closes at about 2 years of age, but occasionally persists into adult life.

Radiology can also assist in the determination of age, from the internal structure of the cancellous bone and cortical thickness of the head of the humerus, for example. Schranz (1959) developed a combination of external visual examination and radiographic features, indicating that the head of the humerus was a better determinator than the corresponding part of the femur. He produced a list of criteria that helped to age a skeleton from about 15 years of age up to one in excess of 75. This was pursued by Nemeskeri and later by him in conjunction with Acsadi (Acsadi and Nemeskeri 1970), to include the proximal parts of both humerus and femur, taking into account the radiological thinning of the cortex and the progressive rarefaction of the apex of the medullary cavity in the head of the bone. The original papers or their full synopsis in Krogman and Iscan's book must be consulted for details.

Histological structure has also received attention and remodelling of Haversion systems seem potentially useful

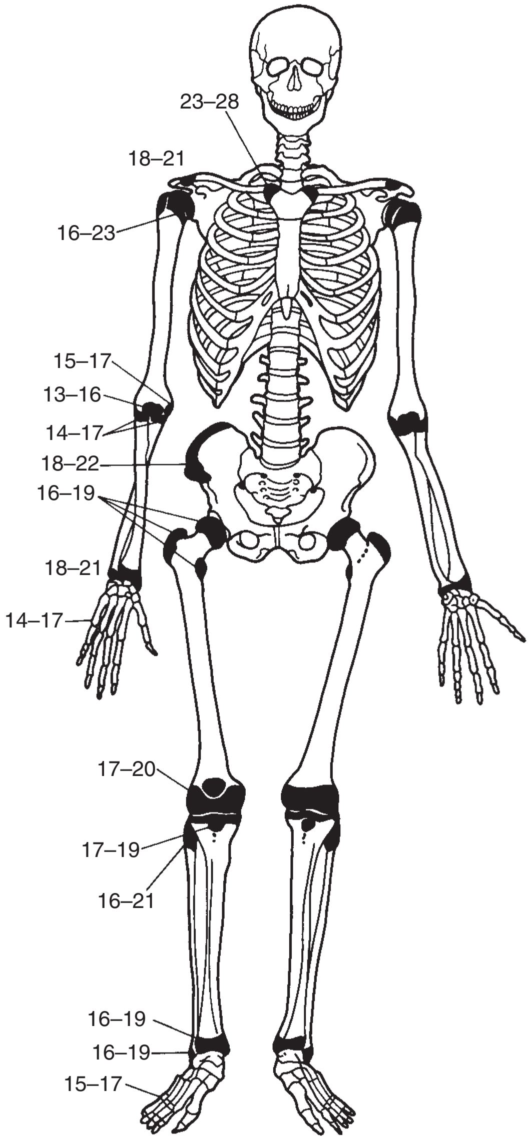

FIGURE 3.16 *A guide to the age of epiphyseal union in the major centres. The commencement and completion of union takes several years. The table is only a guide for male subjects (female slightly earlier) in non-tropical climates; the two dates are partial and complete union (years).*

Head of femur	16–19	Acromion	17–19
Greater trochanter	16–19	Distal femur	17–20
Lesser trochanter	16–19	Proximal tibia	17–19
Head of humerus	16–23	Proximal fibula	16–21
Distal humerus	13–16	Distal tibia	16–19
Medial epicondyle	16–17	Distal fibula	16–19
Proximal radius	14–17	Metatarsals	15–17
Proximal ulna	14–17	Iliac crest	18–22
Distal radius	18–21	Primary elements pelvis	14–16
Distal ulna	18–21	Sternal clavicle	23–28
Metacarpals	14–17	Acromial clavicle	18–21

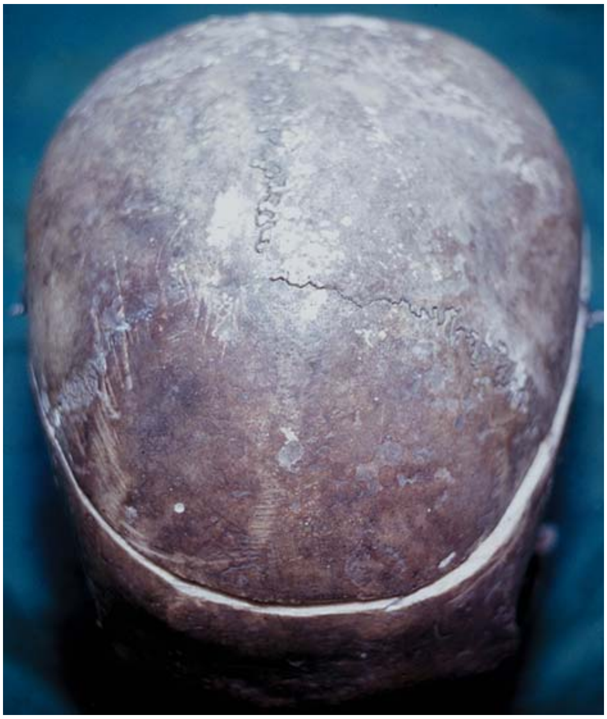

FIGURE 3.17 *The interpretation of cranial suture closure as an index of age is fraught with error. This person was 23 of age and, as would be generally expected, no segments of fusion are shown on the exterior of the skull.*

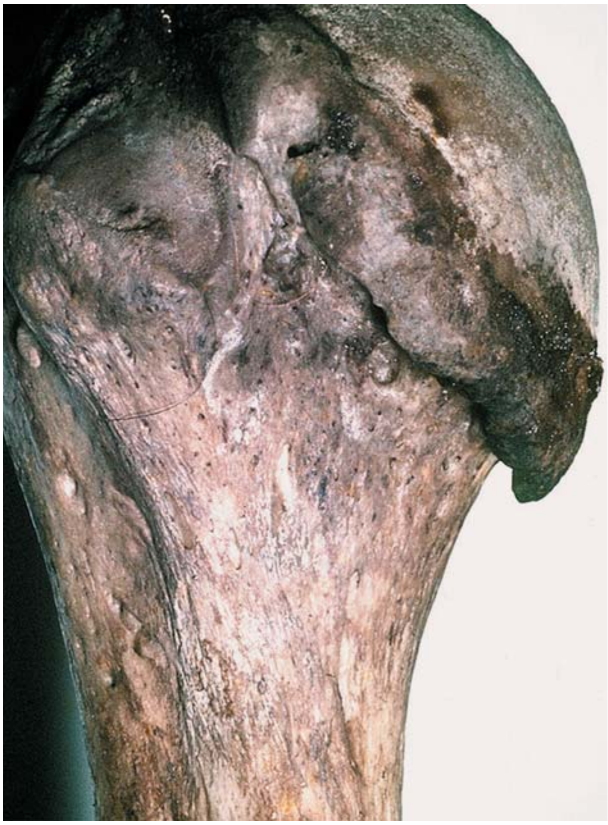

FIGURE 3.18 *Degenerative disease can assist in providing a general estimate of age, or at least exclusion of younger age groups. This grossly arthritic humerus assisted in distinguishing the near-skeletalized bodies of a 66-year-old man and his 22-year-old son.*

for the estimation of personal age. Eriksson and Westermark (1990) have claimed that amyloid inclusions in the choroid plexus can be used as a simple technique.

Chemical methods for assessing personal age lie partly within the field of forensic dentistry, such as the racemization of amino acids in teeth. This has also been applied to bone and depends upon the progressive conversion with advancing age of aspartic acid from the dextro- to the laevo-isomer (Ohtani 1994; Ritz *et al.* 1994).

DETERMINATION OF RACE FROM SKELETAL REMAINS

This is more difficult than any of the preceding investigations, partly because racial traits are not so marked and partly because so much ethnic mixing has taken place, especially in the large-scale immigrations into Europe and North America in recent years.

As usual, those at the extreme ends of the osteological range present little difficulty, as Brothwell (1972) remarks when comparing Eskimos with Australian Aboriginals. Unfortunately, there is a large proportion of racially neutral skeletal material and also some that presents apparently classical racial features but turns out to be totally atypical, so providing a quite incorrect diagnosis. The following criteria are offered, but the usual warnings about dogmatic opinions are even more important in this field.

There are three main racial groups: Caucasian, Mongoloid and Negroid. All others are derived from these and skeletally cannot be distinguished, though when archaeologists and anthropologists are working within a major race, there may well be local criteria that can distinguish racial subgroups with varying degrees of confidence. Most work has been performed in North America and thus most available data are derived from Caucasian and Negroid Americans, and Native Americans.

The skull offers the best evidence on racial origins; Krogman and Iscan claimed that race can be determined in 90–95 per cent of cases. The mandible is excluded from this, apart from the teeth, which will be dealt with in another chapter. It may be mentioned here, however, that one of the most useful pointers to a Mongoloid skull is the present of posteriorly concave, ‘shovel-shaped’, upper incisors, which may be grooved on the rear surface. These were found by Hanihara (1967) to be present in up to 93 per cent of Japanese, 85 per cent of Chinese and 68 per cent of Eskimos, being absent in Caucasians and occurring in only up to 15 per cent in Negroid Americans.

The wide zygomatic arches that give the typical high-cheekboned features of the Mongoloid races can sometimes produce a skull in which, uniquely, the transverse facial width is greater than the width of any other part of the head. Krogman provides a useful table (Table 3.11), outlining the major racial differences in the skull. Elsewhere he subdivides the Caucasian group into its three main subgroups: Nordic, Alpine and Mediterranean. He points out once again that these describe the ‘stereotypes’ of each group and that there is an extensive grey zone of more neutral characteristics. Subjects in the Negroid group tend to be long headed (dolichocephalic) and broad headed (brachycephalic) in the Mongoloid group, with the whole range occuring in the Caucasian group. The Negroid orbit is lower and wider, compared with the higher, rounder eye socket of the Mongoloid. The Negroid nasal aperture is wide, and the prognathism of the lower face and jaw relatively marked.

There are racial differences in the pelvis, but specialized measurements as described by Todd and Lindala (1928) must be made in order to detect these. Iscan suggests that a 79–83 per cent success rate can be achieved in differentiating between Caucasian and Negroid subjects (see References and further reading). The long bones offer some assistance, especially the femur. This is straighter in Negroid groups, the anteroposterior bowing being less. Together with other long bones, especially those of the lower limb, Negroid bones are longer than those of Caucasian or Mongoloid skeletons, even for equal total body length, the leg proportions being greater. The radius and tibia are relatively longer (Table 3.12). Several writers offer various indices of relative bone lengths, such as those of Modi (1957), Munter (1936) and Schultz (1937). The scapula has been extensively studied, but seems a poor discriminator of race.

Table 3.11 *General racial features in the skull (after Krogman 1939)*

Feature	White ('Caucasoid') (excluding Alpine)	Negroid	Mongoloid
Height of skull	High	Low	Medium
Length of skull	Long	Long	Long
Breadth of skull	Narrow	Narrow	Broad
Breadth of face	Narrow	Narrow	Wide
Height of face	High	Low	High
Sagittal contour	Rounded	Flat	Arches
Nasal opening	Narrow	Wide	Narrow
Orbital opening	Angular	Rectangular	Rounded
Lower nasal margin	Sharp	Troughed	Sharp
Nasal profile	Straight	Slanted down	Straight
Palatal shape	Narrow	Wide	Medium

PERSONAL IDENTITY FROM SKELETAL MATERIAL

After the general groups of race, sex, stature and age have been investigated, the bones must be examined for any idiosyncratic features that give them a personal identity. This is invariably a matter of matching up such feature(s) with known pre-existing conditions, and depends upon a potential match being available. Circumstances are infinitely variable and the ways in which this suspected matter comes to the examiner's knowledge are legion. Sometimes there may be one 'good bet' and, indeed, the potential identity may be virtually certain from the circumstances, the anatomical confirmation being almost a formality. In

Table 3.12 *Selected anthropometric dimensions (mm) in White and Black* Americans*

	Men				Women			
	White		Black		White		Black	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Body								
Stature	1706	60.5	1744	60.8	1597	78.1	1586	64.9
Total arm length	760	30.2	798	41.2	695	38.1	713	32.7
Upper arm length	333	16.4	340	21.9	302	22.8	303	15.6
Forearm length	249	12.4	269	15.0	225	16.7	237	14.7
Hand length	187	9.3	199	12.3	172	11.3	179	10.2
Leg length (anterior supine)	938	40.0	984	49.1	871	51.2	892	41.8
Thigh length	497	24.3	515	27.5	461	31.6	467	24.9
Lower leg length	376	19.8	401	23.7	352	25.8	365	19.8
Foot length	244	11.9	256	13.9	215	13.6	231	11.8
Bi-iliac breadth	292	16.9	270	16.5	287	22.5	267	24.7
Bi-spinous breadth	246	17.2	227	18.1	242	25.1	220	22.5
Head								
Maximum length	188	7.5	193	6.1	182	6.5	186	7.7
Maximum breadth	154	6.5	149	6.1	145	5.6	144	6.2
Auricular height	120	5.1	124	5.3	118	5.3	120	4.8
Face								
Bizygomatic breadth	139	6.7	139	6.1	130	7.0	132	8.7
Mandibular breadth	110	7.4	109	8.7	103	8.9	105	11.9
Interocular breadth	33	3.5	35	2.7	31	2.8	34	2.3
Total morphological height	122	9.7	125	7.6	113	9.5	116	11.6
Upper face height	69	6.3	71	5.6	66	6.2	67	4.5
Nasal height	54	4.7	52	4.8	52	3.8	50	4.4
Nasal breadth	35	3.7	43	3.7	32	4.9	40	4.1

*From Todd and Lindala (1928). Compiled by Krogman and Iscan (1986).

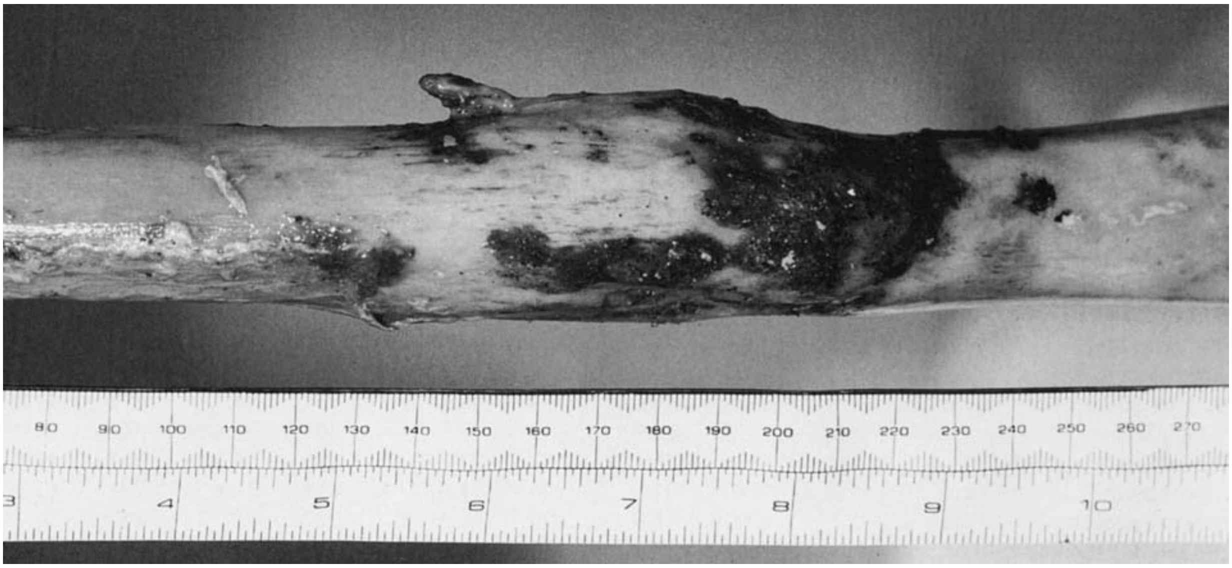

FIGURE 3.19 *Identification of a homicide victim, buried for almost 2 years in a garden, was achieved by matching this callus from a previous femoral fracture, with the medical notes and radiographs obtained from a hospital. The shape and size of the callus, together with the osseous spur due to muscle traction, was identical to follow-up films taken during life.*

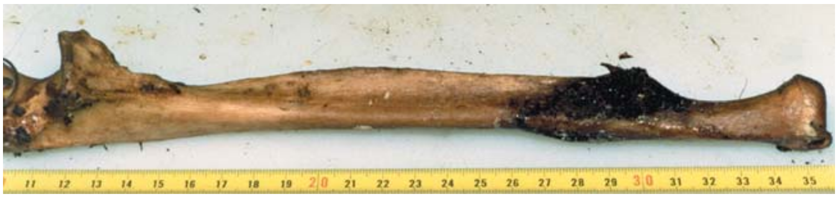

FIGURE 3.20 *Identification of a skeletalized body was achieved by matching this callus from a previous ulnar fracture, with the medical notes and radiographs obtained from a hospital. The shape and size of the callus, was identical to follow-up films taken during life.*

other cases there may be a list of hundreds of potential cases to search through in the hope of finding a match.

Individual characteristics may be of two main types.

- Anatomical shapes that can be matched by radiology, measurement or other means. Examples would be frontal sinus comparison, craniometry and radiological bone architecture.
- Discrete abnormalities such as healing fractures, metal prostheses, bone disease or congenital defects. Some artefacts, such as drill holes or wire, may immediately indicate that the bone is an anatomical specimen.

Anatomical matching depends upon accurate and detailed information about the living subject who is thought to correspond with the skeleton. Such information is usually radiographic and is derived from clinical records. The examples par excellence are dental charts and radiographs obtained from the dentist of the potential match; this is discussed in Chapter 26. Where skull films are available from ante-mortem sources, identity can be almost assured. The lateral view of a head can be matched with similar views taken of the skull, and both obvious anatomical landmarks compared and cranial measurements taken. The matter is further discussed later in relation to radiology, but the profile of the pituitary fossa as well as various intracranial diameters can rapidly exclude a match. In fact, a cursory glance can often confirm that the two skulls are dissimilar; making a positive correlation takes more time and requires accurate orientation of the radiographs.

Frontal sinus identification

This has been extensively studied since Schuller suggested the technique in 1921, as it is particularly useful in mutilated or burnt bodies, such as those from mass disasters like air crashes. The sinuses are well protected from all but the most extreme damage and are unique in that – as first pointed out

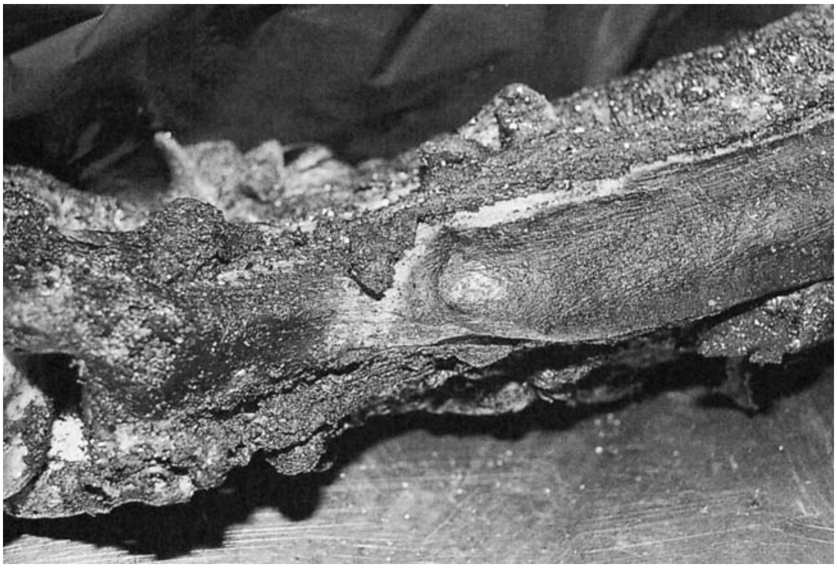

FIGURE 3.21 *Personal identity of this almost skeletalized fire victim was established from this healed callus on the tibia, the site of a well-documented fracture several years before death.*

by Poole in 1931 – no two persons (not even identical twins) have the same profile of these air spaces. They appear in the second year of life and increase in size for the first two decades. They are absent in about 5 per cent of persons and unilateral in another 1 per cent. For the sinuses to be used, an ante-mortem anteroposterior skull radiograph must be available, the most common source being from a previous hospital admission or examination, usually for a head injury. The cadaver skull or head must be X-rayed in exactly the same orientation and degree of enlargement, so that some superimposition technique can be employed. The 'forehead–nose' position was recommended by Schuller (1943), with the tube axis positioned level with the supraorbital margins. The scalloped upper margins of the sinuses are used for comparison, these being smaller and more numerous in the female. Asherson's thesis of 1965 recommends using the Caldwell occipitomental plane for the radiographs, used clinically for investigation of the nasal sinuses. This is preferable to the Wallers occipitofrontal view and the two projections are not necessarily identical in terms of sinus profile. Asherson recommends outlining the sinus shape in black ink on the film or tracing it onto a sheet of paper. Turpin and Tisserand (1942) projected their films on to a cardboard screen and cut out a template, weighing these from both ante- and post-mortem films to determine whether they were identical. Computed tomography of the sinuses is recommended by Reichs (1993).

Other radiological methods of comparing identity include matching of hand and wrist films, matching the profile and structure of the first rib and clavicle, and the craniometric methods of Sassouni (1959) and of Voluter (1959). The methods must be sought in either the original papers or surveys such as the textbook by Evans and Knight (1983). However, in general, it may be said that, wherever ante-mortem X-rays are available, especially of skull, thorax, hip or shoulder, then radiological comparisons of the dead material can almost always exclude an identity and, in many circumstances, can confirm it.

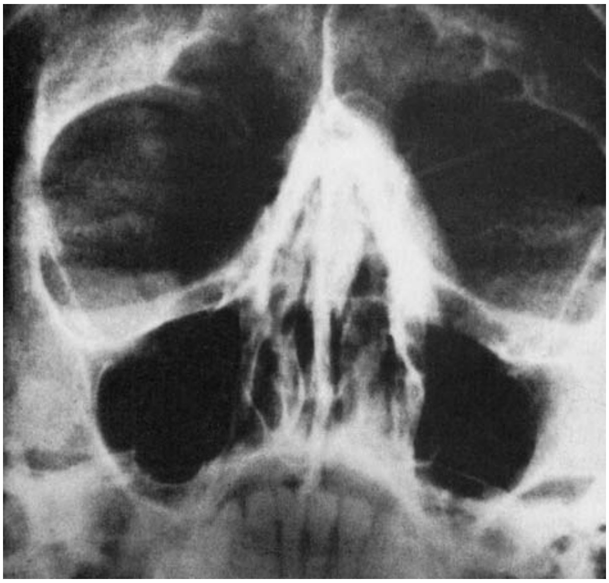

FIGURE 3.22 *The facial and cranial skeleton is unique to every person. If ante-mortem radiographs can be obtained for the person likely to be the victim, matching of the frontal sinus pattern can provide absolute identification. This body, which was washed up on the beach, was eventually matched with hospital radiographs through the sinus pattern seen at the top edge of the picture.*

THE DATING OF SKELETAL REMAINS

One of the issues in identity, as well as in estimating the time since death, is the date (as opposed to the age) of human bones. A knowledge of the time of death, measured in years, decades or even centuries, can assist the investigating authorities considerably. In fact, where skeletal remains are shown to be of considerable antiquity, no investigation at all is required, as even if the death was criminal, the culprit may long since have died. This threshold of investigative interest lies at around 70 years and material older than this may be of concern only to historians and archaeologists.

It is the marginal cases that present difficulties, especially as the methods of dating in this period are so approximate. For example, the skeleton shown in Figure 3.6 was discovered 41 years after the woman had been murdered. When the police went searching for the suspected killer, they found that he had died peacefully 3 years before the discovery, of coronary disease – thus qualifying for the 'perfect murder'.

The major problem in dating bones is that the environment is far more potent than time in changing the state of the bone. In Wales, a historic country where a spectrum of skeletal material has been found dating back some 18 000 years, frequent opportunities present for examining bones from many periods and many conditions of environment. Some Roman or even Bronze Age material, which had been concealed in dry sand dunes or in well-drained ancient burial mounds, appeared almost as pristine as an anatomical demonstration skeleton. On the other hand, modern material buried in coffins in graves marked with headstones only 20 years old have completely vanished because acidic, peaty water repeatedly eluted the bone as the water table in marshy ground rose and fell. An even more striking demonstration of the effect of local conditions was the observation that some limb bones buried vertically in a rock fall in a cave were in excellent condition at the upper ends, yet the lower parts were totally eroded where they had been surrounded by damp rubble.

These facts emphasize that vast errors can be made in dating bone merely by the gross morphological appearances and – as is so often the case – the less experienced the examiner, the more dogmatic his incorrect opinions are likely to be. The following schedule can therefore be nothing more than a rough guide, which must be modified where possible, by a knowledge of the environment in which the bones lay. Certain physicochemical tests are of assistance, but even these are themselves dependent on the environment, with the exception of radiochemical tests that are of little value in the period of forensic concern.

Physical appearance

Recent bones will have soft tissue remnants adhering, in the form of tendon and ligament tags, especially around the joint ends. Periosteum may be visible as fibrous material closely adherent to the shaft surface. Cartilage may also be present on articular surfaces. The time for which these remnants persist is variable, according to the conditions in which the bone has been left. Animal predators may rapidly remove all soft tissue and cartilage, sometimes within days or weeks. If the body has been left in a protected place, such as a vault or closed building, then dried tissue can remain for many years. In temperate climates, bodies left out in the open usually become substantially skeletalized within the first year, though tendon tags and periosteum may survive for 5 or more years. If a body dies in the autumn, then the preservation is likely to be longer through the colder winter months than if it died in the spring or summer.

After all soft tissue has vanished, recent bones may still be differentiated from old material by the density and feel of the bone. For a variable period, depending on storage conditions, a bone will feel slightly greasy to the fingers for several years, sometimes up to a decade if kept indoors. It will also feel heavy compared with older skeletal material because of the preservation of the organic stroma. On sawing a recent bone, it will be hard (especially the shafts of limb bones such as the femur) and will be uniform through the whole thickness. A smell of burning organic tissue will be obvious if vigorous sawing produces heat. In an old bone, the loss of the collagenous stroma will lighten the bone and make it easier to cut. The outer cortex and – to a lesser extent – the zone around the marrow cavity, will lose the stroma first, so a 'sandwich' effect may be seen in which a central ring of hard collagenous bone is layered on each side by a zone of more porous, crumbling material. This is not seen in less than several decades – and sometimes centuries – unless the bone has been exposed to the sun and other elements. The fragile, brittle appearance of old bones is usually first apparent on the ends of long bones, adjacent to the joints, such as the tibial plateau or the greater trochanter of the femur. This is often because the outer layer of compact bone is thinner there than in the shaft, so that the soft cancellous bone of the extremities is more readily exposed. This occurs within a few decades if the bone has been outside, but may not develop for a century in protected material. The aged cortex will feel rough and porous and, in really old material, can be crumbled or pitted with a fingernail.

Another factor that markedly affects the rate of decay of bones is the size and type of the bone itself. Whilst thick, dense bone, such as femoral or humeral shaft, may last for centuries, smaller and thinner items may disintegrate rapidly. Skull plates, tarsal and carpal bones, digits and the thin bones of the facial skeleton will rot more quickly, as will the small bones of fetuses and infants.

Physical tests

In parallel with the physical appearance of the sawn bone, fluorescence in ultraviolet light can be a useful preliminary test. If a shaft is cut across and inspected in the dark under ultraviolet light, such as that from a Wood's lamp, recent bone will shine with a silvery-blue tint right across the whole section. As the bone ages, the outer rim will cease to fluoresce and this will progressively deepen towards the centre. As with visual and tactile examination, a similar zone will work its way outwards from the marrow cavity until only a thin sandwich-filling will survive. This then fragments and eventually vanishes so that all the cut surface is non-fluorescent. The time that this process takes is variable, but total loss of ultraviolet fluorescence takes somewhere of the order of at least 100–150 years to complete. Other physical tests have been described, including density and specific gravity measurements, ultrasonic conduction and thermal behaviour when heated under special conditions. All these criteria depend on the loss of organic stroma and the development of a calcified matrix with a porous structure.

Serological and chemical tests

Positive tests for the presence of haemoglobin may be obtained for a variable time on either the bone surface or powdered bone, this being partly dependent on the sensitivity of the technique. Using the dye-peroxide methods (benzidine is most sensitive, but is now proscribed by Health and Safety regulations, so that the less satisfactory Kastel–Meyer test must be used), positive results may be obtained up to about 100 years.

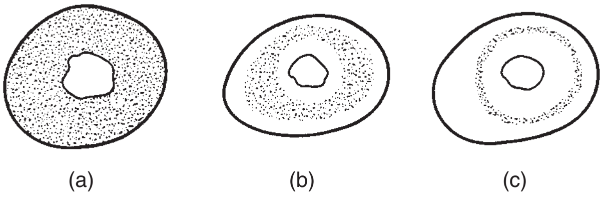

FIGURE 3.23 *Fluorescence of bone in ultraviolet light as a function of time since death. A freshly sawn cross-section of a dense bone such as femoral shaft is to be preferred. Up to several decades after death the whole cut surface fluoresces a bluish colour (a). From 3 to 80 years, greatly depending upon the environmental conditions, the outer zone and the zone around the marrow cavity progressively lose fluorescence (b). After a century or more, the residual fluorescence contracts to a narrowing central sandwich (c) before vanishing completely in the second century. Other methods of attempted dating include microscopic examination. Yoshino* et al. *(1991) claim that microradiography and electron microscopy reveal vacuoles in compact bone left in soil for 5 or 6 years.*

Serological activity lasts only a short time in bones exposed to the weather. Bone powder eluted with weak ammonia and vacuum concentrated, may give a positive reaction with antihuman sera, such as Coombs reagent, for only about 5–10 years – again depending greatly on the environmental conditions.

The recent development of DNA techniques has led to both bone and bone marrow being used to produce profiles for identification. There seems to be a wide range of times during which this is successful. The author (BK) dealt with a case where DNA testing was successful after 8 years' burial, but other more recent skeletons did not yield such success. However, far older archaeological material has given DNA profiles, so, at present, the DNA criteria cannot be used for dating.

Chemical testing seeks to measure the degradation of the proteinous stroma, so knowing the total nitrogen and amino acid content can be useful. Fresh compact bone contains about 4.5 per cent nitrogen, which progressively diminishes with time. If a bone contains more than about 4 per cent nitrogen, it is unlikely to be older than 100 years, but if it has 2.5 per cent or less, it is likely to be older than 350 years.

Estimation of the racemization of amino acids in teeth (laevo/dextro proportions) is being used to determine age, and there is some evidence that it also changes linearly with post-mortem interval. Its use on skeletal structures other than teeth is being investigated.

Residual protein can be converted to its constituent amino acids by prolonged heating in 6 M hydrochloric acid. The digest can then be analysed, either by autoanalytical methods or by two-dimensional chromatography. Fresh bone contains about 15 amino acids, mostly derived from collagen. Glycine and alanine are predominant, but proline and hydroxyproline are more specific markers for collagen. The latter two tend to vanish in about 50 years (as always, depending on the storage environment), so they are useful markers for the time threshold of forensic relevance. The other amino acids diminish over the succeeding scores and hundreds of years, so that a bone containing only four or five amino acids is likely to be relatively ancient. Glycine persists for millennia, being found in palaeolithic material.

Paradoxically, the methods available for archaeologists to use on ancient skeletal remains are proportionately more accurate, but of little or no use in forensic science. The radiocarbon test, which measures the loss of carbon-14 after death, cannot be applied to material less than a couple of centuries old, as the loss of radioactive carbon, which has a half-life of

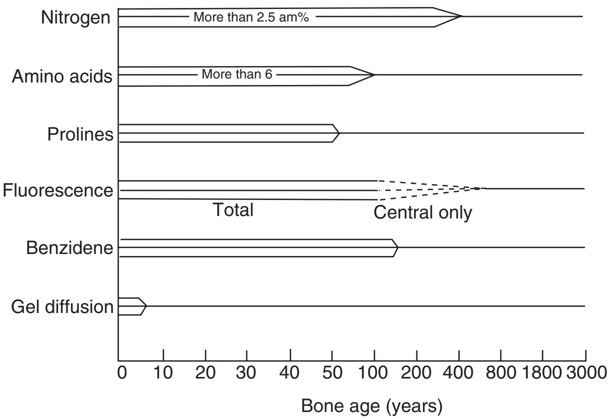

FIGURE 3.24 *Graphical summary of the criteria for estimating the approximate date of human skeletal remains.*

well over 5000 years, is too small to be measured in the half-century or so that interests the forensic pathologist.

Another radioisotope technique is the subject of current research at Aberdeen University, where Maclaughlin-Black *et al.* (1992) are measuring radiostrontium in skeletal remains, the principle being that pre-1945 bones should be free of endogenous strontium-90, derived from nuclear weapons and atmospheric tests, which was at a maximum in the environment in the early 1960s. The possibility exists that atmospheric contamination with other isotopes laid down in bone from dated tests and nuclear accidents may provide a profile that could bring forward the latest date during which the person must have been alive.

PHOTOSUPERIMPOSITION TECHNIQUES FOR IDENTIFICATION

Where potential candidate(s) for the identity of a skull are known to the investigating authority and photographs taken during life are available, a well established technique used for identification is 'photosuperimposition'. In this method, photographs of the skull are taken in exactly the same orientation in three planes as the available photograph. These are then enlarged to exactly the same dimensions as the photograph, and either the negative or a positive print is made on transparent film. This is then laid over the photograph and adjusted in an attempt to match up the major anatomical landmarks such as nasion, supraorbital ridges, angle of the jaw, nasal aperture, external auditory meatus and, especially, teeth. The actual anatomical markers will depend on what is visible, in frontal, lateral or oblique views. The test is mainly an exclusory one, in that, if the match cannot be made, then the skull is not that of the person in the photograph. If the match is good, even excellent, then the two could be the same person, though the standard of proof is low unless some individualistic feature, such as a distinct dental character, is apparent.

The photofit method has been used for many years, one of the most familiar examples being that reproduced in many forensic textbooks, of the skull and face of the wife of Dr Buck Ruxton, a notorious murder in 1935 investigated by Glaister and Brash in Scotland.

A more modern variant of the photosuperimposition technique is with the use of video cameras where two images, one of the photograph and the other of the skull, are mixed on one video display unit. By altering the camera angles and the degree of magnification of the images, superimposition can be tested quickly without the need for laborious photographic processing. This method was used in 1994 to identify a number of the 12 victims of the notorious 'House of Horror' in Gloucester. This is a rapidly developing field of forensic interest, where even the relatively poor-quality images from security cameras can be matched with suspects.

RECONSTRUCTION OF THE FACIAL APPEARANCE FROM THE SKULL

The reconstitution of the visage from the skeletal base has been a goal for workers in many different fields in recent years. The advantages of such a technique are obvious as, when a skull is discovered one prime method of identification would be a reliable reconstruction of the face, so that direct recognition could be obtained from relatives, friends and photographic records.

The first methods were as much artistic as scientific and depended to a great extent on the sculpturing ability of the operator. Gerasimov (1971) was a Soviet pioneer of this method, though much of his work was archaeological and historical, rather than forensic.

The method depends on a pre-knowledge of the usual tissue thickness at a multitude of points on the normal skull, an anatomical exercise that now has quite a large database. Modelling clay is laid on to the unknown skull in layers corresponding to these standard thicknesses, then more imaginative modelling added to 'humanize' the basic shape. The obvious defects in this technique are the lack of knowledge about eyes, lips, nose, ears and head hair, all of which contribute greatly to individual characteristics. Similar methods have been used by graphic artists, rather than sculptors, who use their portraiture talents to create a face on the two-dimensional base provided by the skull profile, plus a knowledge of tissue thicknesses at many anatomical points. Additional information has been provided on tissue thicknesses by radiography of heads.

The method was used with success in the 1988 investigation of the murder of Karen Price in Cardiff. Skeletalized after being buried for 8 years in a carpet beneath a garden, medical artist Richard Neave rebuilt her face upon a skull with sufficient accuracy for its display on public television to be recognized by her parents.

Recently, considerable progress has been made by the use of computer graphic techniques, both in drawing reconstituted heads and in gathering tissue thickness data (Vanezis *et al.* 1989). Some devices are mechanical, measuring the profile of the skull with a device that converts angles and distances into digital data. More recently, a combination of video and laser equipment has allowed 20 000 measurements to be taken and stored within 30 seconds. The data from an unknown skull are then electronically 'clothed' with standard soft tissues from the memory bank and modified on screen to produce various images. These can be rotated electronically so that various profiles can be seen. A variety of stored eyes, ears and noses can be added, and any feature altered almost instantaneously to give a viewer a number of opportunities to recognize the missing person. As with so many techniques in forensic medicine and science, the technology is one for super-specialists at present, but the forensic pathologist should be aware that such methods exist and are increasing in availability and accuracy.

REFERENCES AND FURTHER READING

Acsadi G, Nemeskeri J. 1970. *History of human life span and mortality.* Akademiai Kiado, Budapest.

Allbrook D. 1961. The estimation of stature in British and East African males; based on tibial and ulnar bone lengths. *J Forensic Med* **8**:15–28.

Angel JL. 1919. The bases of paleodemography. *Anthropology* **30**:425–38.

Angel JL, Suchey JM, Iscan MY, *et al.* 1986. Age at death from the skeleton and viscera. In: Zimmerman MR, Angel JL (eds), *Dating and age determination in biological materials.* Croom Helm, London, pp. 179–220.

Asala SA. 2001. Sex determination from the head of the femur of South African whites and blacks. *Forensic Sci Int* **117**:15–22.

Asala SA. 2002. The efficiency of the demarking point of the femoral head as a sex determining parameter. *Forensic Sci Int* **127**:114–18.

Asherson N. 1965. *Identification by frontal sinus prints: a forensic medicine pilot survey.* HK Lewis, London.

Baccino E, Ubelaker DH, Hayek LA, *et al.* 1999. Evaluation of seven methods of estimating age at death from mature human skeletal remains. *J Forensic Sci* **44**:931–6.

Barbet JP, Houette A, Barres D, *et al.* 1988. Histological assessment of gestational age in human embryos and fetuses. *Am J Forensic Med Pathol* **9**:40–4.

Berg S. 1963. *The determination of bone age.* In: Lundquist F (ed.), *Methods in forensic science*. Interscience Publications, New York, pp. 231–52.

Black TK. 1978. A new method for assessing the sex of fragmentary skeletal remains: femoral shaft circumference. *Am J Phys Anthropol* **48**:227–31.

Boström K. 1973. Identification from roentgenograms of sinus frontalis and sella turcica. *Proceedings of Fifth Meeting Scandinavian Society of Forensic Medicine.*

Boucher BJ. 1957. Sex differences in the fetal pelvis. *Am J Phys Anthropol* **15**:581–600.

Boyd JD, Trevor J. 1953. Problems in reconstruction. In: Simpson CK (ed.), *Modern trends in forensic medicine*. Butterworth, London.

Breitinger E. 1937. Zur Berechnung der Körperhöhe aus den langen Gliedmassenknochen. *Anthropol Anz* **14**:249–74.

Brothwell D. 1972. *Digging up bones*, 2nd edn. British Museum, London.

Brown TC. 1950. Medical identification in the Noronic disaster. *Fingerprint Ident Mag* **6**:3–12.

Burnham JT, Preston Burnham J, Fontan CR. 1976. The state of the art of bone identification by chemical and microscopic methods. *J Forensic Sci* **21**:340–2.

Caldwell WE, Molloy HC. 1933. Anatomical variations in the female pelvis. *Am J Obstet Gynecol* **26**:479–505.

Castellano MA, Villanueva EC, von Frenckel R. 1984. Estimating the date of bone remains: a multivariate study. *J Forensic Sci* **29**:527–34.

Cattaneo C, Smillie DM, Gelsthorpe K, *et al.* 1995. A simple method for extracting DNA from old skeletal material. *Forensic Sci Int* **74**:167–74.

Cattaneo C, Craig OE, James NT, *et al.* 1997. Comparison of three DNA extraction methods on bone and blood stains up to 43 years old and amplification of three different gene sequences. *J Forensic Sci* **42**:1126–35.

Cattaneo C, DiMartino S, Scali S, *et al.* 1999. Determining the human origin of fragments of burnt bone: a comparative study of histological, immunological and DNA techniques. *Forensic Sci Int* **102**:181–91.

Cattaneo C, Ritz-Timme S, Schutz HW, *et al.* 2000. Unidentified cadavers and human remains in the EU: an unknown issue. *Int J Legal Med* **113**:N2–3.

Cheevers LS, Ascensio R. 1977. Identification by skull superimposition. *Int J Forensic Dent* **4**:14–16.

Christie AL. 1949. Prevalence and distribution of ossification centres in the newborn infant. *Am J Dis Child* **77**:355–61.

Cobb WM. 1952. Skeleton. In: Lansings AL (ed.), *Cowdry's problems of ageing.* Williams & Wilkins, Baltimore.

Colonna M, Pesce Delfino V, Introna F, Jr. 1980. [Identification by superimposition of the skull on face photography using television: experimental application of a new method.] *Boll Soc Ital Biol Sper* **56**:2271–7.

Colonna M, Introna F, Jr, Potente F, *et al.* 1984. Computer-aided skull-face superimposition by analytical procedures. *Acta Med Leg Soc Liege* **34**:139–49.

Culbert W, Law F. 1927. Identification by comparison of accessory nasal sinuses and mastoid processes. *J Am Dent Assoc* **88**:1634.

Davidson WM. 1960. Sex determination: diagnostic methods. *Br Med J* **4**:1901–7.

Derry DE. 1909. Note on the innominate bone in the determination of sex. *J Anthropol* **43**:266–76.

Derry DE. 1924. On the sexual and racial characters of the ilium. *J Anat* **4**:71–83.

Dixon AD. 1956. Postmortem persistence of sex chromatin. *J Forensic Med* **3**:161–8.

Dixon AD. 1957. Sex determination of human tissues from cell morphology. *J Forensic Med* **4**:11–17.

Dodd BE. 1985. DNA fingerprinting in matters of family and crime [news]. *Nature* **318**:506–7.

Dupertuis CW, Hadden JA. 1951. On the reconstruction of stature from long bones. *Am J Phys Anthropol* **9**:15–54.

Dutra F. 1944. Identification of person and determination of cause of death from skeletal remains. *Arch Pathol* **38**:339.

Dwight T. 1889. The sternum as an index of sex, height and age. *J Anat Physiol* **24**:527–35.

Dwight T. 1890. The closure of the sutures as a sign of age. *Boston Med Surg J* **122**:389–92.

Dwight T. 1904. The size of the articular surfaces of the long bones as characteristic of sex. *J Anat* **4**:19–32.

Earls J, Hester R. 1967. Tattoed sailors: Sociopscychological correlates. *Milit Med* **132**:48.

Eckert W. 1985. The forensic investigation of Josef Mengele [editorial]. *Am J Forensic Med Pathol* **6**:187.

Eckert WG, Teixeira WR. 1985. The identification of Josef Mengele. A triumph of international cooperation. *Am J Forensic Med Pathol* **6**:188–91.

Eckert WG, Willey LR, Blakeslee DJ. 1986. Identification by identi-kit composite from skull appearance. Combined efforts of the anthropologist, forensic pathologist, and criminal investigators. *Am J Forensic Med Pathol* **7**:213–15.

Eliakis C. 1967. Estimating height from the measurement of long bones. *Ann Med Leg* **212**:251–4.

Eliakis EC, Iordanidis PJ. 1963. Determination of sex from the medullary indices of long bones. *Am Med Leg* **43**:326–9.

Eriksson L, Westermark P. 1990. Amyloid inclusions in choroid plexus epithelial cells. A simple autopsy method to rapidly obtain information on the age of an unknown dead person. *Forensic Sci Int* **48**:97–102.

Evans KT, Knight B. 1983. *Forensic radiology*. Blackwell Scientific Publications, Oxford.

Ezra M, Cook SF. 1957. Amino-acids in fossil bone. *Science* **126**:8–86.

Facchini F, Pettener D. 1977. Chemical and physical methods in dating human skeletal remains. *Am J Phys Anthropol* **47**:65–70.

Fawcett E. 1938. Sexing of the human sternum. *J Anat* **72**:633–7.

Fazekas G, Kosa F. 1978. *Forensic foetal osteology*. Akademiai Kiado, Budapest.

Flecker H. 1942. Time of appearance and fusion of ossification centers as observed by roentgenographic methods. *Am J Roentgenol* **47**:97–105.

Francis GC, Werle PP. 1939. The appearance of centers of ossification from birth to 5 years. *Am J Phys Anthropol* **24**:273–86.

Fukushima H, Hasekura H, Nagai K. 1988. Identification of male bloodstains by dot hybridization of human Y chromosome-specific deoxyribonucleic acid (DNA) probe. *J Forensic Sci* **33**:621–7.

Fully G, Pineau H. 1960. Detemination de la stature an moyen du squelette. *Ann Med Leg* **40**:3–11.

Garn SM. 1958. The sex difference in tooth calcification. *J Dent Res* **37**:561–72.

Gatliff BP. 1984. Facial sculpture on the skull for identification. *Am J Forensic Med Pathol* **5**:327–32.

Gatliff BP, Snow CC. 1979. From skull to visage. *J Biocommun* **6**:27–30.

Genovés S. 1954. The problem of sex of fossil hominids. *J R Anthropol Inst* **84**:131–6.

Genovés ST, Messmacher M. 1959. *Valor de los patrones tradicionales para la determinacíon del edad por me las suturas craneos mexicanos. Serie anthropol. No. 7.* Cuadernos de Instit de Hist., Mexico City, Mexico, Vol. 7, pp. 7–53.

George RM. 1987. The lateral craniographic method of facial reconstruction. *J Forensic Sci* **32**:1305–30.

Gerasimov MM. 1971. *The face finder*. Lippincott, Philadelphia.

Gilbert BM, McKern TW. 1973. A method for aging the female os pubis. *Am J Phys Anthropol* **38**:31–8.

Giles E, Elliot O. 1963. Sex determination by discriminant function analysis of crania. *Am J Phys Anthropol* **21**:53–68.

Gill GW. 2001. Racial variation in the proximal and distal femur: heritability and forensic utility. *J Forensic Sci* **46**:791–9.

Glaister J, Brash J. 1937. *The medicolegal aspects of the Ruxton case.* Livingstone, Edinburgh.

Godycki M. 1957. Sur la certitude de determination se sexe après le femur, le cubitus et l'humerus. *Bull Mem Soc Anthrop Paris* Serie **10**:405–10.

Graves WW. 1922. Observations of age changes in the scapula: a preliminary note. *Am J Phys Anthropol* **5**:21–3.

Greulich WW. 1960. Value of X-ray films of hand in human identification. *Science* **1341**:155–62.

Greulich WW, Pye SI. 1959. *Radiographic atlas of the skeletal development of the hand and wrist*, 2nd edn. Stanford University Press, Stanford.

Greulich WW, Thomas H. 1938. The dimensions of the pelvic inlet of 789 white females. *Anat Rec* **72**:45–51.

Greulich WW, Thomas H. 1939. A X-ray study of the male pelvis. *Anat Rec* **75**:289–99.

Hanihara K. 1967. Racial characteristics in the dentition. *J Dent Res* **46**:923–6.

Harrison R. 1968. In: Camps F (ed.), *Gradwohl's legal medicine*, 2nd edn. John Wright, Bristol.

Hellman H. 1929. Racial characteristics in the human dentition. *Am Phil Soc* **67**:157–74.

Herrman L. 1953. Fingerprints and twins. *J Forensic Med* **1**:101–5.

Hill A. 1939. Fetal age assessment by centers of ossification. *Am J Phys Anthropol* **24**:251.

Hoshi, H. 1962. Sex differentiation in the shape of the mastoid process in norma occipalis and its importance to sex determination of the human skull. *Okajima's Folia Anat Jpn* **38**:309–17.

Howard JD, Reay DT, Haglund WD, *et al.* 1988. Processing of skeletal remains. A medical examiner's perspective. *Am J Forensic Med Pathol* **9**:258–64.

Hrdlicka A. 1920. Shovel-shaped teeth. *Am J Phys Anthropol* **3**:429.

Hrdlicka A. 1939. *Practical anthropometry*. Wistar Institute Press, Philadelphia.

Hunt EE, Gleiser I. 1955. The estimation of the age and sex of pre-adolescent children from bones and teeth. *Am J Phys Anthropol* **13**:479–87.

Iordanidis P. 1961. Determination du sexe par les os du squelette (atalas, axis, clavicle, omoplate, sternum). *Ann Med Leg* **41**:28–91.

Iscan MY, Derrick K. 1984. Determination of sex from the sacroiliac; a visual assessment technique. *Fla Sci* **47**:948.

Iscan MY, Loth SR. 1986. Determination of age from the sternal rib in white males: a test of the phase method. *J Forensic Sci* **31**:122–32.

Iscan MY, Miller-Shaivitz P. 1984. Discriminant function sexing of the tibia. *J Forensic Sci* **29**:1087–93.

Iscan MY, Miller-Shaivitz P. 1984. Determination of sex from the tibia. *Am J Phys Anthropol* **64**:53–7.

Iscan MY, Loth SR, King CA, *et al.* 1998. Sexual dimorphism in the humerus: a comparative analysis of Chinese, Japanese and Thais. *Forensic Sci Int* **98**: 17–29.

Iscan MY, Steyn M. 1999. Craniometric determination of population affinity in South Africans. *Int J Legal Med* **112**:91–7.

Iten PX. 1987. Identification of skulls by video superimposition. *J Forensic Sci* **32**:173–88.

Jantz RL. 2001. Cranial change in americans: 1850–1975. *J Forensic Sci* **46**:784–7.

Jedrzejowska ZK. 2001. Craniometry and mathematical calculations as a method for viscero-cranium profile determination. *Forensic Sci Int* **117**:145–51.

Jit I, Singh S. 1956. Estimation of stature from the clavicles. *Indian J Med Res* **44**:137–55.

Jit I, Jhingan V, Kulkarni M. 1980. Sexing the human sternum. *Am J Phys Anthropol* **53**:217–24.

Keen J. 1950. A study of the differences between male and female skulls. *Am J Phys Anthropol* **8**:65.

Keen JA. 1950. Sex differences in skulls. *Am J Phys Anthropol* **8**:479–87.

Kerley ER. 1965. The microscopic determination of age in human bone. *Am J Phys Anthropol* **23**:149–63.

Kerley ER. 1969. Age determination of bone fragments. *J Forensic Sci* **14**:59–67.

Kim MA, Bier L. 1972. Identification and location of the y body in interphase by quinacrine and giemsa. *Humangenetik* **16**:261–5.

Kimura K. 1982. Base wing index for sexing the sacrum. *J Anthropol Soc Nippon* **90**(Suppl):153–62.

Knight B. 1969. Methods of dating skeletal remains. *Med Sci Law* **9**:247–52.

Knight B. 1985. The examination of skeletal remains. In: Wecht C (ed.), *Legal Medicine Annual*, 1985 edn. Praeger Scientific, New York.

Knight B, Lauder I. 1967. Practical methods of dating skeletal remains: a preliminary study. *Med Sci Law* **7**:205–8.

Knight B, Lauder I. 1969. Methods of dating skeletal remains. *Hum Biol* **41**:322–41.

Kobayashi K. 1964. Estimation of age at death from pubic symphysis for prehistoric human remains in Japan. *Zinruig Zassi* **72**:1–16.

Kobayashi R, Nakauchi H, Nakahori Y, *et al.* 1988. Sex identification in fresh blood and dried bloodstains by a nonisotopic deoxyribonucleic acid (DNA) analyzing technique. *J Forensic Sci* **33**:613–20.

Koelmeyer TD. 1982. Videocamera superimposition and facial reconstruction as an aid to identification. *Am J Forensic Med Pathol* **3**:45–8.

Konigsberg LW, Hens SM, Jantz LM, *et al.* 1998. Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology. *Am J Phys Anthropol Suppl* **27**:65–92.

Krogman WCA. 1939. A guide to the identification of human skeletal material. *FBI Law Enforcement Bull* **8**:3.

Krogman WCA. 1946. The skeleton in forensic medicine. *Proc Inst Med* **16**:154.

Krogman WCA, Iscan MY. 1986. *The human skeleton in forensic medicine*, 2nd edn. Thomas, Springfield.

Loth SR, Iscan MY, Scheuerman EH. 1994. Intercostal variation at the sternal end of the rib. *Forensic Sci Int* **65**:135–43.

Lundy JK. 1985. The mathematical versus anatomical methods of stature estimate from long bones. *Am J Forensic Med Pathol* **6**:73–6.

Maclaughlin-Black SM, Herd RJ, Willson K, *et al.* 1992. Strontium-90 as an indicator of time since death: a pilot investigation. *Forensic Sci Int* **57**:51–6.

Mall G, Graw M, Gehring K, *et al.* 2000. Determination of sex from femora. *Forensic Sci Int* **113**:315–21.

Mall G, Hubig M, Buttner A, *et al.* 2001. Sex determination and estimation of stature from the long bones of the arm. *Forensic Sci Int* **117**:23–30.

Maltby J. 1917. Some indices and measurements of the modern femur. *J Anat* **52**:363–82.

Mann RW. 1998. Use of bone trabeculae to establish positive identification. *Forensic Sci Int* **98**:91–9.

Manouvrier L. 1893. Le determination de la taille après les grand os des membres. *Mem Soc Anthropol Paris* **4**:347–402.

Maresh MM. 1940. Paranasal sinuses from birth to adolescence. *Am J Dis Child* **60**:55–78.

Mayler J. 1935. Identification by sinus prints. *VA Med Mon* **62**:517–19.

McBride DG, Dietz MJ, Vennemeyer MT, *et al.* 2001. Bootstrap methods for sex determination from the os coxae using the ID3 algorithm. *J Forensic Sci* **46**:427–31.

McCormick WF. 1980. Mineralization of the costal cartilages as an indicator of age: preliminary observations. *J Forensic Sci* **25**:736–41.

McCormick WF. 1981. Sternal foramena in man. *Am J Forensic Med Pathol* **2**:249–52.

McCormick WF, Nichols MM. 1981. Formation and maturation of the human sternum. I. Fetal period. *Am J Forensic Med Pathol* **2**:323–8.

McKeehan HE. 1970. The restoration of desiccated cadaveric fingers for the purpose of identification. *J Forensic Sci Soc* **10**:115.

McKern TW, Stewart TD. 1957. *Skeletal age changes in young American males, analyzed from the standpoint of identification.* Headq. QM Res & Dev Command, Natick, MA.

Mehta L, Singh HM. 1972. Determination of crown-rump length from fetal long bones: humerus and femur. *Am J Phys Anthropol* **36**:165–8.

Meindl RS, Lovejoy CO, Mensforth RP, *et al.* 1985. A revised method of age determination using the os pubis, with a review and tests of accuracy of other current methods of pubic symphyseal aging. *Am J Phys Anthropol* **68**:29–45.

Mendes-Correa AA. 1932. La taille des portugais d'après des os longes. *Anthropology (Prague)* **10**:268–72.

Messmer JM, Fierro MF. 1986. Personal identification by radiographic comparison of vascular groove patterns of the calvarium. *Am J Forensic Med Pathol* **7**:159–62.

Michelson N. 1934. The calcification of the first costal cartilage in whites and negroes. *Hum Biol* **6**:543–57.

Miller KW, Budowle B. 2001. A compendium of human mitochondrial DNA control region: development of an international standard forensic database. *Croat Med J* **42**:315–27.

Modi JP. 1957. *Medical jurisprudence and toxicology*. Tripathi, Bombay.

Montanari GD, Viterbo B, Montanari GR. 1967. Sex determination of human hair. *Med Sci Law* **7**:208–10.

Moore KL, Barr ML. 1955. Smears from the oral mucosa in the detection of chromosomal sex. *Lancet* **2**:57–8.

Moss JP, Linney AD, Grindrod SR, *et al.* 1987. Three-dimensional visualization of the face and skull using computerized tomography and laser scanning techniques. *Eur J Orthod* **9**:247–53.

Müller G. 1935. Zur Bestimmung der Länge beschädigter Extremitätenknochen. *Anthropol Anz* **12**:70–2.

Munter AH. 1936. A study of the lengths of long bones of the arms and legs in man, with special reference to Anglo-Saxon skeletons. *Biometrika* **28**:84–122.

Myers JC, Okoye MI, Kiple D, *et al.* 1999. Three-dimensional (3-D) imaging in post-mortem examinations: elucidation and identification of cranial and facial fractures in victims of homicide utilizing 3-D computerized imaging reconstruction techniques. *Int J Legal Med* **113**:33–7.

Nagamori H, Ohno Y, Uchima E, *et al.* 1986. Sex determination from buccal mucosa and hair root by the combined treatment of quinacrine staining and the fluorescent feulgen reaction using a single specimen. *Forensic Sci Int* **31**:119–28.

Navani S, Shah JR, Levy PS. 1970. Determination of sex by costal cartilage calcification. *Am J Roentgenol Radium Ther Nucl Med* **108**:771–4.

Nemeskeri J, Harsanyi L, Acsadi G. 1960. Methoden zur Diagnose des Lebensalters von Skelettfunden. *Anthropol Anz* **24**:70–95.

Niyogi S. 1971. A study of human hairs in forensic work. *J Forensic Sci* **16**:176.

Norris SP. 2002. Mandibular ramus height as an indicator of human infant age. *J Forensic Sci* **47**:8–11.

Oettle AC, Steyn M. 2000. Age estimation from sternal ends of ribs by phase analysis in South African blacks. *J Forensic Sci* **45**:1071–9.

Ohtani S. 1994. Age estimation by aspartic acid racemization in dentin of deciduous teeth. *Forensic Sci Int* **68**:77–82.

Ohtani S, Matsushima Y, Kobayashi Y, *et al.* 2002. Age estimation by measuring the racemization of aspartic acid from total amino acid content of several types of bone and rib cartilage: a preliminary account. *J Forensic Sci* **47**:32–6.

Olivier G. 1969. *Practical anthropology*. Thomas, Springfield.

Olivier G, Pineau H. 1958. Determination de l'age du foetus et de l'embyon. *Arch Anat Pathol* **6**:21–8.

Parsons FG, Box CR. 1905. The relation of the cranial sutures to age. *J R Anthropol Inst* **35**:308.

Pearson FG. 1920. Sex differences in the skull. *J Anthropol* **54**:58–65.

Pearson K. 1914. On the problem of sexing osteometric material. *Biomedicine* **10**:479–87.

Pearson K. 1926. On the coefficient of racial likeness. *Biometrika* **18**:105.

Pearson K, Bell JA. 1917. A study of the long bones of the English skeleton. 1. The femur. In: *Drapers Co Res Mem*. Biomedicine, Series X. University of London, London, Chapter 14.

Perper JA, Patterson GT, Backner JS. 1988. Face imaging reconstructive morphography. A new method for physiognomic reconstruction. *Am J Forensic Med Pathol* **9**:126–38.

Pesce Delfino V, Colonna M, Vacca E, *et al.* 1986. Computer-aided skull/face superimposition. *Am J Forensic Med Pathol* **7**:201–12.

Phenice TW. 1969. A newly developed visual method of sexing the os pubis. *Am J Phys Anthropol* **30**:297–301.

Pounder DJ. 1984. Forensic aspects of aboriginal skeletal remains in Australia. *Am J Forensic Med Pathol* **5**:41–52.

Rao NG, Pai LM. 1988. Costal cartilage calcification pattern – a clue for establishing sex identity. *Forensic Sci Int* **38**:193–202.

Reichs KJ. 1993. Quantified comparison of frontal sinus patterns by means of computed tomography. *Forensic Sci Int* **61**:141–68.

Reynolds E. 1945. The boney pelvic girdle in early infancy: a roentgenometric study. *Am J Phys Anthropol* **3**:321–54.

Reynolds E. 1947. The boney pelvis in prepubertal childhood. *Am J Phys Anthropol* **5**:165–200.

Rhine JS, Campbell HR. 1980. Thickness of facial tissues in American blacks. *J Forensic Sci* **25**:847–58.

Rhine JS, Moore CE. 1982. *Facial reproduction tables of facial tissue thicknesses of American caucasoids in forensic anthropology.* Maxwell Museum, Albuquerque.

Rho YM. 1985. Importance of examination of the clothed victim. Fingerprint identification of assailant from skin fragment on knifing victim's clothing. *Am J Forensic Med Pathol* **6**:19–20.

Richardson L, Kade H. 1972. Readable fingerprints from mummified or putrefied specimens. *J Forensic Sci* **17**:325–8.

Ritz S, Turzynski A, Schutz HW. 1994. Estimation of age at death based on aspartic acid racemization in non-collagenous bone proteins. *Forensic Sci Int* **69**:149–59.

Ross AH, Konigsberg LW. 2002. New formulae for estimating stature in the balkans. *J Forensic Sci* **47**:165–7.

Ross AH, Jantz RL, McCormick WF. 1998. Cranial thickness in American females and males. *J Forensic Sci* **43**:267–72.

Sanders CF. 1966. Sexing by costal cartilage calcification. *Br J Radiol* **39**:233–7.

Sassouni V. 1959. Cephalometric identification: a proposed method of identification of war dead by means of roentgenographic cephalometry. *J Forensic Sci* **4**:1–10.

Scammon RE. 1937. Two simple nomograms for determining the age and some of the major external dimensions of the human foetus. *Anat Rec* **68**:221–32.

Scheuer L. 2002. Application of osteology to forensic medicine. *Clin Anat* **15**:297–312.

Schranz D. 1959. Age determination from the internal structure of the humerus. *Am J Phys Anthropol* **17**:273–8.

Schuller A. 1921. Das Röntgengramm der Stirnhöhle. *Monatsschr Ohrenheilkd* **55**:1617–20.

Schuller A. 1943. Note on the identification of skulls by X-ray pictures of the frontal sinuses. *Med J Aust* **1**:554–6.

Schultz A. 1937. Proportions, variability and asymmetry of the long bones of the limbs and clavicles in man and apes. *Hum Biol* **9**:281–328.

Sen NK. 1962. Identification by superimposed photographs. *Int Crim Police Rev* **162**:284–6.

Sharif R, Chandra H. 1972. Histological differentiation as to the origin of bones. *J Indian Acad Forensic Sci* **2**:23–30.

Shiati M. 1983. Estimation of stature by long bones of Chinese male adults in South China. *Acta Anthropol Sinica* **2**:80–5.

Simpson E, Henneberg M. 2002. Variation in soft-tissue thicknesses on the human face and their relation to craniometric dimensions. *Am J Phys Anthropol* **118**:121–33.

Singer R. 1953. Estimation of age from cranial suture closure; a report on its unreliability. *J Forensic Med* **1**:52–9.

Snow CC, Williams J. 1971. Variation in premortem statural measurements compared to statural estimates of skeletal remains. *J Forensic Sci* **16**:455–64.

Snow CC, Gatliff BP, McWilliams KR. 1970. Reconstruction of facial features from the skull: an evaluation of its usefulness in forensic anthropology. *Am J Phys Anthropol* **33**:221–8.

Snow CC, Hartman S, Giles E, *et al.* 1979. Sex and race determination of crania by calipers and computer: a test of the Giles and Elliot discriminant functions in 52 forensic science cases. *J Forensic Sci* **24**:448–60.

Steele D. 1970. *Estimation of stature from fragments of long limb bones.* In: Stewart T (ed.), *Personal identification in mass disasters.* National Museum of Nature History, Washington DC, pp. 85–7.

Steele DG, McKern TW. 1969. A method for assessment of maximum long bone length and living stature from fragmentary long bones. *Am J Phys Anthropol* **31**:215–27.

Stephan CN. 2002. Facial approximation: globe projection guideline falsified by exophthalmometry literature. *J Forensic Sci* **47**:730–5.

Stevenson PH. 1929. On racial differences in stature long bone regression formulae with special reference to stature reconstruction formulae for the Chinese. *Biometrika* **21**:303–18.

Stevenson PM. 1924. Age order of epiphyseal union in man. *Am J Phys Anthropol* **7**:53–93.

Stewart JH, McCormick WF. 1983. The gender predictive value of sternal length. *Am J Forensic Med Pathol* **4**:217–20.

Stewart TD. 1948. Medicolegal aspects of the skeleton. *Am J Phys Anthropol* **6**:315–21.

Stewart TD. 1952. *Hrdlicka's practical anthropometry*, 4th edn. Wistar Institute, Philadelphia.

Stewart TD. 1957. Distortion of the pubic symphyseal surface in females and its effect on a determination. *Am J Phys Anthropol* **15**:9.

Stewart TD. 1959. Bear paw remains closely resemble human bones. *FBI Law Enforcement Bull* **28**:18.

Stewart TD. 1962. Anterior femoral curvature; its utility for race determination. *Hum Biol* **34**:4–57.

Stewart TD. 1968. Evaluation of evidence from the skeleton. In: Camps F (ed.), *Gradwohl's legal medicine*, 2nd edn. John Wright, Bristol, p. 425.

Stewart TD. 1968. Subject: skull sutures and age. In: Camps F (ed.), *Gradwohl's legal medicine*, 2nd edn. John Wright, Bristol, p. 132.

Stewart TD. 1979. *Essentials of forensic anthropology: especially as developed in the United States.* Thomas, Springfield.

Stewart TD, Trotter M. 1954. *Basic readings on the identification of human skeletons: estimations of age.* Wenner-Gren Foundation for Anthropological Research, New York.

Steyn M, Iscan MY. 1998. Sexual dimorphism in the crania and mandibles of South African whites. *Forensic Sci Int* **98**:9–16.

Steyn M, Iscan MY. 1999. Osteometric variation in the humerus: sexual dimorphism in South Africans. *Forensic Sci Int* **106**:77–85.

Straus WL. 1927. The human ilium: sex and stock. *Am J Phys Anthropol* **11**:1–28.

Suchey JM. 1979. Problems in the aging of females using the os pubis. *Am J Phys Anthropol* **51**:467–70.

Suzuki K, Tsuchihashi Y. 1970. New attempt of personal identification by means of lip print. *J Indian Dent Assoc* **42**:8–9.

Swift B, Lauder I, Black S, *et al.* 2001. An estimation of the post-mortem interval in human skeletal remains: a radionuclide and trace element approach. *Forensic Sci Int* **117**:73–87.

Teixeira WR. 1982. Sex identification utilizing the size of the foramen magnum. *Am J Forensic Med Pathol* **3**:203–6.

Teixeira WR. 1985. The Mengele report. *Am J Forensic Med Pathol* **6**:279–83.

Telkkä A. 1950. On the prediction of human stature from long bones. *Acta Anal* **9**:103–17.

Thomas H, Greulich WA. 1940. A comparative study of male and female pelves. *Am J Obstet Gynecol* **39**:56–62.

Thompson DD, Galvin CA. 1983. Estimation of age at death by tibial osteon remodeling in an autopsy series. *Forensic Sci Int* **22**:203–11.

Thomsen JL. 1977. Sex determination of severely burned bodies. *Forensic Sci* **10**:235–42.

Thomsen JL. 1979. Y-chromosome bodies in brain and kidney: the normal variation and an unexpected finding in sudden infant death syndrome. *Med Sci Law* **19**:104–7.

Tibbets GL. 1981. Estimation of stature from the vertebral column in American blacks. *J Forensic Sci* **26**:715–23.

Tobias PV. 1953. The problem of race determination. *J Forensic Med* **1**:113–23.

Todd TW. 1920. Age changes in the pubic bone. *Am J Phys Anthropol* **3**:285–334.

Todd TW. 1921. Age changes in the pubic bone I–IV. *Am J Phys Anthropol* **4**:407.

Todd TW, Lindala A. 1928. Dimensions of the body; whites and American negroes of both sexes. *Am J Phys Anthropol* **12**:35–119.

Todd TW, Lyon DW. 1924. Endocranial suture closure; its progress and age relationship. Part I. *Am J Phys Anthropol* **7**:325–84.

Todd TW, Lyon DW. 1925. Cranial suture closure; its progress and age relationship. Part II. *Am J Phys Anthropol* **8**:23–5.

Todd TW, Lyon DW. 1925. Cranial suture closure; its progress and age relationship. Parts III–IV. *Am J Phys Anthropol* **8**:47–71, 149–68.

Trotter M, Gleser GC. 1958. A re-evaluation of the estimation of stature based on measurements of stature taken during life and on long bones after death. *Am J Phys Anthropol* **16**:79–88.

Trotter M, Gleser GC. 1977. Corrigenda to 'estimation of stature from long limb bones of American Whites and Negroes', American Journal Physical Anthropology (1952). *Am J Phys Anthropol* **47**:355–6.

Turner W. 1886. The index of the pelvic brim as a basis of classification. *J Anat Physiol* **20**:125–43.

Turpin R, Tisserand M. 1942. Etude correlative de sinus. *C R Soc Biol* **136**:203–10.

Tyler MG, Kirby LT, Wood S, *et al.* 1986. Human blood stain identification and sex determination in dried blood stains using recombinant DNA techniques. *Forensic Sci Int* **31**:267–72.

Ubelaker DH. 2001. Artificial radiocarbon as an indicator of recent origin of organic remains in forensic cases. *J Forensic Sci* **46**:1285–7.

Ubelaker DH, Volk CG. 2002. A test of the phenice method for the estimation of sex. *J Forensic Sci* **47**:19–24.

Ullrich H. 1975. Estimation of fertility by means of pregnancy and childbirth alterations at the pubis, ilium and sacrum. *Ossa* **2**:23–39.

Vanezis P, Blowes RW, Linney AD, *et al.* 1989. Application of 3-D computer graphics for facial reconstruction and comparison with sculpting techniques. *Forensic Sci Int* **42**:69–84.

Voluter G. 1959. The 'V' test. *Radiol Clin Basel* **28**:1–32.

Wahl J, Graw M. 2001. Metric sex differentiation of the pars petrosa ossis temporalis. *Int J Legal Med* **114**:215–23.

Walker GF. 1976. The computer and the law: coordinate analysis of skull shape and possible methods of postmortem identification. *J Forensic Sci* **21**:357–66.

Warren MW, Schultz JJ. 2002. Post-cremation taphonomy and artifact preservation. *J Forensic Sci* **47**:656–9.

Washburn S. 1948. Sex differences in the pubic bone. *Am J Phys Anthropol* **6**:199–298.

Wells LH. 1959. Estimation of stature from long bones; a reassessment. *J Forensic Med* **6**:171–7.

Wescott DJ. 2000. Sex variation in the second cervical vertebra. *J Forensic Sci* **45**:462–6.

Wescott DJ, Moore-Jansen PH. 2001. Metric variation in the human occipital bone: forensic anthropological applications. *J Forensic Sci* **46**:1159–63.

Williamson MA, Nawrocki SP, Rathbun TA. 2002. Variation in midfacial tissue thickness of African-American children. *J Forensic Sci* **47**:25–31.

Wiredu EK, Kumoji R, Seshadri R, *et al.* 1999. Osteometric analysis of sexual dimorphism in the sternal end of the rib in a West African population. *J Forensic Sci* **44**:921–5.

Yoder C, Ubelaker DH, Powell JF. 2001. Examination of variation in sternal rib end morphology relevant to age assessment. *J Forensic Sci* **46**:223–7.

Yoshino M, Kimijima T, Miyasaka S, *et al.* 1991. Microscopical study on estimation of time since death in skeletal remains. *Forensic Sci Int* **49**:143–58.

Yoshino M, Miyasaka S, Sato H, *et al.* 1987. Classification system of frontal sinus patterns by radiography. Its application to identification of unknown skeletal remains. *Forensic Sci Int* **34**:289–99.

CHAPTER 4

The pathology of wounds

- Definition of a wound 136
- Mechanism of wounding 136
- Classification of wounds 137
- Forensic anatomy of the skin 137
- Abrasions 138
- Contusions or bruises 143
- Kicking 149
- Lacerations 150
- Blunt penetrating injuries 153
- Incised wounds 153
- Dating of wounds by histology and histochemistry 166
- Survival period after wounding 169
- References and further reading 171

DEFINITION OF A WOUND

A wound or injury can best be defined as 'damage to any part of the body due the application of mechanical force'. Some jurisdictions have a legal definition of a 'wound', which usually requires the integrity of the body surface to be breached. As this obviously excludes bruising and damage to deep organs, it is unrealistic in a medical sense, but there is always some legal alternative, such as 'causing serious bodily harm', which covers any injury to any tissue or organ. Other jurisdictions grade wounds not by their physical nature, but by the perceived risk to life or health of the victim; this can be an extremely difficult prognostic exercise.

MECHANISM OF WOUNDING

The human body is constantly subjected to mechanical forces during the normal course of life, varying from unceasing gravitational forces to the forceful impacts of a sporting contest. The body usually absorbs such forces either by the resilience and elasticity of its soft tissues or the rigid strength of its skeletal framework. It is only when the intensity of the applied force exceeds the capability of the tissues to adapt or resist that a wound or injury occurs.

The intensity of the force obeys the usual laws of physics, in that the force varies directly with the mass of the 'weapon' and directly with the **square** of velocity of impact. Of course, the weapon can be anything, including the body itself, as is often seen in the violent deceleration of a fall or in a traffic accident.

The well-known formula: kinetic energy = ½ mass × velocity^2, means that a one kilogram brick pressed against the scalp will cause no injury, but the same brick thrown against the head at a velocity of 10 m/s may smash the skull. This principle is relevant not only in relation to blunt injury and firearm missiles, but also to stab wounds, as is discussed later.

Another factor of importance is the area over which the force acts. If a plank of wood is struck against the skin, the damage to the tissues will be far greater if the narrow edge is used, than if the impact is from the flat surface. Obviously, the force derived from the same mass and velocity is applied over a smaller area, thus delivering a greater impact to any given unit of tissue. This again is relevant to stab wounds, as all the kinetic energy of a moving knife is concentrated into the tiny area of the tip of the blade; the same energy delivered by the large surface area of a cricket bat might not even cause a bruise.

The effect of excessive mechanical forces on the tissues of the body can cause compression, traction, torsion, and tangential (shear) and leverage stresses. The resultant damage depends not only on the type of the mechanical insult but also on the nature of the target tissue. For instance, violent compression (as in an explosion) may do little harm to muscle, but may rupture lung or intestine, while torsion may leave adipose tissue unaffected, yet cause a spiral fracture in a femur.

The transfer of kinetic energy from the relative movements of the body and the weapon is the source of damage to the tissues. If the attacking object comes to rest in the tissues, then all its kinetic energy must have been transferred. A bullet which passes clean through the tissues thus fails to exchange all its energy into trauma production, which is the reason for the illegal 'dumdum' or more recent, explosive-tipped bullets, the object of which is to expand the missile and prevent its exit from the body, so increasing its 'stopping power' (see Chapter 8). If such energy transfer can be slowed down or spread over a wider area, then the intensity of force applied to the tissues per unit of time is less. This is part of the function of a seatbelt in a vehicle crash, where the stretching of the belt fabric extends the time of energy exchange and the considerable area of the belt surface is a preferable alternative to transferring all the kinetic energy via a few square centimetres of forehead against the windscreen.

Other ways in which the transfer of kinetic energy may be modified is in moving the body in the direction of the force. When a car strikes a pedestrian, that proportion of the momentum of the car that is used in projecting the victim along the road is not then available for damaging the tissues.

Similarly, if a boxer 'rides the punch', he is reducing the differential velocity between his head and his opponent's fist, as well as extending the time over which energy transfer occurs. With a high-speed impact, such as a bullet and – to a lesser extent a stab or violent blow – the inertia of the victim's body prevents any significant loss of energy transfer by movement. If the force does not strike in a linear fashion, but leads to a tangential impact – as in a glancing blow causing a 'graze' – then only part of the kinetic energy is transferred and the damage will be proportionately less.

CLASSIFICATION OF WOUNDS

This is quite arbitrary, but when writing medical reports (especially for legal purposes) it is essential that every doctor should use a standardized nomenclature when describing wounds. All too often reports are seen – especially from clinicians who are not used to medico-legal work – where 'laceration', 'incised wound' and 'abrasion' are used loosely, the accompanying descriptions making it obvious that no clear distinction between them was appreciated by the doctor.

Wounds may also be categorized by the motivation of their infliction, such as 'suicidal or accidental' (the 'manner' index of the WHO International Classification of Disease, ICD E-codes), but a pathologist should classify them by their appearance and the method of causation, such as 'laceration' or 'incised wound'. The most useful classification is:

- Abrasions – colloquially 'grazes or scratches'.
- Contusions – colloquially 'bruises'.
- Lacerations – colloquially 'cuts or tears'.
- Incised wounds – colloquially 'cuts', 'slashes' or 'stabs'.

FORENSIC ANATOMY OF THE SKIN

As most wounds involve the body surface, a reminder of the structure and nomenclature of the skin and subcutaneous tissues is appropriate. Most superficial is the keratinized dead layer of cells, the stratum corneum, which varies greatly in thickness from one part of the body to another. That on the soles and palms is the thickest, while that on protected areas such as the scrotum and eyelids measures

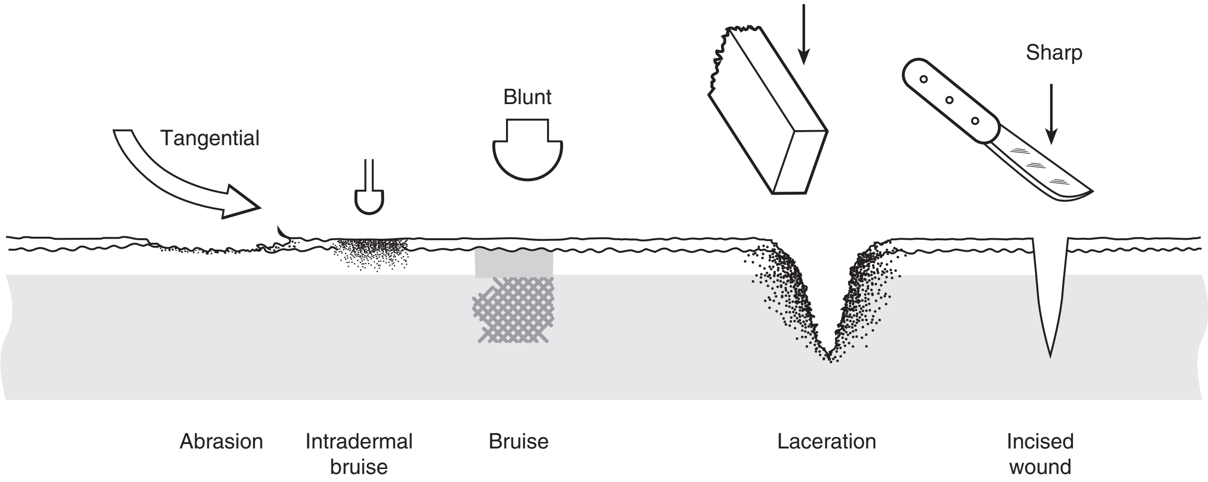

FIGURE 4.1 *Types of injury to the skin.*

only a fraction of a millimetre. This has forensic relevance in the amount of injury that is needed to penetrate the skin and allow bleeding from the underlying tissues.

The living layers of the skin combine with the horny layer to form the epidermis, which has no blood vessels in its thickness. The epidermis is generally corrugated, the under surface by papillae that dip into the dermis. The degree of undulation also varies greatly from place to place, the thinner skin tending to have a flatter junction between dermis and epidermis. The dermis (or 'corium') consists of mixed connective tissue carrying the skin adnexae, such as hair follicles, sweat glands and sebaceous glands. It has a rich network of blood vessels, nerves and lymphatics, and has numerous nerve endings of various types for tactile, pressure and heat sensing. The lower zone of the dermis has adipose tissue and – depending on the site in the body – deep fascia, fatty tissue and muscle will form strata below the skin itself.

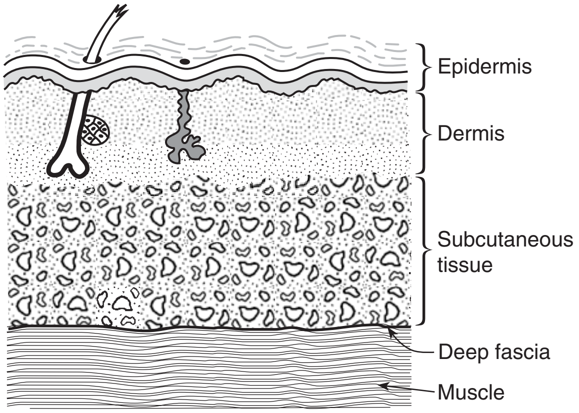

FIGURE 4.2 *General structure of skin.*

ABRASIONS

An abrasion is the most superficial of injuries and, in the most restrictive of definitions, is one that does not penetrate the full thickness of the epidermis. Thus the pure abrasion does not bleed, as blood vessels are confined to the dermis. Because of the corrugated nature of the dermal papillae, however, many abrasions enter the corium and bleeding commonly occurs.

Another definition (all of which are arbitrary) would describe an abrasion as a superficial injury to 'the skin', which would allow penetration of the upper dermis rather than only the epidermis, so that bleeding would fall within the definition.

Abrasions are known in lay terms as 'scratches' or 'grazes', though the former usually indicates a linear mark and the latter a 'brush' abrasion caused by wider tangential impact. Where these brush abrasions are the result of scraping contact with the ground, the lay term is a 'gravel rash'. There are many causes of abrasions, which are common everyday lesions, especially in children whose legs are rarely free from some scratches and bruises. Any contact that rubs across the epidermis and removes the keratinized layer and underlying cells will cause that area to become discoloured and moistened by exuded tissue fluid, even if bleeding does not occur from abraded dermal papillae.

When death ensues soon afterwards, the abrasion becomes stiff, leathery and of a parchment-like brown colour as a result of the drying of the moist exposed surface. This is classically seen, for example, in the ligature mark of hanging or strangulation. It is impossible to tell whether a superficial abrasion occurred immediately before or after

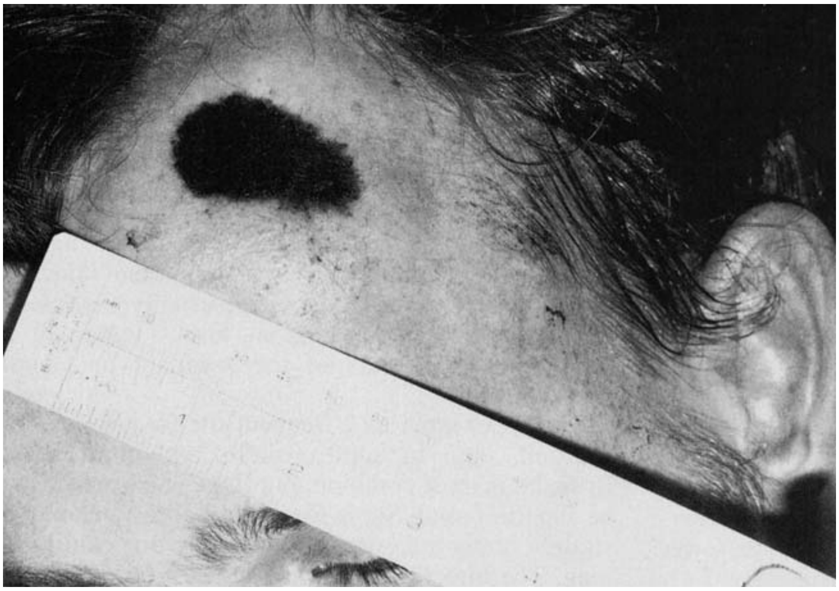

FIGURE 4.3 *Simple abrasion of the skin caused by an almost perpendicular impact of the head against the ground. There is only slight tangential scuffing though most abrasions have some element of sideways contact that damages the epidermis.*

death if there was insufficient time for any inflammatory reaction to occur at the margins. Post-mortem abrasions are common, especially after autopsy when the body is re-examined, because mortuary instruments and the moving of bodies into refrigerators and coffins can often make marks. These abrasions commonly appear yellow and translucent, and are absolutely devoid of any colour change at the edge.

Tangential or brush abrasions

Most abrasions are caused by a lateral rubbing action rather than vertical pressure. Where this tangential component is marked, the direction of the force can often be determined by tags of epidermis dragged to the terminal end of the abrasion. For example, in the 'brush abrasion', which is common in childhood falls and pedestrian accidents, the skidding of the body across a rough surface scrapes linear furrows across the skin. Strands and tags of epidermis may be peeled along these furrows to the further end where contact ceased. Visual examination, using a lens if necessary, can indicate the direction of movement of the body. Similarly, if a victim is struck a glancing blow with a rough object such as a stone or brick, similar epidermal tags may indicate the direction of the blow.

The same type of grooved abrasions can be caused when a victim is dragged along the ground, either in a vehicular accident or by an assailant. It may be important, though difficult, to attempt to differentiate such drag marks as being ante-mortem or post-mortem, a problem which is discussed in another chapter.

Crushing abrasions

Where the impact is vertical to the skin surface, no scraping or tangential marks occur. Instead the epidermis is crushed and an imprint of the impacting object is stamped on the surface. If the impact is substantial and the area of contact small,

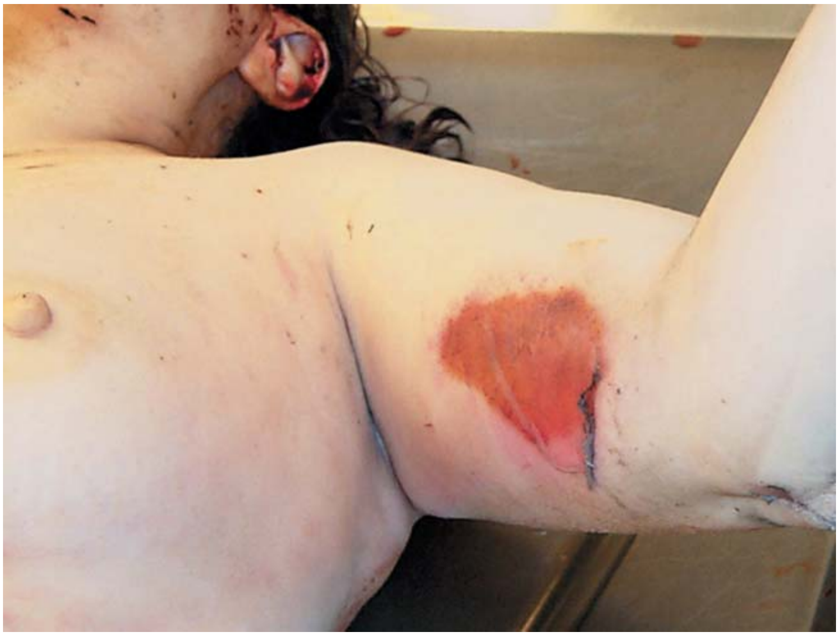

FIGURE 4.4 *A linear abrasion or 'graze', confined to the upper layers of the skin. The tangential direction of impact of the weapon was from above downwards (along the longitudinal axis of the upper arm), as can be determined by the shreds of epidermis peeled towards the lower end.*

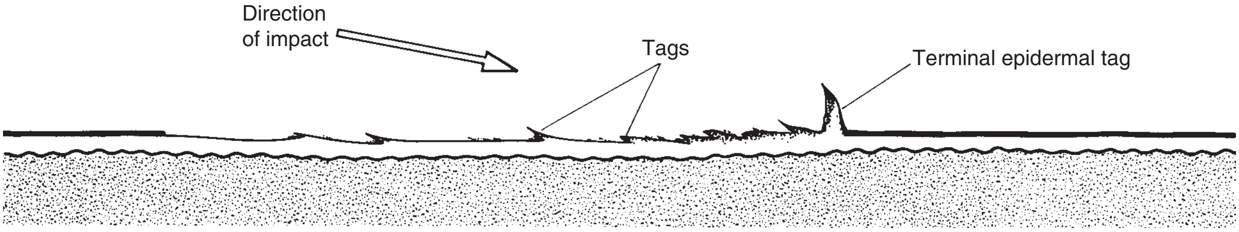

FIGURE 4.5 *Determination of the direction of impact in an abrasion caused by a tangential force. The epidermal tags raised by the impact tend to pile up at the distal end.*

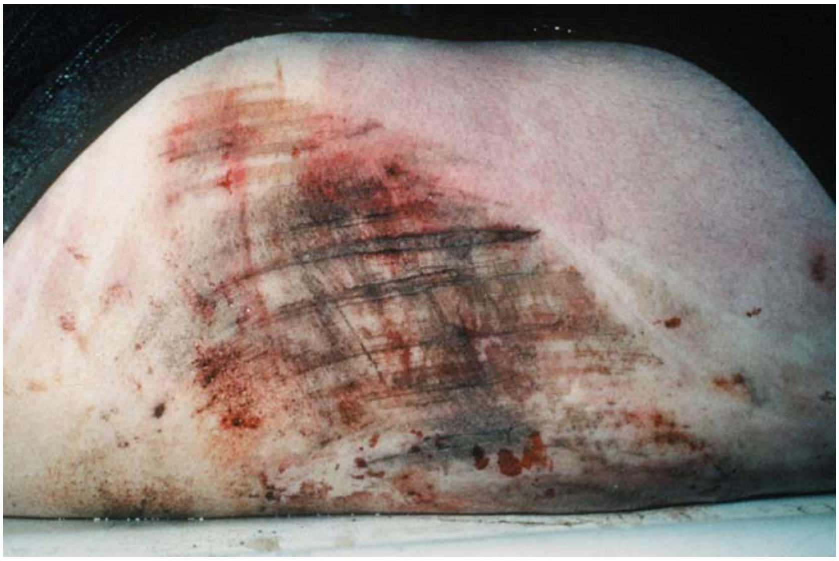

Figure 4.6 *'Brush abrasion' or 'grazes' caused by tangential contact of a moving body with a rough surface. Although most are seen in traffic accidents from contact with the road surface, this example is from a murdered woman dropped down a mine shaft. There are grazes running in two directions at right angles, indicating that she must have struck the wall twice, in different postures.*

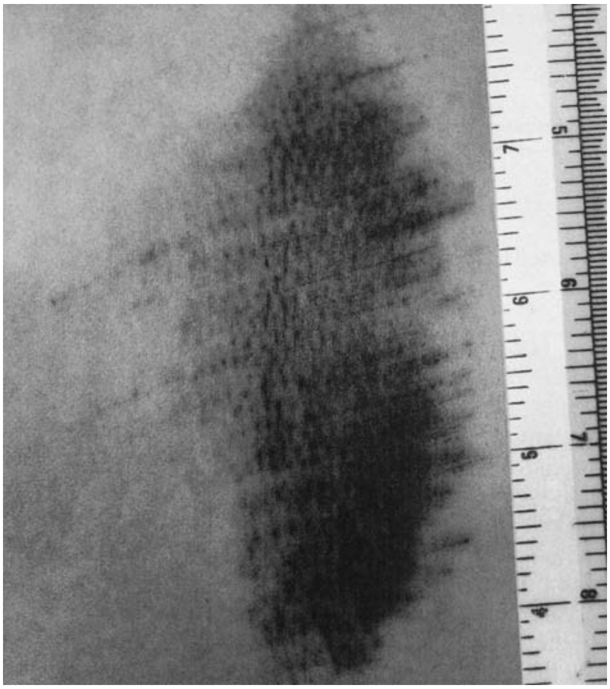

Figure 4.7 *Brush abrasion from skidding contact with the road surface during a drunken scuffle. The lesion is partly abraded and partly bruised. The direction of movement can be discerned, as the scratches begin more abruptly and deeply near the ruler, and tail off superficially at the opposite edge.*

a punctured wound will be made, but otherwise a crushed abrasion will occur. The lesion is slightly depressed below the surface unless an underlying bruise or local oedema bulges the tissues. These abrasions are the ones that most clearly reproduce the pattern of the injuring object. An example

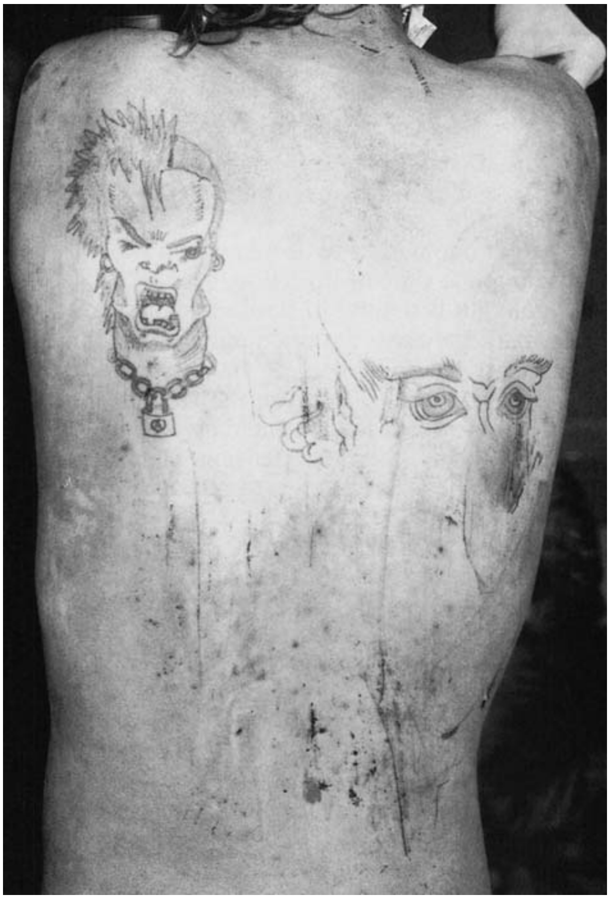

Figure 4.8 *Drag marks on the back of a murder victim. Death was caused by a stab wound of the front of the chest – the body was pulled along the ground for a short distance, obviously in the direction of the feet, as the scratch near the waist is deeper near the feet end and tails off at the opposite end.*

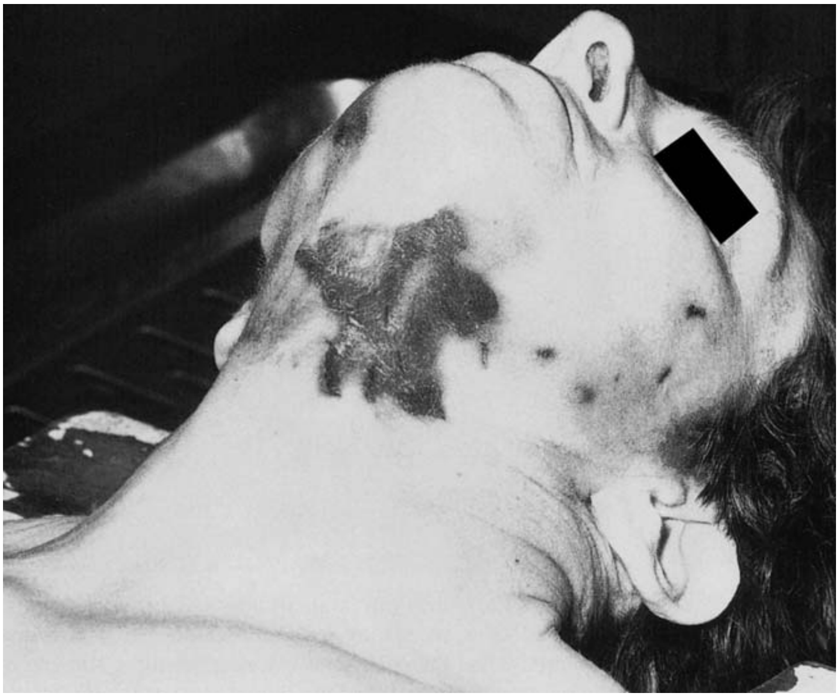

FIGURE 4.9 *Abrasions in manual strangulation. The area is more extensive than is usually seen in such circumstances. The large area of superficial damage to the epidermis is caused by sliding movements of the assailant's hands, the area having become dried and leathery during the post-mortem interval. The smaller marks are fingernail scratches.*

would be the marks of a vehicle radiator on a pedestrian victim or the pattern of a floor grid on to which a person has fallen. If the impact is forcible, then the dermis may also be injured: an area of bruising may underlie the abrasion.

Fingernail abrasions

These are important because of their frequency in assaults – especially child abuse, sexual attacks and strangulation. Often associated with focal bruises, fingernail abrasions are most often seen on the neck, the face, the upper arms and the forearms. They may be linear scratches if the fingers are dragged down the skin, or short, straight or curved marks when the skin is gripped in a static fashion. As women tend to have longer, sharper fingernails than men, they are naturally more often associated with causing such abrasions. A victim resisting a sexual or other attack may rake her nails down her assailant's face, causing linear, parallel scratches that may be several millimetres wide and placed a centimetre or two apart. The expected pattern may be fragmentary, however, as is often seen on the neck when a victim of either manual or ligature strangulation attempts to tear away the attacking fingers or cord. These marks are usually vertical, as opposed to the more random marks that may be inflicted by the nails of the assailant in manual strangulation.

The upper arms are a frequent site for gripping and restraint, both in adult assaults and child abuse. Bruising is most common, but fingernail marks may be superimposed. Static fingernail abrasions may be straight or curved, often about half to one centimetre long. The direction of curvature must be interpreted with care if one wishes to decide which way the hand was held at the time of infliction.

Although it is natural to assume that the concavity of the mark indicates the orientation of the fingertip, experiments by Shapiro *et al.* (1962) have shown that this is often not the case. Because the skin is put under lateral tension when it is indented by the nails, it may distort, so that when the tension is released the elasticity of the skin causes it to return to its original position, carrying the nail mark with it. The curve may then reverse to form either a straight line or a convexity. The shape of the free edge of the fingernail also affects the mark, as pointed nails are more likely than those with straight edges to give these paradoxical results. Once again, the pathologist has to be wary of incorrect interpretation when, for instance, deciding if nail marks on a neck were made by hands approaching from the front or passing around the back of the neck. However, personal experiments with the Shapiro *et al.* contention have shown that it by no means always applies.

Patterned abrasions

Formerly it was often claimed that abrasions retained the pattern of the impacting object more accurately than other injuries such as bruises and lacerations. Though abrasions undoubtedly do preserve such patterns well, many of the examples were not in fact true abrasions, but were intradermal bruises, mentioned in the next section.

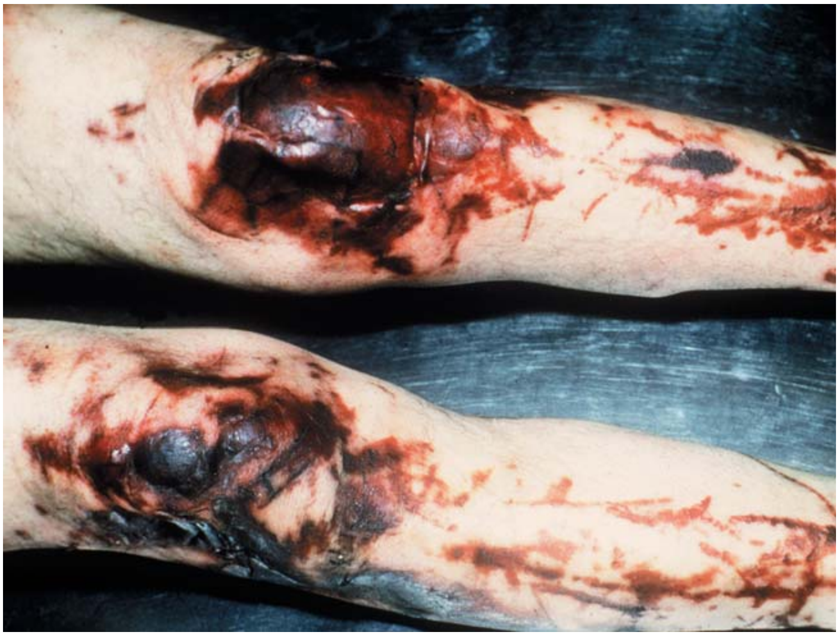

FIGURE 4.10 *Extensive abrasion of the knees and shins, in a drunk who stumbled amongst furniture before dying of a head injury. The damaged skin has exuded tissue fluid, which has dried post-mortem to produce the dark, leathery appearance.*

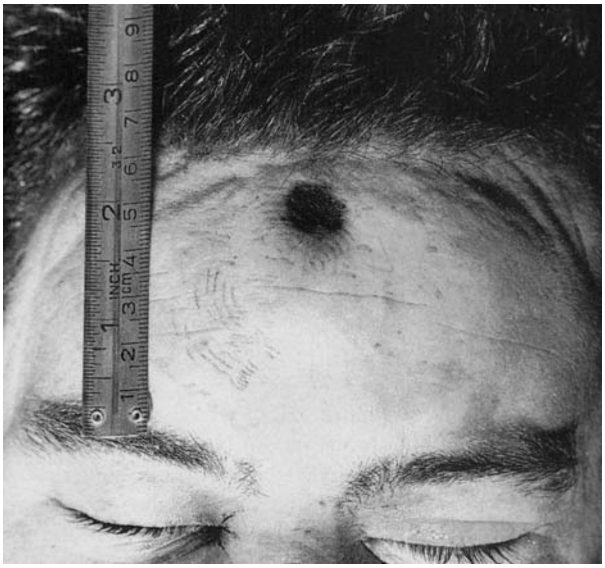

FIGURE 4.11 *Patterned abrasion of the forehead. The victim was struck on the head in a public house with a heavy glass ashtray, the bottom of which had embossed concentric circles moulded into the glass.*

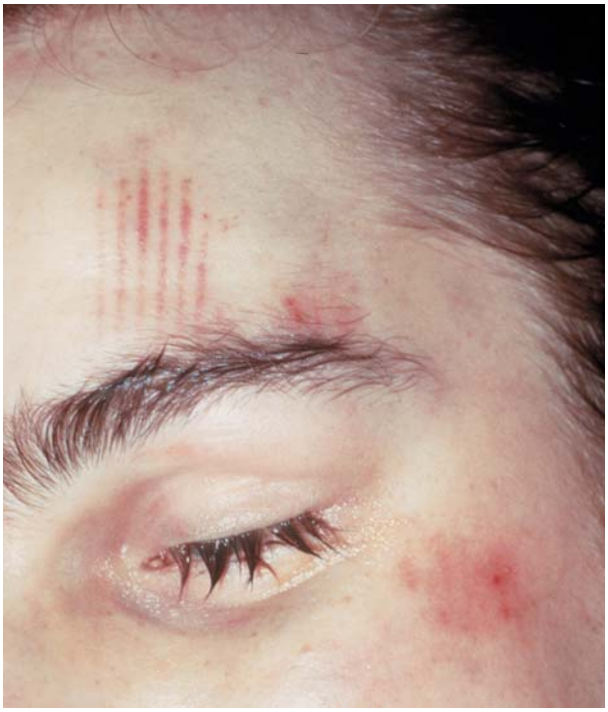

FIGURE 4.12 *Abrasions and intradermal bruising on the forehead of a swimming bath attendant. He was found drowned when there were no witnesses, but the spacing of the marks exactly matched the ridged tiles at the edge of the bath, so it was presumed that he had slipped and struck his head before falling unconscious into the water.*

Patterned injuries occur when the force is applied at or near a right angle to the skin surface, rather than with the skidding impact of a graze. If a weapon with a patterned surface strikes the skin – or the body falls against a patterned surface – the abrasion of the epidermis follows the ridges of the object if it has a profile of varying height. Not only may the epidermis be damaged, but the skin may be compressed into the cavities of the pattern, with consequent capillary damage leading to an intradermal bruise. Probably the best example of this is seen when a motor tyre passes over the skin, leaving a pattern where the skin has been squeezed into the grooves of the rubber tread.

There is little point in trying to list all the possible patterns that can be distinguished in abrasions on a body, but a few

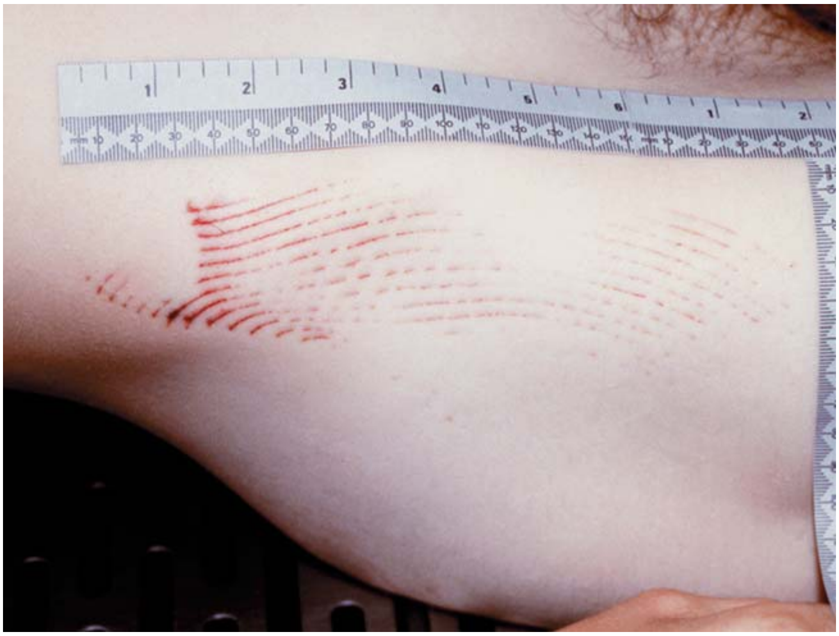

FIGURE 4.13 *Patterned abrasions from the back edge of a serrated 'Rambo' knife. Measurements and photographs with a scale should always be obtained to assist in the identification of the weapon.*

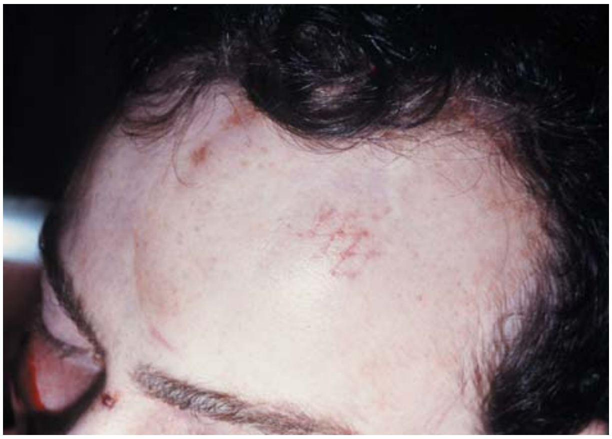

FIGURE 4.14 *Patterned abrasion on forehead due to falling down steps onto a metal grid. Here the abrasions are not produced by tangential force, but imprinted directly by impact at a right angle to the skin.*

have particular medico-legal significance. In former years the honeycomb grid of a motor-vehicle radiator provided many examples of patterned abrasions, but changes in vehicle design have relegated these to historical interest. There may still be projections on vehicles that cause damage, but these are more likely to inflict bruises and lacerations. The muzzle of a firearm can imprint an abrasion on the skin, which is of importance in that it confirms that the discharge was contact in nature. Impact against ribbed ceramic tiles in a bathroom or swimming bath may assist in reconstructing unwitnessed events. Blows from a weapon with a recognizable surface may help to identify that weapon, such as a plaited rope or leather whip, or a solid object with an embossed pattern. Abrasions from objects with a recurring pattern, such as bicycle chain used in gang fights, or a serrated knife, can readily provide a clue to the nature of the weapon.

Post-mortem abrasions

Unlike post-mortem bruises, artefactual abrasions are common. They may have been inflicted after death from a variety of causes, including dragging a corpse or buffeting in moving water. Some post-mortem animal injuries also resemble abrasions such as insect bites, especially by ants.

Other damage may be caused following autopsy. As the post-mortem interval increases, so the skin becomes more fragile. Even the normal procedures of post-autopsy reconstruction and handling in the mortuary may cause dermal damage, especially after washing with hot water. If the pathologist returns for a later examination, or if he is retained to perform a second autopsy for the defence, the appearances should be checked with the original description or photographs, if some injuries suggest a post-mortem origin.

CONTUSIONS OR BRUISES

Although often combined with abrasions or lacerations, a pure bruise lies beneath an intact epidermis and consists of an extravascular collection of blood that has leaked from blood vessels damaged by mechanical impact. An extravasation of blood that is (arbitrarily) larger than a few millimetres in diameter, is usually termed a 'bruise' or 'contusion'. This size overlaps the older and now little used term 'ecchymosis', which is really a small bruise.

Even smaller is the 'petechial haemorrhage', which is the size of a pin head or less. Both ecchymoses and petechiae are not usually caused by direct mechanical trauma and are often seen on serous membranes and conjunctivae as well as on skin. However, moderate pressure, impact or, especially, suction on the skin can produce a patch of localized petechiae.

Bruises are caused by damage to veins, venules and small arteries. Capillary bleeding would be visible only under a microscope and even petechiae originate from a larger order of blood vessel than a capillary.

The word 'bruise' usually implies that the lesion is visible through the skin or present in the subcutaneous tissues, whilst a 'contusion' can be anywhere in the body, such as the spleen, mesentery or muscles. The two words are often interchanged at random, however, though 'bruise' is to be preferred when a doctor gives reports or evidence to a non-medical audience.

Intradermal bruises

These are important but rarely mentioned in most texts. The usual bruise from a blunt impact is situated in the subcutaneous tissues, often in the fat layer. When viewed through the overlying corium and epidermis, the bruise is somewhat blurred, especially at the edge. When a bruise is made by impact with a patterned object, however, the haemorrhage may be far more sharply defined, if it lies in the immediate subepidermal layer. The amount of blood is relatively small, but because of its superficial position and the translucency of the thin layer that overlies it, the pattern is distinct.

Such bruises are especially likely to occur when the impacting object has alternating ridges and grooves, as the skin will be forced into the grooves and be sharply distorted. Intradermal bleeding will occur here and the areas in contact with the raised ridges may remain pale, as the pressure forces the blood from the small vessels. A good example is that of a motor tyre running across the surface. Impacts from whips with patterned thongs may also show the same phenomenon, as do the ribbed rubber soles of 'trainer' shoes.

Factors affecting the prominence of a bruise

Several factors influence the apparent size and prominence of a bruise and, because of these, it is not possible to be dogmatic about the amount of force needed to produce any given bruise.

- As it is a leakage of blood from a vessel, there must be sufficient space outside that vessel for free blood to accumulate. This explains the ease with which bruising appears in lax tissues such as the eye socket or scrotum and its rarity in the sole of the foot or palm of the hand, where dense fibrous tissue and restrictive fascial planes prevent accumulation of blood. Because of the greater volume of soft subcutaneous tissue in fat people, they

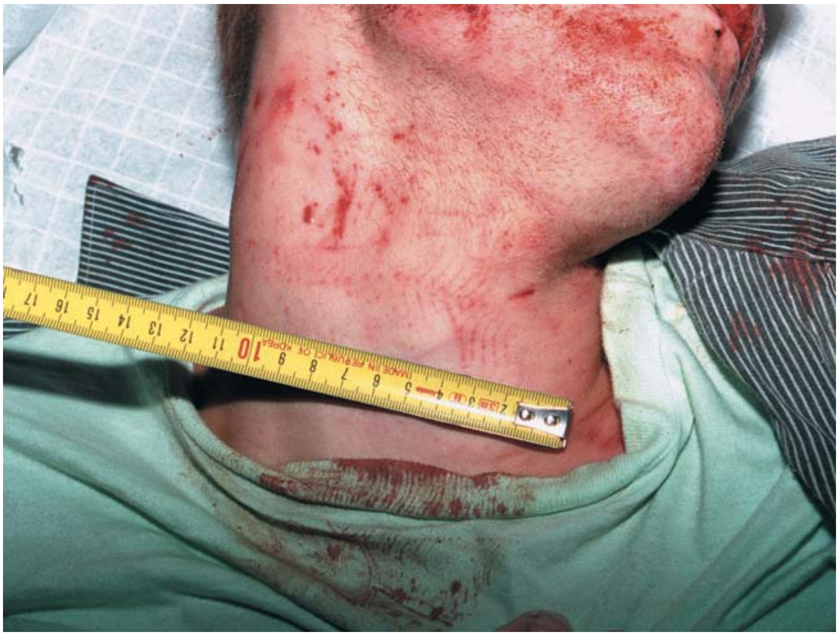

FIGURE 4.15 *Intradermal bruising showing the pattern of rubber soles of 'trainer' shoes on the neck and T-shirt of a homicide victim.*

tend to bruise more easily than thin ones, other factors being equal (such as vessel fragility and senile changes).

- The apparent prominence of a bruise beneath the skin varies with the amount of blood in the extravasation. The size of the haemorrhage depends partly, but not entirely, on the intensity of the injuring force. The size and density of the vascular network varies from place to place and the amount of damage that a given blow causes to local blood vessels is partly a matter of chance.
- Resilient areas, such as the abdominal wall and buttocks, bruise less with a given impact than a region where underlying bone acts an anvil with the skin between the bone and the inflicting force. The head, chest and shins are examples.

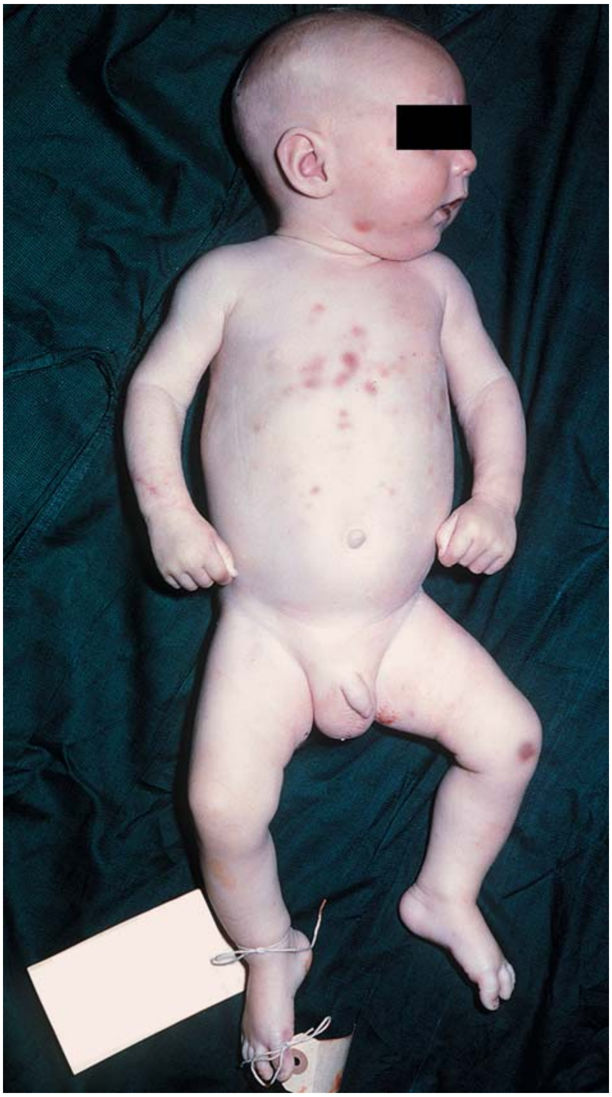

FIGURE 4.16 *Multiple bruises on the trunk of a victim of child abuse. The bruises are of the 'fingertip' type, caused by heavy prodding by adult fingers. The child died of a ruptured liver.*

- The depth at which the bruise is placed affects the apparent severity. A bruise may be placed superficially in the dermis to form the well-patterned intradermal bruise mentioned in the previous section; here, a minute amount of blood will be obvious. Most bruises are in the subcutaneous tissues above the deep fascia and will therefore be fairly obvious, but others can be confined below deeper fascial membranes so that the free blood has to be viewed through the skin and underlying adipose tissue. For a given size of extravasation, this bruise will be less prominent. Some bruises are confined to deep fascial compartments and never become visible without dissection.
- For a given impact, the volume of blood lost into the tissues can depend upon the fragility of the blood vessels and the coagulability of the blood. In old persons, vessel fragility may be extreme and a large bruise may develop from the slightest of knocks. Children tend to bruise more easily than adults, presumably because of the softer tissues and the smaller volume of protecting tissue that overlies the vessels. Any bleeding diathesis resulting from disease, a toxic condition, or certain medication, will also retard the normal clotting process that heals the breach in the bleeding vessels. Those with scurvy and chronic alcoholics bleed easily but, in contrast, certain people (such as boxers) seem able to avoid bruising from blows that would severely damage other people.
- It is common knowledge among lay people, as well as doctors, that a bruise may 'come out' – that is, become more prominent with the passage of hours or days. This is partly caused by continued bleeding from the ruptured vessels, but mainly by percolation of free blood from its origin deeper in the tissues upwards towards the epidermis. Another factor may be haemolysis, when the freed haemoglobin is able to stain the tissues in a more diffuse way and become more noticeable than intact red blood cells. This latter mechanism is certainly the reason not only for the well-known post-mortem phenomenon of bruises becoming more prominent after death, but of new bruises appearing later where none was visible at an autopsy performed soon after death. This is further considered in the chapter on post-mortem changes, but is repeated here because of the importance of recognizing the differences in appearance that can occur between two autopsy examinations spaced a few days apart. The second, usually for a defence opinion, may find new bruises not recorded by the first pathologist, but if the phenomenon of delayed appearance is appreciated, potential disputes may be avoided.

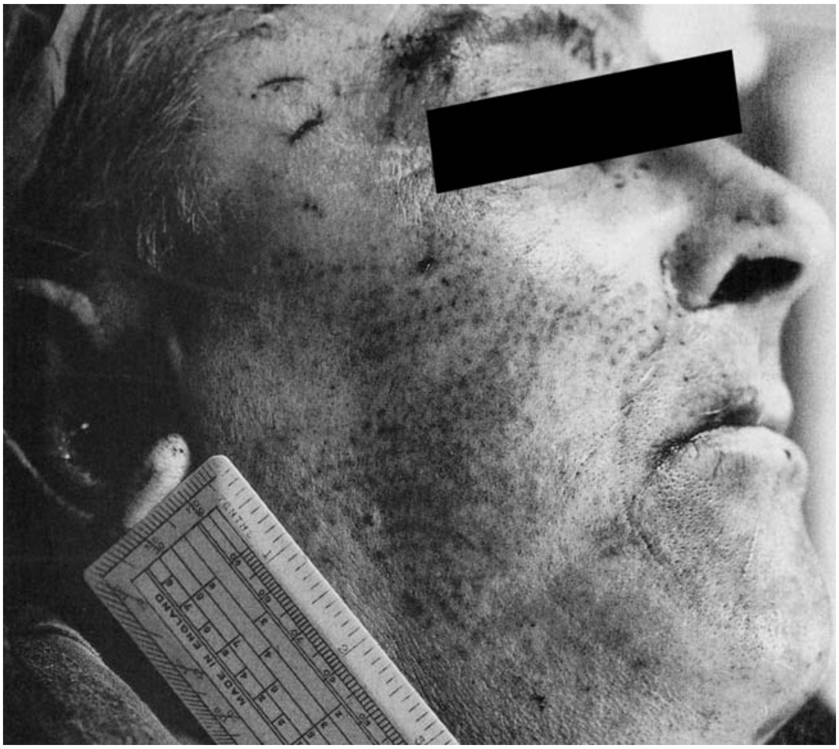

FIGURE 4.17 *Patterned abrasion and intradermal bruising of the face during a fatal armed robbery. The watchman was struck in the face by an assailant wearing a hard-corded driving glove with a coarse-weave pattern.*

Movement of bruises

A bruise may appear at a different place on the surface from the point of impact. When the bruise is superficial, especially intradermal bruising, the lesion appears immediately – or at least rapidly – and is at the point of infliction. When blood extravasates in the deep tissues, however, it may take time to reach the surface (if it ever does), and this may be some distance away because of deflection and obstruction by fascial planes and other anatomical structures.

In addition, bruises may move under gravity. The most frequent example is a bruise or a bleed under a laceration on the upper forehead. If the victim survives for at least some hours, then the subcutaneous haemorrhage can slide downwards over the eyebrow ridge and appear in the orbit, to give a 'black eye', which might be misinterpreted as direct trauma. Similarly a bruise of the upper arm or thigh may surface lower down around the elbow or knee.

Alteration of bruises with time

As already mentioned, bruises often become more prominent some hours or days after infliction because red cells or haemoglobin diffuse closer to the translucent epidermis. There is another temporal series of changes in bruises in the living person, this being part of the healing process. Fresh extravasation of blood is obviously dark red, though when viewed through the skin this may be purple or almost black in appearance. In racially pigmented victims, a bruise may sometimes be undetectable from the surface, apart from swelling caused by a haematoma and tissue oedema.

With the passage of time, the haematoma breaks down under the influence of tissue enzymes and cellular infiltration. The red-cell envelopes rupture and the contained haemoglobin undergoes chemical degradation, which causes a sequence of colour changes. The haemoglobin is broken down into compounds including haemosiderin, biliverdin and bilirubin, which lead the colour changes through a spectrum of purple to bluish brown, to greenish brown to green to yellow, before complete fading.

A small skin bruise in a healthy young adult might be expected to pass through all these stages and vanish in about a week, but there is a tendency to overestimate the length of time needed as shown by Roberts (1983), who observed that 'love bite' bruises in sexual offences could become yellow and vanish within a couple of days. Langlois and Gresham (1991) reviewed the literature on this subject, indicating the wide variation in opinion. They also made careful photographic records of the macroscopic appearance of bruising on 89 subjects, aged between 10 and 100 years. They concluded that the most significant change was the appearance of a yellow colour (in persons less than 65 years of age), which indicated that the bruise could not be less than 18 hours old. Blue, purple and red did not assist in

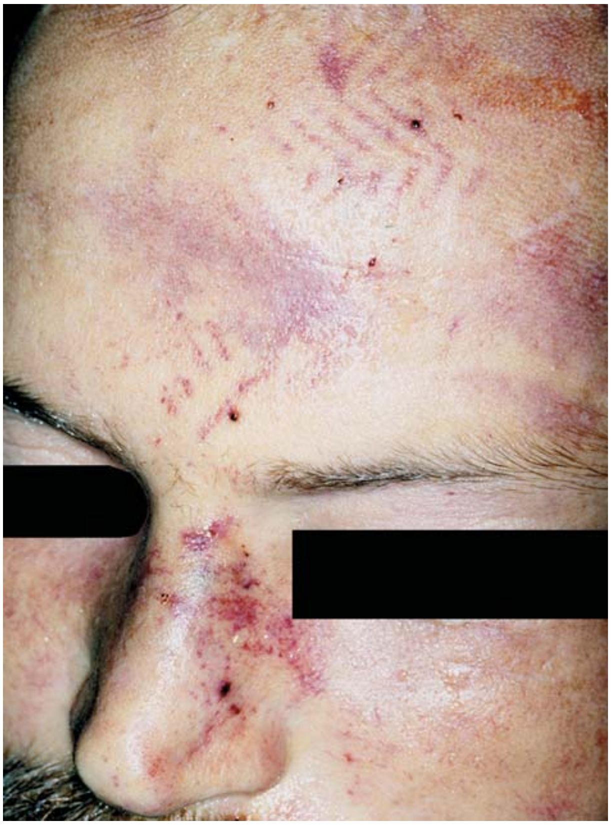

FIGURE 4.18 *Kicking and stamping injury to the face. The nose is bruised from a kick and the patterned rubber sole of the shoe has imprinted intradermal bruising on the forehead. It is essential to obtain accurate photographs and measurements of the shoe-tread pattern, to allow identification of the footwear.*

dating bruises; brown was held to be a mixture of colours and was not considered as useful.

In the present authors' experience, it is impossible to comment on the age of a bruise less than 24 hours since infliction, except to say that it is 'fresh', as no visible changes occur in that time.

It is not practicable to construct an accurate calendar of these colour changes, as was done in older textbooks, as there are too many variables for this to be reliable. These include:

- the size of the extravasation – changes begin at the margin and a larger bruise will take a longer time to be absorbed. A large old bruise may contain all the colours possible – from purple in the centre to yellow at the edges
- the age and constitution of the victim. Aged persons may not heal their bruises at all and carry them for the remainder of their lives

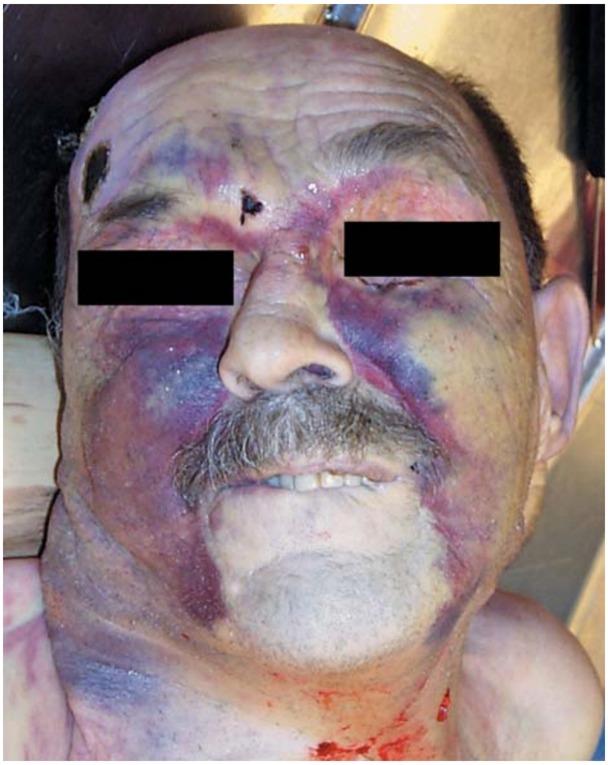

FIGURE 4.19 *Extensive bruising of the face due to hitting, kicking and stamping 6 days earlier. The victim had multiple fractures of the facial bones, bilateral serial rib fractures and a tension pneumothorax.*

- a personal idiosyncrasy in the rapidity with which such changes occur in any one person, including coagulation defects.

Even histological examination is unreliable in the accurate dating of bruises, as is discussed later in this chapter. The appearance of stainable iron, in the form of haemosiderin, however, does not usually appear within the first 2 or 3 days, though Simpson suggests that it may be found as early as 24 hours. In meningeal haemorrhage, it seems to appear from around 36 hours.

Haematoidin, another breakdown product of blood pigment, can appear in old bruises and haematomas after the first week.

Though an absolute date cannot therefore be placed upon a bruise, the following observations are legitimate.

- If a bruise appears fresh over all its area, with no observable colour change, it is unlikely to have been inflicted more than about 2 days before death, except in old persons.
- If the bruise has any green discoloration, it was inflicted not later than 18 hours before death (Roberts 1983; Langlois and Gresham 1991).

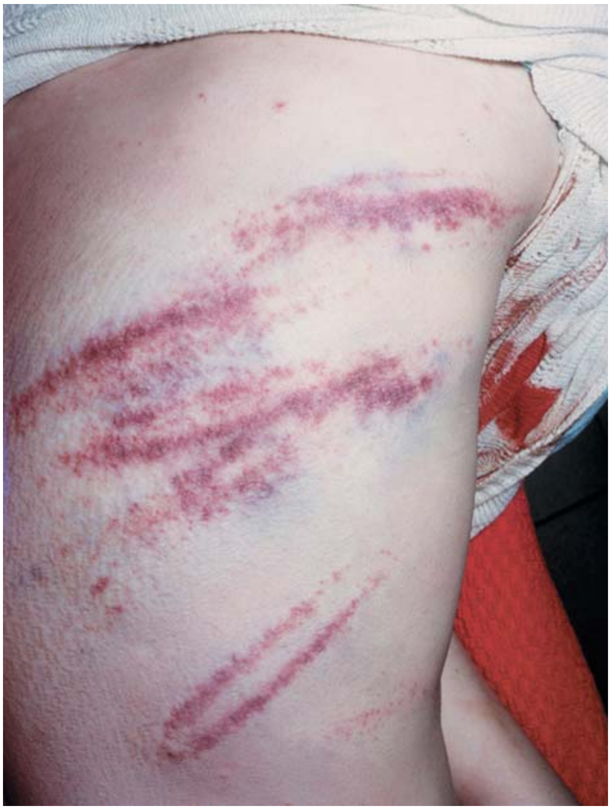

FIGURE 4.20 *Bruises from a beating with a broom handle. They are approximately parallel and several, especially the lowermost, show a double 'tram-line' appearance typical of the impact of a round or square-section rod. The pressure in the centre compresses the vessels so that they do not bleed.*

- If several bruises (of roughly comparable size and site) are present and are of markedly different colours, then they could not have been inflicted at the same time. This is particularly important in suspected child abuse, where intermittent episodes of injury have important diagnostic significance.

Bruising of special significance

Certain types of bruise and bruises at particular sites have a specific significance. Clusters of small discoid bruises of about a centimetre in diameter are characteristic of fingertip pressure from either gripping or prodding. These groups are often seen in child abuse, when an adult hand grips the infant by a convenient 'handle'. Once called 'sixpenny bruises' from their size, the lesions are commonly seen on the forearms or upper arms of the child, or sometimes around the wrist or ankle, though they can occur on the abdomen. Similar bruises from fingertips may be seen on the neck of children or adults in manual strangulation, though there is often additional diffuse bruising caused by a sliding grip on the neck.

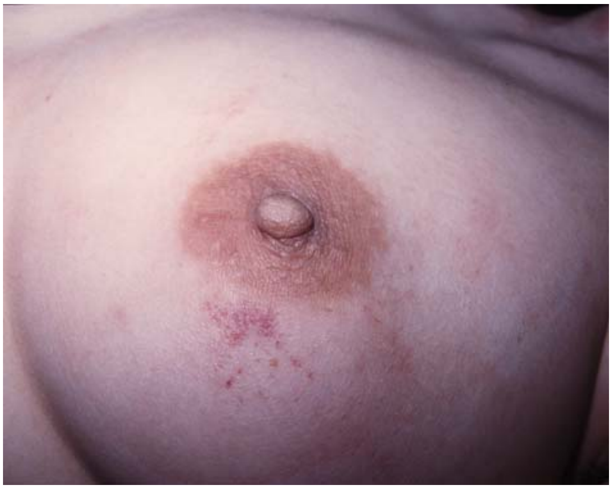

FIGURE 4.21 *Suction marks in the left breast of a 22-year-old homicide victim, who was killed by manual and ligature strangulation.*

When the skin surface is struck by a rod or rectangular sectioned object such as a cane or lath, the consequent bruising may be of the 'tram-line' or 'railway line' type. This appears as two parallel lines of bruising with an undamaged zone in the centre. The mechanism of this double line is that the weapon sinks into the skin on impact so that the edges drag the skin downwards and the traction tears the marginal blood vessels. The centre compresses the skin, which, in the absence of underlying bone, causes little or no damage to the vessels. When the momentary impact is released, blood flows back into the injured marginal zones and leaks into the tissues. Impact from broom handles, narrow planks and wood or metal rods can all cause this characteristic lesion.

As with abrasions, bites can result in bruises; this is dealt with in Chapter 26. It might be noted here that the so-called 'love bites' are often bruises, with or without associated abrasions, being a shower of small petechial lesions caused by oral suction on the skin.

The common 'black eye' is dealt with under 'Head injuries' in Chapter 5, but again it is worth repeating that not all black eyes are true bruising from a blow in the orbit. Some are from fractured orbital roofs and others are the result of gravitational movement of a forehead injury.

A bruise below the ear in a death from subarachnoid haemorrhage needs careful examination of the upper cervical spine and basilo-vertebral arteries, again as discussed under 'Head injuries' in Chapter 5.

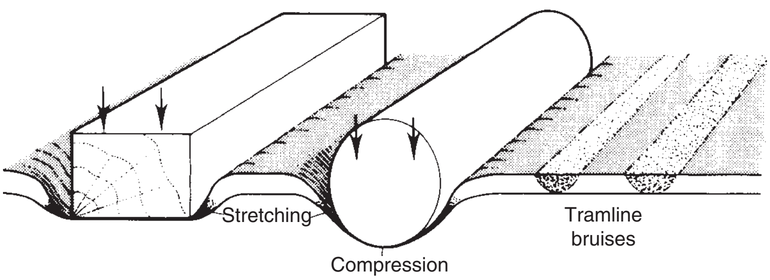

FIGURE 4.22 *Formation of 'tram-line' bruising from the application of a rectangular or cylindrical object.*

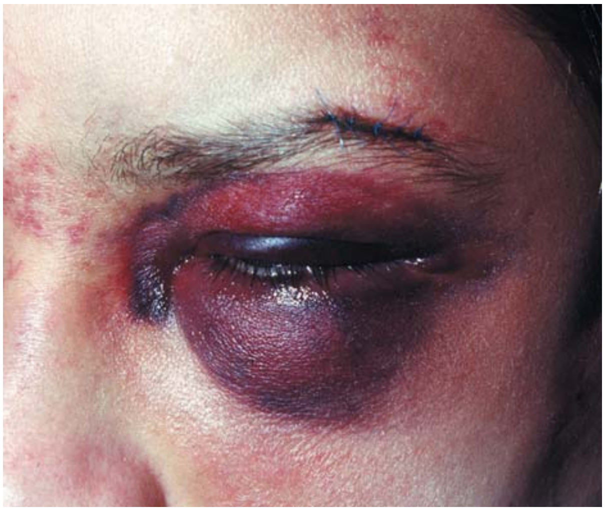

FIGURE 4.23 *A black eye (peri-orbital haematoma) of a live victim from a direct fist blow, which has also bruised the bridge of the nose and lacerated the corner of the eye (sutured).*

KICKING

The shod foot may leave a patterned bruise on the body, most often on the abdomen or chest, though it may be recognizable on the neck or face. A glancing kick is more likely to cause scuffed abrasions and bruising, and a kick from a toecap, a curved abrasion or even laceration. A vertical stamping action may imprint an intradermal bruise corresponding to the sole pattern, especially in these days of rubber 'trainer' shoes. As with a motor tyre, the skin can be forced into the slots in the sole, creating an accurate pattern of superficial bruising. There may also be more diffuse deep bruising if the impact is severe, but this is less useful in delineating an identifiable sole pattern.

Kicking has become increasingly common with the increase in violence and vicious hooliganism in society. The heavy footwear fashionable among some of the least responsible citizens adds to the severity of the lesions.

Though most kicking injuries end up in hospital casualty departments, an increasing number reach the pathologist. Not all victims are drunken youths, as many assaults upon old people by intruders and muggers involve severe kicking.

Obviously, most kicks are delivered with the victim already lying or sitting on the ground, having been brought there by some other means of violence, including pushing, tripping and fist punches. Once prostrate, the 'boot is put in' towards the most convenient targets, which tend to be the concavities of the body. The loin, the groin, the neck and the face are often attacked, and – when the victim assumes a 'fetal position' on his side for protection – swings may be taken into the abdominal area. As an alternative to swinging kicks, the foot may be brought vertically down on the prone victim, the heel being the most damaging part, rather than the toecap. A grinding motion can also be added, which may tear the skin and cause ragged lacerations.

Yet another variant is standing on the body or actually jumping onto it with one or both feet. The chest is the most common target for this practice and fractured ribs or sternum, a flail chest and cardiac injuries can result.

The major characteristic of most kicking injuries is the severity of bruising and underlying damage. Even if the toecap does not cause lacerations, there may be large haematomata and intramuscular contusions from heavy kicks. Abdominal and thoracic viscera can be ruptured and the genitalia are also vulnerable, especially the scrotum and sometimes urethra.

A common danger with kicking to the face is fracture of either (or both) mandible and maxilla, nasal bones and zygoma, though the cranial vault is seldom fractured and kicks to the vertex of the head seem less common. With such facial injuries, death may occur from filling of the pharynx and air passages with blood. A kick on one side of the face can completely detach the lower maxilla – with the upper dental arch and palate – from the rest of the maxilla. A unilateral kick of the jaw can fracture both rami of the mandible, there being no necessity for kicks on either side.

Post-mortem bruising and other artefacts

It is possible to cause what appears to be a bruise by deliberate violence to a dead body, but the amount of force

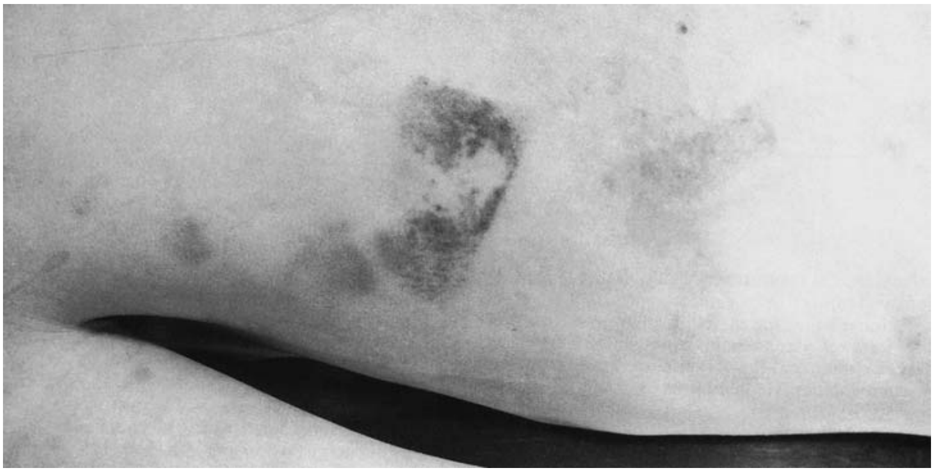

FIGURE 4.24 *A footmark imprinted in superficial bruising on the back. There is another partial mark at waist level, both being made by stamping with the flat of a foot through thin clothing, rather than by swinging kicks.*

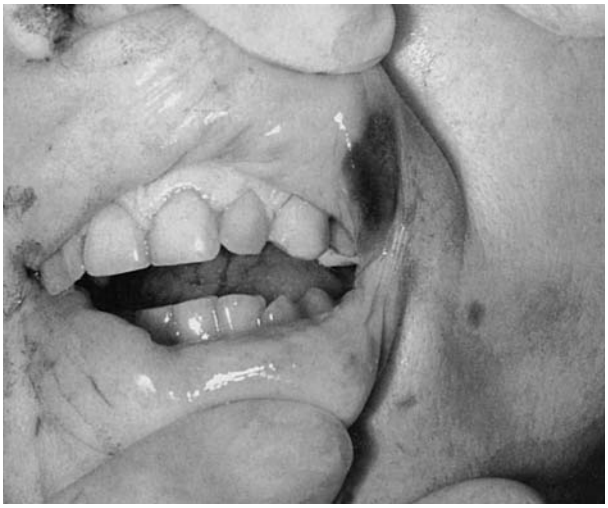

FIGURE 4.25 *Bruising of the interior of the lip from a blow in the mouth. The lip has been impacted on the underlying teeth, but not sufficiently hard to cause a laceration.*

required is great. As there is no internal pressure in the small vessels that have to be ruptured, bleeding is a passive ooze rather than an active extravasation, so the size of the haemorrhage is slight in relation to the amount of effort expended. Thus, post-mortem skin bruising is of little practical significance to the pathologist – in most instances he can be reassured that new or more obvious bruising apparent when he returns several days later for a second examination of a body is genuine. Post-mortem abrasions are common, especially after a first autopsy and additional handling of the body in the mortuary, but these procedures rarely produce artefactual bruises.

More important than dermal artefacts are spurious bruises inside the body. Especially in congestive deaths, such as pressure on the neck, the venous system may be engorged at death and a number of artefactual haemorrhages may occur. One particularly important area described fully elsewhere is the neck, in which collections of blood between the oesophagus and cervical spine may simulate bruising from manual strangulation. These were fully investigated by Prinsloo and Gordon (1951) and, where neck trauma is suspected, it is advisable to remove the brain before dissecting the neck region to allow drainage of the engorged venous plexuses that give rise to the artefact. This phenomenon appears to be related to the post-mortem interval and is more likely to arise if there is a delay in autopsy.

LACERATIONS

The third major type of blunt injury is the laceration, in which the full thickness of the skin is penetrated.

Lacerations differ from incised wounds in that the continuity of the tissues is disrupted by tearing rather than clean slicing, but the distinction is often blurred because some lacerations are caused by jagged projections ripping into the skin in much the same manner as a blunt knife or axe.

Unless great force is used, most lacerations require a firm base to act as an anvil for the skin and underlying tissue to be pinned against. It is unusual for a blunt impact to lacerate the abdomen or buttock, but the scalp, shins, shoulder, face and – to a lesser extent – the thorax, are all prone to lacerated injury. Where a soft area such as buttock, thigh, calf or forearm is lacerated, the lacerating agent is either a projecting point or edge, or a completely blunt object is pulled obliquely against the tension of the skin until it tears. Because of the crushing and tearing components of a laceration, there is usually associated abrasion and bruising,

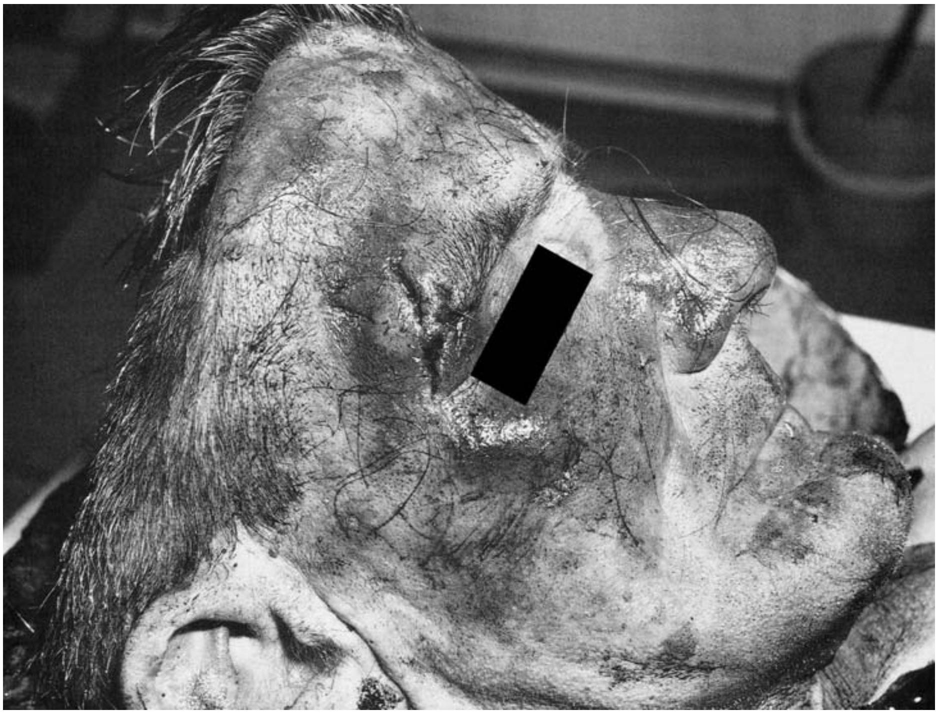

FIGURE 4.26 *Laceration of the eyebrow with surrounding bruising. The wound is stellate as a result of crushing against the underlying bone. The victim fell from a low cliff on to rocks. The sea has washed away the original bleeding.*

though these may be minimal if the lacerating force acts at right angles to the surface.

The scalp offers the best example of a laceration, as the skin and subcutaneous tissue lie over the bony platform of the skull and are crushed against it when the force impacts. In fact, scalp lacerations can resemble incised wounds, because this 'sandwich effect' is so pronounced that a blow from a blunt instrument can cleanly split the tissues against the skull, leaving a linear wound that appears to the inexperienced to be a knife or chopper cut.

A laceration can be distinguished from an incised wound by:

- the bruising and crushing of the margins, though this may be a very narrow zone, requiring inspection through a lens
- the persistence of tissue strands across the interior of the wound, including fascial bands, vessels and nerves. In a wound from a sharp weapon these are divided
- the absence of a sharply linear injury in the underlying bone, especially if it is the skull. A knife or axe is likely to chip or score the base of the wound
- if the area is covered with hair, as on the scalp, intact hairs will survive to cross the wound; an incised injury would divide them.

A flat impact may cause a laceration, especially on the scalp. If the head falls to the ground or is hit by a wide, flat weapon, the skin and subcutaneous tissues are compressed

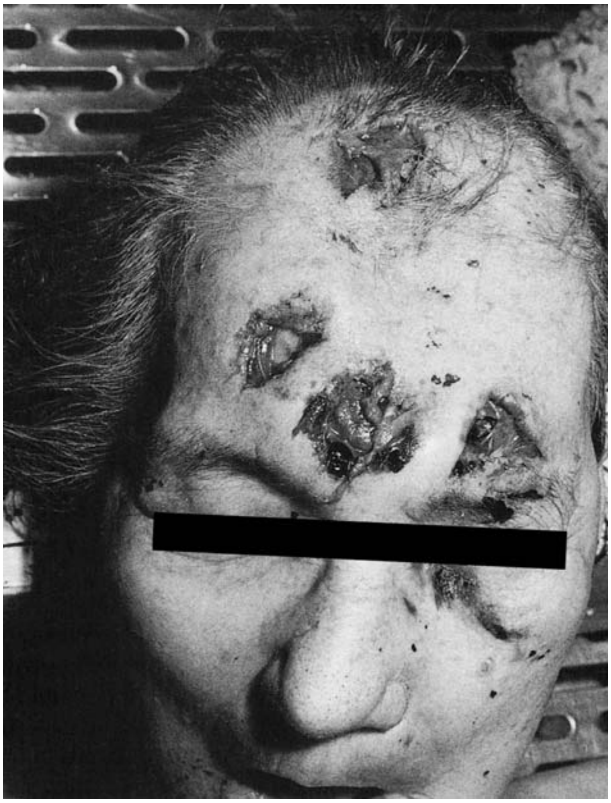

FIGURE 4.27 *Multiple homicidal lacerations of the scalp. The unusual shape is difficult to interpret, but they were caused by a claw hammer. Two are triangular, the upper and lower show tissue bridges because of the gap between the two claws.*

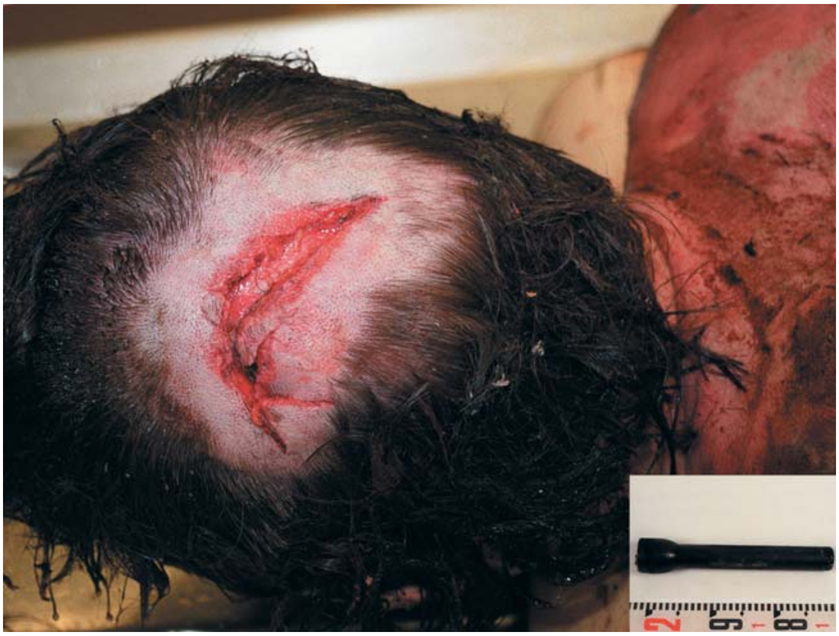

Figure 4.28 *Laceration of the scalp from a 30 cm long heavy torch (insert). The skin over the shoulders is reddish due to first- and second-degree burns. The victim was hit on the head while sleeping in bed, after which the husband poured petrol over the body and set it on fire.*

on the anvil of the skull and may split sideways, often with a zone of surrounding bruising.

Other lacerations are caused by the impact of an edged object, be this a rod, a brick, the edge of a kerbstone or the wing mirror of a vehicle. Though the impact may be virtually perpendicular to the skin surface, there is usually some tangential element and this can cause one edge of the wound to differ from the other. Where there is a glancing element, one edge will be relatively clean-cut, whereas the other will be ripped away from the first, suffering more tearing of the margin and often showing undercutting of the edge. If the impact is significantly asymmetrical, then this undercutting may be pronounced enough to form a flap of skin and underlying tissue, which is peeled off the underlying bone or deep fascia.

In extreme forms, this produces a ‘flaying injury’, where a large area of skin and subcutaneous tissue is rolled off a limb, almost always by the rotary action of a revolving motor wheel. A similar injury can occur on the scalp again from a rotating wheel – though a similar ‘scalping’ lesion can be caused by traction from hair being trapped in machinery. This was formerly a common industrial accident and was the reason for the ‘snood’ hairnet being worn by women factory workers during the last war.

Patterned lacerations

Lacerations do not reproduce the shape of the injuring agent nearly as well as do abrasions and intradermal bruises. As with

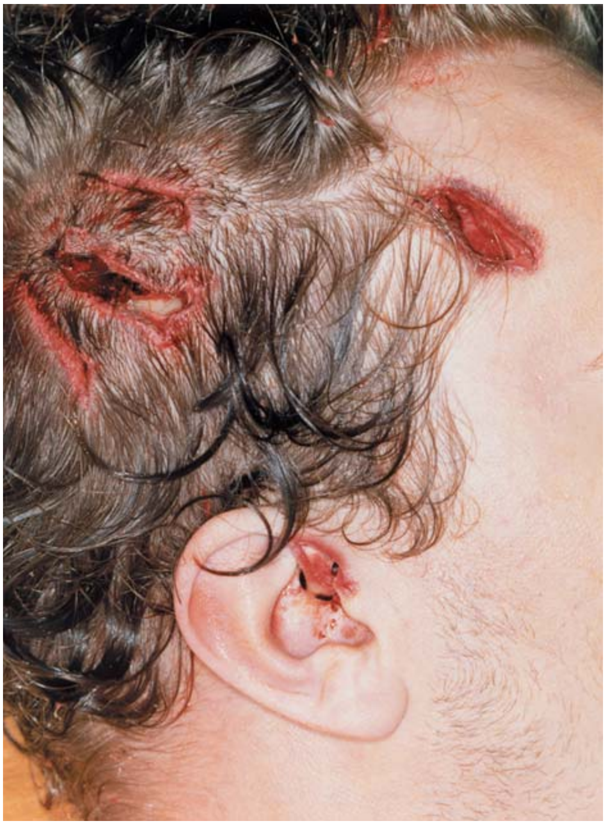

Figure 4.29 *Homicidal lacerations of the scalp penetrating the skull. Although some of the injuries resemble incised wounds, their margins are crushed, and hairs and tissue strands cross the injuries. The damage was probably inflicted with a metal rod with sharp edges.*

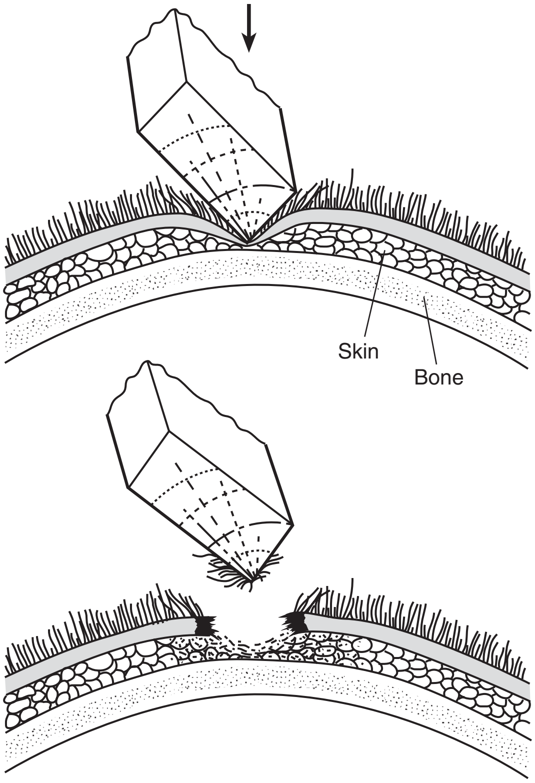

FIGURE 4.30 *Crushing impact of a blunt object on skin supported by bone, such as scalp on skull. The skin is sandwiched between weapon and bone and this causes a lacerated split that has bruised margins and bridges of hair and tissue in the wound.*

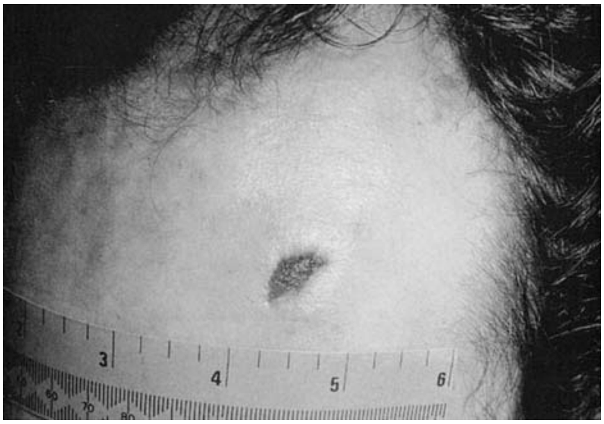

FIGURE 4.31 *A healing shallow laceration of the forehead, about one week after infliction. The base of the injury has filled with granulation tissue and the surrounding skin shows puckering due to contraction of the central scab.*

skull fractures, the magnitude of the force needed to cause a blunt tearing of skin leads to blurring of the impact shape, often with radiating or totally irregular tearing. Sometimes the shape is recognizable and probably the best example is a hammer blow to the head. A circular face may punch out a circle or an arc of a circle, which may also be reproduced in an underlying depressed skull fracture. The exact size of the hammer face may not be accurately reproduced in the laceration, which may be slightly larger than the hammer.

Kicks may cause lacerations, especially if a heavy boot with a hard toecap is used. These tend to be tangential and may peel up a flange of tissue. They can be clean-cut and are sometimes mistaken for knife wounds. Stamping injuries are more likely to cause abrasions and bruises than lacerations, unless the violence is extreme. A heavy focal blow may cause a linear or a stellate laceration, the latter radiating out from the focus of impact. The surface may be abraded and the surrounding tissue deeply bruised, so the final lesion is complex.

Firearm wounds are a special form of laceration and will be dealt with separately.

BLUNT PENETRATING INJURIES

These ‘puncture wounds’ are a hybrid between lacerations and incised wounds, and have more of the features of the latter. Such injuries can occur from metal spikes, wooden stakes, garden or farm forks, and a variety of vehicular and industrial accidents. They are occasionally seen in homicides or suicides when weapons like screwdrivers are used as daggers. The interpretation of such skin wounds can be difficult, as in the ‘Yorkshire Ripper’ cases, when some small stellate stab wounds were thought to have been caused by a Phillips-type screwdriver. A number of fatal wounds, some of them homicidal, have been caused by chisels, usually very sharp woodworking tools with one face of the edge ground to a straight line. The edges of the chisel are squared off and the resulting skin defect may be rectangular.

When a blunt spike is driven through the skin, there will be inversion and abrasion of the edges, though the former may vanish if the weapon has been withdrawn. Foreign material such as rust, dirt or splinters may be left in the wound, and must be carefully preserved for forensic examination if the weapon has not been recovered.

INCISED WOUNDS

Injuries caused by sharp objects are classified as ‘incised’ wounds, though the nomenclature is again somewhat confused. Some authors make these a subcategory of lacerations and, when the weapon is relatively blunt (as in some axe or spike injuries), there may be sufficient crushing and bruising of the wound margins to make differentiation between incised and lacerated wounds difficult.

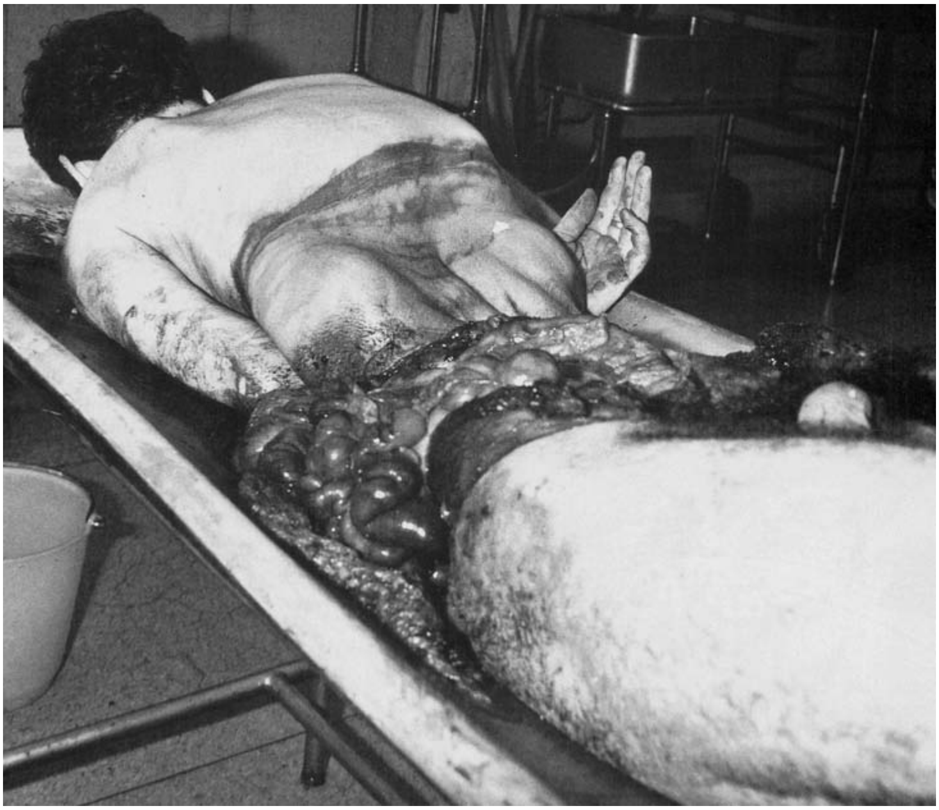

FIGURE 4.32 *Transection of a body into two halves at waist level. The victim jumped from a high-rise building and landed on a fence.*

The term 'incised wound' usually covers all types of injury from, for example, a knife, sword, razor, glass or sharp axe – though some writers would make stab wounds a separate category, reserving incised wounds for injuries that are longer than they are deep. These semantic variations are not important in themselves, but can lead to confusion in medical and autopsy reports, as well as in oral testimony if the doctor fails to clarify his meaning. The essential feature of all incised wounds is the clean division of the skin and underlying tissues so that the margins are almost free from any damage. Microscopically there is always a narrow zone of injury, but this is minimal compared with the tissue destruction caused by blunt injury. Healing is affected, as well as the extent of subsequent scarring.

Cuts or slashes

When an incised wound is longer than it is deep, it may be referred to as a 'cut' or 'slash', though a 'cut' may also be used by lay persons to describe a laceration. Such slashes when inflicted by knife, sword, razor, cleaver, parang, machete, panga, broken glass or bottle, are typical of a fight, when the assailant strikes out with a swiping action, rather than the thrust of a stabbing attack. They are common in gang fights and bar-room brawls and the weapon may be swung at arm's length in a horizontal arc so that if it contacts a body it will slice the skin and tissues as it passes by. They are also seen in suicidal injuries, usually on the wrist or throat. The characteristics of any given wound are a matter of a chance and relative positions. The slash may be deeper at the entry end or the exit point, though it seems true that many such injuries tend to dig in near the point of first contact and become progressively more shallow as the wound approaches the distal end. There may be a superficial 'tail', a shallow scratch, running out of the termination of the slash, as the weapon rises out of the wound and is drawn across the skin surface before leaving the body altogether.

There is a tendency among advocates to visualize a fight as a static confrontation and to attempt to reconstruct events as if the two participants were standing still, the assailant merely moving his arm. This is an unrealistic interpretation, as all fights are dynamic, with constant relative movement of the bodies and limbs of both parties. Therefore any relative angles and depths of contact between the victim and the weapon can occur and overinterpretation and simplification should be avoided.

Slashed wounds are less dangerous than stabs, as the relative shallowness of the wounds is less likely to affect vital organs – especially as the arms and face are the common targets. The rib cage and skull protects the chest organs and

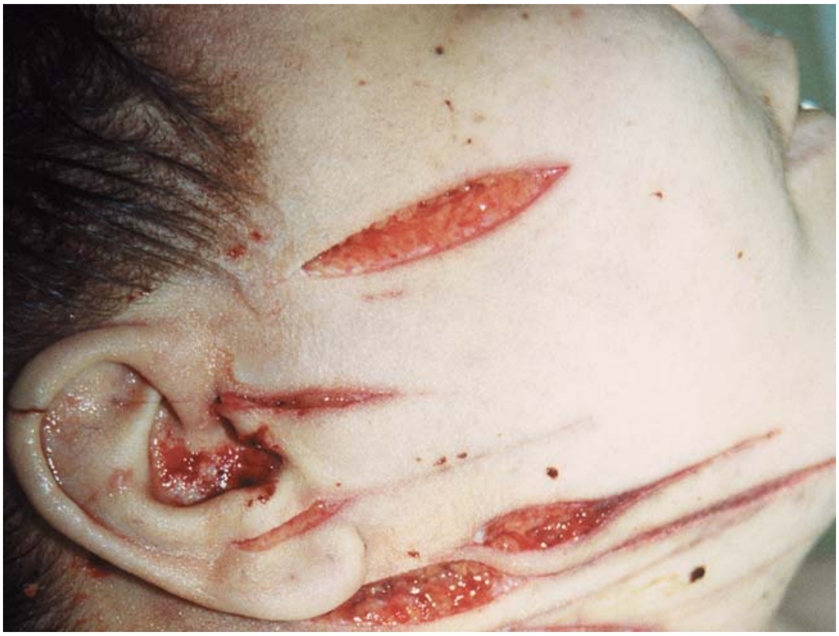

Figure 4.33 *Homicidal slashed wounds caused by knife; the length is greater than the depth, unlike stab wounds. The long tails are due to the knife rising from the skin, thus indicating the direction of the slash.*

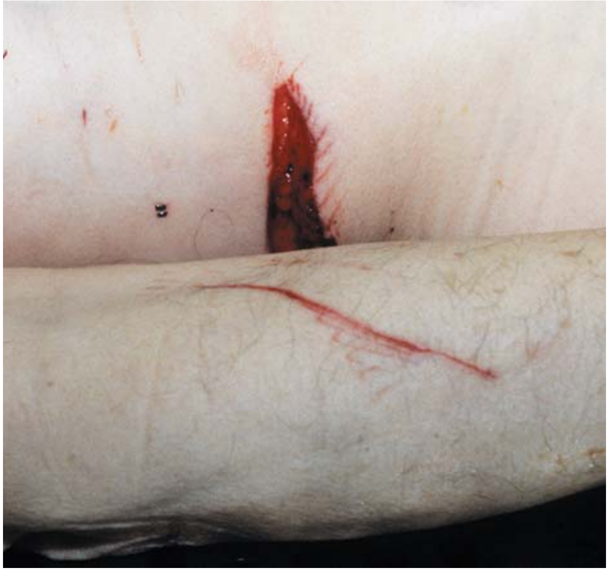

Figure 4.34 *A knife slash of the back showing regular scratches along the margin. This was inflicted with a 'Rambo' knife, which has deep serrations along the back edge that have somehow marked the skin on withdrawal.*

brain and the more vulnerable abdomen is seldom the target for a swung knife. Wounds of the neck are the most dangerous and are discussed below. Bleeding is the most serious complication of any slashed wound, though it will be external and more amenable to immediate treatment than the hidden internal bleeding of a stab wound.

A particular form of slashed wound arises from 'glassing', that is, the use of a broken drinking glass or bottle as a weapon. The glass may sometimes be broken prior to the attack, to produce edges that are exquisitely sharp. There may also be spikes of glass, which can inflict deep, almost stab wounds. Where, more commonly, a glass or bottle is used as a blunt instrument, it may shatter on impact with a victim's head, and the resulting injuries may both be blunt and incised.

Stab wounds and penetrating injuries

These are of major importance in forensic pathology, as they are extremely common in homicide. In Britain, stabbing with a knife is easily the most frequent mode of murder and manslaughter, primarily in domestic disputes and street violence. A stab wound is an incised wound that is deeper than it is wide. Though many penetrate only the skin and subcutaneous tissues, those seen by forensic pathologists are most often fatal, having entered vital deep structures. Several aspects of this important topic need discussion.

THE NATURE OF STABBING WEAPONS

Knives are the weapons most frequently involved and their physical characteristics are important in shaping the wound. A pathologist called upon to examine a knife that may have been used in an assault should note and measure:

- the length, width and thickness of the blade
- whether it is single or double-edged
- the degree of taper from tip to hilt
- the nature of the back edge in a single-edged knife (for example, serrated or squared-off)

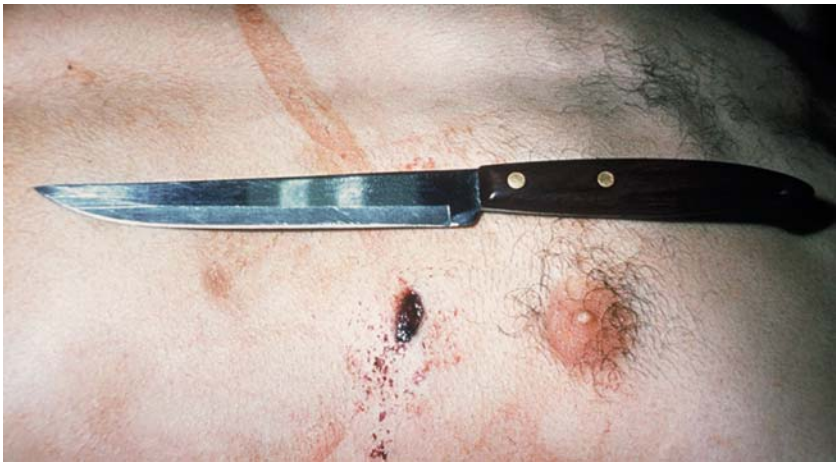

FIGURE 4.35 *A stab wound and the inflicting knife. The wound is slightly shorter than the width of the blade at the depth of penetration because of sideways gaping and the contractile elasticity of the skin.*

- the face of the hilt guard adjacent to the blade
- any grooving, serration or forking of the blade
- most importantly, the sharpness of the edge and especially of the extreme tip of the blade.

Other knife-like weapons include scissors, chisels, swords, open razors and sharp tools of many kinds, including deliberately modified tools such as sharpened screwdrivers. Larger cutting instruments such as axes, choppers, parangs, machetes and pangas and agricultural implements such as shears, bill-hooks, hay-knives and many others have been used as stabbing weapons when used by the point rather the long edge.

Spiked instruments come in many forms, from icepicks to hay-forks, from case-openers to fire-irons. The appearances of their wounds vary with the physical nature of the implement. Some weapons are exquisitely sharp and the wounds made by them exhibit extremely fine division of the tissues. To razors and razorblades may be added surgical and craft knives, often with a disposable blade, together with carpet and general utility and hobby tools, such as the 'Stanley' knife. Equally sharp is broken glass and sometimes porcelain. Both sheet glass and smashed glass utensils can provide edges that equal or exceed surgical scalpels in their cutting ability.

A common weapon in a bar-room fight is a smashed beer tankard, which when held by the intact handle provides a formidable weapon for both slashing and stabbing. Smashed china, such as cups and mugs, can also provide sharp cutting edges. The external glaze can project, if the fracture line is oblique, producing a glass-like edge. From personal experiment, the author (BK) has shown that broken china can easily slice through the full thickness of skin and thus validate a defence that injuries were the result of falling on a broken mug.

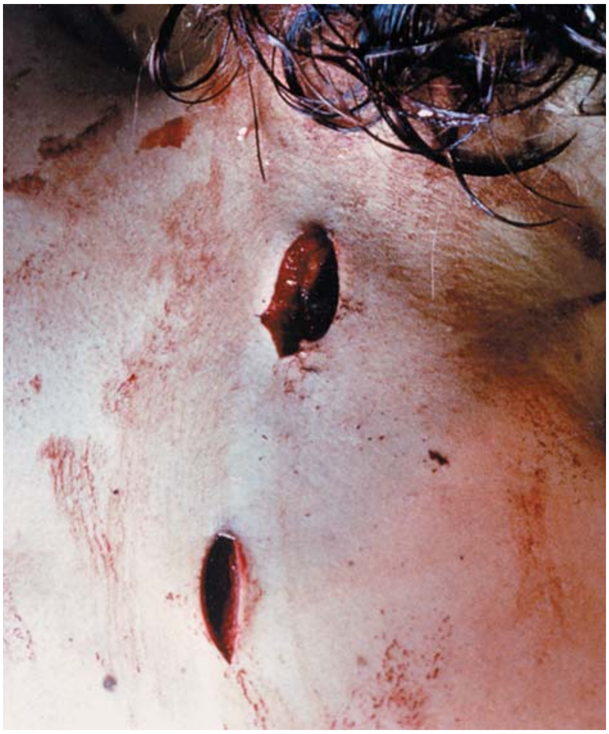

FIGURE 4.36 *Two knife wounds in the back. The wounds have a sharper lower edge compared to a rounder upper edge, due to a one-edged blade.*

THE CHARACTERISTICS OF A STAB WOUND

The surface and internal appearances of a stab wound allow the pathologist to offer an opinion upon:

- the dimensions of the weapon
- the type of weapon
- the taper of the blade

- movement of the knife in the wound
- the depth of the thrust
- the direction of the thrust
- the amount of force used.

Dimensions of the weapon

The dimensions of the knife may be a vital part of the investigation of a homicide when the weapon has been removed from the scene by the assailant. The pathologist can sometimes assist the investigators by telling them, within limits, what size and type of knife to seek. When suspect weapons are recovered, his inspection of them may indicate whether or not they are consistent with having caused the wounds in question. As always in forensic medicine, however, caution must be employed and the cardinal sin of overinterpretation avoided.

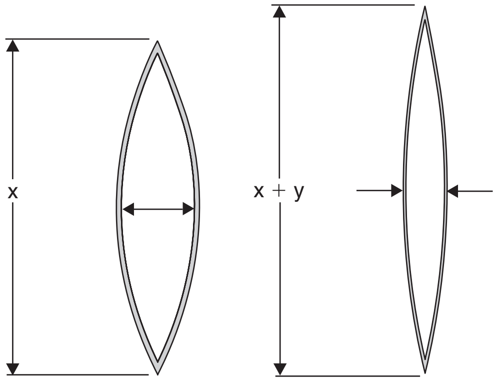

FIGURE 4.37 *A stab wound gapes across its width and shortens in length, especially when across skin or muscle planes.*

The length of the wound should be measured to the nearest millimetre as it lies undisturbed on the skin. In most instances the wound will have gaped across the centre, to form a long ellipse. The extent of gaping will depend on the anatomical situation, for example, over joints or in the axilla or groin, and whether the axis of the stab is in line with or across the tension of Langer's lines or underlying muscle bundles.

When the edges are gently opposed, the length of the wound may then extend slightly and should be measured again in this position, which more accurately approximates to the length when the blade was *in situ*.

An important additional factor, however, must always be taken into account because, when the knife is withdrawn the elasticity of the skin causes it to retract, making the wound smaller than when the blade was *in situ*. This mechanism may be amplified by contraction of the underlying muscles, which, if at right angles to the axis of the wound, may cause the wound to shorten from end to end, while gaping more widely at the centre. Using wound measurements to predict the size of the knife blade, one must take into account the other matters discussed below, such as movement of the blade in the wound, the taper of the blade and the depth of penetration.

Type of weapon

The type of blade usually refers to whether it had a single or double cutting edge. Most knifes have a single sharp edge, the back edge being blunt or otherwise machined. A few dagger-like weapons have both edges ground to sharpness,

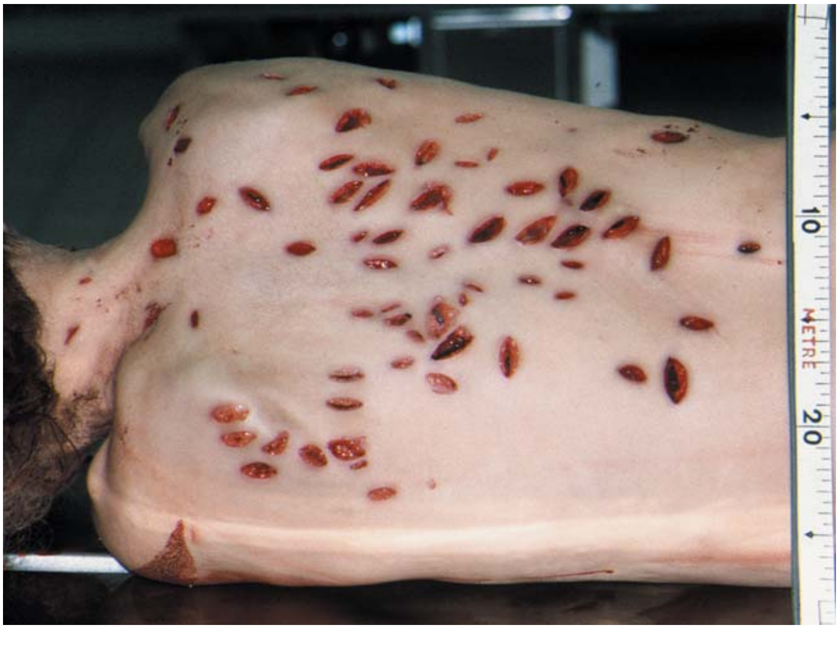

FIGURE 4.38 *Multiple stabs on the back from the same knife, showing differing shapes and sizes.*

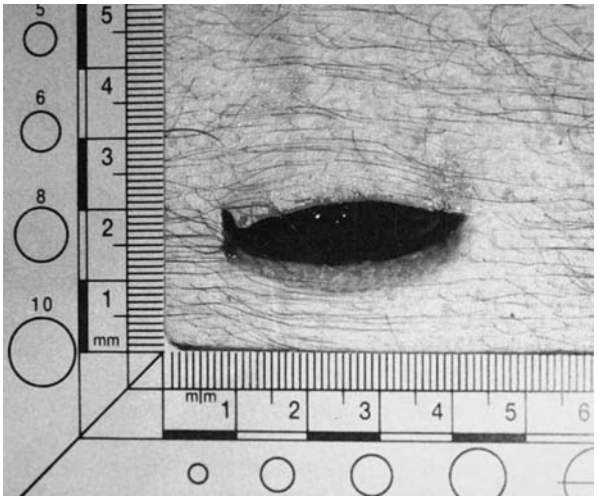

FIGURE 4.39 *A stab wound showing a unilateral 'fish-tail' split caused by the blunt back edge of the knife blade. This is sometimes bilateral due to tearing of the tissues. The other end of the wound is pointed due to the cutting edge of the blade.*

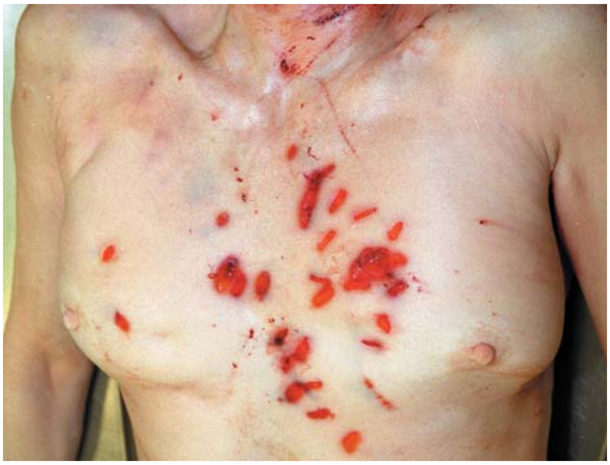

FIGURE 4.40 *Multiple stab wounds from the same knife, showing variations in the size and shape of the injuries. With multiple injuries, it is advisable to number each to facilitate easy reference in the autopsy report and when giving evidence in court.*

some have only the distal part of the back of the blade sharpened. On inspecting the skin wound, the pathologist frequently finds that both ends of the defect appear sharply cut, coming to a fine 'V-point' at the extremities. Unfortunately, this does not necessarily indicate that a knife with two sharp edges was used, as the skin often splits behind the blunt edge to produce a symmetrical appearance.

If, however, there is an obvious difference between one sharply pointed end and a rounded or even square-cut opposite extremity, then it can be said with some confidence that a single-edged knife was employed.

A knife with a thick blade that is squared off to a flat surface on the edge opposite the cutting edge, may sometimes be identified from the wound, where a definite double right-angled end to the skin defect is apparent. The skin may split back slightly from each of these corners, forming the so-called 'fish-tail' appearance.

Where a knife with a serrated back edge has been used, such as is found on some boy scout or other Bowie-type knives (or the recently fashionable 'Rambo' weapons), the back edge of the wound may be torn or ragged and – on some occasions where the knife has entered very obliquely – serrated abrasions may be seen on the skin adjacent to the end of the wound.

If the knife has been plunged into the full extent of the blade, then there may be a hilt bruise or abrasion on the skin surrounding the wound. This may even reproduce the pattern of the guard at the base of the hilt, if there is one. Careful photography, sketches and measurements are needed to capture the exact outline of such a mark, which are, admittedly, unusual.

Taper of the blade

The taper of the blade is naturally related to wound size. If a tapered blade is inserted to 4 cm, then the length of the wound (ignoring skin retraction for the moment) will be that of the width of the blade at that level. If the taper continues to the 8 cm level, then the wound will be correspondingly longer. Only when the blade edges become parallel will the wound size remain constant for further penetration of the knife.

Movement of the knife in the wound

Movement of the knife in the wound can result in loss of evidence as to the size of the blade, but can also add other information. Where a knife is stabbed directly into the body and withdrawn along the same track then, within the limits discussed above, the size of the wound will indicate the minimum width of the blade at the maximum depth of penetration.

If, however, the knife is 'rocked' in the wound, the skin defect will be enlarged, sometimes considerably. The term 'rocking' means a leverage or angulation in the plane of the wound, so that the cutting edge extends the incision. This rocking can occur either by the knife being actively moved in the wound by the assailant – or by the body moving relative to the knife – or by a combination of the two.

It is sometimes a misapprehension among both doctors and (especially) lawyers, that stabbings and other assaults occur in static circumstances, with the victim standing in the anatomical position. Directions of wounds are all too often interpreted as if the only movement came from the hand of the assailant, whereas in most fatal assaults the

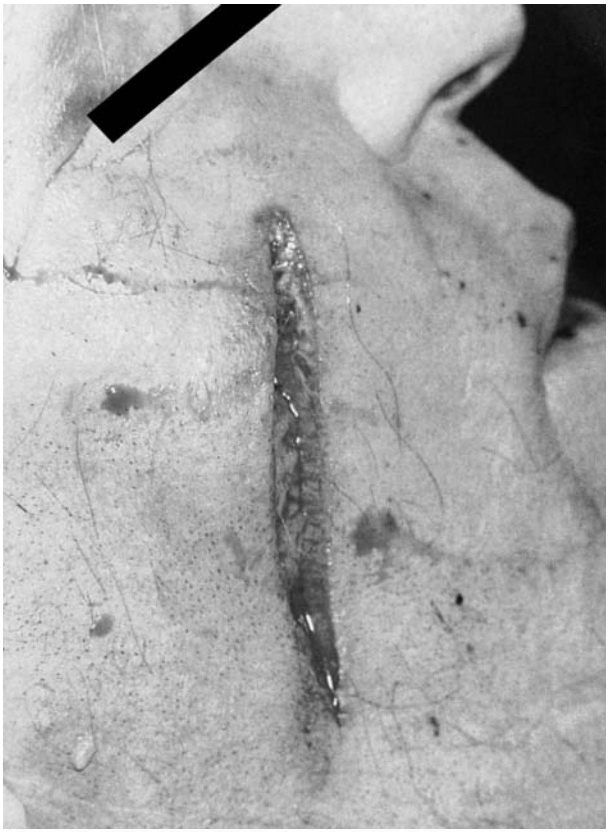

FIGURE 4.41 *A slashed facial wound from a knife, showing undercutting or 'shelving' on the edge nearest the mouth, indicating the angle of the knife relative to the surface. The tailed nature of the lower end also indicates that that knife was drawn downwards across the face.*

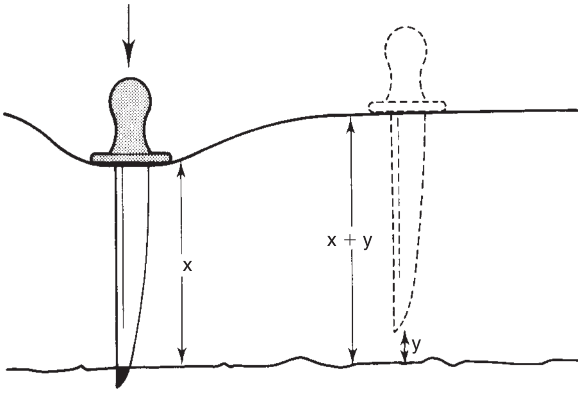

FIGURE 4.42 *Forcible stabbing can indent the body surface so that deep structures can be injured that appear to be beyond the reach of the knife.*

victim bends, turns and twists, and the assailant often follows with similar gyrations. Thus a knife that is thrust into the body may itself move or the body may move against it before it is withdrawn.

This relative movement may extend the wound in a linear fashion, as described above or there may be twisting of the blade in the wound. In the latter case, the resulting skin incision may be V-shaped or totally irregular. Rarely, a single stab may produce multiple skin wounds, such as a tangential stab of the arm, which passes through the superficial tissues and then re-enters the chest wall. Another instance is a stab through a female breast where, especially if the breast is sagging, the knife can pass right through the edge before re-entering the thorax.

Depth of the thrust

The depth of a stab wound may be important in attempting to assess the length of a missing weapon. Again there are pitfalls to be avoided. First, if a knife is driven in up to the hilt, the depth of the wound as measured at autopsy may be greater than the true length of the blade. This is common in the abdomen and to a lesser extent in the chest, because the impact of a forceful stab may momentarily indent the abdominal or chest wall so that the tip of the knife penetrates tissues that apparently should have been out of reach. This is particularly so when the hilt guard of the weapon impinges on the skin. For example, the author (BK) has seen the tip of a broken knife embedded in a thoracic spine, where the distance from that tip to the anterior skin wound was less than the length of the blade when recovered.

As with direction, discussed below, the pathologist must also allow for differences in the position of internal organs as measured at autopsy in a supine body, compared with their position in the living body, often in the upright posture. Such variations in the *in vivo* distance from the skin wound to the organs penetrated must be taken into account when assessing the depth of a wound. In the chest and abdomen, visceral relationships vary with the stage of respiration. The lower ribs move upwards and laterally, and increase the distance between the skin and deeper structures.

Direction of the thrust

The direction of a stab wound is often a matter of contention in homicides, especially as trial lawyers (and their medical advisers) sometimes tend to overinterpret the facts when trying to reconstruct the scene of the fatal assault.

As mentioned above, a stabbing incident is often moving and dynamic, and the victim is rarely in a static, anatomical position. For example, a wound that enters the upper part of the left side of the chest and travels steeply downwards, is not inevitably the result of the killer being taller or situated above the victim, using a downward blow. The victim could equally well have been bending or crouching, the knife then entering on a horizontal plane relative to the floor. All that the pathologist can do is to determine the direction of the wound relative to the axis of the body – it is then a

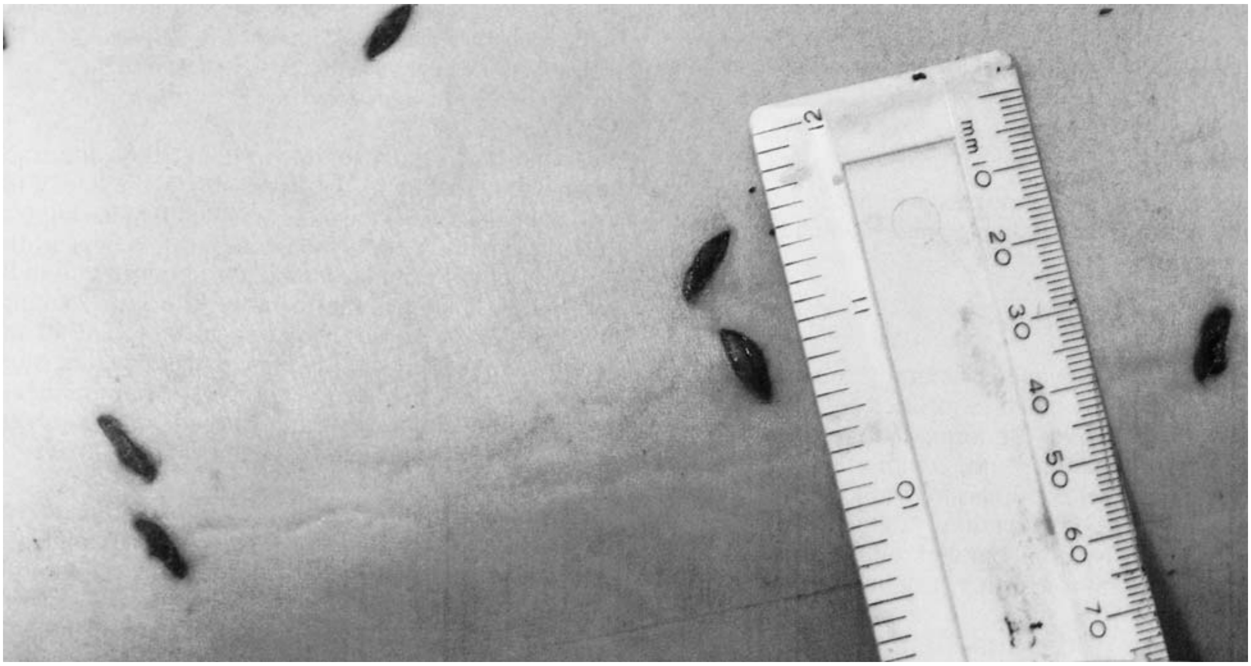

FIGURE 4.43 *Multiple stab wounds from the same knife. Although those at the bottom left are obviously square-ended, the more central ones are almost equally sharp at the extremities – illustrating the dangers of being too dogmatic about describing a missing knife to the investigators.*

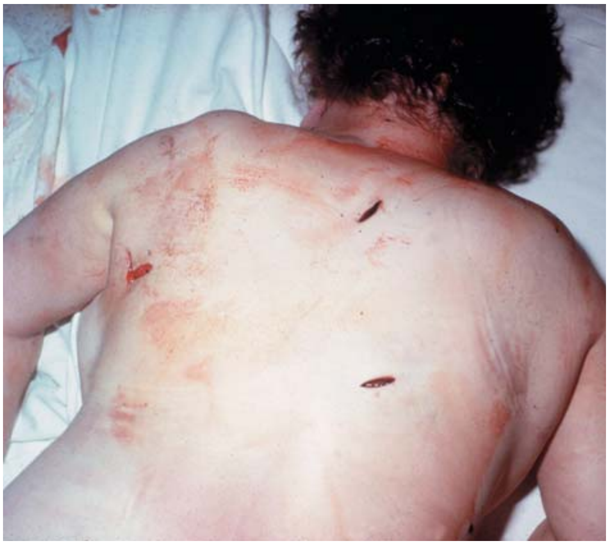

FIGURE 4.44 *Homicidal stab wounds misdiagnosed by an attending doctor as 'haematemesis or epistaxis'. The lower right wound clearly shows evidence of a single-edged blade. The wound in the left axilla is V-shaped because of the twisting knife.*

matter of non-medical evidence to relate that to the posture of the victim when struck.

Determination of the direction relies on both the appearance of the skin wound and the track in the deep tissues. Where a knife penetrates the skin with the plane of the blade at appreciably less than a right angle to the surface, the wound is often 'undercut', being shelved so that subcutaneous tissue is visible below one edge of the wound. Where the knife is plunged in obliquely, but the plane of the blade remains perpendicular to the surface, no such shelving can be seen at the sides of the wound though it may just be visible at the end of the wound.

More information about direction comes from careful examination of the track of the stab wound. This is an anatomical exercise during autopsy, the layers of tissue being examined in sequence from the surface downwards and damage to deep structures and organs compared with the position of the surface wound. For example, a stab wound of the chest may lie below the left nipple. Dissection of the subcutaneous tissues and muscles may reveal that the defect in the chest wall lies in the sixth intercostal space and that, inside the thorax, the right ventricle of the heart has been transfixed, and that there is a wound in the diaphragm that terminates in the liver just to the right of the midline. This information obviously indicates that the direction of the wound was downwards and from left to right. Similar deductions can be made for all other sites, but caution must be exercised, as mentioned earlier, in allowing for the change in position of organs in the supine post-mortem posture compared with the erect, living state. A thorough knowledge of surface anatomy is required, especially of the thorax, as well

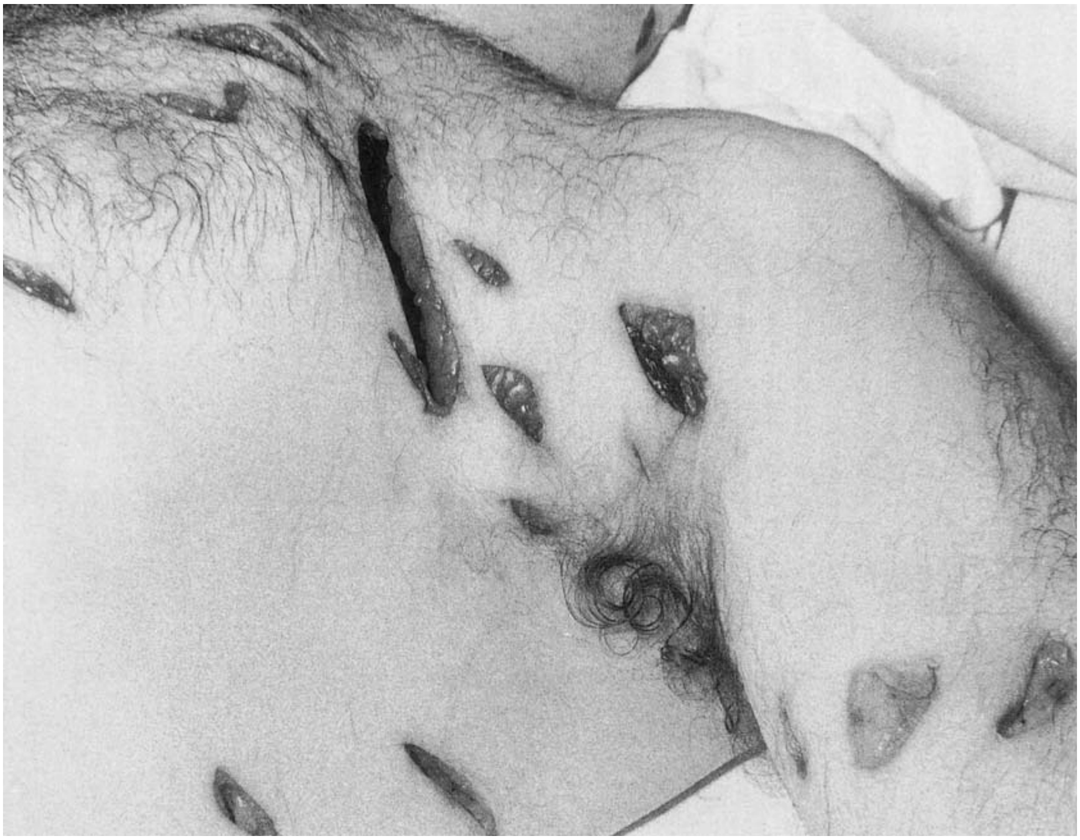

FIGURE 4.45 *Multiple homicidal knife wounds, all inflicted with the same weapon. This shows the marked variation in wound size from the same knife, caused by rocking and twisting movements of either weapon or victim.*

as diagrams of the position of internal organs mapped out in relation to surface markers; samples of such diagrams are given in this book.

Attempts have been made to delineate the track of a stab wound before dissection either by filling the defect with a radio-opaque fluid before taking X-rays, or by filling with a plastic or even metallic substance that will harden to form a cast. In practice, these methods appear to have little advantage over careful dissection. When radio-opaque liquids or pastes are used, there is often leakage from the stab wound into the thoracic or abdominal cavities, leading to a confusing radiological picture. Recently, magnetic resonance imaging (MRI) has been used to visualize the wound track, but such facilities are rarely available for dead victims.

Estimation of the degree of force used in stabbing

The amount of force required to inflict any given stab wound is often a matter of extensive debate in criminal trials. The prosecution naturally gains from showing that a stabbing was inflicted with 'considerable force', sometimes even using pejorative terms such as 'violent' or 'frenzied'. This tends to confirm the intention to stab, whereas the defence proposition is often that the victim inadvertently fell or was pressed against a weapon held passively by the accused. An expert medical witness has difficulty in replying to the almost inevitable question by counsel: 'What amount of force was necessary to cause this wound, doctor?'

Assessment of force is subjective and cannot be quantified in any satisfactory way that is meaningful to the court. Physical units such as dynes/cm^2, even if they can be measured, mean nothing to a judge and jury. The only measure that the medical witness can offer is a broad grading, based on common sense knowledge, such as 'slight pressure', 'moderate force', 'considerable force' and perhaps 'extreme force', the latter being reserved for exceptional cases where the knife has impacted in dense bone, penetrated the skull, or left a marked bruised impression of the hilt guard on the skin. A number of investigators, including the author (BK), have carried out measurements on the force needed to inflict stab wounds, both in animal, cadaver and surrogate tissues. The usual method is to attach some type of dynamometer or transducer to a knife, which records the maximum force applied during penetration, preferably with the duration of impact.

From this work, the following useful generalizations can be made.

- Apart from bone or calcified cartilage, the tissue most resistant to knife penetration is the skin (Knight 1975), followed by muscle where large muscle bundles underlay the fascia.
- The sharpness of the extreme tip of the knife is the most important factor in skin penetration. The cutting edge of the knife, once the tip has penetrated, is of relatively minor importance.
- The speed of approach of the knife is particularly important in achieving penetration. A knife held against the skin, then steadily pushed, requires far more force to penetrate than the same knife launched against the skin like a dart. As in blunt injury, this is an example of the physical law that requires that the force varies directly not with the mass of the weapon, but also with the square of the velocity.
- Stretched skin is easier to penetrate than lax skin. The chest wall, where skin tends to be intermittently supported by underlying ribs, is relatively easy to puncture with a sharp knife as the skin and tissues are stretched over intercostal spaces in the manner of a drum membrane.
- Though the thick skin of the palms and soles is much tougher than on the rest of the body, the variation in resistance of the rest of the skin to a sharp knife is of little importance compared with other factors. Similarly, the skin of the aged, or of women, is not appreciably less resistant to a sharp point than that of men or young persons.
- When a knife-point impacts against skin, the latter dimples and resists until penetration suddenly occurs. The tension developed in the stretched skin appears to act as an ‘elastic reservoir’ and, when the threshold of resistance is exceeded, the knife ‘falls’ through the subcutaneous tissues without any further force being imparted to it unless impeded by bone or cartilage. Thus no additional effort is needed by the assailant to achieve deep or even full penetration up to the hilt. It is sometimes incorrectly claimed by the prosecution that a deep stab wound must imply extreme force or continued pushing after penetration. This is not so and experiments have shown that once penetration occurs, it is difficult or even impossible to prevent deep penetration because of the suddenness of the breakthrough.
- When the knife penetrates the skin rapidly, for example, if the body falls or runs on to the blade, the knife does not need to be held rigidly in order to prevent it being pushed backwards. Its inertia, if the tip is sharp, is quite sufficient to hold it in place while the body spears itself on the blade.

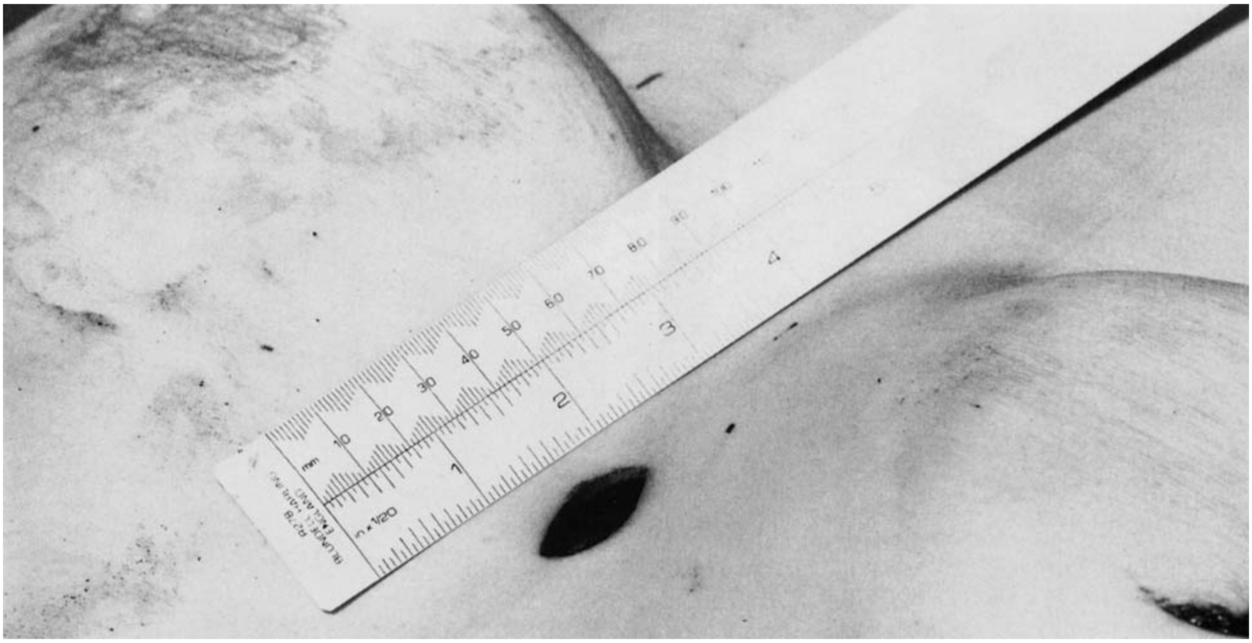

FIGURE 4.46 *A stab wound of the centre of the chest. Although each end of the wound is identical, the weapon was a single-edged kitchen knife. The end of the wound adjacent to the back of the blade has split making it impossible to say if a single- or double-edged knife was used. The sternum beneath was penetrated to reach the heart, so considerable force must have been used.*

It has been wrongly argued against a defence of falling or running onto a knife, that the hilt would have had to have been supported against, for example, the hip of the person holding the knife to immobilize it against the advancing body. Experiments have shown that this theory is invalid.

Uncalcified cartilage, especially that in the costal cartilages of young and middle-aged persons, is easily penetrated by a sharp knife, though naturally more force is required than if the blade passed through an intercostal space. Calcified rib and bone provide a much more resistant barrier, but a forceful stab from a strong, sharp knife can easily penetrate rib, sternum or skull. Firm tissues like myocardium, liver and kidney are easily traversed by all but the most blunt of weapons, and their resistance is far less than that of cartilage or skin.

INJURIES BY WEAPONS OTHER THAN KNIVES

The general features of injuries from knives apply equally to other sharp objects that cause incised and stab wounds. Razors and broken glass have extremely sharp edges so that, when applied tangentially or at a small angle to the skin, undercutting may be a marked feature. The wound may appear as a shallow slice which bleeds profusely and, in a hairy area such as the scalp, it will reveal cut hair bulbs on the shelved surface. Deliberate wounds with razors or sharp knives may exhibit patterns or even words; these may occur in gang fights or sadistic homicides.

Glass can be employed as a cutting weapon, again typically in bar fights or gang vendettas, where a broken bottle may be beld by the neck or a smashed beer tankard by the handle.

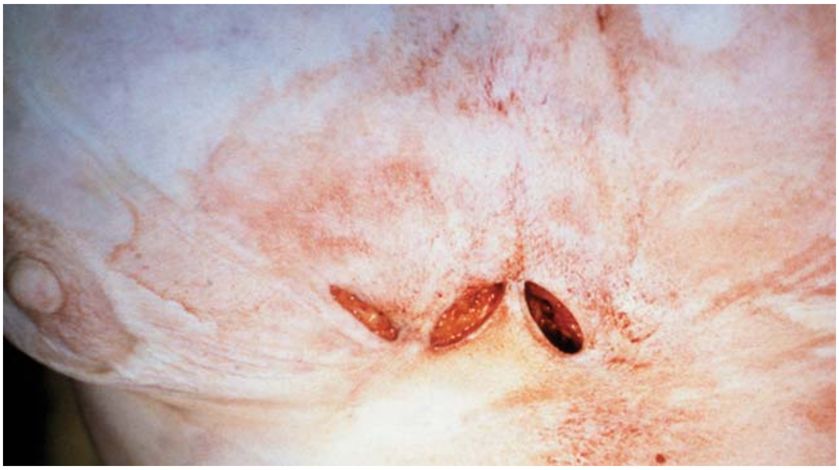

FIGURE 4.47 *Three wounds from a single stab with a knife, which was in place when the body was discovered. The knife had entered obliquely through the inner side of the right breast, emerged into the cleavage and re-entered the mid-line. If the knife had not been in situ, interpretation could have been more difficult.*

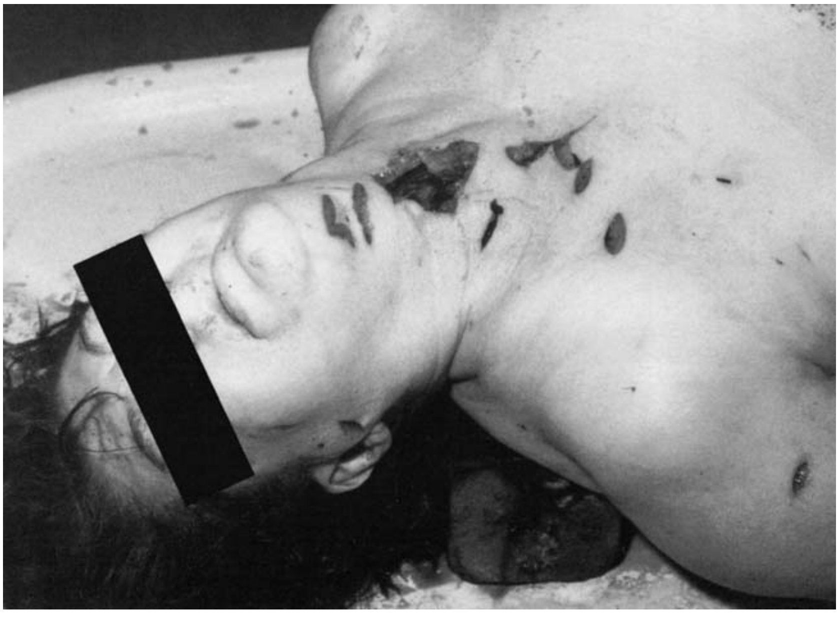

FIGURE 4.48 *Homicidal stab wounds of the throat and head showing the pleomorphism of injuries caused by the same knife. The wound under the chin is V-shaped caused by twisting of the weapon; the large throat wound consists of several superimposed thrusts and over the manubrium there is a 'tail' caused by the tip of the knife slipping out of the wound on withdrawal.*

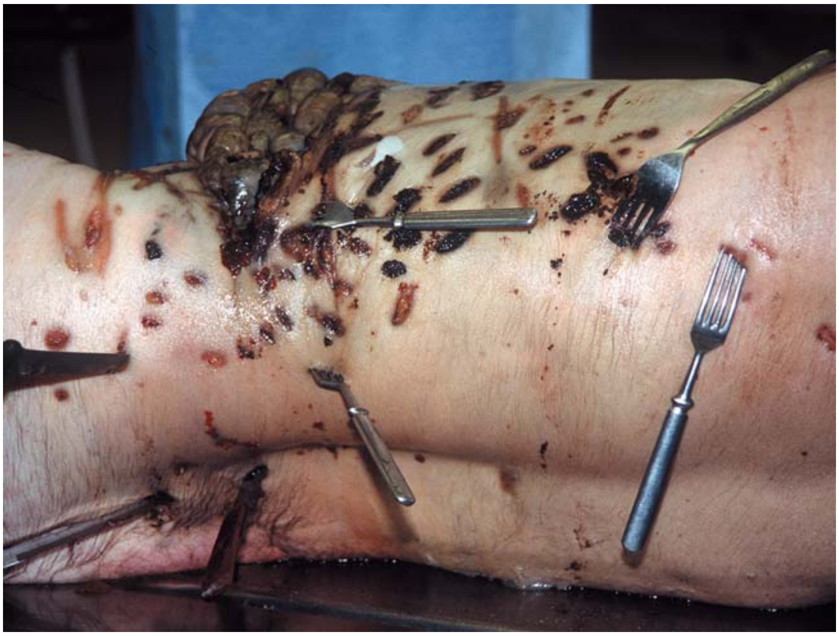

FIGURE 4.49 *Multiple homicidal stab and incised wounds inflicted with various kitchen utensils.*

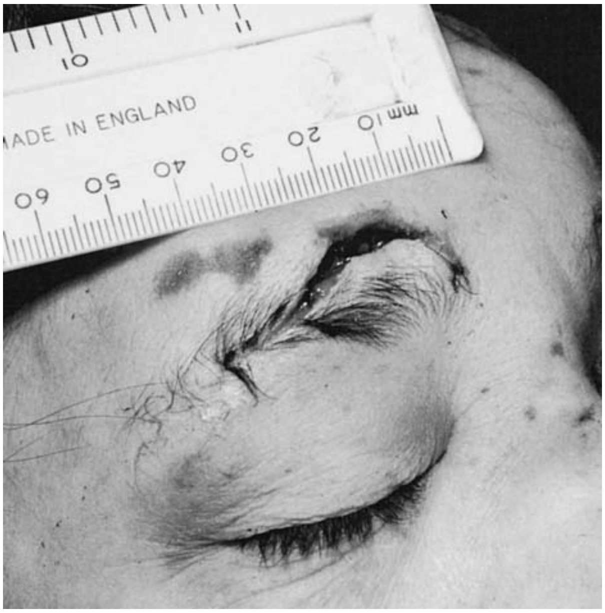

FIGURE 4.50 *A wound over the eyebrow caused by impact from a broken drinking glass. Although this location is commonly the site of laceration from blunt blows, this injury is sharp-edged and has unbruised margins, indicating that it is an incised wound. There is an associated abrasion from non-cutting contact and there is bleeding into the upper eyelid, probably from gravitational seepage from the wound.*

Official estimates in Britain suggest that between 3400 and 5400 offences occur annually in which glass is used as a weapon. Surveys in London and Bristol by Shepherd *et al.* (1990a) and Hocking (1989) showed that bar drinking glasses were the most common weapons, and that 75 per cent of these were from straight-sided pint glasses. Unexpectedly, this research showed that most of the glasses were intact until thrown or thrust at the victim – they had not been smashed prior to the assaults. Most of these non-fatal injuries were to the face. The fragility of these glasses varies greatly from manufacturer to manufacturer, so that there can be a sixfold difference in their propensity to smash on impact.

Though these formidable weapons are usually used for slashing, long slivers of glass can act – sometimes inadvertently – as stabbing agents. The sliver may break off and remain in a deep wound, which may be overlooked if the entry is small.

As many casualty surgeons have learned to their cost – or to the cost of their medical protection society or medical insurers – glass is almost invariably radio-opaque, even though this is either unknown or even denied by the doctor. Numerous civil actions for negligence have occurred – though usually in non-fatal circumstances – but, on occasion, the first person to find a deep fragment of glass in a vital position has been a forensic pathologist.

As mentioned earlier, broken china and porcelain, as from a smashed cup, mug or plate, may have a sharp edge on the exposed glaze and can easily cause a slashed wound.

INJURIES FROM SCISSORS

Stabbing by a pair of scissors is not uncommon and is seen most often in domestic circumstances where a woman uses a weapon upon her husband or consort which is both familiar and near to hand. The appearances of the wound will

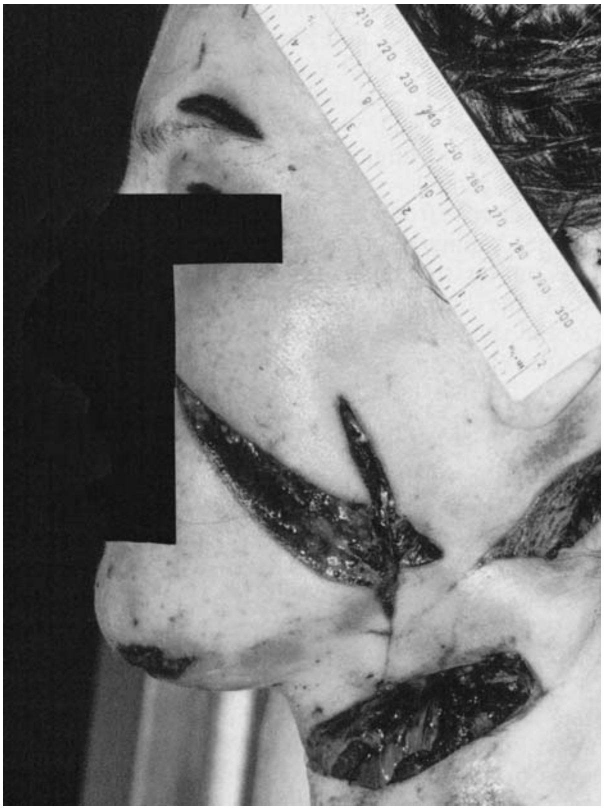

FIGURE 4.51 *Incised wounds, which in spite of their large size are typical of those inflicted by a sharp-edged weapon rather than a crushing laceration. The instrument was a butcher's cleaver.*

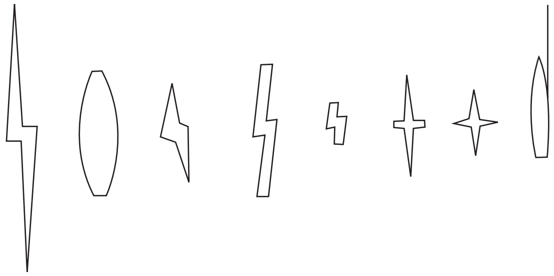

FIGURE 4.52 *Various profiles made by scissors stabbed through the skin. The cross shapes are caused by blade screws or rivets; the one on the right is a full penetration of one blade of an open pair of scissors, the other blade impinging flat on the skin.*

differ according to whether the scissors were used open or closed. If open and one blade is stabbed into the victim, then the appearances will be virtually indistinguishable from a knife wound. Some modern scissors, however, have a two-part blade in which a steel cutting edge is riveted to a carrier continuous with the handles. This may produce a stepped wound that may show a notch or variation in linearity in the margin of the stab wound.

More characteristic is the wound produced when both blades of a closed pair of scissors are plunged through the skin. Some scissors have a double-sharp point, but most domestic, tailors' or dressmaking scissors are fairly blunt at the tips when closed, so considerable force is needed to drive them through the skin – tending to work against a defence of an accidental movement against a passively held weapon.

The skin wound made by closed scissors is typically shaped like a flat 'Z' or the usual impressionist sign for a flash of lightning. The offset of the two blades gives this shallow zig-zag pattern, which is unmistakable when present. Some scissors, however, especially long narrow ones in which the blades close almost completely over one another, may not reproduce this appearance and the wound may look more like that from a thick knife. Sometimes, there are small lateral splits in the wound centre from a projecting hinge screw.

DEFENCE WOUNDS

In assaults of any kind, the natural reaction of the victims is to protect themselves. The limbs used for protection can themselves be injured and these defence wounds may be of considerable medico-legal significance, as they indicate that the victim was conscious, at least partly mobile and not taken completely by surprise. Defence wounds can be sustained from attacks by fists, feet and blunt or sharp instruments. The classic position for them is on the forearms and hands, which are instinctively raised to protect the eyes, face and head. Other defence injuries may be inflicted on the thighs, when attempts are made to shield the genitals.

With attacks from blunt instruments or fists, bruises are the hallmark of defence attempts. They are common on the outer sides of the forearms, the wrists, the backs of the hands and knuckles. The size and shape of the bruises naturally depends upon the attacking object. Abrasions may accompany the bruises but the latter are more common. Fractures of the carpal bones, metacarpals and digits may occur. If the arm is held in front of the face, the inner (palmar) side of the forearm may also receive blows. There may also be defence bruises and abrasions on the thigh, as blows or kicks aimed at the lower part of the body cause the victim, especially men, to protect the genitals. The leg may be brought across the other or the thigh raised, so that its outer side may receive blows or kicks.

The most obvious defence injuries are seen in knife attacks, as the victim often attempts to ward off the thrusts by seizing the weapon. When the fingers are closed around the blade, its withdrawal cuts across the flexures of the phalanges, slicing through the skin and perhaps tendons, or sometimes all four fingers. This may be seen in one, any, or even all of the three sets of flexures.

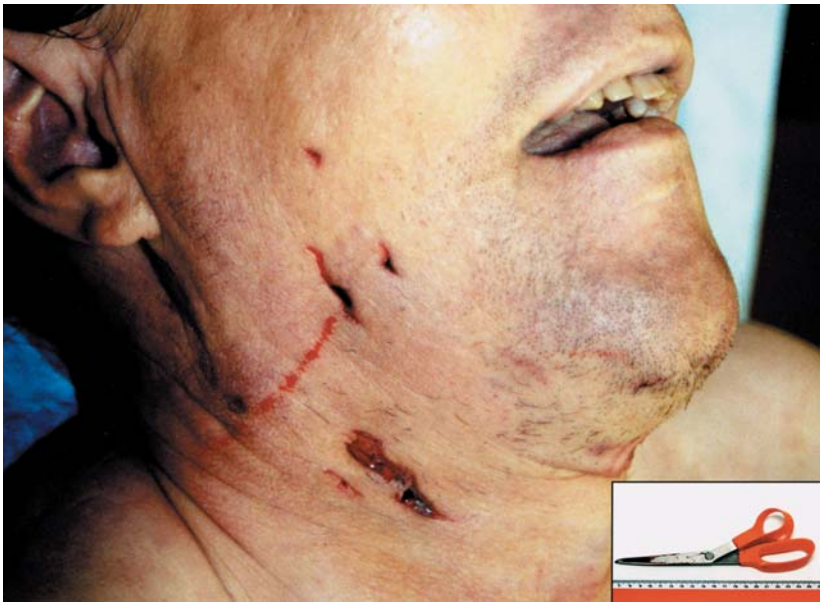

FIGURE 4.53 *Multiple homicidal stab wounds by closed scissors.*

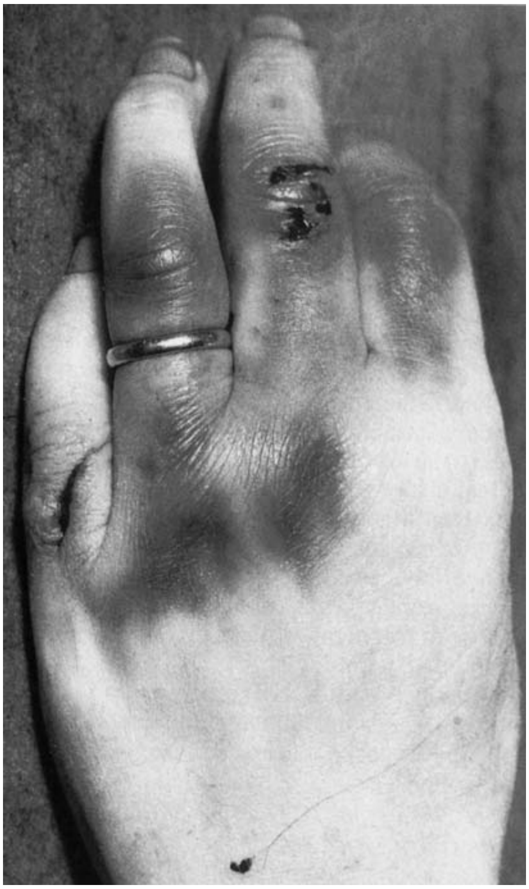

FIGURE 4.54 *Defence injuries: bruising on the back of the hand in an attempt to fend off a blunt instrument in this case a metal poker that was used to inflict fatal head injuries. Such injuries confirm that the victim was conscious and active during the attack, and that it was not made covertly while the victim was unaware.*

The other typical knife defence wound is in the web between the base of the thumb and index finger, when the blade is grabbed in an attempted pincer action. Other knife wounds can be on the backs of the hands or fingers from non-grasping movements trying to ward off the weapon. Defence wounds from knives are often sharply sliced, as the blade is drawn across the tightly applied skin. They are often markedly 'shelved', with loose flaps of skin and copious bleeding.

Defence wounds also occur in firearms injuries, where an arm is raised in a desperate attempt to shield the trunk or head from the blast. This accounts for some shooting deaths where there is an entrance and an exit wound in (often) the upper arm, the missile(s) then repenetrating the trunk.

DATING OF WOUNDS BY HISTOLOGY AND HISTOCHEMISTRY

There are many publications on this subject, as it is a favourite topic for research. As with time of death, it can be an important matter in forensic medical investigations to determine whether a wound found at autopsy was inflicted before or after death and, if inflicted ante-mortem, how long before death it was sustained.

Unfortunately, as with so many problems, biological variability introduces a wide margin of uncertainty, so that a range of probabilities can be offered, but never a definite time interval. Much of the experimental work has been performed on animals and the results are not transferable to man – a common defect in all animal work. The smaller

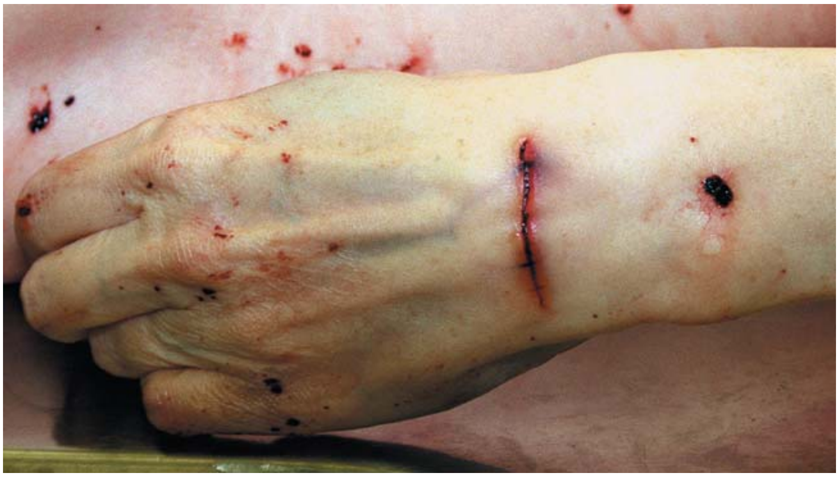

FIGURE 4.55 *Defence wound on the back of the hand from trying to ward off the knife.*

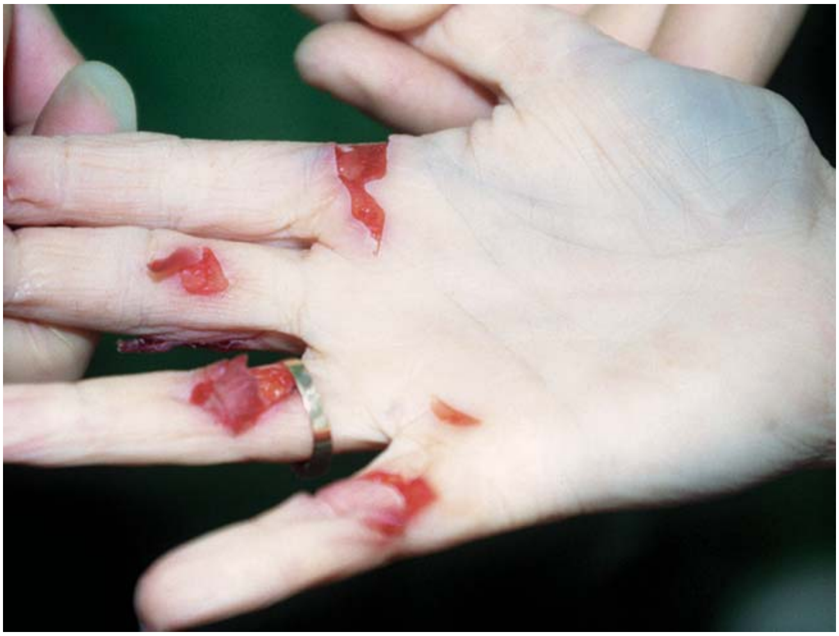

FIGURE 4.56 *Typical defence injuries in a victim of a knife attack. In grasping the blade to deflect it, there have been cuts across the palmar surfaces of the finger joints and a slash between the thumb and forefinger, together with a cut at the base of the thumb.*

the animal, the more rapid the tissue scavenging and reparative changes: more primitive animals have far greater powers of regeneration than man, both in wounds and in regrowing organs and even limbs.

The changes will also vary according to the size of a wound, the type of wound (bruise, abrasion, incision or laceration), the tissue (epidermal or mesodermal), whether there is infection and the age and health of the victim.

The subject is complex and often contradictory and the following schedule of events is offered only as a guideline. The original papers should be consulted if any particular criterion is to be followed; histological appearances, including fluorescence studies may be backed up by histochemical and biochemical assays. The sequence of changes in bruises is far less distinct than in abrasions or lacerations and histology is much less helpful. The most useful criterion in a bruise is probably Perl's reaction for haemosiderin, which becomes positive about 3 days after infliction – though some workers claim an earlier appearance, even down to 12 hours. Haemolysis of red blood cells is irregular and patchy, many intact cells being seen in some bruises after many days, so their rupture and ghosting cannot be used as an index of age.

Chronological histological changes after wound infliction

30 MINUTES–4 HOURS

Margination of polymorph leucocytes in dilated small vessels may occur, a feature which is often completely absent. It is also not a reliable sign of ante-mortem infliction, as

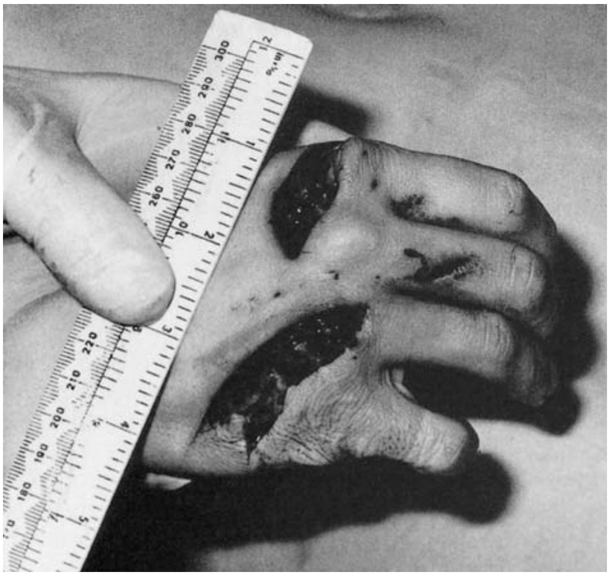

FIGURE 4.57 *Defence wounds of hand in attempts to ward off an assault with a meat cleaver.*

leucocytes may congregate for many hours after death, especially in skin and around aspirated material in the lung.

Some extravascular emigration may begin at the end of this period, but a significant number of polymorphs usually delay for many more hours. Polymorphs tend to appear earlier in the subcutaneous fat than in the upper dermal layers. Basophilic 'mast cells' lose their granules. Fibrin appears in the wound within a few minutes, but this also occurs in post-mortem injuries.

4–12 HOURS

Leucocyte infiltration is likely to be more definite, still mostly polymorphs, but some mononuclear cells as well. A scanty mixed population of lymphocytes and poorly differentiated 'monocytes' appears, usually after 12 hours. Tissue oedema and swelling of vascular endothelium occurs. In small wounds involving the skin, the start of epithelial regeneration may be seen at the sides at the level of the basal layer of epidermis.

12–24 HOURS

Leucocytes tend to demarcate the area of the wound by forming a marginal palisade. The polymorph response declines and the macrophage and mononuclear cell population increases from now on. Removal of necrosed tissue begins, with macrophages evident and a basophilic tinge to the ground substance. Mitoses are visible in fibroblasts from about 15 hours. The epidermis begins to spread across the surface of the scab and down the sides of a cut into the wound.

24–72 HOURS

Leucocyte infiltration reaches a peak at about 48 hours; repair begins concurrently with many fibroblasts appearing, but rarely before 72 hours. New capillaries then begin to bud from vessels, the vascular, infiltrated stroma becoming 'granulation tissue'.

3–6 DAYS

Repair is proceeding apace: collagen begins to form and giant cells may be visible around necrotic debris and foreign matter. The epidermis is actively growing if the wound involves the surface: in animals, it grows laterally at about 200 μm/day. Haemosiderin becomes stainable from about the third day onwards, if there is a bruise or any bleeding into the wound, but is often not obtainable by Perl's reaction until the fifth day even though some claim it can appear on the first or second day.

10–15 DAYS

Cellular reaction subsides in small wounds. Vascularity decreases and the cell population drops, especially leucocytes. Fibroblasts are most active, with collagen being laid down. The epidermis becomes thin and flat, but no papillae ('rete pegs') are reformed for many weeks. Elastic fibres are scarce for a long time and less than in adjacent undamaged tissue.

TWO WEEKS TO SEVERAL MONTHS, DEPENDING ON SIZE AND OTHER FACTORS

Consolidation of the healing tissue continues. The inflammatory response has vanished, unless the wound has become infected. Collagen and elastin increase and a vascular scar is formed, which gradually becomes more dense and avascular. The epithelium remodels and dermal papillae reappear unless the wound is wide and irregular. Skin adnexae do not reappear in the scar unless islands of viable skin survive within the wound area.

Histochemical changes in injured tissues

Histochemical and immunohistochemical methods have been studied widely in recent years, again mainly on animals, the transfer to the human situation being suspect in some cases. A wide variety of markers have been claimed as reliable for both the differentiation between ante-mortem and post-mortem injuries and for dating of ante-mortem wounds.

Electrolytes, such as sodium, zinc, magnesium and calcium, serotonin, esterases, glycophorin A and especially histamine have been used for this purpose.

Mainly as a result of the pioneering work of Jyrki Raekallio in Finland, the histochemical sequence of events in wounds has been actively pursued in recent years. Some techniques need frozen sections, others can be carried out on fixed tissue. Estimations of histamine or serotonin are usually 'tube' methods, rather than microscopic, though some fluorescent techniques have been developed. Betz and his co-workers have also published extensively on the immunohistochemistry of wound dating.

At the present time these methods still remain mainly in the realm of the research laboratory, apart from a few enthusiasts who employ them in casework. For the histochemical novice, it would be unwise to put such procedures into evidence without first gaining extensive experience of control material. Experience emphasizes that strictly standard conditions are required for the production of such histochemical sections, the end result depends largely on laboratory procedures – 'change your technician and you change the answer'.

A list of standard procedures for some of these reactions is given in Appendix 1.

The following are some of the histochemical wound changes described by Raekallio and others:

- In a wound through the skin surface there is a central zone 0.2–0.5 mm wide that will become necrotic and in which enzyme activity rapidly decreases. This may be termed 'negative vital reaction'. Immediately beyond this layer, is a 0.1–0.3 mm zone of reaction and eventual repair, where a number of enzymes and other substances become increased in concentration during the reparative process, compared to the normal level in the areas outside the wound. Enzymatically, this can be called 'positive vital reaction', as no such zone develops in a post-mortem wound.
- Within one hour after injury, esterases and adenosine triphosphatase increase in the positive zone. At around 2 hours, aminopeptidase activities increase and, at 4 hours, acid phosphatase activity increases. Alkaline phosphatase activity is delayed another hour or so though, of course, all these times are relative and subject to the usual biological variability.
- Where death occurs at some stage of this process, the enzyme pattern is 'frozen' at that point and post-mortem changes do not substantially alter the reactions within a few days after death, especially if autolysis is kept at bay by refrigeration. Senility, severe illness and cachexia, as well as widespread multiple injuries, may distort the usual pattern by reducing the ability to produce these reparative enzymes. In contused wounds, the reactions are less useful than in cutaneous injuries, as the damage is more diffuse and there are no definite 'zones'.

  In all wounds, lack of reaction cannot be taken as indisputable proof of a post-mortem origin, though, if the opposite occurs, positive increase in the outer zone must only be the result of a vital reaction. Other enzymes have been used for wound dating: the review by Janssen (1984) in his authoritative book on forensic histology should be consulted for details, as well as the work of Betz.
- Tissue cathepsins are said to increase almost immediately if the stroma is damaged, being demonstrable within 5–10 minutes. Two other substances are of use in establishing that wounds are ante-mortem and giving some idea as to their age. Both are produced during the inflammatory response that accompanies tissue damage – histamine and serotonin (5-hydroxytryptamine). They tend to be complementary to the enzymes, as they appear soon after infliction, within the hour or so before adenosine triphosphatase and esterases become detectable.

  These vasoactive amines appear in maximum concentration about 10 minutes (serotonin) and 20–30 minutes (histamine) after wounding. One of the first demonstrations of this forensic aid was by Fazekas and Varagos-Kis (1965), who showed that a ligature mark in hanging revealed increased serotonin; presumably the hanging was atypical and the victim took 10 minutes to die which is rather unusual. For the employment of the test in human autopsy work, about 2 grams of skin freed from subcutaneous fat are sampled, together with a similar control sample from a nearby normal area of skin from the same corpse. This is to allow for the markedly different amounts found in different people and, indeed, in the same persons at different times.

  To establish that a wound was ante-mortem rather than inflicted after death, the level of histamine in the wound must be at least 50 per cent greater than the control sample – and, for serotonin at least, twice the concentration of the control skin.

SURVIVAL PERIOD AFTER WOUNDING

A pathologist is often asked either by police during an investigation, or by counsel at trial, what was the probable interval between the fatal injuries and death? A supplementary query is how long would the victim have been active, even if still alive?

These are extremely difficult questions to answer, as a number of variables enter the estimation. The doctor should never give a dogmatic answer, unless the nature and severity of injuries are obviously incompatible with continued life or activity.

A person who has his brainstem destroyed by a penetrating injury, or his aortic arch completely transected, will be inactive and clinically dead almost immediately. It would not necessarily be the case if his frontal lobes were damaged or his abdominal aorta was crushed by a railway wheel. Extraordinary instances of survival in both these latter instances are on record.

An opinion should always err on the side of caution in expressing opinions on this subject, as victims can do far more than might be expected after injury. The author (BK) once dealt with a homicide in which the victim was stabbed through the heart, yet ran more than a quarter of mile before collapsing. A wound in the left ventricle can partly seal itself by the contraction of the muscle around the defect and collapse may only occur when sufficient blood has leaked into the pericardium to form a tamponade. Wounds of the right ventricle are often more rapidly fatal as, although the pressure of contained blood is less, the thinner wall is not so effective in preventing the leak.

Head injuries can present many paradoxical instances of prolonged survival, depending partly on what part of the brain is injured. As mentioned, frontal lobes seem remarkably resistant to damage and it is often the generalized head impact, rather than focal damage to this area, that causes most harm. For example, a Finnish man who committed suicide made a home movie of himself firing a pistol into his brain. There was immediate collapse but, just at the end of the 4-minute film, the victim opened his eyes and raised his head, finally dying some uncertain time after the spool ran out.

In the more usual stabbing, head injury, cut throat or shooting, the pathologist must try to assess the nature and severity of the physical damage, and relate that to the age, health and environment of the victim. This is naturally an imprecise exercise. A senile old lady, or someone with severe cardiac or respiratory disease, is less likely to survive multiple injuries for as long as a robust young person – though if the nature of the injury is grossly life-threatening, these factors will make little difference.

A cut artery will lose blood faster than a vein of the same general size, especially if it is only partly severed so that it cannot retract. Other factors, such as the possibility of air embolism occurring in a cut jugular vein, might alter the usual pattern of expected events. The heart is less vulnerable than great thoracic vessels in many stab wounds, as mentioned above. Many wounds of the ventricles are survivable. The author (BK) has twice been warned by police of an imminent autopsy on victims admitted for emergency surgery – only to hear later that, thankfully, they had walked out of the hospital after a successful operation. One had a through-and-through stab wound of the left ventricle that transfixed the heart from front to back.

Where a wound has transected a major coronary artery, then prolonged survival is unlikely, as is one that interrupts a major branch of the conducting system. Other than injuries to the brain or to a large blood vessel, most other injuries can rarely be declared to have caused sudden death or rapid loss of function. In criminal cases, it may be a point of some importance to decide whether the victim could have continued fighting, have run away, have resisted, or even have inflicted injuries on someone else before collapsing and dying. It is difficult to swear that at least some activity was not possible in the great majority of instances. In the adrenaline response of 'fight or flight', the shock element of pain is greatly suppressed – many soldiers have been quite unaware of having sustained severe or lethal wounds until they notice the blood or the battle is over. Thus the usual 'shock' effect may be damped down considerably in an assault and only the sheer physical and haemodynamic sequelae of the injury will eventually lead to a slowing down, then collapse and death. In the interval, injured victims may be able to perform normal physical activity, sometimes to an astonishing degree, albeit followed by sudden deflation of their capacity for exercise.

As in most aspects of forensic medicine, it is unwise of the doctor to be too definite about these matters – the old forensic aphorism 'Seldom say never – seldom say always' is even more true here than usual.

Most issues of this nature revolve around periods of a few minutes, the perimortal or agonal time when few morphological signs are available. The more sophisticated enzyme and serotonin techniques mentioned elsewhere in this chapter may give some clue about the survival period, but are rarely unequivocal enough to be used as hard forensic evidence. If survival is longer, then both gross and histological changes of 'vital reaction', such as thrombosis, inflammation, infection and healing may be useful, but they usually require hours or even days of survival to appear.

The accepted wisdom of wound dating is that polymorph leucocytes begin to appear in wounded tissue within a few minutes, but it has been shown that this can happen even several hours after death, as all leucocytes do not become immobile with cardiac arrest. A good, conventional inflammatory response must be a vital reaction, however, though several hours survival is needed for it to be convincing. The author (BK) has seen a dramatic red flare caused by post-mortem burning at least 30 minutes after undoubted death from strangulation.

Another problem with differentiating 'ante-mortem' from 'post-mortem' injuries is the definition of the moment of death. Again, lawyers, judges and coroners tend to assume that death is an **event**, whereas, in reality, it is a **process** (Chapter 2). Though in most criminal or accidental deaths the moment of death is conventionally – and reasonably – taken to be the moment of cardiac arrest, and thus collapse of blood pressure and cerebral circulation, the cells of the body are still alive and remain so for a variable time – only minutes in the case of neurones, but leucocytes and muscle cells survive for many hours, and connective tissue cells, such as fibroblasts, for days. Thus, it is unreasonable to expect dramatic changes within minutes in skin wounds, etc., until progressive hypoxia alters biochemical processes and enzyme activity. Leucocytes may be motile for more than 12 hours and can aggregate around chemotactically active material, such as gastric contents aspirated into the air passages. This makes the 'vital reaction' a dubiously valid phenomenon in the perimortal period. Much more research is needed into the histology, histochemistry and biochemistry of the perimortal period to clarify these issues, but in the mean time, less dogmatic reliance must be held upon the old criteria of 'ante-' and 'post-' mortem phenomena and the 'vital reaction'.

REFERENCES AND FURTHER READING

Adams VI, Hirsch CS. 1989. Venous air embolism from head and neck wounds. *Arch Pathol Lab Med* **113**:498–502.

Ali TT. 1988. The role of white blood cells in post-mortem wounds. *Med Sci Law* **28**:100–6.

Betz P. 1994. Histological and enzyme histochemical parameters for the age estimation of human skin wounds. *Int J Legal Med* **107**:60–8.

Betz P. 1995. Immunohistochemical parameters for the age estimation of human skin wounds. A review. *Am J Forensic Med Pathol* **16**:203–9.

Betz P, Nerlich A, Wilske J, *et al.* 1992. Comparison of the solophenyl-red polarization method and the immunohistochemical analysis for collagen type III. *Int J Legal Med* **105**:27–9.

Betz P, Nerlich A, Wilske J, *et al.* 1992. The time-dependent rearrangement of the epithelial basement membrane in human skin wounds – immunohistochemical localization of collagen IV and VII. *Int J Legal Med* **105**:93–7.

Betz P, Nerlich A, Wilske J, *et al.* 1992. Time-dependent appearance of myofibroblasts in granulation tissue of human skin wounds. *Int J Legal Med* **105**:99–103.

Betz P, Nerlich A, Wilske J, *et al.* 1992. Time-dependent pericellular expression of collagen type IV, laminin, and heparan sulfate proteoglycan in myofibroblasts. *Int J Legal Med* **105**:169–72.

Betz P, Nerlich A, Wilske J, *et al.* 1993. The immunohistochemical localization of alpha 1-antichymotrypsin and fibronectin and its meaning for the determination of the vitality of human skin wounds. *Int J Legal Med* **105**:223–7.

Betz P, Nerlich A, Tubel J, *et al.* 1993. The time-dependent expression of keratins 5 and 13 during the reepithelialization of human skin wounds. *Int J Legal Med* **105**:229–32.

Betz P, Nerlich A, Tubel J, *et al.* 1993. Localization of tenascin in human skin wounds – an immunohistochemical study. *Int J Legal Med* **105**:325–8.

Betz P, Nerlich A, Wilske J, *et al.* 1993. Analysis of the immunohistochemical localization of collagen type III and V for the time-estimation of human skin wounds. *Int J Legal Med* **105**:329–32.

Betz P, Nerlich A, Wilske J, *et al.* 1993. Immunohistochemical localization of collagen types I and VI in human skin wounds. *Int J Legal Med* **106**:31–4.

Betz P, Nerlich A, Wilske J, *et al.* 1993. The time-dependent localization of ki67 antigen-positive cells in human skin wounds. *Int J Legal Med* **106**:35–40.

Betz P, Lignitz E, Eisenmenger W. 1995. The time-dependent appearance of black eyes. *Int J Legal Med* **108**:96–9.

Biddinger PW. 1987. Postmortem wound dehiscence. A report of three cases. *Am J Forensic Med Pathol* **8**:120–2.

Bohnert M, Baumgartner R, Pollak S. 2000. Spectrophotometric evaluation of the colour of intra- and subcutaneous bruises. *Int J Legal Med* **113**:343–8.

Byard RW, Gilbert JD, Brown K. 2000. Pathologic features of fatal shark attacks. *Am J Forensic Med Pathol* **21**:225–9.

Camps FE. 1952. Interpretation of wounds. *Br Med J* **2**:770–4.

Fatteh A. 1966. Histochemical distinction between antemortem and postmortem skin wounds. *J Forensic Sci* **11**:17–27.

Fatteh A. 1971. Distinction between antemortem and postmortem wounds: A study of elastic fibers in human skin. *J Forensic Sci* **16**:393–6.

Fazekas IG, Viragos-Kis E. 1965. [Free histamine content in the groove caused by hanging as an *in vivo* reaction.] *Dtsch Z Gesamte Gerichtl Med* **56**:250–68.

Fernandez P, Bermejo AM, Lopez Rivadulla M, *et al.* 1994. Biochemical diagnosis of the intravital origin of skin wounds. *Forensic Sci Int* **68**:83–9.

Green MA. 1978. Stab wound dynamics – a recording technique for use in medico-legal investigations. *J Forensic Sci Soc* **18**:161–3.

Grellner W, Madea B, Kruppenbacher JP, *et al.* 1996. Interleukin-1 alpha (IL-1 alpha) and n-formyl-methionyl-leucyl-phenylalanine (FMLP) as potential inducers of supravital chemotaxis. *Int J Legal Med* **109**:130–3.

Gupta SM, Chandra J, Dogra TD. 1982. Blunt force lesions related to the heights of a fall. *Am J Forensic Med Pathol* **3**:35–43.

Hamdy M, May K, Flanagan W, *et al.* 1961. Determination of the age of bruises in chicken broilers. *Poultry Sci* **40**:787–9.

Henn V, Lignitz E, Philipp KP, *et al.* 2000. [Morphology and phenomenology of death by kicking (II).] *Arch Kriminol* **205**:65–74.

Hernandez-Cueto C, Luna A, Lorente JA, *et al.* 1987. Study of cathepsin A, B and D activities in the skin wound edges. Its application to the differential diagnosis between vital and postmortem wounds. *Forensic Sci Int* **35**:51–60.

Hernandez-Cueto C, Luna A, Villanueva E. 1987. Differential diagnosis between vital and postmortem wounds: ions as markers. *Adli Tip Derg* **3**:14.

Hiss J, Hirshberg A, Dayan DF, *et al.* 1988. Aging of wound healing in an experimental model in mice. *Am J Forensic Med Pathol* **9**:310–12.

Hocking MA. 1989. Assaults in south east London. *J R Soc Med* **82**:281–4.

Hou Jensen K. 1969. Some enzyme conditioned vital reactions in the initial phase of wound healing and their medico-legal significance. *Dan Med Bull* **16**:305–8.

Hunt AC, Cowling RJ. 1991. Murder by stabbing. *Forensic Sci Int* **52**:107–12.

Janssen W. 1984. *Forensic histopathology*. Springer Verlag, Berlin.

Kibayashi K, Hamada K, Honjyo K, *et al.* 1993. Differentiation between bruises and putrefactive discolorations of the skin by immunological analysis of glycophorin A. *Forensic Sci Int* **61**:111–7.

Knight B. 1975. The dynamics of stab wounds. *Forensic Sci* **6**:249–55.

Kosa F, Kurucz E. 1986. Immuno-histochemical reactions with monoclonal antileukocyte antibodies in vital injuries. *Acta Med Leg Soc (Liege)* **36**:176–90.

Laiho K. 1967. Immunohistochemical studies on fibrin in vital reactions. *Acta Med Leg Soc (Liege)* **20**:187–91.

Laiho K. 1998. Myeloperoxidase activity in skin lesions. I. Influence of the loss of blood, depth of excoriations and thickness of the skin. *Int J Legal Med* **111**:6–9.

Laiho K. 1998. Myeloperoxidase activity in skin lesions. II. Influence of alcohol and some medicines. *Int J Legal Med* **111**:10–2.

Langlois NE, Gresham GA. 1991. The ageing of bruises: a review and study of the colour changes with time. *Forensic Sci Int* **50**:227–38.

Logoida DM. 1959. Assessment of survival after trauma, by the leucocyte counts in the capillary network of internal organs. *Sud-Med Ekspert* **4**:5–13.

Maeno Y, Takabe F, Mori Y, *et al.* 1991. Simultaneous observation of catecholamine, serotonin and their metabolites in incised skin wounds of guinea pig. *Forensic Sci Int* **51**:51–63.

Mason JK. 1965. The importance of the histological examination in death from accidental trauma. *Med Serv J Can* **21**:316–25.

Maxeiner H. 1998. ‘Hidden’ laryngeal injuries in homicidal strangulation: how to detect and interpret these findings. *J Forensic Sci* **43**:784–91.

McCausland IP, Dougherty R. 1978. Histological ageing of bruises in lambs and calves. *Aust Vet J* **54**:525–7.

Muir R, Niven J. 1935. The dating of contusions. *J Pathol Bacteriol* **41**:183–8.

Niu WY, Hu JZ, Zhang XM. 1991. A new staining method for constriction marks in skin. *Forensic Sci Int* **50**:147–52.

Njau SN, Epivatianos P, Tsoukali Papadopoulou H, *et al.* 1991. Magnesium, calcium and zinc fluctuations on skin induced injuries in correlation with time of induction. *Forensic Sci Int* **50**:67–73.

Ohshima T. 2000. Forensic wound examination. *Forensic Sci Int* **113**:153–64.

Ojala K. 1967. Histochemical and morphological of vital reactions in muscle wounds. *Acta Med Leg Soc (Liege)* **20**:193–4.

Ortiz-Rey JA, Suarez-Penaranda JM, Da Silva EA, *et al.* 2002. Immunohistochemical detection of fibronectin and tenascin in incised human skin injuries. *Forensic Sci Int* **126**:118–22.

Presswalla FB. 1978. The pathophysics and pathomechanics of trauma. *Med Sci Law* **18**:239–46.

Prinsloo I, Gordon I. 1951. Postmortem dissection artefacts of the neck and their differentiation from ante-mortem bruises. *S Afr Med J* **25**:358–61.

Quan L, Fujita MQ, Zhu BL, *et al.* 2000. Immunohistochemical distribution of c-reactive protein in the hepatic tissue in forensic autopsy. *Forensic Sci Int* **113**:177–82.

Rabinovitz A. 1956. Medicolegal conclusions on the form of the knife used based on the shape of the stab wounds received. *J Forensic Med* **6**:160–5.

Raekallio J. 1960. Enzymes histochemically demonstrable in the earliest phase of wound healing. *Nature* **188**:235–45.

Raekallio J. 1963. Histochemical distinction between antemortem and postmortem skin wounds. *J Forensic Sci* **9**:107–10.

Raekallio J. 1966. Enzyme histochemistry of vital and postmortem skin wounds. *J Forensic Medicine* **13**:85–90.

Raekallio J. 1967. Application of histochemical methods to the study of traffic accidents. *Acta Med Leg Soc (Liege)* **20**:171–8.

Raekallio J. 1970. *Enzyme histochemistry of wound healing*. Gustave Fischer Verlag, Stuttgart, Germany.

Raekallio J. 1972. Determination of the age of wounds by histochemical and biochemical methods. *Forensic Sci* **1**:3–16.

Raekallio J. 1980. Estimation of time in forensic biology and pathology. An introductory review. *Am J Forensic Med Pathol* **1**:213–18.

Raekallio J, Makinen PL. 1966. Histamine content as a vital reaction: experimental investigation. *Zacchia* **41**:273–84.

Raekallio J, Makinen PL. 1969. Serotonin content as vital reaction. I. Experimental investigation. *Zacchia* **5**:587–94.

Raekallio J, Makinen PL. 1970. Serotonin and histamine contents as vital reactions. II. Autopsy studies. *Zacchia* **6**:403–14.

Ritz-Timme S, Eckelt N, Schmidtke E, *et al.* 1998. Genesis and diagnostic value of leukocyte and platelet accumulations around 'air bubbles' in blood after venous air embolism. *Int J Legal Med* **111**:22–6.

Roberts R. 1983. Sexual offences. In: *Rape – The new police surgeon supplement.* A. P. S. G. B., Northampton,UK.

Robertson I, Hodge PR. 1972. Histopathology of healing abrasions. *Forensic Sci* **1**:17–25.

Robertson I, Mansfield RA. 1957. Antemortem and postmortem bruises of the skin – their differentiation. *J Forensic Med* **4**:2–10.

Sato Y, Ohshima T. 2000. The expression of mrna of proinflammatory cytokines during skin wound healing in mice: a preliminary study for forensic wound age estimation (II). *Int J Legal Med* **113**:140–5.

Sevitt S. 1970. Reflections on some problems in the pathology of trauma. *J Trauma* **10**:962–73.

Shapiro HA, Gluckman J, Gordon I. 1962. The significance of the nail abrasion of the skin. *J Forensic Med* **9**:17–19.

Shepherd JP, Price M, Shenfine P. 1990. Glass abuse and urban licensed premises. [letter] *J R Soc Med* **83**:276–7.

Shepherd JP, Shapland M, Pearce NX, *et al.* 1990. Pattern, severity and aetiology of injuries in victims of assault. *J R Soc Med* **83**:75–8.

Shepherd JP, Huggett RH, Kidner G. 1993. Impact resistance of bar glasses. *J Trauma* **35**:936–8.

Simpson CK. 1979. *Forensic medicine*, 8th edn. Edward Arnold, London.

Spitz WU, Petty CS, Fisher RS. 1961. Physical activity until collapse following fatal injury by firearms and sharp pointed instruments. *J Forensic Sci* **6**:290–6.

Strauch H, Wirth I, Taymoorian U, *et al.* 2001. Kicking to death – forensic and criminological aspects. *Forensic Sci Int* **123**:165–71.

Takamiya M, Saigusa K, Aoki Y. 2002. Immunohistochemical study of basic fibroblast growth factor and vascular endothelial growth factor expression for age determination of cutaneous wounds. *Am J Forensic Med Pathol* **23**:264–7.

Teare RD. 1960. Blows with the shod foot. *Med Sci Law* **1**:429–36.

Thoresen SO, Rognum TO. 1986. Survival time and acting capability after fatal injury by sharp weapons. *Forensic Sci Int* **31**:181–7.

Thornton RN, Jolly RD. 1986. The objective interpretation of histopathological data: an application to the ageing of ovine bruises. *Forensic Sci Int* **31**:225–39.

Torre C, Varetto L. 1985. Sem study of dermal surface. A new approach to forensic traumatology. *Am J Forensic Med Pathol* **6**:256–70.

Torre C, Varetto L, Mattutino G. 1986. Dermal surface morphology in wound healing. An experimental scanning electron microscope study. *Am J Forensic Med Pathol* **7**:337–43.

Wang D, Zhu J. 1992. Localization and quantification of the nonspecific esterase in injured skin for timing of wounds. *Forensic Sci Int* **53**:203–13.

Watson AJ. 1978. *Kicking, karate and kung-fu.* In: Mason JK (ed.), *Pathology of violent injury*. Edward Arnold, London.

Wilson EF. 1977. Estimation of the age of cutaneous contusions in child abuse. *Pediatrics* **60**:750–2.

Zhong FC, Zhen ZJ. 1991. Localization and quantification of histamine in injured skin as parameters for the timing of wounds. *Forensic Sci Int* **51**:163–71.

CHAPTER 5

Head and spinal injuries

- Injury to the scalp 174
- Facial injuries 178
- Falls 180
- Fractures of the skull 181
- Forensic anatomy of the brain membranes 189
- Extradural haemorrhage 190
- Heat haematoma 192
- Subdural haemorrhage 192
- Subarachnoid haemorrhage 196
- Forensic implications of brain membrane haemorrhage 198
- Ruptured berry aneurysm and trauma 198
- Subarachnoid haemorrhage and alcohol 199
- The rapidity of death in subarachnoid haemorrhage 199
- Rotational trauma to the head and upper neck: basilovertebral artery injury 200
- Head injuries in boxers 204
- Cerebral injuries 204
- Histological diagnosis of early cerebral hypoxia 212
- Secondary brainstem lesions 213
- Spinal injuries 214
- References and further reading 216

Of all regional injuries, those of the head and neck are the most common and most important in forensic practice. Adelson (1974) gives these sound reasons for this dominance of head injuries:

- The head is the target of choice in the great majority of assaults involving blunt trauma.
- When the victim is pushed or knocked to the ground, he often strikes his head.
- The brain and its coverings are vulnerable to degrees of blunt trauma that would rarely be lethal if applied to other areas.

A sound practical understanding of the neuropathology of trauma is more essential to the forensic pathologist than any other aspect of his subject, as head injuries provide the major contribution to death in assaults, falls and transportation accidents.

INJURY TO THE SCALP

The scalp is often, though by no means invariably, damaged in trauma that causes injury to the underlying skull and brain. The usual range of abrasions, contusions and lacerations may be inflicted, though a modifying factor is the presence of hair, which may deflect a tangential blow or partly cushion a direct impact.

When an injury is visible on the forehead, the back of the neck, the lower temple or on a bald area, the examination is no different from elsewhere on the body. In hair-covered areas, care must always be taken at autopsy to palpate the scalp in any case in which there is a possibility of injury, otherwise abrasions, swelling, bruising and even lacerations may be missed. When a lesion is found or suspected, the hair must be carefully shaved away to expose the scalp for further examination and photography.

Forensic anatomy of the scalp

Superficially, the skin carries hair follicles, sebaceous glands and sweat glands. The skin is attached to the aponeurosis (see below) by vertical strands of fibrous tissue that break up the subcutaneous layer into pockets filled with fat. The blood vessels and nerves lie in this layer, above the epicranial aponeurosis (formerly called the ‘galea aponeurotica’). This is a dense sheet of fibrous tissue that lies in the deep layer of the scalp over the whole cranium. It is really a flattened tendon uniting the frontal and occipital bellies of the occipitofrontalis muscle.

Deep to the aponeurosis is a thin layer of loose connective tissue that separates it from the pericranium, which is the exterior periosteum of the skull, the dura being the internal counterpart. Some veins traverse all the layers from the superficial fascia to the pericranium, and go on to penetrate the skull and communicate with the intracranial venous sinuses, thus forming a route for meningitis and sinus thrombosis from infected injuries of the scalp.

Abrasion of the scalp

Brush abrasions are less common than in other sites because of the protective effect of the hair, which also tends to prevent or blur the patterned effect of less severe impacts. Impact abrasions from a perpendicular force are imprinted as usual on to the scalp, though again the intervening hair may reduce the severity. Unless the hair is carefully removed at autopsy, with a sharp scalpel or razor, and care taken not to cause artefactual cuts, lesser degrees of abrasion will inevitably be missed.

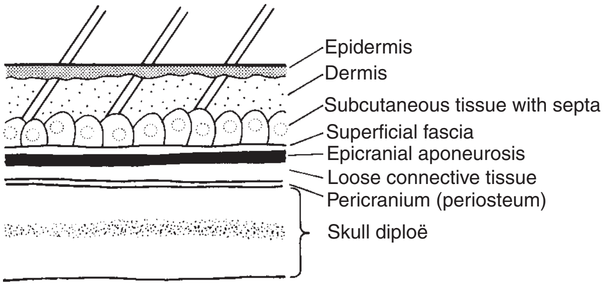

FIGURE 5.1 *Anatomy of the scalp.*

Bruising of the scalp

Bruising may be difficult to detect until the hair has been removed. Marked swelling is a common feature of extensive bruising, as the liberated blood cannot extend downward because of the rigidity of the underlying skull. However, this subsides, or at least diffuses, after death. Commonly, a severe head injury leads to a thick, swollen, indurated layer of blood beneath the scalp, which may extend over a wide area. The blood is sometimes below the aponeurosis, the tough fascial layer of the scalp, but is more often between this and the epidermis.

Blood may also be present beneath the pericranium, the periosteum that is closely applied to the outer surface of the skull. This is often seen in head injuries in infants, usually in association with skull fractures, as the source of the blood is from the fracture line itself. The close attachment of the pericranium to the suture lines in infants may sharply circumscribe the extent of the bleeding.

In addition to frank bleeding beneath the scalp, marked oedema may occur after injury and the layers of the scalp may be greatly swollen and thickened by a jelly-like infiltration of tissue fluid.

As will be discussed under 'black eye', bleeding under the scalp may be mobile, especially under gravity. Thus a bruise or haematoma under the anterior scalp may slide downwards within hours – even minutes – to appear in the orbit,

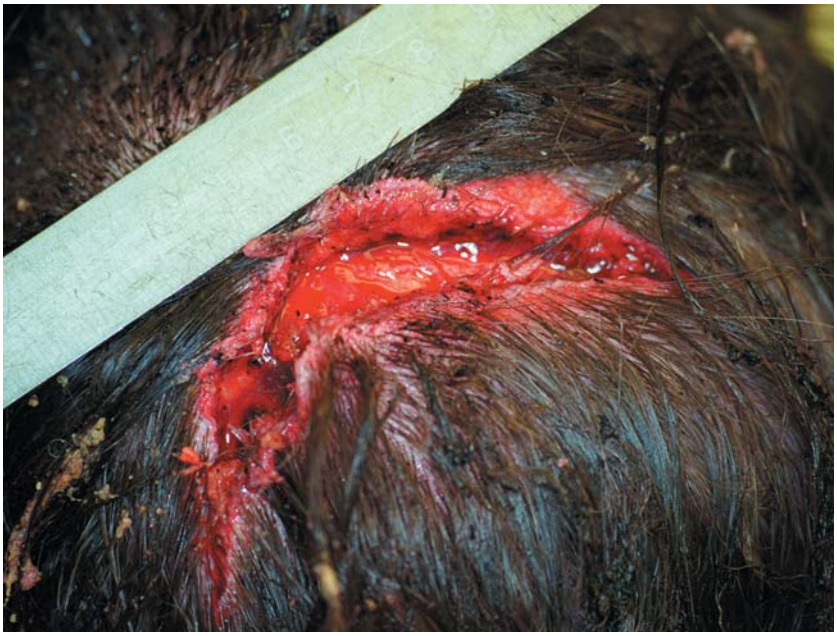

FIGURE 5.2 *A laceration of the scalp caused by a blow from an iron bar. The edges are crushed and bruised, with strands of connective tissue and hairs crossing the gap, indicating that it was not caused by a sharp-edged weapon.*

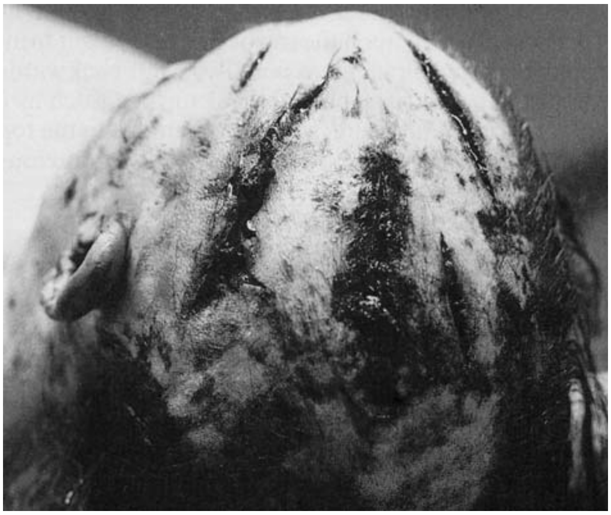

FIGURE 5.3 *Lacerations of the scalp from an iron bar. The margins are bruised and the scalp tissue is extruding in places. The generally parallel direction of the five wounds indicates that the assailant probably delivered the blows in rapid succession with little change of orientation between weapon and head.*

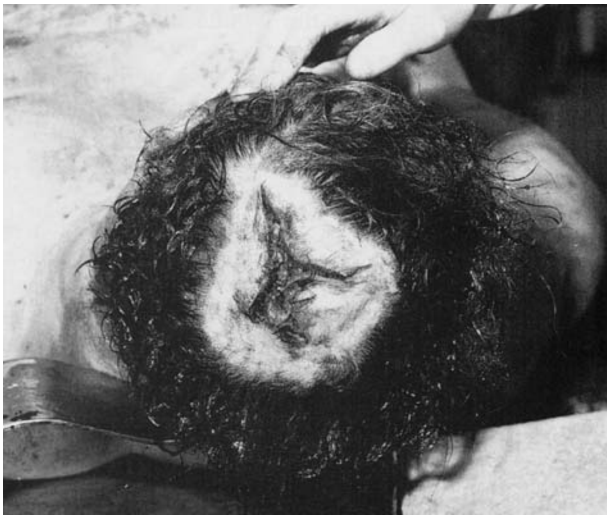

FIGURE 5.4 *Stellate laceration of the scalp caused by a heavy blow with a piece of timber. The support of the underlying skull has caused the tissues to split widely. At autopsy, full clearance of the hair must be made to allow detailed examination and photography.*

simulating a black eye from direct trauma. Similarly, a temporal bruise may later appear behind the ear, suggesting primary neck impact. As with bruises elsewhere, those under the scalp may be obvious immediately after infliction – or their appearance may be delayed, either during life or as a post-mortem phenomenon. They may first become evident, or much more prominent, some hours – or even a day or so – after death. This is caused either by movement of the liberated blood through tissue planes or by haemolysis spreading outwards to stain the subcutaneous tissues, making it always advisable to return to examine the body a day or two following the autopsy.

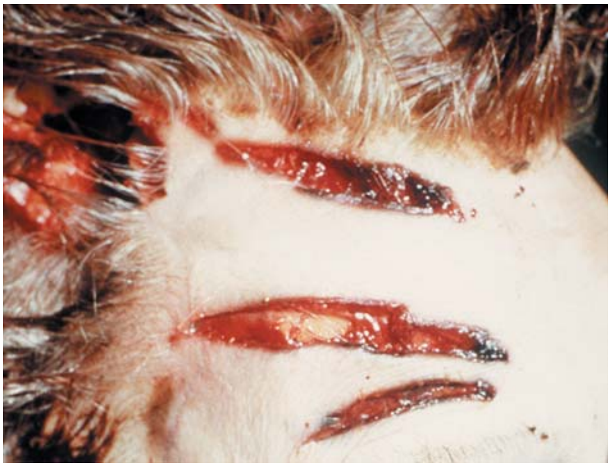

FIGURE 5.5 *Wounds from a metal poker superficially resembling incised cuts, but having edges and tissue bridges within the wounds.*

The shape of an inflicting weapon or object is poorly reproduced on the scalp, again due to the padding effect of the hair. Where the scalp is free of hair, as in the upper forehead or bald areas, all traumatic lesions are similar to elsewhere on the body, with the exception that blunt impact may cause very sharply defined lacerations.

Laceration of the scalp

Lacerations of the scalp bleed profusely, and dangerous and even fatal blood loss can occur from an extensive scalp injury if it is not checked by treatment. The most gross injury is avulsion of a large area of scalp, which can be torn from the head, thereby exposing the aponeurosis or skull. This may happen if the hair becomes entangled in machinery, as was formerly not uncommon in women working in factories. A more common cause nowadays is a traffic accident, where a rotating vehicle tyre comes into contact with the head, causing a 'flaying' injury similar to that seen on limbs.

Scalp injuries may bleed profusely **even after death**, especially if the head is in a dependent position. A post-mortem injury to the head may bleed considerably if inflicted soon after death and these facts may sometimes cause confusion about the ante-mortem or post-mortem nature of the wound, or about the length of time of survival following the injury. There is no reliable way of resolving this difficulty.

Lacerations of the scalp may reproduce the pattern of the inflicting object, even though a random splitting is so common. Severe blows from shaped objects such as hammers or heavy tools may reproduce the profile of the weapon totally or in part. A circular-faced hammer may punch a circle in the scalp, but more often only an arc of a circle is seen. In such cases, the position of the edge that digs in most deeply may give an indication of the angle of the blow. There may be a depressed fracture of the underlying skull of the same shape and size, though the interposition of the dense scalp may cause the skull defect to be slightly larger than the weapon. A depressed fracture in these circumstances is not inevitable, however, and one or more linear fractures may radiate from the impact site.

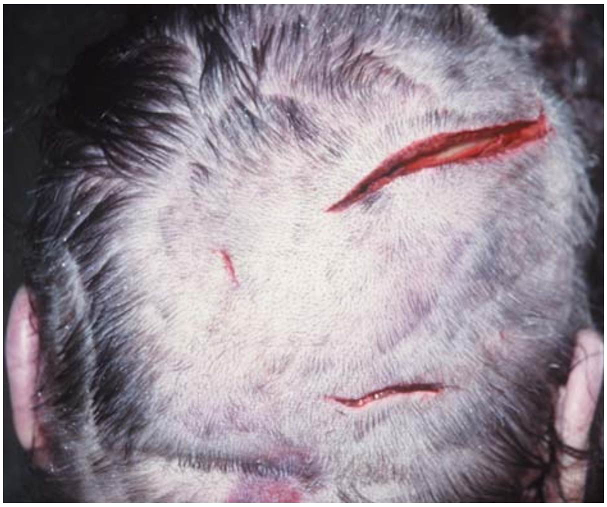

FIGURE 5.6 *Deep linear incised wounds due to a heavy, sharp cleaver. The depth of penetration varies, the large wound overlying extensive skull fractures. The wounds on the neck are due to light contact with the edge of the same blade.*

A major problem in scalp injuries is the differentiation between incised wounds and lacerations from blunt injury.

The scalp is the best example of a surface tissue lying over an unyielding bony support. Violent compression will crush the scalp against the underlying skull, so a blow from a blunt rod-like weapon may split the skin and underlying tissues in a sharply demarcated fashion, which may appear remarkably like a slash from a sharp instrument. Close examination, using a lens if necessary, will show that this blunt laceration has:

- bruised margins, even though this zone may be narrow
- head hairs crossing the wound, which have not been cut
- fascial strands, hair bulbs and perhaps small nerves and vessels in the depths of the wound.

Scalp injuries from falls

It is vital for the pathologist to appreciate that falls on to a flat surface, or a blow from a wide, flat object such as a plank or paving stone, may sometimes leave no external mark whatsoever on the exterior of the head, but commonly such an injury will cause a ragged split which may be linear, stellate or quite irregular.

Such injuries on the back point of the head are commonly caused by falling, especially in inebriated victims. Falls backwards against a ridge, such as a wall or pavement kerb, may cause a transverse laceration, which may be undercut and partly detached from the underlying bone so that a flap of scalp is loosened from the skull.

Falls usually injure the occipital protruberance, the forehead or the parietotemporal areas. Injuries on the vertex

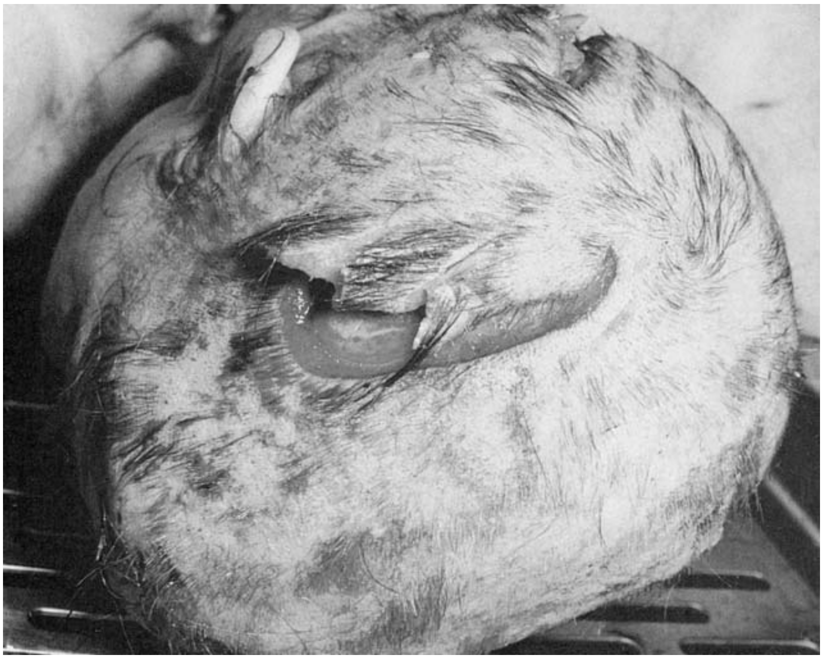

FIGURE 5.7 *Sliced incised wound of the scalp, from a large knife. The wound is markedly undercut, turning a wide flap of scalp. The clean edges, with a lack of any abrasion or bruising, indicate the sharpness of the weapon.*

should always raise the suspicion of assault, as it is unusual to fall upon the top of the head, even from a considerable height. Occasionally, a fall backwards that just happens to reach a vertical surface, such as a wall or piece of furniture, can cause damage to the top of the head, but there is then usually an obvious grazing component to the lesion.

FACIAL INJURIES

Damage to the face is common, but unless gross, with skeletal damage, is rarely fatal in itself unless it leads to bleeding into the air passages. It is often ancillary to fatal cranial damage, or it may be the route by which severe trauma reaches the brain.

The usual range of injuries may be present externally, but all degrees of underlying damage may also occur in the facial skeleton. Because of the complex contours of the face, the various prominences of chin, nose, cheekbones, eyebrows, ears and lips may intercept impacts, with consequent characteristic damage. The eyebrow is particularly vulnerable, being exposed during falls and blows. A blunt impact on the brow often splits the skin and may cause an underlying frontal fracture that can involve the orbital margin.

The distal part of the nose is flexible and often escapes serious damage, though abrasion is common. The bony bridge of the nose is often fractured, which may be detected by movement and crepitus during external palpation – and by dissection at autopsy. Bleeding in the nose is more important than structural damage, as profuse haemorrhage in an unconscious victim may pass back through the posterior nares into the throat and cause fatal airway obstruction.

The maxillae and mandible may be fractured by direct blows and again cause dangerous intra-oral bleeding from associated soft-tissue damage. A heavy blow or kick to one side of the jaw can cause ipsilateral, bilateral or even contralateral fractures. Gross injury to the face, seen in kicking and some transport accidents, may actually detach the facial skeleton from the base of the skull. The lower part of the maxilla, carrying the palate and upper teeth, may be completely separated from the rest of the skull. At autopsy, the best view of the facial skeleton may be gained by dissecting the whole facial skin upwards from the neck incision and reflecting it as far as the orbits, if necessary. Good restoration can be achieved as long as the skin is not penetrated by the knife.

Injuries to the mouth and lips are very common in 'beating-up' incidents, including child abuse. The lips may be bruised or lacerated, much of the damage arising from compression of the lips against the teeth or bony gums. Lacerations on the gingival aspect of the lips may often be exactly matched with the edges of teeth and, as discussed in Chapter 22, rupture of the frenulum inside the upper lip of a child is virtually pathognomonic of a sideswipe across the mouth, if damage from clumsy and forceful attempts at introducing a feeding bottle, dummy or airway can be excluded.

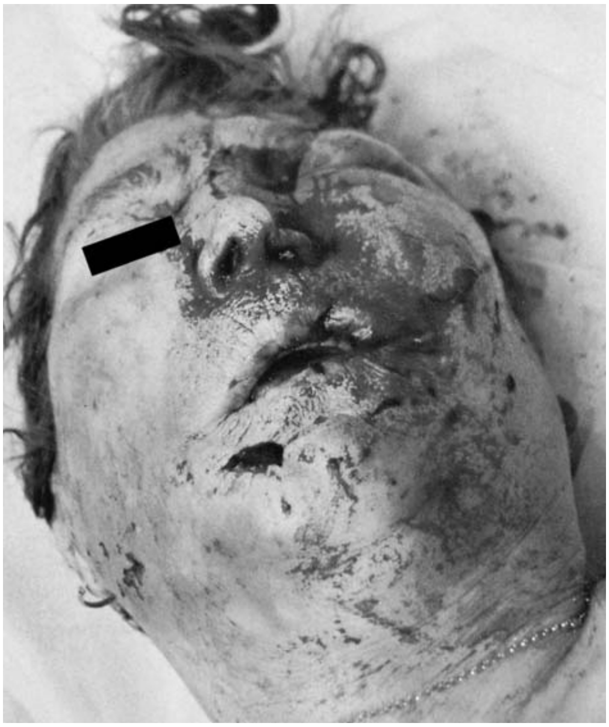

FIGURE 5.8 *Facial injuries caused by kicking. The upper lip is split and the maxilla fractured on that side. The face and orbital region is grossly swollen and bruised. Death was caused by blockage of the air passages by blood.*

Kicking

Kicking of the face is regrettably common and again the prominences suffer most. Bruising, laceration and fractures may result from kicking under the side of the jaw: similar lesions occur on the maxillary area and the eyebrows. Patterned abrasions from boot soles may be seen or crescentic marks from toecaps. Brush abrasions from glancing kicks may be present on the cheeks or forehead as the sole of the shoe scrapes across the skin. It is uncommon to suffer a 'pure' black eye from a kick without other facial injuries, such as scuffed abrasion on the cheekbones, or marks on the brows or bridge of the nose. Teeth may be loosened, broken or detached by both kicks and heavy punches and black eyes and fractures of the nose are common. A kick on the side of the jaw may cause bilateral jaw fractures or even a single contralateral fracture.

It may be difficult or impossible to differentiate in every case, between injuries (especially to the face) caused by kicks

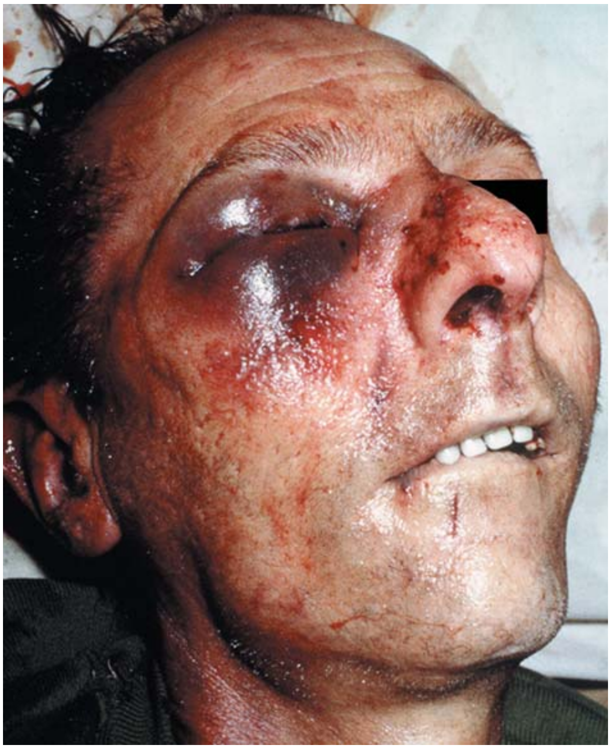

FIGURE 5.9 *Black eye from a direct impact into the orbit from a punch. This is the third mechanism of production of a periorbital haematoma, the others being a fracture of the anterior base of the skull and a frontal scalp wound.*

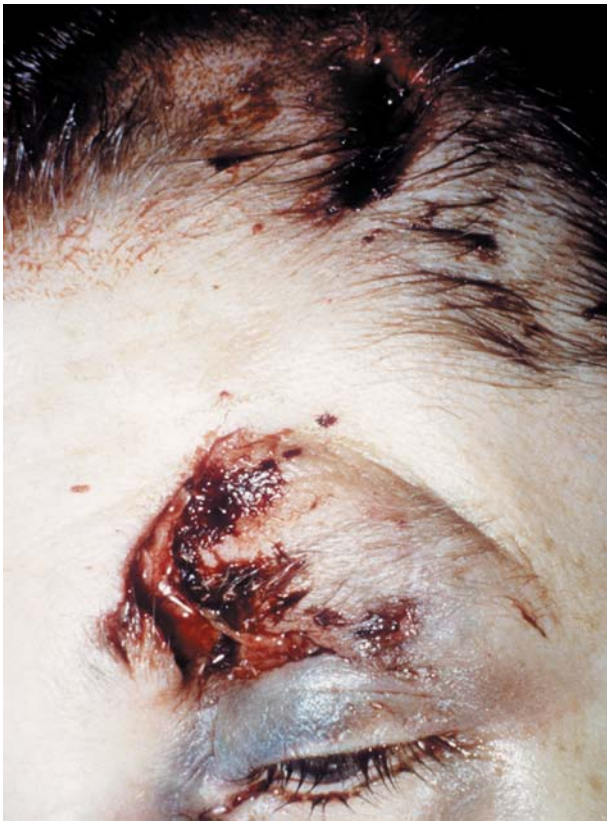

FIGURE 5.10 *Black eye as a result of gravitational seepage of blood downwards from a forehead injury. The woman was struck with a rock on the frontal area and eyebrow, blood then descending during the few hours of coma before death occurred.*

and those from blows from a blunt object. Toecap marks are not all that common, especially since the more flexible rubber 'trainer' shoes have become almost universal footwear.

In stamping, there is the chance that the sole pattern may leave an imprint, but a swing from a toe may leave a non-specific abrasion, bruise or laceration. It may be that the severity of the injury, including underlying bone damage, may be a better indication of a kick than the shape of the injuries, as the force delivered by a swinging foot at the end of a muscular leg is greater than that from a fist.

Black eyes

The usual periorbital haematoma or 'black eye' is usually caused by a direct punch or kick into the eye-socket, but the pathologist must always consider the several alternative explanations. A black eye may be the result of:

- direct violence, which may or may not be associated with abrasion or laceration on the upper cheek, eyebrow, nose or other part of the face
- gravitational seepage of blood beneath the scalp from a bruise or laceration on or above the eyebrow. Survival and at least a partially upright posture of the head must have been maintained for at least some minutes, usually longer, between the time of injury and death. When the scalp lesion is high up on the frontal region, this time will probably be measured in hours
- percolation of blood into the orbit from a fracture of the anterior fossa of the skull. This is often from a contrecoup injury caused by a fall on to the back of the head, leading to secondary fracture of the paper-thin bone of the orbital roof. It is invariably associated with contrecoup contusion of the frontal lobes of the brain, as described later in this chapter.

A simple fall onto the face on a flat surface does not usually cause a black eye, as the prominences of the eyebrow, cheekbone and nose prevent damage to the orbit.

Damage to the ear

The external ear often suffers from blows to the head and is an obvious target in child abuse. Bruising and laceration of

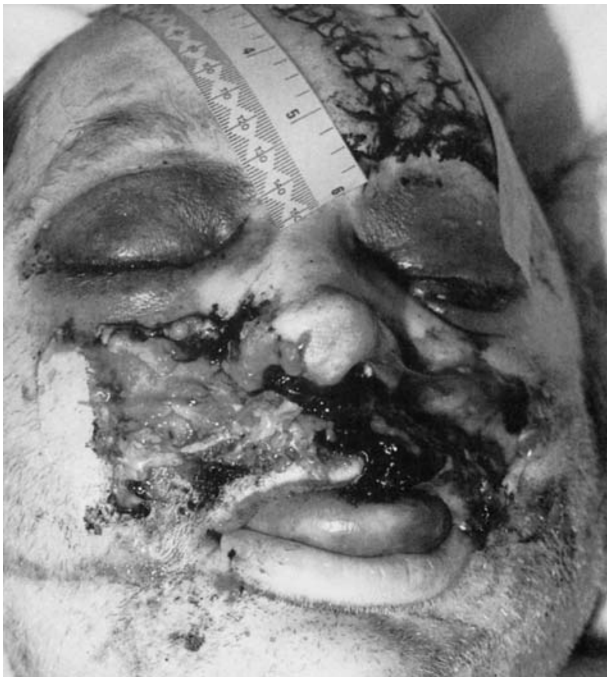

FIGURE 5.11 *Bilateral black eyes caused by leakage of blood into the orbits through comminuted fractures in the floor of the anterior fossa. This homicide victim was struck on the head with a shovel and survived for some days. Brain tissue is escaping from the nostrils through basal skull fractures.*

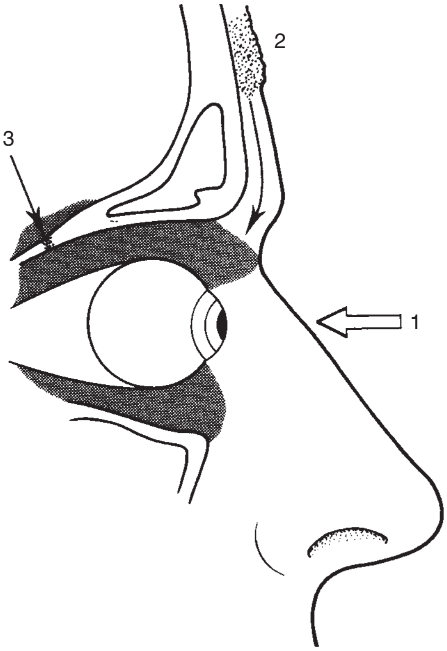

FIGURE 5.12 *The production of a black eye: (1) A direct blow into the orbit. (2) An injury to the front of the scalp, draining down over the supraorbital ridge. (3) A fracture of the base of the skull (direct or contrecoup) allowing meningeal haemorrhage to escape through the orbital roof.*

the pinna is obvious on examination and – in severe trauma – the root of the ear may be detached from the head, usually by a tear at the posterior margin where the ear joins the head. Where gross damage is present, especially with partial avulsion of the pinna, kicking must be considered.

The ear may be bitten and even partly detached, a fate which occasionally is suffered by the nose. In such cases the advice of a forensic odontologist may be invaluable, as teeth marks may form vital evidence.

FALLS

Falls are extremely common, the severity not necessarily being directly related to the distance that the person falls. Many people die after falling from a standing position, yet others sometimes survive a fall of many metres.

Falls from a standing position can occur if a person is drunk, from an assault, during illness (such as a fit or faint) and for many other reasons. Death can follow from a head injury, especially onto the back of the head. An occipital scalp laceration or a fracture of the skull is not necessary for cerebral damage (often frontal contrecoup) to occur. There may also be a subdural or (less often) an extradural haemorrhage, the latter more common from a fall on to the side of the head.

The vexed question of head injury from falls in children is discussed in Chapter 22, but here it may be stated that, although fatal head injury from a fall usually requires a drop of a number of feet, there are well-authenticated instances of skull fractures and brain damage from trivial falls, including some medically witnessed falls from tables and settees. The experimental work of Weber (see Chapter 22) showed that the skulls of small infants could be fractured against a variety of floor surfaces from passive falls of only 34 inches. It is thus invalid for medical witnesses to claim it **cannot** happen, as even one authenticated case creates a precedent. In adults, fractures have certainly occurred from falls onto very hard surfaces from only a foot or so. One such case was a drunk lying on concrete; equally drunken friends attempted to lift him but allowed his head and shoulders to fall back from about half-sitting position, causing occipital fracture.

Falls in old people very frequently cause fractures of the post-cranial skeleton – especially the neck of femur – though ribs, arms and pelvis may also suffer. Osteoporosis is the major reason for the large number of such injuries from falls. More than 47 000 fractured femora occur each year in Britain, with a 25 per cent mortality rate, mainly from subsequent pulmonary embolism or bronchopneumonia.

Falls from a height

Falls from a considerable height, usually from a building, are common in suicide and in some accidents, especially to children. Occasionally deaths from a high fall may be homicidal, again especially in children.

When a person falls or jumps from a height, the trajectory is downwards and outwards, and the distance that the body strikes the ground from the jumping point is variable. Goonetilleke (1980) has published some research on how far from the wall a body is likely to land. Much depends on whether the victims fell passively from near the wall or projected themselves outwards at the top. The body may fall whilst maintaining the same orientation to the ground, but usually turns and twists in an unpredictable manner, the amount of alteration of posture partly depending upon the height of the fall and so the time available for turning. This means that the body may strike the ground in a number of different attitudes – and may also strike some obstruction part of the way down, making interpretation of injuries difficult.

The primary impact is usually the site of the most severe injury, but this is not always the case, as it may strike two areas simultaneously, such as the head and shoulder – or it may bounce or ricochet so that two or more major impacts occur in quick succession. The amount of kinetic energy acquired during the fall has to be fully expended by the time the body comes to rest so that, if only one impact occurs, it is likely to be more damaging than a series of lesser impacts, such as a bouncing, rolling strike.

If the body falls on to the head, there is likely to be a massive fracture, often (but not always) a scalp laceration and possibly extrusion of brain. Both vault and base can fracture and sometimes the base is driven down over the cervical spine, the latter projecting into the posterior fossa. The latter injury is, however, more common with high falls onto the feet when the impact is transmitted up the spinal column and the upper vertebrae – together with a ring of bone around the foramen magnum – are intruded into the skull, causing the classic ‘ring fracture’ of the occipital bone. Where the fall is onto the feet, the deceleration stress can break the axial skeleton at a number of points. The legs can be broken at any point, at tibial or femoral level, often bilaterally. The femoral necks can be snapped off, the hip joints can dislocate and over-ride, or the pelvis may fracture. The latter is often through the sacroiliac joints, the upward force (or more accurately, downward force of the body on to the legs) driving the sacrum down as a wedge into the pelvis.

If the lower limbs and pelvic girdle remain intact, the transmitted force may then fracture the spine, often at mid- or upper thoracic level.

Where the fall occurs onto the side of the body, any combination of injuries can occur. Multiple rib fractures, shoulder girdle or arm fractures, lacerations of back, buttocks or limbs and severe abdominal injuries can occur, with consequent internal lesions, such as rupture of the liver, lungs, heart or spleen.

In falls from high-rise buildings, the injuries can be extreme, as in a suicide from the twentieth floor seen by the author (BK), where the victim fell onto a fence and was completely transected at waist level.

It must also be remembered that biological and circumstantial variability allows for some remarkable escapes from falls; some persons, including children, have fallen from great heights yet have virtually walked away unscathed. Once again, it is very unwise to over-interpret the relationship between observed injuries and the likely length of the fall.

FRACTURES OF THE SKULL

Forensic anatomy

The cranium and facial bones are laid down from membrane in fetal life. The anterior fontanelle closes functionally between 9 and 26 weeks after birth, though is not tightly sealed until about 18 months. The posterior fontanelle closes between birth and 8 weeks of age. Suture lines close by interdigitation during childhood and osseous fusion occurs irregularly at variable dates during adult life.

The adult cranium consists of two parallel tables of compact bone called the ‘diploë’, the outer being about twice the thickness of the inner. They are separated by a central zone of soft cancellous bone, which is often misnamed the diploë. This zone is interrupted at suture lines and vanishes where the bone becomes particularly thin, especially in the floor of the skull.

The cranium varies in thickness in adults and varies from place to place, thin plates being reinforced by stronger buttresses, such as the petrous temporal, the greater wing of the sphenoid, the sagittal ridge, the occipital protruberance and the glabella. This tensile architecture of the skull has been well described and illustrated by Rowbotham (1964). The more vulnerable thin areas lie in the parietotemporal, lateral frontal and lateral occipital zones.

The average frontal and parietal thickness in a young male is between 6 and 10 mm. The thinnest area is in the temporal bone, where it may be only 4 mm, while in the occipital bone in the midline it may be 15 mm or even more. The thickness of the skull is sometimes an issue in courtroom propositions about the special vulnerability of a victim but, unless it is abnormally thin, such theorizing has little forensic relevance as it is well known that fatal brain damage can often occur with an intact skull. It has even been claimed that a thin skull is less likely to fracture,

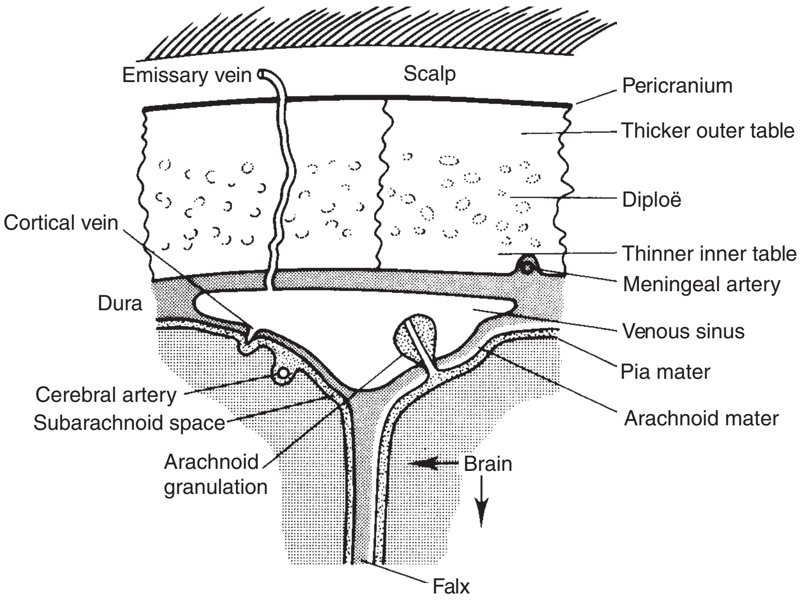
Emissary vein
Scalp
Pericranium
Thicker outer table
Diploë
Cortical vein
Thinner inner table
Meningeal artery
Dura
Venous sinus
Pia mater
Cerebral artery
Subarachnoid space
Arachnoid mater
Arachnoid granulation
Brain
Falx

FIGURE 5.13 *Forensic anatomy of skull and meninges.*

as it is more flexible and can return to normal after distortion without cracking; this is certainly true of the skulls of infants. This may be irrelevant when assessing intracranial damage, however, which is the major issue in head injuries.

It is rarely the skull fracture itself that is a danger to life, but the concomitant effect of transmitted force upon the cranial contents. The presence of a skull fracture is, however, an indication of the severity of the force applied to the head and it is uncommon for a head injury that is sufficiently severe to crack the skull not to cause some intracranial effect, even if it is only transient concussion, though, once again, there are many remarkable exceptions to this generalization.

In skeletal material, fractures of the skull are often seen and the differentiation between ante-mortem injury and post-mortem damage can sometimes be difficult or even impossible in the absence of any soft tissues. Artefactual damage to a skull may be caused during recovery or exhumation, and even stones in the soil can cause erosion and even cracking of the softened, degenerate bone after long burial. Zuo and Zhu (1991) have described scanning electron microscopic details of microfractures and collagen damage, which can differentiate ante- from post-mortem injury, but these are unlikely to assist in old decayed material.

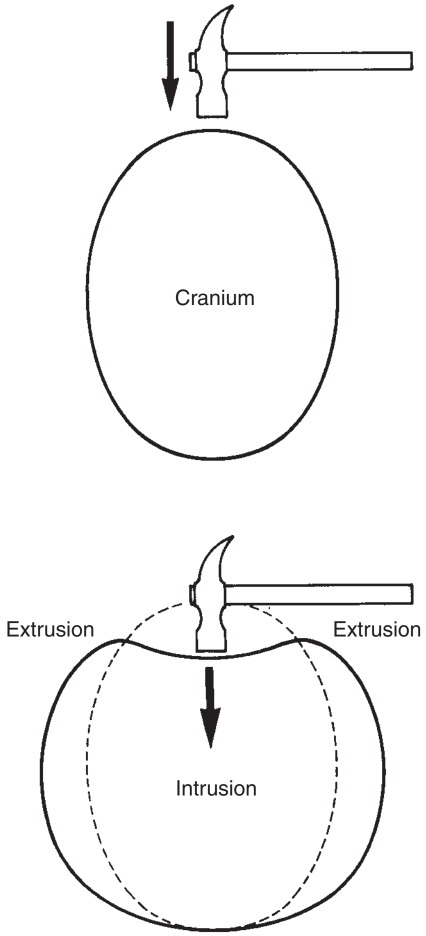
Cranium
Extrusion
Extrusion
Intrusion

FIGURE 5.14 *The 'struck hoop' illustration of impact on the skull. There is a momentary deformation, with the area of impact bending inwards and compensatory bulging elsewhere. This sets up stresses in the inner and outer tables.*

The mechanics of skull fracture

This subject has been extensively studied in living animals, isolated human heads and dried skulls. For details, the writings of Gurdjian, Webster and Lissner (1949, 1950), Weber (1984) and Rowbotham (1964) should be consulted;

a concise summary has also been provided by Shapiro *et al.* (1988). These and other workers have shown that:

- When the skull receives a focal impact there is momentary distortion of the shape of the cranium, the extent of which may be surprisingly large, though transient. Infant skulls, which are more pliable and have flexible junctions at suture lines, may distort much more than the more rigid skulls of adults. The area under the point of impact bends inwards and, as the contents of the skull are virtually incompressible, there must consequently be a compensatory distortion or bulging of other areas – the well-known 'struck hoop' analogy.

  Both these intruded and extruded areas can be the site of fracturing if the distortion of the bone exceeds the limits of its elasticity.
- When the skull is deformed, compression occurs on the concavity of the curved bone and tension (tearing) forces on the convexity. If the latter exceed the elastic threshold, then fracturing takes place. Thus the inner table will fracture where the skull is indented and the outer table will fracture at the margins of the deformed area. If the forces are great enough, a depressed comminuted fracture will occur.
- In the more common circumstance of a wider impact from a blunt injury, the deformation of the skull is less localized but, where the force is sufficient, fractures can still occur from the same mechanism of exceeding the elastic limits. The fractures may be remote from the area of impact, following lines of structural weakness – or may extend from the area of impact, or even commence at a distance and run back to the impact site.

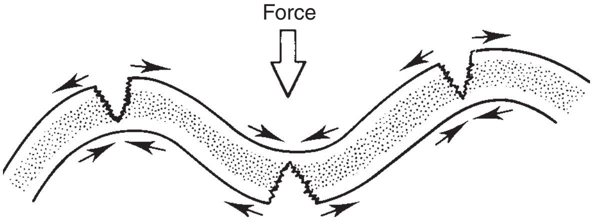

FIGURE 5.15 *The skull is more susceptible to traction forces than compression, so that convexities tend to fracture during the distortion of the 'struck hoop'.*

By using skulls coated with a brittle varnish, Gurdjian showed that stress lines developed in the cranium when it was struck and that these corresponded with the fractures that occurred with heavier impacts.

Blows in certain areas of the skull constantly give rise to fractures in specific localities – for instance, impact on the upper temporal or parietotemporal areas cause fissured fractures running obliquely downwards across the temporal area. If heavier, another fracture line tends to run obliquely contralaterally across the vault of the skull.

A heavy impact on the side or top of the head often leads to the vault fracture running into the base of the skull, usually across the floor of the middle cranial fossa along the anterior margin of the petrous temporal bone, to enter the pituitary fossa. In major injuries, this fracture line may often cross the floor of the skull completely to form a 'hinge fracture', separating the base of the skull into two halves. These fractures do not start at the point of impact unless there is also local

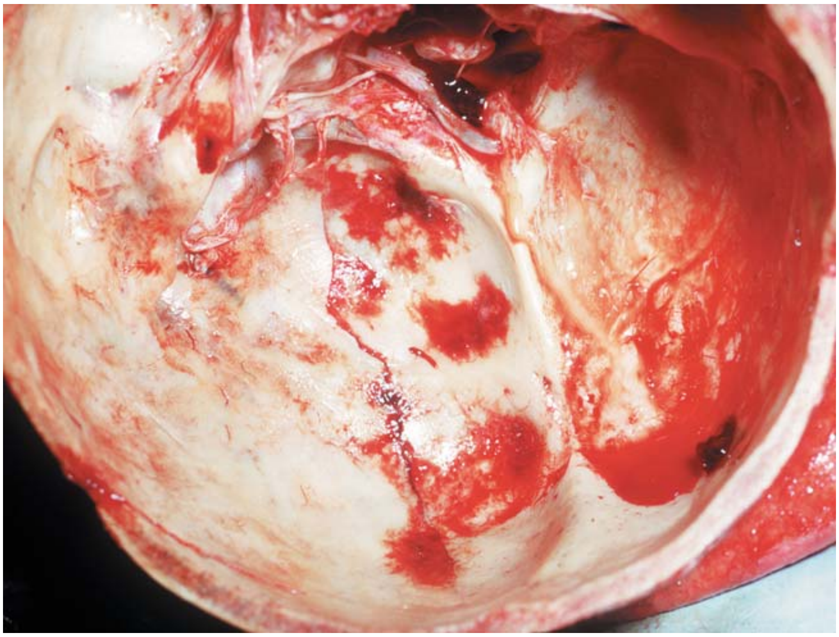

FIGURE 5.16 *Linear fracture of the posterior fossa, due to a fall on the occiput. The fracture typically crosses the thinner bone, avoiding the central buttress and ends near the foramen magnum.*

depressed fracturing; they are initiated at a distance due to the compensatory deformation, but usually run back towards the impact site.

When the frontal area is struck, the usual course for a linear fracture is vertically down the forehead, turning around the orbital margin to run backwards across the floor of the anterior fossa, perhaps into the cribiform plate or air sinuses, or both. A blow or fall onto the occiput may produce a fracture that typically passes vertically or obliquely downwards just to the side of the midline of the posterior fossa, commonly reaching the foramen magnum. In addition, the contrecoup element of an occipital fall may cause fractures of the orbital plates in the anterior fossa, as a result of transmitted force through the brain itself, though the mechanism is not fully understood.

- When severe local impact causes focal and general deformation, a combination of depressed fractures and radial fracture lines may form a 'spider's-web' pattern.
- When the focal impact is severe, the depressed fracture may follow the actual shape of the impacting object, such as a hammer-head. The shape may follow only that part of the object that drives into the skull – for example, the circular head of the hammer may strike at an acute angle, so only a semicircle of bone will be punched inwards, the opposite edge sloping downwards from an irregular crack.

  The deepest part of the depression will indicate where the weapon first struck; there may be 'terracing' of the margins. The author (BK) has seen a recent instance where a victim was struck with a hammer through a thin plastic bag enveloping the head. Parallel terracing cracks had opened for a fraction of a second during the blow, sufficient to trap lines of plastic into the multiple defects in the outer table. Where the impact is from a narrow edge or ring, only the outer table may be fractured, being punched into the softer centre without depression of the inner table.
- The presence of hair and scalp markedly cushions the effects of a blow, so that a far heavier impact is required to cause the same damage, compared to a bare skull. The pattern and nature of the skull fractures are, however, the same. It should be noted that the interposition of scalp and hair may slightly alter the dimensions of the skull lesion from a focal blow. For example, the padding effect of the scalp may add a few millimetres in diameter to the depressed fracture caused by a hammer, compared with the actual measured diameter of the hammerhead.
- Where two or more separate fractures occur from successive impacts and meet each other, the sequence of injuries may be determined by 'Puppe's rule', which is really eponymous common sense. The later fracture will terminate at (that is, not cross) the earlier fracture line, which naturally interrupts the cranial distortion which precedes fracturing (Figure 5.17).

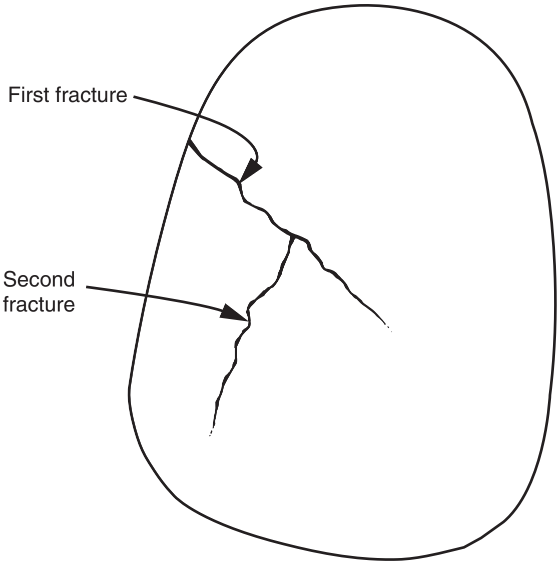

FIGURE 5.17 *Diagram to illustrate Puppe's rule for the sequence of fractures. The course of a later fracture will be interrupted by an earlier pre-existing fracture line.*

Types of skull fracture

Arising from the mechanisms described above, it is conventional to classify skull fractures as follows.

LINEAR FRACTURES

These are straight or curved fracture lines, often of considerable length. They either radiate out from a depressed zone, or arise under or at a distance from the impact area, from bulging deformation. They may involve the inner or outer table, but commonly traverse both.

Depending on the stress contours of that part of the skull and the localization of the impact, they may occur anywhere in the skull, but are especially common in the weak unsupported plates. The temporal, frontal, parietal and occipital plates may all carry single or multiple linear fractures. They may extend downwards into the foramen magnum, across the supraorbital ridges, or into the floor of the skull. A common basal linear fracture is one that passes across the floor of the middle fossa, often following the petrous temporal or greater wing of the sphenoid bone into the pituitary fossa. This frequently continues symmetrically across the other middle fossa separating the base of the

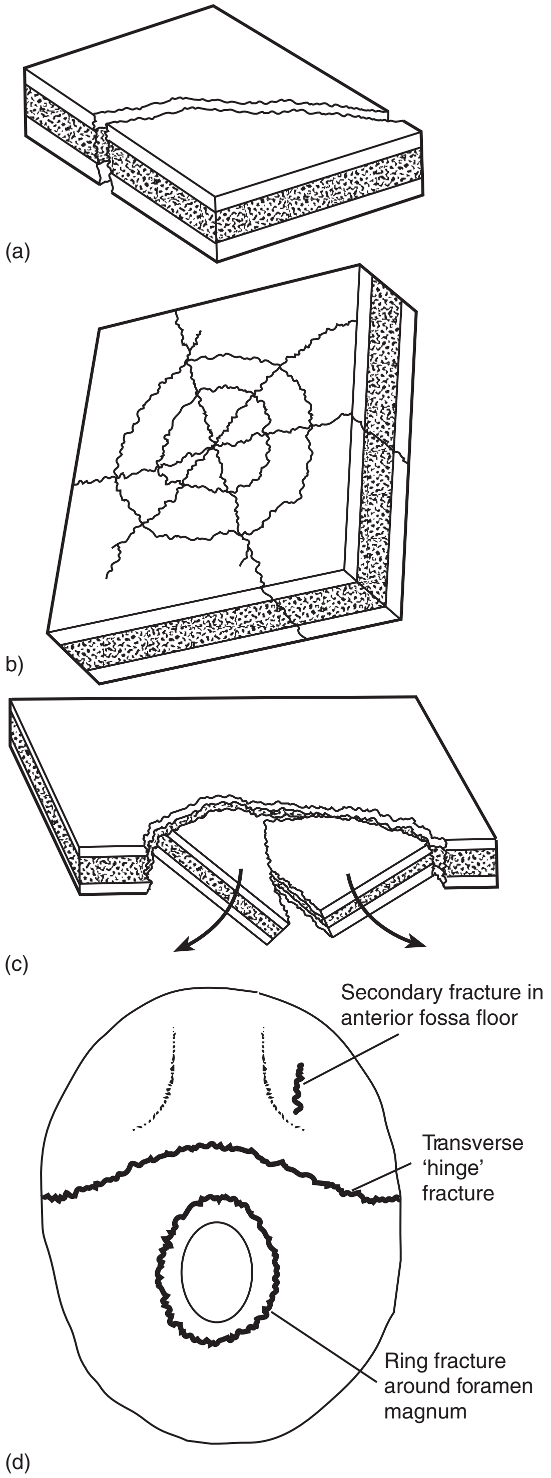
(a)
b)
(c)
Secondary fracture in anterior fossa floor
Transverse 'hinge' fracture
Ring fracture around foramen magnum
(d)

FIGURE 5.18 *Types of skull fracture: (a) linear fracture; (b) 'spider's-web' fracture; (c) depressed fracture; and (d) base fractures.*

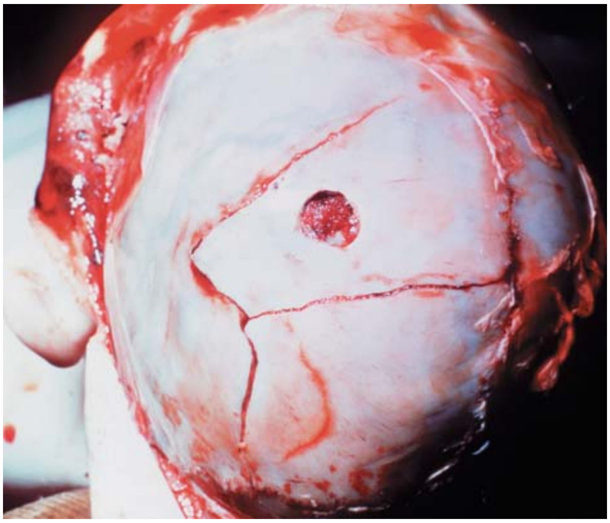

FIGURE 5.19 *Linear fractures of the skull in child abuse. Blunt impact near the vertex has caused a double fracture to run down the parietal bone and continue into the temporal area, the course of the fracture being determined by lines of stress and weakness in the skull. The central burr-hole was for the surgical evacuation of a subdural haemorrhage.*

skull into two halves, usually being caused by a heavy blow on the side of the head; this lesion was sometimes called the 'motorcyclist's fracture' for obvious reasons. Linear fractures may follow a horizontal course around the skull, usually from one temporal area to the other via the occiput rather than around the front.

In children and young adults, a linear fracture may pass into a suture line and cause a 'diastasis' or opening of the weaker seam between the bones. This is most often seen in the sagittal suture between the two parietal bones, but the interfrontal line of weakness left by the earlier fusion of the metopic suture can also reopen under mechanical stress. In infants, especially in the child abuse syndrome, a linear fracture of a parietal bone may reach the sagittal suture and continue across it into the opposite plate. The continuation may be direct or may be 'stepped', so that the two fractures are not in line. This appearance usually means a blow or fall onto the vertex, and the two fractures may be simultaneous but not continuous, explaining the 'stepping'.

RING FRACTURES

These occur in the posterior fossa around the foramen magnum and are most often caused by a fall from a height onto the feet. If the kinetic energy of the fall is not absorbed by fractures of the legs, pelvis or spine, the impact is transmitted up the cervical spine. This may be rammed into the skull, carrying a circle of occipital bone with it.

POND FRACTURES

This is merely a descriptive term for a shallow depressed fracture forming a concave 'pond'. It is more common in

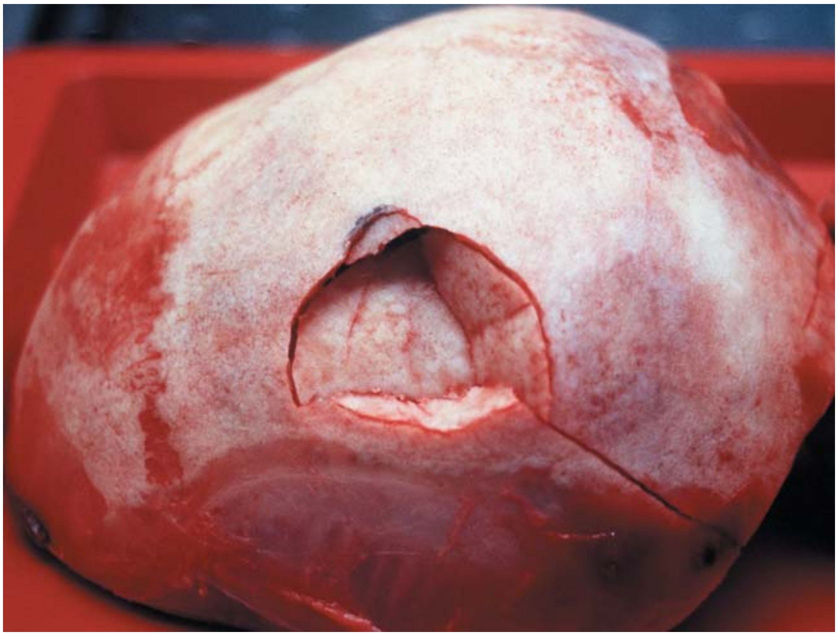

FIGURE 5.20 *Pond fracture.*

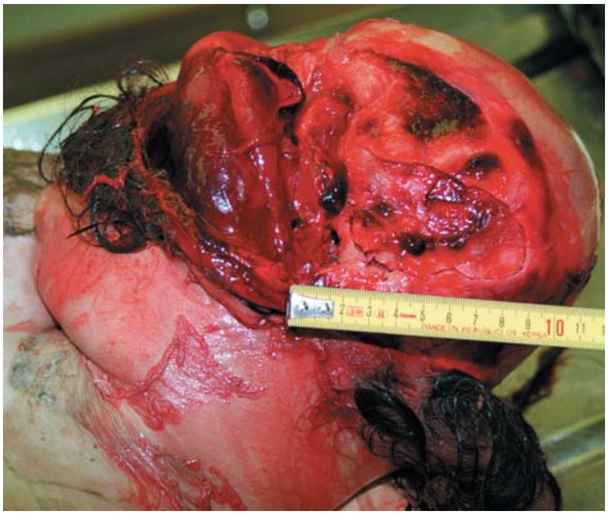

FIGURE 5.21 *Comminuted skull fracture with depression of the central area. This could also be termed a 'pond fracture' in that there is an ovoid depressed zone with radiating linear fractures. The head was struck with a heavy piece of wood along the line of the depressed area.*

the more pliable bones of infants and, indeed, a depression can occur in the absence of a fracture akin to the distortion produced by squeezing a table-tennis ball.

MOSAIC OR SPIDER'S-WEB FRACTURES

As already described, a comminuted depressed fracture may also have fissures radiating from it, forming a spider's-web or mosaic pattern. The degree of actual depression may be minimal or even absent.

DEPRESSED FRACTURES

Focal impact causes the outer table to be driven inwards and, unless absorbed in the diploë, the inner table will also usually be intruded into the cranial cavity with all the dangers of direct damage to the contents. Even sharp-edged weapons, such as heavy knives and axes, which may cause a clean-cut defect externally, usually split and deflect flaps downward from the inner table.

With axes and heavy cutting weapons such as swords, there is a characteristic lesion in the bones, whether skull or elsewhere. The initial impact slices cleanly through the bone on one edge, often burnishing the bone to an ivory-like gloss. The rebound removal of the weapon is at a slightly different angle, either from deliberate intent, or from relative movement between bone and blade. This cracks off an irregular fragment of bone of the opposite face so that the residual defect has one smooth and one rough edge. It is often seen in historical and archaeological material from battles or massacres.

The force required to cause fractures of the skull

Unlike less reliable subjective estimates of the force required to cause other injuries, objective quantitative measurements have obtained for adult skull fractures. Again, the publications of Gurdjian, Webster and Lissner should be consulted for detailed information. The following useful facts arise from their investigations:

- The tensile strength of the adult skull is of the order of 100–150 p.s.i., the compressive strength varying from

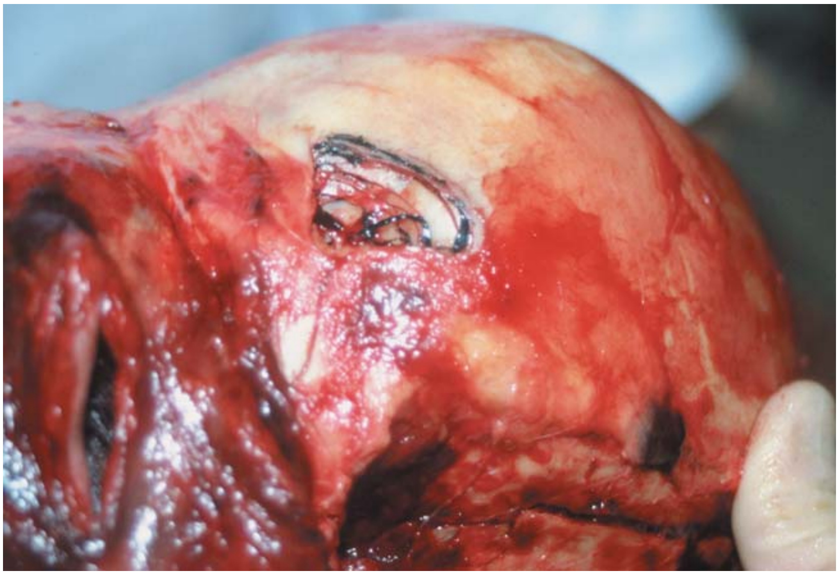

FIGURE 5.22 *A depressed fracture of the skull from a blow with a heavy club hammer. The defect is wedge shaped with a curved anterior border caused by the hammer striking at an angle. The depression is concentrically terraced, with the lowest fragments lacerating the surface of the brain.*

5000 to 31 000 p.s.i. (Gurdjian *et al.* 1949). A simple fissured fracture of the skull can be sustained by walking into a fixed obstruction (Kerr 1954). This requires a force of about 5 foot-pounds (73 N). Fast running into an obstruction produces about 70 ft-lb (1020 N). Falling to the ground from an erect posture also develops at least 60 ft-lb (873 N) and can easily produce skull fractures. A small stone or golf ball weighing about 100 g (4 oz) may also cause a linear fracture when thrown with moderate force against the temporal region (Graham and Lantos 2002).

- The adult head weighs between 3 and 6 kg (7–14 lb), averaging 4–5 kg (10 lb). When falling through about one metre (3 ft), so that the frontal area strikes a hard surface, impact energy of about 35 ft-lb (510 N) develops. This can cause one or two linear fractures or a mosaic fracture (Gurdjian *et al.* 1950).
- Only a small additional amount of energy above that needed to cause a single crack is required to produce multiple fractures.

In spite of these experimental data, it must never be forgotten that, like all biological phenomena, great variation is encountered and skull fractures, though they may be caused by as little as 5 ft-lb (73 N), may be absent when the impact exceeds 90 ft-lb (1314 N). The area of the skull struck, the thickness of the skull, scalp and hair, the direction of the impact and other imponderables, all affect the outcome.

As further discussed in Chapter 22, Weber carried out experiments in which he dropped the bodies of dead infants from a fixed height of only about 85 cm (34 inches) onto various hard and soft surfaces. A high proportion sustained skull fractures, both of the parietal and occipital areas, some of them crossing suture lines.

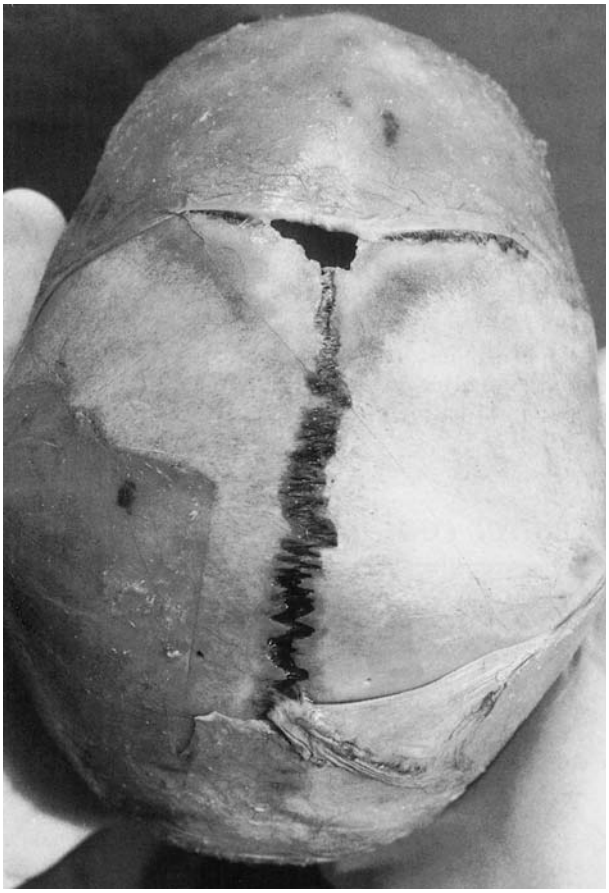

FIGURE 5.23 *Springing of the sagittal suture of the skull of an infant. Though there was no skull fracture, the suture line has widened after a fall that caused a subdural haemorrhage.*

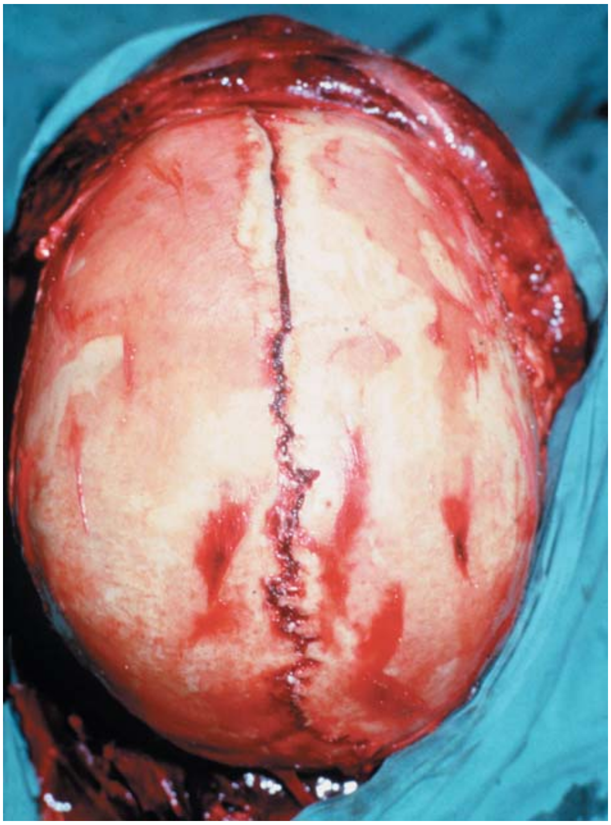

FIGURE 5.24 *Diastasis or 'springing' of the sagittal suture of a young adult who fell from a height onto the vertex of his head. The opening of the interparietal suture has extended anteriorly along the line of weakness of the fused metopic suture between the two halves of the frontal bone.*

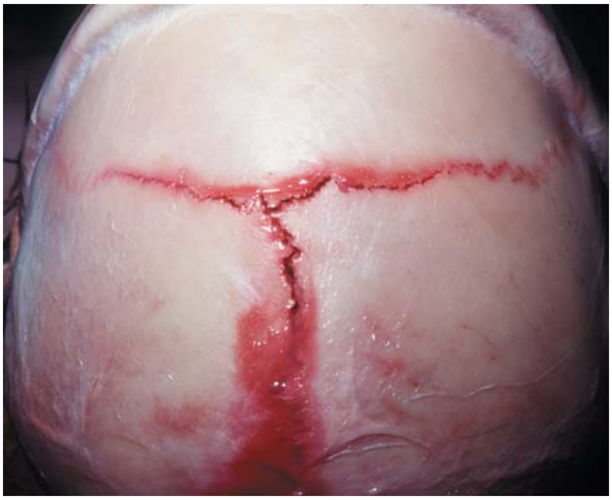

FIGURE 5.25 *Post-mortem diastasis of the sutures as an artefact due to freeze-up of the brain simulating intra-vital skull fracture.*

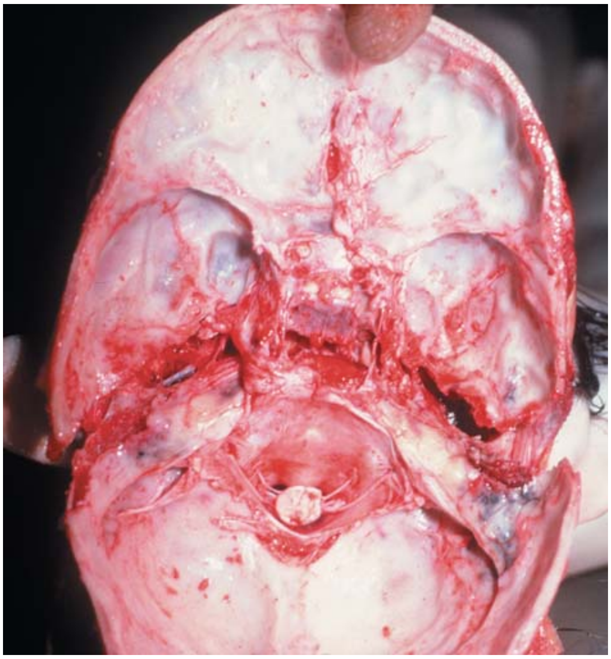

FIGURE 5.26 *A 'hinge' fracture of the base of the skull where the fracture line runs from side to side across the floor of the middle cranial fossa, passing through the pituitary fossa in the midline, following the course of least structural resistance. The victim was a young pedestrian who was hit by a car and then sustained secondary injuries by striking the left side of the head on the ground. This injury is common in road accidents and is sometimes called the 'motorcyclist's fracture'.*

Dangers of fractures of the skull

It has been emphasized that, in the majority of cases, the significance of a fractured skull is an indicator of a substantial insult to the head, with possible injury to the vital contents, rather than the fracture itself being a danger to life.

There are occasions, however, when the fracture itself has dangerous sequelae. The most common is when the crack passes through an embedded meningeal artery, causing a meningeal haemorrhage, which is considered later. A depressed fracture may impinge upon the brain and its membranes, and bone fragments may lacerate or penetrate the brain tissue.

TRAUMATIC EPILEPSY

A late effect of a depressed skull fracture may be 'traumatic epilepsy'. This is of great medico-legal significance, especially in the field of civil litigation where an accident or assault may result in lifelong neurological disability for which very large monetary compensation may be awarded.

Traumatic epilepsy usually manifests as tonic and clonic fits, which may be difficult to differentiate from idiopathic epilepsy, if the injury occurred in early life. When fits begin within weeks, or a year or two of a major head injury in a mature person who had never had fits before, the diagnosis is easier, but all cases need expert neurological examination. There must have been a substantial head injury, usually with a depressed fracture impinging on the underlying cortex, often in the parietotemporal area. The epilepsy appears usually within a range of a few weeks to up to 2 years. It is more common in open head injuries when infection has occurred, or when a spicule of bone has penetrated the meninges, as the mesodermal scarring of the cortex that results is more likely to irritate the brain than the astrocytic reaction found in closed head injuries.

INFECTION FOLLOWING SKULL FRACTURE

Other complications may occur, even in the absence of depression or comminution. The most common is infection of the meninges, or the development of a brain abscess, or both. Infection can gain access via skull fractures:

- by direct spread through a compound fracture, especially where there is a contaminated scalp injury;
- by spread from the nasal cavity when a fracture of the cribriform plate has allowed communication with the anterior fossa. This fracture is not uncommon as part of ‘contrecoup’ lesions, described later. Sometimes the cribriform plate is ruptured by objects entering a nostril and carrying infected material into the cranium: the author (BK) has seen two such fatalities, one from an umbrella ferrule and another from a dirty bamboo garden cane. Another case was seen by PS where a suicide attempt was made with a small calibre pistol. The bullet penetrated the right temple and went through the skull just above the anterior cranial fossa, which was fractured and caused a leakage of the cerebrospinal fluid through the nose. The victim recovered well but, in spite of prophylactic treatment with antibiotics, died one year later from purulent meningitis.
- by spread from fractures that involve a paranasal sinus, such as the frontal or ethmoid, or from the mastoid air cells or middle ear cavity. Basal fractures may allow these spaces, which communicate with the infected outside environment, to reach the meninges, especially when the latter are torn by the traumatic event. A history of leakage of cerebrospinal fluid from the nose or ear must alert both clinician and pathologist to the possibility of communicating basal fractures. In any autopsy on a suspected head injury, however slight, care must be taken to remove the dura from the interior of the vault and base so that a close inspection for fracture lines can be made.

INTRACRANIAL INJURIES

The contents of the skull are the most fragile of the vital organs, necessitating their enclosure in the strong bony box of the cranium. Damage may occur either to the neural tissues or to the rich vasculature that surrounds and penetrates those tissues.

FORENSIC ANATOMY OF THE BRAIN MEMBRANES

The pachymeninges consist of the dura mater and the leptomeninges, the arachnoid and the pia mater (see Figure 5.13). The **dura** is formed of two layers of tough collagenous tissue, the outer of which is firmly attached to the skull and acts as its internal periosteum. The inner layer merges with the arachnoid, so that in reality there is no true subdural space, only a potential cleavage plane.

The dura forms the falx and tentorium, and the cranial venous sinuses run within it. Branches of the meningeal arteries course over and through its substance. The dura is penetrated by bridging veins, especially along the vertex and at the tips of the temporal lobes, also to a lesser extent at the frontal and occipital poles, as well as by random vessels elsewhere. Polypoid invaginations of the dura penetrate the inner walls of the venous sinuses, especially the sagittal sinus, to form the ‘arachnoid granulations’.

The **arachnoid** is a thin, vascular meshwork that is intimately applied to the inner surface of the dura by means of the ‘boundary layer’ so that no subdural space exists in normal conditions, though their junction is so tenuous that they are easily split apart. Sheaths of arachnoid follow vessels into the brain as they penetrate into the neural surface. These vessels and thin strands of connective tissue anchor the brain within the subarachnoid space. This is filled with cerebrospinal fluid and the width of the space varies from less than a millimetre in the young to a centimetre or more in the old, in whom cerebral atrophy has developed. This means that the anchoring strands and the bridging vessels are longer and more vulnerable to shearing and rotatory stresses. Even though anatomically in the subarachnoid space, rupture of these bridging vessels often manifests itself in the subdural space.

The **pia** is not a true membrane, but a surface feltwork of glial fibres that are inseparable from the underlying brain.

EXTRADURAL HAEMORRHAGE

Also known as 'epidural haemorrhage', bleeding between the inner surface of the skull and the dura mater is the least common of the three types of brain membrane haemorrhage. According to Rowbotham (1964), only about 3 per cent of head injuries have an epidural haemorrhage large enough to be of surgical significance; a similar figure of between 1 and 3 per cent was recorded by Tomlinson (1970). Of the 635 fatal head injuries investigated by Adams (see References and further reading) in Glasgow, 10 per cent had extradural haemorrhages. The mortality rate, even with surgical intervention, averages about 11 per cent under the age of 20 years, rising to between 18 and 40 per cent in later life.

The dura is closely applied to the interior of the skull, forming the endocranium or periosteum. It is so tightly applied to the base of the skull that, except in the posterior fossa, extradural bleeding does not occur over the skull floor. In the vault there is a potential space between the dura and the bone, which can be separated by both arterial and – less often – venous leakage. Most extradural haemorrhages are associated with fractures of the skull, but about 15 per cent occur in intact skulls (McKissock 1960). According to Harwood-Nash *et al.* (1971), the incidence in children without a fracture is only 1 per cent, though Adams found that half the children with extradural bleeds in his series of over 600 fatal head injuries had no fracture. About 10 per cent of extradural haemorrhages are associated with subdural haemorrhages. Bilateral epidural haemorrhages are rare, but have been recorded.

The usual site is unilateral in the parietotemporal area, caused by rupture of a branch of the middle meningeal artery where the latter is transected by a fracture line. The posterior branch of this vessel is most commonly involved as it courses diagonally backwards across the squamous temporal bone on the lateral wall of the cranium. The anterior (frontal) branch is rarely the source of bleeding, occurring only twice in Rowbotham's series of 33 cases. The vessel usually lies in a deep osseous groove in the first part of its course. It has been claimed that almost all ruptures occur where the artery is completely roofed over in a bony tunnel so that it is unable to escape damage from a fracture, but observation does not confirm this contention.

Leakage of the high-pressure arterial blood strips back the underlying dura with progressive accumulation of a haematoma, which can reach a volume of several hundred millilitres and cover an appreciable part of the hemicranium. Adams suggests that a minimum volume of 35 ml is needed before clinical signs are apparent, though other writers suggest 100 ml is usually the minimum associated with fatalities.

In the less common occipital and frontal sites, smaller branches of meningeal arteries may be involved or the bleeding may be from torn venous sinuses, in which case there need not be a fracture. When bleeding is venous, the haematomas rarely reach a large size as the pressure is insufficient to tear back much of the dura.

The clinical signs of an epidural haemorrhage are classically those of a 'lucid' or 'latent' interval, as there may be recovery from the initial phase of concussion (see below) before sufficient blood accumulates to cause raised intracranial pressure and consequent relapse into unconsciousness. This classic picture is so frequently absent, however, that no diagnostic reliance can be placed upon it. The coma from the increasing space-occupying lesion formed by the bleed may follow the period of concussion without a break so that there is no temporary phase of recovery. Only 27 per cent of McKissock's series showed the classic history.

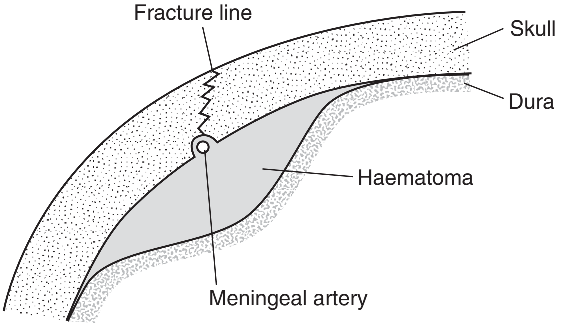

FIGURE 5.27 *Formation of an extradural haemorrhage.*

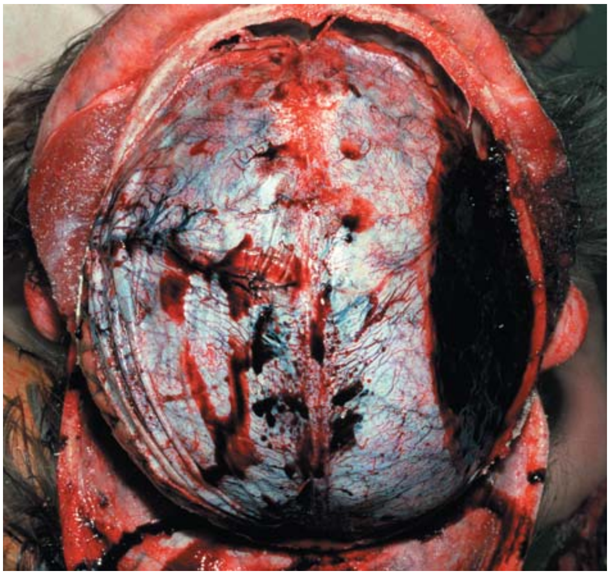

FIGURE 5.28 *A large temporoparietal extradural haemorrhage. The under surface of the scalp on the right shows bruising at the point of impact. There was a linear fracture of the skull passing through the right middle meningeal artery.*

The latent interval may be of variable duration and, as just noted, may not occur if either concussion is prolonged or other brain damage coexists. Half an hour may be enough for the formation of a significant arterial haematoma, though some slow bleeds have taken more than a day to become clinically apparent. In Rowbotham's series, the range was from 2 hours to 7 days, but most were apparent after 4 hours.

More recent investigations using computed tomography, quoted by Adams, suggest that the old concept of worsening clinical symptoms being caused by the progressive accumulation of blood is incorrect, as computed tomograms show that the major volume of blood may appear soon after the injury and that, as with many space-occupying lesions, clinical signs result from other factors, such as cerebral oedema or diffuse neuronal injury.

Medico-legal considerations of extradural haemorrhage

A medico-legal danger is that a victim may be discharged from the care of a doctor or hospital casualty department when he recovers from his transient concussion only to deteriorate and perhaps die at home; negligence may then be alleged against the unsuspecting doctor. Unfortunately, even when the diagnosis is made, the results of surgical intervention are not good, there being a fatal outcome in more than half the cases operated on. Part of the reason for this poor prognosis is that many victims of extradural haemorrhage also have other damage such as cerebral contusion.

When a victim of a head injury recovers from concussion and then lapses into coma within the first 24 hours, the differential diagnosis is between an extradural or subdural haemorrhage, or cerebral fat embolism, especially if there is other skeletal damage. At autopsy the lesion is self-evident, as most haematomas lie in the temporal or parietal areas, and are opened when the usual saw-cut is made to remove

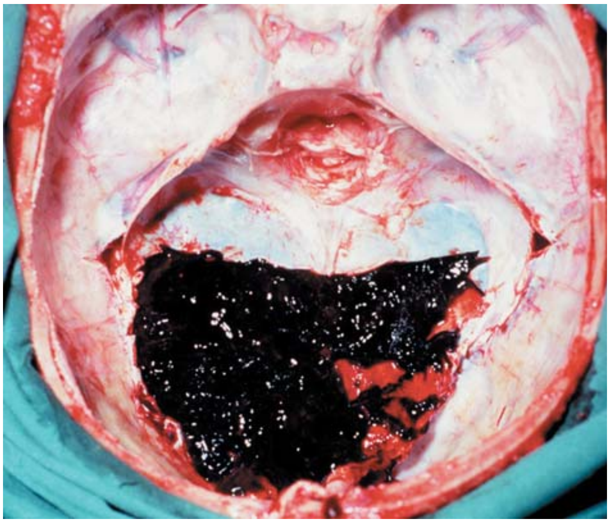

FIGURE 5.29 *An extradural haemorrhage in the posterior fossa. This is an unusual site for this type of lesion, as most occur in the temporal or parietal area as a result of tearing of the middle meningeal artery. In this case a fracture line passed down the posterior fossa towards the foramen magnum, but no bleeding point could be identified.*

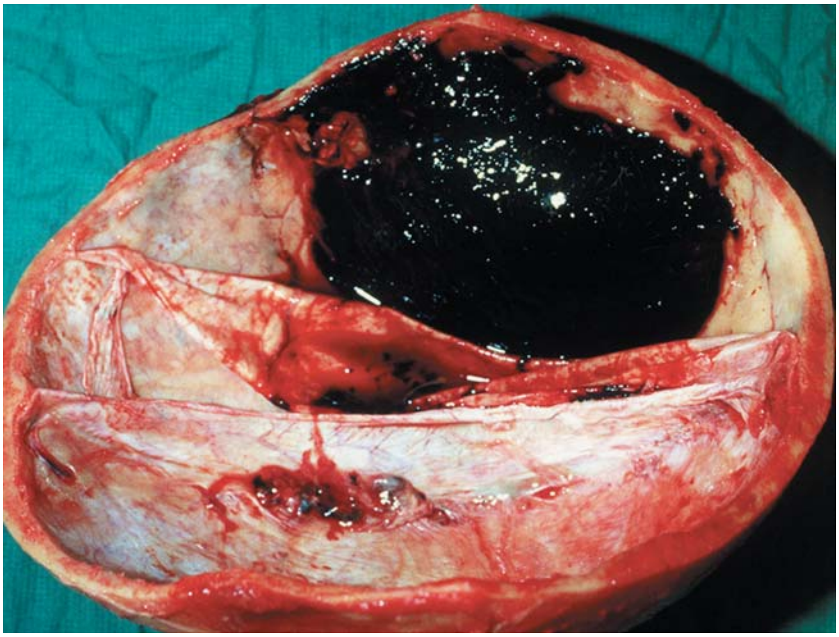

FIGURE 5.30 *A fresh extradural haemorrhage in the temporoparietal area from a fracture crossing the middle meningeal artery. About 85 per cent of such haemorrhages are associated with skull fractures.*

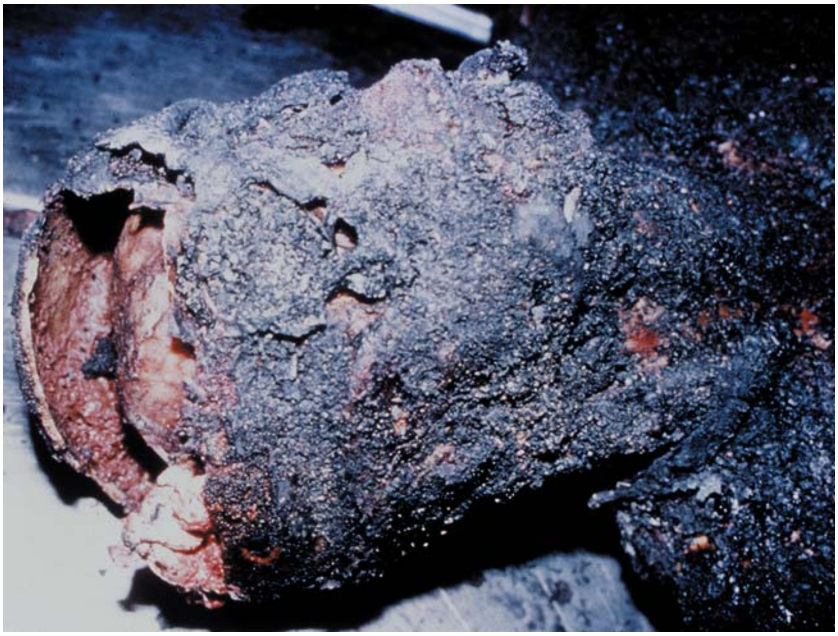

FIGURE 5.31 *Fire victim with extensive defect of the carbonized skull showing reddish-brown 'heat haematoma' on the inner surface of the skull and shrunken brain with brain tissue oozing out through a split in the dura.*

the calvarium. If in the posterior fossa, the dark blood is seen through the raised dura. Wherever it is situated the dimensions and approximate volume should be measured and an estimate made of the inward projection. The brain surface will be flattened or otherwise distorted if the haematoma is of significant size and may have given rise to the usual appearances of raised intracranial pressure, described later. The midline of the brain may be shifted laterally if the bleed is large.

Extradural haemorrhage is never a 'contrecoup' injury, this being purely a cerebral tissue lesion.

HEAT HAEMATOMA

One well-known artefact mimics an extradural haemorrhage. When a head has been exposed to severe external heat sufficient to burn the scalp and perhaps the skull, blood may be extruded from the diploë and venous sinuses into the extradural space to produce a 'heat haematoma'.

The mechanism is obscure, but may be the result of blood being 'boiled' from the diploic layer of bone through emissary veins, or shrinkage of the brain may aspirate blood from the skull. The false haematoma is brown and friable, and the adjacent brain shows hardening and discoloration from the heat.

The importance of the artefact is that it may be mistaken for a true epidural bleed from a head injury, and may mislead the pathologist and investigators into thinking that the fire was started criminally to cover up a fatal assault. As most instances are seen in conflagrations in buildings, there is often a significant level of carboxyhaemoglobin in the body if the death occurred when the fire was in progress. This should be of the same concentration in the heat haematoma as in the peripheral blood; if the victim suffered a head injury before the fire started, then there should be little or no carboxyhaemoglobin in the haematoma.

SUBDURAL HAEMORRHAGE

Bleeding beneath the dura is much more common than extradural haemorrhage. It is also proportionately less often associated with a fractured skull, but in absolute numbers far more fractured skulls cover subdural than extradural haemorrhages. The Glasgow series of 635 fatal head injuries described by Adams included 18 per cent of subdural haematomata.

The lesion is traditionally classified into three types: the acute, the subacute and the chronic. It is unhelpful to subdivide the acute type, however, and only acute and chronic haemorrhage need be considered.

Subdural haemorrhage can occur at any age, but is common at both extremes of life. It is one of the major causes of fatal child abuse and the rediscovery of that syndrome by Caffey (1946) consisted of an association of subdural haemorrhage with long bone fractures. In old people they commonly exist in a chronic form and can be mistaken either for 'strokes' or for senile dementia. The condition is always due to trauma and there is probably no such entity as 'spontaneous subdural haematoma'. Even in states of vascular fragility, such as in senility and in bleeding diatheses, some minimal trauma must precipitate the bleeding even if it was too trivial to be recorded in the history. It is almost certain that minor subdural bleeds, insufficient to give rise to any neurological or clinical symptoms or signs other than a transient headache, occur with the trivial knocks of everyday life. Only when the bleeding is extensive enough to become either a cortical irritant or a space-occupying lesion (probably between 35 and 100 ml) does it become clinically apparent. Of course, many subdural haemorrhages exist in combination with both subarachnoid bleeding and with cerebral damage, making its contribution to the overall symptomatology impossible to assess.

As with the extradural haematoma, the position of a subdural can never be interpreted as a 'contrecoup' lesion and is thus of no use in differentiating a blow from a fall.

Acute subdural haemorrhage

This is a common sequel to any substantial head injury, and the presence or absence of a fracture is immaterial except as an indicator of trauma to the head. Unlike extradural bleeding, a fracture plays no part in the pathogenesis of the haemorrhage, which arises from torn communicating veins that traverse the subdural space between the cortical vessels and the dural sinuses. Less often the sinuses themselves give rise to the haemorrhage.

Naturally in an open head injury or when comminuted fractures penetrate the membranes and perhaps the brain itself, subdural bleeding is merely part of a complex that includes subarachnoid bleeding, and cerebral laceration and contusion.

The lesion is often pure, however, being associated with a closed head injury where the only other signs may be scalp bruising – or even nothing at all, as blunt impacts may leave no signs in the scalp, externally or internally, and no skull fracture.

The latter situation is probably the explanation in most of the cases formerly attributed to shaking of the infant. Many paediatricians and pathologists have enthusiastically adopted the shaking aetiology when there was no overt sign of impact (or sometimes even where there was such evidence!) to such an extent that it is frequently proffered as the favoured diagnosis. However, the concept of the shaken-baby subdural has been strongly challenged recently, as it has been shown that the shearing force (required to rupture subdural vessels) is of the order of 50 times less in shaking than in impact (Duhaime *et al.* 1987). Thus it is very probable that perhaps the majority of allegedly shaken babies have, in fact, had an occult head impact, which has not left any signs on the scalp, subscalp tissues or skull.

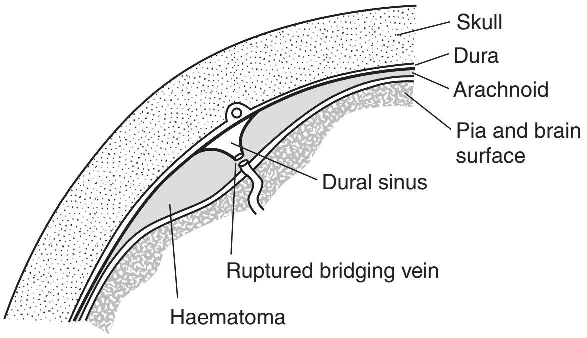

FIGURE 5.32 *Formation of a subdural haematoma.*

Subdural bleeding arises from shear stresses in the upper layers of the cerebrum, which moves the communicating veins laterally sufficiently to rupture their junctions at either the cortical veins or the sinus surfaces. It is very rarely possible to identify the bleeding points. Subdural bleeding is most often over the lateral surface of a cerebral hemisphere, high up in the parasagittal area.

As with most intracranial damage, the mechanical cause is a change of velocity of the head, either acceleration or deceleration, almost always with a rotational component. Where a blunt impact strikes the skull, the subdural bleed need not be situated directly under the impact area – it need not even be on the same side of the head. It is sometimes tempting to attribute a localized subdural to either a 'coup' or to a 'contrecoup' effect (see below), but this is an unsafe interpretation. In addition, a subdural haemorrhage – unlike an epidural – is quite mobile. Lesions that have obviously originated high on the parietal area commonly drain down under gravity and cover the whole hemisphere, with a large accumulation in the middle and anterior fossas, and even through the tentorial opening into the posterior fossa.

The haemorrhage may remain fluid or may clot into a firm mass: both modes are commonly present. If the bleeding is relatively slight, then a 'thin-film' subdural may be found. If the thickness of blood is less than a few millimetres, it cannot be claimed to be a space-occupying lesion, even if the area covered is quite large, as the cerebrospinal fluid in the adjacent subarachnoid compartment can be displaced sufficiently to accommodate an equivalent volume of blood.

Again, Adams suggests that, as in any intracranial space-occupying lesion, a minimum volume of about 35 ml is required to cause neurological signs, though other writers prefer a larger volume, such as 100 ml.

Whether this film of fresh blood is a sufficient irritant to cortical activity – as it undoubtedly is in a subarachnoid haemorrhage – to be a danger to life, is uncertain. It is difficult to claim that a thin layer of blood in the subdural

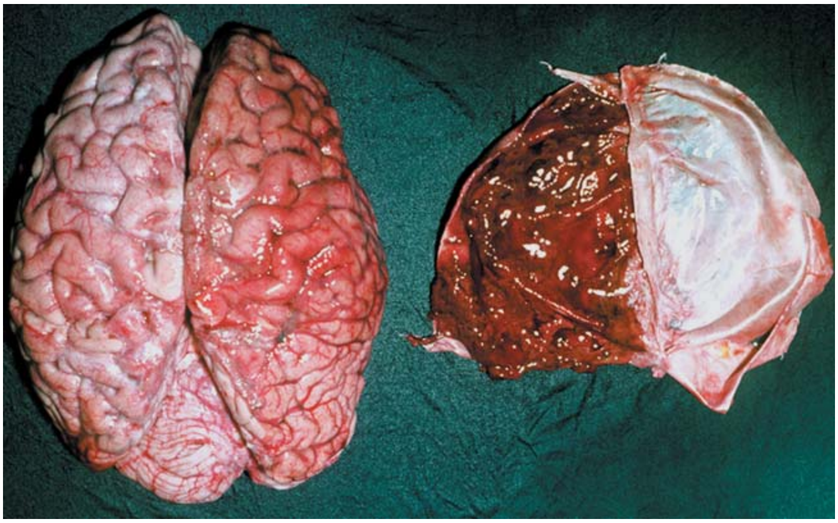

FIGURE 5.33 *Chronic subdural haemorrhage in an old person. Brown liquid escaped from the encapsulated lesion adherent to the meninges leaving a gelatinous outer membrane, as seen. The surface of the right cerebral hemisphere is stained brown from old altered blood and there is some compression of the hemisphere with a midline shift to the left. There was no history of a head injury and no significant neurological deficit.*

space is the sole cause of death, but almost inevitably the force that caused the bleeding will also have had a deleterious effect on the brain tissue, even if this is macroscopically occult such as diffuse axonal damage.

As with the extradural haemorrhage, there may be a latent interval before clinical signs and symptoms appear. After the almost inevitable concussion, which may be very brief, the victim may recover, then relapse into deepening stupor and coma when intracranial pressure rises as the subdural bleeding proceeds. Associated brain damage may, however, cause uninterrupted coma from the time of injury.

When there is a lucid interval, this may be longer than the average 4 hours of the faster arterial bleeding of the epidural haemorrhage. In fact, there is no upper limit to this interval, as the acute subdural haemorrhage merges into the chronic condition, which may recur after weeks or even months. In severe acute bleeds, as are commonly encountered in criminal cases, such intervals tend to be short or non-existent.

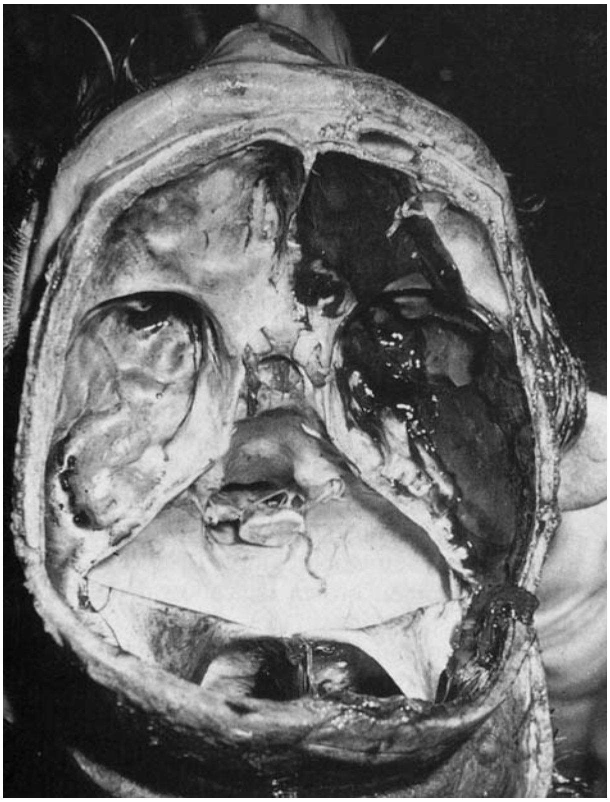

FIGURE 5.34 *A chronic subdural haemorrhage discovered as an incidental finding at an autopsy on an old person dying of an unrelated cause. The blood was brown and gelatinous, but there was no membrane.*

Chronic subdural haematoma

This lesion is most often found in old people, frequently as an incidental finding at autopsy where death was caused by some unrelated condition.

The gross appearance varies with age; recent lesions up to several weeks old are tan or red-brown with a gelatinous membrane covering the surface. The contents are thick but liquid and may have areas of redder, more recent bleeding. An older haematoma, up to months or even a year old, is firmer, with a tough membrane around both surfaces, resembling a rubber hot water bottle filled with jelly or oil. The contents are liquid and may be brown or even straw coloured. Sometimes the interior may be much firmer and variegated in colour due to bleeds of different ages. Loculation is common, with different coloured fluids or ooze in each locule. The underlying brain will be depressed if the haematoma is large (more than 50–100 ml), and is often stained brown or yellow from altered haemoglobin.

Many such subdural bleeds in old people are small and obviously gave rise to no neurological abnormality; the mere finding of such a small lesion at autopsy should not be used to provide a cause of death, which should be sought elsewhere in the body. Some subdurals are substantial, however, reaching a volume of 100–150 ml. They may still have been asymptomatic, but others give rise to neurological symptoms that may have been ascribed to some other pathological cause. Examples are 'strokes', when unilateral signs were thought to be caused by a cerebral thrombosis or haemorrhage; disordered behaviour of the old person may have been blamed upon senile confusion or dementia, whereas in fact a space-occupying haematoma was really responsible.

The chronic haematoma may become large and press down on the cerebral hemisphere sufficiently to dent and distort the surface. This may progress, as does the large acute haemorrhage, to cause signs of hippocampal and cerebellar tonsillar herniation and all the attendant dangers to the vital centres in the brainstem. The chronic haematoma arises from the acute lesion, which, after an interval, becomes sheathed in a capsule of connective tissue. The haematoma may eventually absorb, it may remain dormant at the same size, or it may enlarge at any later date.

The mechanism of the enlargement is controversial. One common explanation, which seems the most reasonable, is that it occurs from repeated further bleeding, perhaps from new blood vessels that penetrate the mass as part of the healing process. The other theory involves osmosis, said to operate because the centre of the haemorrhage commonly liquefies, forming a haemorrhagic fluid that osmotically attracts into it the cerebrospinal fluid from outside the capsule, which acts as a semipermeable membrane.

The first mechanism seems more likely, as areas of fresh bleeding are often found inside a substantial haematoma but, whatever the cause, the final effect is a worsening of the space-occupying effect.

The dating of a subdural haemorrhage

An estimation of the date of onset of a subdural haematoma may have considerable forensic significance, especially if the lesion is obviously mature. There may have been one or more episodes of trauma on record, any of which may have criminal or civil connotations and the opinion of the pathologist will be sought to relate or to eliminate the lesion found from the potential causative event.

For example, in one of the author's cases an old lady was struck on the head by intruders and died several weeks later. At autopsy, a large chronic subdural haematoma was the main finding, but the defence sought to show that the haematoma must have been present before the assault by claiming that the woman's confused behaviour over the previous year was caused by the pre-existing intracerebral lesion. Furthermore, in accidents of any type, there may be doubt about whether the subdural haematoma arose as a result of the injury, or whether a pre-existing lesion caused unsteadiness that may have precipitated the accident.

Unfortunately, in spite of several claims to reliable methods of dating subdural haematomata, such estimations are of doubtful value, partly because repetitive bleeding results in varying ages within the same haematoma.

A subdural haematoma gradually changes from dark red to a brownish colour, first being apparent not before 5 days and sometimes not obvious for 10–12 days (Crompton 1971, 1985).

Reaction to subdural bleeding begins within a few hours of onset, when cellular infiltration begins from the dural surface. A delicate 'neomembrane', histologically composed of thin-walled capillaries and fibroblastic granulation tissue, grows from the periphery to cover the outer (dural) surface of the clot during the next few days and weeks. If no further enlargement occurs, this capsule becomes more and more fibrous, though rarely does it completely absorb the haematoma by fusing with the outer capsule. According to Crompton, the presence of a membrane firm enough to be picked off with forceps makes the subdural haemorrhage at least 12 days old.

Although claims have been made for accurate dating of subdural haematomas by histological criteria (Munro and Merritt 1936), these cannot be depended upon, especially after a few months, as there is considerable personal variation in healing rates. Also, because of the frequency of repeated subsequent fresh bleeding, attempts at estimating the date of the original bleed are unrealistic. As they are of some use as an approximate guide to age, Munro and Merrit's histological criteria are summarized briefly here:

- Fibroblasts appear at the margin of the clot within 36 hours and in 4 days the neomembrane adjacent to the dura is a few cells thick. From 5 to 8 days, the membrane becomes well established and fibroblasts migrate from it into the clot.
- By 8 days the membrane is 12–14 cells thick and is visible to the naked eye. From a few days after onset, there is progressive red-cell lysis and after 5 days haemosiderin-laden phagocytes are present, which may be stained by Perl's reaction.
- From 11 days the clot is subdivided by strands of fibroblasts. By 15 days a membrane is also present on the under surface of the clot and the outer neomembrane is half to one-third the thickness of the dura itself. By day 26 it equals the thickness of the dura, but the inner membrane is still only half as thick.

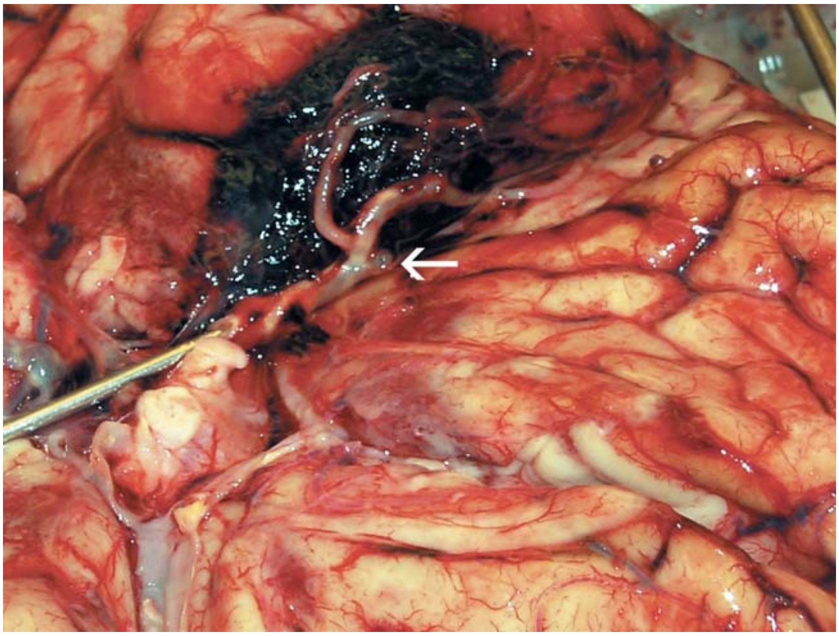

FIGURE 5.35 *Subarachnoid haemorrhage from a tiny berry aneurysm (arrow) of the right middle cerebral artery.*

- Between 1 and 3 months the membrane has lost many fibroblast nuclei and is becoming hyaline. By 6–12 months the membrane becomes thick and fibrous, resembling the dura itself.

A number of later writers have pointed to marked variations in this chronological scheme. Within the first couple of months large sinusoidal vessels appear in the newly formed connective tissue.

From gross appearances, a rough guide is that brown colour changes occur between the first and second week, when a discrete surface membrane becomes obvious. After a month or so a firm capsule develops, forming a cystic cavity containing dark brown, watery fluid. According to Munro, liquefaction of the contents does not occur in less than 3 weeks. Some haematomas remain solid with an organizing blood clot within, often with areas of fresher haemorrhage of different ages.

When a substantial subdural bleed has occurred, the underlying cerebral cortex may be infarcted, caused by either a natural infarct having bled through the arachnoid into the adjacent subdural space or, more often when there has been a head injury, the pressure of the haematoma on cortical blood vessels (Crompton 1971, 1985).

SUBARACHNOID HAEMORRHAGE

The third type of brain membrane bleeding is even more common than subdural haemorrhage, but has a mixed aetiology. Whenever there is damage to the cortex, there will be some degree of subarachnoid bleeding, so all penetrating injuries of the brain, as well as many blunt injuries that give rise to extradural or subdural haemorrhage, will be associated with traumatic subarachnoid bleeding.

It is not common, though not unknown, for traumatic subarachnoid bleeding to occur as a pure lesion where there is no cortical contusion, no neck injury, no deep brain lesion and no other membrane haemorrhage. Slight subarachnoid bleeding probably occurs very frequently after a moderate impact upon the head (as with slight subdural bleeding), but as the vast majority of such victims survive, no autopsy evidence is forthcoming. The increased sensitivity of medical imaging techniques, especially nuclear magnetic resonance, can demonstrate such minor bleeds which hitherto went undetected.

The other complicating issue with subarachnoid bleeding is that it frequently occurs as a result of natural disease, especially rupture of vascular malformations of several types. When trauma is also present, the complex association of either the trauma precipitating the rupture or a rupture causing a fall or other accident leading to the trauma, has to be considered. This is discussed in a later paragraph.

Although the pathology of subarachnoid haemorrhage is of prime forensic importance, because of the relationship to ruptured berry aneurysm, much of the description is offered in the chapter on sudden natural death (Chapter 25), and only the interaction between trauma and this type of bleeding is discussed here.

Appearances and mechanism of formation

The appearance of subarachnoid haemorrhage caused by trauma varies greatly according the nature and extent of the

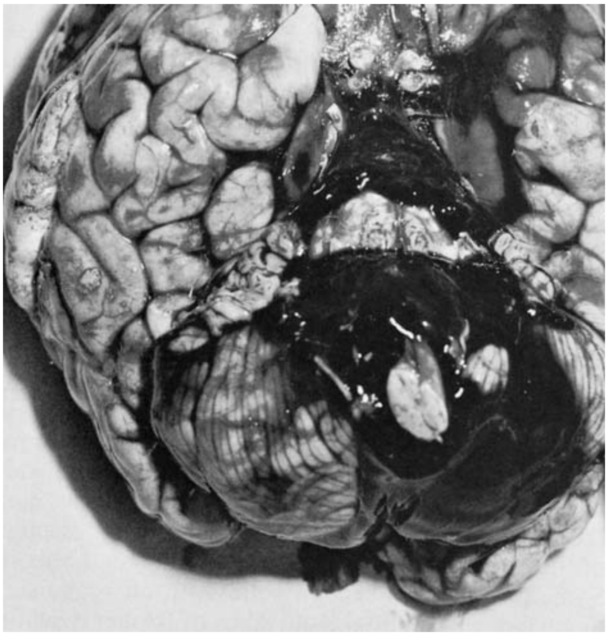

FIGURE 5.36 *An extensive basal subarachnoid haemorrhage arising from a traumatic tear of a basilovertebral artery caused by a blow on the side of the neck.*

injury. Where it is secondary to laceration of the brain or extensive cortical contusion, then its localization and severity depend upon the primary injury. Where it arises from a blunt impact, with or without other membrane bleeding or cortical bruising, its position is not a good localizing sign. Once again, its position cannot be interpreted in the same way as contrecoup contusion of the cortex, though some pathologists use its position to claim that a head injury was sustained either from a fall or from a blow to a mobile head; this is unjustified. Although where the circumstances are known, the meningeal bleeding may be seen in the correct location, this is fortuitous and is equally often sited elsewhere.

Blood in the subarachnoid space mixes with the cerebrospinal fluid, which dilutes it, makes it less ready to clot and allows more mobility. Thus bleeding high over the cerebral hemispheres readily slides down to cover the brain and enter the basal skull fossae, but usually not in a concentration sufficient to form a thick clot. The sulci tend to collect more blood, especially in the insula. Unlike those with subdural bleeding, survivors from subarachnoid haemorrhages rapidly dispose of the blood. Haemolysis turns the cerebrospinal fluid a xanthrochromic yellow and, within weeks, the blood is gone. There may be some residual brown or yellow staining of the pia or arachnoid, similar to that seen when a slight dural bleed leaves its signature on the inner surface of the dura. A positive Perl's reaction for haemosiderin can develop within 36 hours and may remain for months or even years, even after macroscopic discoloration of the membranes has vanished.

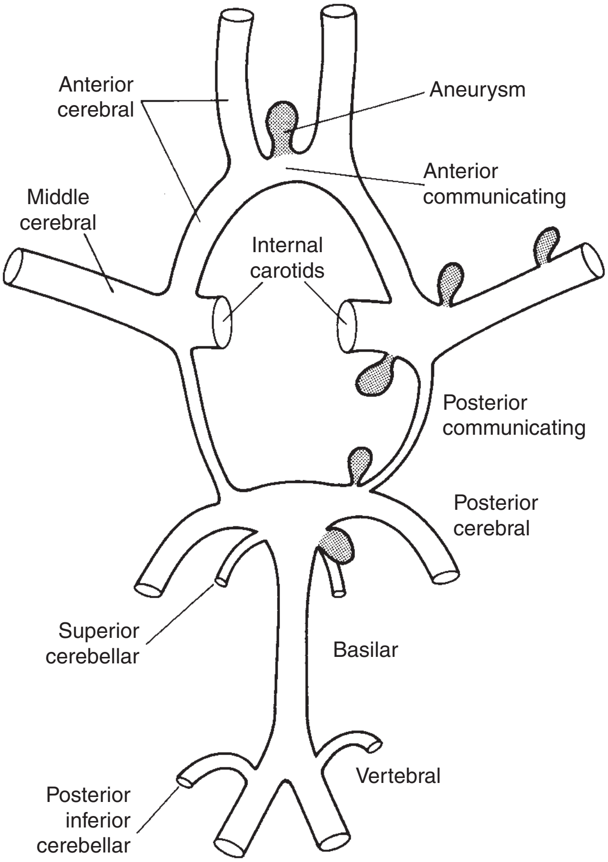

FIGURE 5.37 *Cerebral arteries ('circle of Willis') and common sites for aneurysms.*

This forms a trap for those attempting to date both subdural and subarachnoid haemorrhage in infants, as the common bleeding arising from head-moulding during childbirth may leave a positive haemosiderin reaction for at least a year. Thus if a suspected child abuse head injury occurs in the first months of life, histological attempts to date the bleeding by Perl's reaction may be affected by residual haemosiderin from parturition.

Bleeding in the subarachnoid space is caused by the same mechanism as that in the subdural space, as shear stresses and rotational movements of the brain, described in a later section, shear or rupture the bridging veins that leave the cortex and penetrate the arachnoid en route for the large draining veins and sinuses that lie in the dura. In addition, small cortical arteries may contribute much of the leakage. Where laceration, contusion or infarction of the cortex is present, bleeding will come from cortical veins and small arteries directly into the subarachnoid space. It may also arise from intracerebral bleeding breaking out through the cortex. The subarachnoid haemorrhage allegedly caused by damage to the vertebral arteries near the foramen magnum will be discussed under spinal injuries.

The relevance of a subarachnoid haemorrhage in causing death may be difficult to decide, though this is of more importance in the natural variety caused by a ruptured aneurysm.

In head injuries, death is far more likely to be the result of the other concomitant associated injuries to the brain substance than to a moderate amount of blood in the subarachnoid space.

In the uncommon instances when a substantial subarachnoid haemorrhage seems to be the only sequel of a fatal head injury it must be accepted as the cause of death, though cerebral oedema and microscopic diffuse axonal injury should be excluded as competitors even though this exclusion may be difficult or impossible when death follows rapidly after injury. It is indisputable, however, that sudden death can occur from massive natural subarachnoid haemorrhage in the absence of any traumatic lesions so, by analogy, the pure traumatic bleed – if severe enough – cannot be excluded as a sole cause.

Death can be remarkably rapid when a profuse haemorrhage occurs into the subarachnoid space. The mechanism is not understood, but seems to be confined to those cases where the brainstem is suddenly exposed to a large volume of blood bathing the brainstem in the posterior fossa, usually from a ruptured aneurysm or from a torn basilar or vertebral artery. Numerous instances of victims virtually dropping dead are on record, though many more exhibit the familiar signs of headache, neck stiffness, vomiting and progressive failure of consciousness, before dying at a variable period after the haemorrhage.

FORENSIC IMPLICATIONS OF BRAIN MEMBRANE HAEMORRHAGE

Because of its common occurrence, certain medico-legal problems associated with bleeding into the brain membranes must be considered.

First, the latent or ‘lucid’ interval in both extradural and subdural haemorrhage may lead to civil suits for negligence where a doctor – usually in a hospital emergency department – discharges an apparently well patient who later dies at home or is readmitted **in extremis.** The clinical aspects cannot be considered here, but suffice to say that it is highly unlikely that such patients would not have suffered concussion, albeit transient. Radiology of the skull is not mandatory if the clinical indications are not present, but failure to X-ray can be a legal impediment for the defence. A subdural haemorrhage is commonly present with an intact skull, though this is uncommon with extradural bleeding.

As already mentioned, the dating of subdural haematomas may be crucial in both criminal and civil cases when it is being either maintained or denied that a particular episode of head injury causes the haematoma. Dogmatic adherence to any accurate histological dating regimen is unjustified, though broad distinctions in terms of days, weeks or months may be justified.

The possibility of a previous bleed being the precipitating factor in some later accidental injury, which then caused further bleeding, must be kept in mind.

Using a membrane haemorrhage to localize an impact can be subject to over-interpretation. An extradural haematoma is almost invariably on the same side as the blow, but because of the distortion of the skull described earlier, the fracture that traverses the meningeal artery can be some distance from the point of impact. Using Gurdjian’s work, however, a reasonable estimate of the zone can be offered.

Where subdural or subarachnoid bleeding is concerned, it can be misleading to offer dogmatic opinions about the nature and position of the impact, unless there is scalp or skull damage. Though sometimes it is clear from the presence of a scalp injury that the haemorrhage is either ‘coup’ or ‘contrecoup’, there are so many exceptions when the bleeding is unrelated to the site of the blow, that a firm opinion is not justified.

RUPTURED BERRY ANEURYSM AND TRAUMA

A major cause of medico-legal problems is the association of trauma and subarachnoid haemorrhage from a ruptured ‘berry’ aneurysm. The pathology of the latter is described in another chapter, but here we are concerned with the allegations that an injury led to rupture of a pre-existing aneurysm.

Several variations of this scenario exist. Most commonly, an assault is rapidly followed by the signs of subarachnoid bleeding and subsequently death. Occasionally death may be extremely rapid, as previously described. The question of causation then arises, the prosecution claiming that the head injury mechanically ruptured the aneurysm. Whether this is actually so is virtually beyond absolute proof. Aneurysms of the circle of Willis are remote from the skull surface and lie deeply protected under the buffer provided by the mass of the brain. However, it is hard to deny that a heavy blow to the head, jaw or neck could not rupture, split or weaken the fragile wall of a large, thin-walled aneurysm, but when the bleeding comes from a tiny sessile bulge the evidence is not so convincing.

The complicating factor is that most assaults occur in 'fight or flight' conditions in which both aggressor and victim are physically and emotionally active, so that the adrenal response is likely to be present. Muscle tone, heart rate and blood pressure are increased by catecholamines, and it is likely that raised internal blood pressure in a weak aneurysm is a far more potent reason for rupture than a blow on the head.

A second possibility is that a person with an already spontaneously leaking aneurysm may have a rapidly developing neurological or even behavioural abnormality that leads him into conflict with another person, or into a dangerous physical position, such as a fall or traffic accident. If all this occurs over a short period of time, autopsy may not be able to distinguish the sequence of events and the aneurysmal rupture may be blamed on the trauma instead of the reverse. This may have profound civil as well as criminal legal consequences.

SUBARACHNOID HAEMORRHAGE AND ALCOHOL

A factor that is often said to increase the chance of rupture is alcohol, though there is no objective proof of this claim. A high blood alcohol is said to facilitate bursting of an aneurysm because it dilates cerebral blood vessels, increases cerebral blood flow and raises blood pressure.

The latter is not true because alcohol does not raise systolic pressure, though the pulse pressure – the difference between systolic and diastolic – may widen. The fibrous wall of an aneurysm is incapable of dilating, neither can the major basal arteries do so to any appreciable degree, as they possess little muscle in their walls. The pharmacological evidence that alcohol has any significant effect on the cerebral circulation is extremely weak. There is no evidence that alcohol is associated with completely natural subarachnoid haemorrhage from a ruptured aneurysm, though intense physical activity, such as sport or coitus, certainly does predispose to rupture.

Where alcohol and ruptured berry aneurysm are concerned, a more likely explanation is that the association is coincidental – as most altercations resulting in 'flight, fight and fisticuffs' are catalysed by alcohol. Many of the cases of ruptured aneurysm and violence occur within and on the pavements outside bars and clubs, where a high blood alcohol is virtually inevitable – but not necessarily causative.

The other aspect of alcohol is that it may cause or contribute to unsteadiness, a fall or some other traumatic event, which itself might lead to rupture of a fragile aneurysm. Ataxia and hypotonia are a feature of acute drunkenness and, together with the aggressive behaviour and alcoholic environment of most physical violence, seem sufficient to account for their circumstantial association.

THE RAPIDITY OF DEATH IN SUBARACHNOID HAEMORRHAGE

The immediacy of death is sometimes surprising, especially in association with trauma. It is general clinical experience that the victim of a spontaneous rupture of a berry aneurysm suffers a severe headache, neck stiffness and vomiting, which may resolve or which may progress over hours or days to coma and death. Sudden death is certainly the exception in these clinical cases – but these are from a hospital population, which, by definition, excludes the rapid deaths that reach the mortuary rather than the wards.

In forensic practice, much more rapid demise is not uncommon. For example, the author (BK) recollects that as two men involved in a violent argument in a public house emerged through the outer doors, one struck the other a heavy blow on the head. The victim fell to the ground and never moved again, being certified dead a few minutes later. At autopsy, a typical fresh subarachnoid bleed from a ruptured aneurysm was displayed.

The amount of blood present in most subarachnoid haemorrhages, from whatever cause, often seems insufficient to constitute a space-occupying lesion, especially as, unlike subdural bleeding, it is able to diffuse more widely over the brain and displace cerebrospinal fluid into the spinal theca. The total volume may be considerable, however, and where survival for some hours has occurred, then the typical appearances of raised intracranial pressure may be seen at autopsy, though some of this may be contributed by progressive cerebral oedema.

Most rapid deaths exhibit substantial bleeding into the basal cisterns at autopsy, the brainstem and cranial nerve roots being bathed in a thick layer of blood and clot. As blood in the subarachnoid space seems irritant even in small quantities, it seems possible that such sudden irrigation of the medulla may lead to a rapid cardiorespiratory failure. Where the haemorrhage appears less extensive, the mechanism of death is more obscure.

Crompton points out the sensitivity of cerebral arterioles during surgical operations, when a slight touch upon the cortex leads to blanching spasm of the vessels. Subarachnoid bleeding is undoubtedly irritant and a possible reason for the sudden collapse and death is a widespread vascular spasm, which may have an effect on vital centres in the brainstem.

ROTATIONAL TRAUMA TO THE HEAD AND UPPER NECK: BASILOVERTEBRAL ARTERY INJURY

During the last few years it has been recognized that blows to the side of the neck and/or head can give rise to fatal subarachnoid bleeding. This has been attributed to tearing or dissection of a vertebral artery, allowing blood to track along the upper part of the vessel and enter the cranial cavity where the artery penetrates the dural membrane at the foramen magnum.

The evolution of this concept of traumatic subarachnoid bleeding is interesting. Originally, it was thought that most cases had a fracture of the transverse process of the first cervical vertebra, thus damaging the artery contained in the tunnel-like foramen in this bone – indeed, it was often called the ‘CV-One syndrome’.

The hypothesis was that a blow to the side of the neck caused deep injury, usually manifested by skin bruising and bleeding into the muscles deep in the upper neck. This same injury was considered to fracture the transverse process and lead to tearing or dissection of the wall of the contained vertebral artery; blood then tracked upwards and medially to penetrate the dura and emerge inside the subarachnoid space of the posterior cranial fossa to cause a fatal haemorrhage.

Soon cases were being described with no such atlas vertebra fracture, but it was still accepted that mechanical forces, especially of a tilting and rotational nature, could damage the vertebral artery within the foramina of the upper cervical spine and lead to the same dissecting lesion, and hence to fatal subarachnoid haemorrhage. Until very recently, this was the theory – which is still strongly held by many pathologists – that satisfied the aetiological needs of this injury.

However, though it is likely that this classical mechanism does account for some of the deaths, there are several problems with universal acceptance of the hypothesis:

- Many pathologists – including the authors – find it hard to believe that a tiny dissection of a small artery like the vertebral artery – a lesion which often needs serial microscopic sections to confirm – could allow the torrent of blood necessary to enter the cranial cavity and produce a massive haemorrhage (sometimes well in excess of 100 ml) which may cause death within minutes.
- Several instances of a damaged vertebral **vein** have been reported in association with subarachnoid haemorrhage. Because of the very much lower intravascular pressure in a vein compared with an artery, the possibility of this causing a massive intracranial bleed by percolation through the dural fenestration is even less credible than in the case of the artery.
- Increasing numbers of cases are being described with positive evidence of a blow on the side of the neck or head and a large fatal subarachnoid haemorrhage, but with completely intact vertebral arteries.

Thus, if sudden tilting and rotational forces acting on the upper spine and head can lead to subarachnoid haemorrhage without damage to the vertebral vessels, then the causative link is destroyed. Even where vertebral damage is demonstrable, this may be merely a **concomitant** event, the mechanical forces which primarily led to subarachnoid haemorrhage by direct injury to intracranial vessels also having caused the vertebral artery damage. In other words, the two lesions may in some cases – or for all we know, every case – be a **parallel** phenomenon, not a **cause-and-effect** situation. There may well be a further concomitant effect, in that direct occult brain damage, manifested later by diffuse axonal injury, is the major cause of cerebral dysfunction, both the subarachnoid haemorrhage and the vertebral artery lesion (if there is one) being merely markers of a heavy impact.

In these deaths – and certainly where no lesion at all is discernible in the upper spinal vessels – we need to look elsewhere for an explanation of the subarachnoid bleeding.

Occasionally, this is quite obvious and is situated in the intracranial vertebrobasilar vessels. For example, the author (BK) has seen a wide split in the wall of a vertebral artery **within** the posterior cranial fossa, and another where the vertebral artery was totally avulsed and transected at the internal side of the dural perforation. Both these cases arose as a result of trauma to the side of the neck, and both had a massive subarachnoid haemorrhage with no damage to vessels outside the cranium.

Bostrom, Helander and Lindgren (1992) published details of two cases of rupture of the posterior inferior cerebellar artery due to blunt basal head trauma, and proposed that the term ‘traumatic subarachnoid haemorrhage’ be abandoned and replaced by the nature and localization of the bleeding site.

The demonstration of intracranial bleeding from the vertebrobasilar system is difficult, as the very process of opening the skull at autopsy and removing the brain, however carefully performed, inevitably causes vascular damage. These artefacts cannot be distinguished from original bleeding points due to ante-mortem trauma.

The use of post-mortem angiography by the injection of radio-opaque constrast medium into the lower vertebral and carotid arteries has not lived up to its original claims. There is almost always diffuse leakage of contrast medium from apparently artefactual defects, and localization of true bleeding points is often obscured by a haze of contrast opacity. However, occasionally the method reveals a specific leak, usually within the posterior cranial fossa.

The original descriptions of and claims for vertebral artery damage as the cause of traumatic subarachnoid haemorrhage were published by Cameron and Mant (1972), Coast and Gee (1984), Contostavlos (1971, 1995), Simonsen (167, 1976) and others. Leadbeatter (1994) drew attention to the problems of a universal cause-and-effect relationship; Contostavlos (1995) refuted these claims, but did not satisfactorily dispose of the concomitant hypothesis. Pathologists, both from their dissection experience and reading of the literature must make up their own minds on the matter, but should make a critical evaluation of the practicality of a massive subarachnoid bleed appearing, sometimes within minutes, from a tiny dissection of a small artery outside the dural membrane.

As many pathologists still adhere to the original concept of a subcranial vertebral artery causation for traumatic subarachnoid haemorrhage (and in a small number of instances, the evidence is persuasive), the topic is further pursued here, with the over-riding qualification that most traumatic massive meningeal bleeds are due to intracranial causes consequent upon a tilting, rotatory impact upon the head and/or neck, with or without concomitant damage to a vertebral vessel.

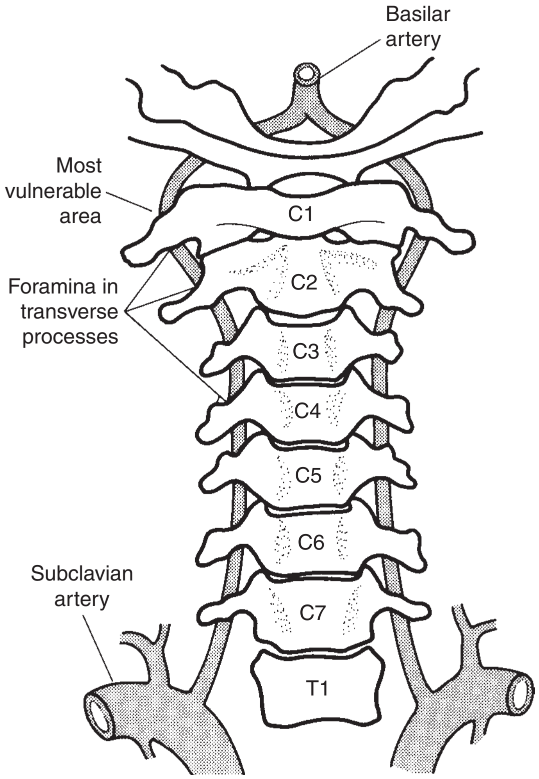

FIGURE 5.38 *The origin and course of the vertebral arteries. They arise from the subclavian arteries at the level of the sternoclavicular joints and ascent via the foramina in the transverse process.*

Forensic anatomy

The two vertebral arteries arise from each subclavian artery in the region behind the sternoclavicular joints. Each artery ascends behind the common carotid to reach the transverse process of the sixth cervical vertebra. It enters the foramen in that process and passes upwards through each similar foramen until it emerges from the upper edge of the second (axis) vertebrae. The artery then bends laterally and enters the final foramen in the atlas vertebrae. Emerging on the upper surface, the artery bends back and medially around the superior articular process, and penetrates the posterolateral aspect of the atlanto-occipital membrane and the underlying spinal dura and arachnoid, emerging on the lateral side of the spinal canal just below the foramen magnum. Both arteries then ascend and converge on the ventral surface of the medulla and pons to fuse in the midline to become the basilar artery. Each vertebral artery is often of markedly different size.

Autopsy appearances

With the caveats expressed earlier, the possibility of vertebral artery trauma, concomitant or otherwise, should always be borne in mind when an external bruise is seen on the side of the neck of the victim of a fatal assault. A blow from a fist, foot or blunt weapon may land in the region between the angle of the jaw to the side of the back of the neck, the area below the ear being the most common place to find an injury.

There may be no external sign at all, but on dissection of the neck, a subcutaneous or deep bruise may be found. Unfortunately, this is an area that is not routinely dissected at autopsy, the usual incision for the removal of the neck organs being too far anterior to reveal many of these injuries, which usually lie in the strong neck muscles. This is probably why the syndrome was not recognized until a few years ago – and also an explanation why so many subarachnoid haemorrhages were said to be due to a berry aneurysm that could not be found because the bursting had destroyed it. No doubt a number of subarachnoid bleeds are due to rupture of non-aneurysmal vessels or to an aneurysm too small to find, but equally, unrecognized head trauma, with or without vertebral artery damage, must have accounted for some of the remainder.

In some autopsies, there is no external neck injury and the cause of death is unknown until the cranium is opened, and a fresh subarachnoid haemorrhage discovered. A decision then needs to be made whether to remove the brain to search for a ruptured berry aneurysm – or begin specialized techniques to seek a rupture of a vertebral artery (described below). If there is a history of assault or other trauma, as opposed to a presumed natural death, then particular methods should be employed. The two avenues are not mutually exclusive, as long as the pathologist is aware of his objectives. The vessels can be clamped off at the base of the brain and the brain removed for a search for an aneurysm or other bleeding-point without spoiling the other techniques.

Mechanism of vertebral artery trauma

When a head is rapidly rotated by a blow that lands at the junction of the head and the neck, there may be a sudden lateral rocking (tilting sideways) at the atlanto-occipital joint, accompanied by rotation of the head. There may also be an element of hyperextension or hyperflexion, the whole episode forming a complex pattern of sudden abnormal movement at the atlanto-occipital junction. It may be that the unexpected impact may allow more unrestrained rotation and angulation of the head, due to absence of anticipatory muscle tensing in the large paravertebral and sternomastoid musles; this may be exacerbated by alcoholic intoxication causing slow protective responses, as most of such episodes occur during altercations related to drinking sessions.

The mechanics are not fully understood and probably differ from case to case but, whatever the mechanism, the vertebral artery can become damaged either:

- in the canal within the first cervical vertebra (whether or not the transverse process is fractured)
- just below the axis, in the space between the transverse processes of the axis and atlas
- as it emerges from the exit of the canal in the atlas to penetrate the spinal dura just below the foramen magnum
- probably much more frequently within the subarachnoid space above the foramen magnum or even higher in its course towards the confluence with the basilar artery – and even the basilar itself. The type of damage is usually a tear or dissection of the wall of the vertebral artery.

Vertebral artery damage was originally said to allow blood to track under arterial pressure within the adventitia and to appear in the subarachnoid space after the vessel has penetrated the dura and arachnoid, though, as discussed above, the likelihood seems slight of these minute lesions allowing the large volume of blood seen in most subarachnoid haemorrhages to flow through the small periarterial window in the dura.

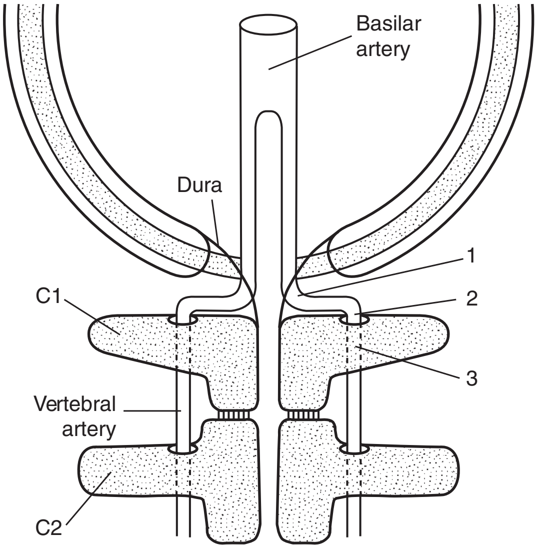

FIGURE 5.39 *Most vulnerable points for trauma to the vertebral arteries: (1) at penetration of dura; (2) at exit from atlas transverse process; (3) in osseous canal in atlas.*

Autopsy demonstration of basilovertebral artery damage

When circumstantial evidence suggests a subarachnoid haemorrhage following trauma or where a bruise is seen on the side of the neck, vertebral artery damage should be suspected, then confirmed or eliminated. The first intimation that it may have occurred may be when the skull-cap is removed and a subarachnoid haemorrhage is discovered.

If the view is taken that most subarachnoid haemorrhage following upper neck trauma is due to intracranial vascular damage, then logically it is unnecessary to use time-consuming and laborious procedures which slow up the completion of the case. In many such deaths, there will be no lesion in the upper cervical spine or extracranial vertebral arteries and even if there is, it is likely to be a concomitant lesion, occurring synchronously, but with no cause–effect relationship with the intracranial vessel rupture.

However, there may be academic satisfaction in demonstrating the concomitant neck lesion, even if it played no role in producing the subarachnoid haemorrhage. In such a case, when the deceased has been in a fight or had some violence applied to the side of his neck, the same routine should be employed.

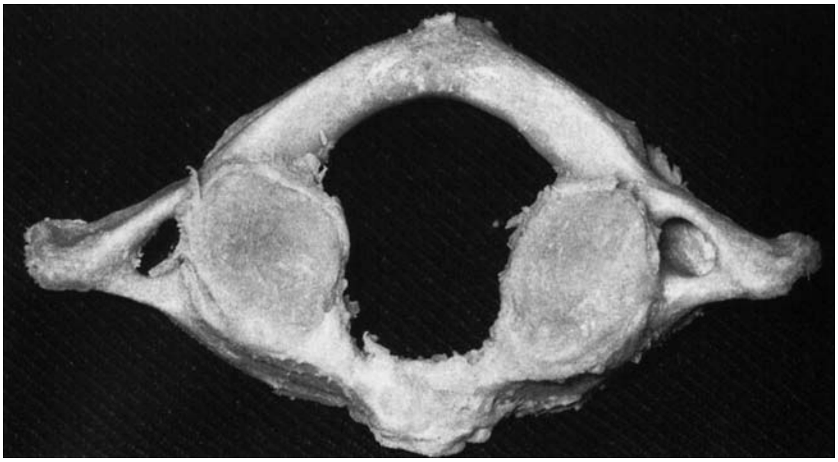

FIGURE 5.40 *The first cervical ('atlas') vertebra viewed from above to show the foramina in the transverse processes through which the vertebral arteries pass.*

Different pathologists have different procedures, but the following would be a reasonable method of investigation:

- Radiographs of the upper cervical region, both anteroposterior and lateral, should be taken as these may (rarely) reveal a fracture of the transverse process of the atlas vertebra. Such a fracture is, however, present in only a minority of vertebral artery injuries – and even when present, straight X-rays may not reveal it.
- Post-mortem angiograms should be taken if facilities exist. There are several methods of performing these. The lower neck is carefully dissected to reach the origins of the vertebral arteries. The best method is to open the subclavian vessels and identify the ostia of the vertebrals, which are the first (most medial) branches of the subclavian arteries. These can be cannulated and contrast media injected, one side at a time, while cervical and skull radiographs are taken. This method tends to fill many intracranial vessels and produce a confused, blurred picture.

  More satisfactorily, the brain should be carefully lifted from the skull and, as soon as the basilar artery is accessible, it should be clamped off with surgical forceps and/or ligated and the brain completely removed as usual after the basilar artery has been transected just above the clamp. Alternatively, each vertebral artery can be clamped or ligated in its terminal course. The angiogram is now performed, either by perfusing up one vessel and forcing contrast medium back down the other via the clamped basilar, or by consecutive injections on either side. Only sufficient medium to fill the system should be introduced to avoid spillage from small branches which obscure the films. The object is to detect any significant leakage from a vertebral vessel in its upper course, usually within the foramen of the atlas or just outside the atlanto-occipital membrane or within the spinal theca.
- As by definition, there will have been a substantial subarachnoid haemorrhage, this sometimes makes it difficult to identify and ligate the basilar or vertebral vessels. The brain should be minutely examined for berry aneurysms or other vascular malformations as, if they are present (and ruptured), then further investigation of vertebral artery damage is pointless. Assuming no such bleeding point is found, the upper cervical region is dissected to determine whether the vertebrals have been injured. When no X-ray or angiographic facilities exist, this is the only method available.

The upper cervical spine should be exposed by a posterior approach, which is continued up to meet the transverse scalp incision. The spine should be freed from the surrounding muscles, taking care to detect and record any muscular bruising in the vicinity. It should then be sawn through at about C4 level or lower.

The occipital bone should then be cut through on each side from the transverse skull saw-cut and prolonged down on each side of the foramen magnum, and then across the clivus to release the central part of the floor of the posterior fossa, carrying the foramen magnum and attached upper spine. Alternatively, using a power saw with a wedge-shaped blade, a square may be cut out of the floor of the posterior fossa to detach the spinal block. If possible, X-rays should be taken of the detached portion, as it may be possible to see a fracture of the transverse process of the atlas better than in the intact body.

The block of bone removed should then be decalcified by prolonged immersion in a large volume of 10 per cent formic acid.

After a week, the spine will be soft enough to slice with a scalpel, so that the lateral parts of the transverse processes can be shaved away, taking care not to get near the arterial foramina. The block is then returned to fresh decalcifying fluid for a further week.

Once the bone has been softened enough to be cut easily by a sharp knife, the transverse processes can be further shaved down on each side to the level of the foramina, so exposing the artery in its whole course. Both sides should be so dissected, the last stages being taken deeper and deeper with extreme care to remove the lateral walls of the canals without damaging the underlying vessel. Naturally, the crucial area is in the upper one or two vertebrae, but though laborious, this method provides the most elegant and convincing demonstration.

Specimens should be taken of any obvious or suspicious breach in the wall for histological examination. The usual type of damage is either a frank tear of the intima and media, or a dissection that allows blood to track through the adventitia.

The papers by Vanezis, Simonsen, and Cameron and Mant, among others, should be consulted for further details of the recommended procedures, but the above is a practical approach to demonstrating vertebral artery lesions. Once again, the relationship must be assessed of any tiny dissection, to the volume of blood found in the cranial meninges – and a decision made as whether this is more likely to be a concomitant injury to intracranial vessels rather than a solitary external vertebral artery lesion.

HEAD INJURIES IN BOXERS

It has become increasingly apparent in recent years that persons who indulge in boxing are at risk of both acute and chronic damage to their brains. In some countries, including parts of Scandinavia, both professional and amateur boxing is banned for this reason. The acute injuries are less common but occur during or soon after the fight itself. A number of boxers have died in the ring or after removal to hospital and, in these fatalities, by far the most common lesion is a subdural haemorrhage. Extradural bleeding almost never occurs, because boxing injuries rarely cause skull fractures. Occasionally a subarachnoid haemorrhage may occur in rare cases where a berry aneurysm is present.

Much attention has been paid to the chronic changes in boxers' brains, which are very common and give rise to what is generally known as the 'punch-drunk' syndrome. It appears that the length of time for which a boxer has been involved in fighting is more important than the number of serious traumatic events he has suffered. It is a cumulative process and many episodes of minor head injury add up to produce the typical lesions described in the now extensive publications on the subject.

There are both anatomical and microscopic lesions in such brains, both in professionals and amateurs exposed over a number of years. The clinical symptoms do not concern the pathologist, but they are explained by the morphological abnormalities. Grossly affected brains may show some cortical atrophy and slight hydrocephalus; the septum pellucidum is characteristically perforated with enlargement of the cavum and tearing of the septal leaves. The fornices and adjacent corpus callosum may be thinned or torn, and throughout the brain substance there may be scars and patches of gliosis. Neurones are lost from the cerebellum and the substantia nigra, the latter often losing pigment.

Another change that has intrigued neuropathologists is the development of an Alzheimer's-like condition, with neurofibrillary tangles throughout the cerebral cortex and brainstem, though no senile plaques are present.

CEREBRAL INJURIES

Though bleeding or infected scalp injuries, depressed fractures, meningitis and substantial meningeal haemorrhage can themselves cause death, in most fatal head injuries it is damage to the substance of the brain itself that is lethal.

In medico-legal practice it is sometimes difficult to convince lawyers that, in most instances, a victim does not die of a simple fractured skull, but that the fracture is evidence only of a substantial head injury, being a 'marker' for concomitant brain damage, which was the real lethal lesion. Similar problems of communication exist with a fracture of the hyoid bone in strangulation, lay persons often being under the misapprehension that a broken hyoid is a mortal injury, instead of merely being an inconstant marker of pressure on the neck.

The neuropathology of brain damage is a large and complex subject, the more subtle varieties requiring both specialist techniques for demonstration and expert knowledge for interpretation. The textbooks of Graham, Adams and Leestma are recommended for detailed description and discussion of cerebral trauma.

The mechanisms of production of some traumatic lesions are matters of conflicting theories, but the forensic pathologist still has to be aware of the general principles of causation in order to offer some interpretation of the injuries.

As always, caution has to be employed, as both the pathological and clinical manifestations of a head injury may appear to be at variance with the degree of force applied to the head. There is a wide range of lesions from a given insult to the head and it is dangerous to be too dogmatic in theorizing about the magnitude of an injury that gave rise to the demonstrable lesions.

It is similarly unwise to hypothesize too firmly about what clinical features must have occurred when the head injury was sustained – for example, quite severe head

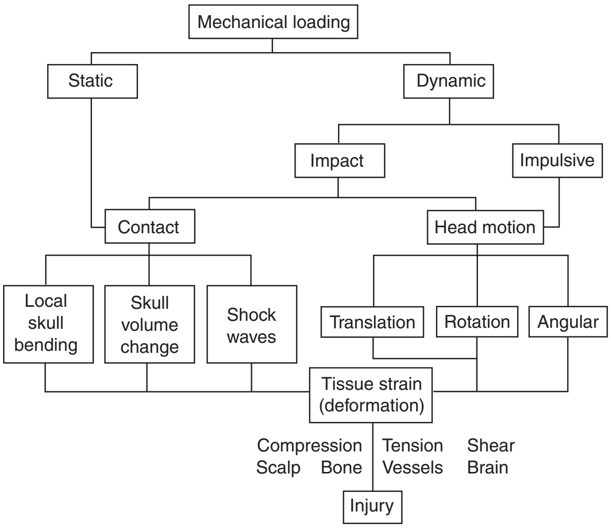

FIGURE 5.41 *Mechanical events that contribute to primary brain injury. (Reproduced from* Greenfield's Neuropathology, *7th edn, 2002 by kind permission of Arnold.)*

injuries have been known to be unaccompanied by concussion, while other apparently slight damage has often been followed by prolonged unconsciousness even ending in death. The well-known aphorism of Munro (1938) must be kept in mind that: 'Any type of head injury can give rise to any type of intracranial damage.'

The mechanism of brain damage

The brain may be injured in the following ways:

- by direct intrusion, either by a foreign object such as a penetrating weapon, bullet or other missile – or fragments of skull in a compound fracture where the skull is disrupted. In these open wounds the mechanism of the damage to the brain is obvious, though of course it may be compounded by the second type of injury described below
- by deformation of the brain in closed head injuries. Here the mechanism of injury is complicated and variable, with several competing theories of causation that have been put forward since the eighteenth century (Ledran 1751, Morgagni 1761). A public debate was held in 1766 in Paris, where rival theories of coup and contrecoup brain damage were hotly defended at the Academy of Surgeons.

The brain is almost incompressible and purely axial impact may give rise to little or no damage. It is extremely rare, however, for an impact not to impart some rotatory movement and it seems agreed that this component is the main culprit in causing brain damage. What is now clear is that no actual blow or fall need be suffered by the head to cause severe and even fatal brain damage. It is the change in velocity – either acceleration or deceleration – with a rotational rather than solely axial element, that leads to damage. The surface of the head need never contact any hard object or surface, though recent research indicates that the quantum of energy delivered by impact is of a far greater order of magnitude that non-impact violence, which lessens the former conviction that shaking, as in child abuse, is a common and potent mechanism in the production of intracranial damage.

In most head injuries – notably traffic accidents and falls – there is marked deceleration of the moving head on contact with a fixed surface but, in many criminal and combat injuries, the head is accelerated by a blow. In either case the initial sudden change in velocity is applied to the scalp and skull, the latter then transmitting the change to the brain via the anatomical suspensory system within the cranium. This system is slightly flexible and consists of the falx and tentorium, which divide the cranial cavity into three major compartments; these contain the two cerebral hemispheres, the cerebellum and the brainstem. When violent relative movements take place between the brain and the dura, forming the partitions of the cranium, the cerebral tissue can become damaged against both the sharp edges and the flat surface of these membranes. In addition, vessels traversing the subdural and subarachnoid spaces can be torn by such relative movements, especially in old people where cerebral atrophy may have widened these spaces.

Among the competing theories of impact brain damage are:

- the rotational shear force theory
- the pressure gradient theory
- the vibration theory
- the transmitted wave force theory
- the brain displacement theory
- the skull deformation theory.

These hypotheses overlap and most are correct in some aspect. Because experiments on primates and mechanical models have been pursued vigorously in recent years, it is now widely accepted that there are marked pressure changes within the cranium on impact (Yanagida *et al.* 1989). When a head falls against the ground, pressure momentarily increases at the impact point but falls to a negative value diametrically opposite. As these suction or cavitation effects are more damaging to neural and meningeal tissue than pressure, this is good evidence for claiming that contrecoup damage is largely a result of this vacuum effect.

The actual physical disruption of cerebral tissue is caused, according to both Gurdjian and Holbourn, by one or more of the following processes:

1. compression of the constituent units, by their being forced together
2. tension of the units, which pulls them apart
3. sliding or 'shear' strains, which move adjacent strata of tissue laterally. The usual homely example is given of a pack of playing cards being displaced, so that each card slides upon its neighbour.

Transient deformation of the skull almost certainly contributes to brain damage (Rowbotham 1964). The area of the skull beneath an impact becomes momentarily depressed even if it does not fracture and therefore may impinge on the underlying brain causing compression, as in 1, above. This is responsible for the typical cone-shaped contusions on the cortex, with the base at the surface, as the impact – possibly via short-lived oscillations of decreasing amplitude – injures the cortex and passes a diminishing force down into the deeper layers.

Simultaneously, other areas of the skull must bulge outward to accommodate the deformation – the so-called 'struck-hoop' action – when it is suggested that a 'rarefaction' remote from the impact may cause tension damage, as in 2, above.

More important is 3, being laminar deformity or 'shear stress' caused by the angular rotation of the head. As the head is pivoted on the first cervical vertebra, almost any impact on jaw, face or cranium will produce an angular momentum, the acceleration being conveyed first to the skull.

Alternatively, if the head is moving and is suddenly arrested, then the skull will decelerate first and the momentum of the brain will cause it to continue in motion, again almost certainly with some rotatory component.

In either the deceleration or acceleration mode, the skull and brain cannot change their velocities simultaneously and the brain will speed up or slow down only by virtue of the restraint provided by the dural septa and the configuration of the interior of the skull. In other words, the brain is either retarded or set into motion secondarily by the skull, especially by the dural septa and the bony prominences.

This restraint will occur first – and with maximum effect – on the most superficial layers of the cortex. These in turn will drag on the next deepest layer and so on until the difference in velocity is equalized – but this will have been at the expense of laminar tearing of the cerebral tissue and its associated blood vessels. In addition to this shearing damage, the brain may be forced against the sharp edge of the tentorial opening and the lower edge of the falx, causing damage to the base of the cerebrum, the corpus callosum and the brainstem. Impact against the side wall of the skull and against the falx may cause diffuse contusion of the cortex. The cerebellum tends to suffer less damage, as it is much smaller and lighter than the cerebrum and there is less room for relative movement in the more tightly enclosed posterior fossa. The configuration of the interior of the cranium is thought to be partly responsible for the common localization of cerebral damage at the tips and undersurface of the frontal and temporal lobes. The rough floor of the anterior fossa, the sharp edge of the wing of the sphenoid and the massive bar of the petrous temporal bone are in contrast to the smooth inner surface of the vault of the skull.

Cerebral contusion

When either linear or, more often, laminar stresses are applied to the cortex, this soft tissue may disrupt. Part of the injury is directly upon the neurocellular structure, but damage to vessels is an important component. If the cortex still retains its shape, but is bruised and swollen, this constitutes 'contusion'. A greater degree of disruption, sufficient to produce macroscopic tearing, is termed 'laceration', but the difference is only one of degree. In gross head injuries, such as crushing, missile wounds and other major penetration, the degree of laceration may lead to partial or even complete extrusion of the brain from the cranial cavity.

In the usual type of cortical contusion seen in a closed head injury, the cortex is blue or red from haemorrhage, though if survival has lasted for some time, there may be added discoloration from associated cortical infarction. The haemorrhage may be diffuse or may be punctate and is often a mottled purplish red when confined to the cortex. Extension into the underlying white matter tends to be pure red in fresh lesions. The lesion is often wedge-shaped, with the base on the surface, tapering away into the deeper layers.

Cerebral laceration

Laceration of the cortex is an extension in severity of contusion in which mechanical separation of the tissue can be seen. When widespread, but relatively superficial, the cortex appears to have a 'red velvet' appearance, which becomes progressively more tattered as the severity increases.

When it is even more severe, the cerebral surface becomes fissured, fragments of cortex may detach and deep lacerations run into the depths of the hemisphere, sometimes reaching even the deep ganglia or ventricles. There may be deep haemorrhage and – especially in the frontal and temporal lobes – the lacerations may be continuous with areas of traumatic haemorrhage. In cerebral lacerations and most contusions, the pia mater and often the arachnoid are torn, so that blood from damaged cortical vessels leaks into the subarachnoid and even subdural spaces. The corpus callosum is commonly torn, especially at its posterior end. This must be distinguished from damage caused by clumsy removal of the brain at autopsy (as must tears of the cerebral peduncles), but when genuine, may represent either a guillotine effect of the free lower edge of the falx or differential lateral movement of one hemisphere relative to the other, again because of unilateral restraint by the falx. Lacerations and contusions are most often found in those areas of the brain where the cortex is most likely to come into contact with irregularities in the internal profile of the skull. The undersurface of the temporal lobes and the orbital surface of the frontal lobes suffer most often.

Traumatic intracerebral haemorrhage

Substantial areas of haemorrhage, either infiltrating the brain tissue or forming actual haematomas, are common in severe head injuries. Some are primary, occurring at the time of impact or soon afterwards; others are secondary and caused by changes in intracranial pressure or bleeding into infarcts caused by vascular damage. These are all seen more often since artificial ventilation has been available, as victims of severe head injuries now survive longer so that there is time for secondary lesions to become apparent.

In the cerebral hemispheres, deep haemorrhage can be caused by coup or contrecoup mechanisms (see below) and may be situated anywhere within the hemispheres. They may rupture into the ventricular system or through the overlying cortex. In some severe contrecoup lesions, there may be large haematomas in one or both frontal lobes with overlying cortical contusion and laceration. These may rupture through the cortex into the meningeal spaces, forming what is sometimes called a ‘burst lobe’. Such haemorrhages sometimes pose a problem for the pathologist when they occur in older subjects, especially those with hypertension and perhaps cerebral atherosclerosis. When a scalp injury is present – and perhaps even a fractured skull – it may be difficult to decide if a head injury (such as a fall) was responsible for the cerebral haemorrhage, or whether a sudden ‘stroke’ caused by a natural cerebral haemorrhage resulted in the fall. The problem is discussed further in the chapter on natural death, but briefly, the presence of left ventricular (cardiac) hypertrophy, a history of hypertension and the site of the (usually solitary) haemorrhage tends to point to a natural bleed. This is especially so if the large size of the lesion seems inconsistent with the degree of head injury sustained. Hypertensive lesions tend to be in the thalamus, external capsule, pons and cerebellum, and are more often occipital than frontal or temporal. Having said that, on occasions it can be impossible to differentiate the two conditions.

Primary brainstem haemorrhage

Secondary brainstem bleeding is dealt with under ‘cerebral oedema’ so here we are concerned with haemorrhage that occurs at the time of injury. As mentioned above, hypertensive haemorrhages not associated with trauma can occur in the midbrain, especially the pons. These tend to be large, explosive lesions that greatly swell the pons and disrupt the central part of the stem, usually with a ragged rim of white matter around the periphery. Traumatic haemorrhage in the brainstem is often a well-circumscribed lesion, sometimes rounded, which lies laterally in the tegmentum, the shape of the midbrain being undistorted (unlike a secondary bleed into an elongated stem).

The typical site is between the aqueduct and the outer end of the substantia nigra. Primary stem haemorrhages are usually associated with occipital impacts and the victim is often unconscious from the time of the injury, as opposed to the lucid interval and gradual decline of those who suffer secondary stem lesions because of a developing cerebral oedema or space-occupying meningeal haematoma.

Coup and contrecoup damage

Whatever the underlying mechanics of cerebral damage, one aspect is of considerable practical importance to the pathologist. When a mobile head is struck with an object, the site of maximum cortical contusion is most likely to be beneath or at least on the same side as the blow. This is the so-called ‘coup’ lesion. When a moving head is suddenly decelerated, as in a fall, though there might still be a ‘coup’ lesion at the site of impact, there is often cortical damage on the opposite side of the brain – the ‘contrecoup’ lesion.

The mechanism of the ‘coup’ and ‘contrecoup’ injuries has long been debated – at least since the time of the famous Paris meeting of 1766. The controversy has been continued, especially by Courville (1942) and by Holbourn (1943), but no satisfactory resolution has been agreed though the work on intracranial pressures by Yanagida *et al.* seems to provide proof that a ‘vacuum’ occurs at the contrecoup site.

The following practical points should be considered:

- There may be no coup damage at all, only contrecoup.
- There need be no fracture of skull, even in the presence of severe coup and contrecoup lesions.
- The most common site for contrecoup injury is in the frontal and temporal lobes. It is often at the tips and on the undersurface of these lobes, and may be symmetrical, if a fall on the occiput has occurred.
- In temporal or parietal impacts, the contrecoup lesions are likely to be diametrically opposite on the

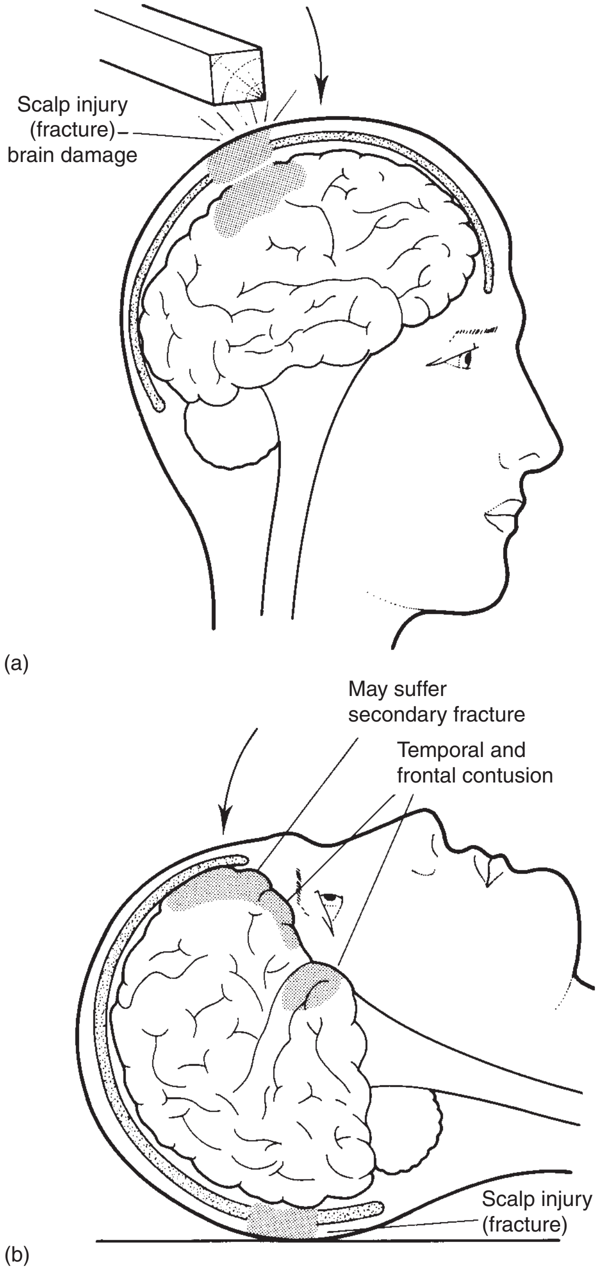
Scalp injury
(fracture) –
brain damage
(a)
May suffer
secondary fracture
Temporal and
frontal contusion
Scalp injury
(fracture)
(b)

FIGURE 5.42 *(a) 'Coup' brain damage to fixed head. (b) 'Contrecoup' brain damage to moving (decelerated) head.*

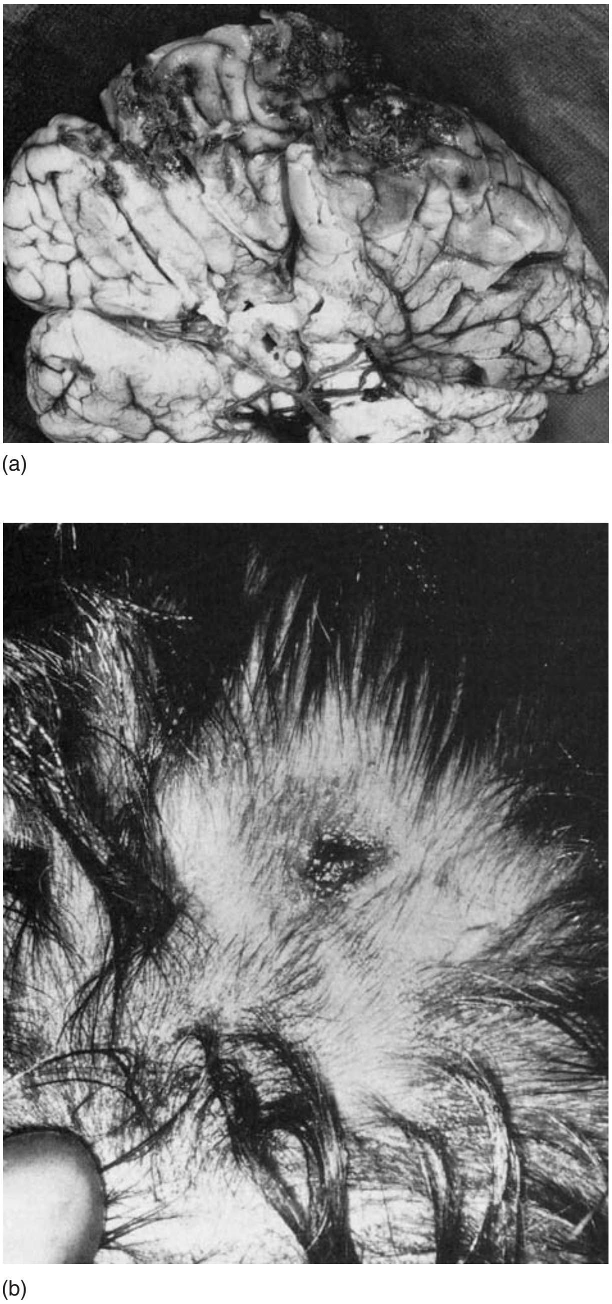

FIGURE 5.43 *(a) Typical contrecoup injury to the tips of the frontal and temporal lobes of a brain. The victim was pushed over in a brawl and fell backwards, striking his occiput and causing a small scalp laceration (b). There was no contusion to the posterior part of the brain.*

contralateral surface of the brain, but exact geometrical correspondence is not necessarily present.

- It is virtually unknown for a fall on the frontal region to produce occipital contrecoup. This is thought to be due to the anatomical configuration of the floor of the cranium, but the reasons are by no means understood.
- In a temporal impact, the contrecoup damage may be not be on the contralateral hemisphere, but on the opposite side of the ipsilateral hemisphere from impact against the falx cerebri.
- The degree of contrecoup damage may be severe, sufficient to cause blood-filled cavitation in the deep cortex and underlying white matter, especially in the frontal lobes and tips of the temporal lobes.

- With severe frontal contrecoup from a fall on the occiput, the transmitted force may be sufficient to fracture the thin bone of the floor of the anterior fossa. Such cracks in the roofs of the orbits may allow meningeal haemorrhage to seep into the orbits and appear as 'black eyes'. In assaults where a fall has occurred, care must be taken not to attribute such periorbital bleeding to direct punches.
- Though contrecoup contusion is classically caused by deceleration of a falling head, it can also occur when a **fixed** head is struck. If the victim is already lying on the ground or against some other unyielding surface, a heavy blow on the upper side may cause typical contrecoup lesions either in the contralateral temporal or parietal cortex, or against the falx on the inner side of the ipsilateral lobe. In these circumstances, there is often coup damage as well. No external scalp injury need be visible.
- The interpretation of contrecoup lesions is most reliable in the form of cortical contusions or lacerations. Meningeal haemorrhage, either subdural or subarachnoid, may also arise in association with a contrecoup lesion, but its diagnostic value is virtually nil compared with cortical damage when interpreting a falling or fixed head injury. Where no associated cortical contusion is present, it is quite unsafe to rely upon a unilateral meningeal haemorrhage as an indicator of the type of head injury.

Concussion

Concussion is a clinical, not a pathological entity, but the pathologist must consider it, as it is related to intracranial lesions and he is often questioned about it in court proceedings. Concussion, according to Wilson (1946), is 'a disorder of cerebral function which follows immediately upon the impact of a force to the head'. A more full definition is offered by Trotter (1914): 'a transient paralytic state due to head injury which is of instantaneous onset, does not show any evidence of structural cerebral injury and is always followed by amnesia from the actual moment of the accident'.

Some neurologists would also include post-concussion symptoms within the definition of concussion, even in the absence of initial coma, following a head injury. There may also be evidence of depressed medullary function, which can affect cardiorespiratory action. Denny-Brown and Russell (1941) showed that the rate of change of velocity of the head was important in producing concussion, which rarely developed if the speed threshold was less than ~8.5 m/s (28 ft/s).

It is an extremely common, but not inevitable, sequel to any significant mechanical insult to the brain. Though in general terms its duration is loosely related to the severity of the injury, there are many exceptions. Gross skull and brain damage have occurred with little or no apparent concussion, though concussion may be so transient that the subject may not even fall to the ground. Relatively minor head injuries have given rise to prolonged unconsciousness so, once again, it is most unwise to be dogmatic about retrospective estimates of concussion.

There is considerable controversy about the cause of concussion, from the unacceptable 'traumatic neurosis' on the one hand (which cannot be true) to claims for the inevitable demonstration of physical lesions on the other.

Courville (1953) has discussed the condition in depth and there seems to be no reason to doubt that some mechanical process does temporarily disrupt the function, if not necessarily the structure, of the neuronic apparatus. Changes in the nucleus and cytoplasm of neurones, the composition of the cerebrospinal fluid and in the electroencephalograph have all been inconstantly reported (see 'Diffuse neuronal and axonal injury' below).

True concussion may last for seconds or minutes. If prolonged unconsciousness extends into hours, days or longer, then there is likely to be some structural brain damage. Occasionally what appears to be simple concussion proves to be fatal, causing respiratory paralysis, though at autopsy no significant lesions are found.

Where a victim of 'simple' concussion dies of some incidental non-neurological condition, autopsy usually reveals no macroscopic damage, though sometimes there is slight cerebral oedema and scattered non-specific petechial haemorrhages may be found. There seems to be a connection between concussion and rotatory movements of the head, which are usually responsible for obvious structural damage because, when a head is fixed before impact, loss of consciousness may not occur. The classic example is trapping of a head against a wall or being jammed between buffers.

That shear stresses are instrumental in causing neurone damage seems confirmed by the frequency in which concussion occurs in boxing contests, where a blow on the jaw is the ultimate in producing a rotational movement of the cranium.

Concussion may be followed by a 'post-concussion state' characterized by headaches, unsteadiness and anxiety. This seems a genuine phenomenon, though it has been pointed out that it may be overlain by a 'compensation syndrome' whilst civil litigation is in progress over responsibility for the accident, which often clears up rapidly once the claim is settled.

Retrograde amnesia is almost inevitably associated with concussion, though, like concussion itself, it may be so transient as to escape notice. A protective mechanism, it seems to be caused by loss of sensory input before the latter is transferred to permanent memory storage in the brain.

Though commonly only of minutes' duration, it can extend to several days before the head injury. Though there is often a later recovery of much of this lost period, the memory of events immediately before the incident rarely returns, which may fortuitously be a protective device. Concussion has been attributed to several causes, including the undoubted vasomotor disturbances that take place after a head injury. Another theory is the impaction of the brain into the foramen magnum or tentorial opening, but the most acceptable hypothesis is 'diffuse neuronal injury'.

DIFFUSE NEURONAL AND AXONAL INJURY

According to Graham *et al.* (2002), diffuse brain damage exists in four principal forms: diffuse vascular injury, diffuse axonal injury, hypoxic brain damage and diffuse brain swelling. Diffuse vascular injury consists of multiple small haemorrhages throughout the brain and is virtually restricted to patients, who die within 24 h, whereas the latter three are encountered in patients who survive long enough to reach the hospital. Recent research and new immunohistochemical methods have shown that traumatic axonal injury is much more common than previously realized, and that axons can also be diffusely damaged by other processes than head injury. Therefore, Geddes *et al.* (2000) suggested new definitions of the terminology to get rid of the incoherent use of terms in the literature:

- Axonal injury (AI) is a non-specific term referring to damage to axons of any aetiology.
- Traumatic axonal injury (TAI), is a damage to axons caused by trauma, which may vary from small foci of axons to more widespread brain damage, diffuse TAI is the most severe form of traumatic axonal damage (originally termed 'DAI').
- Diffuse axonal damage (DAI), first described as a clinicopathological syndrome of widespread axonal damage throughout the brain, including the brainstem, should not be used as a term without reference to the aetiology, because axonal injury may be caused by other pathological processes.

Experimental work has suggested that diffuse axonal injury is primarily a non-impact rotational acceleration–deceleration phenomenon, deformation by stretching probably being the most significant factor. A low level of injury causes transient changes in the permeability of the axolemma, gradually leading to ionic changes, accumulation of fluid and axonal swelling and eventually, with an increasing grade of injury, to intracellular Ca^{2+} accumulation, proteolysis and collapse of the cytoskeleton.

The disruption of axons leads to bulbous and clubbed 'retraction balls or globes' on the axons in the cerebral hemispheres, cerebellum and brainstem. Disturbance of axonal transport causes accumulation of substances in damaged fibres, that can be demonstrated immunohistochemically. At present, β-amyloid precursor protein (βAPP) is considered to be the most reliable indicator of axonal damage, revealing axonal injury within 2–3 hours of the insult, whereas it takes about 12–18 hours for axonal bulbs to become visible on routine or silver stains. However, one should keep in mind that βAPP is not specific to head injury but is an indicator of derangement of fast axonal transport, which has also been demonstrated in other pathological conditions.

It is preferable to examine the brain after proper fixation, usually after 10–14 days in 4 per cent buffered formaldehyde and, where possible, consultation with a neuropathologist with experience in forensic practice. Due to the diffuse nature of the axonal injury, extensive and systematic sampling is essential, as the diagnostic yield is directly related to the number of blocks taken. In addition to any obvious focal pathology, the minumum set of samples should include corpus callosum and parasagittal posterior frontal white matter, splenium of the corpus callosum, deep grey matter to include posterior limb of the internal capsule, cerebellar hemisphere, midbrain including the decussation of superior cerebellar peduncle, and pons including superior or middle cerebellar peduncles. In addition to these, it may also be advisable to include corpus callosum and parasagittal anterior frontal white matter and temporal lobe including hippocampus (Geddes *et al.* 2000).

Retraction globes are most numerous in the corpus callosum, the superior peduncles, the parasagittal white matter, the medial lemnisci and the corticospinal tracts, but can be seen anywhere in the white matter. The retraction globes or axonal bulbs have been shown to lose positive staining for βAPP after about a week but to persist in adjacent varicose axons up to 30 days. After some weeks in those who survive long enough, clusters of microglia, presumably part of a repair process, congregate at the site of ruptured axons. These are best seen in 20 μm thicker sections stained with cresyl violet. One problem with the microscopic evidence of both microglial clusters as well as diffuse axonal injury (as demonstrated by βAPP expression), is that neither of them is specific as a marker for trauma: retraction globes may also be seen around the periphery of natural lesions such as cerebral infarcts and haemorrhages and microglial clusters have been reported in viral/HIV encephalitis, previous global hypoxia and fat embolism. Geddes *et al.* (2000) conclude in their excellent review that 'the demonstration of traumatic axonal damage is likely to be of limited use in most forensic situations, except perhaps to confirm that there has been a head injury'.

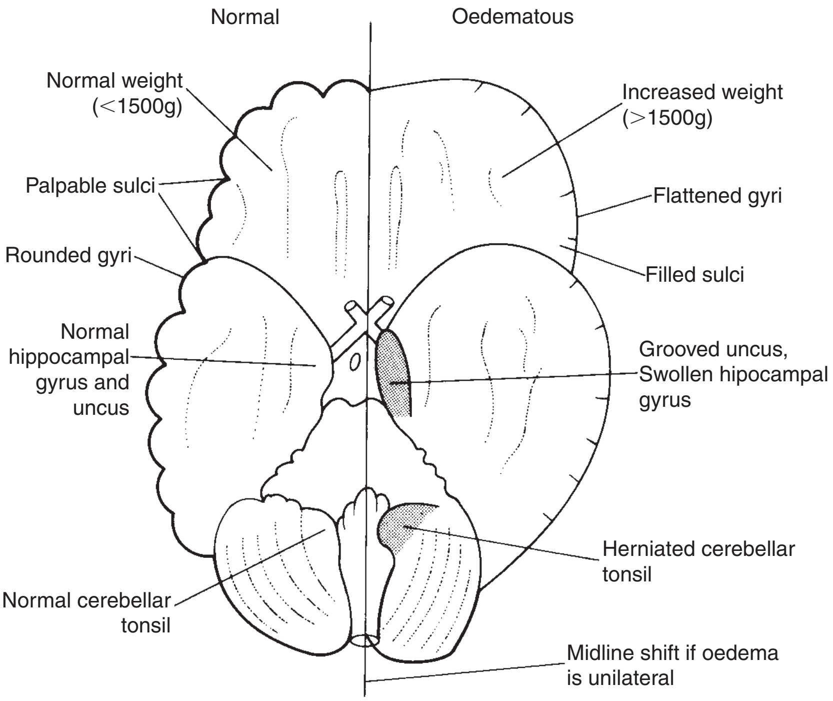
Normal
Oedematous
Normal weight (<1500g)
Palpable sulci
Rounded gyri
Normal hippocampal gyrus and uncus
Normal cerebellar tonsil
Increased weight (>1500g)
Flattened gyri
Filled sulci
Grooved uncus, Swollen hipocampal gyrus
Herniated cerebellar tonsil
Midline shift if oedema is unilateral

FIGURE 5.44 *Signs of cerebral oedema.*

Cerebral oedema

Swelling of the brain tissue may be a local phenomenon around almost any lesion, be it contusion, laceration, tumour or infarct, but here we are more concerned with generalized oedema. Swelling of the brain is extremely common after a substantial head injury, especially in children. Though it is an almost inevitable accompaniment of almost all intracerebral damage as either a local or general phenomenon, it can occur as the sole abnormality – and not infrequently prove fatal, particularly in young victims.

Oedema may well be related to diffuse neuronal injury and to concussion itself. It is the most common cause of raised intracranial pressure, being seen more often than localized space-occupying lesions such as haematomas and tumours, though of course these often coexist with cerebral oedema.

As with concussion, its cause is obscure and hence is the source of considerable controversy. The amount of fluid in the brain increases and the total weight may increase by at least 100 grams, mainly in the white matter. The site of the excess fluid is obscure, as the cut brain surface does not appear wet as do connective tissues elsewhere in an oedematous body. It was formerly thought that there was no true extracellular space in neural tissue and that the fluid must therefore be intracellular, but electron microscopic studies have revealed an extracellular compartment, which is much wider in the white matter (up to 80 nm) than in the grey matter (up to 20 nm), which explains the preference of oedema for the white matter.

The autopsy features of cerebral oedema are readily recognized. On removing the calvarium, the dura is stretched and tense, the brain bulging through the first incision in the membrane. The gyri are pale and flattened, and the sulci filled, giving the normally corrugated cerebral surface a smoothness that can easily be felt at autopsy. The cut surface is pale and, especially in children, the ventricles may be reduced to slits by the swelling of the adjacent white matter.

Severe cerebral oedema causes the larger volume cerebral hemispheres to press down upon the tentorium and herniate through the midbrain opening. The hippocampal gyrus may impact in the opening, lesser degrees causing grooving of the unci. Both these effects may lead to haemorrhage and necrosis at the sites of pressure, especially where the sharp edge of the tentorium cuts into the cerebral tissue. The tonsils of the cerebellum may be impacted or ‘coned’ into the foramen magnum, and sometimes are forced down into the upper part of the spinal canal. The pathologist must be careful not to mistake the normal anatomical grooving that often exists around the cerebellar tonsils for ‘coning’. There should be other signs of brain swelling and true tonsillar herniation will show discoloration or even necrosis of the ischaemic, trapped tissue.

Cerebral oedema may be the only intracranial abnormality found at autopsy after a substantial head injury has occurred. This seems to be more commonly found in children and, in the absence of any other demonstrable lesions, the cause of death has to be attributed to this swelling of the brain, compressing the vital centres in the brainstem.

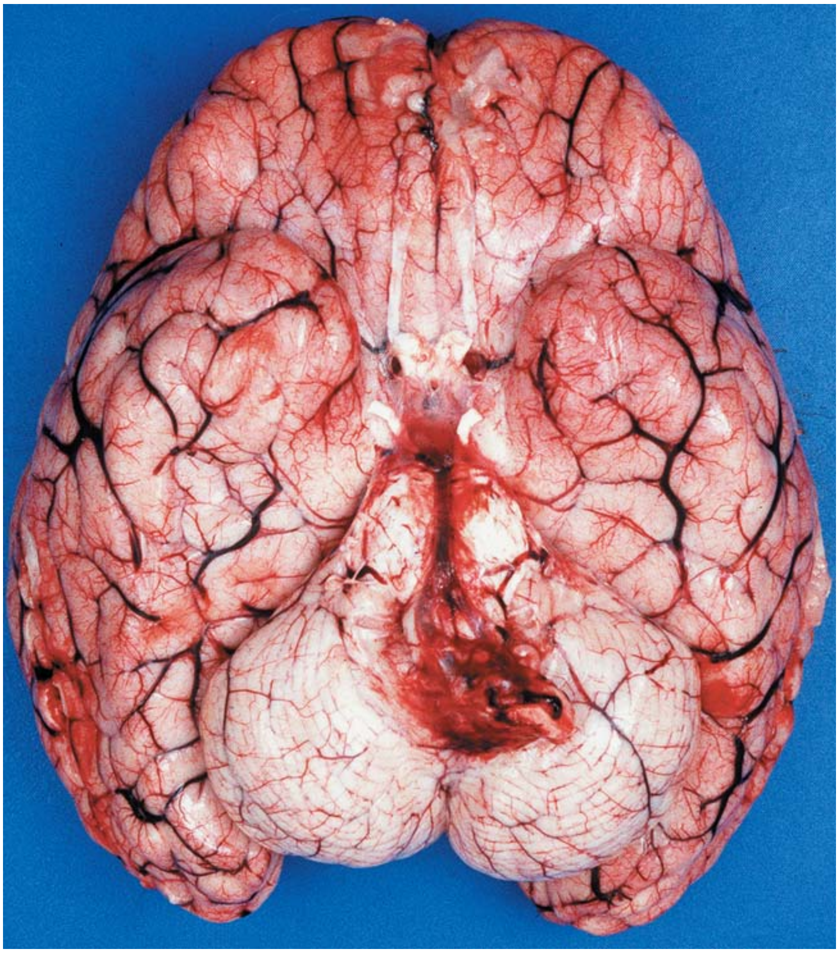

FIGURE 5.45 *Tonsillar herniation as a consequence of lumbar puncture in a patient with increased intracranial pressure. The pressure gradient has forced the cerebellar tonsils into foramen magnum and caused compression of the brainstem and patient's death. (Reproduced by kind permission of Professor H Kalimo.)*

Obviously, many other cases of cerebral oedema resolve either spontaneously or with treatment, so that fortunately no opportunity arises to prove its existence by post-mortem examination. Such oedema is, however, not infrequently found during surgical exploration for a meningeal haemorrhage, the latter sometimes being presented though often only brain swelling is demonstrated.

Cerebral oedema may be self-potentiating, in that once it begins as a result of direct brain trauma, the consequent rise in intracranial pressure then impairs the venous return from the intracranial sinuses. The pressure is insufficient to restrict the arterial inflow, so further congestion and swelling occur. This may lead to worsening cerebral hypoxia and oedema to the stage of actual cerebral infarction and brain death – again, a distressing common syndrome seen especially in child victims of head injuries, usually from road accidents.

In addition to mechanical damage, cerebral oedema can be caused or worsened by hypoxia. Many head injuries may be associated with damage to other parts of the body, such as the thorax – and because the airway may be compromised in many unconscious victims, there may be an element of hypoxia in a considerable proportion of cases of cerebral oedema. Capillary permeability may be increased in a number of states, leading to weakening of the so-called 'blood–brain barrier' formed by the capillary endothelium and basement membrane, together with the astrocyte footplates. Cerebral oedema, either traumatic or hypoxic, can develop with surprising rapidity, especially in children. Macroscopic evidence of brain swelling can be seen at autopsy in cases where the interval between trauma or onset of hypoxia and death was less than one hour.

Several methods are available to reverse oedema, including hyperventilation, which acts by providing full oxygenation and reducing the peripheral carbon dioxide tension, thus causing constriction of arterioles and a reduction in brain volume and transudation.

HISTOLOGICAL DIAGNOSIS OF EARLY CEREBRAL HYPOXIA

Unfortunately for the pathologist, most of the hypoxic conditions seen in forensic practice cause death too quickly for any recognizable histological changes to develop. Acute deaths from strangulation, suffocation and choking occur within minutes or less, even excepting the sudden vasovagal type of cardiac arrest.

There are, however, occasions when a longer period of survival occurs after an acute hypoxic episode and in some of these there may be the opportunity to detect histological changes in the central nervous system. Probably a minimum of 2–4 hours of survival is necessary for unequivocal changes to be observed, though some neuropathologists claim to detect signs after as little as 1 hour. More cautious investigators prefer 4 hours as the minimum. A short post-mortem interval is a considerable advantage in searching for these changes, as they are subtle and can be overlaid by post-mortem autolysis, even if the latter is slight.

It may be helpful to seek the assistance of a neuropathologist in detecting or confirming the relatively minor changes seen in a brain affected by hypoxia, but with experience, a good knowledge of the normal appearances and a laboratory that is able to turn out constantly good-quality sections and uniform staining, any forensic pathologist can build up a considerable expertise in this localized aspect of neuropathology. The parts of the brain most suitable for seeking evidence of hypoxic damage are the hippocampus (especially Sommer’s area), the cerebellar folia, the globus pallidus in the basal ganglia, and the boundary zones of the cerebral cortex where the grey matter is most vulnerable because it lies at the terminal reaches of the arterial supply (for example, at the occipitotemporal junction where the watershed between middle and posterior cerebral arteries lies). The third and fourth cortical layers are best for displaying the histological changes of hypoxia.

The following is a summary of the stages and changes that may be seen using different staining methods. The brain should be prefixed by suspension in formol saline for several weeks, though if this is impossible, small blocks should be wet-cut from the target areas of the brain mentioned above and immediately fixed in a large volume of formalin.

Stage 1

The earliest change is microvacuolation in the neurone cytoplasm, which may be seen in 2–4 hours. The cell outline remains smooth and its size is normal but, internally, small vacuoles appear in the cell body, and the proximal parts of the axon and dendrites. There may be a slight eosinophilia in haematoxylin and eosin stains and a violet colour in Luxol Fast Blue.

The microvacuolar change is said to begin earlier in small neurones and last for a shorter time, being present for up to 4–6 hours after insult in large cells and 2–4 hours in cerebellar Purkinje cells, though these times are arbitrary.

Stage 2

In the next stage, more prolonged hypoxia causes distortion in the shape of the neurones, the cell body becoming shrunken and staining darker with aniline dyes. The Nissl granules become fine and are dispersed more widely throughout the cell. Some vacuoles remain, the cytoplasm shows marked eosinophilia in haematoxylin and eosin, and is bright blue to mauve in Luxol Fast Blue stain. The nucleus becomes triangular and may be placed eccentrically in the cell body. The nucleolus becomes obscured and the nucleus stains a darker blue than normal with Luxol Fast Blue. This change is seen after up to 6 hours survival or later in large cells.

Stage 3

The ischaemic cell stage, described last, becomes worse during the following day or two if survival continues. The cell shrinks further, becoming narrow and tapered in many cases. Eosinophilia and blueness when stained with Luxol Fast Blue persists, but small spherical or irregular bodies begin to adhere to the exterior of the cell membrane. These incrustations persist longer than the remnants of the cytoplasm, which vanish, leaving the bare nucleus. The dark, shrunken nucleus becomes more apparent and survives for a number of days before karyolysis occurs.

Stage 4

Homogenizing cell damage is best seen in the Purkinje cells, in which the cytoplasm is uniformly eosinophilic with no Nissl granules. The nucleus stays dark and triangular until it vanishes and leaves a pale ‘ghost’ of the cell body with a vague perimeter.

There are parallel changes in the glial elements as well as in the neurones. From 4 to 12 hours after a hypoxic insult there may be an increase in the number of astrocytes around damaged neurones. There is a proliferation of astrocytes (which show mitoses) some 4–6 days later if the patient survives. This is most noticeable in the cerebellum and persists for 10–14 days. Fibrous astrocytes proliferate, but not the oligodendroglia. Microglia transform into rod cells, which may lie at right angles to the edge of the cortex, radially in the hippocampus and in the molecular layer of the cerebellar cortex. A later stage of cerebral hypoxic damage is the accumulation of lipid droplets and scavenging of myelin. Groups of microglia around dead neurones (neuronophages) may be obvious even at lower powers of the microscope.

SECONDARY BRAINSTEM LESIONS

When severe brain swelling (or a space-occupying lesion such as a large subdural or extradural haematoma) causes raised intracranial pressure above the tentorium, amongst the possible sequelae is compression of the midbrain against the free edge of the tentorium. This may be unilateral, causing grooving of a cerebral peduncle (called 'Kernohan's notch') associated with subpial petechial haemorrhages and often hemiplegia. When symmetrical, the oedema forces the undersurface of the cerebrum against the tentorium so that the hippocampal gyrus is squeezed into the opening. This elongates the midbrain in an anteroposterior direction as well as grooving the uncus and, in extreme cases, infarcting part of the parahippocampal gyrus. In addition, the distortion stretches, and compresses the paramedian and nigral vessels that supply the midbrain, leading to haemorrhages and infarcts in the upper pons and midbrain. Damage to cranial nerves and to the circulation of the cerebrospinal fluid may be added complications. The calcarine cortex on the medial aspect of the occipital lobe may be infarcted by the posterior cerebral artery being trapped around the edge of the tentorium by the cerebral herniation.

Secondary lesions are almost exclusively midline or paramedian haemorrhages, or haemorrhagic necrosis placed centrally in the upper pons and midbrain, though some may be obliquely placed in the substantia nigra. They may be difficult to distinguish from primary brainstem haemorrhages that arise at or soon after the original head injury – the secondary lesions tend to obscure the primary, rather than the converse. In children, the typical stem haemorrhages may not be present as in adults, but instead the medulla oblongata may be buckled or kinked due to fixation of the upper spinal cord by the denticulate ligaments (see work by Crompton).

SPINAL INJURIES

The spine and head should be thought of as part of the same system in relation to trauma. In recent years, a closer association – both skeletal and neurological – has been acknowledged between the two structures. For instance, interruption of ascending fibres in the cervical cord has been shown to be associated with neuronal chromatolysis in the brainstem and Spicer and Strich have shown that haemorrhage into spinal root ganglia may be associated with head injury. Electroencephalographic changes have been shown to occur in half the victims of cervical spine whiplash injury.

Though all segments of the spine are vulnerable to trauma, the cervical part holds the most interest for the forensic pathologist, mainly because of its close association with head injuries and vehicular accidents. The upper two cervical vertebrae provide most of the rotational movement of the head, whilst the lower neck allows flexion and extension. Violent force applied to the head tends to damage those parts of the neck corresponding with this functional distinction. Spinal damage may be caused by compressional, hyperflexion and hyperextension stress.

Compression damage

This occurs when the victim falls from a height either onto his feet or his head, though in the latter case, head injuries may overshadow damage to the spine, as well as absorbing most of the impact.

When a person falls a long distance onto the feet, the kinetic energy of the deceleration may be absorbed by fractures of the feet, legs and pelvis, but can be transmitted up the spinal column. This can be fractured at one or more points, or the force may cause the upper cervical spine to impinge on the base of the skull and cause a 'ring fracture' around the foramen magnum as the spine is rammed into the posterior fossa.

A fall onto the head may also cause the 'burst atlas' injury, where the impact of the occipital condyles in an axial direction wedge the superior atlantal articulating facets apart and split the ring of the vertebra. The posterior arch can also be fractured in hyperextension by compression between the occiput and the posterior spine of the axis.

Compression fractures of vertebral bodies may occur, most commonly in the lower dorsal and upper lumbar zone, particularly T12 and L1. There is less chance of spinal cord injury compared with the angulation injuries described below, unless posterior extrusion of a disc or backward displacement of fragments of a disrupted vertebral body occurs.

Hyperflexion and hyperextension injury

Of the two, hyperextension is much more dangerous in causing spinal damage, possibly because flexion is protected by contraction of the strong posterior neck muscles, whereas the weak anterior longitudinal ligament is incapable of preserving the integrity of the cervical spine during hyperextension. In frontal or rearward motor vehicle crashes, which comprise 80 per cent of accidents, there is usually a hyperflexion and hyperextension component to the spines of the occupants, though head restraints and seatbelts restrict the range of movement.

Where a car undergoes violent frontal deceleration, the subject's head will swing down into hyperflexion and, unless restrained, will then strike the fascia or windscreen, and rebound into hyperextension.

When the vehicle is struck from the rear, the head will fly into hyperextension first unless a head-rest is available; such collisions often then smash the vehicle into the one in front, when a deceleration hyperflexion takes place.

Whatever the cause, a whole range of lesions can follow, both in the cervical, and to a lesser extent into the thoracic and lumbar segments. Bleeding into the surrounding muscles,

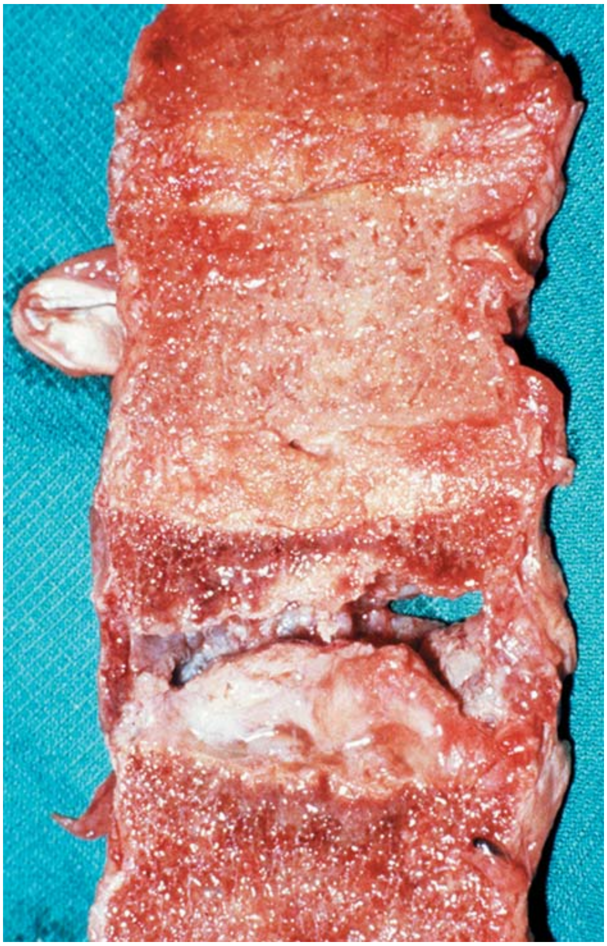

FIGURE 5.46 *Crushing of a thoracic vertebral body with disintegration of the adjacent disc following a violent hyperflexion in a car driver sustaining a deceleration injury. These lesions are best displayed at autopsy by cutting a longitudinal slice down the anterior aspect of the spinal column with an electric saw.*

rupture of the anterior longitudinal ligament, and tearing of intervertebral discs and of the annulus fibrosus may occur.

Nerve roots may be torn or compressed, and the spinal canal may be narrowed, distorted or even almost obliterated by fracture dislocations of the vertebrae. Compression, ischaemia, haemorrhage and even pulping of the spinal cord is the most serious complication. In older persons with cervical spondylosis, further narrowing of the canal by trauma may cause neurological effects that would not occur in younger subjects.

The most common part of the spine to be injured is the region of the upper two cervical vertebrae. In a series of fatal motor car accidents reported by Mant (1978), 35 per cent of passengers and 30 per cent of drivers suffered cervical spine damage. Of these the most common was a dislocation of the atlanto-occipital joint and Mant claims that one-third of all victims of fatal motor vehicle accidents suffer this lesion, even though it is often undetected at autopsy. The mechanism of fracture of the atlas is said to be an axial impact via the head, when the occiput is held rigidly in line with the spine by contracted neck muscles. This most frequently occurs when the vertex of the lowered head violently strikes the windshield of a decelerating vehicle. Fracture or dislocation of the atlas or fracture of the odontoid of the axis can all occur from this type of impact. Hyperextension injuries tend to force the vertebral body forwards and, if there is significant displacement, the arch is intruded into the spinal canal with the consequent risk of cord damage. In addition, hyperextension – especially in the presence of cervical cord enlargement – can cause the ligamentum flavum to corrugate and intrude into the anterior part of the spinal canal to impinge upon the cord. When the injury is due to hyperflexion, dislocation of a vertebra causes it (or its fragments) to tilt backwards, again compromising the lumen of the spinal canal.

Dislocation of cervical vertebra can occur under many conditions. An anterior dislocation can be caused by a fall onto the back of the head and a unilateral dislocation is common when a head strikes a windscreen. Posterior dislocation may be caused by blows to the jaw or face that jolt the head backwards with a hyperextension element. Falling onto the face, especially from a height (such as down a staircase), is a likely cause. Fracture of the odontoid peg of the axis can occur from a variety of violent movements of the head on the neck, and is sometimes associated with fracture of the skull or mandible. Where gross injury to the spinal column is inflicted – such as a relatively high-speed motor vehicle or railway accident – the cord may be transected by a guillotine action of the two displaced fragments. This is most often seen on the upper or mid-thoracic region, from direct impact or gross 'whiplash' effects. The thoracic aorta at this level is often torn at the point where the descending arch meets the spine, even in the absence of a spinal fracture.

Damage to the cervical spine in hanging is rarely seen in the usual suicidal hangings in which sudden death is caused by carotid compression or – much less often – asphyxia. The drop in such cases is small or even absent, the subject merely slumping his weight against the neck restraint. Spinal damage can occasionally occur, however and the author has seen a fractured neck when a heavily built soldier stepped from a lavatory seat with the rope attached to the overhead cistern.

Judicial hanging with a long drop causes a severe fracture dislocation (rather than fracture) of the cervical spine, often with complete severance of the two fragments and transection of the spinal cord. Where such hanging is carried out with the knot of the suspension point beneath the chin, a violent hyperextension is produced that often fractures the axis and dislocates it from the third cervical vertebra. With a heavy person and a long drop, there may be complete decapitation.

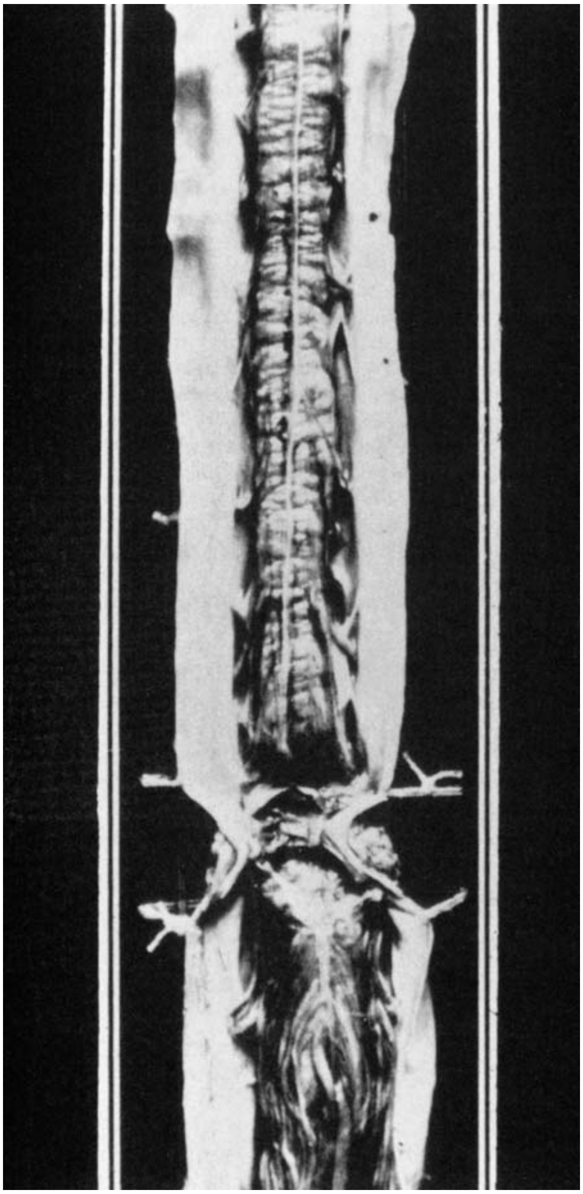

FIGURE 5.47 *Crushing and transection of the lower spinal cord following a hyperflexion fracture of the lower thoracic vertebrae. The spinal cord ends at about the lower border of the first lumbar vertebra, the cauda equina occupying the remainder of the lumbar spinal canal. This injury occurred several years before death, and the dura is constricted and adherent to the remnants of the cord. At autopsy the cord and its membranes should be pinned out and fixed for a proper examination to be made.*

Spinal cord injury

Most damage to the spinal cord arises from intrusion of some part of the spinal column into the canal, be it bony fragments or displacement, ligamentum flavum, disc annulus or extruded nucleus pulposum.

There may be bleeding into the space outside the spinal dura causing a space-occupying lesion in the canal that can compress the cord, or bleeding may occur within the dura, either from ruptured vessels or from haemorrhage in the cord itself. Damage to the cord may also occur in the absence of any apparent intrusion into the canal, in a manner similar to that seen in the brain in closed head injuries without a fracture. Haematomyelia and oedema may develop without any obvious mechanical defect at that level in the spinal column. Such injuries must be attributed to some momentary collision of the cord against the wall of the canal or a transient deformity of the profile of the canal that did not exceed the threshold for either fracture, dislocation or ligament rupture of the vertebral column.

The damage in the cord may extend for several segments above and below the point of impact. A haemorrhage within the cord tends to occupy the central grey matter more than the white columns, because of the softer and more vascular nature of the former tissue. Infarction of the cord can occur either in association with contusion, or because the local blood supply – often the anterior spinal artery – has been damaged.

An originally firm cord may undergo liquefaction ('myelomalacia') over several days or even longer, with progressive worsening of neurological symptoms.

REFERENCES AND FURTHER READING

Adams JH. 1982. Diffuse axonal injury in non-missile head injury. *Injury* **13**:444–5.

Adams H, Mitchell DE, Graham DI, *et al.* 1977. Diffuse brain damage of immediate impact type. Its relationship to 'primary brain-stem damage' in head injury. *Brain* **100**:489–502.

Adams JH, Graham DI, Murray LS, *et al.* 1982. Diffuse axonal injury due to nonmissile head injury in humans: an analysis of 45 cases. *Ann Neurol* **12**:557–63.

Adams JH, Graham DI, Gennarelli TA. 1983. Head injury in man and experimental animals: Neuropathology. *Acta Neurochir Suppl (Wien)* **32**:15–30.

Adams JH, Doyle D, Graham DI, *et al.* 1984. Diffuse axonal injury in head injuries caused by a fall. *Lancet* **2**:1420–2.

Adams JH, Doyle D, Graham DI, *et al.* 1985. Microscopic diffuse axonal injury in cases of head injury. *Med Sci Law* **25**:265–9.

Adams JH, Doyle D, Graham DI, *et al.* 1985. The contusion index: a reappraisal in human and experimental non-missile head injury. *Neuropathol Appl Neurobiol* **11**:299–308.

Adams JH, Doyle D, Graham DI, *et al.* 1986. Gliding contusions in nonmissile head injury in humans [published erratum appears in *Arch Pathol Lab Med* 1986 Nov;**110**:1075]. *Arch Pathol Lab Med* **110**:485–8.

Adams JH, Doyle D, Graham DI, *et al.* 1986. Deep intracerebral (basal ganglia) haematomas in fatal non-missile head injury in man. *J Neurol Neurosurg Psychiatry* **49**:1039–43.

Adams JH, Doyle D, Ford I, *et al.* 1989. Diffuse axonal injury in head injury: definition, diagnosis and grading. *Histopathology* **15**:49–59.

Adams JH, Doyle D, Ford I, *et al.* 1989. Brain damage in fatal non-missile head injury in relation to age and type of injury. *Scott Med J* **34**:399–401.

Adams JH, Graham DI, Gennarelli TA, *et al.* 1991. Diffuse axonal injury in non-missile head injury. *J Neurol Neurosurg Psychiatry* **54**:481–3.

Adams JH, Corsellis JA, Duchen LW. 1992. *Greenfield's neuropathology*, 5th edn. Edward Arnold, London.

Adams JH, Jennett B, McLellan DR, *et al.* 1999. The neuropathology of the vegetative state after head injury. *J Clin Pathol* **52**:804–6.

Adams JH, Graham DI, Jennett B. 2000. The neuropathology of the vegetative state after an acute brain insult. *Brain* **123**(Pt 7):1327–38.

Adams JH, Graham DI, Jennett B. 2001. The structural basis of moderate disability after traumatic brain damage. *J Neurol Neurosurg Psychiatry* **71**:521–4.

Adelson L. 1974. *The pathology of homicide*. Thomas, Springfield.

Blumbergs PC, Jones NR, North JB. 1989. Diffuse axonal injury in head trauma. *J Neurol Neurosurg Psychiatry* **52**:838–41.

Bostrom K, Helander CG, Lindgren SO. 1992. Blunt basal head trauma: rupture of posterior inferior cerebellar artery. *Forensic Sci Int* **53**:61–8.

Bromilow A, Burns J. 1985. Technique for removal of the vertebral arteries. *J Clin Pathol* **38**:1400–2.

Caffey J. 1946. Multiple fractures in the long bones of infants suffering from chronic subdural haematoma. *Am J Radiol* **56**:163–73.

Cameron JM, Mant AK. 1972. Fatal subarachnoid haemorrhage associated with cervical trauma. *Med Sci Law* **12**:66–70.

Casson IR, Siegel O, Sham R, *et al.* 1984. Brain damage in modern boxers. *JAMA* **251**:2663–7.

Coast GC, Gee DJ. 1984. Traumatic subarachnoid haemorrhage: an alternative source. *J Clin Pathol* **37**:1245–8.

Contostavlos DL. 1971. Massive subarachnoid hemorrhage due to laceration of the vertebral artery associated with fracture of the transverse process of the atlas. *J Forensic Sci* **16**:40–56.

Contostavlos DL. 1995. Isolated basilar traumatic subarachnoid hemorrhage: an observer's 25 year re-evaluation of the pathogenetic possibilities. [letter; comment] *Forensic Sci Int* **73**:61–74.

Corsellis JA, Bruton CJ, Freeman Browne D. 1973. The aftermath of boxing. *Psychol Med* **3**:270–303.

Courville CB. 1942. Coup-contre coup mechanism of craniocerebral injuries: some observations. *Arch Surg* **45**:19–43.

Courville CB. 1953. *Commotio cerebri*. San Lucas Press, Los Angeles.

Courville CB. 1964. *Forensic neuropathology*. Callaghan and Co, Mundelein.

Critchley M. 1957. Medical aspects of boxing particularly from a neurological standpoint. *Br Med J* **1**:357–62.

Crompton MR. 1971. Brain stem lesions due to closed head injury. *Lancet* **1**:669–71.

Crompton R. 1985. *Closed head injuries*. Edward Arnold, London.

Dalgaard J. 1957. Brain injuries as a cause of oesophagogastroduodenal ulceration. *J Forensic Med* **4**:110–27.

Dawson SL, Hirsch CS, Lucas FV, *et al.* 1980. The contrecoup phenomenon. Reappraisal of a classic problem. *Hum Pathol* **11**:155–66.

Deck JH, Jagadha V. 1986. Fatal subarachnoid hemorrhage due to traumatic rupture of the vertebral artery. *Arch Pathol Lab Med* **110**:489–93.

Denny-Brown D. 1945. Cerebral concussion. *Physiol Rev* **25**:296–325.

Denny-Brown D, Russell WR. 1941. Experimental cerebral concussion. *Brain* **64**:93–164.

Dowling G, Curry B. 1988. Traumatic basal subarachnoid hemorrhage. Report of six cases and review of the literature. *Am J Forensic Med Pathol* **9**:23–31.

Duhaime AC, Gennarelli TA, Thibault LE, *et al.* 1987. The shaken baby syndrome. A clinical, pathological, and biomechanical study. *J Neurosurg* **66**:409–15.

Duhaime AC, Alario AJ, Lewander WJ, *et al.* 1992. Head injury in very young children: mechanisms, injury types, and ophthalmologic findings in 100 hospitalized patients younger than 2 years of age. *Pediatrics* **90**:179–85.

Edherg S, Angrist A. 1967. Intra-cranial pressure changes following impact of human cadaver heads. *J Forensic Sci* **12**:60–8.

Editorial. 1973. Boxing brains. *Lancet* **2**:1064–5.

Farag AM, Franks A, Gee DJ. 1988. Simple laboratory experiments to replicate some of the stresses on vertebro-basilar arterial walls. An investigation of possible mechanisms of traumatic subarachnoid haemorrhage. *Forensic Sci Int* **38**:275–84.

Fell DA, Fitzgerald S, Moiel RH, *et al.* 1975. Acute subdural hematomas. Review of 144 cases. *J Neurosurg* **42**:37–42.

Flamm ES, Demopoulos HB, Seligman ML, *et al.* 1977. Ethanol potentiation of central nervous system trauma. *J Neurosurg* **46**:328–35.

Ford R. 1956. Basal subarachnoid haemorrhage and trauma. *J Forensic Sci* **1**:117–26.

Freytag E. 1963. Autopsy findings in head injuries from blunt forces: statistical evaluation of 1367 cases. *Arch Pathol* **75**:74–80.

Garfield J. 2002. Acute subdural haematoma in a boxer. *Br J Neurosurg* **16**:96–9; discussion 99–101.

Geddes JF, Whitwell HL, Graham DI. 2000. Traumatic axonal injury: practical issues for diagnosis in medicolegal cases. *Neuropathol Appl Neurobiol* **26**:105–16.

Geddes JF, Hackshaw AK, Vowles GH, *et al.* 2001. Neuropathology of inflicted head injury in children. I. Patterns of brain damage. *Brain* **124**:1290–8.

Geddes JF, Vowles GH, Hackshaw AK, *et al.* 2001. Neuropathology of inflicted head injury in children. II. Microscopic brain injury in infants. *Brain* **124**:1299–306.

Gennarelli TA. 1983. Head injury in man and experimental animals: clinical aspects. *Acta Neurochir Suppl (Wien)* **32**:1–13.

Gennarelli TA. 1993. Mechanisms of brain injury. *J Emerg Med* **11**(Suppl 1):5–11.

Gennarelli TA, Graham DI. 1998. Neuropathology of the head injuries. *Semin Clin Neuropsychiatry* **3**:160–175.

Gennarelli TA, Thibault LE. 1982. Biomechanics of acute subdural hematoma. *J Trauma* **22**:680–6.

Gennarelli TA, Thibault LE, Adams JH, *et al.* 1982. Diffuse axonal injury and traumatic coma in the primate. *Ann Neurol* **12**:564–74.

Gennarelli TA, Thibault LE, Tipperman R, *et al.* 1989. Axonal injury in the optic nerve: a model simulating diffuse axonal injury in the brain. *J Neurosurg* **71**:244–53.

Gentleman SM, McKenzie JE, Royston MC, *et al.* 1999. A comparison of manual and semi-automated methods in the assessment of axonal injury. *Neuropathol Appl Neurobiol* **25**:41–7.

Goonetilleke UK. 1980. Injuries caused by falls from heights. *Med Sci Law* **20**:262–75.

Graham DI. 2001. Paediatric head injury. *Brain* **124**:1261–2.

Graham DI, Adams JH. 1971. Ischaemic brain damage in fatal head injuries. *Lancet* **1**:265–6.

Graham DI, Adams JH. 1972. *The pathology of blunt head injuries.* In: Critchley M (ed.), *Scientific foundation of neurology*. Heinemann Medical Publications, London.

Graham DI, Lantos PL (eds). 2002. *Greenfield's neuropathology*, 7th edn. Arnold, London/New York/New Delhi.

Graham DI, McLellan D, Adams JH, *et al.* 1983. The neuropathology of the vegetative state and severe disability after non-missile head injury. *Acta Neurochir Suppl (Wien)* **32**:65–7.

Graham DI, Lawrence AE, Adams JH, *et al.* 1988. Brain damage in fatal non-missile head injury without high intracranial pressure. *J Clin Pathol* **41**:34–7.

Graham DI, Ford I, Adams JH, *et al.* 1989. Fatal head injury in children. *J Clin Pathol* **42**:18–22.

Graham DI, Clark JC, Adams JH, *et al.* 1992. Diffuse axonal injury caused by assault. *J Clin Pathol* **45**:840–1.

Graham DI, McIntosh TK, Maxwell WL, *et al.* 2000. Recent advances in neurotrauma. *J Neuropathol Exp Neurol* **59**:641–51.

Graham DI, Raghupathi R, Saatman KE, *et al.* 2000. Tissue tears in the white matter after lateral fluid percussion brain injury in the rat: Relevance to human brain injury. *Acta Neuropathol (Berl)* **99**:117–24.

Gurdjian ES, Webster JE. 1958. *Head injuries*. Churchill, London, UK.

Gurdjian ES, Webster JE, Lissner HR. 1949. Studies on skull fracture with particular reference to engineering factors. *Am J Surg* **78**:736–42.

Gurdjian ES, Webster JE, Lissner HR. 1950. The mechanism of skull fracture. *Radiology* **54**:313–39.

Harris LS. 1991. Postmortem magnetic resonance images of the injured brain: effective evidence in the courtroom. *Forensic Sci Int* **50**:179–85.

Harwood-Nash CE, Hendrick EB, Hudson AR. 1971. The significance of skull fractures in children. A study of 1,187 patients. *Radiology* **101**:151–6.

Hijdra A, Vermeulen M, van Gijn J, *et al.* 1984. Respiratory arrest in subarachnoid hemorrhage. *Neurology* **34**:1501–3.

Holbourn AH. 1943. Mechanics of head injuries. *Lancet* **245**:438–41.

Huelke DF, O'Day J, Mendelsohn RA. 1981. Cervical injuries suffered in automobile crashes. *J Neurosurg* **54**:316–22.

Imajo T. 1996. Diffuse axonal injury: its mechanism in an assault case. *Am J Forensic Med Pathol* **17**:324–6.

Imajo T, Kazee AM. 1992. Diffuse axonal injury by simple fall. *Am J Forensic Med Pathol* **13**:169–72.

Imajo T, Roessman U. 1984. Diffuse axonal injury. *Am J Forensic Med Pathol* **5**:217–22.

Imajo T, Challener RC, Roessmann U. 1987. Diffuse axonal injury by assault. *Am J Forensic Med Pathol* **8**:217–19.

Jafari SS, Maxwell WL, Neilson M, *et al.* 1997. Axonal cytoskeletal changes after non-disruptive axonal injury. *J Neurocytol* **26**:207–21.

Jafari SS, Nielson M, Graham DI, *et al.* 1998. Axonal cytoskeletal changes after nondisruptive axonal injury. II. Intermediate sized axons. *J Neurotrauma* **15**:955–66.

Jamieson KG, Yelland JD. 1968. Extradural hematoma. Report of 167 cases. *J Neurosurg* **29**:13–23.

Johnson CP, Lawler W, Burns J. 1993. Use of histomorphometry in the assessment of fatal vertebral artery dissection. *J Clin Pathol* **46**:1000–3.

Johnson CP, How T, Scraggs M, *et al.* 1994. The poor biomechanical response of the vertebral artery applied longitudinal force. *J Pathol* **Suppl**:234.

Johnson DL, Boal D, Baule R. 1995. Role of apnea in nonaccidental head injury. *Pediatr Neurosurg* **23**:305–10.

Johnson J. 1969. Organic psychosyndromes due to boxing. *Br J Psychiatry* **115**:45–53.

Johnson P, Burns J. 1995. Extracranial vertebral artery injury – evolution of a pathological illusion? [letter; comment] *Forensic Sci Int* **73**:75–8.

Kaur B, Rutty GN, Timperley WR. 1999. The possible role of hypoxia in the formation of axonal bulbs. *J Clin Pathol* **52**:203–9.

Kerr D. 1954. *Forensic medicine*, 5th edn. A. & C. Black, Edinburgh, p. 114.

Knight B. 1979. Trauma and ruptured cerebral aneurysm. *Br Med J* **1**:1430–1.

Kubo S, Kitamura O, Orihara Y, *et al.* 1998. Immunohistochemical diagnosis and significance of forensic neuropathological changes. *J Med Invest* **44**:109–19.

Leadbeatter S. 1994. Extracranial vertebral artery injury – evolution of a pathological illusion? [see comments] *Forensic Sci Int* **67**:33–40.

Leadbeatter S. 1995. Letter to editor. *Forensic Sci Int* **73**:78–9.

Ledran H. 1751. *Observatis des chirurgies*. Osmont, Paris.

Leestma JE. 1988. *Forensic neuropathology*. Raven Press, New York.

Li L, Smialek JE. 1994. The investigation of fatal falls and jumps from heights in Maryland (1987–1992). *Am J Forensic Med Pathol* **15**:295–9.

Lindenberg R. 1955. Lesions of the corpus callosum following blunt mechanical trauma to the head. *Am J Pathol* **31**:297–301.

Lindenberg R, Freytag E. 1957. Morphology of cerebral contusions. *Arch Pathol* **63**:23–6.

Lindenberg R, Freytag E. 1960. The mechanism of cerebral contusions: a pathologic-anatomic study. *Arch Pathol* **69**:440–4.

Lindenberg R, Freytag E. 1970. Brainstem lesions characteristic of traumatic hyperextension of the head. *Arch Pathol* **90**:509–15.

Lindsay KW, McLatchie G, Jennett B. 1980. Serious head injury in sport. *Br Med J* **281**:789–91.

Lundberg GD. 1985. Brain injury in boxing. *Am J Forensic Med Pathol* **6**:192–8.

Madro R, Chagowski W. 1987. An attempt at objectivity of post mortem diagnostic of brain oedema. *Forensic Sci Int* **35**:125–9.

Maloney AFJ, Whatmore WJ. 1969. Clinical and pathological observations in fatal head injuries: a five year survey of 173 cases. *Br J Surg* **56**:23–30.

Mansell PW, Hunt AC. 1968. Major head injuries in fights and brawls. *Med Sci Law* **8**:181–7.

Mant AK. 1978. Injuries and death in motor vehicle accidents. In: Mason JK (ed.), *Pathology of violent injury*. Edward Arnold, London.

Mant AK. 1993. Injuries and death in motor vehicle accidents. In: Mason JK (ed.), *Pathology of violent injury*, 2nd edn. Edward Arnold, London.

Marek Z. 1981. Isolated subarachnoid hemorrhage as a medicolegal problem. *Am J Forensic Med Pathol* **2**:19–22.

Marlet JM, Barreto Fonseca JdP. 1982. Experimental determination of time of intracranial hemorrhage by spectrophotometric analysis of cerebrospinal fluid. *J Forensic Sci* **27**:880–8.

Martland MS. 1928. Punch drunk. *JAMA* **91**:1003–5.

Mawdsley C, Ferguson SR. 1963. Neurological disease in boxers. *Lancet* **2**:795–7.

McCormick WF. 1980. The relationship of closed-head trauma to rupture of saccular intracranial aneurysms. *Am J Forensic Med Pathol* **1**:223–6.

McCrory P. 2002. Boxing and the brain. Revisiting chronic traumatic encephalopathy. *Br J Sports Med* **36**:2.

McKissock H. 1960. Extradural haemorrhage; observations on 125 cases. *Lancet* **2**:167–74.

Missliwetz J. 1995. Fatal impalement injuries after falls at construction sites. *Am J Forensic Med Pathol* **16**:81–3.

Mitchell DE, Adams JH. 1973. Primary focal impact damage to the brainstem in blunt head injuries. Does it exist? *Lancet* **2**:215–18.

Miyazaki T, Kojima T, Chikasue F, *et al.* 1990. Traumatic rupture of intracranial vertebral artery due to hyperextension of the head: reports on three cases. *Forensic Sci Int* **47**:91–8.

Moar JJ. 1987. Traumatic rupture of the cervical carotid arteries: an autopsy and histopathological study of 200 cases. *Forensic Sci Int* **34**:227–44.

Morgagni J. 1761. *De sedibus et causis morborum per anatomen indagatis.* Typographia Academia, Lovanii.

Moritz AR. 1942. *Pathology of trauma*. Lea & Febiger, Philadelphia.

Munro D. 1938. *Cranio-cerebral injuries*. Oxford University Press, Oxford.

Munro D, Merritt HH. 1936. Surgical pathology of subdural haematoma. Based on a study of 105 cases. *Arch Neurol Psychiatry* **35**:64–78.

Oehmichen M. 1994. Brain death: neuropathological findings and forensic implications. *Forensic Sci Int* **69**:205–19.

Oehmichen M, Raff G. 1980. Timing of cortical contusion. Correlation between histomorphologic alterations and post-traumatic interval. *Z Rechtsmed* **84**:79–94.

Oehmichen M, Eisenmenger W, Raff G, *et al.* 1986. Brain macrophages in human cortical contusions as indicator of survival period. *Forensic Sci Int* **30**:281–301.

Oehmichen M, Meissner C, Schmidt V, *et al.* 1999. Pontine axonal injury after brain trauma and nontraumatic hypoxic-ischemic brain damage. *Int J Legal Med* **112**:261–7.

Oehmichen M, Theuerkauf I, Meissner C. 1999. Is traumatic axonal injury (AE) associated with an early microglial activation? Application of a double-labeling technique for simultaneous detection of microglia and ai. *Acta Neuropathol (Berl)* **97**:491–4.

Oehmichen M, Ochs U, Meissner C. 2000. Histochemical characterization of cytotoxic brain edema. Potassium concentrations after cerebral ischemia and during the postmortem interval. *Exp Toxicol Pathol* **52**:348–52.

Oehmichen M, Meissner C, Konig HG. 2001. Brain injury after survived gunshot to the head: reactive alterations at sites remote from the missile track. *Forensic Sci Int* **115**:189–97.

Oppenheimer DR. 1968. Microscopic lesions in the brain following head injury. *J Neurol Neurosurg Psychiatry* **31**:299–306.

Povlishock JT. 1992. Traumatically induced axonal injury: pathogenesis and pathobiological implications. *Brain Pathol* **2**:1–12.

Povlishock JT. 2000. Pathophysiology of neural injury: therapeutic opportunities and challenges. *Clin Neurosurg* **46**:113–26.

Povlishock JT, Christman CW. 1995. The pathobiology of traumatically induced axonal injury in animals and humans: a review of current thoughts. *J Neurotrauma* **12**:555–64.

Raisanen J, Ghougassian DF, Moskvitch M, *et al.* 1999. Diffuse axonal injury in a rugby player. *Am J Forensic Med Pathol* **20**:70–2.

Roberts AH. 1970. *Brain damage in boxers*. Pitman Medical, London.

Root I. 1992. Head injuries from short distance falls [see comments]. *Am J Forensic Med Pathol* **13**: 85–7.

Rowbotham OF. 1964. *Acute injuries of the head*, 4th edn. Churchill Livingstone, Edinburgh.

Royal College of Physicians of London. 1969. *Report on the medical aspects of boxing*. Royal Collage of Physicians of London, London.

Schellhas KP, Latchaw RE, Wendling LR, *et al.* 1980. Vertebrobasilar injuries following cervical manipulation. *JAMA* **244**:1450–3.

Schellinger PD, Schwab S, Krieger D, *et al.* 2001. Masking of vertebral artery dissection by severe trauma to the cervical spine. *Spine* **26**:314–19.

Shapiro HA, Gordon I, Benson SD. 1988. *Forensic medicine – a guide to principles*, 3rd edn. Churchill Livingstone, Edinburgh.

Sheffield EA, Weller RO. 1980. Age changes at cerebral artery bifurcations and the pathogenesis of berry aneurysms. *J Neurol Sci* **46**:341–52.

Sherriff FE, Bridges LR, Sivaloganathan S. 1994. Early detection of axonal injury after human head trauma using immunocytochemistry for beta-amyloid precursor protein. *Acta Neuropathol (Berl)* **87**:55–62.

Simonsen J. 1967. Fatal subarachnoid haemorrhage in relation to minor head injuries. *J Forensic Med* **14**:146–55.

Simonsen J. 1976. Massive subarachnoid haemorrhage and fracture of the transverse process of the atlas. *Med Sci Law* **16**:13–6.

Simpson RH, Berson DS, Shapiro HA. 1985. The diagnosis of diffuse axonal injury in routine autopsy practice. *Forensic Sci Int* **27**:229–35.

Skold G. 1985. Injuries to pathologically changed cervical vertebrae. *Am J Forensic Med Pathol* **6**:163–6.

Smith DH, Chen XH, Xu BN, *et al.* 1997. Characterization of diffuse axonal pathology and selective hippocampal damage following inertial brain trauma in the pig. *J Neuropathol Exp Neurol* **56**:822–34.

Smith DH, Nonaka M, Miller R, *et al.* 2000. Immediate coma following inertial brain injury dependent on axonal damage in the brainstem. *J Neurosurg* **93**:315–22.

Somerville A. 1961. Subarachnoid haemorrhage due to trauma without visible external injury. *Med Sci Law* **2**:67–9.

Spicer EJ, Strich SJ. 1967. Haemorrhages in posterior-root ganglia in patients dying from head injury. *Lancet* **2**:1389–91.

Spillane JD. 1962. Five boxers. *Br Med J* **2**:1205–6.

Storey PB. 1969. The precipitation of subarachnoid haemorrhage. *J Psychosom Res* **13**:175–82.

Strassman G. 1949. Formation of haemosiderin and haematoidin after traumatic and spontaneous cerebral haemorrhages. *Arch Pathol* **47**:205–10.

Strich SJ. 1961. Shearing of nerve fibres as a cause of brain damage due to head injury: a pathological study. *Lancet* **2**:443–7.

Strich SJ. 1970. Lesions in the cerebral hemispheres after blunt head injury. *J Clin Pathol Suppl R Coll Pathol* **4**:166–71.

Strich SJ. 1976. Cerebral trauma. In: Blackwood W, Corsellis JAN (eds), *Greenfield's neuropathology*, 3rd edn. Edward Arnold, London, pp. 327–60.

Symonds CP. 1940. Concussion and contusion of the brain. In: Brock (ed.), *Injuries of the skull, brain and spinal cord*. Baillière, London.

Symonds CP. 1962. Concussion. *Lancet* **1**:1–5.

Tatsuno Y, Lindenberg R. 1974. Basal subarachnoid hematomas as sole intracranial traumatic lesions. *Arch Pathol* **97**:211–5.

Terespolsky PS. 1972. Post-traumatic epilepsy. *Forensic Sci* **1**:147–65.

Thiagaraj D, Ming T, Cyn A. 1983. Patterns of injury in deaths due to falling from heights. *Proc First Asian Pacific Congr Legal Med*, p. 399.

Tomei G, Spagnoli D, Ducati A, *et al.* 1990. Morphology and neurophysiology of focal axonal injury experimentally induced in the guinea pig optic nerve. *Acta Neuropathol (Berl)* **80**:506–13.

Tomlinson BE. 1970. Brain-stem lesions after head injury. *J Clin Pathol Suppl R Coll Pathol* **4**:154–65.

Trotter W. 1914. Concussion. *Br J Surg* **2**:271–91.

Vanezis P. 1979. Techniques used in the evaluation of vertebral artery trauma at post-mortem. *Forensic Sci Int* **13**:159–65.

Vanezis P. 1984. *Vertebral artery trauma.* Medical Faculty, University of Bristol, Bristol.

Vanezis P. 1986. Vertebral artery injuries in road traffic accidents: a post-mortem study. *J Forensic Sci Soc* **26**:281–91.

Vanezis P. 1989. *Pathology of neck injury*. Butterworths, London.

Vanezis P, Chan KK, Scholtz CL. 1987. White matter damage following acute head injury. *Forensic Sci Int* **35**:1–10.

Voigt GE. 1981. Small hemorrhages in the brain stem. A sign of injury? *Am J Forensic Med Pathol* **2**:115–20.

Wang H, Duan G, Zhang J, *et al.* 1998. Clinical studies on diffuse axonal injury in patients with severe closed head injury. *Chin Med J (Engl)* **111**:59–62.

Weber W. 1984. Experimentelle Untersuchungen zu Schädelbruchverletzungen des Säuglings. [Experimental studies of skull fractures in infants] *Z Rechtsmed* **92**:87–94.

Wilkinson AE, Bridges LR, Sivaloganathan S. 1999. Correlation of survival time with size of axonal swellings in diffuse axonal injury. *Acta Neuropathol (Berl)* **98**:197–202.

Wilson JV. 1946. *The pathology of traumatic injury.* Livingstone, Edinburgh.

Yamashima T, Friede RL. 1984. Why do bridging veins rupture into the virtual subdural space? *J Neurol Neurosurg Psychiatry* **47**:121–7.

Yanagida Y, Fujiwara S, Mizoi Y. 1989. Differences in the intracranial pressure caused by a 'blow' and/or a 'fall' – an experimental study using physical models of the head and neck. *Forensic Sci Int* **41**:135–45.

Zhang X, Niu W. 1993. A study of enzymohisto-chemistry of cerebral cortical injury. *Forensic Sci Int* **59**:19–24.

Zuo ZJ, Zhu JZ. 1991. Study on the microstructures of skull fracture. *Forensic Sci Int* **50**:1–14.

CHAPTER 6

Chest and abdominal injuries

- Forensic anatomy 222
- Injuries to the chest wall 223
- Haemorrhage and infection in the chest 224
- Pneumothorax 225
- Injuries of the lungs 226
- Injuries of the heart 226
- Haemopericardium and cardiac tamponade 228
- Injuries to great vessels 228
- Foreign bodies in the gut 232
- Injuries to the pelvis and pelvic organs 233
- Complications of abdominal injury 233
- References and further reading 233

The major categories of wound previously described can be inflicted on any part of the body. In forensic practice certain areas are particularly vulnerable or have special medico-legal significance. Head injuries are so important that they have been discussed in a special chapter, but here consideration will be given to other regions, especially the chest and abdomen.

CHEST INJURIES

Damage can be sustained to either the chest wall or to the contents.

FORENSIC ANATOMY

The most important aspect is the relationship of the visceral contents to external landmarks. This is best described by reference to diagrams, where the relationship of pleural cavities, lungs, heart, mediastinum and diaphragm is depicted. It should be appreciated that, from the forensic aspect, the spleen and most of the liver and stomach are thoracic organs in that they lie largely beneath the costal margin, and are vulnerable to both stabbing and blunt injury to the chest.

Penetrating injuries, especially by knife to the lower lateral wall of the thorax, may enter the peritoneal cavity as well as

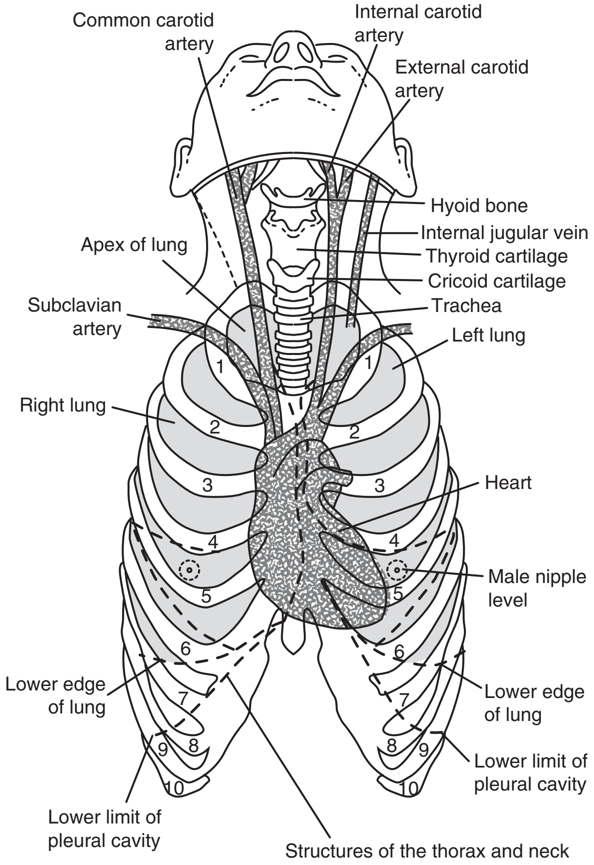

Common carotid artery
Internal carotid artery
External carotid artery
Hyoid bone
Internal jugular vein
Thyroid cartilage
Cricoid cartilage
Trachea
Apex of lung
Subclavian artery
Left lung
Right lung
Heart
Male nipple level
Lower edge of lung
Lower edge of lung
Lower limit of pleural cavity
Lower limit of pleural cavity
Structures of the thorax and neck
1
2
3
4
5
6
7
8
9
10

FIGURE 6.1 *Surface anatomy of heart, lungs and neck structures.*

the pleural spaces, perforating the diaphragm en route. The stomach, being largely within the rib cage, may often be penetrated. The common stab wounds of the heart may also include diaphragmatic and upper abdominal injuries. A knife can enter the front of the lower thorax, pass through either or both ventricles, and exit through the pericardial surface of the diaphragm to enter the upper surface of the liver. The orientation of the heart is often not appreciated after it is removed as an isolated organ, as the tendency is to visualize it as hanging with the apex downwards. In fact, it sits flat on the diaphragm on the lateral edge of the right ventricle, with the inferior vena cava passing immediately downwards into the abdomen.

INJURIES TO THE CHEST WALL

Respiration is dependent on the integrity of the rigid chest wall and, if the expansion is prevented or severely limited, then air entry will be correspondingly diminished. The integrity can be compromised either by severe mechanical failure of the rib cage or by penetration of the pleural cavities.

Fractures of ribs are common, but do not greatly embarrass respiration unless:

- they are so numerous that they prevent expansion of the thorax
- broken ends penetrate the pleura and lungs
- pleural and muscular pain limit respiratory effort.

Where many bilateral fractures are present, especially on the anterolateral sides of the thorax, the condition of ‘flail chest’ may be present, usually with multiple fractures of some ribs and sometimes with added fracture(s) of the sternum. As a result of loss of rigidity of the chest cage, attempts at expanding the thoracic volume during inspiration are impaired. The loose section is sucked inwards during inspiration, this clinical sign being known as ‘paradoxical respiration’. Dyspnoea and cyanosis may develop and extreme degrees of flail chest are rapidly incompatible with life because of progressive hypoxia.

The flail chest is caused by frontal violence, most often sustained in motor vehicle accidents – where the victim is thrown against the steering wheel or fascia – or in stamping assaults, where the shod foot is violently applied to the supine body. In any substantial chest injury, broken rib ends may be displaced inwards, the jagged tips ripping the parietal and visceral pleura. This may cause a pneumothorax or a haemothorax, or both, from penetration of the lungs, with the formation of a bronchopleural fistula. In gross chest injuries there may be compound fractures of ribs that allow a pneumothorax to form from external communication with the atmosphere, but this is rare in civil practice, though common in battle casualties.

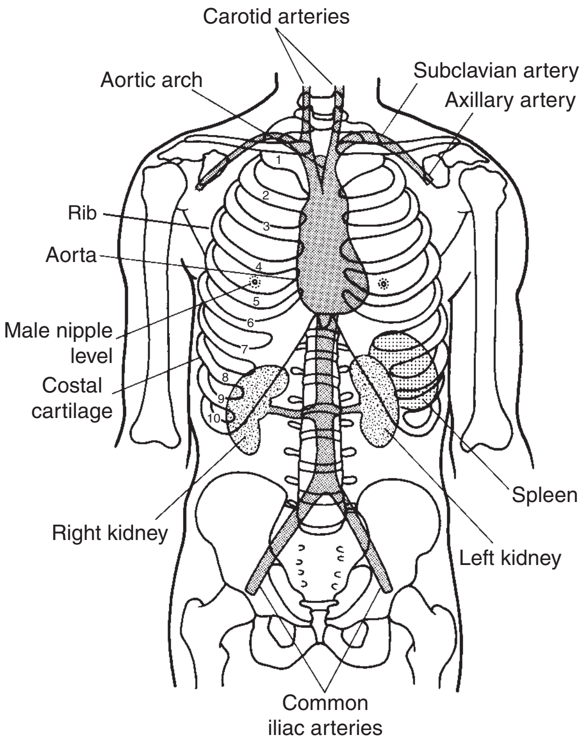

FIGURE 6.2 *Anterior view of trunk showing surface relations of heart and aorta.*

Rib fractures are most often seen in the anterior or posterior axillary lines caused by falls onto the side. The upper ribs are less often fractured, except by direct violence from kicking, heavy punching or traffic accidents. The fracture sites almost always show bleeding beneath the periosteum or the parietal pleura if the fractures occurred during life – though it must be admitted that (rarely) undoubted ante-mortem fractures may be totally bloodless, whereas some post-mortem cracks may exhibit slight oozing from the marrow cavity into the adjacent tissues. Attempts at resuscitation, especially external cardiac massage, now provide a common cause for extensive rib fractures (up to 40 per cent) and make the task of the pathologist much more difficult when trying to differentiate original trauma from the effects of enthusiastic first aid (Leadbeatter and Knight 1988). Bleeding may or may not be seen in these resuscitation fractures and, as the attempts at revival are, by definition, perimortal in timing, it is often impossible to say if they were immediately ante-mortem or post-mortem.

As the bracing action of adjacent intercostal muscles may conceal any mobility of the ribs when being examined at autopsy, it is a useful procedure to slit all the intercostal muscles with a knife when chest injury is suspected to allow any mobility to be detected more easily. In the osteoporosis of senility and some diseases, the ribs may be so fragile as to be breakable by finger pressure. Allowance must be made for this fragility in interpreting the cause of the fractures.

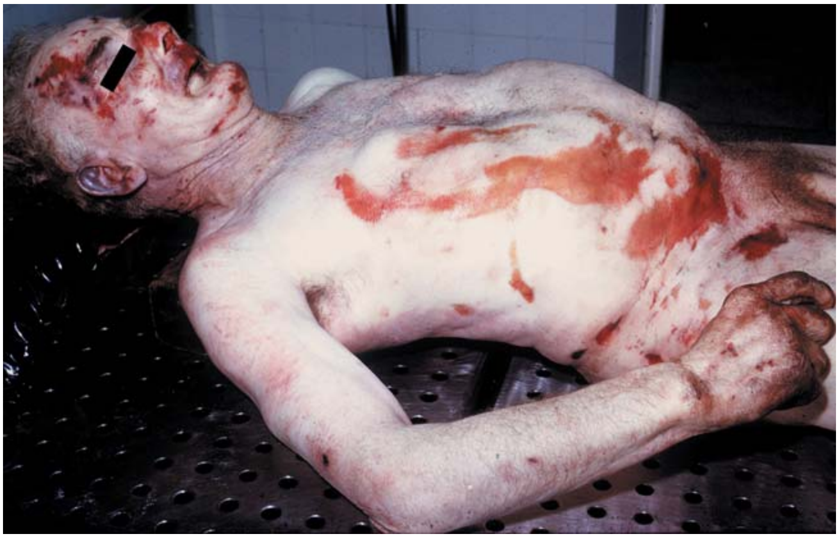

FIGURE 6.3 *Extensive abrasions and deformity of the chest in a man found by the roadside. All his ribs and sternum were fractured, resulting in a 'flail chest'. It was thought to be a road accident until bovine hairs were found on his clothing. He had been crushed by a bull and then tossed over the field wall.*

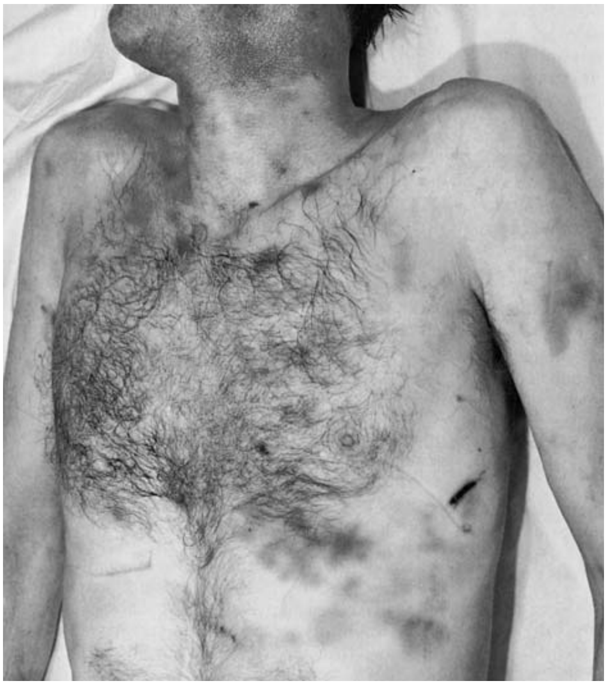

FIGURE 6.4 *Extensive bruising of the chest caused by kicking and stamping, resulting in multiple rib and sternal fractures. Over the left costal margin there is a pattern suggestive of the sole and heel of a shoe.*

In infants, especially victims of child abuse, rib fractures are common and may be an important diagnostic sign of abuse in doubtful cases. Where a small infant is squeezed from side to side, as when adult hands are clamped in each axilla or lower on the lateral sides of the chest, the hyperflexion can easily break ribs in their posterior segments, usually near their necks. The ribs are levered against the transverse processes of the vertebrae by excessive anterior flexion, which explains the tendency to fracture in the paravertebral gutter. Fresh fractures will be obvious, both on radiography and at autopsy. Within about 2 weeks (though this is very variable), callus will form and be visible both on X-ray and by direct post-mortem inspection. It is extremely difficult to date such callus.

It is said by paediatricians and radiologists that anterior rib fractures are rare in infancy other than from child abuse; though this is probably generally correct, care must be taken to exclude bony injury from the now almost universal attempts at resuscitation (even though infant ribs are very pliable). In very young infants, the possibility of older fractures dating back to birth injury cannot be dismissed, though again these are rare.

The sternum may be fractured by stamping or other frontal impacts, but far more force is necessary than with ribs. If posterior displacement of a fragment occurs, the underlying heart or great vessels may be severely damaged.

HAEMORRHAGE AND INFECTION IN THE CHEST

Any injury to the chest wall or lung surface that breaches blood vessels and the pleural lining can lead to a haemothorax. Intercostal and, less often, mammary arteries can bleed into the pleural cavities, but most massive haemorrhage comes from large vessels in the lung or mediastinum. The lung hilum can be torn or penetrated by stabs wounds. Another obvious source of a haemothorax is the heart itself, though there must also be a defect in the pericardial sac before the blood can reach the chest cavity. Several litres of blood may accumulate in the chest, either as liquid or clot, or usually a mixture of both. Death may occur from loss of circulating blood volume, even if there is relatively little external bleeding.

Many intrathoracic haemorrhages may be fatal with virtually no external blood loss, a fact that the police often find

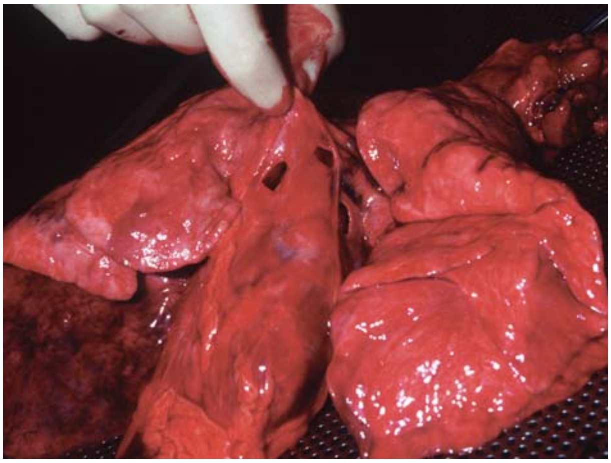

FIGURE 6.5 *Penetrating stab wound of chest, which has traversed the lung. The blade has passed through the upper lobe, crossed the interlobar fissure and re-entered the lower lobe. Death was caused by a haemothorax from severed branches of the pulmonary vessels within the lung.*

hard to credit if there is little blood at the scene of a fatal stabbing. A knife that passes obliquely into the chest through intercostal muscles may puncture a great vessel or heart chamber, allowing a fatal cardiac tamponade or haemothorax, yet the valve-like overlap of the tissues after withdrawal of the blade may seal up the external wound almost completely and prevent significant bleeding, especially as the blood inside the chest is not under any appreciable pressure.

The same may apply to a gunshot wound, where bleeding can be virtually absent if the wound is uppermost after death. Where the body lies above the wound, of whatever type, then considerable post-mortem gravitational leakage can take place.

Whatever the source of the bleeding, post-mortem haemorrhage can add very considerably to the volume found in the chest at autopsy. Due to the great variability of post-mortem coagulation – and subsequent lysis – much of the blood found at autopsy may not have been there at the moment of death. It is impossible to quantify this additional leakage in retrospect, but the ever-present possibility makes it unwise to be dogmatic about the amount of blood loss prior to death.

Infection following a chest wound is uncommon in forensic practice, as most deaths occur from haemorrhage within a relatively short time before infective sequelae have time to be established. Where survival is longer, then effective medical care usually prevents secondary infection. However, cellulitis, pleural inflammation and even empyema may supervene, especially where some dirty weapon is used, or where clothing or other foreign material has been carried into the wound. Infection may be of many types, but staphylococci, *Proteus*, coliforms and *Clostridium perfringens* are commonly found on culture.

PNEUMOTHORAX

There are three types of pneumothorax:

1. a simple type where a leakage through the pleura allows air to enter the pleural cavity, but where the communication rapidly closes. The lung partly collapses, but if death does not supervene the air is soon absorbed. If the communication remains open, then a bronchopleural fistula ensues with air in the pleural cavity but, as it is not under pressure, like type 2, it will not bubble out when the autopsy 'water test' is attempted. Radiology is then the best means of demonstrating the air in the pleural cavity.
2. when the leak in the pleura (or rarely the chest wall) has a valve-like action, air is sucked into the pleural cavity at each inspiration, but cannot escape on expiration. This pumping action leads to a 'tension pneumothorax', which causes complete collapse of the lung onto its hilum and a shift of the mediastinum to the opposite side. It is this type of pneumothorax that may be demonstrated at autopsy by penetrating an intercostal space under water, though radiology is a much better method of detection of all types of pneumothorax.
3. when an injury of the chest wall communicates with the pleural cavity, a 'sucking wound' may form with direct passage of air from the exterior. This type is most often seen in military surgery, and may be complicated by haemorrhage and infection.

A common traumatic cause of pneumothorax is a stab wound of the chest that allows direct communication with the exterior, though usually the layered skin and intercostal muscles closes the track when the weapon is withdrawn. The knife often enters the lung, however, so that air can enter the pleural cavity from the bronchi.

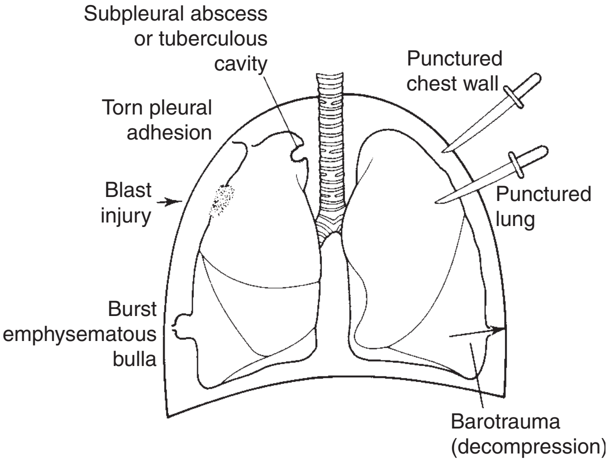
Subpleural abscess or tuberculous cavity
Punctured chest wall
Torn pleural adhesion
Blast injury
Punctured lung
Burst emphysematous bulla
Barotrauma (decompression)

FIGURE 6.6 *Common causes of pneumothorax.*

Natural disease can also cause a pneumothorax, which can lead to sudden death. Common causes are a ruptured emphysematous bulla, a tuberculous lesion at the lung periphery, or a tear at the site of a fibrous pleural adhesion. This may initially appear suspicious, as in a case seen by the author (BK) in which a teenage boy bent down to tie a shoelace and fell dead. At autopsy the only finding was a pneumothorax that had been caused by a single pleural adhesion that had torn the visceral pleura. This illustrates the fact that sometimes a pneumothorax can cause sudden unexpected death, presumably via some vagal cardioinhibitory mechanism.

INJURIES OF THE LUNGS

Bruising of the lungs is common in both open and closed chest injuries. Any substantial impact on the chest can contuse the lung surface or deeper parts. This may be beneath the area of impact or 'contrecoup' damage on the opposite surface. Deceleration injuries are seen in falls and traffic accidents, most commonly along the posterolateral surfaces where a vertical line of subpleural bruising occurs, often in the paravertebral gutter. The outline of ribs may be imprinted in lines of contusion on the pleural surface of the lungs. Bruising may be so severe as to form subpleural blood blisters, which may rupture to release blood or air into the pleural cavities.

In all severe chest injuries the central parts of the lung may show bleeding, sometimes sufficient to form actual haematomas with breakdown of lung tissue. Osborn describes 'pincer contusions' of the lung, where the expanded lower margins of the lungs become trapped in the narrow costophrenic angles.

Laceration of the lung can occur in blunt injuries and even lobes or parts of a lobe may be detached. The hilum may tear and the pulmonary ligament below the hilum is a frequent site of haemorrhage. Vessels in the hilum (especially pulmonary veins) or those more peripherally, may be ripped, causing severe intrapleural or mediastinal haemorrhage. In children, lung injuries can occur without fracturing of the ribs, because of the greater elasticity of the latter and the ability of the chest wall to deform.

Penetrating injuries of the lungs are common, usually from stabbing by knives. The wounds may end in the lung parenchyma or in large vessels, or may be 'through-and-through' injuries that emerge to cause further damage to the heart or great vessels. They assist the pathologist by delineating the track and the direction of the stabbing, but due allowance must be made for the marked variation in topography during inspiration, compared with the collapsed state seen at autopsy.

Blast injury is dealt with elsewhere, but the lungs are the most vulnerable organs to this type of injury due to their large tissue–air interface.

INJURIES OF THE HEART

The heart is vulnerable to both penetrating and blunt injuries. A common form of homicide is a stab wound of the chest which penetrates the heart. The entry point may be anywhere over the praecordium or adjacent areas if the angulation of the track is sufficient. Sometimes the sternum is penetrated by a forceful blow that reaches the underlying heart, but most stab wounds enter via the intercostal spaces, or through a rib or costal cartilage.

Rarely, an upward stab from the abdomen reaches under the costal margin to penetrate the diaphragm. The right ventricle is often injured by a stab wound as it presents the largest frontal area, but the anterior interventricular septum and the left ventricle are also vulnerable. A shallow stab wound may enter the myocardium and not reach the lumen of the ventricle. In such a case there may be little disability unless a coronary vessel is severed, which may either cause death from myocardial insufficiency (if a major artery is transected) or cardiac tamponade.

More often – especially in the right ventricle – the knife passes into the cavity. In the thin right ventricle this usually

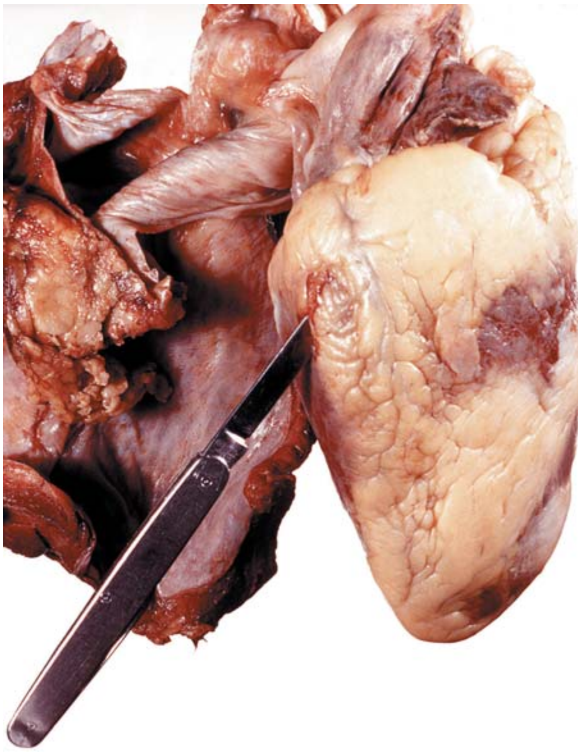

FIGURE 6.7 *Even a small penknife can inflict a fatal wound. This injury was caused by a boy to a police officer when the latter was apprehending him for some minor offence. Though the blade was only about 6 cm long, this is sufficient to enter alongside the sternum and puncture the ventricular wall.*

leads to copious bleeding into the pericardial sac even though the intraventricular pressure is relatively low because of the inability of the thin wall to close the defect by muscle overlap and contraction.

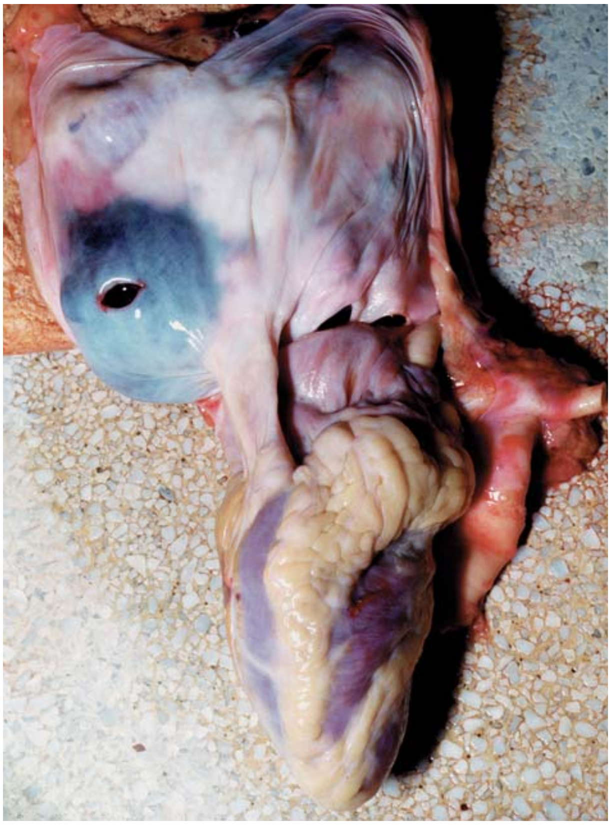

FIGURE 6.8 *Stab wound penetrating the pericardium (reflected upwards) with a wound in the corresponding point on the heart. Death was caused by cardiac tamponade from leakage of blood from the ventricular lumen.*

In the left ventricle, the contraction of the layered thick wall may partly or wholly seal the wound, and bleeding can be slight. It is, however, more common for persistent bleeding to occur and, if the drainage from the pericardial wound is less than the leakage from the ventricle, eventually a tamponade will develop. These variables make it impossible to calculate how long the victim was able to carry on with his activities, often a matter of dispute at a criminal trial. In general, wounds in the right ventricle are more dangerous than in the left because of the absence of the muscular 'self-sealing' effect.

Many stab wounds of the heart are transfixing or 'through-and-through' injuries, the knife entering one wall and emerging through another. If the left ventricle carries both wounds, the rapidity of bleeding may still not be torrential. Some of these wounds pass downwards and transfix the heart, then exit through the lower wall of the right ventricle and pass on via the diaphragm to end in the liver.

Blunt injuries of the heart are seen in civilian practice mainly in traffic accidents, falls from a height and in stamping assaults, though any heavy impact (including a punch) can cause fatal damage. There are usually multiple rib and sometimes sternal fractures, with or without a flail chest. Occasionally there can be heart damage in an intact chest cage, especially in a child with a pliable thorax. Fatal blunt damage of the heart may occur, however, without a mark on the skin of the thorax nor damage to the bony chest cage. The cardiac injuries are usually on the front of the organ, especially to the right ventricle, though posterior bruising and laceration can occur if the heart is compressed against the thoracic spine, as in stamping assaults and steering wheel impacts.

All degrees of damage can occur, from mere epicardial bruising to lacerations that open the ventricular lumen widely.

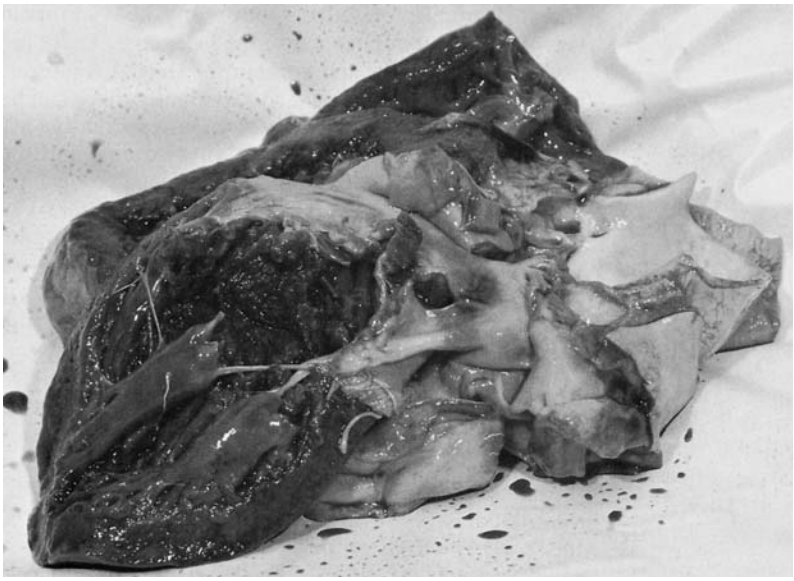

FIGURE 6.9 *Ruptured interventricular septum as a result of impact on the front of the chest. The septum can tear without damage to the outer walls of the heart.*

Internal damage to the heart may be present, without external signs of injury to either the heart or the external chest wall. The interventricular septum may be ruptured as a sole lesion, the mechanics of which seem obscure. In gross injuries, such as aircraft crashes, the whole heart may be avulsed from its root, to be found lying loose in the thorax. Histological damage, especially contraction band necrosis, can be found even when there is no macroscopic abnormality. A comprehensive review of cardiac contusions has been made by Sagall (1971).

HAEMOPERICARDIUM AND CARDIAC TAMPONADE

Bleeding into the pericardial sac may occur from the surface or the cavities of the heart, or from the intrapericardial segments of the roots of the great vessels, particularly the aorta and pulmonary artery. Most causes of haemopericardium are from natural disease, such as a ruptured myocardial infarct or a ruptured dissecting aneurysm of the aorta (Chapter 25), but it is not an uncommon sequel to injury to the chest. When the damage has been caused by a stab that has perforated the pericardial sac, bleeding can escape into the pleural cavities, mediastinum or even abdomen if the diaphragm is penetrated. Death may occur from sheer blood loss if the haemorrhage can escape from the confines of the pericardium, but another common lethal condition is ‘cardiac tamponade’.

In the tamponade, blood accumulates in the pericardial sac faster than it can escape, either because the bleeding rate exceeds the drainage or because the exit hole in the pericardium becomes blocked by blood clot. In cases where the bleeding is from a contusion or laceration of the heart, there is no escape route from the sac. When sufficient blood accumulates, the pressure in the pericardial sac increases and begins to prevent the passive filling of the atria during diastole. The cardiac output falls, as does the systemic blood pressure and the venous pressure rises. If unrelieved, death will follow, though the time that this takes is variable and almost impossible to calculate retrospectively on pathological findings. It has been stated (Moritz 1942) that about 400–500 ml of blood is sufficient to cause death, though this seems a greater volume than is usually seen in tamponades. Where trauma to the heart co-exists, it is difficult to apportion the relative contribution to death of the tamponade as opposed to the other injuries.

INJURIES TO GREAT VESSELS

The most vulnerable vessel is the aorta, which most commonly suffers injury in deceleration trauma from both road and air accidents, as well as from falls from a height. When the thorax is suddenly decelerated, the heart – being relatively mobile in the chest – attempts to continue in the original direction. This causes severe traction on the root of the heart, and a common sequel is complete or partial rupture of the aorta in the descending part of its arch. In falls from a height, Fiddler (1946) claims that the lesion is the result of the abdominal and thoracic viscera being forced caudalwards by the abrupt deceleration when landing on the feet or buttocks. Tannenbaum and Ferguson (1948), however, postulate that the mechanism is a sudden rise in intra-aortic pressure. Lasky (1974) has reviewed the biomechanics of impact injury of the aorta.

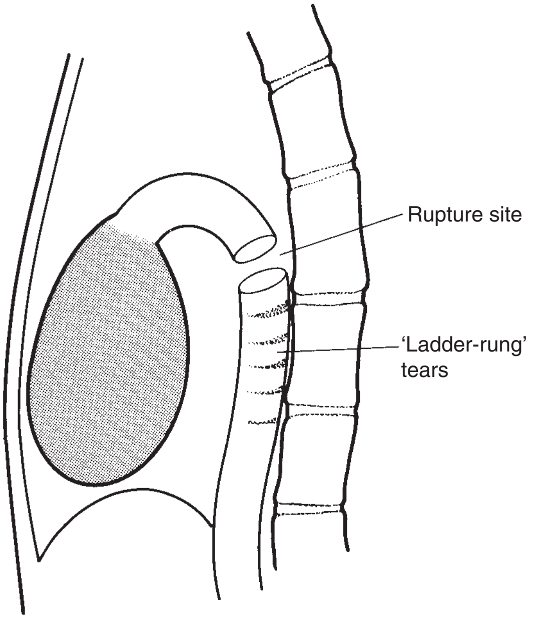

FIGURE 6.10 *The usual site of rupture of the aorta in deceleration injury.*

According to Fiddler, rupture occurs almost constantly at a point 1.5 cm distal to the attachment of the ligamentum arteriosum, the remnant of the ductus. The lower thoracic aorta is closely bound to the anterior longitudinal ligament on the front of the dorsal spine, until it reaches the termination of the arch, where it curves forwards. This appears to be the weak point and transection occurs at this level, sometimes so cleanly that it looks like a surgical incision. The tear is annular and at right angles to the axis of the aorta. Sometimes, there may be multiple parallel intimal tears near the main transection, the so-called ‘ladder-rung tears’. In deceleration trauma, these incomplete tears, which affect only the intima and inner media, may be found without major transection; where death is delayed, false aneurysms and dissections may be diagnosed on aortograms.

The pulmonary artery is much less vulnerable to blunt trauma than the aorta but, in stamping assaults and steering

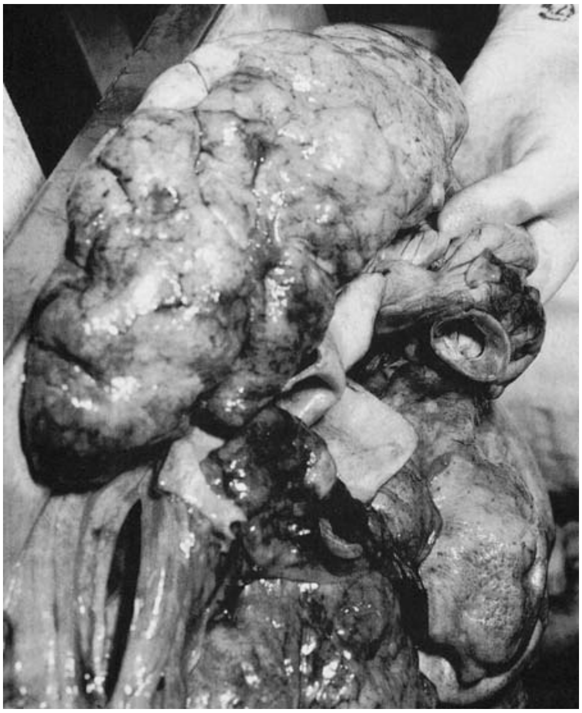

FIGURE 6.11 *Complete transection of the aorta in the distal part of the arch. The victim was a car driver who suffered a severe deceleration impact. There is also some slight pulmonary bruising; the lungs are asthmatic.*

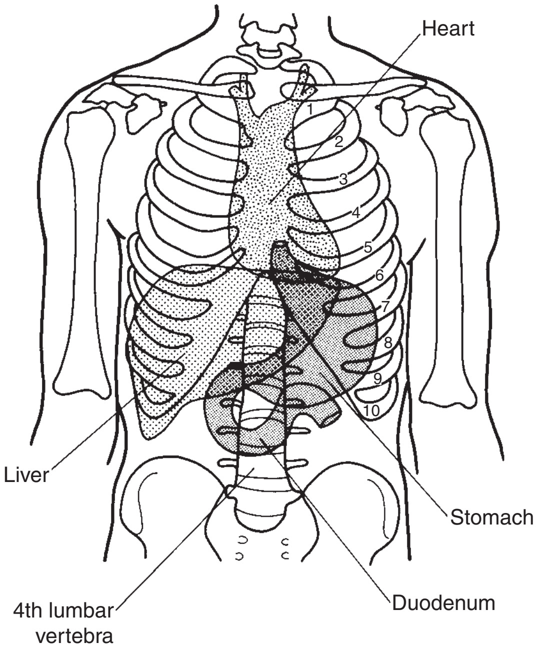

FIGURE 6.12 *Surface anatomy of liver, stomach and heart.*

wheel impacts, it may be damaged by depressed rib cage and sternal fractures. The pulmonary artery and vein branches may also be damaged in the root of the lung, where hilar tears are not uncommon. The great vessels are often involved in penetrating injuries, notably stab wounds. Stabs of the upper part of the chest may pass directly into the arch of the aorta, especially on the right side of the sternum. Here even a shallow injury may reach the aorta and the authors have seen several cases in which a short-bladed penknife has caused death in this way.

Stabs that are either too high, or are directed too laterally to puncture the chambers of the heart may penetrate the ascending aorta or the pulmonary artery. If the wound is below the reflection of the pericardium, a haemopericardium and perhaps cardiac tamponade may result. Other stab wounds may injure the heart valves or may enter the root of the lung to penetrate the large primary branches of the pulmonary artery or veins, causing gross bleeding into the corresponding pleural cavity or the mediastinum.

ABDOMINAL INJURIES

As with the thorax, the damage caused by blunt and penetrating trauma of the abdomen depends upon the location of the injury. In addition, the large area of the anterior abdomen occupied by the intestine provides a target for perforation with consequent chemical or infective peritonitis.

Open or penetrating wounds need little description, as they follow the characteristics of stab-like injuries in general. The liver, and especially the spleen, may bleed extensively, causing a haemoperitoneum. The intestine and mesentery are the other major targets, wounds often being multiple, because of the overlapping nature of the coils and their mesentery. The stomach, being partially protected by the rib margin, is less often penetrated from the abdomen, but is not uncommonly involved in chest stabbings that pass downwards through the diaphragm. The kidneys are rarely stabbed except from a thrust in the back.

Closed or blunt injury to the abdomen is common from both accidents and assault. Impact on the abdomen by a car steering wheel was more common before the widespread use of safety-belts and air-bags and still occurs in severe deceleration accidents. The liver, intestine, spleen and mesentery are most vulnerable. Crushing between two vehicles, or between a vehicle and a wall is another mechanism for abdominal trauma, and may also be seen in railway and industrial accidents where squeezing between two opposing surfaces occurs. In homicides, assaults and child abuse, kicking, stamping and heavy punching can also cause blunt injury to the abdomen.

Whatever the mechanism of infliction, the following features may be present:

- Bruising of the abdominal wall, both of the skin and the underlying muscles, is often (but by no means

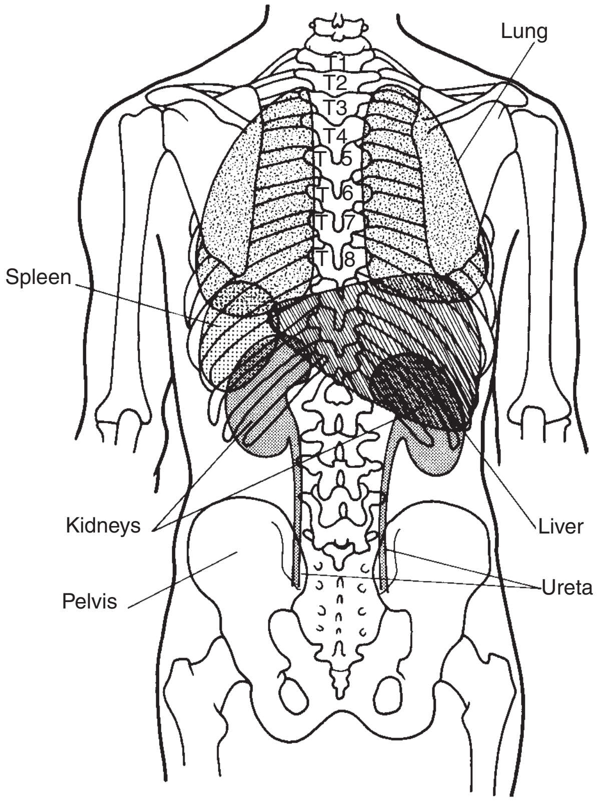

FIGURE 6.13 *Posterior view of trunk to show surface relations of organs.*

invariably) seen in abdominal trauma. Where subcutaneous bleeding is profuse it may track from the initial area of impact to become more diffuse and may cover a large area of abdominal wall, especially in the lower segment. Blood may track down the inguinal canal and appear in the scrotum or labia.

More often there are discrete bruises on the abdominal skin, associated in some instances with surface abrasions. Kicks will usually leave scuffed abrasions if tangential, unless protected by clothing. Fingertip or knuckle bruises may be seen, especially in child abuse. In infants, bruises may be seen on the sides of the abdomen where adult fingers have forcibly gripped or lifted the child, though these are more common in the axillary line of the chest.

Severe or fatal intra-abdominal injury may be present, however, without any mark on the skin. This can occur if clothing protects or if a blunt impact is applied over, a relatively large surface. In child abuse, for example, the liver, mesentery or duodenum can be ruptured with no external sign whatsoever.

- Extensive bleeding into the peritoneal cavity, usually from rupture of a solid viscus or bleeding from the mesentery.
- Bruising or rupture of the stomach and diaphragm. The stomach is less vulnerable than the intestines, but may be lacerated by a heavy blow in the upper abdomen, especially when full of food or fluid.
- The intestine and its mesentery are frequently damaged in abdominal trauma. Extensive bruising of the gut and its vascular mesentery may occur, mainly from being crushed against the prominent lumbar vertebrae in the midline. The duodenum and jejunum are particularly vulnerable to transection from being compressed against the spine, especially in children, where a heavy blow in the central or upper abdomen can cut through the third part of the duodenum almost as cleanly as a scalpel. Laceration of the mesentery is not uncommon both in traffic accidents and assaults. The same mechanism of compression against the lumbar spine causes bruising and tearing of the central parts of the mesentery, which usually occurs towards the intestinal margin of the membrane. Multiple fenestrations may be seen, presumably because folds of mesentery overlie each other and are injured when in apposition.

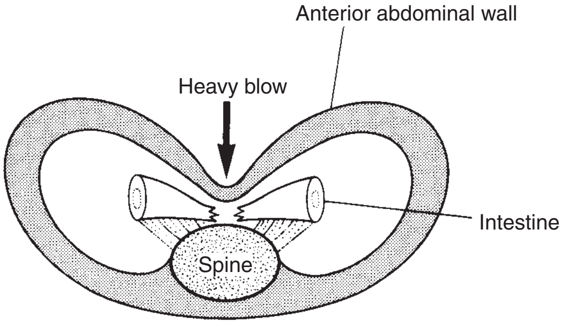

FIGURE 6.14 *Injury or rupture of intestine or mesentery in violent blow to the upper central abdomen. Because of its small size and the thin abdominal wall, the infant duodenum is especially vulnerable.*

Severe and intractable bleeding can occur in circumstances where surgical treatment is often not offered because the condition is unrecognized. Kicks and punches in the abdomen may be sustained by drunken persons who may be apparently unaware of the seriousness of their injuries until collapse and death supervene perhaps hours later.

In a case seen by the author (BK), a man was involved in a fight and then allegedly assaulted the police. He was forcibly restrained and may have been struck in the abdomen by a knee. He remained apparently well, though drunk, in a police cell for more than 5 hours, before collapsing and dying on his way to the lavatory, never having complained of any abdominal pain. Autopsy revealed several litres of blood in his peritoneal cavity, which had escaped from several large tears in the mesentery. Similar bleeding from mesenteric lacerations has been seen in abused infants,

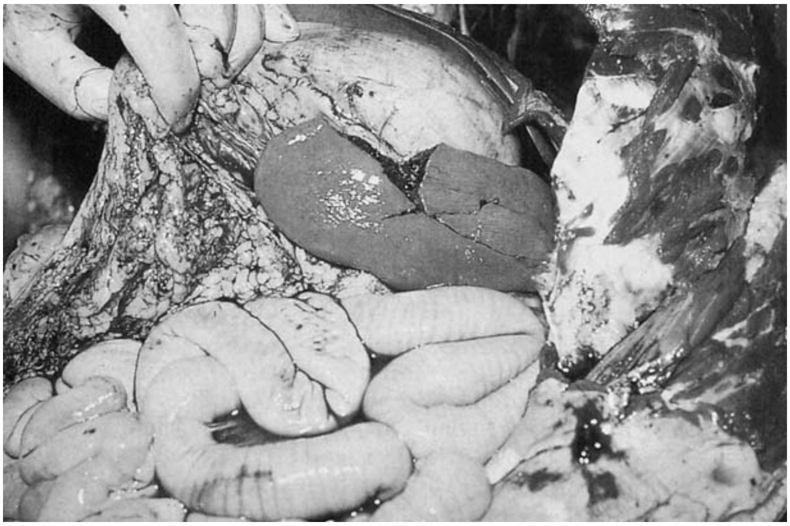

FIGURE 6.15 *Laceration of the spleen following lower chest impact during a road traffic accident.*

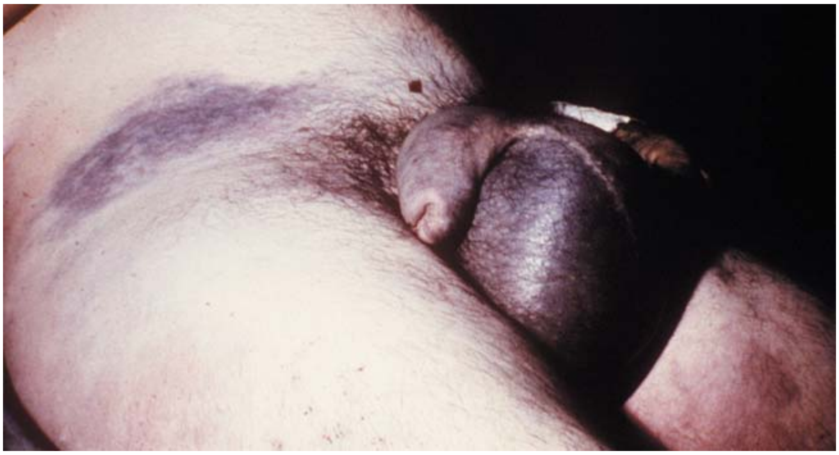

FIGURE 6.16 *Abdominal wall bruising with fatal intraperitoneal haemorrhage from a ruptured mesentery. There is also a large haematocele from a gang attack in which repeated kicking occurred.*

where the thin abdominal wall provides little protection from blows on the abdomen.

Injury to the mesentery may damage local arteries without causing severe bleeding, but may occlude or thrombose them, with infarction of the bowel as a consequence. Perforation may be delayed for a day or two and in children and old people there need not be the dramatic symptoms of the onset of peritonitis that is usually seen in clinical surgical practice.

The colon is rarely injured unless the abdominal injury is gross. One well-known lesion is rupture of the rectum or sigmoid by a high-pressure air hose being placed in or near the anus as a practical joke.

- Rupture of the spleen is a common surgical emergency after trauma but may be first found at autopsy if undiagnosed, or if death occurs from lack of speedy surgical treatment, or if other injuries made recovery impossible. An enlarged spleen is more vulnerable and more fragile than a normal organ. Malaria, glandular fever and other infections increase the risk of rupture. The 'stiletto' dagger was specifically designed in medieval Italy for puncturing the enlarged malarious spleen by an upward blow under the left costal margin.

The spleen can be damaged either from impact on its surface or from traction on its pedicle. The rupture may be immediate or may be delayed if a large subcapsular haematoma and underlying tear are held intact for a time by the capsule. This delay can last for a number of days or even weeks.

- Rupture of the liver is also a common lesion following serious abdominal trauma, such as fall from a height or a crush injury between two wagons. It is seen especially in traffic accidents, either from impact of the driver on the rim or centre of the steering wheel or by the unrestrained passenger being thrown against the fascia. A pedestrian can also suffer a ruptured liver either as a primary impact from a vehicle or as secondary damage from being thrown to the ground.

A more recent cause of liver injury is enthusiastic external cardiac massage. Even if death has already taken place, there may still be a significant amount of blood oozing from the liver tears into the peritoneal

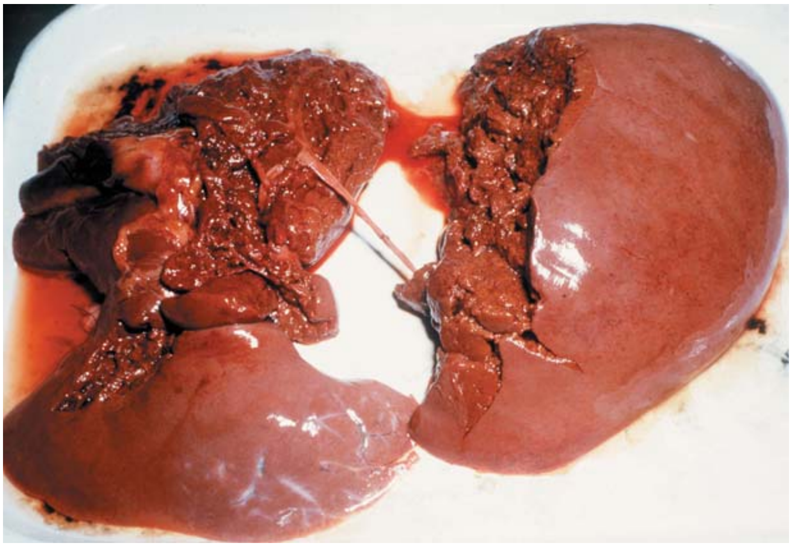

FIGURE 6.17 *Complete bisection of the liver in a high-speed deceleration during a traffic accident. The abdomen of the driver was hurled against the steering wheel as he was projected through the windscreen.*

cavity. Moritz (1942) has described six varieties of hepatic laceration, but such a differentiation is of little practical use. The liver may show one or more linear cracks, most often on the convex upper surface, which may be of any severity from superficial subcapsular tears to complete transection of the organ. Not infrequently, several roughly parallel tears are seen on the upper surface, obviously from the same type of mechanical stress. The surface lacerations may extend deeply into the liver and may even appear on the opposite surface. Occasionally, there may be internal tears that do not communicate with the surface. As in the spleen, a subcapsular haematoma can reach a large size before bursting into the peritoneal cavity, causing a delay in collapse or death.

Infants readily suffer liver damage from trauma and it is a well-known component of the child abuse syndrome. Liver tears can occur during the process of birth, particularly in breech presentation.

- Injuries to the kidney. Being deeply situated at the back of the abdomen in the paravertebral gutter, the kidneys are rarely damaged by frontal trauma, but may be involved in kicks or heavy blows to the loin. Traffic impacts, mainly in pedestrians, can damage the kidney, but kicking is another cause of this rather uncommon injury. When a victim is lying on the ground, a swinging boot can conveniently enter the slightly raised arch between the rib margin and hip, where the kidney is situated.

Perirenal haemorrhage is much more common than damage to the actual organs themselves, but when they are injured, all grades of damage from comminuted pulping through transection to shallow surface lacerations may be seen. Like the mesentery and spleen, damage to vessels may lead to post-injury infarction. The suprarenal vessels can be damaged by impact and, statistically, the right suprarenal artery is more often involved in traffic accidents in Britain because the direction of traffic tends to favour lesions on that side. It is, however, difficult to differentiate direct trauma to the suprarenal vessels from the common post-stress haemorrhages, often bilateral, of the Waterhouse–Frederichsen type. These occur quite commonly a few days after any trauma, surgical operation or infection, and are rarely diagnosed by clinicians (Knight 1980) (see Chapter 13).

FOREIGN BODIES IN THE GUT

Though not usually a feature of trauma, foreign bodies and material are not infrequently found in the alimentary system in forensic practice, both clinically and at autopsy.

Disorders of the mental state may lead to extraordinary numbers of objects being swallowed, which may be present anywhere from mouth to anus, but especially in the stomach. Betz *et al.* (1994) describe two such deaths, where, in one, 2000 cm^3 of debris, including broken glass, was in the alimentary tract; the other suffered intestinal obstruction due to ingestion of a large quantity of his own hair.

Similar aberrations may lead to insertion of foreign objects into the rectum, urethra, bladder and vagina. The authors have seen large objects of various types, e.g. potatoes in the rectum, lengths of broom handle and wood in the colon (one penetrating the liver) and a banana and a beer can in the vagina. Sexual perversions, masturbation and heterosexual or homosexual activities may account for these, as well as frank mental abnormality.

In recent years, the smuggling of narcotics through ports and airports has commonly been accomplished by 'body-packing', the drugs being concealed in condoms, which are then either swallowed or, less often, inserted into the rectum or vagina. Deaths have occurred in these body-packers, usually by rupture of a package within the gut, so that a massive dose of heroin or cocaine is released with all the attendant toxic effects. Rarely, the mass of illicit material may cause intestinal obstruction, leading to a surgical emergency, which the carrier does not wish to be investigated, for obvious reasons.

INJURIES TO THE PELVIS AND PELVIC ORGANS

The bony pelvis suffers a variety of fractures and dislocations in severe trauma:

- Where great pressure is applied to the front of the abdomen or pubic area – as in running over by a vehicle wheel – the pelvis may be splayed open, the symphysis parting and one or both sacroiliac joints becoming dislocated.
- An impact from the side may shatter the superior and, less often, the inferior pubic ramus, and again dislocate the sacroiliac joint on that side.
- A fall from a height onto the feet may transmit the force up the legs either to dislocate the hips or even to drive one or both femoral heads through the acetabulum. If the hip joints remain intact, the pelvic girdle may crack and again the sacroiliac joints may be sheared apart.
- A kick or heavy fall onto the base of the spine may fracture the coccyx or sacrum.

The pelvic organs with the exception of the bladder are generally protected from all but the most severe trauma. When it is full, the bladder can be ruptured by heavy blows or kicks on the lower abdomen. The empty bladder is rarely injured by blunt trauma. The male urethra is also vulnerable to direct trauma, such as falling astride a solid object like a gate, or being kicked in the crutch, both of which compress the urethra against the undersurface of the pubis.

The external genitals may suffer injury, apart from the sexual assaults discussed in Chapter 18. The scrotum is vulnerable to severe bruising, especially from kicks: a large haematocele may develop. Scrotal and vulval injuries can occur from falling astride an object or in traffic accidents. Impact of a bicycle wheel between the legs is one such cause.

COMPLICATIONS OF ABDOMINAL INJURY

The most common fatal sequel to intra-abdominal trauma is haemorrhage from any of the contained organs. The spleen and mesentery tend to bleed most copiously and quickly, though even here there can be a delay of many hours before serious symptoms are obvious – and, in the case of a subcapsular laceration of the spleen, the time can be far longer. The liver tends to ooze more slowly unless a major injury opens a large vessel or an extensive area of hepatic tissue.

The mesentery contains numerous blood vessels that are not covered by parenchymatous tissue as in the liver or spleen, so bleeding is usually brisk from significant tears – although, as illustrated by the case described above, even this injury can lead to delayed death.

Because all bleeding in the abdomen, unless it is from a wound of the aorta, takes time to accumulate, restriction of the victim's activity immediately after the injury need be minimal or absent unless there are other factors involved. It is never safe for a pathologist to declare that rapid immobility must have ensued, unless there are other gross injuries.

As mentioned in relation to intrathoracic bleeding, some of the blood found at autopsy within the peritoneal cavity may have accumulated there after death, adding to the true ante-mortem volume.

In spite of common surgical experience, even severe intra-abdominal injuries leading to bleeding need not necessarily be painful, especially in the commonly associated circumstance of the victim being under the influence of alcohol.

Perforation of the gastrointestinal canal is another serious complication of trauma, though often death ensues from the totality of injuries before infection can become established. As in natural perforated peptic ulcer, penetration of the stomach or duodenum will cause a chemical peritonitis that can be a cause of severe and immediate shock – perhaps a major factor in a rapid death. Rupture of the small or large intestine is less damaging but, if survival continues, will inevitably lead to generalized peritonitis unless vigorously treated. Open wounds of the abdomen, including stabbing, will introduce organisms from the outside into the peritoneal cavity. In addition to infection, trauma to the intestine may cause an intractable ileus; if the pancreas is damaged, there may be widespread fat necrosis in the mesentery and omentum.

REFERENCES AND FURTHER READING

Adams JE, Davis GG, Heidepriem RW, 3rd, *et al.* 2002. Analysis of the incidence of pelvic trauma in fatal automobile accidents. *Am J Forensic Med Pathol* **23**:132–6.

Albrektsen SB, Thomsen JL. 1989. Detection of injuries in traumatic deaths. The significance of medico-legal autopsy. *Forensic Sci Int* **42**:135–43.

Betz P, van Meyer L, Eisenmenger W. 1994. Fatalities due to intestinal obstruction following the ingestion of foreign bodies. *Forensic Sci Int* **69**:105–10.

Bowen DA. 1970. A survey of injuries to the liver and spleen in forensic autopsies. *J Forensic Med* **17**:12–19.

Canfield TM. 1969. Suicidal gunshot wounds of the abdomen. *J Forensic Sci* **14**:445–52.

Cohle SD, Hawley DA, Berg KK, *et al.* 1995. Homicidal cardiac lacerations in children. *J Forensic Sci* **40**:212–18.

Cotton JM, Cooke JC, Monaghan MJ. 2000. Forensic echocardiography: a case in point. *Echocardiography* **17**:193–4.

Darok M, Beham-Schmid C, Gatternig R, *et al.* 2001. Sudden death from myocardial contusion following an isolated blunt force trauma to the chest. *Int J Legal Med* **115**:85–9.

Davis C. 1963. Spontaneous rupture of the stomach. *Arch Surg* **32**:170–6.

Denton JS, Kalelkar MB. 2000. Homicidal commotio cordis in two children. *J Forensic Sci* **45**:734–5.

Fiddler J. 1946. *The pathology of traumatic injury.* Livingstone, Edinburgh.

Frazer M, Mirchandani H. 1984. Commotio cordis, revisited. *Am J Forensic Med Pathol* **5**:249–51.

Gunther WM, Symes SA, Berryman HE. 2000. Characteristics of child abuse by anteroposterior manual compression versus cardiopulmonary resuscitation: case reports. *Am J Forensic Med Pathol* **21**:5–10.

Hale HW. 1965. Pulmonary contusions. *J Trauma* **5**:647–52.

Knight B. 1980. Sudden unexpected death from adrenal haemorrhage. *Forensic Sci Int* **16**:227–9.

Lasky I. 1974. *Human aortic laceration due to impact.* In: Wecht C (ed.), *Legal medicine annual*. Appleton Century Crofts, New York.

Leadbeatter S, Knight B. 1988. Resuscitation artefact. *Med Sci Law* **28**:200–4.

Lloyd RG. 1982. Delayed rupture of stomach after blunt abdominal trauma. *Br Med J Clin Res Ed* **285**:176.

Loop FD, Hofmeir G, Groves LK. 1971. Traumatic disruption of the aortic valve. *Cleveland Clin Q* **38**:187–94.

Mackintosh AF, Fleming HA. 1981. Cardiac damage presenting late after road accidents. *Thorax* **36**:811–13.

Maron BJ, Poliac LC, Kaplan JA, *et al.* 1995. Blunt impact to the chest leading to sudden death from cardiac arrest during sports activities. *N Engl J Med* **333**:337–42.

McCormick GM, 2nd, Young DB. 1995. Spontaneous rupture of the spleen. A fatal complication of pregnancy. *Am J Forensic Med Pathol* **16**:132–4.

Menzies RC. 1978. Cardiac contusion: a review. *Med Sci Law* **18**:3–12.

Moritz AR. 1942. *The pathology of trauma*. Lea & Febiger, Philadelphia.

Murphy GK. 1985. A single fatal penetrating chest wound from shattered wind-blown glass. *Am J Forensic Med Pathol* **6**:332–5.

Noon GP, Boulafendis D, Beall AC, Jr. 1971. Rupture of the heart secondary to blunt trauma. *J Trauma* **11**:122–8.

Ormstad K, Rajs J, Calissendorff B, *et al.* 1984. Difference between in vivo and postmortem distances between anterior chest and heart surface. A combined autopsy and in vivo computerized tomography study. *Am J Forensic Med Pathol* **5**:31–5.

Orr CJ, Clark MA, Hawley DA, *et al.* 1995. Fatal anorectal injuries: a series of four cases. *J Forensic Sci* **40**:219–21.

Ortmann C, Pfeiffer H, Brinkmann B. 2001. Immunohistochemical alterations after intravital and post-mortem traumatic myocardial damage. *Int J Legal Med* **115**:23–8.

Osborn GR. 1943. Findings in 272 fatal accidents. *Lancet* **245**:277–84.

Puffer P, Gaebler M. 1991. [Traumatic diaphragmatic rupture in a forensic medicine autopsy sample.] *Beitr Gerichtl Med* **49**:149–52.

Sagall EL. 1971. Contusion of the heart: medical and legal considerations. In: Wecht C (ed.), *Legal medicine annual*. Appleton Century Crofts, New York.

Simonsen J. 1983. Injuries sustained from high-velocity impact with water after jumps from high bridges. A preliminary report of 10 cases. *Am J Forensic Med Pathol* **4**:139–42.

Stajduhar-Djuric Z. 1960. Penetrating wounds of the ascending aorta and heart. *J Forensic Med* **7**:147–52.

Tannebaum I, Ferguson JA. 1948. Rapid decelaration and rupture of the aorta. *Arch Pathol* **45**:503–5.

Teresinski G, Madro R. 2001. Pelvis and hip joint injuries as a reconstructive factors in car-to-pedestrian accidents. *Forensic Sci Int* **124**:68–73.

CHAPTER 7

Self-inflicted injury

- Suicidal injuries 235
- Self-inflicted injuries other than suicide 240
- References and further reading 243

Injuries that are deliberately self-inflicted are common and their examination is a frequent task for both pathologists and clinical forensic practitioners. These events consist of suicide, attempted suicide and suicidal gestures, the latter lacking the intention to kill though death may inadvertently ensue. In addition, there is non-suicidal self-inflicted trauma, as described below.

One of the most difficult decisions that can face pathologists, medical examiners and legal authorities, such as coroners, is the differentiation between homicide, suicide, accident and other self-inflicted injury. Though it is not the legal function of a pathologist to attribute motive, his experience and training are often the factors that lead the official authority to make a decision as to the classification of the manner of death or injury. One such situation is fatal auto-erotic activities, described in Chapter 15.

In addition, when a pathologist is assisting the police in the early stages of a death investigation, his opinion and decision about the manner of the death may be crucial in initiating or aborting a homicide investigation – an onerous decision which may have serious and expensive consequences if wrong. It is therefore a most important function of the forensic pathologist to decide on the evidence of a death scene and subsequent autopsy whether murder, suicide or accident is the most likely explanation.

Self-inflicted injury, when not suicidal, is often not fatal and the pathologist is usually not consulted unless he also includes clinical forensic medicine in his work. The pathologist may, however, be asked for his opinion, as he is a specialist in trauma and may be able to provide the most useful opinion available. In other cases, self-inflicted injury, which the victim did not intend to be fatal, may end in death from a variety of causes and thus the pathologist will be directly involved.

SUICIDAL INJURIES

This discussion is confined to physical trauma, self-poisoning being considered later in the book.

Suicides may injure themselves in many different ways, some bizarre in the extreme. The pathologist must always be alert to the possibility that such injuries are not true suicides, but some manifestation of peculiar practices that have taken a course unintended by the victim. A prime example is masochistic asphyxia, which is still sometimes mistaken for suicide (Chapter 15). Even more serious is the homicide that appears to be a suicide, though the reverse also occurs.

Suicides, apart from poisoning, may use one (or even more) of the following methods, though the list is by no means exclusive:

- stabbing and cutting
- firearms and explosives
- jumping from a height
- jumping into water
- burning
- suffocation, e.g. plastic bag
- hanging and strangulation
- electrocution
- road and railway injuries.

Most of these modes are discussed in the appropriate chapters.

Suicidal knife wounds

Many incised and stab wounds are self-inflicted, either from motives of self-destruction, from mental aberrations or deliberately for some form of gain. These must be differentiated from each other, and from accidental and homicidal wounding.

Sometimes the distinction may be difficult and even impossible, but the experience of the forensic pathologist is paramount in assisting investigative agencies to come to the correct conclusion. Suicide, attempted suicide and suicidal gestures commonly employ cutting weapons as the means of injury. There are certain features that are quite reliable pointers to these motives, though they are by no means infallible:

- Suicidal knife wounds favour certain 'sites of election', which are predominantly the throat, the wrists and the front of the chest. As with most violent methods, cut throat and chest stabbing is predominantly seen in men – though cutting wrists is not uncommon in women.
- Suicidal wounds are typically multiple, often being characterized by a number of preliminary trial cuts, called 'tentative incisions'. These are most often seen on the throat and wrists, where the person often makes a series of shallow incisions, presumably hesitating while gaining courage to make a final decisive cut. In many suicidal attempts the subject abandons this method after a few trial incisions and uses some other method of self-destruction. Although the presence of tentative incisions is strong presumptive evidence of suicide, exceptions do occur, and the pathologist must take all other aspects of the scene into account before giving an opinion to the investigators. In fact, the almost absolute statements of some older textbooks that tentative incisions must indicate suicide, is to be disregarded. The author (BK) has dealt with two recent murders in which 'typical' suicidal injuries were present on wrist or throat causing considerable doubt to the investigators until other factors clarified the situation.
- A suicidal cut throat usually has these trial incisions; there may be only one or two, or there may be scores of trial cuts. If successful, there will be one or more deep incisions superimposed, which may destroy some of the previous shallow cuts. The classical description of the cut throat is of incisions starting high on the left side of the neck below the angle of the jaw, which pass obliquely across the front of the neck to end at a lower level on the right. This assumes that the victim is right-handed, the obliquity being reversed in a left-handed person. The cuts are said to be deeper at their origin, becoming shallower as they cross the throat, tailing off into surface cuts at the extremity.

  This description, though hallowed by repetition in many textbooks, is often incorrect and many cut throats have horizontal cuts that show no variation in depth at either end. Most suicides appear to raise the chin to provide better access to the throat, so that the skin is stretched when cut. This tends to cause straight-edged incisions, rather than the jagged cut (the so-called 'dentele' toothed incision) seen when a knife is drawn over loose skin. Throwing back the head moves the carotid bundle under the protection of the sternomastoid muscles and, if the cuts are confined to the centre of the front of the neck, only the larynx or trachea may be damaged, rather than large blood vessels.

  Death from a cut throat depends on the nature and extent of local damage to the neck. Severe haemorrhage

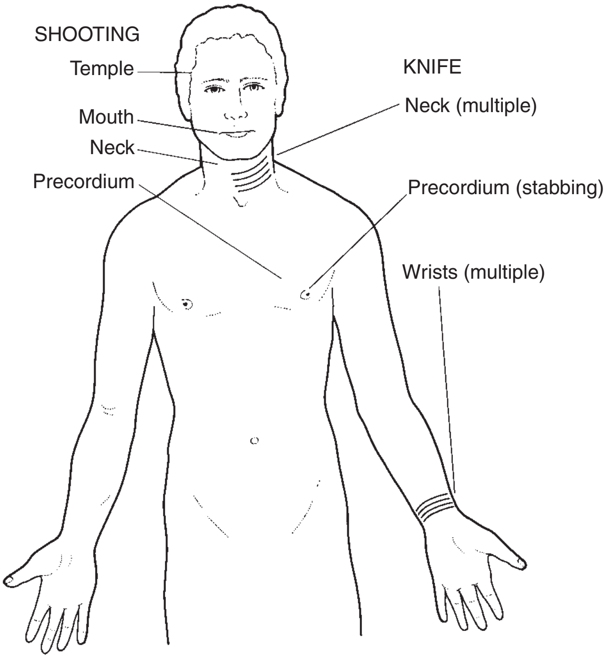

FIGURE 7.1 *Sites of election in violent suicide.*

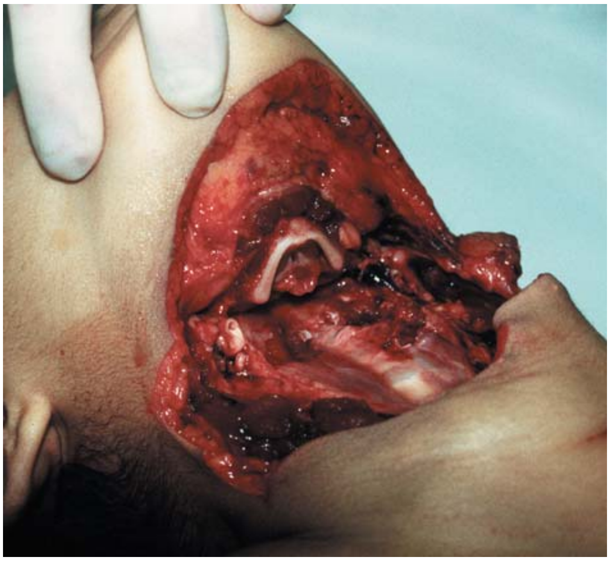

FIGURE 7.2 *Homicidal cut throat, showing single slash with no tentative cuts. Clean-cut edges with no sign of repetitive strokes, severing muscles, larynx and some large vessels.*

from the jugular veins, or less often the carotid arteries, may lead to death from exsanguination. If the larynx or trachea is opened, then even relatively minor haemorrhage from local vessels may cause blockage of the airways by blood and clot, though many slashed air-passage victims survive. A rare cause of death is air embolism, caused by aspiration into cut jugular veins while standing or sitting with the neck at a higher level than the thorax. The possibility that the victim of a cut throat may have died from some unrelated cause must always be borne in mind. Many suicides either use multiple methods of suicide to ensure success – or may abandon the pain of a cut throat for some other mode, such as poison or a fall from a height.

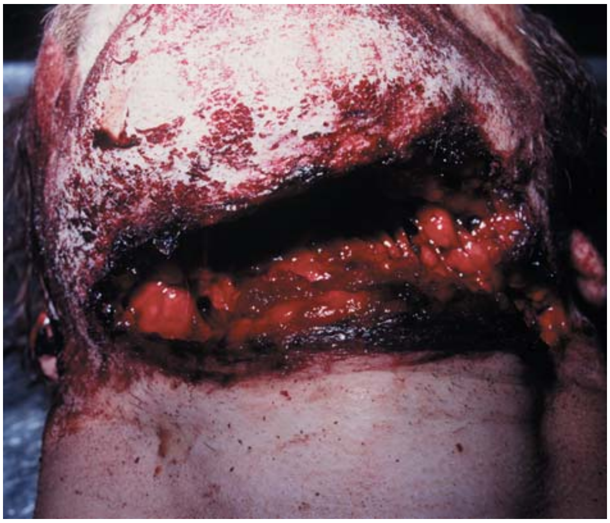

FIGURE 7.3 *Gaping laceration in front of neck of a mental patient, who on leave from a psychiatric hospital, had used a frame saw to commit suicide (see also Figure 7.4).*

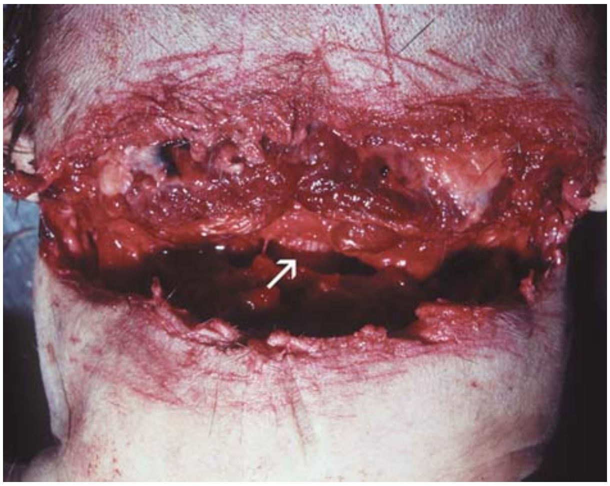

FIGURE 7.4 *Back of neck showing marks of the saw teeth around the laceration and medulla at the bottom (arrow). Mental patient who, on leave from a psychiatric hospital, had used a frame saw to commit suicide (see also Figure 7.3).*

- Deliberate cutting of the wrists is rarely effective as the sole method of suicide, but it is a common injury. Many suicides from other causes are seen at autopsy to have scars on the wrists from previous unsuccessful attempts or gestures. The usual site is on the flexor surface at the level of the skin flexion creases. As with the neck, there may be a number of shallow tentative incisions, but commonly several deep gashes are made without trial cuts. The left wrist is the more common target, because of right-handed dominance (86 per cent in men, 93 per cent in women).

As with the neck, there is a tendency for the victim to hyperextend the wrist before making the cut, which causes the radial artery to slip into the shelter of the lower end of the radius. The knife cuts may then miss major blood vessels and merely divide flexor tendons, though many cuts are too shallow to cause anything

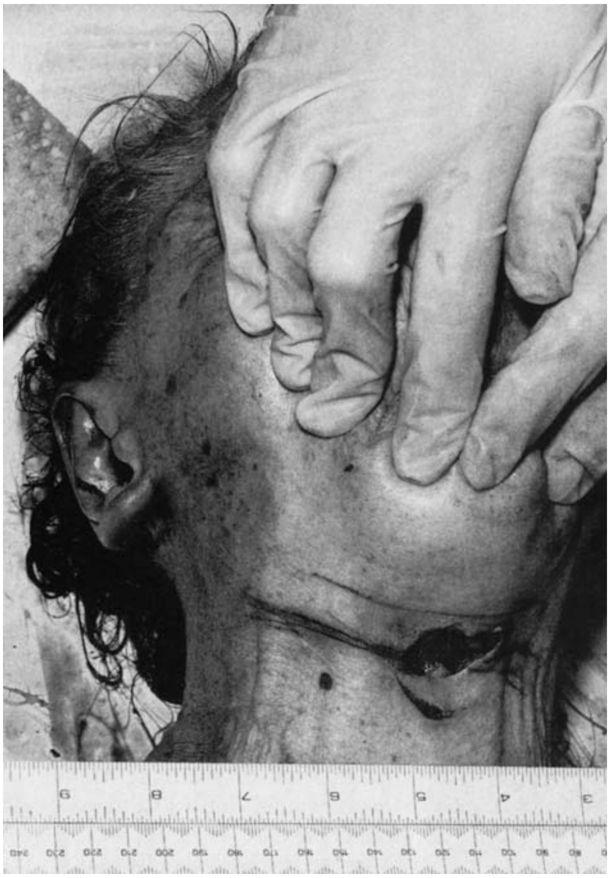

FIGURE 7.5 *Homicidal cut throat, simulating a suicide. There are several apparently 'tentative cuts', but the deep penetrating central laceration and adjacent stab wound is atypical. The face is congested and shows some petechiae from prior pressure on the neck.*

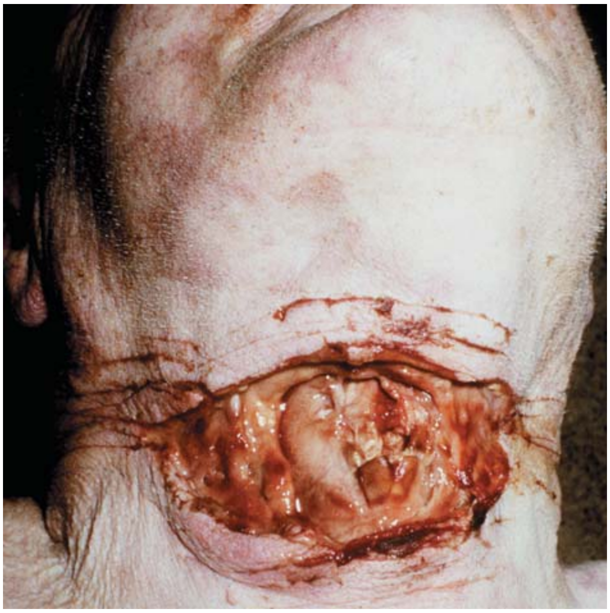

FIGURE 7.6 *Suicidal cut throat. The wounds are horizontal, rather than the more common sloping-down from left to right in a right-handed person. The numerous 'tentative cuts' are the hallmark of self-infliction, though there are rare exceptions. The larynx has been entered and death was due to aspiration of blood into the air passages. The great vessels in the neck were undamaged.*

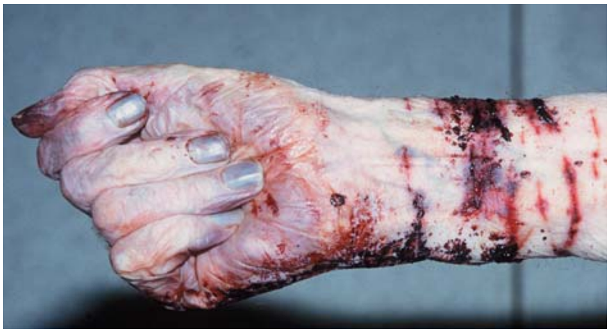

FIGURE 7.7 *Numerous self-inflicted cuts on the wrist, as part of a futile suicide attempt. Most of these attempts fail and the person turns to another method of self-destruction.*

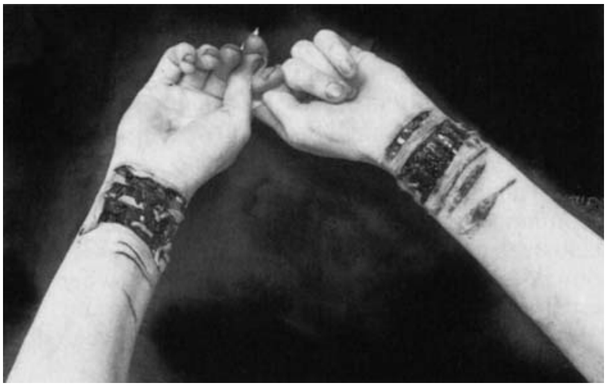

FIGURE 7.8 *Knife wounds of wrists in a suicidal attempt. Even though these are more severe than usually seen, no damage to the radial arteries was effected and death was the result of a subsequent cut throat. The victim was a woman of 31, unusual in that cut wrists and throat are much more common among men.*

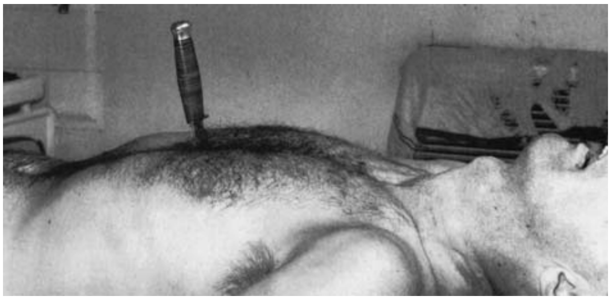

FIGURE 7.9 *Suicidal stab wound of heart. The front of the chest is one of the suicidal sites of predilection; multiple stabs do not exclude self-infliction.*

but superficial damage. Most cut wrists are but an accompaniment to other more effective methods of self-destruction. Once again, the presence of tentative cuts, though very significant, cannot be unquestioningly accepted as an exclusion of homicide.

- Suicidal injuries of the chest are almost always stab wounds. Sometimes linear incised cuts are made over the precordium or more widely over the front of the chest; these may be multiple, parallel and sometimes criss-cross, but rarely do any serious damage. Stab wounds are usually over the left side of the chest where common knowledge places the heart, but they may be alongside the sternum or even on the right side. The wound is often single, but multiple wounds are not uncommon. Though naturally a series of severe wounds always raises the presumption of homicide, many undoubted suicides can inflict a number of injuries upon themselves, each potentially fatal. The author (BK) has seen a suicide in a doctor, where eight severe knife wounds of the lower chest had created a large external defect and detached a portion of liver.

  Suicidal knife wounds are not often made in the abdomen, though they do occur, sometimes to the extent of disembowelling. The author (PS) has seen a suicide of a person, who removed several meters of the gut through a single and relatively small incision before dying of exsanguination. There is usually frank mental abnormality present in these cases, which merge into the self-mutilation described below.
- Stab wounds of the neck are uncommon in suicide, but are recorded (such as those described by Gee and Watson 1989, where actual transfixion of the neck was achieved).

- The position of the clothing in suicidal knife wounds is equivocal. It is more common for the self-stabber to lift the garments to expose the area of chest or abdominal skin to be attacked, but exceptions are common and the point has little diagnostic significance.

Suicidal firearm wounds

The characteristics of gunshot wounds are fully discussed in Chapter 8 but, as far as suicide is concerned, the following salient features may be repeated:

- Women rarely shoot themselves. In Britain, a shot woman is a murdered woman until proved otherwise, though – as always – exceptions do occur, especially in communities where firearms are commonplace, such as farming or hunting areas.
- The weapon must always be present.
- The range must be within arm's length, depending upon the nature of the weapon and excluding some mechanical device to fire the gun. The pathologist may be asked if a dead person could have reached the trigger of a rifle or long shotgun, which may have been placed against the neck, head or chest. Having first established that the entry wound was close contact, the length of the gun from muzzle to trigger can be measured and compared with the distance from wound to tip of finger. It must be remembered, however, that the length of the limb after death, especially if rigor is present, is not necessarily the same as in life when mobile joints and muscles may have allowed a slightly longer reach than that measured after death. Allowance must also be made for lateral or other flexion of the trunk, and mobility of the shoulder girdle before declaring dogmatically that the victim could not have fired the gun and therefore did not shoot himself.
- Sites of election are the temple, the neck, the mouth and the chest. Suicides rarely shoot themselves in the eye or abdomen. The lifting of clothing to expose bare skin, usually in the chest, is very variable, and has even less significance in differentiating suicide from homicide or accident than in knife wounds.
- Shooting in anatomically inaccessible sites cannot be suicide.
- It is not true that suicides always shoot themselves in the head on the same side as their dominant hand.
- It is difficult to differentiate homicide from accident, even though the circumstances may be suggestive. The distance of discharge may be beyond arm's reach in both, the weapon may be missing in both, inaccessible sites may be shot in both and sites of suicidal election may not be used in both.

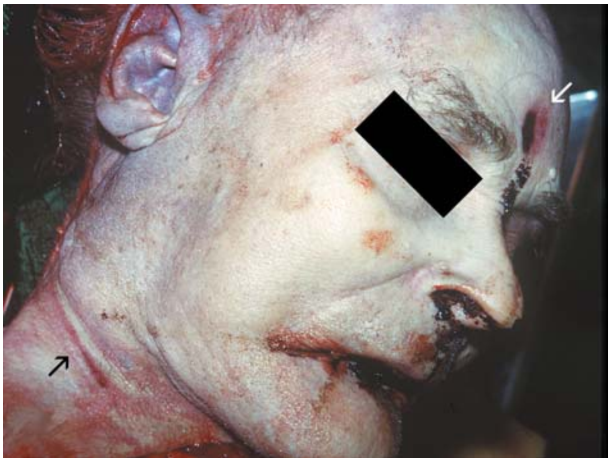

FIGURE 7.10 *A combination of two suicide methods, by firearm and hanging. The man had stood on a stool with a noose around his neck and then shot himself through the forehead. The entrance wound is seen in the forehead (white arrow) and the double ligature mark around the neck shows ecchymoses on the ridge between the two furrows (black arrow).*

Suicide by submersion

This is a common mode of self-destruction, in both men and women. It has a marked geographical variation as a favoured method of self-destruction, naturally depending on the availability of water – it is relatively infrequent in Egypt, but very common in the west of Ireland.

Either large bodies of water, such as rivers, lakes or seas may be used – or the bathtub. The pathology of immersion deaths are dealt with in Chapter 16, but the following features will assist identifying suicides:

- Other means of self-destruction may also be present, such as drug overdose: in baths, associated electrocution is not uncommon, especially in the former West Germany (Bonte 1983).
- Suicide notes may or may not be present, but those who kill themselves usually remove top clothes, hats and spectacles that may be left on the river bank; as usual, exceptions are commonplace.
- Other injuries on the body may be the result of agonal or post-mortem damage in the water.

Suicide by burning

Heat injuries are dealt with in Chapter 11, but suicide from burns has recently increased in Western countries (where it was formerly almost unknown), though in Asia it has always been relatively common. The spread to the West followed political self-martyrdom modelled on that practised by some Vietnamese protesters. There are no particular medical features to distinguish suicide from accident or homicide, this being a circumstantial matter. It is almost invariable, however, for some container of inflammable accelerant to be within reach. The myth of 'spontaneous combustion' is discussed in Chapter 11.

Suicide by 'asphyxia'

The pathology is described in Chapters 14 and 15, but the following facts are pertinent to suicide:

- The masochistic or sexual asphyxias must be carefully distinguished from homicide and suicide. Some asphyxial deaths with sexual aspects appear to be suicidal in nature – or at least, the inevitability of a fatal outcome must have been apparent to the victim (Knight 1979).
- Hanging, though not usually truly asphyxial, is a common mode of suicide, especially (though not exclusively) in men. It is also common as a sequel to homicide, especially in domestic murders, where the perpetrator often commits suicide by this means.
- Plastic bag suicide is increasing in incidence, the means being universally available.
- Suicidal ligature strangulation is not uncommon and sometimes causes difficulty in differentiation from murder. The winding of several turns of ligature and the tying of multiple knots is still consistent with suicide, even though it is hard to understand how, in some cases, death appears to have been caused by reflex cardiac arrest rather than the more florid 'asphyxial' mode of death.

  Sometimes a stick or other object is inserted under a ligature and twisted to tighten the cord against the neck. This is called a 'Spanish windlass', the stick becoming jammed against the body to prevent the tightened ligature from becoming loose once the twisting is completed. Though usually suicidal, this device is sometimes used for murder.

Jumping from a height

'Precipitation' – jumping from a height – is another mode of suicide where the circumstances rather than the autopsy findings determine the motivation. It may sometimes be possible to deduce, by measuring the distance from the jumping point to where the body strikes the ground, the fact that the victim must have actively projected himself outwards rather than fallen close to the wall or cliff face. Some experiments on this aspect have been carried out (Goonetilleke 1980). The injuries may be variable, depending on the attitude of the body when striking the ground, contact with obstacles during the fall and the height of the drop.

Head injuries are common and are described in Chapter 5. If the fall is onto the feet, damage may occur at any point from the ankles to the skull. The legs may fracture, the hip joints may impact, the pelvis may split, the spine may be fractured (especially vertebral body compression, hyperflexion or hyperextension lesions), or the base of the skull may be imploded as a ring fracture.

It is extremely difficult, to the point of impossibility, to deduce the height of the fall from the nature and severity of the injuries, as was evident from the medical testimony in the Helen Smith inquest in Leeds in 1982.

Some persons will be killed from a fall from a standing height, whilst other may make remarkable escapes without injury from a very high fall. The rare survivals in falls from a great height – usually when some cushioning effect, such as snow, exists – may still defy mechanical explanation.

SELF-INFLICTED INJURIES OTHER THAN SUICIDE

Two main classes exist: those where some abnormality of mind leads the victim to mutilate his body – and those in which deliberate injuries are inflicted for motives of gain. Those caused by mental aberration are usually bizarre in either their multiplicity or their site.

Some mentally disordered persons may inflict hundreds of small wounds upon themselves, which may be additional to the actual cause of death, if death has taken place. Forensic physicians, such as police surgeons, may be called upon to examine living victims of this type.

The arms may be covered with scores of parallel superficial incisions from a knife, razor or broken glass. These may cross each other in groups and again are more frequent on the non-dominant side, usually the left. Small stabs may be made on the chest or temples. In a case seen by the author (BK), there were more than 200 knife-point stabs on each temple and the centre of the forehead. Another well-recognized syndrome is self-mutilation of the genitals, almost invariably by men. Paranoid schizophrenics, often with a strong religious flavour to their delusions, are known to attack themselves in this manner. The penis, scrotum and testes may be removed, and death from intractable haemorrhage can follow. In one of the cases illustrated, there was also perforation of the eyeballs and 40 stab wounds on the top of the head, as well as a complete circumferential cutting of the throat, all with the same blunt scissors that performed the castration. The forensic implications arose from

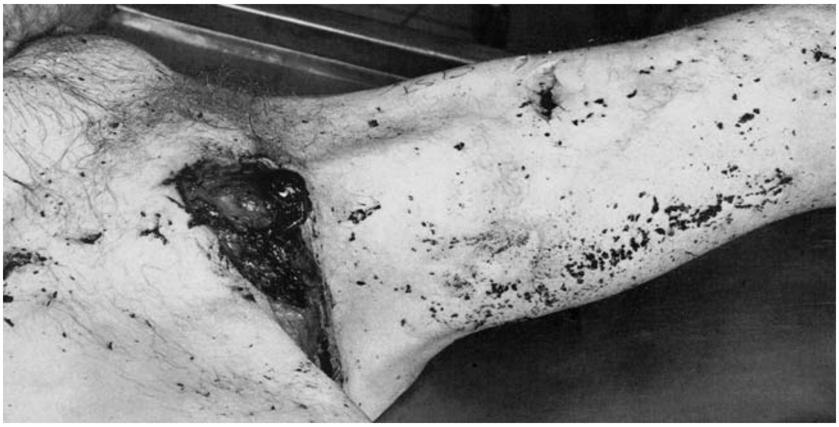

FIGURE 7.11 *Complete self-amputation of the scrotum and penis with scissors.*

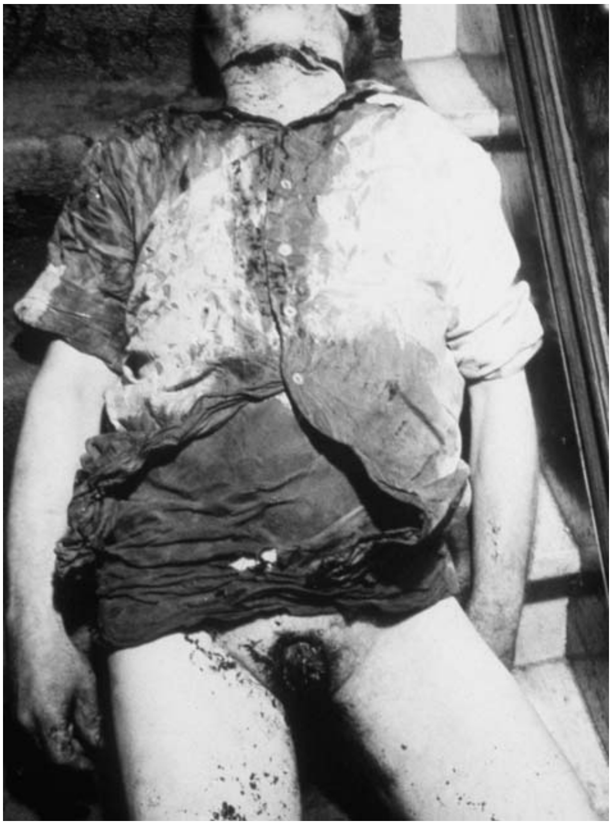

FIGURE 7.12 *Self-mutilation thought by the police to be a bizarre homicide. The genitals have been amputated, the throat cut through 360°, the eyeballs punctured and the scalp stabbed many times, all with scissors. The dead man had paranoid delusions of a religious nature.*

the initial suspicions of the police that a sadistic homicide had taken place.

Self-inflicted injuries, usually non-fatal unless some complication has arisen, may be motivated by some form of gain. The most common is fabrication of injuries to simulate an assault, either to divert attention from the person's own theft or to arouse sympathy. Fabricated accidental injuries are usually a form of malingering, as in the armed forces, or in fraudulent attempts to obtain compensation. In other cases, the 'victim' may claim that a specific person assaulted them, from motives of hate or spite. Others may injure themselves and then waste police time by insisting on a futile search for a non-existent assailant, the motive often being impossible to discover, even by a psychiatrist.

Though these matters are usually seen by the clinical forensic physician, the pathologist may be asked for an opinion where no other medico-legal doctor is available, or if death supervenes. The so-called 'Münchausen syndrome' consists of repeated simulations of illness or the infliction of repeated minor injuries, with the object of gaining admission to hospital, or obtaining medical care and attention.

The following features assist in the recognition of self-inflicted incised injuries:

- The cuts are usually superficial and rarely any danger to life, unless they become infected. Occasionally, they may penetrate the full thickness of the skin, but not in sensitive areas like the face.
- The incisions are regular, with an equal depth at origin and termination, unlike more serious wounds that tend to be deeper at the start and tail off to the surface.
- The cuts are usually multiple and often parallel. They avoid vital and sensitive areas like the eyes, lips, nose and ears, usually being drawn on the cheeks and jaws, temples and forehead, sides of the neck, chest, shoulders, arms and backs of hands and thighs. This is inconsistent with an attack by another person, as the victim is unlikely to stand still to allow these multiple delicate and uniform injuries to be carefully executed.
- In right-handed persons, most of the injuries are on the left side, especially the sides of the face and left hand.
- When the incisions are made on areas covered by clothing, the relevant garments may show either no cuts or cuts that do not match the injuries in position or direction.

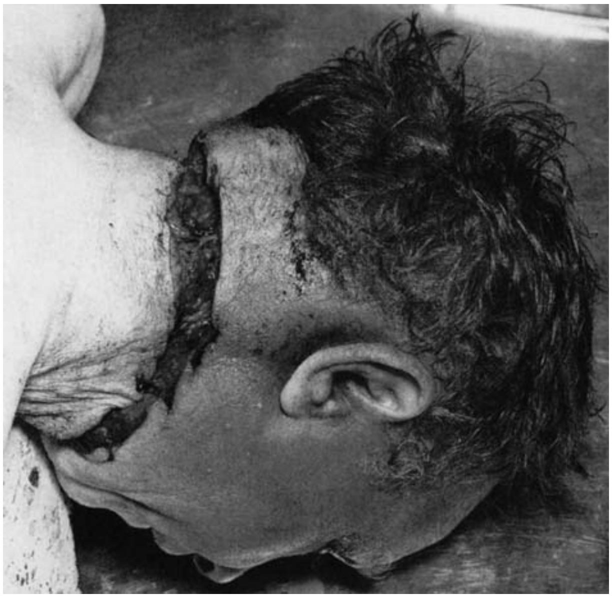

FIGURE 7.13 *Self-inflicted scissors injury around the full circumference of the neck.*

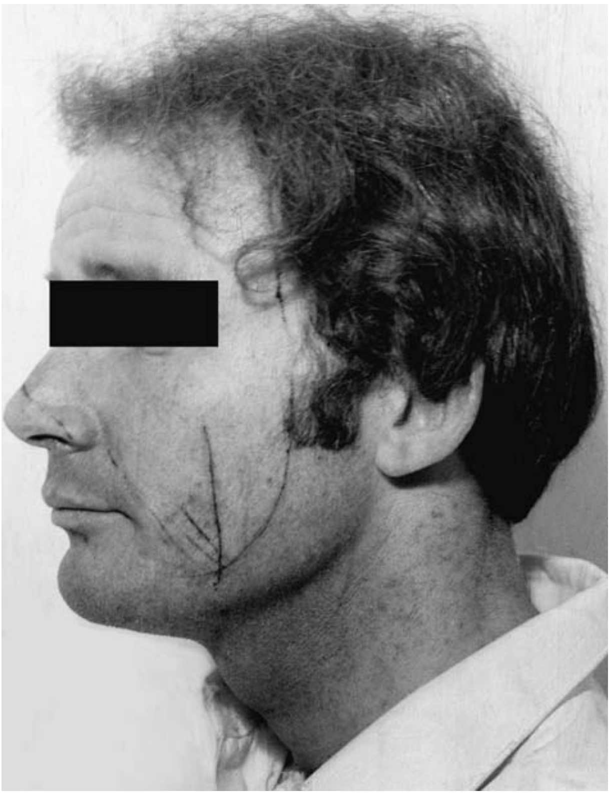

FIGURE 7.14 *Self-inflicted injury in an attempt to produce an excuse during shop-breaking. Surprised by the police, he cut himself with broken glass to claim that he had been assaulted by the robbers whilst trying to defend the premises. There were similar shallow, straight cuts on the other side of the face and on the backs of the hands.*

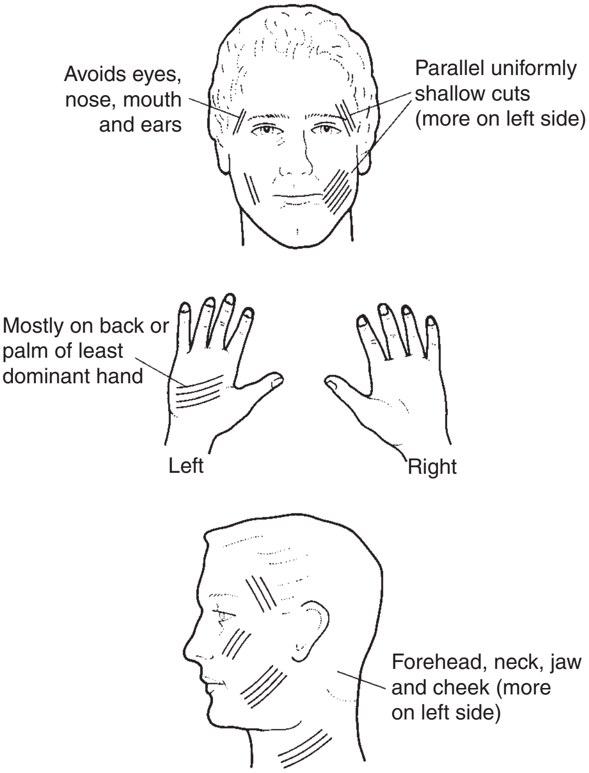
Avoids eyes, nose, mouth and ears
Parallel uniformly shallow cuts (more on left side)
Mostly on back or palm of least dominant hand
Left
Right
Forehead, neck, jaw and cheek (more on left side)

FIGURE 7.15 *Characteristics of non-fatal self-inflicted injuries.*

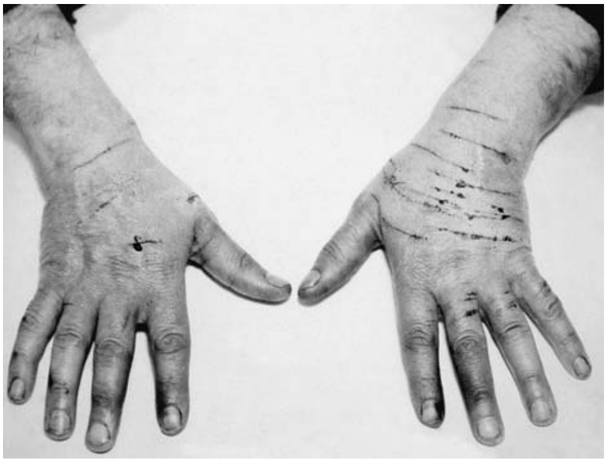

FIGURE 7.16 *Cuts on the backs of the hands in the man shown in Figure 7.14. He was right handed and most of the shallow incisions are on the left hand.*

In summary, the whole range of injuries and surrounding circumstances may suggest to the experienced pathologist that the lesions do not constitute a genuine assault. This diagnosis may be virtually immediate on first looking at the 'victim' and may be almost intuitive after some experience – though this is of little use when giving legal testimony, when every opinion has to be backed up by hard fact and reasoned opinion.

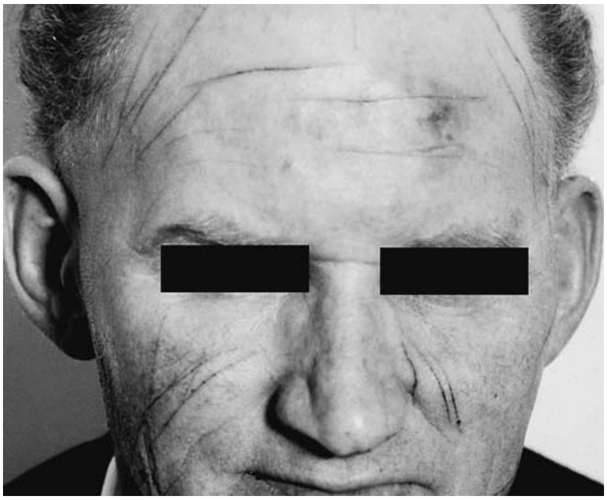

FIGURE 7.17 *Self-inflicted injury in the living. The 'victim' alleged that he had been attacked and robbed. The superficial, uniform, mostly parallel scratches avoid the eyes, nose and mouth.*

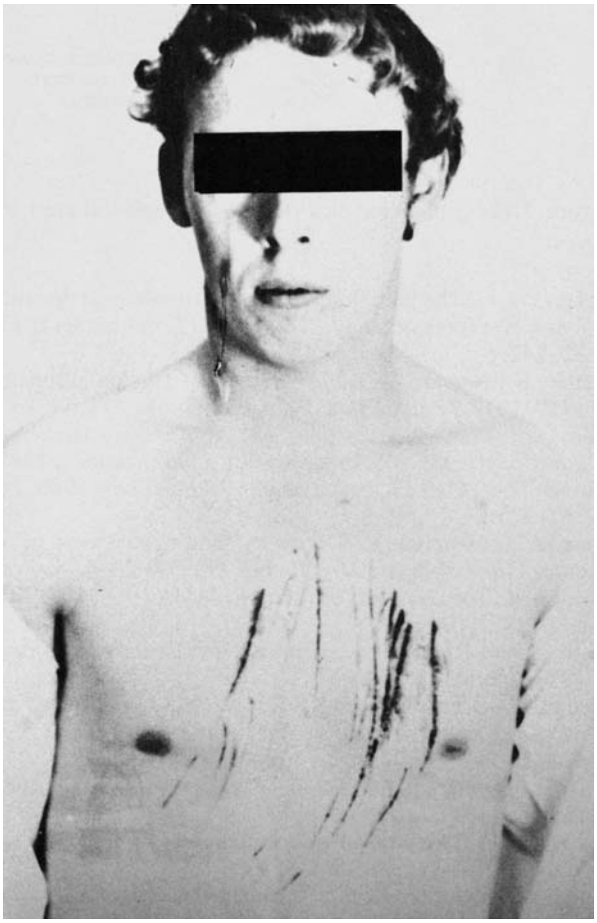

FIGURE 7.18 *Extensive self-inflicted cuts on the chest. The youth persistently claimed to the police that he had been attacked, but the cuts in his shirt were horizontal. He later confessed, explaining that he wanted to get his father's sympathy over some domestic crisis.*

The so-called 'Münchausen syndrome' and the 'Münchausen syndrome by proxy' lie within the realm of the forensic psychiatrist, but a pathologist or clinical forensic physician may be called upon to confirm – or even first detect – that the injuries are self-inflicted.

The Münchausen syndrome refers to adults who injure themselves or feign signs of illness, in order to gain medical attention and often admission to hospital. They can be a plague to casualty departments until recognized. More sinister is the Münchausen syndrome by proxy, where a parent will repetitively injure her child – occasionally fatally – to gain the same attention from doctors and medical facilities. The relationship with child abuse, of which it is a variety, must be recognized.

REFERENCES AND FURTHER READING

Anderson WR, Hudson RP. 1976. Self-inflicted bite marks in battered child syndrome. *Forensic Sci* **7**:71–4.

Asher R. 1951. Munchausen's syndrome. *Lancet* **1**:339–40.

Bonte W. 1978. [Ritual finger amputations. A contribution on the diagnosis of intentional self mutilation using an axe]. *Arch Kriminol* **162**:17–22.

Bonte W. 1983. Self-mutilation and private accident insurance. *J Forensic Sci* **28**:70–82.

Bonte W, Rudell R. 1978. [Accident or planned self mutilation? Probability of accidental injuries during chopping]. *Arch Kriminol* **161**:143–52.

Carney MW. 1978. Self-inflicted bleeding and bruising. *Lancet* **1**:924–5.

Dotzauer G, Iffland R. 1976. Self-mutilations in private-accident-insurance cases. *Z Rechtsmed* **77**:237–88.

Eckert WG. 1977. The pathology of self-mutilation and destructive acts: a forensic study and review. *J Forensic Sci* **22**:242–50.

Faller-Marquardt M, Ropohl D, Pollak S. 1995. Excoriations and contusions of the skin as artefacts in fictitious sexual offences. *J Clin Forensic Med* **2**:129–35.

Fanton L, Schoendorff P, Achache P, *et al.* 1999. False rape: a case report. *Am J Forensic Med Pathol* **20**:374–7.

Gaillard Y, Pepin G. 1998. Case report of an unusual use of lidocaine during episodes of self mutilation. *J Forensic Sci* **43**:235–8.

Gatter K, Bowen DA. 1980. A study of suicide autopsies 1957–1977. *Med Sci Law* **20**:37–42.

Gee DJ, Watson A. 1989. *Lecture notes in forensic medicine*, 5th edn. Blackwell Scientific Publications, Oxford.

Gibbon KL. 1998. Munchausen's syndrome presenting as an acute sexual assault. *Med Sci Law* **38**:202–5.

Goonetilleke UK. 1980. Injuries caused by falls from heights. *Med Sci Law* **20**:262–75.

Greenspan GS, Samuel SE. 1989. Self-cutting after rape. *Am J Psychiatry* **146**:789–90.

Guarner J, Hanzlick R. 1987. Suicide by hanging. A review of 56 cases. *Am J Forensic Med Pathol* **8**:23–6.

Halprin KM. 1966. The art of self-mutilation. I. Neurotic excoriations. *JAMA* **198**:777.

Halprin KM. 1966. The art of self-mutilation. II. Delusions of parasitosis. *JAMA* **198**:1207.

Harris CN, Rai K. 1976. The self-inflicted wrist slash. *J Trauma* **16**:743–5.

Jordan FB, Schmeckpeper K, Strope M. 1987. Jail suicides by hanging. An epidemiological review and recommendations for prevention. *Am J Forensic Med Pathol* **8**:27–31.

Karger B, DuChesne A, Ortmann C, *et al.* 1997. Unusual self-inflicted injuries simulating a criminal offence. *Int J Legal Med* **110**:267–72.

Kleiner GJ. 1984. *Suicide in pregnancy*. John Wright, Bristol.

Knight B. 1979. Fatal masochism – accident or suicide? *Med Sci Law* **19**:118–20.

Knight BH. 1968. Two cases of self-inflicted injuries in the living. *Med Sci Law* **8**:264–6.

Kohr RM. 1990. Suicide by chloroform ingestion following self-mutilation. *Am J Forensic Med Pathol* **11**:324–8.

König H. 1996. [Discrimination of sharp injuries produced in order to feign assaults and such of real cases]. In: Saternus KS, Kernbach-Wighton G (eds), *Research in Legal Medicine*. Schmidt-Römhild, Lübeck.

Koops E, Puschel K. 1990. [Self-mutilation and autophagia.] *Arch Kriminol* **186**:29–36.

Leslie J, Taff ML, Patel I, *et al.* 1984. Self-inflicted ocular injuries. A rare form of self-mutilation. *Am J Forensic Med Pathol* **5**:83–8.

Peschel O, Betz P, Eisenmenger W. 1997. Self-mutilation with needles. *Med Sci Law* **37**:175–8.

Pollak S, Saukko P. 2000. Clinical forensic medicine. Self-inflicted injury. In: Siegel JA, Saukko PJ, Knupfer GC (eds), *Encyclopedia of Forensic Sciences*. Academic Press, San Diego, San Fransisco, New York, Boston, London, Sydney, Tokyo.

Püschel K, Hildebrand E, Hitzer K, *et al.* 1998. [Self-mutilating hand and finger injuries among physicians suspected of insurance fraud]. *Versicherungsmedizin* **50**:232–40.

Riepert T, Schneider P, Urban R. 1999. [Self-injury or attempted murder – analysis of an unusual case.] *Arch Kriminol* **203**:19–26.

Rothschild MA, Raatschen HJ, Schneider V. 2001. Suicide by self-immolation in Berlin from 1990 to 2000. *Forensic Sci Int* **124**:163–6.

Shiono H, Takaesu Y. 1986. Suicide by self-inflicted stab wound of the chest. *Am J Forensic Med Pathol* **7**:72–3.

Sneddon I, Sneddon J. 1975. Self-inflicted injury: a follow-up study of 43 patients. *Br Med J* **2**:527–30.

Tsunenari S, Idaka T, Kanda M, *et al.* 1981. Self-mutilation. Plastic spherules in penile skin in Yakuza, Japan's racketeers. *Am J Forensic Med Pathol* **2**:203–7.

Zlotnick C. 1999. Antisocial personality disorder, affect dysregulation and childhood abuse among incarcerated women. *J Pers Disord* **13**:90–5.

CHAPTER **8**

Gunshot and explosion deaths

- The mechanics of missile injury 245
- Types of weapon 246
- Wounds inflicted by a smooth-bore shotgun 249
- Wounds from rifled weapons 258
- Exit wounds 267
- Firearm damage to internal organs 270
- Estimating the range of discharge 270
- Airgun injuries 271
- Humane (veterinary) killers and industrial 'stud-guns' 272
- Rubber and plastic bullets 274
- Deaths from explosion 274
- References and further reading 277

Though the majority of missile wounds are caused by firearms, other devices such as crossbows, captive-bolt guns, air weapons and even catapults can launch lethal projectiles. In the bombing deaths now so commonly associated with terrorism, missile fragments cause more deaths than the blast effects, so, overall, an understanding of projectile trauma is essential.

THE MECHANICS OF MISSILE INJURY

With the exception of deceleration injuries, all mechanical trauma, whether punching, stabbing or kicking, is caused by the transfer of energy from an external moving object to the tissues and nowhere is this more obvious than in shooting. For damage to occur, some or all of the kinetic energy of the missile has to be absorbed by the target tissues, where it is dissipated as heat, noise and mechanical disruption. When a missile passes completely through soft tissues, it may retain much of its original kinetic energy and fail to transfer any appreciable amount to the tissues, which may remain relatively intact apart from the immediate bullet track. If the latter is in a limb muscle, there may be no serious effects if major blood vessels are not involved, though the same track in brain, lung or heart may prove fatal.

To ensure transfer of energy to the tissues, some missiles are especially designed or modified to slow up or stop within the body. Soft-headed bullets will flatten on impact and some are designed to fragment. The 'dumdum' bullet, which has a scored nose, and the military missile with an air-cavity within the tip are intended to splay open on impact to increase the 'braking' or deceleration effect, and transfer more energy for disruption. Explosive-tipped bullets, such as those used in the assassination attempt on President Reagan, are not designed to cause damage by the tiny detonation, but drastically to deform the missile to cause maximum deceleration. Weapons designed to be fired in confined spaces, such as those for combating hijack attempts in an aircraft, may be of relatively low velocity and have a deceleration feature incorporated to ensure the absence of an exit wound, and thus limited travel of the missile to avoid puncture of the pressurized passenger cabin.

The trajectory of the missile also determines how much and how fast its energy is given up to the target. Shotgun pellets are spherical, so the orientation of impact is not relevant, but all conical bullets may acquire an erratic course in the tissues. They may tumble end-over-end, especially when nearing the limit of range; they may 'wag' or 'yaw' from side to side of their axial trajectory; the base may rotate around the axis, with the tip remaining on the straight path; and they may ‘precess’ or ‘nutate’ with complex spiral or circular movements about the axis.

Whatever the deviation, it offers more contact between the projectile and the tissues, allowing more transfer of energy and thus greater tissue damage. The amount of kinetic energy possessed by a projectile accords to the familiar formula of half the product of the missile mass and the square of its velocity. Modern military science takes advantage of the squaring of the velocity to develop weapons that have a missile of small mass but exceedingly high velocity to provide the maximum kinetic energy for tissue damage.

The mode of injury depends on the velocity of the missile. Relatively slow projectiles are those travelling at up to the speed of sound in air (340 metres/second or 1100 feet/second), which of course includes all non-explosively propelled missiles like crossbow bolts and air-rifle pellets, as well as most revolver bullets. These mechanically thrust aside the tissues along a track only slightly wider than the missile. The tissues are lacerated or crushed, secondary damage occurs from rupture of blood vessels and other structures, and secondary and tertiary damage is caused by displaced bone and cartilage fragments.

Above the speed of sound in air, a missile passing through tissue sends a shock wave of compression ahead of the laceration track, this wave being propagated at about the speed of sound in water (1500 metres/second or 4800 feet/second). Though this wave lasts only for a brief period, it raises the tissue pressure to extreme values, up to thousands of kilopascals. In tissues like brain, liver and muscle, this can cause severe disruption within a wide zone around the bullet track, and can be propagated down hollow fluid-containing vessels to cause distant vascular damage.

High-velocity projectiles produce yet another phenomenon, that of cavitation. The missile accelerates the molecules of the tissues adjacent to the track, so that they continue to move centrifugally outwards even after the missile has travelled onwards. This forms a cavity around the track that is far wider than the diameter of the projectile. It reaches a maximum size within milliseconds and then pulsates with decreasing amplitude so that a fusiform cavity rapidly follows in the wake of the bullet. When the bullet stops or leaves the organ, the cavity rapidly subsides, but the track of damage persists in a tubular zone much wider than the actual missile. In high-velocity bullets from military weapons – for example, 980 metres/second (3185 feet/second) this cavitation effect is the most damaging mechanism; in addition to the physical damage, the near-vacuum sucks infected dirt and clothing fibres into the depths of the wound (Owen-Smith 1981). Solid organs such as brain and liver are affected more than spongy matrices like lung. Those tissues that contain most water are most severely damaged by penetrating missiles and high-velocity missile damage is proportional to the specific gravity of the tissue injured.

In most of the shooting cases seen by forensic pathologists, death will have occurred rapidly but, where it is delayed, secondary damage from infarction, local necrosis of muscle and organs, and infection must always be borne in mind. High-velocity weapons in particular can cause vascular damage at a distance from direct trauma, stretching and thrombosis, leading in turn to ischaemic lesions such as infarcts.

TYPES OF WEAPON

The nature of firearm wounds varies considerably with the type of weapon employed. A general knowledge of the main features of the various types of gun in common use is essential, though there is no need for the pathologist to master the unnecessary mass of detail on firearm construction offered by some forensic medicine textbooks. Indeed, much of this is solely the province of the firearms examiner, not the doctor. The inordinate amount of detail sometimes provided is perhaps symptomatic of the fascination which firearms hold for many men – but rarely women.

In the context of wounding, guns are of two main types.

The smooth-bore weapon or ‘shotgun’

A shotgun consists of one or more metal barrels of relatively wide diameter, which are smooth on the inner surface. They fire a variable number of spherical lead shot (pellets), which emerge from the end (muzzle), from where they gradually diverge in the form of a long, narrow cone. Exceptionally, a shotgun may fire a few large projectiles or even a single slug, but these are rarely met with in forensic practice, the usual load of pellets totalling scores or hundreds. Special types of projectile are described later. A shotgun often has two barrels, either side by side or ‘up-and-under’. One barrel may be parallel sided (the ‘cylinder’) and the other slightly tapered towards the muzzle (the ‘choke’). The latter will produce a narrower cone of shot and this affects estimates of range derived from the size of the wound. There are several gradations of choke, such as ‘improved cylinder’, ‘modified choke’, ‘half-choke’ and ‘full-choke’.

Some modern weapons have the facility for removing one particular type of barrel and replacing it with another degree of choke. Modern accessories can convert a cylinder barrel to a choke barrel – for example, variable choke such as Poly-Choke, Kutts Compensator and Weaver-Choke are available, especially in the USA.

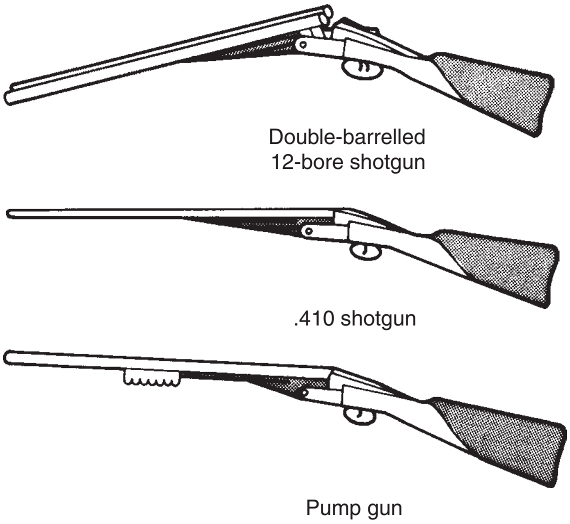

FIGURE 8.1 *Types of smooth-bore weapon.*

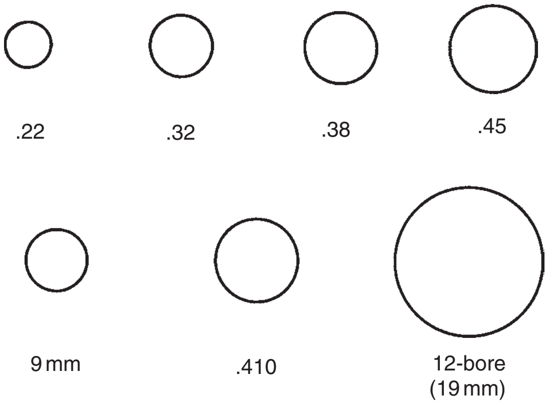

FIGURE 8.2 *Relative calibre of common weapons.*

Another typically American modification is the ability to fire a single projectile through the smooth bore. A number of devices intended for hunting are available, but they may be involved in a human fatality. One of the older types is the solid round 'pumpkin ball' or the Brenneke rifled slug, but more recent devices include the Foster rifled slug and the French Blondeau, which is a dumb-bell-shaped projectile.

Shotguns come in two main sizes, which determine the nature of the wound. The first is the 'twelve-bore', usually called 'twelve-gauge' in North America, which has a diameter of about 19 mm or 0.738 inches. The other is the smaller 'four-ten', usually single-barrelled with a diameter of 10.6 mm or 0.410 inches.

SHOTGUN AMMUNITION

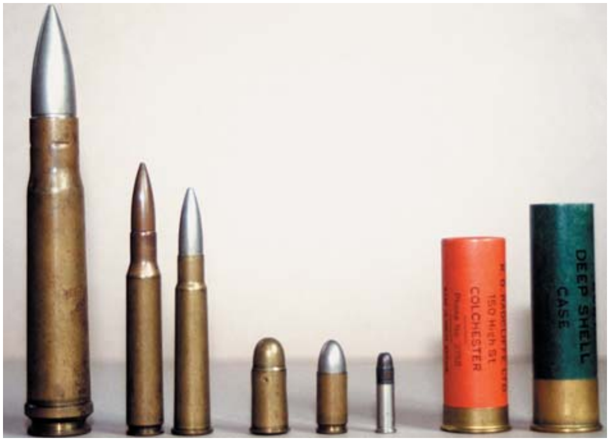

FIGURE 8.3 *Various types of ammunition. From the right, there are two twelve-bore shotgun cartridges, .22 rifle, .38 automatic pistol and .45 revolver bullets and two military rifle shells. On the left is a cannon shell, which is unlikely to be encountered in civilian forensic pathology.*

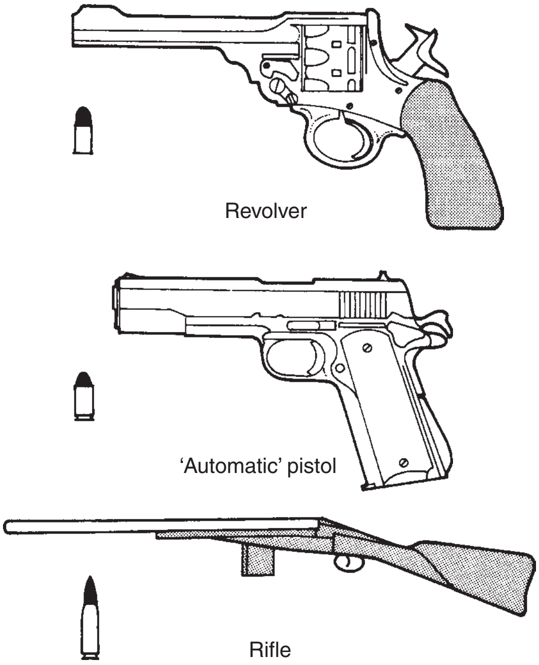

FIGURE 8.4 *Types of weapon and ammunition.*

The ammunition for the shotgun is a cartridge made of a cardboard or plastic cylinder fitted into a metal base. This carries a percussion detonator that is struck by the spring-loaded firing pin when the trigger is pulled. The cartridge contains a charge of propellant, above which are 'wads' or pistons of felt, and cardboard or plastic discs, or both. Above these is the charge of shot, which varies greatly in number and size, finally covered by a card or plastic disc.

This description is of the old, classical cartridge but modern ammunition has a variety of devices incorporated into the cartridge to improve efficiency and accuracy. These include the Power Piston, a proprietary name for one of a class of cartridges which holds the shot inside a polythene cup and which itself may contribute to the wound at short range. Other modern cartridges may have plastic granules as a filler between the shot; this brightly coloured material may be found inside the wounds.

Tampering with the contents of cartridges may markedly alter the spread and other characteristics of the shot and hence the resulting wound. The shot may be fused into a mass by pouring paraffin wax or even molten pitch into the cartridge. The cardboard or plastic cylinder may be partly cut through at the level of the wads, so that, on firing, the upper part of the casing is blown out to restrain the divergence of the pellets.

The propellant has also changed in recent years, so that the old black gunpowder has been largely replaced by cleaner explosives that do not produce the same volume of soot and unburnt propellant, which may be deposited in and around the close-range wound.

‘Country’ guns

In developing countries, in some rural and poor areas of other countries, in subversion and terrorism, and sometimes in juvenile hands, weapons and ammunition may be ‘home-made’. These are common in India, where the name ‘country gun’ is well understood; they are also made in Sri Lanka and other Asian areas, but may be encountered anywhere.

The author (BK) had experience of such weapons in the terrorist campaign in Malaya in the 1950s, where crude weapons were made in the jungle from pieces of water-pipe attached to roughly carved wooden stocks, with a primitive firing mechanism. The guns were muzzle-loaded with black powder and the projectiles were often wood-screws, nuts and bolts and irregular fragments of metal or even stones.

In civilian practice in Britain, two similar weapons have been seen, both made by juveniles out of electrical conduit pipe, using potassium chlorate–sugar propellant. In one, the user was killed when the device exploded; the other one successfully propelled a small screwdriver across several fields to (slightly) injure a cow.

Terrorists in Northern Ireland have also manufactured some of their own weapons, but these were relatively sophisticated, being made in clandestine but well-equipped workshops.

The problem for the pathologist with such guns is that few of the criteria to be described apply to the injuries, especially where random, irregular missiles such as wood-screws are employed. Local knowledge and experience are essential for the interpretation of these wounds.

The rifled weapon

Handguns, rifles, air rifles and military weapons differ from shotguns in that they fire one projectile at a time through a thicker barrel that has spiral grooves machined on its inner surface. The elevations between the grooves are the ‘lands’, which grip the bullet as it passes down the barrel and give it a rotatory movement. This has a gyroscopic effect that increases the stability of the bullet’s trajectory and hence the accuracy.

Handguns, which cause over 50 per cent of all homicides in the USA, comprise the revolver and the automatic pistol. The revolver has a rotating cylinder carrying a variable number of shells, which are brought into the firing position one by one, each time the trigger is pulled. The revolver has a low muzzle velocity of the order of 500 feet/second (150 metres/second). The ‘automatic’, more accurately called a ‘self-loading’ pistol, has its shells in a spring-loaded magazine, each spent round being ejected and a new one positioned by means of the gas pressure developed at each discharge. The muzzle velocity may be 1000–1200 feet/second (300–360 metres/second).

Rifles are long-barrelled guns that may be single-shot, bolt-loaded or self-loading and used for hunting, target-shooting or for military purposes. Military weapons come in a vast range of types, most of them now either ‘self-loading’, which means that the trigger has to be pulled for each shot – or truly ‘automatic’, when the weapon will continue to fire while the trigger is depressed until the magazine is empty. The muzzle velocity varies greatly, from 1500 to 5000 feet/second (450–1500 metres/second).

AMMUNITION FOR RIFLED WEAPONS

The variety of types of rifled ammunition is even greater than the range of weapons designed to fire them, but all conform to a general pattern. There is a metal cylinder, closed at one end, which is the ‘shell’ or cartridge, carrying a percussion detonator in the base, either centrally or peripherally. The shell is loaded with explosive propellant such as nitrocellulose and the bullet is firmly clamped into the open end. The bullet may be composed of a variety of metals, often compound. A lead core may be covered in a nickel or steel jacket, but there are many other variations. The detonator may contain elements such as barium, bismuth mercury or antimony.

Both shotgun and rifled weapon ammunition have a common purpose when detonated – to produce large volumes of hot gas under pressure that expel the bullet or shot from the barrel. One gram of black powder produces about 3000 ml and nitrocellulose produces some 13 000 ml of gas, which consists of carbon dioxide, monoxide, nitrogen, hydrogen sulphide, hydrogen, methane and many other substances, all at high temperatures.

WOUNDS

Though the construction and performance of weapons is of vital interest to the forensic firearms examiner, the relevance to the pathologist is concentrated in those aspects which affect the nature of the wound:

- whether the weapon is smooth-bore or rifled;
- if rifled, the muzzle velocity of the weapon;
- the nature of the projectile(s);
- the nature of the propellant;
- the degree of choke, if any;
- the range of discharge;
- the angle of discharge.

WOUNDS INFLICTED BY A SMOOTH-BORE SHOTGUN

The following constituents of the cartridge emerge on the discharge of a shotgun and all may contribute to the wound:

- lead pellets
- soot in the form of smoke and debris
- unburnt and burning propellant particles
- flame and hot gases under pressure
- carbon monoxide
- wads – either felt, cardboard or plastic
- detonator constituents
- fragments of the cartridge case.

When a shotgun is fired, a compact mass of shot emerges from the muzzle and then begins to disperse, the divergence increasing progressively as the distance lengthens. A tongue of flame and hot gas follows the shot. High pressure and temperature exists just outside the muzzle, but this rapidly expands and cools. This gas is composed of oxides of nitrogen, carbon dioxide, hot air and other compounds, but the one of interest to the pathologist is carbon monoxide. Soot from the combustion of propellant is expelled, along with some flakes or grains of propellant that may be still burning. The wads are also expelled, their nature depending upon the type of cartridge. Some modern ammunition may contain other devices that may contribute to the wound.

Chemical traces of the elements in the detonator or percussion cap cannot be seen, but may be vital laboratory evidence, where barium, antimony and other metals can be recovered on analysis or by means of the scanning electron microscope. Fragments of the cartridge case may also be ejected, especially if it has been tampered with, as described earlier.

The most important deduction that a pathologist is asked to make is the estimation of the range of discharge of the gun, as this may substantially influence the distinction between accident, suicide and homicide. The appearance of the wound is greatly modified by variation in range and it is convenient to describe them in terms of increasing range of discharge.

Shotgun contact wound

SHAPE

Where the muzzle is placed tightly against the surface of the abdomen, thorax, limb or neck, the consequent wound will

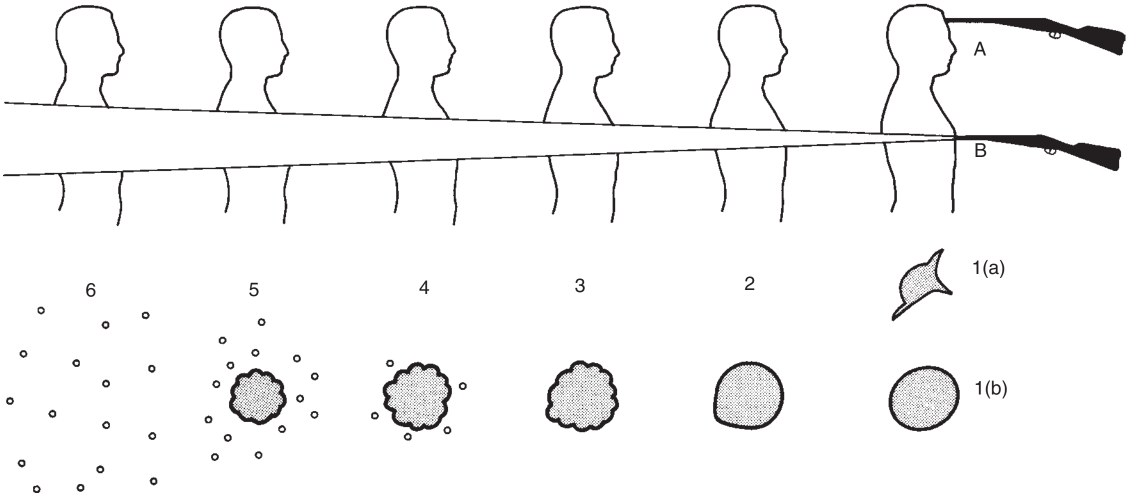

Figure 8.5 *Variation in appearance of a shotgun wound at increasing range of discharge: 1(a), split wound from contact over bone; 1(b), usual round contact wound; 2, close but not contact range up to about 30 cm (variable); 3, 'rat-hole' wound from 30 cm to about a metre (variable); 4, satellite pellet holes appearing over a metre; 5, spread of shot increases, central hole diminishes; 6, uniform spread with no central hole over about 10 m. All these ranges vary greatly with barrel choke, weapon and ammunition.*

be single and circular, of a size approximately equal to the bore of the weapon, though a .410 wound may be smaller because of the relatively greater effect of elastic contraction of the skin. The edges of the wound may be crenated by individual shot, but usually this feature is not noticeable.

SOILING AND BURNING

In a tight contact wound, the skin forms a seal around the muzzle, preventing much escape of hot gas and soot, so that soiling and burning are minimal or absent. The recoil, which takes the muzzle away from the skin, may, however, loosen the seal and, if the muzzle is not pressed firmly, flame, gas and soot may escape sideways and affect the skin in the immediate vicinity. Where clothing is interposed between the muzzle and skin, soot is much more likely to escape sideways and may be found in each layer of fabric, as well as on the underlying skin. The cloth may be singed at the edge of the hole and there may be a ring of burning around the skin wound.

MUZZLE IMPRESSION

There may be a muzzle impression in a tight contact wound made by firm mechanical pressure or impact of the metal rim against the skin. Many textbooks quite wrongly attribute this to ‘recoil’, though of course recoil takes the muzzle away from the skin, not towards it. The true explanation of a muzzle mark is either that the assailant physically kept the weapon pressed hard against the skin or – more often – that the subcutaneous expansion shortly to be described, lifts the skin forcibly up against the muzzle. A muzzle mark is a most useful indication to the pathologist of a contact wound. Actual bruising can occur around the muzzle imprint, though this is more likely to be the result of deep contusion from the effects of the blast. Rarely, a double-barrelled weapon may make a ring-like mark adjacent to the entry wound.

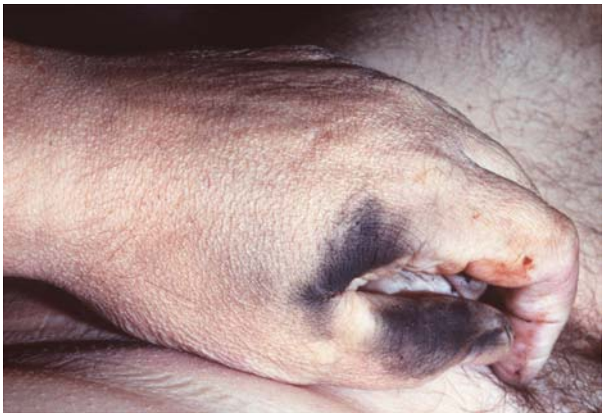

FIGURE 8.6 *Soot-soiling of hand that had grasped a shotgun barrel.*

CARBON MONOXIDE

Carbon monoxide in the gases combines with haemoglobin and myoglobin to give a pink coloration to the interior of the wound track and adjacent tissues. This diminishes in concentration along the track, but can still be present in its depths – and even at an exit wound, if there is one. The presence of carboxyhaemoglobin and myoglobin has been advocated as a test for distinguishing the exit from the entrance wound, especially where decomposition has blurred the morphological appearances. This test must be used with caution, though if quantitative measurements are made it can still be a valid exercise.

CONTACT WOUNDS OVER BONE

When the wound is made over a site with underlying bone, the wound may have a different appearance. Especially in

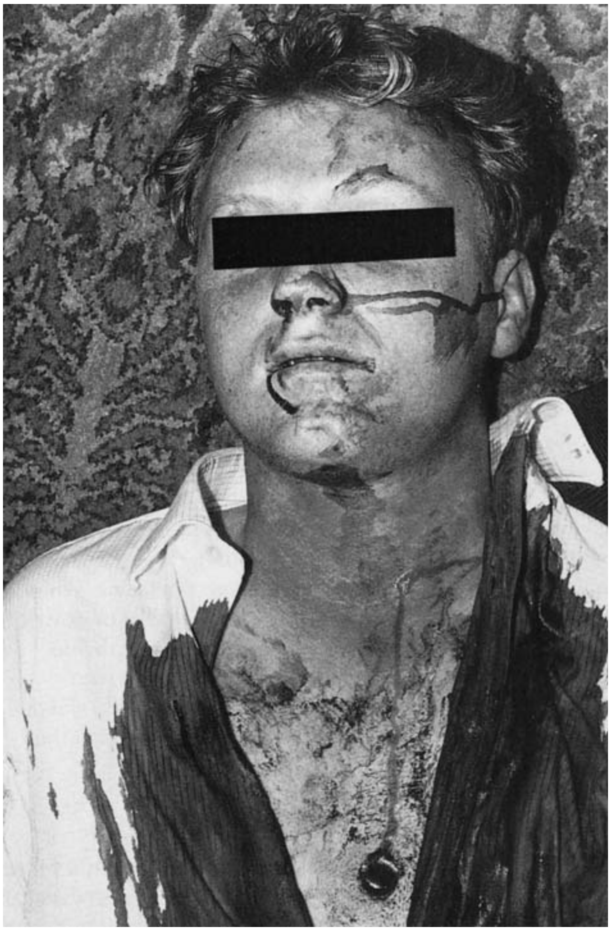

FIGURE 8.7 *Homicidal firearm wound with a twelve-bore shotgun. The wound shows a clear muzzle impression, the barrel having been pressed against the skin at discharge. There is almost no smoke soiling, the muzzle having forced a firm seal against the thin shirt and skin.*

the scalp, but also over the back of the neck, sternum, shoulder, hip and other areas where the soft tissue lies thinly over bone, the large volume of discharge gases cannot dissipate, as they can into the abdomen, chest, or mass of muscle. When the gas is forced through the skin, it is reflected back by the underlying hard layer and momentarily raises a dome of skin and subcutaneous tissue. This is rammed back against the muzzle to increase the pressure of contact further, sometimes producing a muzzle mark. When the gas volume is large, as with a twelve-bore gun, this dome may then split, causing a cruciate, stellate or ragged wound, with skin flaps. This is less likely with a .410 because the cartridge is smaller, but it can occur, as of course it can with rifled weapons.

The close discharge of a shotgun

When the muzzle is held near the skin, but not actually in contact with it, a number of helpful signs are produced. Once again the appearance is naturally modified by clothing, which must always be carefully preserved and subjected to forensic scientific examination. Where clothing is present, then it will trap most, but not necessarily all, of the soot and powder grains. It will reduce the flame effects, though if the fabric is ignited even momentarily, this can diffusely scorch the skin.

Assuming that the body surface is uncovered, a close discharge – one between actual contact and about 15 cm (6 inches) – is likely to show the following features:

- Singeing of hairs around the wound, unless the skin is hairless. The fine downy hairs of the trunk and limbs may be burnt away, though this is unlikely with dense, long, head hair. Where the distance is greater, the keratin of the hair may melt with the flame and then solidify on cooling, causing a ‘clubbed’ appearance of the hairs because of rounded bulges at the tips.
- Burning of the skin, unless protected by head hair. There may be a wide flare or narrow rim of hyperaemia or even blistering from the flame of incandescent gas blown from the muzzle.

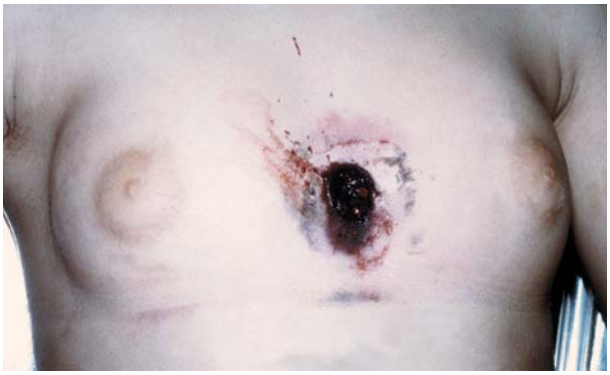

FIGURE 8.8 *Suicidal firearm wound from a twelve-bore shotgun. The wound shows a muzzle impression with soot soiling, the double-barrelled shotgun having been pressed against the skin at discharge.*

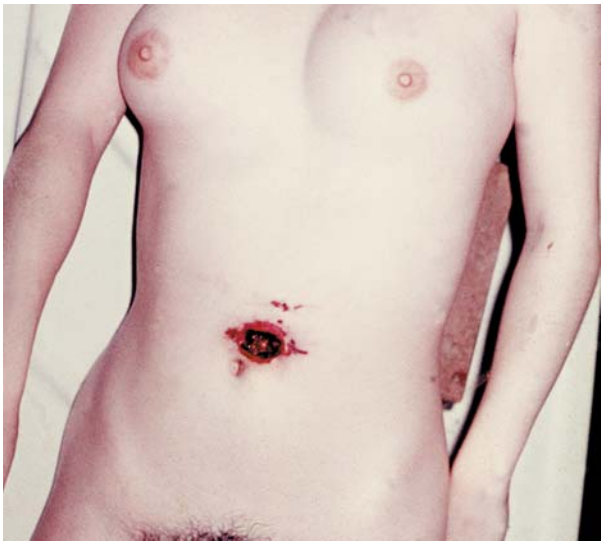

FIGURE 8.9 *Firm contact wound from a twelve-bore shotgun. The clothing has prevented any soot soiling, but there are minor peripheral abrasions from impact of a belt. As the abdomen was distensible, the gas expansion has not caused splitting of the skin at the wound edges.*

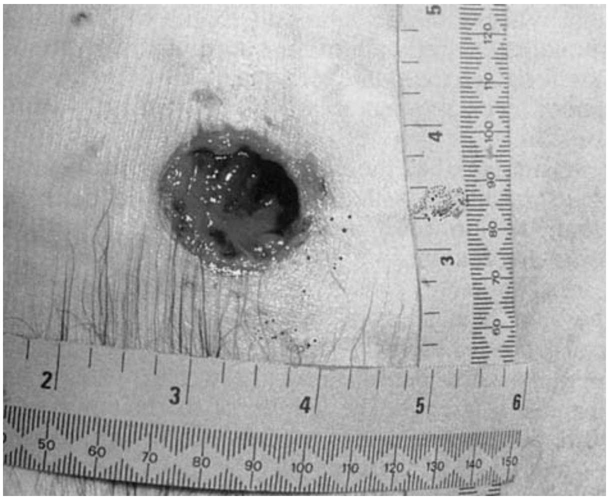

FIGURE 8.10 *Contact wound from a twelve-bore shotgun fired through clothing. There is some abrasion of the wound edges, but no irregularity from the close-packed pellets, which have not begun to disperse. The clothing has protected the skin from smoke and powder soiling. The circular wound indicates discharge at right angles to the body surface.*

- The tissues within and around the wound may be cherry-pink from absorption of carbon monoxide.
- There will be 'smoke soiling' or 'soot staining' of the skin, from carbon deposition. This spreads more widely than powder tattooing, but does not carry as far as the heavier propellant grains. This effect is much less or even absent with modern 'clean' propellants.

  This soot or smoke soiling is easily washed off the skin, and care must be taken at autopsy to obtain all photographs and trace samples before allowing the bloodstains to be washed from the body. Tattooing will not wash off, though adherent unburnt propellant may be physically detached from the tiny burns during any washing or swabbing process.
- Burning, partly burnt and unburnt propellant flakes and grains may pepper the surrounding of the wound. As stated above, these carry further than smoke staining, but usually spread over a smaller area. The incandescent particles cause small burns on the skin, but the unburnt flakes (such as nitrocellulose) may be seen as small glistening particles on the skin, sometimes coloured.
- The wound will be circular if the weapon is held at right angles to the skin and elliptical if slanted, often with undercut edges on one side. Depending partly on the size of the pellets, 'nibbling' or crenation of the edge may be seen, though this is usually imperceptible in close discharges. The edge of the wound will be blackened and there will be soot in the exposed tissues. There may be annular bruising around the wound and distant bruising a few centimetres away as a result of tissue damage from the entry of gases. Other bruises or abrasions may be seen in the vicinity from objects such

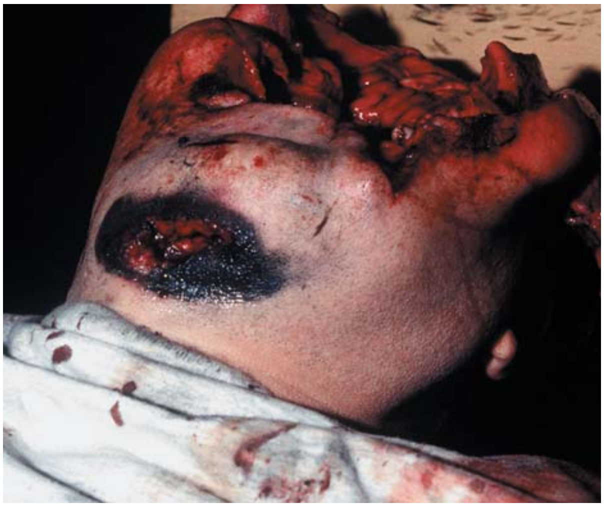

FIGURE 8.12 *Suicidal twelve-bore shotgun wound. The entrance is in one of the 'sites of election', in the neck, and the circular smoke fouling indicated a close or even contact discharge.*

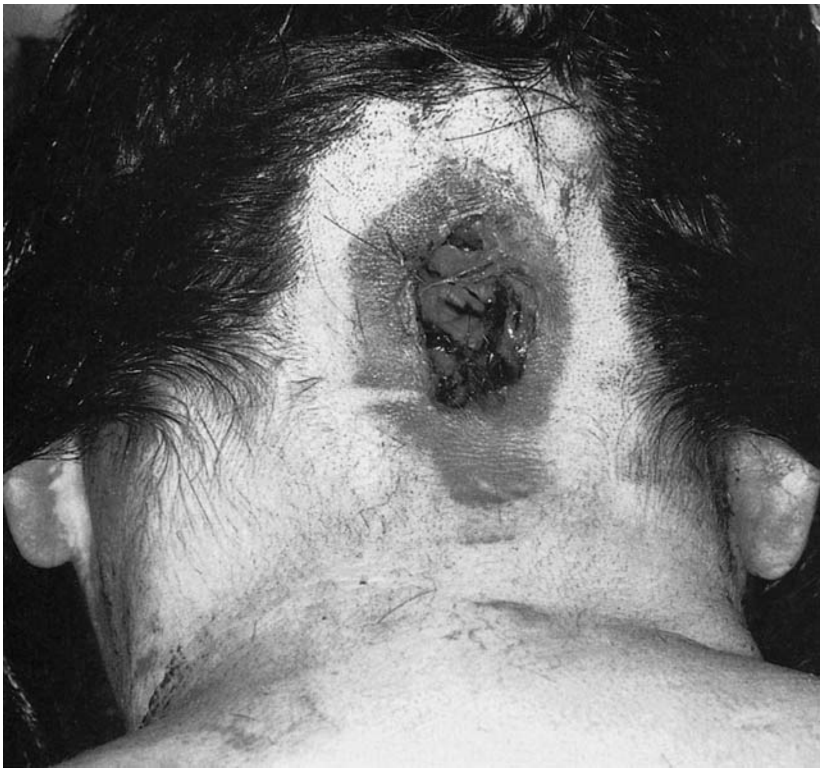

FIGURE 8.11 *Close-range twelve-bore shotgun wound of the head. This homicidal discharge was made through thick layers of fabric, which has filtered out any smoke or powder soiling. However, there is a wide zone of skin burning due to the muzzle gases. The wound is oval due to the downward angulation of the weapon.*

as buttons or buckles being hurled against the skin by the gas pressure.

- Any felt or cardboard wads or plastic cups from the cartridge will be within the depths of the wound.

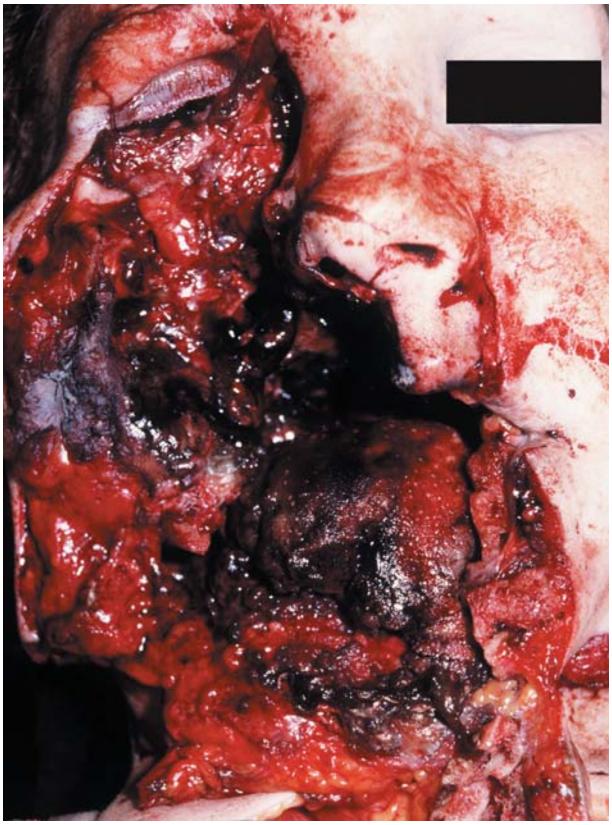

FIGURE 8.13 *Suicidal twelve-bore shotgun wound. The entrance is one of the 'sites of election', in the neck (see Figure 8.14).*

Short- to mid-range shotgun discharge

From about 15 cm (6 inches) to 2 m (6 feet), considerable variation occurs in the appearance of the wound. The shorter range will provide a similar picture to that of close discharge, but the soot soiling diminishes and over 20–40 cm it may vanish. Powder tattooing near the wound periphery may persist for somewhat further than this.

At the upper end of this range the edges of the wound will become crenated and scalloped, especially with larger shot. This is sometimes called the 'rat-hole' or 'rat-nibbling' from its resemblance to rodent teeth marks. From 2 m upwards, the number of satellite pellet holes will progressively increase around the main wound.

It must be emphasized that these are useful generalizations, but that marked variations occur between different weapons, different barrels of the same weapon and different ammunition fired from the same gun. Only test firing can give a measure of the length of the flame, the distance and pattern of soot staining and powder tattooing, the beginning of pellet spreading and the distance over which the wads are projected. As a rough rule-of-thumb, which is often incorrect, hair singeing occurs over the first 30 cm (1 foot), soot staining can be seen for the first half-metre (20 inches) and a single large hole persists for at least 1 m (3 feet).

Mid- to distant-range shotgun wounds

Here, test firing is even more vital for an estimate of range, as the variation is great. The choke as opposed to cylinder barrel on the same gun will provide different appearances at

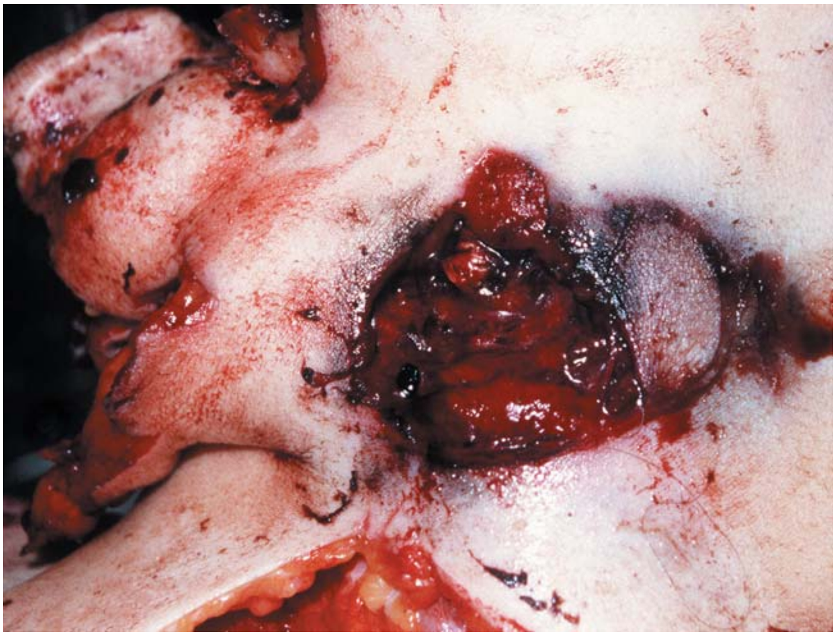

FIGURE 8.14 *Suicidal contact entrance wound in the neck. One barrel of a double-barrelled shotgun was used, the unfired barrel causing a muzzle mark (see Figure 8.13).*

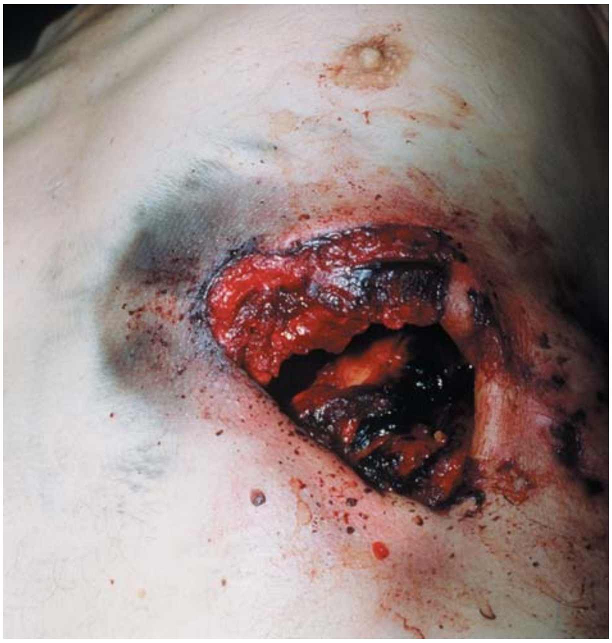

FIGURE 8.15 *Suicidal twelve-bore shotgun wound. Not all such wounds are circular – this one has torn a large defect in the chest wall because of its tangential orientation. The smoke smudging is extensive, having overshot and been deflected from the wound interior. The precordium is another site of election for suicidal shooting.*

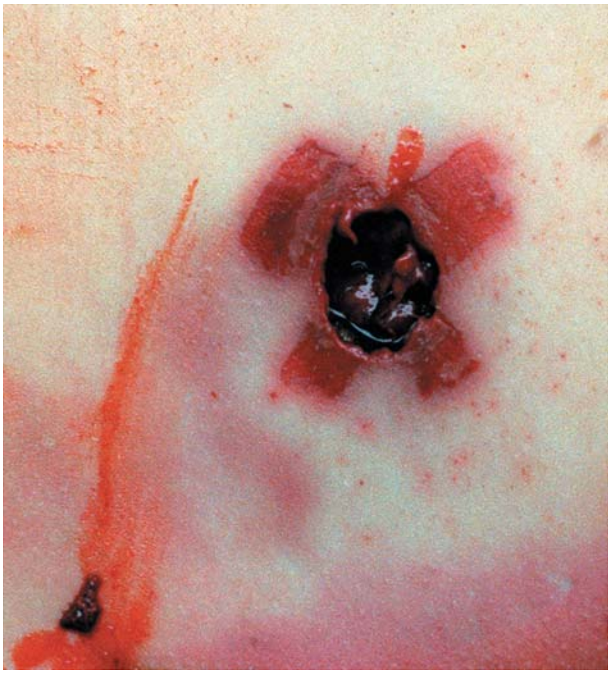

FIGURE 8.16 *Characteristic capital 'X'-shaped bruise caused by opening up of the plastic cup containing the shotgun pellets and surrounding the circular entrance wound from a homicidal short-range discharge.*

the same range. Once beyond 2 m, there will be no burning or smoke staining, rarely will there be powder tattooing and the presence of the wads will be variable. Sometimes the latter fall away within 2 m, but they can sometimes be found in a wound up to 5 m distant. Often the wad takes a lower trajectory and may strike the body below the shotgun wound. It may penetrate the skin, causing a second lacerated wound or it may only bruise the skin. Some of the plastic cup devices now used open up in flight to form a square-edged star or capital 'X' shape. This can strike the skin at or near the shotgun wound and produce a characteristic bruise or abrasion of a similar shape. The spread of pellets, which usually begins at a metre or two range, increases progressively, the central 'rat-hole' diminishing at the same rate.

There is an old (and extremely inaccurate) rule-of-thumb which states that the diameter of the spread in inches is roughly equal to the range in yards (metric: one-third of the spread in centimetres equals the range in metres). This of course applies to a circular wound or the narrow diameter of an oblique elliptical wound; it is of no use other than a quick first check to indicate if suicide can be excluded, but expert examination and test-firing are essential.

At distant ranges, beyond 6–10 m, the central hole may shrink to nothing. There will naturally be no wad injuries, no smoke, flame or tattooing, and no way of determining range except to say that it must be within the maximum discharge distance of that particular weapon, which may be 30–50 m. At such ranges the shot will not be lethal and the pellets, if they penetrate the skin, will lie just in the subcutaneous tissues. Death may occur, however, from an unlucky shot in the eye or from natural disease precipitated by pain and shock.

The direction of a shotgun injury

It has already been stated that, where the discharge has been at right angles to the body surface, the shape of the wound will be symmetrical and circular. A shotgun blast traces out a shallow cone from the muzzle and simple geometry demands that where this cone intersects a plane (the skin), a circle will result only when the cone is at 90° to the plane. In all other positions an ellipse will be traced out, its elongation increasing as the angle between them decreases. This pattern applies not only to the pellet spread but to soot deposition and provides a ready indication of the direction. In addition to the surface features, the internal wound assists in orientating the discharge:

- The wound edges may be shelved, the tissues being undercut below the margin distal to the origin of the

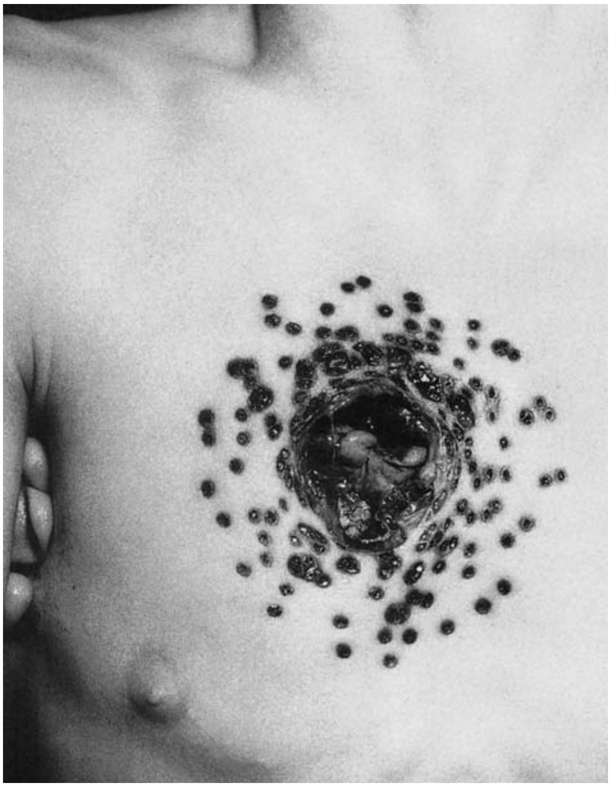

FIGURE 8.17 *Shotgun wound from a twelve-bore weapon. The circular outline indicates that the discharge was perpendicular to the skin surface. The central hole was caused by the entry of a confluent mass of shot, though there were no wads in the wound. The diameter of the peripheral pellet pattern was about 12 cm, which suggests that, approximately, the range was about 4 m. Test firing must always be carried out, however, to obtain a more exact estimate. Whatever the actual range, it is clear that it could not be suicidal.*

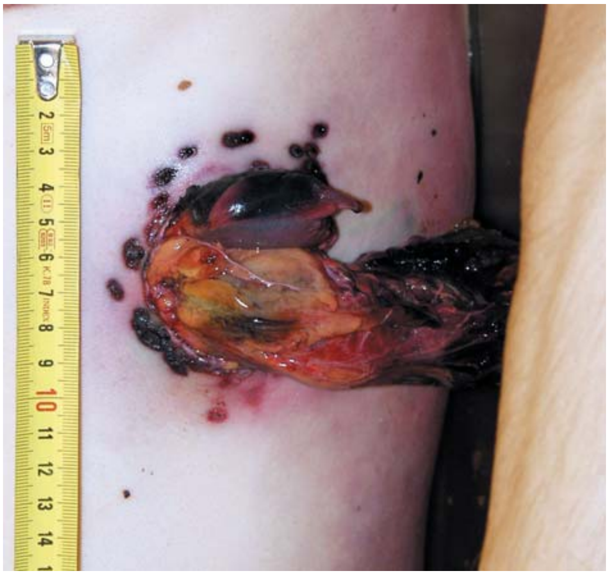

FIGURE 8.18 *Homicidal close-range shotgun wound from a twelve-bore weapon.*

discharge. The appearances are similar to those seen in some knife wounds, where an oblique stab leaves tissue visible at one edge and undercut on the other. In firearm wounds this is better seen in injuries from a single rifled projectile than from the more diffuse mass of shotgun pellets, but is still sometimes apparent in the latter.

- The track of the wound in the deep tissues can be established and this line projected backwards to indicate the discharge direction relative to the body. Again, a missile from a rifled weapon usually gives a clearer picture than those from a shotgun, but a general estimate can be obtained from a knowledge of the positions of the surface wound and the mass of pellets. Examination of a radiograph may be of more assistance that the laborious search from pellets at autopsy. Whatever the type of missile(s), considerable deviation caused by deflection by bone and other tissues can occur, so that the final resting place of the projectile may be far off the initial track. Thus the actual path through the tissues before ricochet must be established, if possible.

Other features of shotgun wounds

- Exit wounds are uncommon in the trunk as the energy possessed by each pellet is small because of its tiny size and the relatively low muzzle velocity of the weapon. The 'four-ten' will almost never exit and even the twelve-bore rarely causes a through-and-through wound of the adult chest or abdomen. The pellets often penetrate the distal chest wall, but are held up beneath the skin of the further side of the trunk. Commonly a bruise may be seen in the deep tissues of the frustrated exit site and lead shot may be felt under the skin, which is tough enough to prevent the final exit phase.

  In the head, neck and limbs, and in children and small thin adults, the twelve-bore commonly causes an exit wound which may be extremely large and ragged, with gross tissue destruction exposed. One of the most common is that seen when a twelve-bore is discharged suicidally into the mouth; this may disrupt the head, and expel most of the brain through a massive skull and scalp defect – the so-called 'burst head'.
- The internal track is more diffuse than that caused by a rifled weapon, though where the discharge is contact or close, the compacted mass of shot travels as a unit

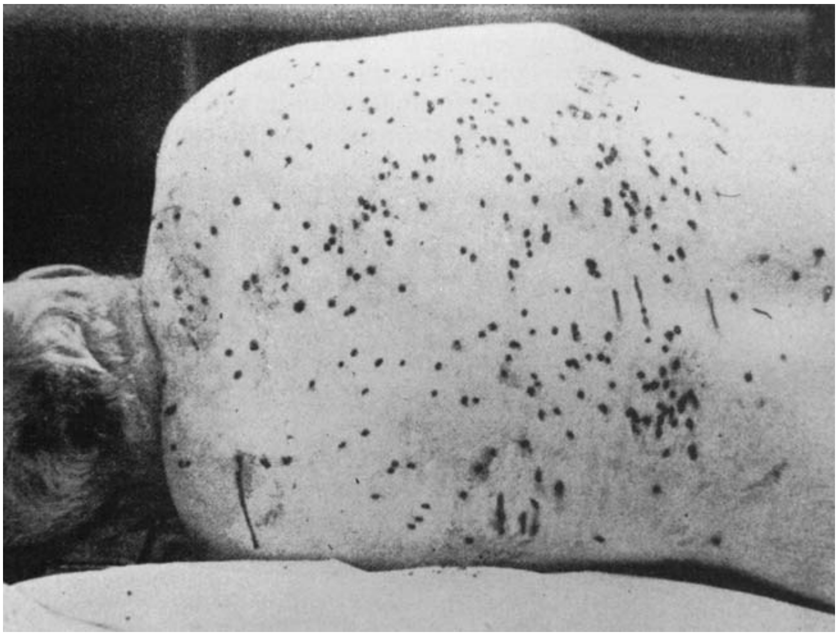

FIGURE 8.19 *Pellet injuries from a distant twelve-bore shotgun discharge. The range was about 25 m and the shot lies just under the skin. Death was caused by coronary heart disease, cardiac arrest being precipitated by the shock of injury.*

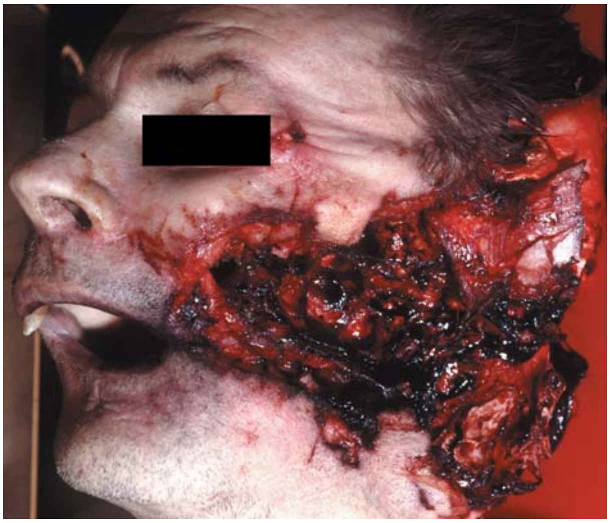

FIGURE 8.20 *Tangential shotgun wound of face. The edge of the dispersal cone of pellets and gas has ploughed through the cheek, and removed the ear. The facial bones and base of the skull were fractured.*

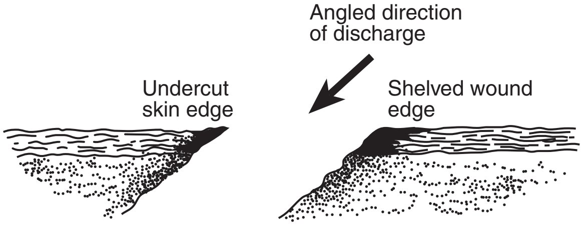

FIGURE 8.21 *Shelving and undercutting of entry wound in oblique discharge.*

for some distance into the tissues before dispersing. In general, the internal damage caused by a shotgun is diffuse and is caused by direct mechanical disruption by the gas and shot, which enter at relatively low velocity. There is no cavitation effect as is seen in high-velocity injuries, the tissue being physically smashed by the impact of a heavy load of pellets and, in the case of near discharges, by the large volume of hot propellant gases.

- Secondary damage is caused by the mass of pellets striking bone and releasing fragments that then act as secondary missiles, damaging adjacent tissues and sometimes emerging to form exit wounds. These may sometimes cause confusion to inexperienced observers, including police officers, who may wrongly suspect multiple shots or the use of some other type of weapon.
- Wad wounds can be caused by the non-metallic contents of a shotgun cartridge. These can vary from trivial bruises to fatal lacerations. The latter may occur when a blank cartridge is discharged at close range, the wad and accompanying gas breaching the body surface. As mentioned earlier, the new devices now often used in shotgun ammunition may cause characteristic wounds, such as the square cruciate mark from the wings of the opened plastic cup which contains the shot. Felt and plastic wads travel a variable distance from the gun muzzle, depending on many factors, such as the type of cartridge, the amount of propellant load and the nature of the wads. The wads fall away much more quickly than the lead shot because of their low weight and high air resistance. At contact and close ranges

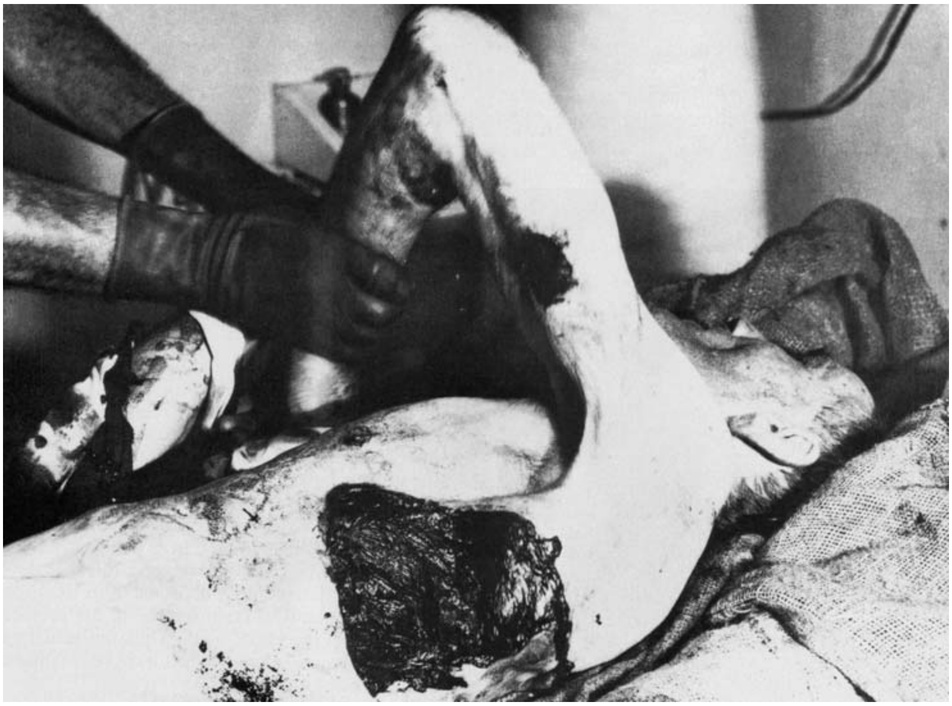

FIGURE 8.22 *A tangential shotgun wound of the side of the chest. When the victim was first found it was not appreciated that he had been shot. The discharge passed under the slightly abducted arm, injuring the upper arm as well as the thorax. Though most of the shot passed to the rear of the victim, some were found in the pleural cavity.*

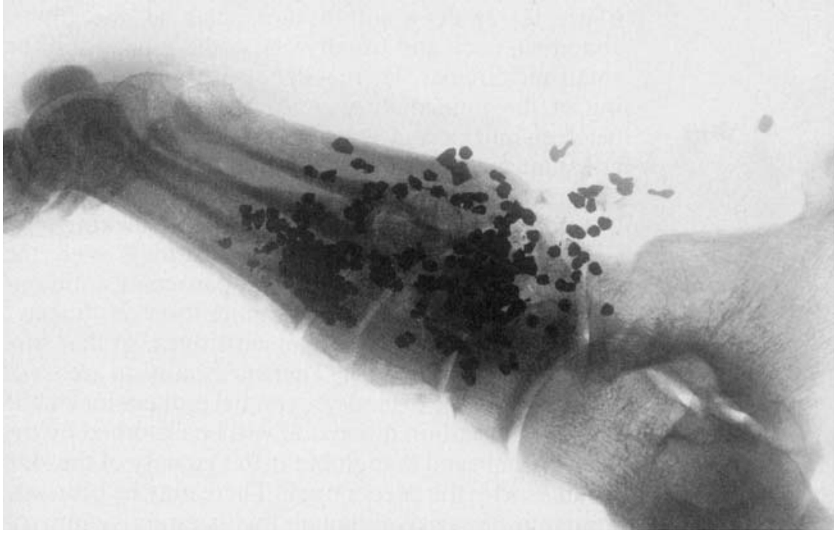

FIGURE 8.23 *Radiography should be carried out before autopsy in all deaths from firearms. The extent of dispersal of shot can be assessed – as in this film – and, in single bullet injuries, the position and possible fragmentation can be detected in order to assist in recovering the missile.*

they will almost always enter the wound and should be recovered at autopsy in order to assist in identifying the type of ammunition.

Beyond a metre or two, the wad may drop sufficiently to miss the main wound and strike the skin at a lower point. From 2 to 5 m, the wads may or may not strike the body, and further than this distance it is almost certain that the wads will fall to the ground. As with all variables associated with firearm wounds, however, it is essential for test firing to be performed wherever possible. The pathologist should not be dogmatic or impractically precise about distances, but offer a bracket of reasonable probability.

- Tangential wounds can offer considerable difficulty in recognition to the inexperienced observer. Some pathologists, as well as clinicians and police officers, have attributed certain wounds to knives or blunt injuries, when in fact they have been gunshot wounds. The shallow cone of a shotgun discharge may pass any part of the body so that the central axis is clear of the tissues, but the periphery strikes the body surface tangentially. The most common position is probably the side of the chest, in the lower axilla. If the arm is raised it may escape damage, but sometimes an opposing injury is seen on the chest wall and inner part of the upper arm. Another area is the side of the face, where a glancing gun blast may gouge the face and perhaps carry away an ear.

The appearance of the wound depends, of course, on the range and the depth to which the cone of shot and gas penetrates. On the side of the chest, a large elliptical area may be ripped from the surface to expose underlying muscle and ribs. If a close discharge, then the nature of the injury will be clear from the usual burning and soiling of the muzzle gases, but a more distant shot may cause a clean wound that can give rise to doubt about causation. Radiography will usually reveal lead shot in the tissues, unless the contact is extremely superficial.

In the chest – even if shot penetration is minimal – severe damage to the pleura and lungs may occur from the impact, and death may occur from a haemothorax with or without lung laceration or contusion. When the side of the head is hit, shattering of skull or facial bones can occur and intracranial damage is common, even though no metallic projectiles enter the cranium. Rifled weapons can also cause tangential wounds and some features of skidding across the surface and re-entry wounds have been mentioned earlier. There is not usually the wide gouging of tissue seen in smooth-bore injuries, but the glancing passage of a high-velocity missile can open up a severe longitudinal laceration. The lateral transfer of energy can cause severe internal damage in both chest and skull, even though the missile does not enter the cavities. The skull may be extensively fractured, with widespread underlying damage, and in the absence of any projectile and a deficient knowledge of the circumstances, some such injuries have erroneously been ascribed to blunt trauma.

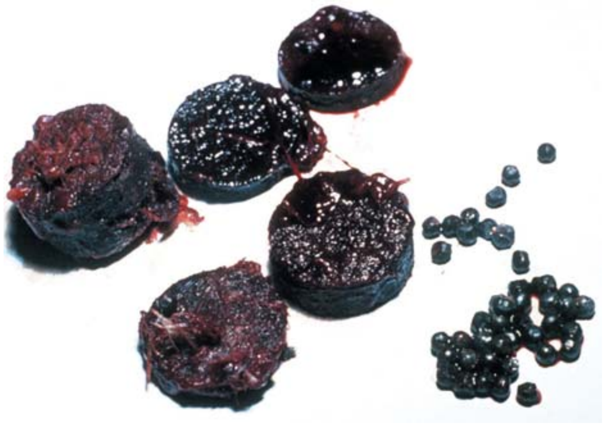

FIGURE 8.24 *Lead pellets and felt cartridge wads from a twelve-bore shotgun wound. It is important to recover at least a sample of pellets and the wads so that the type of cartridge can be identified.*

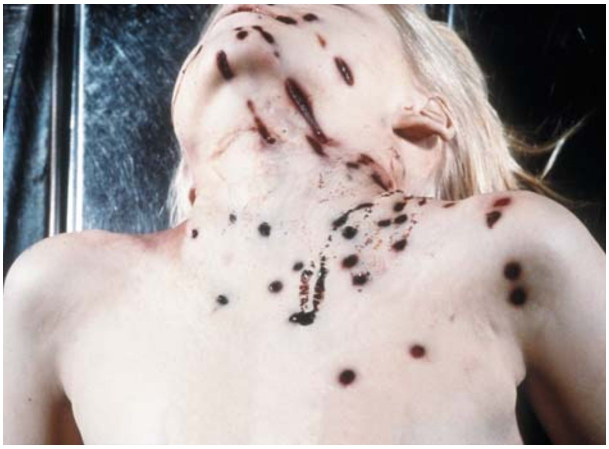

FIGURE 8.25 *Pellet injuries from a distant twelve-bore shotgun discharge (see also Figure 8.26). The child had been walking behind a hunter carrying a shotgun on his shoulder, the barrel pointing down. When the hunter was ducking to avoid branches of a tree the shotgun was triggered by a branch.*

WOUNDS FROM RIFLED WEAPONS

Injuries from rifled weapons vary greatly according to the velocity of the projectile, as discussed at the beginning of this chapter, but there are some characteristics common to all types. Unlike shotguns there is only one missile in each discharge, though an automatic weapon may well cause multiple wounds in close succession which impact upon the same area of the body. Occasionally, because of some defect in the weapon, two bullets may be fired at once, or a fragment of metal may detach from cartridge, barrel or bullet, so that two missiles may strike simultaneously; this is described later.

Contact wounds

When a weapon is fired in contact with the skin, the appearances can vary according to whether the muzzle was firmly pressed against the surface so as to form a tight seal – or whether it was merely held in loose contact, so that the backward jerk of recoil can permit a small gap to appear. Yet another permutation is when clothing is interposed between the muzzle and the skin.

When a rifled weapon is held firmly against the body surface during discharge, the resulting wound is in many respects similar to that from a shotgun. A twelve-bore wound

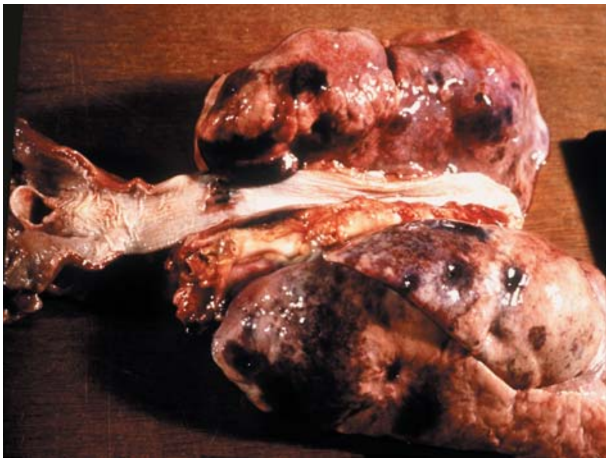

FIGURE 8.26 *Pellet injuries puncturing the oesophagus and both lungs (see also Figure 8.25).*

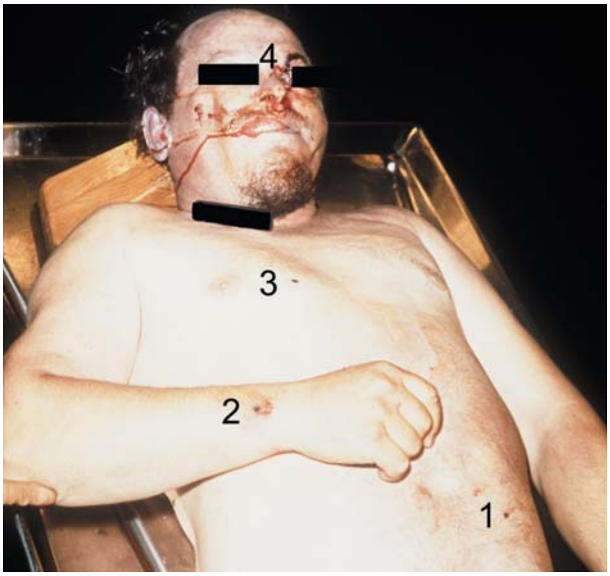

FIGURE 8.27 *Hunting accident in which the man was hit by four pellets from a distant twelve-bore shotgun discharge. Pellet no. 1 punctured the abdominal skin and was found in the subcutaneous fat; pellet no. 2 bruised the right wrist; pellet no. 3 penetrated the chest wall and was found in the pericardial fat tissue without any injury to the heart; pellet no. 4 hit the corner of the left eye and travelled through the left cerebral hemisphere, causing death.*

is usually much larger, but a .410 can closely resemble a large-calibre rifled weapon injury. Over deep soft tissues, such as the limbs, abdomen, neck and usually chest, the wound will be small and circular. There will be burning and blackening of the immediate wound edges, but if the seal between muzzle and skin is tight, there will be little or no sideways escape of flame and soiling unless clothing is interposed.

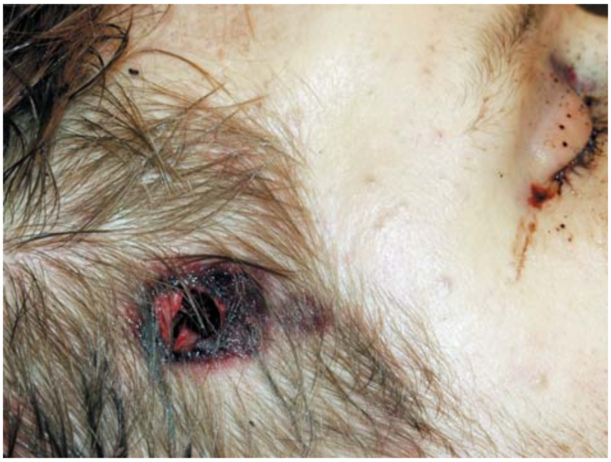

FIGURE 8.28 *Suicidal contact wound with muzzle mark from a 9 mm pistol (see also Figure 8.29 for corresponding exit wound).*

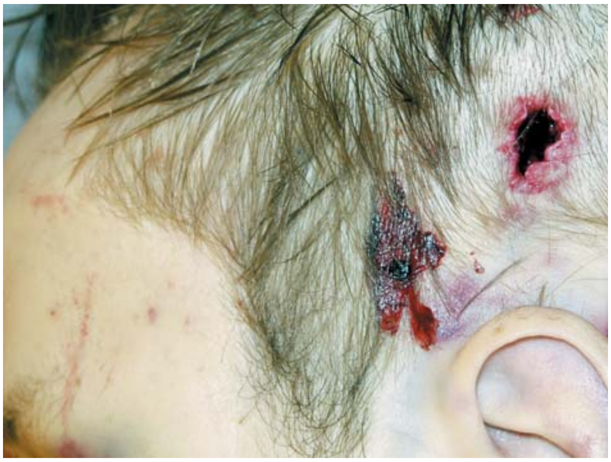

FIGURE 8.29 *Suicidal exit wound from a 9 mm pistol (see also Figure 8.28 for corresponding entrance wound).*

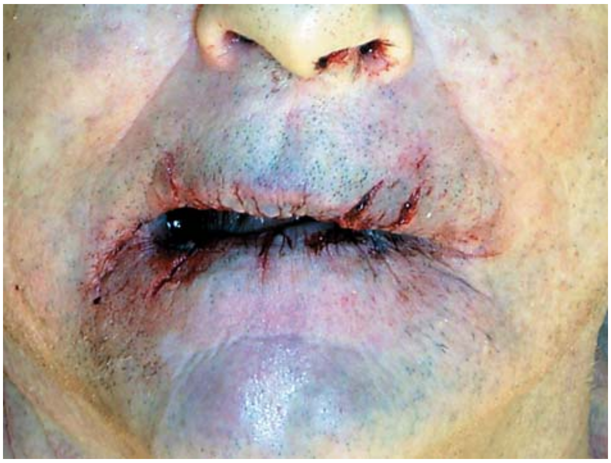

FIGURE 8.30 *Numerous tears in the lips due to expansion of muzzle gases from a 9 mm pistol fired in the mouth.*

Within the wound, there will be some flecking with soot in the tissues, the amount varying with the 'cleanliness' or otherwise of the particular ammunition. Most rifled weapon ammunition is 'clean', compared with many shotgun cartridges, so that soot may be absent altogether. There is usually an areola of hyperaemia that extends beyond the diameter of the muzzle and carbon monoxide will be absorbed by the haemoglobin and myoglobin in the vicinity of the skin wound and in the deeper track. There may be bruising, sometimes extensive, though this is rarely symmetrical and is often absent. The expansion of the tissues by gas entry tends to force the skin even harder against the end of the barrel and a 'muzzle imprint' may be formed. In rifled weapons, especially handguns, there are more features at the end of the barrel than in shotguns, so complex imprints may be made of foresights and mechanisms for self-loading in 'automatics'.

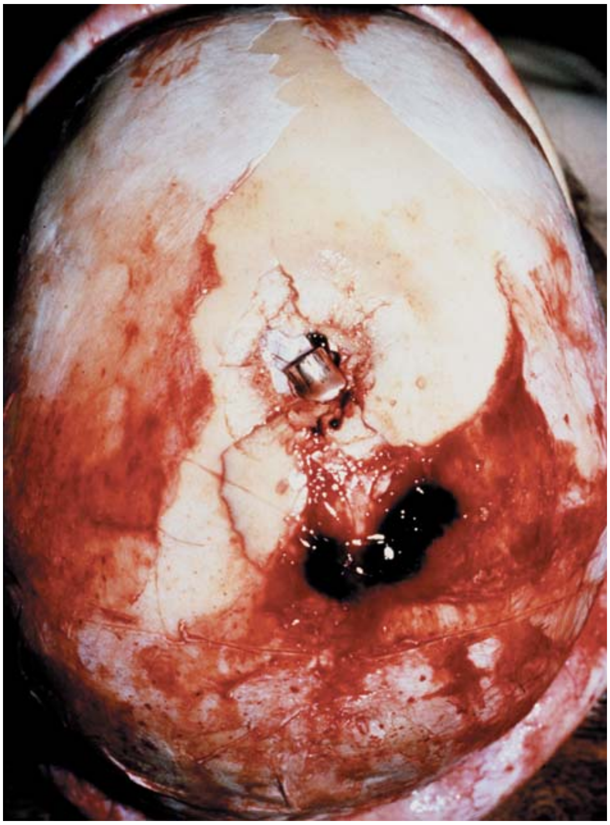

FIGURE 8.31 *Slightly deformed full-jacketed bullet in the middle of the cone-shaped exit wound on top of the skull. Homicidal entrance wound under the chin.*

If the contact wound is over a bony support, especially the skull, then the same phenomenon occurs as in shotgun wounds. The muzzle gases entering the subcutaneous tissues cannot expand by displacing adjacent soft structures and are reflected from the bone to raise a dome of gas that often splits the entry hole. This results in a linear, cruciate or stellate tear that may well destroy the original puncture. As the volume of gas produced is usually less than that formed by the large mass of propellant in a twelve-bore shotgun cartridge, the tear is smaller – though the charge in some older rifles and sporting guns may be comparable.

When the contact is loose, burning and soiling in the immediate vicinity of the muzzle may occur – again modified as in the case of shotguns, by the interposition of clothing between muzzle and skin. The fabric allows a lateral passageway for diffusion of soiling, as the muzzle cannot make such a good seal as against skin.

If the skin of the gun hand remains in firm contact with the weapon steel for a longer period of time, rust stains may develop as seen in Figure 8.34. In experiments carried out by Ulrich and Zollinger (2001) in the living as well as corpses, the formation of rust stains depends on the firmness of the contact, humidity of the skin and the environment, contact time and the state of the weapon surface. The shortest period of time necessary for the formation of rust under the most favourable circumstances, as reported by Ulrich and Zollinger, was 135 minutes for a corpse and 27 minutes for a living person.

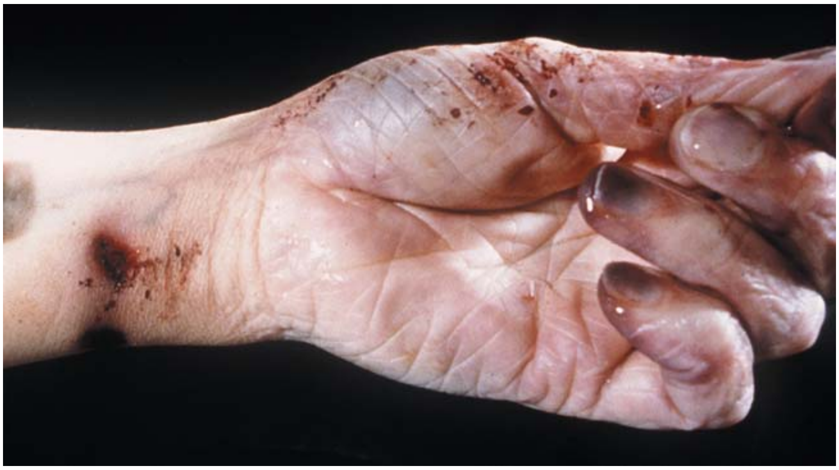

FIGURE 8.32 *Wrist of a homicide victim, who had raised her left hand to ward off the attack, showing a 'clean' exit wound of a full-jacketed bullet in the middle and soot-soiled entrance wound beneath it.*

Blow back into the barrel – 'back spatter'

Contact wounds – and also close-range discharges – may cause blood and tissue fragments to enter the muzzle, sometimes penetrating for several centimetres, this being called 'back spatter'. There appears to be a momentary suction effect after the pressure of the gas blast subsides, possibly as a result of rapid relative cooling in the barrel. This seems to aspirate material into the muzzle and pieces of skin, hair and adipose tissue have been found inside the weapon. Sometimes blood and tissue may soil the hand or arm of the person firing the gun, a matter of considerable significance in forensic science.

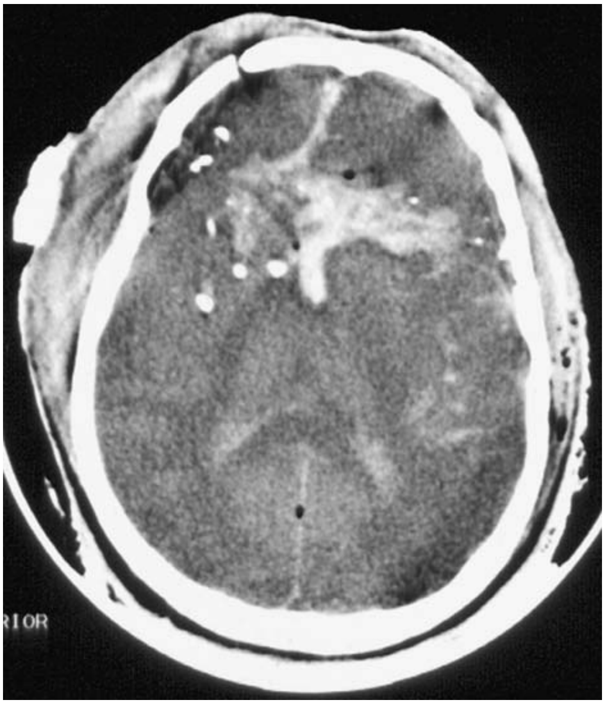

FIGURE 8.33 *Intravital CT scan of a 23-year-old male who in the course of a 'Russian roulette' shot himself in the right temple with a .357 calibre Magnum revolver. Extensive injury of the cranium and brain with several X-ray positive particles probably due to the disintegration of the bullet. (From the files of the Department of Radiology of Turku University Hospital reproduced by kind permission of Professor M Kormano.)*

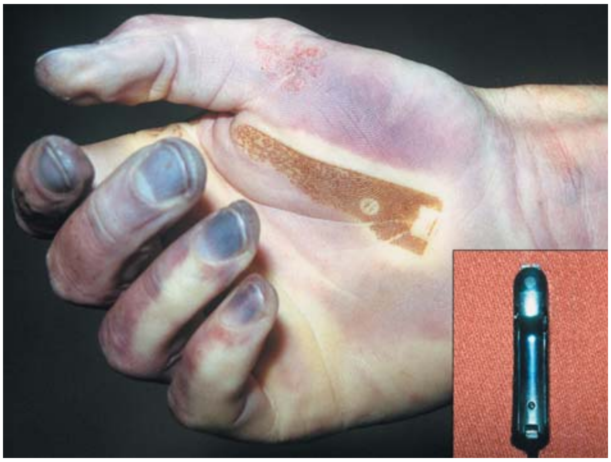

FIGURE 8.34 *Rust stains on the gun-hand of a suicide corresponding to the shape of the grip.*

Close-range wounds

When there is a short distance between the muzzle and body surface, the appearances of the wound will vary accordingly to the type of ammunition used. The wound

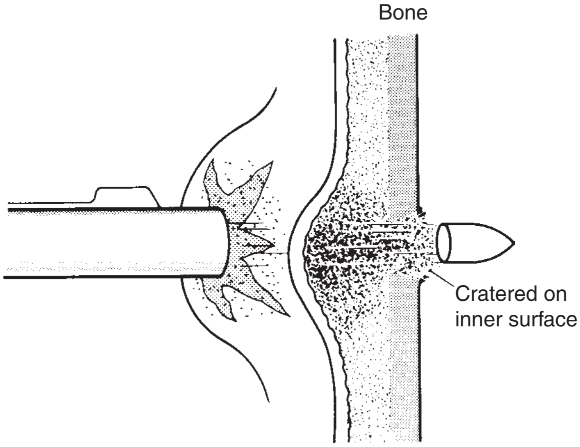

FIGURE 8.35 *A firm contact discharge against tissue overlying bone (such as the skull) causes gas to rebound from the rigid base. This raises a dome under the skin that can split to give a ragged entrance wound. This effect is by no means invariable.*

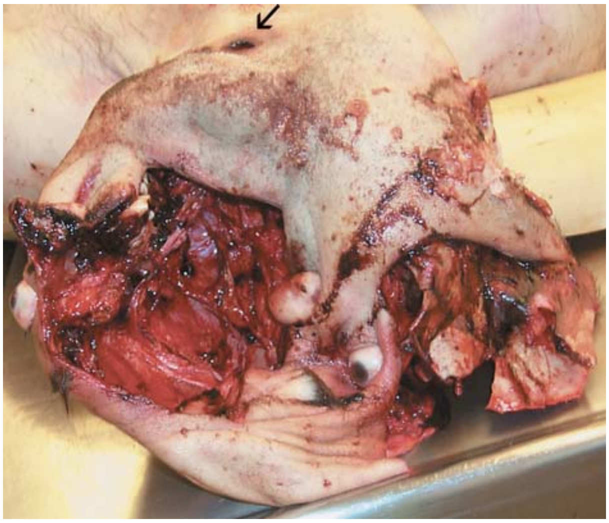

FIGURE 8.36 *Suicidal contact wound in the neck (arrow) from a 7.62 calibre rifle with destruction of the head.*

will almost invariably be circular as the tearing due to gas intrusion is unlikely to occur. The wound edges may be inverted, but at close range the rebounding gases may level up or even evert the margins. As will be described in medium-range wounds, there will be an abrasion collar.

Burning

Modern propellant produces less flame than the black powder of older shotgun cartridges, but within a few centimetres of bare skin there is likely to be burning of the skin with singeing of hairs. Even where hairs are not destroyed, they may be 'clubbed', that is, the keratin melts at the tip and resolidifies as a tiny blob. Hyperaemia around the wound is usual and some bruising can occur, which may be more evident as the post-mortem interval increases.

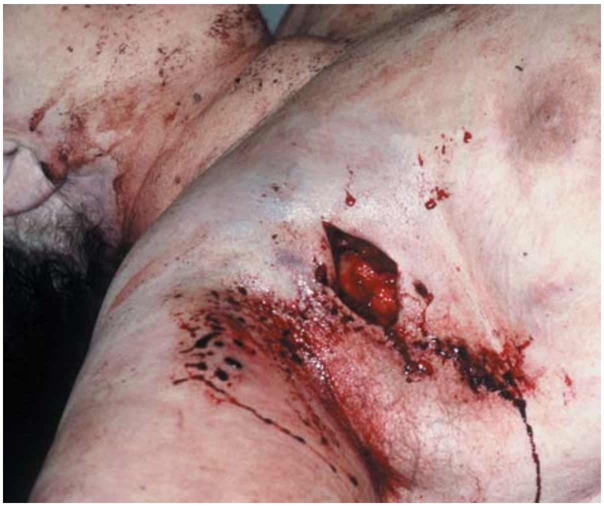

FIGURE 8.37 *Atypical slit-like exit wound of a homicidal firearm wound from a hunting rifle with entrance wound in the back.*

It is sometimes claimed that the flame effects extend from the muzzle a distance equal to the length of the barrel, but the truth of this rule-of-thumb is doubtful, though it may be a very rough approximation.

Soot or smoke soiling and powder tattooing

Though most propellant is now 'clean', some soot may be seen around the wound. In addition, there will be tiny burns from specks of incandescent propellant, the so-called 'powder tattooing'. These may pepper the vicinity of the entry hole in a roughly symmetrical pattern, which may assist in

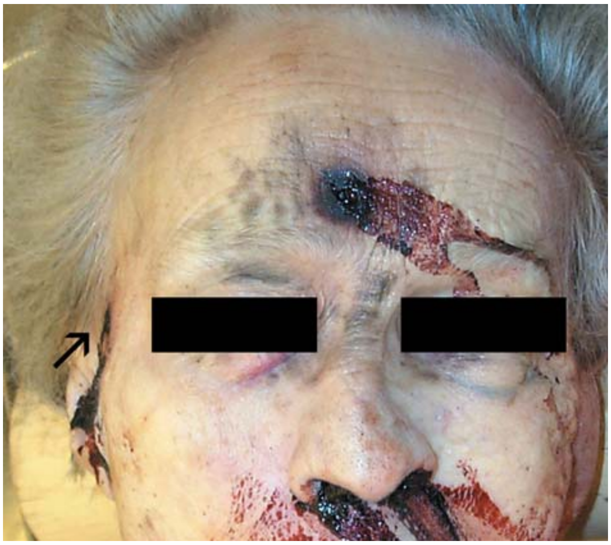

FIGURE 8.39 *Entrance wounds from a 7.65 calibre automatic pistol in the forehead and right side of the face (arrow) of a homicide victim, with soot soiling.*

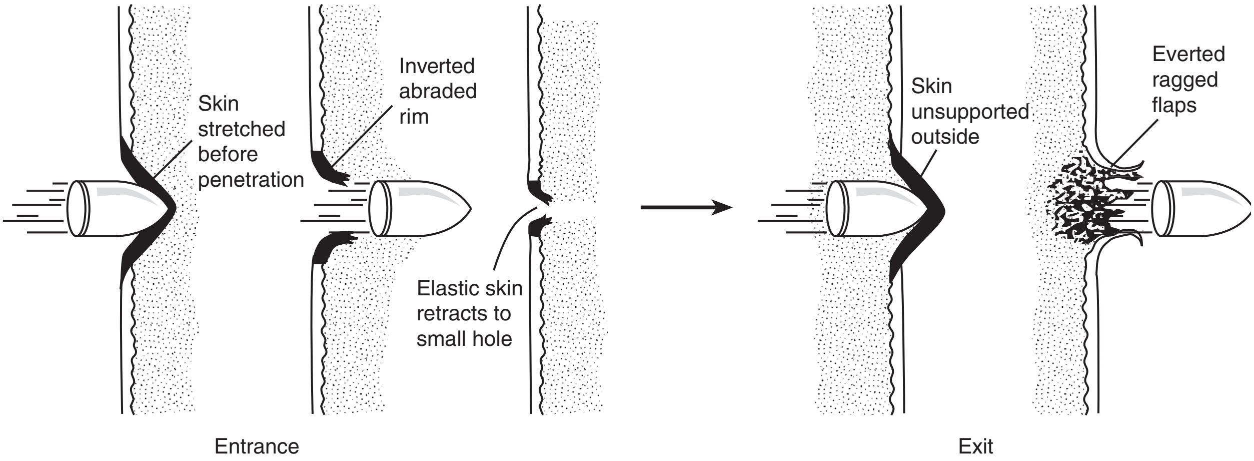

FIGURE 8.38 *Entrance and exit wounds.*

determining the angle of impact. Particles of propellant carry further than smoke, so tattooing around the entrance wound in the absence of soot indicates a greater distance. The soot rarely carries more than 15 cm (6 inches) from a pistol, whereas powder grains will travel 30–45 cm (12–18 inches).

These distances are approximate and can be doubled for a rifle compared with a handgun. The length of the barrel is another variable and a rule-of-thumb for pistols in this respect states that the distance from which powder residues can reach the skin is about twice the length of the barrel. It must be emphasized once again that test firing is essential for any hope of accuracy.

'Fouling' refers to tiny lesions around the entry wound caused by fragments of metal expelled by the discharge. These come either from the surface of the missile or from the interior of the barrel. Friction between the bullet and the rifling may gouge off pieces of lead, jacket, or barrel steel and hurl them at the skin, where they become embedded. These particles, like powder burns, will not wipe off at autopsy, whereas soot soiling can easily be removed with a damp sponge. Even fragments of the shell case may become detached and blown out of the barrel.

Minute flakes of unburnt explosive or inert filler may land on the same area of skin; these tend to glisten and may be orange or blue, depending on the manufacturer. It is important that a forensic scientist or police scene of crime officer has the opportunity to retrieve these before the wound is washed. If no such person is available, the pathologist should swab the area with a water-damped plain swab, or preferably cut out the skin around the wound and preserve it unfixed (refrigerated if necessary). These specimens should be sent for expert forensic examination, as they may indicate the nature of the ammunition and may be matched to that in the possession of a suspect. As with contact wounds, carboxyhaemoglobin and myoglobin will be present in the wound track in diminishing concentrations as the range increases.

The estimation of range is dealt with later, but here it is again emphasized that until test firing is carried out with a similar weapon and ammunition, only an approximate estimate can be offered from the appearances of the wound.

Medium-distance wounds

Unlike the shotgun, once the discharge of a rifled weapon is greater than a metre, there is nothing to indicate increasing range, until perhaps the erratic trajectory of extreme distance suggests maximum range. From perhaps less than half a metre to up to several kilometres the entry wound of a high-velocity rifled weapon appears the same. The typical appearance of the entry wound has been likened to that caused by 'driving a dirty pencil through the skin'. The central aperture is circular and may be inverted – that is, the margins are driven inwards by the passage of the missile – though this feature is often absent.

The size of the hole is rarely equal to the diameter of the missile and therefore the calibre of the weapon cannot instantly be determined by inspection of the wound – in spite of the apparent ease with which television drama doctors and detectives can instantly spot 'a thirty-two' or a 'thirty-eight'. When the bullet strikes the skin, the latter is stretched for a minute fraction of a second until penetration occurs, so that the missile perforates expanded skin. Immediately after entry, the skin undergoes elastic recoil and the resulting hole is usually smaller than the diameter of the bullet unless the range is so close that gas entry – or gas rebound – contributes to the size of the defect.

The abrasion collar

The skin immediately around the central hole is discoloured, the so-called 'abrasion collar'. This may be only a narrow rim or may be equal in width to the central defect. The collar is due to the inversion of the skin during penetration by the missile so that the sides of the bullet are 'wiped' by a tiny tube of skin, the epidermis of which is abraded by heat and friction.

The width of the abrasion collar is determined by the degree of inversion and hence the length of the skin tube. The collar is caused by friction and heat as the missile chafes the skin.

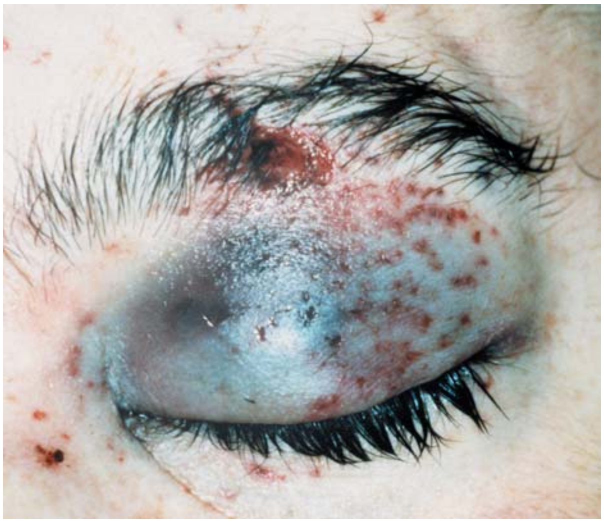

FIGURE 8.40 *Close-range pistol entrance wound. The eyelid and lower forehead show tiny burns from propellant flakes landing on the skin. The eyes are blackened by fractures of the floor of the anterior fossa.*

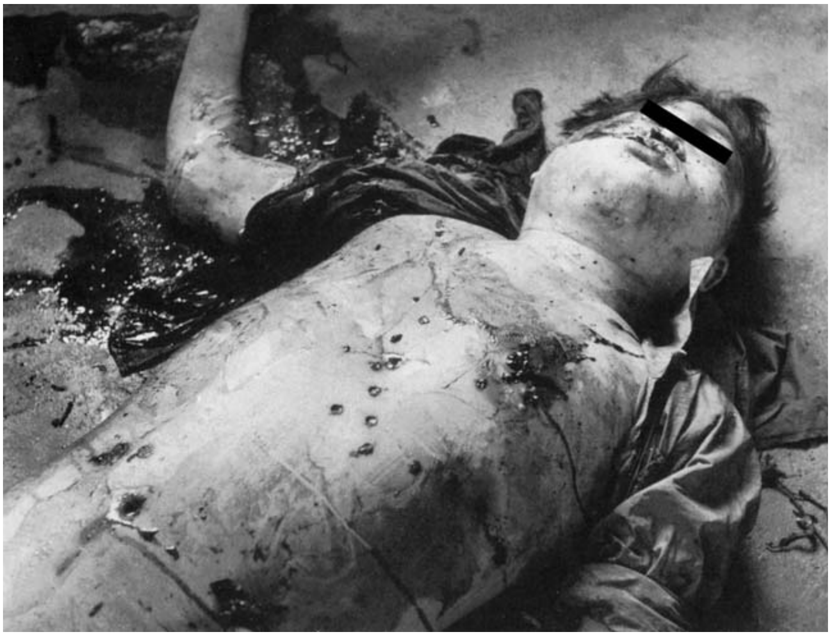

FIGURE 8.41 *Multiple entry wounds from a military self-loading weapon. The wounds are in line and approximately evenly spaced. Each is a typical entry, with an inverted appearance and an abrasion collar. There were massive exit wounds on the back, caused by .303 bullets from a Bren gun used by a Malayan patrol.*

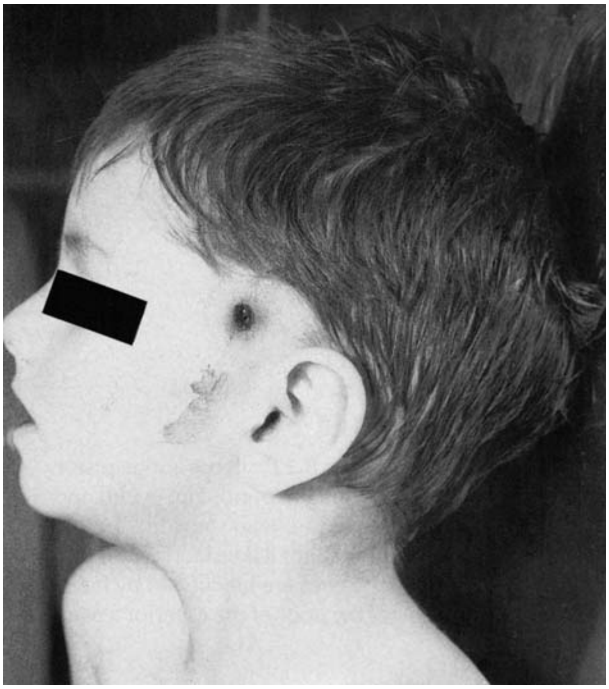

FIGURE 8.42 *Entrance of a .22 rifle bullet at a few inches range. There is slight smoke soiling around the wound, which has a well-marked abrasion collar. In fact, two shots were fired in quick succession through the same hole, which is slightly ovoid because of the non-congruence of the two shots (see Figure 8.49).*

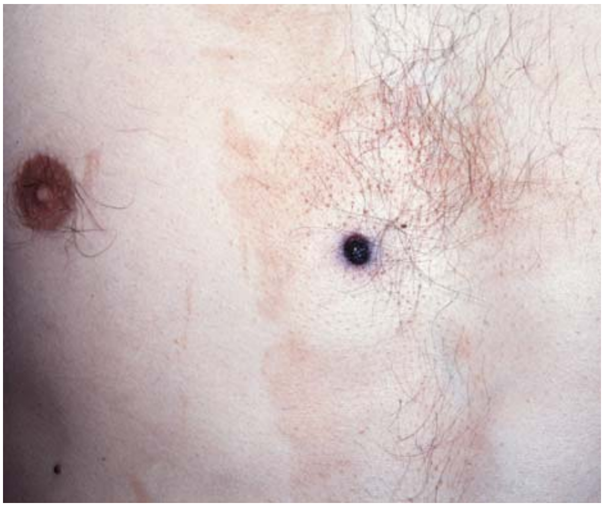

FIGURE 8.43 *Entrance of a .22 rifle bullet in a shooting accident.*

In addition to the true abrasion collar, there is often a slightly wider circle of peeled keratin, where the stratum corneum of the skin is raised to form a slightly frayed edge around the entry wound. This appears due to mechanical separation of the outer layer, due to the impact of the bullet knocking the deeper layers down away from the keratinized surface.

The grease ring

The inner edge of the abrasion collar may be black, as a result of heating effects and to the rubbing-off of dirt, lubricating oil, or grease and metal particles from the bullet. This is often called the 'grease ring' or 'ring of dirt', and may be absent if the missile was clean.

There may also be bruising around the wound from mechanical damage to adjacent subcutaneous blood vessels,

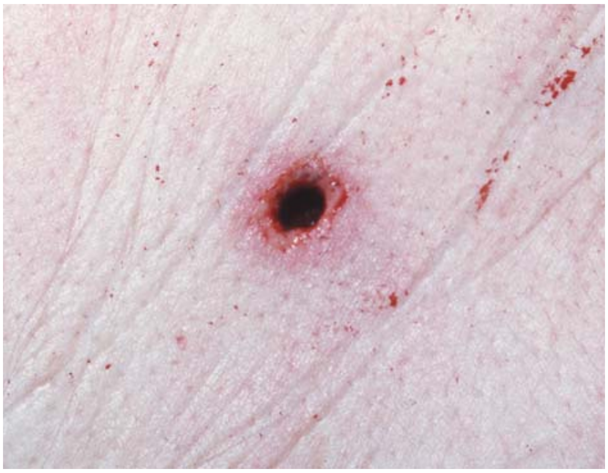

FIGURE 8.44 *Entrance wound in the back of a rifle bullet fired from across a room. There is thus no burning or soiling, and only minimal bruising at the margins of an almost circular inverted entry.*

but no other features, apart from the keratin ring, are present in the surrounding skin. No burns, soot or powder can be found beyond a variable distance that will not exceed a metre in the case of a rifle, and half that distance with a handgun.

Extreme trajectory wounds

Over most of their flight path, rifled weapon missiles maintain a fairly constant attitude. Soon after emergence from the muzzle, the bullet may 'precess', with the tip on the axis of travel and the base rotating in a diminishing circle about the axis. This soon settles down and causes no observable difference in the contact wound. For most of the distance of travel, the bullet remains steady, with only minor variations from axial stability. When the extreme range is approached, however, the reduced velocity allows instability of the flight path. The bullet begins to wobble and yaw, and may even tumble, that is, turn end-over-end. If the missile strikes the body in this phase, the impact may be sideways or even backwards, as well as being in an irregular lateral motion. A bullet striking sideways may leave an almost rectangular entry wound. The wound will then be similarly irregular, and there may be difficulty in recognizing it as a firearm wound or in differentiating it from a laceration caused by other means.

The direction of discharge

As with smooth-bore weapons, a close range discharge may provide information about the angle of the shot because of the geometry of any smoke soiling, powder tattooing, or burns on the skin. An impact at right angles will produce a circular zone of discoloration on the skin, while an oblique angle will result in an elliptical mark, the length of which increases as the angle decreases. As rifled weapons produce relatively little soiling except at close range, this clue is of less assistance than with shotguns.

Again as with smooth-bore guns, the edges of the wound may be shelved or 'undercut' when the discharge enters at an oblique angle. Tissue may be seen in the floor of the wound at one side, which indicates the direction from which the bullet came. Muscular contractions may distort this picture, however and it is not always present, though minute inspection (using a lens if necessary) may reveal shelving in the deeper layers of the skin rather than in the tissues below. More useful is the shape of the abrasion collar, which will be asymmetrical if the bullet enters at an angle, even through the central hole often remains circular. The other means of determining direction, which is not usually available with shotguns, is the trajectory between entrance and exit wounds. These can be joined and projected in direction of approach to indicate the

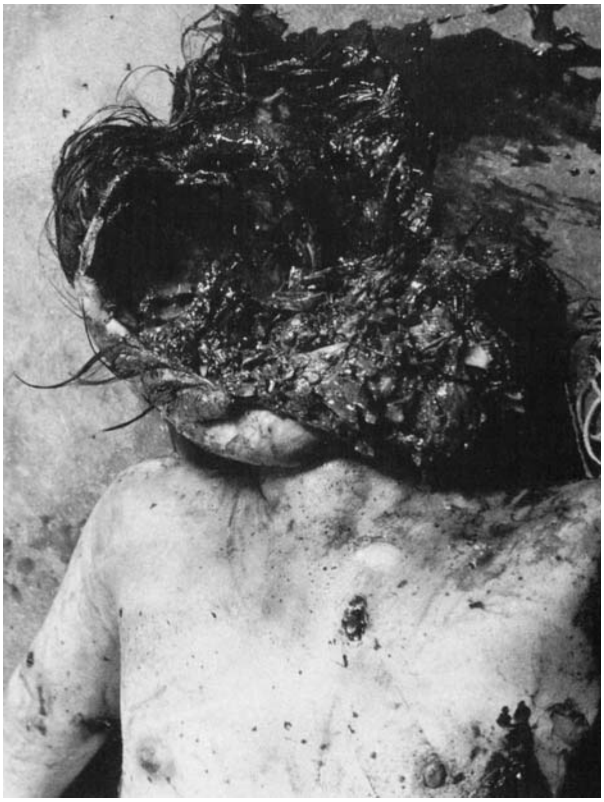

FIGURE 8.45 *Disintegration of the head of a young female terrorist in the Malayan emergency from a single bullet entering the external auditory meatus. The weapon was a F-N rifle with a high muzzle velocity; the impact of the missile on the base of the skull instantly converted the kinetic energy into a virtual explosion.*

site of discharge. Two important conditions attach to this interpretation, however:

- If the bullet strikes a solid object, either bone or even a firm organ, it may be diverted within the body to exit well off the original trajectory.
- The posture of the victim's body at the instant of impact must be taken into account. It is too often assumed, especially by lawyers, that all persons are injured – in whatever fashion – whilst standing passively in the 'anatomical position'. Nothing is further from the truth, as people, especially in conditions of fight, fright or flight, may be moving or dodging in a variety of postures, which change by the second. For example, in one of the author's (BK) cases, a woman accidentally shot by a police officer had the anterior entrance wound much higher than the exit in her back. This did not mean that she was shot from above, but that she was leaning forward when the bullet hit her. Much more bizarre examples occur from time to time, such as another case seen by the author (BK) in which a puzzling high-angle wound in the chest was later proved to have been fired from an upstairs window as the victim was knocking on the front door below.

Bullet wounds in bone

Where a bullet passes through bone, especially the thin bone of the skull, the well-known pattern of 'cratering' can usually be seen, though there are exceptions. Using the cranium as the best example, the initial contact of the missile punches a clean hole through the outer table of the skull. Where the bullet emerges internally, the inner table is then unsupported and a cone-shaped plug is detached, forming a crater that is appreciably larger than the external hole. If the bullet traverses the cranium and penetrates the opposite side, the same pattern occurs, this time with the small hole on the inner table and the crater on the outside. In addition to these penetrating defects, there are often fissured, sometimes comminuted, fractures running away from the central holes.

However, though this is the classical picture, a number of exceptions have been reported in the literature, where the cratering has been reversed or is otherwise atypical (Bhoopat 1995).

A shotgun usually destroys a large area of the skull, but similar shelving may be seen on the edges of bone fragments if a bolus of shot strikes the bone. Individual pellets rarely penetrate, but the massive cranial damage from close range of contact shotgun discharges is largely the result of the rapid and violent increase in intracranial pressure caused by the entry of a large volume of gas, as well as the bolus of shot and wads.

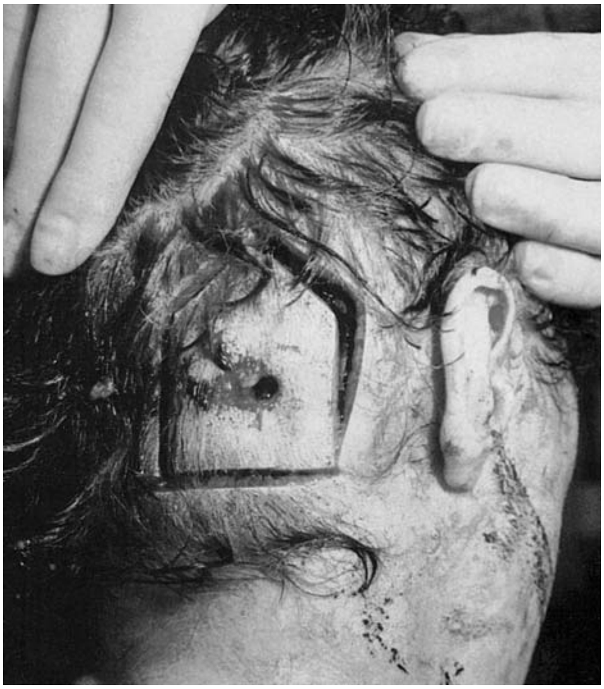

FIGURE 8.46 *Bullet entrance wound, showing the removal of a wide area of skin for forensic examination, the hair having already been removed and preserved before photography. The autopsy incision around the wound is wide, to include any discharge residue and burned skin. The shape of the incision is obviously asymmetrical, to indicate the orientation of the sample after removal from the body.*

Where several bullets have struck the cranium, the sequence of shots can usually be determined using Puppe's rule, which is applicable to any multiple blunt force causing skull fractures. Madea and Staak (1988) have developed this rule in relation to bullet injuries: the test depends on the observations of the fracture lines either when they intersect each other or when they intersect a cratered lesion, so that one can determine which crack or defect must have been formed first.

Where bones other than skull are hit by a bullet or shot, comminuted fractures are frequent and may be the cause of extensive secondary injury. Part of the kinetic energy of the missile is transferred to the bone and fragments may be violently displaced into the adjacent tissues. They may be extruded from the body altogether, leaving secondary exit wounds in the skin. These may give rise to difficulty in interpretation, as both lay persons and doctors may mistake them for lacerated or even stab wounds, unrelated to the firearm incident. Fractures of bone from gunshot impact may be of any variety, and radiography is necessary to detect and orientate fragments of bone and differentiate them from missiles and missile fragments.

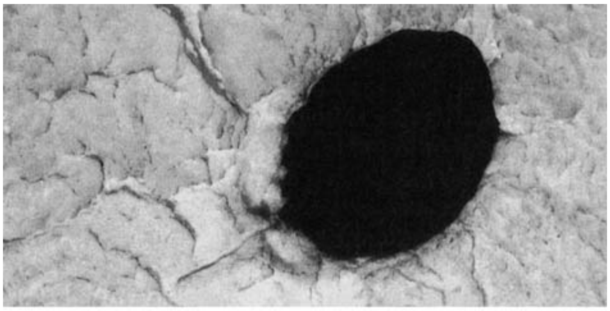

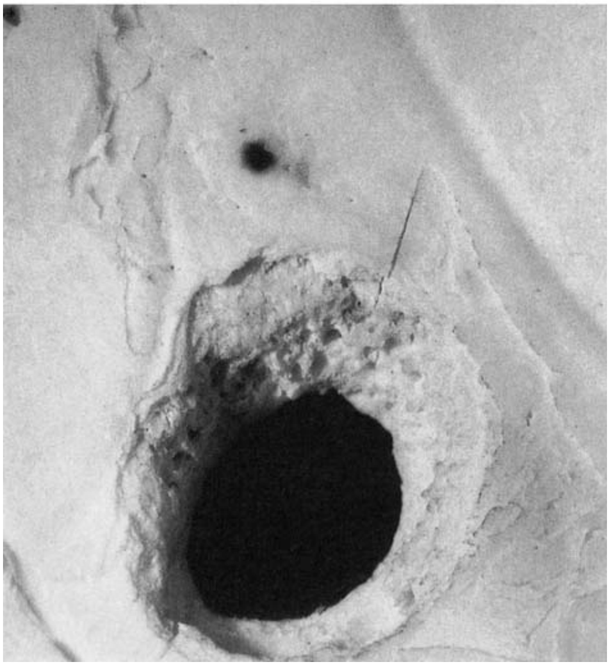

Figure 8.47 *The entrance wound of a bullet into a skull. Externally the defect is clean-cut and slightly ovoid because of the oblique impact of the missile. Internally, the inner table of the cranium is cratered, the bevel being caused by lack of support from the diploë. The exit hole on the opposite side of the skull showed a reversal of this pattern, with the cratering on the exterior surface.*

EXIT WOUNDS

When a gunshot wound completely traverses the part of the body struck, an exit wound will result.

Shotgun exit wounds

When smooth-bore weapons are involved, the residual energy of the individual pellets (which have a small mass and a relatively low initial velocity) is usually insufficient for them to emerge through the wider parts of the body, such as the chest and abdomen. Wounds of the limbs, neck and head may well be 'through-and-through' with a twelve-bore shotgun and occasionally, though rarely, with the small 'four-ten'. The range has an effect on the possibility of a shotgun exit wound, as contact or close-range discharge will allow the compacted mass of shot to act like a single bolus – in addition to which, the large volume of intruded gas assists in disrupting the tissues and clearing a path for the missiles, especially in the head. A shotgun blast into the mouth may well remove much of the back of the head as an exit wound, but this is in part caused by the explosion of the cranium by the gas, the so-called 'burst head', which can occur in any gunshot wound of the head, whether into the mouth or elsewhere.

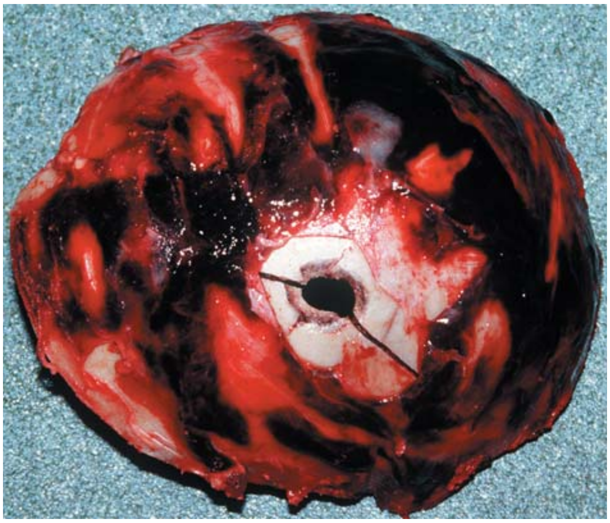

Figure 8.48 *Cone-shaped exit wound of the calvarium, with radial fractures in left parietal bone.*

The appearance of a shotgun exit wound is totally random, depending on the anatomical part injured and the addition of any bone or tissue fragments blown out in the discharge. Typically, it is a jagged, irregular laceration with everted edges that exposes deep tissues and comminuted bone fragments beneath. A shotgun injury of the chest or abdomen often just fails to exit on the opposite side, because the toughest obstruction, other than bone, is the skin. It is not uncommon to be able to palpate lead shot under the skin in the contralateral position after it has traversed the body and been trapped by the skin.

As a contact shotgun discharge may produce a large, ragged, entrance wound, this may resemble the exit wound, if present. Differentiation is usually simple, however, because of the burning and soiling present around the entrance. It has been said that the presence of carbon monoxide in the tissues marks the entrance wound, even in decomposed bodies, but carbon monoxide can in fact be detected right through a close shotgun discharge, though the concentration should be greater near the entrance.

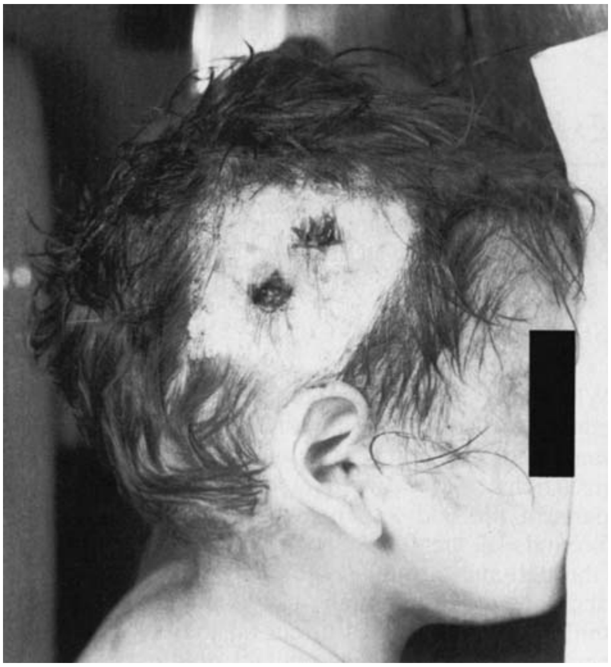

FIGURE 8.49 *The exit wounds in the fatality shown in Figure 8.42. The two bullets have diverged inside the head from their common entrance wound and emerged separately. The exit wounds are typically stellate, everted and clean. With small-calibre missiles, such as a .22, there is often no exit wound, but here the small head and thin skull provide less resistance to complete traversing.*

Exit wounds from rifled weapons

As stated above, many wounds from rifled guns are 'through-and-through', especially from high-velocity military weapons. The muzzle velocity is important in determining whether a single bullet will completely traverse the body. A high-velocity missile (of the order of 800 metres/second or 3000 feet/second) is likely to pass right through the body unless it strikes a large bony structure like the spine, large limb bone or base of the skull. As discussed at the beginning of the chapter, this may be inefficient in terms of wounding capability, as much of the kinetic energy fails to be transferred to the tissues to cause damage, although the cavitating effects of a high-speed missile in vital tissues may be extremely damaging. Much modern military ammunition consists of a small bullet with a large propellant charge to provide maximum muzzle velocity. This, by the familiar physics equation, provides substantial kinetic energy. The bullet is designed to fragment on impact, so that all the energy is transferred and an exit wound may not occur. In some shootings of this nature the pathologist may have difficulty in retrieving any metallic missile material, even after identifying it radiographically, as the fragmentation may reduce the bullet to granules and dust. If the missile does not strike bone or cartilage, it is likely to emerge on the other side of the body, even through the thickness of the chest or abdomen. Of course, in many cases where bone is penetrated (especially the skull or ribs), the bullet will still emerge, but this is considered below.

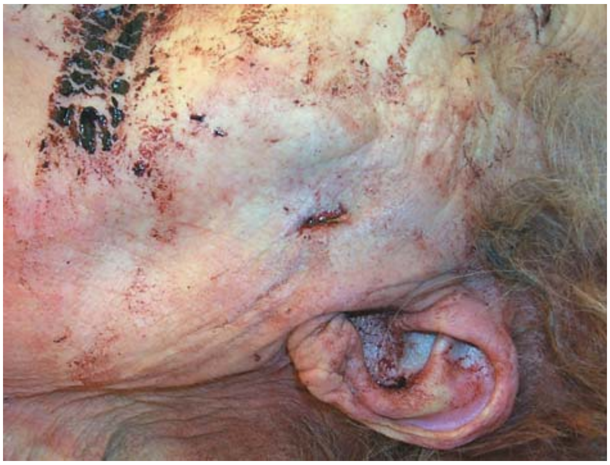

FIGURE 8.50 *Atypical slit-like exit wound of a 7.65 cal. bullet (see also Figure 8.39).*

If a fairly intact bullet emerges, not having been seriously deformed or fragmented, the exit wound is likely to be a small everted defect, either circular or with torn edges. The classical description is of a stellate wound, with triangular skin flaps at the margin, where the bullet has pushed out against unsupported skin. Many wounds may be cruciate or linear slits, however, and others are as circular as the entrance wound, which can give rise to some difficulty in interpretation.

It is naturally important to confirm the direction of the shot, but where a bullet has come from any distance between half a metre and near the extreme range of the weapon, then both entrance and exit wounds can look remarkably similar. There will be no burning or soiling to indicate the entry and identification will depend upon the close examination of the margins of the wounds. Where the entry was covered by clothing, fibres of fabric may have been driven into the wound, which can never occur at the exit.

Though the exit is usually everted, this may not occur if the skin is supported at the moment of penetration. A firm brassière band or an elasticated trouser waistband can press on the skin and prevent eversion, as can a plaster or fibreboard wall, if the victim was leaning against it. In these instances, there may even be a spurious 'abrasion collar', which is really caused by the emerging bullet slamming the wound periphery against the resisting surface. In difficult cases the pathologist or a deputy should examine clothing and the scene, to see if there are any such factors that will help in arriving at a definite answer as to direction.

Usually, the abrasion collar and frayed keratin ring of the entrance wound will be of diagnostic significance, as it is

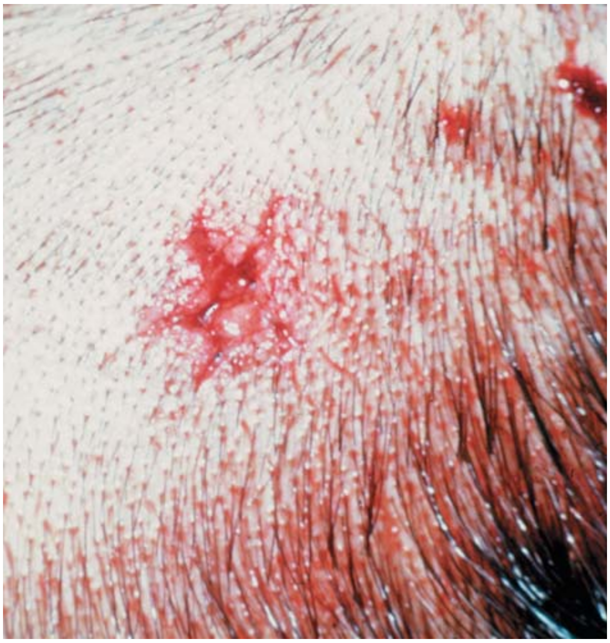

FIGURE 8.51 *Typical exit wound from a 9 mm bullet. The wound in the scalp is everted, stellate and quite clean, the missile having been wiped by its passage through the tissues.*

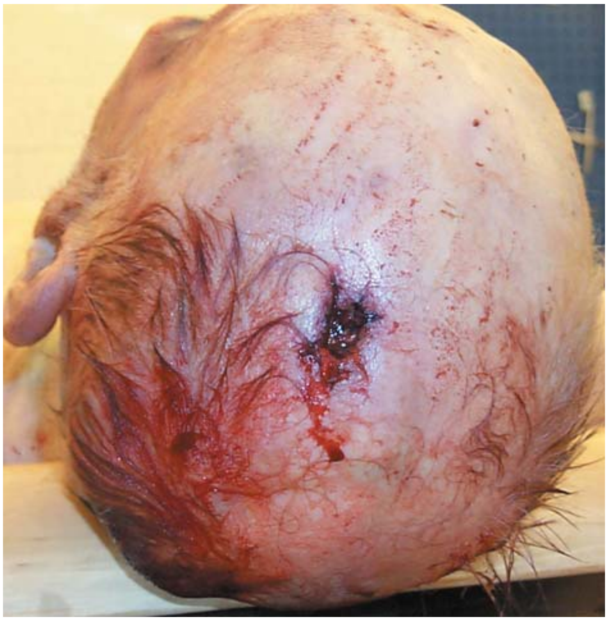

FIGURE 8.52 *Exit wound of a 9 mm pistol fired into the mouth.*

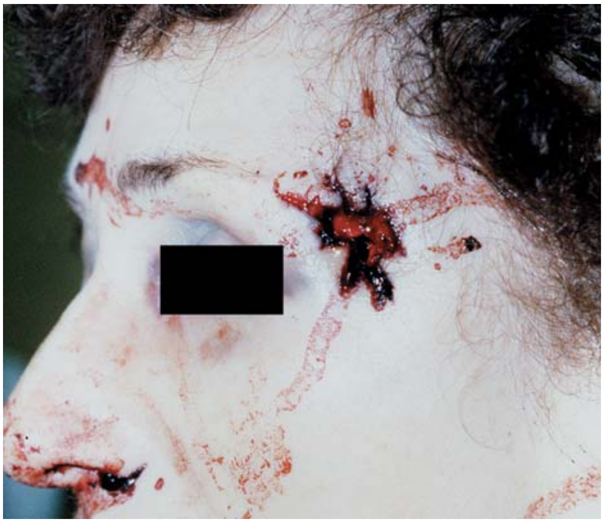

FIGURE 8.53 *Exit wound of the bullet responsible for the wound seen in Figure 8.46, showing the classical everted, split edges, with no soiling of the surrounding skin.*

absent in the exit wound, except as detailed above. Occasionally, however, the abrasion collar of the entrance wound is merely a narrow rim that needs close inspection to detect. Where the bullet has struck bone or fragmented, the exit wound may be of any size or shape, and may be multiple. If bone is shattered, pieces can be driven out through the skin, causing wounds that may be mistaken for lacerations from blunt or even incised weapon violence. Where the bullet has hit bone within the body, the bullet may emerge sideways causing a longer or otherwise distorted exit wound.

Of the many bizarre findings in gunshot wounds, re-entry can pose some problems. One case seen by the author had two entrance and one exit wounds, with a single bullet in the body. The victim had been standing against a wall when shot from the front: the bullet emerged through his back, ricocheted from the wall and re-entered his body.

Penetration of two parts of the body is quite common, a single missile causing several wounds. The most usual is a through-and-through wound of the arm or leg, the bullet then re-entering the abdomen or chest. Placing the limb in the appropriate position immediately clarifies the situation and can considerably aid an estimation of direction.

Glancing or tangential rifled weapon wounds

Though the large gaping wounds caused by a glancing blast from a shotgun are rarely seen with rifled weapons, some strange wounds can occur. A bullet may strike the body surface at a shallow angle, enter and then re-emerge some distance away, having travelled superficially under the skin. Within the skull, a missile may enter, follow the internal curve of the calvarium and exit some distance away. Cases are on record of such wounds causing relatively little damage, having traversed the meninges, rather than the brain substance.

When the body surface is irregular, such as the breast, buttocks or groin, several re-entries and exits can take place.

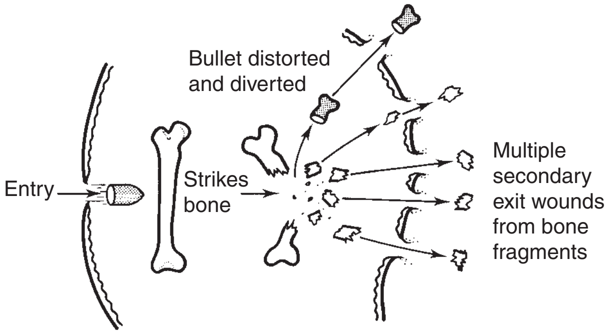
Bullet distorted and diverted
Entry
Strikes bone
Multiple secondary exit wounds from bone fragments

FIGURE 8.54 *Diversion of trajectory of bullet on striking bone. The secondary effects can be more damaging than those from the missile.*

In these instances, extensive bruising may mislead examiners into assuming that some blunt injury has occurred, if the firearm nature of the wound is overlooked.

A bullet striking at a small angle with the surface may follow the curvature of the skull or a rib and emerge some distance away after having made a track in the subcutaneous tissues. Alternatively, a bullet can pass tangentially across the surface, causing a linear graze or even a laceration, which may lead to great difficulties in interpretation if the circumstances are unknown. Gunshot wounds in the axilla or across the side of the thorax may gape widely, looking quite unlike a firearm injury to an unsuspecting medical examiner.

FIREARM DAMAGE TO INTERNAL ORGANS

This may be of any nature, but broadly falls into two categories:

- Contusion and laceration from low-velocity impact. Where a mass of shotgun pellets and gas strike tissues or an organ the damage is simple mechanical disruption, with widespread haemorrhage from local vessels. Similar, though less extensive, damage is likely from a low-velocity rifled bullet or slug, which pushes the tissues apart as it ploughs through an organ or muscle. Secondary damage can occur from fragmentation of bone or bullet, which produces further traumatizing missiles within the tissues. The danger to life that ensues naturally depends upon the target tissues – a clean wound through the thigh may cause only temporary disability, whereas the same wound through heart or brain may well be rapidly fatal.
- High-velocity missiles produce disproportionate damage relative to their diameter because of cavitation effects, described at the beginning of the chapter. This is particularly damaging in solid organs, such as the brain and liver. The track may be many times wider than the diameter of the bullet, and may consist of pulped and haemorrhagic tissue left behind by the pulsating cavitation effect caused by the lateral transfer of energy as the missile passes through.

ESTIMATING THE RANGE OF DISCHARGE

This is one of the most important aspects of the interpretation of firearm wounds. Though guidelines can be offered, it cannot be emphasized too strongly that every weapon and batch of ammunition will vary, so that medical opinion must always defer to the results of test firing with the same gun and similar shells. Allowance must also be made for the effect of interposed clothing.

Rifled weapons

TIGHT CONTACT OVER SOFT TISSUES

- Possible muzzle impression.
- Circular hole with abrasion collar.
- Bruising.
- Local reddening from heat and monoxide.
- Little or no surface burning.
- Little or no propellant soiling or powder tattooing.

TIGHT CONTACT OVER UNDERLYING BONE

- Split or cruciate wound.
- Local reddening and monoxide.
- Bruising.
- Little or no surface burning or propellant soiling.
- Abrasion collar partly lost on skin tags.

DISCHARGE FROM LESS THAN 15 CM (6 INCHES)

- Circular hole with abrasion collar.
- Flame burn on surrounding skin.
- Burnt hairs.
- Soot and smoke soiling (depending on ammunition).
- Small punctate burns from propellant tattooing.
- Unburnt propellant flakes.
- Little or no monoxide in tissues.

DISCHARGE FROM 15 TO 30 CM (6 TO 12 INCHES)

- No soot, but perhaps powder tattooing, depending on barrel length.

- No monoxide.
- Rarely, flame burns.

DISCHARGE FROM MORE THAN 40–60 CM (16–24 INCHES)

- Circular hole with abrasion collar.
- No burning, soiling, burnt hairs or monoxide.

FAR DISTANT DISCHARGE AT LIMIT OF RANGE

- Larger, irregular hole with irregular abrasion rim caused by tumbling bullet.
- No other features.

Smooth-bore shotguns

TIGHT CONTACT DISCHARGE OVER SOFT TISSUES

- Single circular wound about diameter of muzzle.
- Smooth margin.
- Often muzzle imprint.
- Blackened edge.
- No surrounding smoke soiling unless clothing allows leakage.
- Deep bruising.
- Pink tissues.
- Wads in wound.

TIGHT CONTACT DISCHARGE OVER SKULL OR BONY AREA

As last type, but wound may be ragged and split from gas rebound.

DISCHARGE WITHIN A FEW CENTIMETRES

- Circular wound, unless oblique discharge.
- Smooth or slightly crenated margin.
- No satellite pellet holes.
- Surrounding soot soiling.
- Punctate powder burns.
- May be unburnt powder flakes.
- Burning of surrounding skin.
- Burnt hairs.
- Pink monoxide in tissues.
- Wads in wound.

DISCHARGE FROM 30 CM (12 INCHES)

- Circular ‘rat-hole’ wound with nibbled margins.
- No satellite pellet holes.
- Soot soiling may persist.
- Powder tattooing present.
- Little or no monoxide.
- Still burning of skin and hairs.
- Wads in wound.

DISCHARGE FROM 1 TO 5 M (39 TO 195 INCHES)

- Central ‘rat-hole’ wound.
- Satellite pellet holes around periphery.
- No burning.
- No soot.
- Maybe slight tattooing at 1 m.
- No monoxide.
- Wads not in wound at upper part of this range.

DISCHARGE OVER 5 M (195 INCHES)

- Diffuse pellet pattern.
- Probably no central hole.
- No burning, soot, tattooing, monoxide, no wads.

All the foregoing data are very variable, being dependent upon the individual weapon and ammunition. Test firing must be used to validate the pathological interpretation.

AIRGUN INJURIES

Air rifles and pistols produce wounds that have some of the characteristics of both shotguns and explosive rifled weapons. The missile is single and fired through a rifled barrel, but is small and resembles a shaped (waisted) shotgun pellet.

The muzzle velocity is relatively low, but some more sophisticated rifles, such as the German Weirauch, can produce more than 12 foot-pounds of energy and the pellet can easily penetrate a wooden door – and the human skull.

Two calibres are common, the .177 and the .22 inch, the latter naturally being the more dangerous. The .177 can, however, penetrate the thin temporal bone of a child. Wounds from an air weapon are rarely fatal except when the head is struck – and children are the usual victims. The pellet can enter the skull and traverse the whole width of the brain, but never produce an exit wound, being usually found in the contralateral meninges. The track may be surprisingly wide for such a low-velocity missile.

Apart from fatal head wounds, a number of eye injuries occur from airguns, the pellet lodging within the globe, the orbit or the adjacent air-sinuses. Radiology is obviously of prime importance in locating the missile, prior to surgical removal.

There are no particular features of an air weapon wound, as, of course, there are none of the heat, powder or soot effects of an explosive gun. The entry wound is small, but may show an abrasion ring. There is never an exit wound, unless the tissue traversed is extremely narrow.

HUMANE (VETERINARY) KILLERS AND INDUSTRIAL 'STUD-GUNS'

Explosive weapons of a specialized type are used in abattoirs and by veterinary surgeons to dispatch large animals.

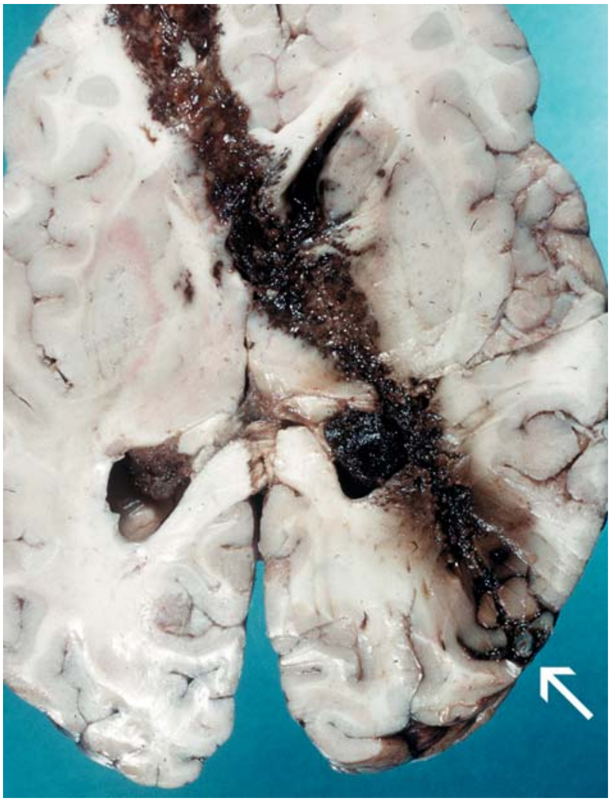

FIGURE 8.55 *Track through a child's brain made by a .177 inch air-rifle pellet (arrowed). Although a relatively low-velocity missile, the track is much wider than the pellet from lateral transfer of kinetic energy.*

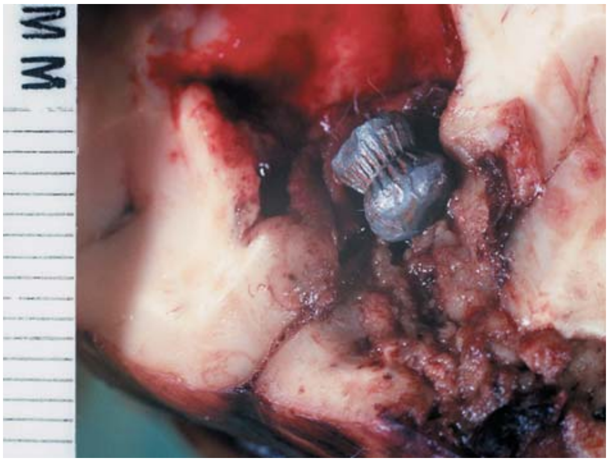

FIGURE 8.56 *Airgun pellet in brain tissue.*

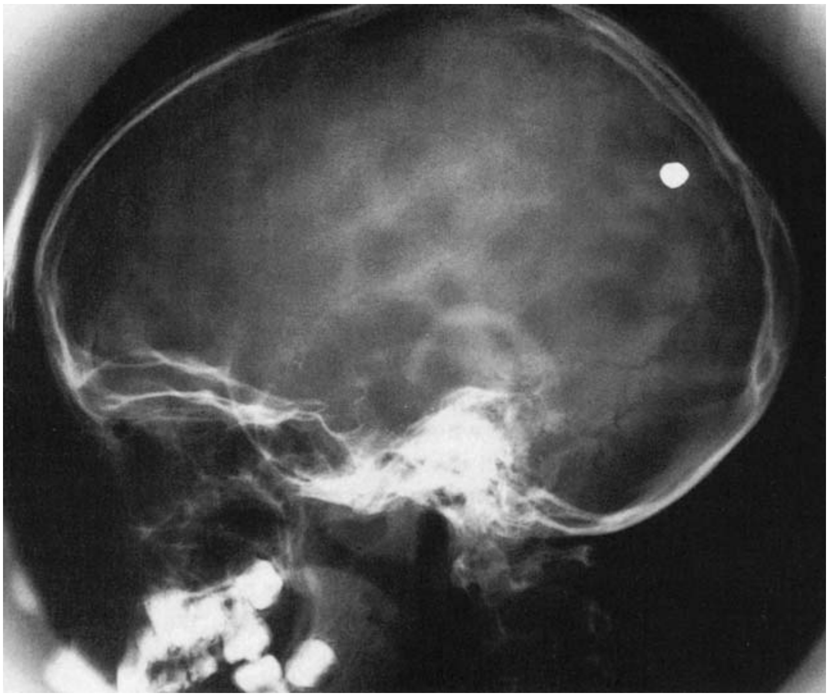

FIGURE 8.57 *Radiography can assist in locating a missile before autopsy, so that random searching does not cause instrument damage to the delicate striation pattern used for weapon identification. Here there is a lead pellet in the occipital lobe, with a bone defect in the extreme front of the skull.*

They may be of the ‘captive-bolt’ type, which consists of a metal rod, the distal end of which is propelled for a few centimetres from the muzzle of the weapon by a blank cartridge inserted in a chamber behind the proximal end. The captive bolt inflicts a clean, penetrating injury, about 5 cm deep, which can be mistaken for a stab-like injury.

Another variety consists of a metal barrel on the distal end of which is a heavy flange, which is placed against the head of the animal. A cartridge carrying a 0.31 inch (8 mm) bullet is inserted into the proximal end of the barrel and a firing pin struck with a mallet. The injuries from this resemble a contact wound from a conventional weapon, though the base of the flange may have embossed concentric circles to prevent slipping, which may produce an unusual patterned imprint on the skin surrounding the wound.

Another type is the Webley veterinary pistol, which is of .32 calibre and is virtually identical with a conventional gun. Humane killers may cause injuries on humans either accidentally or suicidally, homicide being rare. The nature of the weapon will almost always be apparent from the circumstances, including the presence of the weapon. Death may be delayed due to complications, as in one case seen by the author (BK) where a farmer died of tetanus after accidentally shooting himself in the knee with a humane killer.

Devices used in industry and especially the building trade, are similar to humane killers. These are the ‘stud-guns’ used

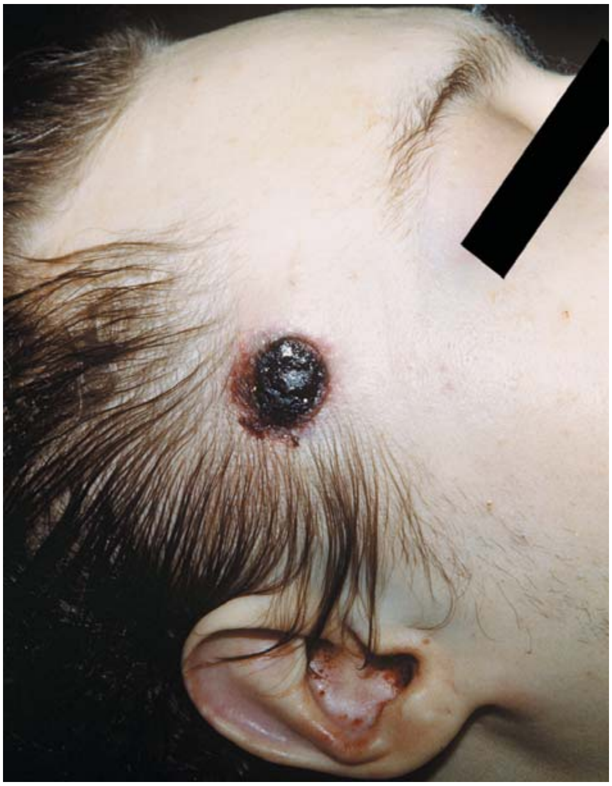

FIGURE 8.59 *Suicidal contact wound from a humane veterinary killer (see also Figure 8.60).*

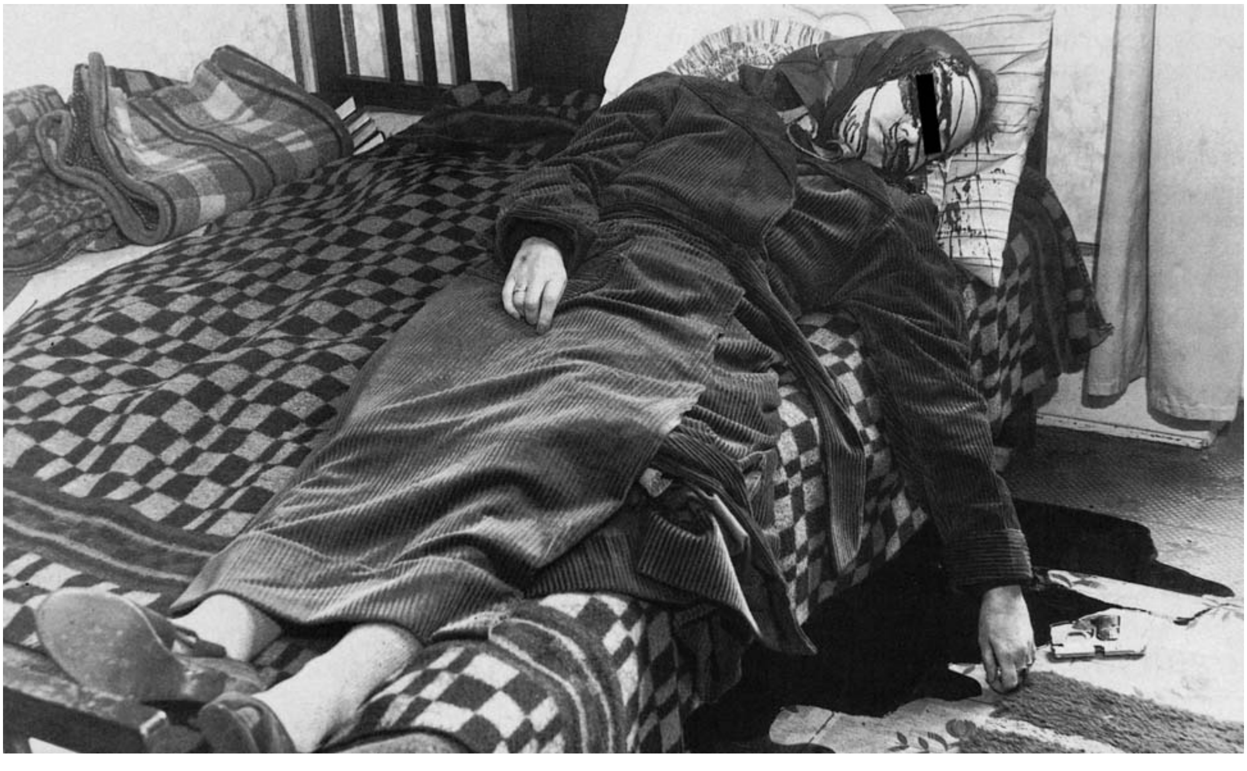

FIGURE 8.58 *A relatively unusual photograph in Britain, in that the suicide is of a woman and an automatic pistol is used. Note the massive haemorrhage on the floor, which illustrates the possible extent of post-mortem bleeding.*

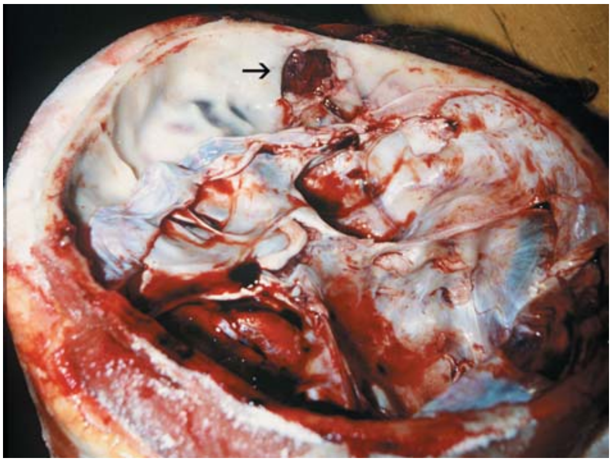

FIGURE 8.60 *Defect of the right frontal bone (arrow) from a humane veterinary killer (see also Figure 8.59).*

for driving a hardened metal pin or threaded stud into masonry, or other wood or metal structures as a rapid and convenient means of attachment. An explosive cartridge, often of .22 calibre, is used to drive a piston that acts as a hammer on the pin or stud. Many accidents and the occasional suicide have occurred from these tools, which resemble a long pistol, with a flat faceplate attached to the muzzle for pressing against the structure. Some accidents occur because the guns are used on relatively flimsy partitions, and the stud traverses the structure and flies on to injure or kill someone on the other side.

RUBBER AND PLASTIC BULLETS

A relatively recent innovation in firearms is the use by police and security forces of projectiles for riot control. Originally rubber bullets were used, being compressed into smooth-bore weapons. These have been abandoned in Britain in favour of 'baton rounds', more commonly known as 'plastic bullets'. They have been used in large quantities in Northern Ireland, some 43 000 having fired up to 1985. These projectiles have not yet been used in the mainland of Britain, though many police forces hold stocks. Elsewhere in the world, they are used extensively for riot control.

The baton round is a solid cylinder of polyvinylchloride 38 mm in diameter and 10 cm long, weighing 135 g. When fired from a smooth-bore weapon it has a muzzle velocity of about 250 km/hour and an effective range of 50–70 m. It should not be fired at a person under about 20 m range. It is used to disperse riots and discourage stone-throwers by striking them sufficiently hard to cause pain and bruising, but no serious injury. The bullet is about the same weight as a cricket ball, but travels twice as fast as the best bowler could deliver it. In Britain, the bullets were made by Pains-Wessex-Schermuly, but later by Brocks Fireworks and also by the Ministry of Defence. Many injuries and some deaths have been reported from their use. Fractures of the face and skull, eye damage including blinding, broken ribs and limbs, and contusions of liver, lung and spleen have occurred.

Police are instructed to fire at the lower half of the body, to avoid the more serious injuries arising from hitting the head or chest. The death rate is said to be 1/18 000 rounds fired, compared with rubber bullets which had a mortality rate of 1/4000 rounds.

DEATHS FROM EXPLOSION

Deaths from the effects of explosive substances or devices occur in both civil and 'military' circumstances, though the latter now include a considerable proportion of terrorist activities rather than conventional wars. Civil tragedies are usually industrial, as in individual incidents in mines and quarries, or on a larger scale, such as the detonation of chemical stores, ships or factories.

The nature of explosive trauma

Several different factors cause injury following an explosion, and the relative importance of each varies considerably with the type of detonation. For example, pure blast effects are far more important with high-explosive projectiles designed for purely military use than with the home-made terrorist bombs, the lethality of which may be primarily caused by flying fragments. Death and injury from explosives is caused by:

- blast effects
- impact of projectiles derived from the explosive device
- impact from surrounding objects and debris impelled by the explosion
- burns from hot gas and incandescent objects
- secondary injuries from falling masonry, beams and furnishings dislodged by the explosion.

A blast consists of a wave of compression passing rapidly through the air – or water in the case of a submarine explosion, such as a mine. The velocity of the shock wave depends on the distance from the epicentre, being many times the speed of sound at the start, but rapidly decreasing as it spreads out. The compression wave is followed by a transient zone of low pressure (below atmospheric), so that a rapid double change in pressure is suffered by the body. The magnitude of the blast varies with the energy released and also with the distance from the epicentre, the intensity obeying the inverse square law. With all but powerful military high explosives, the blast effects of most terrorist devices fall off so rapidly with distance that only those in the immediate vicinity will suffer severe damage from blast. Not uncommonly, this type of victim is the bomb-setter himself, when a premature detonation occurs.

A blast causes the most damage at an interface between tissues in contact with the atmosphere, which is why the lung usually suffers the most. About 100 lb/inch^2 (690 kPa) is the minimum threshold for serious damage to humans. The shock wave can pass through solid, homogeneous tissues such as liver and muscle, causing little or no damage, but in the lung there is marked variation in density between the alveolar walls and the contained air so that damping of the shock wave occurs and energy is absorbed, with disruptive effects.

There is controversy about whether pulmonary damage occurs from direct transmission of the shock wave through the thoracic wall, or through the oronasal orifices and air passages. In any event, the autopsy signs are of subpleural patchy haemorrhages, often in the line of the ribs, intrapulmonary haemorrhage and bullae at the lung margins. The air passages may be filled with bloody froth causing airway obstruction and hypoxia in addition to the primary damage.

Microscopically the alveolar walls can be shredded by the acceleration and consequent tearing of the air-sac wall when the shock waves traverses the lung. Desquamated alveolar and bronchial epithelium is seen lying free. Large areas of haemorrhage are seen, with either the alveolar pattern still visible or the architecture completely disorganized. Pulmonary haemorrhage in explosive incidents is often not caused solely by the blast, but may be from direct impact on the chest from flying objects, aspiration of blood from nasopharyngeal injuries and from bleeding following aspiration of stomach contents.

The ear may suffer severe damage from a blast, but this is not easily discernible at autopsy. Clinical examination and testing (obviously in the living) reveals many lesions, but these are of little relevance to the pathologist.

The gastrointestinal system suffers from the effects of a blast because, like the lung, it contains air and gases, and is thus not a uniform medium for transit of the shock wave. Once again haemorrhage is the most common lesion, usually small foci of the order of a centimetre in size. They may form circumferential bands around the intestine or may be confined to the serous coat as focal haemorrhages. The caecum and colon are more often injured than the ileum, jejunum and stomach, probably because they are larger and often contain more gas. Occasionally ruptures of the gut occur if the blast is violent and the victim in near proximity.

Victims of explosions in water suffer a reverse order of gastrointestinal to lung damage. The former is more common and may be manifested as haemorrhages under the serous coats and mucosa. Perforations may occur and, in survivors, peritonitis is not uncommon. The perforations are often multiple and in any part of the intestine. Lung damage is less common in water and some instances may be caused by impact transmitted up through the diaphragm.

Missile injuries in explosions

It is again emphasized that unless a sophisticated high-explosive military device is involved, or if the victim is virtually adjacent to a lower energy terrorist bomb, blast effects will rarely be the sole cause of death. The body will, however, be vulnerable to impact from solid fragments originating from the bomb casing or a container or conveyance, such as a car in which the bomb was concealed. Fragments of metals, from tiny splinters to large chunks or sheets, will be projected at high speed. The smaller ones will not travel more than a few metres, but larger, heavier pieces can fly over considerable distances and can cause serious or fatal injuries in just the same way as projectiles from a firearm. In the open, debris is scoured away, including dust and dirt, which can impinge on the body to injure and discolour it. At autopsy the body may appear pigmented from this dust blasting, the clothing causing a shadow effect similar to bathing suit protection from sunburn. A more common appearance is a ‘peppering’ as a result of the numerous small missiles causing bruises, lacerations and abrasions. The lacerations are often puncture wounds, of varying size and depth. Burns may also discolour the body, especially those areas unprotected by clothing. Flash burns from the bomb itself only affect those nearby unless the device is massive, though singeing of hair and eyebrows is not uncommon. Other burns may be caused by ignition of clothing, or by the building or vehicle catching alight from the bomb effects –or from gas or petrol ignition.

A massive bomb, or one where the victim is virtually on top of the device, may totally disrupt the body and fling unidentifiable pieces over a wide area. Alternatively, part of the body may be totally destroyed, sometimes the remainder of the victim being remarkably intact. The legs may be blown off or the abdomen disrupted, or the hands and arms torn away. In terrorist attacks these effects may be seen in the person who was planting the bomb or carrying it to the place chosen for detonation. A premature explosion, sometimes during the act of setting the timer, may cause these localized injuries.

The pathologist may be able to assist in reconstructing the events, as localized severe trauma obviously indicates the relative position of the bomb and the victim at the time of detonation. If the lower legs are destroyed, then the person was standing near the device that lay on the ground. Punctured wounds and bruises on the front of the thighs and trunk will indicate that he was facing the bomb. If the thighs, pelvic region and abdomen are damaged, the bomb may have been carried on the lap – and if the hands, chest and face are the most affected areas, the deceased may have been bending over the device.

The autopsy in explosion deaths

The salient features have already been described and much of the autopsy is directed at listing the injuries, as in any other trauma death. The interpretation outlined above, however, may be most important for the investigators.

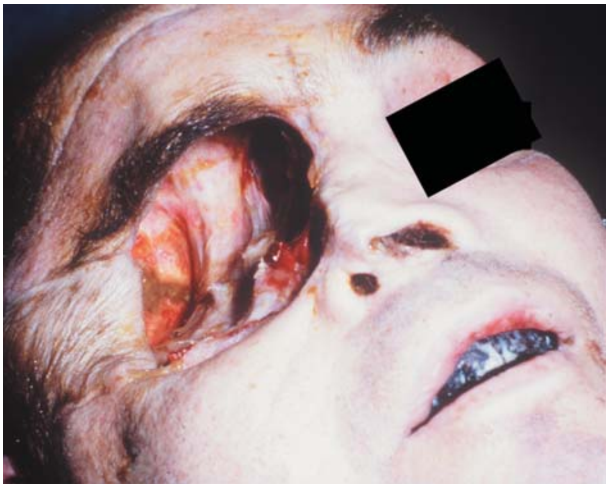

FIGURE 8.61 *Localized injury by an explosive device. The soldier picked up an apparently 'dud' smoke shell, which then detonated. He lived for some weeks, the injury having partly healed, though part of the face and skull were destroyed.*

There are some more specific matters in an explosion death which the pathologist must bear in mind. Firstly, trace evidence may be vital to the forensic scientists and bomb experts, who are trying to trace the maker of a terrorist device. X-rays should be taken of all bodies before autopsy to identify any radio-opaque objects. Not only will this assist in detecting lethal missiles not visible from the surface, but it may reveal small metal objects that form part of the bomb mechanism, such as small springs or contacts from the timer or detonator. These may be invaluable in allowing the experts to recognize the handiwork of a particular bomb-maker or terrorist group. Such radiography may also reveal the unexpected, which in the past has included a bullet, the victim having been blown up after death from shooting. Even prior to radiography, there may be much that the pathologist has to do. Though many explosion victims are relatively intact, where extensive disruption has occurred, the fragments have to be collected and sorted out in the mortuary.

A major initial problem is to discover how many bodies are represented and to try to allot the correct fragments to the right individuals. Where there are a number of victims and a large pile of small fragments, this task may be difficult or impossible, but it is naturally of paramount importance to determine how many victims are involved and to determine the sex of each. This is largely an anatomical exercise, akin to the sorting of multiple skeletal remains. Careful identification of all recognizable structures, such as prostate, uterus, breasts, scalp, eyes, for example, is needed in addition to more gross sorting of limb and trunk fragments. Radiology may again assist in matching contained skeletal structures, but much of

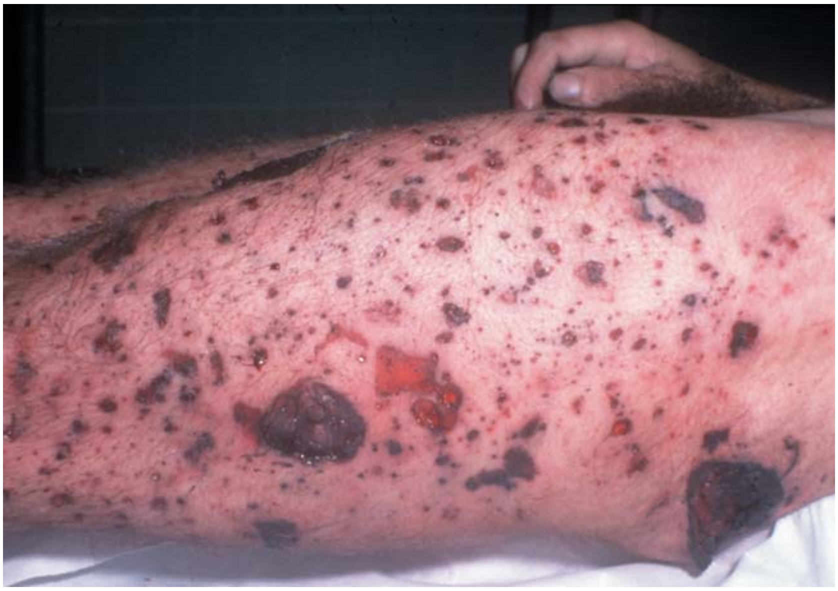

FIGURE 8.62 *Numerous wounds and peppering by debris projected in a terrorist bomb blast. Blast injuries occur only in close proximity to the device – most injury is caused by flying projectiles. (Reproduced by kind permission of Professor TK Marshall.)*

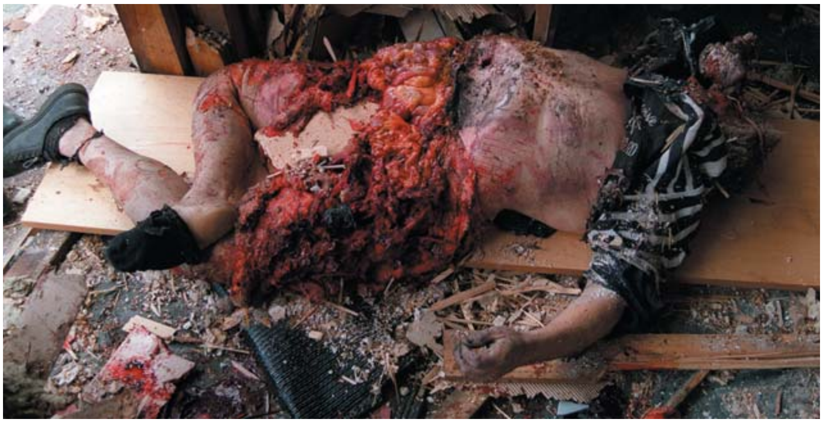

FIGURE 8.63 *Anterior aspect of a man who had constructed an explosive device using dynamite. The device probably detonated prematurely, the blast disrupting the head and abdominal region.*

the debris accumulated after a large explosion consists mainly of skin and attached tissues.

In spite of the most thorough police search of the scene, considerable proportions of some bodies are never recovered, having been disrupted into tiny fragments and mixed with the masonry and other debris of the bomb site. The collected fragments have to be washed clean from the inevitable dirt that coats them and sorted into groups by anatomical similarity. Non-human tissue is discarded, as animal material is not uncommonly admixed. The completed piles of tissue are then resorted by any apparent similarity, such as racial pigmentation, hair colour and sex. Limbs and large joints are then sorted, allotting them to left and right from anatomical considerations.

The difficulties of such a task was emphasized by Professor TK Marshall, whose experiences in Northern Ireland make his writings on the subject the definitive guide to such tragedies. In one of his incidents, eight victims were accounted for, when a single penile fragment forced the inevitable conclusion that there were in fact nine victims, not eight, though no other recognizable part of this ninth man was ever discovered. He had presumably been the bomb-carrier and had been completely disintegrated.

REFERENCES AND FURTHER READING

Adelson L. 1984. Bullet embolism with radiologic documentation. A case report. *Am J Forensic Med Pathol* **5**:253–6.

Aguilar JC. 1983. Shored gunshot wound of exit. A phenomenon with identity crisis. *Am J Forensic Med Pathol* **4**:199–204.

al Alousi LM. 1990. Automatic rifle injuries: suicide by eight bullets. Report of an unusual case and a literature review. *Am J Forensic Med Pathol* **11**:275–81.

Amatuzio JC, Coe JI. 1981. Homicide by exploder ammunition. *Am J Forensic Med Pathol* **2**:111–13.

Anonymous. 1968. Nail-gun injuries. *Br Med J* **1**:462.

Betz P, Pankratz H, Penning R, *et al.* 1993. Homicide with a captive bolt pistol. *Am J Forensic Med Pathol* **14**:54–7.

Bhoopat T. 1995. A case of internal beveling with an exit gunshot wound to the skull. *Forensic Sci Int* **71**:97–101.

Black AM, Burns BD, Zuckerman S. 1942. Wounding mechanism of high velocity missiles. In: Pugh WS (ed.), *War medicine*. F. Hubner & Co, New York.

Breitenecker R. 1969. Shotgun wound patterns. *Am J Clin Pathol* **52**:257–62.

Brissie RM, Collum ES. 1980. Shotgun wounds: multiple probes and shielding effects as adjuncts to determining position of the decreased at time of injury. *J Forensic Sci* **25**:528–32.

Callender GR. 1943. Wound ballistics; mechanism of production of wounds by small bullets and shell fragments. *War Med* **3**:337–45.

Clark MA, Micik W. 1984. Confusing wounds of entrance and exit with an unusual weapon. *Am J Forensic Med Pathol* **5**:75–8.

Clark MA, Smith TD, Fisher RS. 1981. Russian roulette with an exploding bullet. A case report. *Am J Forensic Med Pathol* **2**:167–9.

Coe JI. 1982. External beveling of entrance wounds by handguns. *Am J Forensic Med Pathol* **3**:215–19.

Coe JI, Austin N. 1992. The effects of various intermediate targets on dispersion of shotgun patterns. *Am J Forensic Med Pathol* **13**:281–3.

Conradi SE. 1982. New aluminum-jacketed ammunition: The case of the 'invisible' jacket. *Am J Forensic Med Pathol* **3**:153–5.

Copeland AR. 1985. Concepts in survival from lethal handgun wounds. *Am J Forensic Med Pathol* **6**:175–9.

Cragg J. 1967. Nail-gun fatality. *Br Med J* **4**:784–5.

Dada MA, Loftus IA, Rutherfoord GS. 1993. Shotgun pellet embolism to the brain. *Am J Forensic Med Pathol* **14**:58–60.

de la Grandmaison GL, Brion F, Durigon M. 2001. Frequency of bone lesions: an inadequate criterion for gunshot wound diagnosis in skeletal remains. *J Forensic Sci* **46**:593–5.

DiMaio VJ. 1981. Penetration and perforation of skin by bullets and missiles. A review of the literature. *Am J Forensic Med Pathol* **2**:107–10.

DiMaio VJ, Copeland AR, Besant Matthews PE, *et al.* 1982. Minimal velocities necessary for perforation of skin by air gun pellets and bullets. *J Forensic Sci* **27**:894–8.

Dixon DS. 1984. Pattern of intersecting fractures and direction of fire. *J Forensic Sci* **29**:651–4.

Eckert WG. 1981. Exploding bullets. A hazard to the victim, physician, and investigator [editorial]. *Am J Forensic Med Pathol* **2**:103–4.

Eisele JW, Reay DT, Cook A. 1981. Sites of suicidal gunshot wounds. *J Forensic Sci* **26**:480–5.

Elsayed NM. 1997. Toxicology of blast overpressure. *Toxicology* **121**:1–15.

Freytag E. 1963. Autopsy findings in head injuries from blunt forces: statistical evaluation of 1367 cases. *Arch Pathol* **75**:74–80.

Geurin PF. 1960. Shot-gun wounds. *J Forensic Sci* **5**:294–9.

Glattstein B, Zeichner A, Vinokurov A, *et al.* 2000. Improved method for shooting distance estimation. Part III. Bullet holes in cadavers. *J Forensic Sci* **45**:1243–9.

Goonetilleke UK. 1976. A stud (cartridge) gun suicide (a case report). *Med Sci Law* **16**:181–4.

Green GS, Good R. 1982. Homicide by use of a pellet gun. *Am J Forensic Med Pathol* **3**:361–5.

Guileyardo JM, Cooper RE, Porter BE, *et al.* 1992. Renal artery bullet embolism. *Am J Forensic Med Pathol* **13**:288–9.

Gulmann C, Hougen HP. 1999. Entrance, exit, and reentrance of one shot with a shotgun. *Am J Forensic Med Pathol* **20**:13–16.

Hawley DA, Pless JE, Palmer H. 1987. Tumbling abrasions. Injuries from ricocheting bullets. *Am J Forensic Med Pathol* **8**:229–32.

Hudson P. 1981. Multishot firearm suicide. Examination of 58 cases. *Am J Forensic Med Pathol* **2**:239–42.

Hunt AC, KonVanda M. 1962. The patterns of injury from humane killers. *Med Sci Law* **2**:197–202.

Jacob B, Huckenbeck W, Daldrup T, *et al.* 1990. Suicides by starter's pistols and air guns. *Am J Forensic Med Pathol* **11**:285–90.

James WR. 1962. Fatal air rifle pellet wound of the brain. *Med Sci Law* **2**:153–4.

Janssen W, Kulle KJ, Gehl A, *et al.* 2001. [Unintentional gunshot during police intervention.] *Arch Kriminol* **207**:1–11.

Janssen W, Miyaishi S, Koops E, *et al.* 1996. [Gunshot fatalities in connection with hunting and hunting rifles – causes, prevention and expert evaluation.] *Arch Kriminol* **197**:1–15.

Johnson GC. 1985. Unusual shotgun injury. Gas blowout of anterior head region. *Am J Forensic Med Pathol* **6**:244–7.

Jones AM, Graham NJ, Looney JR. 1983. Arterial embolism of a high-velocity rifle bullet after a hunting accident. Case report and literature review. *Am J Forensic Med Pathol* **4**:259–64.

Kage S, Kudo K, Kaizoji A, *et al.* 2001. A simple method for detection of gunshot residue particles from hands, hair, face, and clothing using scanning electron microscopy/wavelength dispersive X-ray (SEM/WDX). *J Forensic Sci* **46**:830–4.

Karger B, Stehmann B, Hohoff C, *et al.* 2001. Trajectory reconstruction from trace evidence on spent bullets. II. Are tissue deposits eliminated by subsequent impacts? *Int J Legal Med* **114**:343–5.

Kleiber M, Stiller D, Wiegand P. 2001. Assessment of shooting distance on the basis of bloodstain analysis and histological examinations. *Forensic Sci Int* **119**:260–2.

Knight B. 1982. Explosive bullets: a new hazard for doctors [editorial]. *Br Med J Clin Res Ed* **284**:768–9.

Lee KA, Opeskin K. 1995. Gunshot suicide with nasal entry. *Forensic Sci Int* **71**:25–31.

Madea B, Staak M. 1988. Determination of the sequence of gunshot wounds of the skull. *J Forensic Sci Soc* **28**:321–8.

Marshall T. 1988. A pathologist's view of terrorist violence. *Forensic Sci Int* **36**:57–67.

Marshall TK. 1976. Deaths from explosive devices. *Med Sci Law* **16**:235–9.

Marshall TK. 1978. Violence and civil disturbance. In: Mason JK (ed.), *The pathology of violent injury.* Edward Arnold, London.

Marshall TK. 1978. The investigation of bombings. In: Wecht C (ed.), *Legal medicine annual 1979*. Appleton Century Crofts, New York.

McCorkell SJ, Harley JD, Cummings D. 1986. Nail-gun injuries. Accident, homicide, or suicide? *Am J Forensic Med Pathol* **7**:192–5.

Menzies RC, Scroggie RJ, Labowitz DI. 1981. Characteristics of silenced firearms and their wounding effects. *J Forensic Sci* **26**:239–62.

Millar R, Rutherford WH, Johnson S, *et al.* 1975. Injuries caused by rubber bullets: a report on 90 patients. *Br J Surg* **62**:480–6.

Milroy CM, Clark JC, Carter N, *et al.* 1998. Air weapon fatalities. *J Clin Pathol* **51**:525–9.

Murphy GK. 1980. The study of gunshot wounds in surgical pathology. *Am J Forensic Med Pathol* **1**:123–30.

Oehmichen M, Meissner C, Konig HG. 2000. Brain injury after gunshot wounding: morphometric analysis of cell destruction caused by temporary cavitation. *J Neurotrauma* **17**:155–62.

Opeskin K, Cordner S. 1990. Nail-gun suicide [see comments]. *Am J Forensic Med Pathol* **11**:282–4.

Ornehult L, Eriksson A. 1987. Accidental firearm fatalities during hunting. *Am J Forensic Med Pathol* **8**:112–19.

Owen-Smith MS. 1981. *High velocity missile wounds*. Edward Arnold, London.

Petty CS. 1966. Firearms and the forensic pathologists. American Society of Clinical Pathologists. *Council on Forensic Pathology Meeting*. Commission on Continuing Medical Education, Washington.

Petty CS. 1969. Firearms injury research: the role of the practicing pathologist. *Am J Clin Pathol* **52**:227–81.

Petty CS, Hauser JE. 1968. Rifled shotgun slugs, wounding and forensic ballistics. *J Forensic Sci* **13**:114–23.

Pollak S, Ropohl D, Bohnert M. 1999. Pellet embolization to the right atrium following double shotgun injury. *Forensic Sci Int* **99**:61–9.

Quatrehomme G, Iscan MY. 1997. Bevelling in exit gunshot wounds in bones. *Forensic Sci Int* **89**:93–101.

Quatrehomme G, Iscan MY. 1998. Analysis of beveling in gunshot entrance wounds. *Forensic Sci Int* **93**:45–60.

Quatrehomme G, Iscan MY. 1998. Gunshot wounds to the skull: comparison of entries and exits. *Forensic Sci Int* **94**:141–6.

Rao VJ, May CL, DiMaio VJ. 1984. The behavior of the expanding point 25 ACP ammunition in the human body. *Am J Forensic Med Pathol* **5**:37–9.

Ratanaproeksa O, Hood I, Mirchandani H. 1988. Single gunshots with multiple entrances. *Am J Forensic Med Pathol* **9**:212–14.

Rocke L. 1983. Injuries caused by plastic bullets compared with those caused by rubber bullets. *Lancet* **1**:919–20.

Rosenhead J. 1985. Plastic bullets – a reasonable force? *New Scient* **17**:26–7.

Rothschild MA, Liesenfeld O. 1996. Is the exploding powder gas of the propellant from blank cartridges sterile? *Forensic Sci Int* **83**:1–13.

Shaw J. 1972. Pulmonary contusion in children caused by rubber bullets. *Br Med J* **4**:764–5.

Smith OC, Berryman HE, Lahren CH. 1987. Cranial fracture patterns and estimate of direction from low velocity gunshot wounds. *J Forensic Sci* **32**:1416–21.

Spitz WU, Wilhelm RM. 1970. Stud gun injuries. *J Forensic Med* **17**:5–11.

Stahling S, Karlsson T. 2000. A method for collection of gunshot residues from skin and other surfaces. *J Forensic Sci* **45**:1299–302.

Steindler R. 1983. The case of the mysterious dum-dum bullet. *Am J Forensic Med Pathol* **4**:205–6.

Steindler RA. 1982. An expanding .25 ACP bullet. *Am J Forensic Med Pathol* **3**:119–21.

Stephens BG. 1983. Back spatter of blood from gunshot wound observations and experimental simulation. *J Forensic Sci* **28**:437–9.

Stone IC. 1992. Characteristics of firearms and gunshot wounds as markers of suicide. *Am J Forensic Med Pathol* **13**:275–80.

Subke J, Haase S, Wehner HD, *et al.* 2002. Computer aided shot reconstructions by means of individualized animated three-dimensional victim models. *Forensic Sci Int* **125**:245–9.

Thali MJ, Kneubuehl BP, Dirnhofer R, *et al.* 2001. Body models in forensic ballistics: reconstruction of a gunshot injury to the chest by bullet fragmentation after shooting through a finger. *Forensic Sci Int* **123**:54–7.

Thali MJ, Kneubuehl BP, Zollinger U, *et al.* 2002. The 'skin-skull-brain model': a new instrument for the study of gunshot effects. *Forensic Sci Int* **125**:178–89.

Thali MJ, Kneubuehl BP, Zollinger U, *et al.* 2002. A study of the morphology of gunshot entrance wounds, in connection with their dynamic creation, utilizing the 'skin-skull-brain model'. *Forensic Sci Int* **125**:190–4.

Torre C, Varetto L, Ricchiardi P. 1986. New observations on cutaneous firearm wounds. *Am J Forensic Med Pathol* **7**:186–91.

Ulrich U, Zollinger U. 2001. [Development of rust stains on the skin due to contact with a gun.] *Arch Kriminol* **208**:32–41.

Warren MW, Falsetti AB, Kravchenko II, *et al.* 2002. Elemental analysis of bone: proton-induced x-ray emission testing in forensic cases. *Forensic Sci Int* **125**:37–41.

Zuckerman S. 1940. Experimental study of blast injuries to the lung. *Lancet* **2**:219–24.

CHAPTER 9

Transportation injuries

- The dynamics of vehicular injury 281
- Pattern of injury of vehicle occupants 282
- The effect of seatbelts 287
- The vulnerability of children in vehicles 289
- Injuries to motorcyclists 290
- Injuries to pedal cyclists 291
- Injuries to pedestrians 291
- Cause of death in traffic accidents 295
- Suicide and homicide by motor vehicle 297
- Railway injuries 297
- References and further reading 298

Injuries and fatalities occur in all forms of transportation but numerically road traffic accidents account for the great majority worldwide. In developed countries, they are the most common cause of death below the age of 50 years, and in young men this trend is even more marked. The pattern of injury, fatal and otherwise, varies considerably depending upon whether the victim is a vehicle occupant, a motorcyclist, a pedal cyclist or a pedestrian.

THE DYNAMICS OF VEHICULAR INJURY

A number of elementary physical facts help to explain the complex pattern of traffic injuries, especially those sustained by the occupants of a vehicle.

- Tissue injury is caused by a change of rate of movement. A constant speed, however rapid, has no effect whatsoever as is evident from space travel or the rotation of the earth. It is the change of rate that is traumatic – that is, acceleration or deceleration.
- Change of rate is conveniently measured in ‘gravities’ or ‘G forces’. The amount that a human body can tolerate depends greatly on the direction in which the force acts. Deceleration of the order of 300 G can be sustained without injury and even 2000 G can be survived for a short time, if it acts at right angles to the long axis of the body. The frontal bone may resist 800 G without fracture and the mandible 400 G, as can the thoracic cage.
- During acceleration or deceleration the tissue damage produced will depend upon the force applied per unit area, just as a sharp knife penetrates more easily than a blunt one used with the same force. If a car driver is brought to rest from 80 km/hour by striking 10 cm^2 of his head on the windscreen frame, the damage will be vastly more severe than if the same decelerative force was spread over 500 cm^2 of a safety belt.
- Between 60 and 80 per cent of vehicular crashes (either into a fixed structure or into another vehicle) are frontal, causing violent deceleration. Another 6 per cent are rear impacts, which accelerate the vehicle and its occupants. Of the remainder, about half are sideswipes and the rest ‘roll-overs’.
- In the common frontal impact, there is never instant arrest of the vehicle, even when it runs into a massive, immovable structure. The vehicle itself deforms from the front so that there is always a deceleration distance and time, albeit small. In fact, much of the manufacturers’ design research now goes into making deliberate provision for the crumpling or ‘concertinaing’ of the front and rear of the car, leaving a central rigid cell that comprises the passenger compartment. The object is to extend the stopping distance and time, so that the G value acting on the occupants is reduced.
- The value of the G forces can be calculated from the formula: $G = C(V^2)/D$, where V is velocity in km/hour, D is the stopping distance in metres after impact, and C is a constant 0.0039. (If V is in mph and D is in feet, C becomes 0.034.) For example, if a car

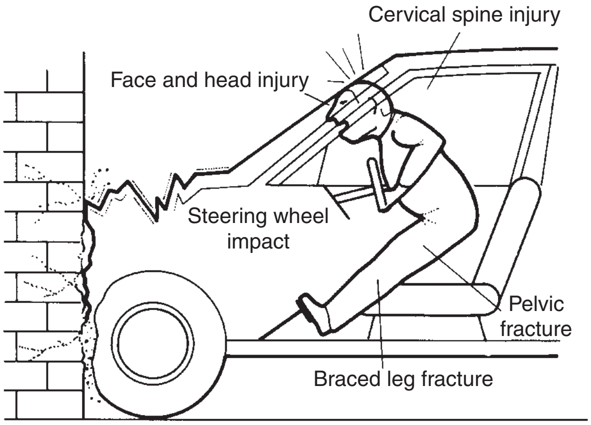

FIGURE 9.1 *Major points of injury to an unrestrained driver of a vehicle in deceleration impact.*

travelling at 80 km/hour runs into a stone wall that it penetrates for 25 cm, plus 50 cm crumpling of the front of the car, the deceleration would amount to about 33 G. If an occupant was rigidly belted into his seat (a practical impossibility), he would also suffer the same deceleration, which would be survivable. If, however, he was unrestrained, he would continue forwards momentarily at 80 km/hour and suffer massive G forces, the magnitude of which would depend on his deformation stopping distance (a few centimetres of tissue compression) when he struck the internal car structures in front of him.

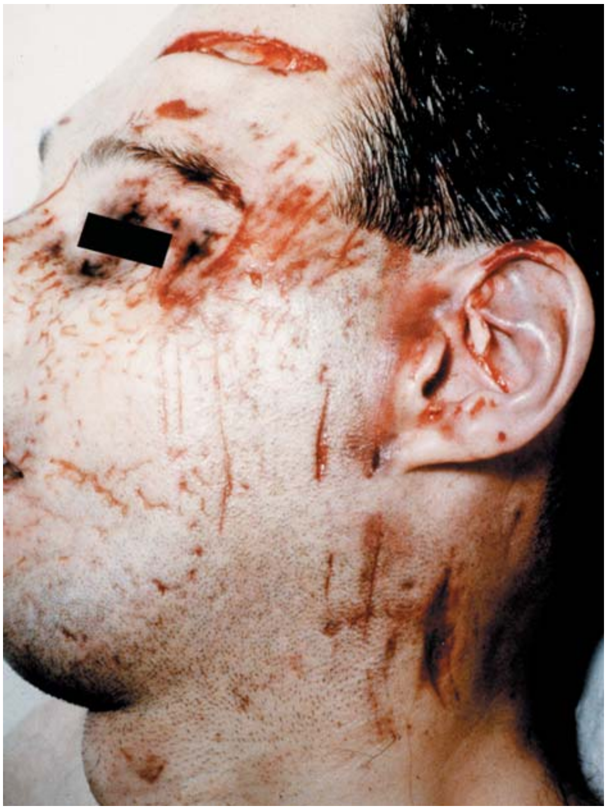

FIGURE 9.2 *Facial lacerations from a shattered windscreen in an unrestrained driver. The toughened glass breaks into small fragments, which produce the characteristic 'sparrow-foot' marks. The laceration on the forehead was made by the windscreen rim.*

PATTERN OF INJURY OF VEHICLE OCCUPANTS

The type of vehicle (other than motorcycles) in theory makes little difference to the mechanism of injury, but most statistical surveys divide them into cars and light vans under 1.5 tonnes, on the one hand and heavier vehicles, such as trucks and buses, on the other, though the latter have different features more akin to passenger aircraft.

Heavy goods vehicles naturally suffer less than cars and light vans in crashes because of their far greater mass and strength, and also due to their height above the ground. Structural damage from impact with other smaller vehicles is less and often sustained below the level of the driver. Given smaller deceleration forces, however, the cab occupants are vulnerable to the same injury patterns.

Light vans are virtually identical to cars with respect to the front-seat occupants. In fact they may be more at risk, as modern vans tend to be flat-fronted and thus have little or no 'crumple' potential to increase the stopping time. Concentrating on cars, the most common vehicular casualty, the pattern of injury varies according to the position of the occupant.

The driver

Numerous investigations have been made by road research organizations and car manufacturers using dummies and actual corpses, together with sophisticated recording equipment and high-speed cinematography. These have established a detailed picture of the sequence of events in automobile crashes. When the most common event – frontal impact – occurs, the unrestrained driver first slides forwards so that his legs strike the fascia/parcel-shelf area, and his abdomen or lower chest contacts the lower edge of the steering wheel. The body then flexes across the steering wheel and begins to rise. The heavy head goes forwards, and there is flexion of the cervical and thoracic spines. The upward and forward component causes the head to strike the windscreen, the upper windscreen rim or the side pillar. The windscreen is often perforated by the head or face, and the whole body may be ejected through the broken glass, to land on the bonnet or even on the roadway ahead.

Another factor causing injury is the intrusion of structural parts into the passenger compartment. Though modern cars are designed to maintain a rigid central passenger compartment, if the impact is gross, the engine or front-wheel assembly may be forced back into the seating area, intruding upon the driver. Similarly, the roof or front corner pillar (the so-called ‘A’-frame) may cave in on top of the driver.

One effect of column, engine, or gearbox intrusion may be to force the floor up and backwards against the driver’s feet and legs. The control pedals also take part in intrusion, and, in the usual desperate braking and declutching, the reflex pressure of feet on rising pedals and floor may cause transmitted force up the legs and into the pelvic girdle. The steering column was formerly a more dangerous item for intrusion, being forced back to ‘stab’ or crush the driver’s chest or abdomen. Modern design has reduced this danger by making the column telescopic, hinged or otherwise collapsible, but injuries still occur – sometimes from the wheel itself breaking and penetrating the chest. Additionally, the door may burst open and the driver, if unrestrained, ejected sideways onto the road, especially in a crash that has a rollover component.

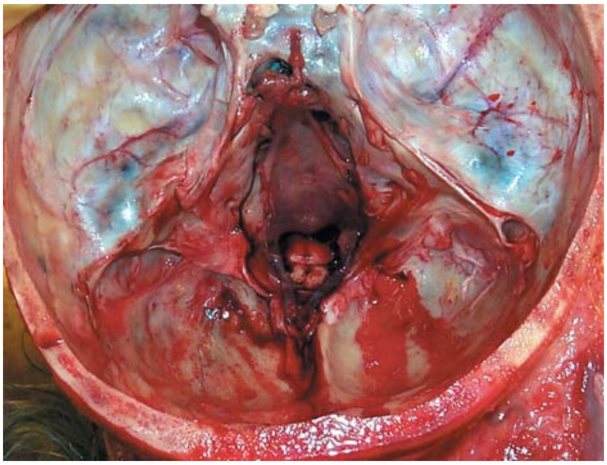

FIGURE 9.3 *Ring fracture around the foramen magnum caused by an impact on the crown of the head in a car driver, who lost the control of his vehicle and crashed into a tree.*

In a rear impact, the driver is violently accelerated and, if no rigid head restraint is fitted to the seat, severe hyperextension of the neck occurs, often followed by the sequence of deceleration events when the car is cannoned into the vehicle or other obstruction in front, causing the popular, if inaccurate name of ‘whiplash’.

In side impacts, the injuries depend upon the amount of intrusion of the driver’s door and side panels. Restraint devices can offer no protection, though modern vehicles usually have strengthened side-impact bars built within the doors.

This range of traumatic events can produce the following lesions in drivers not wearing seatbelts or protected by airbags:

- Impact against the fascia can cause abrasions, lacerations and fractures of the legs around knee or upper shin level.
- Pressure of feet on the floor, especially when it is intruded by the engine, can cause fractures anywhere from foot to femur. The leg can also be injured by violent contact with the fascia or dashboard and the hip joint may be dislocated posteriorly. Not uncommonly the pelvis is fractured, often at one or both sacroiliac joints. In Mant’s (1978) series of 100 driver fatalities there were 22 pelvic injuries and 31 of the lower limb.
- Impact of the abdomen and chest against the steering wheel may cause severe internal injuries, usually rupture of the liver (50 per cent) and, less often, spleen (36 per cent). There may be bruising of the skin surface, but this is often absent even in the presence of

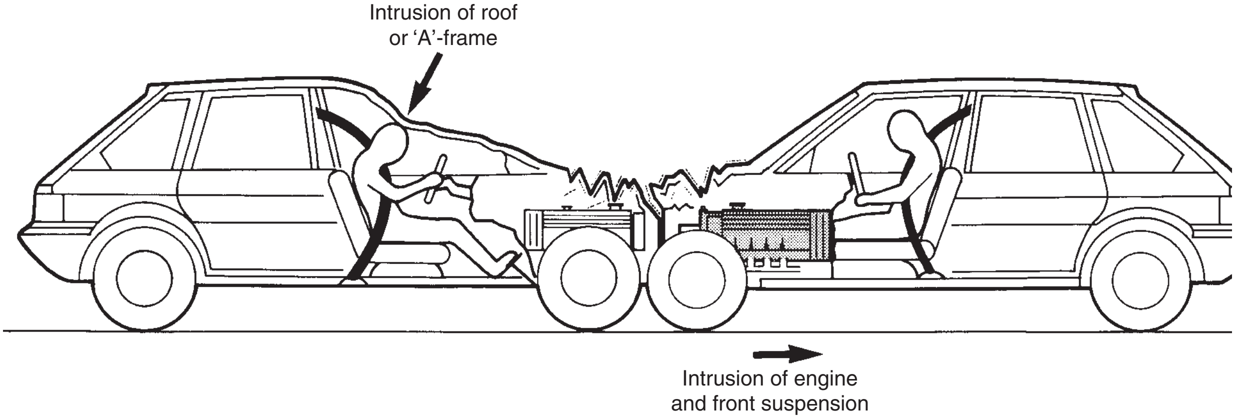

FIGURE 9.4 *When vehicle structures impinge on the occupants even belt restraints offer little protection. The engine, front suspension, roof and ‘A’ frame are frequent intruders.*

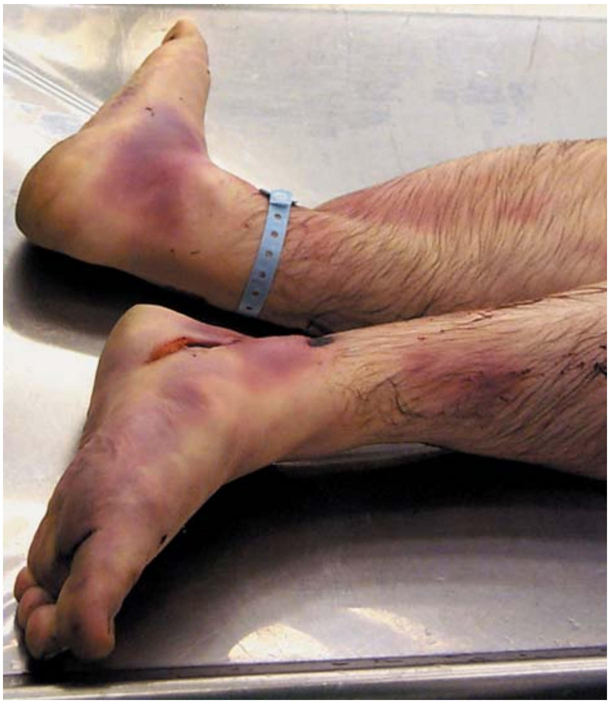

FIGURE 9.5 *Bruising, laceration and bilateral leg fractures of a car driver in a frontal impact.*

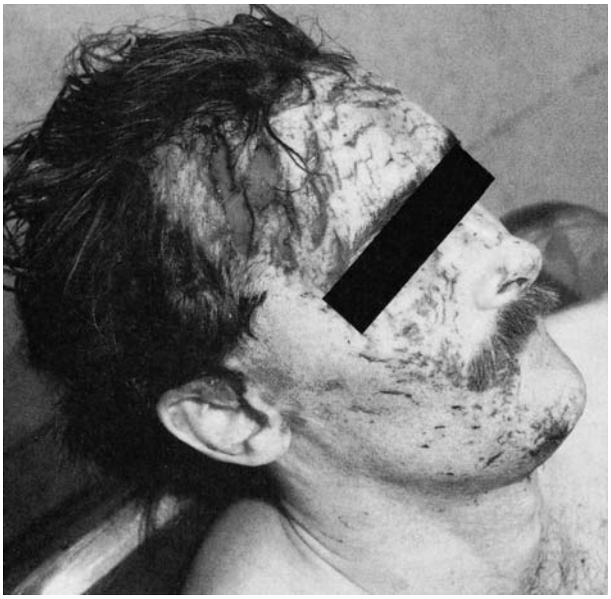

FIGURE 9.6 *Facial injuries in a car driver unrestrained by a seatbelt. Following a deceleration impact his face struck the windscreen, causing the typical small cuts from broken safety glass and lacerations of the temple from striking the windscreen rim or 'A' frame.*

severe internal injuries. Laceration of the skin is rare unless the steering wheel snaps and penetrates the trunk. Other steering-wheel lesions include bruising of the lungs, fractured ribs and sternum, cardiac contusion and haemothorax or pneumothorax or both. Almost 70 per cent of Mant's series had broken ribs.

- Upper limb injuries are less common but may occur from transmitted force through gripping the steering wheel or from impact against the windscreen, pillars, intrusive roof, bonnet or ground when held up in a reflex protective position. Only 19 per cent of Mant's series had arm injuries.
- The most obvious injuries are often those to the face and head as a result of projection against and ejection through, the windscreen. The unrestrained driver rises and flexes forwards so that his forehead and skull are likely to contact the upper rim of the windscreen, leading to lacerations. The face frequently suffers multiple cuts from contact with the shattered safety glass. In most European vehicles the glass is of the toughened, not laminated, variety and, when broken, it shatters into small cubes with relatively blunt edges. These still cause superficial lacerations, often in short 'V-shaped' or 'sparrow-foot' patterns. In themselves they are not a danger to life, but indicate an impact sufficient to hurl the driver on or through the glass. Damage to the eyes is common.
- The impact against the windscreen rim or corner pillar – or after ejection – can cause any type or degree of head injury, including scalp laceration, fractured skull, intracranial haemorrhage or brain damage. In Mant's series there were 42 skull fractures in 100 drivers. This was less than in the front-seat passengers, a figure at variance with Eckert's (1959) series of 300 in the USA, where drivers suffered twice as many head injuries as the passenger, though it is not stated how many accidents were to vehicles occupied only by the driver.
- Hyperflexion of the cervical spine when the head swings can cause fractures or dislocation. There is often a double component in that the hyperflexion of deceleration is followed by a rebound hyperextension when the head strikes an obstruction in front. Rear impacts also cause the double 'whiplash' effect, as already mentioned.

One injury that is frequently overlooked at autopsy is the atlanto-occipital dislocation, which Mant found in a third of his series. Other fractures can occur anywhere in the cervical spine, often at about C5–6. Seatbelt restraint cannot prevent cervical spine damage, though a rigid head restraint can reduce injuries resulting from hyperextension. The thoracic spine is less often damaged, but in unrestrained drivers the same 'whiplash' effect can fracture or dislocate the upper dorsal spine, often around T5–6–7.

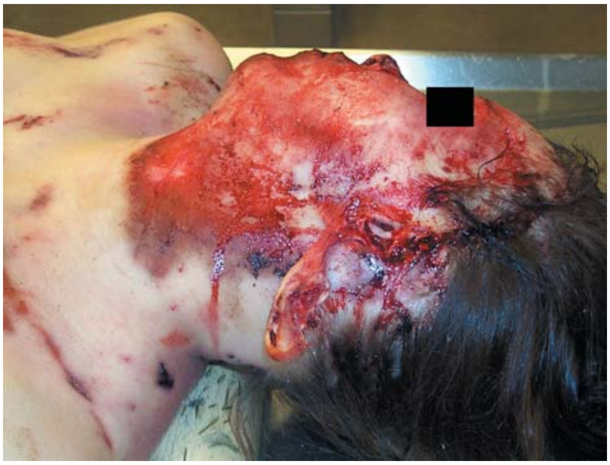

FIGURE 9.7 *Mixed injuries in a restrained car driver from a head-on collision. Death was caused by ruptured aorta.*

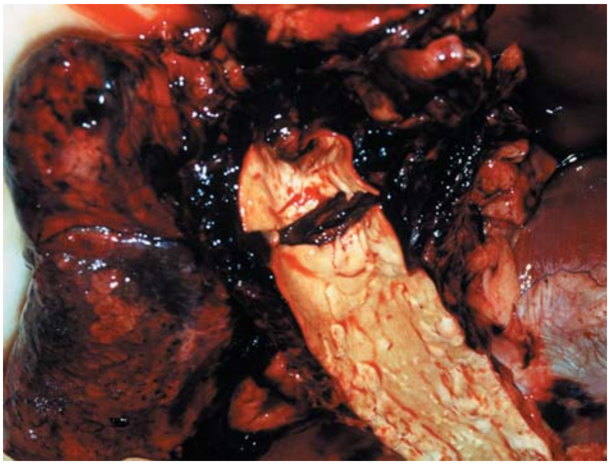

FIGURE 9.8 *Ruptured aorta in which car occupant, unrestrained by seatbelt, suffered severe deceleration. Aorta has torn in the usual place, the distal arch where the curve of the vessel meets the thoracic spine.*

- A more common thoracic injury associated with deceleration is the ruptured aorta. It may be associated with a severe whiplash effect on the thoracic spine, as the aorta is tethered to the anterior surface of the vertebrae where the distal arch joins the straight descending segment. Probably the most common reason for aortic rupture, however, is the 'pendulum' effect of the heart within the relatively pliable thoracic contents. When the thorax is violently decelerated, the heavy cardiac mass attempts to keep moving ahead and may literally pull itself off its basal mountings, the most rigid part of which is the aorta. Separation takes place at the point where the aorta is attached to the spine at the termination of the arch.

  The appearance of the aortic rupture is often of a clean-cut circular break, almost as sharp as if it had been transected with a scalpel. Sometimes there are additional transverse intimal tears adjacent to the main rupture, the so-called 'ladder tears', as they can resemble the rungs of a ladder. These may be present when no actual rupture has occurred and may be found as an incidental finding at autopsy. Sometimes they are deep enough to allow a local dissection of blood to seep into the intima, when death has not been virtually instantaneous. Rarely, a major dissection may lead to delayed death some hours or even days later. Ruptured aorta is a common lesion in traffic accidents – in a two-car crash, the author (BK) has seen three transected aortas among the four fatalities.

  The frequency of such tears is common enough for a warning always to be offered to the autopsy prosector not to use undue force on the neck and thoracic structures when removing the organ pluck from the body. Rough handling during this stage can produce artefactual ladder tears in the aorta.
- Other chest injuries can be caused by impact with the steering wheel, ejection through the windscreen or impact with the road. There may be bruising or laceration on the chest from the steering wheel, though padding, collapsible columns, less fragile wheels, airbags and seatbelts have reduced the incidence of this formerly common lesion. Beneath the skin, sternal and rib fractures are common, though fatal visceral injuries can occur without rib fractures in young people because their ribs are more pliable.
- The heart may be damaged even in the absence of external marks or thoracic cage fractures. Bruising of the epicardium and underlying myocardium is not uncommon and the posterior surface may be damaged from impact against the spine. In high-speed impacts, the heart may be completely avulsed from its base and be found lying loose in the chest. Less severe degrees of damage may lacerate the ventricles or atria, and cause gross haemorrhage. Coronary artery thrombosis has been described following contusion over a coronary artery. Penetrating injuries from sternum, ribs or external objects may lacerate the heart directly. Subendocardial haemorrhages on the left side of the interventricular septum and opposing papillary muscles are not a sign of impact, but an index of catastrophic hypotension. They are also seen in head injuries; they can occur within the space of a few beats, as the author (BK) has seen these prominent lesions in an avulsed heart after a military aircraft crash.
- The lungs are frequently injured, either from stabbing by fractured ribs penetrating the pleura or from blunt

impact. The latter often leads to a line of bruising down the posterior part of the lung where it lies in the paravertebral gutter. There may be air bullae or blood blisters under the pleura overlying the bruised areas and a pneumothorax or haemothorax may result. The interior of the lung may be pulped even in the presence of an intact visceral pleura, from transmitted force or massive variations in intrathoracic pressure during the impact. The lung often shows areas of bleeding under the pleura, which may be from direct contusion, from aspiration of blood from other damaged areas of lung or from blood sucked down the air passages from injuries in the nose or mouth.

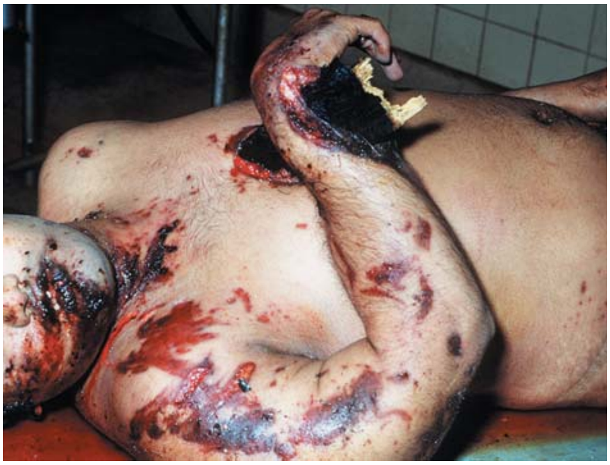

FIGURE 9.9 *Penetration of the wrist and chest by a wooden component of a bus seat. The victim drove his car at speed into a bus and part of the resulting debris penetrated his left ventricle.*

- The major abdominal injury is a ruptured liver, which may be damaged in any part. A common lesion is central tearing of the upper surface, which may extend deeply and even transect the organ. Less serious damage is often seen in the form of shallow, sometimes multiple, parallel tears on the upper surface of the right lobe. Subcapsular tears can occur with the formation of a subcapsular haematoma, which can rupture later. The spleen also shows shallow tears in some accidents, often around the hilum; in rare cases, it may be avulsed from the pedicle. The mesentery and omentum often show bruising and, rarely, there is laceration and fenestration sufficient to cause a lethal haemorrhage.
- Ejection injuries are common, and lethal in both driver and passengers. This is particularly likely to happen in roll-over accidents. Much research has been pursued by manufacturers to develop anti-burst door locks, which have improved safety. Where there has been considerable distortion of the vehicle frame, however, nothing can prevent the doors from opening or even being torn off. It has been shown by Moore and Tourin's study (1954) at Cornell that ejection injuries followed steering-column lesions as the second most frequent type of trauma and, if a victim was ejected, there was a fivefold greater chance of dying than if he was retained in the vehicle. Moore and Tourin found that when doors burst open a third of the car occupants were ejected.

  Almost any kind of injury, usually multiple, may be sustained after ejection, either from contact with the road surface or (in a significant proportion) from being struck by other vehicles, especially on motorways.

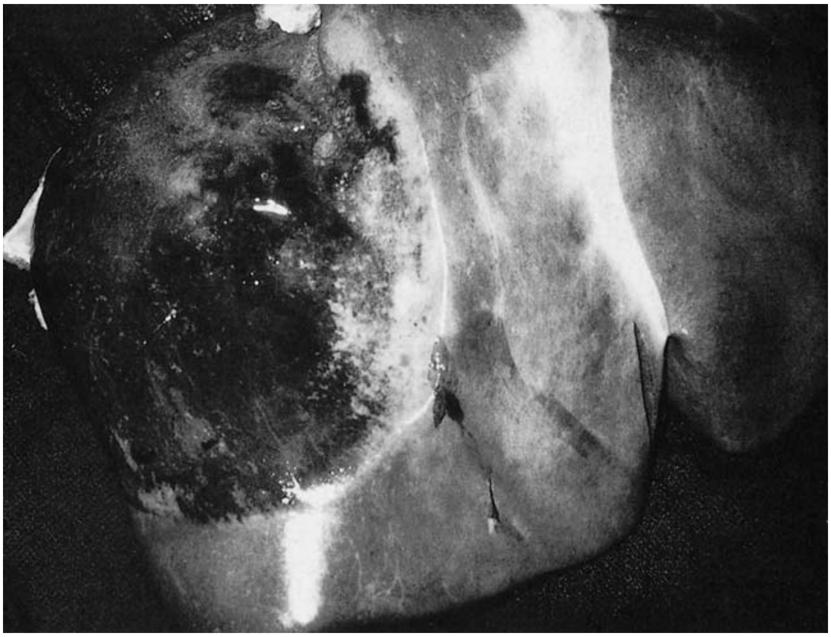

FIGURE 9.10 *Large subcapsular haemorrhage of the liver in a driver who struck the rim of the steering wheel during severe deceleration. Such subcapsular lesions can remain intact for hours or even days, then rupture into the abdominal cavity.*

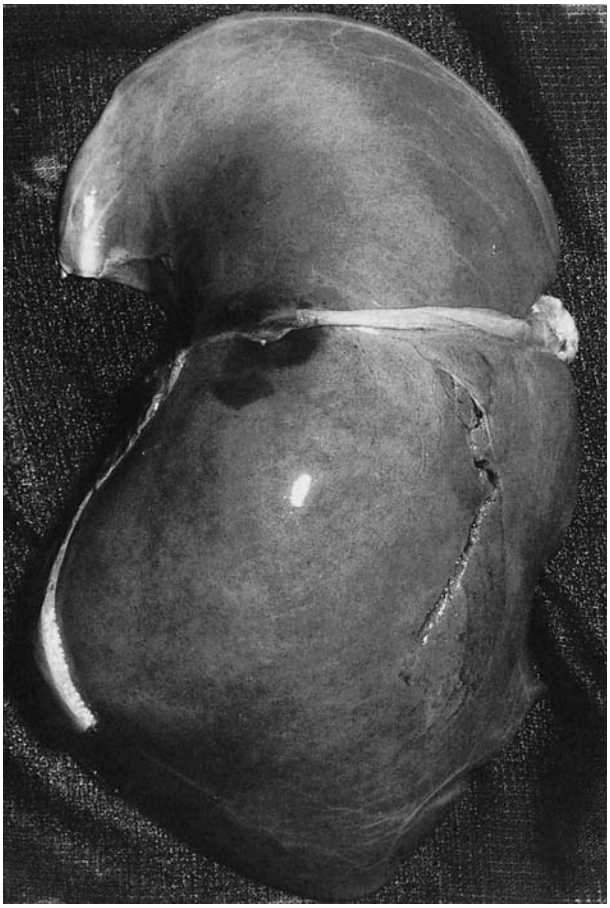

FIGURE 9.11 *Anterior tear through the full thickness of the liver; there is also a haemorrhage around the suspensory ligament. The victim was a car driver who was impacted against the steering wheel in a frontal crash.*

The front-seat passenger

In Western countries, far more drivers than passengers are killed or injured, but this reflects the fact that a high proportion of cars contain only a driver – a third in Mant's series. In countries with a lower ratio of vehicles to population the converse is true and indeed many accidents are due to gross overcrowding of passenger vehicles. The pattern of injuries is similar to that of the driver, but this position in the car is even more dangerous, as indicated by the title of a safety film made by the French Michelin Tyre Company – *La place du mort*.

Though there is no steering wheel to impact into the chest, its absence also denies the slight protection offered to the driver in reducing the collision with the windscreen, perhaps by giving him something to brace against. Another factor may be that the driver gives his attention constantly to the road and so has momentary warning of an impending crash, compared with the passenger who may be oblivious of imminent disaster and fail to 'brace up' ready for the impact. This may explain the greater number of skull fractures and brain damage in unrestrained passengers in Mant's series, the percentages being 55 per cent and 42 per cent, compared with 64 per cent and 53 per cent, respectively. However, these figures were compiled before seatbelts and airbags became commonplace or even mandatory.

Rear-seat occupants

Before the more widespread use of front seatbelts – now mandatory in many countries – it was thought that the rear-seat position was fairly safe, as indeed it is compared with the front seats.

When the widespread use of seatbelts dramatically reduced the deaths and serious injuries amongst front-seat occupants, the dangers of the back seats became much more obvious. One series showed that 49 per cent of rear-seat passengers in car crashes sustained serious or moderately serious injuries. Campaigns and legislation similar to the previous crusade were waged on behalf of the rear-seat occupants, and the use of these has now become mandatory in Britain and some other countries.

During violent deceleration, unrestrained occupants in the rear are projected forwards and strike the back of the front seats, including head-rests where fitted. They may be thrown over the seats, striking and adding further injuries to the front-seat occupants and may even be ejected through the windscreen, which is broken by them or by the people in front.

In roll-over accidents, they share in the general trauma of being churned inside the passenger compartment, when multiple injuries can occur from contact with fitments, such as mirrors, door handles and window winders. Design changes have reduced these hazards by making handles smoother or countersunk and mirrors that easily snap off their mountings. Ejection is another common cause of death and serious injury in rear-seat occupants, a wide range of head, chest and limb injuries being seen.

THE EFFECT OF SEATBELTS

As stated, many countries now have legislation making the wearing of front and rear seatbelts mandatory. Where no laws exist, persuasion seems to have little effect, in spite of the fact that it is uncontested that their use reduces deaths and serious injury by a factor of 20–25 per cent, as seen in the Australian experience in Victoria and New South Wales. A similar reduction was attained in Britain after the introduction of mandatory laws. Not only did the death rate drop substantially, but facial injuries and especially eye damage was dramatically reduced.

Seatbelts are now almost all of the lap-strap and shoulder diagonal type, the so-called 'three-point attachment belt'. The simple lap-strap is now fitted only in aircraft, where it is really only of token effectiveness. Most car belts are now of the 'inertia-reel' type, which allow slow movement but jam at a sudden tug. The advantage, apart from the comfort, is that they automatically tighten up around the body, as a slack belt is not only less effective but can actually constitute a danger. More complex restraints, such as double shoulder harness and crotch strap, are fitted only in light aircraft, gliders and racing cars. Though far more effective, their use would be socially unacceptable in ordinary road vehicles, as would head-band restraints, which are almost the only way of preventing hyperflexion damage to the cervical spine.

The various forms of strap restraints act by:

- Holding the occupant back against the seat, so that forward projection against the steering wheel, windscreen and corner 'A' frame is prevented. The head, though still subject to hyperflexion, is prevented from smashing through the glass and the body cannot be projected through the screen onto the bonnet or roadway. The belt cannot cope with backward intrusion of the engine, floor, roof or corner pillar if those structures reach the occupant sitting in the original seat position. The effectiveness of the belt is also dependent on the secure fixation of the seat to the vehicle floor.
- The belt restrains the occupants within the vehicle in the event of a door bursting, as ejection greatly increases the risk of death or serious injury. A belt is relatively ineffective in a side impact, except in that it reduces injuries from ejection. It has been reported, however, that head injuries were fewer in restrained victims of side-swipes, though the reason is not clear.

  It has been claimed (mainly in insurance disputes) that seatbelts can worsen this sort of injury by holding the occupant in more dangerous proximity to the intruding impact. There would seem few occasions, however, in which even the unrestrained occupant could voluntarily escape from any significant degree of sudden lateral intrusion from a speeding vehicle.
- Extending the deceleration time and distance by substantial stretching of the belt fabric, which may lengthen by many centimetres during a violent arrest. To be effective the belt must be held tightly against the body to get the maximum restraint, either by adjusting the buckle or using an inertia reel. The belt should never be used again, as it cannot stretch further and may break on the next application of tension.
- Spreading the area of application of deceleration forces. As stated earlier, the body has to absorb whatever G forces are applied to it, as calculated by the $G = C(V^2)/D$ equation. If these are absorbed by a focal impact covering a few square centimetres over the skull, fatal injury may well ensue. The same deceleration diffused against the thorax and abdomen by 500 cm^2 of belt may leave no injury or merely surface bruising.

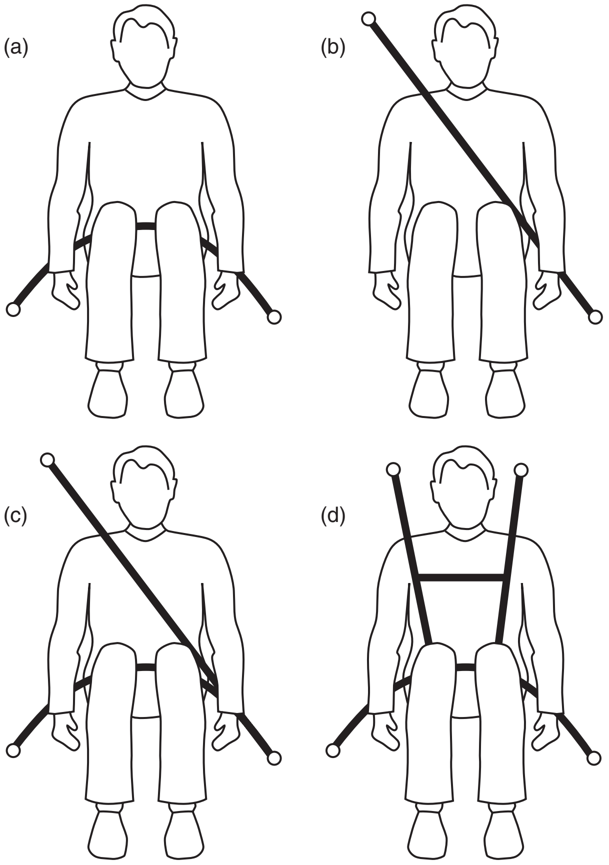

FIGURE 9.12 *Types of seatbelt restraint: (a) simple lap-strap (dangerous to aorta), (b) diagonal only (can slip underneath), (c) diagonal plus lap-strap (usual car type), and (d) shoulder harness (used in aircraft and racing cars).*

Seatbelt injuries

There is no doubt that seatbelts can themselves cause injury, occasionally of considerable severity. This was one of the arguments of the vociferous anti-seatbelt lobby that opposed legislation in several countries, but it is largely illogical, because an impact that causes seatbelt injuries would almost invariably have caused even worse injuries or death if no seatbelt had been worn.

Another invalid criticism of restraints was that they impede escape from a burning vehicle. Fires involving passenger car crashes are, however, quite rare. A report by Bako *et al.* (1970) from Canada showed that, of 1297 vehicle fatalities, only 24 were from burns, a mere three in passenger cars.

Incorrectly adjusted or positioned straps such as a twisted belt, which reduces the area of contact, can increase the danger of injury. Loose straps allow the body to move relative to the belt before sudden restraint occurs, decreasing the distance between the passenger and facing structures.

Where the person is too small for the harness, such as a child or a small woman, the body may slide from under the strap – the so-called 'submarining' – or it may act as a garrotte around the neck. Some women find that the diagonal strap compresses the breast even in normal use, so that the greatly increased tension during deceleration is likely to injure the gland.

Pregnant women also have problems with belts but, although uterine and fetal injuries have been recorded in accidents, the incidence is relatively low and without the belt the consequences would have been as bad or probably worse.

Injuries from seatbelts can vary from the trivial to the fatal. Bruising is the most common and may be seen either under the diagonal or the transverse component of the three-point belt. It is more common with the single aircraft-type lap-strap because of the smaller area of pressure. Bruising may occur in the abdominal or chest wall, but the dangerous lesions are visceral. The abdominal contents suffer most, especially from single lap-straps. Rupture of the mesentery, or the small or large intestine, usually occurs from acute flexion over a lap-strap. The full bladder can be ruptured as can the caecum. The abdominal aorta can be crushed and the lumbar spine suffer a compression fracture or be dislocated through a disc in the midlumbar region. The posterior arch, pedicles or transverse processes may also be damaged.

The diagonal strap usually prevents serious abdominal injury as it prevents hyperflexion, but it may contribute to thoracic injury. Bruising of the skin and underlying muscles and fractures may accompany a broken clavicle or sternum where the belt crosses them. Several authors (listed at the end of this section) have made detailed studies of seatbelt injuries.

Airbags

In recent years, the use of airbags has spread from only the most expensive vehicles to many standard production cars. The device consists of a large fabric bag, which is normally folded into the steering-wheel hub in the case of the driver position and into the fascia in front of the front-seat passenger.

A sensitive deceleration device triggers the ignition of sodium azide, a solid and highly toxic explosive propellant, which is converted in milliseconds to nitrogen gas. The deploying airbag can reach speeds up to over 300 km/h (>200 mph). Deflation is also rapid, so that residual car control and escape from the vehicle is not impeded.

The inflated bag is designed to interpose itself between the occupant and the frontal structures of the passenger cabin, to cushion the impact and prevent forceful contact and hyperflexion.

In addition to the life-saving capabilities, airbags can also cause serious and even fatal injuries either by the airbag itself or the module cover overlying it in the steering wheel or in the fascia on the passenger side. Injuries can be inflected at any stage of the deployment process, depending on the posture and possible objects between the occupant and the deploying airbag or the module cover. They vary from facial bruising, partial or complete amputation of fingers to dislocated and fractured arms or cervical spine and fatal head injuries. Eye injuries are common and range from mild corneal abrasions and chemical burns from contact with unburned sodium azide or the alkaline byproducts of combustion, which are released into the passenger compartment, to globe rupture from blunt trauma or perforation by interposed objects.

THE VULNERABILITY OF CHILDREN IN VEHICLES

This is a particular problem with several aspects. Many countries have brought in laws to protect children in cars because of their special vulnerability.

First, it is an understandable, but highly dangerous, indulgence for adults to allow small children to travel unrestrained in the front seat of a car. Some parents even allow them to stand gripping the fascia edge immediately below the windscreen, an invitation to facial and eye damage even in minor accidents. The seating of a child on the mother's lap is hazardous, as on violent deceleration either mother and child pitch against the windscreen – or the child flies out of the arms of a belt-restrained mother. The close proximity to the fascia and windscreen causes many deaths and facial injuries, especially to the eyes.

Adult-secured seatbelts will not properly accommodate a child (or even a small adult), because the fixation point on the door pillar is too high even if the straps can be shortened sufficiently to be made tight. The diagonal may pass across the throat and, unless special drop-plates are fitted to the door pillars, the restraint is often worse than useless.

Though some European countries already forbid children under 14 years to travel in front seats, the medical profession in Britain campaigned for similar legislation, as they had done for motorcyclists to wear helmets many years before.

As discussed above, a similar campaign has been successfully waged for rear-seat restraints for children and adults. Banished from the front, children were thought to be safe in the back of the car, but many have died and far more have been injured from being projected against the seat backs, front passengers and internal fitments. Special seats secured on top of the regular seating are required and, for a baby, the cot must be lashed down with equally robust restraints. In 1988, legislation was introduced into the British Parliament to make compulsory the restraint of children in rear seats: from 1991, this law also became applicable to adults.

INJURIES TO MOTORCYCLISTS

Though there are fewer motorcycles than four-wheeled vehicles, especially in developed countries, the rate of injury and death amongst motorcyclists is far higher than among car drivers. For example, in England and Wales in 1989 there were more deaths amongst male motorcyclists between the ages of 16 and 24 than with comparable car drivers (343 compared with 323) even though the ratio of motorcycles to cars is very small in Britain. In the 25–44 age group, 192 motorcyclists died compared with 381 car drivers.

The two extremities of the body suffer most in motorcycle accidents, though Larsen and Hardt-Madsen's analysis in Denmark in 1988 also showed high injury rates for chest and abdomen.

- Because the rider inevitably falls to the ground, head injuries are common and often severe, causing 80 per cent of deaths according to Bothwell (1962). Though crash helmets are mandatory in most countries, the severity of the impact often defeats the protective effect of the helmet.

  Impact with the road surface or another vehicle at speed causes skull fractures at any part of the head, but often temporoparietal. A common complication is a basal skull fracture, especially a 'hinge' fracture. This transverse crack across the floor of the skull, crossing the petrous base or behind the greater wing of the sphenoid bones through the pituitary fossa to the opposite side, has also been called 'the motorcyclist's fracture'.

  Another type is the ring fracture around the foramen magnum in the posterior fossa caused by an impact on the crown of the head. The neck suffers quite often and Mant found cervical spine fractures in over a quarter of his series. Brain damage may be severe, even with a helmet in place. Cortical contusion and laceration, sometimes contrecoup, may be gross enough to cause brain tissue to extrude through compound fractures of the skull. In Mant's series of motorcyclists, 60 per cent had skull fractures and almost 80 per cent had brain damage.
- The legs are often injured, either by primary impact with another vehicle or fixed road structures, or by becoming trapped by part of the motorcycle frame. Lacerations, friction burns and fractures – often compound – are common. Mant recorded leg or pelvic fractures in 55 per cent of his cases.
- Any part of the body may suffer injury, but less often than the extremities. Falling from the machine, especially at speed, can cause rib fractures and visceral damage, especially rupture of the liver and spleen.
- An injury common with motorcycles is the 'tail-gating' accident, where a rider drives into the back of a truck so that the machine passes underneath, but the head of the motorcyclist impacts upon the tail-board. Decapitation may occur in the most extreme cases, but severe head and neck injuries are almost inevitable.

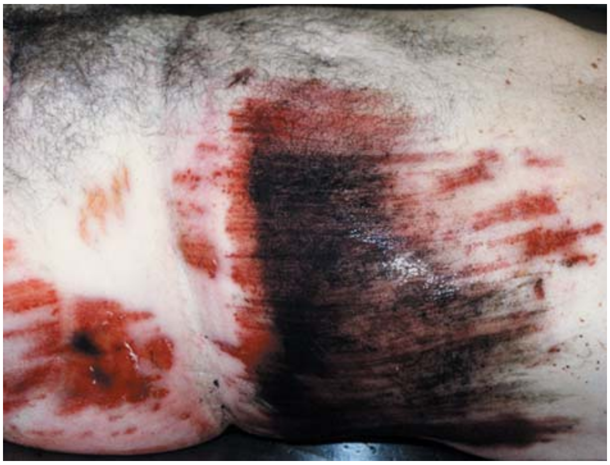

FIGURE 9.13 *Severe brush abrasions or 'friction burns' in a motorcyclist, who came off his machine and skidded across the road surface.*

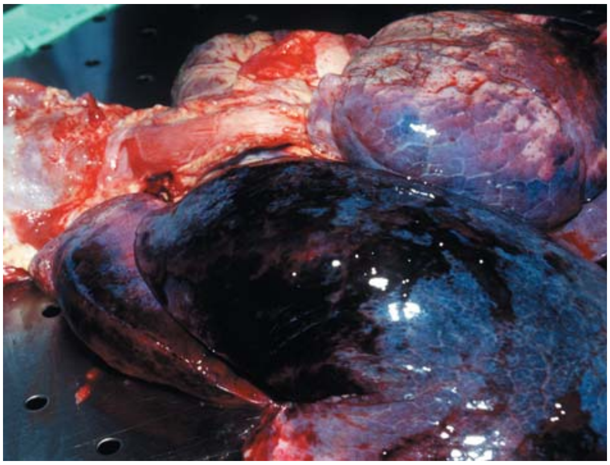

FIGURE 9.14 *Pulmonary contusions in a motorcyclist, who collided with a truck and was hurled into a signpost. Death was caused by a cranial fracture and multiple injuries of the internal organs.*

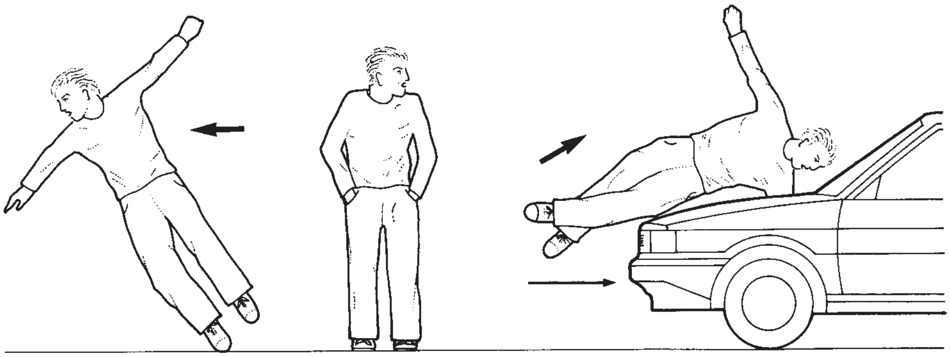

FIGURE 9.15 *A pedestrian struck by the front of a car may be projected forwards or lifted onto the vehicle.*

Trucks in many countries must now have strong bars fitted at the rear to prevent this 'tail-gating', which may also happen to motor cars, the rear of the truck smashing into the windscreen and driver.

Safety helmets act both by providing a rigid barrier against impact, which depends partly on the shock absorbing padding within the helmet and by providing a smooth surface, which is designed to skid across the road surface, thus lengthening the stopping distance and time to reduce the G force of deceleration. Their strength is designed to be finite to control the rate of deceleration but in massive impacts seen at high speeds the helmet may be penetrated or the head and brain damaged by the transmission of blunt force. Crash-bars are another safety measure on motorcycles, being fitted in front of the engine to project on each side and protect the legs if the machine falls over. Unless extremely strong, however, such bars can themselves trap the legs if they bend backwards on impact.

INJURIES TO PEDAL CYCLISTS

These form a less severe counterpart of motorcycle lesions, as the pedal cycle has the same instability but far lower speeds. Once again, head injuries figure largely in accidents, as the height above the ground is considerable and the rider suffers from the passive fall, added to by any forward motion or projection from impact by a motor vehicle. Helmets are now worn by many cyclists and naturally afford considerable protection.

Other injuries are from the primary impact from a striking vehicle, which may hit the rider around thigh, hip or chest level. Secondary damage to the shoulder, chest and arm may occur from striking the ground, when friction grazes are common. (A unique injury, though not fatal, was entrapment of the leg between wheel spokes with compression of the soft tissues of the calf, when the leg penetrated the wheel.)

INJURIES TO PEDESTRIANS

Worldwide these are easily the most common road fatalities, probably accounting for more than 50 per cent of the third of a million road deaths each year. In the densely populated areas of the globe where vehicles are greatly outnumbered by people, such as Southern Asia, parts of Africa, and the Middle East and Central America, pedestrian casualties form a significant part of the total mortality.

Most pedestrians are struck by motor cars or trucks, and the type of vehicle makes a difference to the dynamics of the impact, which – unlike injuries to vehicle occupants – is an **acceleration** not a **deceleration** process.

Primary injuries are caused by the first impact of the vehicle on the victim, while secondary injuries are caused by subsequent contact with the ground. Some writers also use the term 'tertiary injuries' to describe the impact with the ground, reserving 'secondary' for additional contact with the vehicle, as when the pedestrian is hurled up against the windscreen. The usual sequence of events is as follows:

- The height of the car bumper bar ('fender') is well below the centre of gravity of the adult pedestrian, which lies in the abdominal region. Thus the first impact tends to knock the legs from under the victim and rotate them towards the oncoming vehicle. Depending on the profile of the front of the car, the struck pedestrian is either thrown forwards in the direction of travel if the bonnet-front is high and blunt – or scooped up onto the bonnet top, as with many slope-fronted modern vehicles.
- If thrown forward, secondary injuries will be suffered as a result of striking the ground, as well as the primary impact on legs and often the hips. If the car speed is appreciable (anything over 20 km/hour is sufficient), the body can be thrown into the air or knocked down flat with a severe impact. The secondary injuries may fracture the skull, ribs, pelvis, arm or thigh.

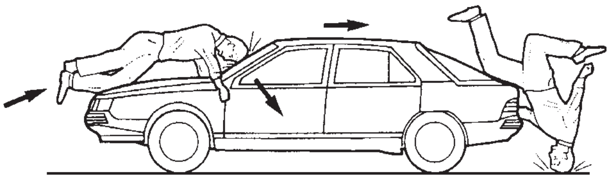

FIGURE 9.16 *At speeds of over 23 km/hour (15 mph) a pedestrian can be 'scooped up' onto a car, suffering head injuries against the windscreen. He may then fall off sideways or – at higher speeds – be thrown over the roof.*

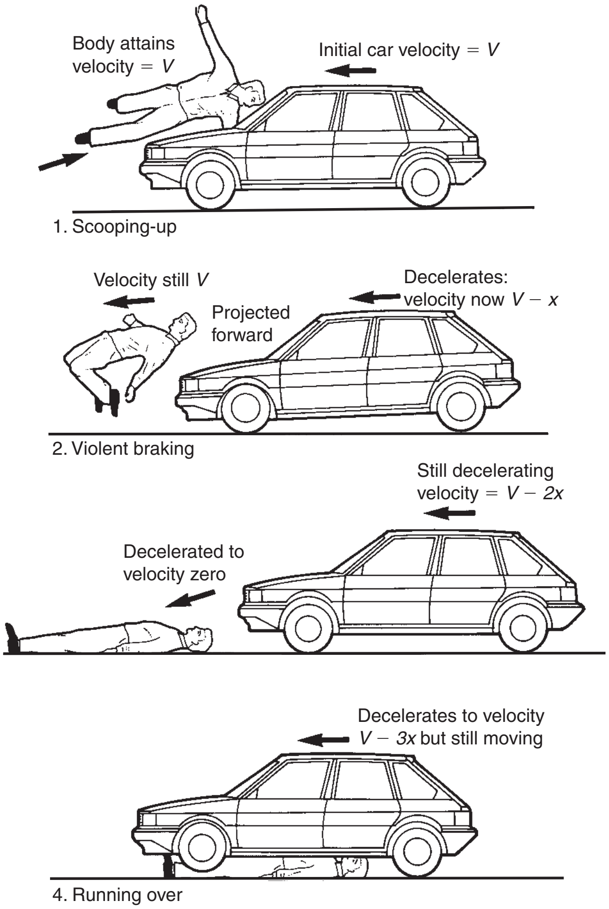
Body attains velocity = V
Initial car velocity = V
1. Scooping-up
Velocity still V
Projected forward
Decelerates: velocity now V − x
2. Violent braking
Still decelerating velocity = V − 2x
Decelerated to velocity zero
Decelerates to velocity V − 3x but still moving
4. Running over

FIGURE 9.17 *Sequence of events when a carried pedestrian is projected forwards during braking.*

A further hazard is being run over by the vehicle if the victim is projected directly in front. Sometimes he may be dragged by the under-belly of the car, and seriously soiled and injured, perhaps appearing at the rear if the vehicle does not stop quickly. Many impacts are on the front corner of the car and the pedestrian may then be knocked diagonally out of the path of the car. If thrown into the centre of the roadway, the person can be run down by a different vehicle overtaking in another lane or by one coming in the opposite direction on a single carriageway.

- If scooped up, the victim will land on either the bonnet or against the windscreen or corner-supporting pillar (the 'A' frame). The flat bonnet usually does relatively little damage, though linear abrasions, brush grazes, or friction burns may be seen. Violent contact with the windscreen, especially the rim or side pillars, is the most frequent cause of severe head injury from primary impact.

  Scooping-up can occur at speeds as low as 23 km/hour (about 15 mph; below 19 km/hour the body will usually be projected forwards). If the speed is high, the victim can be thrown up onto the car roof, sometimes somersaulting so that the head strikes the roof. He can then slide or be flung right over the back of the car, landing behind it in the roadway. This is more likely to happen if the car does not brake, but literally drives from under the body.
- In most cases, the scooped pedestrian falls or is flung off on one side of the car or the other, again to suffer secondary injuries in the road and perhaps be run over by another vehicle. The usual pattern of events is that, at the instant of contact – or even slightly before – the driver will apply the brakes violently. The scooped-up victim will acquire the speed of the car by the time he lands on the bonnet, but then the vehicle decelerates. As the adhesion to the shiny surface is small, the newly acquired velocity of the body will cause it to slide off the front of the car as the latter brakes. The victim then hits the ground in front of the car, sustaining secondary injury – and may even be run over during the residual motion of the vehicle before it finally stops.
- In a high-speed impact, which may be anything over 50 km/hour (31 mph), the body can be flung high in the air and for a considerable distance, either to the side or in the path of the car – or even backwards over the roof. In general, the severity of the injuries – both primary and secondary – will be the more severe the higher the speed.

It is impossible to estimate the speed of impact from the nature of the injuries. These can be fatal even at slow speeds of the order of 10 km/hour (6 mph), yet occasionally high-speed impacts can produce only minor damage. In Ashton's (1975, 1978) series half the deaths occurred at speeds less than 48 km/hour (30 mph).

In child victims, although the general pattern of injuries is similar, their shorter height and smaller weight affects the mechanics of impact. The primary contact is higher up their body, so they tend to be hit forwards rather than rotated upwards, though many do become scooped up onto the bonnet. Children tend to be projected further by impact and may be hurled in

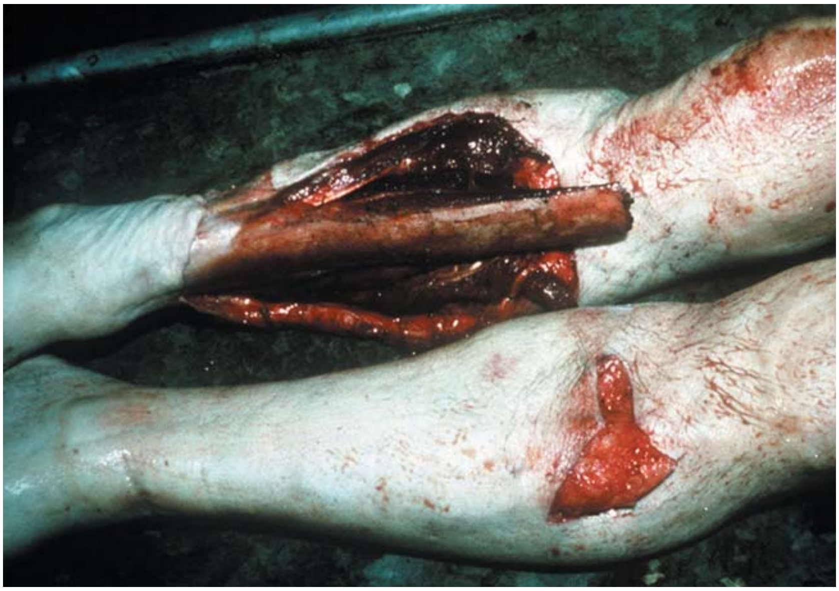

Figure 9.18 *Primary injury to a pedestrian struck by a car. There is damage to both legs at about the same level, with a compound fracture of the right leg. These are sometimes called 'bumper fractures'. The height of the injuries above heel level should always be measured at autopsy.*

the air at lower speeds than with adults. They are also more prone to be run over by reversing vehicles, especially trucks, as they often play between parked vehicles and – being small – are less visible to the driver. Recent safety measures on trucks include audible reversing warnings that are automatically linked to the gear-shift lever.

- When a pedestrian is struck by a larger vehicle, such as a van, truck or bus, the initial point of impact is higher and may cause primary damage to pelvis, abdomen, shoulder-girdle, arm or head. Because of the profile of these vehicles, there is no scooping-up effect, and the victim is usually projected forwards to suffer secondary damage from road contact and sometimes to be run over.
- The nature of pedestrian injuries reflects these dynamic effects.

The most common trauma is to the legs, some 85 per cent of pedestrian casualties having lower limb injuries. Abrasions and lacerations to the upper shin and knee area are typical of car bumper contact, and fractures of the tibia and fibula, often compound, are so common that they are present in a quarter of fatalities, according to Eckert. The femur is fractured less often, but is no rarity. The midshaft may be broken or the head may be driven into the acetabulum, together with a fractured pelvis. In children, because of their small stature, the femur may be fractured by the low bumper bar. At autopsy, the skin of the lower legs should be incised to seek deep bruising, as the clothing often protects the surface from obvious marking.

When a bumper (fender) strikes a leg, the tibia is often fractured in a wedge-shaped manner; the base of the wedge indicates the direction of the impact (often from behind), the front of the wedge pointing away from the side of contact.

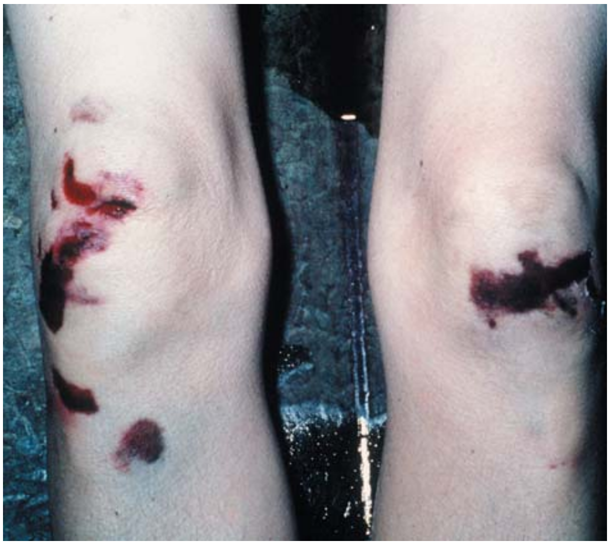

Figure 9.19 *Abrasions and laceration at knee level in a pedestrian struck by a car bumper. Their anterior position indicates that the victim was facing the vehicle at the moment of contact, but this may be a last-instant turn during final awareness of the car's proximity.*

If the leg is weight-bearing at the time of the impact, the tibial fracture tends to be oblique, whereas if not stressed, as when being lifted during walking, the fracture line is often transverse. When both shins are damaged, the level may be different on each side; this indicates that the person was moving at the time, with one leg raised in walking or running. Sometimes the level of injury appears too low for the normal bumper height of most cars, but this may indicate that the vehicle was braking violently at the moment of impact, going down on its suspension as the front wheels decelerated or locked, unless dip compensators were fitted.

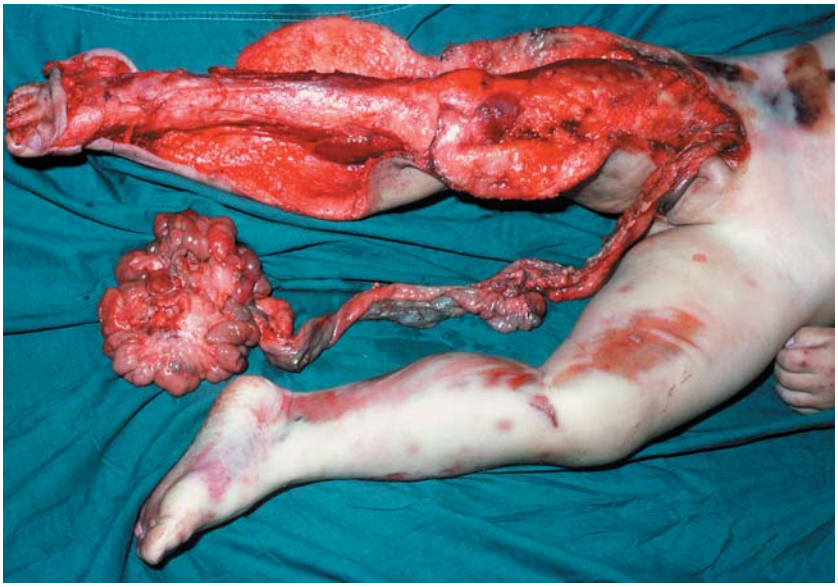

FIGURE 9.20 *Injuries caused by being run over by a bus. The large rotating wheel has 'flayed' the right leg, stripping the skin and subcutaneous tissues from the muscles. The passage of the adjacent double wheel over the abdomen has extruded the intestines through the perineum.*

Because of impact with the windscreen, pillars and roof, together with secondary contact with the ground, the head is the next most frequently damaged region – and the one which leads to most causes of death. Any type of injury may be sustained, as described in Chapter 5. Traffic accidents are the most frequent cause of skull fracture, especially of the base. Fractures of chest, arm and pelvis and injuries to the abdomen follow in frequency. Often the injuries are concentrated on one side, usually on the opposite side to the point of primary impact, because the body was thrown down onto the road. Because of rotation and the variable posture from being thrown off the car structure, however, the injuries are often widespread and may show no particular pattern.

Soft tissue injuries are common and, apart from abrasions, bruises and lacerations, muscle laceration and crushing can occur. A characteristic lesion from running-over, as opposed to knocking-down, is the 'flaying' injury, where a rotating motor wheel tears the skin and muscle from a limb or head. The rotatory effect against a fixed limb may strip off almost all tissue down to the bone. When a wheel passes over the abdomen or pelvis, multiple parallel striae or shallow lacerations may occur near the contact area because of ripping tension in the skin.

When a wheel passes over the pelvis, abdomen or head, there may be great internal damage with little surface injury. The weight of a large vehicle can virtually flatten a head, crushing the cranial vault. Often the brain is extruded through scalp lacerations, as may be the intestine through an abdominal wound. The pelvis may flatten out when run over, the symphysis or superior rami breaking, and one or both sacroiliac joints becoming detached. Any type of intra-abdominal injury may occur from ruptured liver and spleen to perforated intestine, lacerated mesentery and fractured lumbar spine. In the chest, ribs, sternum and thoracic spine may fracture, and heart and lung damage occur from crushing or laceration from jagged ribs. A 'flail chest' is sometimes produced when a heavy wheel runs across the supine body, breaking all the ribs on each side in the anterior axillary line.

Patterned injuries may be important, in that they can assist the police in identifying a vehicle in a 'hit-and-run' accident. The most common is a tyre pattern outlined in intradermal bruising and these should be measured carefully and photographed. These marks are usually caused by the skin being forced into the grooves of the tyre tread, the edge of the raised rubber tracing out the pattern. The elevated parts do not leave bruises, but may imprint dirt on the skin.

Paint fragments and glass shards are also trace evidence that must be carefully retained, as the forensic laboratory may be able to identify the make and model of vehicle involved, and match the fragments when a suspect car is examined.

Parts of the vehicle may leave patterned imprints on the skin, such as headlamps, mirrors or other components. Safety regulations for manufacturers have now almost eliminated the dangerous devices such as bonnet mascots, projecting door handles and non-flexible mirrors that used to adorn older cars. Metallic and plastic objects may still be found in the tissues from time to time, however; the author (BK) has recovered a door handle from the interior of the liver and a chromium bonnet insignia from a cerebral hemisphere. Any such artefacts must be preserved for the police in cases in which the identity of the vehicle is not known.

In Britain, where traffic drives on the left side of the road, it is more common for pedestrians to be struck on the right side as they walk off the pavement. Many, however, are struck facing the vehicle, from turning during last-minute awareness, and some are hit from behind.

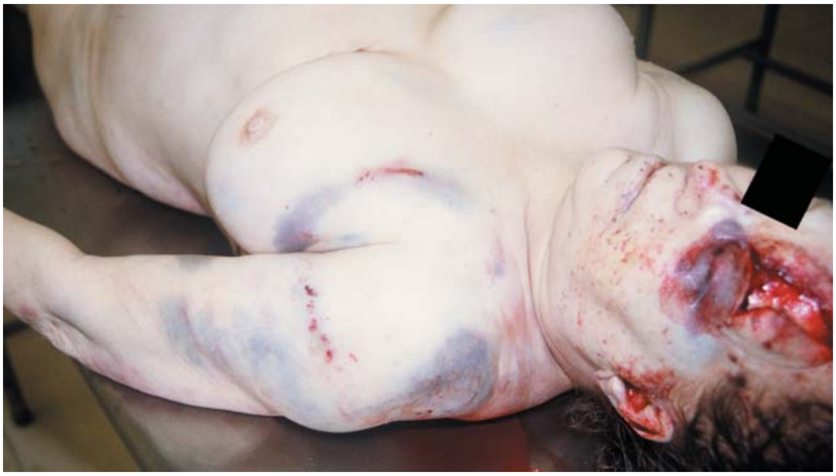

FIGURE 9.21 *A patterned injury in a pedestrian struck by a truck. The circular bruise and abrasion over the front of the shoulder are from a headlamp rim. There is also a well-demarcated imprint around the left eye from some other projection on the vehicle.*

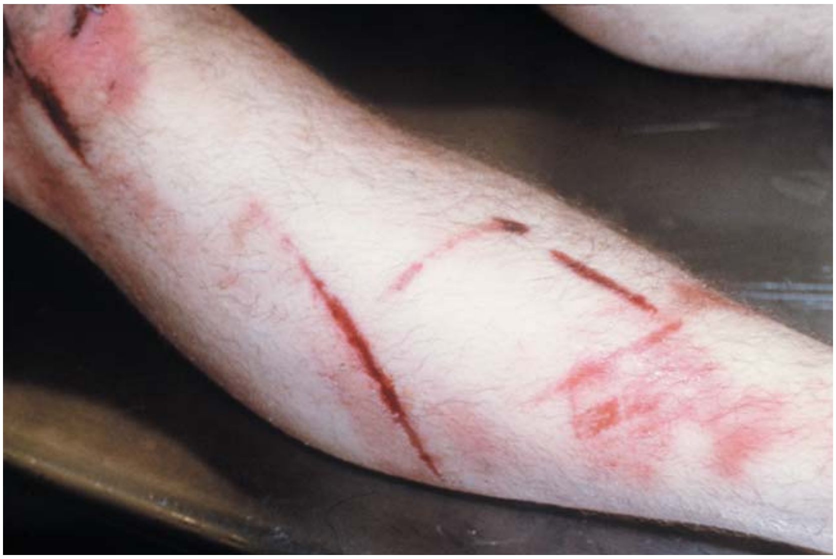

FIGURE 9.22 *Primary pedestrian injury. The pattern is caused by the bumper bar striking the leg. The height from the heel should always be measured for the police to compare with the vehicle – though during extreme braking, many vehicles will dip down at the front, thus lowering their bumper height.*

These factors will all affect the distribution of injuries, but it is too hazardous to try to reconstruct unwitnessed events in any detail from a study of the injuries. Suprarenal haemorrhage is more common in the right gland than the left after a traffic injury in Britain (Johnson, unpublished communication), but this observation must be modified by the fact that many such haemorrhages occur several days after the trauma and are usually the result of general systemic effects, rather than direct impact.

CAUSE OF DEATH IN TRAFFIC ACCIDENTS

In gross injuries this is often obvious, as in the crushed head with extrusion of brain or rupture of the aorta. Often multiple injuries make it difficult to decide which was the most serious and mortal lesion, but in such cases it is quite acceptable to use the term 'multiple injuries', preferably listing several of the most lethal.

When death occurs on the road or soon afterwards, there is usually macroscopic evidence of gross musculoskeletal or organ damage, severe haemorrhage, blockage of air passages from blood, or traumatic asphyxia from fixation of the chest caused by crushing some part of a vehicle.

Delayed death can be caused by continuing bleeding, secondary haemorrhage, renal failure from hypotension and/or extensive muscle damage, fat embolism, local infection, chest or other systemic infections, myocardial or cerebral infarction and other sequelae discussed in Chapter 13.

The presence of natural disease is always an important consideration in all transportation deaths, as a possible cause or contribution to the accident. In pedestrians, a sudden collapse

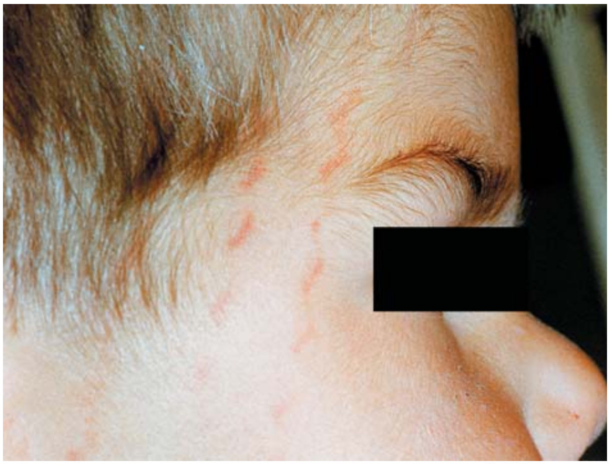

FIGURE 9.23 *Intradermal bruising on the forehead of a live boy who was found wandering in the street with amnesia for recent events. He had been struck by a 'hit-and-run' car, which left identifiable tyre marks on the skin. An astute casualty officer had a photograph taken as the marks faded in a few hours.*

in the roadway may lead to fatal injuries when a vehicle comes along – or even a dead body being run over by the next car.

Similarly, defects in sight or hearing may have contributed to the accident, though this is almost never detectable at autopsy, unless there are gross corneal or lens abnormalities. Of course, the possibility of drug or alcohol intoxication in the pedestrian victim must always be considered.

Where drivers or pilots are concerned – or even shipmasters – the presence of disease or intoxication may be a vital element, with both civil and criminal connotations possible.

Generally speaking, sudden natural disease does not often cause road vehicles to go out of control, as there seems to be a sufficient warning in most instances of cardiac or cerebral disablement, to allow the driver to pull over and stop.

Even the regular medical checks on older drivers legally enforced by some countries are no warranty that sudden collapse will not occur the very next day. The same applies to air pilots, as the VC-10 crash at London Airport proved some years ago, when the hypertensive pilot with coronary artery disease collapsed on take-off after an altercation with his co-pilot.

Schmidt *et al.* (1990) analysed 39 deaths at the wheel in Germany and found that 97 per cent had cardiovascular disease and 90 per cent had coronary heart disease. This was confirmed in a more recent retrospective study covering a 15-year time period in Munich, Germany, ranging from 1982 until 1996. Ischaemic heart disease was the underlying cause of death in 113 (84%) of the 147 natural deaths at the wheel. Morild (1994) found that 14 of 133 traffic deaths in Norway had died of natural disease, again predominantly coronary atherosclerosis.

The autopsy on a road traffic death

The autopsy is in general identical to the usual procedure, but with special attention to the following points:

- As criminal proceedings against a driver may follow, legal matters such as identity of the body and continuity of evidence must be assured, as discussed in Chapter 1.
- The body should be seen clothed, if brought dead to the mortuary or hospital, so that injuries can be matched against soiling and damage to the garments. Often this is not possible, especially if temporary survival allowed admission to a hospital or accident department, but where practicable the clothing should be preserved and examined by the pathologist. In any event, the clothes should be retained by the police for submission to the forensic science laboratory, usually when criminal proceedings are likely.
- Blood samples must be retained for blood grouping and now perhaps even 'DNA fingerprinting' in case a 'hit-and-run' vehicle is found with blood or tissue traces upon it. Sometimes hair samples may be required for the same purpose. Where death occurs within 12 or even 24 hours of the time of the accident, blood analysis for alcohol is essential, whether in the driver or pedestrian (see Chapter 28). Where possible, screening for drugs of dependence and common medicinal substances that might have caused drowsiness should be carried out. In combination with alcohol, even low levels of sedative, hypnotic and antihistamine drugs may be relevant in the causation of an accident. In certain cases, where leakage of carbon monoxide is suspected, the circumstances will suggest analysis of the blood for carboxyhaemoglobin concentration.
- The external examination, as in all trauma deaths, is vital and should be detailed, accurate and fully recorded. The height of major or patterned injuries above heel level must be noted, in order to compare these against dimensions of a vehicle. Patterned injuries must be photographed with a scale in view. Any foreign bodies or particles, either in the clothing, hair, on the skin or in the wounds, must be carefully retained for forensic science examination, especially in a 'hit-and-run' accident, where the identity of the vehicle may be vital.

All types of trace evidence may be found by a pathologist, from paint flakes and glass debris (which may be traced to a certain make, age, type and even individual vehicle) to parts of the vehicle structure. In past years, the author (BK) has retrieved a Renault door handle from inside a liver and an Austin bonnet insignia from within a brain.

- A full autopsy must be carried out, not merely a catalogue of injuries. The presence of any natural disease is relevant, especially if it might have contributed to the accident, either by causing a driver to lose control or ability to drive, or a pedestrian to exercise proper caution or behaviour in the roadway. Old and recent cardiac and cerebral lesions are particularly important, as is any evidence of a fit, such as a bitten tongue, or old meningeal adhesions over cortical damage. It is almost impossible to assess visual acuity at autopsy, but obvious lesions such as lens opacities must be noted. Similarly, it is virtually impossible to give any opinion on acuity of hearing from autopsy findings unless there is a gross neurological abnormality in the auditory tract.

SUICIDE AND HOMICIDE BY MOTOR VEHICLE

There is little that the pathologist can contribute to the elucidation of motivation in traffic accidents, as it is circumstantial and sometimes forensic-laboratory evidence that is more likely to reveal a non-accidental cause. Homicidal traffic deaths are rare, though the author (BK) has been involved in one incident where racial hatred led to the running-down of youths of one ethnic group – and another where a man repeatedly crashed the near-side of his own car in an effort to kill his passenger (his wife). There are no specific pathological features that can assist, except that the incident is likely to occur at relatively high speed, without braking effects. The victims of homicide by other means – or persons rendered unconscious first – may be deliberately placed in motor vehicles which are then crashed, preferably with a subsequent fire. The author (BK) dealt with one such case where a husband placed the body of his strangled wife in his car and secretly pushed it over the edge of a mountain road. Unfortunately (for him), he left the ignition key in the 'off' position. This faking of a vehicular 'accident' to conceal homicide is by no means unknown, the author (BK) being involved in another case where a police officer disposed of the body of his wife who died in suspicious circumstances, by crashing her car and smashing the windscreen with a hammer to add to the effect.

In such circumstances, the pathologist's role is to match the injuries with a traffic accident, to detect any which are atypical. For instance, focal depressed fractures of the skull of the type caused by a weapon are unusual in a car occupant unless there was a localized intrusion of the vehicle roof. The ante-mortem nature of the injuries should be demonstrated, though this is not always possible. When there is a fire, some evidence of ante-mortem burns, soot inhalation or carbon monoxide absorption should be sought, though – as mentioned in Chapter 11 – some flash petrol fires may kill before any monoxide is absorbed. Where the victim was unconscious though not dead, no such differentiation is possible. Full analysis for alcohol and stupefying drugs must be made if there is any suspicion.

Attempts at deliberate self-destruction by the use of a motor vehicle are said to be not uncommon (Selzer and Payne 1962), though this is difficult to prove in most cases. Once again, the evidence is more likely to be based on circumstantial rather than medical evidence – a matter for the investigating authorities rather than the pathologist. Driving at speed into the path of an oncoming truck or into a solid obstruction at the roadside are the methods employed, though it is hard to prove in the absence of definite evidence. It has been said that the imprint of the accelerator pedal on the undersurface of the shoe may be an indication, though this can never be the only indication. A witnessed lack of other causative factors may offer some corroboration, but this is police business, not pathology.

RAILWAY INJURIES

These are not uncommon, especially in countries with many 'level crossings' (called 'grade crossings' in the USA), where a public road crosses a railway track with either no barrier at all or with only a flimsy lifting pole. Many vehicles are struck each year by passing locomotives.

Few rail passengers are killed or injured in moving trains compared with accidents to railway staff and to other types of accident on railway property. Track workers may be run down and some die from electrocution from overhead cables. The pathology of all these is no different from accidents elsewhere, the interest lying in the occupational epidemiology and preventive aspects.

One worrying development in recent years is the malicious damage caused to trains, either by placing objects on the tracks, which may cause a derailment, or the dropping of objects from bridges. The author (BK) has conducted an autopsy on a driver killed by a concrete block dropped from a bridge, which smashed through the windshield of his cab.

The other fairly common railway fatality is the suicide who lays himself in front of an approaching train. Decapitation is the most common injury and the obvious features are the local tissue destruction, usually with grease, rust or other dirt soiling of the damaged area.

The usual search for alcohol and other drugs must be made, as suicides often employ multiple methods to ensure self-destruction. As well as lying down before a locomotive,

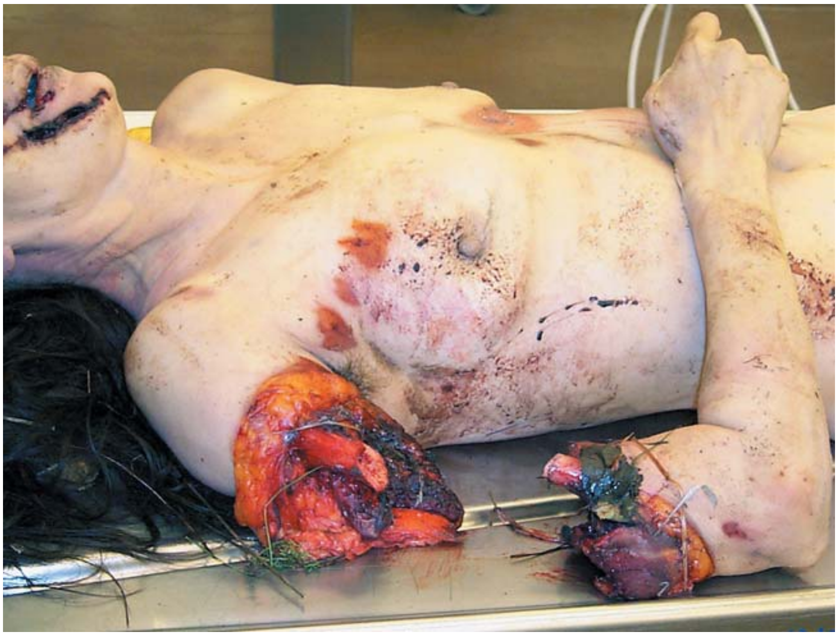

FIGURE 9.24 *Amputation of the right arm and bruising of the face and chest in a pedestrian struck by a passing locomotive.*

another common method of suicide in large cities is to jump from the subway platform of an underground 'tube' or 'metro' system. Here injuries are sometimes complicated by high-voltage electrical lesions, as the typical traction voltage of an electric railway is in excess of 600 volts.

REFERENCES AND FURTHER READING

Ahmed M. 1978. Motorcycle spoke injury. *Br Med J* **2**:401.

AMA Committee on Medical Aspects of Automotive Safety. 1963. Automobile safety belts during pregnancy. *JAMA* **221**:20–1.

Anonymous. 1977. Safety of children in cars [leading article]. *Br Med J* **1**:2.

Anonymous. 1979. Motorcycle and bicycle accidents [special correspondent]. *Br Med J* **1**:39–41.

Anonymous. 1979. Pedestrian accidents [special correspondent]. *Br Med J* **1**:101–4.

Ashton SJ. 1975. The cause and nature of head injuries sustained by pedestrians. *2nd Conference of Biomechemical Serious Trauma*, Lyon, France.

Ashton SJ, Mackay GM. 1978. Pedestrian injuries and death. In: Mason JK (ed.), *The pathology of violent injury.* Edward Arnold, London.

Baik S, Uku JM, Joo KG. 1988. Seat-belt injuries to the left common carotid artery and left internal carotid artery. *Am J Forensic Med Pathol* **9**:38–9.

Baker SO. 1979. Motor vehicle occupant death in young children. *Pediatrics* **64**:860–5.

Baker SP, Spitz WU. 1970. Age effects and autopsy evidence of disease in fatally injured drivers. *JAMA* **214**:1079–88.

Baker SP, Spitz WU. 1970. An evaluation of the hazard created by natural death at the wheel. *N Engl J Med* **283**:405–9.

Bako G, Mackenzie WC, Smith ES. 1970. What is the risk of being burned in a motor vehicle crash? A survey of crash fatalities in alberta. *J Traffic Med* **4**:20–4.

Bothwell PW. 1962. The problem of motor cycle accidents. *Practitioner* **188**:475–6.

Burke DC. 1973. Spinal cord injuries and seatbelts. *Med J Aust* **2**:801–6.

Buttner A, Heimpel M, Eisenmenger W. 1999. Sudden natural death 'at the wheel': a retrospective study over a 15-year time period. 1982–1996. *Forensic Sci Int* **103**:101–12.

Byard RW, Green H, James RA, *et al.* 2000. Pathologic features of childhood pedestrian fatalities. *Am J Forensic Med Pathol* **21**:101–6.

Christian MS. 1975. Non-fatal injuries sustained by back seatbelt passengers. *Br Med J* **1**:320–2.

Christian MS. 1976. Non-fatal injuries sustained by seatbelt wearers: a comparative study. *Br Med J* **2**:1310–12.

Clarke J, Milroy C. 1993. Pedestrian injuries and death. In: Mason JK (ed.), *Pathology of trauma*, 2nd edn. Edward Arnold, London.

Copeland AR. 1991. Pedestrian fatalities. The metropolitan dade county experience, 1984–1988. *Am J Forensic Med Pathol* **12**:40–4.

Cunningham K, Brown TD, Gradwell E, *et al.* 2000. Airbag associated fatal head injury: case report and review of the literature on airbag injuries. *J Accid Emerg Med* **17**:139–42.

Davis GG, Alexander CB, Brissie RM. 1997. A 15-year review of railway-related deaths in Jefferson County, Alabama. *Am J Forensic Med Pathol* **18**:363–8.

Denaner RM, Fitchett VH. 1975. Motorcycle trauma. *J Trauma* **15**:678–81.

DiMaio DJ. 1971. A survey of sudden unexpected deaths in automobile drivers. *Proc 3rd Int Assoc Accident Traffic Med*, pp 75–80.

Eckert W. 1959. Traumatic pathology of traffic accidents; review of 302 cases. *J Forensic Sci* **4**:3–20.

Eckert WB. 1977. *Transportation injuries.* In: Tedeschi L, Eckert WB (eds), *Forensic medicine*. Saunders, Philadelphia.

Edland SF. 1971. The suicide crash. *Proc 3rd Int Assoc Accident Traffic Med*, pp 81–3.

Fisher P. 1965. Injury produced by seatbelts: report of two cases. *J Occup Med* **7**:211–16.

Foster GR, Dunbar JA, Whittet D, *et al.* 1988. Contribution of alcohol to deaths in road traffic accidents in Tayside 1982–6. *Br Med J Clin Res Ed* **296**:1430–2.

Gadd CC. 1966. The use of a weighted impulse criterion to estimating injury hazard. *Proc Tenth Stapp Car Crash Conference*. Society of Automotive Engineers, New York, p. 164.

Gissane W, Bull JP. 1964. Motorway fatalities. *Br Med J* **1**:75–6.

Gissane W, Bull JP, Roberts B. 1970. Sequelae of road injuries. *Injury* **1**:195–200.

Goldberg BA, Mootha RK, Lindsey RW. 1998. Train accidents involving pedestrians, motor vehicles, and motorcycles. *Am J Orthop* **27**:315–20.

Guichon DM, Myles ST. 1975. Bicycle injuries: one-year sample in Calgary. *J Trauma* **15**:504–6.

Hall RR, Fisher AJ. 1972. The influence of car frontal design on pedestrian trauma. *Accid Anal Prev* **4**:47–51.

Hamilton JB. 1968. Seat belt injuries. *Br Med J* **2**:485–6.

Harruff RC, Avery A, Alter-Pandya AS. 1998. Analysis of circumstances and injuries in 217 pedestrian traffic fatalities. *Accid Anal Prev* **30**:11–20.

Hawley DA, Clark MA, Pless JE. 1995. Fatalities involving bicycles: a non-random population. *J Forensic Sci* **40**:205–7.

Hendrickx I, Mancini LL, Guizzardi M, *et al.* 2002. Burn injury secondary to air bag deployment. *J Am Acad Dermatol* **46**:S25–6.

Hotz GA, Cohn SM, Popkin C, *et al.* 2002. The impact of a repealed motorcycle helmet law in Miami-Dade county. *J Trauma* **52**:469–74.

Huelke DF, Chewin WA. 1968. *The energy-absorbing steering column*. Highway Safety Research Institute, Ann Arbor, Michigan.

Huelke DF, Davis RA. 1969. *A study of pedestrian fatalities in Wayne County, Michigan*. University of Michigan, Ann Arbor, Michigan.

Huelke DF, Gikas PW. 1967. Ejection – the leading cause of death in automobile accidents. *Conf Proc Soc Automative Eng*, New York.

Huelke DF, Gikas PW. 1968. Causes of death in automobile accidents. *JAMA* **203**:1100–3.

Huelke DF, Kaufer H. 1975. Vertebral column injuries and seat belts. *J Trauma* **15**:304–18.

Huelke DF, Sherman HW. 1971. Automobile occupant ejection through the side door glass. *Conf Proc Soc Automative Eng*, Detroit.

Huff GF, Bagwell SP, Bachman D. 1998. Airbag injuries in infants and children: a case report and review of the literature. *Pediatrics* **102**:e2.

Jamieson KG, Kelly D. 1973. Crash helmets reduce head injuries. *Med J Aust* **2**:806–9.

Jenkins J, Sainsbury P. 1980. Single-car road deaths – disguised suicides? *Br Med J* **281**:1041.

Jones AM, Bean SP, Sweeney ES. 1978. Injuries to cadavers resulting from experimental rear impact. *J Forensic Sci* **23**:730–4.

Kamdar BA, Arden GR. 1974. Road traffic accidents fatalities – a review of 142 post-mortem reports. *Postgrad Med J* **50**:131–6.

Karger B, Teige K, Buhren W, *et al.* 2000. Relationship between impact velocity and injuries in fatal pedestrian-car collisions. *Int J Legal Med* **113**:84–8.

Kelsch G, Helber MU, Ulrich C. 1996. [Craniocerebral trauma in fall from bicycles – what is the effect of a protective helmet?] *Unfallchirurg* **99**:202–6.

Knight B. 1976. An unnecessary risk to children. *Br Med J* **1**:180.

Larsen CF, Hardt Madsen M. 1988. Fatal motorcycle accidents in the county of Funen (Denmark). *Forensic Sci Int* **38**:93–9.

Lee WB, O'Halloran HS, Pearson PA, *et al.* 2001. Airbags and bilateral eye injury: five case reports and a review of the literature. *J Emerg Med* **20**:129–34.

Lifschultz BD, Donoghue ER. 1994. Deaths due to forklift truck accidents. *Forensic Sci Int* **65**:121–34.

Mackay GM. 1969. Some features of traffic accidents. *Br Med J* **4**:799–801.

Mackay GM. 1975. One engineer's view of human injury. *Injury Br J Accident Surg* **9**:179–83.

Mant AK. 1978. Injuries and death in motor vehicle accidents. In: Mason JK (ed.), *The pathology of violent injury* (and 2nd edition, 1993). Edward Arnold, London.

Mason MA. 1988. Restraining infants in cars. *Br Med J* **296**:1345–6.

McCarrol JR. 1962. Fatal pedestrian automotive accidents. *JAMA* **180**:127–30.

McCarthy M, Gilbert K. 1996. Cyclist road deaths in London 1985–1992: drivers, vehicles, manoeuvres and injuries. *Accid Anal Prev* **28**:275–9.

McDonald QH. 1979. Children's car seat restraints: when top-tether straps are ignored, are these restraints safe? *Pediatrics* **64**:848–55.

Moore J, Tourin B. 1954. *Study of automotive doors opening under crash conditions.* Cornell University, New York.

Morild I. 1994. Traffic deaths in western Norway. A study from the county of Hordaland 1986–1990. *Forensic Sci Int* **64**:9–20.

Morton J. 1968. The suicide driver. *Traffic Saf* **68**:37–40.

Mouzakes J, Koltai PJ, Kuhar S, *et al.* 2001. The impact of airbags and seat belts on the incidence and severity of maxillofacial injuries in automobile accidents in New York state. *Arch Otolaryngol Head Neck Surg* **127**:1189–93.

Murphy GK. 1976. Death on the railway. *J Forensic Sci* **21**:218–26.

Ostrom M, Eriksson A. 2001. Pedestrian fatalities and alcohol. *Accid Anal Prev* **33**:173–80.

Pearlman JA, Eong KG, Kuhn F, *et al.* 2001. Airbags and eye injuries: epidemiology, spectrum of injury, and analysis of risk factors. *Surv Ophthalmol* **46**:234–42.

Petersen PJ, Petty CS. 1962. Sudden natural death among automobile drivers. *J Forensic Sci* **7**:274–7.

Rowe BH, Rowe AM, Bota GW. 1995. Bicyclist and environmental factors associated with fatal bicycle-related trauma in Ontario. *Can Med Assoc J* **152**:45–53.

Schmidt P, Haarhoff K, Bonte W. 1990. Sudden natural death at the wheel – a particular problem of the elderly? *Forensic Sci Int* **48**:155–62.

Selzer ML, Payne CE. 1962. Automobile accidents, suicides and unconscious motivation. *Am J Psychiatry* **119**:237–44.

Sevitt S. 1968. Death after road traffic accidents. *Med Sci Law* **8**:271–87.

Sevitt S. 1968. Fatal road accidents: injuries, complications and cause of death. *Br J Surg* **55**:481–4.

Sevitt S. 1973. Fatal road accidents in Birmingham: times to death and their causes. *Injury* **4**:281–93.

Shennan J. 1973. Seat belt syndrome [letter]. *Br Med J* **4**:786.

Simpson K. 1960. The interpretation of the surface pattern of vehicular injuries. *Med Sci Law* **1**:420–8.

Smith WS, Kaufer H. 1969. Patterns and mechanisms of lumbar injuries associated with lap seat belts. *J Bone Joint Surg Am* **51**:239–54.

Solheim K. 1964. Pedestrian deaths in Oslo traffic accidents. *Br Med J* **1**:81–3.

Transport and Road Research Laboratory. 1974. *Pedestrian injuries*. Department of the Environment, London.

Van der Linden WJ. 2002. Dislocated fracture of the mandibular condylar process after airbag deployment: report of a case. *J Oral Maxillofac Surg* **60**:113–15.

Whent P. 2000. Rail accident investigation. In: Siegel JA, Saukko PJ, Knupfer GC (eds), *Encyclopedia of forensic sciences*. Academic Press, San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, vol 1, pp. 42–47.

Wilks PM. 1967. Safety in car design. *Proc R Soc Med* **60**:955–8.

Wolf RA. 1962. The discovery and control of ejection in automobile accidents. *JAMA* **180**:220–2.

Wyatt JP, Martin A, Beard D, *et al.* 2001. Pedestrian deaths following collisions with heavy goods vehicles. *Med Sci Law* **41**:21–5.

CHAPTER 10

Abuse of human rights: deaths in custody

■ Evidence of fatal physical abuse 301
■ Extrajudicial execution 305
■ Death in custody 305
■ References and further reading 308

Regrettably, abuse of human rights is so widespread on a global scale that some reference must now be made in forensic pathology texts to the physical manifestations of torture. It is stated by Amnesty International (2001) that in a third of the member states of the United Nations torture of one kind or another is practised, in spite of them being signatories to United Nations resolutions banning these atrocities. According to Amnesty International's worldwide report on torture based on a 3-year investigation (2000), currently more than 150 countries routinely torture people compared with 98 countries in their previous report in 1984.

Although most torture does not result in death, there are sufficient cases of physical abuse of human rights that progress to fatality that a number of organizations offer expertise when any alleged case needs to be investigated, usually calling upon forensic pathologists to assist them in such circumstances. These include Amnesty International (based in London), Physicians for Human Rights (PHR), the Medical Foundation for the Treatment of Victims of Torture (London) and many others.

EVIDENCE OF FATAL PHYSICAL ABUSE

In many instances the autopsy appearances of fatal abuse are no different from those by any other homicide, and the confirmation of lethal torture must depend upon circumstantial and other corroborative evidence, which is no concern of this book. For example, if a political detainee is kicked to death, the physical findings may be identical to those of a murder by kicking in a street crime. Similarly, head injuries, shooting and suffocation may show no variation from the same modes of death outside a political context.

Certain features may arouse suspicion or provide definite evidence that the death had political overtones. The word 'political' is used in a wide sense as many abuses of human rights are not directed or even condoned by higher levels of government, though often there is an indifference to being informed about such activities. Many instances of illegal death and injury are perpetrated by the military or the security forces – and even here, the level of command that directs or condones the abuse varies greatly. In general, the civil police and the regular prisons are usually not the perpetrators of gross abuse, though there are many exceptions. It tends to be the armed forces, clandestine security police, and special detention camps that are most often guilty of torture and illegal executions.

One of the features that characterizes torture and abuse is that the victim is usually either in detention or some form of custody, or is temporarily in the power of the authorities, even if that means that soldiers have just burst into a private house. All deaths in custody – discussed later in the chapter – should be examined with care, as much to ensure that the guardians are cleared of any suspicion of ill-treatment as to discover any ill-treatment itself. The fatal event may be quite different from the non-fatal abuse, as a victim who is beaten may be disposed of by shooting. As in all forensic work, the pathological investigations can only be part of the overall enquiry, another example of cooperation and teamwork between all disciplines.

One of the major differences – and great problems – in the investigation of torture and abuse of civil rights is that the authorities in the state where it occurred may be totally uncooperative and, indeed, wholly hostile to impartial enquiry. It is through respected international agencies, such as the Red Cross, Amnesty International and Physicians for Human Rights, that pathologists have the best chance of being allowed access to the relevant material, and the organizations mentioned above are willing to undertake or assist legitimate investigations anywhere in the world.

One particular problem associated with human rights abuse is the frequent long delay before being able to examine the victim. In the living, they may have been detained for months or years after the ill-treatment, or may have been exiled or in hiding before they have the opportunity to be seen by an impartial medical observer, by which time acute injuries will have healed, bruises absorbed and wounds and burns scarred over. In the dead, long-term post-mortem changes are common: the body may have been buried and exhumed before examination.

Even skeletal material is all that is sometimes available, as in the 'lost children' of Argentina or in the killing fields of Uganda. Problems of identification are often difficult. Contemporary photographs of injuries are often of poor quality and the facilities for autopsy are primitive in some of the less developed countries. When the suspicion, lack of cooperation and often open hostility of the local authorities is added, the task for the visiting pathologist is formidable indeed.

The following types of injury may confirm or arouse suspicion of torture, though, as stated, virtually any type of injury can be inflicted deliberately to extract information, punish or degrade the victims or to exterminate them. Repetition of a particular injury, such as 'tramline' bruises from beating, is suspicious, because in ordinary assaults and homicides such excessive and perhaps regularly placed repetitive lesions would be unusual.

- Beating is one of the most common forms of torture and can take many forms, varying both with the weapon used and the part of the body injured. Blows both to the head and to other parts of the body were present in all the cases reported by Hougen (1988). Unless severe and repeated, beating is not often the sole cause of death, though this can occur from haemorrhage, sepsis, injured internal organs, or from sheer exhaustion and pain in an already debilitated victim. The blows may be inflicted by fist or foot, but are more often applied with a weapon. The use of a whip or lath-like instrument is common, but metal or wooden bars, clubs, batons, rifle butts or belts may be used. More recently, lengths of plastic hosepipe or tubing have come into favour in some parts of the world.

Many of these produce a characteristic skin lesion, namely a single- or double-edged linear bruise. The bruise may be a line of confluent petechial haemorrhages or a continuous mark of red skin. The 'tramline' bruise is a double line of parallel marks with a pale unbruised zone between them, caused by the impact of a rectangular or circular-sectioned object. This is described in more detail in Chapter 4.

The bruising may be intradermal, when it reproduces the pattern of the weapon well. If a leather whip with, for example, plaited thongs, is struck against the skin, the pattern may be imprinted clearly on the skin.

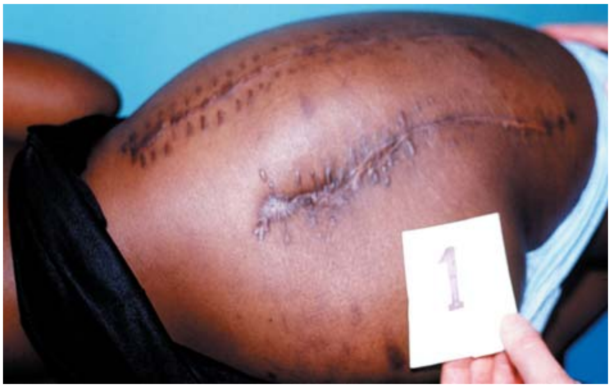

FIGURE 10.1 *A scar about one year after the original injury. The female victim was shot while escaping from an East African concentration camp. The wound was opened surgically and healed imperfectly with keloid formation. The other surgical incision is of more recent origin. Such scars will remain unaltered for life, apart from possible increase in the keloid formation.*

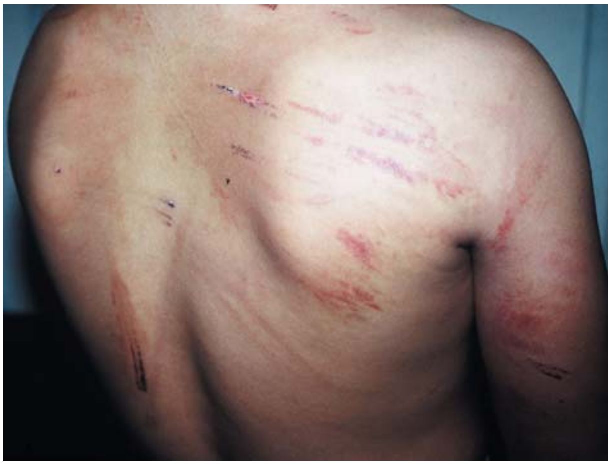

FIGURE 10.2 *Recent bruising inflicted within the previous 2 days, from blows by lengths of plastic tubing about 15 mm wide. The injuries show parallel 'tramlining' typical of impact from a cylindrical weapon. The assailants were vigilantes administering a punishment beating to an alleged collaborator in a Middle Eastern country.*

Buckles on belts and other recognizable artefacts may occasionally be useful in identifying the weapon. Where the bruising is deeper, no such pattern is likely. The skin may be broken, causing abrasions or lacerations and, if the full thickness is breached, healing will cause scars, which may even be recognizable as 'tramlines'.

Repeated beating leaves multiple marks, which though overlapping and often criss-crossing, may have a generally similar orientation that indicates that the attacker stood in a relatively fixed position to the victim. For example, lashing across the back may leave marks running from top right to bottom left, suggesting that a right-handed perpetrator stood to the left of the victim.

The use of a multi-thonged whip, such as a 'cat-o'nine-tails' will leave a series of marks, again generally in the same orientation. The thongs tend to be narrow and may cause linear marks with tramlining. There may be metal tags or knots on the end of each thong, which can cause focal damage. The back is the most frequent target, but whipping and beating may be applied to the buttocks, thighs, front of chest, breasts and abdomen, lower legs, soles of the feet, and even perineum and genitals.

Beating of the soles of the feet with canes or rods is a torture rooted in antiquity and this so-called 'falanga' may leave relatively little to see, even though it is extremely painful and debilitating. The tough tissue and thick fascial planes of the foot do not readily reveal bruising, though it may be found on deep dissection in victims who have been killed by some other means.

If injuries were inflicted months before examination, little or nothing may be found unless the skin was broken, when scarring will have taken place. Sometimes, however, faint red lines may be seen in pale-skinned people and in those with appreciable racial melanin pigment there is often hyperpigmentation along the lines of injury. Where severe damage has been caused and sometimes where there is scar tissue, depigmentation may be seen.

- Burns are unfortunately common and may be either the actual cause of death or visible as recent or scarred evidence of previous torture. All kinds of burns may be suffered. The author (BK) has seen extensive burns from molten rubber dripped onto victims from motor tyres suspended overhead, hot irons applied to skin, ignited kerosene-soaked rags wrapped around limbs and numerous burns from cigarettes pressed into the skin. Knowledge of local practices, such as the burning motor tyre 'necklace', can help identify burns in typical sites. Where molten liquids have been dripped from above onto a bound-up victim, there may be shadow areas that indicate the direction of contact, which makes any innocent explanation highly unlikely. In survivors, scarring occurs in all but the most superficial burns, and may lead to large, unsightly cicatrized areas. In victims of African origin, large keloids may form, and further complicate the damage and attempts at surgical treatment.
- Cutting and stabbing may be inflicted with a variety of weapons, but wounds from knives and bayonets are most common in the context of torture and extrajudicial execution. The features are identical to

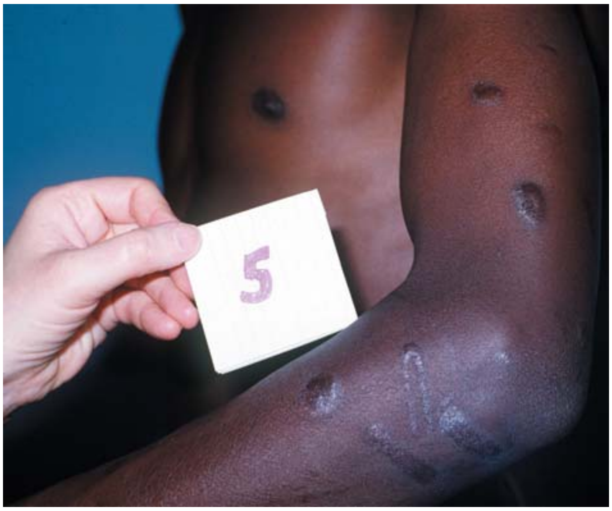

FIGURE 10.3 *Scars from beating with a thin metal rod in an East African torture camp. Though more than a year old, the 'tramline' nature of one of the scars is outlined in keloid.*

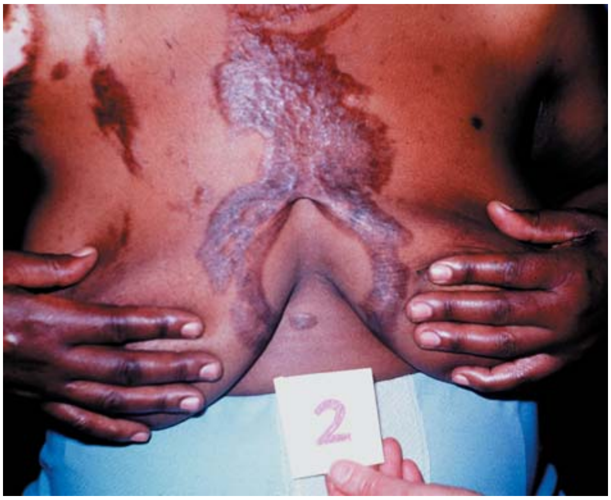

FIGURE 10.4 *Keloid scars from extensive burns during torture. The elderly lady was tied to a chair and a burning motor tyre suspended over her. The shadowed area between the breasts indicates the downward trickle of molten burning rubber. Her face was also extensively scarred.*

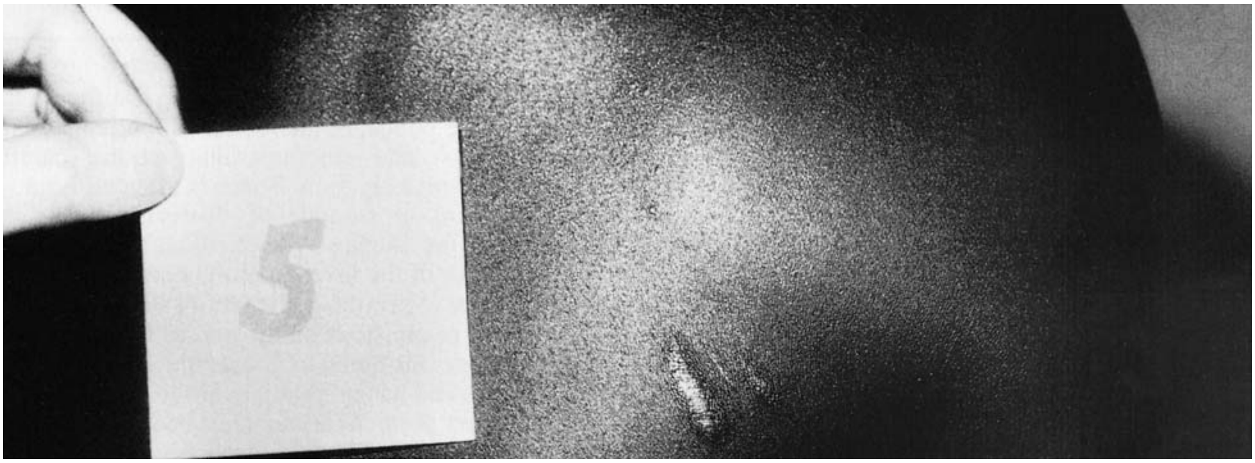

FIGURE 10.5 *The scar of a bayonet stab wound inflicted many months earlier in a Ugandan torture camp. Though scars are usually of little help in interpreting the cause of the injury, here the elliptical shape of the original stab wound can still be seen as a keloid scar and even the sharp upper end of the blade edge, as opposed to blunter lower end.*

those described in Chapter 4. The site on the body may be anywhere, but the chest and upper arms seem to be favourite targets. The scars of old stab wounds, such as those from a bayonet, may still be recognizably elliptical many months or years after infliction.

- Clubbing and blunt injury is extremely common, either to cause death or as a form of abuse. It is sometimes dispensed almost casually, as when guards lay about themselves with rifle butts against the inmates of detention camps. Again the features are non-specific and are described in the chapter on wounding. The head is the most common target, but the legs and knees are also often struck. Blows on the back or side of the neck can be particularly lethal from vertebrobasilar artery damage (Chapter 5).
- Deaths from suffocation and drowning are not common, though non-fatal practices of this kind are well-known methods of abuse. Repeated dipping of the victim's head under water or even foul liquid such as sewage is called 'submarining', and may cause drowning, air-passage occlusion or a later pneumonia. Enveloping the head in an opaque plastic bag is more a means of disorientation than physical torture, but partial suffocation by similar means may eventually prove fatal.
- Electrical torture is common and well documented. Either mains voltage of 110 or 240 V is used – which carries the risk of fatal cardiac arrhythmia as well as local burning – or a magneto delivering high voltage, which is painful but not lethal. The latter will not leave any significant skin marks, as the amperage is so low.

  The later demonstration of electrical injury is difficult, as healed lesions have few characteristic features. The acute injuries which may cause burns and death are described in Chapter 12. Karlsmark and his co-authors have made a special study of the pathology of electrical torture. The current may be applied anywhere, but the genitals are sites favoured by torturers, especially the penis and scrotum. The female nipples are also targets.
- Injuries to the ears can rarely be demonstrated at autopsy, but a favourite torture is the 'telefono', consisting of repeated slapping of the sides of the head by the open palms of the assailant. This may rupture the tympanic membranes and injure the inner ear.

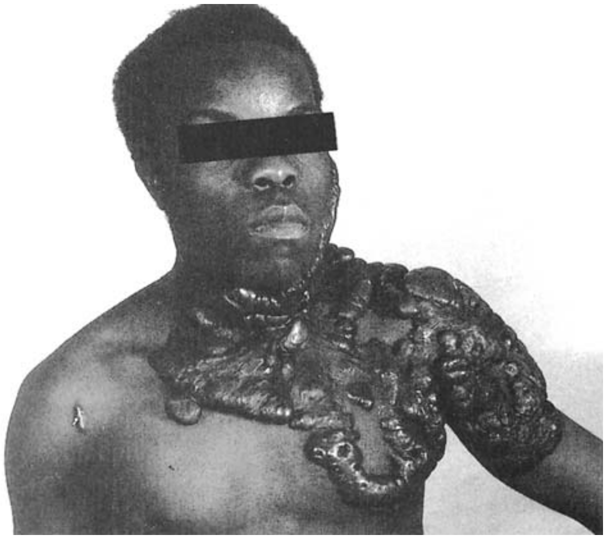

FIGURE 10.6 *Extensive keloid scars following deliberate burning of a 16-year-old student in an East African concentration camp. An electric hot plate was pressed against his neck and kerosene-soaked rags were ignited on his arm.*

- Suspension is a common torture, though again not often fatal in itself. If the victim is seen relatively soon after the ordeal then abrasions, bruises and chafing marks may be found at ligature sites, usually on legs, arms and sometimes genitals.
- Shooting is not a method of torture in the accepted sense, but is a common means of execution or of non-fatal punishment, such as the 'knee-capping' perpetrated in Northern Ireland, where opponents and suspected traitors are shot either through the knee joint or the lower thigh.
- Sexual abuse is also common and in women, rape – sometimes multiple – may leave physical signs as described in Chapter 18.

Medical examiners of alleged victims of torture must be careful not to misinterpret non-torture lesions as evidence of abuse, as such lapses will be seized upon by opponents wishing to discredit the medical opinion, either for political or immigration reasons. Forrest's guide (Forrest *et al.* 1995) lists most of these, which include innocent occupational, sporting or other accidents, tribal marks, skin infections, variations in pigmentation, traditional scarification, healing marks, stretch marks, etc.

EXTRAJUDICIAL EXECUTION

Many hundreds of thousands – indeed probably millions – of victims have been assassinated in recent years by governmental agencies acting quite outside the legal system. Because of ethnic origin, tribal feuds or political views, whole sections of populations have 'disappeared' on a scale sometimes amounting to genocide as in Rwanda, Cambodia, Bosnia and Kosovo. It is common knowledge that these disappearances are usually explained by murder, either on a mass scale or by continuous individual or group extermination by 'extermination squads'. Specific instances are not the concern of a textbook, but as recent years have shown, severe excesses have taken place in most continents leading to several forensic exercises directed towards identification of victims and proving methods of torture.

When vanished people have been presumed to have been killed by troops or security agencies, the evidence is usually very late in coming to light, so that forensically, much of the work consists of skeletal identification and investigations into causes of death. Where the bodies have decomposed or skeletalized, shooting or head injuries may be the only mode of death to leave any telltale evidence on exhumation. Autopsy examination may resemble mass-disaster procedures where there are large numbers of victims. The talents of odontologists and anthropologists, and the techniques of archaeology may be needed in such cases, as described by Clyde Snow in relation to the 'lost children' of Argentina or more recently in Bosnia and Herzegovina, Croatia and Kosovo. The vast numbers of skeletal remains in the Lowero triangle of Uganda have so far had no such attention.

DEATH IN CUSTODY

The occurrence of death while a person is either in the custody of the police, or the inmate of a prison, raises public interest and emotions that require careful handling and investigation. There is often an immediate complaint or rumour of ill-treatment by the relatives or the media. A meticulous autopsy is a necessary part of the investigation needed to dispel – or sometimes confirm – allegations that an act of commission or omission on the part of the custodians has led to, or contributed to, the death. This applies in all countries and has no relationship to the preceding section, when political issues may be suspected as a cause of the death. The two aspects are not always totally separate, however, as countries with excellent human rights records have sometimes been suspected of occasional lapses of standards where custodial deaths are concerned.

Most of the cases dealt with by a forensic pathologist will concern civil police forces and the regular prison service. Many countries, including Britain, have strict legislative rules that make all deaths in custody, whether police or prison, reportable for medico-legal investigation by the appropriate law officer (the coroner in England and Wales), in addition to a searching internal enquiry by officers from a different police force (e.g. in Finland the National Bureau of Investigation).

An autopsy will inevitably be held, regulations in UK making this a case for an accredited forensic pathologist. It is sometimes helpful if the pathologist visits the scene of the death, especially if it was caused by hanging or some form of mechanical trauma.

There are several situations in which death may occur in custody. The deceased may die during or soon after arrest by police officers, as there is often a struggle in which the offender is resisting arrest, sometimes violently. It may be a physical struggle, especially when police officers are attempting to control or overpower a resisting offender. In these situations, the use of so-called personal defence sprays has increased, as they have been shown to be very effective in incapacitating an aggressor. They cause acute ocular irritation, lachrymation, conjunctivitis, blepharospasm and extreme discomfort. Most of these sprays contain *o*-chlorobenzylidene malononitrile (CS), 2-chloroacetophenone (CN), oleoresin capsicum (OC), or a combination of these ingredients as the active agent. Capsacain (8-methyl-*N*-vanillyl-6-nonenamide) is a neurotoxin component of cayenne pepper. It stimulates excitatory afferent sensory neurones, causes hypothermia, neurogenic inflammation and pain, followed by subsequent desensitization. Although these sprays have been considered to be relatively safe, unexpected deaths have been reported after the exposure to oleoresin capsicum (OC) spray devices. Pepper spray has been suspected to induce bronchoconstriction. Chan and co-workers (2002) tested the effect of (OC) spray inhalation on respiratory function in 35 healthy subjects by itself and combined with restraint but could not find evidence of hypoxaemia or hypercapnia in either group. OC exposure did not result in abnormal spirometry, hypoxaemia, or hypoventilation when compared to placebo in either sitting or prone maximal restraint position. In another study on clinical toxicity of OC, in Kansas, Watson *et al.* (1996) reported on 81 emergency department patients who had been investigated after exposure to OC. Ocular burning and redness were the most common presenting symptoms. None of the patients required hospitalization due to OC toxicity. Corneal abrasions and respiratory symptoms occurred in seven and six patients, respectively. In case of death after OC exposure, a thorough autopsy, with ancillary investigations, including complete histology and toxicological analysis with assessment of the circumstances and symptomatology, is necessary to assess whether the exposure to OC spray has been causative, contributory, or unrelated to death.

On other occasions, the offender may threaten police officers with a knife, gun or blunt weapon, and the police have to subdue him either by sheer physical force – truncheons or riot sticks – or by the use of firearms. The most common event is the arrest of a drunken offender and this poses extra problems, discussed below. Other drugs, such as cocaine, cannabis, amphetamine and hallucinogens, rather than opiates such as morphine, heroin and barbiturates, may also be involved.

The physical overpowering of a suspected offender poses definite risks to health and life, and has been the subject of numerous controversial enquiries and litigation in many countries. The police usually outnumber the offender, sometimes by a considerable margin, but even a one-to-one struggle can be dangerous for either party. The following risks exist, though the list is by no means comprehensive.

- Traumatic asphyxia may occur where several policemen fall upon a resisting subject to overpower him. The author (BK) has dealt with several deaths where a number of large police officers piled on top of a prisoner fighting maniacally, to apply handcuffs. When they got up, the man was not breathing and died in hospital shortly afterwards. Death was caused by traumatic asphyxia, the weight of the men on top of him causing chest compression and prevention of respiratory movements. The appearances are described in Chapter 14.
- Arm-locks or neck-holds applied by police officers to resisting persons are other causes of deaths during arrest. These have been particularly notorious in the USA, where police officers have been trained to apply neck-holds as a form of restraint, but similar deaths have been seen elsewhere. The arm-lock is applied either from behind or with the head of the offender tucked under the police officer's arm against his waist. The dangers are compression of the front or sides of the neck, and death can occur either from reflex cardiac arrest or cerebral ischaemia during carotid compression, or asphyxia from airway obstruction, though the latter is unlikely as the sole mechanism. According to American writers such as Reay and Eisele (1982, 1986), there are two types of neck-hold – the 'bar arm control' and the 'carotid sleeper'. The former is alleged to be more dangerous; the officer's forearm is pulled across the front of the larynx to occlude the airway. The carotid sleeper uses the two sides of the 'V' formed by the flexed upper and lower arm to compress the carotids and produce transient cerebral ischaemia. Both varieties are dangerous, however, as there is the ever-present hazard of reflex cardiac arrest from vagal stimulation from the carotid sinus and sheath, as well as the potential danger of neck traction and hyperextension causing subarachnoid haemorrhage from vertebrobasilar artery damage.
- Postural (positional) asphyxia has been reported to have caused sudden deaths in persons after the use of the 'hogtie', 'hobble' or prone maximal restraint and even in situations where a person has been placed in a prone position in the rear compartment of a police car. In an experiment with healthy subjects, the restraint position resulted in a restrictive pulmonary function pattern but did not cause clinically relevant changes in oxygenation or ventilation. It has been postulated that the hogtie restraint position by itself does not cause respiratory compromise to the point of asphyxiation but other factors, e.g. acute intoxication by alcohol or drugs, are responsible for the sudden deaths of individuals placed in this position.
- Blunt injury may occur from the use of fist, arm or leg – or the use of a weapon such as a truncheon, riot stick or pistol butt. During a struggle, usually with one or more police officers who are attempting to subdue a fighting offender, injuries may be received by either side. All types of blunt injury may be sustained, some potentially fatal and these are discussed in Chapter 4. Head injuries may occur during a scuffle from falls either against the ground, or against a wall or other

obstruction. A heavy punch in the face may cause nasopharyngeal bleeding that can block the air passages, especially in a person affected by alcohol. A blow on the side of the neck can cause reflex cardiac arrest or a subarachnoid haemorrhage from vertebrobasilar vascular damage.

A backward blow from the point of an elbow can be damaging, if it strikes the face, neck or abdomen. Kicking and stamping are unusual, but not unknown, in custodial deaths. A blow in the abdomen can also be serious if delivered with sufficient force. Though in an adult a fist blow is not very likely to cause serious harm, it can in a slighter young person. The use of the elbow, knee or a head butt can deliver extreme force, especially from a fit, muscular police officer, as well as from an offender.

The author (BK) has seen fatal intraperitoneal bleeding amounting to three litres in a drunken man arrested after a scuffle with two police officers, in which there were allegations that one struck or fell on the man with his knee. Several hours later the prisoner collapsed and at autopsy there were several large tears in the mesentery.

Alcohol is a frequent cause of death in custody. Not only is it the major factor in provoking aggression and violent resistance, with the consequences mentioned above, but it can have other effects which lead to death while in the care of the police. Acute alcoholic poisoning, described more fully in Chapter 28, may lead to death while the victim is thought to be 'sleeping it off' in a police cell. When blood alcohol levels rise to above 350 mg/100 ml there is an increasing risk of coma and central respiratory depression. Although most responsible police forces have standing orders about placing drunken prisoners in the semiprone position and observing them at frequent intervals, a quiet drunk may still slip into irreversible coma and respiratory arrest.

At lower blood alcohol levels there is still the risk of aspiration of vomit and choking on gastric contents. Though drunken prisoners are placed in the safest posture for drainage, they can still vomit and choke when unobserved, between visits by the often busy police station officers. At autopsy, caution must be observed before ascribing death to aspiration of vomit, as this is a common agonal phenomenon in deaths from other causes. Where an otherwise healthy person dies with a high blood alcohol concentration in these circumstances, however, gross blocking of the trachea and bronchi with vomit forms one of the most convincing arguments for acceptance of aspiration as the cause of death if no other factors can be identified.

Alcohol also contributes to accidents during custody, especially head injuries, which come about because of falls to the ground, falls down steps and stairs, and being run over by traffic, as a result of the ataxia and incoordination of the drunken state. Falls onto a hard surface are often on the occiput and the frequent finding of frontal and temporal contrecoup brain damage at autopsy is good evidence of a deceleration injury rather than an assault with a weapon.

Some falls may occur during custody or in transit from the site of arrest to the police station – others have happened before arrest, but the ill-effects and death may become manifest during the stay in the police cell, when the police are often blamed either for allowing or causing the injury – or for not summoning or providing urgent medical attention. Drugs, especially those causing excitement, such as amphetamine, cocaine or hallucinogens, may also lead to physical damage, but alcohol remains by far the most common.

Drugs are commonly available within prisons and overdose or hypersensitivity deaths are occasionally seen amongst prisoners, as they are in the outside world.

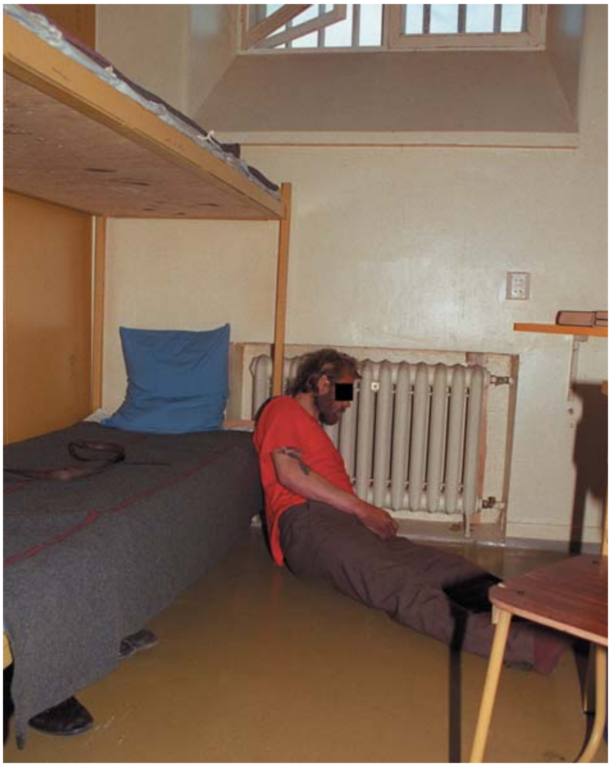

FIGURE 10.7 *Suicidal hanging in a prison cell. This is an infrequent but regular tragedy, which, in spite of strenuous efforts by prison officers and designers of cells, occurs because of the determination of the prisoner to defeat the safeguards. The belt used as ligature has been removed and lies on the bed.*

- Suicide in custody is not uncommon, and often leads to accusations and recriminations from the relatives over lack of supervision. There has recently been a spate of such deaths in Britain, especially amongst young offenders on remand awaiting trial. It is such a well-recognized hazard of custody that most police forces deprive the prisoner of any objects, such as belts, braces (suspenders), cord or even bootlaces, that could be used to hang himself in his cell. In addition, the police cell may be specifically designed to avoid any convenient suspension points, such as hooks, bars or even internal door handles. In spite of these precautions, prisoners regularly manage to find some means of killing themselves. Strips of bedding material, sleeves of clothing and handkerchiefs have all been used for self-suspension. As described in Chapter 14, hanging can be successfully accomplished by traction on the neck at low levels and need not occur from high suspension points, so prisoners have killed themselves by attaching ligatures to bed-heads, chairs and other unlikely objects in the cell.

  Accusations that apparent suicidal hangings were in reality homicides by the custodians can usually be resolved by the autopsy showing no signs of bruising, abrasions or a struggle. It would seem impossible to hang a conscious person against his will without leaving some signs of restraint. The proposition that the hanging was a ligature strangulation is usually disproved by the rising angle of the ligature mark, which most commonly sets under the angles of the jaw and has a defect in the skin mark where the suspension point pulls the ligature away from the surface. A homicidal strangulation usually encircles the neck and a cross-over point is often seen. The level is lower, nearer the horizontal (Chapter 14).

  Where a true hanging mark is caused by a ligature with a slip-knot, there may also be a full circle of skin abrasion and compression. If the suspension point is low, as might happen in a cell using a doorknob or a bed-frame, then the direction of pull can be oblique. If the body also leans over against the direction of pull, the resultant mark can be virtually horizontal and, unless the situation is fully assessed – preferably by a visit to the scene with the body *in situ* – confusion can occur, as with the controversy over the death of Rudolph Hess in 1988, whilst an inmate in Berlin's Spandau prison.
- Death may be from purely natural causes, usually cardiovascular in origin, which happened to have occurred during detention. It must be admitted, though it is almost impossible to provide objective proof, that the emotional and sometimes physical upset of being arrested and confined may have affected the blood pressure and heart rate sufficiently, by an adrenaline response, to have precipitated an acute cardiac crisis in the presence of severe pre-existing disease.

  The presence of diabetes, epilepsy, asthma or other diseases that can potentially cause sudden or unexpected death should be sought by medical history and autopsy appearances (Chapter 25).

REFERENCES AND FURTHER READING

Amnesty International. 2001. *Amnesty international report 2001*. Amnesty International, London.

Billmire DF, Vinocur C, Ginda M, *et al.* 1996. Pepper-spray-induced respiratory failure treated with extracorporeal membrane oxygenation. *Pediatrics* **98**:961–3.

Blaauw E, Vermunt R, Kerkhof A. 1997. Deaths and medical attention in police custody. *Med Law* **16**:593–606.

Bloom AI, Zamir G, Muggia M, *et al.* 1995. Torture rhabdomyorhexis – a pseudo-crush syndrome. *J Trauma* **38**:252–4.

Bork K, Nagel C. 1997. Long-standing pigmented keloid of the ears induced by electrical torture. *J Am Acad Dermatol* **36**:490–1.

Brkic H, Strinovic D, Kubat M, *et al.* 2000. Odontological identification of human remains from mass graves in croatia. *Int J Legal Med* **114**:19–22.

Bro Rasmussen F, Henriksen OB, Rasmussen OV, *et al.* 1982. [Aseptic necrosis of bone following phalanga torture.] *Ugeskr Laeger* **144**:1165–6.

Brown L, Takeuchi D, Challoner K. 2000. Corneal abrasions associated with pepper spray exposure. *Am J Emerg Med* **18**:271–2.

Busker RW, van Helden HP. 1998. Toxicologic evaluation of pepper spray as a possible weapon for the Dutch police force: risk assessment and efficacy. *Am J Forensic Med Pathol* **19**:309–16.

Chan TC, Vilke GM, Neuman T. 1998. Reexamination of custody restraint position and positional asphyxia. *Am J Forensic Med Pathol* **19**:201–5.

Chan TC, Vilke GM, Clausen J, *et al.* 2002. The effect of oleoresin capsicum 'pepper' spray inhalation on respiratory function. *J Forensic Sci* **47**:299–304.

Copeland AR. 1984. Deaths in custody revisited. *Am J Forensic Med Pathol* **5**:121–4.

Corovic N, Durakovic Z, Zavalic M, *et al.* 2000. Electrocardiographic changes in ex-prisoners of war released from detention camps. *Int J Legal Med* **113**:197–200.

Dalton V. 1999. Death and dying in prison in Australia: national overview, 1980–1998. *J Law Med Ethics* **27**:210, 269–74.

Epstein RJ, Majmudar PA. 2001. Pepper spray in the eye. *Ophthalmology* **108**:1712–13.

Fernandez M, Pissiota A, Frans O, *et al.* 2001. Brain function in a patient with torture related post-traumatic stress disorder before and after fluoxetine treatment: a positron emission tomography provocation study. *Neurosci Lett* **297**:101–4.

Fitzpatrick JJ. 1984. Role of radiology in human rights abuse. *Am J Forensic Med Pathol* **5**:321–5.

Forrest D. 1995. The physical after-effects of torture. *Forensic Sci Int* **76**:77–84.

Forrest D, Knight B, Hinshelwood G, *et al.* 1995. A guide to writing medical reports on survivors of torture. *Forensic Sci Int* **76**:69–75.

Frost R, Hanzlick R. 1988. Deaths in custody. Atlanta city jail and Fulton county jail, 1974–1985. *Am J Forensic Med Pathol* **9**:207–11.

Fruehwald S, Frottier P, Eher R, *et al.* 2000. Prison suicides in Austria, 1975–1997. *Suicide Life Threat Behav* **30**:360–9.

Giles HG, Sandrin S. 1992. Alcohol and deaths in police custody. *Alcohol Clin Exp Res* **16**:670–2.

Gniadecka M, Danielsen L. 1995. High-frequency ultrasound for torture-inflicted skin lesions. *Acta Derm Venereol* **75**:375–6.

Goldfeld AE, Mollica RF, Pesavento BH, *et al.* 1988. The physical and psychological sequelae of torture. Symptomatology and diagnosis [published erratum appears in *JAMA* 1988 Jul 22–29;**260**(4):478]. *JAMA* **259**:2725–9.

Goldney RD. 1993. Deaths in custody. *Med J Aust* **159**:572–3.

Gordon E, Knight B. 1985. *Uganda: evidence of torture*. Amnesty International, London.

Granzow B, Puschel K. 1998. [Fatalities during imprisonment in Hamburg 1962–1995.] *Arch Kriminol* **201**:1–10.

Hiss J, Kahana T. 1996. Medicolegal investigation of death in custody: a postmortem procedure for detection of blunt force injuries. *Am J Forensic Med Pathol* **17**:312–14.

Hougen HP. 1988. Physical and psychological sequelae to torture. A controlled clinical study of exiled asylum applicants. *Forensic Sci Int* **39**:5–11.

Jandoo R. 1987. Human rights abuses and the medical profession. *Forensic Sci Int* **35**:237–47.

Johnson HR. 1982. Deaths in police custody in England and Wales. *Forensic Sci Int* **19**:231–6.

Karch SB, Stephens BG. 1999. Drug abusers who die during arrest or in custody. *J R Soc Med* **92**:110–13.

Karlsmark T, Thomsen HK, Danielsen L, *et al.* 1984. Tracing the use of electrical torture. *Am J Forensic Med Pathol* **5**:333–7.

Kirschner RH. 1984. The use of drugs in torture and human rights abuses. *Am J Forensic Med Pathol* **5**:313–15.

Lacroix JS, Buvelot JM, Polla BS, *et al.* 1991. Improvement of symptoms of non-allergic chronic rhinitis by local treatment with capsaicin. *Clin Exp Allergy* **21**:595–600.

Lanphear BP. 1987. Deaths in custody in Shelby county, Tennessee, January 1970–July 1985. *Am J Forensic Med Pathol* **8**:299–301.

Lee RJ, Yolton RL, Yolton DP, *et al.* 1996. Personal defense sprays: effects and management of exposure. *J Am Optom Assoc* **67**:548–60.

Levine LJ. 1984. The role of the forensic odontologist in human rights investigations. *Am J Forensic Med Pathol* **5**:317–20.

Levine M. 1998. Deaths in police custody. *Med Leg J* **66**:97–108.

Lifschultz BD, Donoghue ER. 1991. Deaths in custody. *Leg Med* 45–71.

Lok V, Tunca M, Kumanlioglu K, *et al.* 1991. Bone scintigraphy as clue to previous torture [letter] [see comments]. *Lancet* **337**:846–7.

Luke JL, Reay DT. 1992. The perils of investigating and certifying deaths in police custody. *Am J Forensic Med Pathol* **13**:98–100.

Macdonald HA. 1998. Possible idiosyncratic reaction to OC spray. *N Z Med J* **111**:327.

Mercy JA, Heath CW, Jr., Rosenberg ML. 1990. Mortality associated with the use of upper-body control holds by police. *Violence Vict* **5**:215–22.

Mirzaei S, Knoll P, Lipp RW, *et al.* 1998. Bone scintigraphy in screening of torture survivors. *Lancet* **352**:949–51.

Mirzaei S, Knoll P, Keck A, *et al.* 2001. Regional cerebral blood flow in patients suffering from post-traumatic stress disorder. *Neuropsychobiology* **43**:260–4.

Missliwetz J, Denk W. 1991. [Maltreatment by police officers?] *Arch Kriminol* **187**:1–12.

Morrison S. 1996. Custodial suicide in Australia: a comparative study of different populations. *Med Sci Law* **36**:167–77.

Norfolk G, Cartwright J. 1996. Deaths in police custody are being analysed retrospectively. *Br Med J* **312**:911.

Patel F. 2000. Custody restraint asphyxia. *Am J Forensic Med Pathol* **21**:196–7.

Patel V. 1996. Deaths in police custody. *Br Med J* **312**:56.

Petersen HD, Rasmussen OV. 1992. Medical appraisal of allegations of torture and the involvement of doctors in torture. *Forensic Sci Int* **53**:97–116.

Pollanen MS, Chiasson DA, Cairns JT, *et al.* 1998. Unexpected death related to restraint for excited delirium: a retrospective study of deaths in police custody and in the community. *Can Med Assoc J* **158**:1603–7.

Primorac D, Andelinovic S, Definis-Gojanovic M, *et al.* 1996. Identification of war victims from mass graves in croatia, bosnia, and herzegovina by use of standard forensic methods and DNA typing. *J Forensic Sci* **41**:891–4.

Rainio J, Hedman M, Karkola K, *et al.* 2001. Forensic osteological investigations in Kosovo. *Forensic Sci Int* **121**:166–73.

Rainio J, Lalu K, Penttila A. 2001. Independent forensic autopsies in an armed conflict: investigation of the victims from Racak, Kosovo. *Forensic Sci Int* **116**:171–85.

Rasmussen OV. 1990. Medical aspects of torture. *Dan Med Bull* **37**(Suppl 1):1–88.

Rasmussen OV, Lunde I. 1980. Evaluation of investigation of 200 torture victims. *Dan Med Bull* **27**:241–3.

Rasmussen OV, Skylv G. 1992. Signs of falanga torture [letter]. *Lancet* **340**:725.

Reay DT. 1998. Death in custody. *Clin Lab Med* **18**:1–22.

Reay DT, Eisele JE. 1982. Death from law enforcement neck holds. *Am J Forensic Med Pathol* **3**:253–8.

Reay DT, Eisele JE. 1986. Law enforcement neck holds. *Am J Forensic Med Pathol* **7**:177–87.

Reay DT, Hazelwood RR. 1970. Death in military police custody and confinement. *Milit Med* **135**:765–71.

Reilly CA, Crouc DJ, Yost GS, *et al.* 2001. Determination of capsaicin, dihydrocapsaicin, and nonivamide in self-defense weapons by liquid chromatography-mass spectrometry and liquid chromatography-tandem mass spectrometry. *J Chromatogr A* **912**:259–67.

Reilly CA, Crouch DJ, Yost GS. 2001. Quantitative analysis of capsaicinoids in fresh peppers, oleoresin capsicum and pepper spray products. *J Forensic Sci* **46**:502–9.

Richmond PW, Fligelstone LJ, Lewis E. 1988. Injuries caused by handcuffs [see comments]. *Br Med J* **297**:111–12.

Ross DL. 1998. Factors associated with excited delirium deaths in police custody. *Mod Pathol* **11**:1127–37.

Ruttenber AJ, Lawler-Heavner J, Yin M, *et al.* 1997. Fatal excited delirium following cocaine use: epidemiologic findings provide new evidence for mechanisms of cocaine toxicity. *J Forensic Sci* **42**:25–31.

Savnik A, Amris K, Rogind H, *et al.* 2000. MRI of the plantar structures of the foot after falanga torture. *Eur Radiol* **10**:1655–9.

Skinner M. 1987. Planning the archaeological recovery of evidence from recent mass graves. *Forensic Sci Int* **34**:267–87.

Skylv G. 1992. The physical sequelae of torture. In: Basoglu M (ed.), *Torture and its consequences: current treatment approaches.* Cambridge University Press, Cambridge.

Smith CG, Stopford W. 1999. Health hazards of pepper spray. *N C Med J* **60**:268–74.

Snow CC, Levine L, Lukash L, *et al.* 1984. The investigation of the human remains of the 'disappeared' in Argentina. *Am J Forensic Med Pathol* **5**:297–9.

Steffee CH, Lantz PE, Flannagan LM, *et al.* 1995. Oleoresin capsicum (pepper) spray and 'in-custody deaths'. *Am J Forensic Med Pathol* **16**:185–92.

Stover E, Nightingale EO. 1985. The medical profession and the prevention of torture. *N Engl J Med* **313**:1102–4.

Tedeschi LG. 1984. Human rights and the forensic scientist. *Am J Forensic Med Pathol* **5**:295–6.

Tedeschi LG. 1984. Methodology in the forensic sciences. Documentation of human rights abuses. *Am J Forensic Med Pathol* **5**:301–3.

Thomsen AB, Eriksen J, Smidt-Nielsen K. 1997. [Neurogenic pain following Palestinian hanging.] *Ugeskr Laeger* **159**:4129–30.

Thomsen AB, Eriksen J, Smidt-Nielsen K. 2000. Chronic pain in torture survivors. *Forensic Sci Int* **108**:155–63.

Thomsen JL. 2000. The role of the pathologist in human rights abuses. *J Clin Pathol* **53**:569–72.

Thomsen JL, Helweg Larsen K, Rasmussen OV. 1984. Amnesty International and the forensic sciences. *Am J Forensic Med Pathol* **5**:305–11.

Tunca M, Lok V. 1998. Bone scintigraphy in screening of torture survivors. *Lancet* **352**:1859.

Vesaluoma M, Muller L, Gallar J, *et al.* 2000. Effects of oleoresin capsicum pepper spray on human corneal morphology and sensitivity. *Invest Ophthalmol Vis Sci* **41**:2138–47.

Watson WA, Stremel KR, Westdorp EJ. 1996. Oleoresin capsicum (cap-stun) toxicity from aerosol exposure. *Ann Pharmacother* **30**:733–5.

Zollman TM, Bragg RM, Harrison DA. 2000. Clinical effects of oleoresin capsicum (pepper spray) on the human cornea and conjunctiva. *Ophthalmology* **107**:2186–9.

CHAPTER 11

Burns and scalds

- Heat injury 312
- Classification of the severity of burns 312
- Moist thermal damage – scalds 313
- Burns from dry heat 314
- Gross appearances at autopsy 315
- Spurious 'wounds' in burns 316
- Ante-mortem versus post-mortem burns 317
- Fumes and fires 319
- Inhalation of soot 322
- Atypical localized burning and 'spontaneous combustion' 323
- References and further reading 324

Damage to the tissues arising from the application of heat is commonly encountered in forensic pathology and sometimes provides a challenging problem in the distinction between ante-mortem and post-mortem burning, which may have serious criminal aspects.

HEAT INJURY

This may arise following a defect in body temperature control or, more commonly, from the external application of heat. Mammalian tissues can survive only within a relatively narrow range of temperatures, approximately 20–44°C.

When external heat is applied, the extent of damage depends upon:

- the applied temperature
- the ability of the body surface to conduct away the excess heat
- the time for which the heat is applied.

The temperature/time relationship is important, for it is sometimes forgotten that relatively low temperatures, even as little as 44°C, can cause damage if sustained long enough. This is attested by many negligence lawsuits concerning unconscious patients burned by forgotten hot-water bottles.

The degrees of thermal damage were investigated by Moritz and Henriques (1947), who found that the lowest temperature that would cause damage was 44°C, though it required no fewer than 5 hours before a burn appeared. Only 3 seconds was needed, however, if the object was at 60°C.

Radiant heat can also cause severe damage, as is obvious from overexposure to sunlight or artificial sunlamps. The long-term effects also include malignant changes, now more important because pollution-induced changes in the stratosphere are allowing more ultraviolet radiation to reach the earth's surface. In forensic practice it is more often seen when an old or disabled person falls unconscious within close range of a gas, electric or coal fire, those areas of skin unprotected by clothing becoming burned, often to a blistered state. It can be difficult or impossible to tell how much of such a condition is ante-mortem or post-mortem, as the presence of reddening and blistering cannot always be depended on as criteria of vital infliction. Mottled pigmentation and reddening of the fronts of the legs (erythema ab igne) is a familiar sign in those addicted to sitting before an open fire.

CLASSIFICATION OF THE SEVERITY OF BURNS

An arbitrary, but useful, classification is used (mainly for surgical purposes) to denote the severity of both burns and scalds. The older six-stage classification of Dupuytren has given way to Wilson's three-stage nomenclature:

- **First degree:** erythema and blistering without loss of dermis. There is capillary dilatation and transudation of fluid into the tissues, causing swelling. A split may occur in the epidermis to form a blister with an upper cap of pale skin enclosing fluid, surrounded by a zone of hyperaemia. If small (< 1 cm) this blister may resorb, otherwise it will burst, leaving a reddened base. A first-degree burn will heal without scarring.

- **Second degree:** destruction of the full thickness of skin. The epidermis is coagulated or charred, and a central zone of necrotic tissue is surrounded by first-degree burns or a zone of hyperaemia, or both. The central necrosis sloughs in due course and the epidermis grows in from the margins. The injury cannot heal without scarring, which usually contracts during the healing process, causing puckering and distortion of the surface.
- **Third degree:** destruction of deeper tissues below the skin. This can be of any severity, from damage to subcutaneous fat to loss of muscle, bone and even a whole limb.

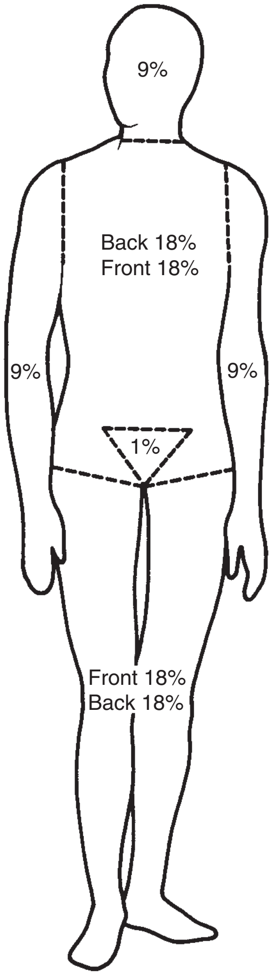

FIGURE 11.1 *The 'Rule of Nines' for calculating the area burned. It does not apply to infants whose body proportions are different from adults. About 60 per cent surface burns can be survived by children and young people up to 20 years of age but this rapidly decreases with advancing age.*

The area involved is traditionally estimated by the surgical 'Rule of Nines', which is sufficient for prognostic purposes, though for the pathologist a more precise anatomical description of the areas burned is essential for his autopsy report.

If the burns are widespread, then large areas of skin may be damaged and functionless. A large area involved may be more dangerous to life than a deeper, more localized burn. It is generally considered that 30–50 per cent involvement of the total body surface is incompatible with survival. Old people can die at considerably lower percentages than this, while children appear more resistant.

MOIST THERMAL DAMAGE – SCALDS

A 'scald' refers to tissue damage from hot liquids, usually water. Other hot fluids include oils, molten rubber, other liquid chemicals and steam. Molten metals are usually at such a high temperature that the results are similar to dry burns. The water scald is a common domestic accident, especially to children and old people, that group at the extremes of life who are vulnerable to so many types of accident.

A scald, unless from some superheated oil, does not cause charring, carbonization or singeing of surface hairs, as does dry heat. It resembles a first-degree dry burn (see later) in

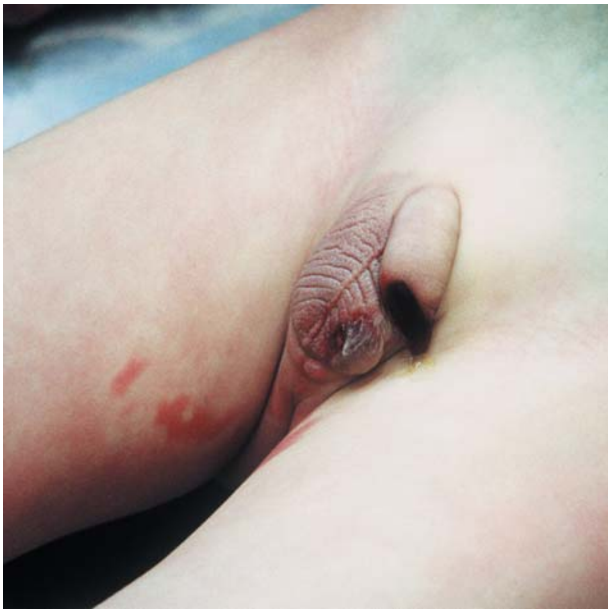

FIGURE 11.2 *First- and second-degree burns on the right thigh and scrotum of a 5-year-old diabetic child, who had been treated by a quack, who had promised the parents to cure the child by hot baths. Death was caused by the untreated diabetes.*

that there is reddening, desquamation and blistering, but the shape of the scald tends to be different. There is usually a sharply demarcated edge, corresponding to the limits of contact of the fluid. If due to immersion in hot water, such as a bath, there will be a horizontal fluid level, though this may be made irregular by splashing. When tipped or splashed, the hot liquid runs under gravity so that trickle patterns may be seen, which may enable the medical examiner to determine the posture of the victim when scalded.

The intensely red base of a severe scald may at first be covered by wrinkled, macerated epidermis. The scalded skin may swell and exude serum. Infection may supervene but where the burns are extensive, death usually results from the systemic effects of shock, fluid and electrolyte disturbance, and secondary chest infections. The severity of a scald depends on the duration of contact with the skin, as well as the temperature. When hot water is splashed, tipped or thrown, it has only a momentary contact, falling away under gravity. The large surface area allows rapid cooling, so damage must occur in a short time, which necessitates the temperature being high if severe damage is to be caused. Often the water is boiling, especially in the kitchen accidents that, sadly, occur to children with such regularity. The interposition of clothing may have two opposite effects; it may protect the underlying skin from the hot liquid, especially if poorly permeable, but it may also hold the hot liquid in contact with the skin for a longer time, especially if the fabric is absorbent.

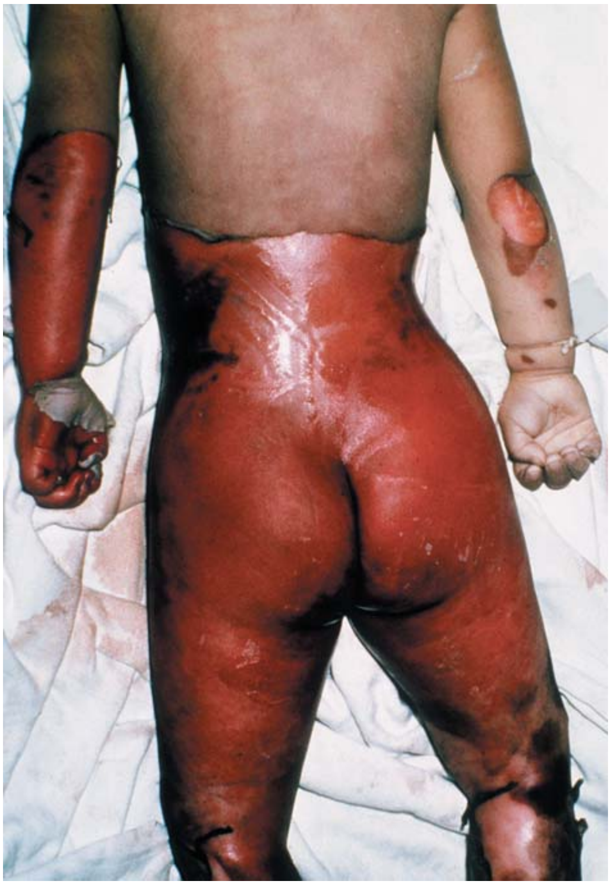

FIGURE 11.3 *Scalds or wet burns in a child accidentally left in a hot bath while her mother went to answer the telephone. This illustrates that time, as well as temperature, is a factor in causing heat damage to skin. The child died some days later from an intercurrent chest infection.*

The appearance of scalds may resemble dry burns except for the distribution and the fact that, if due to immersion, the severity of the skin damage is usually uniform over the whole of the burnt area up to the sharply demarcated margin. When hot or boiling water has been poured or splashed onto the body, the worst scalds will be at the area of initial contact where the fluid was hottest, but decrease in severity as the cooling liquid runs away. In children who pull containers of hot liquid down on themselves, the scalds tend to be on the face, neck, chest and arms, with ‘shadow areas’ of undamaged skin in the axillae and on the back.

BURNS FROM DRY HEAT

More common than scalds, burns due to dry heat may be caused by high temperature applied to the body surface by conduction or radiation. Convection is merely a variant of conduction in this respect; as hot gas impinges on the surface, the molecules transfer their high energy in a similar fashion to the direct contact of a hot solid.

Radiation causes damage through the conversion of infrared frequencies into thermal heat on absorption at the

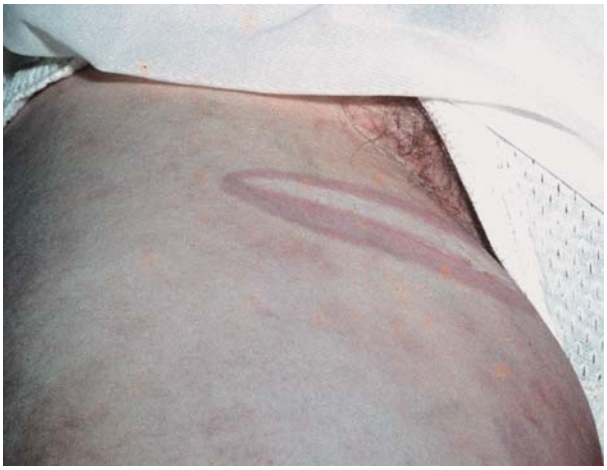

FIGURE 11.4 *A burn on the inner sides of each upper thigh in a young woman. These are suggestive of the application of a hot-water bottle in an attempt at resuscitation after a criminal abortion. The victim died of air embolism following the use of a Higginson syringe.*

skin interface. As with scalds, the tissue damage is a function of temperature and time. A warm water-pipe will cause a burn if left long enough in contact with the same area of skin, but the vast majority of dry burns are due to high temperatures acting for a shorter time.

Where the time is extended beyond that needed for initial burning, the tissues may be charred, carbonized or completely destroyed, as in cremation.

The severity of dry burns

The Wilson classification of severity of burns is given above, with three degrees according to the depth of tissue damage:

- Erythema leads to redness and swelling of the skin. Blisters may form either within the thickness of the epidermis or at the epidermal–dermal junction.
- Destruction of the full thickness of the skin, which may be total or incomplete. If epidermal structures survive, such as the deeper parts of hair follicles, then epithelialization may occur from islands within the burned area. If not, slow coverage with more probability of extensive scarring will occur from invasion of epidermis at the periphery.
- Destruction of subcutaneous tissues, such as fat, muscle and even bone, is the most serious grade of burning, though, as mentioned elsewhere, it may be less dangerous to life than larger areas of more superficial burns.

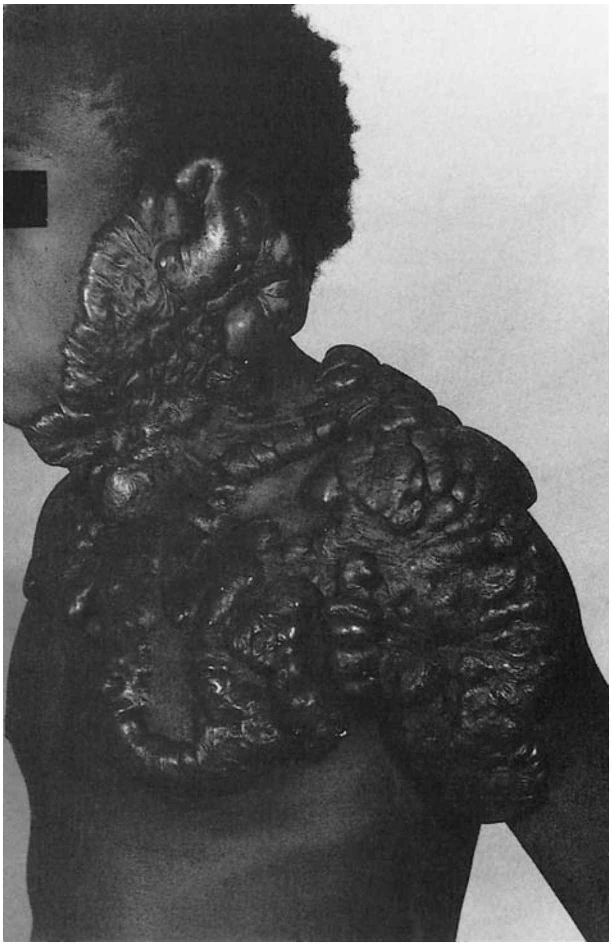

FIGURE 11.5 *Keloid tissue (dense collagenous overgrowth) more often follows burns than other injuries. It is especially prevalent and profuse in negroid persons. This East African youth was tortured in a political camp with hot metal and burning kerosene, resulting months later in exuberant keloid.*

GROSS APPEARANCES AT AUTOPSY

The body may present a wide range of damage from mere reddening over wide areas to almost total cremation, in which a search may have to be made at the scene of the fire to collect or even discover the remnants.

Usually, where leathery coagulation or charring has not occurred, ante-mortem burns will be reddened and often blistered. The latter have a marginal red zone of variable width, usually 5–20 mm across. Blisters may be present either in the main burn or as islands beyond the periphery. Most ante-mortem blisters will have a bright red base when burst and an erythematous areola. The whole of the burned area may form one large blister or be a coalescence of blisters. These are usually collapsed at autopsy, so that sheets and shreds of white epidermis lie across an angry red base.

Where the burns are more severe, the skin may be stiffened, yellow–brown and leathery, a half-way stage to actual carbonization. Drying after death of areas that were weeping serum leads to a stiff, parchment-like surface. Post-mortem heating and deposition of smoke may mimic or overlie the ante-mortem appearances, the surface often being blackened by soot deposition.

Hair is singed or completely burnt away in severe burns. In lesser degrees it may survive and the ends of the hairs may be 'clubbed'. Here the keratin melts at the distal end nearest the heat, then resolidifies on cooling, forming a terminal blob on the shaft like an unlit match. Eyelashes, eyebrows, pubic, axillary and general body hair may all suffer this scorching. Where skin has actually ignited, the subcutaneous fat acting as a fuel, black brittle masses will occupy the tissues merging into cooked, dry muscle beneath, which in turn grades into more moist and normal-looking soft tissue at greater depths. Beneath burns of any degree, the deeper tissues may be affected, especially muscle, which becomes pale, brownish and obviously 'part-cooked'. This is often a post-mortem phenomenon caused by the dead body remaining in a heated environment and such cooked muscle may

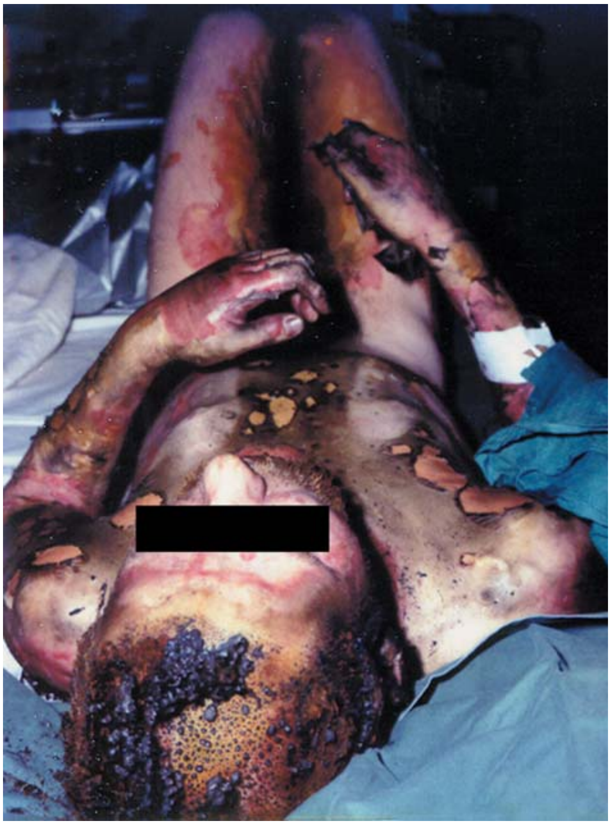

FIGURE 11.6 *Second and third degree burns and singeing of hair on a suicide who doused himself with petrol and set himself to fire.*

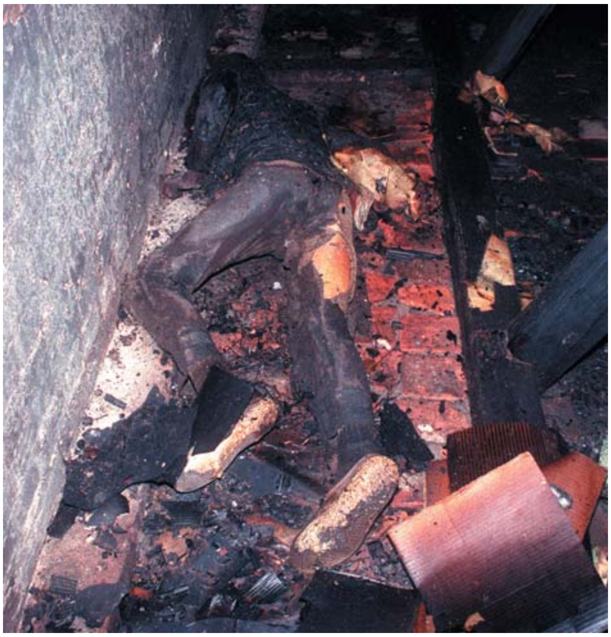

FIGURE 11.7 *Dead bodies in fires pose a problem for pathologists and investigators. Death was shown to have occurred after the fire began from soot in the lungs and carboxyhaemoglobin in the blood. There were no ante-mortem burns and there was a high blood-alcohol concentration.*

extensively underlie normal skin, which shows no evidence of ante-mortem burning, especially if protected by clothing. It is the general high temperature of the environment over a relatively long period which produces this 'parboiled' appearance, similar to the cooking effect of a slow oven.

Where heat has been intense and continuous, all soft tissue down to bone may be consumed and the bone itself may be blackened. The most extreme stage is conversion of the bone to brittle greyish-white splinters, sometimes with loss of skeletal structures, leading to absent feet, hands or limbs. It is rare for the whole body to be completely consumed, even in deliberate cremations, but only a fraction of the original body mass may remain, often almost inseparable from the burned surroundings. Where much smoke has been generated, as in most house fires, the skin will be discoloured where exposed. Sharply demarcated areas of soot deposition may coincide with burning from hot gas. Even the thinnest layer of clothing may protect completely from both these effects.

Muscle contractures are common where substantial heat has reached the body. This is almost always a post-mortem occurrence, as deep-heating effects sufficient to cook muscle are incompatible with life. The muscle becomes shortened by dehydration and protein denaturation. The flexors, being bulkier than the extensors, contract more and force the limbs into a position of general flexion, the so-called 'boxer's' or 'pugilistic' attitude. Contraction of the paraspinal musculature often causes a marked opisthotonos. Froth, often pink-stained, may appear at the mouth and nostrils as a result of pulmonary oedema caused by heat irritation of the air passages and lungs. The tongue frequently protrudes and may be scorched.

SPURIOUS 'WOUNDS' IN BURNS

Heated skin contracts markedly and splits often appear. This may lead inexperienced observers (such as police) to suspect that ante-mortem wounds have been inflicted, the fire being used to cover up a criminal offence. These splits may be anywhere, but are especially seen over extensor surfaces and joints, as well on the head. Some, especially those at elbows and knees when the limbs are flexed across brittle skin, are caused by firemen handling the body during recovery.

The possibility of true wounds being present must always be kept in mind, as many homicides have been concealed in fires. The false split will show no bleeding in the deeper tissues and its position is usually suggestive. The differentiation may be difficult or even impossible, however, especially when severe heat damage in the area makes examination of the underlying tissue impracticable.

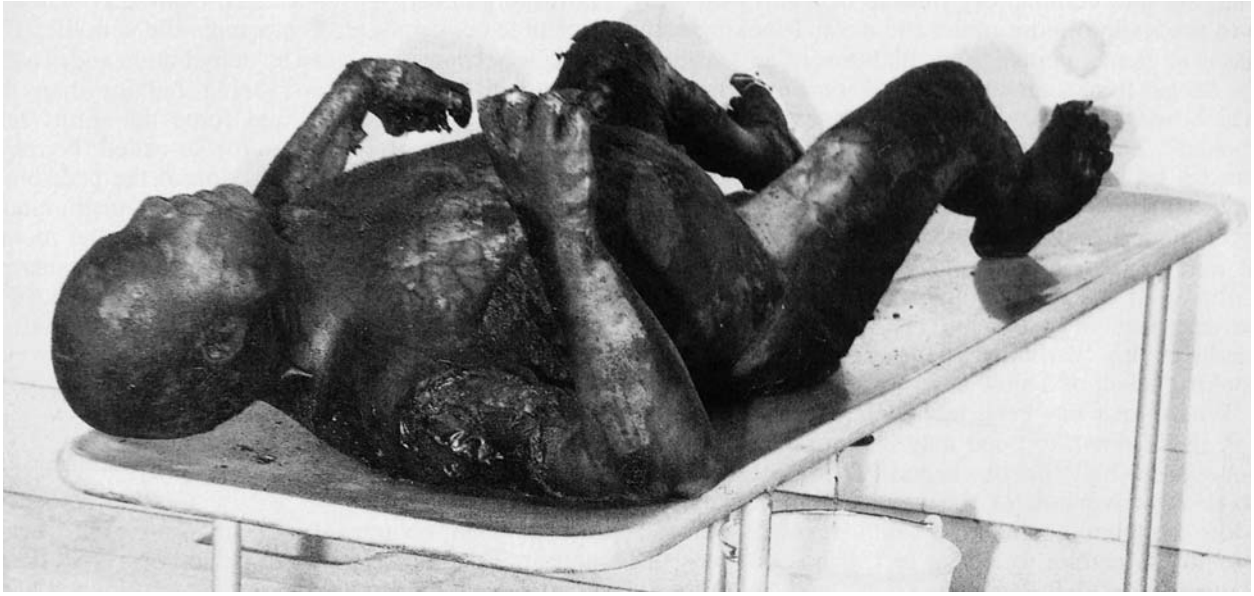

FIGURE 11.8 *Heat flexures of the limbs, part way towards the 'pugilistic attitude' formed when the arms are raised higher. The elbows, knees and wrists are strongly flexed because muscle contraction is stronger in the flexor groups. The victim was the captain of a Russian ship who set fire to his cabin during a bout of heavy drinking.*

The other major false lesion is the 'heat haematoma' in the extradural space. When severe heat is applied to the cranium, a mass of blood resembling a true extradural haematoma may form between the skull and the dura. This may arise either from venous sinuses or be virtually 'boiled' out of the diploic space in the skull through emissary venous channels. The blood is spongy from gas bubbles and tawny or chocolate brown. The exterior of the skull overlying the haematoma is usually charred and the scalp burnt away. If it is a purely post-mortem phenomenon, the heat haematoma will have a similar level of carboxyhaemoglobin to the blood, whereas if it was an ante-mortem lesion sustained before the fire began, it will be free from monoxide, as discussed in more detail below. Where the heat is intense the spurious haematoma may be outside a grossly shrunken dura, which compresses the brain into a cooked mass. The dura may split under the tension and allow brain tissue to ooze out into the large space within the cranium, where it may form a mass of frothy paste (Kondo and Ohshima 1994). These changes may be present without a haematoma, but the latter is often present. In the absence of a skull fracture, other than one due to heat, this haematoma should not be ascribed to trauma.

ANTE-MORTEM VERSUS POST-MORTEM BURNS

In severe conflagrations, either in buildings or vehicles, the terminal state of the body often does not reflect the condition at the time of death. Indeed, many deaths will have occurred before any heat reaches the body, death being caused by the inhalation of smoke. It may be difficult or impossible for the pathologist to determine the extent of ante-mortem damage if the ensuing fire later reaches the body and causes post-mortem burning. The exposed skin surface may be reddened in both ante-mortem and post-mortem burns, the classical distinction of a 'red flare' or 'vital reaction' being unsafe as an index of infliction before death. The author (BK) has seen several cases of undoubted post-mortem burns, one at least 30 minutes after fatal strangulation, where an extensive 'red flare' was caused during attempts at disposal of the body by fire.

Blisters can form post-mortem, but are pale yellow unless on scorched skin. There is rarely a red base or erythematous areola, though this sign cannot be depended upon absolutely. The contained fluid is thin and clear. Traditionally most authors claim that differentiation can be made between an ante-mortem and a post-mortem blister by an analysis for protein and chloride in the fluid. The blister formed in life is said to contain more protein and chlorides, but no absolute figures are offered and the authors have yet to meet a pathologist who does this as a routine. One suspects that the test is another of the apocryphal procedures that have been handed down from textbook to textbook without verification.

Most useful of all in fatalities is the presence of carbon monoxide in the circulating blood and carbon particles in the air passages and lungs, and this is dealt with in the section on inhalation of fumes. They do not necessarily prove that any burns were ante-mortem, but that the victim was still alive when the fire was in progress, which is not the same conclusion. The extradural heat haematoma may be investigated in a different way to determine its time of origin. If it is a true traumatic lesion produced before the fire began, it should contain no carboxyhaemoglobin. A spurious heat haematoma is formed from blood that will contain carboxyhaemoglobin if the victim has absorbed this gas during the fire. It must be appreciated, however, as

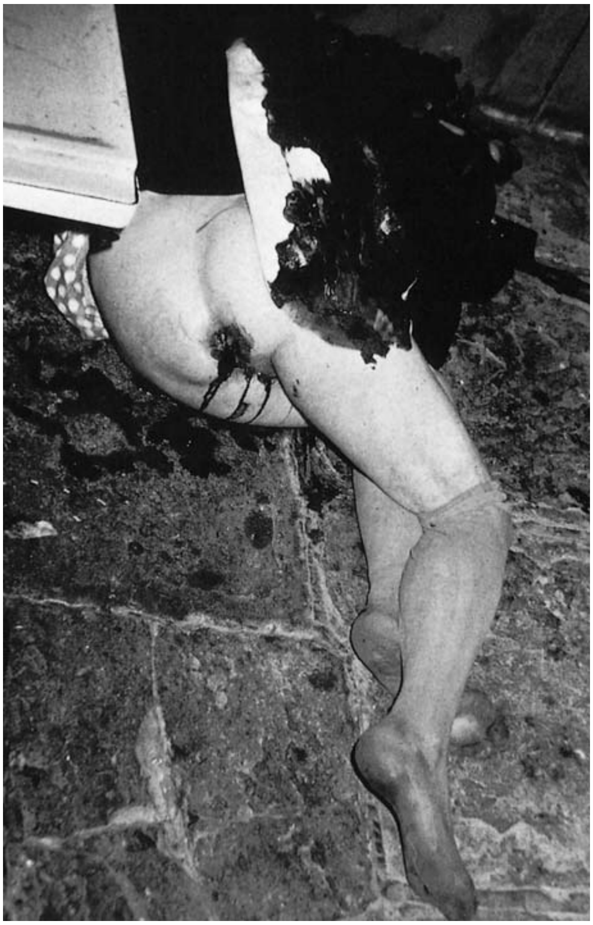

FIGURE 11.9 *Substantial destruction of the body by fire. The upper half of the victim has virtually vanished, including skeletal structures, leaving the lower part intact. The gas stove was ignited with a match after a delay while gas was escaping, causing a near-explosion that consumed the upper part of the body, the fat burning with clothing acting as a 'wick'. Such events have given rise to the myth of 'spontaneous combustion'.*

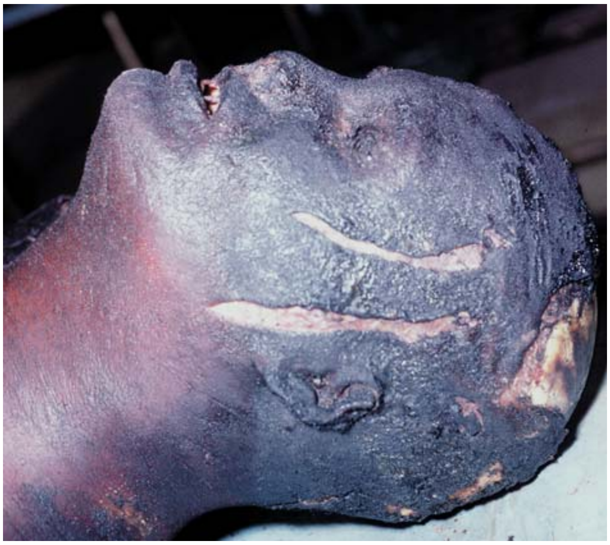

FIGURE 11.10 *Post-mortem injuries caused by heat simulating head injuries. The fire service were concerned about an apparent preconflagration assault, but such skin splits are commonly seen as a result of heat contracture of the tissues. The scalp on the top of the head has been burned through and the skull is charred. Beneath this is a spurious 'heat haematoma', sometimes confused with a traumatic extradural haemorrhage.*

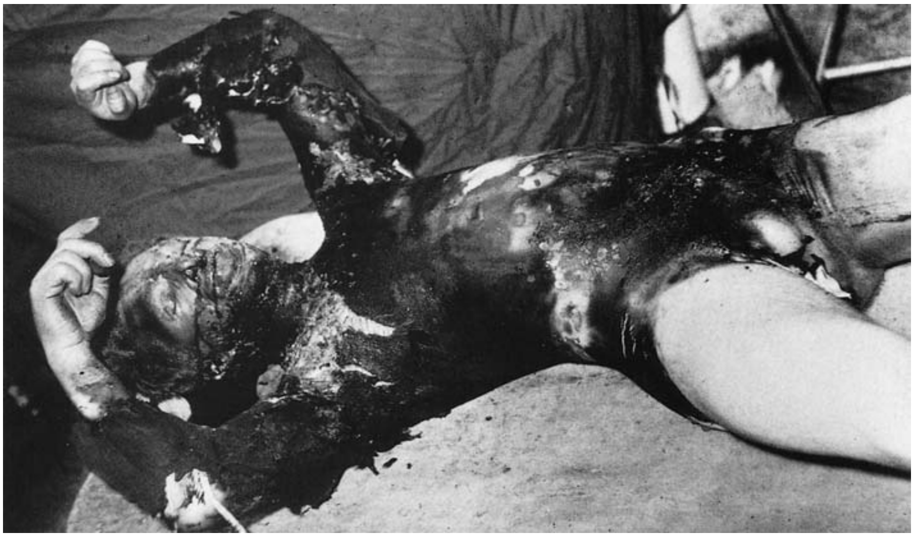

FIGURE 11.11 *Post-mortem burns in a victim recovered from a house fire. Note the raised arms caused by heat contractures and the complete protection against burning afforded by clothing on the legs. There is extensive skin splitting on the chest.*

discussed below, that not all persons alive during a fire accumulate carboxyhaemoglobin in their blood.

If there is carboxyhaemoglobin absorption, then the heat haematoma will also yield it on analysis. The use of this test, again related in most classical texts, is, however, extremely limited.

FUMES AND FIRES

The majority of fatal dry burns in civilian practice occur in conflagrations in buildings, rather than in vehicles or aircraft. In many of these tragedies death is not caused by burns, but to inhalation of fumes produced by the combustion of the building structure and contents. Indeed, most burns seen by forensic pathologists are post-mortem, either because the victim was already dead from smoke inhalation or because severe post-mortem burning obliterates the lesser degrees of burns present up to the moment of death.

Death from inhaling fumes may be caused in several ways:

- Thermal damage to the air passages and lungs from the direct effect of hot gases. At autopsy the tongue, pharynx and especially glottis may be scorched, lesser degrees causing a greyish-yellow blanching of the mucosa. The interior of the larynx, trachea and main bronchi may be thickened and blanched, or reddened and inflamed, if the temperature is too low actually to burn the lining.

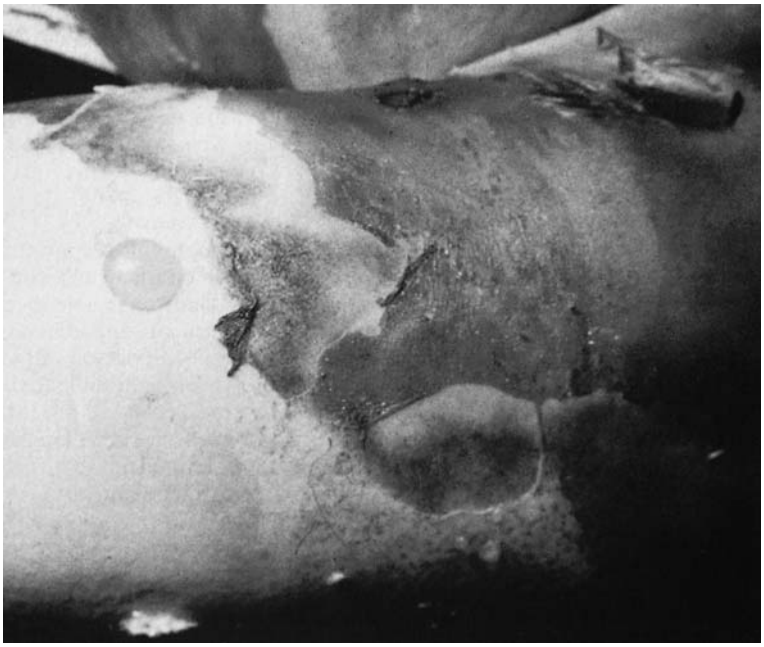

FIGURE 11.12 *It can be difficult or impossible to differentiate between ante-mortem and post-mortem burns when they are sustained near to the time of death. Here the blister with no erythema at the base could either be post-mortem or ante-mortem, as reddening of the margins and adjacent skin can occur for at least an hour after death. It is preferable to call doubtful lesions 'peri-mortal'.*

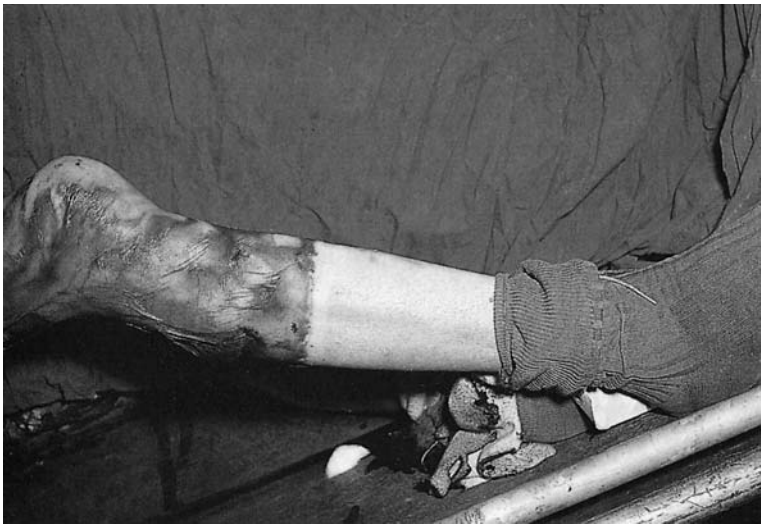

FIGURE 11.13 *Post-mortem smoke soiling and superficial burns on the leg of a house-fire victim who died of smoke inhalation. Note the complete protection of the skin, which had been covered only by a sock.*

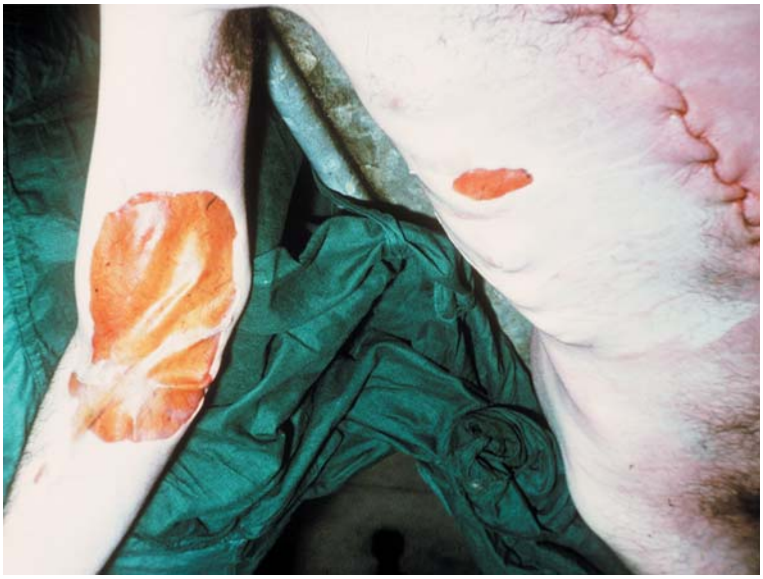

FIGURE 11.14 *Post-mortem burn on the arm and chest caused by a hot-water bottle being left over the heart in an attempt at resuscitation. The lesion is sharp-edged, there is no erythematous margin, and the surface is brown and leathery from post-mortem drying.*

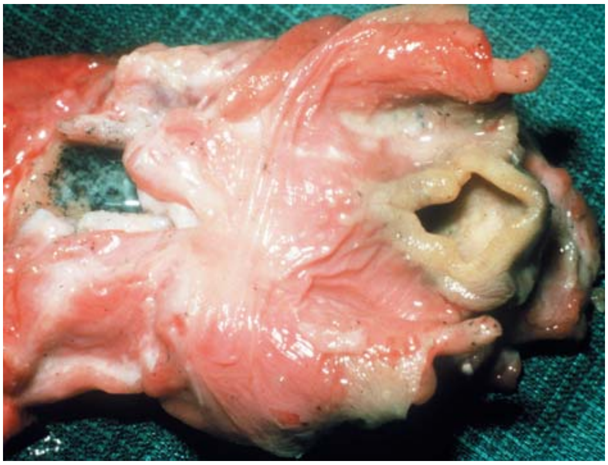

FIGURE 11.15 *Larynx of a house-fire victim showing heat blanching of the epiglottis and glottal entrance from breathing hot gas. The lower cut end of the trachea shows copious soot-containing mucus, also indicating respiration during the fire.*

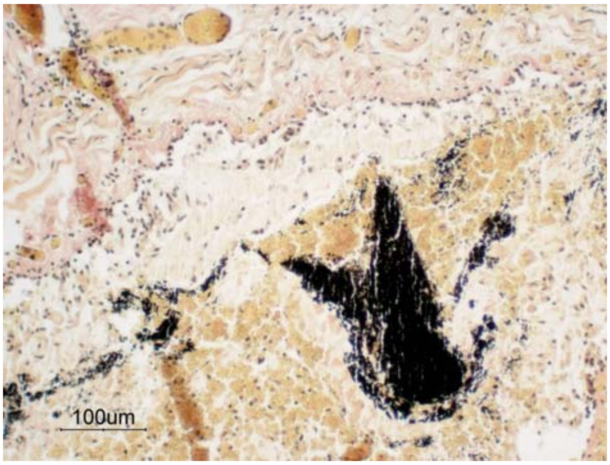

FIGURE 11.16 *A small bronchus from a fire victim, showing histological evidence of soot deposition, together with desquamation of the epithelium and plugging with cellular debris. (Van Gieson; original magnification ×10.)*

Heat effects on the pharynx and epiglottis can occur post-mortem through passive percolation of hot gas through the open mouth. The lungs usually respond to heat damage by marked pulmonary oedema, though this is often present in fire victims even when the inhalation of hot gas has been insufficient to cause visible damage to the bronchial tree.

Carbon monoxide poisoning is an important aspect of most fires – indeed it is the major or even sole cause of death in many victims of conflagrations, especially in house fires. When any combustible material burns in air, most of the carbon in organic material, such as timber, fabric and furnishings, is converted to carbon dioxide. Carbon monoxide is also produced, however, and, where the access of oxygen is limited or exhausted by the ongoing combustion, larger volumes of the monoxide are produced. Slow, smouldering fires with little flame are likely to produce more monoxide, as with burning bedclothes and mattresses. At the other extreme, rapid flash fires with flames fanned by moving draughts and those involving volatile fuels, such as petrol or kerosene, produce relatively little monoxide – though much depends upon the free access of air.

In many house fires, where the seat of the fire is originally remote from the victim, death may occur from carbon monoxide poisoning long before the flames reach the body.

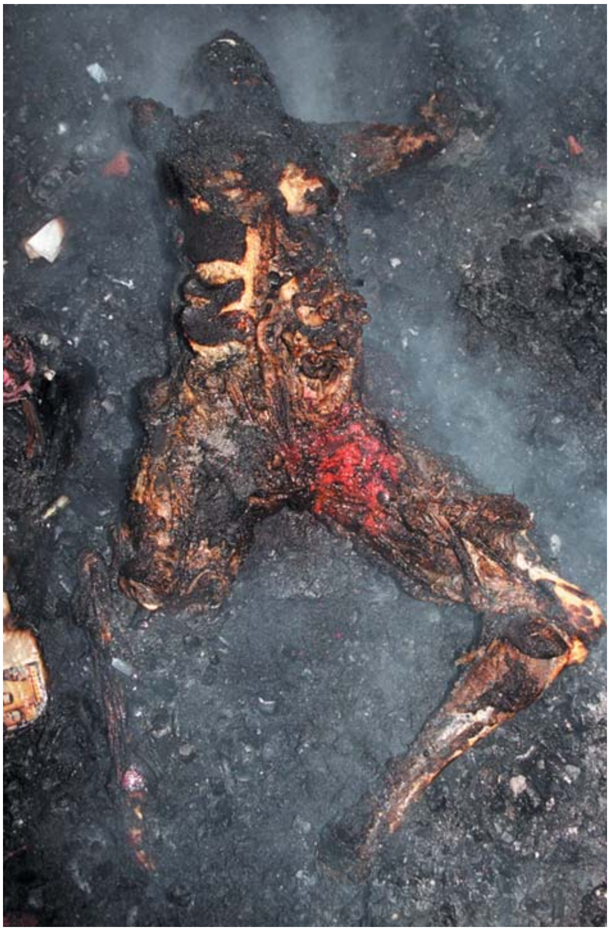

FIGURE 11.17 *Charred body at the scene of fire showing the 'pugilistic attitude' and post-mortem skin splits on the chest.*

A person may be asleep in an apartment or house, and die without ever waking if large volumes of monoxide creep through the rooms. Burns are often post-mortem in such cases and, where extensive charring occurs, the differentiation between ante-mortem and post-mortem damage may be impossible.

It is often a comfort for bereaved relatives to be told that their loved ones were either already dead or unconscious before the agony of burning reached them, and the pathologist may be able to emphasize this at any inquest or inquiry. Unfortunately, where the bodies are found in positions obviously indicating escape or concealment, such an explanation loses its conviction – though even then it may well be that coma from monoxide poisoning overtook them before burning began.

The pathology of carbon monoxide is dealt with in another section of this book, but it may be repeated here that the autopsy signs are primarily that of a cherry-pink coloration of the skin, blood and tissues. Where smoke staining or extensive charring has occurred, there may be little skin to examine, though there is usually some in a protected position under the body. The blood and tissues usually have the characteristic colour but, when the person is anaemic or exsanguinated, this may be hard to detect. A high saturation with monoxide is usually unmistakable at autopsy, though sometimes certain types of artificial light in the autopsy room, such as some fluorescent tubes, make the cherry-pink colour difficult to confirm. The colour may be better seen if blood is diluted with water against a white background, such as a porcelain autopsy table or sink. Nothing replaces laboratory analysis of a sample, however, and this must be done in

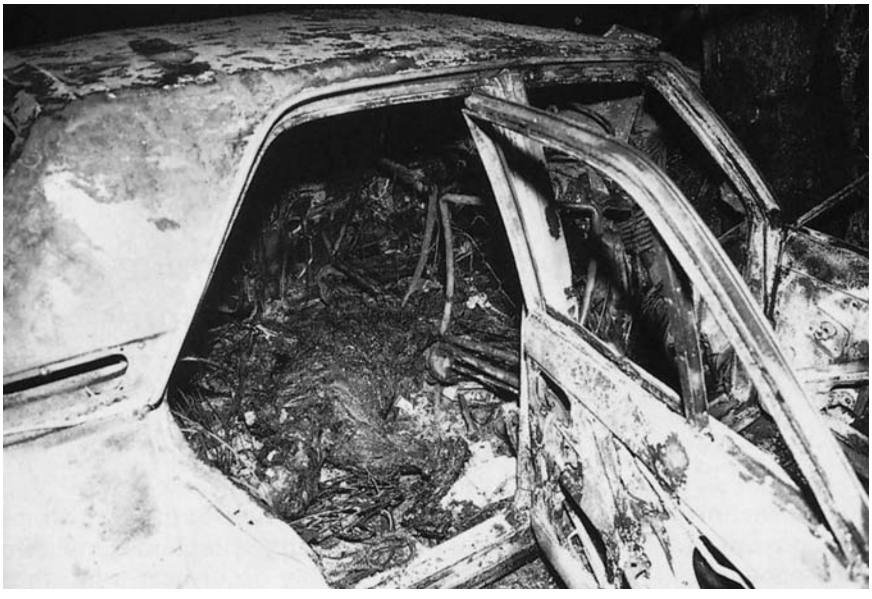

FIGURE 11.18 *The finding of a body in a burnt-out car is always suspicious, but is usually accidental or suicidal. Radiography should always be carried out to exclude firearm injuries. Carboxyhaemoglobin levels are often low or even absent in a rapid flash gasoline fire; this makes the diagnosis of being alive during the fire difficult, especially when there may also be little soot to foul the air passages.*

every case of death from fire, irrespective of the subjective appearance of the skin, blood and tissues.

The saturation of haemoglobin with carbon monoxide varies greatly from one fatal case to another. Many variable factors exist such as the concentration in the atmosphere, the time of exposure and local changes in oxygen content. When carbon monoxide is the sole cause of death, a blood saturation of at least 40 per cent is required, except in old and debilitated persons where deaths have been reported at 25 per cent. Many fatalities will display 50–60 per cent saturation, though levels in general are less than in pure carbon monoxide poisoning, such as car exhaust suicides or industrial exposure, where concentrations of up to 80 per cent may occur. In fires, additional factors such as other toxic fumes and reduction of the atmospheric oxygen may operate, resulting in death at lower carboxyhaemoglobin concentrations.

Yoshida *et al.* (1991) published data from 120 house-fire victims in Japan, in which only nine people had carboxyhaemoglobin (HBCO) concentrations below 10 per cent, the range being from 1 to 95 per cent saturation. Of 31 victims in whom cyanide analyses were made, two had high cyanide and low HBCO.

The variability of monoxide poisoning may be illustrated by the fact that two bodies lying side by side in a burned house may have widely differing blood saturations – indeed, sometimes one victim may have a zero blood monoxide saturation. Children often have much lower levels than adults in the same situation, but even two adults may differ in their carboxyhaemoglobin concentrations, the explanation being unclear. Local variations in draughts, levels above the floor and respiration rates may account for these anomalies.

Carbon monoxide in the blood is a valuable indicator that the victim was alive after the fire began. Where a dead body is disposed of by arson no absorption can take place, as the gas can gain entry only through the pulmonary interface. Thus the presence of more than a smoker's level of 5 per cent saturation of the blood with carboxyhaemoglobin means that breathing occurred after the fire began. It is vital, however, to appreciate that the converse is not true. **The fact that a body in a fire does NOT have carboxyhaemoglobin in the blood does NOT mean that they must have been dead before the fire began.** The records of all forensic departments have examples of this occurrence and therefore it is illogical – and sometimes legally dangerous – to draw the unwarranted inference that such a victim must have been already dead before the fire began, however common the generalization may be. In rapid flash fires, especially where gasoline or kerosene are involved, the monoxide level is more likely to be low or even negative, than in the slower conflagration with restricted access of oxygen that occurs in a burning building.

Toxic substances in fumes

In recent years it has been appreciated that a number of toxic substances other than carbon monoxide may be present in the fumes from fires. Cyanides are predominant and the blood of fire victims may often be shown to have high levels of this toxic compound. Nitric oxide, phosgene and other more complex substances may be liberated, especially when modern plastic polymers are burned. Furnishings, upholstery, paints, lacquers, varnishes and actual structural components are increasingly made from polystyrene, polyurethane, polyvinyl and other plastic materials. These are particularly liable to generate toxic gases when ignited (Anderson *et al.* 1981, 1982).

Caution must be employed when interpreting the analysis of cadaver blood for cyanide, as this substance can be produced in significant concentration by post-mortem decomposition. Particular care must be taken in preservation of samples, which should have fluoride added, and in ensuring there is no delay before analysis so that further decomposition is kept to a minimum. Where cyanides, oxides of nitrogen and other toxic substances are inhaled during a fire, carbon monoxide is inevitably absorbed as well. It then becomes difficult to apportion the relative blame for the death among these different toxic agents – and where there are also ante-mortem burns, the differentiation may be impossible. If the pathologist considers that he cannot identify any ante-mortem burns, then the most logical cause of death is 'inhalation of smoke'. If vital burns are present then, according to the circumstances, these can be given as the cause of death, with or without the addition of 'carbon monoxide or smoke inhalation'.

INHALATION OF SOOT

As well as the inhalation of carbon monoxide, victims in a fire usually breathe in carbon particles present in the sooty smoke. Once again this is more pronounced in a building fire than a vehicle blaze, though there are many exceptions. The combustion of timber floors, roofs, furniture, and the fabric of furnishings and carpets produces large volumes of dense black smoke. Every pathologist attending the scene of a house fire is aware of the thick layer of soot than clings to every surface and may hang in fronds from the ceilings. It is little wonder that such material suspended in the air finds its way into the respiratory passages of the victims.

As a marker of ante-mortem inhalation, it is almost as useful as carbon monoxide. Soot particles may enter the open mouth of a corpse, stain the tongue and pharynx, and may even passively reach the glottis. No significant amount can pass the vocal cords and enter the trachea after death, however, so carbon in the lower respiratory tract is a certain

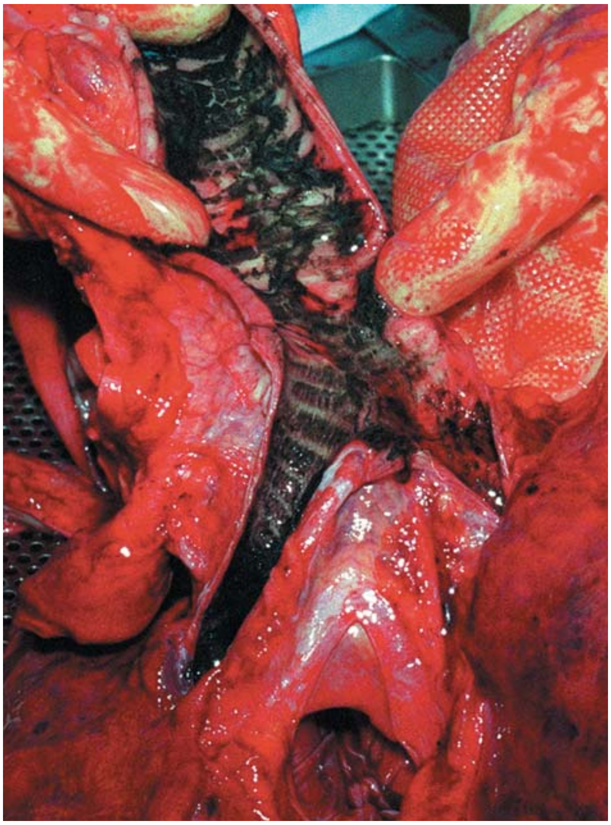

FIGURE 11.19 *Soot in the tracheal mucus, a sign of breathing while the fire was in progress. Though soot can passively reach the glottis after death if the mouth is open, it cannot reach the lower air passages in any significant quantity. Histology of the smaller bronchi offers complete proof of active respiration.*

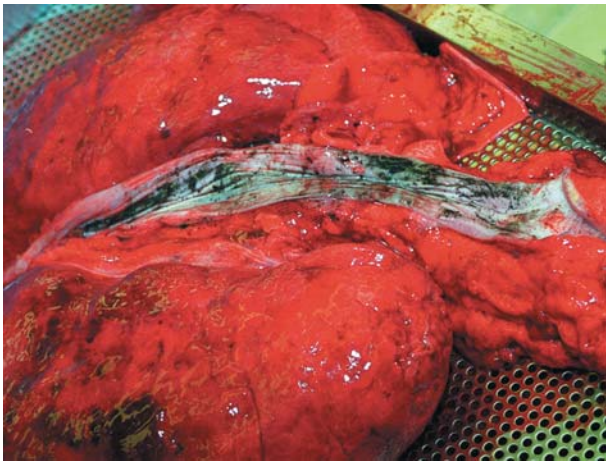

FIGURE 11.20 *Soot in oesophagus of a house-fire victim indicating respiration during the fire.*

indicator of breathing during the fire. Histological demonstration of soot in the more peripheral bronchi, out as far as the terminal bronchioles, is absolute proof of such respiratory function. The carbon is usually mixed with mucus adherent to the tracheal and bronchial walls because of heat irritation of the mucosa. Often there is swallowed soot and mucus in the stomach, and again this is evidence of life during the smoky phase of the fire.

ATYPICAL LOCALIZED BURNING AND 'SPONTANEOUS COMBUSTION'

Some very strange instances of fatal burning occur and most experienced forensic pathologists have recollections of apparently inexplicable cases.

A human body may sometimes burn away almost completely, yet the surrounding fire damage is minimal. These almost invariably occur near a hearth or open fire-grate or chimney. The burning may be confined to the body, its clothing, and a narrow zone of floor or carpet. The author (BK) has seen a circumscribed hole burned through a hearthrug and the floorboards beneath, with an ashed body lying in the space under the floor. He also remembers a pair of undamaged house slippers, with feet still within, the legs burned off at ankle level and no remaining body above, apart from ashed material lying on the adjacent floor. Both these deaths occurred in front of a fireplace with an open chimney.

It seems extraordinary that virtually a whole adult body can be consumed, including the skeleton, with minimal surrounding fire damage, but many such cases are on record and have helped to substantiate the myth of 'spontaneous human combustion'.

Experiments carried out many years ago by Professor David Gee (personal communication) indicated that body fat can burn slowly, using the clothing as a 'wick', similar to a candle flame.

Most cases have in common the fact that there was a source of ignition in an open fire and a chimney which can provide a constant updraught – and often the victims are alcoholics. Though it is not suggested that alcohol plays any part in the burning process, the confusion, instability and lack of judgement which goes with inebriation may well precipitate the ignition process.

As indicated above, the myth of 'spontaneous combustion' is one which refuses to go away, as endless unsubstantiated reports, and newspaper, magazine and television features have continued to fascinate a credulous public ever since Charles Dickens included a description of a case in his novel *Bleak House*.

Numerous allegations have been made of people bursting into flames as they walk down the street, but as yet, no photographs or film has been produced; in fact, the majority of the retrospective claims for this phenomenon have been of bodies next to some source of ignition and where alcohol was involved – as indeed was Dickens' original example.

REFERENCES AND FURTHER READING

Anderson RA, Harland WA. 1982. Fire deaths in the Glasgow area: III. The role of hydrogen cyanide. *Med Sci Law* **22**:35–40.

Anderson RA, Watson AA, Harland WA. 1981. Fire deaths in the Glasgow area: I. General considerations and pathology. *Med Sci Law* **21**:175–83.

Anderson RA, Watson AA, Harland WA. 1981. Fire deaths in the Glasgow area: II. The role of carbon monoxide. *Med Sci Law* **21**:288–94.

Ballantyne B. 1977. In vitro production of cyanide in normal human blood and the influence of thiocyanate and storage temperature. *Clin Toxicol* **11**:173–93.

Bohnert M, Rost T, Faller-Marquardt M, *et al.* 1997. Fractures of the base of the skull in charred bodies – post-mortem heat injuries or signs of mechanical traumatisation? *Forensic Sci Int* **87**:55–62.

Bohnert M, Schmidt U, Perdekamp MG, *et al.* 2002. Diagnosis of a captive-bolt injury in a skull extremely destroyed by fire. *Forensic Sci Int* **127**:192–7.

Brown RF. 1978. Injuries by burning. In: Mason JK (ed.), *The pathology of violent injury.* Edward Arnold, London.

Bull JP. 1963. Burns. *Postgrad Med J* **39**:717–25.

Bull JP. 1971. Revised analysis of mortality due to burns. *Lancet* **2**:1133.

Cattaneo C, DiMartino S, Scali S, *et al.* 1999. Determining the human origin of fragments of burnt bone: a comparative study of histological, immunological and DNA techniques. *Forensic Sci Int* **102**:181–91.

Curry AS, Price DE, Rutter ER. 1967. The production of cyanide in post mortem material. *Acta Pharmacol Toxicol Copenh* **25**:339–44.

Dyer RF, Esch VH. 1976. Polyvinyl chloride toxicity in fires. Hydrogen chloride toxicity in fire fighters. *JAMA* **235**:393–7.

Eckert WG, James S, Katchis S. 1988. Investigation of cremations and severely burned bodies. *Am J Forensic Med Pathol* **9**:188–200.

Gerling I, Meissner C, Reiter A, *et al.* 2001. Death from thermal effects and burns. *Forensic Sci Int* **115**:33–41.

Gormsen H, Jeppesen N, Lund A. 1984. The causes of death in fire victims. *Forensic Sci Int* **24**:107–11.

Hill IR. 1986. The immediate problems of aircraft fires. *Am J Forensic Med Pathol* **7**:271–7.

Karlsmark T, Aalund O, Danielsen L, *et al.* 1988. The occurrence of calcium salt deposition on dermal collagen fibres following electrical injury to porcine skin. *Forensic Sci Int* **39**:245–55.

Katcher ML. 1981. Scald burns from hot tap water. *JAMA* **246**:1219–22.

Kojima T, Nishiyama Y, Yashiki M, *et al.* 1982. Postmortem formation of carbon monoxide. *Forensic Sci Int* **19**:243–8.

Kondo T, Ohshima T. 1994. Epidural herniation of the cerebral tissue in a burned body: a case report. *Forensic Sci Int* **66**:197–202.

Levine B, D'Nicuola J, Kunsman G, *et al.* 1996. Methodologic considerations in the interpretation of postmortem carboxyhemoglobin concentrations. *Toxicology* **115**:129–34.

Lokan RJ, James RA, Dymock RB. 1987. Apparent post-mortem production of high levels of cyanide in blood. *J Forensic Sci Soc* **27**:253–9.

Matsubara K, Akane A, Maseda C, *et al.* 1990. 'First pass phenomenon' of inhaled gas in the fire victims. *Forensic Sci Int* **46**:203–8.

Moritz AR, Henriques FC. 1947. Studies of thermal injury: the relative importance of time and surface temperature in the causation of cutaneous burns. *Am J Pathol* **23**:695–704.

Noguchi TT, Eng JJ, Klatt EC. 1988. Significance of cyanide in medicolegal investigations involving fires. *Am J Forensic Med Pathol* **9**:304–9.

Quatrehomme G, Bolla M, Muller M, *et al.* 1998. Experimental single controlled study of burned bones: contribution of scanning electron microscopy. *J Forensic Sci* **43**:417–22.

Richards NF. 1977. Fire investigation – destruction of corpses. *Med Sci Law* **17**:79–82.

Rothschild MA, Raatschen HJ, Schneider V. 2001. Suicide by self-immolation in Berlin from 1990 to 2000. *Forensic Sci Int* **124**:163–6.

Sajantila A, Strom M, Budowle B, *et al.* 1991. The polymerase chain reaction and post-mortem forensic identity testing: application of amplified D1S80 and HLA-DQ alpha loci to the identification of fire victims. *Forensic Sci Int* **51**:23–34.

Schmidt P, Musshoff F, Dettmeyer R, *et al.* 2001. [Unusual carbon monoxide poisoning.] *Arch Kriminol* **208**:10–23.

Schuberth J. 1997. Gas residues of engine starting fluid in postmortem sample from an arsonist. *J Forensic Sci* **42**:144–7.

Schwerd W, Schulz E. 1978. Carboxyhaemoglobin and methaemoglobin findings in burnt bodies. *Forensic Sci Int* **12**:233–5.

Sigrist T. 1999. ['Crow's feet wrinkles' as a sign of preserved consciousness.] *Arch Kriminol* **203**:103–7.

Soares-Vieira JA, Billerbeck AE, Iwamura ES, *et al.* 2000. Post-mortem forensic identity testing: application of PCR to the identification of fire victim. *Sao Paulo Med J* **118**:75–7.

Strom CM, Rechitsky S. 1998. Use of nested PCR to identify charred human remains and minute amounts of blood. *J Forensic Sci* **43**:696–700.

Sweet DJ, Sweet CH. 1995. DNA analysis of dental pulp to link incinerated remains of homicide victim to crime scene. *J Forensic Sci* **40**:310–14.

Teige B, Lundevall J, Fleischer E. 1977. Carboxyhaemoglobin concentrations in fire victims and in cases of fatal carbon monoxide poisoning. *Z Rechtsmed* **80**:17–21.

Yoshida M, Adachi J, Watabiki T, *et al.* 1991. A study on house fire victims: age, carboxyhaemoglobin, hydrogen cyanide and hemolysis. *Forensic Sci Int* **52**:13–20.

Zhu BL, Ishida K, Quan L, *et al.* 2001. Post-mortem urinary myoglobin levels with reference to the causes of death. *Forensic Sci Int* **115**:183–8.

CHAPTER 12

Electrical fatalities

- Physical factors 326
- Electrical considerations 327
- Effects upon muscle 328
- Mode of death 329
- The cutaneous electric mark 329
- Patterned electric marks 332
- Metallic traces in electrical marks 332
- Internal appearances 333
- Histological appearances 333
- Electrocution in the bathroom 333
- Suicide by electricity 334
- Homicide by electricity 334
- References and further reading 337

The passage of a substantial electrical current through the tissues can cause skin lesions, organ damage and death. This injury is commonly called 'electrocution', though some would use this term only if death occurs. Fatalities are usually accidental, in both a domestic and industrial environment. Suicides from electricity have increased in recent years, especially in Germany. Homicide is rare but is recorded and, in the USA, electricity has again become a means of judicial execution.

PHYSICAL FACTORS

The severity of tissue damage 'including death' is directly related to a number of physical factors, which include current, voltage, resistance and time. For biological damage to occur, the body must be incorporated into an electrical circuit, so that there is a passage of electrons through the tissues. A mere accumulation of electrons in the form of a static charge can do no harm; scientists in the ball of a Van de Graaf static generator may be at a potential of more than a million volts (V), but experience nothing other than their hair standing on end. Similarly, a person outdoors in a thunderstorm may accumulate a high charge from the capacitance effect of an overhead cloud but, unless the insulation of the air breaks down to allow a lightning strike at or near the person, no ill-effects will occur.

In electrocution there must be a pathway for electrons across part of the body which, in fatal cases, contains vital structures. The current enters at one point (most often a hand being used to hold, touch or manipulate some electrical device) and then leaves the body at an exit point, usually to the earth or the neutral conductor of the electricity supply. The pathway of the current will depend mainly on the relative resistance of various potential exit points. It tends to take the shortest route between entry and best exit, irrespective of the varying conductivity of different internal tissues. If a person places a finger on a 240 V conductor while standing with damp shoes on a wet concrete floor, then an appreciable current will pass from hand to feet, with possibly fatal results. If, however, the person is standing on a carpeted upstairs wooden floor, the poor earth return will allow only a small current to flow and all that may be suffered is a painful muscular spasm.

In another variant of the upstairs scene, should the neutral wire of the supply be touching the skin of the same finger a few centimetres away from the live conductor, a severe local burn may occur but no danger to life, because the high resistance through the feet to earth will prevent any significant current flow passing through the thorax.

Should the person upstairs happen to be turning a bath tap with the other hand, the contact with the opposite finger would allow a current to pass to earth via the tap and metal water-pipes from hand to hand across the thorax – an extremely dangerous position.

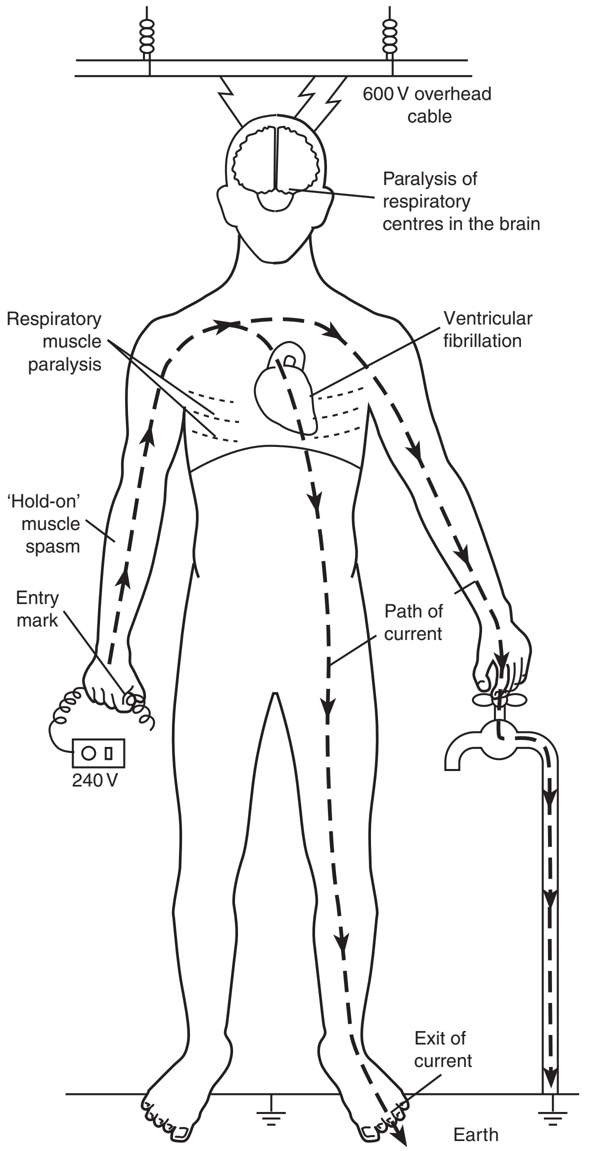

FIGURE 12.1 *Pathways of current in electrocution.*

The pathologist is concerned with fatal electrocution and three major events may occur, which are a threat to life:

- The most common is the passage of a current across the heart, usually when a hand is brought into contact with a live conductor, and the body is earthed either through the feet or the opposite hand. It has been claimed that the most dangerous is contact with the right hand and exit through the feet, as this causes the current to pass obliquely along the axis of the heart. Compared with the other variables of voltage, skin resistance and time, this hypothesis seems immaterial, though this route increases the current across the heart by a factor of 1.5–2.5 compared with left-hand entry. The fatal process is a cardiac dysrhythmia, usually a ventricular fibrillation ending in asystole.
- Less often, the passage of a current across the chest and abdomen may lead to respiratory paralysis from spasm of the intercostal muscles and diaphragm.
- Rarely, the current passes through the head and neck, usually in circumstances when the head of a worker on overhead power lines comes into contact with the conductor. In such instances, there may be a direct effect on the brainstem so that cardiac or respiratory centres are paralysed.

It is commonly said that tolerance can be gained to electric shock and that professional electricians often work on live 240 V conductors with impunity. It seems more likely that expectation of a shock decreases sensitivity, but only for brief contacts, less than would be required for physiological or structural damage.

ELECTRICAL CONSIDERATIONS

Electrons are moved around a circuit by the 'potential difference' between two points, which may be looked upon as the 'pressure' of the electricity, measured in volts (V). The number of electrons flowing constitutes the 'current', analagous to the volume of electricity and is measured in amperes, though in the present discussion, milliamperes (mA) are more relevant to biological effects. The tissues also have 'resistance' to the electrical flow. There is a mathematical relationship between potential difference, current and resistance – the well-known Ohm's law – in which the current varies directly with the voltage and inversely with the resistance. This law has considerable relevance to biological electrical damage.

Alternating and direct current

In spite of the folklore of electricians there is no doubt that direct current (d.c.) is less dangerous than alternating current (a.c.); a current of 50–80 mA a.c. can be fatal in seconds, whereas 250 mA d.c. for the same time is often survived. Alternating current is four to six times as likely to cause death, partly because of the 'hold-on' effect described below, which is the result of tetanoid muscle spasm and prevents the victim from releasing the live conductor.

Alternating current is also much more likely than direct current to cause cardiac arrhythmias. The passage of a.c. at 100 mA for only one-fifth of a second is likely to cause ventricular fibrillation and arrest. High amperage d.c. (above 4 A) may even cause an arrhythmic heart to revert to sinus rhythm, as in medical defibrillation.

The usual frequency of a.c. is 50 cycles/second (cps), though some American and European systems run at 60 cps. Alternating current between 40 and 150 cps is most dangerous in terms of ventricular fibrillation, and regrettably the usual mains supply lies in the centre of this range. Above 150 cps, fibrillation is progressively less likely as the frequency increases: at 1720 cps the heart is 20 times less likely to fibrillate than at 150 cps.

Current

The degree of damage to the tissues is proportional to the actual quantity of electricity flowing through them. This quantity is expressed by the number of electrons per unit time and strictly speaking should be measured in 'coulombs', which is the product of amperes and seconds, though amperes are usually accepted as an index of the current flow. According to Ohm's law, current depends on the applied voltage, the resistance of the tissue and, for tissue damage, the time for which the current is flowing.

In forensic pathology, we are concerned with fatal electrocution and, as most deaths are the result of cardiac dysrhythmias, the most important measure of current is that which leads to acute heart failure. Estimates vary among authors, but it is generally considered that the passage of 50–80 mA across the heart for more than a few seconds is likely to cause death. The most that can be tolerated voluntarily by most people is 30 mA applied to the hand, which results in painful muscle contractions. Consciousness is likely to be lost at about 40 mA and, as stated, currents sustained for some seconds at over 50–80 mA carry a substantial risk of death.

Voltage

To produce a potentially fatal current across the chest, a certain minimum voltage must be applied to the skin surface. As the latter has a relatively high resistance (see below), Ohm's law dictates that an appreciable voltage is required to produce the 50 mA or so needed for ventricular fibrillation. Most fatalities occur with the domestic voltage of 240, though the common alternative in parts of the USA and Europe of 110 V is still lethal in many instances.

It is uncommon to encounter deaths at less than 100 V, mainly because there are few sources of supply between 110 V and the 12 V or 24 V used in vehicle electrical systems. These are almost always innocuous, though Polson (1963) has reported a fatality at 24 V in a man pinned beneath an electrical vehicle for several hours. That case emphasizes the importance of the time element in electrical injury. Extremely high voltages, such as those encountered in power transmission systems and in electronic equipment, may paradoxically be safer on some occasions, as the shock may physically fling the subject off the conductor, thus reducing the contact time below the threshold for cardiac damage.

Resistance

The major barrier to an electrical current is the skin, which has a far higher resistance than internal tissues. That is why skin electric burns occur, as the resistivity causes energy transfer from the electron flow to the skin. Once inside the dermis, the semi-fluid cytoplasm, and especially the vascular system filled with electrolyte-rich fluid, passes the current through the body quite easily. The resistance of skin varies greatly according to the thickness of the keratin-covered epidermis; that on the soles and finger-pads is greater than the thin skin elsewhere. The average resistance is between 500 and 10 000 ohms for areas other than the horny hand and foot pads, which may offer 1 million ohms resistance when dry.

A more potent factor is the dryness or dampness of the skin, which greatly affects its resistance. While dry palm skin may have a resistance of the order of 1 million ohms, when wetted this may fall to only 1200 ohms. Jellinek (1932) found the horny skin of a workman to have a dry resistance of from 1 to 2 million ohms; Jaffe (1928) stated that sweating could reduce skin resistance from 30 000 to 2500 ohms. When the current begins to pass, there is a further marked drop in resistance, as a result of electrolytic changes in the skin, which may fall to only 380 ohms. Thus for a fixed voltage, such as the mains supply of 240 V, the resultant current will be far greater if the skin is wet from sweating or external moisture. This emphasizes the dangers of bathrooms and using electrical equipment in damp surroundings. Where alternating current is concerned (as in most forensic cases), resistance is replaced by 'impedance', but this is not relevant to the pathological aspects.

EFFECTS UPON MUSCLE

One effect of electricity that has practical implications is the spasm that occurs in skeletal muscle if the current reaches between 10 and 40 mA at 50 cps. When, as most often happens, the entry point is in the hand, the stronger flexor muscles of the arm go into spasm and cause a 'hold-on' effect. This means that any object being held in the hand is involuntarily clenched and, as it is likely to be the faulty electrical appliance or a wire, the object cannot be released and the current thus continues to flow. This adds to the time element and progressively worsens the risk of both cutaneous burns and the risk of cardiac or respiratory arrest. Tingling may be felt in the skin with a current of only 1 mA and 'hold-on' can begin at a current as low as 9–10 mA.

MODE OF DEATH

As already stated, most deaths from electricity are from cardiac arrhythmias, usually ventricular fibrillation ending in arrest. This is caused by the passage of current through the myocardium, especially in the superficial epicardial layers and possibly across the endocardium. The current has a profound effect directly upon the myocardial syncytium, the possible dislocation of the pacemaking nodes and conducting systems being ill-understood. When death occurs from cardiac arrest, the body remains either pale or only slightly congested, the autopsy appearances being unhelpful apart from the presence of any external electrical marks.

The second (and far less common) mode of death is respiratory arrest, in which the passage of current through the thorax causes the intercostal muscles and diaphragm to go into spasm, or become paralysed. In either case, respiratory movements are inhibited and a congestive–hypoxic death occurs. The brainstem is affected rarely, when the current enters through the head. Either cardiac arrest or respiratory paralysis can then supervene.

Many electrical deaths are not observed, the person being found dead later, so that the mode of death is unknown. Sometimes the witnessed mode of death is hard to explain on physiological grounds, in that there appears to be a delay (often some minutes in duration) between the shock and the death. In the interval, the victim may be conscious and even apparently recovering. It is difficult to know why a sudden cardiac arrest should take place after the current is switched off, but presumably some fundamental damage has been caused on an intracellular level to cardiac or neural tissue.

Finally, when discussing the mode of death, it must always be remembered that non-electrical trauma is quite common. In a series reported by Bissig (1960), some 15 per cent of cases had injury from falls and other associated trauma. In industrial accidents and when working on power lines, victims of shock may be thrown from a height, or suffer violent muscular spasms that may lead to fractures and other serious injury.

THE CUTANEOUS ELECTRIC MARK

The point of contact on the body surface may leave skin lesions, which are either called 'electrical burns' or 'electrical marks', though the term 'Joule burn' is gaining in popularity. These are the sites of entry of the current, but another mark or marks may also appear where the body was earthed or 'grounded'. It must be emphasized here that fatal electrocution may occur with no skin mark whatsoever, making the diagnosis entirely dependent upon the circumstances of the death. An extreme example is electrocution in the bath, when the large surface area for entry plus the low skin resistance caused by the water abolishes the possibility of a focal burn.

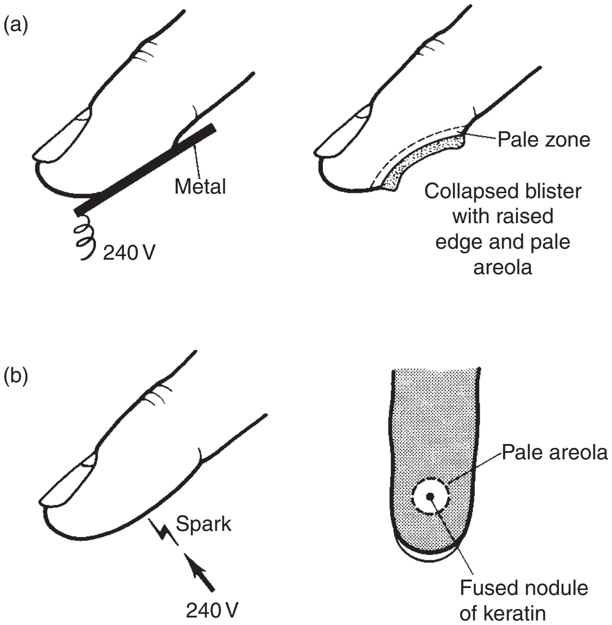

FIGURE 12.2 *(a) Firm contact electric mark, and (b) spark burn across air gap.*

An electrical mark may not be externally obvious, as the current can be applied to the genitals, anus or abdomen in sexual perversions or through the mouth, especially in children. Infants may place a live plug between their lips and sustain electrical burns on the tongue or buccal mucosa, which may not be readily apparent on external examination at autopsy.

When a current passes, there may or may not be a visible lesion, depending upon:

- the density of the current passage in terms of skin area and
- the conductivity, usually varying with the moisture content.

The skin lesion is a thermal burn from heating of the epidermis and dermis as the current passes. Theoretically, the heat generated can be determined from the formula $GC = C2R/4.187$, where GC is the heat in gram calories per second, C is the current in amperes and R the resistance in ohms. If the electron flood passes through a relatively wide area, then the resistance per unit area is small (especially if the skin is damp) and thus the heating effect is proportionately diminished. For example, a person resting the palm of the hand on a flat metal plate which is electrified will pass a far

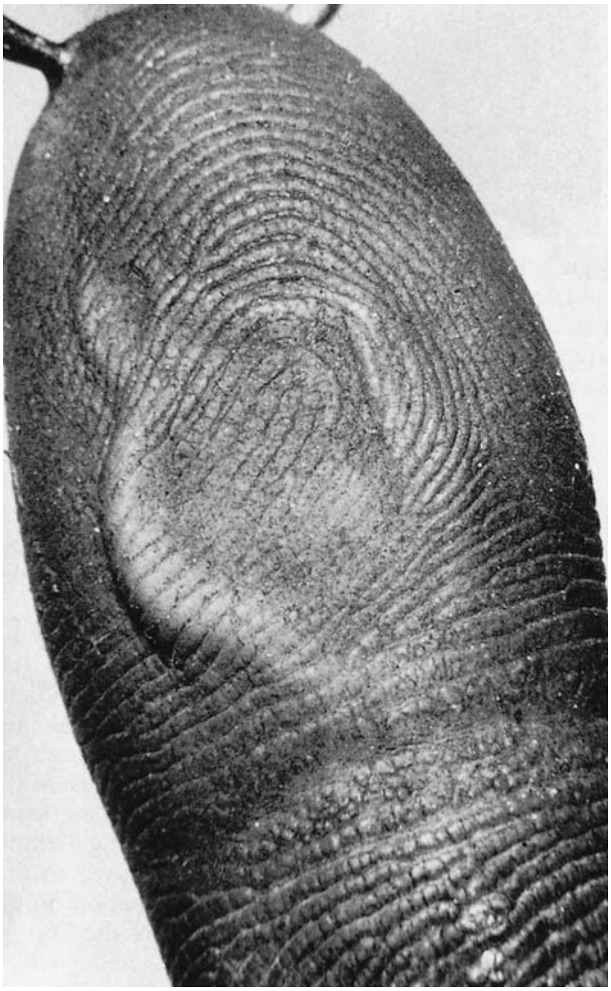

FIGURE 12.3 *Firm contact electric marks on a fingertip. Two blisters have been formed by the heat of local skin resistance; they have collapsed on cooling.*

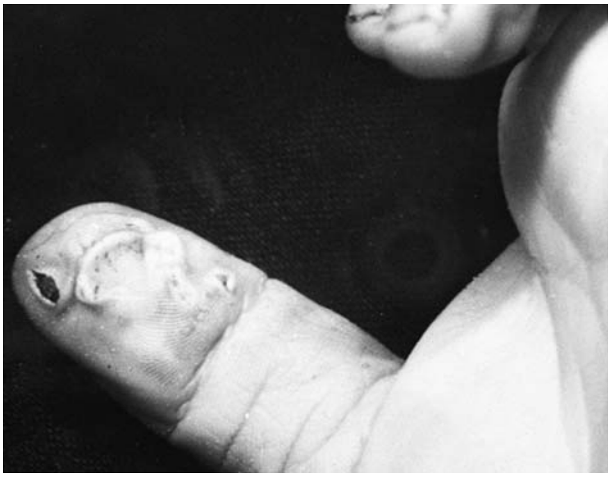

FIGURE 12.4 *Firm contact blister and adjacent spark burn from 240 V mains supply. The victim was holding a faulty electric drill. There were no other marks on the body.*

smaller current per square centimetre of skin than another touching the plate with the tip of a finger. The first is likely to have no demonstrable lesions, while the second will have a blistered burn or a keratinized nodule, depending on the firmness of the contact.

The temperature in the tissues directly under the contact point can easily reach 95°C. Tissue damage can occur within 25 seconds when the temperature reaches a mere 50°C. The focal electric mark is largely a thermal burn and even some of the histological features to be described later, which were once thought to be peculiar to the electrical current, are now generally ascribed to thermal effects. There are some features, however, which are characteristic of an electrical cause.

- When the skin has been in firm contact with an electrical conductor, the passage of the current through the high skin resistance heats up the tissue fluids and produces steam. This may split the layers of the epidermis or the epidermal–dermal junction and produce a raised blister. This may rupture if the current continues or if the area is relatively large.

  When the current ceases, the blister cools and collapses, giving the familiar appearance seen at autopsy. The collapsed blister is often annular, producing a raised grey or white ring with an umbilicated centre. The mark sometimes reproduces the shape of the conductor, especially where this is a linear wire or a shaped metal object. Where the tip of a wire or rod is at right angles to the skin, a focal pit is created, sometimes penetrating quite deeply into the skin.
- Where the contact is less firm, so that an air gap (albeit narrow) exists between skin and conductor, the current jumps the gap as a spark. In dry air 1000 V will jump several millimetres and 100 kV about 35 cm. This is at extremely high temperatures (about 4000°C), as in the sparking plug of a petrol engine, and causes the outer skin keratin to melt over a small area. On cooling, the keratin fuses into a hard brownish nodule, usually raised above the surrounding surface, the so-called ‘spark lesion’.
- In many electrical burns these two types are combined, as a result of movement of the hand or body against the conductor and sometimes because of irregularity of the shape of the conductor. Where the time has been prolonged, the voltage is high, or the conductor is large, the burn may be correspondingly severe with large areas of peeled blistered skin, charred keratin, and a mixture of hyperaemia, deep scorching and shed epidermis.

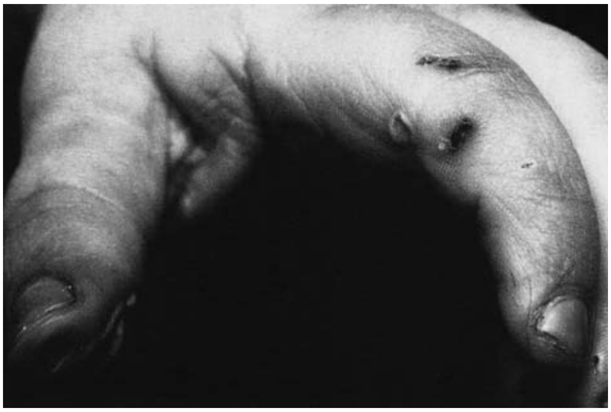

FIGURE 12.5 *Collapsed blisters and a spark burn from holding a faulty power tool.*

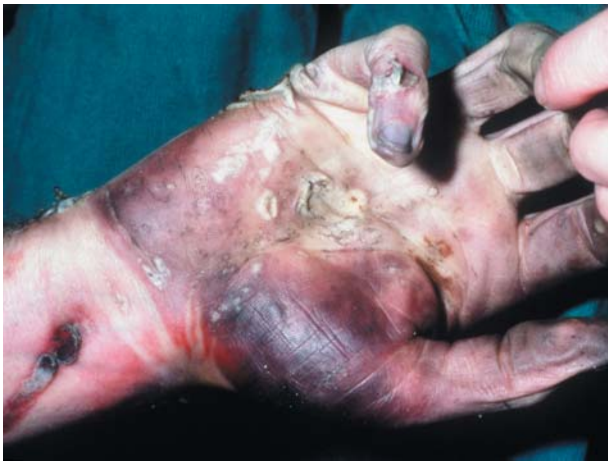

FIGURE 12.6 *Multiple electrical burns on a hand, with blisters showing the typical pale raised margins and areas of peeling epidermis. There is a blackening from metallization as the current was passing for several hours; the victim was an electrician who fell into an air-conditioning plant.*

- A characteristic feature of the electric mark, which is the most useful indicator of the nature of the lesion, is the common occurrence of an areola of blanched skin at the periphery. Presumably because of arteriolar spasm from the direct effects of the current on vessel-wall musculature, the pallor survives death and is virtually pathognomonic of electrical damage. Often there is a hyperaemic border outside the blanching, though reddening may also be seen inside the pale zone, as the outermost rim of the heated burn area. Occasionally, an alternating spectrum of blister-reddening-pallor-reddening can be observed centrifugally from the centre of the lesion. When the burn is linear, such as that from a bare wire pressed against the skin, the 'areola' takes the form of a pale zone parallel to the central burn.
- When the marks are minimal, all that may be seen are tiny white discs representing minute flat blisters where the epidermis has split, but no reddening or areola has formed – or at least, has not persisted until the time of autopsy. These can be hard to find and, as they are likely to be present on the palmar surface of the hands (the usual site from grasping an electrical appliance), the strong flexion of rigor mortis may bring the fingers down to the palms, so obscuring any lesions. It is essential in all autopsies to examine the flexor surface of the fingers by forcible breaking of the rigor – and, where electrocution is a possibility, minute inspection of the hands must be carried out, even if the flexor tendons at the wrist have to be cut to release the rigor clenching of the fingers.
- In high-voltage burns, such as those sustained from high-tension grid transmission cables, where the voltage is in the multi-kilovolt range, sparking may occur over many centimetres. This can cause multiple spark lesions giving rise to a 'crocodile-skin' effect. Both linesmen and copper thieves working on high pylons may suffer non-electrical injuries from being thrown to the ground, or they may sustain gross charring burns or even limb fractures from both direct electrical energy and the muscle spasm caused by a massive bolt of electricity.
- Earthing or 'grounding' lesions are not often seen, but should be looked for on the contralateral hand and on the feet. The lesions may be similar, though less severe. In a case seen by the author (BK), in which a workman pushed a metal wheelbarrow over a live cable lying in a pool of water, the entry burns were on the hands, but on the soles of both feet the electrical burns matched the metal studs in his boots where the current earthed through damp socks and leather.

  In another case, a homicide in a bath, there were no entry burns but the current went to earth by contact of the breasts with the metal taps.
- When the current has flowed for an appreciable time, even at a domestic voltage of 240, the effects may be severe. Charring and more extensive peeling and blistering of skin may occur, with deep muscle damage and cooking of the tissues. This is partly because of the fact that the initial damage lowers skin resistance so that progressively more current flows and therefore more burning and necrosis follows.

Much of this damage seen at autopsy may have occurred post-mortem, if death from cardiac arrest occurs early in the event and, where the victim is alone, there is no one to remove his body from the source of the current. It has been

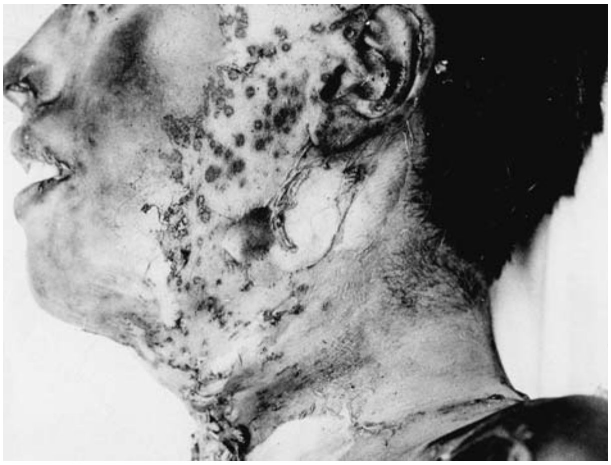

FIGURE 12.7 *Multiple burns from high-voltage (multi-kilovolt) supply lines of the grid system. The appearance is sometimes called 'crocodile skin' and is caused by arcing of the current over a considerable distance. The victim was a boy who went bird-nesting amongst the switch gear of a power station.*

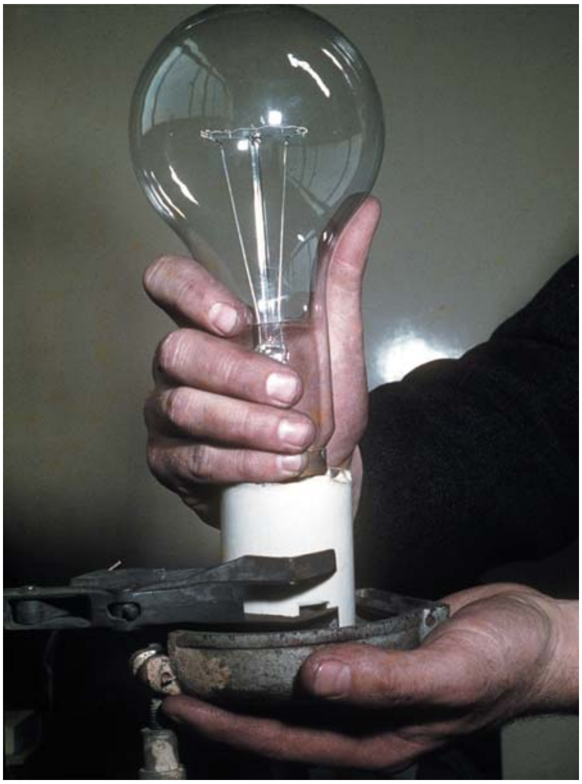

FIGURE 12.8 *Reconstruction of a fatal electrocution. While replacing a bulb a workman dropped dead, even though the light was not illuminated. The circuit had been improperly 'wired through' – that is, the switch had been placed in the neutral wire, instead of the live. The porcelain holder was broken so that his little finger touched the brass base of the bulb, thus completing the circuit across the chest to the other hand, which held an earthed base-plate. The only electrical mark was on the little finger.*

conclusively shown by Polson and others that post-mortem burns can be inflicted on a dead body, the appearances being similar in terms of blistering and burning, though the red flare of 'vital reaction' will be absent if death has occurred some time before.

PATTERNED ELECTRIC MARKS

As with many other injuries, a pattern of the shape of the causative object may sometimes be identified. When the electrical conductor is a wire, a linear burn may occur. The shape and spacing of electrical plugs or contacts may be seen, and faulty electrical equipment may impress its shape into the skin. These marks may be useful when the pathologist tries to reconstruct the events – and indeed, may indicate for the first time that the death was caused by electricity, when the unwitnessed events are particularly obscure.

As electric shock and burns are commonly seen in torture victims in abuse of human rights, a pattern may be useful in determining what object was used and give evidence of deliberate repetition. The deliberate use of a male-pin plug connected to the mains supply, then pressed against the skin, may give a series of regularly spaced marks, consisting of hyperaemia, blistering, areola formation or even charring.

METALLIC TRACES IN ELECTRICAL MARKS

When a current passes from a metal conductor into the body, a form of electrolysis occurs so that metallic ions are embedded in the skin and even in the subcutaneous tissues. This occurs with both d.c. and a.c. because of the combination of metallic ions with tissue anions to form metallic salts. These may be invisible to the eye, but detectable by chemical, histochemical and spectrographic techniques. They persist for some weeks during life and resist a moderate amount of post-mortem change. Where gross they may be observed directly on the skin, and where copper or brass conductors are involved a bright green imprint may be obvious. Where an electric arc forms, vaporized metal may be deposited on the skin, often extensive enough to be visible to the naked eye. In high-voltage contacts, the skin of a wide area may be brown or greyish, partly from heat effects, but partly from metallization.

Recently, using the scanning electron microscope, it has become possible to see tiny globules of molten metal on the skin at and near an electrical mark, deposited there by the almost inevitable 'mini-arcing' that accompanies any electrical contact. Chemical tests for metallic deposits include that devised by Adjutantis and Skalos (1962), which is a simple touch-test using elution on strips of filter paper. Copper, iron, aluminium, zinc or nickel are dissolved in nitric or hydrochloric acid, and the solution tested with a variety of simple but specific reagents.

INTERNAL APPEARANCES

In fatal electrocution, gross findings in the internal organs may be absent and even histological changes are a matter of controversy. Because the internal tissues are largely aqueous and contain conductive electrolytes, the current pathway is usually too diffuse to cause thermal damage. The absence of visceral damage is made up for by physiological and functional abnormalities, mainly in muscle and nervous tissue.

The usual mode of death is cardiac arrhythmia leading to ventricular fibrillation and arrest. In these deaths there is little to find at autopsy apart from the skin lesions. It is claimed that epicardial petechiae may occur, but these are too non-specific to be of any use. The body is either pale or only slightly congested, in contrast to the **few** deaths that occur from respiratory paralysis.

In the latter group the intercostal muscles and diaphragm go into spasm or are paralysed, which leads to marked congestion and cyanosis of the face, with similar changes in the lungs. There may be some petechiae on the pleura, though this again is such a non-specific finding as to be unhelpful as a diagnostic sign. At autopsy, the usual signs of a congestive death are found, with dark blue–red post-mortem hypostasis.

A number of other signs of electrocution have been claimed, but they are rarely confirmed by personal experience. These include intracerebral petechiae, which are probably part of the general congestive state in respiratory paralysis.

HISTOLOGICAL APPEARANCES

These are rather controversial, as some changes formerly claimed to be specific for electrical lesions have been shown to be thermal.

The skin mark consists of vacuolation in the epidermis and sometimes dermis, caused by the gas spaces from the heated tissue fluids splitting the cells apart. The affected tissues become more eosinophilic. The cap of epidermis may be detached and raised into a blister, with a large space beneath. The cells of the epidermis are often elongated, with the nuclei of the lower layers orientated and horizontally stretched; this was once claimed to be an electromagnetic effect, but the same appearance can be seen in purely thermal burns and in hypothermic lesions.

Changes in the brain have been described, even when the cause of death was cardiac. Focal petechial haemorrhages, spaces around small blood vessels and tears in white matter have been described. Electron microscopy reveals a variety of changes, especially in the nuclei of skin cells, which are deformed with clumped chromatin. Janssen (1984) has compiled a review of electrical histological lesions in his book on forensic histology, but it seems evident that there is little that is absolutely pathognomonic of electrical as opposed to purely thermal burns.

Histochemical reactions for metallization can also be carried out, though, again, metals can be transferred into the skin by purely thermal means if hot metallic objects are pressed against the skin.

Scanning electron microscopy appears to be the best way of distinguishing between electrical and thermal damage, as the punctate nature of the deposition and the possibility of chemical analysis by electron microprobes offer a sophisticated means of identifying the metallic deposit. Internal organs reveal no true diagnostic lesions of electricity. The wavy appearance of the myocardial fibres and their fragmentation may be suggestive, but by no means diagnostic. Contraction bands within the fibres, especially of a ‘bark-like’ appearance, have been described but, again, are non-specific, though they are often seen in the subepicardial myocardium after resuscitation that includes electrical defibrillation. However, they are identical to catecholamine effects on the myocardium and so may not be electrical in origin.

ELECTROCUTION IN THE BATHROOM

The bathroom is a common site for electrical tragedies. Accidents, suicides and even homicides occur there because of its vulnerability to electrical shock. The bathroom is electrically the most dangerous place in a house as there is a moist atmosphere, ample water, good earthing through metal taps and pipes, and a wet, unclothed body – all of which are conducive to low electrical resistance and hence high tissue currents. Accidents are common, usually as a result of the careless use of electrical appliances, such as hairdryers and room heaters. Most European countries have strict regulations about electrical installations in bathrooms because of the danger. In Britain, no wall switch is permitted for the light, a ceiling switch operated by an insulating cord being mandatory. No power sockets are allowed except for a shaver socket which incorporates a step-down transformer

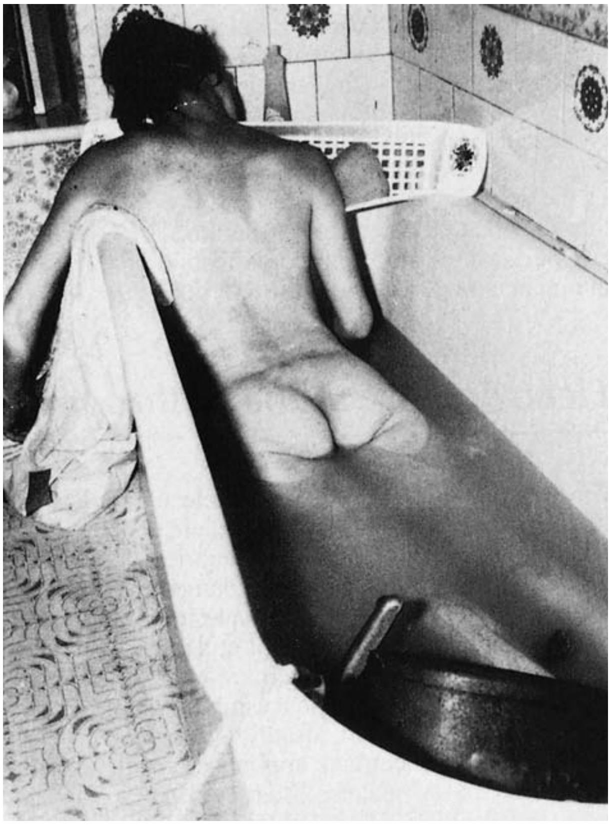

FIGURE 12.9 *Homicidal electrocution in the bath. An electric heater was dropped into the water with the wire to the earth terminal deliberately disconnected. The 240 V supply entered the body diffusely through the water and completed the circuit by earthing through contact with the breasts against the metal taps, leaving electrical marks on the skin.*

with a low maximum current output. In spite of this, people unwisely use extension leads or plug into light sockets to operate a variety of appliances.

SUICIDE BY ELECTRICITY

There has been a remarkable increase in the number of suicidal electrocutions in recent years, especially in the bath. Bonte *et al.* (1986) have described an increasing number of cases in the former West Germany, drawing attention to the unusual demarcation of post-mortem hypostasis in such deaths, where the water level during electrocution appears to limit the subsequent appearance of the hypostasis. Various circuit arrangements are made by suicides in baths: these range from the pulling of an electrical appliance into the water to complex connections to metal soap dishes and wiring up contacts on the body. There may even be arrangements to break the circuit when the bathroom door is opened to safeguard those who find the body. The circumstances are rarely in doubt except that some less sophisticated cases may be difficult to differentiate from the many other accidents that occur in bathrooms. Rarely, other more bizarre deaths associated with electricity occur, as seen in Figure 12.11.

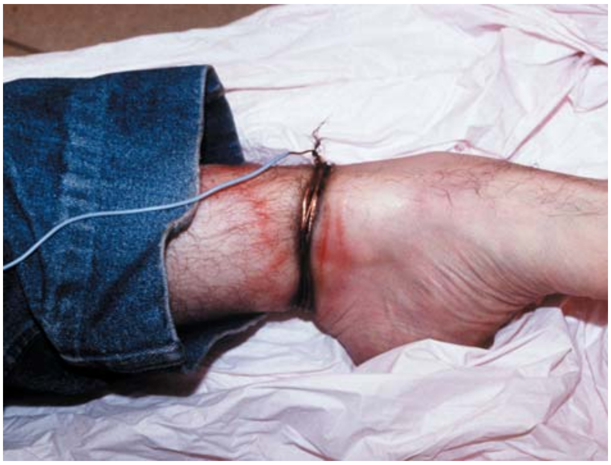

FIGURE 12.10 *Suicide by electricity. Both wrists were wired in this fashion to a 240 V supply socket, the circuit being completed from arm to arm across the chest.*

HOMICIDE BY ELECTRICITY

Homicide is occasionally committed by electricity, the author (BK) having dealt with two such incidents and his co-author (PS) with one. The first took place in the bath and provided good illustrations of the electrical aspects of fatal shock. A young woman was found dead in her bath, slumped forwards in a kneeling position with one breast resting against a chrome tap. The left arm was trailing over the edge of the bath, in which was the usual level of water plus an electric fan-heater immersed near the feet. The heater was connected by a long cable to a 240 V 13 A socket in an adjacent bedroom. Significantly, the third ‘earth’ wire had been deliberately disconnected from the earth pin inside the plug. Subsequent forensic tests showed that the enamel lining of the metal bath was an excellent insulator so that earthing of the current through the bath water could only occur via the chrome waste pipe. The circuit was completed, however, with fatal results, by the woman falling (or being pushed) against the tap, so that she sustained an electrical burn on the breast. She also had another typical burn on the inside of her left arm near the axilla, where the insulating enamel of the bath ceased at the turned-over edge. Forensic tests showed that there was a progressive gradient of voltage down the bath from over 200 V near the heater to virtually zero near the earthed waste pipe. The current had preferentially taken a course

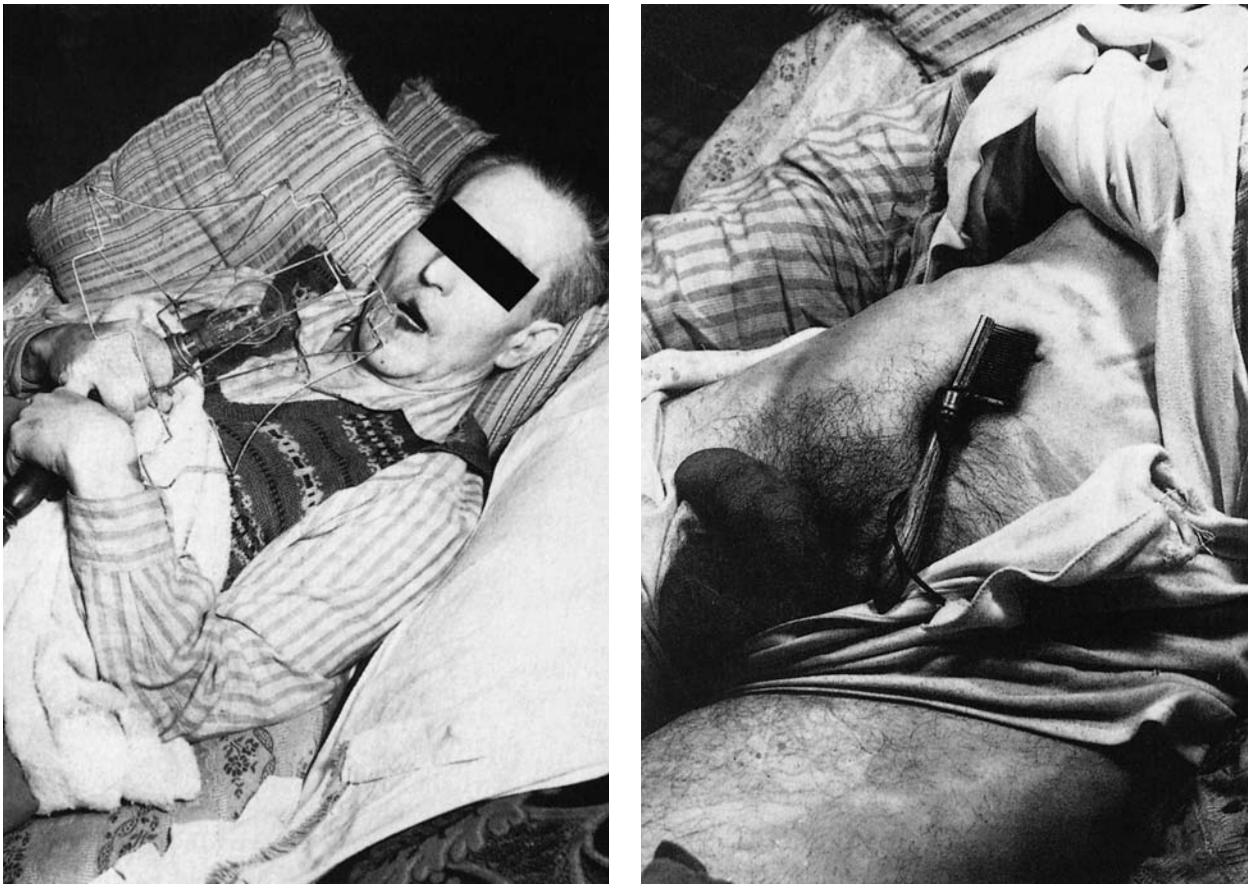

(a) (b)

FIGURE 12.11 *(a) The deceased man had electrical marks on his fingers and lips, which were stained green from metallization from the brass fittings of the table lamp. (b) In addition, he had a metal comb resting over a large electric burn on his abdomen, also wired up to the lamp socket. No circumstantial evidence of suicidal intent could be discovered and some masochistic exercise was suspected.*

through the victim's body from feet to exit through breast and arm, rather than through the higher resistance of the fresh water.

It was also shown that, when the earth safety wire was reconnected, the voltage at the top of the bath fell to negligible levels, even though the fuse did not blow. The woman's husband, after several denials, eventually confessed to dropping the heater into the bath – and also to borrowing a book from the public library the previous day entitled *The do-it-yourself home electrician!*

The other homicide, which could not be proved for lack of evidence or a confession, concerned the wrapping of bare electric wires around a woman's neck. The wire was thick 30 A cooker cable with 40 cm of insulation stripped except for the extreme tips. The woman's husband, a professional electrician, provided the unlikely explanation that he was using this to test his electric shaver!

The third homicide went first undetected as the perpetrator, the husband of the victim, had staged it as an accident. The wife was supposed to have been ironing in the kitchen early one morning and been electrocuted by a faulty cord of the iron, the water-pipe of the sink providing the contact to the earth. In addition to the SOCs, an electrician, but not the pathologist, was called to the scene to investigate the faulty device, but nothing suspicious was detected. Ironically, in spite of the clumsy way in breaking the insulation of the cord and leaving only a couple of millimetres of the naked wire visible, a picture of the faulty cord and iron landed in a textbook for electricians, pointing out the dangers of faulty household appliances. At autopsy the only finding was an electrical burn on the chest. Eight years later, when the husband was in detention in the course of an investigation due to a stolen money transport, he wrote a very detailed confession of the murder of his wife. He had attached two copper electrodes to an extension cord and, while the wife was asleep, pressed them to her chest. In court, in spite of the withdrawal of his confession, he was sentenced to a long imprisonment. Several details, which came out in the course

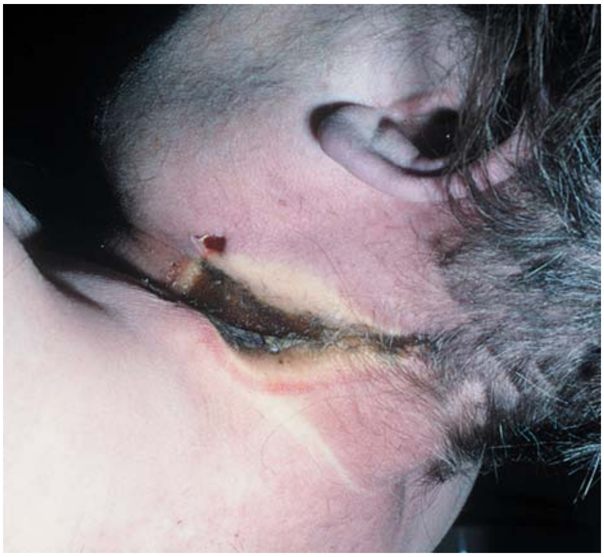

FIGURE 12.12 *Electrical burn around the neck from a mains wire being held against the skin. The typical pale areola and peeled skin blistering can be seen. There was suspicion of homicide, but it could not be proved.*

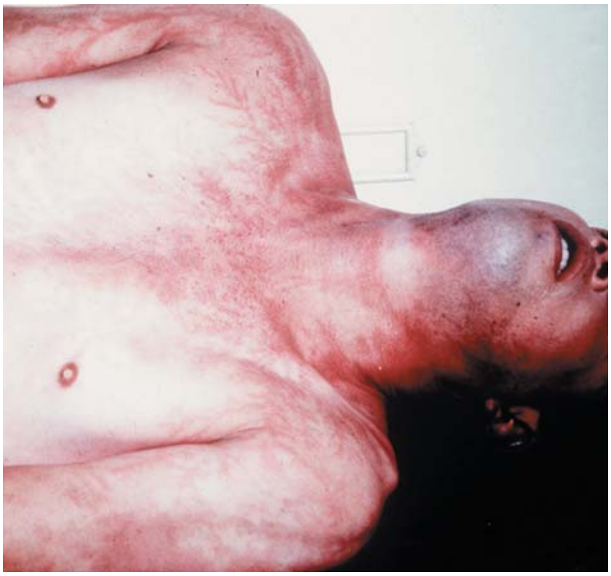

FIGURE 12.13 *The fern-like, arboresque pattern seen over the shoulders and chest in a victim of lightning stroke. Although this is always mentioned in descriptions of lightning deaths, it is seen only in a small minority of cases.*

of the criminal investigation, implied, although it was never proven, that the man had first attempted to kill his wife by feeding her diazepam tablets without her knowledge, and when this had failed, electrocuted her. It is hindsight and fruitless to contemplate what might have happened if the pathologist had been to the scene. The lesson of the case is, however, that if electrocution is suspected, it certainly is worthwhile to take a look at the scene of death.

In tropical and subtropical countries deaths from lightning are not uncommon and, even in higher latitudes, occasional tragedies occur where numbers of people are killed or injured in single episodes. An example was at the Ascot race meeting in 1955. The physics of a lightning strike are complex and not fully understood. Gigantic voltages and amperages are involved when a highly charged thundercloud discharges via a huge arc to the ground.

Lightning deaths cannot be other than accidental and provide no real problems for the forensic pathologist. Occasionally the nature of the death may be uncertain if a dead body is discovered in the open with no marks upon it, but usually there will be reports of a lightning strike nearby and artefacts, such as torn and scorched clothing and magnetized metallic objects in the pockets, may assist in explaining the event.

It is well known that injury from lightning is capricious and unpredictable. Two people can stand side by side during a flash and one may be mutilated and killed while the other is unharmed. The physical damage in fatal cases can vary from virtually nil to gross burning, fractures and tissue destruction. Cutaneous marks may be present, the well-known ‘fern-like’ or ‘arboresque’ pattern, also referred to as ‘Lichtenberg’ figures (because of their similarity to the pattern described by a German physicist, Georg Christoph Lichtenberg, in 1777) being much less common than the standard texts suggest. Irregular red marks, often linear first-degree burns, may follow skin creases, especially if damp from sweating. These marks may be many inches long and generally follow the long axis of the body towards the ground. Frank blistered or charred burns are also present in some cases.

An excellent picture of the Lichtenberg figures was taken by Domart and Garet, published in the *New England Journal of Medicine* in 2000, showing a 54-year-old man struck by lightning. With the exception of numbness and paraesthesia of the left shoulder, flank and leg, he felt well on the arrival at the emergency department. The serum creatine kinase and myoglobin concentrations were slightly elevated. After a 24-hour observation period, the patient was sent home and two days later, on a follow-up visit, the fern-leaf pattern of marks had disappeared.

The clothing may be torn off and this can sometimes raise the suspicion of foul play if the lightning aspect is obscure. The clothing is typically ripped open as if by an internal explosion, and belts and boots may be similarly ruptured. Metal objects in the pockets may be fused or magnetized, as may be metal buttons and tooth fillings. Burns on the skin may be adjacent to metal objects in or on the clothing.

There is often a smell of singeing or burning about the body and its clothing. The hair may be scorched and there is often a head injury, caused either by the lightning strike itself or by falling to the ground.

REFERENCES AND FURTHER READING

Adjutantis G, Skalos S. 1962. The identification of the electrical burn in cases of electrocution by the acro-reaction test. *J Forensic Med* **9**:101–4.

Anders S, Matschke J, Tsokos M. 2001. Internal current mark in a case of suicide by electrocution. *Am J Forensic Med Pathol* **22**:370–3.

Arden GB, Harrison S, Lister J. 1956. Lightning accident at ascot. *Br Med J* **2**:1450–3.

Arnold R, Giebe W, Winnefeld K, *et al.* 2002. [An initially unexplained death during prison sentence. Diagnostic verification by atomic absorption spectrometry.] *Arch Kriminol* **209**:14–19.

Bailey B, Forget S, Gaudreault P. 2001. Prevalence of potential risk factors in victims of electrocution. *Forensic Sci Int* **123**:58–62.

Bissig H. 1960. Über Niederspannungsunfälle. *Elektromedizin* **5**:154–60.

Bonte W, Sprung R, Huckenbeck W. 1986. [Problems in the evaluation of electrocution fatalities in the bathtub.] *Z Rechtsmed* **97**:7–19.

Bornstein FP. 1962. Homicide by electrocution. *J Forensic Sci* **7**:516–17.

Critchley M. 1934. Neurological effect of lightning and electricity. *Lancet* **1**:68–70.

Domart Y, Garet E. 2000. Images in clinical medicine. Lichtenberg figures due to a lightning strike. *N Engl J Med* **343**:1536.

Dutra FR. 1981. Electrical burns of the skin. Medicolegal investigation. *Am J Forensic Med Pathol* **2**:309–12.

Eriksson A, Ornehult L. 1988. Death by lightning. *Am J Forensic Med Pathol* **9**:295–300.

Harrison SH, Arden GP. 1957. Electrical burns. *Br J Clin Pract* **10**:117–20.

Holder JC. 1960. An unusual method of attempted suicide. *Med Leg J* **28**:41–3.

Jacobsen H. 1997. Electrically induced deposition of metal on the human skin. *Forensic Sci Int* **90**:85–92.

Jaffe RH. 1928. Electrical injury. *Arch Pathol* **5**:837–70.

Janssen W. 1984. Injuries caused by electricity. In: *Forensic histopathology*. Springer, Berlin.

Jellinek S. 1932. *Elektrische verletzungen*. Barth, Leipzig.

Klintschar M, Grabuschnigg P, Beham A. 1998. Death from electrocution during autoerotic practice: case report and review of the literature. *Am J Forensic Med Pathol* **19**:190–3.

Lee WR, Zoledziowski S. 1964. Effects of electric shock on respiration in the rabbit. *Br J Ind Med* **21**:135–48.

Marcinkowski T, Pankowski M. 1980. Significance of skin metallization in the diagnosis of electrocution. *Forensic Sci Int* **16**:1–6.

Mellen PF, Weedn VW, Kao G. 1992. Electrocution: a review of 155 cases with emphasis on human factors. *J Forensic Sci* **37**:1016–22.

Moncrief JA, Pruitt BA, Jr. 1971. Hidden damage from electrical injury. *Geriatrics* **26**:84–5.

Morgan EW. 1958. Atrial fibrillation and subdural haematoma associated with lightining strike. *N Engl J Med* **259**:956–9.

Pasetti M, Viterho B. 1965. Ultrastructural modifications in the skeletal muscles of electrocuted mice. *Med Sci Law* **5**:29–36.

Perper J. 1977. *Electrical injuries*. In: Wecht C (ed.), *Legal medicine annual 1976*. Appleton Century Crofts, New York.

Pierucci G, Danesino P. 1980. The macroscopic detection of metallization in the latent current mark. *Z Rechtsmed* **85**:97–105.

Polson CJ. 1963. *Essentials of forensic medicine*. Pergamon Press, Oxford.

Polson CJ, Gee DJ, Knight B. 1985. *Essentials of forensic medicine*, 4th edn. Pergamon Press, Oxford.

Puschel K, Brinkmann B, Lieske K. 1985. Ultrastructural alterations of skeletal muscles after electric shock. *Am J Forensic Med Pathol* **6**:296–300.

Puschel K, Lockemann U, Bartel J. 1995. Postmortem investigation of serum myoglobin levels with special reference to electrical fatalities. *Forensic Sci Int* **72**:171–7.

Robinson DW, Nasters F, Forrest WJ. 1965. Electrical burns an analysis and review of 33 cases. *Surgery* **57**:385–90.

Shaw D, York-Moore ME. 1957. Neuropsychiatric sequelae of lightning strike. *Br Med J* **2**:1152–4.

Somogyi E, Rozsa G, Toro I, *et al.* 1967. Electron microscopic observations on the epidermis of the electrocuted skin of rats. *Med Sci Law* **7**:152–5.

Thomsen HK, Danielsen L, Nielsen O, *et al.* 1981. Early epidermal changes in heat- and electrically injured pig skin. I. A light microscopic study. *Forensic Sci Int* **17**:133–43.

Thomsen HK, Danielsen L, Nielsen O, *et al.* 1981. Early epidermal changes in heat- and electrically injured pig skin. II. An electron microscopic study. *Forensic Sci Int* **17**:145–52.

Torre C, Varetto L. 1986. Dermal surface in electric and thermal injuries. Observations by scanning electron microscopy. *Am J Forensic Med Pathol* **7**:151–8.

Werne AH. 1923. Death by electricity. *NY State Med J* **118**:498–502.

Wright RK, Davis JH. 1980. The investigation of electrical deaths: a report of 220 fatalities. *J Forensic Sci* **25**:514–21.

Zhang P, Cai S. 1995. Study on electrocution death by low-voltage. *Forensic Sci Int* **76**:115–19.

CHAPTER 13

Complications of injury

- Haemorrhage 339
- Infection 340
- Pulmonary embolism 340
- Fat and bone marrow embolism 343
- Adult respiratory distress syndrome 345
- Renal failure following trauma 345
- Disseminated intravascular coagulation 345
- Air embolism 346
- Subendocardial haemorrhage 348
- Sudden death from suprarenal haemorrhage 349
- References and further reading 349

Serious bodily injury, from whatever cause, may either lead to virtually instantaneous death from destruction of vital organs and structures, or may cause delayed death from complications of the original injury. This delay may be short, as in torrential haemorrhage or acute respiratory failure, or it may be progressively longer, stretching into hours, days, weeks or even years. For example, the author (BK) dealt with a death in which a glider crash 15 years previously had to be implicated as the sole reason for the death, as the victim suffered a fractured spine, a paraplegia, paralysed bladder and recurring ascending urinary infection leading to death in renal failure from bilateral pyelonephritis.

Many even more delayed deaths can be found in the records of most experienced forensic pathologists. As long as a direct chain of events can be traced from the injury to the death, then the former must be considered to be the basic cause, a fact which may have profound legal implications both for civil compensation and criminal responsibility. Some of the most difficult problems in forensic pathology – and some of the most arduous testimony and cross-examination in court – concern deaths from which post-traumatic complications are disputed as being causative factors.

HAEMORRHAGE

Bleeding may occur externally through lacerations and incised wounds or via natural passages, such as the bronchi and trachea, oronasal passages, ears, vagina, urethra or rectum. Internal bleeding comprises contusions, which are leakages of blood into tissue spaces from rupture of vessels and free bleeding into body cavities, notably the pericardium, peritoneum, pleura and cranium.

The volume varies greatly, from the tiny petechiae of ruptured venules to the massive amounts seen in a ruptured aorta. As a primary cause of death it is not only the total volume that is lethal, but the rapidity and the site of the bleeding. When the leakage is slow the body can compensate for the loss of a far greater volume, both by adjustment of the vascular bed and by restoration of the blood volume by transfer from other aqueous compartments.

As to site, it is obvious that a small quantity of bleeding into the brainstem is likely to be lethal, whereas the same volume exuded into the pleural cavity would be of little consequence. Bleeding usually begins at the time of injury, though there is often a momentary delay at the instant of infliction. When a sharp blade cuts the skin, it is a matter of common observation that the blood may hesitate for a second or so before welling out. This is caused by transient spasm of local vessels from the stimulus of the injuring object. Bleeding then continues until normal haemostasis plugs the vessels.

In some injuries involving arteries, the musculoelastic vessel may retract and its wall invaginate, so that an almost immediate seal prevents copious haemorrhage. Occasionally even gross injuries, such as the amputation of a limb by a railway wheel, may be almost devoid of significant bleeding, because the crushing effect may combine with arterial wall retraction to seal the cut vessel effectively. It can sometimes be more dangerous partly to transect an artery than to divide it completely, as retraction cannot then take place. Failure of haemostasis is a haematological matter, but may have medico-legal connotations if an injury (which includes a surgical procedure) is performed on someone who has a haemorrhagic diathesis either from natural disease, or from therapeutic anticoagulant treatment, such as heparin or warfarin.

Delayed bleeding may be seen in a number of conditions. In trauma, especially traffic accidents, organs, such as the liver, spleen or (sometimes) lung, may be injured within their capsules or covering membranes, which initially remain intact. A good example is a subcapsular haematoma of the liver, which may be large and grow even larger as continued bleeding strips more capsule from the parenchymal surface. Eventually, the blister may rupture and the now unrestrained bleeding area can pour into the peritoneal cavity. Trauma can also weaken the wall of an artery or vein, leading to a false aneurysm, which can later rupture. An arteriovenous fistula can also form and later burst.

Infections can also develop in trauma sites and involve vessels in the vicinity, so that an abscess or cellulitis can secondarily lead to severe haemorrhage when the vessel wall is eroded. This is rare except among poorly treated accident or battle casualties, as are mycotic aneurysms that result from infected emboli impacting in a branch of an artery and causing local septic necrosis.

It is sometimes difficult to know how much of a haemorrhage found at autopsy may be accounted for by post-mortem bleeding. There is little doubt that the volume may increase after death, but, except in serous cavities, such as the pleura and peritoneum, or externally from the body surface, in most cases this is a small proportion of that which leaked under arterial pressure during life, due to the tissue pressures opposing passive bleeding.

However, a haemothorax from a ruptured aorta may amount to several litres and much of this may be in the form of a large shaped clot. A clot can form post-mortem, so its presence cannot be taken as an index of *in vivo* formation, but the source of the bleeding will usually be obvious and any post-mortem addition will not affect the interpretation. Copious external bleeding can continue after death, especially from the scalp, particularly if the head is dependent after death.

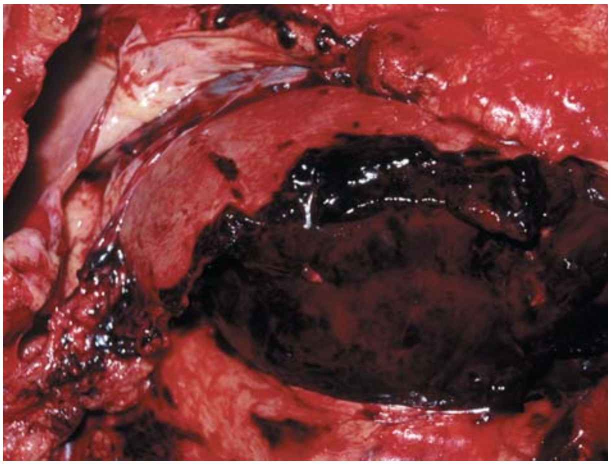

FIGURE 13.1 *Massive intra-abdominal bleeding from a subcapsular haematoma of the liver due to blunt abdominal trauma.*

INFECTION

Infection used to be so common after open wounds that it was the norm rather than the exception. Prior to the introduction of sulphonamides and antibiotics, many criminally inflicted wounds, which in themselves were not a danger to life, became fatally infected so that an assault became a homicide. This illustrates the fallacy of comparing murder rates in the last century with today, as the survival rate because of rapidly available ambulance and resuscitation services, blood transfusion, emergency operations and antibiotics is vastly greater now than in former times, when even a trivial injury was often fatal.

The types of post-traumatic infection are legion and vary greatly from country to country. Purulent wound infection from Gram-positive cocci, Gram-negative bacilli, anaerobes, such as *Clostridium perfringens*, and other more common organisms, is the most frequent, but in some countries tetanus and anthrax are common dangers. The whole matter is one of clinical microbiology, and the main forensic relevance is to prove a chain of causation between the original injury and a death from intercurrent infection. There may be medico-legal issues involved, such as failure to give or delay in giving antibiotic cover, which can have both civil and criminal legal consequences. A criminal assault that ends in death because of a neglected infection does not exonerate the perpetrator from all responsibility, even though there has been a *novus actus interveniens* in the form of defective medical treatment.

PULMONARY EMBOLISM

This is a most important topic in forensic pathology, as the medico-legal implications of a fatal pulmonary embolus are common and profound. Pulmonary embolism is the most underdiagnosed cause of death where no autopsy is performed; it has been estimated that, in the USA – and probably elsewhere – less than half of fatal pulmonary emboli are recognized clinically.

As with infection, an originally non-lethal injury may end in death because of venous thrombosis and pulmonary embolism, making what might be a simple accident or a common assault into a grave legal issue. The victims of many forms of trauma are at risk from pulmonary embolism because:

- Tissue trauma increases the coagulability of the blood for several weeks, the peak being between one and two weeks.
- Injury to the tissues, especially the legs or pelvic region, may cause local venous thrombosis in the contused muscles or around fractured bones.
- The injury may confine the victim to bed, either because of general shock and debility (especially in old people) or because the trauma itself necessitates recumbency, as in head injuries, severe generalized trauma or injury affecting the legs. In either case, recumbency leads to pressure on the calves and immobility causes reduced venous return and stasis because of lessened muscular massage of the leg veins. The common result is thrombosis of the deep veins of the legs, which can extend proximally into the popliteal and femoral vessels, forming a dangerous source of venous thromboemboli. Small emboli may break off and impact in more peripheral branches of the pulmonary arteries, sometimes causing pulmonary infarcts that may be precursors of a massive embolus that impacts in the major lung vessels and causes rapid death.

At autopsy, such large emboli are readily visible and can usually be easily distinguished from post-mortem clot. The latter is dark red, soft and jelly-like, with a shiny, glistening surface. It is often separated into 'chicken-fat' plasma clot and dark red-cell clot by sedimentation after death. When pulled out of the vessels it forms a cast of the branches, albeit shrunken by clot retraction. It is less evident in peripheral branches and, when the lung is sliced post-mortem, clot does not pour out of cut small vessels.

Conversely, ante-mortem embolus (especially if a number of days old) is firm, though brittle and has a dull, matt, striated surface from fibrin lamination. Older thrombus tends to be greyish-red and varies in colour from place to place. Although it may appear to be a cast of the large vessel in which it is impacted, it may often be unravelled to form a long length that obviously originated in a leg vein. Side branches, or the stumps thereof, may be seen that do not correspond to the branches of the pulmonary artery in which they lie.

Post-mortem clot may be adherent to the ante-mortem embolus and sometimes forms a sheath around it, so that the true nature is obscured unless a careful examination is made. On cutting the lung with a knife, ante-mortem emboli may be seen in the more peripheral vessels, often standing up slightly 'proud' above the surface, like toothpaste coming from a tube.

The importance of the differentiation between ante-mortem emboli and post-mortem clot is emphasized, as the legal issues hanging upon the unequivocal diagnosis may be very important. Histological confirmation of an ante-mortem origin must be made if there is any doubt. Pulmonary infarction does not occur from fatal massive pulmonary embolism, as death is too rapid. There may be infarcts present in the lungs but these must be caused by previous smaller emboli, at least a day earlier and probably much longer.

Medico-legal aspects of pulmonary embolism

Pulmonary embolism is the most underdiagnosed condition in British death certification, frequently being unsuspected as the cause of death by clinicians. Several investigations into the medico-legal aspects have been made (Knight 1966; Knight and Zaini 1980; Zaini 1981) and the peak incidence at about 2 weeks after trauma confirmed. In Knight's survey, more than three-quarters of the victims had predisposing factors such as injury, surgical operation or immobility in bed, but the remaining 20 per cent were ambulant and apparently healthy. This has important medico-legal implications because, if fatal pulmonary embolism can strike an appreciable proportion of the population who have not suffered one of the recognized predisposing factors, then the cause–effect relationship after trauma is weakened. If the standard of proof in criminal trials must be 'beyond reasonable doubt', then the fact that up to 20 per cent of pulmonary embolism deaths have not followed trauma or immobility must surely remove the cause–effect relationship from near-certainty to mere probability, which is sufficient for a civil decision, but not for a criminal conviction. This is a legal matter for the judge, however, who may or may not let the matter go to the jury – and if it does, for the jury to decide. There is marked variation in these decisions from case to case.

After the lung appearances have been examined and pulmonary embolism confirmed, the source of the embolus must be sought. In almost all cases this will be found in the vessels draining into the femoral veins, though rarely pelvic vessels are involved (usually in relation to pregnancy or abortion). Here pelvic veins may be thrombosed, with extensions into the iliac system and exceptionally into the inferior vena cava. An even more rare source is from jugular thrombosis, sometimes seen as extensions of intracranial venous sinus thrombosis. Axillary and subclavian vein thrombosis is equally unusual, the legs accounting for the vast majority of emboli.

In the usual leg vein thrombosis, various autopsy techniques are available to seek the site of obstruction. Extensive dissection is favoured by some, in which the femoral vein is exposed through a skin incision, this being continued distally as far as necessary to find the residual thrombosis. Of course, after a massive pulmonary embolus, by definition a large part of the source thrombus has been detached and therefore the pathologist has to search further distally to find the remnants. This makes dissection an extensive and disfiguring process, so alternative techniques can be used.

Zaini found that many thrombi begin far distally, even in the dorsum of the foot. He used a flexible wire with a blunt terminal knob (a Bowden cable from a bicycle brake) inserted down the femoral vein until it was arrested, then cut down from the skin at that point to examine the peripheral veins. Alternatively, transverse or longitudinal incisions can be made into the calf of the leg to examine the deep veins. The soleus and gastrocnemius muscles are transected to view their contained veins, but often the thrombus is found in the interosseous veins between tibia and fibula. If ageing of the thrombus is attempted histologically, then it should not be expressed from the vein, but the vein should be taken out with adjacent muscle, as it is the junction between thrombus and vein wall that offers the most information about the maturity of the thrombus.

Dating of pulmonary emboli and deep vein thrombi

As discussed above, it can be a matter of considerable medico-legal importance to know if a pulmonary embolus arose prior to, or subsequent to, some traumatic event. A major difficulty is that the embolus may be the most recent addition to an extending venous thrombosis that is considerably older.

It is also difficult to use histological criteria to date the free embolus from the lungs, as it is the thromboendothelial junction that provides the most information. The best method, therefore, is to examine the residual thrombus, almost always in the leg veins, to see if the oldest part could have formed as far back in time as the suspected traumatic event. For this, the vein wall with the contained thrombus is required. A segment of thrombosed vein, if necessary with adjacent muscle, should be dissected out of the leg. The presence of thrombus will have been confirmed prior to removal by transverse cuts across the calf or thigh muscles containing the veins. Sometimes the original site of thrombosis may be as far distal as the dorsum of the foot.

The thrombus-containing vein is processed in the usual way and subjected to various histological stains. Though accuracy about dating is impossible, the following is a useful scheme to provide at least an approximate idea of the duration of thrombosis in the section of vein under study – remembering always that other segments or other veins may have thrombi of different dates.

- Platelet and red-cell appearances give no useful information. Red cells begin to haemolyse between 24 and 48 hours to form amorphous masses, but many intact red cells survive for weeks.
- Using phosphotungstic acid-haematoxylin stain (PTAH), fibrin can be seen as purplish strands on the first day, but they aggregate into small masses with a meshwork of thicker strands and sheets by 4 days. After 2 weeks, the fibrin becomes more deeply purple stained, but begins to be absorbed by about the twenty-fifth day. Using Martius Scarlet Blue stain (MSB), the early pink fibrin strands become fringed by scarlet in about a week.
- Endothelial proliferation is most useful in the first week, as buds begin to arise from the vessel wall about the second day and proliferate during the first week. Clefts lined by endothelium are seen around the periphery of the thrombus against the vein wall. Artefactual contraction caused by fixation shrinkage must be excluded by seeking endothelial nuclei. These endothelial buds anchor the thrombus to the vein wall, and represent the first stage of healing and recanalization. They are usually noticeable by about the fourth day when they begin growing into the fibrin mass and breaking it up into compartments.
- Collagen fibres do not appear for about 5–10 days, often much later. Fibroblasts may be seen as early as 2 or 3 days, but tend to appear towards the end of the first week and reach a maximum at 2 or 3 weeks, reaching a maximum at about 4 weeks. Elastic fibres appear late, not before 28 days and often much later. They reach their maximum density in about 2 months, mainly in the walls of recanalizing vessels.
- Haemosiderin, blue granules demonstrated by Perl's reaction, may be seen by the end of the first week and reach a maximum in 3 weeks. It does not seem to be present as early in thrombi as in extravascular haemorrhages, such as bruising or meningeal bleeding, where it can be seen in 2 or 3 days.
- Capillaries begin to appear on the second day as endothelial buds, but do not contain red cells until about 2 weeks. Over the next 3 months, substantial canalization occurs by widening and merging of these channels. The full lumen may be restored within 6–12 months, but endothelial thickening and haemosiderin deposits in the vein wall may form permanent evidence of a former thrombus.
- Leucocytes, both polymorph and mononuclear, are inconsistent markers; sometimes they do not appear at all. Polymorphs may be seen within the first day – sometimes in profusion – but they vanish rapidly, often by the next day when mononuclears take their place.

- Covering of the thrombus surface by endothelium is rapid and may begin on the first day and be complete within a few days. Various authors have given 24–72 hours as the time required. Naturally, the size of the thrombus surface is a variable factor. Also, only sections taken where the thrombus does not fully fill the vein lumen will provide an opportunity for observing the endothelial covering.

FAT AND BONE MARROW EMBOLISM

Another important sequel to trauma – fat embolism – is the subject of considerable controversy about the source of the lipoid; this does not, however, detract from the autopsy significance of its detection.

A great deal has been written about the pathophysiology of fat embolism and it is now evident that it is not just a simple extrusion of marrow or adipose tissue material by mechanical trauma. The work of Sevitt (1962), Bergentz (1961), Edland (1971), Ellis and Watson (1968), Buchanan and Mason (1982), and many others must be studied for detailed accounts of the suggested pathogenesis.

Wherever the fat comes from, embolism is most often seen after injury to bone or fatty tissue. In simplistic terms, where skeletal structures containing fatty marrow are damaged or where subcutaneous fat is compressed or lacerated, fat globules commonly appear in the pulmonary capillaries – and where they are numerous they somehow leak through the lungs into the systemic circulation where they can cause severe disability or death from impaction in vital organs, such as brain, kidney or myocardium. Not only fat, but cellular haematopoietic tissue from bone marrow, can be liberated into the venous system and reach the lungs.

The clinical manifestations of fat embolism depend on the volume of fat reaching the lungs. If substantial, ventilatory problems can ensue because of vascular obstruction and pulmonary oedema. The most dramatic symptoms occur, however, with systemic fat embolism when globules reach the brain and cause a range of neurological abnormalities, usually coma and death from brainstem involvement.

There is usually a delay between trauma and cerebral fat embolism while fat builds up in the lungs, so that a ‘lucid interval’ occurs, which may be confused clinically with the development of an extradural or subdural haemorrhage. Sevitt stated that of patients with multiple bone fractures, no fewer than 45 per cent had pulmonary fat emboli and 14 per cent had cerebral embolism.

Fat embolism is also associated with burns, barotrauma, soft tissue injury, osteomyelitis, diabetes, surgical operations on fatty tissues (especially mastectomy), septicaemia, steroid therapy, acute pancreatitis and the fatty liver of alcoholism, though fractures remain the most potent cause of the condition. In burns, there is some doubt whether pulmonary fat embolism is a true ante-mortem process, or whether fat can be melted out of tissues and organs in the perimortal period. Mason’s (1968) experience in rapid air-crash fires was that fat embolism was not present, but Sevitt’s hospital series revealed embolism in 47 per cent of deaths from burns.

Pulmonary fat embolism

After trauma, fat commonly appears in the lungs and can be demonstrated there histologically in the majority of cases of fractures and injury to fatty parts of the body such as the buttocks. Indeed, Lehman and Moore (1927) showed that half of a series of non-trauma deaths had histological evidence of fat in the lungs. Mason found fat in the lungs of 20 per cent of his series of non-trauma deaths, but emphasized that quantitatively the amount was small in contrast to that found in cases of fatal trauma. He used a simple scale for assessing the histological severity of embolism as seen in Oil Red-O frozen sections of lung:

- Grade 0: no emboli seen
- Grade 1: emboli found after some searching
- Grade 2: emboli easily seen
- Grade 3: emboli present in large numbers
- Grade 4: emboli present in potentially fatal numbers.

In systemic fat embolism, no such grading is possible; they are either absent or scanty, or they are abundant.

In most instances, pulmonary fat embolism is merely a phenomenon and not a clinical syndrome and the difference between the two seems closely related to the amount of fat impacted in the lung vessels. When clinically manifest it appears as acute respiratory insufficiency, the incidence being up to 2 per cent in long bone fractures and as much as 10 per cent in multiple fractures, especially with pelvic injuries. In small to moderate amounts, fat in the lungs gives rise to no disability, but large amounts, easily seen by a cursory glance at any field of a fat-stained histological section, give rise to acute respiratory distress. Marked pulmonary oedema is the pathological marker for this syndrome, but caution must be employed, as cerebral fat embolism can also cause pulmonary oedema by its effect upon the brain.

Systemic fat embolism

Here the fat penetrates the lung capillaries and appears in the major circulation, so that it can be carried to any organ or structure, including the skin, where it can cause petechial-like

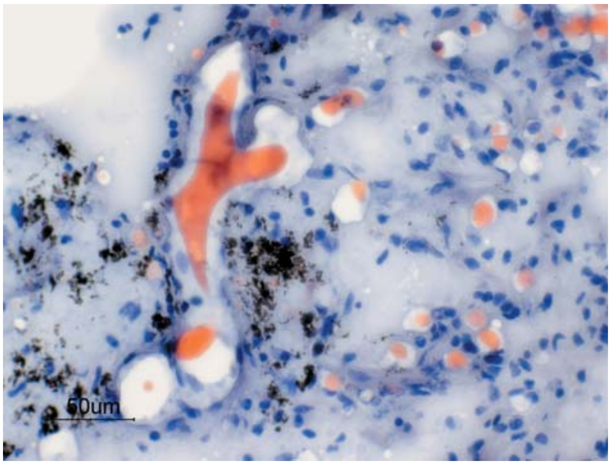

FIGURE 13.2 *Pulmonary fat embolism (Oil Red-O frozen section of lung; original magnification ×20).*

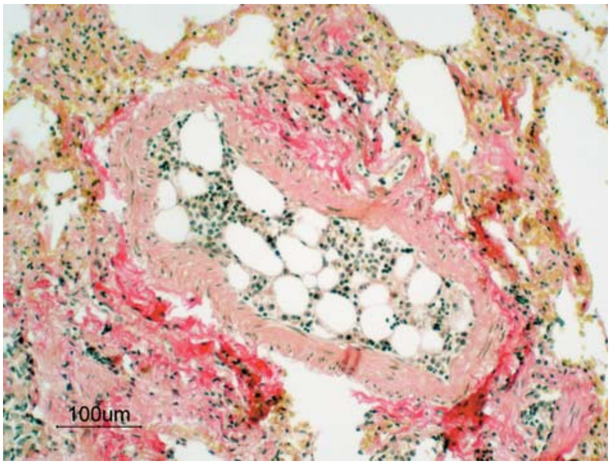

FIGURE 13.3 *Pulmonary bone marrow embolism (van Gieson staining, original magnification ×10).*

lesions. Systemic fat embolism is the fatal manifestation, except where pulmonary loading is so great that it causes respiratory distress. The brain, kidneys and myocardium are the most vulnerable target organs, the brainstem being the one that is most likely to lead to death from impaction of fat globules in its capillaries.

Autopsy findings in fat embolism

When fat is released through the lungs into the systemic circulation, a number of target organs and tissues exhibit abnormalities demonstrable at autopsy, either by gross or histological examination.

Petechial haemorrhages are the typical lesion, caused by the impaction of fat droplets in small venules. These may be seen in the skin of any part of the body, but especially over the front of the chest and on the face and eyelids. Internally they can be widespread but are classically seen in the white matter of the brain, both cerebral and cerebellar hemispheres, as well as the brainstem. Microscopically, when frozen sections are stained by Oil Red-O or other fat stain, the cerebral petechiae will show a central fat globule or globules, and others will be visible without haemorrhage.

In the myocardium, fat may be seen in the interfibre capillaries and in the kidney glomeruli may be stuffed with stained fat. There can be fat in the retina and in the optic nerve, which, in survivors, can cause visual impairment.

Fat embolism is markedly underdiagnosed, both clinically and at autopsy. The more often and the more thoroughly it is sought, the more often it is found. Mild to moderate degrees of pulmonary fat embolism, without systemic overspill, however, must not be given as a cause of death unless there are corroborative clinical features.

Bone marrow embolism

Bone marrow embolism is almost wholly a pulmonary phenomenon, as only in quite exceptional circumstances can cellular material penetrate the lung capillaries. It is quite common in association with fractures of long bones or in multiple bone injuries, though again the frequency with which it is seen is directly related to both awareness and the thoroughness of the search. Where marrow is found, the degree of fat embolism is usually substantial. The common concurrence of fat and cellular embolism is strong evidence of a marrow source for embolic fat, as pointed out by Mason, who found marrow cells in 81 per cent of his cases that had a grade 2 or more lung loading with fat, on a scale of 0–4. According to both Sevitt and Mason, the finding of both substantial fat and marrow emboli in the lungs is definite evidence of ante-mortem infliction of any discovered trauma. For these elements to reach the lungs there must be a competent circulation, albeit for a short time after the trauma.

With the possible exception of burns, any injury inflicted after cessation of effective cardiac function cannot transmit fat or marrow to the pulmonary capillary bed. This is therefore a useful forensic marker on some occasions, such as differentiating between ante-mortem and post-mortem fractures in a body recovered from water. Systemic embolism is not seen in rapid deaths following trauma: there is an interval averaging 24 hours before lung percolation takes place.

Foreign body embolism

A number of other objects may be found as emboli during forensic autopsies and histology. The lung granulomata of

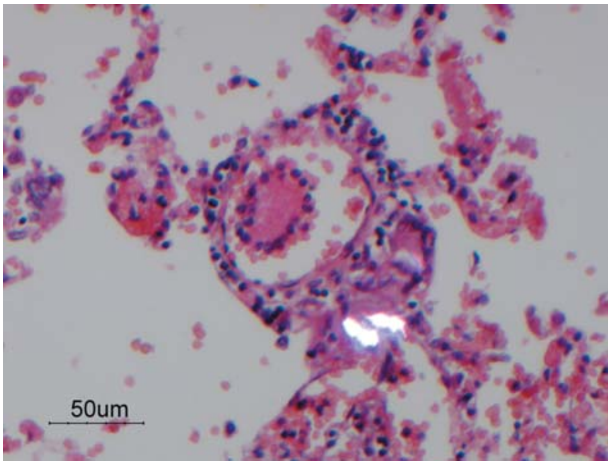

FIGURE 13.4 *Small foreign body granuloma with giant cells and birefringent particles as seen in polarized light in the lung of a drug addict (HE, original magnification ×20).*

talc, etc., seen in the lungs of intravenous drug users, are emboli from contaminants introduced into the venous system by syringe and needle. Shotgun pellets and bullets may enter the circulation via tissue damage and embolize to various sites including coronary arteries. In fat and bone marrow embolism, even fragments of bone and cartilage have been found as emboli in the lungs.

Amniotic fluid embolism is discussed elsewhere.

ADULT RESPIRATORY DISTRESS SYNDROME

Following severe lung injury, such as gross impact upon the thorax or blast injury from explosion, or from aspiration of gastric contents, infections, toxins, systemic shock, irritant gases, near-drowning and many other causes, the lung epithelium may suffer 'diffuse alveolar damage'. Clinically the victim may suffer marked dyspnoea and progressive respiratory failure, becoming hypoxaemic. Pathologically the lungs show a stiff oedema that progresses to a rigid, infiltrated lung if survival is long enough. At autopsy the lungs are hard and retain their shape after removal, having a 'dry-oedema' appearance, being almost double in weight in some instances.

Histologically the initial changes are shedding of type I pneumocytes with intra-alveolar exudates, the formation of hyaline membrane and patchy alveolar haemorrhage. This destructive phase is then replaced after a few days with a proliferative stage, when type II pneumocytes begin to fill the alveoli and a mononuclear response infiltrates the interstitial areas. The alveolar proliferation organizes, if survival continues, and eventual pulmonary fibrosis may seriously incapacitate the half who survive. A typical progression is seen in paraquat poisoning, though here there are virtually no long-term survivors.

A condition that is similar to the appearances of adult respiratory distress syndrome is the so-called 'respirator lung', which develops when a patient dies after an appreciable period of mechanical ventilation. It may appear within a few days, but usually supervenes after many days or weeks on a ventilator. The gross appearances are of a stiff, rigid lung, which, though heavy, does not usually appear obviously oedematous. Microscopically, proteinaceous fluid and a mixture of proliferative cells may be seen in the alveoli.

RENAL FAILURE FOLLOWING TRAUMA

Known for many years, this is a common sequel to extensive muscle damage or to burns that affect considerable areas of skin, as well as from certain poisons. It was formerly ascribed most often to 'acute tubular necrosis', but recent experimental work has thrown some doubt upon this as an accurate clinicopathological entity. When muscle is crushed, severely torn or otherwise rendered ischaemic, where extensive burns are suffered, or where poisons, such as mercuric salts or carbon tetrachloride are given, destruction of the renal tubular epithelium may be seen.

Unfortunately for histological diagnosis, a similar change occurs during post-mortem autolysis. In burns and muscle damage, the tubules may also be blocked with brown casts of myoglobin and these two changes of distal tubular necrosis and casts were thought to cause the profound oliguria or anuria often associated with trauma and burns. In biopsy as opposed to autopsy material, however, the tubular damage has been found to be minimal and the casts may be a sequel to reduced filtration rather than to tubular damage, so the validity of 'acute tubular necrosis' as an explanation of renal failure following trauma has been questioned. The true mechanism is still obscure, but may well involve the juxtaglomerular apparatus and the renin–angiotensin system, as well as disseminated intravascular coagulation effects on the glomeruli.

DISSEMINATED INTRAVASCULAR COAGULATION

The importance and frequency of this condition has been increasingly recognized in recent years, as it may follow a whole range of traumatic, infective and other acute events. Widely known as 'DIC', it has a number of other names indicating that it is a consumption coagulopathy associated with the blood clotting mechanism.

In DIC there is an abnormal activation of the coagulation process within the blood vessels caused by a wide variety of factors related to the blood itself, the vessel walls and to blood flow. Damaged tissue from trauma and burns can trigger thromboplastin-initiated coagulation: entry of tissue cell elements, especially from erythrocytes, brain and placenta are particularly potent. Particulate matter such as micro-organisms or microemboli of all types (including fat and air emboli, especially those from decompression) can precipitate coagulation via factor XII. Vascular endothelial damage and stasis of blood flow can have a similar effect.

Whatever the causes – and they are extremely complex – fibrinogen is consumed and fibrin precipitated in vessels, leading to both vascular obstructive effects and to a haemorrhagic diathesis from depletion of the coagulation system. Platelets are also consumed, adhering to the fibrin thrombi. Fibrinolysis is activated and there is a dynamic contest between intravascular coagulation and its removal.

After death there can be post-mortem fibrinolysis (Mole 1948), which lessens the ease with which fibrin deposits can be detected: staining techniques are also an imperfect method of visualizing the fibrin, so that a careful search must be made in autopsy histology for fibrin remnants. Lungs, kidney, liver and adrenals are the most likely organs to yield positive results.

Microvascular obstruction, leading sometimes to frank infarction as well as reduced function, together with bleeding, form the major dangers of DIC in a forensic context.

AIR EMBOLISM

Most examples of air embolism have a medico-legal significance as the entry of air into the circulation usually comes about as a result of trauma, sometimes surgical or therapeutic, from barotrauma or from criminal intervention.

Air embolism consists of an interruption of the circulatory system by bubbles of air (or other gas) that gain access to the circulation, usually through the venous side. It is an ‘airlock’, familiar to plumbers and owners of diesel engines, where the normal flow of liquid through tubes is wholly or partially blocked by air.

In the animal body the major disability is in the heart, as air is compressible and the contractions of the heart fail to move on the bolus of gas, being designed to pump an incompressible liquid. The air must enter on the venous side to be sucked towards the heart, causing pulmonary air embolism. The air usually remains in the right side of the heart, pulmonary trunk and arteries, and in the pulmonary vessels, rarely emerging on the pulmonary vein side. Death is usually immediate, but can be delayed by up to about 2 hours.

In the surgical context, aspiration of air into neck veins may occur while operating on the head or neck (especially thyroid and neurosurgical operations) with the patient in a sitting position. When the atria are below any breach in a vein, there is a suction effect that can draw in air; this can also happen in suicidal or homicidal cut throat. Refilling of a therapeutic pneumothorax has also led to fatal air embolism, either because the lung is punctured or an adhesion has torn the visceral pleura. In either case air under pressure may enter an exposed vein.

Some accidents have arisen during transfusion or infusion, and may form the basis for a medical negligence action. Now that flexible, collapsible fluid or blood containers are widely used the danger is less, but where rigid bottles are still used, the need for a vent tube may allow air to enter the connecting tubing when a bottle is allowed to empty completely. This in itself is innocuous, but if a fresh bottle is then connected, the new flow will drive all the air in the tubing into the vein.

Even more dangerous is the practice of connecting pressurized air or oxygen to the vent in order to hasten the transfusion speed: if the bottle is allowed to empty, gas under pressure enters the vein.

Arterial air embolism is rare, but can occur in lacerations of the lungs, refilling of an artificial pneumothorax or in some forms of barotrauma where the lung is ruptured in a diver who ascends too rapidly from a great depth. The ‘bends’ are also a type of gas embolism, in which dissolved nitrogen appears in gaseous form in the blood. This is caused by a sudden release of pressure on the return of a diver to atmospheric conditions after being supplied with air at high pressure. Death can be the result of air locks in the cerebral vessels or of ventricular fibrillation from coronary air embolism (Moore and Braselton 1940).

Criminal abortion accounted for many deaths from air embolism in former years (and still does in some parts of the world), from the insufflation of the uterus by a Higginson syringe (Chapter 19). Though therapeutic abortion by suction aspiration of the uterine contents is an extremely safe procedure, several somewhat inexplicable deaths have occurred from air embolism. The mechanism is obscure but may have been caused by muscular rebound of the uterus sucking air into the cervix. Some homicides, including ‘mercy killings’ committed by doctors, have been caused by deliberate injection of air into peripheral veins. The volume of air needed to cause fatal embolism has been hotly debated for years, but no real consensus has emerged, estimates varying from 10 to 480 ml (Polson 1963). If the volume of the right side of the heart is accepted as the minimum space that has to be filled, about 100 ml would appear to be a reasonable volume. In the rare event of air gaining access to the arterial system, presumably much less would be required to be effective; it would have to be introduced into a carotid artery to be of much danger, as this is the only way it could gain access to the cerebral circulation. To reach the coronary vessels – always mentioned in books, but virtually never seen – the blood would have to enter the pulmonary veins to reach the root of the aorta via the left side of the heart. Virtually all of the few cases of arterial air embolism have arisen in relation to dysbarism, where bubbles have either been generated within the vessels by decompression, or have entered the arterial system in the lungs from tears in the lung tissue.

Certainly air injected into a limb artery would do no harm at all, especially as most of the gas is likely to be absorbed very rapidly by the tissues. Thus the ingenious plot in Dorothy L Sayer's novel, *Unnatural Death*, where 'a hypodermic of air was injected into an artery' would have failed on two counts: site and volume!

The autopsy in suspected air embolism

When the circumstances indicate the possibility of fatal air embolism, special precautions must be taken over the autopsy. A pre-autopsy chest radiograph must be taken, as this is by far the best way of demonstrating air in sufficient quantities to be fatal. A chest radiograph is also essential in any kind of barotrauma as it may also indicate a pneumothorax, a lesion so often the cause of air embolism.

Almost the only mechanism of death is 'pump failure' of the right side of the heart. Air fills the great veins, right atrium and right ventricle, causing a froth that cannot be pumped on by the heart in systole because air is compressible. It is unlikely that air will penetrate the pulmonary capillaries in any quantity, unless some vascular shunts are present, so froth is unlikely to be seen in the left side of the heart, except in dysbarism.

Following radiography – in place of it, as second-best – the body must be carefully inspected for any tissue swelling, especially in the upper thorax and neck region, where crepitant gas bubbles may form surgical emphysema from leakage from lung damage or pneumothorax, or both, in dysbarism.

The head should be dissected first, not to detect air bubbles in the cortical veins, as almost every textbook erroneously states, but to look for air in the cerebral arteries. Descriptions and photographs of air segments in the cerebral veins are part of the mythology of forensic pathology, handed on uncritically from one book and one author to another. There is no way in which air bubbles can reach the pial veins on the brain surface. They cannot get there via the venous system, as there is no way in which they could struggle against the flow in the jugular veins, pass through the complex arrangement of intracranial venous sinuses retrograde to the circulation and dispose themselves over the cortical surface. Neither can they reach the same veins by penetrating the capillary bed of the brain, in the rare instances of arterial air embolism. As was demonstrated almost 30 years ago, these bubbles are artefacts and can be seen in many routine autopsies where there can be no question of air embolism.

The basal arteries should be inspected for air, however, which can reach them during episodes of dysbarism where air gains access to the arterial system in the lungs during decompression. The calvarium is removed gently and the dura opened. The brain is carefully lifted and small artery forceps clamped on the intracranial part of both internal carotid arteries and vertebral arteries. The vessels are then cut below these before the brain is removed from the skull in the usual manner. The whole pluck can then be submerged in water and the ends of the vessels watched for escaping bubbles when the clamps are removed.

The thorax is opened in the fashion appropriate for the detection of a pneumothorax, with puncture of an intercostal muscle beneath a pool of water supported by the reflected skin flap. This is done on each side, remembering that a positive result can only be expected in a **tension** pneumothorax.

The abdomen should next be examined, especially where air may have gained access through the uterus. The intestines should be gently displaced and the mesenteric veins examined for air bubbles. The inferior vena cava, the common iliac veins and the various pelvic veins should be scrutinized, and, if necessary, opened under water by pouring water into the opened abdominal cavity.

The sternum and costal cartilages are then removed carefully, preferably keeping the manubrium and upper ribs intact at this stage to avoid damage to the great vessels in the thoracic inlet. The superior vena cava should be inspected for bubbles. The pericardium is opened and the right and left ventricles punctured under water.

Bajanowski *et al.* (1998, 1999) consider that an aspirometer has to be used for the detection, measurement and storage of gas originating from the heart ventricles. It should be filled completely with distilled water containing two drops of Tween 80 to reduce the surface tension of the water to prevent adherence of small air bubbles to the wall of the aspirometer. Subsequently the gas should be analysed by gas chromatography and the results assessed according to the criteria defined by Pierucci and Gherson (1968, 1969) before the diagnosis 'air embolism' is justified.

If air embolism substantial enough to have caused death is present, frothy blood will be quite evident oozing from

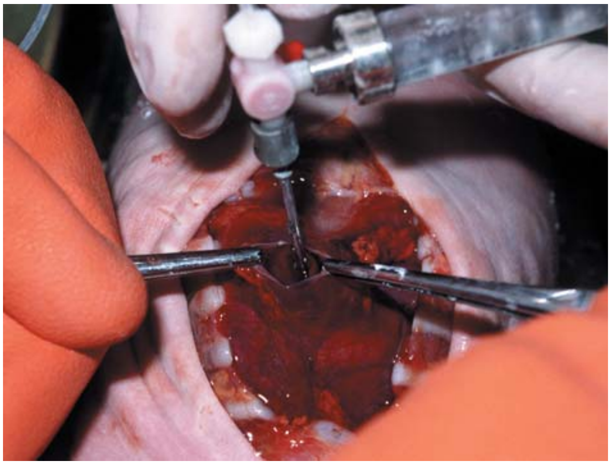

FIGURE 13.5 *An aspirometer should be used to recover air from the heart ventricles for subsequent analysis by gas chromatography.*

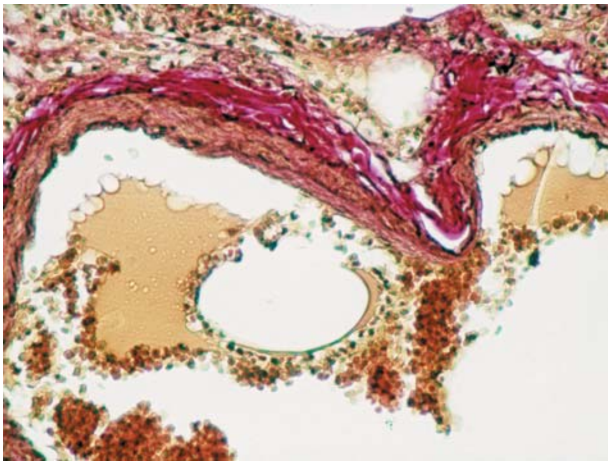

FIGURE 13.6 *Leucocyte and platelet accumulation around an empty space, 'air bubble' in blood, considered by some authors as evidence of venous air embolism. (Reproduced by kind permission of Professor E Lignitz.)*

the ventricular lumen, almost invariably the right. If this is apparent, however, it will almost certainly already have been seen on the pre-autopsy radiograph.

If there has been a considerable delay between death and autopsy, the air can dissipate, presumably dissolving into the tissues. Any degree of decomposition of the body negates a diagnosis of air embolism, because of the production of gases of putrefaction.

SUBENDOCARDIAL HAEMORRHAGE

A striking feature of many forensic autopsies, especially on victims of severe trauma, is the presence of well-marked haemorrhage under the endocardium of the left ventricle. This is often so striking as to provide an indication of some catastrophic event shortly before death.

Though the haemorrhages have been described since the last century, they were specifically studied in the 1930s by Sheehan in cases of abortion and acute haemorrhage associated with pregnancy; they were formerly known as 'Sheehan's haemorrhages'.

The lesions are seen in the left ventricle, on the interventricular septum, and on the opposing papillary muscles and adjacent columnae carneae of the free wall of the ventricle. The haemorrhages are flame-shaped and confluent, not petechial, and tend to occur in one continuous sheet rather than patches. The bleeding is in a thin subendocardial layer, but when severe may actually raise the endocardium into a flat blister that can be palpable on the smooth septum. The mechanism of production is obscure, but they are commonly seen in the following circumstances:

- After sudden, profound hypotension, either from severe blood loss or from 'shock' in the widest sense.
- After intracranial damage, from head injuries, cerebral oedema, surgical craniotomy or large intracranial tumours. Sudden intracranial decompression also seems to be associated with subendocardial bleeding.
- Obstetric catastrophes seem particularly prone to produce these haemorrhages, as found by Sheehan. Death from ante-partum or post-partum haemorrhage, ruptured ectopic gestation, abortions of various types and ruptured uterus often reveal these lesions.
- Various forms of poisoning, especially acute heavy-metal toxicity, particularly arsenic. The most profound subendocardial haemorrhages ever seen by the author (BK) were in a massive suicidal poisoning with arsenious oxide: the endocardium was raised into a large blood blister on the interventricular septum.

The haemorrhages can appear extremely rapidly, within a few heart beats. The author (BK) has seen well-marked lesions in a heart that was avulsed from its base during the crash of a military aircraft, obviously causing virtually instantaneous death. The common factor seems to be sudden hypotension and it has been suggested that, if the intraventricular pressure drops precipitously, the existing blood pressure in the coronary system is then unsupported across the endocardium by an equal pressure within the ventricular lumen, so that rupture of the superficial vessels occurs. There are flaws in this theory, one being the common incidence in intracranial lesions. Experimental work in goats has shown that prior vagotomy prevents the occurrence of subendocardial haemorrhages, suggesting that the phenomenon is mediated by the autonomic nervous system. To support this hypothesis, they are known to be part of Virchow's triad of ‘pulmonary oedema, gastric erosions and subendocardial haemorrhage’ seen in head injuries and cases of raised intracranial pressure.

SUDDEN DEATH FROM SUPRARENAL HAEMORRHAGE

Though not common, haemorrhage into the adrenal glands is a well-recognized terminal event, usually after some trauma to the body not related to direct injury to the suprarenal vessels themselves.

Originally these haemorrhages were thought to be confined to meningococcal septicaemia in children, the lesion being called the ‘Waterhouse–Friderichsen syndrome’. It is now recognized that adrenal haemorrhage is part of a general response to stress, of which meningococcal infection is only one of a wide range.

In a recent 5-year period, the author (BK) found 16 cases of ‘adrenal apoplexy’ (to give it its old name) in the autopsy records of a forensic pathology service that performed 4890 autopsies in that period, giving an incidence of about 0.3 per cent.

Of these 16 cases, more than half (9) were bilateral. There were more male than female subjects, probably because of the greater amount of trauma in men. Three were infants, all with meningococcal infections.

None of the adult cases was diagnosed or even suspected clinically. Most were traffic accidents, the remainder being post-operative, other trauma or poisoning, together with several cases of hypertension and chronic renal failure. The most frequent common denominator was trauma, whether that trauma was direct injury or surgical trauma. The sudden collapse and rapid death usually occurred a few days after the trauma, varying from 2 to 21 days.

At autopsy one or both of the adrenal glands are found to be swollen, sometimes to the size of a large walnut (3–5 cm) the cortex being tightly stretched around a large haematoma occupying the medulla. The haemorrhage is usually dark red and fairly fresh, though occasionally an area of brown bleeding indicates a previous episode.

The late Professor Hugh Johnson, of St Thomas’ Hospital, London, collected between 150 and 200 autopsy cases of suprarenal haemorrhage (unpublished) and considered that they fell into six main groups:

- Trauma of all types. Johnson was of the opinion that the predominance of right-sided gland haemorrhages was a result of the frequency of right-sided impacts on pedestrians in Britain because of the side of traffic flow.
- Perinatal, newborn or stillborn infants, especially when they were breech births or had fetal anoxia.
- Infancy and childhood, especially meningococcal and disseminated intravascular coagulation.
- Infective and post-operative septicaemia, often with Gram-negative organisms. Sometimes there was direct surgical trauma to the gland, as in gastric surgery.
- Tumour: haemorrhage into a secondary deposit or invasion of adrenal vasculature.
- Thrombosis of central vein with adrenal infarction.

REFERENCES AND FURTHER READING

Adams V, Guidi C. 2001. Venous air embolism in homicidal blunt impact head trauma. Case reports. *Am J Forensic Med Pathol* **22**:322–6.

Adkins RB, Foster JH, O’Saile D. 1962. Experimental study of the genesis of fat embolism. *Ann Surg* **156**:515–27.

Allardyce DB. 1971. The postmortem interval as a factor in fat embolism. *Arch Pathol* **92**:248–53.

Bajanowski T, West A, Brinkmann B. 1998. Proof of fatal air embolism. *Int J Legal Med* **111**:208–11.

Bajanowski T, Kohler H, DuChesne A, *et al.* 1999. Proof of air embolism after exhumation. *Int J Legal Med* **112**:2–7.

Bergentz SE. 1961. Studies on the genesis of post-traumatic fat embolism. *Acta Chir Scand* **282**:1–72.

Beyer A. 1979. Shock lung. *Br J Hosp Med* **21**:248–52.

Bonnell H, French SW. 1982. Fatal air embolus associated with pneumatosis cystoides intestinalis. *Am J Forensic Med Pathol* **3**:69–72.

Bowen DA, McKim Sycamore E. 1976. Traumatic air embolism. *Med Sci Law* **16**:56–8.

Buchanan D, Mason JK. 1982. Occurrence of pulmonary fat and bone marrow embolism. *Am J Forensic Med Pathol* **3**:73–8.

Bunai Y, Yoshimi N, Komoriya H, *et al.* 1988. An application of a quantitative analytical system for the grading of pulmonary fat embolisms. *Forensic Sci Int* **39**:263–9.

Butler BD, Hills BA. 1985. Transpulmonary passage of venous air emboli. *J Appl Physiol* **59**:543–7.

Cebelin MS, Hirsch CS. 1980. Human stress cardiomyopathy. Myocardial lesions in victims of homicidal assaults without internal injuries. *Hum Pathol* **11**:123–32.

Chau KY, Yuen ST, Wong MP. 1995. Seasonal variation in the necropsy incidence of pulmonary thromboembolism in Hong Kong. *J Clin Pathol* **48**:578–9.

Chau KY, Yuen ST, Wong MP. 1997. Clinicopathological pattern of pulmonary thromboembolism in chinese autopsy patients: comparison with caucasian series. *Pathology* **29**:263–6.

Dada MA, Loftus IA, Rutherfoord GS. 1993. Shotgun pellet embolism to the brain. *Am J Forensic Med Pathol* **14**:58–60.

Edland J. 1971. Post-traumatic fat embolism. In: Wecht C (ed.), *Legal medicine annual*. Appleton Century Crofts, New York.

Ellis HA, Watson AJ. 1968. Studies on the genesis of traumatic fat embolism in man. *Am J Pathol* **53**:245–51.

Erhen J, Nadvornik F. 1963. The quantitative demonstration of air embolism. *J Forensic Med* **10**:45–8.

Fineschi V, Centini F, Mazzeo E, *et al.* 1999. Adam (MDMA) and eve (MDEA) misuse: an immunohistochemical study on three fatal cases. *Forensic Sci Int* **104**:65–74.

Flute PT. 1970. Coagulation and fibrinolysis after injury. *J Clin Pathol Suppl R Coll Pathol* **4**:102–9.

Freiman DG. 1965. Frequency of pulmonary thromboembolism in man. *N Engl J Med* **24**:1278–82.

Gottlieb JD, Ericsson JA, Sweet RB. 1965. Venous air embolism: a review. *Anesth Analg* **44**:773–9.

Gresham GA, Kuzynski A, Rosborough D. 1971. Fatal fat embolism following replacement arthroplasty for transcenvical fractures of the femur. *Br Med J* **2**:617–18.

Guileyardo JM, Cooper RE, Porter BE, *et al.* 1992. Renal artery bullet embolism. *Am J Forensic Med Pathol* **13**:288–9.

Hamilton PJ, Stalker AL, Douglas AS. 1978. Disseminated intravascular coagulation: a review. *J Clin Pathol* **31**:609–19.

Hartveit F, Lystad H, Minken A. 1968. The pathology of venous air embolism. *Br J Exp Pathol* **49**:81–6.

Havig O. 1977. Deep vein thrombosis and pulmonary embolism. An autopsy study with multiple regression analysis of possible risk factors. *Acta Chir Scand Suppl* **478**:1–120.

Irniger W. 1962–3. Histological determination of the age of thromboses and embolisms. *Virchows Arch A Pathol Anat Histopathol* **236**:220–37.

Jackson CD, Gerendyke RM. 1965. Pulmonary and cerebral fat embolism after closed-chest cardiac massage. *Surg Gynecol Obstet* **120**:25–8.

Janssen W. 1967. [On the pathogenesis and forensic evaluation of brain hemorrhages following cerebral air embolism.] *Dtsch Z Gesamte Gerichtl Med* **61**:62–80.

Johnson HR. 1973. Delayed air embolism. *Forensic Sci* **2**:375–7.

Jörgensen L. 1964. Experimental platelet and coagulation thrombi. *Acta Pathol Microbiol Scand Sect A Pathol* **62**:189–223.

Kerner T, Fritz G, Unterberg A, *et al.* 2003. Pulmonary air embolism in severe head injury. *Resuscitation* **56**:111–15.

King P. 1970. Fat embolism syndrome. *Med J Aust* **2**:1190–1.

Knight B. 1966. Fatal pulmonary embolism: factors of forensic interest in 400 cases. *Med Sci Law* **6**:150–4.

Knight B. 1980. Sudden unexpected death from adrenal haemorrhage. *Forensic Sci Int* **16**:227–9.

Knight B, Zaini MR. 1980. Pulmonary embolism and venous thrombosis. A pattern of incidence and predisposing factors over 70 years. *Am J Forensic Med Pathol* **1**:227–32.

Knight L, Velentine EH. 1962. Spontaneous bilateral adrenal haemorrhage. *JAMA* **14**:1312.

Lau G. 1995. Pulmonary cartilage embolism: fact or artefact? *Am J Forensic Med Pathol* **16**:51–3.

Lehman EP, Moore RM. 1927. Fat embolism including experimental production without trauma. *Arch Surg* **14**:621–6.

Lequire VS. 1959. A study on the pathogenesis of fat embolism based on human necropsy material and animal experiments. *Am J Pathol* **35**:999–1003.

Levy V, Rao VJ. 1988. Survival time in gunshot and stab wound victims. *Am J Forensic Med Pathol* **9**:215–7.

Lignitz E, Lignitz G, Puschel K. 1995. [Lung embolism as the cause of death in legal medicine.] *Versicherungsmedizin* **47**:203–7.

Loehry CA. 1966. Pulmonary emboli in young adults. *Br Med J* **1**:1327–8.

Mackay DG. 1965. *Disseminated intravascular coagulation: an intermediary mechanism of disease.* Hoeber, New York.

Malhotra MS, Wright HC. 1960. Air embolism during decompression under water. *J Physiol* **151**:32–5.

Mason JK. 1968. Pulmonary fat and bone marrow embolism as an indication of ante-mortem violence. *Med Sci Law* **8**:200–6.

Melinek J, Livingston E, Cortina G, *et al.* 2002. Autopsy findings following gastric bypass surgery for morbid obesity. *Arch Pathol Lab Med* **126**:1091–5.

Mole RH. 1948. Fibrinolysin and the fluidity of the blood postmortem. *J Pathol Bacteriol* **60**:413–17.

Moore RM, Braselton CW. 1940. Injections of air and of carbon dioxide into a pulmonary vein. *Ann Surg* **112**:212–18.

Morell MT, Dunhill MS. 1968. The postmortem incidence of pulmonary embolism in a hospital population. *Br J Surg* **55**:5–10.

Nordstrom M, Lindblad B. 1998. Autopsy-verified venous thromboembolism within a defined urban population – the city of Malmo, Sweden. *APMIS* **106**:378–84.

Pierucci G. 1985. [Post-mortem diagnosis of gas embolism.] *Pathologica* **77**:145–55.

Pierucci G, Gherson G. 1968. [Experimental study on gas embolism with special reference to the differentiation between embolic gas and putrefaction gas.] *Zacchia* **4**:347–73.

Pierucci G, Gherson G. 1969. [Further contribution to the chemical diagnosis of gas embolism. The demonstration of hydrogen as an expression of ‘putrefactive component’.] *Zacchia* **5**:595–603.

Polson CJ. 1963. *Essentials of forensic medicine*. Pergamon Press, Oxford.

Rappaport H, Raum M, Horrell JB. 1951. Bone marrow embolism. *Am J Pathol* **27**:407–43.

Robertson H. 1938. A clinical study of pulmonary embolism. *Am J Surg* **16**:3–21.

Rouse DA, Hargrove R. 1992. An unusual case of gas embolism. *Am J Forensic Med Pathol* **13**:268–70.

Sandritter W, Staeudinger M, Drexler H. 1980. Autopsy and clinical diagnosis. *Pathol Res Pract* **168**:107–14.

Sanz P, Reig R, Borras L, *et al.* 1988. Disseminated intravascular coagulation and mesenteric venous thrombosis in fatal amanita poisoning. *Hum Toxicol* **7**:199–201.

Sevitt S. 1962. *Fat embolism*. Butterworth, London.

Sevitt S. 1970. Thrombosis and embolism after injury. *J Clin Pathol* **23**:86–90.

Sevitt S. 1973. Coronary thrombosis following injury and burns. *Med Sci Law* **13**:185–91.

Sevitt S, Gallagher N. 1961. Venous thrombosis and pulmonary embolism: a clinico-pathological study in injured and burned patients. *Br J Surg* **48**:475–81.

Shapiro HA. 1965. Death from delayed air embolism. *J Forensic Med* **12**:3–7.

Shapiro HA, Robertson JK. 1962. The significance of blood in the pleural cavity observed after death. *J Forensic Med* **9**:5–8.

Simpson JG, Stalker AC. 1973. Disseminated intravascular coagulation. In: Douglas (ed.), *Clinics in haematology*, vol. 2. W. B. Saunders, Philadelphia.

Simpson K. 1959. Fat embolism. *J Forensic Med* **6**:19–22.

Slater DN. 1988. Bone marrow emboli [letter]. *Am J Forensic Med Pathol* **9**:357–8.

Stalker AC. 1978. Issue: disseminated intravascular coagulation. In: Mason JK (ed.), *Pathology of violent injury*. Edward Arnold, London.

Stamatakis JD, Kakkar VV, Lawrence D, *et al.* 1978. The origin of thrombi in the deep veins of the lower limb: a venographic study. *Br J Surg* **65**:449–51.

Start RD, Cross SS. 1999. Acp. Best practice no 155. Pathological investigation of deaths following surgery, anaesthesia, and medical procedures. *J Clin Pathol* **52**:640–52.

Stein PD, Henry JW. 1995. Prevalence of acute pulmonary embolism among patients in a general hospital and at autopsy. *Chest* **108**:978–81.

Suzuki T, Ikeda N, Umetsu K, *et al.* 1984. Pulmonary fat embolism as an ante-mortem reaction in traumatic immediate death. *Med Sci Law* **24**:175–8.

Suzuki T, Ikeda N, Umetsu K. 1987. Pulmonary bone marrow embolic phenomena. Antemortem reactions in traumatic immediate death. *Am J Forensic Med Pathol* **8**:283–6.

Thali MJ, Yen K, Plattner T, *et al.* 2002. Charred body: virtual autopsy with multi-slice computed tomography and magnetic resonance imaging. *J Forensic Sci* **47**:1326–31.

Timperley WR. 1978. Disseminated intravascular coagulation in forensic pathology. *Med Sci Law* **18**:108–16.

van Ieperen L. 1985. Gas embolism as a cause of death. *Am J Forensic Med Pathol* **6**:240–3.

Voigt J. 1966. Adrenal lesions in medico-legal autopsies. *J Forensic Med* **13**:3–15.

Williams G. 1956. Experimental studies in arterial ligation. *J Pathol Bacteriol* **72**:569–74.

Wilson D, Cooke EA, McNally MA, *et al.* 2001. Changes in coagulability as measured by thrombelastography following surgery for proximal femoral fracture. *Injury* **32**:765–70.

Zaini MRS. 1981. *Medicolegal aspects of pulmonary embolism.* Medical Faculty, University of Wales, Cardiff.

Zaitoun AM, Fernandez C. 1998. The value of histological examination in the audit of hospital autopsies: a quantitative approach. *Pathology* **30**:100–4.

CHAPTER 14

Suffocation and ‘asphyxia’

- The nature of ‘asphyxia’ 352
- Mechanical ‘asphyxia’ 353
- The classic signs of ‘asphyxia’ 353
- Suffocation 357
- Smothering 358
- Choking 361
- Traumatic ‘asphyxia’ 363
- Postural ‘asphyxia’ 365
- References and further reading 365

Though traditionally every textbook of forensic medicine has a chapter entitled ‘Asphyxia’, it is a partial misnomer, as many of the conditions described under this heading are not truly asphyxial in nature. It is difficult to find a succinct alternative, however, unless each subcondition is described individually – and some deaths have a mixed aetiology. The conventional title is therefore retained in this book, though it is emphasized that ‘asphyxia’ is both inappropriate and inaccurate when applied to many of the mechanisms of death formerly ascribed to this cause.

THE NATURE OF ‘ASPHYXIA’

Common usage has led to the term ‘asphyxia’ being equated with ‘lack of oxygen’, though etymologically the word means ‘absence of pulsation’. Interestingly – if fortuitously – this original sense is sometimes more accurate in certain deaths associated with pressure on the neck, where cardiac arrest – certainly a prime reason for absence of pulsation – is the fatal mechanism, rather than hypoxia.

Reverting to the usual meaning of asphyxia, a whole series of disturbances can lead to lack of oxygen. The basic purpose of respiration is to convey atmospheric oxygen to the peripheral tissue cells. Anything that interferes with oxygen transfer can be called ‘asphyxia’, though other terms such as ‘hypoxia’ or ‘anoxia’ are more accurate, and should be preferred. The following conditions may legitimately be considered to be defects in the chain of respiration and hence are examples of asphyxia:

- Absence or reduction of oxygen tension in the external atmosphere, such as reduced barometric pressure or replacement of oxygen by an inert gas, such as nitrogen or carbon dioxide.
- Obstruction of the external respiratory orifices, as in smothering or gagging.
- Blockage of the internal respiratory passages, at either pharyngeal, laryngeal, tracheal or bronchial level.
- Restriction of the respiratory movements of the thorax, preventing the inspiration of air through patent respiratory passages, as in the so-called ‘traumatic asphyxia’ or in paralysis from brainstem or cord damage, or the use of curare-like drugs.
- Disease of the lungs that prevents or reduces gaseous interchange. Extensive pneumonia, pulmonary oedema, adult respiratory distress syndrome, diffuse fibrosis and many other conditions may lead to hypoxia, but these are rarely of forensic concern.
- Reduction in cardiac function leading to impairment of the circulation of oxygenated blood may be considered as a type of ‘asphyxia’, sometimes called ‘stagnant anoxia’. Again, the condition is rarely relevant to forensic problems.
- A reduced ability of the blood to transport oxygen can also be thought to be asphyxial in nature, such as severe anaemias and the replacement of the oxygen-combining power of haemoglobin by carbon monoxide.
- Finally, an inability of the peripheral tissue cells to utilize the oxygen delivered to them by the bloodstream is seen in conditions such as cyanide poisoning, where the cytochrome-oxidase enzyme systems are inactivated.

Several schemes of classification of asphyxia have been devised, such as those of McIntyre (1969) and Shapiro (1988).

MECHANICAL 'ASPHYXIA'

The usual relevance of asphyxia in a forensic context is in mechanical asphyxia, rather than some of the internal conditions mentioned above, which are more likely to arise as a result of natural disease or toxic conditions.

The normal oxygen content of blood varies according to the age and health of the subject. Young to middle-aged adults have almost complete saturation of their arterial blood with oxygen, at a level of 90–100 mmHg (12–13.5 kPa), whilst persons over 60 may drop to between 60 and 85 mmHg (8–10 kPa). These normal values in the younger group may be contrasted with those sufferers from mild hypoxia of about 60 mmHg (8 kPa), and those in severe to fatal hypoxia when oxygen tension falls to between 40 and 20 mmHg (5–3 kPa).

Unfortunately for pathologists, post-mortem analysis of oxygen levels in the blood is of no value in the retrospective diagnosis of hypoxia, as changes soon after death rapidly distort the distribution of gases and make the results meaningless.

Definition of terms

A number of different names have been used to describe the various types of mechanical 'asphyxia', some of which are confusing or inexact:

- **Suffocation** is a general term used to indicate death from deprivation of oxygen, either from lack of the gas in the breathable environment or from obstruction of the external air passages.
- **Smothering** is more specific, in that it indicates blockage of the external air passages, usually by a hand or soft fabric. A variety of suffocation may be called **gagging** where fabric or adhesive tape occludes the mouth to prevent speaking or shouting. While the nasal passages remain patent, air can enter, but later blockage by mucus or oedema may lead to death.
- **Choking** refers to blockage of the upper airways by some foreign body, but is also used for manual strangulation.
- **Throttling** refers to strangulation, usually by hand, the word rarely being applied to ligature strangulation.
- **Strangulation** is the most specific term, indicating the use either of the hands or a ligature as a means of applying external pressure to the neck. **Garrotting** has been used for ligature strangulation, but more accurately refers only to a Spanish method of judicial execution. Other methods exist in various parts of the world, such as the Indian technique of pressing a flexible stick or 'lathi' across the front of the neck.
- **Mugging** originally meant the application of pressure to the neck by means of an arm crooked around from the rear, but more recent American usage has widened the term to mean any kind of robbery with violence. More recently still the **arm-lock** has been used as means of restraint by law enforcement officers, sometimes with fatal results.

As an aside, in some deaths during arrest, usually where a violent struggle occurs between several police officers and a drunken or drugged offender, the death of the latter occurs where no neck or chest pressure can be implicated. Though the mechanism is obscure and impossible to demonstrate objectively at autopsy, it has been provisionally ascribed to a catecholamine-induced cardiac arrhythmia from an exaggerated adrenal response.

As most of these victims are given energetic cardiopulmonary resuscitation, any markers of this catecholamine effect, such as contraction-band necrosis in the myocardium, cannot be differentiated from resuscitative artefacts.

THE CLASSIC SIGNS OF 'ASPHYXIA'

For many years the autopsy diagnosis of an asphyxial death was made by reference to a set of findings that have come to be known as the 'classic signs of asphyxia'. Unfortunately, it is now quite clear that most of these signs are so non-specific that little reliance can be placed on them in the absence of other confirmatory evidence. They are scathingly referred to by Lester Adelson (1974) as the 'obsolete diagnostic quintet'.

In many cases of undoubted fatal hypoxia these signs are absent and, conversely, they are often present in some degree in conditions that can be shown to be non-hypoxic in origin. The major difficulty for the pathologist is that 'asphyxia' cannot reliably be equated with 'hypoxia'. There are no truly distinctive autopsy signs of pure hypoxia and most of the alleged criteria are caused by factors other than a lack of oxygen.

To confuse the issue further, some apparently hypoxic states lead to sudden or rapid death before there is time for oxygen lack to take effect, such as often happens when a victim enters a space devoid of oxygen, has the trachea suddenly blocked by food or has a plastic bag pulled over the head.

Each of the 'classic features' must be looked at in depth.

Petechial haemorrhages

These are small pin-point collections of blood lying in the skin, the sclera or conjunctivae and under thoracic serous membranes such as the pleura or pericardium. They vary in size from a tenth of a millimetre to about two millimetres. If larger than this, they are more accurately called 'ecchymoses'.

Petechiae are often known as 'Tardieu's spots', but this eponym should be restricted to those lying in the visceral pleura, where they were described by the Parisian Professor Ambroise Tardieu in 1866, in the bodies of infants who he claimed had been 'overlain'. A further common error is to attribute the petechiae to the rupture of **capillaries**, whereas they actually emanate from small venules – capillary bleeding would be invisible to the naked eye.

Petechiae are caused by an acute rise in venous pressure that in turn causes overdistension and rupture of thin-walled peripheral venules, especially in lax tissues, such as the eyelid, and in unsupported serous membranes, such as the pleura and epicardium. It is traditionally claimed that hypoxia of the vein wall is an added factor, but there is no experimental proof of this conjecture, which seems unlikely considering the rapid formation of these lesions when venous pressures rise; for example, petechiae can appear almost instantly after violent sneezing or coughing before any hypoxia is possible.

Petechiae are seen most often in the face and eyes of victims of compression of the neck or fixation of the chest. In the former, where strangulation by either hand or ligature has occurred, occlusion of the jugular veins prevents venous drainage from the head, whilst almost always allowing the arterial supply to continue via the more deeply placed and much less compressible carotid and vertebral arteries. There is a rapid rise in venous pressure in the head, leading to engorgement of the veins. The consequences are swelling of the tissues, both from an increase in intravascular volume and from rapid transudation of fluid into the tissues, followed by showers of petechial haemorrhages in unsupported areas, such as the skin of the upper eyelids, the forehead, the skin behind the ears, the circumoral skin, and the conjunctivae and sclera of the eye. The same venous engorgement often leads to frank bleeding from the nasal mucosa and the external auditory meatus.

In the internal tissues and organs, petechiae are most often seen on serous membranes in the thorax, where the presence of a body cavity leads to lack of support for the superficial venules and hence predisposes to rupture when venous pressure rises. They are almost never seen in the peritoneal serosa and it has been postulated that, in the chest, attempts at breathing against a closed airway may lead to a sudden extreme decrease in intrapleural pressure causing petechiae in the same way as suction on the skin can cause a shower of pin-point haemorrhages. As described by Tardieu, petechiae in the thorax are often seen on the visceral pleura, especially in the interlobar fissures and around the hilum.

Petechiae are also common on the surface of the heart, especially on the epicardium around the atrioventricular groove, particularly on the posterior surface, though these have been shown to appear or worsen as a post-mortem phenomenon. In infants and children, the thymus or thymic remnants may show numerous petechiae. It is claimed that in the sudden infant death syndrome these are confined to the cortex, whereas in other 'asphyxial' conditions they are scattered throughout the gland; this is doubtful.

Petechiae are rarely seen in the parietal pleura or the peritoneum except in haemorrhagic diatheses. In the brain, petechiae occur in the white matter and there may be larger patches of bleeding in the subarachnoid space where superficial vessels have ruptured because of acute venous engorgement. The same mechanism often produces profuse petechiae and ecchymoses under the scalp. In regard to the scalp, caution must be observed as numerous haemorrhages can be caused as a common autopsy artefact during the reflection of the scalp flaps.

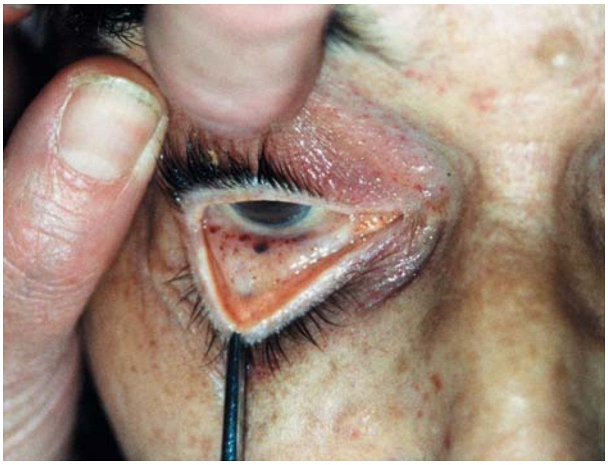

FIGURE 14.1 *Petechial haemorrhages in the eye in manual strangulation. There are also some petechiae on the facial skin.*

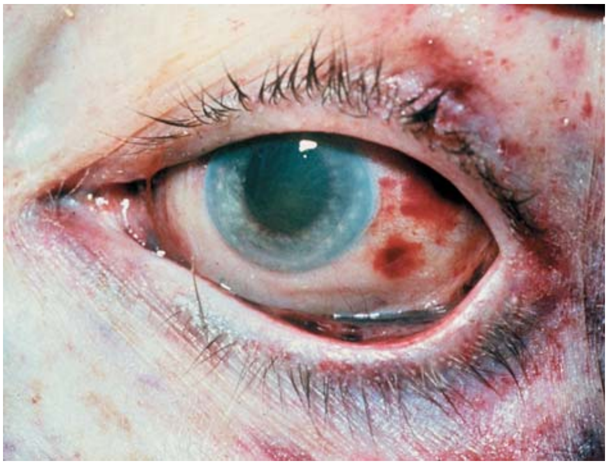

FIGURE 14.2 *Scleral haemorrhages in manual strangulation. These are larger than the usual petechiae, but the latter are present inside the lower lid.*

Petechiae may also disappear with lengthening post-mortem interval. Betz *et al.* (1994) have shown that in putrefaction and freshwater immersion, conjunctival petechiae may vanish.

THE SIGNIFICANCE OF PETECHIAE

A number of factors cause difficulty in the interpretation of petechial haemorrhages. First, there is no doubt that both cutaneous and visceral petechiae, especially the latter, can both appear and enlarge as a post-mortem phenomenon. Gordon and Mansfield (1955) have shown that the presence and number of petechiae is a function of the post-mortem interval, in common with other haemorrhagic artefacts such as those found behind the larynx by Prinsloo and Gordon (1951).

Posture also has an effect on the appearance of haemorrhages. They are commonly seen, along with larger ecchymoses, on the front or back of corpses who have died from a variety of causes in which mechanical asphyxia is absent. They are often present in normal post-mortem hypostasis, especially where the mode of death was congestive as in many types of natural heart disease. Their appearance in the hypostatic areas is also time-related, the longer the post-mortem interval, the more likely – and more prominently – they are to occur.

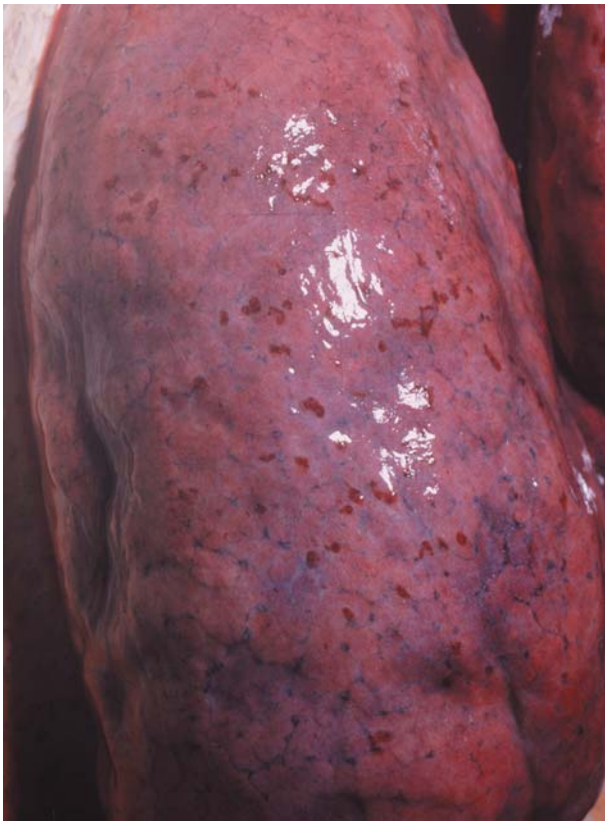

FIGURE 14.3 *Subpleural petechial haemorrhage – the true 'Tardieu spots' – in manual strangulation. These are no longer considered to be specific for 'asphyxia' and a few petechiae can frequently be found in many autopsies unassociated with asphyxia.*

Where some abnormal posture is present – such as the deceased being slumped out of bed or otherwise found with the head lower than the body – marked congestion, cyanosis and petechial haemorrhages are common.

Occasionally the abnormal posture may have occurred during life, such as in overdose victims, drunks and senile persons, who have fallen into a position with the upper part of the body lowermost. This may embarrass respiration and contribute to death and the ante-mortem component may be difficult or impossible to distinguish from post-mortem accentuation of the congestive-petechial appearances.

Another problem in the autopsy interpretation of petechiae is that not all punctate lesions in the pleura are petechiae. Zaini and Knight (1982) showed that many such apparent petechiae were either intravascular venous pools, subpleural microbullae or pigment foci.

In summary, petechial haemorrhages are highly unreliable indicators of an asphyxial process, taking this to mean a hypoxic state. They are the result of venous engorgement, usually from mechanical obstruction of venous return to the heart – or in the thorax, from attempts to inspire against a blocked airway. Petechiae and ecchymoses are common non-specific autopsy findings and many are post-mortem in origin, especially in dependent positions. They

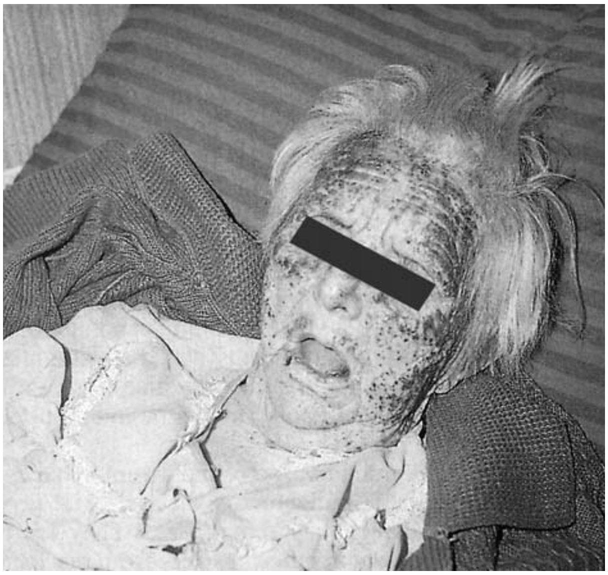

FIGURE 14.4 *Dense confluent skin haemorrhages in the face of an old lady who had slipped from her bed and was found with her head on the floor and her legs on the bed. The signs of this 'postural asphyxia' may well have been accentuated by post-mortem worsening of the hypostatic haemorrhages.*

may occur in many non-asphyxial states and, in the lungs, some petechiae can be found in the interlobar fissures and around the hilum in most routine autopsies.

Conversely, in some types of death where oxygen deprivation is to be expected (such as drowning, plastic-bag suffocation and entering an atmosphere devoid of oxygen), petechiae are seldom demonstrable.

The most significance must be placed on facial and ocular petechiae, because showers of pin-point haemorrhages in the eyelids, conjunctivae, sclera and facial skin require urgent explanation unless the body was face down or head down.

Congestion and oedema

The next 'classic' sign of asphyxia is congestion. This is even more non-specific than petechiae and once again is the result of obstructed venous return. When the neck is compressed, the face, lips and tongue become swollen and reddened. The true colour change of congestion is usually darkened by the onset of cyanosis. Internal organs also become congested and in strangulation this is most notable in the tongue, pharynx and larynx above the level of venous obstruction. In pressure on the chest, failure of respiratory movements cause an intense rise in venous pressure, accompanied by cyanosis.

Congestion is often associated with tissue swelling if the venous obstruction continues. The oedema is the result of rapid transudation through capillary and venule walls, again mainly a function of back pressure in the venous system. Hypoxia of the vascular endothelium is alleged to allow increased permeability, but generalized hypoxia from other causes does not produce the tissue swelling seen in strangulation.

Tissue fluid also escapes rapidly into the brain in strangulation that continues for even a few minutes, though here the well-known effect of hypoxia might be an added factor. Pulmonary oedema, causing excess fluid to enter the alveoli, is often found in hypoxic deaths. Here the mechanism is more obscure, but is probably a combination of hypoxia and raised pulmonary vessel pressure. In strangulation, froth may sometimes be so profuse as to emerge from the mouth and nostrils – yet in other throttlings, it may be absent. Pulmonary oedema is such a common and non-specific phenomenon in a whole range of fatal conditions that it has little diagnostic significance.

Cyanosis

The colour of blood depends upon the absolute quantity of oxyhaemoglobin and reduced haemoglobin present in the erythrocytes. The normal pink colour of well-oxygenated skin may change to purple or blue when oxygen is lacking – indeed, the word 'cyanosis' is derived from the Greek, meaning 'dark blue'. Cutaneous cyanosis, however, depends on the absolute amount of reduced haemoglobin, rather than the proportion of reduced haemoglobin to oxyhaemoglobin. It is not apparent in marked anaemia, even if the ratio of oxyhaemoglobin to reduced haemoglobin is low. There must be at least 5 g of reduced haemoglobin per 100 ml blood before cyanosis becomes evident, irrespective of the total amount of haemoglobin.

In the common forensic event of constriction of the neck, cyanosis almost invariably follows congestion of the face, as venous blood containing much reduced haemoglobin after perfusing the head and neck is dammed back and becomes more blue as the blood accumulates. If and when the airway becomes blocked, then impaired oxygenation in the lungs leads to a diminution in the oxygen content of arterial blood. This will lead to a darkening of all organs and tissues, and will accentuate the cyanosis of the face. This does not happen in the first phases of strangulation, however, and depends on complete or substantial occlusion of the airway, or restriction of the respiratory excursions of the chest.

Cyanosis produced during life may be partly or wholly overshadowed by hypostasis, which may be a deep purple or blue, and may be mistaken for true cyanosis – indeed, some pathologists refuse to use 'cyanosis' in respect of the dead, claiming that it cannot reflect the ante-mortem situation.

'Engorgement of the right heart and fluidity of the blood'

Descriptions of an abnormal fluidity of the blood seen at autopsy in asphyxial deaths are part of forensic mythology and can be dismissed with little discussion. Post-mortem clotting in the heart and venous system is a most erratic process, as is the eventual dissolution of those clots by the action of fibrinolytic enzymes (Mole 1948). It is irrelevant in the diagnosis of asphyxia.

Engorgement of the chambers of the right side of heart and the great veins is also a non-specific autopsy observation, useless as a marker of an asphyxial process. Any type of congestive death, including primary heart failure from many diseases, leads to terminal engorgement of the right ventricle and atrium as part of the generalized rise in venous and intracardiac pressure.

The autopsy diagnosis of 'asphyxia'

Thus there are no specific autopsy findings for 'asphyxia', taking this in its hypoxic sense. As stated earlier, the so-called 'classic signs' have been well described by Adelson (1974) as the 'obsolete diagnostic quintet'. He goes on to observe that 'the co-existence of these findings, in themselves, does not prove that death resulted from mechanical asphyxia. All these phenomena are non-specific and are in no way peculiar to this mode of death. Inasmuch as they are observed frequently in deaths arising from unquestioned natural disease, they are of no value in proving that death resulted from mechanical asphyxia'.

Other salutary cautions on the significance of 'asphyxial signs' may be found in the important publications of Gordon and Turner (1951), Camps and Hunt (1959), Shapiro (1953) and Swann and Brucer (1949).

It is impossible to make a post-mortem diagnosis of acute hypoxia by measuring the blood gases, as rapid changes after death – and indeed agonal changes – make analysis worthless. Only by a careful assessment of the history and circumstances of the death, exclusion of other causes and a cautious evaluation of the signs described above can any conclusion be reached. Most important of all is the finding of a cause for airway obstruction or other local trauma, such as prolonged pressure on the neck or chest, obstruction of the airways, postural causes or occlusion of the external respiratory orifices.

It cannot be emphasized too strongly that the mere finding of any of the non-specific features, such as congestion and petechiae, without firm circumstantial or preferably physical evidence of mechanical obstruction of respiration, is quite insufficient to warrant a speculative diagnosis of asphyxia. If such collateral evidence is not forthcoming, then the cause of death must be left undetermined.

Histological and biomechanical diagnosis of anoxia/'asphyxia'

Many attempts have been made to find markers of hypoxia, anoxia and 'asphyxia' by means of laboratory techniques, including histology, histochemistry and various biochemical methods. Although many claims have been made for their usefulness, the fact that such tests are virtually never put forward in criminal or civil litigation makes it self-evident that they are not reliable.

This is hardly surprising, given the vague and contentious nature of the 'asphyxial' condition, where the mechanisms are diverse and even the nomenclature is unstandardized. If the target for research is narrowed down to severe hypoxia at a tissue or cellular level, then perhaps more success might be expected, by being able to demonstrate cellular damage by one or more of the great battery of techniques now available. However, even this limited objective is not yet attained with the degree of reliability required for legal purposes. The extensive strangling, drowning and gassing of legions of small animals may have produced many theses and papers, but has had no practical impact upon a long-standing problem in forensic pathology. Cell death, especially in neurones or myocytes, due to ischaemic/hypoxic damage is the most common focus for research, but the problem in a forensic context is that a considerable period of hypoxia – usually a minimum of many minutes or even hours – is needed before changes can be detected, even in experimental circumstances. In human autopsy material, the ever-present post-mortem and agonal changes interfere with the subtle early signs of hypoxic damage.

Histological changes in the lung and also chemical markers in blood or vitreous, such as hypoxanthine, have been claimed to indicate generalized hypoxia, but the application of such techniques in routine practice has not been fruitful.

SUFFOCATION

Although not a specific term, suffocation usually refers to a death caused by reduction of the oxygen concentration in the respired atmosphere, formerly called a 'vitiated atmosphere'. It is less often used to include smothering or choking. Reduction of atmospheric oxygen may occur in a wide variety of situations. Decompression, such as cabin failure of aircraft at high altitudes, causes a dramatic fall in the partial pressure of oxygen and hence reduced penetration of the gas through the alveolar wall. Other factors, however, such as the direct effects of a partial vacuum and often mechanical trauma from the near-explosive situation may overshadow the hypoxic aspects.

More common is the reduction of oxygen in the atmosphere by physical replacement by other gases, or chemical changes such as combustion. In fires, the loss of respirable air may be a potent factor in causing death, though other complications, such as the presence of toxic gases such as carbon monoxide, cyanide and many other poisonous substances liberated by the burning of plastics, may cause death more quickly than pure hypoxia. Carbon dioxide, which though not itself poisonous, is irrespirable and may accumulate in fires, and in wells and shafts in limestone. In former years, vagrants sleeping for warmth near limekilns were sometimes suffocated by this heavy gas creeping over them.

Carbon dioxide is also the cause of deaths in a more modern agricultural setting – the grain silo. Here, many tons of grain are stored in gas-tight towers; the seed produces carbon dioxide that settles to the bottom of the tower. When a blockage occurs in the gravity discharge, farm workers may enter the tower to clear the obstruction. Although safety precautions demand venting before the men enter, some workers still suffer a sudden death on encountering an atmosphere rich in carbon dioxide.

A similar hazard exists in ships' tanks or other industrial metal chambers, in which oxygen is replaced by nitrogen. This happens because the damp steel walls become rusty and use up much of the contained oxygen in forming ferric oxides. In all these deaths associated with replacement of oxygen with an inert gas, rapid death is common before hypoxia can have had any physiological effect. For example, the author (BK) has dealt with two deaths in which seafarers entered closed ships' tanks and virtually fell dead off the entry ladder. The presumed mechanism, which was far too quick to be hypoxic, was considered to be some overstimulation of the chemoreceptor system, leading to a parasympathetic 'vasovagal' cardiac arrest.

Even where death was not witnessed as sudden, the 'classic signs' of asphyxia are almost always absent. In domestic circumstances, death may be seen where a heating apparatus has removed oxygen from the atmosphere in the absence of ventilation. Though oxides of carbon are usually formed, a kerosene or natural gas appliance may kill from pure hypoxia, especially where left burning all night in a small room where the occupants lie sleeping. The effect may be accentuated by the victims having blocked up cracks in the doors and windows to keep out draughts. An open wood or coal fire does not present the same hazard, as it requires a flue or chimney in order to burn. In all such deaths, carbon monoxide poisoning must first be excluded by blood analysis, as it is a common accompaniment to lack of oxygen, especially as the reduction in oxygen availability tends to make the heat source produce progressively more monoxide than dioxide during combustion.

In a different variety of hypoxic death, persons – especially children – may asphyxiate by being confined in a small airtight space. Examples include boxes and discarded refrigerators; the danger of the latter is so well known that in Britain it is illegal to dump a refrigerator in a place with public access, unless the self-locking handle is rendered inoperative.

In all such 'hypoxic' deaths (though some are patently due to vasovagal cardio-inhibitory mechanisms) it is very rare to find any petechial haemorrhages, as these are mainly the result of venous obstruction, which is absent in these circumstances. In the true hypoxic deaths, there may be congestion and cyanosis, though even these are often absent and the autopsy findings are essentially negative.

SMOTHERING

This term refers to death from mechanical occlusion of the mouth and nose, though sometimes 'suffocation' is used to include this class of death. The smothering agent is usually fabric, an impervious sheet or a hand, though occasionally (especially in industrial accidents) a mobile solid, such as sand, mud, grain or flour may be responsible for blocking the air passages. In the 1966 Aberfan disaster in South Wales, more than 140 victims – almost all children – were smothered when semi-liquid coal slurry from a collapsing mine tip swept over their school.

In smothering, death may occur either by the occluding substance pressing down upon the facial orifices, or by the passive weight of the head pressing the nose and mouth into the occlusion. Deliberate homicide is seen usually in the old, the debilitated and in infants. It is extremely difficult to prove homicide from objective findings.

In relation to infants, the matter will be further discussed in the chapter on sudden infant death syndrome, but it is essential to appreciate that the smothering of babies, whether intentional or accidental, is both rare and difficult to prove. The so-called 'classic signs' of asphyxia, for what they are worth, are rarely present in proven suffocation – and as intrathoracic petechiae are common in undoubted 'cot deaths', these signs cannot therefore be accepted in isolation as evidence of suffocation.

Pressure marks on the face can rarely be distinguished from post-mortem postural changes, where circumoral and circumnasal pallor is caused merely by passive pressure of the dependent head after death, preventing the gravitational hypostasis from entering these areas.

Even where the head is found supine, variation in colour is still common on the face, with contrasting white and pink patches, which usually change as the post-mortem interval lengthens. Unless there are bruises or abrasion on the cheeks, around the mouth, lips or lesions within the lips or mouth, it is dangerous to overinterpret mere colour variations from

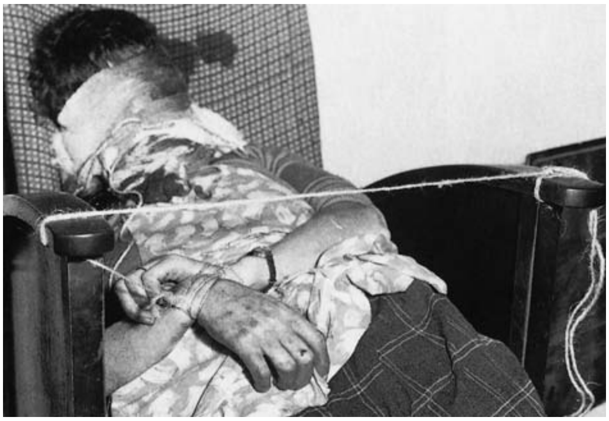

FIGURE 14.5 *Smothering from a towel being tied over the face. Death was contributed to by head injuries, also inflicted by the aged husband, who died within a few hours of natural causes from hypertensive heart disease.*

alterations in the amount of blood in the facial capillary bed, which is almost always a post-mortem phenomenon.

Even a confession from the mother is not always reliable – cases are on record where a mother has falsely confessed to smothering her child to rationalize the inexplicable events of a true sudden infant death syndrome fatality.

A similar situation exists with the elderly, who may be victims of a 'mercy killing', often a euphemism for the exhaustion of patience of long-suffering caring relatives. A pillow placed over the face of a sleeping octogenarian leaves no signs, **unless a struggle develops**, when protracted attempts at respiration against obstruction may lead to congestion, cyanosis and sometimes facial and conjunctival petechiae.

'Overlaying' of infants

This alleged condition has an ancient pedigree, being mentioned in the Old Testament. In Chapter III of the First Book of Kings, Solomon has to adjudicate between two women who both claim the same child, as the other infant was 'overlain'. This event, some 3000 years ago, has been mentioned at intervals through history, as in Wales in AD 1188, when Giraldus Cambrensis records that the Lord punished a woman for preventing her husband joining the Third Crusade, by causing her infant son 'to be overlain in the night'.

When an infant was found dead in the morning in the maternal bed (as separate cots or cribs are a relatively modern invention), it was assumed that the mother had turned over onto the baby in her sleep and suffocated it. When infants began to be placed in cots, the deaths continued unabated and it seems obvious that most of these were victims of the sudden infant death syndrome (SIDS). Whether or not overlaying really exists is doubtful and seems incapable of proof, given that any infant found dead in bed with no physical signs at autopsy could, by definition, be a SIDS.

The recent marked decline in numbers of SIDS in Britain has coincided with a campaign to discourage mothers from putting their infants to sleep face down – but, as yet, no definite cause-and-effect relationship has been proved, so any consequent strengthening of the 'overlaying' hypothesis cannot be substantiated.

Plastic bag suffocation

Although an increasingly common form of suicide in Britain, plastic bag suffocation may also be homicidal or accidental. In all these types, the essential mechanism is that a hood of impervious substance, usually polythene or other plastic, is placed over the head down to neck level. The plastic is usually in the form of an open-ended bag, either transparent or a 'supermarket' shopping bag.

Although many suicides tie the open end of the bag around their neck with cord or a tie, this is not necessary for a fatal result. Indeed, even flat sheets of polythene have killed infants when placed on the face. The mechanism is not understood, as it was formerly but erroneously thought to be the result of the clinging effect of static electricity.

Plastic bag suffocation can be rapid and leave no signs whatsoever. In the author's (BK) series of accidental, suicidal and three homicidal deaths from plastic bags, not one had any petechial haemorrhages or, indeed, any signs of 'asphyxia' at all, the faces being pale and uncongested. In another case, a person was convicted of murder by plastic bag, yet was only arrested after a spontaneous confession 6 weeks after an autopsy that had revealed no signs whatsoever of an asphyxial cause of death. As with hypoxic atmospheres, it seems that the mechanism of death in plastic-bag facial occlusion is some rapid cardio-inhibitory mode, rather than a purely hypoxic process. This conclusion is strengthened by the causes mentioned above, where flat sheets of plastic, not bags, have killed infants by clinging to the face, which of necessity must have been rapid deaths.

At autopsy, unless the bag is still present, these cases can present great difficulty. In fact, as in the homicide mentioned above, unless a confession or other corroborative evidence is forthcoming, the pathologist may never even know that he has dealt with an occult suffocation. Where the bag

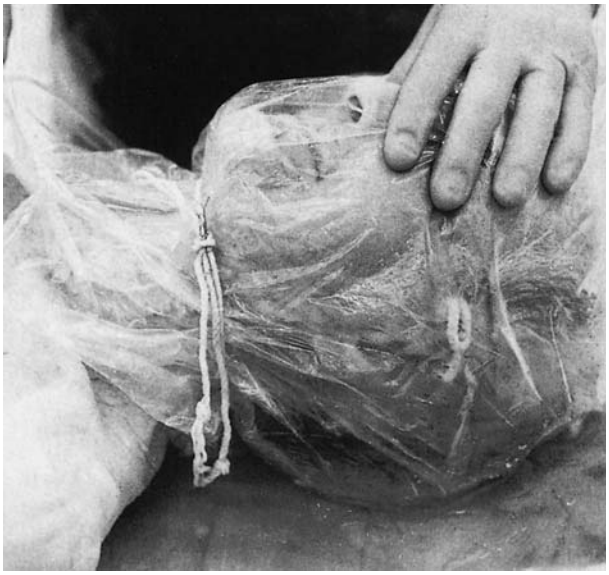

FIGURE 14.6 *Plastic bag suicide. The bag is sometimes tied around the neck, as in this case, but this is not necessary for a fatal outcome. The deaths are not asphyxial in the hypoxic sense; facial congestion and petechiae are almost always absent.*

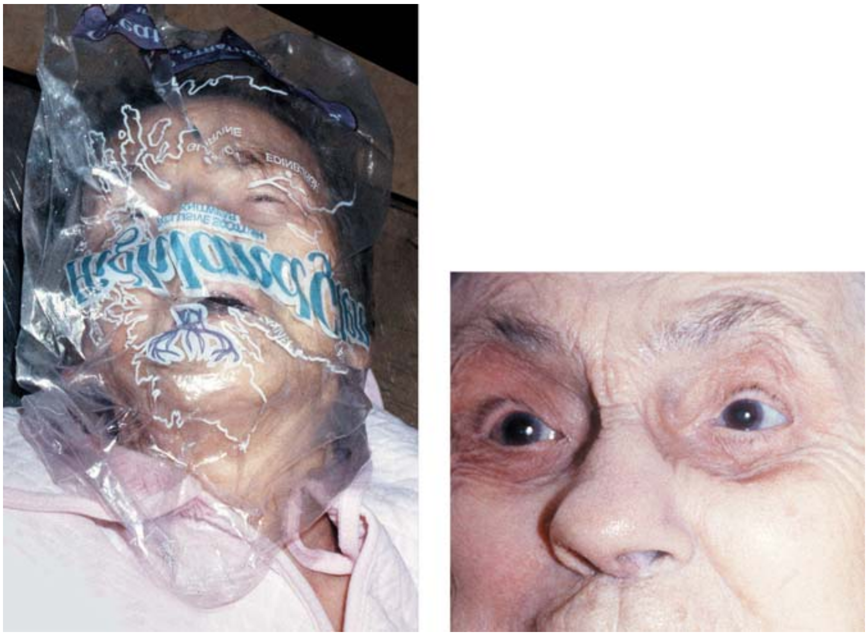

FIGURE 14.7 *Plastic bag suicide, with an open bag placed loosely over the head. As usual there are no congestive petechial signs in the face.*

is still in place, a search must be made for other indications of suicide, such as drug analysis and futile injuries, such as slashed wrists. Masochistic activities, described later in this chapter, are sometimes, though not often, associated with plastic bag suffocation.

It is sometimes said that moisture seen inside the bag indicates that it was put on during life, the water being from respiration. It is rare for this to be an important issue, as placing a bag on a corpse’s head would seem excessively unusual; in any event, the test is invalid, as evaporation from the skin, nose and mouth can produce beads of moisture inside, whether the victim is dead or alive.

Autopsy signs of suffocation

Where smothering is suspected, local signs must be sought to try to substantiate pressure on the face. Such signs include bruising around the mouth, chin and nose, though these are rarely seen except in the more violent incidents. Pressure of the lips on the teeth or dentures may cause the buccal surfaces to be bruised or abraded, though lacerations are rare unless a blow has been administered. It must be remembered that as small infants and many old people have no teeth, these injuries are less likely. The dangers of accepting areas of facial pallor as ‘pressure marks’ when the face shows post-mortem hypostasis has already been mentioned. In the early stages of post-mortem discoloration of the skin, variable colouring may not be truly ‘hypostatic’ as mottling and patchy variation is common due to uneven vasodilatation after death. It takes some time for gravity to pull cutaneous blood down to lower levels.

Smothering may occur accidentally where a person who is incapacitated from one of a variety of causes lies face down on a surface that is impervious to air. Drunkenness, epilepsy, drug overdose, coma or stupor from natural disease, may lead to this. Though it has long been proved in connection with ‘cot death’ that woven fabrics of most types can still admit enough air to support quiet respiration when placed over the mouth and nose, fabric that becomes saturated with saliva, nasal mucus or vomit may form an impervious seal that ceases to allow the passage of air. The weight of the inert head also distorts the nose and presses the nostrils and mouth so hard against a mattress or pillow that this adds to the airway stenosis.

In these instances, congestion and cyanosis are common, but when the body is examined after death, some of these findings may be postural and hypostatic in nature. As always, however, it must be recognized that skin haemorrhages often appear post-mortem in dependent hypostasis.

Gagging

Smothering may also occur where a pad or gag is fixed over the face, as sometimes happens in robbery with violence. A householder, caretaker or night-watchman may be tied up and silenced by having a scarf, tie or other fabric tied around his face. At first this admits air, but when it becomes progressively soaked with saliva or mucus it may become impervious and lead to an asphyxial death, again usually without any petechiae in the face or eyes.

A similar situation arises when a gag is thrust into the mouth to silence a watchman or security officer. Another variant is adhesive tape applied over the mouth. Though the nasal airway may originally be quite patent, any later obstruction from oedema and mucus, or movement of the gag backwards into the nasopharynx, may cause an unforeseen death and turn a robbery charge into a homicide.

CHOKING

This term refers to blockage of the internal airways, usually between the pharynx and the bifurcation of the trachea. Death can be the result of pure hypoxia from occlusion of the airway, when all the attendant signs of congestion, cyanosis and perhaps petechiae may be present, usually where the victim struggles to breathe for an appreciable period. As described below, however, a large proportion of deaths occur suddenly before any possible hypoxic manifestations have time to take effect; these fatalities must be caused by neurogenic cardiac arrest, either purely neurogenic or accelerated by excess catecholamine release from the adrenaline response.

Causes of choking include the following.

Foreign bodies

Objects such as gags, dummies, small toys, table-tennis balls and a multiplicity of objects may be placed in the mouth and inhaled, usually by children and the mentally retarded. Occasionally adults will do the same, accidentally or by design; an example of the latter was a suicide seen by the author (BK), who died (without any asphyxial signs whatsoever) by pushing a closed pill-bottle into his pharynx.

Dentures and haemorrhage

False teeth (especially partial plates), extracted large teeth, blood clots, and frank haemorrhage following dental or ear, nose and throat operations such as tonsillectomy.

Acute obstructive lesions

Lesions of the glottis or larynx, such as the oedema of acute hypersensitivity (including insect stings), irritant vapours, inhalation of hot gas and acute infective conditions. The most dangerous of the latter is diphtheria or *Haemophilus influenzae* epiglottitis in children, which is a medical emergency sometimes requiring an immediate tracheostomy to relieve the airway obstruction. At autopsy, great thickening of the epiglottis and aryepiglottic folds by jelly-like oedema and inflammatory tissue will be found occluding the entrance to the larynx.

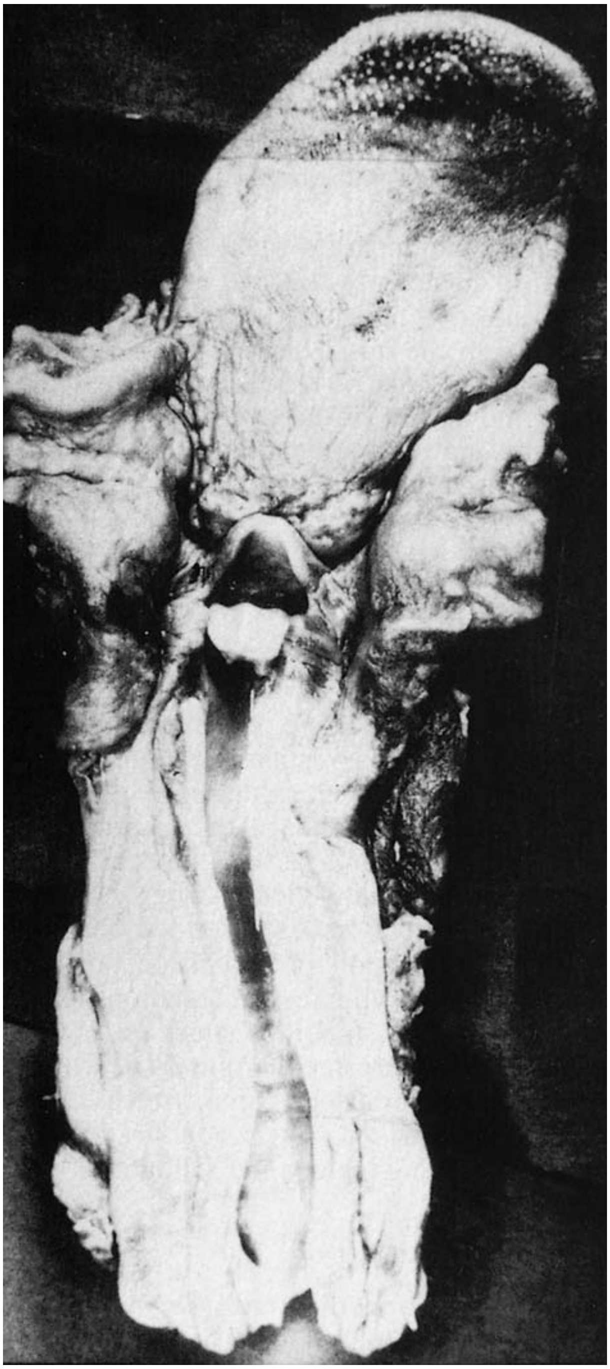

FIGURE 14.8 *Impaction of an extracted molar tooth in the larynx. Though now uncommon with modern dental and anaesthetic practices, it can still happen in less satisfactory circumstances. Blood clot from dental extraction, tonsillectomy, and other ear, nose and throat surgery can also block the airway.*

Food material

This is an important topic, as the cause of death ‘aspiration of vomit’ is used too often without real justification. Food

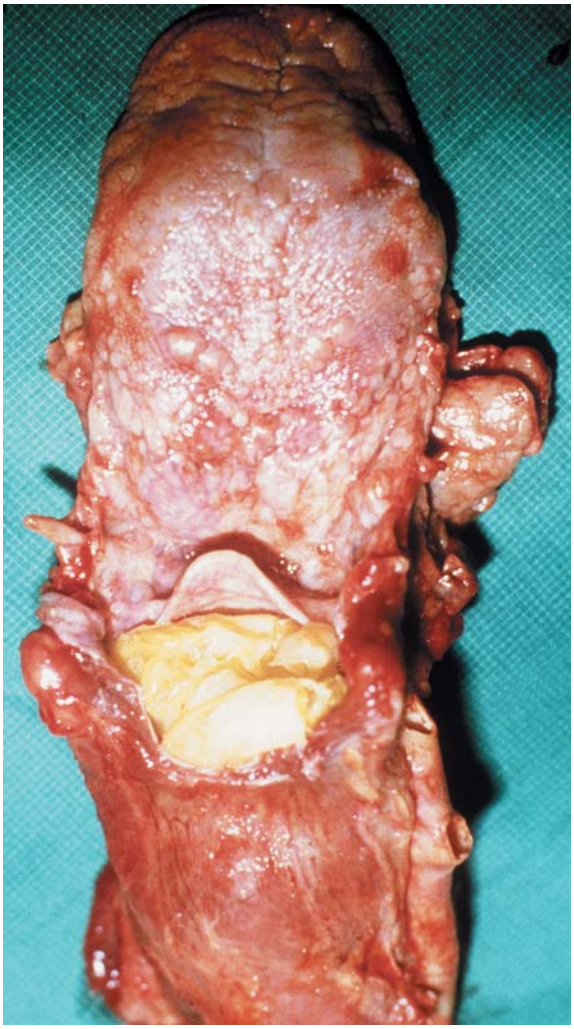

FIGURE 14.9 *Blockage of the larynx by a complete small orange in a mentally disturbed patient. The fruit had just been swallowed and was not regurgitated from the stomach. The face was pale and death was rapid, unaccompanied by symptoms of choking.*

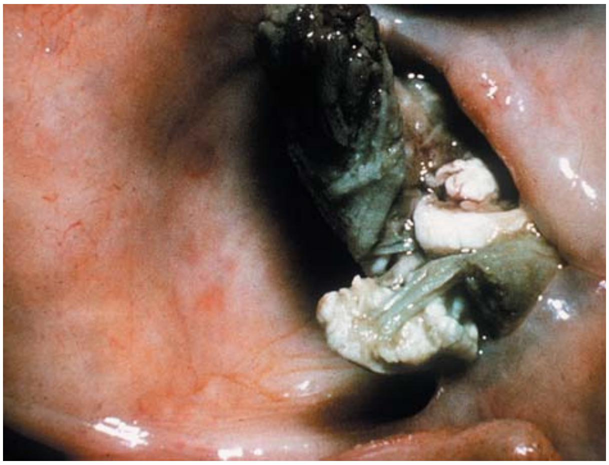

FIGURE 14.10 *Blockage of the larynx by meat. The victim was an old person with senile dementia who gulped food and who eventually died suddenly without asphyxial signs from reflex cardiac arrest.*

may be drawn into the larynx either as it is being taken down from the mouth in the act of swallowing, or it may be regurgitated from the stomach.

- The first instance is rarely misinterpreted, as whole undigested food may be found in the air passages and there is usually a history of death taking place while eating. Though most commonly encountered in old persons and the mentally disturbed, it can occur in any age group. A notable example was the so-called 'café coronary' syndrome, which received considerable publicity in the USA some years ago. The most frequent victims were well-nourished businessmen, who died suddenly and unexpectedly during a meal with no signs of respiratory distress or any of the 'classic signs of asphyxia'. Initially thought to be coronary heart disease, autopsy revealed a bolus of food, often steak, lodged in the pharynx or larynx. In any choking on food, the bolus can be quite large, such as a whole pancake, a whole tangerine or masses of meat, fruit or vegetable. Senile persons in retirement homes and mentally retarded children in institutions are also vulnerable. In such fatalities, the history and mode of death gives no hint of a hypoxic mechanism. Many cases have been observed where the victim merely sits back in their chair, dead – the mode clearly being cardiac arrest, presumably from overactivity of the parasympathetic nervous system from stimulation of the laryngeal or pharyngeal mucosa – the so-called 'vasovagal reflex' or 'reflex cardiac inhibition'.
- The finding of gastric contents in the air passages is by no means as significant as the presence of freshly swallowed food. First, difficulties can arise in distinguishing gastric contents from fresh food material if digestion of the former has not proceeded far – which is not always a function of the time since the last meal, as any physical or psychological disturbance can retard or completely halt digestion. The history, if available, is a better guide, unless the material is obviously partly or wholly digested. In cases of doubt, the smell and acid reaction to pH indicators may be useful.

Gastric contents are commonly found in the larynx, trachea and bronchi at autopsy when no other evidence of aspiration exists, and when there is a clear and unconnected cause of death. Knight (1976) found that no less than a quarter of a series of 100 consecutive autopsies on both adults and children contained some gastric contents in the air passages, and Pullar (1984) observed that this was a low figure compared with his series. These findings were obviously caused in the great majority of cases by agonal or post-mortem spillage.

Gardner (1958) described experiments in which barium was placed in the stomachs of recently dead patients while still in bed in the hospital wards. When X-rays were taken after they were first moved to the mortuary and thence into the autopsy room, most had barium in their tracheobronchial tree, confirming that overspill was common as a post-mortem phenomenon. Even histological evidence of leucocyte clustering around foci of gastric contents deep in the bronchi was shown by Gardner to be an early post-mortem event, not a 'vital reaction'. Leucocyte migration in the skin up to 16 hours post-mortem was demonstrated by Ali (1986).

Almost the only definite evidence of aspiration of gastric contents is either reliable witnessed observation during life or the histological finding of an advanced 'vital reaction' with infection, necrosis and a definite inflammatory reaction. This is a relatively late change and cannot be seen where death occurred within a few hours of aspiration. Finding a few foci of leucocytes around a bronchiole is of no evidential use. There is therefore no reliable method of distinguishing agonal or even early post-mortem overspill from true vital aspiration, unless clinical or other witnessed evidence is available. In most instances, it is not justifiable for a pathologist to claim that death was caused by inhalation of stomach contents without such confirmatory evidence. Unfortunately, where other lesions are absent or unconvincing at autopsy, especially in the sudden infant death syndrome, many pathologists use the presence of gastric contents in the air-passages as the primary and often the only cause of death. This unwarranted assumption may have distressing consequences, both in the field of medical negligence, deaths in custody and especially in the sudden infant death syndrome, where mothers may be misled into thinking that some negligence on their part in failing to observe vomiting, may have caused the death.

The only circumstance where a firmer opinion may be held is in acute alcoholism, though even here certainty is usually elusive. Where an undoubtedly drunken person (usually with a blood alcohol of at least 150 mg/100 ml) is found dead with massive blockage of his air passages by copious gastric contents and where other causes of death can be excluded by autopsy, then it might be reasonable to assume that regurgitation may have killed him, especially if there is other evidence of external vomit on the clothing or immediate surroundings. However, it is not an autopsy diagnosis to be made lightly.

TRAUMATIC 'ASPHYXIA'

This condition is unfortunately misnamed, as the word 'traumatic' could equally be applied to hanging or strangulation. It is well recognized now, however, as meaning mechanical fixation of the chest, and is important both because of its frequency in accidents and because it provides the most extreme demonstration of the 'classic signs' of asphyxia.

Whereas other types of mechanical asphyxia may cause obstruction of air entry into the lungs, 'traumatic asphyxia' acts by restricting respiratory movements and thus prevents inspiration. It was termed 'traumatic' because gross mechanical forces are usually the reason for the fixation of the thoracic cage. Traumatic asphyxia occurs in two main conditions:

- The chest and usually the abdomen are compressed by an unyielding substance or object, so that chest expansion and diaphragmatic lowering are prevented. Burial in earth following the collapse of an excavation is a common cause and may kill workmen even if their heads remain above the fallen soil. Similarly, burial in grain, sand, coal or minerals, may have the same effect, and is usually encountered in industrial, marine or agricultural accidents. An avalanche in a silo, hopper, or other large-scale storage container may bury the worker up to neck level and, unless rescue is rapid, asphyxia will prove fatal. Similar restriction of chest movement may be caused by the victim being pinned under an overturned vehicle, or by falling timber or masonry. Many fatalities have occurred on farms, especially in hill country, by the toppling of a tractor, pinning the driver underneath. Protection in the form of a 'roll-bar' over the driving position or the provision of a rigid cab on a tractor was specifically designed to avoid such accidents.
- Crushing in crowds also leads to traumatic asphyxia, and this has caused some mass disasters, the largest probably being in Mecca. Most of the football-ground tragedies such as Bolton, Ibrox Park (1971), Lima (1964), Hillsborough (1989) and the Heisl Stadium in Belgium (1986) have been the result of crushing in crowds out of control. A similar mechanism led to 173 deaths in wartime London, when a panic on the stairs of Bethnal Green underground station, used as an air-raid shelter, caused the crowd to fall on those beneath.

Other forms of chest crushing arise from trapping between a vehicle and a wall, or between the buffers of two railway trucks.

Individual cases of traumatic asphyxia can occur when one person allows the whole weight of his body to fall upon another for a protracted period. This may happen in sexual intercourse, especially when one or both parties are incapacitated by drink or drugs.

Features of traumatic asphyxia

The appearances just mentioned are the hallmark of traumatic asphyxia and in no other condition, apart from postural dependency, is the degree of congestion and cyanosis so marked. When the chest is fixed, the face, neck and shoulders down to the thoracic inlet are grossly discoloured. Sometimes this colour is more red than purple. It can extend lower than the clavicles, Polson *et al.* (1985) being of the opinion that it often reaches down to the level of the third rib.

The conjunctivae are grossly congested and haemorrhagic. Rather than the petechiae seen in pressure on the neck, the conjunctivae and sclera may be so engorged with blood that the haemorrhagic tissue actually bulges out through the lids, completely obliterating the whites of the eyes. The face, lips and scalp may be swollen and congested, being dotted with petechiae and ecchymoses. There may be copious bleeding from the ears and nostrils. The whole picture is commonly an exaggeration of the appearances seen in slow death from manual strangulation, but local injury is absent and the signs extend down to or beyond the root of the neck. Where the compression has been caused by pinning under a solid object – as opposed to soil, sand or squeezing in a crowd – there may be local bruises and abrasions from the weight of a vehicle or heavy beam, but these are unrelated to the margin of the congestive–haemorrhagic zone.

Internally, the congestion is less marked than on the surface, but the lungs are usually dark and heavy and may well have subpleural petechial haermorrhages, the true ‘Tardieu spots’. The right heart and all the veins above the atria are markedly distended. There may be injuries to the chest wall from the trauma of the fixating object.

It is not clear why there should be such gross venous congestion, but it is usually explained rather unconvincingly by the failure of the pulmonary circulation consequent upon the cessation of the normal expansion and collapse of the pulmonary vasculature. Shapiro suggested that the pressure on the chest forces blood back into the great veins and, as the venous valves in the subclavian vessels prevent displacement into the arms, the extra volume is forced up the valveless jugular system to congest the head and neck.

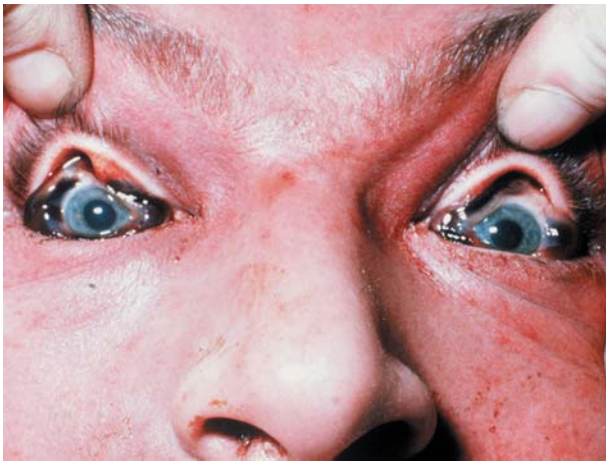

FIGURE 14.12 *Gross conjunctival haemorrhages in traumatic asphyxia. The victim was a workman who was buried up to his chest in an avalanche of ash.*

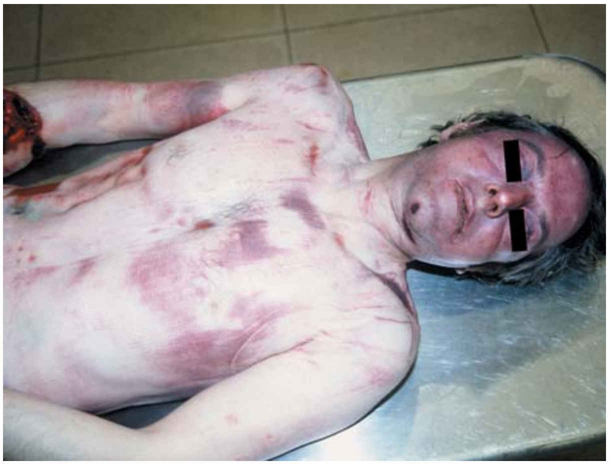

FIGURE 14.11 *Traumatic asphyxia showing gross congestion of confluent petechial haemorrhages of the face and parts of the shoulders. The victim had been buried up to the axillae in an avalanche of iron ore in a factory bunker.*

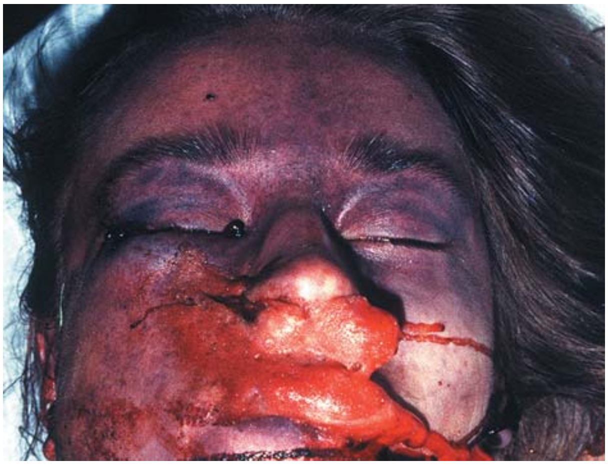

FIGURE 14.13 *Traumatic asphyxia showing gross congestion of the face and red blood-tinged froth exuding from the nostrils and mouth. The victim had lost the control of the tractor he was driving, fell out of the cab and was compressed under the overturned vehicle.*

POSTURAL ASPHYXIA

Closely allied to traumatic asphyxia is the so-called 'postural asphyxia', whose description as a separate entity is quite recent. When a person remains in a certain position for an extended time, either due to being trapped, or being in a drunken or drugged state, there may be a mechanical impediment to adequate respiratory movements. In addition, the normal venous return to the heart may be impaired.

Such positions usually entail inversion, either of the whole body or of the upper half; the syndrome and the pathophysiology are described well by Madea (1993), though most forensic pathologists will have had experience of such situations from time to time.

Persons who have been trapped upside down or even only in a 'jack-knife' position, with the upper half of their body bent acutely downwards from the waist, may have such impairment of their respiratory movements that they become hypoxic and suffer marked disturbance of their circulatory system, especially the venous return to the heart.

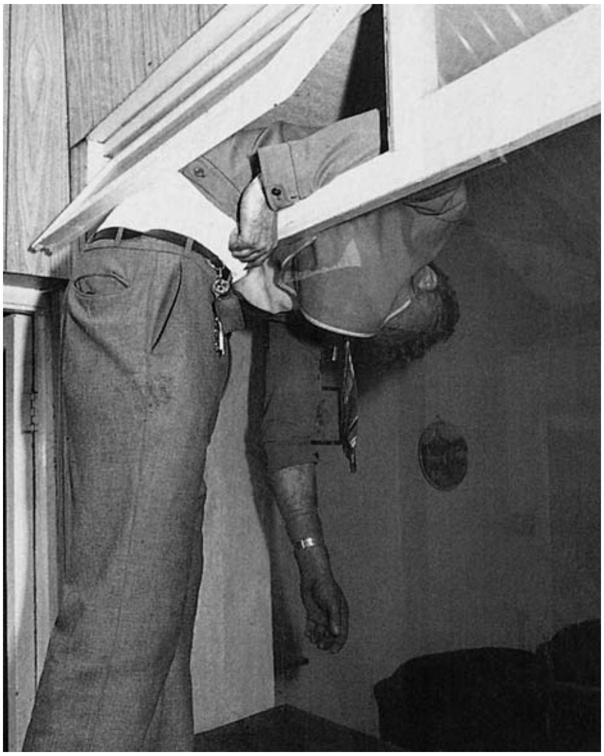

FIGURE 14.14 *Postural asphyxia in an inebriated man attempting to climb through a fanlight window. His feet slipped from the sill and he was unable to make sufficient respiratory movements against the weight of his body. Note the marked difference in congestion and cyanotic lividity between his two hands, the face and lower hand being discoloured at least partly from post-mortem hypostasis.*

The author (BK) has seen two cases in which a victim has become stuck whilst trying to climb through the upper part of a window, one as a thief, the other in trying to get into a house without his key. Others have been seen where, in a state of drunkenness or other disability, they have slipped out of bed, so that their head and shoulders are on the floor, with the legs and pelvis still at a higher level on the bed. These may also suffer the same disturbance of respiratory movements, which when prolonged may lead to death.

Inversion may occur during torture: crucifixion has an element of postural asphyxia. In the case of inverted crucifixion, as in the death of St Peter, it would be the major factor, as inspiration would be impeded by the weight of abdominal viscera upon the diaphragm.

REFERENCES AND FURTHER READING

Adelman HC. 1988. Asphyxial deaths as a result of aspiration of dental appliances: a report of three cases. *J Forensic Sci* **33**:389–95.

Adelson L. 1974. *The pathology of homicide: a vade mecum for pathologist, prosecutor and defense counsel.* W. B. Saunders, Philadelphia.

Ali TT. 1986. *Post-mortem autolytic changes in skin and the role of white blood cells*. Medical Faculty, University of Leeds, Leeds.

Bell MD, Rao VJ, Wetli CV, *et al.* 1992. Positional asphyxiation in adults. A series of 30 cases from the Dade and Broward County Florida medical examiner offices from 1982 to 1990. *Am J Forensic Med Pathol* **13**:101–7.

Betz P, Penning R, Keil W. 1994. The detection of petechial haemorrhages of the conjunctivae in dependency on the postmortem interval. *Forensic Sci Int* **64**:61–7.

Brinkmann B, Püschel K. 1981. [Histomorphological alterations of lung after strangulation. A comparative experimental study (author's transl).] *Z Rechtsmed* **86**:175–94.

Brinkmann B, Püschel K, Bause HW, *et al.* 1981. [Death by obstructive asphyxia: the pathophysiology of respiration and hemodynamics (author's transl).] *Z Rechtsmed* **87**:103–16.

Brinkmann B, Fechner G, Püschel K. 1984. Identification of mechanical asphyxiation in cases of attempted masking of the homicide. *Forensic Sci Int* **26**:235–45.

Bullock MJ, Diniz D. 2000. Suffocation using plastic bags: a retrospective study of suicides in ontario, canada. *J Forensic Sci* **45**:608–13.

Bywaters EG, Beal D. 1941. Traumatic asphyxia. *Br Med J* **1**:427–8.

Camps FE, Hunt AC. 1959. Pressure on the neck. *J Forensic Med* **6**:116–36.

Collins KA. 2001. Death by overlaying and wedging: a 15-year retrospective study. *Am J Forensic Med Pathol* **22**:155–9.

Ely SF, Hirsch CS. 2000. Asphyxial deaths and petechiae: a review. *J Forensic Sci* **45**:1274–7.

Flobecker P, Ottosson J, Johansson L, *et al.* 1993. Accidental deaths from asphyxia. A 10-year retrospective study from Sweden. *Am J Forensic Med Pathol* **14**:74–9.

Fred HL, Chandler FW. 1960. Traumatic asphyxia. *Am J Med* **29**:508–11.

Gardiner EE, Newberry RC, Keng JY. 1990. Avian vitreous humor concentrations of inosine, hypoxanthine, xanthine, uric acid, uracil and uridine as influenced by age and sex: their relevance as indicators of antemortem hypoxia. *Forensic Sci Int* **47**:123–7.

Gardner AN. 1958. Aspiration of food and vomit. *Q J Med* **27**:227–42.

Gordon I. 1976. Medicolegal aspects of rapid deaths initiated by hypoxia and anoxia. In: Wecht C (ed.), *Legal medicine annual 1975*. Appleton Century Crofts, New York.

Gordon I, Mansfield RA. 1955. Subpleural, subpericardial and subendocardial haemorrhages: a study of their incidence at autopsy and of the spontaneous development after death of subepicardial petechiae. *J Forensic Med* **2**:31–50.

Gordon I, Turner R. 1951. Deaths from rapid anoxia. *Am Med Assoc Arch Pathol* **2**:316.

Haddix TL, Harruff RC, Reay DT, *et al.* 1996. Asphyxial suicides using plastic bags. *Am J Forensic Med Pathol* **17**:308–11.

Haugen RK. 1963. The café coronary. *JAMA* **186**:142–4.

Hood I, Ryan D, Spitz WU. 1988. Resuscitation and petechiae. *Am J Forensic Med Pathol* **9**:35–7.

Hunt AC, Camps FE. 1962. Plastic bag deaths. *Br Med J* **1**:378–9.

Keil W, Kondo T, Beer GM. 1998. Haemorrhages in the posterior cricoarytenoid muscles – an unspecific autopsy finding. *Forensic Sci Int* **95**:225–30.

Knight B. 1976. The significance of gastric contents in the air passages. *Forensic Sci Int* **14**:398–402.

Luke J. 1971. The pathology, diagnosis and certain medical-legal aspects of death by homicidal smothering in adults. In: Wecht C (ed.), *Legal medicine annual 1971*. Appleton Century Crofts, New York.

Madea B. 1993. Death in a head-down position. *Forensic Sci Int* **61**:119–32.

McIntyre J. 1969. The classification of asphyxia. *Br J Hosp Med* **2**:1113–15.

Mole RH. 1948. Fibrinolysin and the fluidity of the blood postmortem. *J Pathol Bacteriol* **60**:413–17.

Morita M, Tabata N. 1988. Studies on asphyxia: on the changes of the alveolar walls of rats in the hypoxic state. II. The hypoxic state produced by carbon dioxide and methane gases. *Forensic Sci Int* **39**:257–62.

Ogden RD, Wooten RH. 2002. Asphyxial suicide with helium and a plastic bag. *Am J Forensic Med Pathol* **23**:234–7.

O'Halloran RL, Lewman LV. 1993. Restraint asphyxiation in excited delirium [see comments]. *Am J Forensic Med Pathol* **14**:289–95.

Plum F, Posner JB, Hain RF. 1962. Delayed neurological deterioration after anoxia. *Arch Intern Med* **110**:18–25.

Pollanen MS. 2001. Subtle fatal manual neck compression. *Med Sci Law* **41**:135–40.

Polson CJ, Gee DJ. 1972. Plastic bag suffocation. *Z Rechtsmed* **70**:184–90.

Polson JK, Gee D, Knight B. 1985. *Essentials of forensic medicine*, 4th edn. Pergamon Press, London.

Ponka JL, Lam CR. 1948. Effect of asphyxia on blood coagulation in dogs. *Proc Soc Exp Biol Med* **68**:334–6.

Prinsloo I, Gordon I. 1951. Postmortem dissection artefacts of the neck and their differentiation from ante-mortem bruises. *S Afr Med J* **25**:358–61.

Pullar P. 1984. Mechanical asphyxia. In: Mant AK (ed.), *Taylor's medical jurisprudence*, 13th edn. Churchill, London.

Purdue B. 1992. An unusual accidental death from reverse suspension [see comments]. *Am J Forensic Med Pathol* **13**:108–11.

Rao VJ, Wetli CV. 1988. The forensic significance of conjunctival petechiae. *Am J Forensic Med Pathol* **9**:32–4.

Reay DT, Howard JD, Fligner CL, *et al.* 1988. Effects of positional restraint on oxygen saturation and heart rate following exercise. *Am J Forensic Med Pathol* **9**:16–18.

Reay DT, Fligner CL, Stilwell AD, *et al.* 1992. Positional asphyxia during law enforcement transport [see comments]. *Am J Forensic Med Pathol* **13**:90–7.

Schleyer F. 1963. On adventitial haemorrhages of the thoracic blood vessels. *J Forensic Med* **10**:3–5.

Schmunk GA, Kaplan JA. 2002. Asphyxial deaths caused by automobile exhaust inhalation not attributable to carbon monoxide toxicity: Study of 2 cases. *Am J Forensic Med Pathol* **23**:123–6.

Shapiro H. 1975. Asphyxia. In: Gordon I, Shapiro H (eds), *Forensic medicine: a guide to principles*. Churchill Livingstone, London.

Shapiro H. 1988. Asphyxia. In: Gordon I, Shapiro H, Berson S (eds), *Forensic medicine: a guide to principles*, 3rd edn. Churchill, Edinburgh.

Shapiro HA. 1953. Is asphyxia a pathological entity recognizable at post-mortem? *J Forensic Med* **1**:65–8.

Shapiro HA. 1955. Tardieu spots in asphyxia. *J Forensic Med* **2**:1–4.

Simpson CK. 1943. Traumatic asphyxia. *Lancet* **2**:309–10.

Sturner WQ, Sullivan A, Suzuki K. 1983. Lactic acid concentrations in vitreous humor: their use in asphyxial deaths in children. *J Forensic Sci* **28**:222–30.

Swann HE. 1960. Occurrence of pulmonary oedema in sudden asphyxial deaths. *Am Med Assoc Arch Pathol* **69**:89–93.

Swann HE. 1964. The development of pulmonary oedema during the agonal period in sudden asphyxial deaths. *J Forensic Sci* **9**:360–4.

Swann HE, Brucer M. 1949. Cardiorespiratory and biochemical events during rapid anoxic death: fulminating anoxia. *Tex Rep Biol Med* **7**:511–16.

Swann HE, Brucer M. 1949. Cardiorespiratory and biochemical events during rapid anoxic death: obstructive anoxia. *Tex Rep Biol Med* **7**:593–9.

Tamaki K, Sato K, Katsumata Y. 1987. Enzyme-linked immunosorbent assay for determination of plasma thyroglobulin and its application to post-mortem diagnosis of mechanical asphyxia. *Forensic Sci Int* **33**:259–65.

Tardieu A. 1855. Memoire sur la mort par suffocation. *Ann Hyg Pub Med Leg (Series 11)* **6**:371–82.

Zaini MR, Knight B. 1982. Sub-pleural petechiae and pseudo petechiae. *J Forensic Sci Soc* **22**:141–5.

CHAPTER 15

Fatal pressure on the neck

- Mechanism of death in pressure on the neck 368
- Manual strangulation 371
- Strangulation by ligature 379
- Arm-locks and 'mugging' 383
- Hanging 383
- The sexual asphyxias: autoerotic or masochistic practices 389
- References and further reading 392

This provides one of the most complex and controversial areas of 'asphyxial' deaths, as the mechanism is uncertain and the frequency of such deaths makes them a common problem for both forensic pathologist and jurist.

Pressure on the neck may arise from manual strangulation, ligature strangulation, hanging, direct blows, arm-locks and a variety of accidental lesions, such as entanglement with cords or falling onto the neck.

Strangulation was formerly thought to be a pure asphyxia as a result of 'cutting off the air' by occlusion of the airway during the constriction of the neck. Since the end of the nineteenth century, however, it has been recognized that the rapidity of death in many cases made it impossible for hypoxia to be the sole or even major cause. Many victims died almost immediately and exhibited none of the so-called 'classic' signs of asphyxia.

Others, though showing these signs, still died too quickly for it to be reasonably argued that lack of oxygen had proceeded to a fatal stage. It was common knowledge that most people could hold their breath for over a minute and that some pearl divers could be physically active and yet not breathe for at least 3 minutes, so the much more rapid onset of death seen in many cases required further explanation.

MECHANISM OF DEATH IN PRESSURE ON THE NECK

A number of anatomical and physiological factors must be considered in analysing the effects of pressure on the neck.

Airway occlusion

This may occur either from direct compression of the larynx or trachea, or from a lifting of the larynx so that the pharynx is closed by the root of the tongue being pressed against the soft palate and roof of the mouth. The latter explanation is far more likely, as the strong cartilages of the larynx will resist all but the most extreme compression. Various experiments have been attempted to quantify the force needed to close the airway and blood vessels, but it is difficult to translate the often conflicting results to the degree of violence actually used in fatal events. For example, Brouardel (1897) calculated that the force needed to close the trachea was of the order 15 kg, far more than that required to occlude the blood vessels.

Occlusion of the neck veins

This factor is almost solely responsible for the appearance of the 'classic' signs of congestion, cyanosis, oedema and petechiae above the line of constriction. The external jugular system is most vulnerable, but any significant pressure encircling the neck will also obliterate the lumen of the internal jugular system, causing a rapid rise in venous pressure in the head, especially if the carotid arteries are still patent, which is usually the case. Brouardel's experiments suggest that a ligature with a tension of 2 kg blocks the jugular return; Reuter (1933) found an even lower figure.

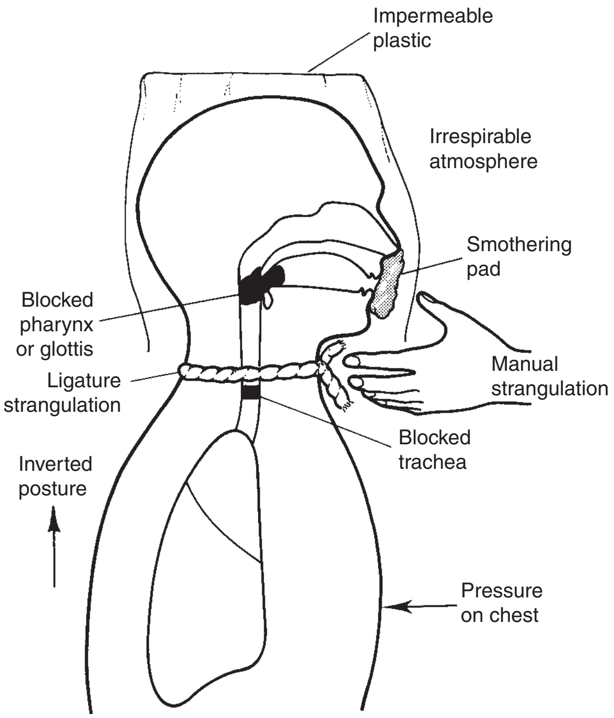

FIGURE 15.1 *Causes of mechanical asphyxia.*

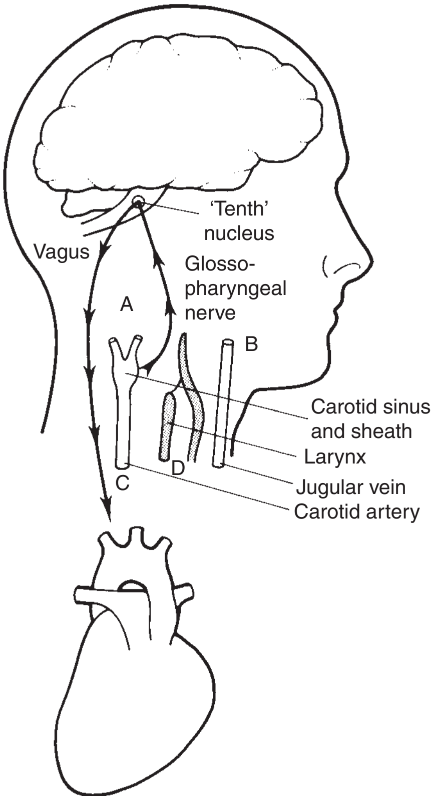

FIGURE 15.2 *Possible effects of pressure on the neck: (A) carotid sinus reflex leading to cardiac arrest; (B) jugular venous compression leading to cyanosis and petechiae; (C) carotid artery compression leading to unconsciousness; and (D) airway obstruction leading to hypoxia.*

Compression of the carotid arteries

This is much less common than venous occlusion, as the higher internal pressure of the carotid arteries resists occlusion and they are situated much more deeply. The carotids are largely obscured by the sternomastoid muscles. If bilateral occlusion of the carotids is achieved, almost immediate unconsciousness will ensue, as the supply of arterial blood to the brain by the vertebral circulation is insufficient to maintain cortical function, which depends mainly upon the anterior and middle cerebral arteries that arise from the carotid supply. Occlusion of the vertebral arteries by neck compression seems virtually impossible despite some claims to the contrary.

If the carotid circulation is totally occluded for an unremitting period of 4 or more minutes, then irreversible cerebral damage may occur. The time for this to happen is variable and the minimum has been disputed for many years, but the consensus of opinion is that permanent brain damage is very unlikely if the supply has been cut off continuously for less than 4–5 minutes. Total recovery has often been recorded after total ischaemia of considerably longer than this, even in normothermic conditions, with 9–14 minutes being quoted. Where the body has been subjected to low temperature, far longer times have been recorded, including an anecdotal case of a boy falling into the cold water of a Scandinavian harbour and being totally submerged for 40 minutes, with total recovery of cardiac and cerebral function.

Nerve effects

Pressure on the baroreceptors situated in the carotid sinuses, the carotid sheaths and the carotid body, can result in bradycardia (slowing of the heart), or in total cardiac arrest. This is an example of the physiological mechanism that monitors and adjusts blood pressure and heart rate.

This mechanism acts through a reflex arc in which the afferent (sensory) nerve impulses arise in the carotid complex of nerve endings, but not – as is sometimes alleged – in the vagus nerve trunk itself.

These impulses pass up to the brain via the glossopharyngeal nerves to the tenth nucleus in the brainstem, then return via the vagus (efferent) supply to the heart and other organs. This reflex arc acts through the parasympathetic side of the autonomic nervous system and is independent of the main motor and sensory nerve pathways. It is often claimed, admittedly without much concrete evidence, that fear, apprehension, struggling and possibly the effect of drugs such as alcohol, may heighten the sensitivity of this vagal mechanism. The release of catecholamines during such adrenal responses may well sensitize the myocardium to such neurogenic stimulation.

The vagal reflex has profound implications in relation to pressure or blows on the neck. Sometimes called 'vagal inhibition', 'vasovagal shock' or 'reflex cardiac arrest', the rapid onset of heart stoppage may antedate any evidence of congestive or 'asphyxial' signs, causing death immediately or within seconds, or at any time thereafter.

It is a matter of some dispute as to whether this reflex can cause immediate cardiac arrest or whether there has to be a period of marked slowing of the heart with negligible cardiac output – or whether an arrhythmia such as ventricular fibrillation precedes such an arrest. Probably any combination can occur, but it is an indisputable fact that collapse and apparent death can occur immediately on the application of pressure to the neck. Overstimulation of nerve endings in the carotid sinus or adjacent arterial sheath may be brought about by direct pressure from fingers, or from a ligature during strangulation or hanging – or from a blow directed at the side of the neck. Severe pain, such as a blow on the larynx or genitals, may also trigger a 'vagal response'.

THE FREQUENCY OF VAGAL CARDIAC ARREST

Though different authors vary in the proportion of such deaths that they attribute to reflex cardiac arrest, they all admit to the existence of such a mechanism. In the author's (BK) own series of fatal pressure on the neck from a variety of causes, the 'classical signs', denoting vascular and perhaps sometimes airway obstruction, were present in slightly less than half the cases. The remaining deaths presented with an absence of congestion, cyanosis and petechiae, the pale faces indicating that cardiac arrest had taken place before the congestive signs had time to appear.

One aspect which is uncertain – and virtually incapable of experimental proof in humans – is whether vagus-mediated cardiac arrest can be spontaneously reversible. Where an arrhythmia leads to arrest, resuscitation by cardiac massage or direct current electrostimulation has an excellent chance of revival – though in most forensic situations, such assistance is often lacking or too late.

However, it is not known whether cardiac arrest caused by the vasovagal reflex, can spontaneously revert to normal rhythm some minutes later – or whether it inevitably leads to death if no timely resuscitation is offered. This may have forensic relevance, as the author (BK) has been involved in several cases where pressure on the neck has lead to a comatose, but heart-beating victim, who later dies on artificial ventilation from irreversible brain damage. In the absence of medical data, the problem then arises as to whether the brain damage was caused by prolonged carotid artery occlusion from neck pressure – or was the result of a momentary neck pressure which caused reflex cardiac arrest, with spontaneous reversion to normal rhythm more than five minutes later, by which time irreversible cerebral damage had occurred.

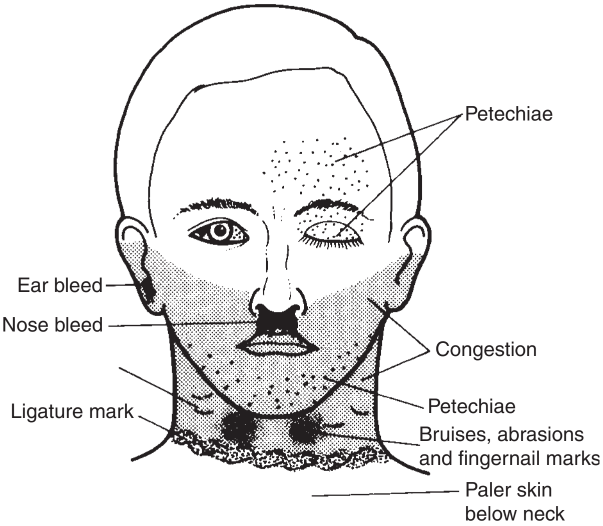

FIGURE 15.3 *Classical features of strangulation when cardiac arrest is delayed.*

Medico-legal aspects

This phenomenon has considerable legal as well as medical significance, as sudden death from 'vagal inhibition' can occur with total unexpectedness even with relatively slight pressure to the neck. Keith Simpson and Polson recorded cases many years ago in which a soldier at a dance playfully 'tweaked' his partner's neck and was mortified to see her drop lifeless to the floor. Many such cases are on record and every forensic pathologist of any experience has examples in his own records. Where it can be shown that the death occurred rapidly and without prolonged manual gripping of the neck, the defence may be raised that neither death nor serious injury was contemplated by the accused. It would be much harder to establish this lack of intent if the grip was maintained long enough to lead to florid congestion and petechiae in the face.

A combined mechanism of injury and the duration of neck compression

It is important to remember that the cardiac arrest mode of death may be mixed with the congestive–petechial mode in that, though the first stages of pressure may continue for long enough for congestive–asphyxial signs to appear (perhaps a minimum of 15–30 seconds), a change in grip may then allow the fingers to impinge on the carotid structures and lead to reflex cardiac arrest. Thus the progression of the pure 'asphyxial' process may be abruptly terminated at any point along its pathway to death by the superimposition of 'vagal inhibition', so that the intensity of the congestive changes may be of any degree in any given death.

The duration for which pressure must be maintained is often a contentious issue in criminal trials, as the 'inadvertent squeeze' which causes rapid vasovagal cardiac arrest is likely to be viewed as less culpable than a prolonged, unremitting gripping of the throat.

It is virtually impossible to measure the average minimum time of gripping that will produce congestion, cyanosis and petechiae from venous occlusion. As so often is the case in forensic medicine, animal experiments are useless for this purpose and obviously few cases of strangulation homicide are ever reliably witnessed, especially by some dispassionate observer with a stopwatch! Even in non-fatal cases – which incidentally, can produce far more florid examples of facial and eye haemorrhages than deaths – there is almost never any independent, accurate measurement of the time for which the neck was compressed.

It has been arbitrarily suggested that 15–30 seconds is probably the minimum period which will give rise to petechiae in the eyelids, conjunctivae and facial skin, but this really is 'picking a figure from the air', with little scientific justification.

Some years ago, the British Home Office and the Royal College of Pathologists set up a small working party to study this problem, composed of several experienced forensic pathologists and a respiratory physiologist – but the project was soon abandoned because of a total lack of reliable data.

Though it is easy to demonstrate that facial congestion develops within a few seconds of total jugular occlusion, petechiae – the only lasting sign after the venous return is re-opened – do not appear as quickly from neck pressure. However, they can occur from sudden, transient rises in venous pressure, such as sneezing or coughing – whooping-cough is a well-known example – and during the Valsalva experiment of forcibly trying to expire against a closed glottis. Eye petechiae and retinal haemorrhages can also occur during the chest compression of energetic cardiac massage during resuscitation.

CAUSES OF REFLEX CARDIAC ARREST

Vagal inhibition of the heart from stimulation of the carotid neural complex may occur in any form of pressure on the neck, but is much more common in manual strangulation than with a ligature, hanging excepted. Fingers seem more prone to dig deeply and find the structures under the anterior edge of the sternomastoid muscle. Perhaps the movements of the fingers, especially during the shifting postures of a struggle, more readily impinge upon the carotid bifurcation than the more static position of a ligature.

The majority of hangings, however, present with a pale face, free from congestive–haemorrhagic signs.

This seems to be caused by the more precipitate impact of the noose on the carotid structures when the victim's weight abruptly bears down, though Polson favoured actual carotid occlusion and hence cerebral ischaemia as the common cause of death in hanging. The scarcity of carotid intimal damage in suicidal hangings does not favour this view.

Blows to the throat

Another cause for sudden cardiac arrest is a blow to the neck or throat. This is the basis of the so-called 'commando punch' and some of the oriental martial arts also contain this in their repertoire – often forbidden because of its potential lethality. The edge of the hand is brought forcibly across the side of the neck or the front of the larynx.

Direct violence to the carotid region naturally causes gross stimulation of the afferent nerve endings. Blows directly to the larynx indirectly stimulate the sinus region or the laryngeal sensory nerve endings may themselves trigger the cardio-inhibitory reflex.

It is well known that the hypopharynx and larynx are particularly sensitive to stimulation, which accounts for the sudden deaths from impaction of food in the larynx, or from the flooding with cold water that causes some sudden immersion deaths. The testicles and uterine cervix also have a similar reputation for leading to sudden cardiac death, if unexpectedly overstimulated, especially when the myocardium is presensitized by catecholamines released by fear or emotion.

MANUAL STRANGULATION

A common method of homicide, manual strangulation is most often encountered when the physical size and strength of the assailant exceeds that of the victim. It is most commonly seen in domestic homicides when a husband kills his wife, in sex-related murders when the victim is again a woman, and in child killings, when the killer is an adult. Manual strangulation is rarely committed by a woman except on a child and a man rarely kills another man of equal physique.

Autopsy appearances in manual strangulation

The autopsy findings fall into two groups, namely, the local signs of violence, and the signs of the mechanism of death, where discernible.

- Bruising on the neck is the result of the assailant's attack, whereas abrasions may be from either victim or

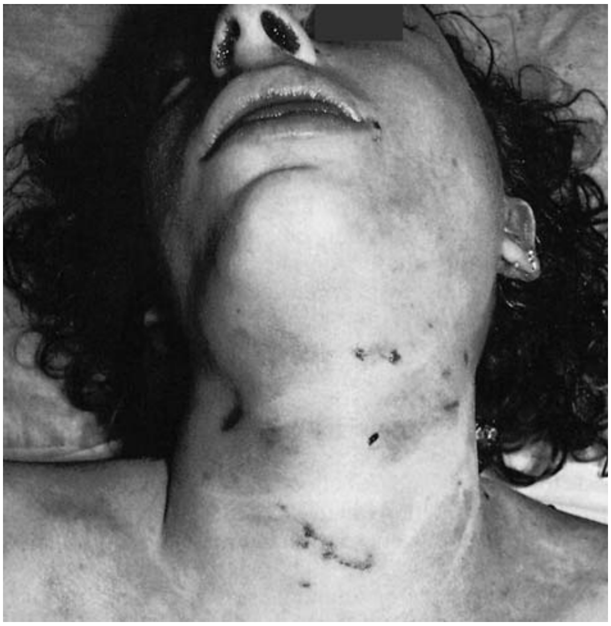

FIGURE 15.4 *Classical manual strangulation with fingernail marks and abrasions on the neck, congestion of the face, slight protrusion of the tongue and bleeding from the nose.*

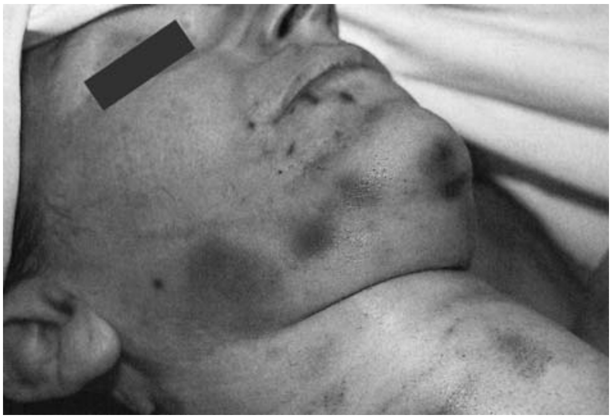

FIGURE 15.5 *Bruising on the neck and jaw margin in manual strangulation. Separate bruises along the jaw are due to high finger pressure – where a larger area of abrasion or bruising is present in this region, consideration must be given to pressure from an arm-lock.*

assailant. The bruises are mainly discoid, but may run together into larger areas of confluent bruising. The discoid marks are from finger-pads and are about 1–2 cm in size, though where the fingers skid across the skin surface, longer, irregular marks may occur, especially along the jaw margins. The bruises tend to cluster at the sides of the neck, often high up under the angles of the jaw. Some may overlap the jaw line and trespass on the chin, but others can be lower on the neck spreading down to the inner ends of the clavicles.

As a result of the shifting grip, often caused by the struggles of the victim, bruising can be anywhere, even at the posterolateral sides of the neck, and on the upper chest over the sternum and collar bones. Some may be seen in the grooves on either side of the larynx, over the anterior edge of the sternomastoid. They are also common over the prominence of the larynx and at the level of the cricoid.

As in most of the cases seen by a pathologist, death will have taken place at or about the time of the attack, most bruises will be fresh, being dark red or purple. Some may be so severe as to form actual superficial haematomas raised slightly above the skin surface. Bruises may often be seen in an asymmetrical pattern, with just one or two on one side of the neck, and a larger group on the opposite side. This may well suggest that a thumb was applied on one side and a group of fingers on the other; this may well be correct, but is not necessarily so and may be fortuitous if a shifting grip caused several reapplications of fingers.

It is sometimes tempting to use such asymmetrical grouping to decide ‘handedness’, that is to suggest that one bruise on the right side of the victim’s neck and a group on the left means that a right-handed assailant gripped from the front. The grouping may not be genuine, however, and because it can be fortuitous, it cannot be definitely determined that it was the right hand nor that it was not placed around the neck from the back. Overinterpretation must be avoided, as in all fields of forensic medicine and ‘Sherlock Holmesian’ opinions as to the handedness, size of hand, and orientation of attacker and victim in strangulation must be suitably restrained if both justice and the credibility of pathologist are to be preserved.

- Abrasions on the neck: scratches may be caused by the assailant or the victim, usually from fingernails. As with bruising, rough finger-pads (especially from a male hand on the delicate skin of a female neck) may abrade the epidermis and underlying bruises may be overlain by diffuse abrasions, again often seen along the margin of the jaw line.

  Linear scratches are the hallmark of fingernails and, when a woman strangles a child, the often longer nails of the woman may leave obvious marks. Whether male or female, the scratches are of two types: when the pressure is static, straight or curved marks up to a centimetre in length are made; when the nails skid down the skin, linear lines may result, sometimes several centimetres in length.

  A static nail mark is often semi-lunar in shape, but caution must be used in interpreting the posture of the hand from this shape. Though it would seem obvious that the concavity of the mark should face the finger-pad, experiments by Shapiro *et al.* in 1962 showed that the converse may be true. This is caused by fixation of the skin at the centre where the nail digs in, with escape at the margins, so that when the traction relaxes, the mark appears inverted, with the convexity facing the finger-pad. However, experiments by others, including the author (BK), show that this phenomenon is by no means invariable and often the marks are the result of the nails being applied in the expected direction.

Scratches are the result of the victim's attempts at pulling away the strangling hand. As most victims are women, the nails may be long and the scratches more severe than those from the assailant, which are often absent altogether. These defence scratches may run in parallel lines from grouped fingers and run in a vertical direction in the long axis of the neck, though they are often random.

In the autopsy on a strangled victim it is good practice to take fingernail scrapings or clippings for full forensic science investigation. These should be preserved and labelled so that each clipping or scraping is individually identified – or failing that, at least the left- and right-hand material should be kept separate. Skin fragments or blood under these nails may provide blood typing or DNA characteristics that can be matched to an alleged assailant – and other valuable trace evidence, such as hairs and fibres might be trapped in a torn fingernail – though admittedly, the number of occasions when such evidence has been gained is extremely small.

The alleged assailant should always be examined medically to correlate any injuries that may have been inflicted on him by the fingernails of the victim, such as facial or hand scratches. If he is examined soon after the offence, his own fingernail scrapings should be taken, especially if the victim has neck abrasions, so that comparison studies can be made between any debris found and the tissue types of the victim.

Internal appearances in manual strangulation

Internally, the bruises described above may be visible to a greater or lesser extent in the tissues of the neck. Often they are quite superficial and are confined to the dermis, but some may be visible to an appreciable depth in the muscles and other structures in the neck.

The platysma muscle may be bruised, but the sternomastoid and the deeper strap muscles that run vertically along the larynx may show patchy bleeding, not always corresponding exactly to the position of surface bruises. They may contain infiltrated blood or even frank haematomata, especially the sternomastoid.

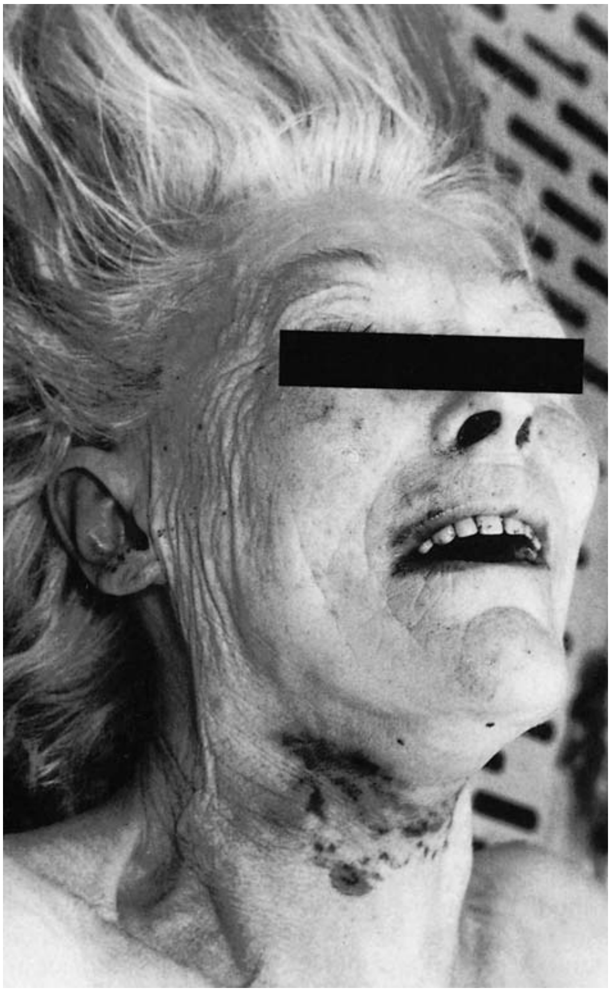

FIGURE 15.6 *Manual strangulation with bruising from the assailant and fingernail abrasions from the victim. The face is pale as a result of rapid vasovagal cardiac arrest before congestive signs could develop.*

At autopsy, it is essential to release the blood in the venous system before dissecting the neck to avoid or reduce artefactual haemorrhages that can occur in the region. It may be easiest to reflect the scalp and remove the brain before dealing with the neck, to release venous engorgement. Alternatively, after the skin of the neck has been carefully flayed off in a wide 'V' incision, the internal jugular vessels can be incised before any further manipulation to release the venous pressure, but removal of the skull is preferable.

Radiography of the neck may be carried out before any dissection to determine the state of the cervical spine and the laryngeal cartilages, though the author (BK) has never found this to be particularly helpful, preferring to X-ray the isolated larynx before dissection.

When examination of the deep neck structures begins, careful removal of the overlying tissues layer by layer is required, seeking genuine haemorrhage as each set of muscles is exposed. There may be bruising into the thyroid capsule and into any of the strap muscles. The vascular bundles must be handled carefully and the large veins opened with scissors. The carotids can be opened *in situ*, or after the laryngo-oesophageal pluck has been removed.

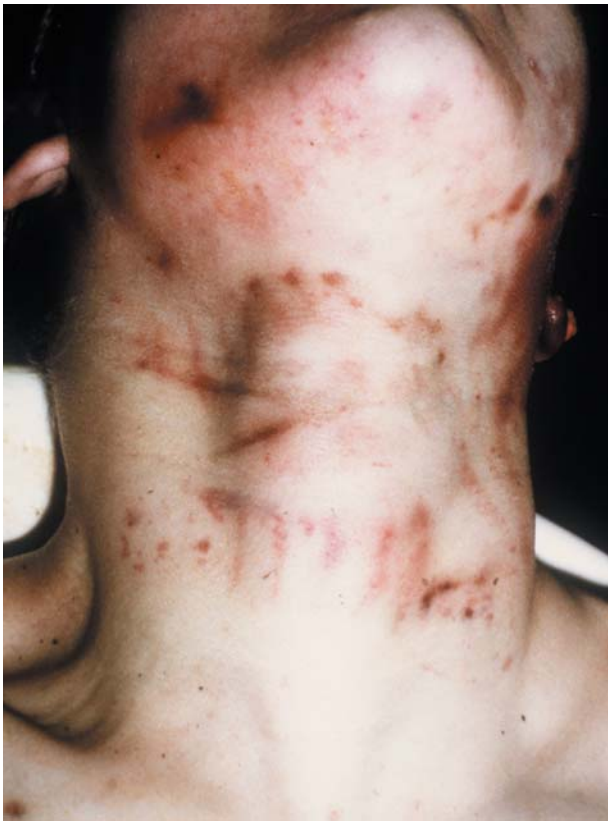

FIGURE 15.7 *Manual strangulation, showing scattered bruises from the fingers of the assailant, but also a pattern of abrasions under the chin and below the larynx, caused by the nails of the victim whilst attempting to pull off the strangling fingers.*

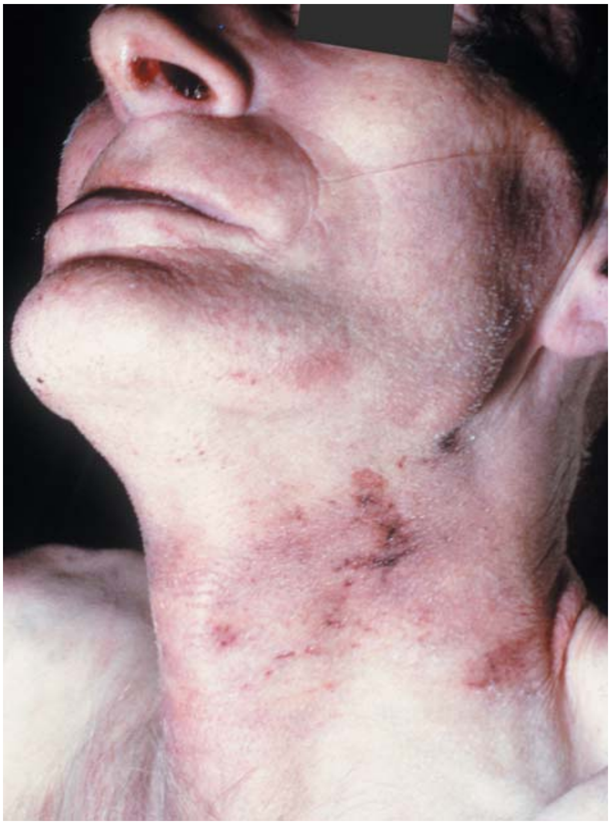

FIGURE 15.9 *Abrasion and bruising in manual strangulation. The face is pale, indicating a rapid sudden death; much of the skin damaged is directly over the position of the underlying carotid sinus.*

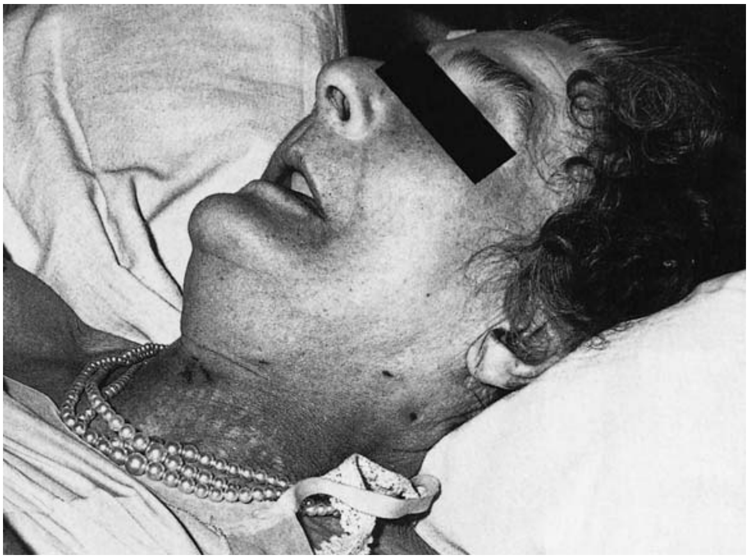

FIGURE 15.8 *Illustration of the need for all doctors to be observant and vigilant. This lady was seen in bed by a family doctor and pronounced a 'heart attack' victim. Later examination revealed facial congestion, fingernail marks on the neck and the impression of her necklace on the skin by strangling hands.*

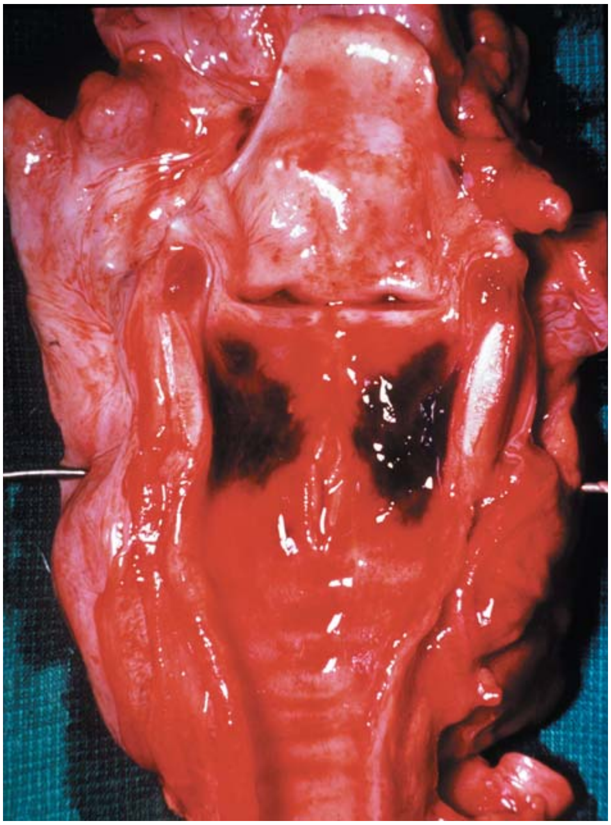

FIGURE 15.10 *Mucosal haemorrhages in the interior of the larynx immediately below the vocal cords, in manual strangulation. This is a common finding and the mechanism is uncertain. It can occur as here – in the absence of generalized congestive–petechial changes – and appears to be caused by local trauma.*

Spurious bleeding behind the larynx

Bleeding over the front and sides of the larynx is usually genuine, if the venous pressure has been released beforehand, but great caution must be employed in interpreting bleeding behind the larynx and pharynx.

In any autopsy, substantial haemorrhage may be seen over the posterior surface of the oesophagus and on the anterior longitudinal ligament of the cervical spine. This is a common post-mortem artefact, described fully by Prinsloo and Gordon (1951) and sometimes known by their name. In the absence of any other neck lesions, especially bleeding in the lateral and anterior parts of the neck, no reliance can be placed on this haemorrhage. It tends to develop more as the post-mortem interval lengthens and even releasing the venous pressure in the neck by removal of the brain, or opening the jugulars early, does not ensure its absence. To avoid the Prinsloo and Gordon artefact, a special technique of dissecting the neck is recommended by Shapiro (1988). Another artefact in the neck, much easier to recognize as false, is ‘banding’ of the oesophagus, especially when the tissues are congested. These bands are pale areas in the mucosa caused by post-mortem hypostasis being prevented from settling by the external pressure of adjacent anatomical structures, including parts of the larynx, trachea and aortic arch. Banding is common in routine non-trauma autopsies, but has been misinterpreted by inexperienced pathologists as evidence of strangulation.

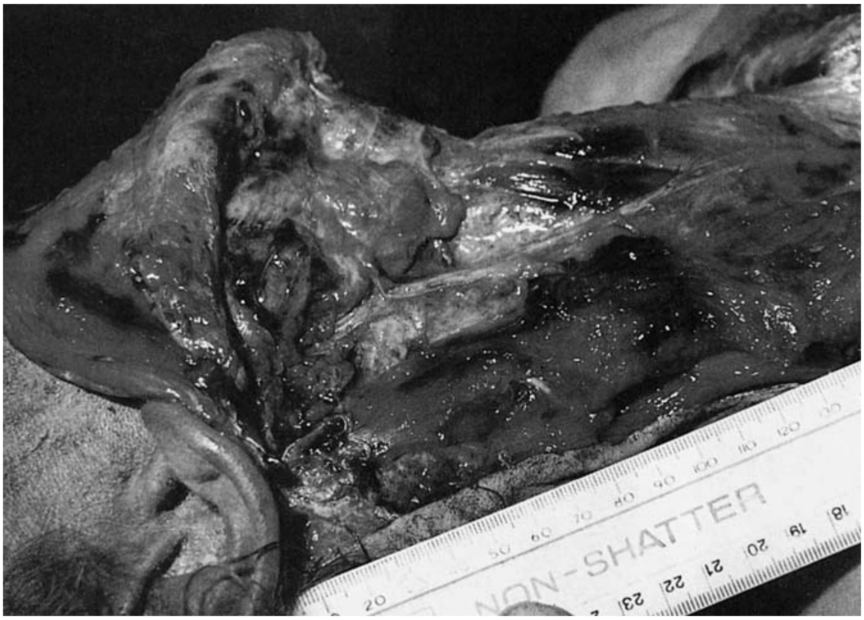

FIGURE 15.11 *Dissection of the neck in a victim of manual strangulation. There are extensive areas of bleeding in the strap muscles, the lower part of the larynx and on the submandibular gland and jaw margin. Removal of the brain before dissection of the neck reduces the risk of artefactual haemorrhages in this region.*

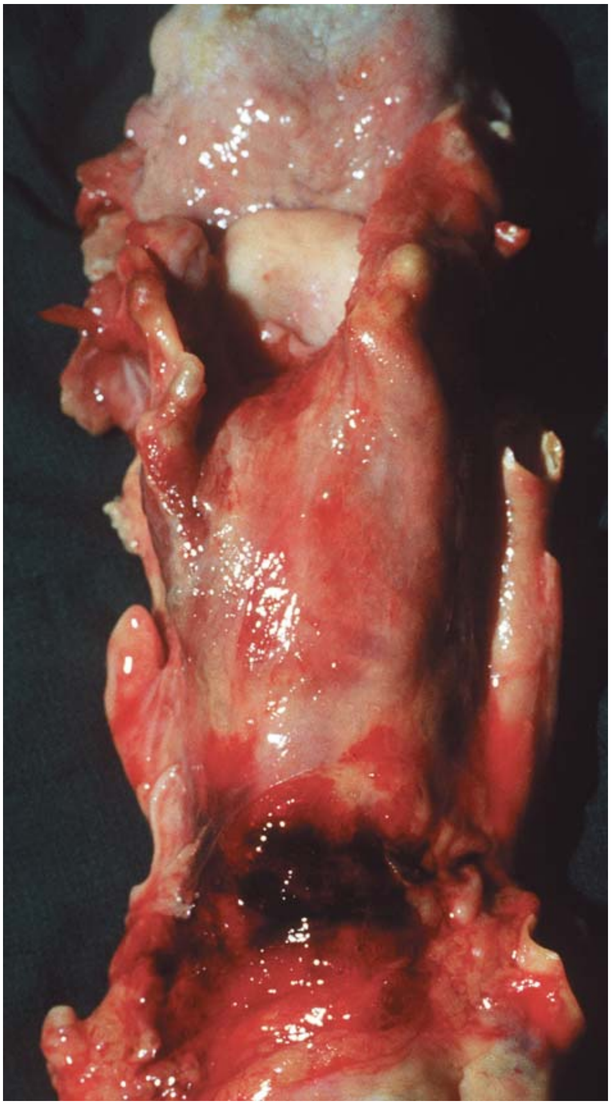

FIGURE 15.12 *Artefactual post-mortem haemorrhage on the posterior surface of the oesophagus in a natural death. This was fully described by Prinsloo and Gordon, and is a reason for draining the blood from the venous system of the head and neck before beginning to dissect.*

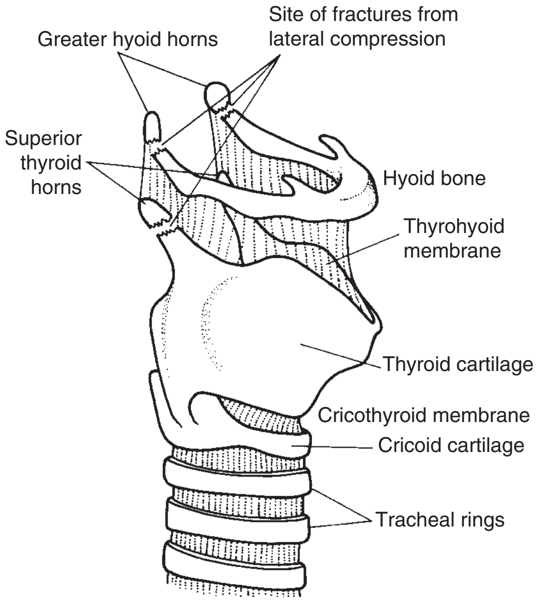

FIGURE 15.13 *The forensic anatomy of the larynx. The hyoid and thyroid bones fracture either from direct lateral compression, or from traction from the thyrohyoid membrane when it is compressed.*

Forensic anatomy of the larynx

The larynx consists of a large ‘V’-shaped thyroid cartilage, which has a prow-like prominence anteriorly and is open at the back. Below this is the smaller cricoid cartilage, which is narrow at the front, but expands posteriorly to occupy the lower part of the space left by the open thyroid cartilage.

At the upper margin of the posterior wing of the thyroid on each side, are the superior horns or ‘cornuae’, which are connected by the thyrohyoid membrane to the greater horns of the hyoid bone which lies immediately above.

The hyoid bone lies at the root of the tongue, and consists of a central horizontal ‘body’, to which are attached two long ‘greater horns’, which sweep backwards and slightly upwards above the upper margin of the thyroid cartilage and superior horns. The hyoid often has natural joints, sometimes even with synovium, which lie at the junction of the body with the greater horns.

There are two lesser horns on the upper surface of the body that have no forensic anatomical significance. The hyoid calcifies at variable times: the body is usually calcified, but the horns may calcify irregularly, both in space and time. In teenagers and young adults they are usually cartilaginous and the joints mobile. In middle and later life, the hyoid and thyroid horns calcify and become more brittle. The cricoid cartilage is a modified upper tracheal ring but can also become partly calcified as age increases; no meaningful ages can be placed on any calcification, but traumatic fractures can occur at any time except in children and most teenagers, in whom fractures are rare.

Injury to the larynx

During manual strangulation the larynx may become damaged in various ways. The pressure is mainly bilateral, so that the sides of the larynx are squeezed. Particularly vulnerable structures are the four ‘cornuae’ or horns, which protect backwards to maintain the patency of the airway around the glottis. Lateral pressure of the fingers can displace any of the

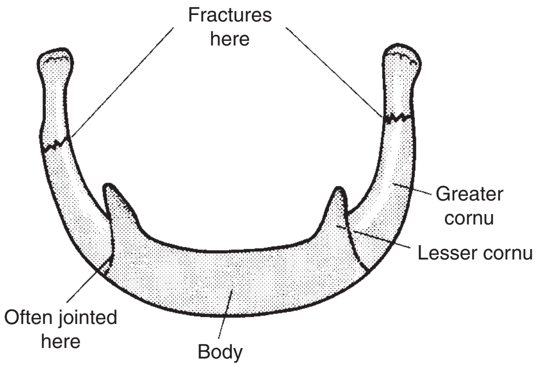

FIGURE 15.14 *The hyoid bone.*

four horns inwards, either by direct pressure or by pressure on the thyrohyoid membrane, which then drags the horns medially. In young persons, the horns are so pliable that they return to their normal position on release of the pressure but, variably beyond the third decade, they may be sufficiently calcified to fracture.

The appearance of a fractured horn is of a loose or even flapping termination. The horn usually displaces medially, but is held by the thyrohyoid ligament. At autopsy it is advisable to cut down this ligament carefully between thyroid cartilage and hyoid to destroy the 'splinting' effect that may support a broken horn. As will be discussed later, however, there will be a haemorrhage at the fracture site that will make the lesion obvious. In the case of the hyoid, care must be taken not to misinterpret a natural joint as a fracture. These are more medial than the usual site of fracture, which is likely to be within a centimetre of the tip. The jagged edge of the fracture line may be exposed, especially in an older person with a brittle horn. Radiography before dissection is an excellent way of confirming a fracture, identifying natural joints and detecting any other fractures in the thyroid or cricoid cartilage.

Though the hyoid bone has received most attention in publications as being the marker of violence to the larynx, in fact the thyroid horns are far more vulnerable. Simpson (1985) found that, in 25 successive deaths from manual strangulation, there were 22 fractures of thyroid horns but only one fractured hyoid. Though this ratio is not typical of the experience of most pathologists, there is no doubt that the superior horn of the thyroid is much more fragile and more vulnerable than the greater horn of the hyoid bone. Although fractures of the horns are more common with advancing age, they can on rare occasions be found even in teenagers. Care must be taken, however, not to confuse mobility at natural joints with a fracture (Evans and Knight 1982).

Conversely, undoubted severe violence to the neck in older persons does not necessarily cause horn fractures. The frequent persistence of joints at the base of the greater horns of the hyoid may allow movement and so avoid fractures when compressed. The cricoid and the main ala of the thyroid cartilage may be cracked, but this is an index of much greater pressure from strangling fingers. The plates may break in a spiral or oblique fashion and the cricoid may crack anteriorly through the narrow bridge rather than at the wider posterior plate. Fractures of these larger areas are also found in direct blunt violence such as punching, kicking or arm-locks, often as part of general severe damage and disruption of the entire larynx.

Significance of laryngeal fractures

Laryngeal cornual fractures are, of course, merely indicators of pressure applied to the neck and are not themselves relevant in terms of a threat to life. This is sometimes misunderstood by laymen, especially police and lawyers, who have been misled by the importance that pathologists attach to the finding of a fractured horn into thinking that the injury itself is a significant contribution to the death.

Such fractures, if proved to be genuine, are certainly significant in proving the application of violence to the neck, but certain precautions must be observed in their interpretation. The sole finding of a fractured horn, where there is no other injury to the skin, subcutaneous tissues, muscles or rest of the larynx is of little value unless there is firm circumstantial evidence of violence to the throat.

When a fracture of a laryngeal horn is found, it must first be proved to be ante-mortem in origin. Post-mortem fractures undoubtedly occur, either from mishandling the body during transit or from incorrect autopsy techniques – though the frequency of the latter has been overestimated.

It is certainly possible to damage the larynx post-mortem by allowing the neck to fall against a hard surface or sharp edge during removal from the place of death, or during handling in the mortuary. Such damage, however, is more likely to occur to the laryngeal plate of the thyroid cartilage or to the cricoid, rather than to the laterally placed horns, though these can be broken. Damage at autopsy is usually caused by an inexperienced pathologist or autopsy-room technician, especially when forensic expertise is lacking. Clumsy removal of the tongue and neck structures can break the thyroid or hyoid cornuae, especially in old persons where they are calcified and brittle and when any natural joints are ankylosed.

This may be one justification for radiography before autopsy, but probably the danger of artefactual breakage has been overestimated, especially where a gentle removal technique is employed.

Proof of ante-mortem laryngeal fracture

This must be obtained by demonstrating haemorrhage at the fracture site. It is usually obvious to the naked eye and can be confirmed histologically. Sometimes the bleeding is prominent, with a blood bleb under the periosteum or perichondrium at the fracture line, and the broken end of a calcified horn may be palpable and even slightly crepitant. Where no macroscopic evidence of bleeding is found, careful histology of the fracture site will almost always reveal some red-cell extravasation, but this evidence is of less value because microscopic – and even slight macroscopic – oozing can occur from a post-mortem fracture of a horn. Bleeding is therefore a 'one-way' criterion: if there is no haemorrhage, the fracture must be post-mortem, but if there is a small bleed, then the lesion can be either ante-mortem or post-mortem. Naturally the findings must be taken in conjunction with other evidence of neck injury and, if there are surface bruises, scratches, muscle bleeding, intralaryngeal and tongue haemorrhage, then a minimal bleed at a fractured horn can be accepted as probably ante-mortem.

Conversely, it must again be emphasized that the solitary finding of a fractured laryngeal horn, even with slight associated bleeding, is not in itself sufficient evidence of ante-mortem trauma to the neck.

Other causes of fractured laryngeal horns

The hyoid and thyroid horns can be broken other than by manual strangulation. Ligature strangulation and hanging can certainly cause these lesions, though not as often as manual throttling. Direct violence can also break the cornuae, but usually only as part of more diffuse laryngeal damage. A punch or kick in the throat can disrupt the larynx and break the horns as part of general damage. Even wide, blunt pressure can fracture the horns, however, such as an arm-lock from behind squeezing the larynx against the cervical spine. It has been said that such direct violence from the front tends to splay the fractured horns laterally, whereas in manual strangulation the horn tends to fall inwards – but this is too tenuous a claim to be of much practical value.

In accidents, the larynx can be damaged. In falls, the larynx is usually protected by the prominences of the chin and chest, but if the fall is onto a ridged object, such as gate or chair-back, then a focal impact can injure the throat. The author (BK) has seen a death in a housebreaker, who slipped when trying to climb through a sash window. He fell with his throat across the top of the window-frame and died – presumably of a vagal cardiac arrest – being found suspended by his neck next morning with fractures of the laryngeal cartilages.

Other laryngeal injuries in strangulation

Other injuries to the larynx include splitting of the ala of the thyroid cartilage, often either in the midline or obliquely across the left or right plate. This may be part of a violent strangulation, seen by Green (1973) in 12 per cent of his series – but is more often seen in blows to the front of the neck, either by fist or the edge of the hand, or a kick or any other kind of substantial trauma directed at the front of the neck.

The cricoid cartilage, a circle shaped like a signet ring that lies below and partly within the thyroid cartilage, may also be fractured by manual throttling when pressure is applied rather lower than usual. Again, this cartilage tends to break either across the front midline or obliquely. Though an uncommon injury, a severe displacement of a broken cricoid – especially if accompanied by mucosal tearing and bleeding – may project backwards into the airway and cause a partial obstruction.

The interior of the larynx may also show indications of pressure on the neck. This usually takes the form of haemorrhage into the mucosal lining and is most often seen immediately below the vocal cords. Bleeding may be intense here, but its cause is obscure. It can occur with no other damage to the cartilages of the larynx and it is then hard to equate it with direct physical pressure as it lies on the protected inner surface.

Similar, but less intense, haemorrhage is sometimes seen on the surface of the epiglottis, which more often shows a shower of petechial haemorrhages if the death was not sudden.

Such haemorrhage inside the larynx is often a bright, fiery red, unlike the dark cyanotic discoloration that is seen in the pharynx, root of tongue and larynx in congestive–asphyxial deaths, and is caused by obstructed venous return rather than direct trauma. At autopsy the base of the tongue should always be sliced, as deep haemorrhages may be present, usually at the sides of the tongue, deep congestion being more central or across the whole posterior part of the tongue.

Damage to the carotid arteries

Though the sudden 'vagal inhibition' type of death caused by pressure on the baroreceptors of the carotid sinus and sheath is quite common, it is unusual to find anatomical confirmation of injury to these structures.

There may be deep haemorrhage in the neck tissues surrounding or adjacent to the bifurcation of the common carotid artery, but this is often absent in these cases of cardiac arrest – and conversely, such bleeding may be found in the slow 'asphyxial' deaths, as part of the generalized diffuse bleeding into the tissues.

Rarely, careful dissection of the carotid arteries may reveal an intimal tear or bleeding into the walls of the vessels at or near the carotid sinus. A delicate technique is required and artefactual damage must be carefully excluded, especially that from the point of scissors used to slit open the artery. Such tears are more often found when a forcible impact or a more focal pressure has been applied, such as from a thin ligature that can cut deeply and abruptly into the neck, especially in hanging. Generally, little or no morphological confirmation is to be expected in the cardiac arrest type of death, following pressure on the neck. The diagnosis is one of exclusion, based on the absence of any congestive–cyanotic–petechial appearances, together with some circumstantial history of sudden death whilst pressure is being applied to the neck.

As already mentioned, one problem is that pressure may be maintained for long enough – perhaps a minimum of 15–30 seconds – to produce these signs, yet not cause death, but death then supervenes abruptly because a vagus-mediated cardiac arrest is superimposed on the early asphyxial mode. In the absence of any structural damage to the carotid apparatus, autopsy cannot differentiate between the two mechanisms and only the history (which is rarely available for obvious reasons) can assist in interpreting the sequence of events.

STRANGULATION BY LIGATURE

Pressure on the neck may be effected by constricting all or part of the circumference of the neck by a ligature. This is sometimes called 'garrotting', though strictly this refers to the tightening of a noose around the neck by twisting a rod within the ligature, a form of judicial execution once employed in Spain, from which also comes the description of the twisting device as a 'Spanish windlass'. This method had a refinement in which the back of the neck was forced against a sharp spike which penetrated the spinal cord.

In forensic practice, if hanging is excepted as a separate entity, most ligature strangulations are homicidal. Some are suicidal and a few accidental, usually in children.

The nature of the ligature

The ligature may consist of a wide variety of objects, some not obviously suitable for the purpose, yet effective in causing death. Cords, wires, ropes and some belts are strong and relatively thin, so that they tend to cut deeply into the neck if the tension is great. Softer fabrics, however, in the form of scarves, ties, towels, stockings, tights and even strips of bed-linen may be used and may cause some problems in interpretation if they are removed from the scene before the investigation begins. The ligature may be applied as one turn around the neck – or even less, as homicides have been perpetrated by the assailant pulling a U-shaped ligature against the front and sides of the neck, while standing at the back. Even a flexible rod, such as bamboo cane or – in India – a flat 'lathi', can act as a ligature if the two ends are forcibly pulled back from behind against the front of the throat.

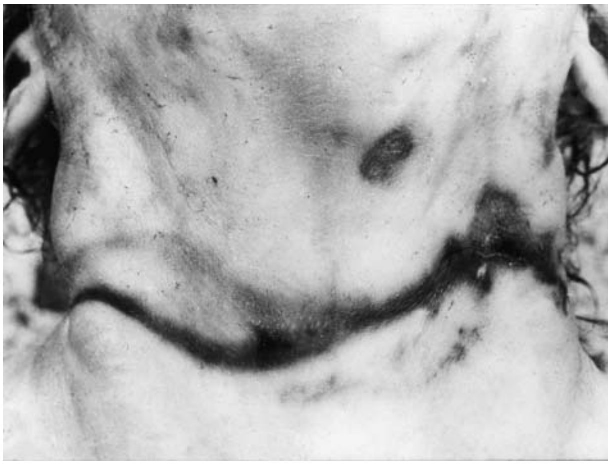

FIGURE 15.15 *Ligature mark from a nylon scarf. Although the fabric was broad, the tension bands from the tightly stretched material have produced a definite line that could be mistaken for a cord or wire.*

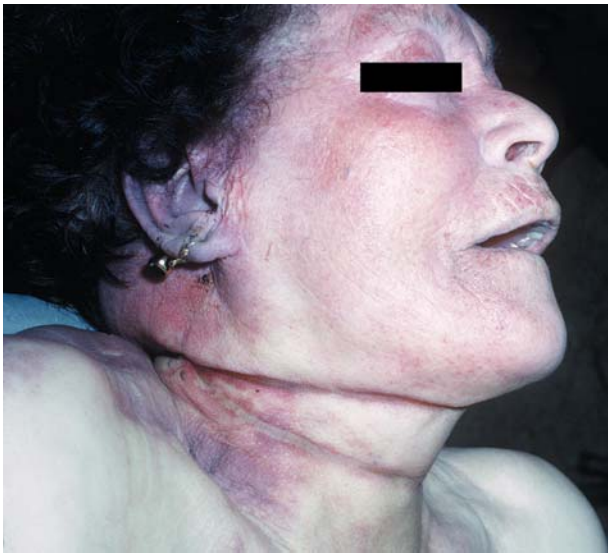

FIGURE 15.16 *Ligature strangulation with a generally horizontal course for the two strands of cord that diverge from the skin at the back of the neck.*

In the majority of cases, however, the ligature is crossed over itself after passing a full circle around the neck – and several turns may be wound around, secured with one or more knots. These multiple turns are not uncommon in

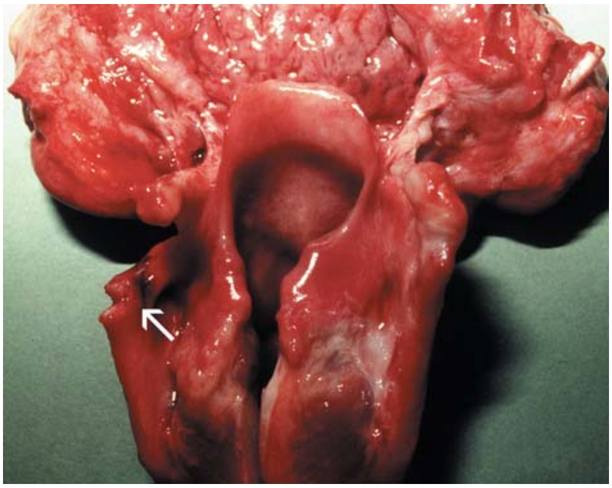

FIGURE 15.17 *Fracture of the left laryngeal horn (arrow) with haemorrhage. The body was recovered from a lake based on the confession of the perpetrator. Further findings were haemorrhage in the fractured cricoid cartilage, bruising on the neck and petechiae in the conjunctivae and larynx.*

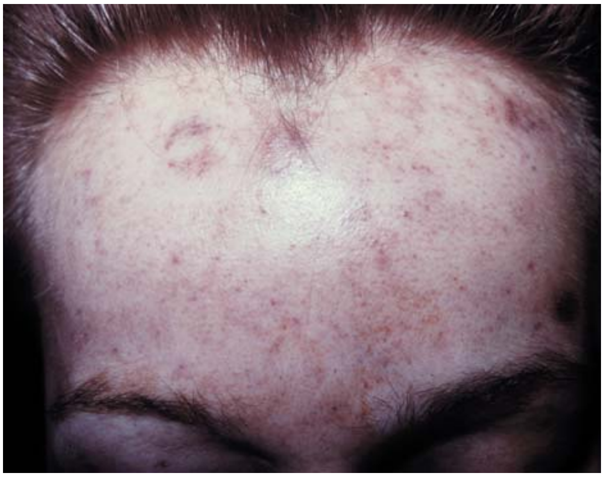

FIGURE 15.18 *Typical appearance of congested, cyanosed forehead, with multiple petechiae in the skin, in a woman strangled manually and by ligature. Circular bruising probably due to knuckle blows. The scarf tied around the neck initially suggested only ligature strangulation.*

suicide, where two, three, or even more circles are wound around, often with complex knots. The presence of these more complicated ligatures by no means confirms homicide – in fact they are more common in suicide, where the determined victim is eager to succeed. The ligature may be tightened even further by means of the 'Spanish windlass' mentioned above, which consists of some rod-like object, which may be a ruler, stick or screwdriver inserted between the ligature and the skin, and twisted around several times. The rod may be held in position by being wedged under the chin or against the shoulder or chest. This device may be seen in either homicides or suicides.

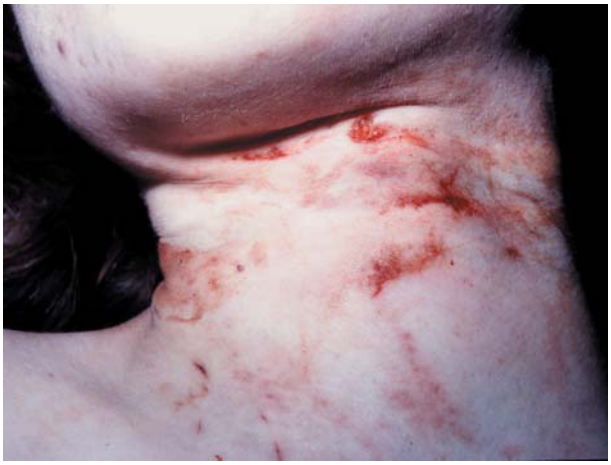

FIGURE 15.19 *In the case shown in Figure 15.18. When the scarf was removed, typical bruises and abrasions of manual strangulation were seen in addition to the ligature mark.*

The ligature mark

The appearance at autopsy naturally depends on the nature and texture of the ligature. When there is a pronounced pattern, such as the weave of a cord or the plaiting of a thong, the same pattern may be imprinted into the skin. In homicide, where the ligature has been removed by the killer, such a pattern may be of great value in tracing its origin. When a fabric has been used, such as a scarf or towel, the marks on the neck are more difficult to interpret. A broad, flat band may leave no mark whatsoever, but it usually leaves one or more linear marks on the skin of the neck, often discontinuous. These may fade if the body is not seen soon after death and if the surface has not been abraded.

A fabric ligature may leave a sharply defined mark, which may be misinterpreted as being caused by a narrow cord or wire. The reason for this is that when a broad piece of cloth is tightly stretched, one or more bands appear that are under greater tension than the rest. It is these that mark the skin, and no sign may be left that a wider area was lying on either side of the most tensed band. These marks are usually less well demarcated at the edges than a cord or rope, but can cause confusion.

When the ligature is still in position when the body is examined, it may appear to be deeply embedded in the skin, sometimes almost out of sight, and on removal a deep groove may be seen in the skin. This embedding may be accentuated by oedema of the tissues, especially above the ligature, which initially may not have been applied so tightly. The swelling can continue to develop to some

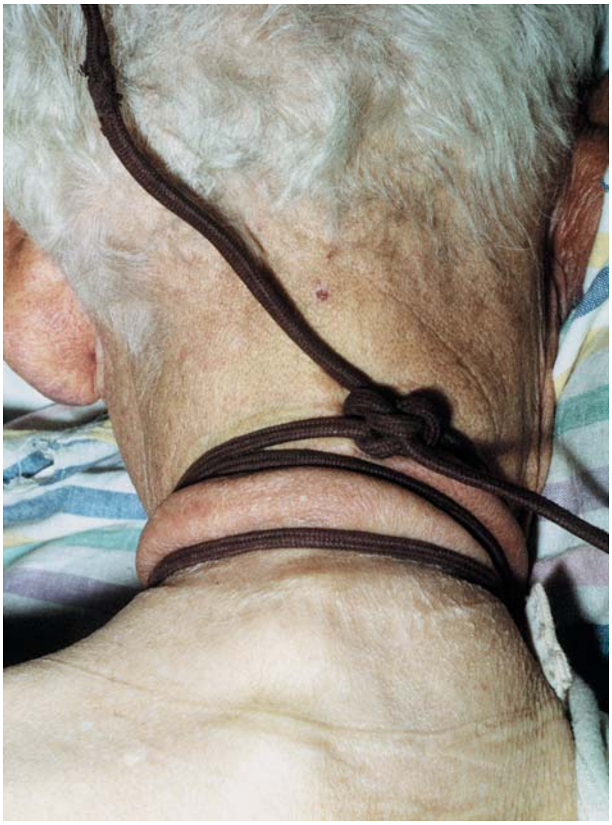

FIGURE 15.20 *Self-strangulation by ligature, with three turns around the neck and a complex knot. There was no doubt from the circumstances that the deceased man had deliberately killed himself with this telephone cord, but the mechanism of apparent vasovagal death before any congestive changes could appear remains obscure.*

extent even after death, accentuating the depth of the groove. Presumably some passive transudation of tissue fluid continues even after the circulation has stopped. In suicides, multiple turns and knots may have been applied and obviously, the degree of tightness as seen later at autopsy could not have been present during the application of the ligature, otherwise incapacity would have been so rapid that the process could not have been completed. In such cases, it is remarkable how often the mode of death still seems to be of the non-asphyxial cardiac arrest type. It might be expected that vagal stimulation could not have occurred once the expected static position was attained with the completion of ligation – but experience indicates otherwise, the mechanism remaining obscure.

The skin mark may remain red, especially if the ligature was of softer material such as cloth, but cords, ropes and wires tend to abrade the surface, which later becomes yellow or brown and parchment-like. This is seen particularly in hanging, when the friction and chafing may be greater.

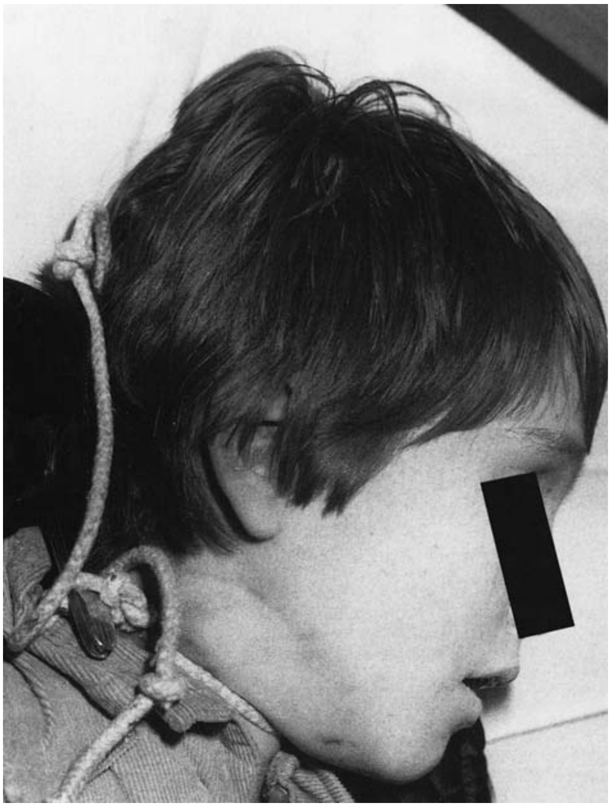

FIGURE 15.21 *Ligature strangulation by means of a 'Spanish windlass'. A penknife has been thrust through the knot to twist the ligature tight.*

The stiff, brownish-yellow appearance occurs post-mortem and tends to become more pronounced as the interval lengthens after death. The mark may be slightly wider, narrower or the same width as the actual ligature, depending partly upon how deeply it cut into the skin. There is often a narrow zone of reddened hyperaemia at either margin of the mark. This used to be taken to indicate that the ligature must have been applied during life – though by analogy with the spurious 'vital reaction' sometimes seen as a post-mortem development at the margin of burns (see Chapter 11), this sign is of doubtful reliability. Experiments by Pollak (personal communication) confirm this by the production of similar marginal reddening by applying ligatures soon after death – it seems to be due to lateral displacement of blood from the squeezed area immediately below the ligature.

THE POSITION OF THE LIGATURE MARK

The geometry of the mark is important in interpreting the fatal events. In strangulation, unlike hanging, the mark tends to encircle the neck horizontally and at a lower level. Typically it crosses immediately above or below the prominence of the

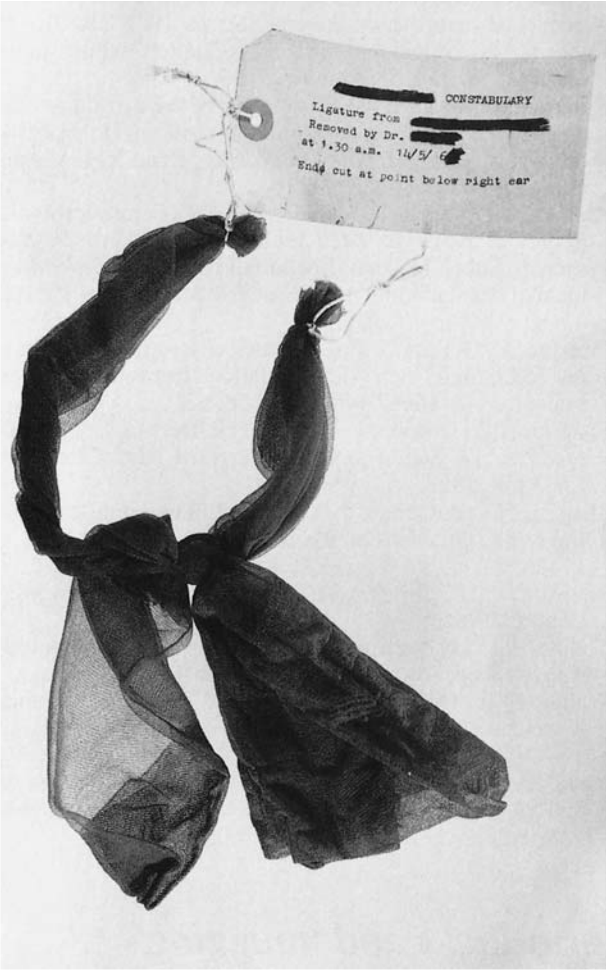
[redacted] CONSTABULARY
Ligature from [redacted]
Removed by Dr. [redacted]
at 1.30 a.m. 14/5/ 6[redacted]
Ends cut at point below right ear

FIGURE 15.22 *At autopsy on a victim of ligature strangulation, the ligature should be cut well away from the knot, if one is present. The knot should never be untied, but preserved for forensic examination. If the ligature is in danger of unravelling, it should be tied at two places and divided between these points. (Reproduced with permission from Robert Hale Publishers.)*

larynx and passes back to the nape of the neck. In homicide, where a single turn is used, there is often a cross-over point where the two ends of the ligature mark overlap. This may be at the front, side or back of the neck, depending on the relative positions of assailant and victim.

When a knot is tied, it may leave a mark on the skin and, of course, if multiple loops are present, some or all of these will be represented on the skin. Unless the killer is pulling upwards, the mark will not be seen to rise nor will there be a gap in the mark at the highest point, as seen in most hangings.

In the later section on hanging, the point will be made that sometimes a hanging mark may resemble a ligature mark, but the converse is almost unknown.

Other signs of local injury in ligature strangulation

There may be scratches on the neck, but these are usually caused by the attempts of the victim to pull away the ligature. Fingernail marks, sometimes linear and vertical, may be present as in manual strangulation, but not focal abrasions or bruises. Internally there may be superficial haemorrhage under the ligature mark, though this is often minimal. Depending on the force with which the ligature was applied, there may be deep damage in the muscles of the neck, but this tends to be less than in manual strangulation. Similarly, there may be laryngeal injury, but less severe than from pressure from the hands. The hyoid bone and thyroid horns may be fractured, especially where the ligature rides at the level of the thyrohyoid ligament, but damage to the laryngeal cartilages is much less common than in manual strangulation. Gonzales found in 24 fatal manual strangulations, four hyoid fractures but nine thyroid and cricoid fractures. In 25 deaths from manual strangulation, Simpson found only one hyoid fracture but 22 with thyroid horn fractures. It is rare for the main thyroid plate or the cricoid to be fractured unless gross violence was applied with excessive pressure by a strong ligature. Uncommonly, a narrow wire may be used, the so-called ‘cheese-cutter’ method. Here, strong pressure over a small area – such as that delivered by a piano wire – may actually lacerate the skin, and even cut into the deeper tissues and cartilages.

Where much bruising and abrasion is seen, especially if scattered and away from the actual ligature mark, then the possibility of a combination of manual and ligature strangulation must be considered. This is not uncommon, the handling of the throat either preceding or accompanying the application of the ligature. Fractures of the laryngeal structures may be caused by this element, rather than the associated ligature. A deeply sunken narrow ligature applied with force may damage the carotid arteries and focal injury to the intima may rarely be found on careful dissection.

The mode of death in ligature strangulation

The mode of death is more often the ‘classic asphyxia’ picture than in manual strangulation, where sudden cardiac death is common before congestive–petechial changes have time to occur. The contrast in the appearance of the skin immediately above and below the ligature mark is often striking, with pale skin below, and a puffy, oedematous, congested, cyanotic and haemorrhagic surface above. Petechiae may abound in the eyelids, conjunctivae and facial skin, and there may be bleeding from the ear and nose. This is by no means invariable, and many ligature strangulations die rapidly from vagal reflex cardiac arrest before any congestive signs have had time to appear.

This can occur in both homicide and suicide and, as mentioned earlier, the mechanism is obscure, especially where the victim – bent on self-destruction – has been able to complete multiple loops around his neck and tie several knots before dying.

Accidental ligature strangulation – which may be actual hanging if the body weight is thrown wholly or partially upon the ligature – is seen in the tragedies than occur to young children, who may become entangled in blind or curtain cords, usually when their cot or playpen is left too near a window.

ARM-LOCKS AND 'MUGGING'

The original meaning of 'mugging' has now been confused by its application, especially in North America, to any form of robbery with violence. The term strictly means throttling by pressure from an arm held around the throat. The attack is usually made from behind, the neck being trapped in the crook of the elbow. Pressure is then exerted either on the front of the larynx, or at one or both sides by the forearm and upper arm. The mechanism is further discussed in Chapter 10, as an arm-lock (or 'choke-hold') is a method of restraint used by police officers for law enforcement, but it is rapidly losing favour because of the number of inadvertent fatalities due either to 'asphyxia' or to reflex cardiac arrest. Following several fatalities during police arrests in the early 1990s, the Association of Chief Police Officers in Britain have issued recommendations that arm-locks be avoided during the restraint of violent prisoners.

The autopsy features are those of ligature strangulation with a broad object, in that signs may be minimal. Some diffuse abrasion may be seen, especially along the margin of the jaw or lower face, sometimes over a considerable area, caused by the friction of the forearm. Internally there may be diffuse bruising, but this again may be slight or even absent. The larynx may also escape damage, though if it is pressed backwards against the spinal column the thyroid horns and even the hyoid may fracture (see also Chapter 10).

HANGING

Hanging is a form of ligature strangulation in which the force applied to the neck is derived from the gravitational drag of the weight of the body or part of the body.

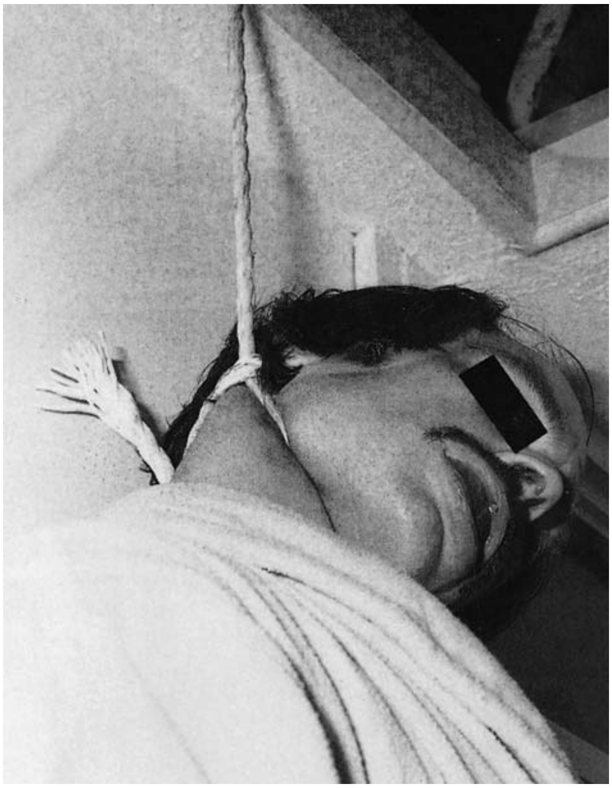

FIGURE 15.23 *Typical posture in hanging with the suspension point rising to the ear, tilting the head to the opposite side. The tongue is projecting because of upward pressure on the larynx and root of the tongue. There are no congestive appearances. As a slip-noose was used, the rope is in contact with the skin throughout the full circumference of the neck.*

Judicial hanging

The modern form of judicial execution is unrelated to the usual suicidal hangings seen in routine forensic practice, as it depends upon severe mechanical disruption of the neck structures. Until the nineteenth century judicial hanging in Britain was carried out even more barbarically by 'ordinary' hanging, where the victim was strangled at the end of a rope by his own weight. At the infamous Tyburn execution site near the present Marble Arch in London, tens of thousands were dispatched, the usual method being the placing of a rope noose around the neck of the condemned person, who stood on a cart or ladder. This support was then removed leaving the victim suspended. Some contemporary accounts suggest that many died with little further movement, but this was not the case for a considerable number and it was not unusual for the victims' relatives to pull on the victims legs to shorten the agony. It appears to have been a desire to increase the rapidity with which death occurred that 'improvements' were made, principally the use of a drop effected suddenly by means of a trapdoor. The use of a drop required the placing of the 'knot' – really a brass eyelet and rubber washer – under the left ear where it was less likely to be pulled over the head than at the occiput, a traditional site. The aim was that, when the rapidly falling body was suddenly arrested, the cervical spine would be dislocated resulting in traction on the spinal cord with consequent spinal cord or brainstem disruption. Exhumation of persons executed by this method shows that fracture was not usual but post-mortem accounts of findings indicate that cervical spine dislocation – occurring at various levels – was common with resulting cord or brainstem damage. However, the effects do seem to be variable with decapitation occurring at drop heights, which, in others, judging by contemporary descriptions of the execution, appear not even to have caused rapid unconsciousness. Though cerebral function presumably ceased immediately on cord or brainstem damage, the heart usually continued to beat, sometimes for up to 20 minutes until hypoxia caused arrest.

Suicidal and accidental hanging

Apart from the now rare 'lynching', hanging is almost always suicidal or accidental, the former being by far the most common. Hanging has many features in common with ligature strangulation. Death is, however, more often caused by reflex cardiac arrest from pressure on the carotid structures. Many more victims of hanging are found to have pale faces, rather than the congested, haemorrhagic appearance of the slower asphyxial type of death. James and Silcocks (1992), in a 15-year survey of hangings in Cardiff, found congestive–petechial features in 27 per cent, being related to the completeness of suspension.

Polson suggested that the usual pale face in hanging is caused by cerebral ischaemia from bilateral occlusion of the carotid arteries rather than a vasovagal effect, though this seems incapable of proof. In either event, the death can be taken to be rapid if no asphyxial signs are present.

Methods of hanging

Most hangings are self-suspension. This may be carried out by a wide variety of methods, but a typical method of self-suspension is to attach a thin rope to a high point such as a ceiling beam or staircase. The lower end is formed into either a fixed loop or a slipknot, which is placed around the neck while the intending suicide stands on a chair or other support. On jumping off or kicking away the support, the victim is then suspended with all or most of his weight upon the rope.

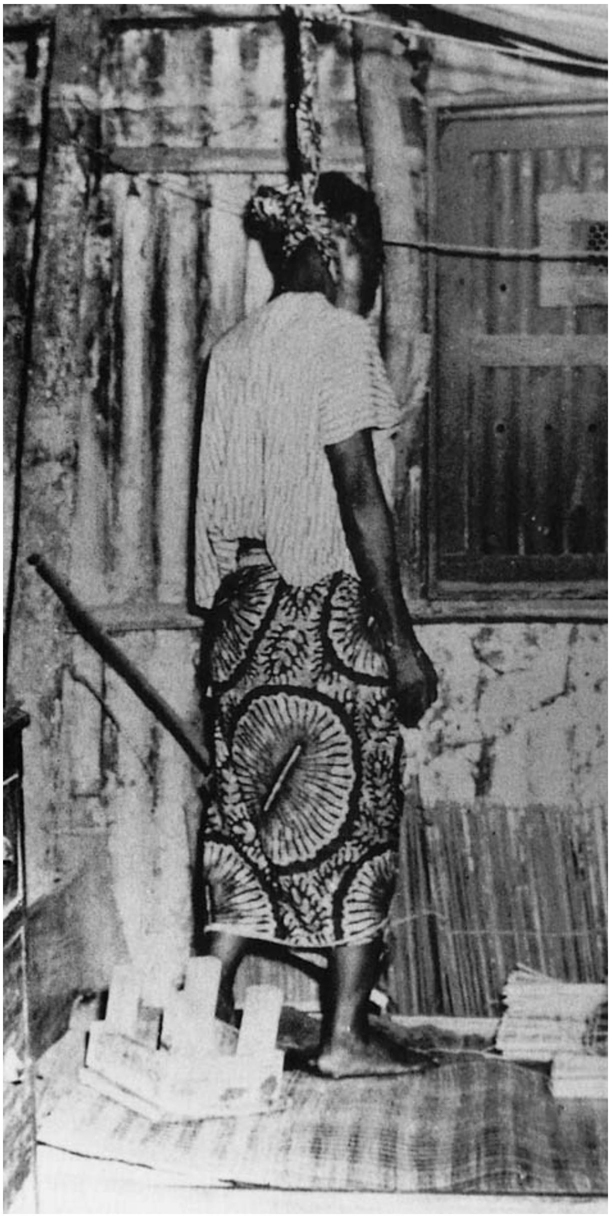

FIGURE 15.24 *Suicidal hanging with the feet flat on the floor. The deceased man stepped off the low stool, but the stretching of the cloth ligature prevented total suspension. Many suicidal hangings are successful at much lower levels, such as doorknobs and bed headboards.*

The many variations of this involve either the ligature or the height of suspension. Wires, string, pyjama cords, belts, braces (suspenders), scarves, neckties, stockings and numerous other devices may be used, depending on availability. In prison or police custody considerable ingenuity may be employed to defeat the efforts of the custodians to remove anything that could be used for self-destruction: shoelaces, stockings and torn bed-sheets have been used in prison cells.

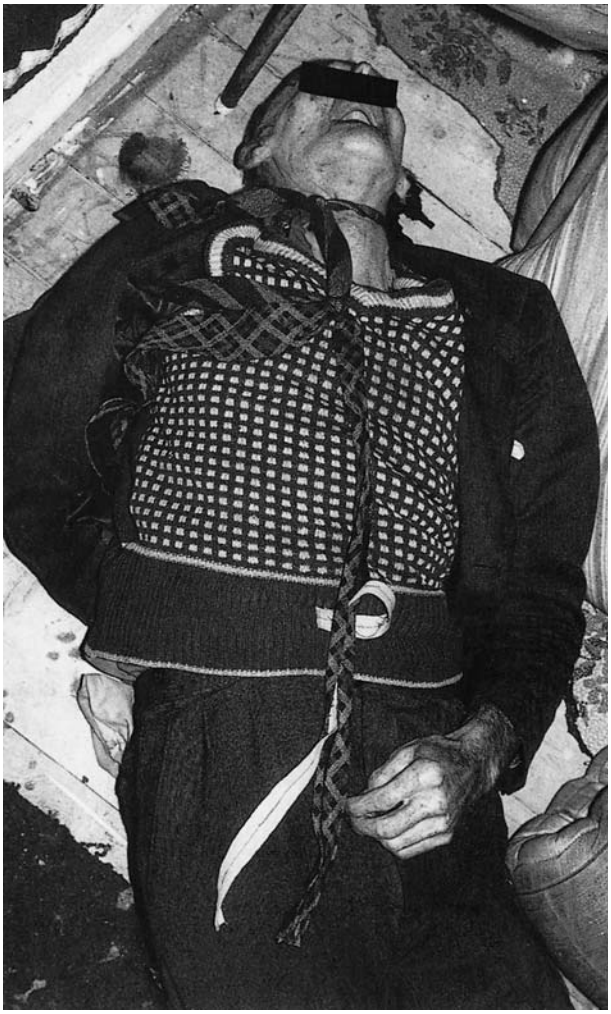

FIGURE 15.25 *Suicidal hanging by means of a necktie. The deceased man had suspended his neck from a hook on the back of a door by tying two neckties together. The hook had pulled out and he was found on the floor. Note the pale, uncongested face, indicating rapid cardiac arrest.*

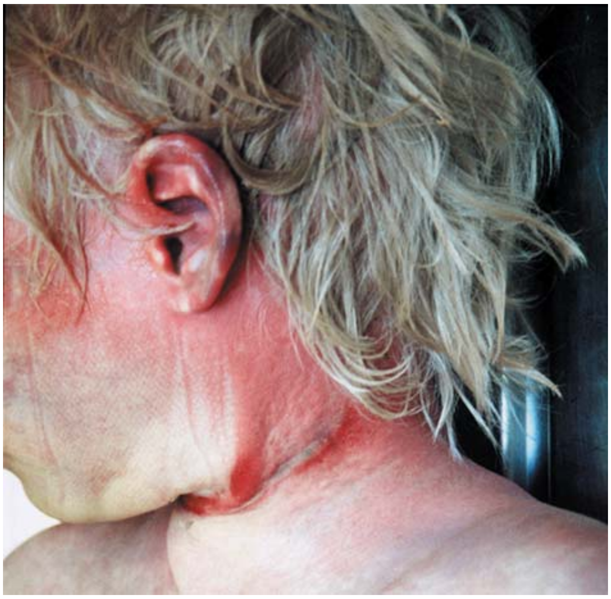

FIGURE 15.26 *Hanging deaths in children are very rare. This 7-year-old boy took his leave from his playmates and hanged himself from a bedpost under their very eyes.*

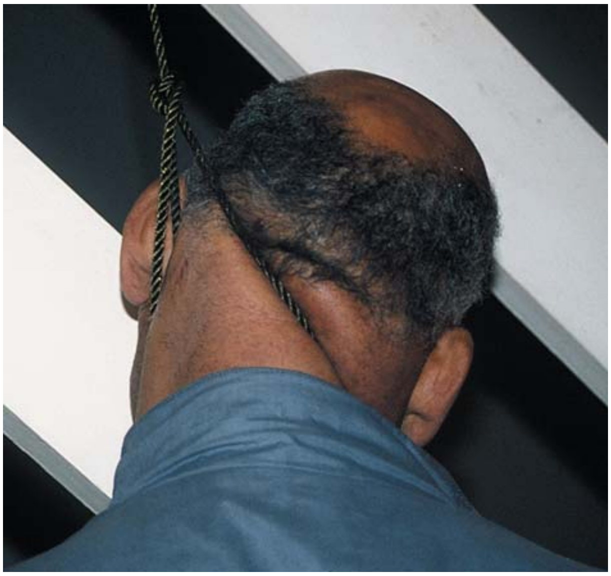

FIGURE 15.27 *Hanging with a fixed knot in the cord, so that there is a segment of skin free from the mark where the cord rises towards the knot. This was part of the common murder–suicide pattern in a domestic homicide.*

Suspension is often not high enough to keep the victim's feet clear of the floor. Commonly, when the person steps from his support, the stretch in the ligature rope is sufficient to allow the feet to reach the ground, but this by no means prevents a fatal outcome. The weight of the upper part of the body leaning into the noose is often more than enough to cause death. Successful hanging can occur from low suspension points, where the person is merely slumped with part of his weight into the ligature. Hanging can take place from doorknobs, bedposts and any other convenient low securing point. The body may be merely slumped against the door or bed or chair, with the legs and buttocks supported on the floor, so that only the weight of the chest and arms is contributing to the fatal pressure within the noose. One 'hanging' seen by the author (BK) was successfully achieved by merely leaning the neck into the shallow curve of a low clothesline stretched between two posts in a garden.

It is unusual for a suicidal hanging to be sufficiently violent for damage to the cervical spine to occur as the length of drop is usually too short. Only occasionally will a person jump from a roof or other high place with a rope around his neck – here severe injury can occur, even decapitation if the

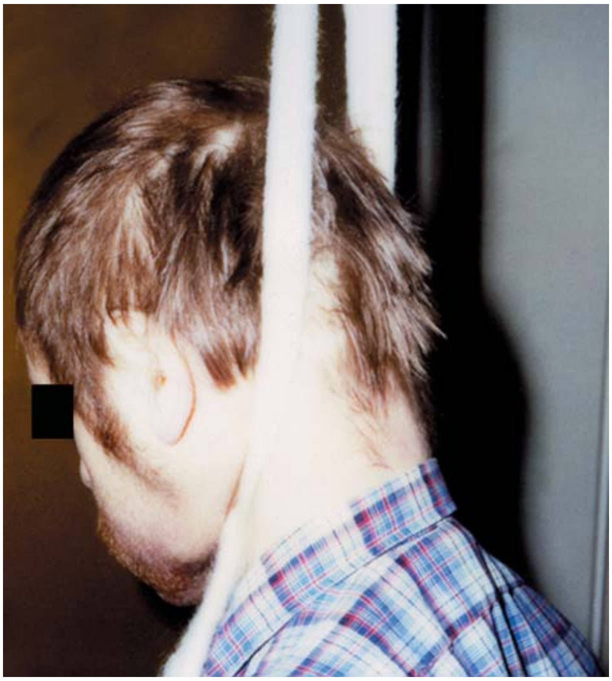

FIGURE 15.28 *This inmate had torn up a sheet and attached both ends to a water tap (not shown) and is leaning into the loop in sitting position with sparing of the back of the neck.*

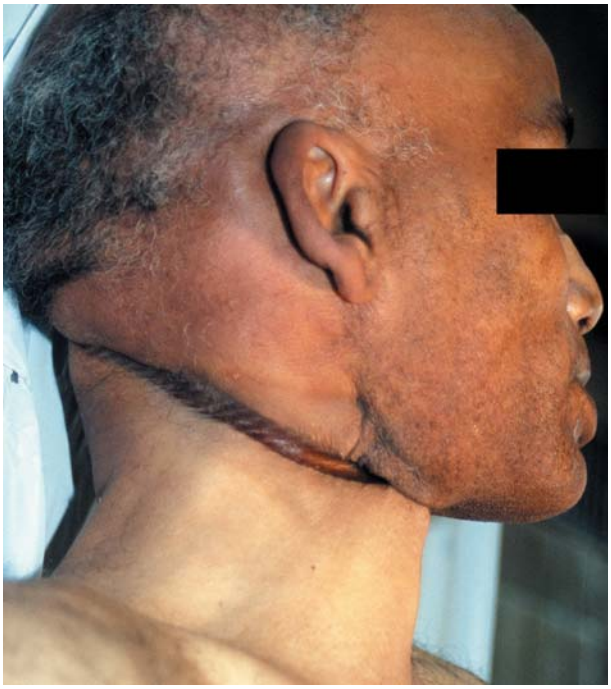

FIGURE 15.29 *A deep hanging mark showing the high position under the chin rising to the back of the neck. The spiral weave of the rope is clearly imprinted in the skin. The dark tint of the face is racial, not congestive.*

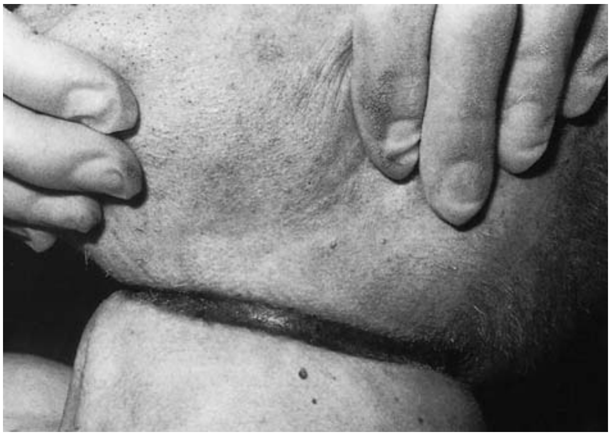

FIGURE 15.30 *A perfectly horizontal mark around the neck in a man who hanged himself from a staircase. A slipknot was used, which tightened so that the usual rising line of a hanging mark did not occur, illustrating the dangers of assuming that the usual is invariable.*

rope is strong enough. More often the jump will be from an attic trapdoor or a tree, sufficient to damage the vertebrae or atlanto-occipital joint. The author (BK) once saw a fractured cervical spine when a heavily built soldier jumped off a lavatory seat after securing a rope to the overhead cistern, but this is unusual.

The hanging mark

The mark on the neck in hanging can almost always be distinguished from ligature strangulation. The circumstances will usually indicate the fact of hanging, but sometimes the rope will break or become detached, and the deceased will be found lying with a ligature around his neck. There may then be doubt as to whether he was hanged or strangled. Obviously a search of the locus for a suspension point and signs of rope attachment will be the task of the investigators.

The hanging mark almost never completely encircles the neck unless a slipknot was used, which may cause the noose to tighten and squeeze the skin through the full circumference of the neck. In most instances the point of suspension is indicated by a gap in the skin mark, where the vertical pull of the rope leaves the tilted head to ascend to the knot and thence to the suspension point. This gap is usually seen at one or other side of the neck or at the centre of the back of the neck. Much less often it is under the chin. The knot itself may be impressed into the skin, especially if at the side of the head, above the gap where the anterior mark begins.

The hanging mark, the features of which resemble those described earlier in strangulation, is usually deepest at the side diametrically opposite the suspension point where the maximum load-bearing occurs. Like ligature strangulation,

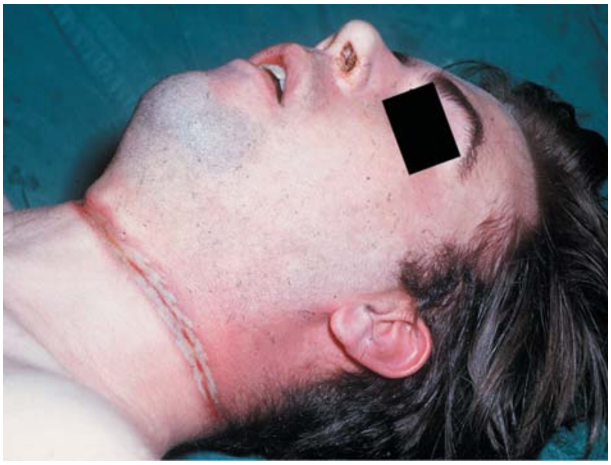

FIGURE 15.31 *A hanging mark showing the plaited imprint of a six-strand cable. There is a central line of abrasion, within a zone of pallor caused by vascular compression, outside which is a narrow band of hyperaemia. The mark is horizontal due to leaning forwards.*

the mark may be abraded, brown and dried to a parchment-like consistency after death. There may be a narrow red zone either above or below (or both sides) the ligature mark. This is not an indicator of vital reaction, as explained in relation to ligature strangulation, but is due to displacement of blood laterally from under the zone of maximum pressure.

THE POSITION OF THE HANGING MARK

The hanging mark is situated higher on the neck than in strangulation, usually being directly under the chin anteriorly, passing round beneath the jawbones and rising up at the sides or back of the neck to the usual gap under the knot.

An exception may be seen where the suspension point is low and part of the body is supported. For example, the victim may sit or slump on a chair, bed or floor with the rope attached to a point only slightly above neck level. Then the pull on the rope is almost at right angles to the axis of the body when the latter slumps away, so that the resulting mark may be almost horizontal and set at a lower level than usual on the neck. In such cases, if the body is cut down before investigators arrive, it may be much more difficult to distinguish hanging from strangulation, a fact which has been argued in more than one criminal trial or controversial death, such as that of Rudolf Hess in Spandau Prison, Berlin in 1987.

On rare occasions, hanging will take place with the suspension point at or above the point of the chin with the rope cutting into the back of the neck. Where the rope passes across the jaw margins to a knot in front of the face, the mechanism of death is somewhat obscure as the laryngeal region suffers little or no compression. Fatalities, however, certainly do occur, presumably as there is still opportunity for pressure on the carotids.

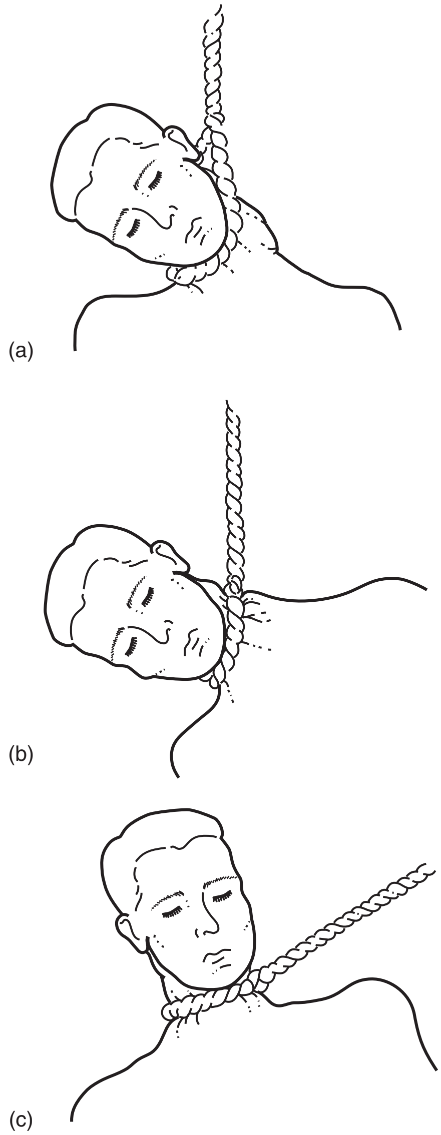

FIGURE 15.32 *The position of the hanging mark on the neck. (a) Usual position with fixed noose and high suspension point. The mark rises high to a gap. (b) If a slipknot is used, the tightness of the deeply impressed loop tends to find the smallest circumference on the neck, and may be lower and more horizontal. (c) If the suspension point is low and the subject leans away, the mark can be horizontal.*

Autopsy appearances in hanging

Apart from the appearances of the hanging mark, there are some other features to note. First, post-mortem hypostasis

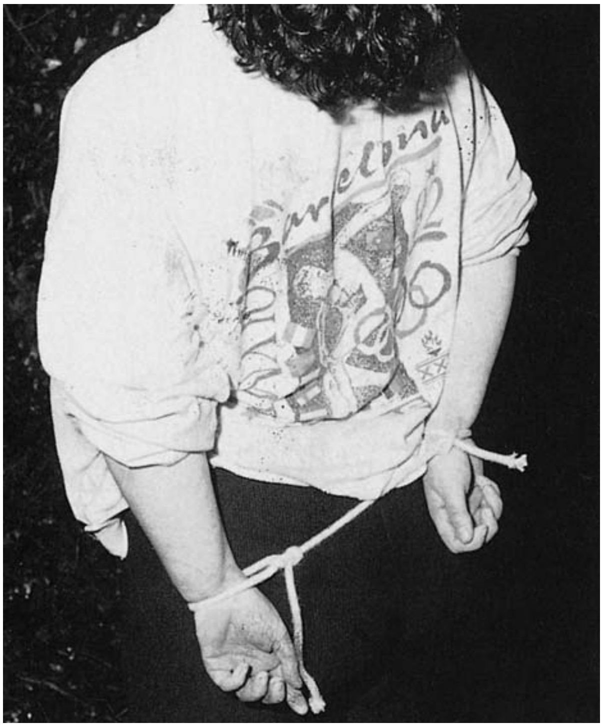

FIGURE 15.33 *The back of the victim in Figure 15.31, showing tying of the wrists. This does not indicate a homicide, unless the ligature could not have been self-applied. Some persons bent on self-destruction will try to ensure that they do not attempt to rescue themselves. A suicide note was left in this case.*

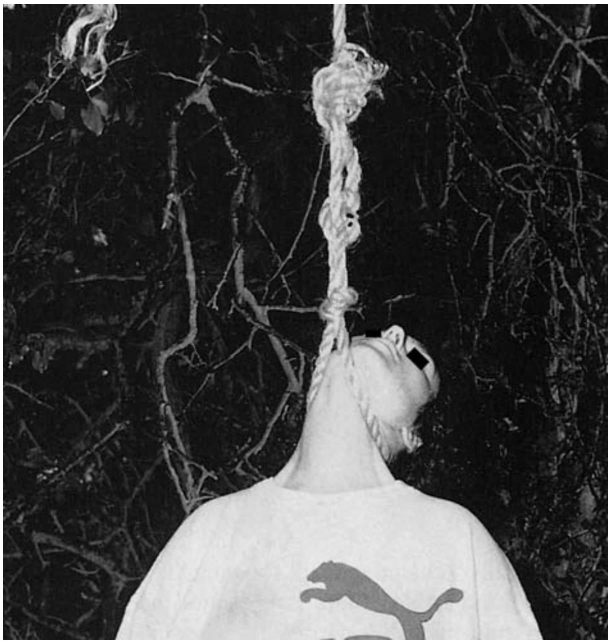

FIGURE 15.34 *Suicidal hanging with the rope in a frontal position. Even though the larynx cannot be compressed, death is due to pressure on the carotid arteries and sinuses. The face is pale, with no so-called 'asphyxial' signs.*

will occur in the legs and hands if the body has been in the vertical position for at least a few hours. When the body is cut down and laid horizontal for a considerable time, some or all of this appearance may flow back into the usual pattern.

Petechial haemorrhages are the exception rather than the rule, most series reporting them in approximately 25 per cent of cases. Such petechiae appear to occur more frequently in incomplete suspension but are frequently present in the absence of significant congestion. Congestion itself is far less usual than a pale face.

In the neck tissues there may be surprisingly little to find with an absence of laryngeal fracture or strap muscle haemorrhage being a common finding, especially if a soft ligature has been used. However, the literature suggests that an average figure for the incidence of soft tissue haemorrhage would be about 20–30 per cent of cases and for laryngeal fractures approximately 35–45 per cent of cases. Fractures of both the hyoid and thyroid may be seen.

Damage to the intima of the carotid arteries, often in the region of the sinuses, may sometimes be found on careful dissection. In hangings with an unusually long drop, severe disruption of the larynx can be found.

Mechanism of hanging

There are a number of mechanisms by which hanging may cause death, which may act independently or in concert. These include: stretching of the carotid sinus causing reflex cardiac arrest; occlusion of the carotid (and possibly vertebral) arteries; venous occlusion; airway obstruction resulting from pushing the base of the tongue against the roof of the pharynx or from crushing of the larynx or trachea; and finally spinal cord–brainstem disruption.

Whilst hanging shares some features with manual strangulation, the majority of victims of hanging are seen with pale faces rather than the congested, haemorrhagic appearance associated with the slower death resulting from pressure on the neck. This probably reflects a different mechanism, with reflex cardiac arrest and carotid occlusion more prevalent in hanging than in strangulation. Such mechanisms cause death rapidly, with unconsciousness resulting from bilateral carotid occlusion 3–11 seconds after the application of circumferential pressure.

However, the precise interplay of different mechanisms, which themselves appear only partially dependent on the completeness of suspension and the location of the noose, is impossible to determine from the pathological findings alone.

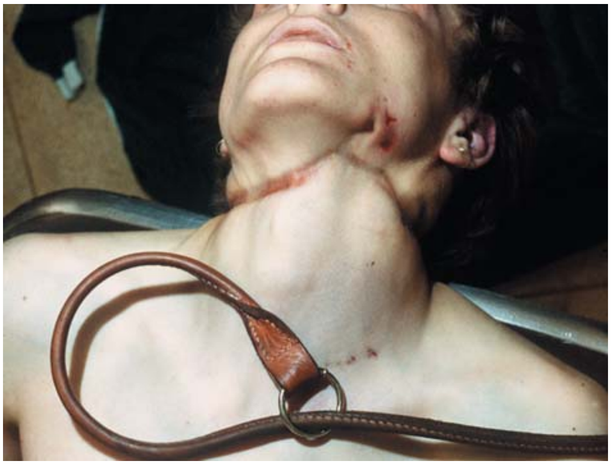

FIGURE 15.35 *Hanging by frontal suspension. A leather noose has been used so that compression has occurred round the whole periphery of the neck. The imprint of the metal ring on the jaw is well seen.*

Accidental and homicidal hanging

Though most hangings are suicidal, accidental hangings occur from two main causes. The first is entanglement in ropes or cords; this is relatively uncommon, and is usually seen in infants and children. The straps of a baby harness, used to support or restrain an infant, have led to death on a number of occasions, the child becoming entangled in leather or plastic reins when no adult was present. Any other cords may present a similar hazard, as in several cases seen by the author (BK), where infants standing in a cot near the bedroom window became entangled in a noose formed by the control cords hanging from a nearby blind or curtain.

On rare occasions, similar tragedies have befallen adults in factories, farms or ships, where a trip or fall has precipitated the victim head-first into machinery or structures, where ropes or cords have caused hanging or strangulation.

More commonly, a totally different situation causes fatal hanging or strangulation, almost exclusively in men and most often in young men, which is described in the next section.

Homicidal hanging is very rare, outside abuse of human rights and 'lynching'. For one individual to hang another, there must be either a disparity in their size and strength – or the victim must be drugged, drunk or otherwise incapacitated by fear, illness or senility. For a conscious, presumably unwilling, victim to be hanged by another, there will inevitably be signs of resistance, such as grip bruises on the arms due to restraint, or signs of binding the arms, wrists or legs.

THE SEXUAL ASPHYXIAS: AUTOEROTIC OR MASOCHISTIC PRACTICES

Though not all of this group are hangings or strangulations, this well-recognized phenomenon is conveniently considered here. The sexual asphyxias occur almost exclusively in men and only a few reports exist of women indulging in this dangerous habit, though Byard *et al.* (1990) have been able to contrast the death scenes in the two sexes, indicating that males tend to use more apparatus, whilst females are usually found naked with only a ligature in evidence.

The age of the male victims can vary widely, but is most often seen in young to middle-aged adults. The deaths that occur are obviously a small proportion of the total incidence of this strange practice. A fatal outcome is usually because of some failure of the equipment and this emphasizes the accidental nature of most of these deaths.

Features of the sexual asphyxias

The basic mechanism of the sexual asphyxias is the production of cerebral hypoxia, which in some men appears to produce hallucinations of an erotic nature.

This hypoxia is most often achieved by constriction of the neck by a ligature, which can be voluntarily tightened to produce vascular obstruction and perhaps airway stenosis. It would seem reasonable to think that vascular obstruction, either by blocking venous return – or even carotid flow – to the brain was the most likely explanation, as this would occur early. The production of cerebral hypoxia by anoxaemic anoxia through airway narrowing would surely be painful, distressing and slow in onset – hardly features which persons intent on self-gratification would welcome.

Some autoerotic procedures use other means of hypoxia, such as anaesthetic agents and a variety of volatile substances, tapering into 'solvent abuse', discussed elsewhere (Chapter 34). Whatever the mechanism, when cerebral hypoxia occurs with its attendant erotic sensations, progressive loss of voluntary control as consciousness fades allows the constrictive device to slacken, so that the subject recovers. As some fatal cases show clear evidence of repeated previous escapades, it is obvious that the mechanism usually functions quite successfully and that death was the result of some unforeseen complication.

A common practice is to place a fixed noose around the neck, so that compression will cease as soon as muscular tension on the free end is relaxed. This free end may be passed down the front or back to be fastened to the ankles, so that by extending the legs, the noose around the neck is

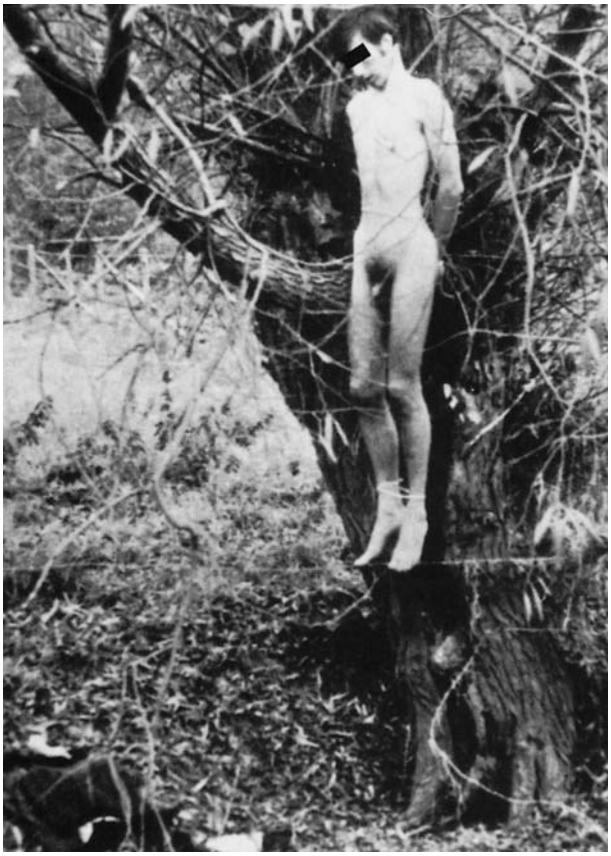

FIGURE 15.36 *Hanging in circumstances which suggest a sexual–masochistic motive. The young man showed no signs of previous depression or suicidal tendency. He was found naked, with arms and legs tied together with wire. Though homicide was suspected at first, the bonds were loose and easily self-applied. No satisfactory explanation was forthcoming.*

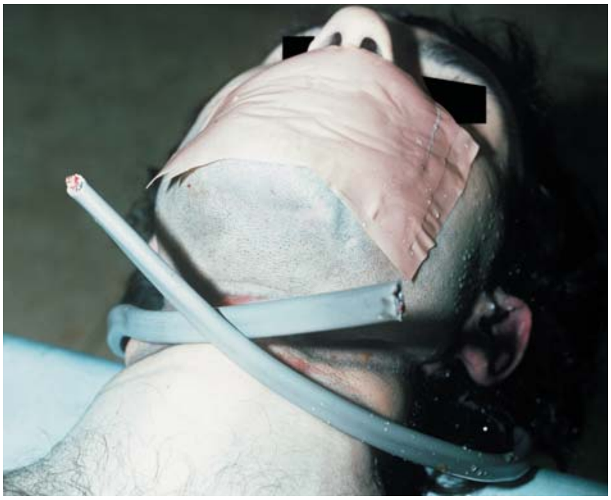

FIGURE 15.37 *A hanging in which it was not clear if there was a true sexual–masochistic element. The victim was suspended by electric cable from an attic trapdoor with his face plastered with surgical tape and the wrists tied together. Homicide could be excluded. In some deaths with an undoubted sexual element, the means used would inevitably lead to a fatal outcome, so there seems to be an overlap between masochism and suicide.*

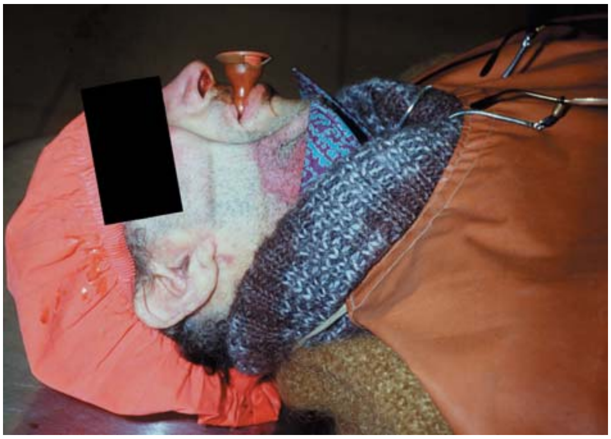

FIGURE 15.38 *Fatal sexual asphyxia in an engineer with a rubber fetish. He was found dead connected to a dental anaesthetic machine, with a rubber face mask strapped on over a rubber teat in the mouth. Bizarre pornographic material devoted to dentistry was in the vicinity. (Figure reproduced by kind permission of Dr S Leadbeatter.)*

tightened. Compression may also be attained by passing the rope over a support and hanging a weight on the end to produce tension.

Other men may secure the end to an overhead support and achieve constriction by slumping against the pull of the rope. In these and similar cases, it is difficult to understand how death can be avoided once consciousness is lost, but evidence of repetition on previous occasions indicates that enough voluntary control must be retained to allow the person to retrieve the situation after obtaining his erotic satisfaction.

Similar hypoxia may be produced by placing the head in a plastic bag, by pushing the head into a confined space, by postural contortions (often associated with bondage) that constrict the airway, or a host of other ingenious methods. One case seen by the author (BK) deliberately constructed a large sack made from impervious material, into which he crawled to attain his hypoxia. Certain features of these deaths make them fairly easy to differentiate from suicides, a distinction which is important, though some cases are impossible to separate from deliberate attempts at self-destruction (Knight 1979).

Bondage is common, with sometimes elaborate and bizarre methods of restriction. The wrists and ankles may be tied together with cord, wire, chains, padlocks or handcuffs, for example, and it is vital for the investigators to ensure

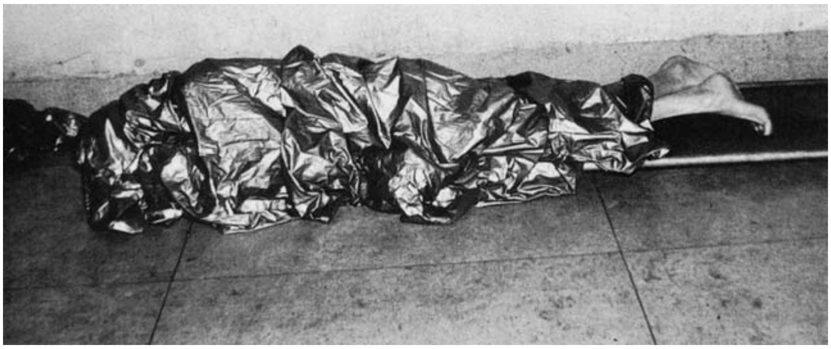

FIGURE 15.39 *A presumed instance of sexual asphyxia. The deceased was naked and had enveloped himself in a large impervious bag, which he had made himself. It seemed impossible for him to save himself once unconsciousness had occurred, but there appeared to be no suicidal tendencies in the history.*

that such bondage could have been secured by the deceased himself to exclude the possibility of homicide, usually by a heterosexual or homosexual partner involved in the ritual.

The bonds may be sexually orientated, with straps around the crutch or constricting the genital organs. In cases seen by the author (BK), one had made a metal sheath for the penis that was padlocked in place and another clamped the penis within a metal shackle.

Masks and gags may be in place, again a manifestation of restraint. The mouth may be sealed by adhesive plaster and the eyes blindfolded. Rubber or leather often features in the masks, which may have eye-slits or completely envelope the face. Transvestism is common, female attire being worn either overtly or under male clothing. False breasts and nipples may be fabricated or brassières padded with cloth. This is a common phenomenon, however, without sexual asphyxia, as any coroner's autopsy service will reveal some men (often elderly) as hidden transvestites, wearing female underclothing beneath their suits. Fetishism is often seen, especially rubber or shiny plastic or leather. Female wigs and make-up are sometimes encountered, though the syndrome does not seem to be overtly associated with homosexual activity.

Pornographic literature is often within view and may be spread around the body at the scene of death. The act is sometimes performed before a mirror and masturbation is common. The mere emission of semen found at autopsy does not confirm sexual activity in itself, however, as post-mortem discharge of semen from the meatus is common in any type of death, not only in asphyxia (as used to be commonly alleged by standard textbooks). Occasionally, lewd writings are left near the body and even upon the body surface. The author (BK) has seen one where obscene graffiti were written with a ball-pen on the abdomen and around the glans penis.

Perhaps more importantly, overt suicide notes are never present, helping to distinguish the cases from definite self-destruction. Another important piece of evidence against suicide is the fact that, in some neck ligatures, the rope may be padded by fabric to avoid leaving a telltale mark on the neck – an act incompatible with an intention of suicide.

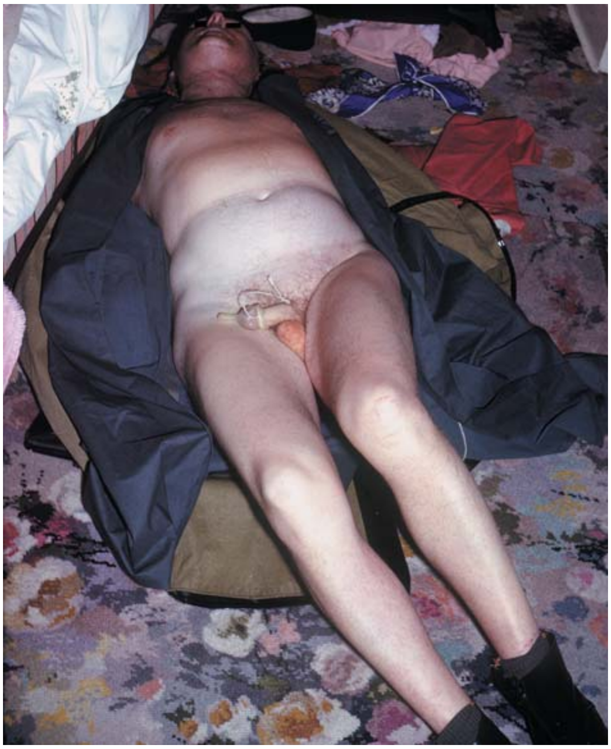

FIGURE 15.40 *A classical sexual asphyxia where the deceased has used a dog collar round the neck (removed before photography), the chain passing down to the ankles, which provide temporary traction. The common features of rubber fetishism and penile bondage were present.*

Rarely, autoerotic gratification may be achieved by the application of electric current, usually low voltage applied to the genitals. Electrical faults, such as in a transformer, may lead to a fatal outcome.

Medico-legal aspects of the sexual asphyxias

Though all these features make the true nature of the death clear to the medical examiner, it can sometimes be difficult

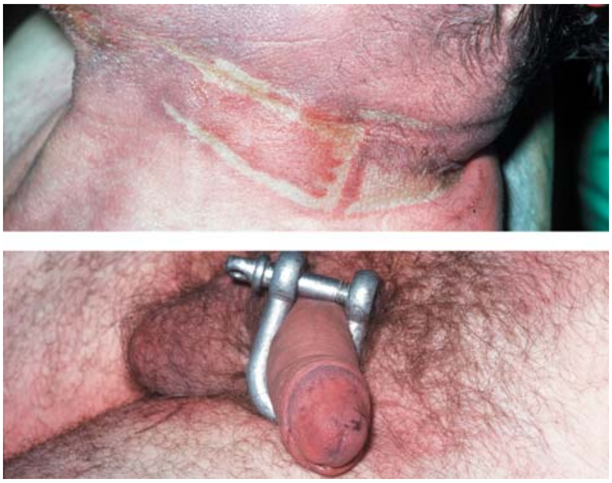

FIGURE 15.41 *A sexual asphyxia from a buckled belt applied tightly to the neck and a metal shackle on the penis. Obscene phrases were written on the glans and abdomen.*

to convince the police, the coroner or judge – and especially the relatives – that death was accidental. The stigma still attached to suicide and the revulsion felt at the perverted sexual element may cause the relatives to prefer homicide as the cause. The judicial authorities often lean towards suicide, especially when they are unaware of the existence of this strange syndrome of the masochistic or sexual asphyxias. The doctor is often the only person able to explain the relatively common occurrence of this phenomenon and assist in reaching the correct conclusion. It has to be admitted, however, that in a small minority of these deaths, it seems inevitable that the victim must have been aware of the fatal outcome, as the mechanics of the asphyxia are such that no escape could have been foreseen. In these instances the circumstances are such that a mixed motivation must have existed. For example, hanging by the neck in free suspension from a tree or from the trapdoor of an attic is a situation from which escape is impossible, even though sexual attributes such as nudity, bondage and masking were present (Knight 1979).

Where the victim is young, motivation is again less clear. Youths of 12 or 14 years are sometimes found hanged, usually without the secondary sexual or bizarre attributes listed above. There is usually nothing in the history to suggest suicide, and quite often the coroner or other legal authority will decide that death was due to 'experimenting with ropes', sometimes alleging that it followed watching lynching in cowboy films on television. It is probable that most of the obscure deaths are also associated with aberrant sexual activity, even if more positive evidence is lacking. Recognition of the true nature of most of these deaths is vital for the medical examiner, as spurious homicide investigations may be initiated if they are misinterpreted. More commonly a false suicide verdict may result, which can have financial implications in respect of life insurance. A mistaken belief in suicide may be preferred by some families, however, rather than the shame of a publicised sexual aberration.

REFERENCES AND FURTHER READING

Adelson L. 1962. Possible neurological mechanisms for sudden death with minimum anatomical findings. *J Forensic Med* **9**:10–16.

Anscombe AM, Knight BH. 1996. Case report. Delayed death after pressure on the neck: possible causal mechanisms and implications for mode of death in manual strangulation discussed. *Forensic Sci Int* **78**:193–7.

Bockholdt B, Maxeiner H. 2002. Hemorrhages of the tongue in the postmortem diagnostics of strangulation. *Forensic Sci Int* **126**:214–20.

Bohnert M, Faller-Marquardt M, Lutz S, *et al.* 2001. Transfer of biological traces in cases of hanging and ligature strangulation. *Forensic Sci Int* **116**: 107–15.

Bowden KM. 1962. The larynx. *J Forensic Med* **9**:10–16.

Brinkmann B, Koops E, Wischhusen F, *et al.* 1981. [Compression of the neck and arterial obstruction (author's transl).] *Z Rechtsmed* **87**:59–73.

Brouardel P. 1897. *La pendaison, la strangulation, la suffocation, la submersion.* Baillière, Paris.

Byard RW, Bramwell NH. 1988. Autoerotic death in females. An underdiagnosed syndrome? *Am J Forensic Med Pathol* **9**:252–4.

Byard RW, Hucker SJ, Hazelwood RR. 1990. A comparison of typical death scene features in cases of fatal male and autoerotic asphyxia with a review of the literature. *Forensic Sci Int* **48**:113–21.

Byard RW, Hucker SJ, Hazelwood RR. 1993. Fatal and near-fatal autoerotic asphyxial episodes in women. Characteristic features based on a review of nine cases. *Am J Forensic Med Pathol* **14**:70–3.

Camps FE, Hunt AC. 1959. Pressure on the neck. *J Forensic Med* **6**:116–36.

Cooke CT, Cadden GA, Hilton JM. 1988. Unusual hanging deaths. *Am J Forensic Med Pathol* **9**:277–82.

Danto BL. 1980. A case of female autoerotic death. *Am J Forensic Med Pathol* **1**:117–21.

DiMaio VJ. 2000. Homicidal asphyxia. *Am J Forensic Med Pathol* **21**:1–4.

Ditto EWD. 1981. Electrocution during sexual activity. *Am J Forensic Med Pathol* **2**:271–2.

Duflou JA, Lamont DL, Knobel GJ. 1988. Homicide in Cape Town, South Africa. *Am J Forensic Med Pathol* **9**:290–4.

Emson HE. 1983. Accidental hanging in autoeroticism. An unusual case occurring outdoors. *Am J Forensic Med Pathol* **4**:337–40.

Evans KT, Knight B. 1982. *Forensic radiology*. Blackwell Scientific Publications, Oxford.

Ford R. 1957. Death by hanging of adolescents and young adult males. *J Forensic Sci* **2**:171–4.

Gonzalez TA. 1933. Fractures of the larynx. *Arch Pathol* **15**:55–62.

Gordon I, Shapiro HA, Taljaard JJ, *et al.* 1976. Aspects of the hyoid-larynx complex in forensic pathology. *Forensic Sci* **7**:161–70.

Green MA. 1973. Morbid anatomical findings in strangulation. *Forensic Sci* **2**:317–23.

Grellner W, Benecke M. 1997. The quantitative alteration of the DNA content in strangulation marks is an artefact. *Forensic Sci Int* **89**:15–20.

Harm T, Rajs J. 1981. Types of injuries and interrelated conditions of victims and assailants in attempted and homicidal strangulation. *Forensic Sci Int* **18**:101–23.

James R, Nasmyth-Jones R. 1992. The occurrence of cervical fractures in victims of judicial hanging. *Forensic Sci Int* **54**:81–91.

James R, Silcocks P. 1992. Suicidal hanging in Cardiff – a 15-year retrospective study. *Forensic Sci Int* **56**:167–75.

Kennedy NM, Whittington RM, White AC. 1995. Suicide by self-strangulation whilst under observation. *Med Sci Law* **35**:174–7.

Khokhlov VD. 2001. Calculation of tension exerted on a ligature in incomplete hanging. *Forensic Sci Int* **123**:172–7.

Knight B. 1979. Fatal masochism – accident or suicide? *Med Sci Law* **19**:118–20.

Koops VE, Puschel K. 1983. Obstruction of the carotid arteries and upper airway by manual strangulation. In: Holczabek W (ed.), *Beiträge zur gerichtlichen medizin*. Verlag Franz Deuticke, Vienna, vol. XLI.

Kornblum RN. 1986. Medical analysis of police choke holds and general neck trauma. *Trauma* **27**(Part I):760; **28**(Part II):13–64.

Kunnen K, Thomas F, Van der Velde E. 1966. Semimicroradiography of the larynx. *Med Sci Law* **6**:218–19.

Leth P, Vesterby A. 1997. Homicidal hanging masquerading as suicide. *Forensic Sci Int* **85**:65–71.

Lew EO. 1988. Homicidal hanging in a dyadic death. *Am J Forensic Med Pathol* **9**:283–6.

Litman RE, Swearingen C. 1972. Bondage and suicide. *Arch Gen Psychiatry* **27**:80–5.

Luke JL, Reay DT, Eisele JW, *et al.* 1985. Correlation of circumstances with pathological findings in asphyxial deaths by hanging: a prospective study of 61 cases from Seattle, WA. *J Forensic Sci* **30**:1140–7.

Mallach HJ, Pollak S. 1998. [Simulated suicide by hanging after homicidal strangulation.] *Arch Kriminol* **202**:17–28.

Maxeiner H. 1998. 'Hidden' laryngeal injuries in homicidal strangulation: how to detect and interpret these findings. *J Forensic Sci* **43**:784–91.

Michalodimitrakis M, Frangoulis M, Koutselinis A. 1986. Accidental sexual strangulation. *Am J Forensic Med Pathol* **7**:74–5.

Minyard F. 1985. Wrapped to death. Unusual autoerotic death. *Am J Forensic Med Pathol* **6**:151–2.

Muller E, Franke WG, Koch R. 1997. Thyreoglobulin and violent asphyxia. *Forensic Sci Int* **90**:165–70.

Perper JA, Sobel MN. 1981. Identification of fingernail markings in manual strangulation. *Am J Forensic Med Pathol* **2**:45–8.

Pollanen MS. 2001. Subtle fatal manual neck compression. *Med Sci Law* **41**:135–40.

Prinsloo I, Gordon I. 1951. Postmortem dissection artefacts of the neck and their differentiation from ante-mortem bruises. *S Afr Med J* **25**:358–61.

Rajs J, Thiblin I. 2000. Histologic appearance of fractured thyroid cartilage and surrounding tissues. *Forensic Sci Int* **114**:155–66.

Reay DT, Eisele JW. 1982. Death from law enforcement neck holds. *Am J Forensic Med Pathol* **3**:253–8.

Reay DT, Holloway GA, Jr. 1982. Changes in carotid blood flow produced by neck compression. *Am J Forensic Med Pathol* **3**:199–202.

Resnick H. 1972. Eroticised repetitive hangings: a form of selfdestructive behaviour. *Am J Psychother* **26**:4–8.

Reuter F. 1933. *Lehrbuch der gerichtlichen Medizin.* Urban and Schwarzenberg, Berlin.

Rogde S, Hougen HP, Poulsen K. 2001. Asphyxial homicide in two Scandinavian capitals. *Am J Forensic Med Pathol* **22**:128–33.

Rothschild M, Maxeiner H. 1992. [How extensive can injury of the larynx in self-choking be?] *Arch Kriminol* **189**:129–39.

Seabourne A, Seabourne G. 2001. Suicide or accident – self-killing in medieval England: series of 198 cases from the eyre records. *Br J Psychiatry* **178**:42–7.

Shapiro H. 1988. *Strangulation*. In: Gordon F, Shapiro H, Berson S (eds), *Forensic medicine: a guide to principles*, 3rd edn. Churchill, Edinburgh.

Shapiro HA, Gluckman J, Gordon I. 1962. The significance of the nail abrasion of the skin. *J Forensic Med* **9**:17–19.

Simonsen J. 1988. Patho-anatomic findings in neck structures in asphyxiation due to hanging: a survey of 80 cases. *Forensic Sci Int* **38**:83–91.

Simpson CK. 1985. *Forensic medicine*, 9th edn. Arnold, London.

Sivaloganathan S. 1985. Catheteroticum. Fatal late complication following autoerotic practice. *Am J Forensic Med Pathol* **6**:340–2.

Spence MW, Shkrum MJ, Ariss A, *et al.* 1999. Craniocervical injuries in judicial hangings: an anthropologic analysis of six cases. *Am J Forensic Med Pathol* **20**:309–22.

Thomas JE. 1969. Hyperactive carotid sinus reflex and carotid sinus syncope. *Mayo Clin Proc* **44**:127–39.

Walter PF, Crawley IS, Dorney ER. 1978. Carotid sinus hypersensitivity and syncope. *Am J Cardiol* **42**:396–403.

Wesselius CL, Bally R. 1983. A male with autoerotic asphyxia syndrome. *Am J Forensic Med Pathol* **4**:341–4.

Wright RK, Davis J. 1976. Homicidal hanging masquerading as sexual asphyxia. *J Forensic Sci* **21**:387–9.

Yada S, Tsugawa N, Uchida H. 1972. Demonstration of thyroglobulin in the heart blood in fatal cases of strangulation. *Acta Crim Jpn* **38**:60–8.

C H A P T E R 16

Immersion deaths

- Signs of immersion 395
- Estimation of duration of immersion 396
- Death from natural causes before entering the water 397
- Natural death whilst in the water 397
- Injuries sustained before entering the water 398
- Injuries sustained in the water 398
- Deaths from immersion other than drowning 399
- Death from drowning 401
- The autopsy signs of drowning 402
- Chemical changes in the blood in drowning 404
- Histological changes in drowning 405
- Diatoms and the diagnosis of drowning 406
- References and further reading 408

Many corpses are recovered from water, but not all have drowned. Of those that have drowned, pathological proof is often difficult or even impossible to obtain. The autopsy diagnosis of drowning presents one of the major problems in forensic medicine, especially when there is delay in recovering the victim.

Bodies retrieved from water may have:

- died from natural disease before falling into the water
- died from natural disease while already in the water
- died from injury before being thrown into the water
- died from injury while in the water
- died from effects of immersion other than drowning
- died from drowning.

All the above may show signs of immersion on examination, but this is rarely helpful, other than confirming that they had indeed been in water. It does not assist in differentiating the mode of death.

SIGNS OF IMMERSION

- Maceration of the skin begins within minutes in warm water, such as death in a bathtub, but in cold water is visible after a variable time, the minimum probably being 4 or 5 hours depending upon temperature. Polson's (1962) experience suggests a much longer time, from 12 to 48 hours, but sodden skin on the hands can undoubtedly occur much earlier than this.

  The first signs tend to be on areas with an appreciable keratin layer (such as the fingertips, palms, backs of the hands and later the soles) where the surface becomes wrinkled, pale and sodden – the so-called 'washer-woman's skin', for obvious reasons. Areas protected by clothing develop these changes later than exposed parts so that the hands usually reveal more marked maceration than the feet during the first few days.

  After prolonged immersion, wide areas of skin assume the same appearance, especially where the stratum corneum is thick or naturally creased, such as the extensor surface of knees and elbows. After some days in warmer water and up to several weeks in cold, the thick keratin of hands and feet becomes detached and eventually peels off in 'glove and stocking' fashion. The nails and hair become loosened at about the same time. Problems of obtaining fingerprints in such circumstances are dealt with later.
- Cutis anserina – or 'goose-flesh' – is a common finding in immersed bodies, but is related to cold rather than warm water. The erector pilae muscles attached to each hair follicle can contract in any type of death and cause a generalized pimpling of the skin. It is often stated that rigor mortis can produce this goose-flesh appearance, but this is doubtful, as rigor does not shorten muscles appreciably. From whatever cause, however, agonal contraction of the erector pilae is common and in immersed bodies has no diagnostic significance.

- The distribution of post-mortem hypostasis is also of no value in immersed bodies. According to Bonte *et al.* (1986), electrocution in the bath seems cause the margin of hypostasis to correspond with the horizontal line of the water level (see Chapter 12). It is usual for most corpses to float or hang in the water with buttocks uppermost, and the head and limbs hanging down, but water movements often roll the body constantly unless in a placid lake or wide river. Other corpses may float upon the back, and gravitational staining of the skin may therefore be in any pattern if the body is subject to frequent changes or posture. The hypostasis of bodies from cold water is frequently pink, but this has no diagnostic value. The reddish colour is caused by the presence of unreduced oxyhaemoglobin in the superficial blood vessels and is the result of cold, whether ante-mortem or post-mortem. It may be seen in any body taken from a mortuary refrigerator or found in a cold environment, irrespective of immersion. In immersion, however, it does not have the typical distribution over extensor surfaces and large joints, nor the rather dusky tinge that is seen in true ante-mortem hypothermia.
- Mud, coal-slurry, oil, silt or sand may be present on or in the body, as well as other artefacts such as seaweed, water weed, algae and small aquatic animal life of many types. Mud may be adherent to the whole body surface and clothing, or may be retained in the hair, mouth, nostrils, ears and other more protected sites. In appropriate circumstances, such deposits should be retained for scientific examination, if they might help establish where the body has travelled from in marine or riverine waters. Sand may be found deep in the respiratory passages and stomach, especially if the body has been rolled by the waves on a beach. Such deep penetration is not evidence of live aspiration. Where a body is decomposed and partly skeletalized, molluscs may colonize the bones. The author (BK) has had several cases recovered from the sea where mussels and barnacles filled the eye sockets and other crevices of the skull.

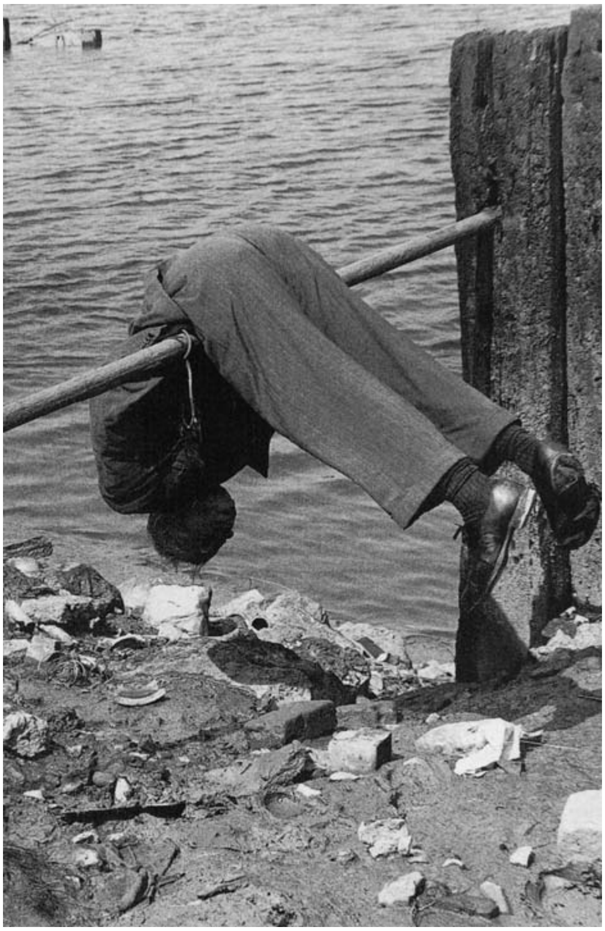

FIGURE 16.1 *The diagnostic problems associated with bodies recovered from the water. In a death such as this, where a body is left by the falling tide, it may be difficult or even impossible to prove drowning, especially if the post-mortem period is long.*

ESTIMATION OF DURATION OF IMMERSION

This is another difficult problem and one that is often too dogmatically answered by doctors with insufficient experience to appreciate the potential errors. The over-riding variable factor is water temperature, which has the most effect upon decomposition. Water pollution has little to do with speed of putrefaction, as most of the decomposing organisms come from the gut of the body itself.

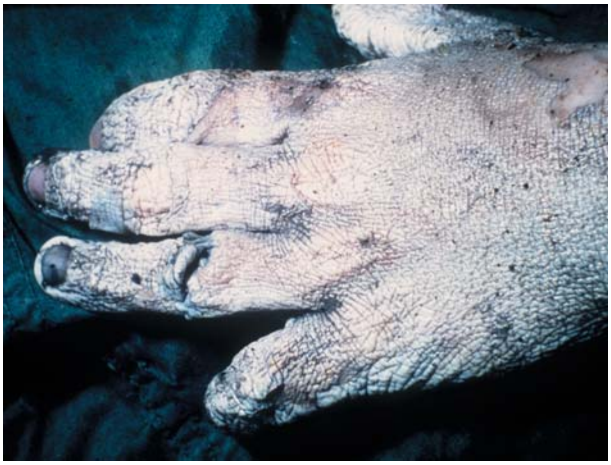

FIGURE 16.2 *Waterlogged skin ('washer-woman's hands') after 2 weeks' immersion in a temperate climate. Fingerprints can still be obtained from hands like these, with appropriate techniques.*

When a body falls into water in average temperate climates, the following is an approximate guide to timing, in conjunction with the other usual signs, such as rigor:

- if no wrinkling of the finger-pads is present, less than a few hours
- wrinkled fingers, palms and feet, progressively from half a day to 3 days
- early decomposition, often first in the dependent head and neck, abdomen and thighs: 4–10 days
- bloating of face and abdomen with marbling of veins and peeling of epidermis on hands and feet, and slippage of scalp: 2–4 weeks
- gross skin shedding, muscle loss with skeletal exposure, partial liquefaction: 1–2 months.

These times may be reduced or exceeded by wide margins according to animal predation, climatic changes and bodily build.

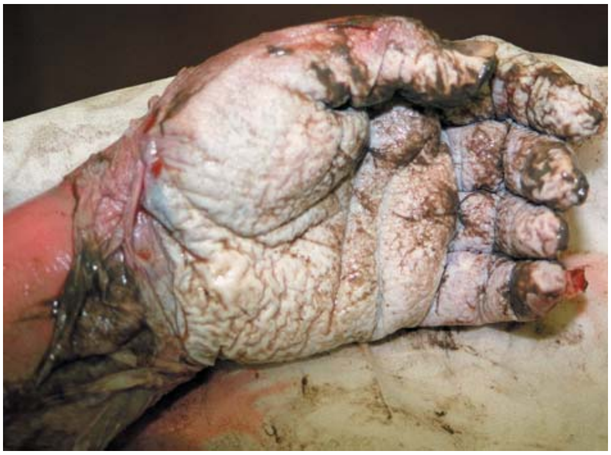

FIGURE 16.3 *Waterlogged skin after one week's immersion in a cold climate.*

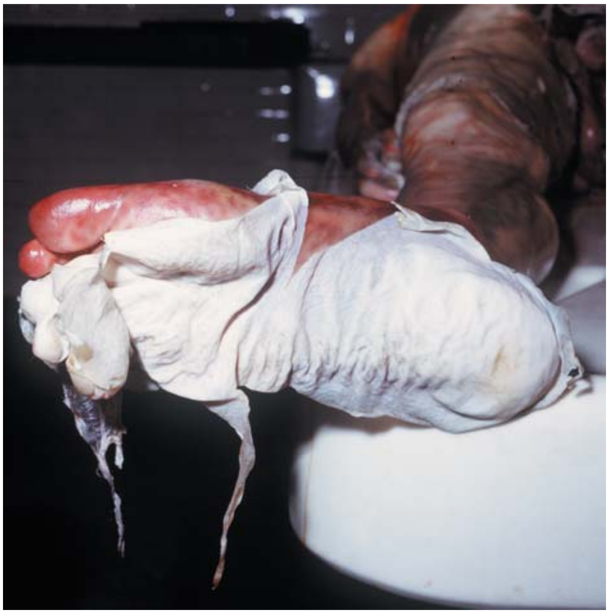

FIGURE 16.4 *Peeling of the epidermis in a 'stocking' fashion after a few weeks' immersion. The rate at which such changes occur is so variable, because of environmental factors, that no timetable can be offered to attempt to date death. In a temperate-zone summer, this stage would require at least 2 weeks and probably much longer.*

DEATH FROM NATURAL CAUSES BEFORE ENTERING THE WATER

This is relatively uncommon, but does occur. The victim may suffer one of the many causes of sudden natural death, most commonly a cardiovascular catastrophe, in circumstances where he involuntarily falls into the water. The last such instance dealt with by the author (BK) was a man returning to his yacht in a rubber dinghy, alone and in the dark. The dinghy was found empty and a search recovered the body, autopsy revealing a well-established myocardial infarct with no evidence of drowning, though the latter is never conclusive. Other instances have involved persons collapsing from natural disease while on boats, a canal tow-path or a river bank.

Of course, in the absence of witnesses it is difficult to separate such cases from those who first fall into the water and then suffer a natural death. The autopsy must be directed towards identifying the pathological lesions and to excluding, as far as is possible, any evidence of drowning.

NATURAL DEATH WHILST IN THE WATER

As discussed above, it may well be impossible to determine whether a person recovered from the water, with a potentially fatal natural disease, suffered this before or after immersion. Because of the problems (to be described later) of positively diagnosing drowning, it is difficult to exclude this as either the primary or contributory cause in someone with a lethal disease.

Cardiovascular disease is the most common reason for death in the water. A pre-existing condition may be exacerbated either by the physical exertion of swimming or struggling, or by the effects of cold. The common belief that swimming after a heavy meal is dangerous probably has some basis in fact, because of circulatory changes such as the 'splanchnic shunt'. When an acute cardiac episode occurs, such as sudden myocardial insufficiency or an arrhythmia, death may follow either from the cardiac lesion itself, or from drowning because of sudden physical incapacity.

INJURIES SUSTAINED BEFORE ENTERING THE WATER

The disposal of a murdered corpse in water is by no means uncommon, as it provides a convenient means of concealing the body, of transporting it from the scene of the crime by means of currents in river or sea, and – if discovery is delayed – of preventing identity by post-mortem decay and damage. Occasionally, accidental or suicidal injuries may be inflicted on a person before he or she enters the water. In marine, air or (rarely) road transport accidents, the victims may be injured or killed before entering the water. Persons falling from docks, bridges or ships may strike masonry or solid obstructions before hitting the water. Suicides may suffer in the same way or in exceptional circumstances, may stab themselves or cut their throats in a position where they fall into water immediately afterwards.

Whatever the means, the difficulties for the pathologist are the same, in that he must try to differentiate them from injuries sustained in the water before death occurs – and from post-mortem injuries. Often this is an impossible task, as such injuries are sustained perimortally, so that the time span involved is far too short for any of the usual criteria of ante-mortem infliction to be applied.

When the nature of the injuries is such that they could not have occurred in water, the issue is clear. Burns, missile wounds, the effect of explosions (other than compression injury) and patterned injuries may be incompatible with infliction in water, and must therefore be related to the effects of a previous assault or some accident on board ship or aircraft.

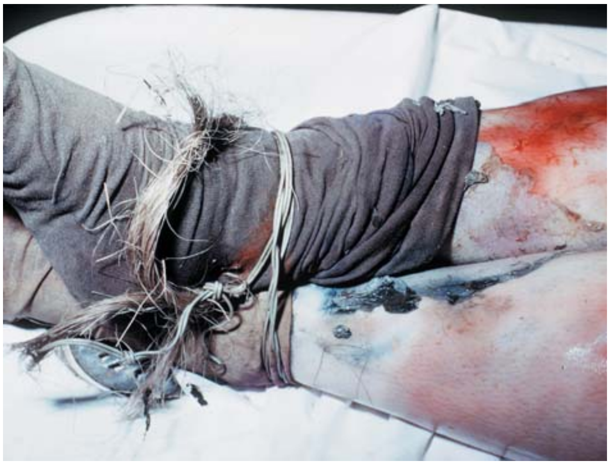

FIGURE 16.5 *Tying of the legs in a body removed from the water is suspicious, but not absolute proof of foul play. Some suicides who are good swimmers deliberately tie their legs to prevent instinctive self-preservation. In this illustration, however, the bonds were used to secure a heavy weight after homicide.*

INJURIES SUSTAINED IN THE WATER

As stated, this category is often impossible to differentiate from injuries inflicted before entering the water. Trauma in the water is common, as both an ante-mortem and a post-mortem phenomenon. Before drowning or another mode of death can supervene, the victim may be washed by waves or current against any kind of obstruction, from bridge pier to wharf, from rocks to weirs. Contact with the rough bottom of a river or a stony beach is another common cause of trauma. Damage from boats, ship's propellers, and marine or animal predators is frequent.

The issue usually becomes one of attempting to distinguish injuries sustained during life from those after death. Water complicates the problem by washing away surface bleeding from open injuries. A laceration inflicted during life will, however, usually show some bleeding into the tissues under the margins of the wound, at least until post-mortem decomposition blurs the appearances. Many decomposed bodies from water reveal green or black areas under the skin, especially in the scalp. It can be difficult or impossible to decide whether these are areas of altered blood, or merely autolytic changes. Even histological examination is often unhelpful in this respect. A similar problem exists with bruises in a decomposed body, which are often mimicked by discoloured areas of putrefaction. Post-mortem damage from animal predators is common in water and is discussed in the chapter on post-mortem changes.

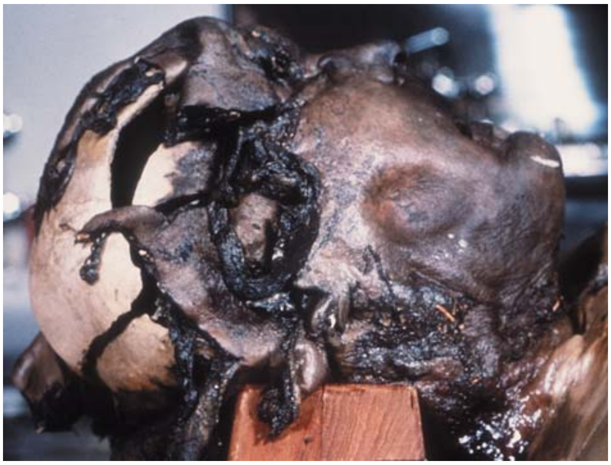

FIGURE 16.6 *Decomposed head and fractured skull of a person, who disappeared in the autumn before the river was covered with ice. The injuries have possibly been caused post-mortem by the ice packs when the body surfaced in the spring.*

In cold climates, if a body enters the cold water in late autumn, the gas formation due to putrefaction may be so slow that the sea, lake, or river will be covered with ice before the body surfaces and this takes place first in the spring when the ice is melting. In rivers, additional injuries may occur if the body gets pinned down between moving ice blocks.

Internally, there is no particular problem distinct from the examination of injured bodies in general, until post-mortem decomposition begins to obscure the appearances. This is more common in immersed bodies because of the frequent prolonged delays before the body is discovered. Bodies that have been suspended in shallow water may bear post-mortem damage from being dragged against the bottom. A common posture is with the head, hands and knees hanging downwards, so that these areas (especially the backs of the hands) may suffer extensive abrasions against a rough stone or gravel bottom.

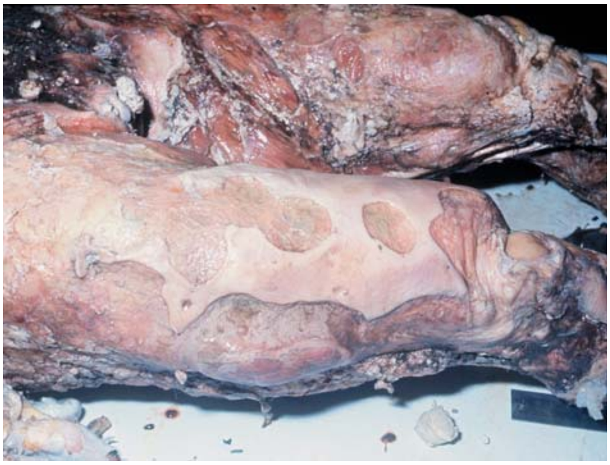

FIGURE 16.7 *Marine predator injuries on a body from the sea. The circular defects in the skin are probably made by crustaceans such as crabs.*

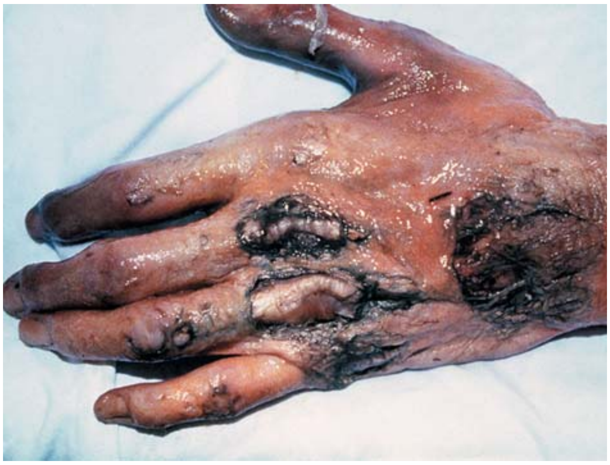

FIGURE 16.8 *Post-mortem injury to the back of the hand in a drowned person. Bodies usually hang suspended in the water with the back upwards and the knees, head and arms hanging down. In shallow water these areas may be dragged against the bottom, causing injuries similar to those shown here.*

Damage from the propellers of ships and smaller boats is common in commercial waterways and where pleasure craft abound. The screw of a large vessel may inflict serious damage – even dismemberment – on a body, but more characteristic is that caused by a small, high-speed propeller, such as an outboard motor or the out-drive of a motor cruiser. These injuries are becoming more common with the increase in water-skiing, and the number of other small pleasure boats on rivers, lakes and off beaches. The characteristic lesions are parallel wounds set a few centimetres apart, often of a sliced, undercut incised nature, usually on the head or back.

DEATHS FROM IMMERSION OTHER THAN DROWNING

Irrespective of natural disease or injury, some persons who die after falling into water do not drown, in the accepted physiological sense. The great problem for the pathologist is that, even in true drowning, there may be no autopsy signs, especially if any appreciable delay has occurred before recovery or autopsy, or both. Therefore it may be difficult or impossible for him to say whether a death is a true drowning or one from the non-drowning mechanisms. All that can then be offered is that death was due to 'immersion' – and even then such a diagnosis is usually one reached by exclusion of natural disease, trauma or toxic conditions, using hearsay circumstantial evidence to arrive at a pathological diagnosis. This is why drowning and immersion deaths present one of the most difficult problems for the forensic practitioner.

In certain circumstances, death may be extremely rapid after falling into water, there being insufficient time for drowning to occur. The typical occasion is familiar to any forensic pathologist with a harbour or dockside within his area. A seaman (often the worse for drink) returns to his ship at night, misses the gangway and falls into the water. The night-watchman immediately raises the alarm and the victim is pulled from the water within a minute or two, but is dead. At autopsy, none of the typical signs of drowning are present.

Other witnessed accidents of a similar nature are relatively common, in which rapid, almost instantaneous death must be presumed, in circumstances where true drowning can be excluded by history and lack of autopsy findings. The mechanism must be attributed to a reflex cardiac arrest, of a type similar to 'vasovagal inhibition', which is discussed elsewhere. Powerful stimulation of sensitive nerve endings can induce sudden overactivity of the parasympathetic system, inducing

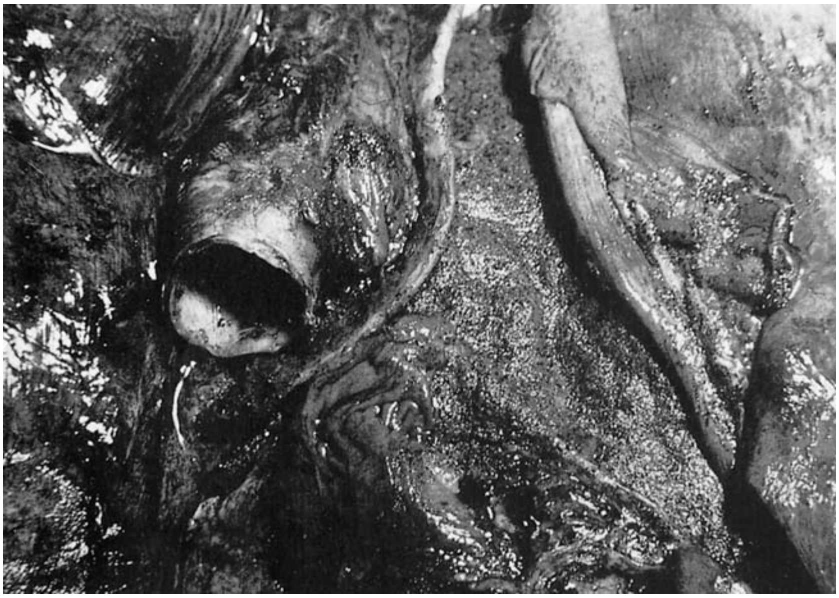

FIGURE 16.9 *The bifurcation of the trachea and main bronchi at autopsy on a body recovered from a beach. The air passages are full of sand, but this is not proof of vital aspiration of water, as foreign material can enter the bronchi by passive percolation after death.*

bradycardia and triggering cardiac arrest via the tenth cranial nerve nucleus and its vagal outflow. The afferent arc of this reflex varies from case to case, but is probably either:

- sudden immersion in cold water, causing intense stimulation of cutaneous nerve endings. Alcohol is known to potentiate such effects, perhaps by the general vasodilation of skin vessels – which would make the contrast with cold water all the more severe – and possibly by some central effects on the vasomotor centre
- sudden entry of cold water into the pharynx and larynx, and perhaps nasal passages can produce powerful stimulation of nerve endings in the mucosa. A bolus of water entering the trachea can also cause reflex cardiac arrest before there is time for airway insufficiency to have any effect.

As with any reflex cardiac arrest, there will be nothing to find at autopsy, a situation that forces this diagnosis to be speculative and based on the circumstances and exclusion of other conditions. Another mechanism that is often postulated as a cause of non-drowning immersion death is 'laryngeal spasm', leading to a hypoxic death from closure of the airway. The evidence for such a condition is tenuous, as such closure would have to operate for many minutes for hypoxia to kill, all the time keeping the larynx closed to prevent entry of water. Cyanosis is an uncommon feature of immersion deaths and in its absence it is difficult to see how laryngeal spasm could be differentiated from 'dry-lung drowning' or vasovagal cardiac arrest, the latter being a more reasonable diagnosis in those found dead after rapid recovery from the water.

Deaths in the bathtub

Fatalities in the bathroom are common, from a variety of causes, some unconnected with immersion. For example, Schmidt and Madea (1995) reported no fewer than 215 deaths in bathtubs from their Cologne Forensic Institute in 13 years, including 11 homicides.

Many suicides take place in the bathroom, from slashing of the throat or wrists, in or out of the bath. The bathroom cabinet or cupboard commonly holds razors for cutting and medicinal poisons for overdose. Electricity is more likely to be fatal in wet surroundings and a number of self-killings from electricity occur in the bathroom. Accidents include falls on slippery, wet floors or bath lining, with hard ceramic surfaces to cause head injuries.

Where gas appliances are used to heat water, the risk of carbon monoxide toxicity from faulty equipment is greater, as is also the case with electrical appliances such as hair-dryers, heaters and other devices. Although most countries have strict regulations about the electrical supply in bathrooms, people often defeat these safety measures by various means. Natural deaths from many causes may occur in the bathroom, but epileptics having unsupervised baths may have a fit and fall or drown.

Even homicide may be committed in bathrooms, by almost any means, but also from electricity and forcible immersion. The notorious multiple murders of 'The Brides in the Bath' perpetrated by George Joseph Smith early this century were probably accomplished by sudden pulling of the feet of the three women, so that their heads slid under the water. There is now a considerable literature about death in the bathtub, and all deaths in these circumstances must be investigated with particular care.

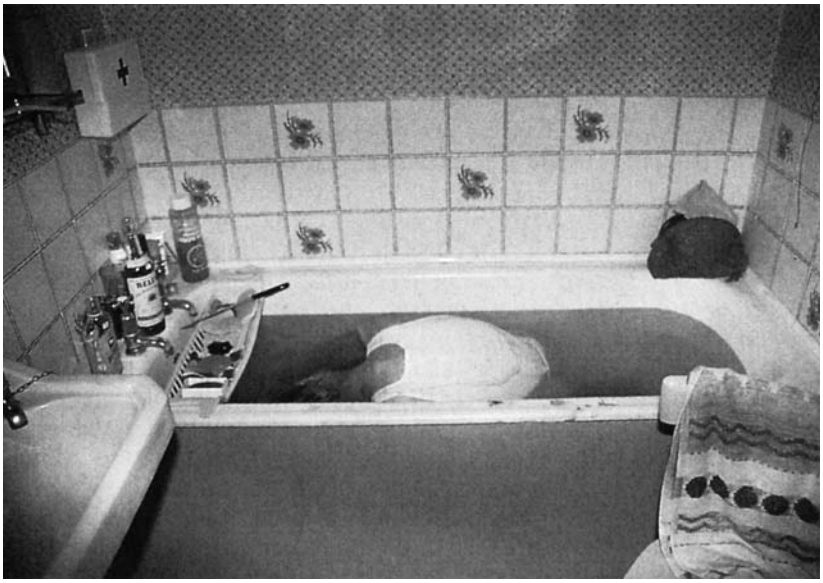

FIGURE 16.10 *Death in the bath is not uncommon and always requires careful investigation. In this case, the water is blood-stained, due to a suicidal cut throat after committing murder.*

DEATH FROM DROWNING

During the past five decades, views on the mechanism of drowning have undergone a radical change and perhaps the pendulum of opinion has swung a little too far. Until the researches stimulated by the Second World War, when large numbers of seamen and airmen were exposed to the risks of cold water, the classic view of drowning was that the victim was plunged into an airless medium and died of lack of oxygen. Following the experiments of Swann and many others (see below and references and further reading), it was recognized that, in animals, fluid and electrolyte disturbances had a major role in causing acute myocardial failure. The hypoxic element was relegated to such a secondary role that some commentators on drowning seemed to ignore the fact that the victims may have suffered 10 or more minutes without drawing a single breath of air. This criticism by no means disputes the undoubted role of fluid and electrolyte changes in drowning, but the effects of hypoxia must be considered in parallel when discussing the pathophysiology of drowning. The biochemical findings in the experimental animals did not seem to be matched by constant changes in human subjects and, in delayed and near-drowning victims, the electrolyte changes in aspiration have been shown to differ considerably from those in experimental animals.

A number of investigators long before Swann showed that blood and plasma were radically affected by drowning. In 1902, Carrara used specific gravity and other methods to demonstrate marked dilution of left heart blood in freshwater drowning and its absence in seawater, work repeated by Planck the following year with similar density techniques. Freimuth in 1955, however, used such specific gravity methods and found similar differences in non-drowning deaths.

The classical work of Swann and his colleagues (Swann *et al.* 1947; Swann and Spafford 1951) and others was conducted on conscious dogs imprisoned in cages and immersed totally under water – which, as pointed out by Crosfill (1956), does not accurately reflect human drowning, in which submersion is intermittent at first, prolonging the process and altering the physiological effects. In summary, the findings of Swann were that in freshwater drowning there was a massive absorption of water through the alveolar membranes, which could amount to 70 per cent of the original blood volume within 3 minutes. This haemodilution caused a relative anaemia and myocardial hypoxia, as well as rapid overloading of the cardiac capacity by hypervolaemia. It was also suggested that haemolysis of red cells released potassium, a powerful myocardial toxin – yet in dogs, unlike man, sodium is the main electrolyte in the erythrocyte. In fact, because of the hypoxia and haemodilution, there was a fall in plasma potassium and sodium, which led to early ventricular fibrillation in dogs that did not occur in seawater. Such fibrillation is said to be rare in man (Rivers *et al.* 1970).

In seawater the hypertonic medium caused a withdrawal of water from the plasma into the lungs and a rise in plasma sodium concentration. This was less deleterious to heart function and helped to explain the longer survival time in seawater immersion.

Pulmonary oedema developed in both types of aqueous medium but was slight and inconstant and, where abnormal, rapidly returned to normal. Death in freshwater typically took some 4–5 minutes, whilst seawater drowning was delayed for 8–12 minutes. Occasional instances have been reported of remarkably long periods of total submersion in the sea, always in cold water where oxygen uptake is reduced: these periods may exceed 20 minutes, and at least one authenticated survival exceeded 40 minutes.

Modern commentators on the researches of Swann, such as Giamonna and Modell (1967), have pointed out the inconsistencies between animal and human responses, and emphasized the role of hypoxaemia as a result of interference with the blood–air interface in the alveoli. This once again indicates the dangers of transposing the results of animal experiments to man and poses the question of how many dogs died in vain.

THE AUTOPSY SIGNS OF DROWNING

At autopsy much depends on the delay, both in recovering the body from the water and in starting the post-mortem examination. Where a considerable period has elapsed between death and retrieval of the body, the positive signs of drowning progressively fade. Depending on temperature, a period of several days in the water is likely to minimize most changes and, once significant putrefaction has occurred, then (apart from the controversial diatom tests) there is little hope of finding positive signs of drowning.

Even when a fresh body is recovered, delay in performing the autopsy reduces the amount of evidence. The author (BK) has seen several drowned bodies soon after retrieval in which a large plume of froth has been present at the mouth, yet when commencing the autopsy several hours later, such froth had vanished and was present only in the deeper air-passages.

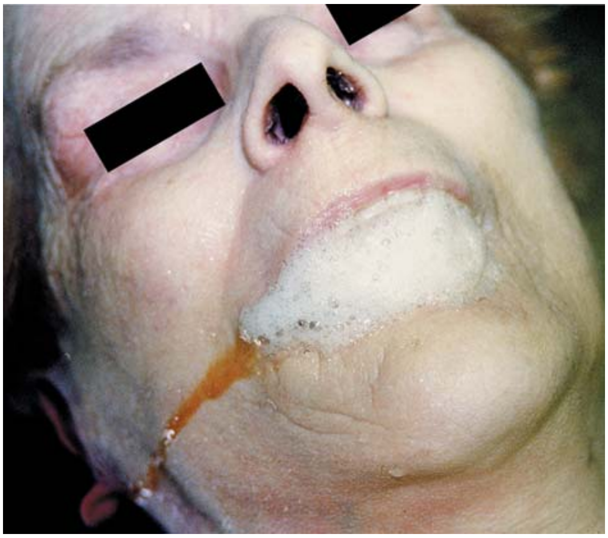

FIGURE 16.11 *Froth exuding from the mouth of a drowned person. Froth disappears with increasing post-mortem interval, though it persists in the internal air-passages for a much longer time.*

Froth in the air passages

The positive signs of drowning, as opposed to mere immersion, are scanty and not absolutely specific. The most useful, has just been mentioned – the presence of frothy fluid in the air passages – which in fresh bodies, often exudes though the mouth and nostrils, sometimes in the form of a plume. The froth is oedema fluid from the lungs, and consists of a proteinaceous exudate and surfactant mixed with the water of the drowning medium. It is usually white, but may be pink or red-tinged, because of slight admixture with blood from intrapulmonary bleeding. It is similar in appearance to the oedema of left ventricular failure that is commonly seen in deaths from cardiac disease such as hypertension, though some drownings exhibit a column of

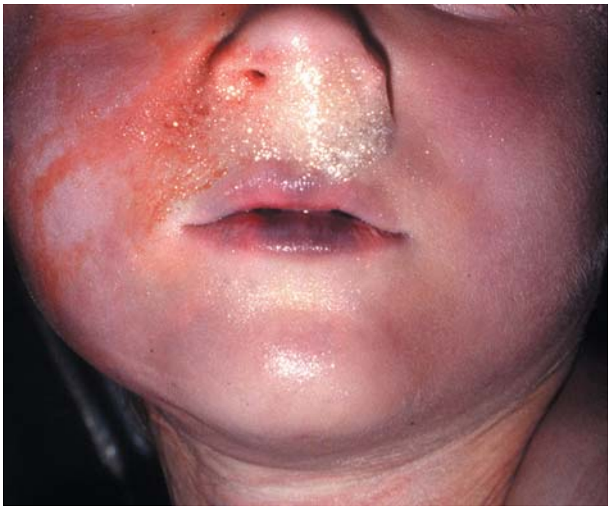

FIGURE 16.12 *Red-tinged froth exuding from the nostrils of a drowned person.*

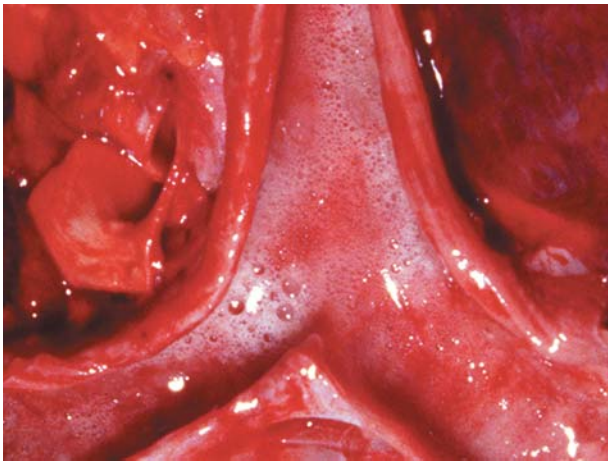

FIGURE 16.13 *Frothy fluid in the trachea and main bronchi of a person who was first strangled and subsequently fell or was thrown into sea and drowned.*

froth that is rarely, if ever, seen in natural pulmonary oedema. The froth extends into the trachea, main bronchi and smaller air passages. The lungs themselves in such cases will inevitably be waterlogged, and frothy fluid will exude from the bronchi when the lung is squeezed and from the cut surfaces when they are sectioned with a knife.

Unfortunately, although such oedema can be taken as diagnostic when combined with the appropriate circumstances and lack of other lethal pathology, the absence of waterlogging by no means excludes true drowning. The so-called ‘dry-lung drowning’ is not uncommon, in which the lungs appear normal in all respects, presumably because all the aspirated water has been absorbed through the alveolar walls into the plasma.

It has been suggested that this is more likely to occur if cardiac arrest takes place after removal from the water or if laryngeal spasm supervenes to prevent further water entry, so that continued circulatory function is able to clear away intra-alveolar fluid into the plasma. This theory, however, often fails to conform to the available facts of the drowning episode.

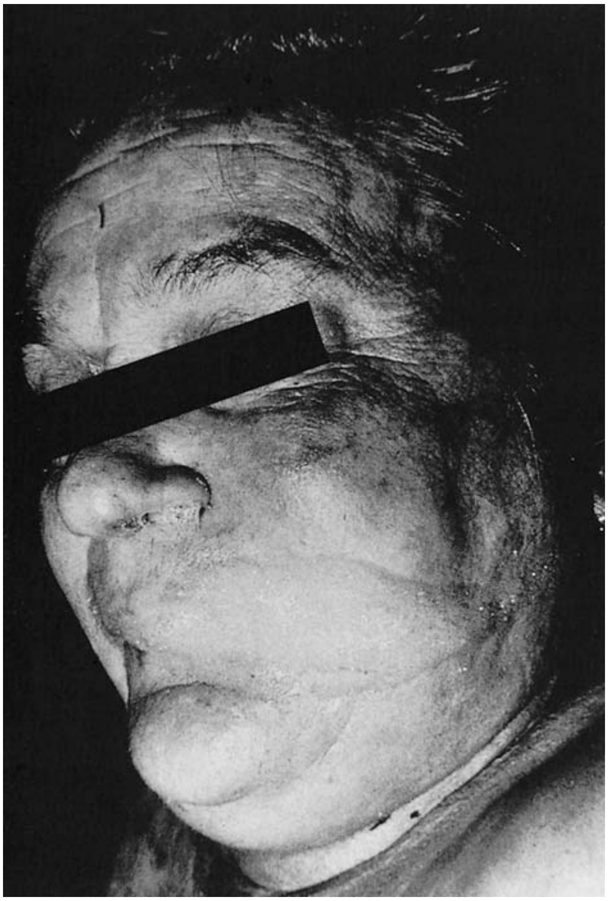

FIGURE 16.14 *Not all froth around the mouth is from drowning. This lady died in bed from hypertensive heart disease with profuse pulmonary oedema.*

Copeland (1985) measured lung weight in non-drowning deaths and deaths from salt-water and freshwater drowning. There was no significant difference in the weights of the last two, though 10–20 per cent of undoubted drownings were ‘dry-lung’ cases with no excess weight. Generally, the weights of a lung in drowning were around 600–700 g, whilst the non-drowned were in the 370–540 g range – though there was a considerable overlap in the two groups. The lung weights were also studied by Kringsholm *et al.* (1991) who claimed that the duration of submersion is related to lung weight. They found that 7 per cent of cases were ‘dry-lung drownings’ with a combined lung weight of less than a kilogram. In the remainder, the average weight of both lungs was 1411 g, compared to 994 g in controls. This was where the body was in the water less than 24 hours. After this, lung weight decreased, but pleural transudate increased. By combining the weight of lungs and transudate, more than 75 per cent of the cases lay between 1000 and 2200 g, up to 30 days.

Overinflation of the lungs

Apart from generalized waterlogging, the lungs may be markedly overinflated, filling the thoracic cavity when the sternum is removed. The normally bare area over the heart may be covered and the lungs may bulge upwards, meeting in the midline to obliterate the anterior mediastinum. The texture is rather pale and crepitant, superficially resembling that shown in asthma – for the same reasons. The older name for this condition is ‘emphysema aquosum’. True pre-existing emphysema may partially mimic the distension of drowning, but in the latter, bullae are never present. The oedema fluid in the bronchi blocks the passive collapse that normally occurs at death, holding the lungs in the inspiratory position. In addition, there is often an element of over-distension caused by the valvular action of the bronchial obstruction – again similar to asthma. In drowning this may be sufficient to mark the lateral surfaces of the lungs with the impression of the ribs, leaving visible and palpable grooves after removal of the organs from the thorax. This may be one of the most valuable positive signs of drowning to be gained at autopsy.

There may be some areas of intrapulmonary haemorrhage, which give the red tinge to the oedema fluid. These areas are seldom large or intense, and the general oedema and distension tend to minimize their prominence. Some may lie near the pleural surface and can be seen as rather blurred blotches on the exterior of the lung, the blurring being partly the result of spreading haemolysis. There are almost never subpleural petechial haemorrhages in drowning, apart from the few non-specific spots that can be found in the fissures and around the hilum in virtually every autopsy, no matter what the cause of death.

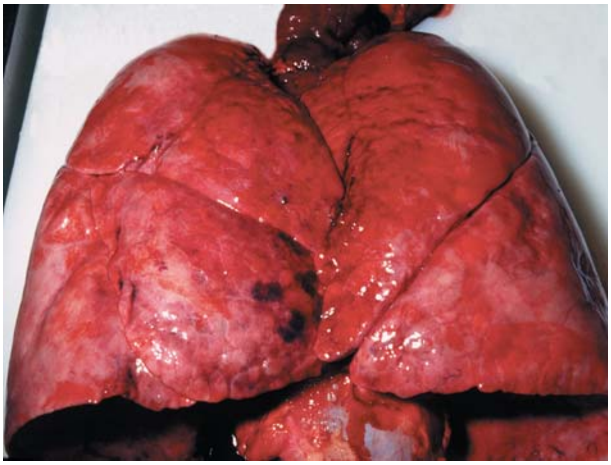

FIGURE 16.15 *Overinflation of the lungs (emphysema aquosum) in drowning with overlapping of the anterior margins and haemorrhages in the right middle lobe.*

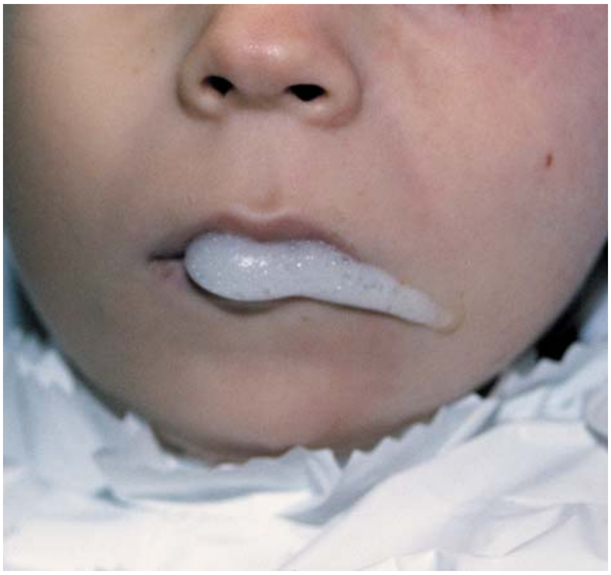

FIGURE 16.16 *A typical plume of froth from the mouth of a recently drowned boy. This will vanish as the post-mortem interval extends.*

Other organs in drowning

There are no other reliable autopsy changes in drowning. The heart and great veins have often been said to be 'dilated and engorged with fluid blood, especially the right side', but this is such a subjective and non-specific finding as to be useless, meriting no further consideration.

Many pathologists have for centuries claimed that the blood remains more fluid in drowning, and the physiological researchers of this century have confirmed the haemodilution in freshwater drowning. As a gross sign at autopsy, however, it is totally subjective and unreliable.

The stomach may contain watery fluid or even foreign material from the water, such as silt, weed or sand, but this cannot be accepted as a positive aid to the diagnosis. Many undoubted drownings show no water in the stomach, yet other cases in which a dead body was immersed reveal copious water in the stomach. The limiting factor is the tone in the oesophagus and cardiac sphincter, not the drowning process. The same applies to foreign materials in the trachea. Haemorrhage into the middle ears has been postulated as a positive sign of drowning, but in the author's (BK) own experience, it is totally unreliable. It can occur in any type of death where drowning can be definitely excluded, yet in classic instances of drowning, such bleeding may be absent. It has been said to be related to the depth to which the body sinks because of the hydrostatic pressure within the ear, but even this does not accord with the experience of most pathologists (Niles 1963).

CHEMICAL CHANGES IN THE BLOOD IN DROWNING

Because of the marked haemodilution that occurs in freshwater drowning and the electrolyte shifts in salt-water drowning, it is reasonable to expect that chemical analyses of the plasma should provide reliable evidence of drowning. Unfortunately, these theoretical hopes have not been realized in practice, mainly because of the biochemical chaos that occurs soon after death from any cause, in which the ability of live cell membranes to partition fluid and electrolytes is rapidly lost. Gettler (1921) published results of analyses of chloride content from the left and right sides of the heart, suggesting that the haemodilution from freshwater drowning would differentially reduce the plasma concentration on the left side, a difference of 25 mg/100 ml being significant. Salt-water drowning was said to produce the opposite effect. Gettler's claims have often been refuted and they are now no longer accepted. It has been shown that post-mortem changes in blood chloride concentrations occur irrespective of drowning and that the rate of change may be different on each side of the heart. Even the differentiation between freshwater and seawater drowning was discredited by the work of Durlacher *et al.* (1953), who found no reliable changes in sodium, potassium and chloride concentration. Similar results were found by Fuller (1963).

Electrolyte measurements in surviving victims of near-drowning showed no regular pattern, according to Rivers *et al.* (1970). Moritz (1944) attempted to use magnesium

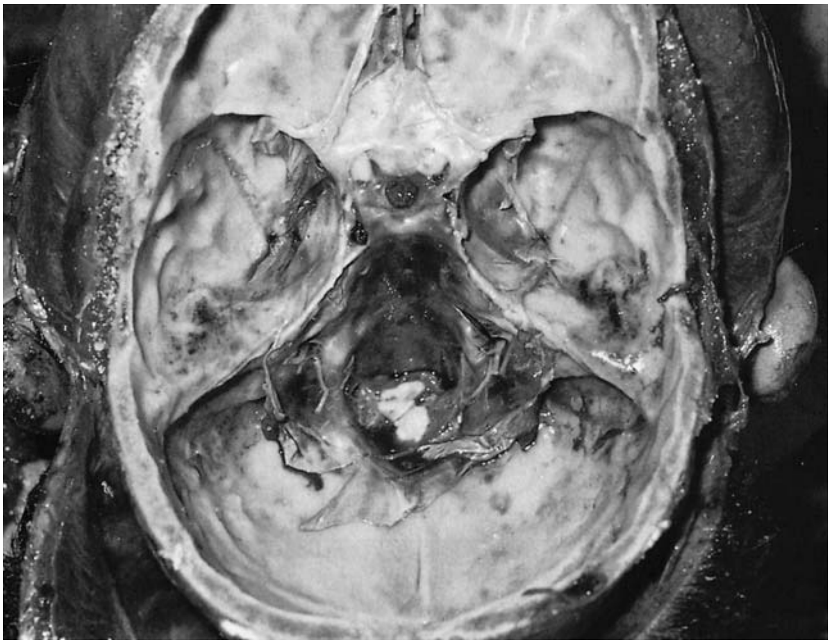

FIGURE 16.17 *Bilateral haemorrhages within the petrous temporal bones in drowning. Though claims have been made for the usefulness of this observation, its occurrence seems random, sometimes being present in non-drowning and absent in proven drowning.*

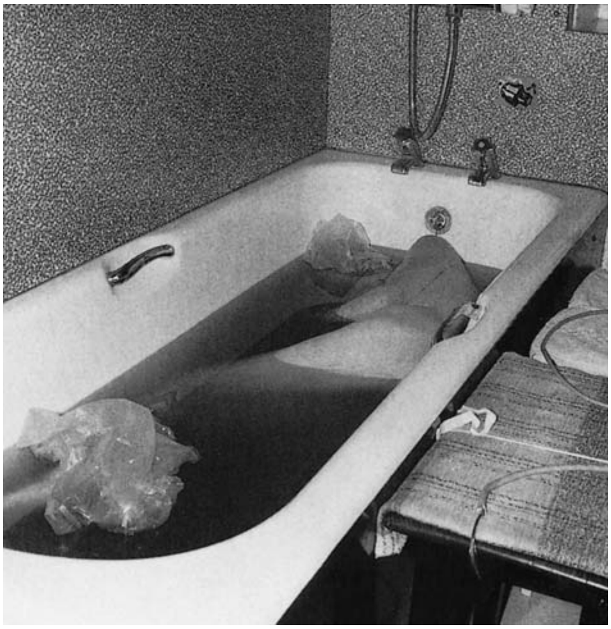

FIGURE 16.18 *Suicide in the bath. The head was enveloped in a plastic bag and an (unused) electrocution device was also present, illustrating the frequent presence of alternative methods in suicide.*

as a similar marker in drowning, but the unreliability of such analyses has prevented their being used as a routine procedure. Once again, the hazards of extrapolating the results of animal experiments to humans is highlighted by these results, as even other mammals like dogs have quite different responses to the same environmental changes.

Other chemical substances have been used as markers of drowning, such as strontium. Azparren *et al.* (1994) claim that differences in the strontium concentration of blood from left and right heart are always greater than 75 μg/l in seawater drowning.

HISTOLOGICAL CHANGES IN DROWNING

Much has been written about both light and electron microscopy of the lungs in immersion deaths, especially in continental Europe. The accounts are confusing, however, and sometimes contradictory, the consensus of opinion being that such changes are inconstant and unreliable.

Much of the work has tried to confirm histologically the genuineness of 'emphysema aquosum' in both fresh and decayed bodies, but the dilatation and rupture of the terminal air spaces can occur not only as a result of drowning, but of passive immersion in water deeper than 4 m. Dilatation of the alveoli, thinning of the walls and compression of the capillaries are easy to observe, but the significance is always ambiguous. Reh (1976) classified such lungs on the basis of reticulin fibres, but Heinen and Dotzauer (1973) showed similar changes in undrowned material. Janssen (1984) has reviewed the subject and concludes that though histological evidence may be helpful and indicative, it is never probative.

The number of macrophages in the alveoli, adjusted to the size of the alveoli where distended, was studied by Betz *et al.* (1993), who claimed a diagnostic usefulness of the method as long as no autolytic changes were present.

The weight of the spleen has been used by Haffner *et al.* (1994) as a diagnostic marker in drowning. Compared with other modes of 'asphyxia', about an 18 per cent reduction in weights and spleen–liver ratios was recorded, with a suggestion that stress, hypoxia, alcohol and cooling were responsible.

The skin changes in maceration are also well recorded, and some have been claimed to have use in dating the period of immersion (Okros 1938; Dierkes 1938). Schleyer (1951) is pessimistic on the reliability of such histological criteria.

The early formation of adipocere can be detected microscopically after 3–4 weeks; this is dealt with in the chapter on post-mortem changes.

DIATOMS AND THE DIAGNOSIS OF DROWNING

Few topics in forensic pathology have given rise to so much argument as the use of diatoms in the diagnosis of drowning. Revenstorf in 1904 was the first to attempt to use diatoms as a test for drowning, though he stated that Hofmann in 1896 was the first to discover them in lung fluid. An excellent review of the diatom controversy was published by Peabody in 1980.

The basic premise is that when a live person is drowned in water containing diatoms (microscopic algae with a silicaceous exoskeleton), many diatoms will penetrate the alveolar walls and be carried to distant target organs such as brain, kidney, liver, and bone marrow. After autopsy, samples of these organs can be digested with strong acid to dissolve the soft tissue, thus leaving the highly resistant diatom skeletons to be identified under the microscope. Alternatively, blood can be used to seek the organisms.

When a dead body is deposited in water or when death in the water is not due to drowning, then, although diatoms may reach the lungs by passive percolation, the absence of a beating heart prevents circulation of diatoms to distant organs.

The great advantage of the diatom test, if it was reliable, would be that a positive diagnosis of drowning could be

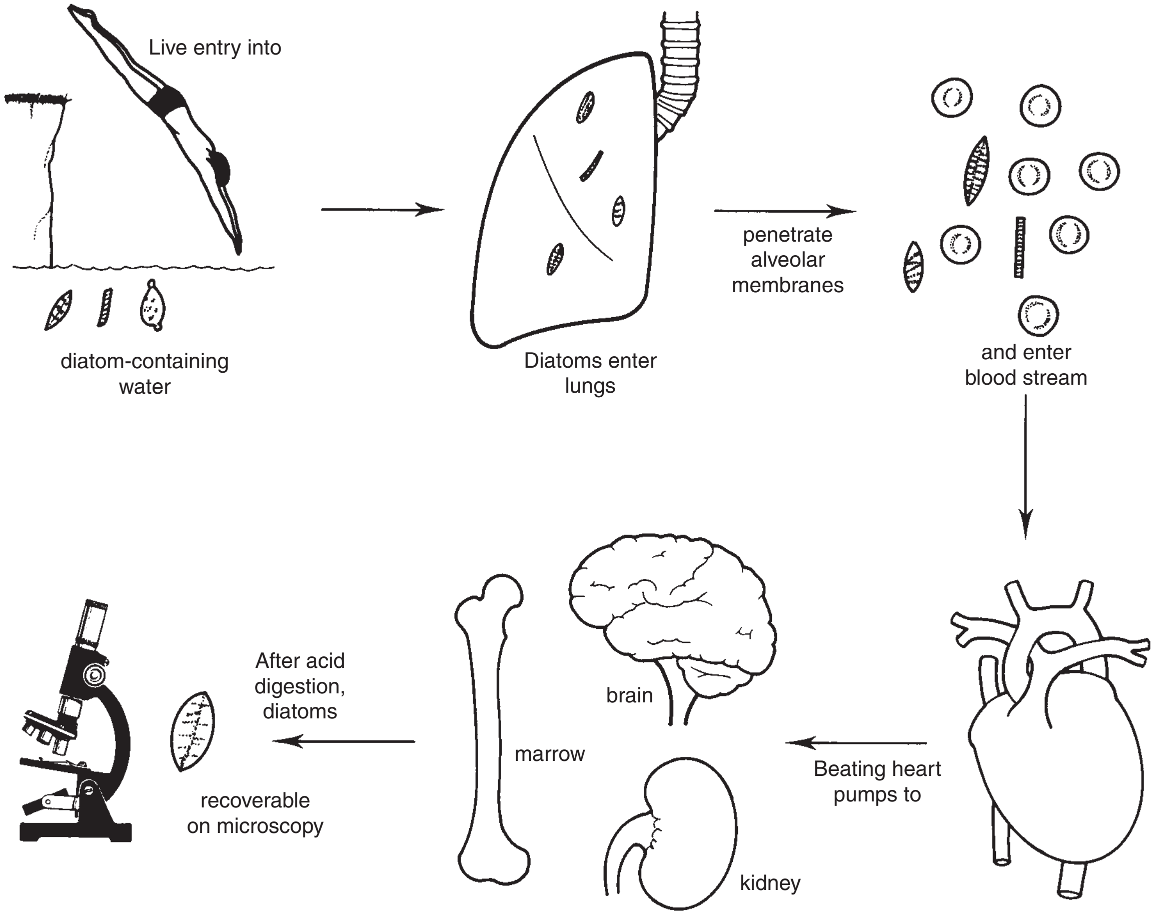

FIGURE 16.19 *The principle of the diatom test for drowning. When a dead body enters the water, diatoms cannot progress further than the lungs.*

made even in the frequently putrefied bodies that are recovered, where no hope of anatomical recognition of drowning is possible. In addition, evidence of the site of the drowning might be obtained from the ecological typing of the diatoms, especially in relation to salt-water or freshwater locations.

Meticulous attention to technique must be employed at autopsy to recover suitable samples of tissue, which must not be contaminated by the body surface or instruments, before digestion is commenced. (Technical details are given in Appendix 1.)

Diatoms belong to the class of plants known as *Diatomaceae* and consist of a box or 'frustule' composed of two valves that fit together to enclose the cytoplasmic contents. They have either radial symmetry ('centric' diatoms) or are elongated ('pennate' diatoms). The valves are highly complex in shape and are extremely resistant to decay, fossil diatoms from Jurassic times being abundant. The floor of the oceans have an area of 11 million square miles of thick diatomaceous ooze and upraised beds on land may be hundreds of feet thick, which emphasizes the vast amount of potential contaminant available.

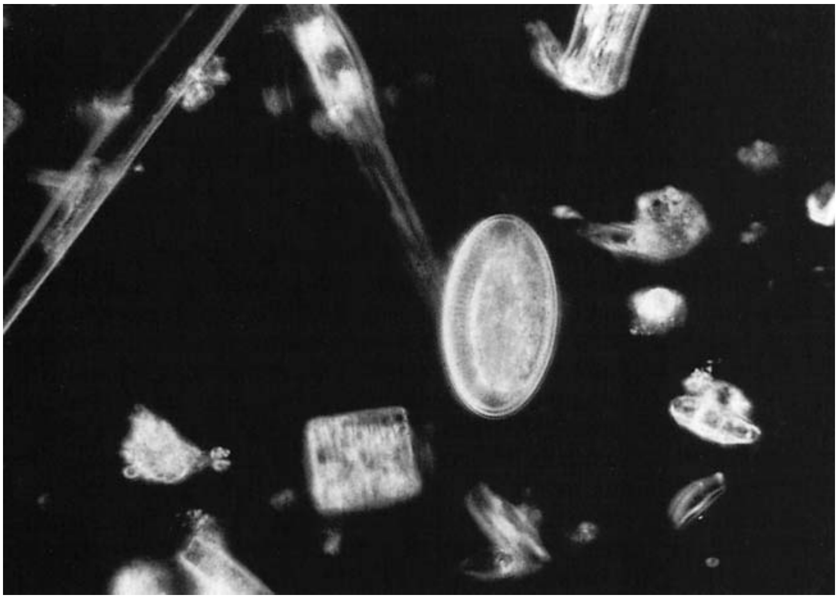

FIGURE 16.20 *Diatoms seen under dark-ground illumination (original magnification ×350). These are from fresh water. There are at least 10 000 different types and, where identification of locus of likely origin is required, the help of a biologist with expert knowledge is essential.*

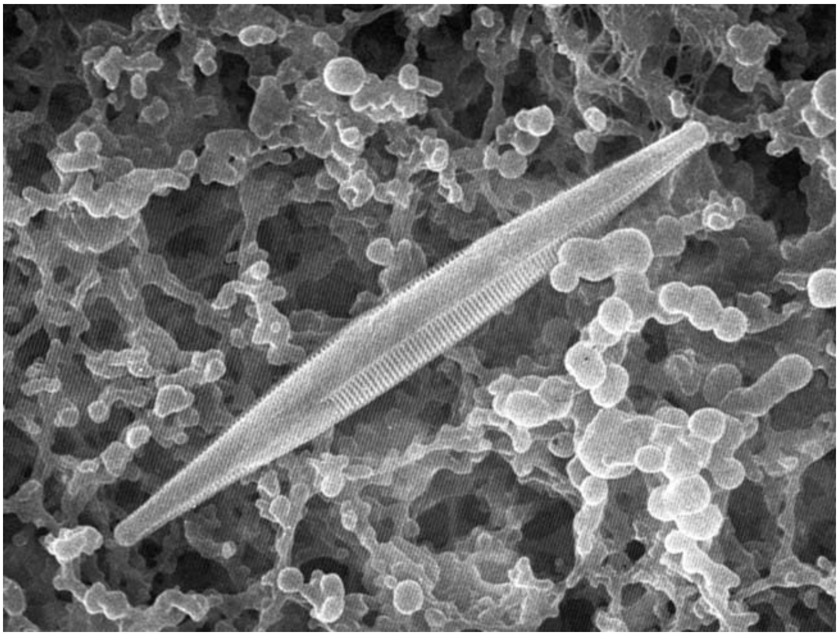

FIGURE 16.21 *Scanning electron micrograph of a diatom (original magnification ×4000, 20 kV).*

There are at least 10 000 species, and the identification of the various types is a matter for an experienced botanist or biologist, but the following general classification is useful: 'oligohalophilic' diatoms live in freshwater with a salinity less than 0.05 per cent, and 'mesohalophilic' and 'polyhalophilic' diatoms live in brackish water and seawater with a salinity higher than 0.05 per cent.

The critics of the diatom test point to certain problems:

- Diatoms are ubiquitous, being present in soil, water supplies and in the air. Many commonly used minerals, such as kieselguhr, are largely fossil diatoms. Manufactured substances like board-chalk and abrasives – even toothpaste – contain diatomaceous raw materials. Mant, however, examined the filters of the air-conditioning system of Guy's Hospital in London and found no diatoms taken from the city atmosphere.
- Though entry into the body is thought to be mainly through the lungs, there seems no reason why they cannot penetrate the intestinal lining, and gain access to the bloodstream and hence any body tissues.
- Certain foods, notably shellfish, contain vast quantities of diatoms that may enter the circulation and reach the tissues.
- Examination of the tissues of both animals and man reveal the presence of diatoms when the cause of death is other than drowning. Schellmann and Sperl (1979) found them in the tissues of 15 out of 16 non-drowned bodies, but other workers found no diatoms in 33 out of 34 bodies, including two drownings.

Foged (1983) made a detailed investigation into drowned and non-drowned bodies in Denmark, and concluded that the diatom test was quite invalid. He quoted many references both for and against the reliability of the technique, and no doubt the controversy will continue. It seems probable that there may be a quantitative difference between the numbers of diatoms recovered from the tissues in drowning and non-drowning deaths, and a careful analysis of the species identification in relation to the locus and circumstances of the death may be useful. At the present time the diatom test should be used only as an indicative aid and not as legal proof of drowning. Recently, other water organisms have been advocated as tests for drowning, including soft-bodied protozoa and crustaceans.

Japanese workers have used other methods to digest autopsy tissues to seek diatoms because it has been shown that strong acid digestion markedly reduces the yield. The new methods introduced for diatoms, such as enzymatic and detergent digestion, allowed soft-bodied microscopic organisms to be recovered and claims for better specificity for drowning have been made (Matsumoto and Fukui 1993; Funayama *et al.* 1987, 2001). Chinese workers have claimed that the spectrofluorophotometric detection in the lungs of chlorophyll derived from plankton is diagnostic, but it is not clear how this can be differentiated from plankton passively entering the lungs after death (Qu and Wang 1992).

REFERENCES AND FURTHER READING

Auer A, Mottonen M. 1988. Diatoms and drowning. *Z Rechtsmed* **101**:87–98.

Azparren J, de la Rosa I, Sancho M. 1994. Biventricular measurement of blood strontium in real cases of drowning. *Forensic Sci Int* **69**:139–48.

Betz P, Nerlich A, Penning R, *et al.* 1993. Alveolar macrophages and the diagnosis of drowning. *Forensic Sci Int* **62**:217–24.

Bhaskar GR. 1965. Diatoms in cases of drowning. *J Indian Acad Forensic Sci* **4**:2–6.

Bonilla-Santiago J, Fill WL. 1978. Sand aspiration in drowning and near drowning. *Radiology* **128**:301–2.

Bonte W, Sprung R, Huckenbeck W. 1986. [Problems in the evaluation of electrocution fatalities in the bathtub.] Probleme bei der Beurteilungen von Stromtodesfällen in der Badewanne. *Z Rechtsmed* **97**:7–9.

Bray M. 1985. Chemical estimation of fresh water immersion intervals. *Am J Forensic Med Pathol* **6**:133–9.

Budnick LD. 1984. Bathtub-related electrocutions in the United States, 1979 to 1982. *JAMA* **252**:918–20.

Budnick LD, Ross DA. 1985. Bathtub-related drownings in the United States, 1979–81. *Am J Public Health* **75**:630–3.

Byard RW, Lipsett J. 1999. Drowning deaths in toddlers and preambulatory children in South Australia. *Am J Forensic Med Pathol* **20**:328–32.

Carrara M. 1902. Electrolytes and osmotic pressure in drowning. *Vrtlschr Gericht Med* **24**:236–40.

Chiaraviglio E, del Carmen, Wolf AV. 1963. Diagnosis of drowning. *Arch Pathol* **75**:337.

Copeland AR. 1985. An assessment of lung weights in drowning cases. The metro Dade County experience from 1978 to 1982. *Am J Forensic Med Pathol* **6**:301–4.

Copeland AR. 1986. Homicidal drowning. *Forensic Sci Int* **31**:247–52.

Copeland AR. 1987. Suicide by drowning. *Am J Forensic Med Pathol* **8**:18–22.

Coutselinis A, Boukis D. 1976. The estimation of Mg^{2+} concentration in cerebrospinal fluid (C.S.F.) as a method of drowning diagnosis in sea water. *Forensic Sci* **7**:109–11.

Crosfill JW. 1956. Mechanism of drowning. *Proc R Soc Med* **49**:1051–4.

Davis JH. 1961. Fatal underwater breath holding in trained swimmers. *J Forensic Sci* **6**:391–6.

Davis JH. 1986. Bodies found in the water. An investigative approach. *Am J Forensic Med Pathol* **7**:291–7.

Devos C, Timperman J, Piette M. 1985. Deaths in the bath. *Med Sci Law* **25**:189–200.

Dierkes K. 1938. Über die Histologie der Waschhaut. *Dtsch Z Gesamte Gerichtl Med* **30**:262.

Durlacher S, Freimuth H, Swan H. 1953. Blood changes in man following drowning. *Arch Pathol* **56**:454–60.

Durwald W. 1964. Elektrotodesfälle in der Badewanne. *Arch Kriminol* **134**:164–71.

Foged N. 1983. Diatoms and drowning – once more. *Forensic Sci Int* **21**:153–9.

Fornes P, Pepin G, Heudes D, *et al.* 1998. Diagnosis of drowning by combined computer-assisted histomorphometry of lungs with blood strontium determination. *J Forensic Sci* **43**:772–6.

Foroughi E. 1971. Serum changes in drowning. *J Forensic Sci* **16**:269–82.

Fukui Y, Hata M, Takahashi S, *et al.* 1980. A new method for detecting diatoms in human organs. *Forensic Sci Int* **16**:67–74.

Fuller RH. 1963. The clinical pathology of human near-drowning. *Proc R Soc Med* **56**:33–8.

Funayama M, Aoki Y, Sebetan IM, *et al.* 1987. Detection of diatoms in blood by a combination of membrane filtering and chemical digestion. *Forensic Sci Int* **34**:175–82.

Funayama M, Mimasaka S, Nata M, *et al.* 2001. Diatom numbers around the continental shelf break. *Am J Forensic Med Pathol* **22**:236–8.

Gardner E. 1944. Death in the bathroom. *Med Leg Criminol Rev* **12**:180–95.

Geertinger P, Voigt J. 1970. Death in the bath. A survey of bathtub deaths in Copenhagen, Denmark, and Gothenburg, Sweden, from 1961 to 1969. *J Forensic Med* **17**:136–47.

Gettler AO. 1921. A method for the determination of death by drowning. *JAMA* **77**:1650–2.

Giammona ST, Modell JH. 1967. Drowning by total immersion. Effects on pulmonary surfactant of distilled water, isotonic saline, and sea water. *Am J Dis Child* **114**:612–16.

Goldhahn RT. 1977. *Scuba diving deaths; a review and approach for the pathologist.* In: Wecht C (ed.), *Legal medicine annual 1976*. Appleton Century Crofts, New York.

Gooden BA. 1972. Drowning and the diving reflex in man. *Med J Aust* **59**:583–7.

Gordon I. 1972. The anatomical signs in drowning. A critical evaluation. *Forensic Sci* **1**:389–95.

Gruspier KL, Pollanen MS. 2000. Limbs found in water: investigation using anthropological analysis and the diatom test. *Forensic Sci Int* **112**:1–9.

Haffner HT, Graw M, Erdelkamp J. 1994. Spleen findings in drowning. *Forensic Sci Int* **66**:95–104.

Heinen M, Dotzauer G. 1973. [The lung of a drowned person.] *Beitr Gerichtl Med* **30**:133–41.

Hendey NI. 1973. The diagnostic value of diatoms in cases of drowning. *Med Sci Law* **13**:23–34.

Holden HS, Crosfill JW. 1955. The significance of foreign bodies in the alveoli of the apparently drowned. *J Forensic Med* **2**:141–50.

Hurlimann J, Feer P, Elber F, *et al.* 2000. Diatom detection in the diagnosis of death by drowning. *Int J Legal Med* **114**:6–14.

Jääskeläinen A. 1967. Diatomenbefunde in Wasserleichen. *Dtsch Z Gesamte Gerichtl Med* **61**:41–7.

Janssen W. 1984. *Forensic histopathology.* Springer, Heidelberg, Germany.

Karkola K, Neittaanmaki H. 1981. Diagnosis of drowning by investigation of left heart blood. *Forensic Sci Int* **18**:149–53.

Knight B. 1981. Immersion or drowning? *Br Med J* **282**:1340–1.

Kobayashi M, Yamada Y, Zhang WD, *et al.* 1993. Novel detection of plankton from lung tissue by enzymatic digestion method. *Forensic Sci Int* **60**:81–90.

Koseki T. 1969. Investigations on the bone marrow as a material in the diatom method diagnosing of death from drowning. *Acta Med Biol* **16**:88–90.

Kringsholm B, Filskov A, Kock K. 1991. Autopsied cases of drowning in Denmark 1987–1989. *Forensic Sci Int* **52**:85–92.

Kvittingen TD, Naess A. 1963. Recovery from drowning in fresh water. *Br Med J* **1**:1315–16.

Lucas J, Goldfeder LB, Gill JR. 2002. Bodies found in the waterways of New York City. *J Forensic Sci* **47**:137–41.

Ludes B, Coste M, North N, *et al.* 1999. Diatom analysis in victim's tissues as an indicator of the site of drowning. *Int J Legal Med* **112**:163–6.

Matsumoto H, Fukui Y. 1993. A simple method for diatom detection in drowning. *Forensic Sci Int* **60**:91–5.

Mikami Y. 1959. Experimental study and practice on the detection of vegetable planktons in the bone marrow of the drowned dead body. *Acta Med Okayama* **13**:259–67.

Modell JH. 1968. The pathophysiology and treatment of drowning. *Acta Anaesthesiol Scand Suppl* **29**:263–79.

Modell JH. 1971. *Drowning and near-drowning.* Thomas, Springfield, USA.

Modell JH. 1978. Biology of drowning. *Annu Rev Med* **29**:1–8.

Modell JH, Davis JH. 1969. Electrolyte changes in human drowning victims. *Anesthesiology* **30**:414–20.

Monantay NK, Rao RD. 1964. The diagnosis of death from drowning with particular reference to 'diatom method'. *J Indian Acad Forensic Sci* **3**:21–7.

Moritz AR. 1944. Chemical methods for the determination of death by drowning. *Physiol Rev* **24**:70–88.

Mueller WF. 1969. Pathology of temporal bone hemorrhage in drowning. *J Forensic Sci* **14**:327–36.

Niles NR. 1963. Haemorrhage in the middle ear and mastoid in drowning. *Am J Clin Pathol* **40**:281–4.

Ohlsson K, Beckman N. 1964. Drowning – reflections based on two cases. *Acta Chir Scand* **128**:327–31.

Okros S. 1938. Annähernde Bestimmung der Todeszeit aus dem Hautzustand. *Dtsch Z Gesamte Gerichtl Med* **29**:497.

Pachar JV, Cameron JM. 1993. The diagnosis of drowning by quantitative and qualitative diatom analysis. *Med Sci Law* **33**:291–9.

Peabody AJ. 1980. Diatoms and drowning – a review. *Med Sci Law* **20**:254–61.

Peabody AJ. 1986. *Drowning cases: a protocol for their submission*. CRE, Aldermaston.

Pearn J. 1985. Pathophysiology of drowning. *Med J Aust* **142**:586–8.

Pearn J, Nixon J. 1977. Bathtub immersion accidents involving children. *Med J Aust* **64**:211–13.

Pearn JH, Brown JD, Wong R, *et al.* 1979. Bathtub drownings: report of seven cases. *Pediatrics* **64**:68–70.

Pollanen MS. 1998. Diatoms and homicide. *Forensic Sci Int* **91**:29–34.

Polson CJ. 1962. *Essentials of forensic medicine.* Pergamon Press, Oxford.

Qu J, Wang E. 1992. A study on the diagnosis of drowning by examination of lung chlorophyll(a) of planktons with a spectrofluorophotometer. *Forensic Sci Int* **53**:149–55.

Reh H. 1976. [About the problem of 'lungs in drowning' (author's transl).] *Z Rechtsmed* **77**:219–21.

Rivers J, Orr G, Lee H. 1970. Electrolytes in near-drowning. *Br Med J* **2**:157–60.

Rushton DG. 1961. Drowning – a review. *Med Leg J* **29**:90–8.

Schellmann B, Sperl W. 1979. [Detection of diatoms in bone marrow (femur) of nondrowned (author's transl).] *Z Rechtsmed* **83**:319–24.

Schleyer F. 1951. Zur Histologie der Waschhaut. *Dtsch Z Gesamte Gerichtl Med* **40**:680.

Schmidt P, Madea B. 1995. Homicide in the bathtub. *Forensic Sci Int* **72**:135–46.

Schmidt P, Madea B. 1995. Death in the bathtub involving children. *Forensic Sci Int* **72**:147–55.

Schneider V. 1985. [Electrocution in the bathtub.] *Arch Kriminol* **176**:89–95.

Schwar TG. 1972. Drowning: its chemical diagnosis. A review. *Forensic Sci* **1**:411–17.

Schwerd W, Lautenbach L. 1960. Mord mit elektrischem Strom in der Badewanne. *Arch Kriminol* **54**:126–49.

Sidari L, Di Nunno N, Costantinides F, *et al.* 1999. Diatom test with soluene-350 to diagnose drowning in sea water. *Forensic Sci Int* **103**:61–5.

Spitz W, Schneider V. 1964. The significance of diatoms in the diagnosis of death by drowning. *J Forensic Sci* **9**:11–18.

Sturner WQ, Balko A, Sullivan J. 1976. Magnesium and other electrolytes in bovine eyeballs immersed in sea water and other fluids. *Forensic Sci* **8**:139–50.

Swann HG. 1956. Mechanism of circulatory failure in fresh water and sea water drowning. *Circ Res* **4**:241–2.

Swann HG, Spafford NR. 1951. Body salt and water changes during fresh and sea water drowning. *Tex Rep Biol Med* **9**:356–62.

Swann HG, Brucer M, Moore C, *et al.* 1947. Fresh water and sea water drowning: a study of terminal cardiac and biochemical events. *Tex Rep Biol Med* **5**:423–37.

Terazawa K, Takatori T. 1980. Isolation of intact plankton from drowning lung tissue by centrifugation in a colloidal silica gradient. *Forensic Sci Int* **16**:63–6.

Thomas F, Van Hecke W, Timperman J. 1961. The detection of diatoms in the bone marrow. *J Forensic Med* **8**:142–4.

Thomas F, Van Hecke W, Timperman J. 1962. [Medicolegal diagnosis of death by submersion by demonstration of diatoms in the marrow of the long bones.] *Ann Med Leg (Paris)* **42**:369–73.

Timperman J. 1962. The detection of diatoms in the marrow of the sternum. *J Forensic Med* **9**:134–6.

Timperman J. 1969. Medico-legal problems in death by drowning. Its diagnosis by the diatom method. A study based on investigations carried out in Ghent over a period of 10 years. *J Forensic Med* **16**:45–75.

Timperman J. 1972. The diagnosis of drowning. A review. *Forensic Sci* **1**:397–409.

Trubner K, Püschel K. 1991. [Fatalities in the bathtub.] Todesfälle in der Badewanne. *Arch Kriminol* **188**:35–46.

CHAPTER 17

Neglect, starvation and hypothermia

- General features of neglect and starvation 412
- Medico-legal problems 414
- Injury caused by cold: hypothermia 414
- The 'hide-and-die' syndrome 418
- Deaths from exposure 419
- Local injury due to cold: frostbite 419
- References and further reading 420

Though, tragically, starvation is one of the most widespread scourges of our planet, we are here concerned with individual neglect with which starvation is often associated, either as a potentially criminal act or as a manifestation of self-neglect. As with so many conditions, the extremes of life are most affected, rather than the middle years. Infants and children are at the mercy of their parents or guardians for the provision of care and sustenance. The old are prey to senile dementias and other disorders of the mind, as well as often being devoid of spouse or caring relatives – or sometimes even the financial means to stay warm and adequately fed.

Many countries have legislation designed to protect children, on the basis that they are totally dependent upon adults for the necessities of life. In Britain, the Children Act places a responsibility upon parents, guardians or the community to care for children. The minimum requirements laid down by law are the provision of food, clothing, shelter and education. Unfortunately, these requirements are evaded in spite of official supervision and, from time to time, scandals explode in the public eye when a child has died, not from physical abuse, but from sheer neglect.

Voluntary organizations, such as the National Society for the Prevention of Cruelty to Children (NSPCC) were founded in the nineteenth century to combat the widespread neglect of children consequent upon the population explosion of the Victorian industrial revolution. Their present-day role, however, has changed radically so that now most of their efforts are directed against physical, sexual and mental abuse, rather than starvation and neglect. Notwithstanding this change of emphasis, deaths still occur from deliberate or negligent lack of provision of proper care and nutrition of children. In general – though there are many exceptions – nutritional neglect and physical battering tend to be mutually exclusive. The abused child is usually well nourished and the emaciated child is usually uninjured.

Neglect and starvation are not synonymous, but are so often closely associated that the two conditions are usually considered together. When fatal, they may come to the attention of the pathologist either as a potentially criminal death or, when self-inflicted, in the usual course of coroners' or similar enquiries. Rarely, victims will starve from being isolated in some catastrophe, such as a mining incarceration or shipwreck, but these days a more common occurrence is a hunger strike or voluntary fast for some political or similar motive.

GENERAL FEATURES OF NEGLECT AND STARVATION

The obvious and outstanding appearance is of emaciation caused by an inadequate intake of food. A minimum of about 1500–1900 calories is needed to maintain body weight in active adults. Life is threatened when more than 40 per cent of the original body weight has been lost, though the speed of loss is relevant. Total deprivation of food will obviously kill faster than a severe reduction. Total lack of food is likely to cause death in about 50–60 days as long as adequate water is available, but this depends on weather and temperature as well as the original fitness – and fatness – of the individual. Deprivation of water will kill in about 10 days – or less in higher ambient temperatures.

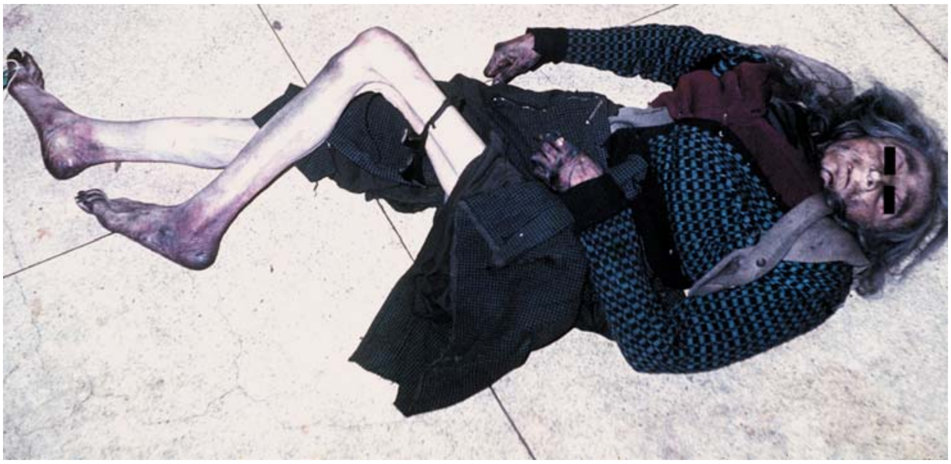

FIGURE 17.1 *Death from neglect and hypothermia in an old woman. The discoloration of the feet and hands was cyanotic.*

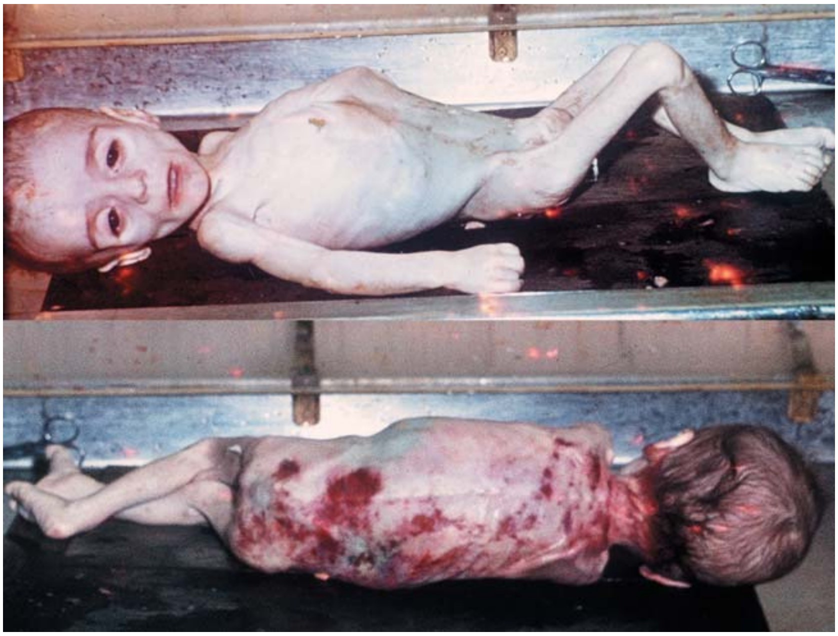

FIGURE 17.2 *Fatal neglect and starvation in an infant. The limbs are emaciated, the abdomen scaphoid and the head appears relatively large. The back shows extensive pressure ulceration. The sunken eyes and cheeks from malnutrition and dehydration constitute the 'Hippocratic facies'.*

In the numerous cases of fatal and non-fatal starvation seen at the end of the 1939–45 war in Nazi concentration camps, two main types were seen. These were the 'dry' type, in which there was emaciation, but only leg oedema, with a body weight up to half the normal. Alive, these showed marked hypotension, feeble pulse and cyanosis. The other 'wet' group had marked oedema of face, trunk and limbs, with ascites and pleural effusions. The oedema of hypoproteinaemia is caused by protein being diverted to produce energy when carbohydrate and fat are grossly deficient in the diet. In some victims, the skeleton accounted for 50 per cent of the total body weight, instead of the usual 15 per cent.

Other features are secondary to this nutritional deficiency, such as intercurrent infections, avitaminoses, skin disorders and nutritional oedema. Dehydration, hypothermia and actual necrosis of the extremities are the most severe manifestations that contribute to death.

Most fatally neglected infants are under one year of age. At autopsy great care must be taken to measure, not estimate, the full indices of size and weight. Crown–heel, crown–rump, head diameter, foot length and exact weight must be measured accurately, so that comparison may be made with standard paediatric percentile growth charts, consonant with age, birth weight, sex and race.

The diameter of limbs at major landmarks such as wrists and ankle should be measured with a flexible tape. Photographs should be taken in all major views, using colour to record skin tints and lesions, as well as the general aspect of the emaciated body. The words 'cachexia', 'emaciation' and 'marasmus' are all synonymous as used today, though 'marasmus' is confined to the description of infant victims. The so-called 'Hippocratic facies' is not confined to the emaciation of neglect, as cachexia from any cause – such as an obstructive carcinoma of the oesophagus – can give rise to the same appearance. The skin of the face is stretched tightly across the cheekbones, the cheeks are sunken and the lines of the jaw are obtrusive. The eyes are deeply sunken from loss of orbital fat, though dehydration may contribute to this manifestation.

On the trunk the ribs are prominent, with concavities in the intercostal spaces and sunken supraclavicular fossae. The abdomen is of the classical ‘scaphoid’ shape, being concave from costal margin to iliac crests, the latter sticking up like wings. The limbs are almost skeletal from loss of fat and muscle, and the head may appear deceptively large by contrast with the narrowed neck.

The skin surface may present differing appearances according to the length of time that nutritional loss has been suffered and the qualitative defect in the diet, especially of certain vitamins. The skin may be pale, lustrous and semi-translucent, with an almost blue tinge in infants, partly as a result of the loss of subcutaneous fat. It may also be coarse and rough, however, with flaking hyperkeratosis, though this is more often seen in aged persons, partly as a result of senile changes. Pigmentation, either diffuse or of a punctate nature, is sometimes seen.

Dehydration is a common feature, especially in infants and may be the major contribution to death. The skin is dry and wrinkled and, when pinched between the fingers, remains ridged because of loss of subcutaneous fat and fluid. The fontanelles may be depressed from loss of pressure of cerebrospinal fluid. In chronic starvation, hypoproteinaemia can lead to oedema, though this is rarely seen in forensic practice. In famines it may be widespread, associated with other nutritional disorders such as kwashiorkor.

Skin infections are common, partly as a result of nutritional loss of resistance to infections and partly because of neglect in care and hygiene, especially in infants. Pressure sores on buttocks, heels and spine may occur in both infants and old people who are lying inert because of weakness, and, in infants, urine dermatitis and lack of any skin care after defecation worsens these. Sores may be seen on any pressure or friction points, such as the elbows, knees, shoulders and occiput. Apart from pressure areas, any part of the skin surface may develop infected lesions, which often crust over after partial healing. Sores on the lips, blepharitis and conjunctivitis may also occur. The hair is dry and brittle, sometimes becoming depigmented and even gingery with chronic malnutrition.

Internally, the same loss of adipose tissue will be evident both subcutaneously and in the internal fat stores, such as the omentum, mesentery and perirenal area. The organs will be small (except the brain) and it is therefore important to weigh the major viscera to compare them with standard tables for sex and age. Organ atrophy will be more apparent in old people, but the problem then exists of distinguishing weight loss caused by nutritional defects from the usual senile atrophy. The differentiation is one of degree, as chronic starvation can reduce organs to unusually low weights. More specific internal signs will include an empty gut, which may be gas-filled and translucent from stomach to colon. Any faeces in the rectum are likely to be inspissated, especially if there has been fluid restriction and dehydration. Faecoliths may ulcerate the intestinal lining. In infant starvation, however, the stools may be fluid and offensive unless dehydration was severe. The gall bladder is distended in starvation if there has been no recent meal to stimulate emptying. The tongue may be thickly coated, but this is such a non-specific finding that it is of no practical use.

MEDICO-LEGAL PROBLEMS

In infants there may sometimes, if rarely, be physical abuse associated with neglect, so radiography prior to autopsy is essential – in fact the whole procedure must be conducted on the lines indicated for the ‘battered child’ (Chapter 22). In both children and adults, a major problem is cause and effect, especially where some pathological disease process is found. Did a cancer lead to cachexia or is it an incidental finding? Did pulmonary tuberculosis cause the emaciation or was the infection secondary to the nutritional deficiency? The relevance of diseases such as diabetes, Addison’s disease, chronic infections and neoplasms must be evaluated. Particularly important in children are familial metabolic diseases that may be associated with failure to thrive and frank emaciation.

Where death from a wasting condition is alleged to be the result of malnutrition, either from wilful withholding of food or from indifference about its provision, a defence may be raised against criminal prosecution that the condition was caused by a natural disease. Indeed, such a claim may be true on occasions. Both the prosecution and defence medical advisers must undertake a thorough investigation using all anatomical, histological and biochemical means at their disposal to discover any such metabolic or pathological cause for fatal wasting, before attributing the death to sheer malnutrition.

INJURY CAUSED BY COLD: HYPOTHERMIA

Until about 40 years ago, injury and death from the effects of low temperature was thought to occur almost exclusively among those subjected to extremes of climate out of doors. Mountaineers, polar explorers and those exposed in some disaster on sea or land were considered to be the prime candidates for frostbite and ‘freezing to death’. Only after some noteworthy medical publications, such as those by

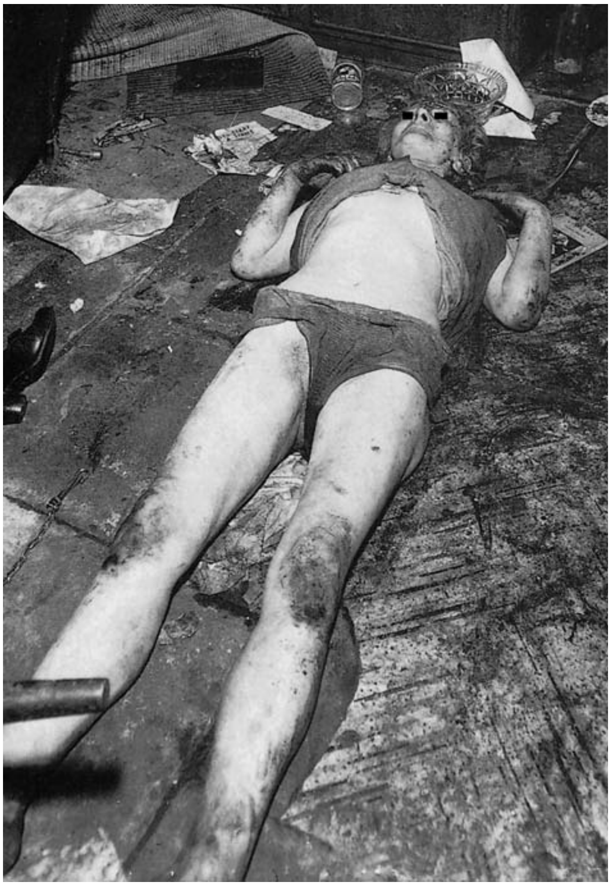

FIGURE 17.3 *Hypothermia in a woman of only 56. She lived in extreme squalor, with no heating and broken windows, even though she had been a state registered nurse. The knees, hands and feet showed hypothermic discoloration as well as dirt.*

Emslie-Smith (1958), Duguid *et al.* (1961) and others, was it generally appreciated that hypothermia was a common and widespread danger in temperate climates and indoors. Recognition was rapid and Special Committees of both the Royal College of Physicians and the British Medical Association produced reports on hypothermia in the 1960s.

As in so many other conditions, the aged and the infants are most vulnerable, especially as the condition is insidious and may not be suspected in time by those more active persons who are the potential guardians of those at the two extremes of life.

Factors involved in hypothermia

LOW ENVIRONMENTAL TEMPERATURE

No precise figures can be given in degrees Celsius for temperatures potentially dangerous to life, as other factors and the personal vulnerability of different individuals vary considerably. Certainly the air temperature need be nothing like as low as was thought in former years, when profound frost was considered essential for the production of lesions. Air temperatures below 10°C are probably low enough to cause hypothermia in vulnerable persons, but air movements such as external wind or internal draughts will make things worse by markedly increasing the rate of body cooling. Damp conditions, especially of clothing, will also aid cooling from the latent heat of evaporation.

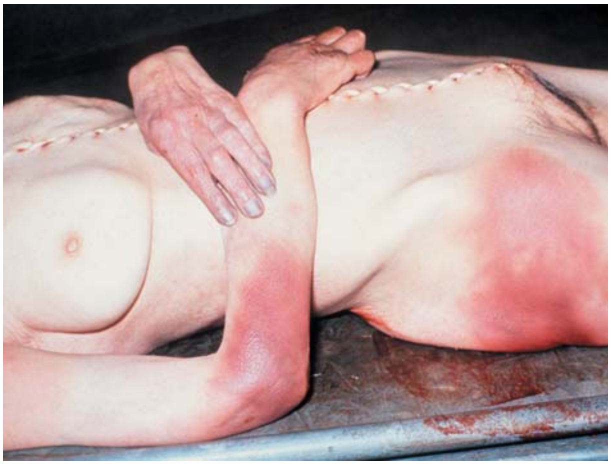

FIGURE 17.4 *A death from hypothermia showing brownish-pink discoloration over the hip and elbow.*

AGE AND PHYSIQUE

As stated, the elderly and the young are at greatest risk. The old person may have poor musculature and a lack of mobility, thus depressing heat production from muscular activity. Often thin, there may be a lack of insulating subcutaneous fat. Poor thyroid function is common, especially in old women, and is discussed below. Cerebral dysfunction – often from degenerative arterial disease – may lead to poor central heat regulation, a disinclination to move around, and apathy concerning personal nutrition, heating and general well-being. Depression and mental illness are not uncommon, as in the early cases reported by Emslie-Smith and by Duguid *et al.*

In infants, the body mass is small and the surface area:mass ratio large, so that heat is lost more easily. In small infants, there is naturally complete dependence on others in regard to clothing and environment.

HYPOTHYROIDISM

Myxoedema, either occult or obvious, is an important factor in hypothermia. Many elderly women have a thyroid deficiency and this is a potent factor in vulnerability to cold. The majority of victims are women over 70 and the environmental conditions in which thyroid dysfunction is a factor may be little below optimum. Cases were first reported in the 1950s by Marquand, soon followed by the paper of Angel and Sash (1960). In hypothyroid subjects, certain drugs may have a deleterious effect by precipitating hypothermia, even when the environment is relatively normal. Imipramine, chlorpromazine and diazepam are particularly hazardous. Barbiturates, phenothiazines and alcohol are also dangerous in this respect.

LACK OF FOOD, ADEQUATE CLOTHING AND INDOOR HEATING

These are factors that may be social or financial, but are often related to the depressed, apathetic state of many old people, especially those with cerebral atherosclerosis. Cause and effect are often hard to separate, as the effects of cold may bring on or worsen mental sluggishness and confusion.

The manifestations of hypothermia

The clinical manifestations are not the concern of the forensic pathologist, though when unsuspected before autopsy the history may give a clue to the fatal process.

It is sometimes claimed that hypothermia in the elderly never occurs unless there is some predisposing disability, such as cerebrovascular disease, hypothyroidism, alcoholism or diabetes, for instance. This seems too restrictive an attitude, though in general is probably true. The rectal temperature is used as the measure of severity in hypothermia. Between 37°C and 32°C there are no ill-effects beyond the normal subjective feeling of being cold, with shivering and constriction of blood vessels.

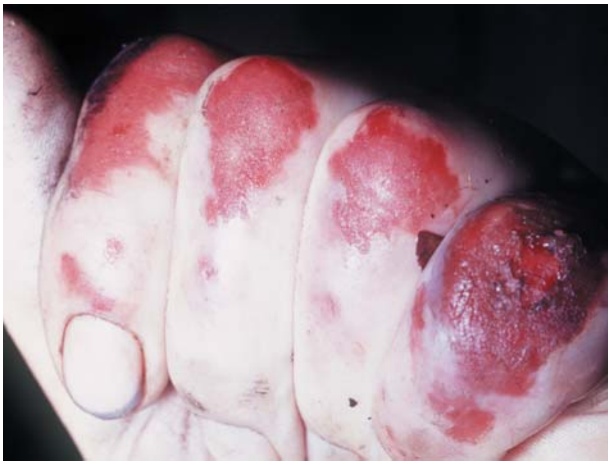

FIGURE 17.5 *Frostbite of the knuckles after being left unconscious during winter in Wales. It does not require Arctic temperatures to cause tissue damage or death.*

When the core temperature falls further below 32°C to 24°C, there is dulling of consciousness, a fall in respiration and heart rate, and a lowering of blood pressure. If maintained without relief, death will ensue and recovery is rare if the rectal temperature falls below about 26°C, though one remarkable case reported by Laufman (1951) recovered after falling to 18°C, albeit with severe mutilating frostbite.

Autopsy findings in hypothermia

There may be no signs whatsoever at autopsy in a death from hypothermia and thus the history may be all-important. The usual difficulties then arise as to how much the pathologist is entitled to rely upon hearsay evidence – as opposed to the objective findings of his own examination – when deciding on a cause of death. This is more fully discussed in the first chapter of this book.

Hypothermia, like some cases of epilepsy, asthma and drowning, provides some of the more difficult pathological dilemmas when there are slight, non-specific or absent autopsy findings. When a victim of hypothermia has been admitted to hospital and 'warmed-up', death may supervene at any time up to a few days later. In such circumstances the autopsy may reveal no positive findings. The skin changes will almost certainly be absent, but intra-abdominal signs may persist.

In a classical death from hypothermia the body will usually be that of an old person – almost half the Royal College of Physicians' series was over 65 years of age; most will be women. The exterior of the body may present suggestive

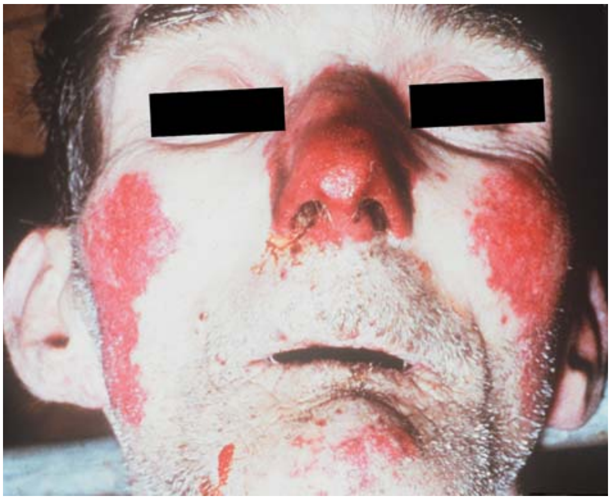

FIGURE 17.6 *Bright red discoloration of the cheeks and nose from hypothermia. The victim suffered from mitral stenosis (where malar flush is one of the clinical signs) and was in bed in an unheated room during a Welsh winter. The tissues of the nose were virtually necrotic, amounting to actual frostbite.*

signs of hypothermia, in that patches of pink to brownish-pink discoloration may be seen in certain areas. They are most often present over extensor surfaces and large joints, such as the outer sides of the hips, the elbows and the knees, and less often on the flanks and face.

The colour is similar to that seen in bodies exposed in cold water or in severe outdoor cold – and to the post-mortem pinkness that often develops in mortuary refrigerators. In hypothermia, however, the distribution over the large joints is distinctive, in addition to the pinkness of hypostasis. There may be pink patches on the cheeks, chin and nose, especially in victims who already had impaired cardiac function, such as the 'malar' flush of mitral stenosis. The slight brownish tinge often seen in the pinkness and the rather indefinite blurred edges of the patches suggest that some haemolysis may have occurred, perhaps in the early post-mortem period as well as during the terminal stages of life. The colour is presumably due to persistent oxyhaemoglobin in the skin capillaries, which, because of the low metabolic activity of the cold tissues, fail to take up the delivered oxygen. Pooling of blood in surface vessels from the breakdown of normal vasomotor control allows this unreduced haemoglobin to show through the skin. Why a normally coloured body placed after death in a mortuary refrigerator should also become pink has never been satisfactorily explained.

In addition to pink areas of skin, the extremities may be cyanosed or they may be white. Sometimes the feet are blue to the ankles, above which is pale. The hands may show blueness of the fingers or nail beds. Oedema may be seen, most often in the feet and lower legs, but many of the victims of hypothermia already have pre-existing congestive cardiac failure. Rarely there may be blistering of the skin, usually in dependent parts such as the buttocks, backs of thighs and

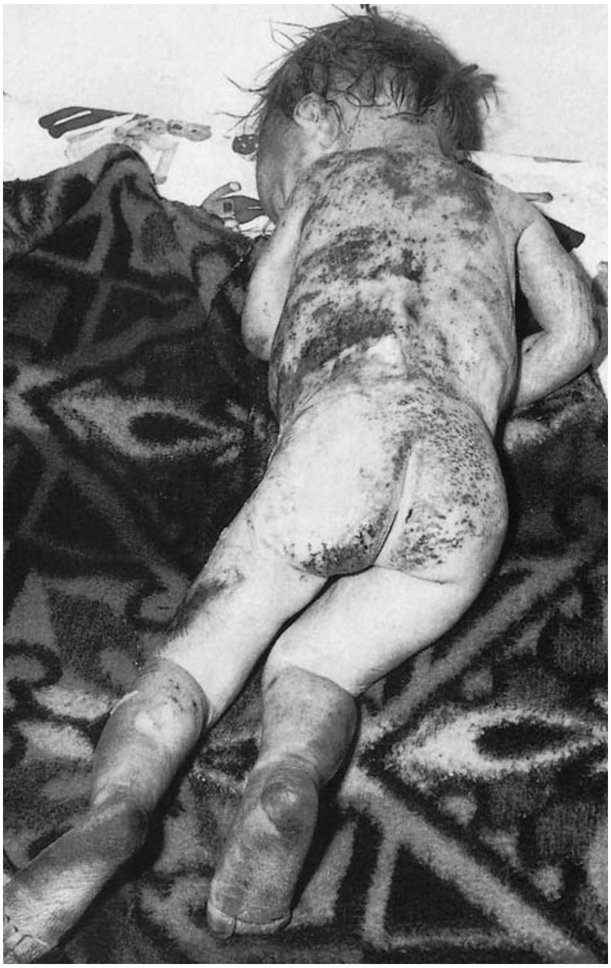

FIGURE 17.7 *Fatal child neglect. The infant has extensive skin ulceration and was underweight. Death was the result of an intercurrent chest infection.*

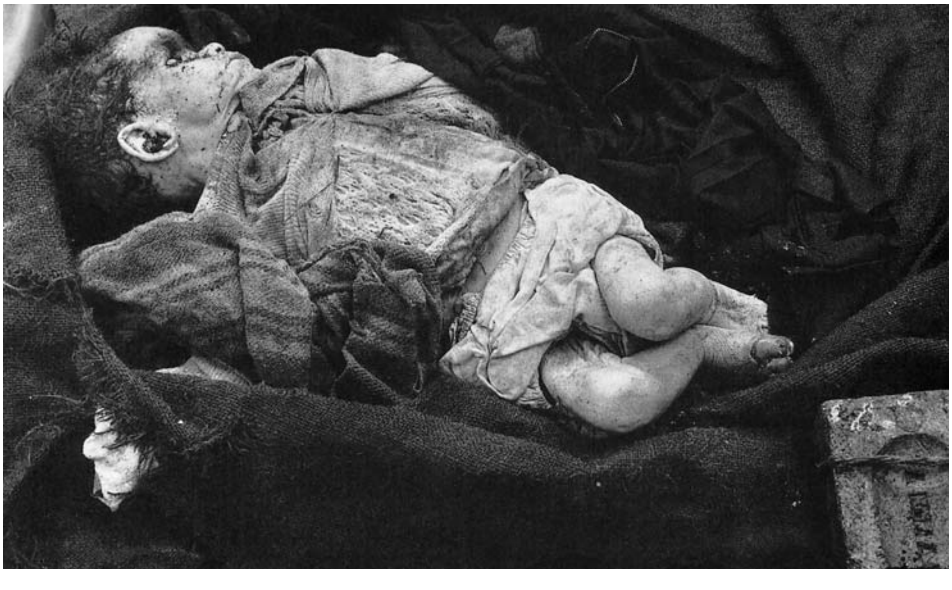

FIGURE 17.8 *This infant died of neglect, hypothermia and frostbite, having been left in an unheated farmhouse bedroom, the cardboard box that acted as a cot having crystals of frozen urine in the bedclothes. After death the child was dropped down the farm well in this sack with a brick as a weight.*

upper arms. This may or may not be in pink areas and seems to be caused by skin oedema which then breaks down.

The face, especially of an elderly woman, may show appearances suggestive of myxoedema or hypothyroidism, with coarsening of the features, puffy eyelids and loss of eyebrow hair. Internally there may be nothing specific to note, even if the external skin changes are obvious. Signs of pre-existing disease are common, especially if one subscribes to the view that hypothermia in the elderly never occurs in the absence of systemic disease. Degenerative arterial disease, senile myocardial degeneration, hypertension, chronic renal lesions and obstructive airways disease are common. More specific lesions are considered below.

ACUTE GASTRIC EROSIONS

In the authors' view, these are found much more often than pancreatitis. The stomach mucosa is frequently studded with numerous shallow ulcers, the floor of each containing a dark brown plug of altered blood. Frank haematemesis is rare, but the stomach contents sometimes contain dark acid-affected blood.

ACUTE PANCREATITIS

This is the lesion most often quoted in textbooks, but is present in less than half the deaths from proven hypothermia. It may have been diagnosed before death from an increase in serum amylase activity. The most obvious feature may be areas of stiff, yellow, fat necrosis when the pancreas is sectioned at autopsy. There may be adjacent areas of fat necrosis in the omentum and mesentery. Haemorrhage into the gland may occur, but caution has to be employed here, as post-mortem autolysis of this organ commonly gives a haemorrhagic appearance in many other modes of death seen at autopsy. Histologically, necrosis, leucocyte infiltration and fat necrosis may be confirmed in some instances, but in others the pancreas, even if haemorrhagic, yields little information on microscopy.

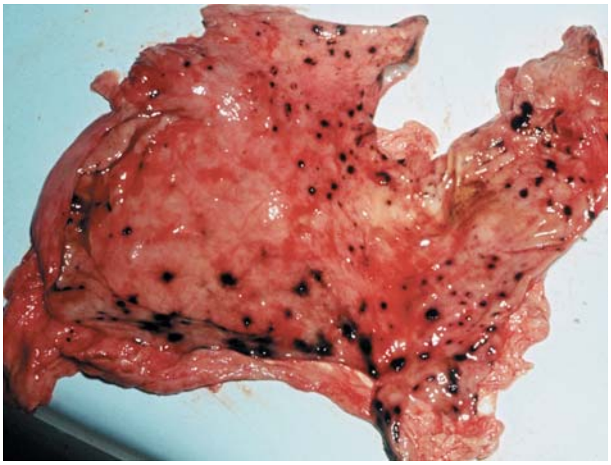

FIGURE 17.9 *Gastric mucosa from a hypothermia death showing numerous acute erosions.*

PULMONARY OEDEMA

This is common, but as it is such a common finding in any death from cardiac failure, its diagnostic value in hypothermia is almost nil.

PERIVASCULAR HAEMORRHAGES

These have been described in the brain, especially in the walls of the third ventricle, but again they are not particularly specific, especially in an autopsy population that has a high incidence of cerebrovascular disease. Sludging of blood in peripheral vessels may occur, partly from cold agglutinins. This contributes to skin oedema and necrosis, though peripheral vasoconstriction is the main cause of the infarction of frostbite. Microinfarcts are common in many organs in hypothermia, presumably from cold agglutinins blocking small vessels. Deep vein thrombosis in the legs and pulmonary embolism are occasionally found, but again cause and effect is hard to establish.

It is said that the good preservation of tissues when examined histologically is characteristic of hypothermia, as autolysis is retarded by the low temperature. In the experience of the authors, this claim is not justified.

THE 'HIDE-AND-DIE' SYNDROME

There is a peculiar aspect to some cases of hypothermic death in that it is associated with the victim undressing and hiding away from sight. The subject is usually an old person, man or woman. When found dead (usually, but not always at home), the body may be partly or even totally naked, even though the environment is cold. In addition, the victim may have burrowed his or her way into some corner or cupboard, or alternatively pulled down furniture and household articles into a heap on top of the body. In fact, the house may be in such a state of disorder that when the police break in and find a corpse amongst such chaos they may naturally suspect a homicide and robbery.

The author (BK) has seen old people who have crawled into small cupboards, the bottom of wardrobes, under low shelves, and into corners of pantries and outhouses. They have been found under piles of books pulled from a bookcase, or hidden under a mound of old clothes, newspapers and other debris.

The signs of hypothermia are usually present and the problem then arises as to whether the victim became hypothermic first, which led to mental confusion that caused the strange behaviour, or whether – as a result of some mental aberration, perhaps related to senility or other cerebral episode – they began behaving abnormally, including shedding their clothes and so became fatally hypothermic.

Though some geriatricians claim that hypothermia is almost always precipitated by some natural illness, including cerebral vascular episodes, the experience of most pathologists is that it is not common to find definite lesions of such predisposing factors at autopsy.

Biochemical markers of hypothermia have been investigated, and catecholamines appear in the blood and then urine in the early stages, but then decline, as the adrenals become exhausted (Hirvonen and Huttunen 1995). There is a variation in the adrenaline:noradrenaline ratio, all these changes being manifestations of stress induced by low temperature.

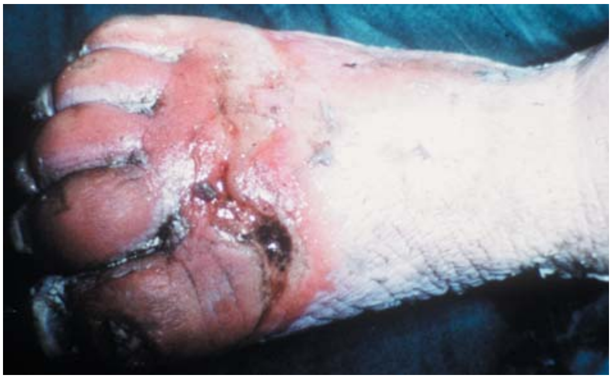

FIGURE 17.10 *Frostbite of the foot from the infant shown in Figure 17.8. The distal part of the foot is infarcted and a line of separation has appeared between the distal slough and the rest of the limb.*

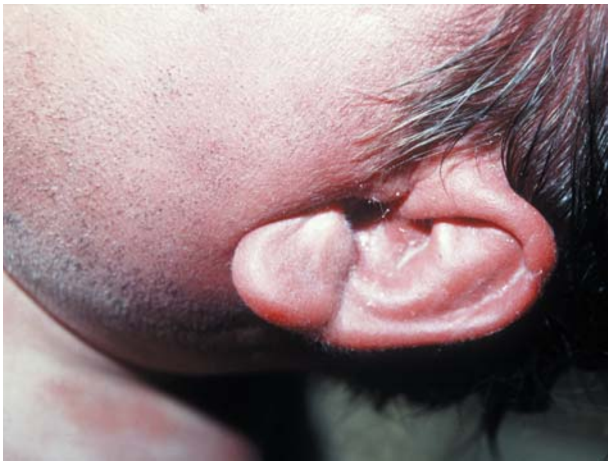

FIGURE 17.11 *Frostbite on left auricle of a man who, heavily intoxicated by alcohol, left for home at night in the Arctic winter, but lost his way at −40°C and was found dead the following day.*

DEATHS FROM EXPOSURE

Fatal hypothermia can occur in healthy adults, as well as infants and old people. This is caused by purely environmental conditions, which need not be extreme. One common group, especially in more northern countries, is the inebriated person who becomes exposed to even moderate cold. The typical circumstances are a person who gets drunk and either falls asleep outdoors in his drunken state or sustains some accident – such as a fall or even mild head injury – which incapacitates him for long enough for him to become hypothermic. Alcohol seems to have a secondary effect, apart from causing incapacity and immobility. Its vasodilating effect on skin allows heat loss to be more rapid and in higher concentrations it may have a direct action on the heat-regulating centres. Apart from drunken victims, many other classes of person may fall prey to hypothermia. Hill-walkers, climbers, yachtsmen and dinghy-sailors, swimmers, and other classes of sportsmen and women can become cold and exhausted. Every year sees deaths and rescue escapades in the hill country of Britain, even when the weather is far from Arctic in severity. Naturally the risks in colder countries and at high altitudes almost anywhere in the world are much greater. For the pathologist, the features are similar to domestic hypothermia, but there may be the added lesions of frostbite.

LOCAL INJURY DUE TO COLD: FROSTBITE

There are a number of different clinical lesions caused by the effect of prolonged cold on the extremities. ‘Immersion foot’ and ‘trench foot’ refer to damp cold damage, whereas ‘frostbite’ is caused by dry conditions below zero. In the context of forensic pathology, they may be found in victims of fatal hypothermia, but rarely kill in themselves unless tissue necrosis becomes infected. Severe damage with sloughing and separation of toes, or part of the foot is likely to be seen only in survivors or victims of some catastrophe on sea or mountainside, but the author (BK) has seen one case where a neglected infant was left in a urine-soaked cot at temperatures below zero for many days. When it died of a mixture of hypothermia and chest infection, the parents placed it in a sack with a brick and dropped it down a well. On recovery of the body, separation of the necrotic distal part of a foot was seen, with a red zone of inflammatory tissue on the proximal margin. Frostbite may affect only the skin or may extend deeply into the underlying tissues. Lesser degrees are reversible, but infarction and necrosis may ensue if the blood supply is cut off by vasoconstriction, sludging and thrombosis, for a sufficient time.

REFERENCES AND FURTHER READING

Angel J, Sash Z. 1960. Hypothyroidism and exposure to cold. *Br Med J* **1**:1855–9.

Benecke M, Lessig R. 2001. Child neglect and forensic entomology. *Forensic Sci Int* **120**:155–9.

Bray M. 1984. The eye as a chemical indicator of environmental temperature at the time of death. *J Forensic Sci* **29**:396–403.

Davis JH, Rao VJ, Valdes-Dapena M. 1984. A forensic science approach to a starved child. *J Forensic Sci* **29**:663–9.

Denmark LN. 1993. The investigation of beta-hydroxybutyrate as a marker for sudden death due to hypoglycemia in alcoholics. *Forensic Sci Int* **62**:225–32.

Duguid H, Simpson R, Strowers J. 1961. Exposure to cold. *Lancet* **2**:1213–19.

Elias AN, Gwinup G. 1982. Glucose-resistant hypoglycemia in inanition. *Arch Intern Med* **142**:743–6.

Emslie-Smith D. 1958. Hypothermia. *Lancet* **2**:492.

Fieguth A, Gunther D, Kleemann WJ, *et al.* 2002. Lethal child neglect. *Forensic Sci Int* **130**:8–12.

Fitzgerald FT. 1980. Hypoglycemia and accidental hypothermia in an alcoholic population. *West J Med* **133**:105–7.

Hirvonen J. 1976. Necropsy findings in fatal hypothermia cases. *Forensic Sci* **8**:155–64.

Hirvonen J. 2000. Some aspects on death in the cold and concomitant frostbites. *Int J Circumpolar Health* **59**:131–6.

Hirvonen J, Huttunen P. 1982. Increased urinary concentration of catecholamines in hypothermia deaths. *J Forensic Sci* **27**:264–71.

Hirvonen J, Huttunen P. 1995. Hypothermia markers: serum, urine and adrenal gland catecholamines in hypothermic rats given ethanol [see comments]. *Forensic Sci Int* **72**:125–33.

Hirvonen J, Lapinlampi T. 1989. Plasma and urine catecholamines and cerebrospinal fluid amine metabolites as hypothermia markers in guinea-pigs. *Med Sci Law* **29**:130–5.

Kelsey RM, Alpert BS, Patterson SM, *et al.* 2000. Racial differences in hemodynamic responses to environmental thermal stress among adolescents. *Circulation* **101**:2284–9.

Knight B. 1976. Forensic problems in practice. XI. Injury from physical agents. *Practitioner* **217**:813–18.

Kortelainen ML, Huttunen P, Lapinlampi T. 1990. Urinary catecholamines in hyperthermia-related deaths. *Forensic Sci Int* **48**:103–10.

Kreyburg L. 1946. Tissue damage due to cold. *Lancet* **1**:338–40.

Laufman H. 1951. Hypothermia. *JAMA* **147**:1201–12.

Mant AK. 1964. Some port-mortem observations in accidental hypothermia. *Med Sci Law* **4**:44–7.

Mimasaka S, Funayama M, Adachi N, *et al.* 2000. A fatal case of infantile scurvy. *Int J Legal Med* **114**:122–4.

Molnar G. 1946. Survival of hypothermia by men immersed in the ocean. *JAMA* **131**:1045–50.

Ortmann C, Fechner G, Bajanowski T, *et al.* 2001. Fatal neglect of the elderly. *Int J Legal Med* **114**:191–3.

Paton BC. 1983. Accidental hypothermia. *Pharmacol Ther* **22**:331–77.

Sadler DW, Pounder DJ. 1995. Urinary catecholamines as markers of hypothermia. *Forensic Sci Int* **76**:227–30.

Steinhauer JR, Volk A, Hardy R, *et al.* 2002. Detection of ketosis in vitreous at autopsy after embalming. *J Forensic Sci* **47**:221–3.

Tanaka M, Tokudome S. 1991. Accidental hypothermia and death from cold in urban areas. *Int J Biometeorol* **34**:242–6.

Teresinski G, Buszewicz G, Madro R. 2002. The influence of ethanol on the level of ketone bodies in hypothermia. *Forensic Sci Int* **127**:88–96.

Wedin B, Vanggaard L, Hirvonen J. 1979. 'Paradoxical undressing' in fatal hypothermia. *J Forensic Sci* **24**:543–53.

Wolf DA, Aronson JF, Rajaraman S, *et al.* 1999. Wischnewski ulcers and acute pancreatitis in two hospitalized patients with cirrhosis, portal vein thrombosis, and hypothermia. *J Forensic Sci* **44**:1082–5.

Young AJ, Castellani JW. 2001. Exertion-induced fatigue and thermoregulation in the cold. *Comp Biochem Physiol A Mol Integr Physiol* **128**:769–76.

Zhu BL, Oritani S, Ishida K, *et al.* 2000. Child and elderly victims in forensic autopsy during a recent 5 year period in the southern half of Osaka city and surrounding areas. *Forensic Sci Int* **113**:215–18.

CHAPTER 18

Deaths associated with sexual offences

- External findings in fatal sexual assaults 421
- Internal examination 424
- Interpretation of minimal findings 425
- Tests for semen 426
- Homosexual offences 428
- Further reading 429

The examination of living victims of sexual assaults is the province of the forensic physician or 'police surgeon', in places where such doctors exist. In many countries, any physician may be asked by the police to examine the victim of an alleged sexual offence, but this difficult and onerous task in the living victim is outside the experience of many pathologists – though in some jurisdictions, this function is combined with that of the forensic pathologist.

In this book, attention will be confined to those cases in which a sexual assault leads to death, almost always rape and homicide. Regrettably, a significant proportion of non-domestic homicides are associated with sexual offences, death occurring either because the woman rejects the sexual approaches or because a sadosexual motive colours an intended murder. The most common mode of killing is either pressure on the neck, head injuries, or, less often, stabbing. In some instances death may be the result of violence associated with the sexual activities themselves, especially in child victims when pelvic trauma may be the immediate cause of death. In this chapter, only the sexual aspects are discussed, as other modes of violent death are covered elsewhere.

EXTERNAL FINDINGS IN FATAL SEXUAL ASSAULTS

As with most forensic autopsies, a meticulous external examination is as important – and usually more important – than the internal dissection. The study of the body surface in sexual assaults is similar in both the living and dead and may be divided into general appearances and those of the perineum. The whole body surface should be searched for sexually orientated injuries.

Mouth and lips

Bruises may be found on the lips and especially inside on the buccal surfaces, from rough kissing. The lips may have been forced back on to the edges of the teeth, causing abrasions, bruises and even laceration of the buccal surface. Swabs should be taken from the mouth in every case before substantial manipulation for examination, to seek evidence of seminal fluid from oral penetration. Though such swabs rarely yield a positive result as to the presence of spermatozoa, modern analytical methods may be able to detect other components of the seminal fluid or the presence of male epithelial cells.

Bites

Other bruises may be found on the neck, shoulders, breasts and buttocks from oral suction or biting. These are somewhat euphemistically termed 'love bites' and may range from minor lesions inflicted from enthusiastic passion to mutilating bites of a sadistic nature.

Suction lesions may comprise a circular or oval area of bruising, in which the damaged zone consists of many small intradermal petechial haemorrhages caused by sucking the skin into the mouth, the reduced air pressure rupturing the small vessels. There may be semilunar marks at the periphery from the lips, and there may be associated teeth indentations or abrasions. Such suction marks are often seen at the sides of the neck, below the ears, on the upper shoulder, and on the upper part of the breasts and around the nipples.

Actual bites may accompany suction marks or may be independent. The characteristics are described in Chapter 26, but the main features are abrasions or bruises matching the dental arches, often with discrete marks from each tooth. These may be linear if the teeth are scraped down the skin in closing the jaws. They may be superficial or deep enough to penetrate the skin as lacerations. The latter are especially common at the nipples in sexual assaults, and the nipple and part of the areola may even be partly or wholly amputated.

In sexual homicides, it is essential to obtain expert forensic odontological advice, if at all possible. In any event, photographs and measurements must be taken of these injuries and, if practicable, a cast of any teeth indentations for later comparison. Before any handling of the lesions, swabs should be taken of the surface in an effort to recover saliva, which – if the assailant was a 'secretor' – may be tested for blood-group-specific substances.

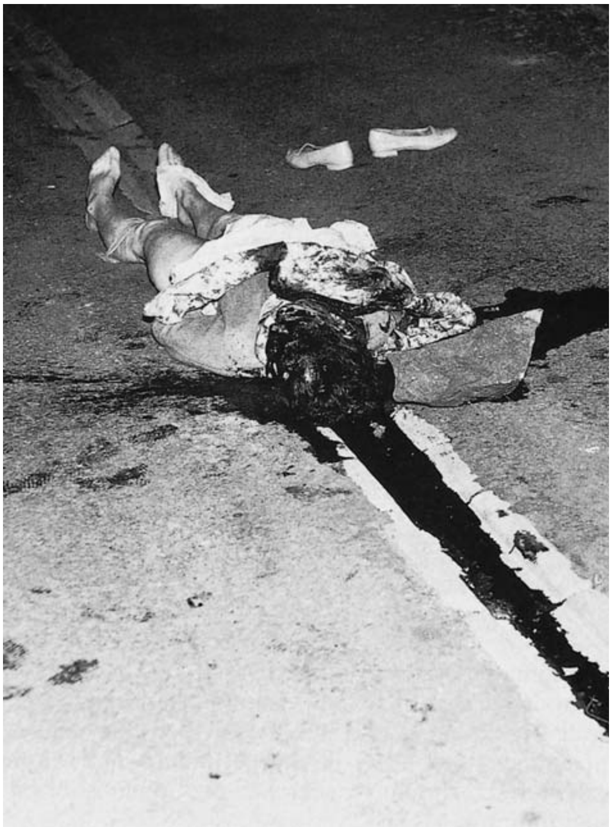

FIGURE 18.1 *Sexual assaults are sometimes associated with extreme violence, often of a sadistic nature. Here, rape has been followed by severe head injury from the large rock seen near the central gutter of this back lane.*

General bruising and abrasion

Apart from bruises that may be associated with the fatal assault, such as throat or head injuries, some bruises are indicative of the sexual motivation. The breasts are often manually squeezed and manipulated, causing discoid bruises of 1–2 cm on any part, especially around the nipples. Linear abrasions, usually from fingernails, may also be present. Bruises may also be seen on the thighs and buttocks, from struggles to achieve intercourse. Both the outer and inner sides of the thighs may be bruised and scratched: a typical area is the inner aspect of the upper thigh, where the legs are manually forced apart. There may be bruising around the anus from fingers opening the buttock cleft to achieve either anal or posterior vulval penetration.

Where the assault takes place on a hard or uneven surface, bruises and abrasions may be seen on the back, especially the shoulders and buttocks. If outdoors, there may be marks from stones or sticks and remnants of vegetation, such as leaves, and grass and green staining may be adherent to the skin, together with earth or dirt. Examination of the clothing is usually a function of the forensic scientist ('criminalist'), but when the pathologist examines the body *in situ*, he will naturally note damage, disarray and foreign material on the garments.

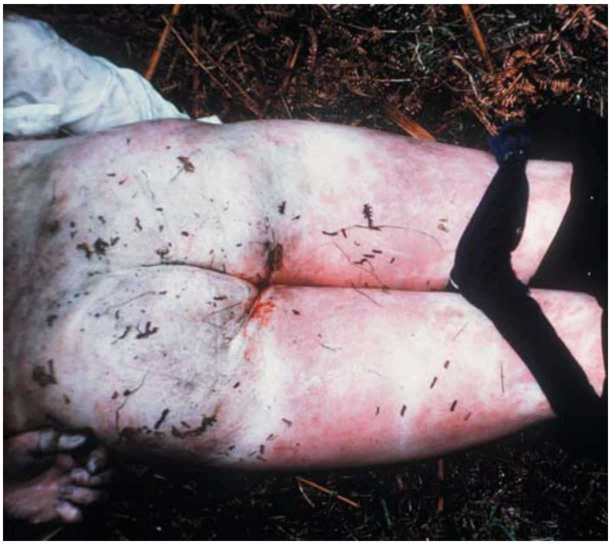

FIGURE 18.2 *In fatal rape the pathologist must note at the scene, or at the mortuary, the position and condition of clothing and presence of foreign material on the skin. Full cooperation with scene of crime officers or forensic scientists is vital in collecting trace evidence.*

Other general injuries

The hands should be carefully examined, as in all assaults. The fingernails, especially the often longer and manicured nails of younger women, may be broken from a struggle. Occasionally, there are hairs or fibres trapped in nail splits, which may have come from the assailant or his clothing. At autopsy, the nails should be cut closely to the junction with the fingers and all parings carefully collected for forensic laboratory examination. It is rarely necessary to collect those from each finger in individual packets, but each hand should be identified separately. Some pathologists or scientific officers prefer to scrape out the space under each nail with a pointed orange-stick or toothpick, carefully retaining any debris. The reason for this interest, especially in sexual assault victims, is that they may have scratched their assailants and have blood, or even skin parings, under the nails that can be identified as to blood group or individually specific DNA characteristics.

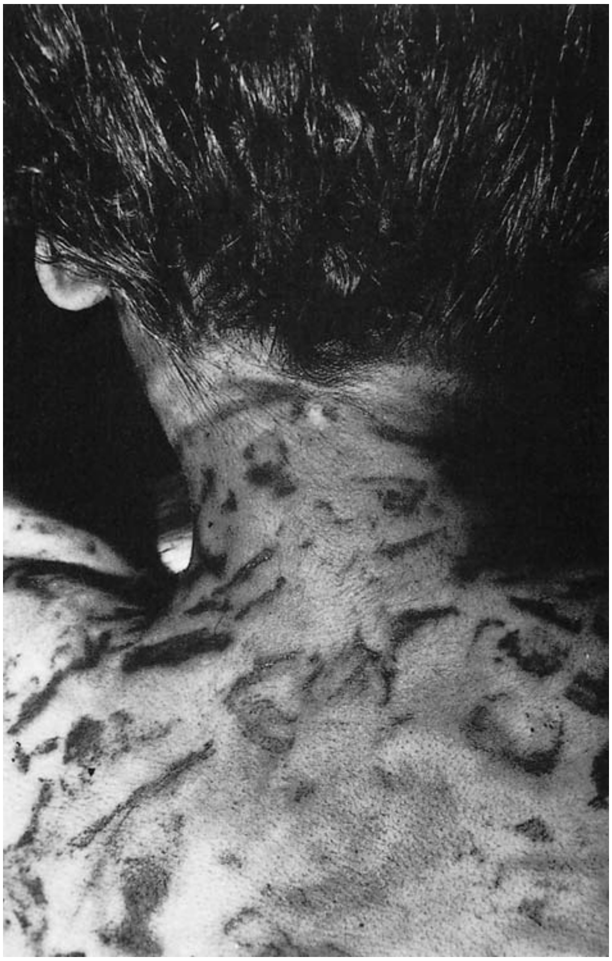

FIGURE 18.3 *Multiple superficial abrasions from undergrowth and ground debris on the back of the victim of rape. This is an indication of pressure of the unclothed body on the ground. There is a ligature mark around the neck with a cross-over point visible.*

Signs of head injury and throttling are dealt with elsewhere, but some fatal trauma may be sexually orientated, especially knife injuries. Sadistic cutting or stabbing may be inflicted on sexually significant areas, especially the breasts, buttocks, perineum and lower abdomen. The vulva itself may be mutilated. These may be slashes or deeply penetrating stabs, often multiple and indeed numerous. They may be arranged in a pattern or may be mutilating, as when whole breasts are amputated. Decapitation is sometimes performed and ritual acts may be performed, with objects placed in definite sites near the body. The notorious 'Jack the Ripper' murders in London in 1888 exhibited the arrangements of disembowelled organs and this type of mutilation is another feature of some sadistic sexual homicides. Intestine or uterus may be removed, either through abdominal wounds or manually through the vagina. General injuries may be continuous with genital trauma, as in knife or broken bottle wounds of the abdomen and perineum.

Genital injuries

Once the pathologist has examined all the external general injuries, attention should be turned to the perineum. As with the clinical examination of sexual offences, a set routine should be used, as the incorrect order may lose valuable forensic evidence.

The vulva and anus should be inspected externally and laceration, swelling, bruising, bleeding and discharge noted. Any blood or suspected semen stains anywhere on the body or clothing should be sampled either by the pathologist or by the forensic scientist or scene of crime police officer. The pubic hair should be examined for foreign material, hairs, vegetation and dried seminal stains, and samples of hair and combings taken. The hair may be combed using a fine comb with the base of the teeth packed with cotton-wool to trap any loose fibres. Dried stains on hair may be cut away and placed in clean folded paper in an envelope, or plastic bag for transit to the laboratory. The vulval labia may show clear signs of trauma, especially in children. In forcible rapes, especially in young persons, there may be external signs of perineal tears, with laceration of the margin of the vaginal introitus or anus, sometimes causing a complete rip between the two orifices.

Caution must be used in interpreting the degree of dilatation of the anus in a dead body, as the sphincter can become patulous and wide open as a normal post-mortem change. Unless the dilatation is very marked, the sole finding of an open anus in the absence of abrasion, bruising, or semen is difficult to sustain as proof of anal penetration. This applies

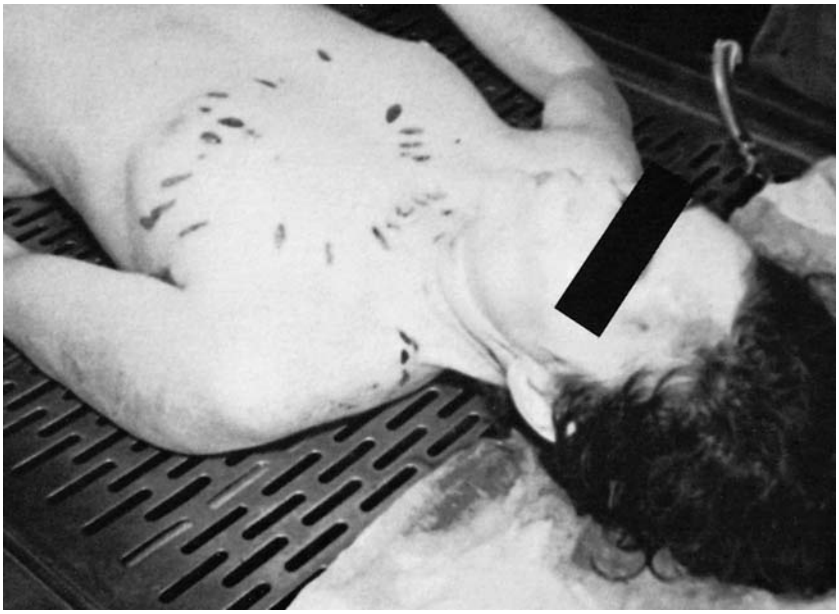

FIGURE 18.4 *Some homicides have obvious sexual connotations even if there is no genital interference. This circle of stab wounds is centred on a breast.*

to children as well as adults; in fact, post-mortem dilatation of the anus is particularly noticeable in children.

Once external examination and assessment has been made, samples should be taken for forensic biological testing, for the presence of semen and venereal infection, as well as samples for DNA profiling, which may be able to distinguish between any semen present and the vaginal secretions. If any fluid is running from the vulva or anus, it should be picked up with clean pipettes and preserved in the smallest available tube, to prevent drying from evaporation. Then cotton-wool swabs on sticks (ensuring that those used for semen detection are plain swabs, not those containing albumin or other media) should be used to take the following samples by touching gently on to the mucosal surfaces:

- the interior of the vulval labia and around the vaginal orifice
- the margins and interior of the anus
- the mid-vagina, using a speculum or broad handle of a dissecting forceps to part the lower vaginal walls gently to allow the swab to reach the area without contamination from the lower vagina
- the upper vagina, cervix and posterior fornix, again using a spatulate instrument to open the canal to give access to the swab. If more fluid contents are seen higher in the vagina, either now or during the later dissection, they should be recovered by pipette.

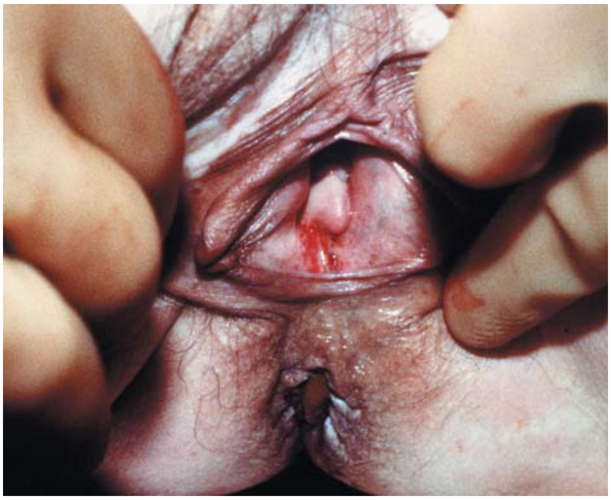

FIGURE 18.5 *Fresh bleeding and mucosal abrasion on the lower wall of the vagina in a sexual assault associated with fatal head injuries in an old lady.*

After samples are taken, the interior of the anus, vulva and vagina, as far as can be seen from the exterior, should be examined. Lacerations, abrasions, bruises and bleeding may be evaluated, though they can be seen in more detail at dissection. The presence and condition of a hymen is noted, with any evidence of recent tearing. Where injuries are gross, especially in children, fistulae between the vagina and the rectum or even peritoneal cavity may be seen. There may be extrusion of intestine.

INTERNAL EXAMINATION

The order of the autopsy may be determined by the nature of the death. The fatal injuries, such as strangulation or head

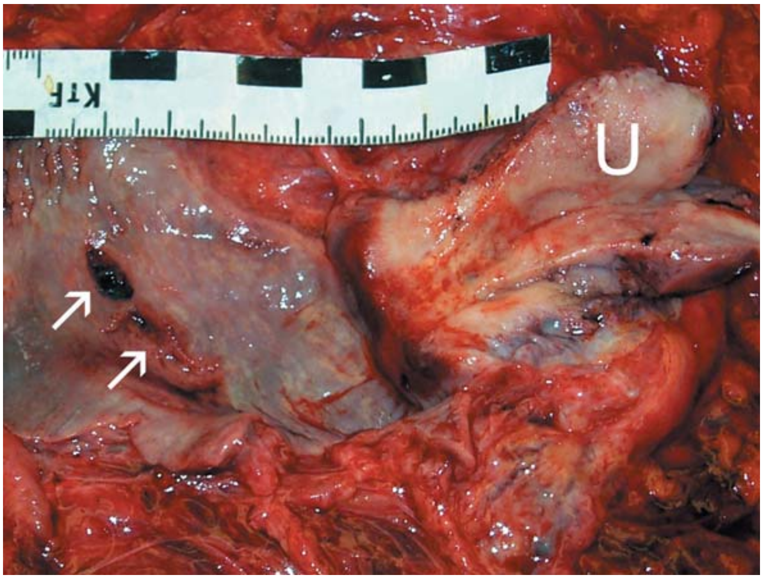

FIGURE 18.6 *Stab wounds (arrows) in the vaginal wall of a homicide victim (U, uterus).*

injuries, may be dealt with first or the pelvic region may command prime attention. The autopsy incision should be similar to that described for the examination of deaths associated with pregnancy, with an incision circumscribing the perineum, removal of the anterior part of the pelvic bones and the extraction of all the pelvic organs in continuity from ovaries to vulva and anus. This block of tissue is then dissected separately. Before this is done, the bladder should first be emptied of urine (either by catheter or through a small incision in the fundus) and the fluid retained for toxicological analysis, especially for alcohol.

The vagina should be opened with large scissors, the track of the cut depending on the assessment of any injuries seen on external examination. If there are tears or bruises in the vulva or vagina, the cut should be orientated to avoid them wherever possible; the anus is later dealt with in a similar way. The vagina is laid open to the posterior fornix and all injuries carefully examined and photographed. The cervix and uterus are examined in the same way.

Injuries may be of all types, from mere reddening or swelling to complete disruption of the vaginal canal. This may occur in small children from sheer brutality of penetration, especially if there is gross disproportion between the adult penis and infantile canal. It may also occur from instrumental injury, as it is by no means uncommon for deliberate incised wounds to be made. The author (BK) has seen abdominal knife wounds carried over the pubis into the vagina in a murdered prostitute; another murder–rape was found to have a broken milk bottle high in the vagina, to the potential hazard of the examining fingers. Vaginal injuries, especially by instrument, may continue up into the abdominal cavity, either via the posterior fornix or lateral vaginal walls. This will have been examined through the abdominal autopsy incision before removal of the pelvic organs.

INTERPRETATION OF MINIMAL FINDINGS

As in clinical forensic practice, there is little difficulty in interpreting gross sexual interference, but problems arise when only minimal evidence is present in a death (traumatic or otherwise) where the circumstances suggest some sexual offence or activity. In such cases, the following questions need to be answered, in the light of all the available findings at the end of the autopsy and ancillary investigations:

- Is there any evidence of sexual intercourse at any time – that is, absence of virginity as determined by an intact hymen? This does not exclude sexual activity short of penetration and technically rape can occur from even the minimal passage of a glans between the labia, which does not affect the hymen. Old intercourse may be assumed from healed hymenal tears with epithelialized ‘carunculae myrtiformes’ at the margins, though the prolonged use of tampons and manual manipulation can also tear a hymen. Evidence of previous pregnancy, such as abdominal striae, old damage to the cervix and breast changes are almost incontrovertible evidence of previous sexual intercourse.
- Is there evidence of recent sexual intercourse? A recently ruptured hymen, with swelling, a raw unepithelialized edge and bleeding may be found, though admittedly it is relatively uncommon except in children and previously virginal young persons. The labia may be red and

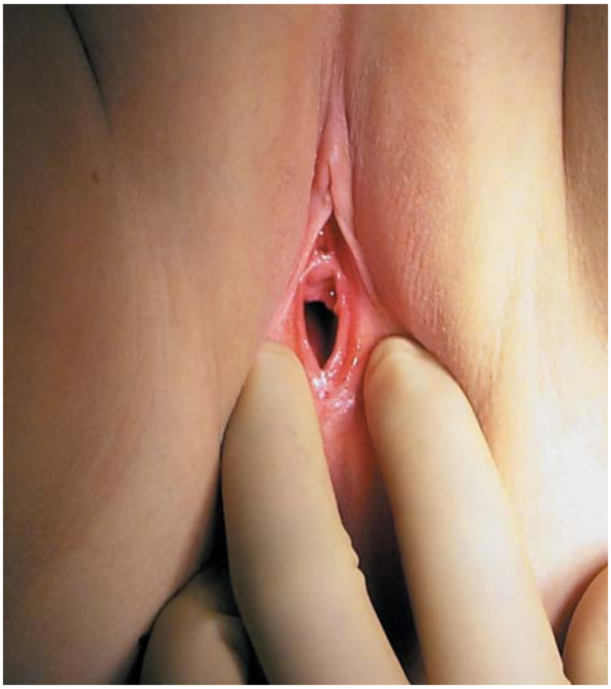

FIGURE 18.7 *External genitals of a live 8-year-old girl, a victim of repeated sexual abuse by an adult living in the neighbourhood. The lack of intact hymen, reddened and larger than normal orifice of the vagina as well as thickened skin and ulcers perianally (not shown) are in keeping with abuse.*

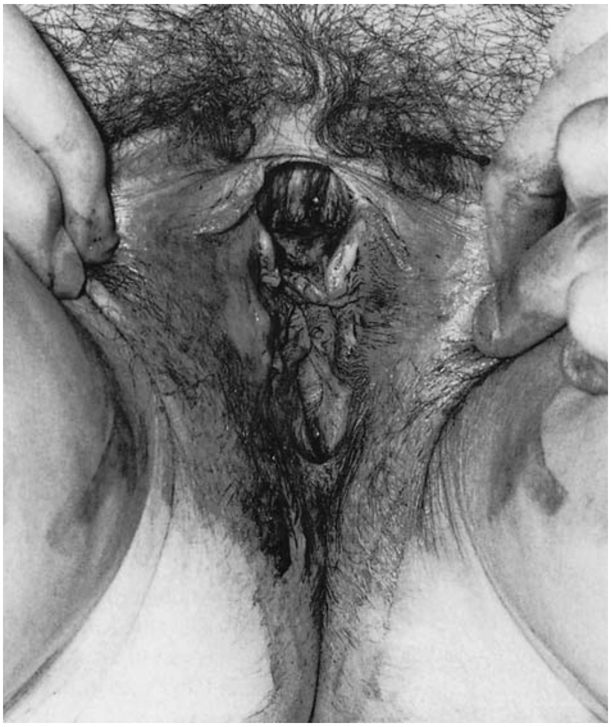

FIGURE 18.8 *Extensive bleeding from inner lacerations of the vaginal walls in a fatal rape. Blood is pooling in the posterior part of the vaginal canal. A full autopsy dissection is required, keeping the perineum en bloc with the genital tract to display any injuries and to exclude natural disease as a contribution to the bleeding.*

inflamed with slight oedema of the vaginal introitus if it is the first episode, or if there is disproportion between an adult man and a young person, even in voluntary sexual connection. The presence of semen on swabs is the best evidence, though with the use of condoms it may be absent. After vasectomy, though no sperms will be present, chemical and enzyme tests for semen remain positive. The presence of venereal disease, especially gonorrhoea, is presumptive evidence of intercourse, though it can be contracted other than by sufficient penile penetration to constitute rape.

- If recent intercourse has taken place, was it by force? This may be all too obvious in the presence of gross injury, especially in small children. Where vaginal or rectal tearing has occurred, or where there is obvious abrasion, bruising or laceration of the vulva, anal margins or perineum, then this can hardly be compatible with voluntary intercourse. The possibility of sexually motivated injury without actual penile penetration must be considered when no semen can be recovered, as equally severe damage can be caused by digital or instrumental trauma. All kinds of foreign objects can be forcibly introduced into the vagina and rectum, and are not infrequently employed by sadistic and perverted killers. The author (BK) has seen candles, bottles, a banana, broken glass and knives introduced into the vagina, and broom-handles and potatoes lying in the rectum.

  Where injury is relatively slight and confined to hyperaemia and oedema of the vaginal or anal entrances – and where abrasion and bruising of the vulva is slight (even including fingernail scratches), although the presumption is that intercourse was by force, the possibility still exists that it was voluntary though overenthusiastic, unless, of course, the victim was a small child.

TESTS FOR SEMEN

In countries with sophisticated forensic science or police laboratories, the pathologist will not be called upon to carry out the actual techniques for the detection of seminal fluid. His role is to be a careful collector of the original samples at autopsy and, to some extent, to be able to interpret the results for the benefit of the police and the courts. He should be conversant with the general principles of the tests and aware of any potential shortcomings. In places where forensic science facilities are deficient or their availability may be delayed, the pathologist may have to conduct the more simple tests himself and the technical details are given in Appendix 1. The detection of semen depends upon the following methods:

- Naked-eye and lens recognition is a screening method which may identify suspicious stains on clothing, skin and pubic hair. Liquid semen leaking from the vagina or anus may be obvious, though leucorrhoeic vaginal discharge may be mistaken for seminal fluid. Dried seminal stains are stiff and rather silvery, depending upon the nature and colour of any fabric they may be upon. On the skin they flake off readily and can easily be lost if not carefully lifted with a blade or needle into a container. Dried stains on hair should be retrieved by cutting away the clump on which they lie. Suspect smears on skin may be collected by gentle rubbing with a plain cotton-wool swab that has been moistened with water or saline. All material should be examined as soon as possible after collection.
- Examination under ultraviolet light causes seminal stains to fluoresce a bluish-silver, but many other biological fluids and vegetable juices give false-positive

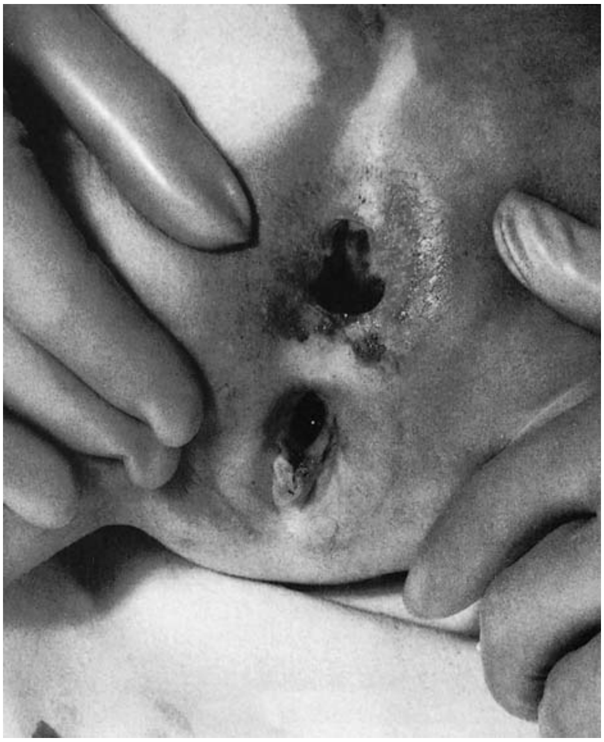

FIGURE 18.9 *Rape and buggery of a small girl who was then strangled. The vulva has a slight posterior tear and there is a large vaginal tear internally, the canal being filled with blood. The anus is widely ripped, especially posteriorly. Swabs were positive for semen in both orifices.*

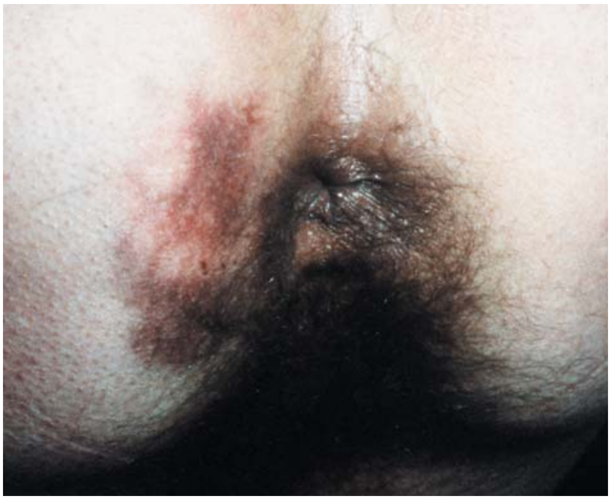

FIGURE 18.11 *Extensive bruising from fingers during a multiple anal and vaginal rape. Though penetrated by several men in a short space of time, the anus shows no injury.*

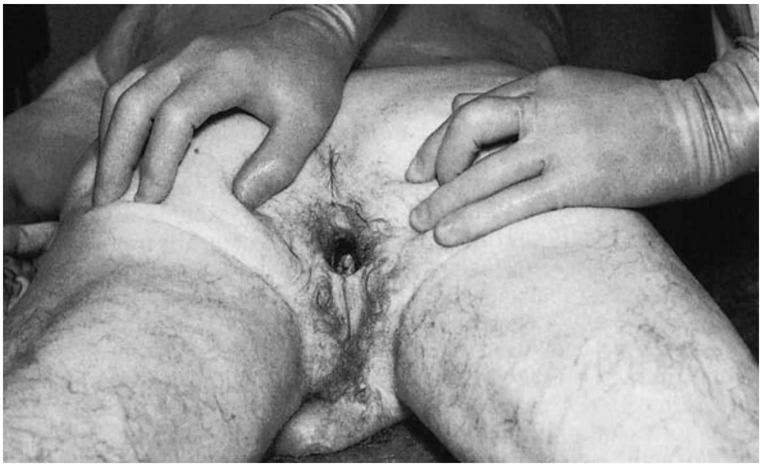

FIGURE 18.10 *A chronically dilated anus in an habitual passive homosexual. After death any anus can appear open and patulous, but this example is too large to be accepted as within normal limits. It also has keratinized, chronically abraded margins.*

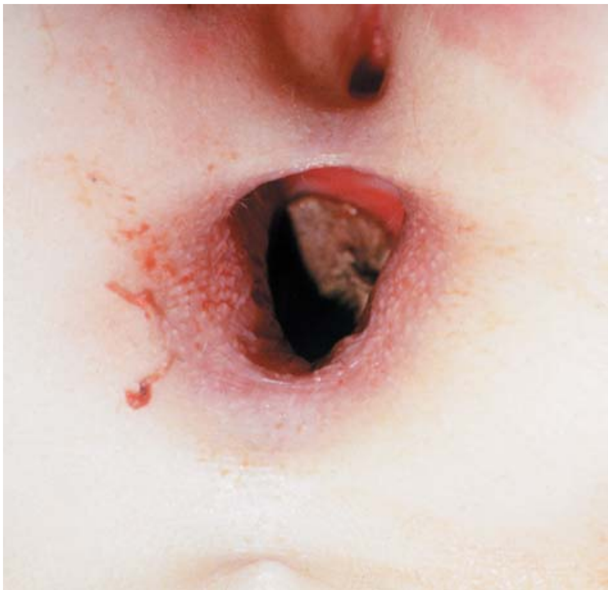

FIGURE 18.12 *The anus of a girl who has suffered penetration. The sphincter is overdilated, there is slight tearing of the anal margin, and some epidermal tags have been abraded from the surface. However, this degree of dilatation in a dead subject should not be used as the sole indication of buggery, as the post-mortem anus can be patulous in quite innocent circumstances.*

results. Many detergents, such as washing powders, also produce a confusing background fluorescence. The method is a useful screening process but cannot be used as definitive evidence of the presence of semen. Areas that fluoresce, especially on fabric, should be outlined to identify them for more specific testing.

- Enzyme reactions are valuable, and though not absolute proof, are strong presumptive evidence of the presence of semen. They depend on detecting high concentrations of acid phosphatase originating in prostatic secretion. Acid phosphatase activity in semen is 500–1000 times greater than in any other normal bodily fluid. Also, vaginal fluid contains endogenous acid phosphatase showing considerable variation in activity depending on various factors, which complicates the interpretation of the findings. There are several methods of detecting phosphatase, including a rapid commercial 'spot' test, which should not be used as hard evidential proof.
- Immunological methods, where animal antihuman semen sera are set up against extracts of the stain. These techniques can only be used by trained laboratory staff.

  Prostate-specific antigen (PSA, p30) is a glycoprotein produced by the prostatic gland and is found in seminal plasma, male urine and blood, but not in any tissues or fluid of the female body. A positive PSA finding is a reliable indicator of semen regardless of the presence of spermatozoa or elevated acid phosphatase level. Simple and sensitive methods have been developed for analysis of the presence of PSA.

  Other serological techniques may be able to determine the blood group of the semen if the ejaculator was one of the 80 per cent of the population who are 'secretors' – those who pass blood-group antigens into their saliva, sweat and semen. A problem for the forensic serologist is the admixture of semen with biological fluids from the victim, in the form of vaginal secretion and blood, but this is a complex technical matter that is usually outside the remit of the forensic pathologist, though he needs to be aware of the potential difficulties.

  Recent developments have now added the DNA technique to the armamentarium of the forensic biologist; this is a remarkable advance, especially in sexual offences, as the previous problems of blood group and enzyme typing caused by the admixture of semen and vaginal secretion can now be solved in many instances. The sensitivity of detection and identification has also been enormously enhanced by amplification techniques, such as PCR (polymerase chain reaction). Y-chromosome short tandem repeat (STR) markers have been succesfully used to analyse mixed stains with a male component.
- Until recent years, the only absolute proof accepted by the courts was the demonstration of spermatozoa in stain extracts and it remains the best evidence, though some of the sophisticated laboratory techniques such as serological and DNA identification are now equally reliable. One modern problem in using spermatozoa retrieval is the widespread and increasing incidence of vasectomy as a method of male contraception, so the seminal fluid contains no sperm; this also occurs naturally in azoospermic men. The phosphatase and some serology tests are unaffected by vasectomy. Fluorescent *in situ* hybridization (FISH) has been suggested as a sensitive and specific test for the detection of male epithelial cells in the post-coital vagina up to one week. Y-chromosome-positive epithelial cells have been identified in vaginal swabs even in cases with no ejaculation.

HOMOSEXUAL OFFENCES

As with heterosexual offences, the cause of death in fatal cases is almost always some form of general trauma, such as strangulation or head injuries. Homosexual activity, however, may be a parallel event; it is a fact that some of the most violent homicides seen by pathologists are among male homosexuals. In addition, quite a number of fatal altercations arise because a heterosexual man becomes violent when importuned by a homosexual.

The pathologist's attention will almost always be directed to the anal region of the passive participant, as the other partner or assailant is rarely dead, and any examination will be carried out by a police surgeon or clinical doctor. The examination of the victim follows closely on that for a female subject of rape, in that general injuries are sought first – including the one that caused death – together with any sexually orientated marks such as scratches anywhere on the body, but especially down the back, on the buttocks or on the thighs. Bruising around the anus and in the cleft of the buttocks is significant, especially if discoid finger bruises are seen. The anus should be inspected and the warning given earlier about the degree of dilatation in corpses should be heeded. Though most marked in children, a patulous anus can be seen in any adult post-mortem. Unless it is grossly dilated, an open anus cannot be used in isolation as an index of homosexual penetration. In recent or forcible anal intercourse, especially by a large penis upon a 'virgin' or young person, the margins of the anus may be reddened, sometimes abraded, and occasionally torn, especially in a child. There might be some eversion of the lower rectal mucosa through the sphincter. In gross anal 'rapes', especially on a young person, there may even be severe perineal tearing. At autopsy, swabs must be taken from the anal margin, and deeper in the anus and rectum. Mouth swabs should also always be taken. The presence of blood or lubricant should also be sought.

The classic description of the chronic passive homosexual has been passed down from textbook to textbook over the years, but is a rarity. The so-called 'funnel anus' is usually an anatomical variant. Some men and women have a deeply set anus as an anatomical variant, sometimes with a bridge of skin dorsally across the cleft of buttocks. This causes the anus to appear as if it was at the bottom of a funnel, but has no significance. The other classic feature was said to be silvery thickening of the skin outside and at the margins of the anus – but this is often due to chronic scratching from pruritis, rather than from prolonged friction during homosexual activities. In life, a lax anus, prolapsed mucosal rim and thickened anal margin may well be genuine signs to the experienced clinical eye, but at autopsy only acute damage and the presence of semen or lubricant is of much evidential value.

FURTHER READING

Allard JE. 1997. The collection of data from findings in cases of sexual assault and the significance of spermatozoa on vaginal, anal and oral swabs. *Sci Justice* **37**:99–108.

Allen M, Saldeen T, Gyllensten U. 1995. Allele-specific HLA-DRB1 amplification of forensic evidence samples with mixed genotypes. *Biotechniques* **19**:454–63.

Berg SP. 1955. A new method for testing for seminal stains. *Crim Pol Rev* **11**:53–7.

Carracedo A, Beckmann A, Bengs A, *et al.* 2001. Results of a collaborative study of the ednap group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389 I and II, DYS390 and DYS393 in a PCR pentaplex format. *Forensic Sci Int* **119**:28–41.

Clery JM. 2001. Stability of prostate specific antigen (PSA), and subsequent Y-STR typing, of *Lucilia (Phaenicia) sericata (Meigen)* (Diptera: Calliphoridae) maggots reared from a simulated postmortem sexual assault. *Forensic Sci Int* **120**:72–6.

Collins KA, Bennett AT. 2001. Persistence of spermatozoa and prostatic acid phosphatase in specimens from deceased individuals during varied postmortem intervals. *Am J Forensic Med Pathol* **22**:228–32.

Concheiro L, Carracedo A, Guitian F. 1982. The use of scanning electron microscopy in the examination of seminal stains. *Forensic Sci Int* **19**:185–8.

Costa MJ, Tadros T, Tackett E, *et al.* 1991. Vaginocervical cytology in victims of sexual assault. *Diagn Cytopathol* **7**:337–40.

Crouse CA, Ban JD, D'Alessio JK. 1993. Extraction of DNA from forensic-type sexual assault specimens using simple, rapid sonication procedures. *Biotechniques* **15**:641–4, 646, 648.

Dahlke MB, Cooke C, Cunnane M, *et al.* 1977. Identification of semen in 500 patients seen because of rape. *Am J Clin Pathol* **68**:740–6.

Davies A. 1978. Evaluation of results from tests performed on vaginal, anal and oral swabs received in casework. *J Forensic Sci Soc* **17**:127–33.

Davies A, Wilson E. 1974. The persistence of seminal constituents in the human vagina. *Forensic Sci* **3**:45–55.

Department of Health. 1988. *Diagnosis of child sexual abuse.* HMSO, London.

Dziegelewski M, Simich JP, Rittenhouse-Olson K. 2002. Use of a Y chromosome probe as an aid in the forensic proof of sexual assault. *J Forensic Sci* **47**:601–4.

Enos WF, Beyer JC. 1978. Spermatozoa in the anal canal and rectum and in the oral cavity of female rape victims. *J Forensic Sci* **23**:231–3.

Enos WF, Beyer JC. 1980. Prostatic acid phosphatase, aspermia, and alcoholism in rape cases. *J Forensic Sci* **25**:353–6.

Enos WF, Beyer JC. 1981. The importance of examining skin and hair for semen in sexual assault cases. *J Forensic Sci* **26**:605–7.

Gabby T, Winkleby MA, Boyce WT, *et al.* 1992. Sexual abuse of children. The detection of semen on skin. *Am J Dis Child* **146**:700–3.

Gill P, Jeffreys AJ, Werrett DJ. 1985. Forensic application of DNA 'fingerprints'. *Nature* **318**:577–9.

Hochmeister MN, Budowle B, Rudin O, *et al.* 1999. Evaluation of prostate-specific antigen (PSA) membrane test assays for the forensic identification of seminal fluid. *J Forensic Sci* **44**:1057–60.

Hooft P, van de Voorde H. 1988. In vitro changes in human spermatozoa exposed to gastric juice: laboratory findings as a support for forensic practice. *Z Rechtsmed* **101**:41–4.

Hooft P, van de Voorde H. 1992. Evaluation of the modified zinc test and the acid phosphatase test as preliminary screening methods in sexual assault case material. *Forensic Sci Int* **53**:135–41.

Hooft PJ, van de Voorde HP. 1997. Bayesian evaluation of the modified zinc test and the acid phosphatase spot test for forensic semen investigation. *Am J Forensic Med Pathol* **18**:45–9.

Jimenez-Verdejo A, Osuna E, Garcia-Olivares E, *et al.* 1994. Study of the enzymatic activity of GGT, LDH, PAP and PSA in semen stains: application to age calculation. *Forensic Sci Int* **68**:7–15.

Keil W, Bachus J, Troger HD. 1996. Evaluation of MHS-5 in detecting seminal fluid in vaginal swabs. *Int J Legal Med* **108**:186–90.

Kind SS. 1964. The acid phosphatase test. In: Curry AS (ed.), *Methods of forensic science*. Interscience, London, Vol. 3, pp. 267–87.

Maher J, Vintiner S, Elliot D, *et al.* 2002. Evaluation of the biosign PSA membrane test for the identification of semen stains in forensic casework. *N Z Med J* **115**:48–9.

Nicholson R. 1965. Vitality of spermatozoa in the endocervical canal. *Fertil Steril* **16**:758–64.

Paparo GP, Siegel H. 1979. Histologic diagnosis of sodomy. *J Forensic Sci* **24**:772–4.

Poyntz FM, Martin PD. 1984. Comparison of p30 and acid phosphatase levels in post-coital vaginal swabs from donor and casework studies. *Forensic Sci Int* **24**:17–25.

Reade DJ. 1985. Early investigations of sexual assault. *Police Surg* **April**:424.

Roberts REI. 1984. Sexual offences. In: McClay W (ed.), *Rape. The new police surgeon*. Association of Police Surgeons of Great Britain, Northampton, pp. 65–81.

Rupp JC. 1969. Sperm survival and prostatic acid phosphatase activity in victims of sexual assault. *J Forensic Sci* **14**:177–83.

Sato I, Sagi M, Ishiwari A, *et al.* 2002. Use of the 'SMITEST' PSA card to identify the presence of prostate-specific antigen in semen and male urine. *Forensic Sci Int* **127**:71–4.

Schiff AF. 1969. Modification of the Berg acid phosphatase test. *J Forensic Sci* **14**:538–44.

Schiff AF. 1975. Sperm identification – acid phosphatase test. *Med Trial Tech Q* **21**:467–74.

Sibille I, Duverneuil C, Lorin de la Grandmaison G, *et al.* 2002. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. *Forensic Sci Int* **125**:212–16.

Simich JP, Morris SL, Klick RL, *et al.* 1999. Validation of the use of a commercially available kit for the identification of prostate specific antigen (PSA) in semen stains. *J Forensic Sci* **44**:1229–31.

Soukos NS, Crowley K, Bamberg MP, *et al.* 2000. A rapid method to detect dried saliva stains swabbed from human skin using fluorescence spectroscopy. *Forensic Sci Int* **114**:133–8.

Steinman G. 1995. Rapid spot tests for identifying suspected semen specimens. *Forensic Sci Int* **72**:191–7.

Takatori T, Sasaki T. 1980. Isolation of spermatozoa in vaginal contents by centrifugation in a colloidal silica gradient. *Forensic Sci Int* **15**:61–5.

Toates P. 1979. The forensic identification of semen by isoelectric focusing of seminal acid phosphatase. *Forensic Sci Int* **14**:191–214.

Vihko P, Mattila K, Ehnholm C. 1981. Radioimmunoassay of human prostate-specific acid phosphatase. A sensitive and specific assay for semen detection in forensic medicine. *Am J Clin Pathol* **75**:219–20.

Willott GM, Allard JE. 1982. Spermatozoa – their persistence after sexual intercourse. *Forensic Sci Int* **19**:133–54.

Willott GM, Crosse MA. 1986. The detection of spermatozoa in the mouth. *J Forensic Sci Soc* **26**:125–8.

Zarrabeitia M, Amigo T, Sanudo C, *et al.* 2002. A new pentaplex system to study short tandem repeat markers of forensic interest on X chromosome. *Forensic Sci Int* **129**:85–9.

CHAPTER 19

Deaths associated with pregnancy

■ Deaths associated with abortion 431
■ The autopsy in abortion deaths 435
■ Amniotic fluid embolism 436
■ References and further reading 437

Though not usually a forensic problem, deaths associated with pregnancy (other than criminal abortions) are intensively investigated in a number of countries. In Britain, the Department of Health has had an ongoing 'Confidential Enquiry into Maternal Deaths' running for many years, which publishes valuable reports at 3-yearly intervals that have helped to clarify both the clinical and pathological aspects of the problem.

The pathologist has an important role in furthering the understanding of the causes of death in pregnancy and after childbirth. A good autopsy is essential, with full histological examination and other ancillary investigations where necessary. Only in this way can the full range of causes of death be recognized, especially amniotic fluid embolism, which is a histological diagnosis.

To give an example of the range of fatal conditions that occur in association with pregnancy, the following causes are taken from the 'Confidential Enquiry' from Britain published in 1998, the figures referring to death during pregnancy and within 6 weeks of parturition in the years 1994–1996. Altogether 376 deaths were reported to or identified by the enquiry in the UK during that period. Of the 376, there were 134 (36 per cent) direct and 134 (36 per cent) indirect maternal deaths. The number of underlying causes of death of the direct maternal deaths is shown in Table 19.1.

TABLE 19.1 *Number of direct maternal deaths by cause, United Kingdom 1994–1996*

Cause of death	*n*	Percentage
Thrombosis and thromboembolism	48	35.8
Hypertensive disease of pregnancy	20	14.9
Amniotic fluid embolism	17	12.7
Early pregnancy deaths*	15	11.2
Sepsis	14	10.4
Haemorrhage	12	9.0
Genital tract trauma	7	5.2
Anaesthesia	1	0.7
Total	134	100.0

* Ectopic pregnancy, spontaneous abortion, or termination of pregnancy, before 24 weeks' gestation.

According to a recent estimation, about 26 million legal and 20 million illegal abortions are carried out yearly throughout the world. Forensic pathological interest in pregnancy revolves almost exclusively around deaths associated with abortion, either criminal or legally induced. The term 'therapeutic abortion' is used in many countries, including Britain, where there is no abortion on demand, as in some countries. The description 'legal termination of pregnancy' is a wider definition for all but illegal (that is, criminal) abortions. The autopsy investigation of deaths from legal abortions has much in common with that into fatalities associated with surgical and anaesthetic procedures.

DEATHS ASSOCIATED WITH ABORTION

Though in many countries medical termination of pregnancy ('therapeutic abortion') is legal, a large area of the world still prohibits any form of abortion, either totally or except for the preservation of life of the pregnant woman. Even in those states where legal termination is possible, criminal abortions are still carried out, albeit on a small scale.

The reasons for this survival are varied, but are either because of restrictive grounds for legally available termination or to some sections of society that do not wish to avail themselves of the legal provisions. Whatever the cause, criminal abortion is associated with a considerable risk of both morbidity and mortality, especially in countries with a lower level of medical and social care.

Legal termination of pregnancy

When carried out with proper facilities, legal abortion has an extremely low mortality rate, being less than the mean death rate associated with pregnancy. The usual methods are vacuum aspiration, dilatation and curettage, or hysterotomy in later pregnancy. A few deaths are reported from time to time, the causes including:

- pulmonary embolism from leg vein thrombosis
- mishaps associated with anaesthesia
- disseminated intravascular coagulation and cerebral damage (including 'butterfly' haemorrhagic infarction in the basal ganglia) when abortion was induced by intrachorionic injection of hypertonic saline or glucose after the twelfth week (Cameron and Dayan 1960)
- air embolism following vacuum aspiration – only two cases have been reported and the mechanism is obscure, one theory being that 'elastic rebound' of the aspirated uterus sucked air into the cavity
- bleeding or infection, which failed to respond to treatment.

Death from illegal abortion

This has a much wider range of causes. The risks vary according to the skill, experience and facilities of the abortionist. When this is carried out by a doctor with aseptic and antiseptic methods, together with antibiotic cover if needed, the risk may be small compared with the crude methods of a lay person using makeshift instruments. The most common methods together with the associated dangers to health and life are as follows.

INSTRUMENTAL INTERFERENCE

The intention is to disturb the pregnancy sac so that, once damaged, it will be expelled by uterine contractions. This usually consists of dilatation of the cervical canal, which in itself also tends to dislodge the pregnancy. All manner of instruments have been used, from surgical dilators to bicycle spokes. A favourite in the hands of paramedical abortionists is the bougie or stiff catheter. When used by doctors or nurses with anatomical knowledge and sterile instruments, the risk is small, but lay persons often have no idea of the relationship of uterus to vagina. The instrument is then often pushed into the posterior fornix in the misguided belief that the cervix lies axially with the vagina. The vault of the vagina can be perforated and the instrument may even be passed through coils of intestine as far as the liver. Penetration of the lower or mid-vagina can also occur. If the cervix is entered, then the canal may be punctured and the instrument emerges through the side.

The external os may be badly injured by repeated, clumsy attempts to introduce too thick an object into the undilated canal. If successfully passed into the cavity of the uterus, it may be pushed right up through the fundus, again to damage the contents of the peritoneal cavity.

The dangers of such instrumentation are bleeding and infection. Perforation of the wall of vagina or uterus may cause severe bleeding, which may be internal or external. Sepsis can supervene in the peritoneal cavity or pelvic tissues either directly from a dirty instrument or from transfer of vaginal, skin or bowel organisms.

Another less common danger of the use of instruments (including syringes) is cervical shock. The mere act of dilating the cervix with an instrument in an unanaesthetized patient may trigger a vagal reflex, the efferent pathway being via the parasympathetic nervous system, causing a cardiac arrest. This is known to be a more potent mechanism in states of fear, apprehension and nervous tension, which obviously will apply to many candidates for a criminal abortion.

INSUFFLATION OF AIR

In Europe, this is becoming more of historical interest, but occasional cases still occur. A rubber pump, usually a Higginson enema syringe, is used to introduce fluid under pressure into the cavity of the uterus. This strips the chorionic sac from the wall of the uterus, exposing the placental bed. If sufficient detachment is achieved, then abortion will occur. It was formerly a popular method of abortion, both by women themselves and by abortionists. The main danger – part from the usual risks of bleeding and infection from damage to the tissues by the stiff nozzle – was air embolism, and in the first half of this century it was a major cause of abortion deaths. The intention was to introduce a fluid such as water–soap solution or disinfectant through the cervix by means of the syringe nozzle, the other end, which carried a one-way valve, being dipped in a receptacle of the fluid.

As the level in the receptacle dropped, the inlet tube rose above the surface and the syringe began to inject air instead of fluid, often as soapy foam. In addition to this danger, unless the bulb of the pump was primed with fluid before use, the first ejection would be the appreciable volume of air

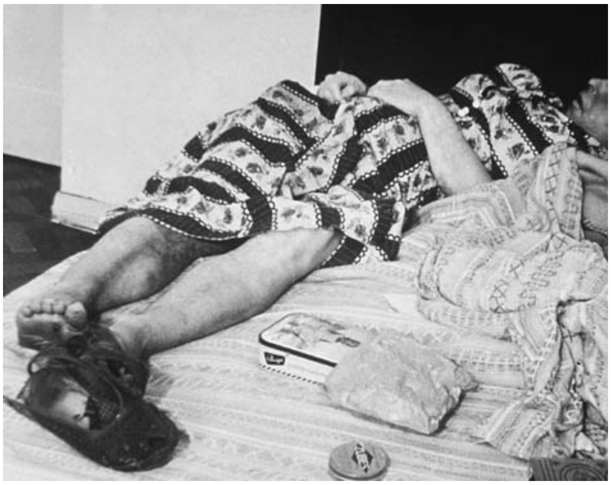

FIGURE 19.1 *The sudden unexpected death of a woman of childbearing age should always arouse the consideration of a complication of pregnancy. This woman insufflated her uterus with a Higginson syringe and had time to clean up the equipment before collapsing with air embolism. This photograph was taken before the British Abortion Act of 1967, as such events are now uncommon.*

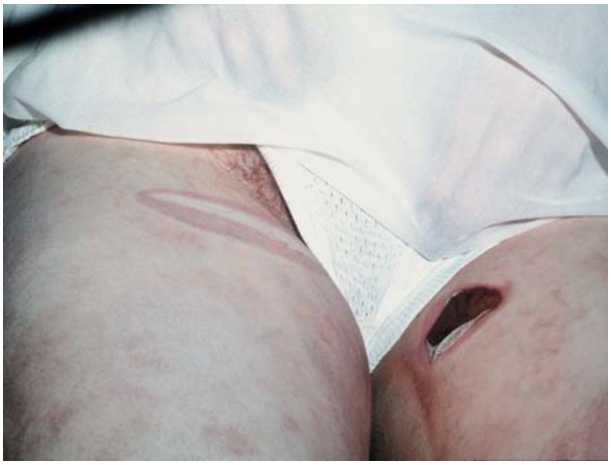

FIGURE 19.2 *Burns from a hot-water bottle on the inner side of the upper thigh in an attempt to resuscitate the victim of a criminal abortion who died from air embolism.*

contained in the empty syringe, which in itself was enough to cause a fatal embolism. Under the considerable pressure available from the instrument (as much as 28 cm of mercury according to Teare 1958), this air was forced into the exposed vascular channels of the placental bed and from there into venous sinuses and pelvic veins to cause cardiac embolism.

Though this occurred almost immediately in most instances, some delay was possible – perhaps because contained air was further squeezed by contraction of the uterus. Some victims were able to leave the premises of an abortionist and survive for up to 2 hours or more before the embolism occurred (Simpson 1951). The addition of various substances, such as phenol, formalin, alcohol and turpentine, to the fluid introduced a toxic element, and in some cases these fluids were extruded into the peritoneal cavity through the uterine tubes, causing a chemical peritonitis.

DILATATION OF THE CERVIX

Another crude method was to introduce a 'tent' into the cervical canal. This was a strip of substance that absorbed water and became greatly enlarged, such as *Laminaria digitale* or slippery elm (*Ulmus fulva Michx*). These vegetable materials are hard and compact when dehydrated, so that a strip about 3–8 cm long could be slid into the cervix. When water was absorbed from the surrounding tissues, the cervical canal became widely dilated and abortion might take place. The risks were perforation of the cervix and also infection, especially if the strip tore into the tissues. The substances used, often crude vegetable material, could be the source of infecting micro-organisms including anaerobes.

PHYSICAL VIOLENCE

Women anxious to lose their pregnancy have resorted to extremes of physical activity and even violence in efforts to dislodge the fetus. Frenzied exercise, horse-riding and severe purging with laxatives were usually ineffective, and some unfortunate women went on to seek violent treatment from husbands or consorts. Punching and kicking of the abdomen were most common, and death from visceral rupture, such as liver, spleen or intestine, has been reported. Ironically, the uterine contents were usually undisturbed.

SYRINGE ASPIRATION

The suction method most often used in therapeutic abortion has also come to be used for illegal operations, mainly by medical or nursing personnel. A large syringe attached to a catheter or length of plastic tubing can produce suction within the uterus sufficient to rupture the chorionic sac and precipitate abortion. The method is safe as long as aseptic methods are used, though, if evacuation is incomplete, some products of conception may be left behind that can form a nidus for infection.

INTRAUTERINE INFECTION

Whatever method is used to evacuate the uterus, the risk of haemorrhage or sepsis is always present. Where an instrument, tent or syringe is used, any remaining fragments of placenta or other products of conception may form a substrate for infection. The organisms involved in sepsis are varied, the most dangerous being non-haemolytic

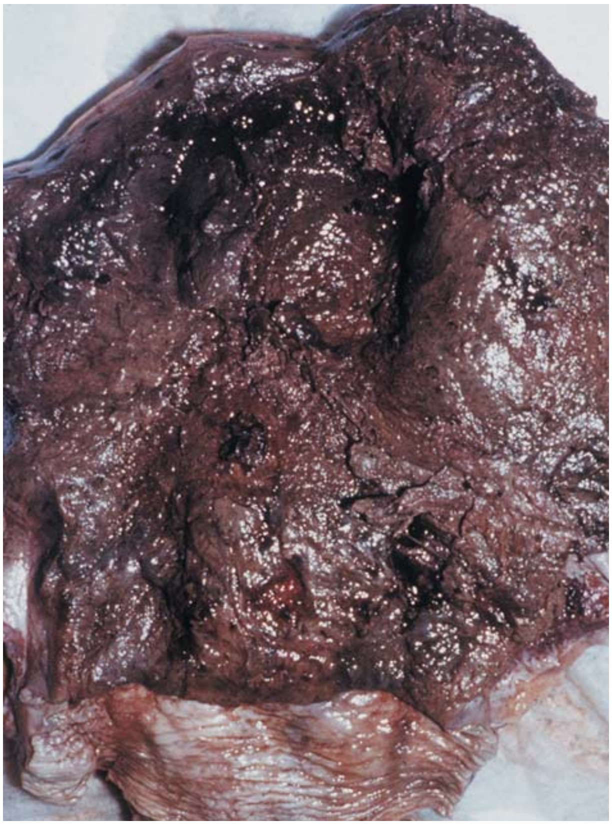

FIGURE 19.3 *Septic, necrotic uterus following an instrumental criminal abortion. The endometrium, containing some retained products of conception, is chocolate brown and foul-smelling, the predominant organism on culture being* Clostridium perfringens.

streptococci and *Clostridium perfringens*, though coliforms and staphylococci may also be responsible.

The uterus becomes swollen, spongy and discoloured. The serosal surface seen at autopsy may be brownish – especially in clostridial infections – and the endometrium may be ragged, foul-smelling and even purulent. Signs of septicaemia may develop with an enlarged, soft spleen, prominent lymph nodes and hepatorenal failure. The kidneys may show bilateral cortical necrosis in extreme cases. In clostridial septicaemias there may be a characteristic bronze coloration of the skin. This may have a mottled or 'raindrop' appearance.

THE USE OF DRUGS AND CHEMICALS

A wide variety of substances, applied locally or taken by mouth, have been used since time immemorial to induce miscarriage. Some have a sound pharmacological basis, others are dangerous, and yet more lie in the realms of folk medicine. Substances applied locally include phenols and Lysol, mercuric chloride, potassium permanganate, arsenic, formaldehyde and oxalic acid. All have their own dangers, both from local corrosion and systemic effects if absorbed. A necrotic pseudomembrane may form in the vagina and severe damage to the cervix may also ensue.

Potassium permanganate was a substance that appeared during the last war and persisted for some years, 650 cases being reported up to 1959, though only a few were fatal. It can cause local necrosis in the vagina and, if absorbed, can have fatal systemic effects including renal failure. Permanganate can cause profuse vaginal bleeding from necrosis, which may give the impression that abortion is threatened and lead an unsuspecting doctor to carry out a curettage. 'Utus paste' is a mixture of soap, myrrh resinoid and potassium iodide, and was used earlier in the century for legitimate abortions, but soon became popular with the criminal abortionist. Another similar product containing elemental iodine, is 'Interruptin'; both were used for late abortions and the induction of labour when the fetus was dead. These pastes are squeezed into the cervical canal through an applicator and the halogen component acts as an irritant which causes eventual expulsion of the chorionic sac. Used medically, they are relatively safe (Barns 1967; Berthelsen and Ostervaard 1959), but fatalities can occur from air or fat embolism or from the toxic effects of absorption.

Substances taken by mouth or injection are legion. The old classification into 'ecbolics' and 'emmenagogues' is now quite redundant and most of the lists of substances are of historical interest only, as their efficacy is either nil or the dose needed to produce abortion is perilously near the fatal level. These include pennyroyal, tansy, rue, savin, laburnum, colocynth, aloes, castor oil, nutmeg, hellebore, cantharides, cotton root, wintergreen and turpentine. Many of these may cause purging, gastrointestinal irritation and general illness if taken in quantity, but have no specific action whatsoever on the uterus. There are other substances that have more chance of causing uterine contractions, though many of them are effective only on a late pregnancy, not at the usual time of 1–3 months when most abortions are sought.

Quinine can be dangerous, as the dosage required for any effect on the uterus is likely to cause cinchonism. Ergometrine has been known from ancient times to lead to abortion, though like most drugs its effectiveness is greater later in pregnancy. Excess dosage may cause peripheral vascular spasm and gangrene. Its availability to doctors and mid-wives is similar to pituitary extract, oestrogens, and – more recently – prostaglandins. Heavy metals, particularly lead, were used for abortion in the past. Plasters coated with a lead compound 'diachylon' were scraped to recover the metallic substances and ingested. Though abortion sometimes occurred, illness and death from acute and subacute lead poisoning was more common; this method has also become of historical interest only. It has been well said that most substances that

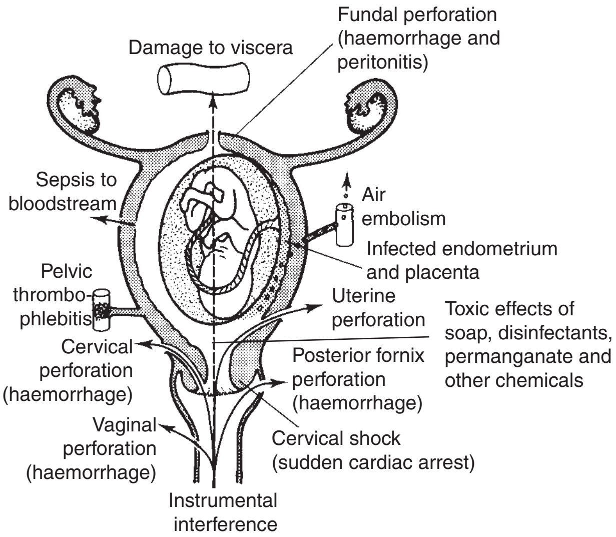
Damage to viscera
Fundal perforation (haemorrhage and peritonitis)
Sepsis to bloodstream
Air embolism
Infected endometrium and placenta
Pelvic thrombo-phlebitis
Uterine perforation
Toxic effects of soap, disinfectants, permanganate and other chemicals
Cervical perforation (haemorrhage)
Posterior fornix perforation (haemorrhage)
Vaginal perforation (haemorrhage)
Cervical shock (sudden cardiac arrest)
Instrumental interference

FIGURE 19.4 *Dangers of a criminal abortion.*

are successful in killing the fetus are likely to kill the mother as well.

A wide variety of medical substances have been used for abortion, including aminopterin, steroids, antimitotic drugs, such as vincristine and ergocornine, and other cytotoxic drugs used in malignant diseases. Even monoamine oxidase inhibitors have been reported as abortifacients. In all these, the autopsy findings in fatal cases will be non-specific or negative, and the history and toxicology provide virtually all the available evidence.

THE AUTOPSY IN ABORTION DEATHS

Where a death in pregnancy or the puerperium occurs, every effort should be made to obtain an autopsy, even if the death does not fall within the usual category of death reportable for medico-legal investigation. When a criminal abortion is suspected, then virtually every jurisdiction will require necropsy.

The usual full autopsy is performed, but a number of additional measures are required, varied to suit the particular circumstances. The fullest possible history is required and, where the death has taken place under medical care, a prior discussion with the clinician is vital. A careful external examination must be carried out, noting especially:

- Abnormal coloration of the skin, such as in the bronzing of clostridial septicaemia and the jaundice of liver damage.
- Signs and duration of pregnancy, such as abdominal swelling and breast changes.
- Signs of injury, including bruising or abrasion of the vulva from instrumentation, and vaginal bleeding.

  Burns on the abdomen or inner sides of the upper thighs may indicate efforts at resuscitation by hot-water bottles when a woman has suddenly collapsed during an illegal abortion.
- The vagina should be examined for signs of recent or current pregnancy, and attempted or successful recent abortion. Any fluid should be removed by pipette to be examined for soap and chemicals, such as antiseptics. A swab should be taken for microbiological culture.
- Where any possibility of air embolism is considered, pre-autopsy radiology of the chest and abdomen must be performed. Some pathologists would consider this obligatory in any death associated with pregnancy, as it is the best method of detecting air embolism by visualizing air bubbles in the heart, great veins in the thorax, inferior vena cava, peritoneal cavity and – possibly – pelvic veins. In the few centres where MRI or computed tomography is available for autopsy material, this may greatly increase the chance of detecting air in the vessels.

Where radiology is not possible, air must be sought by dissection methods, which are not so reliable. The thorax is opened carefully, making sure that removal of the costal cartilages and sternum does not damage the underlying mediastinum or open any of the veins at the thoracic inlet. The cardiovascular system is first examined for the presence of air. The classic method, discussed elsewhere in this book, is to open the heart chambers *in situ* under water poured into the pericardial sac. Escaping bubbles are then said to be indicative of air embolism, though the technique is by no means as reliable as past authors have claimed. In fact, if the post-mortem interval is short, in terms of a few hours, then any fatal air embolism will almost always be quite apparent from the frothy contents of the right ventricle when it is opened in the usual way. Appreciable delay allows absorption of air, as can be proved by the failure to find air in the heart on dissection, even after radiology has clearly indicated its presence. Of course, where any degree of decomposition has set in, which may be on the first day after death in hot countries, gases of putrefaction completely negate any chances of proving air embolism.

Before the organs are disturbed, the great veins of the abdomen should be inspected for bubbles, then the pelvic veins. The uterus itself may be crepitant and bubbles may be seen under the serosal surface or even under the parietal peritoneum of the pelvis. On opening the uterus, gas bubbles may be evident in the wall or placental bed, but this is later in the examination. The usual instructions in most forensic textbooks for opening the head first, carefully removing the calvarium and minutely inspecting the cerebral veins for air bubbles, can be totally disregarded, as this is yet another

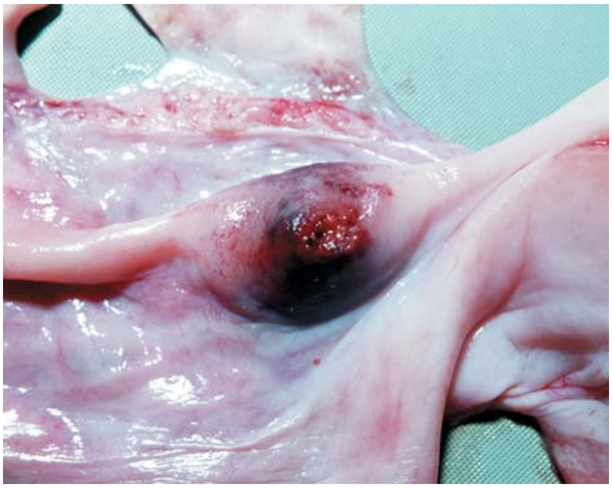

FIGURE 19.5 *Ruptured tuba in (ectopic) pregnancy. Death due to massive intraperitoneal haemorrhage. (Reproduced by kind permission of Professor E Lignitz.)*

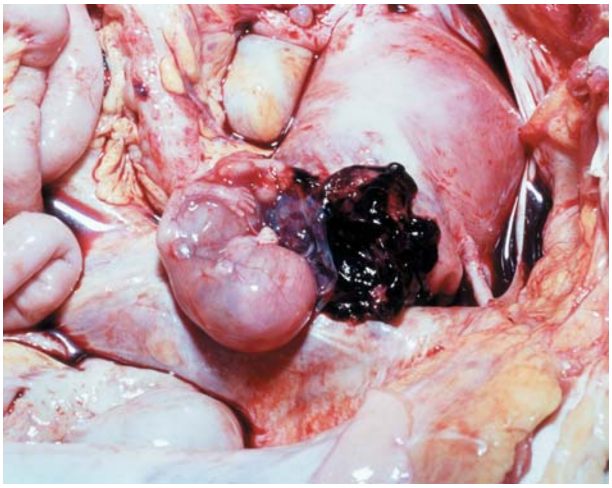

FIGURE 19.6 *Fetus and placenta in the abdominal cavity after ruptured tubal (ectopic) pregnany. Death due to massive intraperitoneal haemorrhage. (Reproduced by kind permission of Professor E Lignitz.)*

example of forensic mythology repeated from one book to the next, without critical evaluation. As discussed in Chapter 13, air embolism cannot occur in the cerebral veins and it was proved many years ago that air bubbles in those veins are artefacts caused by removing the calvarium.

The thorax and upper abdomen should be examined as usual, with particular attention to signs of infection. Blood cultures, and peritoneal and other swabs should be taken as indicated by the circumstances.

The pelvic organs should be removed *en bloc*, as in the examination of fatal rape victims. The lower end of the abdominal incision is carried on each side around the labia and into the cleft of the buttocks to meet beyond the anus. The dissection is taken along the superior pubic rami, which are then carefully sawn through a few centimetres laterally to the symphysis on each side, the cuts being continued down through the inferior rami. The central block of bone is removed and the perineal incision dissected deeply, so that the vagina, rectum and anus, uterus, tubes and ovaries can be lifted out in continuity, and removed for detailed dissection. The vagina is carefully opened along one side to avoid overrunning any midline injuries. Abrasions, bruises, lacerations and any foreign material are noted in the vagina, and any samples kept for chemical or microbiological testing. Any abnormality of the cervix is noted, especially instrument marks, such as forceps or volsellum teeth. The state of dilatation of the cervical canal is noted and the uterus opened, again keeping away from the midline if possible. The colour, size and texture of the uterus is noted, and the state of the interior is naturally vital. If the chorionic sac is still present, its integrity and the attachment to the deciduae are noted. If a fetus is present, this is examined separately for maturity and damage.

If no fetus is present, products of conception are sought and the state of the placental bed noted. Infection, bleeding and air crepitancy are all sought, and the appropriate ancillary investigations made. The tubes are then opened and any foreign fluid collected. The ovaries are examined and the state of any corpus luteum recorded. Finally, extensive histological specimens are taken from all organs and special stains employed where necessary, such as those for seeking amniotic squames in lungs and other organs. Blood, urine, liver and stomach contents are retained for analysis, should this be indicated.

AMNIOTIC FLUID EMBOLISM

Some deaths associated with childbirth or pregnancy are caused by the entry of amniotic fluid into the maternal circulation. The fluid contains fetal squames, lanugo, lipoid from the vernix, meconium, cells from chorion and amnion, and other cellular detritus. The solid elements are usually impacted in the lung capillaries, but rarely have been found in the systemic circulation, including embolization into the kidney, liver and brain. In addition to the solids, the actual fluid itself (which is naturally undetectable histologically) is blamed for the 'allergic' response that may cause such profound collapse, leading to death. Amniotic fluid embolism is also one of the most potent causes of disseminated intravascular coagulopathy, mentioned also in Chapter 13.

A variety of causes exist apart from the pelvic trauma of parturition, including rupture of the uterus, but instrumental interference late in pregnancy may also allow opening of the sinuses in the placental bed with consequent escape of amniotic fluid. Further details of the causes and mechanisms must be sought in obstetric works, but from the autopsy point of view, it is important both to bear the possibility in mind and to have techniques that will reveal the embolism after autopsy. The diagnosis is histological and depends upon identifying the squames detached from the fetal epidermis, which float freely in amniotic fluid. These can be seen in ordinary haematoxylin and eosin stained sections of the lung, but are much better visualized and a greater yield obtained if special staining methods are used, as described in Appendix 1. Fibrin deposition in many organs may occur as a result of the disseminated intravascular coagulation and careful search for this must be made, using fibrin stains. Immunohistochemical techniques have also been used in the lung sections to demonstrate human keratin-, meconium- and amniotic fluid-derived mucin as well as isolated trophoblastic cells in deaths due to AFE (amniotic fluid embolism) syndrome.

REFERENCES AND FURTHER READING

Adebahr G. 1969. [The forensic value of findings in the cadaver.] *Beitr Gerichtl Med* **25**:44–50.

Anonymous. 1999. Abortion in fact: levels, trend and patterns. In: *Sharing responsibility: women, society and abortion worldwide.* Alan Guttmecher Institute, New York, pp. 25–31.

Attwood HD. 1956. Fatal pulmonary embolism by amniotic fluid. *J Clin Pathol* **9**:38–46.

Baekgaard T. 1969. [Penalties for criminal abortion.] *Ugeskr Laeger* **131**:1069–70.

Barns H. 1967. Therapeutic abortion by means of soft soap pastes. *Lancet* **2**:825.

Benson MD, Kobayashi H, Silver RK, *et al.* 2001. Immunologic studies in presumed amniotic fluid embolism. *Obstet Gynecol* **97**:510–14.

Berthelsen H, Ostervaard E. 1959. Techniques and complications in therapeutic abortion. *Dan Med Bull* **6**:105–10.

Bowen DA. 1967. Medical investigation in cases of sudden death. *Br Med J* **2**:33–6.

Bowen DA. 1980. Criminal abortion. Fact or fiction? *Am J Forensic Med Pathol* **1**:219–21.

Breitenecker L, Husslein H. 1968. [Expert testimony as proof of previous pregnancies.] *Wien Klin Wochenschr* **80**:937–42.

Cameron J, Dayan A. 1960. Association of brain damage with therapeutic abortion induced by amniotic fluid replacement; report of two cases. *Br Med J* **1**:1010.

Cera G, Fruttero A, Rua S, *et al.* 1992. Flow cytometric studies in spontaneous abortions. Applications in the medico-legal practice. *Forensic Sci Int* **54**:167–75.

Cimbura G. 1967. Studies of criminal abortion cases in Ontario. *J Forensic Sci* **12**:223–9.

Davies S. 2001. Amniotic fluid embolus: a review of the literature. *Can J Anaesth* **48**:88–98.

Dayan AD, Cameron JM, Phillipp E. 1967. Fatal brain damage associated with therapeutic abortion induced by amniocentesis: report of one case. *Med Sci Law* **7**:70–2.

De Gennaro F. 1969. [Criminal abortion. A clinical case and medicolegal aspects.] *Minerva Ginecol* **21**:1002–5.

Demol R, Vandekerckhove D, Thiery M. 1968. [Various ‘traps’ in diagnosing pregnancy reactions.] *Med Leg Dommage Corp* **1**:282–9.

Eckert WG. 1985. The writings of Sir Bernard Spilsbury: Part II. *Am J Forensic Med Pathol* **6**:31–7.

Eliakis E, Coutselinis A. 1971. [Identification of meconium by the determination of acid phosphatase and spectrophotometric analysis.] *Med Leg Dommage Corp* **4**:163–8.

Fineschi V, Gambassi R, Gherardi M, *et al.* 1998. The diagnosis of amniotic fluid embolism: an immunohistochemical study for the quantification of pulmonary mast cell tryptase. *Int J Legal Med* **111**:238–43.

Garland IW, Thompson WD. 1983. Diagnosis of amniotic fluid embolism using an antiserum to human keratin. *J Clin Pathol* **36**:625–7.

Gastmeier G, Klein A, Falk H. 1969. [Contribution to inverse soap poisoning.] *Dtsch Z Gesamte Gerichtl Med* **65**:96–102.

Gerchow J, Schewe G. 1970. [Legal theory of value assessment and medical aspects of the need to protect the beginning of life.] *Beitr Gerichtl Med* **27**:61–71.

Holzer FJ. 1973. [Abortion using a bicycle pump on the mistress and unusual suicide of a blind man.] *Beitr Gerichtl Med* **30**:187–96.

Hunt AC. 1960. Amniotic fluid embolism. *J Forensic Med* **7**:74–7.

Huntington K. 1976. Forensic gynaecology. *Practitioner* **216**:519–28.

Huxley AK, Sibley MA. 1998. Alleged forgery of sonography report leads to elective abortion of late 23 week-old fetus. *J Forensic Sci* **43**:218–21.

Inoue H, Takabe F, Maeno Y, *et al.* 1989. Identification of fetal hemoglobin in blood stains by high performance liquid chromatography. *Z Rechtsmed* **102**:437–44.

Janssen W. 1967. [On the pathogenesis and forensic evaluation of brain hemorrhages following cerebral air embolism.] *Dtsch Z Gesamte Gerichtl Med* **61**:62–80.

Kenyeres I. 1976. [Various problems concerning maternal mortality during pregnancy.] *Morphol Igazsagugyi Orv Sz* **16**:120–5.

Khong TY. 1998. Expression of endothelin-1 in amniotic fluid embolism and possible pathophysiological mechanism. *Br J Obstet Gynaecol* **105**:802–4.

Kobayashi H, Ooi H, Hayakawa H, *et al.* 1997. Histological diagnosis of amniotic fluid embolism by monoclonal antibody TKH-2 that recognizes neuac alpha 2–6galnac epitope. *Hum Pathol* **28**:428–33.

Kocsis I, Argay I, Toth G. 1972. [Course of disease in an unconscious patient causing several problems.] *Morphol Igazsagugyi Orv Sz* **12**:209–11.

Kopecny J, Martincik M, Chalupa M, *et al.* 1970. [Induction of abortion using high doses of oxytocin.] *Cesk Gynekol* **35**:118–19.

Kumar GP, Kumar UK. 1994. Estimation of gestational age from hand and foot length. *Med Sci Law* **34**:48–50.

Kumar KU, Pillay VV. 1996. Estimation of fetal age by histological study of kidney. *Med Sci Law* **36**:226–30.

Kurucz E, Kosa F, Monostori E, *et al.* 1984. A new method in criminology: use of ELISA to detect AFP on different materials with monoclonal anti-alpha-fetoprotein. *Z Rechtsmed* **93**:117–21.

Lewis G. 2001. *Why mothers die. Report on confidential enquiries into maternal deaths in the United Kingdom 1997–1999.* RCOG Press, London.

Lewis G, Drife JO, Botting BJ, *et al.* 1998. *Why mothers die: report on confidential enquiries into maternal deaths in the United Kingdom 1994–1996.* TSO, London.

Lorente Acosta JA, Lorente Acosta M, Villanueva Canadas E. 1998. [The physician and medicine in the new penal code.] *Med Clin (Barc)* **111**:222–5.

Lunetta P, Penttila A. 1996. Immunohistochemical identification of syncytiotrophoblastic cells and megakaryocytes in pulmonary vessels in a fatal case of amniotic fluid embolism. *Int J Legal Med* **108**:210–14.

Mallach HJ, Pfeiffer KH. 1978. [The importance of air embolism as a primary and secondary cause of death.] *Med Welt* **29**:1391–6.

Marcinkowski T, Przybylski Z. 1968. [Teratoma as a cause of unjustified suspicion of abortion.] *Patol Pol* **19**:135–8.

Möttönen M, Isomäki AM. 1971. Amniotic fluid embolism, diagnosed by a new method. *Med Sci Law* **11**:35–7.

Nishio H, Matsui K, Miyazaki T, *et al.* 2002. A fatal case of amniotic fluid embolism with elevation of serum mast cell tryptase. *Forensic Sci Int* **126**:53–6.

Ohi H, Kobayashi H, Terao T. 1993. [A new histologic diagnosis for amniotic fluid embolism by monoclonal antibody TKH-2 that recognizes mucin-type glycoprotein.] *Nippon Sanka Fujinka Gakkai Zasshi* **45**:464–70.

Patel F. 1993. Artefact in forensic medicine. Fetal congenital abnormality. *Am J Forensic Med Pathol* **14**:212–14.

Rees PO. 1970. A case of poisoning by Epsom salt. *J Forensic Sci Soc* **10**:91–4.

Richard T, Rappolt MD, Sr. 1970. Clinical observation: apparent antagonism of ergotamine on the gravid uterus by LSD-25. *Eur J Toxicol* **3**:138–9.

Ronnau H, Wille R. 1980. [Psychological complications following abortion.] *Beitr Gerichtl Med* **38**:21–4.

Rosenzweig E, Stajduhar-Caric Z, Milicic D, *et al.* 1970. [Incidence of uterine perforations in criminal abortions.] *Rad Med Fak Zagrebu* **18**:221–4.

Russell DH. 1968. Law, medicine and minors. I. *N Engl J Med* **278**:35–6.

Schwerd W. 1965. The detection of soap abortions. In: Curry AS (ed.), *Methods of forensic science*, vol. 4. Interscience Publishers, London.

Shapiro HA. 1966. The diagnosis of death from delayed air embolism. *Acta Med Leg Soc (Liege)* **19**:207–11.

Simpson K. 1951. Air embolism. *Med Leg J* **19**:81–5.

Teare D. 1958. Death in criminal abortion. *Med Leg J* **26**:132–4.

Tesar J, Klir P, Doskocilova L. 1977. [Demonstration of neonatal blood.] *Soud Lek* **22**:9–11.

Toner PG. 1992. *The role of the histopathologist in maternal death. Maternal mortality – the way forward.* Royal College of Obstetricians and Gynecologists, London, Chapter 10.

Uszynski M, Zekanowska E, Uszynski W, *et al.* 2001. Tissue factor (TF) and tissue factor pathway inhibitor (TFPI) in amniotic fluid and blood plasma: implications for the mechanism of amniotic fluid embolism. *Eur J Obstet Gynecol Reprod Biol* **95**:163–6.

Zakharova OA, Bakshinskaia RE. 1968. [Mortality significance of acute renal insufficiency in criminal abortion.] *Sud Med Ekspert* **11**:18–20.

CHAPTER 20

Infanticide and stillbirth

- Stillbirth 439
- Functions of the pathologist in infanticide 440
- Definition of 'live birth' and 'separate existence' 442
- Concealment of birth 442
- The autopsy on a suspicious death of a newborn infant 442
- Infanticide by omission to offer proper care 445
- Proof of a separate existence 445
- Estimation of maturity 448
- References and further reading 449

Historically, infanticide has by no means always been a crime, and has been practised as a social and economic necessity since the dawn of humanity. The pathologist, however, has to deal with the practicalities of infanticide as they exist today. Though legal definitions vary among different countries, the medical concept of infanticide is uniform, being the deliberate killing of a newborn infant by the mother. In England and Wales, infanticide was not accepted as a legal entity in the criminal law until 1922. Before then all child killing was murder – and the convicted mother was liable to be sentenced to death until the Infanticide Act, which was re-enacted in a more satisfactory form in 1938 (see account by Kellett 1992).

In many other countries there is no specific infanticide legislation – even Scotland did not adopt the English statute, having a separate crime of 'child murder', but long before 1922 that country had held a more enlightened attitude towards the mentally stressed mother. In Japan, Funayama *et al.* (1994) have described cases of multiple, repeated infanticides by five mothers.

This book generally avoids territorial legal comparisons, but the foregoing facts serve to illustrate that the killing of a young child by the mother is a well-recognized medical and psychiatric phenomenon that distinguishes it from other types of homicide. The penalties should therefore be different and there is a profound responsibility laid upon the forensic pathologist called to examine such infants, to investigate the death with particular caution.

The substance of infanticide that has relevance to the pathologist is well set out in the English Infanticide Act 1938 (Section 1):

> Where a woman by any wilful act or omission causes the death of a child being under the age of 12 months, but at the time the balance of her mind was disturbed by having given birth or the effect of lactation, she may be dealt with as if she had committed manslaughter.

It should be noted that:

- Only the mother can benefit from this alternative to murder – not the father or any other person.
- The child has to be less than a year old, though in fact the majority of infanticides occur within hours or even minutes of birth.
- It has to be a 'child' – that is, a person with a separate existence living independently fully outside the mother's body.
- It must have died because of a wilful (deliberate) act of omission or commission.

All these provisos have an effect on the pathologist's task. Even in those jurisdictions where there is no formal infanticide legislation, there is usually recognition that the acts of a woman soon after childbirth, often in adverse conditions, are different from common murder.

STILLBIRTH

To clarify later definitions, a stillbirth is defined in English law and, from a medical point of view no doubt has a broadly similar description in other jurisdictions. In England and Wales, a baby is stillborn if after 24 weeks' gestation, 'it did not at any time after being completely expelled from its mother, breathe or show any other sign of life'. Theoretically, this is a medically unsatisfactory definition, as it could be alive when its head was born but die

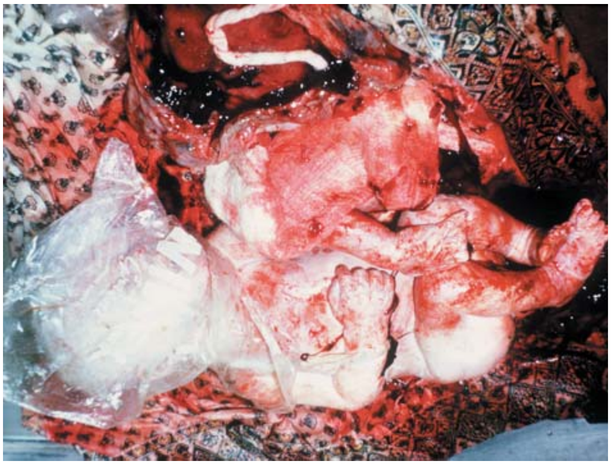

FIGURE 20.1 *Newborn baby with umbilical cord and placenta, abandoned in a car park. It was wrapped in a curtain and though the head was in a plastic bag no signs of suffocation nor of live birth could be detected. The mother was never identified.*

before completion of expulsion, and therefore legally be a stillbirth. However, in practice this would be rare, as most stillbirths either die *in utero* or during the earlier stages of parturition. Stillbirths occur in about one in every 18 legitimate pregnancies in Britain, and are more frequent in illegitimate births and socially disadvantaged families.

FUNCTIONS OF THE PATHOLOGIST IN INFANTICIDE

To assist in identifying the mother, if she is unknown

Newborn infants found dead (especially in clandestine circumstances) are not necessarily the victims of infanticide. Those stillborn or dying naturally or from unintentional lack of care may be hidden or abandoned, a minor crime in Britain called ‘concealment of birth’. The autopsy, as well as artefacts discovered with the body, such as bags, blankets and newspapers can assist to a greater or lesser degree in the police search for the mother.

The apparent ethnic group of the child will be disclosed and blood grouping can be performed, which may assist by eliminating or helping to confirm the consanguinity of any putative mother. In the future, DNA profiling will add greatly to identification. The appearance of the cut end and any ligation of the umbilical cord (see below) may help to decide whether the birth was one where medical, nursing or only amateur attention was available. In spite of these clues, most abandoned newborn corpses fail to be linked to the mother, delay in discovery making both identification and evidence of live birth unattainable.

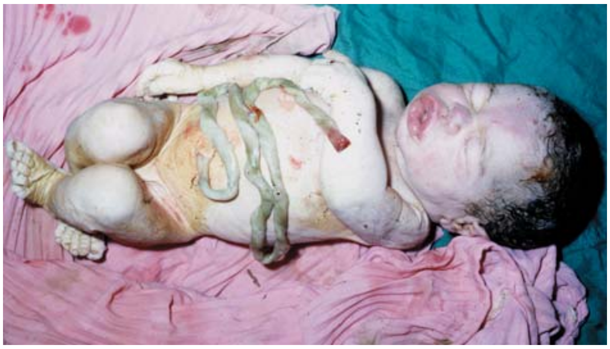

FIGURE 20.2 *Full-term newborn found wrapped in a blanket inside a shopping bag in a refuse container. The cord had been cut with a sharp instrument, but the child was not washed or fed. The lungs showed no positive evidence of respiration.*

To estimate the maturity of the child

Though any newborn infant, whatever the length of gestation, can be the victim of infanticide if born alive, it is obviously relevant to this issue if the fetus is too immature to have survived birth. In another context [Infant Life (Preservation) Act 1929, concerning child destruction], the age of viability in English law was taken as 28 weeks’ gestation, though this has now been reduced to 24 weeks as even younger fetuses can now survive with intensive medical support. If a child is shown to be that premature, there is a strong presumption that it would not long survive a birth away from medical attention.

To determine whether or not the child was stillborn and whether it had a separate existence

There is a presumption in English law that all infants are stillborn and, when a woman is charged with infanticide, the burden of proof is upon the prosecution to demonstrate that the child had a separate existence. Adelson (1974) states that the same legal situation exists in the USA, and gives the following useful and interesting commentary:

> Unless the pathologist has incontrovertible criteria of post-natal survival, e.g. well-expanded lungs, food in the stomach or vital reaction in the stump of the umbilical cord, he is legally bound not to diagnose live birth. Convictions for infanticide have been set aside when there has been any doubt whatsoever that the child was born alive. Many courts have pushed this proposition to the extent that the State has been given the burden of proving that the baby was born alive beyond any

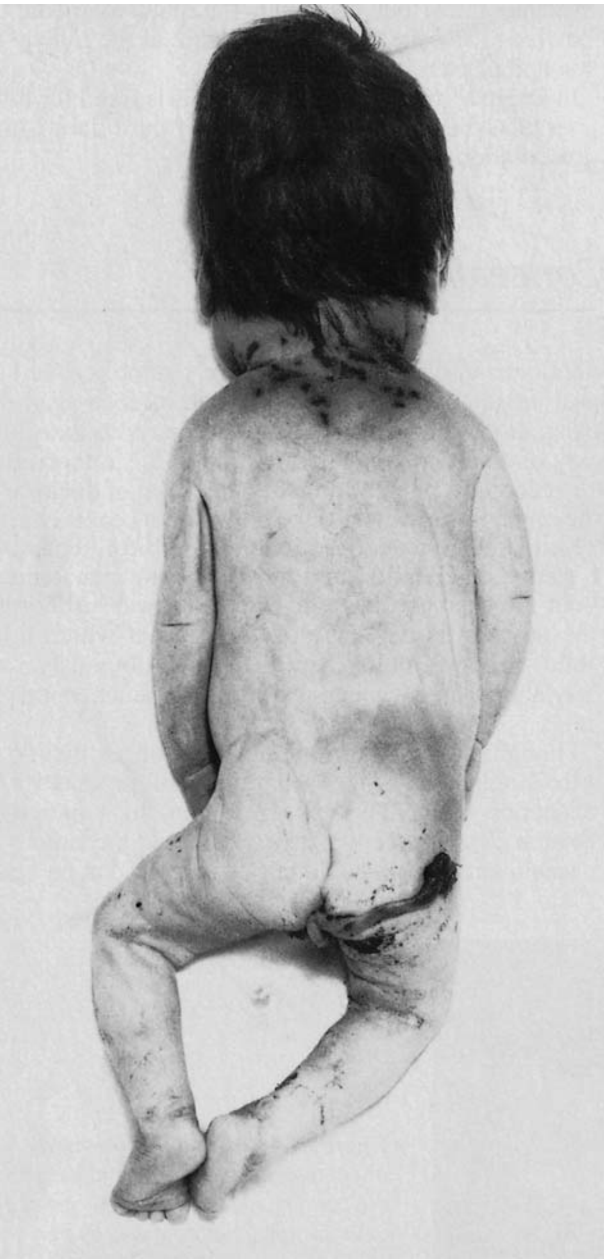

FIGURE 20.3 *A stillborn child concealed by the mother after birth. The injuries around the neck were viewed with suspicion by the police, but are the result of fingernails being used in an attempt to assist self-delivery in an unattended birth.*

possible doubt rather than beyond any reasonable doubt, the latter being the general level of proof required in a criminal prosecution.

Herein lies the most difficult task for the pathologist, one which has given rise to considerable controversy in forensic pathology.

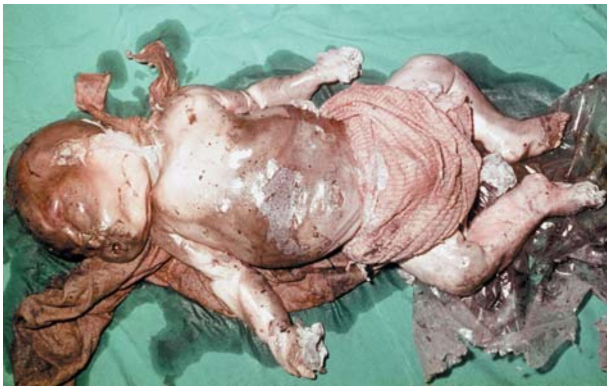

FIGURE 20.4 *A putrefied newborn infant disposed of in a plastic bag in the countryside. There is no hope of determining whether it was live or stillborn with this degree of post-mortem decomposition. There was a strip of fabric knotted around the neck, but it is impossible to decide if that was related to the death, if the child was not stillborn.*

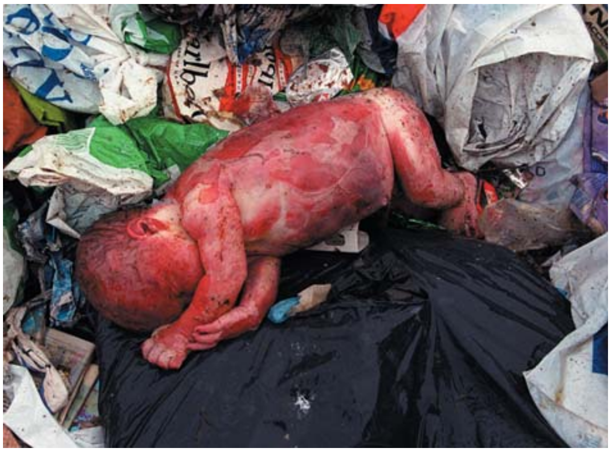

FIGURE 20.5 *A decomposed newborn infant found in a rubbish dump.*

To determine whether its death was due to natural causes or to any act of omission or commission

Where the pathologist claims to have proved separate existence (or at least, live birth), he has an additional burden in that he must show that death occurred from an act of commission or omission, The ‘wilful’ aspect is a matter for the lawyers, but it is for the pathologist either to demonstrate lethal injuries or to prove that some lack of care led to the death – often an impossible task. There is a considerable burden upon the investigating authorities when finding a hidden newborn infant if they wish to pursue an infanticide charge.

First, they must find the mother – often another impossible task. Then the pathologist must prove live birth, which is often impossible, especially in a decomposed body. To prove a separate existence is even more difficult and is rarely a purely medical decision. The firm diagnosis of death from wilful inattention is equally hard to make.

It is small wonder that only a small proportion of suspicious neonatal deaths ever reach the court – where few women are convicted, and those who are, rightly receive lenient and sympathetic treatment, usually consisting of probation and psychiatric support.

DEFINITION OF 'LIVE BIRTH' AND 'SEPARATE EXISTENCE'

There is no definition of live birth in English law, except by inference from the defined 'stillbirth', given earlier. Using the live birth as the converse of a stillbirth, the infant should have attained the gestational age of 24 weeks, but this is not a factor in 'separate existence' with which live birth is almost synonymous.

A child with a 'separate existence' would be an infant that had issued forth from its mother irrespective of its gestational age, and who breathed or showed any other signs of life after being completely expelled from the maternal passages – though the child could still be attached to the placenta within the mother. This is almost synonymous with a live birth, though medically all would agree that a child that was crying lustily, but with a foot still in the vagina, was undoubtedly live born even though it did not satisfy the legal criteria of a 'separate existence'.

Viability

Yet another medical and legal concept is that of viability, which is the potential ability of an infant to survive after birth. An immature fetus may be born alive, but be incapable of maintaining an independent existence. Naturally the stage of gestation at which this occurs varies not only with the condition of the particular fetus, but also with the state of medical knowledge and the facilities available at the relevant time and place.

In English law, a period of 24 weeks is fixed for the onset of viability, for the purposes of the Infant Life (Preservation) Act, 1929.

CONCEALMENT OF BIRTH

Mentioned earlier, this is a relatively minor crime in most jurisdictions and is basically an offence against registration, and perhaps public decency, in that the body of an infant is neither notified to the authorities nor accorded a proper burial or other form of disposal. The crime is the hiding of the body to conceal the fact of birth. It does not depend at all on live birth, separate existence or viability, and the law is not concerned about the cause of death or whether it was stillborn. One problem is the stage of gestation at which the 'child' becomes noticed by the law – obviously an early abortion is of no concern to anyone, except the family.

Though not specified in relation to concealment of birth, English law takes no regard of products of conception under 24 weeks' gestation. In a judicial decision (*R* v *Matthews*), it was held that a charge of concealment of birth could not be sustained at an age where the law did not require the (still)birth to be registered.

Whatever the legal ramifications, the forensic pathologist is not infrequently asked to examine dead infants in which he can neither prove live birth nor a wilful act that might have caused death. The only remaining charge, if he can show an age above that of viability, is 'concealment of birth'. This is the end result of the discovery of most decomposed babies, in the minority where the mother can be traced.

THE AUTOPSY ON A SUSPICIOUS DEATH OF A NEWBORN INFANT

A number of important matters need specific attention, over and above the usual meticulous autopsy procedure on any infant.

The coverings and any other associated articles with the infant must be examined and retained. Though this is mainly police business, the pathologist usually has the task of unwrapping the baby – often in a decomposed and offensive state – from the coverings of newspaper, plastic bags, rags or blankets. He must recover any foreign material or objects that may assist in the identification of the mother when the disposal has been clandestine.

When the baby has been found in a house or other building, he may need to visit the scene, preferably with the child still *in situ*. Numerous cases are on record of the infant being in a lavatory pan and, if there are head injuries or drowning, the circumstances must be evaluated.

The external examination is, as always, important. It is vital to assess the degree of putrefaction because if it is in any way decomposed, it will almost certainly be impossible to determine whether live birth had occurred. Decomposition must be distinguished from intrauterine maceration, as the latter is definite proof of stillbirth. If death occurred within 2–3 days before expulsion from the uterus, the appearances may be fairly normal, apart from general softening and histological evidence of general cellular autolysis.

When it has been dead for many days, the macerated fetus is usually a brownish-pink, rather than the greenish hue of putrefaction. The surface is slimy, blistered, desquamating and sometimes almost jelly-like. The joints are grotesquely loose and the cranial plates may be virtually detached beneath the scalp. Their over-riding, radiologically seen as 'Spaulding's sign', is characteristic, so that the head is shapeless. Rarely, a macerated fetus – which is normally sterile – may become infected if not expelled after the membranes rupture; it will then be born in a putrefied state.

External examination will next be directed at the cord and placenta, if present. The latter must be measured and weighed to estimate the maturity and any abnormalities, such as placental infarcts, noted as possible reasons for fetal death. The cord is important as an indicator of separate existence, if survival has been for an appreciable period. Even where early putrefaction renders evaluation of breathing impossible, vital signs in the cord may indicate live birth if survival reached 24–48 hours. Until this period, no evidence of cord separation is reliable, but after a day or so, a ring of reddening appears around the base of the cord and adjacent abdominal skin. This becomes more marked in the succeeding few days as the cord dries and shrivels, detachment taking place between 5 and 9 days. As most infants die or are killed within hours of birth, however, these signs are rarely of much practical use.

The severed end of the cord should be examined, as it may indicate whether the cord had been broken or actually cut. This is occasionally important if death is the result of head injuries and the defence on behalf of the accused mother claims that the child fell to the ground head first during a precipitate delivery. The length of the cord must be measured, especially if the placental segment is available, so that the distance from the vulva to the floor can be calculated. The average cord length is about 50 cm, but more than twice this length is not uncommon, as are much shorter ones. Naturally, if a ligature is present it indicates either the attendance of a professional or at least knowledgeable person, but this is rare in these cases. The severed end may be examined in the hand, but may also be floated out in water, to better study the detail of the detachment. Morris and Hunt (1966) conducted experiments on cords and determined that they could easily be broken by hand traction. A broken cord can show a clean transverse termination, but is usually ragged. If cut by a sharp instrument such as knife or scissors the cut may be clean, but may also be ragged if the instrument is blunt.

If the baby is examined as found, either at the scene or undisturbed in wrappings, it should be noted if the cord was coiled around the neck. Although it is usually unsafe to claim that death during delivery was from pressure on the neck from the cord, the fact must be noted and evaluated along with all other findings.

The skin may still have vernix caseosa present. This is not a useful observation but, if none is present, it may indicate that the child had been washed, which suggests survival for some time after birth.

Measurements are vital for an estimate of maturity. Weight, crown–heel, crown–rump, head circumference, and foot length are required.

Abnormalities incompatible with continued life may be obvious, such as anencephaly, severe spina bifida or abdominal ectropion. It is, however, often unsafe to assume that actual live birth, as legally defined, could not have taken place.

Injuries are obviously of prime importance as evidence of the 'act of commission', though most newborns seen by the pathologist will be free of such trauma. The types of injury perpetrated by distraught mothers are numerous:

- Strangulation presents the usual features of bruises and abrasions on the neck, though these may be minimal. The classical features of facial congestion, cyanosis, oedema and petechiae may be present, but are often absent, presumably because of the ease with which the vulnerable infant dies. Abrasions on the neck, however, may also be caused by the frantic efforts of the mother to deliver herself, leaving finger-nail marks on the infant's neck. A ligature may be used, which may still be in place. Again caution must be employed, as the fact of some object around the neck, though suggestive,

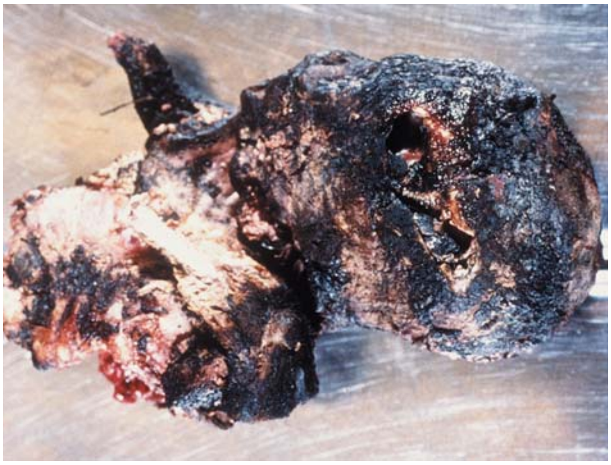

FIGURE 20.6 *Remains of a newborn infant, who was born during a weekend when the mother, a schoolgirl, was alone in her room in a boarding-school. After the birth she wrapped the child in a plastic bag. The following day she left for home and tried to burn the body in a stove but when that did not succeed she buried the partly burned body in a snowdrift where it was later found by a dog.*

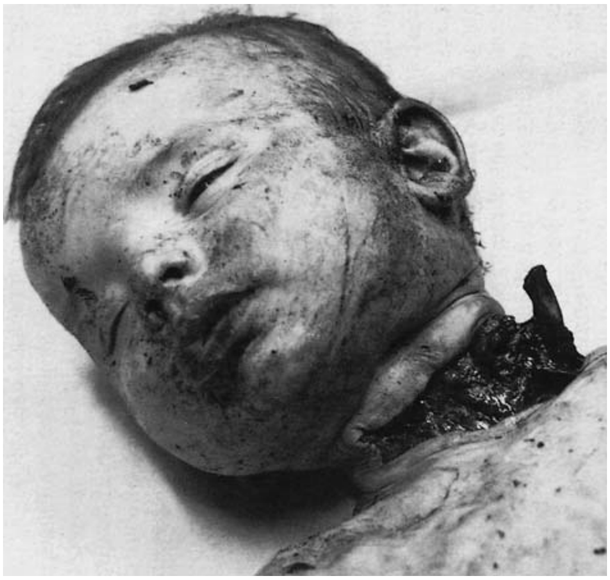

FIGURE 20.7 *Definite infanticide, the child having fully respired before having this severe incised wound of the neck made with a kitchen knife.*

is not absolute proof of strangulation unless corroborative evidence of congestive and petechial signs in the face or skin, and tissue damage beneath the ligature is demonstrated. The explanation has been advanced in the past that a cloth ligature around the neck was an attempt to assist self-delivery by the woman in an obstructed labour.

- Smothering is almost impossible to prove, as the infant will almost never show conjunctival petechiae or other signs that may be seen in strangulation. Only if excessive pressure is applied, so that marks are left on the lips and face, can any evidence be retrieved. These must be definite intradermal or deep bruises, abrasions or marks within the lips and mouth – and not shadowy, vague variations in post-mortem staining. The arguments about criteria of suffocation in infants are further discussed in Chapter 14. Choking by the internal obstruction of the pharynx by various materials has been described in old reports, but it is excessively uncommon these days.
- Cutting and piercing is well known as a means of infanticide. Stabbing with scissors, either in the chest or neck, has been seen or even cutting the throat with a blade. The defence may be offered that the wounds were inflicted accidentally during frenzied efforts by the mother to cut the umbilical cord. Though not seen in Europe, deliberate infanticide has been described in India by clandestine stabbing by a long needle or pin into the spine, fontanelle, eye, or nose.

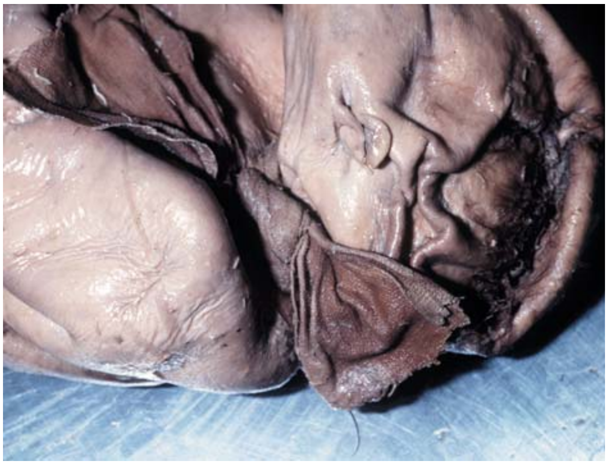

FIGURE 20.8 *A putrefied newborn infant with a scarf knotted tightly around the neck. (Reproduced by kind permission of Professor E Lignitz.)*

- Head injuries are relatively common. The mother may throw the child to the floor, or dash its head against a wall or other obstruction, sometimes by swinging it by the legs. This mode of death is similar to child abuse in older infants. The defence may be raised that the child fell to the ground, either from the mother's arms, but especially during a precipitate birth from the standing or crouching position. Though this defence sounds like a desperate excuse, the author can attest – no doubt in common with all doctors who have practical experience of childbirth – that some deliveries, especially in multiparous women, can occur with considerable speed and force. This is why the length of the cord needs to be measured, as it is obviously an effective brake on such a descent. There is no doubt, however, that cords have broken during precipitate birth, a reason for a careful inspection of the severed ends at every autopsy.

  The lavatory pan has special significance in this respect, as many unexpected births have taken place whilst the woman is sitting on the toilet. A common story is that the pregnant woman, often a teenage girl – is unaware that she is pregnant and goes to the toilet under the misapprehension that she has stomach pain from constipation. There is no doubt that some naive girls are genuinely ignorant of their gravid condition, and are shocked and devastated to the point of blind panic when a baby emerges into the pan.

  The question may then arise as to whether fatal head injuries on the infant can be accounted for by the descent into the lavatory. It is unlikely to produce the severe fracturing sometimes seen – and falling to the

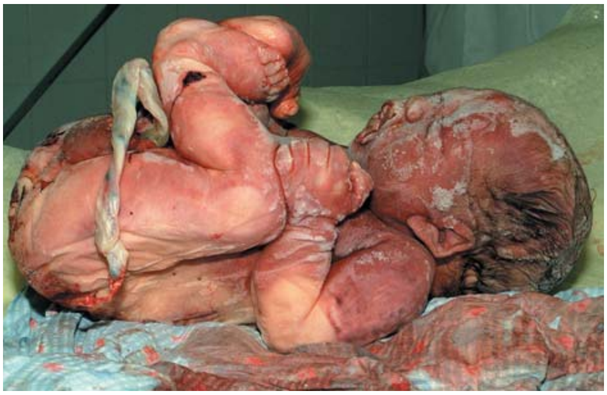

FIGURE 20.9 *Frozen newly born infant found in a freezer by the children of the family. The child and placenta were packed and frozen in separate plastic bags, the umbilical cord is still attached. The family had moved twice after the birth of the child.*

floor in a precipitate delivery is also unlikely to produce fatal damage. Much depends on the severity and nature of the scalp and skull injury, and although it cannot be said that it is impossible for a fatal head injury to be caused in this way, it must be a rare event. Minor fractures of the parietal bones may occur in natural delivery if there is excessive head moulding. These are usually unassociated with internal brain or meningeal damage, and naturally there is no associated scalp damage.

- Drowning is another unusual form of infanticide, but is more often a way of disposing of an already dead child, either stillborn, a natural death or the victim of some other mode of infanticide. It may take place in any form of water, from a washbasin to the open sea. Most often the mother will use a household receptacle such as bowl, bucket or bath, but the infant may be taken out and disposed of in any source of open water, where delay in recovery may add decomposition to the pathologist's difficulties. In a recent case seen by the author (BK), the young parents admitted that the child was live born. After some hours, they wrapped the infant in a blanket with a brick and dropped it into a river. The partial remains, largely converted to adipocere, were recovered some 3 months later, there obviously being no prospect of offering any opinion upon either live birth or the cause of death. The only objective, but equivocal evidence, was the recovery of diatoms from the femoral bone marrow, which resembled those in the river water. Returning to the lavatory pan once again, cases are on record where the birth into the toilet led not to head injuries, but to drowning.

INFANTICIDE BY OMISSION TO OFFER PROPER CARE

This is both rare and almost outside the province of the pathologist's examination. If the infant is older than newborn, then deliberate starvation will leave obvious signs, but the proof of lack of care immediately after birth will be a matter of witnesses and clinical opinion. The only possible signs which the autopsy might reveal would be hypothermia (Chapter 17), and obstruction of the air passages by mucus or amniotic fluid.

PROOF OF A SEPARATE EXISTENCE

Strictly speaking, the pathologist can never demonstrate evidence of a separate existence in the English legal sense of complete expulsion from the mother, but it is usually taken to mean proof of live birth, specifically the attainment of breathing air. In the slightly older child the presence of food in the stomach would obviously prove a separate existence, but most autopsies are upon the recently newborn.

Changes in the cord have already been mentioned, but are of no use in the immediate post-partum period. Polson *et al.* (1985) make mention of the presence of 'extraneous material' in the deeper (secondary) bronchi as evidence of a stillbirth, on the grounds that, if this material, such as soil or sand, gained entry after respiration had commenced, the air in the deeper areas of the lung would prevent its ingress. The practicality of this appears suspect.

The unequivocal demonstration of breathing in a newborn infant is fraught with difficulty. Along with the estimation of the time of death, it has probably provoked more discussion, printed words and controversy than any other topic in forensic medicine. The controversy revolves around the 'hydrostatic test' on the lungs, in which it has been claimed since medieval times that, if the lungs float in water, then the infant breathed. There are modifications of technique, but this is the thrust of the claim. The authors' attitude must be stated at the outset: the test is of limited value, whatever modifications are made and it can at best be a suggestive pointer, but never a definitive test in itself.

There are too many recorded instances when control tests have shown that stillborn lungs may float and the lungs from undoubtedly live-born infants have sunk, to allow it be used in testimony in a criminal trial. Even one such failure negates the whole history of the test and the authors are saddened to contemplate the number of innocent women who were sent to the gallows in previous centuries on the testimony of doctors who had an uncritical faith in this crude technique. As this is such an important issue and one that is still contested today, the words of the late Professor Polson may be recalled from his notable textbook:

> The test was suspect even in 1900 and requires no detailed discussion, because it is now known to have no value. The lungs of the live-born, even those who have been known to live for days, may sink [Dilworth 1900; Randolph 1901] and those which float are not necessarily those of live-born infants . . . . It is therefore pointless to apply the hydrostatic test, which will impair the material for other and more important investigations.

Before dismissing the test so absolutely, there are some points both for and against its use as part of the feeble armamentarium that the pathologist has in determining this vital issue. First, the slightest degree of post-mortem decomposition immediately negates any interpretation of the flotation test. As so many potential infanticides are found hidden, buried or submerged, decomposition is more the rule rather than the exception, so the test cannot even be considered in a large proportion of cases. Even when obvious putrefaction with gas formation is not apparent, slight decomposition will still begin to produce gas on a microscopic scale from the second unrefrigerated day in temperate surroundings – and much earlier in hot climates. In recent years the advent of resuscitation attempts make the evaluation of breathing even more difficult or even impossible. The use of mouth-to-mouth revival, external cardiac massage and the administration of oxygen, completely negate any of the already fragile tests for respiration in a newborn infant.

On the positive side – assuming fresh tissue – the floating of the whole pluck of thoracic organs in water makes one tend to accept that respiration had taken place. This means that the attached heart, mediastinum and neck structures are also buoyed up by the lung – but in this case, it is almost invariable that the lungs are sufficiently aerated for this to be obvious on naked-eye inspection.

This is about the total extent of the usefulness of the hydrostatic test. The complicated instructions offered in many textbooks concerning cutting the lung into lobes and then into pieces, squeezing them with knife blades and even pressing them underfoot on the mortuary floor before floating them, all smatter of black magic and are a complete waste of time. Worse, they can simulate a false sense of scientific validity and even to an eventual miscarriage of justice.

The best way of seeking proof of respiration is to look at, to feel and to listen to the lungs. The lungs of a stillbirth are dark, small, heavy and liver-like, even though they may still float. They lie contracted against the mediastinum, though when the body is intact they must still fill the thoracic cavity, as the pliable rib cage pulls inwards and the diaphragm rises to eliminate the free space in the chest. Though this fact is seldom considered, it must be obvious that there can be no free space in the pleural cavities until the thorax is opened at autopsy. As soon as this is done, air enters the pleural sacs and the lungs then appear much smaller than the expanded chest cavity than is seen when the sternal plate is removed.

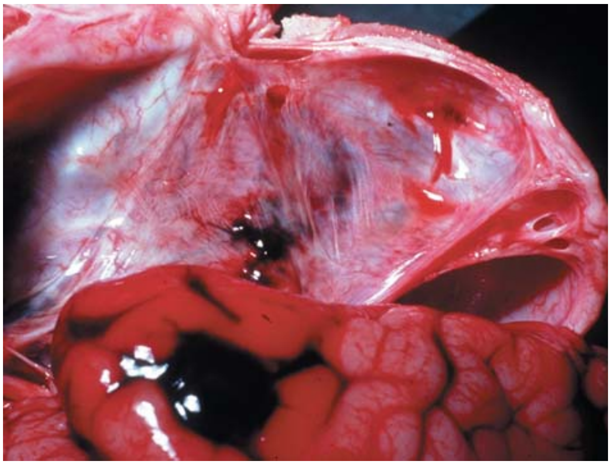

FIGURE 20.10 *A positive finding in a baby whose body was abandoned after death. Respiration had occurred but, on examination of the cranium, a tear in the tentorium was found with an adjacent meningeal haemorrhage, presumably caused by excessive moulding of the head during precipitate delivery.*

Unrespired lungs appear smaller than those that have breathed for a significant time, though there is a continuous range of expansion. According to Polson *et al.* (1985), unrespired lungs weigh something of the order of one-seventieth of the body weight; when aeration begins, the lung weight increases because of the increased vascular volume and become about one-thirty-fifth of the body weight. This seems at variance with actual measurements at autopsy as the combined weight of both lungs in an infant of average birth weight is about 40 g and, although there is a fairly wide variation between individuals, no combined lung weights approach 85 g, which would be one-thirty-fifth the weight of a 3000 g baby.

The texture of an unrespired lung is rubbery and uniform, with no paler, crepitant areas at the margins, which tend to be sharply angled. On slicing, the interior is uniform in colour and texture, being moist and resembling stiff strawberry jelly. On rubbing a small piece between the fingers close to one's ear, no crepitation is heard.

The lung that has respired will change quantitatively from the above appearances depending on the length of time and the depth to which breaths have been taken. Feeble or brief respiration will affect only the margins, whilst more robust or prolonged breathing may lead to total expansion. Even a single good breath can unfold some of the fetal pulmonary tissues – and of course, resuscitation, whether by mouth-to-mouth, chest compression or administration of oxygen, will actively inflate the lungs, dead or alive. Even ordinary handling of the dead baby may aspirate air into the lungs. Fully respired lungs will be obvious on inspection. The organs are pink or mottled, the medial edges overlapping the mediastinum and part of the pericardium, though not as fully as in the older neonate. The weight is greater, but not as much as recorded by Polson and colleagues, as described above. The edges of the lung, including the interlobar fissures, are more rounded by the expansion, compared with the angular margins of the unrespired organ.

The major change is in the colour and texture. It may be pink all over the surface or there may be patches of darker atelectasis under the pleura. The lung is spongy and resembles the familiar adult tissue. On cutting, the aerated texture is apparent, and on holding a piece to the ear, typical crepitance is heard on rubbing gently between the fingers.

The problem arises with the minimally respired lung, where the changes are intermediate between the two extreme states described above. The anterior margin best shows partial expansion, the areas such as the lingula, anterior diaphragmatic margin and the medial edge of the lower lobe sometimes being pinker and more expanded than the posterior parts. The most sensitive criterion is the ear crepitance test, but it is all a matter of degree as to whether the pathologist is impressed enough – having excluded any possibility of decomposition or resuscitation artefact – to consider that significant respiration has occurred.

Any doubts must be resolved in the direction of no breathing and, even in doubtful instances when the pathologist decides – on balance – that respiration has occurred, he should convey his uncertainty in the body of his report. As to the corroborative role of the flotation test, if performed, when the whole chest pluck floats, the descriptions of the lungs given above will usually indicate clearly that the lungs had breathed. A sinking lung will never correspond with a pink crepitant appearance, but what appears to be a dark, fetal lung may float. The test unfortunately tends to mislead more in one direction than the other, in that almost all lungs float irrespective of their appearance. To repeat the advice of Lester Adelson, whose experience and clarity of expression makes his voice one of the most respected in forensic pathology:

> Unless the pathologist has incontrovertible criteria of post-natal survival, e.g. well-expanded lungs, food in the stomach or vital reaction in the stump of the umbilical cord, he is legally bound not to diagnose live birth. Convictions for infanticide have been set aside where there was any doubt whatsoever that the child was born alive. Many courts have pushed this proposition to the extent that the state has been given the burden of proving that the baby was born alive beyond any possible doubt rather than beyond any reasonable doubt, the latter being the general level of proof required in a criminal prosecution.

Histological appearances of newborn lungs

Histology is not as helpful as might be expected in the problem of demonstrating the onset of breathing. It was formerly thought that the shape of the alveoli and the height of the lining epithelium would indicate whether they had been expanded by the passage of air, but this has proved to be a disappointing exercise. According to Shapiro (1977), the height of the epithelium and the shape of the alveoli are more a measure of fetal maturity than respiration at birth. The lung parenchyma develops in the fetus as a branching ramification of the trachea and at an intermediate stage – around 800 g body weight or 4 months' gestation – appear as gland-like structures with a cuboidal or columnar cell lining. The evolution to the thin-walled adult-type alveolus is complete before full term, or when the fetal weight is about 2500 g. The alveolar spaces may be found substantially distended even in fetuses who could not have breathed air, such as the one described by Shapiro that was found free in the maternal abdomen after rupture of the uterus. No intra-alveolar epithelium was visible and the lung could not be distinguished histologically from that of a child that had breathed.

Other authors have described the maturation of the fetal lung (Kuroda *et al.* 1965; Parmentier 1962; Ham and Baldwin 1941) and it is undisputed that fully expanded alveoli can be present before full term, even in infants who have not inspired air. After the fifth month, the glandular nature of the air sacs of the lumen are filled with amniotic fluid. In fact, there is evidence that amniotic fluid may be produced by the nascent pulmonary lining, because it is present even in cases of congenital bronchotracheal atresia. When breathing starts, the alveoli further enlarge, and the fluid is partly expelled through the bronchi and partly absorbed back into the pulmonary circulation. Ham (1950) and Shapiro (1947) point out that the finding of open alveoli with no cuboidal or columnar epithelium means that the fetus has passed the two-thirds stage of its maturation process, not that it has respired air.

Further proof is found in the histological appearance of sequestrated parts of lung tissue in infants, which were not connected to the bronchial system but which still showed alveolar ducts and alveoli (Potter and Buhlender 1941; Potter 1952). Fetal lung transplanted into the anterior chamber of an animal eye also developed alveoli with terminal bronchioles (Waddell 1949). Post-mortem handling has also been incriminated for the entry of air into fetal lungs. Apparently respired alveoli have been found in lung sections from a dead infant taken from the uterus of a dead mother.

Janssen (1977) agrees that resuscitation will lead to post-mortem aeration of the lungs, but claims that such artificial respiration cannot cause the alveoli of stillborn neonates to be perfectly and uniformly aerated. Against this view, however, is the undoubted patchy appearance of the lung expansion of many infants who have undoubtedly breathed for some time, so it is difficult to see how the two situations could ever be differentiated. Janssen's statement about the histological appearances is reasonable:

> According to the present level of knowledge and possibilities for examination, ventilation of the lungs alone cannot be taken as a certain indication of a live birth. Under various circumstances, lungs originally aerated can become devoid of air: conversely, the lungs of stillborn neonates can appear aerated. It is not possible to be certain in all cases.

In general, an infant who has breathed strongly for some time – probably hours in most cases – will have lungs that histologically resemble the older child with full, uniform, expansion of alveoli. In many infants who undoubtedly have lived and breathed for an appreciable time, however, there may be only patchy expansion, which corresponds to the uneven aeration seen macroscopically under the pleural surface. The difficulty arises when paradoxical appearances are seen, examples of which are illustrated in many textbooks. Thus an undoubted stillbirth may reveal quite extensive alveolar expansion, whilst a baby that unequivocally lived for some time may show totally collapsed air sacs.

In medico-legal work, though it is quite in order to refer to the usual state of affairs, when a criminal trial is in progress, the standard of proof is 'beyond reasonable doubt' and the well-documented exceptions just mentioned (as with the flotation test) make it unsafe to be dogmatic over a histological opinion.

ESTIMATION OF MATURITY

This is a necessary part of the autopsy, both as matter of record of the size and presumed age of the infant, but also as a legal requisite in those jurisdictions that require proof of viability before either stillbirth can be recorded or a charge of infanticide brought.

First, it must be accepted that – as at any time of life – morphological measurements are by no means infallible indicators of chronological age. Personal variations are compounded by sex, race and nutritional factors to prevent accuracy. The time of appearance of ossification centres is no longer as uniform as once thought. As with so many biological conditions, there is the typical 'bell-shaped curve' of probability, so that though most subjects will lie in the large central zone, there will be a progressively diminishing number at each extreme of the graph. It should be remembered that a female infant is usually at least 100 g less in weight at full term than the male, though individual differences in either sex can be far larger than this statistical variation. Twins, even those going to full term, are each lighter than a singleton child. The full-term singleton fetus, at **40 weeks' gestation**, will have the following vital specifications, subject to the above cautions:

- a weight of between 2550 and 3360 g
- crown–heel length of 48–52 cm
- crown–rump length of 28–32 cm
- head circumference of 33–38 cm
- the ossification centre in the lower end of the femur will almost always be present, being about 6 mm in diameter
- lanugo is absent or present only over the shoulders; head hair is about 2–3 cm long
- the testes are palpable in the scrotum; the vulval labia closes the vaginal opening
- the umbilicus is midway between xiphisternum and pubis
- there is dark meconium in the large intestine
- an ossification centre in the upper end of the tibia will be present in 80 per cent of full-term infants. The length of finger and toenails is an unreliable guide. More detailed information may be obtained from obstetric and paediatric texts.

At 36 weeks, the crown–heel length is about 45 cm and the weight around 2200 g. There will probably be ossification centres in the cuboid and capitate bones, often in the lower end of the femur and perhaps also in the upper end of the tibia.

At 28 weeks, the fetus is likely to be about 900–1100 g in weight, have a crown–heel length of 35 cm, a crown–rump length of 23 cm, and a foot length of 8 cm.

The old Haase rule-of-thumb for fetal age and size was that, up to the twentieth week, the length in centimetres was the square of the gestation period in (lunar) months. Beyond the twentieth week, the length in centimetres divided by five represents the age in months.

Examination of the ossification centres may be performed radiographically, which requires the advice of a radiologist, as the time of appearance may not be synchronous with visual identification of the centres. At autopsy it is usual to seek them directly, as follows: at the knee, the leg is flexed and a lengthwise cut made over the patella. The bones are pushed forwards through the cut, and transverse

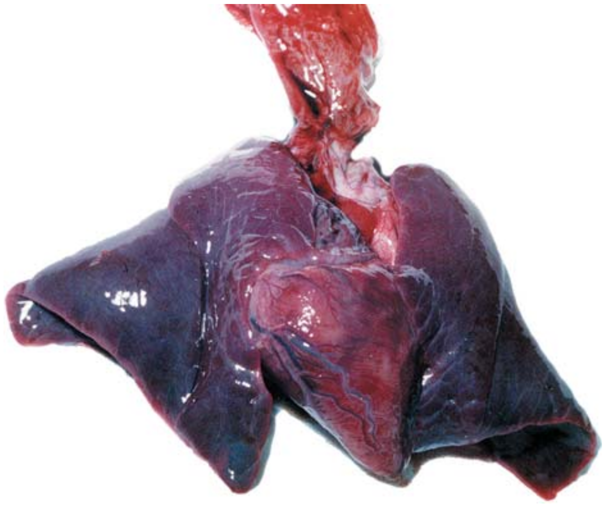

FIGURE 20.11 *The pluck of thoracic organs from a known stillbirth in hospital. The lungs are firm and heavy with no crepitation when squeezed. Margin portions of lung floated in water, however, demonstrating the fallacy of the 'flotation test'.*

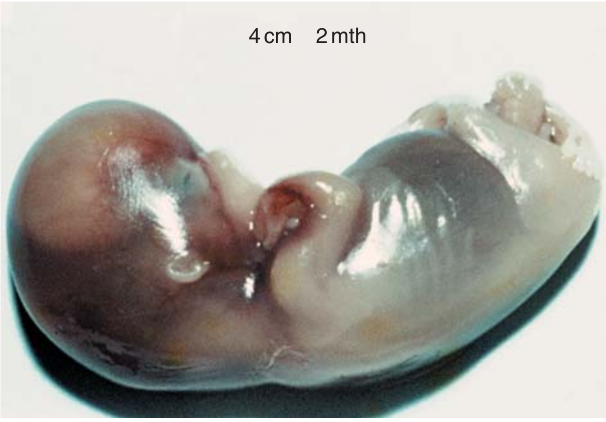

FIGURE 20.12 *Fetal maturity: up to the twentieth week the length of the fetus in centimetres is approximately the square of the age in months (Haase's rule).*

slices made with a knife through the cartilage of the lower end of the femur. If the ossification centre is revealed, the cuts should continue through plain cartilage above it until the diaphyseal centre is reached. This is to avoid mistaking the lower margin of the diaphyseal bone for the separate epiphyseal centre. The upper end of the tibia is then dealt with in a similar way.

In the foot, a cut is made in the axis of the limb up between the third and fourth toes into the ankle joint, to seek the cuboid bone. The talus and calcaneum can be entered by cuts across the upper dorsum of the foot when the latter is plantar-flexed. The centre in the calcaneum appears at about 7 months' gestation, that in the talus at about 7 months', and the cuboid at full term.

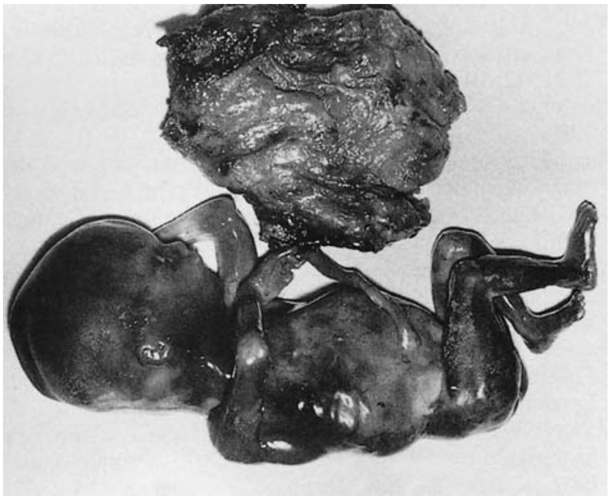

FIGURE 20.13 *Fetal maturity: by Haase's rule this 18 cm crown–heel fetus is about 4.5 months old.*

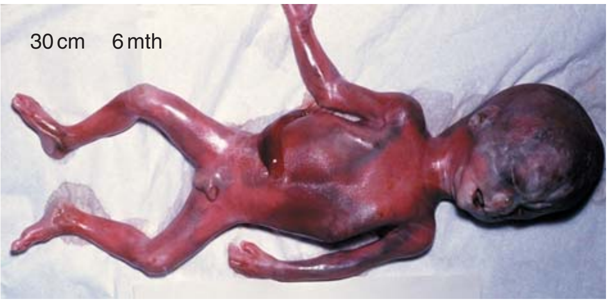

FIGURE 20.14 *Fetal maturity: this infant is 30 cm crown–heel length, which is approximately 6 months' gestation, as the length in centimetres beyond the fifth month is about five times the age in months.*

In assessing skeletal age in the fetus, the reference book of Fazekas and Kosa (1978) is invaluable. Other non-osseous methods of estimating maturity exist, such as the progressive development of surfactant-producing alveolar type II cells in fetal lungs (Betz *et al.* 1992).

REFERENCES AND FURTHER READING

Adelson L. 1974. *Pathology of homicide.* Thomas, Springfield.

American Academy of Pediatrics. 2001. Distinguishing sudden infant death syndrome from child abuse fatalities. *Pediatrics* **107**:437–41.

Betz P, Nerlich A, Wilske J, *et al.* 1992. Determination of fetal age by immunohistochemical estimation of surfactant-producing alveolar type II cells. *Forensic Sci Int* **53**:193–202.

Buris L, Torocsik I, Poczkodi S. 1979. Is the aeration of the lungs a reliable sign of live birth? *Z Rechtsmed* **83**:303–12.

Byard RW, James RA, Zuccollo J. 2001. Potential confusion arising from materials presenting as possible human remains. *Am J Forensic Med Pathol* **22**:391–4.

Castellana C, Kosa F. 1999. Morphology of the cervical vertebrae in the fetal-neonatal human skeleton. *J Anat* **194**:147–52.

Dilworth T. 1900. The flotation test. *Br Med J* **2**:1567.

Driever F, Dettmeyer R, Madea B. 2001. [Forensically relevant unexpected delivery after unrecognized or denied pregnancy.] *Arch Kriminol* **208**:174–81.

Fazekas G, Kosa F. 1978. *Forensic fetal osteology.* Akademiai Kiado, Budapest.

Ferreira GF, Rega RM, Mandarim-De-Lacerda CA. 1990. Allometry of hepatic weight growth in human staged fetuses. *Arch Ital Anat Embriol* **95**:223–8.

Funayama M, Ikeda T, Tabata N, *et al.* 1994. Case report: repeated neonaticides in Hokkaido. *Forensic Sci Int* **64**:147–50.

Gordon I, Shapiro H, Berson S (eds). 1988. *Forensic medicine: a guide to principles*, 3rd edn. Churchill Livingstone, London.

Ham A. 1950. *Histology*. Lippincott, Philadelphia, pp. 448–96.

Ham A, Baldwin K. 1941. A histological study of the development of the lung with particular reference to the development of the alveoli. *Anat Rec* **81**:369–79.

Janssen W. 1977. *Forensic histopathology.* Springer, Heidelberg.

Kellett RJ. 1992. Infanticide and child destruction – the historical, legal and pathological aspects. *Forensic Sci Int* **53**:1–28.

Kuroda S, Nagamori H, Ebe M, *et al.* 1965. Medicolegal studies on the fetus and infant with special reference to the histological characteristics of the lungs of liveborn and stillborn infants. *Tohoku J Exp Med* **85**:40–54.

Kuroki Y, Dempo K, Akino T. 1986. Immunohistochemical study of human pulmonary surfactant apoproteins with monoclonal antibodies. Pathologic application for hyaline membrane disease. *Am J Pathol* **124**:25–33.

Morris JF, Hunt AC. 1966. Breaking strength of the umbilical cord. *J Forensic Sci* **11**:43–9.

Oehmichen M, Gerling I, Meissner C. 2000. Petechiae of the baby's skin as differentiation symptom of infanticide versus sids. *J Forensic Sci* **45**:602–7.

Parmentier R. 1962. L'aeration neonatale due poumon. *Rev Belg Path Med Exp* **29**:121–244.

Perdekamp MG, Bohnert M, Ropohl D. 2000. [Discovery of externally unremarkable case of homicide of an infant.] *Arch Kriminol* **206**:160–7.

Polson C, Gee D, Knight B. 1985. *Essentials of forensic medicine*, 3rd edn. Pergamon Press, London.

Potter E. 1952. *Pathology of the fetus and newborn.* Year Book Publishers, Chicago.

Potter E, Bohlender G. 1941. Intra-uterine respiration in relation to the development of the fetal lung. *Am J Obstet Gynecol* **42**:14–22.

Randolph C. 1901. The flotation test for live birth. *Br Med J* **1**:146.

Shapiro H. 1947. The limited value of microscopy of lung tissue in the diagnosis of live and still birth. *Clin Proc* **6**:149–58.

Shapiro HA. 1977. Microscopy of human fetal lung and the diagnosis of postnatal respiration. In: Wecht C (ed.), *Legal medicine annual 1976*. Appleton Century Crofts, New York.

Tabata N, Morita M, Azumi J. 2000. A frozen newborn infant: froth in the air-passage after thawing. *Forensic Sci Int* **108**:67–74.

Waddell W. 1949. Organoid differentiation of the fetal lung: a histologic study of the differentiation of mammalian fetal lung *in utero* and in transplants. *Arch Pathol* **47**:227–47.

Yamauchi M, Usami S, Ikeda R, *et al.* 2000. Medico-legal studies on infanticide: statistics and a case of repeated neonaticide. *Forensic Sci Int* **113**:205–8.

Zhu BL, Maeda H, Fukita K, *et al.* 1996. Immunohistochemical investigation of pulmonary surfactant in perinatal fatalities. *Forensic Sci Int* **83**:219–27.

Zhu BL, Ishida K, Fujita MQ, *et al.* 2000. Immunohistochemical investigation of a pulmonary surfactant in fatal mechanical asphyxia. *Int J Legal Med* **113**:268–71.

Zhu BL, Ishida K, Quan L, *et al.* 2000. Immunohistochemistry of pulmonary surfactant apoprotein A in forensic autopsy: reassessment in relation to the causes of death. *Forensic Sci Int* **113**:193–7.

CHAPTER 21

Sudden death in infancy

- Definition of the sudden infant death syndrome (SIDS) 451
- Incidence of sudden infant death 451
- Factors influencing the risk of SIDS 452
- The case history in sudden infant death 453
- The scene of death and external findings 453
- Autopsy appearances 454
- Histological findings in SIDS 455
- Theories of causation 455
- The SIDS autopsy 456
- Medico-legal problems in cot deaths 456
- Addresses of counselling organizations 457
- References and further reading 457

This distressing condition has a variety of names and abbreviations. The one now most generally accepted is the 'sudden death in infancy syndrome', often shortened to (and pronounced as) 'SIDS'. An older title was 'SUD', standing for 'sudden unexpected death', but this in itself did not denote the infant connotations. Colloquially, SIDS is known as 'cot death' in Britain and 'crib death' in North America.

DEFINITION OF THE SUDDEN INFANT DEATH SYNDROME (SIDS)

Following one of the international conferences on the condition (held in Seattle in 1969) the following useful definition of the syndrome was proposed by Beckwith and is now generally accepted: 'the sudden death of any infant or young child which is unexpected by history and in whom a thorough necropsy fails to demonstrate an adequate cause of death'. Other definitions have been proposed later, e.g. the definition by the National Institute of Child Health and Human Development (NICHD) in 1989 emphasizing the necessity of the scene of death examination, and a three-tiered definition by Beckwith 1992 in Sydney, during the Second SIDS International Conference, dividing the SIDS into three different categories according to the age, the occurrence of similar deaths in siblings or other close relatives, and the presence of petechial haemorrhages and inflammatory lesions or other abnormalities found during the autopsy. The 'Stavanger definition', proposed during the Third SIDS International Conference in Stavanger, Norway 1994, lays stress on the importance of the investigation of the circumstances of death: sudden death in infancy unexplained after review of the clinical history, examination of the circumstances of death and post-mortem examination. However, when put to the vote during a subsequent SIDS Global Strategy Meeting in Stavanger, the original definition from 1969 received the most votes.

INCIDENCE OF SUDDEN INFANT DEATH

Until recent years, national death statistics were totally unreliable in providing data on the incidence of SIDS, as the nomenclature was confused, and the certifying habits of different pathologists and clinicians were extremely variable. According to a recent review by Rognum (1999), the percentage of sudden deaths in infancy that are diagnosed by various authors as pure SIDS varied between 2.5% and 70%. This situation still applies in many countries, especially those where high infant mortality from many causes overshadows sudden infant deaths. In advanced countries, the progressive and often dramatic reduction in total infant mortality since the beginning of the century has eliminated many infective, nutritional and congenital diseases to the point where the 'law of diminishing returns' has flattened out the mortality curve to a low level. Until 1990, there was little evidence that the SIDS rate had changed significantly for centuries or even millennia, so the pre-1990 straight-line incidence of SIDS became a far greater proportion of the total infant deaths compared with the early years of this century. In many advanced countries, SIDS is still the most common single cause of infant mortality after the perinatal period.

Until about 1990, Britain had a SIDS rate of about 2 per 1000 live births, but the risk varied considerably in various sections of the population. Golding *et al.* (1985) give a comprehensive table in their book of known variations in worldwide incidence before the recent dramatic decline. For example, in the USA there was an overall rate of 2.3/1000, though again this varies considerably from place to place, and among different ethnic and social groups. In Chicago the rate in 1969 was 3.4 compared with 1.4 in upstate New York in 1974. There is far more variation amongst ethnic groups in the USA, the rate being much higher among non-White infants. In Nebraska, for instance, the rate was 1.9/1000 Whites and 5.9/1000 in non-Whites. There is other evidence, however, that social differences are the predominant factor rather than any intrinsically ethnic cause.

As noted above, in Britain and several other European countries the SIDS rate has dropped remarkably in the past few years. In France there was a stagnation for post-neonatal mortality between 1979 and 1993 followed by a sharp decrease to 2/100 in 1995. In England and Wales there was a 69% fall in the sudden infant death rate between 1988 and 1992 from 2.01 to 0.63 and further down to 0.3 in 2000 (see Table 21.1).

The cause for this is obscure. It almost coincides with a vigorous publicity campaign by the Department of Health and the Foundation for the Study of Infant Deaths, which advocated sleeping babies on their back, not on the face and in avoiding overheating and smoking near the baby, as well as exhortations to seek medical advice whenever the infant was unwell.

However, the fall slightly antedated this campaign, so that the contribution of sleeping position, etc., to the decline is not fully understood, though there is other evidence from The Netherlands that recommendations about sleeping supine favourably affects the death rate. However, these same years have shown a series of very mild winters, which may also be a factor.

TABLE 21.1 *Infant mortality rates (per 1000 live births) in England and Wales in 1996–2000*

Infant mortality	1996	1997	1998	1999	2000	No.
Stillbirth	5.4	5.3	5.3	5.3	5.3	3203
Infant (under 1 year)	6.1	5.9	5.7	5.8	5.6	3399
Early neonatal (under 1 week)	8.6	8.3	8.2	8.2	8.2	1760
Neonatal (under 4 weeks)	4.1	3.9	3.8	3.9	3.9	2335
Post-neonatal (4 weeks–1 year)	2.0	3.0	1.9	1.9	1.8	1064
SIDS underlying cause (4 weeks–1 year)	0.6	0.5	0.4	0.4	0.3	203

Source: Office for National Statistics, Health Statistics Quarterly 05, Spring 2000.

FACTORS INFLUENCING THE RISK OF SIDS

Age

The age range in which sudden infant death occurs follows the usual 'bell-shaped curve' of most biological phenomena. As the Seattle definition stated, the ends of the curve lie at 2 weeks and 2 years, but these boundaries are far too wide for the great majority of victims. The peak incidence occurs at around 3 months and the time bracket of 2–7 months will encompass most cases. It is rare after 9 months, but deaths can occur well into the second year which, by all the criteria, can only be called 'cot deaths'.

Similarly, a few small infants die unexpectedly and inexplicably before the first month, but it must be accepted that true SIDS is not a condition of the neonatal period. There is almost always a gap between the deaths that must be attributed as being a sequel to birth, and those that suddenly occur subsequently in babies who have apparently weathered that dangerous period successfully. It is obvious, however, that such an absolute distinction cannot be made, as premature and underweight infants have a higher risk of SIDS, and some perinatal deaths are indistinguishable from SIDS in that autopsy may reveal the same negative findings.

Sex

There is a slight but definite sex bias in the syndrome, showing the usual slight weighting against the male infant, which is a common feature of many conditions and at all ages. Most surveys have shown a ratio of 1:1.3 or thereabouts, with an excess of male infants. Some other investigations, however, have indicated equal numbers and one survey actually showed a slight excess of girls.

Twinning

There is a marked excess of deaths in a member of a twin pair as opposed to singletons. Figures have varied in different surveys, but the risk is at least twofold. In the author's (BK) own hospital district (South Glamorgan), where a total birth survey is maintained, some years ago the increased risk was five times that for single babies. The reason for the excess among twins lies partly in the fact that physically disadvantaged infants, such as premature births and those of low birth weight are at greater risk; these two factors are often present in twins. There is no difference in incidence between identical or binovular twins. There have been at least ten reports of both twins having died of SIDS on the same day, and obviously reported cases are but a small fraction of those occurring. Some pathologists, almost exclusively in the USA, claim that simultaneous death in twins is indicative of homicide, but this view is not generally accepted.

Seasonal incidence

There is a marked seasonal variation in SIDS, the excess occurring in the colder, wetter months in temperate zones. In Europe and North America, most SIDS die between October and April, whilst in Australasia the monthly incidence is reversed. In tropical and subtropical zones, no clear pattern emerges, partly because many of the countries in these latitudes are relatively underdeveloped so have both high infant mortality, and often death certification and statistical records of lesser accuracy. SIDS undoubtedly occurs in the tropics, however, but the incidence is often swamped by that of more obvious disease. Returning to temperate zones, many efforts have been made to relate SIDS to changes of temperature and the incidence of respiratory infections in the community. The results are conflicting, but there seems little doubt that respiratory infections are a precipitating trigger in SIDS.

Social class and housing

This is not the place to expand upon the social epidemiology of cot death, but there is ample evidence that lower levels of social disadvantage, as measured by occupation and housing quality, are strongly related to the incidence of SIDS. It is difficult to reconcile these epidemiological factors with the numerous pathological theories and the few facts concerning the aetiology of sudden infant death.

THE CASE HISTORY IN SUDDEN INFANT DEATH

The history is usually brief and remarkably similar from case to case. The infant is either quite well on the preceding day or (in about half the victims) has trivial symptoms, usually of upper respiratory infection or a bowel upset. Even these symptoms may be overstated, as the history is of necessity taken retrospectively from distraught parents trying to rationalize the tragedy.

The story usually records that the infant was put to sleep in the evening and was found dead in the morning when a parent first visited it. Alternatively, the baby was given an early morning feed when it appeared well, but was later found dead in its sleeping place. Most infants are found dead in the first half of the day, before mid-morning. A minority die (or are found dead) in the afternoon and evening. Death can be rapid and silent. Though most victims die unobserved, the author's (BK) experience can confirm that a number of cases indicate that an infant observed to be quite well at a given moment can be found dead within 5 minutes, and cases are on record of SIDS occurring in hospital and even in a physician's arms.

THE SCENE OF DEATH AND EXTERNAL FINDINGS

The scene of the death is rarely available for examination in its original state. When found, the child is naturally either rushed to a doctor or a hospital, or an ambulance team arrives and attempts resuscitation, again usually hurrying to a hospital casualty department.

When found, a few children have slight oedema fluid exuding from the mouth. This may sometimes be blood-tinged, and frank bleeding – apparently from a congested mouth or pharyngeal vessel – is occasionally seen.

Stomach contents may be present in the mouth, nostrils or on the face, but this must not be assumed to be a cause of death (from aspiration), as is discussed below. The hands of the child are often found to be clenched around fibres from the bedclothes, as if there were spasmodic agonal grasping movements.

Some children are found under the bedclothes, even huddled upside down at the foot of the cot. Again, this has no significance, as thousands of healthy infants habitually sleep in this way without suffering SIDS.

A number of infants have been found moist with sweat and even with a raised body temperature. The significance of this is not known, but part of the recent preventative advice given in Britain, is to avoid overwrapping and overheating.

There are no external findings of significance. The face may be pale or slightly cyanosed or congested: there are no petechial haemorrhages in the face or eyes. The posture at death is not necessarily relevant, even if the initial sleeping position is important in the aetiology – some are found face down and may have a pallid area around the mouth and nose where pressure has prevented post-mortem hypostasis from settling. This develops after death and on no account must be misinterpreted as a pressure mark indicating suffocation.

AUTOPSY APPEARANCES

These have given rise to much controversy over the years. Essentially the gross findings are nil, almost by definition. If any significant pathological lesion is found on naked-eye examination, the death ceases to be a true SIDS, and is categorized as whatever lesion was discovered. The proportion of 'explained' cot deaths varies greatly with the pathologist, though this variation is more in relation to microscopic than gross findings. Of all babies found dead in their sleeping place, that is, 'cot deaths', most will be true SIDS, but a few will have a significant pathological condition. The relative proportions of these two subsets varies from pathologist to pathologist. Many would accept that about 15 per cent reveal some pathological lesions at autopsy, though this is not to say that those lesions – sometimes slight – were a factor in the death. For example, some mild congenital heart lesion or the presence of Down's syndrome is occasionally revealed, but this is not to say that they were causative. Thus almost all SIDS are cot deaths, but not all cot deaths are SIDS.

The external appearances have already been discussed and are essentially negative, apart from inconstant froth at the lips and nostrils, and fibres sometimes clenched in the fists. The length, weight and other physical measurements are of no direct relevance, though naturally they must always be carefully recorded as in any other infant autopsy, for it may transpire that the death was not a typical SIDS.

Internally, the findings are again non-specific in a typical SIDS. A large number of reports now exist on the microscopy, microbiology and biochemistry of the syndrome, but no diagnostic criteria have yet been confirmed. The following naked-eye features may be present at autopsy:

- Petechial haemorrhages on the visceral pleura.
- Petechial haemorrhages in the thymus gland.
- Petechiae or larger ecchymoses on the epicardial surface of the heart, especially posteriorly. These petechiae have been the subject of intense controversy since they were first described in infant deaths by Tardieu in 1855, but they are almost certainly agonal in origin. One explanation suggests that they are the result of forced inspiratory efforts against a closed airway, the latter being either from laryngeal spasm or from a collapsed pharynx in hypotonic infants, but neither of these hypotheses has been substantiated. Petechiae in the thorax are found in about 70 per cent of true SIDS infants. Thymic haemorrhages are often prominent, the organ being studded with both petechiae and larger ecchymoses: it has been claimed recently that in SIDS as opposed to mechanical suffocation, the haemorrhages are in the thymic cortex, rather than the medulla, but few pathologists accept this.

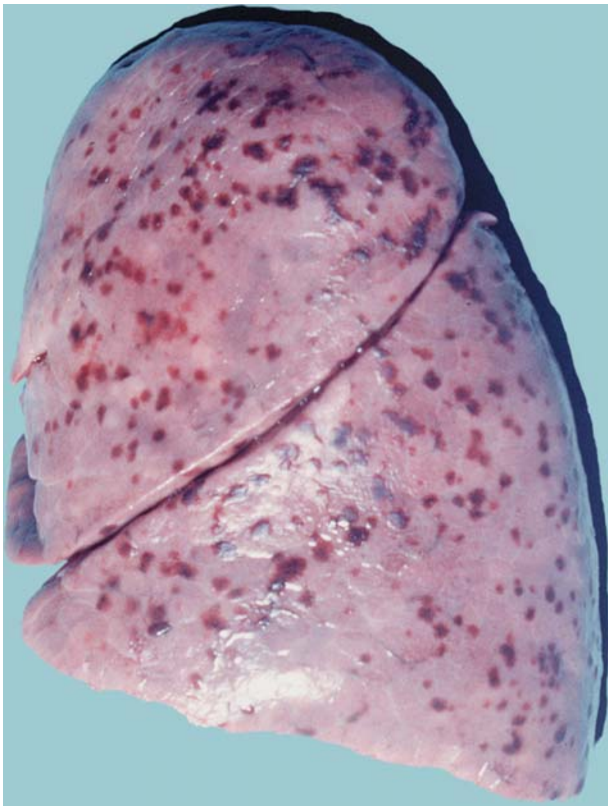

FIGURE 21.1 *Subpleural haemorrhages – the original 'Tardieu spots'. These are far more pronounced than those seen in adult lungs, being taken from a case of SIDS, in which 70 per cent of cases exhibit these lesions.*

- Gastric contents, usually milk-curds, in the air passages. All too often this has been used by some pathologists as the definite cause of death, but there is no evidence that it is the reason for the death. It is either an agonal regurgitation or even a post-mortem phenomenon. In a series of 100 consecutive autopsies, including adults and children, the author (BK) found gastric contents in greater or lesser amount in no less than 25 per cent of cadavers, almost all of which had quite satisfactory pathological lesions to account for death. This aspect is discussed in more detail in Chapter 14.
- Respiratory infections in the form of inflamed laryngeal or tracheobronchial mucosa are not common, but are macroscopically evident in some SIDS victims. If the infection is severe enough to produce pus and especially if there are obvious inflammatory changes in the lung parenchyma, then the death is removed from the SIDS category into a fatal chest infection group. The finding of minimal 'inflammatory' cells on microscopy must be viewed with caution. Some pathologists undoubtedly lay too much emphasis on sparse or small collections of leucocytes in the peribronchial zones. In control infants, small foci of lymphocytes are common and can

be disregarded. It is usually the lung histology that causes the variations in the proportions of 'cot deaths' that are classified as true 'SIDS'.

- Pulmonary oedema is common, though usually moderate in intensity. The surface of the lungs often shows patchy sublobular partial collapse, with areas of blue lung alternating with better-aerated pink zones.

In some infants dying unexpectedly, abnormalities such as Down's syndrome, congenital heart disease or other chronic systemic disease are found. The problem then is to decide whether the death should be ascribed to 'SIDS' or to the obvious disease, even though there is no evidence that such a disease should have suddenly caused death. There is a natural temptation to use an established lesion to explain the death, especially as both doctors and parents tend to be less unhappy with a specific disease process than with the rather nebulous diagnosis of 'SIDS'.

In most cases, however, there is no real reason why an infant with Down's syndrome, perfectly well one night, should be found dead the following morning – and many relatively mild (and often previously undiagnosed) congenital defects similarly are unlikely to cause death in such an abrupt manner. It is probably more logical to assume that SIDS has occurred in a child with Down's or a septal defect, or whatever. It may be accepted that this defect may be one factor among the many involved in SIDS, and therefore it is probably reasonable to enter 'SIDS' in Part I of the certificate and add the other lesion in Part II, as a contributory cause.

HISTOLOGICAL FINDINGS IN SIDS

Again a huge and controversial topic, there have been numerous papers on the microscopic appearances in SIDS. Much of this is pure research material and is often irreproducible by other workers.

Most interest has focused on the lung changes and especially evidence of pulmonary inflammation. The frequency with which positive findings are made depends upon the criteria threshold of the observer, as mentioned earlier. Peribronchiolar cell infiltration is the main finding and some workers claim to find an abnormal cellular response in the majority of SIDS material. Others, more used to the less equivocal appearances in adults, fail to attach significance to marginal increases in cell population in the peribronchial zone.

Many of the other findings are inconstant and claimed by a variety of researchers. For example, Richard Naeye (1973) supported his hypothesis of chronic hypoxia by finding thickened pulmonary artery walls, gliosis in the brainstem, retention of brown fat in the adrenals and abnormalities in the carotid body. Changes in the myocardium, adrenals, parathyroids and liver have all been described, but have either not survived critical assessment, appear incompatible with clinical and epidemiological data, or fail to be reproducible in subsequent investigations.

THEORIES OF CAUSATION

Since the mid-1970s more than 5000 articles have appeared concerning sudden infant death and one now needs review papers, such as those published by Valdes-Dapena, merely to keep up with mainstream research. Theories include allergy to cow's-milk proteins, house-mite allergy, deliberate suffocation, spinal haemorrhages, botulism, calcium deficiency, selenium deficiency, biotin deficiency, vitamin E deficiency, vitamin C deficiency, vitamin D deficiency, thiamine deficiency, hypoglycaemia, hypothyroidism, magnesium deficiency, carbon monoxide poisoning, carbon dioxide poisoning, overlaying, pharyngeal hypotonia, nasal obstruction, tracheobronchitis, respiratory syncytial virus, prolonged sleep apnoea, deficient or abnormal pulmonary surfactant, cardiac conduction defects, sodium overload in feeds, narrow foramen magnum, hyperthermia, hypothermia, influenza infection, metabolic enzyme defects, immunodeficiency, hypogammaglobulinaemia and many others.

In 1989 and again in 1994, press publicity in the UK drew considerable attention to claims that fungal growth on certain antiflammable chemicals used on cot mattress covers produced gases containing arsenic and antimony, and some chemical analyses showed that antimony could be demonstrated in some SIDS tissues. However, a causative link has not been proven and the theory fails to fit the epidemiology – especially as SIDS has been occurring for hundreds and probably thousands of years before mattress covers existed. Reports of cot deaths can be found in sources as ancient as the Old Testament and twelfth century Welsh literature.

In the 1970s and early 1980s, one hypothesis gained widespread favour and was beginning to gain general acceptance. This was the claim that some infants, with poor respiratory drive, suffered from exaggeration of the normal periods of apnoea that all infants have during sleep. A complex mechanism was built up in which progressive hypoxia and failure to respond adequately to hypercapnia and hypoxia drove the sleeping infant into a descending spiral of hypoxia–apnoea–hypoxia, which ended in bradycardia and cardiac arrest. Respiratory infections, sleep, and nasal obstruction contributed to an acute-on-chronic state of hypoxia, which histological findings claimed to substantiate.

This attractive theory was largely deflated, however, when prospective studies of infants with sleep apnoea and other respiratory problems were found to have no greater risk of SIDS than normal controls. The theory led to a great popularity of apnoea alarms, which parents of subsequent children following a SIDS loss have insisted on using. There is no evidence that they have any beneficial effect, other than reassurance of the parents; indeed, in some instances, the stress of waiting for the alarm to sound and the many false alarms, have led to marital problems.

In the present state of knowledge, it seems that SIDS is the final common pathway that leads to death in infants who have been compromised by a number of deleterious factors, which have to come together in one baby at one time in order to summate to a fatal outcome. The factors may be quite different in each case, but all manifest themselves in terminal acute cardiorespiratory failure.

Some of the factors are known such as sleep, which depresses the brainstem, and virus infections – especially of the respiratory tract – which provide a viraemia, reduce oxygenation and may narrow the effective airway lumen by mucus and exudate. Other factors are 'constitutional', in that many SIDS victims are physically compromised in some way, often since fetal life, being premature and of low birth weight (though most SIDS are neither of these). Other factors from the long list given above may operate in certain circumstances, such as botulism which may be relevant in some 5 per cent of SIDS in California.

It is thus futile to search for 'the cause' of cot death, as there is no single cause, but a galaxy of factors that vary from death to death, though the mediating terminal mechanism appears to be the same.

One of the most difficult aspects is the failure of pathological explanations to be compatible with epidemiological findings. It has been shown that non-specific health-care improvements reduce the incidence of SIDS, such as in the excellent ante-natal and post-natal services of countries such as Finland and Holland, which have the lowest SIDS rates – and the Sheffield project, where it was shown that the strongest factor in reducing incidence was the use the family made of the health services.

As far as the pathologist is concerned, his function is to perform a meticulous autopsy, to exclude any overt natural disease or injuries, and to offer sympathetic advice to those parents who want an explanation of the syndrome in comprehensible lay terms. There are several highly proficient counselling charities in Britain, the USA and other countries, who can provide literature and direct help to parents – and who also usually support most of the ambitious research schemes now operating.

THE SIDS AUTOPSY

The routine for an autopsy on a sudden infant death follows that for any infant and is little different from that in suspected child abuse. Indeed, exclusion of child abuse is a prime consideration as, although there is no aetiological connection between the two conditions, sometimes occult abuse – or sometimes genuine accident – may present as a 'cot death'. Some pathologists, both forensic and paediatric, insist on a whole-body radiological survey before commencing the autopsy and, where such facilities exist, this is a wise precaution. Full measurements of the infant are essential, as in all paediatric autopsies, to monitor the size-for-age of the child and to compare it with centile graphs, provided the birth weight is known.

The gross examination is the standard meticulous scrutiny of all external surfaces, orifices and organs. A full histological survey is vital, with even more blocks being taken than may be the norm for many forensic pathologists. The lungs in particular should be extensively sampled and six blocks is a bare minimum for these organs. In addition to histology, swabs should be taken from the major air passages and the lung parenchyma for microbiological culture, and heart blood may also be cultured. The middle ears should always be opened and a swab taken if infection is suspected. A separate piece of lung should be sent for virological culture. Where there has been any history of gut upset before death, a sample of ileal contents should be cultured. It has to be said that the return on such microbiological and virological studies is poor in terms of positive findings, both because of the post-mortem interval and because many of the micro-organisms found may be indistinguishable from commensals or post-mortem contaminants.

Similarly, by the very definition of SIDS, significant histological findings are absent. If a definite pathological lesion capable of causing death is discovered, the case is removed from the SIDS category and classified according to the disease process discovered. The problem arises when 'significant' or 'definite' has to be defined, as the criteria threshold of different pathologists varies considerably, especially where the diagnosis of an inflammatory condition in the lungs is concerned.

MEDICO-LEGAL PROBLEMS IN COT DEATHS

Differentiating SIDS from external suffocation

This is the oldest problem in cot death and one that is virtually insoluble on purely pathological grounds. The signs of infant suffocation are usually nil and the signs of SIDS are nil, so in a completely negative autopsy no reliable criteria exist. The presence of intrathoracic petechiae has little or no diagnostic value, in spite of one recent report that claims to distinguish SIDS from mechanical suffocation by the confinement of petechiae to the thymic cortex in SIDS. In any event, they are usually absent in known suffocation, such as those caused by a plastic bag over the head, yet some 70 per cent of SIDS have serosal petechiae.

In Britain, there are about 1000 cot deaths each year and probably up to 10 000 in the USA. The majority of these were found dead with the face uncovered, and in a posture where the mouth and nose are not obstructed. No one believes that all these children have been smothered by their parents; therefore, as petechiae are valueless for differentiation, it is not possible for an autopsy to state or even suggest that any particular child has been suffocated. This naturally presupposes the absence of facial bruising, scratching or other positive signs that would immediately remove the case from the SIDS category.

It is possible for circumstantial or other non-medical history to suggest or confirm deliberate suffocation, though this is extremely rare and not the concern of the pathologist. A caution must be offered about maternal confessions of smothering, however, as several instances are known where such a confession has proved totally false, the desperately bereaved woman in her grief and recrimination deluding herself in order to rationalize an inexplicable tragedy.

Multiple sudden infant deaths

Again, these pose a formidable problem for pathologist, paediatrician and investigating authority. Statistically, a mother will suffer a second cot death in her family every quarter of a million births, but a number of families are reported where three or even four siblings have died inexplicably. There is always the possibility of foul play in such circumstances, though some familial metabolic or other genetic disease – albeit obscure – is more likely. Even where no trace of the latter can be detected by the most exhaustive medical investigations, if the autopsies reveal no positive evidence of mechanical suffocation or other demonstrable cause of death, the pathologist cannot then substantiate any allegation of mechanical suffocation. He can – and should – express his concern to the investigating authorities but, if their investigations produce no corroborative evidence, it is pointless for him to continue to support a hypothesis that cannot be substantiated by any pathological evidence.

Counselling of parents

Many families, primarily the mother, have an intense desire to know why their baby died, a need that is often unfulfilled by any official contacts, such as with a coroner. Once any legal processes are complete, there is no reason why a pathologist should not meet and offer explanations to the parents of a child upon whom he has performed an autopsy. Often an earlier line of counselling is provided by a paediatrician, but some mothers wish to talk to the pathologist. Both Adelson (1974) and Knight (1983) have advocated that the pathologist should make himself readily available for this purpose, though some pathologists shelter behind an imaginary legal barrier to avoid what should be an ethical obligation.

ADDRESSES OF COUNSELLING ORGANIZATIONS

- The Foundation for the Study of Infant Deaths, Artillery House, 11–19 Artillery Row, London, SW1P 1RT, UK (Tel.: General enquiries: +44 020 7222 8001; 24-hour helpline: +44 020 7233 2090; Fax: +44 020 7222 8002). http://www.sids.org.uk/
- National Sudden Infant Death Syndrome Resource Center (NSRC), 2070 Chain Bridge Road, Suite 450, Vienna, VA 22182, USA (Tel.: (703) 821 8955; Fax: (703) 821 2098). sids@circlesolutions.com
- SIDS International (SIDSI). http://www.sidsinternational.minerva.com.au/

REFERENCES AND FURTHER READING

Adelson L. 1974. *The pathology of homicide: a vade mecum for pathologist, prosecutor and defense counsel.* W. B. Saunders, Philadelphia.

Althoff H. 1980. *Sudden infant death syndrome*. Gustav Fischer Verlag, Stuttgart.

American Academy of Pediatrics. 2001. Distinguishing sudden infant death syndrome from child abuse fatalities. *Pediatrics* **107**:437–41.

Arnestad M, Andersen M, Vege A, *et al.* 2001. Changes in the epidemiological pattern of sudden infant death syndrome in southeast Norway, 1984–1998: implications for future prevention and research. *Arch Dis Child* **85**:108–15.

Arnestad M, Vege A, Rognum TO. 2002. Evaluation of diagnostic tools applied in the examination of sudden unexpected deaths in infancy and early childhood. *Forensic Sci Int* **125**:262–8.

Bartholomew SE, MacArthur BA, Bain AD. 1987. Sudden infant death syndrome in south east Scotland. *Arch Dis Child* **62**:951–6.

Bass M, Kravath RE, Glass L. 1986. Death-scene investigation in sudden infant death. *N Engl J Med* **315**:100–5.

Bergman AB. 1972. Unexplained sudden infant death. *N Engl J Med* **287**:254–5.

Bergman AB, Beckwith RE, Rae CO (eds). 1969. *Sudden infant death syndrome. Proceedings of the Second International Conference, Seattle 1969.* University of Washington Press, Seattle.

Bergman AB, Ray CG, Pomeroy MA, *et al.* 1972. Studies of the sudden infant death syndrome in King County, Washington. 3. Epidemiology. *Pediatrics* **49**:860–70.

Betz P, Hausmann R, Eisenmenger W. 1998. A contribution to a possible differentiation between sids and asphyxiation. *Forensic Sci Int* **91**:147–52.

Bonser RS, Knight BH, West RR. 1978. Sudden infant death syndrome in Cardiff, association with epidemic influenza and with temperature – 1955–1974. *Int J Epidemiol* **7**:335–40.

Boulloche J, Mallet E, Basuyau JP, *et al.* 1986. The value of serum IgE assay in milk aspiration and the sudden infant death syndrome. *Acta Paediatr Scand* **75**:530–3.

Bowden KS. 1953. Sudden or unexpexted deaths in infancy. *J Forensic Med* **1**:19–27.

Byard R, Krous H. 1999. Suffocation, shaking or sudden infant death syndrome: can we tell the difference? *J Paediatr Child Health* **35**:432–3.

Byard RW, Beal SM. 2000. Gastric aspiration and sleeping position in infancy and early childhood. *J Paediatr Child Health* **36**:403–5.

Byard RW, Krous HF (eds). 2001. *Sudden infant death syndrome: problems, progress and possibilities*. Arnold, London.

Camps FE, Carpenter RJ (eds). 1972. *Sudden and unexpected deaths in infancy. Proceedings of the Cambridge Symposium.* John Wright, Bristol.

Cecchi R, Bajanowski T, Kahl B, *et al.* 1995. CMV-DNA detection in parenchymatous organs in cases of SIDS. *Int J Legal Med* **107**:291–5.

Coombs RR, McLaughlan P. 1982. The enigma of cot death: is the modified-anaphylaxis hypothesis an explanation for some cases? *Lancet* **1**:1388–9.

DeSa DJ. 1986. Isolated myocarditis as a cause of sudden death in the first year of life. *Forensic Sci Int* **30**:113–17.

Dettmeyer R, Kandolf R, Schmidt P, *et al.* 2001. Lympho-monocytic enteroviral myocarditis: traditional, immunohistological and molecular pathological methods for diagnosis in a case of suspected sudden infant death syndrome (SIDS). *Forensic Sci Int* **119**:141–4.

Durigon M, Caroff J, Derobert L. 1971. [Pulmonary histology in 123 suspicious or sudden deaths in infants under one year of age.] *Med Leg Dommage Corp* **4**:287–93.

Emery JL. 1983. The necropsy and cot death [editorial]. *Br Med J Clin Res Ed* **287**:77–8.

Emery JL. 1986. Families in which two or more cot deaths have occurred. *Lancet* **1**:313–15.

Ferris JA. 1972. The heart in sudden infant death. *J Forensic Sci Soc* **12**:591–4.

Ferris JA. 1973. Hypoxic changes in conducting tissue of the heart in sudden death in infancy syndrome. *Br Med J* **2**:23–5.

Ferris JA, Aherne WA, Locke WS, *et al.* 1973. Sudden and unexpected deaths in infants: histology and virology. *Br Med J* **2**:439–42.

Fleming PJ, Gilbert R, Azaz Y, *et al.* 1990. Interaction between bedding and sleeping position in the sudden infant death syndrome: a population based case-control study [see comments]. *Br Med J* **301**:85–9.

Forsyth KD, Weeks SC, Koh L, *et al.* 1989. Lung immunoglobulins in the sudden infant death syndrome. *Br Med J* **298**:23–6.

Freed GE, Meny R, Glomb WB, *et al.* 2002. Effect of home monitoring on a high-risk population. *J Perinatol* **22**:165–7.

Froggatt P, James TN. 1973. Sudden unexpected death in infants. Evidence on a lethal cardiac arrhythmia. *Ulster Med J* **42**:136–52.

Froggatt P, Lynas MA, Marshall TK. 1971. Sudden unexpected death in infants ('cot death'). Report of a collaborative study in Northern Ireland. *Ulster Med J* **40**:116–35.

Gardner PS. 1972. Viruses and the respiratory tract in sudden infant deaths. *J Forensic Sci Soc* **12**:587–9.

Golding J, Limerick S, McFarlane A. 1985. *Sudden infant death.* Open Books, Shepton Mallet.

Gruenwald P, Hoang NM. 1960. Evaluation of body and organ weights in perinatal pathology. I. Normal standards derived from autopsies. *Am Clin Pathol* **34**:247–53.

Guilleminault C, Heldt G, Powell N, *et al.* 1986. Small upper airway in near-miss sudden infant death syndrome infants and their families. *Lancet* **1**:402–7.

Guntheroth WG. 1973. The significance of pulmonary petechiae in crib death. *Pediatrics* **52**:601–3.

Hanzlick R. 2001. Pulmonary hemorrhage in deceased infants: baseline data for further study of infant mortality. *Am J Forensic Med Pathol* **22**:188–92.

Harrington C, Kirjavainen T, Teng A, *et al.* 2002. Altered autonomic function and reduced arousability in apparent life-threatening event infants with obstructive sleep apnea. *Am J Respir Crit Care Med* **165**:1048–54.

Hatton F, Bouvier-Colle MH, Blondel B, *et al.* 2000. [Trends in infant mortality in France: frequency and causes from 1950 to 1997.] *Arch Pediatr* **7**:489–500.

Hollander N. 1988. Beta-endorphin in the brainstem, pituitary, and spinal fluid of infants at autopsy: relation to sudden infant death syndrome. *Forensic Sci Int* **38**:67–74.

Iwadate K, Doy M, Ito Y. 2001. Screening of milk aspiration in 105 infant death cases by immunostaining with anti-human alpha-lactalbumin antibody. *Forensic Sci Int* **122**:95–100.

James TN. 1968. Sudden death in babies: new observations in the heart. *Am J Cardiol* **22**:479–506.

Kariks J. 1988. Cardiac lesions in sudden infant death syndrome. *Forensic Sci Int* **39**:211–25.

Kato I, Groswasser J, Franco P, *et al.* 2001. Developmental characteristics of apnea in infants who succumb to sudden infant death syndrome. *Am J Respir Crit Care Med* **164**:1464–9.

Kelly DH, Shannon DC. 1982. Sudden infant death syndrome and near sudden infant death syndrome: a review of the literature, 1964 to 1982. *Pediatr Clin North Am* **29**:1241–61.

Knight B. 1972. Legal and administrative problems in the 'cot death syndrome'. *J Forensic Sci Soc* **12**:581–3.

Knight B. 1983. *Sudden death in infancy: the cot death syndrome.* Faber and Faber, London.

Koch LE, Biedermann H, Saternus KS. 1998. High cervical stress and apnoea. *Forensic Sci Int* **97**:1–9.

Koehler SA, Ladham S, Shakir A, *et al.* 2001. Simultaneous sudden infant death syndrome: a proposed definition and worldwide review of cases. *Am J Forensic Med Pathol* **22**:23–32.

Krous HF, Nadeau JM, Silva PD, *et al.* 2002. Infanticide: is its incidence among postneonatal infant deaths increasing? An 18-year population-based analysis in California. *Am J Forensic Med Pathol* **23**:127–31.

Ladham S, Koehler SA, Shakir A, *et al.* 2001. Simultaneous sudden infant death syndrome: a case report. *Am J Forensic Med Pathol* **22**:33–7.

Langlois NE, Ellis PS, Little D, *et al.* 2002. Toxicologic analysis in cases of possible sudden infant death syndrome: a worthwhile exercise? *Am J Forensic Med Pathol* **23**:162–6.

Lie TJ, Rosenberg H, Erikson E. 1976. Histopathology of the conduction system in the sudden death syndrome. *Circulation* **53**:3–8.

Lundemose JB, Lundemose AG, Gregersen M, *et al.* 1990. Chlamydia and sudden infant death syndrome. A study of 166 SIDS and 30 control cases. *Int J Legal Med* **104**:3–7.

Maron BJ, Fisher RS. 1977. Sudden infant death syndrome (SIDS): cardiac pathologic observations in infants with SIDS. *Am Heart J* **93**:762–6.

Marshall TK. 1972. Epidemiology of cot deaths: the Northern Ireland study. *J Forensic Sci Soc* **12**:575–9.

Mitchell EA, Thompson JM. 2001. Parental reported apnoea, admissions to hospital and sudden infant death syndrome. *Acta Paediatr* **90**:417–22.

Naeye RL. 1973. Pulmonary arterial abnormalities in the sudden-infant-death syndrome. *N Engl J Med* **289**:1167–70.

Naeye RL. 1980. Sudden infant death. *Sci Am* **242**:56–62.

Norvenius SG. 1987. Sudden infant death syndrome in Sweden in 1973–1977 and 1979. *Acta Paediatr Scand Suppl* **333**:1–138.

Oehmichen M, Gerling I, Meissner C. 2000. Petechiae of the baby's skin as differentiation symptom of infanticide versus sids. *J Forensic Sci* **45**:602–7.

Ogbuihi S, Zink P. 1988. Pulmonary lymphatics in SIDS – a comparative morphometric study. *Forensic Sci Int* **39**:197–206.

Opdal SH, Rognum TO, Torgersen H, *et al.* 1999. Mitochondrial DNA point mutations detected in four cases of sudden infant death syndrome. *Acta Paediatr* **88**:957–60.

Oyen N, Markestad T, Skaerven R, *et al.* 1997. Combined effects of sleeping position and prenatal risk factors in sudden infant death syndrome: the Nordic Epidemiological SIDS Study. *Pediatrics* **100**:613–21.

Parish WE, Barrett AM, Coombs RR, *et al.* 1960. Hypersensitivity to milk and sudden death in infancy. *Lancet* **2**:1106–10.

Peterson DR, Beckwith JB. 1973. Magnesium deprivation in sudden unexpected infant death. *Lancet* **2**:330.

Polak JM, Wigglesworth JS. 1976. Letter: Islet-cell hyperplasia and sudden infant death. *Lancet* **2**:570–1.

Protestos CD, Carpenter RG, McWeeny PM, *et al.* 1973. Obstetric and perinatal histories of children who died unexpectedly (cot death). *Arch Dis Child* **48**:835–41.

Rasten-Almqvist P, Eksborg S, Rajs J. 2000. Heart weight in infants – a comparison between sudden infant death syndrome and other causes of death. *Acta Paediatr* **89**:1062–7.

Raven C, Maverakis NH, Eveland WC, *et al.* 1978. The sudden infant death syndrome: a possible hypersensitivity reaction determined by distribution of IgG in lungs. *J Forensic Sci* **23**:116–28.

Richards ID, McIntosh HT. 1972. Confidential inquiry into 226 consecutive infant deaths. *Arch Dis Child* **47**:697–706.

Rintahaka PJ, Hirvonen J. 1986. The epidemiology of sudden infant death syndrome in Finland in 1969–1980. *Forensic Sci Int* **30**:219–33.

Rognum TO (ed.). 1995. *Sudden infant death syndrome, new trends in the nineties*. Scandinavian University Press (Universitetsforlaget AS), Oslo.

Rognum TO. 1999. [Crib death or cot death in the nordic countries. A forensic pathologist's point of view.] *Ugeskr Laeger* **161**:6612–18.

Sadler DW. 1998. The value of a thorough protocol in the investigation of sudden infant deaths. *J Clin Pathol* **51**:689–94.

Saternus KS, Koebke J, von Tamaska L. 1986. Neck extension as a cause of SIDS. *Forensic Sci Int* **31**:167–74.

Sawaguchi T, Fujita T, Sawaguchi A, *et al.* 2000. The epidemiological study on registered cases of sudden infant death syndrome (SIDS) in Tokyo: examination of the effect of autopsy on diagnosis of SIDS and the mortality statistics in Japan. *Forensic Sci Int* **109**:65–74.

Simson LR, Jr, Brantley RE. 1977. Postural asphyxia as a cause of death in sudden infant death syndrome. *J Forensic Sci* **22**:178–87.

Sinclair-Smith C, Dinsdale F, Emery J. 1976. Evidence of duration and type of illness in children found unexpectedly dead. *Arch Dis Child* **51**:424–9.

Southall DP. 1983. Identification of infants destined to die unexpectedly during infancy: evaluation of predictive importance of prolonged apnea and disorders of cardiac rhythm or conduction. *Br Med J* **286**:1092–6.

Steinschneider A. 1972. Prolonged apnea and the sudden infant death syndrome: clinical and laboratory observations. *Pediatrics* **50**:646–54.

Steinschneider A, Weinstein SL, Diamond E. 1982. The sudden infant death syndrome and apnea/obstruction during neonatal sleep and feeding. *Pediatrics* **70**:858–63.

Sturner WQ. 1971. Some perspectives in 'cot death'. *J Forensic Med* **18**:96–107.

Sturner WQ, Dempsey JL. 1973. Sudden infant death: chemical analysis of vitreous humor. *J Forensic Sci* **18**:12–19.

Summers CG, Parker JC, Jr. 1981. The brain stem in sudden infant death syndrome. A postmortem survey. *Am J Forensic Med Pathol* **2**:121–7.

Sunderland R, Emery JL. 1981. Febrile convulsions and cot death. *Lancet* **2**:176–8.

Urquhart GE, Grist NR. 1972. Virological studies of sudden, unexplained infant deaths in Glasgow 1967–70. *J Clin Pathol* **25**:443–6.

Urquhart GE, Logan RW, Izatt MM. 1971. Sudden unexplained death in infancy and hyperimmunization. *J Clin Pathol* **24**:736–9.

Urquhart GE, Izatt MM, Logan RW. 1972. Cot death: an immune-complex disease? *Lancet* **1**:210.

Valdes-Dapena M. 1977. Sudden unexplained infant death, 1970 through 1975 an evolution in understanding. *Pathol Annu* **12**(Pt 1):117–45.

Valdes-Dapena M. 1982. The pathologist and the sudden infant death syndrome. *Am J Pathol* **106**:118–31.

Valdes-Dapena M. 1985. Are some crib deaths sudden cardiac deaths? *J Am Coll Cardiol* **5**:113B–17B.

Valdes-Dapena M. 1986. Sudden infant death syndrome. Morphology update for forensic pathologists – 1985. *Forensic Sci Int* **30**:177–86.

Valdes-Dapena MA. 1973. Editorial: Sudden, unexpected and unexplained death in infancy – a status report – 1973. *N Engl J Med* **289**:1195–7.

Valdes-Dapena MA. 1980. Sudden infant death syndrome: a review of the medical literature 1974–1979. *Pediatrics* **66**:597–614.

Variend S, Pearse RG. 1986. Sudden infant death and cytomegalovirus inclusion disease. *J Clin Pathol* **39**:383–6.

Vawter GF, McGraw CA, Hug G, *et al.* 1986. An hepatic metabolic profile in sudden infant death (SIDS). *Forensic Sci Int* **30**:93–8.

Wedgwood RJ, Benditt EP (eds). 1963. *Sudden death in infancy. Proceedings of the 1963 Seattle Conference.* United States Department of Health and Welfare, Bethesda.

CHAPTER 22

Fatal child abuse

■ General aspects of child abuse 461
■ Modes of death in child abuse 462
■ The range of injuries in child abuse 462
■ Surface bruising 462
■ Skeletal damage in child abuse 464
■ Bone diseases and infant fractures 468
■ Head injury in child abuse 469
■ Visceral injury in child abuse 472
■ Damage to eyes, ears and mouth 472
■ Burns in child abuse 474
■ Bite marks in child abuse 474
■ Other injuries in child abuse 474
■ The autopsy in child abuse 475
■ References and further reading 476

There is now a vast collection of publications on the child abuse syndrome, including international journals devoted to the topic. Most of this interest concerns the clinician, as child abuse is now a major concern in paediatrics. When no intervention is offered, an abused child has up to a 10 per cent risk of having eventually fatal injuries, so unfortunately the forensic pathologist will be involved in this tragic problem from time to time. This chapter will deal solely with the physical manifestations of the child abuse syndrome as seen at autopsy, though naturally these will in large measure coincide with the non-fatal injuries seen by clinicians.

GENERAL ASPECTS OF CHILD ABUSE

Though the modern phase of recognition of the syndrome dates only from the middle decades of this century, almost all the classical features were described by Tardieu in 1860 (see Knight 1986). In 1946, Caffey published his well-known first paper and, from then on, an avalanche of articles led to the present intense awareness of the problem that throws into stark contrast the apparent unawareness of the medical profession during the previous century.

Though definitions are variable, it can be said that the ‘child abuse syndrome’ (also known as the ‘battered baby’ or ‘non-accidental injury in childhood’) exists when an infant or child suffers repetitive physical injuries inflicted by a parent or guardian, in circumstances that exclude accident. Most of the fatal victims are young, more than two-thirds being under 3 years of age (Kempe *et al.* 1963; Lauer *et al.* 1974; Brown 1976; Helfer *et al.* 1977).

The syndrome includes, of course, psychological and sexual abuse, but the forensic pathologist can only be concerned with physical damage, some aspect of which has led to death. Though all homicides of children are ‘child abuse’ in the broad sense, the term is usually reserved for a fatal outcome in those who have suffered repetitive ill-treatment, though admittedly some cases span the borderline between chronic injuries and a single episode of trauma.

The definition of ‘child abuse’ is itself controversial, as what is unacceptable now may well have been considered valid and even desirable discipline in Victorian times. However, serious physical injury must be wrong, irrespective of societal attitudes.

Even so, for a pathologist to label an injury as ‘child abuse’ or ‘non-accidental injury (NAI)’ is to be rather pejorative and judgemental, so perhaps ‘adult-induced injury’ might be a more neutral description, which leaves decisions as to motivation to those who possess a wider picture of the whole circumstances. Having said that, the more gross forms of injury are self-evidently ‘abuse’ but, in lesser degrees of trauma, it may be impossible to differentiate deliberate battering from inexpert rough handling, albeit contributed to by exasperation, panic or even attempts at resuscitation.

Though it is vital to recognize non-fatal child abuse, because of the need for intervention to prevent the 60 per cent recurrence rate and 10 per cent mortality rate, it can be almost as tragic to falsely accuse and/or convict parents or guardians where the injuries were non-culpable.

There has sometimes been perhaps some excess of zeal on the part of paediatricians, radiologists, accident surgeons and pathologists to overinterpret injuries and scenarios that had an alternative and less sinister explanation. Though it is natural and vital for doctors to protect children and their siblings, medical opinion has to remain free from emotive bias and be confined within the bounds of what can be proved in each individual case, or injustice may be done. Some years ago, a situation developed where even innocent parents were afraid to take their accidentally injured child to a doctor, for fear of the euphemism 'non-accidental injury' being applied.

MODES OF DEATH IN CHILD ABUSE

The majority of deaths are caused manually, either by hitting or beating with the hands, shaking, throwing, dropping and – less often – by burning or suffocation. It is exceptional for death to be caused by the impact of a blunt instrument, though non-fatal bruises from beating with a strap, for instance, are sometimes seen. Shooting, strangling and stabbing are characteristic of classical homicide, which is distinct from the child abuse syndrome.

The most common mode of death is head injury. Next in frequency is rupture of an abdominal viscus, leaving a wide range of miscellaneous injuries to account for the small remainder.

THE RANGE OF INJURIES IN CHILD ABUSE

Concentrating on autopsy findings, the various types of trauma will be described, though many will not in themselves have caused or even contributed to death. Their recognition is none the less vital as, at the end of the examination, they may well be the decisive factors that may distinguish an accident from deliberate maltreatment.

SURFACE BRUISING

One of the classic aphorisms in the study of child abuse was stated by the forensic pathologists, Cameron, Johnson and Camps (1966), who said: 'The skin and bones tell a story which the child is either too young or too frightened to tell.'

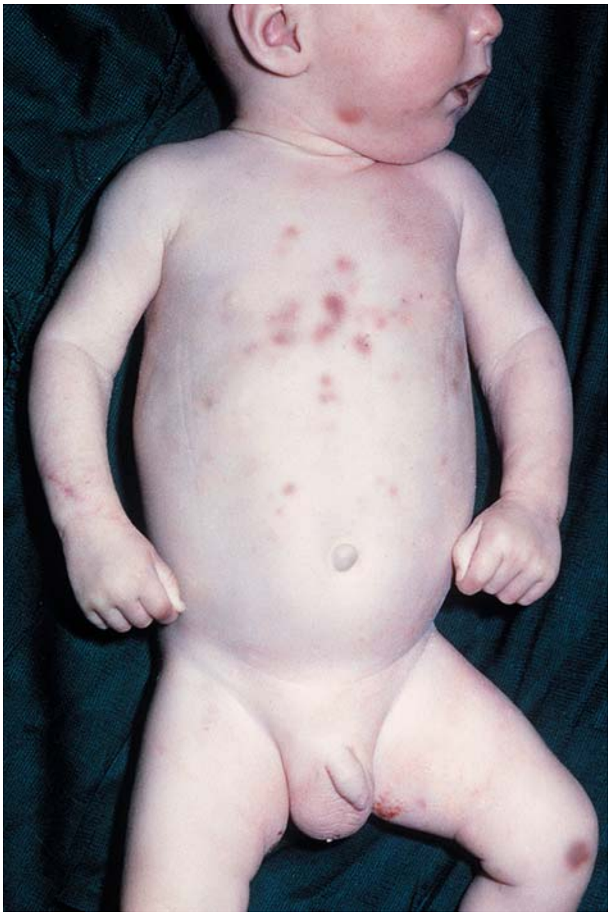

FIGURE 22.1 *Multiple fingertip bruises on the chest and abdominal wall of an infant left with a male babysitter. The child was ill on the mother's return and died during admission to a hospital casualty department (see also Figures 22.4 and 22.5).*

Skin bruising is the most common injury and may be seen almost anywhere on the child's body. There are, however, certain sites of predilection which help to strengthen the diagnosis of abuse:

- Bruising around the limbs, especially the wrist and forearms, upper arms, thighs and – in small infants – around the ankles. These places form convenient 'handles' for an adult to grip the child. In the small infant the lower leg bruises may indicate that the child has been held by the leg or ankle in order to swing it – and may be associated with head injuries. The older child, as an adult, may be gripped by the upper arms in order to be shaken.
- The buttocks are a frequent site of bruising from hand smacks or beating with a strap. Bruises on the thigh are less common, but on the outer side may signify slaps and on the inner, possible sexual interference.

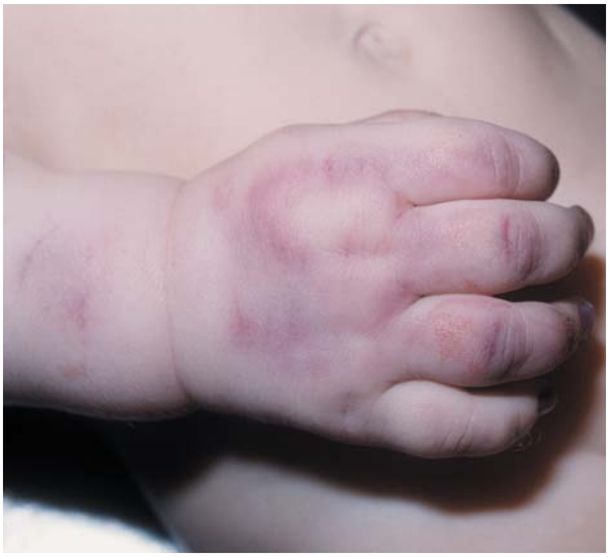

FIGURE 22.2 *Bruising of wrist and back of the hand after maltreatment by mentally ill mother (see also Figures 22.3 and 22.6).*

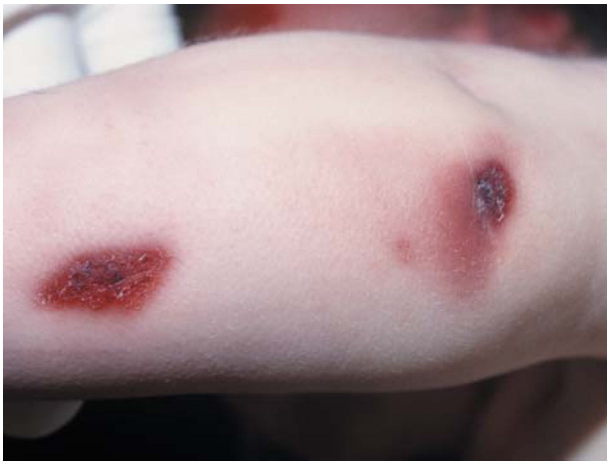

FIGURE 22.3 *Bruising of upper extremity after maltreatment by mentally ill mother (see also Figures 22.2 and 22.6).*

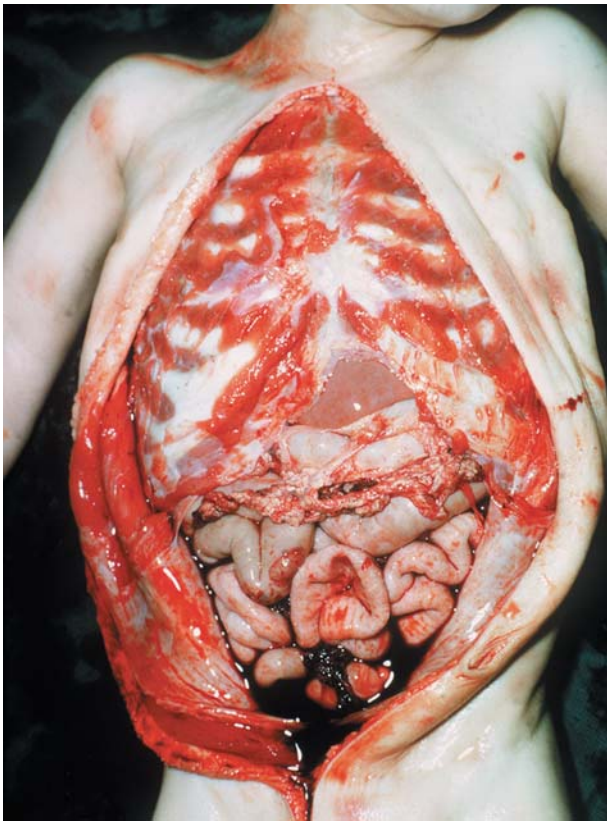

FIGURE 22.4 *Injuries found at autopsy of an infant left with a male babysitter showing a generalized haemoperitoneum.*

- The face is often bruised, especially the cheeks and mouth area, from slaps, which may also be present on the forehead and ears. Associated damage to the mouth and eyes is common. Bruising of the scalp is harder to see because of the hair, but is often part of deeper head injury.
- Bruises on the chest, abdomen and neck are usually from finger pressure rather than slaps or blows. Those on the abdomen and lower chest may be associated with deep visceral injury. Bruises may be of any size or type, but a common variety in child abuse is the small discoid lesion about 1–2 cm in diameter; these were once called 'sixpenny bruises' from the size of coinage at that time. These are caused by impact or pressure from the finger-pads of adults, and may be seen in groups around the limbs and on the neck, chest or abdomen.

Bruises from slapping or punching may be larger and irregular. Occasionally, a partial or even almost whole handprint may be seen outlined in bruises on the buttocks or trunk.

The age of bruises is an important issue in child abuse because:

- The observed age may be at variance with the history given by a parent, so increasing the suspicion of a non-accidental event.
- Bruises of different ages indicate episodes of injury at different times, one of the hallmarks of child abuse that usually continues over a period.

Though it is impossible to be accurate about the absolute age of a bruise, those of markedly different colours cannot have been inflicted during the same episode. The interpretation of patterns and the ageing of bruises are more fully considered in Chapter 4, but any bruise with yellow coloration must be more than 18 hours since infliction.

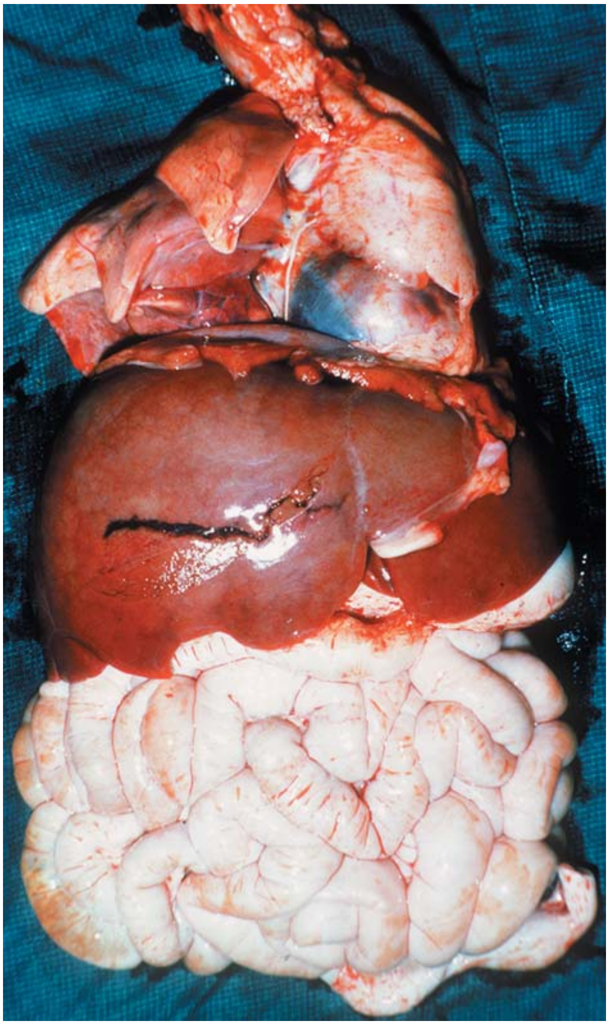

FIGURE 22.5 *The infant shown in Figures 22.1 and 22.4 showing a full-thickness tear extending through the liver. Rupture of an abdominal viscus is the second most common mode of death in child abuse, head injury being the most frequent.*

SKELETAL DAMAGE IN CHILD ABUSE

This is the second part of the aphorism about battered children as, both clinically and pathologically, diagnosis is largely a radiological matter. Just as a whole-body survey is needed in the live victim, so it is essential before an autopsy on a suspected fatal case of child abuse. Some pathologists would recommend a radiological skeletal survey on all infants prior to autopsy if they had not been under constant clinical care for some well-recognized natural disease. This would include all sudden infant deaths but, because of lack of facilities and finance, many pathologists have to confine their requests for such extensive radiology to those deaths in which there was some possibility of abuse, though this decision may lead to injuries being missed.

The range of skeletal injuries seen in child abuse is wide and there are very extensive publications on the subject, including at least one textbook by Cameron and Rae (1975) and a section on the radiology by Evans and Knight (1981).

Worlock *et al.* (1986) compared fractures in 35 abused children with 826 controls and found that the abused were all under 5 years old, whereas 85 per cent of the others were more than 5 years. The abused infants were more likely to have multiple fractures, and to have bruising of the head and neck. Rib fractures were almost confined to abused children, when major chest trauma was excluded. Spiral fractures of the humeral shaft were more common, though classical chipped metaphyses were uncommon. They concluded that one in eight infants under 18 months of age with a fracture had been abused.

The following is a summary of the more important bony lesions.

Fractures of the skull

Skull fractures are common in fatal child abuse, often in association with intracranial haemorrhage, usually subdural, though numerically less than half of infants with subdural haematomata have skull fractures (Harwood-Nash *et al.*). He found that of 4465 childhood head injuries, 1187 had skull fractures, but that there was little correlation with neurological signs and intracranial damage, subdural haemorrhage (SDH) being twice as frequent in the non-fracture cases.

Fractures have been discussed in relation to head injuries in Chapter 5, but some particular features in child abuse must be mentioned, as it is a controversial subject, especially when such cases come to trial.

The most common fracture lies in the occipitoparietal area, but the differentiation from accidental falls is impossible on anatomical or radiological grounds alone, in spite of the dogmatic claims of some radiologists. An infant's skull is more flexible than that of an adult and may absorb some impacts without fracturing. Occasionally, a parietal bone may 'dimple' inwards without cracking, much as a table-tennis ball may be dented with a thumb. The skull of a child is much thinner than that of an adult, even though the brain size and weight is greater relative to the postcranial mass. This means that there are two opposing factors operating:

- Because of the thin skull, less force is required for fracturing.
- Because of the elasticity or 'springiness' of the skull, recovery from distortion caused by impact is greater, so

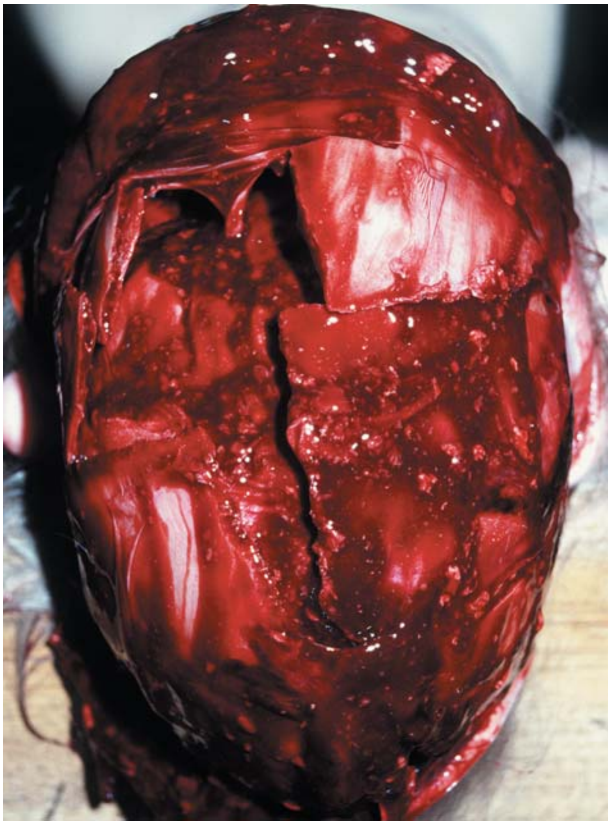

FIGURE 22.6 *Fractured skull of an infant after repeated blows against the floor by mentally ill mother (see also Figures 22.2 and 22.3).*

more force is needed to fracture the pliable bone, relative to its thickness.

Notwithstanding the flexibility, however, it remains a fact that an infant skull will fracture with the application of much less mechanical force than would be needed to fracture a mature skull. In addition, because the cranium is more easily deformed, a momentary depression of the skull can impinge upon the underlying brain (including membranes), damage it and return to its original shape so that, even in the absence of a fracture, brain damage is more likely to occur.

There is now a large literature on infant fractures (especially of the skull) in relation to child abuse, some of it contradictory.

One aspect that arouses much controversy is the height of a passive fall which can (a) fracture a child's skull and (b) cause brain damage. The two injuries are certainly not synonymous, as it is clear that most fractures, both in infants and adults, are not accompanied by any brain damage or neurological effects. However, forces that can cause a skull fracture certainly **can** – though not **must** – cause brain or meningeal lesions, and it is impossible to forecast what will happen following a fall of even minor magnitude.

Though many paediatricians will deny that a passive fall from adult waist level – or even higher – can cause either skull fracture or intracranial damage, there is experimental proof and witnessed cases on record where this has occurred.

In respect of skull fractures, Weber (1984) experimentally dropped dead infants from a horizontal position at only 'table height' (82 cm, 34 inches) onto a variety of surfaces and found that fractures were common. Fifteen infants less than 8 months old were dropped passively from a horizontal position onto concrete, carpet and linoleum; all developed cranial vault fractures in various positions. Though most fractures were in the parietal bones, some extended into the occipital bone. In 1985, Weber again dropped dead infants from 82 cm (34 inches) onto soft cushioning, such as 10 cm (4 inches) thick foam mat and 8 cm (3.2 inches) thick double blanket. Fractures were much less frequent, as was to be expected, but still one occurred on the thick foam and four out of 25 drops onto the folded blanket, all from only 82 cm (34 inches).

Reichelderfer *et al.* (1979) indicated that serious head injuries in playgrounds can occur when the impact force exceeds 50 G; a fall of only 7.5 cm (3 inches) onto concrete can generate 100–200 G and one of 0.3 m (a foot) no less than 475–500 G.

In 246 falls of infants under 5 years from bed or low structures less than 90 cm (36 inches) high, Helfer *et al.* (1977) found no serious injuries or deaths. Of these, 161 occurred at home and 175 had some injury, including two skull fractures. In 85 hospital falls, there were 28 injuries, with one skull fracture. Nimityongskul and Anderson (1987) investigated 76 childhood falls in hospital, 75 of which were younger than 5 years. The height of fall was between 30 and 91 cm (1 and 3 feet), and only one doubtful skull fracture occurred.

Of 398 falls studied by Williams (1991), 106 were observed by independent witnesses. There were 14 severe injuries in falls between 4.5 and 12.20 m (15 and 40 feet), but below 3 m (10 feet), there were no life-threatening injuries, though three skull fractures were detected. Similarly, Reiber (1993) found three skull fractures in infants who had died after witnessed short falls of less than 1 m (3 feet). This was from a series of coroners' cases where all had intracranial bleeding. One 21-month infant fell 1.5–1.8 m (5–6 feet) and had a subdural haemorrhage, but no skull fracture. Another 17-month child fell only 0.6–0.9 m (2–3 feet), but suffered a sub-dural and cerebral contusion. Reiber gives a good short survey of the competing literature, indicating that opinions fall into two groups – the 'major injury–major fall' set and the 'major injury–minor fall' camp.

Hall *et al.*'s (1989) series confirms that severe or fatal damage can occasionally arise from low falls. In a 4-year study, there were 18 deaths following falls from less than 0.9 m (3 feet). Two of these were witnessed by medical care persons, excluding abuse as a covert reason. Sixteen other similar deaths were not so witnessed, but abuse was not corroborated as a cause.

Chadwick *et al.* (1991), analysing 317 childhood falls, saw seven deaths in 100 falls from less than 1.20 m (4 feet), but doubted the veracity of the histories, as there was only one death in those who fell between 3 and 13.7 m (10 and 45 feet). Hobbs (1984) studied 89 children with skull fractures, 29 thought to have been abused. Of 20 deaths, 19 were abused. The characteristics of the battered infant fractures were multiplicity or complex configuration; depressed, wide and 'growing' fractures, the latter meaning widening of the fracture line after infliction; accidental fractures were felt to be narrow, linear and single, usually in the parietal area.

Leventhal *et al.* (1993) investigated 104 infant skull fractures and concluded that abuse was the cause in 34 per cent, accident in 62 per cent, the rest being doubtful. Billmire and Myers (1985) claimed that 64 per cent of all infant head injuries, excluding simple fractures, were caused by abuse and that of serious intracranial damage, 95 per cent were from abuse. These varying figures emphasize the lack of consensus in the literature.

Though the immature skull has more flexibility, once its elastic limit is exceeded during deformation, it will fracture. The pattern of fracturing is similar to that in the adult, but there are some variations because of the presence of open sutures and fontanelles. Fracture lines tend to end at sutures, but if they cross them there is frequently a lateral displacement so that the two limbs of the fracture are not in line. Most instances where what appears to be a fracture line crossing a suture with a 'side-step' are really two independent fractures either approaching or receding from each other, the origins or terminations being slightly offset. The most common example is seen in a child who has been dropped on the vertex of the head. Both parietal bones may then deform and crack transversely from the vertex anterior to the point of impact, the bilateral fracture lines running downwards towards the parietal bosses. The upper ends of both fractures terminate in the sagittal suture, but may be 'staggered' by a centimetre or so.

Another common fracture is a horizontal crack running backwards from the frontoparietal suture, which courses across the parietal bone, often turning down towards the base of the skull. This can be caused by a blow or fall on the side or top of the head. Such fractures can occur bilaterally, which is then an indication of an impact on the vertex of the skull, which causes marked depression of the top of the skull with crack fractures along the lines of maximum stress. These are then more marked on the outer aspect than the inner table – though the diploë of infants' skulls are absent or only partially formed, according to the age. Of course, such bilateral fractures can also be caused by two separate impacts on each side of the head; here multiple bruises on or under the scalp may assist in interpretation.

The frontal bone is less often involved than the parietotemporal area, the frontal suture that is still present in young infants giving more flexibility than in the fused older bone. Occipital fractures may occur from falls onto the back of the head, but again these are much less common than cracks at the vertex or sides of the cranium, though, as the experiments of Weber showed, occipital bone cracks can occur from simple low-level falls in small infants. Wherever the vault fractures, if it is severe enough, extension lines may run down into the base of the skull, but this is not a frequent finding in child abuse injuries.

Sutural 'diastasis' (separation) may occur with or without fractures, the loosely knit skull plates being easily parted by distortion of the calvarium when struck. As mentioned above, some fractures widen after infliction (sometimes due to rising intracranial pressure). These so-called 'growing fractures' are thought by some writers to be more likely to be associated with abuse, though logic would relate it more to severity of impact rather than to motivation.

Limb fractures

These provide some of the most characteristic signs of child abuse, as injuries are common around the metaphyses and epiphyses of growing bones, as well as being the cause of periosteal lesions. Most of the limb injuries are indirect, that is, the bone damage is caused by stresses from abnormal angulation, torsion or traction, rather than from a direct impact upon the bone. Swinging the child by the wrists or ankles, dragging it by an arm or shin and violent shaking using the limbs as 'handles', are the usual mechanisms.

Avulsion of the metaphysis or chipping of the edges of the metaphyses or epiphyses may occur, with small fragments seen isolated on radiographs. Swinging, wrenching or twisting actions can fragment the metaphysis. Small pieces of the adjacent cortex and parts of the provisional zone of calcification may be avulsed from the shaft. The epiphysis may even separate from the metaphysis. Cameron and Rae (1975) suggested that metaphyseal fragments are virtually pathognomonic of child abuse.

The periosteum in infants is only loosely attached to the bone and easily lifts where there is shearing or traction injury. Blood accumulates under the raised periosteum and this rapidly calcifies (usually within 7–14 days) to give a characteristic radiological picture of a bony shell extending along the shaft, often thicker at the extremity, having a lumpy, irregular profile (Evans and Knight 1981). In almost half the cases, this may be seen radiographically within one week (Cameron 1970). The calcified rim may extend around the end of the metaphysis and epiphysis, or between the two, giving a 'bucket-handle' loop effect around the end of the bone, especially the lower end of the femur.

In early infancy it must be appreciated that some breech deliveries have legs that radiologically show long, smooth periosteal thickening secondary to subperiosteal bleeding caused by handling during delivery. It is also accepted that mere growth can produce faint subperiosteal calcification, but this is symmetrical and limited to the central parts of the shaft, keeping clear of the metaphyses.

A spiral fracture of the diaphysis of a long bone must be considered a suspicious injury in infants, as such a lesion is likely to be the result of a twisting strain, unlikely to occur in accidental circumstances.

The lower end of the shaft of a long bone may be 'squared-off', as described by Caffey (1946), by metaphyseal fragments reuniting. Damage to the epiphyseal cartilage may cause permanent growth defects, sometimes with increased growth from hyperaemia following injury. The axis of growth may be deviated and cupping of the metaphysis may occur because of retarded longitudinal growth in the centre.

Although the metaphyseal and epiphyseal lesions described above are classically those seen in child abuse, transverse and spiral fractures of the diaphysis are also common. In a series of 100 cases reported by Kogutt *et al.* (1974), these fractures were more common than the injuries to ends of the bones.

Damage to the chest cage

Apart from occasional perinatal fractures during delivery (Polson, Gee and Knight 1985), accidental fractures of ribs are very uncommon in infants, though the incidence is a matter of considerable controversy. Once again paediatricians, radiologists and pathologists may have differing views and experience.

In child abuse, ribs are commonly fractured, often several consecutive bones being affected on one or both sides. The damage may be fresh or old, giving different radiological appearances. The most characteristic picture of old fractures is of sequential callus formation seen in a vertical line down one or both paravertebral gutters. These form a 'string-of-beads' appearance on X-ray (Figure 22.7), where the ribs have been fractured near their angles, a sufficient time previously for new bone formation to occur – probably a minimum of 10 days, though dating is notoriously variable, and even the claims by experienced radiologists to date fractures have been confounded by a knowledge of the actual time of breakage.

These multiple posterior fractures are more common in small infants who have been picked up by adult hands under the armpits and squeezed from side to side. Whether or not such fractures can be caused by innocent handling, albeit rough or robust, as opposed to angry or impatient violence, is again a matter of dispute, beyond the competence of doctors to resolve – although this does not deter some from expressing strong opinions based on weak facts.

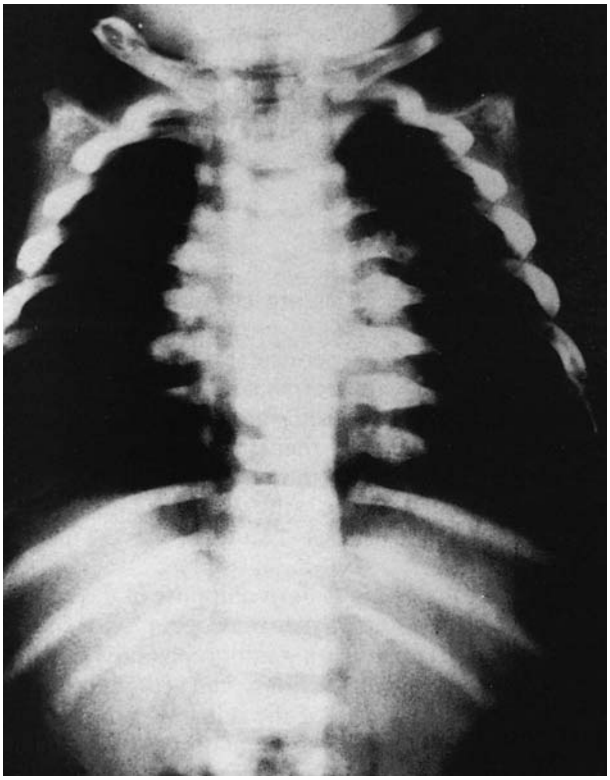

FIGURE 22.7 *Chest radiograph of an infant showing multiple callus formation on the necks of the ribs, giving a 'string of beads' appearance down the paravertebral gutter. These indicate previous multiple fractures caused by lateral squeezing by adult hands.*

These fractures, a centimetre or so from the head of the ribs, are due to the rib being levered against the transverse process until breaking point is reached. They may be very difficult or impossible to see as fresh lesions on radiographs, if the fracture line is not exactly in line with the X-ray beam. This explains the occasional discrepancy between the radiologist's and the pathologist's report; the fractures are much more visible after a week or two, when callus is present.

If such injuries are recent (probably less than 10 days) no visible callus is formed, but the fractures may be seen on X-ray and confirmed at autopsy. However, as with fractures elsewhere – and perhaps more often in ribs than skull – the radiological and dissection diagnosis sometimes does not concur. Fractures seen at autopsy may not be visible radiologically – and conversely, alleged fractures seen on films may not be discoverable at autopsy, even after the most diligent dissection.

Rib fractures in the axillary line are likely to be the result of anteroposterior pressure, rather than side-to-side squeezing, but this mechanism is not exclusive. With fresh fractures found at autopsy, the possibility of chest compression

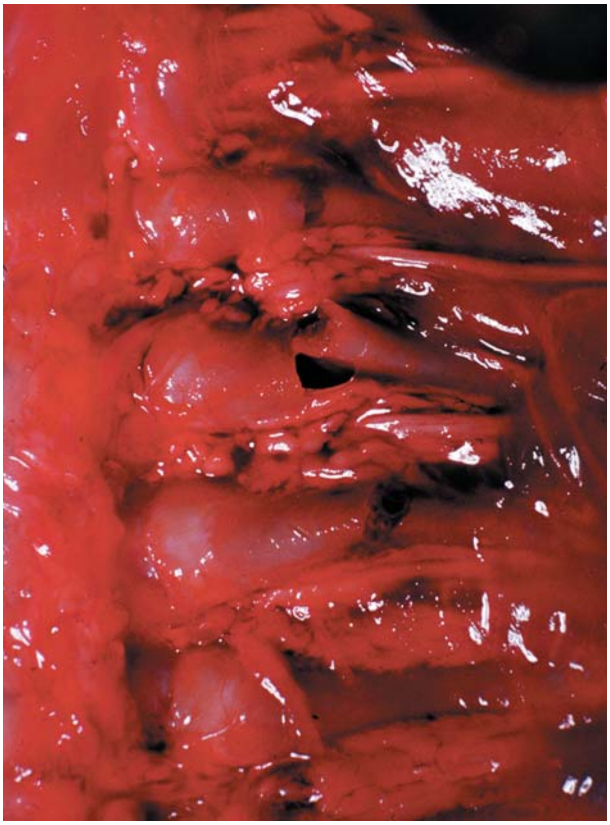

FIGURE 22.8 *Recent fractures of ribs of an abused infant. These breaks are near the necks of the ribs in the paravertebral gutter, caused by side-to-side squeezing of the chest.*

during cardiopulmonary resuscitation must always be considered, even though some paediatricians and radiologists will strenuously deny the possibility of this happening. The literature has different opinions on this matter, though it is admittedly very uncommon. As with other aspects of child abuse, the published material seems to flow more from those who lean towards abuse in the majority of injured children, rather from those who retain a more cautious approach and admit to the possibility of an innocent explanation.

It must be conceded that pliable infant ribs are unlikely to be broken by proper cardiac massage, which in infants should be performed with finger pressure; however, lay persons, especially in the panic of apparent collapse and death, may forcibly pump the small chest using techniques intended for adult resuscitation.

Fractures elsewhere are more likely to arise from direct impact such as a fist blow or kick. The clavicle is sometimes broken, usually from indirect stress from swinging by an arm. Fractures of the scapula or sternum are rare and arouse the probability of 'parental infliction' if some obvious trauma such as a traffic accident can be excluded.

Wherever the site of bony injury, the cardinal features that arouse suspicion are multiplicity and variation in fracture age. Though children often sustain accidental fractures, these are usually single lesions, unless some pathological bone disease is present. Similarly, the radiological demonstration of multiple fractures of varying ages must be accepted as evidence of abuse until proved otherwise.

Dating of fractures

As in the case of skin bruises, though the absolute dating of healing fractures by the state of callus formation is far from accurate (Evans and Knight 1981), marked differences in appearance on radiology usually indicate that the injuries were not sustained at the same time. As it can be shown, however, in control material, that the same person reveals callus of different extent and maturity in different but similar bones fractured contemporaneously, it becomes impossible to be dogmatic about radiological healing rates. Bone healing is rapid in healthy children and within months the abnormalities may have completely resolved, depending on the severity of the original damage.

Histological appearances are also very unreliable in placing an absolute date on fractures, in spite of assertions that this can be done with accuracy.

BONE DISEASES AND INFANT FRACTURES

Both in the differential diagnosis of childhood fractures and in the defence of allegations of child abuse, the several 'brittle bone diseases' that can cause abnormal skeletal fragility must be considered, on the grounds that lesser force, within normal parental handling or even spontaneous movements of the infant, could give rise to the observed fractures.

It has to be said at the outset that only a small minority of injured infants can be shown to have such abnormalities. When this possibility is being run as a defence in a criminal trial or child protection proceedings, it is essential to have an opinion from a radiologist or paediatrician with specialist knowledge of these conditions.

It is also important to have original radiographs, not copies, as the latter may not show marginal, indistinct abnormalities visible on the originals.

Osteogenesis imperfecta

This is a bone dysplasia with four types, numbered I, II, III and IV, of which there are several subgroups. The disease is an inherited disorder of connective tissue, with abnormal collagen, resulting in varying degrees of bone fragility, ligamentous laxity, skin fragility, sometimes blue sclera, hearing defects and dental abnormalities.

Types II and III have obvious bone disease and can hardly be confused with child abuse lesions.

Type I forms 70 per cent of cases and is the classical variety with a family history, blue sclera, Wormian bones and often dental changes, which again is hardly likely to be confused with abuse. It is type IV (only about 5 per cent of the overall disease incidence) that may cause difficulties in diagnosis, though it must be said that of all the medico-legal occasions in which this disease is postulated as a reason for brittle bones as a defence against allegations of abuse, the times upon which it can be proven are few indeed.

Type IVB, the larger subgroup, has dentinogenesis imperfecta. Type IVA has no family history of the disease, no osteoporosis, a high incidence of skull fractures and metaphyseal fractures and normal sclera. Subgroup IVA has Wormian bones, like IVB, but it has been postulated that some alleged cases of type IVA were really abuse.

It has been calculated that the incidence of type IV osteogenesis imperfecta in infants under one year, with fractures, no family history, no Wormian bones or dentinogenesis imperfecta is one in 1–3 million. A case would occur in a city of half a million population only once every 100–300 years.

Recent techniques of fibroblast culture and collagen analysis can provide evidence of defective connective tissue in the absence of overt clinical signs, but this is not necessarily synonymous with brittle bones. The case report by Ojima *et al.* (1994) gives a useful survey of the problem. Bone strength in otherwise normal SIDS infants has been measured by Morild *et al.* (1993) to obtain a baseline for comparison with alleged osteogenesis where this is raised as a defence against child abuse.

Infantile cortical hyperostosis (‘Caffey’s disease’)

Here there is extensive new periosteal bone laid down around the diaphyses, the bones most involved being the ribs, ulnae, clavicles and especially the mandible. The condition is rarely seen over the age of 3 years.

Congenital syphilis

This can mimic the changes seen in child abuse, though the condition is extremely rare in most Western communities. The periosteal thickening is similar to that of trauma, but tends to be symmetrical. The metaphysis may be fractured and even separated, the shaft adjacent to the metaphysis being porotic.

Copper deficiency

Again rare, this condition has radiological features in which periosteal thickening of the shafts of the long bones may be accompanied by symmetrical spurring of the metaphyses, osteoporosis, fractures and cupping of the metaphyses. Abnormalities of the hair may be present in this deficiency disease that is caused by lack of copper impairing the enzyme lysyl oxidase. The radiological appearances may cause confusion with child abuse. Other clinical features include psychomotor retardation, hypotonia, pallor, hypopigmentation of the skin and hair, prominent scalp veins, sideroblastic anaemia and neutropenia.

Menke’s syndrome (‘kinky hair’ syndrome)

This is another rare condition also associated with copper deficiency in which there are abnormal metaphyses. It affects males only, with abnormal hair, Wormian bones in the skull and mental retardation.

HEAD INJURY IN CHILD ABUSE

The most common cause of death in abused children is intracranial damage, with or without skull fracture. As well as being the most frequent fatal condition, brain trauma is responsible for the common tragedy of severe, and often permanent, neurological impairment.

The classic head injury associated with battered children is subdural haemorrhage, which formed part of ‘Caffey’s syndrome’ that he described five decades ago. It was also noted in Tardieu’s famous publication of 1860, when he described bleeding on the side of the cerebrum and other classical features (Knight 1986). Subdural haemorrhage is most common at the extremes of life, the topic being discussed fully in Chapter 5. In infants, it most often occurs from direct impact on the skull – as in a blow or fall. For some years, it has been held that vigorous shaking, without an impact, is also a common cause of subdural bleeding. This is a view strongly held by many paediatricians, who may prefer this mode as a first choice over a blow or fall, but recent research has thrown doubt on the common acceptance of this mechanism.

For example, the excellent textbook of DiMaio and DiMaio (1989) observes:

> Many forensic pathologists have doubted that the shaken baby syndrome exists [Norton 1983]. Most have felt that the injuries to the brain, including the retinal haemorrhages, are due to undetected direct impact

injuries. In all cases seen by the authors in which there has been retinal haemorrhage associated with subdural or subarachnoid bleeding, or some other brain trauma, direct impact injury to the head has been identified.

The forces generated in the cranial cavity by shaking have been shown to be of the order of 50 times less than in the deceleration stresses of an impact (Gennarelli and Thibault 1982; Duhaime *et al.* 1987). These authors investigated 13 fatalities alleged to have occurred from shaking, but though 7 showed no external head injury, all 13 revealed evidence of impact at autopsy. In infant models, they then compared impact with violent shaking; the mean tangential acceleration for 69 shaken babies was less than 10 G, whereas for 60 impacts, the mean was 428 G, some 50 times greater. The mean time for shakes was 106 milliseconds, but for impacts was only 20 milliseconds.

It is the rate of change and the duration of deceleration – the 'strain rate' – that is most damaging, rather than a steady deceleration. A head impacting on an immovable surface after a fall is an example of a high strain-rate injury, whereas low strain-rate injuries are those where the head decelerates over a longer period of time. Non-impact shaking of the head is a low strain-rate mechanical insult. Subdural bleeding from bridging veins is likely to occur from high strain-rate injury, even though this may be a low-energy injury, insufficient to cause cerebral tissue disruption. By contrast, low strain-rate injuries are more likely to cause cerebral contusion, with vascular damage only if concomitant high energy is generated (Howard *et al.* 1993).

Geddes and Whitwell (2004) have recently pointed out that there have been no formal systematic neuropathological studies of infant head injury and the evidence base for diffuse axonal injury (DAI) being a common finding in infant head injury is poor. In fact the idea of the presence of DAI in these cases was widely accepted before the advent of the modern diagnostic methods and before the diagnostic criteria for DAI had been established. Moreover, the findings of Shannon *et al.* (1998), the first formal study of microscopic damage in non-accidental injury (NAI), suggesting that axonal injury in such cases had been ischaemic or vascular in origin, had been largely ignored or misunderstood. In their recent series of 53 fatally injured infants (Geddes *et al.* 2001a, b), diffuse axonal injury was present in only two children, both of whom had multiple skull fractures and severe head injury, whereas the most frequent histological finding was global hypoxic damage in 84 per cent of the cases. In 11 of these cases (21 per cent), however, there was gross and microscopic evidence of injury to the craniocervical junction and concomitant axonal damage in the caudal brainstem long tracts in 8 subjects. This finding is in keeping with cervical hyperflexion/hyperextension injury pattern in adults and suggests as a possible mechanism of death: brainstem injury, apnoea and hypoxic brain swelling. In another systematic study of dural bleeding in 50 paediatric cases, including intrauterine deaths up to infants of 5 months of age, but without head injury, Geddes *et al.* (2003) found fresh haemorrhage into the layers of dura in a majority (36/50) of cases. The authors believe that this is a manifestation of severe hypoxia similar to the haemorrhages found in internal organs in birth asphyxia or prematurity, rather than being caused by trauma, and postulate that hypoxia is sufficient to cause extravasation of significant amounts of venous blood both in and under the dura.

Although shaking presumably may cause subdural haemorrhage (SDH), it is likely that it is a relatively uncommon cause, compared with impact. This situation may well have arisen because a blunt impact upon the head of infant, if spread over a wide area following contact with a flat surface, can leave no external scalp mark, no subscalp bleeding and no fracture of the skull – yet the transmitted forces can still be sufficient to cause high strain-shearing stresses within the cranium leading to subdural bleeding. In addition, as has been discussed elsewhere in this book, death is often not due to the irritative or space-occupying effects of the subdural, but is due to the intrinsic brain damage beneath, even if this is occult. Diffuse axonal injury, cerebral oedema and consequent circulatory disturbances are likely to be the fatal consequences, but detecting neuronal damage may be difficult or impossible histologically, due to the relatively rapid death of the child within 12–24 hours. Also, the microscopic features of diffuse axonal injury seem to be less common and less distinct in infants compared with adults, though the use of immunocytochemical methods, such as the demonstration of β-amyloid precursor protein (βAPP), described in Chapter 5, may provide a marker of use within the first few hours after injury. These factors combine to negate any proof of an impact injury and thus the shaking theory, which requires none of these criteria to be demonstrated, is an easier option.

Howard *et al.* (1993) looked at 28 instances of SDH in infants under 18 months of age. All had a history more consistent with an impact rather than shaking. Six Caucasian infants fell less than 0.9 m (3 feet), such as from a chair or standing adult. Eight non-Caucasians fell from a sitting or standing position or rolled off a bed onto a carpeted floor. There was a history of shaking in three, but these also had signs of impact, the shaking being resuscitation attempts after falling. Of these, 47 per cent of the Caucasians and 20 per cent of the non-Caucasians had a skull fracture. The authors state: 'Our findings do not support shaking as the only cause of SDH and also suggest that non-accidental injury is a less common cause of SDH than it is believed to be.'

There is a suggestion in several papers that Asian children are more susceptible to SDH from minor trauma;

Japanese infants have been reported with SDH after falling from only a sitting position (Aoki and Masuzawa 1984; Rekate 1985).

Most direct impacts in child abuse are caused by the moving head striking a fixed object, rather than the fixed head being struck. The latter certainly occurs, but is usually from an open-handed slap, albeit often of considerable force. It is unusual for the child's head to be struck with a weapon or for a direct punch to be given with the clenched fist. The moving head injury either arises from a fall in circumstances which are non-accidental, or by deliberate swinging or throwing of the child against a hard surface, often with added momentum being gained by holding the infant's ankles or wrist. For example, one such case seen by the author (BK) was a month-old baby whose father grasped it by the ankles to dash its head against stones at the edge of a canal, after which it was stripped so that its clothing could be sold in a street market, the body then being thrown into the canal.

Several other mechanisms of head injury have been described, including some seen by the author (BK) in which fathers were throwing their baby up in the air in a catching game called 'up-and-away', but for some reason missed the child on the downward return, allowing it to fall to the floor. When unwitnessed, this situation can give rise to accusations of deliberate malice, which are difficult to disprove.

Presumably, as vigorous shaking can sometimes rupture vessels in the subdural space, this must still be accepted as an alternative to impact in the causation of subdural haemorrhage. In an infant, the head is relatively large and heavy and the neck musculature undeveloped and often hypotonic. Thus repetitive shaking of the body – often attained by grasping under the armpits – can lead to unrestrained flexion and extension movements of the head with rotational movements of the heavy cerebral hemispheres relative to the skull.

Extradural haemorrhage is uncommon in child abuse, perhaps because the meningeal arteries do not run in deep grooves or even tunnels in the immature skull, as they do in the adult. Subarachnoid haemorrhage is, as always, an inevitable accompaniment of cerebral cortical contusion.

Brain damage may be of any type, as catalogued in Chapter 5, but as in any childhood head injury, cerebral oedema is more common than in the adult. Indeed, a significant proportion of deaths after a head injury in a child are caused by raised intracranial pressure without skull fracture, meningeal haemorrhage or visible brain injury. The typical autopsy indications may be found consisting of a heavy brain, flattened gyri, obliterated sulci and slit-like ventricles. The more florid signs seen in adults, such as herniation of the hippocampal gyrus through the tentorial gap, secondary brainstem haemorrhages and coning of the cerebellar tonsils are seen less frequently in children.

When there has been a severe head injury, with or without a fractured skull, cerebral contusion or laceration may occur. The position of this damage is usually related to the site of external impact. Although many episodes of child abuse involve dropping or throwing to the floor, classical contrecoup lesions are not as common as in similar adult injuries. They certainly do occur, but are often absent when the circumstances would indicate that contralateral contusion might be expected. The smoother internal profile of the infant skull may be a factor.

The distance of a fall sufficient to cause head injury has been mentioned earlier and is a matter of considerable controversy in the literature and amongst expert medical witnesses. There is no doubt, both from reliable anecdotal evidence and from experimental data, such as that by Weber, that skull fractures can occur in infant skulls from very low passive falls, including heights not exceeding chair or table level (Weber 1984: 82 cm; 34 inches). However, as catalogued above, skulls often fracture without any significant internal brain damage – although the reverse is also true. The evidence that low falls may cause brain or meningeal lesions is much less convincing than that proving skull fractures – but the possibility exists and cannot be dismissed by inflexible, dogmatic opinion. The problem with assessing the mechanical basis of head injury is that human experimentation is virtually impossible – Weber's work on dead infants probably could not be repeated in the present ethical climate – and obviously no data can be derived from the living. Animal models are useless, so the frequent difficulties for pathologist, lawyers and the courts have to rest on available statistical evidence – which can never give a definitive answer in any one case. If a phenomenon has definitely occurred once, then it can occur again, whatever the statistical weight is against its frequency, as a precedent has been established which cannot be gainsaid as a possibility.

One of the problems is the 'standard of proof', as in the criminal courts, it is 'beyond reasonable doubt', whereas in civil matters, it only has to pass the test of 'balance of probabilities', where statistical evidence is far more persuasive.

To rely on the 'experience' of medical witnesses is also fraught with difficulty, as their anecdotal memories of previous cases will provide unsatisfactory data in respect of the causation of the injury. Where adults are accused of malice or neglect, they are very likely to fabricate or distort the circumstances of the injury, and claim that the infant fell from the chair, the arms or the bed. Even in genuine accidents, feelings of guilt, clouded recollection from grief and panic, will often lead to an inaccurate account of the circumstances, so the doctor's past experience is often of lesser value than the accused adults, the lawyers and the court imagine, due to inadequate or false knowledge of the true circumstances of previous cases.

VISCERAL INJURY IN CHILD ABUSE

Damage to internal organs is almost always confined to the abdomen, as the heart and lungs are rarely injured. As stated earlier, rupture of an abdominal viscus is the second most common cause of death in child abuse after head injury. Forcible impacts on either the lower chest or the abdominal wall are responsible. Direct punching or heavy 'prodding' are the usual mechanisms rather than being dropped or thrown, which is a common cause of head injury. Often there will be an excuse that the child fell or tripped upon some protruding obstruction, such as a toy or piece of furniture; occasionally this may be true and it is a matter of fact and interpretation as to whether the circumstantial evidence is compatible with the medical findings. The liver is frequently injured, the most common lesion being a deep tear in either lobe, sometimes being a complete penetration or, more rarely, actual detachment of hepatic tissue. A haemoperitoneum results and in fatalities, may be the proximate cause of death. A heavy blow over the lower ribs or xiphisternal area is the most likely site of impact.

The small intestine is the other common target organ, the duodenum or jejunum being the most frequent site of damage. The second part of the duodenum is vulnerable to blows in the central abdomen, as beneath this area, the duodenum crosses the midline and is liable to be 'sandwiched' between the compressed anterior abdominal wall and the promontory of the lumbar spine. As the tissue thickness is small in a child, the gut can be virtually 'guillotined' at this point and may be completely transected, appearing almost as clean-cut as if it had been done with a surgical knife. Alternatively the gut wall may be damaged and, though it may not leak or rupture at the time of impact, ischaemic and necrotic changes over the next few hours or even days may lead to a delayed rupture, with the later onset of peritonitis.

The same lesions, though rarely as sharply cut, can occur in the jejunum, often accompanied by laceration of the mesentery and haemorrhage into the peritoneal cavity instead of – or in addition to – leakage of intestinal contents.

Rupture of the stomach can occur, but must be carefully differentiated from gastric rupture from non-traumatic causes and (these days) especially from external cardiac massage, which produces linear tears in the mucosa and even rupture of the complete wall.

DAMAGE TO EYES, EARS AND MOUTH

As part of general facial damage, black eyes, and scleral and conjunctival haemorrhages are often seen, but damage to the interior of the globe is much less apparent at autopsy. Some considerable time after the general recognition of the child abuse syndrome, it was discovered that a considerable proportion of victims, both live and dead, had significant internal damage to the eyes. Direct violence to the head and allegedly shaking, can cause bleeding into the vitreous humour, dislocation of the lens, retinal detachment and retinal haemorrhages (Gilliland *et al.* 1994).

The latter can also be caused by cardiopulmonary resuscitation, where vigorous compression of the chest has been applied; the literature on this topic is controversial, but numerous references confirm this mechanism, yet others,

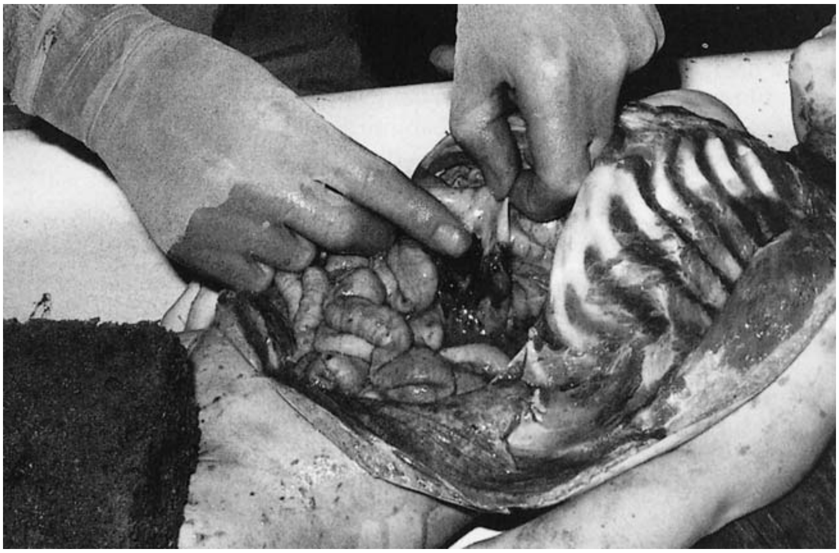

FIGURE 22.9 *A ruptured duodenum in an abused child who suffered a blow in the central abdomen. This relatively common mode of death in fatal child abuse is caused by the duodenum being guillotined against the lumbar spine by an impact that indents the thin anterior abdominal wall. The gut may be severed immediately or may become infarcted because of damaged blood supply over the ensuing few days, then rupture into the peritoneal cavity.*

such as Gilliland and Luckenbach (1993), deny it. Once again, there are hotly contested viewpoints in the position taken by different doctors over the aetiology of various lesions in child abuse, both by clinicians and pathologists. They are also, of course, often secondary to raised intracranial pressure from head injuries, rather than due to primary direct violence. During life, these can be diagnosed by ophthalmoscopy, but at autopsy they must be sought by dissection. As disfigurement of the body of a child must be avoided wherever possible, examination for eye lesions may be achieved either by enucleation and replacement with glass prostheses – or by opening the orbits from the anterior cranial fossae and removing the posterior two-thirds of the globes.

Injury to the external ear is frequent, as abuse of a child is often directed at the obvious targets, such as face, head, arms, buttocks and ears. The pinnae may be bruised or even lacerated, usually by an open-handed slap. Damage to the internal ear from heavy impacts on the head may be diagnosed clinically, though often in retrospect from later deafness. At autopsy, damage to such minute structures is rarely visible, though sometimes haemorrhage into the petrous temporal bone may be found when routinely examining this area for middle-ear sepsis.

Injury to the lips and mouth is so common as to be one of the prime diagnostic signs of the battered child. Infants and older children are commonly slapped across the mouth, the swipe usually having a tangential component that moves the tissues laterally. This causes the most reliable indicator of child abuse, the torn frenulum (or ‘frenum’) beneath the upper lip. This small median band of mucosa that joins the upper gum to the upper lip is ripped as the lip is jerked sideways, and is virtually pathognomonic of battering. The other possible cause of a ruptured frenulum is the forcible entry of a feeding bottle teat into the mouth of a reluctant infant. Instead of entering between the gums, the teat is rammed up between the gum and lip, causing a tear. The lesion heals rapidly, as do all injuries in the mouth and, within a few days, nothing may be seen except perhaps a loose mucosal tag or an absent frenulum, or both.

Damage to the lips themselves is common, the type of injury varying somewhat according to whether or not the child is old enough to have teeth. In any event, bruising, swelling, abrasion and sometimes external laceration may be seen. The inner mucosal aspect of the lips may be bruised even if the child has no teeth and, if the blow is hard enough, laceration can occur. This is far more likely to happen if the lips are forced against the teeth, especially the cutting edge of the incisors.

The teeth themselves may be damaged, especially in older children. Total avulsion from the sockets, loosening and breakage may occur. In small infants, developing teeth may be forced back into the sockets, like pegs into holes. Expert dental advice should be sought wherever possible when injuries to teeth and jaws are found, both in live children and at autopsy.

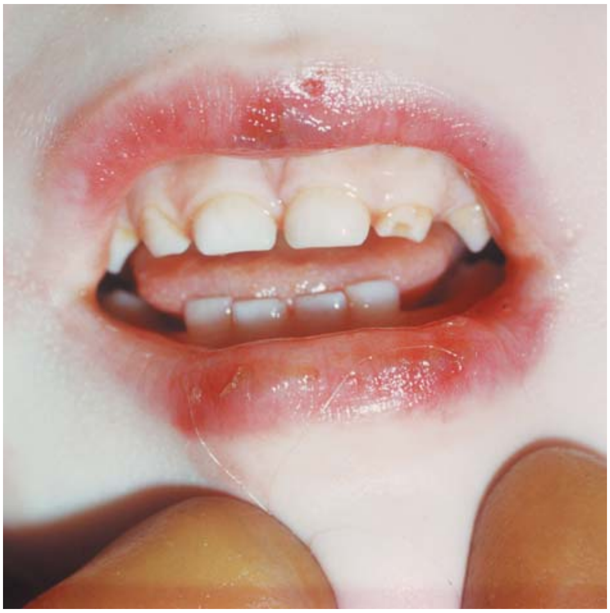

FIGURE 22.10 *Bruising of the lips in child abuse caused by smacking in the mouth. When the child is old enough to have teeth, lacerations may occur inside the lips. The frenulum behind the midline of the upper lip (not shown) was ruptured in this child, a strong pointer to a blow across the upper lip.*

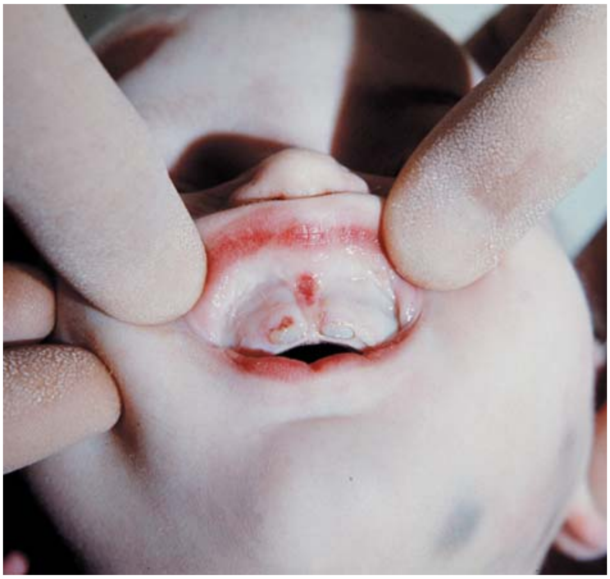

FIGURE 22.11 *Tearing of the frenulum of the upper lip of a child, a classical injury from a tangential slap across the mouth.*

BURNS IN CHILD ABUSE

Unfortunately, thermal injuries are by no means uncommon in abused infants. These may be moist scalds, dry burns or, exceptionally, electrical injuries.

Scalds result from dipping in hot fluid and often involve over-hot bath water. The line between accident and abuse may be blurred in some cases, but it can occur as a form of chastisement. In one case, a child was deliberately held in water at over 80°C as a punishment for persistently complaining about being cold. Other scalds may be from boiling water from a kettle or saucepan deliberately being poured over the child.

Dry burns can be inflicted in innumerable ways as a deliberate and sadistic act. Children have been forcibly sat on electric cooking rings and hotplates, branded with hot shovels and electric soldering-irons, or pressed against the bars of a fire. A particular type of burn seen relatively often in abused children is never the cause of death but may be found incidentally. This is the cigarette burn, seen most often on skin not normally covered by clothing, such as the hands, arms, neck and head. Such burns are usually circular, but not always; if the cigarette has been held obliquely against the skin, the mark may be triangular. The regular shape and size usually indicate the nature of the burn, though some skin diseases may simulate an old burn, such as a small patch of impetigo. Fresh cigarette burns are red, sometimes with a narrow rim of a deeper red. When healing, they become pink and later have a silvery sheen on the surface.

BITE MARKS IN CHILD ABUSE

Most bites on the skin surface are seen either in sexual offences or in child abuse. In the author's own experience, all bitten infants have had these inflicted by the mother, though published reports indicate that both male and female guardians can bite their infant charges. The topic is discussed in more detail in Chapter 26, but as they are relatively common in child abuse, a summary is repeated here.

Bites may be seen anywhere on the child's body. The arms, back of the hands, cheeks, shoulders, buttocks and abdomen are favoured sites. A bite mark usually consists of two opposing semicircles, which may be incomplete. Individual teeth marks may be visible or the dental arch may be represented by a continuous line. The mark is either an abrasion or a bruise, or a combination of both. There may be confluent petechiae in the centre caused by suction, but this is much more common in sexual assaults – the so-called 'love bite'. In child abuse, the centre is more often undamaged. The size of the dental arch should be carefully measured to determine whether it was formed by an adult, another child or even an animal, as is sometimes alleged in defence.

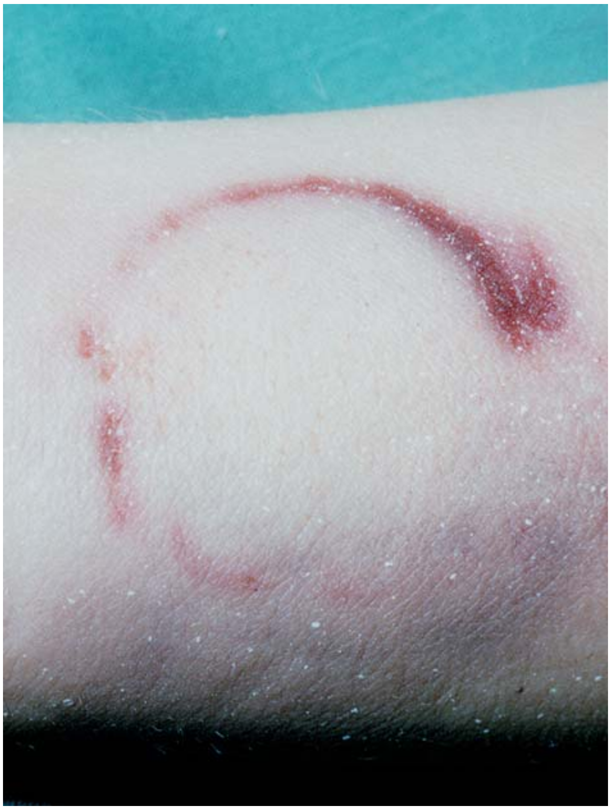

FIGURE 22.12 *Bite mark on baby's arm in a case of child abuse. The culprit was the mother. The teeth marks are deficient over one quadrant, the gap on the other side being smaller than usual.*

In older children, the possibility of self-biting must be borne in mind – the author has seen multiple self-bites all the way up one arm in a young person.

As with all dental matters, the advice of a dentist experienced in forensic problems is highly desirable. Not only can he confirm or exclude doubtful lesions, but he can identify individual teeth marks, make casts of any residual impression in the skin and take swabs for saliva identification. He can also examine any suspects to compare their dentition with the marks – a common excuse of the parents is that another child or even the family dog caused the bite. If no forensic odontologist is available, however, the pathologist must make the best examination possible and details are given in Chapter 26.

OTHER INJURIES IN CHILD ABUSE

A variety of injuries, some bizarre and sadistic, occur from time to time in battered children. The hair may be pulled out in clumps (epilation), leaving pseudoalopecia on the

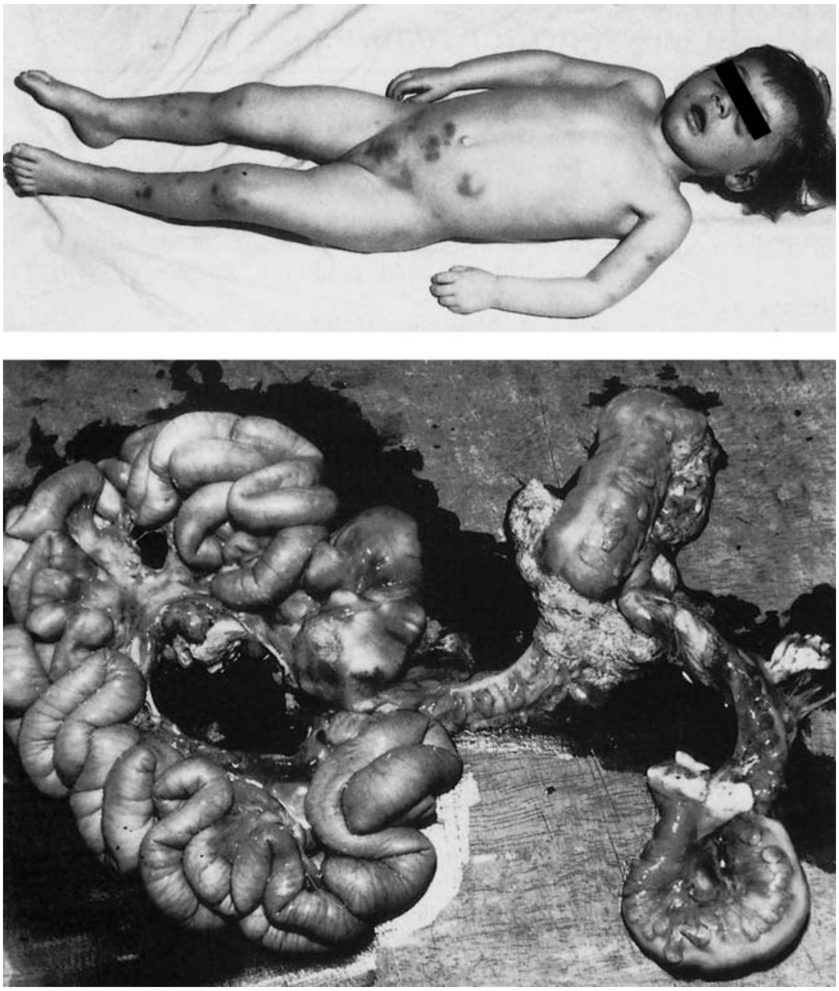

FIGURE 22.13 *A child with numerous bruises on the body, totalling 80 on the face, neck, abdomen, back and legs. The male adult in the house claimed that she had 'slipped while being bathed'. At autopsy, there was a massive haemoperitoneum arising from multiple ruptures of the mesentery.*

scalp. Fingers may be broken by hyperextension and fingertips crushed by blows or other means. Patterned marks from straps and ropes may be seen on the skin, usually on buttocks, thighs and abdomen. 'Pinch' or 'tweaking' marks are not uncommon, made by the skin being nipped between adult fingernails. These appear as two small opposing semi-circular or triangular bruises, often with a clear zone between them.

THE AUTOPSY IN CHILD ABUSE

This follows the usual routine for any forensic autopsy, but certain additional matters must be given close attention. Because of the emotive nature of child abuse, especially where a fatality is concerned, the pathologist must pay meticulous attention to all aspects of his examination and report. A full description of the autopsy in child abuse may be found in the chapter by Knight in *Paediatric Forensic Pathology* (Knight 1989). The following matters must be borne in mind:

- The scene should be visited wherever possible, though usually the child will already have been removed, usually to a hospital accident department. Even a retrospective visit can be helpful, however, to assess such matters as the nature of the floor and its coverings, where falls are alleged, the height of settees and chairs, and the general size and space between walls, doors and furniture, as claims of colliding with household structures are a constant line of defence – and may well be true in some instances.
- The child should be examined in its original clothing where possible.
- The appearances before washing and other cosmetic manoeuvres should be recorded; when children die in hospital accident departments, such compassionate acts may remove useful evidence, and the balance between

sympathy and justice can sometimes be hard to reconcile.

- Careful enquiries should be made about any resuscitatory attempts, as these can introduce a number of artefacts.
- Where any suspicion of abuse is present, a full radiological skeletal survey should be carried out before the autopsy is begun.
- Full physical measurements must be made including weight, crown–heel length, crown–rump length, foot length and head diameter.
- A meticulous external examination must be made, covering every square centimetre of the body, including all body orifices. The interior of the mouth and lips must be searched.
- A full photographic record must be made of any abnormality discovered. Where bruising is suspected, but not definitely apparent, the use of both infrared- and ultraviolet-sensitive film has been advocated, though experience is required in their interpretation, as artefacts are common.
- Extensive dissection of subcutaneous tissues is required wherever bruising is seen or suspected. Incisions must be made into doubtful areas or the skin flayed away to reach such areas, if disfigurement is to be avoided.
- Full sampling of blood, urine, stomach contents, liver, ocular fluid and cerebrospinal fluid must be taken, according to the circumstances. It is best to be liberal with the range of samples – it is easy to discard unwanted samples, but difficult or impossible to obtain them at a later date.
- Where appropriate, skin samples for fibroblast culture and swabs for microbiological culture are taken. Blood samples for grouping and other blood characteristics may be needed and a sufficient amount for DNA testing.
- Bite marks may need to be swabbed for saliva and casts made of tooth indentations. The cooperation of a forensic odontologist is then vital (see Chapter 26).
- A full internal dissection of all organs is required, with full sampling for histological purposes. Further radiographs may need to be taken on isolated bones or on the chest cage. The latter can be removed entirely by careful dissection if multiple rib fractures are present. X-rays can then be taken of the detached thoracic cage free of soft tissue and further histological blocks taken for decalcification in an attempt to date the lesions.
- If there is obvious or suspected brain damage, the brain should be removed carefully and suspended in formalin until firmly fixed. On no account should 'wet cutting' take place if intracranial damage or cerebral oedema is present.
- The globes of both eyes should be removed for histological examination: they can be taken entire and false eyes replaced for cosmetic purposes – or the posterior three-quarters of the globes can be removed via a superior approach through the base of the anterior fossa.

REFERENCES AND FURTHER READING

Ablin DS, Greenspan A, Reinhart M, *et al.* 1990. Differentiation of child abuse from osteogenesis imperfecta [see comments]. *AJR Am J Roentgenol* **154**:1035–46.

Adams PC, Strand RD, Bresnan MJ, *et al.* 1974. Kinky hair syndrome: serial study of radiological findings with emphasis on the similarity to the battered child syndrome. *Radiology* **112**:401–7.

Aoki N, Masuzawa H. 1984. Infantile acute subdural hematoma. Clinical analysis of 26 cases. *J Neurosurg* **61**:273–80.

Bacon CJ, Sayer GC, Howe JW. 1978. Extensive retinal haemorrhages in infancy – an innocent cause. *Br Med J* **1**:281.

Barlow B, Niemirska M, Gandhi RP, *et al.* 1983. Ten years of experience with falls from a height in children. *J Pediatr Surg* **18**:509–11.

Billmire ME, Myers PA. 1985. Serious head injury in infants: accident or abuse? *Pediatrics* **75**:340–2.

Brown RH. 1976. The battered child syndrome. *J Forensic Sci* **21**:65–70.

Bush CM, Jones JS, Cohle SD, *et al.* 1996. Pediatric injuries from cardiopulmonary resuscitation. *Ann Emerg Med* **28**:40–4.

Caffey J. 1946. Multiple fractures in the long bones of infants suffering from chronic subdural hematoma. *Am J Roentgenol* **56**:163–72.

Caffey J. 1972. On the theory and practice of shaking infants. Its potential residual effects of permanent brain damage and mental retardation. *Am J Dis Child* **124**:161–9.

Caffey J. 1974. The whiplash shaken infant syndrome: manual shaking by the extremities with whiplash-induced intracranial and intraocular bleedings, linked with residual permanent brain damage and mental retardation. *Pediatrics* **54**:396–403.

Cameron JM. 1970. The battered baby. *Br J Hosp Med* **4**:769–74.

Cameron JM, Rae LJ. 1975. *Atlas of the battered child syndrome.* Churchill Livingstone, Edinburgh.

Cameron JM, Johnson HR, Camps FE. 1966. The battered child syndrome. *Med Sci Law* **6**:2–21.

Carty H. 1988. Brittle or battered. *Arch Dis Child* **63**:350–2.

Case ME, Graham MA, Handy TC, *et al.* 2001. Position paper on fatal abusive head injuries in infants and young children. *Am J Forensic Med Pathol* **22**:112–22.

Chadwick DL, Chin S, Salerno C, *et al.* 1991. Deaths from falls in children: how far is fatal? *J Trauma* **31**:1353–5.

Chapman S. 1990. Radiological aspects of non-accidental injury. *J R Soc Med* **83**:67–71.

Cohle SD, Hawley DA, Berg KK, *et al.* 1995. Homicidal cardiac lacerations in children. *J Forensic Sci* **40**:212–18.

Cumming WA. 1979. Neonatal skeletal fractures. Birth trauma or child abuse? *J Can Assoc Radiol* **30**:30–3.

Department of Health and Social Security. 1988. *Diagnosis of child sexual abuse.* HMSO, London.

DiMaio D, DiMaio V. 1989. *Forensic pathology.* Elsevier, New York.

Duhaime AC, Gennarelli TA, Thibault LE, *et al.* 1987. The shaken baby syndrome. A clinical, pathological, and biomechanical study. *J Neurosurg* **66**:409–15.

Duhaime AC, Alario AJ, Lewander WJ, *et al.* 1992. Head injury in very young children: mechanisms, injury types, and ophthalmologic findings in 100 hospitalized patients younger than 2 years of age. *Pediatrics* **90**:179–85.

Eisenbrey AB. 1979. Retinal hemorrhage in the battered child. *Child's Brain* **5**:40–4.

Evans KT, Knight B. 1981. *Forensic radiology.* Blackwell Scientific Publications, Oxford.

Feldman KW, Brewer DK. 1984. Child abuse, cardiopulmonary resuscitation, and rib fractures. *Pediatrics* **73**:339–42.

Fontana VI. 1964. The neglect and abuse of children. *NY State J Med* **64**:215–20.

Friendly D. 1971. Ocular manifestations of physical child abuse. *Trans Am Acad Ophthalmol Otolaryntol* **75**:318–32.

Geddes JF. 1997. What's new in the diagnosis of head injury? *J Clin Pathol* **50**:271–4.

Geddes JF, Whitwell HL. 2001. Head injury in routine and forensic pathological practice. *Curr Top Pathol* **95**:101–24.

Geddes JF, Whitwell HL. 2004. New thoughts on inflicted head injury in infants. *Forensic Sci Int* accepted for publication.

Geddes JF, Whitwell HL, Graham DI. 2000a. Traumatic axonal injury: practical issues for diagnosis in medicolegal cases. *Neuropathol Appl Neurobiol* **26**:105–16.

Geddes JF, Whitwell HL, Graham DI. 2000b. Traumatic or diffuse axonal injury? Author's response. *Neuropathol Appl Neurobiol* **26**:491.

Geddes JF, Hackshaw AK, Vowles GH, *et al.* 2001a. Neuropathology of inflicted head injury in children. I. Patterns of brain damage. *Brain* **124**:1290–8.

Geddes JF, Vowles GH, Hackshaw AK, *et al.* 2001b. Neuropathology of inflicted head injury in children. II. Microscopic brain injury in infants. *Brain* **124**:1299–306.

Geddes JF, Tasker RC, Hackshaw AK, *et al.* 2003. Dural haemorrhage in non-traumatic infant deaths: does it explain the bleeding in 'shaken baby syndrome'? *Neuropathol Appl Neurobiol* **29**:14–22.

Gennarelli TA, Thibault LE. 1982. Biomechanics of acute subdural hematoma. *J Trauma* **22**:680–6.

Gilliland MG, Folberg R. 1992. Retinal hemorrhages: replicating the clinician's view of the eye. *Forensic Sci Int* **56**:77–80.

Gilliland MG, Luckenbach MW. 1993. Are retinal hemorrhages found after resuscitation attempts? A study of the eyes of 169 children. *Am J Forensic Med Pathol* **14**:187–92.

Gilliland MG, Luckenbach MW, Chenier TC. 1994. Systemic and ocular findings in 169 prospectively studied child deaths: retinal hemorrhages usually mean child abuse. *Forensic Sci Int* **68**:117–32.

Goetting MG, Sowa B. 1990. Retinal hemorrhage after cardiopulmonary resuscitation in children: an etiologic reevaluation [see comments]. *Pediatrics* **85**:585–8.

Goldstein B, Kelly MM, Bruton D, *et al.* 1993. Inflicted versus accidental head injury in critically injured children. *Crit Care Med* **21**:1328–32.

Gornall P, Ahmed S, Jolleys A, *et al.* 1972. Intra-abdominal injuries in the battered baby syndrome. *Arch Dis Child* **47**:211–4.

Grunebaum M, Horodniceanu C, Steinherz R. 1980. The radiographic manifestations of bone changes in copper deficiency. *Pediatr Radiol* **9**:101–4.

Hahn YS, Raimondi AJ, McLone DG, *et al.* 1983. Traumatic mechanisms of head injury in child abuse. *Child's Brain* **10**:229–41.

Hall C, Shaw D. 1991. Non-accidental injury. *Curr Imaging* **3**:88–93.

Hall JR, Reyes HM, Horvat M, *et al.* 1989. The mortality of childhood falls [see comments]. *J Trauma* **29**:1273–5.

Harcourt B, Hopkins D. 1971. Ophthalmic manifestations of the battered-baby syndrome. *Br Med J* **3**:398–401.

Harwood-Nash CE, Hendrick EB, Hudson AR. 1971. The significance of skull fractures in children. A study of 1,187 patients. *Radiology* **101**:151–6.

Helfer RE, Slovis TL, Black M. 1977. Injuries resulting when small children fall out of bed. *Pediatrics* **60**:533–5.

Hobbs CJ. 1984. Skull fracture and the diagnosis of abuse. *Arch Dis Child* **59**:246–52.

Hood I, Ryan D, Spitz WU. 1988. Resuscitation and petechiae. *Am J Forensic Med Pathol* **9**:35–7.

Howard MA, Bell BA, Uttley D. 1993. The pathophysiology of infant subdural haematomas. *Br J Neurosurg* 7:355–65.

Ingraham F, Heyl L. 1939. Subdural haematoma in infancy and childhood. *JAMA* **112**:198–204.

Ingram DM, Everett VD, Ingram DL. 2001. The relationship between the transverse hymenal orifice diameter by the separation technique and other possible markers of sexual abuse. *Child Abuse Negl* **25**:1109–20.

Kanter RK. 1986. Retinal hemorrhage after cardiopulmonary resuscitation or child abuse. *J Pediatr* **108**:430–2.

Kaur B, Taylor D. 1990. Retinal haemorrhages [see comments]. *Arch Dis Child* **65**:1369–72.

Kaur B, Taylor D. 1992. Fundus hemorrhages in infancy. *Surv Ophthalmol* **37**:1–17.

Kempe CH. 1971. Paediatric implications of the battered baby syndrome. *Arch Dis Child* **46**:28–37.

Kempe CH, Silverman FN, Steele BF, *et al.* 1963. Battered child syndrome. *JAMA* **181**:17–20.

Kiffney G. 1964. The eye of the battered child. *Arch Ophthalmol* **72**:231–3.

Kirschner RH, Stein RJ. 1985. The mistaken diagnosis of child abuse. A form of medical abuse? *Am J Dis Child* **139**:873–5.

Kleinman P. 1987. *Diagnostic imaging of child abuse.* Williams and Wilkins, Baltimore.

Kleinman PK, Marks SC, Blackbourne B. 1986. The metaphyseal lesion in abused infants: a radiologic-histopathologic study. *AJR Am J Roentgenol* **146**:895–905.

Kleinman PK, Marks SC, Adams VI, *et al.* 1988. Factors affecting visualization of posterior rib fractures in abused infants. *AJR Am J Roentgenol* **150**:635–8.

Knight B. 1986. The history of child abuse. *Forensic Sci Int* **30**:135–41.

Knight B. 1989. *The autopsy in child abuse.* In: Mason JK (ed.), *Paediatric forensic pathology*. Chapman and Hall, London.

Kobayashi M. 2001. Infant abuse in Osaka: health center activities from 1988 to 1999. *Pediatr Int* **43**:197–201.

Kogutt MS, Swischuk LE, Fagan CJ. 1974. Patterns of injury and significance of uncommon fractures in the battered child syndrome. *Am J Roentgenol Radium Ther Nucl Med* **121**:143–9.

Kravitz H, Driessen G, Gomberg R, *et al.* 1969. Accidental falls from elevated surfaces in infants from birth to one year of age. *Pediatrics* **44**(Suppl):869–76.

Lauer B, ten Broeck E, Grossman M. 1974. Battered child syndrome: review of 130 patients with controls. *Pediatrics* **54**:67–70.

Leventhal JM, Thomas SA, Rosenfield NS, *et al.* 1993. Fractures in young children. Distinguishing child abuse from unintentional injuries. *Am J Dis Child* **147**:87–92.

Mann K, Chan K, Yue C. 1986. Skull fractures in children: their assessment in relation to developmental skull changes and acute intracranial haematomas. *Child's Nerv Syst* **2**:258–61.

Mason JE. 1990. *Paediatric forensic medicine and pathology.* Chapman and Hall, London.

Maxeiner H. 2001. Demonstration and interpretation of bridging vein ruptures in cases of infantile subdural bleedings. *J Forensic Sci* **46**:85–93.

McClelland CQ, Rekate H, Kaufman B, *et al.* 1980. Cerebral injury in child abuse: a changing profile. *Child's Brain* **7**:225–35.

McCort J, Vaudagna J. 1964. Visceral injuries in battered children. *Radiology* **124**:681–4.

Milroy CM. 1999. Munchausen syndrome by proxy and intra-alveolar haemosiderin. *Int J Legal Med* **112**:309–12.

Morild I, Gjerdet NR, Giertsen JC. 1993. Bone strength in infants. *Forensic Sci Int* **60**:111–19.

Mushin AS. 1971. Ocular damage in the battered-baby syndrome. *Br Med J* **3**:402–4.

Nimityongskul P, Anderson LD. 1987. The likelihood of injuries when children fall out of bed. *J Pediatr Orthop* **7**:184–6.

Nokes L, Roberts A, Knight B. 1995. The use of the gadd severity index in forensic medicine: a case study. *Forensic Sci Int* **76**:85–90.

Norman MG, Smialek JE, Newman DE, *et al.* 1984. The postmortem examination on the abused child.

Pathological, radiographic, and legal aspects. *Perspect Pediatr Pathol* **8**:313–43.

Norton L. 1983. Child abuse. *Clin Lab Med* **3**:321–42.

O'Connor J, Cohen J. 1987. Dating of fractures. In: Kleinman P (ed.), *Diagnostic imaging of child abuse.* Williams and Wilkins, Baltimore.

Ogata M, Tsuganezawa O. 1995. An isolated perforation of the jejunum caused by child abuse. A case report. *Am J Forensic Med Pathol* **16**:17–20.

Ojima K, Matsumoto H, Hayase T, *et al.* 1994. An autopsy case of osteogenesis imperfecta initially suspected as child abuse. *Forensic Sci Int* **65**:97–104.

Parsons MA, Start RD. 2001. Acp best practice no. 164: necropsy techniques in ophthalmic pathology. *J Clin Pathol* **54**:417–27.

Pearn J. 1990. Physical abuse of children. In: Mason JK (ed.), *Paediatric forensic medicine and pathology.* Chapman and Hall, London, pp. 204–20.

Pena SD, Medovy H. 1973. Child abuse and traumatic pseudocyst of the pancreas. *J Pediatr* **83**:1026–8.

Pollanen MS, Smith CR, Chiasson DA, *et al.* 2002. Fatal child abuse – maltreatment syndrome. A retrospective study in Ontario, Canada, 1990–1995. *Forensic Sci Int* **126**:101–4.

Polson CJ, Gee D, Knight B. 1985. *Essentials of forensic medicine*, 4th edn. Pergamon Press, Oxford.

Rao N, Smith RE, Choi JH, *et al.* 1988. Autopsy findings in the eyes of fourteen fatally abused children. *Forensic Sci Int* **39**:293–9.

Reiber GD. 1993. Fatal falls in childhood. How far must children fall to sustain fatal head injury? Report of cases and review of the literature. *Am J Forensic Med Pathol* **14**:201–7.

Reichelderfer TE, Overbach A, Greensher J. 1979. Unsafe playgrounds. *Pediatrics* **64**:962–3.

Rekate HL. 1985. Subdural hematomas in infants [letter]. *J Neurosurg* **62**:316–17.

Rivara FP, Kamitsuka MD, Quan L. 1988. Injuries to children younger than 1 year of age. *Pediatrics* **81**:93–7.

Root I. 1992. Head injuries from short distance falls [see comments]. *Am J Forensic Med Pathol* **13**:85–7.

Sezen F. 1971. Retinal haemorrhages in newborn infants. *Br J Ophthalmol* **55**:248–53.

Shannon P, Smith CR, Deck J, *et al.* 1998. Axonal injury and the neuropathology of shaken baby syndrome. *Acta Neuropathol (Berl)* **95**:625–31.

Silverman FN. 1972. Unrecognized trauma in infants, the battered child syndrome, and the syndrome of Ambroise Tardieu. Rigler lecture. *Radiology* **104**:337–53.

Smith FW, Gilday DL, Ash JM, *et al.* 1980. Unsuspected costo-vertebral fractures demonstrated by bone scanning in the child abuse syndrome. *Pediatr Radiol* **10**:103–6.

Spevak MR, Kleinman PK, Belanger PL, *et al.* 1994. Cardiopulmonary resuscitation and rib fractures in infants. A postmortem radiologic-pathologic study. *JAMA* **272**:617–18.

Tardieu A. 1860. Etude medico-legale sur les services et mauvais traiments exerces sur les enfants. *Ann Hyg Pub Med Leg* **13**:361–98.

Thomas PS. 1977. Rib fractures in infancy. *Ann Radiol Paris* **20**:115–22.

Thomas SA, Rosenfield NS, Leventhal JM, *et al.* 1991. Long-bone fractures in young children: distinguishing accidental injuries from child abuse. *Pediatrics* **88**:471–6.

Tomasi LG, Rosman NP. 1975. Purtscher retinopathy in the battered child syndrome. *Am J Dis Child* **129**:1335–7.

Touloukian RJ. 1968. Abdominal visceral injuries in battered children. *Pediatrics* **42**:642–6.

Vowles GH, Scholtz CL, Cameron JM. 1987. Diffuse axonal injury in early infancy. *J Clin Pathol* **40**:185–9.

Wagner GN. 1986. Crime scene investigation in child-abuse cases. *Am J Forensic Med Pathol* **7**:94–9.

Weber W. 1984. [Experimental studies of skull fractures in infants.] *Z Rechtsmed* **92**:87–94.

Weber W. 1985. [Biomechanical fragility of the infant skull.] *Z Rechtsmed* **94**:93–101.

Weedn VW, Mansour AM, Nichols MM. 1990. Retinal hemorrhage in an infant after cardiopulmonary resuscitation. *Am J Forensic Med Pathol* **11**:79–82.

Williams RA. 1991. Injuries in infants and small children resulting from witnessed and corroborated free falls. *J Trauma* **31**:1350–2.

Wilson E. 1977. Estimation of the age of cutaneous contusions in child abuse. *Pediatrics* **60**:750–5.

Wooley PV, Evand WA. 1955. Significance of skeletal lesions in infants resembling those of traumatic origin. *JAMA* **158**:539–43.

Worlock P, Stower M, Barbor P. 1986. Patterns of fractures in accidental and non-accidental injury in children: a comparative study. *Br Med J Clin Res Ed* **293**:100–2.

Yukawa N, Carter N, Rutty G, *et al.* 1999. Intra-alveolar haemorrhage in sudden infant death syndrome: a cause for concern? *J Clin Pathol* **52**:581–7.

CHAPTER 23

Deaths associated with surgical procedures

- The autopsy on deaths associated with surgery and anaesthesia 480
- The mode and cause of death after surgery and anaesthesia 482
- Deaths associated with anaesthesia 484
- Malignant hyperthermia 485
- Autopsy procedure 486
- References and further reading 486

In most jurisdictions, deaths that occur during or within a short time after surgical operation, invasive diagnostic procedure or an anaesthetic, become the subject of a medicolegal investigation. Similarly, any death thought to be caused or contributed to, by any of these procedures – irrespective of the interval – may be enquired into if the medical attendants or the relatives consider that a causal relationship exists. To be effective such investigations must include an autopsy, and the pathologist involved in these has a difficult and often professionally sensitive task.

The pathologist should, wherever possible, be independent of the institution in which either the death occurred or which carried out the procedure. This is partly why the services of a forensic pathologist should be retained by the investigating authority, rather than a clinical or histopathologist belonging to the hospital, or associated institution. It is important that a separation should exist – and be seen to exist – between the loyalties of these two elements, not because there is any substantial likelihood of collusion, but because the public and private interest is best served by displaying the independence of the pathologist. In addition, there may be professional embarrassment when a close colleague of a surgeon or anaesthetist is placed in a potentially critical position within the same institution.

It is also sometimes useful to have the technical advice and expert opinion of an independent clinical consultant who is unconnected with the hospital, and who is not a colleague of the team involved in incident.

THE AUTOPSY ON DEATHS ASSOCIATED WITH SURGERY AND ANAESTHESIA

A number of difficulties exist, that are peculiar to this type of death. The following matters need particular attention:

- The morphological findings, especially in so-called ‘anaesthetic deaths’, may be minimal or even absent so, more than in any other type of case, expert advice and full clinical information are essential.
- Technically the autopsy on a post-operative death may be difficult as a result of the surgical intervention and its sequelae, especially in abdominal and thoracic procedures. Exudate, sepsis, adhesions, haemorrhage, oedema and distortion of the normal anatomy may make the dissection difficult, especially if the details of the surgical procedure are not fully known. Post-mortem changes can further complicate the appearances: for example, recent suture lines in the intestine or stomach may appear to be leaking, but this may be caused by autolysis, and even handling the tissues at autopsy may tear devitalized and autolytic structures even further.
- Numerous surgical and anaesthetic devices may have been introduced into the patient during the procedure, such as airways, endotracheal tubes, indwelling needles, intravascular cannulae, self-retaining catheters, wound

drains, chest tubes, monitoring electrodes, and metal or plastic prostheses. It is essential that none of these be removed before autopsy, as their proper placement and patency may need to be checked. Standing orders should be issued to hospitals within the jurisdiction of any coroner or medical examiner to the effect that no post-mortem interference with the body be made by members of the medical, nursing, technical or portering staff, as it has happened in the past that the single vital piece of evidence about some technical mishap has been discarded by some well-meaning person.

- Before any further disturbance of the body is made (for instance, by turning it over to look at the back) and certainly before a knife is picked up, the position of any endotracheal tube that may remain after anaesthesia must be checked. If the clinical information raises any doubt whatsoever about its malposition, then a pre-autopsy radiograph should be obtained, both anteroposterior and lateral.

  Oesophageal intubation is not all that uncommon in anaesthetic mishaps and should be investigated carefully. The tube may be palpable through the skin if properly in the trachea – if there is doubt, a small primary incision in the midline of the neck with a further careful cut into the trachea can resolve the matter. Sometimes the intubation has been into the oesophagus, then been rectified after the damage has been done. Here the pathologist may find a ring of oedematous oesophageal mucosa still present at a level equal to that of the tube in the trachea. When nitrous oxide or other anaesthetic gas has previously been passed down a tube in the oesophagus there may be distension of the stomach and intestines. A sample of the contained gas may reveal the agent on analysis.
- The hospital laboratory should be requested to retain any ante-mortem blood or body-fluid samples sent to them so that they remain available for analytical checks, such as blood grouping in transfusion mishaps, or creatine phosphokinase activity in malignant hyperthermia.
- Even more than usual, the fullest information is needed before beginning the autopsy. The patient's notes are essential, together with any other relevant information. Sometimes the nursing records may be more helpful

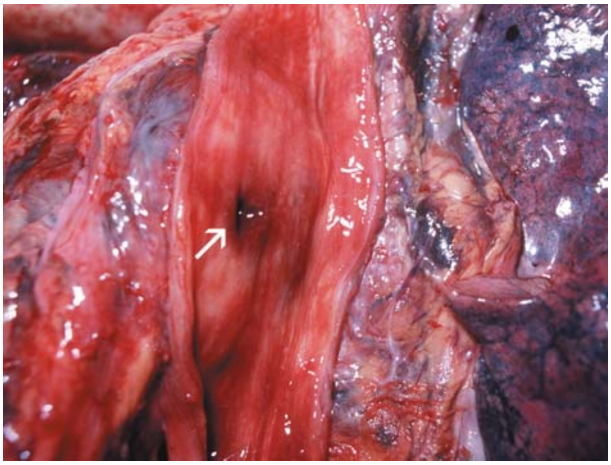

FIGURE 23.1 *Aorto-oesophageal fistula (arrow) due to ruptured anastomosis causing fatal haematemesis 18 days after replacement of the thoracic aorta with a prosthesis.*

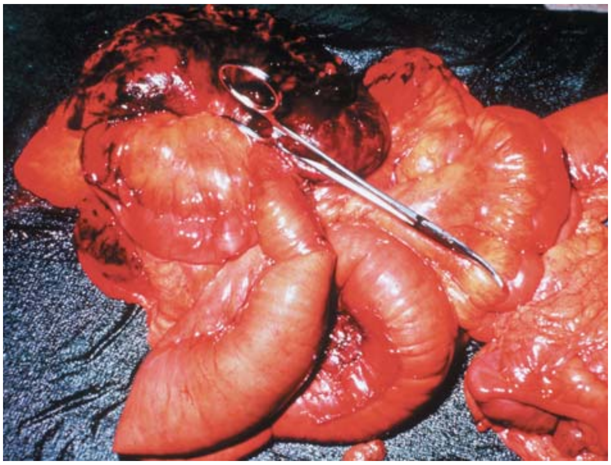

FIGURE 23.2 *Gut strangulated through the handle of forceps left in the abdomen during surgery and detected during an unrelated autopsy six weeks later for pneumoconiosis assessment.*

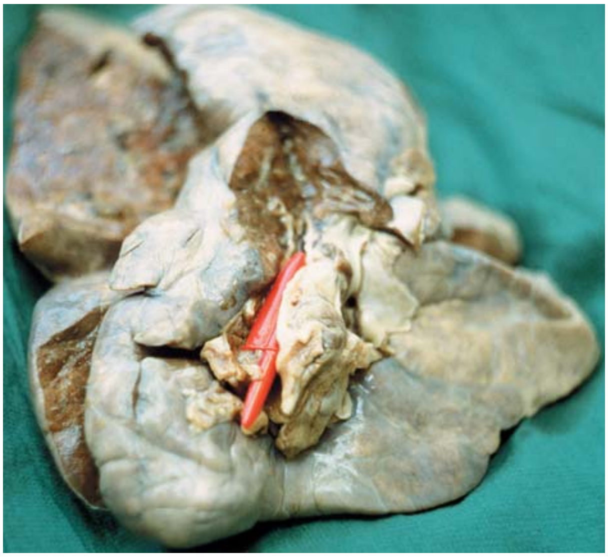

FIGURE 23.3 *Bronchial obstruction by the cover of a ballpoint pen, used by a midwife to hold down the tongue of a pregnant woman, who died days after childbirth from pneumonia.*

than the medical notes, as they are frequently more detailed and recorded at shorter intervals.

Equally important is the attendance of clinicians at the autopsy. Indeed, in deaths associated with anaesthesia, this is usually more useful than the post-mortem dissection, which frequently reveals little or no morphological evidence. Sadly, attendance at autopsies has declined markedly in recent years, as have requests for clinical autopsies, but every effort should be made to secure the presence of someone who knows the circumstances of the case well and who can point out to the pathologist the procedures that were carried out. A discussion across the autopsy table is of paramount importance and it is of little use if a junior doctor is delegated to attend, especially if he confesses to knowing little or nothing about the patient.

Especially in anaesthetic-related deaths, this dialogue with a clinician – in this case, the anaesthetist – may provide virtually all the data upon which to base a cause of death. Discussion between pathologist, surgeon and anaesthetist may arrive at an amicable conclusion that will be the best consensus of opinion to offer the investigating authority in cases where a bare autopsy might reveal little or nothing upon which to base any interpretation of the fatal processes.

- During the autopsy, care must be taken to detect any surgical emphysema, pneumothorax or air embolism. When surgical operations have been conducted with the patient in the sitting position, as in posterior fossa neurosurgery or some thyroidectomies, the possibility of air aspiration into open veins or venous sinuses always exists.

  Where infusion or transfusion mishaps may be a possibility, especially where gas pressure has been used to hasten the infusion rate, the possibility of air embolism must always be borne in mind.

THE MODE AND CAUSE OF DEATH AFTER SURGERY AND ANAESTHESIA

The rather loose terms ‘operative deaths’ and ‘anaesthetic deaths’ are usually inaccurate descriptions. When used without due consideration, they may lead to medico-legal enquiries and even civil litigation, which is not justified by the actual circumstances.

Fatalities associated with surgical intervention or invasive diagnostic procedures can be separated into several categories:

- those directly caused by the disease or injury for which the operation or anaesthetic was being carried out
- those caused by a disease or abnormality other than that for which the procedure was being carried out
- those resulting from a mishap during, or a complication of, the surgical or diagnostic procedure
- those resulting from a mishap during, or a complication of, the anaesthetic being administered.

These groups must be considered in more detail.

Deaths directly caused by disease or injury for which surgery is performed

Many deaths during a surgical or diagnostic procedure or an anaesthetic are caused by the disease process or injury for which the procedure was being performed. The death may be virtually inevitable in some ‘heroic’ operations where urgent intervention is the only hope of saving life. Presumably there must have been some chance of success or even palliation, otherwise the intervention would not have been justified, but the decision not to operate may be difficult. When death occurs from the effects of such disease or injury, some medico-legal systems will certify the original condition as the cause of death and omit any reference to the intervention.

The American Society of Anesthesiologists have a classification system for deaths during surgical procedures, as follows:

- ASA 1: those with no serious disease
- ASA 2: those who may have a serious disease, but have no limitations on their activities
- ASA 3: those with serious disease and some limitation of their activities
- ASA 4: those with a serious disease that is a constant threat to their health and that limits their activities
- ASA 5: those with such serious disease that their death can be expected within 24 hours, with or without treatment.

The Society recommended that classes 1–3 require full investigation, as they were expected to survive; those in class 4 that were elective and not emergency procedures, with no expectation of death, are also included. The general principle is that unexpected deaths are investigated, though in the British coroner and Procurator Fiscal system there are no exceptions to such investigations, once reported by doctors to the appropriate legal authority.

Pathologists in some jurisdictions have problems in the nomenclature of certification following an autopsy. The strict test to be applied should be: ‘Would death have occurred when it did, if the operation had not taken place?’ Sometimes this is impossible to answer and the balance of probabilities may be evenly poised. If a patient has a leaking aortic aneurysm, which catastrophically ruptures whilst the surgeon is still making his skin incision, most would agree that the original disease process was the primary cause of death. Even here, however, it may be argued that the added disturbance of preparation and anaesthesia contributed something to a possible slight rise in blood pressure that further split the aneurysm wall.

Where a person dies of a pulmonary embolus 7 days after an elective gastrectomy for peptic ulcer, the connection between the operation and the death can hardly be denied, even though in up to a fifth of fatal pulmonary embolisms, depending on the study, no predisposing factors have been found (see Chapter 13).

When both trauma and surgical operation have taken place, then it can be impossible to separate the relative contributions of each to a death. A common example would be death a few days after the pinning of a fractured neck or femur in an old person. A fatal bronchopneumonia, pulmonary embolism or acute cardiac failure might all be caused either by the original trauma, by the stress of operation or by the subsequent immobilization in bed.

Death due to a disease or disability other than that for which surgery is performed

Where death is due to a disease or disability other than that for which the operation was performed, a distinction has to be drawn between those conditions that were known before the operation and those which were unsuspected.

When some disease is known beforehand, it must be evaluated in deciding whether or not the operative procedure is justified. For example, a patient with a non-obstructing hernia who has chronic obstructive airways disease might be thought to be too poor a risk for an elective, prophylactic operation. If that same patient's hernia becomes strangulated and he develops potentially fatal intestinal obstruction, however, the risk:benefit ratio dramatically changes and the danger of a perioperative death may have to be accepted. Naturally, operative and anaesthetic techniques may have to be modified to take account of the known adverse conditions.

It is different when an occult disease process is not suspected by the clinicians. It may be that they had no diagnostic means of appreciating the risk – such as an unsuspected phaeochromocytoma – but a medico-legal problem can arise if they took no reasonable precautions in detecting common risk factors, such as hypertension, lung disease or ischaemic heart disease, which in retrospect could have been diagnosed.

Death as a result of failure of surgical technique

Exceptionally, death is the result of a failure of surgical technique. This may be inadvertent, from a true 'accident', sometimes caused by unusually difficult operative circumstances, to anatomical abnormalities or even failure of equipment. When it is the result of error or incompetence, then a legal action for negligence may ensue and the pathologist must be even more meticulous than usual in producing a detailed, objective and impartial report. In all these cases, the presence of the surgeon is essential, so that agreement can be reached on what actually was found at autopsy.

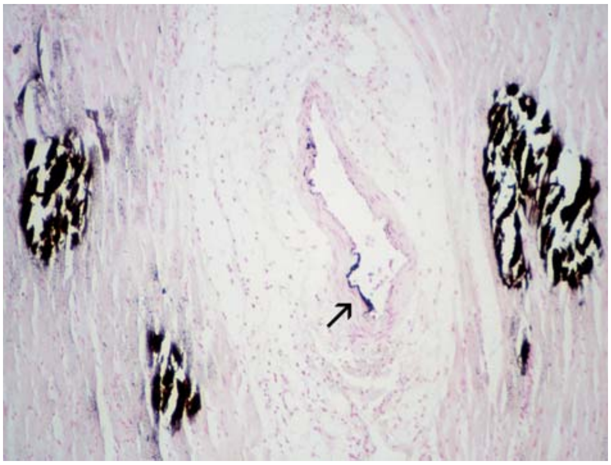

FIGURE 23.4 *Calcification of myocardium and vascular endothelium (arrow) due to postoperative rhabdomyolytic myoglobinuria and acute renal failure after reconstructive hand surgery and the interruption of blood circulation to the corresponding arm for the time of a nerve reconstruction. (Von Kossa staining, original magnification ×100.)*

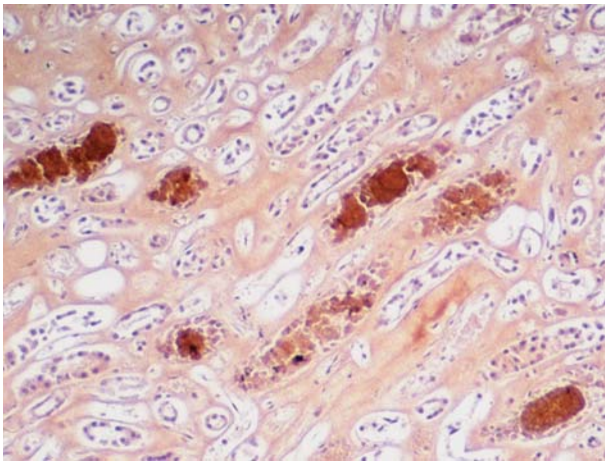

FIGURE 23.5 *Immunohistochemical demonstration of myoglobin in the renal tubuli with Anti-Human Myoglobin (Dakopatts a/s, Denmark).*

When a failure of equipment may be responsible, then expert examination and advice is essential. Anaesthetic machines, gas supply, compatibility of connections and all the sophisticated hardware of operating theatres must be subjected to the most rigorous inspection if a malfunction is suspected. This is no part of the pathologist's concern but, as he is responsible for the eventual decision about the cause of death, any conclusions about an equipment failure must be communicated to him.

Deaths due to anaesthetic administration

The foregoing categories account for most perioperative deaths and fatalities following surgery, even if delayed. True 'anaesthetic deaths' are comparatively rare, but have such potential medicolegal importance that they must be discussed in more detail.

DEATHS ASSOCIATED WITH ANAESTHESIA

Even more than is the case with surgeons, any autopsy on an anaesthetic-related death must be a cooperative process with the anaesthetist. Not only are the objective autopsy findings in true anaesthetic deaths meagre or absent, but the pathologist's training and experience of the complex techniques of modern anaesthesia is insufficient for him to appreciate, analyse and criticise constructively without the expert knowledge of the anaesthetist.

Most anaesthetic-related deaths are not caused by the anaesthetic agent itself, but due to other aspects of the procedure. Patients have anaesthetics because they have some significant – often very serious – disease or injury and, as has already been discussed, it is more likely that the fatality was the result of these or of some co-existent condition rather than the anaesthetic. There is a tendency on the part of relatives – and sometimes their lawyers – to assume that because someone died 'under' an anaesthetic, then this must be the major or even sole factor in the death because it was so closely related in time.

When factors other than anaesthesia can be excluded, certain features can be identified. A survey was conducted by Lunn and Mushin (1982) for the Association of Anaesthetists. This indicated that, although 1 in every 166 patients died within 6 days of a surgical operation, only 1 in 10 000 die solely as a consequence of the anaesthetic. Anaesthesia contributed to (but not totally caused) the death of 1 in every 1700 patients, many of these deaths being potentially avoidable.

The survey also showed that the causes have not changed significantly during the last 30 years, in spite of marked changes in anaesthetic technology and that anaesthesia may contribute to deaths that occur more than 24 hours after the administration. Many patients suffer from intercurrent disease unrelated to their surgical condition, but the implications for the anaesthetist are all too often ignored. The major causes of true anaesthetic fatalities include inexperienced and insufficiently supervised junior staff. Allowing periods of hypoxia to occur appears to be one of the common faults leading to a fatal outcome.

As far as the pathologist is concerned, the survey complained that the autopsy is of limited value in investigating the death. All too often irrelevant findings are offered, such as 'bronchitis'. Ischaemic heart disease was recorded in 39 per cent of the series. It was emphasized that the investigation of such deaths cannot be the sole province of the pathologist, but must rely heavily upon clinical data and cooperation with other doctors. Earlier investigations indicated that acute cardiovascular failure was the most common cause of unexpected death during anaesthesia. Lahey and Ruzicka in 1950 measured the risk of death during anaesthesia as about 1/1000, from all causes. Cardiac arrest was the most common single mode of death, being seen on average once or twice a year in most busy operating suites even under the most careful surgical and anaesthetic regimens. The autopsy is usually unhelpful in determining the cause, if coronary insufficiency can be excluded. Most cardiac arrest occurs under relative light anaesthesia and thus tends to occur at either the start or finish of the surgical procedure.

Reid and Brace (1940) showed that the arrest is mainly neurogenic and that any irritation of the respiratory tract, such as laryngoscopy or intubation, may cause a lightly anaesthetized patient to have a cardiac arrest. Inexperienced anaesthetists may not keep the patient at a level of unconsciousness appropriate to the manipulations of the surgeon, as it has been shown that such procedures can also precipitate arrest during light anaesthesia.

Hypoxia is also a potent precipitating factor in cardiac arrest. Deaths from respiratory failure also occur, and hypoxia is again a prime example – either from faults in the apparatus or more commonly to inexperience of the anaesthetist, especially in handling equipment with which he is not familiar. An overdose of the anaesthetic agent may depress the respiratory centres and begin a descending spiral of hypoxia. Excessive premedication and the use of muscle paralysing agents may also predispose towards respiratory failure unless appropriate measures are taken, such as assisted respiration.

Airway obstruction is another danger – from blood, teeth, dentures, faults in the connecting tubing, laryngeal spasm, swabs and an abnormal posture of the neck. Though regurgitation of gastric contents is a real danger, the use of

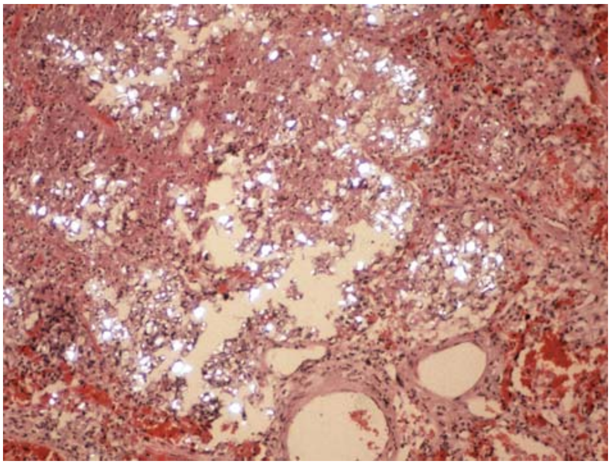

FIGURE 23.6 *Birefringent material in the alveoli due to contrast medium aspiration (barium sulphate).*

cuffed endotracheal tubes has greatly reduced the hazard. From the pathologist's point of view, the finding of gastric contents in the airways must be backed up by some clinical evidence of its ante-mortem origins, as a quarter of bodies have some gastric contents in some part of the air passages at autopsy, mostly as a result of agonal or even post-mortem transfer that is a consequence of the dying process, not a cause (Chapter 14).

Rarely, some physical fault in the anaesthetic equipment may cause death. Mention has been made of faults in the connecting tubing, which may be internal delamination, not visible from outside. Flow-meter errors occur, but a more common one is the confusion of bottled or piped gas supplies, and the inadvertent use of the wrong substance or the connection of an empty cylinder. Two such fatalities, where the wrong gas was introduced into the theatre piped supply, occurred in Hong Kong in recent years.

Fires are not uncommon, the most frequent danger being the ignition of spirit-based skin antiseptics ignited by cautery. Any electrical apparatus is potentially dangerous, and defective cauteries, defibrillators and diathermy equipment have all caused death.

Explosions from inflammable gases and vapours, such as cyclopropane and ether, have been catastrophic on occasions.

Hazards of local and epidural anaesthesia

Local anaesthetics rarely cause death, though the overall rate of complications is about 1/2000. Hypersensitivity and escape of contained adrenergic drugs form the major fatal hazards. It is said that most cases of 'hypersensitivity' are really overdosage, but there are undoubtedly some patients who have an abnormal sensitivity to the cocaine-like active constituents. Legal actions have been brought against doctors for failure to give a test skin dose before injecting the major quantity. Diffusion away from the operative site of adrenaline-like substances used to vasoconstrict the area and thus retain the anaesthetic effect may cause a sudden cardiac arrest.

Epidural and spinal anaesthesia has returned to favour, especially in obstetrics, after a period in which spinal injection of anaesthetic agents fell into disrepute because of a number of lawsuits brought because of spinal cord damage as a result of contaminants or excessive dosage.

The upwards diffusion of the active agents may cause paralysis, and the effect upon the vasomotor system may lead to a precipitous fall in blood pressure. Total spinal anaesthesia from upward diffusion blocks all sympathetic efferent impulses so, if the concentration of the anaesthetic agent in the upper thoracic and cervical spinal fluid level is high enough, efferent transmission to all motor fibres of the respiratory muscles will be blocked.

Epidural anaesthesia has now replaced intrathecal administration, but even this has occasional tragedies, and is said to have a 1/10 000 incidence of permanent nerve paralysis. This usually comes about either from an overdose or from puncture of the theca, instead of the anaesthetic solution remaining in the epidural space. In the rare autopsies on such deaths, a sample of cerebrospinal fluid should be taken for analysis to ascertain the concentration of anaesthetic substance. An attempt should be made to find the puncture in the dura, which is easier said than done. If an epidural catheter is still in place for continuous administration of the active agent, dye or Indian ink can be injected to see if it enters the subdural space.

MALIGNANT HYPERTHERMIA

This is a familial condition in which certain agents, including some anaesthetics and muscle relaxants, precipitate a metabolic change in skeletal muscle with the production of energy, and hence a rapid and sometimes fatal rise in temperature. The basic condition is an autosomal dominantly inherited trait associated with sudden uncoupling of oxidative phosphorylation, accompanied by massive energy production from the muscles with consequent overheating.

When malignant hyperthermia is diagnosed, the family members should be investigated, preferably by muscle biopsy, to assess the risk of their suffering a similar crisis if subjected to anaesthesia. In 70 per cent of carriers, the creatine phosphokinase and aldolase activities are elevated, though the basic abnormality appears to be a defect in calcium transfer and phosphorylation within the muscle fibres. This is one reason why hospital laboratories should be asked to retain any pre-mortem blood samples, as the creatine phosphokinase activity can be estimated retrospectively.

Suxamethonium and halogenated anaesthetic agents seem particularly prone to precipitate the condition. Halothane is one of the agents associated with malignant hyperthermia, either alone or in combination with suxamethonium. The condition occurs about once in every 10 000 anaesthetics, with various estimates ranging from 1/5000 to 1/70 000. There are virtually no autopsy features of any significance and the diagnosis is made on clinical grounds. The features of malignant hyperthermia include a sudden rise in temperature, sometimes reaching 43°C, with stiffening of the skeletal muscles, rapid heart rate and respiratory distress.

Halothane anaesthesia has acquired the reputation for causing liver injury known as halothane hepatitis and, rarely, a fulminant hepatic failure. Its incidence varies between 1/35 000 and 1/600 000 and if halothane anaesthesia is repeated within one month, 1/6000–1:22 000. It is a typical example of immune-mediated adverse drug reactions. Other drugs used in anaesthesia are occasionally responsible for death. Barbiturates used in induction, such as thiopentone, may give rise to cardiorespiratory failure, often because the quantity used has been excessive. Trichlorethylene and atropine have been involved in fatalities, the usual mode being sudden circulatory failure.

AUTOPSY PROCEDURE

In most autopsies on deaths during anaesthesia, toxicological investigations are generally unrewarding, except where overdose with specific drugs such as barbiturate or adrenaline is involved. It is usually impossible to make any quantitative assessment of the amount of volatile agents present, such as nitrous oxide, halothane or cyclopropane, even though some American writers advocate clamping a main bronchus and retaining a lung in a plastic – or preferably nylon – bag, so that head-space gas can be analysed. Neither is it practical to try to prove hypoxia by post-mortem measurement of blood gases. Usually the function of the autopsy is to discover or exclude natural disease, and mechanical blockages, as the pharmacological aspects are usually beyond investigation. Even when it seems obvious that a surgical error has taken place, one should be careful and not jump to conclusions nor let this distract from a thorough investigation. A full range of specimens for histological examination should be taken to exclude occult conditions such as a lymphoma or myocarditis, as well as to investigate the severity of the disease for which the surgical procedure was being performed. This includes also the investigation of any surgical specimens taken during the operation and sent to the histopathologist. Histological examination of the brain (fixed before cutting wherever possible) must be carried out. Though most mishaps will have occurred too quickly for much hope of any stigmata of cerebral hypoxia to be seen, changes can classically occur in Sommer's area of the hippocampal gyrus. Perhaps an even better marker is the distal parts of the cerebellar folia, where pallor and rapid loss of Purkinje cells can be seen within a few hours.

The work of Plum (Plum and Posner 1984) is relevant in this respect, as he found morphological changes in the brains of victims of only slight hypoxic periods who survived for long periods after an anaesthetic. At autopsy he recorded diffuse, severe leucoencephalopathy of the cerebral hemispheres, with sparing of the immediate subcortical connecting fibres and – usually – of the brainstem. Demyelination and obliteration of axons was observed and sometimes infarction of the basal ganglia. The cortical neurones were spared, however, and the damage appeared confined to the white matter. Plum attributed this damage to greater glycolysis in the white matter during hypoxia, compared with the grey.

Some new immunohistochemical techniques have been used in animal experiment to demonstrate structures of the cytoskeleton, e.g. in detecting ischaemic/hypoxic damage. Great caution must be exercised, however, when using these methods in autopsy material and it is therefore crucial to examine alterations occurring in the cytoskeleton as a result of post-mortem delay before conclusions can be drawn from the possible changes observed. Irving *et al.* (1997) investigated the post-mortem distribution of microtubule-associated proteins (MAP) in various regions of rat brain using immunohistochemistry. They showed that each MAP underwent unique changes that were dependent both on post-mortem interval and the brain region examined. Following long post-mortem delays, some of the changes in these proteins were similar to those seen in rodent models of cerebral ischaemia.

REFERENCES AND FURTHER READING

Adelstein A, Loy P. 1979. Fatal adverse effects of medicines and surgery. *Popul Trends* **17**:17–21.

Beecher HK, Todd DP. 1954. A study of deaths associated with anesthesia and surgery: based on a study of 599 548 anesthesias in ten institutions 1948–1952. *Ann Surg* **140**:2–12.

Briely JB. 1963. Cerebral injury following cardiac operations. *Thorax* **19**:291–6.

Britt BA, Locher WG, Kalow W. 1969. Hereditary aspects of malignant hyperthermia. *Can Anaesth Soc J* **16**:89–98.

Campling EA, Devlin HB, Lum JN. 1990. Gynaecology and the national confidential enquiry into perioperative deaths. *Br J Obstet Gynaecol* **97**:466–7.

Campling EA, Devlin HB, Lunn JN. 1989, 1990, 1992. *The report of the National Confidential Enquiry into*

Perioperative Deaths, *1989, 1990, 1992.* Royal College of Surgeons, London.

Campling EA, Devlin HB, Lunn JN. 1992. Reporting to NCEPOD. *Br Med J* **305**:252.

Caplan RA, Ward RJ, Posner K, *et al.* 1988. Unexpected cardiac arrest during spinal anesthesia: a closed claims analysis of predisposing factors [see comments]. *Anesthesiology* **68**:5–11.

Carr NJ, Burke MM, Corbishley CM, *et al.* 2002. The autopsy: lessons from the national confidential enquiry into perioperative deaths. *J R Soc Med* **95**:328–30.

Corby C, Camps FE. 1960. Therapeutic accidents during the administration of barium enemas. *J Forensic Med* **7**:206–20.

Denborough MA, Lovell RR. 1960. Anaesthetic deaths in a family. *Lancet* **2**:45–6.

Dinnick OP. 1964. Deaths associated with anaesthesia: a report on 600 cases. *Anaesthesia* **19**:536–44.

Edelstein JM. 1988. Sudden death following administration of radio contrast media. *J Forensic Sci* **33**:734–7.

Edwards G, Morton HJ, Pask EA, *et al.* 1957. Deaths associated with anaesthesia: a report on 1000 cases. *Anaesthesia* **11**:194–202.

Ermenc B. 2000. Comparison of the clinical and post mortem diagnoses of the causes of death. *Forensic Sci Int* **114**:117–19.

Falk E, Simonsen J. 1979. The histology of myocardium in malignant hyperthermia: a preliminary report of 11 cases. *Forensic Sci Int* **13**:211–20.

Hunt AC. 1958. Anaesthetic causes of death on the table. *J Forensic Med* **5**:141–5.

Irving EA, McCulloch J, Dewar D. 1997. The effect of postmortem delay on the distribution of microtubule-associated proteins tau, map2, and map5 in the rat. *Mol Chem Neuropathol* **30**:253–71.

Isaacs H, Barlow MB. 1970. The genetic background to malignant hyperpyrexia revealed by serum creatine phosphokinase estimations in asymptomatic relatives. *Br J Anaesth* **42**:1077–84.

Keats AS. 1988. Anesthesia mortality – a new mechanism [editorial]. *Anesthesiology* **68**:2–4.

Lau G. 1996. Perioperative deaths: a comparative study of coroner's autopsies between the periods of 1989–1991 and 1992–1994. *Ann Acad Med Singapore* **25**:509–15.

Lau G. 2000. Perioperative deaths: a further comparative review of coroner's autopsies with particular reference to the occurrence of fatal iatrogenic injury. *Ann Acad Med Singapore* **29**:486–97.

Leading article. 1964. Halothane re-examined. *Br Med J* **3**:325.

Loghmanee F, Tobak M. 1986. Fatal malignant hyperthermia associated with recreational cocaine and ethanol abuse. *Am J Forensic Med Pathol* **7**:246–8.

Lunn J, Mushin W. 1982. *Mortality associated with anaesthesia.* Nuffield Provincial Hospital Trust, London.

Lunn JN. 1994. The national confidential enquiry into perioperative deaths. *J Clin Monit* **10**:426–8.

Lunn JN, Devlin HB, Hoile RW, *et al.* 1993. Editorial. II – The National Confidential Enquiry into Perioperative Deaths. *Br J Anaesth* **70**:382.

Macrae WA, Miller KM, Watson AA. 1979. Malignant hyperpyrexia. *Med Sci Law* **19**:261–4.

Mazzia VD, Simon A. 1979. Another look at malignant hyperthermia. In: Wecht C (ed.), *Legal medicine annual 1978*. Appleton Century Crofts, New York.

National Confidential Enquiry into Early Post-operative Deaths (NCEEPOD): various reports 1989–95. Royal College of Surgeons, London.

Plum F, Posner JB. 1984. *Diagnosis of stupor and coma*, 3rd edn. F. Davis & Co., Philadelphia.

Ranklev E, Fletcher R, Krantz P. 1985. Malignant hyperpyrexia and sudden death. *Am J Forensic Med Pathol* **6**:149–50.

Reid L, Brace D. 1940. Irritation of the respiratory tract and its reflex effect upon the heart. *Surg Gynecol Obstet* **70**:157–62.

Relton JE, Britt BA, Steward DJ. 1973. Malignant hyperpyrexia. *Br J Anaesth* **45**:269–75.

Rognum TO, Vege A. 1997. [Forensic autopsy after possible medical malpractice. A 3-year material from the institute of forensic medicine in Oslo, 1993–95.] *Tidsskr Nor Laegeforen* **117**:2969–73.

Shanks JH, McCluggage G, Anderson NH, *et al.* 1990. Value of the necropsy in perioperative deaths. *J Clin Pathol* **43**:193–5.

Start RD, Cross SS. 1999. Acp. Best practice no. 155. Pathological investigation of deaths following surgery, anaesthesia, and medical procedures. *J Clin Pathol* **52**:640–52.

Truemner KM, White S, Vanlaudingham H. 1960. Fatal embolization of pulmonary capillaries. Report of a case associated with routine barium enema. *JAMA* **173**:1089–90.

Warden JC, Horan BF, Holland R. 1997. Morbidity and mortality associated with anaesthesia. *Acta Anaesthesiol Scand* **41**:949.

CHAPTER 24

Dysbarism and barotrauma

- Physical damage 488
- The autopsy in dysbarism and barotrauma deaths 489
- Further reading 490

Damage to the body from changes in external pressure usually lies in the province of specialized occupational or armed services medicine, rather than that of the forensic pathologist. The major exception is recreational and sporting diving, using SCUBA (self-contained underwater breathing apparatus) equipment. When death occurs in such circumstances, however, though criminality is rarely a factor, there is usually some form of official enquiry, be it quasi-judicial or an internal investigation. Civil suits for compensation may arise and in these circumstances a forensic pathologist may well become involved as the expert witness providing the autopsy evidence.

Injurious pressure effects are virtually always the result of a decrease in atmospheric pressure from a previous level. This can occur either from a decrease of high pressure to normal atmospheric pressure or a decrease from normal to a low pressure. An example of the first is the decompression of a diver ascending to the surface, while the second is seen in depressurization of an aircraft at high altitude. The first is by far the most common as a source of injury and death, mainly because of the great increase in recent years of diving activity associated with the offshore oil industry, which in Britain now forms the most dangerous type of commercial employment.

Dysbarism is a term that covers all adverse effects of pressure, while barotrauma describes the mechanical damage from gas released into the tissues.

In diving, each 10 m depth of water increases the pressure on the body by one atmosphere – 101 kiloPascals (kPa) – so that, in a non-rigid diving suit, the air supply has to be provided at the appropriate pressure relative to the working depth. This depth depends on the equipment used and is limited by legal constraints – for example, in Britain, dives of more than 50 m (164 ft) are not permitted using a nitrogen/oxygen mixture, a gas such as helium being mandatory. Similarly, surface decompression is only allowed at less than this depth of 50 m (164 ft).

When air is supplied at increased pressure, the contained nitrogen (and, to a much lesser extent, carbon dioxide) will progressively dissolve in the plasma and tissue fluids, the amount being governed by Henry's law (the solubility of a gas in a liquid is proportional to its partial pressure). This itself can produce nitrogen narcosis, a state resembling drunkenness in which orientation, judgement and rational behaviour may be diminished. This can occur at depths greater than 30 m. A more physical danger, however, is when a diver who has been at high pressure for long enough to have appreciable nitrogen dissolved in his tissues is returned too quickly to normal atmospheric pressure. The dissolved gas will then come out of solution, almost as champagne effervesces when the cork is released. Gas bubbles appear in the circulation, tissues and joint cavities, and can cause 'decompression sickness' with a wide range of clinical signs and symptoms.

Bubbles in the circulation can cause gas emboli that block small vessels and give rise to infarction, especially in the central nervous system, such as the spinal cord. In addition, such bubbles can interfere with the coagulation system, causing platelet aggregation and disseminated intravascular coagulation. Subcutaneous emphysema can appear as a result of gas under the skin.

PHYSICAL DAMAGE

When sudden, or too rapid, decompression occurs, there may also be physical damage from volume changes. The gas in body cavities expands markedly as the diver rises and, unless these can be progressively vented to the exterior, pressure effects from the contained gas will occur. Such effects may develop in the paranasal sinuses, if blocked, and even in the teeth, if small vacuoles of gas are trapped in the pulp cavities. The middle ear may also be involved and ruptures of the drum can occur. None of these lesions in themselves is likely to be fatal, however, unless the pain and disability interfere with performance – then drowning is the usual cause of death.

A more serious volume effect is in the thorax because, if the air in the lungs is not vented as the diver rises, the alveolar walls may be ruptured by the gas pressure. This leads to interstitial emphysema – air bursting through the respiratory membranes into the interalveolar, interlobular and interlobar septa. Bullae may appear on the pleural surface and, if these rupture, a pneumothorax is formed. Air may reach the mediastinum and track up the tissues to appear in the neck. Apart from a disabling pneumothorax, air can break into the lung capillaries and veins with the danger of air embolism to the left side of the heart, as it cannot move back down the pulmonary arteries. The filling of the left atrium and ventricle interrupts cardiac pump function and, if the volume is sufficient, can be fatal in itself.

Even though this does not occur in most cases of air entry, smaller volumes will be swept into the systemic arterial circulation and impact in the arterioles and capillaries of target organs, especially the myocardium, spinal cord and brain, causing microinfarcts, haemorrhagic necrosis and loss of function of vital tissues. This pulmonary barotrauma is not necessarily confined to deep diving, and may occur in only 2 or 3 m of water.

Many incidents occur to sporting divers, rather than to professionals, in relatively shallow water. The SCUBA diver is now a ubiquitous user of the seas, especially in the warmer, clearer waters of many parts of the world. It is estimated that there are about 80 000 recreational divers in the UK who complete more than two million dives each year. In 1987 there were 600 000 dives, which produced 162 incidents serious enough to be reported. Of these, 115 were of a ‘medical’ nature and included 8 deaths. There were 6 instances of air embolism, 69 of decompression sickness, 4 of hypothermia, 4 of nitrogen narcosis and 4 cases of ear damage. The incidence of medical incidents was 1/5000 dives, with 1 death in every 75 000 dives.

THE AUTOPSY IN DYSBARISM AND BAROTRAUMA DEATHS

Many of the vast quantity of reports on this subject deal with the clinical effects. Most deaths associated with diving are caused by drowning, though dysbaric disablement may have caused or contributed to the drowning. The autopsy must be directed to investigating the drowning aspect (Chapter 16), and to excluding or confirming any evidence of dysbarism or barotrauma.

The autopsy should not begin until some expert advisers are present, including someone with detailed knowledge of the equipment. If the victim has died in hospital or in a decompression chamber, then obviously much original information will be lost, as no *in situ* examination of the suit and breathing apparatus is possible. In any event, as few forensic pathologists are likely to be expert in this specialized field, every effort should be made to have the attendance of a naval or commercial person who is an authority on the equipment being used at the time of the death.

It has been recommended that the autopsy itself should be conducted in a decompression chamber in order to prevent the post-mortem dissipation of gas that must take place if the body is brought into atmospheric pressure. This is hardly likely to be a practical procedure except where the autopsy takes place under the aegis of armed-service or specialized commercial facilities, who have expert medical staff who are well versed in such procedures.

However, it must be recognized that the very act of decompression is likely to allow bubbles to form within the body, even if they were not there when death took place. This is an insurmountable difficulty for the pathologist unless, as just mentioned, the autopsy itself can be performed under pressure. Thus the finding of gas in the vessels and tissues has to be interpreted in the knowledge that they may be an artefact; part of the differentiation is quantitative, as a mass of froth distending the heart is unlikely to have accumulated post-mortem merely from the static fluid and tissues in the immediate vicinity, and is therefore more likely to have been brought there to form an ‘airlock’ by a vital circulation.

The autopsy should not begin – nor the body be undressed – until all necessary examination is made of the diving apparatus by an expert. Full photography at every stage, and radiology of the chest and major joints is essential before any dissection is attempted. It has been recommended that air samples should be taken by external lung puncture using a greased syringe so that chemical analysis can be performed to determine the oxygen:nitrogen ratio. This should only be performed under expert guidance. Externally, attention should be paid to the skin colour, as hypothermia is sometimes an associated hazard of diving. Unfortunately, many bodies recovered from water have a bright pink colour that may be post-mortem in origin, much as refrigerated bodies often have a pink hue. Red linear marks may be seen in those who have suffered ‘squeezes’ inside a neoprene drysuit when the external pressure has become excessive. Crepitance of the skin – usually of the head, neck and thorax – is an indication of barotrauma. The eardrums should be examined with an otoscope to detect any recent ruptures.

The brain should be removed before opening the body cavities, using a special technique to detect air in the cerebral arteries. Artery forceps should be placed on the middle cerebral, basilar and vertebral arteries before they are cut through; the brain should then be placed under water and the clamps removed to observe whether air bubbles escape from the cut ends of the vessels. The detection of air bubbles in the cerebral veins is of no significance, always being an artefact, as it is impossible to remove the calvarium without aspirating some air into the pial vessels. In any case, there is no way in which air in the arterial side of the cerebral circulation can penetrate the brain capillaries in sufficient volume to be visible in the cortical veins.

In the central nervous system, petechial haemorrhages may be seen in any part of the system, including the spinal cord. If the victim survives for a day or more, early infarcts may be seen. Histologically, ring-shaped haemorrhages around vessels may be seen in the white matter. In long-term survivors, living for some weeks, softening of the lateral and dorsal columns of the thoracic spinal cord may be seen if air emboli have impacted there.

Subcutaneous, mediastinal, retroperitoneal and subpleural emphysema must be sought. In the lungs, rupture of alveolar membranes allows air to track through the interstices from where it either enters the circulation or reaches the periphery of the lung to form subpleural bullae, or both. It can cross to the other lung via the hilum; in addition, local bronchial stenosis may trap air in a single lobe, causing a local distension.

The chest should be examined for pneumothorax, first by radiology and then by puncturing an intercostal space under a pool of water held in the lateral skin flap. Similarly, the heart should be examined radiologically for the presence of bubbles. If the air has emerged from the general tissue mass, it will have been conducted centrally by the venous drainage and so more air will be found in the right chambers than in the left side of the heart. If lung damage has occurred, then entry into the pulmonary venous system will bring it to the left side. It is hardly worth attempting to open the heart under a pericardium full of water, as froth in the chambers will be readily apparent if in sufficient volume to have caused death. Some authorities, however, advocate placing clamps on all the major thoracic vessels and the trachea before removing the thoracic pluck of organs. This is then placed under water and the clamps released to see if and from where air bubbles emerge.

Histologically, there are no particularly specific changes in spite of some claims. Lipoid-containing tissues, such as adipose tissue and adrenal cortex, may show microbubbles with a foamy appearance. There may be fatty change in the liver, possibly from air emboli. In lungs in which barotrauma has taken place, oedema, patchy haemorrhage, torn alveoli and focal emphysema confirm the macroscopic findings. The myocardium may show small foci of necrosis.

Fat embolism may occur, both in post-descent shock and in decompression in both water and high altitude. Not only fat, but cellular bone marrow may be found in target organs such as lung, myocardium, brain and kidney. There is controversy over the origin of fat emboli in decompression states, some claiming that it is caused by air-bubble disruption of fat depot tissues, while others favour a more complex genesis involving a redistribution of blood lipids, inter-relating with fibrin production and disseminated intravascular coagulation.

FURTHER READING

Brown CD, Kime W, Sherrer EL, Jr. 1978. Postmortem intravascular bubbling: a decompression artifact? *J Forensic Sci* **23**:511–18.

Busuttil A, Obafunwa J. 1995. A review of the forensic investigation of scuba diving deaths. *Sci Justice* **35**:87–95.

Calder IM. 1985. Autopsy and experimental observations on factors leading to barotrauma in man. *Undersea Biomed Res* **12**:165–82.

Calder IM. 1985. A method for investigating specialised accidents with special reference to diving. *Forensic Sci Int* **27**:119–27.

Calder IM. 1986. Dysbarism. A review. *Forensic Sci Int* **30**:237–66.

Calder IM. 1987. Use of postmortem radiographs for the investigation of underwater and hyperbaric deaths. *Undersea Biomed Res* **14**:113–32.

Davis OD. 1961. Physiological hazards of skin diving. *Med J Aust* **2**:1035–41.

Elliott D, Davis J. 1982. The causes of underwater accidents. In: Bennett PB, Elliott D (eds), *Physiology and medicine of diving*. Baillière Tindall, London.

Fryer DI. 1962. Pathological findings in fatal subatmospheric decompression sickness. *Med Sci Law* **2**:110–14.

Giertsen JC, Halvorsen JF. 1972. [Fatal scuba diving accidents. A discussion of 8 fatal cases.] *Tidsskr Nor Laegeforen* **12**:924–8.

Giertsen JC, Sandstad E, Morild I, *et al.* 1988. An explosive decompression accident. *Am J Forensic Med Pathol* **9**:94–101.

Goldhahn RT. 1976. Scuba diving deaths: a review and approach for pathologists. In: Wecht C (ed.), *Legal medicine annual 1976*. Appleton Century Crofts, New York.

Hart AJ, White SA, Conboy PJ, *et al.* 1999. Open water scuba diving accidents at Leicester: five years' experience. *J Accid Emerg Med* **16**:198–200.

Hayman J. 1985. Autopsy method for investigation of fatal diving accidents. *South Pacific Underwater Med Soc J* **15**:8–16.

Morild I, Mork SJ. 1994. A neuropathologic study of the ependymoventricular surface in diver brains. *Undersea Hyperb Med* **21**:43–51.

Obafunwa JO, Busuttil A, Purdue B. 1994. Deaths of amateur scuba divers. *Med Sci Law* **34**:123–9.

Smith N. 1995. Scuba diving: how high the risk? *J Insur Med* **27**:15–24.

Williamson JA, King GK, Callanan VI, *et al.* 1990. Fatal arterial gas embolism: detection by chest radiography and imaging before autopsy [see comments]. *Med J Aust* **153**:97–100.

CHAPTER 25

The pathology of sudden death

- Sudden or unexpected deaths 492
- Sudden death from cardiac disease 493
- Coronary atherosclerosis 493
- Sites of coronary stenosis and occlusion 495
- Myocardial infarction 495
- Complications of myocardial infarction 502
- Cause of death in coronary insufficiency 503
- Hypertensive heart disease 505
- Sudden death in aortic valve disease 506
- Sudden death from the cardiomyopathies 507
- Death in old age – the senile myocardium 509
- Sudden death from rupture of an aneurysm 509
- Pulmonary thromboembolism 513
- Sudden death in epilepsy 514
- Sudden death in bronchial asthma 515
- Respiratory obstruction 515
- Haemoptysis 516
- Gastrointestinal haemorrhage 516
- Fatal abdominal catastrophes 516
- Genitourinary system 517
- The relationship between trauma and disease 517
- References and further reading 520

Virtually all forensic pathologists deal not only with criminal, suspicious, accidental and suicidal deaths, but with a wide range of deaths from natural causes. Many of these are sudden, unexpected, clinically unexplained or otherwise obscure, even though there need be no unnatural element in their causation.

It is good that such a large substrate of natural deaths is available to most forensic pathologists: the situation where they deal exclusively with trauma and crime is professionally unhealthy, as they become progressively more out of touch with morbid anatomy, and lose daily contact with disease processes and uninjured tissues and organs. Involvement with natural death means frequent professional intercourse with clinicians and non-forensic pathologists, with all the consequent benefits of cross-fertilization of knowledge and ideas. To work in a totally forensic vacuum is to lose touch with pathological and clinical reality, which is essential for a medico-legal expert to retain a sense of proportion and an awareness of contemporary medical advances.

Another indispensable benefit of sustained experience in natural disease is the fact that some of the most difficult problems in criminal and litigious cases arise not out of gross, rapidly fatal, trauma, but in deaths where concurrent natural disease or complications after trauma lead to a fatal outcome. The assaulted victim that dies later from a stroke or the negligent minor accident that has a fatal pulmonary embolism – these can pose far greater difficulties over causation than a gunshot wound or a stabbing.

In this chapter no attempt is made to duplicate the detailed descriptions of disease processes provided in a score of illustrious textbooks of pathology, but a survey will be offered of the spectrum of causes of sudden or unexpected death as commonly encountered by forensic and 'coroner's' pathologists.

SUDDEN OR UNEXPECTED DEATHS

The definition of a sudden death varies according to authority and convention. The World Health Organization definition is of death within 24 hours from the onset of symptoms, but this is much too long for many clinicians and pathologists; some will only accept death within one hour from the onset of illness. We have to also bear in mind that a death may appear sudden and unexpected to an outsider but need not have been so from the point of the pathological disease process. The deceased may have been symptomless and utterly unaware of his chronic disease or he may have had symptoms but interpreted them as harmless. Also, fear, lack of human contact or his own disposition may have prevented him mentioning symptoms to anyone, including a doctor.

In many jurisdictions, deaths may only be certified by an attending physician if he has seen the patient recently and is satisfied that the death was caused by a potentially lethal disease from which he was aware the patient suffered. The fact that, without autopsy, this physician is wrong in his belief in between 25 and 50 per cent of cases cannot concern us at the moment, but the relevance is that, where a clinical doctor cannot so certify, the death is usually reported for medico-legal investigation. In many countries such notifications form by far the largest proportion of medico-legal autopsies, and in England and Wales they account for some 80 per cent of coroner's autopsies, the remainder being suicide, accident and homicide.

The description 'sudden' or 'unexpected' is not always accurate, as 'unexplained' is an equally common reason for medico-legal investigation. Here the clinician is unable to offer a cause for the death, though the patient was under medical care. Even after autopsy, the cause of death may still not be revealed and this problem of the obscure autopsy is discussed elsewhere.

In sudden death, the immediate cause is almost always to be found in the cardiovascular system, even though topographically the lesion is not in the heart or great vessels. Massive cerebral haemorrhage, subarachnoid bleeding, ruptured ectopic pregnancy, haemoptysis, haematemesis and pulmonary embolism, for example, join with heart disease and aortic aneurysms to contribute most of the vascular system reasons for sudden, unexpected death.

SUDDEN DEATH FROM CARDIAC DISEASE

Ischaemic heart disease is easily the most common cause of sudden death in Western nations. The term is employed rather loosely and inaccurately, often being used as if it was synonymous with 'coronary atherosclerosis'. Certainly the latter is the largest contributor to ischaemic heart disease, but not to the exclusion of other conditions.

As a further complication of nomenclature, 'coronary atherosclerosis' or 'coronary atheroma' is the proper title of the common degenerative disease that causes most deaths, but the term 'coronary artery disease' is so ingrained into popular usage that it is accepted as synonymous with coronary atherosclerosis, even though there are other coronary diseases.

Ischaemic heart disease comprises:

- coronary atherosclerosis
- hypertensive heart disease
- aortic valve disease
- anomalies of the coronary circulation
- other coronary artery diseases, such as polyarteritis
- cardiomyopathic enlargement
- some congenital heart disease.

CORONARY ATHEROSCLEROSIS

Sometimes called 'the Captain of the Men of Death', this is certainly the most frequent cause of sudden death in Western societies. The basic mechanism is stenosis or occlusion of one or more major branches of the coronary arteries by atheromatous lesions, or one of the complications of such a lesion. The severity of stenosis before death occurs is debatable: cardiac pathologists claim that at least 80 per cent of the normal lumen must be lost before myocardial necrosis occurs. Most forensic pathologists would admit to blaming coronary atheroma for death in autopsies where a significantly smaller percentage of the lumen was lost, when the symptomatology and circumstances were strongly suggestive. Areas of myocardial fibrosis and – rarely – even recent infarcts, may be seen in the presence of relatively small degrees of coronary atheroma. Part of the problem lies in quantifying the amount of stenosis at autopsy. Because of the empty vessels lacking the normal intraluminal blood pressure, the walls are lax, collapsed and unstretched. During life, the same lumen will be larger and only fixation at arterial pressure will retain a semblance of the original size. The pathologist sees all coronary arteries in the dead state, however, so the relative degree of stenosis is still a reliable observation, as both normal and diseased vessels can be compared, even if their absolute luminal size is incorrect.

Coronary atheroma may be focal, with irregular plaques that vary both in size and in stenosing effect from place to place. They may be few and localized, with a virtually normal lumen in the rest of the system. This means that every part of the major vessels must be examined at autopsy, with transverse cuts at no more than 3 mm intervals. It is unjustifiable to make a few cuts at wide intervals and claim that this provides a reliable estimate of the state of the coronary system. A coronary artery is like a chain – just as the latter is only as good as its weakest link, so a vessel is only as good as its narrowest point.

Two cuts, if even half a centimetre apart, may pass either side of a grossly stenosing plaque or a thrombus. Opening the vessels lengthwise obviates this danger, but this is outweighed by the disadvantage of not being able to measure accurately the degree of stenosis by making frequent cross-cuts. A compromise is the best answer to this dilemma, as described in Chapter 1.

Plaques are often eccentric, leaving a crescentic residual lumen, but another variety of coronary atheroma is a more diffuse lesion, where the vessel is concentrically narrowed by a more longitudinal degeneration, forming a tube-like segment with a narrow, sometimes 'pin-hole' central lumen. The author (BK) has the impression that this is more common in younger victims and that such lesions seem less likely to be ulcerated and calcified. Both types of lesion are often found in the same patient and there is presumably no fundamental difference in their pathogenesis.

Complications in coronary atheromatous lesions

ULCERATED PLAQUES

The simple endothelial thickening develops to involve the media and usually becomes infiltrated with lipids. Whilst the covering endothelium remains intact, the danger to life is confined to the luminal reduction from the bulge of the enlarging plaque. When the fibroendothelial cap begins to break down under the pressure and erosion of the central necrosing process, the plaque may rupture into the lumen. This has several consequences, which may precipitate acute symptoms or even death.

First, a large plaque may suddenly disgorge its pultaceous contents into the lumen. This can mechanically block the vessel or seriously stenose it. Thrombosis may be precipitated at that spot by the slowed blood flow plus thrombogenic tissue elements. The cap of the plaque may be projected across the lumen, forming a sudden and perhaps complete occlusion. This will be more dangerous if the open end of the ruptured plaque faces 'upstream' into the direction of the blood flow, so that the intimal cap is lifted up, causing a valve-like obstruction. When the blood flow sweeps across the cap in the opposite direction, the cap will not be displaced, though the soft, grumous contents of the plaque will be washed downstream. Sometimes the flap can be seen in histological sections, though it is difficult to know if this is a processing artefact. It is best demonstrated directly at autopsy, when two cuts are joined longitudinally and the intima examined with a lens or low-power microscope.

The contents of the plaque may be washed downstream by the blood flow, eventually impacting in small branches and occluding them. These can be seen histologically using lipid stains and the multiple small infarcts that often ensue can be seen either in classical histological sections or, more readily, by enzyme or fluorescent techniques. They are probably responsible for many of the small fibrotic areas commonly seen in the myocardia of sufferers from long-standing coronary disease.

HAEMORRHAGE

Haemorrhage often occurs within an atheromatous plaque, usually into the softened, necrotic centre. This 'subintimal haemorrhage' may give rise to a sudden reduction in the blood-carrying capacity of a coronary artery and cause sudden death. The source of the bleeding is somewhat controversial, but the best explanation is that it comes from rupture of small blood vessels in the periphery of a plaque. A normal coronary artery does not have a blood supply to the intima but, in the disorganization frequently associated with the distorted microanatomy of a diseased coronary artery, such vessels may lie within the intima and be eroded by extension of the degenerative process.

Haemorrhages, both fresh and old, can often be demonstrated in histological sections of atheromatous plaques. It is reasonable to presume that such a bleed may be precipitated by some sudden rise in blood pressure from exertion or emotion, though, if a vessel is sufficiently eroded by atheroma, such a precipitation is not strictly necessary for such an event to take place.

The sudden release of blood into an atheromatous lesion can rapidly enlarge it and may raise the cap of the plaque towards the other side of the already stenosed lumen. It may even rupture the plaque: some subintimal haemorrhages may track circumferentially, causing a 'minidissection'. Subintimal haemorrhage as a factor in acute coronary occlusion was investigated many years ago by Patterson (1938), who believed that they played a major part in coronary thrombosis. In the same year, Wartmann demonstrated occlusion by subintimal haemorrhage without rupture of the overlying intima. Though English and Williams (1943) doubted if haemorrhage could compress the lumen against arterial pressure, other authors, including Davies and Pomerance (1975) believe that the haemodynamics within the coronary vessels are such that the pressure varies markedly from place to place, and that there may even be a contributing factor to subintimal haemorrhage by a suction effect in certain conditions. Other authors, such as Drury (1954), consider that the blood within the plaque is not caused by haemorrhage from mural vessels, but from blood percolating from the lumen of the coronary artery at the time of rupture of a plaque. This seems a minority opinion, however, as many haemorrhages can be seen (some older ones containing haemosiderin) within plaques which have a perfectly intact overlying intima.

Whatever the precise origin of the blood, it is clear that subintimal haemorrhage is a potent factor in rapidly reducing the available lumen of a coronary artery – and sometimes precipitating thrombosis by further stretching and damaging the overlying intimal cap.

CORONARY THROMBOSIS

The atheromatous plaque may not undergo these more dramatic changes, but progressive internal necrosis may erode the luminal surface and expose the fibrofatty contents. This loss of normal covering endothelium then forms a nidus for thrombus formation, which may gradually accrete in layers so further reducing the lumen or even occluding it, especially if a combination of lesions causes the plaque to expand at the same time as roughening the surface. Thus mural thrombus may completely block or severely narrow the residual lumen, with all the consequences of reduced blood flow to the distal myocardium.

A narrow lumen is by no means essential to the formation of thrombosis, as it may occur in the abnormally wide, virtually aneurysmal, coronary arteries sometimes seen in aged people. The lumen may be up to a centimetre wide yet be firmly thrombosed, presumably solely because of damage to the intima. Thrombosis often occurs in recanalized vessels, secondary thrombosis taking place after organization and re-establishment of a lumen through the previous block. Multiple coronary thromboses are by no means unusual. Many are post-infarct, the original thrombus causing myocardial necrosis and the resulting stasis in circulation, together with the thrombogenic effect of tissue damage leading to sluggish flow of readily coagulable blood.

It has also been noted that coronary thrombosis may be accompanied by thrombotic lesions elsewhere in the body. For example, a coronary thrombosis may be followed by a pulmonary embolus from thrombosed leg vein – and the converse may also occur in non-fatal pulmonary embolism. The coagulability of the blood, together with circulatory stasis, aided by immobility in bed, are obvious factors.

SITES OF CORONARY STENOSIS AND OCCLUSION

The whole length of the coronary arterial system is not uniformly vulnerable to atheromatous lesions. First, the major trunks are most affected where they lie subepicardially, often in the fatty surface tissue. Once the arteries dip down into the myocardium, these more distal intramuscular branches become much less prone to significant atheroma, especially of the grumous, degenerative type, though intimal thickening may still be seen.

Second, there are sites of predilection in the three major arteries. The most common site of occlusion is in the first 2 cm of the anterior descending branch of the left coronary artery, which is more frequent than in the common trunk. The next most frequent site is in the right coronary artery, but here the thrombosis is more distal than in the left vessel, usually seen as the vessel courses around the right margin of the heart in the atrioventricular groove part-way between aorta and the beginning of the posterior descending branch.

The third most common place is the proximal part of the left circumflex artery, soon after the bifurcation from the common trunk. The latter is then the next most frequent site, in the short segment (sometimes absent) between the aorta and the birfucation into descending and circumflex branches. At autopsy, it is always necessary to transect the proximal part of both coronary arteries right up to the coronary ostia, as occlusion can sometimes be present in the first few millimetres.

MYOCARDIAL INFARCTION

There is a different emphasis laid upon myocardial infarction by the forensic pathologist and by the clinician. To the latter, the clinical signs of chest pain and shock mean an infarct and certainly, a much higher proportion of patients who reach hospital beds do have myocardial infarcts compared with those who are taken straight to the mortuary. Recently the Joint European Society of Cardiology/American College of Cardiology Committee for the Redefinition of Myocardial Infarction produced a consensus document examining the scientific and societal implications of a new definition for myocardial infarction from seven points of view: pathology, biochemistry, electrocardiography, imaging, clinical trials, epidemiology and public policy.

In the pathology of sudden death, overt infarcts are the exception, rather than the rule. The incidence in autopsy material from sudden deaths varies considerably, partly because of the different methods of demonstrating the muscle necrosis, but speaking of obvious naked-eye infarcts, the author (BK) would estimate that, in his material considerably less than a quarter of deaths attributable to coronary atherosclerotic disease have myocardial infarction. In Finnish autopsy material of the author (PS) the 5-year average (1987–1991) of infarcts among deaths due to coronary atherosclerotic disease was 13.5 per cent. The relevance of this observation in relation to the mechanism of sudden cardiac death is discussed later.

Almost all myocardial infarcts are caused by atheromatous lesions and their complications. A few are the result of other types of coronary obstruction, such as polyarteritis, other vasculitides, embolism of various types, ostial occlusion by syphilis, severe aortic stenosis, some congenital anomalies of the coronary arteries or great vessels, dissecting aneurysms at the aortic root, and tumour or sarcoidosis affecting the coronary vessels.

Most infarcts are caused by super-added coronary thrombosis, but even this statement needs further analysis. A proportion of infarcts have no demonstrable thrombus in the supplying vessel, but this proportion can be reduced by a more careful search. There is no doubt, however, that muscle necrosis can follow severe narrowing of the supplying vessel by subintimal haemorrhage, a ruptured plaque or simple severe stenosis from atheroma. In addition, grumous debris from a ruptured plaque can produce embolic occlusions with 'microinfarcts' in the distal myocardium. The fact that some thromboses are post-infarct has already been mentioned.

It is said that the original lumen must be reduced to 20 per cent or less before the ischaemia in the distribution zone is sufficient to cause myocardial necrosis. Most pathologists with considerable experience of sudden death autopsies will, however, have numerous experiences of undoubted infarction in the absence of an 80 per cent stenosis. The embarrassing situation occasionally occurs where an infarct exists with virtually normal coronary arteries.

The contrary is much more common: the finding of complete thrombosis of a major vessel with no sign of infarction. In this case, either an effective collateral circulation provided an alternative blood supply – or the victim died before signs of infarction had time to develop.

There are several types of myocardial infarct recognizable at autopsy:

- A laminar infarct in which the subendocardial region of much – or even all – of the left ventricle is involved, sometimes extending through half or more of the thickness of the wall. This is due to a reduction in perfusion pressure to the inner zones, as all the coronary supply comes from the epicardial surface. Laminar infarcts are the result of generalized stenosis in the major branches of the coronary vessels, but there is usually a second factor, in that a drop in blood pressure or the oxygenation of the blood compromises the already poor supply so that the outer zones of the ventricular wall consume the available oxygen and nutrients, leaving little for the inner zone. This is discussed further under hypertension and aortic stenosis.
- A regional or focal infarct is more common in pure coronary artery disease, and is caused by a localized occlusion or severe stenosis in a coronary artery. These are true myocardial infarcts, as the definition of an infarct requires occlusion of the vascular supply, which strictly excludes some laminar infarcts when perfusion pressure or relative insufficiency is caused by hypertension, or aortic valve disease. The regional infarct is a topographically demarcated zone of muscle necrosis, the size and position depending on the site of vascular occlusion, though any collateral supply may modify these. Almost all infarcts are in the left ventricle. Wartmann and Hellerstein (1948) found that between 6 and 9 per cent involved the right ventricle and that 7 per cent involved the atria, though Cushing's (1942) figure was 17 per cent for atria. Most of these infarcts, however, were overlapping from infarcts in the left ventricle.

The relative immunity of the right ventricle and atria is presumably because of the relatively thin walls, which do not require so much blood, do not have the perfusion gradient of the thick left ventricle and which can more easily obtain oxygen and other constituents from the blood in the lumen of the ventricle. The presence of a conus artery, small branches to foci, such as to the sinus node, and the presence of anastomotic channels between the left circumflex and right coronary arteries must all contribute to the better survival of these areas. The site of a left ventricular infarct does not always correspond strictly to the site of the coronary occlusion or thrombus because of the presence of a collateral circulation.

It was also shown by Wartmann and Souders (1950) that many infarcts involve only certain layers of muscle, which cannot be supplied by a particular vessel. There is obviously no simple explanation for the topography of a myocardial infarct – this is further confirmed by the patchy nature of myocardial damage on microscopic study, as the alternate necrosis and survival of adjacent fibres or even segments of the same fibre cannot be explained solely by coronary perfusion.

The autopsy diagnosis of early myocardial injury and infarction

The autopsy demonstration of an acute lesion in the heart, such as an early myocardial infarct, can have profound medico-legal implications. In a fatal traffic accident, or even rail or air crash, the proof of an acute disabling myocardial lesion in the driver may be vital in the investigation of the event and the apportionment of legal liability. In potentially criminal deaths, the presence of a recent infarct again may be relevant in causation or as a contribution to the death. It is thus important to make a full histological search for evidence of myocardial fibre damage, using all the methods available, including histochemical and fluorescent techniques.

The macroscopic appearances of myocardial infarction are described with a considerable lack of uniformity in most pathology texts, partly because of the varying ages of infarct that the authors depict. The age of an infarct is notoriously difficult to establish in the human, as the onset of clinical symptoms, however dramatically abrupt, are often much later than the onset of the pathological lesion precipitated by a coronary occlusion. In animal experiments, a coronary vessel can be ligated at zero time and serial sacrifices made at different intervals to gain an accurate estimate of the age of the infarct.

In the human, the time of chest pain and shock cannot be used in a similar fashion. When a victim of coronary disease dies say, 8 hours after the onset of acute symptoms, though one might expect an early infarct to be visible histologically or histochemically, not infrequently a demarcated yellow or tigroid area of necrosis is present, which must be several days old. The following appearances are typical of the stages of myocardial infarction, though marked divergences from this time scale are not uncommon:

- For the first 12–18 or even 24 hours, no definite naked-eye changes are visible. The first sign towards the end of this variable period is oedema of the affected area of muscles, which causes pallor as the swollen fibres squeeze the blood from the vessels that lie between them. The normal moist lustre of the cut surface becomes more granular and dull when cut by a sharp knife. Caution must be used with this feature, because if the scissors are used to cut even normal muscle, the crushed edges of the fibres will give a similar dry, matt appearance.
- From about the end of the first day progressively through the second and third, the area becomes better demarcated and turns yellow. With breakdown of the myocytes, streaks of red appear, being both dilated vascular channels and areas of interfibre haemorrhage. This gives a 'tigroid' appearance suggestive of tiger stripes across the area, though sometimes the yellow element is virtually uniform or the red streaks may fade after a few days.

  Demarcation becomes more sharply apparent and there may be a red zone in the less damaged muscle around the periphery. The size of the infarct depends on the means with which it is visualized; the naked-eye size is less than the area detectable by classical histology and this in turn is smaller than that demonstrable by enzyme studies, where the more peripheral fibres are sick but potentially recoverable.
- After a few days, progressing through weeks, the infarct becomes softer and more friable, justifying the old name 'myomalacia cordis'. It is at this stage, from about the second or third day onwards, that ruptures occur into the pericardial sac. Some pathology textbooks are in error in stating that such cardiac rupture occurs in early infarcts.
- From the third week and later, the centre of the infarct becomes gelatinous, the colour fading to a translucent grey that often subsides below the level of the cut surface. This rather mucoid stage may still have old haemorrhage within it, but the stage of healing slowly develops and, during the next month or two, depending upon the size of the infarct, fibrosis replaces the dead muscle to form a scar. Even after this stage, there is often a narrow zone around the young fibrous tissue that shows muscle necrosis, though special staining or enzyme techniques may be needed to demonstrate this.

The laminar infarct goes through the same cycle of changes, but this is often less intense. Eventual fibrosis may be widespread but remains thin, often subendocardial, especially on the left ventricular aspect of the interventricular septum where a wide glistening sheet may obscure the underlying muscle. The apex also may show widespread fibrosis and all muscle may be replaced in this region, sometimes leading to a cardiac aneurysm on the free wall. Infarcts may be transmural, extending from epicardium to endocardium, or they may be confined to the inner zone. It is almost impossible to have an infarcted area confined to the outer subepicardial zone because of the topography of the coronary supply. The papillary muscles are usually involved, being particularly vulnerable to ischaemia as they are at the end of the line of coronary supply. The central part of the muscle may necrose and even rupture. Infarction usually spares the immediate subendocardial zone, the three or four most superficial layers of fibres surviving, even though they may show ischaemic damage. They presumably receive enough oxygen and nutrients from the ventricular blood to survive, though this does not seem to prevent deposition of mural thrombus over the infarcted area.

Microscopical appearances of myocardial injury and infarction

There are a vast number of publications on this subject dealing with morphological changes, classical histological methods, enzyme histochemistry, fluorescent studies, immunohistochemical methods and transmission and scanning electron microscopy. Only a brief summary can be offered here, though some of the methods of detecting very early infarction are detailed in Appendix 1.

MORPHOLOGICAL CHANGES

There is marked variation from case to case, just as in the gross appearances. The same difficulties exist about assessing the time of onset and hence the age of the infarct. This may sometimes have considerable medico-legal importance,

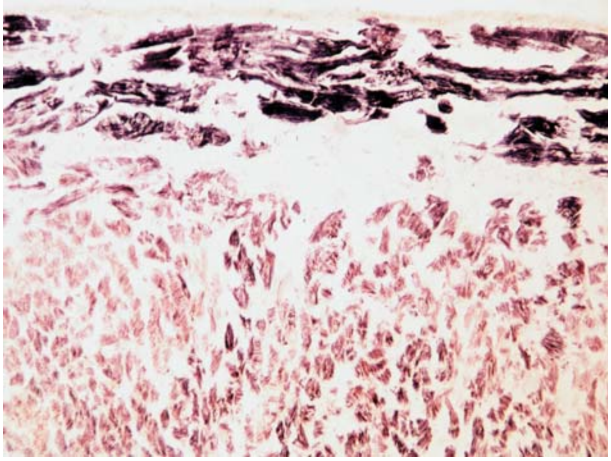

FIGURE 25.1 *Enzyme histochemistry of early myocardial damage. This photomicrograph shows strong succinate dehydrogenase reaction in the myofibres adjacent to the endocardium and moderate decrease of reaction elsewhere suggesting early injury. (Original magnification ×100.)*

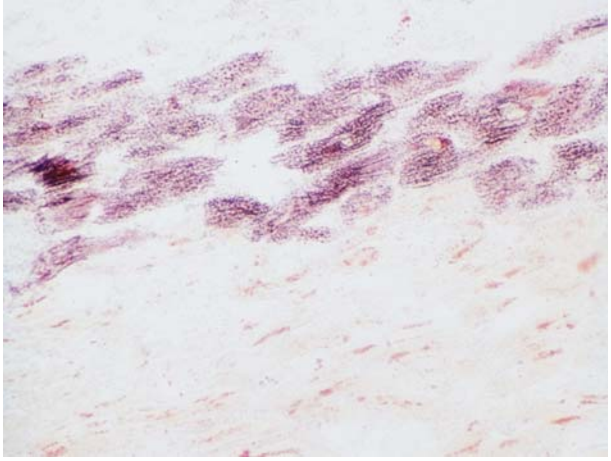

FIGURE 25.2 *Fresh myocardial infarction with β-hydroxybutyrate dehydrogenase staining showing positive reaction in the myofibres adjacent to a larger vessel seen above as an empty space. The lower part of the picture shows infarcted myofibres with some lipofuscin but no enzyme reaction. (Original magnification ×160.)*

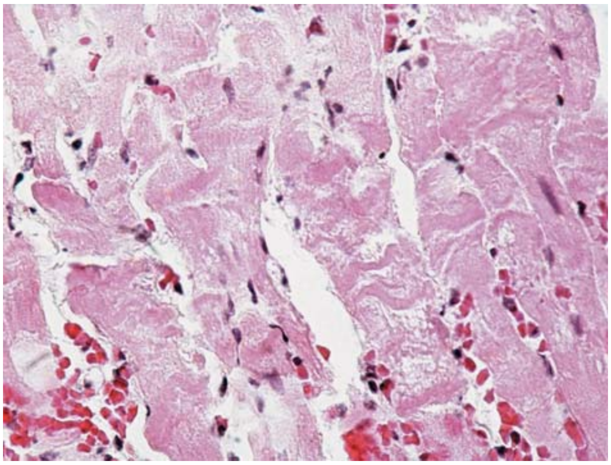

FIGURE 25.3 *Contraction bands in myocardium. (HE, original magnification ×500.)*

especially when cardiac disease may have caused a traffic or other transportation accident – and is sometimes relevant in criminal circumstances. In animal experiments, changes in myocardial fibres become apparent within 30 seconds of coronary ligation, the first change being a loss of glycogen as the cell shifts from aerobic to anaerobic respiration. This cannot be observed in human material, for obvious reasons. The first observable change is subjective and different pathologists given the same material will offer different opinions on whether or not ischaemic damage is visible. Some of the earliest changes are mimicked by autolysis, especially granularity of the cytoplasm. The infarcted area becomes swollen, so that the interfibre clefts become obliterated, and the overall appearance is pinker and more solid in routine sections. Care must be taken to assess sections of uniform thickness, as material from other laboratories may not conform to the observer's usual quality. Along with swelling, the cytoplasm becomes granular, the 'cloudy swelling' of the old texts.

Early morphological changes seen in light microscopy are unspecific: interstitial oedema, congestion and small haemorrhages. An overstretching of myofibres with a sarcomere length of about 2.51 μm is possible only when the fibre is injured. In experimental infarction in the rat and the dog sarcomere, overstretching has been observed 15 minutes after coronary ligation. Cellular infiltration is variable and neutrophil infiltration is seen in most infarcts during the first few days being replaced by a mononuclear response within a week or so.

Contraction bands, known also as myofibrillar degeneration, and coagulative myocytolysis, are commonly described in connection with different pathological conditions. The change signifies irreversible myocyte injury. Contraction bands are irregular, dense and in haematoxylin and eosin (HE) staining eosinophilic transverse bands alternate with lighter staining granular zones in the myocyte, possibly firstly described by Smith (1904) in the periphery of myocardial infarction. These lesions have been observed in various forms of experimental heart muscle injury and have been also seen, e.g. in association with potassium deficiency, magnesium deficiency, in cases of malignant hyperthermia, in the hearts of drug addicts and in cases of sudden cardiac death.

They may also be seen after blunt trauma to the chest or even cardiopulmonary resuscitation, both where mechanical massage has been applied and especially where cardiac stimulants like noradrenaline have been injected. In fact, catecholamines seem able to cause this fibre disruption, which may also appear during electrocution and other conditions

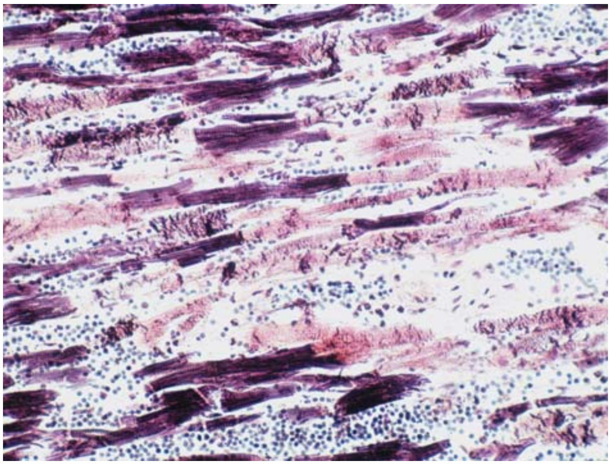

FIGURE 25.4 *Contraction bands in fresh myocardial infarction. (PTAH, original magnification ×250.)*

of cardiac stress. In sections stained with haematoxylin and eosin, pale areas in the fibre are common as a result of fibrocytolysis and the sarcolemma is often seen to be deficient at these sites. Although the change signifies fresh and irreversible myocyte injury and is more frequently observed in cardiac deaths, it is not specific to the cardiac death.

Contraction bands are best seen by phosphotungstic acid haematoxylin stain and all myocardial sections should be stained in this way as well as – or even in preference to – haematoxylin and eosin. Phosphotungstic acid haematoxylin stain does not reveal ischaemic damage at an earlier stage, but renders it far more obvious. Large areas can be scanned under low power of the microscope and the danger of missing abnormal areas is far less than with haematoxylin and eosin. The striations become distorted and break up, forming either a general 'sandy' appearance or the striking 'contraction bands' where fragmentation occurs in transverse lines across the fibre. The contractile material aggregates into thicker, darker masses, sometimes like 'Chinese writing', at first within an intact sarcolemma, but later the cell membrane ruptures and the basophilic contents scatter extracellularly.

Zonal lesions occur as a region of supercontraction of myocytes at the intercalated discs. These lesions have been observed both in the dog and in humans after haemorrhagic shock and the changes have been considered ubiquitous in and pathognomonic for hypovolaemic shock. These possibly reversible changes have been observed in experimentally induced shock in animals 15 minutes after the induction. The presence of catecholamines is also considered necessary.

Fragmentation of myocardial fibres is, at least to some extent, practically always present in adult hearts and more frequently in the wall of the left ventricle. Although it has generally been associated with post-mortem changes, it has also been shown to be more frequently seen in myocardial samples from sudden cardiac death victims. Early injury possibly increases the brittleness of myocardium, resulting in increased fragmentation of myofibres after the fixation and cutting process. This phenomenon has, however, no value as a positive marker of myocardial injury.

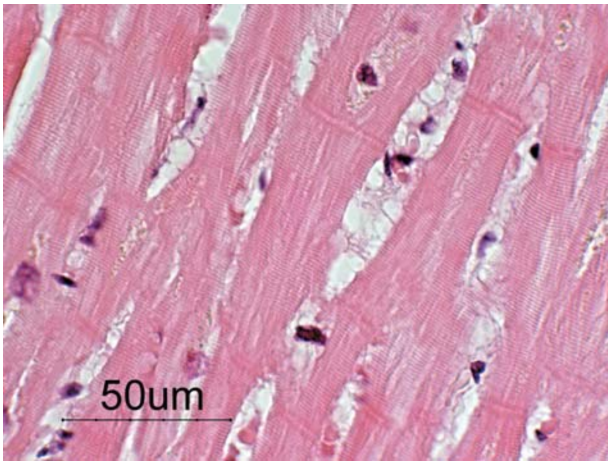

FIGURE 25.5 *Zonal lesions in haemorrhagic shock (HE).*

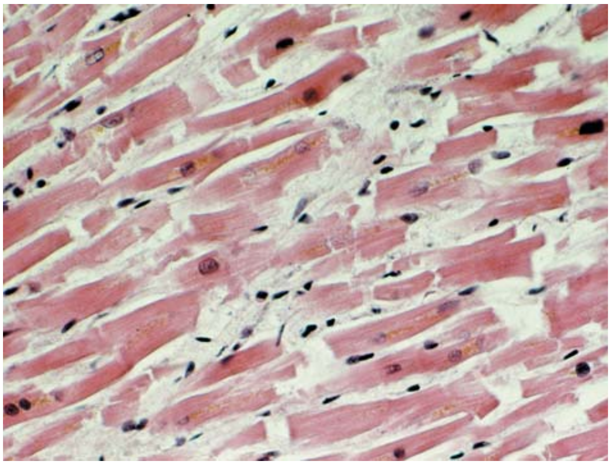

FIGURE 25.6 *Fragmentation of myocardial fibres. (HE, original magnification ×250.)*

Waviness of fibres was noted by Glogner (1903), who saw corkscrew-shaped myofibres in the heart and skeletal muscle in some cases of beri-beri. Similar patchy deformation of myofibres in a wavelike fashion was seen in association with fatty degeneration of the myocardium and in early and also more advanced human myocardial infarction. Both in cardiac deaths and non-cardiac deaths it can be observed more frequently in the wall of the right ventricle. It has no diagnostic significance as to the cause of death and it signifies only agonal injury.

LATER STAGES

The later stage of infarction follows the 18–24-hour period and, as by now macroscopic changes will be apparent, the histology is merely a confirmatory procedure. Even here, there is great variation in appearances, especially in cellular infiltration. There is a progressive degeneration of the fibres, best seen in longitudinal section. Eosinophila becomes stronger for a time until the internal architecture breaks down. The cellular oedema subsides and interfibre oedema takes over, separating the myofibres, a process that becomes more marked as the latter shrink and fragment.

Loss of nuclei is not an early phenomenon but, by the second to fourth days, the nuclei become hollow, shadowy and eventually ghost-like, though nuclear remnants can sometimes still be seen in completely necrotic infarcts. Cellular infiltration is variable and sometimes does not occur at all, again in contradiction to the established descriptions in most textbooks. In general, however, there is a neutrophil infiltration in most infarcts during the first few days, which subsides gradually and is replaced by a mononuclear response within a week or so. These are mainly macrophages clearing up the debris and fibroblasts coming in to lay down new collagen during repair.

By the end of the first week, disintegration of muscle fibres is well under way, and new capillaries and fibroblasts are appearing, though there are many exceptions to this timetable. In general, infarcts are older than most observers think, so that their age is usually underestimated. By the fourth week, depending upon the total size of the infarct – early fibrosis is evident but still with lacunae of necrotic, unrepaired muscle, which heals unevenly and slowly. From this point on there is a variable population of cells, leucocytes diminishing in favour of fibroblasts, macrophages and new vessels.

It is difficult, if not impossible, to date such an infarct even within weeks, and different parts of the same infarct may show quite different appearances. Only when sufficient time has elapsed for all the lesions to catch up with the end stages of firm fibrosis can one say that probably a minimum of about 3 months has elapsed – though again this is dependent on the size and other more imponderable factors, such as age and other concomitant disease.

HISTOCHEMICAL METHODS

In haematoxylin and eosin (HE) staining, increased cytoplasmic eosinophilia is an early form of cellular injury partially resulting from increased binding of eosin by cytoplasmic proteins. At the pH conventionally used in staining, eosin is negatively charged. The binding may increase because of greater exposure of positively charged reactive sites along polypeptide chains after their denaturation. The loss of cytoplasmic basophilia usually reflecting detachment and scattering of polysomes from rough-surfaced endoplasmic cisternae is considered as another cause for the increased eosinophilia.

The intensity of eosinophilia increases, sometimes uniformly, but often in patches with advancing injury. It is earlier and more distinctly seen in HE-stained cryosections and can be enhanced by placing a green filter in the light path of the microscope. Patchy hyperchromasia can be observed in all kinds of deaths and therefore it is of limited diagnostic value. Care must be taken to exclude the occasional thickened areas made by cutting artefacts, which look pinker because of a greater depth of cytoplasm. These early changes are not usually visible in the first 8–12 hours after the (presumed) onset of infarction.

Acid fuchsin (AF) was suggested to be useful for demonstration of early myocardial infarction in man or in animal experiments, but was found to be inconsistent and misleading.

Haematoxylin/basic fuchsin/picric acid stain (HBFP) was claimed to demonstrate early myocardial ischaemia and to be unaffected by post-mortem autolysis. In experiments with isolated cardiac myocytes in suspension made anoxic by complete oxygen deprivation, basic fuchsin was taken up by contracted or damaged myocytes, which, according to their morphology in suspension, revealed irregular contractions, but neither by undamaged nor necrotic myocytes. However, after the first promising reports, contradictory results started to accumulate and the method has been shown to produce inconsistent results and to be unreliable and non-specific and without value for diagnostic purposes.

FLUORESCENT METHODS

Fluorescent methods have been applied to demonstrate early myocardial infarctions and myocardial degeneration in animal experiments as well as in human heart either by using a fluorescent dye, e.g. acridin orange to stain unfixed cryostat sections or paraffin sections or utilizing the fluorescent properties of eosin in the HE-stained myocardium. Intravenously or intraperitoneally injected tetracycline has also been used in experimental infarction for demonstration of the perfused region in myocardium. Carle (1981) claimed that only hypereosinophilic cells autofluoresced, but Badir and Knight (1987) found that myocardium that had normal eosinophilic staining also fluoresced yellow in ultraviolet light; this has been confirmed by Saukko and Knight (1989).

Acridin orange stained cryosections of intact myocardium show golden brown fluorescence which turns into greenish fluorescence with increasing ischaemia time whereas eosin fluorescence of normal myocardium in paraffin-embedded samples show olive-green fluorescence which turns into yellow in injured tissue. Post-mortem autolysis does not seem to have any significant effect on the fluorescence but the high percentage of wrong positive samples indicates that at least eosin fluorescence is obviously too sensitive injury marker capable of demonstrating agonal ischaemic changes.

A major handicap common to all classical histochemical staining methods is that the basis of these colour reactions is poorly understood. Considering the possible legal implications, it is not reasonable to use such diagnostic methods for medico-legal purposes, when their diagnostic significance, to say the least, is questionable and one does not know for sure what the methods are actually measuring.

ENZYME HISTOCHEMISTRY

After Rutenburg *et al.* (1953) had found large amounts of succinate dehydrogenase in the myocardium of various species using blue tetrazolium as hydrogen acceptor, this method was also applied to autopsy material and the inactivation of histochemically demonstrable succinate dehydrogenase was observed in human myocardial infarction. The enzyme histochemical methods became soon widely applied in animal experiments and also to diagnose early human myocardial infarction in autopsy material. Various oxidative enzymes were demonstrated either by incubating thick myocardial slices in a medium containing a tetrazolium salt with or without exogenous substrate and coenzyme (macromethods) or to stain cryosections observed under the microscope (micromethods). It was assumed that enzyme histochemistry might be, at least, a 'semi-quantitative' indicator of the actual biochemical activity of a given enzyme and, therefore, a scientifically more solid biochemical indicator of myocardial metabolism and injury than the previously discussed traditional histochemical methods. In all, the results between individual authors and laboratories have been extremely variable, suggesting that the situation is not that simple. Hiltunen *et al.* (1985) assessed the enzyme histochemical methods, used by the author (PS) in autopsy material, comparing the enzyme histochemical reaction with the actual biochemical acitivity of the enzymes in a global ischaemia model in an animal experiment. The results pointed out that, although the enzyme histochemical methods have been named after the enzymes, the enzyme histochemical reactivity need not necessarily correlate with the actual biochemical activity of the enzyme at all, suggesting that other factors, such as the presence of auxiliary enzyme systems or other co-factors of the histochemical reaction, may be rate-limiting. This brings about the same dilemma as with the conventional histochemistry: unless the biochemical basis of any given enzyme

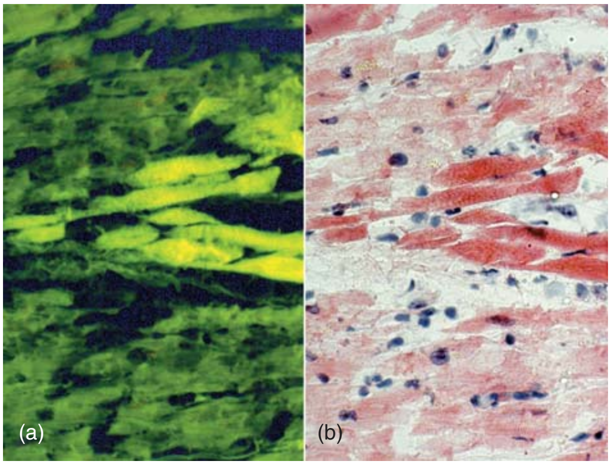

FIGURE 25.7 *Bright yellowish eosin fluorescence of damaged myofibres (a) corresponding exactly with eosinophilia as seen in the HE-stained cryosection (b) of a fresh myocardial infarction.*

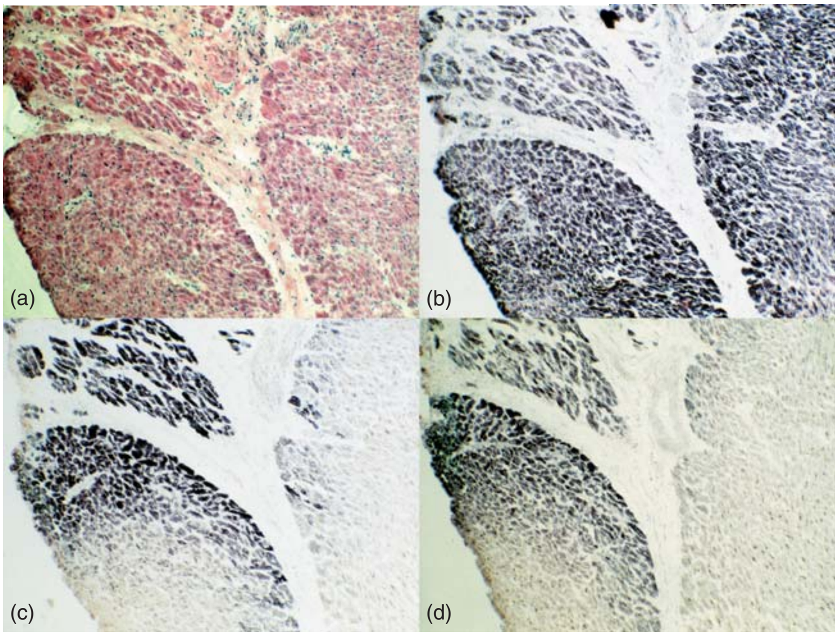

FIGURE 25.8 *The size of a myocardial injury depends upon the method used to display it. These are four serial sections of a heart from autopsy, stained in different ways. Haematoxylin and eosin stained cryosection (a) does not reveal any damage; the adjacent section (b) is stained for malate dehydrogenase activity, and shows strong and uniform reaction. Section (c) is stained for succinic dehydrogenase and reveals a varying degree of enzyme loss in almost all the section, the section (d) on the right stained for β-hydroxybutyrate dehydrogenase activity reveals slightly larger area of damage. (Original magnification ×100.)*

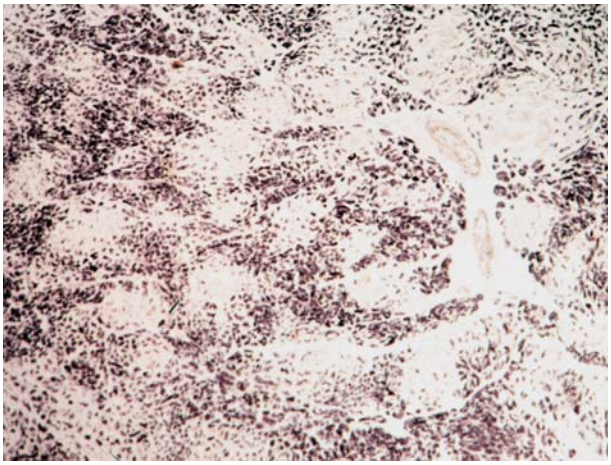

FIGURE 25.9 *Patchy loss of malate dehydrogenase reaction simulating early injury. In fact, the injury was caused by artificially prolonged agonal period due to respiratory treatment and cardiac pacing. It was equally well visible in HE staining emphasizing the necessity of control staining with conventional methods always when more sensitive methods are used.*

histochemical method is thoroughly understood, it is precarious to use it for medico-legal purposes.

IMMUNOHISTOCHEMICAL METHODS

The advances in immunohistochemistry and good results in clinical pathology as well as the availability of a great number of new antibodies have resulted in numerous publications reporting on potential markers of early myocardial injury in both animal experiments and autopsy material. As it has been shown that ischaemia damages cell membrane, contractile proteins, cytoskeleton and subcellular organelles of the myocardium relatively quickly, the search for better diagnostic methods has been directed to various components of the myofibre, e.g. basement membrane (collagen IV, fibronectin, laminin), cytoskeletal proteins (actin, desmin, alpha- and beta-tubulin), cell-matrix focal adhesion molecules (vinculin, talin), membrane-associated proteins (dystrophin, spectrin), terminal complement complex (C5b-9) and fatty acid binding protein (FABP), among many others. In spite of the plethora of publications, the results have been rather contradictory and/or their practicability difficult to assess due to shortcomings in test material or design.

In spite of better insights into the pathophysiology of myocardial injury, the practical problem as to its diagnostic significance as a cause of death has not changed: whenever myocardial injury has been detected, either macroscopically or by using whatever conventional or more sophisticated methods, **it does not necessarily prove that the injury is the cause of death, unless other causes of death have been excluded**. The rule-of-thumb is that the more sensitive methods are used the greater the probability that agonal period and autolytic changes may be difficult to differentiate from intravital changes.

COMPLICATIONS OF MYOCARDIAL INFARCTION

Ruptured heart

Ruptured heart is the most common cause of a haemopericardium and cardiac tamponade, the rupture always occurring through an infarct. The softened, necrotic muscle gives way from the internal pressure of the ventricular blood during systole, there being no equalizing rise in external pressure. Hypertension will increase the risk, but a more potent factor is a senile, soft myocardium, so that the elderly woman is a common victim of a ruptured heart. This by no means excludes younger men if the infarct is extensive and transmural. The most common area for rupture is the more distal part of the free wall of the left ventricle. The septum occasionally ruptures and the consequent left–right shunt, whilst the patient survives, provides a classical diagnostic sign for the stethoscopes of clinicians.

The rupture does not take place in the early stages of a new infarct, but after a day or two when necrotic softening is well established. The blood usually tracks through tortuous channels between muscle bundles, rather than bursting a direct fistula from ventricle to pericardial sac. The infarcted area may not always be obvious, as the haemorrhagic patch may obscure it, but histologically the ragged tissues and the periphery may be seen to be necrosed.

Haemopericardium is the pathological condition found at autopsy and is not quite synonymous with 'cardiac tamponade', which is a clinical state caused by the progressive accumulation of blood within the closed pericardial sac. As the external pressure rises, the heart cannot fully expand in diastole to allow filling from the great veins. As input volume falls, so does stroke output. The venous drainage is dammed back so that congestion and cyanosis of the face and neck occur, until a fatal endpoint is reached.

Mural thrombosis

Mural thrombosis almost inevitably occurs when an infarct is based upon the endocardium of the left ventricle. Though there is usually a thin layer of viable (though sick) cells immediately beneath the endocardium, they do not seem to prevent deposition of platelets and fibrin.

The thrombus is often entwined between the muscular bands lining the distal ventricle, which aids its adherence. When extensive, the thrombus may fill the entire apical area of the ventricle. The layers nearest the endocardium are naturally the oldest, and visible strata (lines of Zahn) are often visible. The youngest layers on the surface are less organized and more friable, so fragments tend to break off and form emboli that enter the arterial circulation, and can cause infarcts in kidney, brain, spleen and even the myocardium itself.

Pericarditis

Pericarditis occurs with full-thickness (transmural) infarcts that form the most common cause of pericarditis in Western communities. The visceral pericardium becomes purple–red as a result of the infarct beneath, with a vascular blush on the surface. Granular fibrin deposit causes the surface to lose its glistening sheen, and strands of fibrin may link the visceral and parietal layers together. When healed, fibrous adhesions may form which can obliterate part or even all of the pericardial sac.

Myocardial fibrosis

Myocardial fibrosis has already been described, healed infarcts being replaced by dense collagen, as new muscle cannot be formed. Large plaques from healed infarcts can be anywhere in the left ventricle, depending on the vessel(s) that were occluded. The septum and the posterior wall are more common sites than the anterolateral wall, though the latter is by no means exempt. Often the fibrotic plaques are multiple or irregular, being scattered throughout the ventricular wall.

Diffuse fibrosis is common in elderly people, even when there is negligible coronary narrowing. This may be in part the result of ventricular hypertrophy in hypertension, which causes a relative ischaemia, but many old hearts with no enlargement (indeed, atrophy) reveal patchy fibrosis, either macroscopically or microscopically. In the latter case, the fibrosis is often perivascular in situation.

Cardiac aneurysms

Cardiac aneurysms occur where a large area of fibrosis replaces a previous transmural infarct, usually on the distal part of the free wall of the left ventricle. During systole, the cavity blood presses outwards onto the unsupported area and gradually herniates it into the pericardial sac. A saccular aneurysm develops and may undergo various changes, including calcification of the wall, adherence to the parietal pericardium, or filling with laminated thrombus. The aneurysm wall is tough and fibrous and, though lacking in elasticity, almost never ruptures – again contradicting some of the dicta of pathology textbooks.

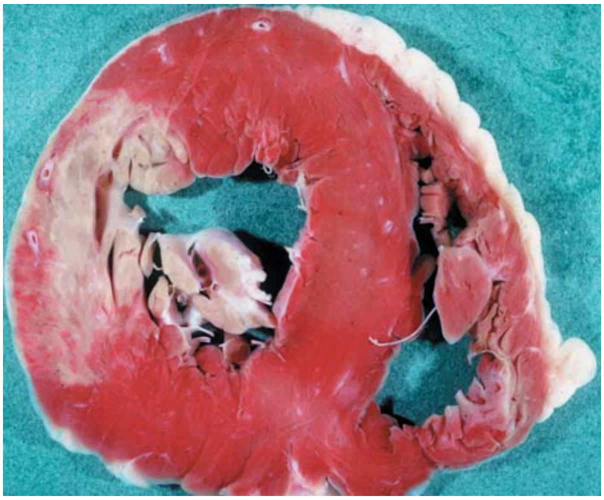

FIGURE 25.10 *An infarct of the lateral wall of the left ventricle of between 5 and 10 hours' duration. It has been rendered visible by staining the heart slice with triphenyltetrazolium chloride, which colours the normal myocardium red by the activity of tissue dehydrogenases. The infarcted area remains unstained. The infarct is virtually transmural, though some subepicardial fibres survive; the papillary muscle is particularly vulnerable, being at the distal end of the coronary supply.*

CAUSE OF DEATH IN CORONARY INSUFFICIENCY

Most sudden deaths from coronary insufficiency do not have myocardial infarction, even when the most sophisticated techniques are employed for its detection. Neither do the majority have a coronary thrombosis, though severe coronary stenosis is by definition present.

The rare case of myocardial infarction with only moderate or even minimal coronary atheroma has to be relegated to the realms of mystery in the present state of knowledge, unless some embolic cause can be found. Some pathologists invoke 'spasm' of the coronary arteries when the vessels show minimal atheromatous stenosis, but this is a hazardous diagnosis for a pathologist to make. Though spasm is certainly seen by clinicians when performing cardiac catheterizations or cardiac operations, it can never be a morphological diagnosis at autopsy, for obvious reasons.

In most sudden 'coronary' deaths, the usual autopsy finding is of severe, long-standing stenosis sometimes with foci of complete occlusion from a ruptured plaque or a subintimal haemorrhage. There may often be partial thrombosis, in that a plaque or narrow segment may have mural thrombus that further narrows the lumen, though does not totally occlude it.

There is macroscopic and microscopic evidence of myocardial fibrosis in many of these cases. This may either be a large plaque of scar tissue at the site of a previous focal infarct or more diffuse fibrosis. There may be fibrosis subendocardially and in the papillary muscles. In many of these sudden deaths there is no naked-eye or even microscopic evidence of any new lesion in the coronary system or myocardium, so why do they die?

Death must be ascribed to a rhythmical defect, rather than loss of mechanical pumping power. Myocardium and especially the pacemaking and conducting system, is vulnerable to ischaemia and hypoxia. Disturbances of rhythm, from ectopic beats to atrial fibrillation through to ventricular fibrillation and cardiac arrest, may occur, as well as heart block and the many other clinical manifestations of ischaemic heart disease from whatever cause. In addition, many victims are hypertensive and left ventricular hypertrophy exacerbates their relative coronary insufficiency. Anyone with a heart weighing more than 450 g is a candidate for sudden death, with or without coronary artery stenosis. The mechanism of this sudden type of death, where there is no large infarct, seems to be a state of electrical instability from chronic hypoxia, so that sudden stresses (such as exercise or emotion) can suddenly cause the arrhythmias mentioned above, though it must be said that death can occur without any of these factors while the victim is asleep or at rest. It is known that the myocardium is sensitive to catecholamines, this being a physiological response to intrinsic adrenergic hormones. Excess catecholamines given parenterally can cause arrhythmias and even contraction-band damage to the myofibrils. Certain other drugs, such as organic solvents, can sensitize the myocardium to even physiological amounts of noradrenaline, leading to ventricular fibrillation and sudden death.

Those with fibrosis in the myocardium may have the additional risk of interruption of the conducting system, which may give some degree of heart block or defect in the spread of the contraction impulse, rendering them more prone to arrhythmias and arrest. This is especially so where a large fibrotic plaque occupies the interventricular septum, as here the left bundle branch penetrates to supply the apical region and may be strangled by dense fibrosis.

Where there is a recent coronary thrombosis, the absence of an infarct may mean that insufficient time has elapsed for it to become apparent, but of course the pumping efficiency will be compromised and the risks of electrical disturbance will already be operating, so that all the problems listed above exist. The other reason for lack of an infarct is an adequate collateral circulation but, even though necrosis is avoided, a new coronary occlusion can only further worsen the ischaemic and hypoxic state of the myocardium, with all the potential consequences of electrical instability.

If there is an infarct, then again all the above dangers exist, but there is the added liability of a loss of mechanical pumping function, which, with a large infarct, may be a substantial proportion of the whole cardiac output, leading to overt low-output failure during which sudden death can supervene at any time. In addition, the risks of sequelae of myocardial infarction, such as rupture and embolism, are ever present.

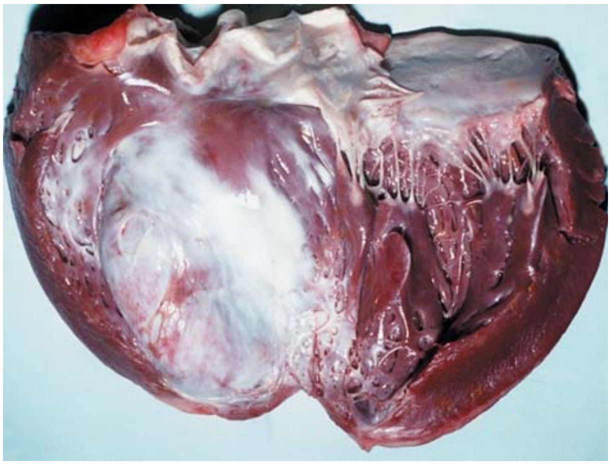

FIGURE 25.11 *Extensive myocardial fibrosis occupying the interventricular septum and apex of the heart with severe occlusive disease of the anterior descending branch of the left coronary artery. Near the apex all the muscle has been replaced by scar tissue. This type of lesion is particularly prone to cause sudden death, as well as clinically apparent dysrhythmias, as the septal fibrosis interrupts the left bundle branch of the cardiac conducting system.*

The role of the conducting system in sudden cardiac deaths is one which is attracting increasing interest, especially as new histological and immunohistochemical techniques are becoming available. A revised sampling technique, described by Song *et al.* (1997), uses longitudinal sectioning, reduces the workload and allows observation of continuity between different components of the cardiac conduction tissue. It demonstrates the sino-atrial (SA) node, the atrio-ventricular (AV) node and the distal part of the His bundle and the bundle branches of the cardiac conduction system in 4–5 blocks. However, although a significant number of abnormalities can be recognized, it is usually difficult or impossible to relate morphology to function. The fact that an AV or SA node, or the bundle of His or its branches shows fibrosis or some other lesion, does not necessarily mean that this played any part in the death. It has been shown that the conducting system is remodelled with advancing age – and many deaths from obvious extracardiac causes reveal lesions in the conducting system when studied at autopsy. Nevertheless, new knowledge now being accumulated about the conducting system may become useful in investigated sudden cardiac deaths, as it is obvious that a high proportion of these fatalities, which have no recent acute lesion demonstrable in the coronary arteries or myocardium, are due to arrhythmias and functional arrest, in which abnormalities of the propagation of the pacing impulse may be important. James (1996) has suggested that apoptosis, the programmed cell death, might be a logical explanation and responsible for histological abnormalities of the cardiac conduction system in sudden deaths due to progressive development of complete heart block and various arrhythmias.

HYPERTENSIVE HEART DISEASE

Hypertension may kill in a number of ways, such as by renal failure, ruptured aneurysm or cerebral haemorrhage, but here we are concerned with primary heart failure, a quite common cause of death. During life, the clinical syndrome of 'cardiac asthma' or 'paraoxysmal nocturnal dyspnoea' is caused by pulmonary oedema from hypertensive left ventricular failure. The same oedema is a marked feature of the autopsy on fatal hypertensive heart disease.

Because there is an aetiological connection between hypertension and coronary atheroma, the two conditions are often present together in a victim of sudden death, making it difficult to separate the myocardial consequences of each component. Sufficient 'pure' deaths from hypertensive heart disease are available where there is no concurrent coronary disease, however, to establish the effect of the former on the myocardium. When the left ventricle has to work against a higher pressure in the systemic arteries, the muscle fibres hypertrophy. They cannot increase in number, but they increase in length and thickness, with irregularity and enlargement of nuclei. This fibre enlargement leads to increase in the mass of the left ventricle, the well-known 'concentric hypertrophy' of hypertension.

If 360–380 g is taken as the upper limit of heart weight for an average size man, then hypertensive disease may produce hearts of 500–700 g. Hearts larger than this usually have some other cause, such as valvular disease or a cardiomyopathy. Though it is sometimes denied that hypertension can cause sudden death, most forensic pathologists would strongly disagree, all having performed numerous autopsies where no other reasonable explanation existed. The usual picture is of marked pulmonary oedema fluid running freely from the cut surfaces of the lungs because of a terminal failure of the left ventricle slightly prior to right ventricular arrest. This causes a rapid rise in pulmonary artery pressure and consequent transudation across the pulmonary alveolar membranes.

There may not always be evidence of hypertension in the often sparse history available at sudden death, but the presence of concentric hypertrophy in the absence of valve disease or a cardiomyopathy – together with characteristic changes in other vessels and organs, such as the kidney – is strong evidence of raised blood pressure. Sometimes the overall weight of the heart may be normal, yet there is relative left ventricular thickening; this again suggests hypertension. In the hearts of aged people, the usual atrophy – with tortuous surface vessels – may be concealed by a left ventricular hypertrophy, keeping the heart at normal weight.

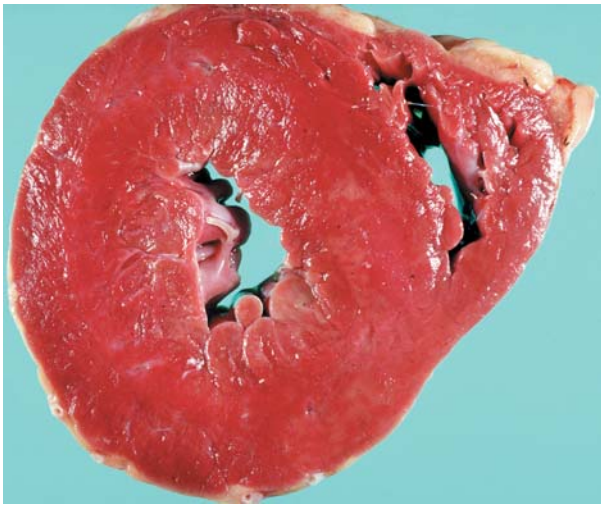

FIGURE 25.12 *Concentric hypertrophy of the left ventricle in a person with cardiomegaly from hypertension. The slice has been stained with triphenyltetrazolium chloride to reveal dehydrogenase activity; some mottled areas, especially in the septum, show pallor from reduced enzyme activity as a result of relative ischaemia of the increased muscle mass.*

The existence of hypertensive heart disease as a specific entity was strengthened by some investigations published by the author (BK) in 1973, in which it was shown that gross enzyme deficiencies exist in the inner part of the wall of the hypertrophied left ventricle. This lesion is not specific to hypertension, but is seen whenever the left ventricle is enlarged, notably in aortic valve disease. In severe instances, the defect in dehydrogenase activity extends from endocardium almost to the epicardium, there being only a narrow zone of normal activity. In some ways this defect is similar to the 'laminar infarct' of coronary artery disease.

In pure hypertension and aortic valve disease, the cause is a relative ischaemia of the inner zone of the ventricular wall caused by insufficiency of coronary blood supply compared with the demands of the thickened ventricular mass. The reason for the laminar distribution is the anatomy of the coronary supply, graphically described by Farrer-Brown (1968, 1977). The various levels or strata of the ventricular wall have different arterial patterns. The subendocardial zone is supplied by arteries with many branching terminations that provide a profuse and concentrated blood supply. A second set of vessels passes straight through the ventricular wall without division – all the coronary supply coming, of course, from the epicardial surface. This set serves the papillary muscles and trabeculae carnae, while a third group tends to run circumferentially parallel and below the epicardium, giving branches both outwards and inwards.

Farrer-Brown found that the termination of these vessels were of two types, as they divided into arterioles and then capillaries. The direction of the tree-like branching in the subendocardial zone continued in the general line of the main artery. The area of myocardium supplied by these branches was small. This was in contrast to the pattern in the mid-zone of the wall, where the branches turned at an angle to the main artery and then divided to supply a much wider area of myocardium that was supplied by the terminal branch in the subendocardial zone. These anatomical differences suggest that the vessels in the mid-zone each have to supply a larger volume of muscle than in the other layers. This might well explain the greater vulnerability of this central zone to hypoxic changes, as indicated by reduced enzyme activity and the occurrence of laminar necrosis. In actual practice, the inner zone also suffers, apart from the immediate subendocardial layers, but the superficial subepicardial zone is almost always spared. The rationale of damage to the myocardium in hypertensive heart disease, as in aortic stenosis, would appear to be a relative coronary insufficiency. In pure coronary artery disease, part of a normal-sized myocardial mass is rendered ischaemic by localized reduction of blood flow by a coronary stenosis or occlusion, the resulting necrosis being distributed in a focal pattern according to the location of the block.

In hypertensive heart disease, normal coronary arteries attempt to supply a much larger muscle mass, so that there is the same imbalance between supply and demand. As there is no focal blockage, however, the resulting lesion is diffuse and the areas that suffer most are those in the mid- and inner zones, probably because of the microanatomy of the vessels described above. The same type of lesion may be seen in any severe sustained hypotensive episode of extracardiac cause, though the early onset of death in many of these will preclude the development of morphological evidence of laminar myocardial damage.

SUDDEN DEATH IN AORTIC VALVE DISEASE

A similar situation exists with aortic valve disease in which stenosis rather than incompetence can lead to sudden death. Most of the lesions are primarily degenerative, the commonest being ‘idiopathic calcific aortic stenosis’, most often seen in elderly men. Aortic stenosis of rheumatic origin always affects the mitral valve before the aortic, so a solitary aortic lesion cannot be ascribed to rheumatism and is therefore of the primary degenerative type. Rheumatic valve disease is becoming rare, both with the decline in rheumatic fever and the frequency with which valve lesions are treated surgically at an earlier stage, before marked ventricular hypertrophy and its myocardial lesions are evident.

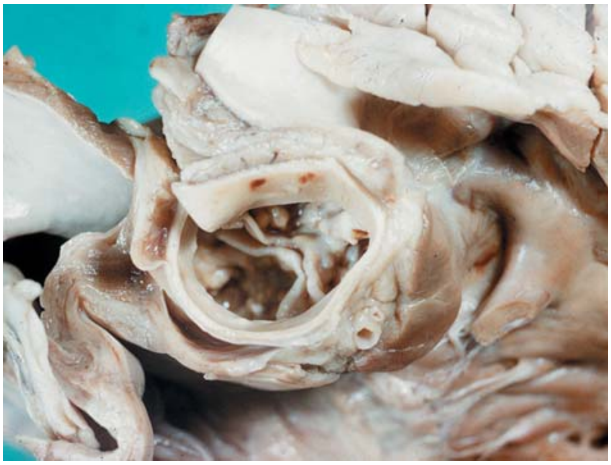

FIGURE 25.13 *Calcific aortic stenosis, a common cause of sudden unexpected death especially in older men. Though the stenosed valves are often biscuspid, the stenosis often deforms normal valves, as in this illustration. Marked left ventricular hypertrophy usually ensues, the large muscle mass becoming ischaemic, especially in the inner layers.*

In the common calcific disease, the valve is thickened and rigid, with fusion of the commissures in most cases. There may be large, irregular excrescences on the cusps and in the sinuses behind. At a later stage, the whole valve may be an almost unrecognizable, chalky mass, with a lumen barely wide enough to admit a pencil. This type of valve may have been biscuspid from birth, a condition that undoubtedly encourages calcific degeneration. Many stenotic and apparently bicuspid valves are not truly bicuspid, however, but appear so because the degenerative process has destroyed or obliterated one of the commissures between the cusps. This may often be demonstrated by the fact that a trace of the commissure may still be seen, or by the fact that the two surviving cusps are not equal in size, the commissures being placed at 120° rather than 180°.

A tight aortic valve, from whatever cause, obstructs the outflow tract of the left ventricle and causes the muscle to hypertrophy in order to eject the same stroke volume through a narrower orifice. In addition, if there is associated regurgitation, further extra work has to be performed to attempt to throw out the refluxed blood.

The effect of a severe aortic stenosis in relation to sudden death is to enlarge the left ventricle to sizes even greater than those seen in hypertension. Some of the biggest hearts (excepting some cardiomyopathies) are seen in aortic valve disease, going up to 800 g or – exceptionally – 1000 g.

Another effect of a tight valve is to lower the perfusion pressure in the coronary arteries, which is made worse if there is also an element of regurgitation. It is the diastolic pressure in the proximal aorta that feeds the coronary supply, as during systole the contraction of the heart prevents flow in the intramural arteries. A less common effect may be the distortion of the coronary ostia by calcific masses in the sinuses of Valsalva and the wall of the root of the aorta, which can partly block the entry into the coronary vessels. The combination of a large myocardial mass and decreased coronary flow is often made worse by concomitant coronary atherosclerosis, so the propensity of victims of aortic valve disease to drop dead is considerable. Elderly men are the most frequent victims. Lesions of other heart valves seldom cause sudden death as an isolated cause, though the more chronic cardiac dysfunction – mainly congestive cardiac failure – caused by them may end life rather abruptly on occasions. Rheumatic mitral stenosis is one such lesion, but perhaps a more direct mitral cause of sudden death is the 'floppy' valve usually seen in old age.

SUDDEN DEATH FROM THE CARDIOMYOPATHIES

There is a heterogenous group of diseases of myocardium associated with cardiac dysfunction known as the 'cardiomyopathies', which, though of mixed and often uncertain aetiology, have sufficient in common pathologically to merit a special identity. As they account for the second largest number of sudden deaths after coronary artery disease, i.e. about 10–15 per cent of all sudden deaths of cardiac origin, they are briefly described here, though specialized texts and papers on cardiac pathology should be consulted for details. The outstanding feature is a large heart in the absence of hypertension or valve lesions. The victims are usually young adults, partly because in older groups in which there is atherosclerotic coronary disease and also hypertension, these overlay the cardiomyopathy and make the diagnosis difficult or impossible. The only exception is if there is definite histological evidence or if the heart weight is excessive (over 700 g) in the absence of a valve defect. Cardiomyopathies are the most common cause of heart failure and are an important cause of death in children and adults. According to the Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the Definition and Classification of Cardiomyopathies, they are classified into the following main types: dilated, hypertrophic, restrictive and arrhythmogenic right ventricular cardiomyopathy.

Dilated cardiomyopathy (DCM) is the most common cause of congestive heart failure. The annual incidence varies, depending on the diagnostic criteria, from two to eight cases per 100 000. Around 30–40 per cent of them are familial (FDCM), but it may also be idiopathic, viral, immune or toxic, e.g. alcoholic. Idiopathic dilated cardiomyopathy (IDCM) is characterized by dilated ventricular chamber and reduced contractility without coronary, valvular or pericardial disease. Histology is non-specific; however, viral genome has been identified in 10–34 per cent of patients with dilated cardiomyopathy.

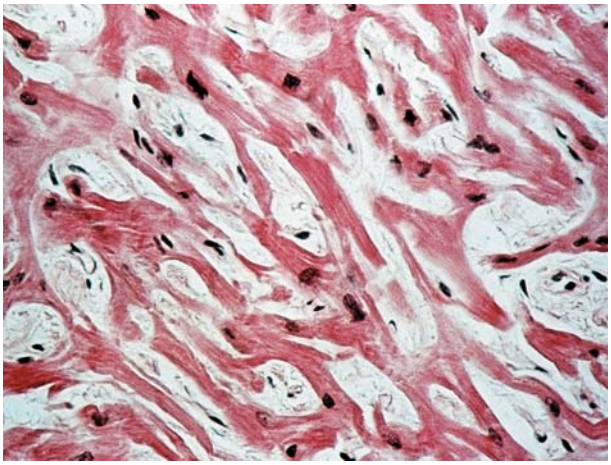

FIGURE 25.14 *Myofibre disarray in hypertrophic cardiomyopathy.*

Hypertrophic (obstructive) cardiomyopathy was first described in medicolegal autopsy material by a London forensic pathologist, Donald Teare, in 1958. Although he noted and described asymmetrical hypertrophy of the interventricular septum, so that a bulging pad of myocardium partly obstructed the left ventricular outflow tract, this cardiomyopathy is predominantly (75 per cent) a non-obstructive disease and, hence, hypertrophic cardiomyopathy (HCM) is now the preferred name. It is a primary sarcomere disorder characterized by left and/or right ventricular hypertrophy and, in most cases, inherited as a predominantly autosomal disease. It is caused by mutations of any of the genes encoding the proteins of the cardiac sarcomere.

Between 20 and 30 per cent are familial, and the condition may occur at any age from infancy to senility, though the majority of deaths occur in early middle age. The weight of the heart may range from a high normal at 400 g, to 800 g or even a kilogram. More striking than the total weight is the left ventricular preponderance, which is often maximum in the proximal septum, as described above.

Histologically the myocardium shows irregular-sized, large fibres with nuclear pleomorphism. Though once thought to be a rhabdomyoma or a hamartoma, or aberrant adrenergic tissue, the septal bulge usually shows a bizarre pattern of fibres, with loss of the orderly strata and replacement by a criss-crossing, confused architecture of muscle bundles. Interstitial fibrosis is common, and – though there may be an increase in interstitial cellularity – inflammatory infiltrate is not seen, as it would be in an active myocarditis.

Restrictive cardiomyopathy is, according to the WHO report, 'characterized by restrictive filling and reduced diastolic volume of either or both ventricles with normal or near-normal systolic function and wall thickness. Increased interstitial fibrosis may be present'. It may be idiopathic or associated with other diseases such as cardiac amyloidosis, cardiac haemochromatosis or endocardial fibroelastosis with or without hypereosinophilia.

Arrhythmogenic right ventricular cardiomyopathy (ARVC) is a primary myocardial disorder characterized by progressive fibrofatty degeneration of right ventricular myocardium, arrhythmias and risk of sudden death. In addition to familial autosomal-dominant disease with variable penetrance and polymorphic phenotype, two autosomal-recessive forms of ARVC are known.

The WHO classification includes further two groups of cardiomyopathies:

- specific cardiomyopathies that are associated with specific cardiac or systemic (ischaemic, valvular, hypertensive, inflammatory, metabolic, general systemic e.g. connective tissue or neuromuscular disorders) and
- unclassified cardiomyopathies, which do not fit readily in any group, e.g. fibroelastosis or isolated non-compaction of the ventricular myocardium (INVM), a rare disorder characterized by excessively prominent trabecular meshwork, ventricular arrhythmia and systemic embolism.

Myocarditis

Many infective diseases produce an acute myocarditis, which may be the immediate cause of death. A prime example is diphtheria, but this has virtually no forensic relevance. More important is 'isolated myocarditis', where the condition is primary and usually of unknown aetiology. It was formerly [and some recent results from an ongoing study by the author (PS) suggest, that it still is] a much-used and overused autopsy diagnosis, especially where no gross lesions could be found. Known by a variety of names, such as 'Fiedler's' or 'Saphir's' myocarditis, its incidence was said to be appreciable in victims of sudden death – and it was incriminated as the cause of accidents on the road and in the air that had been attributed to driver or pilot incapacity. In forensic pathology some 20–30 years ago, it became a 'fashionable' diagnosis on rather equivocal histological criteria. It was then found that many victims of trauma where death was rapid and where cardiac incapacity could not possibly be a causative factor in the accident, also had foci of mononuclear cells in the myocardium.

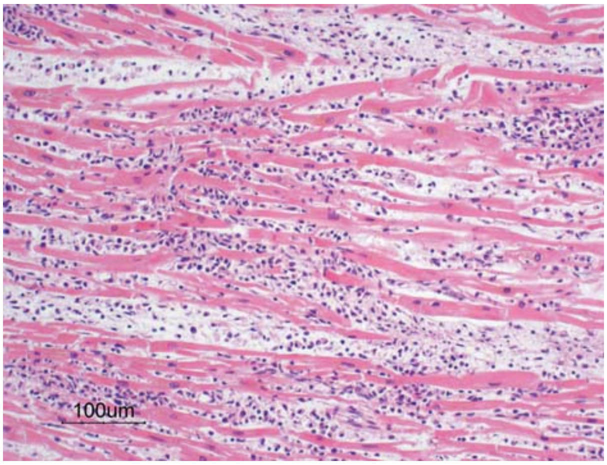

FIGURE 25.15 *Severe diffuse lymphocytic myocarditis with destruction of myocytes. (HE, original magnification ×10.)*

Stevens and Ground (1970), after finding six cases of myocarditis in 263 air pilots, when none of them could have been affected by a cardiac event, looked at a considerable number of other accident victims in circumstances in which disease was obviously unrelated to the trauma. They found about 5 per cent of young men had lesions otherwise histologically acceptable as 'myocarditis' and concluded that it was an unsafe diagnosis to use for the cause of death. Their histological minimum criteria for acceptance was one focus of 100 or more cells or many foci of half that size. Groups of lymphocytes, especially subepicardial of 40 or fewer cells, were disregarded.

In Australia, Tongue *et al.* (1972, 1977) did similar studies, but used whole transverse sections of ventricles at up to 18 levels in each heart. Their results showed that the more extensive the search, the greater the number of foci discovered in deaths that had no circumstantial relation to a sudden cardiac death. Using 18 full sections, they found that 90 per cent of the accident victims had 'myocarditis' lesions and suggested that, if the search was extensive enough, it could probably be brought up to 100 per cent.

The Dallas Classification System was proposed by eight cardiac pathologists in an attempt to provide the practicing pathologists with easily applicable and uniform criteria for the diagnosis of myocarditis in endomyocardial biopsies (Aretz *et al.* 1987). By this definition, the diagnosis of myocarditis can be made only if myocyte necrosis or degeneration or both are associated with an inflammatory infiltrate adjacent to the degenerating or necrotic myocytes. A semi-quantitation of the amount of inflammatory infiltrate was suggested as mild, moderate, or severe and its distribution characterized as focal, confluent, or diffuse. The composition of inflammatory infiltrate, i.e. the presence of lymphocytes, neutrophils, eosinophils or giant cells was suggested for differential diagnosis of the aetiology. According to Feldman and McNamara (2000) the Dallas criteria probably underestimate the true incidence of myocarditis.

Among the great variety of infectious agents, viruses have been considered as important cause of myocarditis. The molecular biological techniques such as polymerase chain reaction (PCR) and *in situ* hybridization have provided new diagnostic possibilities and viral genome has been detected both in myocardial biopsies as well as in autopsy material. However, the detection rate and the viruses identified have varied markedly according to the study. In addition, drugs may cause myocarditis, either through direct toxic or immune-mediated mechanism.

Probably, the position now to be accepted is that unless multiple, florid foci are found, with myofibril necrosis seen, using special stains if necessary, bland mononuclear foci that are purely interstitial and not involving muscle fibres must be disregarded. The presence of myofibre necrosis, however localized, would appear the most important confirmatory criterion. Where stenotic coronary artery disease exists, it can be difficult or impossible to differentiate the resulting cellularity, fibrosis and fibre necrosis from that of a myocarditis.

A disease process that may well have been underestimated in the pathology of sudden death is cardiac sarcoidosis. Unless routine histology is carried out on every sudden death, this diagnosis will be missed (Fleming 1988) as will a considerable number of myocardial amyloidoses, which need special stains.

DEATH IN OLD AGE – THE SENILE MYOCARDIUM

In many autopsies on old persons, no specific lesions can be discovered that provide a clear cause of death. In the increasingly aged population of the developed nations, this problem becomes a more frequent task for the pathologist. The history of the mode of death then becomes particularly important, as it may give some clue as to which area of the indeterminate lesions should be given priority in deciding on a cause of death. For instance, if there is moderate coronary artery disease, then if the person died with sudden breathlessness or chest pain, rather than slowly fading away, more credence would be given to the coronary lesions. Many old people have good, patent coronary arteries, however, even if the walls may be calcified – if they did not have such good vessels, they probably would not have lived to 88 or 95 years, anyway.

Where nothing other than general senile atrophy of most organs is found and the history is unhelpful as to a specific mode of death, as long as the pathologist can exclude any unnatural cause, then it is quite legitimate to ascribe the death to ‘myocardial degeneration due to senility’. The WHO’s ICD-10 (International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Geneva 1992) provides a code for such a disease even though many cardiac pathologists seem reluctant to agree to its existence, wishing always to find some more specific disease process. In the authors’ view, senile myocardial degeneration is quite acceptable, given exclusion of other causes. Senile hearts are usually small, but hypertension may have enlarged the ventricles long ago and the residual effect of this may be to sustain the heart at a normal weight. More important is the appearance of the heart, which is brown on the surface and in the muscle and is flabby and soft. The thumb can be pushed into such muscle at autopsy without difficulty. The weight may be down to 300 g or even 250 g, if no previous hypertrophy has existed. A good indication of cardiac atrophy is the presence of tortuous coronary vessels on the epicardial surface, especially the anterior descending branches. These may meander markedly, as the myocardium has shrunk, the apex has contracted nearer the base and so the vessels have to corrugate to fit into the diminished space.

Microscopically, the myocardial fibres are uneven in size, there may be a fine, diffuse fibrosis, which is not ischaemic in origin and the nuclei of the myocytes have prominent clumps of lipofuscin pigment at their poles. Unfortunately, none of these features is diagnostic of the cause of death, as other old persons dying of quite unrelated causes, such as trauma, might have equally poor heart muscle.

The justification for considering that senile myocardial degeneration is a valid cause of death comes from the species life span of different animals. Even without lethal specific disease, each species has a fairly uniform life span, the human being some 70–90 years. No one lives to 150 years, so there must be some limiting factor; as cardiac function is the most immediate arbiter of survival, it seems reasonable to indict heart contractibility as the primary factor. All tissues age, but senility of the spleen or thumbs does not have the immediacy for life support possessed by the myocardium.

SUDDEN DEATH FROM RUPTURE OF AN ANEURYSM

The most frequent extracardiac cause for sudden death arising in the cardiovascular system is a rupture of an aneurysm, almost always of the aorta or a cerebral vessel. Aortic aneurysms are of three types, all capable of catastrophic rupture.

The atheromatous aneurysm

The atheromatous aneurysm is the most common and is seen mainly in the abdominal segment of the aorta. Though atherosclerosis affects the whole length of the aorta, it is usually worse below the diaphragm, probably because of the haemodynamic effects of the larger branches that come off in this segment. Turbulence and 'Venturi' effects are known to localize the development of fatty streaks and atheromatous plaques, as they are seen first around the ostia of the intercostal vessels and become particularly severe at the major bifurcation into common iliac arteries. Though aneurysms can develop at any point in the aorta, most are below the diaphragm. Excepting dissecting aneurysms, however, any aneurysm in the thoracic segment is still more likely to be atheromatous than syphilitic, even though syphilis is virtually confined to the aortic arch. Most atheromatous aneurysms are fusiform or saccular, the bulge being either symmetrical about the axis of the vessel or more commonly bulging out more on one side than the other. Progressive destruction of the media by the fibrolipid degenerative processes of atherosclerosis leads to weakening of the aortic wall. Bulging caused by the continuous internal blood pressure begins, aided by the hypertensive tendencies of later life. The damage or complete destruction of the intima leads to platelet and fibrin deposition, so that mural thrombosis is laid down in the expanding aneurysm. This may partly or wholly fill the sac, being many centimetres thick in some instances, with well-marked stratification called the 'lines of Zahn'. The wall of the aneurysm may contain calcific fragments similar to the adjacent artery wall.

Most aneurysms remain intact throughout life and are found as incidental findings at autopsy. Naturally the incidence of ruptured aneurysms is greater in medico-legal autopsies by reason of the selective population of sudden deaths. When rupture occurs, the weakest point of the wall is penetrated by the contained blood and this usually leaks out into the retroperitoneal tissues, rather than free bleeding into the peritoneal cavity. The blood tracks behind the root of the mesentery and around the kidney. A large perirenal haematoma may form but, more often, the obvious site of bleeding is in the centre of the back of the abdomen. The diagnosis is usually quite obvious as soon as the autopsy begins for, as soon as the abdomen is opened and the intestine moved aside, a dark red mound may be seen projecting forwards over the lumbar spine, bulging at the root of the mesentery.

The development of a successful technique for the repair of many such arterial defects has reduced the death rate, though there is a substantial perioperative mortality, often because the patient is virtually moribund when the operation begins, or because gross calcification and degeneration of the remaining aortic wall makes it technically impracticable.

The dissecting aneurysm

Rupture of a dissecting aneurysm of the aorta is much less common than the atheromatous variety and seems to have decreased even more during the past decade. It is still by no means an uncommon cause of sudden death, however, and is the second most frequent cause of a haemopericardium and cardiac tamponade. Whereas atheromatous aneurysms are usually in the abdomen, the dissecting aneurysm can span the whole length of the vessel from iliac artery to aortic valve, though the main effects are usually manifest in the thoracic segment.

The basic lesion is a degeneration of the aortic media, the so-called 'medionecrosis'. The thick elastic media becomes degenerate and cystic in the central layers, producing a vessel wall like a sandwich with a soft cleavage plane in the middle. This defect is of unknown aetiology and affects both the elastic and muscular elements of the media. It is relatively common over the age of 50, especially in men. Many aortas seen at autopsy have this degeneration, but no aneurysms, as they never develop because the blood fails to enter the potential cleavage space from the lumen of the aorta.

In most cases where a dissecting aneurysm has caused death, the blood has broken through from the lumen through a tear in an atheromatous plaque, which is a quite separate disease process existing in parallel with the medionecrosis. When blood enters the medial cleft under high arterial pressure, the two layers of the media are split apart and the dissection may rapidly travel both upwards and downwards in the aortic wall. The upper component may split its way around the arch and beyond the attachment of the pericardium. Here, just above the aortic ring, the dissecting haemorrhage often bursts through the remaining outer media and adventitia into the pericardial sac. This causes a massive haemopericardium and the cardiac tamponade described earlier in relation to ruptured myocardial infarct. The distal extension may travel down to the iliac and even femoral arteries, sheathing them in blood. Rarely, in a patient who survives a dissection, the secondary lumen formed in the cleavage plane may rupture back into the main aortic lumen, thus forming a 'double-barrelled' aorta, a curiosity occasionally found at autopsy. Death usually occurs from cardiac tamponade, but some dissections do not burst into the pericardium. Here death is by a less obvious mechanism, which may be partly the sudden narrowing of the aortic lumen and pressure on the coronary ostia and roots of the coronary arteries. In persons under 50 years of age, there is a much less common alternative cause of medionecrosis, known as Erdheim's degeneration. This is sometimes called 'mucoid medial degeneration' and is associated with Marfan's syndrome, an inherited condition characterized by arachnodactyly and optic, aural and bony lesions.

The syphilitic aneurysm

Syphilitic aneurysms are now uncommon, because of the relative rarity of tertiary lesions as treatment is available for the

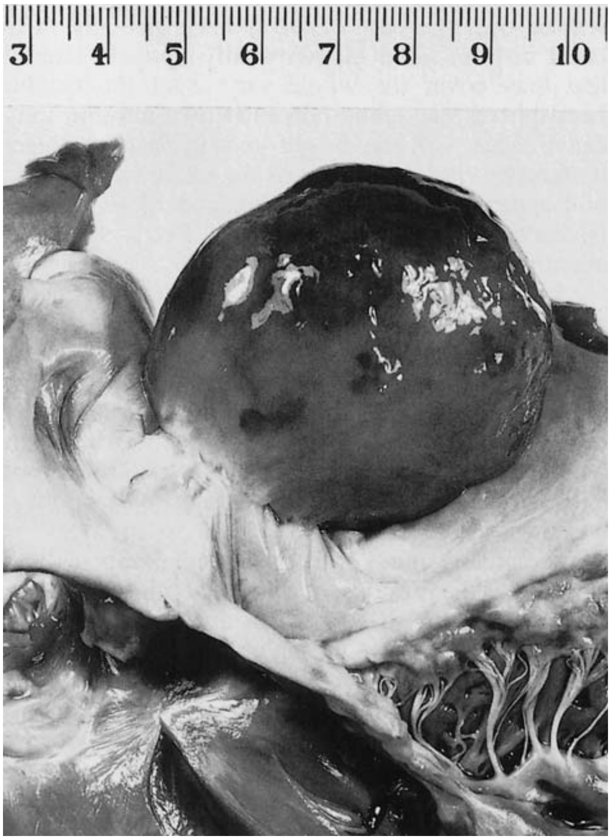

FIGURE 25.16 *Sudden unexpected death caused by a myxoma of the left atrium. The ball-shaped mucoid tumour impacted in the mitral valve during a postural change. No other cause of death was discernible.*

earlier stages of the disease. In some areas of the world, however, tertiary syphilis is not the rarity it has become in Europe and America.

The syphilitic aneurysm of the aorta is almost always in the thoracic segment and usually in the arch. Syphilitic aortitis, which precedes an aneurysm, may be recognized in the lining of the thoracic and abdominal aorta by its irregular corrugated pattern, though admixture with atheroma can obscure the classical appearances. There tend to be sinuous linear folds, like coarse wrinkles, which have been likened to the bark of an oak tree. If the aorta dissected free of adherent connective tissue is held up to a strong light, 'Gough's test' may be applied, in which translucent windows can be seen in the media. When confluent, these weak patches allow the internal blood pressure to blow out an aneurysm and the syphilitic bulge can be the largest of all, sometimes almost filling the upper part of the chest. The wall is thin and fragile, and there may be several separate bulges from overlapping aneurysms.

Pressure may erode bronchi, oesophagus, pulmonary vessels, lung and even the thoracic cage. The latter may allow the aneurysm to become subcutaneous and, in former years, the external bursting of a pulsating chest tumour must have been a dramatic event. The usual mode of rupture is into the pleural cavity or into the oesophagus or bronchus, giving rise to a sudden, massive haematemesis or haemoptysis. Histologically, it may be impossible to confirm syphilitic aortitis, as the reactive phase of the disease may have been long burnt out, but the gross pathological appearances are diagnostic in themselves.

Fatal aneurysms of other vessels

Fatal aneurysms of other vessels are rare, apart from the cerebral arteries. Atheromatous bulges can occur on the iliac and femoral vessels, and sometimes in the mesenteric arteries. Polyarteritis nodosa can produce mini-aneurysms from the inflammatory process eroding the wall, but death is not caused by the direct effects of rupture, but by vascular problems in the coronary, renal and other arteries. Infective mycotic aneurysms are rare in civilian practice, as are traumatic arteriovenous fistulae, but are not uncommon in war casualties.

Sudden death from ruptured cerebral aneurysm

The relationship of head and neck injury to subarachnoid haemorrhage is discussed in Chapter 5, but here we are concerned solely with spontaneous rupture of an aneurysm of the circle of Willis at the base of the brain. This is one of the most common causes of death in young to middle-aged adults, if coronary disease is excluded. In women, who are relatively immune from coronary occlusion up to the fifth decade of life, ruptured berry aneurysm is proportionately much more common than in men. It is a useful rule-of-thumb, when presented with the sudden death of a woman of child-bearing age, to consider a complication of pregnancy, self-poisoning, pulmonary embolism and subarachnoid haemorrhage as the first choices.

Subarachnoid bleeding can cause virtually instantaneous death, even though the mechanism is obscure. Numerous cases have been described where a previously fit person was seen to collapse and was apparently already dead when attended by onlookers. There must be an element of cardiac arrest in these examples, caused by the sudden bathing of the brainstem in blood from a jet of arterial blood impinging on the base of the brain. As with coronary disease, arterial spasm has been invoked as a cause of the sudden death in ruptured berry aneurysm, without any proof. Most cases of subarachnoid haemorrhage have a much longer course, with clinical symptoms and signs allowing either surgical intervention or, often, spontaneous resolution.

We are by definition here, however, concerned with sudden or rapid fatalities. In the usual coroner's case, the victim is either found dead with no available history, or has died rapidly and inexplicably, or has expired after suggestive symptoms like severe headache and rapid coma. Many die after physical or emotional exertion, especially sexual intercourse or strenuous sporting activity. Those that occur during or soon after some assault or altercation are dealt with in the chapter on head injuries (Chapter 5), and can pose a major forensic problem concerning causation.

At autopsy, the diagnosis of subarachnoid haemorrhage is self-evident. As the usual point of bleeding is in the circle of Willis, the most dense haemorrhage will be over the base of the brain, especially in the basal cisterns. The blood usually spreads laterally and may cover the whole surface of the cerebral hemispheres, the hindbrain and down into the spinal canal. This will be bright red in a fresh bleed; if survival lasts a week or so, a brownish tinge will appear as the haemoglobin undergoes changes. Haemosiderin can be detected by Perl's stain after about 3 days.

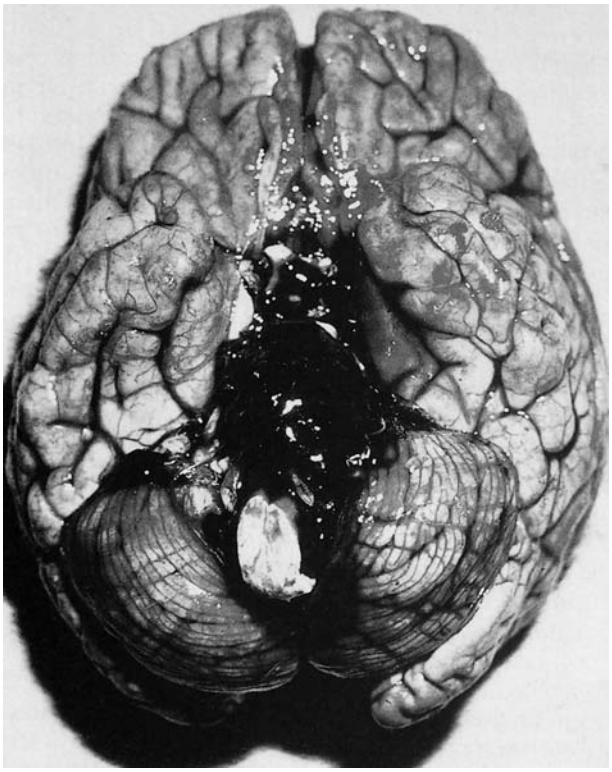

FIGURE 25.17 *A dense basal subarachnoid haemorrhage that caused rapid death. The blood clot filling the basal cisterns conceals a ruptured berry aneurysm of the circle of Willis; at autopsy, this must be removed under a stream of water, using careful blunt dissection to avoid further damage to the vessels.*

The detection of the source of the haemorrhage is sometimes difficult. An aneurysm is present in about 85 per cent of spontaneous subarachnoid haemorrhages, but the remainder reveal no aneurysm, even after an exhaustive search. This may be because of destruction of a small aneurysm at the time of rupture, but it is just as likely to be a leak from a vessel without an aneurysm, as a result of the existence of a point of weakness as described below.

A relatively rare cause of bleeding, especially in children, is the presence of a vascular malformation in the arterial system, such as an angioma, an arteriovenous anastomosis, or an aberrant vessel in the meninges. Post-mortem angiography may be of assistance in tracing intracranial bleeding from natural lesions, as in trauma. Karhunen *et al.* (1990) describe the techniques required.

Though berry aneurysms are often called 'congenital', they are not themselves present at birth, but a defect in the wall of a cerebral artery is probably present from late fetal life. The circle of Willis condenses out of an irregular embryonic meshwork of arteries over the base of the brain by unwanted vessels undergoing atrophy. Where these redundant vessels originally joined the surviving main arteries, a 'window' in the elastic and collagenous coat is left, usually at the junction of two vessels. This window is a weak point where rising blood pressure in adult life may begin to form a 'blow-out', a thin-walled aneurysm, which may be single or multiple. At a certain stage, any sudden extra rise in blood pressure or flow rate may rip the already tense sac, and cause extravasation of blood at full arterial pressure into the subarachnoid space. Where there is no aneurysm, it must be assumed that a weak point in a vessel wall has given way *ab initio*, without the prior formation of a sac.

The search for a small aneurysm at autopsy may be difficult because of the thick layer of blood clot that is trapped in the meninges and vessels. Blunt dissection should be used, employing the handle of a scalpel or the nose of a pair of forceps. The blood should be constantly washed away in a continuous stream of water. It is possible to inject water into one of the cut ends of a vertebral artery, having carefully tied or clamped the other vessel and the two cut terminations of the carotid arteries, to see where the water leaks from. The leaks are often multiple, however, from artefactual tearing of small vessels during autopsy removal of the brain. The search for an aneurysm is best carried out on the fresh unfixed brain, as formalin fixation hardens the blood clot so much that it cannot be removed without the danger of tearing the underlying vessel and any aneurysm. It is essential to carry out blunt dissection progressively in a continuous stream of water. Aneurysms are most often found at the bifurcation of the middle cerebral and posterior communicating arteries, at the bifurcation of the basilar arteries, on the middle cerebral in the Sylvian fissure, on the anterior communicating artery, or where the posterior communicating artery joins the posterior cerebral vessels. An aneurysm is sometimes on the cortical aspect of the artery and may be part-buried in the cerebral surface, making it hard to find. If the swelling (which may be completely collapsed at autopsy, especially when ruptured) is not seen on superficial examination of the circle of Willis, the vessels should be gently lifted away from the brain surface with a blunt elevator so that the underside can be inspected. Sometimes a buried aneurysm will rupture mainly into the cortex, causing a lesion that may be mistaken for an intracerebral haemorrhage. Berry aneurysms are frequently multiple and of varying size, from a few millimetres to several centimetres, though the usual diameter is of the order of 3–8 mm.

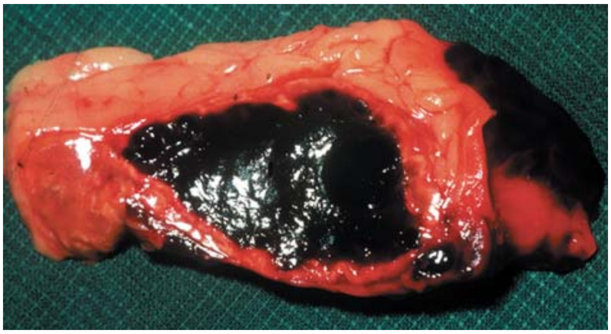

FIGURE 25.18 *Adrenal haemorrhage occurring several days after multiple injuries sustained in a traffic accident. There was no apparent abdominal injury. The haemorrhage is fresh and occupies the medulla, with the cortex stretched around the periphery.*

PULMONARY THROMBO-EMBOLISM

This is discussed fully in Chapter 13, but here it is sufficient to emphasize that pulmonary embolism is the most underdiagnosed cause of death, according to Cameron and McGoogan (1981). Their figures show that less than half the deaths shown at autopsy to have been caused by a pulmonary embolism were so diagnosed by the attending clinician. Paradoxically, many other presumed deaths from this condition were found at autopsy not to have been embolic. Thus the overall statistics are not so much in error – but in relation to the wrong victims!

Pulmonary embolism is more common with advancing age, with obese subjects and in the majority of instances there is a predisposing factor such as trauma, surgical operation, confinement to bed or immobility from another cause. Even prolonged sitting can lead to deep vein thrombosis, as was seen in persons sleeping in deckchairs in air-raid shelters in the last war and, more recently, it has been described as a hazard of long air flights.

A significant proportion, however, estimated by the author (BK) at up to 20 per cent, occur unexpectedly in the absence of any of these usual factors, making the legal problem of causation difficult. This is especially so in criminal cases where a high standard of proof has to be

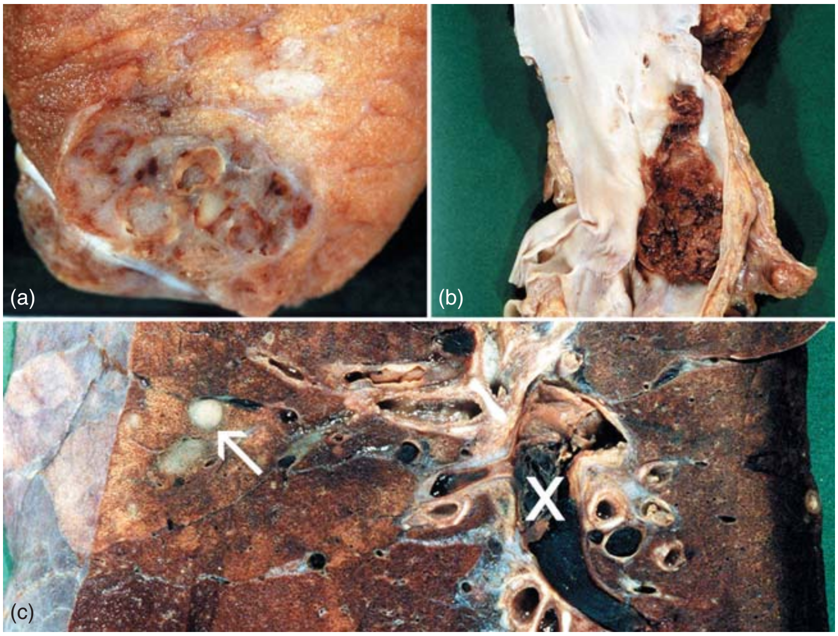

FIGURE 25.19 *A rare cause of pulmonary embolism. Metastasizing embryonal carcinoma + seminoma of the testis (a) and thrombosis attached to the wall of inferior vena cava due to metastases of the local lymph nodes (b) and (c) pulmonary metastasis (white arrow) and massive pulmonary embolism (white X).*

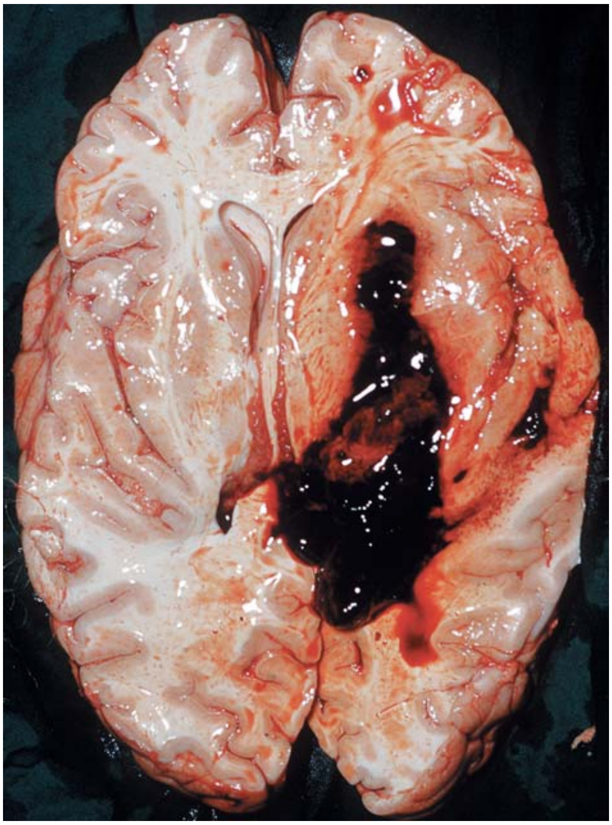

FIGURE 25.20 *A large recent intracerebral haemorrhage in a person with hypertension. The bleeding has originated in the region of the external capsule from a lenticulostriate branch of the middle cerebral artery then broken through into the posterior horn of the lateral ventricle.*

attained, rather than the balance of probabilities required for a civil action.

SUDDEN DEATH IN EPILEPSY

Every coroner's pathologist or the equivalent in other jurisdictions, has experience of sudden deaths in epileptics, when no morphological lesion can be found at autopsy. Epileptics can die unexpectedly without being in status epilepticus – or indeed, not even in a typical fit. It is true that many such deaths are not witnessed to prove that no such fit occurred, but sufficient evidence exists to indicate that a rapid sudden death can certainly occur.

There are, of course, many deaths in epileptics where the mode is apparent, such as asphyxia during a fit in bed when the face is pressed into the pillow, and saliva and mucus form an airtight seal against the fabric around the nose and mouth. Epileptics also drown in the bathtub and suffer other traumatic deaths because of fits occurring when they are in some vulnerable position.

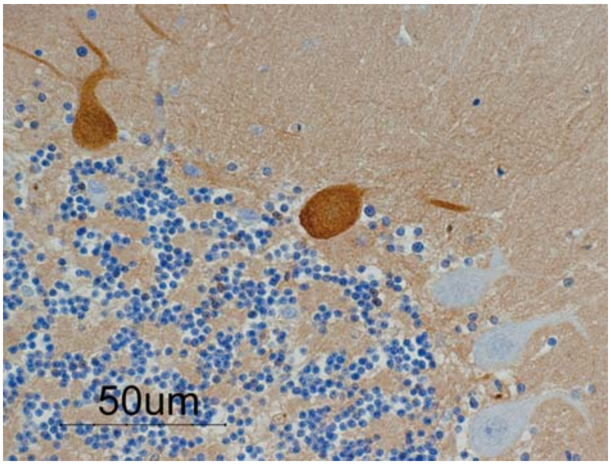

FIGURE 25.21 *Two fibrinogen-positive Purkinje cells and three negative cells in a sudden unexpected death with a history of epilepsy. (Peroxidase-DAB immunostaining, original magnification ×20.)*

These causes apart, the victim of epilepsy seems to be able to die with no apparent immediate cause and the mechanism is obscure. It has been suggested that some massive cerebral electric discharge or neurone storm occurs, leading to cardiac arrest, but this is mere hypothesis. When all investigations have been completed (including toxicological screening for antiepileptic drug overdose), after an essentially negative autopsy, then the pathologist is justified in recording 'epilepsy' as an acceptable cause of death.

The autopsy should always include a search for bites on the tip and distal sides of the tongue, which suggest a fit. A careful examination of the brain is essential, to search for any cause of post-traumatic epilepsy. In some epileptics, without coronary stenosis, patchy myocardial fibrosis has been recorded (Falconer and Rajs 1976), said to be the result of episodic hypoxia from apnoea during fits. Based on the observation of Sokrab *et al.* (1988) that Purkinje cells become labelled by plasma proteins in experimental blood–brain-barrier (BBB) breakdown, Ikegaya *et al.* (2003) investigated immunohistochemically the presence of fibrinogen-positive Purkinje cells in 50 sudden unexpected deaths. There were 24 subjects with and 26 without known history of epilepsy. Regarding $\geqslant 30$ per cent of fibrinogen-positive Purkinje cells as diagnostic limit indicating recent major BBB breakdown, they found that 22 of the 24 cases with high percentage of fibrinogen-positive Purkinje cells had a history of epilepsy. One of the two remaining high-fibrinogen subjects had unknown leucoencephalopathy and the other had essential hypertension. Only two of the low-fibrinogen cases had a history of epilepsy. The immunopositivity of Purkinje cells is not specific for epilepsy but can occur in any conditions causing BBB breakdown, most likely due to uptake of fibrinogen

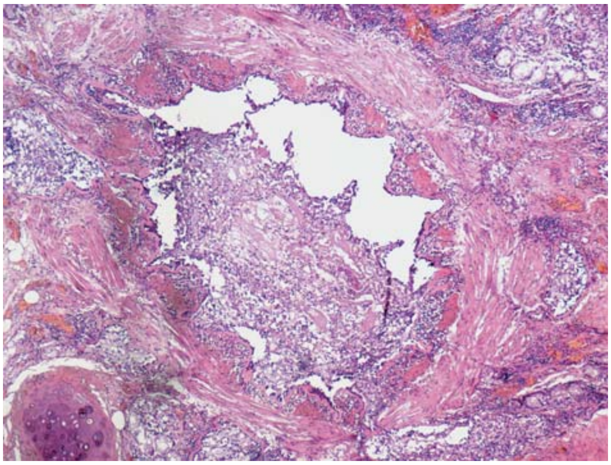

FIGURE 25.22 *Cross-section of a bronchus in sudden asthma death showing cell-laden mucus, folding of the soft tissues of the bronchial wall, thickening of the basement membrane and hypertrophy of the muscle. (HE, original magnification ×10.)*

from the cerebrospinal fluid irrigating the cerebellar surface. The time interval needed for the protein accumulation in man is not known. It is probable that, in addition to fibrinogen, other plasma proteins can also be used as a marker and an auxiliary method in assessing the probability of epileptic seizures contributing to the cause of death in sudden unexpected deaths.

SUDDEN DEATH IN BRONCHIAL ASTHMA

As in epilepsy, sufferers from bronchial asthma may die suddenly and unexpectedly, without necessarily being in status asthmaticus or even in an acute asthmatic attack. The mechanism is obscure, but experience proves that such fatalities are not uncommon. About 20 years ago, there was a marked increase in sudden deaths in asthmatics, but it was recognized that this was caused by the overuse of inhalers containing bronchodilators. These adrenergic drugs, used in excess, have a direct action upon the myocardium causing tachycardia, arrhythmias and ventricular fibrillation. Awareness among prescribing doctors rapidly reduced this danger but sudden deaths still occur, even when these substances are not used to excess.

The question of sudden death in asthma was discussed by Morild and Giertsen (1989). They considered that a number of factors contributed to death. Hypoxia and respiratory acidosis occur in asthma and increase myocardial irritability. Drugs such as theophylline and sympathomimetic agents can provide ventricular fibrillation. The β_2-antagonists so commonly used for the relief of bronchospasm are not so prone to cause cardiac arrhythmias, but they are by no means safe drugs in this respect. Of the 11 deaths in this Norwegian series, seven victims were found dead with a β_2-inhaler still clutched in their hands. Overdosage of the substance appears to be common. The use of theophylline derivatives combined with sympathomimetic agents tends to accentuate the cardiotoxicity of the bronchodilators.

Some of the deaths appeared to be in instances where the drug had been exhausted from the inhaler, but the propellant was still active. This may be another possible mechanism for some of these deaths, similar to the vagal inhibitory fatalities seen in butane misuse, where cold gas impinges on the pharynx.

At autopsy little or nothing is found, except confirmation of the chronic asthmatic state. On opening the chest cavity, the lungs are seen to fill the thorax, not collapsing as is usual and they have a pale spongy texture, forming stable pits when pressed with the fingers. The cut surface may reveal thick-walled prominent bronchi, often plugged with thick, clear mucus. This is not an explanation for the sudden collapse and death, however, and the mechanism remains obscure.

The topic of sudden death from asthma in childhood has been investigated by Champ and Byard (1994), who consider that it occurred only in children with significant chronic disease.

RESPIRATORY OBSTRUCTION

Mechanical obstruction by foreign bodies is dealt with in Chapter 14, but some rapid – if not actually sudden – deaths can be due to fulminating natural disease.

Rupture of a retropharyngeal abscess (‘quinsy’), when pus and necrotic material pour into the pharynx and larynx, is now a rare cause. Diphtheria is also virtually unknown in advanced countries so that the toxaemia, a fatal myocarditis and the possibility of laryngeal obstruction from a detached pharyngeal membrane are rarely considered by most doctors.

More common – and a cause of legal actions for negligence against doctors – is the fulminating epiglottitis caused by *Haemophilus influenzae* in children. This is a major paediatric emergency and any child with stridor of recent onset should be seen urgently by a paediatrician and anaesthetist. The gross infective oedema of the laryngeal entrance can lead to death within hours of onset of the illness, and may require rapid intubation or tracheostomy to save life.

HAEMOPTYSIS

Now that pulmonary tuberculosis is less common and treatable in advanced countries, deaths from haemoptysis are rare, as most massive respiratory haemorrhages were from this cause. However, warning signals are now evident about a revival of tuberculosis, even in 'advanced' countries. Part of the return of this scourge is due to opportunist infection in HIV-infected persons, but AIDS apart, phthisis seems to be on the march again, with worrying concerns about increasing resistance to the drugs, which revolutionized its treatment 30–40 years ago.

Bronchial tumours rarely cause fatal haemoptyses, unless a large vessel is eroded; syphilitic aortitis perforating the aorta or other large vessel is now almost a matter of history.

The bleeding rarely causes death from exsanguination except where the aorta is eroded, but smaller bleeds may fill the air passages, and cause an asphyxial-type death or a 'vagal inhibition'-type death from sudden flooding of the larynx.

GASTROINTESTINAL HAEMORRHAGE

Bleeding may take place at any point along the alimentary tract and though modern transfusion and resuscitatory methods make deaths unusual these days, some are seen in medico-legal practice when people living alone or otherwise remote from assistance, die without any medical intervention.

Bleeding from surgical operations in the mouth or pharynx, such as tooth extraction or tonsillectomy, rarely causes death, but can occasionally do so from blood and clot blocking the airways. In the oesophagus, a penetrating carcinoma may erode the aorta or other great vessel in the mediastinum, but the most frequent sources of bleeding are varices at the lower end, caused by the portal hypertension associated with hepatic fibrosis. It is often difficult to identify the varices at autopsy, as they have collapsed, but the absence of gastric or duodenal ulceration, and the congested, bluish vessels around the cardia often give away the lesion, especially in the presence of liver cirrhosis and splenomegaly.

Other causes of massive gastrointestinal bleeding are gastric and duodenal peptic ulcers, which can erode large vessels in their base. Acute erosions, though shallow, can produce considerable blood in the stomach and may require surgical resection. They are rarely the cause of lethal bleeding.

Carcinoma of the stomach can reach a large size and the crater variety may perforate a major vessel to produce severe haemorrhage. Bleeding from the small intestine is rare, except in certain haemorrhagic diatheses. More common is bleeding from colonic lesions, such as carcinomas and ulcerative colitis.

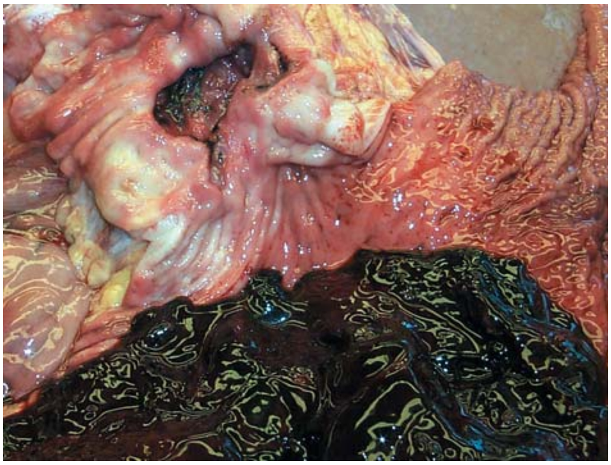

FIGURE 25.23 *Massive lethal bleeding from eroded vessels of a gastric ulcer.*

Polyps, sometimes malignant, may also give rise to melaena and even massive bleeding *per rectum*, though these are now rarely fatal unless medical intervention is delayed or absent.

FATAL ABDOMINAL CATASTROPHES

Though usually surgical emergencies, some abdominal conditions go untreated because the victim is either a solitary person living alone with no opportunity to call for help, or declines to seek medical assistance. They may well be found dead with no history to suggest a cause of death.

Mesenteric thrombosis and infarction is one such condition. The thrombosis may in fact be an embolus, rather than a thrombosis *in situ*, the most common cause of both being atherosclerosis of the aorta and its mesenteric branches. The length of small intestine supplied by the vessel becomes dark and necrotic, and may involve almost the whole length of the jejunum and ileum. Even if precipitated by an embolus, much of the rest of the arterial system may then thrombose and, on cutting into the mesentery at autopsy, the cut ends of arteries may show plugs of firm ante-mortem thrombus. The gut will be blue–red or even almost black, with loss of the serosal lustre and a friable necrotic wall.

Strangulated intestine is yet another condition that may present as sudden or rapid death, where medical attention is lacking, and even sometimes when it is negligently offered.

Hernias, both femoral and inguinal, and internal strangulation beneath fibrous bands in the peritoneal cavity, may all twist the intestine so that its blood supply is cut off, with subsequent necrosis. The gut above may develop ileus and become paralysed and dilated.

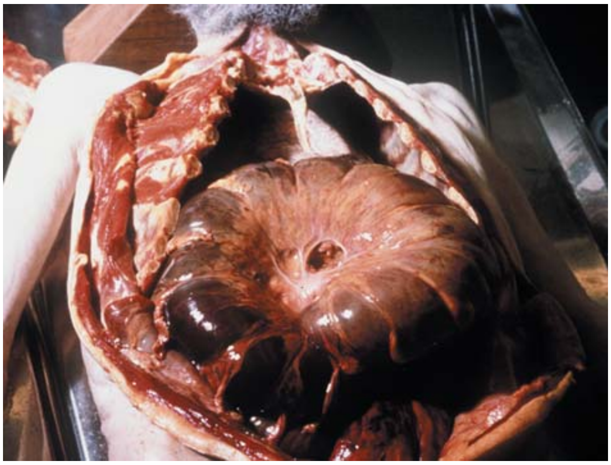

FIGURE 25.24 *Great distension of colon due to volvulus of sigmoid colon. Deceased found dead with history of abdominal pain.*

Caution is needed at autopsy to differentiate the dark red colour of loops of intestine due to post-mortem hypostasis from true infarction. Hypostasis is seen to be interrupted when the gut is stretched out, due to alternate dependent loops, whereas real necrosis is usually continuous and the serosa is dull and friable.

Fulminating peritonitis may be seen in a variety of causes, again leading to death if undiagnosed or untreated. Death from acute appendicitis and appendix abscess is now uncommon, but perforation of the colon through a diverticulum, whether inflamed or merely distended and eroded by faecoliths, is common.

Tearing of the senile gut where it is adherent to other structures, such as uterus or tube, can also lead to peritonitis, and a carcinoma (often difficult to differentiate macroscopically from chronic diverticulitis) may also lead to perforation with loss of intestinal contents into the peritoneum.

A perforated peptic ulcer, usually duodenal, leads to a chemical peritonitis. The autopsy appearances are straightforward, turbid fluid being found in the peritoneal cavity, which may show early inflammatory changes, if the perforation occurred many hours previously. The perforation is most commonly found on the posterior surface of the duodenum.

GENITOURINARY SYSTEM

There is very little in the urinary system that can give rise to a sudden or rapid death, but the female genital organs certainly make up for this. Complications of pregnancy are discussed elsewhere (Chapter 19), but it should be repeated here that ectopic pregnancy, usually in the tubes, can rupture with massive intraperitoneal haemorrhage. Induced abortions, unless under reputable medical control, are another source of death from haemorrhage, air embolism, perforation of vagina or uterus, infection and the use of toxic substances.

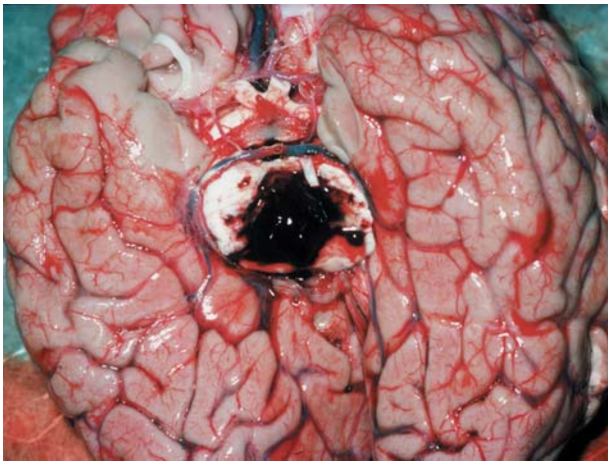

FIGURE 25.25 *A massive pontine haemorrhage in a person with hypertension. The circumstances and the irregular distribution of the bleeding within the brainstem distinguish this primary haemorrhage from the secondary lesions seen in raised intracranial pressure. Sometimes a natural intracranial haemorrhage may precipitate an accidental fall or traffic accident, however, and the resulting head injury may make it more difficult to identify the nature of the brain haemorrhage.*

THE RELATIONSHIP BETWEEN TRAUMA AND DISEASE

One of the most difficult problems in forensic medicine occurs when a death takes place in a person who has suffered trauma but who has either pre-existing natural disease or where some apparently natural disease has supervened after the trauma. The relative contributions of trauma and disease may then become an acute medico-legal problem.

For example, the author (BK) conducted an autopsy on an old man of 82 years, previously in excellent health, who had been tied up and slightly injured by robbers of a sub-post office. Within 2 hours of release from his bonds the old gentleman became ill and then hemiplegic, dying of a massive cerebral haemorrhage shortly afterwards. The attackers were charged with homicide, in addition to robbery, but that charge was rejected by magistrates on the grounds that a cerebral haemorrhage could be a natural disease, unrelated to the assault. The prosecution obtained a private Bill of Indictment to circumvent the magistrates, and at the Crown Court the accused were convicted of manslaughter on the grounds that the physical and emotional stress of being assaulted and tied up probably (beyond reasonable doubt) led to an elevation of blood pressure that precipitated a cerebral haemorrhage in a previously fit person, within such a short time span.

This marginal case illustrates the problems that can occur when trauma and natural disease coexist in the same person. A common dilemma exists in relation to subarachnoid haemorrhage from a ruptured berry aneurysm when there has been a head injury.

In autopsies when an injury has been sustained by a person with substantial natural disease, the following problems must be addressed:

- whether death was caused entirely by the disease and would have occurred irrespective of the injury
- whether death was caused entirely by the injury and would have occurred whether or not the disease was present
- whether the death was caused by a combination of these two processes.

In practical terms, the most common situations involve coronary artery disease, pulmonary embolism and subarachnoid haemorrhage.

Coronary atherosclerosis

The disease has invariably been present for months or probably years before the traumatic episode, and it is therefore indisputably obvious that the trauma can have no relationship to the genesis of the disease – though relatives and some lawyers are often hard to convince on this point. Furthermore, unless there was a direct blow onto the front of the thorax, which could physically have damaged the heart surface and caused a direct traumatic lesion of the coronary system (by dislodging an atheromatous plaque – but see end of this section – causing a subintimal haemorrhage or precipitating a coronary thrombosis, a rare event), there is no way in which the injury can be accused of worsening the state of the coronary arteries. If the state of the coronary system was already poor, thus it could justifiably be claimed by the defence that death could have occurred at any time. The standard of proof in criminal cases is high and must be proved ‘beyond reasonable doubt’, whereas in civil disputes, only ‘the balance of probabilities’ must be attained.

As opposed to showing that the trauma directly worsened the physical state of the coronary vessels, it is easier to claim that the physical and emotional stress associated with the traumatic event caused increased demands upon a weakened heart and caused it to fail. Recent investigations (see end of this section) have strongly confirmed the association between acute myocardial infarction, cardiac arrest and exertion.

The adrenal response of ‘fight or flight’ may be invoked, it being accepted that released endogenous catecholamines such as noradrenaline can send a chronically hypoxic heart into ventricular fibrillation and arrest. There may even be morphological evidence in some cases, by the finding of contraction bands in the cardiac myofibrils, especially in the superficial subepicardial layers.

It is often an assessment of the circumstances, in addition to the autopsy findings, that decides a court’s attitude to the relationship of trauma to disease. Where a man who has had no previous cardiac symptoms drops dead immediately after some assault or strenuous activity, the court is more likely to assume an association than with some cardiac cripple, who has been hovering on the brink of heart dysfunction for a long time before the traumatic event.

The average jury is likely to decide that it would be too much of a coincidence for an assault to cause a man to drop dead on the spot, even if it was shown that his coronary artery stenosis had been there for several years. Much depends upon the individual circumstances, especially where there is only a short interval between the traumatic episode and death.

The strict test is: ‘Would he have died when he did die if the assault had not taken place?’ This is often unanswerable on medical grounds, but the court is entitled to take a commonsense view of coincidence in terms of immediacy in time and apply the ‘beyond reasonable doubt’ test to it.

The fact that a person is in a parlous state of health and might die from relatively minimal trauma is no defence in law, which states that ‘an assailant must take his victims as he finds them’. In other words, it is immaterial whether the attacker knows that his victim is ill or not – if death results from an illegal act against him, the charge of homicide may stand. This dictum is not adhered to slavishly by public prosecutors, however, as it is a matter of practical politics whether a criminal prosecution can be maintained with any chance of success (using public funds), against an energetic defence that can show medically that the victim had such a degree of natural disease that he might have dropped dead spontaneously at any time. The prevalence of coronary disease is so widespread that a close association in time must be shown before any causal connection can be accepted. The actual death itself need not occur very soon after the assault, however, if it can be established that the latter could have precipitated an infarct in a myocardium already compromised by coronary stenosis. It is not only a blow or some physical trauma that may precipitate a myocardial infarct or arrhythmia; as stated above, even the emotional upset that accompanies injury, or even the threat or fear of an injury, can lead to death. The blow may never actually be struck or it may fail to land, yet the threatened person may suffer a transient hypertension or tachycardia that may precipitate a subintimal haemorrhage, arrhythmia, or cerebral or subarachnoid haemorrhage that leads to death. Such an event, however, with little or no morphological evidence to be presented as evidence, has a poor chance of acceptance in court unless the circumstances are clear-cut. These cases pose a problem for police and pathologist, in that it may be difficult to decide whether to arrest and charge the potential culprit with a criminal offence.

A typical case was once seen by the author (BK) in which an elderly man was involved in an altercation with another, concerning parking a car. There was a scuffle, in which only trivial blows were exchanged, but one participant immediately developed chest pain and breathlessness, and soon died. The other man was arrested, but as autopsy showed marked cardiac enlargement, extensive myocardial fibrosis and gross occlusive coronary atherosclerosis, the prosecution decided to offer no evidence and the charge was dropped, even though the answer to the 'acid test' of whether the victim would have been likely to die at that particular time had the altercation not taken place was probably in the negative.

In civil matters, considerable sums of money may hang upon such decisions by way of damages and insurance payments. Here the standard of proof is much lower, as the plaintiff only has to show that there was a 51 per cent or better chance of the association being present, rather than the much higher standard of criminal liability.

An actual example of the civil problems was a workman who was removing a truck wheel from his employer's vehicle with a wheelbrace. Under full pressure, the threaded stud in the hub suddenly snapped and the man was thrown to the ground under the force of his own muscular effort. He was shocked and soon developed chest pain and died the next day. Autopsy revealed no fresh infarct, but there was longstanding left ventricular hypertrophy, myocardial fibrosis and stenosing coronary atheroma. The employer's liability was disputed, but eventually a compromise settlement was reached.

Recent very careful and statistically meticulous research on large cohorts, both in the USA and Germany, have shown an indisputable relationship between exertion and sudden cardiac death. Mittleman *et al.* (1993) investigated 1228 patients with acute myocardial infarcts and showed that there was a sixfold increase in the incidence of infarction during or within one hour of heavy physical exertion, such as snow shovelling, jogging or sexual activity. In Germany, Willich *et al.* (1993) researched 1194 patients with the same condition and showed a twofold increase in risk. Both surveys showed that the risk was greater in those who were otherwise sedentary and that long-term moderate exercise was undoubtedly associated with a reduced risk of infarction (Curfman 1993). Several mechanisms were proposed for the increased risk during sudden severe exertion, including the splitting and dislodgement of an atheromatous plaque. It was also claimed that increased platelet activation occurred in sedentary people who suddenly exerted themselves. In an earlier publication, Siscovick *et al.* (1984) showed the same trend in primary cardiac arrest, exertion in otherwise sedentary persons increasing the risk, whilst habitual exercise protected.

Subarachnoid haemorrhage

Subarachnoid haemorrhage is another difficult area in which trauma and natural disease interact. When a berry aneurysm on the cerebral arteries ruptures, the allegation is often made that some traumatic occurrence precipitated that event. The matter is discussed in Chapter 5, but to recapitulate here, the problem is to decide whether the trauma itself was a mechanical factor in causing the arterial blister to burst.

A rupture is a relatively common event without external trauma, this cause of death being well-known in young to middle-aged persons who are not indulging in any strenuous activity whatsoever. It is known to occur, however, during innocent activities, such as jogging, intercourse and sporting exertion, probably because of a transient rise in blood pressure and pulse rate. Once again, a potent element may be in the adrenal response, catecholamines pushing up the blood pressure. Thus when a ruptured aneurysm occurs during or shortly after a fight or altercation, it is arguable whether a mechanical blow to the head or merely the emotional stress of a dispute is the most potent factor in causing rupture. There has been much medical argument about the role of trauma in causing the rupture, as many would maintain that the aneurysm is so deep-seated within the skull that internal pressure from hypertension is far more likely to be the reason.

Certainly, where a small aneurysm a few millimetres in size breaks, it is difficult to accept that a blow could have caused a tear, but in the more exceptional cases of a large tissue-thin globule, then it is easier to accept that intracranial stresses may have played a significant part. The role of alcoholic intoxication – another hotly disputed aspect – has been discussed in Chapter 5. The legal problem is similar in berry aneurysm deaths to that in coronary disease. The time interval is naturally important, though in most instances, bleeding occurs immediately or soon afterwards.

Autopsy is not all that helpful in assessing the relative contribution of assault versus pre-existing disease. Unless the aneurysm is found to be large and fragile, the morphological findings are mainly of use in confirming the diagnosis, evaluating any injuries and excluding any other contributory disease. It used to be the usual practice in England and Wales not to prosecute assailants when an aneurysm was demonstrated at autopsy, though in Scotland and on the continent of Europe no such reluctance was evident. Following a case in Gibraltar in 1978, however, in which a British sailor was convicted of homicide after kicking another in a drunken brawl, the latter eventually dying of subarachnoid haemorrhage from a ruptured aneurysm, such charges are now usually brought (Knight 1979).

Pulmonary embolism

The other major example of the interaction of trauma and natural disease is pulmonary embolism. This is discussed in Chapter 13, but is so important in medico-legal casework that the forensic aspects may be repeated here with advantage.

Deep vein thrombosis, almost always in the leg veins, is a common sequel to injury and consequent immobility. Virchow's triad of slowing of the circulation, local injury to vessel walls and increase in blood coagulability are the major factors, though deep vein thrombosis can occur in the absence of all three. When the leg is the site of the injury, it is more likely to occur in the ipsilateral limb, though it can occur contralaterally or bilaterally. The fact that a fatal embolus came from the contralateral limb when the other leg has been injured by no means rules out a causative element.

Pulmonary embolism typically occurs about 2 weeks after an injury or surgical operation, but the range of time during which a cause-and-effect mechanism operates can be anything from about 2 to 90 days. It becomes progressively more difficult to maintain a causative relationship, however, when the interval exceeds a few weeks.

Forensically, it is vital to show that the deep vein thrombosis post-dated the traumatic event, if an association is to be established. If a person suffers a fatal embolism a week after injury, yet histologically the deep vein thrombosis appears to be several weeks old, then obviously the injury could not have initiated the process. The embolus may, however, be much younger than the most distal part of the leg vein thrombosis (which may be as far down as the foot) as progressive proximal extension of a deep vein thrombosis undoubtedly occurs. It could still be argued, therefore, that although the original thrombosis antedated the injury, the latter encouraged extension and hence the eventual breaking off of an embolus. The case becomes weaker, however, than if the whole process was entirely subsequent to the traumatic event.

Histological dating of the embolus should always be undertaken, though it is less rewarding and more difficult than that of the leg vein thrombi, which must always be taken for microscopy with the vein wall intact. It is the thrombomural junction that gives the most information about the age of any thrombus. Thrombosis of the deep leg veins is far more common than pulmonary thromboembolism. In 100 consecutive coroner's autopsies, Knight and Zaini (1980) found 32 instances of deep vein thrombosis, but only 10 had pulmonary emboli, not all fatal.

The legal problem, outlined in Chapter 13, is that though trauma, surgery and immobility are potent factors in the development of deep vein thrombosis, the latter condition frequently occurs in the absence of all three. Thus it is difficult to maintain that any traumatic event was the cause of a subsequent fatal pulmonary embolism, as many fatal emboli occur in the absence of trauma. In a series of coroner's and hospital autopsies studied by Knight (1966), 25 per cent of fatal pulmonary emboli came 'out of the blue' in ambulant people who had no previous history of trauma or surgical operation. Later, Knight and Zaini studied 38 000 autopsy reports and found that 10 per cent of deaths due to pulmonary embolism had no history of any predisposing factors. The discrepancy in these proportions is caused by the variable standard of history reports and also by marked alterations in pathologist's reporting habits over the years – in fact, in the large series mentioned above, there was not a single fatal pulmonary embolism recorded in the 20 years from 1908 until 1928!

Even disregarding the dubious statistical value of such investigations, there can be no doubt that a significant proportion of fatal pulmonary emboli cannot be related to previous trauma. This provides a line of defence in those cases where the prosecution allege that a 'cause-and-effect' relationship exists between some criminal assault and death. It becomes a matter of law for the judge to decide whether he will let the issue go to the jury, and a matter of fact for the jury to decide whether this relationship can be 'beyond reasonable doubt' in circumstances where even conservatively, at least 10 per cent of pulmonary emboli are not related to trauma. Several court cases are within the author's (BK) experience in which quite different verdicts have been reached on essentially identical pathological evidence.

In civil cases, usually related to road traffic accidents or industrial personal injuries, the burden on the plaintiff is much lighter, as the 'balance of probabilities' test can be much more readily satisfied.

REFERENCES AND FURTHER READING

Al Rufaie HK, Florio RA, Olsen EG. 1983. Comparison of the haematoxylin basic fuchsin picric acid method and the fluorescence of haematoxylin and eosin stained sections for the identification of early myocardial infarction. *J Clin Pathol* **36**:646–9.

Anderson HN. 1964. An evaluation and comparison of macroscopic enzyme techniques for the autopsy demonstration of myocardial infarction. *J Pathol* **127**:93–8.

Ansari A, Maron BJ, Berntson DG. 2003. Drug-induced toxic myocarditis. *Tex Heart Inst J* **30**:76–9.

Aretz HT. 1987. Myocarditis: the Dallas criteria. *Hum Pathol* **18**:619–24.

Aretz HT, Billingham ME, Edwards WD, *et al.* 1987. Myocarditis. A histopathologic definition and classification. *Am J Cardiovasc Pathol* **1**:3–14.

Badir B, Knight B. 1987. Fluorescence microscopy in the detection of early myocardial infarction. *Forensic Sci Int* **34**:99–102.

Bajanowski T, Rossi L, Biondo B, *et al.* 2001. Prolonged QT interval and sudden infant death – report of two cases. *Forensic Sci Int* **115**:147–53.

Balachandra AT, O'Conner R, Bowden DH. 1987. Sudden unexpected death in asthmatics. *J Can Forensic Sci* **20**:227.

Baroldi G. 1975. Different types of myocardial necrosis in coronary heart disease: a pathophysiologic review of their functional significance. *Am Heart J* **89**:742–52.

Bateman JR, Clarke SW. 1979. Sudden death in asthma. *Thorax* **34**:40–4.

Bjork VD. 1964. *Cardiomyopathies.* Ciba Foundation Symposium. Churchill, London.

Bostrom H. 1966. Acute epiglottitis as a cause of sudden death. *Dtsch Z Gesamte Gerichtl Med* **61**:53–9.

Bouchardy B, Majno G. 1971. A new approach to the histologic diagnosis of early myocardial infarcts. *Cardiology* **56**:327–32.

Bouchardy B, Majno G. 1974. Histopathology of early myocardial infarcts. A new approach. *Am J Pathol* **74**:301–30.

Cameron HM, McGoogan E. 1981. A prospective study of 1152 hospital autopsies: I. Inaccuracies in death certification. *J Pathol* **133**:273–83.

Cardell BS, Pearson RS. 1959. Death in asthmatics. *Thorax* **14**:341–52.

Carle BN. 1981. Autofluorescence in the identification of myocardial infarcts. *Hum Pathol* **12**:643–6.

Champ CS, Byard RW. 1994. Sudden death in asthma in childhood. *Forensic Sci Int* **66**:117–27.

Chan AC, Dickens P. 1992. Tuberculous myocarditis presenting as sudden cardiac death. *Forensic Sci Int* **57**:45–50.

Clark JC. 1988. Sudden death in the chronic alcoholic. *Forensic Sci Int* **36**:105–11.

Clawson BJ. 1928. Myocarditis. *Am Heart J* **4**:1–7.

Cochrane GM, Clark JH. 1975. A survey of asthma mortality in patients between ages 35 and 64 in the greater London hospitals in 1971. *Thorax* **30**:300–5.

Copeland AR. 1986. Asthmatic deaths in the medical examiner's population. *Forensic Sci Int* **31**:7–12.

Corby C. 1960. Isolated myocarditis as a cause for sudden obscure death. *Med Sci Law* **1**:1–40.

Crawford T. 1977. *Pathology of ischaemic heart disease.* Butterworth, London.

Curfman GD. 1993. Is exercise beneficial – or hazardous – to your heart? [editorial; comment] [see comments]. *N Engl J Med* **329**:1730–1.

Cushing EH. 1942. Infarction of the cardiac auricles. *Br Heart J* **4**:17–26.

Davies M, Pomerance A. 1975. *Pathology of the heart.* Blackwell Scientific Publications, Oxford.

Davies MJ. 1981. Pathological view of sudden cardiac death. *Br Heart J* **45**:88–96.

Davies MJ, Popple A. 1979. Sudden unexpected cardiac death – a practical approach to the forensic problem. *Histopathology* **3**:255–77.

Decastello A, Remenar E, Toth J, *et al.* 1977. Post mortem detection of early myocardial infarction by determination of the tissue K^+/Na^+ ratio. *Acta Morphol Acad Sci Hung* **25**:289–96.

Derias NW, Adams CW. 1979. The non-specific nature of the myocardial wavy fibre. *Histopathology* **3**:241–5.

Derias NW, Adams CW. 1982. Macroscopic enzyme histochemistry in myocardial infarction: use of coenzyme, cyanide, and phenazine methosulphate. *J Clin Pathol* **35**:410–13.

deSa DJ. 1986. Isolated myocarditis as a cause of sudden death in the first year of life. *Forensic Sci Int* **30**:113–17.

Dettmeyer R, Schlamann M, Madea B. 1999. Immunohistochemical techniques improve the diagnosis of myocarditis in cases of suspected sudden infant death syndrome (SIDS). *Forensic Sci Int* **105**:83–94.

Dettmeyer R, Kandolf R, Schmidt P, *et al.* 2001. Lymphomonocytic enteroviral myocarditis: traditional, immunohistological and molecular pathological methods for diagnosis in a case of suspected sudden infant death syndrome (SIDS). *Forensic Sci Int* **119**:141–4.

Dettmeyer R, Reith K, Madea B. 2002. Alcoholic cardiomyopathy versus chronic myocarditis – immunohistological investigations with LCA, CD3, CD68 and tenascin. *Forensic Sci Int* **126**:57–62.

Drury RA. 1954. The role of intimal haemorrhage in coronary occlusion. *J Pathol Bacteriol* **67**:207–15.

Edston E. 1997. Evaluation of agonal artifacts in the myocardium using a combination of histological stains and immunohistochemistry. *Am J Forensic Med Pathol* **18**:163–7.

Edston E, Kawa K. 1995. Immunohistochemical detection of early myocardial infarction. An evaluation of antibodies against the terminal complement complex (C5b-9). *Int J Legal Med* **108**:27–30.

English JP, Williams FA. 1943. Haemorrhagic lesions of the coronary arteries. *Arch Intern Med* **71**:594–602.

Falconer B, Rajs J. 1976. Post-mortem findings of cardiac lesions in epileptics: a preliminary report. *Forensic Sci* **8**:63–71.

Farrer-Brown G. 1968. Normal and diseased vascular pattern of myocardium of human heart. I. Normal pattern in the left ventricular free wall. *Br Heart J* **30**:527–36.

Farrer-Brown G. 1977. *A colour atlas of cardiac pathology.* Wolfe Medical Publications, London.

Feibel JH, Campbell RG, Joynt RJ. 1976. Myocardial damage and cardiac arrhythmias in cerebral infarction and subarachnoid hemorrhage: correlation with increased systemic catecholamine output. *Trans Am Neurol Assoc* **101**:242–4.

Feldman AM, McNamara D. 2000. Myocarditis. *N Engl J Med* **343**:1388–98.

Feldman S, Glagov S, Wissler RW, *et al.* 1976. Postmortem delineation of infarcted myocardium. Coronary perfusion with nitro blue tetrazolium. *Arch Pathol Lab Med* **100**:55–8.

Ferris JA. 1974. Conducting tissue changes in sudden death. *Med Sci Law* **14**:36–9.

Ferris JA, Friesen JM. 1979. Definitions of ischaemia, infarction and necrosis. *Forensic Sci Int* **13**:253–9.

Ferris JA, MacLennan JR. 1973. A simplified method for examining the conducting tissues of the heart. *Med Sci Law* **13**:285–8.

Ferris JA, Rice J. 1979. Drug-induced myocarditis: a report of two cases. *Forensic Sci Int* **13**:261–5.

Fleming HA. 1988. Death from sarcoid heart disease: United Kingdom series 1971–1986, 300 cases with 138 deaths. In: Grassi C (ed.), *Sarcoidosis and other granulomatous disorders.* Excerpta Medica, Amsterdam, pp. 19–33.

Fleming HA, Bailey SM. 1981. Sarcoid heart disease. *J R Coll Physicians Lond* **15**:245–6, 249–53.

Fothergill DF, Bowen DA, Mason JK. 1979. Dissecting and atherosclerotic aneurysms: a survey of post-mortem examinations, 1968–77. *Med Sci Law* **19**:253–60.

Fraser PM, Speizer FE, Waters SD, *et al.* 1971. The circumstances preceding death from asthma in young people in 1968 to 1969. *Br J Dis Chest* **65**:71–84.

Fukumoto H, Naito Z, Asano G, *et al.* 1998. Immunohistochemical and morphometric evaluations of coronary atherosclerotic plaques associated with myocardial infarction and diabetes mellitus. *J Atheroscler Thromb* **5**:29–35.

Glogner. 1903. Cited by Stamer. 1907. *Virchows Arch Pathol Anat* 171.

Gore I, Saphir O. 1947. Myocarditis: a classification of 1042 cases. *Am Heart J* **34**:827–41.

Gormsen H. 1955. Sudden unexplained death due to myocarditis. *Acta Pathol Microbiol Scand* **105**(Suppl):303.

Gregersen M. 1979. Myocardial changes in sudden unexpected coronary deaths. *Forensic Sci Int* **13**:183–6.

Hansen SH, Rossen K. 1999. Evaluation of cardiac troponin I immunoreaction in autopsy hearts: a possible marker of early myocardial infarction. *Forensic Sci Int* **99**:189–96.

Hiltunen JK, Saukko P, Hirvonen J. 1985. Correlations between enzyme histochemical reactions and respective enzyme activities in global ischaemic rat hearts. *Br J Exp Pathol* **66**:743–52.

Hirvonen J, Hiltunen JK, Saukko P. 1987. Oxidative enzyme activities and respective histochemical reactions in ischemic rat myocardium. *Forensic Sci Int* **35**:231–6.

Holmbom B, Lindstrom M, Naslund U, *et al.* 1991. A method for enzyme- and immunohistochemical staining of large frozen specimens. *Histochemistry* **95**:441–7.

Hougen HP, Valenzuela A, Lachica E, *et al.* 1992. Sudden cardiac death: a comparative study of morphological, histochemical and biochemical methods. *Forensic Sci Int* **52**:161–9.

House RK. 1948. Diffuse interstitial myocarditis in children. *Am J Pathol* **24**:1235–48.

Hu BJ, Chen YC, Zhu JZ. 1996. Immunohistochemical study of fibronectin for postmortem diagnosis of early myocardial infraction. *Forensic Sci Int* **78**:209–17.

Hudson RE. 1963. The human conducting system and its examination. *J Clin Pathol* **16**:492–8.

Hudson RE. 1970. The cardiomyopathies: order from chaos. *Am J Cardiol* **25**:70–7.

Ikegaya H, Heino J, Laaksonen H, *et al.* 2003. Accumulation of fibrinogen in Purkinje cells in sudden death in epilepsy. *Forensic Sci Int* accepted for publication.

Inman WH, Adelstein AM. 1969. Rise and fall of asthma mortality in England and Wales in relation to use of pressurised aerosols. *Lancet* **2**:279–85.

Iwadate K, Tanno K, Doi M, *et al.* 2001. Two cases of right ventricular ischemic injury due to massive pulmonary embolism. *Forensic Sci Int* **116**:189–95.

Jaakkola K, Jalkanen S, Kaunismaki K, *et al.* 2000. Vascular adhesion protein-1, intercellular adhesion molecule-1 and P-selectin mediate leukocyte binding to ischemic heart in humans. *J Am Coll Cardiol* **36**:122–9.

Jääskeläinen AJ. 1968. Histochemical observations on phosphorylase, cytochrome oxidase and succinic dehydrogenase in experimental and human fresh myocardial infarcts. *Ann Univ Turku* **Series A II**(Suppl 39):1–64.

Jackson RT, Beaglehole R, Rea HH, *et al.* 1982. Mortality from asthma: a new epidemic in New Zealand. *Br Med J Clin Res Ed* **285**:771–4.

James TN. 1996. Long reflections on the QT interval: the sixth annual Gordon K. Moe lecture. *J Cardiovasc Electrophysiol* **7**:738–59.

Johnson AJ, Nunn AJ, Somner AR, *et al.* 1984. Circumstances of death from asthma. *Br Med J Clin Res Ed* **288**:1870–2.

Karhunen PJ, Penttilä A, Erkinjuntti T. 1990. Arteriovenous malformation of the brain: imaging by postmortem angiography. *Forensic Sci Int* **48**:9–19.

Kherdar A, Nevins MA. 1973. Right ventricular infarction. *Med Soc New Jersey* **70**:374–8.

Knight B. 1965. The postmortem detection of early myocardial infarction. *Med Sci Law* **5**:31–46.

Knight B. 1966. Fatal pulmonary embolism: factors of forensic interest in 400 cases. *Med Sci Law* **6**:150–4.

Knight B. 1967. Early myocardial infarction. Practical methods for its post-mortem demonstration. *J Forensic Med* **14**:101–7.

Knight B. 1971. The value of enzyme techniques in medicolegal pathology. In: Wecht C (ed.), *Legal medicine annual 1971*. Appleton Century Crofts, New York.

Knight B. 1973. The myocardium in sudden death from hypertensive heart disease or aortic stenosis. *Med Sci Law* **13**:280–4.

Knight B. 1976. Investigation of sudden deaths from myocardial ischaemia. *Forensic Sci* **8**:33–6.

Knight B. 1979. A further evaluation of the reliability of the HBFP stain in demonstrating myocardial damage. *Forensic Sci Int* **13**:179–81.

Knight B. 1979. Trauma and ruptured cerebral aneurysm. *Br Med J* **1**:1430–1.

Knight B, Zaini MR. 1980. Pulmonary embolism and venous thrombosis. A pattern of incidence and predisposing factors over 70 years. *Am J Forensic Med Pathol* **1**:227–32.

Knight BH. 1966. *The postmortem diagnosis of early myocardial infarction*. Cardiff University of Wales, Cardiff.

Lachica E, Villanueva E, Luna A. 1988. Comparison of different techniques for the postmortem diagnosis of myocardial infarction. *Forensic Sci Int* **38**:21–6.

Leadbeatter S, Stansbie D. 1984. Postmortem diagnosis of familial hypercholesterolaemia. *Br Med J Clin Res Ed* **289**:1656.

Leadbeatter S, Wawman H, Jasani B. 1989. Immunocytochemical diagnosis of early ischaemic/hypoxic damage in myocardium. *Acta Med Leg Soc (Liege)* **39**:187–8.

Leadbeatter S, Wawman HM, Jasani B. 1990. Further evaluation of immunocytochemical staining in the diagnosis of early myocardial ischaemic/hypoxic damage. *Forensic Sci Int* **45**:135–41.

Leestma JE, Kalelkar MB, Teas SS, *et al.* 1984. Sudden unexpected death associated with seizures: analysis of 66 cases. *Epilepsia* **25**:84–8.

Leestma JE, Hughes JR, Teas SS, *et al.* 1985. Sudden epilepsy deaths and the forensic pathologist. *Am J Forensic Med Pathol* **6**:215–18.

Lie JT, Holley KE, Kampa WR, *et al.* 1971. New histochemical method for morphologic diagnosis of early stages of myocardial ischemia. *Mayo Clin Proc* **46**:319–27.

Luke JL, Helpern M. 1968. Sudden unexpected death from natural causes in young adults. A review of 275 consecutive autopsied cases. *Arch Pathol* **85**:10–17.

Macdonald JB, Seaton A, Williams DA. 1976. Asthma deaths in Cardiff 1963–74: 90 deaths outside hospital. *Br Med J* **1**:1493–5.

Maisch B, Ristic AD, Portig I, *et al.* 2003. Human viral cardiomyopathy. *Front Biosci* **8**:S39–67.

Malik MA. 1973. Emotional stress as a precipitating factor in sudden deaths due to coronary insufficiency. *J Forensic Sci* **18**:47–52.

Malik MO. 1979. Sudden coronary deaths associated with sexual activity. *J Forensic Sci* **24**:216–20.

Mallory GK, White PD, Salcedo-Salgar J. 1939. The speed of healing of myocardial infarction. *Am Heart J* **18**:647–71.

Manion WC. 1966. Myocarditis: a review of the development of the concept. *Med Ann D C* **35**:405–13 passim.

Mann JI. 1984. Familial hypercholesterolaemia: renewed interest in an old problem [editorial]. *Br Med J Clin Res Ed* **289**:396.

Maron BJ. 1983. Myocardial disorganisation in hypertrophic cardiomyopathy. Another point of view. *Br Heart J* **50**:1–3.

Maron BJ. 2002. Hypertrophic cardiomyopathy: a systematic review. *JAMA* **287**:1308–20.

Marshall TK. 1970. Asymmetrical hypertrophy of the heart. *Med Sci Law* **10**:3–6.

Martin AB, Webber S, Fricker FJ, *et al.* 1994. Acute myocarditis. Rapid diagnosis by PCR in children. *Circulation* **90**:330–9.

Martin AM, Jr, Hackel DB. 1966. An electron microscopic study of the progression of myocardial lesions in the dog after hemorrhagic shock. *Lab Invest* **15**:243–60.

Martin AM, Jr, Hackel DB, Entman ML, *et al.* 1969. Mechanisms in the development of myocardial lesions in hemorrhagic shock. *Ann NY Acad Sci* **156**:79–90.

Mittleman MA, Maclure M, Tofler GH, *et al.* 1993. Triggering of acute myocardial infarction by heavy physical exertion. Protection against triggering by regular exertion. Determinants of myocardial infarction onset study investigators [see comments]. *N Engl J Med* **329**:1677–83.

Morild I, Giertsen JC. 1989. Sudden death from asthma. *Forensic Sci Int* **42**:145–50.

Möttönen M. 1970. Myocardial infarction and coronary atherosclerosis in forensic autopsy material. *Med Sci Law* **10**:115–19.

Nayar A, Olsen EG. 1974. The use of the basic fuchsin stain in the recognition of early myocardial ischaemia. *Cardiovasc Res* **8**:391–4.

Noren GR, Staley NA, Bandt CM, *et al.* 1977. Occurrence of myocarditis in sudden death in children. *J Forensic Sci* **22**:188–96.

Northcote RJ, Ballantyne D. 1983. Sudden cardiac death in sport. *Br Med J Clin Res Ed* **287**:1357–9.

Olsen EG. 1987. *Atlas of cardiovascular pathology.* MTP Press, Lancaster.

Osborn GR. 1963. *The incubation period of coronary thrombosis.* Butterworth, London.

Otsuka N, Hara T. 1965. Gross demonstration of the mammalian atrioventricular bundle by a periodic acid-Schiff procedure. *Stain Technol* **40**:305–8.

Park HY, Weinstein SR. 1990. Sudden unexpected nocturnal death syndrome in the Mariana islands [see comments]. *Am J Forensic Med Pathol* **11**:205–7.

Patterson JC. 1938. Capillary rupture with intimal haemorrhage as a causative factor in coronary thrombosis. *Arch Pathol* **25**:474–87.

Paul M, Schulze-Bahr E, Breithardt G, *et al.* 2003. Genetics of arrhythmogenic right ventricular cardiomyopathy – status quo and future perspectives. *Z Kardiol* **92**:128–36.

Paz Suarez-Mier M, Aguilera B. 1998. Histopathology of the conduction system in sudden infant death. *Forensic Sci Int* **93**:143–54.

Pedersen PK. 1980. Determination of potassium/sodium ratio in heart tissue. Evaluation of its use as an index of myocardial ischaemic damage. Comparison with the nitro-BT test. *Forensic Sci Int* **16**:271–80.

Penttilä A. 1980. Sudden and unexpected natural deaths of adult males. An analysis of 799 forensic autopsies in 1976. *Forensic Sci Int* **16**:249–59.

Phillips LH, Whisnant JP, Reagan TJ. 1977. Sudden death from stroke. *Stroke* **8**:392–5.

Piano MR. 2002. Alcoholic cardiomyopathy: incidence, clinical characteristics, and pathophysiology. *Chest* **121**:1638–50.

Piro FR, di Gioia CR, Gallo P, *et al.* 2000. Is apoptosis a diagnostic marker of acute myocardial infarction? *Arch Pathol Lab Med* **124**:827–31.

Plueckhahn VD, Cameron JM. 1968. Traumatic 'myocarditis' or 'myocarditis' in trauma. *Med Sci Law* **8**:177–80.

Preston HV, Bowen DA. 1987. Asthma deaths: a review. *Med Sci Law* **27**:89–94.

Priemer F, Keil W, Kandolf R. 1999. Hydrocution in a case of coxsackie virus infection. *Int J Legal Med* **112**:368–71.

Ramkissoon RA. 1966. Macroscopic identification of early myocardial infarction by dehydrogenase alterations. *J Clin Pathol* **19**:479–81.

Rammer L, Jansson O. 1976. Determination of electrolytes in the myocardium as a tool for the post-mortal diagnosis of recent infarction. *Forensic Sci* **8**:127–30.

Rampazzo A, Beffagna G, Nava A, *et al.* 2003. Arrhythmogenic right ventricular cardiomyopathy type 1 (ARVD1): confirmation of locus assignment and mutation screening of four candidate genes. *Eur J Hum Genet* **11**:69–76.

Randall B. 1980. Fatty liver and sudden death. A review. *Hum Pathol* **11**:147–53.

Reichenbach D, Benditt EP. 1970. Myofibrillar degeneration: a common form of cardiac muscle injury. *Ann NY Acad Sci* **156**:164–76.

Richardson HL, Graupner KI, Richardson ME. 1966. Intramyocardial lesions in patients dying suddenly and unexpectedly. *JAMA* **195**:254–60.

Richardson P, McKenna W, Bristow M, *et al.* 1996. Report of the 1995 World Health Organization/ International Society and Federation of Cardiology Task Force on the Definition and Classification of Cardiomyopathies. *Circulation* **93**:841–2.

Rubin E. 1979. Alcoholic myopathy in heart and skeletal muscle. *N Engl J Med* **301**:28–33.

Rump AF, Theisohn M, Klaus W. 1995. The pathophysiology of cocaine cardiotoxicity. *Forensic Sci Int* **71**:103–15.

Rus HG, Niculescu F, Vlaicu R. 1987. Presence of C5b-9 complement complex and S-protein in human myocardial areas with necrosis and sclerosis. *Immunol Lett* **16**:15–20.

Rutenburg AM, Wolman M, Seligman AM. 1953. Comparative distribution of succinic dehydrogenase in six mammals and modification in the histochemical technic. *J Histochem Cytochem* **1**:66–81.

Sahai VB, Knight B. 1976. The post-mortem detection of early myocardial infarction by a simple fluorescent method. *Med Sci Law* **16**:17–20.

Sakurai S, Inoue A, Ohwa M, *et al.* 1995. [Immunohistochemical analysis of adhesion molecules in directional coronary atherectomy specimens.] *J Cardiol* **26**:139–47.

Sandritter M, Jedstadt R. 1958. Triphenyltetrazolium (TTC) als reduktionsindikator zur makroscopischen diagnose des frischen herzinfarktes. *Zent Allg Pathol* **97**:188–9.

Saphir O. 1942. Myocarditis: general review with analysis of 240 cases. *Arch Pathol* **33**:88–99.

Saphir O, Wile SA, Reingold IM. 1944. Myocarditis in children. *Am J Dis Child* **67**:294–312.

Saraste A, Pulkki K, Kallajoki M, *et al.* 1999. Cardiomyocyte apoptosis and progression of heart failure to transplantation. *Eur J Clin Invest* **29**:380–6.

Särkioja T, Hirvonen J. 1984. Causes of sudden unexpected deaths in young and middle-aged persons. *Forensic Sci Int* **24**:247–61.

Saukko P. 1983. Evaluation of diagnostic methods for early myocardial injury in sudden cardiac deaths. *Acta Univ Oul D 107, Anat Pathol Microbiol* **17**:53.

Saukko P, Knight B. 1989. Evaluation of eosin-fluorescence in the diagnosis of sudden cardiac death. *Forensic Sci Int* **40**:285–90.

Saukko P, Lignitz E. 1990. [Sudden death caused by malignant testicular tumors.] *Z Rechtsmed* **103**:529–36.

Schwender LA, Troncoso JC. 1986. Evaluation of sudden death in epilepsy. *Am J Forensic Med Pathol* **7**:283–7.

Seligman AM, Rutenburg AM. 1951. The histochemical demonstration of succinic dehydrogenase. *Science* **113**:317–20.

Sepulchre MA, Fechner G. 1992. [Detection of ischemic myocardial damage.] *Beitr Gerichtl Med* **50**:161–7.

Siboni A, Simonsen J. 1986. Sudden unexpected natural death in young persons. *Forensic Sci Int* **31**:159–66.

Simonsen J, Falk E. 1980. A case of sudden cardiac death in connection with *Salmonella typhimurium* infection. *Forensic Sci Int* **16**:283–7.

Simpson CK. 1947. Pathology of sudden death. *Lancet* **2**:745–7.

Siscovick DS, Weiss NS, Fletcher RH, *et al.* 1984. The incidence of primary cardiac arrest during vigorous exercise. *N Engl J Med* **311**:874–7.

Smith AJ. 1904–5. On the histological behaviour of the cardiac muscle in two examples of organization of myocardial infarct. *Univ Pennsylvania Med Bull* **17**:227–34.

Sokrab TE, Johansson BB, Kalimo H, *et al.* 1988. A transient hypertensive opening of the blood-brain barrier can lead to brain damage. Extravasation of serum proteins and cellular changes in rats subjected to aortic compression. *Acta Neuropathol (Berl)* **75**:557–65.

Song Y, Zhu J, Laaksonen H, *et al.* 1997. A modified method for examining the cardiac conduction system. *Forensic Sci Int* **86**:135–8.

Song Y, Laaksonen H, Saukko P, *et al.* 2001. Histopathological findings of cardiac conduction system of 150 finns. *Forensic Sci Int* **119**:310–17.

Sopher IM. 1974. Myocarditis and the aircraft accident. *Aerosp Med* **45**:963–7.

Speizer FE, Doll R, Heaf P, *et al.* 1968. Investigation into use of drugs preceding death from asthma. *Br Med J* **1**:339–43.

Speizer FE, Doll R, Heaf P. 1982. Observations on recent increase in mortality from asthma. *Br Med J* **285**:1251–5.

Stamer A. 1907. Untersuchungen über die Fragmentation und Segmentation des Herzmuskels. *Beitr Pathol* **42**:310–53.

Stevens PJ, Ground KE. 1970. Occurrence and significance of myocarditis in trauma. *Aerosp Med* **41**:776–80.

Stolley PD, Schinnar R. 1978. Association between asthma mortality and isoproterenol aerosols: a review. *Prev Med* **7**:519–38.

Tada T, Okada H, Okada N, *et al.* 1997. Membrane attack complex of complement and 20 kDa homologous restriction factor (CD59) in myocardial infarction. *Virchows Arch* **430**:327–32.

Teare D. 1958. Asymmetrical hypertrophy of the heart in young adults. *Br Heart J* **20**:1–8.

Thomas AC, Knapman PA, Krikler DM, *et al.* 1988. Community study of the causes of 'natural' sudden death. *Br Med J* **297**:1453–6.

Thomsen H, Held H. 1994. Susceptibility of C5b-9(m) to postmortem changes. *Int J Legal Med* **106**:291–3.

Thomsen H, Held H. 1995. Immunohistochemical detection of C5b-9(m) in myocardium: an aid in distinguishing infarction-induced ischemic heart muscle necrosis from other forms of lethal myocardial injury. *Forensic Sci Int* **71**:87–95.

Tonge JI, O'Reilly MJ, Davison A, *et al.* 1972. Traffic crash fatalities. Injury patterns and other factors. *Med J Aust* **2**:5–17.

Tonge JI, O'Reilly MJ, Davison A, *et al.* 1977. Traffic-crash fatalities (1968–73): injury patterns and other factors. *Med Sci Law* **17**:9–24.

Towbin JA, Bowles NE. 2002. The failing heart. *Nature* **415**:227–33.

Varga M, Zsonda L. 1988. A simple method for postmortem detection of acute myocardial infarction. *Forensic Sci Int* **37**:259–63.

Velisheva LS, Vikhert AM, Shvalev VN, *et al.* 1981. [Sudden cardiac death in alcoholic cardiomyopathy.] *Arkh Patol* **43**:32–7.

Wartmann WB. 1938. Occlusion of the coronary arteries by hemorrhage into their walls. *Am Heart J* **15**:459–66.

Wartmann WB, Hellerstein HK. 1948. The incidence of heart disease in 2000 consecutive autopsies. *Ann Intern Med* **28**:41–50.

Wartmann WB, Souders JC. 1950. Localization of myocardial infarction with respect to the muscle bundles of the heart. *Arch Pathol* **50**:329–38.

Whitehead R. 1965. Isolated myocarditis. *Br Heart J* **27**:220–4.

Willich SN, Lewis M, Lowel H, *et al.* 1993. Physical exertion as a trigger of acute myocardial infarction. Triggers and mechanisms of myocardial infarction study group [see comments]. *N Engl J Med* **329**:1684–90.

Wilson JD, Sutherland DC, Thomas AC. 1981. Has the change to beta-agonists combined with oral theophylline increased cases of fatal asthma? *Lancet* **1**:1235–7.

Wisten A, Forsberg H, Krantz P, *et al.* 2002. Sudden cardiac death in 15–35-year olds in Sweden during 1992–99. *J Intern Med* **252**:529–36.

World Health Organization. 1985. Sudden cardiac death. Report of a WHO Scientific Group. *WHO Tech Rep Ser* **726**:5–25.

Zhang JM, Riddick L. 1996. Cytoskeleton immunohistochemical study of early ischemic myocardium. *Forensic Sci Int* **80**:229–38.

Zugibe FT, Conley TL, Bell P, Jr, *et al.* 1972. Enzyme decay curves in normal and infarcted myocardium. *Arch Pathol* **93**:308–11.

CHAPTER 26

Forensic dentistry for the pathologist

- Bite marks 527
- Identification of the dead from the dentition 532
- Pink teeth 538
- References and further reading 538

Forensic odontology – the application of dentistry to forensic problems – is a discipline in itself and requires special dental expertise not possessed by pathologists qualified only in medicine. It is therefore imperative that, wherever possible, problems involving the teeth and jaws are referred to a dentist at an early stage, preferably at the time of the original autopsy or other examination of the subject.

A number of different circumstances need to be taken into account in this respect, however. First, it has to be accepted that by no means all dentists are either interested or competent in forensic problems. It is the same with pathologists, anatomists or any other specialists, in that the lack of enough motivation and experience can make their participation in legal matters of little value. A forensic odontologist naturally provides the best expertise, as he is professionally involved in the subject and will have training or experience, or both, of dento-legal problems, even though few forensic odontologists are occupied full time in this super-speciality, most being teachers or researchers in some other branch of dental science.

In many parts of the world, there are no forensic odontologists available. In developing countries there may be no forensic dentists at all – or the only expert is in the medical school of the capital city, inaccessible to distant areas. Even in more advanced states, problems of distance and travel may make it impracticable to call an odontologist to examine the body *in situ*. Dental expertise may be made available later, if material is retained for examination for transmission to the expert. In yet other circumstances, many problems of availability, cost and the lack of perceived importance of the case may make it impossible to involve an odontologist at an early stage – or even at all. In all these instances the pathologists or other medico-legal doctor may have no choice but to handle the dental aspects as best he can, either in the initial stages or even throughout the investigation.

This chapter is meant to be nothing more than a guide for the doctor placed in this position and is certainly not intended to encourage pathologists to attempt to replace or dispense with a good odontological opinion. It cannot be sufficiently emphasized that, where the issues are serious, every effort should be made to obtain expert forensic dental assistance. The remainder of this chapter must be looked on solely as a 'first-aid primer' for cases in which the ideal investigation cannot be attained for one reason or another. If no expert dental advice can be obtained at the outset, then the pathologist should do all he can to assemble and preserve both detailed records and physical specimens, which can be examined later by a forensic odontologist, if one can be found further along the investigative pathway.

The assistance that dentistry can render falls into two broad headings:

- the interpretation of bite marks
- personal identification, either individually or in the context of mass disasters.

BITE MARKS

One of the two major interests of forensic odontologists is one that has direct relevance to pathologists in that it concerns the interpretation of trauma to the body surface. For the sake of completeness, however, mention must first be made of bite marks on inanimate objects, usually foodstuffs.

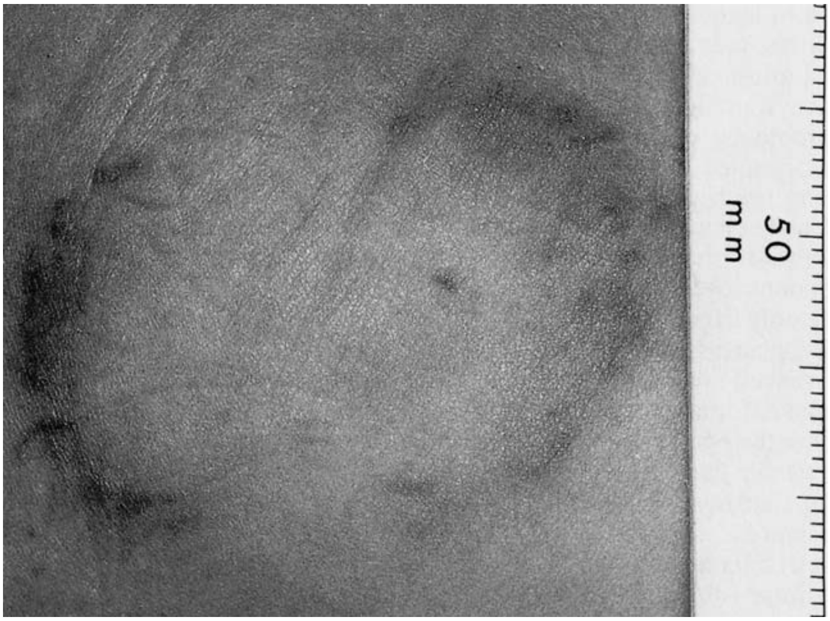

FIGURE 26.1 *Bite mark on skin. The two opposing bruises from the upper and lower dental arches are separated as usual by a gap, as only the front pans of the two arches contribute to the injury.*

This is mainly within the province of forensic science in the sense of 'criminalistics', where dental evidence is used to identify the perpetrators of a crime who happen to have left their teeth marks in some substance left at the scene. This rarely concerns forensic pathologists, except in the rare instance of their being asked by the police to examine an object, such as an apple or piece of cheese, to see if some obvious dental abnormality of a suspect could have caused the unusual bite mark.

The doctor may also be asked how the bite mark may be best preserved until the object can be examined by a forensic odontologist at a later date. Where the substance is 'plastic', such as butter, cheese, lard, wax or chocolate, for instance, it should be stored in a refrigerator to prevent melting or gradual flowing. It should not be deep frozen, as this may cause brittleness and cracking. Forensic science advice should be taken if at all available. Fruit, especially apples, seems prone to be bitten at scenes of crime, and Marshall *et al.* (1974) recommend preservation in Campden solution, a metabisulphite fluid used for fruit bottling. Alternatively, 5 per cent acetic acid in 40 per cent aqueous formaldehyde solution can be used. Ordinary histological formalin is not satisfactory – nor is refrigeration, which allows the fruit to shrivel. Whatever preservation is recommended to the police, the object should be adequately photographed with the film plane at right angles to the bite and a scale placed in the focal plane. If appropriate, swabbing for saliva traces should be carried out (see below) and then the fruit preserved for future examination.

Returning to the human body, bites are relatively common, especially in cases of child abuse and in adult sexual assault. In the former, bites may be anywhere on the infant, favourite sites being the arms, hands, shoulders, cheeks, buttocks and trunk. In the author's (BK) experience almost all bites on small children have been inflicted by the mother, but this is by no means invariable in the generality of child abuse. A common excuse from the parents is that the infant was bitten by another sibling or by the family dog – and occasionally, this explanation may be true. Others may be self-inflicted. It is therefore vital that the bite mark be properly examined to determine whether it is of a size consistent with adult dentition, or whether it is small enough to have come from another child – or is of a different shape, indicative of an animal.

The other common circumstances in which human bite marks occur are rape or other sexual offences. Here the pathologist will need to examine bites as part of the autopsy in a sexual murder – or even sometimes in live victims of an assault, if it is part of his duties to deal with clinical examinations or where no 'police surgeon' or forensic physician is available.

In this type of crime, bites may be sexually orientated or be distributed on any part of the body. Common sites are the breasts and nipples, but the neck, shoulders, thighs, abdomen, pubis and even vulva may be attacked. As noted in the chapter on sexual homicide, care must be taken to recognize that some so-called 'love bites', especially with suction petechiae, may be part of acceptable (if over-enthusiastic) sexual intercourse, but where real damage occurs, especially to breasts and nipples, then a violent or sadistic element is likely.

Bites may also be inflicted on police officers when attempting to arrest resisting offenders. They are also suffered in sporting events, especially football and some forms of wrestling, and during assaults when the victim manages to bite the assailant. In these instances bites may be

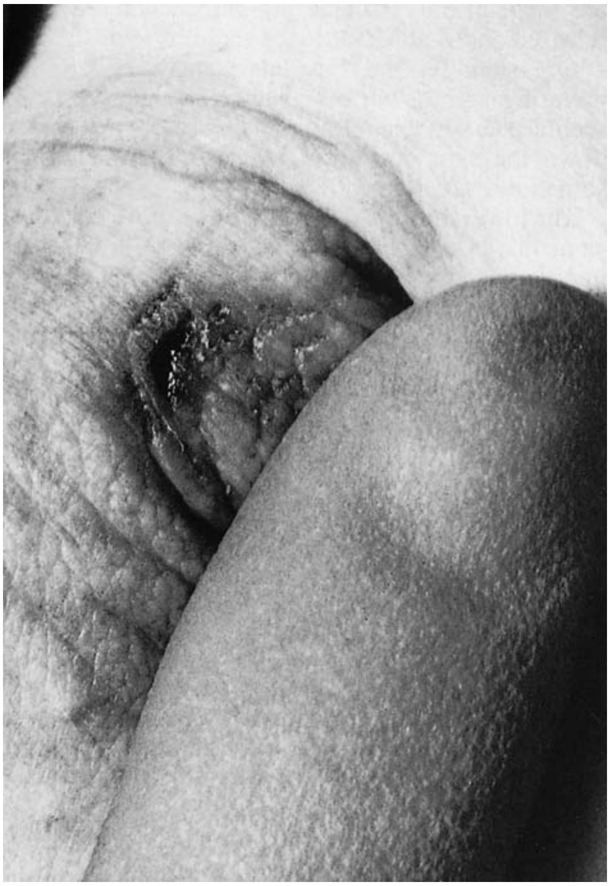

FIGURE 26.2 *Bite mark on a nipple during sexual homicide. The injury is too small for any dental matching, but a saliva sample might provide the blood group of the assailant, if he is a secretor.*

inflicted anywhere, but the hands, fingers, nose, forearms, ears and even lips may be the targets.

Some bite marks are self-inflicted; falls onto the face or a fit may cause the tongue and lips to be badly bitten. Other persons deliberately bite themselves, sometimes to fabricate injuries for a variety of motives ranging from gain to psychiatric disorder. Multiple bite marks (especially of the suction type), which are seen on accessible areas of the shoulder and arms, raise the suspicion of self-infliction, especially in older children and teenage girls.

The nature of the bite mark

Though called a 'bite mark', some of the components of the application of the mouth to skin may not be from the actual teeth. The lips can transiently mark the skin if forcibly nipped, especially on children, though the marks are short-lived and rapidly fade in life, not persisting after death unless associated with petechiae. Suction can produce a crop of punctate haemorrhages, either small petechiae or larger ecchymoses merging into a confluent central bruise.

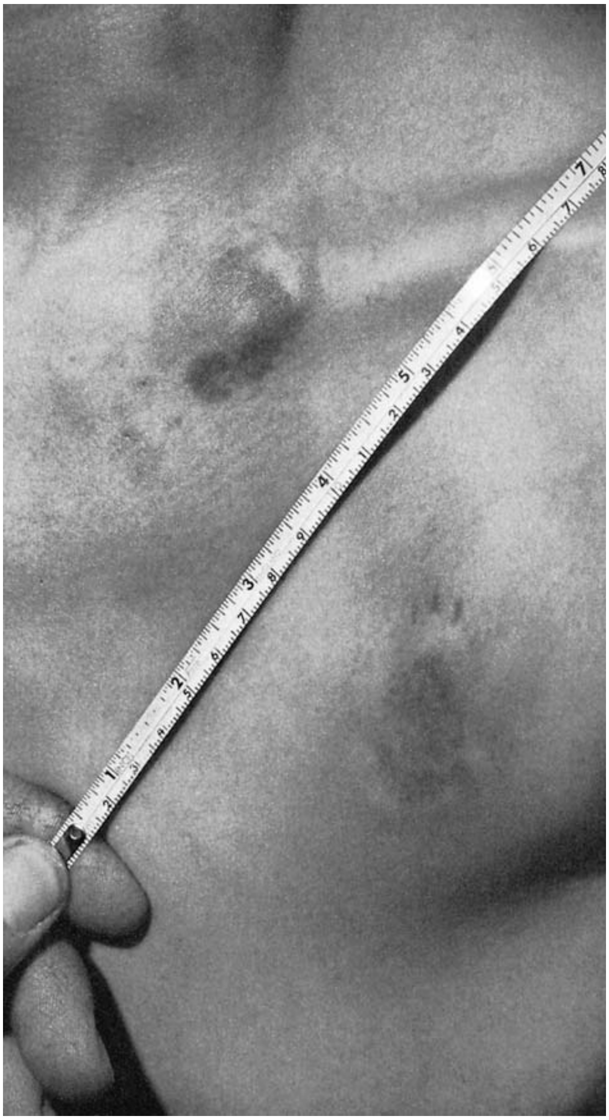

FIGURE 26.3 *Bite marks made during a sexual assault. The upper mark shows no specific features but its position is typical of a sexually orientated bite. The other shows central confluent petechiae and red bruising from suction and tongue pressure. There is a lower semi-circular pale zone corresponding with the lower lip, but the striking feature is the group of linear abrasions at the top of the mark made by the upper incisors dragging across the skin as the jaw is closed. Both bites are surrounded by zones of diffuse bruising.*

A human bite mark may present only a small part of the dental arcade, caused by the front teeth from canine to canine with an almost invariable gap at either side representing the separation of upper and lower jaw. A human bite is near circular or a shallow oval. A deep parabolic arch or a 'U-shape' can only be animal in origin. The teeth may cause clear, separate marks or they may run into each other to form a continuous or intermittently broken line. As time passes, originally clear tooth marks spread out and blur, progressively losing their definition – though abrasions on the skin surface retain their shape until scabbed healing is complete.

Teeth marks may be abrasions, bruises or laceration – or a combination of any two or three. The clarity of a bite mark depends on a number of factors. If the contour of the part bitten is irregular or markedly curved, then only part of the dental arch may contact the tissues. If the bite is forcible, then extensive subcutaneous bruising may spread laterally and blur the outline. If the bite was inflicted many days before, then healing of abrasions and lacerations, and absorption of bruising will leave progressively less detail. Where teeth have been forcibly applied, the typical appearance is of two 'bows' with their concavities facing each other and a gap at each end. Within this may sometimes be suction petechiae, which are often present without teeth marks, in the so-called 'love bite'. They are caused by the firm application of the lips, which form an airtight seal against the skin, then a sucking action reduces the air pressure over the centre. This causes a shower of petechial haemorrhages to appear from rupture of small venules in the superficial layers of the subcutaneous tissues. If forcible, the petechiae are confluent and a frank bruise, or even haematoma develops. Added to this is pressure from the tongue, pushing the tissues against the palate. This type of lesion is most often seen on the side of the neck and on the breasts of women, either in love play or after a sexual assault. They are by no means rare in child abuse, however, and are sometimes self-inflicted. Such a suction lesion must be human in origin and the claim that one is caused by a household pet can be immediately discounted.

Bites may be inflicted by the teeth closing down on a relatively flat skin surface, but there is usually an element of indrawing into the mouth, so that the teeth close down onto a parallel-sided or elliptical block of skin. In sexual bites, especially of the breast or nipple, the tissue may be actually sucked into the mouth before the jaws close upon it. This will naturally affect the shape of the resultant bite mark when the skin is released and flattens out once more.

Occasionally, the bite mark may not be two opposing arches, but more linear in pattern. This is seen especially where the upper incisors are scraped down the skin, leaving a series of parallel tracks, sometimes several centimetres long.

The lower incisors may leave either a curved line of static marks, or a straight or interrupted line below the upper incisor scrapes, as the lower teeth dig in and anchor the skin whilst the upper teeth gouge downwards towards them, during the act of closing the mouth.

The major problem with bite marks is the identification of the perpetrator and hence the problem cannot really be tackled without the expertise of a dentist fully experienced in this specialist task. Unless the dentition has some really characteristic features, especially in the incisors and canines, then a non-dentist is a poor witness if the matter comes to a legal dispute. Some odontologists would claim that a bite mark can only be identified in a negative sense – that is, suspects can be excluded if their dentition is obviously incapable of inflicting the bite mark under investigation, but that no positive match can be claimed. Though many forensic odontologists would dispute this conservative claim, it is probably the best attitude for a pathologist involved in such a problem, unless there is some really spectacular and unique feature in the front teeth of a suspect.

Missing teeth, grossly displaced teeth and substantially damaged teeth may suggest a match with a person – or the mark may be quite inconsistent, which is helpful to the investigators if they eliminate a person or persons, especially in the restricted range of a family in which a child has been abused.

It is probably best for a pathologist to realize his limitations in this field, however, and to confine himself to taking the best physical evidence for later study by a forensic odontologist, where this is at all possible.

The investigation of a bite mark

As stated, every effort should be made to obtain the best evidence for future specialist examination. First, the bite mark should be carefully and fully photographed. If police facilities are to be used, the pathologist – as in other aspects of forensic pathology – should direct them as to the important features required on film. Often such photographers, though otherwise expert in the technical aspects of their craft, do not appreciate the problems that inadequate or inappropriate views can cause – especially at a later date, when experts reviewing the case and lawyers presenting it in court have to rely on photographs without the benefit of having seen the original lesions.

The photographs should be taken from several different angles, but especially from a directly perpendicular viewpoint, with the plane of the film at right angles to that of the lesion. Some police photographers persist in taking tangential shots that foreshorten the true shape. An accurate scale should always be adjacent to the lesion, as close as possible, but not impinging upon it or obscuring any detail.

Bites are often on a curved surface, such as the face, breast or arm, and thus can never be reproduced exactly on a flat surface, as there is bound to be slight foreshortening at the ends, but several views at slightly different angles can overcome this problem. Small lens apertures and short focal lengths will obviate the blurring that results from the lesion curving out of the focal plane. Too short a focal length will itself produce image distortion.

The lighting is important, as perpendicular lighting may lead to a flat rendering with no capture of detail. Side

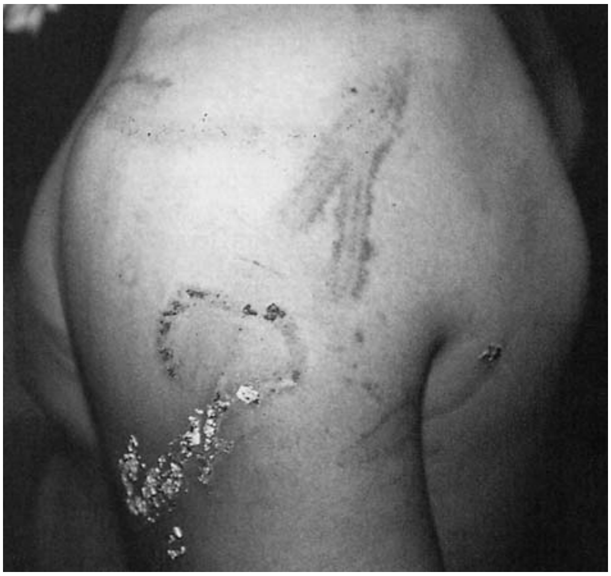

FIGURE 26.4 *A clear bite mark on the shoulder, a few days after infliction. This was from political abuse of human rights, there also being linear bruises on the shoulder and back from beating with a hose-pipe.*

lighting may throw small irregularities into relief, especially if there are tooth indentations in the skin. Both monochrome and colour photographs should be taken, with particular attention to sharp focus and correct exposure. The use of infrared-sensitive film has been recommended to reveal occult bruising, but it also appears to demonstrate artefacts. The lesion should almost fill the camera frame in some shots to capture as much detail as possible, either by short lesion-to-camera distance or the use of long-focus lenses. More general, wider shots should also be taken, however, to orientate the bite mark in relation to anatomical landmarks. During photography, care should be taken not to heat up the skin by the close proximity of high-power tungsten lamps. These should be placed at a distance or used only for short periods. Flash illumination will obviate this danger of heating the skin, which can distort the bite mark.

When photography is completed, swabs of the bite should be taken to try to recover saliva. Though not often successful, this technique can occasionally be vitally important in helping to identify or exclude the assailant, if he or she is one of the 80 per cent of people who are ‘secretors’, that is, who exude their blood-group substance in their saliva. Plain cotton-wool swabs are gently rubbed onto the bite – some experts recommend slightly moistening them first with water or saline. They should then be deep frozen unless sent straight to the serology laboratory.

Following photography and swabbing, there is usually little more that the pathologist can do in the absence of a dentist. If facilities are available and someone has the expertise, an impression of the bite can be made. This consists of laying a plastic substance over the bite mark, which then hardens, so as to produce a permanent negative cast of the lesion. It is usually made with a rubber- or silicone-based medium containing a catalytic hardener. Less satisfactory substances are water-based pastes, such as plaster of Paris, which are put on wet and allowed to dry before removal. These have the disadvantage of potential damage to the actual bite, if that is required for further evidential examination. Though it is unlikely that a pathologist will be able to carry out these procedures, it may be that a non-forensic dentist or a forensic scientist used to making casts of other evidence may be able to make satisfactory impressions that can then be preserved for later examination by an odontologist.

After autopsy, it is also possible for the whole area of skin carrying the bite to be removed and preserved in formalin for future examination. The shrinkage and distortion that are virtually inevitable, however, make these specimens of limited value for detailed tooth matching, though if good photographs with accurate scales are also available, they may be a useful addition to the dental evidence.

It is recommended that, as with all bruises and abrasions, the body be re-examined a few days after the first autopsy, as the appearances may be markedly enhanced. In a bite mark, faint or even absent original lesions may become more prominent, or even appear for the first time a day or two after death. The removal of skin carrying the lesion should be delayed if it is wished to see if this enhancement will occur, unless climatic conditions or lack of refrigeration would make this delay undesirable.

Matching the bite mark with the suspect’s dentition

For a pathologist, this is such a specialist enterprise that it would rarely be attempted, unless there was no prospect whatsoever of a forensic dentist being available, even at a later date. As mentioned earlier, unless there is a striking dental feature in the bite mark or in the teeth of the suspect, the best that can be done is to try to exclude a limited number of potential assailants on the basis of lack of correspondence of their dentition with the bite mark. If neither mark nor teeth show any particular distinctive appearances, then even this should be avoided, without expert dental opinion. Where circumstances dictate that only the pathologist will ever be available to help the investigation, he should confine himself to the following routines:

- The teeth of those who are either suspected by the police or who had access to the victim should be examined. In most jurisdictions it is vital that fully informed consent should be obtained from the person beforehand. The doctor must explain what he intends doing and the reason for doing it. It must be made clear that this is to be done for the purposes of a legal investigation and that any information gained might be used as evidence – and that it is not related to diagnosis or treatment for the subject's own welfare.

- Any refusal must be a bar to any further action. Where children are concerned, usually in the setting of child abuse, the consent of the fully informed parents or guardian must be obtained. All such consent is preferably obtained in writing and many police forces have special forms for the purpose. If not written, then a dubious second best is getting oral consent; at least one witness to this must be obtained, if possible from independent persons other than relatives or police officers.
- With consent, the dentition is then examined and the following points determined and recorded by diagram and writing. Photographs should also be taken if there is an issue of possible correspondence between the teeth and the bite – or where exclusion is legally important.
  - The presence of full or partial denture – and, if so, were they worn at the time of the incident?
  - The number of teeth in upper and lower jaw.
  - A charting of any missing teeth, especially incisors and canines.
  - An estimate of the bite overhang, or whether there is edge-to-edge occlusion or an undershoot projection of the lower teeth.
  - Recording of any broken teeth or teeth with significant individual abnormalities; these must be charted and described.
  - A record of any irregularity or marked variation in the cutting edge profile of any front teeth.
  - An evaluation of the size and prominence of any teeth, especially in the canines and incisors and any developmental abnormalities, such as an extra interposed front tooth.
  - Recording of any abnormality in the orientation of any tooth or teeth, such as twisting (rotation) in the anteroposterior tilting or double row of teeth. Gaps and irregular spacing are vital factors.

Though it is unlikely that a pathologist will have either the expertise or materials to make an impression of the bite, this might be attempted if no dental assistance of any kind is available – though the aid of a local dental practitioner would be useful in obtaining a wax or other bite registration. If this is impossible, then a bite impression into any available plastic substance might be attempted. The use of such substances as modelling clay, Plasticine, or beeswax might be better than nothing if some unique feature required demonstration. Finally, whether it be from visual inspection of the teeth or from comparison with a cast, an attempt should be made to compare the characteristics of the bite mark with the dentition of the suspect or suspects. It is assumed that an animal bite will have already been excluded by this stage.

This comparison can be performed in various ways, different odontologists having their own techniques. Where marked abnormalities exist in the teeth, these may be present in the bite mark if the affected teeth happen to have registered in the lesion. Some forensic dentists prefer to match photographs of the mark with photographs or tracings of the teeth, the former being printed to a 1:1 magnification. Tracings can be made from positive casts of a bite impression, inking the cutting edges of the front teeth and transferring these to transparent sheets, which can then be laid over the photographs to determine correspondence. Others use a negative photograph of the teeth laid over a positive photograph of the bite, again ensuring exact correspondence of magnification.

In a bite mark it is said that a tooth impression is better evidence than an absent mark – in other words, the presence of a tooth mark means that the tooth was actually in the jaw, whereas a gap in the injury could either mean that the tooth was missing or merely that the occlusal edge had not marked the skin, perhaps because of slight shortening, unevenness or wear. In all matching procedures between a bite mark and the suspect teeth, allowance must be made for distortion of the skin surface during the biting process and the angle of attack of the teeth. Again it is emphasized that exclusion is probably a much safer exercise than an allegation of positive matching, unless there is some unique feature in the dentition. Unlike fingerprints, fewer points of definite correspondence are required to claim correspondence. This applies only in the hands of experts, where the number needed for a confident identification depends on the strength of the idiosyncratic features, but may be inferred from three or four identical points.

In practice, it is the six upper and six lower front teeth that give the most information. The canines may provide particular help when prominent and pointed. Premolars and molars are rarely useful, as might be expected from their lower profile and posterior position in the jaws.

IDENTIFICATION OF THE DEAD FROM THE DENTITION

The major contribution of forensic odontology is in the field of identification, especially in mass disasters, such as aviation and marine catastrophes. In air crashes, dental investigation is the most successful single procedure leading to identification of mutilated and burned bodies, as the passenger manifest lists provide a circumscribed population for whom dental records can be obtained in the majority of cases. In such disasters the involvement of dentists is imperative, but the size of the subject is far too large to be attempted in a book of this nature.

Apart from mass casualties, forensic odontology is frequently used in problems of individual identity that are the direct concern of the forensic pathologist, as accident, suicide and murder form the majority of such unidentified bodies.

As with skeletal remains (of which dental evidence is part), discussed in Chapter 3, there are two prime avenues of investigation:

- general or reconstructive identity, which attempts to classify the unknown person by age, sex and race
- comparative methods, which confirm or exclude the personal identity of the individual against ante-mortem dental records.

Once again, the pathologist can only act for the dentist in a 'second-best' way, when no forensic odontological expertise is available. Unlike bite marks, the material is likely to be much more permanent and, wherever possible, the pathologist should retain photographs, chartings and even the actual dentition against the time when expert dental opinion might become available. When time presses in a criminal inves-tigation, however, or where no dental help is ever likely to be forthcoming, then the pathologist has to do the best he can, though many special techniques such as radiology or tooth sections may be quite beyond his capabilities.

General or reconstructive identity

Unlike skeletal remains, the human origin of dental material is rarely in doubt. In badly decomposed or skeletalized bodies the jaws usually survive intact, though in dry skeletons, teeth may become loose and fall out, especially the single-rooted canines and incisors.

Even in fragmented bodies and skeletons, the jaw remnants and teeth are readily recognizable, even by lay persons. Where teeth have dropped out and been recovered independently of a body, they are still usually recognizable as human, as opposed to most domestic or farm animals. In countries where large primates exist, there may be some confusion, but this is a rare problem.

Having established the human origin, the next determination is sex and here teeth have a poor discriminating value, though the intact jaw is more helpful, as discussed in Chapter 3. Male teeth are usually larger, but this is generally unhelpful. The difference in size between the upper lateral and upper central incisors is often greater in women, the male incisors being more equal in size. The female canines are usually smaller and more pointed relative to the male, more especially in the mandible than the maxilla. Girl's teeth tend to calcify and erupt earlier than boys. If the skeletal age is known, then more advanced tooth eruption in young persons is an indication of being female, though in these circumstances there are usually far better indicators of sex available elsewhere in the skeleton. The mandibular first molar often lacks a fifth cusp in the female, which is almost always present in the male. Extraction of pulp tissue from a tooth, even up to many months after extraction or death, can provide material for fluorescent staining for the female intranuclear F-body, as described in Chapter 3, but this has now been superseded by sex determination via DNA, if recoverable from the pulp.

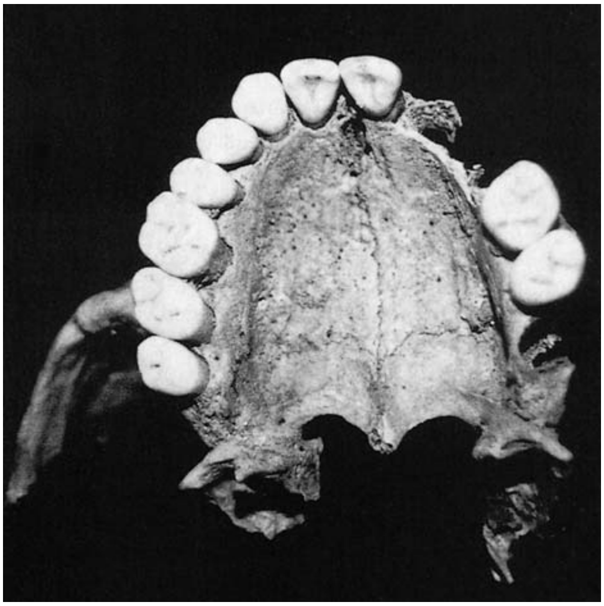

FIGURE 26.5 *Upper dental arch of a skeleton found buried on a small island in the Bristol Channel. The central incisors are 'shovel-shaped' and have a depressed groove on their posterior surface. This strongly suggests that the deceased was of mongoloid race.*

Race is also a difficult criterion to determine from teeth. The best-known feature is the 'shovel-shaped' upper central incisors of Mongoloid races, first described in Leipzig by Muhlreiter in 1870. The posterior surfaces of these teeth have a depression centrally, with two marginal bars, causing the back of the tooth to appear like a coal shovel with turned-up edges. The feature is found mainly amongst Chinese, Mongols, Eskimos and Japanese, but is also found amongst non-Mongoloid races in lesser numbers. Some 91 per cent of Chinese, Japanese and Tibetans have such teeth, 95 per cent of Native Americans, 84 per cent of Eskimos, 46 per cent of Palestinian Arabs – and 90 per cent of Finns. It is rare amongst Negroids and Australian Aboriginals. In Caucasian races, the lateral incisors in the upper jaw are usually smaller than the central, especially in women, a feature absent or less marked in Negroid or Mongoloid races. Caucasians also have long pointed canine roots, a feature not seen in Mongoloids. Enamel pearls, small nodules of enamel on the tooth surface, are much more frequent in Mongoloid teeth. Small nodules on the lingual surface of maxillary molars, called 'Carabelli's cusp', are most common in Caucasian races and rare in the other major racial groups. The condition of bull-tooth or 'taurodontism' is most common in Mongoloid peoples: here the pulp cavity of molars is wide and deep, and the roots are fused and bent. A congenital lack of the third upper molar is most common in Mongoloids, but can occur in any race. Negroid races tend to have large teeth and often have more cusps on their molars, even up to eight, with two lingual cusps on the mandibular first premolars as an additional common finding.

The age of the person is one of the most useful findings disclosed by the teeth, especially in the first two decades of life. The sequence of deciduous or 'milk' teeth is well known, overlapping the appearance of the permanent dentition, shown in Figure 26.6. This is only an average timetable, however, and is modified by several factors, such as sex, race and climate. Again, dental expertise is needed to refine the accuracy of such estimations. The determination of age from fetal teeth is also a matter for embryologists or dentists with specialized knowledge of this period. After the third molar has erupted in the third decade of life, then age determination becomes much more difficult. Much research has been expended in forensic odontology and the name of Gustafson (see references and further reading) is well known in this respect. His criteria for age in adult life comprised six factors:

- occlusal attrition of the tip of the tooth
- secondary dentine deposition in the apex of the pulp cavity
- apical migration of the attachment of the periodontal membrane
- increase in root transparency – the best single indicator
- root resorption
- accumulation of cementum around the root.

This method, as later modified by Johanson (1971), is said to give an age accuracy within 5 years either side of the true age.

It seems pointless to rehearse these methods in advice to pathologists, however, as specialized techniques, equipment and knowledge are needed for these procedures. The standard textbooks on forensic odontology and original papers should be consulted for the details. Returning to general features of age, obvious pointers are the state of the teeth in respect

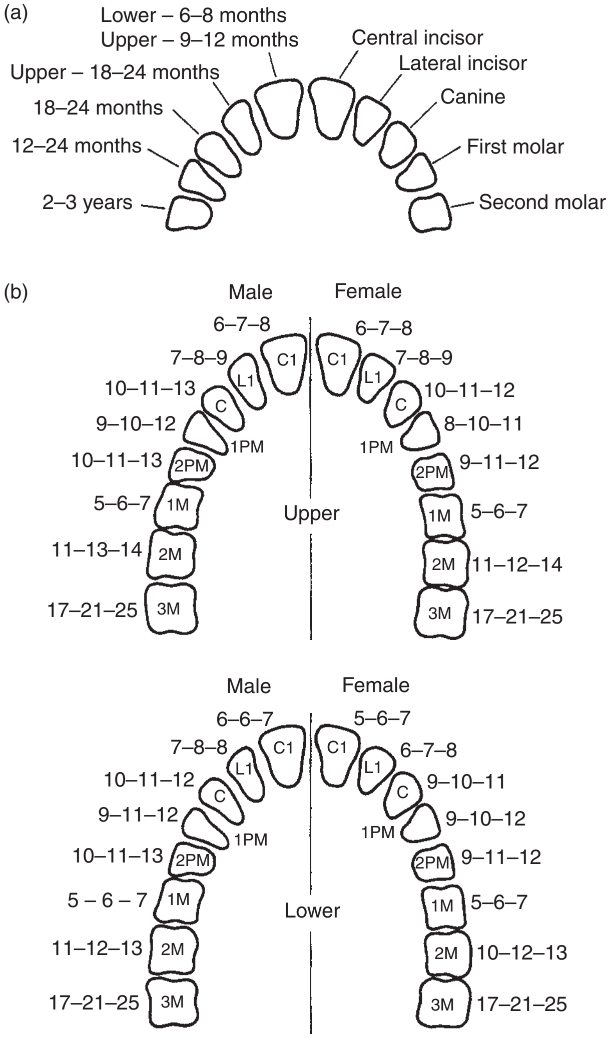
(a)
Lower – 6–8 months
Upper – 9–12 months
Central incisor
Lateral incisor
Upper – 18–24 months
18–24 months
Canine
12–24 months
First molar
2–3 years
Second molar
(b)
Male
Female
6–7–8
6–7–8
7–8–9
7–8–9
C1
C1
L1
L1
10–11–13
10–11–12
C
C
9–10–12
8–10–11
1PM
1PM
10–11–13
2PM
2PM
9–11–12
5–6–7
1M
Upper
1M
5–6–7
11–13–14
2M
2M
11–12–14
17–21–25
3M
3M
17–21–25
Male
Female
6–6–7
5–6–7
7–8–8
6–7–8
10–11–12
9–10–11
9–11–12
9–10–12
1PM
1PM
10–11–13
2PM
2PM
9–11–12
5 – 6 – 7
1M
Lower
1M
5–6–7
11–12–13
2M
2M
10–12–13
17–21–25
3M
3M
17–21–25

FIGURE 26.6 *(a) Dates of eruption of deciduous teeth (average times). (b) Times of eruption of permanent teeth; three dates are shown (years) for early, average and late eruption, rounded to the nearest year.*

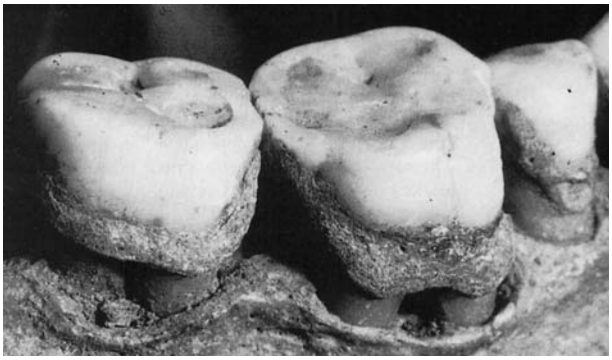

FIGURE 26.7 *Severe occlusal attrition in a jaw from the early nineteenth century. The crowns are worn down to expose the dentine by a rough diet that probably contained stone dust in the flour from contemporary milling methods.*

of wear, hygiene and colour, which may deteriorate with advancing age. Much depends upon the care with which they have been maintained, however, though marked occlusal attrition tends to go with increasing age, unless a rough diet has accelerated the wear. In Western Europe, gross occlusal attrition, sometimes down to gum level, is seen in old skeletal material. This usually indicates that the bones and teeth came from someone alive in the mid-nineteenth century or earlier, before modern milling methods removed abrasive stone dust from flour.

Edentulous jaws also suggest advancing age but, especially in former years before more conservative dentistry, even young adults often had total tooth clearances for caries. Once the teeth have gone, there is a general atrophy of the alveolar margins, but this is a poor criterion of age because of the great variability in the time when teeth are lost.

Newer techniques for age estimation include the variation with age of racemization of amino acids, especially aspartic acid, but this is a very specialized area.

Comparative identification from teeth

Establishing personal identity requires the matching of observed features with pre-existing dental records, the latter almost always obtained from previous diagnostic and therapeutic surveillance. For this method to be applied, there must be:

- Some collateral evidence to indicate either who the unknown body might be, so that records can be sought.
- Alternatively, a circumscribed population must be searched for records that may match the unknown. Such a population may be the known passengers on an aircraft or ship or a cohort of missing persons maintained on some register. It is manifestly impossible to search a large population, such as a whole country or even a city. Attempts were made in Britain to computerize dental records from the National Health Service so that a wide search could be made, but the completeness and quality of data was such that the scheme was found to be impracticable.
- The unknown person must have had dental attention in the past. The dentist or hospital must be known; the records must be traceable and, when found, must contain sufficient clinical information to provide adequate identifying features. The recovery of dental radiographs is a most useful adjunct. Unfortunately, these criteria are not always satisfied.

The dental records are needed to provide a description of the dentition and jaws at a date as late as possible before the finding of the body. Work may have been done on the deceased during life since the last record was made, if the person had been treated elsewhere and the information not recorded. Where a record does not conform to the dental state of a body who was expected to be a match, it is obvious that such discrepancies can be of two types:

1. If the record indicates some condition that is irreversible but which does not exist in the body, then that excludes matching. For example, if the record states that certain teeth have been extracted, yet they are still present in the jaws, then any hope of correspondence must be abandoned.
2. If fillings are present in the teeth, which are not shown in the records, then these may have been made later and not recorded. Of course, no discrepancies of point 1 must be present for those of point 2 to be acceptable.

Charting the teeth

In the absence of a forensic odontologist, the next best person to record the state of dentition of a dead body would naturally be a dental practitioner, to whom this task would be an everyday routine. The chart made by him could then be used for future reference in any identification procedure. This chapter is concerned with the imperfect yet sometimes inevitable circumstance where the pathologist is the only person available. Unfortunately, there are a host of different methods of charting the contents of the jaws, and, in spite of sustained efforts, no universal internationally accepted system has yet been adopted. For our purposes, any careful record of the number, position and state of the teeth made by a pathologist can be converted into whatever system of charting is required at a later stage.

Two main methods of recording data are in use: the first is a diagrammatic chart commonly used in large dental services such as the British National Health Service, seen in Figure 26.8. The other system uses a more pictorial representation of the teeth, so that all features can be recorded in a more exact topographical manner. Both systems, as well as many variants, have a notation which describes the position of the teeth, almost always in four quadrants, right upper, left upper, right lower and left lower. Unfortunately, there is considerable variation in the sequence of numeration, especially where the transition points from left to right, and from upper to lower are concerned, these being confusing and often not interchangeable. The pathologist need not concern himself too much with this problem, as his graphic representation can always be turned into whatever system of numeration is required at a later time. For example, if an unknown body has its teeth charted on one

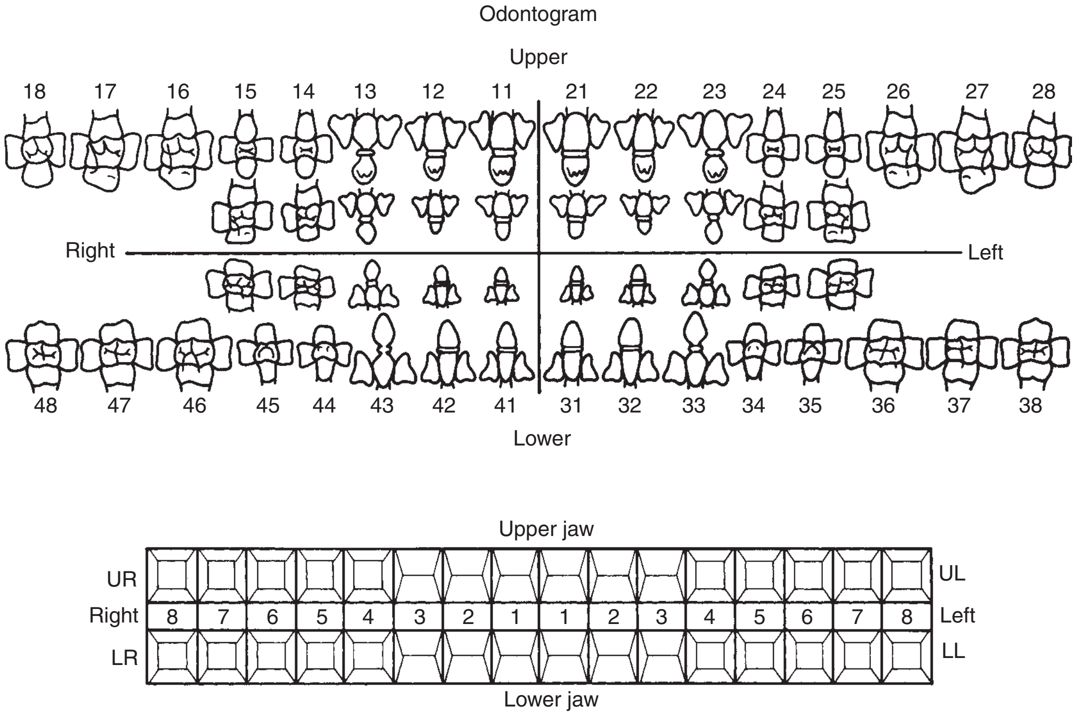

FIGURE 26.8 *Two of the many types of chart used for recording dental characteristics. The upper is the 'Odontogram' designed for Interpol. The chart displays each surface of each tooth, including the deciduous teeth. The lower chart is that used in the British National Health Service, which is a simpler design. In both charts the buccal aspects are at the top and bottom and the lingual sides of the teeth face each other centrally.*

of the common diagrams and later the dental records of a potential 'identitee' become available, the chart can be transposed into the same system as the clinical records. The notation derived from this can then be transmitted by telephone, telex, fax or other form of communication, if the records are not available at the place where the body lies. This is particularly important in transportation fatalities such as air crashes, where the home base of the victims may be far removed from the scene of death.

On the diagrammatic or anatomical charts, each tooth is represented by a pictorial symbol that provides the same number of tooth surfaces as those on the same teeth in the mouth. The incisors and canines have four surfaces represented, while the premolars and molars have an additional facet, the occlusal surface. On these diagrams, the positions of fillings, crowns, caries and damage are marked, and, of course, missing teeth are deleted.

In charting the teeth, the pathologist needs good access to the mouth and this presents the first problem. Rigor mortis in a relatively fresh cadaver may make it impossible to open the mouth without great effort. Excessive force should not be used, especially leverage with a metal instrument, because of the danger of damaging the teeth. Where the rigor cannot be broken by sustained firm pressure on the chin – or where time prevents waiting for rigor to pass off – it may be necessary to extend the autopsy incision into a neck 'V' and dissect the skin off the lower part of the face, to gain access to the masseter and temporalis muscles, which can then be divided above their insertion into the mandible, to allow the jaw to become mobile. Care must be taken not to disfigure the face during this procedure.

When the body is decomposing, no rigor will be present. If badly rotted, then cosmetic considerations will not apply and more radical removal of the jaws may need to be performed. The same holds for badly burned bodies, where heat contraction of the facial muscles may make it impossible to open the jaws without dissecting away tissues. In both these instances, where identity is a prime consideration, both mandible and maxilla may need to be removed for retention and later examination. The mandible can be disarticulated at the temporomandibular joints and removed intact. The lower maxilla, comprising the tooth-bearing jaw, the palate and the inferior part of the facial skeleton, can be removed after the mandible has been taken. This is done by sawing horizontally across the maxilla at the level of the lower margin of the nasal aperture, taking care to saw above any tooth

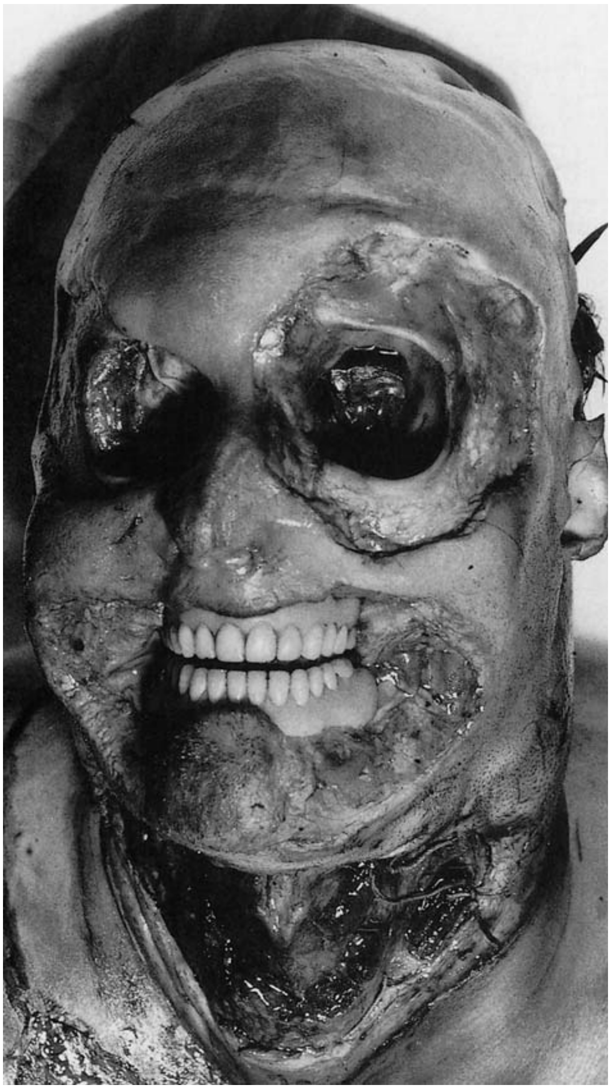

FIGURE 26.9 *Use of dentures to identify a decomposed body from the sea. Some dental prostheses carry makers' or patients' record numbers embedded in the plate material. The loss of tissue around eyes and mouth was caused by marine predators. The cut throat was made by a nylon hawser that threw this ship's officer overboard.*

roots. After photography and charting, the jaws can be preserved in formalin or a freezer. With modern serological and DNA identification techniques, the need to keep some unfixed tissue for serology and other investigations should be borne in mind. Naturally, the presence of any full or partial dentures or any other type of dental prosthesis will have been recorded and the prosthesis carefully retained for examination. Even edentulous persons may show signs of having worn dentures by the presence of pressure marks on gums or palate. Wherever possible, radiographs should be taken of the head of an unidentified body (as discussed in Chapter 3) where non-dental considerations apply, such as craniometry or frontal sinus visualization. At the same time, the radiographs may show unexpected foreign bodies in the head and some of these may have unique dental importance, such as broken drill, broken roots or congenital abnormalities, which may be matchable with clinical films, or a note in the ante-mortem records. Radiography has a special place in forensic odontology, but specialist techniques and knowledge are likely to lie outside the abilities of a pathologist.

When, by whatever means necessary, the dental arches have been made available for examination, the following features are sought and recorded on the chart:

- Extractions – whether recent or old – should be noted from the condition of the socket.
- Fillings – their number, position and composition.
- Artificial teeth – gold, porcelain or stainless steel.
- Other prosthetic work in the mouth, such as bridgework or braces.
- Crowned teeth.
- Broken teeth.
- Pathological conditions in teeth, jaws or gums.
- Congenital defects such as enamel pearls, Carabelli's cusps or ectopic teeth.
- Malpositioned teeth – rotated or tilted for example.
- The general state of care and hygiene including caries, plaque, tobacco staining and gingivitis.
- Racial pointers, such as shovel-shaped upper central incisors or multi-cusped molars.

When all available information has been discovered and entered on the chart, the process of comparison with any ante-mortem records can be made. Much will depend on the quality and the date of these records, which are frequently less detailed and exact than the autopsy charting. Allowance must be made for the clinical records to be substantially out of date on occasions, as mentioned earlier. The process of matching is again really the work of an experienced dentist, but the major task of checking the above list against the previous records can be carried out adequately in respect of missing teeth, filling, prostheses and major trauma and other anomalies.

Anything on the clinical chart that is not represented in the actual jaws almost always excludes a matching – fillings cannot go away and extracted teeth cannot return, even though some remarkable feats of dental treatment now exist. The degree of correspondence that is acceptable for a positive match is a matter of common sense, and naturally other non-dental aspects must be taken into account – it is useless trying to claim a match on a woman's dental characteristics if other information clearly indicates that the missing person was a man. Finally, and at the risk of being over-repetitive, it must again be emphasized that, though the pathologist may be a collector of data and physical evidence, where the issues

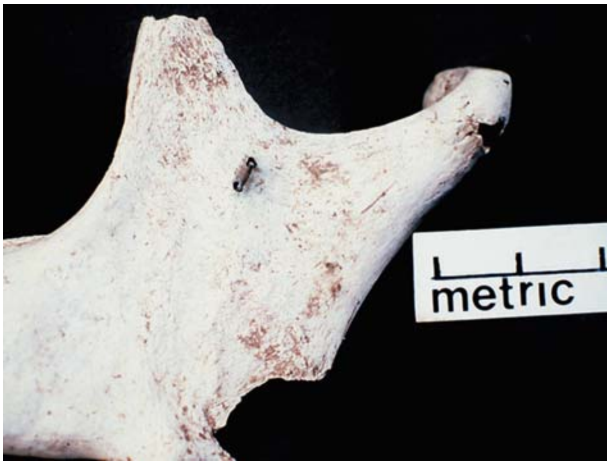

FIGURE 26.10 *A short-cut in excluding homicide. A skeleton was found buried in a suburban garden. Initial suspicions were allayed when the ramus of the mandible was seen to carry a bronze wire for attaching a spring to suspend the jaw from the skull. Police enquiries confirmed that the house had previously belonged to a member of the medical school teaching staff.*

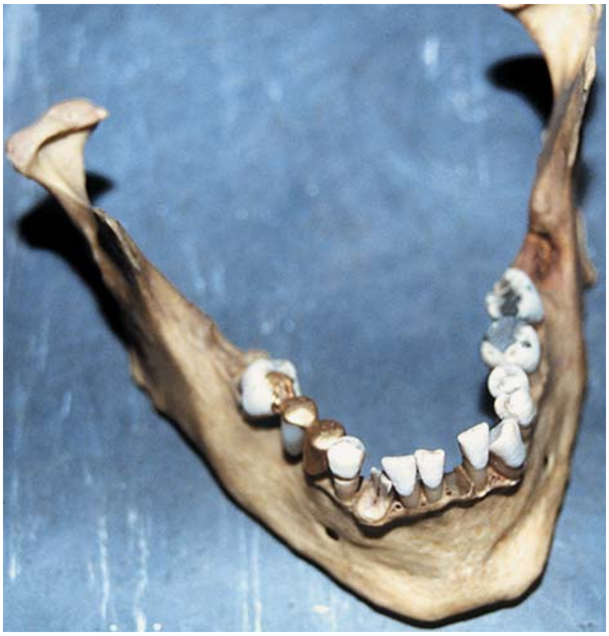

FIGURE 26.11 *A mandible from a skeletalized body found in suspicious circumstances. Identity was established solely from dental features, such as the gold bridge and acrylic tooth on the right side of the jaw and the numerous amalgam fillings. Comparison of dental radiographs obtained from a hospital confirmed the correspondence. The jaw, with the prominent square symphyseal region, is typically male.*

are vital (as in a homicide investigation), every effort must be made for the material to be evaluated by a forensic odontologist, even if the details have to be sent by air mail to another country.

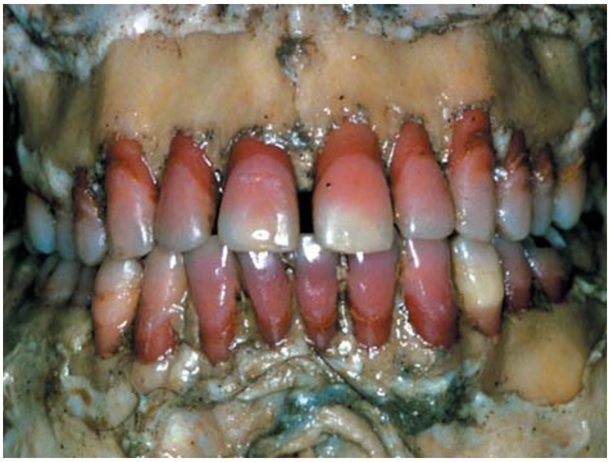

FIGURE 26.12 *Pink teeth. The jaw came from a decomposed body dumped in the countryside after death from narcotic overdose. No carbon monoxide was present in the body, the pinkness being caused by haemoglobin products staining the dentine. (Reproduced by kind permission of Professor David Whittaker.)*

PINK TEETH

At autopsy on putrefied bodies, the teeth are sometimes seen to be a marked pink colour, especially near the gum line. Considerable research has been carried out into this interesting phenomenon and it is clear that the colour is the result of the dentine being stained by haemoglobin products. It was once thought that carbon monoxide played a part in the coloration, but this has been disproved. Claims that the phenomenon is a pointer to an asphyxial death can be discounted, though possibly a very 'congestive' mode of death – albeit a very non-specific condition – encourages pinkness because of retention of blood in the pulp.

REFERENCES AND FURTHER READING

Ajmal M, Mody B, Kumar G. 2001. Age estimation using three established methods. A study on Indian population. *Forensic Sci Int* **122**:150–4.

Bernitz H. 2001. Identification by means of denture marking. *SADJ* **56**:368–9.

Bernitz H, Kloppers BA. 2002. Comparison microscope identification of a cheese bitemark: a case report. *J Forensic Odontostomatol* **20**:13–16.

Borrman HI, DiZinno JA, Wasen J, *et al.* 1999. On denture marking. *J Forensic Odontostomatol* **17**:20–6.

Bowers CM, Johansen RJ. 2002. Photographic evidence protocol: the use of digital imaging methods to rectify angular distortion and create life size reproductions of bite mark evidence. *J Forensic Sci* **47**:178–85.

Bowers CM, Johansen RJ. 2002. Digital imaging methods as an aid in dental identification of human remains. *J Forensic Sci* **47**:354–9.

Brkic H, Strinovic D, Kubat M, *et al.* 2000. Odontological identification of human remains from mass graves in Croatia. *Int J Legal Med* **114**:19–22.

Brondum N, Simonsen J. 1987. Postmortem red coloration of teeth. A retrospective investigation of 26 cases. *Am J Forensic Med Pathol* **8**:127–30.

Cameron JM, Sims BG. 1973. *Forensic dentistry.* Churchill Livingstone, Edinburgh.

Commission on Armed Forces Dental Services. 1967. *Identification by dental means*, Vol. Redaction I–IV. Federation Dentaire International, Paris.

DeVore DT. 1977. Radiology and photography in forensic dentistry. *Dent Clin North Am* **21**:69–83.

Du Chesne A, Benthaus S, Brinkmann B. 1999. Manipulated radiographic material – capability and risk for the forensic consultant? *Int J Legal Med* **112**:329–32.

English WR, Robison SF, Summitt JB, *et al.* 1988. Individuality of human palatal rugae. *J Forensic Sci* **33**:718–26.

Fearnhead RW. 1961. Facilities for forensic odontology. *Med Sci Law* **1**:273–6.

Fielding CG. 2002. Nutrient canals of the alveolar process as an anatomic feature for dental identifications. *J Forensic Sci* **47**:381–3.

Forbes G, Watson AA. 1975. *Legal aspects of dental practice.* John Wright, Bristol.

Furness J. 1976. Positive dental identification. *Int J Forensic Dent* **3**:2–5.

Furness J. 1981. A general review of bite-mark evidence. *Am J Forensic Med Pathol* **2**:49–52.

Garn SM. 1958. The sex difference in tooth calcification. *J Dent Res* **37**:561–4.

Glass RT, Jordan FB, Andrews EE. 1975. Multiple animal bite wounds: a case report. *J Forensic Sci* **20**:305–14.

Gustafson G. 1950. Age determination on teeth. *J Am Dent Assoc* **41**:45–50.

Gustafson G. 1966. *Forensic odontology.* Staples Press, London.

Haines DH. 1972. Racial characteristics in forensic dentistry. *Med Sci Law* **12**:131–8.

Hanihara K. 1967. Racial characteristics in the dentition. *J Dent Res* **46**:923–6.

Harvey W. 1976. *Dental identification and forensic odontology.* Henry Kimpton, London.

Heliman H. 1929. Racial characteristic in the human dentition. *Am Phil Soc* **67**:157–74.

Hurme VO. 1948. Standards of variation in the eruption of the first six permanent teeth. *Child Dev* **19**:213–22.

Ito S. 1976. Age determination based on tooth crown. *Int J Forensic Dent* **3**:9–14.

Jakobsen JR, Keiser Nielsen S. 1981. Bite mark lesions in human skin. *Forensic Sci Int* **18**:41–55.

Johanson G. 1971. Age determination from human teeth. *Ondontol Rev* **22**(Suppl):22–8.

Kemkes-Grottenthaler A. 2001. The reliability of forensic osteology – a case in point. Case study. *Forensic Sci Int* **117**:65–72.

Kolltveit KM, Solheim T, Kvaal SI. 1998. Methods of measuring morphological parameters in dental radiographs. Comparison between image analysis and manual measurements. *Forensic Sci Int* **94**:87–95.

Koshy S, Tandon S. 1998. Dental age assessment: the applicability of Demirjian's method in South Indian children. *Forensic Sci Int* **94**:73–85.

Lund H, Mornstad H. 1999. Gender determination by odontometrics in a swedish population. *J Forensic Odontostomatol* **17**:30–4.

Luntz L, Luntz P. 1973. *Handbook for dental identification: techniques in forensic dentistry.* Lippincott, Philadelphia.

Maples WR. 1978. An improved technique using dental histology for estimation of adult age. *J Forensic Sci* **23**:764–70.

Marshall W, Potter J, Harvey W. 1974. Bite marks in apples – forensic aspects. *Criminol* **9**:21–6.

Masters PM. 1986. Age at death determinations for autopsied remains based on aspartic acid racemization in tooth dentin: importance of postmortem conditions. *Forensic Sci Int* **32**:179–84.

McKenna CJ, Haron MI, Taylor JA. 1999. Evaluation of a bitemark using clear acrylic replicas of the suspect's dentition – a case report. *J Forensic Odontostomatol* **17**:40–3.

Meredith HV. 1946. The order and age of eruption of the deciduous dentition. *J Dent Res* **25**:43–5.

Mertz CA. 1977. Dental identification. *Dent Clin North Am* **21**:47–67.

Metzger Z, Buchner A. 1980. The application of tetracyclines in forensic dentistry. *J Forensic Sci* **25**:612–18.

Miles AEW. 1963. Dentition in the estimation of age. *J Dent Res* **42**:253–5.

Muller M, Lupi-Pegurier L, Quatrehomme G, *et al.* 2001. Odontometrical method useful in determining gender and dental alignment. *Forensic Sci Int* **121**:194–7.

Myers JC, Okoye MI, Kiple D, *et al.* 1999. Three-dimensional (3-D) imaging in post-mortem examinations: elucidation and identification of cranial and facial fractures in victims of homicide utilizing 3-D computerized imaging reconstruction techniques. *Int J Legal Med* **113**:33–7.

Nambiar P, Carson G, Taylor JA, *et al.* 2001. Identification from a bitemark in a wad of chewing gum. *J Forensic Odontostomatol* **19**:5–8.

Ogino T, Ogino H, Nagy B. 1985. Application of aspartic acid racemization to forensic odontology: postmortem designation of age at death. *Forensic Sci Int* **29**:259–67.

Ohtani S. 1994. Age estimation by aspartic acid racemization in dentin of deciduous teeth. *Forensic Sci Int* **68**:77–82.

Ohtani S, Yamada Y, Yamamoto I. 1998. Improvement of age estimation using amino acid racemization in a case of pink teeth. *Am J Forensic Med Pathol* **19**:77–9.

Pretty IA, Smith PW, Edgar WM, *et al.* 2002. The use of quantitative light-induced fluorescence (QLF) to identify composite restorations in forensic examinations. *J Forensic Sci* **47**:831–6.

Ritz-Timme S, Cattaneo C, Collins MJ, *et al.* 2000. Age estimation: the state of the art in relation to the specific demands of forensic practice. *Int J Legal Med* **113**:129–36.

Rogers SL. 1988. *The testimony of teeth: forensic aspects of human dentition.* Thomas, Springfield.

Ruddick RF. 1974. A technique for recording bite marks for forensic studies. *Med Biol Illus* **24**:128–9.

Sheasby DR, MacDonald DG. 2001. A forensic classification of distortion in human bite marks. *Forensic Sci Int* **122**:75–8.

Sims BG, Grant JH, Cameron JM. 1973. Bite-marks in the ‘battered baby syndrome’. *Med Sci Law* **13**:207–10.

Sognnaes RF. 1980. Hitler and Bormann identifications compared by postmortem craniofacial and dental characteristics. *Am J Forensic Med Pathol* **1**:105–15.

Solheim T. 1980. Unusual dental forensic cases in Norway. *Am J Forensic Med Pathol* **1**:197–203.

Solheim T. 1993. A new method for dental age estimation in adults. *Forensic Sci Int* **59**:137–47.

Stevens PJ, Tarlton SW. 1966. Medical investigation in fatal aircraft accidents. The role of dental evidence. *Br Dent J* **120**:263–70.

Sweet D, Shutler GG. 1999. Analysis of salivary DNA evidence from a bite mark on a body submerged in water. *J Forensic Sci* **44**:1069–72.

Tramini P, Bonnet B, Sabatier R, *et al.* 2001. A method of age estimation using raman microspectrometry imaging of the human dentin. *Forensic Sci Int* **118**:1–9.

Valenzuela A, Martin-de las Heras S, Marques T, *et al.* 2000. The application of dental methods of identification to human burn victims in a mass disaster. *Int J Legal Med* **113**:236–9.

Whittaker D, McRonald R. 1989. *Atlas of forensic dentistry.* Wolfe Medical Publications, London.

Whittaker DK, Brickley MR, Evans L. 1998. A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis. *Forensic Sci Int* **92**:11–20.

Whittaker DK, Richards BH, Jones ML. 1998. Orthodontic reconstruction in a victim of murder. *Br J Orthod* **25**:11–14.

Wright FD. 1998. Photography in bite mark and patterned injury documentation – part 2: a case study. *J Forensic Sci* **43**:881–7.

Xu XH, Philipsen HP, Jablonski NG, *et al.* 1991. Preliminary report on a new method of human age estimation from single adult teeth. *Forensic Sci Int* **51**:281–8.

Yaacob H, Nambiar P, Naidu MD. 1996. Racial characteristics of human teeth with special emphasis on the mongoloid dentition. *Malays J Pathol* **18**:1–7.

CHAPTER 27

Poisoning and the pathologist

- The concept of the fatal dose 541
- Variations in published fatal poison concentrations 542
- The autopsy in suspected poisoning 543
- Poisoning deaths in hospital 543
- The collection of autopsy samples for toxicological analysis 543
- Containers for toxicology samples 545
- Preservatives in samples 545
- The site of sampling for toxicological analysis 546
- The technique of obtaining autopsy samples 547
- References and further reading 550

Toxicology has a number of different aspects and, in such a huge subject, various specialists have different interests. The clinical toxicologist is most concerned with diagnosis and treatment of the living patient; the analytical toxicologist has the complex task of laboratory investigation; and the pathologist is concerned with evaluating poisons as a cause or a contribution to death. Though obviously linked, these various aspects are substantially different and few people can claim to be proficient in all three. As far as the pathologist is concerned, his main task is to exclude or confirm other non-toxic factors in the death. He has then to collect suitable samples for analysis and, when the laboratory results are available, to interpret them in the light of his knowledge of the history, clinical features and autopsy appearances.

The pathologist inevitably needs the expertise of the laboratory analyst and the latter's knowledge of the therapeutic, toxic and fatal levels of the substances under consideration. Such data must, however, be evaluated in the knowledge of other pathological and physiological conditions present, so that it is the pathologist, rather than the laboratory toxicologist, who should provide the final opinion upon the proximate cause of death. This does not always happen and some laboratory report forms may be seen that unequivocally – and unwisely – state that a particular drug **caused** the death.

Where, as so often happens, the toxic levels found at post-mortem are not in a potentially fatal or even toxic range, then the pathologist should seek the advice of a **clinical** toxicologist to determine whether any of the symptoms or signs during life may assist in deciding on the cause of death. As so often happens in forensic problems, the investigation of a fatal poisoning must be a cooperative effort, especially between pathologist and laboratory analyst. Even in apparently obvious cases, such as a blood saturation of 50 per cent carboxyhaemoglobin, it is not for the analyst to declare a definite cause of death, as the victim may also have had a fractured skull – but equally, the pathologist has an obligation to provide the laboratory with the best possible samples in the best possible condition, as well as good information about the circumstances of the case.

THE CONCEPT OF THE FATAL DOSE

Many persons, including some doctors, are under the firm misapprehension that, for most toxic substances, there are relatively constant quantities that will cause death. Not only the lay public, but lawyers, police, coroners and others assume that there is a more or less linear relationship between the amount of poison that enters the body, the resulting levels in blood and tissues and the degree of disability caused – the ultimate disability being death.

In addition, it is often thought that back-calculation from blood and tissue levels can arrive at a definite assessment of how much poison was originally administered. This aspect is of particular concern to coroners and similar officers, who have to decide on motive in potential suicides, where the magnitude of any overdose may assist in distinguishing between accident and self-administration.

The pathologist has his own difficulties in respect of 'the fatal dose', as quantitative results from the laboratory have to be matched against a knowledge of published blood and tissue levels for that substance in relation to its potential toxicity. Though numerous tables of toxic levels have been published, there is considerable variation between the levels recorded. Reasons for this are explored later. It is obvious that there is no 'fatal dose' in the sense of a single threshold concentration above which a person dies and below which he survives. Instead, there is a range of levels, the upper and lower margins of which vary from one authority to another, which encompasses most deaths – but even here there are many exceptions, instances being recorded where survival occurs well above the upper limit and death occurs below the lower margin. In such cases, the task of the pathologist – which cannot always be satisfactorily discharged – is to evaluate all other non-toxicological data to see if they can modify the circumstances sufficiently to allow an acceptable explanation for the death. It is sometimes difficult to explain these concepts of great biological variation to lawyers and police officers, who expect more definite decisions, and might even feel that the pathologist is being evasive or obstructive.

The concept of the 'LD_{50}' is sometimes raised by the more knowledgeable lawyer, but this does not assist materially in any individual case. The LD_{50} is a device used by pharmacologists and toxicologists in a statistical sense in animal experiments. Over a large number of tests, it provides a toxic level at which half the animals will be expected to die. Though the indicative value in a general sense in comparing the toxicity of one substance against another, there is no way of knowing whether the human victim of poisoning lies at the upper or lower end of the classical bell-shaped curve that characterizes most biological responses.

Far more useful is the cumulative record of actual laboratory results from toxicology centres that deal with human poisoning, which progressively build up a large database of blood and tissue levels, and correlate these with records of the clinical state, toxic effects and fatal outcome. Even here the variations are wide, as the many published tables testify, but at least general guidance can be obtained.

VARIATIONS IN PUBLISHED FATAL POISON CONCENTRATIONS

In the many publications that offer therapeutic, toxic and lethal ranges for a wide variety of poisonous substances, there is considerable variation in the quoted levels. These can be a source of confusion and sometimes dismay to the pathologist who, even with the advice of his local toxicologist, may find it difficult to decide whether a death can justifiably be attributed to a particular drug or other toxic substance. Even if he can so satisfy himself, he may be subjected to keen questioning, interrogation, doubt or criticism from colleagues, coroners, police, lawyers and others, who have access to different versions of toxic and fatal levels. The ultimate challenge may come in a criminal court, where opposing counsel may openly defy the pathologist's interpretations in cross-examination.

Dr A R W Forrest, Chemical Pathologist and Toxicologist at Sheffield's Royal Hallamshire Hospital (personal communication) points out some reasons for the variations found in such published data. He expresses the view that it is rather surprising that the order of disagreement is not greater, given the opportunities for disparity.

First, many of the published series are small, some being only individual case reports. Statistically, this is not a good foundation for establishing reference ranges, which are much better obtained from a database derived from the cumulative results of a large laboratory service, such as the British Home Office Forensic Science Service, which maintains a central computer store of all results from its laboratories.

Second, analytical techniques vary widely, both in method and accuracy. The specificity varies from laboratory to laboratory so that there may be a lack of uniformity about what is actually being measured. For example, paracetamol may be measured by a non-specific method that picks up its metabolites as well as the native drug, which will then offer a different blood level in a fatal case from that found by more specific methods. In other words, one is not comparing 'like with like'.

Third, as discussed elsewhere in this chapter, the site of sampling may introduce wide errors. With some substances, a several-fold variation in concentration may be found between femoral vein and cardiac cavity blood.

Last, errors occur because a 'fatal' level may be attributed to one substance without taking into account the level – or even the existence – of other toxic substances that the deceased may have taken, and of which the pathologist or analyst may not even have been aware. For example, the newer more potent benzodiazepines may be missed in a simple toxicological screen, but could well have contributed to the toxicological overload that caused the patient to die. In such a case, the level of the recognized drug would be blamed for the death, whereas in fact it may not have been a lethal dose in itself, though was recorded as such in any database or tables.

These facts highlight the dangers of limiting a request for analysis solely to the substance known or thought to have been taken. Often an efficient screen for other substances will reveal other unsuspected compounds, sometimes more toxic than the one originally suspected.

It has to be recognized, however, that, in many jurisdictions, the availability and expense of toxicological investigations may make it impossible to pursue a full analytical survey, especially if this is to be undertaken on a speculative basis rather than for quantification of a known or strongly suspected toxic agent. A full screen may only be practicable and justified in a suspected homicide, if laboratory facilities and fiscal support are severely restricted. In many countries homicide by poison is relatively rare and the funds to investigate accidental, suicidal and iatrogenic poisoning exhaustively may just not be available.

THE AUTOPSY IN SUSPECTED POISONING

The autopsy in these circumstances can be amongst the most difficult of problems faced by a forensic pathologist – not in the technical procedure of the examination, but in the final evaluation of all the available information.

The nature of the poisoning autopsy in Western countries has changed dramatically since the last century, when poisoning was a common method of homicide. There has been a marked change in the nature of poisons used in murder, suicide and accident. The corrosives, heavy metals and alkaloids commonly ingested in former years became relatively easy to detect, either by gross autopsy appearances or by straightforward analytical methods. Further refinements of toxicological techniques, instead of the old methods in which large samples had to be tested because of the insensitivity of laboratory tests, allow the detection of nanogram quantities.

Acids, alkalis, phenols, arsenic, antimony and strychnine, for example, became easy to detect, and in the Western world these gave way to compounds that leave little or no gross, or even histological changes in the body. Most are pharmaceutical or agrochemical substances, active in low dose compared to the old ‘blockbuster’ poisons. As for the majority of drugs used in legitimate medical therapy, an added problem arises when low post-mortem levels are found – is this merely a therapeutic dose or the tail-end of a declining lethal dose?

In some parts of the world, such as South-east Asia, Africa and the Indian subcontinent, poisoning remains common and more physically damaging substances continue to be seen that leave obvious autopsy lesions. Some of these are described under the appropriate headings in the succeeding chapters, but in most poisonings the major function of the autopsy is to evaluate any other conditions present, both from trauma and natural disease – but also to collect suitable material for laboratory analysis. The proper retention of optimal samples, their correct preservation and dispatch to the toxicologist are of such fundamental importance that they are discussed in detail.

POISONING DEATHS IN HOSPITAL

A considerable proportion of those who die from suspected poisoning will have died in hospital, and it is of prime importance that the medical records be obtained and studied before the autopsy begins. Even if poisoning was not confirmed or even suspected by the clinicians treating the patient, later information may have come to the pathologist to raise this possibility. Whether known or not, the results of ante-mortem investigations may be of considerable use to the pathologist. If poisoning was known or suspected before death, there may well have been toxicological analyses performed and the results of these may be of great value.

Whether poisoning was suspected or not, there may still be ante-mortem blood or urine samples (taken for biochemical or haematological tests) stored in the hospital laboratory, which may be rescued for retrospective analysis. Such ante-mortem samples are likely to be of greater use than fluids drawn off at autopsy, because of sampling defects, post-mortem changes and because toxic levels are likely to have been higher during life, representing more accurately the maximum toxic concentrations.

In addition, many patients dying in hospital from drug overdoses will have indwelling catheters in place. It is the common practice for the nursing staff to remove these at death and discard the urine. If it is possible to establish a practice in the wards where such terminal urine samples are saved, valuable material for analysis can be obtained.

THE COLLECTION OF AUTOPSY SAMPLES FOR TOXICOLOGICAL ANALYSIS

The investigation of a death from suspected poisoning may stand or fall upon the correctness or otherwise of the sampling of fluids and tissues from the body. Unsuitable samples, inadequate amounts, incorrect sampling sites, poor containers, inadequate preservation methods, and delayed or unsatisfactory storage and transport to the laboratory may frustrate or distort proper analysis. The final outcome may be wrong, either in failing to detect a poison actually present, in measuring only part of that originally present or – in some cases – even producing falsely high results that then lead to an incorrect cause of death. Not only must samples be in the optimal condition, but the accompanying information from the pathologist to the analyst needs to be as accurate and comprehensive as possible, so that the most appropriate techniques are used, and allowance made for any interfering substances that may be present.

In decomposed bodies infested with maggots, and in the absence of tissues or fluids normally taken for toxicology, Diptera and other arthropods can be used as alternative specimens for toxicological analyses. This relatively new field of forensic entomology is called entomotoxicology.

The time of sampling

It is obvious that the shorter the delay between death and the removal of samples, the better. Though some toxic substances, such as carbon monoxide, form stable compounds in the body, many others (especially volatile substances and some pharmaceutical products) will be broken down by post-mortem autolysis and decomposition. When an autopsy cannot be performed quickly after death, in terms of a few hours, then mortuary refrigeration is the first line of defence to slow up putrefactive and autolytic processes. If delay is foreseen, usually because of administrative problems in obtaining consent or authority for autopsy, it may be possible to obtain a sample of blood through the body surface, such as puncturing the femoral vein by needle and syringe. The blood can then be kept in optimal conditions, with preservative where needed, and perhaps with the serum or plasma separated from the cells to avoid haemolysis. Similarly, urine could be drawn off by catheter or even suprapubic puncture, unless strict regulations forbid this as anticipating autopsy permission.

In some jurisdictions, authority for autopsy may be particularly hard to obtain, from religious, financial or administrative reluctance. Where poisoning is suspected, permission may be granted only for external examination and sampling; here venous blood, urine and perhaps vitreous humour may have to suffice for all investigations.

Information supplied to the laboratory

When samples are submitted to the toxicologist, they should be accompanied by the best possible information relevant to the case. It is both counterproductive and an unprofessional discourtesy merely to record the personal details of the deceased and list the samples, with a terse demand such as 'Any poisons?' on the request form. Such a demand could legitimately be refused by the toxicologist, as it is quite inadequate information upon which he can be expected to function effectively and safely. The following information should be supplied and where necessary, supplemented by direct discussion either in person or by telephone:

- The personal details of the deceased, including age, sex and where thought relevant, the occupation (especially if in agriculture or industry).
- Brief details of symptoms, if any, and length of illness.
- The post-mortem interval before samples were obtained, and the actual date and time of sampling.
- The name, address and telephone number of the pathologist.
- A list of all samples provided, with an indication of the sampling site for each.
- The nature of any preservative in each of the samples.
- If there has been a delay in submitting or transporting the samples, a note of the condition under which they have been stored (for example, refrigeration or deep-freeze).
- Any special risk associated with the samples **must** be communicated to the laboratory. The most obvious are infective conditions, especially hepatitis B or C virus or HIV infection in the deceased, though other diseases such as tuberculosis, tetanus, anthrax, gas gangrene or any other bacterial or viral condition must also be specifically reported. In relation to hepatitis and HIV, even if these are not definitely confirmed, the toxicology laboratory must be told if the deceased was in a high-risk group, such as a drug addict or a homosexual. In many areas, including Britain, some toxicology laboratories will not accept samples from such high-risk groups until a blood sample has been screened for hepatitis B and HIV antibodies; if positive, they may decline to carry out the analysis or perform it only under strictly controlled conditions.

  Similar warnings must be given to the analyst if there is any possibility of certain harmful substances, such as radioactive isotopes or certain war gases, being present in the samples.
- When a death has criminal aspects, such as a murder or manslaughter, then the usual strict precautions must be taken for continuity of evidence. Each container must be carefully labelled and preferably countersigned by the pathologist. Some jurisdictions will require actual seals on the containers themselves or the package into which they are placed for transport, or both. Accompanying signed 'exhibit labels' with serial numbers corresponding exactly with numbers on the jars may be required. The containers must be given by the pathologist to a named person, usually the 'exhibits officer' of the police investigating team (or to forensic scientists if they attend the autopsy). The police officer must hand the samples personally to a member of the laboratory staff and a record of this chain of evidence must be kept, so that there can be no criticism levelled at anyone when the matter comes to court, raising doubts about the correct identity of the sample.

CONTAINERS FOR TOXICOLOGY SAMPLES

There will be considerable local variation in the type of containers used for the collection and transmission of samples to the laboratory. Some laboratories will issue their own kits, an example being those provided by the Home Office Forensic Science Service in England and Wales. These consist of a plastic bag containing a sheet of instructions about the most suitable samples, together with a large plastic pot with tightly self-sealing lid for liver, a smaller one for stomach contents, several plastic or glass 30 ml universal containers for blood and urine, a small vial containing fluoride for blood-alcohol estimation and a sterile syringe and several needles. Variations of such kits may be provided by different laboratories, but the general nature of the containers is the same. In other areas, the pathologist will provide the containers himself and they should conform to the following general specifications:

- They should either be new or, if previously used for other samples, have been rigorously cleaned and sterilized. Even when they are new, it is preferable for containers to be washed and sterilized before use unless the manufacturer's specifications clearly make this unnecessary. All containers must be chemically clean, not just apparently clean to the naked eye. This applies particularly to any rubber, plastic or other seal within the lid or cap, which may be a trap for any debris or even residue from a previous sample.
- Blood should be collected in screw-capped universal containers of about 30 ml or, where less will suffice, in plastic-capped tubes of about 5 ml. Urine is best held in 30 ml universal containers, as is bile. Stomach contents may be retained in glass or plastic jars with a volume of at least 250 ml, the lid being either a tight screw-thread with a cardboard or plastic liner, and a self-sealing plastic top. Some laboratories require the stomach wall as well as the contents, as toxic substances may be held in higher concentrations in the rugae and crypts of the mucosa, or even in the blood in the actual stomach wall. The size of the container may need to be larger if the organ itself is retained. Liver is usually required, preferably a piece weighing a few hundred grams, rather than the whole organ. The total weight of the liver should be communicated to the laboratory if only part is sent. It may be preserved in a larger glass or plastic container. When the laboratory wishes the whole organ, this container needs to be large enough to hold about 3 litres. If only small samples are needed, such as vitreous humour or cerebrospinal fluid, bijou bottles of 5 ml capacity or small tubes of similar size, are needed.

In Britain, the government forensic laboratories provide special bottles for blood-alcohol estimations, for use in both drink–driving offences in the living and for autopsy samples. These are short, wide bottles of about 10 ml capacity, with a fixed top consisting of a plastic diaphragm secured by a metal flange with a central aperture. They are filled by syringe, the needle being pierced through the diaphragm, but have the disadvantage of building up pressure when the blood is injected, so that, when the needle is removed, blood may spurt back and even contaminate the operator – the reverse of the intention of these vials to be safe. The danger can be avoided by either inserting a second needle through the diaphragm to equalize the pressure, or to suck out air from the bottle with the empty syringe and needle before taking the sample so that the introduced blood replaces the lost air. Intestinal contents, with or without the gut itself, can be placed in large jars or plastic pots, similar to those used for liver.

Plastic containers, especially polypropylene, are increasingly used and have the advantage of not smashing when dropped. Some analyses, however, may be affected by the plasticizer in either the container wall or the cap – cocaine is an example. The advice of the toxicologist with whom each pathologist normally collaborates should be sought in respect of his individual preferences for containers. Where lungs or other tissues are to be submitted for analysis for volatile substances, as in solvent abuse, ordinary polythene bags are not suitable containers. Nylon bags (such as those used by arson investigators) should be used, as they are not permeable to such substances.

PRESERVATIVES IN SAMPLES

Though many samples for analysis are best sent in their original state, others require additives to maintain them in optimum condition until they reach the laboratory. In addition to such preservatives, some require anticoagulants to keep the blood fluid. The prime example of sample preservations is that for blood alcohol, as improper care of samples may distort the original alcohol level in either an upward or downward direction. The usual preservative is sodium or potassium fluoride, which is essential if the sample is not to be analysed within a few hours of withdrawal from the body.

Errors caused by post-mortem changes before the sample is removed cannot be avoided. These are variable, depending partly on the conditions of the post-mortem environment and partly on the microbiological population of the corpse. In warm conditions, fermentation by yeasts and other alcohol-producing flora can produce appreciable quantities of alcohol: it has been reported that as much as 150 mg/100 ml have been generated within 24 hours after death.

Once the blood or urine has been withdrawn from the body, however, further changes can be arrested by preservatives. Much has been published on this subject and a variety of concentrations of fluoride have been recommended by different authors. Glendening and Waugh (1965) used 100 mg of sodium fluoride/10 ml blood and found no change in the alcohol content when samples were kept at room temperature for up to 3 months. Plueckhahn (1968) found that at least 5 mg/ml of sodium fluoride was required to inhibit alcohol dehydrogenase activity, which destroys alcohol. He found, however, that this concentration failed to inhibit alcohol-forming organisms completely, such as when yeasts were introduced into the samples. Together with Ballard, he investigated the use of mercuric chloride as a preservative (Plueckhahn and Ballard 1968). This substance was also used by Bradford (1966) who added 0.5 mg of sodium citrate and 0.1 mg of mercuric chloride/ml of blood to ensure that the samples remained liquid and sterile. When he replaced the mercury by fluoride, significant changes occurred, usually a loss of alcohol, though occasionally it increased.

In most cases the delay before analysis is relatively short, being measured in days – often with refrigerated storage – so the long-term experiments mentioned here do not apply. When the ambient temperature is high, much more caution needs to be used. It would seem that for general use, a concentration of 10 mg/ml of sodium or potassium fluoride is satisfactory. Fluoride should also be added to urine and vitreous humour if alcohol estimations are required. Cocaine and its metabolites are also labile *in vitro*, and fluoride should be added to samples submitted for analysis for this drug.

Cyanide may be formed in considerable quantities in plain blood samples, which are of little use for cyanide estimation. Fluoride should be added to such specimens, as well as for carbon monoxide (carboxyhaemoglobin) if the analysis is to be delayed. In all analyses for pharmaceutical drugs, two samples of blood should be submitted, one plain in large volume of at least 25 ml and another smaller sample in fluoride.

When insulin assay is required on a blood sample, special precautions should be taken. Haemolysis of the red cells releases enzymes which will reduce the S–S bonds in insulin and destroy its immunoreactivity, so the sample should be centrifuged as soon as it is obtained to separate the serum. A heparinized sample should also be taken for glycosylated haemoglobin estimation and fructosamine assay. Vitreous humour, blood and urine samples should be placed in fluoride and sent for glucose estimation, as well as tissue samples from around any putative injection site, together with a control tissue sample from a site elsewhere on the body where insulin is unlikely to have been recently injected.

THE SITE OF SAMPLING FOR TOXICOLOGICAL ANALYSIS

It is now obvious that, in the past, serious errors were made in toxicological analyses from lack of care or consideration concerning the source of body-fluid samples. Significant variation can be found in the concentration of many substances depending on the place from which sampling was carried out. In life there may be variations when arterial as opposed to venous blood is used, as tissues may take up the compound from the arterial supply, the concentration then being lower in the venous return. Similarly, portal blood may have a substantially higher concentration of a substance that is being absorbed from the intestine, before it is extracted by passage through the liver. After death, most variation is caused by uneven destruction by enzymatic and microbiological activity – and by diffusion from sites of higher concentration. The barriers formed by living cell membranes break down after death and small molecules in particular may move easily through the tissues into vascular channels.

Post-mortem levels of many substances are unreliable because of this diffusion effect, making the interpretation of physiological components, such as sodium, potassium, calcium, glucose, urea and many others extremely difficult, if not impossible. Applying these facts to toxic compounds, the concentrations may vary considerably according to the sampling site. As an illustration, the Moorgate Tube disaster in London showed that blood-alcohol levels may vary widely between different sampling points. The driver of an underground train was killed, along with a number of passengers, but his body could not be recovered from the warm environment for several days. As it was obviously vital to know if drink had contributed to the accident, four samples were taken from various sites in the body. On analysis, a fourfold variation from between 20 and 80 mg/100 ml was obtained, presumably because of variable rates of putrefactive alcohol production, as there was no evidence to show that the driver had imbibed alcohol before death.

Most research in respect of sampling variation has been directed at alcohol. One point of dispute, which is relevant even when post-mortem changes are minimal, is whether alcohol in the stomach can diffuse after death to neighbouring organs and produce a falsely elevated alcohol concentration in blood at those sites.

The relevance is clear, in that, if a person drinks alcohol immediately before death, there will be insufficient time for it to be absorbed. Thus ethanol remaining in the stomach after death cannot have contributed to his ante-mortem blood-alcohol level, and hence his cerebral function and consequent behaviour. If however, analysis of a sample of blood taken from the heart cavities is contaminated by post-mortem diffusion from the adjacent stomach, then a falsely high reading will be obtained, which may have serious legal implications, if the error is not appreciated.

Several workers have investigated this problem, with conflicting results. Gifford and Turkel (1956) introduced alcohol into the stomachs of cadavers and found that subsequent heart blood levels increased by 25 up to 106 mg/100 ml. They then compared femoral vein blood with heart blood voided into the pericardial sac in alcohol-positive autopsies and found that the concentration in the latter was higher. Plueckhahn and Ballard (1968), however, criticized this work on the grounds that pericardial fluid alcohol was usually about 20 mg/100 ml higher than that in the heart or peripheral vessels. Plueckhahn went on to instil alcohol into the stomach of cadavers, taking multiple site samples at times from 6 to 50 hours later. His results indicated that there was a significant diffusion of alcohol to the pericardial and pleural fluids, but that the increase was minimal in intraventricular heart blood.

Pounder and Yonemitsu (1991) carried out experiments with dead bodies in which they introduced a slurry containing alcohol, dextropropoxyphene and paracetamol into the trachea to simulate aspiration of stomach contents. After a delay of 48 hours, an autopsy was carried out and drug concentrations measured in samples taken from various sites. Whilst femoral vein blood remained clear of these substances, samples from vessels in the thorax showed up to 130 mg/100 ml of alcohol with a mean of 58 in a pulmonary vein – and up to 1934 mg/l of paracetamol (mean 969), both obviously spurious due to post-mortem diffusion. This situation could have occurred from agonal regurgitation of stomach contents with recently swallowed drugs back into the air passages and shows the dangers of inappropriate sampling sites, with the femoral the place of choice.

Winek *et al.* (1995) have shown that, in the sampling practice of transthoracic needling, especially in cases of trauma where the gastrointestinal tract has been damaged, false elevations of blood alcohol can occur. These workers used heart blood as their control, which itself seems suspect as a satisfactory autopsy sample.

For the reasons given above, **peripheral** vein blood should be used whenever possible, avoiding heart blood as a potential source of error for alcohol and presumably other easily diffusible substances.

THE TECHNIQUE OF OBTAINING AUTOPSY SAMPLES

Blood

There are several ways of obtaining blood samples at autopsy and perhaps the most useful advice is what **not** to take. Blood should never be obtained from body cavities after evisceration, as it is almost certain to be contaminated with other body substances. The practice of scooping ‘blood’ – or more accurately, bloody fluid – from the paravertebral gutters or the pelvis is always unacceptable, as urine, intestinal contents, gastric contents, lymph, pleural and ascitic fluid and general tissue ooze will always find their way into such a sample and negate the reliability of analysis.

When a large haemothorax or haemopericardium is present, it may be grudgingly acceptable to use such blood or clot, if a clean sample is taken immediately on opening the chest, before any dissection or disturbance of organs is made. This is only second best to obtaining intravascular blood – and, in the case of alcohol and other diffusible substances, the results cannot be relied upon, as the sample may be contaminated by post-mortem diffusion from the stomach, as discussed in the preceding section. Carbon monoxide estimations may also be made on clean cavity blood, if nothing else is available, as sometimes happens in a badly incinerated body. As it is absorbed through the lungs, it cannot diffuse from the gastrointestinal tract to give a false result. Even so, it is a rare autopsy that cannot provide a small sample of blood from some peripheral vein, if enough care and persistence is used.

To return to the general state of affairs, the most satisfactory way of obtaining a venous blood sample is venepuncture of the femoral vein by direct puncture in the groin before the autopsy begins. Practice is required as, unlike a living patient, the vein is not usually palpable. Once the lumen is entered, 25 ml can usually be drawn off without trouble, though a slow flow may be improved by massaging the leg to drive blood proximally or the leg may be raised, if rigor allows. Though less satisfactory than a leg vein, for reasons discussed earlier, a neck vein can also be punctured through the skin, and even after death the jugular may sometimes be both visible and palpable, especially in congestive deaths. The subclavian vein can also be tapped.

When collecting blood samples by percutaneous puncture, some pathologists prefer to use a needle designed for aspirating samples from rubber-capped vials, rather than a needle meant for clinical aspiration or injection. The same needles can be used for collecting vitreous humour. An example of this type of needle is the Becton Dickinson purple-hubbed 16 gauge one-inch or a similar American equivalent.

In some jurisdictions, especially in countries where the system allows only a low autopsy rate, transcutaneous sampling is virtually the only post-mortem investigation possible and the doctors in these areas become adept at cadaver venepuncture. Many pathologists, the authors included, have been brought up to obtain blood samples during the actual performance of autopsy and this can be as satisfactory as external venepuncture, as well as saving the expense of a new syringe and needle every time. A common procedure after evisceration is to hold a container (such as the 30 ml universal) under the severed end of the subclavian vein, within the upper part of the empty thorax. The arm is then elevated and massaged if necessary, to express blood into the container. Similarly, and perhaps preferably, the femoral or external iliac vein may be transected near the brim of the pelvis, and a container held under the cut while the leg is elevated, or massaged, or both. Either of these methods will provide ample, clean, peripheral venous blood, though care must be taken to avoid any contamination from the body cavities. This initial collection should be transferred into one or more new tubes, with preservative added if appropriate, as the first receptacle will be soiled externally during the collection process.

An alternative method of collecting a venous sample, which always gives a copious flow of blood, is to incise the internal jugular vein during the initial stages of the autopsy. As soon as the main autopsy incision is made, the flaps of the neck skin can be dissected aside and the jugular exposed, if necessary by dividing the sternocleidomastoid muscle. As in the previous technique, a bottle can be held against the vein as it is transected and the flow collected directly into it. The blood may overflow into the pouch formed laterally by the reflected skin, which forms a reservoir where a considerable volume may collect. If the flow is scanty, it can be improved by raising and lowering the head, which will cause blood to return from the upper venous drainage areas.

Even more can be expressed by pressing on the chest, but here there is the potential disadvantage of forcing up blood through the superior vena cava from the heart, which may distort the value of alcohol and other drugs that may have diffused port-mortem from the stomach. Blood should not be taken from the heart cavities, the inferior vena cava, or the portal or hepatic veins, as these may also give concentrations that are at variance with those in the peripheral vascular system. A full 30 ml of blood should be submitted to the laboratory wherever possible, together with a 5 or 10 ml fluorided sample.

Urine

Prior to, or in the absence of, an autopsy, urine can be obtained by catheter or suprapubic puncture with syringe and long needle. At autopsy it is usual to wait until evisceration has been carried out before dealing with the bladder. A sample can be obtained by puncturing the fundus with syringe and needle. Alternatively, the bladder can be stretched by pulling the fundus upwards with the fingers, then a sagittal incision made with a knife on the ventral surface, which must be free from blood soiling. The urine that wells out can be collected directly into a small container such as a 30 ml universal. If only a small amount is present, then the incision may have to be enlarged and the residual urine sucked out under direct vision, using a syringe without a needle.

Bile

This can be useful for some analyses, such as those for morphine and chlorpromazine, which are concentrated by the liver and excreted into the gallbladder. Direct collection into a bottle is advised, as bile is usually too viscous to be sucked through a needle.

Cerebrospinal fluid

This is not often required for toxicological analysis, though it may be needed for microbiological and virological studies. If needed, it should be collected by lumbar or cisternal puncture, as in the living patient. It can be difficult or impossible to obtain in this way because of the lack of any intrathecal pressure.

The body should be turned on to its side before the autopsy begins and flexed as much as possible by an assistant. If an infant, it should be sat up and again flexed forwards to curve the spine as much as possible. A needle on a syringe should then be passed between two lumbar or lower thoracic spines, and stopped as soon as any penetration of the theca is felt. Moderate suction should be maintained on the syringe piston to compensate for the lack of internal pressure. Alternatively, the needle should be passed into the midline just below the occiput and advanced upwards until the skull is contacted just posterior to the foramen magnum. The needle is slightly withdrawn and re-advanced until it slips through the posterior part of the atlanto-occipital membrane into the basal cistern when cerebrospinal fluid may be aspirated. It is of little use taking cerebrospinal fluid from the bowl of the posterior fossa after the brain has been removed as the concentration of many substances is different in blood compared with the fluid, so blood contamination distorts any results. Clear cerebrospinal fluid may, however, sometimes be obtained from the lateral ventricles, either by needle puncture or cutting down through the cortex.

Vitreous humour

This is sometimes useful, especially in bodies with appreciable post-mortem decomposition, as the fluid in the eye resists putrefaction longer than other body liquids. Vitreous fluid is also used for estimating the time since death (see Chapter 2).

A puncture should be made through the sclera at the outer canthus with a fine-gauge needle. This should be placed as far laterally as possible, pulling the lid out, so that when released it returns to cover up the puncture mark for cosmetic reasons. The fluid should be sucked out by syringe, but it will often come only slowly because of its viscosity. If the best restoration is needed, water should be reinjected through the same needle to reinflate the globe, which tends to collapse on suction.

Stomach contents

As stated earlier, these can be collected directly into wide-mouth glass or plastic pot of at least 250 ml volume. To collect them, the exterior of the stomach should be washed clean of blood and other contamination, and pulled with attached organs to the edge of the dissecting board or sink. The greater curvature should be opened cautiously with large scissors, and the jar held underneath so that the contents flow directly into it. When most have been expressed, the greater curvature can be opened up fully and the gastric lining examined. Any further contents are scraped out and any powder, capsule or tablets picked off, and either added to the main jar or placed in a separate small container. Such focal deposits form a more concentrated sample for the laboratory. Any undissolved tablets or capsules should be carefully preserved, as the laboratory or a pharmacist may be able to identify them by their appearance and colour. After the mucosa has been examined, including that of the duodenum, the stomach wall can be dissected off. Some laboratories require this for their analysis, either added to the jar of contents or sent separately. The wishes of each individual toxicologist should be discovered in advance.

Intestinal contents

These are not routinely required for analysis unless some particular gastrointestinal poison is suspected, such as one of the heavy metals. Again, the toxicologist should be consulted. If required, both ends of the small gut should be ligated by string sutures at duodenum and ileum. The intestine is cut through at these points and stripped out by cutting through the mesentery in the usual way. The intact gut can be sent to the laboratory in a large container (such as that used for liver) or – if the mucosa needs to be examined first – it should be opened in a large clean tray. One suture can be cut and the contents milked from one end to the other into a suitable jar, the gut then being opened and sent with the contents to the laboratory. The large intestine and contents are rarely required, as most toxicology laboratories are not keen on handling and storing large volumes of faeculent material. In heavy-metal poisoning, such as arsenic and antimony, however, some analysis may be required.

Vomit

This is rarely collected at autopsy, unless a large quantity has been found in the air passages. However, vomit is not infrequently collected by ambulance crews and police at the scene of the illness or agonal event, and brought with the deceased to hospital or the mortuary. If thought relevant in a case of suspected poisoning, it should be properly contained, labelled and forwarded with other samples for toxicological examination.

Other fluids

Other fluids are rarely collected, but on occasions may be useful, especially if clean blood cannot be obtained. Pericardial fluid, synovial fluid, pleural effusion and ascitic fluid can be used for qualitative analysis to identify a range of substances, though these levels can rarely be related to established ranges of blood concentrations. They may be collected and transported in the same type of container and with the same preservatives as vitreous humour.

Analysis for volatile substances

In solvent abuse and deaths from gaseous or volatile substances, the toxic material may be isolated from a whole lung. As soon as the thorax is opened at autopsy, a lung is mobilized and the main bronchus tied off tightly with a string ligature. The hilum is then divided and the lung placed immediately into a nylon bag, which is sealed and sent as soon as possible to the laboratory. Ordinary plastic (polythene) bags are not suitable for this task, as they are permeable to volatile substances. Nylon bags, as used by arson investigators to collect samples that may have volatile accelerants, do not suffer from this defect.

In collecting blood samples for volatile substances, plastic tubes or tube/syringe combinations of the 'monovette' type are unsuitable, as the concentration of such compounds as toluene or other solvents will decrease considerably over a few days in storage. Vials with a rubber septum as a seal are also unsuitable, as the volatile substance can escape through the rubber. Suitable tubes are made of glass with an aluminium foil-lined cap or a polytetrafluorethylene (Teflon) liner. This type of tube is often used in laboratories for scintillation counting and can be employed for the collection of solvent samples. Such tubes should be available in any busy autopsy room where solvent abuse deaths may be handled. The tubes should be filled to the top to avoid loss of volatile substances into the head space and should be stored at 4°C, rather than being frozen solid.

Body tissues

Body tissues may be needed for some toxicological analyses. Liver is the most usual, as it concentrates many substances and their metabolites, which may then be recoverable long after the blood and urine levels have declined. Either the whole organ is saved or an aliquot of 50–100 g, according to the wishes of the particular laboratory. If only a part is retained, it should be taken from the periphery, away from major vessels and bile ducts.

Brain and kidney may sometimes be required and again a 50–100 g aliquot is usually sufficient. As with liver, the total original weight of the organ should be recorded and notified to the toxicologist.

Where a toxic substance may have been injected into the subcutaneous tissue or muscle, these tissues should be excised and submitted to the laboratory. The usual method of identifying the site is from a needle puncture mark, and a zone of skin and tissue should be removed circumferentially around this or an ellipse cut away, so that the defect can be sewn up on the body. A few centimetres in diameter around the mark is usually sufficient. The depth of the sampling depends on how deeply the needle track extends and may need to go into the underlying muscle. Radiography may assist in rare cases, as in the George Markov murder in London, where a political assassination was carried out by the introduction of a tiny sphere, presumably injected from an air weapon concealed in an umbrella. The sphere, drilled out to carry a potent toxin, probably ricin, was located by X-ray at autopsy after a small skin puncture was found. Where an injection site is sampled, it is essential that a control area from a remote part of the body is also sent to the laboratory. It is usual to take this from a symmetrical zone on the contralateral side, but caution must be used, as in both drug dependence and insulin usage, alternate sides may be used at frequent intervals for injection, so another more remote site might be preferable.

The substances which may have been injected are numerous, but insulin, morphine, heroin, cocaine and other illicit drugs are most common. In the notorious Coppolino case in the USA, products of succinylcholine (suxamethonium) were identified from around a needle track in the buttock of an exhumed body of a woman, leading to the conviction of her anaesthetist husband for murder.

REFERENCES AND FURTHER READING

Baselt RC. 2000. *Disposition of toxic drugs and chemicals in man*, 5th edn. Chemical Toxicology Institute, Foster City, CA.

Bourel B, Fleurisse L, Hedouin V, *et al.* 2001. Immunohistochemical contribution to the study of morphine metabolism in Calliphoridae larvae and implications in forensic entomotoxicology. *J Forensic Sci* **46**:596–9.

Bradford LW. 1966. Changing attitudes in drinking driver enforcement as reflected by the actions of researchers, courts, prosecutors and society. *J Forensic Sci Soc* **6**:204–8.

Cordonnier J, Heyndricks A, Piette M. 1986. Drug levels in plasma, vitreous humor, bone marrow, liver and bile. *Police Surg* **30**:60–7.

Drummer OH, Gerostamoulos J. 2002. Postmortem drug analysis: analytical and toxicological aspects. *Ther Drug Monit* **24**:199–209.

Fermer RE. 1995. *Forensic pharmacology*. Oxford University Press, Oxford.

Gagliano-Candela R, Aventaggiato L. 2001. The detection of toxic substances in entomological specimens. *Int J Legal Med* **114**:197–203.

Gifford A, Turkel M. 1956. Diffusion of alcohol through stomach wall after death. *JAMA* **161**:866–8.

Glendening BL, Waugh TC. 1965. The stability of ordinary blood alcohol samples held at various periods of time under different conditions. *J Forensic Sci* **10**:192–200.

Goff ML, Lord WD. 1994. Entomotoxicology. A new area for forensic investigation. *Am J Forensic Med Pathol* **15**:51–7.

Grellner W, Glenewinkel F. 1997. Exhumations: synopsis of morphological and toxicological findings in relation to the postmortem interval. Survey on a 20-year period and review of the literature. *Forensic Sci Int* **90**:139–59.

Hilberg T, Rogde S, Morland J. 1999. Postmortem drug redistribution – human cases related to results in experimental animals. *J Forensic Sci* **44**:3–9.

Introna F, Campobasso CP, Goff ML. 2001. Entomotoxicology. *Forensic Sci Int* **120**:42–7.

Jenkins AJ, Levine BS, Smialek JE. 1995. Distribution of ethanol in postmortem liver. *J Forensic Sci* **40**:611–13.

Jones AW, Lofgren A, Eklund A, *et al.* 1992. Two fatalities from ingestion of acetonitrile: limited specificity of analysis by headspace gas chromatography. *J Anal Toxicol* **16**:104–6.

Klatt EC, Beatie C, Noguchi TT. 1988. Evaluation of death from hypoglycemia. *Am J Forensic Med Pathol* **9**:122–5.

Langford AM, Pounder DJ. 1997. Possible markers for postmortem drug redistribution. *J Forensic Sci* **42**:88–92.

Levine BS, Smith ML, Froede RC. 1990. Postmortem forensic toxicology. *Clin Lab Med* **10**:571–89.

McCurdy WC. 1987. Postmortem specimen collection. *Forensic Sci Int* **35**:61–5.

O'Neal CL, Poklis A. 1996. Postmortem production of ethanol and factors that influence interpretation: a critical review. *Am J Forensic Med Pathol* **17**:8–20.

Palmeri A, Pichini S, Pacifici R, *et al.* 2000. Drugs in nails: physiology, pharmacokinetics and forensic toxicology. *Clin Pharmacokinet* **38**:95–110.

Paterson SC. 1985. Drug levels found in cases of fatal self-poisoning. *Forensic Sci Int* **27**:129–33.

Plueckhahn VD. 1968. Alcohol levels in autopsy heart blood. *J Forensic Med* **15**:12–21.

Plueckhahn VD, Ballard B. 1968. Factors influencing the significance of alcohol concentrations in autopsy blood samples. *Med J Aust* **1**:939–43.

Polson CJ, Green MA, Lee MR. 1983. *Clinical toxicology*, 3rd edn. Pitman, London.

Pounder DJ, Jones GR. 1990. Post-mortem drug redistribution – a toxicological nightmare. *Forensic Sci Int* **45**:253–63.

Pounder DJ, Yonemitsu K. 1991. Postmortem absorption of drugs and ethanol from aspirated vomitus – an experimental model. *Forensic Sci Int* **51**:189–95.

Prouty RW, Anderson WH. 1990. The forensic science implications of site and temporal influences on postmortem blood-drug concentrations. *J Forensic Sci* **35**:243–70.

Shapiro HA. 1953. The concept of the fatal dose. *J Forensic Med* **1**:129–31.

Sperhake J, Tsokos M, Sperhake K. 1999. Perimortem fixation of the gastric and duodenal mucosa: a diagnostic indication for oral poisoning. *Int J Legal Med* **112**:317–20.

Stead AH, Moffat AC. 1983. A collection of therapeutic, toxic and fatal blood drug concentrations in man. *Hum Toxicol* **2**:437–64.

Wennig R. 2000. Threshold values in toxicology – useful or not? *Forensic Sci Int* **113**:323–30.

Winek CL, Jr, Winek CL, Wahba WW. 1995. The role of trauma in postmortem blood alcohol determination. *Forensic Sci Int* **71**:1–8.

CHAPTER 28

Forensic aspects of alcohol

- Mode of action of alcohol 552
- Absorption of alcohol 553
- Elimination of ethanol 554
- Alcohol concentrations: units and various drinks 555
- Physiological effects of alcohol 556
- Modes of death in acute alcoholism 557
- Chronic alcoholism 557
- References and further reading 558

Ethyl alcohol is the most commonly used drug in the world and has such numerous points of contact with medico-legal pathology that it has to be considered separately from all other substances. Its abuse is a prime factor in many accidents – transport, domestic and industrial – and the majority of homicides are catalysed by alcohol intake. It acts as an adjuvant to many other toxic substances, combining to cause a fatal outcome where often the other drug alone would not have caused death. The chronic abuse of alcohol leads not only to definite pathological changes in a number of target organs, but also contributes to deaths from neglect, hypothermia and burns.

This central role in forensic practice makes it imperative that a comprehensive knowledge of its metabolism and effects is held by every forensic pathologist, not only for an understanding of autopsy appearances, but to be able to assist investigative authorities to place the death into its alcoholic context.

MODE OF ACTION OF ALCOHOL

Ethanol is a small molecule that is easily miscible with water and therefore quickly diffuses through the whole of the aqueous compartment of the body.

There are slight variations in the intracellular concentration in different tissues – for instance, the red blood cells contain somewhat less alcohol than plasma, so that a whole-blood concentration is slightly less than that in separated plasma or serum. Most serous fluids are in equilibrium, however, though Plueckhahn (1968) claims that pleural and pericardial fluids have a higher concentration than that of plasma. The aqueous humour of the eye is in equilibrium with blood, as is cerebrospinal fluid, given sufficient time for equilibration.

It is said above that alcohol disperses rapidly through the aqueous compartment and this excludes the adipose tissues, as ethanol is almost insoluble in fat. This fact has important practical significance, as people with large fat stores will produce a higher blood-alcohol level for a given intake of alcohol than persons of the same weight who are lean, as the aqueous compartment is smaller. This is particularly relevant to women, who, by virtue of their **panniculus adiposus**, may develop blood-alcohol concentrations at least 25 per cent higher than men of the same body weight after similar drinks. This must be allowed for in all calculations that attempt to relate blood-alcohol levels to intake of drink.

The behavioural effect of alcohol is caused by its action on brain cells and is a function of the concentration in the blood, the alcohol readily passing the 'blood–brain barrier' to bathe neurones via the cerebral extracellular fluid. The action of alcohol is entirely depressant, a fact not always appreciated by those who observe the apparent excitant effect of drinking. Alcohol acts on neural cells in a way similar to hypoxia, reducing their activity. In lower concentrations, this action is confined to the more specialized and sensitive cells of the cerebral cortex, leaving lower brain functions relatively unaffected. The consequent depression of these higher areas releases their inhibitory tone and thus 'takes the brakes off' more primitive and unrestrained behaviour. This was aptly described by Thomas de Quincey (1785–1859) when he wrote that 'sobriety disguiseth man'!

With increasing concentrations of ethanol, progressively lower levels of brain function are depressed and, when the vital centres in the midbrain and medulla are affected, there is a danger of fatal cardiorespiratory failure. Alcohol also causes changes in the heat regulatory mechanism, both centrally and by a vasomotor effect. Generalized vasodilatation occurs, especially in the skin, which may lead to marked heat loss and dangerous hypothermia. There is an increase in heart rate with low concentrations but, when alcohol levels approach the dangerous ranges in excess of 300 mg/100 ml, bradycardia may develop. A slight and inconstant increase in blood pressure may occur, more often in the systolic level, which increases the pulse pressure between systolic and diastolic. There is no evidence to support the contention that alcohol dilates the coronary arteries and claims that it causes dilatation of cerebral arteries (and thus worsens intracranial haemorrhage) are also equivocal. Alcohol also has a diuretic effect and, when combined with large quantities of fluid, as in excessive beer drinking, may lead to electrolyte disturbances. In healthy test subjects short-term alcohol administration causes transitory hypoparathyroidism, accounting, at least in part, for the transient hypocalcaemia, hypercalciuria and hypermagnesiuria. In alcoholic patients especially hypomagnesaemia is common.

ABSORPTION OF ALCOHOL

For all practical purposes, it may be assumed that the only route of absorption of alcohol is by mouth as, although it can be inhaled, situations where this occurs must be rare indeed. One case is known to the author (PS), in which a man suffering from a tumour of the base of the tongue, probably causing difficulties in swallowing, had instilled fruit brandy rectally via a plastic tube using a pump, and dying as the result of acute alcohol poisoning.

Almost as soon as the alcohol is swallowed, it begins to be absorbed into the blood – and as soon as that blood reaches the liver, the alcohol begins to be eliminated. Therefore the blood level (and hence the brain concentration) is a dynamic balance between absorption and elimination, the peak determining the maximum behaviour effect. Such a balance is often represented graphically by the 'blood-alcohol curve' (BAC), which indicates the intensity and duration of physiological effects.

Ethanol is capable of being absorbed by any part of the gastrointestinal canal, but in practice this is confined to the stomach and upper small intestine, as only a little alcohol remains to pass through the wall of the ileum or colon. When alcohol is taken with food, however, there seems to be a deficit in the amount absorbed, as some never appears in the bloodstream. Some of this lost alcohol may be excreted in the faeces, but when absorption is slow, more may be destroyed by the liver directly from the portal blood, never surviving to enter the systemic circulation. Nickolls (1956) estimated this lost alcohol taken with food to be between 17 and 20 per cent and Alha (1951) gave the range as 10–20 per cent. Recent studies suggest that it is probably much less, around 5 per cent.

Owing to a thinner mucosa, a better blood supply and a larger surface area, the upper small intestine – the duodenum and jejunum – has the maximum capacity for absorption, compared with the gastric mucosa. This has practical implications, as drink taken by mouth will be absorbed more quickly when:

- a gastrectomy or gastroenterostomy has been carried out previously, as the drink will pass rapidly through to the upper small intestine
- the stomach is empty, as fluid will pass through the pylorus with almost no delay.

Conversely, when the stomach contains food, the drink will be held up until digestion has proceeded sufficiently for the contents to be released into the duodenum. A fatty meal will slow this process down even more and milk feed also has a marked delaying effect. Obviously some absorption will still occur in the interim via the gastric mucosa, but not at the same high rate as in the duodenum and jejunum. As well as delaying emptying, a full stomach will retard absorption by mixing with the alcohol and physically reducing its access to the gastric lining where transit into the blood takes place.

Another factor in the speed of absorption is the concentration of the alcohol. A strength of about 20 per cent is optimum for rapid absorption, which is met (in an empty stomach) by sherry or port wine, or spirits diluted with a 'mixer', such as gin and tonic or whisky and soda. It is also said that carbonated drinks (those containing dissolved carbon dioxide, such as champagne, tonic or soda water, or lemonade) hasten absorption, perhaps because the bubbles greatly increase the surface area carrying alcohol.

Dilute drinks, such as beer (with a concentration of about 4 per cent alcohol) will be absorbed much more slowly – probably because the large volume impedes access of the alcohol molecules to the stomach lining. Beer may take twice as long to absorb as stronger drinks, though part of the delay is caused by contained carbohydrates, which is another factor that slows absorption. For example, when whisky is diluted to the same strength as beer, absorption is more rapid and the peak is higher than drinking the same amount of alcohol in the form of beer.

Very strong drink slows the rate of transfer into the bloodstream. Neat spirits or liqueurs, which may be in excess of 40 per cent alcohol, cause:

- pyloric spasm and hence retarded emptying into the duodenum

- irritation of the gastric lining, forming a barrier of mucus, which slows absorption
- reduced gastric motility, which also retards emptying.

Given an empty stomach and an optimum concentration of alcohol, most of the drug will have entered the bloodstream between 30 and 90 minutes after drinking. According to research by Wilkinson *et al.* (1977), the average times to reach peak blood concentration were 22, 40, 55 and 60 minutes after drinking 11, 22, 36 and 45 g of alcohol in a bolus of similar volumes.

It has been calculated that 98 per cent of alcohol drunk would be absorbed within 10 minutes if it went straight into the small intestine – most of the delay is due to hold-up in the stomach. Rates vary greatly among different people and even in the same person at different times, irrespective of food being taken, but an acceptable mean time would be that 60 per cent of the imbibed alcohol would be absorbed within 60 minutes and 90 per cent within 90 minutes.

Food in the stomach can, however, at least double these times and a large fatty meal can delay total absorption for a number of hours. This has an important effect on the dynamic state between absorption and elimination, as the rate of the latter is relatively constant (see below) and can therefore deal with the slow delivery of alcohol from the portal blood so effectively that the peak of the blood-alcohol curve is low – indeed the curve becomes a long shallow curve instead of a sharp hillock. Where there is a legal threshold for blood alcohol in relation to driving, the taking of food can easily cause one person who has had a large meal to remain well under this limit, whereas another person who drank the same amount at the same time may rapidly exceed the threshold if he drank on an empty stomach. Incidentally, the same effect can occur, notwithstanding food, where a man and woman drink the same amount, as described earlier.

Some drugs will affect absorption rates, by modifying the speed of stomach emptying. Atropine, chlorpromazine, tricyclic antidepressants, procyclidine, amphetamines, morphine, antidiarrhoea compounds (e.g. Lomotil), codeine, methadone, heroin, pethidine, etc., will delay gastric transit, whilst the antiemetics cisapride and metoclopramide, as well as the antibiotic erythromycin, will hasten stomach emptying.

ELIMINATION OF ETHANOL

Almost all alcohol is detoxified by the liver, only 2–10 per cent being excreted unchanged. This means that a heavy drinking session places a great metabolic burden upon the liver and is the cause of hepatic damage after long-standing drinking. The elimination mechanism is an oxidation of alcohol by liver enzymes, through acetaldehyde to acetate. The first stage is performed by the enzyme alcohol dehydrogenase but, as the second stage is much more rapid, little acetaldehyde has time to accumulate. The acetic acid is rapidly oxidized further to carbon dioxide and water. Some may also be broken down by a microsomal oxidase system.

The rate of elimination is of crucial importance to the shape of the blood-alcohol curve, the height of its peak and to the duration of the alcoholaemia. It is also central to any attempts at retrospective calculations of blood, breath or urine levels, as discussed later. Whereas the rate of absorption is variable and is affected by a number of factors, the speed of detoxification in the liver is much more constant and relatively independent of external influences. This is not to say that it is fixed and immutable, even in the same person at different times, but it is capable of reasonable approximation. Much experiment and research has been devoted to this topic and the results, though variable, lie within a fairly narrow band.

Holford (1987) reviewed 11 different studies and found average rates varying between 12.6 and 26.8 mg/100 ml/hour. Neuteboom and Jones (1990) looked at levels in 1300 drivers stopped by the police, and calculated rates between 12 and 38 mg/100 ml/hour in 95 per cent of subjects. Several had rates over 48 mg/100 ml/hour. The elimination rate appeared to increase as the net blood concentration increased, for which several explanations are possible.

As a generalization, it has traditionally been assumed that blood alcohol declines, after the peak is reached, at a rate of around 15 mg/100 ml/hour, but more recent research, especially that of Holford, indicates that this should be raised to 18.7. This applies to healthy adults who are not habituated drinkers, and includes light to moderate drinkers and those ‘binge drinkers’ who may indulge heavily but intermittently. The enormous number of publications on alcohol must be consulted for details of the range of variation, but basically, elimination can vary from about 12 to 27 mg/100 ml/hour.

Taking the mean as around 18 mg, a man of average size can therefore destroy about 9 g alcohol/hour, with a range variation of between 7 and 16 g. This is about the same as the ‘unit’ of alcohol, a concept devised for convenience in estimating the daily or weekly intake of drinkers. A ‘unit’ is of 10 g and is contained (approximately) in half a pint of beer, one single measure of spirits, or a standard glass of table wine.

The habituated drinker, the ‘chronic alcoholic’, can eliminate far faster than the average person, at least until he suffers severe liver damage in the later stages of his addiction. Though there is some controversy about the experimental evidence, workers such as Bonnichsen (1966) have shown that the rate is greater in older people, at higher blood-alcohol levels, and in habituated drinkers. The rate in the latter may be in excess of 40 mg/100 ml/hour and work at

Lion Laboratories in Wales, who manufacture equipment for breath testing, have found rates over 50 mg/100 ml/hour in chronic alcoholics.

Between 90 and 98 per cent of ingested alcohol is removed from the blood by the liver, leaving a small residue to be excreted unchanged by the kidneys, lungs, sweat, salivary and mammary glands. The ethanol in glomerular filtrate is in equilibrium with plasma, but as water is absorbed in the renal tubules, the urine concentration is higher than the blood level at the time of filtration, the ratio being approximately 123:100. This means, for example, that the legal limit for driving in Britain of 80 mg/100 ml blood is taken to be 107 mg/100 ml in urine.

Unfortunately, it is obvious that, except in the highly unlikely circumstances of ureteric catheterization, the urine concentration can never accurately represent the blood concentration at any given time. The blood concentration is almost never static, but is either rising or falling, so the amount of alcohol in the glomerular filtrate is also constantly varying. It is being mixed in the bladder, however, with previously filtered urine and will also have that which is filtered later added to it, until the bladder is emptied – so it can only provide an average concentration for the time between two micturitions. An added error is that urine produced before drinking began (and which was therefore alcohol-free) may have already been in the bladder and will dilute the alcoholic urine. In many countries, where urine is used for drink–driving testing, the subject is instructed by the police to empty his or her bladder before collection is made over the subsequent hour, in order to avoid the dilution factor.

Breath is now used by many jurisdictions to measure alcohol intake, either as a screening test before blood is taken for analysis, or as an evidential method instead of blood or urine. There is still some controversy over the scientific accuracy of this method, but usually the results are so high that errors are immaterial – or in marginal results, more accurate blood testing is indicated.

Alveolar air at 37°C is in equilibrium with the pulmonary capillary plasma alcohol, the ratio being about 2300:1, volume to volume, for blood as against breath. There is some dispute as to the true ratio, which lies somewhere between 2100 and 2400. If sufficiently deep exhalation is made to drive out dead-space air, then the collected sample can be analysed to give a measure of the blood alcohol, though slight errors occur if there is incomplete elimination of dead-space air and a drop in temperature as the air travels through the dead space. In most countries using evidential breath testing, however, the offence is not in having a breath-alcohol level in excess of an equivalent blood level, but in having excess alcohol in the breath. This obviates defence ploys that would attempt to throw scientific doubt on the relationship between the two concentrations.

ALCOHOL CONCENTRATIONS: UNITS AND VARIOUS DRINKS

The concentration (often called the ‘level’) of alcohol in blood, urine and breath is expressed by a variety of metric units, which may lead to some confusion. The index most widely used for blood, urine and other body fluids is the weight of alcohol per volume of diluent – for example, milligrams per hundred millilitres (mg/100 ml). The expression ‘decilitre’ may be used instead of 100 ml (mg/dl). In some countries in continental Europe, alcohol concentration is expressed as ‘pro mille’, which is grams per litre (g/l), equivalent to milligrams per millilitre (mg/ml). Elsewhere, especially in the USA, a ‘percentage’ system is common, but can be ambiguous as it does not intrinsically state whether the percentage is volume/volume, volume/weight, weight/weight or weight/volume. Unless otherwise stated, it is assumed to be a weight/volume.

Breath is almost universally measured as micrograms per hundred millilitres (μg/100 ml). The matter of weight and volume is important in respect of alcohol concentrations. The specific gravity of alcohol is 0.79, the compound being appreciably lighter than water. In alcoholic drinks, the manufacturer’s description and labelling is almost always ‘volume/volume’ (v/v), but physiological calculations are made via the weight of alcohol in a given volume of body fluid (w/v). Therefore, especially for stronger alcoholic drinks, a conversion has to be made. For example, many spirits, such as whisky, may be labelled as 40 per cent v/v, but this would be only about 32 per cent weight/volume. For weak drinks, such as beer, it is hardly worth correcting the 4 per cent v/v, as calculations have a far greater intrinsic error from other factors.

Approximate strengths of common drinks are as follows (all v/v):

- beer, lager, stout: 3–5 per cent
- cider (variable): 5 per cent
- table wines (unfortified): 9–12 per cent
- fortified wines (sherry, port, vermouth): 18–20 per cent
- spirits (brandy, gin, whisky, rum, vodka): 37–42 per cent
- liqueurs (variable): 15–55 per cent.

The practice of using ‘units of alcohol’ has become popular in recent years, not so much for calculating concentrations, but for approximate estimates of intake, in relation to excessive drinking and the long-term medical consequences of alcohol consumption. A ‘unit’ is of the order of 10 per cent ethanol and very approximately delivered by ‘one drink’, where this is either a half pint of beer, one glass of table wine or one small measure of spirits. For example, it has been recommended that men should not exceed about 20 units per week and women 14, to avoid the risk of liver damage. It has recently been claimed that from statistical analysis of forensic autopsy material, the risk of coronary heart disease can be reduced by drinking 2 units a day (Thomsen 1995).

Calculation of blood levels from drink taken and the converse

The most important statement in this respect is to stress the utter unreliability and inaccuracy of attempting back-calculations in either direction. Only gross approximations can be achieved and no pretence at accuracy must be offered. In this book, we are not concerned with the controversial problems of trying to estimate blood or breath levels in living vehicle drivers at some time prior to an accident or other event, but with similar problems that can arise in fatal cases, especially in relation to drink and driving. In both criminal and civil disputes, evidence is often sought as to the alcoholic state of the deceased at some material time, based on calculations made from blood or urine alcohol analyses taken at autopsy. Less often, aviation, railway, diving and industrial fatalities may present the same potential problem. Criminal proceedings may arise because of alleged reckless driving on the part of another, when the drunken state of the deceased victim may offer some defence. In civil matters, often involving insurance companies, a significant blood-alcohol level may be used as contributory negligence.

Whatever the reason, the pathologist must offer interpretations of alcohol levels found at autopsy with caution, especially where retrospective calculations are requested. Less often, the pathologist may be asked what blood or urine levels might be expected at a certain time (for example, at the time of death) given a description and timetable of alcoholic drinks taken by the deceased. The same cautions against overprecise calculations must be offered here. In calculating approximate blood levels from a knowledge of the drink taken, there are several methods in use: either the well-known ‘Widmark factor’ or other calculations, which are really modifications or simplifications of the Widmark technique.

Widmark, in 1932, produced his well-known formula for calculating the total amount of alcohol in the body, from which knowing the body weight and assuming equilibration throughout the water compartment, the blood-alcohol level could be derived.

The Widmark equation is: $A = R \times P \times C$, where A is the total body alcohol, C the blood concentration, P the body weight in kilograms and R a factor, which is 0.68 in men and 0.55 in women. The sex difference is due to the different fat:water ratios, men having about 54 per cent and women 44 per cent water partition by weight.

Much research and modification of Widmark’s factors has been carried out. Hume and Fitzgerald (1985) claimed that the use of body water distribution was too complex. Gullberg and Jones (1994) have published data on a re-evaluation of the Widmark equation, in which they claim to be able to estimate from a single blood-alcohol determination, the amount of alcohol consumed, within an error of ±20 per cent. Pounder and Kuroda (1994) claim that the use of vitreous fluid (which is sometimes used in decomposed or damaged bodies in place of blood) to predict the blood-alcohol concentration, is too variable to be of practical use.

A useful approximate calculation, derived from Widmark, is that an intake of 0.2 g of alcohol per kilogram body weight is likely to result in a blood-alcohol concentration in men of about 25 mg/100 ml.

The following facts are of use:

- The average rate of decline of blood alcohol after the peak of the curve is reached, may be taken as about 15 mg/100 ml/hour, though recent research suggests 18 as more accurate.
- The weight of alcohol imbibed may be calculated from knowledge of the v/v strength of the liquor and the amount taken. For example, if a ‘double’ British measure of 40 per cent v/v whisky is drunk, then 15 ml will contain $(40 \times 0.8) = 32$ per cent alcohol w/v in $15\text{ ml} = 4.8\text{ g}$.
- The weight of alcohol/kg body weight is calculated.
- A ratio of 0.2 g alcohol/kg body weight will produce a blood level of approximately 25 mg/100 ml in a man, assuming an empty stomach.
- In women, the level so produced may be 20–25 per cent higher.
- If only beer is drunk, the peak will be considerably less, sometimes only 50 per cent that produced from wine or spirits.
- Drinking during or after a meal markedly flattens the blood–alcohol curve.

PHYSIOLOGICAL EFFECTS OF ALCOHOL

The pathologist is frequently asked in either written opinions or in court testimony to give an estimate of the behavioural state of a victim at a certain level of blood alcohol or after having taken a specified amount of drink. Though he is usually not qualified as an expert on alcoholism in any

TABLE 28.1 *General spectrum of behaviour (mg/100 ml of alcohol in blood)*

Above 30	Impairment of complex skills such as driving
30–50	Definite deterioration in driving ability
50–100	Objective signs such as loquaciousness, progressive loss of inhibitions, laughter and some sensory disturbance
100–150	Slurred speech, unsteadiness, possible nausea
150–200	Obvious drunkenness, nausea, staggering gait
200–300	Stupor, vomiting, possibly coma
300–350	Stupor or coma, danger of aspirating vomit
Over 350	Progressive danger of death from respiratory centre paralysis

clinical sense, he will be a medical practitioner with general knowledge and some personal experience of alcoholic behaviour, from his pre-pathology years. He thus can give a general opinion to assist the court, but unless he has special experience of the matter, he should not extend himself into detailed clinical expositions, which are the province of the psychiatrist with an interest in alcoholism, a police surgeon or a casualty officer, all of whom deal frequently with drunken patients. A general level of knowledge can be offered to the lawyer, police or court, however, especially in respect of the usual level of capability and consciousness at different blood-alcohol levels (Table 28.1).

The list in Table 28.1 is probably 'worst case' in nature, as many people show less effects at given blood levels. Women are usually more affected at lower levels, though the 'world record' for survival may be a woman who survived a blood alcohol of 1510 mg/100 ml (Johnson *et al.* 1982). Many drivers stopped by the police at random road blocks in Australia had blood levels over 500 mg/100 ml.

MODES OF DEATH IN ACUTE ALCOHOLISM

Death from alcoholic poisoning is not uncommon and can occur at blood levels in excess of about 300 mg/100 ml. Johnson *et al.* (1982) claimed that some deaths could be attributed to alcohol at even lower concentrations. Death can be caused either by the direct depressive effects upon the brainstem, mediated via the respiratory centres – or through secondary events such as aspiration of vomit. As discussed in Chapter 14, the use of 'aspiration of vomit' as a cause of death must be used with great caution unless there is ante-mortem eyewitness evidence. The major exception to this proviso is in acute alcoholism, where if copious inhalation of stomach contents right down to the secondary bronchi is confirmed, then in the absence of significant natural disease, injury or other toxicity, a high blood-alcohol level may reasonably be incriminated as the probable cause. Many such fatalities occur during police custody, when considerable outcry, publicity and disciplinary investigations are the usual outcome.

Drunken persons are often involved in fatal trauma, which may be of many types. The majority of homicides are triggered by the aggressive behaviour engendered by alcohol. Road accidents, either caused by drunken drivers (often upon themselves) or by drunken pedestrians walking into traffic, are commonly related to alcoholic vulnerability. Falls are extremely frequent and often fatal. Drunken persons may fall down stairs or steps and suffer head injuries. Falls from high places are less common, but do occur from drunken carelessness or unsteady gait – exemplified by the notorious inquest on Helen Smith in Leeds in 1982.

Death from burns or carbon monoxide poisoning may occur in drunken persons who smoke when intoxicated. A common scenario is for a drunk to go to bed and fall asleep whilst smoking, the cigarette igniting the bedclothes. Sometimes, a gas, electric or kerosene heater may be knocked over during drunken staggering, which again starts a fatal fire.

Drowning is seen occasionally, especially in river or dockland areas. A typical happening is for a drunken sailor to return to his ship late at night, and fall from a bridge or gangway into the water. Death is sometimes not caused by drowning, but by sudden vagal cardiac arrest from the shock of hitting cold water or having cold water suddenly flood the pharynx and larynx. The drunken state seems to sensitize the victim to such vasovagal shock, perhaps because of the marked cutaneous vasodilatation encouraged by alcohol. The author (BK) has seen a number of such tragedies in Cardiff dockland, including a sailor and his woman companion, who both fell off the ship after carousing.

CHRONIC ALCOHOLISM

The pathological features of this condition are extensive and can only be surveyed briefly here. In this context, chronic alcoholism refers to the steady, regular abuse of drink, rather than intermittent 'binge' drinking, which gives the tissues time to recover between bouts of acute alcoholism.

At autopsy, there may be signs of general neglect and malnutrition, but many chronic alcoholics are obese or even oedematous, the latter because of chronic heart failure. The specific lesions are in the liver, heart and brain, though they may be difficult to identify as unequivocally caused by alcohol. The early stages of liver damage cause fatty change, usually with enlargement. The normal weight, according to the sex and build, is between 1300 and 1600 g, but a fatty liver may be well in excess of 2000 g.

The surface is pale and greasy, though this may not be a uniform change, especially in early or less severe cases. Patchy yellowish areas may be visible within normal hepatic parenchyma.

If the abuse continues, then the fatty change may eventually give way to fibrosis, the liver surface becoming rippled beneath its capsule. Such cirrhosis is fairly fine, with nodules of the order of 5–10 mm in diameter. In the later stages the liver becomes smaller and contracts to a hard, greyish-yellow block of only 800–1200 g.

Without a history of long-term alcohol abuse, it is difficult or impossible to be definite about the aetiology purely on autopsy appearances, though suspicion may be strong. A similar liver may develop as a sequel to hepatitis – and less often as an end result of certain dietary or metabolic defects. The spleen may be enlarged and firm, and portal varices may be present at the gastro-oesophageal junction, but these are both manifestations of portal hypertension and do not assist in determining the precise aetiology of the hepatic fibrosis. A useful index of liver damage and the progression or remission of alcoholic impairment is the level of the enzyme γ-glutamyl transpeptidase in the serum. Normal levels are less than 36 units, whereas liver damage can elevate this by a factor of many times. Alcoholic cardiomyopathy is certainly a real entity and can be diagnosed clinically. Whether it can be definitely identified on histological appearances alone is a matter of dispute. The heart is enlarged and shows patchy fibrosis with a variable mixed cellular infiltrate, hypertrophy of muscle fibres, patchy necrosis, hyalinization, oedema and vacuolization. Nuclear enlargement and polymorphism complete the range of changes, but none of these are specific, being found in hypertensive heart disease, coronary stenosis and other types of myocarditis. Combined with a definite history of chronic alcoholism, however, these relatively non-specific changes can be ascribed to alcohol if other causes can be excluded (see also Chapter 25). More specific myocardial damage has been caused by cobalt added to commercial beers and several outbreaks are on record. Systemic fat embolism has also been recorded in victims of alcoholic fatty liver. Microinfarcts in myocardium and brain are possible, though this aspect has so far been rather neglected in research. Fat stains are not usually employed in routine investigations and it is not known how the victims of a fatty liver compare with control subjects in respect of diffuse target organ embolism.

REFERENCES AND FURTHER READING

Al Lanqawi Y, Moreland TA, McEwen J, *et al.* 1992. Ethanol kinetics: extent of error in back extrapolation procedures [see comments]. *Br J Clin Pharmacol* **34**:316–21.

Alha A. 1951. Blood alcohol and inebriation in Finnish men. *Am Acad Sci (Fen Series A)* **V**:Medica.

Bonnichsen R. 1966. Oxidation of alcohol. *Q J Stud Alcohol* **27**:554–60.

Brownlie AR, Walls HJ. 1985. *Drink, drugs and driving*, 2nd edn. Sweet and Maxwell, London.

Buono MJ. 1999. Sweat ethanol concentrations are highly correlated with co-existing blood values in humans. *Exp Physiol* **84**:401–4.

Chikasue F, Yashiki M, Miyazaki T, *et al.* 1988. Abnormally high alcohol concentration in the heart blood. *Forensic Sci Int* **39**:189–95.

Dick GL, Stone HM. 1987. Alcohol loss arising from microbial contamination of drivers' blood specimens. *Forensic Sci Int* **34**:17–27.

Freireich AW, Bidanset JH, Lukash L. 1975. Alcohol levels in intracranial blood clots. *J Forensic Sci* **20**:83–5.

Gibson AG. 1975. Alcohol can be absorbed through the respiratory tract (a case report). *Med Sci Law* **15**:64.

Grant SA, Millar K, Kenny GN. 2000. Blood alcohol concentration and psychomotor effects. *Br J Anaesth* **85**:401–6.

Gullberg RG, Jones AW. 1994. Guidelines for estimating the amount of alcohol consumed from a single measurement of blood alcohol concentration: Re-evaluation of Widmark's equation. *Forensic Sci Int* **69**:119–30.

Helander A, Jones AW. 2002. [5-HTOL – a new biochemical alcohol marker with forensic applications.] *Lakartidningen* **99**:3950–4.

Holford NH. 1987. Clinical pharmacokinetics of ethanol. *Clin Pharmacokinet* **13**:273–92.

Hume DN, Fitzgerald EF. 1985. Chemical tests for intoxication: what do the numbers really mean? *Anal Chem* **57**:876A–878A, 882A, 884A passim.

Johnson RA, Noll EC, Rodney WM. 1982. Survival after a serum ethanol concentration of 1½% [letter]. *Lancet* **2**:1394.

Jones AW. 2000. Aspects of in-vivo pharmacokinetics of ethanol. *Alcohol Clin Exp Res* **24**:400–2.

Jones AW, Jonsson KA, Kechagias S. 1997. Effect of high-fat, high-protein, and high-carbohydrate meals on the pharmacokinetics of a small dose of ethanol. *Br J Clin Pharmacol* **44**:521–6.

Kalant H. 2000. Effects of food and body composition on blood alcohol curves. *Alcohol Clin Exp Res* **24**:413–14.

King LA. 1983. Nomograms for relating blood and urine alcohol concentrations with quantity of alcohol consumed. *J Forensic Sci Soc* **23**:213–17.

Kubo S, Dankwarth G, Püschel K. 1991. Blood alcohol concentrations of sudden unexpected deaths and non-natural deaths. *Forensic Sci Int* **52**:77–84.

Lahti RA, Vuori E. 2002. Fatal alcohol poisoning: medico-legal practices and mortality statistics. *Forensic Sci Int* **126**:203–9.

Lieber CS. 2000. Ethnic and gender differences in ethanol metabolism. *Alcohol Clin Exp Res* **24**:417–18.

Lieber CS, Davidson CS. 1962. Some metabolic effects of ethyl alcohol. *Am J Med* **33**:319–27.

Logan BK, Case GA, Distefano S. 1999. Alcohol content of beer and malt beverages: forensic consideration. *J Forensic Sci* **44**:1292–5.

Mason JK, Blackmore DJ. 1972. Experimental inhalation of ethanol vapour. *Med Sci Law* **12**:205–8.

Neuteboom W, Jones AW. 1990. Disappearance rate of alcohol from the blood of drunk drivers calculated from two consecutive samples; what do the results really mean? *Forensic Sci Int* **45**:107–15.

Nickolls LC. 1956. *The scientific investigation of crime.* Butterworths, London.

Paton A. 1988. *ABC of alcohol.* British Medical Journal publication, London.

Piette M, Timperman J, Vanheule A. 1986. Is zinc a reliable biochemical marker of chronic alcoholism in the overall context of a medico-legal autopsy? *Forensic Sci Int* **31**:213–23.

Plueckhahn VD. 1968. Alcohol levels in autopsy heart blood. *J Forensic Med* **15**:12–21.

Plueckhahn VD, Ballard B. 1968. Factors influencing the significance of alcohol concentrations in autopsy blood samples. *Med J Aust* **1**:939–43.

Pounder DJ, Kuroda N. 1994. Vitreous alcohol is of limited value in predicting blood alcohol [see comments]. *Forensic Sci Int* **65**:73–80.

Roine R. 2000. Interaction of prandial state and beverage concentration on alcohol absorption. *Alcohol Clin Exp Res* **24**:411–12.

Sunter JP, Heath AB, Ranasinghe H. 1978. Alcohol associated mortality in Newcastle upon Tyne. *Med Sci Law* **18**:84–9.

Thomsen JL. 1995. Atherosclerosis in alcoholics. *Forensic Sci Int* **75**:121–31.

Thomsen JL, Felby S, Theilade P, *et al.* 1995. Alcoholic ketoacidosis as a cause of death in forensic cases. *Forensic Sci Int* **75**:163–71.

Wilkinson PK, Sedman AJ, Sakmar E, *et al.* 1977. Pharmacokinetics of ethanol after oral administration in the fasting state. *J Pharmacokinet Biopharm* **5**:207–24.

Winek CL, Eastly P. 1976. Factors affecting contamination of blood samples for ethanol determination. In: Wecht C (ed.), *Legal medicine annual 1976.* Appleton Century Crofts, New York.

Winek CL, Esposito FM. 1985. Blood alcohol concentrations; factors affecting predictions. In: Wecht C (ed.), *Legal medicine annual 1985.* Praeger Scientific, New York.

Winek CL, Jr, Winek CL, Wahba WW. 1995. The role of trauma in postmortem blood alcohol determination [see comments]. *Forensic Sci Int* **71**:1–8.

CHAPTER 29

Carbon monoxide poisoning

■ Causes of carbon monoxide poisoning 560
■ The autopsy in carbon monoxide poisoning 562
■ Blood analysis 563
■ References and further reading 564

CAUSES OF CARBON MONOXIDE POISONING

After alcohol and some pharmaceutical drugs, carbon monoxide poisoning is probably the most common toxic condition to be met with in routine forensic pathology. The widespread introduction of natural gas (which contains no carbon monoxide) as a replacement for 'coal gas' as a heating fuel has removed a major source of the poison. Monoxide still provides lethal dangers in many other ways, however. It is produced whenever fossil fuels are incompletely oxidized to carbon dioxide and, because of its great affinity for haemoglobin, even low concentrations can be cumulative.

There are a vast number of publications on carbon monoxide toxicity, both in an environmental and industrial context. Once again, this discussion is confined to the pathologist's involvement in fatalities. A body coming to autopsy with suspected (or sometimes unsuspected) carbon monoxide poisoning, will have suffered that toxic condition by the following means.

Motor vehicle exhaust gases

Since the replacement of coal gas (containing up to 7 per cent monoxide) with natural gas, a major means of suicide has been removed. 'Putting the head in the gas oven' was the most common form of self-destruction in Britain and many other countries until this changeover. In 1961 in the UK, for example, there were 2711 suicides and 1014 accidental deaths from carbon monoxide. Since then, suicides have obtained carbon monoxide to kill themselves from the internal combustion engine. Petrol (gasoline) engines produce up to 5–7 per cent carbon monoxide in their exhaust fumes, and more if the engine is idling, defective or improperly tuned. Diesels produce far less monoxide than petrol engines. Normally, the exhaust pipe vents the gases into the atmosphere, where it may contribute to considerable low-level monoxide contamination in large cities, such that policemen on traffic duty may have up to 10 per cent saturation of their haemoglobin. When the exhaust fumes are confined to a small space, then a dangerous or lethal level can build up in a short space of time. It is calculated that a 1.5 litre petrol engine, idling in a closed single garage, can produce a lethal concentration in the atmosphere within 10 minutes.

Some suicides will merely sit in the garaged car with a window open and allow the gas to overtake them. More commonly, some device is fitted to pipe the gas into the interior of the car and this may be done outside a garage, often in a remote parking spot. The flexible tube from a vacuum cleaner seems a favourite means, though ordinary hose-pipe, pushed inside the tail-pipe and led in through a window, is a common alternative. Apart from suicide, accidental poisoning sometimes occurs in relation to internal combustion engines. A mechanic or car owner may work on a vehicle in a closed garage, especially in cold weather, and be overcome by monoxide before he is able to remove himself from the danger. The insidious nature of the toxic effects may make him unaware of the supervening stupor and coma.

Carbon monoxide can also affect drivers of a moving vehicle, usually because of a defective exhaust system, which allows gas to percolate through the floor or engine bulkhead into the interior. Rarely, a strong following wind blows the external exhaust gas into the open doors of a van or truck. Another motoring cause is a leak in the heat exchanger in those vehicles that use a direct air supply from around the exhaust manifold to provide passenger heating.

In aircraft – usually light planes with the engine immediately adjacent to the cockpit – a leakage of exhaust gas from the motor compartment can lead to disablement of the pilot and death from a crash rather than from fatal toxicity. The same applies to those road-vehicle drivers who are overcome by the gas to an extent that makes them incapable of driving safely. The differential diagnosis in survivors is between carbon monoxide and alcohol, as the clinical symptoms of both are quite similar at one stage of the toxic process.

Deaths have also been reported in 'scuba' divers, whose air cylinders have been contaminated with carbon monoxide during refilling from faulty petrol-driven compressors.

Domestic appliances

Domestic heating appliances can produce carbon monoxide in addition to the usual dioxide from a restriction of their air supply, so that incomplete combustion of the fuel takes place. The fuel can be natural gas, which, though itself free from monoxide, is only partially oxidized from some defect of design, maintenance, or patency of the exhaust flue. Solid-fuel boilers used for central heating may have some restriction of air entry or a partial blockage of the chimney system. Paraffin (kerosene) heaters may burn with inadequate oxygen ingress and any other form of hydrocarbon fuel appliance can malfunction so that part of the products of combustion are monoxide.

The author (BK) has dealt with several deaths that occurred because the faulty installation of natural-gas fires led to absent or inadequate ventilation, causing monoxide to back-diffuse into living rooms. Gas appliances, especially water heaters, are one reason why the bathroom is such a dangerous place. Apart from the extra hazards of electricity, a bath for drowning, wet surfaces for slipping, tablets in the cabinet and sharp instruments such as razor blades, the small-sized room and the frequent installation of a gas water heater or 'geyser' makes this room a frequent locus for unnatural death. Blockage of the exit or 'flue' pipe is a common fault, sometimes from bad installation or because it becomes blocked by soot or by birds' nests.

Structural fires

A common cause of monoxide deaths arises from structural fires in houses and other buildings. As described in Chapter 11, the majority of deaths in house conflagrations are caused not by burns, but by inhalation of smoke. These fatalities are largely caused by carbon monoxide poisoning, though other lethal gases such as cyanide, phosgene and acroleins are partly responsible. Many victims of house fires die remote from the flames, and may be overcome in different rooms or even on different floors, the monoxide percolating considerable distances and killing persons either asleep or trapped elsewhere in the building.

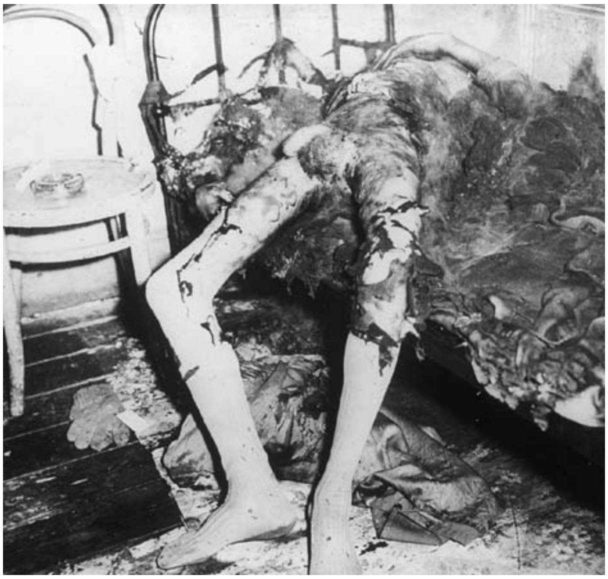

FIGURE 29.1 *Carbon monoxide poisoning in a chronic alcoholic. This is a relatively common event when a drunken person collapses with a lit cigarette across his bed (note the ashtray and cigarette packet). The burns are post-mortem, death being caused by a high level of carboxyhaemoglobin generated by the smouldering bed coverings.*

Industrial processes

Many industrial processes may lead to monoxide poisoning, especially in iron and steel works, where producer gas and water gas are deliberately formed and stored as part of the manufacturing process. Water gas can contain up to 40 per cent carbon monoxide and in former days was added to town gas for the domestic supply, which greatly enhanced the monoxide content of some 7 per cent from coal gas. Many other industrial processes, such as the Mond method of nickel production, use carbon monoxide, as well as the universal danger of any heating process producing the gas during combustion. In coal-mining, carbon monoxide is one of the gases that presents a constant threat; it escapes from the coal seams themselves but is also produced from the frequent small fires that smoulder in the recesses of the mines.

Incomplete combustion

Incomplete combustion of a gas flame from any gaseous fuel can produce carbon monoxide. Where such a flame impinges on a cold metal surface or where that surface is coated with soot, partial oxidation of the fuel supply leads to monoxide production. In appliances fuelled from the common butane or propane cylinder sources, as in caravans, campers and boats, maladjustment or restriction of ventilation can lead to slow but insidious production of monoxide. Deaths of whole families have occurred under such circumstances, as they are exposed overnight to slow accumulation of carbon monoxide from refrigerators and other appliances.

THE AUTOPSY IN CARBON MONOXIDE POISONING

As always, an adequate history may give the clue to the cause of death. Where the circumstances are obvious, such as a death in a car with a tube leading from the exhaust, the pathologist's attention will be directed towards carbon monoxide from the outset. There are many other circumstances, however, in which the history may be obscure, and only the vigilance of the pathologist and his staff will pick up the possibility of this type of toxicity. In fact, many a case has first been recognized by the mortuary technician who commented on the colour of the skin or tissues. A number of fatal monoxide poisonings have already been certified as 'natural causes' by uncritical clinicians (especially in general practice) when in fact carbon monoxide poisoning, either suicidal or accidental, was the true reason for the death. Failure to examine the body fully is the usual cause, as the pink coloration may only be noticeable in the areas of postural hypostasis not normally visible in a body in bed, though the sides and back of the neck are reasonably accessible even to cursory examination. The author (BK) recollects at least one suicide where the relatives placed the body back in bed and had it certified as 'coronary thrombosis' by an unsuspecting family doctor.

At autopsy the most striking appearance of the body is the colour of the skin, especially in areas of post-mortem hypostasis. The classical 'cherry-pink' colour of carboxyhaemoglobin is usually evident if the saturation of the blood exceeds about 30 per cent. Below this, familiarity and good lighting are needed and below 20 per cent, no coloration is visible. As these low concentrations are rarely fatal, however, little is lost. Sometimes, darker cyanosis tends to mask the skin colour, but the margins of the hypostasis and the internal tints are usually apparent.

When the victim is anaemic, the colour may be faint or even absent because insufficient haemoglobin is present to display the colour. In racially pigmented victims the colour may obviously be masked, though may still be seen on the inner aspect of the lips, the nail-beds, tongue, and palms and soles of the hands and feet. It is also seen inside the eyelids, but rarely in the sclera.

Rarely, there may be blistering of the skin of dependent areas, such as the calves and buttocks, and around wrists and knees, similar to the so-called 'barbiturate blisters'. These are not specific to carbon monoxide toxicity and are the result of cutaneous oedema in any profound coma where there is total immobility and lack of venous return from muscle movement. As most carbon monoxide deaths are relatively rapid, such blisters are rare. Internally the most noticeable feature is again the colour. Blood and muscle will be pink as a result of carboxyhaemoglobin and carboxymyoglobin. In relatively low concentrations, in poor light and in some artificial lighting in autopsy rooms, the cherry-pink colour may be difficult to see. It can be enhanced by diluting the blood with water against a white background, as in a porcelain sink or an enamel scale-pan, when the pinkness will be more evident. The pinkness of hypothermia or refrigeration is a different colour. Unfortunately, some pathologists have a red-colour visual impairment, which makes it difficult for them to differentiate between subtle changes in redness, so laboratory analysis is always required for objective confirmation and quantification of the monoxide concentration. Other suggestive indications of carbon monoxide are that, when tissues are placed in formol saline for preservation of histology, they do not decolourize as quickly as normal tissues and remain pink for a long period. If carbon monoxide poisoning is suspected at autopsy, a quick test is to add a few drops of blood to some 10 per cent sodium hydroxide solution on a white tile or in a tube against a white background. The normal blood will immediately become brownish-green but, if significant monoxide is present, the colour will remain pink, as no methaemoglobin is formed. One has to bear in mind though, that unlike adult blood, fetal blood up to the age of 6 months is more resistant to alkali and the colour change can take hours to develop. Another simple test, mentioned in 1857 by Hoppe, is to take 3 ml blood into a test tube and dilute it with 6 ml tap water and to heat it for a short while with care in a water bath. Carbon monoxide blood turns brick red whereas control blood becomes greyish-brown. However, these crude tests are not recommended as an alternative to proper analysis.

Other signs at autopsy are non-specific. Pulmonary oedema is usually present. The author (BK) has the impression that the white matter of the brain remains unusually firm and that the whole brain after removal from the skull keeps its shape better, being rather stiff and rigid. This may be caused either by oedema or by some minimal fixation change in the brain tissue.

None of these rather idiosyncratic tests can replace proper laboratory investigation, however. Blood should be taken for analysis in the usual way, preferably from a peripheral vein. In contrast to most toxicological investigations, even foul samples can still be useful for carboxyhaemoglobin estimations. Heart blood, blood from body cavities and even from bone marrow when the bones are split open, can still provide valid material for estimating the percentage of haemoglobin converted to carboxyhaemoglobin. It is the

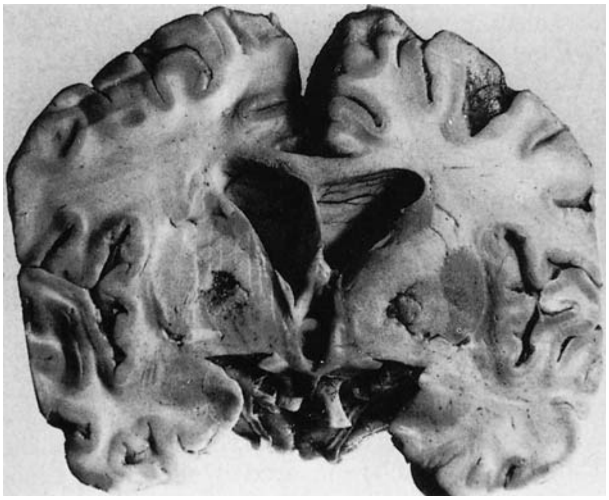

FIGURE 29.2 *Basal ganglia degeneration after carbon monoxide poisoning. The patient survived for a number of weeks, but succumbed to bronchopneumonia. There are cystic lesions in the globus area of both cerebral hemispheres.*

ratio of haemoglobin to the carboxy-type that is required, rather than an absolute concentration in the blood, so virtually any blood-containing fluid will provide this answer. When a body has been badly damaged by fire, fluid blood may be hard to obtain, so that any sanguineous body fluid or bone marrow can provide material for analysis. If analysis is to be delayed more than a day, it is recommended that fluoride be added as a preservative.

There are no other autopsy features in acute deaths, which form the majority seen by forensic pathologists. Where survival occurs there are a number of neurological lesions that follow severe carbon monoxide exposure. Necrosis and cavitation of the basal ganglia in the brain, notably the putamen and globus pallidus, has been known for well over a century. Within about 5 days, histological changes occur here, with 'gitterzellen' scavenging cellular debris, the foam cells and microglia presenting an appearance characteristic of tissue breakdown in the central nervous system. There may also be damage in the substantia nigra of the brainstem. In delayed deaths, petechiae and ring-shaped haemorrhages may occur in the cerebral white matter. All these changes, together with the clinical symptomatology, are exhaustively described in neurological publications.

There are also tiny focal necroses in the myocardium in delayed deaths from this cause, and similar lesions were described by Korb and David (1962) even in acute deaths. Frank myocardial infarction has often been reported after severe monoxide exposure and relative hypoxia, usually in the presence of pre-existing coronary disease, is probably the reason.

BLOOD ANALYSIS

Carbon monoxide has an affinity for haemoglobin that is between 200 and 300 times greater than oxygen. Therefore even small concentrations of monoxide in the inspired air will progressively displace oxygen from the erythrocytes and this will lower the oxygen-carrying capacity of the blood. It was formerly thought that all the toxic properties of carbon monoxide lay in this hypoxic action, but more recently it has been shown to interfere with other ferroproteins such as myoglobin and various enzymes including members of the cytochrome family.

The major effect is undoubtedly the reduction in oxygen transport and, for this reason, it is the percentage saturation of the total available haemoglobin that is important rather than an absolute quantity of carboxyhaemoglobin in the blood. Anaemic persons will not show the external cherry-pink colour if insufficient haemoglobin is available for the monoxide component to become visible. It is the residual non-combined haemoglobin still available for oxygen transport that is important in maintaining life. For example, an anaemic person with only 8 g/100 ml of haemoglobin having four of those occupied with carbon monoxide (50 per cent saturation) is in much worse a state than a person with a total of 14 g of which four are monoxide-occupied. There are many tables available that relate saturation levels to the clinical symptoms. There is a wide margin of variation in these, as there is with lethal levels.

Broadly speaking, the saturation levels found in fatal cases reflect both the physical state of the victim and the rapidity with which the poisoning took place. Robust, healthy adults under 60 years of age rarely die at saturation level under 50–60 per cent unless the concentration of carbon monoxide in the inspired air was so great that death occurred before gas exchange in the lungs was able to progress sufficiently to enter all available monoxide into the red cells. Death may then supervene with post-mortem saturation levels of as low as 40 per cent. Where absorption is slow and life persists longer (for there seems to be a relationship between air concentration of monoxide and length of survival), autopsy samples may show 80 per cent saturation.

Old people may die at relatively low concentrations, such as 30 per cent and, in some cases, no other cause can be found when the carboxyhaemoglobin level is only 25 per cent. This may be the result of anaemia, so that there is less reserve for oxygen carrying when part of the haemoglobin is occupied by monoxide. In many cases, the senile myocardium is already in a fragile state and any extra hypoxia will cause it to fail. Infants also seem to die at relatively low levels, perhaps because their higher respiration rate allows a more rapid absorption. The variation in fatal concentrations is wide and quite irregular. This is best exemplified when two or more victims have died together in the same environment. The carboxyhaemoglobin concentrations of bodies lying alongside each other can sometimes be totally different, even in persons of the same age group and physical health. Any disease process can contribute to death at lower monoxide concentrations. Coronary artery disease, respiratory insufficiency and other debilitating conditions allow death to occur earlier, when the toxic level is still relatively low. Conditions that are worsened by hypoxia, such as myocardial insufficiency or obstructive airway disease, are particularly potent. In addition to disease, any separate toxic state is additive, such as hypnotic drugs or alcohol.

The method of analysis has traditionally been by reversion spectroscopy, but this is a crude technique in which accuracy cannot be attained to within 10 per cent of the true value. For most purposes related to fatalities, this is good enough, as more exact results do not add anything to the investigation. The only problem can be at the lower end of the scale, when it is difficult to be certain about positivity under about 15 per cent. Modern analytical methods include non-dispersive infrared spectrophotometry and gas chromatography. Carboxyhaemoglobin is stable and can be detected even in putrefied bodies a long time after death, as long as sophisticated laboratory techniques are used. Monoxide cannot enter a body post-mortem to any significant extent. Bodies burnt after death do not absorb any monoxide and thus a significant level (more than 10 per cent) in a body from a fire means that respiration must have been proceeding whilst the conflagration was in progress. As emphasized in Chapter 11, the converse is often not true, in that deaths from rapid fires, especially flash fires caused by petrol, for example, may have no carboxyhaemoglobin in the blood. The possibility of carboxhaemoglobin being formed in the blood post-mortem has been investigated as a possible source of error (Kojima *et al.* 1982) and fluorided samples are recommended.

A minor source of local carboxyhaemoglobin and myohaemoglobin is gunshot wounds, where the propellant gases (which are rich in the compound) are blown into the wound from contact or short-range discharges. The tissues absorb monoxide, especially around the entrance wound. It has been claimed that this is a way of differentiating the entrance from the exit wound, but in fact the whole track of the missile(s) may be sheathed in a zone of monoxide transfer into blood and muscle from a near-discharge, though theoretically there should be a gradient of concentration from one end of the track to the other (see Chapter 8).

Many other war, military and terrorist circumstances are associated with exposure to carbon monoxide. Large quantities are released from the firing of guns and, in enclosed spaces, can become a hazard. Detonation of explosives in places where there is little ventilation also allows monoxide to accumulate and it may persist long after the physical dangers of the explosion have passed.

REFERENCES AND FURTHER READING

Annotation. 1962. Carbon monoxide from paraffin stoves. *Br Med J* **1**:312.

Bohnert M, Weinmann W, Pollak S. 1999. Spectrophotometric evaluation of postmortem lividity. *Forensic Sci Int* **99**:149–58.

Bowen DAL. 1986. Acute renal failure in carbon monoxide poisoning. *J Forensic Med* **7**:78–82.

Chang MY, Lin JL. 2001. Central diabetes insipidus following carbon monoxide poisoning. *Am J Nephrol* **21**:145–9.

Chaturvedi AK, Smith DR, Canfield DV. 2001. Blood carbon monoxide and hydrogen cyanide concentrations in the fatalities of fire and non-fire associated civil aviation accidents, 1991–1998. *Forensic Sci Int* **121**:183–8.

Garland H, Pearce J. 1967. Neurological complications of carbon monoxide poisoning. *Q J Med* **36**:445–55.

Goldbaum LR. 1977. Is carboxyhaemoglobin concentration the indicator of carbon monoxide toxicity? In: Wecht C (ed.), *Legal medicine annual 1976*. Appleton Century Crofts, New York.

Goldsmith JR, Landaw SA. 1968. Carbon monoxide and human health. *Science* **162**:1352–9.

Gottfried JA, Chatterjee A. 2001. Carbon monoxide-mediated hippocampal injury. *Neurology* **57**:17.

Helpern M, Strassman G. 1943. Differentiation of fetal and adult human hemoglobin. *Arch Pathol* **35**:776.

Hoppe F. 1857. Über die Einwirkung des Kohlenoxydgases auf das Hämatoglobin. *Virchows Arch* **11**:228.

King LA. 1983. Effect of ethanol in fatal carbon monoxide poisonings. *Hum Toxicol* **2**:155–7.

Kojima T, Nishiyama Y, Yashiki M, *et al.* 1982. Postmortem formation of carbon monoxide. *Forensic Sci Int* **19**:243–8.

Kojima T, Okamoto I, Yashiki M, *et al.* 1986. Production of carbon monoxide in cadavers. *Forensic Sci Int* **32**:67–77.

Korb A, David K. 1962. Fluorescenzmikroskopische und elektronmikroskopische Untersuchungen am

Herzmuskel der Ratte nach Leuchtgasvergiftungen. *Deutsch Z Gesamt Gericht Med* **52**:549–55.

Kunsman GW, Presses CL, Rodriguez P. 2000. Carbon monoxide stability in stored postmortem blood samples. *J Anal Toxicol* **24**:572–8.

Lee CW, Yim LK, Chan DT, *et al.* 2002. Sample pre-treatment for co-oximetric determination of carboxyhaemoglobin in putrefied blood and cavity fluid. *Forensic Sci Int* **126**:162–6.

Levine B, Moore KA, Fowler D. 2001. Interaction between carbon monoxide and ethanol in fire fatalities. *Forensic Sci Int* **124**:115–16.

Lo Menzo G. 1966. Il comportomento dell'attivita citrocromossidasica a succinodeidrogenase di taluni organi nell'avvelenamento acute da co. *Folia Med* **8**:607–12.

Mant AK. 1960. Accidental carbon monoxide poisoning: A review of 100 consecutive cases. *Med Leg J* **28**:30–6.

Miro O, Casademont J, Barrientos A, *et al.* 1998. Mitochondrial cytochrome *c* oxidase inhibition during acute carbon monoxide poisoning. *Pharmacol Toxicol* **82**:199–202.

O'Donnell P, Buxton PJ, Pitkin A, *et al.* 2000. The magnetic resonance imaging appearances of the brain in acute carbon monoxide poisoning. *Clin Radiol* **55**:273–80.

Pearce J. 1968. Coal gas and the brain [leading article]. *Br Med J* **1**:767.

Petelenz T. 1972. Badania doswiadczaine dzilania. *Arch Med Sadowej Kryminol* **12**:219–25.

Seto Y, Kataoka M, Tsuge K. 2001. Stability of blood carbon monoxide and hemoglobins during heating. *Forensic Sci Int* **121**:144–50.

Simpson K. 1955. Carbon monoxide poisoning. *J Forensic Med* **2**:5–13.

Somogyi E, Balogh I, Rubanyi G, *et al.* 1981. New findings concerning the pathogenesis of acute carbon monoxide (CO) poisoning. *Am J Forensic Med Pathol* **2**:31–9.

Thom SR. 1993. Leukocytes in carbon monoxide-mediated brain oxidative injury. *Toxicol Appl Pharmacol* **123**:234–47.

Thomsen JL, Kardel T. 1988. Intoxication at home due to carbon monoxide production from gas water heaters. *Forensic Sci Int* **36**:69–72.

Uemura K, Harada K, Sadamitsu D, *et al.* 2001. Apoptotic and necrotic brain lesions in a fatal case of carbon monoxide poisoning. *Forensic Sci Int* **116**:213–19.

Walker E, Hay A. 1999. Carbon monoxide poisoning. *Br Med J* **319**:1082–3.

CHAPTER 30

Agrochemical poisoning

- Paraquat poisoning 566
- Organophosphorus pesticides 567
- References and further reading 568

In the last half century there has been a technological revolution in farming methods, which has spread to almost all parts of the world. The developing nations (who in terms of area of land usage and proportion of the population who work on that land, are numerically greatest at risk) suffer most and this is sometimes exacerbated by the less stringent controls and less safe working methods that exist in those countries.

The potent chemicals used in agriculture may harm persons by accidental exposure, either during their application to crops, or due to incorrect or careless storage. Another major source of human poisoning is through self-administration, when the easily available substances are used for suicide. For example, in Sri Lanka many thousands of hospital admissions each year are for agrochemical poisoning (16 649 in 1983) with over a thousand deaths annually (1521 in 1983). Of these, about three-quarters were self-administered, the remainder being accidental and occupational.

The main dangers from agricultural chemicals lie in the pesticides, especially organophosphorus compounds, such as parathion and herbicides, such as paraquat.

PARAQUAT POISONING

Paraquat is a herbicide that is sprayed on unwanted weeds and other vegetation before planting crops. It is absorbed by the foliage and rapidly kills the plant, but is inactivated when in contact with the soil, so cannot harm the seeds or young plants that are placed in the same ground a short time later. Paraquat (and the less popular diquat) are dipyridyl compounds, paraquat chemically being 1,1′-dimethyl-4,4′-bipyridyldiylium. It is produced commercially as a brownish concentrated liquid of the dichloride salt at 10–30 per cent strength, under the trade name Gramoxone and, for horticultural use, as brown granules called Weedol at about 5 per cent concentration.

Toxicity occurs almost exclusively through ingestion, though some cases are reported where inhalation through spraying has led to a fatal outcome. A suicide from intravenous injection of paraquat has been reported by Fernandez *et al.* (1991).

Only about 5 per cent of the substance is absorbed orally. Poisoning usually occurs either by deliberate self-ingestion of the liquid concentrate, or by accidental drinking from unmarked or incorrectly marked bottles. It is said that one swallowed mouthful of Gramoxone will almost always be fatal and certainly 5 ml has caused death. The decanting of the concentrate from the original manufacturer's containers leads to many accidental deaths. Because of the relatively high cost, paraquat is often taken from the large containers and kept in lemonade or other soft drink bottles. Such decanting is illegal in some countries such as the Republic of Ireland, because of the dangers of unlabelled bottles. Unsuspecting people, especially children, then take a drink from such bottles and immediately suffer the effects of the potent toxin.

Paraquat in concentrated form, is irritant to all epithelial tissues. The lips, mouth, pharynx and oesophagus are superficially eroded, though there is nothing like the dangerous structural damage of strong acids or alkalis. The major danger of paraquat is to the lungs and liver. The lungs are affected mainly by direct aspiration, either as the irritative substance is swallowed or by regurgitation from the stomach during vomiting.

Of that absorbed through the gastrointestinal tract, the major danger is to the liver, where paraquat causes a centrilobular necrosis, with giant mitochondria and crystalline inclusion bodies seen on electron microscopy. In the kidney, renal failure may develop within 2 or 3 days from diffuse tubular damage.

The striking lesion, almost characteristic of paraquat, is in the proliferative damage to the lungs. If rapid death from acute hepatorenal failure does not occur, either because of low dosage or energetic clearance of the residual poison, then progressive lung damage may still lead to death within the next 2 weeks. When paraquat reaches the distal air spaces, diffuse pulmonary oedema and haemorrhages occur, followed by stimulation of the alveolar lining cells. Prior to this, in the first day or two, there is damage to the pneumocytes with vacuolation, desquamation and necrosis. A hyaline membrane is often visible. The mechanism appears to be that paraquat reacts with tissue elements to produce peroxides, including hydrogen peroxide, which is responsible for the damage, which can be worsened by oxygen therapy.

Within a few days repair begins, the mesenchymal interstitial cells and the alveolar lining cells dividing rapidly and filling the alveoli. Both granular and membranous pneumocytes (types I and II) are involved. Within the first week, gas exchange begins to suffer as the air spaces become occluded by mononuclear cells forming rounded-up ‘fibroblasts’. If survival continues, the alveoli begin to fibrose, with reticulin and collagen being laid down to form a rigid, stiff lung.

The autopsy in paraquat poisoning

There may be ulceration around the lips and mouth from escape of paraquat concentrate. The mucosa of the mouth may be reddened or desquamated, and the oesophagus may show worse changes, including casts of shed epithelium. None of these changes is by any means inevitable, however, and the upper gastrointestinal tract may be normal.

Similarly, the stomach may show erosion and patchy haemorrhages, or may be unremarkable. The liver may show pallor or mottled fatty change to the naked eye. It is unusual for any gross charges to be visible. Other organs show no specific changes, apart from the lungs; the kidneys may reveal cortical pallor if there is renal failure.

If the victim has lived a week or more, then the lungs may reveal typical changes in that they are large and stiff, keeping their shape when removed from the chest. There may be a fibrinous pleurisy and sometimes slight bloody pleural effusions. The cut surface reveals oedema and patchy haemorrhage, though this may have subsided in cases that survive for some time. The main appearances are microscopic and, unless the history is known, the lung may be mistaken for a diffuse pneumonia. The author (BK) saw a homicide by paraquat, in which the victim was in hospital for 2 weeks with a presumptive diagnosis of ‘virus pneumonia’, until a confession revealed the true cause of the lung changes. Samples required at autopsy, apart from full histology, include the usual blood samples, urine, stomach contents, lung and liver.

In most instances, the relative slowness of the death allows the diagnosis to be made on history and clinical grounds, so that ante-mortem toxicology is usually well documented.

Toxicology interpretation

Blood levels of paraquat of 0.2 mg/l are sufficient to cause lung damage. Survival has been recorded, however, with plasma levels of up to 1.6 mg/l. Paraquat is excreted over a long period and can be detected in urine at autopsy many days after ingestion. Concentrations in excess of 0.07 mg/l have been found 26 days later.

The concentrations found on analysis after autopsy naturally depend upon the dose taken and the time that has elapsed since ingestion. On the first day, ranges from 0 to 63 mg/l (average 15) have been found in the blood, with urine levels of 20–1210 mg/l (average 462). Seven days later, the blood level comes down to an average of 0.8 mg/l with 4.5 in the urine. Two to three weeks after ingestion, the blood concentration averages 0.5 and the urine level 0.6 mg/l.

ORGANOPHOSPHORUS PESTICIDES

Used in huge quantities throughout the world, these substances cause thousands of deaths in southern Asia, Africa and elsewhere. The major example is parathion, with malathion and dichlorvos used far less extensively. These chemicals act on insects and other arthropods by inhibiting cholinesterase and their toxic effects on man are caused by the same mechanism.

Parathion (nitrostigmine) is extremely toxic and can be absorbed through the skin, conjunctivae, lungs and gut. Because of the inhibitory effect on plasma and erythrocyte cholinesterases, acetylcholine builds up at neuromuscular junctions and other neurotransmitter sites, resulting in hyperexcitability of both voluntary and involuntary muscle.

Other organophosphorus compounds include malathion, which is much less toxic than parathion and is used more in horticulture: dithion, diazinon, hexaethyltetraphosphate (HETP), tetraethylpyrophosphate (TEPP), octamethylpyrophosphamide (OMPA) and demeton are other pesticides of this group.

Toxic signs and symptoms appear when the cholinesterase level drops to 30 per cent of its normal activity. Fatalities begin to occur after the ingestion of 125–175 mg, though ingestion of much larger amounts can be survived. Death can occur in less than an hour after ingestion, though usually several hours elapse in those who are not going to survive.

Autopsy appearances are more helpful than with many other compounds, in that parathion is dissolved in a kerosene base, which can often be detected by its smell. A greenish colouring agent is often added to the commercial products. The stomach therefore may be seen to contain an oily, greenish scum. The gastric mucosa may be haemorrhagic, though this is too non-specific to be of much use, as is the common finding of haemorrhagic pulmonary oedema.

The author (BK), when practising pathology in South-east Asia, noted a useful diagnostic sign in some less modern mortuaries that were infested with flies and bluebottles. When these insects alighted on an opened cadaver at autopsy, many fell dead or dying onto the autopsy table when the case in progress was a parathion poisoning.

Toxicology interpretation

Post-mortem toxicology is usually conclusive in acute poisoning. The substance resists post-mortem autolysis well and can be recovered from putrefied bodies. The range of blood parathion concentrations in fatal cases may be from 0.5 to 34 mg/l with an average of 9.0. Urinary excretion ranges from 0.4 to 78 mg/l with an average of 10 mg/l. The liver may contain from 0.1 to 120 mg/kg, averaging 11.

Malathion is much less toxic but deaths certainly occur given a greater dose. Blood levels in fatalities ingesting from 25 to 70 g malathion ranged from 100 to 1880 mg/l with a mean of 815. The liver contained from 200 to 1700 mg/kg, with an average of 1300.

Other pesticides and insecticides include chlorinated compounds such as aldrin, DDI (dicophane), lindane (Gammaxane) and dieldrin. Other types include the carbamates (opxymyl, landrin and aldicarb), dinitrophenols, pyrethrum and derris. The phenolic pesticides include DNOC (dinitro-orthocresylphosphate) and dinitrobutylphenol. None of these has any specific autopsy features and the diagnosis rests upon the history, circumstances, and post-mortem analysis of blood, urine and liver.

REFERENCES AND FURTHER READING

Baselt RC. 2002. *Disposition of toxic drugs and chemicals in man*, 6th edn. Biomedical Publications, Foster City, CA.

Beebeejaun AR, Beevers G, Rogers WN. 1971. Paraquat poisoning-prolonged excretion. *Clin Toxicol* **4**:397–407.

Carson DJ, Carson ED. 1976. The increasing use of paraquat as a suicidal agent. *Forensic Sci* **7**:151–60.

Copeland AR. 1988. Organophosphate related fatalities – a violitional biohazard? *Forensic Sci Int* **39**:155–62.

Cravey RH. 1979. Poisoning by paraquat. *Clin Toxicol* **14**:195–8.

Crowley WJ, Jr, Johns TR. 1966. Accidental malathion poisoning. *Arch Neurol* **14**:611–16.

Davies DS. 1987. Paraquat poisoning: The rationale for current treatment regimes. *Hum Toxicol* **6**:37–40.

Davies DS, Hawksworth GM, Bennett PN. 1977. Paraquat poisoning. *Proc Eur Soc Tox* **18**:21–6.

De Alwis LB, Salgado MS. 1988. Agrochemical poisoning in Sri Lanka. *Forensic Sci Int* **36**:81–9.

Dearden LC, Fairshter RD, McRae DM, *et al.* 1978. Pulmonary ultrastructure of the late aspects of human paraquat poisoning. *Am J Pathol* **93**:667–80.

Fernandez P, Bermejo AM, Lopez Rivadulla M, *et al.* 1991. A fatal case of parenteral paraquat poisoning. *Forensic Sci Int* **49**:215–24.

Fernando R. 1989. *Pesticides in Sri Lanka.* Friedrich-Ebert Stifting, Colombo.

Gage JC. 1960. The detection of organic phosphorus compounds. *Med Sci Law* **1**:137–44.

Grant H, Lantos PL, Parkinson C. 1980. Cerebral damage in paraquat poisoning. *Histopathology* **4**:185–95.

Harsanyi L, Nemeth A, Lang A. 1987. Paraquat (gramoxone) poisoning in south-west Hungary, 1977–1984. Toxicological and histopathological aspects of group intoxication cases. *Am J Forensic Med Pathol* **8**:131–4.

Karalliedde L, Senanayake N. 1988. Acute organophosphorus insecticide poisoning in Sri Lanka. *Forensic Sci Int* **36**:97–100.

Klein-Schwartz W, Smith GS. 1997. Agricultural and horticultural chemical poisonings: mortality and morbidity in the United States. *Ann Emerg Med* **29**:232–8.

Kuo TL, Kuo CY. 1988. Determination of paraquat from formalin-fixed tissue. *Forensic Sci Int* **38**:243–9.

Neoral L, Dusek J, Smysl B. 1977. [A contribution to the pathogenesis of lethal paraquat poisoning (author's transl)]. *Z Rechtsmed* **80**:1–7.

Parkinson C. 1980. The changing pattern of paraquat poisoning in man. *Histopathology* **4**:171–83.

Peiris JB, Fernando R, De Abrew K. 1988. Respiratory failure from severe organophosphate toxicity due to absorption through the skin. *Forensic Sci Int* **36**:251–3.

Petty CS. 1958. Organic phosphate insecticide poisoning. *Am J Med* **23**:467–70.

Przybylska A. 1999. [Intoxications caused by plant protection chemicals in 1997.] *Przegl Epidemiol* **53**:121–8.

Rebello G, Mason JK. 1978. Pulmonary histological appearances in fatal paraquat poisoning. *Histopathology* **2**:53–66.

Reif RM, Lewinsohn G. 1983. Paraquat myocarditis and adrenal cortical necrosis. *J Forensic Sci* **28**:505–9.

Russell LA, Stone BE, Rooney PA. 1981. Paraquat poisoning: toxicologic and pathologic findings in three fatal cases. *Clin Toxicol* **18**:915–28.

Stewart MJ, Moar JJ, Mwesigwa J, *et al.* 2000. Forensic toxicology in urban South Africa. *J Toxicol Clin Toxicol* **38**:415–19.

Teare RD. 1976. Poisoning by paraquat. *Med Sci Law* **16**:9–12.

Tompsett SL. 1970. Paraquat poisoning. *Acta Pharmacol Toxicol (Copenh)* **28**:346–58.

Vercruysse A. 1964. Acute parathion poisoning; forensic problems. *J Forensic Med* **11**:107–19.

Wesseling C, van Wendel de Joode B, Monge P. 2001. Pesticide-related illness and injuries among banana workers in Costa Rica: a comparison between 1993 and 1996. *Int J Occup Environ Health* **7**:90–7.

Zivot U, Castorena JL, Garriott JC. 1993. A case of fatal ingestion of malathion. *Am J Forensic Med Pathol* **14**:51–3.

CHAPTER 31

Poisoning by medicines

- General considerations 570
- Autopsy appearances 570
- Results of laboratory tests 571
- Analgesics 571
- Paracetamol 572
- Antidepressant drugs 573
- The benzodiazepines 573
- The phenothiazines 573
- The barbiturates 573
- Insulin poisoning 574
- References and further reading 575

GENERAL CONSIDERATIONS

In most countries, especially those with advanced economies and relatively sophisticated medical services, poisoning with medicinal compounds is common. In many such countries, poisoning with therapeutic substances exceeds deaths from other types of toxic agents, especially in suicidal and accidental poisoning.

This is explained by the ease of access of such substances, whether they are obtainable from a doctor on prescription or on demand across the counter of a pharmacy. Where state-sponsored health services exist, the cost to the recipient may be minimal or absent and this ease of access contributes to opportunities for self-poisoning, whether it be deliberate self-destruction, suicidal gestures or accidental ingestion – the latter especially in children. Unfortunately, overprescribing or the supply of too great a quantity of drugs at one time allows excessive stocks of drugs to be easily available to the public.

Though only a minority of victims of medicinal poisoning fail to recover, there are still an appreciable number of deaths. These come to the attention of the investigative authorities and hence to pathologists. The autopsy investigation of a fatality from a therapeutic substance can be difficult, for a number of reasons:

- The nature of the substance may be uncertain or unknown.
- There may be more than one such substance involved.
- There may be a delay between ingestion and death sufficient to allow blood, urine and tissue concentrations to decline below fatal, toxic or even therapeutic levels.
- Analysis may be difficult to arrange because of lack of facilities.
- Information about fatal levels may be unobtainable.
- Most medicinal poisons leave virtually no characteristic features at autopsy, so diagnosis depends upon laboratory findings.
- Post-mortem changes may make analysis difficult, inaccurate or impossible.
- Where death is delayed after taking the substance, none may be recoverable from the stomach (which has emptied) or even from the intestine.
- The original substance may be rapidly metabolized into one or more breakdown products, adding to difficulties in identification and interpretation.

AUTOPSY APPEARANCES

The lack of characteristic autopsy appearances is often very frustrating for the pathologist. Unless there is an indicative or suggestive history as to which drug was taken – a matter for the investigators in respect of circumstances and recovery of containers – then an autopsy may have to be performed ‘blind’. Where no significant morphological lesions can be discovered, then a full toxicology screen must be considered, which in some jurisdictions may be difficult or impossible to obtain, or be extremely expensive.

The majority of modern medicinal substances are, by design, bland and non-irritant to the tissues and gastrointestinal tract. Most of those met with in forensic practice are taken orally and, though the active constituents may be potent in their pharmacological effect on target organs and tissues, the medicine will cause no erosion or damage to the alimentary tract. Thus little or no physical evidence can be obtained from a gross or even microscopic examination of the gastrointestinal tract or other organs. Much of the physical bulk of modern tablets or capsules is merely the vehicle for introducing the active component into the body and is thus unlikely to have any adverse effect.

When a medicinal compound causes death, the mode of death is most often some form of cardiorespiratory failure, often secondary to depressive effects on the central nervous system. This mode of death causes only non-specific changes discernible at autopsy, which are usually of no use in indicating the basic reason for the death. Acute congestive cardiac failure, pulmonary oedema, sometimes cerebral oedema, generalized organ congestion, scattered petechiae on serous membranes – none of these is of any real use to the pathologist, who has to rely on the results of toxicological analysis for a definitive answer.

There are some therapeutic substances, which, though not the cause of specific lesions, may suggest themselves from their autopsy appearances. An example is the widespread membrane ecchymoses sometimes seen in aspirin poisoning. This could not in any way be sufficient to provide a legally acceptable cause of death, however, unless absolutely reliable circumstantial evidence existed, together with the finding of a large bolus of undissolved tablet remnants in the stomach. Even then it might be contested that the substance was not aspirin – or not aspirin alone – unless analytical confirmation was obtained.

RESULTS OF LABORATORY TESTS

As discussed in Chapter 27, the investigation of fatal poisoning is a collaborative investigation between pathologist and toxicologist. The autopsy excludes, confirms or evaluates any trauma or natural disease, and provides suitable material for analysis. The toxicology laboratory conducts the technical assays, and produces qualitative and quantitative results. The toxicologist/analyst interprets those results to the pathologist, by providing an indication of the therapeutic, toxic and fatal ranges of concentrations in various body fluids and tissues, and by pointing out problems such as decline from post-ingestion survival, conversion to metabolites and many others. The pathologist then collates this information with his own knowledge of the history and autopsy findings to offer the best interpretation of the investigation for judicial authorities. Most problems arise either because the information about the medicine (especially if it is newly developed or where its toxicity is low) is incomplete in terms of toxic blood and tissue levels – because post-ingestion survival has allowed the originally lethal levels to have subsided to therapeutic or even lower limits. What is offered in the remainder of this chapter is a digest of information about such potentially lethal levels, culled from a variety of sources. The ranges are often wide as most of the data are of necessity derived anecdotally, and the problems of uncertainty of dosage, variation in post-ingestion survivals and wide individual biological variation make it impossible to lay down strict thresholds between therapeutic, toxic and fatal concentrations.

Wherever possible, the advice of the analytical toxicologist should be taken about each case – but where this is impracticable, then the following data and similar material, which is constantly being updated in forensic and toxicological publications, may be of assistance. The choice of substances is arbitrary, but represents the most common medicines seen in suicidal and accidental poisoning.

ANALGESICS

Aspirin (acetylsalicylic acid) and salicylates

Aspirin is the most widely used therapeutic drug, being analgesic, antipyretic and anti-inflammatory. It was formerly very common as an agent of self-poisoning, both accidental in children and suicidal in adults. In Britain in the last two decades, its use as a self-poisoning agent has declined remarkably, so that fatalities are now rarely seen.

The therapeutic dose is usually 325–975 mg, that is, 1–3 tablets. Rarely, persons with an aspirin hypersensitivity may become ill or even die after therapeutic doses, suffering urticaria, angioneurotic oedema, hypotension, vasomotor disturbances and laryngeal and glottal oedema.

Patients on long-term salicylate therapy for arthritic or rheumatic diseases may take 3–5 g/day and slowly reach blood concentrations which would be in the lethal range if caused by acute overdoses. Those on 3 g/day have blood levels varying between 44 and 330 mg/l.

Apart from deaths caused by hypersensitivity, death in an adult is unlikely with the ingestion of fewer than about 50 tablets, that is, about 16 g. The blood concentration (measured as total salicylate), from a medicinal dose of 975 mg, ranges from about 30 to 100 mg/l (with a mean of 77) 2 hours after ingestion. There is a rapid fall to around 25 mg/l some 8 hours later.

At autopsy, aspirin is one of the few medicines that may cause some gross abnormalities, though they are not particularly specific. Externally there is nothing to see, unless vomiting has taken place, when dark or even red, bloody gastric contents may be expelled. Rarely, there may be some haemorrhagic manifestations in the form of skin petechiae.

Internally, the stomach may still contain a mass of fused, unabsorbed tablets. These tend to start dissolving and then aggregate into a grey or dirty-white mass if a large number (several hundred) have been swallowed. The gastric mucosa may be eroded by the irritative, acidic substance. This may be localized in the stomach or spread widely across the fundus and cardia. The lining may be dotted with acute erosions, sometimes causing bleeding that may amount to a frank haematemesis. Black altered blood may lie in the stomach and pass into the intestine to form a melaena if survival is long enough. Mucosal petechiae and ecchymoses in the stomach, without actual erosive destruction, may be seen as part of a haemorrhagic result of the anticoagulant action of aspirin. Similar petechiae may be spread through other organs and especially serous membranes, particularly on the parietal pleura and epicardium. Even single tablets may cause a small ulcer if they stick to the stomach lining – this is sometimes seen incidentally at autopsy in non-poisoning cases, where a tablet has been swallowed shortly before death.

Post-mortem toxicology requires the usual samples of blood, urine, stomach contents and liver. The mass of aspirin in the stomach may remain for several days in life, forming a partly insoluble concretion that may retard the absorption of the drug. This is why it is always worth washing out the stomach of a live victim, as a large proportion of the aspirin may be removed before it can cause systemic effects. With the advent of soluble aspirin or effervescent preparations, this aspect is lost as no such insoluble bolus forms. At autopsy, part of such a mass can be sent for analysis, whilst quick ‘spot’ tests can be carried out on another part. Such quick chemical confirmation may be performed in the autopsy room itself, by using a 10 per cent solution of ferric chloride. If a small quantity is added to a urine sample or to the surface of the tablet mass, an immediate purple–blue colour suggests aspirin. This is by no means specific, but is merely suggestive. If negative, however, then aspirin can be virtually discounted. These are only rapid screening methods and by no means replace proper laboratory analysis.

Toxic blood levels (measured as total salicylate) begin at about 300–500 mg/l, though both death and survival are consistent with far higher or lower levels. Blood concentrations in fatal cases may range from about 60 to 7300 mg/l, some authorities suggesting that 500 mg/l is an average minimum level. The liver concentration in fatalities varies from 2.5 to 1000 mg/kg and urine salicylate from 20 to 1350 mg/l, emphasizing the wide levels that are compatible with life. Salicylate has a rather slow clearance rate from the blood, the half-life being up to a day in some massive overdoses.

Aspirin poisoning is dangerous in that sudden cardiac arrest can occur in the absence of any toxic symptoms. This accounts for the deaths that take place after patients have been discharged fit from the emergency departments of hospitals. They may have seemed quite well and symptom free, but suffer a fatal collapse up to a day or so later. Fatal cardiac arrhythmias can supervene without warning and make it advisable, where possible, to admit patients with aspirin overdoses for observation for a day or so.

PARACETAMOL

Paracetamol is also known as acetominophen, *N*-acetyl-*p*-aminophenol or 4′-hydroxyacetanilide. It is an analgesic and antipyretic, without the anti-inflammatory properties of aspirin, for which it is often used as an alternative because of the lack of gastric irritation.

In wide use, especially in combination with other drugs such as codeine and dextropropoxyphene, paracetamol is one of the most common agents in self-induced poisoning by medicinal products. It is used alone in therapeutic doses of up to 500 mg. Overdoses of 20 g or more are potentially lethal, but much less is needed in combination with other drugs, such as propoxyphene. Paracetamol is a potent liver poison, as a small proportion is converted by the liver enzyme ‘P450’ (microsomal mixed function oxidase) into a toxic compound, probably *N*-acetyl-*p*-benzoquinone. Normally, glutathione and other sulphydryl compounds detoxify this substance, but in overdose, these are exhausted and the toxic agent causes a profound centrilobular hepatic necrosis. The concurrent administration of other drugs, such as phenobarbitone or phenytoin in epileptics or chronic alcoholism, activate the P450 enzyme and worsen the toxicity through this mechanism.

At autopsy there is nothing specific to observe in the gastrointestinal tract. In massive overdosage, there can rarely be rapid death from a direct depressive action on the central nervous system, but most deaths are delayed 2–4 days whilst liver failure develops. At autopsy, the liver may be enlarged, but is often under the normal weight of 1500 g. It may be pale yellow or tan, or the damage may be only visible histologically, when a centrilobular necrosis is seen. Renal changes in the form of tubular necrosis may also be present and occasionally myocardial fibre damage may be visible histologically.

Analytically, the therapeutic range of plasma concentrations 6 hours after a 324 mg dose is 2–6 mg/l, though some

TABLE 31.1 *Fatal concentrations of some antidepressants (mg/l or mg/kg)*

	Blood (average)	Urine (average)	Liver (average)
Amitriptyline	2.7–4.7 (3.7)	0.4–7.9 (3.4)	13–317 (130)
Dothiepin	0.3–2.5 (1.5)	0.4–5.1 (3.2)	2.0–14 (8)
Imipramine	6–8.5 (7.3)	0.6–54 (20)	33–381 (166)
Tranylcypromine	3.7	25.0	7.3

records claim peak levels of up to 25 mg/l. The plasma half-life is a guide to hepatotoxicity, it being dangerous to have a half-life of more than 2 hours at a level of 300 mg/l at 4 hours after ingestion. Typical blood levels in overdoses when at least 10–15 g have been taken are 100–400 mg/l, with an average around 250. The urine may contain 150–800 mg/l, but all levels depend on dose and survival time. Paracetamol may exhibit post-mortem redistribution.

The inclusion of other drugs, especially dextropropoxyphene (and, of course, alcohol), may markedly reduce the level needed for a fatal outcome. Paracetamol was the most frequently detected substance in a compilation of post-mortem femoral blood concentrations of drugs (Druid and Holmgren 1997) based on a selection of 15 800 samples sent to the Department of Forensic Chemistry in Linköping, Sweden, during 1992–95. In their series of 139 fatal intoxications with paracetamol in combination with other drugs and/or alcohol, the median concentration of paracetamol was 170 mg/l (range 90–320).

ANTIDEPRESSANT DRUGS

The tricyclic antidepressants are frequently involved in self-poisoning, partly associated with the type of patient for whom they are prescribed.

Amitriptyline, dothiepin, doxepin and trimipramine have additional sedative properties. Those with little or no sedating action include protriptyline, nortriptyline, imipramine, clomipramine, iprindole, lofepramine, desipramine and butriptyline.

Tetracyclic antidepressants include maprotiline and mianserin. Other types include the monoamine oxidase inhibitors, which are well known to have dangers related to the concurrent ingestion of other drugs and foods, especially those with sympathomimetic action and tyramine content, such as rich cheese, yeast extracts, red wine and beans. Dangerous hypertension may ensue with the risk of cerebrovascular haemorrhage. Drugs of this class include phenoxypropazine, tranylcypromine, isocarboxazid and phenelzine.

TABLE 31.2 *Fatal concentrations of three benzodiazepines (mg/l or mg/kg)*

	Blood	Urine	Liver
Chlordiazepoxide	20	8	10
Diazepam	5–19	3	16
Nitrazepam	1.2–9.0	1–10	0.7–4.0

TABLE 31.3 *Fatal concentrations of two phenothiazines (mg/l or mg/kg)*

	Blood (average)	Urine (average)	Liver (average)
Chlormethiazole	10–214 (55)	5–114 (43)	42–190 (94)
Chlorpromazine	6.6	1.2	84

THE BENZODIAZEPINES

These widely used drugs are employed for their sedative and tranquillizing effects. A large number of both 1,4- and 1,5-benzodiazepines are available, divided into short-acting, intermediate-acting and long-acting compounds.

Long-acting benzodiazepines include: flurazepam, nitrazepam, diazepam, ketazalam, chlordiazepoxide, clobazam, chlorazepate, medazepam and alprazolam. Intermediate-acting benzodiazapines include: loprazolam, lormetazepam, temazepam, flunitrazepam, lorazepam, bromazepam and oxazepam. A short-acting benzodiazepine is triazolam.

THE PHENOTHIAZINES

This group of tranquillizer drugs includes: haloperidol (butyrophenone), chlormethiazole, chlorpromazine, fluphenazine, diphenylbutylpiperidine, promazine, trifluoperazine and prochlorperazine.

The autopsy appearances are non-specific and toxicology may resolve any diagnostic problems if the death has occurred fairly soon after ingestion – which may often not be the case, when history and ante-mortem investigations can provide the only answer.

THE BARBITURATES

The massive problem posed by therapeutic administration of barbiturates until about 20 years ago has largely abated in countries with a responsible medical profession, which voluntarily refrained from the prescription of these drugs except where specifically indicated. Their use as sleeping tablets and general soporific sedative agents led to widespread abuse, so that at one time they were easily the most common agent of drug addiction. The development of non-barbiturate hypnotics, such as the benzodiazepines, helped to remove the need for the older and more lethal compounds. Unfortunately, barbiturates are still widely available on the illicit market, either alone or in combination with other substances such as amphetamines.

Barbiturates exist in many forms, the best classification (which relates to their degree of toxicity) being their speed of action:

- Long-acting barbiturates: barbitone, phenobarbitone and phenytoin, which is still prescribed for epilepsy.
- Intermediate-acting barbiturates: amylobarbitone, sodium amytal, pentobarbitone, allobarbitone, butobarbitone and pentobarbitone.
- Short-acting barbiturates: hexobarbitone, cyclobarbitone, secobarbital and thiopentone.

Much lower blood levels will be found in fatal poisonings in the short-acting group as death may occur more quickly from the usual mode of action, a central depression of the respiratory centres. The author (BK) has knowledge of a death within 20 minutes of taking a massive overdose of 'Seconal'.

At autopsy, the signs are of general cardiorespiratory failure, with often a cyanotic, congestive appearance. Though non-specific, probably the congested lungs in acute barbiturate poisoning are more intense than in any other condition. These organs may be almost black and the whole venous system is engorged with dark, deoxygenated blood. There may be 'barbiturate blisters' on dependent parts of the skin surface, especially buttocks, backs of thighs, calves and forearms, though as discussed in the chapter on carbon monoxide poisoning, these blisters are common to all states of deep coma.

Internally there may be local signs of erosion from the drug itself. The gastric mucosa may be badly damaged from the alkaline attack of drugs such as sodium amytal which, being the sodium salt of a weak organic acid, hydrolyses in the stomach. The fundus may be thickened, granular and haemorrhagic. The cardia and lower oesophagus may be eroded from reflux and, if the victim regurgitates, then black, altered blood may appear at the nose and mouth.

The capsules of certain barbiturates also leave characteristic traces in the mouth, oesophagus and stomach. The colour varies with the manufacturer, but the turquoise–blue of sodium amytal capsules may stain the stomach contents and even be visible through the wall of the intestine when the abdomen is opened. Other pigmented gelatine capsules can be red, yellow or blue. As with so many other drugs, combination with alcohol greatly increases the dangers of a fatality.

TABLE 31.4 *Fatal concentrations of barbiturates (mg/l or mg/kg)*

	Blood	Urine	Liver
Phenobarbitone	55–144	–	–
Amylobarbitone	29–68	210	106–580
Secobarbitone (or quinal barbitone)	5–52	–	15–330

INSULIN POISONING

Formerly a rarity, death from parenteral administration of insulin is now not uncommon. The Beverley Allitt case in Britain in recent years showed that multiple deaths can occur, especially where medical or nursing staff are concerned, as well as those in proximity to diabetics, as both groups may have access to insulin.

Another change is the relative ease with which analysis of body fluids and tissue can be carried out now, compared with the great technical difficulties of only a few years ago, so that fewer such deaths may be missed from lack of reliable insulin assays.

Fatal insulin toxicity may be accidental, suicidal or homicidal. The accidental fatalities are usually examples of medical error, mostly from misreading the label on the box or ampoule. A former student of the author (BK), on the fourth day of her first intern post, gave ten times the dose of insulin to a patient during a pituitary function test, by wrongly assuming that the stated number of units written on the box were the total ampoule contents, instead of being the strength per millilitre.

Suicide by insulin is not uncommon; the author (BK) has seen it given into a saline drip by a doctor and also injected into the abdominal wall by a non-diabetic who stole it from her diabetic neighbour's refrigerator.

As stated, homicide and attempted homicide have given rise to some notorious cases, both in UK and the USA. In one case known to the author (PS), a previously healthy 48-year-old man was delivered unconscious to the emergency unit because of a suspicion of decompression sickness. The treatment was aborted as the patient was found to be hypoglycaemic (nadir serum glucose 0.3 mmol/l) and treatment and diagnostics of hypoglycaemia commenced. Serum samples drawn at admittance were stored frozen, whereby it was possible to show retrospectively that, while the concentration of insulin in serum was high (75 mU/l, increasing further to over 240 mU/l in the next few hours), concentration of C-peptide was low (below detection limit of 0.1 nmol/l) at the hypoglycaemic stage, suggesting that the patient had received exogenous insulin somehow, and the police were informed. Due to severe hypoglycaemic brain damage, the patient remained in a vegetative state for 2 months before dying of multiorgan failure. Circumstantial evidence obtained during the ensuing criminal investigation was considered by the court to prove the patient's wife (a nurse) guilty of murder (Koskinen *et al.* 1999).

Insulin is, of course, inactive orally and has to be given by injection to perform its hypoglycaemic effect. At autopsy, where either from the circumstances or the finding of needle marks, insulin is a possibility, peripheral blood samples and skin and underlying tissue from the injection site should be carefully preserved, together with control skin from another site.

The fine needles usually used by diabetics may leave virtually no mark on the skin. The author (BK) has tested some such needles on cadaver skin and found that, often, the mark cannot be seen immediately after withdrawal, unless a small vessel has been damaged.

Although insulin has been recovered many days, even weeks, after death, the sooner the better as far as collecting samples is concerned. Serum should be separated from red cells and the former frozen until sent to the analysts, unless whole blood can be sent straight away. Skin and tissue samples should either be frozen or kept in the refrigerator – not fixed. Porcine or bovine insulin can be detected as such but, if therapeutic insulin is of human origin, then it cannot be distinguished on analysis from the patient's own insulin. As well as immunoassay of the insulin itself, the measurement of C-peptide, produced on a one-to-one basis by the pancreas, assists in distinguishing endogenous from exogenous insulin. All such interpretations are a matter for specialists in this field, upon whose advice the pathologist must rely.

Attempting to prove insulin-induced hypoglycaemia by measuring glucose levels in human post-mortem fluids is impracticable, due to the unreliability of such estimations after death.

Very low vitreous humour glucose levels may strongly suggest hypoglycaemia, but are not absolutely acceptable.

REFERENCES AND FURTHER READING

Baselt RC. 2002. *Disposition of toxic drugs and chemicals in man*, 6th edn. Biomedical Publications, Foster City, CA.

Bauman WA, Yalow RS. 1981. Insulin as a lethal weapon. *J Forensic Sci* **26**:594–8.

Beastall GH, Gibson IH, Martin J. 1995. Successful suicide by insulin injection in a non-diabetic. *Med Sci Law* **35**:79–85.

Brodsgaard I, Hansen AC, Vesterby A. 1995. Two cases of lethal nitrazepam poisoning. *Am J Forensic Med Pathol* **16**:151–3.

Cardauns H, Iffland R. 1973. [Fatal intoxication of a young drug addict with diazepam (author's transl).] *Arch Toxikol* **31**:147–51.

Chaturvedi AK, Hidding JT, Rao NG, *et al.* 1987. Two tricyclic antidepressant poisonings: levels of amitriptyline, nortriptyline and desipramine in post-mortem biological samples. *Forensic Sci Int* **33**:93–101.

Crifasi J, Long C. 1997. The GCMS analysis of tranylcypromine (parnate) in a suspected overdose. *Forensic Sci Int* **86**:103–8.

Druid H, Holmgren P. 1997. A compilation of fatal and control concentrations of drugs in postmortem femoral blood. *J Forensic Sci* **42**:79–87.

Engelmann L, Kohler H. 1971. [Clinical aspects of chlordiazepoxide poisoning.] *Dtsch Gesundheitsw* **26**:435–9.

Fatteh A, Blanke R, Mann GT. 1968. Death from imipramine poisoning. *J Forensic Sci* **13**:124–8.

Fermer R. 1995. *Forensic pharmacology*. Oxford University Press, Oxford.

Finkle BS, McCloskey KL, Goodman LS. 1979. Diazepam and drug-associated deaths. A survey in the United States and Canada. *JAMA* **242**:429–34.

Fletcher SM. 1983. Insulin. A forensic primer. *J Forensic Sci Soc* **23**:5–17.

Forbes G, Pollock Weir W, Smith H, *et al.* 1965. Amitriptyline poisoning. *J Forensic Sci Soc* 5:183–7.

Giusti GV, Chiarotti M. 1979. Lethal nitrazepam intoxications, report of two cases. *Z Rechtsmed* **84**:75–8.

Griffiths GJ. 1973. Overdose of parstelin (tranylcypromine). *Med Sci Law* **13**:93–4.

Haibach H, Dix JD, Shah JH. 1987. Homicide by insulin administration. *J Forensic Sci* **32**:208–16.

Iwase H, Kobayashi M, Nakajima M, *et al.* 2001. The ratio of insulin to C-peptide can be used to make a forensic diagnosis of exogenous insulin overdosage. *Forensic Sci Int* **115**:123–7.

Junge M, Tsokos M, Puschel K. 2000. Suicide by insulin injection in combination with beta-blocker application. *Forensic Sci Int* **113**:457–60.

Kanto J, Sellman R, Haataja M, *et al.* 1978. Plasma and urine concentrations of diazepam and its metabolites in children, adults and in diazepam-intoxicated patients. *Int J Clin Pharmacol Biopharm* **16**:258–64.

Kernbach-Wighton G, Puschel K. 1998. On the phenomenology of lethal applications of insulin. *Forensic Sci Int* **93**:61–73.

Koskinen PJ, Nuutinen HM, Laaksonen H, *et al.* 1999. Importance of storing emergency serum samples for uncovering murder with insulin. *Forensic Sci Int* **105**:61–6.

Langford AM, Taylor KK, Pounder DJ. 1998. Drug concentration in selected skeletal muscles. *J Forensic Sci* **43**:22–7.

Ledda F. 1968. [Acute chlordiazepoxide poisoning in a drug addict, followed by withdrawal syndrome.] *Clin Ter* **44**:167–71.

Levy WJ, Gardner D, Moseley J, *et al.* 1985. Unusual problems for the physician in managing a hospital patient who received a malicious insulin overdose. *Neurosurgery* **17**:992–6.

Lutz R, Pedal I, Wetzel C, *et al.* 1997. Insulin injection sites: morphology and immunohistochemistry. *Forensic Sci Int* **90**:93–101.

Mackell MA, Case ME, Poklis A. 1979. Fatal intoxication due to tranylcypromine. *Med Sci Law* **19**:66–8.

Marks V. 1995. Hypoglycaemia – real and unreal, lawful and unlawful: the 1994 Banting lecture. *Diabet Med* **12**:850–64.

Marks V, Teale JD. 1999. Hypoglycemia: factitious and felonious. *Endocrinol Metab Clin North Am* **28**:579–601.

Mirchandani H, Reich LE. 1985. Fatal malignant hyperthermia as a result of ingestion of tranylcypromine (parnate) combined with white wine and cheese. *J Forensic Sci* **30**:217–20.

Noirfalise A, Dodinval P, Quiriny J, *et al.* 1987. Death through injection of barbiturates. *Forensic Sci Int* **35**:141–4.

Patel F. 1992. Fatal self-induced hyperinsulinaemia: a delayed post-mortem analytical detection. *Med Sci Law* **32**:151–9.

Patel F. 1995. Successful suicide by insulin injection in a non-diabetic. *Med Sci Law* **35**:181–2.

Phillips AP, Webb B, Curry AS. 1972. The detection of insulin in postmortem tissues. *J Forensic Sci* **17**:460–3.

Polson C, Green M, Lee M. 1983. *Clinical toxicology*, 3rd edn. Pitman, London.

Pounder DJ, Owen V, Quigley C. 1994. Postmortem changes in blood amitriptyline concentration. *Am J Forensic Med Pathol* **15**:224–30.

Price LM, Poklis A, Johnson DE. 1991. Fatal acetaminophen poisoning with evidence of subendocardial necrosis of the heart. *J Forensic Sci* **36**:930–5.

Schneider V, Dulce HJ. 1979. [Suicidal poisonings with insulin and their demonstration on the corpse (radioimmunoassay).] *Arch Kriminol* **164**:142–52.

Sticht G, Kaferstein H. 1980. [Results of toxicological investigations on vesparax-poisonings (author's transl).] *Z Rechtsmed* **85**:169–75.

Sturner WQ, Putnam RS. 1972. Suicidal insulin poisoning with nine day survival: recovery in bile at autopsy by radioimmunoassay. *J Forensic Sci* **17**:514–21.

Williams KR, Pounder DJ. 1997. Site-to-site variability of drug concentrations in skeletal muscle. *Am J Forensic Med Pathol* **18**:246–50.

Wilson Z, Hubbard S, Pounder DJ. 1993. Drug analysis in fly larvae. *Am J Forensic Med Pathol* **14**:118–20.

Winston DC. 2000. Suicide via insulin overdose in nondiabetics: the New Mexico experience. *Am J Forensic Med Pathol* **21**:237–40.

Yacoub M, Cau G, Faure J, *et al.* 1968. [Medico-legal diagnosis of fatal imipramine poisoning.] *Med Leg Dommage Corp* **1**:211–13.

Yonemitsu K, Pounder DJ. 1993. Postmortem changes in blood tranylcypromine concentration: Competing redistribution and degradation effects. *Forensic Sci Int* **59**:177–84.

CHAPTER 32

Death from narcotic and hallucinogenic drugs

- Morphine and other opioid drugs 578
- Toxicological results 579
- Methadone 580
- Hallucinogens 580
- Cannabis 581
- The amphetamines 581
- Cocaine 581
- References and further reading 582

These substances are discussed solely from the point of view of the pathologist dealing with fatal cases.

Drugs of dependence may be absorbed orally, by intravenous, subcutaneous or – rarely – intramuscular injection, by smoking, or by nasal sniffing. The routine at autopsy, in respect of obtaining samples for toxicological analysis, is altered according to the route of administration. As mixing of drugs and addition of non-narcotic drugs is common, it is the usual practice to take a wide range of samples even if the primary route is known with some degree of certainty. For example, an addict dying 'on the needle' where intravenous injection is obvious, will still have stomach contents taken for investigation. The standard samples should be taken, as described in a previous chapter, comprising several samples of venous blood (one with fluoride), stomach and contents, liver and urine. In some circumstances, additional samples such as bile, cerebrospinal fluid and vitreous humour may be taken, as well as brain or kidney. The great advances in the analytical techniques allow the analysis of drugs also in other biological samples, such as saliva, sweat and hair. Hair analysis can also provide evidence of long-term exposure to drugs (weeks, months or years), because most drugs, if not all, incorporate in hair and are relatively stable. At least 50 mg of hair should be collected, cutting about a pencil thickness of strands of hair as close to the skin as possible from the back of the head, dried and stored in a sealed plastic bag or tube at room temperature.

When the drug has been injected, then an ellipse of skin around the injection mark, extending down through the subcutaneous tissue to the muscle, should be excised, along with a control area of skin from another non-injected site. These should be refrigerated, not fixed in formalin, until delivery to the laboratory can be arranged. Full histology should always be taken, especially if drugs have been injected, as foreign substances may be discovered as embolic particles, especially in the lungs. Pulmonary granulomata are well-known histological features of 'mainlining' addicts taking impure drugs intravenously, as the lung capillaries filter out coarse particulate matter used to dilute the active narcotic. Talc is particularly prone to form granulomata, sometimes with foreign body giant cells. Under polarized light, doubly refractile particles may be seen in the centre of the reactive nodules. Sometimes, strands of cotton may form foreign bodies, derived from the cloth strainer used to filter particles crudely from the drug solution before injection.

Siderophages were also claimed to be increased in these lungs, compared with healthy young adult controls. However, Lockemann and Püschel (1993), studying lungs stained for iron with Perl's reaction, found conflicting results in two large series and conclude that the pathognomonic value of these cells in drug takers is uncertain.

If drugs may have been taken by the nasal route, such as cocaine and heroin, then dry swabs from each nostril should be taken.

MORPHINE AND OTHER OPIOID DRUGS

Morphine is the major representative of the general group of opioids, which comprise natural opium and a whole series of chemically related derivatives. They may be taken orally or injected and several, such as crude opium and heroin – may be absorbed by inhaling smoke. Morphine itself is poorly absorbed from the gastrointestinal tract; heroin can be taken via the nasal mucosa.

The group consists of opium, morphine, heroin (diacetyl morphine), codeine (dimethyl morphine), dihydrocodeine (DF 118), etorphine (Immobilon), methadone, papaverine, pethidine, dipipanone, dextromoramide, dextropropoxyphene, pentazocine, cyclazocine, diphenoxylate, buprenorphine, tramadol, fentanyl and many more.

Autopsy appearances

The autopsy findings in deaths from all these drugs are relatively non-specific. Toxicological analysis and expert interpretation of the results are necessary for the proper elucidation of the deaths, but certain features can be useful pointers. The first is the presence of injection marks. When fresh, they look just like any other needle mark commonly seen from therapeutic or diagnostic procedures. They are commonly on the arms, either in the classical position in the antecubital fossa on the front of the elbow, or into one of the prominent veins of the forearms or dorsum of the hand. The left side is favourite as most people are right-handed, but in habitual users, sclerosis of the veins may lead to the arms being used randomly. The veins of the dorsum of the foot may be used when the hands and arms have become unusable because of thrombosis and scarring. Less common sites are in the thighs but here, as with the abdominal wall, the injections may be subcutaneous, rather than intravenous. This mode of injection is known as 'skin-popping' and can lead to areas of subcutaneous sclerosis, fat necrosis, abscesses and, if the injections are deeper into the muscle, to chronic myositis.

Other external signs may be tattoos, often bizarre and connected with the drug subculture. One specific type is tattooing, often of numerals, such as '13', on the buccal (inner) surface of the lower lip. Where chronic addiction has taken its toll, the body may be emaciated, dirty and show signs of infection, especially in the form of skin ulceration. Rarely, there may be necrosis or even loss of phalanges from thrombotic or septic emboli. Old injection marks, sometimes with associated bruising, may be found, the bruising undergoing the usual spectrum of colour changes if not recent. The veins may show overlying fibrosis where phlebitis has occurred, or old venous thrombosis with firm cord-like vessels under the skin.

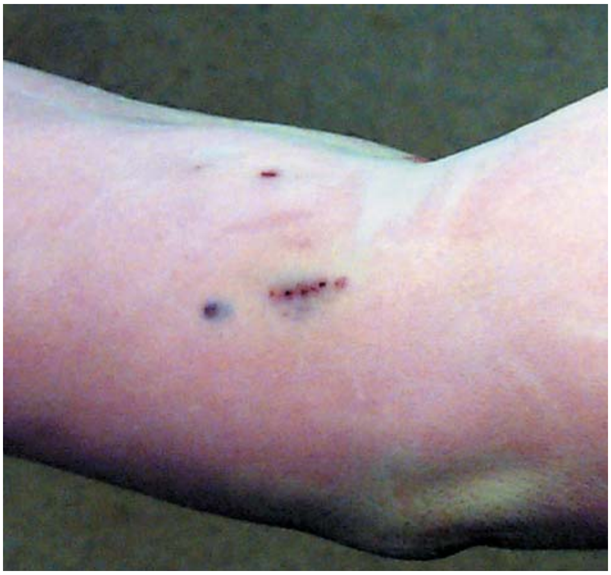

FIGURE 32.1 *Fresh injections marks on the arm of a drug addict.*

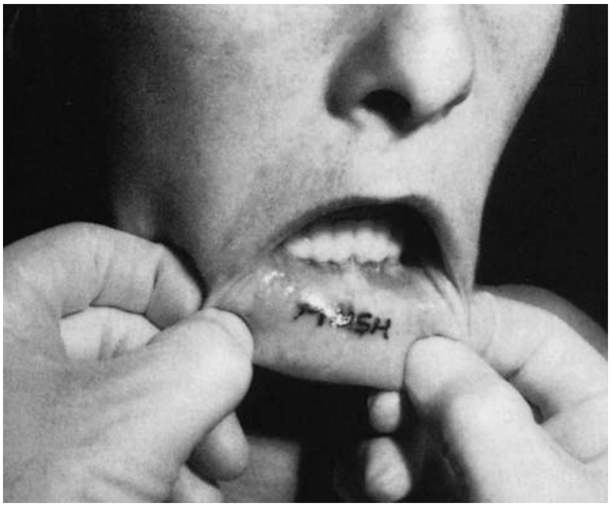

FIGURE 32.2 *Tattoo inside the lower lip of a narcotic drug dealer.*

Where sudden death has occurred in habituated addicts, there may be gross pulmonary oedema, with a plume of froth exuding from the mouth or nostrils, suggestive of drowning. As Polson *et al.* (1983) point out, this may cause some confusion when the victim is found dead in a bathroom, a common venue (along with toilets) for drug administration. This pulmonary oedema is sometimes a striking feature of rapid death in those who are habituated to opioid drugs, especially heroin. It does not seem to occur in novices to the habit, who tend to die in a different way, from a sudden primary cardiac arrest. The oedema may be blood-tinged, again causing confusion with drowning. The cause is not known, but is often attributed to some 'allergic' phenomenon, a rather unsatisfactory explanation. It has been claimed that it sometimes may be caused by adulterants in the drugs, such as quinine, but this is doubtful.

The death may be so rapid that the needle may still be found in the vein when the body is discovered, often in a public lavatory or sometimes in the presence of other addicts, who may then be charged with criminal offences in relation to the death as well as with possessing and using illicit drugs.

There are no other more specific autopsy appearances. Though clinically, a constricted pupil is a prominent sign of morphine administration, after death, any kind of pupil alteration may occur. They may remain small, dilate, or become quite unequal – this applies to any form of death, so no specificity exists for opioid poisoning.

Among new recruits to drug dependence, some die at the first injection of a parenteral dose of heroin or morphine. The mode of death appears to be a cardiac arrest following an arrhythmia and ventricular fibrillation, but no morphological signs can be found. It may be related to sensitization of the myocardium to catecholamines by the drug, the excitement and apprehension being similar to the triggers that seem to cause sudden death from falling into cold water or having a criminal abortion.

The other aspect of autopsies on drug-dependent victims is the risk of infection to pathologist and mortuary staff. Addicts form a high-risk group for both hepatitis B and C, HIV and AIDS. The way in which such autopsies are dealt with in different jurisdictions must vary widely, but it is common practice for HIV and hepatitis testing to be carried out prior to autopsy, unless the policy of the particular autopsy service is to accept all cases with appropriate safety techniques. The analytical laboratory must be informed of any high-risk samples being sent to them, and some will be reluctant to handle blood or tissue samples with positive serology. The pathologist may well refuse to carry out an autopsy on a suspect body until a negative result has been returned on a sample of blood taken (with adequate precautions) from the femoral vein. If positive for hepatitis B or C virus, and especially if the victim shows any signs of hepatic involvement such as jaundice, then many pathologists would decline to carry out a full autopsy unless there are pressing reasons. In that case, scrupulous safety precautions must be taken (see Chapter 1).

In relation to AIDS and HIV positivity, opinions and practices vary widely in the present state of knowledge. It seems obvious that the HIV virus is not in the same category of infectivity as hepatitis. Though autopsies on patients dying of clinical AIDS are routinely performed, there is a difference in the potential risk of a literally 'cold' autopsy being performed on a hospital death after a day or two in the mortuary refrigerator, compared with the urgent activities of a forensic pathologist who may be called upon to conduct an autopsy on a still-warm corpse within a few hours of a suspicious death. However, as mentioned in Chapter 1, HIV may remain viable for many days after death.

TOXICOLOGICAL RESULTS

As with all deaths from toxic substances, the interpretation of laboratory analytical results may present considerable difficulties. There may be a long delay between the administration of the drug and death, during which time the blood, urine and even tissue levels may decline, or even disappear. Many drugs break down rapidly in the body and their metabolites may be the only recognizable products of their administration. In some cases, data on lethal blood levels may be imperfectly known and great variations in personal susceptibility may make the range of concentrations found in a series of deaths so wide as to be rather unhelpful.

As mentioned above, some persons die rapidly after the first episode of taking a 'normal' dose of a drug because of some ill-understood personal idiosyncrasy and here quantitative analysis may not assist.

Where habituation and tolerance has developed, drug users may have concentrations in their body fluids and tissues far higher than lethal levels published for non-habitués. In general, the great usefulness of toxicological analysis is both qualitative and quantitative. The former will show what drugs have been taken in the recent past; the length of time that drugs or their metabolites – persist in different fluids and tissues varies widely.

The quantitative analysis can be useful, especially when the results reveal high levels – into the toxic or lethal ranges. These ranges are usually obtained anecdotally from surveys of large numbers of deaths but, as stated, can differ in terms of minimum and maximum values from different laboratories. The problems of idiosyncratic sensitivity and tolerance cause such published ranges to act only as a general yardstick and deaths that occur outside the ranges (usually off the lower end) cannot be excluded from having been caused by the drug in question if other factors may have been involved. Such factors include the presence of other drugs or alcohol, or both, delayed death and abnormal sensitivity.

Thus the analysis is not the final arbiter of the cause of death, although it is a highly important component of the whole range of investigation. The pathologist has the duty to correlate and interpret all known facts. He must fit the circumstances, the presence of natural disease, trauma and other toxic substances with the laboratory findings, to arrive at the most reasonable cause of death.

The advice of the toxicology laboratory is vital in this process, especially in relation to known lethal ranges and the significance of metabolites, but the analysts should not become the sole arbiter of the cause of death.

METHADONE

Methadone has come to be used primarily as a treatment for established heroin dependency, but unfortunately appears to substitute one form of addiction for another, albeit somewhat less lethal, though recent experience in Scotland (1995) has shown that deaths from methadone may exceed those from heroin, whose use it was intended to supplant. Its clinical uses are as an analgesic more powerful than morphine, with the advantage that it is almost as potent by oral administration as by injection. It has strong and often undesirable sedative properties, however, and is virtually as addictive as the other opioids.

Many deaths occur from its misuse, when obtained either legitimately for the replacement of heroin dependency or on the large illegal market. Most deaths follow oral administration and appear to be due to lack of tolerance, in the absence of which even 50 mg may be lethal, though some recipients appear to die at the outset of the replacement treatment (Drummer *et al.* 1992).

In post-mortem samples, the liver has the highest concentration. The fatal levels found on analysis overlap those found in persons on maintenance doses for heroin replacement, making interpretation difficult. The blood levels published by Manning in 10 fatal cases showed: blood – mean 1.0 mg/l (range 0.1–1.8); liver – mean 3.8 mg/l (range 1.8–7.5); bile – mean 7.5 mg/l (range 2.9–18.0); and brain – mean 1.0 mg/l (range 0.5–1.4).

HALLUCINOGENS

Few of the hallucinogenic drugs are primary causes of death, but some may lead to traumatic deaths because of the abnormal behaviour of the person who is under their influence. For example, the victim of a lysergic acid diethylamide (LSD) 'trip' may be under the impression that he can fly and thus project himself from a high window. Some drugs, however, may have direct toxic effects, though, as with so many drugs of dependence, the autopsy findings are negative or totally non-specific.

Phencyclidine

Chemically this is 1-(1-phenylcyclohexyl) piperidine hydrochloride, known as PCP or 'angel dust' among many other nicknames. It is often used in combination with other drugs of addiction. A number of deaths have been reported from its use, a variety of modes including hyperthermia, intracranial haemorrhage and high output cardiac failure.

TABLE 32.1 *Blood and tissue concentrations (mg/l or kg) of narcotic drugs from various series of fatalities (Baselt* et al. *1975; Stead and Moffat 1983)*

	Baselt *et al.*		
	Range	**Mean**	**Stead and Moffat**
Heroin			
Blood	0.05–3.0	0.3–0.4	>0.1 (mean 0.43)
Urine	0.7–86	5–18	
Bile	3.2–119	32	
Morphine			
Blood	0.2–2.3	0.7	>0.2 (mean 0.7)
Urine	14–18	3.0	
Liver	0.4–18	52	
Cocaine			
Blood	0.9–21	5.2	3.1 (one case)
Urine	1.4–215	47	
Liver	0.1–20	4.3	
Amphetamine			
Blood	0.5–41	8.6	>0.5
Urine	25–700	237	
Liver	4.3–74	30	
Phencyclidine			
Blood	0.3–25	4.8	>0.3 (mean 2.9)
Urine	0.4–120	35	
Liver	0.9–170	23	
Pethidine (meperidine), orally			
Blood	8–20	12	>2 (mean 14)
Urine	150	150	
Liver	5–10	7	

Phencyclidine is excreted in the urine for a long time after ingestion and autopsy samples may be positive for up to a week.

Lysergic acid diethylamide (LSD)

The well-known LSD, which takes its acronym from the German 'Lysergsäurediäthylamid', is a powerful hallucinogenic which is not fatal in itself, as mentioned above. It is an indole alkaloid derivative, of which other members are psilocybin and psieocin, contained in the Mexican mushroom (*Psilocybe mexicana*).

Mescaline, obtained from the cactus *Lophophora williamsii* of central America, is chemically trimethoxyphenethylamine, again not in itself a lethal substance.

CANNABIS

Cannabis, with all its various names in different parts of the world, has at least six active chemical constitutents, the most important of which is one of the tetrahydrocannabinols. Though its effects are a matter of controversy, there seems to be general agreement that used alone, it cannot be blamed for a single death. Proof of the intake of cannabis can be obtained from analysis of urine, blood, and swabs from lips and fingers.

THE AMPHETAMINES

Now rarely used in reputable medical practice because of the undesirable side effects, the amphetamine group of drugs was formerly employed for the abolition of fatigue and for the suppression of appetite. Methylamphetamine sulphate and dexamphetamine were the commonest analogues of amphetamine itself and are still widely available on the illicit market. Death is uncommon from overuse of the amphetamines alone. At autopsy there are no specific findings, apart from the rare possibility of a cerebral or subarachnoid haemorrhage from the induced hypertension.

Toxicology data suggests that after a therapeutic dose of 10 mg, the blood concentration at 2 hours is about 0.035 mg/l. After 30 mg, a peak plasma level of about 0.11 mg/l was observed at 2–5 hours, declining to 0.084 mg/l at 4.5 hours. Chronic abusers consuming large amounts may have blood levels of up to 2–3 mg/l. In the rare fatalities, there is a wider range of blood levels at autopsy, from 0.5 to 41 mg/l (average 8.6), depending on the size of the overdose and the time until death. Liver levels of 474 mg/kg (average 30) and high urine excretion, ranging from 25 to 700 mg/l, with an average of 237 mg/l, are common.

Methylenedioxymethamphetamine (MDMA) or ‘ecstasy’

In recent years, new derivatives of amphetamine have appeared in the drug abuser’s pharmacopoeia, though the substances have been available for medical use for many years. The most commonly encountered is MDMA (3,4-methylenedioxymethamphetamine), known as ‘ecstasy’, ‘XTC’ or ‘ADAM’. Others of the groups are MDA (methyl 3,4-methylenedioxyamphetamine) and MDM (*n*-methyl-MDA). MDMA or ‘ecstasy’ was developed as long ago as 1914 and for some time had a role in psychotherapy as a consciousness-altering agent. It is now widely used as an illegal hallucinogenic drug, and though deaths have been few, they are now being reported from both the USA and Britain. The author (BK) conducted an autopsy on the first fatality in the British Isles, in a man who swallowed a considerable supply on being challenged by the police. There were no autopsy signs apart from a marked generalized cyanotic congestive state and the diagnosis was made purely on toxicological evidence.

Since the first edition of this book, MDMA deaths have increased markedly due to its ubiquitous availability to young people, especially at ‘rave’ parties: hyperthermia, dehydration and myocytolysis are some of the potentially fatal consequences.

COCAINE

Along with heroin, cocaine and its associated drugs like ‘crack’ form the core of the hard-drug problem, from which fatalities are likely to arise. As cocaine is rapidly destroyed when given orally, it is usually taken by injection or sniffed. Recently, a fatality has been reported from absorption from administration via the rectum. This is in addition to some deaths in which cocaine smugglers have died after packets of the drug hidden in their alimentary tract have broken, causing a massive overdose. In India, cocaine has been used by means of urethral instillation and may also be used in the vagina by prostitutes.

In areas where cocaine usage is common, a significant proportion of fetal deaths are associated with the narcotic. In a series reported in New York by Morild and Stajic in 1990, of 103 fetal deaths, toxicology revealed cocaine in 64. The study suggested that fetal death, *abruptio placentae* and abortion were caused by maternal use of the drug.

Death has been recorded by as little as 20–30 mg being applied to the nasal mucosa, but a gram taken by mouth may not be lethal. As with other narcotics, marked habituation and tolerance occurs in chronic users, making estimates of dangerous dose levels difficult to forecast. Intravenous doses are of the order of 100 mg, a common lethal dose being 10 times greater, though far larger amounts can be tolerated by habitual users. Absorption via the nasal mucosa is less effective and larger doses are needed for the same effect than when used parenterally. Though ulceration and even perforation, of the nasal septum in chronic users is mentioned in most standard texts, they are extremely rare lesions.

Death may occur with considerable rapidity in cocaine overdose or hypersensitivity. It may be quite sudden, from cardiac arrest, in first-time users.

At autopsy, there are no specific features. The pulmonary oedema often seen in heroin deaths is not present with cocaine, though the mode of death is also a dysrhythmia.

The diluent used in ‘cutting’ the drug for street sale may be found in the injection sites, the regional lymph nodes, in the lungs and in other organs. Adulterants may be similar to those used with heroin and include talc, starch, quinine, lactose and dextrose. Particles of cocaine itself may also be found as microemboli. Strychnine has been used because of its bitter taste, like quinine. Deliberate poisoning by a supplier has been known by either increasing the strychnine content or by supplying pure cocaine or heroin, leading to a massive overdose.

Cocaine is an excitant and stimulator of the autonomic nervous system. Sudden dramatic rises in blood pressure can occur, sometimes to over 300 mmHg and cerebral haemorrhage is a possible complication of this acute hypertension.

As with the opioids, autopsy may reveal complications of the septic methods used for injection, apart from the already stated risks of virus infection, such as hepatitis and HIV. In former years, in New York and California, transmission of malaria was a well-known problem amongst those who shared syringes and needles – indeed, one suggestion for the use of quinine as a diluent, was as a treatment for the *Plasmodium*.

Pyogenic infections are most common, with phlebitis and distant embolic abscesses occurring. The injection sites may ulcerate, there can be regional lymphadenitis, but more serious sequelae that may be fatal include endocarditis. This can affect any heart valve, including those on the right side, which are not usually affected in post-rheumatic endocarditis. Many organisms are involved, mainly haemolytic and non-haemolytic streptococci, *Streptococcus faecalis*, *Staphylococcus aureus* and *Pseudomonas aeruginosa* and some fungi. Post-mortem blood cultures frequently show such a mixed flora from contamination that isolation of the true causative organism may be difficult, but a heavy predominant growth may be significant. The advice of a microbiologist should be enlisted in assessing the relevance of a positive blood culture.

Any drug used intravenously, if it has particulate matter admixed, such as starch or talc, may cause foreign-body granulomata in the lungs, when the undissolved components are filtered out in the pulmonary capillary bed. Such granulomata are characteristic of intravenous drug abusers – when viewed under a polarizing microscope, refractile elements may be easily visualized.

As cocaine is so commonly used by sniffing, swabs should always be taken from each nostril using a plain cotton-wool swab. An unused swab should be sent to the laboratory as a control. The full range of blood, urine, stomach contents, liver and vitreous samples should routinely be taken at autopsy. Blood levels in fatal cases vary widely, but typical ranges extend from 1 to 21 mg/l, with a mean of 5.2 mg/l according to Baselt *et al.* (1975).

REFERENCES AND FURTHER READING

Baden MM. 1972. Narcotic abuse: a medical examiner’s view. *N Y State J Med* **72**:834–40.

Baden MM. 1972. Homicide, suicide, and accidental death among narcotic addicts. *Hum Pathol* **3**:91–5.

Baselt RC, Allison DJ, Wright JA, *et al.* 1975. Acute heroin fatalities in San Francisco. Demographic and toxicologic characteristics. *West J Med* **122**:455–8.

Billman GE. 1995. Cocaine: a review of its toxic actions on cardiac function. *Crit Rev Toxicol* **25**:113–32.

Burgess C, O’Donohoe A, Gill M. 2000. Agony and ecstasy: a review of MDMA effects and toxicity. *Eur Psychiatry* **15**:287–94.

Carter N, Rutty GN, Milroy CM, *et al.* 2000. Deaths associated with MBDB misuse. *Int J Legal Med* **113**:168–70.

Chakko S, Myerburg RJ. 1995. Cardiac complications of cocaine abuse. *Clin Cardiol* **18**:67–72.

Davis WM, Hatoum HT, Waters IW. 1987. Toxicity of MDA (3,4-methylenedioxyamphetamine) considered for relevance to hazards of MDMA (ecstasy) abuse. *Alcohol Drug Res* **7**:123–34.

Doss PL, Gowitt GT. 1988. Investigation of a death caused by rectal insertion of cocaine. *Am J Forensic Med Pathol* **9**:336–8.

Dowling GP, McDonough ET, Bost RO. 1987. ‘Eve’ and ‘ecstasy’. A report of five deaths associated with the use of MDEA and MDMA. *JAMA* **257**:1615–17.

Drummer OH, Opeskin K, Syrjänen M, *et al.* 1992. Methadone toxicity causing death in ten subjects starting on a methadone maintenance program. *Am J Forensic Med Pathol* **13**:346–50.

Edland JF. 1972. Liver disease in heroin addicts. *Hum Pathol* **3**:75–84.

Finkle BS, McCloskey KL. 1978. The forensic toxicology of cocaine (1971–1976). *J Forensic Sci* **23**:173–89.

Garriot JC, Sturner WQ. 1973. Morphine concentrations and survival periods in acute heroin fatalities. *N Engl J Med* **289**:127–8.

Gill JR, Hayes JA, deSouza IS, *et al.* 2002. Ecstasy (MDMA) deaths in New York city: a case series and review of the literature. *J Forensic Sci* **47**:121–6.

Giroud C, Menetrey A, Augsburger M, *et al.* 2001. Delta(9)-THC, 11-OH-delta(9)-THC and delta(9)-THCCOOH plasma or serum to whole blood concentrations distribution ratios in blood samples taken from living and dead people. *Forensic Sci Int* **123**:159–64.

Gowing LR, Henry-Edwards SM, Irvine RJ, *et al.* 2002. The health effects of ecstasy: a literature review. *Drug Alcohol Rev* **21**:53–63.

Greene MH, Luke JL, Dupont RL. 1974. Opiate 'overdose' deaths in the District of Columbia. I. Heroin-related fatalities. *Med Ann D C* **43**:175–81.

Hayner GN, McKinney H. 1986. MDMA. The dark side of ecstasy. *J Psychoactive Drugs* **18**:341–7.

Helpern M, Rho YM. 1966. Deaths from narcotism in New York city. Incidence, circumstances, and post-mortem findings. *N Y State J Med* **66**:2391–408.

Henry JA. 2000. Metabolic consequences of drug misuse. *Br J Anaesth* **85**:136–42.

Huestis MA, Henningfield JE, Cone EJ. 1992. Blood cannabinoids. II. Models for the prediction of time of marijuana exposure from plasma concentrations of Δ^9-tetrahydrocannabinol (THC) and 11-nor-9-carboxy-Δ^9-tetrahydrocannabinol (THCCOOH). *J Anal Toxicol* **16**:283–90.

Isner JM, Estes NA, Thompson PD, *et al.* 1986. Acute cardiac events temporally related to cocaine abuse. *N Engl J Med* **315**:1438–43.

Karch SB, Billingham ME. 1988. The pathology and etiology of cocaine-induced heart disease. *Arch Pathol Lab Med* **112**:225–30.

Karch SB, Stephens B, Ho CH. 1998. Relating cocaine blood concentrations to toxicity – an autopsy study of 99 cases. *J Forensic Sci* **43**:41–5.

Katsumata S, Sato K, Kashiwade H, *et al.* 1993. Sudden death due presumably to internal use of methamphetamine. *Forensic Sci Int* **62**:209–15.

Kringsholm B. 1988. Deaths among drug addicts in Denmark in 1968–1986. *Forensic Sci Int* **38**:139–49.

Kringsholm B, Christoffersen P. 1987. Lung and heart pathology in fatal drug addiction. A consecutive autopsy study. *Forensic Sci Int* **34**:39–51.

Kringsholm B, Christoffersen P. 1987. Lymph-node and thymus pathology in fatal drug addiction. *Forensic Sci Int* **34**:245–54.

Kringsholm B, Voigt J, Dalgaard JB, *et al.* 1981. Deaths among narcotic addicts in Denmark in 1978 and 1979. *Forensic Sci Int* **18**:19–30.

Laposata EA, Mayo GL. 1993. A review of pulmonary pathology and mechanisms associated with inhalation of freebase cocaine ('crack'). *Am J Forensic Med Pathol* **14**:1–9.

Ling LH, Marchant C, Buckley NA, *et al.* 2001. Poisoning with the recreational drug paramethoxyamphetamine ('death'). *Med J Aust* **174**:453–5.

Lockemann U, Püschel K. 1993. Siderophages in the lung of drug addicts. *Forensic Sci Int* **59**:169–75.

Manno JE, Manno BR, Kemp PM, *et al.* 2001. Temporal indication of marijuana use can be estimated from plasma and urine concentrations of Δ^9-tetra-hydrocannabinol, 11-hydroxy-Δ^9-tetra-hydrocannabinol, and 11-nor- Δ^9-tetra-hydrocannabinol-9-carboxylic acid. *J Anal Toxicol* **25**:538–49.

Martin TL. 2001. Three cases of fatal paramethoxy-amphetamine overdose. *J Anal Toxicol* **25**:649–51.

Mirchandani HG, Rorke LB, Sekula Perlman A, *et al.* 1994. Cocaine-induced agitated delirium, forceful struggle, and minor head injury. A further definition of sudden death during restraint. *Am J Forensic Med Pathol* **15**:95–9.

Moore C, Negrusz A, Lewis D. 1998. Determination of drugs of abuse in meconium. *J Chromatogr B Biomed Sci Appl* **713**:137–46.

Morild I, Stajic M. 1990. Cocaine and fetal death. *Forensic Sci Int* **47**:181–9.

O'Connor A, Cluroe A, Couch R, *et al.* 1999. Death from hyponatraemia-induced cerebral oedema associated with MDMA ('ecstasy') use. *N Z Med J* **112**:255–6.

Parr MJ, Low HM, Botterill P. 1997. Hyponatraemia and death after 'ecstasy' ingestion. *Med J Aust* **166**:136–7.

Paties C, Peveri V, Falzi G. 1987. Liver histopathology in autopsied drug-addicts. *Forensic Sci Int* **35**:11–26.

Polson CJ, Green MA, Lee M. 1983. *Clinical toxicology*, 3rd edn. Pitman, London.

Price KR. 1974. Fatal cocaine poisoning. *J Forensic Sci Soc* **14**:329–33.

Ramcharan S, Meenhorst PL, Otten JM, *et al.* 1998. Survival after massive ecstasy overdose. *J Toxicol Clin Toxicol* **36**:727–31.

Rammer L, Holmgren P, Sandler H. 1988. Fatal intoxication by dextromethorphan: a report on two cases. *Forensic Sci Int* **37**:233–6.

Richards RG, Reed D, Cravey RH. 1976. Death from intravenously administered narcotics: a study of 114 cases. *J Forensic Sci* **21**:467–82.

Siegel H, Helpern M, Ehrenreich T. 1966. The diagnosis of death from intravenous narcotism. With emphasis on the pathologic aspects. *J Forensic Sci* **11**:1–16.

Stead AH, Moffat AC. 1983. A collection of therapeutic, toxic and fatal blood drug concentrations in man. *Hum Toxicol* **2**:437–64.

Steentoft A, Worm K, Christensen H. 1988. Morphine concentrations in autopsy material from fatal cases after intake of morphine and/or heroin. *J Forensic Sci Soc* **28**:87–94.

Stevens BC. 1978. Deaths of drug addicts in London during 1970–4: toxicological, legal and demographic findings. *Med Sci Law* **18**:128–37.

Suarez RV, Riemersma R. 1988. 'Ecstasy' and sudden cardiac death. *Am J Forensic Med Pathol* **9**:339–41.

Watling R. 1983. Hallucinogenic mushrooms. *J Forensic Sci Soc* **23**:53–66.

Weinmann W, Bohnert M. 1998. Lethal monointoxication by overdosage of MDEA. *Forensic Sci Int* **91**:91–101.

Wetli CV. 1987. Fatal cocaine intoxication. A review. *Am J Forensic Med Pathol* **8**:1–2.

CHAPTER 33

Corrosive and metallic poisoning

- Cyanide 585
- Corrosive acids, alkalis and phenols 587
- Oxalic acid and oxalate salts 588
- Ethylene glycol poisoning 589
- Metallic poisons 589
- References and further reading 592

Toxicology is a vast subject, most of it concerned with the nature, occurrence, symptomatology, biochemistry, mode of action and treatment of a wide range of poisonous substances. Many forensic medicine textbooks, especially those from Asia, devote a major part of their text to all these aspects of hundreds of toxic substances, many of which are seldom – if ever – encountered by a pathologist in most parts of the world. As the autopsy appearances of most poisons are non-specific, it seems fruitless to offer a repetitive catalogue, and therefore the descriptions selected here refer to some of those that either have specific features or are encountered more often. This chapter describes the autopsy appearances in a range of poisons that can broadly be classed as 'corrosive', even if this is not necessarily their main lethal mode of action. In addition, several toxic heavy metals will be discussed, again from the point of view of autopsy findings and relevant toxicological laboratory findings.

CYANIDE

Cyanide is a relatively common poison, both in suicide, accident and, occasionally, homicide. It forms part of lethal toxicity of many fires in buildings, where smoke inhalation kills the majority of victims, rather than burns (Chapter 11). Although the autopsy diagnosis of acute cyanide poisoning is rarely in doubt, toxicological analysis may be difficult to interpret because of both the destruction and production of cyanide in the dead body and even in stored blood samples awaiting analysis. Acute cyanide poisoning is most often self-administered (70 per cent in one series), in which case usually the sodium or potassium salt is swallowed. It may be accidental or industrial, in which case either salts may be involved, or it may be the free gas liberated from some commercial process.

Homicidal poisoning is rare, except for the mass homicides which still occur, such as the Jonesville tragedy in Guyana, or the use of cyanide as a weapon of war against civilians in the Middle East. It has also been used for judicial execution in parts of the USA, a practice which seems to be reviving in recent years.

Cyanide acts only as free hydrogen cyanide and therefore swallowed salts need to meet either water or gastric acid before liberating hydrocyanic acid, a process that takes only a few seconds. The fatal dose of cyanide is small, of the order of 150–300 mg, which allowed it to be used as hidden suicide pills by prominent Nazis at the end of the last war. Recovery has been recorded from far greater doses, however, such as 2–4 g of potassium cyanide. Much depends on the purity of cyanides, as they tend to decompose in storage and old samples may contain only half the weight as active cyanide.

Autopsy findings in cyanide poisoning

Cyanide acts by linking with the ferric iron atom of cytochrome oxidase, preventing the uptake of oxygen for cellular respiration. Cyanide cannot combine directly with haemoglobin, but can do so through the intermediary

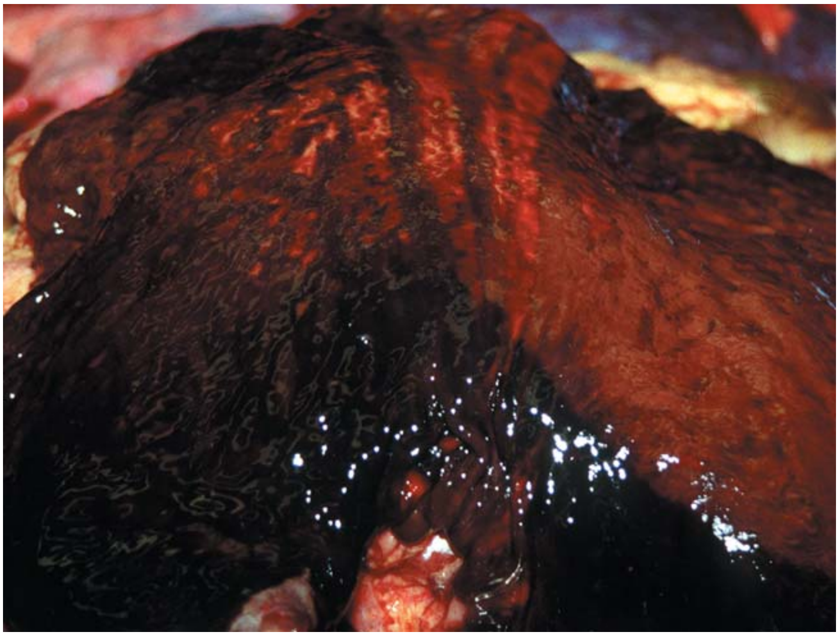

FIGURE 33.1 *Haemorrhagic gastric mucosa in cyanide poisoning.*

compound methaemoglobin. Cyanides are moderately corrosive through their alkaline nature, causing local tissue damage that is unrelated to their more general toxicity via enzyme inhibition.

Externally there can be wide variations in the appearance. Traditionally, the hypostasis is said to be brick-red, due to excess oxyhaemoglobin (because the tissues are prevented from using oxygen) and to the presence of cyanmethaemoglobin. Many descriptions refer to a dark pink or even bright red skin, especially in the dependent areas, which can be confused with carboxyhaemoglobin. The few cases seen by the authors have shown a marked dark cyanotic hypostasis, perhaps caused by lack of oxygenation of the red cells by paralysis of the respiratory muscles. There may be no other external signs apart from the colour of the skin and possibly black vomit around the lips.

There may be a smell of cyanide about the body, though it is well known that many persons cannot detect this, the ability being a sex-linked genetic trait. This may be of importance to pathologists and mortuary staff, as corpses dead of cyanide poisoning can present a health hazard. A former colleague of the author (BK) became ill and was temporarily disabled shortly after conducting an autopsy on a suicide who had swallowed a massive amount of potassium cyanide. Presumably he had inhaled hydrogen cyanide from the stomach contents when examining the viscera.

Internally the tissues may also be bright pink caused by the oxyhaemoglobin that cannot be utilized by the tissues – which is probably more common than the presence of cyanmethaemoglobin. The stomach lining may be badly damaged and can present a blackened, eroded surface, by altered blood staining the stripped mucosa. This is mainly because of the strongly alkaline nature of the hydrolysed sodium or potassium salts of cyanide; hydrogen cyanide itself causes no such damage. In less severe cases, the stomach lining will be streaked with dark red striae, where the rugae have been eroded while leaving the intervening folds relatively unharmed. The stomach may contain frank or altered blood from the erosions and haemorrhages in the walls. If the cyanide was in dilute solution, there may be little damage to the stomach, apart from pinkness of the mucosa and perhaps some petechial haemorrhages. There may also be undissolved white crystals or powder, with the almond-like smell of cyanide mentioned above.

As death is usually rapid, little of the contents will have passed into the intestine. The oesophagus may be damaged, especially the mucosa of the lower third, though some of this may be a post-mortem change from regurgitation of the stomach contents through the relaxed cardiac sphincter after death. The other organs show no specific changes and the diagnosis is made by history, smell and the reddish colour of the internal tissues, and often skin.

Toxicological analysis

The usual blood, stomach contents, urine and any vomit should be submitted to the laboratory, taking particular care that the samples present no hazard to those packing, transporting or unpacking them. The laboratory should be warned in advance that a possible cyanide case is coming their way.

If death was possibly caused by the inhalation of hydrogen cyanide fumes, a lung should be sent intact, sealed in a nylon (not polyvinylchloride) bag.

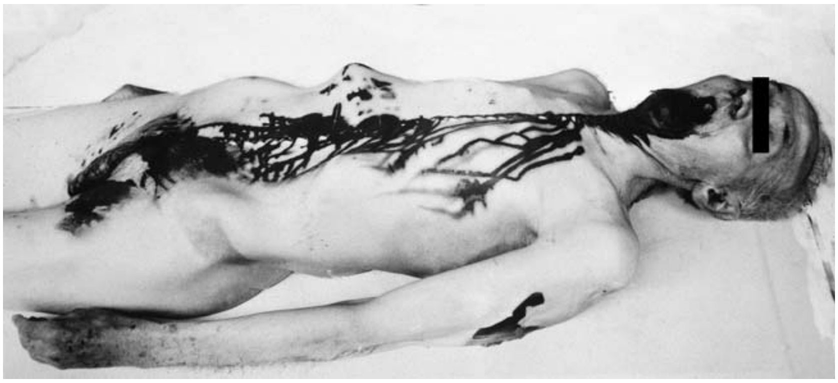

FIGURE 33.2 *Skin burns on a victim of lysol suicide. The trickle pattern makes it obvious that the man was either standing or sitting erect when he drank the fluid. Lysol burns are usually leathery and brownish-purple.*

It is important to get the samples to the laboratory as soon as possible (in terms of days) to avoid the spurious formation of cyanide in stored blood samples. This usually occurs at room temperature so, if there is to be a delay, refrigeration is essential. By contrast, some positive samples may actually decrease on storage, as described by Curry (1969). Up to 70 per cent of the cyanide content may be lost after some weeks, from reaction with tissue components and conversion to thiocyanate.

The amount found on analysis naturally depends on the amount taken and the time between administration and death. Though the latter is usually measured in minutes, low dosage – or treatment – may allow survival for hours or even days.

Assuming that no spurious cyanide is formed, any significant amount found is evidence of cyanide ingestion, which in itself is abnormal and presumably confirmatory evidence of poisoning. However, Karhunen *et al.* (1991) have reported a case in which post-mortem burning of a homicide victim led to a blood cyanide level of 10 mg/l, presumably due to passive diffusion of cyanide through body cavities opened by the fire.

Typical blood levels in one series of fatal cases following ingestion of the poison range from 1 to 53 mg/l, with an average of 12 mg/l. The spleen always has the highest tissue concentration, presumably because it contains so many red cells; in the same series, the spleen level was between 0.5 and 398 mg/l, with a mean of 44 mg/l. In another series, mean blood levels were 37 mg/l.

CORROSIVE ACIDS, ALKALIS AND PHENOLS

Corrosive poisons were formerly common suicidal agents, though they are now relatively rare in Western countries, probably because of the ease of obtaining less painful substances. In some parts of the world, mineral acids are still often used for homicide, assault ('vitriol-throwing') and suicide. In Malaya, reagents used in rubber production, such as formic and acetic acids, were often taken as a means of self-destruction by young women, especially Tamil rubber workers.

In Britain, acids and alkalis are now almost unknown as agents of death. Even the occasional use of sulphuric 'battery acid' as a weapon of assault rarely causes death. The phenolic corrosives, however, such as carbolic acid and lysol, are occasionally encountered as suicidal agents. Toxicologically, none of these presents much problem, as the damage is often structural rather than poisonous, unless the victim survives long enough to have complications such as renal failure or chest infections. All the corrosive substances have the following features in common:

- There may be spillage of the fluid on the exterior of the body, corroding the skin in a pattern which may be helpful in reconstructing the posture of the victim at the time of drinking the substance. The lips may be burnt, and trickle and splash marks may run from the mouth down the chin, neck and chest. The pattern of burns at the mouth may sometimes indicate the shape of the container from which the poison was drunk, as the wide brim of a cup may mark the cheeks, while a bottleneck may sit more cleanly in the mouth. If the person was standing or sitting, then these runnels of fluid may pass down the chin onto chest and abdomen. If lying, then they may run across the face and cheeks and pass to the back of the neck. Further spillage may come from the nostrils due to spluttering and gagging. The hands may also be affected if the hands are instinctively brought up to the face.
- The interior of the mouth may be eroded, and the tongue swollen or shrivelled, according to the nature of the corrosive agent. The pharynx, larynx and oesophagus are all eroded, and if survival lasts more than a few minutes the glottis area may become oedematous. Spillage into the larynx and air passages may allow the respiratory mucosa to be damaged, and aspiration of liquid or vapour into the lungs can cause rapid pulmonary oedema and haemorrhages.

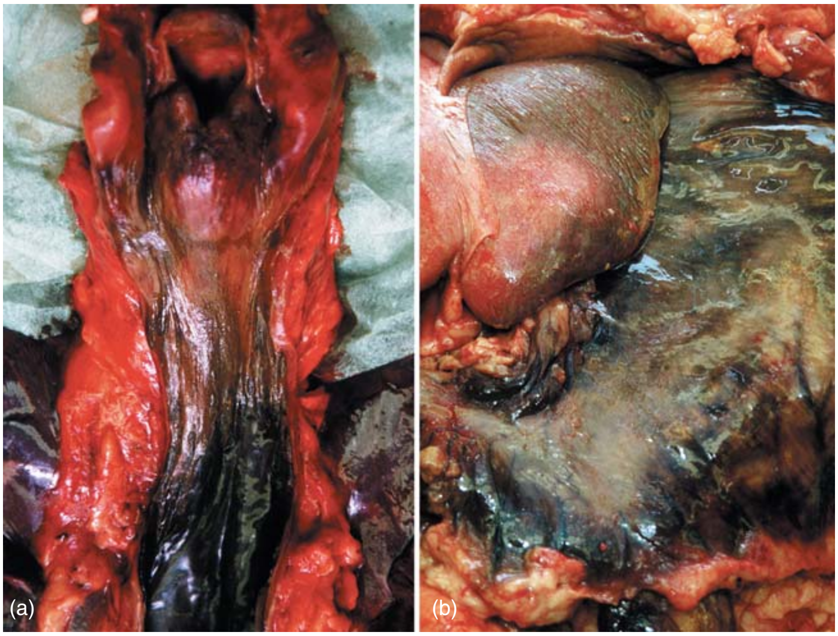

FIGURE 33.3 *Corrosion of larynx, oesophagus (a) and stomach (b) in suicidal ingestion of concentrated hydrochloride acid with alcohol.*

- The lower oesophagus and stomach rapidly become damaged, with discoloration, desquamation and sometimes perforation. Attempts at passing a stomach tube may themselves penetrate the softened wall of the oesophagus or stomach. If survival lasts long enough, some corrosive may pass through to damage the small intestine, but this is rare because of the time factor and spasm of the pylorus.
- All may cause death by pulmonary oedema from spillage into the lungs: if survival lasts a day or more, then a fulminating bronchopneumonia may be the terminal event.

The different corrosive agents have different actions on soft tissues, which can sometimes be differentiated by appearance and smell, though the mineral acids are not all that different. The phenolic compounds can usually be detected by smell alone. Strong acids act by dehydrating the tissues, coagulating the proteins and converting haemoglobin to haematin.

Sulphuric acid, in concentrated form, is extremely corrosive and produces great heat in contact with water or tissues. The tissues are grey to black, rather dry and dehydrated. They may actually be charred into a blackened crust by the generated heat. The gastric lining may be grey, dark brown or black, or mixtures of all colours, depending upon the amount of altered blood in each part. Perforation may occur.

The oesophagus and stomach may be grey and swollen, depending on the amount of acid and the amount of food already in the stomach. The tongue may be grey or black and distorted.

Nitric acid is similar, but has a brownish-yellow cast to its mucosal damage. There may be yellow or brown sharp-edged patterns on the skin burns of the face, with the usual trickle marks coming down from the corners of the mouth. Yellow fumes may arise from the stomach contents if a considerable quantity is present. The internal appearances are of yellowish-brown sloughing, though perforation seems less common than with sulphuric acid.

Hydrochloric acid has similar effects, especially on mucous membranes, but is not so injurious to intact skin as sulphuric and nitric acids. The stomach may be converted into a slimy, softened mass and can perforate. The colour is greyish to black, depending on the amount of altered blood.

Sodium hydroxide in concentrated form is also a corrosive, but soft, slippery slime is the characteristic appearance and feel to tissues damaged by caustic soda. The colour is dirty white to grey.

Phenol and lysol are also damaging and affect the tissues in much the same way as acids and alkalis. Carbolic acid (pure phenol) tends to stiffen the tissues and bleach them so that hard, cracked, whitish surfaces are seen on the face and skin. Internally, the same stiffness is noted in the oesophagus and stomach. Lysol is a soapy solution of phenol and cresols. It discolours the tissues a brownish purple, but is otherwise similar to phenol in its action.

OXALIC ACID AND OXALATE SALTS

These are not so corrosive as the mineral acids, but are poisonous and often act quickly, death occurring within minutes or the hour, from shock or hypocalcaemia. The acid is locally corrosive, but also has a systemic effect that may well be fatal even if the local damage is non-lethal.

At autopsy, if an appreciable amount of either the white crystals or a strong solution has been swallowed, the local effect is a bleaching, the mucosa of mouth, pharynx and oesophagus being white, though local haemorrhage can streak this with red. The stomach contains altered blood from the damaged mucosa and is dark brown or black from acid haematin, the wall studded with acute erosions. Calcium oxalate crystals may be seen in the stomach contents or in scrapings from the mucosa.

In those who have survived the acute phase, death may be caused by abnormalities of the muscle function (including the myocardium) from the hypocalcaemia, caused by the precipitation of body calcium as insoluble calcium oxalate. More common is renal failure, death occurring 2–10 days later. The renal tubules suffer necrosis, primarily in the proximal convoluted tubules. This is not caused by the presence of calcium oxalate crystals, though these can be demonstrated histologically in the kidney.

ETHYLENE GLYCOL POISONING

Though in no sense a corrosive poison, ethylene glycol has certain features in common with oxalate poisoning and is so common relative to death from mineral acids that it cannot be omitted. The glycols are used widely as antifreeze agents in motor engines and as solvents in industry, so they are easily available. Because of their chemical inclusion in the alcohol group, they are abused as a source of intoxication, as well as being accidental and suicidal agents. At least 40–60 deaths a year are reported and this is probably an underestimate. The compounds involved are ethylene, diethylene, propylene and hexylene glycols. These do not have the same toxic effects (in fact, propylene glycol is virtually non-toxic), but ethylene glycol is the most commonly encountered. When drunk in excess of 100–200 ml, it is almost certain to be fatal unless specific treatment is given, such as dialysis and competition with alcohol.

The first effects resemble drunkenness, but this passes into coma and death often within the first day. The glycol is metabolized in the body, a small but significant amount (about 1 per cent) being converted to oxalic acid, via the process glycol–glyoxal–glycolic acid–formic acid–glyoxylic acid–oxalic acid. It is not clear which of these compounds causes the most damage to the tissues.

At autopsy there is no local damage, but widespread precipitation of the sheaf-like doubly refractile crystals of calcium oxalate into the tissues can be rendered visible microscopically by the use of polarized light. It is a matter of controversy whether this crystal deposition is the cause or merely a side effect of the lethal action of glycols. There may be cerebral oedema and a chemical meningoencephalitis. In the kidney there is a tubular necrosis similar to that in oxalate poisoning, and the crystals can be seen in the interstitial tissues and the tubules. Diffuse liver damage can also occur.

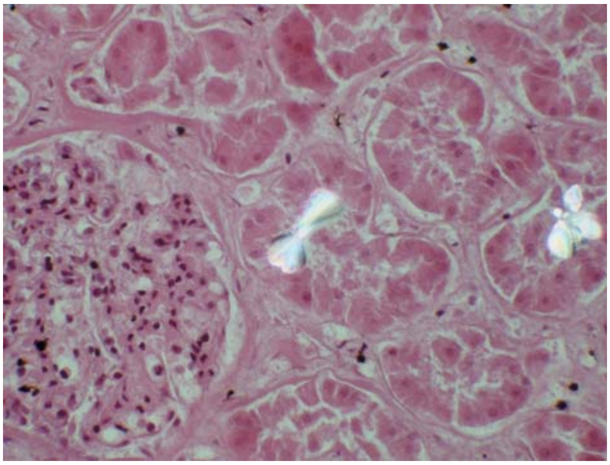

FIGURE 33.4 *Birefringent calcium oxalate crystals in kidney in ethylene glycol poisoning as seen in polarized light under a microscope.*

Laboratory analysis in ethylene glycol poisoning

The blood levels naturally depend on the time that has elapsed since ingestion. It is usually greater than 300 mg/l and the mean is as high as 2400 mg/l in samples taken before the second day. The brain may contain from 300 to 4000 mg/kg and the urine from 600 to 10 000 mg/l, with a mean of 5700 mg/l.

METALLIC POISONS

There is a whole range of metallic poisons, most of them from the upper reaches of the Periodic Table, accounting for their usual description as ‘heavy metals’. The vast majority of toxic effects come from environmental and occupational poisoning, both chronic and acute. Acute poisoning by suicide, accident and homicide is becoming much less common, both because of the availability of other toxic agents, and because of greater awareness and controls on the environmental and industrial hazards of heavy metals. In former years, especially the nineteenth century, heavy-metal poisoning was common in homicide, but is now rarely seen in Western countries, mainly because they are now known to be easily detectable.

As the subject is dealt with in great detail by so many textbooks and monographs, only one such poison – arsenic – will be discussed, together with a mention of any significant differences with other metals.

Arsenic

A constituent of all animal tissue, arsenic is the twelfth most abundant element on earth. This emphasizes the need for strict controls when making analysis for arsenic in human fluids or tissues, as the excretion of a healthy person on a diet rich in fish (especially shellfish) can exceed that seen in chronic arsenical poisoning. Similarly, where an exhumation is performed after allegations of poisoning, full control samples of soil and grave water must be taken to ensure that arsenic found in the body could not have arisen from local contamination (Chapter 1).

Arsenic metal is not poisonous, only its compounds. These interfere with cellular respiration by combining with the sulphydryl groups of mitochondrial enzymes, especially pyruvate oxidase and certain phosphatases. Arsenic has a particular target in vascular endothelium, accounting for the many lesions caused by increased permeability, tissue oedema and haemorrhage, especially in the alimentary canal.

Arsenical poisoning may arise from the ingestion of arsenious oxide, a tasteless white powder – from copper, sodium and potassium arsenites, arsenates of lead and calcium, arsenic sulphides and gaseous arsine (confined to industry). In forensic practice, the rare cases of arsenic poisoning are now usually from arsenious oxide or one of the arsenites.

Arsenical poisoning may be acute or chronic, the latter being the presentation of most environmental and occupational toxicity. Suicides are obviously invariably acute, whereas the uncommon homicidal cases may be either acute or chronic.

FATAL DOSAGES IN ARSENICAL POISONING

If taken on an empty stomach, especially in solution, only about 150 mg may be fatal, but usually some 250–300 mg are needed as a minimum lethal dose. Much larger quantities have been survived and there is some evidence that tolerance to arsenic can be attained. With large doses, much may be vomited. Death can be rapid – within hours from 'shock' and cardiorespiratory failure – or may be delayed for many days, when hepatorenal failure is the mode of death. In chronic poisoning, no lethal dosage can be indicated, as if the ingestion exceeds the small normal excretion rate, then a cumulative build-up of arsenic will occur.

Autopsy findings

In acute poisoning the findings may be minimal, if death occurs within hours. There may be some mild irritation of the upper gastrointestinal tract, such as reddening of the gastric mucosa, especially along the top edges of the rugae. The description of 'red velvet' has been applied to some stomach linings in acute arsenic poisoning. There may be mucus coating and granules of the poisonous agent may be trapped on the lining – a reason for sending both contents and stomach wall for analysis, as in most suspected poisonings of any type. The intestines are usually normal in acute poisoning.

The only other lesion commonly seen is subendocardial haemorrhage on the left ventricular wall. This, of course, is a common finding in any severe shock condition when sudden hypotension occurs. It is seen in any gross injury, with loss of blood volume, blood pressure and neurogenic shock. Head injuries and raised intracranial pressure are other conditions in which these lesions are prominent (Chapter 13). The phenomenon can be rapid, the author (BK) having seen them in the victims of military air crashes where the heart was avulsed from its base at the moment of impact. One of the best descriptions of subendocardial haemorrhages was given by Sheehan in 1940 in obstetrical catastrophes. The haemorrhages are situated on the upper part of the interventricular septum and on the opposing papillary muscles. The most severe the author (BK) has ever seen was in a massive suicidal overdose of arsenious oxide, where the haemorrhages were actually raised blood-filled blisters under the endocardium. These suggested the true diagnosis before analysis was carried out, albeit with an element of speculative good luck.

In chronic arsenical poisoning, the picture is quite different, though unless some suggestive history or circumstantial evidence is available, the diagnosis may be extremely difficult. As many a clinician has discovered to his chagrin in the past years, many a chronic arsenical poisoning has been misdiagnosed as a gastroenteritis.

Part of forensic mythology surrounds the alleged preservation of corpses dying from arsenic poisoning. This has been endlessly discussed, but there is no real evidence that it is true. A more likely explanation is that the dehydration from diarrhoea in chronic poisoning retards the usual moist putrefaction.

Externally there may be a diffuse abnormality of the skin, with a dry, scaly, hyperkeratotic surface. Clinically, there is a 'rain-drop' punctate pigmentation, but this may not be apparent after death unless really marked. It is more common in skin flexures and over the forehead and neck. There may be some hair loss. Puffy thickening and oedema of the face has been described, suggestive of myxoedema.

Internally the stomach may be normal or may show a chronic gastritis with thickening of the mucosa and subserous coats. Mucus formation may be noticeable and there may be a patchy inflammatory redness of some of the rugae. Sometimes there is a haemorrhagic gastritis with acute and chronic erosions. The small intestine is dilated and generally reddened, with thickened mucosa, the whole picture being of a non-specific congestion oedema so commonly seen in any enteritis.

There is rarely any mucosal ulceration. The contents may be copious and fluid, the usual description of 'rice-water' being applied. The large intestine may show minimal changes or be normal: the contents may be fluid and similar to the small bowel. The liver may reveal fatty change or more severe necrosis, sometimes at the periphery of the lobule. Severe liver damage may be associated with externally apparent jaundice. The kidney is damaged in chronic toxicity, there being non-specific tubular necrosis. The myocardium may also show myofibril damage, interstitial collection of cells and sometimes fatty degeneration.

TOXICOLOGY SAMPLING AND ANALYTICAL RESULTS

In acute poisoning, the major requirement is the stomach and contents, and preferably the small intestine, tied off at each end. Blood, urine and liver should also be taken.

In chronic poisoning, especially if the diagnosis is not firmly established from circumstantial and gross autopsy findings, a much wider range of samples is needed:

- blood from peripheral veins
- stomach and contents
- small intestine and contents
- sample of large bowel contents
- urine
- bile
- whole liver
- one kidney
- nail clippings or whole nails
- hair samples – whole length of at least 20 hairs, including roots.

Arsenic levels in blood are elevated only for a short time following absorption, unless exposure is continuous. The highest concentrations of arsenic are found tissues rich in sulphydryl (SH-) groups, such as skin, hair and nails. Typical levels of arsenic found in healthy persons unexposed to extra amounts over and above that normally (but variably) present in their environment are as follows: urine, <2–20 μg/l; blood, <0.5–4 μg/l; liver, <60 μg/kg (dry weight); nails, $<10^3$ μg/kg (dry weight).

It was formerly thought that it took a week or two for ingested arsenic to find its way into the keratinized tissues such as hair and nails. More sensitive analytical techniques have, however, shown that it can appear there within hours of ingestion by mouth. The explanation was originally offered that arsenic was rapidly secreted into sweat, and spread to the hair and nails by surface diffusion. This has recently been disproved by neutron activation analysis and the mechanism remains obscure. The test for estimating the timing and duration of arsenic administration by analysing different lengths of the hair shaft has thus been vindicated. Arsenic can remain stored in keratin for several years after administration ceases.

TABLE 33.1 *Levels in acute arsenious oxide poisoning from 49 fatalities (Rehling 1967)*

	Range	Average
Blood (mg/l)	0.6–9.3	3.3
Liver (mg/kg)	2.0–120	29.0
Kidney (mg/kg)	0.2–70	15.0
Spleen (mg/kg)	0.5–62	8.8
Brain (mg/kg)	0.2–4.0	1.7

Antimony

Antimony is similar to arsenic in many respects. Apart from industrial exposure, the usual vehicle of poisoning is 'tartar emetic', antimony potassium tartrate. This has been used suicidally and homicidally, as well as in some accidental deaths. The pathology and toxicology are similar to that of arsenic, the minimum fatal dose of the tartrate being about 150 mg, though this was formerly the normal intravenous dose in the treatment of bilharzia. Vastly greater amounts of tartar emetic have been survived, so the minimal lethal dose is not a particularly useful concept, as with many other poisons.

Occupationally unexposed people have either zero or not more than 0.01 mg/l of antimony in their blood and a tenth of this concentration in their urine.

Baselt (2002) quotes a case of acute poisoning from antimony trichloride where the following concentrations (in mg/l or mg/kg) were found on analysis: blood, 4–6; liver, 45; kidney, 32; bile, 404.

Thallium

Though other heavy metals have declined in forensic importance over the past century, thallium has several times been used homicidally in recent years, sometimes for multiple poisonings. Thallium is used as a rat killer and widely employed in industry, especially in glass manufacture. It has curious aspects in relation to forensic pathology, in that it can be seen radiologically in the intestine and deposited in the liver, so in the rare event of a pathologist suspecting thallium poisoning,

TABLE 33.2 *Fatal thallium concentrations (mg/l or mg/kg) collected by Baselt (2002)*

	Range	Average
Blood	0.5–11	4.0
Urine	1.7–11	5.2
Kidney	6–20	11.0
Liver	5–29	15.0

X-rays of the body should be taken before autopsy. The other unique aspect is that it is probably the only homicidal agent to be confirmed after cremation. In the Young case in Britain in 1971, where two people died and several others were non-fatally poisoned by Young, thallium was detected retrospectively in the ashes of one of his victims.

The fatal dose is somewhere about 1 g, depending upon the type of thallium compound employed, as there are several different salts available, such as the acetate, sulphate or nitrate. The estimates vary from 6 to 40 mg/kg body weight, with an average of about 12 mg/kg.

Autopsy appearances are variable and non-specific, but pallor and streaking of a pale, degenerate myocardium have been recorded. Hair loss is one of the clinical signs that arouses suspicion of thallium poisoning, as it was formerly used as a depilatory. This effect begins about a week after administration, but may not be noticeable for twice that time. Large tufts tend to come away, rather than a general thinning. Loss of the outer third of the eyebrows is said to be a significant sign, though these are also the areas that are lost in hypothyroidism.

Examination of the roots of surviving hairs may show a black coating, caused not by the thallium itself, but by an excess of melanin. Liver necrosis and renal tubular necrosis are non-specific findings in those who survive for some time.

REFERENCES AND FURTHER READING

Ballantyne B. 1976. Changes in blood cyanide as a function of storage time and temperature. *J Forensic Sci Soc* **16**:305–10.

Barillo DJ, Goode R, Esch V. 1994. Cyanide poisoning in victims of fire: analysis of 364 cases and review of the literature. *J Burn Care Rehabil* **15**:46–57.

Baselt RC. 2002. *Disposition of toxic drugs and chemicals in man*, 6th edn. Biomedical Publications, Foster City, CA.

Baud FJ, Barriot P, Toffis V, *et al.* 1991. Elevated blood cyanide concentrations in victims of smoke inhalation. *N Engl J Med* **325**:1761–6.

Blackwell M, Robbins A. 1979. Arsine (arsenic hydride) poisoning in the workplace. *Am Ind Hyg Assoc J* **40**:A56–61.

Bogusz M. 1980. Vitreous humour as reliable material for ethylene glycol determination. *Forensic Sci Int* **16**:75–6.

Bowen DA, Minty PS, Sengupta A. 1978. Two fatal cases of ethylene glycol poisoning. *Med Sci Law* **18**:101–7.

Broor SL, Kumar A, Chari ST, *et al.* 1989. Corrosive oesophageal strictures following acid ingestion: clinical profile and results of endoscopic dilatation. *J Gastroenterol Hepatol* **4**:55–61.

Chaturvedi AK, Smith DR, Canfield DV. 2001. A fatality caused by accidental production of hydrogen sulfide. *Forensic Sci Int* **123**:211–14.

Cheng JT, Beysolow TD, Kaul B, *et al.* 1987. Clearance of ethylene glycol by kidneys and hemodialysis. *J Toxicol Clin Toxicol* **25**:95–108.

Chikasue F, Yashiki M, Kojima T, *et al.* 1988. Cyanide distribution in five fatal cyanide poisonings and the effect of storage conditions on cyanide concentration in tissue. *Forensic Sci Int* **38**:173–83.

Christesen HB. 1995. Prediction of complications following caustic ingestion in adults. *Clin Otolaryngol* **20**:272–8.

Curry A. 1969. *Poison detection in human organs.* Thomas, Springfield, IL.

Curry AS. 1963. Cyanide poisoning. *Acta Pharmacol Toxicol* **20**:291–4.

Curry AS, Price DE, Rutter ER. 1967. The production of cyanide in post mortem material. *Acta Pharmacol Toxicol (Copenh)* **25**:339–44.

Davis LE, Standefer JC, Kornfeld M, *et al.* 1981. Acute thallium poisoning: toxicological and morphological studies of the nervous system. *Ann Neurol* **10**:38–44.

Dilawari JB, Singh S, Rao PN, *et al.* 1984. Corrosive acid ingestion in man – a clinical and endoscopic study. *Gut* **25**:183–7.

Erkens M, Kellner H. 1981. [Toxicologic course studies in thallium poisoning.] *Beitr Gerichtl Med* **39**:157–61.

Fernando GC, Busuttil A. 1991. Cyanide ingestion. Case studies of four suicides [published erratum appears in *Am J Forensic Med Pathol* 1993 Dec;**14**(4):353]. *Am J Forensic Med Pathol* **12**:241–6.

Ferrari LA, Arado MG, Giannuzzi L, *et al.* 2001. Hydrogen cyanide and carbon monoxide in blood of convicted dead in a polyurethane combustion: a proposition for the data analysis. *Forensic Sci Int* **121**:140–3.

Galvan-Arzate S, Santamaria A. 1998. Thallium toxicity. *Toxicol Lett* **99**:1–13.

Gebel T. 2000. Confounding variables in the environmental toxicology of arsenic. *Toxicology* **144**:155–62.

Hawley CK, Harsch HH. 1999. Gastric outlet obstruction as a late complication of formaldehyde ingestion: a case report. *Am J Gastroenterol* **94**:2289–91.

Herrero F, Fernandez E, Gomez J, *et al.* 1995. Thallium poisoning presenting with abdominal colic, paresthesia, and irritability. *J Toxicol Clin Toxicol* **33**:261–4.

Hoffman RS. 2000. Thallium poisoning during pregnancy: a case report and comprehensive literature review. *J Toxicol Clin Toxicol* **38**:767–75.

Howell JM. 1986. Alkaline ingestions. *Ann Emerg Med* **15**:820–5.

Ikegaya H, Iwase H, Hatanaka K, *et al.* 2000. Postmortem changes in cytochrome *c* oxidase activity in various organs of the rat and in human heart. *Forensic Sci Int* **108**:181–6.

Ikegaya H, Iwase H, Hatanaka K, *et al.* 2001. Diagnosis of cyanide intoxication by measurement of cytochrome *c* oxidase activity. *Toxicol Lett* **119**:117–23.

Jacobsen D, Akesson I, Shefter E. 1982. Urinary calcium oxalate monohydrate crystals in ethylene glycol poisoning. *Scand J Clin Lab Invest* **42**:231–4.

Jacobsen D, Ostby N, Bredesen JE. 1982. Studies on ethylene glycol poisoning. *Acta Med Scand* **212**:11–15.

Kamijo Y, Soma K, Asari Y, *et al.* 1998. Survival after massive arsenic poisoning self-treated by high fluid intake. *J Toxicol Clin Toxicol* **36**:27–9.

Karhunen PJ, Lukkari I, Vuori E. 1991. High cyanide level in a homicide victim burned after death: evidence of post-mortem diffusion. *Forensic Sci Int* **49**:179–83.

Kazsuba A, Vitez A, Gall J, *et al.* 2000. Gastric hyalinization as a possible consequence of corrosive injury. *Endoscopy* **32**:356–8.

Lander H, Hodge PR, Crisp CS. 1965. Arsenic in hair and nails: its significance in acute arsenical poisoning. *J Forensic Med* **12**:52–67.

Lokan RJ, James RA, Dymock RB. 1987. Apparent post-mortem production of high levels of cyanide in blood. *J Forensic Sci Soc* **27**:253–9.

Makela JT, Laitinen S. 1997. Corrosive oesophageal injury. A study of four cases. *Ann Chir Gynaecol* **86**:360–3.

Marquet P, Clement S, Lotfi H, *et al.* 1996. Analytical findings in a suicide involving sodium azide. *J Anal Toxicol* **20**:134–8.

McCallum R. 1989. The industrial toxicology of antimony. The Ernestine Henry lecture 1987. *J R Coll Physicians Lond* **23**:28–32.

Meggs WJ, Hoffman RS, Shih RD, *et al.* 1994. Thallium poisoning from maliciously contaminated food. *J Toxicol Clin Toxicol* **32**:723–30.

Metter D, Vock R. 1984. [Structure of the hair in thallium poisoning.] *Z Rechtsmed* **91**:201–14.

Mollhoff G, Schmidt G, Bosche J. 1979. ['Thallium poisoning' – neurologic and forensic aspects.] *Arch Kriminol* **163**:1–13.

Moore D, House I, Dixon A. 1993. Thallium poisoning. Diagnosis may be elusive but alopecia is the clue. *Br Med J* **306**:1527–9.

Naon H, Shaul D, Mahnovski V, *et al.* 1996. Fibroproliferative disorder of the antrum after an alkali ingestion. *Am J Gastroenterol* **91**:383–4.

Neil MW. 1957. Immune to cyanide? *J Forensic Med* **4**:17–19.

Nolte KB, Dasgupta A. 1996. Prevention of occupational cyanide exposure in autopsy prosectors. *J Forensic Sci* **41**:146–7.

Pense SC, Wood WJ, Stempel TK, *et al.* 1988. Tracheoesophageal fistula secondary to muriatic acid ingestion. *Burns Incl Therm Inj* **14**:35–8.

Peterson CD, Collins AJ, Himes JM, *et al.* 1981. Ethylene glycol poisoning: pharmacokinetics during therapy with ethanol and hemodialysis. *N Engl J Med* **304**:21–3.

Poklis A, Saady JJ. 1990. Arsenic poisoning: acute or chronic? Suicide or murder? *Am J Forensic Med Pathol* **11**:226–32.

Questel F, Dugarin J, Dally S. 1996. Thallium-contaminated heroin. *Ann Intern Med* **124**:616.

Raekallio J, Jääskeläinen AJ, Mäkinen PL. 1967. The simple demonstration of calcium oxalate crystals in kidneys of victims of ethylene glycol poisoning. *J Forensic Sci* **12**:238–40.

Rehling CJ. 1967. Poison residues in human tissues. *Prog Chem Toxicol* **3**:363–86.

Saha JC, Dikshit AK, Bandyopadhyay MK, *et al.* 1999. A review of arsenic poisoning and its effects on human health. *Crit Rev Environ Sci Technol* **29**:281–313.

Shabalina LP, Spiridonova VS. 1979. Thallium as an industrial poison (review of literature). *J Hyg Epidemiol Microbiol Immunol* **23**:247–55.

Sheehan HL. 1940. Subendocardial haemorrhages in shock. *Lancet* **1**:831–2.

Stajduhar-Caric Z. 1968. Acute phenol poisoning. Singular findings in a lethal case. *J Forensic Med* **15**:41–2.

Stoeppler M, Vahter M. 1994. Arsenic. In: Herber RFM, Stoeppler M (eds), *Techniques and instrumentation in analytical chemistry*, vol. 15, *Trace element analysis in biological specimens*. Elsevier, Amsterdam, London, New York, Tokyo, pp. 291–320.

Tromme I, Van Neste D, Dobbelaere F, *et al.* 1998. Skin signs in the diagnosis of thallium poisoning. *Br J Dermatol* **138**:321–5.

Villanueva E, Hernandez-Cueto C, Lachica E, *et al.* 1990. Poisoning by thallium. A study of five cases. *Drug Saf* **5**:384–9.

Walton EW. 1978. An epidemic of antifreeze poisoning. *Med Sci Law* **18**:231–7.

Weinig E, Schmidt G. 1966. [On the distribution of thallium in the organism in fatal thallium poisoning.] *Arch Toxikol* **21**:199–215.

Westveer AE, Trestrail JH, Pinizzotto AJ. 1996. Homicidal poisoning in the United States: an analysis of the uniform crime reports from 1980 through 1989. *Am J Forensic Med Pathol* **17**:282–8.

Winek CL, Fusia E, Collom WD, *et al.* 1978. Cyanide poisoning as a mode of suicide. *Forensic Sci* **11**:51–5.

Winek CL, Bricker JD, Fochtman FW. 1980. Lithium intoxication. A case study. *Forensic Sci Int* **15**:227–31.

Winek CL, Esposito FM, Cinicola DP. 1990. The stability of several compounds in formalin fixed tissues and formalin-blood solutions. *Forensic Sci Int* **44**:159–68.

Winek CL, Rozin L, Wahba WW, *et al.* 1995. Ingestion of lye. *Forensic Sci Int* **73**:143–7.

Wormald PJ, Wilson DA. 1993. Battery acid burns of the upper gastro-intestinal tract. *Clin Otolaryngol* **18**:112–14.

Wu ML, Tsai WJ, Yang CC, *et al.* 1998. Concentrated cresol intoxication. *Vet Hum Toxicol* **40**:341–3.

Zamir O, Hod G, Lernau OZ, *et al.* 1985. Corrosive injury to the stomach due to acid ingestion. *Am Surg* **51**:170–2.

CHAPTER 34

Deaths from organic solvents

■ Solvent abuse 595
■ Substances used in abuse 596
■ The autopsy on fatal solvent-abuse victims 597
■ References and further reading 598

A wide range of organic compounds used in industry and in the home as solvents may cause injury or death, especially when used carelessly or for improper purposes. Other organic substances, often gaseous, can also be fatal when used or abused in the same circumstances.

SOLVENT ABUSE

This phenomenon arose, spontaneously and independently, in several parts of the world in the 1960s. It consists of deliberately inhaling a variety of substances, especially organic solvents, in order to achieve a distortion of consciousness. The effects vary from a state resembling alcoholic intoxication through euphoria (sometimes of an erotic nature) and distortion of perception, to actual hallucinations. The practitioners of this activity are predominantly, though by no means exclusively, male and most are between the ages of about 14 and 22 years, though again there are a few beyond either end of this age range. The majority of deaths occur in solitary users, some 70 per cent being reported by Anderson *et al.* (1985) as having happened when the young person was alone. Sniffing can certainly be a group activity, however, as is seen in schools and even the streets of some cities.

At first, the most common substance inhaled was a toluene-based adhesive and therefore the name 'glue-sniffing' soon became attached to the habit, though in fact many substances now used are not adhesives. The more accurate title of 'solvent abuse' has now been applied to this widespread and dangerous habit. Although (like the sexual asphyxias, with which there is some cross-relationship) the ratio of deaths to non-fatal escapades is relatively small, the large number of children and young people who indulge in the habit means that a considerable death toll occurs annually.

The usual method of abuse is for a quantity of the solvent or other chemical to be placed in a plastic bag – which may be any kind from a large black polythene garbage sack to a used potato-crisp packet. The open end is then placed against the nose and mouth, and the air inside re-breathed. The warmth of the hand holding the bag, as well as the warm breath, encourages vaporization of the solvent, which enters the tidal air stream and is absorbed through the pulmonary membranes into the blood. When the substance is not a viscous glue, it may be placed on a handkerchief or rag, and inhaled direct as a pad over the nose and mouth.

When gaseous substances are abused, they may be introduced directly into the mouth or nose. A common variety is butane or propane, taken from either a large cylinder used for camping, heating and cooking, or from the small ampoule cylinders sold for refilling gas cigarette lighters. Others are used directly from pressurized aerosol cans, including pain-relieving sprays used for the relief of muscular conditions. Yet again, sniffing may be attained by merely inhaling directly from the neck of a container. Tins, jerry-cans and the petrol-fillers of motor vehicles may be used in this way. Some fire extinguishers contain an organic bromine derivative which is also abused for inhalation. This has a particular propensity for sensitizing the myocardium to catecholamines.

Dangers of solvent abuse

SUDDEN DEATH

The major cause of death appears to be sudden cardiac arrest, following an arrhythmia. Any of the solvents appears to have the ability to sensitize the myocardium to the action of catecholamines, such as noradrenaline. Any sudden 'flight-or-fight' stimulus, even some considerable time after sniffing solvents, has the ability to precipitate ventricular fibrillation and hence sudden death. A recent instance seen by the author (BK) involved a youth whose habit was to steal fire extinguishers from trains parked in sidings and inhale the bromine-based contents. Having been in trouble with the police before, he took fright on seeing a constable approaching, though in fact the officer was merely walking his dog. The boy ran from his group of friends and did not reappear, being found dead in some bushes some time later, though he had not 'sniffed' that day. The autopsy and toxicology were negative, apart from residual traces of organic substance in his blood.

In addition to this physiological mechanism, actual myocarditis can occasionally occur, as in a 15-year-old girl who dropped dead whilst inhaling from the petrol tanks of vehicles. At autopsy all that the author (BK) could find was a florid, diffuse interstitial myocarditis with minimal myofibril necrosis.

CHEMICAL EFFECTS

The mechanical means of obtaining concentrated solvent vapour is in itself dangerous. Persistent rebreathing can produce hypoxia and hypercapnia, which is additive to the toxic effects of the solvent itself. When large plastic bags are used, the abuser sometimes places his head within the bag and this risks the dangers of rapid death from plastic bag asphyxia (Chapter 14). It is difficult to know if some of these cases are at least partially instances of erotic hypoxia, the effects of both oxygen lack and the hallucinogenic effect of the solvent combining to give both sexual pleasure and a heightened risk of death.

Another potent risk is vomiting while the effects of the solvent incapacitate the victim sufficiently to prevent his reflexes from protecting or clearing his glottis or air passages. The usual warning must be given, however, about uncritically accepting the finding of gastric contents in the air passages at autopsy. This may well have been an agonal regurgitation in a victim who was dying of the common cardiac (ventricular fibrillation) effect of solvents on the myocardium. This has medico-legal relevance in that a fellow abuser or even some other bystander may be accused of failing to render assistance in clearing the air passages, when in fact this was the effect, rather than the cause of death. This is not to say that aspiration of vomit cannot be the cause of death, as in alcoholic intoxication, but the autopsy finding should be backed up with some circumstantial evidence that it occurred before the agonal period.

PHYSICAL DANGER

When gaseous substances are used, such as butane or propane from cylinders, or propellants from aerosol cans, a physical danger is added to the chemical effects. The gases are commonly directed or sprayed into the open mouth while inhalation is performed. The release of high-pressure gas causes rapid cooling by the 'refrigerator effect' and the impact of this on the palate, pharynx and larynx may cause a reflex cardiac arrest. The author (BK) has dealt with two such cases, one in a school yard, where the users of small propane refuellers virtually dropped dead in the presence of other youths. The mechanism of this phenomenon is disputed, but freezing of the sensitive pharyngolaryngeal area is surely a parallel to the sudden deaths seen when cold water enters the nasopharynx in sudden immersion deaths.

DANGER TO LONG-TERM USERS

Long-term users can suffer liver damage, typical of halogenated hydrocarbon exposure, with fatty change in the liver and sometimes associated renal damage.

SUBSTANCES USED IN ABUSE

Toluene

Toluene was one of the first substances recognized as giving rise to 'glue-sniffing', as it is the solvent for many adhesives, such as the proprietary glue 'Evostick'. It is an aromatic petrol hydrocarbon used widely in industry as a solvent and thinner for adhesives and paints. It resembles benzene in many of its properties, another substance that can be used in solvent inhalation. Toluene can cause acute and chronic intoxication in industrial situations. Exposure to relatively high concentrations in the air, between 10 000 and 30 000 p.p.m. can cause drunkenness, mental confusion and coma within a few minutes.

In non-fatal toluene abuse, blood levels of 0.3–7.0 mg/l were measured by Bonnichsen *et al.* (1966), with urine excretion of up to 5 mg/l. Those with blood concentrations of 1.0–2.5 mg/l showed some sign of intoxication, while half of those between 2.5 and 10 mg/l were admitted to hospital with marked symptoms. Those who were unconscious or dead had blood levels in excess of 19 mg/l. Nomiyama and Nomiyama (1979) found blood levels of between 50 and 80 mg/l in three fatalities. Baselt (2002) quotes a fatal range of between 10 and 48 mg/l with an average of 22 mg/ml for fatalities, the average in the lung being 12 mg/l and in the brain 47 mg/l. Brain damage has been reported in long-term abusers, with electroencephalographic changes, encephalopathy and occasional cerebral atrophy.

Petrol (gasoline), xylene and benzene

These closely resemble toluene in behaviour, though benzene is said to be more potent in causing intoxication. As with toluene and most other solvents, they can burn the skin if left in contact with it for some time. The odour threshold for benzene has been reported as 1–5 p.p.m. in air. In chronic exposure, benzene can cause bone-marrow depression and fatal aplastic anaemia, but this is not a feature of the shorter exposures in solvent abuse.

Methylene chloride and ethylene dizchloride

These are found in paint strippers and, as these solvents, in many other products. One used for abuse is the solvent for typewriter correction fluids, such as Tippex.

Butane and propane

These have been mentioned earlier, being used direct from pressurized containers. They are light hydrocarbons, being the upper fractions of oil distillation.

Fluorocarbons

These are the substances now blamed for having the worst effect on the ozone layers of the upper atmosphere. They are widely used as refrigerants and propellants in aerosol containers. It is in this latter role that they are widely abused for euphoric and hallucinatory effects. The various compounds are numbered, rather than named, two fluorine atoms being incorporated into molecules that otherwise resemble carbon tetrachloride. Those used for propellants are FC-11 and FC-12.

A similar compound incorporating bromine instead of fluorine is used as a fire extinguisher, as is carbon tetrachloride itself. As long ago as 1970, Bass published a series of 110 sudden deaths, including many from fluorocarbons. As mentioned earlier, these substances sensitize the myocardium to circulating catecholamines. The range of blood concentrations in fatal cases in respect of FC-12 is 0.6–12 mg/l, with a mean of 3.0 mg/l and for FC-11, 1.2–32 mg/l, with a mean of 12 mg/l. In the lung, FC-12 may be present at levels of 0.9–134 mg/kg, with a mean of 33 mg/kg; for FC-11 the corresponding figures are 5.8–94 mg/kg, with a mean of 43 mg/kg.

Carbon tetrachloride

Used as a degreaser, dry-cleaner and fire extinguisher, carbon tetrachloride is easily available from retail outlets as a spot remover. It is quite toxic, though it (unwisely) used to be used in human pharmacy to treat intestinal worms. As little as 3–5 ml can be fatal. Much of its toxicity is industrial, but it has also been used in suicide and solvent abuse. Chronic exposure can cause liver and renal damage, worsened by the simultaneous use of alcohol. Addiction can occur, even from initial exposure during industrial use.

As with most halogenated hydrocarbons, the liver suffers from exposure to tetrachloride. There is a centrilobular necrosis, usually preceded by fatty change if the poisoning is low level and lasting. In fatal cases there may well be virtually total acute yellow atrophy. The liver damage is markedly worsened by a high alcohol intake. Renal changes consist of a tubular necrosis and diffuse fatty degeneration in the cortex.

Post-mortem blood levels vary greatly, but a fatal case has been recorded with a concentration of 260 mg/l. Korenke and Pribilla (1969) found tissue levels at autopsy a week after inhalation as being 142 mg/kg in the liver and 39 mg/kg in the lung.

Other halogenated hydrocarbons

Other halogenated hydrocarbons have similar effects to carbon tetrachloride, but the level of toxicity varies with the number of chlorine atoms in the molecule. The ascending order of danger is methyl chloride, trichlormethane, chloroform and carbon tetrachloride. Others include trichlorethylene, used widely in both anaesthesia (Trilene) and in dry-cleaning. Several other similar compounds have been the cause of fatalities in industry as well as in solvent abuse. The pathological features are similar in most of them, the effects being on the central nervous system in acute massive overdose, whether by ingestion or inhalation, and upon the liver and kidneys in more chronic intoxications.

THE AUTOPSY ON FATAL SOLVENT-ABUSE VICTIMS

The scene should be visited if possible, though it is rare to find the body there, as strenuous resuscitation and removal by ambulance is always a prime consideration in children and young persons found apparently dead. The paraphernalia of sniffing may be seen in its original state.

At autopsy, the clothing should be examined, even though it has almost always been removed for resuscitation attempts. Any soiling with adhesive or solvent stains should be detected and, though volatile, should be kept for forensic science investigation. The best means is to pack them as soon as possible in a nylon bag to retain any vapour; these solvents will penetrate a polyvinylchloride bag and be lost.

Externally, the face should be inspected for any signs of chronic or recent solvent abuse. The act of holding a bag against the lower face, often contaminating it with glue or solvent, may leave skin lesions. These can be red erythematous pimples or actual excoriations from the irritant action of the solvent. The lesions may become infected or scratched, becoming crusted. The rest of the autopsy may be unrewarding from the point of view of gross pathology. Rarely, there may be a fatty liver from long-standing damage from the solvent, especially the halogenated hydrocarbons. Full samples for histological examination should be taken to determine the state of the myocardium, liver and brain. More immediate is the need to obtain samples for analysis.

Vigorous resuscitation, including mouth-to-mouth and chest compression, together with mechanical ventilation and oxygen administration, will obviously blow off much of the solvent in the lungs. If it has already been absorbed, however, some will come back into alveolar air from the lung tissue and the sensitivity of the head-space gas chromatography used for its detection is often sufficient to identify the substance, which can also be detected in blood and tissues.

Blood samples should be taken in the usual manner, including a tube with fluoride for alcohol, which is sometimes also present. A lung should be taken intact for the laboratory. It was formerly the practice to aspirate air from the trachea before removal of viscera, using a 50 ml syringe, which was then sealed and sent to the laboratory. This is impracticable, and the best method is to open the pleural cavity and pass a string ligature round the hilum of a lung, pulling it tight to occlude the main bronchus. This is then drawn tight and the hilum transected. In actual fact, little is lost if the lung is cut off without this procedure, as long as it is put straight into a nylon bag without delay and without squeezing out the contained air. The nylon bags are those that arson investigators use to save material for fire-accelerant examination, as they are impervious to organic solvents.

A urine sample, and unfixed tissue from liver and brain should also be supplied to the toxicology laboratory, though local discussion should be made to determine the requirements of each laboratory.

REFERENCES AND FURTHER READING

al-Alousi LM. 1989. Pathology of volatile substance abuse: a case report and a literature review. *Med Sci Law* **29**:189–208.

Anderson HR, Macnair RS, Ramsey JD. 1985. Deaths from abuse of volatile substances: a national epidemiological study. *Br Med J Clin Res Ed* **290**:304–7.

Backer RC, Pisano RV. 1978. Gas chromatography-mass spectrometry of fluorocarbons 11 and 12 in biologic specimens. *Clin Toxicol* **12**:69–75.

Barrowcliff DF. 1978. Chronic carbon monoxide poisoning caused by methylene chloride paintstripper. *Med Sci Law* **18**:238.

Baselt RC. 2002. *Disposition of toxic drugs and chemicals in man*, 6th edn. Biomedical Publications, Foster City, CA.

Bass M. 1970. Sudden sniffing death. *JAMA* **212**:2075–9.

Bonnichsen R, Maehly AC, Moeller M. 1966. Poisoning by volatile compounds. I. Aromatic hydrocarbons. *J Forensic Sci* **11**:186–204.

Chao TC, Lo DS, Koh J, *et al.* 1993. Glue sniffing deaths in Singapore – volatile aromatic hydrocarbons in post-mortem blood by headspace gas chromatography. *Med Sci Law* **33**:253–60.

Clark MA, Jones JW, Robinson JJ, *et al.* 1985. Multiple deaths resulting from shipboard exposure to trichlorotrifluoroethane. *J Forensic Sci* **30**:1256–9.

Claydon SM. 1988. Myocardial degeneration in chronic solvent abuse. *Med Sci Law* **28**:217–18.

Editorial. 1982. Solvent abuse. *Lancet* **2**:1139–40.

Escobar A, Aruffo C. 1980. Chronic thinner intoxication: clinico-pathologic report of a human case. *J Neurol Neurosurg Psychiatry* **43**:986–94.

Fuke C, Berry CL, Pounder DJ. 1996. Postmortem diffusion of ingested and aspirated paint thinner. *Forensic Sci Int* **78**:199–207.

Garriott J, Petty CS. 1980. Death from inhalant abuse: Toxicological and pathological evaluation of 34 cases. *Clin Toxicol* **16**:305–15.

George B. 1960. Kerosene poisoning. *Med Sci Law* **1**:14–52.

Harada K, Ichiyama T, Ikeda H, *et al.* 1999. A fatal case of oral ingestion of benzine. *Am J Forensic Med Pathol* **20**:84–9.

Harrer G, Kisser W, Pilz P, *et al.* 1973. [Three cases of trichloroethylene resp. Carbon tetrachloride ‘sniffing’ with fatal outcome.] *Nervenarzt* **44**:645–7.

Heath MJ. 1986. Solvent abuse using bromochlorodifluoromethane from a fire extinguisher. *Med Sci Law* **26**:33–4.

Inoue H, Iwasa M, Maeno Y, *et al.* 1996. Detection of toluene in an adipoceratous body. *Forensic Sci Int* **78**:119–24.

Isenschmid DS, Cassin BJ, Hepler BR, *et al.* 1998. Tetrachloroethylene intoxication in an autoerotic fatality. *J Forensic Sci* **43**:231–4.

Kamijo Y, Soma K, Hasegawa I, *et al.* 1998. Fatal bilateral adrenal hemorrhage following acute toluene poisoning: a case report. *J Toxicol Clin Toxicol* **36**:365–8.

Kim NY, Park SW. 2000. The comparison of toluene determination between headspace-solid phase microextraction and headspace methods in glue-sniffer's blood and urine samples. *J Forensic Sci* **45**:702–7.

Kimura K, Nagata T, Hara K, *et al.* 1988. Gasoline and kerosene components in blood – a forensic analysis. *Hum Toxicol* **7**:299–305.

King MD, Day RE, Oliver JS, *et al.* 1981. Solvent encephalopathy. *Br Med J Clin Res Ed* **283**:663–5.

Korenke HD, Pribilla O. 1969. [Suicide by single inhalation of carbon tetrachloride (CCl_4), with resulting leukoencephalopathy.] *Arch Toxikol* **25**:109–26.

Kringsholm B. 1980. Sniffing-associated deaths in Denmark. *Forensic Sci Int* **15**:215–25.

Levine B, Fierro MF, Goza SW, *et al.* 1981. A tetrachloroethylene fatality. *J Forensic Sci* **26**:206–9.

Morinaga M, Kashimura S, Hara K, *et al.* 1996. The utility of volatile hydrocarbon analysis in cases of carbon monoxide poisoning. *Int J Legal Med* **109**:75–9.

Nomiyama K, Nomiyama H. 1979. [Health effects of trichloroethylene in human subjects (author's transl).] *Sangyo Igaku* **21**:311–34.

Park SW, Kim N, Yang Y, *et al.* 1998. Toluene distribution of glue sniffers' biological fluid samples in Korea. *J Forensic Sci* **43**:888–90.

Poklis A. 1975. Determination of fluorocarbon 11 and fluorocarbon 12 in post-mortem tissues: a case report. *Forensic Sci* **5**:53–9.

Ramsey JD, Flanagan RJ. 1982. The role of the laboratory in the investigation of solvent abuse. *Hum Toxicol* **1**:299–311.

Reed BJ, May PA. 1984. Inhalant abuse and juvenile delinquency: a control study in Albuquerque, New Mexico. *Int J Addict* **19**:789–803.

Ridgway P, Nixon TE, Leach JP. 2003. Occupational exposure to organic solvents and long-term nervous system damage detectable by brain imaging, neurophysiology or histopathology. *Food Chem Toxicol* **41**:153–87.

Sourindrhin I. 1985. Solvent misuse [editorial]. *Br Med J Clin Res Ed* **290**:94–5.

Standefer JC. 1975. Death associated with fluorocarbon inhalation: report of a case. *J Forensic Sci* **20**:548–51.

Suwaki H. 1983. A follow-up study of adolescent glue-sniffers in Japan. *Br J Addict* **78**:409–13.

Tranthim-Fryer DJ, Hansson RC, Norman KW. 2001. Headspace/solid-phase microextraction/gas chromatography-mass spectrometry: a screening technique for the recovery and identification of volatile organic compounds (VOC's) in postmortem blood and viscera samples. *J Forensic Sci* **46**:934–46.

Trochimowicz HJ, Azar A, Terrill JB, *et al.* 1974. Blood levels of fluorocarbon related to cardiac sensitization: II. *Am Ind Hyg Assoc J* **35**:632–9.

Wang CC, Irons SV. 1961. Acute gasoline intoxication. *Arch Environ Health* **2**:715–20.

Watson JM. 1977. 'Glue-sniffing' in profile. *Practitioner* **218**:255–9.

Watson JM. 1979. Morbidity and mortality statistics on solvent abuse. *Med Sci Law* **19**:246–52.

Winek CL, Collom WD. 1971. Benzene and toluene fatalities. *J Occup Med* **13**:259–61.

Winek CL, Wahba WW, Huston R, *et al.* 1997. Fatal inhalation of 1,1,1-trichloroethane. *Forensic Sci Int* **87**:161–5.

APPENDIX 1

Technical methods

Staining techniques 600
Diatom technique in suspected drowning 602
Useful conversion tables 603
Samples and controls needed for DNA profiling 608

STAINING TECHNIQUES

Periodic acid–Schiff (PAS) for myocardium

In established early infarcts of probably at least 28–24 hours' duration, damaged myofibres stain a pale purple-blue with PAS, compared with the pinker colour of healthy fibres. This stainable material is diastase resistant and is probably a mucoprotein.

(a) Bring sections to water.
(b) Immerse in 1 per cent periodic acid for 5 minutes.
(c) Wash in water for 5 minutes.
(d) Rinse in distilled water.
(e) Immerse in Schiff's reagent for 10 minutes.
(f) Wash in water for 10 minutes.
(g) Stain in haematoxylin for 30 seconds.
(h) Blue in tap water and mount in usual way.

Malic acid dehydrogenase (MDH)

Enzyme histochemistry is the most reliable method of detecting early myocardial infarction. Succinic, lactic, malic, β-hydroxybutyric dehydrogenases and cytochrome oxidase are among those used by many workers and each has its advocates. The author (BK) prefers malate dehydrogenase, though β-hydroxybutyrate is also recommended. Fixed tissue cannot be used, as frozen cryostat sections are required. The method for malate dehydrogenase is given, though the technique for other enzymes is similar, using different substrates; details are found in standard histochemistry texts.

MALATE DEHYDROGENASE METHOD FOR MYOCARDIUM

Normal myocardium stains dark blue–black, mostly concentrated on intracellular mitochondria, with some staining of the cytoplasm. Fibres damaged by anoxia or ischaemia progressively lose their staining properties and may be completely devoid of colour when totally infarcted. There is almost always perivascular survival of a few layers of cells and the subendocardial and subepicardial layers usually also survive: this acts as a built-in control for the staining reaction.

Stock solutions

Buffer

0.2 M for pH 7.0 (store at 0–4°C).

Solution A: 2.4 g trometamol (TRIS) in 100 ml distilled water	25 ml
Solution B: 0.1 M hydrochloric acid	45 ml
(Use ANALAR concentrated HC1 diluted 1/12 for molar, then dilute further 10 times for 0.1 M)	
Distilled water	30 ml

Malic acid substrate

L-Malic acid 6.7 g in 50 ml distilled water.
Add 12 g TRIS to buffer to pH 7.0.
Store frozen.

Sodium cyanide

0.1 M solution: weigh out 0.25 g sodium cyanide in 100 ml distilled water.
Add 8–8.5 ml M hydrochloric acid to buffer at pH 7.2.
Store at 0–4°C.

Tetrazolium salt

Make up nitroblue tetrazolium at 1 mg/ml strength in distilled water. Keep frozen.

Nicotinamide adenine dinucleotide (NAD) – keep solid desiccated in freezer.

Method

Cut cryostat sections from myocardium frozen either in liquid nitrogen or in isopentane/solid carbon dioxide slush.

(a) Place slides carrying sections in 37°C incubator.

(b) Make up incubating medium as follows:

Sodium cyanide solution (0.1 M at pH 7.2)	0.1 ml
NAD coenzyme	6 mg
Nitroblue tetrazolium solution	0.25 ml
TRIS buffer pH 7.0	0.55 ml
Malic acid substrate (1 M at pH 7.0)	0.1 ml

(Check pH and adjust with a small amount of TRIS)

(c) Warm the incubating medium to 37°C and place sufficient on each section with a pipette, covering the section completely. Leave for 30 minutes, mount in glycerine jelly or an aqueous mountant. A negative control can be set up as required by omitting the malate substrate from the incubating medium.

Haematoxylin–eosin (H and E) autofluorescence

Routine formalin-fixed H and E sections can be examined under ultraviolet light for autofluorescence of those fibres that have anoxic or ischaemic damage. Fibres with early infarction show a shift of their secondary emission towards the yellow, away from the usual olive-green of healthy fibres. It has been said that only fibres that reveal increased eosinophilia in ordinary tungsten light will fluoresce in this way, but the authors and co-workers have noted yellow shift in fibres that cannot be detected in tungsten light as being eosinophilic.

Acridine-orange fluorescent stain for myocardium

Either formalin-fixed paraffin sections or unfixed frozen sections can be stained by this technique. The former must be brought to water first. After staining and mounting in water (no other mountant can be used, not even Fluormount), the sections are examined immediately under ultraviolet light. They can be photographed if required. If the cover-slip is removed and the sections allowed to dry, however, they can be retained indefinitely. To review, they need only be wetted and a cover-slip replaced. Normal myocardium is a golden brown or sometimes yellowish brown depending on the optical system used. Damaged fibres show a colour shift to green when viewed under ultraviolet light.

Method

A stock solution of 1 per cent acridine-orange in water is kept refrigerated. This lasts for several months. For use, it is diluted 1/10 with distilled water.

(a) Bring sections to water.

(b) Agitate in the 0.1 per cent stain for approximately 5 seconds, though the time is not critical.

(c) Briefly wash, wipe off the excess water from around the section.

(d) Apply a cover-slip.

They are then ready for immediate examination. The pH of the washing water is not critical and tap water can be used if the pH range is between 6.0 and 7.0.

Martius scarlet blue (MSB) stain for fibrin

This trichrome stain is useful for examining thrombi and emboli and for seeking fibrin in disseminated intravascular coagulation.

(a) Bring sections to water.

(b) Stain with celestine blue for 3 minutes and Mayer's haematoxylin for 1 minute.

(c) Blue in water.

(d) Rinse in 95 per cent ethanol and then stain for 2 minutes with 0.5 per cent Martius yellow in 95 per cent ethanol containing 2 per cent phosphotungstic acid.

(e) Rinse in water and stain with 1 per cent brilliant crystal scarlet in 2.5 per cent glacial acetic acid for 10 minutes.

(f) Rinse in water and treat with 1 per cent aqueous phosphotungstic acid for 5 minutes.

(g) Rinse in water and stain with 0.5 per cent soluble blue (Acid Blue 93) in 10 per cent aqueous acetic acid for 10 minutes.

(h) Rinse and mount.

This method stains fibrin red, erythrocytes yellow and collagen blue.

Lendrum's stain (phloxine-tartrazine)

Though this stain has a number of applications, such as the demonstration of viral inclusion bodies, its major forensic use is in the detection of amniotic fluid embolism deaths associated with pregnancy. The fetal squames are rendered visible much more readily than in haematoxylin–eosin stain.

(a) Stain with Mayer's haematoxylin for 6 minutes and rinse in running water for 5 minutes.
(b) Stain with 0.5 per cent phloxine in 0.5 per cent calcium chloride solution for 20 minutes in a Coplin jar.
(c) Rinse in water and drain almost dry.
(d) Flood with a saturated solution of tartrazine in cellusolve ($C_4H_{10}O_2$; ethylene glycol monomethyl ether), which differentiates counterstain. This takes about 7 minutes and should be monitored microscopically so that red cells appear red–brown.
(e) Rinse by two changes in 95 per cent ethanol and two changes in absolute ethanol.
(f) Transfer straight to xylol and mount.

The keratin of amniotic squames is stained red; nuclei blue and cytoplasm yellow.

The 'WHO' method

The 'WHO' method is similar for demonstrating keratin and mucin-like substances in amniotic fluid embolism.

(a) Bring sections to water.
(b) Stain with celestine blue/Mayer's sequence, rinse.
(c) Differentiate quickly in 0.5 per cent acid–alcohol, then blue.
(d) Stain with erythrosin (1 per cent) for 5 minutes and rinse.
(e) Differentiate quickly in 95 per cent alcohol, rinse.
(f) Stain, with alcian green solution for 5 minutes.
(g) Rinse – the section should now be slightly pink.
(h) Rinse and treat with 95 per cent alcohol.
(i) Stain with ethanolic saffron for 5 minutes.
(j) Mount without going through water.

The alcian green is made by mixing 50 ml of 1 per cent aqueous solution of alcian green with 50 ml of 1 per cent acetic acid then adding 20 mg of thymol and filtering. Ethanolic saffron is made by cutting up strands of saffron and warming in 95 per cent alcohol until no more will elute. This should be a deep yellow solution.

DIATOM TECHNIQUE IN SUSPECTED DROWNING

There are several versions of this method, each with its own advocates. More recent methods have used suction of the filtrate through a millipore membrane to trap the diatoms, but the author (BK) has encountered early clogging of the membrane by debris and prefers the older centrifugation. References to the published methods will be found in Chapter 16.

Take a sample of water from the site where the body was recovered or where the immersion was first thought to have occurred. Several litres should be collected in a clean container and allowed to settle overnight. The top 90 per cent volume is then carefully decanted and the remainder centrifuged. The deposit should be examined under phase-contrast or dark-ground microscopy to see if any diatoms are present. If not, it is pointless to examine the autopsy tissues. If present, they should be retained for species comparison.

If the water contains diatoms, then the material retained at autopsy can be processed. The following procedure should have been followed:

(a) The sternum or a long bone such as the femur, or both, should be removed from the body in the usual way. It should be taken to a part of the mortuary or a side room distant from the cadaver and well-washed externally in a clean bowl or sink. Tap water can be used, if it is checked at intervals to ensure that it is diatom-free. Some water supplies at irregular times contain a few diatoms, though these are rarely numerous enough to cause problems in interpretation.

When the exterior of the bone has been well washed to remove any contamination from the exterior of the body, a clean saw (not used in that autopsy) is employed to open the marrow cavity of femur or sternum. With a clean scalpel or gouge, several grams of marrow are removed into a glass universal (30 ml) container. Obtain red marrow if possible, as fatty yellow marrow causes problems during digestion.
(b) Whole kidney, lung, liver or brain is also washed in this way and 1 × 1 cm cubes (several from each organ) cut with a clean scalpel from the deeper tissue. The organs can then be returned to the autopsy area for the usual gross examination.
(c) Each container, holding its separate tissue, is placed in a fume cupboard and covered with five times the volume of concentrated nitric acid. If metal caps are used on the containers, they may need to be

shielded by polythene sheet to prevent dissolution. Digestion is allowed to proceed at room temperature for 1–2 days or, if the matter is urgent, they can be heated in a water-bath for a number of hours or overnight. When the solid cubes have dissolved into a slurry the containers are gently centrifuged, the supernatant acid poured off and replaced with distilled water. This process is repeated twice more to dilute the acid, knocking up the deposit, and recentrifuging, the final spin being as hard as possible to produce a small button of deposit. This may require transferring to a smaller conical centrifuge tube.

(d) A portion of this is transferred by pipette or loop to a slide and covered with a cover-slip to examination under phase-contrast or dark-ground illumination.

A prolonged search under medium power of the microscope is required before the sample is declared negative. The lung sample should be examined first, as if it is negative it is unlikely that other organs can be truly positive, for diatom-containing water can percolate passively down the air-passages even after death.

Opinion is divided about how many diatoms can be accepted as presumptive evidence of drowning when found in distant organs like brain, bone marrow and kidney. One solitary organism is certainly insufficient and the author (BK) requires five as a minimum under one coverslip before acceptance. The tissue examined may make a difference to the yield. Peabody strongly recommends brain as the choice and condemns femoral marrow, but bone marrow was the tissue used by the pioneers of the technique and seems to the authors to be satisfactory. The type of diatom found in the tissues should correspond to those found in the water sample, though it must be admitted that most diatoms seem to the non-botanical eye to be ovoid or boat-shaped. Expert advice from a suitable biologist should be sought if the issue is important.

USEFUL CONVERSION TABLES

TABLE A1.0 *Conversion factors*

To convert from	To	Multiply by
Centimetres	Inches	0.394
Inches	Centimetres	2.540
Grams	Pounds	0.002
Pounds	Grams	453.592

To convert Centigrade to Fahrenheit, multiply by 1.8 and add 32.
To convert Fahrenheit to Centigrade, subtract 32 and divide by 1.8.

TABLE A1.1 *Approximate weight of bodies*

Stones	Pounds	Kilograms
1	14	6.3
2	28	12.7
3	42	19.0
4	56	25.4
5	70	31.7
6	84	38.1
7	98	44.4
8	112	51.0
9	126	57.0
10	140	63.0
11	154	70.0
12	168	76.0
13	182	82.0
14	196	89.0
15	203	92.0
18	252	114.0

TABLE A1.2 *Height of bodies*

Feet and inches	Inches	Centimetres
4′10″	58″	145.0
4′11″	59″	147.5
5′	60″	150.0
5′1″	61″	152.5
5′2″	62″	155.0
5′3″	63″	157.5
5′4″	64″	160.0
5′5″	65″	162.5
5′6″	66″	165.0
5′7″	67″	167.5
5′8″	68″	170.0
5′9″	69″	172.5
5′10″	70″	175.0
5′11″	71″	177.5
6′	72″	180.0
6′1″	73″	182.5
6′2″	74″	185.0

TABLE A1.3 *Approximate breath–blood-alcohol equivalents*

µg/100 ml breath	mg/100 ml blood
10	23
20	45
30	68
40	91
50	114
60	137
70	160
80	182
90	205
100	228

1 µg/100 ml in breath = 2.28 mg/100 ml in blood.
1 mg/100 ml in blood = 0.43 µg/100 ml in breath.

Table A1.4 *Length conversion of wounds, etc.*

Inches	Centimetres
0.25	0.63
0.5	1.27
1.0	2.54
1.5	3.81
2.0	5.08
2.5	6.35
3.0	7.62
4.0	10.16
5.0	12.70

Table A1.5 *Normal ranges of biochemical blood analyses*

Sodium	133–147 mmol/l
Potassium	3.4–5.0 mmol/l
Urea	2.5–7.5 mmol/l
Calcium	2.2–2.6 mmol/l
Chloride	95–106 mmol/l
Bicarbonate	22–28 mmol/l
Glucose (fasting)	3.6–5.6 mmol/l
Lactate	0.8–2.0 mmol/l
Phosphate (inorganic)	0.8–1.4 mmol/l
Magnesium	0.6–1.0 mmol/l
Serum protein (total)	60–80 g/l
Albumin	35–48 g/l
Bilirubin	up to 17 μmol/l
Alkaline phosphatase	30–90 IU/l
Amylase	70–300 IU/l
Creatinine	50–100 μmol/l
Creatine kinase (CPK)	Up to 240 IU/l

Table A1.6 *Body weight and length 15–30 years.* In this table, three weights (without clothing) are given for each body length: (1) weight for persons of average build = middle entry; (2) weight for lightly built persons = top entry; (3) weight for heavily built persons = bottom entry*

Women					Men				
	Weight (kg)					Weight (kg)			
Body length (cm)	15 yrs	20 yrs	25 yrs	30 yrs	Body length (cm)	15 yrs	20 yrs	25 yrs	30 yrs
142.5	40.8	43.0	44.0	45.3	150	41.7	45.8	47.6	49.4
	45.3	47.6	49.0	50.3		46.2	50.7	53.0	54.8
	51.2	53.0	55.3	56.6		51.6	57.1	59.3	61.6
145	41.2	43.5	44.8	46.2	152.5	42.6	46.7	48.5	50.3
	45.8	48.5	49.8	51.2		47.1	51.6	53.9	55.7
	51.6	53.9	56.2	57.5		53.0	58.0	60.7	62.5
147.5	41.7	44.4	45.8	47.1	155	43.5	47.6	49.4	51.2
	46.2	49.4	50.7	52.1		48.5	53.0	54.8	56.6
	52.1	55.7	57.1	58.4		54.4	59.3	61.6	63.4
150	42.6	45.3	46.7	47.6	157.5	44.9	48.9	50.7	52.1
	47.1	50.3	51.6	53.0		49.8	54.4	56.2	58.0
	53.0	56.6	58.0	59.8		56.2	61.2	63.0	65.2
152.5	43.5	46.7	47.1	48.5	160	46.2	50.3	52.1	53.5
	48.5	51.6	52.6	53.9		51.2	55.7	58.0	59.3
	54.4	58.0	59.3	60.7		57.5	62.5	65.2	66.6
155	44.8	47.6	48.5	49.8	162.5	47.6	51.6	53.9	55.3
	49.8	53.0	53.9	55.3		53.0	57.5	59.8	61.2
	55.3	59.8	60.7	62.1		59.3	64.8	67.0	68.9
157.5	46.2	48.9	50.3	51.2	165	49.4	53.5	55.7	56.6
	51.2	54.4	55.7	56.6		54.8	59.3	61.6	63.0
	57.5	61.2	62.5	63.9		61.6	66.6	69.3	70.7

(*continued*)

TABLE A1.6 *Body weight and length 15–30 years.* (continued)*

Women					Men				
	Weight (kg)					Weight (kg)			
Body length (cm)	**15 yrs**	**20 yrs**	**25 yrs**	**30 yrs†**	**Body length (cm)**	**15 yrs**	**20 yrs**	**25 yrs**	**30 yrs†**
160	47.1	50.3	51.2	52.5	167.5	51.2	55.3	57.1	58.4
	52.5	55.9	57.1	58.4		56.6	61.2	63.4	64.8
	59.3	62.5	6.43	65.7		63.4	68.9	71.1	72.9
162.5	48.9	51.2	52.5	53.9	170	52.5	56.6	58.9	59.8
	54.4	57.1	58.4	59.8		58.4	63.0	56.2	66.6
	61.2	64.3	65.7	67.5		65.7	70.7	73.4	74.7
165	50.7	53.0	54.4	55.7	172.5	54.4	58.4	60.2	61.6
	56.2	58.9	60.2	61.6		60.2	64.8	67.0	68.4
	63.4	66.1	67.5	69.3		67.5	73.0	75.2	77.0
167.5	52.1	54.9	55.7	57.1	175	55.7	59.8	62.1	63.9
	58.0	60.7	62.1	63.4		62.1	66.6	68.9	70.7
	65.2	68.4	69.8	71.6		69.8	74.7	77.5	79.3
170	53.9	56.2	57.5	58.9	177.5	58.0	61.6	63.9	65.7
	59.8	62.5	63.9	65.2		64.3	68.4	71.1	72.9
	67.5	70.2	71.6	73.4		72.0	77.0	79.7	82.0
172.5	55.3	57.5	59.3	60.2	180	59.8	63.9	66.1	68.0
	61.6	63.9	65.7	67.0		66.6	70.7	73.4	75.7
	69.3	72.0	73.8	75.7		74.7	79.3	82.4	85.2
175	57.1	59.3	60.7	61.6	183	62.1	65.7	68.4	70.7
	63.4	65.7	67.5	68.4		68.9	72.9	76.1	78.4
	71.6	73.8	75.7	77.0		77.5	82.0	85.6	87.9
178	59.3	60.7	62.1	63.4	185.5	63.9	68.0	71.1	72.9
	65.7	67.5	68.9	70.2		71.1	75.2	78.8	81.1
	73.9	76.1	77.5	78.9		79.7	84.3	88.3	91.1
180	61.2	63.0	63.4	64.8	188	66.1	69.8	73.0	75.7
	68.0	69.8	70.7	72.0		73.4	77.5	81.1	83.8
	76.1	78.4	79.7	81.1		82.4	87.0	91.1	94.2

* From Basel JR, Geigy AG. 1953. *Wissenschaftliche Tabellen*, p. 199. By permission.
† Fourth decade weight should be maintained throughout life.

TABLE A1.7 *Weight of fetus (g) in relation to age (month)**

Lunar month	Streeter	Scammon and Calkins
Second	1.1	3.5
Third	14.2	14.3
Fourth	108.0	86.8
Fifth	316.0	260.9
Sixth	630.0	551.6
Seventh	1045.0	971.4
Eighth	1680.0	1519.0
Ninth	2378.0	2196.1
Tenth	3405.0	2998.8

* From Potter EL. 1961. *Pathology of the fetus and infant*, 2nd edn. Chicago, Year Book Medical Publishers. By permission.

TABLE A1.8 *Length (cm) of fetus in relation to age (month)**

	Crown–rump length		Crown–heel length	
Lunar month	**Streeter**	**Scammon and Calkins**	**Dietrich**	**Scammon and Calkins**
Second	2.3	–	3.0	–
Third	7.4	5.1	9.8	7.0
Fourth	11.6	10.7	18.0	15.5
Fifth	26.4	15.5	25.0	22.7
Sixth	20.8	19.7	31.5	29.2
Seventh	24.7	23.6	37.1	35.0
Eighth	28.3	27.1	42.5	40.4
Ninth	32.1	30.5	47.0	45.4
Tenth	36.2	33.6	50.0	50.2

* From Potter EL. 1961. *Pathology of the fetus and infant*, 2nd edn. Chicago, Year Book Medical Publishers. By permission.

TABLE A1.9 *Weights of the organs at various ages (in grams)**

	Lungs		Brain		Heart		Kidneys		Liver		Spleen	
	Men	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
New born	51.7	50.9	353	347	19	20	24	24	124	125	8	6
0–3 mth	68.8	63.6	435	411	–	–	–	–	–	–	–	–
3–6 mth	94.1	93.3	600	534	–	–	–	–	–	–	–	–
6–9 mth	128.5	114.7	877	726	41	36	60	52	300	240	26	25
9–12 mth	142.4	142.1										
1–2 yr	170.3	175.3	971	894	54	48	72	65	400	390	35	34
2–3 yr	245.9	244.3	1076	1012	63	62	85	75	460	450	42	41
3–4 yr	304.7	265.5	1179	1076	73	71	93	84	510	500	48	47
4–5 yr	314.2	311.7	1290	1156	83	80	100	93	555	550	53	52
5–6 yr	360.6	319.9	1275	1206	95	90	106	102	595	595	58	57
6–7 yr	399.5	357.5	1313	1225	103	100	112	112	630	635	62	62
7–8 yr	365.4	404.4	1338	1265	110	113	120	123	665	685	64	67
8–9 yr	405.0	382.1	1294	1208	122	126	128	135	715	745	68	71
9–10 yr	376.4	358.4	1360	1226	132	140	138	148	770	810	73	77
10–11 yr	474.5	571.2	1378	1247	144	154	150	163	850	880	82	85
11–12 yr	465.6	535.0	1348	1259	157	168	164	180	950	950	91	93
12–13 yr	458.8	681.7	1383	1256	180	188	178	195	1050	1080	101	103
13–14 yr	504.5	602.3	1382	1243	202	207	196	210	1150	1180	111	112
14–15 yr	692.8	517.0	1356	1318	238	226	212	222	1240	1270	121	120
15–16 yr	691.7	708.8	1407	1271	258	238	229	230	1315	1330	135	127
16–17 yr	747.3	626.5	1419	1300	282	243	244	236	1380	1360	145	134
17–18 yr	776.9	649.5	1409	1254	300	247	260	240	1450	1380	152	140
18–19 yr	874.7	654.9	1426	1312	310	250	270	244	1510	1395	157	146
19–20 yr	1035.6	785.2	1430	1294	318	251	282	247	1580	1405	160	151
20–21 yr	953.0	792.8	–	–	322	252	290	248	1630	1415	162	155

* Boyd E. 1962. In: Altman and Dittmer (eds), *Growth, including reproduction and morphological development*. Biological Handbooks, Federation of American Societies for Experimental Biology, Washington, pp. 346–8.

TABLE A1.10 *Heart: men and women, weight in relation to body weight by sex**

Men				**Women**			
Body weight		**Heart weight (g)**		**Body weight**		**Heart weight (g)**	
Pounds	**Kilograms**	**Mean**	**Range**	**Pounds**	**Kilograms**	**Mean**	**Range**
				90	40	162	135–193
				95	43	171	143–204
				100	45	180	150–215
105	47	205	165–241	105	47	189	158–226
110	50	215	173–253	110	50	198	165–237
115	52	225	181–264	115	52	207	172–248
120	54	235	190–276	120	54	215	180–259
125	56	245	198–287	125	56	225	188–268
130	58	255	206–299	130	58	234	195–277
135	60	265	213–310	135	60	244	203–286
140	63	274	221–322	140	63	253	211–295
145	65	284	229–333	145	65	262	219–304
150	68	294	237–345	150	68	272	225–313

(*continued*)

TABLE A1.10 *Heart: men and women, weight in relation to body weight by sex* (continued)*

Men				Women			
Body weight		Heart weight (g)		Body weight		Heart weight (g)	
Pounds	Kilograms	Mean	Range	Pounds	Kilograms	Mean	Range
155	70	304	245–356	155	70	282	233–322
160	72	313	253–368	160	72	288	240–330
165	74	323	261–370	165	74	297	247–337
170	77	333	268–371	170	77	306	255–343
175	79	343	280–372	175	79	315	283–350
180	81	353	288–373	180	81	324	301–356
185	83	363	296–382	185	83	333	309–361
190	86	373	304–392	190	86	342	317–366
195	88	382	312–402	195	88	351	325–371
200	90	392	320–412				
Average weight of adult male heart: 294 g				Average weight of adult female heart: 250 g			

* Data from Smith HL. 1928. The relation of the weight of the heart to the weight of the body and of the weight of the heart to age. *Am Heart J* **4**:79–93.

NB There is considerable variation in published tables of heart weights: probably hypertensive hearts account for some of the upper ranges of 'normal'. However, the authoritative publication of Hangartner *et al.* gives considerably higher normal maximum weights. (Hangartner J, Marley N, Whitehead A, Thomas A, Davies M. 1985. The assessment of cardiac hypertrophy at autopsy. *Histopathology* **9**:1295–306.)

TABLE A1.11 *Heart: normal weight in relation to body length*

Body length (cm)	Heart weight (g)		Body length (cm)	Heart weight (g)		Body length (cm)	Heart weight (g)	
	Males	Females		Males	Females		Males	Females
	σ = ±40	σ = ±30		σ = ±40	σ = ±30		σ = ±40	σ = ±30
135	254	219	157	296	258	179	338	297
136	256	220	158	298	259	180	340	299
137	258	222	159	300	261	181	342	300
138	260	224	160	302	263	182	344	302
139	262	226	161	304	265	183	346	304
140	264	227	162	306	267	184	348	306
141	266	229	163	308	268	185	349	307
142	268	231	164	310	270	186	351	309
143	270	233	165	311	272	187	353	311
144	272	235	166	313	274	188	355	313
145	273	236	167	315	275	189	357	315
146	275	238	168	317	277	190	359	316
147	277	240	169	319	279	191	361	318
148	279	242	170	321	281	192	363	320
149	281	243	171	323	283	193	365	322
150	283	245	172	325	284	194	367	323
151	285	247	173	327	286	195	368	325
152	287	249	174	329	288	196	370	327
153	289	251	175	330	290	197	372	329
154	291	252	176	332	291	198	374	331
155	292	254	177	334	293	199	376	332
156	294	256	178	336	295	200	378	334

σ = standard deviation.

From Zeek PM. 1942. Heart weight. I. The weight of the normal heart. *Arch Pathol* **34**:820–32.

Copyright 1942, American Medical Association.

NB But see also Hangartner *et al.* 1985 (details above).

Table A1.12 *Brain: adults**

Age (years)	Brain weight (g) Men Mean	Men Range	Women Mean	Women Range
17–19	1340	1170–1527	1242	1120–1420
20–29	1396	1158–1620	1234	1057–1565
30–39	1365	1075–1685	1233	1038–1440
40–49	1366	1069–1605	1240	995–1543
50–69	1375	1113–1665	1200	820–1477
60–69	1323	1018–1610	1178	920–1372
70–85	1279	1039–1485	1121	832–1370

* Data from Sunderman FW, Boerner F. 1949. *Normal values in clinical medicine.* W. B. Saunders Company, Philadelphia.

Table A1.13 *Brain: children**

Age	Body length (cm)	Brain weight (g)
Birth–3 days	49	335
3–7 days	49	358
1–3 wks	52	382
3–5 wks	52	413
5–7 wks	53	422
7–9 wks	55	489
3 mths	56	516
4 mths	59	540
5 mths	61	644
6 mths	62	660
7 mths	65	691
8 mths	65	714
9 mths	67	750
10 mths	69	809
11 mths	70	852
12 mths	73	925
14 mths	74	944
16 mths	77	1010
18 mths	78	1042
20 mths	79	1050
22 mths	82	1059
24 mths	84	1064
3 yrs	88	1141
4 yrs	99	1191
5 yrs	106	1237
6 yrs	109	1243
7 yrs	113	1263
8 yrs	119	1273
9 yrs	125	1275
10 yrs	130	1290
11 yrs	135	1320
12 yrs	139	1351

* Data from Sunderman FW, Boerner F. 1949. *Normal values in clinical medicine.* W. B. Saunders Company, Philadelphia.

SAMPLES AND CONTROLS NEEDED FOR DNA PROFILING

The essence of DNA profiling in forensic work is comparison between two samples. One is the 'test' sample from the living or dead subject (or objects with which he or she has been in contact, such as clothing) and the other the 'control', a somewhat inaccurate term for the samples obtained from a suspect, a group of suspects or a screening programme which may have hundreds or even thousands of items.

Since the beginning of the DNA analysis, significant advances have been made in methodology allowing forensic identification of individuals from minute amounts of DNA. Microgram amounts of relatively undegraded DNA is required for multi-locus typing of the restriction fragment length polymorphism (RFLP) analysis and hundreds of nanograms for single-locus typing. Polymorphic DNA sequences can be detected from the root of shed or freshly plucked single human hairs using the polymerase chain reaction (PCR). Mitochondrial DNA (mtDNA) has been detected even in a sample from a single hair shaft.

Blood

As red cells have no nuclear DNA, sufficient blood must be obtained to extract DNA from the much sparser leucocytes. At least 1 ml and preferably 5 ml are taken into an EDTA tube, which extracts metallic ions and not only prevents clotting, but inhibits enzymes in blood or micro-organisms, which may break down DNA during storage.

The blood sample, in a plastic (not glass) tube, should be frozen solid in a deep-freeze or the ice-making compartment of an ordinary refrigerator. As frozen blood is useless for some other forensic investigations, the DNA sample should be clearly marked as such. Other samples, for grouping and many other serological and toxicological tests, must be taken and stored at only 4°C.

If a blood sample is to be taken directly to the laboratory for DNA profiling, then freezing is not necessary, but wherever there is delay, either freezing or, at second best, effective cooling should be carried out.

Most of the laboratories are using nowadays FTA paper for the collection of blood samples. FTA paper is an absorbent cellulose-based paper that contains chemical substances to protect DNA molecules from nuclease degradation and preserve the paper from bacterial growth (Seah and Burgoyne 2000). As a result, DNA on FTA paper is stable at room temperature over a period of several years. The use of FTA paper simply involves adding a spot of blood to the paper and allowing the stain to dry.

Blood stains, as opposed to liquid blood, should either be sent intact on surfaces, kept as cool as possible before and during transit to the laboratory – or rubbed with a cotton-wool swab moistened with water. This swab is then air-dried without heat and frozen. Although not required for DNA testing, always provide the laboratory with control swabs for conventional blood grouping, including swabs rubbed on an area of the fabric away from obvious stains, as well as an unused swab.

Dried blood stains on hard surfaces can be scraped off with a scalpel into a small plastic container and sent as they are, kept as cool as possible.

Buccal cells

A disposable toothbrush may also be used for collecting buccal cells in a non-threatening manner. This method can be very helpful when samples need to be collected from children and this is the method actually used for most of paternity testing labs. After the buccal cells have been collected by gently rubbing a wet toothbrush across the buccal surface, the brush can be tapped onto the surface of FTA paper for sample storage and preservation.

Seminal and vaginal fluids

Swabs from vagina, rectum, mouth, etc., should be air-dried as quickly as possible, but not heated. They should then be stored in deep-freeze, unless sent straight to the forensic science laboratory. Liquid semen found in the vagina or elsewhere should be recovered with a fine pipette, placed in a small plain tube and frozen solid.

Seminal stains on small items of fabric or any other small objects may be frozen or kept as cool as possible during transit. Bulk clothing cannot be frozen and should merely be kept cool, transit to the laboratory being made as soon as possible, using strong sterilized paper bags, not impervious polythene, which encourages mould formation. Alternatively, damp swabs may be taken of suspect stains, which are then dried and frozen. As with blood, suspect dried seminal stains on hard surfaces may be scraped off and transmitted dry to the laboratory.

Mouth swabs may be taken either to seek semen in suspected oral intercourse (though positive recoveries are very rare) or to obtain buccal mucosa for DNA identification. In the latter case, the yield is small, so at least three to four swabs must be rubbed hard against the inside of the cheek – or the mouth lining scraped with an instrument and then smeared onto a swab. All such swabs should be frozen in the usual manner.

Hairs – head, pubic, etc.

Collection of hairs – head, pubic, eyebrow, axillary – from the victim or at the scene of crime is extremely important as a 2 cm piece of single hair is actually sufficient for mtDNA analysis. It is of paramount importance to avoid cross-contamination and to submit the hairs in sterilized paper bags, well identified.

Autopsy tissue samples

At least 0.5 g of tissue should be cut from the parenchyma of an organ and placed in a small plastic tube with no fixative or preservative. This should be frozen solid. Spleen is said to be one of the best organs for DNA recovery, though liver, muscle, kidney and brain may also be used. The fresher the tissue the better, and totally putrefied material was formerly of little use, as the nuclear chromatin was degraded. However, rapidly developing new techniques are allowing the recovery of recognizable DNA from old material.

For identification of skeletal remains using DNA technology, a well-preserved tooth is the tissue of election. In addition, long bones (femur) or ribs can also be used.

Other general remarks

The sensitivity of DNA technology is changing the strategy of sample collection. Very sensitive low-copy-number DNA methods are being developed allowing the detection of subnanogram amounts of DNA. For this reason, it is extremely important to avoid cross-contamination from the personnel collecting the samples.

DNA can even be obtained from personal objects, fingerprints, sweat and of course from saliva recovered from bite marks.

Finally, as for other type of forensic evidence, the chain of custody must be extremely well-preserved.

FURTHER READING

Badir B, Knight B. 1987. Fluorescence microscopy in the detection of early myocardial infarction. *Forensic Sci Int* **34**:99–102.

APPENDIX 2

Council of Europe Committee of Ministers Recommendation No. R (99) 3

- Recommendation No. R (99) 3 of the Committee of Ministers to Member States on the harmonisation of medico-legal autopsy rules 610
- Appendix to Recommendation No. R (99) 3 616

RECOMMENDATION NO. R (99) 3 OF THE COMMITTEE OF MINISTERS TO MEMBER STATES ON THE HARMONISATION OF MEDICO-LEGAL AUTOPSY RULES[1]

(*Adopted by the Committee of Ministers on 2 February 1999 at the 658th meeting of the Ministers' Deputies*)

The Committee of Ministers, under the terms of Article 15.*b* of the Statute of the Council of Europe,

Considering that the aim of the Council of Europe is to achieve a greater unity between its members;

Having regard to the principles laid down in the Convention for the Protection of Human Rights and Fundamental Freedoms and, in particular, the prohibition of torture or inhuman or degrading treatment or punishment, and the right to life;

Conscious that it is normal practice for autopsies to be carried out in all Council of Europe member States to establish the cause and manner of death for medico-legal or other reasons or to establish the identity of the deceased;

Considering the importance of compensation for victims and families in criminal and civil proceedings;

Underlining the need for investigation, description, photographic documentation and sampling during medico-legal autopsy to follow primarily medical and scientific principles and simultaneously consider legal requirements and procedures;

Conscious that the increasing mobility of the population throughout Europe and the world, as well as the increasing internationalisation of judicial proceedings, require the adoption of uniform guidelines on the way autopsies are to be carried out and on the way autopsy reports are to be established;

Considering the Council of Europe Agreement on the Transfer of Corpses (European Treaty Series No. 80) and having regard to the difficulties often experienced by the receiving country when a dead body is repatriated from one member state to another;

Aware of the importance of proper autopsy procedures, in particular with a view to bringing to light illegal executions, and murders perpetrated by authoritarian regimes;

---

[1] When this recommendation was adopted, the Representatives of Denmark and the Netherlands, in application of Article 10.2*c* of the Rules of Procedure for the meeting of the Ministers' Deputies, reserved the right of their Governments to comply or not with paragraph 2 (scope of the recommendation) of the present recommendation.

When this recommendation was adopted, the Representative of Germany, in application of Article 10.2*c* of the Rules of Procedure for the meeting of the Ministers' Deputies, reserved the right of his Government to comply or not with paragraph 2 *f* and *h* of the present recommendation.

Underlining the need to protect the independence and impartiality of medico-legal experts, as well as to make available the necessary legal and technical facilities for them to carry out their duties in an appropriate way and to promote their training;

Considering the importance of national quality control systems to ensure the proper performance of medico-legal autopsies;

Underlining the need to strengthen international co-operation with a view to the progressive harmonisation of medico-legal autopsy procedures at a European level;

Having regard to Recommendation 1159 (1991) on the harmonisation of autopsy rules adopted, at its 43rd Ordinary Session, by the Parliamentary Assembly of the Council of Europe;

Having regard to the Model Autopsy Protocol of the United Nations, endorsed by the General Assembly of the United Nations in 1991;

Taking into account the 'guide on disaster victim identification' adopted by the International Criminal Police Organisation (Interpol) General Assembly in 1997,

1 Recommends the governments of member states:
  - **(i)** to adopt as their internal standards the principles and rules contained in this recommendation;
  - **(ii)** to take or reinforce, as the case may be, all appropriate measures with a view to the progressive implementation of the principles and rules contained in this recommendation;
  - **(iii)** to set up a quality assurance programme to ensure the proper implementation of the principles and rules contained in this recommendation.

2 Invites the governments of member states to inform the Secretary General of the Council of Europe upon his or her request of the measures taken to follow up the principles and rules contained in this recommendation.

PRINCIPLES AND RULES RELATING TO MEDICO-LEGAL AUTOPSY PROCEDURES

Scope of the recommendation

1 In cases where death may be due to unnatural causes, the competent authority, accompanied by one or more medico-legal experts, should where appropriate investigate the scene, examine the body and decide whether an autopsy should be carried out.

2 Autopsies should be carried out in all obvious or suspected unnatural death, even where there is a delay between causative events and death, in particular:
  - **a** homicide or suspected homicide;
  - **b** sudden, unexpected death, including sudden infant death;
  - **c** violation of human rights such as suspicion of torture or any other form of ill treatment;
  - **d** suicide or suspected suicide;
  - **e** suspected medical malpractice;
  - **f** accidents, whether transportational, occupational or domestic;
  - **g** occupational disease and hazards;
  - **h** technological or environmental disasters;
  - **i** death in custody or death associated with police or military activities;
  - **j** unidentified or skeletalised bodies.

3 Medico-legal experts must exercise their functions with total independence and impartiality. They should not be subject to any form of pressure and they should be objective in the exercise of their functions, in particular in the presentation of their results and conclusions.

Principle I – Scene investigation

A *General principles*
  - **1** In case of obvious or suspected unnatural death, the physician who first attended the dead body should report to the competent authorities, the latter deciding whether an examination should be carried out by a qualified medico-legal expert or by a physician familiar with medico-legal examination.
  - **2** Particularly in cases of homicide or suspicious death, medico-legal experts should be informed without delay and, where appropriate, go immediately to the place where the body is found and have immediate access there. In this respect, there should be an adequate structure of co-ordination among all persons involved and, in particular, among judicial bodies, medico-legal experts and police.

B *Examination of the body*
  - **1** Role of the police

    The following tasks, among others, should be carried out by police officers:
    - **a** record the identities of all persons at the scene;
    - **b** photograph the body as it is found;
    - **c** make sure that all relevant artifacts are noted, and that all exhibits, such as weapons and projectiles, are seized for further examination;
    - **d** in agreement with the medico-legal expert, obtain identification of the body and other pertinent information from scene witnesses, including those who last saw the decedent alive, where available;

  - **e** protect the deceased's hands and head with paper bags, under the control of the medico-legal expert;
  - **f** preserve the integrity of the scene and surroundings;
- **2** Role of the medico-legal expert
  The medico-legal expert should without delay:
  - **a** be informed of all relevant circumstances relating to the death;
  - **b** ensure that photographs of the body are properly taken;
  - **c** record the body position and its relation to the state of the clothing and to the distribution pattern of rigor mortis and hypostasis, as well as the state of *post-mortem* decomposition;
  - **d** examine and record the distribution and pattern of any blood stains on the body and at the scene, as well as other biological evidence;
  - **e** proceed to a preliminary examination of the body;
  - **f** except where the body is decomposed or skeletal, note the ambient temperature and deep-rectal temperature of the body, and estimate the time of death by recording the degree, location and fixation of rigor mortis and hypostasis, as well as other findings;
  - **g** make sure that the body is transported and stored in a secure and refrigerated location in an undisturbed state.

Principle II – Autopsy physicians

Medico-legal autopsies should be performed, whenever possible, by two physicians, of whom at least one should be qualified in forensic pathology.

Principle III – Identification

In order to ensure that proper identification of the body is carried out in accordance with the disaster victim identification guide adopted by the General Assembly of Interpol in 1997, the following criteria should be considered: visual recognition, personal effects, physical characteristics, dental examination, anthropological identification, fingerprints and genetic identification.

- **1** Visual identification
  Visual identification of a body should be carried out by relatives or persons who knew and have recently seen the decedent.
- **2** Personal effects
  A description of clothing, jewellery and pocket contents should be recorded. These may assist correct identification.
- **3** Physical characteristics
  Physical characteristic should be recorded through an external and an internal examination.
- **4** Dental examination
  Where appropriate, the examination of teeth and jaws should be carried out by a dentist with medico-legal experience.
- **5** Anthropological identification
  Whenever human material is skeletised or in an advanced stage of decomposition, an anthropological identification should be carried out, if necessary.
- **6** Fingerprints
  Where appropriate, fingerprints should be taken by police officers. A close collaboration should exist between all experts involved.
- **7** Genetic identification
  Where appropriate, genetic identification should be carried out by an expert in forensic genetics.

It is appropriate to take biological samples from the deceased in order to assist genetic identification. Measures should be taken in order to avoid contamination and guarantee appropriate storage of biological samples.

Principle IV – General considerations

- **1** Medico-legal autopsies and all related measures must be carried out in a manner consistent with medical ethics and respecting the dignity of the deceased.
- **2** Where appropriate, the closest relatives should be given an opportunity to see the corpse.
- **3** Before beginning the autopsy, the following minimum rules should be applied:
  - **a** record the date, time and place of autopsy;
  - **b** record the name(s) of the medico-legal expert(s), assistant(s) and all other persons present at the autopsy with indication as to the position and role of each one in the autopsy;
  - **c** take colour photographs or video, where appropriate, of all relevant findings and of the dressed and undressed body;
  - **d** undress the body, examine and record clothing and jewellery, verify the correspondence between injuries on the body and clothing;
  - **e** where appropriate, take X-rays, particularly in cases of suspected child abuse, and for identification and location of foreign objects.
- **4** Where appropriate, before beginning the autopsy, body orifices should be appropriately swabbed for the recovery and identification of biological trace evidence.

5 If the decedent was hospitalised prior to death, admission blood specimens and any X-rays should be obtained as well as hospital records.

Principle V – Autopsy procedures

I EXTERNAL EXAMINATION

1 The examination of the clothing is an essential part of the external examination and all findings therein are to be clearly described. This is especially important in those cases where the clothing has been damaged or soiled: each area of recent damage must be described fully and relevant findings are to be related to the site of injuries on the corpse. Discrepancies in such findings are also to be described.

2 The description of the body following an external examination must include:

  - **a** age, sex, build, height, ethnic group and weight, nutritional state, skin colour and special characteristics (such as scars, tattoos or amputations);
  - **b** post-mortem changes, including details relating to rigor and post-mortem hypostasis – distribution, intensity, colour and reversibility – and putrefaction and environmentally induced changes;
  - **c** findings on a primary external inspection and description which, if required, include sampling of stains and other trace evidence on the body surface and a reinspection after removal and cleaning of the body;
  - **d** inspection of the skin of the posterior surfaces of the corpse;
  - **e** description and careful investigation of the head and the facial orifices includes: colour, length, density and distribution of hair (and beard); nasal skeleton; oral mucosa, dentition and tongue; ears, retro-auricular areas and external meati; eyes: colour of irises and sclerae, regularity and appearance of pupils, sclerae, conjunctivae; skin (presence and absence of petechiae to be described); if fluids have been evacuated from facial orifices, their colour and odour;
  - **f** neck: checking for excessive mobility, presence and absence of abrasions, other marks and bruising (including petechiae) over the entire circumference of the neck;
  - **g** thorax: shape and stability; breasts; aspect, nipples and pigmentation;
  - **h** abdomen: external bulging, pigmentation, scars, abnormalities and bruising;
  - **i** anus and genitals;
  - **j** extremities: shape and abnormal mobility, abnormalities; injection marks and scars; palmar surfaces, finger and toe nails;
  - **k** material findings under fingernails.

3 All injuries, including abrasions, bruises, lacerations and other marks have to be described by shape, exact measurement, direction, edges, angles and location relative to anatomical landmarks. Photographs should be taken. Bite marks shall be swabbed, and casts made where necessary.

4 Signs of vital reaction around wounds, foreign particles inside wounds and in their surroundings and secondary reactions, such as discolouration, healing and infections must also be described.

5 The investigation of cutaneous and sub-cutaneous bruising may require local skin incision.

6 Where appropriate, specimens from wounds must be removed for further investigations, such as histology and histochemistry.

7 All signs of recent or old medical and surgical intervention and resuscitation must be described. Medical devices must not be removed from the body before the intervention of the medico-legal expert.

8 A decision has to be taken at this stage as to the strategies of investigation and the necessity of documentation by X-rays and other imaging procedures.

II INTERNAL EXAMINATION

A General

1 All relevant artifacts produced by the dissection and from sampling procedures, must be documented.

2 All three body cavities – head, thorax and abdomen – must be opened layer by layer. Where appropriate, the vertebral canal and joint cavities should be examined.

3 Examination and description of body cavities include: an examination for the presence of gas (pneumothorax), measurement of volume of fluids and blood, appearance of internal surfaces, intactness of anatomical boundaries, external appearance of organs and their location; adhesion and cavity obliterations, injuries and haemorrhage.

4 The demonstration and dissection of the soft tissues and musculature of the neck have to be components of all medico-legal autopsies (see the paragraph concerning special procedures).

5 All organs must be examined and sliced following established guidelines of pathological anatomy. This includes opening of all relevant vessels, for example, intracranial arteries, sinuses, carotid arteries, coronary arteries, pulmonary arteries and veins, aorta and vessels of the abdominal organs, femoral arteries and lower

limb veins. Relevant ducts have to be dissected, for example, central and peripheral airways, biliary ducts and ureters. All hollow organs have to be opened and their content described by colour, viscosity, volume (samples should be retained, where appropriate). All organs have to be sliced and the appearance of the cut surface described. If injuries are present, the dissection procedure may have to vary from the normal one: this should be appropriately described and documented.

6 All internal lesions and injuries must be precisely described by size and location. Injury tracks must be described in order to include their direction as regards the organ anatomy.

7 The weight of all major organs must be recorded.

B Detailed

1 *Head*

- **a** Before opening the skull, the periosteum must be scraped off in order to display or exclude any fractures.
- **b** The head examination procedure must allow the inspection and description of the scalp, external and internal surfaces of the skull and of the temporal muscles.
- **c** The thickness and appearances of the skull and sutures, the appearances of the meninges, the cerebrospinal fluid (CSF), the wall structure and contents of cerebral arteries and sinuses must be described. The description of the bones must also include an examination of their intactness, including the connection between the skull and the first two vertebrae.
- **d** In obvious or suspected head injury (for example, if a detailed examination is required or if autolysis or putrefaction is present) fixation of the whole brain is recommended before its dissection.
- **e** Middle ears must be always opened and nasal sinuses where indicated.
- **f** The soft tissue and skeleton of the face is dissected only in relevant cases, using a cosmetically acceptable technique.

2 *Thorax and neck*

The opening of the thorax must be performed using a technique which allows the demonstration of the presence of pneumothorax and the inspection of the thorax walls, including the postero-lateral regions. *In situ* dissection of the neck must display the details of its anatomy.

3 *Abdomen*

The opening procedure of the abdomen must allow an accurate examination of all layers of the walls, including the postero-lateral regions. *In situ* dissection is necessary in certain cases, particularly for the demonstration of injury tracks and evacuation of fluids. Dissection of organs should observe anatomical continuity of systems, where possible. The whole intestine must be dissected and its contents described.

4 *Skeleton*

- **a** The examination of the thoracic cage, the spine and the pelvis must be part of the autopsy procedure.
- **b** Where appropriate traumatic deaths need a precise dissection of the extremities, possibly complemented by X-ray examination.

5 *Special procedures*

- **a** If there is any suspicion of neck trauma, the brain and thoracic organs are to be removed prior to the dissection of the neck, to enable detailed dissection to take place in a bloodless field.
- **b** If there is a suspicion of air embolism, pre-autopsy radiology of the thorax must be performed. The first stage of the autopsy in such a case must be a careful partial opening of the thorax and dislocation of the lower three-quarters of the sternum with the subsequent opening of the heart under water, allowing the measurement and sampling of escaping air or gas.
- **c** For the demonstration of particular injury patterns, deviation from the normal procedure of dissection has to be accepted, provided that such procedures are specifically described in the autopsy report.
- **d** The dissection in traumatic deaths must include a full exposure of the soft tissues and musculature on the back of the body. The same procedure must be applied to the extremities (so called ‘peel-off’ procedure).
- **e** In suspected or overt sexual assaults, the sexual organs are to be removed *en bloc* together with the external genitalia, rectum and anus, before they are dissected. Relevant swabs of orifices and cavities must be taken prior to this procedure.

6 *Sampling*

The scope of the sampling procedure is to be case-dependent. However, the following minimum rules should be applied:

- **a** in all autopsies, the basic sampling scheme includes specimens from the main organs for histology and peripheral blood sampling (such as for alcohol and drug analyses and genetic identification), urine and gastric contents. All blood samples must be peripheral blood and not heart or thoracic;
- **b** if the cause of death cannot be established with the necessary degree of certainty, sampling includes

additional specimens and fluids for metabolic studies and thorough toxicology. This includes blood, vitreous humour, CSF, bile, hair samples and further relevant tissues;

c if death is related to physical violence, sampling includes the injuries, for example to determine wound age and any foreign materials in the wounds;
d if reconstructions are desirable, the removal of bones and osseous compartments may become necessary;
e if identification is the predominant aim, the removal of jaws and other bones may be necessary;
f if strangulation or the application of physical force to the neck is suspected or diagnosed, the entire neck structures, musculature and neurovascular bundles must be preserved for histology. The hyoid bone and the laryngeal cartilages must be dissected very carefully;
g biological samples must be collected in tightly closed jars, properly preserved and placed under seal and transported to the laboratory in perfect safety;
h certain specimens and fluids need to be sampled in a special way and analysed without delay.

7 *Release of the body*
After a medico-legal autopsy has been carried out, medico-legal experts should ensure that the body is returned in a dignified condition.

Principle VI – Autopsy report

1 The autopsy report is as important as the autopsy itself, as the latter is of little value if the findings and opinions of the medico-legal expert are not communicated in a clear, accurate and permanent document. The autopsy report should be an integral part of the autopsy procedure and be drafted carefully.

2 The report should therefore be:
a full, detailed, comprehensive and objective;
b clear and comprehensible not only to other doctors, but also to non-medical readers;
c written in a logical sequence, well-structured and easy to refer to in various sections of the report;
d be in a legible and permanent form, with hard paper copy even if it is retained in electronic storage;
e be written in a discursive 'essay' style;

3 When drafting an autopsy report, the following minimum content should be included:
a legal preface to fulfil statutory requirements, if needed;
b serial number, computer retrieval coding and International Classification of Disease Code (ICD) code;
c full personal details of deceased (including name, age, sex, address and occupation) unless unidentified;
d date, place and time of death, where known;
e date, place and time of autopsy;
f name, qualifications and status of medico-legal expert(s);
g persons present at the autopsy and their function;
h name of the authority commissioning the autopsy;
i person(s) identifying the body to the medico-legal expert;
j name and address of the medical attendant of the deceased;
k a synopsis of the history and circumstances of the death, as given to the medico-legal expert by the police, judges, relatives or other persons, as well as information contained in the file, where available;
l description of the scene of death, if attended by the medico-legal expert; reference should be made to the provisions contained in Principle I above;
m external examination; reference should be made to the provisions of Principle V above;
n internal examination by anatomic systems, together with a comment on every organ. Reference should be made to the provisions of Principle V above;
o a list of all samples retained for toxicology, genetic identification, histology, microbiology and other investigations should be included; all such specimens should be identified and attested by the medico-legal expert according to the legal system of the state concerned, for continuity of evidence;
p results of ancillary investigations, such as radiology, odontology, entomology and anthropology should be included, when such results are available;
q one of the most important parts of the autopsy report is the evaluation of the significance of the accumulated results by the medico-legal expert. After termination of the autopsy, evaluation is usually provisional because later findings and later knowledge of other circumstantial facts can necessitate alteration and modification. Medico-legal experts must interpret the overall findings so that the maximum information and opinion can be offered. Also questions that have not been raised by the competent authority must be addressed if they could be of significance;
r based on the final interpretation, the cause of death (in the International Classification of Disease should be given. Where several alternatives for the cause of death exist and the facts do not allow a differentiation between them, the medico-legal expert should describe the alternatives and, if possible, rank them in order of probability. If this is

not possible, then the cause of death should be certified as 'Unascertained';

s the report should be finally checked, dated and signed by the medico-legal expert(s).

4 The date of the autopsy and the date of the provisional report should never be more than a day or two apart. The date of the autopsy and the date of the final report should be as close together as possible.

APPENDIX TO RECOMMENDATION NO. R (99) 3

SPECIFIC PROCEDURES (SELECTED EXAMPLES)

1 *Constriction of neck (hanging, manual and ligature strangulation)*
The examination of the scene where the body was found is extremely important: for example the presence of a chair or similar platform; fastening of the strangulation device; technique of tying of the knot; adhesive taping of hands and objects for trace evidence:

- Strangulation marks: depth, width, intermediate rings, direction, suspension point, raised ridges of skin, zones of hyperaemia, presence of duplicate strangulation marks; further specific neck injuries: dried excoriations due to slippage of the implement, marks due to textile weave pattern and structure, distribution of petechiae in the skin, bruising, scratch marks, blisters in the strangulation mark.
- Bleeding from facial orifices. Differences in widths of the pupils, localization of hypostasis, presence and distribution of congestion.
- Injuries due to convulsions, defensive injuries, injuries due to being held forcibly.

Dissection of the soft tissues, of the musculature and of the organs of the neck in a bloodless field is essential.

2 *Drowning/Immersion*
Note carefully the following findings: foam at the mouth, cutis anserina, maceration, mud and algae, lesions due to water animals, injuries due to surroundings (for example rocks and ships), loss of nails, skin, localization of livor mortis.

Technique: sampling of gastric contents, precise description of the lungs (weight, measurement, extent of emphysema), sampling, lung fluid, liver and other tissues, for the possible demonstration of diatoms and other contaminants.

If required, sampling of drowning medium (for example river, bath water) should be carried out.

3 *Sexually motivated murder*
The inspection and documentation of the scene of crime, e.g. relative to the injury pattern, is especially important. All injuries must be photographed together with a scale. If required, the body surfaces must be investigated under UV light and taped. Search for and sampling of foreign biological material must include pubic hairs and secretions on the body surface as for instance originating from bites. Such material must be preserved carefully for DNA investigation and protected against contamination. 'En bloc' dissection of the genital organs is strongly recommended. It is also necessary to proceed to the careful removal and sampling of material under the fingernails and control hairs.

4 *Death from child abuse and neglect*
State of nutrition and general care, thorough description and documentation of external injuries and scars, thorough examination for bone fractures (X-ray), must be evaluated.

Consider the removal of a variety of tissues: for example all injuries, regional lymph nodes in malnutrition, endocrine organs, immuno-competent tissues, specimens from different parts of the intestine.

5 *Infanticide/still-birth*
Special techniques of dissection are necessary to expose the falx cerebri and the tentorium cerebelli; describe the site of caput succedaneum; remove all fractures 'en bloc'; investigate all bone centres of ossification (size and presence). Special care is to be applied to the thoracic organs: degree of inflation of the lungs, flotation test 'en bloc' and 'en detail'. However, the limitations of the flotation test must be appreciated. All malformations must be described. As regards abdominal organs, gas content of the intestine must be investigated. The umbilical cord and the placenta must be subject to morphological and histological examination.

6 *Sudden death*
A subdivision into three main categories relative to the further strategy after gross examination is useful:

a findings that obviously explain the sudden occurrence of death (for example haemopericardium, aortic rupture). Cases belonging to this category can usually be regarded as sufficiently solved;

b findings that could explain the death but allow other explanations. Cases belonging to this category necessitate the exclusion of , for example,. poisoning and possibly histological proof of recent or chronic alterations relative to the cause of death;

c findings are either nil/minimal or do not explain the occurrence of death. Cases belonging to this category will usually require extensive further investigations.

This is especially so with sudden infant death cases. In such cases a more comprehensive investigative scheme is essential.

7 *Shooting fatalities*

The following should be carried out:

- extensive account on the scene of the incident, of weapons involved, of types of bullets, of sites of 'environmental' damage, of cartridge cases and of relative positions of persons involved;
- thorough examination of the clothing and description of relevant damage and careful sampling;
- thorough investigation and documentation of any blood (splashes) on the body surfaces (including clothing and hands);
- precise description of bullet entry and exit wounds relative to anatomical landmarks and distances from the soles of the feet and bullet tracks within the body;
- description of any impression marks of the muzzle;
- excision of uncleaned skin specimens surrounding entry and exit wounds;
- X-ray before and/or during autopsy (where necessary);
- determination of bullet tracks and their direction(s);
- final determination of direction(s) of fire, of the succession of shots, of intra-vital occurrence, of the victim's position (s).

8 *Death caused by explosive devices*

a As well as evaluating the cause of death, autopsy is essential to assist in reconstructing the nature of the explosion and identifying the type and maker of the explosive device, especially in aircraft sabotage or other terrorist actions.

b Full X-ray of the body must be made to detect and localise any metallic objects, such as detonator components, which may lead to the identification of the explosive device.

c The pattern of injury may indicate that the dead person was a perpetrator of the explosion, for example maximum injury in the lower abdominal region suggests that he or she carried the device on his or her lap during a premature explosion.

d At autopsy, all foreign objects in the tissues, identified on X-rays, must be carefully preserved for forensic examination.

e Samples of tissues, clothing, etc., must be retained for chemical analysis to identify the type of explosive.

9 *Blunt and/or sharp force injuries*

The following should be carried out:

- examination of the weapons or objects that are possibly involved (especially their dimensions);
- extensive examination and inspection of clothing (including damage, stains);
- careful dissection and description of all tracks (layer by layer) including their dimensions and weapon-related traces, signs of vitality.

10 *Fire Deaths*

The following should be carried out:

- examination of remains of clothing, - specific types and shapes of skin combustions;
- search for heat-related alterations and peculiarities;
- demonstration/exclusion of fire accelerants.;
- search for signs of vitality: carbon monoxide, HCN, soot inhalation, skin lesions.

11 *Suspicion of intoxication (General Outlines)*

11.1 Where anatomical findings do not reveal a cause of death and/or there is vague suspicion of poisoning, basic sampling should include peripheral blood, urine, stomach contents, bile, liver and kidney.

11.2 If specific suspicion arises, sampling should be group-related as follows:

- hypnotics, sedatives, psycho-active drugs, cardiac drugs and analgesics, pesticides: as aforementioned under (11.1);
- drugs of abuse: as aforementioned under (11.1) and additionally cerebrospinal fluid, brain tissue, injection marks, hairs;
- volatile fat-soluble substances such as fire accelerant and solvents: as aforementioned under (11.1) and in addition: blood from left ventricle, brain tissue, subcutaneous fat tissue, lung tissue, clothing;
- nutritional intoxication: as aforementioned under (11.1) and in addition: intestinal contents, if possible taken from 3 different sites;
- suspicion of chronic intoxication (heavy metals, drugs, pesticides etc.) as aforementioned under (11.1) and in addition: hairs (tufts), bones, fat tissue, intestinal contents.

12 *Decomposed bodies*

The presence of decomposition does not remove the need for a full autopsy.

Radiological examination will exclude bony injury, the presence of foreign bodies, for example bullets.

Toxicological studies (particularly estimation of alcohol concentrations) should be carried out but interpreted with great caution.

Index

References to autopsy, death and homicide are limited in the index, as these are major subject of the book. Readers are advised to seek more specific references. *vs* denotes differentiation or comparison.

- abdomen
  - anatomy, forensic 229
  - bruising, in child abuse 463
  - compression, traumatic ‘asphyxia’ in 363
  - examination in air embolism 347, 435–6
  - exposure/incisions at autopsy 17
  - internal examination at autopsy 18
  - kicking injuries 230
  - in neglect/starvation 413, 414
  - surface relations of organs 229
- abdominal injuries 229–33
  - child abuse 230, 463, 472
  - closed/blunt 229
  - complications 233
  - foreign bodies 232–3
  - haemorrhage 230, 232, 233
  - illegal abortion 433
  - open/penetrating 229
  - suicidal knife wounds 238
  - in vehicular accidents 286
    - drivers 283–4
    - pedestrians 294
- abdominal organs 229
  - examination at autopsy 18
  - removal at autopsy 21–2
- abdominal wall, bruising 229–30
- abortion
  - autopsy 435–6
    - external examination 435
    - internal findings 435–6
  - criminal/illegal 431, 432
    - air embolism 346, 432–3, 434, 435
    - causes of death 432–5
    - cervix dilatation 432, 433
    - dangers 435
    - drugs/chemicals 434–5
    - infections 433–4
    - infections/bleeding 432
    - instrumental interference 432
    - intrauterine infection 433–4
    - physical violence 433
    - syringe aspiration 433
  - deaths associated 431–5
  - septic 57, 432
  - skin colour 346
  - sudden death due to 517
  - therapeutic 346, 431
- abrasion collar 263–4, 268–9
- abrasions 137, 138–43
  - in bruising *vs* hypostasis 59
  - brush (tangential) 139, 140
    - face 178
    - motorcyclists 290
    - scalp 175
  - crushing 139–40
  - defence wounds 165–6
  - definitions and lay terms 137, 138
  - fingernail *see* fingernail, abrasions
  - footmarks 144, 150
  - impact, of scalp 175
  - intradermal bruising with 140, 141, 142, 146
  - linear (grazes) 138, 139
  - neck
    - infanticide 441, 443
    - strangulation 140, 372
  - patterned 141–3, 144
    - child abuse 475
    - facial 178
    - firearms muzzle 143
    - pedestrians 294, 295
    - tyres, vehicular injuries 142, 143, 294, 296
  - post-mortem 139, 143, 149–50
  - putrefied corpse 39
  - scalp 175
  - sexual offences 422
  - strangulation marks 141, 148, 372–3
- acceleration injuries 205, 281, 291
- accidents *see* trauma; vehicular accidents
- acetabulum, sex differentiation 111
- acetaminophen (paracetamol) poisoning 572–3
- acetylsalicylic acid *see* aspirin poisoning
- acid fuchsin (AF) staining 500
- acid phosphatase 428
- acquired immunodeficiency syndrome (AIDS) 579
- acridine-orange fluorescent stain 601
- actin 62

Index

- actomyosin 62
- acute tubular necrosis 345
- ADAM (Ecstasy) 581
- adenosine diphosphate (ADP) 62
- adenosine triphosphate (ATP) 61, 62
- adipocere
  - chemistry/characteristics 70
  - decomposition forms mixed with 72
  - definition 64, 70
  - formation 68, 70–2
    - bacteria 71
    - in drowning 406
    - importance 71–2
    - timing 71
- adipose tissue
  - decomposition *see* adipocere
  - loss, in neglect/starvation 414
- ‘adrenal apoplexy’ 349
- adrenal glands, autopsy 27, 349
- adrenal haemorrhage 349, 513
- adrenaline:noradrenaline ratio 419
- β_2-adrenergic antagonists 515
- adult respiratory distress syndrome 345
- African features 122, 123
- age
  - hypothermia relationship 415
  - infant, sudden infant death syndrome and 452
- age estimation
  - by dentition 534–5
  - external examination at autopsy 11
  - eyes 104
  - infants 440, 443, 448–9
  - non-skeletalized bodies 104–5
  - skeletal remains 116–22
    - adult 118
    - child/young adult 118
    - factors affecting 118
    - fetus/young infant 117–18, 449
    - later years 118
    - pubic symphysis 118–20
    - radiography/radiology 117, 121
    - skull sutures 120–2
    - sternal rib methods 120
  - teeth 104–5, 534–5
- agrochemical poisoning 566–9
  - organophosphorus pesticides 567–8
  - paraquat 345, 566–7
- AIDS 579
  - *see also* HIV infection
- air
  - in diving 488
  - insufflation in illegal abortion 432–3, 435
  - movement, effect on cooling curve 81
- airbags 289
- aircraft
  - carbon monoxide poisoning 560–1
  - crashes
    - identification of dead from dentition 532–3
    - necessities in planning 42–3
  - pilots 296, 508
- air embolism 346–8
  - air volume for 346–7
  - arterial 346
  - aspirometer 347, 348
  - autopsy in 347–8, 435–6
  - barotrauma 346, 488, 489
  - in cerebral veins, myth 347, 435–6, 490
  - in cut throats 237, 346
  - divers 488
  - gas chromatography 347, 348
  - homicide 346
  - illegal abortion 346, 432–3, 434
    - autopsy findings 435
  - legal termination of pregnancy 432
  - mechanism of death 346, 347
  - medico-legal aspects 346
  - radiograph 347
  - transfusion/infusion/surgical accidents 346
- air guns 271
  - injuries 271–2
- air-rifles, injuries 271–2
- airways
  - damage, resuscitation artefacts 41
  - froth in, in drowning 402–3
  - gastric contents in 362–3
  - obstruction
    - anaesthetic-related deaths 484–5
    - choking 361–3
    - gagging 360–1
    - infections 361
    - pressure on neck causing 368
  - occlusion, fatal pressure on neck 368, 369
  - thermal damage 319–20
- Al-Alousi’s ‘temperature ratio’ method 83, 87
- alcohol 552–9
  - absorption 552, 553–4
    - drugs affecting 554
    - food effect 553, 554
  - behavioural effect 552, 556–7
  - binge drinkers 554
  - blood-alcohol curve (BAC) 554
  - blood levels 19–20, 553
    - ante-mortem *vs* post-mortem 546–7
    - calculation from drinks 556
    - deaths in custody 307
    - peripheral vein blood 547
    - road traffic deaths 296, 557
    - sample collection/containers 20, 545
    - sample preservation 545, 546
    - women 552
  - breath 555
  - carbon monoxide poisoning and 557, 560, 561
  - cardiomyopathy associated 558
  - choking association 363

alcohol (*contd.*)
  chronic abuse 552, 557–8
  concentrations of drinks 553–4, 555–6
  deaths from exposure 419
  deaths in custody 307
  as depressant 552–3
  diffusion (post-mortem) and sampling concerns 546–7
  driving and 554
    legal limit in UK 555
    urine levels 555
  elimination/excretion 554–5
  heat regulation changes 553
  hypothermia and 553
  hypoxia 552
  metabolism 554
  mode of action 552–3
  modes of death 557–8
    acute alcoholism 307, 557
    chronic alcoholism 557–8
  physiological effects 552–3, 556–7
  putrefactive production 545–6
  spontaneous combustion association 323
  subarachnoid haemorrhage and 199, 519
  sudden death from berry aneurysm and 199, 519
  sudden immersion in cold water 400, 557
  units 554, 555–6
    specific drinks 555–6
  volume/volume 555–6
alcoholism 554–5
  acute 307, 557
  chronic 552, 557–8
  falls 557
alkali, deaths from 587–8
alkaline phosphatase 169
alveoli
  alcohol levels in air 555
  in newborn lung 447–8
  in paraquat poisoning 567
Alzheimer's-like condition, in boxers 204
ambient temperature *see* temperature, ambient
American Society of Anesthesiologists 482
amino acids
  dating of bones 122, 127
  in teeth 105, 535
amitriptyline 573
ammunition 247
  rifled weapons 248
  shotgun 246, 247–8
  see also bullets
amnesia, retrograde 209–10
Amnesty International 301
amniotic fluid embolism 436–7
amphetamine abuse 580, 581
β-amyloid precursor protein 470
anaemia, carbon monoxide poisoning in 562, 563
anaerobes, adipocere formation 71
anaesthesia-related deaths, medico-legal aspects 482, 483
anaesthetic-related deaths 480–5
  airway obstruction 484–5
  autopsy procedure 480–2
  cardiac arrest 484
  causes 482–4
  equipment failure 485
  histology 483, 485, 486
  hypoxia in, brain changes after 484, 486
  incidence/risk 484
  local/epidural anaesthesia 485
  mode/causes of death 482–4
anaesthetist 482, 484
analgesics, poisoning 571–2
anal intercourse 427, 428–9
anatomy, forensic
  abdomen 229
  chest 222–3, 223
  circle of Willis 512, 513
  cranium 181
  heart 223
  larynx 376, 377
  scalp 174–5
  skin 137–8, 174
  skull and meninges 108, 181–2, 182, 189
  vertebral arteries 201
aneurysms
  aortic 509–10
    atheromatous 509–10
    diagnosis 510
    dissecting 510
    fusiform/saccular 510
  berry *see* berry aneurysm
  cardiac 503
  cerebral arteries 196, 512–13
    rupture 511–13
  circle of Willis 197, 198, 511
  mycotic 511
  rupture, sudden death 509–13, 519–20
  syphilitic 510–11
angel dust (phencyclidine) 580
angiography, post-mortem 26, 200–1, 203, 512
animal bones 107, 108
animal predation
  buried body decomposition 69
  dating of bones and 126
  immersion deaths 396, 397, 398
  maggot infestation 65, 66, 74–5, 77
  marine predation 68, 75–6, 396, 398, 399
  mummified corpse 73
  post-mortem damage 73–6, 74, 75–6, 126
  putrefied corpse 67
  types/types of damage 74
  see also insects
anoxia 52, 53
  stagnant 352
ante-mortem *vs* post-mortem injury
  burns 317–19, 320

- in fire 7
- fractures 182, 344
- joint splits 7
- laryngeal fractures 378
- skull fractures 182
- ‘undertaker’s fracture’ 21
- wounds 167–9, 170–1

anthropometric dimensions 123
anticoagulants 545
antidepressants, poisoning 573
antifreeze agents 589
antimony poisoning 591
ants, post-mortem damage 75
anus
- bruising
  - homosexual offences 429
  - sexual offences 422, 427
- dilatation
  - homosexuals 427
  - interpretation 423–4, 428
  - sexual offences 427, 428
- examination at autopsy 16
- funnel-shaped 16, 429
- in homosexuals 427, 429
- injuries, in sexual offences 423
- patulous 427, 429

aorta
- anatomy 223
- aneurysms *see* aneurysms, aortic
- atherosclerosis 509–10
- deceleration injuries 228, 229
- double-barrelled 510
- examination at autopsy 25
- impact injury 228, 229
- injuries 228
  - vehicular accidents 285, 289
- ‘ladder tears’ 21, 228, 285
- medionecrosis 510
- ruptured 228, 340
  - vehicular accidents 285
- transection 228

aortic valve disease 506–7
- sudden death 506–7

aortic valve stenosis 506–7
aortitis, syphilitic 511
aorto-oesophageal fistula, surgical procedures, deaths associated 481
apex, myocardial infarction 497
apnoea, in sudden infant death syndrome 455–6
apnoea alarms 456
aponeurosis 174–5
apoptosis, conducting system, sudden death due to 505
appendicitis, acute 517
appendix abscess 517
arachnoid 189
arachnoid granulations 189
arcus senilis 104
arm, injuries in vehicular accidents 284, 286
arm-locks 353, 383
- deaths in custody 306
- fractured laryngeal horns 378

arrest, deaths after/during 305–7, 353
- *see also* custody, deaths in

arrhythmias *see* cardiac arrhythmias
arsenical poisoning 590–1
- acute/chronic 590
- autopsy findings 590–1
- blood/tissue levels 591
- corpse preservation 590
- fatal dosages 590
- sources 590
- toxicology sampling/results 591

arsenious oxide poisoning 590, 591
artefacts *see* post-mortem artefacts; resuscitation artefacts
arterial spasm 503
arteriovenous anastomosis 512
arthritis 104
asbestos/asbestosis 24–5, 103
‘asphyxia’ 352–67, 368, 369
- autopsy diagnosis 356–7
- cautions on diagnosis 357, 358, 368
- classic signs 353–7
  - absent in suffocation 358
  - cautions on use 353, 357, 368
  - congestion and oedema 356
  - cyanosis 356
  - engorgement of right heart 356
  - fluidity of blood 356
  - petechial haemorrhage 353–6, 364
- conditions/examples 352
- definitions/terminology 352, 353
- histological/biomechanical diagnosis 357
- hypostasis colour confusion 57, 58, 356
- ligature strangulation 382
- masochistic 235, 240
- mechanical 353, 369
  - *see also* choking; smothering; strangulation; suffocation
- nature of 352
- ‘obsolete diagnostic quintet’ 353, 357
- plastic bag 390, 391, 596
- postural (positional) 355, 365
  - deaths in custody 306
- sexual *see* sexual asphyxia
- in solvent abuse 596
- sudden death in epileptics 514
- suicide by 240
- traumatic 352, 363–4
  - cyanosis 364
  - deaths in custody 306

aspiration, gastric contents *see* gastric contents
aspirin poisoning 571–2
- hypersensitivity 571
- spot tests 572

aspirometer, autopsy in air embolism 347, 348
assaults, berry aneurysm rupture and 198, 519–20

asthma, sudden death 515
astrocytes, in cerebral hypoxia 213
atherosclerosis 104
  aortic 509–10
  coronary *see* coronary atherosclerosis
atlanto-occipital dislocation, in vehicular accidents 284
atria, examination at autopsy 25
auditory meatus temperature 7
Austrian decree 1
autoerotic procedures 389–92
  see also sexual asphyxia
autolysis 36, 59–60, 89
  early myocardial infarction *vs* 498
  post-mortem 345, 500–1
automatic pistols 247, 248, 273
automatic weapons 248
autopsy
  clinical (academic) 2
  forensic *see* autopsy, medico-legal
  history 1
  legislation 2
  types 2
autopsy, medico-legal 2
  abortion 435–6
  abrasions caused by 143
  in air embolism 347–8, 435–6
  in 'asphyxia' 356–7
  attendance at 3–4
  authorisation/consent 3
  blind 10
  brain examination 28–9
  bruise development 145
  burns 315–16
  case history 9–10
  child abuse 475–6
  computed tomography (CT) 31
  consent documents 3
  consent issues 2
  in decompression chamber 489
  dentures examination 15
  DNA profiling sample 604
  drowning *see* drowning, autopsy signs
  epilepsy 514
  examination of organs *see individual organs*
  external examination *see* external examination at autopsy
  failure rates 45
  fatal abuse, evidence 301
  hazards for staff 10–11
    cyanide poisoning 586
    HIV/hepatitis infection 4, 10–11
  identification of body 8–9
  incisions 16–17, 17, 425
  infanticide 442–5
  internal examination *see* internal examination at autopsy
  legislation 2
  magnetic resonance imaging (MRI) 31
  mass disasters 44
  objectives 3
  obscure 44–7, 493
    mode of death 46–7
    samples 45–6
    syndrome 45
  photography *see* photography
  plastic bag suffocation 359–60
  post-mortem artefacts *see* post-mortem artefacts
  precautions 10–11
  procedures 2–4
  putrefied corpse 39–40
    see also putrefied corpse
  radiology *see* radiography/radiology
  reasons 493
  Recommendation No. R (99) 3 610
  report 33–5
    body sketches 35
    cause of death 10
    differential diagnosis 10
    discussions/conclusions 35
    external examination 34
    format/information 33–5
    information required 34
    internal examination 34–5
    'original,' legal requirements 35
    timescale 35
  resuscitation artefacts *see* resuscitation artefacts
  retention of tissues/organs 3
  samples/ancillary investigations 29–30
  standards 1–2
axonal injury, diffuse
  in child abuse 470
  concussion in 210

back injuries, in human rights abuse 303
'back spatter' 261
bacteria
  adipocere formation 71
  in putrefaction 64–5, 68
'barbiturate blisters' 562, 574
barbiturates 486, 573–4
  poisoning 573–4
barium, in tracheobronchial tree 363
baroreceptors 369
barotrauma 346, 488–91
  air embolism 346, 488, 489
  autopsy 30, 489–90
  definition 488
  physical damage 488–9
  pulmonary 489
Barr bodies 104
basal arteries 46, 347
'base wing index' 112
basilar artery 201, 203
  see also basilovertebral artery injury

basilovertebral artery injury 200–4
  autopsy appearances 201–2
  hypothesis and problems with 200
  see also vertebral artery
basisphenoid synchondrosis 121
bathroom, deaths in 400, 401, 405
  electrical fatalities 326, 333–4, 334–5, 400, 405
  homicide 400
baton rounds 274
battered baby *see* child abuse
battery acid 587
beard, growth after death 61
beatings, torture and human rights abuse 302–3
beetles 77–8
behaviour, alcohol effect 552, 556–7
'bends' 346
benzene abuse 597
benzidine 127
benzodiazepines, poisoning 542, 573
berry aneurysm 196
  autopsy 512
  'congenital' 512
  multiple/sizes 512
  ruptured 511
    alcohol and 199, 519
    trauma/disease interaction 198–9, 519, 520
  spontaneously leaking 199
  sudden death from, mechanism 198–9, 199, 511
'beyond reasonable doubt' test 341, 471, 518, 520
bile, sample collection 548
biochemistry
  post-mortem interval 89–90
  rigor mortis 61–2
  vitreous humour 88, 89–90, 575
birds, post-mortem damage 76
birth
  concealment 440, 441, 442
  live, definition/proving 440–1, 442, 443
  newborn, lungs 447–8
bites/bite marks 527–32
  autopsy/re-examination 528, 531
  bruises associated 146
  child abuse 474, 476, 528
  common sites 528
  ear 180
  in foodstuffs 527–8
  identification of suspect 530
  impression/casts of mark 531
  investigation 530–1
  love bites 146, 421, 474, 528, 530
  matching with suspect's dentition 532
  multiple 529
  nature of mark 529–30
  nipples 528, 529, 530
  photography 528, 530–1, 532
  preservation 528
  sample collection 528, 531
  self-inflicted 529
  in sexual offences 421–2, 528, 529, 530
black eye, causes 178, 179, 180
  direct violence 149, 179
  fractures of anterior fossae of skull 179, 180
  gravitational seepage 179
  movement of bruises 146, 148, 175–6, 179
bladder
  injuries 233, 289
  urine collection at autopsy 548
blast injury 226, 274–5
bleeding
  absence in crush injuries 339
  deaths in custody 307
  in illegal abortion 432
  post-mortem 225, 233, 340
  scalp injuries 175, 176
  spurious, behind larynx 58, 150, 375
  see also haemorrhage
blisters
  barbiturate 562, 574
  in burns 315
    first degree 312–13
    post-mortem/'peri-mortal' 315, 317
  in carbon monoxide poisoning 562
  in electrical fatalities 330, 331
  post-mortem artefacts 36
  in putrefaction 64, 65
blood
  alcohol levels *see under* alcohol
  in carbon monoxide poisoning 562, 563–4
    see also carbon monoxide
  carboxyhaemoglobin *see* carboxyhaemoglobin
  changes in drowning 404–5
  clotting 58, 341–2, 356
  culture, samples 29
  DNA profiling 106, 604
  flow, retinal, sign of death 54
  fluidity in asphyxia 356
  groups 428, 440
  from mouth/nostrils, post-mortem artefacts 12, 15, 36
  pooling, in hypothermia 417
  pressure, alcohol effect 553
  purging in putrefaction 36, 39, 65
  sample collection *see* sample collection, blood
  sample preservation 546
  sludging, in hypothermia 418
  splashes 5–6
    exclamation mark 6
  spurious formation of cyanide 587
  volume in cardiac tamponade 228
blood-alcohol curve (BAC) 553, 554
blood–brain barrier 212

- blood vessels
  - calcification 26
    - *see also* calcification
  - examination at autopsy 25–7
- blowfly 76
- bluebottles 76–7
- ‘blue scar’ of coal-miners 103
- blunt injuries
  - abdomen 229
  - deaths in custody 306–7
  - human rights abuse 304
  - penetrating 153
  - scalp 177
  - skull fractures 177, 183
  - *see also* abrasions; bruises/bruising
- body
  - buried *see* buried bodies
  - clothed, examination at autopsy 8, 296
  - dimensions
    - effect on cooling curve 80
    - male/females (tables) 108
    - race determination 123
  - fate after death 53
  - fluids, sample collection 19–21, 549
  - fragments, in explosion deaths 275–6
  - identification *see* identification
  - internal examination at autopsy 16–23
  - labels/tags 9
  - length/height *see* height; length
  - posture, hypostasis pattern 56
  - retrieval at mass disasters 43–4
  - scene of death *see* scene of death
  - size, racial/ethnic characteristics 105
  - sketches, autopsy report 35
  - transection 154
  - weight *see* weight
- body bag 7
- body secretions
  - clothing 8
  - purging, in putrefaction 36, 39, 65
  - release, external examination of body 12
  - sample collection *see* sample collection
  - in sudden infant death syndrome 453
  - *see also* frothy fluid; specific secretions
- bombs, injuries from 274–5
- bondage, autoerotic procedures and 391
- bone(s)
  - animal origin 107, 108
  - bullet wounds 266–7, 270
  - decay, factors affecting 126–7
  - density, dating of bones 126
  - disease, infant fractures due to 468–9
  - dried, stature determination 114, 117
  - fragments, identification 107–8
  - haversian architecture 108
  - human origin, determination 107–8
  - injuries, bone marrow embolism 344
  - long
    - lengths in women/men (table) 113
    - race determination 122–3
    - sexing 112–13
    - stature determination 114
  - recent *vs* old, differentiation 126
  - recognition and confirmation 106–8
  - sex determination from 108–14
  - shotgun contact wounds over 250–1
  - stature determination from 114–16
  - wormian 469
  - wounds from rifled weapons 266, 267, 270
  - *see also* fractures; skeletal remains; *specific bones*
- bone dysplasias 468–9
- bone marrow embolism 344
- boxers, head injuries 204
- ‘boxer’s attitude’ 316
- boxing, kinetic energy transfer 137
- bradycardia 369
- brain
  - air gun pellets 272
  - autopsy
    - examination 28–9, 46
    - fixation 28, 29, 512
    - meningeal bleeding 28
    - removal 22–3, 150
    - serial sections preparation 28–9
    - weight 28
    - wetcutting 28
  - in barotrauma/dysbarism 489–90
  - in carbon monoxide poisoning 562, 563
  - cerebral aneurysms 512–13
  - damage 204
    - carotid artery occlusion 369
    - child abuse 192, 205, 465, 471, 476
    - impact, theories 205
    - mechanism 205–6
    - motorcyclists 290
    - pressure on neck causing 370
    - *see also* cerebral injuries
  - deformation
    - closed head injuries 205, 206
    - laminar 206
    - transient 206
  - in electrical fatalities 333
  - petechiae 354
  - putrefied corpse 40
  - sample collection 550
  - swelling *see* cerebral oedema
  - *see also entries beginning* cerebral
- brain death 52–3, 212
  - medico-legal aspects 53
- brain membranes
  - forensic anatomy 181, 189
  - haemorrhage 28, 188, 190–2, 211

    forensic implications 198
    see also extradural (epidural) haemorrhage; subarachnoid haemorrhage; subdural haemorrhage
  infections 189
brainstem
  damage 53
  death 53
  in electrical fatalities 327, 329
  haemorrhage 207, 214, 339
  lesions, secondary 213–14
breasts
  bites 528, 529, 530
  bruising, in sexual offences 422
breath, alcohol levels 555, 603
breathing, evidence in newborn 445–8
brittle bone diseases 468–9
bromine, abuse 597
bronchial asthma, sudden death 515
bronchial obstruction, surgical procedures, deaths associated 481
bronchial tumours 516
bronchodilators, 515
bronchopleural fistula 223, 225
bronchus, thermal damage in fires 320
bruises/bruising 143–50
  abdominal 229–30
  abrasions with 140, 141, 142, 146
  alteration with time 145, 146–8
  bite marks 148, 530
  cerebral 206
  chest/heart/face, resuscitation artefacts 40
  in child abuse 148, 462–3, 475
    age of bruises 148, 463
    fingertip 145, 148, 462, 463
  colour changes 147
  'coming out' 145
  dating, variables affecting 147
  defence wounds 165, 166
  definition 143–4
  depth/size and severity 145
  development 145
  factors affecting prominence/size 144–5
  fingertip 145, 148, 462, 463
  haemosiderin reaction 167
  hilt of knife causing 158
  hypostasis differences 59–60
    post-mortem *vs* 59–60
  in immersion deaths 398
  intradermal 140, 141, 142, 144, 146
    human rights abuse 302–3
  kicking causing 147, 149
  lacerations with 150–1, 151
  lips 150
    in sexual offences 421
  lungs 226
  movement 146, 175–6
    see also black eye
  muzzle imprint in shotgun wounds 250
  neck
    ligature strangulation 382
    manual strangulation 371–2, 374
  post-mortem artefacts 13, 36, 149–50
  prominence after death/autopsy 145, 176
  putrefied corpse 40
  scalp 175–6, 463
  seatbelts causing 289
  in sexual offences 421, 422, 427
  sixpenny 148, 463
  of special significance 148–9
  subcutaneous tissue 144–5
  'tramline' ('railway line') 148, 149, 302–3
  vessel fragility affecting 145
  vessels involved 144
brush abrasions *see* abrasions, brush
buccal cells, DNA profiling sample 604
'bucket-handle' effect, femur 466
buggery 427
bullets
  dumdum 137, 245
  explosive-tipped 137, 245
  modifications for kinetic energy transfer 137, 245
  'precessing' 265
  rifled weapons 248
  rubber and plastic 274
  soft-headed 245
  wounds *see* gunshot wounds; rifled weapons, wounds
'bumper' fractures 293
burial, traumatic 'asphyxia' in 363
buried bodies
  decomposition 68–9, 126
  deep burial 69
burning
  atypical localized (spontaneous combustion) 323–4
  suicide by 239–40
burns 312–25
  in acute alcoholism 557
  ante-mortem 315
    post-mortem burns *vs* 317–19, 320
  autopsy appearances 315–16
  child abuse 474
  classification of severity 312–13
  corrosive acids/alkalis 587
  deaths from lightning 336
  dry heat 314–15
    severity 315
  electrical 329–32
  fires 316, 318
  first degree 312–13
  flash 275
  friction *see* abrasions, brush
  hot water bottle 314, 320, 433, 435
  human rights abuse 303
  identification of body 99

burns (*contd.*)
  infections 314
  missile injuries from explosions 275
  moist heat (scalds) 313–14
  molten rubber 303
  muscle stiffening of body 63
  post-mortem 331–2
  renal failure after 345
  in rifled weapon wounds 259, 262
  'Rule of Nines' 13, 313
  scars, in identification 102
  second degree 313
  shotgun wounds 250, 251, 252
  spurious 'wounds' 316–17
  temperature/time relationship 312
  third degree 313
burnt bone fragments, identification 107–8
'burst atlas' injury 214
'burst head' 255, 267
'burst lobe' 207
butane 561, 595, 597
butcher's cleaver, wounds 165
buttocks, bruising, in child abuse 462, 463

cachexia 413
cadaveric spasm 61, 62–3
'cafe coronary' syndrome 362
Caffey's syndrome 469
calcification
  aortic valve 506, 507
  coronary arteries 26
  cricoid cartilage 376
  under periosteum 466
calcium oxalate 589
Calliphora 77
callus, in identification of remains 124, 125
callus formation 224, 467, 468
calvarium, removal 22, 347
cameras 32–3
  see also photography
Campbell de Morgan spots 11, 104
cannabis 581
capillaries
  bleeding (invisible to eye) 144
  dating emboli/thrombi 342
  permeability 212
capsaicin, injury 306
'captive-bolt' weapons 272–4
Carabelli's cusp 534
carbolic acid 587, 588
carbon dioxide, suffocation associated 357–8
carbon monoxide
  blood analysis 563–4
    sample collection 547
    sample preservation 546, 563
  blood levels
    anaemia 562, 563
    burns/fires 317, 322, 562
    road traffic deaths 296
    see also carboxyhaemoglobin
  in flash fires 320, 322, 564
  haemoglobin saturation 322, 563–4
  inhalation in fires 320–2, 561
    alive when fire started 564
    death due to 320–2
  poisoning 320, 560–5
    in aircraft 560–1
    alcoholics/alcoholism 557, 560, 561
    autopsy 562–3
    in bathroom 400
    blisters 562
    causes 560–2
    cherry-pink hypostasis 57, 321, 562
    children 322
    coal-mining 561
    divers 561
    elderly and infants 563
    neurological lesions after 563
    signs 321
    sodium hydroxide test 562
    suffocation in 357–8
    suicide 560
    vehicle drivers 560–1
  post-mortem formation 564
  in shotgun wounds 250, 252, 267
  sources 560–2
    fires 561
    heating appliances 560, 561
    incomplete combustion 561–2
    industrial processes 561
    motor vehicles 560
carbon tetrachloride 597
carboxyhaemoglobin 563–4
  analytical methods 564
  carbon monoxide poisoning 322, 562–3
  levels
    deaths in fires 562
    heat haematoma/peripheral blood 192, 317, 318, 319
  samples 20
  in shotgun wounds 250, 564
cardiac aneurysm 503
cardiac arrest 53, 55
  anaesthesia association 484
  asphyxia in 353
  aspirin poisoning 572
  choking due to food 362
  in electrical fatalities 328, 329
  manual strangulation *vs* 374, 378–9
  reflex 358, 362, 369–70, 370–1
    carotid artery damage 378–9
    causes 371
    criminal abortion 432
    deaths in custody 306
    frequency 370

    - hanging and/*vs* 384, 388
    - immersion deaths 400, 403
    - solvent abuse 595–6
- cardiac arrhythmias
  - aspirin poisoning 572
  - catecholamines 504, 579, 596
  - electrical fatalities 327, 328, 329, 333
  - sudden death due to 503
- cardiac asthma 505
- cardiac massage, external 231–2
- cardiac sarcoidosis 509
- cardiac tamponade 228
  - aortic aneurysm rupture 510
  - chest and heart injuries 227, 228
  - myocardial infarction complication 502
- cardiomegaly 505
- cardiomyopathies
  - alcoholic 558
  - arrhythmogenic right ventricle 508
  - definition 507
  - dilated 507
    - familial 507
    - idiopathic 507
  - heart weight 507
  - hypertrophic obstructive 507
  - sudden death 507–9
  - World Health Organization classification 508
- cardiopulmonary resuscitation (CPR)
  - child abuse *vs* 467–8, 472–3
  - myocardial contraction bands after 498–9
  - *see also* resuscitation artefacts
- cardiorespiratory failure 52, 53
  - sign of death 53, 55
- cardiovascular disease, death whilst immersed in water 397–8
- carotid arteries
  - compression 369
  - damage
    - in hanging 388
    - in manual strangulation 378–9
    - reflex cardiac arrest 378–9
  - examination/dissection 24, 379
- carotid sinus reflex 369
- 'carotid sleeper' 306
- cartilage
  - dating of bones 126
  - stab wounds 163
- case history, in autopsy 9–10, 453
- catecholamines 41, 361, 419
  - arrhythmias caused by 504, 579, 596
  - resuscitation artefacts 41, 353
  - sudden death from subarachnoid haemorrhage 519
- cathepsins, in wounds 169
- 'cat-o'nine-tails' 303
- Caucasian features 122, 123, 534
- cause of death *see* death, cause
- cavitation, high-velocity missiles/projectiles 246, 270
- cellular death 52, 63–4
- central nervous system, in barotrauma/dysbarism 490
- cerebellar tonsils, herniation/coning 211
- cerebral aneurysm, rupture 511–13
  - *see also* berry aneurysm
- cerebral arteries 197
- cerebral arterioles, sensitivity 199
- cerebral contusion 206
- cerebral dysfunction, hypothermia with/and 415
- cerebral fat embolism 343
- cerebral haemorrhage *see* intracerebral haemorrhage
- cerebral hypoxia 212–13
  - histology 212–13
  - in sexual asphyxia 389
  - stages and changes 213
- cerebral infarction 214
- cerebral injuries 204–12
  - in child abuse 192, 205, 465, 471, 476
  - concussion after 209–10
  - coup/contrecoup damage 207–9
  - deceleration 205, 206
  - intracerebral haemorrhage 207
  - mechanism 193, 206
  - rotatory movements of head and 209
  - *see also* brain, damage
- cerebral lacerations 206–7
- cerebral oedema 211–12
  - autopsy features 211
  - cause 211
  - child abuse 470, 471
  - hypoxia in 212
- cerebral veins, air embolism myth 347, 435–6, 490
- cerebrospinal fluid
  - anaesthetic-related deaths 485
  - leakage 189
  - sample collection method 21, 548
  - in subarachnoid haemorrhage 197, 199
- cervical shock, in illegal abortion 432
- cervical spine
  - autopsy 21, 203
  - damage/dislocation in hanging 215, 384, 385
  - hyperflexion/hyperextension injuries 214–15, 215, 284
  - seatbelt damage 284
  - vertebral artery trauma 203
- cervical vertebrae
  - dislocation 215, 384, 386
  - injuries 214–15, 215
- cervix
  - dilatation, in illegal abortion 432, 433
  - perforation 433
- chemicals, illegal abortions 434–5
- chemical tests, dating of bones 127–8
- chest
  - bruising, in child abuse 463
  - compression, traumatic 'asphyxia' in 363
  - crushing 363–4
  - flail 223, 224, 294
  - forensic anatomy 222–3

chest (*contd.*)
  haemorrhage 224–5
    external bleeding absent 225
  infection 224–5
  injuries 222–9
    cardiac tamponade 227, 228
    self-inflicted 243
    vehicular accidents 283–4, 285
  radiograph, in air embolism 347
  stab/sucking wounds 225
    see also stab wounds
  suicidal knife wounds 238
  see also thorax
chest wall
  damage, in child abuse 467–8
  injuries 223–4
  rib fractures *see* rib fractures
'chicken-fat' plasma clot 341
child abuse 11, 414, 461–79
  abdominal injuries 230, 463, 472
  abrasions, patterned 475
  autopsy 475–6
    radiology 32, 476
  axonal injury, diffuse 470
  bites/bite marks 474, 476, 528
  bone diseases *vs* fractures due to 468–9
  bruising *see under* bruises/bruising
  burns 474
  cardiopulmonary resuscitation *vs* 467–8, 472–3
  chest wall damage 467–8
  cigarette burns 474
  definitions 461
  duodenum rupture 472
  eye/ear/mouth injuries 472–3, 476
  facial injuries 178, 463, 472–3
  falls *vs* 465
  fractures 224
    dating in 468
    spiral 467
  general aspects 461–2
  haemoperitoneum 463, 475
  hair pulled, in 474–5
  head injuries 193, 469–71
    brain damage 192, 205, 465, 471, 476
    extradural haemorrhage 471
    mechanisms 469–70, 471
    moving 471
    subarachnoid haemorrhage 197, 470–1
    subdural haemorrhage 192, 193, 469, 470, 471
  injuries 462–3
    miscellaneous 475–6
  liver damage 232, 472
  medico-legal aspects 471
  modes of death 462
  non-fatal 462
  periosteum shearing 466
  psychological 461
  radiography/radiology 32, 476
  rib fractures 224, 467–8
  sample collection 476
  scalds 474
  scalp 463
  scene of death visit 475
  sexual 461
    anal dilatation 423–4, 428
    anal rapes 427, 429
    internal examination 425
    rape 425–6, 427, 429
  shaking and subdural haemorrhage 192, 193, 469, 470, 471
  skeletal damage 464–8
    chest cage 224, 467–8
    dating of fractures 468
    limb fractures 466–7
    rib fractures 224, 467–8
    skull fractures 185, 464–6
    see also skull fractures
  subdural haemorrhage 192, 193, 469, 470, 471
  teeth, avulsion/breakage 473
  visceral injury 472
children
  accidental hanging 389
  accidental injury, child abuse concerns 462
  adult-induced injury *see* child abuse
  age estimation from skeletal remains 118
  air gun injuries 271
  carbon monoxide poisoning in fires 322
  cerebral oedema 212
  hanging in 385
  head injury from falls 180
  neglect 412, 417
  pedestrians, injuries 292–3
  scalds 314, 474
  seatbelts 289
  skull fractures 180, 185, 466
  suffocation 358
  vulnerability in vehicular accidents 289–90
  see also infant
china, stab/slash wounds from 156, 164
chisel wounds 153
chlordiazepoxide 573
chloride 89, 317, 404
chloroform 597
chlorophyll 408
chlorpromazine 548
choke-holds *see* arm-locks
choking 361–3
  causes 361–3
  deaths in custody 307
  definition 353
  haemorrhage 361
  infanticide 443–4
  suicide 361
cholinesterase, inhibition 567
cigarette burns, child abuse 474

circle of Willis 197, 198, 512, 513
  see also berry aneurysm
circumcision 104, 105
cirrhosis of liver 558
cisternal puncture 548
civil cases, disease and trauma 519, 520
clavicle
  fractures, in child abuse 468
  medial end, epiphyseal fusion 118
clay, modelling 128
clinical toxicologist 541
Clostridium perfringens 57, 434
  adipocere formation 71
clothing
  body secretions on 8
  body temperature measurement and 82
  effect on cooling curve 81
  examination
    sexual offences 422, 423
    solvent abuse 597–8
  inadequate, hypothermia and 416
  pathologists' 4
  position in suicidal knife wounds 8, 239
  race/ethnic characteristics, in identification 104
  removal methods 8
  ripped, in deaths from lightning 336
  road traffic accidents 8, 296
  scene of death 8
  soiling/burning in shotgun wounds 8, 250, 251
  spontaneous combustion 323
  vehicular accidents 296
clotting, blood 58, 341–2, 356
clubbing injuries, in human rights abuse 304
coagulative myocytolysis 498
coal-mining, carbon monoxide poisoning 561
cocaine
  blood/tissue levels 580
  deaths associated 581–2
  toxicological sample preservation 546
coffin
  decomposition of bodies 69, 72
  exhumation 37, 38
  measurement 103
'cogitation' *see* hypostasis, post-mortem
cold
  injury caused by 414–18
    see also hypothermia
  local injuries 416, 419
  muscle stiffening 63
collagen 168, 342
collateral circulation 496
collodiaphyseal angle 113
colon
  perforation 517
  removal at autopsy 21
colour of skin *see* hypostasis; skin, colour
coma, epidural haemorrhage 190, 191
combustion, spontaneous 323–4
'commando punch' 371
comparisons, technique of 106
complications of injuries 339–51
  adult respiratory distress syndrome 345
  air embolism 346–8
  disseminated intravascular coagulation 345–6
  fat and bone marrow embolism 343–5
  haemorrhage 339–40
  infections 340
  pulmonary embolism 340–3
  renal failure 345
  subendocardial haemorrhage 348–9
  suprarenal haemorrhage 349
  see also under individual complications
compression wave, explosions 274
computed tomography (CT) 31
  gunshot wound 261
computer graphic techniques, facial appearance reconstruction 128–9
concealment
  birth 440, 441, 442
  mummification in 72
conception, products of 434, 436, 442
concussion 209–10
  diffuse neuronal/axonal injury 210
  extradural haemorrhage and 190, 191
conductors, electrical 326
congenital deformities 12
congenital heart disease 455
congestion, in 'asphyxia' 356
  traumatic 363
conjunctival haemorrhage 472
containers, sample collection 30, 545
continuity of evidence 8, 544
contrecoup lesions 179, 189, 207–9
  black eyes and 179
  contusion 207, 209
  subdural haemorrhage 193
  vacuum effect 205, 207
contusions 59, 137, 143–50
  cerebral 206
  contrecoup 207, 209
  definition 143–9
  firearm wounds 270
  see also bruises/bruising
conus artery 496
conversion tables 603–4
cooling, post-mortem 79–80
  'core' temperature 79
  curves 79
    factors affecting 79, 80–1
    nomogram method 83
    plateau 79–80, 82–3
  historical aspects 78–9
  measurement methods 81–2
  multiple-site serial measurements 83

cooling, post-mortem (*contd.*)
  rectal temperature 79–80
  'temperature ratio' concept 83, 87
  see also post-mortem interval; temperature, body
copper deficiency 469
Coppolino case 550
corium (dermis) 138
cornea, signs of death 54
corneal reflex 54
cornuae *see under* larynx
coronary artery
  examination at autopsy 26, 46
  in hypertensive heart disease 506
  luminal size estimation 493
  occlusion 494, 495
  spasm 503
  stenosis 493, 503
  thrombosis 285, 495, 496, 503
    without infarction 504
  vessel types 505–6
  wounds, survival period after 170
coronary artery disease 493
coronary atherosclerosis 493–5
  atheromatous plaques 493–4, 495
    detection 493–4
  complications 494–5
    haemorrhage 494–5
    thrombosis 285, 495, 496, 503
    ulcerated plaques 494
  diffuse disease 494
  in elderly 509
  haemorrhage 494–5
  hypertension and 505
  mechanism/pathology 493–4
  sites 495
  trauma relationship 518–19
    'beyond reasonable doubt' test 518
    time association 518
    violence/threats precipitating 518–19
  see also myocardial infarction
coronary insufficiency 505
  cause of death 503–5
coroner 3
corpus callosum, torn 206–7
corrosive acids, deaths from 587–8
costochondral junction, in age estimation 120
cot deaths *see* sudden infant death syndrome (SIDS)
coulombs 328
counselling, in sudden infant death syndrome 456, 457
'country guns' 248
coup lesions 207–9
coverings, body, effect on cooling curve 81
crack cocaine 581
cranial suture closure 120–1
craniometry 110–11, 124–5
cranium, anatomy 181
cratering, bullet wounds 266
creatinine 89
crenation
  edges of bites 75, 76
  shotgun wound edges 251, 253
crib deaths *see* sudden infant death syndrome (SIDS)
cribiform plate, ruptured/fractured 189
cricoid cartilage 376
  displacement 378
  in strangulation 376, 378, 382
crocodile-skin 331, 332
crown–heel length 448, 449
crown–rump length 118
crucifixion 365
crushing, crowds, traumatic 'asphyxia' in 363
crushing abrasions 139–41
crush injuries
  abdominal 229
  bleeding absent 339
crustaceans 76, 78
custody, deaths in 305–8, 353, 384, 386
  alcohol intoxication 307
  blunt injuries 306–7
  drugs 581
  hanging 308, 384, 386
  natural causes 308
  suicide 308
  see also torture
cutis anserina 395
cuts 154–5, 444
  human rights abuse 303–4
cut throats 236–7, 238
  homicidal 236, 237
CV-One syndrome 200
cyanide
  in fumes from fires 322
  inhalation 586
  poisoning 585–7
    autopsy findings 57, 585–6
    homicide 585
    sample collection 586–7
    suicide 585
    toxicology 586–7
  sample preservation 546
  smell 586
  spurious formation in blood 587
cyanmethaemoglobin 586
cyanosis 356
  carbon monoxide poisoning 562
  electric fatalities 333
  fatal pressure on neck 368
  hypothermia 417
  immersion deaths 400
  traumatic 'asphyxia' 364
cytochrome P450 572
cytological examination, sex determination 104

dagger, 'stiletto' 231
damp conditions, hypothermia in 415
dartos muscle 61
dating
  bruises, variables affecting 147
  emboli/thrombi 342–3, 520
  fractures 468
  myocardial infarction 496–7
  pulmonary embolism 342–3, 520
  skeletal remains 126–8
    environment effect 126
    physical appearance 126–7
    physical tests 127
    serological/chemical tests 127–8
  subdural haemorrhage 195–6, 197
  wounds 166–9, 170, 171
death
  brain *see* brain death
  cause 9–10
    mode *vs* 55
    obscure autopsy 45–7
  cellular 52, 63–4
  in custody *see* custody, deaths in
  fate of body after 53
  gunshot/explosive, radiology 31
  indications 53–5
  interval between injuries and 169–71
  manner 55
  mode 46–7, 55
  pathophysiology 52–97
  scene of *see* scene of death
  signs 53–5
  somatic 52, 53
  sudden *see* sudden death
  time of/since *see* post-mortem interval
  traumatic 8
  types 52–3
death certificates 10, 55, 341
death certification, by physicians 493
decalcification 203
decapitation
  hanging 384
  railway suicides 297
  sexual offences 423
  'tail-gating ' motorcycle accident 290
deceleration 281
  G forces 281
  seatbelts effect 288
deceleration injuries 214, 215, 281, 287
  aorta 228, 229
  cerebral 205, 206
  liver 232
  lungs 226
  spinal 214, 215
  subdural haemorrhage 193, 470
  tissue damage 281
  see also vehicular injuries
decomposition 63–78
  adipocere formation *see* adipocere
  buried bodies 68–9, 126
  changes in 64
  charting teeth and 536–7
  coffin 69, 72
  damage by predators 67, 69, 73–6
  dry (mummification) *see* mummification
  identification of corpse 99
  immersed bodies 67–8, 396–7, 398–9
  infant 441, 442–3, 446
  maggot infestation 65, 66, 69
  mixed forms in body 72
  putrefaction *see* putrefaction
  tattoos in identity establishment 100–1
  time scale 64
decompression
  aircraft 357
  divers 488, 489
  physical damage 488–9
decompression chamber, autopsy in 489
decompression sickness 488
deep vein thrombosis 341, 513
  dating 342–3, 520
  in hypothermia 418
  medico-legal aspects 341–2, 520
  sudden death from pulmonary embolism 513, 520
defence wounds 165–6, 167, 168
defibrillator, post-mortem artefacts 12, 36, 41
degenerative disease, age estimation 122
dehydration, in neglect/starvation 414
delayed deaths 339, 563
  see also complications of injuries
dental records 535
  diagrammatic/anatomical charts 535–6
  features to record 537
  odontogram 536
dentine, staining 538
dentistry, forensic 527–40
  bite marks 474, 527–32
    see also bites/bite marks
  identification of dead 532–8
    age determination 122, 534–5
    charting teeth 535–8
    general/reconstructive identity 533–5
    human origin 533
    mass disasters 44
    personal identity/comparative 535
    positive match 537
    race determination 533–4
    sex determination 533
  need for expert odontologists 527
  photography role 537
  pink teeth 538
  radiography role 537
  see also teeth

dentures
  choking due to 361
  examination at autopsy 15
  identification by 537
dermis 138
diabetes mellitus, post-mortem interval estimation 90
'Diachylon' 434
diastasis of suture lines 185, 187, 188
diatoms, types/species 407–8
diatom techniques *see under* drowning
diazepam 573
DIC (disseminated intravascular coagulation) 345–6, 432, 436, 437
diethylene 589
digestion 85, 86–8
diphtheria 361, 508, 515
diploe 181
Diptera 76
diquat poisoning 566
disasters *see* mass disasters
discoloration *see* skin, colour
discriminant function analysis 110
disease, natural
  deaths before immersion 397–8
  deaths in custody and 308
  deaths whilst in water 397–8
  trauma relationship 198–9, 517–20
  vehicular injuries and 295–6
  see also individual diseases
dismembered corpses, identification 99
dissection techniques 16–23
  see also autopsy, medico-legal
disseminated intravascular coagulation (DIC) 345–6, 432, 436, 437
divers/diving 488
  'bends' 346
  carbon monoxide poisoning 561
  pressure changes/effects 488, 489
DNA 105
  hair root cells 99
  hypervariable regions 106
  probes 106
  redundant segments 106
DNA fingerprinting/profiling 105–6
  dating of bones 127
  'hit-and-run' accidents 296
  identity establishment 105–6
  infanticide 440
  method 106
  samples/controls 106, 604
  seminal/vaginal fluids 106, 424, 428
  sex determination 104, 533
doctors, fingerprints 101
dogs, post-mortem damage 74
dothiepin 573
Down's syndrome 455
drag marks 140
drivers, medical checks 296
drowning 10, 401–2
  in acute alcoholism 557
  adipocere formation 406
  autopsy signs 402–4
    froth in air passages 402–3, 404
    other organs 404
    overinflation of lung 403, 404
  barotrauma and 489
  cadaveric spasm and 63
  chemical changes in blood 404–5
  decomposition of bodies 67–8
  diagnosis and diatom test 406–8, 602–3
    problems/criticisms 408
    techniques 406–7, 408
  'dry-lung' 400, 403
  electrolytes changes in 401, 404–5
  in epilepsy 514
  freshwater 401
  haemodilution 401, 404
  haemorrhage 404
  histological changes 405–6
  human rights abuse 304
  infanticide 445
  lungs 402–3, 404, 405
  mechanism of death 401–2
  petechial haemorrhage 403
  pulmonary haemorrhage 403
  seawater 401–2
  sodium, release in 401
  see also immersion deaths
drug(s)
  alcohol absorption and 554
  deaths in custody and 307, 581
  illegal abortions 434–5
  overdoses, in hospital 543
  poisoning/overdose *see under* poisoning
drug addicts/abuse 577–84
  autopsy appearances 578–9
  autopsy risks for pathologist 579
  intravenous 577, 582
  'mainlining' 577, 582
  see also specific drugs
drugs of dependence 577
'dum-dum' bullet 137, 245
duodenal ulcer, perforated 517
duodenum
  rupture, in child abuse 472
  transection 230
Dupertuis and Hadden's table, stature estimation 116
dura mater 189
dust blasting 275
Dwight's table, sex determination 112
dysbarism 488–91
  autopsy 489–90

  definition 488
  physical damage 488–9

ear
  barotrauma 488, 489
  bites 180
  blast injuries 275
  bruise below 148
  haemorrhage, in drowning 404
  injuries 179–80
    child abuse 473
    human rights abuse 304
ear crepitance test 447
'ecbolics' 434
ecchymoses 144, 353
  see also bruises/bruising; petechial haemorrhage
'ecstasy' 581
ectopic pregnancy 436
elderly
  carbon monoxide levels in blood 563
  chronic subdural haematoma 194, 195
  falls 180
  hypothermia 415, 416–17
  neglect 412
  subdural haemorrhage 192–3
  sudden death (senile myocardium) 509
  tearing of gastrointestinal tract 517
electrical conduit pipe weapon 248
electrical fatalities 326–38
  in bathroom 326, 333–4, 334–5, 400, 405
  cardiac arrhythmias 327, 328, 329, 333
  cutaneous burns/marks 329–32
    areola of blanched skin 331
    charring 331
    earthing/'grounding' lesions 331
    firm contact (blisters) 329, 330, 331
    high-voltage 331
    histology 333
    spark burn 329, 330, 331
    temperature and heat generation 329–30
  cyanosis 333
  effects on muscle 328–9
  electrical considerations 327–8
  histological appearances 333
  homicide 334–6
  hypostasis, post-mortem 56, 333
  internal appearances 333
  metallic traces in marks 332–3, 335
  mode of death 329
  non-electrical trauma in 329
  pathway for current 326, 327
  patterned electric marks 332, 336
  petechial haemorrhage 333
  physical factors 327
  post-mortem hypostasis 56, 333
  requirements for/events in 327, 328, 329
  resistance of skin 328
  rigor mortis timing 61
  suicides 239, 326, 334
  time, importance 327, 328, 331
electrical torture, in human rights abuse 304
electric current 327–8, 328
  alternating (AC)/direct (DC) 327–8
  application to genitals 391
  frequency 327
  mA leading to death 328
electric shock 326, 327
  tolerance to 327
electrocution 326
  see also electrical fatalities
electrolytes, changes in drowning 401, 404–5
electronic cameras, use at autopsy 32–3
electrons 326, 327
emaciation 412, 413
emboli, sources 343
embolism 341
  air *see* air embolism
  amniotic fluid 436–7
  bone marrow 344
  fat *see* fat embolism
  foreign body 344–5
  pulmonary *see* pulmonary embolism
'emmenagogues' 434
emphysema
  in barotrauma/dysbarism 490
  interstitial 489
emphysema aquosum 403, 405
enamel pearls, on teeth 534
endocarditis, post-rheumatic 582
endothelial proliferation 342
endotracheal tube, position 481
energy transfer, kinetic 137, 245, 246
English Infanticide Act (1938) 439
entomology 76–8
  collection of material 78
  insect succession 76
  insect types 76
  rationale 76
  see also insects
enzyme histochemistry 498–9, 500, 501–2
enzymes
  dating of wounds 169
  in myocardial infarction 498
eosinophilia, myocardial infarction 498, 500
epicardial petechiae 333
epicranial aponeurosis 174–5
epidermis 138
  tags 139
epidural anaesthesia, deaths related 485
epidural haemorrhage *see* extradural (epidural) haemorrhage
epiglottis, haemorrhages 378
epiglottitis 361, 515

epilation 474–5
epilepsy
  autopsy 514
  sudden death 514–15
    Purkinje cells in 514–15
  traumatic/post-traumatic 188–9, 514
epiphyseal fusion, age estimation 118
epiphyseal injuries/lesions, in child abuse 466, 467
equipment
  failure, surgical/anaesthetic deaths 484, 485
  internal examination 18
  scene of death 4–5
Erdheim's degeneration 510
ergometrine, illegal abortions 434
erythema, in burns 315
erythema ab igne 312
ethnic characteristics
  hair structure 99, 105
  identity establishment 105
  scars 102
  skin colour 99
  stature estimation from skeletal remains 114
  tattoos 100, 105
  see also race
ethyl alcohol *see* alcohol
ethylene 589
ethylene chloride 597
ethylene glycol poisoning 589
Europe, sudden infant death syndrome in 452
execution, extrajudicial 305
exertion, sudden death and 519
exhumation 36–9, 68
  arsenical poisoning 590
  coffin identification 37, 38
  legal procedures 37
  photography 37
  poisoning 38–9
  reasons for 37
  transportation 38
exit wounds *see* rifled weapons, wounds; shotgun wounds
explosions
  anaesthetic-related deaths 485
  deaths from 274–7
    autopsy in 276–7
    body fragments 275–6
    identification 276–7
    missile injuries 275–6
    nature of trauma 274–5
    radiography/radiology 31, 276
explosives, detonation, carbon monoxide poisoning 563
exposure, deaths from 419
external examination at autopsy 11–16
  age/race/sex 11
  anatomical diagrams 14
  external marks 12
  procedures/routine 11–16
  report 34
  vomit/secretions 12–13
  weight/length 11–12
external jugular occlusion 368
extradural (epidural) haemorrhage 190–2
  blood volume 190
  in child abuse 471
  clinical signs and latent interval 190–1, 198
  formation and sites 190
  heat haematoma *vs* 192, 317
  incidence/mortality rate 190
  'lucid' (latent) period 190, 191
  medico-legal aspects 191–2, 198
  in posterior fossa 191–2
  post-mortem artefacts resembling 192
extradural space, heat haematoma 317
eye(s)
  in age determination 104
  black *see* black eye
  colour, identity establishment 99
  examination 14–15
  globe tension decrease at death 54
  injury, in child abuse 472–3, 476
  laceration of eyebrow 151
  petechial haemorrhages 354, 371
  signs of death 54–5
  in traumatic 'asphyxia' 364
eyebrow, injuries 151, 178
eyelids, sign of death 54

face/facial appearance
  abrasions 178
  bloating in putrefaction 65
  bruising, in child abuse 463
  congestion 370
  examination, in solvent abuse 598
  external examination 13
  features in adipocere formation 71, 72
  in hanging 384
  hypothermia 416, 417
  identity establishment 99
  in ligature strangulation 380
  petechial haemorrhages 354, 355, 356
  pressure marks 358, 360
  in putrefaction 65
  racial/ethnic characteristics 105
  recognition in mummification 72
  reconstruction from skull 128–9
facial injuries 178–80
  brain removal at autopsy 22
  in child abuse 178, 463, 472–3
  from kicking 149, 178–9
  self-inflicted 242, 243
  in vehicular accidents 282, 284
faeces, purging, in putrefaction 65
faecoliths 414
'Falanga' 303
falls 180–1

  acute alcoholism and 557
  contrecoup lesions 208
  deaths in custody 307
  elderly 180
  head injuries 181, 471
  head injury in children 180, 187, 465
  from height
    brain damage in children 465
    cadaveric spasm and 63
    injuries 181, 240
    injury to aorta 228
    onto feet 181, 214, 240
    onto head 181, 471
    pelvic injuries 233
    skull fractures 184, 187
    suicidal 240
  hip, dislocation in 233
  occipital, skull fractures 184
  scalp injuries 177–8
  spinal injuries 181, 214
  from standing position, injuries 180
  trajectory 181
falx, mechanism of brain damage 207
fat
  alcohol insolubility 552
  effect on cooling curve 80
  putrefaction 67
fat embolism 343–5
  autopsy findings 344
  in barotrauma/dysbarism 490
  cerebral 343
  in chronic alcoholism 558
  conditions associated 343
  marrow source 344
  petechial haemorrhage 344
  pulmonary 343
  systemic 343–4
feet, beating of soles 303
femoral heads, sexing from 112
femoral vein, blood samples 19–20, 547
femur
  fractures, in pedestrians 293
  sexing from 112–13
    'mathematical sexing' 112
    shaft and angle 112–13
  stature estimation 114, 115, 117
fetishism 391
fetus
  age estimation from skeletal remains 117–18, 449
  deaths, cocaine use 581
  decomposition 64
  delivery after death 65
  maturity estimation, ossification centres 444, 445
  measurements 448, 449
  mummified 72–3
  skeletal age 449
  see also infanticide; stillbirth
fibrin
  amniotic fluid embolism 437
  dating of emboli/thrombi 342
  staining 342, 601
fibrinogen 346
fibrinolysis 346
fibroblasts 168, 195
fibula, fractures 293
Fiedler's myocarditis 508
'fight or flight' response 170
fights, dynamic nature 154
finger, desquamated casts 102
fingernail
  abrasions 141
    concavity of mark 141, 373
    curvature/orientation 141
    ligature strangulation 382
    manual strangulation 141, 372–3, 373
  examination, in sexual offences 423
  putrefaction 65
  scraping/clippings 373, 423
finger pads, scarring 102
fingerprints
  desquamated casts 102
  identity establishment 101–2
    putrefying corpse 39, 102
  immersion deaths 395, 396
fingertips, bruising by 148
  in child abuse 144, 462, 463
fire
  anaesthetic-related deaths 485
  ante-mortem *vs* post-mortem injury 7
  autopsy appearance after 315–16
  carbon monoxide poisoning 320–2, 561, 564
  cause of death 319–20, 322
  deaths, carboxyhaemoglobin levels 562
  evidence of being alive at 322, 564
  flash 320, 322, 564
  heat haematoma 191, 192, 317
  injuries to body 7
  suffocation in 357
  vehicular accidents 288, 297
  see also burns; smoke
firearms 246–8
  air guns 271
  calibre 247
  country guns 248
  muzzle, impression/patterned abrasion 143, 250
  range of discharge estimation 270–1
  rifled weapons 248
  shotguns (smooth-bore) 246–8
  veterinary pistols and 'stud-guns' 272–4
  wounds *see* gunshot wounds
  see also bullets; rifled weapons; shotguns
fire extinguishers 595
flail chest 223, 224, 294

'flaying' injury 152, 176
  scalp 176
  vehicular accidents 294
flotation test 445–6, 447
fluid, body *see* body secretions
fluorescence 127
  myocardial infarction 500–1
  sex determination 104
fluorocarbons, abuse 597
fontanelles, anterior/posterior 181
food
  alcohol absorption and 553, 554
  bite marks 527–8
  choking due to 361–3, 362
  diatoms in 408
  lack, hypothermia and 416
  total deprivation 412
foot, beating of sole 303
footprints
  abrasions/bruises 144, 150
  identity establishment 101–2
foramen magnum, ring fracture around 181, 185, 214, 283, 290
force
  assessment, stab wounds 161–3
  intensity, in wounding 136
  in skull fractures 186–7
forehead, sex determination from skull 109
forehead injury, movement of bruises 146
foreign bodies
  abdominal injuries 232–3
  choking due to 361
  embolism 344–5
  in gut 232
  respiratory obstruction 515
  sand/mud in immersion deaths 396, 404
  in vagina/rectum 426
forensic autopsy *see* autopsy, medico-legal
formalin fixation 28, 29, 30
formalin inflation, lungs 24–5
fractures
  ante-mortem *vs* post-mortem 182, 344
  'bumper' 293
  child abuse 224
  dating 468
    in child abuse 468
  heat, post-mortem artefacts 36
  infant, bone diseases causing 468–9
  'motor cyclist's' 185, 188, 290
  spiral, in child abuse 467
  undertaker's 21
  in vehicular accidents 283, 284, 293
  see also other specific fractures; rib fractures; skull fractures
frenulum of lip, rupture/tearing 178, 473
friction burns *see* abrasions, brush
frontal lobe
  'burst lobe' 207
  contrecoup lesions 208, 209
frontal sinus identification 124–5
frostbite 416, 419
frothy fluid, from mouth/nostrils
  burns 316
  drowning 402–3, 404
  morphine addicts 578
  sudden infant death syndrome 454
fruit, bite mark preservation 528
fume inhalation 319–22
  causes of death 319–20, 322
  toxic substances 322
  see also smoke

gagging 352, 353, 360–1
gags, in sexual asphyxia 391
galea aponeurotica 174–5
gall bladder
  examination at autopsy 27
  in starvation 414
garrotting 353, 379
  see also strangulation, ligature
gas appliances, carbon monoxide poisoning 561–2
gas chromatography 598
  autopsy in air embolism 347, 348
gas formation 66
  decomposition of immersed bodies 68
  putrefaction 64, 65, 435
gasoline 560, 597
gastrectomy 553
gastric contents
  agonal regurgitation 547
  aspiration 362–3, 557
    in anaesthesia 484–5
    medico-legal aspects 596
    sudden infant death syndrome 453, 454
  examination at autopsy 24
  mouth examination 15
  resuscitation artefacts 41
  sample collection 27, 545, 549
    containers 545
gastric emptying
  alcohol absorption and 553, 554
  factors effecting 87–8
  investigation methods 87
  post-mortem interval estimation 83, 85, 87–8
gastric erosions, acute, in hypothermia 418
gastric lining, alcohol irritation 553
gastric transit time 88
gastroenterostomy 553
gastrointestinal haemorrhage, sudden death 516
gastrointestinal tract
  barbiturate poisoning 574
  blast injuries 275
  foreign bodies 232
  paraquat poisoning 566
  perforation 233, 517

  tearing in elderly persons 517
  see also intestine; stomach
gastromalacia 36
genitals
  examination at autopsy 16
  injuries 233
    sexual offences 423–4
  sample collection, in sexual offences 423, 424
  self-mutilation 240, 241
genitourinary system, sudden death 517
G forces 281–2, 288
glass, wounds from 155, 156, 163–4
'glassing' 155
glottis, examination at autopsy 24
glucose 89–90, 575
'glue-sniffing' 595, 596
γ-glutamyl transpeptidase 558
glycine 127
glycogen 61, 62
  depletion 63
glycol poisoning 589
glycophorin A 40
'goose-flesh' appearance 61, 395
Gough's test 511
grade crossings, injuries at 297
grain silo, suffocation in 357–8
gramoxone 566
granulation tissue 168
gravel rash 138
gravity 281
grazes 137, 138, 139, 142
  see also abrasions
grease ring 264–5
great vessels, injuries 228–9
gunpowder 248
gunshot residues, clothing 8
gunshot wounds 249–77
  air guns 271–2
  ammunition *see* ammunition; bullets
  carbon monoxide levels 250, 252, 267
  computed tomography (CT) 261
  contusions 270
  defence wounds 166
  exit 267–70
  external bleeding absent 225
  internal organs 270
  mechanics 245–6, 268
  radiology 31, 276
  range of discharge estimation 270–1
  scars 102
  shock wave from missile 246
  suicidal 239, 252, 253, 254, 255
  velocity of missile 246, 258, 268
  weapons 246–8
  see also missile injuries; rifled weapons, wounds; shotgun wounds
Gustafson technique 105

Haase's rule 448
haematoidin 147
haematoma 146
  heat 191, 192, 317, 318, 319
  peri-orbital *see* black eye
  perirenal 510
  subcapsular liver 286, 340
haematoxylin basic fuchsin-picric acid (HBFP) stain 500
haematoxylin–eosin (H and E) autofluorescence 601
haematoxylin-eosin (HE) staining 500
haemodilution, in drowning 401, 404
haemoglobin 57
  carbon monoxide saturation 563–4
  dating of bones 127
  degradation in bruises 145, 146
  glycosylated 546
haemolysis, in subarachnoid haemorrhage 197
haemopericardium 228, 229, 502
  blood sample collection 547
  dissecting aortic aneurysm 510
haemoperitoneum, child abuse 463, 475
Haemophilus influenzae 361, 515
haemoptysis, sudden death from 516
haemorrhage 339–40
  abdominal 230, 232, 233
  adrenal 349, 513
  autopsy in strangulation case 373
  behind oesophagus, hypostasis artefact 59, 150, 375
  brainstem 207, 214, 339
  chest 224–5
  choking due to 361
  in coronary atherosclerosis 494–5
  delayed deaths due to 339–40
  drowning 404
  epidural *see* extradural (epidural) haemorrhage
  epiglottis 378
  infections, due to 340
  intracerebral *see* intracerebral haemorrhage
  intracranial *see* intracranial haemorrhage
  laryngeal, in manual strangulation 377, 378
  laryngeal fractures 378, 380
  mesenteric 230–1, 233
  perirenal 232, 510
  perivascular 418
  petechial *see* petechial haemorrhage
  post-mortem 225, 233, 340
  post-mortem artefacts 57, 59, 150, 375
  pulmonary 275, 286, 403
  Sheehan's 348
  spinal cord 216
  subarachnoid *see* subarachnoid haemorrhage
  subdural *see* subdural haemorrhage
  subendocardial 285, 348–9, 590
  subintimal 494
  suprarenal 232, 295, 349
  see also bleeding
haemorrhagic diatheses 145, 354, 516

haemorrhagic shock, zonal lesions 499
haemosiderin 147, 167, 168
  dating emboli/thrombi 342
  subarachnoid haemorrhage 197, 512
haemostasis, failure 339–40
haemothorax 224, 225, 258, 340
  blood sample collection 547
  source 224
hair
  arsenical poisoning 591
  clubbed 251, 262, 315
  colour
    age determination 104
    identity establishment 99
  in copper deficiency 469
  DNA profile 99
    sample 604
  drugs of dependence 577
  effect on skull fractures 184
  loss, in thallium poisoning 592
  microscopic examination 99–100
  pubic *see* pubic hair
  pulled, in child abuse 474–5
  racial/ethnic characteristics 99, 105
  scalp injuries and 174, 175, 176
  shaving 14
  singeing 315
    deaths from lightning 337
    shotgun wounds 251, 253
    wounds from rifled weapons 262
  structure, identity establishment 99–100
  texture, in identification 105
  wetting 17
hair follicles, rigor mortis 61
hallucinogens, deaths from 580
halogenated hydrocarbons 597
halothane 486
handedness, determination 372
handguns 248
hands
  examination 12, 16
    sexual offences 423
  self-inflicted injuries 242
  sudden infant death syndrome 453
hanging 383–9
  accidental 384, 389
  autoerotic procedures and 390, 392
  autopsy appearances 387–8
  carotid arteries damage 388
  cervical spine damage 215, 384, 385
  in children 385
  in custody 308, 384, 386
  decapitation 384
  homicide 389
  hypostasis, post-mortem 56, 387–8
  judicial 383–4
  knots 384, 385, 386
  laryngeal fracture 388
  marks 308, 381–2, 386–7
    position 387
  mechanism of death 388
  methods 384–6
  petechial haemorrhage 388
  serotonin from ligature mark 169
  signs 371
  suicidal 240, 384, 388
head
  in air embolism 347, 435–6
  rotational trauma 200–4
  weight 187
head injuries 174–221
  air-gun 271, 272
  in boxers 204
  cerebral oedema *see* cerebral oedema
  child abuse *see* child abuse
  closed 206
  deaths in custody 306–7
  examination at autopsy 14
  falls 181, 471
  infanticide 444–5
  intracranial *see* intracranial injuries
  nature of impact, and haemorrhages 198
  petechial haemorrhages 14
  post-mortem 176
  post-mortem heat injuries *vs* 317, 318
  sexual offences 423
  shotgun wounds 258
  subarachnoid haemorrhage and 196–8
  survival period after 170
  vehicular accidents 283
    drivers 284
    motor cyclists 188
    pedestrians 294
head-moulding, in childbirth 197
heart
  air embolism 346, 347
  anatomy 223
  avulsion 228, 285
  in barotrauma/dysbarism 490
  bruising, in vehicular accidents 285
  conducting system, sudden death due to 504–5
  in drowning 404
  in electrocution 327
  examination at autopsy 25–7
  incisions at autopsy 25
  injuries 226–8
    blunt 227
    cardiac tamponade 227, 228
    stab wounds 223, 226–7, 229, 238
    valves 229
  orientation 223
  'pendulum' effect 285
  petechiae 354
  putrefaction 66–7

  rhythm disturbances *see* cardiac arrhythmias
  right, engorgement 356
  rigor mortis 61
  rupture 502
  sudden infant death syndrome 454
  suicidal stab wounds 238
  weight 26, 505, 507, 509
    cardiomyopathies 507
  wounds, survival period after 170
  see also entries beginning cardiac; left ventricle; right ventricle
heat
  convection/radiation 314–15
  flexures 317
  generation, cutaneous electric marks 329–30
  injuries 312
    see also burns; fire; scalds
  loss by body 79, 80–1, 81
    see also cooling, post-mortem
  muscle stiffening 63
  radiant, injuries 312
  regulation, alcohol effect 553
heat fractures, post-mortem artefacts 36
heat haematoma 191, 192, 317, 318, 319
heating, inadequate, hypothermia and 416
heating appliances, carbon monoxide poisoning 560, 561
heavy metal poisoning 549, 585, 589–92
height
  body 603
  estimation from skeletal remains 114–16
  identity establishment 103–4
  see also length
Heimlich manoeuvre, resuscitation artefacts 41
helmets, cyclists/motorcyclists 290, 291
Henry's law 488
Henssge's nomogram 83, 84–5, 86
hepatitis B, risk 4, 10–11, 544, 579
herbicides, poisoning 566–7
hermaphroditism 104
hernia 516
heroin 580
hexylene glycol 589
'hide-and-die' syndrome 418–19
Higginson syringe 432, 433
high-velocity projectiles/missiles 246, 270
high-voltage electrical burns 331
hip, dislocation in falls 233
Hippocratic facies 413–14
histamine, in dating wounds 169
histochemistry 498, 499, 500
  enzyme 498, 501–2
  myocardial infarction 498–9, 500
  wounds 168–9
histology
  asphyxia diagnosis 357
  cerebral hypoxia 212–13
  changes after wound infliction 167–8
  sample collection 30
  sudden infant death syndrome 455, 456
  surgical/anaesthetic-related deaths 483, 485, 486
history of case, in autopsy 9–10, 453
'hit-and-run' accidents 8, 294, 296
HIV, infectivity 11, 579
HIV infection
  pre-autopsy test 10
  risk at autopsy 4, 10–11, 579
  risk in toxicology samples 544
  tuberculosis and sudden death in 516
Home Office Pathologists 4
homicide
  acute alcoholism 557
  air embolism in 346
  in bathrooms 400
  cyanide poisoning 585
  electrocution 334–6
  'faking' of vehicular accidents 297
  hanging 389
  injuries before immersion in water 398
  insulin poisoning 574–5
  ligature strangulation 379
  scalp lacerations 152
  second autopsy 3–4
  sexual offences 421, 423, 424
  shotgun wounds 250, 252
  slash wounds 155
  smothering 358
  stab wounds 155, 160, 161, 163
  suicide *vs* 235, 236, 238, 239
  vehicular 297
  see also strangulation
homosexual, anus in 427, 429
homosexual offences 428–9
hospital, poisoning deaths 543
housefly 77
humane killers, weapons 272–4
human immunodeficiency virus *see* HIV infection
human rights abuse 301–11
  evidence of fatal abuse 301–5
  extrajudicial execution 305
  organizations 301, 302
  see also custody, deaths in
Human Tissue Act 1961 3
humeral head, sexing from 112
humerus
  sexing from 113
  stature estimation 117
humidity, effect on cooling curve 81
hunting accident, twelve-bore shotguns 259
hyaline membrane 196
hydrocarbons, halogenated 597
hydrochloric acid 588
hydrogen cyanide 586
hydrogen peroxide 567
hydrolysis, fat 70, 71
hydrostatic test 445–6, 447

hydroxyproline 127
hygiene, examination at autopsy 12
hymen 424, 425
hyoid bone 376–7, 377
  fracture *see* larynx, fractures
  in manual strangulation 377
hyperaemia, in rifled weapon wounds 260, 262
hypercapnia, in solvent abuse 596
hyperchromasia 500
hyperextension injuries 214–15
hyperflexion injuries 214–15, 284
hyperkeratosis 414
hyperpigmentation, human rights abuse 303
hypersensitivity
  anaesthetics 485
  aspirin 571
hypertension
  evidence 505–6
  intracerebral haemorrhage 207
hypertensive heart disease 505–6
hyperthermia
  effect on cooling curve 80
  malignant 485–6
hypertrophic obstructive cardiomyopathy 506
hypocalcaemia 588, 589
hypoproteinaemia 414
hypostasis, post-mortem 55–60, 333
  abdominal catastrophes *vs* 517
  artefact, haemorrhage behind oesophagus 59, 150, 375
  'asphyxia' confusion 57, 58, 356
  blotchy patches 55
  bruising *vs* 59–60
  cherry-pink, carbon monoxide poisoning 57, 321, 562
  clotting of blood 58
  colour 57, 99, 562
  cyanide poisoning 57, 586
  distribution 56, 57
  electrocution 56, 333
  in hanging 56, 387–8
  immersion deaths 396
  organs 18, 59
  petechial haemorrhage 355
  skin colour 99
  skin haemorrhages in 57
  sudden infant death syndrome 454
  testing 59–60
  timing/permanence 57–8, 58
hypotension, in subendocardial haemorrhage 348
hypothermia 10, 414–18
  alcohol consumption and 553
  autopsy findings 416–18
    skin colour 416, 417
  biochemical markers 419
  blood pooling 417
  case history 10
  cyanosis 417
  death in hospital after 416
  effect on cooling curve 80, 81
  elderly 415, 416–17
  face/facial appearance 416
  factors involved 415–16
  in 'hide-and-die' syndrome 418–19
  infants 415
  manifestations 416
  thrombosis 418
  vitreous glucose 90
hypothyroidism, hypothermia in 415–16, 417
hypovolaemic shock 499
hypoxaemia, in drowning 402
hypoxanthine 89
hypoxia 352, 353
  acute-on-chronic 455
  alcohol effect 552
  anaesthetic-related deaths 484, 486
  cerebral *see* cerebral hypoxia
  cerebral oedema and 212
  in coronary insufficiency 504
  episodic, myocardial fibrosis in epileptics 514
  oedema in asphyxia 356
  post-mortem diagnosis difficulty 357
  solvent abuse 596
  suffocation in airtight spaces 358
  total 53
  see also 'asphyxia'

identification 8–9, 98–135
  callus 124, 125
  characteristics used in 99–106
  coffin, exhumation 37, 38
  corpses/material requiring 98–9
  dentition used in *see under* dentistry, forensic
  DNA characteristics 105–6
  explosion deaths 276–7
  facial, after reconstruction from skull 128–9
  fatal abuse 302
  fingerprints *see* fingerprints, identity establishment
  general groupings and personal identity 98–9
  mass disasters, serial/pathology numbers 43
  of mother 440
  need for, reasons 98
  non-skeletalized bodies
    age determination 104–5
    racial/ethnic characteristics 105
    sex determination 104
    stature 103–4
  photosuperimposition techniques 128
  putrefied corpse 101, 102
  skeletal remains *see* skeletal remains
  tattoos 100–1
  see also age estimation; sex determination; stature
iliac veins, blood samples 20
imipramine 573

immersion deaths 395–411
  in bath 400, 401
  cooling curve 81
  cyanosis 400
  death from natural causes before 397
  deaths from immersion 399–400
  decomposition of bodies 67–8, 396–7, 398–9
  estimation of duration 396–7
  fingerprints 395, 396
  hypostasis, post-mortem 396
  injuries before entering water 398
  injuries sustained in water 398–9
  low temperature 369, 397, 400
  maceration 395, 397, 406
  mode of death 400
  natural death whilst in water 397–8
  position of body 399
  predator damage 396, 397, 398
  reflex cardiac arrest 400, 403
  sand/mud in 396, 404
  signs 395–6
  sudden immersion in cold water 400
  suicidal 239
  see also drowning
immobility, venous thrombosis in 341, 520
immunocytochemistry, axonal injury 210
immunohistochemistry, myocardial infarction 502
immunological methods, semen testing 428
impedance 328
incised wounds 137, 153–66
  cuts/slashes 154–5
  lacerations *vs* 150, 151, 177
  stabs/penetrating 155–66
    see also stab wounds
industrial processes, carbon monoxide poisoning 561
infant
  accidental hanging 389
  age/maturity estimation 440, 443, 448–9
    from skeletal remains 117–18
  carbon monoxide poisoning 563
  decomposition 441, 442–3, 446
  hypothermia 415
  live birth determination 440–1, 443
  measurements 448–9
  neglect 412, 413, 414, 417
  plastic bag suffocation 359
  sample collection 21
  separate existence 445–8
  skull flexibility 183
  smothering (‘overlaying’) *see* smothering
  sudden death *see* sudden infant death syndrome (SIDS)
  viability 440, 442
  vitreous humour chemistry 88
  see also children
infanticide 439–50
  autopsy 442–5
  choking 443–4
  concealment of birth 440, 441, 442
  death by act of omission/commission 441–2, 443, 445
  decomposition of infant 441, 442–3, 446
  DNA fingerprinting/profiling 440
  drowning 445
  evidence of breathing 445–7
  functions of pathologist 440–2
  head injuries 444–5
  identification of mother 440
  infant maturity estimation 440, 443, 448–9
  injuries 443
  investigations, pathologists role 440–2
  lavatory pan and 444–5
  ‘live birth’ 440–1, 442, 443
  methods 443–5
  by omission to offer care 441–2, 445
  respiration determination 445–8
  ‘separate existence’ 440–1, 442, 443
    proof 445–8
  stillbirth determination 440–1
  strangulation 443–4
infantile cortical hyperostosis 469
infections
  after skull fractures 189
  airway obstruction 361
  burns 314
  chest 224–5
  delayed deaths due to 340
  haemorrhage due to 340
  in illegal abortion 432, 433–4
  medico-legal aspects 340
  post-traumatic 340
  risk at autopsy 4, 10–11, 579
  skin, in neglect/starvation 414
  see also HIV infection
infinite cylinder model 79, 83
inflammatory response after injuries 170
infrared monitoring, temperature 82
infusion accidents 482
inhumation 36–7
insecticides, poisoning 567–8
insects
  infestation of putrefied corpse 39
  larval length 77
  life histories 76, 77
  post-mortem damage 74–5, 75, 76
  sample collection 78
  types and succession 76
  see also entomology
insufflation of air, abortion 432–3, 435
insulin 546, 574–5
insurance, trauma/disease relationship 519
intercostal muscles 224–5
internal examination at autopsy
  abdominal organ removal 21–2
  body cavity exposure 17

internal examination at autopsy (*contd.*)
  body fluids *see* body secretions; sample collection
  dissection techniques 16–23
  equipment 18
  incisions 16–17
  neck structures removal 21
  pelvic organs removal 22
  report 34–5
  thorax 17–18, 21
  see also specific anatomical regions/organs
International Death Certificate 10
interpol document, mass disasters 44
'Interruptin' 434
intersex 104
interstitial fibrosis 507–8
interventricular septum, rupture 227, 228
intestine
  child abuse injuries 472
  contents, sample collection 549
  examination at autopsy 28
  injuries 230–1, 231, 233
  obstruction, narcotics causing 233
  removal at autopsy 21
  rupture 233
  strangulated 516
  see also colon; gastrointestinal tract
intracerebral haemorrhage 207, 513, 514
  cocaine addicts 582
  traumatic 207
intracranial haemorrhage 517
  in child abuse 192, 193, 464, 469–70
intracranial injuries 189–215
  child abuse 193, 466, 469–71, 476
  subendocardial haemorrhage in 348
  see also individual injuries
intracranial pressure, raised 211, 212, 213–14, 473
intrapulmonary haemorrhage 403
intrauterine infection 433–4
intubation 481
iris, changes after death 61, 99
iron works, carbon monoxide poisoning 561
ischaemia 52, 496, 506
ischaemic heart disease 493, 505–6
  see also aortic valve disease; coronary atherosclerosis
ischio-pubic index 111

jaw
  edentulous 535
  fractures 178
  see also mandible; maxilla
jejunum, injuries, in child abuse 472
joint splits, ante-mortem injuries 7
joule burn 329
judicial hanging 383–4
jugular veins
  blood samples 20, 548
  cut 236–7
jugular venous compression 368, 369
jumping from heights, suicide 154, 240

Kastel–Meyer test 127
keloid formation 304, 315
Kernohan's notch 214
kerosene heaters 561
kicking injuries 149
  abdominal 230
  bruises/abrasions 149
  facial 178–9
  kidneys 232
  lacerations 153
kidney
  in ethylene glycol poisoning 589
  examination at autopsy 27–8
  failure, after trauma 345
  injuries 232
  paraquat poisoning 566
  sample collection 550
kinetic energy transfer 137, 245, 246
kinky hair syndrome 469
'knee-capping' 305
knife
  blade type 157–8
  dimensions from stab wound 157
  examination and measurements 155–6
  movement within wounds ('rocking') 156, 158–9, 161
  penetration of skin 162–3
  serrated edge 158
  stab wounds by 156
    see also stab wounds
  taper of blade 158
knife wounds 155, 156
  defence 165–6, 167
  'dentele' toothed incision 236
  scars from 102
  slashed 154–5, 155
  suicidal 236–9
  'tentative incisions' 236, 237, 238
  see also stab wounds
kwashiorkor 414

labels/tags, body identification 8
laboratory, toxicology *see* toxicology
lacerations 137, 150–3
  cerebral 206–7
  firearm wounds 270
  incised wounds *vs* 150, 151, 177
  lung 226
  patterned 152–3, 176–7, 294, 295
  'sandwich effect' 151
  scalp 151, 152, 175, 176–7
  stellate 176
  types of impact causing 151–2
  vehicular accidents 284
lactic acid 62

'ladder tears,' aorta 21, 228, 285
Laminaria digitale 433
lanugo 448
larder beetles 78
laryngeal spasm 400, 403
larynx
  blockage 362
  bruising around, in manual strangulation 372
  compression 368
  cornuae (horns) 376–7, 377
  forensic anatomy 376, 377
  fractures of horns
    ante-mortem, proof 378
    autopsy damage causing 377
    bleeding 378, 380
    causes 378
    in hanging 388
    ligature strangulation 378, 382
    manual strangulation 377
    post-mortem 377
    significance 377
  haemorrhages 378
    manual strangulation 377, 378
    mucosal 375
  injuries
    ligature strangulation 382
    manual strangulation 376–7
  spurious bleeding behind 59, 150, 375
  thermal damage in fires 320
lavatory, infanticide 444–5
laxatives, illegal abortions 433
LD50 542
lead poisoning 434
lead shot (pellets) 246, 255, 256, 258, 259
left circumflex artery 495
left coronary artery 495
left ventricle
  aortic stenosis 506–7
  in arsenical poisoning 590
  hypertrophy 61, 504, 505
  stab wounds 227
  wounds 170
leg
  injuries in pedal cycle accidents 291
  injuries in vehicular accidents
    drivers 283, 284
    motorcyclists 290
    pedestrians 293–4, 295
  see also limbs
Lendrum's stain 602
length
  assessment 11–12, 103–4
  see also height
Letulle's method, examination of viscera at autopsy 21
leucocytes
  air embolism 348
  dating emboli/thrombi 342
  dating of wounds 167–8, 170, 171
  sex differences 104
level crossings, injuries 297
life-support machines 53
ligamentum arteriosum 228
ligature strangulation *see* strangulation, ligature
lightning, deaths from 336–7
limbs
  bruising, in child abuse 462
  diameter, in neglect/starvation 413
  fractures *see* fractures
  length after death 239
  see also leg
lips
  autopsy findings in sexual offences 421
  bruises/bruising 150
  in child abuse 178, 473
  frenulum of, rupture/tearing 178, 473
  injuries 178
  pressure marks 360
  prints 101–2
liver
  alcohol detoxification 554
  central/parallel tears 286, 287
  in chronic alcoholism 557–8
  cirrhosis 558
  fatty changes 557–8
  injuries in child abuse 232, 472
  necrosis, halothane association 486
  paracetamol poisoning 572–3
  paraquat poisoning 566
  rupture 231–2, 232, 286
  sample collection 545, 550
  in solvent abuse 596
  subcapsular haematoma 286, 340
  temperature 82
  vehicular injuries 231, 232, 286, 287
lividity *see* hypostasis, post-mortem
'love-bites' 146, 421–2, 474, 528, 530
'lucid interval' 343
lumbar puncture 548
lung(s)
  autopsy, examination 24
  in barotrauma/dysbarism 489, 490
  diatoms in 406
  in drowning 402–3, 404, 405
  'dry-oedema,' in adult respiratory distress syndrome 345
  fat in 343–4
  fetal, maturation 447–8
  formalin inflation 24–5
  granuloma, in cocaine addicts 582
  haemorrhage *see* pulmonary haemorrhage
  hypostasis 59
  injuries 226
    bruising 226
    deceleration 226
    penetrating 226

lung(s) (*contd.*)
  injuries (*contd.*)
    'pincer contusions'/lacerations 226
    vehicular accidents 285–6
  newborn 447–8
  overinflation 403, 404
  paraquat poisoning 566–7
  post-mortem aeration in newborn 447
  sample collection, for volatile substances 549, 598
  stillbirth 446–7
  sudden death in asthma 515
  surgical/anaesthetic-related deaths 486
  thermal damage 320
  weight 403
  see also entries beginning pulmonary, respiratory
Luxol Fast Blue stain 213
lynching 384
lysergic acid diethylamide (LSD) 580
lysol 587, 588

maceration 64, 442–3
  immersion deaths 395, 397, 406
macrophage 168, 405–6
maggot infestation 65, 66, 74–5, 77
  see also animal predation
magnesium 404–5
magnetic resonance imaging (MRI) 31, 161
malar flush 416, 417
malate dehydrogenase 502
malathion 567, 568
malic acid dehydrogenase (MDH) 600
mandible
  fractures 178
  race determination from skeletal remains 122
  removal/examination 536
  sex determination 108, 109–10
marasmus 413
marbling 65
Marfan's syndrome 510
marine predation 68, 75–6, 396, 398, 399
  see also animal predation
Martius scarlet blue (MSB) stain 601
masks, in sexual asphyxia 391
masochistic practices 389–92
  see also sexual asphyxia
Mass Casualty Plan 42
mass disasters
  autopsy 44
  forward planning 41–2
  identity establishment
    from dentition 532–3
    frontal sinus comparison 124–5
  interpol document 44
  mortuary, facilities 42, 43
  necessities in planning 42–3
  pathologist's role 41–4
  pathology investigations 42–3
  retrieval of bodies 43–4
  Royal College of Pathologists guidelines 44
  temporary mortuaries 43
  toxicology samples 44
mastoid process, sex determination 109
masturbation 391
maternal deaths *see* abortion; pregnancy
maxilla
  fractures 178
  kicking injuries 149
  removal/examination 536–7
MDMA (methylenedioxymethamphetamine) 581
meals, post-mortem interval estimation and 85, 87–8
mechanical forces, wounding 136–7
medical records, poisoning deaths 543
medico-legal aspects
  air embolism 346
  aspiration of gastric contents 596
  brain death 53
  child abuse 471
  extradural (epidural) haemorrhage 191–2, 198
  infections 340
  myocardial infarction 496, 518–19
  neglect/starvation 414
  pressure on neck 370
  pulmonary embolism 340, 341–2, 520
  sexual asphyxia 391–2
  solvent abuse 596
  subarachnoid haemorrhage and berry aneurysm 198–9, 519–20
  sudden infant death syndrome (SIDS) 456–7
  surgery/anaesthesia-related deaths 482, 483
  trauma and disease relationship 517–18
medico-legal autopsy *see* autopsy, medico-legal
medionecrosis 510
meningeal haemorrhage 188, 190–2, 197
meninges *see* brain membranes
meningococcal septicaemia 349
Menke's syndrome 469
mentally disordered persons, self-inflicted injuries 240
mercuric chloride 546
mercy killings 346, 359
mescaline 580
mesentery
  haemorrhage 230–1, 233
  infarction 516
  injuries in vehicular accidents 286, 289
  lacerations/injuries 230–1, 231
  rupture 289
  thrombosis 516
metallic ions, in electrical marks 332–3, 335
metallic poisons 589–94
  see also arsenical poisoning
metaphysis, avulsion/lesions 466, 467
methadone, misuse 580
methaemoglobinaemia 57

methylamphetamine sulphate 581
methyl chloride 597
methylene chloride 597
methylenedioxymethamphetamine (MDMA) 581
metopic suture 121
microbiological culture, autopsy samples 20
microbiology samples 20, 45–6
  collection at autopsy 29
  contamination at autopsy 29
  see also sample collection
microvacuolation 213
microwave thermography 82, 83
middle meningeal artery, in haemorrhages 190, 192
miliary plaques 67
military weapons 248
minerals, diatomaceous materials in 408
miners, identity establishment 103
minisatellite DNA regions 106
missile injuries
  in explosions 275–6
  mechanics 245–6, 268
  trajectory of missile 245–6
  velocity of missile 246, 258, 268
  see also firearms; gunshot wounds
mitral stenosis 507
molluscs 396
Mond method 561
Mongoloid races 122, 123, 533, 534
morphine 548, 578–80
  autopsy appearances of addicts 578–9
  blood/tissue levels 580
  sudden death 578
mortuary, mass disasters
  facilities 42, 43
  temporary 43
motorcycles, modern safety features 290
motorcyclists, injuries 290–1
'motor cyclist's fracture' 185, 188, 290
motor vehicle accidents *see* vehicular injuries
motor vehicles
  carbon monoxide sources 560
  see also vehicles
mottling, in hypostasis 59
mould formation, mummified corpse 72, 73
mouth
  autopsy findings in sexual offences 421
  blood from, post-mortem artefacts 12, 15, 36
  child abuse 473
  examination at autopsy 13, 15
  injuries 178, 473
  paraquat poisoning 566, 567
  shotgun wounds 267
mucoid medial degeneration 510
mucosal ecchymoses, in aspirin poisoning 571, 572
mucus formation, in arsenical poisoning 591
mud, on body, in immersion deaths 396
'mugging' 353, 383
mummification 64, 72–3
  adipocere with 72
  artificial 72
  factors affecting 72
  injury preservation 73
  timing 73
Munchausen syndrome 241, 243
Munchausen syndrome by proxy 243
muscle
  cadaveric spasm 62–3
  contractures, burns 316
  damage, renal failure after 345
  electricity effects on 328–9
  flaccidity as sign of death 53, 60
  heat/cold stiffening 63
  rigor mortis 60, 61
  sample collection 550
  shortening, rigor mortis 62, 103–4
  spasm, electric current effect 328–9
mutilated corpses 31, 99
mutilation, in sexual offences 423
mycotic aneurysms 511
myelomalacia 216
myocardial failure, in drowning 401
myocardial fibres, in electrical fatalities 333
myocardial fibrosis 497, 503, 504, 509
  cause of death 503–4
  diffuse 503
  in epilepsy 514
myocardial infarction 493, 495–502
  age/dating 496–7
  atheroma/thrombosis causing 495–6
  autopsy diagnosis 496–7
  cause of death 503–4
  complications 502–3
    aneurysms 503
    mural thrombosis 502–3
    myocardial fibrosis 503
    pericarditis 503
    ruptured heart/tamponade 502
  enzyme histochemistry 498, 501–2
  fluorescent methods 500–1
  histochemical methods 498–9, 500
  immunohistochemical methods 502
  incidence 495
  laminar 496, 497
  macroscopic appearances/stages 496–7
  medico-legal aspects 496, 518–19
  microscopic appearances 497–502
    morphological changes 497–9
  nuclei loss 500
  regional/focal 496
  stages 497
  topography 496
  transmural 497
  transportation injuries 496, 498
  trauma/exertion relationship 518–19

myocardial infarction (*contd.*)
  types 496
  see also coronary atherosclerosis
myocardial ischaemia 496, 506
myocardial necrosis 496, 563
myocarditis 508–9
  chronic alcoholism 558
  solvent abuse 596
myocardium
  blood supply 505–6
  contraction bands 498–9
  discoloration, post-mortem artefacts 36
  examination at autopsy 27
  fat embolism 344
  fatty degeneration 499
  fragmentation of fibres 499
  hypertrophic obstructive cardiomyopathy 507
  hypostasis 59
  senile 509
  staining 498, 499, 500, 600–1
  waviness of fibres 499
myocytes 499
  nuclei 509
myofibrillar degeneration 498
myosin 62
myxoedema 415–16, 417
myxoma 511

nails *see* fingernail
narcotic drugs, deaths from 577–84
  autopsy appearances 578–9
  idiosyncratic sensitivity/tolerance 579
  mode of death 579
  toxicology 579–80
narcotics, body-packing 233
nasal aperture/bones, in sex determination 109
nazi concentration camps 413
neck
  abrasions, manual strangulation 372–3
  bruising
    child abuse 463
    ligature strangulation 382
    manual strangulation 371–2, 373, 374
  constriction, cyanosis 356
  dissection, manual strangulation 375, 377
  examination at autopsy 21, 24
  fatal pressure on 368–94
    cyanosis 368
    duration effect 370–1
    medico-legal aspects 370
    petechial haemorrhage 371
    see also hanging; strangulation
  fatal pressure on, mechanism of death 368–71
    airway occlusion 368, 369
    carotid artery compression 369
    combined/duration 370–1
    nerve effects 369–70
    reflex cardiac arrest 369–70, 370–1
  haemorrhage, manual strangulation 378
  injury to side of, vertebral artery trauma 201, 202
  post-mortem artefacts 59, 150, 375, 377
  putrefaction 65
  rotational trauma 200–4
  self-inflicted injury 242
  suicidal stab wounds 238
neck-holds, deaths in custody 306
necropsy *see* autopsy
negative autopsy *see* autopsy, medico-legal, obscure
neglect 412–14
  children 412, 417
  elderly 412
  general features 412–14
  infants 412, 413, 414, 417
  medico-legal problems 414
negligence, medical 483
  surgical procedures, deaths associated 481
Negroid features 534
nerves, pressure on neck and 369–70
Netherlands, sudden infant death syndrome prevention 452
neurological lesions, carbon monoxide, poisoning 563
neuronal injury, diffuse 209–10, 470
neurones, in cerebral hypoxia 213
Newton's Law of Cooling 79
nickel production, carbon monoxide poisoning 561
nipples, bite marks 528, 529, 530
Nissl granules 213
nitrazepam 573
nitric acid 322, 588
nitrocellulose 248
nitrogen, dating of bones 127
nitrogen narcosis 488
nitrostigmine (parathion) 567, 568
non-accidental injury in childhood *see* child abuse
nortriptyline 573
nose
  bleeding 178
  blood from, post-mortem artefacts 12, 15, 36
  fracture of bridge 178
nuclear chromatin, in sex determination 104
nylon bags, for samples 545, 549, 586, 598

obscure autopsy *see under* autopsy, medico-legal
obscure syndrome 45
obstetric catastrophes, subendocardial haemorrhage 348
obturator foramen 111
occipital bone, 'ring fracture' 181, 185, 214, 283, 290
occupational stigmata, identity establishment 103
odontogram 536
odontologists, forensic 527
odontology, forensic *see* dentistry, forensic
oedema
  in 'asphyxia' 356
  in hypothermia 417
  myocardial 500

  post-mortem dependent 56
  see also cerebral oedema; pulmonary oedema
oesophagus
  artefactual haemorrhage behind 59, 150, 375
  'banding' 375
  carcinoma 516
  corrosive acids/alkalis effect 587, 588
  cyanide poisoning 586
  intubation 481
  soot inhalation 323
  varices 516
Ohm's law 327, 328
Oil Red-O 343, 344
oleoresin capsicum (OC) 306
omentum, injuries 286
opioids 578–80
organic solvents *see* solvent abuse
organophosphorus pesticides, poisoning 567–8
organs
  atrophy, in neglect/starvation 414
  examination at autopsy 23–8
  firearm damage 270
  hypostasis 59
  position in stab wounds 159
  putrefaction 66–7
ossification centres 117, 118, 119, 121
  fetal maturity estimation 444, 445
  time of appearance (table) 119
osteogenesis imperfecta 468–9
osteometric board 117
osteoporosis 180
oxalate salts, deaths from 588–9
oxalic acid, deaths from 588–9
oxygen
  blood content 353
  deprivation 352, 357–8
oxygen:nitrogen ratio 489
oxyhaemoglobin 356, 396, 586

pachymeninges 189
pancreas, post-mortem artefacts 36
pancreatitis, acute, in hypothermia 418
papillary muscles, myocardial infarction 497
paracetamol poisoning 572–3
paradoxical respiration 223
paraffin heaters 561
paraquat poisoning 345, 566–7
parathion 567, 568
paroxysmal nocturnal dyspnoea 505
paternity testing 106
pathologists 2
  fingerprints 101
  infanticide investigations 440–2
  sudden death examination 492
  surgical procedure-related deaths 480
patient's notes 481–2
patterned abrasions *see* abrasions, patterned
patterned electric marks 332, 336
patterned lacerations 152–3, 176–7, 294, 295
PCP (phencyclidine) 580
Pearson and Bell's table, 'mathematical sexing of femur' 112
Pearson's table, stature calculation 116
pedal cyclists, injuries 291
pedestrians
  drug/alcohol intoxication 296
  injuries 291–5
    children 292–3
    primary/secondary/tertiary 291
    scooping-up 291, 292
    sequence of events 291–2, 292
    speed of impact 293
    types 293–4
pelvic 'indices' 111
pelvic organs, removal at autopsy 22, 436
pelvis
  fractures 233
  injuries 233, 294
  race determination from skeletal remains 122
  sex differentiation 111–12
penetrating injuries 155–66
penis, in putrefaction 65
penknife, wounds 226, 229
peppering, in explosions 275
pepper spray, injury 306
peptic ulcer, perforated 517
peribronchiolar cell infiltration 455
pericardial sac, bleeding into 228
pericarditis 503
pericardium, examination in air embolism 347
pericranium 175
periodic acid–Schiff (PAS) stain 600
periosteum
  blood and calcification under 466
  dating of bones 126
  shearing, in child abuse 466
perirenal haemorrhage/haematoma 232, 510
peritoneal cavity, bleeding into 230
peritonitis 233, 517
  chemical 433
perivascular haemorrhage, in hypothermia 418
Perl's reaction 167, 197
persistent vegetative state 53
personal defence spray, injury 305–6
personal identity, from skeletal remains 123–5
pesticides, organophosphorus 567–8
petechial haemorrhage 144
  'asphyxia' 354–6, 364
    as unreliable indicator 355
  barotrauma/dysbarism 490
  causes 354
  distribution 354
  drowning 403
  electric fatalities 333
  facial appearance 354, 355, 356

petechial haemorrhage (*contd.*)
  fatal pressure on neck 371
  fat embolism 344
  in hanging 388
  head injuries 14
  love bites 528, 530
  lungs, examination 24
  post-mortem artefacts 36
  posture affecting 355
  scleral/eye 354
  significance 355–6
  strangulation 354, 382–3
  subpleural 355, 403
  sudden infant death syndrome 454, 456–7
  traumatic ‘asphyxia’ 364
pethidine, blood/tissue levels 580
petrol 560
pets, post-mortem damage 76
Pfitner’s table, bodily dimensions 108
pharynx, thermal damage 320
phencyclidine 580
phenols 588
  deaths from 587–8
phenothiazines, poisoning 573
phloxine–tartrazine stain 602
phosgene, in fumes from fires 322
phosphotungstic acid-haematoxylin stain (PTAH) 342, 499
photography
  at autopsy 13, 15, 32–3
  bite marks 528, 530–1, 532
  cameras 32–3
  child abuse 476
  exhumation 37
  film type 32
  identification by dentition 537
  illumination 32, 530–1
  scene of death 6
    composition 33
photosuperimposition techniques 128
physical abuse
  in custody *see* custody, deaths in
  fatal, evidence 301–5
  see also child abuse
physical activity
  cadaveric spasm and 63
  sudden death and 519
  timing of rigor mortis 60–1
physique, hypothermia relationship 415
pial veins 347
pia mater 189
piercing, infanticide 444
pigmentation, mottled, in heat injury 303, 304
pistols, injuries 271–2
placenta, examination in infanticide 443
plastic bag
  sexual asphyxia 390, 391, 596
  suffocation 359–60, 596
plastic bullets 274
plastic containers, for samples 545
plastic polymers, fumes from fires 322
platelets 342, 346
platysma muscle, bruising 373
pleural adhesions 21, 225, 226
pneumoconiosis, formalin inflation of lungs 24–5
pneumonia, paraquat poisoning *vs* 567
pneumothorax 223, 225–6
  autopsy 18, 31
  barotrauma/dysbarism 489, 490
  causes 225
  natural/unexpected deaths 226
  resuscitation artefacts 41
  ‘sucking wound’ 225
  tension 225
  types 225
poisoning 541–51
  agrochemicals 566–9
  antimony 591
  arsenic *see* arsenical poisoning
  autopsy 543
  corrosive acids/alkalis and phenols 587–8
  cyanide 585–7
  detection of poison/chemical 543
  drugs of dependence 577
  ethylene glycol 589
  exhumation 38–9
  fatal concentrations, variations 542–3
  fatal dose concept 541–2
  hallucinogenic drugs 577–84
  hospital deaths 543
  medicines 8, 543, 570–6
    autopsy appearances 570–1
    autopsy difficulties 570–1
    laboratory test results 571
    mode of death 571
    see also specific drugs/medicines
  metals 589–92
  narcotic drugs 577–84
  organophosphorus pesticides 567–8
  oxalic acid and oxalate salts 588–9
  paraquat 566–7
  ricin injection 550
  sample collection *see* sample collection
  samples, ‘exhibit labels’/‘exhibits officer’ 544
  subendocardial haemorrhage in 348
  thallium 591–2
polaroid cameras, use at autopsy 32
police
  autopsy request 3
  bite marks on 528
police cells, suicide in 308
police surgeons 421
politics, human rights abuse and 301
polyarteritis nodosa 511
polymerase chain reaction (PCR) 428

polymorphonuclear leucocytes, infiltration 168, 170
polyps, sudden death from 516
pontine haemorrhage 517
post-concussion state 209
post-mortem artefacts 36
  abrasions 139, 143, 149–50
  aortic tears 285
  bleeding behind larynx 59, 150, 375
  blisters/bloating/discoloration 36
  blood from mouth/nostrils 12, 15, 36
  bruises/bruising 13, 36, 149–50
  burns 331–2
  defibrillators 12, 36, 41
  epidural haemorrhage *vs* 192
  gastromalacia 36
  haemorrhage 57, 59, 150, 375
  haemorrhage patches 36
  myocardium discoloration 36
  neck, haemorrhage 59, 150, 375, 377
  pancreas 36
  petechiae 36
  petechial haemorrhage 36
  putrefied corpse 36, 65
  resuscitation *see* resuscitation artefacts
post-mortem bleeding 225, 233, 340
post-mortem blood clots 341
post-mortem changes 56–63
  burns 317
  hypostasis 56–60
  immersion deaths 67–8, 396–7, 398–9
  surgery/anaesthesia-related deaths 480
  toxicological 545–6, 547
  see also decomposition; hypostasis, post-mortem; rigor mortis
post-mortem cooling *see* cooling, post-mortem
post-mortem dependent oedema 56
post-mortem examination *see* autopsy
post-mortem hypostasis *see* hypostasis, post-mortem
post-mortem interval
  cerebral hypoxia diagnosis 213
  entomology and 76
    see also entomology
  estimation 78–90
    body cooling/temperature 7, 79–83
    gastric emptying 83, 85, 87–8
    multiple site temperatures 83
    post-mortem chemistry 89–90
    rectal temperature 81–2
    retinal changes 54
    at scene of death 6–7
    ‘temperature ratio’ concept 83
    thermometry 82–4, 83
    vitreous humour chemistry 88–9, 89–90
    see also cooling, post-mortem; rectal temperature
  hypostasis relationship 57
  rigor mortis and 60, 61
  ‘spot-check’ for 61
post-mortem overspill 363
postural asphyxia 355, 365
posture
  drunken prisoners 307
  effect on cooling curve 80–1
  petechial haemorrhages and 355
  sudden infant death syndrome 453
potassium 88, 89
  release in drowning 401
potassium cyanide 585, 586
potassium fluoride 545, 546
potassium permanganate 434
pre-auricular sulcus 111
precipitation (jumping from height) 240
predators, post-mortem damage *see* animal predation; insects
pregnancy 65
  deaths associated 431–8, 517
    see also abortion
  ectopic 436
  evidence from skeleton 114
  legal termination 431, 432
  seatbelt problems 289
premedication, excessive 484
pressure, adverse effects 488–91
pressure marks 359, 360
pressure sores 414
Prinsloo and Gordon artefact 59, 150, 375
procurator fiscal 3
propane 561, 595
propeller-induced injuries 399
property, personal 8
  mass disasters 43–4
  see also clothing
propylene 589
prostaglandins, illegal abortions 434
protein
  ante-mortem *vs* post-mortem burns 317
  dating of bones 127
  identification of bone fragments 108
pubic bone, sex differentiation 111
pubic hair 104
  examination in sexual offences 423
‘pubic scars’ 114
pubic symphysis, age estimation 118–20
‘Pugilistic’ attitude 63, 316, 321
pulmonary artery, injuries 228–9
pulmonary barotrauma 489
pulmonary damage, explosions 275
pulmonary embolism 24, 340–3, 513, 520
  after legal termination of pregnancy 432
  after surgery 483, 520
  appearances 341–2
  dating 342–3, 520
  medico-legal aspects 340, 341–2, 520
  post-mortem blood clot *vs* 341
  predisposing factors 341, 520

pulmonary embolism (*contd.*)
  rare causes 513
  sources of embolus 341, 520
  sudden death, deep vein thrombosis 513–14, 520
  trauma/disease interaction 340–1, 341, 520
pulmonary fat embolism 343
pulmonary granuloma 577
pulmonary haemorrhage 275, 286
  drowning 403
  vehicular accidents 286
pulmonary inflammation 455
pulmonary ligament 226
pulmonary oedema 343, 356, 455
  carbon monoxide poisoning 562
  deaths from corrosives 587
  drowning 402–3
  hypertensive heart disease 505
  hypothermia 418
  morphine addicts 578–9
pulmonary thromboembolism 513–14, 520
pulmonary tuberculosis, sudden death 516
pulsation, absence 352
'punch-drunk' syndrome 204
punching, in child abuse 472
puncture marks, autopsy resuscitation artefacts 40–1
puncture wounds 153
pupils
  'cadaveric position' 54
  morphine administration effect 579
  sign of death 54
Puppe's rule 266
purging of body fluids 36, 65
Purkinje cells, sudden death in epilepsy 514–15
putrefaction 64–9
  bacteria 64–5, 68
  blisters 64, 65
  definition 64
  facial bloating 65
  fingernail 65
  gas formation 64, 65, 435
  inhibition by adipocere formation 71
  internal organs 66–7
  purging due to 36, 65
  sequence of changes 64–7
  timing 64, 67, 69
    temperature effect 64, 67, 68
  see also decomposition
putrefied corpse
  bruising, glycophorin A detection 40
  external/internal examination 39–40
  fingerprints 39, 102
  identification 39, 101, 102
  insect infestation 39
  post-mortem artefacts 36, 65
  sex determination 104
pyloric spasm 553
pyogenic infections, in cocaine addicts 582
quinine, illegal abortions 434
quinsy 515

race 11
  determination
    dentistry, forensic 105, 533–4
    from skeletal remains 122–3
  identity establishment 105
  sex determination 108, 110
  skull features 110, 122, 123
  subdural haemorrhage in child abuse 470–1
  see also ethnic characteristics
radiation, burns 314–15
radiocarbon test 127–8
radiography/radiology
  age estimation in skeletal remains 117, 121
  air gun injuries 272
  ante-mortem, identity establishment 125
  in barotrauma 31
  bullet wounds 266
  child abuse 32, 476
  gunshot/explosion deaths 31, 276
  identification by dentition 537
  identification in explosion deaths 276–7
  mass disasters 44
  medico-legal autopsy 32
  mutilated remains 31
  neck, in manual strangulation 373–4
  pneumothorax 31
  vertebral artery trauma 203
radiostrontium 128
radius, in stature estimation 117
railway injuries 297–8
Rambo knife 143, 155
rape
  anal 427, 428, 429
  children 425–6, 427, 429
  crime scene 5
  deaths associated 421
  evidence 425–6
  genital injuries 424
  see also sexual offences, deaths associated
rats, post-mortem damage 75
razors, stab/incised wounds 163
rectal temperature
  cooling curve 79–80
  in hypothermia 416
  measurement methods and controversy on 81–2
  post-mortem interval estimation 81–2
    Henssge's nomogram 83, 84–5, 86
  at scene of death 7, 81–2
  'temperature ratio' concept 83, 87
rectum
  foreign bodies 426
  rupture 231
'red flare' 317

reflex cardiac arrest *see* cardiac arrest, reflex
refrigeration, infectivity effect 11
'refrigerator effect' 596
relatives
  attendance at autopsy 3
  information on autopsy 3
renal failure, after trauma 345
report, autopsy *see* autopsy, medico-legal, report
respiration
  cessation as sign of death 55
  changes in lung in infants 446–7
  chest wall injuries and 223
  evidence in newborn 445–8
  obstruction, foreign bodies 515
'respirator lung' 345
respiratory arrest, in electrical fatalities 328
respiratory infections 454, 456
respiratory insufficiency, in fat embolism 343
respiratory paralysis, in electrocution 327
resuscitation artefacts 36, 40–1, 224, 353
  β-adrenergic catecholamines 41, 353
  airway damage 41
  bruising chest/heart/face 40
  categories of damage 40–1
  defibrillators 12, 36, 41
  gastric contents 41
  haemorrhages 40, 41
  Heimlich manoeuvre 41
  pneumothorax 41
  puncture marks 40–1
  see also cardiopulmonary resuscitation (CPR)
rete pegs 168
retina
  child abuse 472–3
  sign of death 54
retinal haemorrhages 41, 472–3
retinal vessels, 'trucking' 54–5
retraction balls/globes 210
retrograde amnesia 209–10
retropharyngeal abscess, rupture 515
revolvers 247, 248
  see also rifled weapons, wounds
rheumatic valve disease 506, 507
'Rhyl mummy' 72
rib fractures 223–4
  callus formation 467, 468
  child abuse 224, 467–8
  pneumothorax 223
  respiration affected 223
  'string of beads' 467
  vehicular accidents 285, 290
ricin 550
rifled weapons 248
  ammunition 248
  barrel length 263
  muzzle velocity 268
  test firing 263, 270
rifled weapons, wounds 258–67
  abrasion collar 263–4, 268–9
  blowback into barrel ('back spatter') 261
  in bone 266, 267, 270
  burning/blackening 259, 262
  close-range 261–2
  contact 258–60
  direction of discharge 265–6, 268
  entrance wounds 262, 264, 265, 266, 267
  exit wounds 260, 262, 263, 268–9
  extreme trajectory 265
  fouling 263
  gas rebound 263
  grease ring ('ring of dirt') 264–5
  hyperaemia around 260, 262
  mechanics 245–6
  medium-distance 263
  muzzle gases 259
  muzzle imprint 260
  posture affecting 266
  powder tattooing 262–3
  range of discharge estimation 263, 270–1
  rust stains 260, 261
  shelving/undercutting 265
  soot/smoke soiling 262–3
  stellate 268
  suicidal 259, 261
  tangential/glancing 269
  'through-and-through' 268
  unburnt explosive/filler on skin 263
rifles 247, 248
right coronary artery 495
right ventricle
  hypertrophy 25
  stab wounds 226
rigidity, cadaveric *see* rigor mortis
rigor mortis 60–3
  biochemistry 61–2
  'breaking' 62
  charting teeth and 536
  duration 60
  electrical fatalities 61
  external examination 13
  false, heat and cold stiffening 63
  fingerprint taking 101
  goose-flesh in 61, 395
  gross effects 62
  height of body 103–4
  instant/cadaveric spasm 63
  movement of body after 62
  non-skeletal muscle/tissues 60, 61
  removal of clothes 8
  testing 60, 61
  timing/sequence 60, 61
    factors affecting 60–1
'ring fracture' 181, 185, 214, 283, 290
road traffic accidents *see* vehicular accidents; vehicular injuries

Royal College of Pathologists 30, 44
rubber bullets 274
'Rule of Nines' 13, 313

sacrum, sex characteristics 112
salicylate poisoning 571–2
saliva, samples 422, 528
sample collection
  in abortion deaths 436
  at autopsy 19–21, 547–50
  bile 548
  bites/bite marks 528, 531
  blood 19–20, 547–8
    containers 545
    for cultures 29
    for DNA testing 106, 604
    road traffic deaths 296
    sites not suitable 19–20, 547
    sites/procedures 19–20
  body fluids 19–21, 549
  cerebrospinal fluid 21, 548
  in child abuse 476
  containers for 30, 545
  in cyanide poisoning 586–7
  in drug overdoses 543–4
  entomological 78
  exhumation 37
  intestinal contents 549
  microbiological 20, 29, 45–6
  microbiological contamination 29
  obscure autopsy 45–6
  in poisoning deaths 38–9, 543–4
  retention, legal issues 3
  saliva 422, 528
  scene of death 6
  semen 106, 423, 424
  in sexual offences 106, 423, 424
  in solvent/volatile substance abuse 549–50, 597–8
  stomach contents 27, 545, 549
  in sudden infant death syndrome 456
  in surgical/anaesthetic-related deaths 486
  tissues 550
  toxicological 29–30, 46, 486, 543–4
    avoiding operator contamination 545
    containers 30, 545
    information for laboratory 544
    mass disasters 44
    post-mortem changes 545–6, 547
    preservation 544, 545–6
    request form 30
    sampling site 546–7
    timing 544
  toxicology kit 20
  urine 20, 548
  vitreous humour 20–1, 88, 548–9
  vomit 549
samples
  ancillary investigations 29
  DNA profiling 106
  histology 30, 46, 486
  labelling and continuity of evidence 544
  microbiological culture 20
  post-mortem changes 545–6, 547
  tape 6
sand, in body, in immersion deaths 396, 404
Saphir's myocarditis 508
sarcoidosis, cardiac 509
saw, internal examination of thorax 18
scalds 312–14
  child abuse 474
scalp
  bleeding 176
    after death 176
  bleeding below 175
  in child abuse 463
  forensic anatomy 174–5
  injuries 174–8
    abrasions 175
    bruising 175–6, 463
    from falls 177–8
    'flaying' 176
    lacerations 151, 152, 175, 176–7
  petechiae 354
scalping injuries 152, 176
scanning electron microscopy 332–3
scapula
  fractures 468
  sexing from 113
scars
  age of 102–3
  appearance/coloration 102, 103
  ethnic differences 102
  histology 103
  human rights abuse 102, 303, 304
  identity establishment 102–3
  knife and bullet wounds 102
  surgical 102
  traumatic 102
scene of death 4–7
  attendance by pathologist 4
  body
    blood splashes 5–6
    clothing/property 8
    position 5
    removal 7
  equipment/thermometers 4–5
  examination 4–7
  personnel at 5
  photographs 6, 33
  protection 6
  samples 6
  solvent abuse 597

  sudden infant death syndrome 453
  time since death estimation 6–7
    see also post-mortem interval, estimation
sciatic notch, greater 111
scissors, injuries from 164–5, 166
scleral haemorrhage, in child abuse 472
scratches 137, 141
  see also abrasions; fingernail, abrasions
screwdriver wounds 153
scrotum
  injuries 233
  putrefaction 64, 65
  rigor mortis 61
SCUBA diving 488, 489, 561
season, putrefaction timing 64, 67
seasonal incidence, sudden infant death syndrome 453
seatbelts 287–9
  cervical spine damage despite 284
  children/small persons 289
  criticisms 288
  injuries due to 288–9
  injury reduction 287–8, 288
  kinetic energy transfer reduced 137
  lap-strap type 287, 288
  legislation 287
  mechanism of action 287
  pregnancy 289
  types 287, 288
self-biting, in children 474
self-inflicted injuries 235–44
  non-suicidal 235, 240–3
    characteristics 242
    features assisting recognition 241
    for gain 241
  wound position and clothing 8
self-martyrdom 240
semen
  collection/sampling 106, 423, 424
  post-mortem ejaculation 12–13, 61, 391
  testing 106, 424, 426–8
    'spot' test 428
seminal fluid, DNA profiling sample 604
seminal stains 427–8
senile myocardium 509
septicaemia, in illegal abortion 433–4
septum, ruptured 502
serological tests, bones/bone fragments 108, 127–8
serotonin, in dating wounds 169
sex determination 11
  bodily dimensions 108
  cytological examination 104
  dentistry, forensic 533
  DNA analysis 104, 533
  fluorescence 104
  identification from teeth 533
  non-skeletalized bodies 104
  skeletal remains 108–14
    accuracy 108
    bone length 112, 113
    long bones 112–13
    other bones 113–14
    pelvis characteristics 111–12
    pregnancy evidence 114
    sacrum characteristics 112
    skull 108–11
  teeth 109
sex differences
  infant, sudden infant death syndrome and 452
  stature estimation 114
sexual asphyxia 389–92, 596
  features 389–91
  mechanism 389
  medico-legal aspects 391–2
  plastic bag 390, 391, 596
  suicide *vs* 390
sexual intercourse, evidence 425–6
sexual offences, deaths associated 421–30
  bites/bite marks 421–2, 528, 529, 530
  bruising 421, 422, 427
  children 106
    see also child abuse, sexual
  decapitation 423
  DNA profiling 106, 424, 428
  external examination/findings 421–4
  fingernail examination 423
  homicide 421, 423, 424
  homosexual 427, 428–9
  human rights abuse 305
  internal examination 424–5
  interpretation of minimal findings 425–6
  rectal temperature cautions 82
  sample collection 106, 423, 424
  semen testing 106, 424, 426–8
shaken-baby syndrome 193, 469–70, 471
shaking, subdural haemorrhage after 193, 469–70, 471
shark attacks 76
shaving, head 14
shear stress 193, 206, 209
Sheehan's haemorrhage 348
ships, suffocation in tanks 358
shock 170
shock waves, explosions/missiles 246, 275
shooting, human rights abuse 305
shotguns 246–8
  ammunition 246, 247–8
  cylinder and choke barrels 246
  test firing 253
  twelve-bore 247
shotgun wounds 249–58
  appearance 249, 251–3, 255–8
  burns 250, 251, 252
  carbon monoxide levels 250, 252, 267

shotgun wounds (*contd.*)
  cartridge constituents contributing 249
  close discharge 251–3
  contact wounds 249–51
    entrance *vs* exit wound 267
    muzzle impression 250, 251
    over bone 250–1
  direction of discharge/injury 254–5
  exit wounds 255, 259, 267
  features 251–3, 255–8
  foreign body embolism 345
  hair singeing 251, 253
  internal tract 255–6
  mid-to-distant range 253–4
  radiography 257, 258
  range of discharge estimation 249, 271
  'rat-hole' appearance 249, 253
  secondary damage 256
  shape 249–50, 254
  shelving/undercutting 254–5, 256
  short-to-mid-range discharge 253
  skull 266, 267
  soiling/burning 249–50, 251, 252
  suicidal 253, 254, 255
  tangential 256, 257–8
  tattooing in 252
  'through-and-through' 267
  track 255
  twelve-bore shotguns 251, 252, 253, 254, 255, 256, 258
    hunting accident 259
  wad wounds 254, 256–7
siderophage 577
SIDS *see* sudden infant death syndrome (SIDS)
sixpenny bruises 148, 463
skeletalization 67
skeletal remains 106–8
  age determination from *see* age estimation
  are the remains bones? 107
  dating *see under* dating
  evidence of pregnancy from 114
  extrajudicial execution 305
  facial appearance reconstruction 128–9
  fatal abuse 302
  general categorization 107
  human origin, determination 107–8
  identification 99, 106–8, 537
    personal identity 123–5
    see also under dentistry, forensic
  photosuperimposition techniques 128
  race determination 122–3
  sex determination *see* sex determination
  stature estimation *see under* stature
skeleton, reassembly and layout 108
skin
  in age determination 11, 104
  blisters *see* blisters
  burning *see* burns
  colour 12
    abortion deaths 346
    barotrauma 489
    burnt bodies 99
    carbon monoxide poisoning 57, 562
    cyanide poisoning 586
    ethnic characteristics 99
    hypostasis effect 99
    hypothermia 416, 417
    putrefaction 64, 65
  crepitance 489
  destruction, second degree burns 313
  forensic anatomy 137–8, 174
  haemorrhages, in hypostasis 57, 58
  heated, splits in 316, 318
  infections, in neglect/starvation 414
  maceration *see* maceration
  in mummification 72
  in neglect/starvation 413–14
  pigmentation, in identity establishment 99, 105
  'rain-drop' punctate pigmentation 590
  resistance to electrical current 328
  stab wounds 162–3
  structure 138
  temperature 79, 82
  thickness, stab wounds 162
'skin popping' 578
skin-slippage 64, 65, 99
skull
  bullet/shotgun wounds 266, 267
  children 464–5, 466
  examination at autopsy 22
  facial appearance reconstruction 128–9
  flexibility, in infants 183
  forensic anatomy 108, 181–2
  identification from
    age estimation 120–2
    personal identity 124–5
    race determination 110, 122, 123
    sex determination 108–11
  muscle ridges 109
  removal at autopsy 19
  stress lines 183
  suture lines 120–2, 181
    diastasis (springing) 185, 187, 188
  tensile strength 186–7
  thickness 181–2
skull fractures 181–9
  ante-mortem *vs* post-mortem 182
  anterior fossae, black eyes 179, 180
  axes/cutting weapons 186
  basal linear 184–5
  base of skull 185
  blood below pericranium 175
  blunt injuries 177, 183
  child abuse 185, 464–6
  children 180, 185, 466

- comminuted, bullet wounds 266, 267
- comminuted depressed 183, 186
- dangers 188–9
- depressed 177, 184, 186, 187, 188
  - traumatic epilepsy after 188–9
- extradural haemorrhage associated 190–2
- falls 184, 187
- force required for 186–7
- frontal bone 466
- ‘growing’ 466
- ‘hinge’ 183–4, 188, 290
- horizontal cracks 466
- infections 189
- linear 183, 184–5
- mechanics 182–4
  - ‘struck hoop’ analogy 182, 183, 206
- meningeal haemorrhage after 188
- motorcyclists 185, 188, 290–1
- occipital 184, 466
- occipitoparietal 464
- pond fracture 185–6, 186
- Puppe’s Rule 184
- ‘ring fracture’ 181, 185, 214, 283, 290
- shotgun wounds 258
- significance 182, 204
- sites 183–4
- ‘spider’s web’ (mosaic) 184, 186
- ‘staggered’ 466
- stepped 185
- types 184–6
- vehicular accidents 188–9, 284, 287

slashes 154–5, 155

slippery elm 433

smoke
- autopsy appearance of burns 315–16
- inhalation 317, 322
- soiling 252
- *see also* fume inhalation

smothering 358–61
- accidental 360
- definition 353
- infants 358, 359, 444, 456–7
  - cot deaths *vs* 456–7

sodium 89
- release in drowning 401

sodium fluoride 545, 546

sodium hydroxide, burns 588

sodium hydroxide test, carbon monoxide poisoning 562

soft tissue injuries, pedestrians 294

soil, decomposition of buried bodies 69

solvent abuse 595–9
- asphyxia 596
- autopsy 597–8
- chemical effects 596
- dangers 595–6
- effects 595
- facial examination 598
- medico-legal aspects 596
- methods of abuse 595
- physical danger 596
- reflex cardiac arrest 595–6
- substances used 589, 596–7
- sudden death 595–6

somatic death 52, 53

Sommer’s movements 62

soot
- inhalation 322–3
- staining, of skin 252, 262–3

space-occupying lesions, intracranial 193

‘Spanish windlass’ 240, 379, 380, 381

spark burns/lesions 329, 330, 331

spasm, cadaveric 61, 62–3

Spaulding’s sign 443

specimen collection *see* sample collection

spermatozoa, retrieval 428

spiked instruments, stab wounds 156

spinal anaesthesia, deaths related 485

spinal column, stature estimation 116

spinal cord
- haemorrhage 216
- hyperflexion/hyperextension injuries 216
- infarction 216
- injuries 216
- removal/examination at autopsy 23, 24

spinal injuries 214–16
- compression damage 181, 214
- deceleration 214, 215
- examination at autopsy 23
- fractures in falls 181, 214
- hyperflexion/hyperextension 214–15

splanchnic shunt 397

spleen
- cyanide poisoning 587
- damage, in chronic alcoholism 558
- examination at autopsy 27
- haemorrhage 231, 233
- injuries 231, 286
- rupture 231
- vehicular accidents and 286
- weight, in drowning 406

spontaneous combustion 323–4

stab wounds 153, 155–66
- amount of force used 161–3
- characteristics 156–64
- chest 160, 162, 222–3, 225, 226–7, 229
  - haemothorax and 224–5, 225
  - heart 223, 226–7, 229, 238
  - lung injuries 226
  - suicidal 238
- depth 159
- direction 158–9, 159–61
- homicide 155, 160, 161, 163
- human rights abuse 303–4
- infanticide 444

stab wounds (*contd.*)
  kinetic energy transfer 136
  kitchen utensils 164
  knife dimensions and blade type 156, 157–8
  length 157
  movement of knife within ('rocking') 156, 158–9, 161
  multiple 157, 158, 160
  nature of weapons 155–6
  non-suicidal self-inflicted 241
  pneumothorax 225
  razors/glass 163–4
  sexual offences 423
  'shelving' (undercut) 159, 160
  skin 162–3
  skin thickness 162
  suicidal 238
    see also suicidal injuries, knife wounds
  survival period after 170
  'through-and-through' 226, 227
  track 160
  unilateral 'fish-tail' split 158
  'V-shaped' 160
'stagnant anoxia' 352
'staining' *see* hypostasis, post-mortem
staining techniques 600–2
stamping injuries 147, 149, 153
  flail chest 224
  great vessels 228–9
  head 179
  heart 227
starvation 412
  'dry'/'wet' types 413
  general features 412–14
  infanticide 445
  medico-legal problems 414
stature
  estimation from skeletal remains 114–16
    calculation systems 115–16, 117
    dried bones 114
    fragmentary bones 116
    length of long bones 114, 115, 116
    spinal column 116
  intact body, identity establishment 103–4
  tables, female/male 115, 116
'Stavanger definition,' sudden infant death syndrome (SIDS) 451
steel workers, identity establishment 103
steel works, carbon monoxide poisoning 561
steering wheel injuries 223, 229, 283, 287
  abdominal 229
sternal rib methods, age estimation 120
sternum
  autopsy in air embolism 347
  fractures 224, 468
  sexing from 113
stillbirth 439–40
  decomposition of fetus 64
  definition 439
  determination 440–1
  legal criteria 440–1
  proof of separate existence and 445, 446
stomach
  in aspirin poisoning 572
  in barbiturate poisoning 574
  carcinoma 516
  contents *see* gastric contents
  in cyanide poisoning 586
  deaths from corrosives 588
  in drowning 404
  emptying *see* gastric emptying
  examination at autopsy 27
  injuries 229–30
  in paraquat poisoning 566, 567
  'red velvet' in arsenical poisoning 590
  rupture 230, 231, 472
strangulation 369
  abrasion marks 141, 148, 372–3
  cricoid cartilage 376, 378, 382
  deaths in custody 308
  definition 353
  features 370
  hypostatic artefacts of oesophagus *vs* 59, 373
  infanticide 443–4
  ligature 353, 379–83
    accidental 383
    asphyxia 382
    embedded 380–1
    facial appearance 380
    hanging mark *vs* 381–2, 386
    homicide 379
    infanticide 443–4
    knots 382
    laryngeal horn fracture 378, 782
    ligature mark 141, 308, 379, 380–2
    mode of death 381, 382–3
    nature of ligature 379–80, 381
    position of mark 381–2
    signs of local injury 382
    suicidal 240, 380
  manual 371–9
    autopsy appearances 371–3
    bruising on neck 371–2, 373, 374
    cardiac arrest 374, 378–9
    carotid artery damage 378–9
    fingernail abrasions/marks 141, 372–3, 373
    internal appearance 373–4
    laryngeal fractures 377
    laryngeal fracture significance 377
    laryngeal injuries 376–7, 378
    spurious bleeding behind larynx 375
    vagal inhibition 371, 378

  petechial haemorrhage 354, 382–3
  scleral haemorrhage in 354
stress
  coronary atherosclerosis and trauma 518
  effect on digestion 88
  ruptured berry aneurysm and 199, 512
stroke, cerebral haemorrhage and 207
strontium 405
struck-hoop' action 206
'stud-guns' 273–4
subarachnoid haemorrhage 196–8
  alcohol and 199
  appearances/mechanism 196–8
  bleeding sources 197, 512
  blood volume 199
  bruise below ear 148
  in child abuse 470–1
  haemolysis 197
  investigation methods 203
  mechanism and relevance 198
  medico-legal implications 198–9, 519–20
  rapidity of death 198, 199, 511
  resuscitation artefacts 41
  rotational trauma to head/neck 200, 202
  ruptured berry aneurysm 198–9, 511, 519–20
  sudden death due to 196, 198, 511–12, 519–20
  trauma and 196–7, 198, 519
  trauma/disease interaction 519–20
  vertebral artery injury 200, 202, 203
subarachnoid space 189
subclavian veins, blood samples 20, 548
subcutaneous fat, in burns 315
subcutaneous tissue
  bruising 144–5
  sample collection 550
subdural haemorrhage 192–6
  acute 193–4
  blood volume 193
  in boxers 204
  child abuse 192, 193, 469, 470, 471
  chronic 194–5
    enlargement mechanism 195
  closed head injuries 193
  colour changes 194, 196
  dating 195–6, 197
  elderly 192–3
  forensic implications 195, 198
  formation 193, 195
  latent interval and coma 194, 197–8
  'neomembrane' formation 195
  racial aspects 470–1
  shaking causing 193, 469–70, 471
  'spontaneous' 192
subdural space 189
subendocardial haemorrhage 286, 348–9, 590
'submarining' 289, 304
subpleural petechial haemorrhage 355
succinate dehydrogenase 498, 501
succinylcholine 550
sucking wounds, chest 225
suction lesions, bites in sexual offences 421–2
suction petechiae 530
  love bites 148, 421–2, 474, 528, 530
sudden death 492–526
  abdominal catastrophes 516–17
  abortion 517
  aneurysm rupture 509–13, 519–20
  aspirin poisoning 571, 572
  bronchial asthma 515
  cardiac disease 493–511
    aortic valve disease 506–7
    cardiomyopathies 507–9
    coronary insufficiency 503–5
    exertion relationship 519
    hypertensive 505–6
    myocarditis 508–9
    senile myocardium 509
    trauma relationship 518–19
    see also coronary atherosclerosis; myocardial infarction
  definition 492–3
  electrical fatalities 329
  in epilepsy 514–15
  gastrointestinal haemorrhage 516
  genitourinary system 517
  haemoptysis 516
  in morphine addicts 578
  in old age (senile myocardium) 509
  pathologists 492
  pneumothorax causing 226
  pulmonary thromboembolism 513–14, 520
  respiratory obstruction 515
  solvent abuse 595–6
  subarachnoid haemorrhage 196, 198, 511–12, 519–20
  suprarenal haemorrhage 349, 510
  trauma/disease relationship 517–20
  tuberculosis 516
  vagal cardiac arrest 370
  women of child-bearing age 511
sudden infant death syndrome (SIDS) 451–60
  autopsy
    failure 45, 456–7
    procedure 456
  autopsy findings 359, 360, 454–5
    external 453
    histological 455, 456
  body secretions 453
  case history 10, 453
  causation theories 455–6
  congenital disorders and 455
  counselling 456, 457
    organizations 457
  definition 451

sudden infant death syndrome (SIDS) (*contd.*)
  factors influencing risk 452–3
  gastric contents in air passages 363
  histology 455, 456
  incidence 451–2
  medico-legal problems 456–7
  multiple 457
  negative autopsy 45, 456
  petechiae 354
  petechial haemorrhage 454, 456–7
  posture at death 453
  publicity campaigns and 452, 455
  scene of death 453
  seasonal variation 453
  smothering of infants and 359
  'Stavanger definition' 451
  suffocation *vs* 456–7
  symptoms 453
sudden unexpected death (SUD) *see* sudden infant death syndrome (SIDS)
suffocation 352–67
  autopsy signs 360
  carbon monoxide, poisoning 357–8
  children 358
  cot deaths *vs* 456–7
  definition 353, 357
  human rights abuse 304
  plastic bag 359–60, 596
suicidal gestures 235
suicidal injuries 235–40
  fire-arm wounds 239
    automatic pistol 273
    rifled weapons 259, 261
    shotguns 239, 252, 253, 254, 255
  humane veterinary killer 273
  jumping from heights 240
  knife wounds 236–9
    abdominal 238
    chest 238
    clothing 8, 239
    cut throats 236–7, 237
    neck 238
    wrist 237–8
  multiple methods 237, 239
  sites 236
  stab wounds 238
  'tentative incisions' 236, 238
suicide
  agrochemical poisoning 566
  'asphyxia' 240
  in bathroom 400, 401, 405
  burning 239–40
  carbon monoxide poisoning 560
  choking due to foreign bodies 361
  corrosive poisons 587, 588
  in custody 308
  cyanide poisoning 585
  electrical fatalities 239, 326, 334
  falls from height 240
  hanging 240, 385, 388
  homicide *vs* 235, 236, 238, 239
  injuries before immersion in water 398
  by insulin 575
  ligature strangulation 240, 380
  medicine overdose 8
  methods 235
  by motor vehicle 297
  multiple methods 297
  notes 239, 391
  plastic bag suffocation 359, 360
  railway 297–8
  sexual asphyxia *vs* 390, 392
  submersion 239
sulphuric acid 588
supraorbital ridges, in sex determination 109
suprapubic angle 111
suprarenal artery, damage 232
suprarenal haemorrhage 232, 295, 349, 510
suprarenal vessels, damage 232
surgical procedures, deaths associated 480–7
  autopsy 480–2, 486
  bronchial obstruction 481
  classification system 482
  disease process causing 482–3
  histology 483, 485, 486
  medico-legal aspects 482, 483
  mode/causes of death 482–4
  pulmonary embolism 483, 520
  sudden death in gastrointestinal haemorrhage 516
  technique/equipment failure 483–4
  see also anaesthetic-related deaths
suspension, in torture 305
sutural diastasis 466
suxamethonium 486
syphilis 469, 511

'tache noire' 54
talc, granuloma 577
tape, samples 6
Tardieu's spots 354, 355, 364, 454
tartar emetic 591
tattooing, in gunshot wounds 252, 262–3
tattoos
  decomposition 101
  ethnic characteristics 100, 105
  identity establishment 100–1
  initials 100
  morphine addiction 579
  pigments 100, 101
  removal 101
  significance 100
taurodontism 534

teeth
  absence 535
  age determination 104–5, 534–5
  avulsion/breakage, in child abuse 473
  bite marks 530, 531–2
  bull-tooth (taurodontism) 534
  Carabelli's cusp 534
  choking due to 361
  eruption times 533, 534
  fillings 535, 537, 538
  impressions 531, 532
  loosened/broken 178
  milk 534
  occlusal attrition 534, 535
  pink 530, 538
  racemization of amino acids 105, 535
  racial differences 105, 533–4
  sex determination 109
  see also dentistry, forensic
'telefono' 304
Telkkä's table, stature calculation 116
temperature
  heat injury 312
  putrefaction 64, 67, 68
  scalds severity 314
  skin 79, 82
  tissue, electrical fatalities 330
  water, immersion deaths 369, 396, 400
temperature, ambient
  decomposition of buried bodies 68
  decomposition of immersed bodies 67
  effect on cooling curve 81
  low
    cerebral function recovery after immersion 369
    deaths from exposure 419
    hypothermia 415
    immersion deaths 396, 697, 698
  post-mortem interval estimation 7
  putrefaction 68
  rigor mortis timing 60
  succession of insects on corpses 76
temperature, body
  adipocere formation 70
  auditory meatus 7
  'core' 79, 416
  defects 312
  diurnal variations 80
  gradient for cooling 79
  initial, effect on cooling curve 80
  malignant hyperthermia 485
  measurement methods 81–2
  post-mortem interval estimation 7, 79–83
    see also cooling, post-mortem
  raised, sudden infant death syndrome 453
  rectal *see* rectal temperature
  skin 79, 82
  taking at scene of death 7
  'temperature ratio' concept 83, 87
  vitreous humour chemistry 88
temporal bone, petrous, haemorrhage 405
temporal bruise 176
tendon tags 126
tension pneumothorax 225
terrorist explosions 274–5, 277
terrorist weapons 248
testes
  fetal maturity estimation 448
  stimulation and sudden cardiac death 370
thallium poisoning 591–2
theophylline 515
thermal burns *see* burns
thermocouples 81, 82
thermometers 4, 81, 82
thermometry, post-mortem interval 82–4, 83
thighs, bruising, in sexual offences 422
thorax
  autopsy in air embolism 347, 436
  barotrauma 489
  contents, removal at autopsy 21
  internal examination at autopsy 18
  see also chest
throat
  blows to 371
  cut 236–7, 238
    air embolism in 346
thromboembolism, pulmonary 513–14
thrombosis
  coronary artery *see* coronary artery
  mesenteric 516
  mural 495, 502–3
  see also deep vein thrombosis; venous thrombosis
throttling, definition 353
'through-and-through' wounds 226, 227, 267, 268
thymic haemorrhage 454
thymus, petechial haemorrhages in 354
thyroid cartilage 376
  injuries, in manual strangulation 378
thyroid horn
  fractures 382, 388
    see also larynx, fractures
  injuries, in manual strangulation 377
tibia
  fractures 293
  stature estimation 117
tissue, sample collection 550
toluene abuse 596–7
tongue
  examination 23
  haemorrhages 378
tonsillar herniation 212
torture 301
  beatings 302–3

torture (*contd.*)
  electrical 304
  inversion in and postural asphyxia 365
  patterned electric marks 332
  scars 102, 303
  see also custody, deaths in
toxicology 541
  analyses 542–3
  laboratory, information for 544
  laboratory records 542
  mass disasters 44
  poisoning/drug abuse
    amphetamine 581
    antidepressants 573
    arsenic 590–1
    barbiturates 573–4
    cyanide 586–7
    ethylene glycol 589
    medicines 571, 572
    methadone 580
    narcotics 579–80
    organophosphorus pesticides 568
    paracetamol 572–3
    paraquat 567
  sample collection *see* sample collection
  solvent abuse 596–7
  surgical/anaesthetic-related deaths 481, 486
  toluene abuse 596
  see also poisoning
trachea
  compression 368
  soot inhalation 322–3
tracheobronchial tree, barium in 363
traffic accidents *see* transportation injuries; vehicular injuries
tranquilizer drugs, poisoning 573
transfusion accidents 482
transplant surgery 53
transport, body 7, 38
transportation injuries 281–300
  cause of death 295–7
  motorcyclists 290–1
  myocardial infarction in 496, 498
  pedal cyclists 291
  pedestrians *see* pedestrians
  railway 297–8
  suicide/homicide by 297
  vehicular *see* vehicular injuries
transsexualism 104
transvestism 391
tranylcypromine 573
trauma
  in acute alcoholism 557
  berry aneurysm rupture, and 198–9
  complications after *see* complications of injuries
  disease relationship 198–9, 517–20
    see also coronary atherosclerosis
  pulmonary embolism after 341, 520
  pulmonary fat embolism after 343
  subarachnoid haemorrhage and 196–7, 198, 519–20
  suprarenal haemorrhage after 349
traumatic asphyxia *see under* 'asphyxia'
trichloroethylene 486, 597
trichloromethane 597
tricyclic antidepressants, poisoning 573
Trotter and Gleser system 115
tuberculosis, sudden death 516
twinning, sudden infant death syndrome and 452–3
tyres, bruises/abrasions from 142, 143, 144

UK, sudden infant death syndrome in 452
ultraviolet light
  dating bones 127
  myocardial infarction 500
  semen testing 427–8
umbilical cord
  examination in infanticide 443
  post-partum changes 443, 445
unconsciousness 53
'undertaker's height' 11, 21, 103
unexpected death *see* sudden death
United States, sudden infant death syndrome in 452
urea 89
urethra, injuries 233
urine
  alcohol levels 555
  purging, in putrefaction 65
  sample collection methods 20, 548
  sample preservation 546
  toxicological analysis 543, 544
'Utus paste' 434

vagal inhibition 363, 370–1, 378, 399–400
  see also cardiac arrest, reflex
vagal reflex *see* cardiac arrest, reflex
vagina
  in abortion deaths 346, 432
  examination at autopsy 16
  in fatal sexual offences 426, 427
  foreign bodies 426
  perforation, in illegal abortion 432
  tearing/injuries 426, 427
vaginal discharge 12
vaginal fluid, DNA profiling sample 604
Valsalva experiments 371
vasectomy 428
vasovagal cardiac arrest *see* cardiac arrest, reflex
vasovagal reflex/shock *see* cardiac arrest, reflex
vegetative state 52–3
vehicles
  airbags 289
  exhaust fumes 560–1
  fragments, hit-and-run accidents 294, 296

  modern design/safety features 283, 286, 287–8, 289
    reversing warnings 293
    see also seatbelts
  rear-seat restraints 287
  seatbelts *see* seatbelts
  windscreen glass 282, 284
vehicular accidents
  in acute alcoholism 557
  concertinaing 281
  deceleration and G forces in 281
  delayed bleeding 340
  drivers 283–4
  ‘faking’ 297
  fires in 288, 297
  a frame 283
  frontal 281, 283
  ‘hit-and-run’ 294, 296
  intrusion of structures 283
  pedestrians in *see* pedestrians
  rear impact 281, 283, 284
  roll-overs 281, 283, 286, 287
  side-impacts 283
  suicide 297
  ‘tail-gating’ 290–1
  types 281
  unrestrained 282
vehicular injuries 281–91
  aorta ruptured 285
  autopsy 296–7
  brush abrasions 139, 290, 294
  cervical spine hyperflexion 215, 284
  clothing examination in 8, 296
  delayed death 295
  disease, natural 295–6
  dynamics 281–2
    see also deceleration injuries
  ejection 286, 287
    seatbelt effect 287
  facial 282, 284
  flail chest 223, 294
  ‘flaying’ 294
  head *see under* head injuries
  hyperflexion/hyperextension injuries 214, 215
  liver injuries 231, 232, 286, 287
  lung 285–6
  multiple 295
  natural deaths and 295–6, 296
  pattern abrasions from tyres 142, 143, 294, 296
  pattern of injuries 282–7
    driver 282–6
    front-seat passenger 287
    rear-seat occupants 287
  to pedestrians *see* pedestrians
  pulmonary haemorrhage 286
  radiator injuries 143
  ruptured aorta 228, 229, 285
  seatbelt effects *see* seatbelts
  steering wheel *see* steering wheel injuries
  suicidal and homicidal 297
  vulnerability of children 289–90
  see also deceleration injuries; *other specific injuries*
venereal disease 424, 426
venous congestion, in traumatic ‘asphyxia’ 364
venous obstruction, congestion in asphyxia 356
venous thrombosis 340–1, 341–2
  dating 342–3
  see also deep vein thrombosis
ventilation
  inadequate, carbon monoxide poisoning 561
  mechanical, adult respiratory distress syndrome 345
ventricular fibrillation
  drowning 401
  electrical fatalities 328, 329, 333
vernix caseosa 443
vertebral artery
  forensic anatomy 201
  occlusion 369
  trauma 200–4
    autopsy appearances 201–2
    autopsy demonstration 202–4
    investigations 203
    mechanism 202
    spinal blocks 203
    vulnerable points 202
vertebral bodies, compression fractures 214, 215
vertebro-occipital block 203
vessel fragility, bruising and 145
veterinary pistols 273
vials, sample collection 549
video cameras 128
Virchow’s triad 348–9, 520
virus infections, myocarditis 509
virus infections, in sudden infant death syndrome 456
viscera, examination at autopsy 21, 23
‘vital reaction’ 169, 171, 317, 332, 363
vitreous humour
  chemistry, post-mortem interval estimation 89–90
  glucose levels 90, 575
  post-mortem interval estimation 88–9
  potassium levels 88, 89
  sample collection 20–1, 88, 548–9
  sample preservation 546
volatile substances
  sample collection 549–50, 597–8
  see also solvent abuse
voltage 328
voluntary organizations 412, 457
vomit
  aspiration 361–2, 557
  external examination at autopsy 12
  sample collection method 549
vomiting, in solvent abuse 596

vulva
  examination at autopsy 16
  injuries 233
    in sexual offences 423, 424

Wales, dating skeletal remains in 126
walkers, deaths from exposure and 419
'washer-woman's skin' 395, 396
wasting conditions 414
water
  adipocere formation 70
  carbon monoxide poisoning 561
  deprivation 412
  explosions in 275
  immersion deaths and *see* immersion deaths
  in plastic bags 360
Waterhouse–Friederichsen syndrome 349
Webley veterinary pistol 273
Weedol 566
weight
  body 603
    external examination 12
  brain, at autopsy 29
  heart *see* heart, weight
wetcutting, brain, at autopsy 27
whip, multi-thonged 303
whiplash injury 215, 283, 284
whipping, torture and human rights abuse 303
WHO method, staining 602
Widmark equation/factor 556
Wilson's classification of burns 312–13, 313
World Health Organization (WHO)
  cardiomyopathies classification 508
  cardiomyopathies definition 507
  death certificates 55
  staining method 602
  sudden death definition 492–3
  'undetermined death' 10
  wound classification 137
wounding, survival period after 169–71
wounds
  'ante-mortem' *vs* 'post-mortem' 167–9, 170–1
  chronological histological changes 167–8
  classification and types 137
  dating 166–9, 170, 171
  defence 165–6, 167, 168
  definition 136
  gunshot/firearm *see* rifled weapons, wounds; shotgun wounds
  healing 103, 167–8
  histochemical changes 168–9
  mechanism 136–7
  negative/positive vital reactions 169
  pathology 136–73
  razors/glass and china 163–4
  scissor 164–5, 166
wrists
  slashing, scars in identification 102
  suicidal cutting 237–8

XTC (Ecstasy) 581
xylene abuse 597

Zahn, lines of 503, 510
zonal lesions 499
zygomatic process, sex determination 109

Copyrighted Material

John V. Forrester • Andrew D. Dick
Paul G. McMenamin • Fiona Roberts • Eric Pearlman

THE EYE

BASIC SCIENCES IN PRACTICE

Fourth Edition

ELSEVIER

Copyrighted Material

The Eye

Basic Sciences in Practice

To Anne, Lindsey, Christine and Lucy

Senior Content Strategist: Jeremy Bowes
Content Development Specialist: Helen Leng
Project Manager: Andrew Riley
Designer: Miles Hitchen
Illustration manager: Amy Naylor
Illustrator: Electronic Publishing Services, Inc.

4th Edition

The Eye

Basic Sciences in Practice

John V. Forrester, MB ChB MD FRCS(Ed) FRCP(Glasg) (Hon) FRCOphth(Hon) FMedSci FRSE FARVO
Professor of Ophthalmology and Head of Department of Ophthalmology, University of Aberdeen, Aberdeen, UK; Section of Immunology and Infection, University of Aberdeen, UK; Ocular Immunology Program, The University of Western Australia, Australia; Centre for Experimental Immunology, Lions Eye Institute, Western Australia, Australia

Andrew D. Dick, BSc MB BS MD FRCP FRCS FRCOphth FMedSci FARVO
Professor of Ophthalmology and Head of Academic Unit of Ophthalmology
University of Bristol
Bristol, UK

Paul G. McMenamin, BSc MSc(MedSci) DSc (Med) PhD
Director of Centre for Human Anatomy Education, Department of Anatomy and Developmental Biology, Monash University, Melbourne, Victoria, Australia

Fiona Roberts, BSc MB ChB MD FRCPath
Consultant Ophthalmic Pathologist and
Honorary Senior Lecturer in Pathology
University Department of Pathology
Southern General Hospital
Glasgow, UK

Eric Pearlman BSc PhD
Director, Institute of Immunology
Professor, Departments of Ophthalmology and Physiology, University of California, Irvine
Professor and Director of Research in the Department of Ophthalmology and Visual Sciences, Case Western Reserve University, Cleveland, Ohio

ELSEVIER
EDINBURGH LONDON NEW YORK OXFORD PHILADELPHIA
ST LOUIS SYDNEY TORONTO 2016

ELSEVIER

© 2016 Elsevier Limited. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the publisher (other than as may be noted herein).

First edition 1996
Second edition 2002
Third edition 2008
Fourth edition 2016

ISBN 978-0-7020-5554-6

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book is available from the Library of Congress

NOTICES
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

With respect to any drug or pharmaceutical products identified, readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of practitioners, relying on their own experience and knowledge of their patients, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions.

To the fullest extent of the law, neither the publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ELSEVIER your source for books, journals and multimedia in the health sciences
www.elsevierhealth.com

ELSEVIER
Book Aid International
Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org

The publisher's policy is to use **paper manufactured from sustainable forests**

Printed in China

Preface

The fourth edition of *The Eye* is upon us and, in the intervening years since the third edition, much has happened in the life sciences which have a direct bearing on the basic science of the eye. For instance, the massive strides being taken by the new genetics and functional genomics based on the Human Genome Project, the new understanding of how the microbiome affects all aspects of immunology, the remarkable new imaging technology which is applied to anatomy and neurophysiology, the exciting new molecular, and other, diagnostic methodologies being used in microbiology and pathology, have collectively brought a wealth of new knowledge to students and practitioners in the fields of ophthalmology and visual science. For these reasons alone, we have felt that there is strong need to update the text and allow our continuing and new readership to view these developments as they inform the various scientific disciplines in how they affect the eye and its workings.

And so, *The Eye Fourth Edition* has been extensively revised with much new text as well as many new figures. The aim of the book, as always, has been to provide a concise text which provides as much information in the simplest and most readable format which can be used by a diverse readership. This includes optometrists and ophthalmologists in training, and in practice, as well as new recruits to visual and ophthalmic science who are embarking on a career in the basic science of the eye. We expect readers will take from the text those aspects of knowledge and information which are directly relevant to them and hope that they will also dip into areas which might not seem so immediately important to them but remain part of the whole. As indicated in the preface to the third edition, the purpose is to produce a "basic science of the eye" handbook which is readily accessible.

The book retains its familiar overall organisation in terms of subject matter. Excitingly, however, in this internet dawn, *The Eye* is also being produced as an online text, with links to additional information as well as video clips prepared by the authors which are aimed to help explain and expand on aspects of the basic science. We hope that our new, as well as our previous, students and readers enjoy the new product.

John V. Forrester

Andrew D. Dick

Paul G. McMenamin

Fiona Roberts

Eric Pearlman

Acknowledgements

We would like to thank Professor William R. Lee, who co-authored the first two editions of *The Eye*, for his guidance and support and generous provision of material for the current edition. We are also grateful to the anonymous reviewers who commented on the draft text.

Paul McMenamin would like to thank the following for discussions relating to the revisions for the third edition: Professor Alan Harvey (the anatomy of the visual pathways), Dr Joseph Demer (anatomy of extraocular muscles) and Professor Brian Hall (embryology of the head and neck).

John V. Forrester

Andrew D. Dick

Paul G. McMenamin

Fiona Roberts

Eric Pearlman

Chapter 1

Anatomy of the eye and orbit

- Anatomical terms of reference
- Osteology of the skull and orbits
- Structure of the eye
- Orbital contents
- Cranial nerves associated with the eye and orbit
- Ocular appendages (adnexa)
- Anatomy of the visual pathway

Anatomical terms of reference

The internationally accepted terminology for description of the relations and position of structures in the body requires reference to a series of imaginary *planes* (Fig. 1-1). Thus, relative positions of anatomical structures are referred to in terms of: *medial* (nearer the median or mid-sagittal plane) and *lateral* (away from this plane); *anterior* and *posterior* refer to the front and back surfaces of the body; *superior* (cranial or rostral) or *inferior* (caudal) refer to position in the vertical; *superficial* and *deep* specify distance from the surface of the body. A combination of terms can be used to describe the relative position of structures that do not fit exactly any of the other terms, e.g. ventrolateral, posteromedial, etc.

Osteology of the skull and orbits

GENERAL ARRANGEMENT AND FEATURES OF THE SKULL

The skull is divided into two parts: an upper part shaped like a bowl, which contains the brain, known as the *cranium* or *neurocranium*; and a lower part, the *facial skeleton* or *viscerocranium*. The cranium can be further subdivided into the cranial vault and cranial base.

The skull is composed of a large number of separate bones that are united by *sutures* (fibrous immovable joints). The cranium consists of eight bones (only two are paired), the facial skeleton consists of 14 bones, of which only two are single (Fig. 1-2A,B). The skull contains a number of cavities that reflect its multiple functions:

- *cranial cavity* – houses, supports and protects the brain
- *nasal cavity* – concerned with respiration and olfaction
- *orbits* – contain the eyes and adnexa
- *oral cavity* – start of gastrointestinal tract, responsible for mastication and initial food processing; houses taste receptors.

Many of the cranial bones contain air-filled spaces, the paranasal sinuses (Fig. 1-3). Most of the anatomical features of the whole skull relevant to the study of the eye and orbits are indicated in Figure 1-2A and B (*norma frontalis* and *norma lateralis*). (See video 1-1).

OSTEOLOGY OF THE ORBIT

The orbital cavities, situated between the cranium and facial skeleton, are separated from each other by the nasal cavity and the ethmoidal and sphenoidal air sinuses (Fig. 1-3A–C). Each bony orbit accommodates and protects the eye and adnexa, and serves to transmit the nerves and vessels that supply the face around the orbit. Parts of the following bones contribute to the walls of the orbit: maxilla, frontal, sphenoid, zygomatic, palatine, ethmoid and lacrimal (Figs 1-4 and 1-5A,B). Each orbit is roughly the shape of a quadrilateral pyramid whose base is the *orbital margin* and whose apex narrows at the *optic canal*. Each orbit has a floor, roof, medial wall and lateral wall (Fig. 1-4). (See video 1-2).

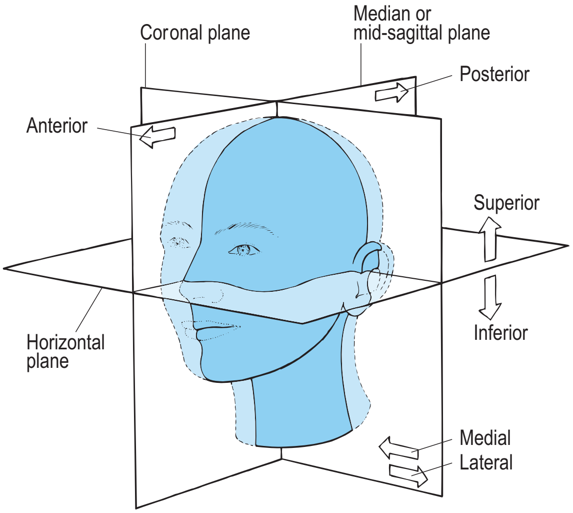

FIGURE 1-1 Diagram illustrating anatomical planes of reference.

The floor tapers off before reaching the apex; therefore the apex of the pyramid is triangular in shape. The orbit is widest approximately 1.5 cm behind the orbital margin. The medial walls are approximately parallel to the mid-sagittal plane, while the lateral walls are oriented at an angle of approximately 45° to this plane. The *orbital aperture* is directed forwards, laterally and slightly downwards, a characteristic of primates and indeed predators which require binocular vision. Nerves and muscles passing from the apex into the orbit pass forward and laterally (Fig. 1-3A,B). The orbit is approximately 40 mm in height, 40 mm in width and 40 mm in depth. The volume is approximately 30 mL, of which one-fifth is occupied by the eye.

The walls of the orbit

The bones that make up the roof, floor and medial and lateral walls are summarized in Figure 1-4.

Features of the orbital roof

- *Fossa for the lacrimal gland:* lies in the anterolateral aspect of the roof behind the zygomatic process of the frontal bone.
- *Trochlear fossa (fovea):* lies in the anteromedial aspect of the roof, 4 mm from the margin, and is the site at which the trochlea (small pulley) is attached. The tendon of the superior oblique passes through the trochlea.
- *Anterior and posterior ethmoidal canals* Positioned at the junction of roof and medial wall above the frontoethmoidal suture (Fig. 1-5A). They transmit the anterior and posterior ethmoidal nerves and vessels.

Relations. The roof, which is thin and translucent except at the lesser wing of the sphenoid, separates the orbit from the anterior cranial fossa and frontal lobes of the brain. Anteriorly, the frontal sinus lies above the orbit.

Features of the medial orbital wall

- This wall is oblong in shape and thin (0.2–0.4 mm). The four bones that comprise this wall are separated by vertical sutures (Figs 1-4 and 1-5A).
- *Lacrimal fossa* for the lacrimal sac: it is bound by anterior and posterior lacrimal crests and is continuous below with the *nasolacrimal canal* (Fig. 1-5B).

Relations. This is the thinnest of the walls and is largely transparent or semitransparent – the ethmoidal air sinuses can easily be seen through this wall in a dried skull (Fig. 1-5A,B). Medial to this wall in an anterior to posterior sequence lie the anterior, middle, posterior ethmoidal air cells and the sphenoidal sinus.

Features of the orbital floor

- The floor slopes slightly downwards from the medial to the lateral wall.
- It is crossed by the *infraorbital groove*, which runs forward from the *inferior orbital fissure*. Before it reaches the orbital margin this fissure becomes the *infraorbital canal*, which opens as the *infraorbital foramen* 4 mm below the orbital margin on the anterior surface of the maxilla (Figs 1-3C, 1-4 and 1-5A).

Relations. The floor separates the orbit from the maxillary sinus, the bone being only 0.5–1 mm in thickness (Fig. 1-3C).

Features of the lateral orbital wall (Fig. 1-4)

- *Spina recti lateralis:* a small bony spine on the greater wing of the sphenoid near the apex of the orbit which gives origin to part of the lateral rectus.
- *Zygomatic foramen:* transmits zygomatic nerve and vessels to temporal fossa and cheek (zygomaticotemporal nerve and zygomaticofacial nerve) (Fig. 1-5B).

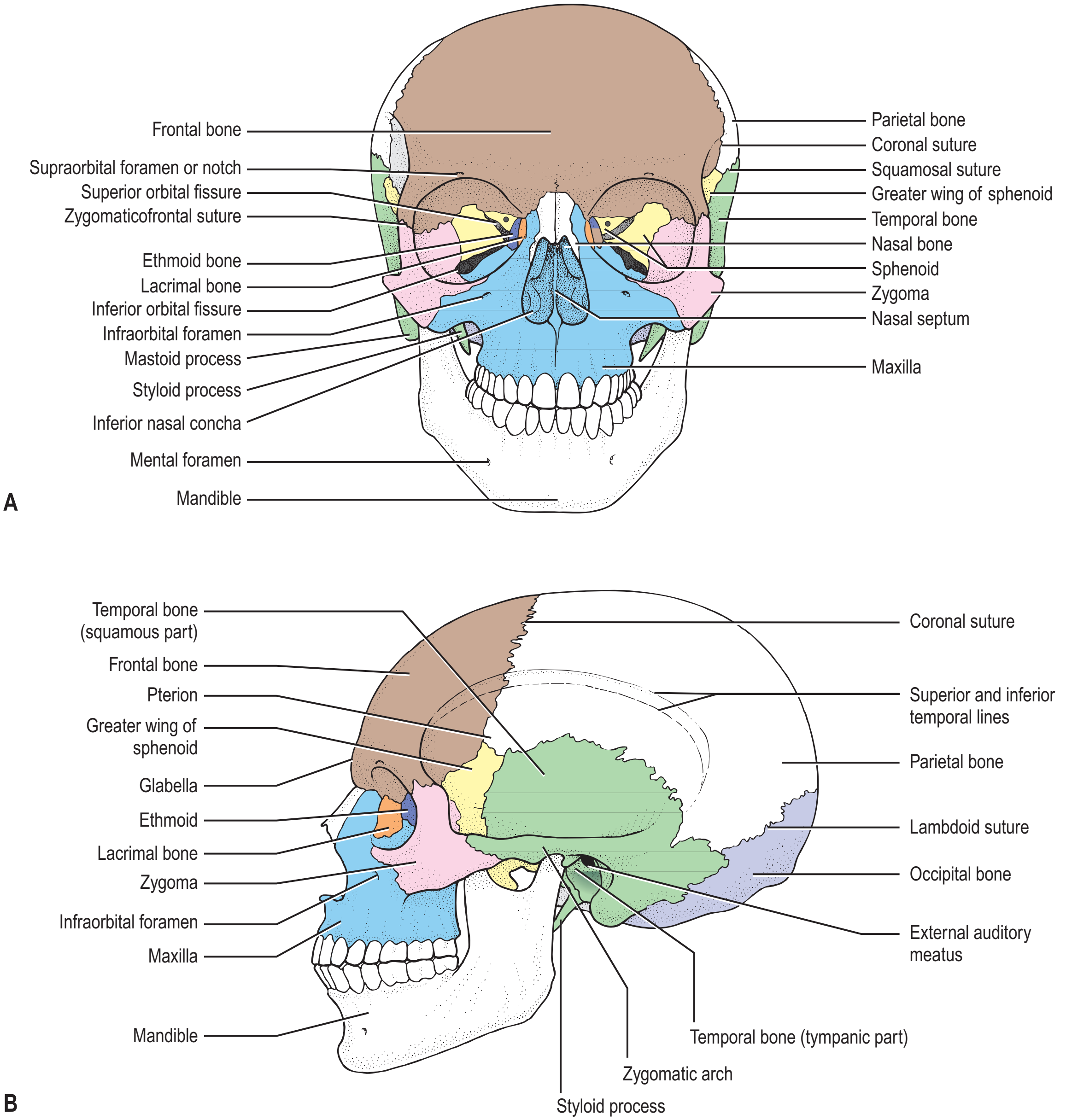
Frontal bone
Supraorbital foramen or notch
Superior orbital fissure
Zygomaticofrontal suture
Ethmoid bone
Lacrimal bone
Inferior orbital fissure
Infraorbital foramen
Mastoid process
Styloid process
Inferior nasal concha
Mental foramen
Mandible
A
Parietal bone
Coronal suture
Squamosal suture
Greater wing of sphenoid
Temporal bone
Nasal bone
Sphenoid
Zygoma
Nasal septum
Maxilla
Temporal bone (squamous part)
Frontal bone
Pterion
Greater wing of sphenoid
Glabella
Ethmoid
Lacrimal bone
Zygoma
Infraorbital foramen
Maxilla
Mandible
B
Coronal suture
Superior and inferior temporal lines
Parietal bone
Lambdoid suture
Occipital bone
External auditory meatus
Temporal bone (tympanic part)
Zygomatic arch
Styloid process

FIGURE 1-2 Osteology of the skull. Two views of the skull: (**A**) norma frontalis; and (**B**) norma lateralis to illustrate the individual bones and important anatomical landmarks.

- *Lateral orbital tubercle:* forms the attachment of the check ligament of the lateral rectus, suspensory ligament (Lockwood's) of the eye, superior transverse ligament (Whitnall's) and aponeurosis of levator palpebrae superioris.
- Foramina for small veins that communicate with middle cranial fossa.

Relations. Laterally – skin, temporal fossa and middle cranial fossa in an anterior–posterior sequence (Fig. 1-3A).

Orbital margin, fissures and optic canal

Orbital margin. This is a thickened rim of bone that helps protect the orbital contents. It is made up of three bones: the frontal, zygomatic and maxilla

BOX 1-1 CLINICAL CORRELATES

Orbital cellulitis

This condition may be a consequence of infection spreading from the air sinuses to the orbit via the paper-thin medial wall (*lamina papyracea*) that separates the two. An example is shown of a patient with orbital cellulitis following creation of a drainage fistula (**A**). A coronal CT (**B**) illustrates the communication with the ethmoidal sinuses and nasal cavity (arrow).

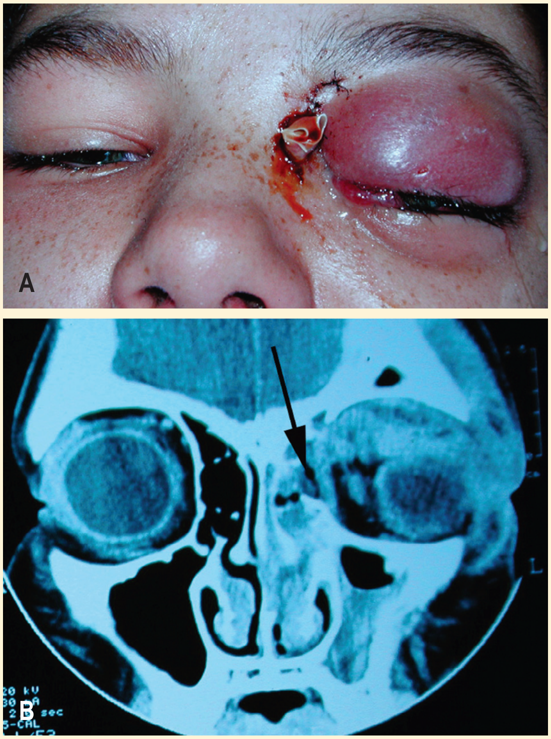

(Images courtesy of Dr Alan McNab.)

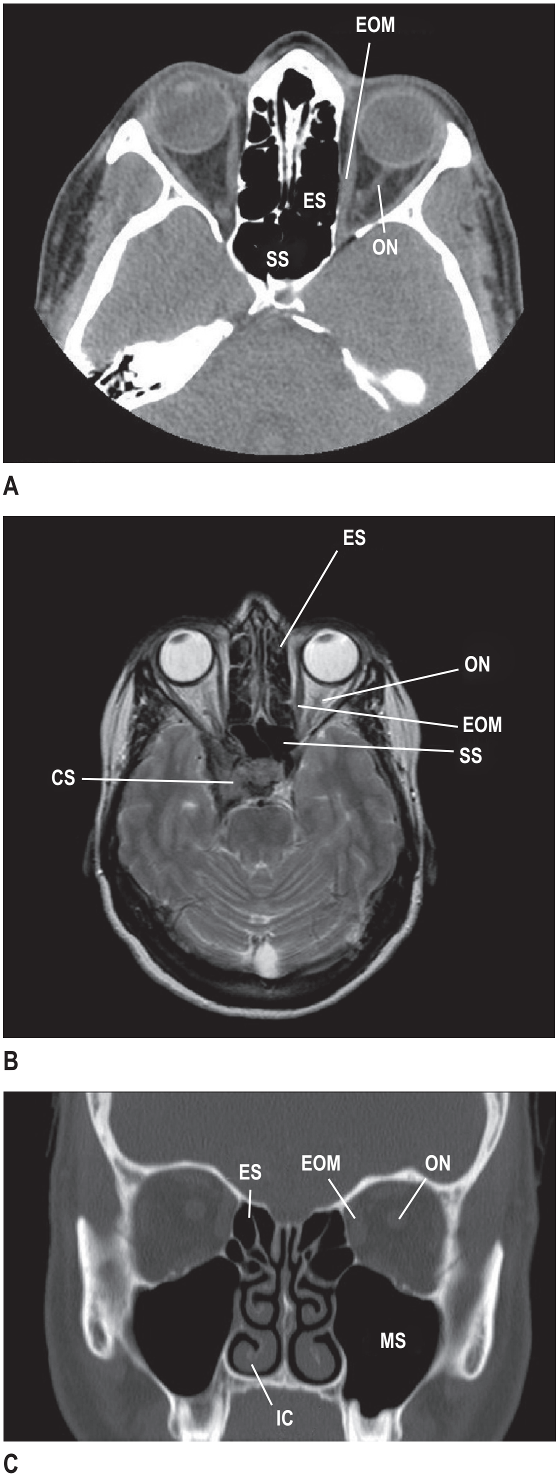

FIGURE 1-3 Transverse (**A,B**) and coronal (**C**) computed tomography (CT: **A** and **C**) and magnetic resonance imaging (MRI: **B**) scans of the head displaying the major relations of the orbits. Features identifiable in the scans include ethmoid air cells/sinuses (ES), maxillary sinus (MS), sphenoid sinus (SS), nasal cavity (NC), inferior nasal concha (IC), extraocular muscle (EOM), optic nerve (ON) and cavernous sinus (CS).

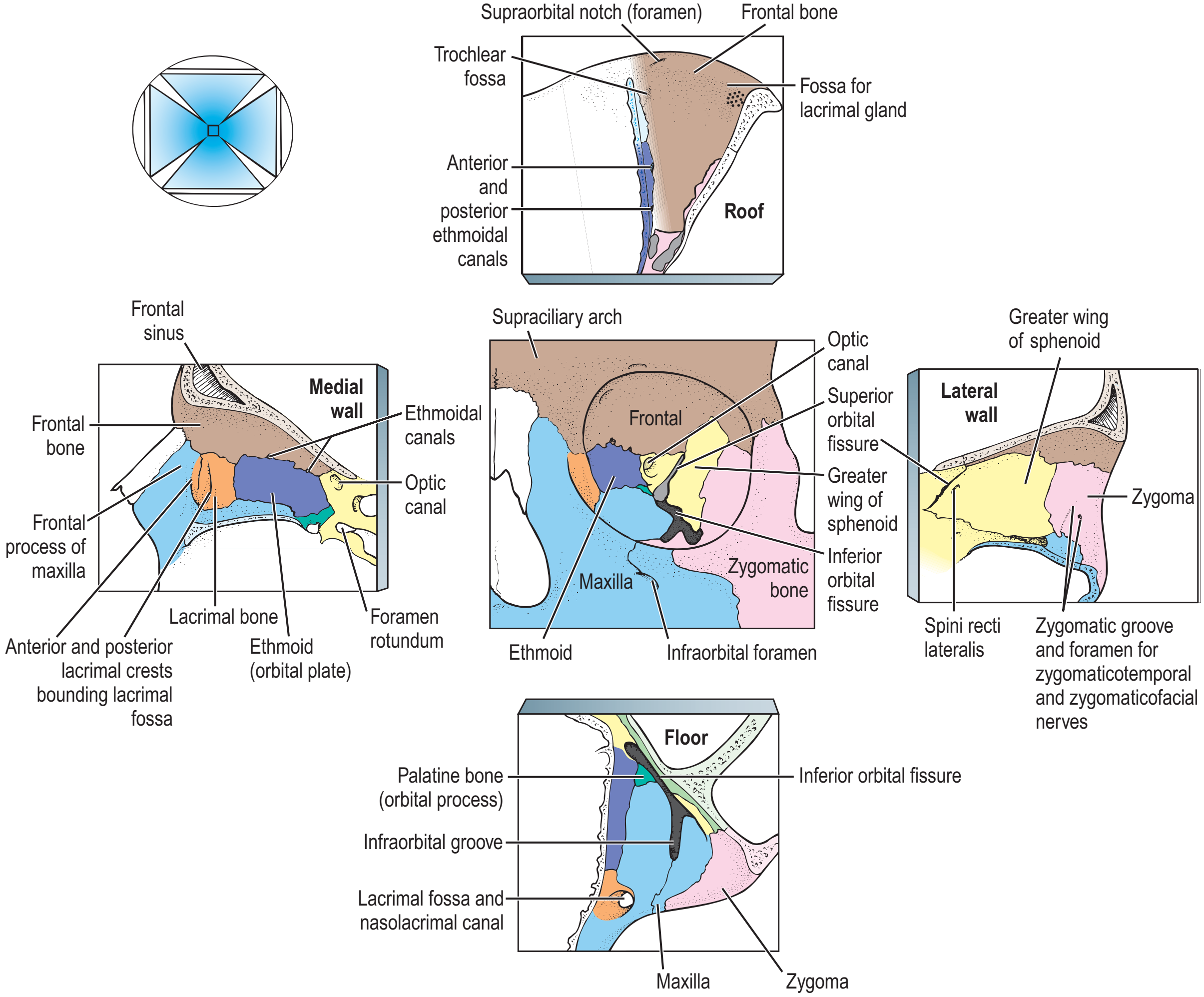
Supraorbital notch (foramen)
Frontal bone
Trochlear fossa
Fossa for lacrimal gland
Anterior and posterior ethmoidal canals
Roof
Frontal sinus
Medial wall
Frontal bone
Ethmoidal canals
Optic canal
Frontal process of maxilla
Lacrimal bone
Foramen rotundum
Anterior and posterior lacrimal crests bounding lacrimal fossa
Ethmoid (orbital plate)
Supraciliary arch
Frontal
Optic canal
Superior orbital fissure
Greater wing of sphenoid
Inferior orbital fissure
Maxilla
Zygomatic bone
Ethmoid
Infraorbital foramen
Greater wing of sphenoid
Lateral wall
Zygoma
Spini recti lateralis
Zygomatic groove and foramen for zygomaticotemporal and zygomaticofacial nerves
Floor
Palatine bone (orbital process)
Inferior orbital fissure
Infraorbital groove
Lacrimal fossa and nasolacrimal canal
Maxilla
Zygoma

FIGURE 1-4 Osteology of the orbit. The central diagram illustrates the anterior view of the intact left orbit, the four surrounding diagrams ('exploded orbit' – see inset, top left) show the individual bones which form the roof, floor, medial and lateral walls and other noteworthy features. *Roof:* orbital plate of the frontal bone and small area of lesser wing of the sphenoid. *Medial wall:* frontal process of the maxilla, lacrimal bone, orbital plate of the ethmoid and the body of the sphenoid. *Floor:* orbital plate of the maxilla, orbital surface of the zygoma and the orbital process of the palatine bone. *Lateral wall:* orbital surfaces of greater wing of sphenoid posteriorly and zygomatic bone anteriorly.

(Figs 1-4 and 1-5A,B). The lateral margin does not reach as far anteriorly as the medial margin (see Figs 1-2B and 1-5B). The medial margin is sharp and distinct in its lower half because of the anterior lacrimal crest, but is indistinct superiorly (Figs 1-4 and 1-5B).

Superior orbital fissure. This communication between the orbital and cranial cavities lies between the roof and lateral wall of the orbit and is bounded by the lesser and greater wings of sphenoid (Figs 1-4, 1-5A and 1-6). It is wider at its medial end and narrowest at its lateral end. It is around 22 mm long and is separated from the optic foramen above by the posterior root of the lesser wing of the sphenoid. The part of the *common tendinous ring* that gives origin to the lateral rectus spans between the narrow and wide parts of the fissure. Structures passing above or outside the tendinous ring or annulus include the lacrimal nerve, frontal nerve, trochlear nerve, superior ophthalmic vein and recurrent branch of the lacrimal

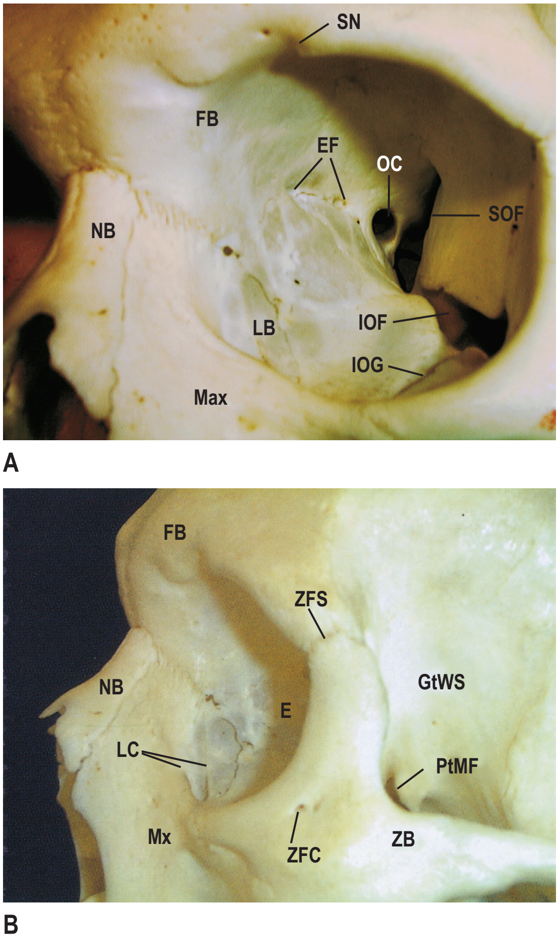

FIGURE 1-5 (**A**) Anterior view of the bony orbit showing important osteological features of the apex including the relation of the superior orbital fissure (SOF), optic canal (OC) and inferior orbital fissure (IOF). IOG, inferior orbital groove; LB, lacrimal bone; EF, anterior and posterior ethmoidal foramina; SN, supraorbital notch; FB, frontal bone; NB, nasal bone; Max, maxilla. (**B**) Lateral view of the orbit. ZB, zygomatic bone; PtMF, pterygomaxillary fissure; GtWS, greater wing of the sphenoid; ZFS, zygomaticofrontal suture; ZFC, zygomaticofacial canal; LC, lacrimal crest; E, ethmoid bone; Mx, maxilla.

artery. The latter anastomoses with the orbital branch of the middle meningeal artery and may more commonly travel in a small cranio-orbital foramen lateral to the superior orbital fissure. Structures passing within the ring, and thus within the apex of the muscle cone, include the oculomotor nerve (superior and inferior divisions), abducent nerve, nasociliary nerve, sympathetic root of the ciliary ganglion, and variably the inferior ophthalmic vein (Fig. 1-6).

BOX 1-2 CLINICAL CORRELATES

Orbital blow-out fractures

The floor, although thicker than the medial wall, is more often involved in orbital blow-out fractures, probably because it lacks the buttress-like supports of the ethmoidal air cells and the protection of the nose. Tumour spread to or from the maxillary sinus may occur via the floor of the orbit.

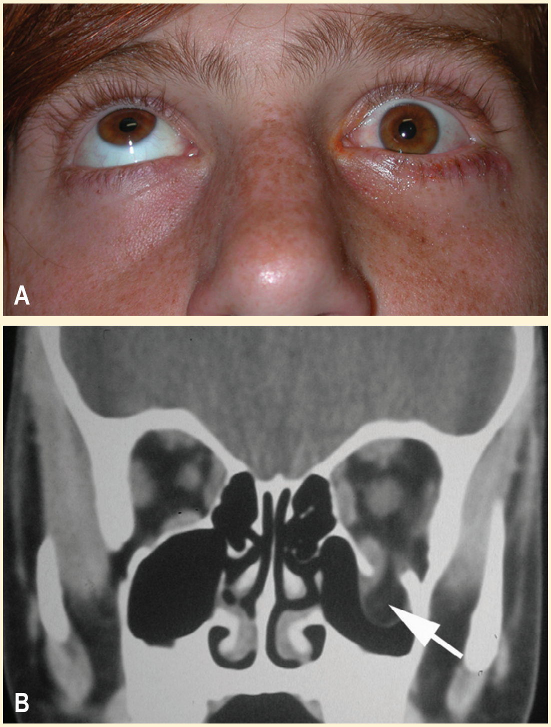

An example of ocular motility being compromised in left eye following blow-out fracture (**A**) and a coronal CT (**B**) showing the herniation of orbital contents into the roof of the maxillary sinus (arrow).

(Images courtesy of Dr Alan McNab.)

Inferior orbital fissure. This fissure lies between the lateral wall and floor of the orbit below the superior orbital fissure. It forms a communication between the orbit and the infratemporal fossa and pterygopalatine fossa. It runs forward and laterally for approximately 20 mm and ends 20 mm from the orbital margin (Figs 1-4 and 1-5A). The fissure is narrowest in the middle section and in life is covered by periorbita and a sheet

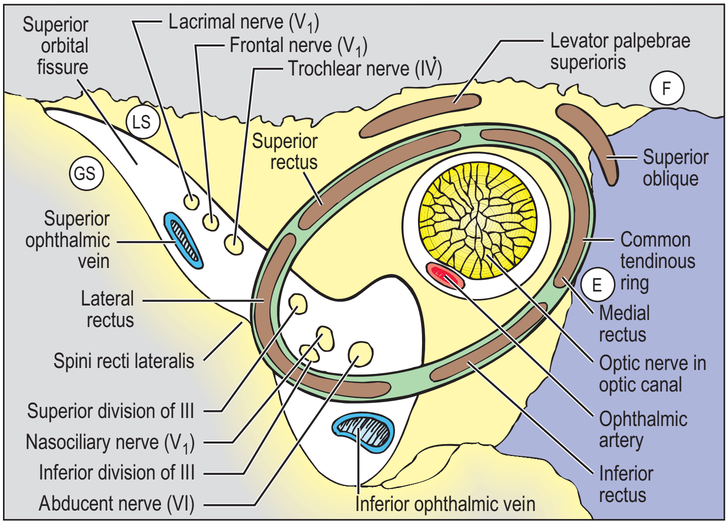
Superior orbital fissure
Lacrimal nerve (V_1)
Frontal nerve (V_1)
Trochlear nerve (IV)
Levator palpebrae superioris
F
LS
GS
Superior rectus
Superior oblique
Superior ophthalmic vein
Common tendinous ring
E
Lateral rectus
Medial rectus
Spini recti lateralis
Optic nerve in optic canal
Superior division of III
Ophthalmic artery
Nasociliary nerve (V_1)
Inferior division of III
Inferior rectus
Abducent nerve (VI)
Inferior ophthalmic vein

FIGURE 1-6 Schematic diagram of the superior orbital fissure and optic canal in the right orbit. Note the origins of the extraocular muscles from the common tendinous ring and the relative position of the cranial nerves and vessels as they enter or exit the orbit. GS, greater wing of sphenoid; LS, lesser wing of sphenoid; F, frontal bone; E, ethmoid. The positions of the veins are variable. The first letters of each of the structures passing through the superior orbital fissure (LFTSNIA) form a well-known mnemonic.

of smooth muscle of unknown function, the orbitalis or 'muscle of Müller'. It transmits the infraorbital nerve, zygomatic nerve, branches from the pterygopalatine ganglion and the inferior ophthalmic vein may communicate with the pterygoid venous plexus below.

Optic canal. This is a bony channel in the sphenoid that passes anteriorly, inferiorly and laterally (36°) from the middle cranial fossa to the apex of the orbit. The canal is formed by the two roots of the lesser wing of the sphenoid. The optic canals are 25 mm apart posteriorly and 30 mm anteriorly. Each is funnel-shaped and narrowest anteriorly where its opening into the orbit is oval with sharp upper and lower borders and a prolonged roof (10–12 mm in length). The opening at the cranial aspect is oval with a prolonged floor. The sphenoidal and posterior ethmoidal air sinuses are important medial relations, and the olfactory tracts are superior relations of the canal. The canal transmits the *optic nerve* with its meningeal coverings and the *ophthalmic artery*, which lies below and then lateral to the nerve within the dural sheath for part of its course (Fig. 1-6). Sympathetic nerve fibres accompany the artery.

PARANASAL SINUSES

The paranasal sinuses (Figs 1-3, 1-7 and Video 1.3) comprise the frontal, ethmoidal, sphenoidal and maxillary sinuses. They are air-filled cavities in the skull that are in communication with the nasal cavity via a series of apertures. Infection commonly spreads from the nasal cavity into the sinuses. The sinuses function to warm and moisten the air, add resonance to the voice and lighten the skull. They vary in size and shape between individuals.

Additional content available at https://expertconsult.inkling.com/.

CRANIAL CAVITY (FIG. 1-8A,B AND EFIG. 1-1)

The cranial cavity houses the brain, cerebral vessels, the meninges, meningeal vessels and the intracranial portions of the cranial nerves. The base of the cranial cavity can be subdivided for descriptive purposes into three fossae: anterior, middle and posterior. Accounts of the detailed anatomy of these fossae can be found in any standard anatomy text; therefore only features of relevance to the eye and orbit in the anterior and middle cranial fossae will be described.

For discussion of posterior cranial fossa see https://expertconsult.inkling.com/.

Cranial fossae

Anterior cranial fossa. The anterior cranial fossa is limited in front and laterally by the frontal bone and posteriorly by the lesser wing of the sphenoid. Its floor is formed by the orbital plate of the frontal bone, the cribriform plate of the ethmoid (with a median

FRONTAL SINUSES (FIGS 1-4 AND 1-7)

The frontal sinuses are paired and lie behind the superciliary arches within the frontal bone (eFig. 1-1). They are separated from each other or further subdivided by thin bony septa that are not necessarily in the midline. The sinuses may extend as far laterally as the zygomatic process of the frontal bone. Each is approximately triangular and extends highest above the medial end of the eyebrow (Fig. 1-7). Each sinus opens into the middle meatus of the nasal cavity, either through the ethmoidal infundibulum or directly via the frontonasal duct. The mucosal lining is supplied by the supraorbital nerves and vessels; hence, referred pain from frontal sinusitis is experienced along the course of the supraorbital nerve.

Recent geometric morphometric studies (elliptic Fourier analysis) of the outlines of frontal sinuses from large numbers of radiographic images have confirmed a long-held belief that each individual's frontal sinus is distinct and unique. This may have important applications for personal identification in the context of forensics.

ETHMOIDAL SINUSES (AIR CELLS) (FIGS 1-3A–C AND 1-7)

These thin-walled sinuses are for the most part situated in the lateral mass of the ethmoid, although frontal, maxillary, lacrimal, sphenoidal and palatine bones contribute to the walls. They are variable in number and are grouped into anterior, middle and posterior. The general pattern of drainage of the sinuses is as follows: the anterior opens into the hiatus semilunaris, the middle on to the bulla ethmoidalis (both middle meatus) and the posterior into the superior meatus. They are related to the frontal sinus anteriorly, the sphenoidal sinus posteriorly, the nasal cavity medially and below, and laterally the orbit (Fig. 1-7A).

SPHENOIDAL SINUS (FIGS 1-3A,B AND 1-7)

This sinus lies within the body of the sphenoid bone and possesses an indented roof because of the pituitary fossa that lies above and houses the pituitary gland (Fig. 1-8). It may be divided by a variable midline septum. A transverse ridge in the lateral wall marks the position of the internal carotid artery (within the cavernous sinus). Other important relations of the sinus include the optic chiasma and nerves above, the nasal cavity below, the ethmoidal sinuses anteriorly, and the paired cavernous sinuses laterally (Fig. 1-9C). The sphenoidal sinus drains into the superior meatus or sphenoethmoidal recess. Surgical access to the pituitary gland may be gained via the nasal cavity and sphenoidal sinus; hence, surgeons must be aware of the above-mentioned relations.

MAXILLARY SINUS (FIGS 1-3C AND 1-7)

These are the largest of the paranasal air sinuses. They are pyramidal in shape and lie within the body of the maxilla. The *base* forms part of the lateral wall of the nasal cavity and the *apex* is within the zygomatic process. Each sinus is in communication with the middle meatus of the nasal cavity via an aperture, the maxillary hiatus, on its base, which empties into the lower part of the hiatus semilunaris. The opening is positioned high on this wall and therefore does not facilitate gravitational drainage in the upright position. The *nasolacrimal duct* lies in a thin bony canal in the anterior part of the base. The orbital plate forms the *roof* of the sinus and floor of the orbit.

ORBITAL FLOOR FRACTURES

Rapid traumatic compression of the orbital contents, such as occurs during squash ball injuries, can lead to blow-out fractures; orbital contents may herniate into the maxillary sinus. It was once thought that orbital contents, including extraocular muscles, became trapped in the fractured floor, thus restricting range of movement and explaining the diplopia suffered by these patients. However, recent studies have indicated that in many cases only orbital fibroadipose tissue is trapped in the damaged floor of the orbit (see p. 6).

The *floor* of the maxillary sinus is formed by the alveolar process housing a variable number of the roots of the first and second molars that protrude into the sinus, and may be separated from the sinus by only a thin covering of bone or mucous membrane. Thus sinusitis may present as referred pain such as toothache and vice versa. In addition, abscesses in the maxillary sinus may result from infection of these roots. The *anterior/lateral wall* is directed on to the face, and access for drainage of maxillary obstructions or other surgical procedures in the sinus may be gained by this route. The *posterior wall* faces the infratemporal fossa.

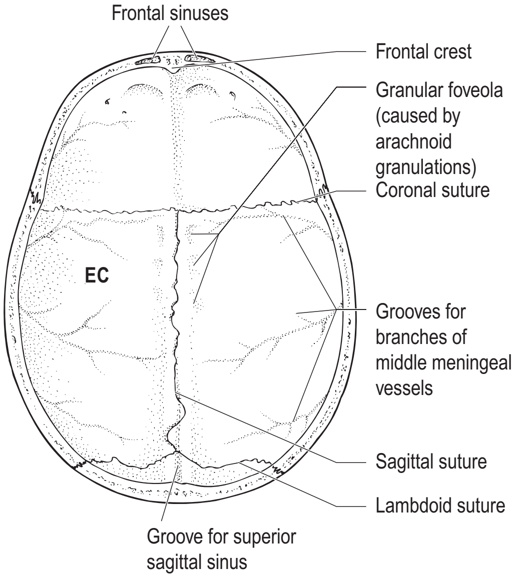
Frontal sinuses
Frontal crest
Granular foveola (caused by arachnoid granulations)
Coronal suture
EC
Grooves for branches of middle meningeal vessels
Sagittal suture
Lambdoid suture
Groove for superior sagittal sinus

eFIGURE 1-1 Features of the vault interior.

POSTERIOR CRANIAL FOSSA

This is the deepest of the three cranial fossae, its floor lying below the level of the middle fossa. Its roof is formed by the tentorium cerebelli. It lodges the hindbrain: the cerebellum, pons and medulla oblongata. The fossa is bound anteriorly by the superior border of the petrous temporal bone and the dorsum sella, and surrounds the foramen magnum, the cerebellum being housed in the cerebellar fossae on the squamous part of the occipital bone. Features and openings on the floor of the posterior cranial fossa are not as relevant to the eye and orbit as those in the anterior or middle fossae; however, readers should be able to identify the following: foramen magnum, jugular foramen, hypoglossal canal, internal acoustic meatus, grooves for the sigmoid and transverse sinuses, internal occipital protuberance and clivus (Fig. 1-8A,B).

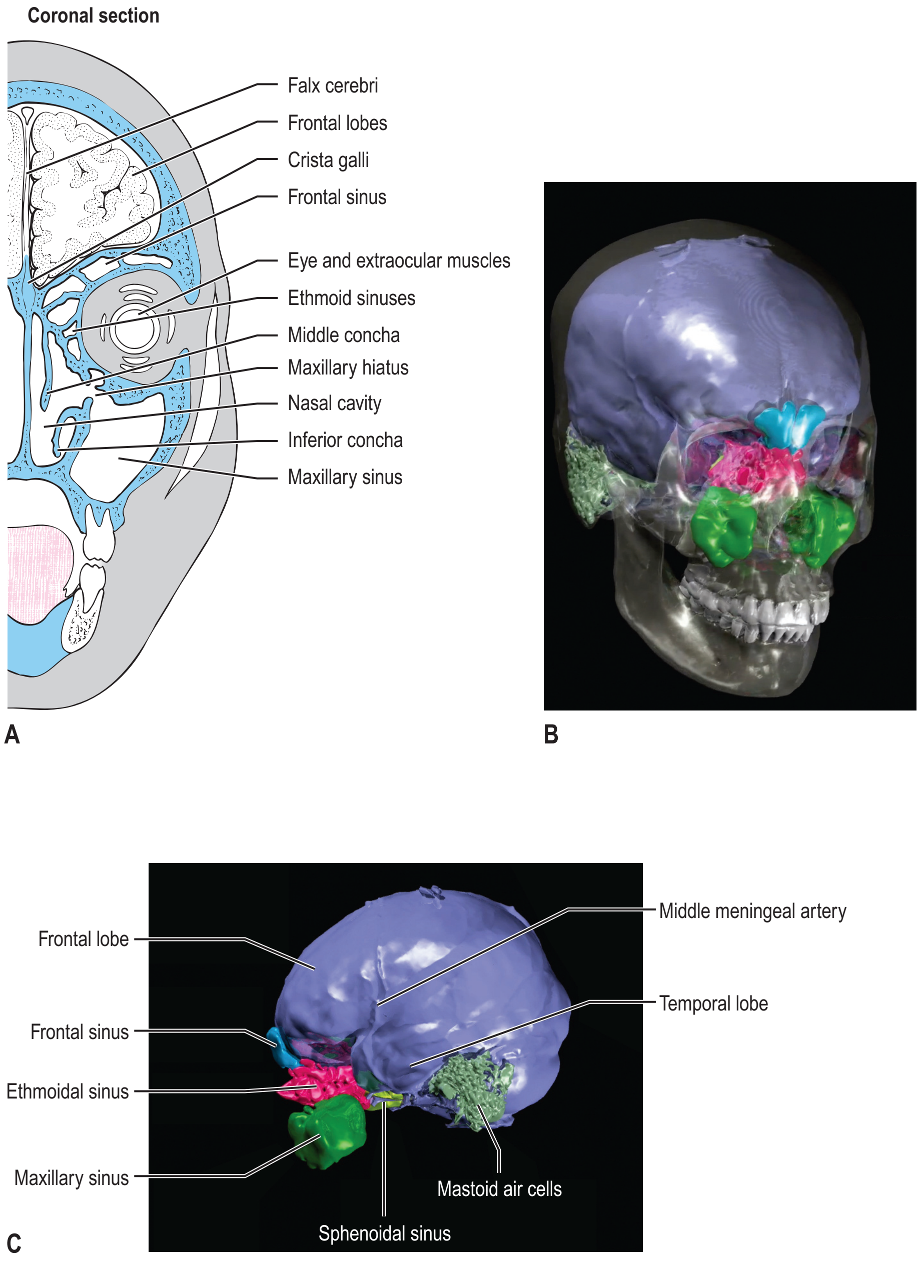
Coronal section
Falx cerebri
Frontal lobes
Crista galli
Frontal sinus
Eye and extraocular muscles
Ethmoid sinuses
Middle concha
Maxillary hiatus
Nasal cavity
Inferior concha
Maxillary sinus
A
B
Frontal lobe
Frontal sinus
Ethmoidal sinus
Maxillary sinus
Middle meningeal artery
Temporal lobe
Mastoid air cells
Sphenoidal sinus
C

FIGURE 1-7 (**A**) Diagram of a coronal section of the head revealing most of the paranasal sinuses except the sphenoid sinus. (**B**) 3D visualization of paranasal sinuses as seen from anterior aspect and (**C**) lateral aspect. See http://www.oucom.ohiou.edu/dbms-witmer/3D_human.htm or https://expertconsult.inkling.com/ for 3D pdf movies of skull and more details of paranasal sinus anatomy.

crest-like ridge, the crista galli, which forms the anterior attachment of the falx cerebri), and the lesser wings and anterior part of the body (jugum) of the sphenoid. The perforations of the cribriform plate transmit the olfactory nerves. The orbital plate of the frontal bone separates the orbit below from the frontal lobes of the cerebral hemispheres, whose sulci and gyri cause surface impressions on the bone. Projecting posteriorly from the lesser wings of the sphenoid are the anterior clinoid processes that overhang the middle cranial fossa and give attachment to the free edge of the tentorium cerebelli.

Middle cranial fossa. The middle cranial fossa lies at a lower plane than the anterior cranial fossa but is higher than the posterior cranial fossa. Its floor is shaped like a butterfly, with a narrow central or median part and expanded lateral parts. It is bound anteriorly by the posterior free edge of the lesser wing of the sphenoid, the anterior clinoid processes, and the anterior margin of the sulcus chiasmaticus (Fig. 1-8A,B). Posteriorly it extends to the superior borders of the petrous temporal bones and dorsum sellae of the sphenoid, and laterally it is bound by the squamous part of the temporal bone, part of the parietal bones, and the greater wings of the sphenoid. Features and foramina of the floor of the middle cranial fossa and the structures that they transmit are summarized in Table 1-1.

Pituitary fossa. The *pituitary fossa* (hypophyseal fossa) is an indentation in the roof of the body of the sphenoid bone in the middle cranial fossa. It is bound anteriorly by the *tuberculum sellae*, in front of which lies the *sulcus chiasmatica*, and posteriorly by the *dorsum sellae*, a ridge of bone at either end of which lies the *posterior clinoid processes*. The pituitary fossa houses the pituitary gland or *hypophysis cerebri*. This is connected by a thin stalk – the pituitary stalk (or tuber cinereum) – to the brain. The fossa is roofed by a sheet of dura mater, the *diaphragma sella* (Fig. 1-9C,D) which is attached in front to the tuberculum and behind to the dorsum sellae. The pituitary stalk passes through a small opening in the roof. The right and left cavernous sinuses are important lateral relations (Fig. 1-9C).

The meninges (Fig. 1-9A,B and eFig. 1-2A)

The brain and spinal cord are surrounded by three layers of meninges: a tough *pachymeninx*, the dura mater, and the *leptomeninges* consisting of the arachnoid mater and pia mater. Between the arachnoid and pia is the *subarachnoid* space filled with cerebrospinal fluid.

Dura mater. The *dura mater* is theoretically 'divided' into an endosteal layer (really the periosteum on the

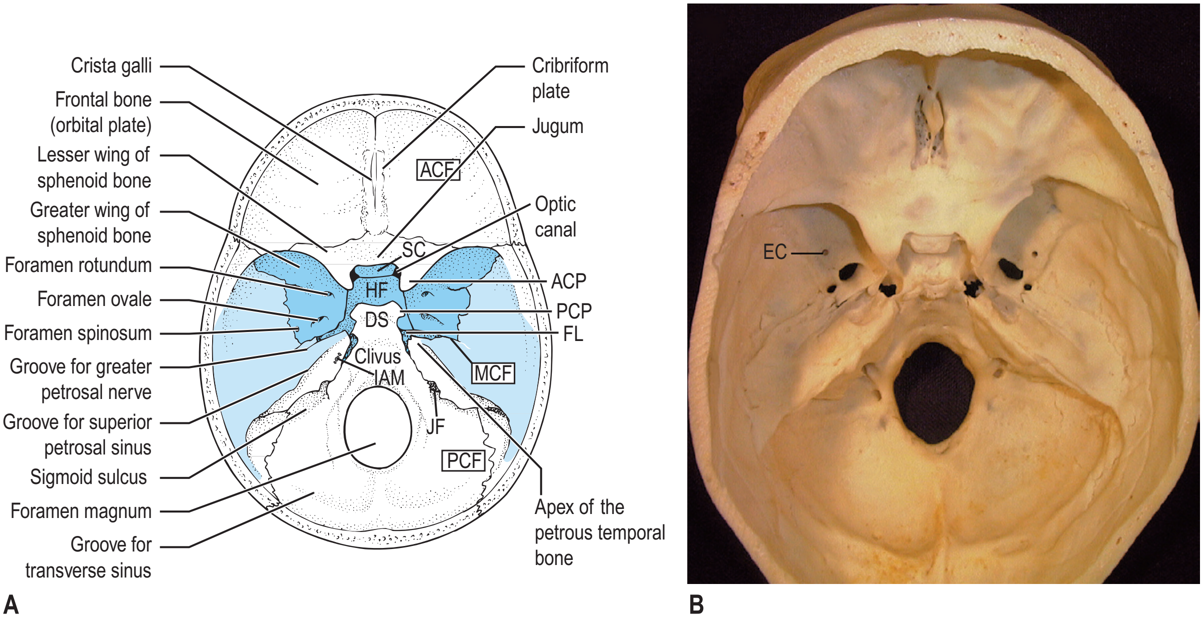
Crista galli
Frontal bone (orbital plate)
Lesser wing of sphenoid bone
Greater wing of sphenoid bone
Foramen rotundum
Foramen ovale
Foramen spinosum
Groove for greater petrosal nerve
Groove for superior petrosal sinus
Sigmoid sulcus
Foramen magnum
Groove for transverse sinus
Cribriform plate
Jugum
Optic canal
ACP
PCP
FL
Apex of the petrous temporal bone
ACF
SC
HF
DS
Clivus
IAM
MCF
JF
PCF
A
EC
B

FIGURE 1-8 Osteology of the cranial cavity. (**A**) The boundaries of the anterior (ACF), middle (MCF) and posterior (PCF) cranial fossae together with major foraminae and important landmarks on the base of the skull. HF, hypophyseal fossa; DS, dorsum sella; SC, sulcus chiasmaticus; IAM, internal auditory meatus; JF, jugular foramen and fossa; ACP, anterior clinoid process; PCP, posterior clinoid process; FL, foramen lacerum. (**B**) Photograph of cranial cavity to illustrate features shown in (**A**): EC, emissary canal.

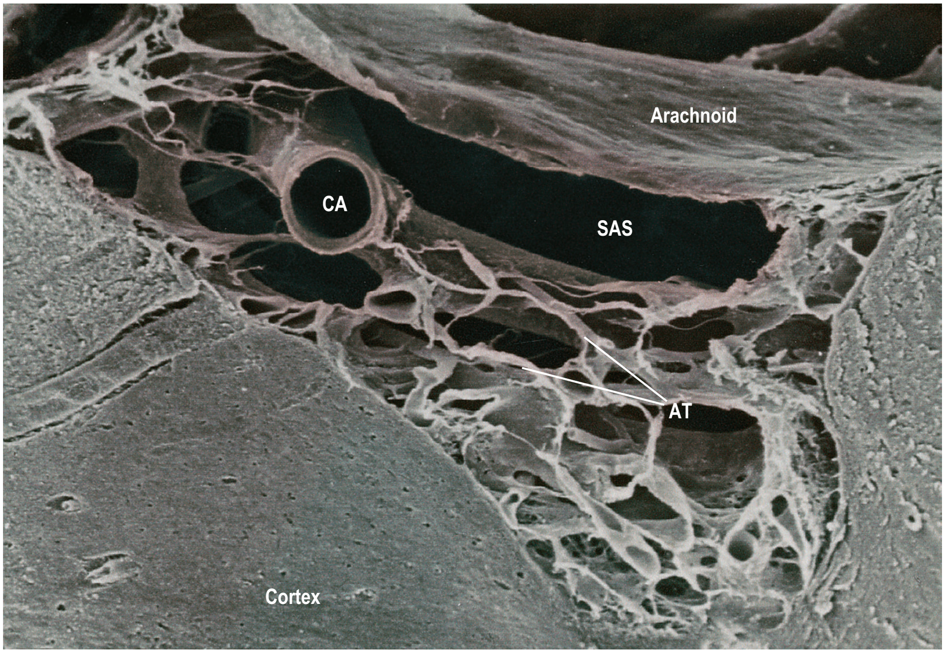
Arachnoid
CA
SAS
AT
Cortex

A

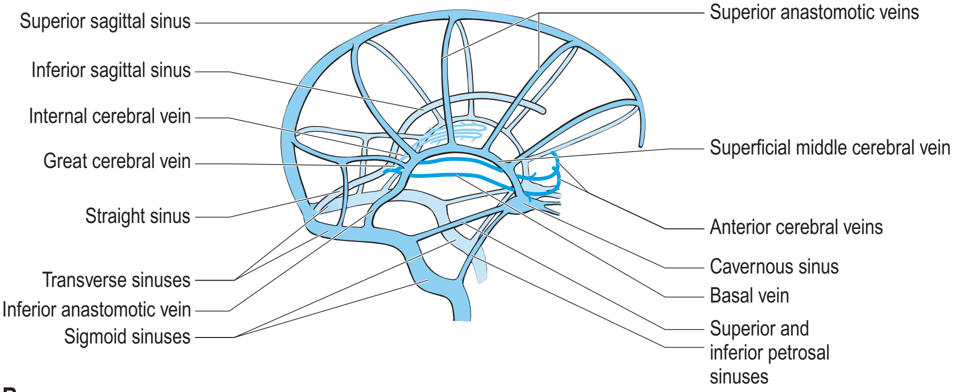
Superior sagittal sinus
Inferior sagittal sinus
Internal cerebral vein
Great cerebral vein
Straight sinus
Transverse sinuses
Inferior anastomotic vein
Sigmoid sinuses
Superior anastomotic veins
Superficial middle cerebral vein
Anterior cerebral veins
Cavernous sinus
Basal vein
Superior and inferior petrosal sinuses

B

eFIGURE 1-2 (**A**) Scanning electron micrograph of the meninges and cortex of the brain showing the arrangement of the arachnoid trabeculae (AT) supporting the cerebral arteries (CA) as they course through the subarachnoid space (SAS) (× 100). (**B**) Schematic diagram of the dural venous sinuses and their connections with cerebral veins.

TABLE 1-1 Summary of features on the floor of the middle cranial fossa

Feature/foramen	Position	Relevance
Sulcus chiasmaticus	Between the two optic canals anterior to tuberculum sella	Only rarely does optic chiasma lie in contact with this region
Sella turcica ('Turkish saddle')	Central part of sphenoid body between the two cavernous sinuses	The central hollow, the hypophyseal fossa, houses the pituitary gland. Anterior and posterior clinoid processes give attachment to the free and attached margins of the tentorium cerebelli
Optic canal	Between the two roots of the lesser wing of the sphenoid	Transmits optic nerve, ophthalmic artery, sympathetic nerves and meningeal coverings
Superior orbital fissure	Between the lesser and greater wings of the sphenoid. Lies at apex of cavernous sinus	Transmits trochlear, abducent and oculomotor nerves and terminal branches of ophthalmic nerve
Foramen rotundum	Pierces greater wing of sphenoid	Transmits maxillary nerve and small veins from cavernous sinus
Foramen ovale	Pierces greater wing of sphenoid	Transmits mandibular nerve, accessory meningeal artery and occasionally the lesser petrosal nerve
Foramen spinosum	Posterolateral to foramen ovale	Transmits middle meningeal artery and vein and meningeal branch of the mandibular nerve
Foramen lacerum	At apex of petrous temporal bone	The upper end transmits the internal carotid artery before it enters the cavernous sinus. Also transmits sympathetic nerves and a small plexus of veins. The lower end is covered by connective tissue and pierced only by small branches of the ascending pharyngeal artery
Trigeminal impression	Anterior surface of petrous temporal bone behind foramen lacerum	Occupied by trigeminal ganglion in trigeminal cave. Joined on lateral aspect by grooves for the greater and lesser petrosal nerves
Tegmen tympani and arcuate eminence	Tegmen is a thin plate of temporal bone over middle ear cavity. Arcuate eminence is produced by superior semicircular canal in petrous temporal bone	Infections in middle ear may spread through thin plate of bone to middle cranial fossa and temporal lobe of the brain

inner surface of the skull) and a meningeal layer; however, on the whole they are fused except where they separate to form *dural venous sinuses* and *dural folds* (Fig. 1-9A,B). The latter are connective tissue septae that extend into the cranial cavity and serve to subdivide it into compartments. In association with the cerebrospinal fluid they aid in providing physical support and protection for the brain. The position and form of the dural folds are summarized diagrammatically in Figure 1-9A.

The *dural venous sinuses* are valveless, highly specialized, firm-walled veins within the cranial cavity which drain venous blood from the brain and cranial bones (Fig. 1-9B). In common with other veins the sinuses are lined by vascular endothelial cells; however, their walls contain no smooth muscle cells. The arrangement of the sinuses is summarized in Fig. 1-9A and eFig. 1-2B.

Of particular note to those studying the eye and orbit is the pair of cavernous sinuses lying either side of the body of the sphenoid (Fig. 1-9C,D).

The importance of the *cavernous sinuses* (Fig. 1-9C,D) lies in their *position, relations* and extensive *communications*. Each cavernous sinus is around 2–3 cm long in the sagittal plane and consists of a series of incompletely fused venous channels or a single venous channel partially subdivided by *trabeculae*. It has walls of dura mater, like other venous sinuses.

Position. There is one cavernous sinus on either side of the body of the sphenoid. The sinus extends from the superior orbital fissure in front to the apex of the petrous temporal bone behind.

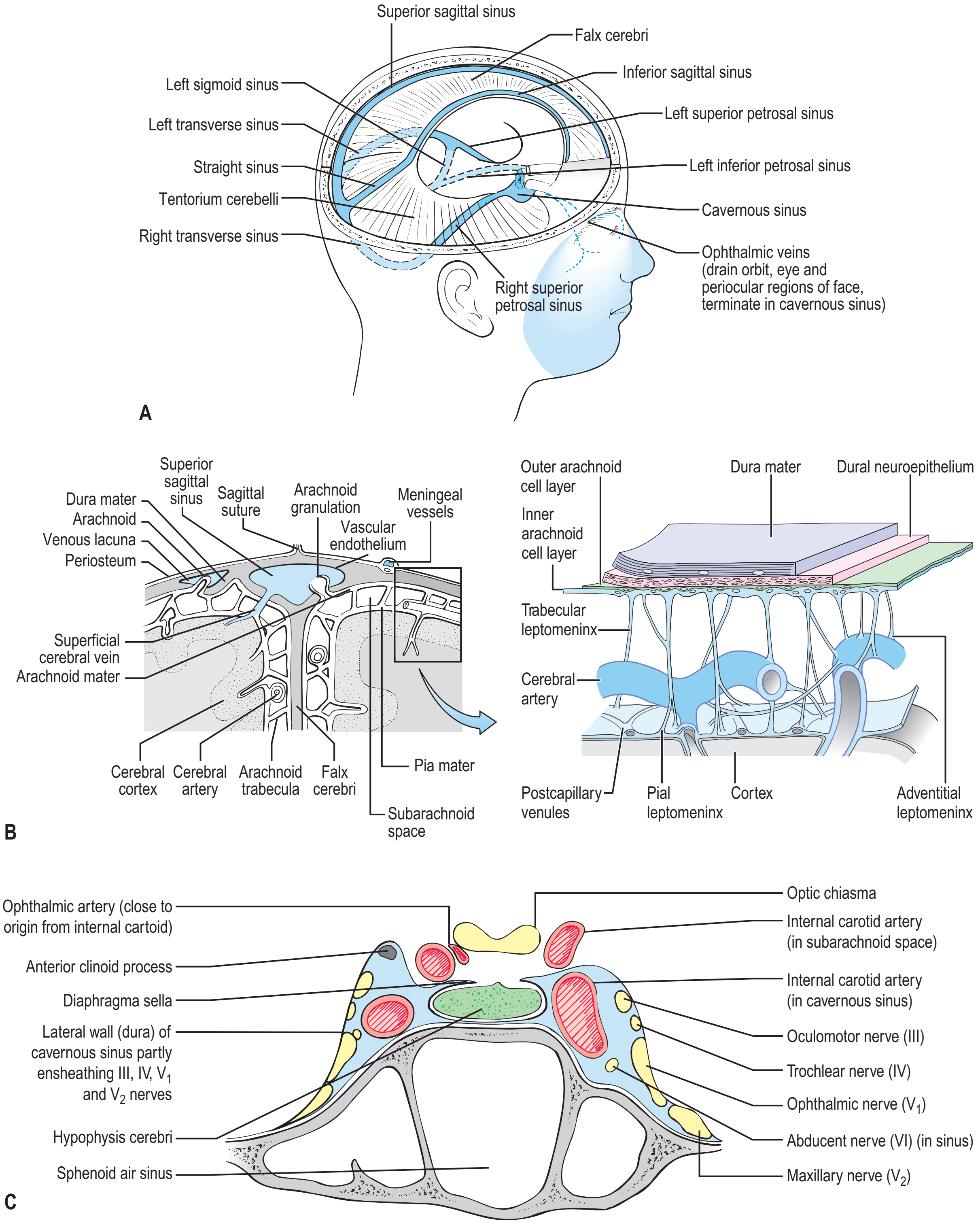
Superior sagittal sinus
Falx cerebri
Inferior sagittal sinus
Left sigmoid sinus
Left transverse sinus
Left superior petrosal sinus
Straight sinus
Left inferior petrosal sinus
Tentorium cerebelli
Cavernous sinus
Right transverse sinus
Ophthalmic veins (drain orbit, eye and periocular regions of face, terminate in cavernous sinus)
Right superior petrosal sinus
A
Superior sagittal sinus
Sagittal suture
Arachnoid granulation
Meningeal vessels
Dura mater
Arachnoid
Venous lacuna
Periosteum
Vascular endothelium
Superficial cerebral vein
Arachnoid mater
Cerebral cortex
Cerebral artery
Arachnoid trabecula
Falx cerebri
Pia mater
Subarachnoid space
B
Outer arachnoid cell layer
Dura mater
Dural neuroepithelium
Inner arachnoid cell layer
Trabecular leptomeninx
Cerebral artery
Postcapillary venules
Pial leptomeninx
Cortex
Adventitial leptomeninx
Ophthalmic artery (close to origin from internal cartoid)
Anterior clinoid process
Diaphragma sella
Lateral wall (dura) of cavernous sinus partly ensheathing III, IV, V_1 and V_2 nerves
Hypophysis cerebri
Sphenoid air sinus
Optic chiasma
Internal carotid artery (in subarachnoid space)
Internal carotid artery (in cavernous sinus)
Oculomotor nerve (III)
Trochlear nerve (IV)
Ophthalmic nerve (V_1)
Abducent nerve (VI) (in sinus)
Maxillary nerve (V_2)
C

FIGURE 1-9 (**A**) The interior of the cranium with the brain removed to illustrate the arrangement of the dural folds and some of the related dural venous sinuses. (**B**) The meninges as seen in coronal section in the region of the superior sagittal sinus. Inset higher-power diagrammatic representation of the meningeal layers. (**C**) Coronal section approximately midway along the body of the sphenoid bone to reveal the paired cavernous sinuses, one on either side. Note the position of the cranial nerves (II, III, IV, V_1, V_2 and VI), internal carotid artery (cut in two places, within and above the sinus) and hypophysis cerebri (pituitary gland).

Continued

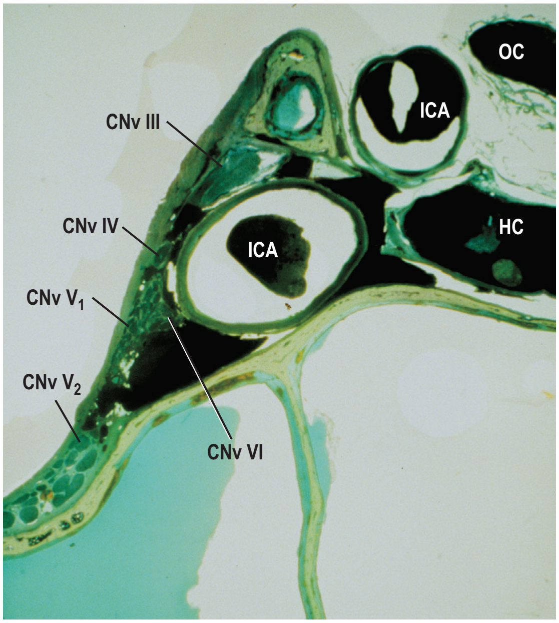

FIGURE 1-9, cont'd (**D**) High-power view of the left cavernous sinus (coronal plane, 100 μm thick section of low-viscosity nitrocellulose resin-embedded specimen) upon which eFig. 1-2B was based. OC, optic chiasma; HC, hypophysis cerebri; ICA, internal carotid artery.

Relations. These are summarized in Figure 1-9C and D (coronal section).

Communications. The sinuses communicate with each other via the anterior and posterior intercavernous sinuses. Tributaries draining into the sinuses anteriorly include the superior and inferior ophthalmic veins (which drain the eye and orbit as well as areas of skin around the periorbital region of the face and nose), and the sphenoparietal sinuses. The superficial cerebral vein from the brain drains into the sinus from above (eFig. 1-2B).

Blood from each sinus may, depending on relative pressures, drain via the superior and inferior petrosal sinuses either directly to the internal jugular veins (inferior petrosal) or to the transverse sinuses and thus to the internal jugular veins. Other exits include venous plexi around the internal carotid artery or veins traversing the foramen ovale or sphenoidal emissary foramen to communicate with the pterygoid plexus and other veins in the region of the skull base. Communications with the vertebral venous plexus in the epidural space also exist via the basilar venous plexus on the clivus.

The major dural folds (Fig. 1-9A) are falx cerebri, tentorium cerebelli, cavum trigeminale and diaphragma sella.

Additional content available at https://expertconsult.inkling.com/.

The meningeal arteries lie within the inner (or periosteal) layer of dura with their accompanying veins (Fig. 1-9B) and are responsible for the many fine grooves that ramify over the inner surface of the cranium (see Fig. 1-8B and eFig. 1-1).

The largest and most important of these is the middle meningeal artery, which enters the skull through the foramen spinosum. These arteries supply the meninges and diploë (bone marrow of cranial bones), but they *do not* supply the brain.

Arachnoid mater (eFig. 1-2A)**.** The arachnoid (*Gk.* spider) is a delicate fibrocellular layer beneath the dura (separated by potential subdural space) that is connected to the pia mater covering the brain by numerous fibrocellular bands that cross the cerebrospinal fluid-filled subarachnoid space. This arrangement has led some to consider the leptomeninges as a conjoined pia–arachnoid membrane. The arachnoid bridges over the sulci, gyri and other irregularities on the brain surface, thus creating the *subarachnoid cisterns* or enlargements in the subarachnoid space (Fig. 1-9B, inset, and eFig. 1-2A).

Specialized regions of arachnoid, the *arachnoid villi* and *granulations* (fibrous aggregations of villi), project into several of the dural venous sinuses (Fig. 1-9B) and act as one-way pressure-sensitive valves allowing cerebrospinal fluid to drain from the subarachnoid space into the dural venous sinuses. Structures passing to and from the brain to the skull or its foramina, such as cranial nerves, must traverse the subarachnoid space. In addition, all cerebral arteries and veins lie in this space (eFig. 1-2A).

Since the arachnoid fuses with the perineurium of cranial nerves, the cerebrospinal fluid-containing subarachnoid space extends for a short distance around all cranial nerves. In particular it surrounds the optic nerve in a cuff-like manner as far as the posterior surface of the eye.

Falx Cerebri

A sickle-shaped fold with its attached border in the midsagittal plane from the crista galli to the tentorium cerebelli behind. It lies in the vertical fissure between the two cerebral hemispheres, its lower border lying above the corpus callosum. The superior sagittal sinus is situated in the attached border and the inferior sagittal sinus is in the lower free border of the falx cerebri.

Tentorium Cerebelli. This fold lies approximately in a horizontal plane at 90° to the falx, although it is elevated centrally (hence 'tent-like'). It separates the occipital lobe of each cerebral hemisphere above from the cerebellum in the posterior cranial fossa below. The free edge forms the boundary of the tentorial notch, which separates the forebrain from the hindbrain and 'houses' the midbrain. Sinuses related to the tentorium include the straight sinus, right and left transverse sinuses, superior petrosal sinuses and cavernous sinuses.

Cavum Trigeminale. A blind-ended dural recess whose entrance is in the posterior cranial fossa. It is formed by an invagination of the dura beneath the free edge of the tentorium and is roofed by dura on the floor of the middle cranial fossa. It houses the trigeminal ganglion, which sits in a shallow hollow on the apex of the petrous temporal bone, and some accompanying vessels. The ganglion is surrounded by cerebrospinal fluid continuous with the subarachnoid space of the posterior cranial fossa.

Diaphragma Sella. A small circular fold of dura over the sella turcica that is pierced centrally by the infundibulum (Fig. 1-9C). It blends laterally with the roof of the cavernous sinus (Fig. 1-9D).

The area above the tentorium is known as the supratentorial compartment; that below is the infratentorial compartment. The cranial dura of the supratentorial compartment is innervated by sensory branches of the trigeminal nerve, and stimulation of these nerves (stretching, inflammation, compression) gives rise to frontal or parietal headache. The infratentorial compartment is supplied by branches of the upper cervical nerves, and stimulation of these sensory nerves may therefore manifest as occipital and neck pain.

The neck rigidity accompanying acute meningitis of the infratentorial region is most likely the result of reflex contractions, or spasm, of posterior neck musculature in response to stretching of the inflamed cranial and spinal cord meninges.

BOX 1-3 CLINICAL CORRELATES

Arteriovenous Fistulas in the Cavernous Sinus

These cause a variety of symptoms including pulsating protrusion of the globe and congestion of the vessels of the lids and conjunctiva owing to raised venous pressure. Patients complain of hearing noises resembling rushing water, probably because of increased flow rates in the labyrinthine plexus, which is in communication with the cavernous sinus via the superior petrosal sinus.

Cavernous Sinus Thrombosis

Cavernous sinus thrombosis as a sequel to infection spreading to the sinus, from such diverse initial sites as the nose, lids, behind the ear, bony labyrinth, pharynx and temporomandibular joint, can give rise to a variety of symptoms explainable on the basis of structures affected in and around the sinus. Facial pain may be the result of the involvement of the ophthalmic nerve (V_1). Lateral rectus paralysis may follow involvement of the abducent nerve. Involvement of the other oculomotor nerves is less common because they are more protected in the lateral wall of the sinus. Thrombosis is usually bilateral because of the communications via the intercavernous sinuses. Papilloedema may result from obstruction of central retinal venous return.

Pia mater (Fig. 1-9B and eFig. 1-2A). The pia mater, a vascular fibrocellular membrane that is thicker than the arachnoid, closely follows the contours of the brain. Vessels entering or leaving the brain parenchyma carry a pial sheath with them. Pial tissue is rich in astrocytes, which extend along the vessel walls and form an important component of the blood–brain barrier. The perivascular spaces surrounding these vessels, the so-called Virchow–Robin spaces, are potentially in communication with the cerebrospinal fluid of the subarachnoid space and may be dilated in pathological states.

Structure of the eye

BOX 1-4 CLINICAL CORRELATES

Extradural and subdural haematomas

Damage to middle meningeal vessels, especially the frontal branch of the middle meningeal artery and vein (the latter lying closest to the bone), may result from blows to the head, especially in the temporal region (the pterion; see Figs 1-2B, 1-5B, 1-8B) where the bones are thinnest and most likely to fracture. Slow venous, or more rapid arterial, bleeding will lead to an extradural (epidural) haematoma with a resultant rise in intracranial pressure. Coma and death will occur if such a haematoma is not drained as soon as possible after symptoms of raised intracranial pressure, including papilloedema, manifest. A subdural haematoma may occur if the trauma results in brain laceration or tearing of intradural veins.

The eye (Fig. 1-10) is a highly specialized organ of photoreception, the process by which light energy from the environment produces changes in specialized nerve cells in the retina, the rods and cones. These changes result in nerve action potentials, which are subsequently relayed to the optic nerve and then to the brain, where the information is processed and consciously appreciated as vision. All the other structures in the eye are secondary to this basic physiological process, although they may be part of the system necessary for focusing and transmitting the light on to the retina, for example cornea, lens, iris and ciliary body, or they may be necessary for nourishing and supporting the tissues of the eye, for example the choroid, aqueous outflow system and lacrimal apparatus.

GENERAL SHAPE, SIZE AND POSITION OF THE EYE

The eye is approximately a sphere 2.5 cm in diameter with a volume of 6.5 mL. However, in reality it is the parts of two spheres, a smaller one anteriorly, the cornea, that has a greater curvature than the sclera, which constitutes the large sphere. The cornea forms one-sixth of the circumference of the globe and has a radius of 7.8 mm; the remaining five-sixths is formed by the sclera, which has a radius of 11.5 mm. There is variation in size between individuals but the average axial length of the globe is 24 mm (range 21–26 mm). The diameter is 23 mm and the horizontal length approximately 23.5 mm. Small eyes (<20 mm) are hyperopic or hypermetropic, while large eyes (26–29 mm) are myopic. The eye is situated in the anterior portion of the orbit, closer to the lateral than the medial wall and nearer the roof than the floor. The eye is made up of three basic layers or coats, often known as tunics (Fig. 1-10). These are the fibrous (corneoscleral) coat, the uvea or uveal tract (composed of choroid, ciliary body and iris), and the neural layer (retina). The coats surround the contents, namely the lens and the transparent media (aqueous humour and vitreous body).

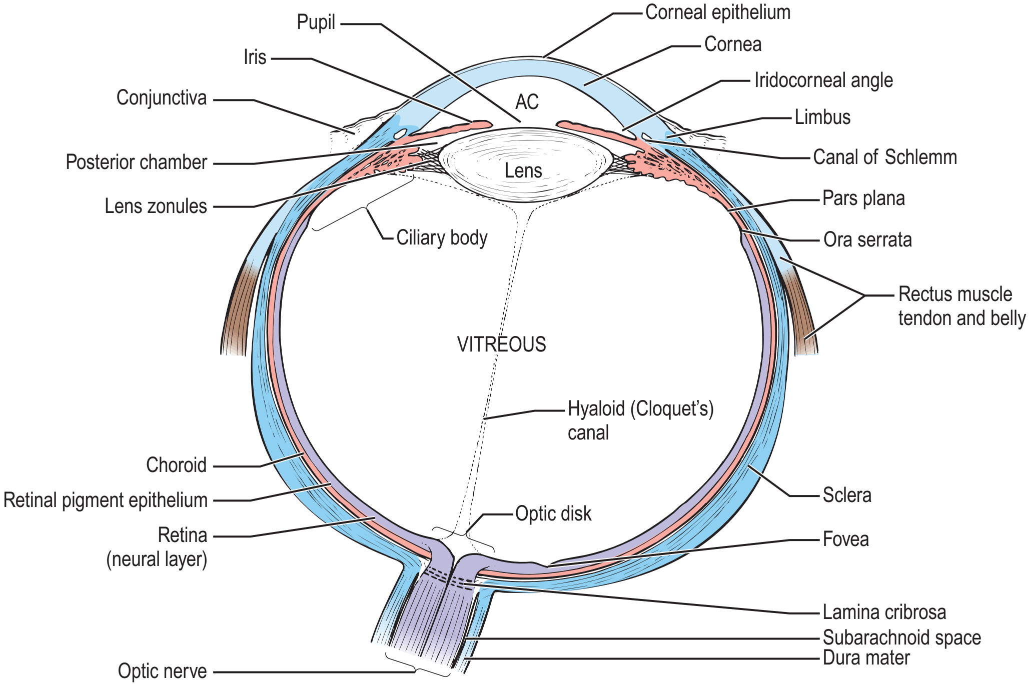
Pupil
Iris
Conjunctiva
Posterior chamber
Lens zonules
Ciliary body
Corneal epithelium
Cornea
Iridocorneal angle
Limbus
Canal of Schlemm
Pars plana
Ora serrata
Rectus muscle tendon and belly
AC
Lens
VITREOUS
Hyaloid (Cloquet's) canal
Choroid
Retinal pigment epithelium
Retina (neural layer)
Optic disk
Sclera
Fovea
Lamina cribrosa
Subarachnoid space
Dura mater
Optic nerve

FIGURE 1-10 Schematic diagram of the human eye in horizontal section revealing the major components and the arrangement of the three layers. AC, anterior chamber. The corneoscleral envelope (blue), the uveal tract (orange/red) and the inner neural layer (purple).

The cornea and sclera together form a tough fibrous envelope that protects the ocular tissues. The fibrous coat also provides important structural support for intraocular contents and for attachment of extraocular muscles. The cornea meets the sclera at a region known as the limbus or corneoscleral junction.

THE CORNEA

The surface of the cornea (air–tissue interface) and associated tear film is responsible for most of the refraction of the eye. The transparency of the cornea is its most important property, although because of its highly exposed position it must also present a tough physical barrier to trauma and infection. Corneal transparency is the result of a number of related factors: its avascularity; the regularity and smoothness of the covering epithelium; and the regular arrangement of the extracellular and cellular components in the stroma, which is dependent on the state of hydration, metabolism and nutrition of the stromal elements.

Shape

The cornea is smaller in the vertical (10.6 mm) than in the horizontal (11.7 mm) diameter; however, viewed from behind, the circumference appears circular. The central radius is 7.8 mm with the peripheral corneal curvature being less marked. The cornea is also thicker at the periphery (0.67 mm) than in the centre (0.52 mm).

Structure

The cornea is composed of five layers (Fig. 1-12A).

Corneal epithelium (Fig. 1-12B). The corneal epithelium is a *stratified* (possessing five or six layers) *squamous non-keratinized* epithelium (the superficial cells are flattened, nucleated and non-keratinized). It is 50–60 μm in thickness and adjacent cells are held together by numerous *desmosomes* and to the underlying basal lamina by *hemidesmosomes* and anchoring filaments (Fig. 1-12B). The anterior surface of the corneal epithelium is characterized by numerous microvilli and microplicae (ridges) whose glycocalyx coat interacts with, and helps stabilize, the precorneal tear film. New cells are derived from mitotic activity in the limbal basal cell layer (see p. 211) and these displace existing cells both superficially and centripetally. The corneal epithelium responds rapidly to

BOX 1-5 CLINICAL CORRELATES

ASTIGMATISM

Astigmatism is usually the result of differences in the radius of curvature in the vertical and horizontal meridians. Abnormalities in corneal curvature can be readily demonstrated by computerized video keratography, which applies the principle of projecting placido rings onto the corneal surface from which topographic maps can be constructed (Fig. 1-11).

repair disruptions in its integrity by amoeboid sliding movements of cells on the wound margin followed by cell replication.

The basal epithelial cells rest on a thin, but prominent, *basal lamina* (lamina lucida, 25 nm; lamina densa, 50 nm). Corneal epithelial adhesion is maintained by a basement membrane complex, which anchors the epithelium to *Bowman's layer* via a complex mesh of anchoring fibrils (type VII collagen) and anchoring plaques (type VI collagen), which interact

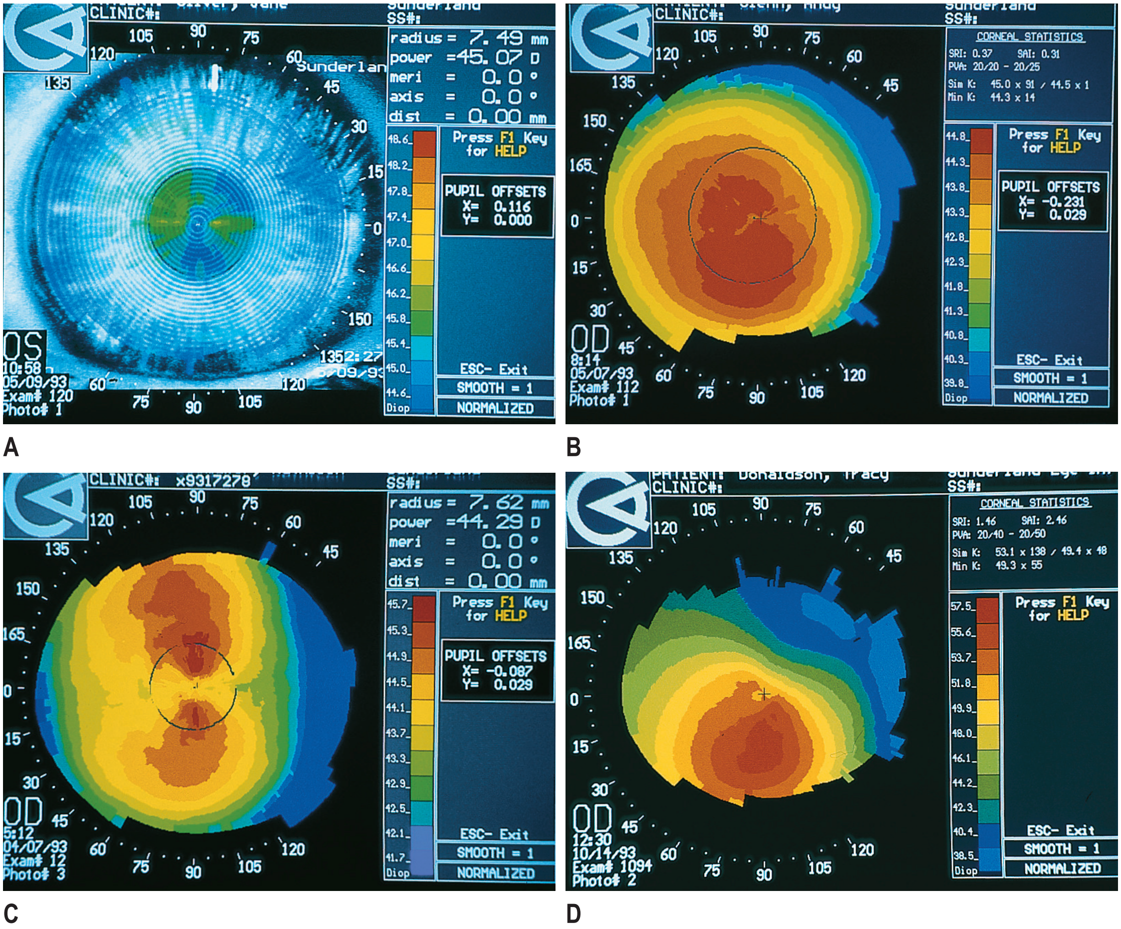
A
B
C
D
CLINIC#:
SS#:
radius = 7.49 mm
power =45.07 D
meri = 0.0 °
axis = 0.0 °
dist = 0.00 mm
Press F1 Key for HELP
PUPIL OFFSETS X= 0.116 Y= 0.000
ESC- Exit
SMOOTH = 1
NORMALIZED
Diop
OS
05/09/93
Exam# 120
Photo# 1
CORNEAL STATISTICS
SRI: 0.37 SAI: 0.31
PVA: 20/20 - 20/25
Sim K: 45.0 x 91 / 44.5 x 1
Min K: 44.3 x 14
PUPIL OFFSETS X= -0.231 Y= 0.029
OD
05/07/93
Exam# 112
Photo# 1
radius = 7.62 mm
power =44.29 D
PUPIL OFFSETS X= -0.087 Y= 0.029
OD
04/07/93
Exam# 12
Photo# 3
PATIENT: Donaldson, Tracy
SRI: 1.46 SAI: 2.46
PVA: 20/40 - 20/50
Sim K: 53.1 x 138 / 49.4 x 48
Min K: 49.3 x 55
OD
10/14/93
Exam# 1094
Photo# 2

FIGURE 1-11 Computerized video keratography (CVK). (**A**) This method involves the projection of over 6000 points of light on to the corneal surface in the form of placido rings. The images are analysed by the computer and complex colour-coded topographical/dioptric maps can be constructed. The scale or key is shown alongside: 'hotter' colours represent higher dioptric values. (**B**) A normal or round topographic map. (**C**) Regular 'with the rule' astigmatism in a normal healthy cornea with +1.5 diopters of astigmatism at 90°. (**D**) Corneal topographic map of a patient with early keratoconus. CVK analysis is particularly useful in identifying early keratoconus. In this case the higher dioptric values are concentrated in the infratemporal region. (Photographs kindly provided by Prof. C. McGhee.)

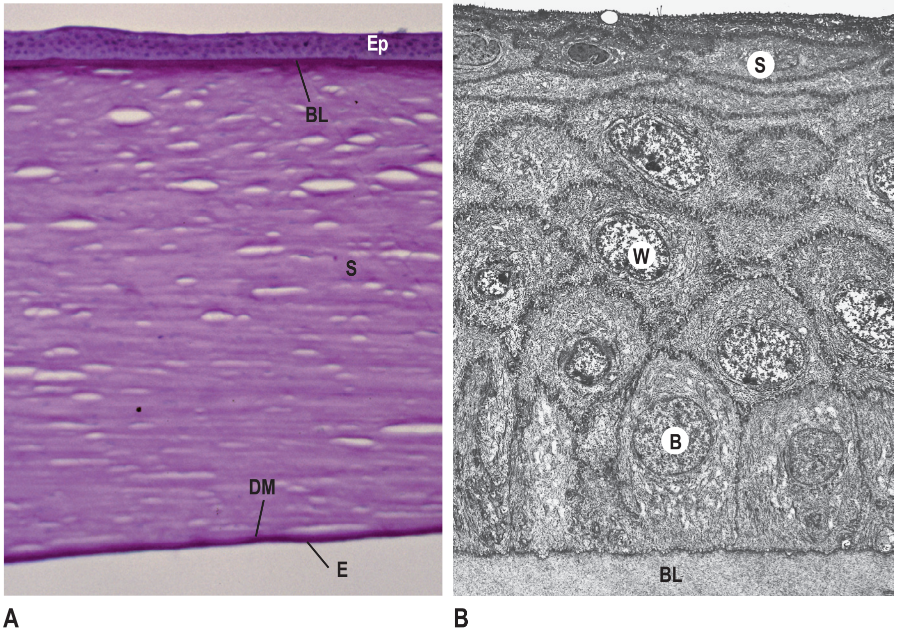

FIGURE 1-12 Histology and ultrastructure of the cornea and its constituent layers. (**A**) Low-power light micrograph showing the five layers of the human cornea. Ep, epithelium; BL, Bowman's layer; S, substantia propria or stroma; DM, Desçemet's membrane; E, endothelium. (**B**) Electron micrograph of the corneal epithelium. B, basal cell layer; W, wing cells; S, superficial cells; BL, Bowman's layer. Original magnifications: **A**, × 80; **B**, × 3000. (Part B courtesy of W.R. Lee and D. Aitken.)

with the lamina densa and the collagen fibrils of Bowman's layer. The corneal epithelium is devoid of melanocytes. Myeloid-derived major histocompatibility complex (MHC) class II antigen-positive dendritic cells (Langerhans cells) are present in the limbus and peripheral cornea (Fig. 1-12), but decline sharply in density in a centripetal gradient, and are rare in the central cornea. However, MHC class II-negative dendritic cells have been identified in the mouse central cornea and recent *in vivo* confocal microscopy (IVCM) (Fig. 1-13) suggests that the normal human central corneal epithelium contains dendritic cells although their immunophenotype cannot be ascertained from IVCM. The comparative paucity of potential antigen-presenting cells, such as dendritic cells, and the avascular nature of the cornea are considered factors crucial to the success of corneal grafting (see Ch. 7).

Anterior limiting lamina (Bowman's layer). Bowman's layer (a modified acellular region of the stroma; 8–12 μm thick) consists of fine, randomly arranged, collagen fibrils (20–30 nm diameter, types I, III, V and VI). The anterior surface is well delineated and is separated from the epithelium by the thin basal lamina, while the posterior boundary merges with the stroma (Fig. 1-12A). Bowman's layer terminates abruptly at the limbus.

Substantia propria or corneal stroma (Fig. 1-14 A–C). The corneal stroma is a dense connective tissue of remarkable regularity. It makes up the vast majority of the cornea and consists predominantly of 2 μm thick, flattened, *collagenous lamellae* (200–250 layers) oriented parallel to the corneal surface and continuous with the sclera at the limbus. Between the lamellae lie extremely flattened, modified fibroblasts known as *keratocytes*. These cells are stellate in shape with thin cytoplasmic extensions containing conspicuously few distinctive organelles (Fig. 1-14A–C) when viewed in conventional cross-sections. However, frontal sections reveal an abundance of organelles and a novel network of fenestrations on their surface which may facilitate the diffusion of metabolites or the mechanical 'anchoring' or attachment of collagen bundles (Fig. 1-14A). The density of keratoctyes in the anterior stroma is 20 000–24 000 cells/mm^2 and that density decreases

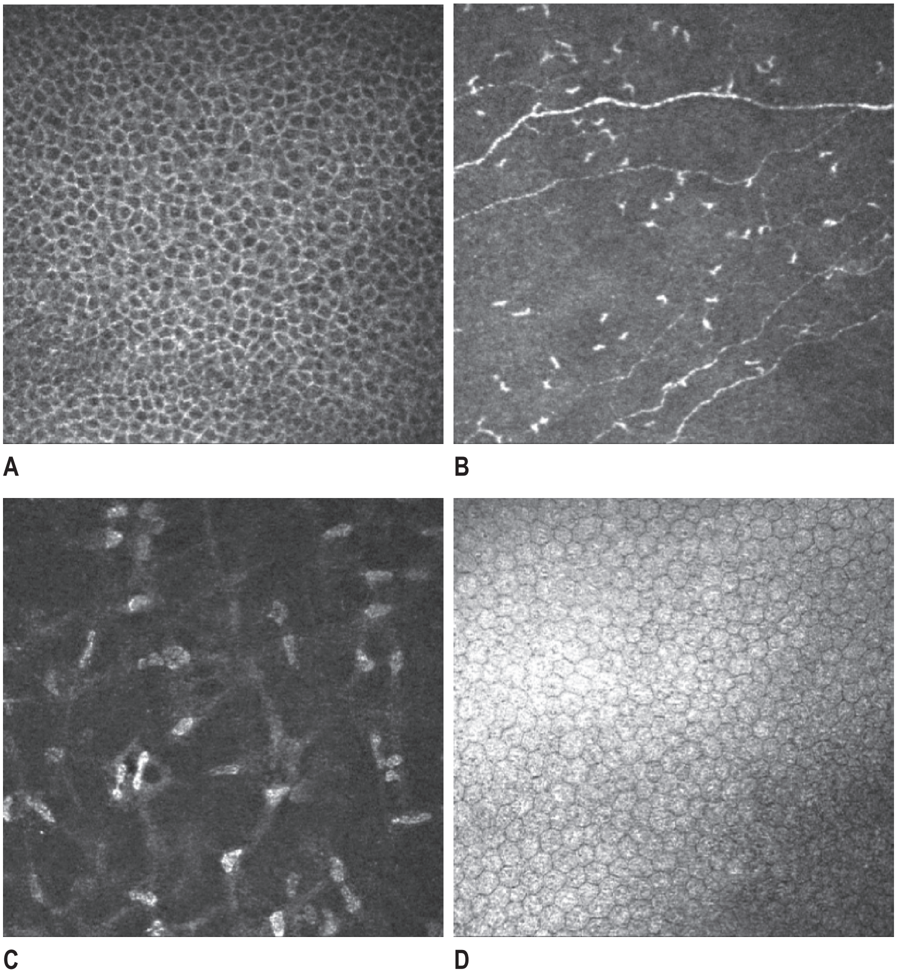

FIGURE 1-13 *In vivo* confocal microscopy (IVCM) of the human cornea is a powerful non-invasive instrument used in the clinical evaluation of corneal abnormalities and normal structure of the tear film, cornea and conjunctiva. IVCM images are obtained by performing ‘optical sections’ of the cornea using non-coherent white light. Cells and matrix components with differing reflective properties within the transparent cornea can be imaged. The advantage of a ‘confocal’ approach is that only information in a narrow focal plane, approximately 4–25 μm in thickness, is analysed or collected by the microscope and scattering of light from structures outside the focal plane is thus minimized. The optics allow the light beam to be scanned (in the *x* and *y* axes) in a narrow area at one focal plane before shifting in depth to another plane of ‘focus’ (*z* axis) where the scan is repeated. Thus a series of optical ‘slices’ of high lateral resolution (1–2 μm) can be obtained from the entire cornea and, because of small differences in brightness/contrast, cellular detail can be visualized. This provides information that is normally the realm of conventional light microscopic and *ex vivo* laser scanning fluorescence confocal microscopic studies of processed tissues and whole mounts. The images are ‘slices’ at differing depths in the cornea from superficial to deep: (**A**) epithelium; (**B**) sub-basal nerve plexus and dendriform cells, which may represent Langerhans cells; (**C**) keratocytes in the posterior stroma; (**D**) corneal endothelial cells. (Images courtesy of Prof. C. McGhee.)

posteriorly before increasing again near Desçemet’s membrane (Fig. 1-13C). Keratocytes are connected by gap junctions to their neighbouring cells and arranged in a corkscrew pattern spiralling from the epithelium to the endothelium. The collagenous lamellae form a highly organized orthogonal ply, adjacent lamellae being oriented at right angles, with the exception of the anterior third in which the lamellae display a more oblique orientation. The collagen fibres (Fig. 1-14C, inset) are predominantly of type I (30 nm diameter, 64–70 nm banding) with some type III, V and VI also present. The transparency of the cornea is highly

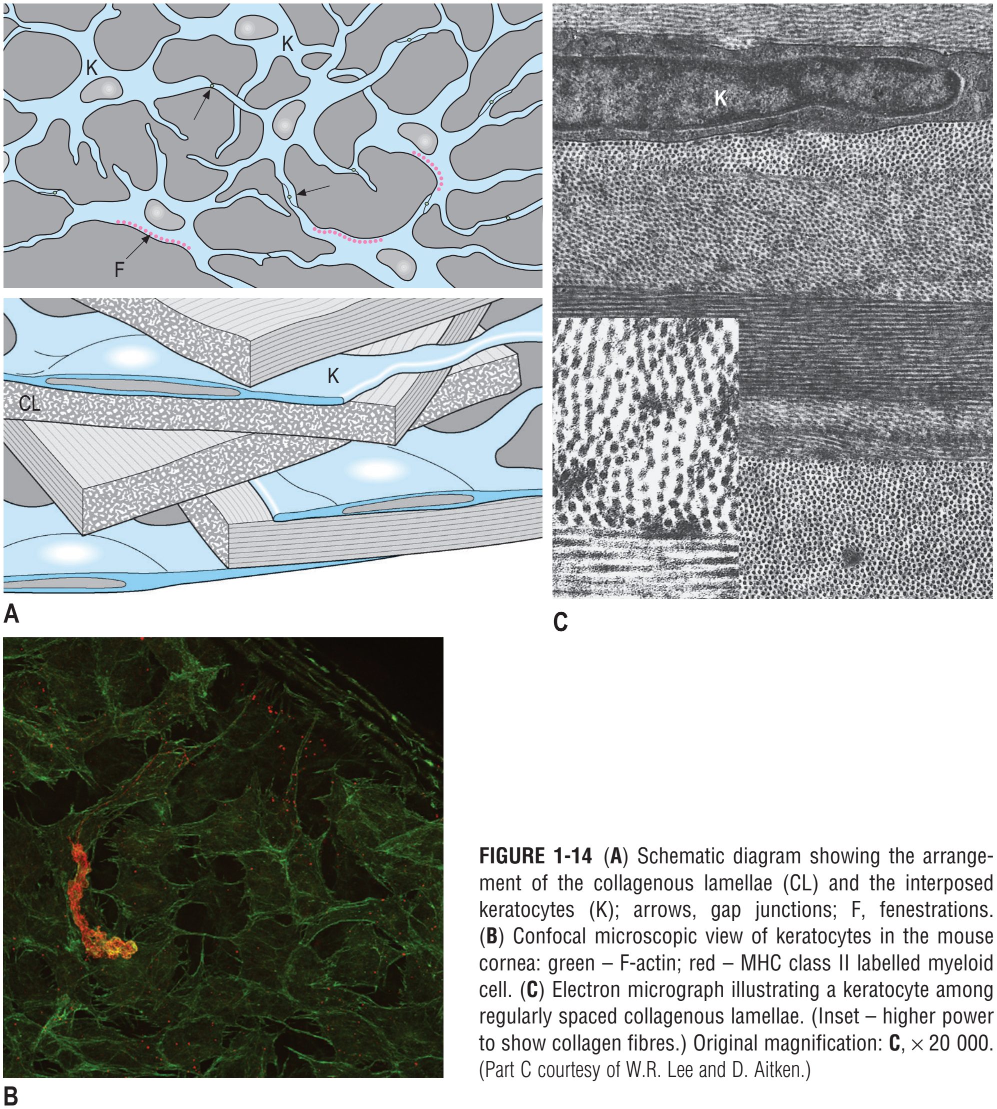

FIGURE 1-14 (**A**) Schematic diagram showing the arrangement of the collagenous lamellae (CL) and the interposed keratocytes (K); arrows, gap junctions; F, fenestrations. (**B**) Confocal microscopic view of keratocytes in the mouse cornea: green – F-actin; red – MHC class II labelled myeloid cell. (**C**) Electron micrograph illustrating a keratocyte among regularly spaced collagenous lamellae. (Inset – higher power to show collagen fibres.) Original magnification: **C**, × 20 000. (Part C courtesy of W.R. Lee and D. Aitken.)

dependent on the regular diameter (influenced by the presence of type V collagen in particular) and spacing of the collagen fibres (interfibrillary distance), which in turn is regulated by glycosaminoglycans (GAG) and proteoglycans forming bridges between the collagen fibrils. The GAGs in the human cornea are predominantly keratan sulphate and chondroitin (dermatan) sulphates (see Ch. 4). The corneal stroma normally contains no blood or lymphatic vessels, but sensory nerve fibres are present in the anterior layers en route to the epithelium (see below and Fig 1.13). Studies using transgenic mice in which eGFP (enhanced green fluorescent protein) is expressed on all CX_3CR1 positive monocyte derived cells has revealed extensive populations of resident tissue macrophages throughout the corneal stroma, some of which have recently described membrane nanotube cell–cell communications (Fig. 1-15).

Posterior limiting lamina (Desçemet's membrane) (Figs 1-12A and 1-16A). This is a thin, homogeneous, discrete, periodic acid–Schiff-positive layer between the posterior stroma and the endothelium, from which it can become detached. It is 8–12 μm in thickness and represents the modified basement membrane of the corneal endothelium. It consists of two parts, an anterior third that is banded and a homogeneous or non-banded posterior two-thirds. It is rich in basement membrane glycoproteins, laminin and type IV collagen. The anterior banded region is reported to contain type VIII collagen. Types V and VI collagen may be involved in maintaining adherence at the interface of Desçemet's membrane with the most posterior lamellae of the stroma. Desçemet's membrane is continuous peripherally with the cortical zone of the trabeculae in the trabecular meshwork. Microscopic wart-like protuberances (Hassall–Henle bodies) containing 'long banded' (100 nm) deposits of unknown nature appear in the periphery of Desçemet's membrane with age. It is frequently thickened at its peripheral termination (Schwalbe's line, the anterior limit of the trabecular meshwork). If disrupted, Desçemet's membrane tends to curl inwards towards the anterior chamber.

Corneal endothelium. The corneal endothelium, a *simple squamous epithelium* on the posterior surface of the cornea, has a critical role in maintaining corneal hydration and thus transparency.

Fluid is constantly being lost via evaporation at the ocular surface, a fact illustrated by increased corneal thickness after a night of lid closure and when an impermeable lens is placed over the epithelium. The endothelial cells rest on Desçemet's membrane (Fig. 1-16A) and form an uninterrupted polygonal or hexagonal array, or mosaic (Fig. 1-16B), which can be clearly seen *in vivo* with the aid of *specular microscopy* and *in vivo confocal microscopy* (Fig. 1-13). The cells are 5–6 μm in height and 18–20 μm in diameter. Their lateral surfaces are highly interdigitated and possess apical junctional complexes that, together with abundant cytoplasmic organelles including mitochondria (Fig. 1-16A), are indicative of their crucial role in active fluid transport.

In the normal human cornea endothelial cells are generally considered to have low regenerative capacity and lost cells are quickly replaced by spreading of adjacent cells. However, there is some evidence that putative endothelial stem cells are located in centripetally arranged 'niches' or columns in the extreme peripheral cornea. Damage to corneal endothelial cells and density below 800 cells/mm^2 leads rapidly to oedema and swelling of the stroma, with resultant loss of transparency (see Ch. 4, p. 213).

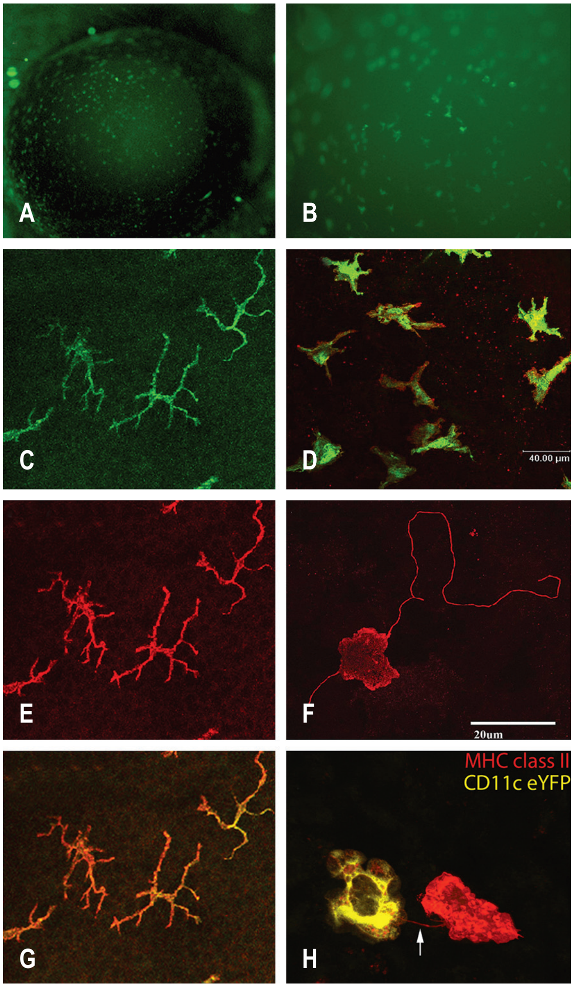

FIGURE 1-15 (**A**, **B**) Low- and high-power *in vivo* fluorescent microscopy of the cornea of a normal CX3CR1-GFP transgenic mouse in which all myeloid cells (macrophages and dendritic cells) are labelled with green fluorescent protein (GFP). Note the regular array of corneal stromal macrophages. (**C-H**), Confocal microscopy of immune cells in the normal cornea. C, E, G is the same field of corneal epithelium from a corneal flatmount from a CX3CR1-GFP mouse: C – GFP (green); E – MHC class II (red); and G – combined image. D, CD11b$^+$ (red) stromal macrophages in the same mouse cornea. F – MHC class II$^+$ putative dendritic cell [DC] (see Ch. 7 for details of DC function) with a membrane nanotube extending from the cytoplasm. H, Putative DC joined to a MHC class II$^+$ macrophage.

BOX 1-6 AGEING CHANGES

There are approximately 350 000 endothelial cells per cornea (3000–4000 cells/mm^2 at birth, falling to 2500 cells/mm^2 in middle age and 2000 cells/mm^2 in old age). Consequently, with age, the dense, regular hexagonal arrangement typical of the young cornea is replaced by fewer cells of more heterogeneous size and shape.

A density lower than 1500 cells/mm^2 in a potential donor cornea is considered unsuitable for transplantation.

Nerve supply of the cornea

The cornea is richly supplied by sensory fibres derived from the ophthalmic division of the *trigeminal nerve*, mainly via the *long ciliary* nerves. Occasionally the inferior cornea receives some branches from the maxillary division of the trigeminal. Nerve bundles enter the peripheral cornea in a radial manner and as they travel centrally below the anterior one-third of the stroma and approximately 1 mm from the limbus they lose their perineurium and myelin sheaths. This alteration in myelination is thought to be related to the importance for transparency. They divide into smaller branches and begin to change direction towards the epithelium where they must pierce Bowman's layer, whereupon they further divide into smaller bundles to form the *subepithelial* or *subbasal plexus* in the interface between Bowman's layer and the basal aspect of the corneal epithelium. There are apparently no specialized end organs associated with these terminal axons, which are predominantly within the size range 0.1–0.5 μm consistent with A-delta and C fibres that function to transmit the sensory modalities of pain and temperature. Individual beaded fibres penetrate the epithelial layers and terminate in the superficial layers in the form of an intraepithelial plexus. There are approximately 7000 nociceptors per mm^2 in the human corneal epithelium.

THE SCLERA

The sclera (Fig. 1-17A-D) forms the principal part of the outer fibrous coat of the eye and functions both to protect the intraocular contents and to maintain the shape of the globe when distended by intrinsic intraocular pressure. The globe shape is maintained

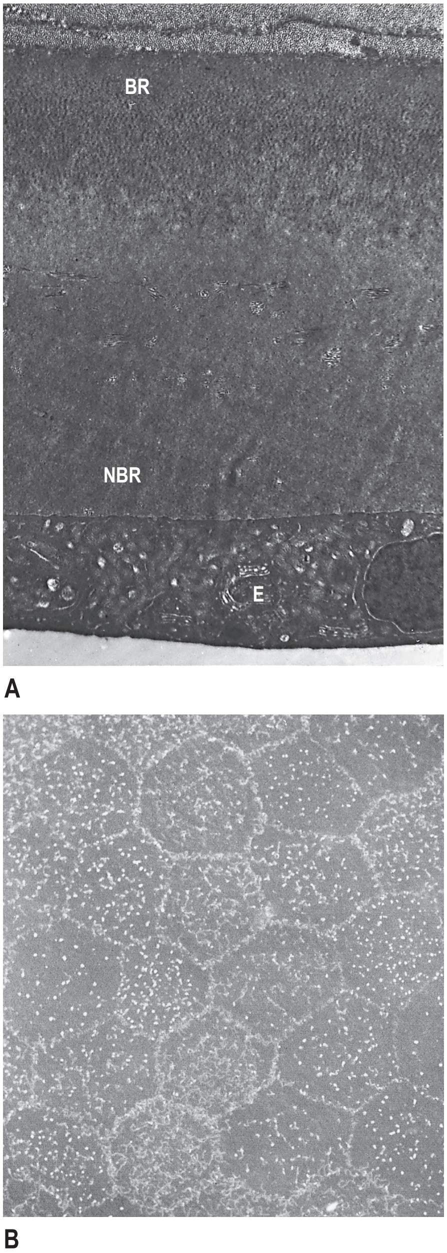

FIGURE 1-16 (**A**) Electron micrograph of Desçemet's membrane and corneal endothelium (E) to illustrate the banded region (BR) and non-banded region (NBR). (**B**) An *en face* view of the inner surface of the corneal endothelium as seen by scanning electron microscopy. Note the homogeneous hexagonal array of endothelial cells. A similar but less detailed view of the endothelium can be achieved in the living patient with the aid of specular microscopy. Original magnifications: **A**, × 7000; **B**, × 1200. (Courtesy of W.R. Lee and D. Aitken; Part A courtesy of Springer-Verlag.)

BOX 1-7 CLINICAL CORRELATES

Larger myelinated nerve fibres can often be seen during slit-lamp examination as fine whitish fibres radiating into the cornea from the limbus. *In vivo confocal microscopy* has greatly added to our understanding of the distribution of corneal nerves in healthy and diseased corneas (see Fig. 1-13). Due to their position in the anterior stroma and their radial arrangement, many are damaged during refractive procedures such as *LASIK* (Excimer laser *in situ* keratomileusis). Damage to the corneal epithelium and intraepithelial nerve terminals can cause a great deal of pain.

Reactivation of latent herpes simplex virus in the trigeminal ganglion occurs following damage to nerve terminals (cold, exposure to ultraviolet light, trauma, corticosteroids), and activated virus is transmitted to the cornea along sensory nerve branches, leading to recurrent herpes simplex keratitis and superficial corneal ulceration.

BOX 1-8 CLINICAL CORRELATES

Buphthalmos

The corneoscleral envelope of children with congenital glaucoma responds to raised intraocular pressure by irreversibly stretching, owing to the immaturity of the collagen fibres, thus producing the characteristically enlarged buphthalmos ('ox-eye') of this condition.

BOX 1-9 CLINICAL CORRELATES

The yellowing of the eyeball in jaundice is the result of bilirubin deposition in the conjunctiva and not the sclera. Abnormal thinning of the sclera, such as occurs in some connective tissue disorders, e.g. Ehlers–Danlos syndrome, may also lead to a blue tinge. Localized thinning of the stromal collagenous layers may lead to staphyloma (bulging).

even during contraction of the extraocular muscles, whose tendons insert on its surface. The sclera is relatively avascular and in adults appears white externally. The viscoelastic nature of the sclera (great tensile strength, extensibility and flexibility) allows only limited distension and contraction to accommodate minor variations in intraocular pressure.

The sclera is thickest posteriorly (1 mm) and thinnest (0.3–0.4 mm) behind the insertions of the aponeurotic tendons of the extraocular muscles. It is covered by the *fascia bulbi* posteriorly and the conjunctiva anteriorly. The sclera consists of dense irregular connective tissue comprising extracellular matrix and matrix-secreting fibroblasts. The matrix consists principally of collagen type I, although types III, IV, V, VI, VIII, XII and XIII have been identified (Rada et al., 2006). Unlike the cornea, the scleral collagenous lamellae are irregularly arranged (Fig. 1-17D) and are interspersed with elastic fibres, each consisting of an elastin core surrounded by longitudinally arranged microfibrils composed of a number of glycoproteins including fibrillin. The opaque nature of the sclera, in contrast to the transparency of the cornea, can be partly ascribed to this irregular arrangement of the collagen fibres (Fig. 1-17D), but also to the variable fibre diameter (25–250 nm), variable and irregular fibrillar spacing, higher water content, and the reduced coating of GAGs on collagen fibres. Indeed, the sclera contains one-quarter of the proteoglycan and GAG content of the cornea. Dermatan sulphate and chondroitin sulphate proteoglycans are the most abundant in the sclera.

Collagen fibrils take up tensile force and are aligned with the direction of greatest tensile strength. The arrangements of scleral collagen can be studied using the 'split-line' technique, which has revealed that the collagen fibrils in the outer sclera are arranged in bundles that course in whorls, loops and arches, particularly around the muscle insertions and optic nerve (Fig. 1-17C). The collagen fibrils on the internal aspect of the sclera are arranged in a rhombic pattern (Fig. 1-17D).

The sclera extends anteriorly from the *limbus* to the *lamina cribrosa* posteriorly (Fig. 1-17A and Fig. 1-42). The scleral collagen fibrils are arranged in circles or figure-of-eight patterns at the lamina cribrosa. Structures that transverse the sclera are shown in Figure 1-17A. Histologically the sclera has three layers: the lamina fusca, stroma and episclera (Fig. 1-17B).

LIMBUS AND AQUEOUS OUTFLOW PATHWAYS (Fig. 1-18)

It is becoming increasingly appreciated that the *limbus* (Fig. 1-18A–C) is more than the border zone between the cornea and sclera; it has multiple functions including nourishment of the peripheral cornea, corneal wound healing, immunosurveillance of the ocular surface and hypersensitivity responses; it contains the pathways of aqueous humour outflow and is thus involved in the control of intraocular pressure.

Surgical incisions to access the anterior chamber for cataract and glaucoma surgery are made at the limbus. The limbus is 1.5–2.0 mm in width and the change in the radius of curvature between the sclera and cornea produces a shallow *external scleral sulcus* and an *internal scleral sulcus*; the latter is deepened by the *scleral spur* and houses the canal of Schlemm and trabecular meshwork. The longitudinal ciliary muscle fibres attach to the posterior aspect of the *scleral spur*, and its anterior surface gives rise to the corneoscleral trabeculae.

Several important transitions take place at the limbus (Fig. 1-18A–C).

- The regularly arranged corneal lamellae give way to the more random array of lamellae in the sclera. The corneal termination is V-shaped (Fig. 1-18B,C).
- The stratified squamous non-keratinized corneal epithelium with its parallel internal and external surfaces gives way to conjunctival epithelium, characterized by a folded basal surface and interdigitating subepithelial connective tissue (sometimes forming distinct papillae) (Fig. 1-18B).
- The conjunctival epithelium contains goblet cells and a rich network of MHC class II$^+$ CD11c$^+$ dendritic (Langerhans) cells (see section on conjunctiva, p. 83).
- Loops or arcades of conjunctival capillaries (derived from the anterior ciliary arteries) and lymphatic capillaries terminate at the limbus. The smaller vessels are not under neuronal control and are particularly susceptible to the

BOX 1-10 AGEING CHANGES

The ‘blue’ sclera of infants is the result of the underlying choroidal pigment showing through the thin collagenous stroma. In elderly individuals, fat deposition in the sclera may produce a yellowish hue.

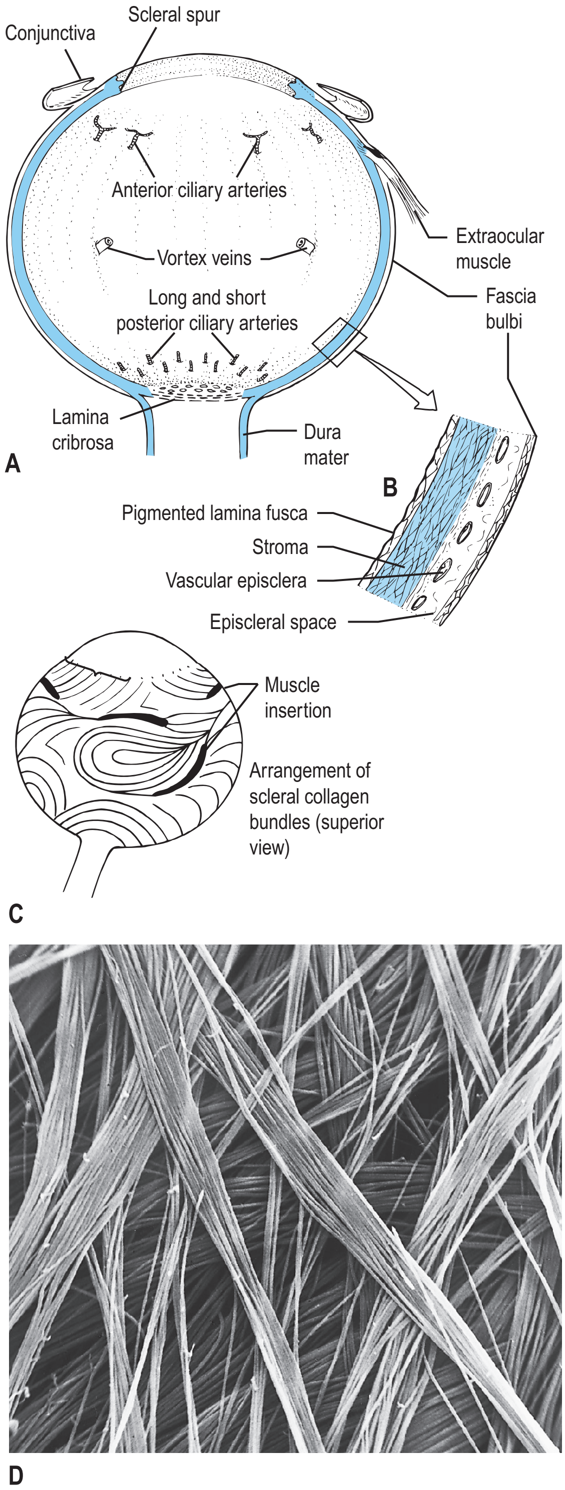
Conjunctiva
Scleral spur
Anterior ciliary arteries
Extraocular muscle
Vortex veins
Fascia bulbi
Long and short posterior ciliary arteries
Lamina cribrosa
Dura mater
A
B
Pigmented lamina fusca
Stroma
Vascular episclera
Episcleral space
Muscle insertion
Arrangement of scleral collagen bundles (superior view)
C
D

FIGURE 1-17 (**A**) Schematic diagram of the isolated sclera and structures that blend with it (muscle tendons and optic nerve dura) or traverse its substance. (**B**) The scleral layers. (**C**) The pattern of orientation of the collagen bundles in the scleral stroma in relation to the extraocular muscle tendinous insertions. (**D**) Collagen bundles in sclera. Original magnification: **D**, × 7000 (Part D courtesy of Dr A. Thale, from Thale and Tillmann, 1993.)

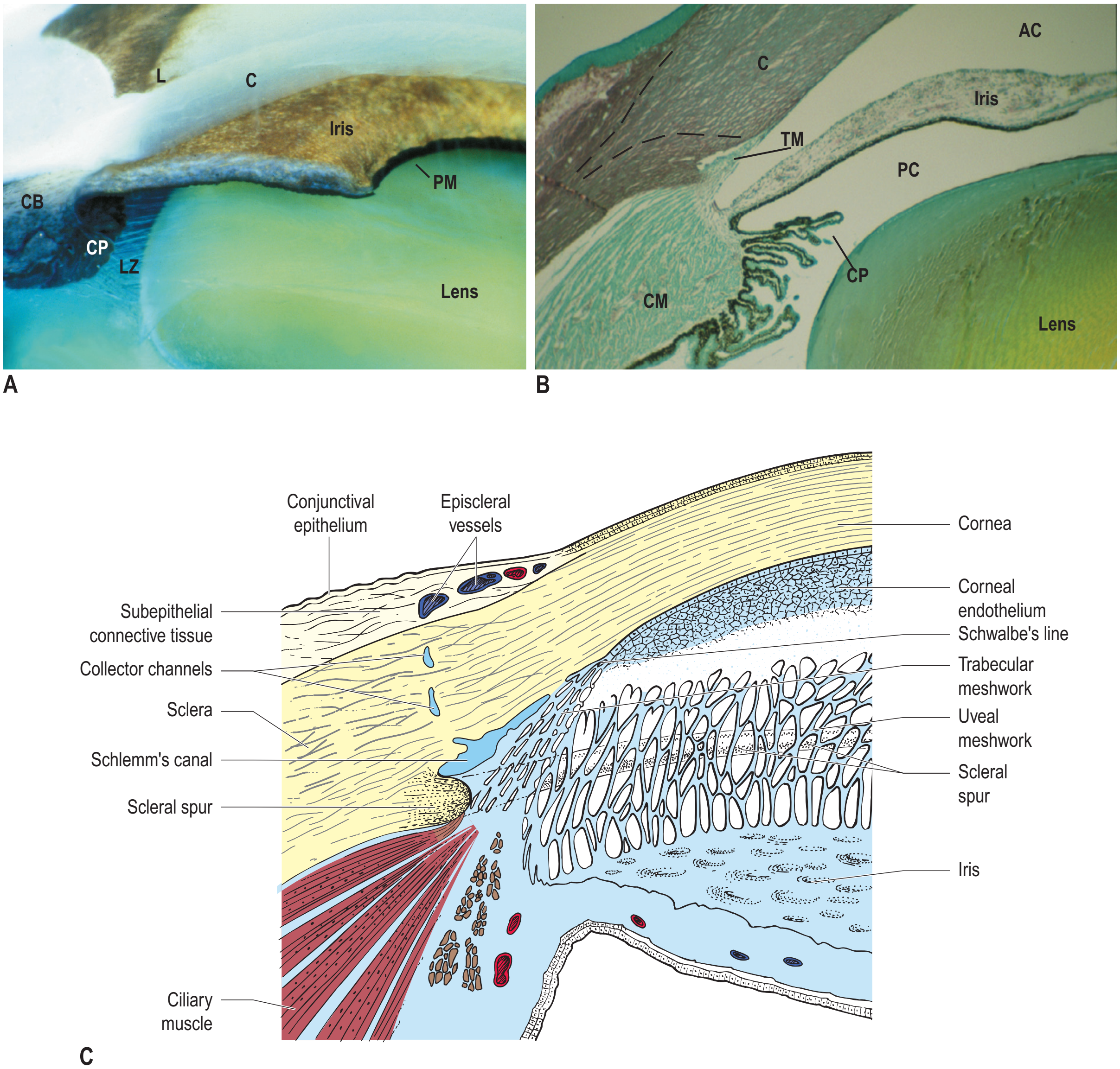
L
C
Iris
PM
CB
CP
LZ
Lens
A
AC
C
Iris
TM
PC
CP
CM
Lens
B
Conjunctival epithelium
Episcleral vessels
Cornea
Subepithelial connective tissue
Corneal endothelium
Schwalbe's line
Collector channels
Trabecular meshwork
Sclera
Uveal meshwork
Schlemm's canal
Scleral spur
Scleral spur
Iris
Ciliary muscle
C

FIGURE 1-18 (**A**) Macroscopic photograph of the anterior segment of a primate eye, which is almost identical to the human eye. Note the heavily pigmented limbus (L) characteristic of primates: C, cornea; CB, ciliary body; CP, ciliary processes; LZ, lens zonules; PM, pupillary margin. (**B**) Histology section of the primate anterior segment (Van Gieson stain). Dotted line indicates the corneoscleral junction: AC, anterior chamber; CM, ciliary muscle; PC, posterior chamber; TM, trabecular meshwork; C, cornea; CP, ciliary processes. (**C**) Three-dimensional schematic diagram of important features in and around the iridocorneal angle and corneoscleral limbus. Original magnifications: **A** and **B**, × 50.

effects of vasoactive amines (e.g. histamine, leukotrienes, prostaglandins) released by local immune cells (see below).

- Desçemet's membrane and Bowman's layer terminate in this region.
- The loose conjunctival subepithelial vascularized connective tissue (substantia propria), containing immunocompetent cell types such as mast cells, plasma cells and lymphocytes, tapers off at the limbus and is absent in the cornea (Fig. 1-18B).

BOX 1-11 DEFINITIONS

Anatomical and surgical limbus

Definitions of the limits and markings of the limbus vary among anatomists, pathologists and surgeons. The anatomical (histological) limbus is defined by a line that follows the V-shaped transition of corneal lamellae to scleral lamellae (Fig. 1-18B). Pathologists define the limbus as a block of tissue bordered anteriorly by a line passing through the termination of Schwalbe's line and the junction of the conjunctival and corneal epithelium (corneolimbal junction), and posteriorly by a line from the scleral spur perpendicular to the tangent of the external surface. Surgeons usually cut close to the blue–grey transition zone seen on external examination, and incisions made here will pass anterior to the trabecular meshwork and Schlemm's canal (Fig. 1-18B,C).

Aqueous outflow pathways

In the chamber or iridocorneal angle, partially nestled in the internal scleral sulcus, lies a complex wedge-shaped circumferential band of specialized, sponge-like, connective tissue, the *trabecular meshwork*, with the *canal of Schlemm* (sinus venosus sclerae) on its outer aspect (Figs 1-18B,C and 1-19A). The base of the trabecular meshwork is formed posteriorly by the scleral spur, the anterior face of the ciliary muscle and the iris root. The apex of the meshwork terminates anteriorly at Schwalbe's line and the adjacent innermost corneal lamellae (Figs 1-18 and 1-19A). The trabecular meshwork can be further subdivided into three anatomical zones: the innermost *uveal meshwork* with cord-like trabeculae; the *corneoscleral meshwork* with flattened sheet-like trabeculae (Fig. 1-19D); and the outermost *cribriform meshwork* beneath the inner wall of Schlemm's canal (Fig. 1-19E). The cribriform meshwork, unlike the rest of the trabecular meshwork, is not arranged in lamellae but consists of trabecular cells enmeshed in a loose extracellular matrix of collagen (types I, III and IV), elastic-like fibres and proteoglycans. This layer is thought to be the main site of resistance to aqueous outflow. The elastic cores of the trabeculae are continuous with the elastic fibres in the cribriform meshwork, which are in turn connected to the inner wall of Schlemm's canal via 'connecting fibres'. The anterior ciliary muscle fibres terminate in the elastic cores of the trabeculae. As the ciliary muscle contracts and moves inwards, the three-dimensional trabecular meshwork can be expanded, which results in an increase in the amount of 'free' spaces in the cribriform meshwork. This in turn allows greater aqueous outflow, thus increasing aqueous outflow facility.

Aqueous humour passes from the anterior chamber through the *intertrabecular* and *intratrabecular spaces*, which are lined by trabecular cells. These cells envelop the trabeculae (Fig. 1-19D) and maintain the state of hydration of the connective tissue core in a similar manner to corneal endothelium. In addition, trabecular cells are also phagocytic, trapping and removing debris from the aqueous humour as it percolates through the tortuous intertrabecular and intratrabecular spaces which narrow as Schlemm's canal is approached (Figs 1-18C and 1-19A).

Schlemm's canal (*sinus venosus sclerae*) is an endothelium-lined 36 mm long circumferential channel filled with aqueous humour. It measures 200–400 μm in the anteroposterior axis, is seldom more than 50–60 μm deep and is often septate. The canal is drained by 25–35 *collector channels* (20–90 μm in diameter) and between two and eight *aqueous veins* (of Ascher) (up to 100 μm in diameter). These either join deep, intrascleral and episcleral venous plexuses which drain into conjunctival veins or, in the case of aqueous veins, may drain directly into superficial conjunctival veins. The majority of aqueous humour (70–90%) leaves the anterior chamber through the trabecular meshwork and Schlemm's canal (*'conventional' outflow pathways*). The inner wall of the canal is characterized, in well-preserved and properly fixed eyes, by transcellular channels or giant vacuoles (Fig. 1-19B,C). There is good evidence to suggest that these intracellular vacuoles, with openings on both the trabecular and luminal aspects, function to drain the great bulk of aqueous humour. The number and size of vacuoles and their openings or pores vary in a pressure-sensitive manner. Small quantities of aqueous may also pass between endothelial cells in the canal wall.

A proportion (10–30%) of aqueous humour drains via the *'non-conventional' aqueous outflow pathways*. This route is not pressure sensitive and consists of the intercellular spaces between ciliary muscle fibres and the loose connective tissue of the suprachoroidal space. From here, aqueous traverses the sclera via the

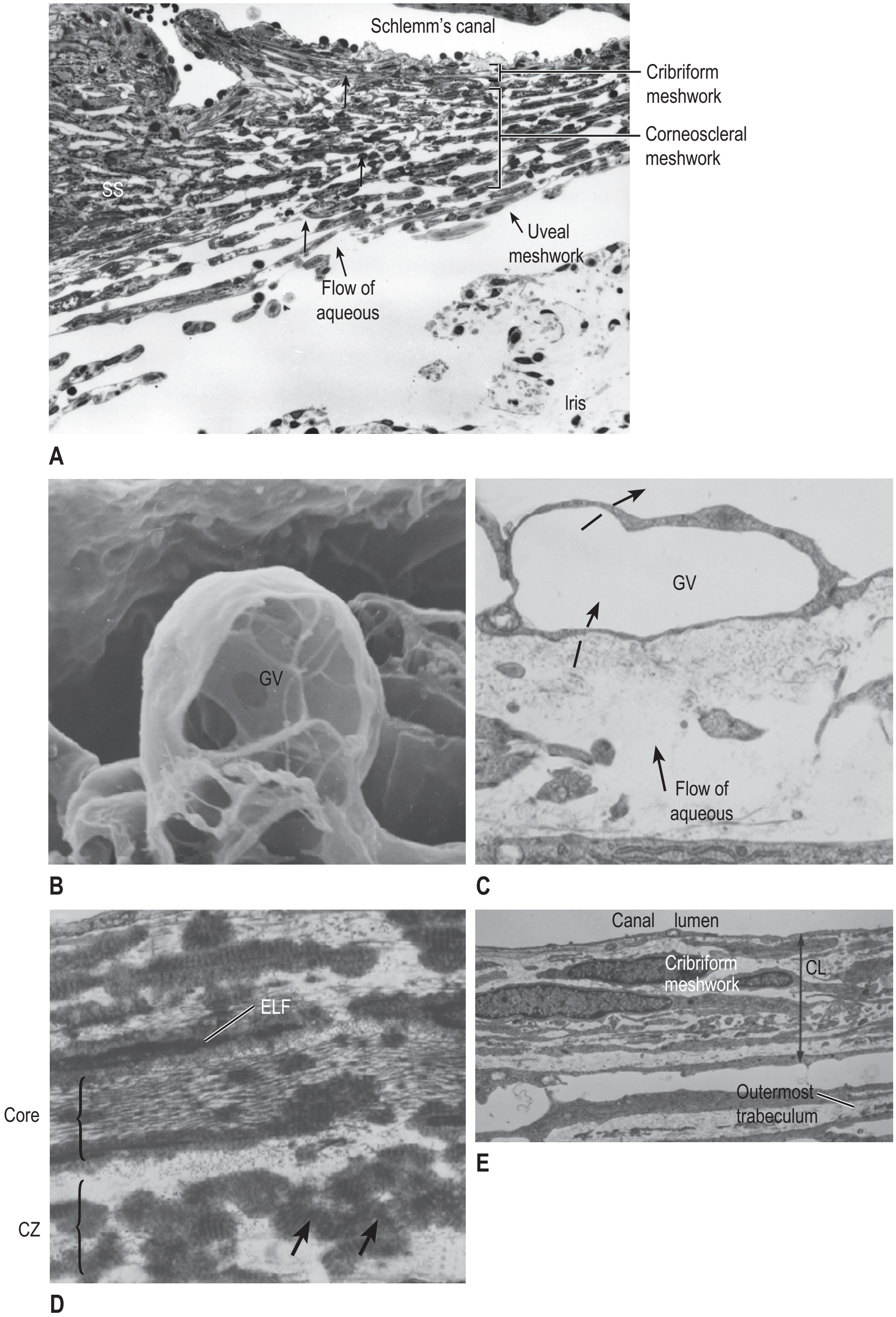
Schlemm's canal
Cribriform meshwork
Corneoscleral meshwork
SS
Uveal meshwork
Flow of aqueous
Iris
A
GV
B
GV
Flow of aqueous
C
ELF
Core
CZ
D
Canal lumen
Cribriform meshwork
CL
Outermost trabeculum
E

FIGURE 1-19 Histology and ultrastructure of the trabecular meshwork and Schlemm's canal. (**A**) Light micrograph showing the scleral spur (SS), Schlemm's canal and the three zones of the meshwork: the uveal, corneoscleral and cribriform meshworks. The path of aqueous through the inter- and intratrabecular spaces is indicated by arrows. (**B**) 'Giant vacuole' (GV) in the inner wall of Schlemm's canal as seen by scanning electron microscopy. (**C**) Transmission electron micrograph of a 'giant vacuole'. The flow of aqueous is indicated by the arrows. This particular section does not include the basal or luminal pores seen in some 'giant vacuoles'. (**D**) Electron micrograph of a trabecula cut in cross-section showing the layered arrangement of the extracellular components: CZ, cortical zone; ELF, elastic-like fibres; arrows, 'long spacing collagen'. (**E**) High-power micrograph of the cribriform meshwork or layer showing the lack of trabecular organization. Fibrocyte-like cells are loosely arranged in various types of extracellular matrix. Original magnifications: **A**, × 340, **B** and **C,** × 7000, **D,** × 13 000, **E,** × 5000. (From McMenamin, Lee and Aitken, 1986, *Ophthalmology*, with permission.)

connective tissue sheaths of nerves and vessels that pierce its substance (see Fig. 1-17A).

Uveal tract or uvea

The uveal tract (L. *uva* = grape), the middle vascular pigmented layer of the eye, consists of the *iris, ciliary body* and *choroid* (see Fig. 1-10). These three components are continuous with one another and have an opening anteriorly, the *pupil*, and posteriorly the choroid is deficient at the *optic nerve canal*. The uveal tract is analogous to the vascular pia-arachnoid of the brain and optic nerve, with which it anastomoses at the optic nerve head. The choroid is described on p. 55; the iris and ciliary body are described below.

THE IRIS

The iris (Fig. 1-20A,B) is a thin, heavily pigmented, contractile circular disk analogous to the diaphragm of a camera. It is suspended in the frontal or coronal plane anterior to the lens and ciliary body, and is surrounded by aqueous humour. The iris separates the anterior and posterior chambers, which are in continuity through an opening, the *pupil*, which lies slightly inferonasal to the centre of the iris. The iris is attached by its *root* at the angle (iridocorneal) of the anterior chamber where it merges with the ciliary body and trabecular meshwork. The free edge is known as the *pupillary margin.*

The iris is 12 mm in diameter with a circumference of 37 mm. It is cone-shaped with the pupil margin positioned more anteriorly than the root. The pupil margin rests on the lens, without whose support, for example in aphakic patients, it becomes tremulous (iridonesis). The size of the pupil regulates the amount of light entering the eye and is dependent on the state of contraction of the intrinsic pupillary muscles, the *dilator* and *sphincter* pupillae. The pupil may vary from 1 to 8 mm in diameter and there may be a slight degree of asymmetry between right and left eyes in normal individuals.

BOX 1-12 CLINICAL CORRELATES

GLAUCOMA

Glaucoma is defined as a progressive optic nerve neuropathy. For most forms of glaucoma, elevated intraocular pressure and ageing remain important risk factors, although low tension or normal tension forms of glaucoma are common. However, in many forms of this condition pathological changes in the trabecular meshwork and Schlemm's canal may be responsible for increased resistance to aqueous outflow and raised, or diurnal fluctuations in, intraocular pressure (IOP). In *congenital glaucoma* there is malformation of the complex three-dimensional arrangement of the trabeculae and excess extracellular matrix in the outer meshwork. Physical blockage of the inner surface of the chamber angle by the iris occurs in *closed-angle glaucoma*; this may be a primary or a secondary process (see p. 512). Various forms of obstruction in the trabecular meshwork may give rise to *open-angle glaucoma* (see p. 512); the cause of the primary form of this condition is unknown, although there is evidence to indicate that excessive deposition of extracellular elements may occur in the cribriform meshwork. Secondary forms of open-angle glaucoma may be the result of debris such as lens proteins, melanin, macrophages and haemorrhagic products physically obstructing the intertrabecular and intratrabecular spaces, thus causing raised IOP (see p. 512).

Structure (Fig. 1-20A,B)

The pupil margin and iris root are thin, and hence more susceptible to tearing in contusion injuries (iridodialysis). The anterior surface is divided into two zones, the *ciliary zone* and *pupil zone*, by the thickened region known as the *collarette*. The anterior surface is characterized by *radial streaks* (straight when the pupil is contracted and wavy when dilated) and *contraction furrows* (more noticeable in dilated irides). The surface of the iris appears smooth in dark irides, in which the intrastromal melanocytes are heavily pigmented, and more irregular in blue irides, which have a less heavily pigmented stroma. The blue appearance of some irides is because of absorption of long wavelengths and reflectance of shorter wavelengths, especially by the collagenous stroma.

Microscopic anatomy

The iris consists of four layers: anterior border layer, stroma, dilator pupillae muscle and posterior pigment epithelium.

Anterior border layer (Figs 1-20B and 1-21A). This layer is not covered by a layer of epithelial cells, as early anatomists believed, but is in fact made up of modified stroma consisting of a dense collection of fibroblasts, melanocytes and a few interspersed

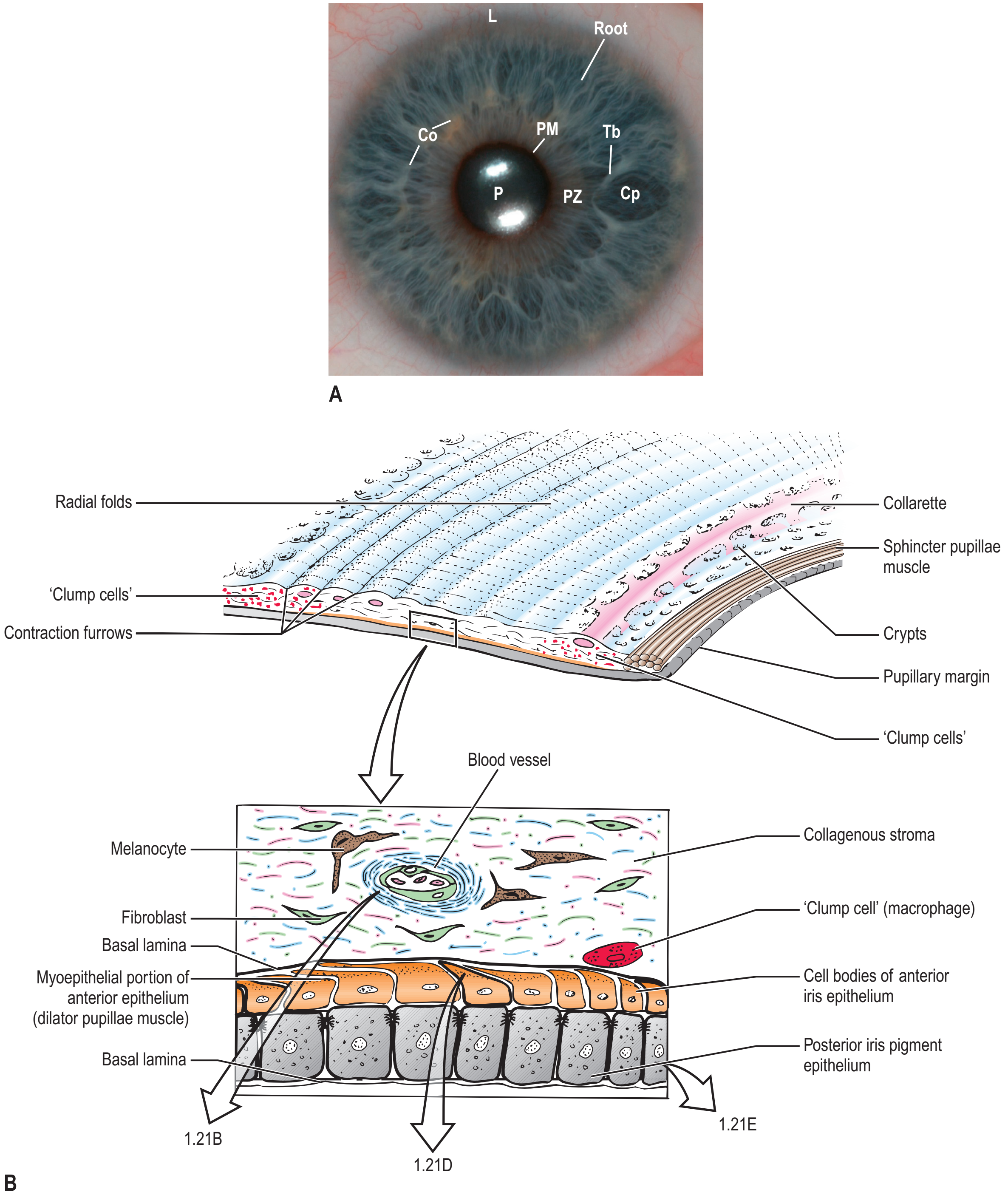
L
Root
Co
PM
Tb
P
PZ
Cp
A
Radial folds
'Clump cells'
Contraction furrows
Collarette
Sphincter pupillae muscle
Crypts
Pupillary margin
'Clump cells'
Blood vessel
Melanocyte
Fibroblast
Basal lamina
Myoepithelial portion of anterior epithelium (dilator pupillae muscle)
Basal lamina
Collagenous stroma
'Clump cell' (macrophage)
Cell bodies of anterior iris epithelium
Posterior iris pigment epithelium
1.21B
1.21D
1.21E
B

FIGURE 1-20 The structure of the iris. (**A**) Clinical macroscopic image of a blue iris. Note the pupil (P), pigmented pupil margin (PM) and pupillary zone (PZ) separated from the ciliary portion by the collarette (Co). The root is attached at the iridocorneal angle deep to the limbus. (L). Note the fine collagenous trabecula (Tb), some of which form the boundaries of ovoid crypts (Cp). (**B**) Top diagram: surface features of a portion of the iris. Bottom diagram: high-power exploded view summarizing the arrangement of the pigmented (posterior) and non-pigmented (anterior) layers.

collagen fibres. This layer is deficient in areas; consequently the iris stroma is in free communication with the aqueous humour in the anterior chamber. Larger deficiencies in the anterior border layer are evident macroscopically as *crypts*. Aggregates of heavily pigmented melanocytes in this layer appear as *naevi*.

Stroma (Fig. 1-21C–E). This consists of loose connective tissue containing fibroblasts, melanocytes and collagen fibres (types I and III). The loose nature of this tissue, and its free communication via openings in the anterior border layer, allows fluid to move in and out of the stroma quickly during dilation and contraction. The iris stroma in humans contains numerous mast cells and macrophages, many of which are perivascular (see Fig. 1-21C and Ch. 7). Many of the macrophages are heavily pigmented, and a subgroup may form large ovoid 'clump cells' (of Koganei), which tend to accumulate near the iris root and sphincter pupillae muscle (Fig. 1-20B). Lying free within the stroma close to the pupil margin is the *sphincter pupillae* muscle, a circumferential ring of smooth muscle fibres about 1 mm in width. The sphincter muscle consists of muscle bundles, each comprising six to eight smooth muscle cells, which are continuous via gap junctions and surrounded by a basal lamina. This muscle is innervated by parasympathetic nerve fibres derived from the oculomotor nerve (postganglionic fibres from the ciliary ganglion travel via the short ciliary nerves) although sympathetics also terminate in this muscle. The unusual embryological origin of this muscle from the neuroectoderm is described in Chapter 2.

Dilator pupillae muscle. This is a layer of myoepithelial cells derived from the *anterior iris epithelium*. The basal processes of this epithelium are 4 μm in thickness and extend up to 50–60 μm in a radial direction, while the apices of the myoepithelial cells are lightly pigmented and closely apposed to the apical aspect of the *posterior pigment epithelium* (Figs 1-20B and 1-21A,E). The *dilator pupillae muscle* is innervated by non-myelinated sympathetic fibres whose cell bodies are situated in the superior cervical sympathetic ganglion. Its parasympathetic innervation seems less significant. The dilator pupillae extends only as far centrally as the outer margin of the sphincter pupillae.

Posterior pigment epithelium (Fig. 1-21E). This heavily pigmented layer consists of large cuboidal epithelial cells that appear black macroscopically on examination of the posterior surface of the iris. The posterior layer is derived from the inner neuroectodermal layer of the optic cup (see Ch. 2). The cells extend for a short distance on to the anterior iris surface at the pupillary margin; this forms the black ruff seen on the pupil margin during slit-lamp examination of the eye (Fig. 1-20A). The posterior pigmented epithelial layer forms a series of radially arranged furrows (most evident near the pupil margin) and circumferential contraction folds (most evident in the periphery).

Pupil movements

Mydriasis (dilation) occurs in conditions of low light intensity and in states of excitement or fear. It is a result of the action of the dilator pupillae muscle.

Miosis (contraction) occurs in more illuminated conditions, during convergence, and while sleeping. It is the result of the action of the sphincter pupillae muscle.

Blood supply of the iris

The iris has a rich blood supply and extensive anastomoses. At the root there is an incomplete major 'circle' of the iris that is derived from anterior rami of the anterior ciliary arteries. Branches from here pass centripetally and form an incomplete minor arterial 'circle' at the level of the collarette.

The arteries have an unusual coiled form to accommodate the variable states of contraction of the iris. Veins lie close to the arteries, with larger veins primarily in the anterior stroma and smaller veins in the deeper layers. Veins drain posteriorly/centrifugally into the ciliary body and eventually the vortex veins (see Fig. 1-39A,E).

Iris capillaries are characterized by non-fenestrated endothelial cells that have a high density of endocytotic vesicles and tight junctions. This makes them less permeable to a variety of solutes than normal somatic vessels (hence they do not normally leak in fluorescein angiography). These vessels thus constitute an important component of the *blood–ocular barrier.* The basal lamina of the endothelial cells is thickened (0.5–3 μm) and further strengthened by perivascular collagenous/hyalinized layers (Fig. 1-21A,B). Periarteriolar smooth muscle cells are rare and elastic fibres are absent.

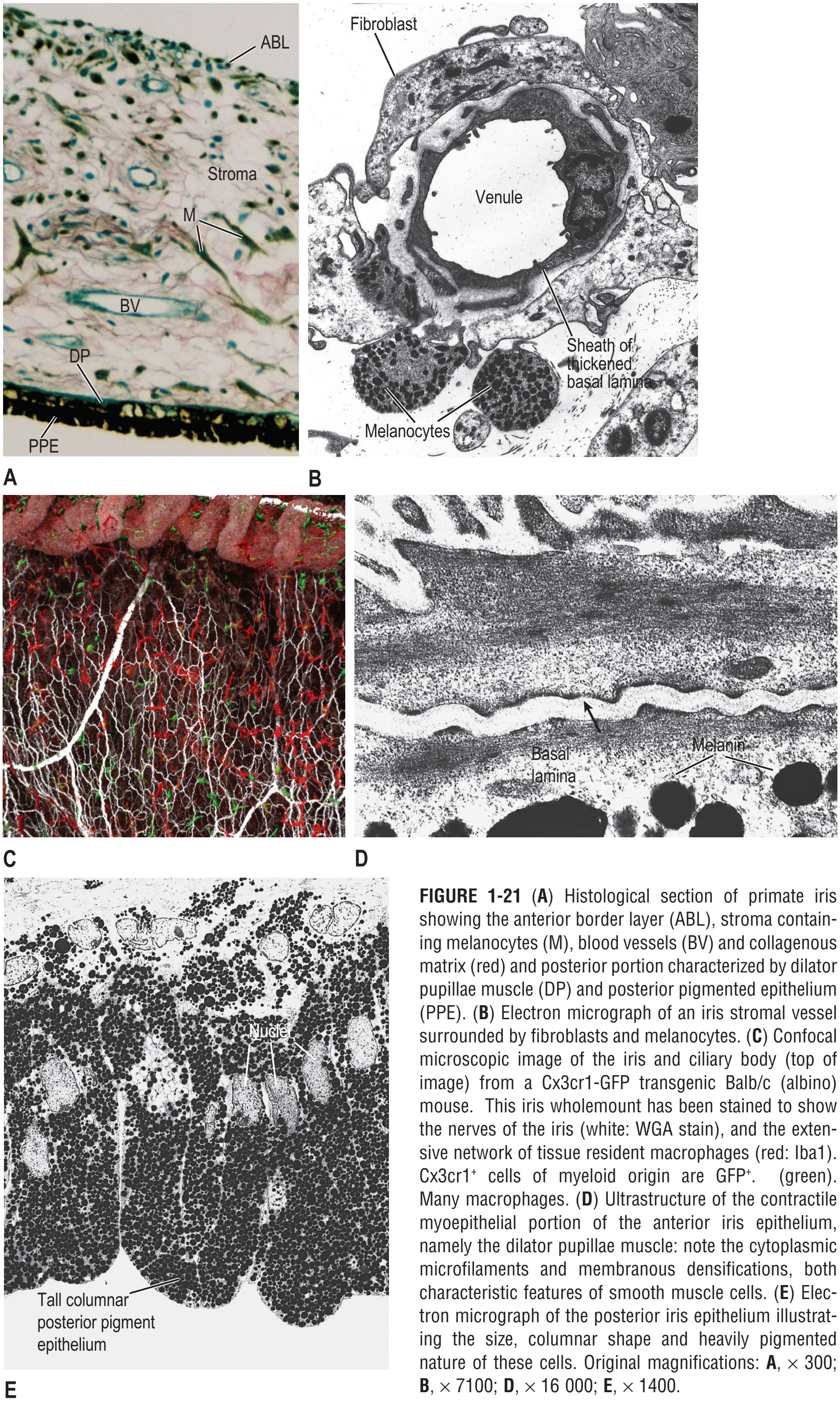
ABL
Stroma
M
BV
DP
PPE
A
Fibroblast
Venule
Sheath of thickened basal lamina
Melanocytes
B
C
Basal lamina
Melanin
D
Nuclei
Tall columnar posterior pigment epithelium
E

FIGURE 1-21 (**A**) Histological section of primate iris showing the anterior border layer (ABL), stroma containing melanocytes (M), blood vessels (BV) and collagenous matrix (red) and posterior portion characterized by dilator pupillae muscle (DP) and posterior pigmented epithelium (PPE). (**B**) Electron micrograph of an iris stromal vessel surrounded by fibroblasts and melanocytes. (**C**) Confocal microscopic image of the iris and ciliary body (top of image) from a Cx3cr1-GFP transgenic Balb/c (albino) mouse. This iris wholemount has been stained to show the nerves of the iris (white: WGA stain), and the extensive network of tissue resident macrophages (red: Iba1). $Cx3cr1^+$ cells of myeloid origin are GFP^+. (green). Many macrophages. (**D**) Ultrastructure of the contractile myoepithelial portion of the anterior iris epithelium, namely the dilator pupillae muscle: note the cytoplasmic microfilaments and membranous densifications, both characteristic features of smooth muscle cells. (**E**) Electron micrograph of the posterior iris epithelium illustrating the size, columnar shape and heavily pigmented nature of these cells. Original magnifications: **A**, × 300; **B**, × 7100; **D**, × 16 000; **E**, × 1400.

Nerve supply

The iris possesses a rich three-dimensional nerve plexus of myelinated and non-myelinated nerves. The sensory nerves are branches of the long and short ciliary nerves, themselves branches of the nasociliary nerve (ophthalmic division of the trigeminal). The autonomic innervation of the iris muscles is discussed above.

THE CILIARY BODY

The ciliary body (Fig. 1-22A–D) is an approximately 5–6 mm wide ring of tissue that extends from the scleral spur anteriorly to the ora serrata posteriorly. Temporally it measures 5.6–6.3 mm, and nasally 4.6–5.2 mm. It is divided into two zones, an anterior *pars plicata* (corona ciliaris) and a posterior *pars plana* (Fig. 1-22A). The ciliary body is approximately triangular in cross-section; its base faces the anterior chamber and the apex blends posteriorly with the vascular choroid. The pars plicata is 2 mm wide and consists of 70 radially arranged folds known as *ciliary processes* (Fig. 1-22B), each of which is 0.5–0.8 mm high and 0.5 mm wide. The tips are paler as a result of decreased pigmentation. Minor ciliary processes may be present in the valleys between the major processes. The pars plana is an approximately 4 mm wide zone stretching from the posterior limits of the ciliary processes to the *ora serrata*, the sharp serrated or dentate junction where non-pigmented ciliary epithelium undergoes a sharp transition to become the neural retina.

The ciliary body can be divided histologically into the ciliary epithelium, ciliary body stroma and ciliary muscle (Fig. 1-22A). Each is described below in the context of the three principal functions of the ciliary body: (1) accommodation; (2) aqueous humour production; and (3) production of lens zonules, vitreal glycosaminoglycans and vitreal collagen.

Accommodation

The anterior two-thirds of the ciliary body is occupied by the *ciliary muscle* which, in conjunction with the *lens zonules* (suspensory ligament) and the natural elastic nature of the lens fibres and capsule, functions to alter the refractive power of the lens. Histologically, the ciliary muscle in meridional sections consists of three groups of smooth muscle fibre bundles embedded in a vascular connective tissue stroma (Figs 1-18 and 1-22A). The stroma contains melanocytes, fibroblasts, and occasional immune cells such as mast cells, macrophages and lymphocytes.

The *outer longitudinal muscle fibres* are attached to the scleral spur and therefore indirectly to the corneoscleral trabeculae anteriorly; posteriorly they are anchored to the inner aspect of the sclera. The *middle oblique* or *radial muscle fibres* are continuous with inner corneoscleral trabeculae. The *inner circular muscle fibres* (one of the three so-called 'Müller's muscles' associated with the eye and adnexa) appear as cross-sectional profiles in conventional meridional sections of the ciliary body (Figs 1-18C and 1-22A). A three-dimensional scheme has been proposed to explain the interrelationship between all three groups of muscle bundles; it appears that the fibres are all components of one interwoven fibre network (see Fig. 1-23 for summary).

The mode of action of the ciliary muscle is still controversial; however, it is generally agreed that during accommodation there is some degree of forward and inward shift of the ciliary body, which serves to slacken the tension on the zonules, thus increasing the refractive power of the lens (see pp. 35 and 217).

Action of the ciliary muscle on aqueous outflow. There is some evidence that contraction of the longitudinal ciliary muscle fibres causes an inwards and posterior movement of the scleral spur as well as an effect via extensions of these tendons to the inner wall of Schlemm's canal, hence distending the inter- and intratrabecular spaces of the trabecular meshwork and preventing collapse of Schlemm's canal during periods of high pressure. These have been proposed as modes of action of miotic drugs such as pilocarpine, which are used to increase aqueous outflow facility in patients with *glaucoma*.

Production of aqueous humour

Aqueous humour is a clear colourless fluid actively secreted by the ciliary processes (for details of aqueous constituents see Ch. 5). The ciliary processes consist essentially of delicate finger-like protrusions of loose vascular connective tissue covered by a bilaminar cuboidal/columnar neuroepithelium, the outer pigmented and inner non-pigmented ciliary epithelium, both derived embryologically from the double neuroectoderm of the optic cup (see Ch. 2). The morphological basis of the blood–aqueous barrier is depicted

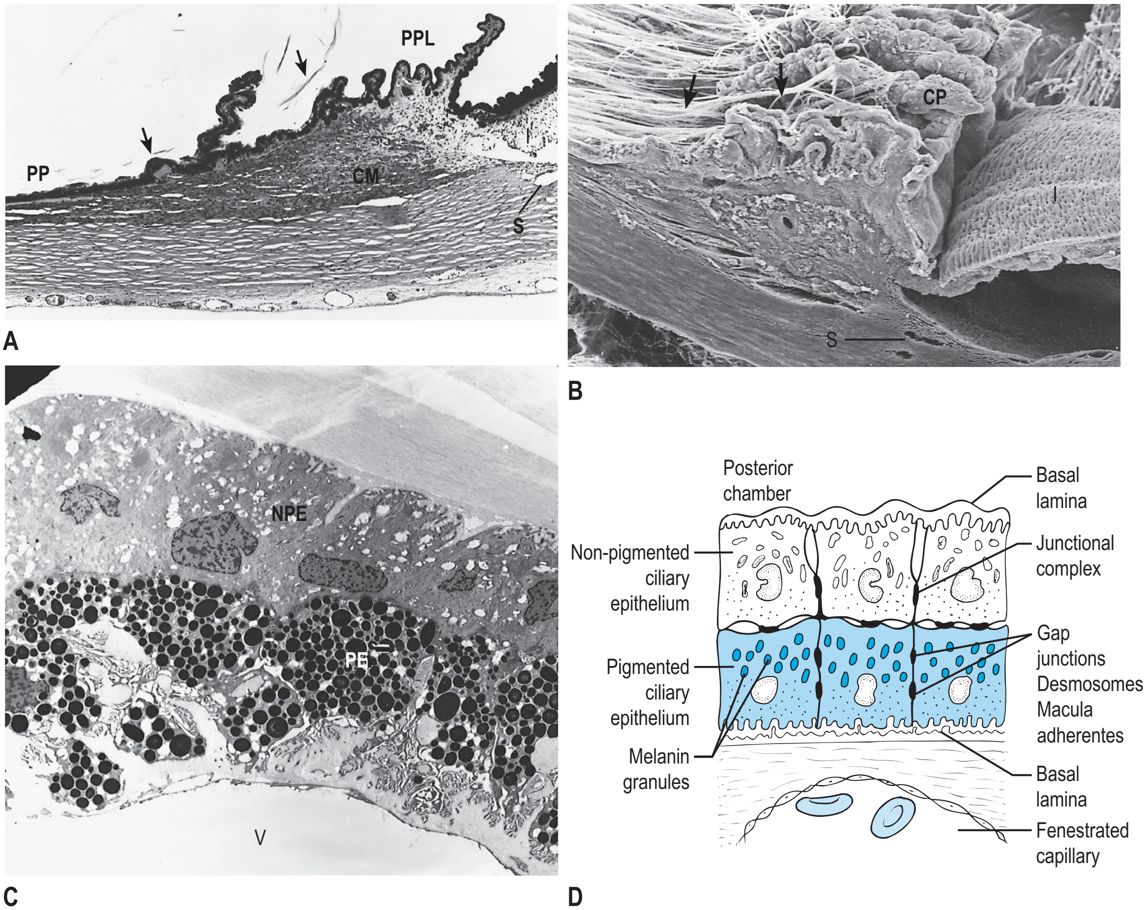
PPL
PP
CM
S
A
CP
I
S
B
NPE
PE
V
C
Posterior chamber
Basal lamina
Non-pigmented ciliary epithelium
Junctional complex
Pigmented ciliary epithelium
Gap junctions
Desmosomes
Macula adherentes
Melanin granules
Basal lamina
Fenestrated capillary
D

FIGURE 1-22 (**A**) Histological section of the ciliary body showing the two major regions; the pars plana (PP) and the pars plicata (PPL) which includes the ciliary processes: CM, ciliary muscle; I, iris; S, Schlemm's canal. (**B**) Scanning electron micrograph of the inner surface of the ciliary processes (CP) and iris (I). Arrows in (**A**) and (**B**), zonular fibres. (**C**) Low-power electron micrograph of the pigmented (PE) and non-pigmented ciliary epithelia (NPE) of a ciliary process. Note the large fenestrated blood vessel (V) and the lens zonules blending with the basal lamina of the non-pigmented epithelium. (**D**) Diagrammatic representation of this double layer of epithelium, which constitutes the major site of the blood–aqueous barrier. Original magnifications: **A**, × 40; B, × 45; **C**, × 1600. (Part A courtesy of W.R. Lee and Springer-Verlag.)

in Figure 1-22B. Aqueous humour is actively secreted by the inner non-pigmented ciliary epithelium whose apices are connected by junctional complexes including tight junctions. Macromolecules, having filtered through the highly permeable fenestrated stromal capillaries, pass between pigment epithelium cells held together only by punctate adherentes (macula adherentes)-like junctions or permeable band-like junctions (zonula adherentes), and are prevented from passing into the posterior chamber. The non-pigmented ciliary epithelium has morphological features characteristic of secretory epithelium, namely numerous mitochondria, and histochemically can be shown to contain enzymes, such as carbonic anhydrase and ATPase, necessary for active fluid transport (see Ch. 4, p. 217).

Production of zonules, vitreal collagen and vitreal hyaluronic acid

The non-pigmented ciliary epithelial cells of the pars plana are cuboidal, columnar or irregular, depending on age and location. It is likely that these cells play a role in secretion of zonular fibres, the extracellular vitreal components, i.e. collagen and hyaluronic acid, and the inner limiting membrane, especially during embryonic development.

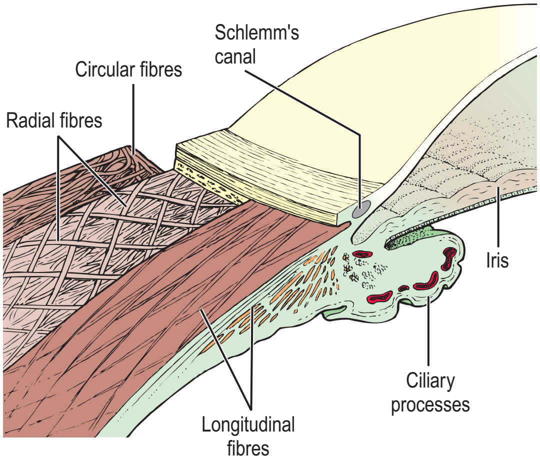

FIGURE 1-23 Arrangement of the ciliary muscle fibres as seen on external view with the sclera removed. (After Rohen and Ciliankorpër, 1964.)

Some inner non-pigmented cells 'tilt' anteriorly, suggesting traction by the zonular fibres. The lens zonules (of Zinn) or suspensory ligament of the lens merge with the fibrous basal lamina material of the non-pigmented ciliary epithelium (Fig. 1-22C). The precise mode of zonule synthesis is still unclear. Tall non-pigmented ciliary epithelial cells in the pars plana have ultrastructural and histochemical features that indicate active hyaluronic acid secretion.

Blood supply of the ciliary body

The blood supply of this region is derived from long posterior ciliary arteries and anterior ciliary arteries. The two long posterior ciliary arteries arise from the ophthalmic artery and after piercing the sclera near the optic nerve head they travel forward in the choroid in the medial and lateral horizontal plane and divide in the ciliary body before anastomosing with anterior ciliary branches, thus forming the major 'circle' of the iris (see p. 217, and Fig. 1-36). From this circle arises muscular and recurrent choroidal branches and the numerous branches that form the vascular plexus in the ciliary processes. Venous return occurs predominantly posteriorly through the system of vortex veins and less so through the anterior ciliary veins.

Nerve supply of the ciliary body

The ciliary body has rich parasympathetic, sympathetic and sensory innervations.

Parasympathetic innervation. Preganglionic neurone cell bodies are located in the Edinger–Westphal nucleus which lies posterior to the main oculomotor nucleus in the rostral midbrain at the level of the superior colliculus. Their axons travel in the oculomotor (III) nerve and synapse with postganglionic cell bodies located in the ciliary ganglion. The postganglionic fibres travel to the eye in the short ciliary nerves and terminate as an extensive plexus in the ciliary muscle. The action is mediated by acetylcholine on muscarinic receptors (see Ch. 6, p. 355).

Sympathetic innervation. Preganglionic neurone cell bodies are situated in the lateral grey horn of the first thoracic segment of the spinal cord. Preganglionic fibres relay in the superior cervical ganglion (adjacent to vertebrae C2 and C3, behind the internal carotid artery). Postganglionic fibres leave the ganglions as the internal carotid nerve and plexus. These fibres may reach the orbit as either direct branches of the internal carotid plexus or by joining the ophthalmic division of the trigeminal and its main branch in the orbit, the nasociliary nerve. Sympathetic fibres may either pass directly to the retrobulbar plexus behind the eye or through the ciliary ganglion uninterrupted. From the ciliary ganglions, fibres are distributed via the short ciliary nerves to the blood vessels of the eye, including the ciliary body. Some terminal filaments of the internal carotid plexus may also be distributed via the ophthalmic artery and its branches. The sympathetic action is mediated by the action of norepinephrine on two subclasses of receptors, α_1 and β_2 adrenoceptors, both of which are inhibitory (see Ch. 6, p. 358).

Sensory innervation. Sensory innervation is derived from the nasociliary nerve; however, the function of these fibres is unknown.

THE LENS AND ZONULAR APPARATUS

The lens (Fig. 1-24) is a highly organized system of specialized cells (so-called lens 'fibres'), which constitutes an important component of the optical system of the eye and fulfils the important function of altering the refractive index of light entering the eye to focus on the retina. While it has less refractive power (15 dioptres) than the cornea, the lens has the ability to change shape, under the influence of the ciliary muscle, and thus alter its refractive power. The range of dioptric power diminishes with age (8 at 40 years, 1–2 by 60 years). The transparency of the lens is due to the shape, arrangement, internal structure and biochemistry of the lens cells or lens fibres.

Position, size and shape

The lens, enclosed in its capsule, lies behind the iris and in front of the vitreous body. It is encircled by the ciliary processes and held in position by the zonular fibres laterally, the anterior vitreous face posteriorly (patellar fossa), and the iris anteriorly (see Fig. 1-10). It is normally transparent and avascular following regression of the pupillary membrane and tunica vasculosa lentis late in fetal development (see Ch. 2). It receives its nourishment from the aqueous and vitreous humours. It is a biconvex, ellipsoid structure with differing radius of curvature on the anterior and posterior surfaces. The anterior curvature is approximately 10 mm (range 8–14 mm) and the posterior curvature is approximately 6 mm (range 4.5–7.5 mm). The centre points of these surfaces, described as the *anterior* and *posterior poles*, are connected by an imaginary *axis*. The anterior pole lies 3 mm from the posterior corneal surface. The anterior and posterior surfaces are separated by the *equator*, which has a ridged (indented) appearance caused by the zonular fibres. In the adult eye the lens measures approximately 10 mm in diameter and has an axial length of 4 mm. The lens continues to grow (0.023 mm per year) and alters shape throughout life. It becomes rounder with age, especially after the age of 20 years.

Structure

The lens comprises three parts: (1) the capsule; (2) anterior or lens epithelium; and (3) the lens fibres (Fig. 1-24A,D).

Lens capsule. The *lens capsule* is a thickened, smooth, basement membrane produced by the lens epithelium and lens fibres. It completely envelops the lens and has regions of variable thickness, being thickest pre- and post-equatorially (17–28 µm) and thinner at the posterior (2–3 µm) than at the anterior pole (9–14 µm). Ultrastructural examination reveals a fibrillar or lamellar appearance. The interfibrillar matrix consists of basement membrane glycoproteins (type IV collagen) and sulphated GAGs which are responsible for its prominent periodic acid–Schiff-positive staining properties in histological sections. It possesses elastic properties and, when not under tension of the zonules, the capsule together with the cortex causes the lens to assume a more rounded shape.

Lens epithelium. This is a simple cuboidal epithelium (Fig. 1-24A,D) restricted to the anterior surface of the lens. The cells become more columnar at the equator. As they elongate, the apical portion comes to lie deeper to other, more anteriorly positioned, lens cells. These elongated lens cells are known as lens ‘fibres’ (Fig. 1-24D). The manner in which equatorial lens epithelial cells are transformed into lens fibres is depicted in Figure 1-24A. The cell nucleus and cell body sink deeper into the lens as further cells are laid down externally. Mitotic activity is maximal in the pre-equatorial and equatorial lens epithelium, known as the germinative zone (Fig. 1-24A).

Lens fibres. While each lens fibre is only a 4 × 7 µm hexagonal prismatic band in cross-section (Fig. 1-24C), it may be up to 12 mm (12 000 µm) in length. The apical portion of the elongated lens cell (or lens ‘fibre’) passes anteriorly, the basal portion posteriorly. The cell nucleus migrates anteriorly as the cell is pushed deeper in the lens, hence creating the anteriorly oriented *lens bow* (Fig. 1-24D). The meridionally oriented lens fibres extend the full length of the lens, meeting at the anterior and posterior sutures (Fig. 1-24A). Deeper (hence older) lens fibres are anucleate. Continual growth of the lens, by addition of superficial strips of new cells, produces a series of concentrically arranged laminae, similar to the layers of an onion (best seen by dissecting a fixed or frozen lens). In life, the outer *cortex* of the lens has a softer consistency than the hard central *nucleus.*

Lens fibres are tightly packed with little intercellular space. Neighbouring cells are linked by ball-and-socket cytoplasmic interdigitations (Fig. 1-24B,C) and numerous gap junctions. The junctions may aid maintenance of centrally positioned cells (via intercellular and molecular coupling or metabolic cooperation) some distance from the source of nutrition (aqueous humour). Superficially located lens fibres are rich in ribosomes, polysomes and rough endoplasmic

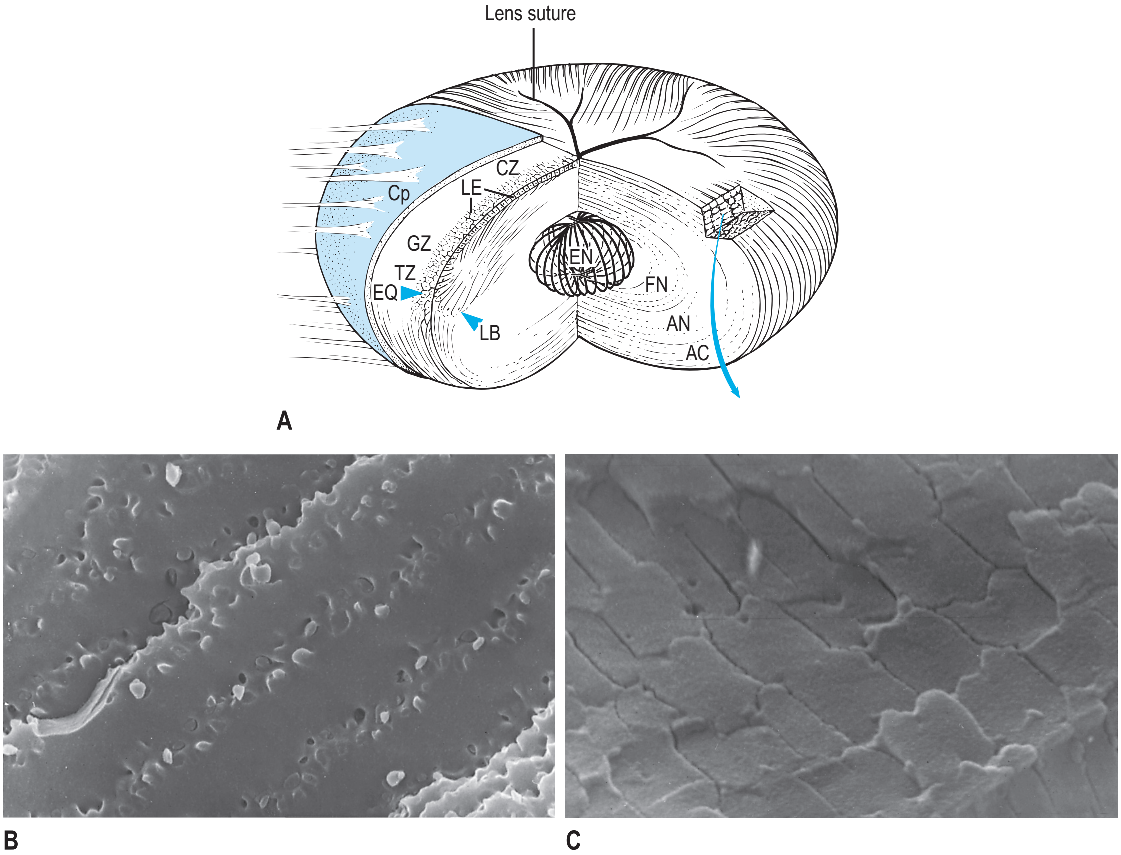

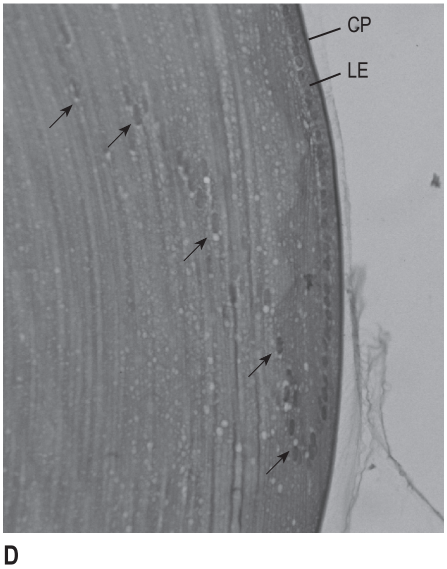

FIGURE 1-24 (**A**) The structure of the lens: CZ, central zone; GZ, germinative zone; TZ, transitional zone; LE, lens epithelium; EQ, equator; LB, lens bow; Cp, capsule; AC, adult cortex; AN, adult nucleus; FN, fetal nucleus; EN, embryonic nucleus. (**B** and **C**) Scanning electron micrographs of lens fibres in longitudinal section (**B**) and in cross-section (**C**). Note the 'tooth and peg' arrangement interlocking the adjacent lens fibre surfaces. (**D**) Histological section of the equatorial region of the lens showing the orientation of the lens fibres, the lens capsule, lens epithelium and lens bow (arrows). Original magnifications: **B**, × 2500; **C**, × 5000; **D**, × 150.

reticulum, and actively synthesize unique lens proteins, lens crystallins (see Ch. 4); however, the cytoplasm of mature lens fibres appears homogeneous. Lens fibres are rich in cytoskeletal elements oriented parallel to the long axis of the cell.

Lens zonules (zonular apparatus)

The lens is held in position by a complex three-dimensional system of radially arranged zonules (zonules of Zinn or the suspensory ligament of the lens) (Fig. 1-25). These delicate fibres are attached to the lens capsule 2 mm anterior and 1 mm posterior to the equator, and arise from the pars plana region of the ciliary epithelium and pass forward closely related to the lateral surfaces of the ciliary processes (Fig. 1-25A,B). In humans, the zonules transmit the forces that flatten the lens, allowing the eye to focus on distant objects. However, the zonules are also present in non-accommodating species, where they presumably play a role in lens centration.

The zonules consist of dense, glassy bundles 5–30 μm in diameter. Each bundle consists of a series of fine fibres (0.35–1 μm in diameter), themselves composed of 10–12 nm diameter microfibrils. Biochemically, the zonules are unique, most closely resembling the periodontal ligament of the teeth. Proteomic analysis has revealed that the zonules are composed principally of fibrillin, a 350 kDa cysteine-rich glycoprotein. The zonular fibres are synthesized by cells of the non-pigmented ciliary epithelium. At their proximal end they emerge from the basal lamina of the ciliary epithelial cells. The distal portions of the zonular fibres connect to the lens capsule near the lens equator. In humans, the fibres appear to insert directly into the capsule and their tips can be seen to terminate below the capsular surface. The lens zonules are synthesized and maintained by cells in the non-pigmented ciliary epithelium. The zonule is synthesized during embryonic and early postnatal development and it likely that zonular proteins turn over slowly, if at all.

Accommodation. In the non-accommodated state, the ciliary body maintains tension on the zonules. During accommodation, movement of the ciliary body causes slackening of the zonules, whereupon the lens assumes an increased anterior curvature, with resultant increase in refractive power, owing to elasticity of the lens capsule and the outer cortical layers. Some authorities believe that there are two sorts of zonules: main zonules and 'tension' zonules, the latter being placed under tension during accommodation.

Posterior zonules are closely associated with the collagenous material of the anterior hyaloid membrane. Zonules running perpendicular to the main zonule stream form circumferential bands near the base of the ciliary processes or in the pars plana (posterior zonular girdle) and over the apices of the ciliary processes (anterior ciliary girdle).

ANTERIOR AND POSTERIOR CHAMBERS

The cavity anterior to the lens and lens zonules is divided into two chambers by the iris. These two chambers, the larger anterior and smaller posterior, communicate through the pupil. The boundaries of these two chambers are shown in Figure 1-26.

Posterior chamber

This is a very small irregularly shaped space whose size varies during accommodation. It is approximately triangular with it's apex at the pupil margin; the base is formed by the ciliary processes; the posterior border is the lens and zonular apparatus; and the anterior border is the posterior surface of the iris. Aqueous humour secreted by the ciliary processes continually

BOX 1-13 CLINICAL CORRELATES

CATARACT

Cataract is the loss of normal lens transparency, besides the normal age-related yellowing, and may be caused by cumulative damage by ultraviolet light or perturbations in lens fibre biochemistry. Opacification may be the result of damage or disruption of the capsule, the lens fibre configuration or the lens epithelium (see Ch. 9).

BOX 1-14 CLINICAL CORRELATES

Ectopis lentis in Marfan syndrome

In humans, mutations in the fibrillin-1 gene cause Marfan syndrome, a condition characterized by disturbances in connective tissue and skeletal elements. In these patients the zonules are disrupted and the lens dislocated or malpositioned within the eye – a condition known as ectopis lentis.

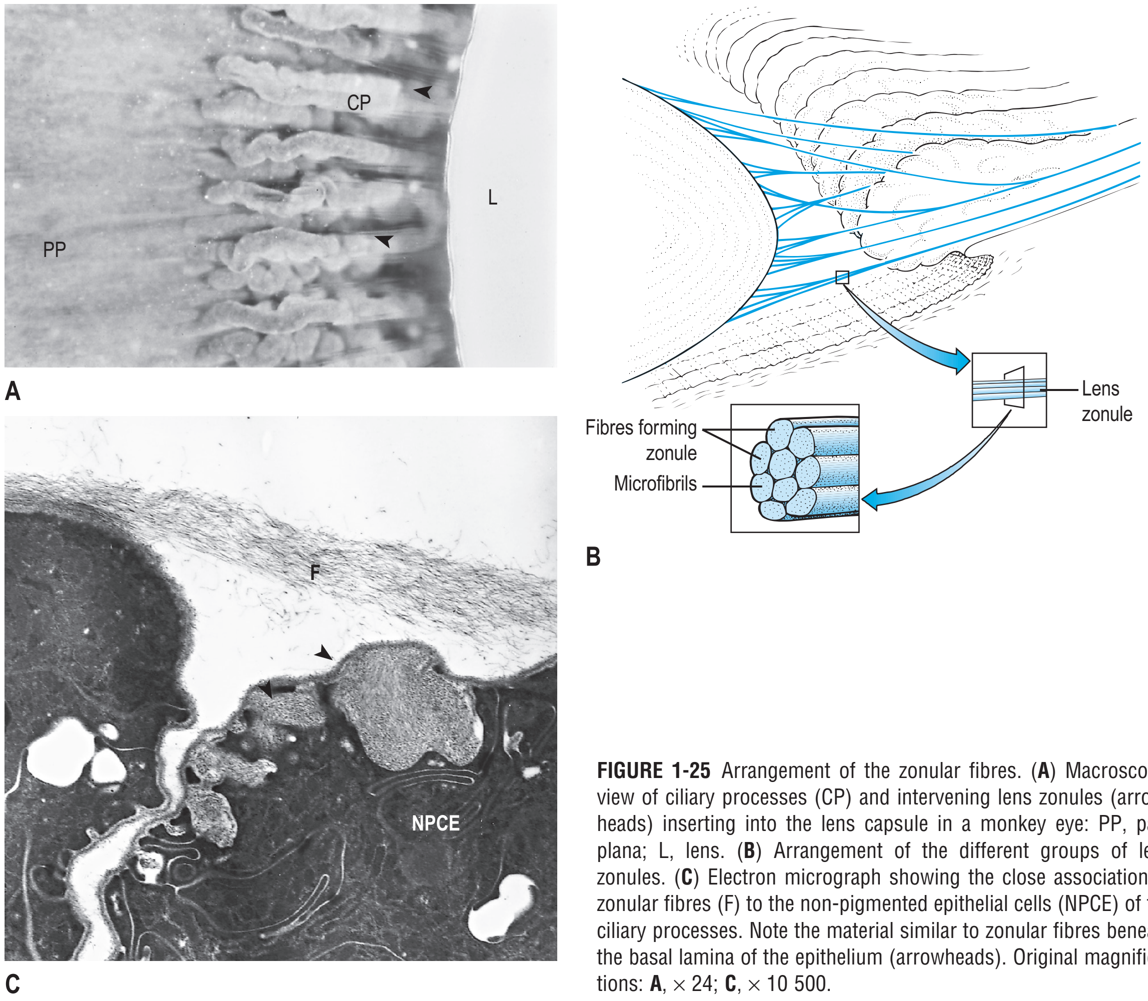

FIGURE 1-25 Arrangement of the zonular fibres. (**A**) Macroscopic view of ciliary processes (CP) and intervening lens zonules (arrowheads) inserting into the lens capsule in a monkey eye: PP, pars plana; L, lens. (**B**) Arrangement of the different groups of lens zonules. (**C**) Electron micrograph showing the close association of zonular fibres (F) to the non-pigmented epithelial cells (NPCE) of the ciliary processes. Note the material similar to zonular fibres beneath the basal lamina of the epithelium (arrowheads). Original magnifications: **A**, × 24; **C**, × 10 500.

enters this chamber before passing through the pupil into the anterior chamber.

Anterior chamber

The anterior chamber is bound in front by the cornea and posteriorly by the anterior iris surface and the pupillary portion of the lens. The lateral recess of the anterior chamber is formed by the iridocorneal angle occupied by the trabecular meshwork. The anterior chamber is deepest centrally (3 mm) and contains approximately 250 μL of aqueous humour. Aqueous humour is produced at around 2–4 μL/minute. A little passes back into the vitreous; however, the bulk flow of aqueous is from the posterior chamber through the pupil into the anterior chamber. The aqueous humour is drained from the anterior chamber via the conventional and non-conventional routes (see p. 21).

BOX 1-15 CLINICAL CORRELATES

Ageing changes in the eye: presbyopia

This condition may develop around the age of 40–50 years when the elasticity of the lens markedly decreases and there is associated atrophy of the ciliary muscle fibres; consequently, the lens fails to change shape sufficiently during accommodation. This becomes evident as decreasing ability to read, i.e. use near vision.

Aqueous humour has two principal functions. It is the medium by which the necessary metabolites are

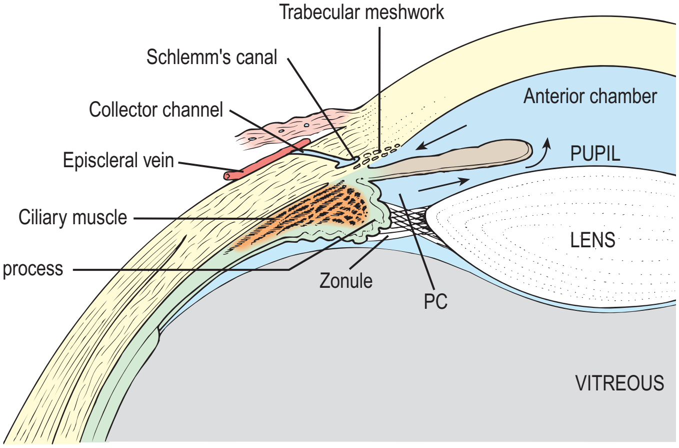
Trabecular meshwork
Schlemm's canal
Anterior chamber
Collector channel
Episcleral vein
PUPIL
Ciliary muscle
LENS
Ciliary process
Zonule
PC
VITREOUS

FIGURE 1-26 Diagram of the anterior segment of the eye. The major pathway followed by aqueous humour from the posterior chamber (PC) to the anterior chamber and outflow pathways is shown by arrows.

transported to the avascular lens and cornea. It also removes toxic metabolic waste products of the cornea and iris. Second, it has a hydromechanical function in maintenance of intraocular pressure. This pressure depends on the balance between the *rate* of production of aqueous humour and its *resistance* to drainage. These two properties of aqueous humour, together with the fact that it is transparent, are necessary for the normal functioning of the eye (see p. 222).

THE VITREOUS

The vitreous cavity (Fig. 1-27) is the largest cavity of the eye (two-thirds the volume of the eye, weight 3.9 g) and contains the vitreous humour or vitreous. It is bound anteriorly by the lens, posterior lens zonules and ciliary body, and posteriorly by the retinal cup. The vitreous is a transparent viscoelastic gel that is more than 98% water, with a refractive index of 1.33. Its viscosity is two to four times that of water. The main constituents of the vitreous, besides water, include hyaluronan (hyaluronic acid), collagens type II and IX, fibronectin, fibrillin and opticin. The gel structure of the vitreous body is dependent on the collagenous constituents and not the hyaluronan. The fine-diameter type II collagen fibres (8–12 nm in diameter) entrap large coiled hyaluronan molecules (see p. 240).

The vitreous is shaped like a sphere with an anterior depression, the *hyaloid fossa* (also known as the patellar or lenticular fossa) (Fig. 1-27A). The vitreous is traditionally regarded as consisting of two portions: a *cortical zone*, characterized by more densely arranged collagen fibrils, and a more liquid *central vitreous*. The vitreous can be further subdivided for descriptive purposes into three major topographical zones, as shown in Figure 1-27A.

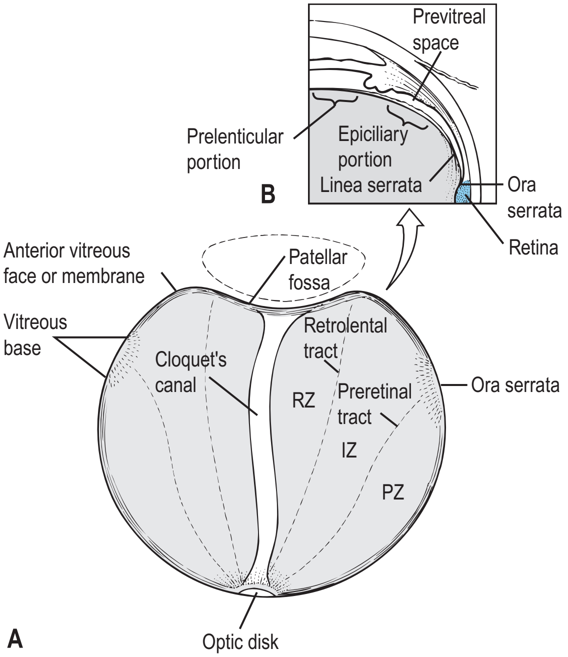
Previtreal space
Prelenticular portion
Epiciliary portion
Linea serrata
B
Ora serrata
Retina
Anterior vitreous face or membrane
Patellar fossa
Vitreous base
Retrolental tract
Cloquet's canal
RZ
Preretinal tract
Ora serrata
IZ
PZ
A
Optic disk

FIGURE 1-27 (**A**) Diagram summarizing the anatomical zones of the vitreous (PZ, preretinal zone; IZ, intermediate zone; RZ, retrolental zone) and the major vitreal condensations (the retrolental tract, preretinal tract, vitreous base, anterior hyaloid face, vitreous cortex). (**B**) Higher power of the anterior vitreous face and its relation to the ciliary body and iris.

The cortical vitreous is attached by condensation of fine collagen fibrils at several points around its margin to (Fig. 1-27A,B):

- the peripheral retina and pars plana via the *vitreous base*, a 3–4 mm wide band
- the posterior lens capsule (ligamentum hyaloide capsulare)
- the retina along the margins of the optic disk (base of the hyaloid canal), although this is disputed (Sebag, 2004)
- the inner limiting membrane of the retina, especially near retinal vessels (most variable and weakest of attachments).

The *central vitreous* possesses less collagen than cortical vitreous. It is traversed by a central fluid-filled canal (hyaloid or Cloquet's canal), which represents the remnants of the course taken by the hyaloid artery that supplied the vitreous and lens in the fetal eye (see Ch. 2). The retrolental and intermediate zones in the human eye are semi-liquid. The existence of an ordered, organized, vitreal structure is still controversial owing to the problems of studying a gel that is 98.5–99.7% water.

The human vitreous begins to degenerate at adolescence, leading to the appearance of liquid-filled cavities and fibrillar strands, such as the retrolental, preretinal and other named tracts of significance only to vitreal specialists. Most of the central tracts (except preretinal) are mobile and change during eye movements. The posterior vitreous cortex may possess zones of reduced density or 'cortical holes' or pockets that, if present, occur close to the fovea, retinal vessels and any developmental anomalies. While these are normal features, secondary pathological holes may develop following various disease processes.

Vitreous cells (Fig. 1-28A,B)

The vitreous is essentially acellular; however, occasional isolated cells may occur in the cortex, particularly near the vitreous base, optic disk and retinal vessels. Cells known as hyalocytes, which have the morphological, ultrastructural, immunophenotypic and functional characteristics of bone marrow-derived macrophages, are the main cell type in the vitreous. A marked vitreal cellular infiltrate is indicative of pathological or inflammatory processes in adjacent tissues, e.g. uveoretinitis.

BOX 1-16 CLINICAL CORRELATES

Posterior vitreous detachment

The thin potential (subhyaloid or sublaminar) space between the surface of the cortical vitreous and the retina may fill with fluid in cases of vitreous detachment. The vitreous may detach relatively easily in the posterior segment where it is less weakly bound to the retina. The fluid may accumulate rapidly in the case of rhegmatogenous vitreous detachment and more slowly where the cortex is not ruptured (arrhegmatogenous vitreous detachment). The former is an age-related change in the vitreous. Vitreous detachment may predispose to retinal detachment (see Ch. 9).

RETINA AND RETINAL PIGMENT EPITHELIUM (Fig. 1-29)

The retina is the innermost of the three coats of the eye. This layer is in the focal plane of the eye's optical system and is responsible for converting relevant information from the image of the external environment into neural impulses that are transmitted to the brain for decoding and analysis. It consists of two primary layers: an *inner neurosensory retina* and an outer simple epithelium, the *retinal pigment epithelium* (RPE). These two layers can be traced embryologically to the inner and outer layers of the invaginated optic cup (see Ch. 2). In the adult they are continuous anteriorly with the epithelial layers over the ciliary processes and posterior iris surface (see p. 28). Between the neural retina and RPE is a potential space, the *subretinal space*, across which the two layers must adhere. The neural retina is firmly attached only at its anterior termination, the ora serrata, and at the margins of the optic nerve head.

The retina is bound externally by Bruch's membrane and on its internal aspect by the vitreous (Fig. 1-26). It is continuous with the optic nerve posteriorly, the site of exit of ganglion cell axons from the eye.

Regions of the retina

Before considering the histological structure of the individual retinal layers (Fig. 1-29) and their constituent cell types, it is important that the reader appreciates the regional or topographical variations of the human retina (Fig. 1-30A,B).

There is often confusion regarding the terminology of the regions of the retina owing to the use of differing

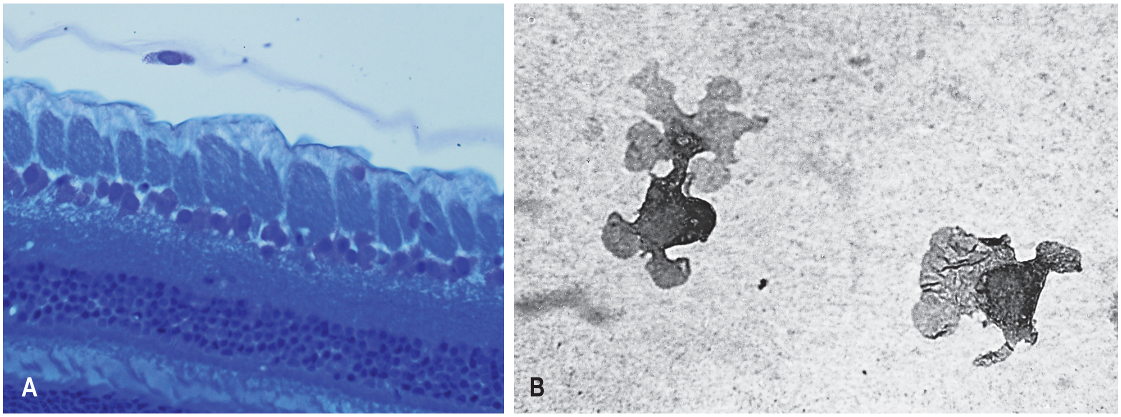

FIGURE 1-28 (**A**) Section of the inner retina in primate eye showing a hyalocyte. (**B**) CD169$^+$ hyalocytes in the rat subhyaloid space (plan view, retinal whole mount).

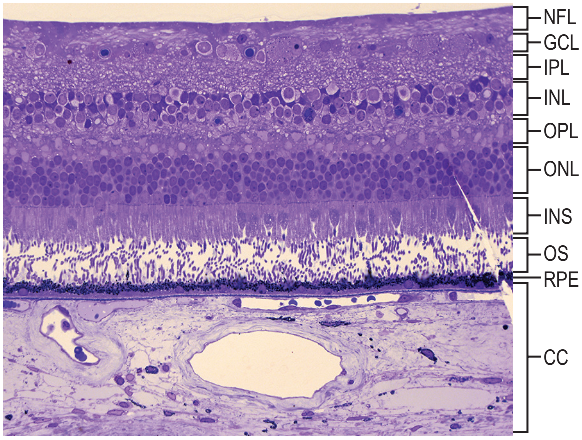

FIGURE 1-29 Low-power micrograph of the human retina (resin histology): arrows, retinal vessels. Original magnification: ×150. Abbreviations: see Fig. 1-32A.

terms by clinicians and anatomists. Figure 1-30A gives a summary of both terminologies.

1. The *posterior pole* or *central retina* (anatomically, area centralis) is a 5–6 mm diameter circular zone of retina situated between the superior and inferior temporal arteries. This region is cone-dominated and is characterized histologically by the presence of more than a single layer of ganglion cell bodies.
2. The *macula lutea* (anatomically, fovea) is a 1.5 mm diameter area in the posterior pole, 3 mm lateral to the optic disk. It is partly yellow as a result of yellow screening xanthophyll carotenoid pigments (zeaxanthin and lutein) in the cone axons. This may serve to act as a short wavelength filter protecting against UV irradiation.
3. The *fovea centralis* (anatomically, foveola) is a central 0.35 mm wide zone in the macula, consisting of a depression surrounded by slightly thickened margins. Cone photoreceptors are concentrated here at maximum density to the exclusion of rods. The inner retinal layers in the margins of the pit (clivus) are displaced laterally (Fig. 1-30C). The foveal retina is avascular and relies on the choriocapillaris for nutritional support.
4. The *optic disk* lies 3 mm medial to the centre of the macula (fovea). There are no normal retinal layers in this zone (blind spot) as ganglion cell axons from the retina pierce the sclera to enter the optic nerve. This pale pink/whitish area is 1.8 mm in diameter with a slightly raised rim. The central retinal vessels emerge at the centre of the optic disk, pass over the rim, and radiate out to supply the retina (Figs 1-30A and, below, 1-34A). The vein usually lies lateral to the artery.
5. The *peripheral retina* is the remainder of the retina outside the posterior pole. The distance from the optic disk to the ora serrata is 23–24 mm on the temporal aspect and approximately 18.5 mm on the nasal aspect. The peripheral retina is 110–140 μm in thickness, rich in rods, and possesses only one layer of ganglion cell bodies.

BOX 1-17 CLINICAL CORRELATES

Retinal detachment

In this condition the neural retina separates from the retinal pigment epithelium (RPE), thus reopening the embryonic intraretinal space or optic ventricle (analogous to the ventricles of the brain), known in the adult as the *subretinal space.* Proteinaceous exudate tends to accumulate in the newly formed space (see Ch. 9). Adhesion of the neural layer and RPE is normally maintained by negative pressure, viscous proteoglycans in the subretinal space and electrostatic forces.

6. The *ora serrata* is the scalloped or dentate anterior margin of the sensory retina. At this transition zone, the neuroretina is continuous with the columnar non-pigmented epithelial cells of the pars plana. The ora serrata is around 1 mm closer to the limbus on the nasal than on the temporal side.

For descriptive purposes, the retina is divided into nasal and temporal halves by a vertical line through the fovea. The optic nerve head is often used as a central point to describe the retina as having supero- and inferonasal and supero- and inferotemporal quadrants. The area of the retina is approximately 1250 mm^2 and varies in thickness from 100 μm (periphery) to 230 μm (near the optic nerve head) (Fig. 1-30B).

Retinal pigment epithelium (Fig. 1-31)

The retinal pigment epithelium (RPE) is a continuous monolayer of cuboidal/columnar epithelial cells, which extends from the margins of the optic nerve head to the ora serrata, where it is continuous with the pigment epithelium of the pars plana. This cell layer has many physical, optical, metabolic/biochemical and transport functions, which play a critical role in the normal visual process (Fig. 1-31A). These include: maintaining adhesion of the neurosensory retina; providing a selectively permeable barrier between the choroid and neurosensory retina; phagocytosis of rod, and to a lesser extent cone, outer segments; synthesis of interphotoreceptor matrix; absorption of light and reduction of light scatter within the eye, hence improving image resolution; and transport plus storage of metabolites and vitamins (especially vitamin A). The complex morphology of this neuroectoderm-derived epithelium reflects these multiple functions.

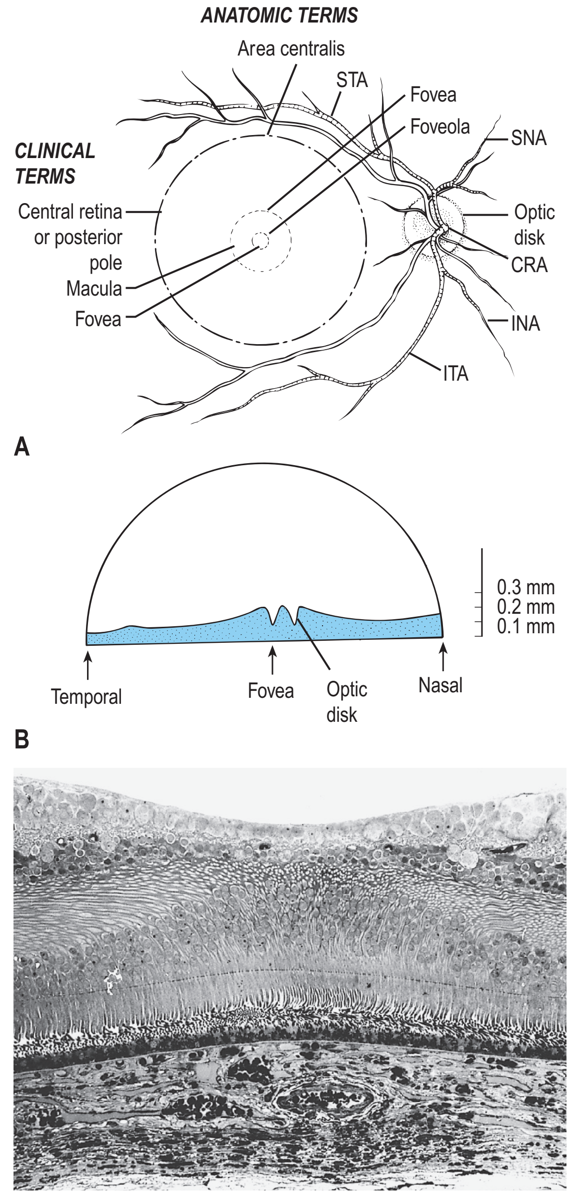

FIGURE 1-30 (**A**) Anatomical and clinical terminology used to describe the regions of the retina: STA, superior temporal artery; ITA, inferior temporal artery; SNA, superior nasal artery; INA, inferior nasal artery; CRA, central retinal artery. (**B**) Regional variations in retinal thickness. (**C**) Section of the retina at the fovea. Original magnification: × 150. (Courtesy of D. Aitken and W.R. Lee.)

Cell size, shape and structure (Fig. 1-31). The RPE cells vary in size and shape depending on age and location, being more columnar in the central retina (14 μm tall, 10 μm wide) and more flattened (10–14 μm tall, 60 μm wide) in the peripheral retina. The basal aspect of the cells lies on *Bruch's membrane* and their apical surface is intimately associated with the photoreceptor outer segments (Fig. 1-31A,B). When examined *en face* they form a highly organized hexagonal pattern of homogeneously sized cells (Fig. 1-31C). The number of RPE cells per eye varies from 4.2 million to 6.1 million. The correlation between structure and function of this monolayer is summarized diagrammatically in Figure 1-31A.

Neurosensory retina (Figs 1-29 and 1-32)

The neurosensory retina is a thin transparent layer of neural tissue that in life has a red/purple tinge due to the presence of visual pigments; however, after death and in fixed specimens, it is white or opaque and often detached from the underlying RPE. Light stimuli are converted into neural impulses in the retina. These impulses are then partially integrated locally before transmission to the brain via the ganglion cell axons in the optic nerve. An appreciation of the anatomy of the neurosensory retina is crucial to an understanding of the physiology of vision (see Ch. 5).

The retina consists of several cell types among which neural cells predominate; other cell types include glial cells, vascular endothelium, pericytes and microglia. The three principal neurone cell types that relay impulses generated by light are photoreceptors, bipolar cells and ganglion cells, and their activity is modulated by other cell types such as horizontal cells, amacrine cells, and possibly by non-neuronal elements. It is the culmination of this neural processing concerning the visual image that is eventually transmitted to the brain along the optic nerve. Retinal cells are arranged in a highly organized manner, and in histological sections appear as eight distinct layers that include three layers of nerve cell bodies and two layers of synapses. The arrangement as seen in conventional histological section is shown in Figure 1-29 and the ultrastructural appearance is shown in Figure 1-32A. A simplified schematic diagram of retinal circuitry is shown in Figure 1-32B. Each of the principal cell types is described briefly below.

BOX 1-18 AGEING CHANGES IN THE RETINAL PIGMENT EPITHELIUM

The retinal pigment epithelium (RPE) has low regenerative capacity in the normal human eye; therefore, cell loss is accommodated by hyperplasia of adjacent cells. Thus, in older eyes, the regular hexagonal array is lost and a heterogeneous mixture of sizes and shapes is more evident. Also evident with age is an increase in lipofuscin within the RPE which displays autofluorescent properties.

Photoreceptors (Fig. 1-33)

There are two types of photoreceptor in the human eye: rods and cones. They are situated on the outer or 'sclerad' aspect of the retina. There are approximately 115 million rods and 6.5 million cones in the human eye. Rods are responsible for sensing contrast, brightness and motion, while cones subserve fine resolution, spatial resolution and colour vision (see Ch. 5). The density of rods and cones varies in different regions of the retina, the periphery being rod-dominated (30 000/mm^2) while cone density increases nearer the macula (150 000/mm^2 at the fovea), the fovea being exclusively cones.

Each photoreceptor consists of a long narrow cell with an inner and outer segment joined by a connecting stalk consisting of a modified cilium (Fig. 1-33). These inner and outer segments are 'separated' from the cell body by the outer limiting membrane. The nucleus is situated in the outer nuclear layer of the retina and axons pass into the outer plexiform layer where they form synaptic terminals (cone pedicle or rod spherule) with bipolar cells and interneurones (horizontal cells) (Fig. 1-32B). The outer segments of rods and cones are shaped precisely as their name implies. They contain the visual pigments that are responsible for absorption of light and initiation of the neuroelectrical impulse.

Rod cells. Rods are long (100–120 μm) slender cells whose outer segment contains the visual pigment rhodopsin sensitive to blue-green light (maximal spectral sensitivity 496 nm). Rods are highly sensitive photoreceptors and are used for vision in dark-dim conditions. The rhodopsin is contained within the membrane-bound lamellae or disks (up to 1000 per cell, 10–15 nm thick) that are enclosed by a single cell

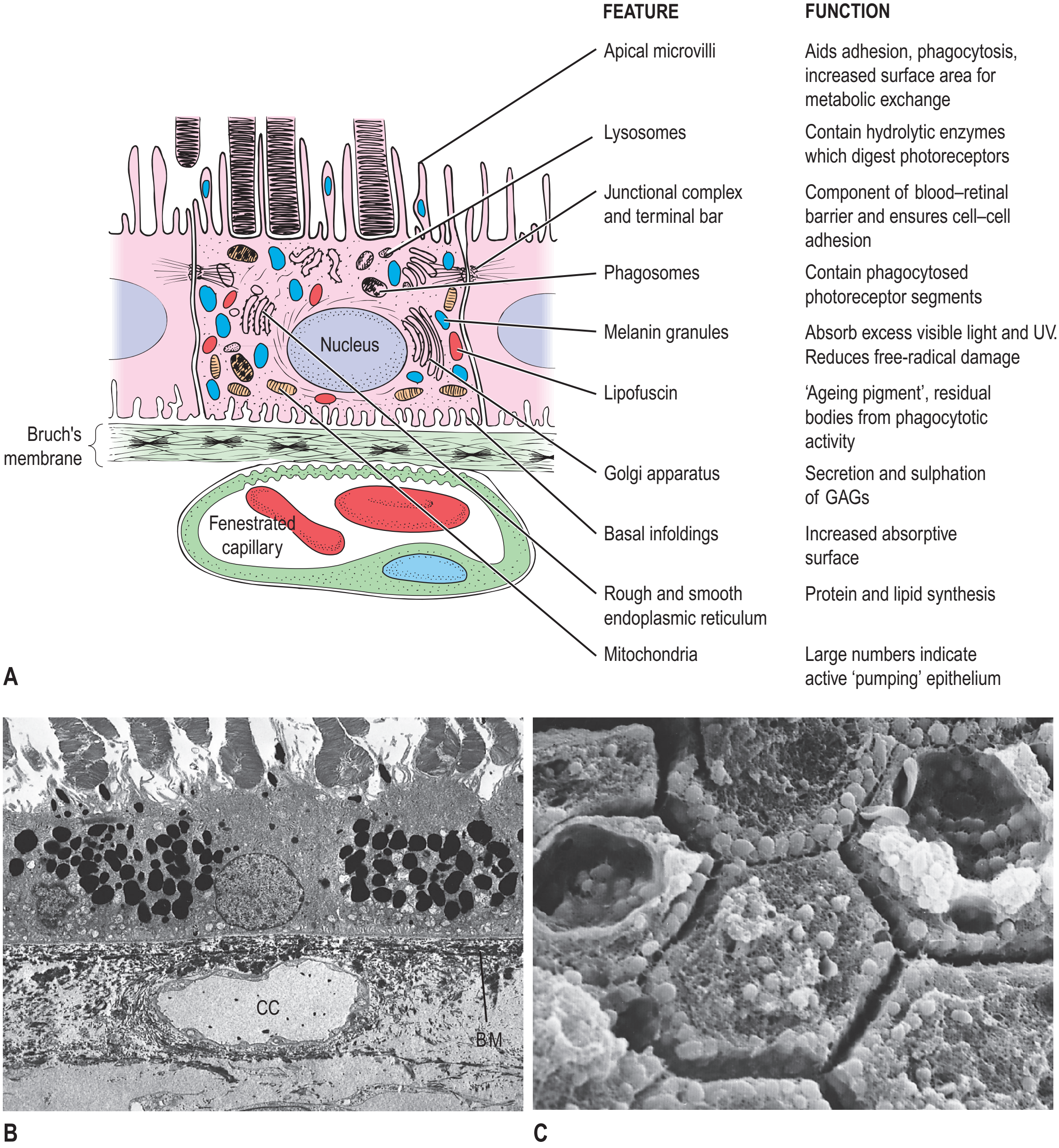

FEATURE	FUNCTION
Apical microvilli	Aids adhesion, phagocytosis, increased surface area for metabolic exchange
Lysosomes	Contain hydrolytic enzymes which digest photoreceptors
Junctional complex and terminal bar	Component of blood–retinal barrier and ensures cell–cell adhesion
Phagosomes	Contain phagocytosed photoreceptor segments
Melanin granules	Absorb excess visible light and UV. Reduces free-radical damage
Lipofuscin	'Ageing pigment', residual bodies from phagocytotic activity
Golgi apparatus	Secretion and sulphation of GAGs
Basal infoldings	Increased absorptive surface
Rough and smooth endoplasmic reticulum	Protein and lipid synthesis
Mitochondria	Large numbers indicate active 'pumping' epithelium

FIGURE 1-31 (**A**) Diagram summarizing the main ultrastructural features of the retinal pigment epithelium (RPE). (**B**) Transmission electron micrograph of human RPE layer: CC, choriocapillaris; BM, Bruch's membrane. (**C**) Scanning electron micrograph of the apical surface of the retinal pigment epithelium. Note the hexagonal shape and the ovoid melanin granules, only visible because of post-mortem-induced disruption of the apical cell membrane. Original magnifications: **B**, × 2600; **C**, × 3600. (Parts B and C courtesy of D. Aitken and W.R. Lee.)

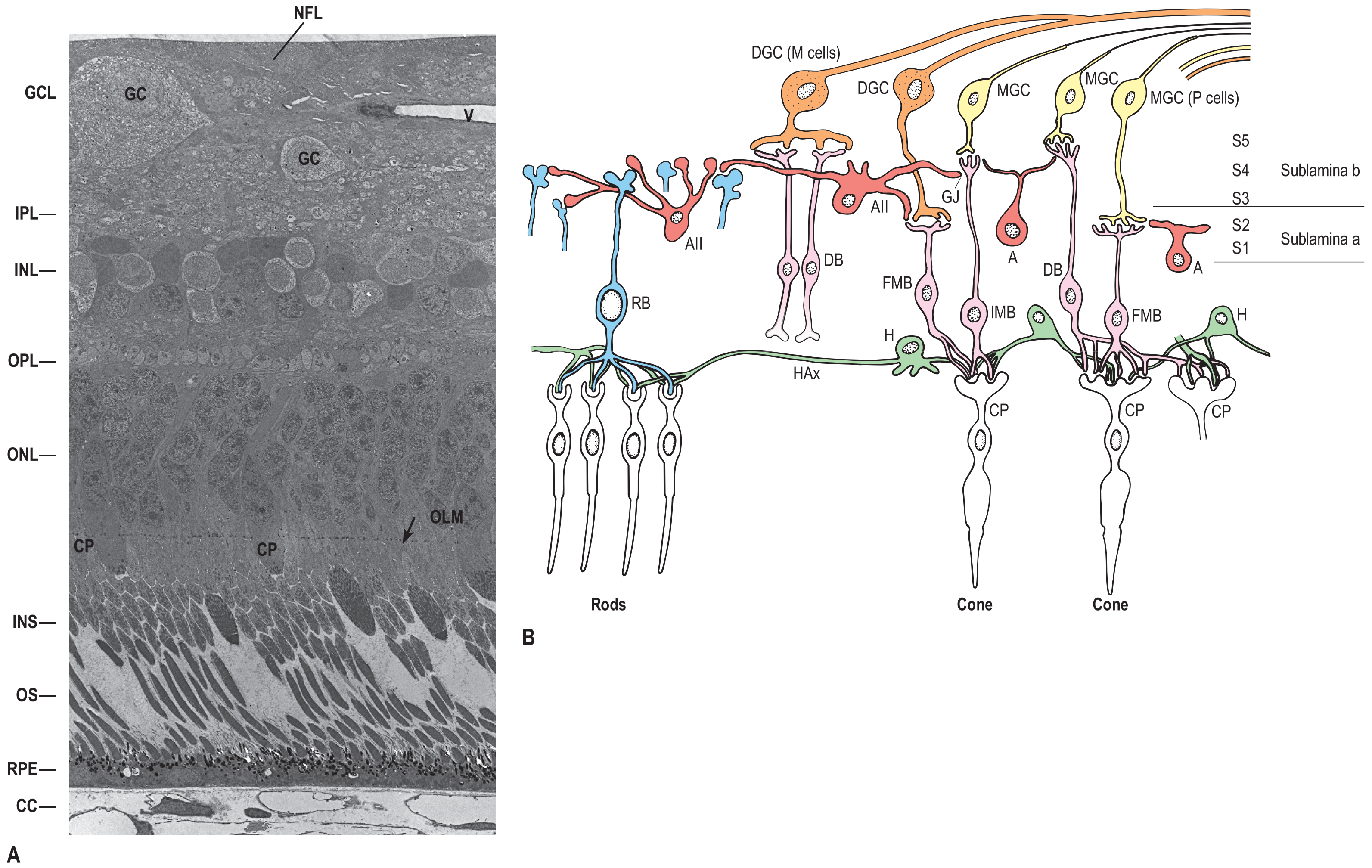
NFL
GCL
GC
GC
V
IPL—
INL—
OPL—
ONL—
OLM
CP
CP
INS—
OS—
RPE—
CC—
A
DGC (M cells)
DGC
MGC
MGC
MGC (P cells)
S5
S4
Sublamina b
S3
S2
S1
Sublamina a
AII
AII
GJ
A
A
DB
DB
RB
FMB
IMB
FMB
H
H
HAx
CP
CP
CP
Rods
Cone
Cone
B

FIGURE 1-32 (**A**) Low-power transmission electron micrograph of the primate retina demonstrating the layered arrangement: CC, choriocapillaris; RPE, retinal pigment epithelium; OS, outer segment; INS, inner segments; ONL, outer nuclear layer; OPL, outer plexiform layer; INL, inner nuclear layer; IPL inner plexiform layer; GCL, ganglion cell layer; GC, ganglion cell; NFL, nerve fibre layer; V, retinal vessel; OLM, outer limiting membrane; CP, cone pedicle. Original magnification: × 930. (**B**) Schematic diagram showing the arrangement and relations of the major cell types in the retina: RB, rod bipolar cell; CP, cone pedicle; H, horizontal cell; HAx, horizontal cell axon; A, cone amacrine cell; AII, AII (rod) amacrine cell; IMB, invaginating midget bipolar cell; FMB, flat midget bipolar cell; DB, diffuse bipolar cell; MGC midget ganglion cell. DGC, diffuse ganglion cell; GJ, gap junction.

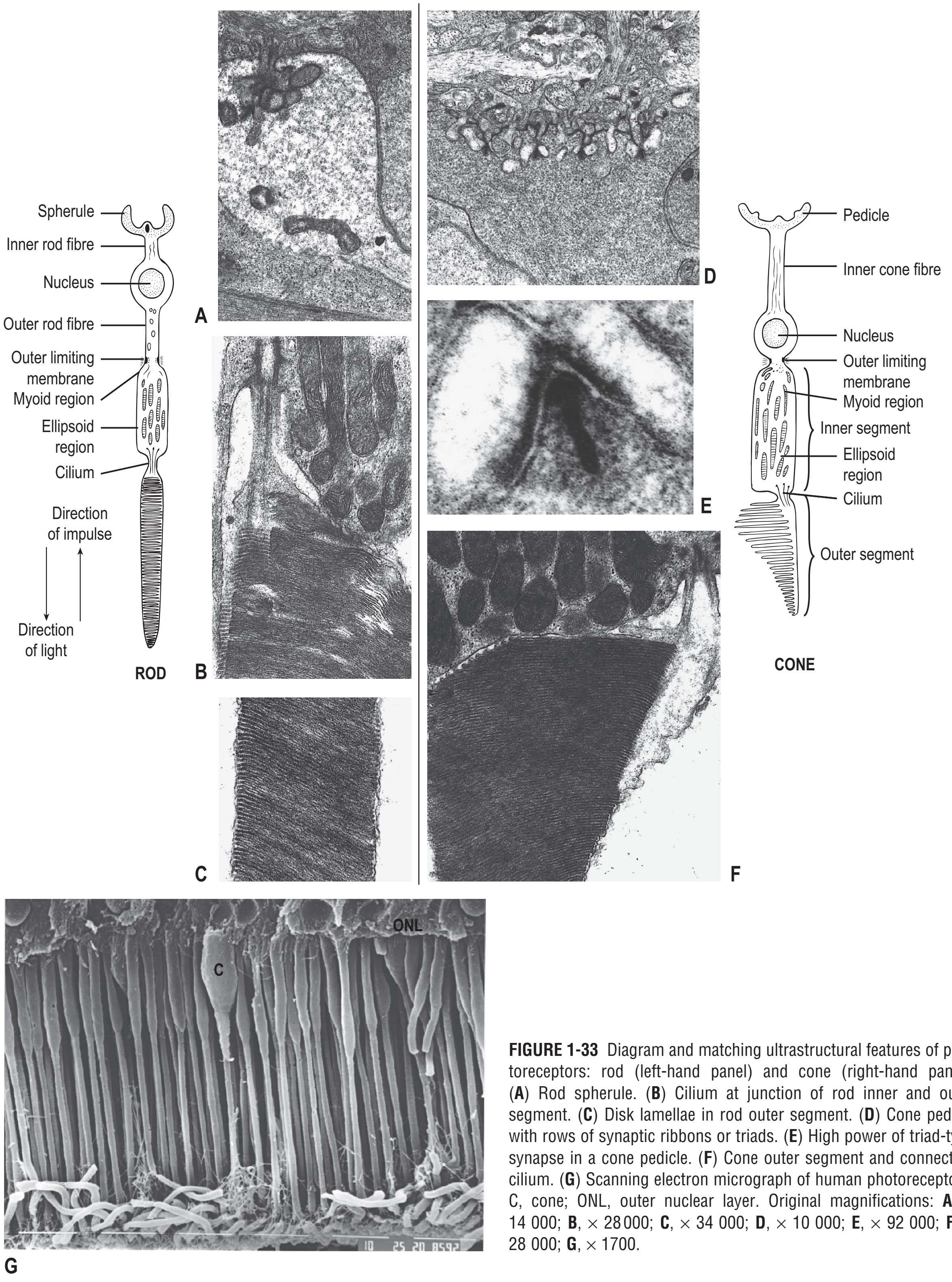
Spherule
Inner rod fibre
Nucleus
Outer rod fibre
Outer limiting membrane
Myoid region
Ellipsoid region
Cilium
Direction of impulse
Direction of light
ROD
A
B
C
D
E
F
Pedicle
Inner cone fibre
Nucleus
Outer limiting membrane
Myoid region
Inner segment
Ellipsoid region
Cilium
Outer segment
CONE
ONL
C
G

FIGURE 1-33 Diagram and matching ultrastructural features of photoreceptors: rod (left-hand panel) and cone (right-hand panel). (**A**) Rod spherule. (**B**) Cilium at junction of rod inner and outer segment. (**C**) Disk lamellae in rod outer segment. (**D**) Cone pedicle with rows of synaptic ribbons or triads. (**E**) High power of triad-type synapse in a cone pedicle. (**F**) Cone outer segment and connecting cilium. (**G**) Scanning electron micrograph of human photoreceptors: C, cone; ONL, outer nuclear layer. Original magnifications: **A**, × 14 000; **B**, × 28 000; **C**, × 34 000; **D**, × 10 000; **E**, × 92 000; **F**, × 28 000; **G**, × 1700.

membrane (a conceptually useful analogy is to liken them to coins stacked inside a stocking). Each outer segment is only 1–1.5 μm in width (Fig. 1-33A–C,G) and 25 μm in length. The disks are produced at the base of the outer segment (the ciliary connection) and over the course of 10 days travel to the tips, which are enclosed by the apical microvilli of the RPE. Here they are phagocytosed by the RPE cells in a circadian manner (predominantly shed in the early morning). The rods are separated by a modified extracellular ground substance known as interphotoreceptor matrix, which contains a 135 kDa glycolipoprotein, interphotoreceptor binding protein (IRBP). The inner half of the inner segment is known as the *myoid*, the outer half the *ellipsoid* (3 μm in length). The ellipsoid is connected to the outer segment by a modified cilium (nine doublet microtubules without a central pair) (Fig. 1-33B) whose basal body is situated in the ellipsoid. The cilium represents the embryological vestige of the ciliated neuroepithelial cells that line the primitive retinal or optic ventricle (see Ch. 2). The cilium acts as a conduit for metabolites and lipids between the inner and outer segments. The remainder of the ellipsoid contains numerous mitochondria, indicative of the high metabolic activity of these cells. The myoid region contains numerous organelles including Golgi apparatus, smooth endoplasmic reticulum, microtubules and glycogen, evidence of a metabolically and synthetically active cell.

Cones (Fig. 1-33D–F). In most diurnal animals two spectrally distinct cone types exist (one maximally sensitive to short wavelengths and one to long wavelengths; known as dichromatic retina); however, in diurnal Old World primates, apes and humans, a third type exists (trichromatic retina). The three types are generally referred to as blue, green and red (or the short, medium and long wavelength) sensitive cones. Cone outer segments are generally shorter than rods and are so called because they are generally conical (6 μm at the base, 1.5 μm at the tip); however, in the fovea they are long, slender and tightly packed (see Fig. 1-30C). The lamellae or disks in cones are not surrounded by a plasma membrane in the same manner as rods but are in free communication with the interphotoreceptor space. Cone disks have a greater lifespan than rods and are not produced in the same manner; in addition, they do not undergo circadian phagocytosis by the RPE cells. They are surrounded by the long villous melanin-containing apical processes of the RPE.

Cones are about 60–75 μm in length. The outer segment is connected to the mitochondria-rich ellipsoid region (containing around 600 mitochondria per cell) of the inner segment by a cilium similar to that described in rods (Fig. 1-33F). The cell body of the cone can be easily identified histologically in the sclerad aspect of the outer nuclear layer because of its large pale-staining nucleus and perinuclear cytoplasm (Figs 1-29 and 1-32A).

The cell bodies of rods and cones are connected by an *inner fibre* to specialized expanded synaptic terminals, known as *spherules* and *pedicles*, respectively. These synapse with bipolar and horizontal cells and contain many highly specialized presynaptic vesicles.

Rod spherules lie more sclerad than the cone pedicles and are deeply indented by bipolar and horizontal cell processes (telodendria). A specialized region known as a *synaptic ribbon* is present between two adjacent nerve fibres. The horizontal cell telodendria penetrate deeply into the spherule; the bipolar cell dendrites (from one to four cells) have a shallower penetration. Up to five processes may be embedded in one spherule (Fig. 1-33A). There is no apparent contact between rod spherules; however, cone pedicles may be connected by gap junctions.

Cone pedicles are broader than rod spherules (7–8 μm) and have a pyramidal shape. In the cone pedicle there are up to 12 indentations, each of which contains three neuronal terminals (*triad*) (Fig. 1-33D,E). The central process in each triad is a midget bipolar cell dendrite (each may have multiple contacts with the same pedicle; Fig. 1-32B). The laterally disposed processes in the triad are horizontal cell processes that may also be involved in several triads on the one pedicle. Thus there may be up to 25 synaptic ribbons in each pedicle (Figs 1-32B and 1-33D,E). Each cone is usually contacted by all the horizontal cells (four to six) in the immediate area or field. Each pedicle also has numerous shallow indentations or synapses with flat diffuse bipolar cells (Fig. 1-32B).

Bipolar cells (Figs 1-32B and 1-34A)

The retina contains approximately 35.7 million bipolar cells, which comprise several functional and morphological subtypes. They are primarily responsible for transmitting signals from photoreceptors to ganglion cells, between which they are interposed. Their cell bodies lie in the *inner nuclear layer* and are oriented in a radial fashion parallel to the photoreceptors. Their single or multiple dendrites pass outwards to synapse principally with photoreceptors (but also with horizontal cells), while their single axon passes inwards and synapses with ganglion and amacrine cells. In the foveal region of the central retina the ratio of cones : bipolar cells : ganglion cells can be 1 : 1 : 1, whereas in the peripheral retina one bipolar cell receives stimuli from up to 50–100 rods. In intervening regions the ratio corresponds to the decreasing visual acuity present in the peripheral retina. This summation of stimuli is a crucial factor in the sensitivity of the rod system to low levels of illumination. Bipolar cells have been subdivided in humans into nine morphological subtypes (Figs 1-32B and 1-34A): one rod bipolar type and eight types of cone bipolars. The latter group can be subdivided into five types of diffuse cone bipolars and three types of midget bipolars.

Rod bipolar cells. *Rod bipolar cells* have a receptive field or dendritic tree, which is small in the central retina (15 μm wide, 10–20 rods) and larger in the peripheral retina (30 μm and 30–50 rods). These represent 20% of all bipolar cells and are most dense around the fovea. In the periphery they contact up to 50 rods and synapse with AII amacrine cells; only rarely do they synapse directly with diffuse ganglion cells.

Diffuse cone bipolar cells. *Diffuse cone bipolar cells* are concerned with converging information from many cones. Their dendrites fan out (up to 70–100 μm) to end in clusters of between five and seven cone pedicles (can be as high as 15–20). The overlap of adjacent cells of this type is extensive in the perifoveal region.

Midget bipolar cells. *Invaginating midget bipolar cells* are the smallest of the bipolar cells whose dendrites penetrate the base of a single cone pedicle (occasionally two) to form a central element in triads. The near 1 : 1 ratio of midget bipolar cells to cones decreases peripherally. Their dendrites may synapse with amacrine cells and midget ganglion cells.

Flat midget bipolar cells connect single cones with single midget ganglion cells or more rarely may connect several cones. They are similar to invaginating midget bipolar cells except their dendrites do not invaginate deeply into the cone pedicle. Potentially, therefore, most cones can be in contact with these two types of midget bipolar cells (flat and invaginating midget) as well as the diffuse type.

Blue cone-specific bipolar cells (blue S-cone) are present in primates and humans and appear to make invaginating contact with only a limited number of cones in their territory, the suggestion being that these are specifically blue cones.

Ganglion cells

The cell bodies of most ganglion cells are located in the innermost nucleated layer of the retina (ganglion cell layer) situated between the nerve fibre layer and the inner plexiform layer (Figs 1-29 and 1-32); however, 'displaced' ganglion cells have been identified in the inner nuclear layer. Ganglion cells are the last neuronal link in the retinal component of the visual pathway (Fig. 1-32B). Their axons form the nerve fibre layer on the innermost surface of the retina and synapse with cells in the lateral geniculate nucleus of the thalamus. The axons form bundles separated and ensheathed by glial cells (Fig. 1-34E). The bundles leave the eye to form the optic nerve. Upon exiting through the *lamina cribrosa*, the axons become myelinated with oligodendrocytes. There are up to seven layers of ganglion cell bodies in the central retina or fovea (ganglion cell layer is 60–80 μm thick) and as few as one cell layer in the peripheral retina (10–20 μm thick). There are approximately 1.2 million ganglion cells per retina; thus, theoretically, there are approximately 100 rods and four to six cones per ganglion cell. While they are functionally diverse, ganglion cells are characterized morphologically by a large cell body, abundant Nissl substance (arrays of rough endoplasmic reticulum) and a large Golgi apparatus (Fig. 1-34E). They are classified into different types on the basis of cell body size, dendritic tree spread, branching pattern and branching level in the five strata of the inner plexiform layer (Fig. 1-32B,C). Some ganglion cells in the macular area may contain yellow (xanthophyll carotenoid) pigment in the cytoplasm, although the cone axons and Müller cells are thought to also

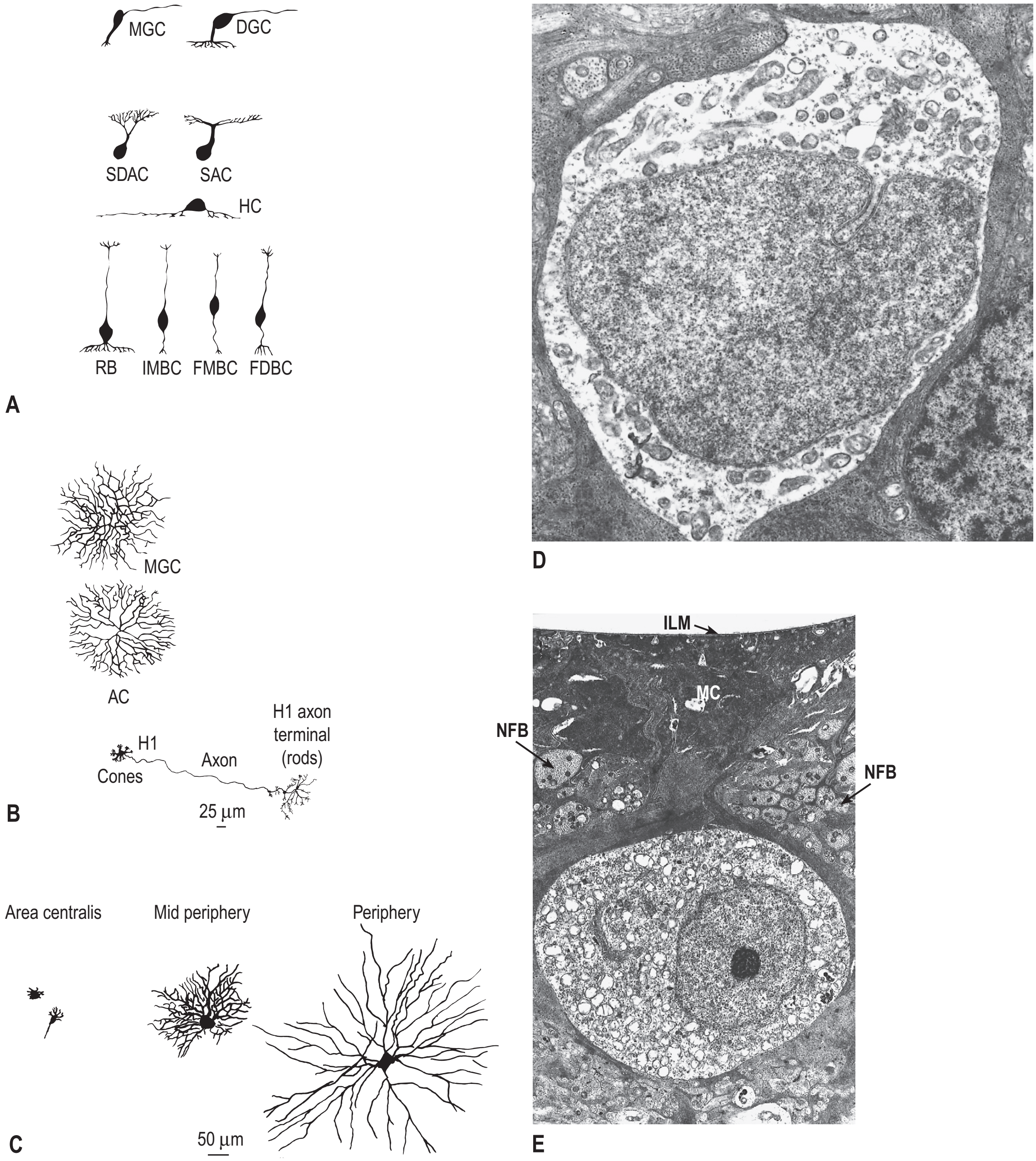
MGC
DGC
SDAC
SAC
HC
RB
IMBC
FMBC
FDBC
A
MGC
AC
H1
Cones
Axon
H1 axon terminal (rods)
25 µm
B
Area centralis
Mid periphery
Periphery
50 µm
C
D
ILM
MC
NFB
NFB
E

FIGURE 1-34 (**A**) Diagram showing the arrangement and location of major neuronal cell types as seen in Golgi-stained preparations in relation to their position in the layers of the retina: DGC, diffuse ganglion cell; FDBC, flat diffuse bipolar cell; FMBC, flat midget bipolar cell: HC, horizontal cell; IMBC, invaginating midget bipolar cell; MGC, midget ganglion cell; RB, rod bipolar; SAC, stratified amacrine cell; SDAC, small diffuse amacrine cell. (**B**) Shape of dendritic fields of midget ganglion cell (MGC), amacrine cell (AC) and H1 horizontal cell (HC) as seen in Golgi preparations or single cell injections in retinal whole mounts. (**C**) Cat ganglion cells increasing in dendritic tree span with increasing eccentricity from the fovea. (Parts B and C, after Kolb et al., 1992, 1994.) (**D**) Electron micrograph of amacrine cell. (**E**) Ganglion cell body and adjacent nerve fibre bundles (NFB): ILM, inner limiting membrane; MC, Müller cell. Original magnifications: **D**, × 7100; **E**, × 4200.

contain these pigments in the macular region. Impulses are received primarily from bipolar cells and amacrine cells via axodendritic and axosomatic synapses, the former occurring predominantly in the inner plexiform layer (Fig. 1-32B) where their dendrites repeatedly branch to form the 'dendritic tree', whose form and size varies considerably and may be correlated with location in the retina and therefore function (receptive field size) (Fig. 1-34B,C). The morphological diversity of ganglion cells (up to 25 types in mammalian and 18 types in human retinas) has prompted classification of these cells into categories, α, β and γ, or X, Y and W types, predominantly based on research in the cat.

Recently a non-rod, non-cone photoreceptive pathway, arising from a population of retinal ganglion cells, was discovered first in nocturnal rodents and then in primates. These ganglion cells express the putative photopigment melanopsin and by signalling gross changes in light intensity serve the subconscious, 'non-image-forming' functions of circadian photoentrainment and pupil constriction. The primate retina, in addition to being intrinsically photosensitive, is strongly activated by rods and cones to signal irradiance over the full dynamic range of human vision. Thus, in the diurnal trichromatic primate, 'non-image-forming' and conventional 'image-forming' retinal pathways are merged, and the melanopsin-based signal might contribute to conscious visual perception.

Midget ganglion cells. These cells synapse exclusively with amacrine cells and one midget bipolar cell (and thus usually one cone) (Fig. 1-32B). Dendritic spread is around 5–10 μm in diameter in the central retina; however, this increases 10-fold in a zone of 2–6 mm eccentricity and attains a maximum of over 100 μm (Fig. 1-34C). Neighbouring midget ganglion cell dendritic fields do not overlap but form mosaics. In humans they are also known as P-cells because they project to the parvocellular layer of the lateral geniculate nucleus (LGN).

Diffuse (parasol) ganglion cells. These comprise a large synaptic field with all types of bipolar cells except midget bipolar cells. They occur in the central retina and their cell bodies (soma) are 8–16 μm in diameter with 30–70 μm dendritic fields, these being smaller nearer the fovea than the periphery. They are also known as M-cells because they project to the magnocellular layer of the LGN.

The finding that midget ganglion cells synapse exclusively with the midget bipolar cells, and that both are common near the fovea, provides the anatomical basis for the observation of small receptive fields and high visual acuity in this region. There are five types of diffuse ganglion cell, classified on the basis of morphology. The anatomical basis of antagonistic fields surrounding receptive fields is complex, although they do not appear to vary much in size from within an 8 mm radius of the fovea. The basis of the antagonist field may be the lateral extensions of the amacrine cell, with its extensive interconnections with ganglion cell dendrites and bipolar cells as well as fellow amacrine cells.

Association neurones (amacrine and horizontal cells) (Figs 1-32B and 1-34A,B)

Horizontal cells. These cells derive their name from the extensive horizontal extensions of their cell processes. There are two distinct morphological varieties in the retina of most species, of which the cat is the most extensively studied: type A is a large sturdy axonless cell with stout dendrites that contact only cones; type B has a smaller bushier dendritic tree that contacts cones exclusively but, in addition, has an axon up to 300 μm in length that ends in extensive arborization that is postsynaptic only to rods (Figs 1-32B and 1-34B). Type A cells have much larger receptive fields than type B. In primates it appears that the two types of horizontal cell, HI (approximates to type B) and HII (approximates to type A), both possess axons. A third type (HIII) has been described in the human

BOX 1-19 CLINICAL CORRELATES

Proliferative vitreal retinopathy

When injured, the retina frequently responds by forming astroglial scars. Indeed, normal age-related degenerative processes in the peripheral retina (microcystoid degeneration) are accompanied by astrocyte proliferation. Disruption of the inner limiting membrane can lead to astrocyte proliferation in the subhyaloid space and in the vitreous itself.

retina. Each rod has connections with at least two horizontal cells and each cone with three or four horizontal cells of each type. In primates the stout dendrites of HI cell soma processes contact around seven cones near the fovea (dendritic tree covering 15 μm); this number increases to as many as 18 further from the fovea (dendritic tree covering 80–100 μm). The axon from HI cells passes laterally and terminates up to 1 mm away in a thickened axon terminal bearing a fan-shaped protrusion of lollipop-like endings in rod spherules (up to 100) (Fig. 1-34B). HII dendritic trees are more spidery and contact about twice as many cones. Their axons are generally shorter (100–200 μm) and contact cone pedicles by small wispy terminals. The manner of their insertion is depicted in Figure 1-32B. Their cell bodies are located primarily in the outer part of the inner nuclear layer. They have few distinctive cytoplasmic organelles except the crystalloids, a series of densely stacked tubules with associated ribosomes. Their processes ramify in the outer plexiform layer close to the cone pedicles. The overlap between horizontal cells is considerable and any one area of retina may be served by up to 20 horizontal cells. Horizontal cells have an integrative role in retinal processing and release inhibitory neurotransmitters, mainly γ-aminobutyric acid (GABA). Recent evidence suggests that there is some colour-specific wiring for the three types of horizontal cells in the human retina.

Amacrine cells (Figs 1-32B and 1-34D). These association neurones were thought to lack axons; however, recent studies have shown that some do indeed possess an axon. They are located in the vitread or inner aspect of the inner nuclear layer (bipolar cell layer) and are distinguishable as a result of their larger size (12 μm) and oval shape. They display a remarkable degree of diversity. There are at least 25 different types in the monkey and human retina. Their cell body is usually flask-shaped and the numerous dendritic processes of these cells ramify and terminate predominantly in the synaptic complexes formed by the bipolar and ganglion cell processes, namely the inner plexiform layer. The shape of their dendritic fields is highly variable and a few examples are shown in Figure 1-34A,B. They can be divided into subtypes on several criteria such as the stratification of their dendrites in the inner plexiform layer or their shape; for example, diffuse, starburst and stratified. Diffuse types can cover narrow fields (approximately 25 μm wide), their fibres being cone-shaped. Other types may spread their axon-like processes several millimetres. They may also be classified on the basis of their neurotransmitters. Amacrine cells may be GABAergic and dopaminergic or can release acetylcholine indicating, together with their morphology, that these cells play a role in modulation (most probably inhibitory) of signals reaching ganglion cells. A subclass of amacrine cells are also thought to be the principal source of the peptide somatostatin, an important neuroactive peptide, in the retina. It may function as a neurotransmitter, neuromodulator or trophic factor.

Retinal neuroglia

Astrocytes. Astrocytes are not the principal or predominant glial cell in the retina. This role is fulfilled by Müller cells, which are analogous to central nervous system oligodendrocytes. Astrocytes are predominantly located in the nerve fibre layer, ganglion cell layer, inner plexiform layer (site of cell bodies), and their outer limit is the vitread aspect of the inner nuclear layer in humans. They form an irregular honeycomb scaffold between vessels and neurones perpendicular to the Müller cells. They may occur as fibrous (elongated) or protoplasmic (rounded) astrocytes. They both contain abundant cytoplasmic structural fibrils (10 nm in diameter) consisting of glial fibrillary acid protein (GFAP) (Fig. 1-35B). Astrocytes are often oriented perpendicular to the direction of the neurone cell bodies or processes, such as in the nerve fibre layer (Fig. 1-35). It seems their role may be to isolate the receptive surfaces of neurones in the retina, thus preventing unwanted signals or effects in neighbouring neurones. They have abundant intracytoplasmic glycogen and form 'gap junctions' with neighbouring astrocytes. Their pedicles or foot processes were once believed to constitute an important functional component of the blood–retinal barrier (see below).

Müller cells. Müller cells (Fig. 1-36A,B) are the principal supporting glial cells of the retina and are considered analogous to central nervous system radial glial or ependymal cells. They have a radial orientation and extend through the depths of the retina from the inner surface, where their expanded 'foot process' lies

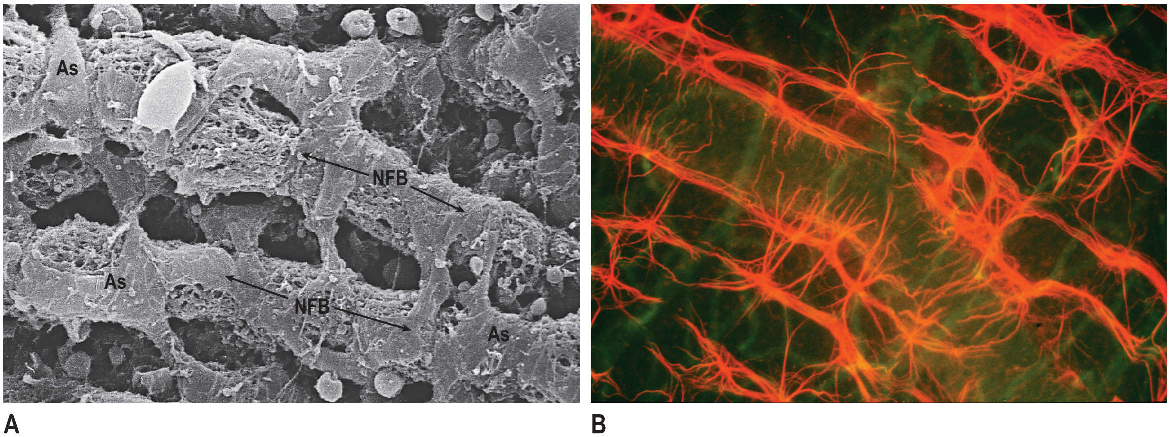

FIGURE 1-35 (**A**) Scanning electron micrograph (viewed from the vitreous aspect) of astrocytes (As) surrounding nerve fibre bundles (NFB) in the inner retina (the inner limiting membrane has been removed to expose the underlying nerve fibre layer). (**B**) Double-colour immunofluorescence illustrating the relations of astrocytes shown with an antibody to glial acidic fibrillary protein (GFAP) (red) and lectin-stained vessels (pale green). Vn, retinal vein; C, capillaries. Original magnifications: **A**, × 1500; **B**, × 150. (Part B courtesy of Dr T. Chan-Ling.)

adjacent to the inner limiting membrane, to their outer limit where they have adherens junctions with photoreceptor inner segments to form the external limiting membrane. They envelop blood vessels, neuronal cell bodies and processes, creating glial 'tunnels' via a series of cytoplasmic processes, as shown in Figure 1-36A,B. Müller cells in humans contain little glycogen, in contrast to species with avascular retinae. Their cytoplasm contains abundant endoplasmic reticulum and microtubules, reflecting their role in protein synthesis, intracellular transport and secretion (Fig. 1-36C). These cells may help to nourish and maintain the outer retina, which lacks a direct blood supply.

While there is extensive coupling between astrocytes and Müller cells, which allows the exchange of tracer molecules, recent studies have demonstrated an absence of significant spread of spatial buffer current between retinal glial cells. Both astrocytes and Müller cells have high K$^+$ membrane conductances, and most spatial buffer current will flow out through these conductances rather than spreading into neighbouring glial cells through gap junctions. In contrast to electrical coupling, chemical coupling between astrocytes is sufficiently strong to mediate propagation of intercellular signals such as the spread of metabolites and ions between glial cells. Coupling between glial cells, therefore, could serve to enhance the transport of key

BOX 1-20 AGEING CHANGES IN MACROPHAGE POPULATIONS IN THE RETINA

With age there is an increasing tendency in animals for microglia to assume a more amoeboid form and activated phenotype and migrate towards the subretinal space. In humans this is evident especially in the peripheral retina close to age-related cystoid retinal degeneration.

FIGURE 1-36 (**A**) Micrograph of a horseradish peroxidase (HRP) filled Müller cell in the rabbit retina. The dark band at the top of the micrograph is composed of Müller cell endfeet and the labelled axons of ganglion cells in the nerve fibre layer. The Müller cells possess side-processes that form different strata in the inner plexiform and they send numerous processes to wrap around the somata of photoreceptors: NFL, nerve fibre layer; GCL, ganglion cell layer; IPL, inner plexiform layer; INL, inner nuclear layer; OPL, outer plexiform layer; ONL, outer nuclear layer; IS, inner segments. (**B**) Diagram of the shape and position of a Müller cell: RPE, retinal pigment epithelium. (**C**) Morphology of a Müller cell (MC) within the outer nuclear layer. Note the intracytoplasmic microfilaments. (**D**) Microglial cells (specialized macrophages) in a retinal whole mount from a transgenic mouse in which eGFP is expressed alongside the locus for the chemokine receptor CX$_3$CR1. All microglia in these animals express CX$_3$CR1 and thus appear fluorescent green in confocal microscopy. (**E**) 3D-rendered retinal microglial network in a retinal wholemount in which retinal vessels have been highlighted (red) by perfusion of a vascular dye. Original magnifications: **A**, bar 10 mm; **B**, × 4400; **C**, × 200. (Part A courtesy of Dr. S. Robinson.)

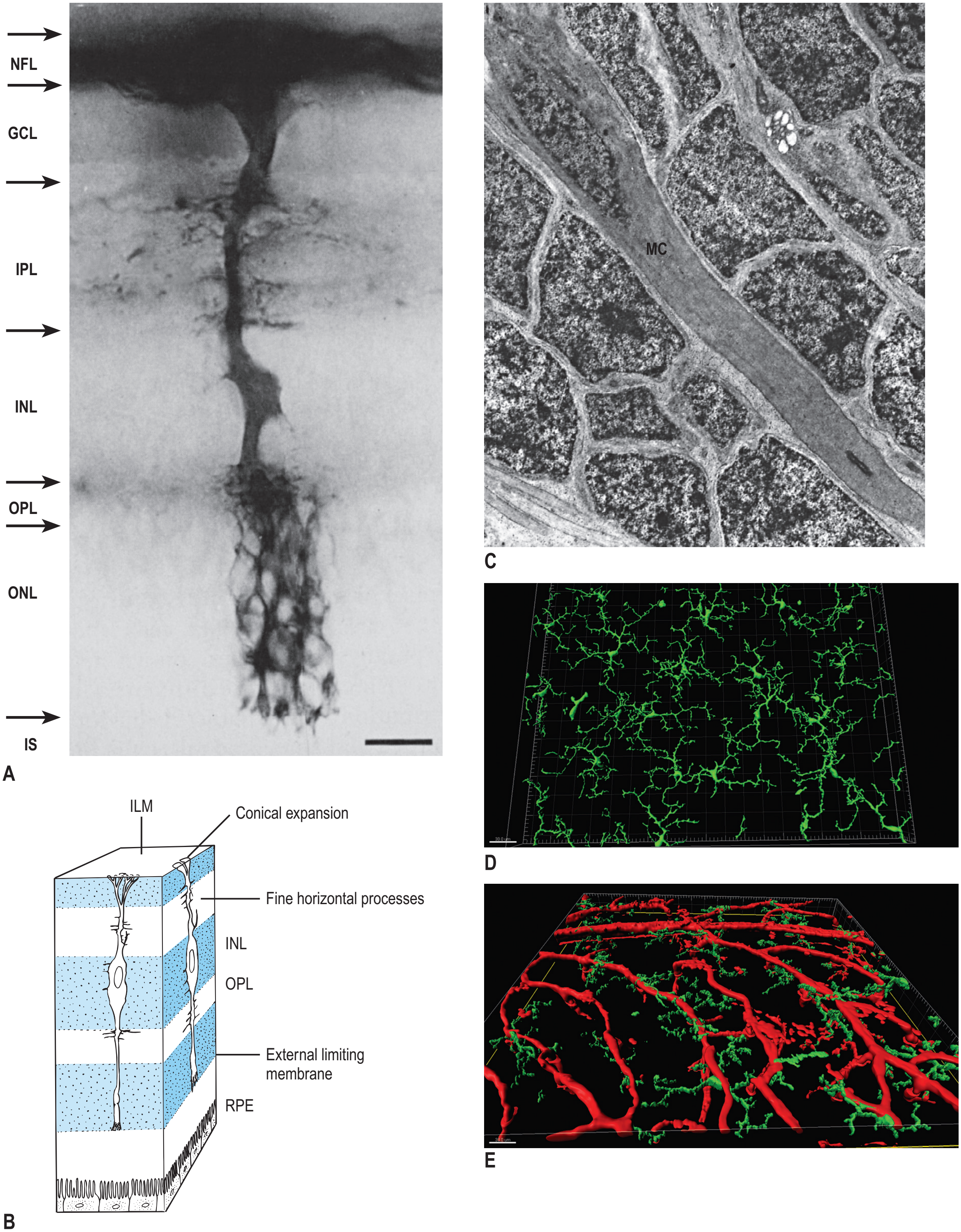
NFL
GCL
IPL
INL
OPL
ONL
IS
A
ILM
Conical expansion
Fine horizontal processes
INL
OPL
External limiting membrane
RPE
B
MC
C
D
E

FIGURE 1-36
For legend see opposite page.

metabolites, such as glutamate, glutamine and lactate, both into and out of glial cells, by allowing them to diffuse between neighbouring cells in the glial syncytium. Clearly the underlying arrangement of both astrocytes and Müller cells reflects this function.

Microglia. Microglia (Fig. 1-36D,E) are a highly specialized subpopulation of the mononuclear phagocyte system that reside in the parenchyma of the central nervous system. These cells most likely arise from yolk sac precursors during early development and may be replenished in adulthood by bone marrow-derived monocytes. They are characterized by an extremely arborized morphology and an immunophenotype of resting macrophages. In the retina, their cell bodies are located largely in three strata, one at the nerve fibre layer–ganglion cell layer interface, one in the inner nuclear layer and another in the outer plexiform layer, although the latter is more obvious in the rodent than human retina. Their processes form a lateral and vertical three-dimensional network within the retina extending only as far as the outer limiting membrane in the normal eye. Less arborized subtypes, sometimes referred to as perivascular macrophages, which closely resemble homologous cells in the parenchyma of the brain, are associated with the perivascular space of retinal capillaries (see Fig. 1-38C), although they are less numerous than brain perivascular macrophages. Retinal microglia share many properties with brain microglia including tissue homeostasis and host defence. Their highly arborized processes are constantly on the move sampling their immediate microenvironment. Upon injury to the retina these cells become activated and assume the role of wandering phagocytes. Activated microglia play a role as immune effectors, via the release of chemokines and cytokines; however, their role as potential antigen-presenting cells (they are predominantly MHC class II^-), or indeed as immunomodulators limiting leucocyte infiltration of the retina, is controversial.

Blood supply of the retina (Figs 1-30A and 1-37)

The retina is an extremely metabolically active sheet of neural tissue with the highest oxygen consumption (per weight) of any human tissue. Like the brain, the retina has a highly selective blood–tissue barrier, which serves primarily to regulate the optimal extracellular environment to facilitate neural transmission. It also regulates the passage of pathogens and intravascular leucocytes, thus partly protecting the neural environment from 'surveillance' by immune cells. In humans, the retina has a dual blood supply (holangiotic), the inner two-thirds being nourished by branches from the central retinal vessels, while the outer one-third is nourished by the choroidal circulation. The choroidal circulation has a high flow rate (150 mm/s), low oxygen exchange and a fenestrated capillary bed; the retinal circulation has a low flow rate (25 mm/s) and high oxygen exchange. The blood–retinal barrier is defined by two sets of characteristics. The first is the structural character of the endothelial and RPE cells located at the endothelial and RPE tight junctions, and the second is the membrane-associated transport characteristics.

Central retinal artery and branches (Figs 1-30A, 1-37 and 1-38). This vessel (0.3 mm in diameter) arises from the ophthalmic artery either in the optic canal or close to the optic foramen where the ophthalmic artery lies bound to the dural covering of the nerve (see p. 59). The central retinal artery then travels forward on the undersurface of the nerve within its dural covering. About 1–1.5 cm behind the eye it pierces the inferomedial aspect of the remainder of meningeal coverings to pass through the subarachnoid and then pierces the nerve. As it passes forward in the centre of the optic nerve, the artery is accompanied by the central retinal vein and a few sympathetic fibres. It resembles other muscular arteries and indeed is affected by conditions such as atheroma and giant cell arteritis. It pierces the papilla centrally, having passed through a constriction or gap in the lamina cribrosa (see Fig. 1-42E). This is a potential site for partial or complete occlusive disease. It branches into superior and inferior branches, which subdivide into nasal and temporal arteries, a pattern best appreciated and investigated clinically by fluorescein angiography. A small vessel, the cilioretinal artery, may be present near the optic nerve head and provide a small anastomotic connection between the choroidal and retinal circulations. The central retinal artery diameter decreases to 100 μm upon emerging from the disk. The large retinal arterial branches travel in the nerve fibre layer beneath the inner limiting membrane.

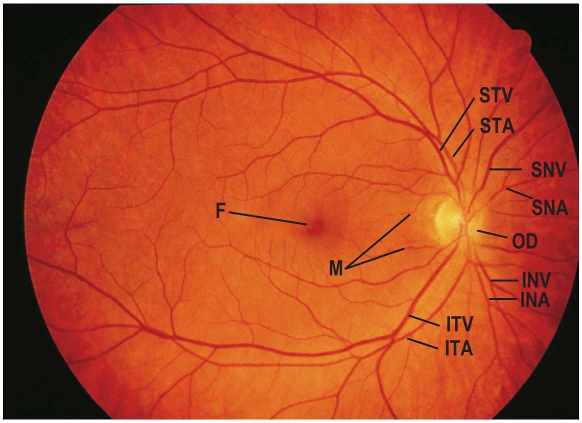

A

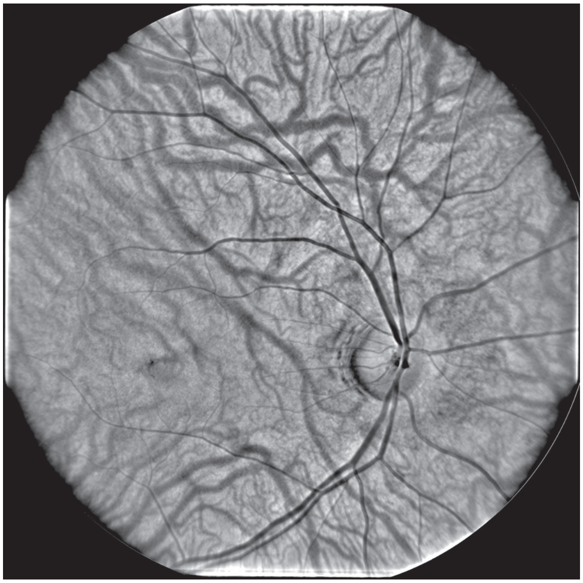

B

FIGURE 1-37 (**A**) Wide-field photograph of the normal human fundus: F, fovea; OD, optic disk; M, macular vessels; STV and STA, superior temporal vein and artery; ITV and ITA, inferior temporal vein and artery; INV and INA, inferior nasal vein and artery; SNV and SNA, superior nasal vein and artery. (**B**), Multispectral digital ophthalmoscopic image of retinal and choroidal circulations. This new technique captures high-resolution image data through the retinal and subretinal layers and hence shows the larger vessels of the choroidal circulation. It both expands the examination wavelength range to include image data from invisible wavelengths of light and also generates the probe wavelengths to separate specific spectral regions for enhanced visibility and discrimination. No intravascular contrast is used in this method of visualizing the fundus. (Part A courtesy of C. Barry; Part B, courtesy of Annedis (Canada).)

These vessels have no internal elastic lamina (which is lost at the optic disk) and are thus not affected in temporal arteritis. They possess a well-developed muscularis, and numerous pericytes lie within the endothelial basal lamina. Each of its four major branches (Figs 1-30A and 1-37) supplies a sector of the retina between which there is no overlap, i.e. they are *functional end-arteries*. The superior and inferior temporal arteries curve above and below the macula and foveal region. Arteries pass over veins and may in some pathological situations cause 'nipping' or narrowing of the veins. There are two main levels of capillary networks, which spread like a vast cobweb throughout the retina (Fig. 1-38B). The inner plexus is situated at the level of the ganglion cell layer and the outer plexus at the level of the inner nuclear layer. The concept of these two laminae is not universally accepted, although patterns of vascular disease support the concept (see Ch. 9). There may be up to four layers of capillaries in the peripapillary zone, and single layers in the perifoveal region and at the ora serrata.

In the human, retinal capillaries pass only as far as the sclerad margin of the inner nuclear layer, the outer retina being normally avascular. Capillaries are most dense in the macula but are absent from the fovea itself (capillary-free zone 500 μm in diameter), which is thus dependent on the choriocapillaris for nutritional support. Larger arterioles are surrounded by a capillary-free zone. Capillary network density decreases towards the peripheral retina.

Retinal capillaries are characterized by complete circumferentially oriented endothelial cells joined by non-leaky tight junctions (zonulae occludentes); however, the high number of endocytotic vesicles suggests that they are more permeable than brain capillaries (Fig. 1-38C,D). They are surrounded by a thick basal lamina, pericytes and astrocyte foot processes (Fig. 1-38C), which are four times more numerous around retinal vessels than brain capillaries and may act as a second front in the blood–retinal barrier and thus compensate for the more permeable nature of retinal vascular endothelium. The numbers of these supportive cells decrease in diabetes, macroglobulinemia and other ischaemic diseases (see p. 503)

The luminal diameter of retinal capillaries (3.5–6 μm) is somewhat smaller than that of conventional capillaries. There is very little extravascular

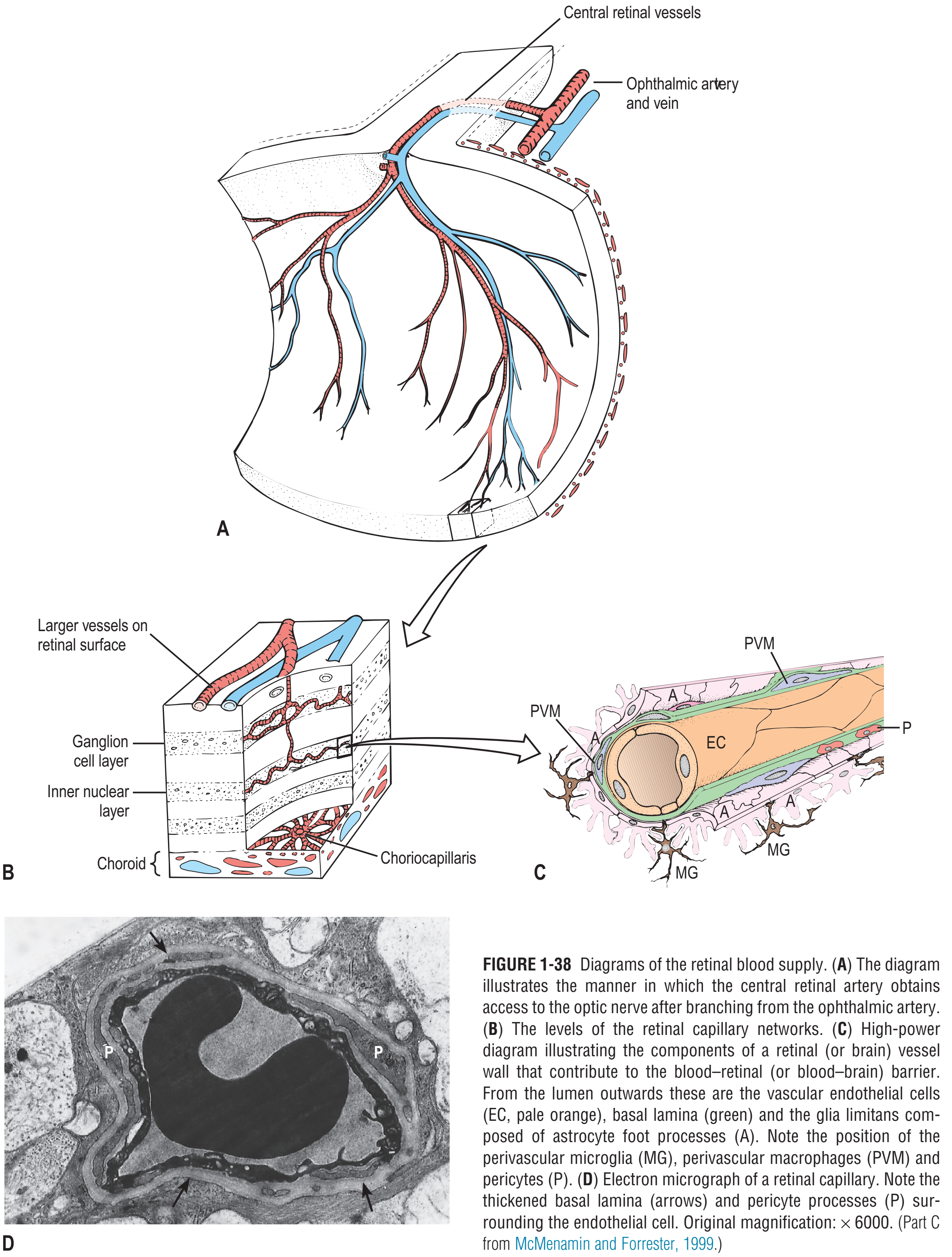
Central retinal vessels
Ophthalmic artery and vein
A
Larger vessels on retinal surface
Ganglion cell layer
Inner nuclear layer
Choroid
Choriocapillaris
B
PVM
PVM
A
A
A
A
EC
P
MG
MG
C
P
P
D

FIGURE 1-38 Diagrams of the retinal blood supply. (**A**) The diagram illustrates the manner in which the central retinal artery obtains access to the optic nerve after branching from the ophthalmic artery. (**B**) The levels of the retinal capillary networks. (**C**) High-power diagram illustrating the components of a retinal (or brain) vessel wall that contribute to the blood–retinal (or blood–brain) barrier. From the lumen outwards these are the vascular endothelial cells (EC, pale orange), basal lamina (green) and the glia limitans composed of astrocyte foot processes (A). Note the position of the perivascular microglia (MG), perivascular macrophages (PVM) and pericytes (P). (**D**) Electron micrograph of a retinal capillary. Note the thickened basal lamina (arrows) and pericyte processes (P) surrounding the endothelial cell. Original magnification: × 6000. (Part C from McMenamin and Forrester, 1999.)

connective tissue around retinal vessels. Mast cells, a common perivascular element in other tissues, including the choroid, are absent in the retina, which has a high threshold of tolerance to histamine. There are *no lymphatic* vessels in the retina.

BOX 1-21 CLINICAL CORRELATES

Besides the capillaries described above, a further lamina fans out over the nerve fibre layer in the peripapillary region. This unique radial capillary network may be more vulnerable to raised intraocular pressure in glaucoma because of the long course of these vessels (over 1000 μm), infrequent arterial input and lack of anastomoses. Flame-shaped haemorrhages (due to hypertension or papilloedema) or cotton-wool spots (in ischaemic disease) occur predominantly in this unusual capillary network.

THE CHOROID

The choroid (Fig. 1-39A–G) is the posterior portion of the middle vascular coat of the eye, the *uveal tract*. It is homologous to the pia–arachnoid of the brain. The choroid is a thin, highly pigmented, vascular, loose connective tissue situated between the sclera and the retina, whose principal function is to nourish the outer layers of the retina. It also acts as a conduit for vessels travelling to other parts of the eye and may also have a thermoregulatory role. Furthermore, absorption of light by choroidal pigment aids vision by preventing unwanted light from reflecting back through the retina as occurs in some nocturnal species that possess a tapetum. The regulation of blood flow in the choroid may also influence intraocular pressure by affecting perfusion rates of the ciliary processes.

The choroid extends from the optic nerve margins to the ciliary body and, although its thickness is probably dependent on blood flow dynamics and has a diurnal variation, it is quoted as being approximately 220 μm at the posterior pole and 100 μm anteriorly. Its inner surface is smooth and forms part of Bruch's membrane beneath the RPE. The outer surface, the suprachoroid, is irregular and firmly attached to the lamina fusca of the sclera. Histologically the human choroid consists of Bruch's membrane, the choriocapillaris, a vascular layer and the suprachoroid (Fig. 1-39C).

BOX 1-22 CLINICAL CORRELATES

Central retinal artery occlusion

This condition is a vivid reminder of the 'functional end-artery' status of the retinal blood supply. In complete central retinal artery occlusion, irreversible changes occur after 1–2 hours and the inner retina becomes white and oedematous except at the fovea, which survives owing to the underlying choroidal circulation, which shows through as a round red patch.

Bruch's membrane (lamina vitrea)

This modified connective tissue layer is 2–4 μm thick and histologically appears as an acellular glassy membrane beneath the RPE (Figs 1-31A,B and 1.39F,G). Bruch's membrane comprises five layers: the RPE basal lamina (0.3 μm thick) (not truly part of the choroid); an inner collagenous zone; a middle elastic layer (incomplete interwoven bands or perforated sheets of elastic 'fibres'); an outer collagenous zone (which blends with the stroma between the choriocapillaris); and the basement membrane of the endothelial cells in the choriocapillaris. Age-related changes in Bruch's membrane lead to areas of diffuse or discrete thickening known as drusen (see Ch. 9, pp. 514).

Choriocapillaris

This is an extraordinarily rich bed of wide-bore fenestrated capillaries that extends only as far anteriorly as the ora serrata and functions to provide nutritional support for the outer retina, especially the photoreceptors. The capillary 'network' is more akin to a perforated vascular 'net' than a network of capillaries (Fig. 1-39B,D,E).

The bore of the capillaries (20–40 μm) and the density of the 'net' (Fig. 1-39D) are greatest near the macula. The capillaries are fenestrated (75–85 nm diameter) on their retinal aspect (Fig. 1-39F,G) and these fenestrae occur at a density of approximately 46 per μm^2. Smooth muscle cells are not usually present in this layer. This sheet or net of capillaries is fed from arterioles, from the layer composed of arterioles and venules (Sattler's) in the manner depicted in Figure 1-39B, i.e. hexagonal patches or 'lobules' of choriocapillaris are fed by a central precapillary arteriole that runs perpendicular to the flat choriocapillaris. This lobular pattern is clinically significant because

BOX 1-23 COMPARATIVE ANATOMY

Tapetum is the cause of 'eye shine'

Many mammalian (carnivores, ruminants, cetaceans, seals) and non-mammalian (fish, crocodiles) species possess a reflective tapetum that may serve to increase photoreception in low light conditions. This may be located in the choroid or the retinal pigment epithelium (RPE). The choroidal tapetum may be cellular or fibrous and may occupy only part, usually the upper portion, of the globe. In carnivores it generally consists of several layers of flattened cells containing reflective material (e.g. guanine, zinc cysteine). In ruminants (e.g. cows, sheep) and cetaceans (dolphins) it is fibrous (tapetum fibrosum) and consists of fine regularly arranged collagen bundles which cause diffractive patterns depending on their orientation. A retinal tapetum generally consists of lipid (e.g. opossum) or guanine (fish, crocodiles) deposits within the RPE.

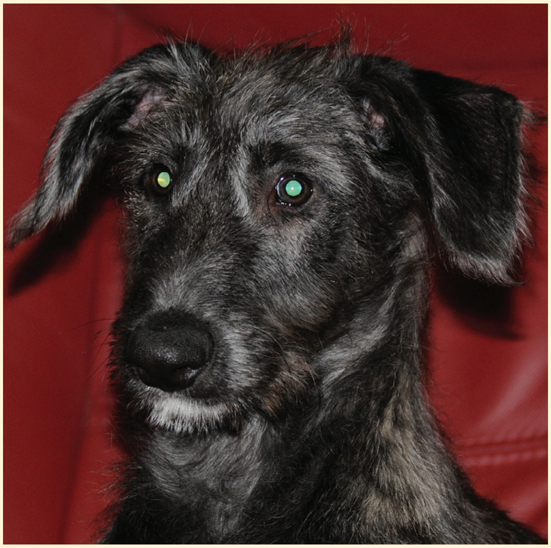

choroidal ischaemia often occurs as pale hexagonal patches (mosaic pattern). The venous channels drain the periphery of these lobules (Fig. 1-39B).

Vascular layer

This layer (Fig. 1-39C) lies beneath the choriocapillaris and can be subdivided into an inner layer of intermediate-sized vessels (arterioles and venules, Sattler's layer) and an outer component (major arteries and veins, Haller's layer). The blood supply of the choroid is chiefly from the long and short posterior ciliary arteries, although recurrent branches from the anterior ciliary arteries anastomose with anterior choroidal vessels (Fig. 1-39A). Venous drainage occurs via a series of large vortex veins (venae vorticosae), of which there are usually four (but may be up to six), each draining a sector of the choroid (Fig 1-38B). These large veins pierce the sclera through emissary canals (Figs 1-17A, and 1-39A,E) and drain into the superior and inferior ophthalmic veins in the orbit.

The choroidal stroma consists of randomly arranged collagen fibres (type I), flattened ribbon-like elastic fibres, fibrocytes and numerous melanocytes (Fig. 1-39C,F). The extent of choroidal pigmentation influences the appearance of the fundus, with the highly pigmented choroid of darker-skinned people showing through more than in the fundus of a person with less skin pigmentation, in which the red–orange reflex is primarily the result of the choroidal vasculature. The choroid, being a connective tissue, contains resident populations of immunocompetent cells including occasional plasma cells and lymphocytes, numerous perivascular mast cells (Fig. 1-40A, B) and networks of resident tissue macrophages and dendritic cells (Fig. 1-41).

FIGURE 1-39 (**A**) Diagram of the uveal tract blood supply. (**B**) Schematic representation of the hexagonal units in the choricapillaris fed by small arterioles. (**C**) Semi-thin resin section of the outer retina and choroid in the primate eye. Note the heavy degree of pigmentation and the layers of the choroid. (**D**) Resin vascular cast of the choriocapillaris viewed from the retinal aspect. (**E**) Resin vascular cast of a quadrant of the choroid viewed from the external aspect showing a large vortex vein (VV). (**F**) Low-power electron micrograph of the retinal pigment epithelium, Bruch's membrane (BM) and the closely related choriocapillaris in a primate eye. (**G**) Higher-power electron micrograph of the basal aspect of the RPE with its basal lamina forming part of Bruch's membrane. Note the fenestrated capillaries (FC) in the endothelial lining of the choriocapillaris. Note the capillary endothelial cells are characterized by fenestrae (FC) on the retinal aspect adjacent to Bruch's membrane. RPE, retinal pigment epithelium; CC, choriocapillaris; HL, Haller's layer; SL, Sattler's layer; M, melanocyte; SPCA, short posterior ciliary arteries; LPCA, long posterior ciliary arteries. Original magnifications: **C**, × 350; **D**, × 35; **E**, × 120; **F**, × 2400; **G**, × 14 000.

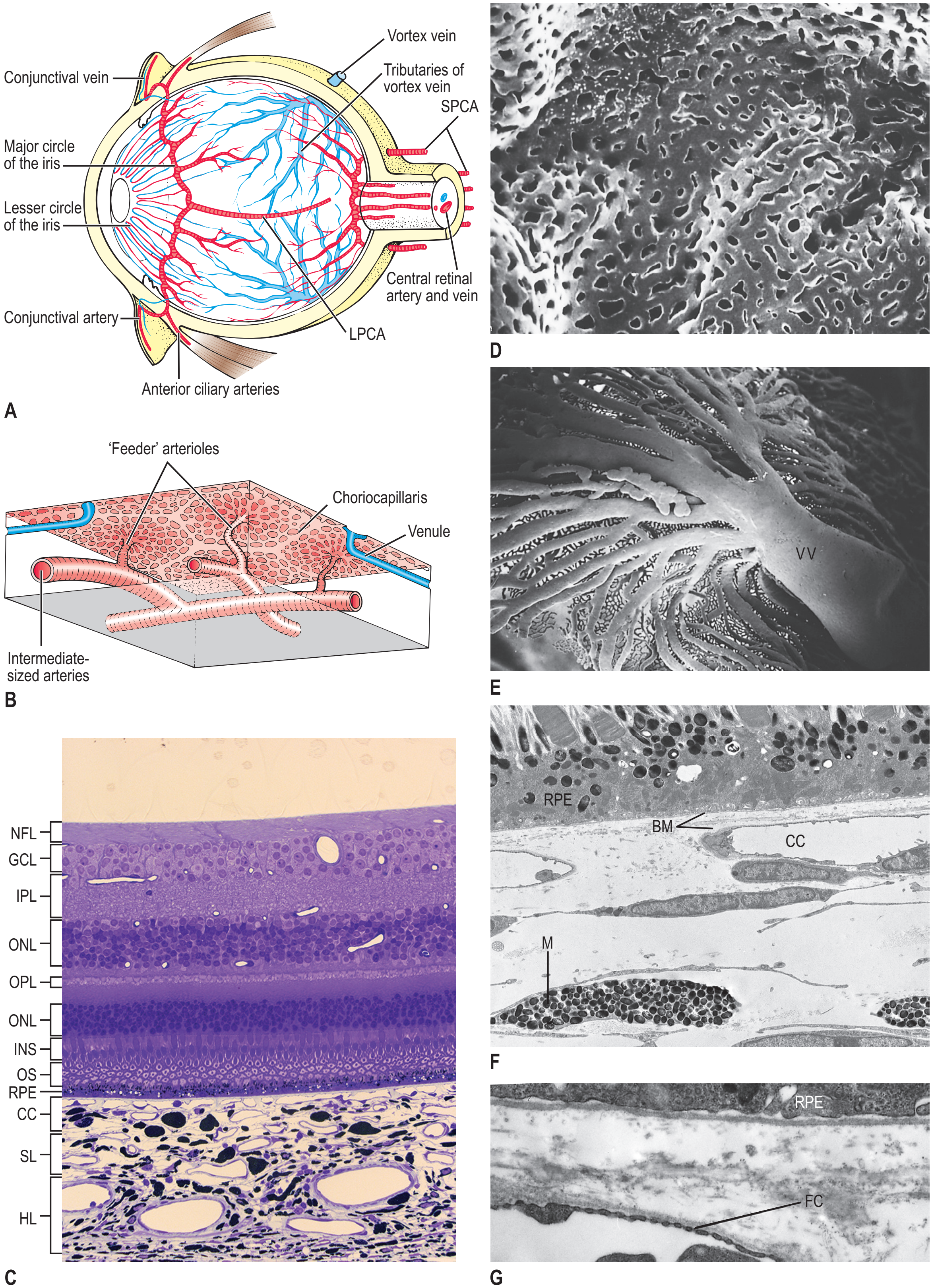
Conjunctival vein
Major circle of the iris
Lesser circle of the iris
Conjunctival artery
Anterior ciliary arteries
LPCA
Vortex vein
Tributaries of vortex vein
SPCA
Central retinal artery and vein
A
'Feeder' arterioles
Choriocapillaris
Venule
Intermediate-sized arteries
B
NFL
GCL
IPL
ONL
OPL
ONL
INS
OS
RPE
CC
SL
HL
C
D
VV
E
RPE
BM
CC
M
F
RPE
FC
G

FIGURE 1-39
For legend see opposite page.

Suprachoroid

This is a 30 μm thick transition zone between the choroid and the sclera. It consists of thin interconnected lamellae of melanocytes, fibroblasts and connective tissue fibres separated by a thin 'potential' (supra- or perichoroidal) space, which in pathological conditions may become separated by fluid and blood. It is frequently artefactually enlarged in histological preparations. It is an avascular layer, the only vessels being those that traverse the suprachoroid entering or leaving the choroid. The lamellae blend with the choroid and the lamina fusca of the sclera. The suprachoroidal space is continuous with the supraciliary space anteriorly. Recent research has unveiled a previously unrecognized, highly organized network of non-vascular smooth muscle cells in the suprachoroid. These networks were particularly evident behind the fovea, around the entry points of the posterior ciliary arteries and nerves and in bundles running parallel to vessels travelling anteriorly from the posterior pole as far as the exit points of the vortex veins. The function of this network of smooth muscle cells in the human choroid remains speculative.

Nerve supply of the choroid

The choroid is innervated by the *long and short ciliary nerves*. The long ciliary nerves (from the nasociliary branch of V_1) pass through the choroid and transmit sensory fibres to the cornea, iris and ciliary body. Sympathetic fibres are also carried in these nerves to the dilator pupillae (see p. 28). The *short ciliary nerves* arise from the ciliary ganglion and carry sensory (from nasociliary), sympathetic and parasympathetic fibres (derived predominantly from nerve III, but also from VII). The latter have already synapsed in the pterygopalatine ganglion. Both long and short ciliary nerves pierce the sclera in the form of a ring 2–3 mm anterior to the optic nerve sheath, along with the long and short posterior ciliary arteries. The nerve terminals

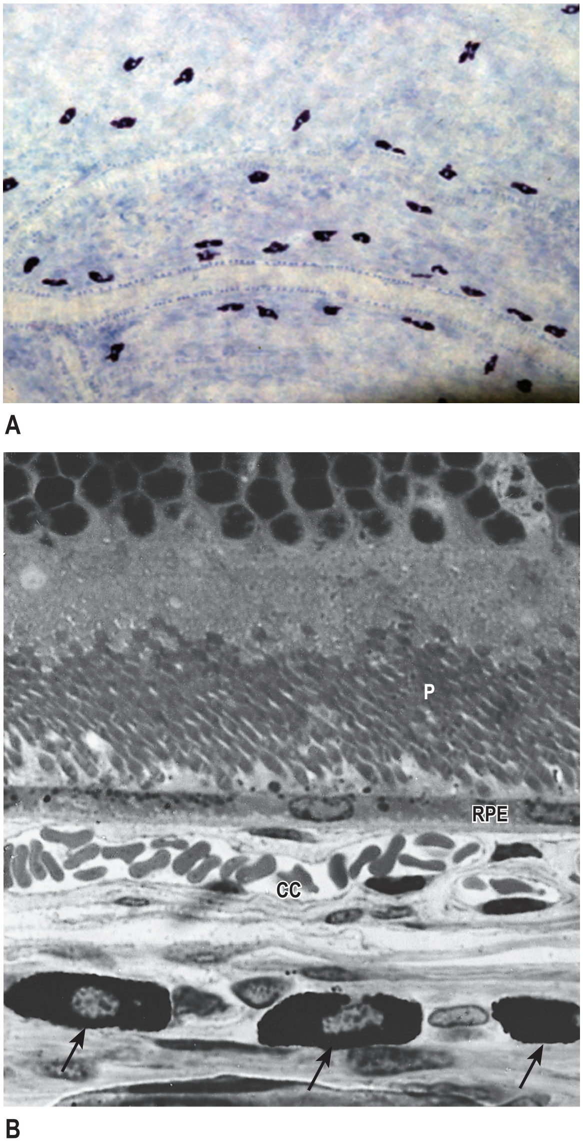

FIGURE 1-40 Mast cells in the choroid. (**A**) Low-power view of rat choroidal whole mount stained with toluidine blue which demonstrates the perivascular arrangement of mast cells: A, artery. (**B**) Semi-thin resin section of rat outer retina and choroid stained with toluidine blue showing three mast cells (arrows): P, photoreceptors; RPE, retinal pigment epithelium; CC, choriocapillaris. Original magnifications: **A**, × 75; **B**, × 900.

BOX 1-24 AGEING CHANGES IN THE EYE

Choroidal immune cells and age-related macular degeneration (AMD)

The deposition of lipid, complement and immunoglobulin G that accompanies the senescent changes in Bruch's membrane may be augmented by impaired macrophage recruitment and/or decreased homeostatic scavenging by the resident macrophages and dendritic cells in the choroid and thus they may contribute to the accumulation of debris and the formation of drusen in 'dry AMD' and to the eventual choroidal neovascularization in 'wet AMD'.

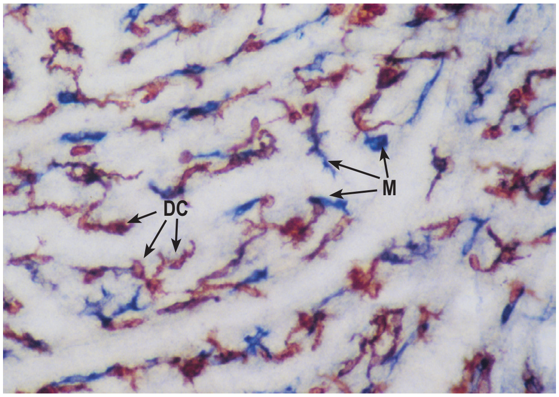

FIGURE 1-41 Double-colour immunohistochemistry of rat choroidal whole mount stained with monoclonal antibodies specific for macrophages (M; blue) and major histocompatibility class II-positive dendritic cells (DC; red). This method of examining stained tissue provides a 'plan view' that clearly demonstrates the distinct networks of both these cell types in the choroid. Original magnification: × 600.

branch extensively and form plexi of unmyelinated fibres in the choroid and suprachoroid adjacent to vascular smooth muscle cells; however, they do not extend into the choriocapillaris. Fibres containing vasoactive intestinal peptide (VIP) and neuropeptide Y have been identified in the choroid and probably act as vasodilator and vasoconstrictor agents, respectively. Multipolar and bipolar ganglion cell bodies immunoreactive for nitric oxide synthase (NOS) and VIP have been recently identified in the choroid and their axons may supply the choroidal vasculature (vasodilatory) or non-vascular smooth muscle cells. Their structure and immunohistochemical characteristics suggest that they may have a mechanosensory role.

OPTIC NERVE

The optic nerve (Fig. 1-42) is unique anatomically as it is the only tract in the central nervous system to leave the cranial cavity. Furthermore, it is subdivided into fascicles by connective tissue and glial septae and is surrounded by cerebrospinal fluid. It is also unique in that it is the only central nervous system tract that can be visualized clinically.

The optic nerve is formed by convergence of ganglion cell axons at the optic disk, the commencement of the nerve. Foveal/macula fibres constitute around 90% of all axons leaving the eye and form the distinct *maculopapillary bundle*. From the disk, the axons extend along the nerve through the orbit to traverse the optic canal in the sphenoid bone.

BOX 1-25 CLINICAL CORRELATES

Choroidal infarctions

These appear on angiograms to take the form of triangular areas near the equator, with the apex pointing towards the optic disk. There is probably less functional anastomosis between choriocapillaris lobules than was once suspected; however, some degree of anastomosis in the subcapillary arterioles exists. Peripheral retinal cobble or paving-stone degeneration represents chronic focal ischaemic changes in the anterior choroid.

The optic nerve can be divided into four main portions: *intraocular* (1 mm in length), *orbital* (25–30 mm), *intracanalicular* (4–10 mm) and *intracranial* (10 mm). The latter portion is discussed on pp. 92 in the context of the visual pathways.

Intraocular portion

The *intraocular portion* (Fig. 1-42A) extends from the surface of the optic disk to the posterior margins of the sclera. The nerve fibres are not myelinated in this portion. It can be further subdivided into three regions: the retinal (pars retinalis), choroidal (pars choroidalis) and scleral (pars scleralis) portions. Myelination commences approximately level with the termination of the subarachnoid space at the posterior limits of the lamina cribrosa. As the fascicles of nerve fibres pass posteriorly from the optic disk into the intraocular portion, the glial cells become more common; columns of glial cell nuclei are especially prominent in the scleral portion, where they account for up to 40% of the tissue mass (Fig. 1-42A). The commencement of the optic nerve, the *optic disk*, varies depending on the method of measurement but is approximately between 1.7 and 2.8 mm in diameter, although variations both within a population and between races are observed and this variation has been linked to susceptibility to glaucoma. The layers of the retina and the choroid terminate at the edge of the disk as specialized regions of glial tissue, the intermediary tissue (of Kuhnt) and marginal border tissue (of Elschnig). The absence of retinal tissue in this region explains the *'blind spot'* phenomenon. As the 1.2 million ganglion cell axons in the nerve fibre layer become crowded towards the disk, they create a raised area or *papilla*, which is thickest on the lateral aspect owing to the large number of fibres in the

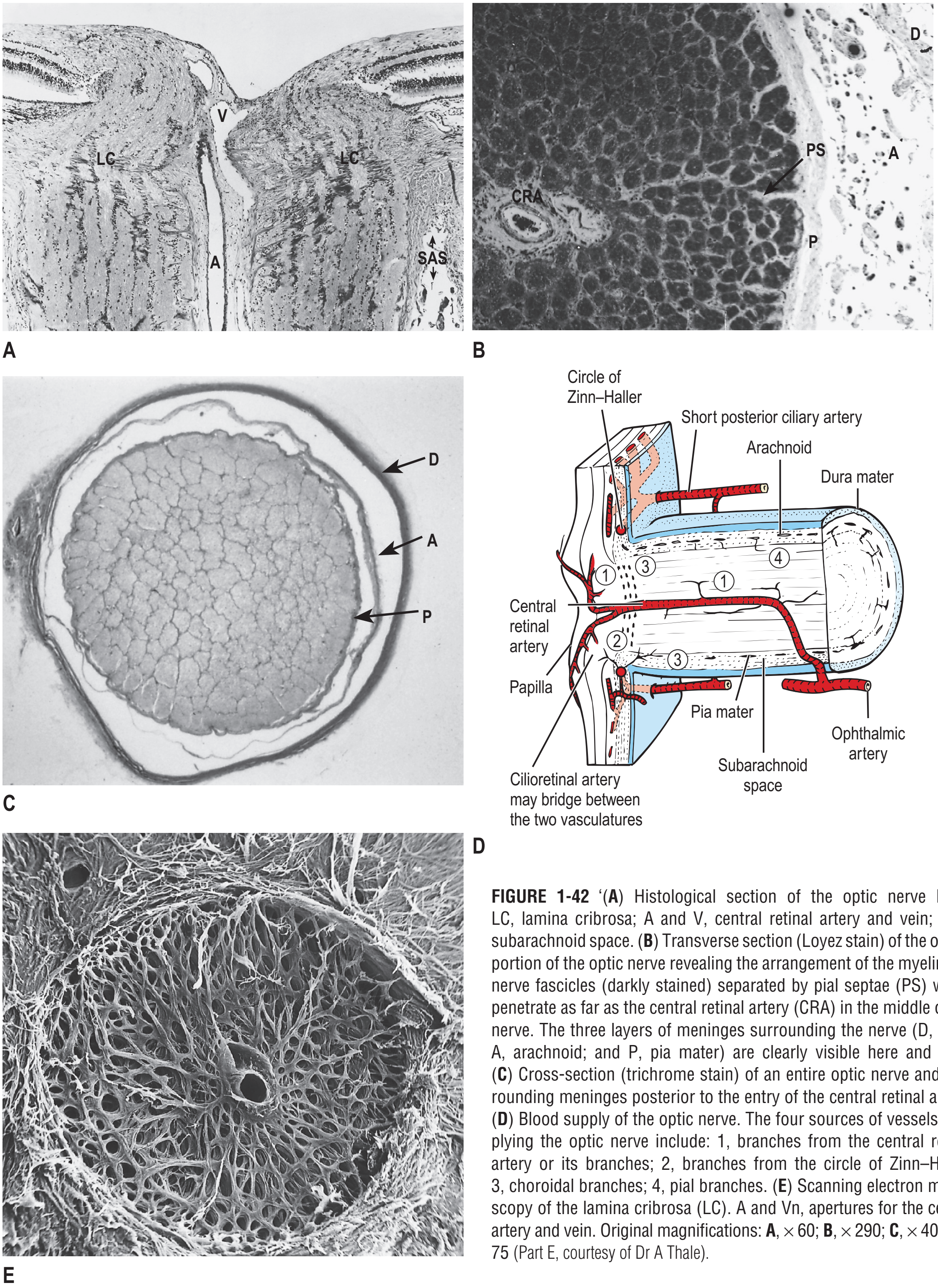
V
LC
LC
A
SAS
A
D
PS
A
CRA
P
B
D
A
P
C
Circle of Zinn–Haller
Short posterior ciliary artery
Arachnoid
Dura mater
①
③
④
①
Central retinal artery
②
③
Papilla
Pia mater
Ophthalmic artery
Subarachnoid space
Cilioretinal artery may bridge between the two vasculatures
D
E

FIGURE 1-42 '(**A**) Histological section of the optic nerve head: LC, lamina cribrosa; A and V, central retinal artery and vein; SAS, subarachnoid space. (**B**) Transverse section (Loyez stain) of the orbital portion of the optic nerve revealing the arrangement of the myelinated nerve fascicles (darkly stained) separated by pial septae (PS) which penetrate as far as the central retinal artery (CRA) in the middle of the nerve. The three layers of meninges surrounding the nerve (D, dura; A, arachnoid; and P, pia mater) are clearly visible here and in C. (**C**) Cross-section (trichrome stain) of an entire optic nerve and surrounding meninges posterior to the entry of the central retinal artery. (**D**) Blood supply of the optic nerve. The four sources of vessels supplying the optic nerve include: 1, branches from the central retinal artery or its branches; 2, branches from the circle of Zinn–Haller; 3, choroidal branches; 4, pial branches. (**E**) Scanning electron microscopy of the lamina cribrosa (LC). A and Vn, apertures for the central artery and vein. Original magnifications: **A**, × 60; **B**, × 290; **C**, × 40. **E**, × 75 (Part E, courtesy of Dr A Thale).

maculopapillary bundle. The raised margin of the optic disk surrounds an indentation, the *physiological cup*. As the fibres pass posteriorly, they pierce the sieve-like connective tissue mesh, the *lamina cribrosa*, which fills the posterior scleral foramen. The lamina cribrosa is formed by irregular collagen fibre bundles continuous with the sclera. These bundles are arranged in the form of circles or a figure of eight (Fig. 1-42E). Elastic tissue from the choroid and Bruch's membrane is continuous with and 'anchored' to the adventitia surrounding the central retinal artery and vein. The collagenous bundles in the lamina cribrosa are separated from the axons by a covering of glial tissue, which may protect the nerve fibres as they pierce the irregular openings. The *scleral canal* is some 0.5 mm long and may vary in shape from cone-like (narrowest portion nearest the disk) to double cone or funnel-like. Posterior to the pars scleralis, the nerve fibres become myelinated by *oligodendrocytes* (Fig. 1-42B), causing a doubling of the thickness of the optic nerve.

Orbital portion

The *orbital portion* of the optic nerve (Fig. 1-42C,D) extends backwards and medially from the back of the eye to the optic canal in the sphenoid at the apex of the orbit. It is covered by three layers of meninges: pia, arachnoid and dura. The dura and arachnoid blend with the sclera, and the subarachnoid space around the nerve terminates at the posterior surface of the sclera in the form of a fluid-filled ring (Fig. 1-42A). The central retinal vessels must cross the subarachnoid space and are therefore vulnerable, particularly the vein, in cases of raised intracranial pressure.

The majority of the axons in the nerve are 1 μm in diameter and approximately 10% are between 2 and 10 μm. The glial septae between fascicles present in the intraocular portions extend into the orbital portion but become less distinct as the orbital apex is approached. The orbital portion of the optic nerve has a slight S-shaped bend, which allows a full range of ocular movement without stretching the nerve. As the optic nerve approaches the orbital apex it is surrounded by the tendinous annulus, which gives origin to the rectus muscles.

Intracanalicular portion

The *intracanalicular portion* of the optic nerve passes through the optic canal (foramen), accompanied by the ophthalmic artery and sympathetic nerves. The dura surrounding the nerve splits at the orbital opening, the majority continuing as the dural sheath of the nerve inside the canal and a thinner portion blending with the periorbita (Fig. 1-43).

BOX 1-26 CLINICAL CORRELATES

Glaucoma

The intraocular portion is that part of the optic nerve damaged in glaucoma. Axonal damage may be a consequence of either interference with blood flow or interruption of axonal transport and raised intraocular pressure is a significant risk factor. No single hypothesis has been proposed that adequately explains why specific regions of the nerve are more likely to be damaged than others, resulting in the characteristic visual field defects or *scotomas*. Recent research has highlighted the differential pressure gradient across the lamina cribrosa between the cerebrospinal fluid pressure and the intraocular pressure (IOP) (the translaminar pressure gradient (TLPG)) and how it may influence central retinal venous pressure as it traverses the lamina cribrosa, including the 'arteriolization' of the vessel wall and endothelial lining. It appears that other factors such as the buffering effects of orbital tissue, pia mater and the conformation of the lamina itself may further influence the TLPG and this previously unsuspected pressure gradient may in part be responsible for the outward bowing of the lamina in the optic cup. Increased TLPG and its relationship to IOP may have a role in the progression of optic nerve damage and glaucoma.

Blood supply

The optic nerve has a complex blood supply, which has been extensively investigated because of its importance in the pathogenesis of glaucoma (see Ch. 9). The intraocular portion is supplied by branches from four sources: central retinal vessels and their branches, scleral vessels (the circle of Zinn–Haller), choroidal vessels and pial vessels (see Fig. 1-42D). The first three are derived from the ophthalmic or central retinal artery, and pial vessels from the adjacent branches of the internal carotid artery. The majority of capillaries pierce the nerve and course longitudinally within the nerve via the glial septae.

Orbital contents

GENERAL ARRANGEMENT

The orbits are a pair of bony sockets, with each orbital cavity having a volume of about 30 cm^3.

PERIORBITA AND ORBITAL FIBROADIPOSE TISSUE (Figs 1-43 and 1-44)

The orbital contents are bound together and supported by fibroadipose tissue. This connective tissue has classically been divided into separate components.

- *Periorbita* or periosteum of the orbit. This layer of connective tissue is frequently described as having a dense outer layer and a looser inner layer, which invests orbital nerves and the lacrimal gland. It is tightly bound to the bones only at the sutures, fissures and foraminae in the orbital walls, and also to the posterior lacrimal crest where it covers the lacrimal sac and is continuous with the fibrous lining of the nasolacrimal duct. It forms a dense membrane over the inferior and superior orbital fissures, with sufficient gaps for transmission of nerves and vessels. It is continuous with the periosteum lining the optic foramen and with the sheath of the optic nerve, itself an extension of dura mater of the brain. The periorbita is firmly attached at the orbital margins anteriorly where it becomes continuous with the orbital septum (palpebral fascia) in the eyelids (see pp. 84).
- *Bulbar fascia (Tenon's capsule)*, a thick fibrous sheath enclosing the globe but separated from it by a layer of loose connective tissue.
- *Muscular fascial sheaths* that surround the extraocular muscles and blend with the bulbar fascia.
- *Medial and lateral check ligaments.*
- *Suspensory ligament (of Lockwood).*

The fibrous intermuscular membrane connecting the four rectus muscles helps create the intraconal space (best developed in the anterior part of the orbit;

BOX 1-27 CLINICAL CORRELATES

PAPILLOEDEMA

Papilloedema is the swelling of the papillary fibres, which appears as a raised white disk margin, and is partly the result of the lack of Müller cells in this region. These cells serve to bind the nerve fibres together in the remainder of the retina. Papilloedema may be a cardinal sign of raised intracranial pressure which is transmitted to the subarachnoid space which envelopes the optic nerve as far anteriorly as the sclera surrounding the optic nerve.

MENINGITIS

The continuation of the subarachnoid space from the cranial cavity along the nerve may facilitate the spread of infection or tumours from the orbit to the cranium and vice versa.

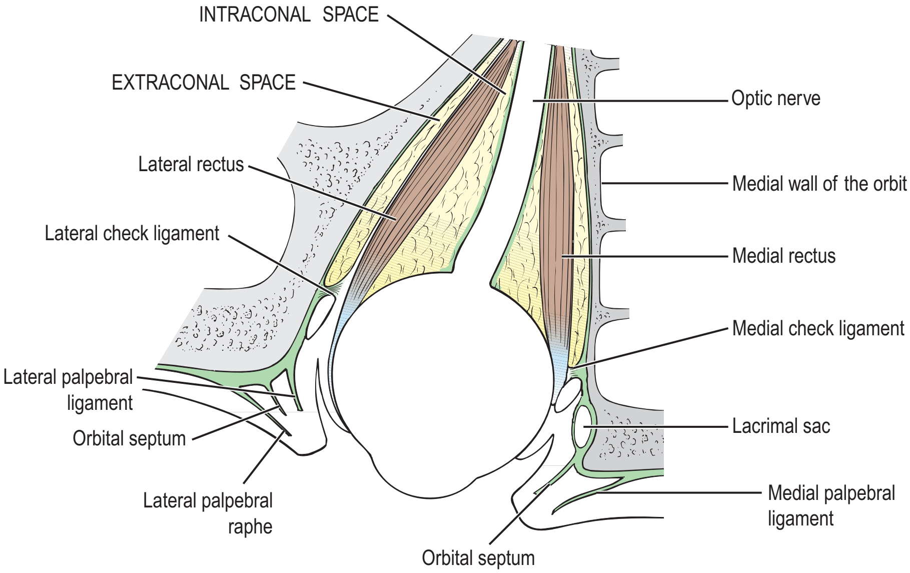

FIGURE 1-43 Diagram of a horizontal section through the orbit illustrating the formation of an intra- and extraconal space by the four rectus muscles (only medial and lateral rectus shown in this section).

incomplete behind the globe) (Fig. 1-44B,C). Besides the 'check' ligaments there are other specific attachments via fibrous bands to the orbital walls throughout their course (Fig. 1-44A–C). A theoretical framework, known as the 'active pulley hypothesis' postulates a crucial role for these connective tissue bands, known as 'pulley suspensions', in understanding the kinematics of extraocular muscle action. These suspensions pass between the orbital wall and the 'pulley sleeve' of each muscle, which is described as a ring-like extension of the connective tissue from Tenon's capsule posteriorly around the muscle. The tone of the pulleys is possibly under neuronal control because of the presence of smooth muscle fibres. The 'active pulley hypothesis' proposes that the rectus muscles have a so-called 'orbital layer' of fibres that are continuous with (or 'blend with' or 'insert into') these sleeves (and thus also into the pulley suspensions), in essence one part of a bifid insertion. The inner half or 'global layer' of the rectus muscle continues through the sleeve and bulbar fascia to insert directly into the sclera. The newly postulated function for orbital connective tissue in the 'active pulley hypothesis', which states in effect that this dual insertion allows the pulleys to act as a second 'origin' and thus influence the direction of pull of the extraocular muscles, has gained wide acceptance but some investigators have questioned the anatomical evidence of orbital muscle fibres terminating in connective tissue other than the sclera.

There are well-recognized but variable amounts of smooth muscle within the orbital connective tissue, including the sleeves of some of the recti muscles, whose functions (besides the superior and inferior palpebral muscles) are presently unclear. It has been suggested that like the smooth muscle covering the inferior orbital fissure (orbitalis or Müller's muscle) they may represent a redundant evolutionary remnant.

Veins passing through the orbit are supported by connective tissue septae (Fig. 1-44B); arteries, instead, travel among the fat locules and frequently pierce the septae. A thickened band of orbital fibrous tissue

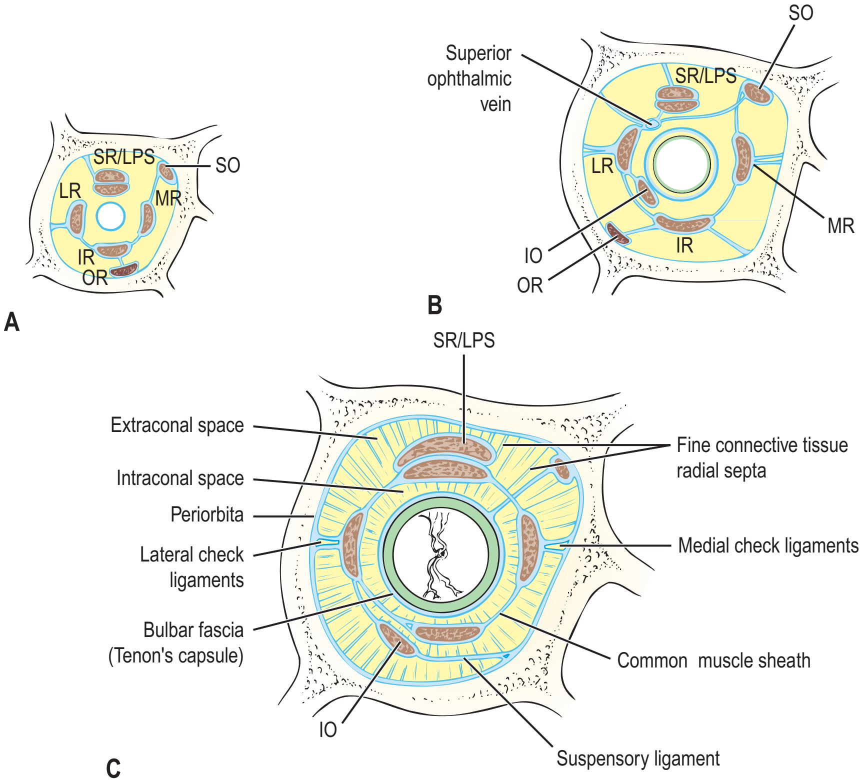

FIGURE 1-44 Schematic diagrams of the connective tissue septae associated with each extraocular muscle at three levels in the orbit: (**A**) near orbital apex; (**B**) posterior part of the globe; (**C**) close to the equator of the globe. SO, superior oblique; MR, medial rectus; LR, lateral rectus; IR, inferior rectus; SR, superior rectus; LPS, levator palpebrae superioris; IO, inferior oblique; OR, orbitalis. (Modified from Koornneef, 1982.)

connects the superior rectus and the levator palpebrae superioris. This aids in coordinating lifting of the eyelid when the eye is directed upwards by the superior rectus.

The complex and interlinked nature of the fibroadipose system of connective tissue septae may explain why patients with orbital floor 'blow-out' fractures display vertical ocular mobility problems. It is not necessary to invoke the incarceration of the inferior rectus and inferior oblique muscles in the fracture to explain the symptoms.

EXTRAOCULAR MUSCLES

There are six true extraocular muscles responsible for movements of the globe. In addition there is one further 'orbital' muscle, the levator palpebrae superioris, which originates at the orbital apex and inserts into the tarsal plate and upper eyelid (see pp. 81).

The true extraocular muscles comprise *four rectus muscles*, which arise from the tendinous ring at the apex of the orbit and insert into the sclera about 4–8 mm behind the limbus, and *two oblique muscles* (superior and inferior), whose tendons approach the globe from in front and insert into the posterior aspect of the sclera (Figs 1-45 and 1-48). Details of the six true extraocular muscles, including their innervation, origin, insertion, tendon length (important in the surgical management of strabismus), length of muscle belly, the angle subtended by the muscle axis to the vertical, and the size of the motor units, are provided in Table 1-2. The origins of the muscles are shown in Figure 1-47 and the pattern of insertion into the sclera is shown in Figure 1-48. The collagen bundles of the tendons blend with the scleral collagen as shown in Figure 1-17C. The relations of the orbital muscles to each other and to the orbital nerves are summarized diagrammatically in Figures 1-45 and 1-46. Movements are discussed in Chapter 5 (pp. 326).

Microscopic anatomy of extraocular muscle

Histologically, extraocular muscle differs from skeletal muscle in the following respects (compare Fig. 1-49A with Fig. 1-49B).

- The epimysium or muscle sheath of extraocular muscle is generally very thin by comparison with other muscles.
- The fibres are not tightly packed but are separated by unusually large amounts of connective tissue (perimysium) rich in reticulin and elastic fibres.
- The muscle fibres are rounded or oval in shape with small fibres (5–15 μm) around the periphery of the muscle and larger fibres (10–40 μm) in the centre.
- Extraocular muscle is the most vascular in the body, next to myocardium. The most vascularized region is the orbital aspect.
- In normal extraocular muscle there often appear to be histopathological or ultrastructural changes normally associated with myopathy, i.e. mild mononuclear cellular infiltrate, centrally placed nuclei, disorganization of the sarcolemma, disruption of the Z lines, and mitochondrial clumping.
- Extraocular muscle contains large numbers of specialized sensory or proprioreceptive endings, including large muscle spindles up to 1 mm long (nuclear bag fibres, nuclear chain fibres and annular nerve terminals). Golgi tendon organs are also numerous and are generally found within the tendons of extraocular muscles in greater numbers than in skeletal muscle (Fig. 1-49). The afferent fibres from extraocular muscles are transmitted initially for part of their course in the respective cranial nerve innervating the muscle (either III, IV or VI); however, they leave these nerves and join the ophthalmic division of the trigeminal, either in the cavernous sinus or in the brainstem. Their cell bodies are situated in the mesencephalic nucleus, although some muscle afferents have been traced to Purkinje cells in the cerebellum, and play an important role in positional sense and control of ocular movements (both saccadic and tracking).

The structural differences between extraocular and skeletal muscle outlined above are not surprising in light of the fundamental differences in function, namely the constancy of activity (even during sleeping) and the rapidity and fine gradation of contraction of extraocular muscle required to fixate subjects of interest on the fovea. Since both eyes must move together, both sets of six muscles must be highly coordinated and move simultaneously (see Ch. 5). Up to six types of muscle fibre have been identified morphologically, but functionally there appear to be three main types (Table 1-3).

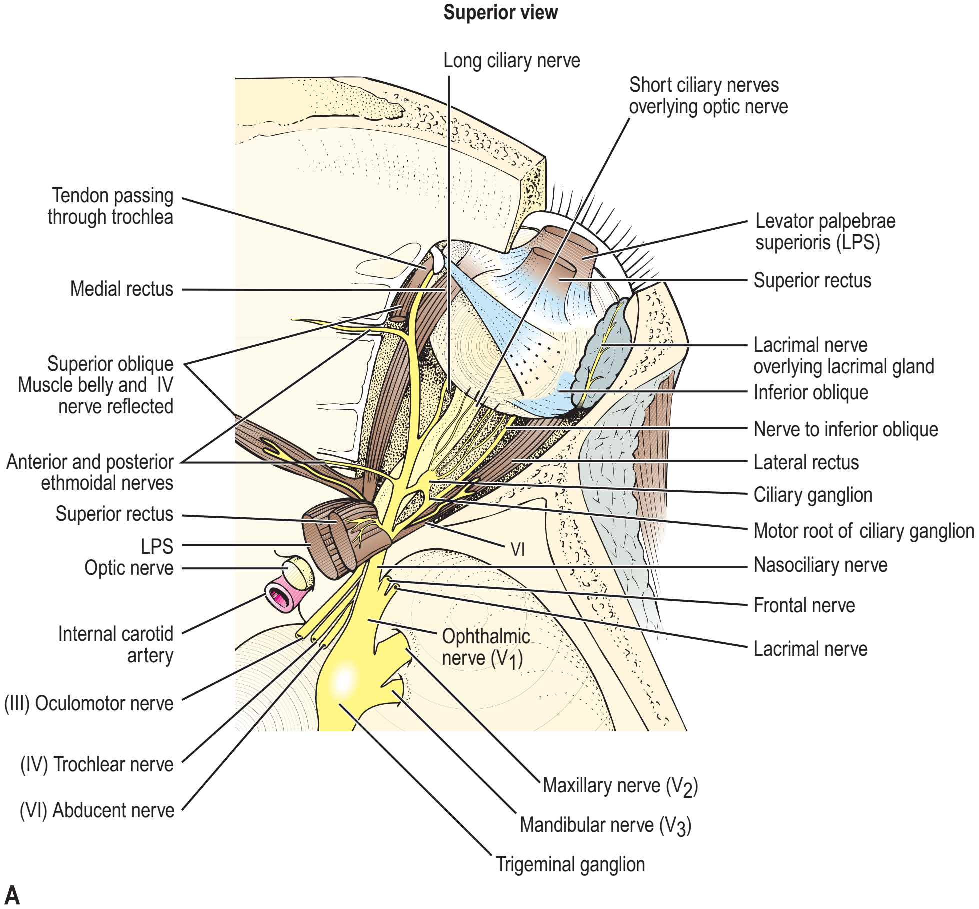
Superior view
Long ciliary nerve
Short ciliary nerves overlying optic nerve
Tendon passing through trochlea
Medial rectus
Superior oblique Muscle belly and IV nerve reflected
Anterior and posterior ethmoidal nerves
Superior rectus
LPS
Optic nerve
Internal carotid artery
(III) Oculomotor nerve
(IV) Trochlear nerve
(VI) Abducent nerve
Levator palpebrae superioris (LPS)
Superior rectus
Lacrimal nerve overlying lacrimal gland
Inferior oblique
Nerve to inferior oblique
Lateral rectus
Ciliary ganglion
Motor root of ciliary ganglion
VI
Nasociliary nerve
Frontal nerve
Lacrimal nerve
Ophthalmic nerve (V1)
Maxillary nerve (V2)
Mandibular nerve (V3)
Trigeminal ganglion
A

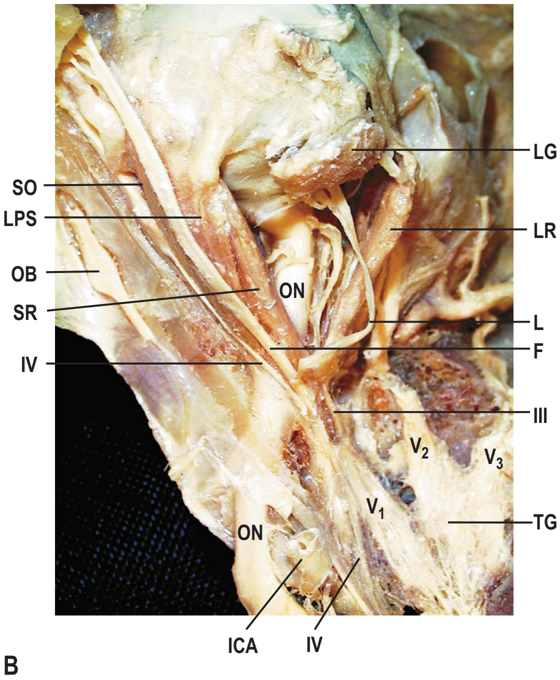
SO
LPS
OB
SR
IV
ON
LG
LR
L
F
III
V2
V3
V1
TG
ON
ICA
IV
B

FIGURE 1-45 Diagram (**A**) and prosected specimen (**B**) of the orbit viewed from above, revealing the relations of the orbital nerves and extraocular muscles. Orbital fat and vessels have been excluded for the purposes of clarity.

- The roof of the orbit and superior orbital fissure have been removed and the periorbita divided.
- In (**A**) the lacrimal (L), frontal (F) and trochlear (IV) nerves have been cut. In (**B**) these nerves are intact as they pass external to the tendinous ring or annulus.
- In (**A**) only one long ciliary nerve is shown arising from the nasociliary nerve.
- The sensory root of the ciliary ganglion emerges from the nasociliary. The motor root (parasympathetic fibres) arises from the branch of the oculomotor supplying inferior oblique.
- In (**A**) and (**B**) the lateral dural covering of the cavernous sinus has been removed to expose the cranial nerves before they pass through the superior orbital fissure: TG, trigeminal ganglion; V_1, V_2, V_3, divisions of the trigeminal nerve; OB, olfactory bulb; ICA, internal carotid artery; III, oculomotor nerve; VI, abducens nerve; SO, superior oblique; LR, lateral rectus; LG, lacrimal gland; LPS, levator palpebrae superioris; SR, superior rectus; ON, optic nerve.
- The optic canal has not been opened.

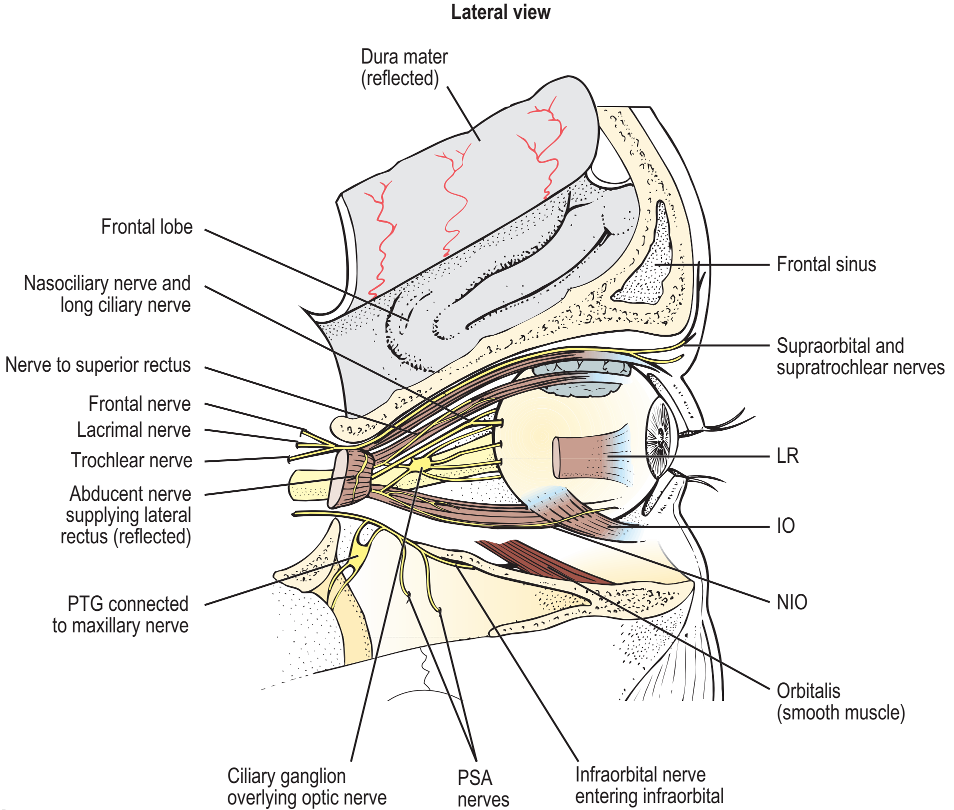
Lateral view
Dura mater (reflected)
Frontal lobe
Nasociliary nerve and long ciliary nerve
Nerve to superior rectus
Frontal nerve
Lacrimal nerve
Trochlear nerve
Abducent nerve supplying lateral rectus (reflected)
PTG connected to maxillary nerve
Frontal sinus
Supraorbital and supratrochlear nerves
LR
IO
NIO
Orbitalis (smooth muscle)
Ciliary ganglion overlying optic nerve
PSA nerves
Infraorbital nerve entering infraorbital groove

A

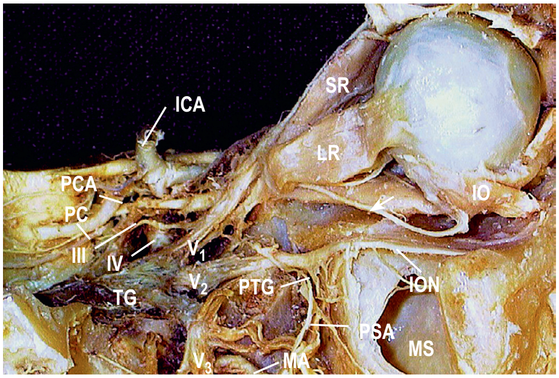
ICA
SR
LR
PCA
PC
III
IV
V_1
V_2
V_3
TG
PTG
IO
ION
PSA
MS
MA

B

FIGURE 1-46 (**A**) Diagram illustrating a lateral view of a dissected orbit revealing the relations of the orbital nerves and extraocular muscles (vessels have been excluded for the purposes of clarity). (**B**) Dissection of the orbit similar to the diagram above except that lateral rectus has not been cut and the course of the orbital nerves within the cavernous sinus is also shown (by removal of the lateral dural wall). ICA, internal carotid artery; PCA, posterior communicating artery; PC, posterior cerebral artery; MA, maxillary artery; TG, trigeminal ganglion; V_1, V_2, V_3, divisions of the trigeminal nerve; PTG, pterygopalatine ganglion; PSA, posterior superior alveolar nerves; ION, infraorbital nerve; IO, inferior oblique; LR, lateral rectus; SR, superior rectus; arrow, nerve to superior oblique branch of oculomotor nerve (III); IV, trochlear nerve; MS, maxillary sinus.

- The lateral wall has been removed and the infratemporal fossa has been dissected to expose the pterygomaxillary fissure and pterygopalatine fossa.
- The cranial cavity has been opened to reveal the dura (reflected) covering the frontal lobe.
- The lateral rectus has been divided and reflected to expose the optic nerve and other cranial nerves entering the orbit through the tendinous ring. Note the abducent nerve entering its bulbar surface.
- The ciliary ganglion lies between the lateral rectus and the optic nerve. Note the motor root and sensory root as seen in the superior view. The nerve to inferior oblique (NIO) is a useful landmark for finding the ciliary ganglion. The short ciliary nerves emerge from the ganglion and enter the globe around the optic nerve.
- Note the three nerves which enter the orbit outside the tendinous ring: lacrimal, frontal and trochlear.
- The nerve to superior rectus (branch of superior division of III nerve) pierces the muscle and enters the levator palpebrae superioris, which it supplies, from below.
- Branches of the pterygopalatine ganglion (PTG) enter the orbit through the inferior orbital fissure and contribute to the formation of the retrobulbar plexus (not shown).
- Inferior oblique passes backwards, laterally and superiorly beneath the inferior rectus.
- Orbitalis (Müller's muscle), a band of smooth muscle, covers the inferior orbital fissure.

TABLE 1-2 **Summary of anatomical features of human extraocular muscles**

Muscle	Innervation	Origin	Insertion (mm from cornea)	Tendon length (mm)	Length of muscle belly (mm)	Size of motor unit	Comment
Medial rectus	III (inf)	Tendinous ring	5.6	3.6	40	1:1.7–1:4	Largest of the ocular muscles
Inferior rectus	III (inf)	Tendinous ring	6.6	5	40	1:2–1:6	
Lateral rectus	VI	Tendinous ring via two heads	7.0	8.4	40	1:3–1:6	Opening between the two heads bridging the medial end of superior orbital fissure
Superior rectus	III (sup)	Tendinous ring	7.8	5.4	41	1:4	Lies beneath levator palpebrae superioris
Superior oblique	IV	Superomedial to optic canal	Lateral aspect of posterosuperior quadrant	Tendon forms 10 mm before winding around trochlea	32	1:5–1:6	The only extraocular muscle with a fusiform shape. IV nerve enters the muscle on its upper border
Inferior oblique	III (inf)	Behind orbital margin lateral to nasolacrimal canal	Posterolateral quadrant, mostly below horizontal	Very short tendon; muscle fibres almost reach the sclera	34	1:7	The only extraocular muscle not to originate at apex of orbit. Passes between the eye and lateral rectus

Values are approximate means based on several studies; for full details and ranges see Eggers (1982). III (inf), inferior division of oculomotor nerve; III (sup), superior division of oculomotor nerve.

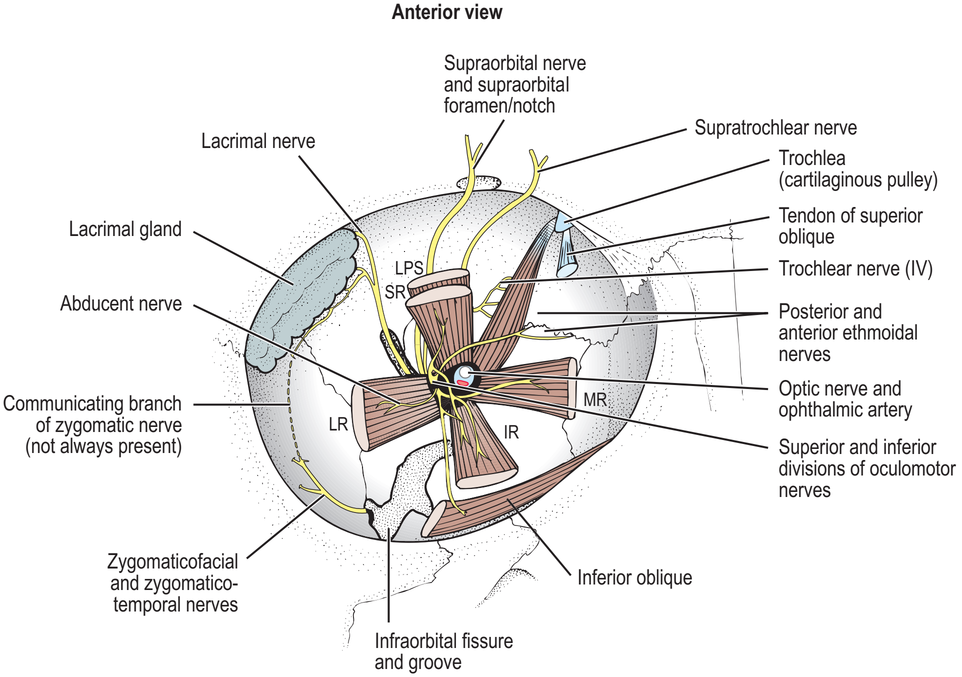
Anterior view
Supraorbital nerve and supraorbital foramen/notch
Lacrimal nerve
Supratrochlear nerve
Trochlea (cartilaginous pulley)
Lacrimal gland
Tendon of superior oblique
Trochlear nerve (IV)
LPS
SR
Abducent nerve
Posterior and anterior ethmoidal nerves
Optic nerve and ophthalmic artery
Communicating branch of zygomatic nerve (not always present)
MR
LR
IR
Superior and inferior divisions of oculomotor nerves
Zygomaticofacial and zygomatico-temporal nerves
Inferior oblique
Infraorbital fissure and groove

FIGURE 1-47 The orbit from in front with the globe removed to show the origins of the extraocular muscles and the orbital nerves (vessels and fat are not included in this diagram).

BLOOD VESSELS OF THE ORBIT (Fig. 1-50)

The orbital contents are supplied chiefly by the ophthalmic artery, which usually arises from the internal carotid artery shortly after it emerges from the roof of the cavernous sinus. It commences its course beneath the optic nerve, closely bound to the dura while in the optic canal, then winds around its lateral aspect, and finally passes above the nerve. It then proceeds forward above the medial rectus and under the superior oblique. It ends its tortuous course by dividing into dorsal nasal and supratrochlear branches. The branches are summarized in Figure 1-50. There are several important points of anastomosis between arteries derived from the internal carotid and the external carotid arteries (Table 1-4).

These anastomoses may be important during occlusive vascular disease of the ophthalmic artery by serving as alternative routes of blood supply to the eye and orbit. The veins that accompany the above arteries, in common with most veins of the head and neck, lack valves and thus there are several sites of communication between veins on the upper face and lids with intraorbital veins (superior and inferior ophthalmic veins) which drain posteriorly into the cavernous sinuses. The inferior ophthalmic vein may drain via the inferior orbital fissure into the pterygoid venous plexus. These communications are important clinically as they act as potential routes for the spread of infection from the face around the nose and eye to the cavernous sinuses and the cranial cavity.

Cranial nerves associated with the eye and orbit

GENERAL FUNCTIONAL ARRANGEMENT

Cranial nerves contain a diversity of functional components. Besides those found in spinal nerves (somatic efferents, somatic afferents, general visceral efferents, general visceral afferents), cranial nerves also contain additional functional categories including special visceral efferents (branchiomotor), special somatic afferents (special senses, hearing and balance), and special visceral afferents (taste and smell). The functional classification of cranial nerve components and their target organs and tissues are summarized in eTable 1-1.

eTABLE 1-1 Functional analysis of cranial nerve components

Functional classification	Modality/target (ontogeny/phytogeny)	Present in cranial nerves
Somatic efferent (general motor)	Supplies skeletal muscle of somatic origin (preotic somites – extraocular muscles; occipital somites – tongue musculature)	III, IV and VI XII
Somatic afferent (general sensory)	Pain, temperature and touch. Supplies skin and mucous membranes of the head and neck	Predominantly in V but several minor elements in VII, IX and X
General visceral efferents* (parasympathetic)	Supplies smooth muscle (viscera), cardiac muscle, glands, blood vessels and intrinsic eye muscles (ciliary muscle and sphincter pupillae)	III, VII, IX, X and XI. X (vagus) is largest parasympathetic nerve in the body
General visceral afferents	Pain and sensibility of viscera	VII, IX and X
Special visceral efferents (branchiomotor)	Skeletal muscles of mastication and facial expression (i.e. pharyngeal arch or visceral evolutionary origin)	V, VII, IX, X and XI
Special somatic afferent (special senses concerned with body position, excluding vision)	Maintenance of balance and reception of sound (vestibulocochlear organ)	VII
Special visceral afferent (special visceral senses, taste and smell)	Olfactory epithelium in nasal cavity and taste receptors in tongue and palate	Olfaction (smell) in I; taste in VII, IX and X

*Sympathetic nerve fibres, originating from upper thoracic segments of the spinal cord and synapsing in cervical ganglia may ‘hitch-hike’ with various cranial nerves and/or blood vessels to reach ocular and obital structures, e.g. dilator pupillae muscle and tarsal muscle (see Fig. 1-65).

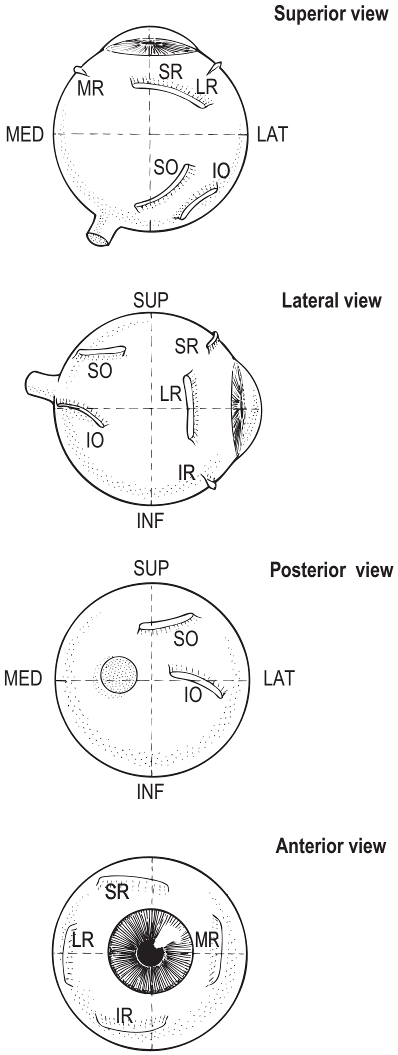

FIGURE 1-48 Four views of the right globe to demonstrate the insertions of the extraocular muscles.

The sites of origin of the cranial nerves from the brainstem are illustrated in Figures 1-51 and 1-52.

OCULOMOTOR NERVE (CRANIAL NERVE III) (Fig. 1-52)

This is the largest of the extraocular nerves and supplies all the extraocular muscles except the lateral rectus and superior oblique.

Origin

The oculomotor nerve nuclei comprise two main types:

- a complex of five individual motor (*somatic efferent*) nuclei containing the cell bodies of the multipolar motor neurones whose axons directly innervate their respective extraocular muscles
- a *general visceral efferent* nucleus, the Edinger–Westphal nucleus, containing small spindle-shaped preganglionic (first-order) *parasympathetic* neurones.

The oculomotor nuclei lie at the level of the superior colliculus in the ventral region of the periaqueductal grey matter and extend cranially for a short distance into the floor of the third ventricle. The *medial longitudinal fasciculus* lies lateral to the nucleus and contains the axons of internuclear neurones that pass vertically between the brainstem nuclei of the III, IV and VI nerves. The fibres emerge from the oculomotor nuclei, pass anteriorly through the tegmentum of the midbrain and red nucleus, and emerge medial to the cerebral peduncle at the upper border of the pons (Figs 1-51 and 1-52).

Intracranial and intracavernous course

The nerve passes forward, laterally and slightly downward in the interpeduncular fossa (one of the enlargements of the subarachnoid space or cisterns) lateral to the posterior communicating artery (Fig. 1-52). It passes between the posterior cerebral artery (above) and the superior cerebellar artery (below). It grooves the posterior clinoid process and courses forward before it passes through the dural roof of the cavernous sinus (Fig. 1-46B). The nerve runs forward in the upper part of the lateral wall of the cavernous sinus (Fig. 1-9C and eFig. 1-2B) and enters the intraconal space of the orbit through the superior orbital fissure within the tendinous ring (Figs 1-5 and 1-6), where it divides into superior and inferior divisions.

Intraorbital course

In the orbit the nasociliary nerve is interposed between the two divisions of the oculomotor nerve. The superior supplies the superior rectus, which it pierces to reach levator palpebrae superioris. The inferior division splits into several branches which supply the medial rectus and inferior rectus, and a long branch passes forward on the lateral aspect of inferior rectus to reach inferior oblique (Fig. 1-46B). It is from this latter branch that the stout motor root (preganglionic parasympathetic fibres) passes to the ciliary ganglion, the site of postganglionic parasympathetic (second-order) neurones (Fig. 1-45A). Axons from the

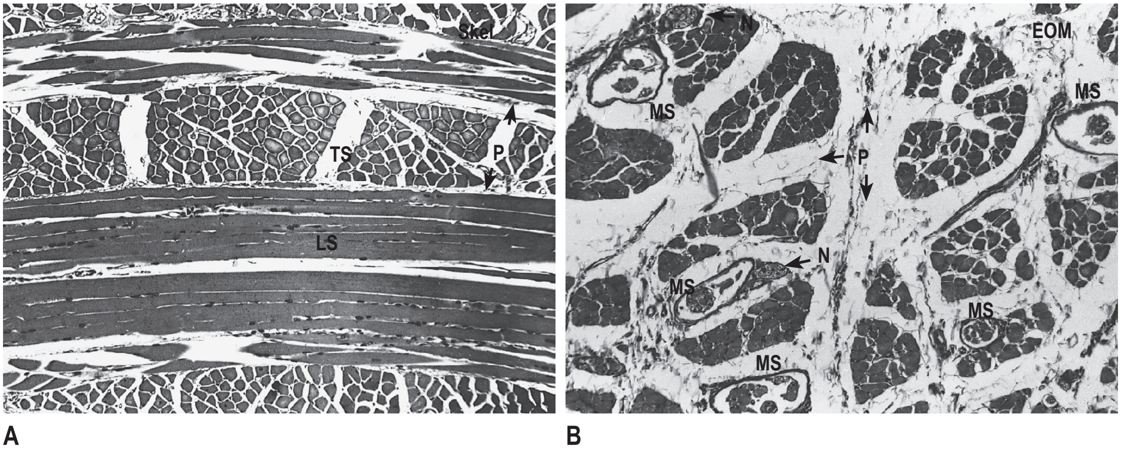

FIGURE 1-49 Histological section of (**A**) normal human skeletal muscle fibres (SKel) and (**B**) extraocular muscle fibres (EOM): LS, longitudinal section of fibres; TS, fibres in transverse section; P, perimysium; MS, muscle spindles; N, nerve. Original magnifications: **A** and **B**, × 100.

TABLE 1-3 Classification of mammalian extraocular muscle fibres

Type A	**Type B**	**Type C**
Large diameter	Intermediate diameter	Small diameter
Single end-plate	Multiple end-plates	Small *en grappe* plates
Fast twitch	Slow twitch	Tonic contractions
Required for saccadic movements	Needed for smooth pursuit movements	Function to align both visual axes, i.e. fine local contractions

TABLE 1-4 Sites of anastomosis between branches of the internal and external carotid arteries

External carotid branch	**Internal carotid branch**	**Region of anastomosis**
Angular artery (facial)	Dorsalis nasi (ophthalmic)	Medial palpebral margin
Transverse facial artery (superficial temporal)	Lacrimal artery (ophthalmic)	Lateral palpebral margin
Middle meningeal artery and deep temporal artery	Lacrimal artery (ophthalmic)	Orbit

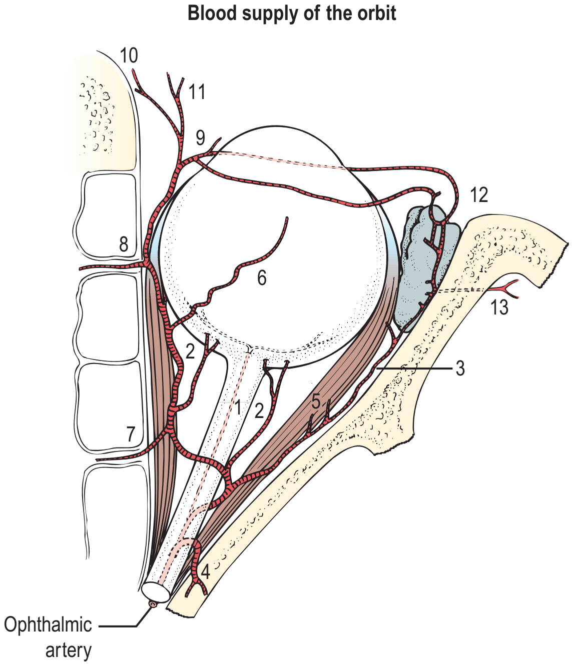

FIGURE 1-50 Diagram summarizing the blood supply of the orbit as seen in a superior view: 1, central retinal artery; 2, posterior ciliary arteries (usually emerge as two trunks that divide into the short posterior ciliary arteries (seven or more) and the long posterior ciliary arteries (usually two, medial and lateral)); 3, lacrimal artery; 4, recurrent branches (to meninges); 5, muscular branches (give rise to anterior ciliary arteries); 6, supraorbital artery; 7, posterior ethmoidal artery; 8, anterior ethmoidal artery; 9, superior and inferior medial palpebral arteries; 10, dorsalis nasi; 11, supratrochlear; 12, superior and inferior lateral palpebral arteries; 13, zygomatic branches of the lacrimal artery.

BOX 1-28 CLINICAL CORRELATES

Intracavernous lesions of cranial III nerve

The oculomotor, like other cranial nerves coursing through the cavernous sinus, can become involved in pathological processes such as venous thrombosis or aneurysms of the internal artery. Pituitary enlargements more commonly affect the oculomotor and trochlear nerves than the abducent, which is protected by the internal carotid artery. Meningioma or expanding lesions in the region of the superior orbital fissure can also compress the nerve.

BOX 1-29 CLINICAL CORRELATES

Lesions of the oculomotor nerve

Complete lesions of the oculomotor nerve (e.g. trauma) result in:

- inability to look upwards, downwards or medially
- lateral or *external strabismus* because of unopposed action of lateral rectus
- diplopia
- complete ptosis (paralysis of levator palpebrae superioris and unopposed orbicularis oculi)
- dilated non-reactive pupils (unopposed dilator pupillae)
- lack of accommodation.

Incomplete lesions – some of the symptoms above may be present.

Internal ophthalmoplegia – loss of parasympathetic components only. This may be the first sign of nerve palsy as the parasympathetic fibres are located superficially in the nerve and they may be damaged first in intracranial lesions; thus, pupil dilation is a crucial sign of compression within the cranial cavity following head injury.

External ophthalmoplegia – the loss of extraocular muscle supply.

Intracranial lesions affecting the oculomotor nerve – aneurysms of adjacent arteries around the brainstem may cause compression of the nearby nerve. Meningitis can involve the nerve along its course in the subarachnoid space.

postganglionic neurones travel in the short ciliary nerves to supply the choroid, sphincter pupillae of the iris and the ciliary muscle (Fig. 1-52).

The nerves that supply extraocular muscles generally pierce the muscle one-third of the way along the muscle belly on the bulbar aspect.

TROCHLEAR NERVE (CRANIAL NERVE IV) (Fig. 1-53)

This is the only somatic efferent nerve to emerge from the posterior aspect of the central nervous system. It is also unusual in that it decussates before leaving the brainstem. It supplies only one extraocular muscle, the *superior oblique*.

BOX 1-30 CLINICAL CORRELATES

Intracranial lesions affecting the trochlear nerve

The trochlear nerve is rarely paralysed alone, although it is particularly vulnerable at its posterior exit from the brainstem and as it winds round the midbrain. Lesions causing compression on the undersurface of the tentorium may affect the trochlear nerve. It may also be involved in pathological processes in the cavernous sinus. Patients suffering paralysis of the superior oblique because of trochlear nerve lesions suffer diplopia when looking down and have difficulty in looking down when the eye is adducted because the superior oblique is the only depressor in the adducted state. Patients characteristically carry the head tilted to the non-affected side with the chin lowered to compensate for the overaction of the inferior oblique producing unopposed torsion on the eye.

Origin

The nucleus lies in the anterior part of the periaqueductal grey matter at the level of the inferior colliculus (caudal to III nerve nucleus) in line with the other oculomotor nuclei. The fibres first pass anteriorly and laterally towards the tegmentum before turning and passing posteriorly around the periaqueductal grey matter and into the superior medullary velum (part of the roof of the fourth ventricle) where they decussate before emerging from the posterior surface of the brainstem in the posterior cranial fossa (Fig. 1-53).

Intracranial and intracavernous course

The trochlear nerve winds around the crus of the midbrain (cerebral peduncles) above the superior cerebellar artery and the pons and below the posterior cerebral artery. It continues anteriorly immediately beneath the free edge of the tentorium cerebelli (Fig 1-53). It pierces this dura and enters the lateral wall of the cavernous sinus beneath the oculomotor nerve (Figs 1-9C and eFig. 1-2B; 1-45B and 1-46B). The trochlear nerve then passes upwards, thus coming to lie above the oculomotor nerve before entering the orbit outside the tendinous ring in the lateral part of the superior orbital fissure.

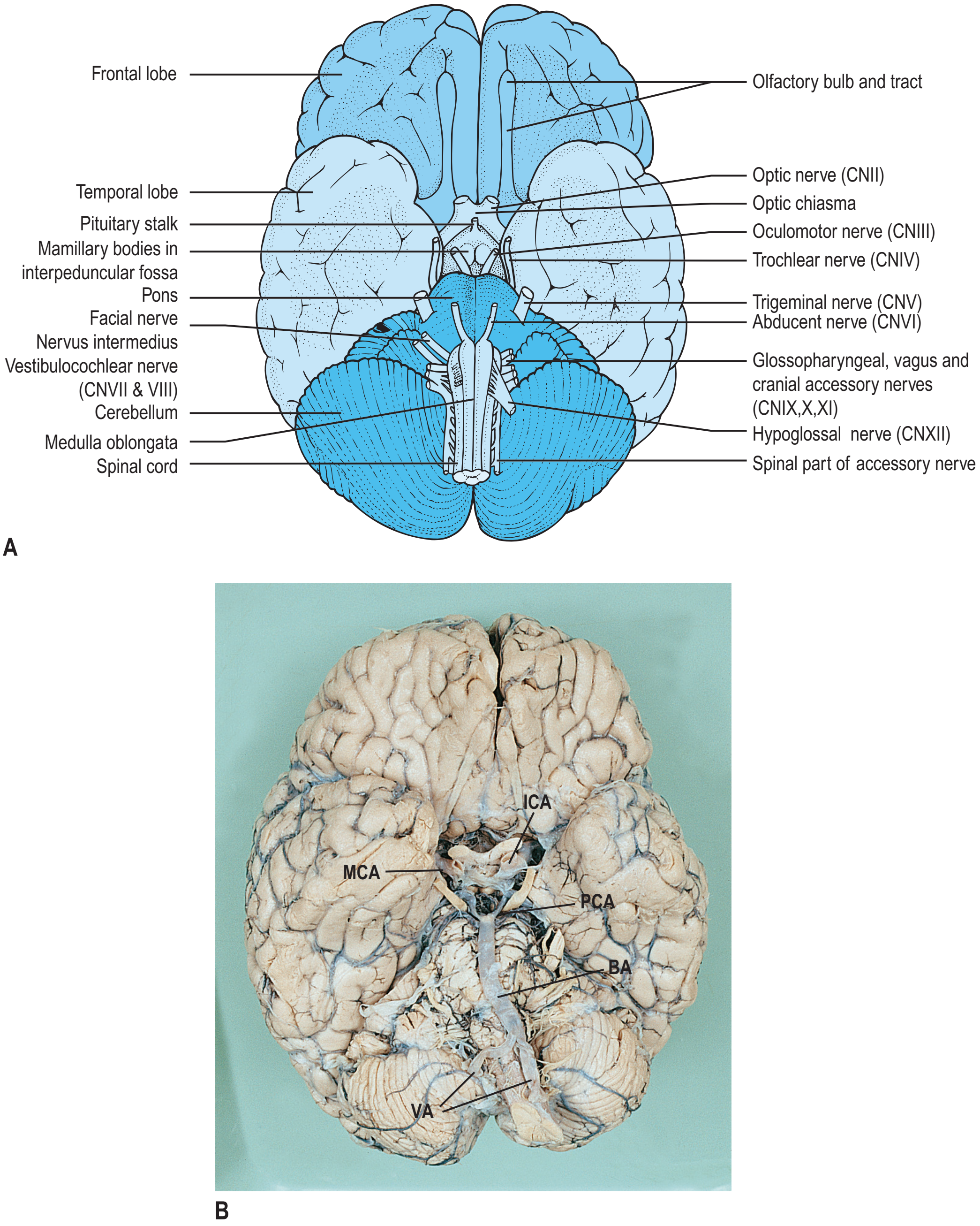
Frontal lobe
Temporal lobe
Pituitary stalk
Mamillary bodies in interpeduncular fossa
Pons
Facial nerve
Nervus intermedius
Vestibulocochlear nerve (CNVII & VIII)
Cerebellum
Medulla oblongata
Spinal cord
Olfactory bulb and tract
Optic nerve (CNII)
Optic chiasma
Oculomotor nerve (CNIII)
Trochlear nerve (CNIV)
Trigeminal nerve (CNV)
Abducent nerve (CNVI)
Glossopharyngeal, vagus and cranial accessory nerves (CNIX,X,XI)
Hypoglossal nerve (CNXII)
Spinal part of accessory nerve
A
ICA
MCA
PCA
BA
VA
B

FIGURE 1-51 Ventral views of the brain to demonstrate the origin of the cranial nerves: (**A**) diagram without vessels; (**B**) photograph of a whole brain with vessels. ICA, internal carotid artery; MCA, middle cerebral artery; PCA, posterior cerebral artery; BA, basilar artery; VA, vertebral artery.

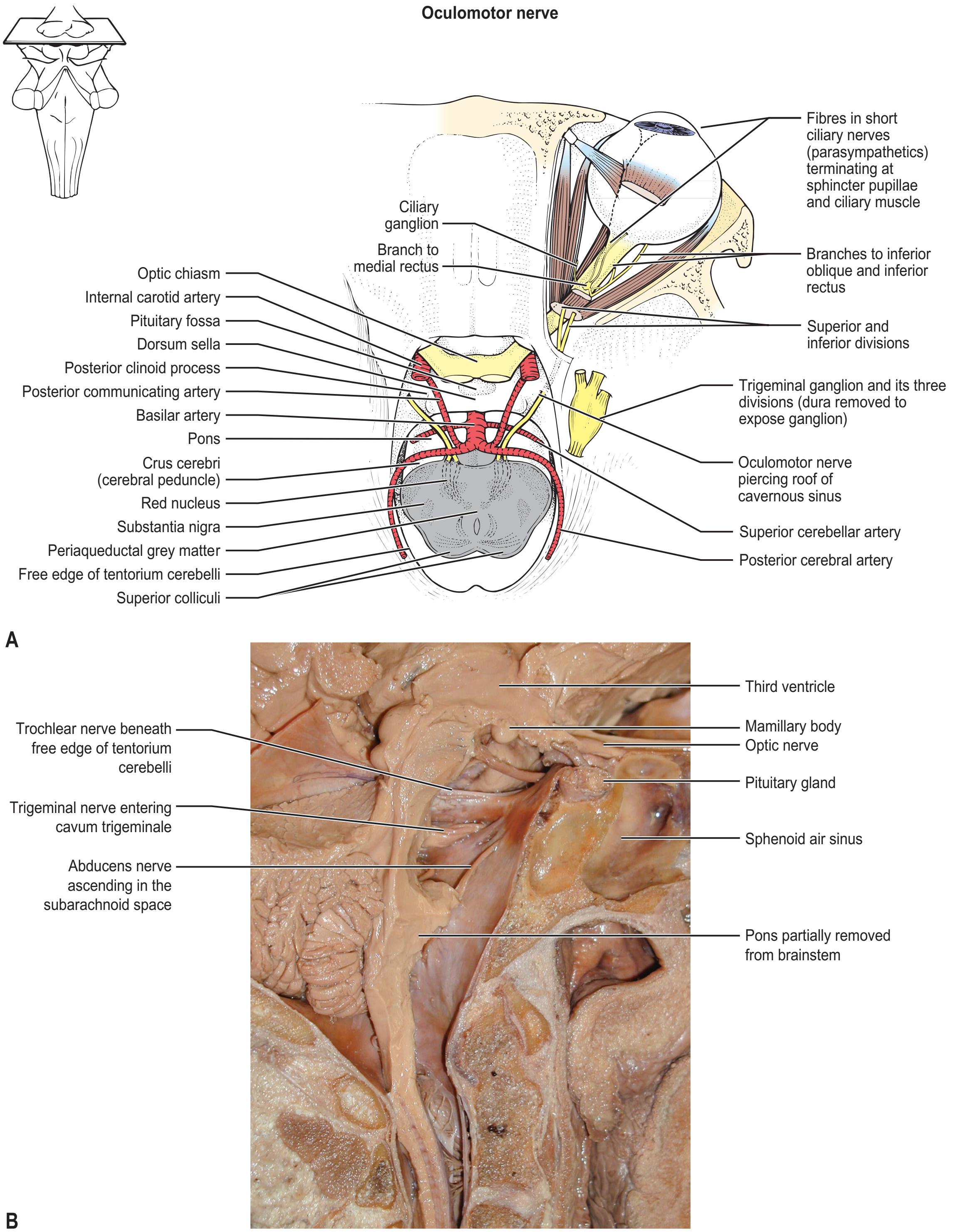
Oculomotor nerve
Ciliary ganglion
Branch to medial rectus
Fibres in short ciliary nerves (parasympathetics) terminating at sphincter pupillae and ciliary muscle
Branches to inferior oblique and inferior rectus
Superior and inferior divisions
Optic chiasm
Internal carotid artery
Pituitary fossa
Dorsum sella
Posterior clinoid process
Posterior communicating artery
Basilar artery
Pons
Crus cerebri (cerebral peduncle)
Red nucleus
Substantia nigra
Periaqueductal grey matter
Free edge of tentorium cerebelli
Superior colliculi
Trigeminal ganglion and its three divisions (dura removed to expose ganglion)
Oculomotor nerve piercing roof of cavernous sinus
Superior cerebellar artery
Posterior cerebral artery
A
Third ventricle
Mamillary body
Optic nerve
Pituitary gland
Sphenoid air sinus
Pons partially removed from brainstem
Trochlear nerve beneath free edge of tentorium cerebelli
Trigeminal nerve entering cavum trigeminale
Abducens nerve ascending in the subarachnoid space
B

FIGURE 1-52 (**A**) Diagram summarizing the brainstem origin (inset shows level of section), intracranial, intracavernous and intraorbital course of the oculomotor nerve. (**B**) Brainstem cranial nerves.

Intraorbital course

It passes forward above the origin of levator palpebrae superioris close to the bone, to enter the upper free edge of the superior oblique (Figs 1-45A,B; 1-47 and 1-53).

ABDUCENT NERVE (CRANIAL NERVE VI) (Fig. 1-54)

The abducens nucleus lies in the mid-pons beneath the floor of the upper part of the fourth ventricle, close to the midline beneath the facial colliculus. The fibres pass anteriorly to emerge on the lower border of the pons above the medulla near the midline (Figs 1-51 and 1-54).

Intracranial and intracavernous course

The abducent nerve has the longest intracranial course of any cranial nerve. It courses upwards in the pontine cistern between the brainstem and the clivus, either side of the basilar artery. It is crossed or 'bound down' to the brainstem close to its origin by the anterior inferior cerebellar artery. It may pierce the dura early in its upward course upon the clivus close to the inferior petrosal sinus (2 cm below the posterior clinoid process). On reaching the upper border of the apex of the petrous temporal bone, it crosses the inferior petrosal sinus from medial to lateral and changes direction sharply from a vertical to a horizontal course and runs forward beneath the *petrosphenoidal ligament (of Gruber)* and *superior petrosal sinus* (Fig. 1-54). The abducent nerve passes forward *within the cavernous sinus*, surrounded by venous spaces and suspended by fine connective tissue trabeculae. It lies *lateral* to the ascending portion of the internal carotid artery and then inferolateral to its horizontal portion (Fig. 1-9C and **eFig. 1-2B**). The abducent nerve enters the intraconal space of the orbit by passing within the tendinous ring (Figs 1-5 and 1-6).

Intraorbital course

The abducent nerve has a short intraorbital course. It enters the bulbar surface of the lateral rectus one-third of the way from its origin. This is the only muscle supplied by the abducent nerve.

Sensory endings in oculomotor nerves

The cell bodies of proprioceptive fibres in the extraocular muscles are located in the mesencephalic nucleus of the trigeminal nerve. The mesencephalic nucleus also receives proprioceptive terminals from neck and face musculature. The coordination of simultaneous movements of the head and eyes is dependent on conjugation of sensory (proprioceptive) information from the musculature of the neck and eyes, and input from cerebellar oculomotor centres (see Ch. 5).

TRIGEMINAL NERVE (CRANIAL NERVE V) (Fig. 1-55)

The trigeminal is the largest of the cranial nerves. It is the main sensory nerve of the head, supplying the mucous membranes of the oronasal cavities, middle ear, paranasal sinuses, skin of the face, the teeth, the cornea, the temporomandibular joint, and the dura of the anterior and middle cranial fossae (Fig. 1-55A,B). The dermatomes corresponding to its three major subdivisions and the cutaneous branches of the ophthalmic and maxillary nerves are shown in Figure 1-55A.

The mandibular division, besides its sensory component, supplies motor fibres to the muscles of mastication and also receives proprioceptive fibres from these muscles, together with the muscles of facial expression.

The trigeminal nerve arises from the brainstem in the posterior cranial fossa as a large *sensory root* and a small *motor root*. The sensory root consists of the central processes of pseudo-unipolar sensory neurones whose cell bodies lie in the large *trigeminal ganglion* located in the cavum trigeminale, a bony depression near the apex of the petrous temporal bone lined by evaginating dura mater from the edge of the tentorium cerebelli and roofed over by dura of the middle cranial fossa. The ganglion is partly surrounded by cerebrospinal fluid, which is continuous with the subarachnoid space of the posterior cranial fossa. The ganglion is homologous to a dorsal root or sensory ganglion of a spinal nerve. It is from the anterolateral convex surface of this flattened ganglion that the three named branches emerge: the ophthalmic (V_1), maxillary (V_2) and mandibular (V_3) (Fig. 1-55B).

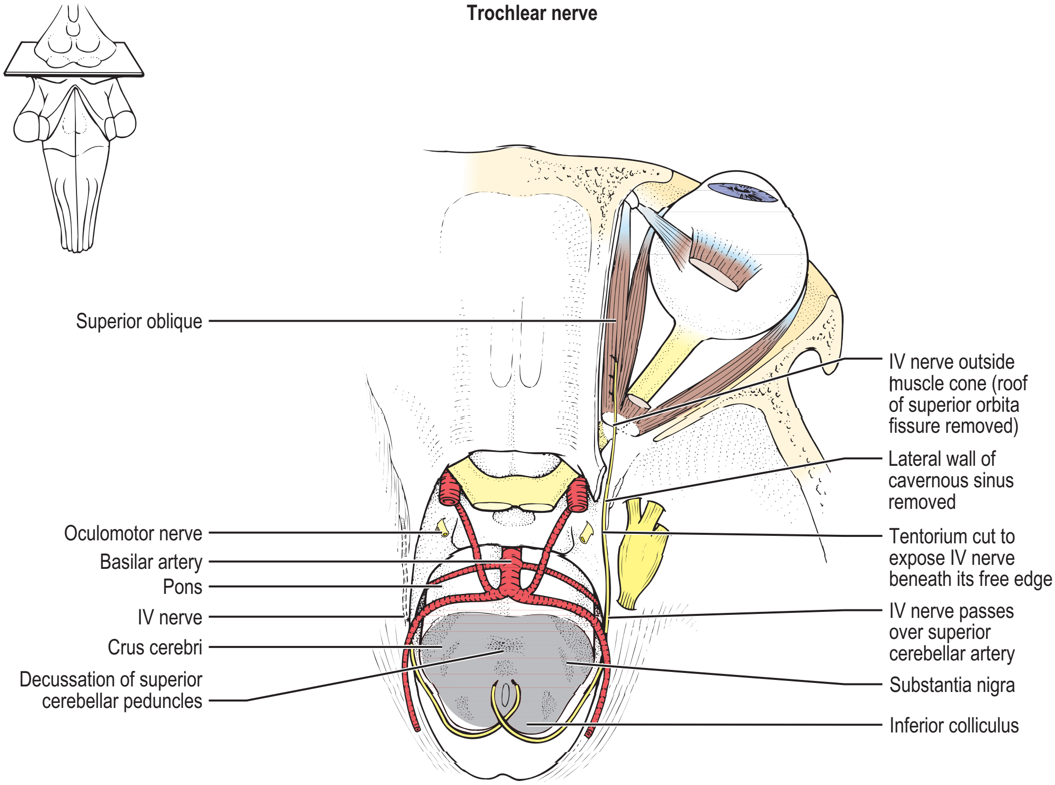
Trochlear nerve
Superior oblique
Oculomotor nerve
Basilar artery
Pons
IV nerve
Crus cerebri
Decussation of superior cerebellar peduncles
IV nerve outside muscle cone (roof of superior orbita fissure removed)
Lateral wall of cavernous sinus removed
Tentorium cut to expose IV nerve beneath its free edge
IV nerve passes over superior cerebellar artery
Substantia nigra
Inferior colliculus

FIGURE 1-53 Diagram summarizing the brainstem origin (inset shows level of section), intracranial, intracavernous and intraorbital course of the trochlear nerve.

The *ophthalmic division* or nerve (Fig. 1-56A) splits in the anterolateral portion of the cavernous sinus into three main branches: the *lacrimal, frontal* and *nasociliary.* The pathway and termination of these nerves are summarized in Figures 1-45–1-47, 1-55B and 1-56A.

The *maxillary nerve* (Figs 1-55B and 1-56B) passes through the *foramen rotundum* and spans the *pterygopalatine fossa* before entering the orbit through the *inferior orbital fissure* as the *infraorbital nerve.* It lies beneath the periorbita and is thus not truly an orbital content. The nerve passes forward from the inferior orbital fissure to the *infraorbital groove*, which becomes the *infraorbital canal*, the nerve eventually emerging through the *infraorbital foramen* (Fig. 1-57). Here it radiates out as a number of cutaneous branches supplying the lower eyelid, the nose, upper lip and cheek (Fig. 1-55A). The *zygomatic nerve*, a branch of the infraorbital, runs along in the inferior orbital fissure, beneath the periorbita, to the lateral wall of the orbit where it pierces the zygomatic bone as two branches, the *zygomaticotemporal* and *zygomaticofacial* nerves (both cutaneous). Traditionally, a communicating branch is described as passing up the lateral wall of the orbit to join the lacrimal nerve; however, the presence and importance of this nerve has been disputed. Other branches of the maxillary nerve, which pass through the *pterygopalatine ganglion* (without synapsing), supply the nasal cavity, upper alveolar arch and hard and soft palate (Fig. 1-55B). More details of these branches and those of the mandibular division are provided in standard anatomical texts.

The four parasympathetic ganglia of the head and neck (ciliary, pterygopalatine, otic and submandibular) are associated with the branches of the trigeminal (Fig. 1-55B). They are generally connected by short

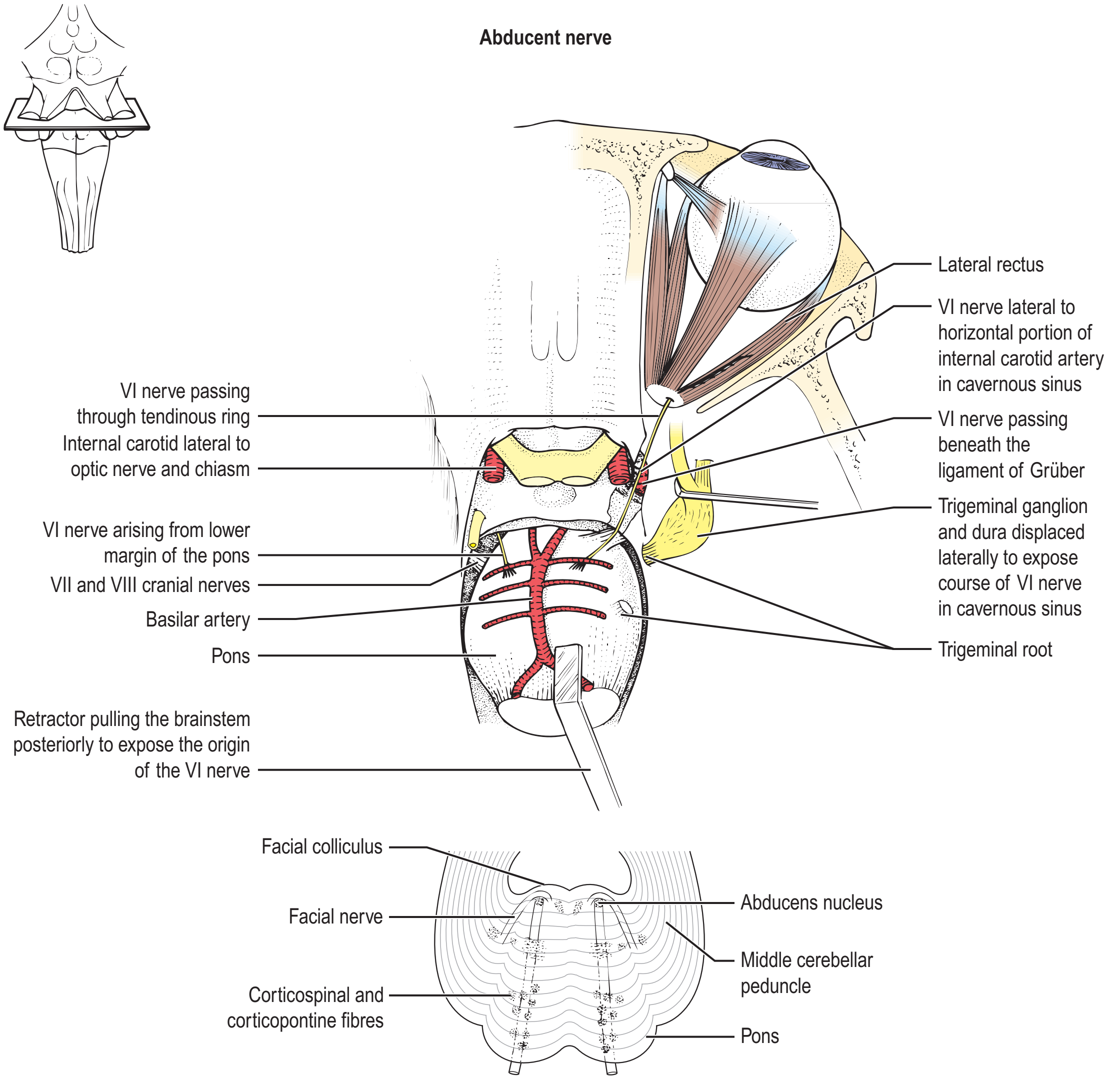
Abducent nerve
Lateral rectus
VI nerve lateral to horizontal portion of internal carotid artery in cavernous sinus
VI nerve passing through tendinous ring
Internal carotid lateral to optic nerve and chiasm
VI nerve passing beneath the ligament of Grüber
Trigeminal ganglion and dura displaced laterally to expose course of VI nerve in cavernous sinus
VI nerve arising from lower margin of the pons
VII and VIII cranial nerves
Basilar artery
Trigeminal root
Pons
Retractor pulling the brainstem posteriorly to expose the origin of the VI nerve
Facial colliculus
Abducens nucleus
Facial nerve
Middle cerebellar peduncle
Corticospinal and corticopontine fibres
Pons

FIGURE 1-54 Main diagram summarizes the intracranial, intracavernous and intraorbital course of the abducent nerve. The smaller lower diagram shows the neurohistological features at the level of the abducens nuclei (level of mid-pons on brainstem).

BOX 1-31 CLINICAL CORRELATES

Intracranial lesions affecting the abducent nerve

The abducent nerve is considered the ‘weakling’ of the cranial contents. It is very susceptible to damage in head injuries, such as fractures to the base of the skull or any type of expanding cerebral lesion, owing to its long intracranial course. If the brainstem is displaced downwards (due to raised intracranial pressure) the nerve may be compressed against the inferior cerebellar artery or severed where it bends sharply over the apex or crest of the petrous temporal bone. Within the *cavernous sinus* the nerve, not being protected by the dura of the lateral wall, is more susceptible than the other intracavernous nerves. For example, atheromatous changes in the internal carotid artery may compress the abducent nerve. Lesions of the abducent nerve result in paralysis of lateral rectus; thus, the patient is unable to abduct the eye and suffers esotropia (*internal strabismus*) as a result of the unopposed action of the medial rectus.

stalks containing pre- and postganglionic fibres, and the terminal fibres are distributed with the branches of the trigeminal – they may share the same perineural sheath for part of their course ('hitch-hikers'). Sympathetics may also hitch-hike with branches of the trigeminal for part of their course and be distributed to the territories of this large nerve.

Sensory nuclei of the trigeminal nerve

There are three sensory nuclei in the brainstem associated with the trigeminal nerve:

- *Mesencephalic nucleus.* This receives sensory (proprioceptive) information from muscles of mastication, muscles of facial expression and extraocular muscles. The peripheral processes are distributed with all three divisions of the trigeminal.
- *Pontine nucleus.* This is concerned with discriminative tactile information from the face. Fibres from all three divisions enter this nucleus.
- *Spinal nucleus.* This is continuous with the substantia gelatinosa of the posterior grey horn of the spinal cord. It is concerned with tactile, nociceptive and thermal information from the territories of the three divisions.

The afferent fibres terminating in the last two nuclei are the central processes of sensory neurones whose cell bodies are located in the trigeminal (Gasserian) ganglion. The peripheral processes terminate in appropriate sensory receptors in the territories of the three divisions.

FACIAL NERVE (CRANIAL NERVE VII)

The facial nerve contains a number of functional components (eTable 1-1). The *facial nerve proper* (containing *branchiomotor* or *special visceral efferents*) supplies the muscles of facial expression, stapedius (small muscle in middle ear), stylohyoid, and the posterior belly of the digastric. The second component, the *nervus intermedius*, contains secretomotor or *parasympathetic* fibres (*general visceral efferent*), which synapse in the pterygopalatine ganglion and supply the lacrimal gland and choroid in addition to other glands in and around the nose and mouth. The nerve also transmits taste fibres (*special visceral afferents*) from the anterior two-thirds of the tongue. The extracranial branch of cranial nerve VII exits the skull through the *stylomastoid foramen* and pierces the parotid gland to emerge at the anterior border of that gland (Fig. 1-58). The facial nerve is important to the eye and orbit primarily because of its parasympathetic supply to the lacrimal gland (and some intraocular branches) and its motor supply to the periorbital facial muscles (especially orbicularis oculi). It is the most frequently paralysed of all peripheral nerves.

There are three reflex arcs in the brainstem involving the facial nerve, of which the corneal reflex is an important clinical test (Table 1-5).

Ocular appendages (adnexa)

MUSCLES OF THE EYELIDS AND ADJACENT FACE

The muscles that are primarily responsible for movement of each eyelid include the *orbicularis oculi* (a muscle of facial expression), which is responsible for lid closure, and the *levator palpebrae superioris* (an extraocular muscle), which raises the lid.

Other *muscles of facial expression* of primary interest around the eye and eyelids are the *corrugator supercilii* and *occipitofrontalis.*

Orbicularis oculi (Fig. 1-58)

Shape. The orbicularis oculi muscle is a broad, flat, sheet of skeletal muscle with orbital, palpebral and lacrimal portions. The circular orientation of the fibres is a reflection of the sphincter-like function of this muscle. The *orbital portion* arises from the medial palpebral ligament and the adjacent orbital margin

BOX 1-32 CLINICAL CORRELATES

Trigeminal neuralgia ('tic douloureux')

This is a condition characterized by excruciating pain in the territory of one or more of the divisions of the trigeminal nerve. Thus, the clinician should be familiar with structures supplied by each division. Many causes have been suggested, including osteitis of the petrous temporal bone or compression of the root or ganglion in the cavum trigeminale by enlarged or engorged vessels; however, in many cases the aetiology is unknown. The territory of the maxillary nerve is most frequently involved, then the mandibular, and less commonly the ophthalmic nerve. Three commonly performed surgeries are glycerol rhizotomy, stereotactic radio-surgery (Gamma Knife) and endoscopic vascular decompression.

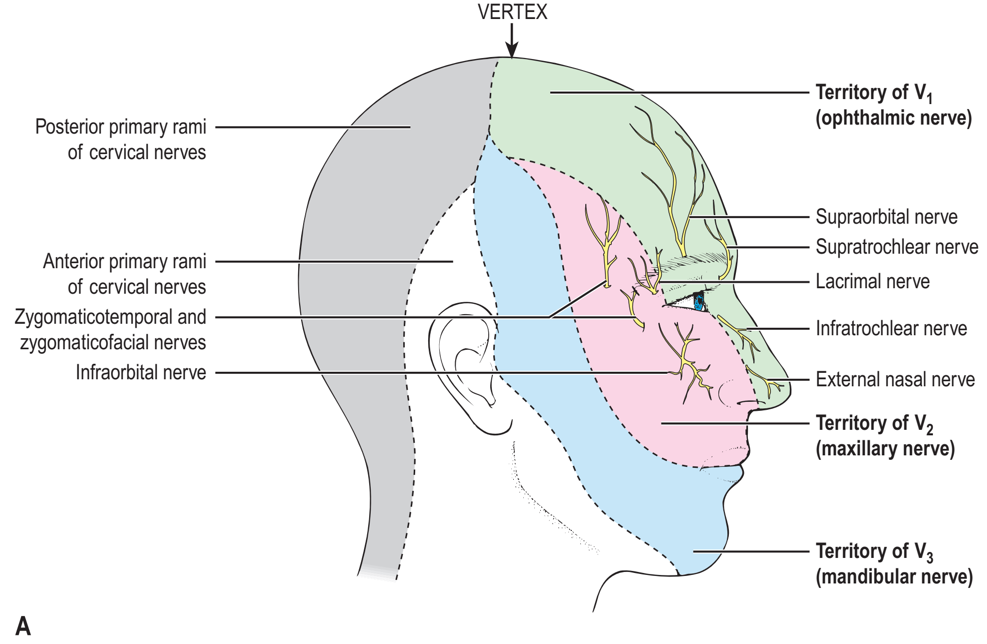
VERTEX
Posterior primary rami of cervical nerves
Anterior primary rami of cervical nerves
Zygomaticotemporal and zygomaticofacial nerves
Infraorbital nerve
Territory of V1 (ophthalmic nerve)
Supraorbital nerve
Supratrochlear nerve
Lacrimal nerve
Infratrochlear nerve
External nasal nerve
Territory of V2 (maxillary nerve)
Territory of V3 (mandibular nerve)
A

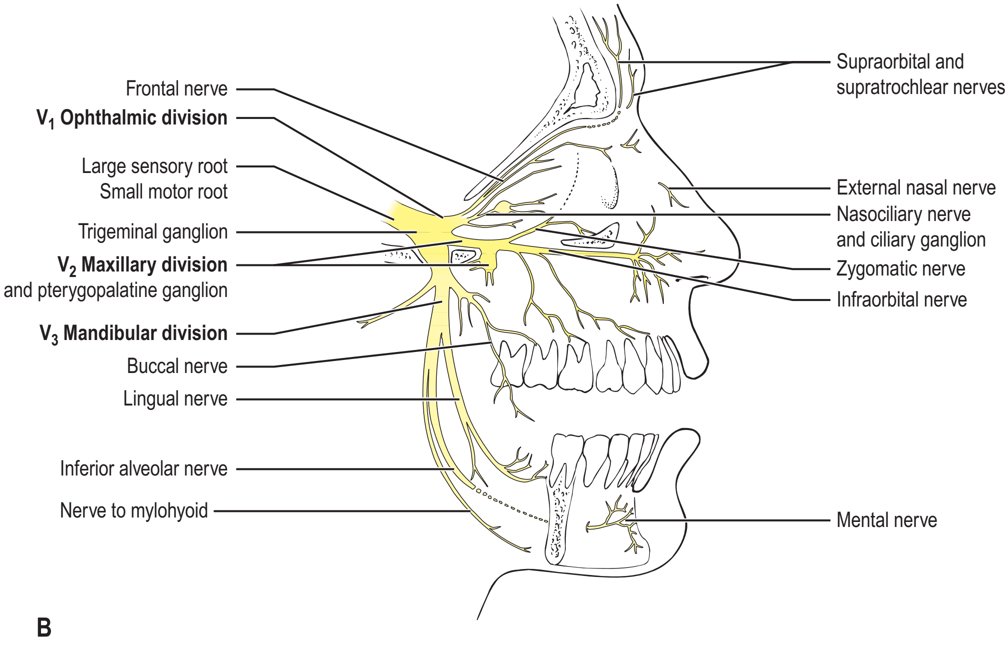
Frontal nerve
V1 Ophthalmic division
Large sensory root
Small motor root
Trigeminal ganglion
V2 Maxillary division and pterygopalatine ganglion
V3 Mandibular division
Buccal nerve
Lingual nerve
Inferior alveolar nerve
Nerve to mylohyoid
Supraorbital and supratrochlear nerves
External nasal nerve
Nasociliary nerve and ciliary ganglion
Zygomatic nerve
Infraorbital nerve
Mental nerve
B

FIGURE 1-55 (**A**) Sensory 'map' of the head and neck. Note the limits of the territories of the skin of the face and scalp supplied by the three divisions of the trigeminal nerve (ophthalmic, maxillary and mandibular). (**B**) Stylized diagram showing the origin of the motor and sensory roots of the trigeminal, the position of the pterygopalatine and ciliary ganglia, and the territories of some of the major branches of the three divisions.

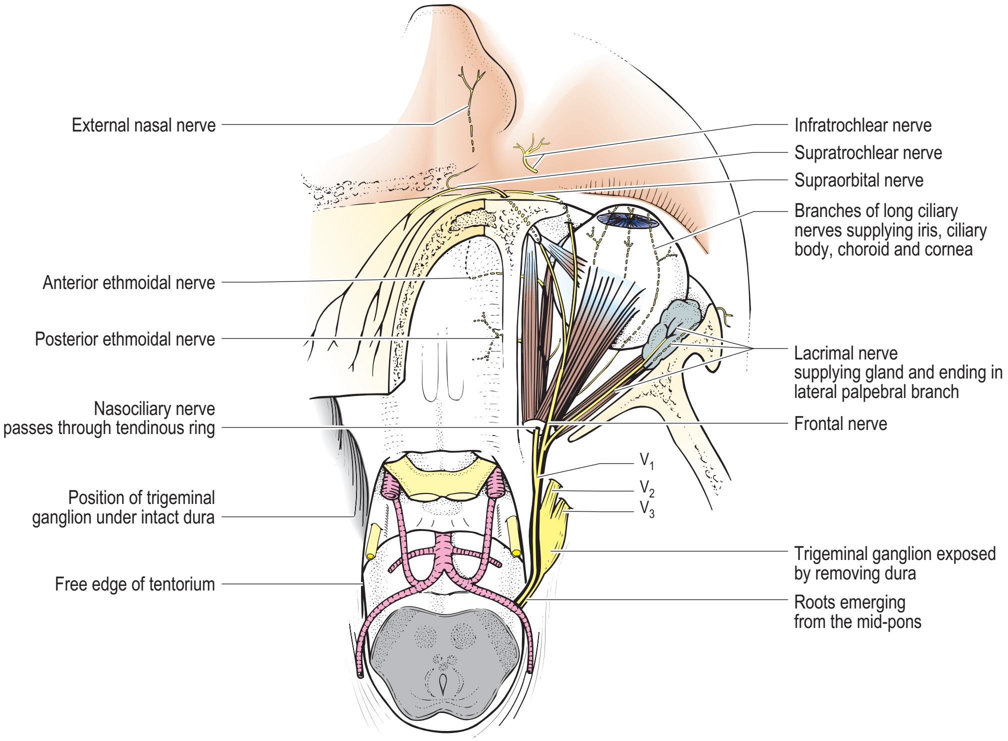
External nasal nerve
Anterior ethmoidal nerve
Posterior ethmoidal nerve
Nasociliary nerve passes through tendinous ring
Position of trigeminal ganglion under intact dura
Free edge of tentorium
Infratrochlear nerve
Supratrochlear nerve
Supraorbital nerve
Branches of long ciliary nerves supplying iris, ciliary body, choroid and cornea
Lacrimal nerve supplying gland and ending in lateral palpebral branch
Frontal nerve
V1
V2
V3
Trigeminal ganglion exposed by removing dura
Roots emerging from the mid-pons

FIGURE 1-56 Diagram summarizing the origin, intracranial, intracavernous and intraorbital course of the ophthalmic nerve and its three main branches.

- *The frontal nerve* passes obliquely above the levator palpebrae superioris and divides into the supraorbital nerve and the supratrochlear nerve.
- The supraorbital nerve passes through the supraorbital notch or foramen and ascends in the subcutaneous tissue of the forehead to supply the skin.
- The supratrochlear nerve passes along the medial wall of the orbit to pass above the trochlea and supply the skin at the root of the nose, upper eyelid and the medial part of the forehead.
- The *lacrimal nerve* travels forward along the upper border of the lateral rectus and sends fibres to the gland before ending as *lateral palpebral branches* to the conjunctiva and skin of this region.
- Branches of the *nasociliary nerve* include the *sensory root* to the ciliary ganglion (from which the short ciliary nerves emerge), *long ciliary nerves, posterior ethmoidal nerve* (supplies ethmoidal and sphenoidal sinuses), the *anterior ethmoidal nerve* and *infratrochlear nerves*.
- The anterior ethmoidal briefly re-enters the cranial cavity at the cribriform plate (beneath the dura) before piercing the bone to exit the cavity to terminate as the *medial* and *lateral internal nasal branches* (supply the nasal cavity), the latter of which ends as the *external nasal branch* which supplies the skin on the lower half of the nose.
- The infratrochlear nerve runs along close to the medial orbital wall, passes beneath the trochlea and supplies the skin at the angle of the eye and upper part of the skin of the nose.

(Fig. 1-58, inset). Its fibres run circumferentially in an elliptical fashion around and beyond the orbital margin. Most pass round the lateral orbital margin without interruption, although some fibres (known as *depressor supercilii*) are inserted into the skin and connective tissue of the eyebrow. The *palpebral part* is an extremely thin muscle that originates from the *medial palpebral ligament.* Its fibres pass laterally within the eyelid anterior to the orbital septum and tarsal plate (see below), and interlace to form the *lateral palpebral raphe.* The small *lacrimal component* of the muscle passes deep to the medial palpebral ligament and is attached to the posterior lacrimal crest (behind the lacrimal sac) as two muscle slips (upper and lower).

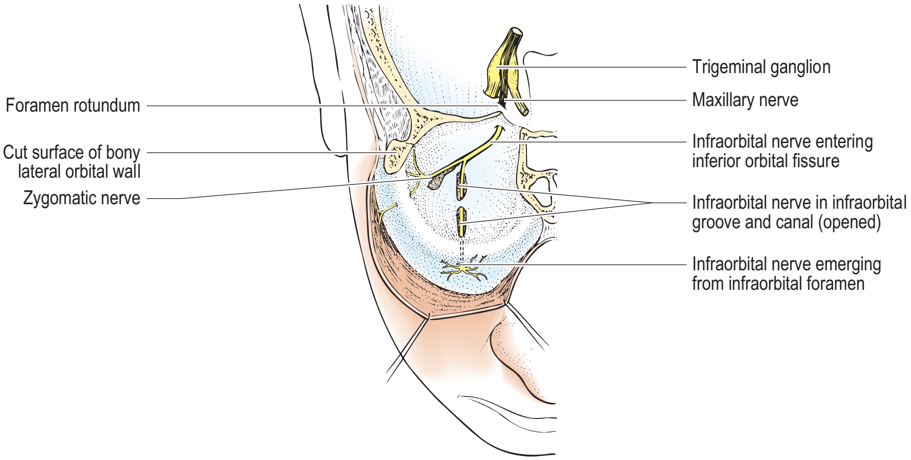
Foramen rotundum
Cut surface of bony lateral orbital wall
Zygomatic nerve
Trigeminal ganglion
Maxillary nerve
Infraorbital nerve entering inferior orbital fissure
Infraorbital nerve in infraorbital groove and canal (opened)
Infraorbital nerve emerging from infraorbital foramen

FIGURE 1-57 Diagram of the orbital floor and the course of the maxillary and infraorbital nerves.

BOX 1-33 CLINICAL CORRELATES

Lesions of the facial nerve

Supranuclear lesions – caused by vascular stroke in which descending corticonuclear and corticospinal fibres are damaged in the internal capsule. The upper facial motor nucleus (supplying upper half of facial muscles) receives input from 'face' areas of both the ipsilateral and contralateral motor cortices. The lower part of the facial nucleus has only contralateral input. The effect of a stroke, therefore, is to cause contralateral paralysis or weakness of the limbs and lower face. The upper face survives because of the bilateral supranuclear supply to the upper part of the facial nucleus.

Nuclear lesions – direct damage to the facial nucleus, such as thrombosis of the pontine branches of the basilar artery, results in complete paralysis of structures supplied by the facial nerve (and abducent nerve; see Fig. 1-54) together with motor weakness of the limbs on the opposite side (owing to pyramidal decussation occurring below this level).

Infranuclear lesions – Bell's palsy involves direct neuritis of the facial nerve in the bony canal within the temporal bone and results usually in complete facial paralysis. The patient is unable to move the lips (saliva and food drools from the corner of the mouth), eyebrows or close the eyelids (lids may be lax, causing epiphora), and suffers hyperacusis (due to paralysis of the stapedius). Some patients may also have reduced lacrimal and salivatory secretions and loss of taste to the anterior two-thirds of the tongue. Other causes of infranuclear lesions include multiple sclerosis, tumours of the cerebellopontine angle (acoustic neuromas), middle ear disease and tumours of the parotid gland.

These fibres are inserted laterally into the tarsi close to the lacrimal canaliculi; they help draw the eyelids and lacrimal papillae medially and in addition dilate the lacrimal sac during blinking. This helps to suck tears into the lacrimal punctum from the lacus lacrimalis.

Nerve supply. Temporal and zygomatic branches of the facial nerve.

Action. The *orbital* portion, owing to its elliptical form and medial attachments, acts like a purse string, drawing the skin of the forehead, temple, cheek and orbital margin towards the medial angle of the orbit, firmly closing the lids (for example, when in very bright light). The *palpebral* portion of orbicularis oculi can act under both voluntary and involuntary control to close the eyelids during normal blinking (and sleeping). This blinking reflex (Table 1-5) is essential to the integrity of the ocular tear film and function of the cornea.

Corrugator supercilii

This is a small pyramidal muscle at the medial aspect of the eyebrow (Fig. 1-58), beneath the occipitofrontalis and orbicularis oculi. It draws the eyebrow downwards and medially (frowning), producing vertical skin furrows on the forehead. It assists in

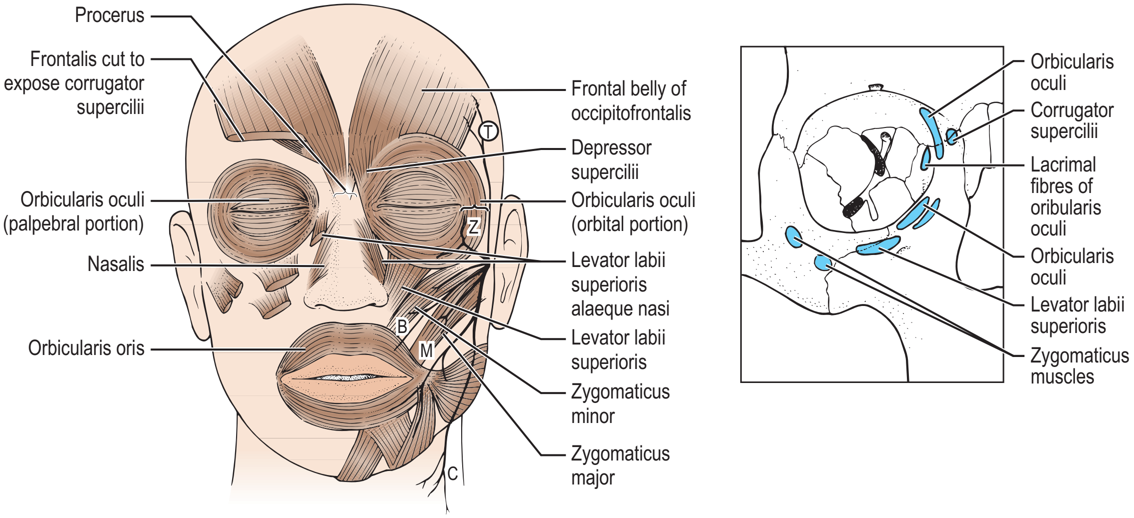
Procerus
Frontalis cut to expose corrugator supercilii
Orbicularis oculi (palpebral portion)
Nasalis
Orbicularis oris
Frontal belly of occipitofrontalis
Depressor supercilii
Orbicularis oculi (orbital portion)
Levator labii superioris alaeque nasi
Levator labii superioris
Zygomaticus minor
Zygomaticus major
T
Z
B
M
C
Orbicularis oculi
Corrugator supercilii
Lacrimal fibres of oribularis oculi
Orbicularis oculi
Levator labii superioris
Zygomaticus muscles

FIGURE 1-58 The muscles of facial expression, particularly those of relevance to the eye and orbit. Inset shows the bony origins of some of these muscles around the orbit. Some of orbicularis oculi and frontalis muscle fibres have been removed on the right side of the face to display underlying musculature. The extracranial motor branches of the facial nerve that supply the muscles are shown on the right: T, temporal or frontal branches; Z, zygomatic branches; B, buccal branches; M, marginal mandibular branch; C, cervical branches.

protecting the eyes in bright light. It is supplied by small subdivisions of the temporal branch of the facial nerve.

Occipitofrontalis

This fibromuscular layer covers the dome of the skull from the eyebrows to the nuchal lines. It consists of two occipital bellies posteriorly attached via a thick fibrous layer, the *galea aponeurotica*, to the two frontal bellies. Only the frontal part is of relevance to the eye and orbit. Its fibres form a thin quadrangular sheet that is attached to the superficial fascia above the eyebrows. The medial fibres are continuous with procerus (Fig. 1-58), the intermediate fibres with corrugator supercilii and orbicularis oculi, and the lateral fibres with orbicularis oculi. The frontal belly is supplied by the temporal branches of the facial nerve. Upon contraction it draws the scalp backwards and elevates the eyebrows, causing transverse wrinkles on the scalp as in expressions of surprise, horror, fright, or when glancing upwards.

Levator palpebrae superioris

This muscle lies within the orbit and is responsible for opening the eyelids and, upon relaxation, allows lid closure due to gravity. The muscle has its origin from the lesser wing of sphenoid, above and in front of the optic foramen, blending with the origin of superior rectus. The muscle belly passes horizontally forward above superior rectus, close to the orbital roof (Fig. 1-45). Behind the orbital margin it curves downwards into the lid where it becomes aponeurotic. The aponeurosis fans out, on either side, to form medial and lateral horns, which extend the whole width of the eyelid. The levator palpebrae superioris inserts into the skin of the upper lid (causing the horizontal palpebral sulcus or furrow) and the anterior surface of the tarsal plate (see Fig. 1-60C). The lateral horn of the aponeurosis forms the *lateral palpebral ligament*, which inserts into the lateral orbital tubercle, and the medial horn forms the *medial palpebral ligament*, which inserts into the frontolacrimal suture. The levator palpebrae superioris is supplied by the superior division of the oculomotor nerve (see p. 69 and Fig. 1-52). Upon contraction it elevates the upper lid, thereby opening the palpebral fissure. On the inferior aspect of the levator palpebrae superioris is a small band of smooth muscle, the *superior tarsal* or *Müller's muscle*. It is attached anteriorly to the upper surface of the tarsal plate and conjunctival fornix (see

TABLE 1-5 **Reflexes involving the facial nerve**

	Corneal reflex	Blinking to light or fast-approaching object	Blinking to noise
Receptor	Sensory ending in corneal epithelium	Retina	Cochlea
Afferent pathway	Long ciliary nerves, nasociliary nerve, ophthalmic nerve	Optic nerve	Vestibulocochlear nerve
First synapse	Spinal nucleus of trigeminal	Superior colliculus	Inferior colliculus
Second synapse	Facial nucleus	Facial nucleus	Facial nucleus
Efferent pathway	Temporal and zygomatic branches of facial nerve	Temporal and zygomatic branches of facial nerve	Temporal and zygomatic branches of facial nerve
Effector muscle	Orbicularis oculi	Orbicularis oculi	Orbicularis oculi

Fig. 1-60C). It has a sympathetic innervation and upon contraction assists the levator in elevating the eyelid. While there is no equivalent muscle to the levator in the lower lid, there is a small group of smooth muscle fibres (inferior tarsal muscle) that originate from the fascial sheath of the inferior rectus and insert into the lower tarsus.

THE EYELIDS

The eyelids are thin curtains of skin, muscle, fibrous tissue and mucous membrane that serve to protect the eyes from injury and excessive light and also to distribute tears over the ocular surface during blinking. The upper lid, when open, normally just overlaps the corneoscleral junction, and it is this lid that undergoes most displacement during eyelid closure, the lower lid moving only minimally during normal blinking.

On external examination (Fig. 1-59) each lid is seen to be divided into *orbital* and *tarsal portions* by a horizontal *palpebral sulcus*, which is most evident on the upper lid. The upper lid is limited superiorly by the eyebrow, whereas the lower lid blends with the skin of the cheek. The upper and lower lids meet at the *medial* and *lateral canthi* or angles, and are separated from one another by an elliptical opening – the *palpebral fissure*. The lateral canthus is an acute angle (60°) and lies close to the eyeball; the medial canthus is rounded, elongated medially, and lies 6 mm from the eyeball. It is separated from the eye by a triangular zone, the *lacus lacrimalis* (lake of tears), in which a small raised red swelling, the *curuncula lacrimalis*, is situated. There are obvious racial differences in the shape and form of the eyelids and canthi, the most conspicuous being the vertical *epicanthal fold* in Oriental and Asian races.

The *eyelid margins* (Figs 1-59 and 1-60) are approximately 30 mm in length, 2 mm in thickness, and relatively square in profile along most of their length, except the medial one-sixth, which is rounded and lacks eyelashes. Eyelashes are modified, thick, stiff hairs that occur as double or triple rows close to the anterior lid margin. They curl away from the lashes of the opposite lid. Notable features on the lid margins include:

- *Lacrimal puncta* located at the medial ends of the upper and lower lids. These drain tears from the lacus lacrimalis. The puncta are more easily identified if tension is placed on the lids, causing the papillae to blanch.
- Openings of *tarsal (meibomian) glands* are visible to the naked eye as a row of minute openings on the lid margin posterior to the eyelash follicles. There are around 30 in the upper lid and slightly fewer in the lower lid.
- The *skin/conjunctival transition zone* mucocutaneous junction occurs at the level of the opening of the tarsal glands.
- The *grey line* marks the anterior boundary of the tarsal plate and is a useful landmark for surgical incisions.

The histological structure of the eyelid is summarized in Figures 1-60 and 1-62. Note that the fibrous framework of the lids is formed by the *orbital septum* arising from the orbital margin (Fig. 1-43) and the *tarsal plates*. The *tarsal plates* are modified regional thickenings of the orbital septum that provide rigidity to the upper and lower lids and separate the orbit and its contents from the lids.

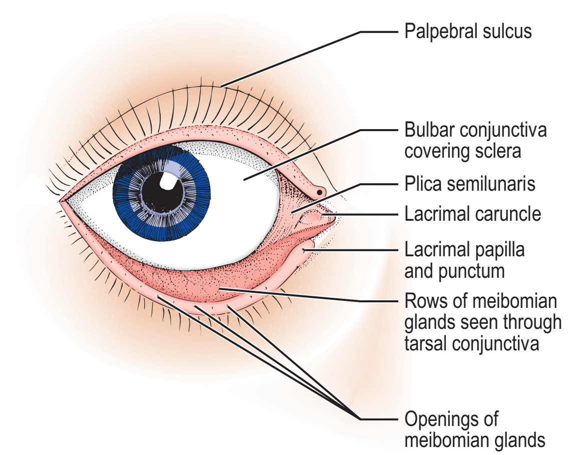
Palpebral sulcus
Bulbar conjunctiva covering sclera
Plica semilunaris
Lacrimal caruncle
Lacrimal papilla and punctum
Rows of meibomian glands seen through tarsal conjunctiva
Openings of meibomian glands

FIGURE 1-59 Surface anatomy of the eye and eyelids. The bottom lid has been slightly everted to reveal the inner surface of the lid, lacrimal papillae and puncta and the openings of the meibomian glands along the lid margin.

BOX 1-34 AGEING CHANGES

Herniation of orbital fat through weakened regions of the septum occurs in the elderly, producing bulging, sagging lids (blepharochalasis).

The tarsal plates consist of dense fibrous connective tissue and are approximately 25–30 mm from the medial to lateral borders. They are 1 mm thick and the upper plate is greater in height (10–12 mm) than the lower plate (5 mm). They are attached at either end via their continuations, the *medial* and *lateral palpebral ligaments*. Skin moves freely over their anterior surface, although the conjunctiva is tightly bound to the posterior surface. External examination of an everted eyelid reveals vertical rows of yellowish tarsal glands (see inset, Box 1-35). They are embedded in the matrix of the tarsal plate and consist of modified sebaceous glands. Histologically the acinar cells are replete with lipid droplets that are secreted in a holocrine manner on to the eyelid margin, which functions to retain tears in the conjunctival sac and contributes to the lipid layer of the precorneal tear film.

The blood supply and nerve supply of the lids and surrounding areas are summarized in Figure 1-61. Lymphatics drain to the superficial parotid or submandibular lymph nodes. The pretarsal portion derives its arterial supply from the superficial temporal and facial arteries (branches of the external carotid), while the post-tarsal portion is supplied by branches of the ophthalmic artery (branch of internal carotid artery) (see Fig. 1-50). Venous drainage follows a similar pattern to the arterial supply (see p. 70 for consideration of anastomoses between the internal and external carotid arteries). The post-tarsal venous drainage is via the ophthalmic veins to the cavernous sinus.

BOX 1-35 CLINICAL CORRELATE

Chalazion, hordeolum and ectropion

A *chalazion* is a localized, painless swelling in the lid due to obstruction and chronic inflammation of a tarsal gland (see Fig. 1-60B).

A hordeolum (sty) may be an acute infection of an eyelash follicle or its sebaceous gland, infection of a ciliary sweat gland (external hordeolum) or acute infection of a tarsal gland (internal hordeolum).

Ectropion is drooping of the lower lid owing to paralysis of orbicularis oculi. Because of paralysis of the fibres of orbicularis that enclose the lacrimal sac, the puncta no longer suck up tears, which may thus pass over the lid margin (*epiphora*).

Movements of the eyelids

The eyelids close as a result of the action of the palpebral fibres of the orbicularis oculi and relaxation of the levator palpebrae superioris. Opening of the lids occurs via the pull of levator palpebrae superioris on the skin, tarsal plate and forniceal conjunctiva. The nerve supply to these muscles is from three sources: orbicularis oculi – the facial nerve (VII); levator palpebrae superioris – the oculomotor nerve (III); while its smooth muscle component, superior tarsal (Müller's) muscle, is supplied by sympathetic nerves. The latter is important in times of fear or excitement when the width of the palpebral fissure is further increased.

THE CONJUNCTIVA

The conjunctiva (Figs 1-60A,C and 1-62) is a thin translucent mucous membrane that derives its name from the fact that it attaches the eyeball to the lids. It consists of a superficial conjunctival epithelium overlying a loose connective tissue stroma. The epithelium is continuous with the corneal epithelium at the limbus and with the skin at the mucocutaneous

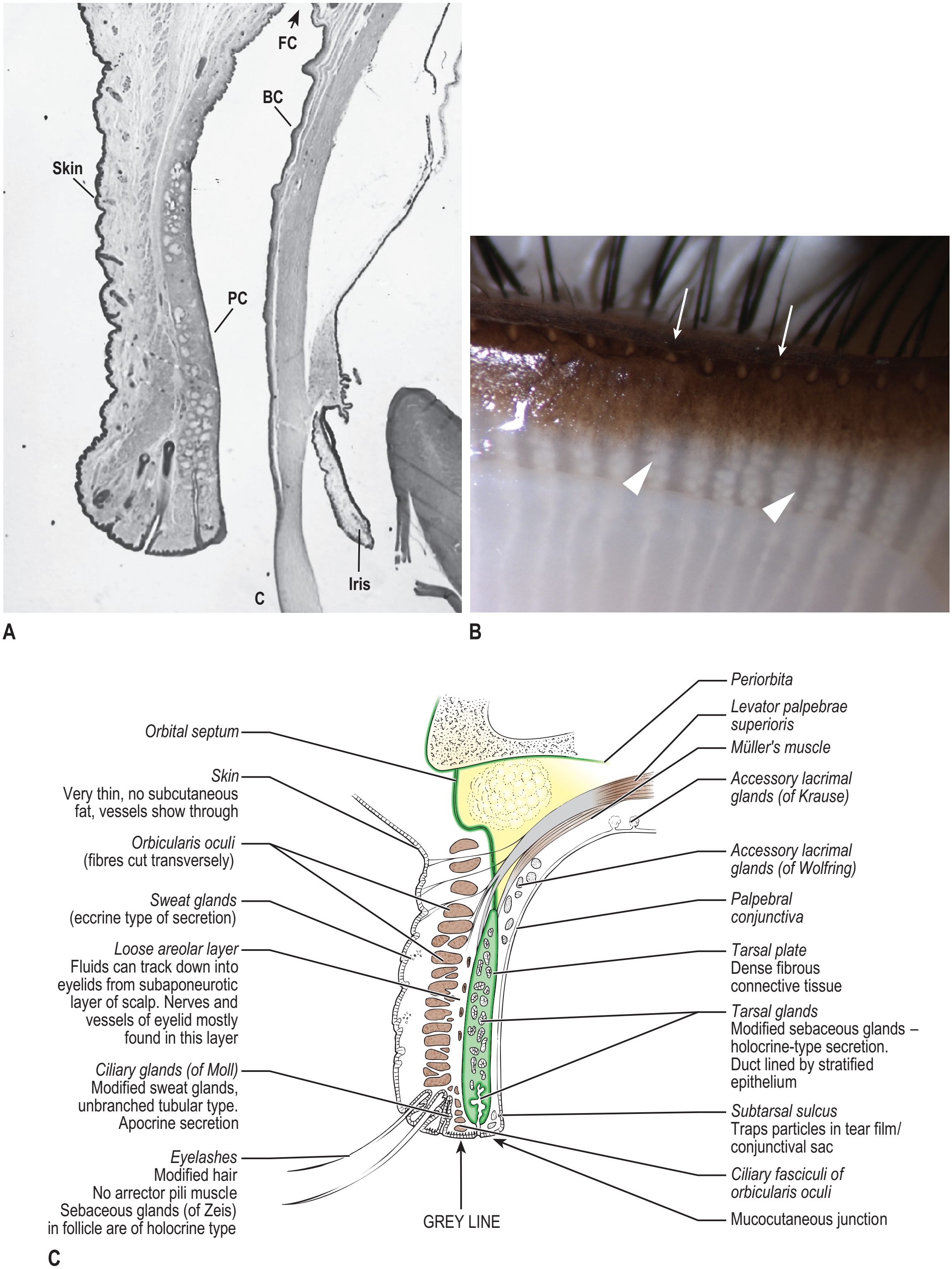
FC
BC
Skin
PC
Iris
C
A
B
Periorbita
Levator palpebrae superioris
Müller's muscle
Accessory lacrimal glands (of Krause)
Accessory lacrimal glands (of Wolfring)
Palpebral conjunctiva
Tarsal plate
Dense fibrous connective tissue
Tarsal glands
Modified sebaceous glands – holocrine-type secretion. Duct lined by stratified epithelium
Subtarsal sulcus
Traps particles in tear film/ conjunctival sac
Ciliary fasciculi of orbicularis oculi
Mucocutaneous junction
GREY LINE
Orbital septum
Skin
Very thin, no subcutaneous fat, vessels show through
Orbicularis oculi
(fibres cut transversely)
Sweat glands
(eccrine type of secretion)
Loose areolar layer
Fluids can track down into eyelids from subaponeurotic layer of scalp. Nerves and vessels of eyelid mostly found in this layer
Ciliary glands (of Moll)
Modified sweat glands, unbranched tubular type. Apocrine secretion
Eyelashes
Modified hair
No arrector pili muscle
Sebaceous glands (of Zeis) in follicle are of holocrine type
C

FIGURE 1-60
For legend see opposite page.

BOX 1-36 CLINICAL CORRELATES

Horner syndrome

Because of the complex neuroanatomy of the sympathetic nervous system, Horner syndrome, which is characterized by classical symptoms of unilateral ptosis, miosis and dry facial skin (anhidrosis) and blushing on the affected side, may result from a wide variety of lesions in the central and peripheral nervous system. These include iatrogenic interruption of the sympathetic chain in the neck, dissection of the internal carotid artery, cervical disk dislocation and the lysis of the first rib affecting the stellate ganglion associated with Pancoast tumour. Other symptoms may include heterochromia and enophthalmos, although the latter is debatable in humans.

junction on the lid margin. The conjunctiva is reflected from the anterior portion of the sclera at the superior and inferior fornices on to the tarsal surface of the eyelids. Thus, when the lids are closed a potential sac, the *conjunctival sac*, is formed. The volume of this sac is approximately 7 μL, which explains the tendency for eye drops from commercial dispensers (volume 50–70 mL) to overflow unless the lower lid is held away from the globe. The conjunctiva is responsible for the production of the mucous component of the tear film and, in common with other mucous membranes, has a variety of immunological defence mechanisms that protect the ocular surface from infection (see Ch. 7). For descriptive purposes the conjunctiva can be divided into three main regions.

Palpebral conjunctiva (Fig. 1-60A,C)

This part lines the inner surfaces of the eyelids. It is tightly bound to the tarsal plate, the subepithelial connective tissue stroma being thin in this region. The lacrimal puncta open on to the palpebral conjunctiva; thus the conjunctival epithelium is continuous with the lining of the inferior meatus of the nasal cavity, which explains the manner in which infection spreads between these two sites. A small *subtarsal sulcus*, close to the lid margin, is important in trapping and removing foreign particles and debris on the ocular surface.

BOX 1-37 CLINICAL CORRELATES

Allergic responses in the conjunctiva

Clinical examination of the everted lid for signs of ocular allergy and infection is a common procedure. Two major types of abnormal accumulations of immune cells may occur: *follicles*, which are similar to mucosal-associated lymphoid follicles elsewhere and consist primarily of lymphocytes; and *papillae*, which are focal aggregates of chronic inflammatory cellular infiltrates and accompanying vascular changes. These are usually associated with allergic conditions and irritation of the ocular surface, such as in contact lens wear.

Forniceal conjunctiva (Fig. 1-60A)

The superior and inferior fornices are continuous at the medial and lateral canthi, thus forming a circular *cul de sac*. It is into the superolateral fornix that the ducts of the main lacrimal gland and the bulk of accessory lacrimal glands empty. The forniceal conjunctiva is loosely attached to the fascial sheaths of levator palpebrae superioris and the rectus muscles, and thus moves slightly with the eye during contraction of these muscles.

Bulbar conjunctiva

The white sclera is visible through the normal translucent bulbar conjunctiva (see Figs 1-20A and 1-59). It clothes the anterior part of the eyeball including the extraocular muscle insertions and Tenon's capsule. Near the limbus the conjunctiva is tightly bound to the globe, but further from the limbus there is a loose episcleral tissue layer (Fig. 1-62A) within which lies the pericorneal vascular plexus (Fig. 1-62D). These vessels can become dilated and conspicuous as a result of physical and inflammatory stimuli.

There are two specializations of the conjunctiva in the medial fornix. First, the semilunar fold (*plica semilunaris*), which is probably homologous to the nictitating membrane of lower mammals and many non-mammalian vertebrates. It is highly vascular and

FIGURE 1-60 (**A**) Histological preparation of the upper eyelid, conjunctiva and anterior segment: PC, palpebral conjunctiva; FC, forniceal conjunctiva; BC, bulbar conjunctiva; C, cornea. (**B**) Primate upper eyelid as viewed from the inner aspect to reveal the rows of tarsal glands (arrowheads) and their openings (arrows) on the lid margin. (**C**) Schematic diagram of the upper eyelid in longitudinal section (sagittal plane). Original magnification: **A**, ×15. (Part A courtesy of W.R. Lee and Springer-Verlag.)

rich in goblet cells and interstitial immunocompetent cells. The function of this loose fold may be to facilitate lateral movement of the eye. Second, the *caruncle* (*caruncula lacrimalis*) is a highly vascular nodule of modified skin in the medial corner of the eye containing large nests of accessory lacrimal and sebaceous glandular tissue.

Structure of the conjunctiva

Histologically the conjunctival epithelium varies in structure, depending on location, from a stratified squamous non-keratinizing epithelium (close to the lid margin) to a stratified columnar epithelium (bulbar). In general it consists of between two and seven layers of epithelial cells that are organized into three main types: basal, intermediate and superficial. There is no

BOX 1-38 CLINICAL CORRELATES

Systemic disease evident in the conjunctiva

The conjunctiva may manifest signs of several important systemic diseases: pathognomonic signs are present in sickle cell anaemia (comma sign), jaundice (scleral icterus (yellowing)) and vitamin A deficiency (Bitot's spots).

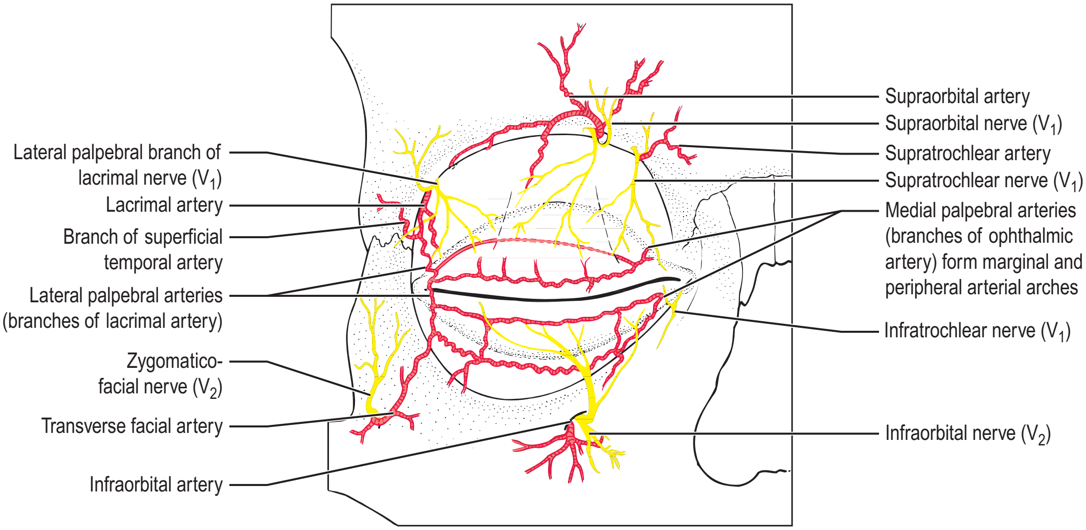
Lateral palpebral branch of lacrimal nerve (V_1)
Lacrimal artery
Branch of superficial temporal artery
Lateral palpebral arteries (branches of lacrimal artery)
Zygomatico-facial nerve (V_2)
Transverse facial artery
Infraorbital artery
Supraorbital artery
Supraorbital nerve (V_1)
Supratrochlear artery
Supratrochlear nerve (V_1)
Medial palpebral arteries (branches of ophthalmic artery) form marginal and peripheral arterial arches
Infratrochlear nerve (V_1)
Infraorbital nerve (V_2)

FIGURE 1-61 Summary of the blood supply and sensory nerve supply of the eyelids and adjacent areas from cutaneous branches of the ophthalmic (V_1) and maxillary (V_2) divisions of the trigeminal nerve.

FIGURE 1-62 Histology of the conjunctiva. (**A**) Low-power light micrograph of human conjunctiva (periodic acid–Schiff (PAS) stain) showing the irregular nature and goblet cell content (purple and red PAS$^+$ profiles) of the epithelium (compare to corneal epithelium; Figs 1-12, 1-14 and 1-16). Note the accumulations of lymphoid cells in the highly vascular connective tissue stroma, a common feature in eyes of elderly patients. CT, subepithelial connective tissue or connective tissue stroma. (**B**) Electron micrograph of goblet cells (GC) in the conjunctival epithelium: Ep, epithelium. (**C**) 'Plan view' of $^+$ intraepithelial dendritic cells (sometimes called Langerhans cells (LC)) in the limbal/conjunctival epithelium in wholemount preparation from Cx3cr1-GFP transgenic mouse (green – Cx_3cr1^+ myeloid-derived cells; red – MHC class II; blue – DAPI-stained nuclei). Note the Langerhans cells are Cx_3cr1^+ MHC class II$^+$. The Z-profile shows how some of the cell processes act as 'periscopes' and project towards the superficial aspect of the conjunctival/limbal epithelium. (**D**) Toluidine blue-stained conjunctival whole mount illustrating the orientation and distribution of mast cells around limbal vessels (V) and in the bulbar conjunctiva (Conj) where they are more rounded. (**E**) Toluidine blue-stained semi-thin resin section of the limbal region showing a mast cell (MC) adjacent to a large venule (V). (**F**) Primate conjunctiva (histological preparation, H&E) showing melanocytes in the basal layer of the conjunctival/limbal epithelium. Note the intraepithelial melanin granules throughout the conjunctival epithelial layers. (**G**) Melanocytes as seen in a limbal whole mount preparation. Note their highly dendriform shape (arrows) and how they form a halo (dotted lines) of melanin granules within the adjacent epithelial cells ('epithelial-melanin unit'). Original magnifications: **A**, × 150; **B**, × 3000; **C**, × 200; **D**, × 100; **E**, × 650; **F**, × 150; **G**, × 160. (Part B courtesy of W.R. Lee.)

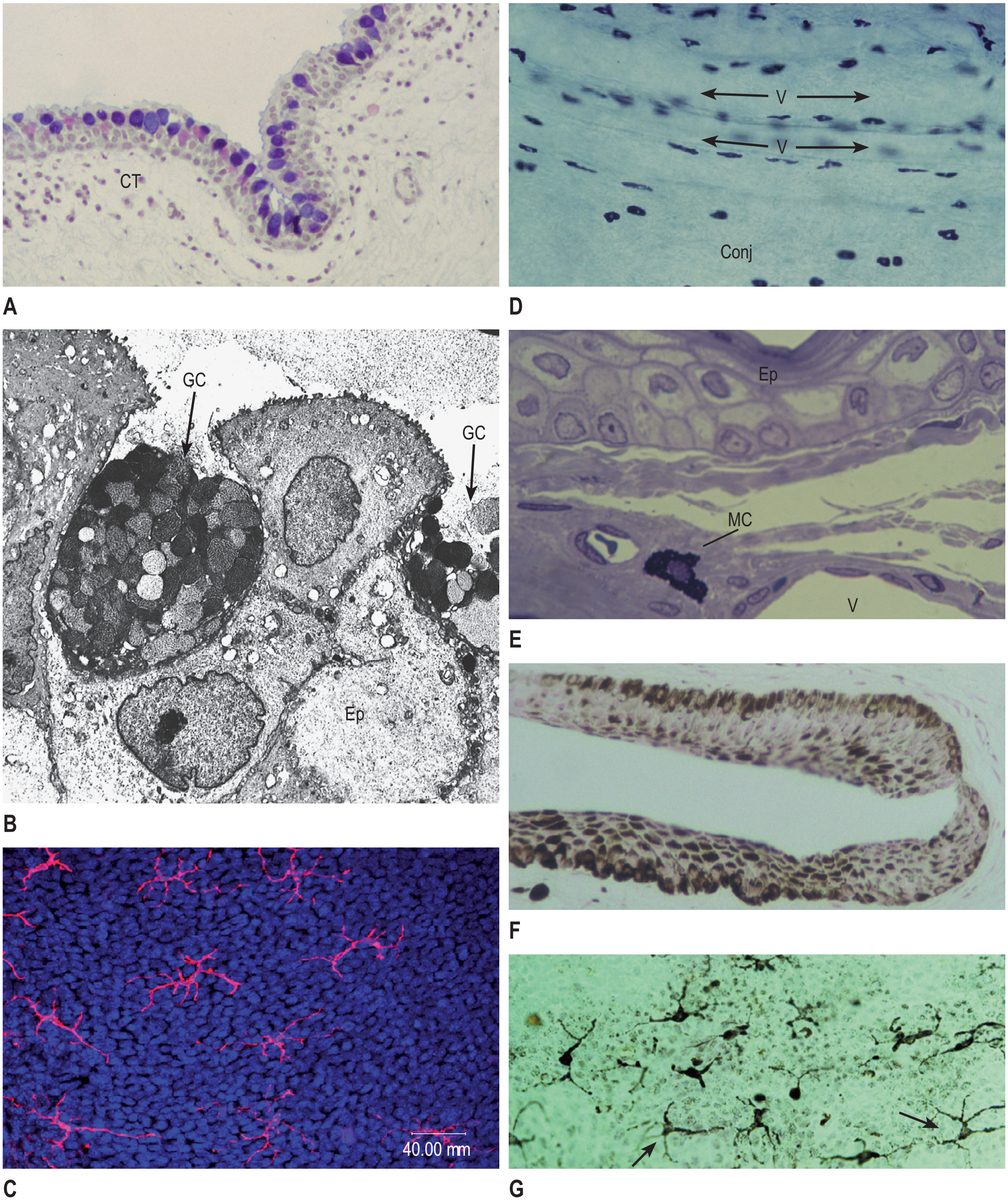
CT
A
V
V
Conj
D
GC
GC
Ep
B
Ep
MC
V
E
F
40.00 mm
C
G

FIGURE 1-62
For legend see opposite page.

'prickle' layer as found in corneal epithelium, indicating that there are fewer desmosomes between the conjunctival epithelial cells. There are numerous other cell types resident within the epithelium, reflecting its protective function, including the following:

- *Goblet cells* (Fig. 1-62A,B) – unicellular mucus-secreting cells that vary in density in different regions of the conjunctiva, being most numerous in the fornices and plica semilunaris. They are responsible for the secretion of the majority of conjunctival mucins.
- *Melanocytes* (Fig. 1-62F,G) – degree of melanization varies dependent on race, although melanocytes are present in all eyes. Melanosomes are synthesized within the melanocyte before exocytosis and subsequent uptake by surrounding epithelial cells as occurs in the epidermis.
- *Intraepithelial MHC class II-positive dendritic cells* (sometimes referred to as Langerhans cells because of their similar morphology to analogous dendritic cell populations in the epidermis of the skin) (Fig. 1-62C) – function as 'sentinels' on the ocular surface and are responsible for trapping and internalizing antigens and transporting these signals to either local lymph nodes (such as the preauricular nodes) or conjunctival associated lymphoid tissue (CALT) or follicles, where they are capable of presenting antigens to naive T cells and inducing primary immune responses or driving antigen-specific B-cell maturation and immunoglobulin production (see Ch. 7).
- *Intraepithelial lymphocytes* – a feature of normal conjunctiva, but increased numbers occur in inflammatory conditions and close to subepithelial lymphoid accumulations. These are predominantly CD3$^+$ T cells, although occasional B cells are present.

The epithelium has an irregular basal aspect adjacent to the underlying connective tissue which is sometimes described as having a looser lymphoid layer and a deeper fibrous layer. Distinct papillae, finger-like protrusions of connective tissue stroma that project into the epithelium, are found only near the limbus. The subepithelial connective tissue contains numerous immunocompetent cells such as mast cells (Fig. 1-62D, E), eosinophils, plasma cells and lymphocytes scattered among the matrix. In some eyes, particularly of older individuals, these may form diffuse or discrete aggregates. Some of these follicles contain pale germinal centres and represent the local mucosal lymphoid tissue (MALT) or CALT. There is no clear evidence of the specialized antigen-transporting intraepithelial M cells that are typically found in other similar MALT such as Peyer's patches or tonsils. The diffuse subepithelial aggregates along with intraepithelial lymphocytes form the efferent arm of the immune system and aid in immunological protection of the ocular surface. The topographical distribution of the small (~0.3 mm) lenticular lymphoid follicles and epithelial crypts in the tarso-orbital conjunctiva suggests that they may approximate to the cornea during eye closure and thus function as an 'immunological cushion'.

The loose connective tissue stroma contains a rich vascular network, similar to the eyelids. In addition, it also receives blood from the anterior ciliary arteries (see Fig. 1-50).

The sensory nerve supply of the palpebral conjunctiva is almost entirely from branches of the ophthalmic division of the trigeminal (supraorbital, supratrochlear and lacrimal nerves). These contain the neurotransmitters substance P, calcitonin gene-related peptide (CGRP) and gallanin. Only the medial portion of the inferior forniceal and palpebral conjunctiva derives its nerve supply from the maxillary division (infraorbital nerve) (see Fig. 1-61). The long ciliary nerves supply the bulbar conjunctiva. Parasympathetic nerves from the pterygopalatine ganglion (containing the neurotransmitters acetylcholine and VIP) and sympathetics (containing norepinephrine and neuropetide Y) travelling with branches of the ophthalmic artery are also present in the conjunctiva. Both parasympathetics and sympathetics have been identified around goblet cells, whereas the sensory nerve endings occur only among the stratified squamous epithelial cells.

Glands in the conjunctiva

Besides the unicellular mucous glands (goblet cells) distributed throughout the conjunctiva, there are several small collections of named glands, some of which are accessory lacrimal glands (glands of Krause in the upper fornices, glands of Wolfring in the upper border of the tarsus); others secrete mucus (glands of Henle). The accessory lacrimal glands are under sympathetic stimulation and are responsible for baseline tear production.

LACRIMAL APPARATUS (Fig. 1-63)

The lacrimal apparatus consists of the *lacrimal gland, lacrimal puncta, lacrimal canaliculi, lacrimal sac* and *nasolacrimal duct.* The lacrimal apparatus functions to produce tears that moisten the ocular surface, thus preventing desiccation of delicate ocular cells and tissues, and facilitating non-friction-bearing movements of the lids on the globe. Tears are thus essential in maintaining the functional integrity of the eye.

Tear film. The tear film (7–9 μm) is composed of three layers: an outer oily or lipid layer (from meibomian and Zeis glands), a middle aqueous layer containing protein, electrolytes and water (mainly from lacrimal glands but also small contributions from conjunctival epithelia and cornea) and a deep hydrophilic mucin layer (from goblet cells and conjunctival epithelial cells and some from the corneal epithelium) associated with the microplicae-rich surface of the conjunctival epithelium (Fig. 1-62B).

Lacrimal gland

The lacrimal gland measures approximately 20 × 12 × 5 mm and weighs approximately 78 mg. It is divided by the lateral horn of the aponeurosis of the levator palpebrae superioris into a large *orbital* and small *palpebral portion,* which are continuous via a small isthmus around the lateral border of the aponeurosis. The orbital portion is shaped like an almond with a convex outer surface that is lodged in the *lacrimal fossa* (see Fig. 1-47). The concave inferior surface is moulded around the tendons of levator palpebrae superioris and lateral rectus (Fig. 1-63). The palpebral portion of the gland is approximately one-quarter of the total gland and its inferior surface lies close to the eye; indeed the gland can usually be seen when the upper lid is everted. In the orbital portion, fine interlobular ducts unite to form three to five main excretory ducts which then traverse the palpebral portion, joining a further five to seven from this part of the gland before entering the superotemporal conjunctival fornix. As a result of this arrangement, removal of the palpebral portion renders the entire gland non-functional.

Histological structure (Fig. 1-64A–C). The lacrimal gland is a branched tubuloacinar gland of the serous type. It is composed of many lobules separated by interstitial fibrovascular septae that are continuous with the poorly developed capsule. On section, each lobule contains numerous acini separated by abundant loose intralobular connective and adipose tissue. Histologically the acini resemble those of the parotid gland and appear as a series of rounded profiles in cross-section (Fig. 1-64A,B). Each acinus or tubuloacinar unit consists of a single layer of cuboidal or columnar cells whose apices are directed towards a central lumen (Fig. 1-64B–D). A layer of stellate-shaped myoepithelial cells surrounds each acinus. The central lumen of several acini unite to form intralobular ducts (Fig. 1-64A), which eventually form

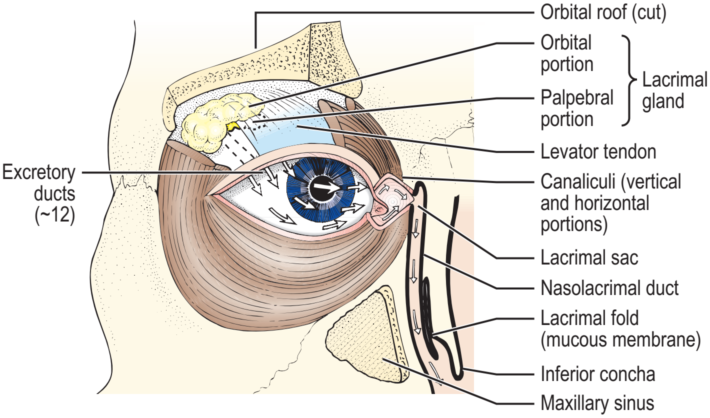

FIGURE 1-63 Schematic diagram summarizing the entire lacrimal apparatus. Arrows indicate the direction of tears from the site of production to the site of drainage.

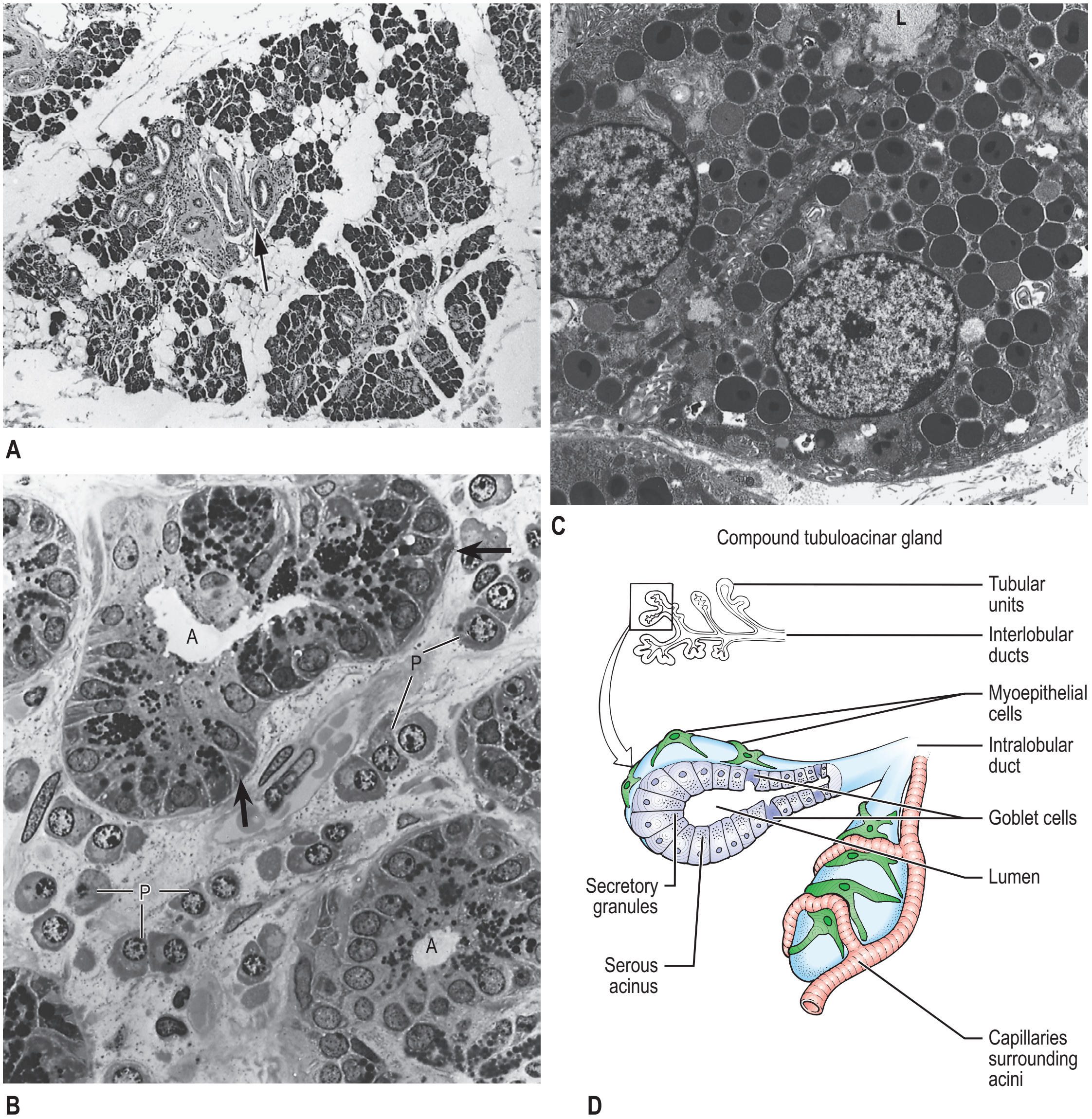
A
L
A
P
P
A
C
Compound tubuloacinar gland
Tubular units
Interlobular ducts
Myoepithelial cells
Intralobular duct
Goblet cells
Lumen
Secretory granules
Serous acinus
Capillaries surrounding acini
B
D

FIGURE 1-64 Histology of the lacrimal gland. (**A**) Low-power micrograph of an entire lobule containing a series of large intralobular ducts (arrow). (**B**) Semi-thin section illustrating the arrangement of the glandular epithelial cells, containing numerous secretory granules, in acinar units (A). The vascular intralobular connective tissue is extremely rich in mature plasma cells (P). Arrows, myoepithelial cells. (**C**) Electron micrograph revealing the ultrastructure of a few pyramidal-shaped acinar cells whose apices are directed toward the central lumen (L). Note the numerous electron-dense zymogenic granules in the apical portion of the cells. An intracellular canaliculus is indicated (arrow): N, nucleus. (**D**) Three-dimensional diagram summarizing the arrangement of the epithelial cells, myoepithelial cells and capillaries in the lacrimal gland. Original magnifications: **A**, × 40; **B**, × 630; **C**, × 4400. (Part B courtesy of W.R. Lee.)

larger interlobular ducts that unite to form the main excretory duct system. The glandular epithelial cells have the characteristic histological and ultrastructural appearance of serous cells, namely basophilic cytoplasm, owing to large numbers of round or oval secretory ('zymogen') granules. The epithelial cells have been subdivided into various subtypes depending on the size and electron density of these granules; however, there is still debate as to whether these are functional subtypes or different stages in the life

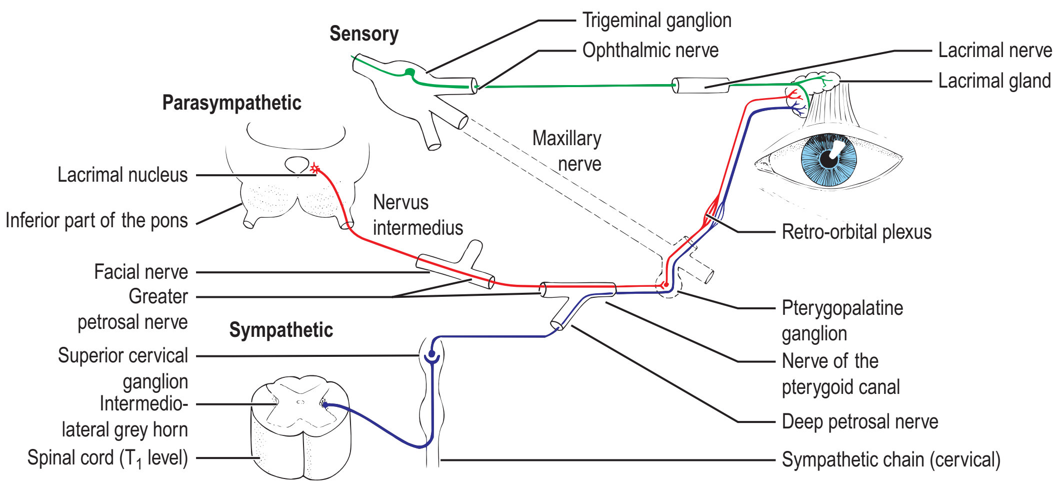
Sensory
Trigeminal ganglion
Ophthalmic nerve
Lacrimal nerve
Lacrimal gland
Parasympathetic
Maxillary nerve
Lacrimal nucleus
Inferior part of the pons
Nervus intermedius
Retro-orbital plexus
Facial nerve
Greater petrosal nerve
Sympathetic
Pterygopalatine ganglion
Nerve of the pterygoid canal
Superior cervical ganglion
Intermedio-lateral grey horn
Deep petrosal nerve
Spinal cord (T1 level)
Sympathetic chain (cervical)

FIGURE 1-65 Diagram summarizing the sensory (trigeminal nerve – green), secretomotor (facial nerve – red) and sympathetic (blue) innervation of the lacrimal gland.

cycle of one cell type. The presence of true intracellular canaliculi (Fig. 1-64C), as observed in salivary glands, is also still controversial. The secretion of the gland is primarily proteinaceous, although some granules contain glycosaminoglycans and the lumen on histological examination contains strongly eosinophilic and mucoid-like secretory material. The secretion also contains lysozymes, lactoferrin, B-lysin and immunoglobulin A (IgA), which are important in defence of the ocular surface against microbial infection. The IgA is derived from numerous plasma cells, present along with other immunocompetent cells such as lymphocytes and mast cells, in the intralobular connective tissue (Fig. 1-64B). The number of these cells increases with age concomitant with increased fibrosis and fatty infiltration and a decrease in the acinar elements especially in the orbital lobe.

Nerve supply. The nerve supply of the lacrimal gland and the complex course of the secretomotor fibres is summarized in Figure 1-65. The lacrimatory nucleus of the facial nerve lies at the rostral end of the general visceral efferent column of cell bodies in the brainstem, which include the superior and inferior salivatory nuclei. The cells are under the influence of the hypothalamus via descending autonomic pathways, thus explaining the neuronal pathways involved in excess lacrimation, which accompanies various emotional states. Reflex excess lacrimation occurs following irritation of the cornea, conjunctiva and nasal epithelia (afferent pathways in ophthalmic and maxillary divisions of the trigeminal). Interneurones connect the trigeminal sensory nuclei with the lacrimatory nucleus.

Blood supply. The lacrimal gland derives its blood supply principally from the lacrimal artery, an early branch of the ophthalmic artery, although a variable branch from the infraorbital artery (originating indirectly from the external carotid) may also aid in its supply. Venous blood drains posteriorly, usually to the superior ophthalmic vein in the orbit, and lymph drains to the preauricular node.

Collecting portion of the lacrimal apparatus (Fig. 1-64)

The collecting system serves to drain normal tears that have not evaporated (normally only a very small quantity) and those produced in times of increased lacrimation. Excess tears are drained from the medial aspect of the conjunctival sac via the canaliculi into the lacrimal sac and nasolacrimal duct, to empty into the inferior meatus of the nasal cavity (Fig. 1-63).

The *puncta* are small openings (visible to the naked eye on the medial margin of each lid) at the summit of small swellings, the *papillae lacrimalis* (see Fig. 1-59). Tears that enter the puncta from the *lacus lacrimalis* during blinking pass into the *lacrimal canaliculi* situated in the upper and lower lid behind the medial palpebral ligament. Each canaliculus is about 10 mm long (0.5 mm in diameter) and has vertical and horizontal components (Fig. 1-63) that may unite to form a common canaliculus before entering the *lacrimal sac*. The canaliculi are lined by stratified squamous non-keratinizing cells. They enter the lacrimal sac by piercing the fascial covering. This sac is 12 mm long and in its upper portion the walls are usually in apposition. It lies in the lacrimal fossa (see p. 2; Fig. 1-5B), protected by the medial palpebral ligament anteriorly and the lacrimal fibres of orbicularis oculi posteriorly. It is related medially to the ethmoidal air cells and the middle meatus of the nasal cavity. The walls of the lacrimal sac consist of fibroelastic tissue and are lined by a mucous membrane consisting of stratified cuboidal/columnar epithelium containing goblet cells. This epithelium is continuous with that lining the canaliculi and the nasolacrimal duct inferiorly.

The nasoclacrimal duct empties into the anterior part of the inferior meatus of the nasal cavity, the opening being protected by a flap of mucous membrane, which prevents air and debris passing up the duct during nose 'blowing'. The duct lies in a bony nasolacrimal canal formed by the maxilla, the lacrimal bone and the inferior nasal concha. The duct is lined by a stratified columnar ciliated epithelium, which rests upon a vascular substantia propria. Knowledge of the position of a variety of constrictions and mucous membrane folds or 'valves' along the course of the canal is important during reconstruction of congenitally malformed lacrimal drainage systems.

Anatomy of the visual pathway

The visual pathway is made up of the *retina, optic nerves, optic chiasma, optic tracts, lateral geniculate bodies, optic radiations* and *visual cortex* (summarized in Fig. 1-66). There are other areas of the cortex also associated with vision such as the frontal eye fields (see Ch. 5 for full description of visual physiology). The visual pathway is effectively a tract within the central nervous system because the retinae develop as evaginations of the diencephalon (see Ch. 5) and, as discussed above, the optic nerves are covered by layers of meninges; even the corneoscleral envelope and uveal tract of the eye itself can be considered as homologous to the dura mater and pia–arachnoid, respectively.

The retina has been described on p. 38. The intraocular, orbital and intracanalicular portions of the optic nerve were described on p. 59. Description of the visual pathway will commence at the intracranial portion of the optic nerve.

INTRACRANIAL PORTION OF THE OPTIC NERVE (Figs 1-51 and 1-66)

The optic nerves leave the cranial end of the optic canal and pass medially, backwards and slightly upwards within the subarachnoid space of the middle cranial fossa. They end by forming the optic chiasma in the floor of the third ventricle. Important relations include the olfactory tracts, frontal lobe (gyrus rectus) and the anterior cerebral arteries above. Each internal carotid artery as it emerges from the roof of the cavernous sinus lies lateral to the junction of the optic nerve and chiasma (Fig. 1-45). Below the optic nerves lies the jugum of the sphenoid and the sulcus chiasmaticus or optic groove.

OPTIC CHIASMA

The optic chiasma (Figs 1-66 and 1-67) is situated at the junction of the anterior wall and the floor of the third ventricle, approximately 5–10 mm above the diaphragma sella and the hypophysis cerebri. It is a flattened quadrangular bundle of nerves measuring 12 × 8 mm whose anterolateral angles are continuous with the optic nerves, and its posterolateral angles form the optic tracts. It usually lies just behind the optic groove or sulcus chiasmaticus, but may rarely lie partly within the sulcus. The tuber cinereum (a sheet of grey matter that forms a median eminence around the base of the pituitary stalk or infundibulum) lies behind and below the chiasma between the mamillary bodies. The anterior perforated substance is an important lateral relation. The anterior communicating artery passes between the two anterior cerebral arteries, and lies above the chiasma. The partial crossing of optic nerve fibres in the optic chiasma is an essential requirement for binocular vision. The fibres from the nasal hemiretina of each eye cross the midline to enter the contralateral optic tract after taking a short loop in the ipsilateral tract or into the contralateral optic nerve. Nerve fibres from the

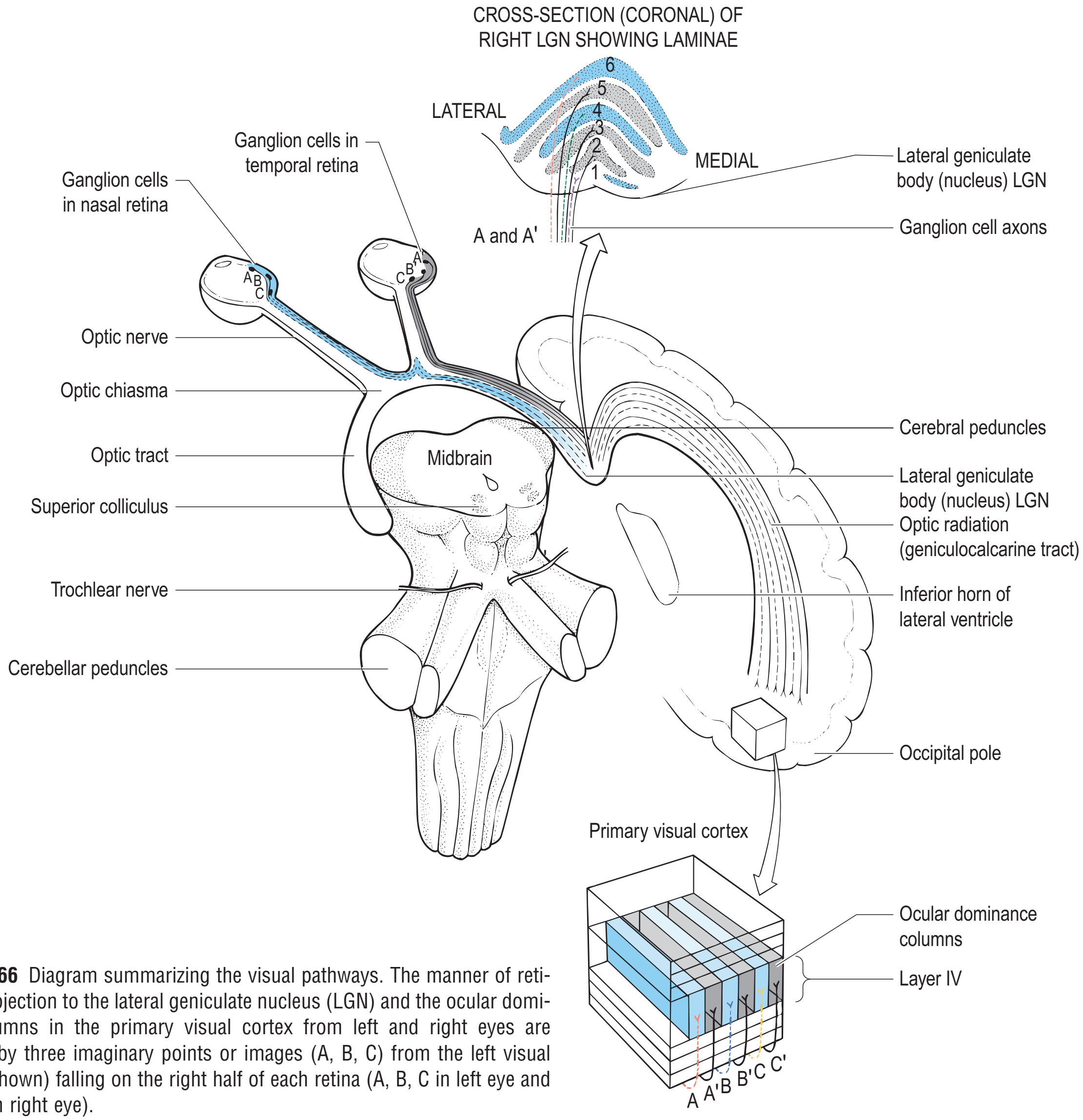
CROSS-SECTION (CORONAL) OF
RIGHT LGN SHOWING LAMINAE
6
5
4
3
2
1
LATERAL
MEDIAL
Lateral geniculate body (nucleus) LGN
A and A'
Ganglion cell axons
Ganglion cells in temporal retina
Ganglion cells in nasal retina
A
B
C
A
B
C
Optic nerve
Optic chiasma
Optic tract
Midbrain
Superior colliculus
Trochlear nerve
Cerebellar peduncles
Cerebral peduncles
Lateral geniculate body (nucleus) LGN
Optic radiation (geniculocalcarine tract)
Inferior horn of lateral ventricle
Occipital pole
Primary visual cortex
Ocular dominance columns
Layer IV
A A'B B'C C'

FIGURE 1-66 Diagram summarizing the visual pathways. The manner of retinotopic projection to the lateral geniculate nucleus (LGN) and the ocular dominance columns in the primary visual cortex from left and right eyes are illustrated by three imaginary points or images (A, B, C) from the left visual field (not shown) falling on the right half of each retina (A, B, C in left eye and A′, B′, C′ in right eye).

temporal hemiretina do not cross at the chiasma (Fig. 1-66).

OPTIC TRACTS

The optic tracts (Figs 1-66 and 1-67) wind round the cerebral peduncles of the rostral midbrain and each divides into a large *lateral root*, which terminates posteriorly in the lateral geniculate body and is concerned with conscious visual sensation, and a smaller *medial root.* The medial root is connected both to the *pretectal area* and *superior colliculus* by the superior brachium and carries around 10% of tract fibres, which functionally are not concerned with conscious vision. They contain six groups of fibres, three of which target the superior colliculus (involved in the visual grasp reflex, automatic scanning of images and visual association pathways); the remaining three enter either the pretectal nucleus (serve the pupillary light reflex), the parvocellular reticular formation (arousal function), or the retinohypothalamic tract, which terminates in the suprachiasmatic nucleus of the hypothalamus (possibly involved in photoperiod regulation and has been invoked to account for the beneficial effect of bright artificial light or sunshine on mood).

The *superior colliculi* are two small rounded elevations located on the dorsal surface of the midbrain above the inferior colliculi, visible on external examination of the brainstem. The pineal body lies between and above the superior colliculi. The two pairs of colliculi are referred to collectively as the *tectum*. The mesencephalic or tectal termination of optic tract fibres is phylogenetically older than the forebrain termination (visual cortex).

The *lateral root of the optic tract* passes backwards, a little upwards, and terminates in the lateral geniculate nucleus (LGN), part of the thalamus (a relay station for ascending sensory information). The lateral root does not lie completely free because its medial aspect is attached to the outer wall of the third ventricle by a narrow band of tissue. It rotates slightly on its own axis (90° inward twist) as it passes round the cerebral peduncles. It runs above the dorsum sella and crosses the third nerve from medial to lateral. Below and parallel to the optic tract runs the posterior cerebral artery (Figs 1-52 and 1-71). The middle portion of the tract is overlapped by the uncus and parahippocampal gyrus.

LATERAL GENICULATE BODIES

Each lateral geniculate body (Figs 1-66 and 1-67) is distinguishable on the surface of the brain as an ovoid projection on the posteroinferior aspect of the thalamus, partly obscured by the overhanging temporal lobe (Fig. 1-67). It consists of a body, head, spur and hilum. The hilum is continuous with the groove between the medial and lateral root of the optic tract, which enters its anterior aspect. It lies at the anterior aspect of the pulvinar, which also partly surrounds it, particularly from above. The LGN in which the great majority of the optic tract fibres terminate consists of six laminae or cell layers (numbered 1 to 6 beginning at the hilum), oriented in a dome-shaped mound similar to a stack of hats (Fig. 1-66). On coronal section, the layers of cell nuclei (approximately 1 million) are separated by white matter (optic tract fibres). Nerve fibres derived from the contralateral eye (crossed fibres from the nasal half of the retina) terminate on cell bodies in layers 1, 4 and 6. Those of the ipsilateral eye (uncrossed) terminate in layers 2, 3 and 5. Thus each LGN receives information from both retinae. Each retinal ganglion cell axon may terminate on up to six geniculate cells; however, these are located

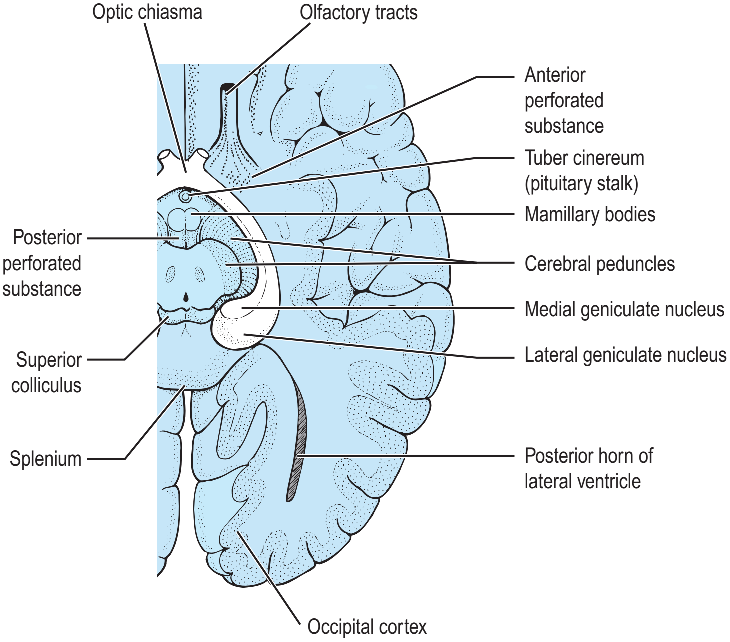
Optic chiasma
Olfactory tracts
Anterior perforated substance
Tuber cinereum (pituitary stalk)
Mamillary bodies
Posterior perforated substance
Cerebral peduncles
Medial geniculate nucleus
Lateral geniculate nucleus
Superior colliculus
Splenium
Posterior horn of lateral ventricle
Occipital cortex

FIGURE 1-67 The base of the brain with the brainstem and cerebellum removed to expose the optic nerves, optic chiasma, optic tracts and lateral geniculate nucleus and their relations.

in one lamina. Fibres from the upper quadrants of peripheral retinae synapse on the medial aspect of the LGN and those of the lower quadrant on the lateral aspect. The *macula* projects to a disproportionately large central wedge of the LGN. The posterior aspect of the LGN is dome-shaped, and it is from here that the geniculate cell axons that form the optic radiation emerge. The bulk of the LGN sends its fibres via the optic radiation to the visual cortex (area 17). The LGN has input from areas 17, 18, 19, oculomotor centres and the reticular formation.

OPTIC RADIATIONS (GENICULOCALCARINE TRACTS)

These tracts (Fig. 1-66) consist of nerve fibre bundles whose cell bodies lie in the LGN. Their axons terminate in the visual (striate) cortex. The fibres form a wide forward and inferiorly directed fan-shaped loop (of Meyer), firstly into the retrolenticular portion of the internal capsule (posterior to sensory fibres and medial to auditory fibres). The fibres then pass into the temporal lobe around the inferior horn of the lateral ventricle (Fig. 1-68). Each tract then passes posteriorly along the lateral aspect of the posterior horn of the lateral ventricle before turning medially to enter the visual cortex. The optic radiations are of major clinical importance as they are frequently involved in cerebrovascular disturbance or tumours (Fig. 1-76). Not all fibres loop to the same degree (Fig. 1-68A,B). Those destined for the lower half of the visual cortex take a wider sweep into the loop around the tip of the inferior horn of the lateral ventricle than those designed for the upper half of the visual cortex (Fig 1-68B). The fibres that swing furthest into the loop are associated with peripheral retina; those that pass more directly posteriorly originate closer to or within the macula region of the retina.

PRIMARY VISUAL CORTEX (AREA 17)

The myelinated fibres of the geniculocalcarine tract (containing fibres from both eyes) enter the primary visual cortex, which lies within the depths of the calcarine sulcus and extends both above and below its margins on the medial surface of the occipital cortex, extending as far posteriorly as the occipital pole (Fig. 1-69B) and as far anteriorly as the parieto-occipital sulcus. The area above the fissure is known as the *cuneus gyrus* and below is the *lingual gyrus*.

Fibres from the superior retinal quadrants (representing the inferior visual field) pass to the upper lip of the calcarine sulcus. The fibres representing the macula account for one-third of the visual cortex (posterior portion of area 17). The myelinated fibres of the geniculocalcarine tract entering this area of cortex create the conspicuous white line or stria (of Gennari). This represents layer IV in the cortex (Fig. 1-70). The six basic layers of the primary visual cortex are shown in Figure 1-70. This region of cortex, although thinner (1.5 mm), is more cellular than other areas of cortex, the predominant cell type not being pyramidal but small stellate cells. Alternating *ocular dominance columns* of these cells receive input from right and left eyes (Fig. 1-66). The geniculocalcarine projection is ordered in a manner whereby matching points from the retinae of both eyes are registered side by side in contiguous columns (see Ch. 5 for a discussion of binocular vision, colour vision, etc.). The cells in laminae II and III project to the secondary visual cortex (areas 18 and 19; Fig. 1-69B,C). Those in lamina V project to the superior colliculus, and those of VI are a major source of 'feed-forward' to the LGN.

SECONDARY VISUAL ASSOCIATION AREAS (AREAS 18 AND 19)

These association areas (Fig. 1-69B,C), which lack the characteristic 'extra' stria found in the primary visual cortex, lie above and below area 17 and extend on to the lateral surfaces of the cerebral hemispheres. They possess the usual six layers, although layer IV is less extensive. Areas 18 and 19 receive afferent input fibres from area 17, the thalamus and pulvinar, together with other regions of the cerebral cortex. The connections of areas 18 and 19 mainly follow dorsal and ventral pathways (Fig. 1-69C). Outputs to area 7 in the parietal cortex are mainly involved in stereopsis and movement. Ventral outputs to the inferotemporal cortex are concerned with analysis of colour and form, and connections to area 37 are associated with recognition of faces. Area 18 is also likely to be involved in sensory–motor eye coordination, as this is known to be linked to the frontal eye fields and oculomotor nuclei via descending pathways. This area also integrates information from two halves of the visual field via commissural fibres crossing the midline in the splenium of the corpus collosum (Fig. 1-67).

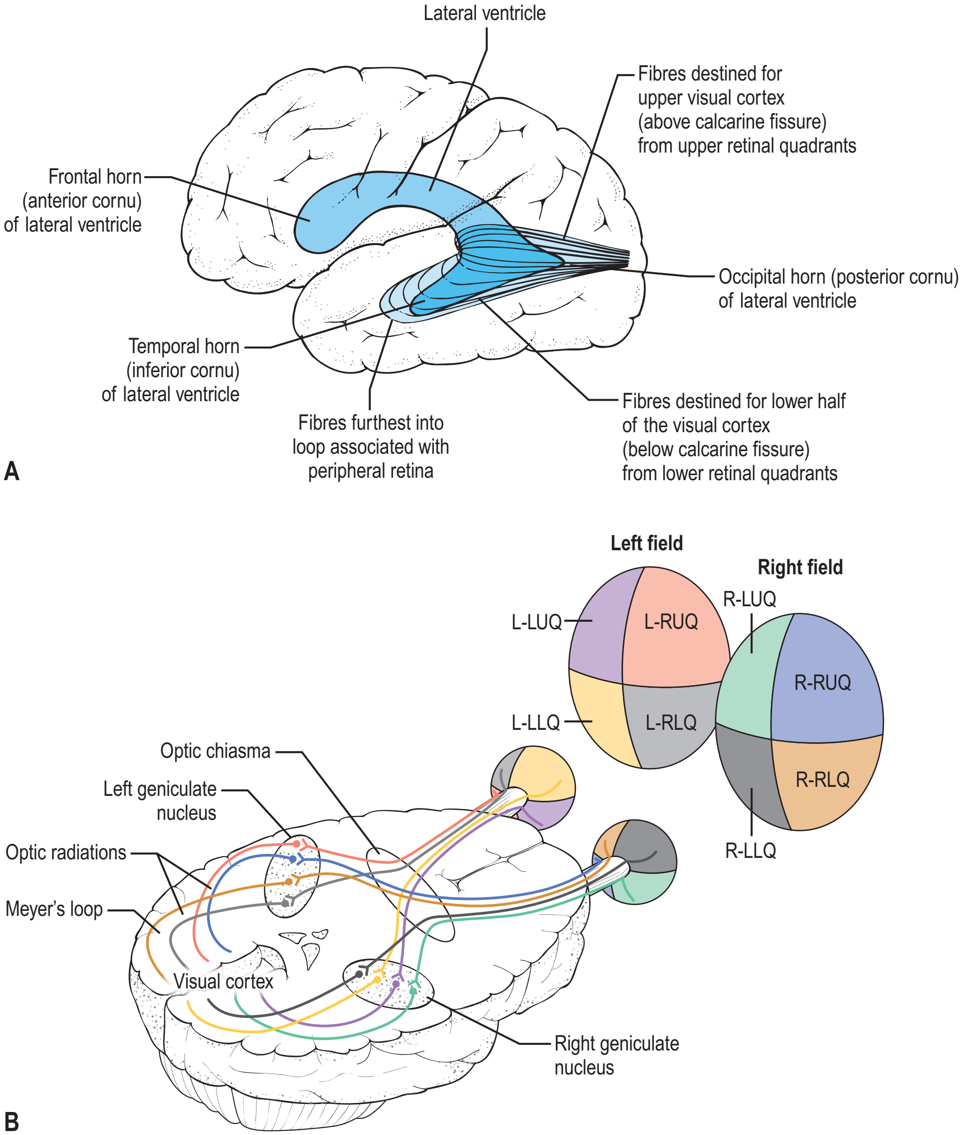
Lateral ventricle
Fibres destined for upper visual cortex (above calcarine fissure) from upper retinal quadrants
Frontal horn (anterior cornu) of lateral ventricle
Occipital horn (posterior cornu) of lateral ventricle
Temporal horn (inferior cornu) of lateral ventricle
Fibres furthest into loop associated with peripheral retina
Fibres destined for lower half of the visual cortex (below calcarine fissure) from lower retinal quadrants
A
Left field
Right field
L-LUQ
L-RUQ
L-LLQ
L-RLQ
R-LUQ
R-RUQ
R-RLQ
R-LLQ
Optic chiasma
Left geniculate nucleus
Optic radiations
Meyer's loop
Visual cortex
Right geniculate nucleus
B

FIGURE 1-68 (**A**) The optic radiation and its relation to the lateral ventricle, viewed from the left side. (**B**) Diagram showing the retinotopic organization of fibres within the optic radiations. (Part B courtesy of Wikimedia Commons.)

FRONTAL EYE FIELD

This frontal area (Fig. 1-69C) corresponds to Brodmann's areas 6, 8 and 9, and is concerned with voluntary control of eye movements (saccades). Fibres pass from here to the superior colliculus, and in turn are connected to the 'extraocular' cranial nerve nuclei (III, IV and VI) and anterior horn cells (motor neurones) in cervical spinal cord segments, thus allowing coordination of head and neck movements with eye movements.

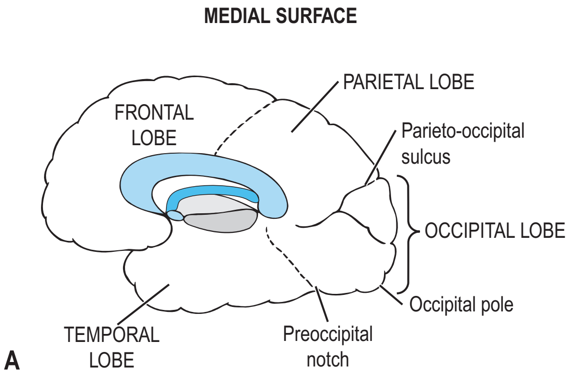
MEDIAL SURFACE
FRONTAL LOBE
PARIETAL LOBE
Parieto-occipital sulcus
OCCIPITAL LOBE
Occipital pole
TEMPORAL LOBE
Preoccipital notch
A
HIGH POWER OF OCCIPITAL LOBE (MEDIAL SURFACE)
Parieto-occipital sulcus
19
18
Secondary visual areas (areas18 & 19)
17
Primary visual area (area 17)
18
19
Secondary visual areas (18 & 19)
Calcarine sulcus
B
Sensory speech area (of Wernicke)
6
8
9
5
7
1
3
2
40
46
44
45
39
22
41
42
17
18
19
37
21
38
20
Motor speech area (of Broca)
C

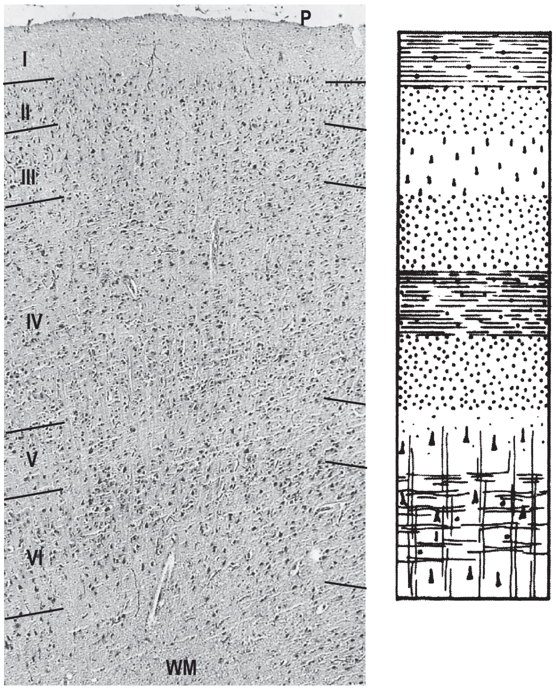
P
I
II
III
IV
V
VI
WM

FIGURE 1-70 Histology and cytoarchitecture of the primary visual cortex. Layers I–VI are indicated on the micrograph and adjacent diagram: P, pia; WM, white matter. Original magnification: × 35.

FIGURE 1-69 (**A**) Simplified diagram showing the boundaries of the lobes of the brain as viewed from the medial aspect. (**B**) The medial surface of the right occipital lobe indicating the sites of the primary and secondary visual areas. (**C**) The cytoarchitectural areas of the cortex as described by Brodmann. Higher visual projections are from area 17 of the left hemisphere. Those to and from area 20/21 are concerned with detail and colour. Projections to and from area 7 are associated with stereopsis and movement, while those to and from area 39 are concerned with recognition of letters and numbers.

RETINOTOPIC ORGANIZATION OF THE VISUAL PATHWAY AND VISUAL PATHWAY DISTURBANCES (Figs 1-71–1-77)

A large amount of neurobiological research in primates and non-primates, together with observations of visual dysfunction or abnormalities in human subjects by neuro-ophthalmologists, has led to a considerable body of knowledge regarding the position along the visual pathway of fibres originating from various points on the retina. This information has been crucial to our understanding of the physiology of vision (see Ch. 5), but in addition helps to explain the specific patterns of visual field disturbances following localized lesions in the pathway. Examples of these lesions and the resultant visual field loss are provided in Figures 1-72–1-77.

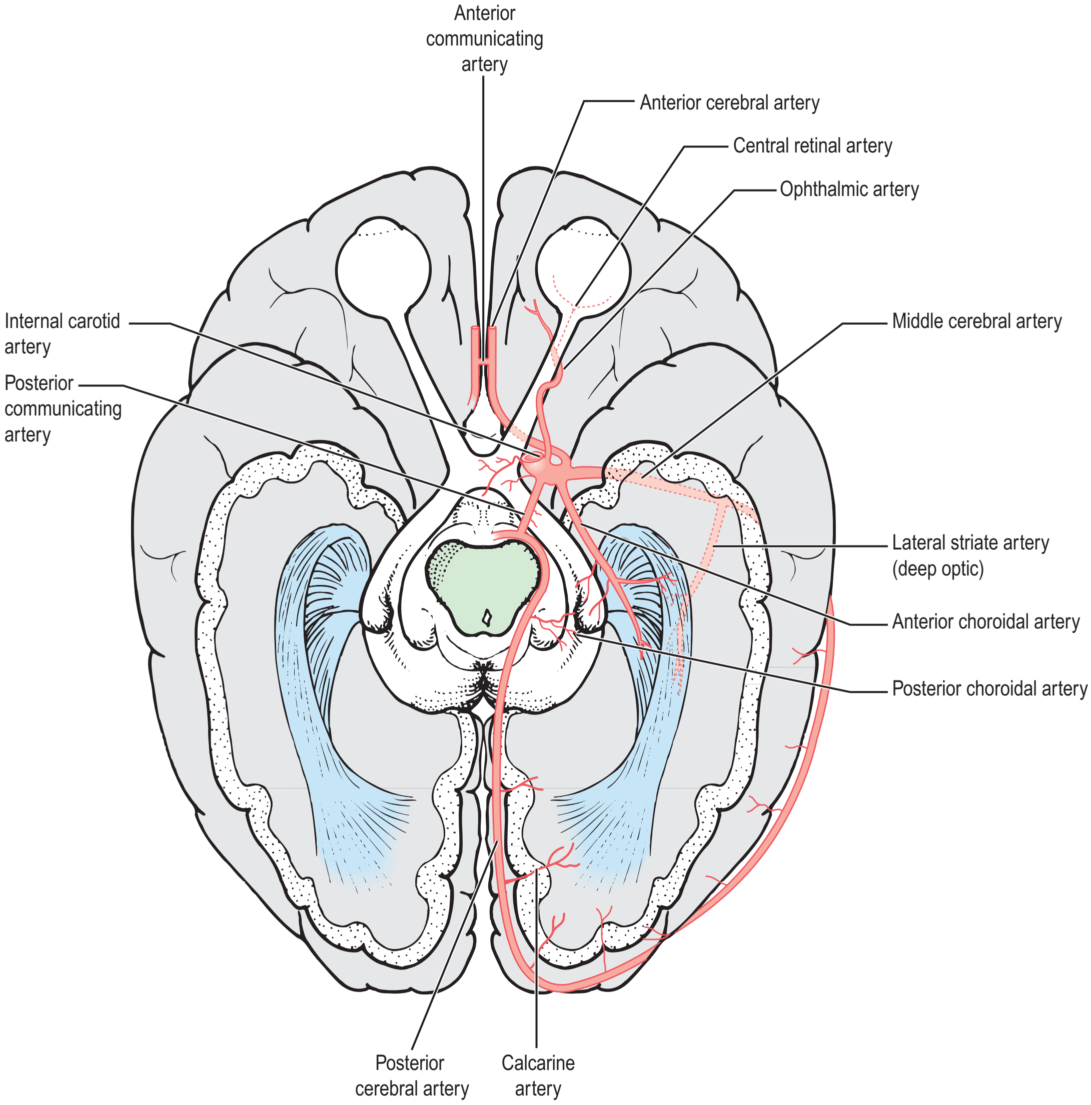
Anterior communicating artery
Anterior cerebral artery
Central retinal artery
Ophthalmic artery
Internal carotid artery
Middle cerebral artery
Posterior communicating artery
Lateral striate artery (deep optic)
Anterior choroidal artery
Posterior choroidal artery
Posterior cerebral artery
Calcarine artery

FIGURE 1-71 Diagram of a brain dissected to display the visual pathways as seen from the ventral aspect. The blood supply to the various parts of the visual pathways is shown in red on the right-hand side of the diagram (corresponding to the *left side of the brain*). Note the blood supply to the following areas:

- Intracranial optic nerve: ophthalmic artery (an important inferior relation) and pial branches of the hypophyseal artery.
- Optic chiasma: adjacent related vessels including the superior hypophysial, internal carotid, posterior communicating, anterior cerebral and anterior communicating artery.
- Lateral root of the optic tract: anterior choroidal artery.
- Lateral geniculate body: anterior choroidal artery and branches of posterior cerebral artery.
- Commencement of the optic radiation (geniculocalcarine tract): anterior choroidal artery.
- Posteriorly directed fibres: lateral striate (deep optic) branch of the middle cerebral artery.
- Termination of geniculocalcarine tract and visual cortex: perforating branches of cortical arteries, principally the calcarine branch of the posterior cerebral although the middle cerebral may anastomose and aid in the supply of the cortex at the anterior end of the calcarine sulcus and at the posterior pole.

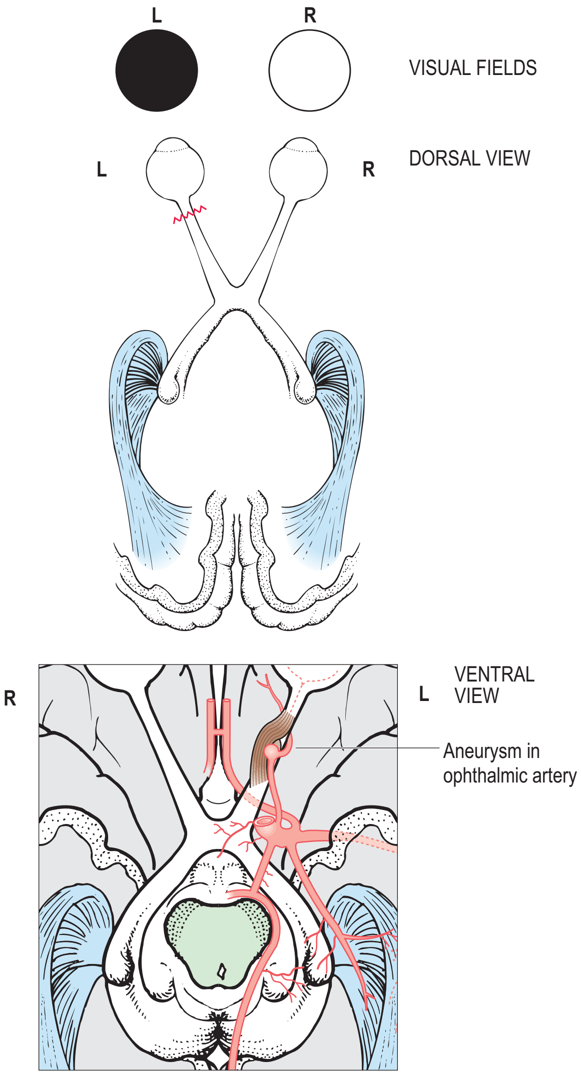
L
R
VISUAL FIELDS
L
R
DORSAL VIEW
R
L
VENTRAL VIEW
Aneurysm in ophthalmic artery

FIGURE 1-72 Blindness in the left eye (top panel shows visual field deficit) caused by a lesion in the left optic nerve (middle panel). An example of such a lesion, an aneurysm in the ophthalmic artery, is illustrated (bottom panel: ventral view of the brain).

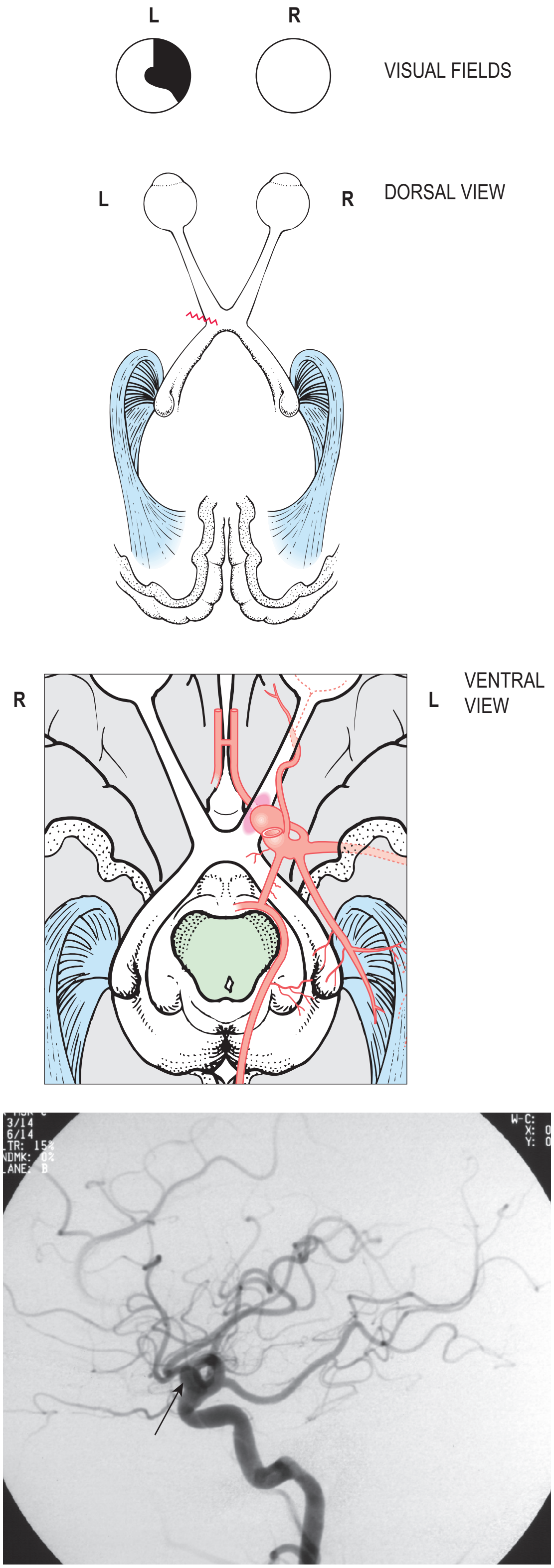
L
R
VISUAL FIELDS
L
R
DORSAL VIEW
R
L
VENTRAL VIEW

FIGURE 1-73 (see right) Incongruous ipsilateral nasal hemianopia (top panel) caused by a lesion on the left side of the optic chiasma (middle panel). An example of such a lesion is an aneurysm of the terminal portion of the internal carotid artery (bottom panel). The radiographic image shows the digitally subtracted arterial phase of a carotid arteriogram of a 48-year-old patient suffering incongruous ipsilateral nasal hemianopia due to such an aneurysm (arrow). (Radiographic image courtesy of Prof. T. Chakera, Royal Perth Hospital.)

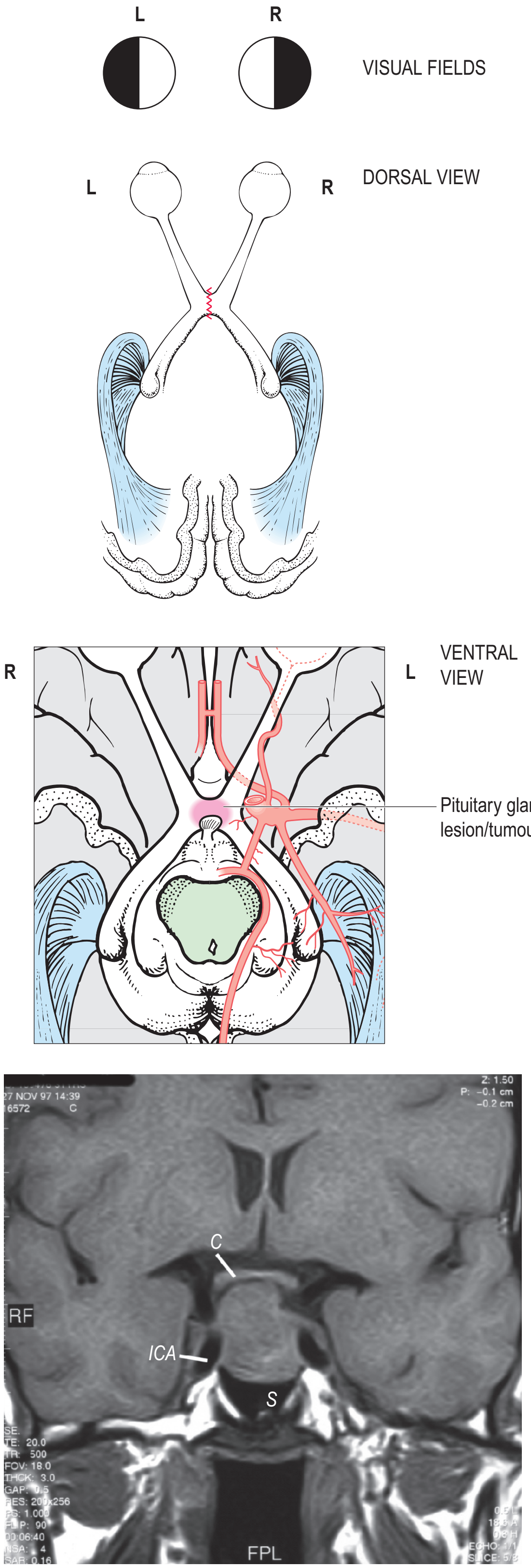
L
R
VISUAL FIELDS
DORSAL VIEW
L
R
R
L
VENTRAL VIEW
Pituitary gland lesion/tumour
27 NOV 97 14:39
Z: 1.50
P: −0.1 cm
−0.2 cm
C
RF
ICA
S
TE: 20.0
TR: 500
FOV: 18.0
THCK: 3.0
GAP: 0.5
FPL

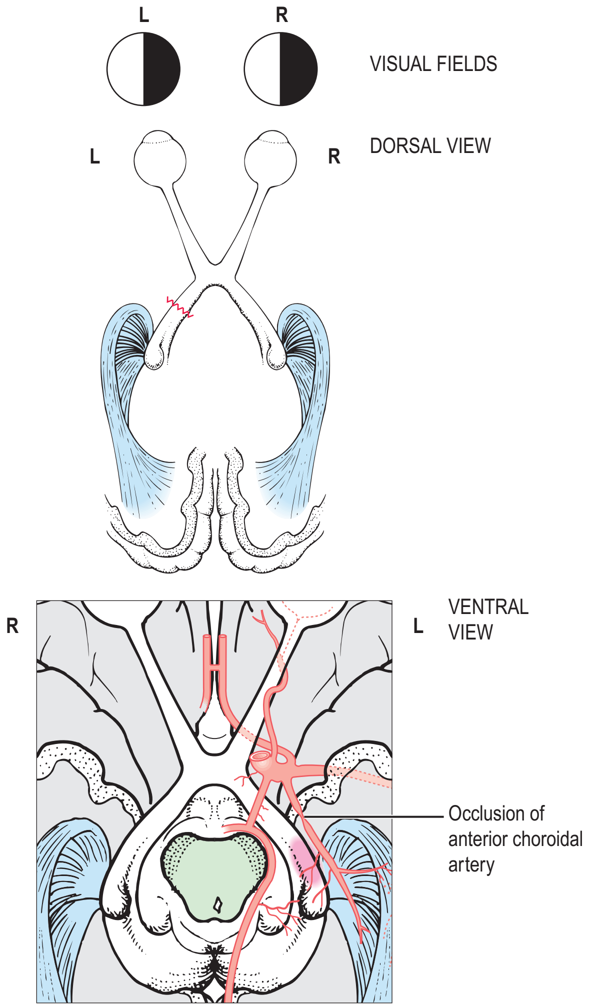
L
R
VISUAL FIELDS
DORSAL VIEW
L
R
R
L
VENTRAL VIEW
Occlusion of anterior choroidal artery

FIGURE 1-75 Contralateral homonymous hemianopia (top panel) caused by a lesion in the left optic tract (middle panel). Such a lesion would damage uncrossed fibres from the temporal retina of the left eye and crossed fibres from the nasal retina of the right eye and therefore cause disturbances in the right visual field. Causes of this type of deficit may include vascular disturbances such as occlusion of the anterior choroidal artery (bottom panel).

FIGURE 1-74 (see left) Contralateral bitemporal homonymous hemianopia (top panel) caused by interruption or damage to the nasal retinal fibres decussating in the optic chiasma (middle panel). A common cause of such deficits is pituitary tumours (bottom panel). Radiograph: coronal MR image of the sellar region in a 31-year-old patient who presented with bitemporal hemianopia. A large pituitary tumour can be seen compressing the optic chiasma (C). S, sphenoid sinus; ICA, internal carotid artery. (Radiographic image courtesy of Prof. T. Chakera, Royal Perth Hospital.)

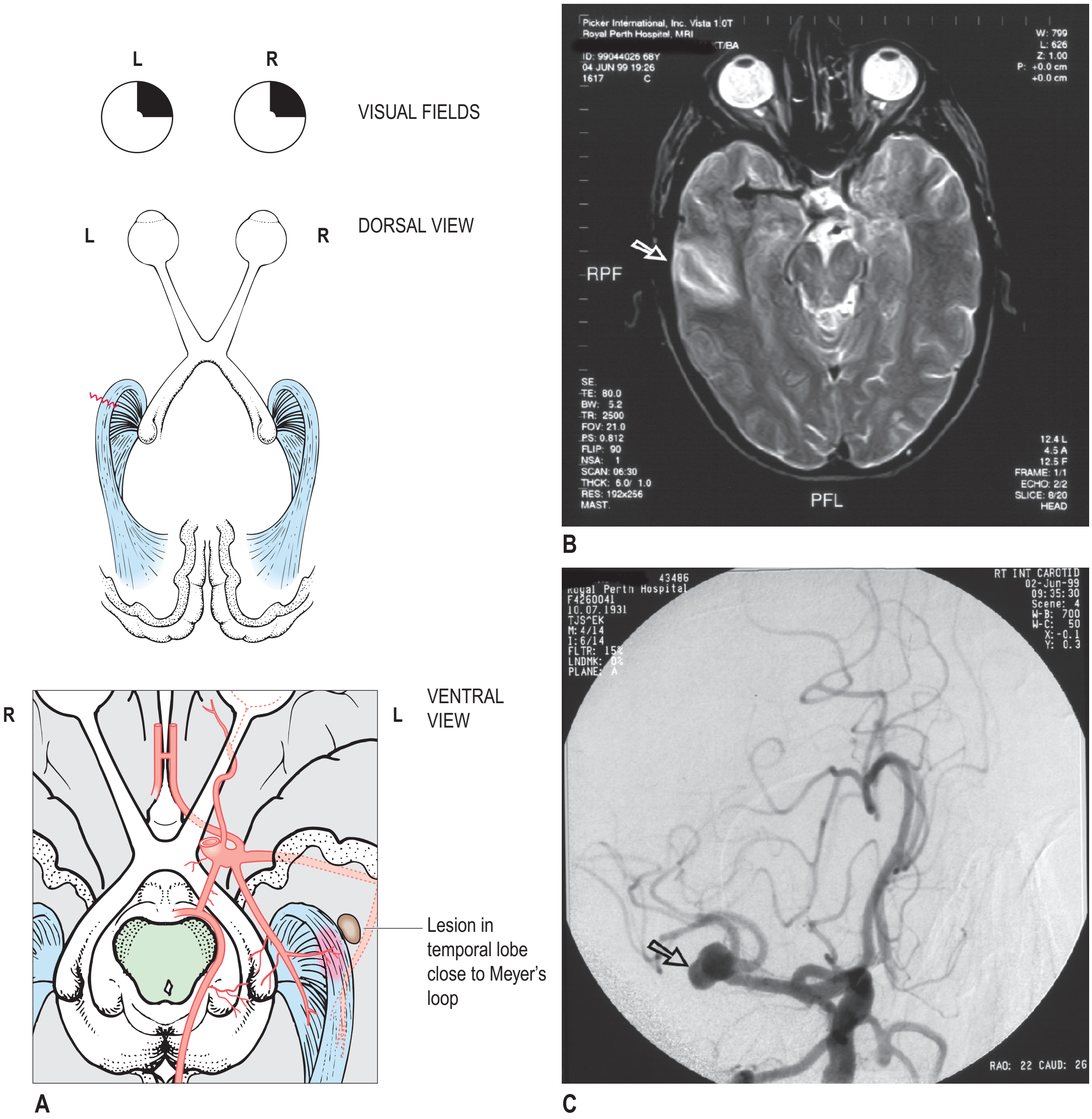
L
R
VISUAL FIELDS
L
R
DORSAL VIEW
R
L
VENTRAL VIEW
Lesion in temporal lobe close to Meyer's loop
A
Picker International, Inc. Vista 1.0T
Royal Perth Hospital, MRI
ID: 99044026 68Y
04 JUN 99 19:26
RPF
PFL
B
RT INT CAROTID
02-Jun-99
Royal Perth Hospital
RAO: 22 CAUD: 26
C

FIGURE 1-76 (**A**) Contralateral homonymous superior quadrantanopia (top panel) – so called 'pie in the sky' defects – may result from lesions in the temporal lobe affecting the fibres furthest into the optic radiation (middle panel) (see Fig. 1-68B), which are derived from the inferior retinal quadrants and therefore cause deficits in the superior visual field. (**B**) Radiographic image of the brain of a 68-year-old patient with a stroke in the right middle cerebral artery (MCA) territory. Axial MR scan (diffusion weighted) shows the size and limits of the right temporoparietal infarct (arrow). (**C**) The digital subtraction carotid arteriogram (anteroposterior projection) shows the aneurysm (arrowhead) at the bifurcation of the right MCA. The extent of the ischaemic changes in the temporal lobe (involving the optic radiation) are consistent with the patient's loss of the left visual field causing him to bump into objects. (Radiographic images courtesy of Prof. T. Chakera, Royal Perth Hospital.)

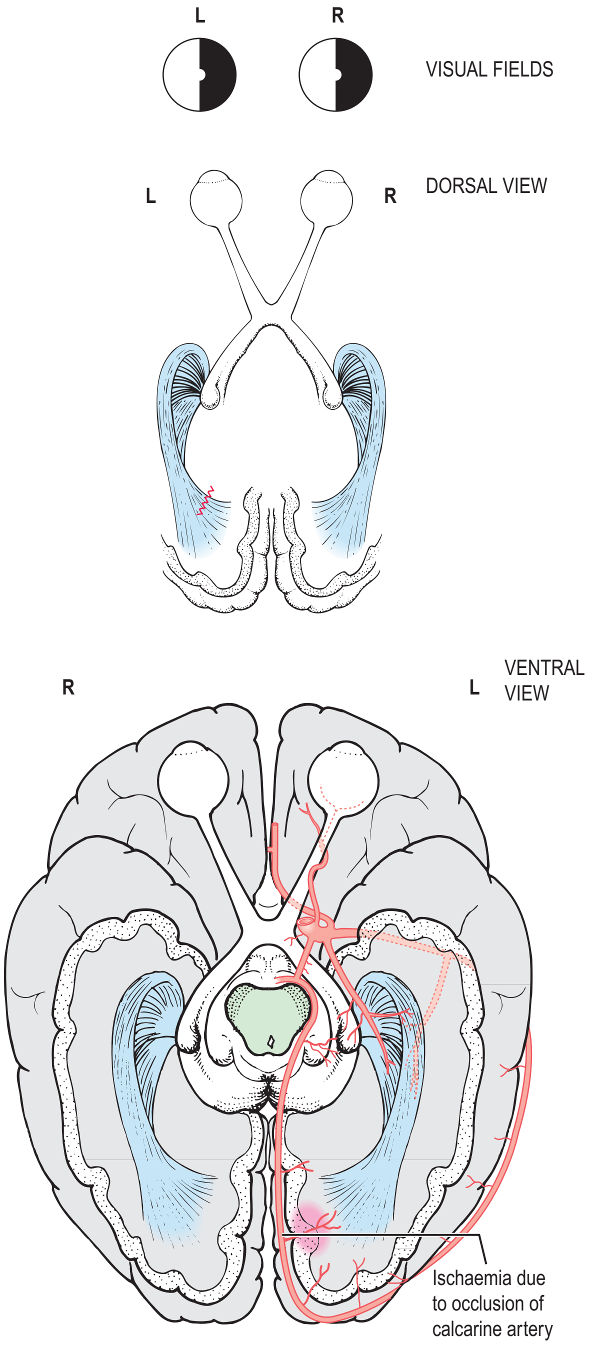
L
R
VISUAL FIELDS
L
R
DORSAL VIEW
R
L
VENTRAL VIEW
Ischaemia due to occlusion of calcarine artery

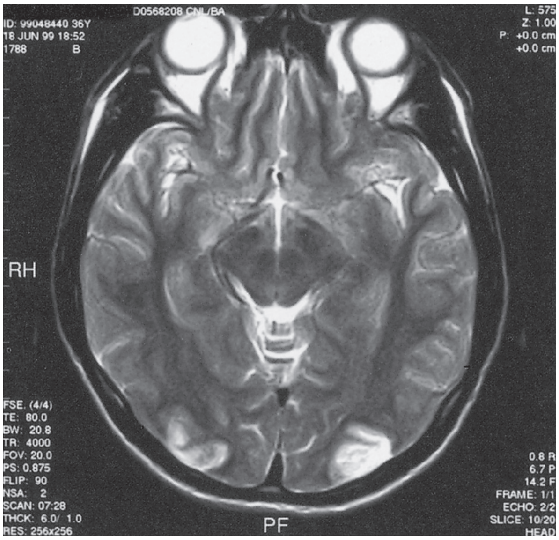
D0568208 CNL/BA
ID: 99048440 36Y
18 JUN 99 18:52
1788 B
L: 575
Z: 1.00
P: +0.0 cm
+0.0 cm
RH
FSE. (4/4)
TE: 80.0
BW: 20.8
TR: 4000
FOV: 20.0
PS: 0.875
FLIP: 90
NSA: 2
SCAN: 07:28
THCK: 6.0/ 1.0
RES: 256x256
PF
0.8 R
6.7 P
14.2 F
FRAME: 1/1
ECHO: 2/2
SLICE: 10/20
HEAD

FIGURE 1-77 Contralateral homonymous hemianopia with macular sparing (top left panel). This may be the result of lesions affecting portions of the occipital cortex, such as tumours or infarcts (bottom panel). Lesions affecting the entire occipital cortex can cause complete blindness. One such case is shown: MR scan of the brain in a 36-year-old intravenous drug user who presented with occipital headaches after injecting speed (amfetamine sulphate) with a dirty needle and waking 2 days later with total blindness and occipital headaches. Early CT scan (not shown) showed no sign of infarction but MR scan and diffusion-weighted imaging (not shown) clearly show bilateral occipital lobe infarcts which would explain the bilateral cortical blindness in this subject. (Radiographic image courtesy of Prof. T. Chakera, Royal Perth Hospital.)

BLOOD SUPPLY OF THE VISUAL PATHWAY

The blood supply of the visual pathways is summarized in Figure 1-71.

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

FURTHER READING

Ahnelt, P., Kolb, H., 1994. Horizontal cells and cone photoreceptors in human retina: a Golgi-electron microscopic study of spectral connectivity. J. Comp. Neurol. 343, 406–427.

Ambati, J., Anand, A., Fernandez, S., Sakurai, E., Lynn, B.C., Kuziel, W.A., et al., 2003. An animal model of age-related macular degeneration in senescent Ccl-2- or Ccr-2-deficient mice. Nat. Med. 9 (11), 1390–1397.

Balazs, E.A., 1982. Functional anatomy of the vitreous. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 425–440.

Bishop, P.N., 2000. Structural macromolecules and supramolecular organisation of the vitreous gel. Prog. Retin. Eye Res. 19, 323–344.

Bishop, P.N., Takanosu, M., Le Goff, M., Mayne, R., 2002. The role of the posterior ciliary body in the biosynthesis of vitreous humour. Eye 16 (4), 454–460.

Butler, T.L., McMenamin, P.G., 1996. Resident and infiltrating immune cells in the uveal tract in the early and late stages of experimental autoimmune uveoretinitis. Invest. Ophthalmol. Vis. Sci. 37, 2195–2210.

Ceelen, P.W., Lockridge, A., Newman, E.A., 2001. Electrical coupling between glial cells in the rat retina. Glia 35, 1–13.

Chan-Ling, T., 1994. Glial, neuronal and vascular interactions in the mammalian retina. Prog. Ret. Eye Res. 13, 357–389.

Chinnery, H.R., Ruitenberg, M.J., Plant, G.W., Pearlman, E., Jung, S., McMenamin, P.G., 2007. The chemokine receptor CX3CR1 mediates homing of MHC class II-positive cells to the normal mouse corneal epithelium. Invest. Ophthalmol. Vis. Sci. 48, 1568–1574.

Christensen, A.M., 2004. Assessing the variation in individual frontal sinus outlines. Am. J. Phys. Anthropol. 127, 291–295.

Cruzat, A., Witkin, D., Baniasadi, N., Zheng, L., Ciolino, J.B., Jurkunas, U.V., et al., 2011. Inflammation and the nervous system: the connection in the cornea in patients with infectious keratitis. Invest. Ophthalmol. Vis. Sci. 52 (8), 5136–5143.

Dacey, D.M., Liao, H.W., Peterson, B.B., Robinson, F.R., Smith, V.C., Pokorny, J., et al., 2005. Melanopsin-expressing ganglion cells in primate retina signal colour and irradiance and project to the LGN. Nature 17 (433), 749–754.

Dartt, D.A., 2002. Regulation of mucin and fluid secretion by conjunctival epithelial cells. Prog. Ret. Eye Res. 21, 555–576.

Demer, J.L., 2004. Pivotal role of orbital connective tissue in binocular alignment and strabismus. Invest. Ophthalmol. Vis. Sci. 45, 729–738.

Demer, J.L., Poukens, V., Miller, J.M., Micevych, P., 1997. Innervation of extraocular pulley smooth muscle in monkeys and humans. Invest. Ophthalmol. Vis. Sci. 38, 1774–1785.

Diaz-Araya, C.M., Provis, J.M., Penfold, P.L., Billson, F.A., 1995. Development of microglial topography in human retina. J. Comp. Neurol. 363, 53–68.

Dick, A.D., Carter, D., Robertson, M., Broderick, C., Hughes, E., Forrester, J.V., et al., 2003. Control of myeloid activity during retinal inflammation. J. Leuk. Biol. 74 (2), 161–166.

Eggers, H.M., 1982. Functional anatomy of the extraocular muscles. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 783–834.

Eisner, G., 1982. Clinical anatomy of the vitreous. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 391–424.

Ertrurk, M., Kayalioglu, G., Govsa, F., Varol, T., Ozgur, T., 2005. The cranio-orbital foramen, the groove on the lateral wall of the human orbit, and the orbital branch of the middle meningeal artery. Clin. Anat. 18, 10–14.

Fitzgerald, M.J.F., 1992. Neuroanatomy. Basic and clinical, second ed. Baillière Tindall, London.

Forrester, J.V., 2003. Macrophages eyed in macular degeneration. Nat. Med. 9, 1350–1351.

Gausas, R.E., 2004. Advances in applied anatomy of the eyelid and orbit. Curr. Opin. Ophthalmol. 15, 422–425.

Gebert, A., Pabst, R., 1999. M cells at locations outside the gut. Semin. Immunol. 11, 165–170.

Guymer, R.H., Bird, A., Hageman, G., 2005. Cytoarchitecture of choroid capillary endothelial cells. Invest. Ophthalmol. Vis. Sci. 45, 1660–1666.

Halfter, W., Dong, S., Schurer, B., Ring, C., Cole, G.J., Eller, A., 2005. Embryonic synthesis of the inner limiting membrane and vitreous body. Invest. Ophthalmol. Vis. Sci. 46 (6), 2202–2209.

Hamrah, P., Zhang, Q., Liu, Y., Dana, M.R., 2002. Novel characterization of MHC class II-negative population of resident corneal Langerhans cell-type dendritic cells. Invest. Ophthalmol. Vis. Sci. 43, 639–646.

Hingorani, M., Metz, D., Lightman, S.L., 1997. Characterisation of the normal conjunctival leukocyte population. Exp. Eye Res. 64, 905–912.

Ihanamaki, T., Pelliniemi, L.J., Vuorio, E., 2004. Collagens and collagen-related matrix components in the human and mouse eye. Prog. Ret. Eye Res. 23, 403–434.

Jakobiec, F.A., 1982. Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia.

Kabil, M.S., Eby, J.B., Shahinian, H.K., 2005. Endoscopic vascular decompression versus microvascular decompression of the trigeminal nerve. Minim. Invasive Neurosurg. 48 (4), 207–212.

Knop, E., Knop, N., 2005. The role of eye-associated lymphoid tissue in corneal immune protection. J. Anat. 206, 271–285.

Kolb, H., Linberg, K.A., Fisher, S.K., 1992. Neurons of the human retina: a Golgi study. J. Comp. Neurol. 318, 147–187.

Kolb, H., Fernandez, E., Schouten, J., Ahnelt, P., Linberg, K.A., Fisher, S.K., 1994. Are there three types of horizontal cell in the human retina? J. Comp. Neurol. 343, 370–386.

Koornneef, L., 1982. Orbital connective tissue. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 835–857.

Last, R.J., 1968. Wolff's anatomy of the eye and orbit. Lewis, London.

Lockwood, C., 1886. The anatomy of the muscles, ligaments, fascia of the orbit etc. J. Anat. Physiol. 20, 1.

Lütjen-Drecoll, E., 1998. Functional morphology of the trabecular meshwork in primate eyes. Prog. Retin. Eye Res. 18, 91–119.

Mariani, A.P., 1990. Amacrine cells of the rhesus monkey retina. J. Comp. Neurol. 301, 382–400.

May, C.A., 2005. Non-vascular smooth muscle cells in the human choroid: distribution, development and further characterization. J. Anat. 207, 381–390.

May, C.A., Neuhuber, W., Lutjen-Drecoll, E., 2004. Immunohistochemical classification and functional morphology of human choroidal ganglion cells. Invest. Ophthalmol. Vis. Sci. 45, 361–367.

McClung, J.R., Allman, B.L., Dimitrova, D.M., Goldberg, S.J., 2006. Extraocular connective tissues: a role in human eye movements? Invest. Ophthalmol. Vis. Sci. 47, 202–205.

McMenamin, P.G., Lee, W.R., Aitken, D.A.N., 1986. Age-related changes in the human outflow apparatus. Ophthalmol. 93, 194–208.

McMenamin, P.G., Forrester, J.V., 1999. Dendritic cells in the central nervous system and eye and their associated supporting tissues. In: Lotze, M.T., Thomson, A.W. (Eds.), Dendritic cells: biology and applications. Academic Press, San Diego, pp. 205–248.

Muller, L.J., Pels, L., Vrensen, F.J.M., 1995. Novel aspects of the organisation of human corneal keratocytes. Invest. Ophthalmol. Vis. Sci. 36, 2557–2567.

Muller, L.J., Marfurt, C.F., Kruse, F., Tervo, T.M.T., 2003. Corneal nerves: structure, contents and function. Exp. Eye Res. 76, 521–542.

Nimmerjahn, A., Kirchhoff, F., Helmchen, F., 2005. Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo. Science 308 (5726), 1314–1318.

Obata, H., Yamamoto, S., Horiuchi, H., Machinami, R., 1995. Histopathologic study of human lacrimal gland. Ophthalmol. 102, 678–686.

Penfold, P.L., Madigan, M.C., Provis, J.M., 1991. Antibodies to human leucocyte antigens indicate subpopulations of microglia in human retina. Vis. Neurosci. 7, 383–388.

Qiao, H., Hisatomi, T., Sonoda, K.H., Kura, S., Sassa, Y., Kinoshita, S., et al., 2005. The characterisation of hyalocytes: the origin, phenotype, and turnover. Br. J. Ophthalmol. 89 (4), 513–517.

Rada, J.A., Shelton, S., Norton, T.T., 2006. The sclera and myopia. Exp. Eye Res. 82, 185–200.

Raviola, G., 1977. The structural basis of the blood–ocular barriers. Exp. Eye Res. (suppl.) 25, 27–63.

Robinson, S., Dreher, Z., 1990. Müller cells in the adult rabbit retinae: morphology, distribution and implications for function and development. J. Comp. Neurol. 292, 178–192.

Rohen, J., 1964. Ciliarkörper. In: von Möllendoff, W. (Ed.), Das Auge und seine Hilfsorgane. Ergänzung zu Band 111/2. Haut und Sinnesorgane. 4 Teil. Handbuch der mikroskopischen Anatomie des Menschen. Springer, Berlin, pp. 189–238.

Rowe, M.H., 1991. Functional organisation of the retina. In: Dreker, B., Robinson, S.R. (Eds.), Neuroanatomy of the visual pathways and their development. Vision and visual dysfunction (Cronly-Dillon J, general ed.), Vol. 3. Macmillan UK, Basingstoke, pp. 1–68.

Rowland, L.P., 1985. Blood–brain barrier, cerebrospinal fluid, brain edema, and hydrocephalus. In: Kandel, E.R., Schwartz, J.H. (Eds.), Principles of neuroscience. Elsevier, New York, pp. 837–844.

Ruskell, G.L., 1971a. Facial parasympathetic innervation of the choroidal blood vessels in the monkey. Exp. Eye Res. 12, 166–172.

Ruskell, G.L., 1971b. The distribution of autonomic postganglionic nerve fibres to the lacrimal gland in monkeys. J. Anat. 109, 229–242.

Ruskell, G.L., 2004. Distribution of pterygopalatine ganglion efferents to the lacrimal gland in man. Exp. Eye Res. 78, 329–335.

Ruskell, G.L., Kjellevold Haugen, I.B., Bruenech, J.R., van der Werf, F., 2005. Double insertions of extraocular rectus muscles in humans and the pulley theory. J. Anat. 206, 295–306.

Sebag, J., 2004. Seeing the invisible: the challenge of imaging vitreous. J. Biomed. Opt. 9 (1), 38–46.

Snell, R.S., Lemp, M.A., 1989. Clinical anatomy of the eye. Blackwell Scientific, Boston.

Stewart, P.A., Tuor, U.I., 1994. Blood–eye barriers in the rat: correlation of ultrastructure with function. J. Comp. Neurol. 340, 566–576.

Stiemke, M.M., Watsky, M.A., Kangas, T.A., Edelhauser, H.F., 1995. The establishment and maintenance of corneal transparency. Prog. Ret. Eye. Res. 14 (1), 109–140.

Thale, A., Tillman, B., 1993. The collagen architecture of the sclera: SEM and immunohistochemical studies. Ann. Anat. 175, 215–220.

Thermos, K., 2003. Functional mapping of somatostatin receptors in the retina: a review. Vision Res. 43 (17), 1805–1815.

Van Buskirk, E.M., 1989. The anatomy of the limbus. Eye 3, 101–108.

Waring, G.O., Bourne, W.M., Edelhaauser, H.F., Kenyon, K.R., 1982. The corneal endothelium. Normal and pathologic structure and function. Ophthalmol. 89, 531–590.

Watson, P.G., Young, R.D., 2004. Scleral structure, organisation and disease. A review. Exp. Eye Res. 78, 609–623.

Williams, P.L., Warwick, R., 1980. Gray's anatomy, thirty-sixth ed. Churchill Livingstone, Edinburgh.

Wotherspoon, A.C., Hardman-Lea, S., Isaccson, P.G., 1994. Mucosa-associated lymphoid tissue (MALT) in the human conjunctiva. J. Pathol. 174, 33–37.

Yemelyanov, A.Y., Katz, N.B., Bernstein, P.S., 2001. Ligand-binding characterization of xanthophyll carotenoids to solubilized membrane proteins derived from human retina. Exp. Eye Res. 72 (4), 381–392.

Young, R.W., 1967. The renewal of photoreceptor cell outer segments. J. Cell Biol. 33, 61–72.

Zhang, H.R., 1994. Scanning electron-microscopic study of corrosion casts on retinal and choroidal angioarchitecture in man and animals. Prog. Ret. Eye Res. 13, 243–270.

Zhivov, A., Stave, J., Vollmar, B., Guthoff, R., 2005. In vivo confocal microscopic evaluation of Langerhans cell density and distribution in the normal human corneal epithelium. Graefes Arch. Clin. Exp. Ophthalmol. 243 (10), 1056–1061.

Chapter **2**

Embryology and early development of the eye and adnexa

- Introduction
- General embryology
- Ocular embryology: general introduction
- Periocular mesenchyme is derived from a mixture of neural crest and mesoderm
- The neural retina and retinal pigment epithelium are derived from neuroepithelium
- Optic nerve and disk development
- Development of the fibrous coat of the eye
- Development of the intraocular contents
- Development of the uveal tract
- Development of the anterior chamber angle and aqueous outflow pathways
- Development of the extraocular muscles
- Development of the eyes and surrounding structures is influenced by the pattern of development of the skull, pharyngeal arches and face

Introduction

This chapter aims to provide an embryological basis for understanding the anatomy of the eye and adnexa. Basic embryological and developmental events from the earliest formation of diverticula in the forebrain at the beginning of the fourth week to the maturation of the various components of the eye in fetal life that influence the final differentiation and functional specialization of the adult eye are described. The contribution of the neural ectoderm, surface ectoderm and periocular mesenchyme to the final configuration of the adult eye and surrounding tissues of the head is emphasized. The cellular interaction and influence of gene expression and transcription factors in determining cell fate will be outlined and examples are given where disruption of gene function underlies disturbances in these embryological events and the interactive processes between the various embryonic tissue and cell types lead to congenital abnormalities in humans.

General embryology

Before considering eye development, which commences in the fourth week, it may be useful for some readers to review the embryological events of the first 3 weeks following fertilization (see eFig. 2-1 and Video 2-1).

Additional content available at https://expertconsult.inkling.com/)

Ocular embryology: general introduction

Although eye development can be considered from an embryological perspective to commence around day 22, when the *optic sulci* (optic primordium) appear as shallow grooves or pits in the inner aspect of the neural plate or neural folds (Fig. 2-1A) and the embryo is around 2 mm in length with eight somites, the group of cells that constitute the eye primordium or eye field have already begun to express a set of 'eye field transcription factors' (EFTFs) that are highly conserved in our evolutionary ancestory (see below). The neural folds have commenced fusion to form the neural tube but the optic sulci form before they have completed their closure rostrally and caudally. When the folds in this area fuse shortly afterwards they give rise to the future *diencephalon* region of the *prosencephalic (forebrain) vesicle*. The optic sulci evaginate to form hollow diverticulae – the *optic vesicles* (Fig. 2-1B).

By about day 25 (20-somite stage) the hollow optic vesicles enlarge and become ensheathed by mesenchymal cells, except at the apex of the vesicle, which is

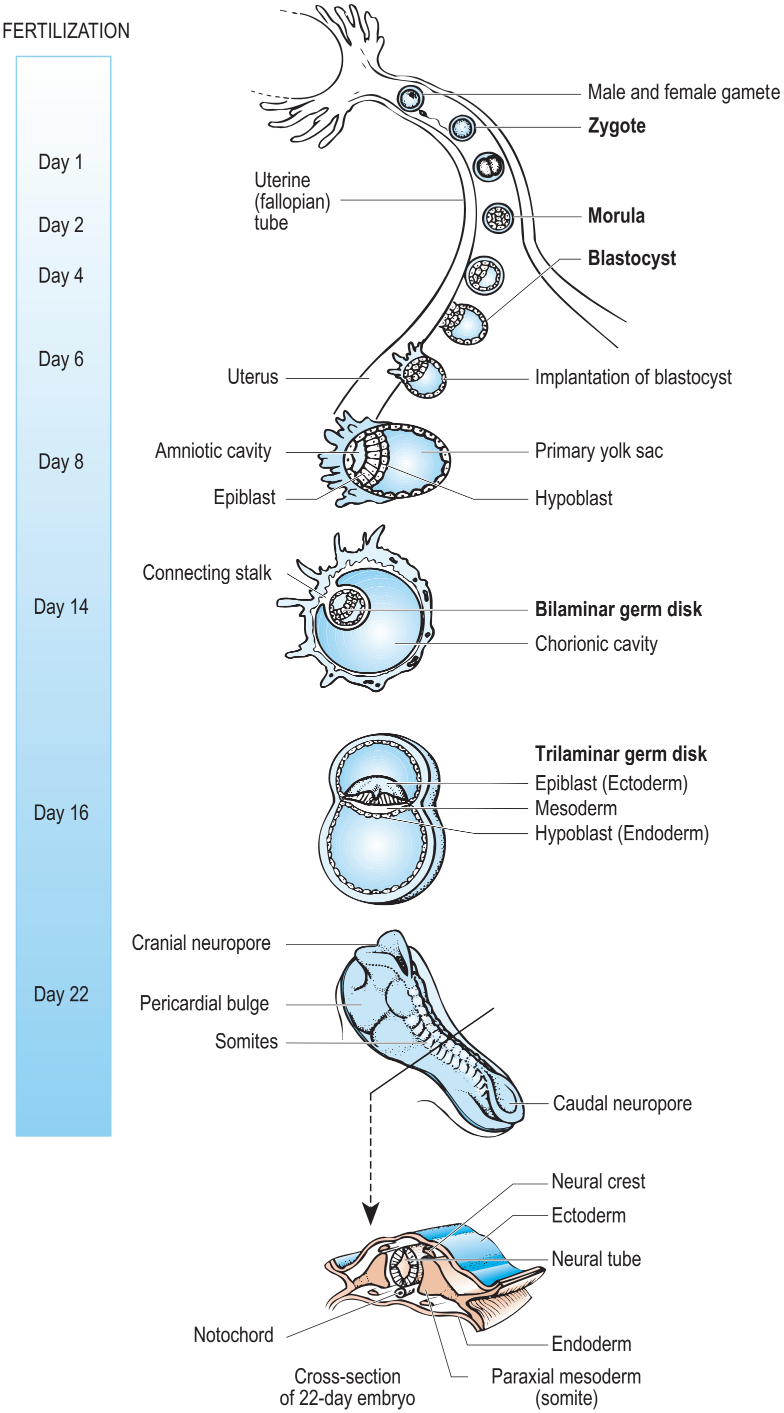
FERTILIZATION
Day 1
Day 2
Day 4
Day 6
Day 8
Day 14
Day 16
Day 22
Male and female gamete
Zygote
Uterine (fallopian) tube
Morula
Blastocyst
Uterus
Implantation of blastocyst
Amniotic cavity
Primary yolk sac
Epiblast
Hypoblast
Connecting stalk
Bilaminar germ disk
Chorionic cavity
Trilaminar germ disk
Epiblast (Ectoderm)
Mesoderm
Hypoblast (Endoderm)
Cranial neuropore
Pericardial bulge
Somites
Caudal neuropore
Neural crest
Ectoderm
Neural tube
Notochord
Endoderm
Cross-section of 22-day embryo
Paraxial mesoderm (somite)

eFIGURE 2-1 Summary of embryological events occurring in the first 3 weeks following fertilization.

The male and female *gametes* unite at fertilization, which generally occurs within the oviduct or uterine tube, to form the *zygote*. The newly formed diploid cell undergoes *cleavage* as it travels towards the uterus. Within the solid mass of cells (30-cell stage) or *morula*, the *blastocyst cavity* appears. The earliest *differentiation* of the embryo occurs around this stage when two groups of cells are formed: a peripheral *outer cell mass* or trophoblast and a central *inner cell mass* or embryoblast. The *embryoblast* will give rise to the embryo proper and some of its attached membranes. By day 5 or 6 the embryo is a hollow spheroid of around 100 cells, known as the *blastocyst*. The *embryoblast* is evident as a heterogeneous collection of cells at one pole of the blastocyst; the remainder forms the *trophoblast*. This differentiates into the syncytiotrophoblast and the cytotrophoblast, which contribute to the fetal component of the placenta. Around this time the blastocyst enters the uterus and commences implantation into its rich endometrial lining.

At the beginning of the second week two cavities appear within the embryoblast. One forms the *amniotic cavity*, lined by epiblast cells, while the other forms the *yolk sac*, lined by cells derived from the hypoblast. Where the two cavities impinge there is a double discoid layer of cells, the *epiblast* or *primary ectoderm* and the *hypoblast* or *primary endoderm*. These two flat disks of cells constitute the *bilaminar germ disk*, which will develop into the embryo proper.

The third week of development commences with formation of the *primitive streak* and *primitive knot* or node at the caudal (tail) end of the epiblast. It is here that epiblast cells detach and migrate laterally and cranially into the potential space between the epiblast and the hypoblast to form the *intraembryonic mesoderm* or third germ layer. Some of the epiblast cells also replace the original hypoblast to form the *definitive* or *secondary hypoblast*. The epiblast is now known as the *ectoderm*. Formation of the primitive streak establishes the craniocaudal axis and bilateral symmetry of the future embryo. The formation of the *notochord* by budding from the primitive knot induces the formation of the *neural plate*. A series of *paraxial mesodermal* condensations or *somites* form along each side of the neural plate as it folds to form the *neural tube*, the precursor of the central nervous system (eFig. 2-1). A special group of cells, the *neural crest* cells, detach or delaminate from the margins of the neural folds and undergo extensive migration throughout the embryo where they differentiate into a remarkable variety of cells and tissues (eFig. 2-2). Derivatives of trunk neural crest cells are shown in eFigure 2-2. Derivatives in the head are more extensive.

Anatomically, the human embryo is described as being in the prone (face-down) position. The terms dorsal and ventral correspond to posterior and anterior in the adult. The terms rostral and caudal correspond to superior (head end) and inferior (tail end).

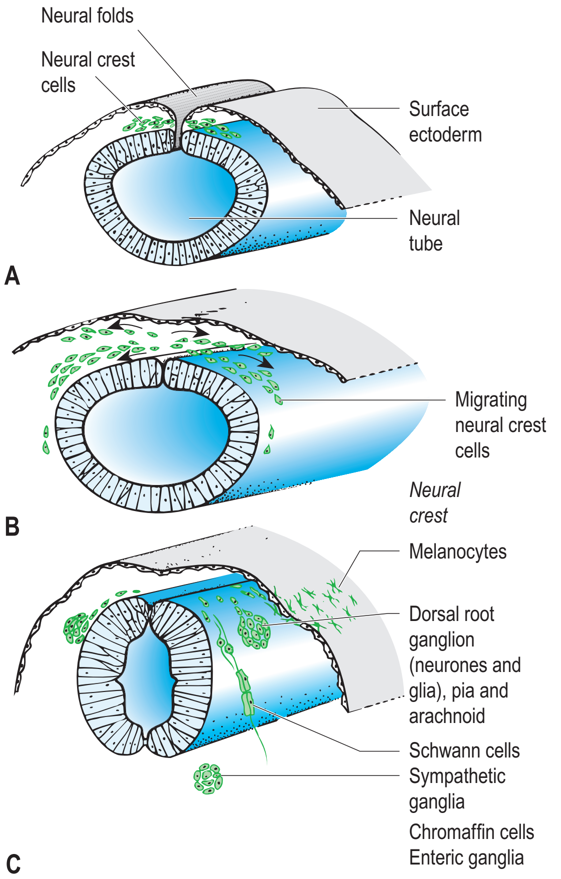
Neural folds
Neural crest cells
Surface ectoderm
Neural tube
A
Migrating neural crest cells
B
Neural crest
Melanocytes
Dorsal root ganglion (neurones and glia), pia and arachnoid
Schwann cells
Sympathetic ganglia
Chromaffin cells
Enteric ganglia
C

eFIGURE 2-2 Diagrammatic summary of the origin (**A**), migratory pathway (**B**), and derivatives (**C**) of neural crest cells in the trunk.

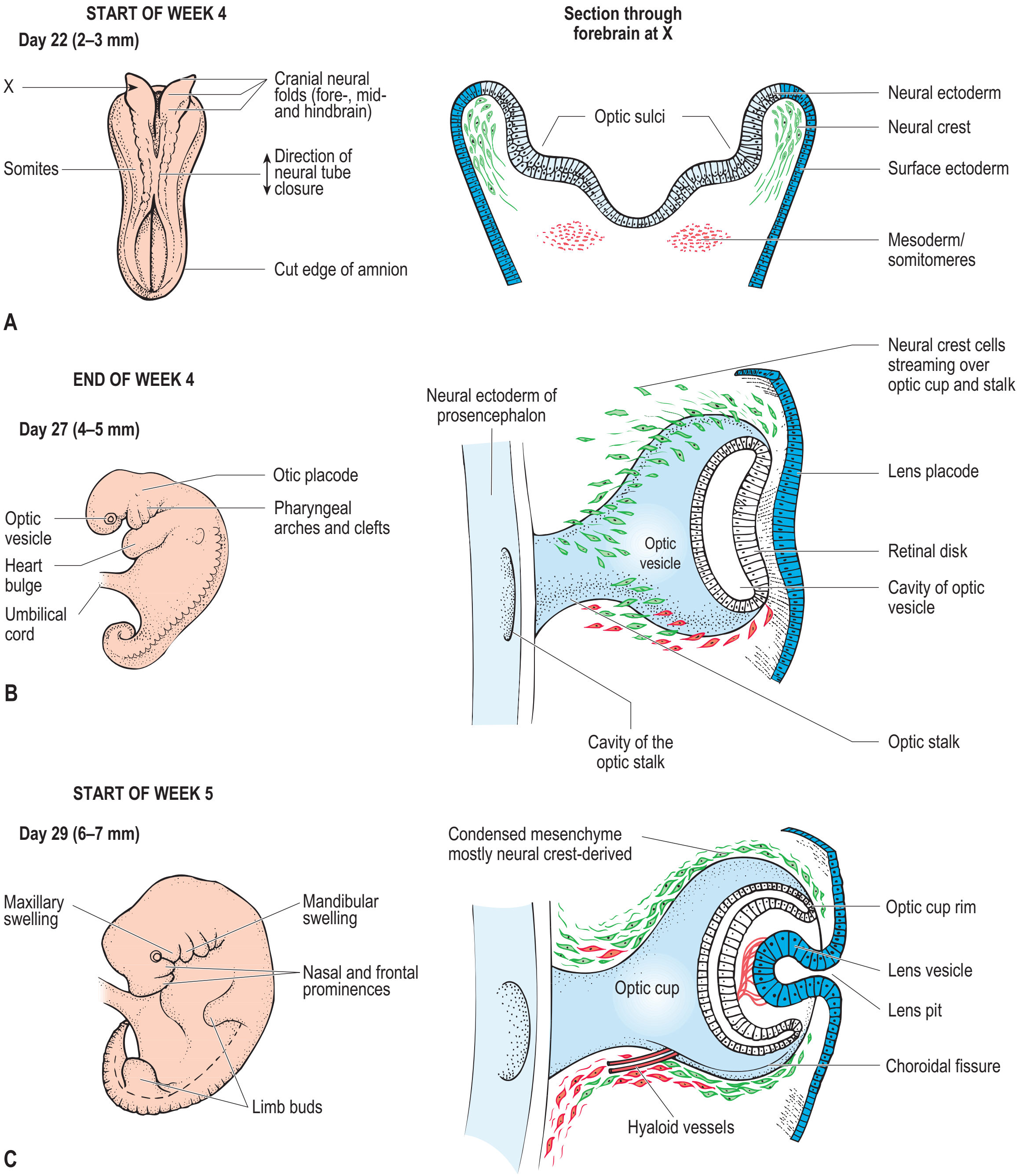
START OF WEEK 4
Day 22 (2–3 mm)
X
Cranial neural folds (fore-, mid- and hindbrain)
Somites
Direction of neural tube closure
Cut edge of amnion
Section through forebrain at X
Optic sulci
Neural ectoderm
Neural crest
Surface ectoderm
Mesoderm/ somitomeres
A
END OF WEEK 4
Day 27 (4–5 mm)
Otic placode
Pharyngeal arches and clefts
Optic vesicle
Heart bulge
Umbilical cord
B
Neural ectoderm of prosencephalon
Neural crest cells streaming over optic cup and stalk
Optic vesicle
Lens placode
Retinal disk
Cavity of optic vesicle
Cavity of the optic stalk
Optic stalk
START OF WEEK 5
Day 29 (6–7 mm)
Maxillary swelling
Mandibular swelling
Nasal and frontal prominences
Limb buds
Condensed mesenchyme mostly neural crest-derived
Optic cup rim
Optic cup
Lens vesicle
Lens pit
Choroidal fissure
Hyaloid vessels
C

FIGURE 2-1 Diagrammatic summary of ocular embryonic development from day 22 to week 8. The external appearance of the whole embryo at the equivalent period is shown on the left. The various ‘germ layers’ are colour-coded to illustrate their origin and final contribution to the eye and periocular tissues.

START OF WEEK 6

Day 37 (8–11 mm)

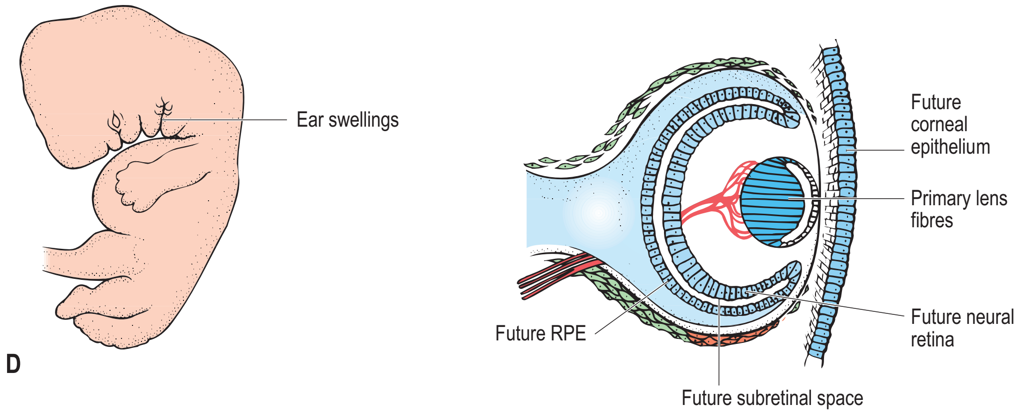
Ear swellings
Future corneal epithelium
Primary lens fibres
Future neural retina
Future RPE
Future subretinal space
D

START OF WEEK 7

Day 44 (13–17 mm)

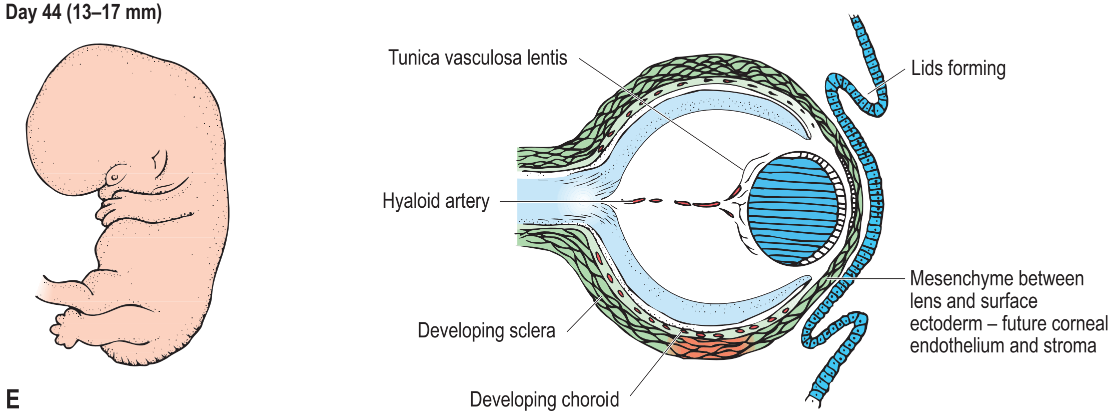
Tunica vasculosa lentis
Lids forming
Hyaloid artery
Mesenchyme between lens and surface ectoderm – future corneal endothelium and stroma
Developing sclera
Developing choroid
E

WEEK 8 (20–30 mm)

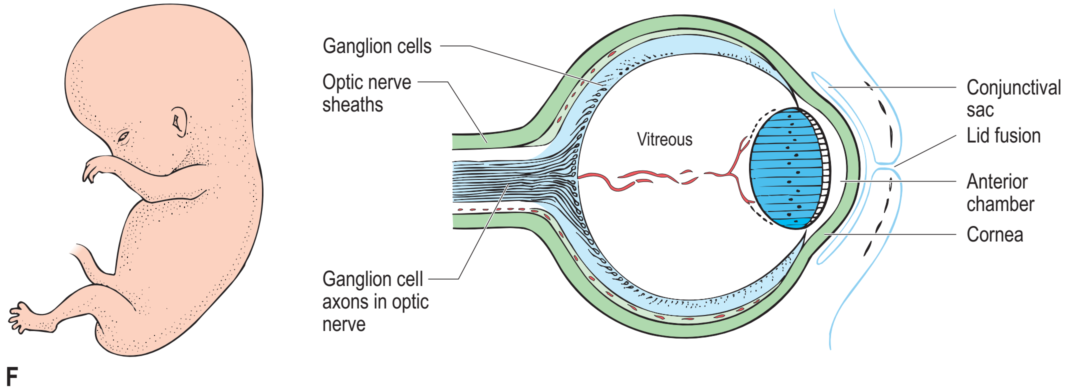
Ganglion cells
Optic nerve sheaths
Vitreous
Conjunctival sac
Lid fusion
Anterior chamber
Cornea
Ganglion cell axons in optic nerve
F

FIGURE 2-1, cont'd

closely apposed to the *surface ectoderm* on the lateral aspect of the developing head. This mesenchyme is derived from a mixture of cephalic neural crest and both para-axial and prechordal mesoderm (see eBox 2-1).

A disk-shaped thickening of the neural ectoderm, the *retinal disk* (future neural retina), lies beneath a localized thickening of the surface ectoderm, which on day 27 is recognizable as the *lens placode* (Fig. 2-1B). Formation of the lens placode directly adjacent to the underlying neural ectoderm is regarded as one of the best examples of *induction* in developmental biology. Lens placode formation coincides with the formation of a constriction in the optic vesicle at its attachment to the wall of the forebrain to form the *optic stalk*. The *cavity of the optic vesicle* or *optic ventricle* (future subretinal space) is continuous, via the lumen in the optic stalk, with the future third ventricle (Figs 2-1B and 2-2A).

The single-layered spheroidal *optic vesicle* undergoes active invagination to fold in on itself creating two 'nested' layers of the goblet-shaped *optic cup*. The distal part of the optic vesicle becomes the inner layer or presumptive neural retina, whereas the proximal part of the optic vesicle becomes the outer layer of the

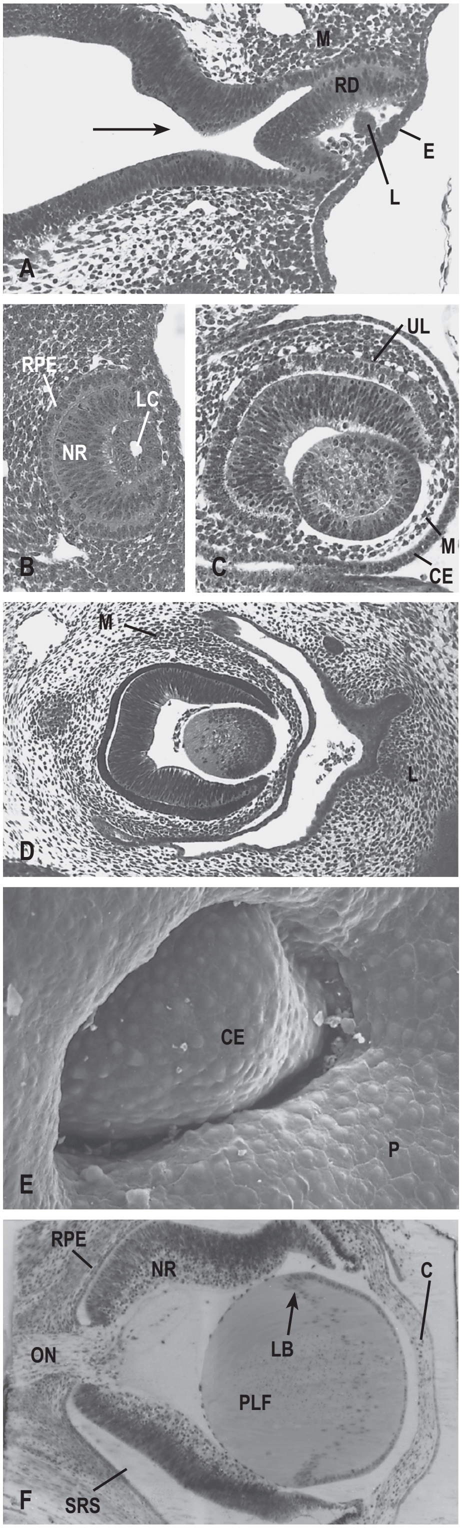

FIGURE 2-2 Histological and scanning electron micrographs of early mammalian eye development (chronological sequence **A–F**). (**A**) The cavity of the optic vesicle is clearly in continuity via the cavity of the optic stalk with the forebrain ventricle (arrow). M, mesenchyme condensation; E, surface ectoderm/periderm; L, lens vesicle; RD retinal disk. (**B**) The lens vesicle, containing a distinct lens cavity (LC), fills the optic cup, which consists of two layers, an outer retinal pigment epithelium (RPE) and neural retina (NR). (**C**) Mesenchyme has condensed around the optic cup and migrated over the cup margin to form the future corneal endothelium/stroma (M) beneath the surface ectoderm-derived corneal epithelium (CE). Beneath the RPE the vascular mesenchyme has already formed a distinct row of vessels, the uveocapillary lamina (UL) or presumptive choroid. (**D**) Pigment is identifiable in the RPE layer. The lentoretinal space contains vascular mesenchyme. The mesenchyme (M) around the developing eye has now formed two layers, an outer dense avascular layer (future sclera) and an inner vascular layer (future choroid). Beneath the newly formed lids (L) lies the conjunctival sac. (**E**) Scanning electron micrograph of the embryonic corneal epithelial surface (CE) and periderm (P) just before lid closure. (**F**) Late embryonic eye with well-developed cornea (C), large lens consisting of primary lens fibres (PLF) whose nuclei form the lens bow (LB). The neural retina (NR) is artefactually detached from the RPE, producing a large subretinal space (SRS). Axons have commenced migration along the optic nerve (ON). Original magnifications: **A**, × 120; **B**, × 180; **C**, × 160; **D**, ×55; **E**, × 100; **F**, ×65. (From McMenamin and Krause, 1993, with permission.)

eBox 2-1

Mesenchyme

This is a term used to describe the tissue occupying the embryo between the surface ectoderm (and derivatives such as neuroectoderm) and the endoderm-derived epithelial layers. It is a loose tissue consisting of stellate amoeboid mesenchymal cells embedded in a matrix rich in glycosaminoglycans. Mesenchymal cells may be derived from several sources, namely mesoderm (dermatome or sclerotome component of the somites or lateral plate mesoderm) or neural crest. Thus, the descriptive term 'mesenchyme' does not imply an origin from any particular embryonic germ layer.

optic cup, namely the presumptive retinal pigment epithelium. The thickened retinal disk at the tip or distal part of the vesicle (Fig. 2-1B,C) and the ectoderm-derived lens placode invaginate via a combination of differential growth (cell elongation and mitosis) and buckling to form the dorsal hemisphere of the optic cup and the lens vesicle. This combined invagination may also be aided by temporary fine cellular bridges between the lens placode and the retinal disk and is mediated by by the extracellular matrix protein fibronectin-1 (*Fn1*), whose extent and circumferential limit beneath the lens placode is regulated by *Pax6* expression (Fig. 2-3). The active growth of the optic cup is not uniform around the circumference, which leads to the development of a groove at the distal and ventral aspect where the margins form the *choroidal* or *optic fissure*. By day 29 invagination of the retinal disk and lens placode is almost complete (Figs 2-1C and 2-2A). A small *lens pit* can be identified just before the surface ectoderm seals over the site of lens placode invagination. By the start of day 36 the *lens vesicle* separates from the surface ectoderm. The lens epithelial cells enclose the *lens cavity* and are surrounded externally by a basal lamina, which will form the future *lens capsule*. The longitudinal groove of the optic fissure, which extends into the optic stalk, acts as a temporary deficiency in the expanding and invaginating cup through which vascular mesenchyme and a branch of the ophthalmic artery, the *hyaloid artery*, become incorporated into the fissure and thus gain access to the lentoretinal space. By the end of the sixth week the growing edges of the choroidal or optic fissure meet and fuse; thus the hyaloid vessels and associated mesenchyme become situated in the centre of the optic stalk and form the future central retinal artery and vein (Fig. 2-1D,E). The fusion or closure of the optic fissure commences at the mid-portion of the optic stalk and continues both proximally and distally. It is completed distally at the margins of the optic cup that will eventually form the pupil.

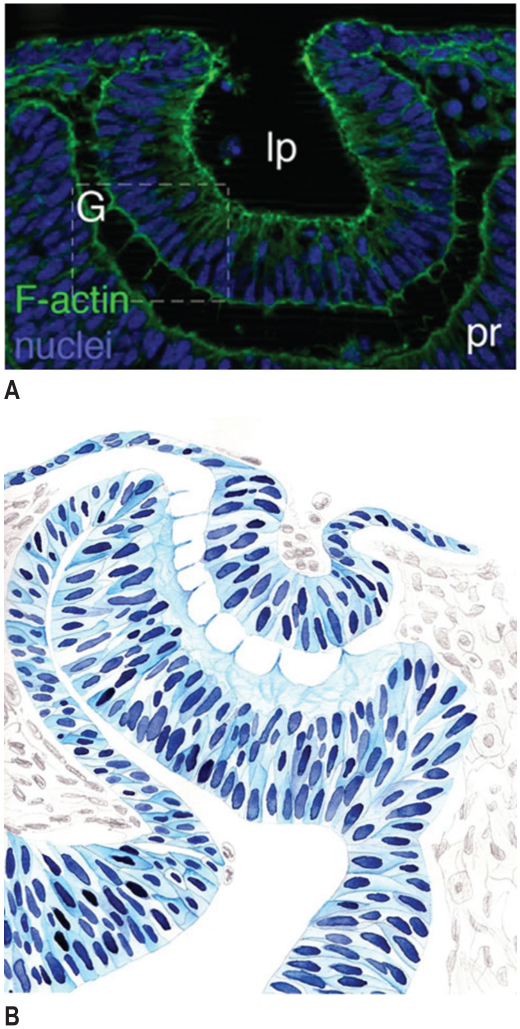

FIGURE 2-3 Transient, F-actin-filled processes connect lens and retina during lens pit–optic cup invagination in the mouse eye (sections from cytoplasmic protrusions interepithelial processes containing filamentous actin (F-actin, green) are clearly seen between the lens and future retina from E9.5 to E11.0 (embryonic days/post-conception). Equivalent interepithelial processes were described in the early human eye many years ago by Ida Mann. lp, lens pit; pr, presumptive retina. Scale bars: 20 μm. Nuclear staining (Hoechst 33258, blue). (Part A reproduced from Chauhan et al., 2009, with permission; Part B redrawn from Mann, 1928)

Following separation of the lens vesicle from the surface ectoderm this layer regenerates and closes the lens pit and ectodermal layer in this region forms the future corneal epithelium (Fig. 2-2E). Around this time (day 39) a 'wave' of mesenchyme passes over the rim of the optic cup, directly beneath the surface ectoderm (Figs 2-1D,E and 2-2C,D). The cells in the first of the three waves of mesenchyme, which lie posteriorly closest to the lens, become flattened and form apicolateral contacts. These contacts become continuous bands of junctional complexes and thus form an endothelium. The other

BOX 2-1 CLINICAL CORRELATES

Coloboma

The word coloboma ('a mutilation') in clinical practice is the term is used to describe a defect in the inferonasal quadrant in the iris, ciliary body or choroid. Colobomas are usually sporadic and bilateral and do not lead to significant complications. They are the result of a failure of closure of the (inferonasal) optic fissure. The consequence is an interference with the normal induction and formation of uveal tissues. A coloboma of the iris appears as an *inferonasal defect* in the stroma, the smooth muscle and the pigment epithelium whereas a ciliary body coloboma is characterized by absence of ciliary processes and presence of a diminutive muscle. The adjacent lens is indented, owing to a failure of formation of the zonular fibres. In the colobomas associated with complex malformations and abnormalities (e.g. trisomy) there may be an ingrowth of mesodermal tissue into the retrolental space with formation of fat and cartilage.

Colobomas involving the retina may be large and may extend to the disk. In the retina adjacent to the coloboma, proliferation of neuroblastic tissue leads to the formation of rosettes. The retinal pigment epithelium and Bruch's membrane and choroidal tissue are absent at the site of the coloboma, although the underlying sclera is normal. In many cases of retinal coloboma there is glial and vascular ingrowth from the retina across the bed of the coloboma. Colobomas have been described in association with mutations in the *CHD7* gene (see Table 2-1). Mutations in the *CHD7* gene are found in approximately 60–65% of individuals with CHARGE syndrome (*C*oloboma, *H*eart disease, *A*tresia choanae, *R*etardation of growth and/or development, *G*enital hypoplasia, *E*ar malformation). Most mutations lead to the production of an abnormally short, non-functional CHD7 protein, which appears to disrupt chromatin remodelling and the regulation of gene expression.

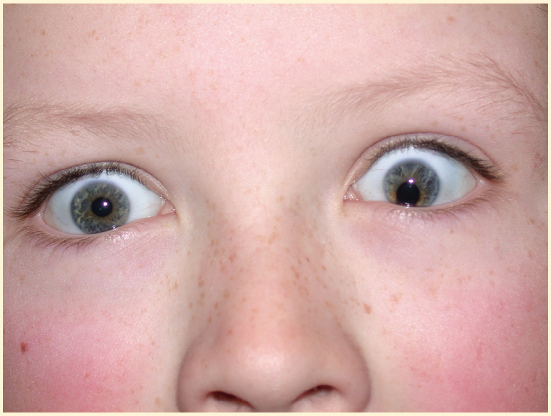

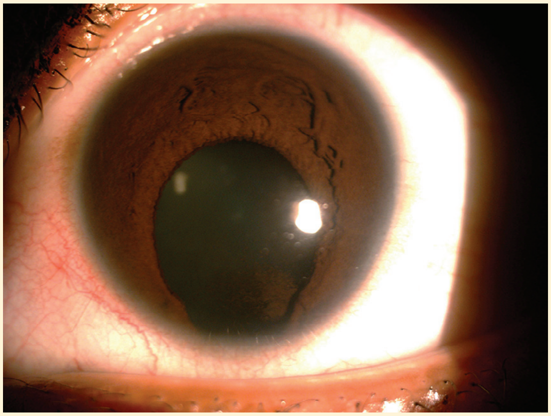

waves of mesenchyme will form the remainder of the cornea, iris stroma and iridocorneal angle mesenchyme.

By the end of the embryonic period (defined as the end of week 8 in humans) the retina can be clearly differentiated into a thin outer layer, which will form the *retinal pigment epithelium* (RPE) and a much thicker inner *neural retina* (Figs 2-1F and 2-2F). These two layers are separated by a narrow *intraretinal or subretinal space*, the remains of the almost obliterated ventricular cavity of the optic vesicle. Melanin first appears in the RPE around 5 weeks (Fig. 2-2C,D) and is visible on external examination of embryos of this gestational age. The neural retina commences its centrifugal differentiation into an *inner* and *outer neuroblastic layer* near its continuity with the optic stalk. The lens cavity disappears as the posterior cells elongate to form the *primary lens fibres* (Figs 2-1D and 2-2D). Mesenchyme, mostly derived from neural crest cells, condenses around the external surface of the optic cup. The innermost layer of this mesenchyme is loose and highly vascular and will form the *choroid* (uveocapillary lamina) (Fig. 2-2C,D). It lies adjacent to a distinct basement membrane formed by the RPE, which is continuous with a similar membrane around the forebrain. Indeed, the choroid is homologous in its embryonic origin with the *pia mater* and *arachnoid* investing the brain. The outer layer of the condensed

TABLE 2-1 **Critical genes in ocular development**

Mouse gene	Human gene	Expression pattern	Ocular defects in mouse mutants	Ocular defects in humans with known mutations in related gene
Sox1, Sox 2, Sox3	*SOX1, SOX2 SOX3*	Central nervous system, sensory placodes, *Sox2* in anterior neural ectoderm and lens placode	Micro-ophthalmia, cataract	Anophthalmia
Otx2	*OTX2*	Anterior neural ectoderm Future forebrain		Range of phenotypes from bilateral anophthalmia to retinal dystrophy
Rax	*RAX*	Anterior neural plate, optic vesicle, developing retina, photoreceptors	Eyeless	Anophthalmia
Pax-6	*PAX6*	Anterior neural plate, optic sulcus/cup and stalk Surface ectoderm (future lens and corneal/ conjunctival epithelium) Weakly expressed in mesenchymal cells	Micro-ophthalmia, cataract, hypoplastic iris, incomplete separation of cornea and iris, corneal defects	Anophthalmia, anterior segment dysgenesis, congenital glaucoma, Peters' anomaly, Axenfeld–Rieger syndrome (aniridia)
Six3	*SIX3*	Presumptive eye field	Lack of neural retina	Holoprosencephaly
Pitx3	*PITX3*	Developing lens vesicle	Persistent lens stalk, malformed lens	Congenital cataract, leucoma, Peters' anomaly
Chd7	*CHD7*	Neuroectoderm, lens vesicle	CHARGE-like features, keratoconjunctivitis sicca	CHARGE syndrome (see Box 2-1)
Maf	*MAF*	Lens placode, lens vesicle, primary lens fibres (transcription factors for α-crystallin gene along with *Sox1*)	Failure of lens fibres to elongate, lens vesicle fails to separate from surface ectoderm	Defects in lens, cornea and iris (coloboma), Peters' anomaly
Foxe3	*FOXE3*	Lens placode	Failure of lens to separate from surface ectoderm	Peters' anomaly, posterior embryotoxon, cataract
Pitx2 Fox1	*PITX2 FOXC1*	Periocular mesenchyme (presumptive cornea, eyelids, trabecular meshwork, extraocular muscle)	Anterior segment abnormalities	Iridogoniodysgenesis, Axenfeld–Rieger syndrome, 50% develop juvenile glaucoma
Crya, Cryb, Cryg	*CRYA, CRYB, CRYG*	Lens	Various forms of cataract	Various forms of cataract

mesenchyme will form the *sclera*, which is homologous to, and indeed continuous with, the dura mater around the optic nerve and brain posteriorly.

In the eighth and last week of the embryonic period (Figs 2-1F and 2-2F), ganglion cell axons grow from the inner retina towards the optic stalk. The axons travel within the stalk towards the brain, thus forming the optic nerve. Other major landmarks in development occurring in the eighth week include formation of secondary lens fibres, lens sutures and the secondary vitreous.

In summary, by the end of the embryonic period the eye comprises a double-layered neural ectoderm-derived optic cup containing a surface ectoderm-derived lens, both enveloped by condensed mesenchyme comprising a dense outer layer (the bulk of the future cornea and sclera) and an inner vascular layer which will form the choroid and stroma of the iris and ciliary body. At this stage the human embryo is 30 mm in length (crown–rump length) and the developing eye is 1.5–2.0 mm in diameter (Figs 2-1F and 2-2F).

GENETIC REGULATION OF EYE DEVELOPMENT

It is now clear that *epigenetic development*, the process by which invertebrates and vertebrates develop their definitive characteristics through the gradual alteration of simpler precursors, is regulated by *cascades of gene expression*. That is, early acting regulatory genes initiate developmental processes and induce the expression of *'downstream' genes*, which may subsequently lead to expression of further genes and so on until genes encoding actual structural and functional characteristics of specific cells and tissues are activated. There is overwhelming evidence that these cascades have been conserved throughout evolution from insects to fish to mammals. The *Drosophila*, or fruit fly, zebra fish and mouse are the most intensively studied experimental species in each respective group and recent research has shed light on some of the critical genes in ocular development, some of which have been remarkably conserved in evolution. The clues from animal studies have aided clinical geneticists to screen the DNA from patients with abnormalities and have helped reveal how mutations in single genes can cause congenital abnormalities, and the eye has provided science with some of the more elegant examples of the relevance of such genetic studies (Table 2-1) (see Ch. 3).

The basic body plan of all animal embryos is initially established by a class of regulatory genes called *selector* or *switch genes* which, like the *maternal effect genes* of the fruit fly, establish longitudinal or antero-posterior (head–tail), dorsoventral and left–right axes. A further class of genes, the *zygotic genes*, which includes *segmentation genes*, is switched on later, after the maternal effect genes. The changes induced by the segmentation genes cause the expression of another class of selector genes, the *homeotic genes*, which encode a region of DNA called a *homeobox* (termed *Hox* genes). These subsequently regulate many downstream genes and thus act as *master control genes*.

Hox gene activation plays a critical role in the differentiation of what initially appear as identical segments in the embryo (induced by earlier expression of segmentation genes) into, for example, cervical, thoracic, abdominal and sacral regions as well as the segmentation of the head and neck, particularly the pharyngeal arches and subdivision of the brain.

Paired-box (Pax) genes encode transcription factors involved in the regulation of several aspects of vertebrate and non-vertebrate early development and, as such, are also considered as *'master control genes'*. In vertebrates two *Pax* genes, *Pax-6* and *Pax-2*, are important eye field transcription factors.

EYE FIELD TRANSCRIPTION FACTORS

In mammals (and one must appreciate that much of this research is performed in mice) the eye field transcription factors (EFTFs) include *Pax-6, Rax, Six3* and *Lhx2* and these are expressed in an overlapping fashion in the anterior neural plate or future forebrain at the paired sites that define the site of the eye primordium or eye field (Fig. 2-4). There is evidence that *Otx2*, a transcription factor essential for forebrain development, may cooperate with *SOX2*, a neural ectoderm transcription factor, to progressively activate *Rax* expression in the eye field, which may in turn activate *Pax-6, Rax, Six3* and *Lhx2*. Evidence of the importance of these genetic transcription factors in early developmental events can be seen by considering the effects of mutations in both mouse models and humans (see Table 2.1).

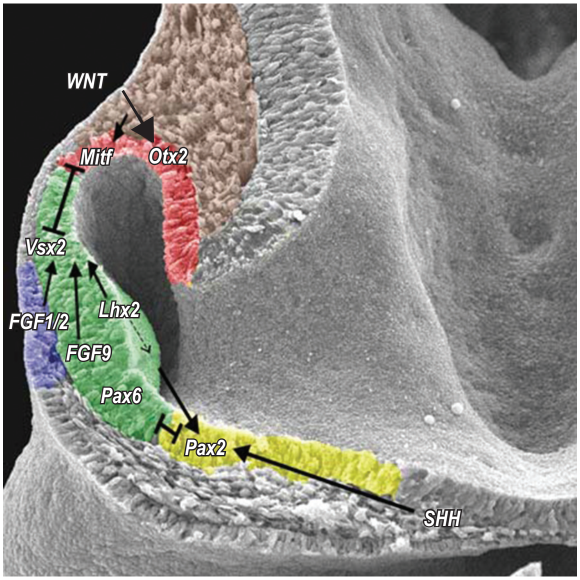

FIGURE 2-4 Signalling networks establish boundaries in the optic vesicle. Dorsal is to the top, and distal is to the left (note the cavity of the optic stalk communicating with the forebrain vesicle (left)). The optic vesicle is regionalized into prospective RPE (red, dorsal), neural retina (green, central) and optic stalk (yellow, ventral). Extracellular signals organize the optic vesicle in part through the activation of transcription factors that specify the tissue type in which they are expressed. These transcription factors cell-intrinsically regulate optic vesicle organization through mutual repression of one another. The lens placode, which expresses fibroblast growth factor (FGF) ligands important for neural retinal specification, is shown in blue. (Reproduced from Heavner W, Pevny L. Eye development and retinogenesis. Cold Spring Harb Perspect Biol. 2012;4:a008391 with permission.)

The primordial eye field is initially one region of anterior neural ectoderm that has two hemispheres that later split into two respective fields, a left and a right, and at least two molecules (*sonic hedgehog (Shh)* and *Six3*) are involved in this early morphogenesis, which, if disrupted, leads to a failure in the splitting, and thus to cyclopia or a single midline eye.

Pax-6 is expressed in the anterior neuroepithelium and lens placode-forming epithelium. Thus it appears that the gene defines a field of cells *competent to differentiate into eye tissues* (neural retina, retinal pigment epithelium, iris epithelium, ciliary body epithelium, lens and corneal epithelium). It also appears to be required to *maintain growth and proliferation* of cells in the optic vesicle (and other regions of the central nervous system). Mice with mutations in *Lhx2* can generate optic vesicles but never form optic cups and it appears as if both this transcription factor and *Pax-6* are needed to form a proper eye cup and together induce *Six6* in the optic vesicle, which is essential for retinal differentiation. The boundaries between the parts of the optic vesicle that ultimately differentiate into neural retina, RPE and optic stalk are predetermined by the differential expression of cell signalling pathways along the dorsoventral and proximal–distal axes, which are in turn regulated by transcription factors (Fig. 2-4). For example, a mutually antagonistic relationship exists between expression of *Pax-2* and *Pax-6*. *Pax-2* is largely restricted to the future optic stalks on the ventral aspect where it appears to be necessary for successful closure of the *choroid fissure,* whereas *Pax-6* is expressed more dorsally and is more involved with formation of the RPE and neural retina. These cell-intrinsic signalling pathways are not the only factors that determine the fate of cells or patterning in the optic vesicle. There are further extrinsic factors including members of the transforming growth factor-β (TGF-β), fibroblast growth factor (FGF) and Wnt families and *Shh* that signal to functionally compartmentalize the optic vesicle. For example, ectopic expression of FGF9 in the presumptive RPE promotes neural retina formation and inactivation of canonical Wnt signalling in the future RPE causes it to transdifferentiate into neural retina.

Periocular mesenchyme is derived from a mixture of neural crest and mesoderm

For many years it was widely held that the middle germ layer, the *mesoderm*, gave rise to most of the mesenchyme and its derivatives in the head and neck region in a similar manner to the pattern of differentiation in the trunk (see eFig. 2-2).

A large body of experimental evidence from avian and mammalian studies has now shown that mesenchyme in the head region is derived from two sources, neural crest (*mesectoderm*) and mesoderm. Neural crest cells originate at the neuroectoderm–surface ectoderm junction of the fore-, mid- and hindbrain regions before fusion of the neural folds. They migrate ventrally into the pharyngeal arches and rostrally around the forebrain and developing optic cup and into the facial region in a highly ordered manner,

BOX 2-2 CLINICAL CORRELATES

Malformations of the neural tube and optic vesicle and genetic factors

These occur in the first month of embryonic life and include:

- *Anophthalmia.* Extremely rare and the result of a failure of formation of the optic vesicle. The orbits do not contain ocular tissue, but the extraocular muscles (mesoderm) and lacrimal gland (ectoderm) are present. Mutations in the *RAX* and *SOX2* genes have been associated with anophthalmia (Table 2-1).
- *Nanophthalmia and microphthalmia* Formation of the optic vesicle without proper subsequent development produces a rudimentary eye in the orbit – nanophthalmia (or dwarf eye). In microphthalmia there is a small but recognizable eye that contains recognizable elements, e.g. lens, choroid and retina. The mouse mutant small eye (*Sey*) has a haploinsufficiency in *Pax-6*. Human mutations in one copy of *PAX6* have aniridia (iris abnormalities) and microphthalmia but homozygous loss causes anophthalmia. Mutations in *SOX2* can also cause microphthalmia.
- *Cyclopia Mutations* in *SHH or SIX3* result in midline defects including cyclopia.
- *Synophthalmia.* Fusion of the two eyes may result from a malformation of the mesenchymal tissue between the optic vesicles or faulty inductive processes. Only rarely is a single eye (cyclops) formed by this mechanism and in most cases there are two recognizable corneas and lenses, and identifiable parts of the iris and ciliary body. The midline sclera and uveal tissue may be absent and the optic nerve may be single or duplicate. This malformation may be associated with a deletion of chromosome 18.

guided in their migration by components of the extracellular matrix, such as fibronectin and glycosaminoglycans (Fig. 2-5). The pattern of appearance and migration of cranial neural crest is closely associated with the expression products of the *homeobox* (*Hox*) gene family within the rhombomeres of the hindbrain. In the face region, neural crest cells contribute significantly to mesenchyme-derived tissues, such as bone, cartilage, connective tissues, meninges and ocular and periocular connective tissues (which normally do not arise from trunk neural crest), as well as giving rise to the usual crest derivatives: melanocytes, dorsal root ganglia equivalents (sensory ganglia of V, VII, IX and X) and parasympathetic ganglia (ciliary, otic, pterygopalatine and submandibular).

At the same time as cephalic neural crest cell migration commences, the optic stalks begin to constrict, thus creating a pathway between the stalk and surface ectoderm into which predominantly mesencephalic crest cells migrate. The migration of these cells ceases when they reach the choroid fissure on the ventral aspect of the optic cup (for further information see eBox 2-2).

Paraxial mesoderm (somites and the less distinct rostrally situated *somitomeres*) forms most of the walls and floor of the brain case, all voluntary muscles of the craniofacial region (including extraocular muscles), all vascular endothelial cells, the dermis and connective tissues of the dorsal region of the head, and the meninges caudal to the prosencephalon. The seven rostrally situated somitomeres differentiate in a segmental manner within the pharyngeal arches and they play an important role in eye development by influencing neural crest cell migration and differentiation and by directly contributing to the periocular mesenchyme (Fig 2-6).

Thus, mesodermally derived mesenchyme contributes more to the periocular connective tissues than experimental studies in birds had previously suggested.

The neural retina and retinal pigment epithelium are derived from neuroepithelium

The thickened portion of the optic vesicle that invaginates, the *retinal disk*, is destined to differentiate into the *neural retina*, while the thinner outermost layer of the optic vesicle is destined to form the RPE. These layers are continuous at the optic cup margin, where a sharp transition in morphology is evident (Figs 2-1B,C and 2-2A–D). The *optic cup margin* will later be the site from which the neuroepithelial component of the iris and ciliary body will arise and ultimately form the *pupil margin*. Because of the invagination of the optic cup the apical aspect of the primitive neural retina comes to lie adjacent to the apical surface of the RPE, thereby obliterating the intraretinal space (Fig. 2-1C).

eBox 2-2

Experimental methods of mapping neural crest cell migration and fate

The fate of neural crest cells from various regions of the neural plate has been mapped in birds (for reviews, see Noden, 1982, 1988; Creuzet et al., 2005); it has also been mapped in mouse and rat embryos (Erickson et al., 1989; Fukiishi and Morriss-Kay, 1992; Trainor and Tam, 1995). Avian experiments have used a variety of cell-tracing methods, including transplanted radiolabelled donor crest cells, chick–quail transplant chimeras and subsequent tracking of quail cells using natural nucleolus marker or anti-quail nuclear antigenic determinant (for review, see Creuzet et al., 2005).

Tracking the fate of neural crest cells in mammals has been aided by the production of mouse mutants with neural developmental anomalies and more recently by utilizing transgenic mice in which a transgene (bacterial *lac Z* 'reporter' gene which codes for β-galactosidase) is introduced into the mouse genome in a position where it will be co-expressed alongside proteins specific for the cell types under investigation, such as peripherin or retinoic acid receptor, which are expressed on neural crest cells during and after migration. The migration pathways of the cells carrying the transfected genome can thus be visualized with appropriate chemical substrates for bacterial β-galactosidase (Mendelson et al., 1994). Other techniques that are unveiling the destiny and differentiation of various cell types in embryonic development include micromanipulative cell grafting and labelling using fluorescent cell markers (Dil and DiO), *in situ* hybridization and chimeric and mosaic mouse models (Osumi-Yamishita et al., 1990; Trainor and Tam, 1995; Collinson et al., 2004) (see Fig. 2-6).

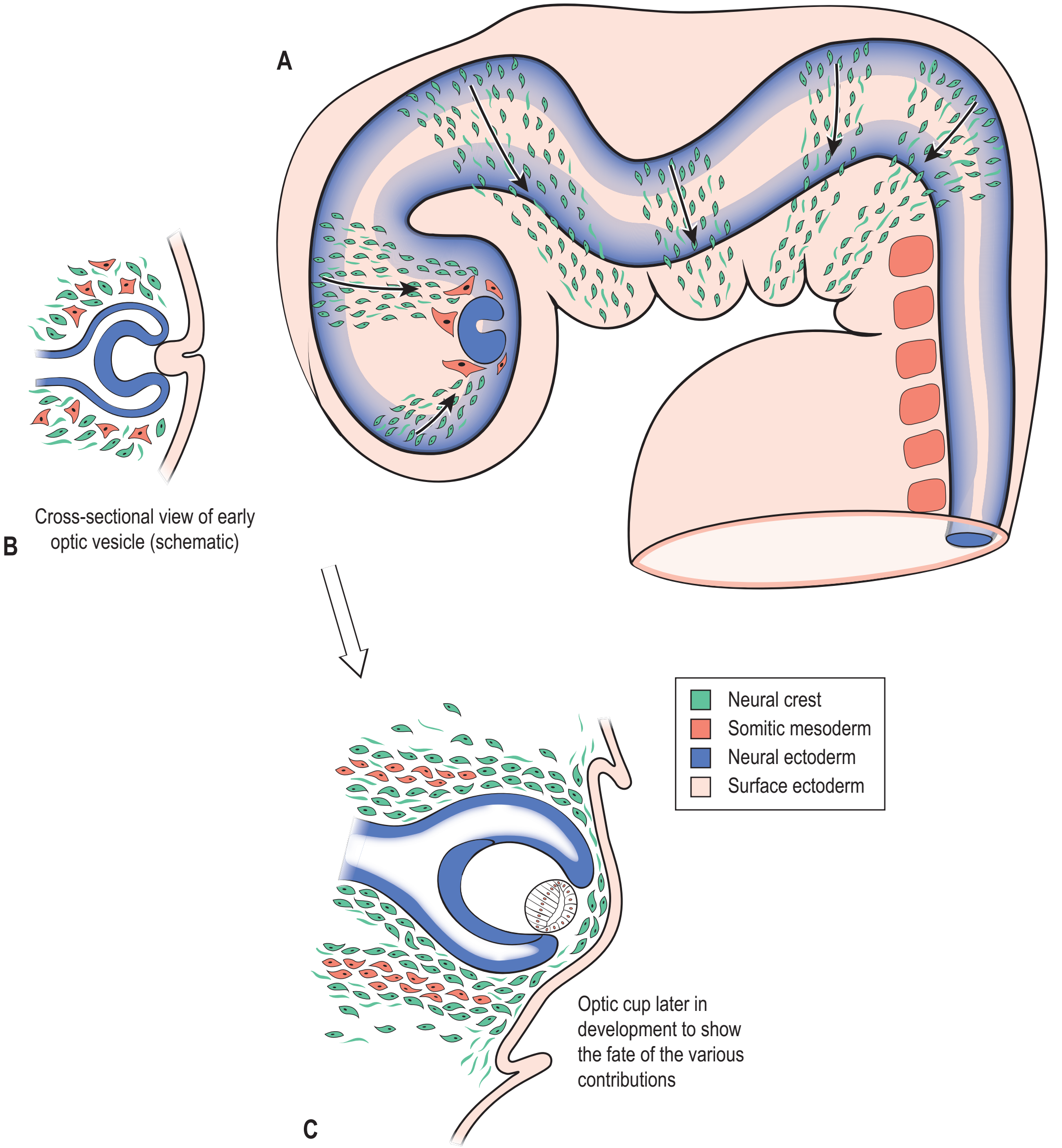
A
B
Cross-sectional view of early optic vesicle (schematic)
Neural crest
Somitic mesoderm
Neural ectoderm
Surface ectoderm
Optic cup later in development to show the fate of the various contributions
C

FIGURE 2-5 Diagramatic summary of the interactions between surface ectoderm, neural ectoderm, neural crest and mesoderm early in development and their final contributions to the eye and periocular tissues.

Just as the ependymal cells that line the developing (and adult) ventricular spaces in the brain are ciliated, the apposing surfaces of the primitive neural retina and future RPE are also ciliated. The cilia of the neural retina are important later in development, in the formation of rods and cones. The cilia of the RPE degenerate.

AXES IN THE NEURAL RETINA

We now understand more clearly how the axes in the retina which are critical to establishing the pattern of retinotopic projection become organized during development. Mouse studies have shown that when the inner layer of the optic cup invaginates to form the neural retina it must establish three primary axes:

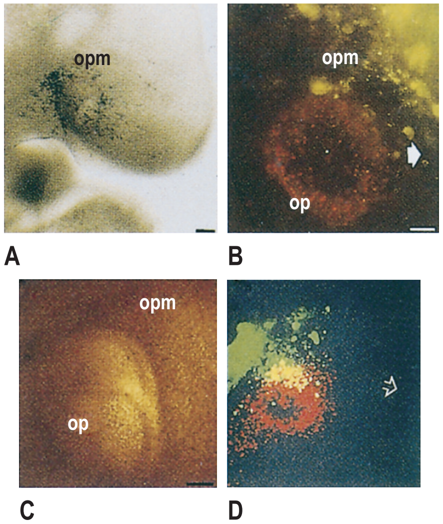

FIGURE 2-6 Colonization of the periocular mesenchyme (opm) by cells derived from somitomere I in the mouse embryo. In (**A**) the cells have been labelled with X-gal (blue) and in (**B**) with DiO (green fluorescence). In (**B**) the red cells are DiI-labelled neural crest cells. (**C**) A bright field view of the same specimen showing the orientation of the optic vesicle (op) and periocular mesenchyme (opm). (**D**) A confocal image showing the co-distribution of the two cell populations showing that neural crest cells are found in both the neural epithelium of the optic vesicle and in the surrounding mesenchyme where they share territory with somitomeric-derived mesenchyme (yellow indicates areas of overlapping distribution but not double-labelled cells). Arrow points rostrally. Bar, 500 µm. (Reproduced from Trainor and Tam, 1995, with permission.)

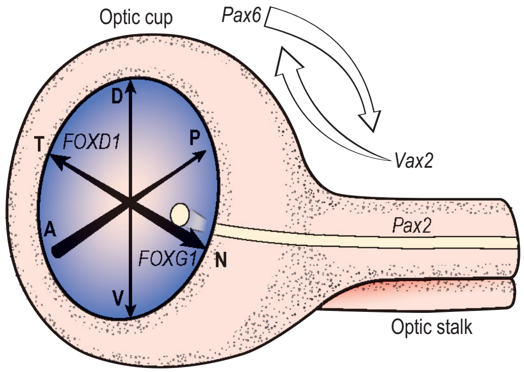

FIGURE 2-7 Regulation of the axes in the optic cup and optic stalk by differential expression of transcription factors which determine the regional specificity or fate of the retina along dorsal-ventral (D-V), naso-temporal (N-T) and antero-posterior (A-P) axes.

dorsoventral, nasal–temporal and anterior–posterior. In addition, the optic stalk must form inferiorly relative to these axes (Fig. 2-7). The signalling that is important in the dorsoventral axes are mediated through the *Pax-6* (highest expression dorsally) and *Pax-2* (highest expression ventrally) transcription factors in addition to the *VAX* family of homeodomain transcription factors. *Vax2*, in particular, has a steep ventral–dorsal gradient, wheras *Vax 1* is expressed on the optic stalk. Naso-temporal patterning is regulated by the forkhead transcription factors *FOXD1* and *FOXG1* (Fig. 2-7).

RETINAL MORPHOGENESIS (Fig. 2-8)

The primitive neural retina consists of an *outer nuclear zone* and an *inner acellular* or *marginal zone*. The outer nuclear zone is homologous with the proliferative neuroepithelium of the neural tube. Both the inner and outer layers of the optic cup rest on their respective basal laminae: that of the inner layer becomes the *inner limiting membrane*, the outer is incorporated into *Bruch's membrane*. Differentiation of the retinal layers commences at the posterior pole and progresses in a centrifugal manner; thus a gradient of retinal differentiation can be seen within an individual eye. Miotic activity in the primitive neural retina is greatest in the outer part of the nuclear zone (Fig. 2-8A). Around 7 weeks of gestation (16–20 mm) newly formed cells migrate in a vitread direction into the marginal zone to form the *inner neuroblastic layer*. The outer nucleated zone is now referred to as the *outer neuroblastic layer* (Figs 2-8B and 2-9A). The two neuroblastic layers are separated by an acellular zone – the *transient layer of Chievitz*. The earliest differentiated cells, which form the inner neuroblastic layer, are future ganglion cells, Müller cells (radial glial) and amacrine cells. Elegant studies using chimeras and mosaic mouse models have revealed that clones of cells radiate in a vitread direction and, as the retina begins to stratify, proliferation and differentiation cells that have arisen from these original 'clones' appear as columns. Subsequently some cells disperse laterally.

The nerve fibre layer becomes identifiable on the inner aspect of the inner neuroblastic layer owing to growth of ganglion cell axons that converge towards the optic stalk. A zone where the processes of cells from the inner neuroblastic layer intermingle (the

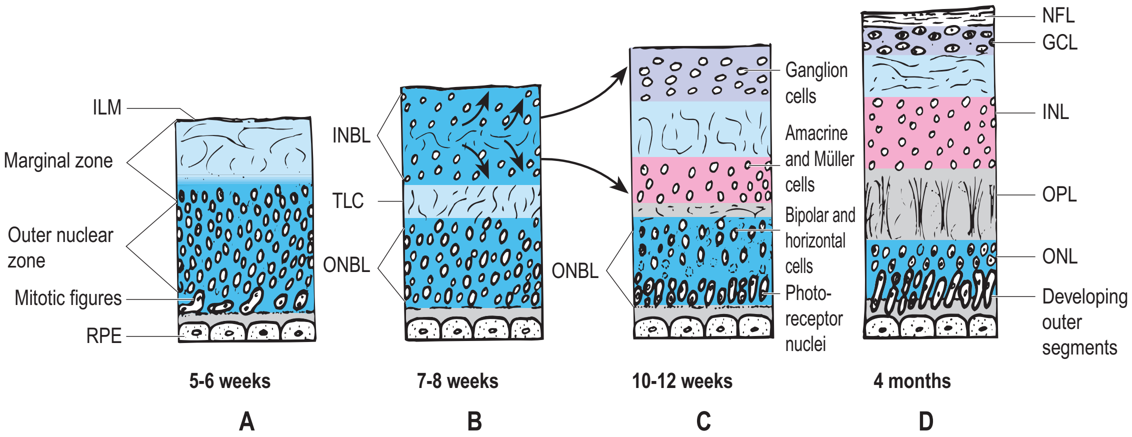

FIGURE 2-8 (**A–D**) Summary of early retinal morphogenesis in the human eye. Arrows indicate major patterns of cell movements. ILM, inner limiting membrane; RPE, retinal pigment epithelium; INBL, inner neuroblastic layer; ONBL, outer neuroblastic layer; TLC, transient layer of Chievitz; NFL, nerve fibre layer; GCL; ganglion cell layer; INL, inner nuclear layer; OPL, outer plexiform layer; ONL; outer nuclear layer.

inner plexiform layer) becomes identifiable at approximately 10.5 weeks of gestation, thereby obliterating the transient layer of Chievitz (Figs 2-8C and 2-9A). A new intermediate nucleated layer, the *inner nuclear layer*, becomes identifiable in the posterior pole retina and already contains the amacrine and Müller cell bodies, and shortly afterwards the bipolar and horizontal cells differentiate from the outer neuroblastic layer and migrate into this new nucleated layer (Fig. 2-8C). The remaining components of the outer neuroblastic layer will form the *outer nuclear layer* containing the cell bodies of the photoreceptors (rods and cones). The zone where fibres from this layer intermingle with those of the inner nuclear layer constitutes the new *outer plexiform layer* (Fig. 2-8D). The *external limiting membrane* (not a membrane *per se*) of the retina is identifiable in the earliest stages as rows of tight junctions between adjacent neuroblasts (Fig. 2-10B).

Further important landmarks in retinal development include:

- synaptogenesis in cone pedicles at approximately 4 months, but not in rod spherules until 5 months
- photoreceptor outer segment formation commences around the fifth month
- horizontal cells become distinguishable around the fifth month
- microglia (resident tissue macrophages) invade the retina via the retinal vasculature and peripheral subretinal space (10–12 weeks onwards)
- the terminal expansions of the Müller cells beneath the inner limiting membrane mature around 4.5 months, at around the same time as their processes can be identified between the rods and cones.

The onset of cellular differentiation in the neural retina is partly dependent on Shh (sonic hedgehog) and FGF signalling. In addition, a number of bHLH (basic helix-loop-helix, transcriptional activators) genes are likely involved in determining neuronal fate. NOTCH1 signalling appears to be involved in regulating the fate of the glial cells, particularly Müller cells.

MACULA AND FOVEAL DEVELOPMENT

Maculogenesis is first evident as a localized increase in ganglion cell density temporal to the optic disk at around 4.5 months. By 6 months, the ganglion cell layer may be eight or nine cells deep in this region. The thickened immature outer nuclear layer consists predominantly of immature cones. In the seventh month, there is a displacement of ganglion cells and formation of a *foveal depression*. There are approximately two layers of ganglion cells in the foveal region in the eighth month and at birth this is reduced to one. By 4 months postpartum the inner nuclear and

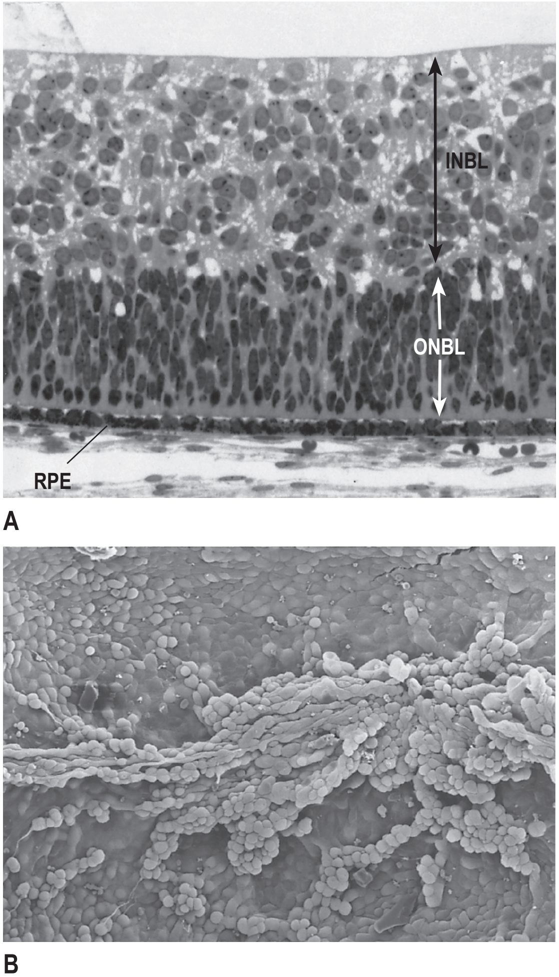

FIGURE 2-9 (**A**) Human fetal retina (12–13 weeks) showing inner and outer neuroblastic layers (INBL, ONBL). The inner neuroblastic layer has commenced differentiation and the transient layer of Chievitz is obliterated. RPE, retinal pigment epithelium. (**B**) Scanning electron micrograph of 'cords' of endothelial and supporting cells (retinal vessel precursors) ramifying on the retinal surface. Original magnifications: **A**, × 115; **B** × 55.

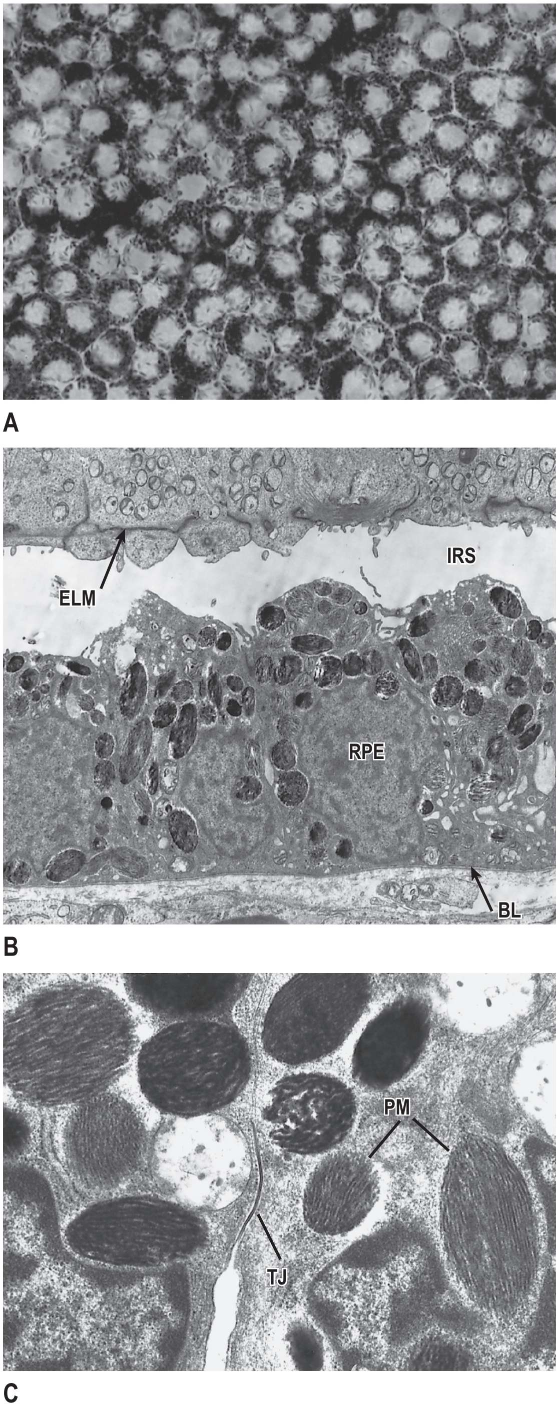

FIGURE 2-10 (**A**) Human fetal RPE (15 weeks) viewed *en face* to demonstrate the regular hexagonal arrangement. (**B**) Transmission electron micrograph of human fetal RPE (12 weeks). ELM, external limiting membrane; IRS, intraretinal or subretinal space; BL, basal lamina. (**C**) Premelanosomes (PM) and tight junctions (TJ) near apices of human fetal RPE (22 weeks). Original magnifications: **A**, × 100; **B**, × 4100; **C**, × 16 000.

ganglion cell layers have receded to the margins of the fovea, leaving only cone nuclei in the foveal region. Elongation of the inner and outer segments occurs over the next few months.

PERIPHERAL RETINA

Until approximately 10–12 weeks of gestation the periphery of the retina extends to within 50–100 μm of the optic cup margin (Fig. 2-2F). By 14 weeks the retina terminates immediately posterior to the newly formed ciliary folds, with minimal pars plana. However, a definite pars plana and a poorly formed ora serrata are present by 6 months. The pars plana and the region from the ora serrata to the equator of the eye continue to grow after birth with continued growth of the eyeball, which occurs up to 2 years of age. The area of the retina is approximately 600 mm^2 at birth and reaches 800 mm^2 by 2 years.

DEVELOPMENT OF RETINAL VASCULATURE

The vessel incorporated into the choroidal fissure is the *hyaloid artery*, a branch of the ophthalmic artery, itself a branch of the internal carotid artery (Fig. 2-11A). The hyaloid artery, upon emerging from the centre of the optic stalk, spreads between the lens surface and the marginal zone of the primitive neural retina (lentoretinal space). With growth of the optic cup and formation of the vitreous cavity, the hyaloid artery elongates and courses through the primitive vitreous, within the *hyaloid canal*, to reach the posterior lens surface.

Early in the fourth month of development, temporal clusters, or angiogenic buds, develop from the hyaloid vessels at the optic disk. These strands consist of endothelial cells, future glial cells and macrophages (Fig. 2-9B). The endothelial cells are canalized and form new vessels that course along the nerve fibre layer towards the peripheral retina at approximately 0.1 mm per day, to reach the ora serrata by the eighth month. At the same time, dividing vascular endothelial cells penetrate the depth of the neural retina to the outer border of the outer nuclear layer, a process not completed until the ninth month. Here they form a polygonal network of vessels, the outer retinal plexus.

The intraneural portion of the hyaloid vessels becomes the *central retinal artery*. Developing capillaries are united by immature punctate tight and gap junctions and their basal laminae are incomplete.

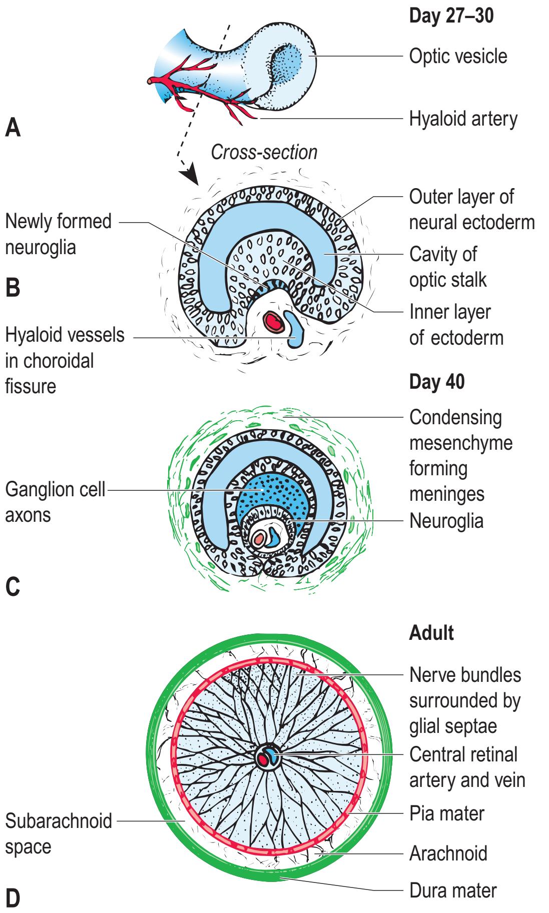
Day 27–30
Optic vesicle
Hyaloid artery
A
Cross-section
Newly formed neuroglia
B
Hyaloid vessels in choroidal fissure
Outer layer of neural ectoderm
Cavity of optic stalk
Inner layer of ectoderm
Day 40
Condensing mesenchyme forming meninges
Neuroglia
Ganglion cell axons
C
Adult
Nerve bundles surrounded by glial septae
Central retinal artery and vein
Pia mater
Arachnoid
Dura mater
Subarachnoid space
D

FIGURE 2-11 Diagram summarizing the major events in the development of the optic stalk and optic nerve.

DEVELOPMENT OF THE RETINAL PIGMENT EPITHELIUM

One of the most dramatic events in eye development is the appearance of melanin in the embryonic RPE, which occurs as early as 28 days after fertilization (6–7 mm embryo). The RPE arises from the proximal cells of the optic vesicle which go on to form the outer layer of the optic cup (Fig. 2-1B,C). The presumptive RPE expresses the bHLH transcription factor *Mitf* (Fig. 2-4) and this appears to be dependent on canonical WNT signalling as evidenced in mice where inactivation of this signalling pathway causes the RPE to revert

BOX 2-3 CLINICAL CORRELATES

Retinopathy of prematurity

Altered oxygen tension has complex effects on the formation of capillary networks in the retina. Placing premature infants in high oxygen environments (to aid in their respiration due to the immature state of their lungs) can result in a delay or reduction in retinal vascularization. Upon returning to normal oxygen tension, the retinal tissues experience a relative hypoxia which induces the release of angiogenic factors (VEGF), resulting in episodes of abnormal neovascularization within the retina and vitreous, known clinically as *retinopathy of prematurity* (ROP) or *retrolental fibroplasia*. Progress in modern neonatal care in many countries has made possible the survival of premature infants born at gestational ages and birth weights that were once considered to offer little hope of viability. The incidence of ROP increases in prematurity as both a function of reduced birth weight and lower gestational age from approximately 40% in infants with birth weights 1101–1200 g to more than 90% in those with birth weights 501–600 g (image courtesy of Dr S. Tahija.)

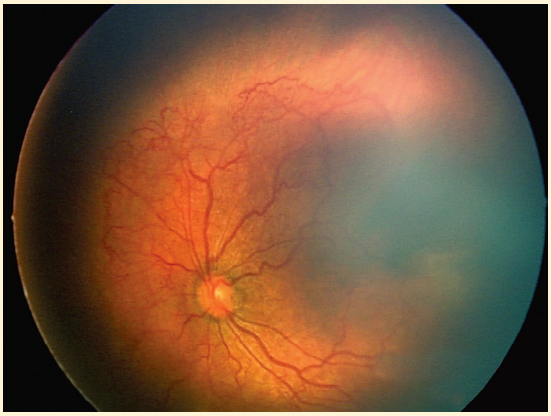

to neural retina. Initially the RPE comprises a mitotically active pseudostratified columnar ciliated epithelium (Fig. 2-2B,C). The cilia disappear as melanogenesis commences. The RPE cells become hexagonal in shape and homogeneous in size (Fig. 2-10A) and in section appear as simple cuboidal epithelium (Fig. 2-10B), although a columnar morphology is maintained in the peripheral retina for a longer period. By the fourth month the RPE has only minimal apical microvilli, few or no basal infoldings, primitive basolateral interdigitations (Fig. 2-10B), mature apical junction complexes and intracytoplasmic premelanosomes (Fig. 2-10C). Mitotic activity appears to take place early in development and is reputed to have ceased by birth; therefore, growth of the eye, and consequently of the RPE itself, is accommodated by hypertrophy or enlargement of the existing cells. A component of the future five-layered Bruch's membrane, the RPE basal lamina, is recognizable at the optic cup stage. Collagen fibrils are subsequently laid down beneath the basal lamina around 10 weeks; the first evidence of the elastic fibre layer can be detected around 3.5 months and by mid-term the elastic layer forms a fenestrated sheet.

Optic nerve and disk development

The hollow optic stalk forms a connection between the cavity of the forebrain (future third ventricle of the brain) and the cavity of the developing optic vesicles (Figs 2-1B and 2-2A). It is formed by the constriction of the proximal portion of the vesicle, particularly on the dorsal aspect concurrent with the expansion of the distal part (Fig. 2-11A). At this stage of morphogenesis (26–28 days) the central hollow fluid-filled stalk is lined by neuroectodermal cells. Invagination of the optic stalk at the 'choroidal' fissure, on the ventral aspect, which occurs simultaneously with invagination of the optic vesicle, results in a double layer of neuroectoderm with narrowing and eventual obliteration of the intervening fluid-filled cavity (Fig. 2-11B). The invagination process in the distal and ventral portions of the stalk leads to the incorporation of the hyaloid vessels and surrounding mesenchyme (Fig. 2-11B). The lips of the optic stalk start closing over the hyaloid vessels near the forebrain (5–6 weeks) and gradually extend distally. This fusion lags behind that of the cup. The optic stalks lie at approximately 65° to the mid-sagittal plane, compared with 40° in the adult.

Axons from developing retinal ganglion cells grow towards the optic stalk and, upon reaching the optic disk, change direction and course towards the brain among the inner neuroectodermal cells of the developing optic nerve. The choroidal fissure closes soon after, and by 6 weeks the optic nerve contains numerous axons that surround the hyaloid artery and vein. The outer neuroectodermal layer of the stalk differentiates into the peripheral glial mantle and the glial component of the lamina cribrosa. A cone-like structure,

BOX 2-4 CLINICAL CORRELATES

Malformations of the optic nerve head

When there is a failure of closure of the posterior part of the optic fissure, the optic nerve head is deformed by a *coloboma*, located inferonasally and associated with bulging of the sclera (scleractasia). The coloboma may take the form of a small recess (*optic pit*) at the rim of the disk where herniation of the retina occurs into the meninges and adjacent optic nerve. The clinical importance of an optic pit lies in its association with visual loss as a result of leakage from the pit and exudation of fluid beneath the macula.

AXIAL COLOBOMA OR 'MORNING GLORY SYNDROME'

There are numerous names for a symmetrically enlarged and excavated optic disk – a condition that may be unilateral or bilateral. The most extreme axial malformation is the 'morning glory syndrome', so called because of the similarity to the American flower of the same name. This malformation is complicated by severe visual dysfunction and characterized by retrodisplacement of the optic disk into the meninges of the optic nerve. The abnormality is the result of a defect in mesodermal organization in the disk; the lamina cribrosa is not formed and there is fat and smooth muscle in the meninges.

Bergmeister's papilla, consisting of glial cells and the remnants of hyaloid vessels, may persist at the optic nerve head in some individuals. An outer layer of condensed mesenchyme forms the optic nerve dura, which blends with the sclera. The glial septae surrounding the nerve bundles are composed of astroglia that differentiate from the cells of the inner layer of the optic stalk. The latter also gives origin to the oligodendroglia that surround the individual axons and are myelinated as far as the posterior margin of the lamina cribrosa. The nerve is displaced nasally during the third month by enlargement of the temporal side of the eye.

Development of the fibrous coat of the eye

Around weeks 6 and 7, periocular mesenchyme, derived from the neural crest and mesoderm (Fig. 2-5), begins to condense around the optic cup (Figs 2-1D,E and 2-2C,D). This mesenchyme can be differentiated into an *inner vascular layer* (uveocapillary lamina), which forms the stroma of the choroid, ciliary body and iris, and an *outer fibrous layer*, which will form the sclera and cornea.

DEVELOPMENT OF THE SCLERA

Mesenchymal condensation is most conspicuous at the future site of insertions of the extraocular muscles (limbal–equatorial region) (Fig. 2-2D). In the third month, active fibroblasts are already embedded in an irregular matrix of collagen, elastic fibrils and glycosaminoglycans. By 12 weeks a well-formed fibrous coat envelops the eye posteriorly as far as the optic nerve where the connective tissue forms a perforated plate, the lamina cribrosa, through which glia-covered ganglion cell axons pass.

DEVELOPMENT OF THE CORNEA

The *surface ectoderm* that seals over the lens pit forms the future corneal epithelium. It is a stratified squamous epithelium of three or four layers, the basal layer of which rests on a thin basal lamina. The first 'wave' of mesenchymal cells that passes over the optic cup margin migrates centripetally in the space between the anterior surface of the lens and the surface ectoderm to form the *corneal endothelium* (around 33 days) (Fig. 2-2C). Around day 49 a *second 'wave' of mesenchyme* commences migration from the optic cup margin and penetrates the space between the basal surface of the corneal epithelium and endothelium to form the corneal stroma (Fig. 2-2D). Both waves of mesenchyme are derived from the neural crest. The epithelium, which is continuous with the surface ectoderm, becomes stratified (three or four layers) and over the next few weeks the eyelids form and fuse (week 9–10) (Fig. 2-2D). Around 8 weeks the first evidence of loosely arranged collagen fibres can be detected amidst the actively synthetic fibroblasts, now known as *keratoblasts*. Within the corneal epithelium an intermediate layer of wing cells does not appear until the fourth or fifth month (Fig. 2-12A). The endothelium, which until now has been a double layer, becomes initially a simple cuboidal and ultimately a simple squamous layer resting on a thick basal lamina – the precursor of *Desçemet's membrane* (Fig. 2-12B). By this time all the corneal layers are present with the exception of *Bowman's membrane*, which becomes identifiable by 5 months as an acellular collagenous zone beneath the epithelium (Fig. 2-12A). The stromal collagen bundles become organized into highly oriented *lamellae*, and the keratoblasts mature into long flattened *keratocytes*. This maturation process commences first in the

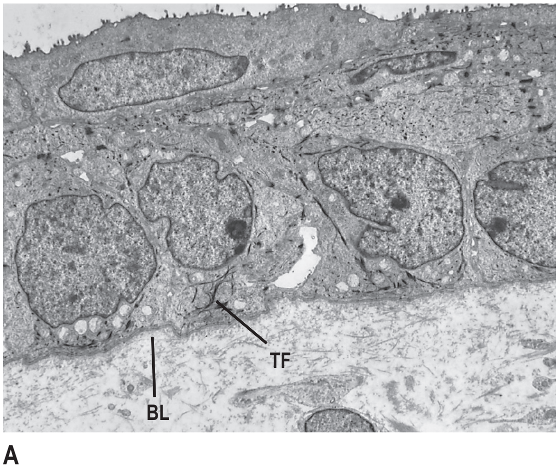

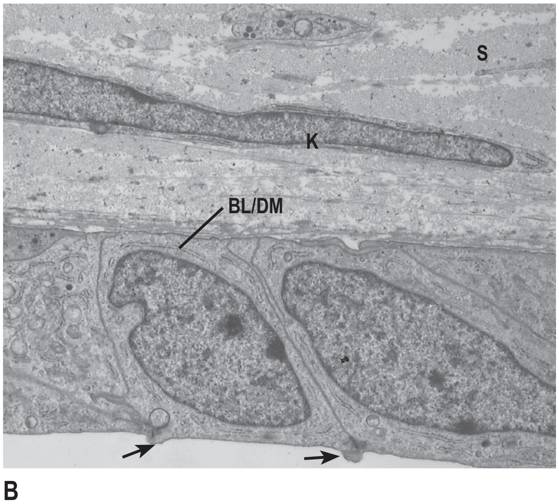

FIGURE 2-12 Transmission electron micrographs of human fetal cornea (16 weeks). (**A**) Epithelium consists of three cell layers joined by desmosomes resting upon a thin basal lamina (BL). Note the electron-dense intracytoplasmic tonofilaments (TF). (**B**) Developing endothelial cells. Note that the cells are cuboidal with apical junctional complexes (arrows) and rest on the basal lamina (BL/DM), which is already showing evidence of thickening to form the future Desçemet's membrane. K, keratoblast; S, collagenous stroma. Original magnifications: **A**, × 2300; **B**, × 5000.

posterior or deeper layers of the cornea and progresses more anteriorly or superficially.

Corneal thickness and diameter continue to increase throughout development by both interstitial growth (thickening of lamellae) and appositional growth (addition of new lamellae). The glycosaminoglycan constituents are known to alter during development and are thought to underlie the initial swelling of the cornea (when lids are fused) and its subsequent thinning, which occurs during eyelid opening (24 weeks). Corneal transparency is gradually attained before birth owing to maturation of the superficial lamellae and the hydration activity of the endothelial cells. Innervation of the cornea commences at 3 months and reaches the epithelium at 5 months.

In summary, the cornea develops from the interaction of a surface ectoderm-derived epithelium and neural crest-derived mesenchyme, which gives rise to the deeper layers including Bowman's layer, stroma, endothelium and its thick basal lamina, Desçemet's membrane.

Development of the intraocular contents

LENS DEVELOPMENT (Fig. 2-13)

The thickened disk of ectodermal cells that forms the *lens placode* can be identified at 27 days. Differential elongation of these cells and contraction of their apical terminal bar causes the placode to invaginate, producing a *lens vesicle* with a central depression, the *lens pit*, leading into a hollow *lens cavity* that is connected briefly to the amniotic cavity via the *lens pore* (Figs 2-1C and 2-2B). As the vesicle, surrounded by its basal

BOX 2-5 CLINICAL CORRELATE

Corneal malformation (corneal leucoma)

Corneal opacification may be the result of a failure of the keratocytes to produce collagen fibres arranged in a lamellar array: instead, the pattern resembles sclera (scleralization of the cornea). *Peter's anomaly* is a term used to describe a posterior axial stromal defect associated with incarceration of the pupillary part of the iris at the edge of the defect. A mild form of malformation is the presence of thickening at the periphery of Desçemet's membrane (Schwalbe's line). When this is visible clinically (by gonioscopy) it takes the form of a bow: hence the term *embryotoxon* (Gk. *toxon* – a bow). Broad strands of tissue derived from the iris are sometimes seen in the chamber angle. In *Axenfeld's anomaly*, iridocorneal strands are localized to Schwalbe's line. When 'iris hypoplasia' is present, the malformation is known as *Rieger's (or Axenfeld–Rieger's) anomaly* (see Table 2-1 for gene mutations associated with these abnormalities).

lamina, detaches from the surface ectoderm (10 mm, 33 days) it sinks into the underlying rim of the optic cup (Figs 2-1D and 2-2B). Occasionally, degenerating cells (epitrichial or periderm cells) are seen within the lens cavity (Fig 2-3). It now appears that lens induction is a multi-stage process requiring signalling by bone morphogenetic protein (Bmp) and also probably fibroblast growth factor (FGF) combined with the expression of a number of transcription factors, the most important of which appears to be *Pax-6*. More recently it has been shown that a disk-shaped 'pad' of modified extracellular matrix, consisting of fibronectin, produced by the lens placode, is critical to ensuring that the thickening of the surface ectoderm, destined to become the lens placode, is restricted to the area in close apposition to the optic vesicle and that disruption of this matrix leads to lateral spreading of the placode and failure to develop a lens placode – the so-called 'restricted expansion hypothesis'. These alterations in the surface ectoderm are *Pax-6* dependant as demonstrated by conditional knockout mice experiments.

The posterior cells of the lens vesicle elongate to form the *primary lens fibres* (Fig. 2-13A) and commence synthesis of a new group of intracytoplasmic proteins known as *crystallins*. Fibroblast growth factor, as well as being an inductive signal, is likely to be involved in lens fibre differentiation. The base of each elongating lens cell remains anchored to the basal lamina posteriorly and their apices grow towards the *anterior lens epithelium*, thereby obliterating the lens cavity (Figs 2-1D,E, 2-2C and 2-13B,C). The nuclei migrate forward within the elongated cell body to produce a *lens bow, a row of cell nuclei* with a conspicuous forward convexity (Figs 2-2F and 2-13C,H). Subsequent lens fibres arise from mitotic activity within the *anterior lens epithelium* at the equatorial zone and are known as *secondary lens fibres* (Fig. 2-13D,H). The tips of the secondary fibres extend around the primary fibres and meet at the Y-shaped anterior and posterior *lens sutures* (Fig. 2-13D). Every subsequent generation of fibres throughout embryonic, and indeed later, life is added superficial to the previous layer. Early in embryonic development the lens is nearly spherical or possibly longer in its antero-posterior axis; however, as secondary fibres are added at the equator the lens becomes more ellipsoid, a trend that continues until birth and then into adulthood. Basal lamina material is continually deposited by the lens epithelium on its external aspect and encases the lens in a membranous non-cellular envelope, the *lens capsule*. During embryonic and fetal development the lens receives nourishment via an intricate vascular net, the *tunica vasculosa lentis* (Fig. 2-13E–G), which completely encompasses the lens by approximately 9 weeks.

BOX 2-6 CLINICAL CORRELATE

Malformations of the lens

The lens is particularly susceptible, during its formation and early growth, to intrauterine toxic insults such as rubella (German measles). The disorganization that ensues causes degeneration in the fibres and visible opacities in the lens (congenital cataract). If the lens fibre cells recover, the opacities become buried in the inner part of the cortex by the newly formed lens fibres.

Small (microphakic) or round (spherophakic) lenses do not exhibit a strikingly abnormal histology. Axial bulges on the anterior and posterior surfaces of the lens (anterior and posterior lentiglobus) are presumed to be the result of abnormalities in the lens epithelium and the lens capsule.

DEVELOPMENT OF THE VITREOUS AND HYALOID SYSTEM

At 5 weeks of gestation the lentoretinal space is narrow and occupied by the *primary vitreous* (Figs 2-1C,D, and 2-14A), which consists of the *hyaloid artery* and its branches, the *vasa hyaloidea propria*, which become incorporated into the optic cup through the choroidal fissure (Fig. 2-13G).

The *tunica vasculosa lentis* has two sources. The first (visible at around 5 weeks) is a series of capillaries that arise from the hyaloid vessels and form a palisade-like network of vessels, the *capsulopupillary vessels*, around the equator of the lens (Figs 2-13F and 2-14B). This capillary network anastomoses with the anterior component of the tunica, the *pupillary membrane* (*lamina iridopupillaris*), on the anterior lens surface (Fig. 2-14B). The pupillary membrane vessels are derived predominantly from branches of the long posterior ciliary arteries, which form an annular vessel close to the optic cup margin and whose branches pass over the rim of the optic cup to supply the anterior portion of the lens (Fig. 2-14B). The hyaloid system has no veins and venous drainage occurs anteriorly via the

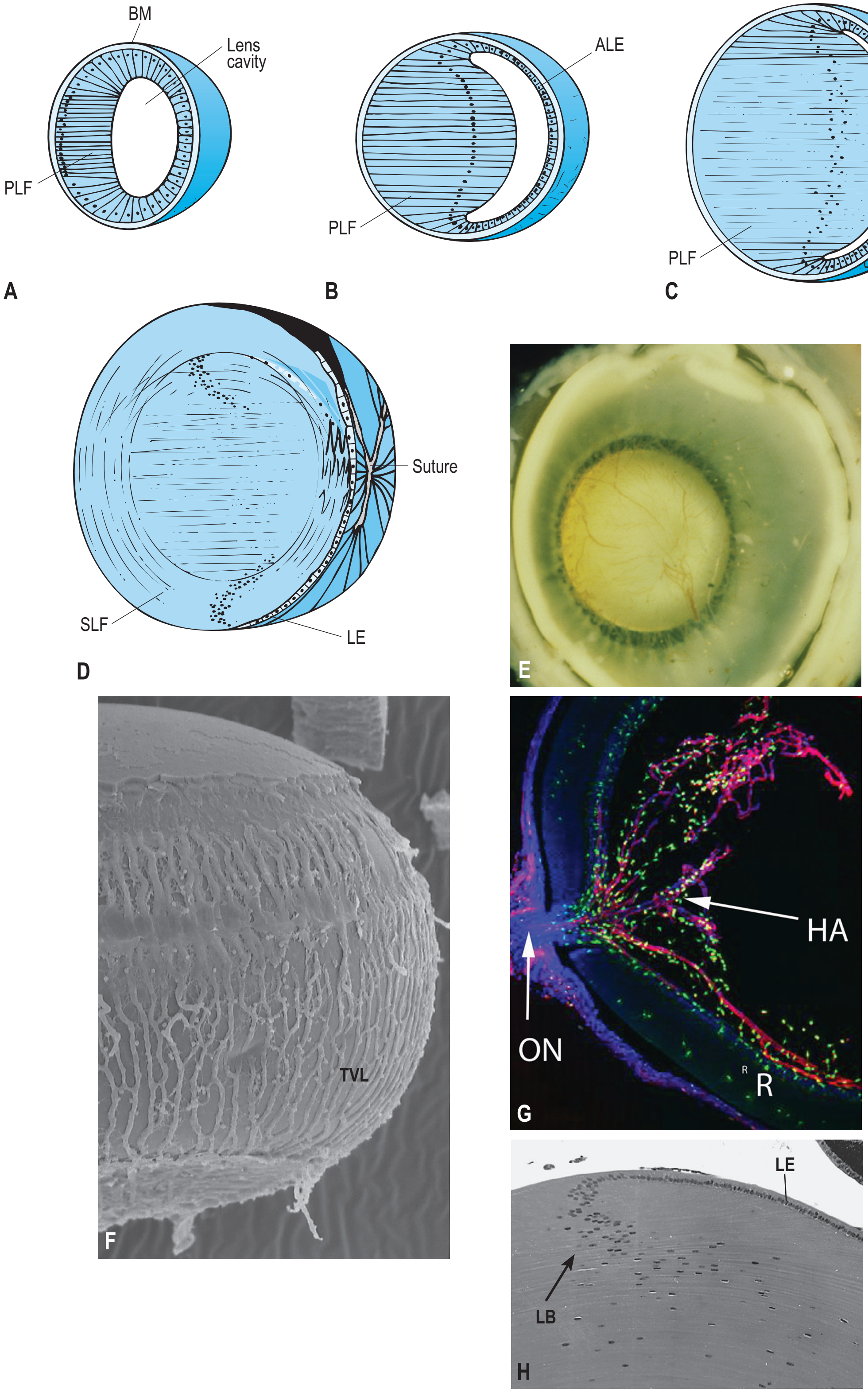
BM
Lens cavity
PLF
A
ALE
PLF
B
PLF
C
Suture
SLF
LE
D
E
TVL
F
HA
ON
R
R
G
LE
LB
H

FIGURE 2-13
For legend see opposite page.

FIGURE 2-13 (**A–D**) Diagrammatic summary of lens morphogenesis. BM, basement membrane of lens cells; ALE, anterior lens epithelium; PLF, primary lens fibres; SLF, secondary lens fibres; LE, lens epithelium (former anterior lens epithelium). (**E**) Macroscopic view of human fetal eye (20 weeks) with posterior segment removed to reveal the hyaloid artery (HA) and tunica vasculosa lentis (TVL) around the lens. Original magnification: × 10. (**F**) Scanning electron micrographic view of a fetal rat lens surrounded by the fine vessels of the TVL. Note the small spherical macrophages associated with the vessels on the lens surface. Original magnification: × 95. (**G**) Frozen section of P0 mouse eye (day of birth) showing hylaoid artery emerging from the optic nerve and ramifying in the vitreous. The lens has been removed. The mouse is a Cx$_3$cr1-GFP transgenic mouse in which all myeloid-derived cells (monocytes, macrophages, microglia) are fluorescent green. Vessels are stained red and blue represents nuclei: × 30 (Courtesy of Dr Wai Wong). (**H**) Fetal lens showing the lens bow (LB) arising from the equatorial region. LE, lens epithelium. Original magnification: × 90.

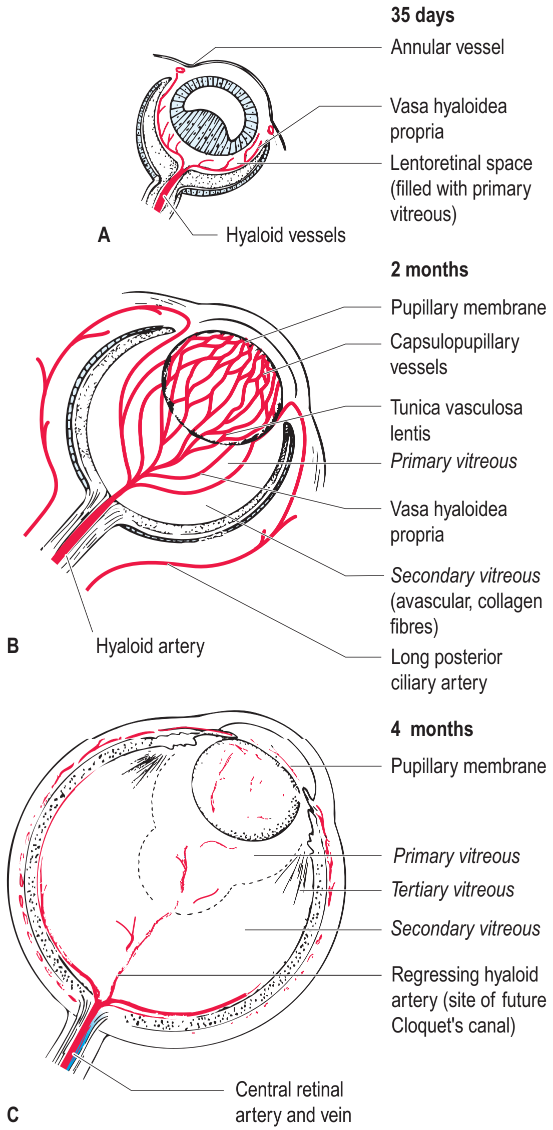
35 days
Annular vessel
Vasa hyaloidea propria
Lentoretinal space (filled with primary vitreous)
A
Hyaloid vessels
2 months
Pupillary membrane
Capsulopupillary vessels
Tunica vasculosa lentis
Primary vitreous
Vasa hyaloidea propria
Secondary vitreous (avascular, collagen fibres)
B
Hyaloid artery
Long posterior ciliary artery
4 months
Pupillary membrane
Primary vitreous
Tertiary vitreous
Secondary vitreous
Regressing hyaloid artery (site of future Cloquet's canal)
Central retinal artery and vein
C

FIGURE 2-14 Early development of the vitreous and hyaloid system.

pupillary membrane and the uveal vessels. The expression of vascular endothelial growth factor by the lens is thought to be a critical molecular event in the formation of the hyaloid vascular system.

An avascular *secondary (definitive) vitreous* composed of finely fibrillar material is deposited behind the primary vitreous between 5.5 and 12 weeks. The primitive hyalocytes, widely recognized as belonging to the mononuclear phagocyte system, arise at this time and likely derive from the same population as the macrophages associated with the hyaloid artery and tunica vasculosa lentis (Fig. 2-13G). These cells also give rise to the microglia of the retina and are all derived in embryonic life from yolk sac precursors – namely the origin of haematopoeitic cells prior to this role being assumed by the liver and later the bone marrow. Mesenchymal cells in the adventitia of vitreal vessels likely contribute to the vitreous matrix. Much of the hyaluronan and collagen (type II) in the vitreous gel is added after birth.

Around the end of the third month distinct condensations of secondary vitreous become evident in the space between the optic cup margin and the lens. The fibres, which are firmly attached to the inner limiting membrane of the retina in the developing pars plana region, are sometimes referred to as the *tertiary vitreous* and, when mature, will form the *vitreous base* (Fig. 2-14C). It seems likely that the non-pigmented ciliary epithelial cells in this region are responsible for synthesis of the tertiary vitreous and zonular fibres.

During the fourth month the remains of the primary vitreous, the hyaloid vessels and the vasa hyaloidea propria, together with the tunica vasculosa lentis, begin to atrophy (Fig. 2-14C). The regression occurs contemporaneously with the formation of the retinal vasculature. The remnant of the course these vessels took through the vitreous is evident in the adult as a narrow fluid-filled central channel, Cloquet's canal. Macrophages play an important scavenging role in the regression of the hyaloid vessels (Fig. 2-13F,G) and may also be responsible for the induction of apoptosis of vascular endothelial cells. Small portions of the pupillary membrane may persist in otherwise normal newborn eyes.

BOX 2-7 CLINICAL CORRELATES

Malformations of the vitreous and hyaloid artery system

Normally, the hyaloid system of vessels vanishes completely but in some disorders regression does not occur.

Persistent Tunica Vasculosa Lentis

Persistence of the anterior part of the tunica vasculosa lentis or the pupillary membrane causes deformation of the iris.

Persistent Hyperplastic (Anterior) Primary Vitreous

If the embryonic fibrovascular tissue in the anterior vitreous face persists, the ciliary processes are drawn internally, providing a valuable clinical diagnostic feature that is visible when the pupil is dilated. The lens is opaque in persistent hyperplastic anterior primary vitreous, because a retrolental fibrovascular mass erodes the posterior lens capsule and penetrates the lens cortex. The white retrolental mass (leucocoria) produced by this malformation can lead to a mistaken clinical diagnosis of retinoblastoma.

Persistent (Posterior) Hyperplastic Primary Vitreous

A persistent hyaloid artery and the condensed posterior primary vitreous project from the optic disk and the adjacent retina. Distortion of the disk by prepapillary and preretinal fibrous membranes is associated with radial or falciform folds in the retina. Although many cases appear sporadic, there has been some evidence for mutations in *LRP5* and *FZD4* genes and also defects in apoptosis and WNT signalling indicating these processes may be critical to the normal regression of the hyaloid vascular system. In addition, mice lacking the Arf tumour suppressor protein (usually produced in response to sustained mitogenic activity and important in controlled apoptosis) fail to resorb the hyaloid system and this leads to persistence of pericyte-like cells that proliferate and destroy the lens and retina, causing blindness.

Development of the uveal tract

THE CHOROID

The choroid arises very early in development from the loose vascular layer of mesenchyme that surrounds the optic cup (Fig. 2-1C,D). A palisade layer of vessels lies immediately external to the RPE and forms the basis of the future choriocapillaris. Fenestrations in the endothelium become evident very early in development. This layer of vessels forms communications with the precursors of the posterior ciliary arteries at around 2 months. A second layer of vessels forms at around 4 months external to the future choriocapillaris and consists of thin-walled venous channels that will eventually unite to form rudimentary vortex veins and branches of the long and short posterior ciliary arteries. An intermediate or middle layer (future Sattler's layer) of mainly arterioles forms between the larger vessels (Haller's layer) and the choriocapillaris. The choroidal vessels are initially embedded in a loose collagenous stroma; however, elastic fibres form later in development in the outer choroid (future lamina suprachoroidea) and pigment-bearing melanocytes appear around the seventh to eighth month of gestation.

THE CILIARY BODY

The development of the ciliary body has similarities to iris development because it involves an interaction between mesenchyme and neuroectoderm. Ciliary body and iris development commence in the 11–12th week with indentation of the outer pigmented layer of the neuroectoderm (presumptive pigmented ciliary epithelium) near the optic cup rim by small capillaries in the inner vascular mesenchyme (Fig. 2-15A,B). Investigations of *Hox* gene expression during early ocular development in the mouse revealed a highly restricted expression of *Hox 7.1* in the non-pigmented or inner neural epithelium just behind the optic cup margin at the site of the prospective ciliary body and iris. This gene is expressed 2 days before any morphological evidence of these structures is evident and thus it may be an early molecular marker for regional specialization and differentiation in the eye.

Initially, the inner non-pigmented ciliary epithelium is flat, but as the vascular sprouts enlarge they push inwards to form primitive radial folds. This arrangement of a vascular connective tissue core

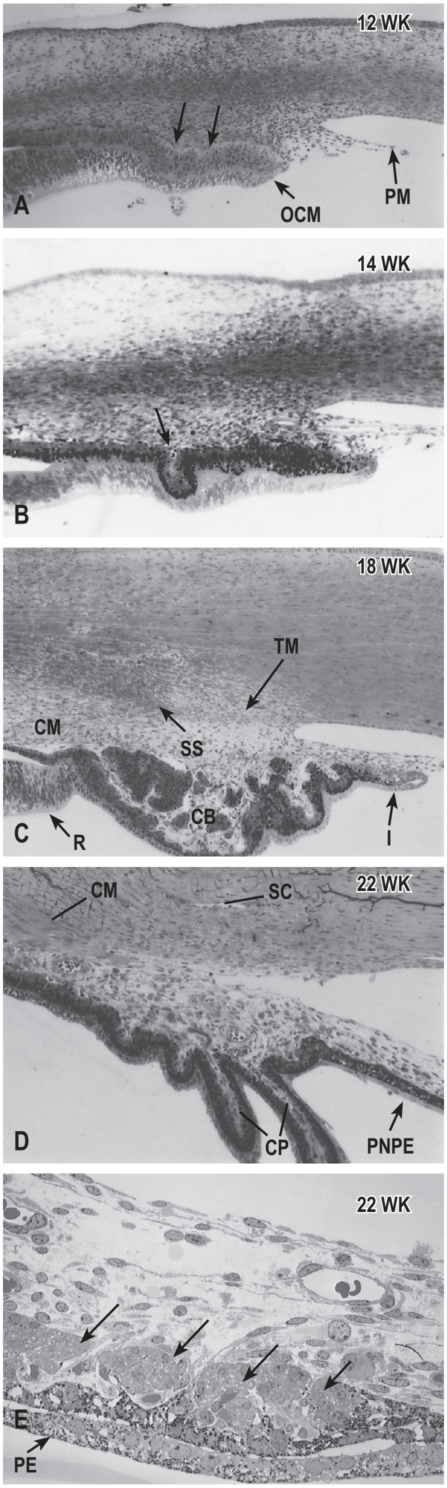

FIGURE 2-15 (**A–D**) Development of the ciliary body and iris in the human fetal eye from 12 to 22 weeks (the lens has been removed in all specimens). Note that the earliest evidence is the vascular mesenchyme indenting the outer neuroectoderm layer (arrows) near the optic cup margin (OCM) in 12- and 14-week specimens. PM, pupillary membrane; CM, ciliary muscle; SS, scleral spur; TM, trabecular meshwork anlage; I, iris; CB, ciliary body; R, retina; SC, Schlemm's canal; PNPE, posterior non-pigmented epithelium of the developing iris; CP, ciliary processes. (**E**) Electron micrograph of iris margin in a 22-week-old human fetal eye to illustrate sphincter pupillae smooth muscle bundles (arrows) differentiating from the anterior pigmented iris epithelium. Note that the posterior epithelium (PE) is showing early evidence of melanogenesis. Original magnifications: **A**, × 75; **B**, × 95; **C**, × 110; **D**, × 80; **E**, × 420.

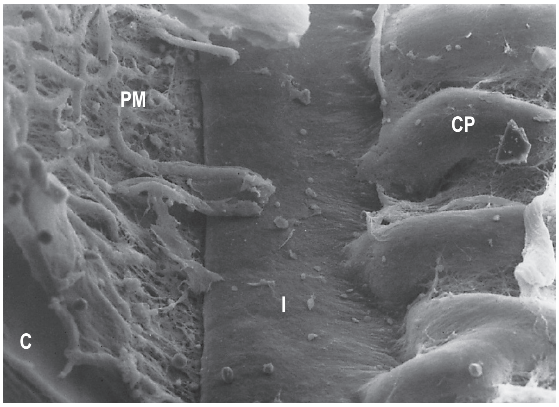

FIGURE 2-16 Scanning electron micrograph of the developing human ciliary processes and iris (20-week fetus). Note the smooth outline of the ciliary processes (CP), the short iris (i), and vessels of the pupillary membrane (PM). C, cornea. Original magnification: × 120.

overlaid by a double layer of ciliary epithelial cells forms the basis of adult ciliary process anatomy (see Ch. 1). The 70–75 radial folds that develop in this manner appear initially as smooth undulations (Figs 2-15B and 2-16); however, between weeks 14 and 22 they increase in height and complexity (Fig. 2-15C,D). Early in development the primitive neural retina terminates immediately posterior to the ciliary folds (Fig. 2-15C), but later a smooth area, the future *pars plana*, separates the two regions and continues to expand during the remainder of gestation with continued growth of the eye. The ciliary epithelium may commence aqueous production as early as 20 weeks, coinciding with concomitant changes in the iridocorneal angle.

The ciliary muscle differentiates at around 15 weeks' gestation from the mesenchyme between the neuroectoderm and scleral condensation external to the early ciliary folds, namely the region that will form the future ciliary body stroma. The longitudinally oriented smooth muscle fibres terminate in the region of the trabecular anlage during fetal development. Circular or radial ciliary muscle fibres do not differentiate until much later in development and indeed are not fully formed until about 1 year of age.

THE IRIS

Until 12–13 weeks of development there are no morphological signs of iris differentiation at the rim of the optic cup, and the cup margin lies posterior to the lateral recess of the anterior chamber (Fig. 2-15A). At around 14 weeks there is an expansion or growth of neuroepithelial cells at the cup margin anterior to the presumptive ciliary body (Fig. 2-15B,C). The optic cup neuroectoderm grows in a centripetal manner between the mesenchyme that has formed the cornea and the anterior lens surface. As it grows it incorporates some of the vessels of the lamina iridopupillaris or pupillary membrane, which lie on the anterior surface of the lens. This vascular mesenchyme is effectively split by the centrally growing neuroepithelial cells into vessels of the iridopupillary membrane, which now face the anterior chamber and form the future iris stroma (Fig. 2-15C,D), and deep to the epithelium the vessels of the capsulopupillaris.

The smooth muscles of the iris, the *sphincter* and *dilator pupillae muscles*, are unique in embryological terms because they differentiate directly from neuroectoderm. The sphincter pupillae differentiation commences before that of the dilator pupillae. Around 13–14 weeks, anterior iris pigment epithelial cells delaminate, lose their melanin, develop intracytoplasmic microfilaments (actin) and dense bodies, and deposit a basal lamina. These are classic characteristics of smooth muscle cells. Cell-to-cell contact (gap junctions) among the smooth muscle cells in this circumferential muscle band is not fully established until 7 months (Fig. 2-15E), and the muscle becomes free in the stroma at around 8 months. The dilator pupillae muscle develops much later, around 6 months, as basal extensions of the anterior or pigmented epithelial layer of the iris (Fig. 2-17), and continues to develop even after birth. These basal extensions are arranged radial to the pupil. This muscle never becomes fully independent of the epithelium because, even in adulthood, it is composed of modified basal processes of the neuroepithelial cells (Fig. 2-17).

During development the posterior (inner) iris epithelium is largely amelanotic. It is continuous with the non-pigmented ciliary epithelium and thus the neural retina. Intracytoplasmic melanin increases in the fourth month, initially near the pupil margin (Fig. 2-15E), and by months 7–8 this layer is heavily pigmented (Fig. 2-17) and the anterior layer has lost its pigment.

Iris innervation, both adrenergic and cholinergic, is not established until late in development. In common with the choroid, pigment-bearing melanocytes are not identifiable in the iris stroma until late in development, around birth or later (Fig. 2-17). The thickness of the stroma and degree of melanogenesis are determining factors in eye colour at birth, and indeed full pigmentation and the pattern of the anterior surface are not complete until a few years postpartum. Blue irides are the result of interference and reflection of light from stromal collagen, whereas a thin stroma

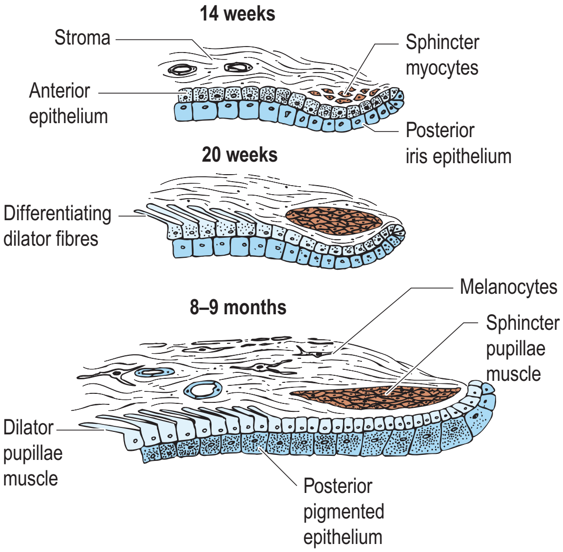
14 weeks
Stroma
Sphincter myocytes
Anterior epithelium
Posterior iris epithelium
20 weeks
Differentiating dilator fibres
Melanocytes
8–9 months
Sphincter pupillae muscle
Dilator pupillae muscle
Posterior pigmented epithelium

FIGURE 2-17 Summary of human iris development showing the differentiation of the sphincter and dilator pupillae from the anterior iris epithelium (neuroectoderm).

BOX 2-8 CLINICAL CORRELATES

Aniridia

Aniridia is a rare autosomal dominant bilateral disease in which there is an apparent absence of the iris. The term is a misnomer because histologically the abnormal iris is seen as a stump of hypercellular stroma, often with an abnormal proliferation of the pigment epithelium. Malformation or hypoplasia of the outflow system occurs in aniridia, as do anterior and posterior cortical lens opacities. The lens may dislocate (ectopia lentis) and the optic nerve may be hypoplastic. It is now well known that aniridia is caused by mutations of the *Pax-6* gene.

BOX 2-9 CONGENITAL GLAUCOMA

Whilst there have been many theories on the aetiology of congenital glaucoma it appears to be a failure of differentiation or alterations in differential growth rates. This is a strong candidate because the few pathological accounts of congenital glaucoma describe the tissue as undifferentiated, or as lacking the typical organized trabeculae and intertrabecular spaces, especially in the outer or cribriform zone. Mutations in *CYP1B1*, *LTBP2* and *MYOC* have been found in a cohort of primary congenital glaucoma patients and the *CYP1B1* gene, which encodes cytochrome P450 1B1, a member of the cytochrome P450 superfamily of enzymes. It may metabolize a signalling molecule involved in eye development, possibly a steroid; hence, mutations may lead to congenital glaucoma.

Chamber Angle Malformation (Goniodysgenesis)

A failure of remodelling of the progenitor tissue in the angle leaves very obvious strands between the iris stroma and the trabecular meshwork, or between the iris and cornea. Abnormalities in the outflow system, 'goniodysgenesis', include hypoplasia of the scleral spur with extension of the ciliary muscle into the outer part of the trabecular meshwork and an excessive amount of trabecular tissue. In addition, there are a number of more generalized abnormalities of the mesenchyme and mesoderm, which display altered iridocorneal angle structure and may manifest as infantile glaucoma. These include posterior embryotoxon, Axenfeld syndrome, Rieger's anomaly and Peters' anomaly. A number of mutations or deletions have been associated with iridogoniodysgenesis (Table 2-1).

may allow the brownish colouration of the posterior epithelium to show through. Later in life, brown irides are the result of heavily pigmented melanocytes within the stroma.

Development of the anterior chamber angle and aqueous outflow pathways

As early as the 12th week a roughly wedged-shaped distinctive mass of mesenchyme, the trabecular anlage, can be identified at the junction of the pupillary membrane and lateral margins of the cornea, namely the future anterior chamber angle (Figs 2-15A and 2-18). The trabecular anlage consists of a dense collection of stellate mesenchymal cells (neural crest-derived) and some loosely arranged extracellular matrix (Fig. 2-18A,D). The deep aspect of the wedge-shaped anlage is characterized by a row of small capillaries (Fig. 2-18A), which most probably have grown in from the capillary plexus on the external surface of the eye (future episcleral plexus) and are thus lined by mesoderm-derived vascular endothelial cells. By weeks 20–22 (fifth month) the connective tissue matrix of the trabecular anlage consists of flattened 'trabecular' endothelial-lined sheets and cords (early trabeculae) separated by intervening spaces (Fig. 2-18). On the deep aspect of the fetal trabecular meshwork the collection of small capillaries fuses to form a single elongated slit-like vessel, the *canal of Schlemm*, lined by endothelial cells that are continuous with those of the collector channels and episcleral vessels. The characteristic 'giant vacuoles' (eFig. 2-3B) in the canal endothelium that are responsible for the passage of aqueous across the inner wall of the canal (see Ch. 1) appear at around 18–20 weeks of gestation.

During the remainder of fetal development the meshwork becomes further specialized into cord-like inner uveal trabeculae, numerous intermediate layers of lamellar corneoscleral trabeculae, and a deep loosely arranged cribriform meshwork (Fig. 2-18B). The scleral spur is formed by month 4–5 (Fig. 2-15C). Scanning electron microscopy of the developing angle reveals that the inner aspect of the developing uveal meshwork is incomplete and numerous perforations allow communication between the anterior chamber and the spaces of the developing meshwork from 15 weeks onwards (Fig. 2-18C).

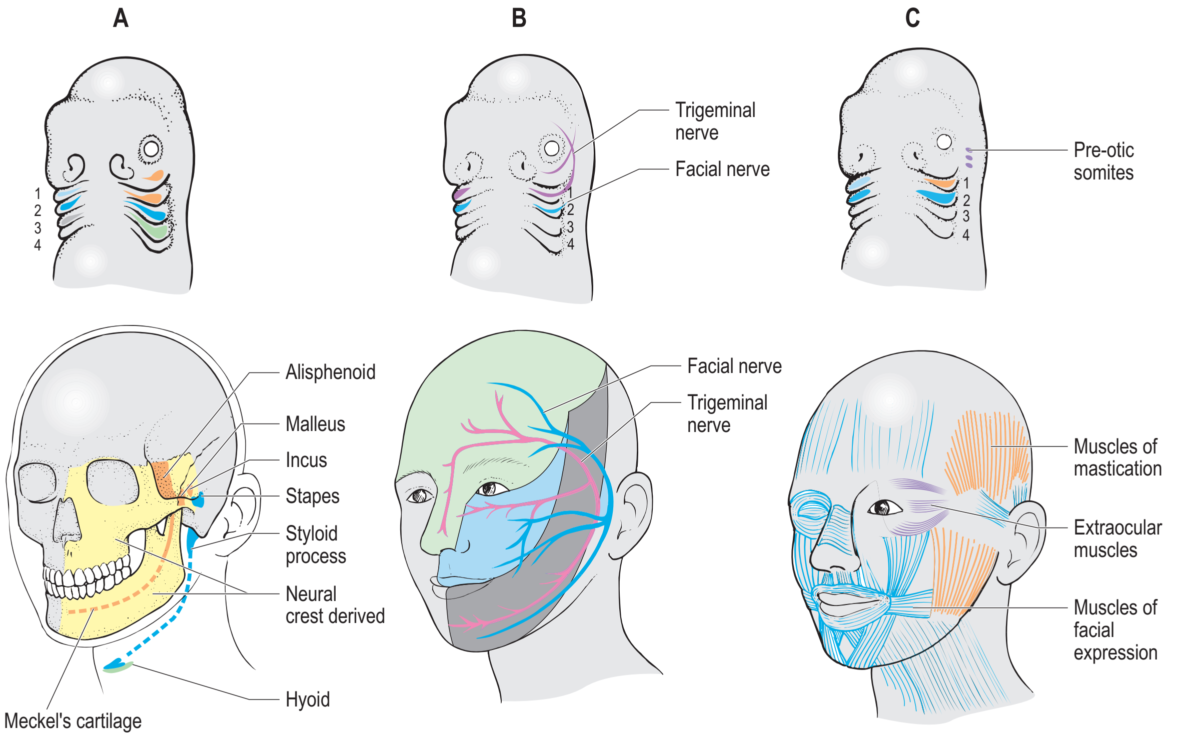
A
B
C
1
2
3
4
Trigeminal nerve
Facial nerve
Pre-otic somites
Alisphenoid
Malleus
Incus
Stapes
Styloid process
Neural crest derived
Hyoid
Meckel's cartilage
Facial nerve
Trigeminal nerve
Muscles of mastication
Extraocular muscles
Muscles of facial expression

eFIGURE 2-3 Summary of the embryonic pharyngeal arches and their derivatives in the adult. (**A**) Arch cartilages (yellow shaded areas represent bones derived from the neural crest); (**B**) arch nerves and the sensory territories of the three divisions of the trigeminal nerve shaded green (V_1), blue (B_2) and grey (V_3) (see also Fig. 1-55); and (**C**) arch musculature.

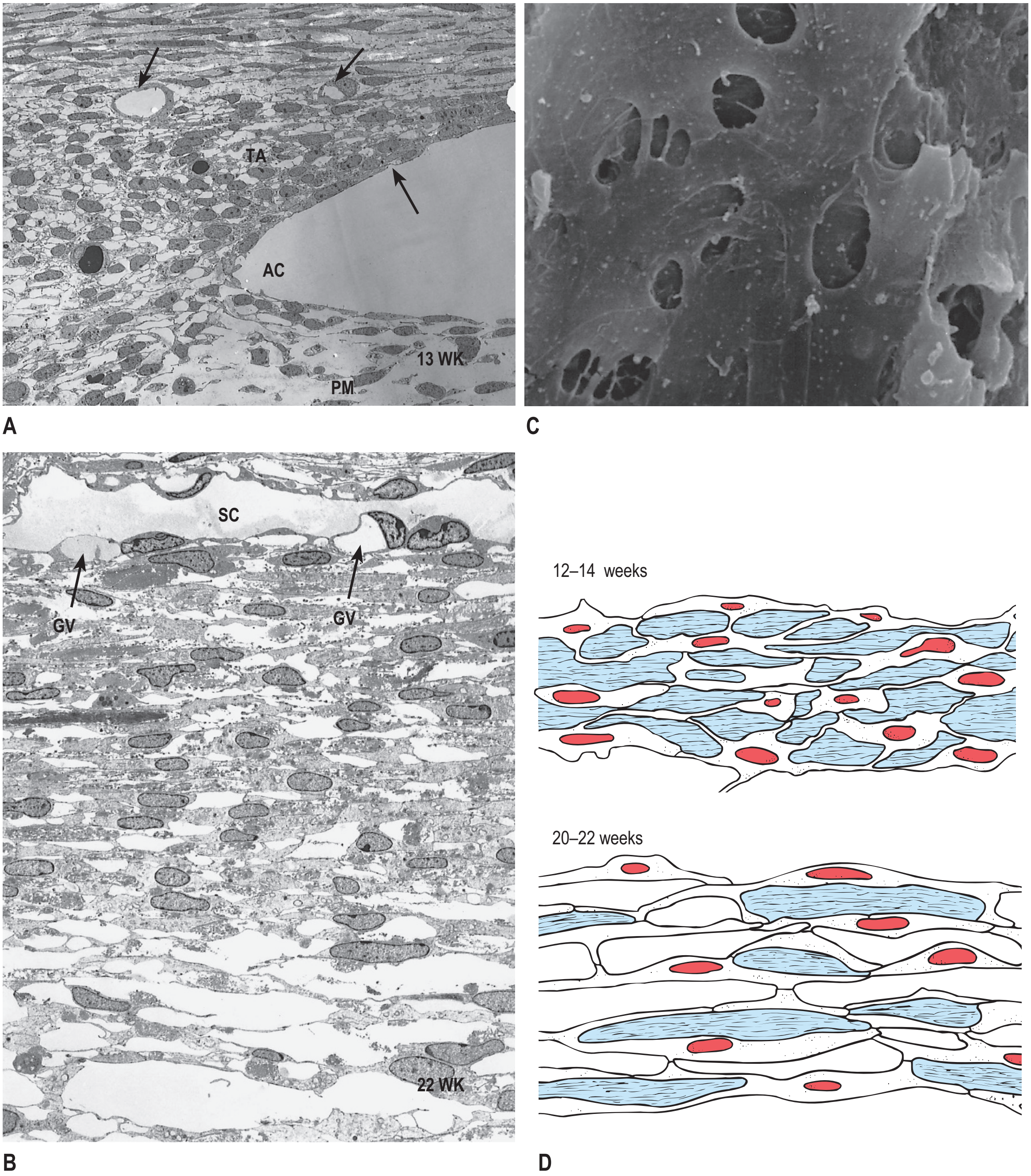

FIGURE 2-18 Development of the iridocorneal or chamber angle (see also Fig. 2-16). (**A**) Electron micrograph of the trabecular anlage (TA) in a 13-week-old human fetus. Note the two small capillaries (arrows) on the deep aspect of the anlage, the high density of trabecular cells, and the poorly developed extracellular matrix. AC, anterior chamber; PM, pupillary membrane (future iris stroma). Original magnification: × 500. (**B**) Electron micrograph of the trabecular meshwork in a 22-week-old human fetus. Note the enlarged intratrabecular spaces separated by well-formed connective tissue trabeculae and the size of Schlemm's canal (SC), which possesses giant vacuoles (GV) in its inner wall. Original magnification: × 100. (**C**) Scanning electron micrograph of the inner surface of the trabecular anlage in a 13-week-old human fetus revealing the incomplete nature of the endothelial cells facing the anterior chamber. These perforations most likely allow free passage of cells and fluids from the anterior chamber to the developing meshwork from this early stage onwards. Large arrow in (**A**) indicates the perspective from which the scanning micrograph was obtained. Original magnification: × 1600. (**D**) Summary of the morphogenetic changes that occur during remodelling of the loose mesenchyme of the trabecular anlage to form the trabecular meshwork. (Parts A–C from McMenamin, 1989, 1991; with permission.)

Development of the extraocular muscles

The extraocular muscles are some of the few periocular tissues that have been shown not to be of neural crest origin (Fig. 2-6). They are thought to arise from presumptive myocytes in the *preotic region (paraxial mesoderm)* in the area of the prochordal plate (eFig. 2-3).

They migrate ventrally and caudally around the developing eye. The presumptive myocytes concentrate particularly in the equatorial zone external to the mesenchymal condensation, which forms the sclera. Here they proliferate and differentiate. Their flattened connective tissue tendons, which are of neural crest origin, eventually fuse with the sclera.

Investigations of *MyoD* gene expression in transgenic mice (using the *LacZ* reporter gene) have shown evidence of myogenesis *in situ* around the developing eye as early as E10.5 (day 10.5 of embryogenesis), at about the time when myocytes are appearing in the hyoid arch mesenchyme.

The extraocular muscles appear in approximately the following sequence: lateral rectus, superior rectus and levator palpebrae superioris (week 5), superior oblique and medial rectus (week 6), followed by inferior oblique and inferior rectus (common primordium).

The axons of the *general somatic efferent* neurones of cranial nerves III, IV and VI, which innervate these muscles, are ‘dragged’ behind the migrating myocytes from the site of their cell bodies in the developing brainstem to the periocular region.

Development of the eyes and surrounding structures is influenced by the pattern of development of the skull, pharyngeal arches and face

A detailed account of the embryology of the head, skull and face pertinent to understanding the development of the eye and orbit can be found at https://expertconsult.inkling.com/. See also Video 1-2.

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

DEVELOPMENT OF THE SKULL

The orbit lies at the junction of the neurocranium (skull vault) and viscerocranium (facial skeleton), and has also evolved partly from the primitive sensory capsules around the eyes. Therefore, to understand the embryology and development of the orbit, it is essential to appreciate that the entire craniofacial skeleton is formed by a combination of several components.

CHONDROCRANIUM

The *chondrocranium* forms the skull base initially as cartilaginous precursors that develop in a rostral to caudal sequence, namely the prechordal plate, hypophyseal and parachordal cartilages. These subsequently ossify to form the midline bones in the base of the skull from the interorbital region (body of the sphenoid) to the occipital region. The *sensory capsules* that evolved to support the olfactory organs, eyes and inner ear develop separately alongside this midline basal cartilage. These capsules develop initially as cartilage and in humans are represented by bones in the nasal cavity (ethmoid from prechordal cartilages), orbits (body of the sphenoid – hypophyseal cartilages, lesser wing, and medial part of the greater wing of the sphenoid) and part of the temporal bone.

MEMBRANOUS BONES

The *membranous bones*, evolved from ‘dermal’ bones like those seen in fossils of primitive placoderm fishes, ossify directly from mesenchyme (‘in membrane’). They are derived from neural crest and parachordal (head) mesoderm and form the calvaria or cranial vault in the human skull. They are represented in the orbit by the orbital plate of the frontal bone.

Viscerocranium

The *viscerocranium* evolved to support the *branchial (gill) arches* in fish. In humans there are five pairs of *pharyngeal arches*, either side of the foregut tube, whose derivatives contribute to the viscerocranium (see below).

The Pharyngeal Arches

The pharyngeal arches play an important role in the morphogenesis of the head and neck; the first and second arches in particular are important to the development of the periocular region. Pharyngeal arches form in a craniocaudal sequence and are not all at similar stages of development at any point in time. In humans, five arches develop that correspond to arches 1, 2, 3, 4 and 6 of their evolutionary precursors in fish. The first or mandibular arch has two components, one forming the upper jaw the other the lower jaw; the second or hyoid arch has evolved to support the jaw, tongue and larynx (eFig. 2-3). The third arch also contributes to the hyoid. The fourth and sixth arches help form the larynx. Each arch has an inner covering of endoderm (separated by endodermal *pharyngeal pouches*) and an outer covering of ectoderm (separated by ectodermal *pharyngeal clefts*), a cartilaginous component (e.g. Meckel’s cartilage), an arch nerve and an arch artery, together with a core of mesenchyme (eFig. 2-4). This mesenchyme is a mixture of somatic mesoderm and neural crest-derived mesectoderm (derived from rhombomere 2 (first arch), rhombomere 4 (second arch) and rhombomere 6 (third arch); (Hunt et al., 1991; Noden, 1991; Maden et al., 1992; Hall, 2005)). eTable 2-1 and eFigure 2-3 summarize the skeletal, neuronal and muscular derivatives from the pharyngeal arches. The migration of neural crest cells into each of the pharyngeal arches is like all axial organization (i.e. neural tube, somitic mesoderm and endoderm) controlled by the *Hox* code. Only a few elements of the viscerocranium, such as the medial part of the greater wing of the sphenoid (alisphenoid), incus, malleus and stapes, arise directly from the cartilaginous component (eFig. 2-3A); the majority of the cartilaginous elements of the pharyngeal arches regress and become encased within membrane bones that ossify directly from neural crest-derived mesenchyme. Such membranous bones include the maxilla, zygoma, squamous temporal bone and dentary (described in most textbooks as the mandible) (eFig. 2-3A).

Some of the cranial nerve sensory neurones (part of V, VII, IX and X), like the dorsal root ganglia of the trunk, arise from the neural crest, as do the four parasympathetic ganglia of the head and neck.

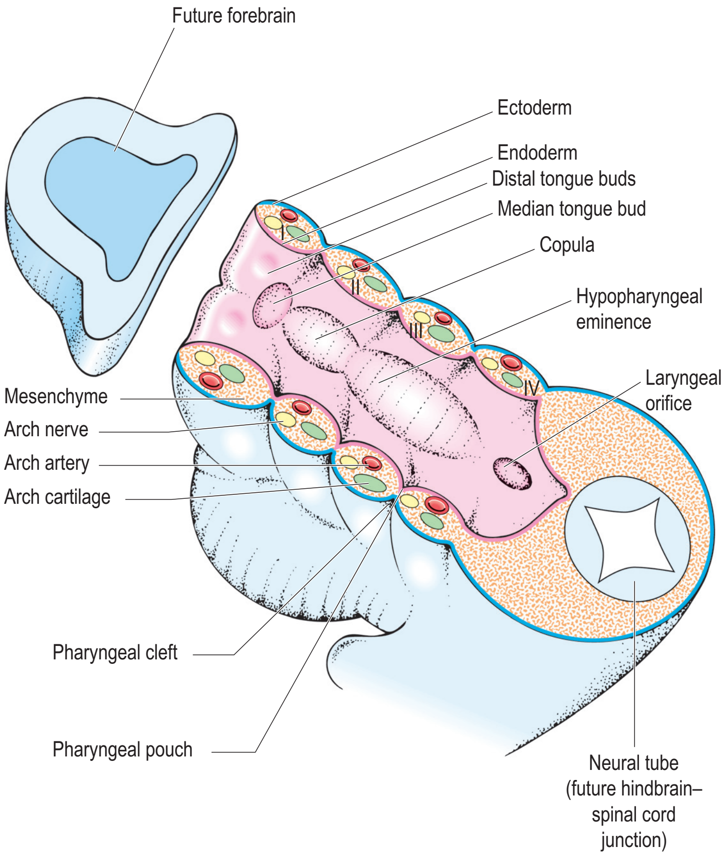
Future forebrain
Ectoderm
Endoderm
Distal tongue buds
Median tongue bud
Copula
Hypopharyngeal eminence
Laryngeal orifice
I
II
III
IV
Mesenchyme
Arch nerve
Arch artery
Arch cartilage
Pharyngeal cleft
Pharyngeal pouch
Neural tube (future hindbrain–spinal cord junction)

eFIGURE 2-4 Schematic diagram of the floor of the embryonic mouth as viewed from above in a horizontal section of the head. Note the pharyngeal arches (I–IV) have been cut to expose their neural (yellow), arterial (red), cartilaginous (green) and mesenchymal (orange stipple) contents. The arches are lined externally by ectoderm (blue) and internally by endoderm (purple).

eTABLE 2-1 Pharyngeal arch derivatives (arch arteries and their derivatives not included)

	Pouch endoderm	Skeletal elements		Cranial nerve	Muscles
Pharyngeal arch		**Ossify in cartilage**	**Ossify in membrane**		
1	Tympanic cavity Pharyngotympanic tube (*cleft ectoderm – external auditory meatus*)	Arch cartilage – palatoquadrate bar: alisphenoid, incus Meckel's cartilage – malleus	Arch mesenchyme: *upper part* – maxilla, zygoma, squamous temporal bone *lower part* – mandible	Maxillary division of trigeminal (V) (upper part of arch 1) Mandibular division of trigeminal (V) (lower part of arch 1)	*From cranial somitomere 4*: Muscles of mastication (temporalis, masseter, pterygoids) *plus* mylohyoid, anterior belly of digastric, tensor tympani, tensor veli, palatini
2	Epithelial lining of tonsillar crypts and tonsillar fossa	Reichert's cartilage: stapes, stylohyoid ligament, upper part of hyoid		Facial nerve (VII)	*From somitomere 6*: Muscles of facial expression *plus* posterior belly of digastric, stylohyoid, stapedius
3	Dorsal wing – inferior parathyroid Ventral wing – epithelioid cells of thymus (Hassall's corpuscles and epithelial reticulum)	Lower part of hyoid		Glossopharyngeal nerve (IX)	*From somitomere 7*: Stylopharyngeus
4	Dorsal wing – superior parathyroid Ventral wing – ultimobranchial body: C cells in thyroid gland		Upper laryngeal cartilages (mesoderm)	Superior laryngeal branch of vagus nerve (X)	*From occipital somites 2–4*: Pharyngeal constrictors, cricothyroid, levator veli palatini
6			Lower laryngeal cartilages (mesoderm)	Recurrent laryngeal nerve (X)	Intrinsic muscles of the larynx

The accessory nerve has all the hallmarks of a branchial arch nerve as its cell bodies in the brainstem and cervical spinal cord are in line with other branchial efferent cell bodies (branchial efferent column); however, its evolutionary history, or more specifically the muscles it supplies (sternocleidomastoid and trapezius), is controversial, as are many of the musculoskeletal elements of the neck and shoulder (Matsuoka et al., 2005).

DEVELOPMENT OF THE FACE

(eFig. 2-5)

Development of the face commences around the fourth week and is largely complete by week 10. The basic arrangement of the face is the result of fusion of five swellings around the stomodeum (primitive mouth), nasal pits and eyes. There are paired maxillary and mandibular processes plus an unpaired fronto-nasal process (eFig. 2-5A). The mesenchyme of the fronto-nasal process, a conspicuous swelling over the developing forebrain vesicles, arises from the neural crest and does not appear to have any association with pharyngeal arch development. The fronto-nasal process is innervated by the ophthalmic division of the trigeminal nerve and forms the tissues above the eye as far back as the vertex, thus explaining the sensory cutaneous distribution in the adult. The fronto-nasal process also gives rise to medial and lateral nasal swellings around the nasal pits (eFig. 2-5B). The maxillary processes grow medially beneath the developing eyes and form the lower eyelids. Concomitant with this growth, the eyes also move from their position on the lateral aspect of the embryonic head to the front of the face. The *nasolacrimal groove* on each side lies along the line of fusion of the maxillary process with the lateral nasal swelling (eFig. 2-5C). The ectoderm of the groove invaginates (10 mm stage) into the surrounding mesenchyme to form the *nasolacrimal duct*, which at this stage consists of a solid cord of ectodermal cells that grows upwards into the lids to form the canaliculi and downwards into the nose. Canaliculization of the solid columns of cells commences first in the vicinity of the lacrimal sac (3 months). Mesenchyme condenses and ossifies around this cord of cells to form the bony walls of the nasolacrimal canal.

The upper eyelid develops from the fronto-nasal process and both lids are visible early in the second month (Fig. 2-1E). The primitive lid folds fuse in weeks 9–10 of gestation, enclosing a surface ectoderm-lined cavity, the conjunctival sac (Fig. 2-2D). Myocytes differentiate in the mesenchyme of both lids and eventually form the orbicularis oculi muscle (second arch derived – nerve supply, facial nerve). Meibomian glands, sebaceous glands and eyelashes develop as invaginations of the conjunctival epithelium or epidermis and are therefore ectodermal derivatives. In the seventh week the future *lacrimal gland* arises as a bud of epithelial cells from the region of the upper temporal conjunctiva sac.

At birth and in the infant, the eyes in humans are in a comparatively advanced state of development relative to the rest of the face; the orbits are therefore large, although the remainder of the facial skeleton is small by comparison with the adult. During infancy and childhood, as a consequence of the development of teeth and growth of the paranasal sinuses, the facial skeleton steadily increases in size relative to the neurocranium. With the exception of the frontal sinuses, the paranasal sinuses develop as diverticulae from the nasal passages around the fifth month of gestation and are rudimentary at birth. The frontal sinuses appear in the fifth to sixth postnatal year as evaginations from the ethmoid sinuses or middle meatus.

CONGENITAL MALFORMATIONS

Classification of congenital malformations or abnormalities is difficult for several reasons. First, the aetiology is often unknown and, even where a single genetic or environmental cause is suspected, it is often difficult to ascribe full responsibility to such agents or events because exposure to widely different teratogenic agents, such as drugs or trauma, can result in identical developmental defects. Second, defects may be strongly associated with chromosomal abnormality (e.g. trisomy 13).

Furthermore, disturbances in basic cellular events such as neural tube closure may lead to multisystem pathologies. In common with more generalized developmental defects, the consequences of exposure to teratogenic agents are highly dependent on the timing of major embryonic or fetal developmental events occurring at the time of exposure: there are important known *periods of vulnerability*, or critical (sensitive) periods, during which morphogenesis of particular systems and organs may be at risk. In this respect the eye is highly vulnerable for a long period of embryonic and fetal development because there are crucial developmental landmarks from the point of formation of the optic pits (day 22) until late in gestation, with such events as retinal vascularization and pupillary membrane regression occurring close to birth.

Congenital malformations affecting ocular tissues may have the following causes:

Genetic Causes

- *Chromosomal anomalies:* deletions (*cri-du-chat* syndrome, Turner syndrome), trisomies (trisomy 13 or Patau syndrome, Down syndrome, Klinefelter syndrome) and triploidy (see Table 2-1)
- *Hereditary:* either sporadic mutations (aniridia with Wilms' tumour) or dominant/recessive inheritance.

Environmental Causes

- *Drugs:* alcohol, tobacco, anticonvulsants, thalidomide
- *Vitamins and minerals:* excess (e.g. hypervitaminosis A) or deficiency (e.g. folic acid, zinc)
- *Infection:* e.g. rubella, syphilis, toxoplasmosis
- *Radiation:* X-rays.

Maternal Age

There is an increased incidence of genetic abnormalities in the oocyte with advancing maternal age.

eBox 2-3

Craniofacial abnormalities

Skeletal development can be considered to have three essential steps: (1) classic induction (as the result of mesenchyme–epithelium interactions mediated by highly conserved signalling molecules such as Bmp (bone morphogenetic protein) and FGF (fibroblast growth factor)); (2) condensation of mesenchyme; and (3) overt differentiation. Interestingly, 65% of skeletal abnormalities of the head and neck are the result of defects in this first signalling step (Hall, 1998).

It is now recognized that a group of craniofacial abnormalities, the mandibulofacial dysostoses, including Treacher Collins and Hallerman–Streiff syndromes, are caused by deficits in neural crest cell migration and differentiation in the first and second pharyngeal arches. These are manifest as abnormal ear development, hypoplasia of the maxilla and mandible, and lower lid defects. These and more generalized disturbances of neural crest migration, e.g. Rieger syndrome, Pierre Robin syndrome, and conditions affecting primarily the periocular region such as Peters' anomaly, are increasingly being classified as *neurocristopathies* because of their proposed link to disturbances in neural crest cell migration, proliferation and differentiation.

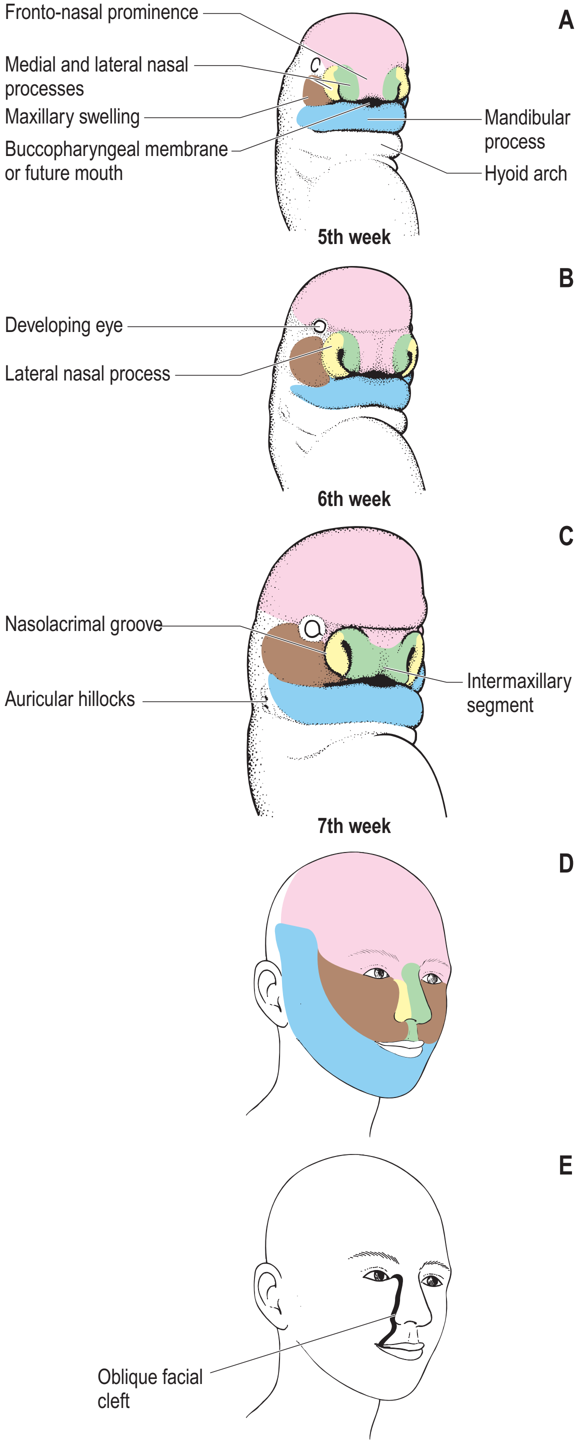
A
Fronto-nasal prominence
Medial and lateral nasal processes
Maxillary swelling
Buccopharyngeal membrane or future mouth
Mandibular process
Hyoid arch
5th week
B
Developing eye
Lateral nasal process
6th week
C
Nasolacrimal groove
Auricular hillocks
Intermaxillary segment
7th week
D
E
Oblique facial cleft

eFIGURE 2-5 Early development of the face (**A–C**), the adult form (**D**), and one of the many congenital abnormalities, oblique facial cleft (**E**), that results from incomplete fusion of the facial swellings.

eBox 2-4

Developmental anomalies of the nasolacrimal duct

Disturbances in morphogenetic processes may lead to multiple canaliculi and punctae, abnormal diverticulae, and blockage of the nasolacrimal duct, possibly because of debris from the degenerating central cells producing a mucocoele (not uncommon in the first few weeks after birth).

There is a range of *congenital facial defects*, of which cleft lip and palate are the most common. These are a consequence of complete or partial failure of fusion of the various processes and swellings in the face. Oblique facial clefts occur along the course of the nasolacrimal duct (eFig. 2-5E). The aetiology of these conditions is multifactorial, although maternal exposure to anticonvulsant drugs (phenytoin) and vitamin A are known teratogenic agents that produce these defects. The spatial and temporal expression of retinoic acid receptors during embryological development is currently the subject of much research. While details are still emerging, it is clear that the patterns of expression partially explain the effects of retinoic acid (a vitamin A derivative) as a potent morphogen. Retinoic acid appears sequentially to activate genes of the *Hox* cluster, thereby influencing the normal segmental pattern of hindbrain development (rhombomeres), neural crest migration pathways into pharyngeal arches, and development of the face (Maden et al., 1992).

eBox 2-5

Trisomy 13 (Patau syndrome)

The ocular pathology in trisomy 13 illustrates various forms of malformation. The cornea and chamber angle are malformed and persistent hyperplastic primary vitreous is common. An anterior coloboma is present and is characterized by a fibrous ingrowth that contains nodules of cartilage. Retinal dysplasia is extensive. Optic nerve malformation is limited to hypoplasia. The systemic malformations are not compatible with survival and are extreme forms of brain malformation (arrhinencephaly) with cardiac and renal malformation.

eBox 2-6

Trisomy 21 (Down syndrome)

The systemic disturbances in this disorder are well known. In ophthalmology the important components are a high incidence of keratoconus and cataract. Small nodules are formed by spindle cells on the iris (Brushfield's spots); myopia and the attendant complication of retinal detachment may require surgical intervention.

FURTHER READING

Bartelmez, G.W., Blount, M.P., Gage, P.J., Rhoades, W., Prucka, S.K., Hjalt, T., 1954. The formation of neural crest from the primary optic vesicle in man. Contr. Embryol. 35, 55–71.

Blankenship, T., Peterson, P.E., Hendricks, A.G., 1996. Emigration of neural crest cells from macaque optic vesicles is correlated with discontinuities in its basement membrane. J. Anat. 188, 473–483.

Bohnsack, B.L., Gallina, D., Thompson, H., Kasprick, D.S., Lucarelli, M.J., Dootz, G., et al., 2011. Development of extraocular muscles requires early signals from periocular neural crest and the developing eye. Arch. Ophthalmol. 129, 1030–1041.

Chan-Ling, T., 1997. Glial, vascular, and neuronal cytogenesis in whole-mounted cat retina. Microsc. Res. Tech. 36, 1–16.

Chauhan, B.K., Disanza, A., Choi, S.Y., Faber, S.C., Lou, M., Beggs, H.E., et al., 2009. Cdc42- and IRSp53-dependent contractile filopodia tether presumptive lens and retina to coordinate epithelial invagination. Development 136, 3657–3667.

Collinson, J.M., Hill, R.E., West, J., 2004. Analysis of mouse eye development with chimeras and mosaics. Int. J. Dev. Biol. 48, 793–804.

Creuzet, S., Vincent, C., Couly, G., 2005. Neural crest derivatives in ocular and periocular structures. Int. J. Dev. Biol. 49, 161–171.

Cvekl, A., Tamm, E.R., 2004. Anterior eye development and ocular mesenchyme: new insights from mouse models and human diseases. Bioessays 26, 374–386.

Erickson, C.A., Loring, J.F., Lester, S.M., 1989. Migratory pathways of HNK-1-immunoreactive neural crest cells in the rat embryo. Dev. Biol. 134, 112–118.

Fukiishi, Y., Morriss-Kay, G.M., 1992. Migration of neural crest cells to the pharyngeal arches and heart in rat embryos. Cell Tissue Res. 268, 1–8.

Gage, P.J., Rhoades, W., Prucka, S.K., Hjalt, T., 2005. Fate maps of neural crest and mesoderm in the mammalian eye. Invest. Ophthalmol. Vis. Sci. 46, 4200–4208.

Gilland, E., Baker, R., 1993. Conservation of neuroepithelial and mesodermal segments in the embryonic vertebrate head. Acta Anat. 148, 110–123.

Hall, B.K., 1998. Evolutionary developmental biology. Kluwer Academic Publishers, Dordrecht.

Hall, B.K., 2005. Bones and cartilage: developmental and evolutionary skeletal biology. Elsevier Academic Press, London.

Heavner, W., Pevny, L., 2012. Eye development and retinogenesis. Cold Spring Harb. Perspect. Biol. 4, pii: a008391.

Hunt, P., Wilkinson, D., Krumlauf, R., 1991. Patterning the vertebrate head: murine *Hox2* genes mark distinct subpopulations of premigratory and migratory cranial neural crest. Development 112, 43–50.

Lang, R.A., 2004. Pathways regulating lens induction in the mouse. Int. J. Dev. Biol. 48, 783–791.

Lang, R.A., Bishop, J.M., 1993. Macrophages are required for cell death and tissue remodelling in the developing mouse eye. Cell 74, 453–462.

Larsen, W.J., 1997. Human embryology, second ed. Churchill Livingstone, New York.

Maden, M., Horton, C., Graham, A., Leonard, L., Pizzey, J., Siegenthaler, G., et al., 1992. Domains of cellular retinoic acid-binding protein I (CRABP I) expression in the hindbrain and neural crest of the mouse embryo. Mech. Dev. 37, 13–23.

Mann, I., 1928. The development of the human eye. Grune and Stratton, New York.

Mann, I., 1964. The development of the human eye, third ed. Grune and Stratton, New York.

Matsuoka, T., Ahlberg, P.E., Kessaris, N., Iannarelli, P., Dennehy, U., Richardson, W.D., et al., 2005. Neural crest origins of the neck and shoulder. Nature 436, 347–355.

McAvoy, J.W., Chamberlain, C.G., de Jongh, R.U., 1999. Lens development. Eye 13, 425–437.

McMenamin, P.G., 1989. A morphological study of the inner surface of the anterior chamber angle in pre- and post-natal human eyes. Curr. Eye Res. 8, 727–739.

McMenamin, P.G., 1989. The human foetal iridocorneal angle: a scanning electron microscopic study. Br. J. Ophthalmol. 73, 871–879.

McMenamin, P.G., 1991. A quantitative study of the prenatal development of aqueous outflow system in the human eye. Exp. Eye Res. 53, 507–517.

McMenamin, P.G., Djano, J., Wealthall, R., Griffin, B.J., 2002. Characterisation of the macrophages associated with the tunica vasculosa lentis of the rat eye. Invest. Ophthalmol. Vis. Sci. 43, 2076–2082.

McMenamin, P.G., Krause, W., 1993. Development of the eye in the North American opossum (*Didelphis virginiana*). J. Anat. 183, 343–358.

Mendelson, C., Larkin, S., Mark, M., LeMeur, M., Clifford, J., Zelent, A., et al., 1994. RARβ isoforms: distinct transcriptional control by retinoic acid and specific spatial patterns of promotor activity during mouse embryonic development. Mech. Dev. 45, 227–241.

Monaghan, A.P., Davidson, D.R., Sime, C., 1991. The Msh-like homeobox genes define domains in the developing vertebrate eye. Development 112, 1053–1061.

Noden, D., 1982. Periocular mesenchyme: neural crest and mesoderm interactions. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 97–119.

Noden, D.M., 1988. Interactions and fate of avian craniofacial mesenchyme. Development 103 (Suppl.), 121–140.

Noden, D.M., 1991. Vertebrate craniofacial development: the relation between ontogenetic process and morphological outcome. Brain Behav. Evol. 38, 190–225.

O'Rahilly, R., 1966. The prenatal development of the human eye. Exp. Eye Res. 21, 93–112.

Osumi-Yamashita, N.O., Noji, S., Nohno, T., Koyama, E., Doi, H., Eto, K., et al., 1990. Expression of retinoic acid receptor genes in neural crest-derived cells during mouse facial development. FEBS Lett. 264, 71–74.

Ozanics, V., Jakobiec, F.A., 1982. Prenatal development of the eye and its adnexa. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 11–96.

Patapoutian, A., Miner, J., Lyons, G.E., Wold, B., 1993. Isolated sequences from the linked *Myf-5* and *MRF4* genes drive distinct patterns of muscle-specific expression in transgenic mice. Development 118, 61–69.

Reese, B.E., Necessary, B.D., Tam, P.P.L., Faulker-Jones, B., Tan, S.S., 1999. Clonal expansion and cell dispersion in the developing mouse retina. Eur. J. Neurosci. 11, 2965–2978.

Robinson, S.R., 1991. Development of the mammalian retina. In: Dreher, B., Robinson, S.R. (Eds.), Neuroanatomy of the visual pathways and their development; Cronly-Dillon JR, series ed. Vision and visual dysfunction, vol. 3. Macmillan, Basingstoke, pp. 69–128.

Saha, M.S., Spann, C., Grainger, R.M., 1989. Embryonic lens induction: more than meets the optic vesicle. Cell Differ. Dev. 28, 153–171.

Saint-Geniez, M., D'Amore, P., 2004. Development and pathology of the hyaloid, choroidal and retinal vasculature. Int. J. Dev. Biol. 48, 1045–1058.

Schulz, M.W., Chamberlain, C.G., de Longh, R.U., McAvoy, J.W., 1993. Acidic and basic FGF in ocular media and lens: implications for lens polarity and growth patterns. Development 118, 117–126.

Spaeth, G., Nelson, L.B., Beaudoin, A.R., 1982. Ocular teratology. In: Jakobiec, F.A. (Ed.), Ocular anatomy, embryology and teratology. Harper and Row, Philadelphia, pp. 955–1080.

Sullivan, C.H., Braunstein, L., Hazard-Leonards, R.M., Holen, A.L., Samaha, F., Stephens, L., et al., 2004. A re-examination of lens induction in chicken embryos: in vitro studies of early tissue interactions. Int. J. Dev. Biol. 48, 771–782.

Trainor, P.A., Tam, P.P.L., 1995. Cranial paraxial mesoderm and neural crest cells of the mouse embryo: co-distribution in the craniofacial mesenchyme but distinct segregation in the branchial arches. Development 121, 2569–2582.

Zieske, J.D., 2004. Corneal development associated with eyelid opening. Int. J. Dev. Biol. 48, 903–911.

Chapter **3**

Genetics

- Chromosomes and cell division
- Molecular genetics (DNA and genes)
- Chromosome defects and gene mutations
- Clinical genetics
- Population genetics
- Understanding the human genome: DNA analysis
- Molecular biology and clinical medicine
- Molecular and cell biology: controlling cell destiny
- Molecular genetics and ophthalmology

Genetic disorders are often thought of as interesting but rare. However, in a population of 100 000, there will be 10 children born with a genetic disorder, a further 10 who will develop genetic disorders in later life, and more than 10% of adults will have a chronic illness with a strong genetic basis. The advances in medical and molecular genetics have increased our understanding of disease processes, which can be illustrated particularly with examples of inherited ophthalmic disorders. Moreover, the increasing ability to individually, or within disease populations, define the genome, polymorphisms in gene expression, control of translation and the impact of environment on genetic expression or heritable as well non-heritable non-DNA influence on gene expression (epigenetics) has profoundly influenced our interrogation of disease pathogenesis and targeted therapeutic intervention.

Chromosomes and cell division

CHROMOSOMES

Human cells contain 46 chromosomes (diploid), consisting of 22 almost identical pairs (homologous autosomes) and a pair of sex chromosomes, which constitute the karyotype of an individual. Chromosomes are numbered 1–22, as they decrease in size. The chromosome consists of two arms: the short arm is designated ‘p’ and the long arm ‘q’. During somatic division, or mitosis, each chromosome, and therefore gene, replicates precisely. Gamete formation involves a special reduction division (meiosis), in which the homologous chromosomes separate from each other so that each cell contains 23 chromosomes (haploid). Since chance determines which chromosome of a pair ends up in a particular gamete, there is a possibility of 2^{23} combinations in the gametes. The different stages of the two forms of cell division are summarized below. The genes are located in a linear order along the chromosome, each gene having a precise location or *locus*. Each of a pair of chromosomes (homologous chromosomes) carries matching genetic information, i.e. they have the same sequence of gene loci. However, at any locus they may have slightly different forms, which are called *alleles*. Cytogenetic analysis of the chromosomes can be performed by looking at the banding pattern of individual chromosomes. Giemsa staining highlights the horizontal banding of the chromosome, by the characteristic pattern of light and dark staining (G-banding). More specific patterns of staining can be performed by looking at *high-resolution banding* with G-banding patterns at the early stages of mitosis (prophase; see below). *Fragile sites* are non-staining gaps that can be demonstrated only when the chromosomes are cultured in specific conditions (e.g. thymidine and folic acid deprivation) and then stained.

CELL DIVISION

Mitosis (Box 3-1)

During mitosis each chromosome, and with it every gene along its length, replicates to produce two

BOX 3-1 STAGES OF MITOSIS

Interphase	Resting phase
Prophase	Chromosomes identifiable in nucleus
Metaphase	Chromosomes aligned in centre of nucleus
Anaphase	Centromere of chromosome divides with chromatid separation
Telophase	Daughter chromosomes separate

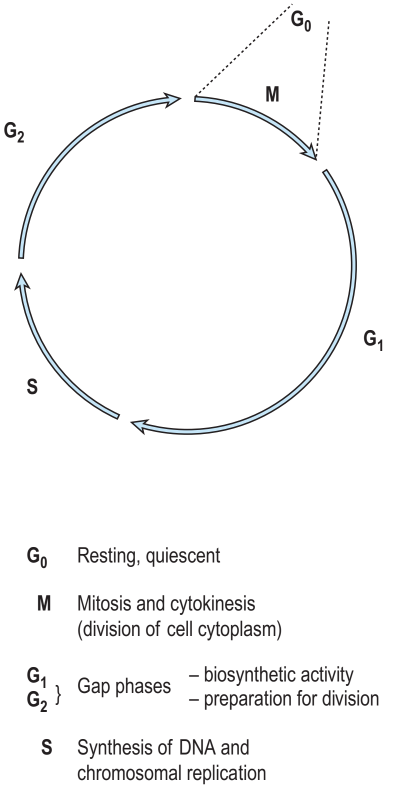
G$_0$
M
G$_2$
G$_1$
S
G$_0$ Resting, quiescent
M Mitosis and cytokinesis (division of cell cytoplasm)
G$_1$ G$_2$ } Gap phases – biosynthetic activity – preparation for division
S Synthesis of DNA and chromosomal replication

FIGURE 3-1 The cell cycle: S, synthesis of DNA; M, mitosis and cytokinesis; G_1 and G_2, gap phases.

identical daughter cells, so that every nucleated cell (except gametes) has 46 chromosomes (diploid). In culture, the cell cycle of most mammalian cells varies but is usually about 24 hours. Mitosis itself occupies approximately 1 hour of the total time, whereas the time taken to synthesize DNA required for replication is 6–8 hours (Fig. 3-1). The gap phases G_1, between mitosis and synthesis of DNA, and to a lesser extent G_2, between synthesis and mitosis, govern the turnover of the cells because some cells can rest in G_1 for days. It appears that late in G_1 the cell passes a *restriction point* after which it will proceed through the rest of the cell cycle at a standard rate. During the S phase each chromosome replicates, each with its own individual pattern, although homologous pairs of chromosomes replicate synchronously. During G_2, the cell gradually enlarges, doubling its total mass before the next mitosis. Some cells, however, do not divide once they are fully differentiated (e.g. neurones and erythrocytes) and are permanently arrested in a phase known as G_0. Control of cell cycle is represented as a schematic in Box 3-2.

Meiosis: formation of gametes

Meiosis takes place in two stages, both having the same stages as mitosis. In the first stage of meiosis there is a prolonged prophase, in which cell division of only the chromosomes occurs so that each subsequent daughter nucleus contains half the number of chromosomes (23), resembling mitosis but in the haploid state. The prophase of the first meiotic division is complex and may be divided into a further five stages.

- *Leptotene*. Starts with the first appearance of the chromosome.
- *Zygotene*. The homologous chromosomes pair and bind closely to form bivalents.
- *Pachytene*. The main stage of chromosomal thickening; each chromosome consists of two chromatids and, because they are bivalent, has four strands. During this stage, chromatids of the chromosomes exchange material (crossovers), further ensuring a random assortment of paternal and maternal homologues.
- *Diplotene*. The bivalents start to separate (the centromere of each remains intact).
- *Diakinesis*. The bivalent chromosome formations separate and coil tightly.

The chromosomes then undergo *metaphase* and *anaphase*, where the bivalents disjoin, going to each equatorial plane of the cell, and as the cytoplasm divides each cell has 23 chromosomes, each with a pair of chromatids. The second meiotic division follows the first without an interphase and resembles mitosis.

BOX 3-2 CONTROL OF CELL CYCLE

Cyclins are regulatory subunits within the cell that are periodically synthesized and degraded, regulating cyclin-dependent kinases (CDK) essential for cell cycle control.

Cyclins (of which to date there are nine; A–I) bind to CDKs, forming an active complex that in a controlled fashion drives the cell through its stages via phosphorylation of unique protein substrates, such as retinoblastoma protein, Rb, allowing progression into the next phase of the cycle. Interaction of cyclin–CDK complexes with the retinoblastoma (Rb) protein family (Rb, p107, p130) is regulated by proteins that block cyclin–CDK activity (e.g. p21). p53, a cell cycle regulator, leads to increased p21 expression.

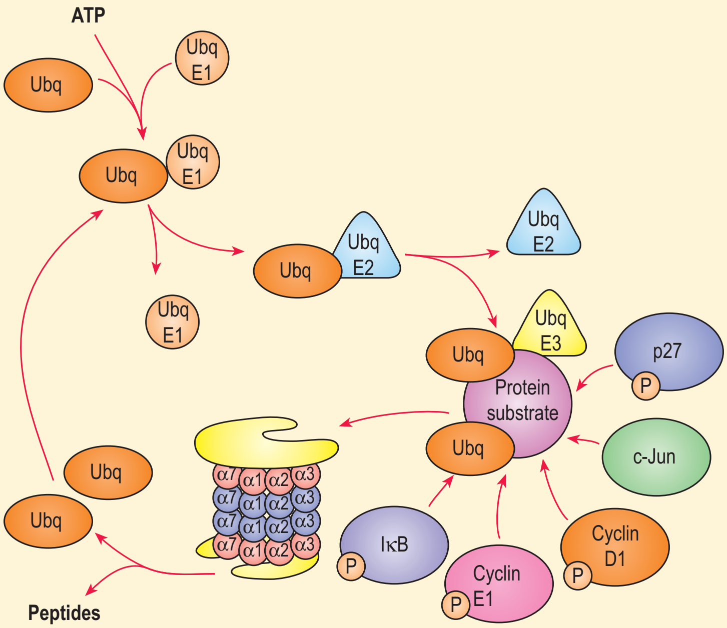

Thus meiosis gives rise to gametes that contain only one representative of each homologous pair of chromosomes and there is random selection of paternal and maternal homologues.

Genetic makeup

The genetic makeup of each individual or of a specific patient population is a result of mutations and natural selection in the past. During cell division, the division of chromosomes may be imperfect. A mutation is a sudden change in the gross or fine structure of DNA, such as might be caused by a replication error or a crossover defect between misaligned chromosomes at meiosis, giving rise to deletion or duplication of DNA. The mechanisms by which chromosomal abnormalities may occur are discussed below. A mutation may affect a specific gene and thus the protein structure it codes for. Equally, the mutation may alter the structure of the protein in such a way that the function of the protein is changed only slightly but the metabolism remains unaffected. If the gene is destroyed altogether, including the nucleotide sequence, which is critically important for the protein that it encodes, then the individual will be at an immediate disadvantage. It is therefore not surprising that it is genetic diseases that occur as a result of harmful mutations, as opposed to diseases with multifactorial inheritance patterns, which have a predictable mode of inheritance.

Molecular genetics (DNA and genes)

The basis of inherited disease can be understood from information about chromosomal makeup, specific genes, and how proteins are encoded. The human haploid genome consists of 3×10^9 base pairs of double-stranded DNA (dsDNA). The genetic

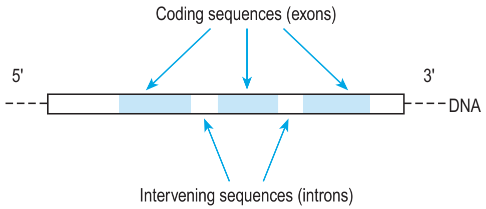

FIGURE 3-2 Components of a gene.

information that determines the sequence of amino acids in peptide chains is stored in the DNA by the order of the nucleotide bases: adenine (A), cytosine (C), guanine (G) and thymine (T). Three bases (a codon) code for a particular amino acid. A gene is part of the DNA that directs the sequencing of polypeptide chains. The gene itself consists of both coding regions (*exons*) and variable-length intervening regions (*introns*). There are also enhancer regions of the gene, which carry out regulatory activities, particularly the expression of genes in tissues. Similarly there is a family of regulatory molecules on the gene that activate or suppress both pairs of homologous genes (Fig. 3-2).

HOW GENES WORK: DECODING DNA

All proteins are encoded in DNA and, as mentioned above, each amino acid is represented by a triplet code of bases (codon). Since the number of possible codons is greater than the number of amino acids, most amino acids, except methionine and tryptophan, may be encoded by more than one codon. The code AUG (methionine) acts as a starter signal for protein synthesis (Box 3-3), and there are also codes that act as polypeptide chain terminators (stop codons).

Polypeptide chain synthesis occurs in two processes: *transcription* and *translation* (Fig. 3-3). During transcription one strand of DNA forms a template for the synthesis of messenger RNA (mRNA), catalysed by the enzyme RNA polymerase. RNA polymerase binds to a specific DNA sequence known as the *promoter*, which signals the unwinding of the DNA helix, which acts as a template for the complementary base pairs. The process moves along the DNA molecule, extending the RNA molecule in the 5′ to 3′ direction. This process continues until it meets the termination signal on the DNA molecule. The first transcript is an exact replica of the single strand of DNA. Before leaving the nucleus the introns are excised and the exons are spliced together. On leaving the nucleus the mRNA acts as a template on cytoplasmic ribosomes for protein synthesis. The mRNA 5′ end is blocked before the whole molecule is transcribed with 7-methylguanine joined 5′ to 5′ and a 3′ poly-A tail, which aids transport into the cytoplasm. The poly-A tail is added after transcription by *poly-A polymerase*. Translation of nucleotide sequences is performed by specific transfer RNA (tRNA): amino acids, each specific for three nucleotide bases, are attached before polymerization into polypeptide chains. They are attached by their C-terminal ends and are transformed into high-energy molecules, from which peptide bonds form to produce polypeptides. Each tRNA (anticodons) is linked with the appropriate codons on the mRNA. This process of translation is controlled by a set of enzymes called *aminoacyl-tRNA synthetases*, which couple each amino acid to the appropriate tRNA molecule. The mRNA sequence is read three nucleotides at a time from the 5′ to 3′ direction along the mRNA. The polypeptide chain is constructed from its N-terminal end to its C-terminal end, and the growing C-terminal end of the protein remains activated by covalent attachment to the tRNA (peptidyl-tRNA molecule). The mRNA is coded with specific start and stop codons, which thus determine the length of the polypeptide chain. The initiation of polypeptide chain synthesis is under the control of initiation factor 2 (IF-2), which catalyses the reaction between a specific initiator tRNA molecule and the mRNA. The growing polypeptide chain must be

BOX 3-3 SUMMARY OF PROTEIN SYNTHESIS

Messenger RNA is produced from DNA transcription under control of the enzyme 5′-3′ RNA polymerase. Translation begins in the cell cytoplasm by the binding of mRNA to ribosomal subunits, to which an initiator tRNA molecule binds. Amino acid-specific tRNAs have anticodons specific for a triplet of nucleotides (codon), forming the aminoacyl-tRNA molecule, which binds to the A site on the ribosome under the control of an enzyme, aminoacyl-tRNA synthetase. Each amino acid is added to the C-terminal of the growing polypeptide chain, progressing from codon to codon in the 5′ to 3′ direction of the mRNA. The length of the polypeptide chain is regulated by specific start and stop codons.

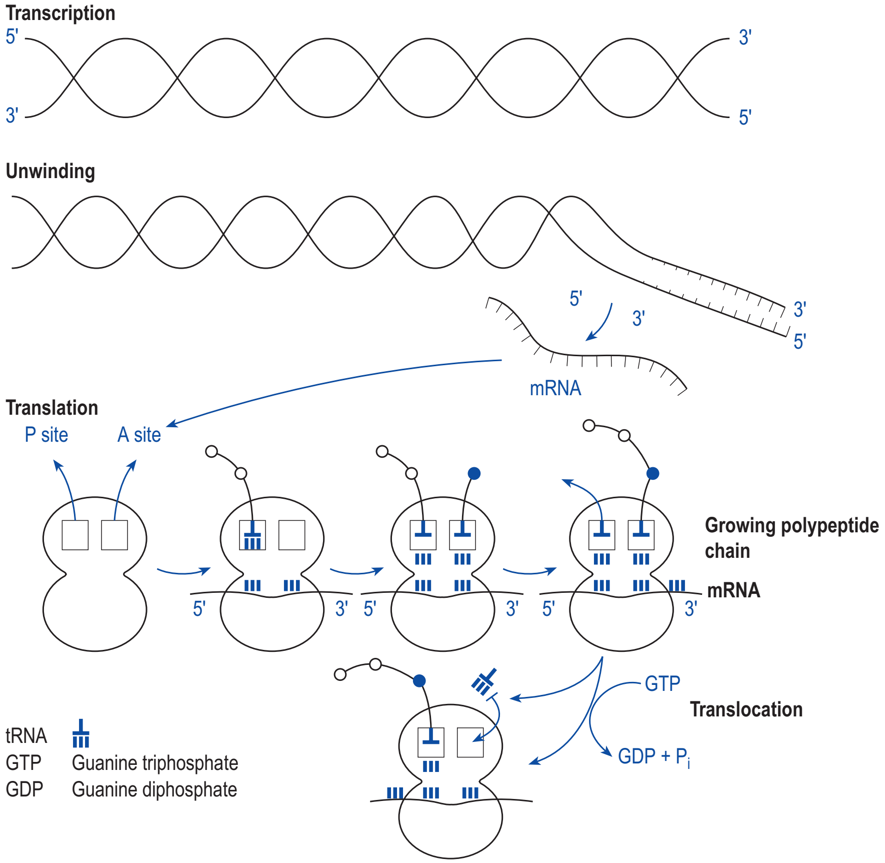

FIGURE 3-3 Protein synthesis: transcription and translation.

exactly three nucleotides along the mRNA after the addition of each amino acid, which requires a multi-enzyme system incorporated within the ribosome. Ribosomes are half RNA by weight and have a groove that accommodates a growing polypeptide chain and a groove that accommodates each mRNA molecule. On the ribosome there are two different binding sites for tRNA. One site holds the tRNA and growing polypeptide chain, the peptidyl-tRNA-binding site (P site), and the other site holds the incoming tRNA molecule, the aminoacyl-tRNA-binding site (A site). The process of polypeptide chain elongation involves two steps. The first step involves an aminoacyl-tRNA molecule becoming bound to the A site adjacent to an occupied P site. The second step involves uncoupling the C-terminal end of the polypeptide chain from the tRNA in the P site and joining it to the tRNA molecule in the A site. The reaction is catalysed by *peptidyl transferase*, which is present on the ribosome. The new peptidyl-tRNA molecule is then transferred to the P site, releasing the now unoccupied tRNA molecule, and the whole process is repeated. The stop codon on the mRNA initiates a protein called *release factor* to bind to the A site, causing hydrolysation of the peptidyl-tRNA molecule on the P site and subsequent release of the polypeptide chain into the cell cytoplasm.

Genes are regulated by adjacent sequences on the genome. Upstream of the gene is the *promoter*, which is involved in the attachment of RNA polymerase to the DNA strand. Promoters have an element called a TATA box, which specifies the correct 5′ end of the RNA precursor. Several promoter-specific transcription factors have now been isolated, each specific for a gene promoter which, when bound, activates transcription. TATA-binding protein is a general transcription factor for RNA polymerases I, II and III. *RNA polymerase III* is responsible for the synthesis of a variety of small RNA molecules including tRNA and small subunits of ribosomal RNA. *RNA polymerase II* carries out the transcription of genes, and *RNA polymerase I* synthesizes the large subunits of ribosomal RNA.

REGULATING GENE EXPRESSION IS CONSTANTLY ACTIVE, HAS MULTIPLE PATHWAYS AND IS INFLUENCED BY ENVIRONMENT (EPIGENETICS)

Ordering of DNA and accessibility for transcription

Histones are proteins that package DNA within nucleosomes. They act to form a structure that DNA spools around, are key to gene regulation and are influenced by environmental change (*epigenetics – changes in gene expression not caused by changes in DNA sequence*). Several families of histones exist and the formation and structure of the nucleosome, which in turn designates which areas of DNA are open for transcription, is influenced by key enzymatic modification, such as methylation, acetylation, phosphorylation, ubiquitination and SUMOlyation. For example, methylation of histone 3 at lysine residue 4 (H3K4Me3) at the promotor of active genes is associated with active transcription, as opposed to methylation at lysine 27, H3K27Me3, which results in gene repression.

Similar mechanisms may occur at DNA level. DNA methylation, in particular *CpG* repeats, confers an 'off-switch' where genes remain repressed. Methylation is essential for cell cycle and cell differentiation and alteration (which is part of epigenetic target) in methylation profile or sites of methylation can alter cell differentiation, fate and gene function.

Non-coding RNA and RNA-binding proteins constantly regulate gene expression

Gene expression can be further regulated during or after transcription (post-transcriptional regulation of gene expression). For example, encoded by nuclear DNA are microRNA (miRNA), which are small non-coding RNA molecules (around 22 nucleotides) that have complementary sequences with mRNA and usually result in regulation via gene silencing or via RNA-binding proteins, which are nuclear or cytoplasmic proteins that regulate RNA transcripts (for example within the spliceosome).

More work is still required on the understanding of RNA-binding proteins and biological and disease relevance; however, we do appreciate that miRNA have multiple mRNA targets, while each mRNA may also be targeted by several miRNA. Given that our genome codes for over a 1000 miRNA, there is therefore a rich network by which regulation of gene expression occurs, while aberrant expression of miRNA has been implicated in disease, in particular inflammation and cancer. Such miRNA has become very topical in the pathogenesis of degenerative disorders such as age-related macular degeneration. The formation of RISC (RNA-induced silencing complex) incorporates miRNA and activates enzyme RNA endonuclease III Dicer 1 that aids formation of small interference RNA (siRNA) within RISC and inhibition of gene expression (Fig. 3-4).

Spliceosomes

When the gene is transcribed (Fig. 3-3), the precursor mRNA is then further modified in spliceosomes where non-coding sequences are removed and coding sequences (exons) are ligated. Recognizing this process is important, as some proteins undergo 'alternative' splicing and thus form other active mRNA with the consequence of altered protein action and cell response and behaviour. Understanding mechanisms of splicing and the spliceosome therefore remain important, as altered activity is implicated in disease as well as targets for therapeutics.

Chromosome defects and gene mutations

During cell division the division of chromosomes may be imperfect. The extent, however, of genetically determined disease remains undetermined. The risk of genetic abnormality detectable at birth or in infancy is approximately 1 in 40. However, most fetuses with chromosomal aberrations are spontaneously aborted and it has been estimated that the frequency of genetic defects in live births is 0.6% and in stillbirths 5%. Chronic diseases with a significant genetic contribution occur in about 10% of the adult population. Abnormalities of the chromosomes are usually

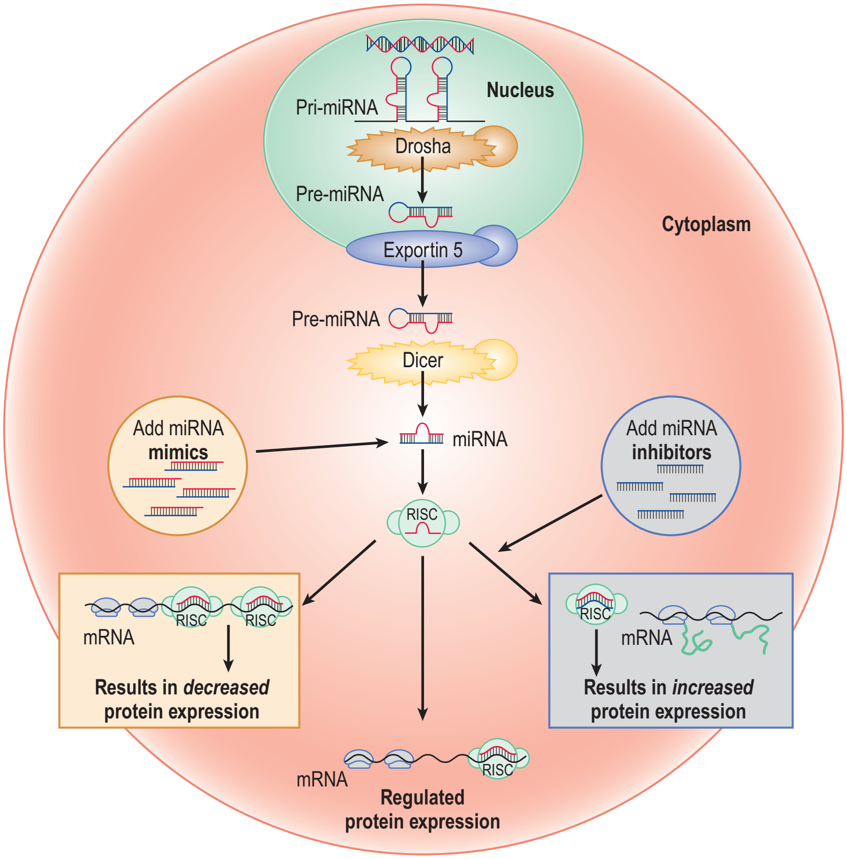
Nucleus
Pri-miRNA
Drosha
Pre-miRNA
Exportin 5
Cytoplasm
Pre-miRNA
Dicer
Add miRNA
mimics
miRNA
Add miRNA
inhibitors
RISC
RISC
RISC
mRNA
Results in *decreased* protein expression
RISC
mRNA
Results in *increased* protein expression
RISC
mRNA
Regulated protein expression

FIGURE 3-4 Micro-RNA pathway (adapted from an illustration by Michael Hamers, Lightspeed Design and Branding, Boulder, CO, USA, with permission from miRagen Therapeutics http://www.miragentherapeutics.com).

classified as numeric abnormalities, where somatic cells contain an abnormal number of normal chromosomes, or as structural abnormalities, where somatic cells contain one or more abnormal chromosomes, which in turn may affect either sex or autosomal chromosomes.

Further information on numerical chromosomal abnormalities is available at https://expertconsult.inkling.com/.

STRUCTURAL CHROMOSOMAL ABNORMALITIES

Structural chromosomal abnormalities result from chromosomal breakage. Normally single breaks are repaired quickly, but if more than one break occurs repair mechanisms may cause random rejoining of the wrong ends. Spontaneous breakage increases with exposure to mutagenic chemicals and ionizing radiation. The following structural abnormalities may occur:

- translocation – chromosomes break and exchange segments
- inversion – segment of chromosome is inverted in sequence
- deletion – section of chromosome is lost
- mutation – point mutation occurs with a change in a single base of a triplet code of a gene.

Translocation usually results in no loss of DNA so that individuals may appear clinically normal. Translocation may be reciprocal if exchange of chromosomal segments distal to the breaks occurs. *Robertsonian* translocation arises from breaks at or near the centromere in two acrocentric chromosomes (where centromere is located nearer one end of the chromosome). For example translocation of segments on

NUMERICAL CHROMOSOMAL ABNORMALITIES

Aneuploidy arises from failure of paired chromosomes to disjoin or as a result of delayed movement during anaphase, thus producing an extra copy of the chromosome (trisomy) or a missing copy of the chromosome (monosomy). Meiotic aneuploidy can occur during either of the stages of meiosis. Polyploidy describes a complete extra set of chromosomes (69 in triploidy), which usually results from fertilization of the ovum by two sperm cells. This usually results in miscarriage.

Autosomal Trisomies (eFig. 3-1)

Trisomy describes the individual chromosome that is somatically triplicated in the cell. Triplication of individual chromosomes gives rise to characteristic syndromes with multisystem disorders. Examples of autosomal trisomies include trisomy of chromosome 21 (Down syndrome), chromosome 18 (Edwards syndrome) and chromosome 13 (Patau syndrome). In each of the autosomal trisomies, mosaicisms (two or more cell lineages derived from a single zygote) may influence the clinical picture markedly, particularly when the normal cell line is present in a critical tissue. Trisomy, as mentioned above, may result from a failure of the separation of chromosome pairs during cell division (nondisjunction) or (in approximately 4% of cases) from translocation of one chromosome onto another inherited from one of the parents, thus giving the offspring a normal number of chromosomes but with a translocated copy of, for example, chromosome 21 (Robertsonian translocation).

Meiotic Aneuploidy

Similar chromosomal anomalies exist with the sex chromosomes, whereby an extra X or Y chromosome gives rise to a phenotype not distinguished in the normal population. Loss of an X chromosome gives rise to Turner syndrome (45 X0), which again may not present with the characteristic phenotype because of the prevalence of mosaic forms and therefore normal sexual development may be present. This is also true for Klinefelter syndrome, in which there is an extra X chromosome (47 XXY).

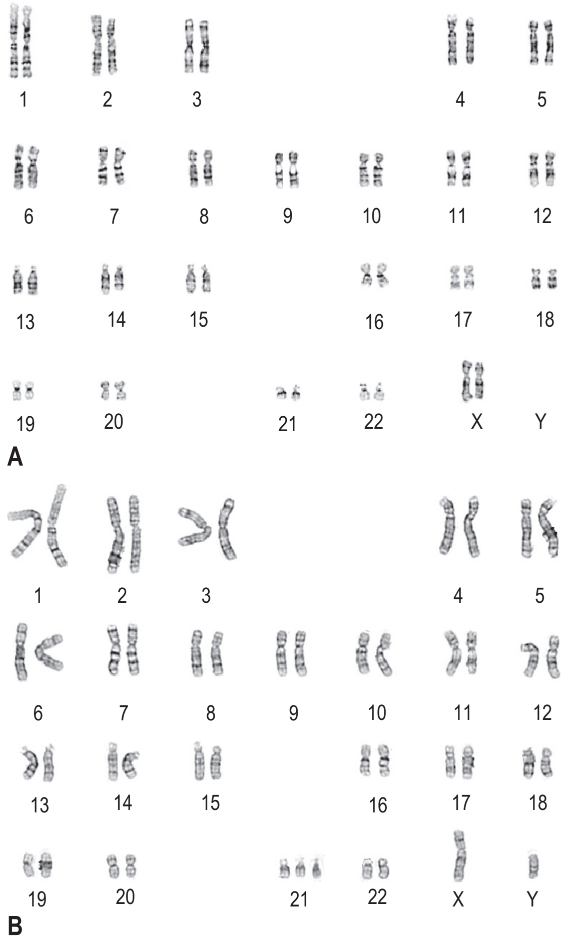

eFIGURE 3.1 (**A**) Normal human karyotype: 22 autosome pairs and sex chromosomes. (**B**) Trisomy of chromosome 21.

chromosomes 14 and 21 may occur and, during meiosis, a trivalent is formed which will then lead to a mosaic of normal and abnormal karyotype during anaphase. Insertional translocation requires the occurrence of three breaks in one or two chromosomes, resulting in deletion of one segment and insertion of another into the gap in the first chromosome. Inversions arise from two chromosomal breaks and the segment being inverted through 180° between the breaks. Inversions interfere with the pairing of chromosomes during meoisis and crossovers are suppressed, generating unbalanced gametes. Deletion may also occur when only part of the chromosome is lost. This can occur between two breakpoints or as a result of breaks and loss of segments in both arms of the chromosome. Point mutations, on the other hand, occur when a single nucleotide base is replaced by another, which may in turn alter the amino acid coding for that protein.

GENE MUTATIONS

Many types of mutation may occur throughout the alleles at each locus. Within the normal population, and in examples of inherited disease, mutations can occur that range from a single base pair to deletion of large gene segments involving many millions of base pairs. The description of these mutations has led to increasing availability of diagnostic and screening tools for genetic disease. Box 3-4 details the diversity of gene mutations that occur within the human genome.

BOX 3-4 MUTATIONS WITHIN THE HUMAN GENOME

Point Mutation (Nucleotide Substitution)

A single nucleotide change in the transcribed codon leading to a different amino acid substitution in the polypeptide chain – *missense mutation.*

A single nucleotide change which results in a stop codon, giving rise to a shortened polypeptide chain, or vice versa, resulting in an elongated polypeptide chain – *nonsense mutation.*

A single nucleotide base change which alters a critical splice junction and leads to abnormal RNA processing or a reduction in the normal gene product, or prevents the efficient addition of the poly-A chain for effective transcription – *splice site mutation.*

Deletions and Insertions (Frameshifts)

Deletion of one or two nucleotide bases leading to incorrect reading during translation and resulting in an inappropriate amino acid sequence and premature termination of the polypeptide chain.

Codon Deletions and Insertion (Repeat Expansions)

Insertion of a repeated codon (three nucleotide bases) that, as a consequence, interrupts the coding sequence. These mutations, known as *triplet repeats,* are found in myotonic dystrophy, the fragile X syndrome (X-linked mental retardation), Huntington's disease and spinocerebellar ataxia.

Clinical genetics

There are several main types of genetic disorders: chromosomal (see above), mitochondrial, multifactorial, somatic cell genetic and single gene (autosomal and X-linked inheritance) disorders (Box 3-5).

AUTOSOMAL INHERITANCE

A trait that is determined by a gene on an autosome may be *dominant* or *recessive*. Heterozygotes are individuals with different alleles at the corresponding locus on the pair of homologous chromosomes. Homozygotes have the same allele. As such, autosomal dominant refers to the situation where a monogenetic disorder manifests clinically in the heterozygous state, and inheritance is usually from one parent only. Autosomal recessive traits refer to disorders that, while they can affect both sexes, require both abnormal genes for the disease to be clinically expressed.

Additional content available at https://expertconsult.inkling.com/

X-linked disorders

X-linked inheritance refers to the pattern of inheritance carried by the genes on the sex chromosomes. This inheritance therefore carries a characteristic family pedigree. An X-linked recessive trait carried on the X chromosome is manifest only in females when the homozygous state exists, because in heterozygotes the normal dominant gene would be expressed. Such individuals may express disease traits. Traits are thus transmitted from healthy female carriers or affected males. An affected male would pass the trait to all his daughters, who would then be heterozygous carriers, but he would be unable to pass the disorder

AUTOSOMAL DOMINANT DISORDERS

The overall incidence of autosomal dominant inheritance (eFig. 3-2A) of disease in the UK is 7 per 1000 live births. Usually there is one affected parent, unless the trait occurs as a new mutation, in which case neither parent expresses the disease. Affected individuals have a 1 in 2 chance of passing it on to their child. This form of inheritance is not related to sex chromosomes and male-to-male transmission may occur. In contrast, X-linked disorders cannot be transmitted by males (see below). In clinical practice the detection of autosomal dominantly inherited disease is complicated by variation in expression of the gene and by new mutations. As such the clinical manifestation in an affected individual may not be apparent, giving rise to the view that the condition has skipped a generation. Most autosomal dominant disorders present clinically with features in later life, often after the carriers of the gene have completed their families. This raises difficulties in genetic counselling, and for this reason consideration is now being given to developing predictive tests based on gene tracking with DNA probes. The more severe the disorder the less likely the individual is to reproduce, so that the proportion of individuals affected as a result of new mutations will increase. Both new mutations and variable gene expression can combine to create difficult clinical decisions for the affected individuals as well as the clinical geneticist.

AUTOSOMAL RECESSIVE DISORDERS

Autosomal recessive traits affect both sexes but the trait is manifest only if both abnormal genes are present (eFig. 3-2B), i.e. the patient has no normal allele at the affected locus. Usually both parents are heterozygous carriers of the gene in question and are clinically normal. Thus there is a 1 in 4 chance of any offspring being affected, a 1 in 2 chance of producing a heterozygous carrier, and a 1 in 4 chance of being normal. In rare recessive traits there is usually a strong family history of consanguinity between first cousins, who share 1 in 8 of their genes by virtue of their common ancestry. In practical terms the incidence of severe autosomal recessive disorders is low, making the risk to the sibling of first-cousin parents small. In the UK the overall incidence of autosomal recessive disorders is 2.5 per 1000 live births, one of the commonest being cystic fibrosis.

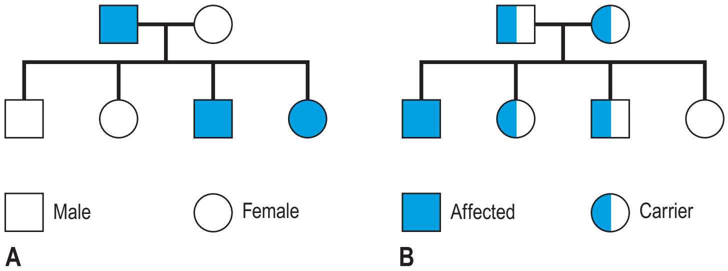

eFIGURE 3.2 Pedigree of (**A**) autosomal dominant inheritance, (**B**) recessive inheritance.

BOX 3-5 CHARACTERISTICS OF MENDELIAN INHERITANCE

AUTOSOMAL DOMINANT

- Vertical transmission
- 50% of offspring affected
- Males and females affected equally
- Variable expressivity
- Unaffected persons do not pass on trait
- Homozygotes are more severely affected

AUTOSOMAL RECESSIVE

- Recessive gene does not cause disease in heterozygotes
- Disease expressed in homozygotes
- Males and females affected equally
- Constant expressivity
- Affected individuals have children who are carriers

X-LINKED RECESSIVE

- Vertical transmission
- Usually only males are affected
- All daughters of affected males are carriers
- Heterozygous females may be affected because of 'lyonization'
- Variable expressivity

BOX 3-6 LYONIZATION

Lyonization is the process of inactivation of one member of a pair of X chromosomes in every female somatic cell. The Barr body (densely stained mass of chromatin, which is the inactivated X chromosome) occurs in the trophoblast by day 12 after fertilization. Inactivation occurs only in somatic cells; it is random and then fixed for all subsequent generations of that cell line. The inactive X chromosome is not transcribed, except for a small region at the tip of the short arm. If loss of material occurs in one X chromosome, then the structurally abnormal chromosome is inactivated. About 30% of cells from buccal smears show a Barr body and 1–10% of female neutrophils will also show an inactivated X chromosome (drumstick). Females have random expression of either paternal or maternal X chromosome, which clinically gives rise to patchy expression of mutant X-linked genes in carrier females (e.g. retinitis pigmentosa and choroideraemia).

on to his sons, who receive only his Y chromosome. In some X-linked recessive disorders a proportion of female heterozygotes are affected more as a trait. The explanation lies in the fact that only one of the X chromosomes in a cell is active (Lyon hypothesis; see Box 3-6). This inactivation occurs early in ontogeny and thereafter the descendants of the cell have the same inactive X chromosome. By chance, therefore, females may inactivate the healthy chromosome and thus manifest the disease.

In the UK the incidence of X-linked inherited disorders is around 1 per 1000 live births. The term X-linked dominant has been used to describe X-linked inheritance where the heterozygous female is regularly affected. In these disorders males are frequently more severely affected and do not survive gestation (e.g. Aicardi syndrome: agenesis of the corpus callosum and chorioretinopathy). If the affected male does reproduce, none of the sons will be affected and all the daughters will be. Occasionally these pedigree patterns may be mimicked by autosomal traits that display sex limitation. If the affected males of an autosomal dominant trait with sex limitation are infertile, the pedigree pattern is identical to that of X-linked recessive traits, except that carrier females exhibit lyonization. The family pedigree will also show a smaller proportion of affected females than would be expected in X-linked dominant traits.

MITOCHONDRIAL INHERITANCE

Cells have multiple copies of an extranuclear chromosome, contained in mitochondria. The mitochondria contain about 10 single circular chromosomes, which replicate independently of the nuclear genome. Within their genome there are the genes for tRNA and ribosomal RNA necessary for mitochondrial protein synthesis and peptides involved in cellular oxidative phosphorylation. As the acrosome of the sperm is lost during fertilization, the inheritance of mitochondrial chromosomes is exclusively through the female ova, which gives rise to a characteristic form of cytoplasmic inheritance with the following features: there is no transmission from males to their offspring, although both males and females can receive the defective gene from their mother; if all the mitochondria carried the gene, then all offspring of the affected woman would also be affected. However, the precise pattern of the disease may not be recognized if only a proportion of the mitochondrial genes are affected.

MULTIFACTORIAL INHERITANCE

There are many conditions that have a familial incidence, yet the incidence of affected siblings is less than can be accounted for by unifactorial Mendelian inheritance. This is because the clinical presentation is influenced by several genes and by environmental factors. In multifactorial inheritance both genetic and environmental factors combine in varying proportions in different individuals to permit the disease to be manifest. Each trait is determined by the interaction of a number of genes at different loci. Multifactorial inheritance may be *continuous* or *discontinuous*. In discontinuous multifactorial traits the risk within affected families is raised above that in the general population, but is low compared with that of single-gene defects. In continuous traits there is a range of disease expression, with individuals falling between two extremes (for example, blood pressure).

Multifactorial discontinuous traits have an increased incidence in affected families compared with that in the general population. If multifactorial inheritance is suspected, twin concordance and family correlation studies are undertaken. For instance, in multifactorial inheritance the incidence of the trait in monozygotic twins will exceed that in dizygotic twins. Some discontinuous multifactorial traits will also show unequal sex distribution, indicating that for one of the sexes the threshold for disease penetrance is higher than for the other sex, i.e. a higher proportion of underactive genes is required before the disease is expressed.

DISEASE ASSOCIATIONS

TABLE 3-1 **HLA-linked disease associations**

Disease	HLA allele	Relative risk
Ankylosing spondylitis	B27	90–100
Acute anterior uveitis	B25	8
Reiter syndrome	B27	40
Behçet syndrome	B51	4–10
Birdshot chorioretinopathy	A29	50–224
Intermediate uveitis	DR15	6
Sympathetic ophthalmia	DR4	14
Vogt–Koyanagi–Harada syndrome	DR4	12

One approach to identifying disease associations with a specific genotype is to examine major genetic polymorphisms in a population and search for associations. Strong associations have been obtained between ankylosing spondylitis and *HLA-B27*, and birdshot chorioretinopathy and *HLA-A29*. In practice it is perhaps the absence of such allelic correlations that is more useful in predicting disease occurrence. Disease associations can be expressed as relative risk by comparing the incidence of the allele within a control population with allelic expression in patients. Table 3-1 gives a list of some of the commoner associations. Currently, high-resolution typing that characterizes the nucleic acid sequence of major histocompatibility complex alleles has meant that, for example, classical DR4 associations are now further subtyped, e.g. the association of *DRB1 *0404* with sympathic ophthalmia and Vogt–Koyanagi–Harada syndrome.

Genetic polymorphisms

Within the human genome many gene loci possess commonly occurring alleles, which can be used to distinguish populations into discrete phenotypes. This was first described for components of human blood, particularly ABO and MNS phenotypes, where allelic differences in the former determine terminal sugar residues on membrane glycoproteins and those in the latter determine the amino acid sequence of membrane-bound glycoproteins. This is also true for the Rhesus system where polymorphisms exist for the Rhesus polypeptide on the red blood cell membrane. Genetic polymorphisms are defined as the occurrence of multiple alleles at a locus, where at least two alleles occur with a frequency greater than 1%. The existence of different versions of the same genetic material in different people has prompted their use as genetic markers. This is particularly the case for human enzymes and other proteins in which polymorphisms are detectable in population groups. The use of polymorphisms in genetics will be highlighted below when we discuss the types and uses of molecular polymorphisms. To summarize at this point, their practical use in medical genetics is seen for mapping genes to individual chromosomes by linkage analysis (see below), presymptomatic and prenatal diagnosis of genetic disease, evaluation of high-risk and low-risk persons with a predisposition to common adult disorders (e.g. diabetes mellitus), and tissue typing for both tissue and organ transplantation.

Population genetics

Population genetics is the study of the distribution of genes within populations and how these gene frequencies are changed or maintained. This form of study is important for the understanding of *linkage analysis* (linked genes that have their loci within measurable distances of one another on the same chromosome) and *linkage disequilibrium* (the association of two linked alleles more frequently than would be expected by chance), which is described later.

With our increased understanding of the regulation of gene expression that we have discussed above, it is now appreciated that gene expression and cell function (differentiation) is influenced by *epigenetics*.

GENE FREQUENCY

When contemplating the genetic makeup of whole populations, in particular when relating the frequency of people who are heterozygotes at a particular locus to the frequency of homozygotes, the *Hardy–Weinberg equilibrium* is employed. This calculation is based on the assumption that in a given population where random mating occurs, in the absence of mutation and selection, the genetic constitution remains the same from one generation to the next.

It is important to remember that genes are paired in alleles, so that within the Hardy–Weinberg equilibrium, p is the proportion of normal alleles and q is the proportion of abnormal alleles, so that $p + q = 1$. Since a pair of alleles occurs at each gene locus, the relative proportions of normal homozygotes, heterozygotes and abnormal homozygotes are given by the equation:

$$p^2 + 2pq + q^2 = 1.$$

A few important points can be derived from this equation.

- Although the gene frequency determines the chance of any one chromosome carrying the abnormal allele, individuals have two alleles for each gene, which thus doubles the chance of having the abnormal allele on one or other chromosome.
- The heterozygote frequency is often twice the gene frequency.
- In autosomal recessive inherited disorders the gene frequency is the square root of the abnormal homozygote frequency.

It is also important to note that the Hardy–Weinberg equilibrium is concerned only with alleles at a single gene locus; in practical terms the relationship of gene loci to each other along the length of the chromosomes is of greater importance.

GENETIC LINKAGE AND LINKAGE ANALYSIS

Two genes may be transmitted more frequently than independent assortment would suggest.

During meiosis a crossing over between homologous chromosomes may occur at any point along the length of the chromosome. Therefore, the closer two genes are to each other, the more likely they are to be transmitted together. Gene loci on the same chromosome are therefore said to be linked when their alleles do not show independent segregation during meiosis (i.e. their loci are within measurable distance of one another on the same chromosome). Although genes may be mapped on the chromosome by techniques such as *in situ* hybridization (see below), genetic linkage may also be shown by family studies, confirming a genetic background to the disorder. In a family linkage study two loci are considered: one for the disease, the other for the trait.

One way of performing linkage analysis is to study family members who are heterozygous at each of the gene loci. One must first establish the linkage phase, determining which allele at locus 1 to be studied is on the same chromosome (of a pair) as a particular allele at locus 2 (marker locus). The family must of course be informative for the loci being considered (i.e. express the phenotype). The loci most informative for linkage analysis are those that are highly polymorphic and thus heterozygous in a large proportion of the population. The probability of meiotic crossing over between them is in effect a means of describing the distance between linked loci. Gene loci that are sufficiently well separated on the same chromosome will always have several crossovers between them and thus exhibit a 50% recombination rate (as if they were situated on different chromosomes and segregating independently). To establish the probability of two gene loci being linked, the number of opportunities for recombination and the proportion in which they have occurred is counted. This may be given as a ratio of the probability of observing the data in the sibship, based on the assumption that there is no linkage between the two loci. As large sibships are required to assess probability of linkage, it is calculated by computer and given as a logarithmic ratio, known as the Lod score (derived from log odds). In most situations there is significant linkage if the Lod score is greater than 3, and linkage can be ruled out if it reaches −2.

Linkage equilibrium and disequilibrium

Linkage equilibrium exists when, in the presence of random mating and given enough generations, the various combinations of alleles that could occur are expressed with equal proportion. *Linkage disequilibrium*, on the other hand, describes the association of two alleles more frequently than would occur by chance. This linkage may arise from mutation of one allele that has not yet established equilibrium because of a selective advantage or disadvantage. The same principles of equilibrium and disequilibrium can apply to alleles of any linked gene loci. At this point it should be emphasized that close linkage between a disease-specific gene locus and a gene showing considerable polymorphism does not necessarily mean association between the two particular alleles of the gene loci. The association between disease and human leucocyte antigens (HLA), for example, arises as a result of both genetic linkage between *HLA* allele and the disease susceptibility gene, and linkage disequilibrium involving the particular allele at the *HLA* locus. Linkage disequilibrium is important clinically as a means of identifying disease-causing genes and identifying the origin and spread of mutations. As mentioned above, genetic polymorphism is the occurrence in the same population of two or more discontinuous traits at a frequency where the rarest trait would not be maintained by recurrent mutation alone. The most commonly accounted genetic polymorphisms exist within blood groups (ABO) and cell surface antigens (HLA) but more recently, with the advent of molecular biological techniques, DNA sequences of genes have shown that many are polymorphic (see below).

Mendelian randomization identifies causal modifiable environmental risk to disease

Studying the influence of environmental risk factors influenced by gene mutations for complex multifactorial disorders such as cardiovascular disease to move from just association (as may be determined through genome-wide association studies, GWAS) toward causality involves *Mendelian randomization*. Mendelian randomization uses known variation in genes of known function to examine the causal effect of a modifiable exposure on disease. The studies rely on data from genetic association studies, as misleading conclusions on causality can occur because of linkage disequilibrium, genetic heterogeneity, pleiotropy or population stratification. In order to test for supposed causality, for example, the method takes common genetic polymorphisms which have indirect biological consequences on environmental exposure patterns (e.g. nicotine addiction) or consequences generated from modifiable risks (e.g. raised blood pressure).

EPIGENETICS (Box 3-7)

The term *epigenetics* is now largely understood as a heritable phenotype resulting from changes in a chromosome without alterations in DNA sequence. Mechanisms that give rise to epigenetic inheritance in which a stably inherited cell memory exists and altered gene expression without altering DNA sequence include: DNA methylation and chromatin remodelling, RNA-binding proteins and miRNA (as discussed above).

BOX 3-7 EPIGENETICS

Genetic imprinting on offspring results from heritable patterns and is different from both mother and father.

Epigenetic regulation of gene expression and cancer may result from epigenetic carcinogens or alteration in DNA methylation patterns (including DNA repair mechanisms – BRCA1 and breast cancer).

Cancer therapy is currently identifying histone deacetylases or methyltransferases that, as therapies, can alter the cell differentiation and transcriptome profile of malignant cells.

Study of the epigenome is currently ongoing in understanding of methylation status in tumours. We are able to interrogate changes by assays such as:

- chromatin immunoprecipitation sequencing (CHiP-Seq)
- RNA-Seq (identifying RNA expression or coding levels, and can include non-coding siRNA).

Understanding the human genome: DNA analysis

The advent of modern molecular biological techniques has benefited many branches of medicine, including ophthalmology. In particular the techniques described below have increased our fundamental understanding of such conditions as X-linked ophthalmic disorders, retinoblastoma and Leber's hereditary optic neuropathy. Some of the techniques used to examine genes and genetic linkage are shown, with examples of how this has helped in the identification and understanding of the molecular and genetic basis of ophthalmic diseases.

CLONING

Cloning (Fig. 3-5) is an *in vivo* technique that produces identical copies of DNA sequences, which can then be used as gene probes. Cloning is achieved by inserting the specific fragment of DNA to be studied into another DNA molecule, which can then be replicated quickly in bacteria (recombination). Fragments of DNA can consistently be obtained by cleaving the DNA with naturally occurring enzymes, known as *restriction endonucleases*. These restriction enzymes cut DNA at specific recognition sites, usually into lengths of four to six nucleotide base sequences. Reproducing large quantities of the same DNA fragment (i.e. cloning) can, however, be performed only by transferring the genetic material into a bacterium via a vector. The common vectors are bacteriophages or modified bacterial plasmids (extrachromosomal, self-replicating, circular DNA). The choice of vector is dependent on the size of DNA fragment to be cloned: bacteriophages have a capacity of 5–20 kilobases (kb) and plasmids a smaller capacity of up to 10 kb. These vectors replicate their own DNA independently of the host DNA and also have single restriction enzyme sites, which allow insertion of recombinant molecules. After slicing the DNA molecule, the recombinant DNA is prepared by incubation of the plasmid with the fragment in the presence of an enzyme, *DNA ligase*, which ligates or binds the DNA to the cut ends of the plasmid. To date, over 200 restriction enzyme sites have been recognized. Restriction enzymes are named according to the organism of origin, e.g. E-cor is derived from *Escherichia coli*.

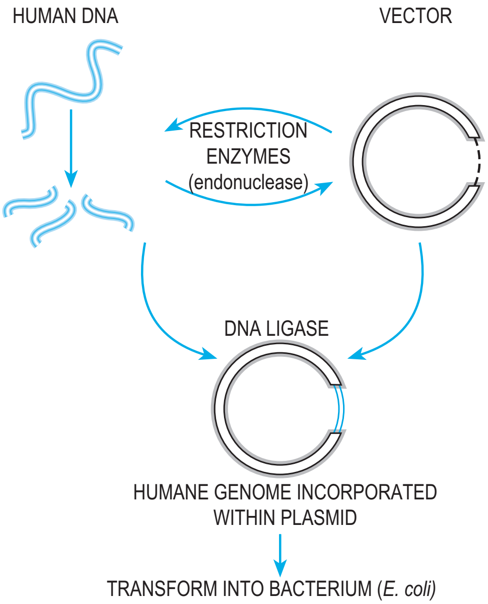
HUMAN DNA
VECTOR
RESTRICTION ENZYMES (endonuclease)
DNA LIGASE
HUMANE GENOME INCORPORATED WITHIN PLASMID
TRANSFORM INTO BACTERIUM (*E. coli*)

FIGURE 3-5 Recombinant DNA technology: cloning.

REFINING TECHNOLOGY: MOLECULAR ANALYSIS OF HUMAN GENES AND MUTATIONS

Southern and Northern blotting

The total DNA from a sample of an individual's cells may be cut into fragments using restriction endonucleases. The fragments can then be analysed according to their molecular weight by electrophoresis. The DNA fragments are denatured into single-stranded DNA and blotted on to nitrocellulose. Single-strand copies of a segment of DNA that has been cloned and labelled with a radioisotope (^{32}P), known as a DNA probe, are then hybridized with the complementary nucleotide sequences by incubating the radioactive probe with the blotted DNA in the nitrocellulose membrane. The complementary sequences are revealed as bands, by autoradiography (Southern blotting; Fig. 3-6). In a similar way, mRNA fragments can be detected by Northern blotting, in which labelled DNA probes are hybridized with electrophoresed and blotted mRNA fragments. Both the size and abundance of the mRNA from a specific gene in a sample of RNA can be determined.

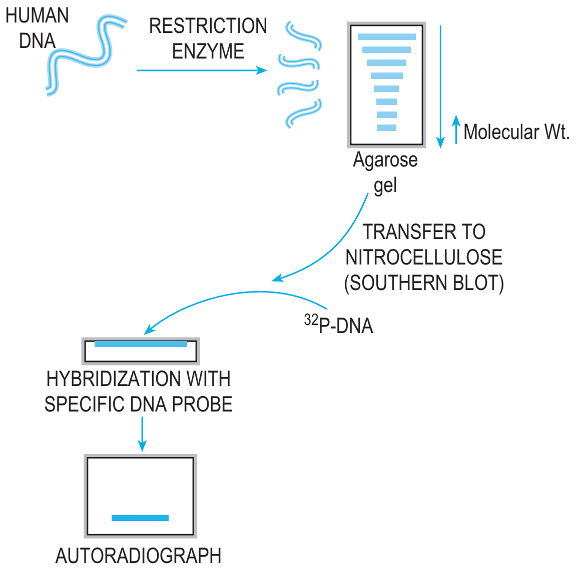
HUMAN DNA
RESTRICTION ENZYME
Agarose gel
Molecular Wt.
TRANSFER TO NITROCELLULOSE (SOUTHERN BLOT)
32P-DNA
HYBRIDIZATION WITH SPECIFIC DNA PROBE
AUTORADIOGRAPH

FIGURE 3-6 Recombinant DNA technology: Southern blotting.

DNA polymorphisms

Polymorphisms in DNA segments (genes) can be detected by many techniques, two of which are restriction fragment length polymorphism (RFLP) and variable number of tandem repeats (VNTR) polymorphism.

Restriction fragment length polymorphism

RFLP occurs because of point mutations in the human genome, which normally occur in the non-coding regions of the DNA molecule. As a result of these acquired polymorphisms, recognition sites for the restriction endonucleases are created or abolished. The consequence is that restriction enzymes will produce variable lengths of DNA fragments, which give characteristic banding on Southern blotting. If an individual is heterozygous for an RFLP, one restriction band on the electrophoretic gel corresponds to one chromosome and the other band to the other chromosome of the pair. This allows one to track the transmission of a single chromosome region through the family and perform classic linkage studies. The discovery of DNA polymorphisms has greatly increased the extent to which individual copies of particular genes are recognized as being truly unique. In fact, DNA polymorphisms are simply the molecular manifestations of variations in the genome, which have been recognized for a long time from studies of protein polymorphisms.

Variable number of tandem repeats polymorphism

Also within the genome are many tandem repeats of sequences that vary for each individual. The sequences act as spacers between restriction enzyme sites. VNTR polymorphisms are short sequences of DNA and the distance between the VNTR depends on the number of repeats. VNTR are heterozygous and can be detected using probes for the core sequence that gives rise to these repeats. VNTR are highly polymorphic and almost unique for each individual (genetic fingerprinting). VNTR markers are highly informative for genetic linkage analysis as well as for individual identification (have been used prior to gene sequencing for paternity and forensic testing). Within the genome there are several classes of repetitive DNA, whose nucleotide sequence is repeated several hundred times within the genome. These tandem repeats are again highly polymorphic and can be used for genetic linkage studies.

POLYMERASE CHAIN REACTION

The polymerase chain reaction (PCR) is a technique that allows small amounts of DNA to be amplified for detection and analysis. PCR requires two flanking sequences to the sequence that is to be amplified. DNA primers are formed from short sequences of DNA with these flanking sequences. Amplification of the DNA sequence uses heat-resistant DNA polymerase to extend the short DNA primers of about 20 base pairs, which are hybridized at either end. Cycles of denaturation with heat and polymerization with cooling occur up to 50 times. As a result the target DNA is amplified 10^5–10^6 times. This process, which is largely automated, is proving a great asset in detecting specific DNA where the amount of starting material is very small, e.g. neonatal diagnosis and the identification of infective organisms in tissue samples, and as a research tool. However, quality control of false-positive rates must be established before its full clinical efficacy can be ascertained.

GENOMICS, TRANSCRIPTOMICS AND PROTEOMICS

The genome contains genes that, as we have already described in this chapter, are transcribed to produce

RNA and ultimately generate proteins. What has been realized since the findings of the Human Genome Project is that there are very many more proteins in the *human proteome* than there are genes in the *human genome* (*c.* 400 000 proteins and 22 000 genes). The study of the *genome* (*genomics*) has been made possible by the development of *gene arrays* (Fig. 3-7). These are DNA microarrays (small solid platform devices with thousands of DNA sequences corresponding to genes, immobilized on a variety of surfaces – nylon membranes or glass) which give us the ability to look at many genes at one time. As such, we can get a picture of the possible interactions among thousands of genes. The array is an orderly arrangement of gene-specific DNA that is hybridized to mixtures of labelled RNA in samples. Therefore, based on the amount of

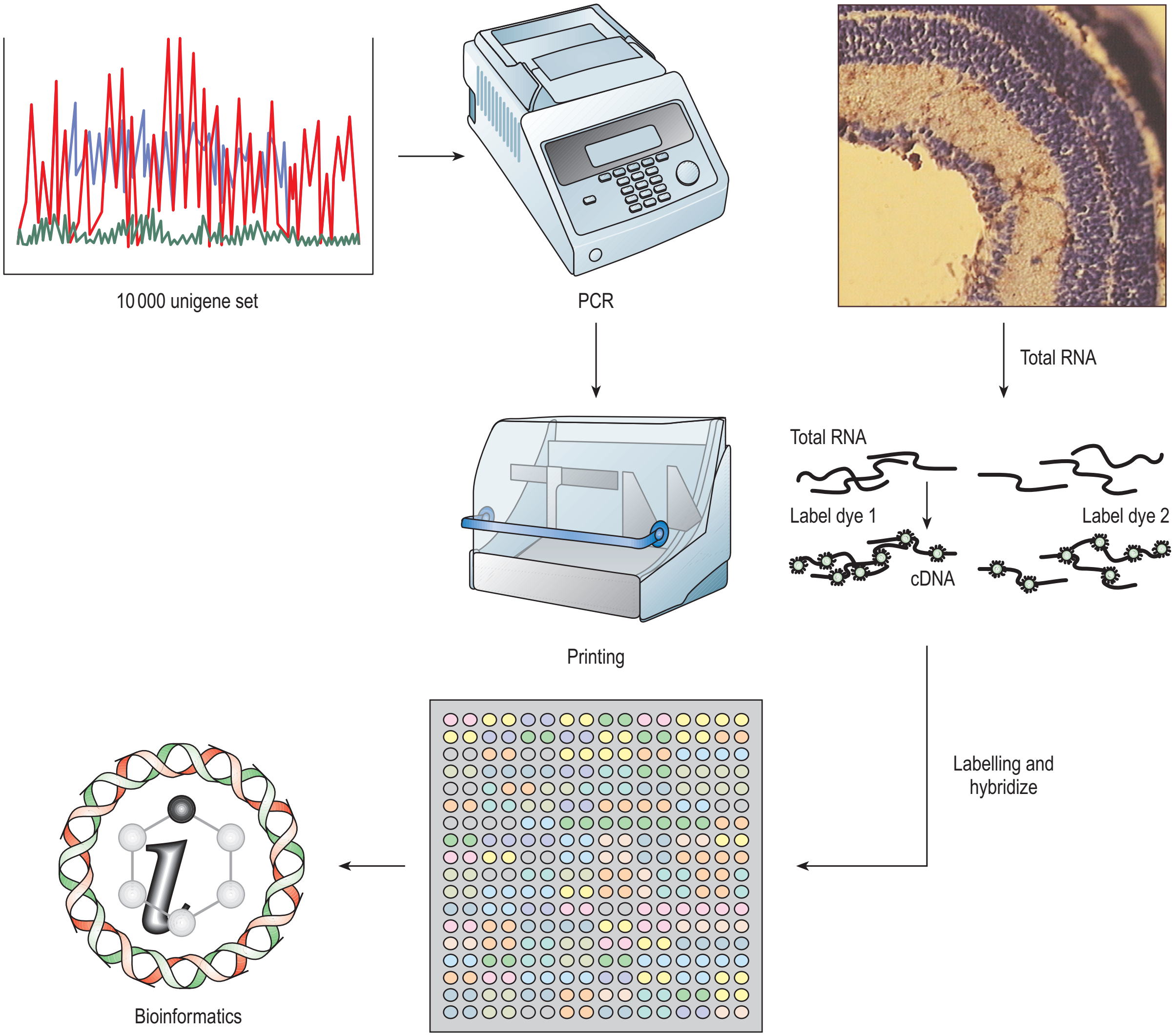

FIGURE 3-7 Gene array and bioinformatics. The array hybridization generates many thousands of computational parameters describing upregulation and downregulation of genes that have been targeted and require quantification. (Courtesy of UOB Transcriptomics Facility and Dr Chungui Lu, School of Biological Sciences, University of Bristol, UK.)

signal from successful hybridization, we can determine the relative expression of the corresponding gene. With such a capability we can now compare the extent of gene expression within a sample or of specific genes between samples (see Box 3-8). For population studies and most recently with the advent of high-throughput sequencing, programs of work have undertaken genome-wide association studies (GWAS). This platform allows for examination of genetics variants from study of single nucleotide polymorphisms in major diseases to be undertaken on a large scale (Box 3-9). Currently there is also exome sequencing, a process where sequencing of coding regions of the genome can be targeted instead of whole genome sequencing. All this has become more accessible with the advent of next-generation sequencing (NGS) technology.

BOX 3-8 GENE ARRAYS (AFFYMETTRIX GENECHIPS®)

- Isolate RNA from sample or samples to be compared
- Convert to DNA with reverse transcriptase – complementary or cDNA
- Hybridize the labelled cDNA (e.g. ^{32}P) to arrays with fixed DNA sequences (spot sizes on plate of 300 µm)
- Remove unhybridized cDNA
- Detect and quantify hybridized cDNA.

BOX 3-9 GWAS HAS ILLUMINATED SUSCEPTIBILITY OF DISEASE, POTENTIAL CAUSAL RELATIONSHIPS AND TARGETS FOR THERAPIES*

- Polymorphisms of the complement gene *CFH* have a high association in age-related macular degeneration
- IL-23R-IL12RB2 and IL-10 linked as Behçet's disease susceptibility loci
- Multiple sclerosis risk following anti-TNF therapy mirrors TNFR1 genetic variant
- Genome-wide meta-analysis identified new associated genetic susceptibility loci for refractive error and myopia.

*(See www.genome.gov.)

However, the genome is only a source of information and the initial step to provide proteins involves the transcription to RNA, where the *transcriptome* is the complete set of RNA transcripts. *Transcriptomics* is the ability to study the dynamic *transcriptome*, which varies depending upon the cell, the stage of cellular development and activation, and environmental influences. *Transcriptomics* allows us to investigate the control of RNA transcripts by activating or inactivating transcription factors that control RNA transcription, and looking at gene expression patterns under various conditions or comparing normal with pathology. The dynamics of the system are further exemplified by the fact that there are very many more proteins than genes. Also, ultimately, cell and tissue behaviour is defined by protein interactions both intracellularly and extracellularly. *Proteomics* studies these dynamic consequences of gene expression. The study relies on an ability to separate and identify proteins by, for example, mass spectrometry, gel quantification (two-dimensional electrophoresis) and protein sequence analysis. *Structural proteomics* determines three-dimensional structure by X-ray crystallography and nuclear magnetic resonance spectroscopy. How proteins interact can be assessed by affinity chromatography and fluorescent resonance energy transfer (FRET). Finally, given the diversity of proteins observed but not accounted for at the gene level, their post-translational modifications can be studied by looking at the extent of phosphorylation or glycosylation (Fig. 3-8).

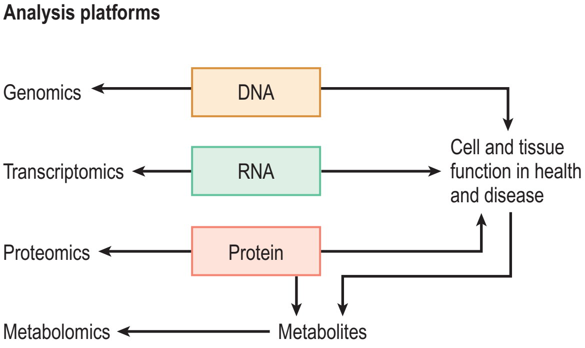

FIGURE 3-8 The era of the 'OMICS' analysis. Analysis platforms for detection of gene defects and relation to cell and tissue functional consequences.

FINDING AND TRACKING THE GENE THROUGH FAMILIES

The use of restriction endonuclease mapping (gene mapping) is a tool for analysing genetic disease. DNA obtained from tissues or cells (usually peripheral blood leucocytes) is treated with restriction enzymes and analysed by Southern blotting using a radioactively labelled gene probe. By using different restriction enzymes and orientation of the fragments, it is possible to build up restriction enzyme maps of sections of the human genome. Any normal or abnormal gene for which a specific probe has been generated can be analysed by this technique. Potential probes are chosen from a central library and screened for their ability to detect RFLPs by hybridizing them to Southern blots of DNA from a panel of unrelated individuals and by searching for restriction enzyme sites that give variable patterns. A disease gene can be mapped by collecting large pedigrees and comparing the segregation of the disease with the segregation of the RFLP. Once linkage has been clearly established, the probe is assumed to map to a chromosome region close to the disease gene and can be used to track the gene through families.

Genetic analysis has largely been superseded by techniques such as high-throughput genetic sequencing and exome analysis. There is a high demand for high-throughput sequencing (next-generation sequencing) that generates hundreds of thousands of sequences at once and is increasingly cost-effective. *Exome* sequencing is an efficient strategy that sequences the coding regions, although this risks missing possible disease-causing mutations as changes in coding regions are estimated to account for around 85% of disease-causing mutations.

Molecular biology and clinical medicine

GENE PROBES

Gene probes can be produced in several ways and fall broadly into three types: gene-specific probes, oligonucleotide probes and polymorphic probes. Gene-specific probes are produced from specific mRNA by the enzyme reverse transcriptase, which synthesizes a complementary DNA copy (cDNA) from mRNA. If radioactive bases are added to the reaction mixture, the cDNA will be labelled and can thus be used as a hybridization probe to look for the complementary sequences. Probes can be used in *dot–blot hybridization*, where serial dilutions of DNA samples are held on DNA-binding membranes and the complementary radioactively labelled probes are hybridized *in vitro*, so that the amount of radioactive signal is proportional to the amount of target DNA present. The cDNA can also be cloned by synthesizing a second DNA strand from cDNA using a bacterial DNA polymerase, which is incorporated into a plasmid and grown in bacterial cells. These oligonucleotide probes recognize short sequences of DNA that correspond to the sequence known to occur in the gene. With a probe of this length, a single mismatched base pair is sufficient to impair hybridization and can be used to detect changes to a single base (point mutations). Similarly, probes can be developed that recognize the various DNA polymorphisms within the non-coding sequences, for example RFLP and VNTR. DNA probes can also be used to identify abnormal genes or gene products at the molecular level within cell cytoplasm or nucleus, by *in situ hybridization*. This technique utilizes labelled DNA or mRNA probes which hybridize to the expressed genes in the cell in a manner similar to that used for immunohistochemistry. *In situ* hybridization can thus establish whether the genomic material of interest is present in the DNA of the cell *in vitro*.

Molecular and cell biology: controlling cell destiny

Cell death can occur by either necrosis or apoptosis (programmed cell death). Apoptosis is a process whereby developmental or exogenous environmental signals trigger specific intracellular genes, which results in cell death. Ligation of cell surface receptors such as Fas ligand (Box 3-10), which is commonly associated with death domains that signal and stimulate caspases, disrupts mitochondrial membrane channel permeability. Apoptosis is essential for normal development and many of the genes that control apoptosis (Box 3-10) have been highly conserved throughout evolution. Histologically, apoptosis is associated with chromatin condensation and nuclear DNA fragmentation. Only finally when caspase enzyme activation has occurred is membrane integrity affected and

BOX 3-10 GENE REGULATION OF APOPTOSIS

- p53 – tumour suppressor protein which functions to activate DNA fragmentation and apoptosis, thus regulating tumour development
- *Bcl-2* – gene for family of proteins which *inhibit* apoptosis
- *BAX* – gene for family of proteins which *activate* apoptosis
- Fas and TNFRI – death domain-containing proteins which when cross-linked with ligand (Fas ligand and tumour necrosis factor-α, respectively) induce apoptosis.

BOX 3-11 CELL DESTINY AND FUNCTION, MUTATIONS AND DRUG DEVELOPMENT

Some cutaneous melanomas have an activating mutation in the serine-threonine protein kinase *B-RAF* oncogene (V600E mutation) which regulates the MAP kinase pathway that controls cell proliferation. By inhibiting BRAF with new-targeted drugs inhibiting the enzyme B-RAF (vemurafenib) benefits in outcomes of stage IV melanomas are now possible.

In cystic fibrosis there is abnormal or complete loss of function of cystic fibrosis transmembrane conductance regulator gene (*CFTR*). CFTR is an ABC transporter-class ion channel regulating the constituents of sweat, mucus and digestive fluids. In a small number of cytstic fibrosis patients (5%) there is a specific mutation in the *CFTR* gene (G551D) that causes impaired ion transport but CFTR remains expressed on epithelial surfaces. Drugs have been developed (ivacaftor) to restore or enhance function to the CFTR in this targeted group of patients, improving their outcome.

the cell eliminated via phagocytosis and the reticuloendothelial system without the secondary inflammatory response that occurs when cells undergo necrosis. Understanding control of cell homeostasis (see eBox 3-1) – or where it goes wrong – is important in understanding disease pathogenesis and cancers and the development of therapies (Box 3-11).

USING MOLECULAR BIOLOGICAL TECHNOLOGY TO DETECT APOPTOSIS

Observed, and therefore measurable, changes in apoptotic cells include intranucleosomal cleavage of DNA, which generates both monomers and multimers of DNA that can be detected by DNA electrophoresis (DNA laddering). Early changes can be detected by labelling DNA fragments at the 3′ OH ends with terminal deoxynucleotidyl transferase either immunohistochemically (TUNEL staining) or by flow cytometric analysis. Changes in apoptotic cell membranes can be detected by increased flip-flopping of the plasma membrane and exposure of phosphatidylserine residues on its outer leaflet. Phosphatidylserine is detected by annexin V via either *in situ* histochemistry or flow cytometry. Apoptosis ultimately increases intracellular caspase activity and cell lysates can be tested for levels of caspase enzymatic activity by simple fluorometric analysis.

GENE THERAPY

Gene therapy is the transfer of selected genes into a host with the intention of alleviating or curing disease. There are many gene therapy strategies that can be used; which one to choose depends upon the pathogenesis.

- *Gene augmentation therapy* For diseases caused by loss of gene function and which will supply more copies of normal gene in the hope of restoring normal phenotype, e.g. cystic fibrosis, haemophilia, severe combined immunodeficiency syndrome (SCID), retinitis pigmentosa.
- *Targeted killing of specific cells* Genes are directed to specific cell types and incorporated into the genome, expressed so that protein interferes with cell cycling and survival, thus killing cells.
- *Targeted mutation correction* Can be performed using *ribozymes* (which cleave and repair mRNA), *triple helix oligonucleotides* (block gene transcription) and *anti-sense oligonucleotides* (that block mRNA translation).
- *Targeting inhibition of gene expression* When diseases display novel gene products or excessive expression of gene product, then blocking at a single gene level (either DNA or RNA) or blocking protein can be possible to attain specific inhibition of expression.

Achieving gene therapy depends upon the size of the DNA fragments to be transferred (*transfection/transduction*). Techniques can include injection of naked DNA, which is not limited by size or number of genes but is inefficient (*gene gun*). Often, therefore, the gene to be transferred is not conventional and the

eBox 3-1

Ubiquitin (Ubq)–proteosome pathway maintains cellular homeostasis

Proteosomes are nuclear and cytosolic protease complexes that degrade proteins, which become covalently linked to Ubq via a cascade of enzymatic reactions (Ubq 1–3). Ubq is a highly conserved 76-amino-acid protein that has been implicated in the pathogenesis of genetic disease and malignancies. It is integral in the destruction of phosphorylated cyclins and cyclin inhibitors, such as p27, an inducible inhibitor of cyclin-dependent kinase activity (see Box 3.2); transcriptional proteins such as STAT 1; and p53, a tumour suppressor gene product. Viruses (e.g. human papillomavirus) can upregulate p53 degradation in the Ubq–proteosome pathway, a process implicated in tumour formation. Proteosome inhibitors such as lactacystin can arrest cycle–cycle progression and therefore offer potential as chemotherapeutic agents.

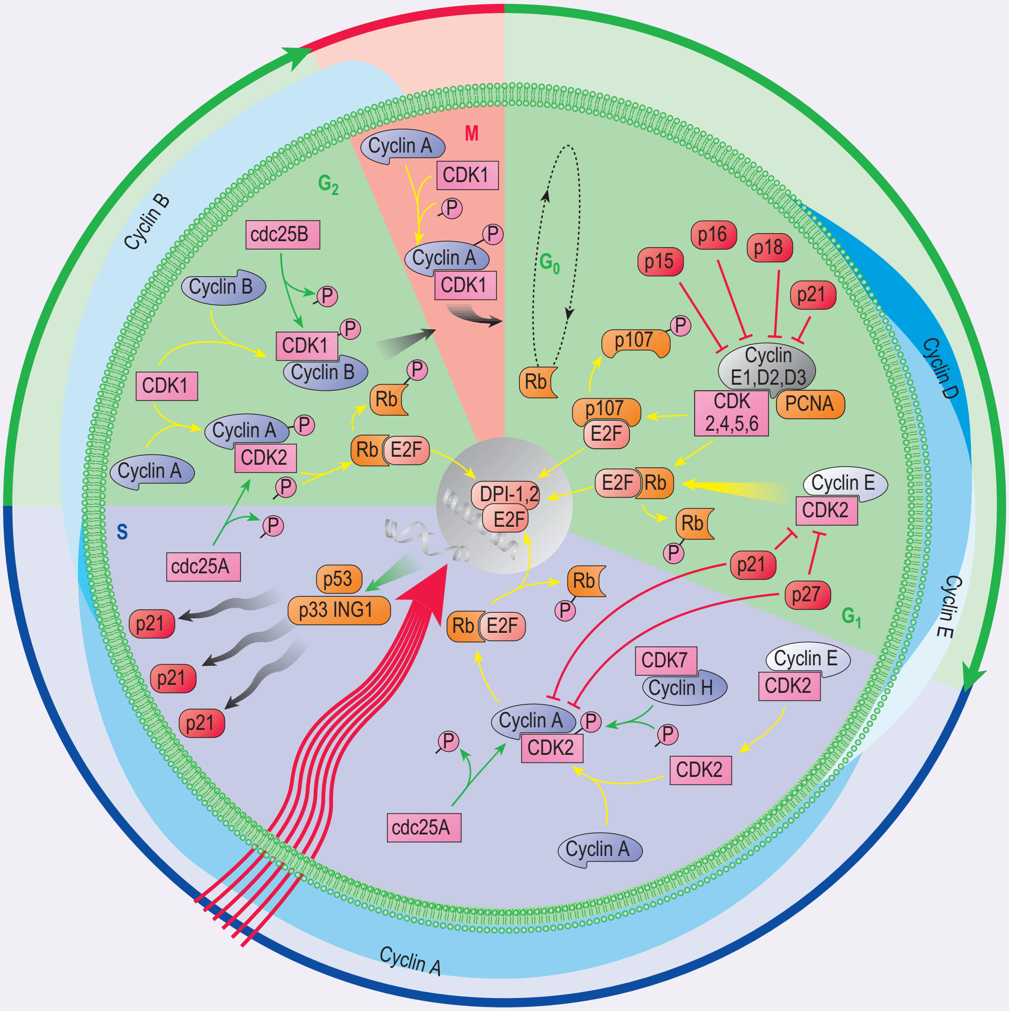
M
Cyclin A
CDK1
P
Cyclin A
P
CDK1
G_2
cdc25B
Cyclin B
P
CDK1
P
Cyclin B
Cyclin B
CDK1
Rb
P
Cyclin A
P
CDK2
Rb
E2F
Cyclin A
P
P
S
cdc25A
p53
p33 ING1
p21
p21
p21
DPI-1,2
E2F
G_0
Rb
p107
P
p15
p16
p18
p21
Cyclin E1,D2,D3
CDK 2,4,5,6
PCNA
p107
E2F
E2F
Rb
Cyclin E
CDK2
Rb
P
p21
p27
G_1
Cyclin D
Cyclin E
Rb
P
Rb
E2F
CDK7
Cyclin H
Cyclin E
CDK2
Cyclin A
P
CDK2
P
P
CDK2
cdc25A
Cyclin A
Cyclin A

BOX 3-12 VECTORS

VIRAL VECTORS

Retroviruses – e.g. lentiviruses (human immunodeficiency virus; simian immunodeficiency virus). Although RNA viruses, they possess reverse transcriptase so can generate complementary DNA.

Adenoviruses – have been engineered to remove replicatory ability and also antigenic coats, particularly to reduce the immune response to these vectors.

Adeno-associated viruses (AAV) – a single strand of DNA virus without viral genes and therefore reduced immunogenicity. They only accommodate smaller DNA inserts but do establish long-term expression.

NON-VIRAL VECTORS

Liposomes – spherical lipid bilayer vesicles. No limit to size of DNA to be carried but efficiency of transfer is low.

BOX 3-13 NOT ALL GENE THERAPY REQUIRES REPLACEMENT OF THE GENE: *OPTOGENETICS*

Recent developments to try and restore vision in the degenerate retina utilizes viral transduction of, for example, channel rhodopsins to excite neurones, creating a rudimentary but significant and potential functional non-photoreceptor voltage-sensitive system to create artificial vision. Such systems could be coupled with a functional prosthetic (optobionics).

coding DNA is engineered to be flanked by regulatory sequences to ensure high expression dependent upon the tissue. By whatever method, the aim is to insert the DNA into the chromosome of the cell. To assist entry of DNA into the cell, *vectors* are used (Box 3-12). These are currently attenuated viruses that have been engineered so that the replication and disease-causing components have been removed. Even with this manipulation they retain an efficient ability to enter cells – so-called *transduction* (Box 3-13).

GENES, CELL DIFFERENTIATION AND CELL-BASED THERAPIES

Stem cells have the capacity to self-renew, proliferate and have no limitation to their potential differentiation. Progenitor cells can divide but have restricted differentiation potential. In the adult vertebrate central nervous system, neural progenitor cells (NPC) have been identified and shown to generate neurones and glia. Developments from molecular and cellular understanding of cell development as well as understanding the control of gene expression during cell differentiation and tissue organization in rodents has opened opportunities to develop cell-based therapies for degenerative disorders. Cepko's fate mapping of retinal cells in rodents determined the timing of differentiation from which many researchers since have ascribed specificities of genes, transcription factors and growth factors for cell fate determination (Fig. 3-9). During development there are two waves of progenitor cells that exit their cell cycle and commit to their post-mitotic mature cell fate. Together with increasing understanding of pivotal gene expression during cell differentiation and retinal development, this allows recapitulation experimentally to generate cell sources for cell transplantation.

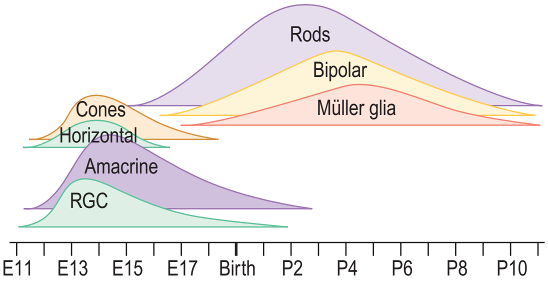

FIGURE 3-9 Two waves of retinal cell-type differentiation. The curves depict the relative number of retinal progenitors that exit the cell cycle and commit to a specific fate over time. Retinal ganglion cells (RGCs) appear at E11 and peak at E13.5, as do amacrine cells, horizontal cells, and cone photoreceptor cells. The other cell types, bipolar cells, Müller glia and rod photoreceptor cells, appear later as indicated. (Adapted from Cepko et al, 1996 and Wang et al, 2002.)

How do you generate cells for retinal reconstruction? The opportunities come from several sources. First embryonic stem cells (ES cells), which are derived from the inner cell mass of the blastocyst, can be conditioned in culture to generate stem cells that can be driven toward neural differentiation and then photoreceptors. Second, the recent Nobel Prize-winning creation of inducible pluripotential stem cells (iPS cells) from non-pluriopotential cells (including adult somatic cells such as fibroblasts) creates great possibilities to derive cells from the host cells (Fig. 3-10).

However, the developing and the postnatal vertebrate animal retina contain NPC, which divide, generate *neurospheres* and undergo neuronal and glial differentiation (Fig. 3-11). In humans, central nervous system-derived NPC have been successfully cultured

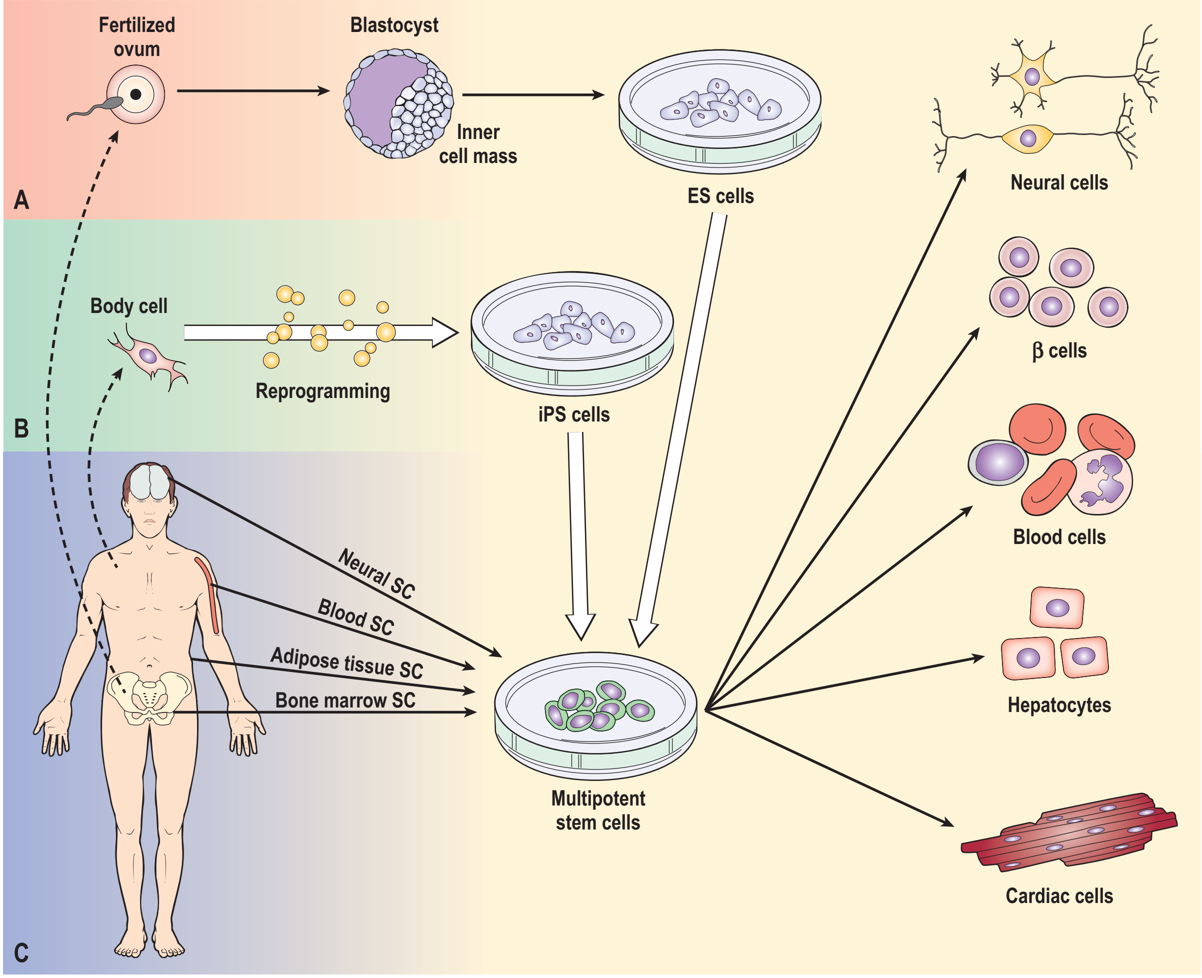
Fertilized ovum
Blastocyst
Inner cell mass
ES cells
A
Body cell
Reprogramming
iPS cells
B
Neural SC
Blood SC
Adipose tissue SC
Bone marrow SC
Multipotent stem cells
C
Neural cells
β cells
Blood cells
Hepatocytes
Cardiac cells

FIGURE 3-10 Pluripotential stem cell sources for delivery of cell based therapies. (A) Generation of embryonic stem (ES) cells. (B) Generation of inducible pluripotent stem cells (iPS). (C) Harvesting of lineage precursor cells. (Reproduced from Power and Rasko, 2011.)

from the adult brain and more recently adult retina; and also NPC have been isolated from the human retina while it is still immature and undergoing development. One phenotypic marker of stem cells and NPC is nestin, which is an intermediate filament. The adult human retina contains nestin-positive neuronal and glial cells, as does epiretinal scar tissue. This evidence suggests that NPC exist throughout life or are inducible in the human retina. Human retinal NPC are currently under extensive investigation to look at their capacity to renew damaged retina via, for example, transplantation with potentially gene-modifying cells that have undergone gene therapy before administration. To date the research has determined an ability to create retinal pigment epithelial cells from human ES cells, which have now been developed and are currently utilized in several clinical trials for age-related macular degeneration (see www.clinicaltrials.gov). Alongside this, there are great advances in the ability, at least experimentally, to develop ways of transplanting photoreceptor precursors that are able to integrate into the retina and develop synapses and generate action potential.

With all the developments discussed, we can see that there are many exciting current and potential ways forward for the treatment of inherited retinal

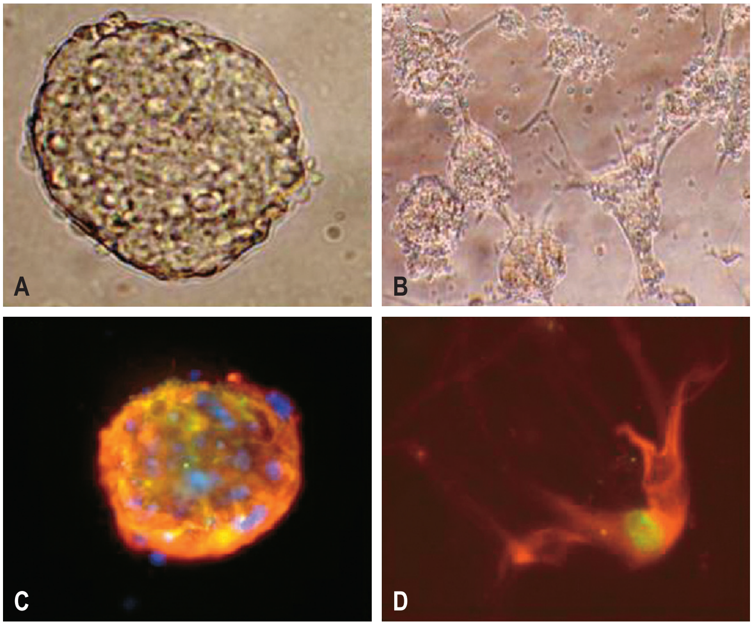

FIGURE 3-11 Neurospheres generated from cell suspensions derived from the adult human retina and pars plana differentiate after dividing into neurones. Phase contrast photography showing (**A**) free-floating neurosphere at 3 weeks, which may attach to a fibronectin-coated coverslip (**B**). Neurospheres were generated at a rate of one neurosphere for every 200 × 10^4 vital cells in the retinal cell suspension. Neurospheres contained nestin-positive (red) cells, some of which colabelled with green GFAP (**C**). Primary and secondary (passaged) neurospheres were exposed to BrdU at 1 week. After dissociation, primary neurospheres incubated in the presence of BrdU (green nucleus) express neurofilament M (red) in monolayer cultures (**D**). GFAP, glial fibrillary acid protein; BrdU, bromodeoxyuridine. (Figure courtesy of Dr Eric Mayer and Dr Debbie Carter, Academic Unit of Ophthalmology, University of Bristol, UK.)

degeneration. Figure 3-12 shows possible routes for intervention, depending on the stage at which treatment may be required.

Molecular genetics and ophthalmology

With the advent of modern molecular biological techniques, which have and will improve the understanding of the pathology of disease at both the cellular and molecular levels, considerable progress has been made, in particular in the study of X-linked disorders, several of which are of interest to the ophthalmologist.

The problem of how to find and identify an abnormal gene among the millions of genes in the human genome has already been discussed. To summarize briefly, the problem can be tackled principally by two approaches. The first is to find a gene marker in the human genome that is close to the causative gene defect. This approach, as described above, is the basis of genetic linkage and requires the analysis of DNA from affected families. Using the genetic marker, linkage analysis will locate the gene to loci on the same chromosome. Fragments of DNA from that region of the chromosome are cloned and sequenced to identify mutations in the gene.

The second approach is to identify candidate genes specifically expressed in the tissue, or which are known to code for proteins important in that tissue. Patients are then screened for mutation in these genes. By this method the discovery of mutant genes gives insight into the underlying pathophysiology of the condition, if the functions of the proteins from the genes studied are already known. This method of analysis is known as *candidate gene analysis*. Both of these methods have been used to study ophthalmic disorders and enhanced in current studies by the new platforms of next-generation sequencing discussed above.

The ability to detect mutations has led to several consortia for 'new gene detection' but also genetic testing. The National Eye Institute in the USA has developed the eyeGENE® network with list of genes currently available for testing (http://www.nei.nih.gov/eyegene/genes_eyegene.asp) (Table 3-2).

THE X CHROMOSOME

Historically, many molecular genetic studies have been based on an investigation of the X chromosome and of X-linked disorders. Mapping of disorders such as red–green colour blindness (Xq22–28), blue cone

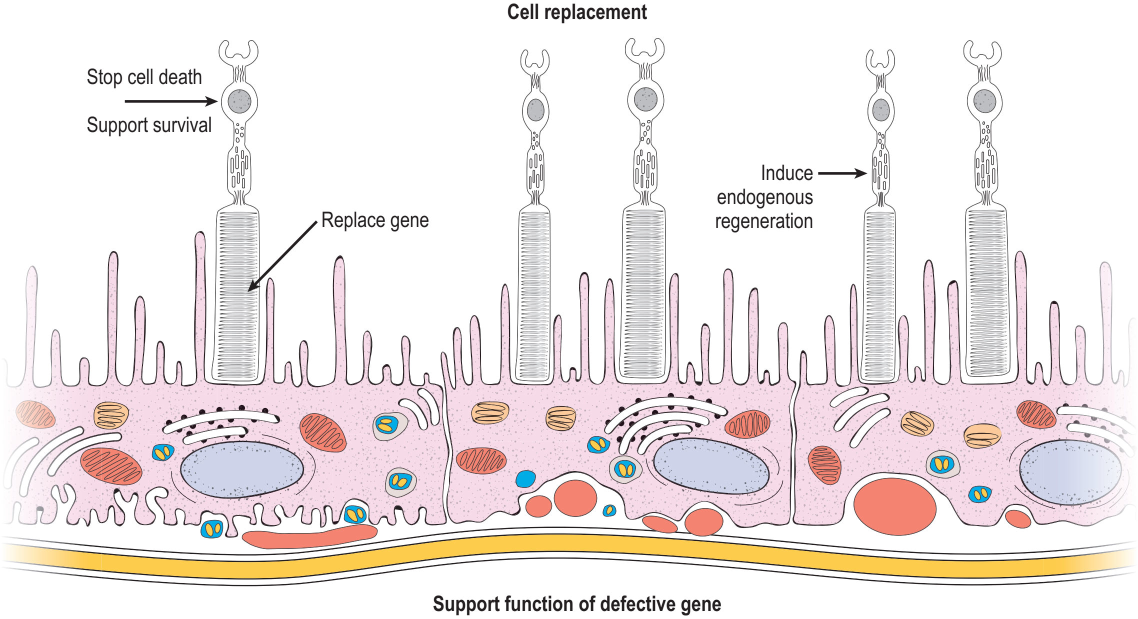
Cell replacement
Stop cell death
Support survival
Replace gene
Induce endogenous regeneration
Support function of defective gene

FIGURE 3-12 The schematic represents possible approaches to either maintain function or restore function in face of photoreceptor degeneration. The approach is derived from information that from early in diagnosis therapeutic support may be developed to prevent the cell death that occurs via apoptosis (see the explanation of the exponential cell death theory where the risk of cell death remains constant throughout life in Clarke et al., 2000). Similarly it is assumed that until the cells die they receive survival support via growth factors to maintain function and therefore pharmacological approaches may support function for longer. Ultimately, gene therapy can be applied to replace the mutant, defective or lost gene. However, late in the disease process, when loss of cells via apoptosis has occurred, an approach using cell replacement via stem cells (see Fig. 3-10) or one stimulating endogenous retinal precursor cells to replace cells are being pursued. Function may be partially restored byoptogenetics (using gene therapy approaches to deliver channel rhodopsins) to substitute photoreceptor function.

monochromacy (Xq28) and congenital stationary night blindness (Xp11) have come about with the use of RFLP and recombinant DNA technology.

RETINITIS PIGMENTOSA (Fig. 3-13)

Retinitis pigmentosa is the term used to describe a heterogeneous group of rod–cone dystrophies that have a variety of clinical appearances by virtue of varying inheritance patterns. Studies have documented the frequencies of the various modes of inheritance. Approximately 43% are inherited by autosomal transmission, 20% by autosomal recessive transmission, and between 8% and 25% by X-linked recessive transmission. Figures again vary, but approximately 20–25% of cases of retinitis pigmentosa appear to be isolated, or at least have unidentifiable patterns of inheritance. As the disease may be classified according to Mendelian inheritance patterns, the possibility of single-gene defects is high and this has prompted an energetic search for the isolation of such genes.

X-linked retinitis pigmentosa

Patients with this form of retinitis pigmentosa present with symptoms of night blindness from childhood; they have progressive constriction of visual fields and loss of vision in mid-life, although the severity of the disease does vary. The gene for this condition has been mapped to Xp11.3 (short arm of the X chromosome). Further evidence also maps the X-linked gene to Xp21, particularly in families where the female carrier demonstrates the golden tapetoretinal reflex. Recently these gene loci have been designated *RP2* (Xp11.3) and *RP3* (Xp21.1). Currently, probes are available for identifying both loci and may be used for prenatal diagnosis and genetic counselling.

TABLE 3-2 Ocular disease-causing genes

Diagnoses eligible for inclusion	Genes offered for testing
Achromatopsia	CNGA3, CNGB3
Albinism	*Recessive*: TYR, OCA2, TYRP1, SLC45A2 *X-linked*: GPR143 (OA1)
Aniridia and other developmental eye anomalies	PAX6, WT1#, DCDC1#, ELP4# (# del/dup testing only)
Axenfeld–Rieger syndrome	PITX2, FOXC1
Best disease	BEST1
Bietti crystalline corneoretinal dystrophy	CYP4V2
Choroideraemia	CHM
Chronic progressive external ophthalmoplegia (CPEO)	POLG
Cone rod dystrophy	ABCA4, RPGR, CRX, GUCY2D *(codon R838)*
Congenital cranial dysinnervation diseases (CCDD)	KIF21A, CHN1, SALL4, TUBB3, HOXA1, PHO2A, ROBO3, HOXB1
Familial/congenital nystagmus (X-linked cases only)	FRMD7
Congenital stationary night blindness/Oguchi disease	GPR179, RHO, NYX, TRPM1, SAG
Corneal dystrophy	TGFBI, KRT3, KRT12
Doyne honeycomb dystrophy	EFEMP1
Familial exudative vitreal retinopathy	FZD4, LRP5, NDP, TSPAN12
Fundus albipunctatus/Bothnia retinal dystrophy	RDH5, RLBP1
Glaucoma (juvenile open angle and congenital only)	CYP1B1, OPTN, MYOC
Hermansky–Pudlak syndrome	HPS1 and HPS3
Juvenile X-linked retinoschisis	RS1
Kearns–Sayre syndrome (KSS), mitochondrial encephalopathy, lactic acidosis and stroke-like episodes (MELAS), myoclonis epilepsy associated with ragged-red fibres (MERRF), neuropathy, ataxia and retinitis pigmentosa (NARP)	Mitochondrial gene panel
Leber's hereditary optic neuropathy (LHON)	LHON panel (MT-ND4, MT-ND1, MT-ND6/mutations 11778G>A, 3460G>A, 14484T>C and 14459G>A)
Lowe syndrome	OCRL
Microphthalmia and anophthalmia	RAX, SOX2, OTX2, VSX2, STRA6 and SIX6del/dup analysis
Neurodegeneration with brain iron accumulation (NBIA)	FA2H, MMIN, PANK2, PLA2G6
Occult macular dystrophy	RP1L1 (R45W)
Optic atrophy, dominant	OPA1, OPA3
Papillorenal syndrome	PAX2
Pattern dystrophy	PRPH2
Retinitis pigmentosa (RP) and retinal degenerations	*Dominant*: (panel including RHO, PRPH2, RP1, IMPDH1, PRPF8, NR2E3, PRPF3, TOPORS, PRPF31, RP1, KLHL7, SNRPN200), CA4, CRB1, CTRP5 *X-linked*: RPGR, RP2
Retinoblastoma^ (12/50 enrolled for 2013)	RB1
Sorsby fundus dystrophy	TIMP3
Stargardt disease	ABCA4, ELOVL4, RDS
Stickler syndrome^ (6/50 enrolled for 2013)	COL2A1
Usher syndrome^ (limit exceeded for 2013)	Usher panel (CDH23, CLRN1, DFNB31 (WHRN), GPR98, MYO7A, PCDH15, USH1C, USH1G, USH2A)

The genes highlighted in this table emphasise the increasing number of genes that may be tested to detect or diagnose disorders. Whilst not a complete set and moreover, what genes are included for testing vary world-wide, the advent of exome and whole genome deep sequencing illuminates continually other disease-causing genes.

(Courtesy National Eye Institute, National Institutes of Health (NEI/NIH). Adapted from http://www.nei.nih.gov/eyegene/genes_eyegene.asp.)

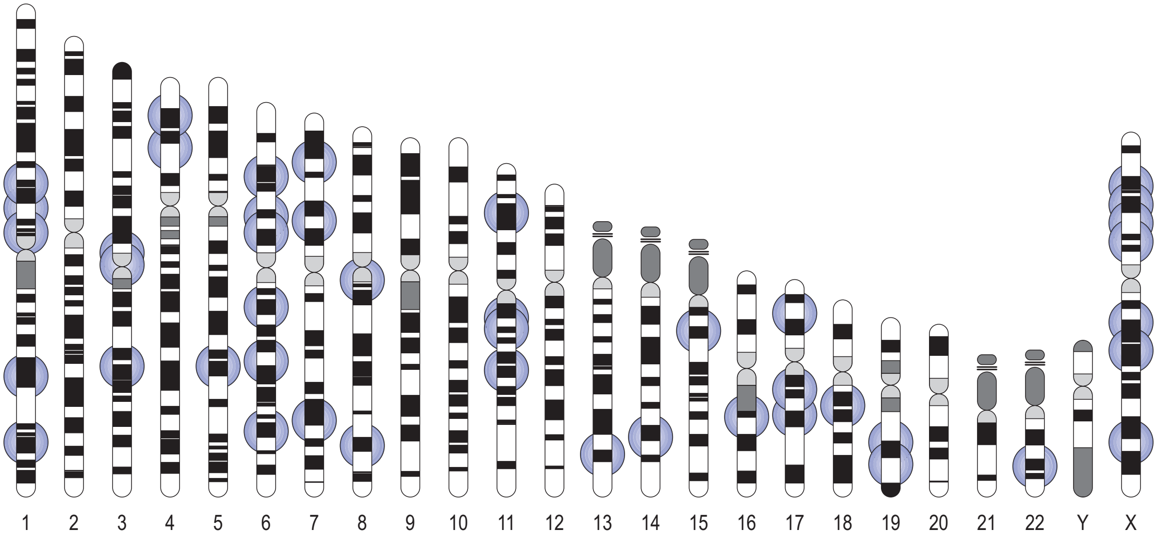

FIGURE 3-13 A schematic representation of the human chromosomes, showing the various loci implicated in retinal disorders. (Figure courtesy of Dr Z. Mohamed, PhD thesis, University of Aberdeen, UK.)

Autosomal dominant retinitis pigmentosa

Several mutations have been found in candidate genes in up to 30% of patients with autosomal dominant retinitis pigmentosa. Mutations in two genes have been studied in particular. These are the rhodopsin gene on chromosome 3q (accounting for 20% of all cases) and the peripherin gene on chromosome 6p. The rhodopsin molecule, composed of 348 amino acids, exists as a seven-loop transmembrane protein in the rod outer segment. The C-terminus of the protein is in the cytoplasm and the N-terminus of rhodopsin is in the intradiscal space. Throughout the protein, several regions are affected by mutations, which fall into three main groups: (1) mutations affecting amino acids in the intradiscal space; (2) mutations affecting amino acids in the transmembrane domain; and (3) mutations affecting amino acids in the cytoplasm. Most of these mutations probably destroy the three-dimensional (tertiary) conformation of the protein and in some way affect protein function. To date, over 150 different mutations have been reported, the majority of which are point mutations, although deletions have also been discovered.

The peripherin gene codes for the retinal degeneration show (RDS) protein found in rodents (see Ch. 9, p. 515), which is a component of the rod outer segment disk membranes. More than 20 mutations have been discovered in this gene in association with autosomal dominant retinitis pigmentosa and other retinopathies, for example retinitis pigmentosa albescens and hereditary maculopathies. Recently, other genes have been identified in autosomal dominant retinitis pigmentosa, and these include one at the centromere of chromosome 8 and both the long and short arms of chromosome 7.

Autosomal recessive forms of retinitis pigmentosa have been less well studied; to date, only one gene mutation has been identified in a rhodopsin gene. Recently there has been a report that certain patients with autosomal dominant retinitis pigmentosa have defects in the gene coding for cyclic guanosine monophosphate phosphodiesterase (see Ch. 4, p. 261) (Box 3-14).

CHOROIDERAEMIA

Choroideraemia presents in childhood with night blindness, leading to eventual loss of vision in later life. Fundal changes show granular pigmentary changes early in the course of the disease and choroidal atrophy in the late stages. Like X-linked recessive retinitis pigmentosa with the *RP3* gene defect, carrier females also demonstrate patchy non-progressive

BOX 3-14 PHOTORECEPTOR DEGENERATIONS AND CILIOPATHIES

Examples of photoreceptor degeneration are via genetic mutations that lead to cell functional deficits. It was thought that the rate of cell death or dysfunction increases over time. However, mathematical modelling has determined that it is not a cumulative damage that occurs but that the risk of cell death remains constant and the genetic mutation results in random cell death over time – 'one-hit' biochemical model. Such examples of neuronal degeneration are photoreceptor degenerations.

- Ciliopathies define a dysfunction of cilia function in retinal photoreceptors.
- Mutations, for example, that may occur in genes such as *CEP290* and *RPGR* give rise to photoreceptor dysfunction and degeneration as well as being ubiquitous in many cells for centrosome function.
- Retinitis pigmentosa related to ciliary dysfunction can be an isolated feature or a part of a syndrome such as Bardet–Biedl syndrome (BBS).

equatorial pigmentary changes, as a consequence of lyonization. The gene locus for choroideraemia has been identified at the locus Xq21. Several point mutations, deletions and RFLPs resulting from aberrant splicing at the exon–intron junction have been associated with the locus and the disease is due to deletion to *Rab1* gene.

NORRIE'S DISEASE

Norrie's disease affects males from birth. Features include bilateral congenital blindness secondary to retinal dysplasia and retinal vascular anomalies, impaired hearing, mental retardation, retinal detachment and cataract. This condition is similar to another congenital blinding condition, namely familial exudative vitreoretinopathy, which also shows abnormal vascular development, particularly of the retinal vessels. The gene locus for Norrie's disease has been identified as Xp11.1, Xp11.3. Similar gene loci have been suggested for familial exudative vitreoretinopathy, suggesting that these two disorders are related by expressing different mutations at the same gene loci, producing the two phenotypes. Mutations and deletions have been demonstrated at the gene loci for Norrie's disease. The gene's official symbol is *NDP*, a gene for a protein called norrin integral to specialization of retinal cells (Table 3-3).

TABLE 3-3 **Subretinal delivery of gene therapy from identification of gene mutations**

Disease	Vector-gene
Leber's congenital amaurosis due to RPE65 mutation	AAV2-hRPE65v2
Choroideraemia	rAAV2-Rep1
Usher type 1B	Ushstat®

RETINOBLASTOMA

Retinoblastoma is a tumour of primitive photoreceptor cells. It is the commonest ocular malignancy of childhood with a prevalence of approximately 1 in 20 000, showing an equal sex distribution. Approximately 40% of cases are inherited. Unilateral tumour, however, is nearly always sporadic, with no family history. Bilateral cases usually have a strong family history of retinoblastoma and are inherited by autosomal transmission. These patients are also at greater risk in later life of developing osteosarcoma. It has been shown that retinoblastoma may be secondary to a deletion of 13q14 in about 4% of patients. The gene locus that encodes the enzyme d-esterase is also situated on chromosome 13, and patients with retinoblastoma show a much reduced level of this enzyme, indicating that the d-esterase and retinoblastoma loci are closely linked. Knudson proposed a hypothesis to explain the fact that inherited cases are usually bilateral, multifocal and of early onset, whereas sporadic cases are unilateral and solitary. The 'two-hit hypothesis' proposes two mutational events in inherited retinoblastoma. The first mutation is present in the germinal cell and would therefore be present in every cell. The second 'somatic' mutation must occur to induce tumour growth in cells with the initial mutation by releasing suppression or regulation of the retinoblast. Two somatic mutations to one retinoblast must occur in sporadic cases of retinoblastoma and are therefore likely to be solitary and unilateral. This retinoblastoma gene has therefore been proposed as a *tumour suppressor gene*, whose presence in normal retinoblasts prevents uncontrolled mitosis. The DNA sequence of the retinoblastoma gene has now been identified, and the surrounding complex gene locus identified as a *retinoblastoma predisposition gene*, which is structurally altered in patients with retinoblastoma, although its true function is unknown. As mentioned above, patients who survive retinoblastoma have an increased risk of developing osteosarcoma and, interestingly, in isolated patients with osteosarcoma the retinoblastoma gene has been found to be deleted.

ALBINISM

Albinism is a cause of poor visual acuity and nystagmus in children and is divided broadly into two groups: oculocutaneous and ocular albinism. Patients with oculocutaneous albinism may be further differentiated into tyrosinase-producing (tyrosinase-positive) and tyrosinase-non-producing (tyrosinase-negative) groups, shown in children over the age of 4 years by hair bulb incubation in tyrosinase solution (Fig. 3-14).

Oculocutaneous albinism

This form of albinism is inherited in an autosomal recessive Mendelian fashion and within the tyrosinase-negative group gives rise to severe disease with profound visual loss, photophobia and nystagmus, as well as the classic features of iris transillumination, absent fundal pigmentation and absent foveal reflex. Most of the optic nerve fibres cross at the chiasma (90%), and further neuronal disorganization occurs within the lateral geniculate body.

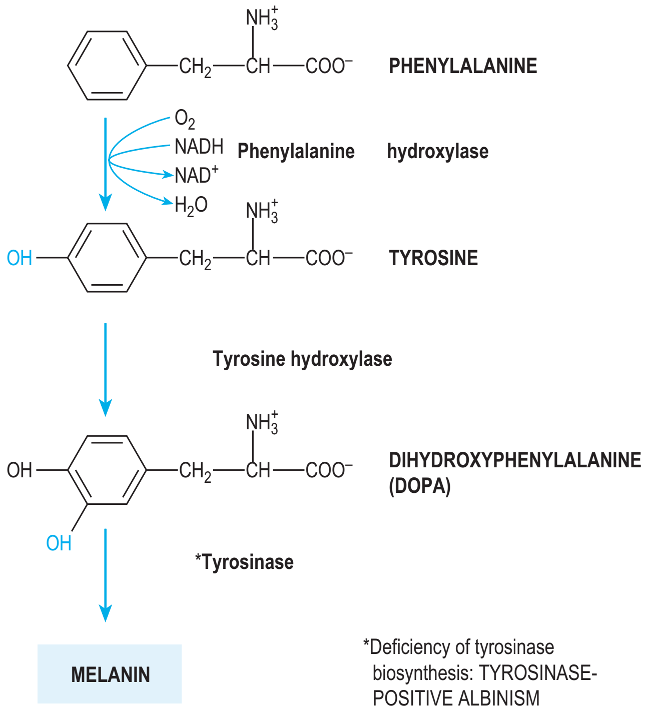

*Deficiency of tyrosinase biosynthesis: TYROSINASE-POSITIVE ALBINISM

FIGURE 3-14 Enzyme defects in albinism.

Ocular albinism

Ocular albinism is present when most of the hypopigmentation (hypomelanosis) is confined to the ocular structures. Ocular albinism may be inherited in an autosomal recessive (Nettleship–Falls syndrome) or an X-linked recessive pattern. X-linked recessive ocular albinism gives rise to ocular disease of moderate severity, with a prevalence of approximately 1 in 50 000. Affected males have reduced visual acuity, nystagmus, strabismus and iris translucency. Fundal examination shows classic hypopigmentation and foveal hypoplasia. In this condition giant melanosomes are present in the retinal pigment epithelium and are also found in skin biopsies; they are similar to the melanosome aggregates found in one of the albinoid syndromes, Chédiak–Higashi syndrome (associated with phagocytic dysfunction). Carrier females, whose visual acuity is normal, also demonstrate iris transillumination, retinal pigment epithelial granularity and a preponderance of giant melanosomes in the skin. At-risk females, however, can pose a diagnostic problem, and accurate genetic counselling will be available only when future genetic diagnostic tests and techniques for identification of the candidate gene have been developed.

MYOTONIC DYSTROPHY

Myotonic dystrophy presents with progressive muscle weakness early in adult life. The condition is characterized by expressionless face, frontal balding, gonadal atrophy and myotonia when shaking hands. Patients often develop cataracts and may also develop a pigmentary retinopathy. Myotonic dystrophy is transmitted as an autosomal dominant trait with an incidence of 1 in 20 000. Gene loci for this disease have been localized with the use of RFLP and DNA probes to chromosome 19. However, because of the recent discovery of an unstable DNA mutation consisting of an increased number (more than 50) of a nucleotide triplet (CTG repeats), whose protein product is a member of the protein kinase family, family studies can confirm or exclude those at risk.

MITOCHONDRIAL INHERITANCE

As has been stated above, mitochondria contain specific circular DNA that replicates separately from the nuclear DNA and is inherited solely from maternal mitochondria. Recently, some inherited disorders have been identified as having a mitochondrial mode of transmission, because they do not follow classic Mendelian patterns of inheritance.

Leber's hereditary optic neuropathy

This condition is characterized by rapid onset of visual failure particularly in boys, but it may affect either sex. The result of the initial hyperaemic disk swelling and peripapillary telangiectasia is optic atrophy and visual failure. Mothers characteristically pass the disease to their sons, but sons never transmit it (i.e. there is no male-to-male transmission). A characteristic point mutation causing histidine to be inserted instead of arginine at the 340th amino acid of NADH in complex I of the respiratory chain has been demonstrated in patients with this type of optic neuropathy. Other point mutations in mitochondrial DNA have also been documented. However, there has been no explanation as to why males are predominantly affected in this disorder, which cannot be explained purely on the basis of a single mitochodrial gene defect. With the advent of mitochondrial DNA analysis, investigation of patients with optic neuropathy of uncertain aetiology can be carried out to determine whether Leber's neuropathy is the cause. Also, recent studies have shown that the genotype of the condition is associated with a variable phenotype, in that some families with specific gene mutations demonstrate recovery of vision in up to 50% of patients. In addition, other gene mutations are linked to Leber's hereditary optic neuropathy and are associated with generalized neurological abnormalities.

Other mitochondrial disorders

Other mitochondrial inherited disorders also affect the eye. These include the mitochondrial myopathies, of which the most documented is Kearns–Sayre syndrome. This syndrome occurs secondary to multiple point mutations within the mitochondrial genome, which in turn lead to multiple deletions of varying size. The heterogeneity of the mutations accounts for the variance of the clinical signs encountered, which include a pigmentary retinopathy and progressive myopathy, involving cardiac and proximal limb muscles as well as a progressive external ophthalmoplegia.

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

FURTHER READING

Cepko, C.L., Austin, C.P., Yang, X., Alexiades, M., Ezzeddine, D., 1996. Cell fate determination in the vertebrate retina. Proc. Natl. Acad. Sci. USA. 93(2), 589–595.

Chu, C.J., Barker, S.E., Dick, A.D., Ali, R.R., 2012. Gene therapy for noninfectious uveitis. Ocul. Immunol. Inflamm. 20, 394–405.

Clarke, G., Collins, R.A., Leavitt, B.R., Andrews, D.F., Hayden, M.R., Lumsden, C.J., et al., 2000. A one-hit model of cell death in inherited neuronal degenerations. Nature 406, 195–199.

Davey Smith, G., 2010. Mendelian randomization for strengthening causal inference in observational studies: application to gene × environment interactions. Perspectives on Psychological Science 5, 527–545.

Jacobson, S.G., Cideciyan, A.V., 2010. Treatment possibilities for retinitis pigmentosa. N. Engl. J. Med. 363, 1669–1671.

Klug, W., Cummings, M., Spencer, C., Palladino, M. Concepts of genetics plus mastering genetics with eText – Access Card Package. International Edition, tenth ed. 2011.

Lipinski, D.M., Thake, M., MacLaren, R.E., 2013. Clinical applications of retinal gene therapy. Prog. Retin. Eye Res. 32, 22–47.

MacLaren, R.E., Pearson, R.A., MacNeil, A., Douglas, R.H., Salt, T.E., Akimoto, M., et al., 2006. Retinal repair by transplantation of photoreceptor precursors. Nature 444, 203–207.

Petersen-Jones, S.M., Annear, M.J., Bartoe, J.T., Mowat, F.M., Barker, S.E., Smith, A.J., et al., 2012. Gene augmentation trials using the Rpe65-deficient dog: contributions towards development and refinement of human clinical trials. Adv. Exp. Med. Biol. 723, 177–182.

Power, C., Rasko, J.E., 2011. Promises and challenges of stem cell research for regenerative medicine. Ann. Intern. Med. 155, 706–713.

Snead, M.P., Yates, J.R., 1999. Clinical and molecular genetics of Stickler syndrome. J. Med. Genet. 36, 353–359.

Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676.

Wang, S.W., Mu, X., Bowers, W.J., Klein, W.H., 2002. Retinal ganglion cell differentiation in cultured mouse retinal explants. Methods 28(4), 448–456.

West, E.L., Pearson, R.A., MacLaren, R.E., Sowden, J.C., Ali, R.R., 2009. Cell transplantation strategies for retinal repair. Prog. Brain Res. 175, 3–21.

Chapter **4**

Biochemistry and cell biology

- Introduction
- Cells and tissues
- Biochemical pathways that affect ocular function
- The ocular surface
- Cornea and sclera
- Uveal tract
- Aqueous humour dynamics
- The lens
- The vitreous
- The retina
- The chemistry of the visual response
- Conclusion

Introduction

The eye is a miracle of self-organization. While many tissues such as muscle and kidney comprise predominantly a restricted set of more or less specialized cell types, almost all cell types are represented in the eye including secretory cells, neuronal cells, vascular cells, specialized fibroblasts, tissue myeloid cells, and supporting cells, and the matrices contain all of the molecular components found in other tissues. Uniquely, the cells and tissues have been customized for the eye.

Aristotle understood the phenomenon of self-organization as it applies to many organic and inorganic systems as 'in the case of all things … in which the totality is not, as it were a mere heap, but the whole is something besides the parts, there is a cause' (Sasai, 2013). That the process of self-organization as it applies to the eye results in an organ specialized for the transmission, reception and conversion of light energy into cellular signals remains a remarkable feat of differentiation and development. Even more remarkable is the observation that the forces leading to this event appear to be intrinsic to the cells, since optic cup and eye morphogenesis can be induced in embryonic stem cells *in vitro* given the right conditions (Fig. 4-1).

The unique feature of cells and tissues in the eye is that they are organized for the transmission, reception and conversion of light energy into cellular signals. Cells respond to stimuli in a remarkably similar manner. What differentiates one cell from another are the stimuli each cell responds to and the mechanisms it uses to respond. Cells have the genetic potential to express any type of receptor but their unique specialization is down to the limited set of membrane receptors they express. Through these receptors, cells respond to a specific stimulus by activating an intracellular second messenger system that has a limited generic range, i.e. the same set of signalling molecules is frequently activated by a wide range of ligand–receptor interactions: the specificity lies in the ligand binding. This produces a programmed response in the cell, resulting in an effect, e.g. aqueous secretion from acinar lacrimal gland cells after stimulation by adrenaline (epinephrine), an ocular muscle action potential after neurotransmitter release from nerve endings, or rhodopsin activation by a photon of light. Remarkably, the process of signal induction and transmission is based on a very limited set of biochemical reactions often involving an ATP-energy-driven mechanism, such as the addition or subtraction of a phosphate group to the signalling molecule, respectively via enzyme activity provided by a kinase or a phosphatase.

Cells are organized, singly or in groups (tissues), to receive information from the environment (via membrane receptors), to signal this information to

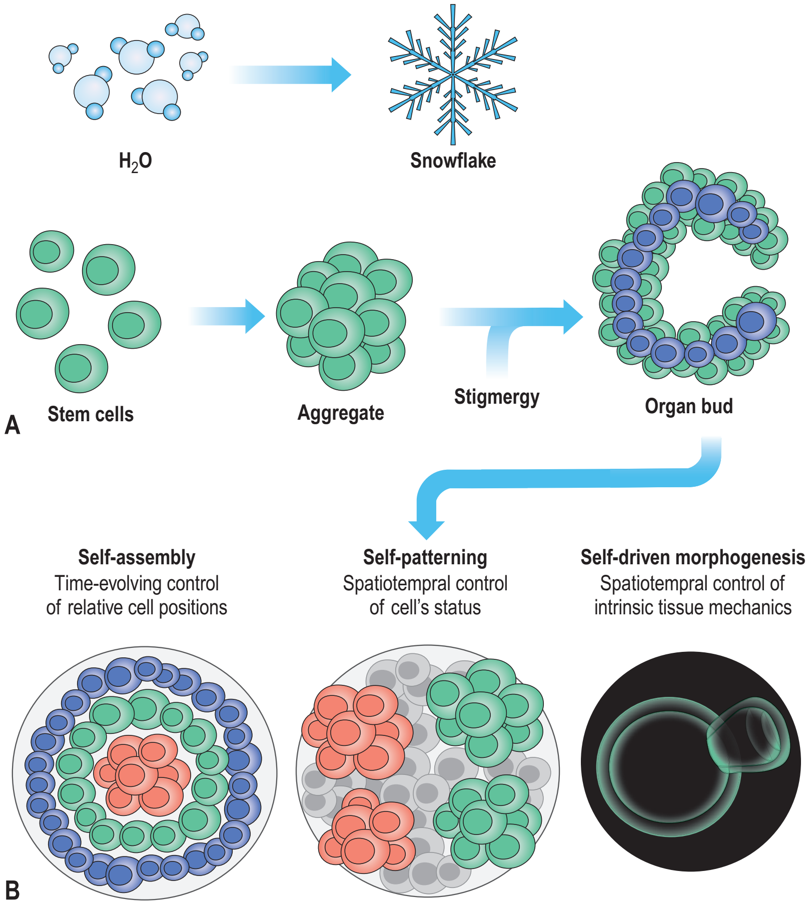
H_2O
Snowflake
A
Stem cells
Aggregate
Stigmergy
Organ bud
Self-assembly
Time-evolving control of relative cell positions
Self-patterning
Spatiotempral control of cell's status
Self-driven morphogenesis
Spatiotempral control of intrinsic tissue mechanics
B

FIGURE 4-1 Self-assembly and self-organization is a general biological phenomenon. Individual particles aggregate as in water molecules transforming into snowflakes (**A**) or in stem cells which then develop by intrinsic forces into morphogenetically identifiable structures, a process that underpins embryogenesis (**B**). (Images from Sasai, 2013, with permission.)

the intracellular compartment (via signalling networks), to convert the message into cellular responses (gene activation and protein transcription) and to relay this information to the outside world (e.g. changes in cell behaviour, tissue function, secretion, etc.). Innumerable molecules and genes are involved in a single response by the cell through these networks, in which thousands of molecular interactions are connected through molecular 'hubs' (similar to hubs regulating aircraft traffic) (see next section). Cells may also respond simultaneously to several stimuli through several receptors, and controlling the flow of this information can be difficult. In recent years, this function seems to be attributed to that mysterious cellular constituent chromatin, whose histone tails may act as a signal storage and converter device, similar to digital–analog converter devices in the electronic industry. In this way accumulated repeated rapid ON–OFF signals over time can induce epigenetic and indeed phenotypic change in cells and tissues, which not only influence cellular functions as they age but also can be transmitted to the next generation.

Additional content available at https://expertconsult.inkling.com/.

These concepts have been derived from the vast amount of information made available through full genetic analysis of various organisms (genomics) and the use of novel methods of investigation including microarray technology and informatics.

Studies in mice have shown that epigenetic changes to the gene pool of the fetus can lead to damage to either the somatic cells (genes not transmitted) or the germ-line cells (genes passed on). If the damage is not sufficient to kill the fetus the first-generation progeny (F1) will only pass on their mutated genes to the next generation (F2) if the germ-line genes have been affected (eFig. 4-1). If, however, there is epigenetic germ-line damage and it is stably expressed in the adult, it may be passed on as part of the normal germ-line genes (i.e. is passed down several generations – transgenerational) unless reprogramming occurs in the embryo before implantation. Reprogramming can be viewed as a type of fail-safe repair mechanism.

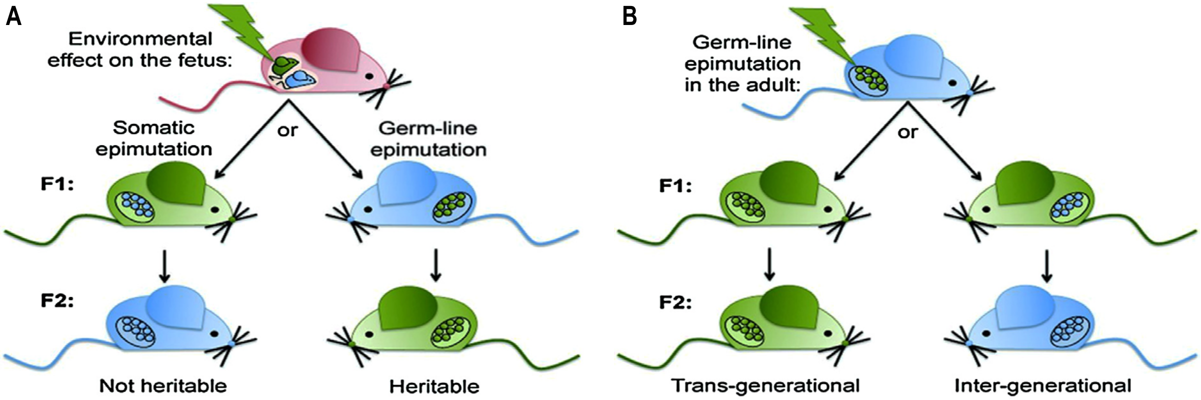
A
Environmental effect on the fetus:
Somatic epimutation
or
Germ-line epimutation
F1:
F2:
Not heritable
Heritable
B
Germ-line epimutation in the adult:
or
F1:
F2:
Trans-generational
Inter-generational

eFIGURE 4-1 Epigenetic inheritance could occur through several mechanisms: (**A**) the fetus (FI) might be affected by environmental factors such as chemicals or drugs which would lead to either somatic (blue cells) or germ-line (green cells) mutations. In this case only the somatic mutations are heritable. (**B**) If a germ-line mutation is inherited, it can either be inherited over many generations or the germ-line cells can be reprogrammed back to normal cells and the genetic defect only lasts for one or two generations (inter-generational). In the figure, pink and blue mice are normal, green mice have the genetic defect.

(From Stringer et al., 2013.)

THE CONTINUING ADVANCE OF ‘OMICS’

The development of systems biology, which is based on microarray techniques and high-throughput technology, has rendered the complexity of molecular interactions, such as those involved in signalling or in transcriptional regulation, amenable to analysis. This is the science of ‘omics’, a term applied to a body of work, knowledge or data, and includes genomics, transcriptomics, metabolomics, proteomics, signalomics and the microbiome. This has generated a vast amount of information leading to some further ‘omic’ subdivisions such as cancer genomics or toxicogenomics. Genomics examines the many genes which may be involved by increased or decreased expression, while transcriptomics studies the many transcription factors which may be activated or deactivated in any one cellular behaviour, such as cell division, and metabolomics investigates the many biochemical pathways which may be utilized, or not, in conversion of one molecular species to another in the process of energy generation and consumption. For the epigenetics researcher, there is even an epigenome, which provides information on DNA methylation, histone modifications and chromatin remodelling. Meanwhile, the microbiome is a term applied to the databank relating to gut commensals and has considerable influence on the immune system (see Ch. 7, p. 373). In several of these ‘-omes’, molecular networks are entrained, which in themselves reveal the extensive interdependence that one system has on another. In addition, the notion of central (‘hub’-based) molecular species without which the entire network would collapse allows a hierarchy of importance to be applied to molecules. This is demonstrated in the genetic mutagenesis studies in which certain molecules, such as transforming growth factor-β, are lethal to the embryo when deleted, while others, such as plasminogen activator inhibitor 1, barely alter the murine phenotype.

Signalling networks are a prime example of how cellular information is transmitted. It is now recognized that there are hundreds to thousands of signalling receptors in the cell membrane interacting with around 10 second messenger ‘hubs’ in large interacting intracellular networks of several thousand cellular proteins (see eFig. 4-2 and eBox 4-1).

Additional content available at https://expertconsult.inkling.com/.

Examples of typical general second messenger systems include:

- receptor tyrosine kinase-linked receptor systems (RTK)
- ion channels and pumps
- G protein-driven messengers
- seven-transmembrane loops.

And there are several more generic types of receptor.

Each of these may interact with other intracellular signalling systems and the signalling systems themselves may be ‘customized’ to respond selectively depending on the conditions (Fig. 4-2). In addition, each receptor may be represented many times in a single cell (it is estimated that there are between 5000 and 10 000 major histocompatibility (MHC) class II molecules on a dendritic cell), while several different

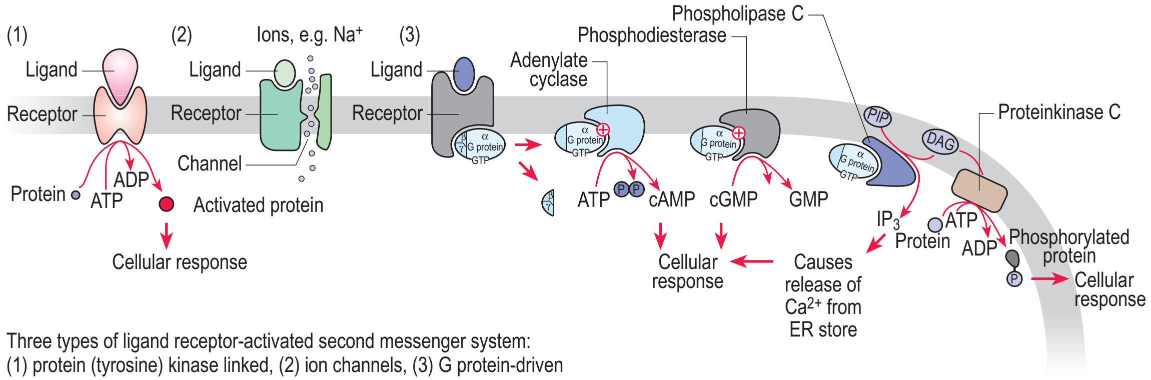
(1)
Ligand
Receptor
Protein
ADP
ATP
Activated protein
Cellular response
(2)
Ions, e.g. Na⁺
Ligand
Receptor
Channel
(3)
Ligand
Receptor
α
β
γ
G protein
GTP
Adenylate cyclase
Phosphodiesterase
Phospholipase C
ATP
P P
cAMP
cGMP
GMP
Cellular response
PIP
DAG
Proteinkinase C
IP3
Causes release of Ca2+ from ER store
Protein
ATP
ADP
Phosphorylated protein
P
Cellular response

Three types of ligand receptor-activated second messenger system:
(1) protein (tyrosine) kinase linked, (2) ion channels, (3) G protein-driven

FIGURE 4-2 Ligand–receptor-activated second messenger systems.

Many of the thousands of molecules and genes are involved in a single response by a cell through intracellular networks; in this process immense numbers of molecular interactions are connected through a system in which some molecules form many interactions with others in 'hubs' while other molecules make very few interactions and are on the periphery of the network (eFig. 4-2). These concepts are derived from the vast amount of information made available through full genomic analysis of various organisms and the use of novel methods of investigation including microarray technology and informatics. Furthermore, hub molecules are frequently transient in activity, one molecule acting as a hub during activation of one signalling pathway while the same molecule acts merely as a relay station during activity of another pathway. Information usually proceeds from 'outside in' to the cell, but on occasion information initiated outside the cell can be relayed back to extracellular targets (eFig. 4-2). Many of the cellular proteins may not be directly involved in signalling but may act as adaptors or amplify/diminish the overall response. In addition, other proteins act as 'chaperones' to protect proteins and signalling molecules for optimal function (eBox 4-1).

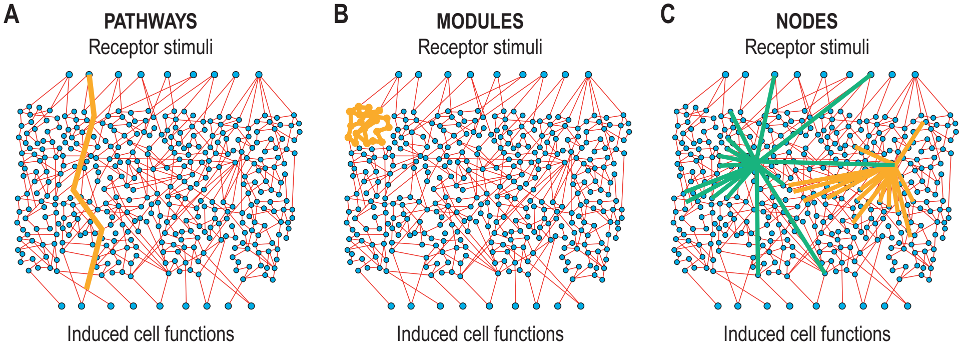

eFIGURE 4-2 Three concepts that are useful for describing signalling networks. Cell signalling is initiated by receptor stimuli. Each connection point reflects a signalling protein or second messenger, with lines indicating functional interactions. (**A**) Linear signalling pathways. (**B**) Modular structures within the network. (**C**) Nodes, which can be proteins or second messengers. Nodal points are regulated by many upstream events and/or regulate many downstream events. (From Meyer and Teruel, 2003, with permission from Elsevier.)

eBox 4-1

Molecular chaperones in the regulation of signalling:

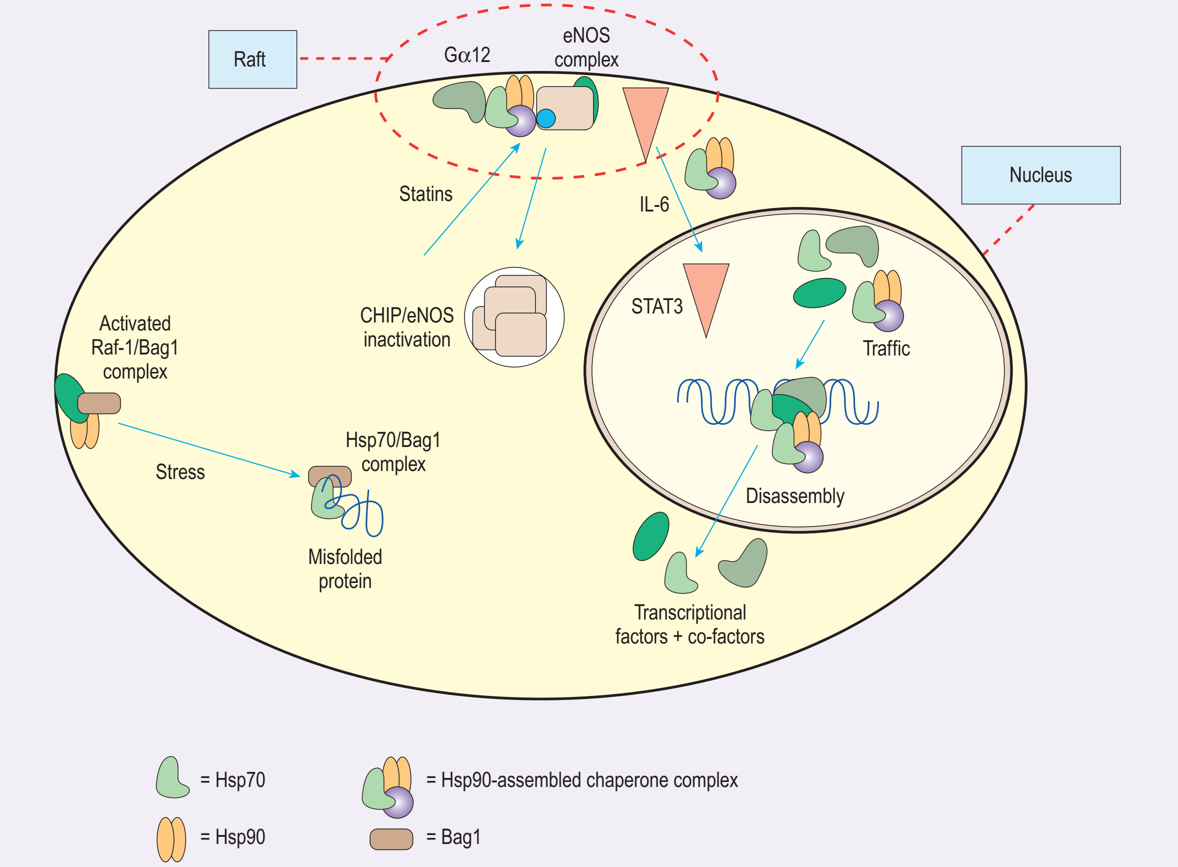
Raft
Gα12
eNOS complex
Nucleus
Statins
IL-6
STAT3
Traffic
CHIP/eNOS inactivation
Activated Raf-1/Bag1 complex
Hsp70/Bag1 complex
Stress
Misfolded protein
Disassembly
Transcriptional factors + co-factors
= Hsp70
= Hsp90-assembled chaperone complex
= Hsp90
= Bag1

Chaperone proteins are everywhere in the cell, shepherding essential functional proteins such as enzymes and signalling molecules to ensure their proper functioning. Many different types of chaperone exist and some come into their own in certain circumstances, such as heat-shock proteins (Hsp), while others are constitutively functioning, such as lens crystallins (Fig. 4-2). Chaperones play an essential role in the activation of protein kinases: for instance, Bag1, the co-chaperone of Hsp70, which can activate the Hsp90-dependent process. Stress is known to inhibit cell proliferation and sequestration of Bag1 may be how this occurs. Chaperones such as Bag1 and Hsp70 play an essential role in the maturation and activation of hundreds of protein kinases, regulating, for instance, cell proliferation in response to stress. Chaperones participate in raft-dependent signalling of molecules such as eNOS, G-proteins and STATs. Chaperones also help the subnuclear trafficking and disassembly of transcriptional factors and related complexes.

(From Söti et al., 2005, with permission from Elsevier.)

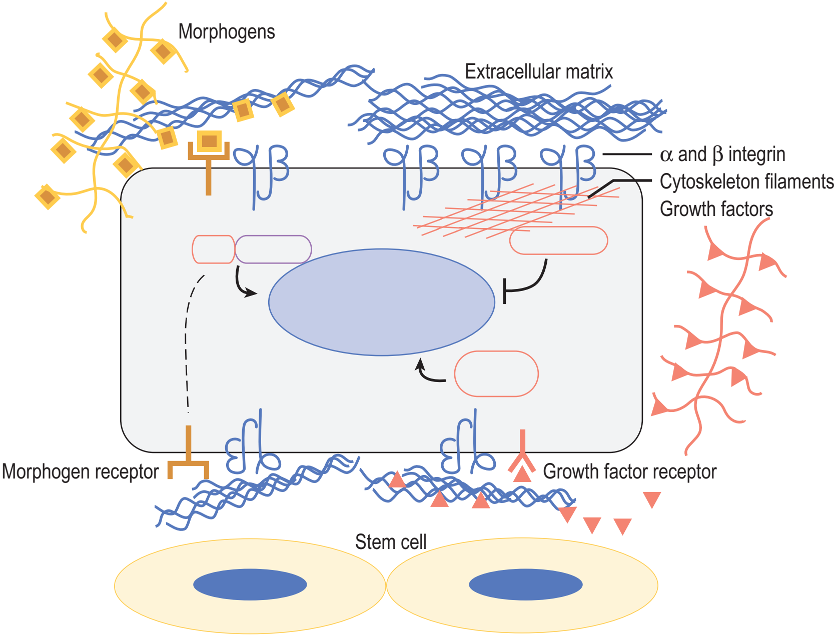

FIGURE 4-3 Suggested model for how matrix componenets (ECM), integrins and transmembrane receptors engage in 'cross-talk' in the stem cell niche. (1) The stem cell is physically anchored to the ECM, which activates growth factor receptors on the cell surface through the mediation of integrins; (2) the ECM also acts as a reservoir for growth factors and other mediators. Meanwhile the integrins also signal via the intracellular signalling molecules such as ERK 1/2m Akt, and SMADs which retain the stem cell properties of the cell; (3) the ECM integrin coupling process allows the cells to sense biomechanical stiffness in the surrounds and transmits this to the cytoskeleton. Growth factors and their receptors; morphogens and their receptors; extracellular matrix; α and β integrin subunits; cytoskeleton filaments; SC, stem cell. (From Brizzi et al., 2012.)

ligand–receptor pairs may act in clusters at the cell surface, as for instance in the T-cell receptor synapse (see Ch. 7, p. 427).

Signalling networks behave similarly to other biological networks, such as metabolic and gene transcriptional networks, and probably represent a basic biological organizing system. Three basic concepts underpin a signalling network: signalling pathways, signalling modules and signalling nodes (eFig. 4-2).

In addition, receptors exist not only on the cell surface but intracellularly on endosomes and in the nucleus and ligands may have to be transported intracellularly to interact with their receptor. Information transmitted via ligand/receptor interactions can also be bidirectional, as for instance in the regulation of stem cells in stem cell niches through integrin molecule binding to extracellular matrix proteins (Fig. 4-3).

Cells and tissues

THE CELL

General structure

Technology drives science and there is no better example of this than the discovery of the cell as the basic unit of living organisms by Anton van Leeuwenhoek using a compound optical microscope. The basic structure of the mammalian cell can be illustrated by the retinal pigmented epithelial (RPE) cell (Fig. 4-4) as it contains most of the recognized cellular structures and intracellular organelles. While all cells have the potential and machinery for mitosis and motility, many adult tissue cells such as the RPE cell are considered terminally differentiated, non-motile cells, except under pathological conditions. The RPE cell is an example of a bidirectional transporting epithelial cell with polarity, i.e. an apical surface with

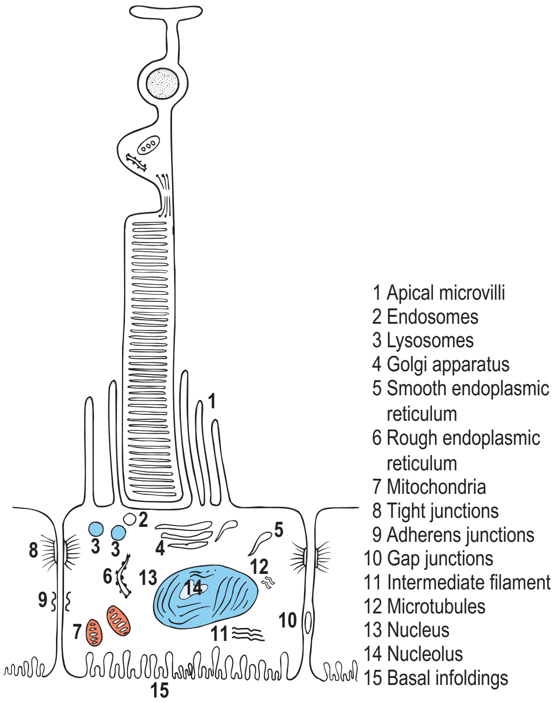

FIGURE 4-4 Diagram of RPE and photoreceptor cells. The photoreceptor outer segment lies in close apposition to the RPE cell, enclosed in a sheath of apical microvilli. The RPE cell is a terminally differentiated epithelial cell with several functions, one of which is to transport fluid across the cell towards the basal infoldings and into the choroid sink.

microvillous processes and a basal surface with numerous infoldings. The RPE is also an example of how the basic structure of the cell has been modified extensively, as in several types of specialized cells in the eye. Dysfunction of this critically important cell underlies the pathology of age-related macular degeneration (AMD), one of the commonest causes of blindness in developed nations (see Ch. 9, p. 513).

The plasma membrane

The plasma membrane, which surrounds all cells, is a selective two-way barrier to passive diffusion, which also has active transport mechanisms subserved by specialized proteins (for instance ion channels, pumps and suspended transporters) floating in a lipid bilayer, composed of phosphoglycerides, sphingolipids and sterols, which forms spontaneously due to its content of phospholipids. The barrier function of the plasma membrane applies even to ions and is the basis of the electrical potential that occurs across cell membranes and which is energy-driven, for instance by which is energy-driven, for instance by ATPase enzymes. Other proteins are also suspended in the plasma membrane, such as receptors for hormones, neurotransmitters, viruses and other cells. Many of these receptors have a three-part structure with an extracellular, variably sized, component, a transmembrane component, and a short intracellular section coupled to the second messenger system. The plasma membrane is a variably thick lipid bilayer 'crowded' with many membrane proteins whose functions not only serve transmembrane traffic via receptors and transporters but also have a direct role to play in the physiology and indeed the survival of the cell: this is in part determined by the types of motion and tension which different membrane proteins exert (Box 4-1).

Many of these proteins are held in patchy distributions within the lipid bilayer, termed lipid (micro) domains, which themselves have variable constitution. The lipid domains also move within the membrane like rafts, and indeed if large enough are known as lipid rafts. Other microdomains decorate the cell surface in the shape of irregularly pitted invaginations and evaginations such as specialized structures for endocytosis (clathrin-coated pits) and caveolae, embedded in a glycoprotein-rich matrix (glycocalyx). These are also involved in a plethora of functions including cell signalling, protein trafficking, cell movement, waste disposal (exocytosis) and even cell survival (Fig. 4-5). Microdomains are frequently detergent-resistant and usually contain a specific protein such as caveolin in caveolae, or several proteins, as occur in lipid rafts, areas specialized for specific functions such as the immunological synapse in antigen-presenting cells (see Ch. 7, p. 421). Other microdomains include tiny domains (nanodomains), which contain GTP-binding protein (inhibitory) (GPI)-anchored proteins (important in some types of signalling) and glycosphingolipids, transient confined zones of varying size, and small regions composed of more fluid lipids.

Cells of the nervous system are especially rich in lipid rafts, which contain a high content of cholesterol, and sphingolipid rafts appear to have an organizational function either as discrete functional elements in which both the ligand and receptor are present, or

BOX 4-1 MEMBRANE PROTEINS HAVE MANY WAYS OF MOVING

Free diffusion (**A**) of proteins in the lipid bilayer allows random or directional movement depending on external triggers. Many proteins are, however, anchored or tethered either to cytoskeletal proteins or to extracellular matrix proteins (**B**). Examples include GPI-tethered proteins. Some proteins are only transiently tethered during a particular function, for instance in the immunological synapse of immune cell activation (**C**). Some of the proteins can escape from the synapse or corral and become internalized. In contrast some proteins have limited movement and are confined to a region of the plasma membrane, as in confined diffusion (**D**).

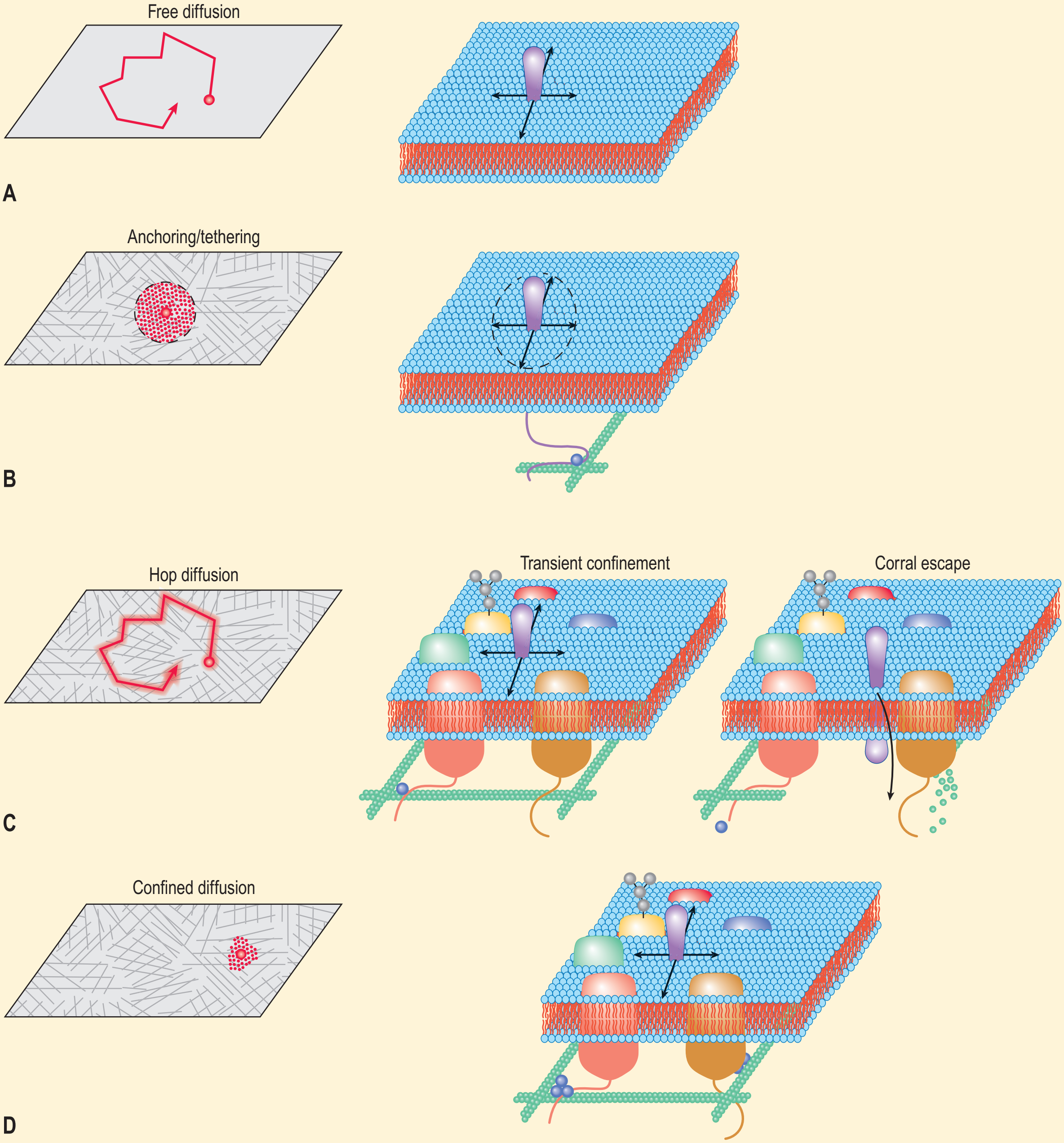
Free diffusion
A
Anchoring/tethering
B
Hop diffusion
Transient confinement
Corral escape
C
Confined diffusion
D

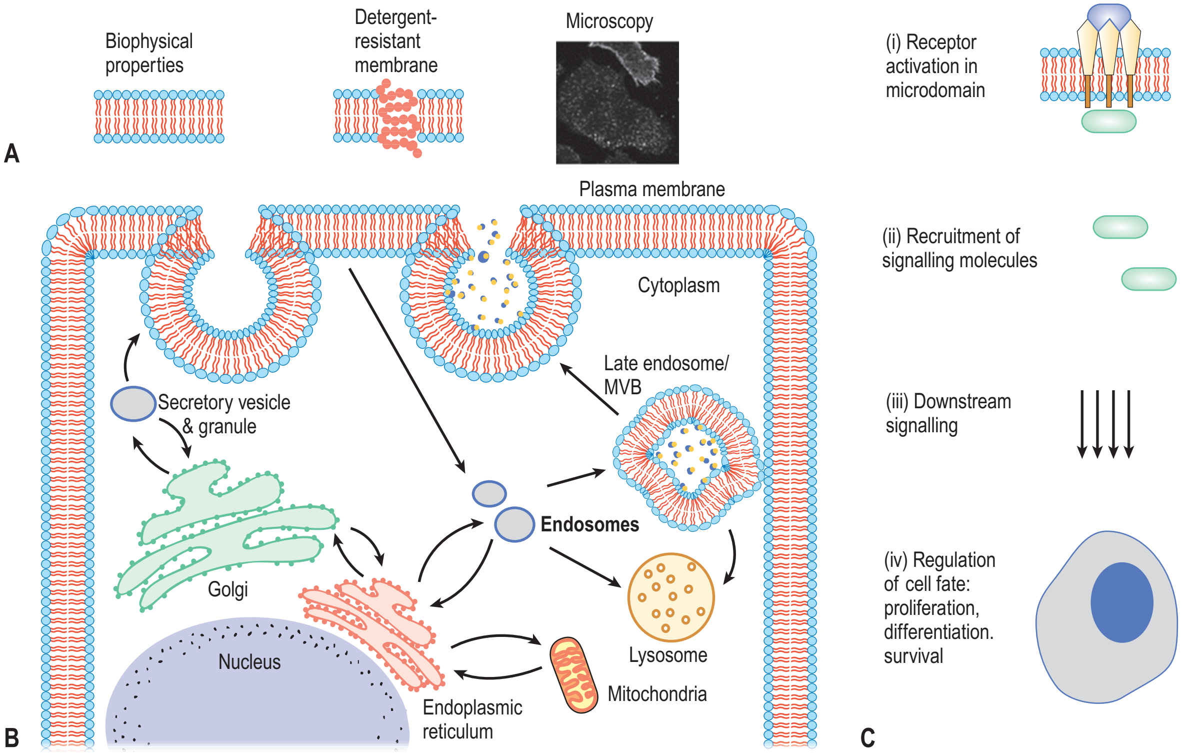
Biophysical properties
Detergent-resistant membrane
Microscopy
A
Plasma membrane
Cytoplasm
Secretory vesicle & granule
Late endosome/ MVB
Golgi
Endosomes
Nucleus
Lysosome
Endoplasmic reticulum
Mitochondria
B
(i) Receptor activation in microdomain
(ii) Recruitment of signalling molecules
(iii) Downstream signalling
(iv) Regulation of cell fate: proliferation, differentiation. survival
C

FIGURE 4-5 Plasma membrane microdomain dynamics. Membrane lipids such as phospholipids, sphingomyelin and cholesterol separate into two distinct phases: a highly mobile phase (Ld) allowing free diffusion and a highly ordered phase (Lo) greatly restricting their mobility. Sphingomyelin and cholesterol are in the latter group. These differences greatly affect membrane function such as endosome formation, cell proliferation and many other processes. The diagram in (**A**) shows the techniques involved in demonstrating lipid mobility in plasma membranes during different cell activities; (**B**) shows different types of membrane with different lipid membrane microdomains; (**C**) shows how separation of different microdomains yields information on the various protein receptors and signalling protein molecules which can be isolated from separated lipid membranes domains. (From Inder et al., 2013.)

where the activation of the receptor is dependent on recruitment to the raft of the effector ligand. It is likely that the enormous complexity associated with microdomain function will only be revealed using a systems biology approach such as a combination of proteomics and transcriptomics.

Many other specializations occur in the plasma membrane, depending on the cell type, such as junctional complexes, gap junctions, desmosomes, hemidesmosomes and contact sites with the basement membranes (see below). In the eye these membrane specializations are developed to a high level. For instance, the photoreceptor cell (Fig. 4-4) (see Ch. 1, p. 41) is a highly polarized structure comprising a receptor component, a nucleus and a synapse. The rod photoreceptor (specialized for scotopic vision; see Ch. 9) develops as an evagination of the plasma membrane, which folds upon itself many times to form stacks of membranous disks by fusion of the peripheral disk membrane. The plasma membrane is typical of any cell, i.e. it comprises a lipid bilayer containing a high concentration of membrane proteins. The lipid bilayer is a self-assembling sheet of phospholipid that adopts the bilayer format because of the physicochemical properties of the polar phospholipids, ensuring that the polar groups are external and the hydrophobic groups form the inner layer of the leaflet. The photoreceptor can adopt this special arrangement because it has more cholesterol in its bilayer, not only making it less fluid but also preventing crystallization of the membrane by inhibiting possible phase transition of the hydrocarbons. The synaptic terminal of the photoreceptor interacts with mobile Ca^{2+} channels in the bipolar cell which appear to function within lipid rafts confined in their mobility by attachments to the cytoskeleton (see Part B of the figure in Box 4-1).

Endoplasmic reticulum and Golgi apparatus

A wide variety of cell organelles are embedded in a cytoplasmic gel, which is traversed by a system of membranes, the endoplasmic reticulum (ER). The ER, a series of thin bilayered membranes, is a flowing dynamic system constantly forming and reforming. Cisternal, tubular and vesicular elements exist. The rough ER (RER) is distinguished from the smooth ER as a ribosome-studded structure that is highly developed in secretory cells such as the lacrimal gland acinar cell, and is specialized in other cells, e.g. the sarcoplasmic reticulum of striated (including extraocular) muscle. The RER is arranged *en face* in rows or rosettes of ribosomes (polysomes). Newly synthesized proteins come off the ribosomes and are threaded through the lipid bilayer into the interior of the ER where they are post-translationally folded, ready for secretion via the Golgi apparatus by vacuolar budding and fusion with the plasma membrane for exocytosis (Box 4-2). The smooth ER is also the site of synthesis of molecules such as lipids, triglycerides and steroids, and is prominent in cells such as the RPE and meibomian gland cells; the Golgi apparatus is a membranous stack of flattened cisternae which receives proteins, now equipped with a leader sequence required for secretion, for sorting and exocytosis from the smooth ER; the smooth ER also plays a role in lipid trafficking and modification.

The ER and the Golgi apparatus have other functions: they are involved in signalling, for instance during mitogen activation of cells for proliferation, via the small GTPases which are activated at the plasma membrane and also signal in the ER and Golgi body (Box 4-3). In fact, small GTPases and other molecules, such as phosphoinositides, provide a signature for each organelle and are involved in the specific lipid membrane folding that characterizes each organelle (Box 4-3). The ER also provides a regulatory role in ensuring quality control of good protein secretion (GPS). This is regulated by a process of ubiquitinylation by which the level of ubiquitin molecules added to the protein determines whether the protein is targeted for secretion or degradation in the proteasome (see eFig. 4-3).

Additional content available at https://expertconsult.inkling.com/

If proteins are not folded correctly they are unable to function and the ER generates a ‘stress response’, also known as the unfolded protein response (UPR), which is one form of ER stress. Together with a second fail-safe mechanism, mediated by a kinase known as the mammalian target of rapamycin (mTOR), the UPR and the mTOR pathways control many cellular processes including programmed cell death (apoptosis), protein translation, waste disposal (autophagy), energy supply (ATP), and response to inflammatory stimuli.

Additional content available at https://expertconsult.inkling.com/.

The ER also forms the nuclear envelope during telophase, when a series of flat vesicles surround the chromosomes and fuse at their edges. The envelope contains many *nuclear pores*, which are composed of eight cylindrical filamentous structures in a highly organized arrangement. Pores act as molecular sieves, permitting rapid passage of small 4.5 nm (4.5 kDa) particles and slower passage of larger molecules (12–70 kDa). The outer aspect of the nuclear envelope in secretory cells is lined with ribosomes and polysomes, while the inner surface is in contact with a nuclear filamentous matrix.

Mitochondria

Mitochondria are small (2 μm long) oval-shaped organelles comprising a two-membrane system of compartments, the inner one of which is composed of two domains, the inner boundary membrane (IBM) and the invaginations, folded into structures termed cristae (Fig. 4-6). At the cristae junction with the IBM, the mitochondrion inner membrane organizing system (MINOS) is located and separates the two populations of proteins: those in the IBM assist in protein movement (translocases) and those in the cristae contain the proteins involved in the respiratory chain, the F1FO-ATP synthase and carrier proteins for ADP/ATP (see below).

The intermembrane space contains carrier proteins that are responsible for the transport of metabolites

The way a protein is folded critically determines its function. This process is monitored by a ubiquitous cellular protein called ubiquitin (Ub) which is bound to the protein as it emerges from the endoplasmic reticulum. As more and more Ub molecules are added to the protein, it protects it from degradation, while as more Ub molecules are removed by the protein, it becomes susceptible to degradation and recycling/removal. This process is tightly regulated (see eFig. 4-3).

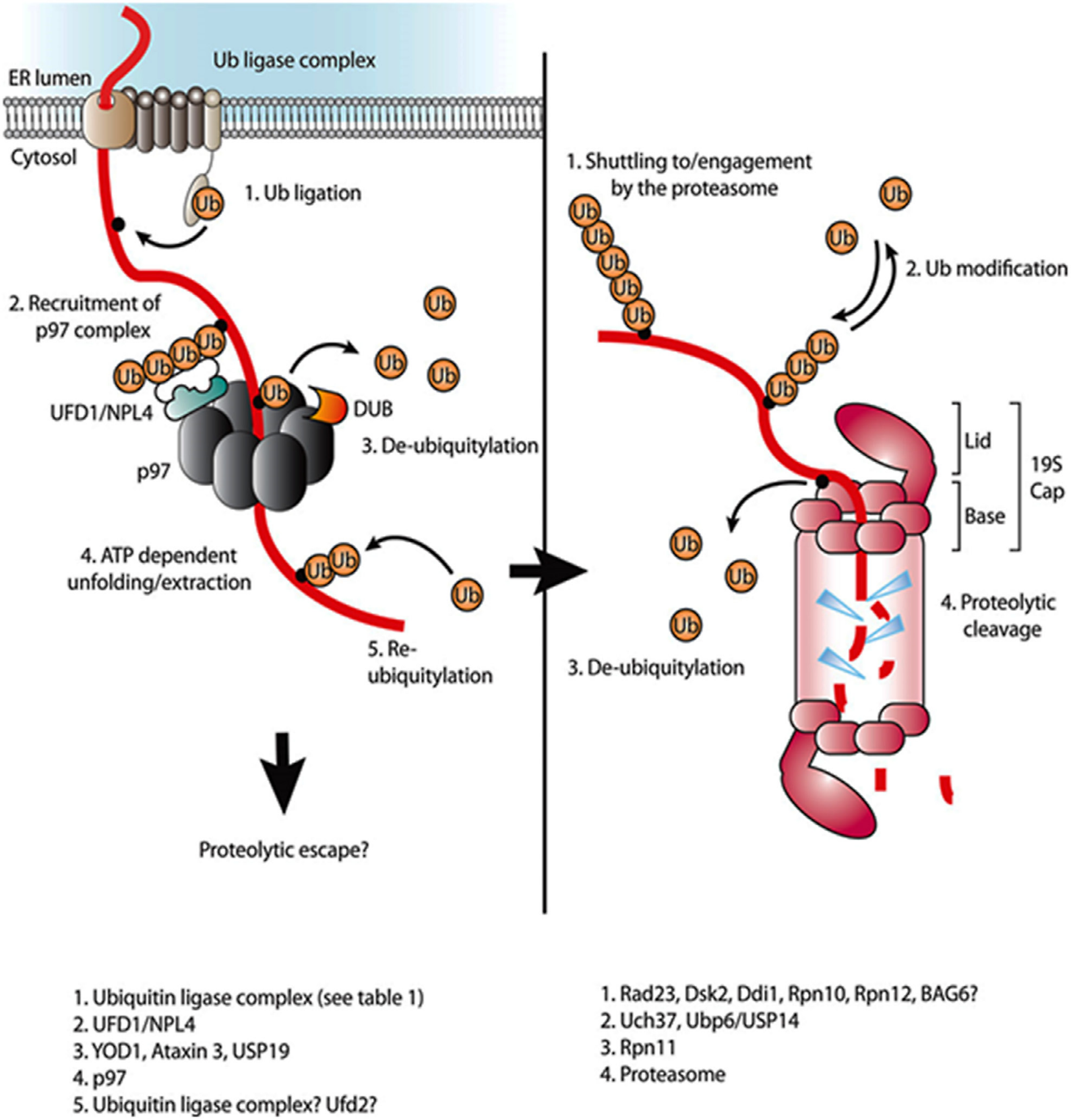
ER lumen
Ub ligase complex
Cytosol
Ub
1. Ub ligation
2. Recruitment of p97 complex
Ub
Ub
Ub
Ub
UFD1/NPL4
Ub
DUB
3. De-ubiquitylation
p97
Ub
Ub
Ub
4. ATP dependent unfolding/extraction
Ub
Ub
Ub
5. Re-ubiquitylation
Proteolytic escape?
1. Shuttling to/engagement by the proteasome
Ub
Ub
Ub
Ub
Ub
Ub
Ub
2. Ub modification
Ub
Ub
Ub
Lid
19S Cap
Base
Ub
Ub
Ub
3. De-ubiquitylation
4. Proteolytic cleavage
1. Ubiquitin ligase complex (see table 1)
2. UFD1/NPL4
3. YOD1, Ataxin 3, USP19
4. p97
5. Ubiquitin ligase complex? Ufd2?
1. Rad23, Dsk2, Ddi1, Rpn10, Rpn12, BAG6?
2. Uch37, Ubp6/USP14
3. Rpn11
4. Proteasome

eFIGURE 4-3 The function of a protein can be disabled both by dislocation and degradation. In the left panel a dislocated protein is ubiquitinylated at the ER membrane and entered into a complex with p97 via NFD1/NPL4. Ubiquitin is cleaved, which allows the protein to be threaded through the central pore of the protein complex, p97. The protein is then re-ubiquitinlyated for proteasomal targeting and final degradation as shown in the right panel. (From Claessen et al., 2012.)

Cells respond to unfavourable conditions by demonstrating a stress response. For instance this occurs where there is an excess of free radical generation by the cell but can also occur when normal physiological mechanisms are tested, as in a condition known as ER stress. Stress is dealt with by the cell through a variety of mechanisms, including the ER stress responses and a cell signalling system known as the mammalian target of rapamycin (mTOR) response, which is implicated in many conditions involving inflammatory responses in the eye (eFig. 4-4).

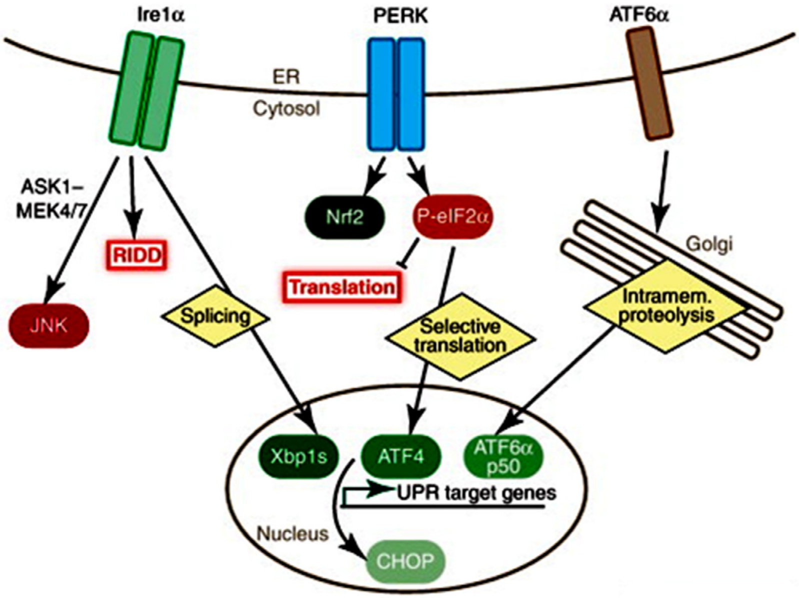
Ire1α
PERK
ATF6α
ER
Cytosol
ASK1–
MEK4/7
RIDD
JNK
Splicing
Nrf2
P-eIF2α
Translation
Selective translation
Golgi
Intramem. proteolysis
Xbp1s
ATF4
ATF6α p50
UPR target genes
Nucleus
CHOP

eFIGURE 4-4 IRE-1α is an ER stress sensor which signals to the nucleus to activate the unfolded protein stress response (UPR) and maintain cellular physiology. However, if the cell fails to respond, convergent signals through PERK activate a nuclear transcription factor (CHOP) which initiates apoptosis. A similar signal can also be initiated via ATF6α to activate CHOP

BOX 4-2 ENDOCYTOSIS, EXOCYTOSIS AND EXOSOMES

Endocytosis is generally achieved via incorporation of ligand–receptor complexes in clathrin-coated vesicles. This applies to soluble proteins and to small and large particles such as viruses, which frequently use constitutive cell surface receptors to enter cells.

Clathrin-coated vesicles start as small pits on the cell surface. When the vesicle is fully intracellular it loses its clathrin coat and becomes an endosome, which fuses with primary lysosomes that have a high content of acid hydrolases and other proteases. These lead to degradation of the ingested material, and further processing depending on the cell type. Certain cell surface receptors are recycled to the cell membrane during this process to engage further extracellular ligand (**A**). Clathrin-coated pits are normally restricted to the region of the plasma membrane by the cortical cytoplasm actin organization. Relaxation of this actin assembly by proteins such as latrunculin B allows movement of the coated pits.

Exocytosis occurs by a similar process but in reverse. Cytoplasmic granules such as those in exocrine glands or in granulocytes are lipid vesicles containing material for extrusion. Secretory vesicles bud from the Golgi apparatus and are transported towards the plasma membrane by the cytoskeleton. There, they fuse in a lipid microdomain assisted by proteins such as SNARE and the SNARE-binding proteins in a 'targeting patch' (**B**). Fused secretory vesicle membranes act as targets for further secretory granule fusion in cells such as mast cells. In addition, vesicles termed exosomes may be 'pinched' off cells such as macrophages and immune dendritic cells in the same way as platelets are pinched off megakaryocytes. Exosomes convey material (information) from one cell to another and the nature of the information varies from cell to cell.

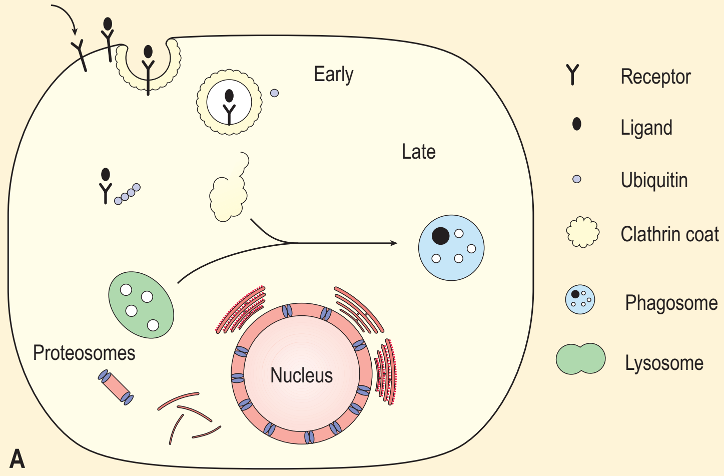

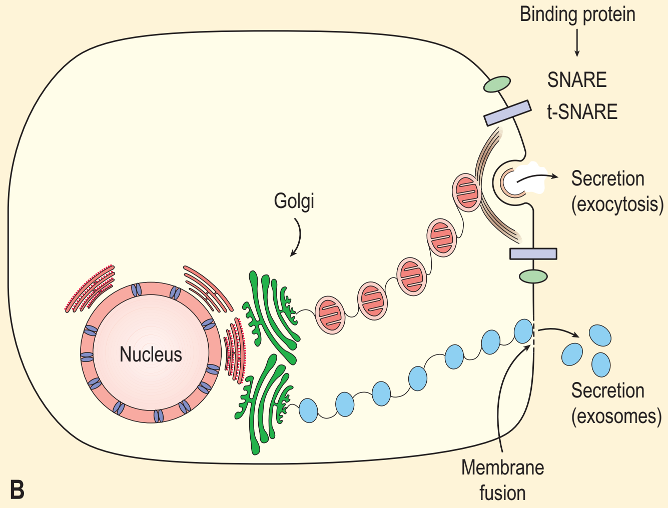

BOX 4-3 SCHEMATIC REPRESENTATION OF HOW MITOGENS ACTIVATE CELLS

Ligand binding to the tyrosine kinase receptor (PTKR) activates the small GTPases, Ras, through a complex of signalling molecules (Grb2/SOS). Activated Ras on endosomes from the plasma membrane induces signalling in this organelle as well as in the Golgi apparatus (indirectly via phospholipase CgCa and another protein known as GRPI) while inhibition of the small Ras occurs in the endoplasmic reticulum via an inhibitor protein ERII (see figure). In this example of a signalling network, interstin and kinase suppressor of Ras serve as scaffold proteins while p14 acts as an adaptor protein in Ras-independent activation of the kinase, MEK-1 (mitogen-activated protein (extracellular signal-regulated (ER kinase) kinase), by the endosome. The integration of organelles and signalling networks with cytosolic proteins is thus central to proper functioning of the cell in response to an external stimulus such as a mitogen.

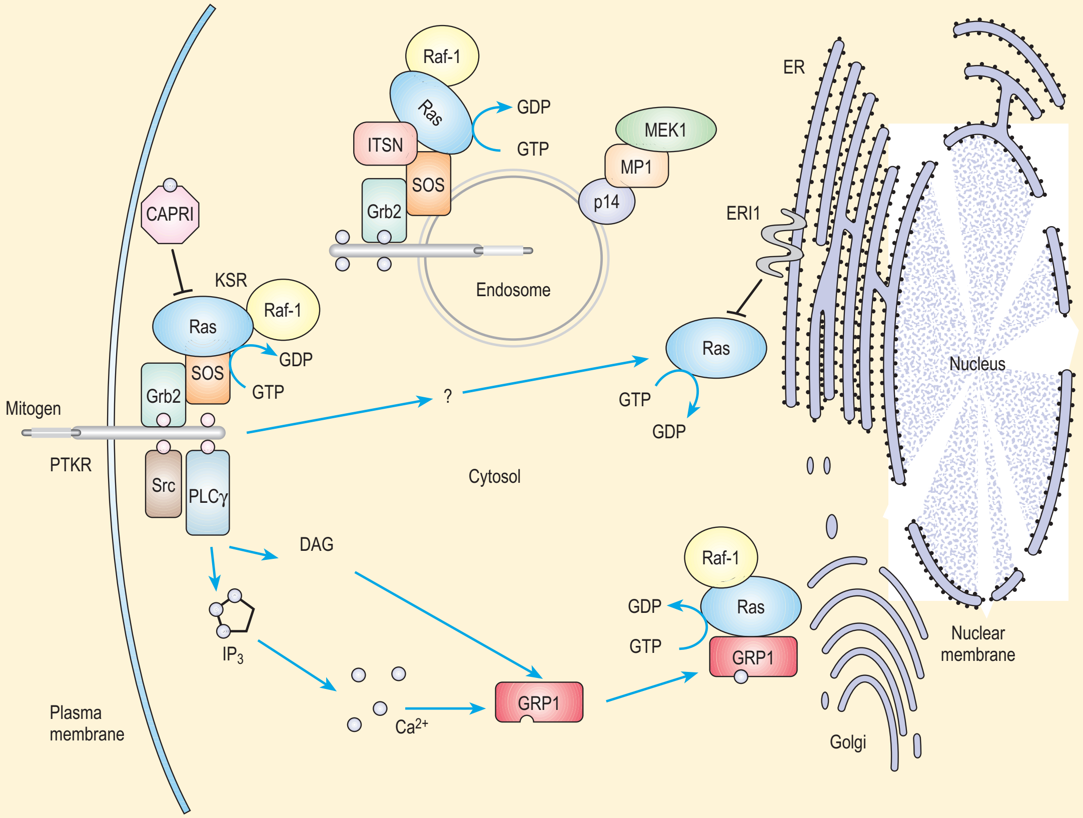

between the two compartments and also between the cytosol and the outer compartment. Their transport systems include antiport, aspartate/glutamate, ornithine/citrulline, maleate/citrate, symport, pyruvate/H$^+$, and urea and porphyrin synthesis. Mitochondria are the powerhouses of the cell and have several essential metabolic functions as they contain all the elements for the respiratory assembly, for the citric acid cycle and for fatty acid metabolism. Their main functions therefore are to act as the site of energy-rich adenosine triphosphate (ATP)/guanosine triphosphate (GTP) formation, to function as a calcium store mainly in the form of calcium phosphate, to engage in the uptake of energy-rich substances, and to facilitate the oxidative breakdown of ATP. Mitochondria and the ER form an integrated system known as the ER-mitochondria organizing network (ERMIONE), enabling the transport of lipids and calcium between the compartments.

Mitochondria are powerhouses of nutrient handling and energy storage, essential for cell growth and proliferation. Uncontrolled, this can lead to tumours,

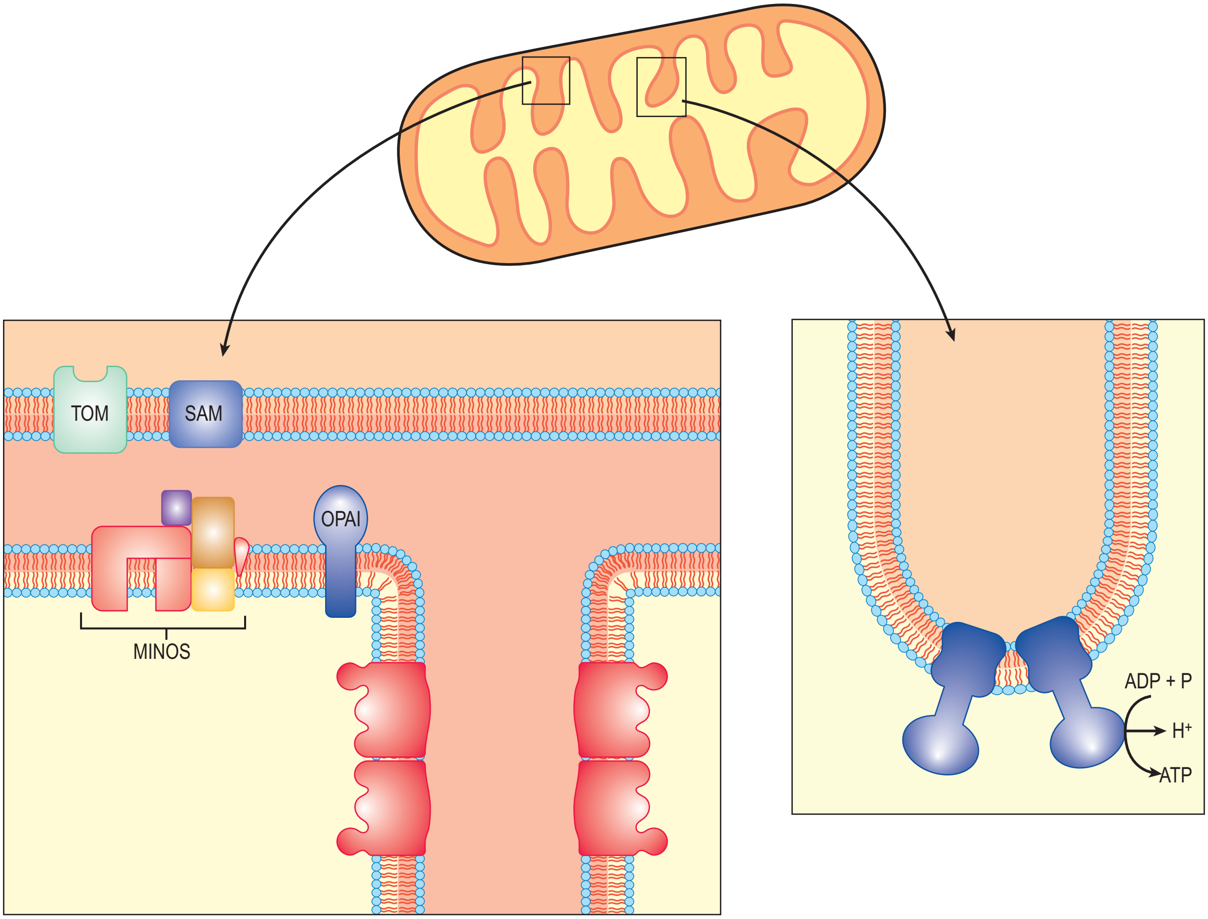

FIGURE 4-6 Mitochondrial membrane organization. Mitochondria are composed of an extensive intraorganelle membrane structure to maximize their numerous functions in metabolism, cell death, autophagy and energy control. Membrane organization is regulated by protein complexes such as MINOS, which occurs on the inner mitochondrial membrane, mostly at cristae junctions. MINOS is composed of several proteins and interacts with others such as the inner membrane morphology proteins OPA1, TOM and SAM.

and so a balance between homeostatic metabolism and cell death is essential. Mitochondria are therefore regulators of cell death through several routes (Box 4-4).

The central inducer of cell death is mitochondrial cytochrome c, which is released into the cytoplasm and complexes with Apaf-1 and caspase 9 (the apoptosome) and ultimately leads to caspase 3 and 7 as the final executioners. More recently, mitochondria have been recognized as sensors and regulators of cytosolic Ca^{2+} levels, through Ca^{2+} transporters which have central physiological roles in signalling, proliferation, metabolism and ultimately cell survival.

There are many 'death receptors' (receptors which initiate the cell programme leading to death of the cell, i.e. apoptosis), including FasL, TNFR and granzyme B, as well as environmental conditions which promote cell death, such as hypoxia, genotoxic damage, cytokine or nutrient deprivation (linked to the tricarboxylic acid cycle and cytochrome c) (Fig. 4-7), excess steroid exposure, UV exposure and toxic drug exposure, and the balance between life and death is maintained by pro- and anti-apoptotic Bcl-2 proteins (see Ch. 7).

Mitochondria have their own complement of DNA (but no histones, and thus are not susceptible to epigenetic changes; see Fig. 4-1) as well as ribosomal RNA/transfer RNA, and generate a series of mitochondrion-specific proteins associated with mutations and a

BOX 4-4 MITOCHONDRIA AND APOPTOSIS: THE SIGNALLING PATHWAYS

Cytochrome c (red spots in Figure 4-7) in mitochondria controls many apoptosis pathways involving caspases, granzymes and perforin molecules as well as several pro- and anti-apoptotic genes such as Bax, Bcl2, BH3, Bid and Bim. Many of these, plus numerous other proteins, form a complex known as the apoptosome.

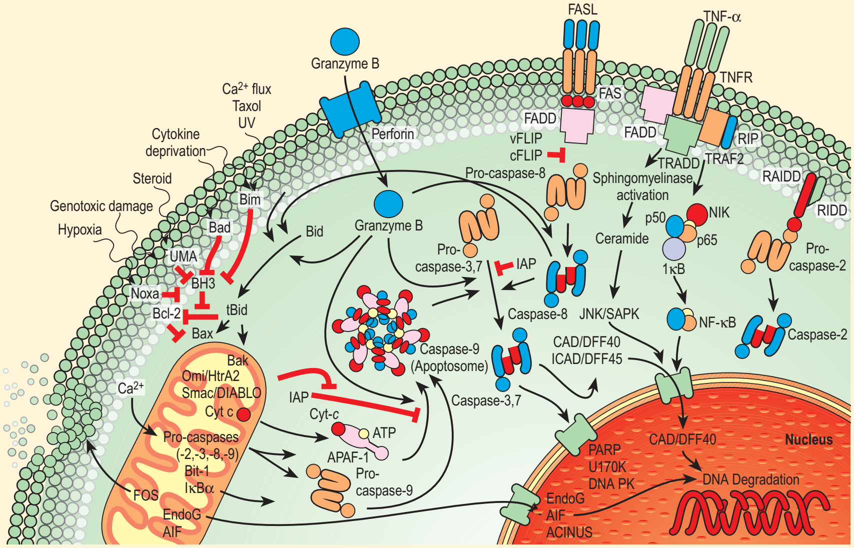
Granzyme B
Ca²⁺ flux
Taxol
UV
Cytokine deprivation
Steroid
Genotoxic damage
Hypoxia
Perforin
Bim
Bad
UMA
BH3
Noxa
Bcl-2
tBid
Bax
Bid
Granzyme B
Bak
Omi/HtrA2
Smac/DIABLO
Cyt c
Ca²⁺
Pro-caspases (-2,-3,-8,-9)
Bit-1
IκBα
FOS
EndoG
AIF
IAP
Cyt-c
ATP
APAF-1
Pro-caspase-9
Caspase-9 (Apoptosome)
Pro-caspase-3,7
IAP
Caspase-3,7
FASL
FAS
FADD
vFLIP
cFLIP
Pro-caspase-8
Caspase-8
TNF-α
TNFR
FADD
RIP
TRADD
TRAF2
Sphingomyelinase activation
Ceramide
JNK/SAPK
p50
NIK
p65
1κB
NF-κB
RAIDD
RIDD
Pro-caspase-2
Caspase-2
CAD/DFF40
ICAD/DFF45
PARP
U170K
DNA PK
EndoG
AIF
ACINUS
CAD/DFF40
Nucleus
DNA Degradation

(Figure reproduced with permission of EMD, Madison, WI, USA, www.calbiochem.com.)

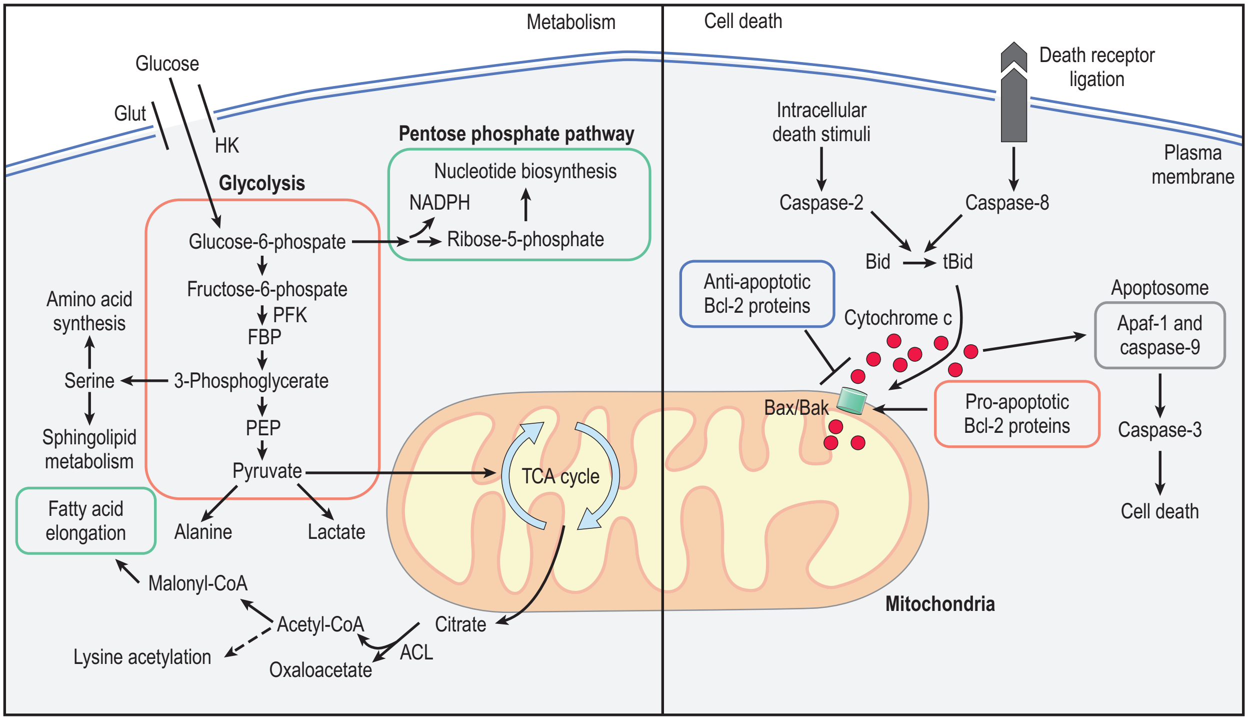
Metabolism
Cell death
Glucose
Glut
HK
Glycolysis
Pentose phosphate pathway
Nucleotide biosynthesis
NADPH
Glucose-6-phospate
Ribose-5-phosphate
Fructose-6-phospate
PFK
FBP
Amino acid synthesis
Serine
3-Phosphoglycerate
Sphingolipid metabolism
PEP
Pyruvate
Fatty acid elongation
Alanine
Lactate
TCA cycle
Malonyl-CoA
Acetyl-CoA
Citrate
ACL
Lysine acetylation
Oxaloacetate
Death receptor ligation
Intracellular death stimuli
Plasma membrane
Caspase-2
Caspase-8
Bid
tBid
Anti-apoptotic Bcl-2 proteins
Cytochrome c
Apoptosome
Apaf-1 and caspase-9
Bax/Bak
Pro-apoptotic Bcl-2 proteins
Caspase-3
Cell death
Mitochondria

FIGURE 4-7 Cell metabolism and cell death (apoptosis) are intricately linked through the tricarboxylic acid cycle. Cross-talk between the signalling pathways for these two systems provides cross-regulation and determines whether a cell will live or die. Much of this cross-talk goes on in the mitochondria and is dependent critically on cytochrome c, which is central to both pathways in generation of ATP for energy and caspase-1 for apoptosis. (Andersen and Kornbluth, 2013)

number of discrete syndromes, some with ophthalmological consequences (see Ch. 3, p. 156). Since mitochondria originate only from ova, transmission of these genetic defects is purely maternal.

The nucleus

The nucleus is contained within a perforated sac, formed by pores (nuclear pore complexes). These allow transport of proteins and other signalling molecules passage between the transcriptional machinery and the appropriate instructions from outside. Both import and export of materials are receptor-mediated in an 'address-label'-like system based on small GTPases (see Box 4-3).

The main nuclear component is chromatin, a complex structure of highly extended DNA, RNA and protein in the interphase (non-dividing) cell, which becomes greatly condensed (by 400-fold) to form chromosomes during cell division. Packing of chromatin is achieved by interaction between negatively charged DNA and certain basic proteins (histones) which carry a positive charge at the pH of the cell; euchromatin is less packed than heterochromatin, the proportion of which varies from nucleus to nucleus and may be characteristic of certain cell types, e.g. the 'clock-face' heterochromatin of plasma cells detectable in histological sections. Genome stability is dependent on minimizing DNA damage, which might occur for instance during cytokinesis and chromatin condensation, and chromatin regulators exist to ensure stability. Most of these are histone-modifying enzymes such as acetyl and methyl transferases and deacetylases. Much of gene modification is mediated via reversible nucleotide methylation and acetylation.

The nuclear membrane also contains receptors for ligands, which may be synthesized in the cytoplasm or may have been endocytosed through plasma membrane receptors. Typical nuclear membrane receptors include steroids, growth factors such as fibroblast growth factor and novel groups of proteins known as the peroxisome proliferator-activated receptors (PPARs), which are involved in many cell processes such as lipid and glucose homeostasis, wound healing, and inflammation generally. PPARs are unique receptors that allow integration of signals mediated by lipophilic ligands with plasma membrane-derived signals. Nuclear membrane receptors may be organized for induction or suppression of genes in a coordinated fashion (Fig. 4-8). In addition, a major function of the nucleus is to manage the packaging of mRNA into ribonucleoprotein particles called mRNPs. These are associated with very large numbers of proteins involved in the transcription of mRNA in the nucleus to translation and degradation of mRNA in the cytoplasm (Fig. 4-9).

The nucleolus is essentially composed of RNA and fibrillar material, and is the site of ribosomal RNA synthesis and intense transcriptional activity. In addition, it may contain non-coding regions of RNA involved in the stress response. The nucleolus develops during the late stages of mitosis in association with specific regions on the chromosomes, known as the nucleolar organizer centre.

The intracellular matrix

The cytoplasm is a highly viscous aqueous medium that has deformability (elasticity). Physically, it exists at different times as a gel or as a sol. The cortical cytoplasm (ectoplasm) is more akin to a gel structure, while the endoplasm is usually more fluid. Thus the cortical cytoplasm restricts movement of organelles such as coated vesicles (see Box 4-2). The gel-like properties of the cytoplasm are the result of the binding of 'structured' water molecules and Ca^{2+} ions to polymeric filaments; in fact, water, as the main constituent of cells, has a significant role even within the molecular interstices of cell membrane receptors where in its 'ordered' state it contributes to the tertiary structure of the molecule as shown by rhodopsin molecules in the photoreceptor disks (Fig. 4-10).

Three main types of intracellular filaments are described, microfilaments, such as actin; intermediate filaments; and group 3 cytoskeletal fibres, such as tubulin and myosin (see Box 4-5). More recently, a fourth type has been recognized as a cytoskeletal element, although this family of molecules was discovered over 40 years ago, namely septins.

Microfilaments. Microfilaments (5–7 nm), such as actin, tropomyosin and troponin, are universal constituents of cells and are involved in almost every cellular activity, including cell motility, contractility, endo- and exocytosis (see Box 4-2) and maintaining

BOX 4-5 DIAGRAM OF CELLULAR STRUCTURES INDICATING THE VARIOUS CYTOPLASMIC CONTENTS

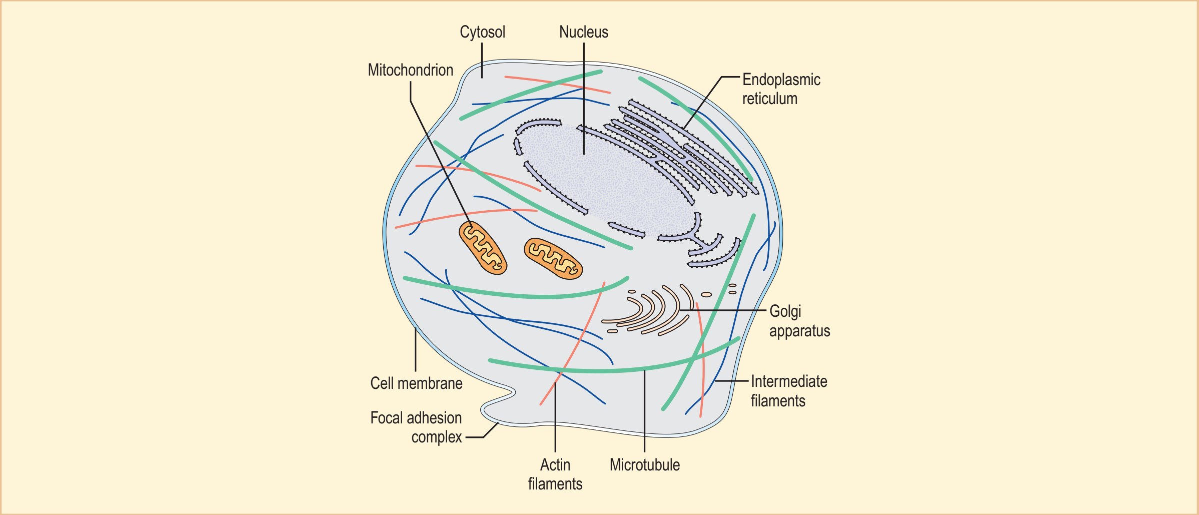
Cytosol
Nucleus
Mitochondrion
Endoplasmic reticulum
Golgi apparatus
Cell membrane
Focal adhesion complex
Intermediate filaments
Actin filaments
Microtubule

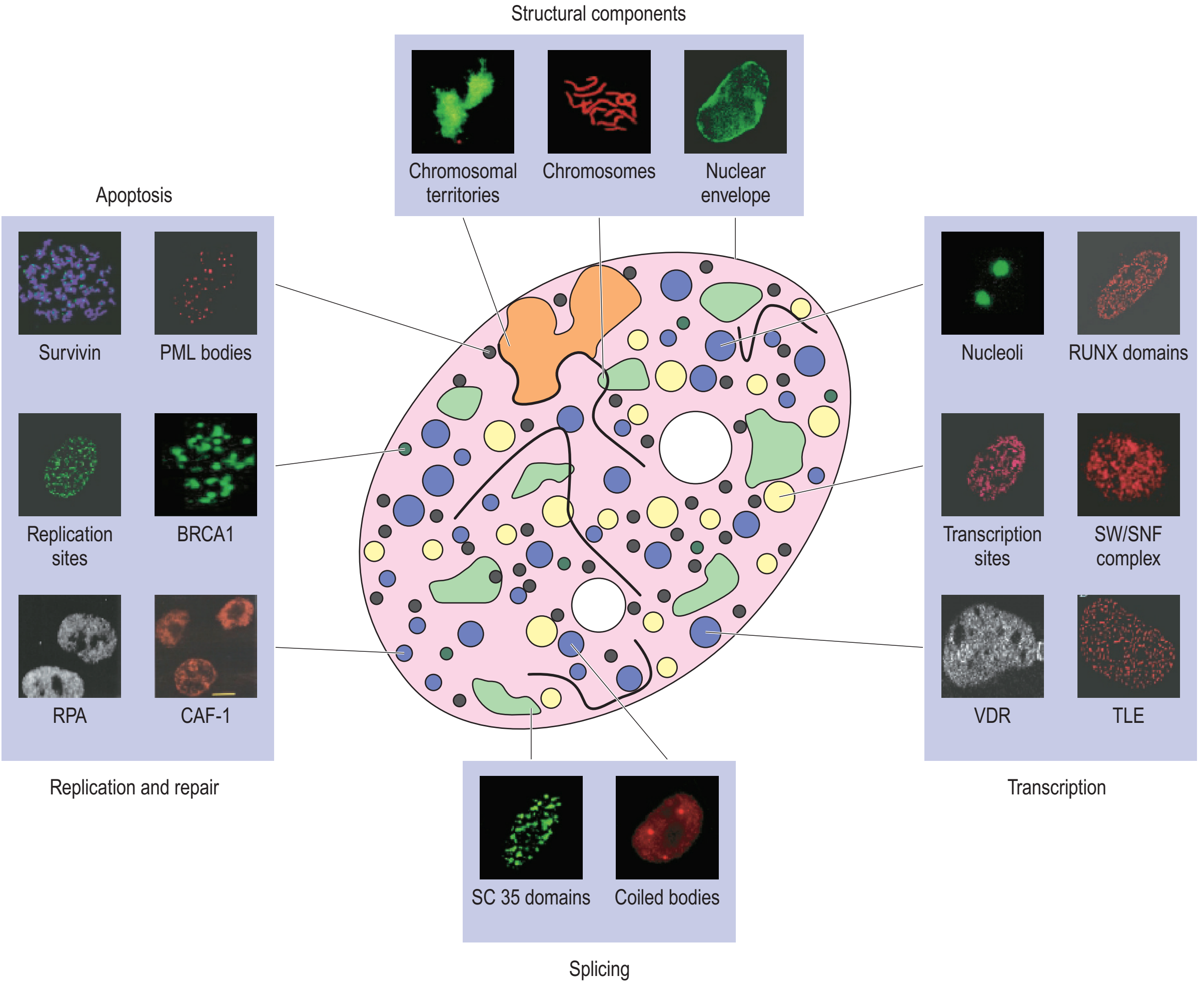
Structural components
Chromosomal territories
Chromosomes
Nuclear envelope
Apoptosis
Survivin
PML bodies
Replication sites
BRCA1
RPA
CAF-1
Replication and repair
Nucleoli
RUNX domains
Transcription sites
SW/SNF complex
VDR
TLE
Transcription
SC 35 domains
Coiled bodies
Splicing

FIGURE 4-8 Subnuclear organisation of different nuclear structures. The nuclear architecture is functionally linked to the organization and sorting of regulatory information. Immunofluorescence microscopy of the nucleus *in situ* has revealed the distinct non-overlapping subnuclear distribution of vital nuclear processes. (From Stein et al., 2003, with permission from Elsevier.)

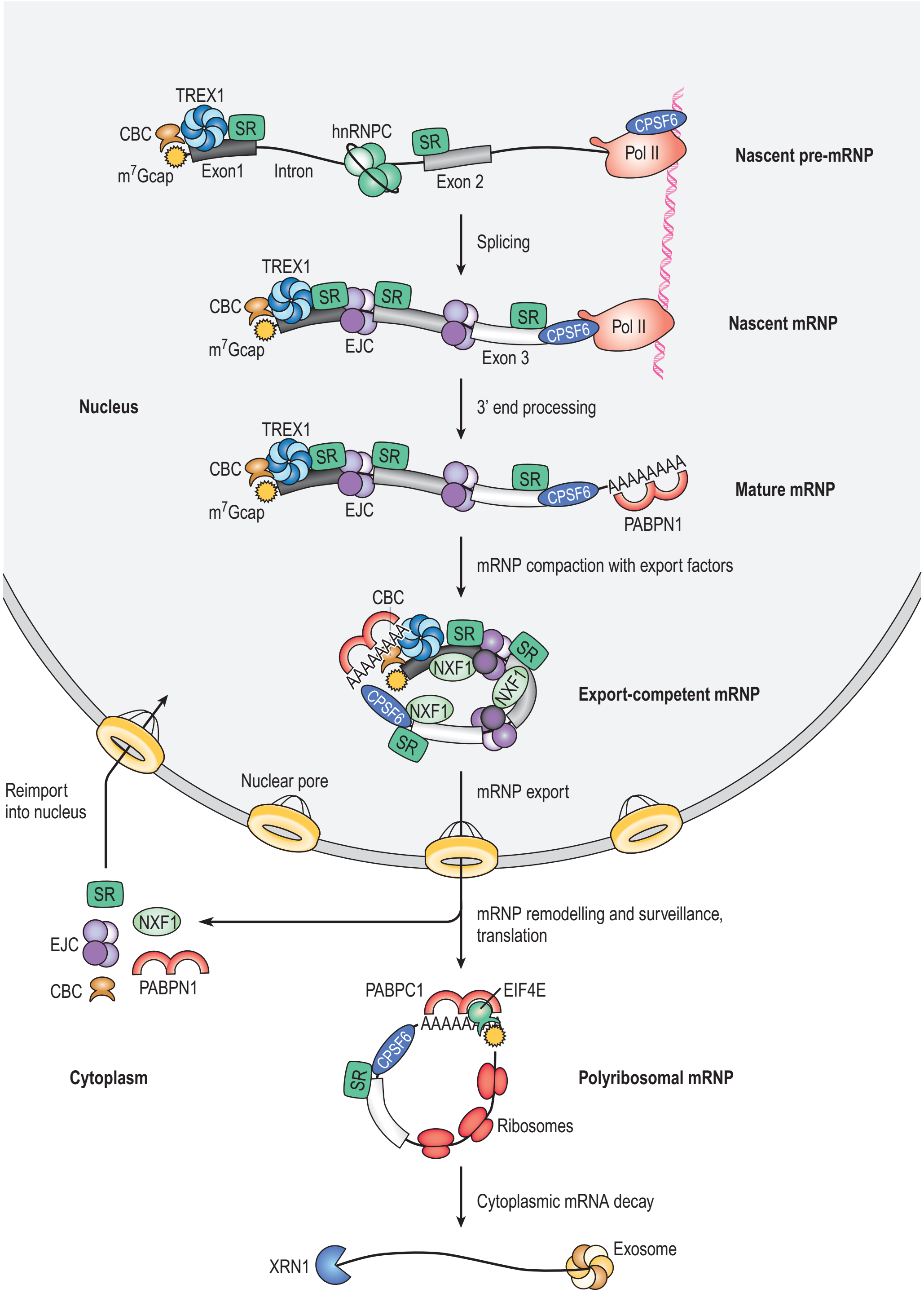
TREX1
CBC
SR
hnRNPC
SR
CPSF6
Pol II
Nascent pre-mRNP
m7Gcap
Exon1
Intron
Exon 2
Splicing
TREX1
CBC
SR
SR
SR
CPSF6
Pol II
Nascent mRNP
m7Gcap
EJC
Exon 3
Nucleus
3' end processing
TREX1
CBC
SR
SR
SR
CPSF6
AAAAAAAA
Mature mRNP
m7Gcap
EJC
PABPN1
mRNP compaction with export factors
CBC
SR
SR
AAAAAAAA
NXF1
NXF1
CPSF6
NXF1
SR
Export-competent mRNP
Reimport into nucleus
Nuclear pore
mRNP export
SR
NXF1
EJC
CBC
PABPN1
mRNP remodelling and surveillance, translation
PABPC1
EIF4E
AAAAA
CPSF6
SR
Cytoplasm
Polyribosomal mRNP
Ribosomes
Cytoplasmic mRNA decay
XRN1
Exosome

FIGURE 4-9 Summary diagram of how messenger ribonucleoprotein particles are assembled and remodelled. It is now realized that there is considerable diversity in how these moelcules are generated. In the diagram one example is shown of an RNP cap binding complex (CBC) facilitating splicing of the 3′-end formation and export together with bound mRNAs through the nuclear pores into the cytoplasm. Finally, it is then bound to cytoplasmic proteins in ribosomes and becomes actively translated mRNA. (From Muller-McNicoll and Neugebauer, 2013.)

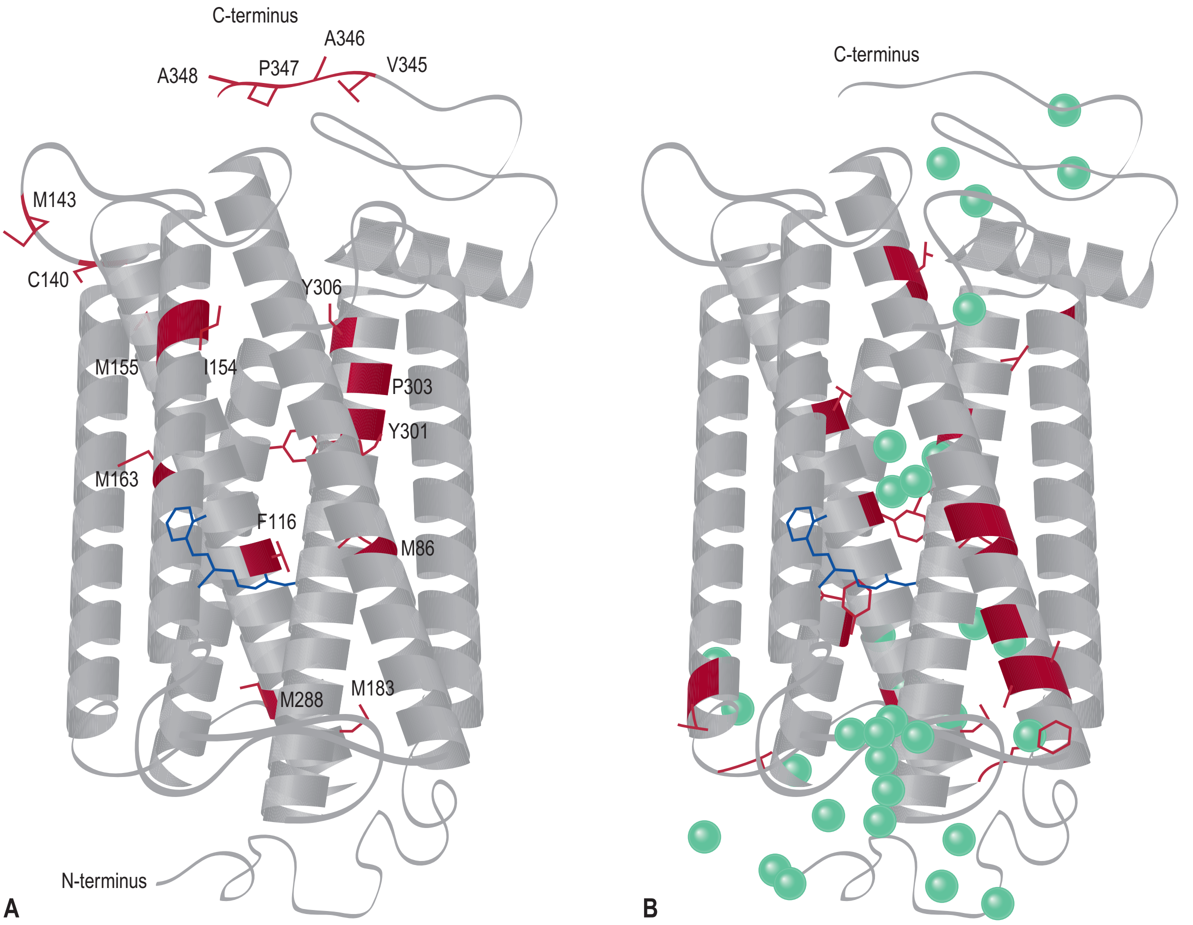

FIGURE 4-10 Ribbon diagram of the rhodopsin molecule as it flows in the photoreceptor disk membrane. Vitamin A bound to rhodopsin is shown in blue in (**A**). In (**B**), water molecules represented as green spheres, interact intimately with the binding of vitamin A to rhodopsin and are bound and released in concert with vitamin A as it is bound and released from the rhodopsin scaffold. It is believed that the water molecules modify, if not determine, how vitamin A interacts with rhodopsin. (From Orban et al., 2010)

structural integrity, where they insert into cell adhesion junctional complexes (see eBox 4-2).

Actin bundles are brought to a highly developed level in muscle cells. Actin occurs in several forms within the cell, depending on its associated protein, e.g. as a fine lattice meshwork or as a sheaf of fibres (stress fibres). Stress fibres are often the most prominent cytoskeletal features in a cell: they not only generate force but they respond to mechanical tension by reinforcement and in some respects are similar to sarcomeres of muscle cells.

The cellular distribution and, particularly, the degree of polymerization of microfilaments are determined by the nature of proteins that bind to them, i.e. the microfilament-associated proteins (Table 4-1). Monomeric soluble actin (G-actin) is converted to gel-phase polymerized fibrils (F-actin) by association with certain proteins. In smooth muscle cells and fibroblasts, filamin assists polymerization; in contrast, in lymphocytes, profilin and thymosin maintain G-actin in the depolymerized state, presumably to facilitate flexibility in cell shape during rapid migration within tissues. The specificity of these actin-binding proteins is remarkable; examples include the ankyrin–spectrin–actin combination, which determines the red cell biconcave shape, and the spectrin–peripherin–actin

eBox 4-2

Filaments and junctions

Intercellular connections or junctions are structures that make cells into tissues and tissues into organs. There are several types, each of which is composed of junction-specific proteins.

Tight junctions – have no detectable 'space' between the cell membranes. They are also known as zonulae occludens; they form a barrier to paracellular diffusion of all molecules, including water and ions. They are also involved in regulation of epithelial cell proliferation and differentiation. They occur at such sites as the blood–aqueous barrier of the ciliary body and the blood–retinal barrier at the apex of the RPE cell. They consist of a system of ridges and grooves, as seen by freeze–fracture studies. There are four major classes of tight junction proteins:

- transmembrane proteins – occludin, claudin
- adaptors – ZO-1, cingulin, MUPPI
- transcriptional and post-transcriptional regulators – AP-1
- signalling proteins – αPKC, CDK4.

Claudins have unique abilities to selectively permit transport of charged ions and vary from tissue to tissue (**A**).

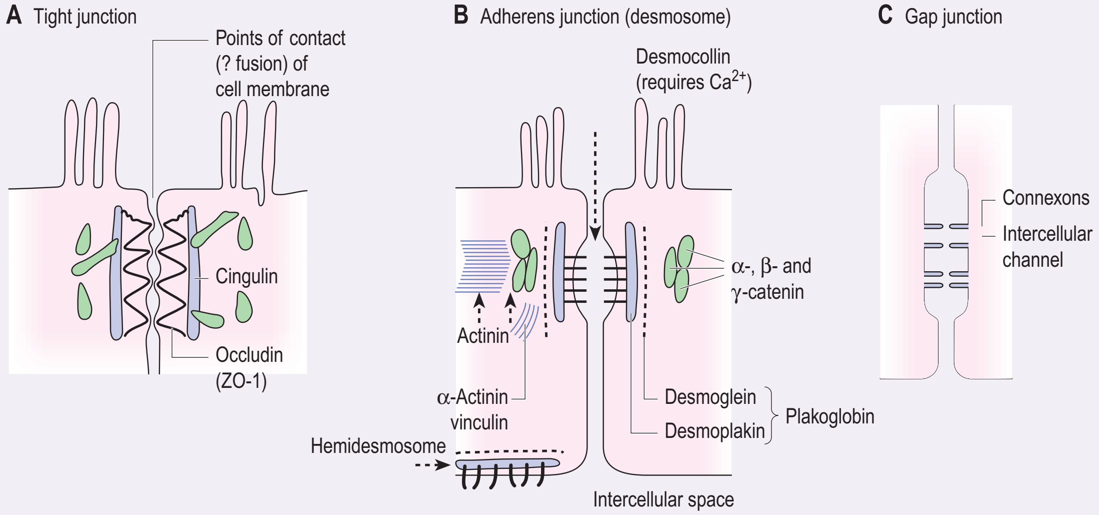
A Tight junction
Points of contact (? fusion) of cell membrane
Cingulin
Occludin (ZO-1)
B Adherens junction (desmosome)
Desmocollin (requires Ca^{2+})
α-, β- and γ-catenin
Actinin
α-Actinin vinculin
Hemidesmosome
Desmoglein
Desmoplakin
Plakoglobin
Intercellular space
C Gap junction
Connexons
Intercellular channel

Desmosomes – form specialized 'adherens' junctions of 20 nm width between cells. Two types are described: spot desmosomes (at single site) and belt desmosomes (as a ring round the apex of the cell). The latter are also known as zonulae adherens. Their probable role is mechanical adhesion: cytoplasmic filaments insert into spot (cytokeratin) and belt (actin) forms (**B**). Several proteins are involved in different regions of the desmosome, including desmocollin, desmoglein and desmoplakin, forming a subfamily of cadherins. The cadherins bind to β-catenin and then to actin-based cytoskeletal proteins such as vinculin and α-actinin, ZO-1 and actin itself. The function of tight junctions (above) is dependent on the integrity of the adherens junctions. Hemidesmosomes have a similar plaque formed in this case by bullous pemphigoid antigen (BPAG1).

Gap junctions – so called because there is a 2 nm gap at this adhesion site between cells. Gap junctions occur in the basal regions of epithelial and other cells. They are a highly organized structure composed of 'connexons', plasma membrane domains containing connexins, which have a role in permitting the passage of larger ions such as Ca^{2+}, and other signalling molecules, so permitting a coordinated response by a group of cells, as in ocular or cardiac muscle (**C**).

Synapse – a specialized form of junction between nerves, or between nerves and muscles, and characterized by synaptic vesicles in the axon terminal and pre- and post-synaptic thickening of the plasma membrane (see Ch. 5).

TABLE 4-1 **Function of actin-binding proteins**

Function	Protein	Cell/structure
Gelation	Filamin	Smooth muscle/ fibroblasts
Bundling	α-actin	Muscle
	Fimbrin	Microvilli
	Talin	All cells
Severing	Gelsolin	Macrophages
	Villin	Microvilli
	β-actinin	Skeletal muscle
Depolymerizing	Profilin	Lymphocytes
	Thymosin	All cells
	Actobinin	All cells
Membrane-binding	Vinculin	Adhesion sites
	Spectrin	Red cells, photoreceptors
Receptor transport	Capping proteins	Leucocytes
Junctional complex	Radixin	Liver cells

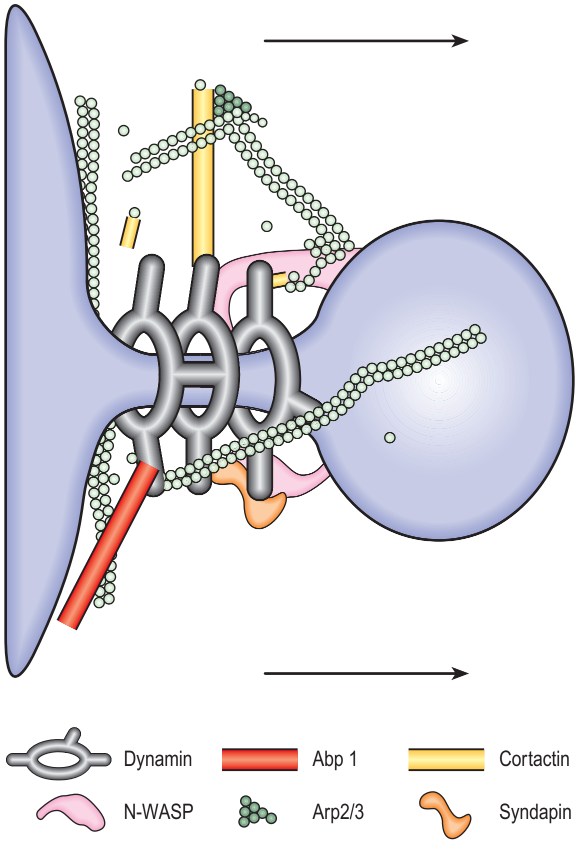

FIGURE 4-11 Dynamin-mediated tubulation and vesiculation of secretory and endocytic compartments. Model showing dynamin-mediated tubulation/vesiculation of a membrane as it might occur at the plasma membrane. Dynamin, its binding protein and associated proteins are likely to function together during vesicle formation. The molecular 'pinchase' activity of dynamin, together with the enhanced actin filament nucleation at the membrane interface, results in the tubulation and severing of the membranous vesicle necks (see text for details). Large black arrows indicate the generation of force and the movement of nascent vesicles. (From Orth and McNiven, 2003, with permission from Elsevier.)

combination between rod outer segment disks and plasma membrane. Actin polymerization and depolymerization is a highly regulated process requiring addition of actin monomers to one end of the microfilament and removal at the opposite end, each of which has separate K_{on} and K_{off} constants. This process is under the control of actin-depolymerization factor (cofilins) which alter these rate constants as necessary, leading to changes in the twist of the molecule and promoting severance of the actin filament.

Regulation of coordinated changes to the actin cytoskeleton is under the control of intracellular enzymes, particularly the Rho family of GTPases (this includes enzymes such as Rho, Rac and Cdc42, the last of which is a cell cycle-related protein). Certain kinases, known as the p-21 activated kinases (PAKs), are involved in regulating some of the diverse changes induced by Rac and Cdc42. PAKs may determine such cellular responses as polarity in epithelial cells and motility in fibroblasts. For example, coordinated functional regulation of the interaction between dynamin, cortactin, actin-binding protein 1, neuronal Wiskott–Aldrich syndrome protein 1 (N-Wasp-1), profilin and actin occurs during the pseudopodial extension and vesicle formation of migrating fibroblasts (Fig. 4-11). Central to this process is the ARP 2/3 complex which is required for actin nucleation. For instance the Arp 2/3 complex is also involved in a relatively recently described family of actin-binding proteins, the coronins, which are involved in leucocyte migration.

Several isoforms and at least three families of actin exist (α, β, γ) and some of these are located in specific parts of the cell, e.g. β-actin in the cell cortex.

Intermediate filaments. Intermediate filaments (10–12 nm) are coiled α-helices which act as stretchable components of the cytoskeleton scaffold. They occur in cell-, tissue- and differentiation-specific distribution in both cytoplasmic and nuclear compartments of the cell, and are classified into five groups depending on domain and sequence homology. Their major function is to protect cells from mechanical and non-mechanical damage (Fig. 4-12), and gene mutations account for around 30 different diseases in man, mostly related to skin, muscle and nerve dysfunction. In general, intermediate filaments are important for the correct positioning and function of cell organelles such as mitochondria and ER.

There are five classes of cytoplasmic intermediate filaments, often used to characterize cells in tissue culture or tumours, and one class of nuclear intermediate filaments:

- *keratins* – found in epithelial cells; over 50 individual members grouped into two types, I and II, exist as heterodimers
- *vimentin* – found in mesenchymal cells
- *desmin* – interconnects myofibrils of muscle cells, at site of Z disk and M line, thus maintaining their register
- *glial fibrillary acidic protein* (glia)
- *neurofilaments* – e.g. S100 protein in neuroectoderm; connect with microtubules via small projections
- *lamins* – line the inner surface of the nuclear envelope as a fibrous lattice (karyoskeleton); more than five types.

One function of intermediate filaments is to ensure normal cytoplasmic positioning and function of different organelles. Lamins undergo considerable molecular disruption during mitosis and may communicate with cytoplasmic intermediate filaments, but how this occurs is not clear. Others, such as keratins, provide mechanical strength to junctional structures such as desmosomes (see eBox 4-2).

Cellular junctions are highly specialized: desmosomes and hemidesmosomes ('junctions' between the cell and basement membrane in epithelia) have some ultrastructural similarity, but at the molecular level there are clear differences between the constituent proteins. Desmosomes are formed by a series of proteins spanning the cell membrane and the intercellular space (desmoglobin, desmoplakin), while hemidesmosomes contain other proteins such as the $\alpha_6\beta_4$ integrin receptor, which binds to laminin in the basement membrane, plus other proteins such as the bullous pemphigoid and the pemphigus antigens.

They may also have a function in the positioning of the nucleus in the cell in a cage-like bundle of fibrils. Intermediate filaments maintain cell homeostasis by dealing with stress in its various guises (Fig. 4-13), and the cell responds in a number of ways such as generating 'inclusion bodies', producing new and regenerating intermediate filements, or by reorganizing the filaments into variously sized bundles.

Interactions between intermediate filaments and microfilaments are mediated by plectin, a >500 kDa dumb-bell-shaped protein that can self-associate and in addition can bind at both ends of the hemidesmosomal protein $\alpha_6\beta_4$ integrin and probably to other junctional proteins.

Group 3 cytoskeletal fibres. A third group of cytoskeletal fibrillar elements also exists but the fibres are less easy to categorize. Several thick filaments occur as part of the cytoskeleton, such as myosin (myosin comprises 25% of cytoplasmic protein in striated muscle) and microtubules. Microtubules are cylindrical structures about 24 nm wide, comprising 13 globular elements composed of the heterodimer ($\alpha\beta$) tubulin. Stable microtubules occur in flagellae and cilia, while labile microtubules are found in structures such as the muscle spindle. Microtubules are involved in movement and cell motility, including intracellular transport as in axoplasmic flow. They are often required to switch very rapidly from an extending to a shrinking fibril, as in processes such as cell division where a multiprotein structure termed the kinetochore engineers chromosome segregation by facilitating chromosome attachment to spindle microtubules. Rapid growth to shrinkage in the microtubule is achieved through GTP to GDP hydrolysis, mediated through the Rho family of small GTPases; this is known as 'catastrophe', while the reverse change is termed 'rescue' (Fig. 4-14).

These changes occur in any microtubule-related function, including cell movement, and drugs such as colchicine (used in Behçet's disease) and taxol (proposed as prophylaxis for proliferative vitreoretinopathy) disrupt microtubule organization and inhibit cell motility and cytokinesis.

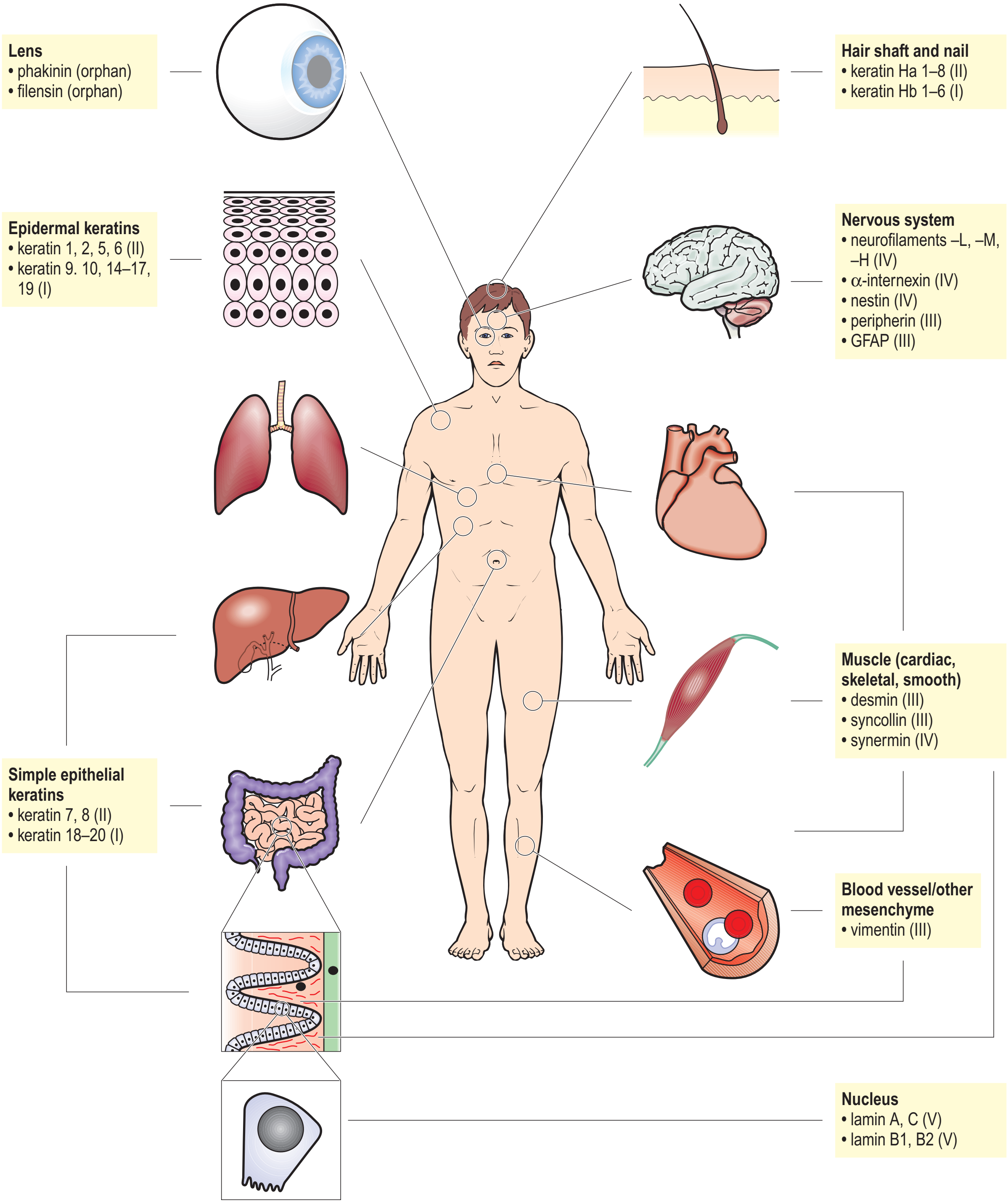
Lens
• phakinin (orphan)
• filensin (orphan)
Hair shaft and nail
• keratin Ha 1–8 (II)
• keratin Hb 1–6 (I)
Epidermal keratins
• keratin 1, 2, 5, 6 (II)
• keratin 9. 10, 14–17, 19 (I)
Nervous system
• neurofilaments –L, –M, –H (IV)
• α-internexin (IV)
• nestin (IV)
• peripherin (III)
• GFAP (III)
Muscle (cardiac, skeletal, smooth)
• desmin (III)
• syncollin (III)
• synermin (IV)
Simple epithelial keratins
• keratin 7, 8 (II)
• keratin 18–20 (I)
Blood vessel/other mesenchyme
• vimentin (III)
Nucleus
• lamin A, C (V)
• lamin B1, B2 (V)

FIGURE 4-12 Distribution of intermediate filament (IF) proteins in the human body. IF proteins include five major types (I–V) and a separate 'orphan' category (IF type and category are listed in parentheses). Lamins (type V) are found in the nucleus of most mammalian cells, whereas the remaining IFs (types I–IV) are cytoplasmic and expressed in a cell-tissue-selective manner. For each tissue, representative major IFs are listed for the principal cell type. (From Toivola et al., 2005, with permission from Elsevier.)

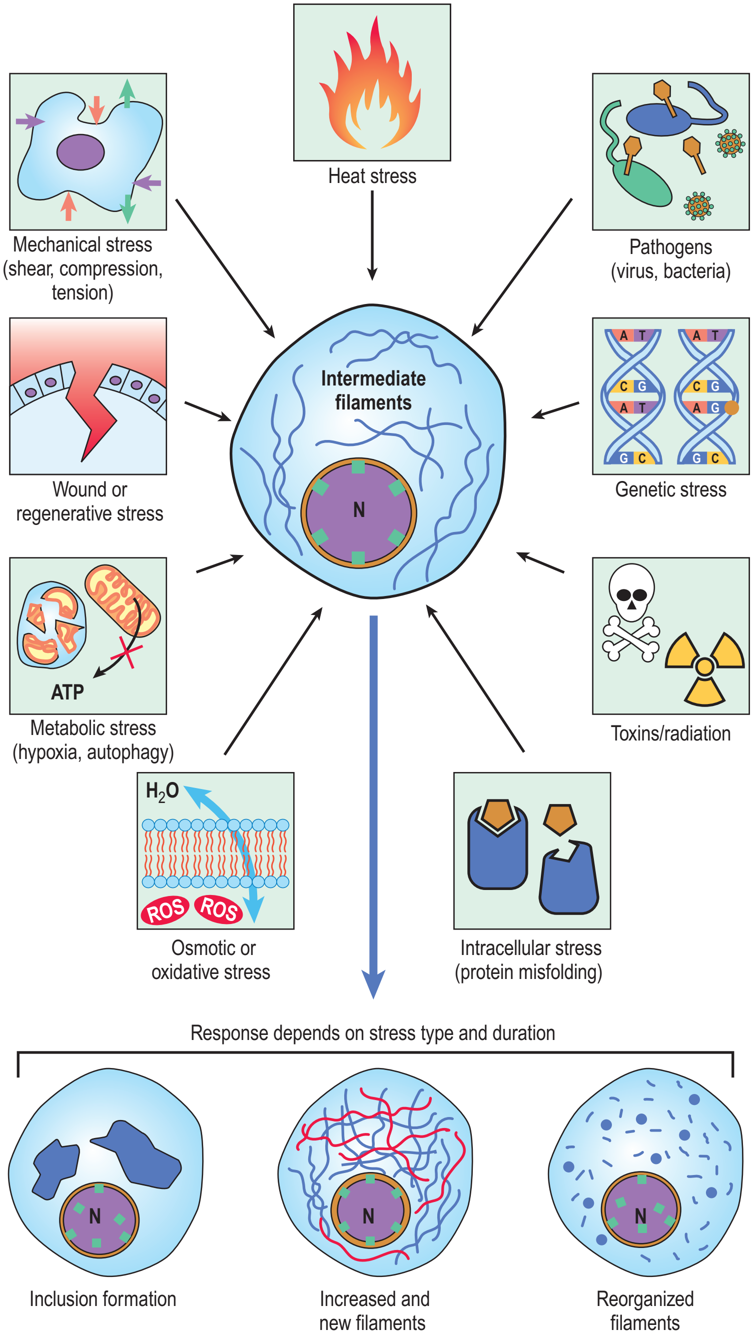
Heat stress
Mechanical stress (shear, compression, tension)
Pathogens (virus, bacteria)
Intermediate filaments
N
Wound or regenerative stress
Genetic stress
A T
C G
A T
G C
A T
C G
A G
G C
ATP
Metabolic stress (hypoxia, autophagy)
Toxins/radiation
H_2O
ROS
ROS
Osmotic or oxidative stress
Intracellular stress (protein misfolding)
Response depends on stress type and duration
N
Inclusion formation
N
Increased and new filaments
N
Reorganized filaments

FIGURE 4-13 Cells respond to stress through their intermediate filaments by (a) forming inclusion bodies, (b) making new filaments or (c) reorganizing their filaments. (From Toivola et al., 2010.)

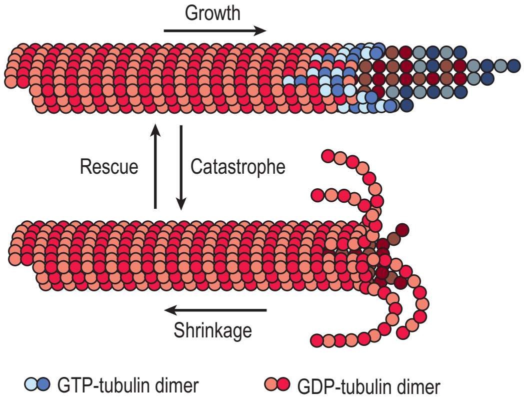

FIGURE 4-14 Microtubules switch between phases of growth and shrinkage generated by GDP-GTP-mediated incorporation and removal of tubulin dimers in the region of the tubulin cap. (From Bowne-Anderson et al., 2013.)

Group 3 fibres are thus a class of superfine filaments which combine to form an intracellular meshwork in which proteins do not exist in solution as previously surmised, but are attached to and transported along the filaments with other structures such as 'free' ribosomes and small vesicles (polysomes). This arrangement has special relevance to highly organized cells such as lens fibres. Transport of molecules along microtubules involves two families of molecular motors, kinesins and dyneins.

Certain discrete cellular structures are composed of microtubules, such as the centriole (the microtubular organizing centre, a cylindrical structure comprising nine groups of triplet microtubules) and the mitotic spindle. Cilia are remarkably constant structures in eukaryotic cells, composed of nine peripheral and two central bundles of three fused microtubules. Movement occurs by sliding of the outer arm of dynein along the core of tubulin.

Septins. Septins are a family of proteins expressed in all cell types which are integrators of functions of other microfilaments such as the ATPase role of actin and the dynamic functions of microtubules. There are 13 septins in humans, mutations of which are known to cause specific diseases. Septins are important in cells with high migratory functions, such as lymphocytes. In addition, septin 8, which lies on chromosome 5q31, is associated with retinal degeneration. Their main function in the cell appears to be supportive, e.g. as a scaffold for protein recruitment and as an intracellular compartment organizer, and they are called septins for this reason. For example, several septins in one cell combine to form a hexameric filament of around 25 nm long which acts as an intracellular 'septa' (barrier to diffusion) (Fig. 4-15).

In conclusion, therefore, there are extensive interconnections between the four filamentous systems orchestrating many fundamental types of cell behaviour. In addition, the cytoplasm also contains several storage products such as glycogen granules, lipid droplets and melanin in melanosomes, often inside endosomes, which, as indicated above, are transported around the cell by MT-associated proteins and made to function by active fibre contraction as in actin-mediated exocytosis (see Box 4-2). With age, some cells, such as the RPE cell, accumulate unwanted intracellular bodies such as lipofuscin granules.

Intracellular signalling mechanisms

As stated above, cells respond to external stimuli by means of cell surface receptors, which convert the external stimulus to a series of intracellular signals (second messengers) directed towards specific cellular functions such as protein transcription for growth control or ion-channel gating in neural responses (see Ch. 5, pp. 288–295).

Second messenger systems are based on a network of reactions involving an agonist, a receptor and an interacting set of coupled proteins. Cyclic adenosine monophosphate (cAMP) is the archetypal second messenger, and the result of such a response is the phosphorylation of a regulatory intracellular protein via a kinase, as for instance serine-threonine kinase or tyrosine protein kinases, which demonstrate great substrate specificity. In contrast the phosphatases, such as protein phosphatase-1 (PP1), which usually bring the kinase reaction to a close, are much more widespread in their range of activity and, consequently have many fewer genes expressed.

During the last 30 years, a family of ligand receptors known as WNTs has been shown to have widespread roles in embryonic development, cancer and wound healing. These molecules define the interaction between matrix and cells via intra- and inter-cell signalling and are essential for developmental processes

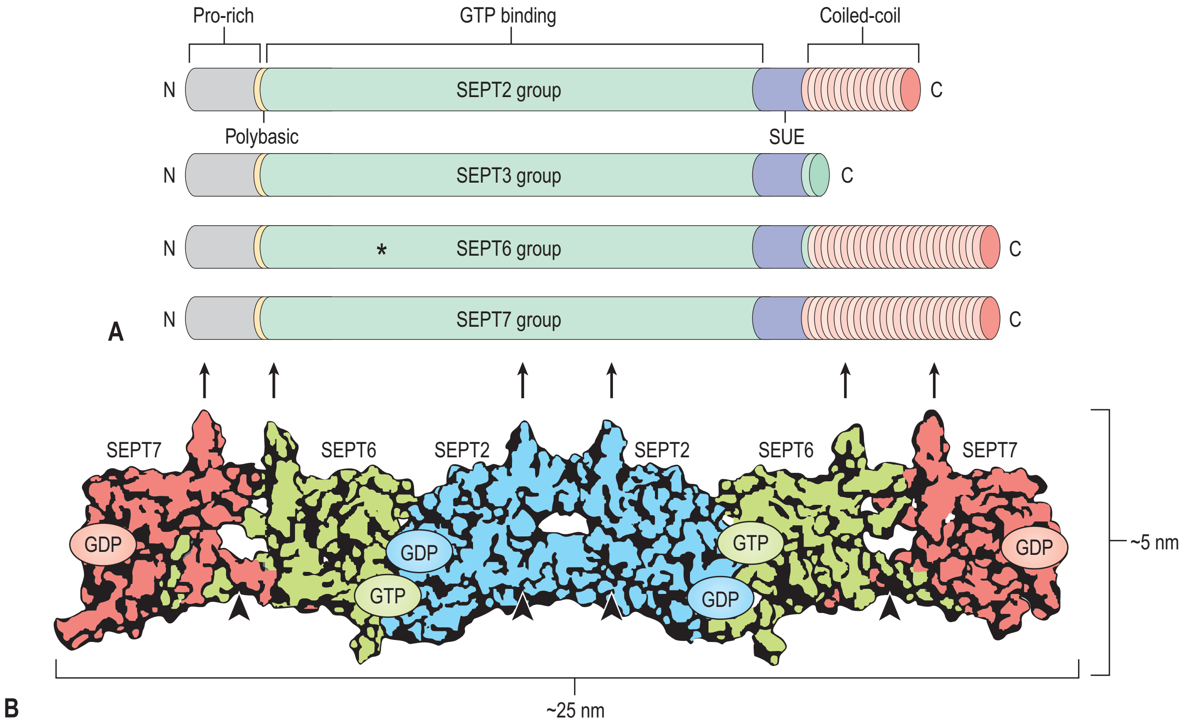

FIGURE 4-15 There are 13 human septins classified into four groups (SEPTS 2,3, 6 and 7). They have three conserved domains: a phosphoinositide-binding region; a GTP-binding region; and the septin unique element (SUE). (From Mostowy and Cossart, 2012.)

such as patterning. The name WNT derives from the discovery that the int1 (integration gene) and the (Wg) Wingless gene in *Drosophila* were identical and there are now known to be several Wnt genes with multiple effects on extracellular matrix proteins.

Additional content available at https://expertconsult.inkling.com/.

THE EXTRACELLULAR MATRIX

Cells exist within a structural framework, the extracellular matrix (Box 4-6), which is secreted by the cells such as the myofibroblast during wound healing and consists of several classes of macromolecules, the most abundant of which is collagen (accounting for 30% of total protein in the organism) (Fig. 4-16 and Video 4-1).

Genetic and protein sequence analysis of extracellular proteins has shown that, despite the great variety of matrix molecules, extensive sequence homology exists between them with recurring structural motifs (Box 4-6). For instance, the 'epidermal growth factor' motif appears to be present in several apparently different extracellular matrix proteins.

Extracellular matrix proteins determine the structural nature of the tissue

Tissues are defined by their extracellular matrix constituents. For instance, the cornea contains type I collagen filaments, which do not form fibrils of diameter greater than 5 nm; this is important in corneal transparency. The vitreous contains an isoform of type IX collagen, which is different in certain characteristics from cartilage type IX collagen.

Collagen fibres are composed of a triple helix of three polypeptide chains, of which there are 46 different types and which assemble into 29 distinct collagen molecules (Table 4-2).

At least 22 of these have been detected in the developing and adult eye. The eye and cartilaginous tissues share six types of collagen not found very frequently in other tissues; indeed, the collagen triple helix (Fig. 4-17 and Box 4-6) has been found as a domain of

WNT signalling pathways are central to many developmental and cell biological functions (eFig. 4-5), including planar cell polarity (PCP) mediated by Wnt /Frizzled interactions; β-catenin-dependent signalling which is extensively involved in cell–cell interaction and cell adhesion and includes interaction of Wnt with the low-density lipoprotein receptor (LRP)5,6; and Wnt-calcium (Wnt-Ca^{2+}) signalling which is important calcineurin /NFAT signalling and is involved in cell fate and cell migration. This latter pathway is activated in dendritic cells but not macrophages and induces apoptosis in DC (see Ch. 7, p. 373–380).

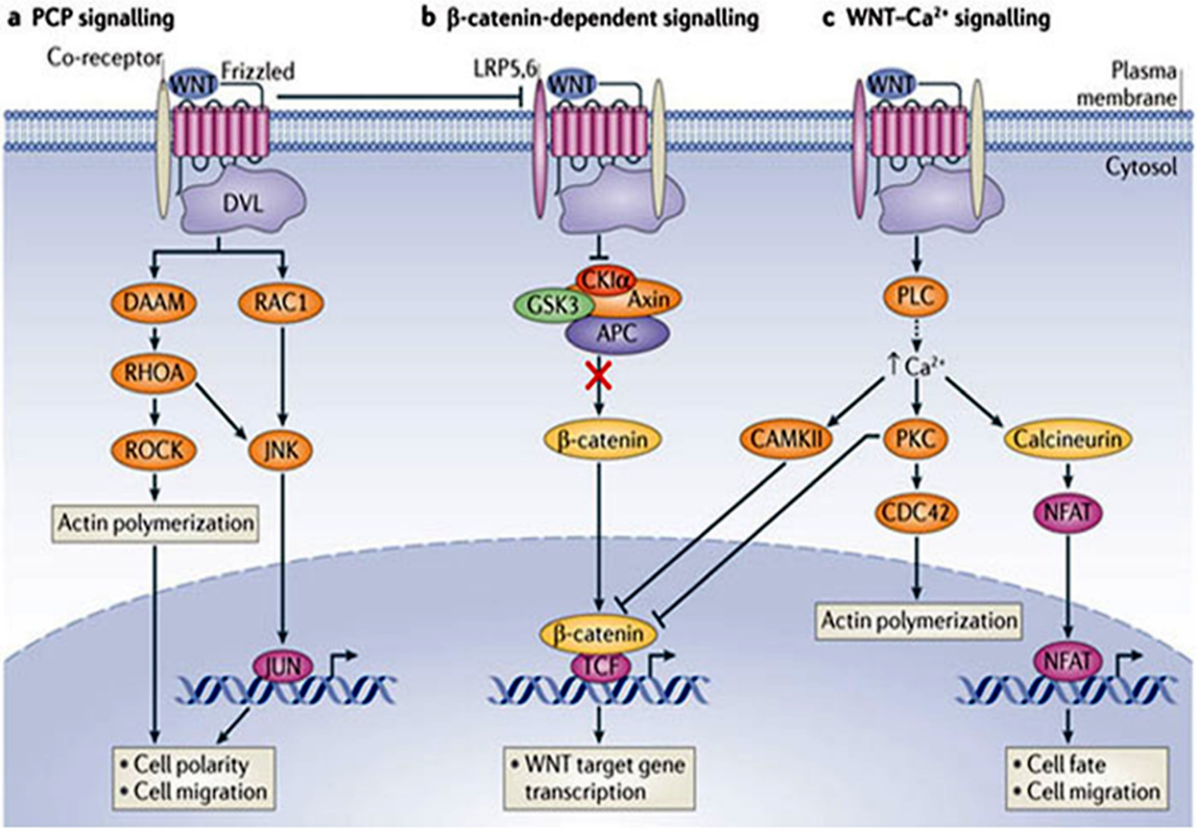
a PCP signalling
b β-catenin-dependent signalling
c WNT–Ca²⁺ signalling
Co-receptor
WNT
Frizzled
LRP5,6
WNT
WNT
Plasma membrane
Cytosol
DVL
DAAM
RAC1
RHOA
ROCK
JNK
Actin polymerization
JUN
• Cell polarity
• Cell migration
CKIα
GSK3
Axin
APC
β-catenin
β-catenin
TCF
• WNT target gene transcription
PLC
↑Ca²⁺
CAMKII
PKC
Calcineurin
CDC42
NFAT
Actin polymerization
NFAT
• Cell fate
• Cell migration

eFIGURE 4-5 Wnt signalling involves the co-receptor Frizzled to promote actin polymerization and cell polarity (planar cell polarity, PCP), the low-density lipoprotein receptor (LRP) 5,6 to mediate β-catenin signalling to the nucleus, and Wnt-Ca^{2+} signalling, which activates the calcineurin-NFAT pathway determining cell fate. (From Niehrs, 2012.)

BOX 4-6 EXTRACELLULAR MATRIX PROTEINS

Extracellular matrix proteins appear to be made up of building blocks (‘structural motifs’), which are protein domains with extensive homology to existing protein structures. For instance, three fibronectin domains are reproduced in many other proteins; the epidermal growth factor-like domain also appears in modular form in proteins such as plasminogen; and the arginine–glycine–aspartate (RGD) cell adhesion site is present in many molecules (**A**).

Fibronectin is a dimer containing discrete domains for the attachment of other molecules (**A**). The modular structure of fibronectin shows that it consists of 12 type I modules (rectangles), two type II modules (violet ovals) and 15–17 type III modules (ovals). The alternatively spliced domains IIIB, IIIA and the V region are shown in yellow. Binding domains for fibrin, collagen, cells and heparin are indicated; dimer forms via cysteine pair at the C-terminus (SS).

Laminin is a cross-shaped trimer in which the B1, B2 and A chains form a triple helix in the stem of the cross. Laminin also has several discrete domains for attachment of various molecules. In addition, it contains a cryptic cell adhesion-binding site, which becomes exposed after partial proteolysis (**B**). Fibrillar collagen is organized in tissues by association with smaller non-fibrillar collagens, e.g. type I with type XII collagen, and type II with type IX collagen (**C**).

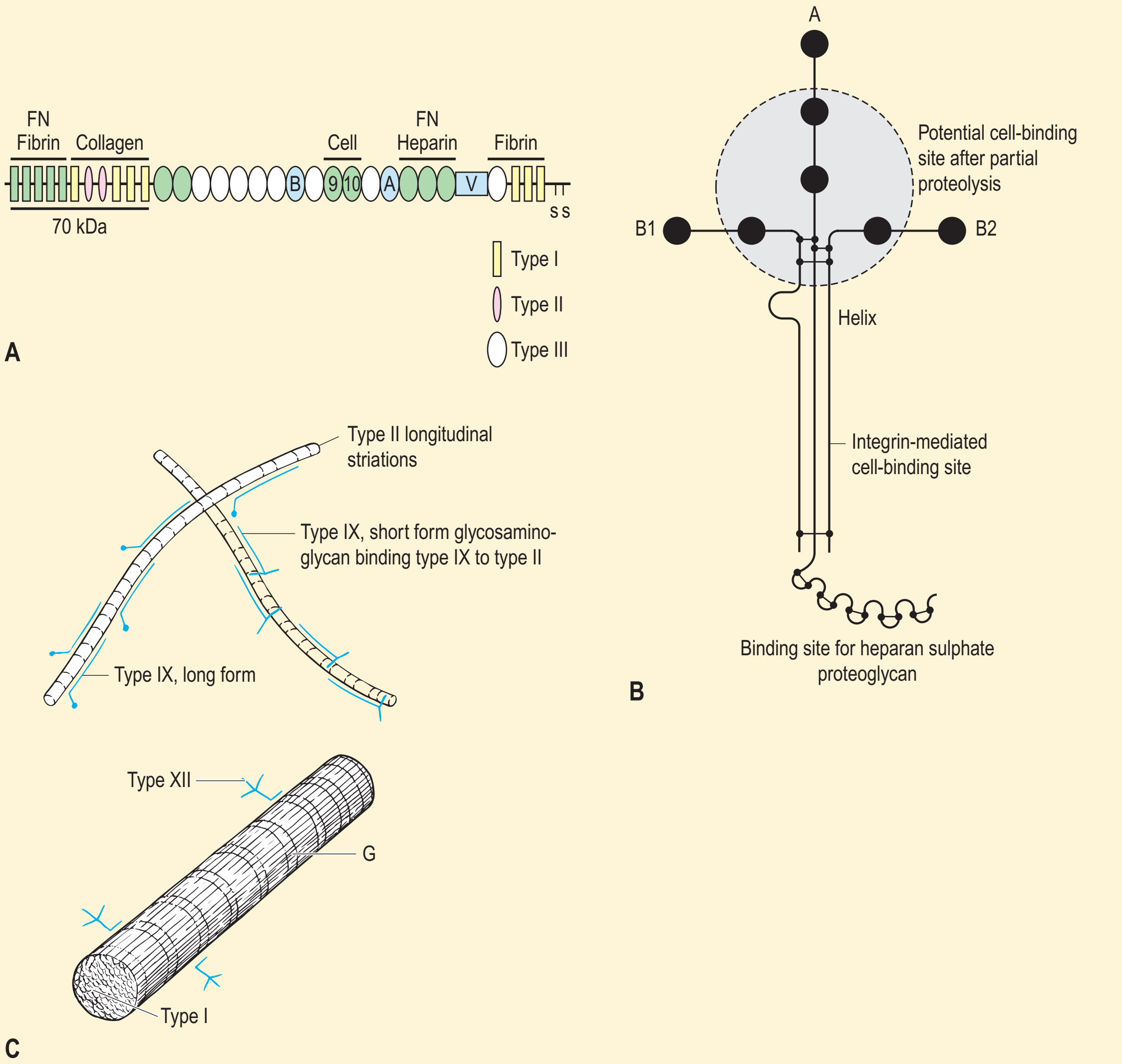
FN
Fibrin
Collagen
Cell
FN
Heparin
Fibrin
B
9
10
A
V
S S
70 kDa
Type I
Type II
Type III
A
A
Potential cell-binding site after partial proteolysis
B1
B2
Helix
Integrin-mediated cell-binding site
Binding site for heparan sulphate proteoglycan
B
Type II longitudinal striations
Type IX, short form glycosamino-glycan binding type IX to type II
Type IX, long form
Type XII
G
Type I
C

(Part A reproduced from Mao and Schwarzbauer, 2005, with permission from Elsevier; Parts B and C courtesy of Yamada and Miyamoto, 1995.)

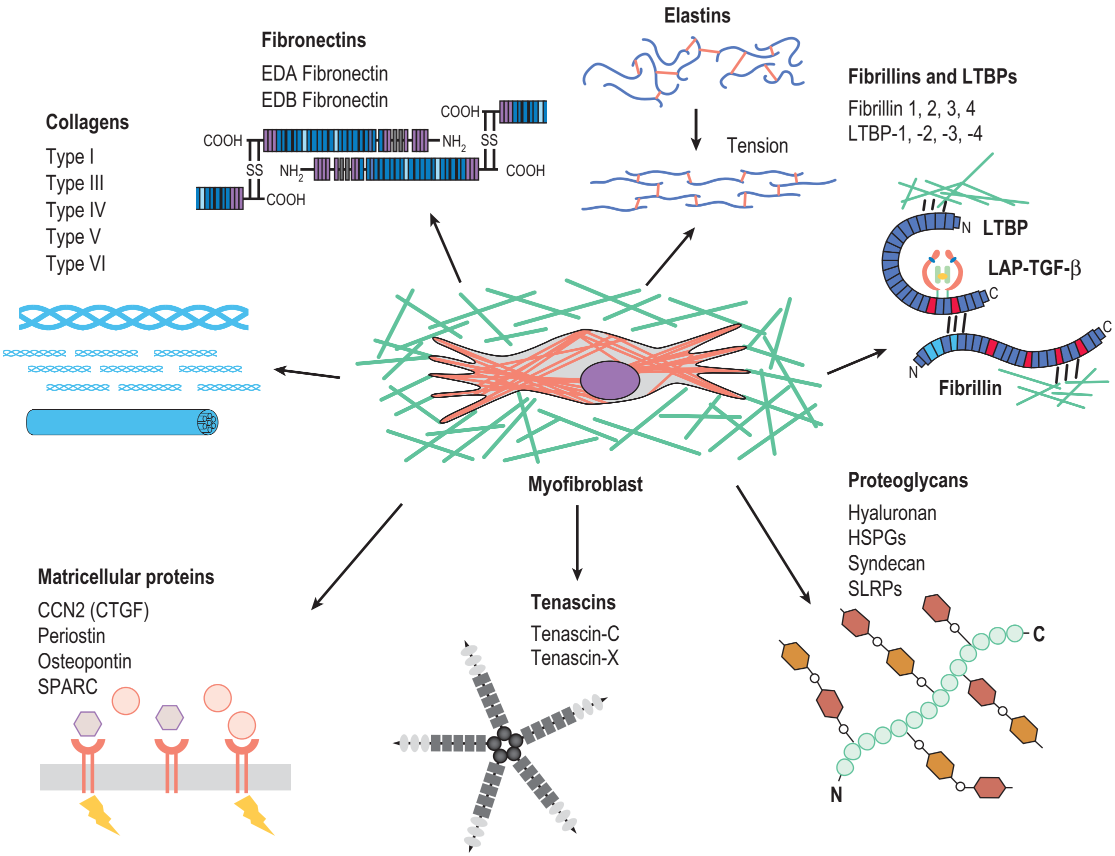
Elastins
Tension
Fibronectins
EDA Fibronectin
EDB Fibronectin
COOH
SS
NH2
COOH
NH2
SS
COOH
COOH
Fibrillins and LTBPs
Fibrillin 1, 2, 3, 4
LTBP-1, -2, -3, -4
Collagens
Type I
Type III
Type IV
Type V
Type VI
N
LTBP
LAP-TGF-β
C
C
N
Fibrillin
Myofibroblast
Proteoglycans
Hyaluronan
HSPGs
Syndecan
SLRPs
C
N
Matricellular proteins
CCN2 (CTGF)
Periostin
Osteopontin
SPARC
Tenascins
Tenascin-C
Tenascin-X

FIGURE 4-16 The myofibroblast matrix. Schematic of some of the ECM molecules relevant to tissue fibrosis. The myofibroblast (centre, with red stress fibres containing α-smooth muscle actin) lies enmeshed in its ECM (green). Components of the ECM are depicted (clockwise, from the 12 o'clock position): elastins, fibrillins and LTBPs, proteoglycans, tenascins, matricellular proteins, collagens and fibronectins. The myofibroblast encounters, signals, and modulates the expression of these various components as outlined in the text. (From Klingberg et al., 2013.)

TABLE 4-2 **Some collagen types**

Fibrillar	Anchoring	FACIT**	Transmembrane	Beaded filament	Network-forming	Multiplexin
I, II, III, V XI, XXIV XXVII	VII	IX, XII XIV, XVI XIX, XX XXI, XXII	XIII XVII XXIII XXV	VI, XXVI XXVIII	IV, VIII, X	XV, XXVI XXVIII

**FACIT, Fibril-associated collagen with interrupted triple helices.

many proteins and it has been suggested that all such proteins should be included in the collagen family. Collagen types are determined by the combination of the three types of chain forming the α helix core. For instance, type I collagen triple helix is made up of two unique α_1 chains and a unique α_2 chain, coded as $[\alpha_1(I)]_2\ \alpha_2(I)$; type II collagen is composed of three identical unique α_1 chains, $[\alpha_1(II)]_3$. Each collagen therefore has a set of unique chains that make up the triple helix.

In certain collagens, particularly the non-fibrillar collagens such as types IV and IX, the protein is

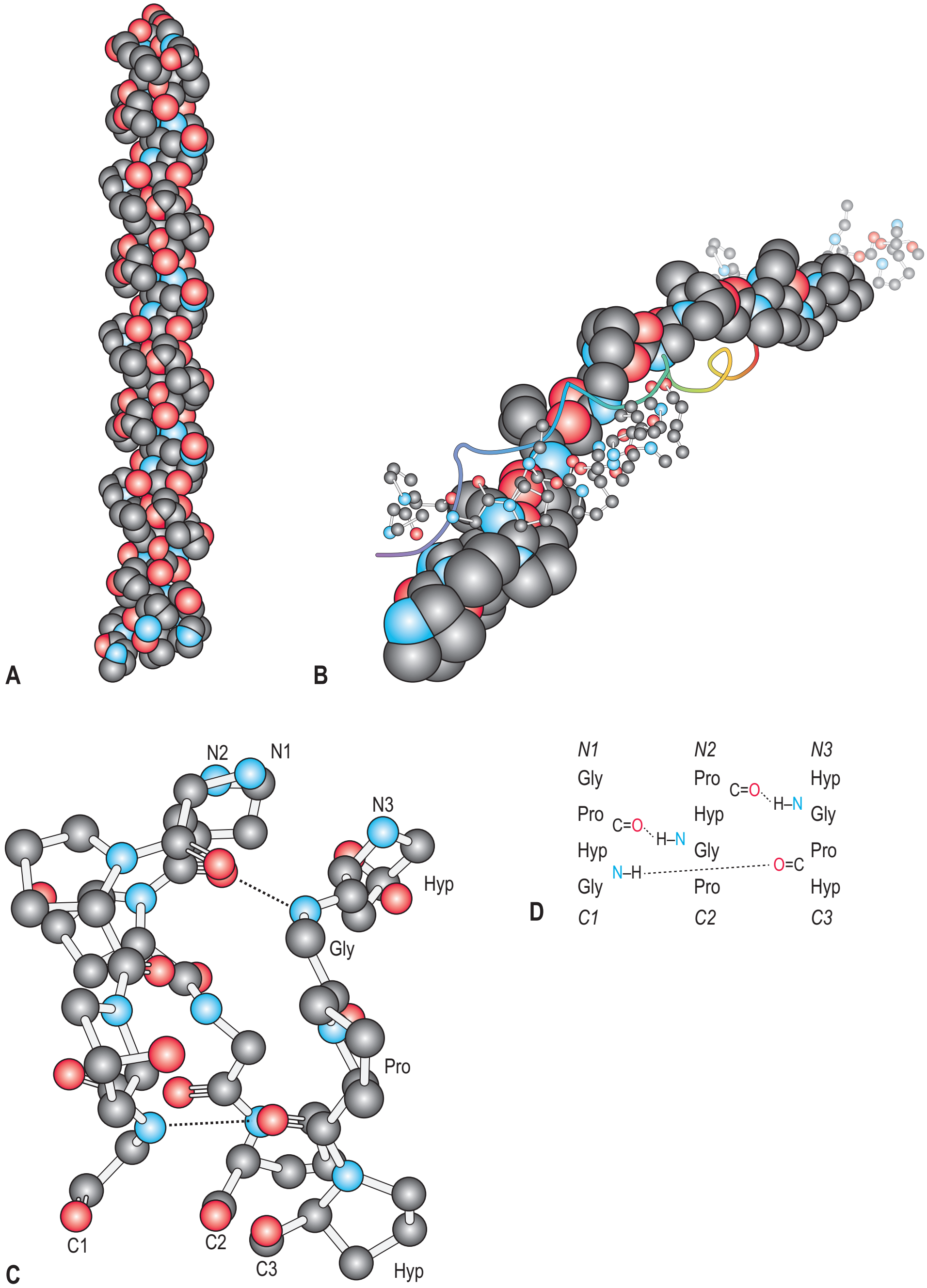
A
B
N2
N1
N3
Hyp
Gly
Pro
C1
C2
C3
Hyp
C
N1
N2
N3
Gly
Pro
Hyp
C=O
H–N
Pro
Hyp
Gly
C=O
H–N
Hyp
Gly
Pro
N–H
O=C
Gly
Pro
Hyp
D
C1
C2
C3

FIGURE 4-17 Overview of the collagen triple helix. (**A**) First high-resolution crystal structure of a collagen triple helix, formed from (Pro-HypGly)4–(ProHypAla)–(ProHypGly)5 [Protein Data Bank (PDB) entry 1cag (19)]. (**B**) View down the axis of a (ProProGly)10 triple helix [PDB entry 1k6f (7)] with the three strands depicted in space-filling, ball-and-stick, and ribbon representation. (**C**) Ball-and-stick image of a segment of collagen triple helix [PDB entry 1cag (19)], highlighting the ladder of interstrand hydrogen bonds. (**D**) Stagger of the three strands in the segment in panel **C**. (From Shoulders and Raines, 2009.)

composed of short segments of triple helix (COL1, -2, -3, etc.) interspersed with sections of non-collagenous (NC1, NC2, etc.) protein. These are also sometimes referred to as FACIT collagens (fibril-associated collagens with interrupted triple helices) (Table 4-2). These proteins usually act as bridges or networks for binding other proteins and forming complex protein aggregates such as basement membranes. Collagens in essence are the structural components that hold organs and tissues together and the mechanism by which they form is interesting: the three polypeptide chains form a monomer by self-assembly from a small nucleus in a zipper-like fashion, much in the way that crystallization occurs, and the formation of fibrils from monomers is also entropy driven. Indeed the semi-crystalline packing of collagen fibres, achieved through lysine hydroxylysine cross-links, confers tensile strength to tissues.

A subcategory of collagenous proteins is included as components of transmembrane proteins, including types XIII, XVII, XXIII and XXV (Table 4-2), and remain following cleavage of the soluble fragment of the cytokine or adhesion molecule by enzymes known as sheddases. They exist as homotrimers of a collagen-specific α-chain. Some transmembrane collagen-like molecules occur as specific receptors such as the macrophage scavenger receptors. Sheddases are of several types and include the ADAMTs (proteinases of the adisintegrin and metalloproteinase family), which are involved for instance in the shedding of molecules such as soluble tumour necrosis factor-α-converting enzyme (TACE) from leucocytes (see Ch. 7). This general process is important in establishing the soluble regulatory constituents of particular extracellular matrices.

The transmembrane collagen XVII (originally described as bullous pemphigoid antigen 180, BP180) is important in cellular adhesion because it binds α_6 integrin and laminin 5 extracellularly and β_4 integrin, plectin and BP230 intracellularly. Degradation products of collagens, such as the non-fibrillar collagens XVIII and XV, which yield the anti-angiogenic products endostatin and restin after cleavage, have major roles in regulating cell function through inhibition of matrix metalloproteinases.

The eye contains a wide variety of different collagens. For instance, the cornea contains types I, V and VI in the stroma, while types IV and VII are present in the subepithelial layer. In addition, type XVI transmembrane linker collagen has been found in the basal epithelial cell matrix. The iris contains collagen types I, III and IV, while the zonule contains type IV. The lens contains only type IV, while the vitreous contains collagens II, IX and XI, complexed to the extracellular matrix protein fibrillin. This complex has an important role in vitreous matrix organization.

Certain types of collagen are unique to ocular tissue in that their structure has been modified. Examples include type VIII in Desçemet's membrane, types II and IX in vitreous (both similar but not identical to cartilage collagens) and type III in the distensible tissue of the choroid. Several of the newer collagens contain domains similar to type IX, and it has been suggested that these may represent a subfamily of type IX collagens (Table 4-2). The more recently described types XV, XVIII and XIX are located in basement membranes and may have a role in the formation of blood vessels.

Collagen does not simply fill the space between cells but engages four types of specific receptors on cells: integrins, discoidin domain receptors (DDRs 1 and 2, a subfamily of receptor tyrosine kinases involved in signalling to the cell), glycoprotein VI (GPVI) receptor (on platelets) and leucocyte-associated immunoglobulin-like receptor-1 (LAIR-1). DDRs and integrins are exquisitely specific in the collagen domain to which they bind and their main function is to maintain tissue stability during intra-tissue cell migration, while GPVI and LAIR-1 are involved in haemostasis after tissue injury.

Elastin is the second major insoluble protein of the extracellular matrix. While collagen occurs in all tissues, elastin is present only in deformable tissues such as blood vessel walls, lung parenchyma and the zonule of the lens. Unlike collagen, elastin does not contain any methionine residues and is therefore separable from collagen after digestion with cyanogen bromide. Elastin is formed from soluble tropoelastin monomers, secreted from fibroblasts and smooth muscle cells, which initially self-associate through hydrophobic domains in a process termed co-acervation *in vitro* which *in vivo* corresponds to coalescence into spherical globules of 1–6 μm size. This process of self association aligns lysine residues to form cross-links, via allysine intermediates, desmosine and isodesmosine, a process which is facilitated by the elastin

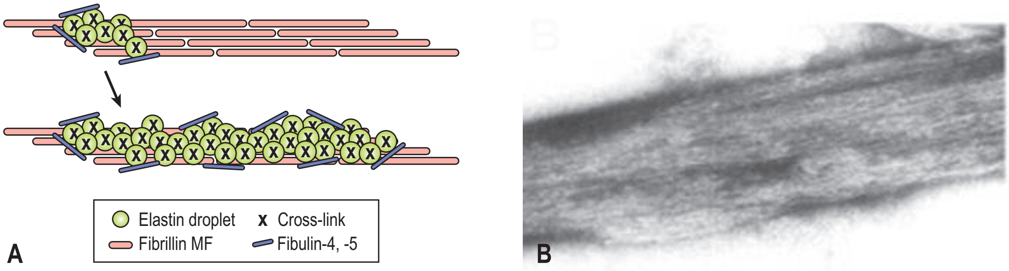

FIGURE 4-18 Elastin 'droplets' are deposited on fibrillin microfibrils and cross-linked to the fibrillin molecule via lysyl oxidase in association with fibulin molecules. (From Muiznieks and Keeley, 2012.)

co-acervated droplets depositing onto fibrillin microfibrils together with fibulin (see below) (Fig. 4-18). The final fibre is 90% elastin and 10% fibrillin and cross-links are mediated via a set of lysyl oxidase enzymes, mutations in which have been associated with the pseudo-exfoliation syndrome, the commonest cause of secondary glaucoma and a risk factor for successful cataract surgery due to the role of elastin in formation of the zonule. Elastin has little order to its tertiary structure; instead, it forms a random coil, which tends to become more ordered during deformation.

As for collagen, elastin degradation products have a role in regulating cell behaviour in the matrix and under certain circumstances may be pathogenic, e.g. in promoting tumour invasion and angiogenesis. Even smaller molecules derived from a range of matrix proteins (matrikines) have regulatory activity for connective tissue cells; for instance, the tripeptide GHK, which has stimulatory activity for several cell types and promotes collagen biosynthesis and wound healing overall.

Fibrillins are a family of proteins serving as components of extracellular elastin microfibrils that are responsible for the biomechanical properties of tissues. Cysteine-rich glycoproteins, composed of multiple repeats of a Ca^{2+}-binding epidermal growth factor-like domain, are secreted in a proform and polymerize extracellularly. They are present in vitreous (see above) and are an important component of the zonule. Fibrillins have a structural role in long-range elasticity of tissues and are a component of elastic fibres (see above). They also have a role in the fine-tuning of growth factor signals such as those involving transforming growth factor-β (TGF-β) and particularly latent TGF-β-binding proteins (LTBP) involved in morphogenesis. Mutations in *FN1* cause a dominant form of Marfan syndrome.

Tenascins are extracellular matrix proteins best described as adhesion modulating, matricellular proteins which do not form major fibrillary structures such as collagen or elastin. They are module-built proteins comprising EGF-like repeats, fibronectin-type III repeats and fibrinogen domains (Fig. 4-19). Several forms are described (tenascin-C, -X, -W, etc.) and phenotypes in knockout mice are broadly normal although there may be some subtle defects in wound healing. Tenascin-Y appears to be restricted to neural tissue, while mutations in tenascin-X have been linked to Ehlers–Danlos syndrome, in conjunction with known defects in pro-collagen and elastin biosynthesis. Tenascins are widely distributed and appear to have an anti-adhesive role, particularly antagonizing the effects of fibronectin. Interestingly, fibronectin is found in the anterior lens capsule, while tenascin is found in the posterior capsule.

Laminins are an integral part of basement membranes and have a characteristic hetero trimeric cross-shaped structure composed of an α chain, a β and a γ chain constructed to shape and an α coiled rod (see Box 4.6 and eFig. 4-6).

There are 16 isoforms variously made up for one α (α1–α5), 3β (β1–β3) and 3γ (γ1–γ3) chains that are specific for different cell types (e.g. kalinin in epithelial cells) and are named according to their chain components. For instance laminin-332 (i.e. α3, β3 and γ2 chains) is specific for epithelial cells and binds to the $\alpha_3\beta_1$ and $\alpha_6\beta_4$ integrins.

Additional content available at https://expertconsult.inkling.com/.

The infrastructure of the basement membranes is composed of type IV collagen in a highly organized

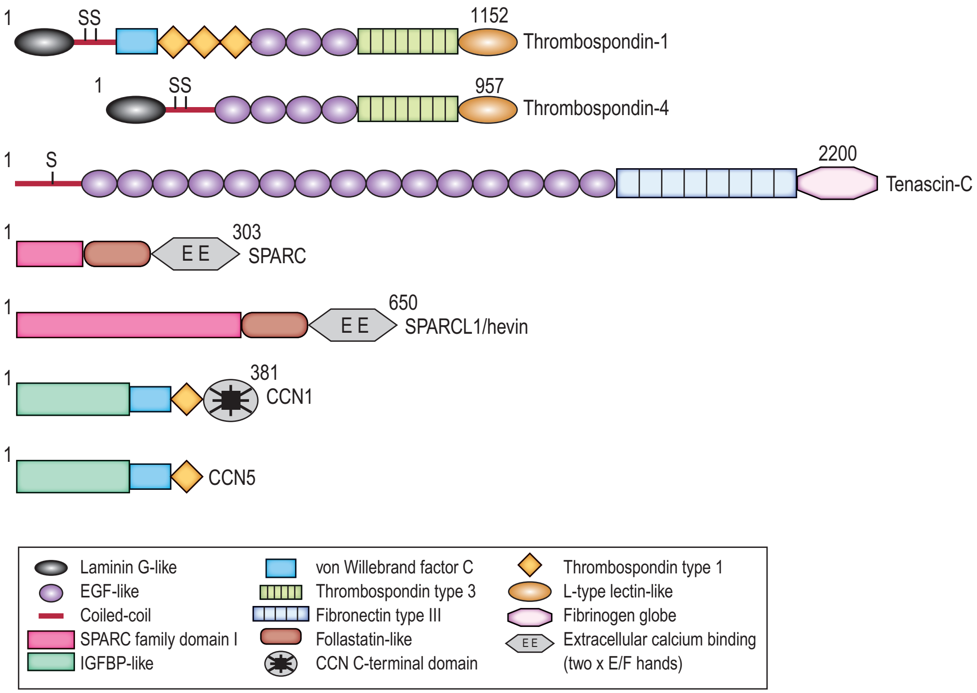
1
SS
1152
Thrombospondin-1
1
SS
957
Thrombospondin-4
1
S
2200
Tenascin-C
1
E E
303
SPARC
1
E E
650
SPARCL1/hevin
1
381
CCN1
1
CCN5
Laminin G-like
EGF-like
Coiled-coil
SPARC family domain I
IGFBP-like
von Willebrand factor C
Thrombospondin type 3
Fibronectin type III
Follastatin-like
CCN C-terminal domain
Thrombospondin type 1
L-type lectin-like
Fibrinogen globe
E E
Extracellular calcium binding (two x E/F hands)

FIGURE 4-19 Schematic diagrams of the domain organization of representative members of the protein families discussed in this chapter. Not to scale. (From Mosher and Adams, 2012.)

lattice network to which the complex of laminin–nidogen is also bound. The proteoglycan then acts as a space-filling molecule in the basement membrane (Fig. 4-20). Cells are bound to the basement membrane via anchoring fibrils containing type VII collagen and transmembrane proteins such as types XI and XVIII collagen.

Most laminins are produced by the cells adhering to the basement membrane in which they are found; interestingly, the laminins of the internal limiting membrane (ILM) of the retina are synthesized by ciliary body and lens cells. ILM laminin and fibronectin are substrates for ocriplasmin, a therapeutic enzyme with potential for use in vitreoretinal disease.

Non-structural proteins

Many other proteins are distributed throughout the extracellular matrix, which, although not having a direct structural role, have important functions in cell–matrix and cell–cell interactions. Many of these proteins are engaged in self-organization with other ECM proteins (see p. 178) to form fibrillar networks. Probably the best known of these is fibronectin, a 250 kDa heterodimer that has multiple cell and molecule binding domains (see Box 4-5).

Additional content available at https://expertconsult.inkling.com/).

For instance, one of the type III fibronectin domains contains the ubiquitous cell adhesion domain Arg-Gly-Asp-Ser (RGDS in single-letter amino acid nomenclature), found in many other proteins. Such domains bind to other proteins which traverse the cell membrane, anchoring the cell to the matrix, including the integrins. Multiple forms of fibronectin occur by alternate splicing. Two major forms exist: plasma fibronectin, secreted by hepatocytes, and cellular fibronectin, secreted by fibroblasts and forming the ECM fibrillar network. Fibronectin may be involved in reverse integrin-mediated cell signalling and, through the actin-binding protein profilin, may regulate stress fibre formation in endothelial cells and fibroblasts. Fibronectin also has high-affinity binding sites for fibrin(ogen), thus promoting incorporation of fibrin into the extracellular matrix during wound healing. Fibrin binds

Fibronectin. This protein probably initiates the process by binding integrin receptors on the cell surface and entraining the mechanical forces of the cell through transmembrane connections with the dynamic acting cytoskeleton. Matrix assembly is critically dependent on interactions between matrix proteins and the cells which produce them. Such proteins self-organize (see p. 158) through specific domains on cell surface protein receptors such as $\alpha_5\beta_1$ integrin which connect to the intracellular cytoskeleton proteins such as actin via linking proteins such as talin and vinculin. Cell contractility then fashions the shape, contour and architecture of the matrix. Once in place, fibronectin then contributes to the assembly of other matrix proteins through its many molecule-specific domains (eFig. 4-7).

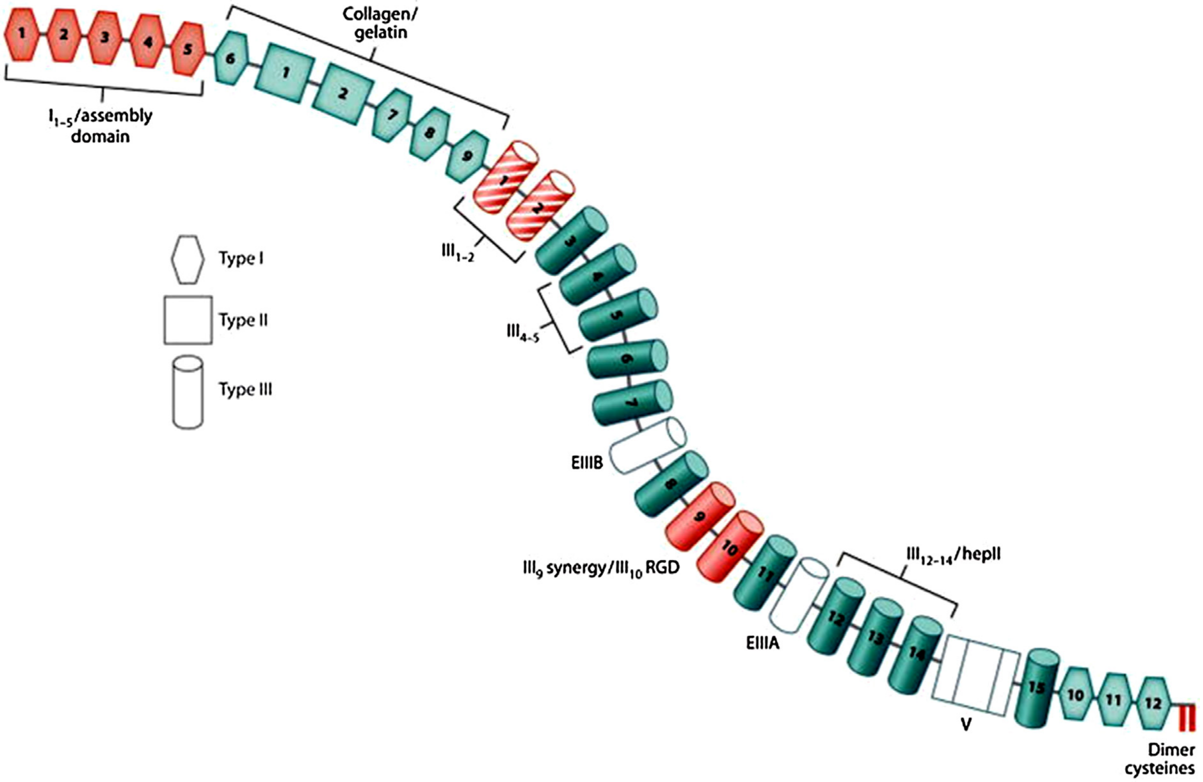
1
2
3
4
5
6
1
2
7
8
9
Collagen/ gelatin
I_{1-5}/assembly domain
Type I
Type II
Type III
III_{1-2}
1
2
3
4
5
6
7
III_{4-5}
EIIIB
8
9
10
11
III_9 synergy/III_{10} RGD
EIIIA
12
13
14
III_{12-14}/hepII
V
15
10
11
12
Dimer cysteines

eFIGURE 4-7 The fibronectin molecule: Each subunit comprises three types of repeat involved in Fn assembly (including the cell-binding RGD domain), the collagen-binding domains, and glycosaminoglycan-binding domains (heparin and syndecan). (From Singh et al., 2010.)

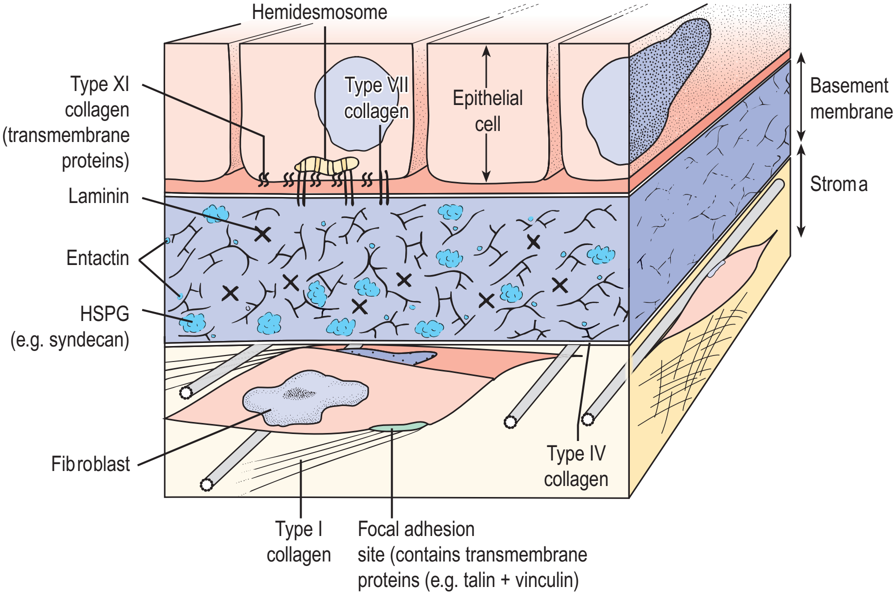

FIGURE 4-20 Structure of basement membrane.

growth factors such as fibroblast growth factor 2 and vascular endothelial cell growth factor, providing sources for angiogenic stimulation.

Thrombospondins (Fig. 4-19) are a group of extracellular matrix proteins that were first identified as platelet release proteins but that are now known to be secreted by endothelial cells and other cell types. Together with SPARC (secreted protein, acidic and rich in cysteine), tenascins and CCN proteins they are considered to be adhesion-modulating proteins. Five forms of thrombospondin exist, two of which appear to be alternately spliced forms of the parent molecule. Like extracellular matrix proteins, thrombospondins are composed of building blocks, each with specific cellular and molecular adhesive properties. Thrombospondins are involved in the regulation of angiogenesis, for instance in preventing angiogenesis associated with ocular inflammatory disease (see Ch. 9). In this regard, they form complexes with a ubiquitous plasma protein, histidine-rich glycoprotein (HRG), which has many regulatory functions in wound healing, cell migration and immune cell function. In addition, thrombospondins are better regarded as matricellular proteins that regulate a variety of processes, including cell adhesion and collagen fibrillogenesis. SPARC (Fig. 4-19) is a component of basement membranes but appears to be essential only for normal lens physiology since SPARC-deficient mice develop age-related cataract and lens rupture at 6 months of age but have no other phenotype. Less is known about CCN (for CYR61/CTGF/NOV) proteins (Fig. 4-19), apart from von Willibrand factor, which is involved in haemostasis.

Several other proteins are present in the extracellular matrix, such as proteases (e.g. plasminogen) and their inhibitors (plasminogen activator inhibitor 1 (PAI-1), α_2-macroglobulin, etc.). PAI-1 is present at high concentrations around various cells in the quiescent state and is considered important in the regulation of cell migration by controlling the level of cell-associated plasminogen activator required to initiate the degradation of basement membrane proteins. Other important proteins include the matrix metalloproteinases (MMPs) and their inhibitors (TIMP-1 to -3). Some of these proteins are secreted by the cells themselves, while others are synthesized predominantly in the liver and reach the extracellular matrix via the plasma.

Some proteins have their action at a distance from the cell, such as fibrillin, a 350 kDa microfibrillar protein that is involved in the assembly of elastin fibrils (Fig. 4-18). Fibrillin is found in the tertiary vitreous and the zonule (see previous section). Mutations in fibrillin are found in Marfan syndrome, a disease of elastic tissue in which dislocation of the lens is a central feature.

Glycosaminoglycans occur in the extracellular matrix bound to core proteins as proteoglycans

Glycosaminoglycans (GAGs) occur in a variety of forms, essentially based on a repeating disaccharide structure (Box 4-7).

The prototype proteoglycan was described in relation to cartilage, in which a series of proteoglycans are linked to a hyaluronic acid backbone (Fig. 4-21). This, however, is not relevant to ocular proteoglycans; for instance, hyaluronic acid in the vitreous is not associated with other GAG-containing proteoglycans but is linked to collagen type IX, which in this situation is regarded as a proteoglycan.

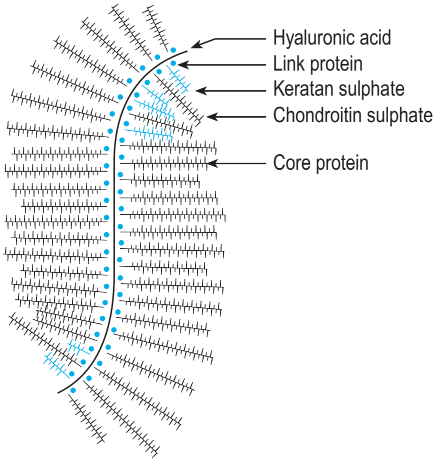

FIGURE 4-21 Artist's impression of hyaluronic acid molecule. (After Rosenberg, 1975.)

In addition, keratan sulphate proteoglycan occurs in the cornea in the absence of hyaluronan, where it appears to regulate collagen fibril diameter (see below). In other tissues, hyaluronan has both structural and cell-regulatory activities. For instance, two receptors of the LINK protein family, LYVE-1 and stabilin-1, are important in specific functions of lymphoid cells and endothelial cells, respectively. The main cell surface receptor for hyaluronan is CD44 and after internalization hyaluronan may even have some intracellular functions.

The number of proteoglycans that have been described in relation to other cellular functions has increased greatly (Table 4-3). Proteoglycans (PGs) are thus now described in three families of proteins: transmembrane, hyaluronan-binding and collagen fibril-regulating proteoglycans. Transmembrane PGs have a direct role in specific intercellular interactions, hyaluronan-binding PGs function as links for space-filling effects (e.g. in the vitreous cavity), while collagen fibril-regulating PGs have specific roles, e.g. force-generating in tendons, light transmission in the cornea, mechanical support in bone and deformability in heart muscle. In the cornea, four PGs (decorin, lumican, keratocan and mimecan) are the major keratan sulphate moieties that, together with collagens type VI and XII, are essential for maintaining corneal transparency.

Extracellular matrix molecules are intimately involved in cell adhesion

Some proteoglycans such as the syndecans participate constitutively in cellular adhesion and its associated cell signalling, acting as co-receptors with other receptors such as integrins in the receptor cluster, and thus

TABLE 4-3 **Some tissue proteoglycans**

Proteoglycan	GAG	Tissue	Function
Aggrecan	CS, KS	Cartilage	Structural support
Versican	CS	Fibroblasts	Cell migration, support
Decorin	CS/DS	Fibroblasts	Fibrillogenesis
Fibromodulin	KS	? Keratocytes	Fibrillogenesis
α_2 (IX) collagen	CS	Vitreous, cartilage	Collagen binding
Syndecan	CS, HS	Epithelia, fibroblasts	Morphogenesis
Basement membrane	HS	Basement membrane	Support
CD44	CS	Lymphocytes, epithelia, retina	Cell–cell interactions

BOX 4-7 STRUCTURE OF GLYCOSAMINOGLYCANS

Glycosaminoglycans are long chains of repeating disaccharides based on a common structure: (**A**) heparan sulphate; (**B**) chondroitin 4,6-sulphate. Chondroitin 6- and 4,6-sulphates have additional sulphate groups at the appropriate C atoms (*); dermatan sulphate has the GlcA (glycosamino-) residue replaced in variable lengths of the chain by IdoA (iduronic acid). Heparan has many more sulphate groups and a higher content of IdoA than heparin. Keratan sulphate has galactose instead of glucose, and hyaluronic acid has no sulphate groups.

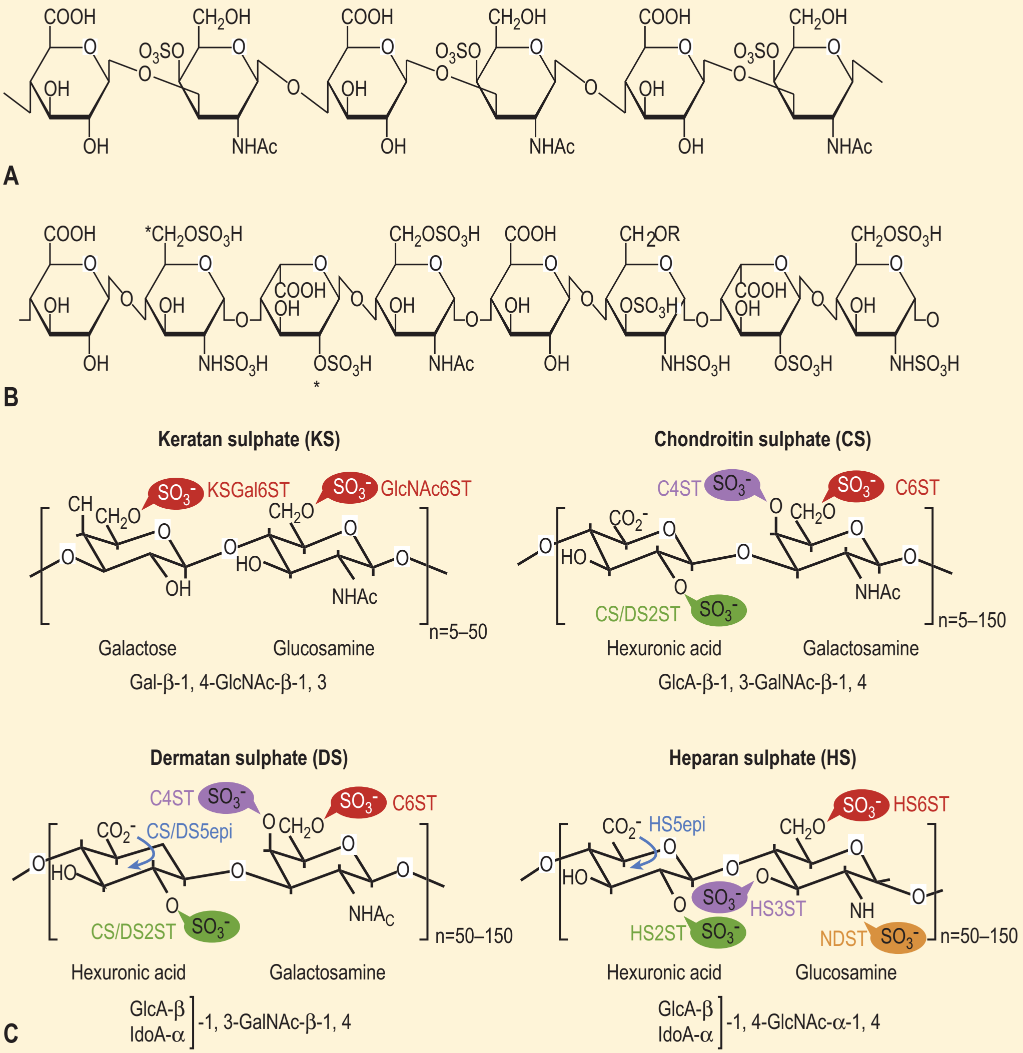
COOH
CH_2OH
O_3SO
OH
NHAc
A
*CH_2OSO_3H
CH_2OSO_3H
CH_2OR
COOH
OSO_3H
$NHSO_3H$
*
B
Keratan sulphate (KS)
SO_3^- KSGal6ST
SO_3^- GlcNAc6ST
CH
CH_2O
HO
OH
NHAc
n=5–50
Galactose
Glucosamine
Gal-β-1, 4-GlcNAc-β-1, 3
Chondroitin sulphate (CS)
C4ST SO_3^-
SO_3^- C6ST
CO_2^-
CS/DS2ST SO_3^-
n=5–150
Hexuronic acid
Galactosamine
GlcA-β-1, 3-GalNAc-β-1, 4
Dermatan sulphate (DS)
CS/DS5epi
n=50–150
GlcA-β
IdoA-α
-1, 3-GalNAc-β-1, 4
Heparan sulphate (HS)
SO_3^- HS6ST
HS5epi
SO_3^- HS3ST
HS2ST SO_3^-
NH
NDST SO_3^-
Glucosamine
-1, 4-GlcNAc-α-1, 4
C

Variations in the common structure of disaccharide units of glycosaminoglycans (GAGs) underpin the different GAGs. Dermatan sulphate contains iduronic acid, while chondroitin sulphate contains glucuronic acid. C4ST, CS-4O-sulphotransferase; C6ST, CS-6O-sulphotransferase; CS/DS2ST, CS/DS-2O-sulphotransferase; Gal, galactose; GalNAc, N-acetyl-galactosamine; GlcA, glucuronic acid; GlcNAc, N-acetyl-glucosamine; GlcNAc6ST, N-acetyl-glucosamine-6O-sulphotransferase; HS2ST, HS-2O-sulphotransferase; HS3ST, HS-3O-sulphotransferase; HS6ST, HS-6O-sulphotransferase; IdoA, iduronic acid; KSGal6ST, KS-galactose-6O-sulfotransferase; NDST: N-deacetylase-N-sulfotransferase (Adapted from Bulow and Hobert, 2006.)

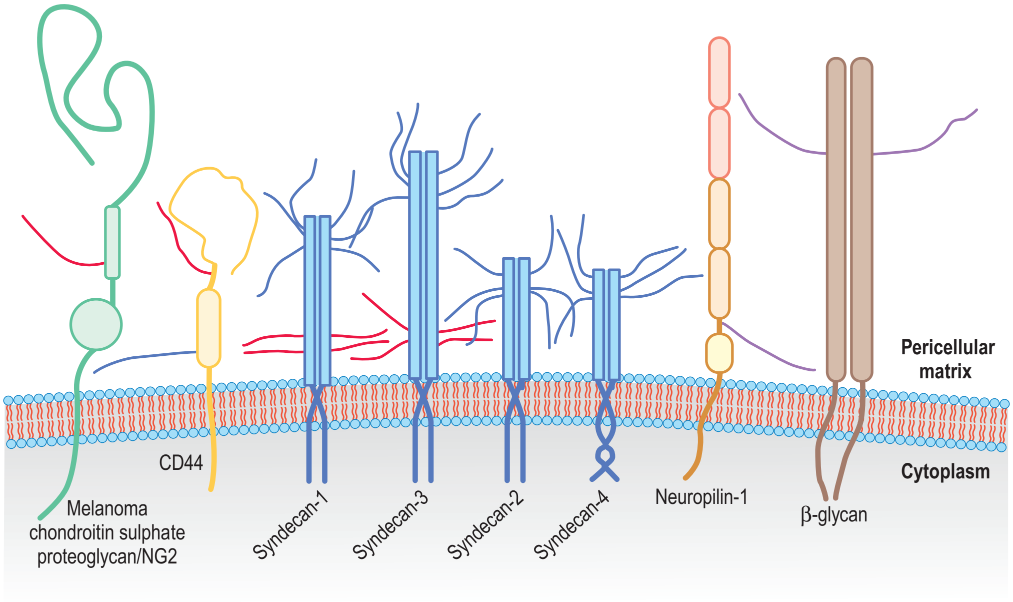

FIGURE 4-22 The transmembrane proteoglycans. The four syndecans are usually substituted with heparan sulphate chains (dark blue), while syndecans-1 and -3 have additional chondroitin sulphate chains (pink). The melanoma chondroitin sulphate proteoglycan/NG2 (green) has one chondroitin sulphate chain. The other proteoglycans also have other functions and are considered ‘part time’ proteoglycans. (From Couchman, 2010.)

connect through to the cell cytoskeleton (Fig. 4-22). Other proteoglycans interact variably with glycosaminoglycans such as CD44 and neuropilin, both of which are involved in leucocyte–matrix interactions and may be dependent on the state of leucocyte activation.

Cell adhesion is fundamental to many biological processes, such as morphogenesis, development, immune reactions to foreign proteins and many other processes. Cells adhere both to the matrix and to other cells, and usually do so via specific receptors in the cell membrane, i.e. integrins. Integrins are heterodimeric proteins (they have α and β chains, some of which are common to more than one integrin type), and are described in terms of their chain composition, e.g. $\alpha_3\beta_5$ and $\alpha_1\beta_6$ integrins (Table 4-4; see also Ch. 9 and the role of integrin receptors in leucocyte adhesion). Each cell type adheres preferentially to particular extracellular matrix proteins, depending on the type of integrin receptor it happens to express (see Table 4-4). In addition, cells may express different integrins at different times, depending on their state of activation, as, for instance, with the (myo)fibroblast during wound healing (see below). Differential adhesion by cells is shown clearly in the lens where anterior epithelial cells use the $\alpha_5\beta_1$ integrin to bind fibronectin, while the equatorial and posterior fibre cells express the $\alpha_6\beta_1$ integrin.

TABLE 4-4 **Integrin binding to extracellular matrix proteins**

- Integrins $\alpha_1\beta_1$, $\alpha_2\beta_1$, $\alpha_6\beta_1$, $\alpha_7\beta_1$, $\alpha_v\beta_4$, $\alpha_6\beta_4$ bind to laminin
- Integrins $\alpha_4\beta_1$, $\alpha_5\beta_1$, $\alpha_v\beta_1$, $\alpha_v\beta_3$, $\alpha_6\beta_4$ bind to fibronectin
- Only integrin $\alpha_3\beta_1$ binds to both laminin and fibronectin

Binding of the cell to the matrix via integrin receptors not only has a structural role but also is involved in transmembrane signalling, which may be two-way, i.e. the matrix may modify the behaviour of the cell and the cell may transmit information to other cells via the matrix (inside-out signalling).

Additional content available at https://expertconsult.inkling.com/.

This is an exquisitely sensitive system and is tunable, i.e. the integrins can vary their affinity for adhesion to matrix proteins, as directed by cellular inside-out signalling processes. The molecular subtlety is so exquisite that integrins embedded in discoid membrane structures (nanodiscs) are bent and become extended when they bind ligand. Integrins are kept in a loosely bound state by 'clasps' on either side to the membrane. The orientation (tilt) of the integrin molecule precisely determines the interaction with ligand such as fibrinogen/fibrin. Intracellular talin mediates inside-out signalling, which is regulated by kindlins under the control of the small signalling molecule Rap1 (eFig. 4-8).

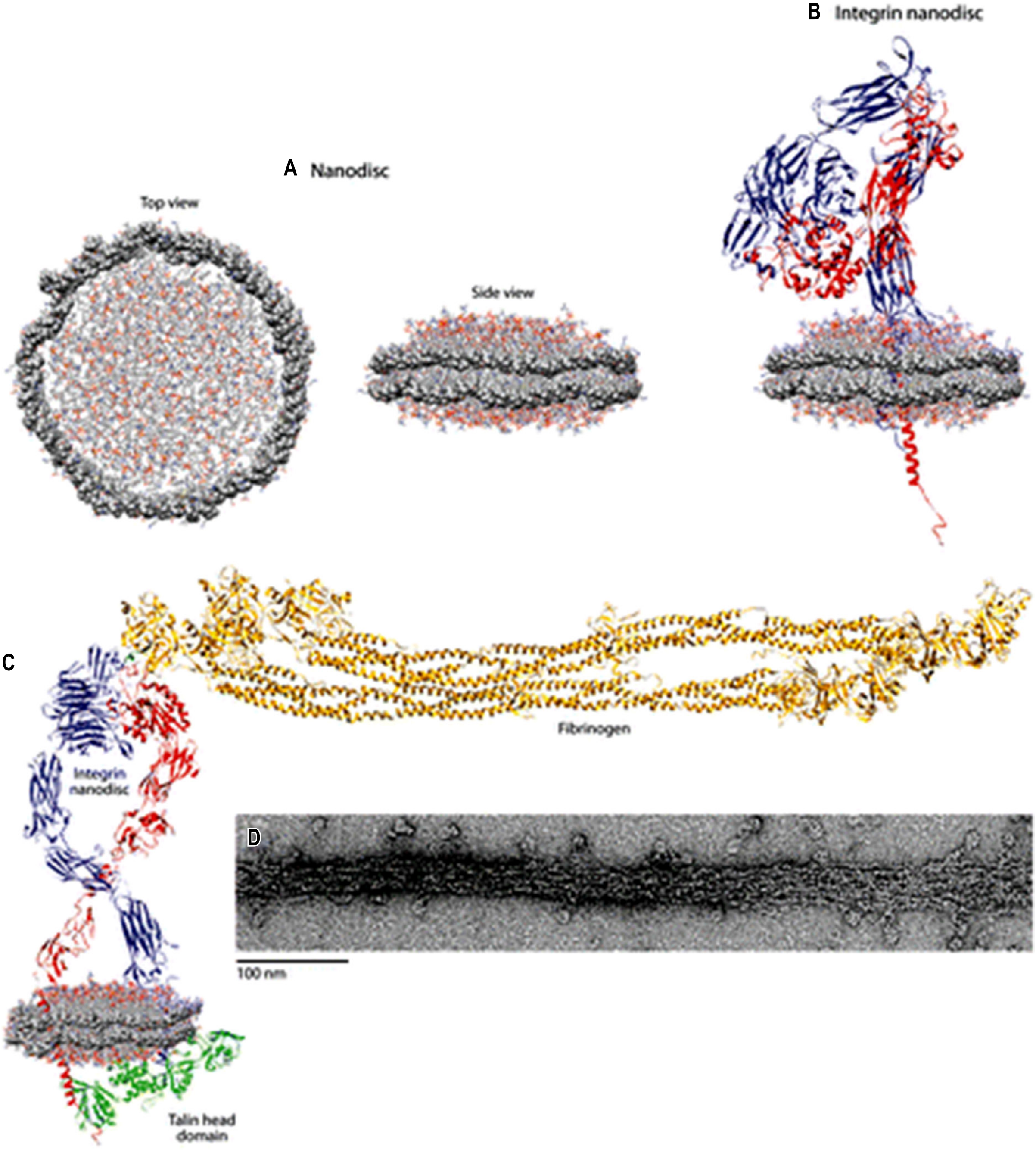
B Integrin nanodisc
A Nanodisc
Top view
Side view
C
Fibrinogen
Integrin nanodisc
D
100 nm
Talin head domain

eFIGURE 4-8 (**A**) Nanodiscs with lipid molecules are shown (coloured) and membrane scaffolding proteins (grey) are shown in solid surface view. The nanodiscs range from 10 to 13 nm in diameter. (**B**) Reconstituted integrin nanodisc containing an integrin molecule (ribbon diagram, α subunit in blue and β subunit in red). (**C**) The head of a talin domain (green), integrin nanodisc (red and blue), and an integrin ligand (in this case, fibrinogen, yellow). (**D**) A negatively stained electron microscopy image of the reconstituted inside-out integrin activation system using fibrin as a ligand. (From Kim, et al., 2011)

TABLE 4-5 **Proteins associated with cell junctions**

Adherens junction	Desmosome
α, β, γ Catenin	Plakoglobin
Plakoglobin	Desmoplakin I and II
Vinculin	Desmoplakin IV
α Actinin	Desmocollin
Tenuin	Lamin B-like protein
Plectin	Plectin
Radixin	Desmoglein

Epithelial and endothelial cells bind via transmembrane complexes to each other and to the basement membrane

Epithelial and endothelial cells rest on a highly organized basement membrane (see above). Binding of epithelial-type cells occurs predominantly via adherens-type junctions and desmosomes/hemidesmosomes, in which several distinct proteins have been identified (Table 4-5). Some of these proteins are members of what is known as the cadherin family of proteins (cell adhesion proteins), a group of transmembrane proteins that regulate intercellular adhesion (see eBox 4-2).

These proteins are not only important in the mechanical support of intercellular interactions but also play a role in inter- and intracellular signalling. Gap junctions in particular facilitate and indeed may permit amplification of signalling events which are important in coordinated cell behaviour, as for instance in an epithelial cell sheet.

Mesenchymal cells bind to the matrix via focal adhesion sites

Mesenchymal cells, such as fibroblasts, keratocytes, chondrocytes and stromal cells generally bind to the extracellular matrix via specific focal adhesion sites that contain transmembrane cytoskeletal proteins. Actin stress fibres bind via α-actinin and talin to the cytoplasmic side of the integrin receptor and to vinculin, while the extracellular component of the integrin receptor binds to extracellular matrix proteins such as collagen and fibronectin (see Fig. 4-20 and eFig. 4-8). Direct reverse signalling takes place through these sites using the signalling protein, focal adhesion kinase, one of the protein tyrosine kinases. Also involved is a further intracellular protein termed VASP (vasodilator-stimulated phosphoprotein) which integrates profilin–actin binding with talin, vinculin, F-actin and the integrin signalling complex. Together with known ion channel-dependent mechanosensors, focal adhesion proteins such as talin act as mechanosensors responding by induced unfolding of their tertiary structures. Similarly, cofilin acts to sever actin filaments in a negative-dependent fashion based on tension in the actin filament at focal adhesion sites. Certain other mesenchymal cells, such as muscle, have specific proteins in their focal adhesion sites such as dystrophin and paxillin, not only at the site of attachment of the extracellular matrix but also at the neuromuscular junction. A mutation in the dystrophin gene has been identified in patients with muscular dystrophy. An integrin-linked kinase appears to be central to regulation of outside-in and inside-out signalling in adherent cells.

Biochemical pathways that affect ocular function

The eye, particularly the outer retina, is highly metabolically active and requires large amounts of ATP, the universal energy storage molecule, for this purpose. Cells cash in on this energy bank by hydrolysing ATP and coupling this event to cell-specific enzymic reactions that are otherwise energetically unfavourable. ATP is generated by oxidative metabolism, particularly of glucose but also of other molecules such as fats and proteins.

Hydrolysis of ATP is not without risk to the cell: free electrons are produced (H^+), which are normally mopped up by nicotinamide adenine dinucleotide (NAD). Indeed, NADH is the main electron carrier in the oxidation of glucose and other molecules, while NADPH (see below) is used to reduce certain molecules such as free fatty acids to permit them to enter the metabolic pathways.

Oxidative consumption of glucose requires coenzyme A, which is a carrier of acyl groups. Other coenzymes are also required for active metabolism, many of which are derived from vitamins.

GLUCOSE METABOLISM AND TISSUE GLYCATION

Glycolysis is the conversion of glucose to pyruvate in the absence of oxygen, and is accompanied by a net

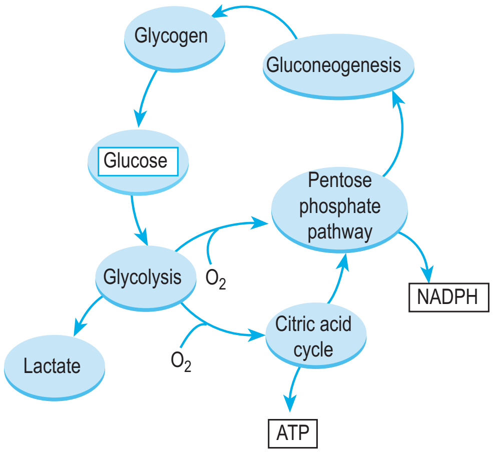

FIGURE 4-23 Summary of glucose pathways.

gain of two molecules of ATP. In the presence of oxygen, pyruvate enters the citric acid cycle with the production of 24 molecules of ATP. NAD$^+$ is regenerated in the mitochondrial electron transport chain. Glucose and other fuel molecules entering the citric acid cycle may thus be consumed to provide energy for synthesis of the building materials for other molecules such as amino acids.

Not all the fuel molecules are utilized immediately, but instead they are diverted to produce molecules that act as sources of power rather than energy. This is achieved via the pentose phosphate pathway, in which NADPH and ribose are produced, the former for use in reductive processes and the latter in the biosynthesis of nucleotides for RNA and DNA. These different pathways are briefly summarized in Figure 4-23.

The pentose phosphate pathway also permits glucose formation from unrelated precursors (gluconeogenesis) but does not occur in the brain owing to lack of glucose-6-phosphatase; this probably also applies to the retina. However, the pentose phosphate pathway is active in the cornea and lens (see below).

Finally, glucose may be stored in the liver and muscle as glycogen, whence it can be retrieved for energy purposes by glycogenolysis. The Müller cells in the retina also contain stores of glycogen, which may be essential to maintain retinal function (see Chs 1 and 5).

Energy metabolism is controlled globally by nutrient intake, use and storage and is regulated by a range of signalling pathways targeting food excess such as insulin, insulin-like growth factor-1 and mammalian target of rapamycin (mTOR); food restriction is likewise controlled by a set of pathways including the ATP-consuming anabolic pathways and ATP-generating catabolic pathways, which are under strict regulatory control by a 'master-switch' enzyme system, adenosine monophosphate-activated protein kinase (AMPK), which controls the overall whole body energy metabolism and is involved in loss of neuroprotection in the brain, and possibly the retina, in ischaemic states. In addition, sirtuins, a family of seven histone-deacetylase proteins in mammals, serve as sensors of energy (ATP) levels and also have wider transcriptional and other regulatory roles such as protection from toxicity of reactive oxygen species via upregulation of AMPK.

Glucose enters cells by facilitated diffusion

Glucose transporters (GLUTs 1–14) are members of the SLC2 (solute carrier 2) gene family and are membrane proteins which permit the facilitated diffusion of hexoses and polyols, including myoinositol, glucosamine and ascorbate, into and out of cells down their concentration gradient. Different cells and tissues have different GLUTs, which is relevant to how each tissue handles glucose (eFigs 4-9 and 4-10).

Additional content available at https://expertconsult.inkling.com/.

Thus skeletal muscle utilizes GLUT 4, which is dependent for its function in the plasma membrane on insulin; in contrast, brain and retina, which are considered to be insulin-independent tissues, express GLUTs 1 and 3. The liver and adipose tissue also have GLUTs 2 and 7, with several of the other transporters. GLUTs 1, 3 and 4, particularly GLUT 1, can be upregulated in response to hypoxia via the gene hypoxia-inducible factor-1, which also induces other genes such as vascular endothelial cell growth factor, which is important in retinal vessel permeability and new vessel growth. GLUT 4 in muscle is critical for whole body glucose homeostasis; GLUT 7 is involved in gluconeogenesis and glycogenolysis in the liver, where it is located in the endoplasmic reticulum in association with glucose-6-phosphatase, the essential enzyme for these reactions.

The glucose transporters have a closely defined mechanism for shifting glucose molecules across the cell membrane (eFig. 4-9) and are specific to tissues and cells (eFig. 4-10). Glucose is transported across the membrane in aqueous phase propelled by hydrophobic interactions between glucose molecules and residues on the amino acids in the transporter protein which maintain phase separation of the molecules, while other amino acids are dedicated to promoting binding and thus carriage of the molecule across the glucose gradients that exist. About one-third of the glucose is transported through the gut via Glu2 and Glut 5 (for fructose) and is carried via red blood cells due to their very high concentration of GLUT1. In the resting state most of the glucose is utilized by the brain and nervous tissues while during exercise muscle uses most of the circulating glucose. Adipose tissue regulates glucose utilization via production of hormones termed adipokines.

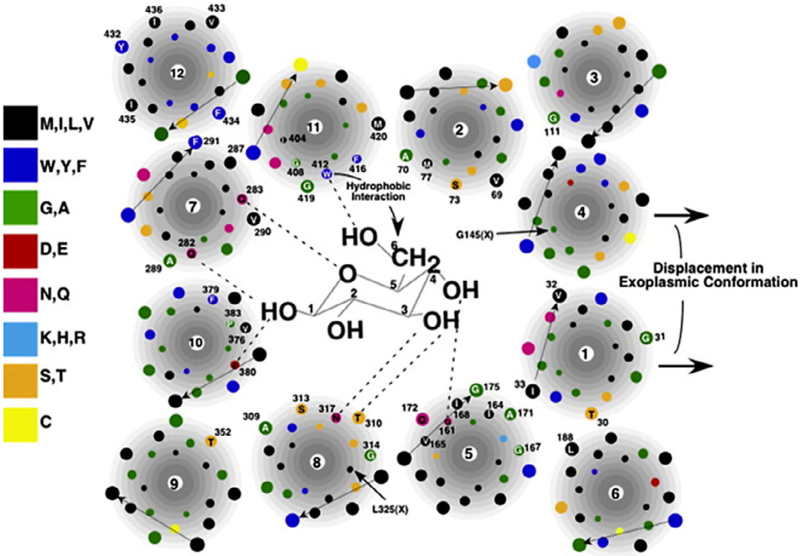

eFIGURE 4-9 Model of the exoplasmic substrate-binding site of GLUT1. The glucose molecule in the centre of the diagram is not drawn to scale. The helices surrounding amino acids in the glucose transporter are shown in a simplistic fashion for clarity. Amino acid residues that are in contact with solvent in the aqueous cavity are numbered and identified by the single-letter code. Dotted lines represent putative hydrogen bonds. (From Mueckler and Thorens, 2013.)

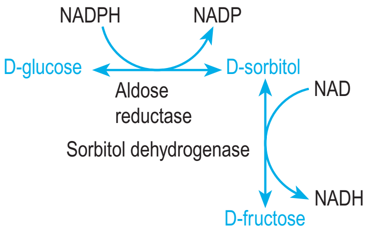

FIGURE 4-24 Aldose reductase pathway.

Excessive levels of glucose may impair cellular metabolism

Glucose metabolism is tightly regulated by insulin, IGF-1 and glucagon release from the pancreatic β cells. In diabetes mellitus, in which there is impaired secretion or utilization of insulin, hyperglycaemia leads to excessive uptake of glucose into cells, despite negative feedback on the expression of GLUTs on the cell surface. This has the effect of overloading the metabolic pathways with activation of alternative routes for glucose handling such as the aldose reductase pathway (Fig. 4-24).

In certain tissues, such as the lens, the effect of increased sorbitol may be to cause damage by osmotic dysregulation because sorbitol cannot be transported out of the cell easily. In this case, even high glucose can have direct osmotic damaging effects on the cell owing to raised intracellular $[Ca^{2+}]_i$ perhaps as a result of cell shrinkage with consequent activation of the stretch receptors. Alternatively, the excessive utilization of NADPH in this pathway might have deleterious effects on the levels of myoinositol, which is required for intracellular signalling, or on the increased generation of reactive oxygen species directly via activation of the phosphoinositol pathway.

It has therefore been suggested that activation of the aldose reductase pathway is not the direct cause of cellular damage but is merely a coincidental perturbation in the cell; instead, excessive production of free radicals may be important during oxidation of the high concentrations of glucose. Interestingly, aldose reductase has been implicated in inflammatory processes by promoting lipid peroxidation and free radical generation (see below). Whatever the mechanism, aldose reductase is likely to be involved in the pathophysiology of the lens at least, because inhibitors of this enzyme prevent the development of cataracts.

Chronic hyperglycaemia has also been shown to alter cellular metabolism via modulating insulin post-receptor intracellular signalling, involving the cascade insulin receptor substrate 1 (IRS-1)/phosphatidylinositol 3-kinase (PI3K)/Akt, as the result of production of high levels of hexosamine in the cell. This may be one mechanism of induction of insulin resistance in diabetes mellitus. A more general mechanism for the effects of high ambient glucose concentration on cell behaviour has been the induction of the cytokine TGF-β with its widespread effects on cell function, particularly increased proteoglycan synthesis and consequent basement membrane thickening. Patients with type II diabetes are known to have increased serum levels of TGF-β.

Excessive levels of glucose lead to glycation of proteins

Addition of sugar moieties to proteins can occur by enzymatic conjugation (termed glycosylation) or non-enzymatic conjugation (termed glycation). Glycation of proteins is classically considered to occur in the presence of high concentrations of glucose in two phases: an early reversible phase during which the protein forms a Schiff base and Amadori products, followed by a later irreversible phase in which advanced glycation end products (AGEs) appear (Fig. 4-25). However, AGEs can be produced in the early phase because free glucose degrades to α-oxoaldehydes, which are potent glycating agents. AGEs can thus be formed at any stage in the process of glycation, including through binding of Schiff bases and fructosamine. In addition, α-oxoaldehydes such as glyoxal, methyl glyoxal and 3-deoxyglucosone are produced as glycolytic intermediates and also during lipid peroxidation (see below), and lead directly to AGE formation (Fig. 4-25).

Glycation of proteins occurs at lysine, arginine and cysteine amino acids, thus affecting proteins such as collagen (pentosidine cross-links) and haemoglobin and occurs as part of ageing and pathologically in diabetes mellitus; in addition, important extracellular matrix proteins, such as PAI-1 (see above) may be glycated, leading to defective control of cell behaviour, including cell migration and activation. Glycation of cell membrane proteins such as Ca^{2+} channels in pericytes may impair their function, thus rendering them

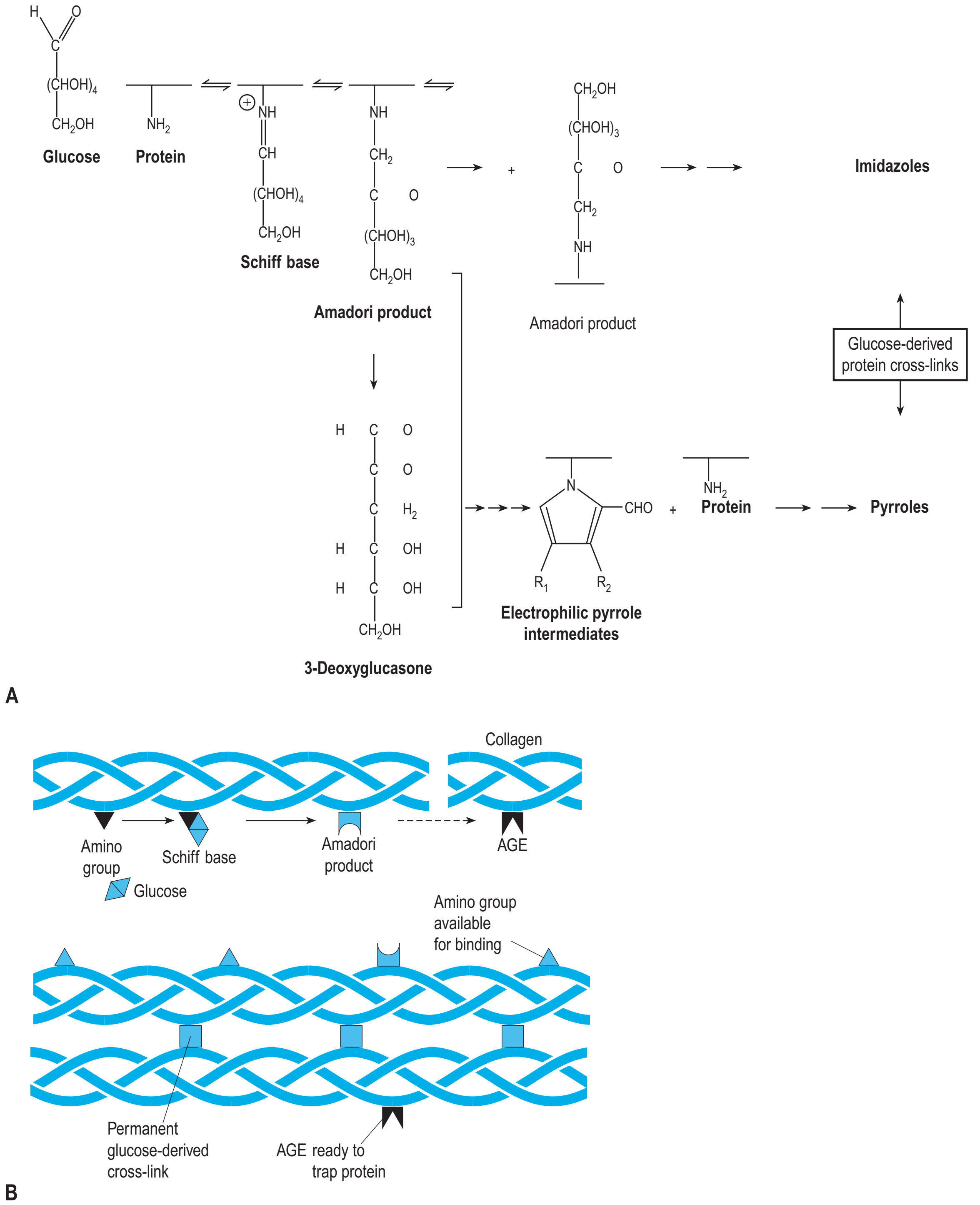
H
O
C
(CHOH)4
CH2OH
Glucose
NH2
Protein
NH
CH
(CHOH)4
CH2OH
Schiff base
NH
CH2
C
O
(CHOH)3
CH2OH
Amadori product
+
CH2OH
(CHOH)3
C
O
CH2
NH
Amadori product
Imidazoles
Glucose-derived protein cross-links
H C O
C O
C H2
H C OH
H C OH
CH2OH
3-Deoxyglucasone
N
CHO
R1
R2
Electrophilic pyrrole intermediates
+
NH2
Protein
Pyrroles
A
Collagen
Amino group
Schiff base
Amadori product
AGE
Glucose
Amino group available for binding
Permanent glucose-derived cross-link
AGE ready to trap protein
B

FIGURE 4-25 Process of non-enzymatic glycation. (**A**) Biochemical pathway via Schiff base and Amadori products leading to glucose-derived cross-links and ultimately advanced glycation end products (AGEs). (**B**) Diagrammatic representation of molecular interactions involved in AGE-mediated collagen cross-links. (Courtesy of A. Ceramin and the publishers of Scientific American.)

less responsive to endothelin-1-induced contractility. The two processes of non-enzymatic glycation and auto-oxidation of proteins are also directly interrelated and may be important in glucose-derived oxidative stress. This last effect is a direct inducer of apoptosis and mediates neuronal and endothelial cell death in diabetic neuropathy and vasculopathy, through binding of AGEs to a specific receptor, RAGE.

A number of enzymes are known to repair proteins in the early stages of glycation, including fructosamine-3-kinase and glyoxalase and the term ‘carbonyl stress’ is used to describe imbalance in glycation versus deglycation.

Interestingly, the free radical scavenger vitamin B_6 may inhibit AGE production via the inhibition of oxidative degradation of Amadori intermediates and the trapping of reactive oxygen products (see below).

Glucose and lipid metabolism

Glucose and lipid metabolism are also intimately linked via the metabolite acetyl coenzyme A. In the presence of excess glucose, production of fatty acids and cholesterol through acetyl coenzyme A is increased with consequent increases in phospholipids (see below) and circulating levels of very low-density lipoproteins. While this has well-known implications for the development of atherosclerosis, more recently this form of metabolic dysfunction has been suggested to underlie pathologies such as age-related macular degeneration, not simply through the production of abnormal lipid deposits in areas where their removal is difficult, such as the subretinal space, but also by a concomitant increase in lipid-based free radicals (see below).

OXIDATION/REDUCTION AND FREE RADICAL PRODUCTION

Oxidative metabolism and the generation of ATP storage energy molecules are conducted via the cytochrome enzyme system. However, during the process of oxygen consumption, a small amount of oxygen (<5%) is metabolized by alternative pathways. Univalent reduction of oxygen produces highly reactive free radicals, namely the superoxide anion and the hydroxyl radical, and the toxic molecule hydrogen peroxide (Fig. 4-26).

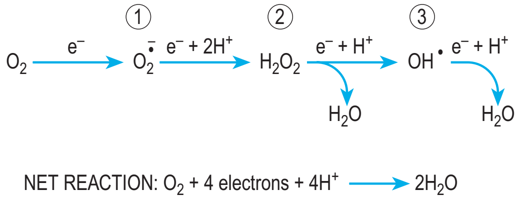

FIGURE 4-26 Free radical production: (1) superoxide anion; (2) hydrogen peroxide; (3) hydroxyl radical.

TABLE 4-6 **Detoxification of free radicals**

Mechanism	Agent
Superoxide anion degrading enzymes	Superoxide dismutases
Antioxidants	Ascorbate
Free radical scavengers	Vitamins A and E
H_2O_2 degrading enzymes	Catalase, peroxidase

The superoxide anion is generated in mitochondria in the ubiquinone cytochrome b system and in other membranes by auto-oxidation via the cytochrome P_{450}-linked reductases (also part of the drug-metabolizing system), which are rich in the vitamin flavoprotein. Superoxide is then converted by superoxide dismutase to hydrogen peroxide (H_2O_2), which diffuses into the cytosol. Superoxide is also generated in the cytosol itself via recycling of redox enzymes such as xanthine oxidase; conversion to H_2O_2 in the cytosol is rapid. Xanthine, in particular, accumulates in certain tissues during ischaemia, and subsequent reperfusion leads to massive release of free radicals, producing extensive tissue damage. This can be prevented by free radical scavengers.

Free radicals can be generated by other mechanisms. Reduction of Fe^{3+} to Fe^{2+}, by reducing agents such as ascorbate, catalyses the conversion of H_2O_2 to $OH^\bullet$, and auto-oxidation of other compounds such as thiols and catecholamines produces free radicals. The Fe^{2+} ion has particular relevance to the problem of retained intraocular metallic foreign bodies. Tissue damage during inflammation is in a large part attributable to the production of free radicals by phagocytic cells during the respiratory burst (see Ch. 7). These interactions are illustrated in Box 4-8.

Free radicals are detoxified by a variety of enzymatic and non-enzymatic mechanisms (Table 4-6). In

BOX 4-8 GENERATION OF FREE RADICALS FROM HYDROGEN PEROXIDE

The oxidation of hydrogen peroxide (H_2O_2) leads to glutathione consumption via the pentose phosphate pathway and requires a supply of glucose to maintain homeostasis.

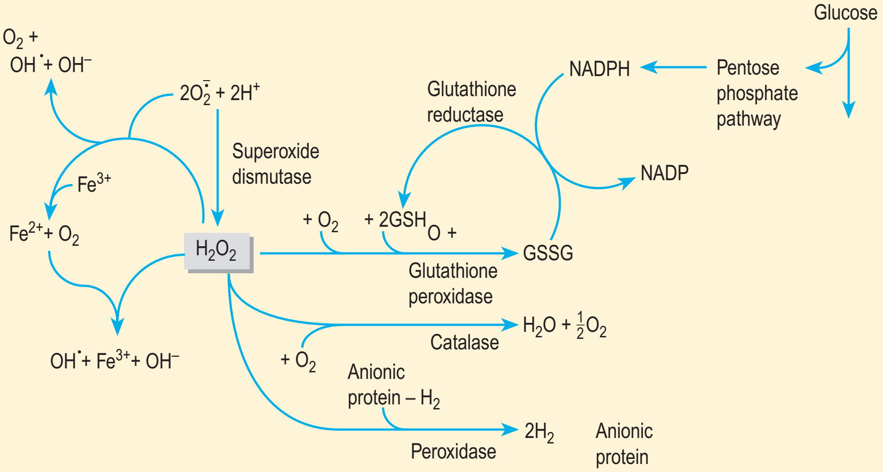
Glucose
O_2 +
$OH^{\bullet}$ + OH^-
$2O_2^{\bar{\cdot}}$ + $2H^+$
Glutathione reductase
NADPH
Pentose phosphate pathway
Superoxide dismutase
Fe^{3+}
NADP
Fe^{2+} + O_2
+ O_2
+ 2GSH
O +
H_2O_2
GSSG
Glutathione peroxidase
H_2O + $\frac{1}{2}O_2$
Catalase
$OH^{\bullet}$ + Fe^{3+} + OH^-
+ O_2
Anionic protein – H_2
$2H_2$
Anionic protein
Peroxidase

addition, a critical redox regulatory system involving glutathione and the enzyme glutathione S-transferase is central to cellular homeostasis (Fig. 4-27). Superoxide dismutases require divalent ions such as Mn^{2+}, Cu^{2+} and/or Zn^{2+} for normal activity; some of these have been demonstrated in ocular tissues. Further reduction of H_2O_2 is either by enzymatic or non-enzymatic means, and may involve several mechanisms directed towards removal of reactive oxygen species (see Box 4-8).

The mechanism of cell damage varies with each molecular species. Chelated metal ions are important in non-enzymatic degradation of H_2O_2, producing the highly reactive $OH^{\bullet}$ radical, which is particularly damaging to cell membranes. H_2O_2 causes damage by inhibiting glycolysis and glucose uptake and is the most stable of the reactive oxygen species as well as being generated by several intracellular processes (see eFigs 4-10 and 4-11).

Additional content available at https://expertconsult.inkling.com/.

Oxidative stress is considered one of the major mechanisms underlying DNA damage and has been correlated with changes in telomere length associated with ageing. Cells respond to DNA damage using a protein termed protein kinase ataxia telangiectasia mutated (ATM) which mediates DNA repair, cell cycle arrest and apoptosis. Oxidative stress can directly activate dimeric ATM (usually the monomer is the active agent), which seems to act as a cellular sensor for potential oxidative damage.

LIPIDS AND LIPID PEROXIDATION

Lipids are the structural basis on which cell membranes are built and are generally composed of a hydrophobic tail of two fatty acid chains bound to a hydrophilic (polar) head (Box 4-9). Fatty acids, thus, have at least three roles:

- as integral components of phospholipids and glycolipids
- as functional molecules, e.g. hormones and second messengers
- as energy stores in the form of triacylglycerols.

In addition to phospholipids and glycolipids, cholesterol is the third major lipid component of eukaryotic cell membranes; cholesterol is predominantly hydrophobic but has a hydrophilic group on carbon 3

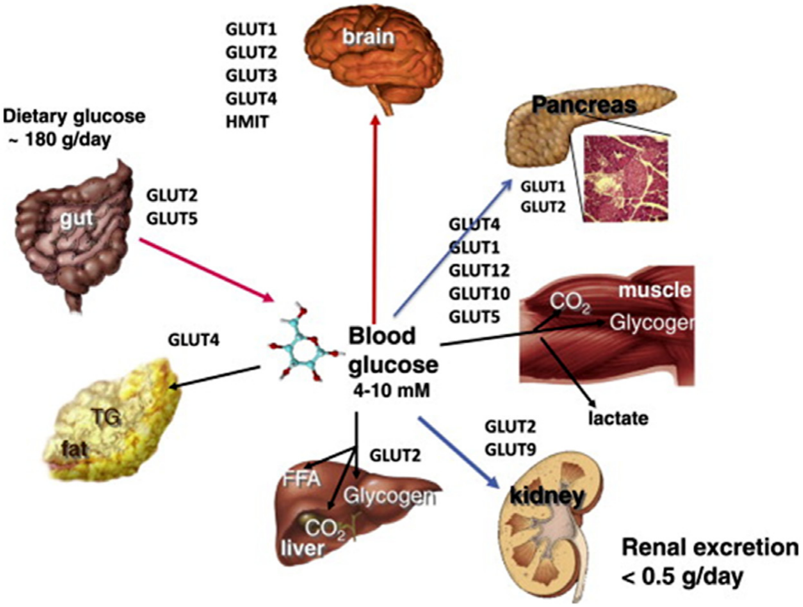
GLUT1
GLUT2
GLUT3
GLUT4
HMIT
brain
Dietary glucose
~ 180 g/day
Pancreas
gut
GLUT2
GLUT5
GLUT1
GLUT2
GLUT4
GLUT1
GLUT12
GLUT10
GLUT5
CO2
muscle
Glycogen
GLUT4
Blood
glucose
4-10 mM
TG
fat
lactate
GLUT2
GLUT9
GLUT2
FFA
Glycogen
CO2
liver
kidney
Renal excretion
< 0.5 g/day

eFIGURE 4-10 Role of GLUT proteins in the maintenance of glucose homeostasis. The various GLUT proteins expressed by different tissues are shown. (Mueckler and Thorens, 2013.)

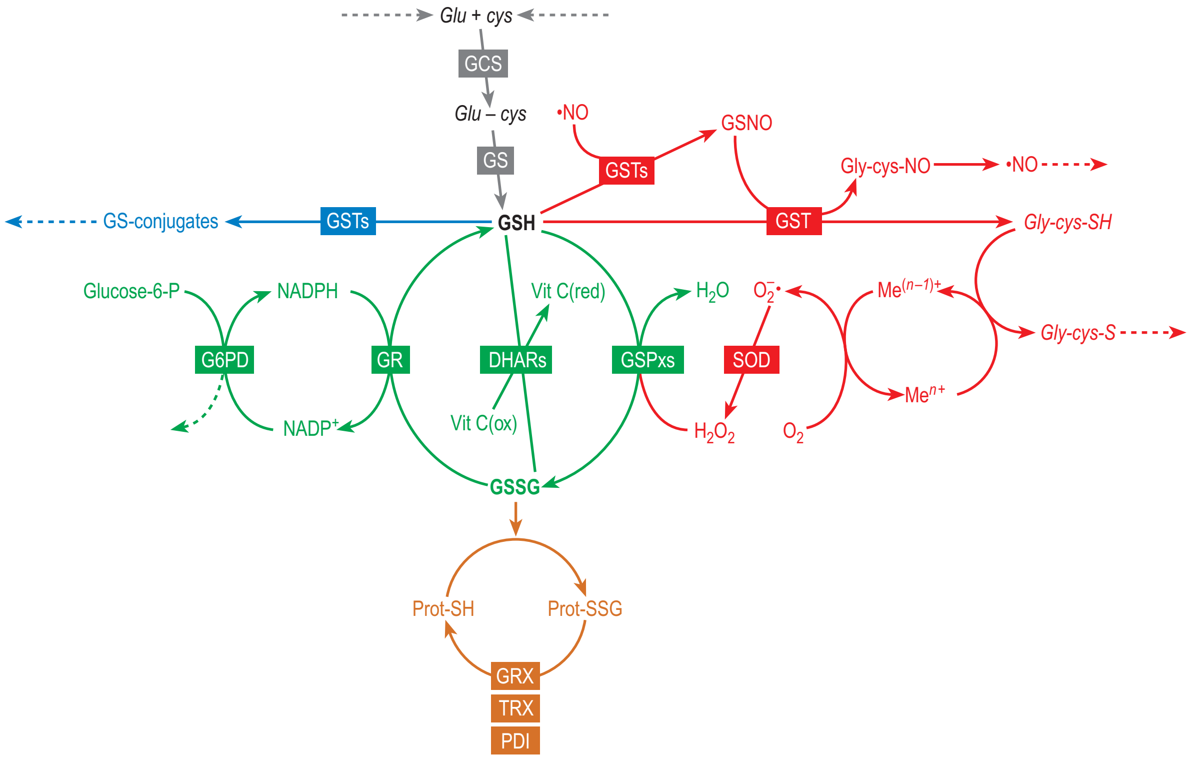

FIGURE 4-27 Outline of the interrelations connecting the various roles played by GSH in cellular homeostasis: antitoxic (blue), antioxidant (green), pro-oxidant (red), modulator (yellow). γ-GCS, γ-glutamyl-cysteine synthetase; DHARs, dehydroascorbate reductases; G6PD, glucose-6-phosphate dehydrogenase; GPxs, glutathione peroxidases; GR, glutathione reductase; GRX, glutaredoxin; GS, glutathione synthetase; GSNO, S-nitrosoglutathione; GSTs, glutathione *S*-transferases; Me, metal; PDI, protein disulphide isomerase; SOD, superoxide dismutase; TRX, thioredoxin. (From Pompella et al., 2003, with permission from Elsevier.)

(Fig. 4-28). In an aqueous medium, lipids arrange themselves in such a way that the polar (hydrophilic) groups face the aqueous phase, whereas the hydrophobic groups face each other (a micelle). However, owing to the bulky nature of the two chains of fatty acids, membrane lipids do not readily form micelles but group together as a lipid bilayer (Fig. 4-28). Phospholipid bilayers can reach macroscopic dimensions and are barriers to the diffusion of aqueous solutes but remain quite fluid themselves. Thus, they form an ideal material to act as biological membranes. The fluidity of lipid membranes is related to the length of the fatty acid chains, the number and nature of double bonds in the chain and, in eukaryotes, the content of cholesterol. Proteins suspended in lipid bilayers, such as rhodopsin, have very rapid lateral mobility within the membrane unless they are anchored across the membrane to cytoskeletal and/or matrix proteins; in contrast the polarity of the protein in the membrane is fixed.

The formation of lipid bilayers is a spontaneous event, due to the physical hydrophobic interactions between the hydrocarbon tails, assisted by polar interactions between the aqueous phase and the hydrophilic head, and between adjacent polar groups. Furthermore, lipid bilayers form closed compartments in which defects are self-sealing, even after disruption. In many membranes, however, proteins, especially receptors and co-receptors, are distributed in clusters while the lipids form lipid rafts which differ in their lipid composition from adjacent regions (see p. 162).

Fatty acids undergo physiological degradation during energy consumption by oxidation and polyunsaturated fatty acids (PUFAs) are particularly

BOX 4-9 GENERAL STRUCTURE OF PHOSPHOLIPIDS

The general composition of lipids is a hydrophobic tail consisting of fatty acids bound to a hydrophilic head, which characterizes the molecule (**A**). Glycolipids differ from phospholipids in having a sugar moiety rather than a phosphorylcholine linked to the fatty acid chain. Sphingomyelin is intermediate between these (**B**).

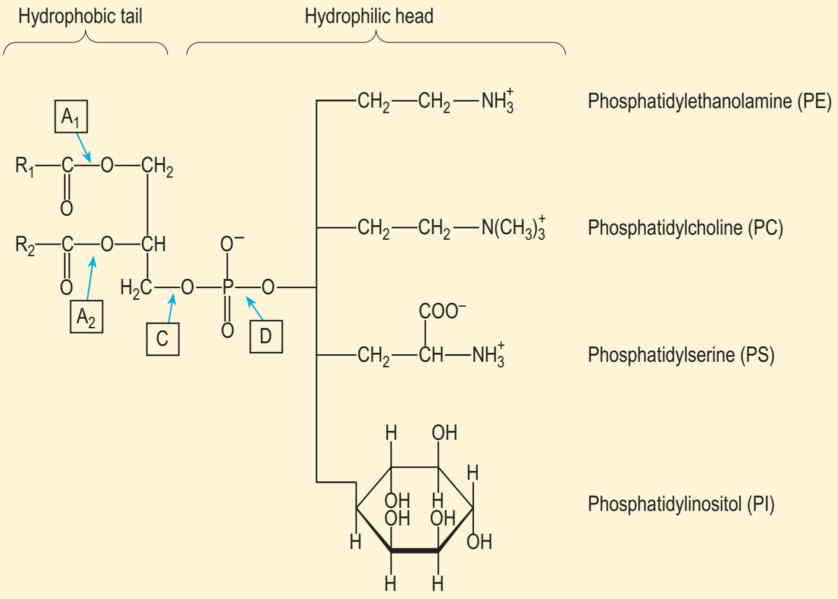
Hydrophobic tail
Hydrophilic head
A1
A2
C
D
R1
R2
Phosphatidylethanolamine (PE)
Phosphatidylcholine (PC)
Phosphatidylserine (PS)
Phosphatidylinositol (PI)

R_1 and R_2: Fatty acids (usually saturated and unsaturated, respectively)

$\boxed{A_1}$ $\boxed{A_2}$ $\boxed{C}$ $\boxed{D}$: Sites of phospholipase activity

A

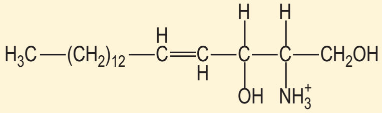

Sphingosine

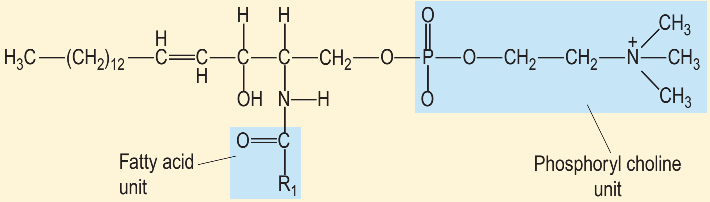
Fatty acid unit
Phosphoryl choline unit

Sphingomyelin

B

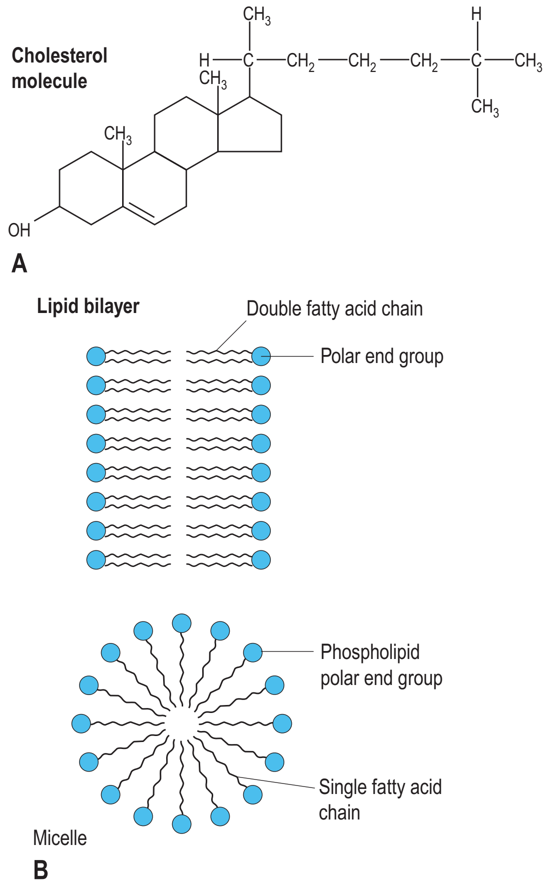
Cholesterol molecule
CH3
H
H—C—CH2—CH2—CH2—C—CH3
CH3
CH3
CH3
OH
A
Lipid bilayer
Double fatty acid chain
Polar end group
Phospholipid polar end group
Single fatty acid chain
Micelle
B

FIGURE 4-28 Structure of the cholesterol molecule (**A**) and the lipid bilayer (**B**).

BOX 4-10 OXIDATION OF LIPID MEMBRANES

Phospholipids, especially those containing arachidonic acid (20:4) and docosohexanoic acid (22:6), a fatty acid abundant in photoreceptor cells, are oxidized to lipid free radicals, which are then conjugated to form dienes. These are then converted to lipid peroxides.

$$RH \rightarrow R^{\bullet} + O_2 \rightarrow ROO^{\bullet}$$

Peroxidation of lipids in cell membranes leads to the release of large quantities of arachidonate, which acts as a substrate for eicosanoid production.

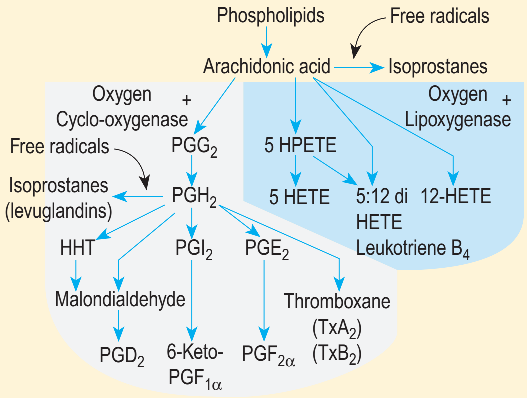
Phospholipids
Free radicals
Arachidonic acid
Isoprostanes
Oxygen +
Cyclo-oxygenase
Oxygen +
Lipoxygenase
Free radicals
PGG2
5 HPETE
Isoprostanes (levuglandins)
PGH2
5 HETE
5:12 di HETE
12-HETE
Leukotriene B4
HHT
PGI2
PGE2
Malondialdehyde
Thromboxane (TxA2) (TxB2)
PGD2
6-Keto-PGF1α
PGF2α

susceptible. Similarly, lipids in cell membranes can be oxidized. Indeed, peroxidation of lipids is one of the main sources of cell damage, and polyunsaturated fatty acids are particularly susceptible to this form of non-enzymatic damage.

The lipid peroxides oxidize neighbouring fatty acids and set up chain reactions in the membrane. Phospholipase A_2, by releasing free fatty acids from the membrane, is thought to detoxify this reaction. However, the arachidonate released is then converted to prostaglandins and other eicosanoids (Box 4-10; see also Ch. 7, p. 382). Arachidonates may also be modified by interaction with free radicals in a non-enzymatic reaction. This produces a series of prostanoids, prostaglandin-like chemicals termed isoprostanes. In fact, isoprostanes are produced in much larger quantities than prostaglandins and are detectable *in vivo* as a measure of lipid peroxidation, and thus indirectly tissue damage. Isoprostanes occur at low levels in normal healthy individuals, thus reflecting a level of oxidative injury not completely suppressed by natural antioxidants. However, as biomarkers of disease hydroxyoctadecadienoic acids (HODEs), hydroxyeicosatetraenoic acids (HETEs), and hydroxycholesterols appear to be sufficiently stable end products of lipid peroxidation, as are isoprostanes and neuroprostanes. Isoprostanes have biological and potentially pathological effects on cell function, thus directly contributing to disease.

Release of carbonyl compounds from both carbohydrates and lipids can lead to protein modification by glycation (AGEs, see above) and by lipoxidation (ALEs, advanced lipo-oxygenation end products). These are seen especially in uraemic patients on dialysis. Vitamin B_6 (pyridoxamine) may also inhibit ALE production as well as AGE production (see above).

Lipid peroxidation can be limited by vitamin E, a lipid-soluble free radical scavenger that is present within the bilayer (especially photoreceptors) and reacts directly with the lipid free radical to produce the phenoxy radical, or with ascorbic acid in the aqueous phase (see above).

Lipids also occur extracellularly and are of particular importance as lipoprotein carriers in the circulation, and also as lubricants on surfaces. This is particularly important for the ocular surface.

The ocular surface

The ocular surface comprises the conjunctival and corneal non-keratinized epithelium and is bathed by the tear film. However, integrity of the surface depends on other structures such as the apposition of the eyelid margins, the meibomian gland and the lacrimal gland secretions. Collectively, this is termed the lacrimal functional unit (LFU). Evaporation of the tear film is clinically measurable as the tear film break-up time, and tear film integrity is restored by blinking. Blink rate and break-up time are therefore linked. Tears provide lubrication for lid closure, assist in smoothing out the irregularities in the ocular surface (which otherwise may have transient effects on light transmission), and also have an antibacterial function.

THE TEAR FILM

The tear film is a protective covering for the cornea composed of three layers: a surface oily layer, an aqueous layer and a deep mucous layer. Previous estimates of tear film thickness suggested that the aqueous layer comprised more than 95% of the tear film, but this has been reduced to 60%, with a larger component being provided by the mucous layer based on evidence from *in vivo* studies using the confocal corneal microscope. In addition, clinical evaluation of the tear film involves an assessment of the tear meniscus height (TMH). The TMH reflects the volume of tears that collects at the contact line between the eyelid margin and the bulbar conjunctiva and has several other names, including the inferior marginal strip and the tear prism or rivus.

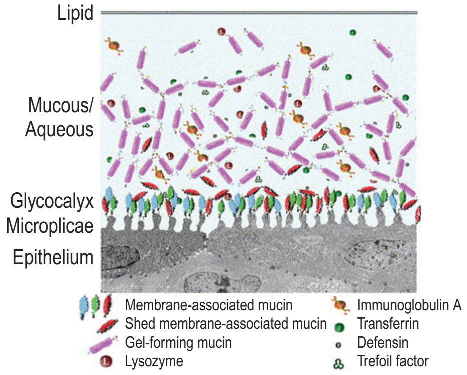

FIGURE 4-29 Diagram of the tear film–glycocalyx interface demonstrating tear components and the apical surface glycocalyx with its membrane-tethered mucins. (From Govindarajan and Gipson, 2010.)

The cornea presents a hydrophobic non-wettable surface, which is made wettable by possessing a layer of mucus on its surface. This is overlaid by the aqueous component of the tear film and lastly by a layer of lipid, which prevents evaporation of the tears (Fig. 4-29). Soluble mucus in the aqueous layer and meibomian lipid combine to ensure stability of the tear film by lowering the surface tension and permitting spreading of the tear film on blinking. The break-up time is therefore thought to represent the time it takes for aqueous to evaporate and meibomian lipid to come in contact with the hydrophobic epithelial cell layer.

The blink reflex removes up to 70% of the surface aqueous–mucous layer while the remaining 30–40% thick mucous layer in contact with the cornea is not wiped away on blinking. Instead, it is thinned by blinking and may take 30 minutes to reconstitute itself. The mucous layer is in part composed of the glycocalyx of the epithelial cells and also by an additional layer of tear mucins produced by the conjunctival goblet cells. Reduced tear film break-up time in dry eye disorders of various types therefore reflects a disturbance in tear mucin/aqueous protein interaction and may be associated with a reduction in goblet cell density.

Tear film lipids have unique characteristics

The lipid layer of the tear film is derived from meibomian gland secretions and is very thin (0.1 μm). It is composed of a mixture of polar and neutral lipids (Box 4-9) with a melting point (35°C) that ensures it is always fluid on the ocular surface. The polar lipids are in contact with the aqueous phase of the tear film and

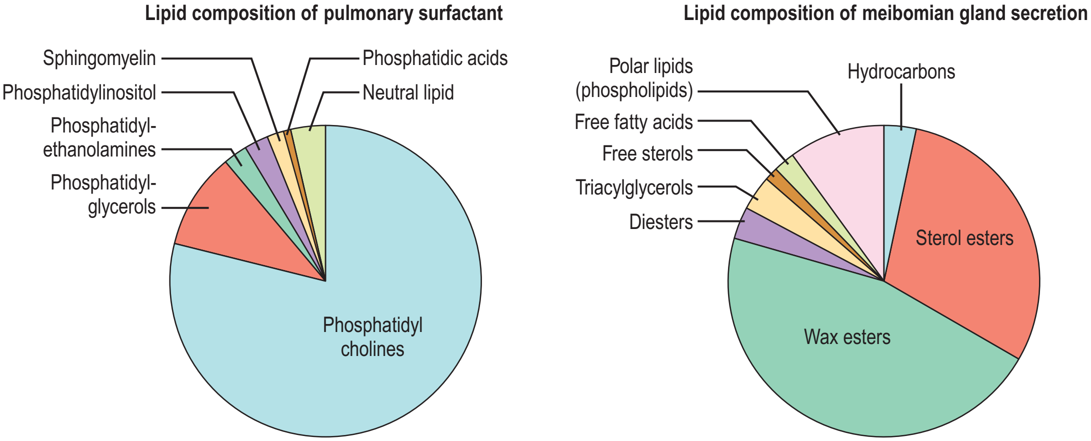

FIGURE 4-30 Proportion of different lipid moieties compared between composition of pulmonary surfactant, a lipid-rich surface layer convering the lung mucosa, and the mebomian gland secretion, the major contribution to tear film lipid. Note the diversity of lipids in the tear film, particularly the wax esters. (From Panaser and Tighe, 2012)

TABLE 4-7 Composition of the tear film

Physical properties	Solutes (μmol/L)	Proteins	Enzymes/inhibitors
98% H_2O_2	Na 120–160	Lysozyme	Glycolytic
6–9 μL volume	Cl 118–135	Lactoferrin	Amylase
pH 7.5	HCO_3 20–25	Tear-specific prealbumin (lipocalin)	Plasminogen activator
310–334 mOsm	K 20–42	G protein	(α_2-macroglobulin)
	Mg 0.7–0.9	(IgA, IgG)	(α_1-antitrypsin)
	Ca 0.5–1.1	(Ceruloplasmin)	
	Glucose 0.5–0.7	(Albumin)	
	Retinol	(Orosomucoid)	
	Urea		

provide structural stability to the tear film, while the non-polar lipids are at the air interface and provide barrier function and thixotropic properties. The lipids include unsaturated and branched-chain fatty acids and alcohols, 8–32 carbon chains in length. Tear lipids are under study by way of lipidomics and different sets of wax esters, oleates and cholesterol esters have been identified. Interestingly, in comparison with lung surfactant, there is a more heterogeneous composition with prominence of wax esters (Fig. 4-30).

Tear lipids promote movement of water into the aqueous phase during formation of the tear film, where they bond to tear lipocalins. The functions of the meibomian lipid layer are:

- to prevent evaporation of tears
- to prevent spillover of tears at the lid margin
- to prevent migration of skin lipid onto the ocular surface
- to provide a clear optical medium.

Lacrimal gland secretion provides the aqueous component of tears

The lacrimal gland and its accessory glands (see Ch. 1, p. 89) are classic exocrine acinar glands secreting a dilute aqueous solution containing proteins and small molecular weight components and electrolytes (Table 4-7). Although over 1500 proteins have been identified in tear fluid by preoteome analysis, the principal proteins in tears are immunoglobulin A (IgA), lactoferrin, G protein, tear-specific prealbumin (lipocalin) and lysozyme (Fig. 4-31). Tear-specific prealbumin is a member of the lipocalin superfamily and together with

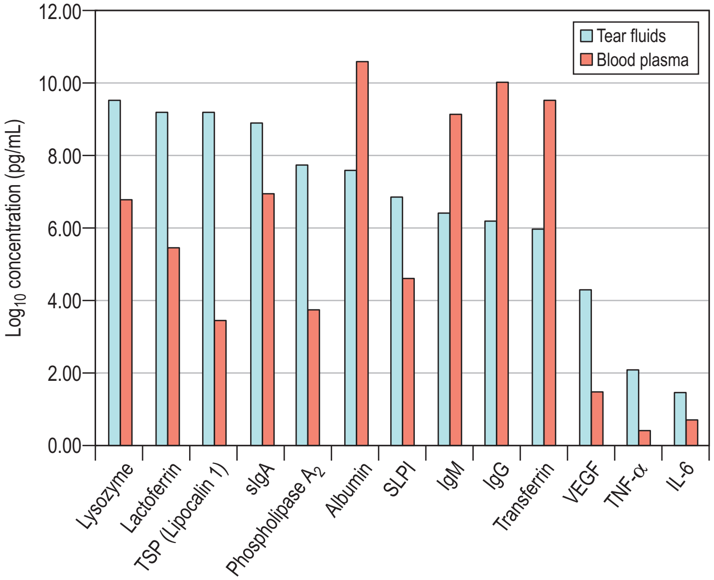

FIGURE 4-31 Comparison of protein concentrations in human tear fluid and plasma to show the remarkable difference between tear fluid and plasma. The proteins shown in this figure represent a wide dynamic range (from mg/mL to pg/mL). (From Zhou, 2012.)

a further lipocalin, apolipoprotein D, is secreted by the lacrimal gland. The function of tear lipocalin is to interact with meibomian gland lipid and induce surface lipid spreading. It may also have a role in removing harmful lipophilic molecules.

Tear secretion by the lacrimal gland is under neural control: basal tear secretion occurs at a rate of 1.2 μL/min, but massive tear production can be induced by a variety of mechanical and psychophysical stimuli. Recent studies in mice also show that tears contain the first discovered soluble male pheromone. Neural control is mediated by the autonomic nervous system (see eFig. 4-11).

Additional content available at https://expertconsult.inkling.com/.

The aqueous component contains several proteins with an antibacterial activity (e.g. lactoferrin and lysozyme), and the high levels of immunoglobulins may have an immunological role (see Ch. 7). Lactoferrin synergizes with lysozyme in its antibacterial action by binding to lipoteichoic acid on the bacterial surface and allowing access of lysozyme to the peptidoglycan. The aqueous component also has anti-adhesive and lubricant properties possibly attributable to lipocalin, thus ensuring that protein and debris generally do not adhere to the corneal surface (e.g. during prolonged periods of lid closure while sleeping).

The mucous layer stabilizes the aqueous layer by providing a hydrophilic contact surface

The mucous layer is composed of the glycocalyx of the epithelial cell surface and an additional layer of tear-specific mucoproteins (Fig. 4-29). Further tear mucins are secreted by the conjunctival goblet cells.

Mucus imparts viscosity to the tear film and is an inclusive term for the entire secretion from the goblet cells (i.e. glycoproteins, proteins and lipoproteins). Mucins are the glycoprotein components of mucus and vary greatly in molecular size (up to 50 × 10^6 kDa). The molecular structure of mucin has been likened to a bottle-brush, where the hairs of the brush are represented by multiple *O*-glycosylated short oligosaccharide chains on an elongated protein 'handle' containing many threonine and serine residues (Fig. 4-32). This structure has the capacity to form many interactions with other hydrophobic (lipid) as well as hydrophilic (protein) and charged molecules, thus forming mucus.

There are many mucins produced by cells of mucosal surfaces and there is some tissue specificity concerning the type of mucin on each surface. Mucins are classified

Moreover, it is likely that free radicals, and in particular H_2O_2, involved in cell damage in various ways. They are not only a cause of acute damage and the stress response but also have a cumulative effect that may contribute to ageing. This may be more than simply a by-product of damage: some of the mitochondrial production of H_2O_2 is generated by an enzyme involved in the apoptosis pathway which induces oxidative stress and may accelerated ageing (eFig. 4-11). Thus the genes that control production may have a say in determining lifespan.

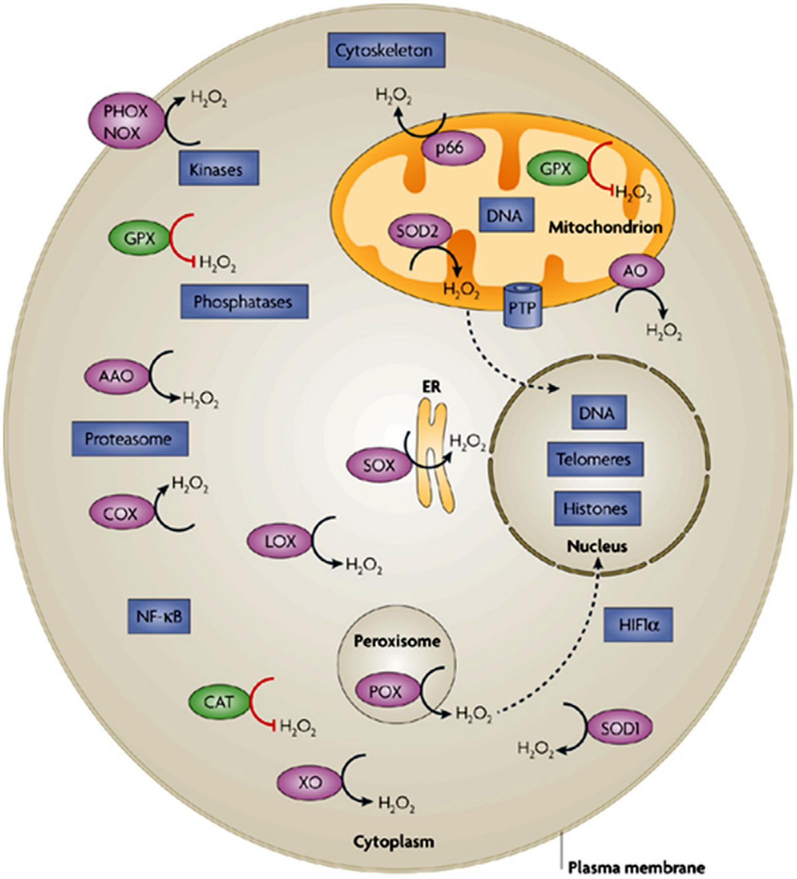
Cytoskeleton
PHOX
NOX
H_2O_2
Kinases
H_2O_2
p66
GPX
H_2O_2
DNA
Mitochondrion
GPX
H_2O_2
SOD2
H_2O_2
PTP
AO
H_2O_2
Phosphatases
AAO
H_2O_2
ER
Proteasome
SOX
H_2O_2
DNA
Telomeres
Histones
Nucleus
H_2O_2
COX
LOX
H_2O_2
NF-κB
HIF1α
Peroxisome
POX
H_2O_2
CAT
H_2O_2
SOD1
H_2O_2
XO
H_2O_2
Cytoplasm
Plasma membrane

eFIGURE 4-11 (1) H_2O_2 is produced by several enzymatic systems (indicated in purple) which generate H_2O_2 in different cellular compartments including phagocytic oxidases (PHOX) and NADP/H oxidases (NOX) in the plasma membrane; superoxide dismutases (SOD2), mitochondrial p66^{Shc} (p66) and amine oxidase (AO) in mitochondria; peroxisomal oxidases (POX) in peroxisomes; sulphydryl oxidase (SOX) in the endoplasmic reticulum (ER); and amino acid oxidases (AAO), cyclooxygenease (COX), lipid oxygenase (LOX), xanthine oxidase (XO) and superoxide dismutases (SOD1) in the cytosol. (2) H_2O_2 disrupts cellular homeostasis by targeting regulatory macromolecules (in blue) including actin, myosin, tubulin (cytoskeleton), different kinases, several Ser/Thr and Tyr phosphatases, the proteasome, the mitochondrial permeability transition pore (PTP), mitochondrial and nuclear DNA, transcription factors such as HIF1 or nuclear factor (NF)-B, and histones and telomeres. Catalase (Cat) and glutathione peroxidase (GPX) enzymes buffer H_2O_2 to avoid cumulative oxidative events. (From Giorgio et al., 2007.)

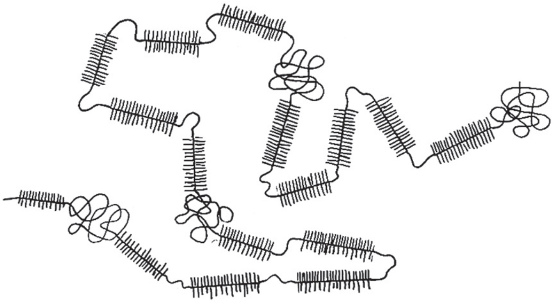

FIGURE 4-32 Bottle-brush mucin. (Courtesy of J Tiffany.)

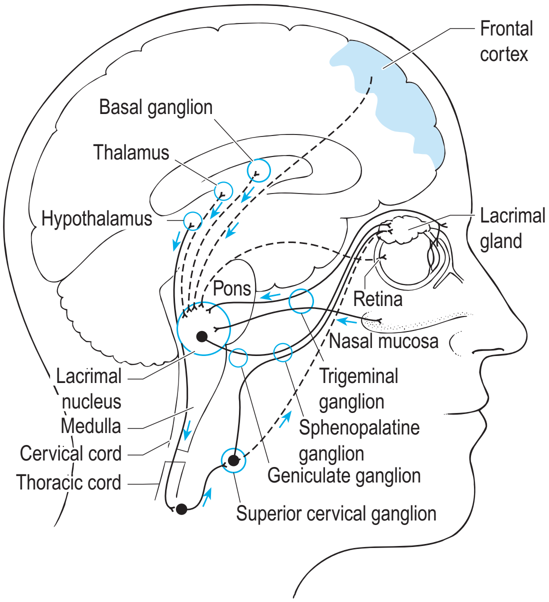

FIGURE 4-33 Neural control of tear secretion.

according to whether they are secreted or membrane tethered: in tears, the main secreted mucins are produced by goblet cells (MUC5AC), by lacrimal acinar cells (MUC7) and by corneal and conjunctival epithelium. Thus, MUC5AC is secreted along with trefoil peptides TFF1 and TFF3. In contrast, MUC1, MUC4 and MUC16 and the sialomucin complex are associated with the glycocalyx and, although regarded as membrane mucins, they may be released into the aqueous phase of tears through activity of ADAM-TS-1 (a disintegrin-like metalloproteinase with thrombospondin type 1 motif) metalloproteases. The ocular surface membrane-tethered mucins are considered to have anti-adhesive properties, particular MUC16 for bacterial adherence. The importance of mucin in tears has been emphasized by the realization that tears are not predominantly aqueous but are probably a type of mucin gel, formed especially by MUC5AC.

An interesting physiological feature of the tear meniscus occurs at Marx's line, a lissamine green and fluorescein staining border of the tear meniscus at the mucocutaneous junction between the conjunctiva and the eyelid skin. Evaluation of this line is assuming greater importance in dry eye disease and it has been suggested that this line represents a zone of tear hyperosmolarity due to the relative immaturity of rapidly proliferating conjunctival epithelial cells, induced at this site by the mechanical forces generated by the repetitive action of eyelid closure. Such cells may produce less tear mucins such as MUC16, which weakens the barrier to penetration by dyes such as lissamine green.

Tear secretion is under psychoneuroendocrine control

The regulation of tear secretion is shown in Figure 4-33. Most of the control is through the autonomic system and thus tear secretion is affected by instillation of drugs that modulate this system (e.g. pilocarpine, atropine). Apart from direct effects on the acinar cells (see above, intracellular second messengers), the parasympathetic system can markedly increase tear flow via its effects on the myoepithelial cells surrounding the acinar cells (see Ch. 1 and eFig. 4-13).

Hormonal control of tear secretion is well recognized but not clearly defined. Reduction in tear flow occurs in women after the menopause, while testosterone stimulates secretion of certain tear components, such as IgA.

THE CONJUNCTIVA

The conjunctiva is richly endowed with a variety of specialized cells that allow it to act as a base for and the source of many of the constituents of the tear film. These include immune cells (T and B cells, mast cells, dendritic cells; see Ch. 7) and specialized mucus-secreting goblet cells.

The epithelium is more than a simple covering layer for the conjunctiva

The conjunctival epithelium can be regarded as intermediate in type between the keratinized squamous epithelium of skin and typical non-keratinized columnar mucosal epithelia, for instance of the respiratory and gastrointestinal tracts. However, it resembles the latter more closely and may require the multifunctional protein clusterin to inhibit keratin production. Clusterin is found in all body fluids including tears and on the surface of cells lining body cavities. It is involved in transport of lipoproteins, inhibiting complement-mediated lysis, and in modulation of cell–cell interactions.

Although the conjunctiva is non-keratinized, keratins are expressed by the conjunctival epithelium in a typical paired combination (K3/K12). However, there is much more K3/K12 in corneal than in conjunctival epithelium. Several other cytokeratins present in non-keratinized stratified (K4 and K13) or simple (K8 and K19) epithelia are also found.

The epithelium contains numerous innate immune cells (macrophages, mast cells, NK cells and dendritic cells) but also has a rich population of intraepithelial and stromal T cells organized to form the conjunctiva-associated lymphoid tissue (CALT), similar to the mucosa-associated lymphoid tissue (MALT) of other organs (see Chs 2 and 6). Even within the CALT, the distribution of lymphocyte subsets varies from tarsal to bulbar to forniceal conjunctival regions (see Ch. 1). In addition to its function as a barrier to external organisms, the epithelium is a major source of tear mucins derived from the intraepithelial goblet cells, which, like the lacrimal gland, are under neuroendocrine control (see eFigs 4.12 and 4.13).

Additional content available at https://expertconsult.inkling.com/.

The turnover and health of the epithelial and goblet cells are markedly dependent on vitamin A and the retinoids, and become abnormal in vitamin deficiencies, leading to a severe form of dry eye syndrome. In addition, certain accumulations of goblet cells occur, e.g. on the tarsal (lid) conjunctiva where they occur in crypts (Henle's crypts) and on the bulbar conjunctiva a few millimetres nasal of the limbus (Manz's glands).

In addition to goblet cells there are other types of conjunctival epithelium, distinguished by ultrastructural appearances: secretory epithelia appear in mature and immature forms depending on their content of secretory granules and the presence of a Golgi complex; in contrast, other cells with a high content of RER and/or mitochondria are presumably involved in transport and epithelial regeneration.

The conjunctival stroma is highly vascular and contains aqueous veins

The stroma of the conjunctiva has a superficial 'lymphoid' layer and a deep layer containing a rich plexus of vessels, including lymphatic vessels, that acts as a watershed between the intraocular circulation and the external circulation of the eye and lids. Through these vessels waste materials from the anterior chamber of the eye are transported to the pre-auricular draining lymph nodes and venous drainage systems in the neck.

THE LIDS

The lids function to protect the cornea and adnexal structures and have a highly specialized structure that ensures they are properly apposed to the surface of the globe (see Ch. 1, p. 82). Indeed, defects in lid apposition, as occur in diseases such as trachoma where there is scarring and deformation of the lids, lead to significant corneal exposure, ulceration and blindness (see Ch. 9).

Closure of the lids leads to compression of the lipid layer of the tears such that it increases in thickness to about 1.0 mm at the lid margin. When the eyes open, the lipids are dispersed to form a lipid bilayer from the lower lid upwards as the eyelid opens, with the hydrophilic groups on the phospholipid molecules interacting with the aqueous compartment.

Control of lid movement is via the VII nerve for motor function and the V nerve for afferent input via mechanisms such as the blink reflex (see Ch. 1, pp. 77–78).

The lid also contains accessory lacrimal glands (glands of Krause in the fornix and glands of Wolfring in the tarsal lid margin), specialized mucus-secreting glands (they secrete a different type of mucin, rMUC4, important in eyelid opening during development, from that secreted by the goblet cells (rMUC5AC, see above) in the conjunctiva) and oil-secreting glands (meibomian glands) that contribute to the makeup of the tear film. Lashes and their hair follicles, as well as sweat glands, are also important specialized structures in the lid, where they function to protect the eye from foreign particles.

Tight neural control of lacrimal secretion is essential to healthy functioning of the tear film. The three major components of the aqueous secretion (water, electrolytes and proteins) are each secreted through different mechanisms, but there is overall neural regulation of the final secreted product so that a coordinated response can be achieved. Sensory input comes via the trigeminal nerve endings in the cornea and conjunctivae to induce efferent parasympathetic and sympathetic nerves that innervate the lacrimal gland (eFig. 4-12) to release neurotransmitters which effect the release of the tear fluid (eFig. 4-13).

Tear secretion is dependent on several autonomic receptors and ligands apart from the more well-recognized adrenergic and cholinergic mediators (eFig. 4-13).

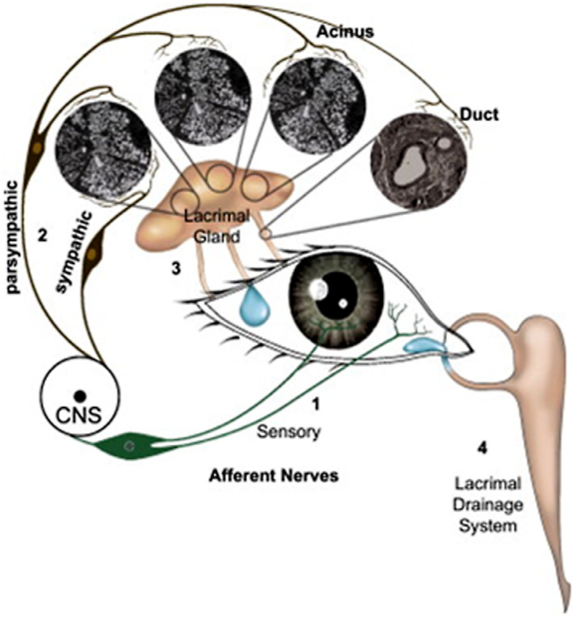
Acinus
Duct
parsympathetic
2
sympathic
3
Lacrimal Gland
CNS
1
Sensory
Afferent Nerves
4
Lacrimal Drainage System

eFIGURE 4-12 The neural anatomy and physiology of lacrimal gland tear secretion. (From Dartt, 2009.)

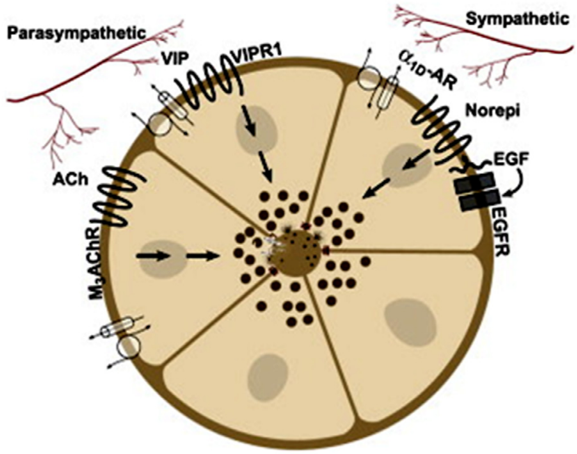
Parasympathetic
VIP
VIPR1
Sympathetic
α_{1D}-AR
Norepi
EGF
EGFR
ACh
M_3AChR

eFIGURE 4-13 Release of acetylcholine occurs from parasympathetic nerve endings as well as vasoactive intestinal peptide (VIP) while noradrenalinr is released by sympathetic nerves. Additional mediators such as EGF and a1d-AR contribute to the overall response. (From Dartt, 2009)

Incomplete eyelid closure may be physiological during sleep. Control of eyelid closure is under both reflex supranuclear and learned/conditioned (cerebellar) control.

Cornea and sclera

The outer coat of the eye (cornea/sclera) is a tough non-compressible layer of connective tissue that can withstand considerable deformation and pressure.

THE CORNEA

The clear cornea functions as the main optical lens of the eye, focusing transmitted light rays onto the retina. Its cellular and extracellular matrix components are of the same basic chemical composition as other opaque tissues in the body which scatter, rather than transmit, light rays. Scattering of light in opaque tissues is the result of the large disparity in refractive index (RI) between matrix components such as collagen (RI = 1.55 in the dry state) and glycosaminoglycans (RI = 1.35). The cornea's ability to transmit light is, in essence, a function of how the cells and matrix components are organized within the tissue to reduce this RI disparity.

Corneal transparency is a function of its relative acellularity and matrix structure

The epithelium. The six-cell-thick stratified layer that is the corneal epithelium (see Ch. 1, p. 14) presents the first refracting interface to transmitted light. Most of the light-absorbing properties of the cornea take place in this layer, mainly for short-wavelength light. However, the majority of light on the visible spectrum is transmitted through the epithelium.

The cells are typical keratin-expressing epithelial cells containing integrin receptors for basement membrane components such as fibronectin, laminin and collagen. Corneal epithelial cells express a particular combination of paired 55/64 kDa keratins (keratin 3/keratin 12). The 54 kDa protein may have a role in inflammatory eye disease, while the 64 kDa protein appears to be useful as a marker for differentiating cells of the central cornea from limbal stem cells (see below). Keratin 12 may be important in corneal epithelial junctions because K-12 knockout mice are prone to recurrent erosion (epithelial cell loss).

The cells are organized to present few interfaces, the most prominent being at the interface between the basal cells and the basement membrane. Hemidesmosomes affect the adhesion between these cells and the basement membrane. The hemidesmosome is bound to the anterior corneal stroma through a band of anchoring fibrils, which pass through the lamina densa of the basement membrane proper into a densely woven network of collagen fibres (see Box 4-6 and Fig. 4-20) known as Bowman's layer (see Ch. 1, p. 16) (Fig. 4-34). Bowman's layer is about 12 μm thick and composed of type VII collagen. In addition, trans-plasma membrane collagen type XVI supports firm adhesion in these basal cells.

The epithelium presents an effective barrier to fluid transport, which is achieved by extensive close contacts and tight junctional complexes eBox 4.2. The site

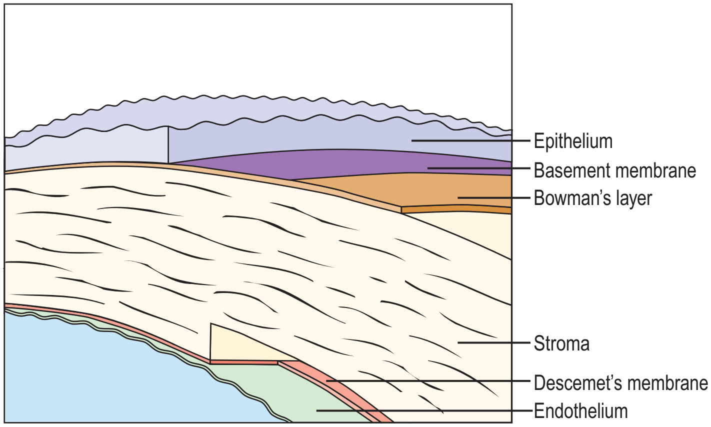

FIGURE 4-34 Diagram demonstrating the layers of the human cornea: the anterior basement membrane, Bowman's layer, the stroma, Desçemet's membrane and the endothelium. (From Last et al., 2009.)

of the barrier is in the suprabasal epithelium and is mediated by high expression of the tight-junction protein claudin. Spot desmosomes are numerous and differences in the content of desmosomal proteins have been observed depending on the site. For instance, desmoglein and desmocollin are absent from basal limbal epithelial cells, which may have functional significance in their role as putative stem cells. Limbal stem cells arise from pouches or stem cell niches in the peripheral corneal epithelium.

Matrix factors affecting transparency: collagen. The stroma in the human cornea accounts for 90% of the corneal thickness and thus contributes most to the optical function of the tissue. Several different types of collagen are present in the cornea. In addition to normal basement membrane types IV and VIII collagens at the epithelial and endothelial cell layers, the two specialized corneal regions, Bowman's layer and Desçemet's membrane, contain collagens not normally found in other matrices. Bowman's layer is a condensation of type I/VI with a high proportion of type III in a matrix containing chondroitin and dermatan sulphate, while Desçemet's membrane contains high levels of novel collagens (types V, VIII, IX and XII) organized in a lattice arrangement. This provides elasticity and deformability to the cornea while maintaining high levels of light transmission. Desçemet's membrane also imparts strength and resilience to the corneal stroma and is the main resistance to normal intraocular pressure.

The orthogonal lamellar arrangement of stromal type I collagen (which accounts for 50–55% of stromal collagen; see Ch. 1) fibrils is considered to be important in determining corneal transparency. Transparency was initially attributed by Maurice (1957) to 'destructive interference' in which light is scattered by neighbouring fibrils in predictable and opposing directions, which tend to cancel each other out except along the primary visual axis. However, this concept cannot apply to light transmission in Bowman's layer where the fibrils are irregularly displayed, suggesting that the arrangement of the fibrils is less important than their size. An alternative view therefore is that significant light scatter does not occur within the cornea because the fibril diameter does not exceed 30 nm and the interfibrillar distance is around 55 nm. It is only when the distance between the regions of different refractive index becomes greater than 200 nm that light scatter occurs, as for instance when corneal swelling occurs. In one sense this is essentially a paraphrase of Maurice's theory because in both situations the critical factor is interfibrillar distance.

Type V (approximately 10%) and some type III (1–2%) collagen also exist in the corneal stroma, while the remainder is made up of type VI collagen. Types I, III and V collagen are fibrillar collagens, but type VI has large non-helical globular polypeptides at both the

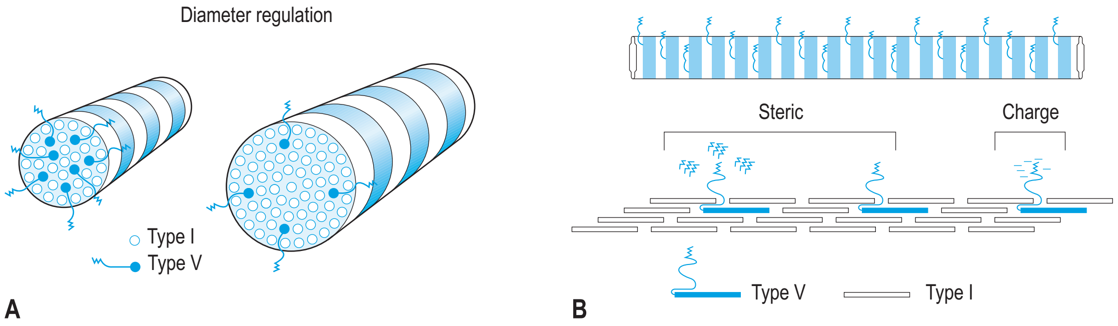

FIGURE 4-35 Model for type V regulation of collagen fibril diameter. Collagen fibrils within the cornea contain a high percentage of type V collagen and have small diameters (**A**), while fibrils in other tissues possessing low levels of type V collagen have a large diameter (**A**). The experimental reduction of type V collagen generates fibrils that have the characteristics of those found in tissues with low type V levels. A mechanism by which type V collagen may limit fibril growth is shown in (**B**). Type V N-terminal domains project onto the fibril surface, and when sufficient numbers have accumulated, they block further accretion of collagen monomers and thereby limit growth in diameter. The N-terminal domain is large and possesses a number of acidic residues, and so may affect this block using steric and/or electrostatic hindrance. (Courtesy of David E. Birk.)

C- and N-termini of the helical protein backbone (see Table 4-2). Collagen fibrils are heterotypic (i.e. they contain both type 1 and V collagen) and type V is thought to initiate fibril formation and regulate fibril thickness (Fig. 4-35). However, fibril thickness is also dependent on the nature of the stromal glycosaminoglycans and particularly on the specific proteoglycan (see below).

The collagen fibril lamellae are arranged in parallel, running at oblique angles to each other (see Ch. 1, p. 17). Apart from their small size, their uniformity of thickness is likely to be a major factor in light transmission. The parallel arrangement of the central corneal fibrils extends to the periphery where the fibrils adopt a concentric configuration to form a 'weave' at the limbus with some transversely running fibrils fusing with the circumferential collagen fibrils. More recently, it has been proposed that scleral collagen fibrils take a curved orientation across the peripheral cornea, thus acting as anchoring fibrils which flatten the peripheral cornea at this point (Fig. 4-36). This imparts considerable strength to the peripheral cornea and permits it to maintain its curvature and thus its optical properties.

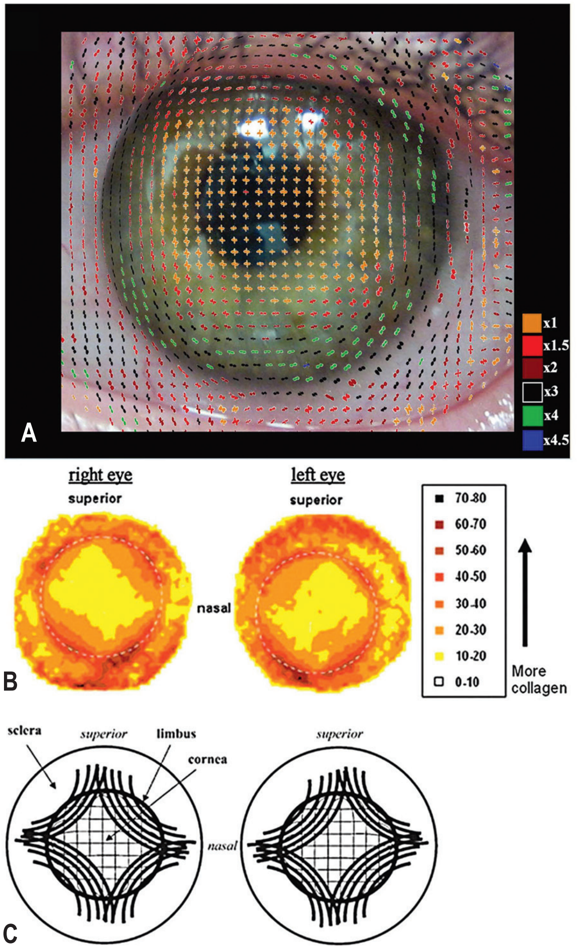

FIGURE 4-36 (**A**) X-ray diffraction data showing the preferred directions of stromal collagen lamellae across the human cornea, limbus and anterior sclera. The scale of the X-ray vector plots (indicated in the colour key) reflects the degree of collagen alignment. (**B**) Maps of X-ray scatter intensity showing the distribution of preferentially aligned lamellae in a pair of eyes from the same human donor. The limbus is denoted by a broken line. Note the symmetry between left and right eyes. (**C**) Theoretical model showing the net course of lamellae and based on the data shown in **A** and **B**. Lamellae are thought to change direction in the corneal periphery by a process of splitting and interweaving. (Figure modified from Meek and Boote, 2009.)

Glycosaminoglycans. The corneal stroma is unusual in that it contains no hyaluronan, except at the limbus where there is a gradual increase in concentration towards the sclera. The major corneal glycosaminoglycan is keratan sulphate; in the central cornea nonsulphated chondroitin is also present, while towards the periphery chondroitin sulphate is the second major GAG. Chondroitin-4-sulphate and dermatan sulphate are almost identical (see Box 4-7) and many believe that the second major GAG is not chondroitin but dermatan sulphate.

Corneal GAGs exist in the native state as proteoglycans (PGs); four major forms exist in the cornea: decorin, lumican, keratocan and mimecan. Decorin- and lumican-deficient mice have reduced corneal transparency due to disorganization of fibril arrangement, while keratocan maintains overall corneal thickness and less so mimecan. Both corneal dermatan sulphate and keratan sulphate proteoglycans are considered to belong to the class of small non-aggregating PGs (small leucine-rich PGs known also as SLRs).

The critical region in proteoglycans is their linkage sites; for keratan sulphate the link is an *N*-glycosidic bond between *N*-acetylglucosamine and asparagine in the core protein. The terminal sites of the branched oligosaccharide structures contain fucose or mannose, while chondroitin sulphate contains xylose residues.

The interaction between proteoglycans and collagen fibrils has been elegantly demonstrated using cupromeronic blue and $MgCl_2$ at a tightly controlled

(3.0 mM) concentration (the 'critical electrolyte concentration'). Both dermatan sulphate proteoglycan and keratan sulphate proteoglycan bind to the collagen arrays at specific binding sites (one proteoglycan to one binding site), suggesting that these sites are essential to the spacing of the fibrils and to the thickness of the interfibrillar space. Keratan sulphate proteoglycans appear to bind to the step regions of the fibrils while dermatan sulphate proteoglycans bind to the gap (Box 4-11).

BOX 4-11 ORGANIZATION OF TISSUE PROTEOGLYCANS

The organization of proteoglycans in the collagen matrix is probably as shown below: KS-PG (keratan sulphate proteoglycan) double complexes maintain the lateral interfibrillar distance, while CS-PGs (chondroitin sulphate proteoglycans) regulate the overall longitudinal arrangement of the fibrils by spanning three fibrils (**A**). Proteoglycans bind directly to the collagen fibrils, probably through the 'minor' collagens such as V or VI that co-distribute with the type I fibrils and have non-collagenous polypeptide domains to interact with the proteoglycans (**A** and **B**).

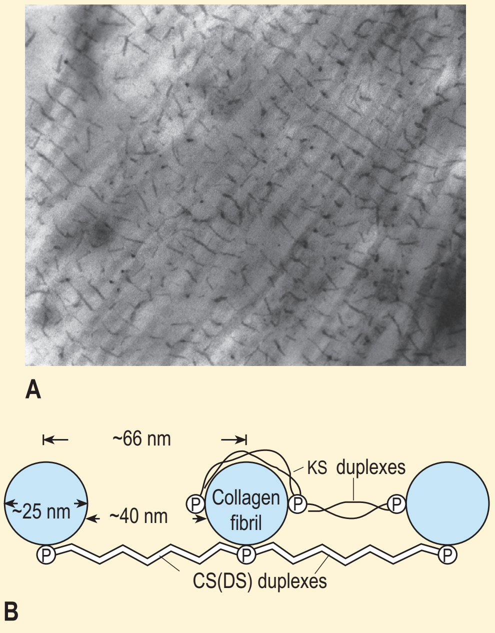

(Figure courtesy of J Scott and IRL Press.)

Different binding 'maps' occur in different species, such as the mouse. Therefore, several variations on the theme of proteoglycan–collagen interaction appear to be compatible with transparency. However, SLRs appear to be essential. Studies in human tissue developed this model to reveal a regular hexagonal arrangement of six proteoglycans per collagen fibril interacting with the 'next but one' fibril (Fig. 4-37).

Swelling pressure versus hydration. The cornea is about 80% hydrated. This is higher than other tissues such as the sclera, which is about 70% hydrated. Despite this, the corneal stroma 'imbibes' water if it is placed in a solution of saline (this is well demonstrated by the injured, lacerated cornea which swells and becomes opaque); the water-attracting 'hygroscopic' properties of the cornea are the result of its high content of GAGs. The cornea has been described as 'a

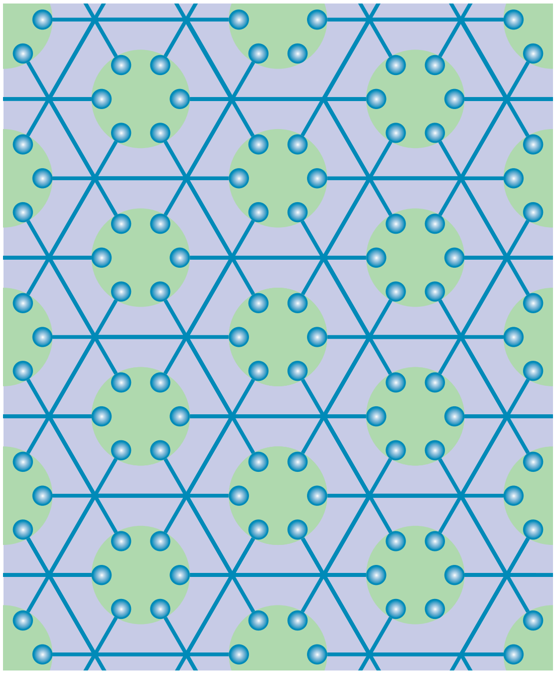

FIGURE 4-37 Schematic visualization of the basic components of a new model for the corneal collagen lattice. At equidistant sites along their circumference, six core proteins of proteoglycans are attached to the hexagonally arranged collagen fibrils. GAGs of the proteoglycans as stained by CB are connecting their next nearest neighbour collagen fibrils and form a ring-like structure around each collagen fibril. (From Muller et al., 2004, with permission from Elsevier.)

slice of water stabilized in three dimensions by a meshwork of fibrils and soluble polymers'. The cornea thus has a swelling pressure and a metabolic pump (the endothelium) designed to maintain it. The swelling pressure generates a level of interfibrillar tension and this may be the biophysical mechanism whereby the fibrils are maintained in their normal arrangement. In addition, the swelling pressure itself may reciprocally activate chloride channels and other transporters that maintain a balance of excess ions in the aqueous humour.

The swelling pressure of the cornea is an important clinical concept and attempts have been made to measure swelling pressure using instruments such as the Ocular Response Analyzer® (ORA), an instrument similar to an applanation tonometer used for measuring intraocular pressure. The ORA measures corneal hysteresis, which is a function of biomechanical strength of the corneal tissue related to the viscosity of the matrix and its response to deformation. However, whether the ORA truly measures hysteresis independently of ocular pressure is still unclear, despite its considerable relevance for corneal refractive surgical procedures.

Cellular factors affecting transparency: (1) keratocytes. Corneal fibroblasts (keratocytes) are important in maintaining transparency because they are the source of stromal collagens and proteoglycans. Although most of the changes that occur in the assembly of the matrix are post-translational, the enzymes that promote these changes are present in the keratocytes, in which the essential specific genes have been induced (specific enzyme defects are associated with corneal opacification as in the mucopolysaccharidoses). Keratocyte dysfunction may also underlie the corneal haze seen after corneal refractive surgery.

Collagen turnover in early postnatal life is about 24–50 hours, but there is little information on adult collagen metabolism. For both collagen and GAGs, studies on cultured keratocytes have not been very informative because these cells produce a range of GAGs not found *in vivo*. In contrast, organ cultures of cornea produce a panel of GAGs more akin to that found *in vivo*. The preferential production of KS-PG to CS/DS-PG (dermatan sulphate proteoglycan) in corneal cells from different species has been attributed to the relatively hypoxic conditions and to anaerobic glycolysis, which favours the former. In rabbit cornea, the development of transparency correlates with a dramatic increase in the concentration of KS-PG in the early postnatal period.

Keratocytes also express a range of corneal 'crystallins', so called because they are thought to play a role in light transparency by reducing light scatter and not because of homology to lens crystallins (see below). Corneal crystallins include trans-ketolase and aldehyde dehydrogenase and expression varies from species to species.

Cellular factors affecting transparency: (2) leucocytes. The cornea also contains a population of resident stromal and intraepithelial leucocytes which are more frequently found in the peripheral cornea (see Chs 1 and 7). These cells include macrophages and dendritic cells which are thought to be important in the maintenance of ocular immune privilege (see Ch. 7), as well as rare haemopoeitic stem cells. These are distinct from limbal corneal stem cells of mesenchymal origin, which reside in epithelial stem cell niches (see below).

Cellular factors affecting transparency: (3) the endothelium. The endothelial pump determines the level of hydration of the GAGs and thus transparency. Despite the greater than normal level of hydration of the corneal stroma, its water binding is unsaturated, a condition achieved by an endothelial pump (Box 4-12), which transports water out of the cornea towards the anterior chamber. This is known as the pump–leak mechanism and is an energy-driven ATPase-dependent mechanism (see p. 208). Transport of Cl^- ion in this model probably depends on the cystic fibrosis transmembrane conductance regulator (CFTR), a Cl^- channel which is expressed in many fluid transporting epithelia in the body.

Other mechanisms may be also be operative in the transport of fluid across the endothelium. These include an Na^+/H^+ exchanger protein which drives an electrogenic coupling between Na^+ and HCO_3^- ions. This 'antiport' is essential for maintenance of the intracellular pH by exchanging Na^+ (in) with H^+ (out) in the cell (Box 4-12). This may be one component of the electro-osmosis concept for fluid transport in the

BOX 4-12 CORNEAL ENDOTHELIAL CELL PUMP

The corneal endothelium transports water out of the stroma by an ATP-driven ion pump mechanism.

Na^+ and HCO_3^- are transported across the endothelium from the stroma to the aqueous, mediated by a Na^+/K^+ -dependent ATPase and a HCO_3^- -dependent ATPase, probably involving carbonic anhydrase. The Na^+/K^+ ATPase is located in the plasma membranes but the HCO_3^- -dependent ATPase is in the mitochondria, where its major role in ion transport may be to generate the ATP required for the Na^+/K^+ ATPase. Carbonic anhydrase may also be involved in the Na^+/H^+ antiport for maintaining intracellular pH.

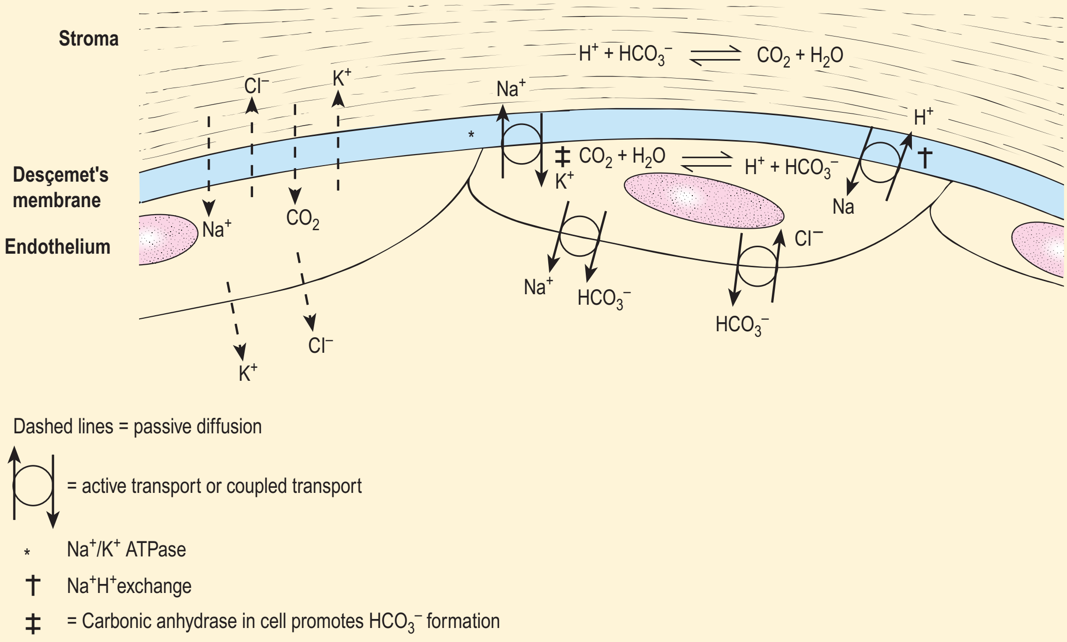
Stroma
Desçemet's membrane
Endothelium
H⁺ + HCO₃⁻ ⇌ CO₂ + H₂O
Cl⁻
K⁺
Na⁺
H⁺
*
‡ CO₂ + H₂O ⇌ H⁺ + HCO₃⁻
K⁺
Na⁺
CO₂
Na
Cl⁻
Na⁺
HCO₃⁻
HCO₃⁻
K⁺
Cl⁻
†
Dashed lines = passive diffusion
= active transport or coupled transport
* Na⁺/K⁺ ATPase
† Na⁺H⁺exchange
‡ = Carbonic anhydrase in cell promotes HCO₃⁻ formation

cornea. In addition, the corneal endothelial cell membrane contains a water transporter/channel, namely aquaporin 1, which is involved in bulk transport of water molecules (see below under Lens) and perhaps also in CO_2 transport. This, however, seems to be mainly a passive transporter, responding to sudden changes in stromal water content and endothelial cell volume rather than having a constitutive physiological role.

Recently, a metabolic component to fluid transport across the cornea has been suggested based on the high content of lactate in the stroma and the steep gradient towards the anterior chamber. A bulk movement of lactate and water could be facilitated via transcellular monocarboxylate co-transporters in the endothelium.

Transport of water out of the cornea is a dual process. In addition to the transport of water from the stroma to the aqueous humour, there is a net flux of ions and water towards the epithelium and the tears. For instance, there is a Cl^- pump which appears to be modulated by several receptors including β-adrenergic and serotonergic receptors coupled to the adenylate cyclase and Ca^{2+} second messenger systems, and receptors that involve protein kinase C. Dopamine and α-adrenergic receptors may also be involved in Cl^- transport, but importantly many of these receptor systems are coupled to the CFTR Cl^- channel expressed in corneal and conjunctival epithelia (see eBox 4.3).

Ion transport into the cornea mediated by Na^+/K^+ ATPase and Ca^{2+}/Mg^{2+} ATPase in the basolateral plasma

eBox 4-3

Agonist receptors in the corneal epithelium

Adrenergic and cholinergic receptors are present in the corneal epithelium and are coupled to a variety of second messenger systems including cAMP, phosphatidyl inositol phosphate 2 (PIP_2), and protein kinase C (PKC) (**A**). They are involved in the transport of fluid out of the cornea in an anterior direction by enhancing the function of the HCO_3^-/Cl^- pump (**B**). Active pumps for outward extrusion of Cl^- (activated by β-adrenergic and other receptors) and inward transport of Na^+ and HCO_3^- are present. Dashed line, passive transport; circles, coupled/active transport. There is also a separate water transporter, aquaporin 5, in the corneal epithelium, which is important for the maintenance of normal corneal stromal thickness.

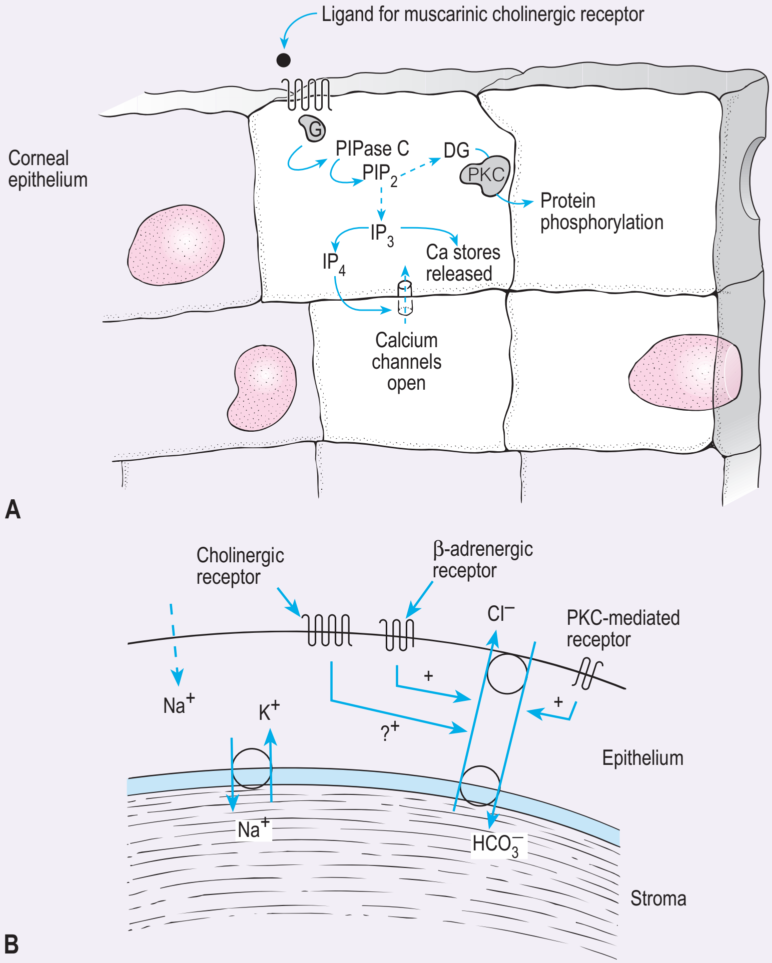
Ligand for muscarinic cholinergic receptor
Corneal epithelium
G
PIPase C
PIP_2
DG
PKC
Protein phosphorylation
IP_3
IP_4
Ca stores released
Calcium channels open
A
Cholinergic receptor
β-adrenergic receptor
Cl^-
PKC-mediated receptor
+
+
Na^+
K^+
$?^+$
Epithelium
Na^+
HCO_3^-
Stroma
B

Laminin-332 (eFig. 4-6) is specific to the epithelial basement membrane and is central to epithelial sheets bound to mesenchyme which have to withstand external forces as in the skin and other mucosal layers. Matrix metalloproteinases are involved in post-translational processing of laminins as they incorporate themselves into the basement membrane to form anchoring adhesions and reciprocally, MMPs (MMP-3, -12, -19 and -20) are activated by laminin degradation products to process the γ2 chain and induce epithelial migration during wounding. Laminin in the basement membrane is tightly bound to nidogen (entactin), a protein that mediates the binding of laminin to heparin-containing proteoglycans (heparin sulphate proteoglycan) such as perlecan and syndecan in the basement membrane matrix.

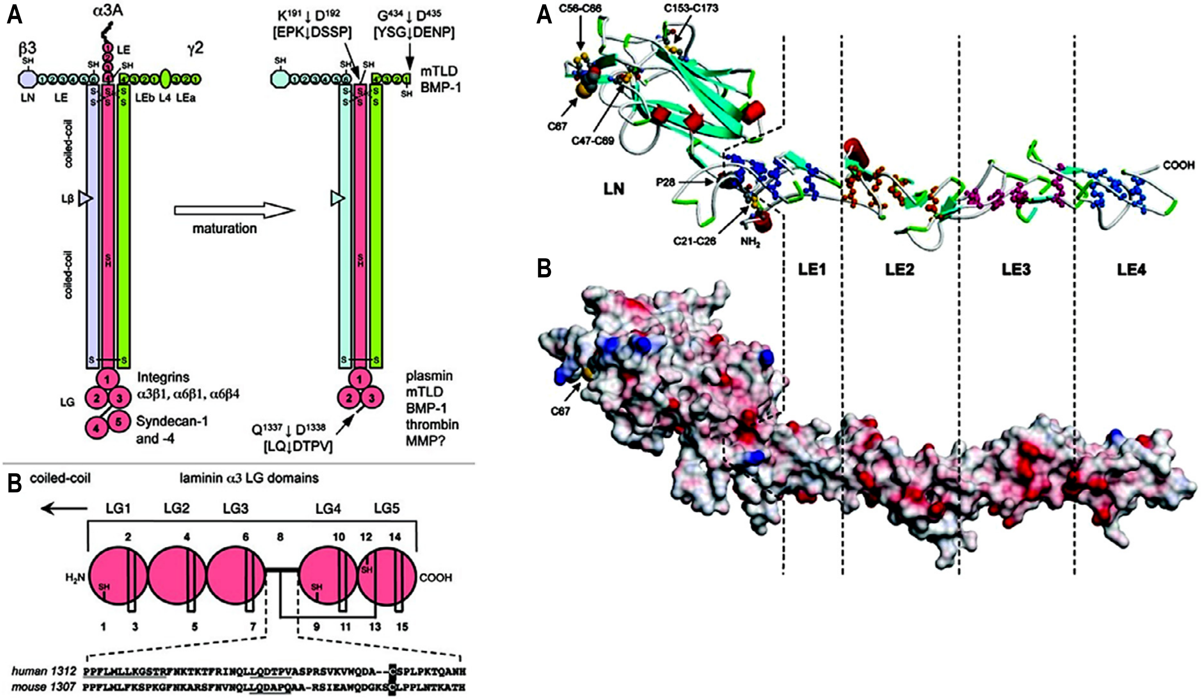
A
α3A
β3
SH
LE
γ2
LN
LE
LEb L4 LEa
coiled-coil
Lβ
coiled-coil
maturation
Integrins
α3β1, α6β1, α6β4
LG
Syndecan-1
and -4
K191 ↓ D192
[EPK↓DSSP]
G434 ↓ D435
[YSG↓DENP]
mTLD
BMP-1
Q1337 ↓ D1338
[LQ↓DTPV]
plasmin
mTLD
BMP-1
thrombin
MMP?
B
coiled-coil
laminin α3 LG domains
LG1
LG2
LG3
LG4
LG5
H2N
COOH
2
4
6
8
10
12
14
1
3
5
7
9
11
13
15
human 1312 PPFLMLLKGSTRFNKTKTFRINQLLQDTPVASPRSVKVWQDA--CSPLPKTQANH
mouse 1307 PPFLMLFKSPKGFNKARSFNVNQLLQDAPQAA-RSIEAWQDGKSCLPPLNTKATH
A
C56-C86
C153-C173
C67
C47-C69
LN
P28
C21-C26
NH2
COOH
B
LE1
LE2
LE3
LE4
C67

eFIGURE 4-6 Structure of human laminin. Human laminin is known as laminin 332 on the basis of its three subunit chains (α3A, β3 and γ2). Each chain is composed of different domains that are indicated in (**A**). An odd number of cysteine groups allows SH bonds to form in the L domains after maturation of the molecule. The first three LG domains (LG1–3) interact with α3β1, α6β1 and α6β4 integrins while the last two (LG45) contain binding sites for syndecan-1 and -4. Newly synthesized laminin 332 undergoes maturation by proteolytic processing at the α3A chain N- and C-terminus as well as at the γ2 chain N-terminal extremity. A schematic structure of laminin domains is shown in (**B**), with numbered cysteine groups and the disulphide bridges.

membrane of the epithelium is also operational in corneal ion shifts. Na$^+$ movement from the tears into the epithelium is by passive diffusion down a concentration gradient, but from the epithelium into the stroma active transport is required via the Na$^+$K$^+$ ATPase to which Cl$^-$ transport in the opposite direction is coupled, again by CFTR channels. This transport system is sensitive to the eicosanoid metabolite 12(R)HETE (compound C) (see Box 4-10). In addition, this electrolyte flux across the epithelium accounts for the electrical potential difference from (−) outside to (+) inside the cornea of ~25–40 µV.

Optical factors affecting image formation. The curvature of the cornea plays a major part in refracting and focusing light to produce an image on the retina. Even with a completely clear cornea, the image can be distorted by abnormalities in curvature, producing the various forms of regular and irregular astigmatism. A certain degree of astigmatism is present in many eyes since the corneal curvature is rarely perfectly spherical. These account for some refractive errors which are common in the healthy population. Refractive errors can also be caused by variations in the overall ocular dimensions of the eye.

Additional content available at https://expertconsult.inkling.com/.

Severe degrees of astigmatism can be caused by disease of the cornea such as keratoconus (‘cone-shaped cornea’), and the late effects of scarring from wounds such as corneal incisions following cataract surgery, when it can be very difficult to restore preoperative curvature. While these conditions do not directly affect the transparency of the cornea, they can increase the amount of spherical and chromatic aberration, and diffraction, thus degrading the image.

Metabolism of corneal cells

Oxidative metabolism and glucose utilization. The epithelium takes up most of its glucose from the stroma and converts it to glucose-6-phosphate, after which 85% is metabolized via the glycolytic pathway to pyruvate. The bulk of this is then metabolized to lactic acid, but some is diverted into the citric acid cycle to produce ATP as an energy store. The pentose phosphate pathway accounts for the remainder of glucose utilization by the epithelium, producing an important resource for free radical control, namely NADPH (see Fig. 4-23). This is the main mechanism of generation of reducing agents such as glutathione and ascorbic acid. The corneal epithelium has also developed a unique nuclear ferritin-based mechanism to minimize DNA damage from ultraviolet (UV) light-induced free radical damage. Lactoferrin in tears may also assist in this process.

The metabolism of keratocytes is mostly concerned with generating sufficient energy to produce and maintain stromal components (see above). In the steady state, these cells are not highly metabolically active and there is a high stromal lactate production, possibly contributing to fluid transport across the endothelium (see above).

The endothelium has large energy requirements to sustain its ATPase-dependent pump mechanism and is about five times as active as the epithelium. The major metabolic pathway in the endothelium is anaerobic glycolysis, with the citric acid cycle and the pentose phosphate pathway also playing a significant role.

Oxygen handling by the cornea. The epithelium obtains its oxygen from the preocular tear film at a rate of 3.5–4.0 µL per cm^2 per hour. The endothelium, however, and the keratocytes in the deep stroma receive their oxygen supply from the circulation via the aqueous humour. Corneal function and health are therefore dependent on local conditions at the surface of the eye and on systemic factors such as cardiopulmonary capacity.

Oxygen is consumed in the citric acid cycle, generating 36–38 molecules of ATP per molecule of glucose. The utilization of the citric acid cycle versus the glycolytic pathway is determined by the energy demands of the tissue, specifically the need for ATP. Thus the endothelium makes greater use of the citric acid cycle than the epithelium. In addition, oxygen consumption by the cornea increases almost twofold when acidosis prevails, as occurs in contact lens wear. This is in part the result of the activation of pH regulatory mechanism, including Na$^+$/H$^+$ exchange, which then stimulates Na$^+$/K$^+$ ATPase activity.

As discussed above, excess oxygen can be detrimental to the organism if it is converted to the

Briefly, a ‘normal-sized’ eye, anteroposterior dimension 22–24 mm, is termed an emmetropic eye and is one in which the image is focused sharply on the retina by the combined refracting properties of the cornea and the lens. Myopia or short-sightedness occurs in eyes with above average length and the image therefore is focused in front of the retina; conversely, long-sighted eyes have a shorter than normal length and the image is focused behind the retina. Myopia can be corrected by placing a convex lens in front of the cornea, which thus causes some divergence of the light rays and with the correct lens, the image can be formed on the retina correctly. Myopia can also be corrected by ‘reshaping’ the cornea (see p. 213) to cause some degree of ‘flattening’ of its curvature and thus reduce its refractive power. Conversely, hyperopia can be corrected using a convex lens and similarly corneal surgery can be performed to correct hyperopia, although it is generally less successful. Further information on refraction with links to optometry literature is given in Chapter 5 (see eFig. 5-4).

superoxide radical and then to hydrogen peroxide (see Box 4-8 and Fig. 4-26). In both the epithelium and the endothelium, redox systems involving glutathione and its two enzymes (glutathione reductase and glutathione peroxidase) depend on the generation of NADPH and thus on a supply of glucose. When the intracellular levels of glutathione are reduced in the cornea by one-third, the clarity of the cornea and its ability to pump fluid decline dramatically.

The aqueous contains high levels of H_2O_2, perhaps by virtue of the reduction of oxygen in the aqueous via ascorbate usage. Free radical damage to the corneal endothelium induces apoptosis and thus may account for the progressive endothelial cell loss associated with age.

Free radical damage to the cornea can also be induced by therapeutic interventions such as that incurred during phacoemulsification for cataract and by the use of UV radiation to preform corneal collagen cross-linking as part of refractive surgery (see p. 213).

Several receptors for neurotransmitters and other agonists are present in the corneal epithelium. These are probably involved in nutrient handling and regulation. In addition, since the cornea is richly innervated, they probably contribute to a healthy nutritive neurogenic environment, preventing neuropathic keratopathy.

Effects of contact lens wear on corneal physiology. Contact lenses are common optical devices which correct the refractive errors of the eye to achieve emmetropia. The tear film bathes both surfaces of the contact lens (Fig. 4-38), but corneal epithelial function is still relatively compromised. The corneal epithelium layer receives its oxygen from the tears and its glucose from the circulation via the aqueous and the limbal vessels (see above). Contact lenses reduce the direct availability of oxygen to the epithelium, thus shifting the balance from aerobic to anaerobic metabolism. The already high lactate levels in the cornea are doubled with contact lens wear and carbon dioxide production is increased. The induced acidosis has a direct effect on stromal hydration by impairing deturgescence mechanisms (see above).

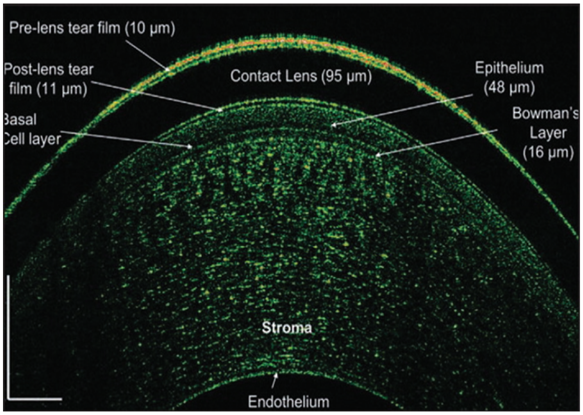

FIGURE 4-38 Ultra-high resolution optical coherence tomography image of the central cornea with a PureVision (Bausch & Lomb, Rochester, NY) lens after instillation of artificial tears. The central cornea was imaged with 6 mm scan on the horizontal meridian. The image was taken immediately after lens insertion and instillation of one drop of artificial tears. The epithelium, including the basal cell layer, and Bowman's layer, are evident in addition to the pre-lens and post-lens tear films. Total corneal thickness was measured at 526 μm. Bars = 250 μm. (Reprinted with permission from Wang J, Jiao S, Ruggeri M, Shousha MA, Chen Q. In situ visualization of tears on contact lens using ultra-high resolution optical coherence tomography. *Eye & Contact Lens* 2009;2:44–49.)(Wang, 2011)

Hard (rigid) contact lenses are usually made from polymethylmethacrylate (PMMA) and have the greatest effect on corneal function; in addition to restricting oxygen availability, hard lenses deplete glycogen stores, even though the level of glucose availability is not reduced. It has been suggested that hard lens-induced inhibition of aerobic enzymes such as hexokinase reduces direct glucose utilization by the cornea. Prolonged wear of hard contact lenses is therefore not possible, owing to the damaging effect on corneal transparency induced by the disturbed metabolism. Soft contact lenses are made from polymers of hydroxyethyl methacrylate (HEMA), poly(HEMA/vinylpyrrolidones), silicone or other similar materials, and permit extended wear of the lens owing to their permeability to oxygen and carbon dioxide. However, there is still some degree of lactate accumulation with soft lenses and prolonged use appears to affect the function of the endothelium. HEMA-based lenses are hydrophilic, while silicone-based lenses are hydrophobic: accordingly, there is less protein deposition on the latter but greater levels of denatured proteins. However, corneal inflammatory episodes are if anything more frequent with silicone lenses. Manufacturers of contact lenses continually produce new 'biomimetic'-type lenses with increasing water content (up to 59%) in attempts to support normal corneal physiology (hydrogel lenses). In addition, incorporating other material into contact lenses such as polyethylene glycol and cross-linked hyaluronan are other possibilities under consideration to improve biocompatibility.

A popular compromise in contact lens type is the gas-permeable rigid lens, which combines the reduced toxicity of PMMA with high gas-transfer capability. The wide variety of lens types and materials has led to their being characterized on the basis of their oxygen flux, defined as the *DK* value:

$$\text{Oxygen flux} = DK/L \times \Delta P$$

where D is the diffusion coefficient, K is the solubility, and L is the thickness of the lens material. ΔP is the change in the partial pressure of oxygen across the material. HEMA and PMMA have a low oxygen flux, while hydrogels and silicones have a high flux. Both the thickness of the lens and the *DK* value determine its suitability for use in terms of its gas permeability. The actual amount of oxygen that reaches the cornea is the most important factor in the design of a contact lens, and most practitioners describe contact lenses in terms of their equivalent oxygen performance (EOP).

Contact lenses may have deleterious effects on the epithelium, causing thinning, reduction in the hemidesmosome density and the number of anchoring fibrils, and reduced adhesion of the epithelium to the basement membrane. This may be a direct effect of low O_2-transmitting lenses on basal epithelial cell proliferation. This is especially true of extended-wear hydrogel lenses. In severe cases, excessive use of contact lenses produces epithelial oedema and keratopathy in the form of punctate epithelial erosions. Rigid contact lenses also produce tear film instability by causing damage to the epithelium and the mucin layer in particular.

Contact lens wear may also induce changes in the corneal stroma (thickening) and the endothelium (polymegathism).

Cell turnover and wound healing in the cornea

The epithelium. The epithelium is constantly being regenerated by mitotic activity in the basal layer of cells. However, after epithelial abrasion, the initial response of the basal epithelium at the edge of the defect is to migrate as a flattened sheet of single-cell thickness across the stroma. These cells adhere by forming hemidesmosomes and intercellular contacts and more superficially placed cells tumble over the adherent cells into the available space, leading to rapid filling of the epithelial defect (see Video 4-1).

Additional content available at https://expertconsult.inkling.com/.

Further proliferation of the basal cells, combined with reorientation of nerve endings into the wound edge, leads to differentiation of the newly divided cells into the five- to seven-cell layered mature epithelium. Closure of corneal epithelial defects occurs in a predictable manner which can be visualized clinically using fluorescein staining of the cornea; here, the initial wound edges, in geometric patterns are filled in and meet in the centre of the cornea.

Migration of epithelial cells is achieved by cytoskeletal and cell-shape changes involving redistribution of actin–myosin fibrils and changes in actin-binding proteins (e.g. fodrin cell adhesion molecules such as E-cadherin), under genetic regulation via growth factors. Migration of the cells also depends on outside-in intracellular signalling via matrix components such as fibronectin/fibrin, laminin and collagen peptides through cell surface integrins (see pp. 178–183). The role of fibronectin/fibrin in corneal epithelial resurfacing may be to facilitate healing where the normal basement membrane and, in particular, its laminin component have been lost but are not essential for wound healing.

Adhesion of epithelium to the basement membrane and Bowman's layer is normally mediated via hemidesmosomes, the lamina densa and the anchoring type VII collagen fibrils (see Fig. 4-20). However, while hemidesmosomes form during the early stages of re-epithelialization (18 hours), many days elapse before anchoring fibrils reappear, and many months pass before full ultrastructural integrity is restored. This may explain in part the phenomenon of recurrent erosion where there has been damage to Bowman's/superficial stromal layers of the cornea. Proteolytic activity in repairing epithelial defects is also important – both urokinase-type plasminogen activator and matrix metalloproteinases have been implicated.

Limbal corneal stem cells. The above process describes corneal epithelial repair after abrasion. However, the epithelium is constantly being renewed (in the mouse the entire corneal epithelium is renewed in 7 days) under the control of stem cells located in

This has been visualized directly in cultures of corneal epithelium (see Video 4-1) in which superficial epithelial cells appear to rush to fill the defect ahead of the basal more slowly centrally migrating epithelial cells. In addition, the differentiation of the corneal epithelium into a multilayered structure after wounding requires neural input from the corneal nerve endings which orient themselves perpendicular to the wound edge after the initial single layer of cells has closed the defect and are probably required to promote full differentiation of the six to ten layers of cells (eFig. 4-14). Thus, in conditions of neuropathy or when the nerves fail to extend their axons, as after herpes infections of the cornea, full epithelial healing frequently fails and recurrent breakdown of the defect occurs.

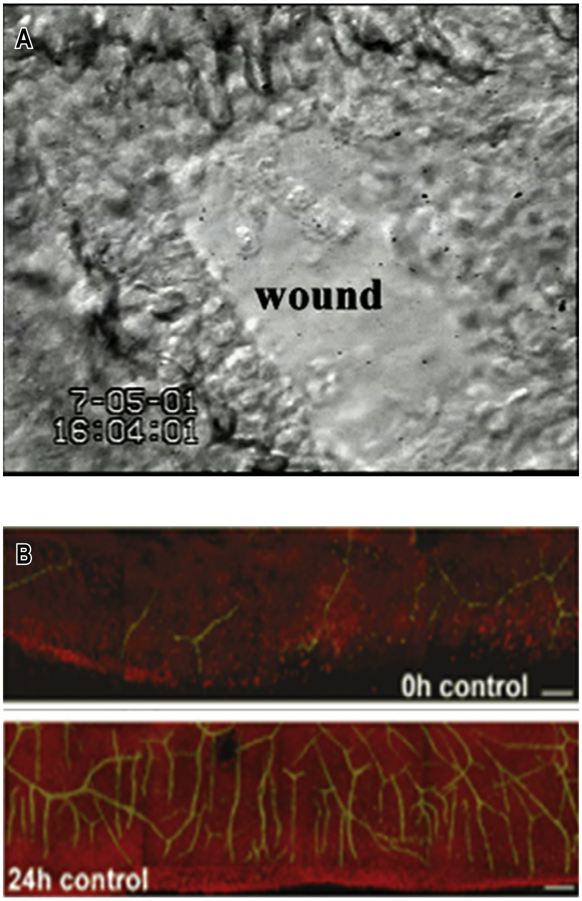

eFIGURE 4-14 (**A**) Time-lapse video sequence of corneal epithelial wound healing showing 'tumbling' of superficial cells to fill the defect; (**B**) orientation of regenerating nerve endings towards the wound edge (dark area in lower section) as the wound closes. (From Song et al., 2004.)

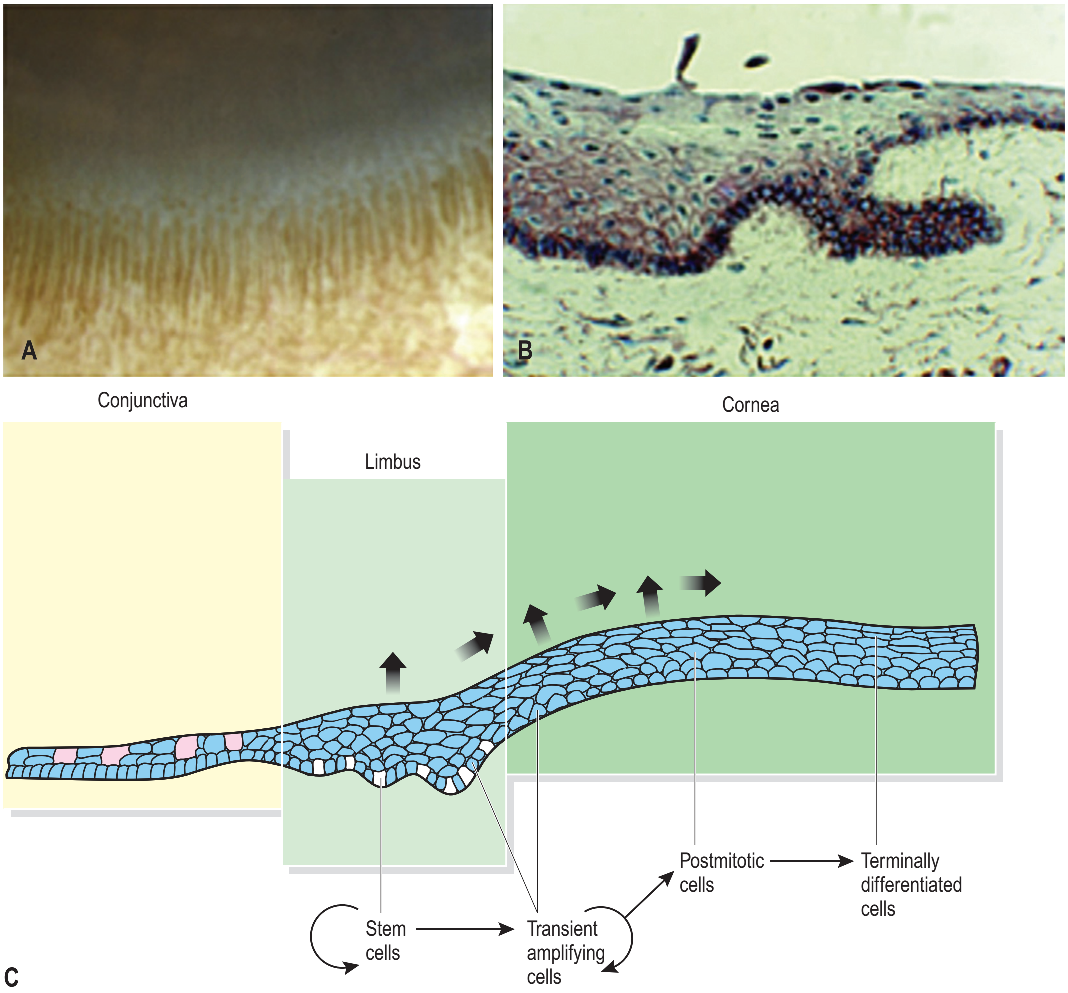

FIGURE 4-39 Concept of limbal location of corneal stem cells and transient amplifying cells. Stem cells (white) are exclusively located in the basal limbal epithelium at the bottom of the epithelial papillae forming the palisades of Vogt. Transient amplifying cells occur in the basal epithelia of limbus and peripheral cornea. Postmitotic and terminally differentiated cells make up the suprabasal and superficial layers. (From Schlötzer-Schrehardt and Kruse, 2005, with permission from Elsevier.)

an epithelial 'stem cell niche' at the limbus in the palisades of Vogt (Fig. 4-39). While some renewal of epithelial cells takes place from the basal layer to the surface throughout the cornea, the main source of cell renewal is limbal stem cells. Patients who have severe damage to the limbal region of the cornea, as for instance with chemical burns, suffer from stem cell deficiency and are prone to conjunctivilization of their cornea in which the conjunctival epithelium migrates onto the cornea to fill the defect, bringing with it stromal tissue and blood vessels. The outcome is corneal opacification and blindness. Attempts to promote corneal epithelial wound healing using growth-promoting agents such as epidermal growth factor and retinoic acid have not met with great success.

The stroma. Incisional wounds of the cornea that involve the stroma may be accidental or intentional. The immediate effect is to cause wound gape and imbibition of water from the tears by corneal GAGs (see p. 186). This causes localized opacification (light scatter) and initiates a series of events in the cornea directed at closing the wound. These include deposition of fibrin within the wound, rapid epithelialization of the wound incision, and activation of the keratocytes to divide and synthesize collagen and GAGs. During the early phase of corneal wound healing, there is loss of specialization in the keratocytes such that they revert to a fibroblast-like function and lay down collagen and GAGs found in any typical wound: e.g. hyaluronan acid, type I and type III collagen, and matrix glycoproteins. In addition, the size and arrangement of the fibrils are not regular, further contributing to the corneal opacity. In extensive wounds, this opacification remains permanently; however, in smaller, well-defined wounds there is an attempt by the cornea to restore clarity by producing normal corneal matrix components.

Surgical wounds to the cornea aim to minimize the risk not only of causing wound-related corneal opacities but also of inducing shape change to the cornea and thus astigmatism. Accordingly, cataract surgeons are using increasingly small wounds to remove the lens, the most recent innovation being optical coherence tomography (OCT)-guided femtosecond laser surgery, in this case where the laser is used to perform the incision in the lens capsule rather than the cornea. Since the corneal curvature is a function of tension in its circumferential fibres (see Fig. 4-36), restoration of the normal curvature will not be achieved unless the edges of the wound are apposed by surgical reconstruction. This is the basis of refractive surgery where, initially, partial thickness radial keratotomy was perfomed by making 'relaxing' incisions in the peripheral cornea to release the circumferential tension and thus 'flatten' the corneal curvature. The wounds are intentionally left to heal in a gaping configuration. Various types of laser surgery, such as argon-F1 (excimer = excited dimer), ultraviolet laser energy and more recently femtosecond laser, are used to produce precise customized 'ablation' of the anterior stroma performed directly on the exposed tissue after lifting a central, hinged flap of stroma. Ablation is thought to be caused by photon–photon interactions derived from thermal reactions or directly by photoablation, whereby covalent cross-links in the collagen fibrils are disrupted. This is termed laser *in situ* keratomileusis (LASIK) in which the surface of the cornea is reshaped (usually flattened) by raising a corneal flap, laser treating the exposed stromal bed and restoring the corneal flap without sutures. Both these and conventional surgical corneal incisions are fully epithelialized in the normal manner, with epithelial migration into the depths of the wound sometimes producing excessive layers of cells.

Refractive surgery continues to expand both in quantity and in novelty of technique, combined with greater precision using methods such as eye-tracking to control the incision. Wavefront-guided surgery combined with anterior segment OCT imaging are further innovations in this ever-expanding field. Wavefront technology is a term used to describe the fundamental quality of an optical system, for instance in the field of astronomy, in terms of wavefronts of higher and lower order, and the eye lends itself well to such analysis. For instance, images formed by the healthy eye may be blurred for three main reasons: light scatter, diffraction and optical aberrations, e.g. chromatic or spherical. Using wavefront technology, optical aberrations from monochromatic sources such as myopia are selectively addressed and treated with wavefront-based LASIK.

Refractive surgery, including photorefractive keratometry (PRK), LASIK and full-thickness penetrating corneal graft, has complications including sometimes severe residual astigmatism and even progressive ectasia (widening) of the initial wound. Alternative approaches have been tried to restore normal corneal curvature; these include topography-guided laser surgery using keratometry, sometimes including restoration of collagen cross-linking induced by riboflavin and UV light. Collagen cross-linking helps to 'stiffen' the stromal bed (Fig. 4-40).

The endothelium. The corneal endothelium does not normally undergo mitosis in humans even after direct injury as in a perforating corneal wound. With age there is a decline in the number of endothelial cells with an increase in their size and variable morphology (polymegathism). The response to direct wounding is to undergo enlargement and 'cell slide', as occurs in the epithelium in the early stages of migration. If sufficient numbers of endothelial cells are lost, the cell layer cannot perform its pumping action and the cornea imbibes water (decompensates) and becomes opaque.

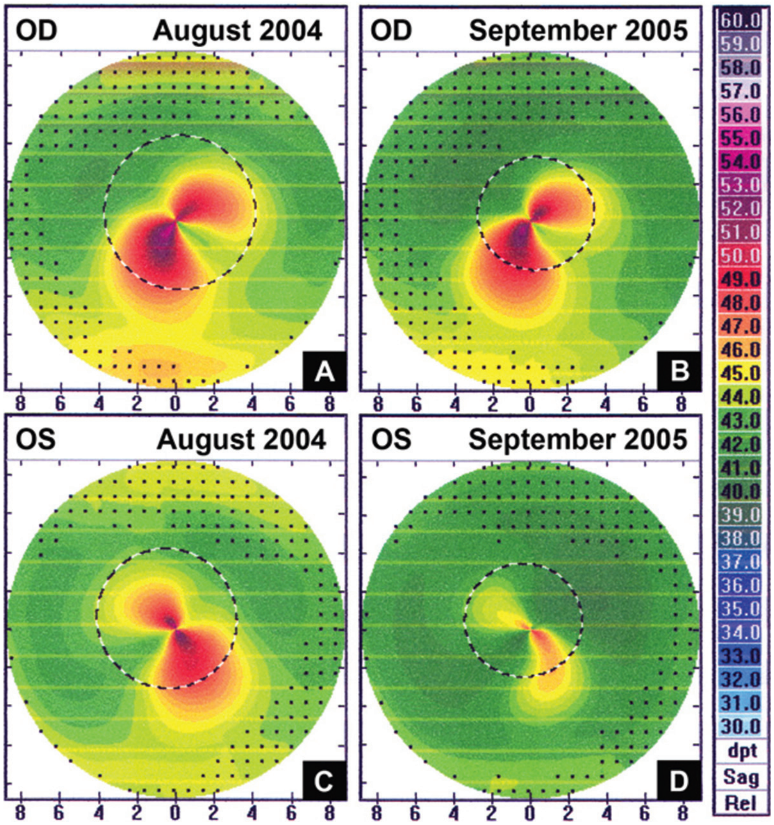

FIGURE 4-40 Corneal topographic maps showing the stabilizing effect of corneal cross-linking on corneal curvature (compare lower right and left panels). (From Hafezi et al., 2007.)

Vascularization. The normal cornea is avascular, although a few vessels may be found at the limbus. Blood vessels from the conjunctiva or the deep episcleral plexus may invade the periphery of the cornea beyond the limbus during healing of wounds or infected ulcers. When corneal epithelial or stromal defects fail to close promptly, often as a result of infection or during the severe inflammatory response of chemical injury, the continued release of proteolytic enzymes from inflammatory cells as well as damaged corneal cells, causes degradation of the stroma and increases the risk of spontaneous perforation. Matrix metalloproteinases such as matrilysin and stromelysin and MMP-9 as well as plasminogen activators (uPA and tPA) and pro-inflammatory cytokines, such as interleukin 1 (IL-1), IL-6 and IL-8, tumour necrosis factor-α and TGF-β, macrophage inflammatory proteins (MIP) 1α and β, and granulocyte–macrophage colony-stimulating factor, stimulate further ingress of inflammatory cells and initiate a vascularization response (see Ch. 7). Growth factors such as VEGF, FGF and HGF (hepatocyte growth factor) are also released, of which VEGF is considered most active. Vessels advance across the cornea to the site of injury or infection and contribute to the eventual opaque 'leucoma' of the healed cornea.

Inflammation also induces new lymphatic vessels from the limbal regions, from lymph vessel precursors in the conjunctiva, especially after herpes simplex virus infections of the cornea, which participate in the overall immune response. Lymphatic vessels are induced by the isoform VEGF-C. Inhibitors of angiogenesis are co-released during the process of inflammation, and include endostatin, a degradation product of collagen XVIII, and other products of collagen, fibronectin and even KI5, which is a fragment of plasminogen itself. In addition, thrombospondins (1 and 2) in synergy with a scavenger receptor CD36 are important anti-angiogenic factors present in the normal corneal stroma. The lack of corneal vascularity is determined during development and appears to be under the control of the transcription factor FoxC1.

Vitamin A and the cornea. Deficiency of vitamin A leads to impaired corneal and conjunctival epithelial function, with loss of corneal lustre, Bitot's spots, punctate erosions and xerophthalmia, partly as a result of loss of goblet cells in the conjunctiva. Vitamin A or retinol is required for control of epithelial keratin expression and the synthesis of cell surface glycoproteins involved in the glycocalyx. Deficiency of vitamin A leads to a form of keratinization of the corneal epithelium.

Vitamin A is also essential for normal corneal wound healing. A simple abrasion or ulcer that would be dealt with rapidly by a healthy cornea is likely to be complicated by a stromal melting response (keratomalacia) in vitamin A-deficient humans and animals. Experimentally it has been shown that topical retinoic acid can reverse the effects of vitamin A deficiency. In addition, retinol (Vitamin A) itself promotes the synthesis of the α-1 proteinase inhibitor, which inhibits a wide range of proteolytic enzymes. Vitamin A has also been shown to have protective antioxidant effects on corneal endothelial cells in culture while retinoic acid is important in promotion of T regulatory cells for control of immune, especially autoimmune, responses.

Vitamin A can be converted from traditional β-carotenoids taken in developing countries but is less efficiently converted to vitamin A than from carotenoids in other material such as lycopene, lutein and zeaxanthin.

THE SCLERA

The sclera is non-transparent and tough because of its acellularity and matrix components.

Matrix factors. The sclera is essentially acellular, containing only a few fibroblasts and non-branching traversing vessels. Recent studies have shown that there are some contractile fibroblasts (myofibroblasts) in the sclera and choroid that may have a role in refractive properties of the eye. It is opaque for the opposite reasons that the cornea is transparent, i.e. that the type I and III collagen fibres are of variable diameter and their distribution is irregular. There are also several other minor collagens in sclera (types V, VI, VIII, XII, XIII). The proteoglycans are predominantly proteodermatan and proteochondroitin sulphate of the SLR type (see p. 186) and they are localized to the collagen fibres in a similar manner to corneal proteoglycans. However, there are no proteokeratan sulphates. Other proteoglycans present in sclera include aggrecan, PRELP (proline-arginine-rich and leucine-rich repeat), decorin and biglycan among others. In addition, the amount of proteoglycan in the sclera is considerably less than that in the cornea, with the effect that it is much less hydrated (70%). The sclera, unlike the cornea, also contains elastic fibres around a fibrillin core, accounting for about 2% of total fibril content in the adult.

The sclera also contains a certain amount of large aggregating proteoglycans, such as versican, neurocan and brevican, combined with hyaluronan. There is a considerable turnover of extracellular matrix constituents in the sclera, and this may determine the shape and size of the eye and thus refraction itself.

Increases in the anteroposterior dimension of the globe occur with age, and the adult globe size is reached around age 7–10 years. Continuing increases in the size of the globe underlie the development of myopia (short-sightedness) and considerable research has been devoted to understanding this phenomenon. Recently it has been suggested that the choroid may serve as a paracrine tissue, regulating collagen and proteoglycan synthesis, particularly hyaluronic acid, in the sclera and consequently its overall dimensions.

Bulk fluid transport and the uveal effusion syndrome. Although most of the bulk transport of fluid out of the eye takes place through the anterior chamber drainage angle and/or the uveoscleral meshwork (see Ch. 1, p. 21), there is appreciable transretinal transport of fluid towards the choroid. Some of this is drained via the normal choroidal vessels (10% of the ocular fluid is drained via the vortex veins) but a proportion is drained directly trans-sclerally. The effect of this trans-scleral flow is to 'suction on' the retina to its adjoining RPE layer and maintain retinal apposition.

Fluid flowing across the sclera is absorbed by the matrix proteoglycans. Thus, the sclera is maintained in its normal state by having proteoglycans with a low water-binding capacity. In some conditions, such as the rare uveal effusion syndrome, and in nanophthalmia, the sclera contains high levels of abnormal proteoglycans, especially dermatan-sulphate-containing proteoglycans, which bind and trap large volumes of water. Thus the sclera thickens and may secondarily obstruct the choroidal venous drainage, causing further swelling and water retention.

Uveal tract

The uveal tract, comprising the iris, the ciliary body and the choroid, is a continuous layer of which the major functions are to regulate the pupil size for optimal visual function and to act as the lymphovascular tissue of the eye. Each component, however, has several other functions.

THE IRIS

Physiology

The iris is derived from neuroectodermal and mesodermal tissue and is designed to function as the lens aperture of the eye. This is achieved by the opposing actions of the sphincter pupillae and the dilator pupillae muscles (see Ch. 1, p. 26). The sphincter is an annular band of true smooth muscle that inserts close to the pupil margin at the pigment epithelium. The dilator is a highly unique series of myoepithelial cells representing the continuation of the outer layer of ciliary body pigmented epithelial cells. Some consider these cells rather to be myofibroblasts which do not have full complement of muscle proteins including desmin. The non-pigmented ciliary body epithelial cells are continuous with the posterior pigmented iris epithelium (see Ch. 1). Between the origin at the ciliary body and the insertion near the sphincter pupillae, the dilator has several side insertions into the stroma of the iris which allow it to mobilize the iris during dilation.

The functions of the pupil are:

- to regulate the amount of light entering the eye (it increases 16-fold on dilation of the pupil from 2 to 8 mm)
- to increase the depth of focus for near vision
- to minimize optical aberrations.

These functions are mediated by light and near reflexes, whose neural pathways involve autonomic parasympathetic (constriction) and the sympathetic (dilation) mechanisms (see Ch. 6). Unusually large pupil diameters have been linked to the development of myopia through increased optical aberrations.

The neuromuscular junctions of the iris are susceptible to direct pharmacological manipulation by agents that induce miosis (cholinergic agents, sympathetic antagonists) and mydriasis (anticholinergic agents, sympathomimetic agents) (see Ch. 6). In addition, pharmacological agents that induce the release of neurotransmitters from synaptic terminals (e.g. substance P is released after nitrogen mustard exposure) can have a marked effect on pupil responses. There is some evidence that dual innervation exists for iris muscles with excitatory and inhibitory input to each; thus the action of any individual drug may not be entirely predictable, depending on the state of activity at the time of administration of the drug.

Sympathetic activity in the dilator muscles appears to be mediated mainly by α-adrenergic receptors because the effects can be blocked by phenoxybenzene, but some β activity also exists. In some species, such as the cat, the action of sympathetic agents on the sphincter muscle appears to be mediated by β receptors only; in the monkey the action appears to be strongly α-mediated in the sphincter and dilator, thus producing antagonism. Studies on isolated human iris dilator indicate that the excitatory innervation is α-adrenergic while the inhibitor is cholinergically mediated.

Pigmentation of the iris contributes to iris colour, as does backscatter of short-wavelength light reflected into the iris stroma; thus, iris thickness is a significant determinant of iris colour.

Blood flow in the iris

About 5% of the total ocular blood flows through the iris. Iris blood vessels, derived from the major vessel circle, are contained as radial coils within tube-like formations of the stromal tissue, an arrangement that allows them to remain patent when the iris is fully dilated. Iris vessels have tight junctions and lack fenestrations; this renders them relatively impermeable to large molecules, as demonstrated by anterior segment fluorescein angiography. They constitute a second component of the blood–aqueous barrier (see below). Imaging techniques using high-resolution ultrasound allow visualization of iris and ciliary body vessels and indicate that flow velocities as low as 0.6 mm/second can be detected (Fig. 4-41).

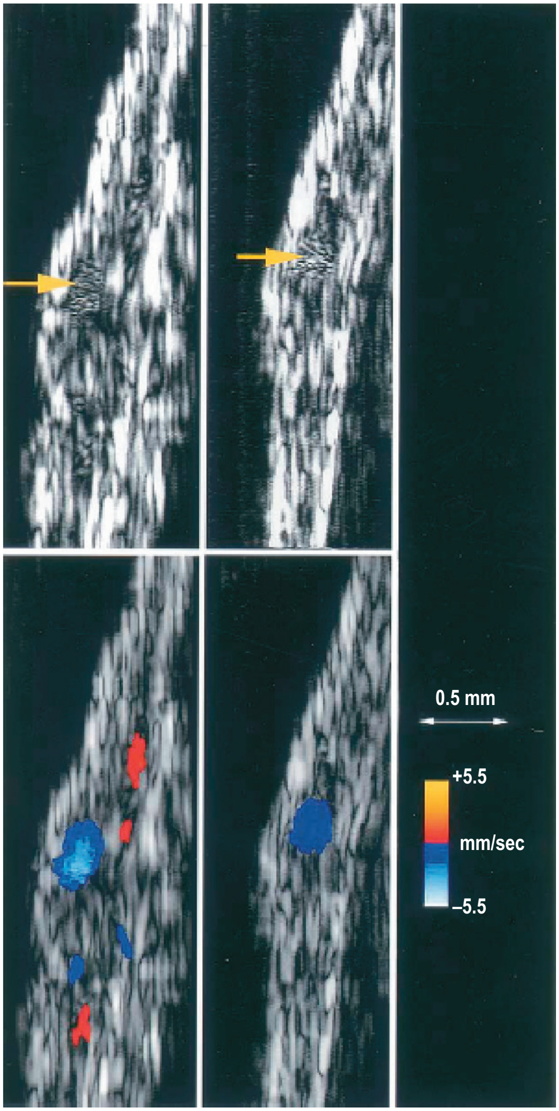

FIGURE 4-41 Top panels: swept-mode images of iris and ciliary processes at 37°C (left) and 4°C (right). Arrows indicate position of major arterial circle. Bottom panels: colour flow images derived from swept-mode data shown above. (From Silverman et al., 2002, with permission from the World Federation for Ultrasound in Medicine and Biology.)

THE CILIARY BODY

Functions of the ciliary body

The ciliary body has multiple functions:

- It provides the blood and nerve supply to the anterior segment.
- It maintains intraocular pressure by secretion of aqueous.
- It constitutes the major portion of the blood–aqueous barrier.
- Its musculature underlies the process of accommodation.

Blood flow in the ciliary body

The ciliary body receives blood vessels from the long posterior ciliary arteries and the major iris circle (see Ch. 1, p. 28). Blood flow through the ciliary body is about 7% of total ocular flow. The vessels are highly fenestrated, leaking most of their plasma components into the stroma. Blood flow in the iris and ciliary body is autoregulated like that of the retina (i.e. it does not alter significantly with changes in perfusion pressure), but it is also under autonomic control and can be modified by a variety of adrenergic and muscarinic inputs. Under normal circumstances, aqueous production is independent of ciliary body blood flow until the latter declines to <75% of normal.

Ciliary muscle and accommodation

In humans, relaxation of the zonule is induced by contraction of the ciliary muscle, which moves forward, thereby allowing the lens to adopt a more spherical shape owing to the elasticity of the lens capsule (Fig. 4-42). Accommodative range in emmetropic, non-presbyopic (see next paragraph) individuals is 5–6 dioptres. There has also been some evidence that contraction of the ciliary muscle steepens the corneal curvature, thus increasing its refractive power, as occurs in lower vertebrates. The parasympathetic neurones mediating this response are carried in the III cranial nerve via the long ciliary nerves (see Ch. 1, p. 30). It is not clear which of the three sets of ciliary muscle fibres is responsible for the major action in inducing this forward movement but, as for the iris muscles, the effects can be blocked by anticholinergic drugs. There is also a small sympathetic inhibitory component and this is increased in late-onset myopes. Accommodation produces greater ability to focus on

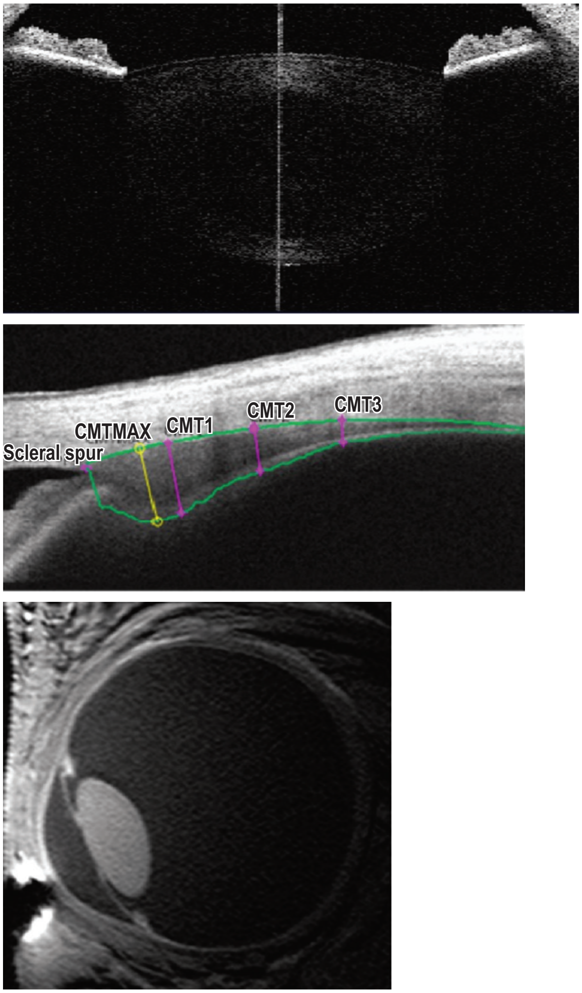

FIGURE 4-42 OCT images of the crystalline lens (top) and ciliary muscle (middle), and MR imaging of eye (bottom). Ciliary muscle image analysis shows cross-sectional ciliary muscle thickness (CMT) at 1, 2, and 3 mm posterior to the scleral spur as well as maximum thickness. (From Richdale et al., 2013.)

near objects owing to the increased refractive (dioptric) power of the lens. Epidemiological and theoretical quantitative analyses have shown that intense near work for prolonged periods disrupts emmetropization associated with eye growth and induces myopia. In addition, young children sleeping in dimly lit, as opposed to completely dark, rooms are at risk of developing myopia because of the persistence of poorly focused images through thin eyelid skin.

With age, accommodative ability declines due to changes in the deformability of the lens, while the ciliary muscle remains fully contractile. Reduced lens deformability is due to increases in lens size as well as changes in crystallin protein, combining to produce a ‘hardened’ lens (beginning stages of cataract formation). It has been estimated that the lens equatorial diameter decreases by 0.055 mm for every dioptre decline in accommodative ability.

Blood–aqueous barrier

Aqueous humour, secreted by non-pigmented ciliary epithelial cells, is derived from plasma but contains different concentrations of electrolytes and other small molecules and a restricted set of proteins in low concentration. These differences have led to the concept that a barrier exists between the plasma transudate in the ciliary body stroma and the aqueous in the posterior chamber of the eye which prevents the free diffusion of molecules over a certain size from gaining access to the posterior and anterior chambers which freely communicate via the pupil. The barrier is erected by the tight junctions between non-pigmented ciliary epithelial cells (see Ch. 1, p. 30). In contrast, extensive gap junctions between pigmented and non-pigmented cells allow the two layers of the ciliary epithelium to act as a metabolic and transport syncytium. However, in the iris, where tight junctions between the epithelial cells do not exist, it has been assumed that the barrier is formed by tight junctions between the vascular endothelial cells. Recently, this concept has been challenged. The differences in protein composition between aqueous and plasma are not absolute and thus the barrier is leaky, since some high molecular weight proteins are present in aqueous. Instead there is evidence that the protein transudate in the ciliary body stroma diffuses anteriorly into the iris stroma and thence into the anterior chamber down a concentration gradient (Fig. 4-43). Thus, aqueous in the anterior chamber always contains a small amount of protein, while aqueous in the posterior chamber is protein-free (Fig. 4-44).

Breakdown of the blood–aqueous barrier occurs in many conditions, including inflammation and vascular disease. In these circumstances the ciliary body/iris vessels become highly permeable and the stromal transudate increases with a great excess of proteins diffusing into the aqueous. The aqueous humour becomes visibly cloudy (seen as ‘flare’ by slit lamp

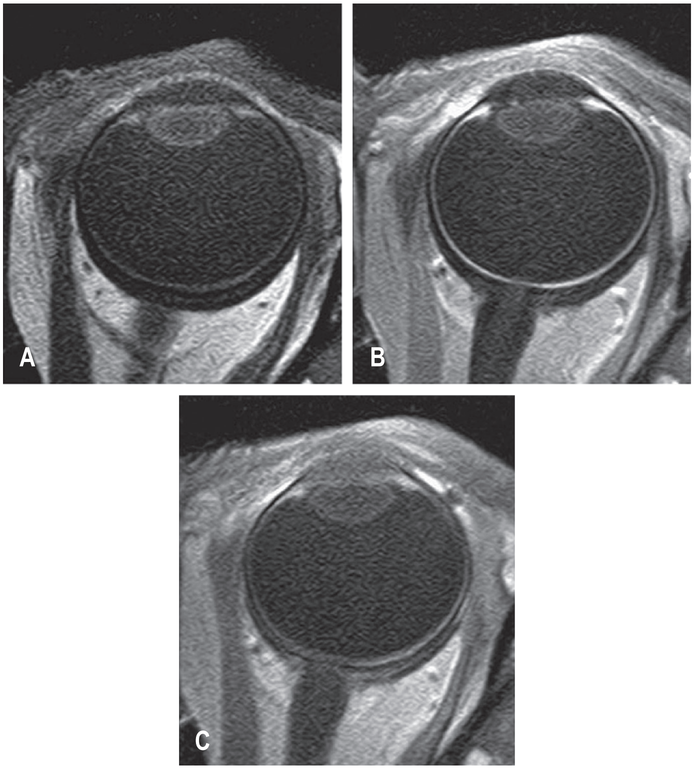

FIGURE 4-43 Contrast-enhanced MRI images of the eye. (**A**) Pre-contrast image shows details of tissues in the anterior segment of the eye, including iris, ciliary body and both the anterior and posterior chambers. (**B**) Within 2 minutes of contrast infusion, there is clear enhancement of the ciliary body and the choroid, but the anterior chamber, posterior chamber and vitreous body show no enhancement. (**C**) After 90 minutes, the enhancement in the ciliary body and choroid has begun to diminish. There is clear enhancement in the anterior chamber but the posterior chamber and the vitreous remain unchanged from **A** and **B**. (From Freddo, 2013.)

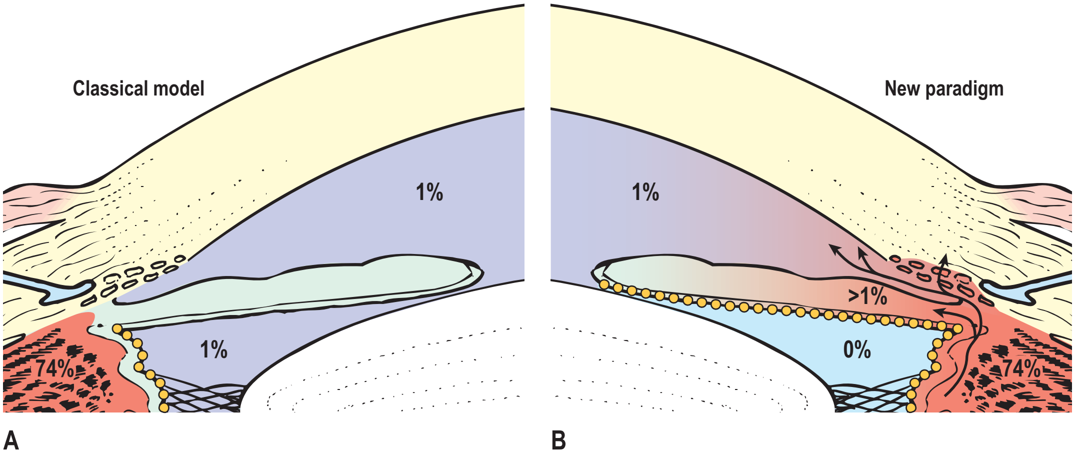

FIGURE 4-44 (**A**) In the classical model, the tight junctions of the non-pigmented ciliary epithelium and of the iris vascular endothelium are considered to be the main components of the blood–aqueous barrier, by preventing the passage of molecules greater than10 kDa into the anterior and posterior chambers. The iris stroma is presumed to be free of plasma proteins and the concentration of plasma proteins in the aqueous is uniform throughout the anterior and posterior chambers. Increases in aqueous humour plasma protein concentrations are considered the result only of an increase in barrier permeability. (**B**) In the new concept (paradigm) of blood–aqueous barrier physiology, the small amount of plasma-derived protein present in aqueous humour diffuses from the ciliary body stroma to the root of the iris, accumulates in the iris stroma and is then released into the aqueous humour of the anterior chamber only (arrows). Thus, the posterior chamber is free of protein. Some of the protein delivered to the iris root immediately enters the trabecular outflow pathways (uveo-scleral outflow, arrows). The tight junctions of the non-pigmented ciliary epithelium and of the iris vasculature endothelium still provide the main barrier function. However, an additional key element becomes the tight junctions of the posterior iris epithelium, which prevent the protein in the iris stroma from diffusing posteriorly, when combined with the one-way valve created by the pupil resting on the anterior lens capsule, and the continuous forward flow of aqueous humour through the pupil. (From Freddo, 2013.)

biomicroscopy) because of plasma proteins in anterior chambers – it may even become 'plasmoid' owing to the presence of fibrinogen and other proteins. Inflammatory cells are also likely to be present when the blood–aqueous barrier breaks down. If clotting occurs, as in severe uveitis, the aqueous becomes 'plastic'.

Eicosanoids in the iris/ciliary body

Prostaglandins were first discovered in the eye in 1957 by Ambache, who demonstrated the biological activity in aqueous and named the factor 'irin'. Eicosanoids is the generic term to describe prostaglandins (PGs) and leukotrienes, both of which are metabolites of arachidonic acid (see p. 197). Prostaglandins are synthesized in large amounts after trauma or inflammation involving the iris/ciliary body, from arachidonic acid released from esterified sites in membrane phospholipids. Other neuropeptides are involved in this response. For instance, release of substance P from the iris leads to receptor-mediated breakdown of PIP_2 (see Ch. 6, p. 362) and the formation of large amounts of arachidonic acid in the iris sphincter and synthesis of PGE_2. In the ciliary body, the cyclo-oxygenase pathway is also active in microsomes. PGE_2 is involved in miosis, while $PGF_{2\alpha}$ is involved in the control of intraocular pressure. Interestingly, breakdown of the blood–aqueous barrier in response to PGE_2 agonists is impaired in PGE-receptor knockout mice. $PGF_{2\alpha}$ is also known to increase vasodilatation and capillary permeabilization in the anterior segment of the eye. Many other peptides are present in the iris/ciliary body and the aqueous, including neuropeptide Y, vasoactive intestinal peptide, somatostatin and calcitonin gene-related peptide (CGRP). Nitric oxide is also released during activation of iris/ciliary body tissues. Many of these mediators modulate normal iris/ciliary body functions such as miosis and aqueous humour production. For instance, CGRP relaxes iris dilator smooth muscle via cAMP mechanisms. They also have other functions such as the immunosuppressive role of vasoactive intestinal peptide in ocular immune privilege (see Ch. 7, p. 457).

Drugs that inhibit the cyclo-oxygenase pathway (see p. 197), such as indomethacin and aspirin, may be useful in ocular inflammation. However, steroids act at the level of phospholipase A_2 and may have a more global effect on the response (see Ch. 6, p. 363). In addition, the lipoxygenase pathway is active in the anterior uvea with synthesis of leukotrienes B_4, C_4 and D_4, and the chemotaxis of polymorphonuclear leucocytes (see Ch. 7). Penetration of drugs into the intraocular environment after topical application occurs more readily through the conjunctiva and the sclera than the cornea, partly as a result of the numerous transport mechanisms available in the conjunctival epithelium but also because of the greater transepithelial permeability of the conjunctiva compared to cornea, which is impermeable even to low molecular weight compounds (<1000 kDa).

Detoxification and antioxidation in the anterior segment

The cytochrome P_{450} system is the major drug detoxification (CYP) system in the eye. Microsomes contain a group of proteins known as the cytochrome P_{450} proteins, which catalyse the transfer of a single oxygen atom to endogenous and exogenous substances destined for excretion and/or detoxification, such as steroids, phenobarbital, etc. (see Ch. 6, p. 344). Their main effect is to convert hydrophobic compounds to hydroxylated hydrophilic compounds, which are then more easily metabolized.

The cytochrome P_{450} system is present in the ciliary body (at about 5% of the concentration in liver) where it acts to detoxify many compounds. It does this by either converting the hydroxylated, highly reactive compound to a glucuronide via UDP-glucuronyl transferase or by conjugating it to glutathione via glutathione *S*-transferase (Box 4-13). Several CYP family enzymes have been identified and/or purified from the ciliary body, in particular the non-pigmented epithelium. For instance the enzyme CYP2D6 is important for metabolizing the topical anti-glaucoma β-adrenergic blocker timolol. There is considerable genetic variation in the induction of the cytochrome P_{450} system in the eye, perhaps explaining the variable toxic effects of drugs in individuals.

The ciliary body is the main source of antioxidant systems in the anterior segment. Although antioxidant systems exist in the lens (see below) and the cornea (see above), the ciliary body is especially rich in antioxidant systems with the highest concentrations of catalase, superoxide dismutase, and glutathione peroxidase types I and II. Type I is selenium-dependent

BOX 4-13 CYTOCHROME P_{450} AND DRUG DETOXIFICATION

The cytochrome P_{450} system detoxifies compounds by utilizing the glutathione *S*-transferase system and degrading compounds to mercapturic acid.

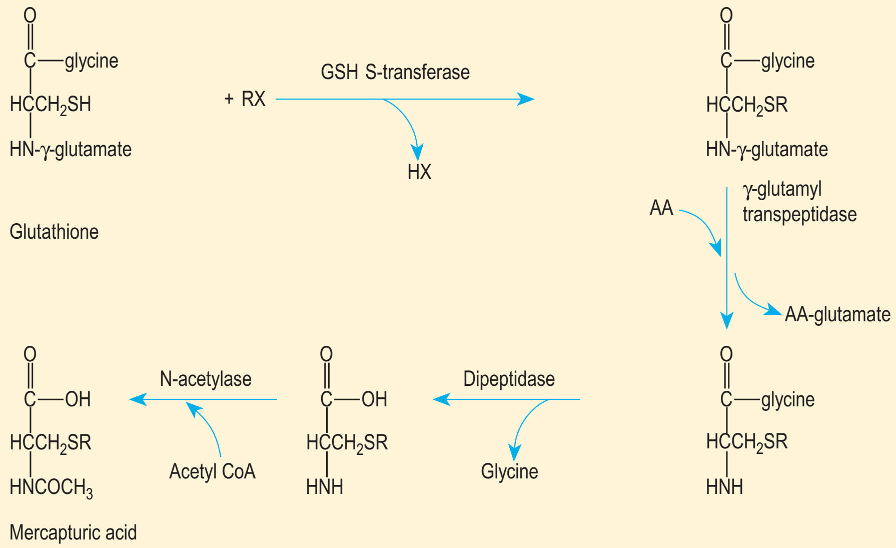
O
C—glycine
HCCH₂SH
HN-γ-glutamate
Glutathione
+ RX
GSH S-transferase
HX
O
C—glycine
HCCH₂SR
HN-γ-glutamate
AA
γ-glutamyl transpeptidase
AA-glutamate
O
C—glycine
HCCH₂SR
HNH
Dipeptidase
Glycine
O
C—OH
HCCH₂SR
HNH
N-acetylase
Acetyl CoA
O
C—OH
HCCH₂SR
HNCOCH₃
Mercapturic acid

while type II is selenium-independent. Type I is closely linked to glutathione reductase whose main function is the reduction of oxidized glutathione (GSSH) produced by the detoxification of peroxides (see Box 4-13).

Hydrogen peroxide (H_2O_2) is present in normal aqueous, most of it derived from the non-enzymatic interaction between reduced ascorbate and molecular oxygen, and it is reduced to H_2O by glutathione secreted by the ciliary epithelium. Most of these studies have been performed in experimental animals and it is not clear how relevant they are to the human eye. It has been suggested that oxidized ascorbate (via the superoxide anion) is more important in degrading H_2O_2 in humans. Melatonin, a neuropeptide involved in biological circadian rhythms, is also an H_2O_2 scavenger. A role for xanthine oxidase has also been suggested. H_2O_2 can induce noradrenaline (norepinephrine) release from the iris/ciliary body in the aqueous and has recently been implicated in cataract formation.

Ciliary body tissue also contains a peroxiredoxin, a constituent of a widely distributed family of antioxidant enzymes, whose amino acid sequence and tissue distribution are now known. Their role is to degrade H_2O_2 and alkyl peroxides.

THE CHOROID

Functions of the choroid

The function of the choroid is to act as the lymphovascular supply to the posterior segment of the eye.

Vascular function. The choroid is almost entirely composed of vessels embedded in a loose connective tissue matrix which has a high content of type III collagen, typical of an expansile or spongy tissue. The blood supply to the choroid has several interesting features:

- Some 98% of the blood to the eye passes through the uveal tract, of which 85% is through the choroid.
- Blood flow occurs at a rate of 1400 mL/min per 100 g tissue, which is higher than the perfusion of blood through the kidney.
- The choriocapillaris is organized in a lobular architecture, collecting into larger vessels and finally into four vortex veins, one in each quadrant of the globe (see Ch. 1).

- Venous blood draining from the choroid is not desaturated, only 5–10% oxygen having been extracted during passage through the eye. The choroid supplies the outer retina, where the partial pressure of oxygen ($P\mathrm{o}_2$) is highest, rapidly falling towards the retinal inner segments and then rising again, less so, towards the inner retina.
- Blood vessels in the choroid are highly fenestrated and leaky, like ciliary body vessels.
- Choroid and ciliary body blood vessels are sensitive to the $P\mathrm{o}_2$ and $P\mathrm{co}_2$; in conditions of high $P\mathrm{co}_2$ the vessels expand greatly in a forward direction, altering the forward position of the retina, and exerting pressure on structures such as the vitreous gel and lens/iris diaphragm.
- Although previously considered to have considerable autoregulation in relation to perfusion pressure, choroidal blood flow, especially around the optic nerve, is sensitive to the effects of nitric oxide and endothelin and other as-yet unidentified vasoconstrictors.

Non-vascular functions. The choroid has additional functions:

- It is involved in ocular temperature control by dissipating heat from the eye.
- It secretes growth factors which control scleral thickness and may have a role in emetropization (see eFig. 4-13).
- It drains aqueous fluid from the anterior segment of the eye via uveoscleral outflow, which may be as much as 40% of aqueous outflow in humans (see p. 211).
- Higher primates, including man, have intrinsic choroidal neurones which may have a role via nitric oxide in controlling vascular diameter and blood flow to the choroid.
- In humans and other primates, non-vascular smooth muscle cells may be found in some cases forming discrete rings around blood vessels and even in the region of the fovea.

Lymphoid function. Intraocular structures lack a recognized lymphatic system. However, through the uveoscleral outflow, drainage of intraocular fluid into conjunctival (aqueous veins) and orbital lymphatics occurs. Some reports of suprachoroidal lymphatic lacunae have been documented. The choroid contains a rich network of immune cells including mast cells, macrophages and dendritic cells which adopt this migration route to the draining lymph node, which in the human are the pre-auricular and submandibular lymph nodes of the eye (see Ch. 1). This tissue can respond massively to intraocular inflammation (see Ch. 7, p. 459). In addition, choroidal and ciliary body/iris melanocytes are potential antigenic targets for autoimmune disease (see Ch. 7).

Aqueous humour dynamics

A fundamental physiological function of the eye is to maintain an intraocular pressure (IOP) between 10 and 20 mmHg. This is achieved by the circulation of aqueous humour secreted by the ciliary body into the posterior chamber and circulated through the pupil towards the anterior chamber angle where it drains via the outflow apparatus into the episcleral veins (see Ch. 1). Factors affecting IOP include:

- circadian rhythms
- episcleral venous pressure
- rate of secretion and flow of aqueous humour
- neural (cranial nerves V and VII) and hormonal influences.

The intraocular pressure, as measured clinically, actually represents the balance between the inflow and outflow of aqueous and is altered by changes in the gradient of pressure between the posterior chamber and the anterior chamber, and eventually by the episcleral pressure. The uveoscleral outflow is greatly affected by alterations in this gradient of pressure as appears to occur in glaucoma (Fig. 4-45).

AQUEOUS HUMOUR IS SECRETED BY THE CILIARY BODY EPITHELIUM

The rate of aqueous humour formation is about 2–3 μL/min. Aqueous humour is formed by the transport of water and electrolytes from the leaky fenestrated capillaries of the ciliary processes to the epithelial syncytium and thence across the plasma membrane of the non-pigmented epithelium (see Box 4-14).

Composition of aqueous humour

The aqueous humour is composed predominantly of electrolytes and low molecular weight compounds

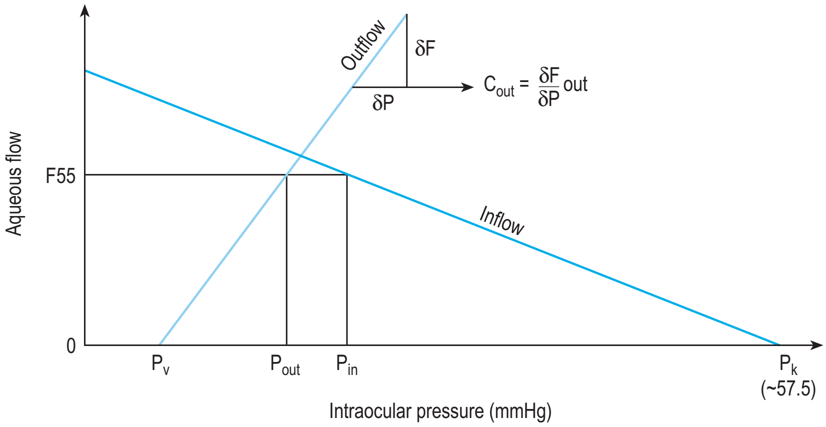

FIGURE 4-45 Aqueous inflow declines as intraocular pressure (IOP) rises towards P_k (about 57.5 mmHg). Goldmann estimates pseudofacility as $P_k = 0.58 \times$ brachial BP. The gradient for inflow is $(F_{55}/P_k - P_{55}) = C_{in} = C_{ps}$ where F_{55} is steady-static flow. C_{ps} is the pseudofacility. (Courtesy of D. Woodhouse.)

with some protein, and there are significant differences from plasma in several of the components. Aqueous receives contributions from a variety of sources, including the corneal endothelium and the iris/lens (see p. 211), in addition to active secretion from the ciliary body. Several trace compounds are also present in aqueous humour, including steroid sex hormones, enzymes such as carbonic anhydrase, lysozyme and plasminogen activator, and cytokines such as basic fibroblast growth factor (bFGF) and TGF-β (see Ch. 7, p. 457). It is likely that most of the high molecular weight components are present only in the anterior chamber. Low levels of catecholamines (adrenaline, noradrenaline and dopamine), prostaglandins and cyclic nucleotides are present in normal aqueous, but the source of these compounds is uncertain.

The protein content of aqueous is very low (about 1/500 of plasma), and the major species is albumin. Since immunoglobulins are relatively large molecules, it is unlikely that they gain access to the aqueous via the ciliary body epithelium; rather, diffusion through the ciliary body /iris stroma is more plausible; alternatively, local production via iris lymphocytes or plasma cells is a possible source since local antibody production in the eye has been well recorded. Small amounts of fibronectin are also produced locally.

Aqueous contains detectable amounts of hyaluronic acid, derived as breakdown oligomers from vitreous hyaluronic acid during the normal process of GAG renewal (see p. 187). However, some of the aqueous hyaluronan is of higher molecular weight than vitreous, suggesting that it is produced in the anterior segment. Aqueous hyaluronic acid may have a role to play in regulation of IOP because perfusion of the anterior chamber with hyaluronidase leads to a marked drop in pressure. However, it is also possible that intracameral hyaluronidase affects trabecular meshwork cells and extracellular matrix, thus leading to a lowering of the IOP.

NEURAL/AUTONOMIC CONTROL OF AQUEOUS SECRETION

Adrenergic and cholinergic agonists and receptors are present in the iris and ciliary body, and autonomic innervation of this tissue occurs in muscle, vessels and epithelial cells. Adrenergic receptors are present in the ciliary epithelium and regulate IOP via the adenylate cyclase system. β-Adrenergic antagonists and α_2-selective adrenergic agonists both suppress aqueous flow. Muscarinic receptors linked to the PIP_2 second messenger system are also present in the ciliary epithelium.

Cholinergic mechanisms are not involved in control of intraocular pressure

M_3 muscarinic receptors in the ciliary epithelium linked to phosphatidylinositol in the cell membrane have been identified, but they do not appear to have a significant role in IOP control. Despite this,

BOX 4-14 SECRETION OF AQUEOUS HUMOUR

Classic theory suggests that passive diffusion of water and ions from the fenestrated vessels of the ciliary body is followed by active transport of Na^+ and Cl^- across the ciliary body syncytium. This is an active secretory process involving Na^+/K^+ ATPase and carbonic anhydrase type II activity. In some respects this has been viewed as ultrafiltration of ions and water and eventually leads to the formation of aqueous humour, which is secreted into the posterior chamber. However, the oncotic pressure of the ciliary stroma is greater than the hydrostatic pressure difference across the ciliary epithelium, so tending towards absorption of water into the ciliary body from the posterior chamber. Thus, active transport of ions in the opposite direction is the main mechanism of aqueous humour formation. This process is also under adrenergic receptor control at the level of the ciliary epithelial cells and possibly also by regulation of blood flow to the ciliary body. The possibility has been raised that H_2O transport may also be achieved by aquaporins. The ciliary epithelium has been shown to express two aquaporins, AQP1 and AQP2, but these do not seem to play a major role in aqueous humour production. In addition, some of the K^+ channels are Ca^{2+}-sensitive and can be activated by Ca^{2+} entry via Ca^{2+} channels. More recently, the role of the Cl^- ion in aqueous humour formation has assumed greater importance through its transport via the cystic fibrosis transmembrane receptor. This may play a greater role than the HCO_3^- carbonic anhydrase-mediated mechanisms and may also be facilitated by adenosine receptors.

Aquaporin water channels are also present in both the secretory machinery (AqPO4 and AqPO1) and the outflow channel (trabecular meshwork endothelium) and appear to contribute to bulk flow of water because intraocular pressure is reduced in AqPO4/1 knockout mice.

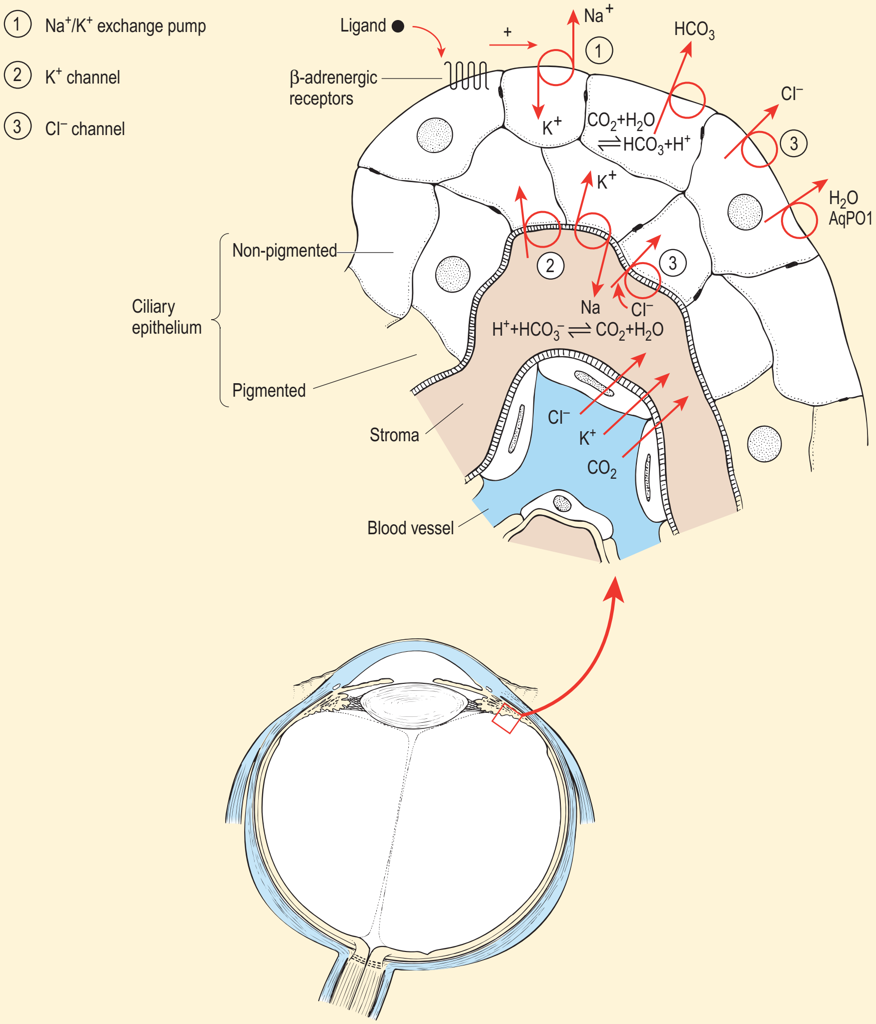
① Na^+/K^+ exchange pump
② K^+ channel
③ Cl^- channel
Ligand ●
+
Na^+
HCO_3
β-adrenergic receptors
1
K^+
$CO_2+H_2O \rightleftharpoons HCO_3+H^+$
Cl^-
3
H_2O
AqPO1
K^+
2
3
Non-pigmented
Ciliary epithelium
Pigmented
Na
Cl^-
$H^++HCO_3^- \rightleftharpoons CO_2+H_2O$
Stroma
Cl^-
K^+
CO_2
Blood vessel

cholinergic agents such as pilocarpine are thought to have some action via reduction in aqueous secretion, although the experimental evidence is weak. Most of the action of pilocarpine appears to be mediated via its effect on outflow resistance and uveoscleral flow (see below).

Adrenergic receptors regulate IOP via adenylate cyclase

The majority of α receptors in the ciliary body are α_2, while more than 90% of β receptors are β_2. Stimulation of α_2 receptors lowers the IOP via a reduction in aqueous humour production through inhibition of adenylate cyclase. Adrenaline, a preferential α-adrenergic agonist, stimulates prostaglandin synthesis, particularly that of PGE_2 and $PGF_{2\alpha}$, the latter having potent ocular hypotensive activity.

Stimulation of β receptors, particularly β_2 receptors, also leads to an increase in aqueous secretion via activation of adenylate cyclase.

The dual control of aqueous secretion through activation (β) or inhibition (α) of adenylate cyclase is mediated by their respective stimulatory and inhibitory G proteins (see Ch. 6, p. 345). Thus, IOP can be lowered by α_2 agonists (e.g. clonidine) or β_2 antagonists (β-blockers, e.g. timolol). The α_2 receptors are also linked to vasoactive intestinal peptide receptors, which are co-stimulated and lead to a reduction in cAMP levels, which in itself may also lower IOP. β antagonists have no effect on aqueous flow when aqueous production is at its lowest, whereas α_2 agonists and carbonic anhydrase inhibitors do.

The mechanism whereby changes in intracellular cAMP levels alter aqueous secretion is not known but appears to involve transport of the HCO_3^- ion across the cells. In addition, a number of other components appear to be involved in aqueous secretion, such as protein kinase C, which is linked to adenylate cyclase activation and thus may act as part of an intracellular signalling network connecting the two main second messenger systems (see Ch. 6, p. 345).

Nucleotides and nucleotide receptors

Low molecular weight nucleotides such as adenosine are present in aqueous humour in significant quantities (4^{-4} μM) and are likely to play a role in IOP control. The wide-ranging effects of nucleotides, particularly adenosine, are mediated via receptors present on nerve endings, including PTX and PTY receptors. Topical application of some nucleotides to the eye leads to lowering of the IOP.

Interestingly, melatonin has some IOP-lowering effects and the melatonin analogue 5-methoxycarbonylamino-N-acetyltryptamine (5-MCA-NAT) has similar effects which are thought to be due to induction/modulation of adrenergic receptors in the eye.

Circadian regulation of aqueous humour formation

It has long been known that there is a diurnal variation in IOP, possibly due to melatonin-based mechanisms. This is in part the result of a circadian regulation of aqueous humour secretion. Secretion in humans occurs at a rate of 2.6 mL/min during the day and falls to 1.0 mL/min at night. Both β-adrenergic receptor-mediated and neuropeptide-mediated mechanisms, particularly that of vasoactive intestinal peptide, are involved. Activation of G protein-coupled adenylate cyclase leads to cAMP production, which activates protein kinase A, thus regulating the cation channels. The process is terminated by hydrolysis of cAMP by phosphodiesterase.

Is guanylate cyclase involved in intraocular pressure control?

Large amounts of brain natriuretic peptide (BNP) as well as atrial natriuretic peptide (ANP) are found in the iris/ciliary body and in the aqueous humour of rabbits and humans. The receptor for ANP is linked to membrane-bound guanylate cyclase, and *in vitro* studies have shown that this enzyme can be stimulated by ANP in the ciliary body. In the rabbit this is accompanied by a reduction in IOP but its relevance in humans is not clear because similar levels of peptides are found in both normal subjects and in glaucoma patients.

AQUEOUS HUMOUR OUTFLOW FROM THE EYE

Flow of aqueous humour from the eye is controlled at several different levels, including at the site of the ‘conventional’ (C_t) outflow pathway via the trabecular meshwork, at the uveoscleral (C_u) outflow system, and, outside the globe, at the site of the episcleral veins (see Ch. 1).

Control of outflow at the trabecular meshwork

Resistance to the outflow of aqueous occurs at the level of the endothelium covering the trabecular meshwork, which is an extension of the corneal endothelium, but also within the matrix of the meshwork itself. The hydraulic conductivity at this site is around 10^{-7} cm^{-2} s^{-1} g^{-1} , which is several times greater than for other lining endothelia and has been attributed to a process of water transport termed 'transcytosis', in which membrane-bound water-containing vesicles are transported through micron-sized pores in the cell, across the meshwork. Formation of transcellular pores may be a pressure gradient-induced mechanoreceptor-based mechanism involving the actin–myosin cytoskeleton and appears less effective when the cells are stiffened, as occurs with agents which increase outflow resistance such as thrombin and sphingosine-1-PO_4.

Despite the potential contribution of the trabecular endothelial cell, the juxtacanalicular cribriform meshwork is considered to account for much of the resistance to aqueous flow into Schlemm's canal. This has been attributed to matrix components, especially the GAGs, present in this region. The trabecular beams, composed of type I collagen, with a significant proportion of type III and IV collagen, and other matrix constituents such as laminin, fibronectin and elastin, are separated by GAG-filled spaces, particularly hyaluronan (Fig. 4-46) that retard the flow of fluid by virtue of hydrophilic properties and large hydrodynamic volume. A wide range of GAGs have been identified in the trabecular meshwork matrix—namely, hyaluronan, chondroitin sulphate, dermatan sulphate, keratan sulphate and heparan sulphate—with significant variation between species. Some unidentified proteoglycan material has also been detected. Trace amounts of types V and VII collagen have also been detected in the trabecular meshwork.

Apart from the mechanosensor or stretch receptor function of trabecular meshwork (TM) endothelial cells discussed above, TM endothelial cells have special characteristics: namely, active phagocytic properties, high levels of cytoskeletal actin, which in cultured cells is particularly sensitive to cytochalasin B, and lower levels of microtubules, which appear to be relatively non-responsive to colchicine. These cells also contain vimentin and desmin, thus showing some similarity to smooth muscle cells. Taken together, these findings suggest that trabecular meshwork endothelial cells are specialized for both endocytic transport of water and solutes, and contractility. Actin mobilization appears to be mediated via adrenergic receptors, probably of the β_2 type, which are highly responsive to adrenaline. Energy metabolism in the trabecular meshwork endothelium is predominantly glycolytic rather than oxidative, although both enzyme systems are present and functional. Transport of water may be achieved not merely by passive transportation of H_2O packets but by activation of a water channel protein, aquaporin-1 (AQP01), found in corneal endothelial cells.

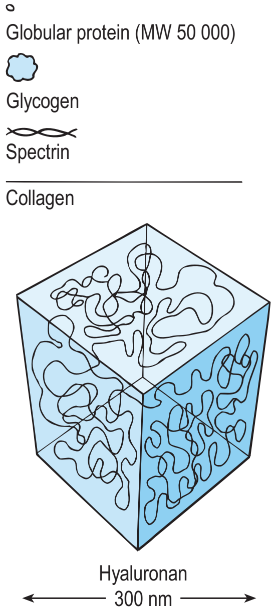

FIGURE 4-46 Model of hyaluronan, showing size compared to other 'typical' molecules. (Courtesy of J. Alberts.)

Metabolism of trabecular meshwork cells

The matrix components in the trabecular meshwork are thus synthesized and degraded by the endothelial cells. In addition, these cells have high levels of surface tissue plasminogen activator (tPA), higher even than in vascular endothelial cells, and this is likely to play a role in maintaining patency and reducing the resistance of the outflow passages. Phagocytic activity of trabecular meshwork cells is associated with several other enzymatic activities such as GAG-degrading enzymes and acid phosphatase.

Trabecular meshwork cells have receptors for a variety of agents including adrenaline (β_2 adrenergic receptors, decrease phagocytosis) and glucocorticoids. Both steroids and oxidative damage induce the expression of the trabecular meshwork inducible glucocorticoid response (TIGR) protein. Mutations in the *TIGR* gene, now known as the myocillin gene, have been found in patients with glaucoma. Myocillin is found widely in ocular tissues and may be a component of exosomes (see p. 165) which could potentially obstruct TM outflow if released in excess or not removed after release. Other susceptibility genes such as the optineurin and the WD repeat domain (*WDR36*) genes have been found but account for only a small proportion of patients with primary open angle glaucoma and the *CYP1B1* gene (drug detoxifying gene, see p. 221) in congenital glaucoma. In addition, TGF-β appears to be involved in the glucocorticoid cell response and in the production of trabecular meshwork extracellular material. Steroids inhibit prostaglandin by trabecular meshwork cells at concentrations as low as 10^{-8} mol/L. Prostaglandin synthesis by trabecular meshwork accounts for a significant proportion of its arachidonic acid metabolism (70% compared with less than 5% in other cells), suggesting that prostaglandins play a major role in trabecular meshwork cell physiology. In addition to substantial amounts of PGE and $PGF_{2\alpha}$, leukotriene B_4 appears to be produced in high amounts.

Trabecular meshwork cells contain the free radical and hydrogen peroxide detoxifying enzyme systems present in other tissues such as the ciliary body (see above). Both a catalase and a glutathione-dependent system are active in handling hydrogen peroxide, which can reach levels as high as 25 μmol/L in the aqueous humour. Ageing is associated with increasing free radical damage and mitochondrial dysfunction and clearance of waste products by the ubiquitin–proteasome system, and other lysosomal and non-lysosomal enzyme systems such as the calpains of TM cells becomes less efficient. Waste products thus accumulate in the TM itself (‘the garbage catastrophe theory of ageing’) and contribute to increased outflow resistance, ocular hypertension and glaucoma. This age-related process will also be affected by genetic susceptibility as shown in the recent identification of two mutations in the lysyl oxidase gene in patients with secondary glaucoma associated with pseudoexfoliation syndrome. Lysyl oxidase is required for elastin production (see p. 183) and presumably vitreous/zonule components are more readily degraded and washed through to the TM in this syndrome.

Uveoscleral drainage

A variable proportion of aqueous (up to 40%) drains directly into the anterior uvea at the ciliary body immediately posterior to the cornea and thence into the suprachoroidal space and towards the posterior pole of the eye (see Ch. 1). The anterior uvea at this point is incompletely lined with endothelial cells. In addition, the localization of MMP-1 suggests a role for this enzyme in uveoscleral outflow. Uveoscleral drainage is possible because the pressure in the suprachoroid is 2–4 mmHg lower than in the anterior chamber; this can be reversed after trabeculectomy and can lead to choroidal effusions. This pressure differential is also less with age, leading to greater risk of choroidal effusion in such patients. Prostaglandins may decrease the intraocular pressure by increasing the uveoscleral outflow. Several possible mechanisms have been proposed, including relaxation of the ciliary muscle, cell shape changes, cytoskeletal rearrangements or compaction of the trabecular meshwork matrix.

Episcleral circulation

It was shown many years ago that dye-stained aqueous fluid would not drain out of the eye into the episcleral veins if the IOP was less than 15 mmHg. Thus, this represents the combined episcleral venous pressure and the oncotic pressure in the perivenous tissues of the episcleral veins. Aqueous humour draining via the canal of Schlemm into the aqueous veins does so by passing through large transcellular channels and giant vacuoles on the meshwork side of the canal. The canal has direct vascular communications on its outer wall with a network of intrascleral collector channels that drain into the scleral veins (see Ch. 1). An alternative model for aqueous outflow has been proposed which involves a mechanical pumping mechanism generated in response to small changes in IOP and linked to the ocular pulse. Pumping of aqueous from Schlemm's canal into the collecting veins and episcleral veins is assisted by small valves in this model.

The episcleral venous pressure can be measured using a non-invasive manometer (the EV-310) and is normally around 8–10 mmHg.

An intraocular pressure-independent link between aqueous production and aqueous outflow?

Bestrophins are anion channels known to be involved in fluid transport in the gut. Recently, similar proteins have been identified in the ciliary body, including the HCO_3^- channel bestrophin-2 which is selectively expressed in non-pigmented epithelial cells. Unexpectedly, bestrophin-2-deficient (best2 –/–) mice have a higher rate of aqueous inflow but a reduced IOP. This appears to be due to the very high level of a soluble adenylate cyclase, in the non-pigmented epithelium, which is acutely sensitive to HCO_3^-, but is absent from the outflow pathway cells. This suggests that the ciliary body may produce substances via HCO_3^--sensitive soluble adenylate cyclase, which directly influences outflow independent of pressure (Lee, 2011).

Does aqueous contain components that contribute to flow resistance?

Although aqueous has the same viscosity as isotonic saline, its passage through microporous filters *in vitro* is slower than that of saline. This effect can be abolished by proteolytic agents and detergents, but not by hyaluronidase. It has therefore been suggested that some forms of glaucoma may be caused by a build-up of a surfactant-like material with age.

The lens

The transparency of the lens is a function of the highly ordered state of its cells and extracellular matrix. In essence, the extracellular matrix of the lens is confined to its capsule, while the cells form a syncytium with interlocking cellular processes. The syncytial arrangement is brought about by the extensive communicating gap junctions between the cells, but the cells are not fully apposed, leaving a nanometer-sized space between cells which permits an important and essential circulation within the lens.

LENS TRANSPARENCY

The light transmission properties of the ocular media vary to some degree depending on the nature and the age of the tissue. The cornea, aqueous humour and lens all transmit long-wavelength light well above the limit of visible light (about 720 nm). However, short wavelengths below 300 nm are absorbed by the cornea but wavelengths above 300nm are transmitted through the aqueous; the lens further filters the majority of the short wavelengths below 360 nm and is an absolute barrier to light below 300 nm (Fig. 4-47).

The epithelium

The single-celled epithelium of the lens and its capsule do not scatter or reflect light, essentially because the combined refractive index is the same as that of the aqueous humour (1.336). However, the epithelium is of great importance in the maintenance of fluid and electrolyte balance of the lens syncytium via ion-pump mechanisms (see below). Thus, any agent that disturbs epithelial function and/or viability (such as ionizing radiation to the lens bow region) will have significant effects on lens clarity. This applies to all aspects of lens structure and function.

The organization of the lens fibre cells underpins the transmission properties of the lens

It might be expected that the plasma membranes of the lens fibres would produce interference diffraction patterns that would affect the ability of the lens to

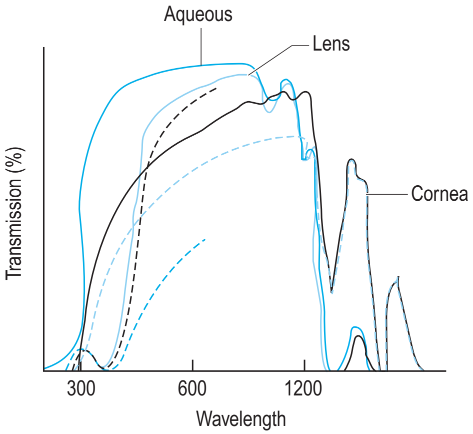

FIGURE 4-47 Optical density curves for various ocular components. (With permission of Drs Jaffe and Horwitz and JB Lippincott.)

transmit light. However, it has been shown that the weak diffraction rings that are produced occur in a repeating pattern with a period of the same dimension as the thickness of an individual lens fibre in the anteroposterior axis. This reduces any scatter by the plasma membranes of the normal epithelium; it has been estimated that the amount of scatter by the epithelium in the human lens is about 5% of the transmitted light.

Lens fibres differentiate from mitotic equatorial lens epithelial cells, which migrate posteriorly and centrally as new fibre cells are generated. This produces a zone of differentiating fibre cells which not only lose their nuclei but also most of their intracellular organelles as they become mature cells towards the lens centre (nucleus) (Fig. 4-48). This loss of cellular organelles as well as their anteroposterior orientation assist in allowing light transmission. The fibres are organized in a densely packed cellular arrangement, with interdigitations like pieces in a three-dimensional jigsaw puzzle (see Ch. 1, p. 32) in which extensive intercellular communication exists via the lens gap-junction, generated by a set of heterodimeric proteins composed of connexins (Cx43, Cx46 and Cx50) (Fig. 4-49) as well as the gap-junction-like protein, aquaporin-0 (AQP0). Lens fibres essentially become crystallin-replete (over 90% of the total cellular protein) sacks in which crystallins are embedded within a complex cytoskeletal matrix, some components of which are also lens-specific (e.g. the beaded intermediate filament protein). The polydisperse nature of the crystallins prevents spontaneous crystallization in the packed arrangement. The high refractive index (RI) of the lens is caused by the crystallins; at the periphery of the lens the RI is slightly less (1.38) than at the nucleus (1.41). The water content of the lens is also greater at the periphery (75–80%) than at the lens nucleus (68%).

As the lens fibres differentiate towards the centre (lens nucleus), this packing arrangement becomes modified and less compact (Fig. 4-50). The presence of the crystallins is in itself insufficient to explain the transparency of the lens. Transparency is predominantly the result of the packing of the crystallins in very high concentration such that they resemble a dense liquid or a glass because of the high level of 'short-range spatial order'; this means that the scatter of light from each individual molecule is related to the scatter from its immediate neighbours and that they tend to cancel each other out. At a macroscopic level, the arrangement of the crescent-like fibre cells in end-to-end concentric shells around a polar axis provides a highly ordered architecture. A series of coaxial refractive surfaces, thus created, promotes transparency of the multicellular structure. However, this general view does not apply to all species and in most there are two types of fibre cell, an S-shaped cell and a concentric cell. Overlapping tails of S-shaped cells form the lens sutures, which paradoxically lie along the visual axis and can affect optical quality and produce variability in focus.

In addition, the circulation within the lens provided for by the nanometer space between the gap

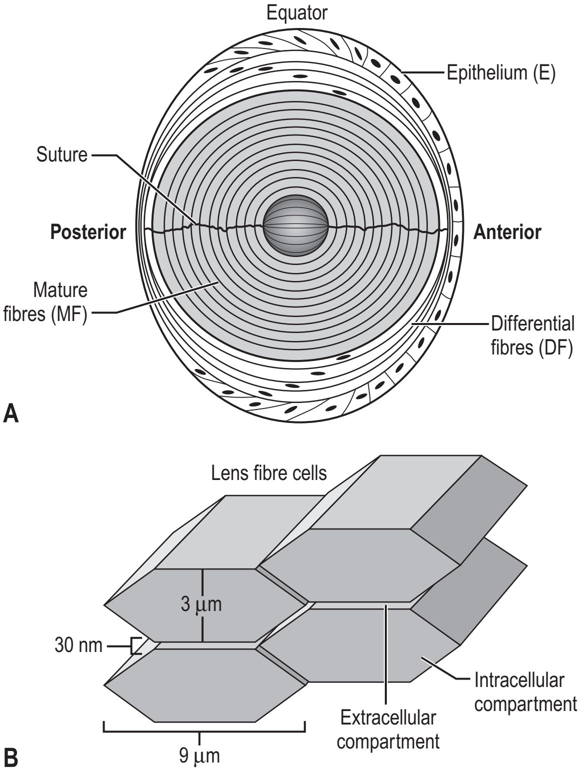

FIGURE 4-48 (**A**) Three zones comprise the lens structure: the epithelium (function: transport of proteins), the anucleate mature lens fibres (function: light transmission) and the differentiating fibres (retain some organelles: function: nutrient transport link between epithelium and lens fibres). (**B**) The lens fibres appear as flattened hexagons when cut in cross-section. (From Mathias et al., 2010.)

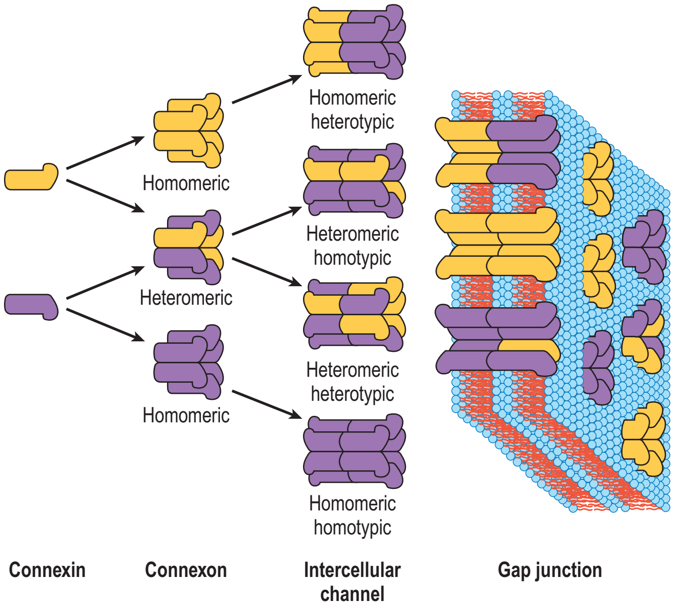
Homomeric
heterotypic
Homomeric
Heteromeric
homotypic
Heteromeric
Heteromeric
heterotypic
Homomeric
Homomeric
homotypic
Connexin
Connexon
Intercellular channel
Gap junction

FIGURE 4-49 Gap junctions are a major feature of cellular connections in the lens: they comprise many closely packed channels each formed by two hemichannels (connexons), one in each of the connecting cells. (From Mathias et al., 2010.)

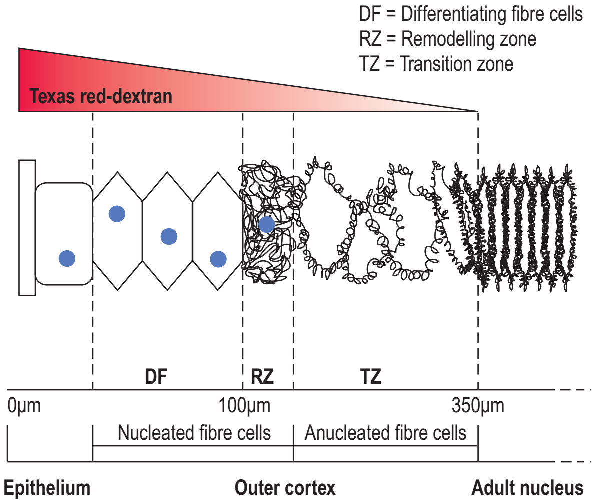
DF = Differentiating fibre cells
RZ = Remodelling zone
TZ = Transition zone
Texas red-dextran
DF
RZ
TZ
0μm
100μm
350μm
Nucleated fibre cells
Anucleated fibre cells
Epithelium
Outer cortex
Adult nucleus

FIGURE 4-50 Summary diagram of the changes in the lens as adult lens fibres differentiate from lens epithelium. (From Borchman and Yappert, 2010.)

junctions in fibre cells is essential for lens fibre metabolism even at is basal level. Mutations in gap junction connexin proteins account for many different types of congenital cataract.

The crystallins

Crystallins make up 90% of the water-soluble proteins of the lens; there are three types in mammals (see eBox 4-4).

Some crystallin protein also coextracts with the urea-soluble protein, indicating that a fraction of the crystallins is strongly bound to the cytoskeletal (urea-soluble) proteins (5% total protein). The water/urea-insoluble protein represents membrane protein (2% total protein) and some crystallin is also found in this fraction when it is solubilized in detergent.

The αA and αB crystallins show about 50% sequence homology. The molecules exist as polydisperse globular proteins in aggregates organized in three concentric layers or as a protein 'micelle'. However, their true quaternary structure is unknown. αA and αB crystallins belong to the family of small heat-shock proteins and display chaperone-like activity. Studies in knockout mice have shown that αA lens-specific crystallin is required for normal lens differentiation and transparency, while αB, which is expressed in neural tissue and upregulated under conditions of stress, is not essential for lens transparency. Homology of lens crystallins to certain enzymes such as aldehyde dehydrogenase class 3 enzymes in corneal epithelial cells suggests that this form of 'gene sharing' is quite widespread.

The crystal structure of some small invertebrate heat-shock proteins such as Hsp16.5 has allowed a model for lens αA crystallin micelles to be developed (Fig. 4-51). In this model it has been shown that neither αA or αB is necessary to develop the molecular

eBox 4-4

The crystallins

Three types of crystallin have been identified in mammals, α, β and γ, mainly on the basis of molecular weight (see figure). The δ crystallins have also been detected in birds. Several other 'taxon-specific' crystallins (ε, τ, ρ, χ, μ, λ, ζ, SIII) are recognized in other species, based on the criterion that they account for at least 10% of the total water-soluble protein. More recent data suggest that in vertebrates there are only two classes of crystallins, α and γβ. Native α crystallin is of two types, αA and αB, each with a molecular weight of about 20 kDa. In the native state, however, the α crystallins form large multimeric aggregates of 300–1200 kDa (average 800 kDa), held together by non-covalent interactions.

β Crystallins range in molecular weight from 23 to 35 kDa and occur in several subtypes: βB_1, βB_2 and βB_3; βA_2, βA_3 and βB_4. Mixed aggregates of between 50 and 200 kDa occur naturally. The γ crystallins are monomeric in the native state; there are six types (*γA–E, and γS (formerly known as βS)), differentiated by charge. Not all types are present in human lenses at all ages, some such as γS and γC being present at higher concentrations in fetal than in adult lenses. The relative amounts of α:β:γ crystallin also vary greatly depending on age and other factors; in the 'typical' lens the ratio of α:β:γ is of the order 40:35:25. Protein sequence analysis has shown homology between the β and γ crystallins; in addition, all three proteins exist as β-pleated sheets.

The structure of the human lens (**A**) and the major soluble lens crystallins (α, **B**), (β, **C**) and (γ, **D**) is shown below.

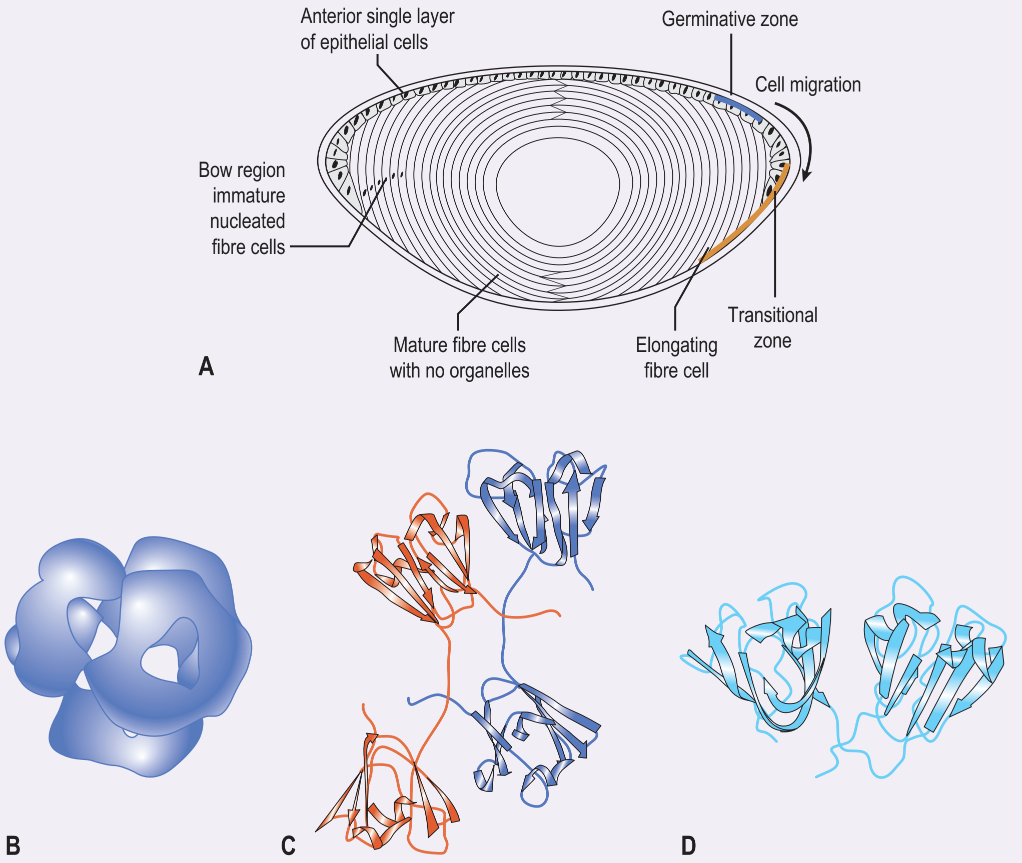
Anterior single layer of epithelial cells
Germinative zone
Cell migration
Bow region immature nucleated fibre cells
Mature fibre cells with no organelles
Elongating fibre cell
Transitional zone
A
B
C
D

(From Moreau and King, 2012.)

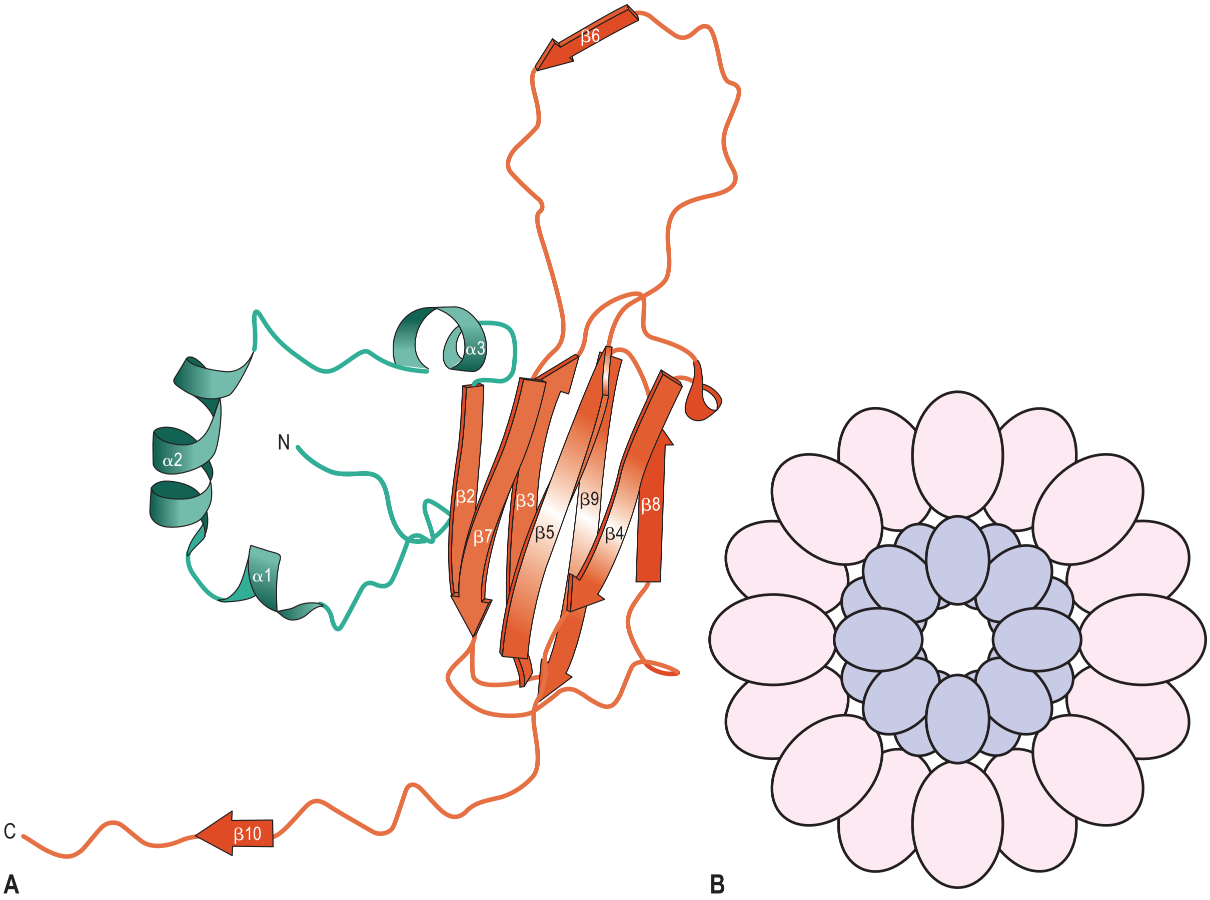

FIGURE 4-51 (**A**) Secondary structure of the Hsp16.9 subunit with an ordered N-terminus. The N-terminal domain (green) contains three helical segments, as shown, in half of the subunits. The remaining subunits in Hsp16.9 and all subunits in Hsp16.5 have unstructured N-termini. The α crystallin domain (brown) consists of seven-stranded β-sandwich, an interdomain loop containing one β-strand (β6 at the top) and a C-terminal extension (at the bottom), which is largely unstructured except for the short V10-strand. (**B**) A possible micelle-like structure for α crystallin. The subunits contain two domains and assemble into large aggregates through interactions between their hydrophobic N-terminal domains (lilac), which are located in the centre of the aggregate. The hydrophilic C-terminal (α crystallin) domains (not pink) are on the surface of the assembly. (Part A is reproduced with permission from Nature Publishing Group. Part B is from Augusteyn, 2004, with permission from the Optometrists Association of Australia.)

arrangement but that some combination of either will suffice. However, the tertiary and quaternary structure of human crystallins remains elusive due to expected difficulties in achieving crystallization.

Phosphorylation of the αA_2 and the αB_2 chains produces the αA_1 and the αB_1 chains. Spontaneous non-enzymatic cleavage of the molecules also occurs, as does high molecular weight aggregation, especially with age. The α crystallins, acting as molecular chaperones, 'trap' other crystallins and proteins such as intermediate filaments, which may be undergoing denaturation and unfolding. In this way they maintain lens transparency by preventing disruption of the highly ordered structure of the crystallin packing.

The γβ crystallins are thought to have a similar structure: four repeating antiparallel β sheets in the form of 'Greek key' motifs. γ Crystallin is a highly stable molecule, attributed to its extensive internal symmetry; recent studies have revealed a link between mutations in congenital cataract and cataract of old age (Fig. 4-52).

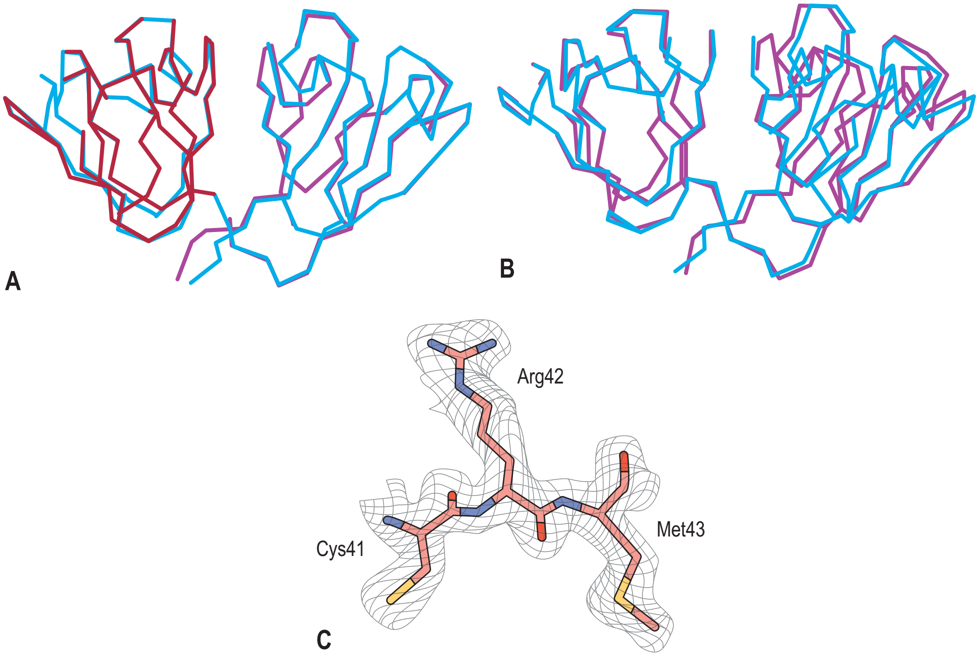

FIGURE 4-52 Crystal structure of human W42R γD-crystallin, a mutant form of human γ crystallin which has only minor differences to normal γ crystallin but is protease-sensitive and is associated with certain forms of congenital cataract. (**A**) Superposition of the normal and mutated forms of γ crystallin. (**B**) Best fit superposition. (**C**) Electron densitiy of residue Arg42 mutated from Trp42 in the normal γD crystallin. (From Ji et al., 2013.)

Multimeric complexes of β crystallins tend to form between the acidic molecules (βA_2, βA_3 and βA_4) and the basic molecules (βB_1, βB_2 and βB_3), followed by association between similar heterodimers. Homology among the various β crystallins both within and between species is quite variable and sequence analysis is still in progress for most of the human proteins. However, on the basis of γβ sequence homology a predicted structure of β crystallin has been suggested and X-ray crystallography has shown it to have some basis. In this model, the two γ-like structures are joined by a connecting peptide.

Although the molecular packing of the crystallins and their high refractive index contribute extensively to lens transparency, biophysical studies of molecular interactions between proteins indicate that the crystallins in themselves are not essential – proteins that can adopt the correct state of phase transition and osmotic pressure would do equally well. In essence it is related to the volume exclusion properties of proteins which are at their highest with γD crystallins, the most abundant crystallin in the lens.

Cytoskeletal proteins of the lens

Cytoskeletal proteins are usually to be found in the urea-extractable fraction of lens proteins. In addition to the usual complement of microfilaments, such as actin, vimentin and spectrin, and intermediate filaments (see section on cells and tissues above), there are certain lens-specific intermediate filaments such as beaded filaments. Vimentin is the major intermediate filament in the lens cell, and is present in epithelial and cortical fibre cells but not in nuclear fibre cells. A similar distribution has been found for microtubules in lens cells. Cytokeratins are not found in the adult lens. Some differences occur in relative proportions of cytoskeletal elements. Thus, talin, α-actinin and the signalling proteins are at high concentration in lens equatorial epithelium, while vinculin is prominent in stable fibre cells with strong cell–cell contacts.

Beaded filaments are specific to the lens fibre. Two main species have been identified: beaded filament-specific proteins 1 and 2 (BFSP1, BFSP2, also known as filensin and CY49, respectively). In addition, there are link proteins, the plakins, which together are likely to be involved in crystallin packing and density distribution, perhaps by offering attachment sites for crystallin molecules. CP49 is also known as phakinin. Both proteins co-assemble with α crystallin but not with vimentin.

Lenses with targeted deletions of phakinin and filensin are opaque even though lens fibre morphology is normal, indicating that these two proteins, which co-assemble to form beaded filaments, are essential for lens transparency through lens fibre cytoskeletal organization.

Membrane lipids and proteins

Membrane lipids in lens cells are highly saturated, the main phospholipid being the extremely stable dihydrosphingomyelin. In cell membranes they are densely populated with proteins, thus restricting their mobility. Cholesterol is also abundant, which in view of its exclusion from protein 'rafts' (see p. 162) leads to patches of pure cholesterol in the membrane (cholesterol bilipid domains). Lens cell membrane proteins are extractable from lens membranes in detergents such as sodium dodecyl sulphate (SDS), a major protein being lens fibre cell-specific junctional complex protein aquaporin-0 (AQP0), previously known as major lens intrinsic protein 26. In fact this was the first aquaporin (or water channel) identified and they act as osmoreceptors or cell volume regulators. At least 100 of these genes have been described, 11 of which are present in mammalian systems and five involved in fluid transport through various ocular tissues in and out of the eye (Fig. 4-53).

The *AQP0* gene is located on the cen-q14 region of the long arm of chromosome 12. The protein is not only involved in water transport but also in intercellular communication and movements of ions as part of the microcirculation of the lens. In the closed configuration it acts as an adhesive protein for the posterior lens capsule, maintaining the lens electrical dipole. AqP0 is important for lens transparency; mutations in this gene causing cataracts. AqP0 is absent from lens epithelial cells and its relationship to gap junction proteins in other cells is unclear. However, AQP1 is present in lens epithelium.

Other membrane proteins include numerous enzymes such as ATPases, and cytoskeletally attached proteins such as calpactin-1 and *N*-cadherin. Several other high molecular weight proteins also exist in the lens fibre plasma membrane such as plakoglobin, plectin, periplakin and desmocollin, members of the spectrin family with a role in maintaining cell shape as for spectrin in red cells. Lensin is a lens-specific cytoskeletal protein linking beaded filaments to cell membrane.

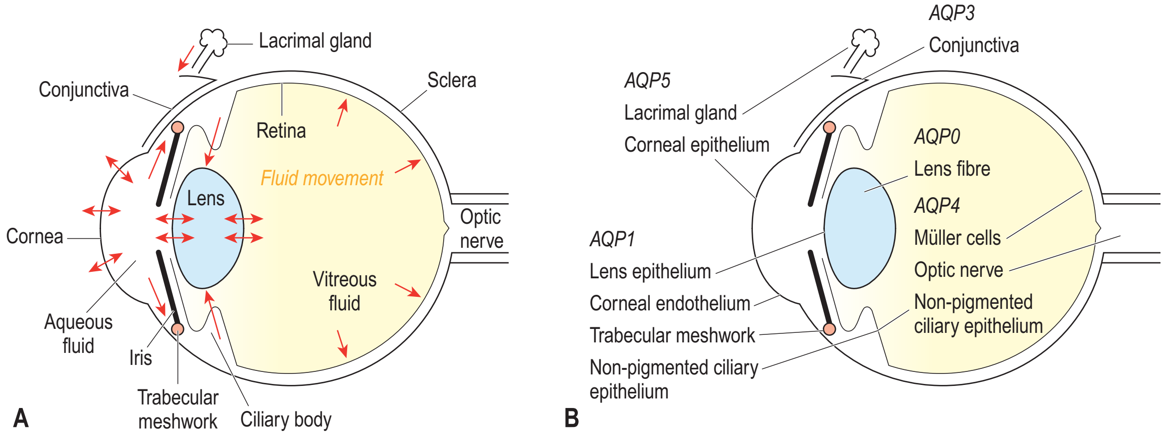
Lacrimal gland
Conjunctiva
Sclera
Retina
Fluid movement
Lens
Cornea
Optic nerve
Vitreous fluid
Aqueous fluid
Iris
Trabecular meshwork
Ciliary body
A
AQP3
Conjunctiva
AQP5
Lacrimal gland
Corneal epithelium
AQP0
Lens fibre
AQP4
Müller cells
Optic nerve
Non-pigmented ciliary epithelium
AQP1
Lens epithelium
Corneal endothelium
Trabecular meshwork
Non-pigmented ciliary epithelium
B

FIGURE 4-53 Fluid transport and aquaporin expression in the eye. (**A**) Routes of fluid movement showing secretion by lacrimal gland and ciliary body, absorption by trabecular meshwork and retinal pigment epithelium, and bidirectional movement in cornea and lens. (**B**) Sites of aquaporin (AQP) water channel expression in ocular tissues. (From Verkman, 2003, with permission from Elsevier.)

Extracellular matrix

The only extracellular matrix of any importance in the lens is the capsule. The capsule is constructed as for any epithelial cell basement membrane of type IV collagen and heparan sulphate proteoglycan, and acts as a diffusion barrier for the lens. Fibronectin is localized to the anterior capsule, while tenascin is present in the posterior capsule. Tenascin is one of a family of matricellular proteins which includes thrombospondin and SPARC (sialo-protein associated with rods and cones, see p. 183 and Fig. 4.19), the last being required for lens transparency. The $\alpha_5\beta_1$ integrin is present in the anterior lens epithelium, while the $\alpha_6\beta_1$ integrin receptor for laminin is present in equatorial and lens fibre cells, both of which are migratory.

Semipermeable membranes and physiology of the lens

As indicated above, the lens behaves like a very large syncytium or single cell, both electrically and chemically. Active pumping mechanisms, based on Na^+/K^+ ATPases exist to pump Na^+ ions out of the lens, while chloride and water are transported into the cell (see Box 4-15). The pump is located in the epithelium with the highest concentration at the lens equator, while lens fibres do not have a pump function. As a result, electrical current density across the epithelium is highest at the equator, generating a flow of current from the equator to the lens centre, and this appears to be critical for intra-lens microcirculation (see Box 4-15).

Barriers to ion and solute transport occur at the capsule and at the plasma membranes of the epithelial and fibre cells. The capsule is permeable to small molecular weight proteins (<50 000 Da), including low molecular weight crystallins, but prevents diffusion of large molecules.

At the epithelial barrier, the cells show the typical polarization of other epithelial cells but lack tight junctions at the lateral cell surface. Instead there is an extensive system of gap junctions, which permits rapid intercellular communication, thereby allowing the cells to behave as a syncytial sheet. At the junction between the epithelium and the fibre cell, the main transport mechanism is rapid endocytosis via coated vesicles, while the very extensive system of gap junctions between each lens fibre permits rapid interfibre cell movement of metabolites, amounting in effect to a microcirculation (see Figs 4-48 and 4-49). Indeed, 50% of the fibre cell plasma membrane protein consists of AQP0, while the gap junctions in lens are to some degree specialized (more efficient) than other cell types in that AQP0 appears to assist in channel formation between the cells.

BOX 4-15 TRANSPORT OF MOLECULES ACROSS THE LENS SURFACE

The lens behaves like a syncytium in which K^+ is transported into the lens and Na^+ is transported out via Na^+/K^+ ATPase present in the lens epithelium (**A**). Interestingly, the concentration of the ATPase is highest in the equatorial zone of the lens epithelium where much of the ionic transport occurs. This sets up electrical gradients with differential electrical potential differences between the lens equator and the lens poles, in part accounting for the electrical dipole that occurs in the lens (**B**). The lens also contains specific glucose transporters and transporter molecules for ascorbate and water, which ensure adequate metabolism and minimize free radical damage.

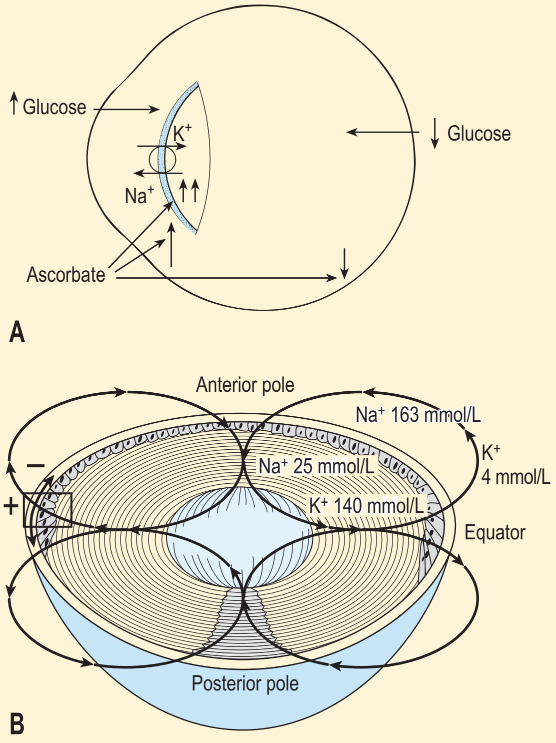

(Part B reproduced from McCaig et al., 2005.)

The Na^+/K^+ ATPase pump in the epithelium actively exchanges Na^+ (pumped out) for K^+ (pumped in). The Na^+ passively diffuses anteriorly down a concentration gradient present in the vitreous, across the posterior lens capsule into the lens body, where it rapidly diffuses to the anterior epithelium and is pumped out into the aqueous. K^+ ions are handled in the reverse direction, eventually diffusing passively across the posterior capsule into the vitreous. Inward currents at the anterior and posterior poles occur in the extracellular space while outward currents driven by the ATPase pump occur at the epithelium, particularly the equatorial epithelium. Critical to this process is the presence of K^+ channels in the epithelium but not in the lens fibres, which allow an overall Na^+/K^+ exchange.

While this simple pump–leak model serves to explain ion transport across the lens body there are several unanswered questions, such as the mechanism of Ca^{2+}/Mg^{2+} transport, which also occurs via a specific ATPase. The Ca^{2+}/Mg^{2+} ATPase is most abundant in the lens cortex. In addition, specific transporter proteins for glucose (GLUT1 in epithelium, an Na^+-dependent SGLT transporter in lens fibres) and amino acids exist in the plasma membrane of lens fibre and epithelial cells.

Transport of water across the lens has also been shown to occur in an anteroposterior direction at a rate of about 10 μL/h, probably involving AQP0, thus participating in the overall lens microcirculation, and helping to rid the lens of waste products and maintain transparency.

LENS METABOLISM

Carbohydrate

Glucose from the aqueous humour is the main source of energy for lens metabolism. Glucose enters the cell via an insulin-dependent glucose transporter located in the plasma membrane (GLUT1). Both the glycolytic and the pentose phosphate pathways are used, and under conditions of excess glucose the sorbitol pathway is entrained (Fig. 4-24). About 80% of glucose is consumed by the lens via anaerobic glycolysis. The pentose phosphate pathway uses about 10% of the remaining glucose, providing sugar residues for nucleotide synthesis. Aerobic glycolysis via the citric acid cycle occurs only in the lens epithelium because these are the only lens cells to possess mitochondria. The epithelium also possesses most of the aldose reductase, indicating that any metabolic activity occurring via the sorbitol pathway takes place in this cell. Under normal circumstances, less than 5% of glucose is used up in the sorbitol pathway. Indeed, it is unlikely that sorbitol has a significant role in the induction of complications of high ambient glucose (diabetes). Aldose reductase is apparently induced by osmoreceptors (such as aquaporins) and interferes with NADH-binding proteins, thus disturbing regulatory free radical scavenger mechanisms (see below). In addition, aldose reductase has recently been shown to have $PGF_{2\alpha}$ synthase activity, is involved in the COX2 pathway, and is regulated by IL-1β. Indeed, inhibition of aldose reductase is reported to prevent allergic rhinitis by blocking the PI3 kinase/Akt/GSK intracellular signalling pathway (see Ch. 7, p. 445). Thus, it has a more direct role in inflammatory mechanisms, which are also implicated through other routes in the complications of diabetes.

Protein

Synthesis of new protein ceases with lens fibre cell formation, and all changes that occur to lens proteins after this stage are post-translational modifications. Phosphorylation of many proteins occurs, including crystallins, cytoskeletal proteins and AQP0. Several phosphorylation systems exist, including a cAMP-dependent protein kinase A and a phospholipid-dependent protein kinase C. Certain drugs enhance phosphorylation of intermediate filaments, including β-adrenergic compounds.

Numerous enzyme activities have been detected in lens protein extracts but the level of enzyme protein is very low. Some of the taxon-specific crystallins have apparent enzyme activity such as ε crystallin from duck lens (lactic dehydrogenase), ρ crystallin from frog (aldose reductase) and ε crystallin from frog (lung prostaglandin F synthase). These findings are of evolutionary rather than physiological significance but because several of these enzymes are induced in cells undergoing stress it has been suggested that stress responses may be the common denominator in these homologies (see above, heat-shock protein and α crystallin). Stress proteins and long-lived lens crystallins may require similar properties to maintain stability and durability in anaerobic conditions. Thus, crystallins, especially the α and γβ series, are among the most conserved proteins known and interestingly are not restricted to the lens (αβ has been found in heart, lung, brain and retina). The promoter sequence of the αA gene has been shown to be lens-specific and has the capability of driving foreign genes selectively into its sequence. This has been proposed as an explanation for the interchangeability of function for apparently identical proteins from widely divergent sources. Similarly, the promoters for χ crystallins are lens-specific. During development and growth there is differential expression of the various crystallin genes in a highly regulated manner.

In addition to the evidence for enzymic activity in the crystallin proteins, the lens has several other proteolytic enzymes, including endo- and exopeptidase activity and membrane-associated proteases. Historically, the denucleation of the lens fibre has been described as a form of attenuated apoptosis, taking place over several days rather than hours. In addition, some of the 'death' enzymes such as caspases are activated during this process, although mice deficient in caspases 3 and 6 have normal lenses. In addition, a key component of apoptosis, namely the phosphatidylserine 'inside-out' membrane flip, does not occur. Despite these caveats, it is likely that much of the signalling machinery involved in apoptosis and lens fibre differentiation overlaps in function. In addition, autophagy does not play a part in this process since Atg5 –/– (autophagy defective) mice also have normal lenses.

The neutral endopeptidases calpain I and II (cysteine Ca^{2+}-dependent enzymes) and their inhibitors, calpactins, have also been detected in lens cells. Substrates for these enzymes include cytoskeletal proteins and crystallins, and their role is probably related to protein turnover. Calpain I is present in the epithelium and lens cortex but not in the nucleus. Dysregulation of calpain genes has been suggested as a cause for age-related cataract.

Increased degradation of proteins occurs with age, particularly that of MIP26, which may have significance for coordinated intercellular functions of lens fibres and contribute to cataract. Of interest is the fact that ubiquitin conjugation, a protein degradation system in which the small 8.5 kDa ubiquitin molecule binds to proteins before degradation, is markedly reduced in aged versus young lens nuclei.

Lipid

The unusually high concentrations of sphingomyelin, cholesterol and saturated fatty acids impart rigidity to the cell membrane (see above), which may be important in maintaining intercellular connections. In particular, the cholesterol bilipid domains ensure a high content of cholesterol in the neighbouring regions, thus maintaining stability of the membrane even if there are significant changes in the membrane phospholipids (Fig. 4-54).

In addition, high levels of phosphatidylinositol are also found in the lens, suggesting significant receptor-mediated second messenger activity in lens cells, e.g. responsiveness to hormones and catecholamines.

Redox systems in the lens microenvironment

The lens is constantly exposed to attack by oxidative agents; indeed there is a high level of hydrogen peroxide in normal aqueous and peroxidase activity is also present in the lens itself. Several enzyme systems are available to minimize or buffer the effects of oxidants, including catalase, superoxide dismutase, glutathione peroxidase and glutathione *S*-transferase. The lens contains high levels of glutathione (3.5–5.5 μmol/g wet weight), with the highest concentration in the epithelium, and detoxification via the mercapturic acid pathway is an important pathway in the lens. Glutathione is produced from the interaction between glutamate and cysteine in lens cells. According to one theory glutathione diffuses towards the centre of the lens, where lens fibre cells are essentially non-metabolizing and thus unable to generate glutathione. Diffusion would occur in the opposite direction to the proposed microcirculation in the lens (see p. 229) which is dependent on ion and water transport and this difference of viewpoint has not yet been resolved. However, the current flow and the lens dipole measurements (see Box 4-15) support a pump-driven circulation rather than diffusion.

Glutathione is also important in protecting thiol groups in proteins, especially cation-transporting membrane proteins in the lens, which additionally accounts for its unusually high concentration in this tissue. More than 95% of glutathione is in the reduced state.

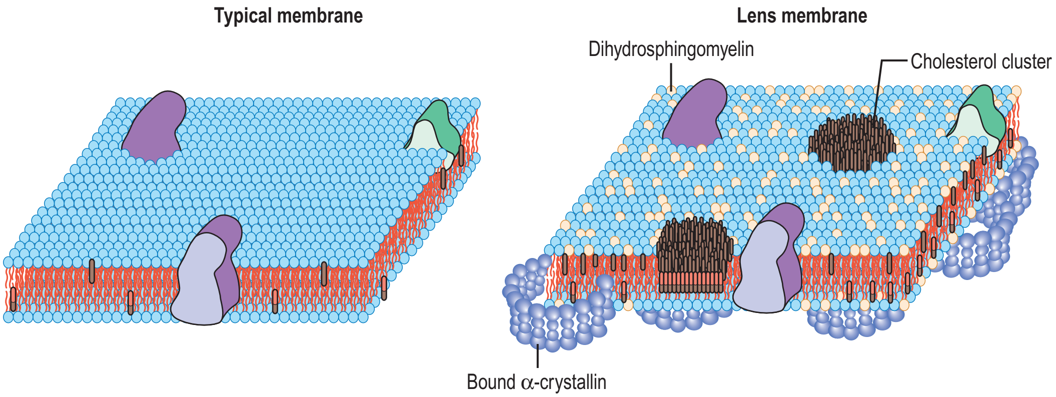

FIGURE 4-54 Figure on the right shows a 'typical' membrane containing few cholesterol molecules. On the left is shown a human lens membrane which contains many cholesterol molecules while most of the lipid is associated with crystallins and other lens-specific molecules such as aquaporin. The lipid-saturated membranes are highly ordered and stiff. The major lipid is dihydrosphingomyelin. (From Borchman and Yappert, 2010.)

Catalase and low levels of superoxide dismutase have also been identified in lens epithelium, indicating that these systems are also probably important.

AGEING IN THE LENS AND CATARACT FORMATION

The transmission of light decreases with age, especially for the lower wavelengths (up to a factor of 10), to the point that at low levels of illumination an apparent tritanopia can occur (see Ch. 5, p. 302). Morphologically, the cells lose cytoskeletal organization and develop vacuolation and electron-dense bodies. An increase in sodium concentration is accompanied by a decline in the membrane potential of the lens, suggesting ion channel dysfunction. Enzymatic activities decline in the lens nucleus but not in the cortex or epithelium. In addition, the appearance of water clefts in the lens as an early sign of cataract suggests decreased function of AQP0 (MIP26) and fluid transport.

Ageing of the lens and cataract formation are not synonymous. In age-related nuclear cataract there is extensive oxidation of cystine and methionine residues on lens proteins, while in aged lenses without cataract, oxidation is much less. Glutathione SH (oxidized glutathione) (see Box 4-8, p. 194) is the key. Post-translational modification of lens proteins continues throughout life. In addition to cross-linking and degradation, which occur in any stable protein system, non-enzymatic glycation is a conspicuous event. In general, γD crystallins are synthesized in young lenses, while production of γσ and β crystallins increases with age. In addition, most of the α crystallin is lost from the water-soluble compartment to the water-insoluble compartment, as are some of the β and γ forms.

Non-enzymatic glycation of crystallins occurs at the ε amino groups of lysine, especially the high molecular weight aggregates of α crystallin. *In vitro*, this reaction produces a yellow fluorescent pigment similar to that seen in the ageing human lens. Interaction between various amino groups and aldehydes released from free radicals, especially through lipid, (per)oxidation produces fluorophores and ceroid/lipofuscin. In spite of the colour changes, the amount of protein that is glycated is less than 5% in an aged lens, which is considerably less than for other long-lived proteins such as haemoglobin and collagen. Lens crystallin glycation is more likely to be the result of its interaction with oxidized ascorbic acid than glucose on the basis of intra-lens concentrations, and it is possible that glutathione, by maintaining ascorbic acid in its reduced state, inhibits this process.

AQP0 also undergoes modification with age, losing a 5000 Da peptide to become MIP22 in increasing concentration. Cleavage occurs at both the C- and N-terminal ends of the molecule.

Reduced vision in cataract is caused by increased light scatter by lens proteins

The transmission of light by the lens is reduced when the ordered packing of the lens crystallins is disturbed. This can be induced in many ways, such as increased water accumulation within the lens, formation of high molecular weight lens protein aggregates, and vacuole formation within the lens fibres with age.

Certain metabolic conditions are associated with cataract, the best known being the cataract of diabetes and a similar lens opacity in galactosaemia. In these forms of 'sugar' cataract, accumulation of water in the lens fibres was previously thought to result from the accumulation of non-degradable polyols such as sorbitol and galacticol in the lens fibre cells. High glucose/galactose concentrations in the aqueous lead to increased intracellular accumulation of glucose, which saturates the normal anaerobic glycosis pathways. Accordingly, aldose reductase is thought to fail in regulating the polyol pathway and polyols (such as sorbitol) accumulate in the cells, thereby increasing the osmotic drag of water into the cell via activation of the osmoreceptor AQP0. However, aldose reductase also affects $PGF_{2\alpha}$ synthesis and inflammatory processes via $PGF_{2\alpha}$ (see p. 197) and thus the precise role of aldose reductase in cataract formation is unclear.

Whatever the mechanism, the dysregulation in cellular metabolism, with reduction in the levels of cellular ATP and glutathione, and secondary damage to the cell, plus the increased water content causes phase separation between protein-rich and protein-poor regions of the cells and increased light scatter (cataract). As individual cells loosen their interdigitations with neighbouring cells, water clefts and vacuoles appear within the lens substance. As cells die, there is progressive increase in opacification, which in the lens cortex is seen as 'spoke-like' opacity and in the nucleus is characterized by the accumulation of insoluble protein aggregates and chromophores, causing the nucleus to change colour from yellow to red to black.

Cataract formation is caused by any insult to the lens

Since the lens is designed for the transmission of light, it responds to any insult that disturbs normal development or metabolism by opacification, even if this is only for a temporary period. Thus, certain congenital cataracts appear to affect only the fetal nucleus; radiation cataract may be limited if only a discrete area of the lens bow region is affected; sunflower cataract of trauma may be the result of shearing forces momentarily separating lens fibre cells, which then restore their interconnections; and certain forms of cataract such as the 'feather' cataract after vitrectomy are thought to be the result of large volumes of fluid transfused through the vitreous cavity at too low a temperature or an incorrect electrolyte composition. Interestingly, such cataracts may be reversible, suggesting that the lens microcirculation is temporarily unable to cope with increased fluid load but can recover when bulk fluid flow is normalized. Similarly, 'cold' cataract is induced in young animals and is caused by the reversible precipitation of γ crystallins by phase separation in the fetal nucleus.

Many forms of cataract have been attributed to specific mutations in lens crystallins probably related to impaired molecular packing as well as several other proteins. For instance, a mutation in the Wolfram gene (*WFS1*) has been identified as the cause of congenital nuclear cataract while another gene, the *FCO1* gene, has also been implicated in a recessive form of congenital cataract.

Age-related cataract formation is multifactorial

A great wealth of studies during the past 20 years has shown many biochemical changes associated with age-related cataract. In summary, these include an increase in the insoluble components of the lens, an increase in chromophores, increased protein cross-linking and aggregation, and oxidation of amino acid groups. There is a concomitant decrease in antioxidant enzyme systems and increased proteolytic activity. The level of glutathione is also reduced. Normal lenses contain a trypsin inhibitory activity, which may regulate age-related proteolytic activity. The major protein change in cataractous lens is the loss of αA crystallin and the selective loss of γS crystallin. In addition there are numerous degradation peptides detectable in the water-soluble component. These mechanisms are summarized in Figure 4-55.

The possibility that UV light might cause or hasten some of the effects of age has been suggested by the observation that age changes appear to be more marked in the region of the visual axis than in the equatorial region of the lens. Current views are that oxidative events are the most likely mechanism of cataract formation. Near-UV light is absorbed by tryptophan, which in sunlight is converted to *N*-formyl-kynurenine, a fluorescent chromophore similar to 3-hydroxy-kynurenine, a second UV absorbent molecule in the lens. Both these compounds can act as photosensitizers and lead to the production of the free radical singlet oxygen (see Box 4-8). Free radicals downregulate the function of critical lens enzymes such as Na^+/K^+ ATPase and lead to lens swelling and opacification, at least in the rat model. Other free radicals generated by near-UV light such as hydrogen peroxide have been implicated in the dysfunction of hexokinase, an enzyme central to glucose utilization in the lens.

Oxygen increases the rate of photo-oxidation, and vitamin E, ascorbic acid and glutathione reduce the effects of light damage (Fig. 4-55).

The role of UV light in human cataract is unclear, although exposure to UV-B is reportedly associated with an increase in cortical and posterior subcapsular cataract but not in nuclear cataract formation. Interestingly, aged human lenses appear to absorb more UV-A and even visible light than young lenses.

Certain trace metals and compounds are associated with cataract. Experimental depletion or excess of selenite leads to cataracts by a mechanism that appears to be closely interwoven with Ca^{2+} homeostasis. In contrast, cyanate induces carbamylation of lens proteins and cataract, a process that can be prevented experimentally with aspirin. Interestingly, aspirin usage may also delay the onset of cataracts in humans.

BOX 4-16 FREE RADICAL DAMAGE AND THE LENS

Free radical damage in the lens may occur through oxidative metabolism but is considered mostly to be the result of UV damage by activation of endogenous photosensitizers.

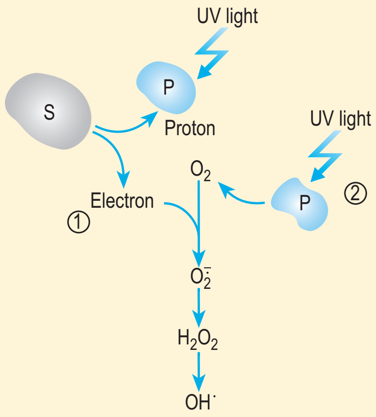
UV light
P
S
Proton
UV light
O_2
Electron
①
P
②
O_2^-
H_2O_2
$OH^\cdot$

Photosensitizers act by one of two mechanisms: (1) triplet sensitizer absorbs a proton from a substrate, leading to the production of free radicals (SO_2 and H_2O_2); (2) sensitizer reacts with O_2, leading to singlet O_2. Typical photosensitizers include riboflavin, tryptophan and kynurenine (all present in the lens).

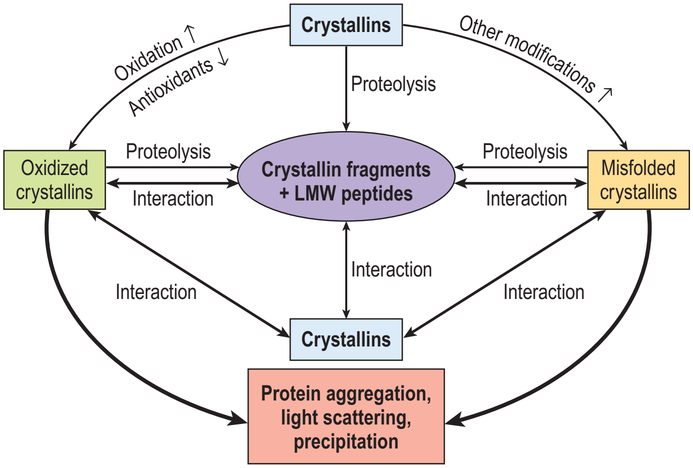
Crystallins
Oxidation ↑
Antioxidants ↓
Other modifications ↑
Proteolysis
Oxidized crystallins
Proteolysis
Interaction
Crystallin fragments + LMW peptides
Proteolysis
Interaction
Misfolded crystallins
Interaction
Interaction
Interaction
Crystallins
Protein aggregation, light scattering, precipitation

FIGURE 4-55 Schematic representation of the proposed role of crystallin fragments in ageing of the lens and development of cataract. (From Sharma and Santhoshkumar, 2009.)

Inhibition of cholesterol synthesis also leads to cataract in experimental animals. Potentially, reduced levels of magnesium may promote cataract formation.

Metabolites and chemicals such as pyruvate and caffeine have been shown to prevent effects of UV and selenium *in vitro*, respectively.

In summary, although many factors may contribute to cataract formation, probably the most important is glutathione consumption and unrestricted oxidation of intrinsic lens protein.

Mechanism of age-related cataract formation: a failure of chaperone function

As shown above, there are several processes which can lead to damaged lens proteins. Oxidation, carbamylation, deamidation and other perturbations of βγ crystallins lead to their progressive inability to sustain normal intermolecular interactions. As they denature and precipitate, the α crystallins bind the unfolded proteins, but unlike true chaperones they do not have the ability to refold the βγ crystallins. As a result, the chaperone capacity of the α crystallins is consumed and the complexes precipitate within the lens fibre cells, forming the insoluble protein fraction that increases with age.

The vitreous

THE VITREOUS BODY IS A TRUE CONNECTIVE TISSUE CONTAINING COLLAGEN, GAGS AND CELLS

The vitreous is 98% water, 1.0% macromolecules, and the rest solutes and low molecular weight materials.

The matrix

The vitreous transmits light by the same mechanism as the cornea, i.e. its collagen fibrils (10–20 nm) are thinner than half the wavelength of light, and the interfibrillar space is filled with GAGs (hyaluronan) at intervals that reduce the effects of diffraction in the system. Collagen imparts the gel structure to the vitreous body, is predominantly type II, similar but not identical to cartilage type II collagen (vitreous collagen has more galactosyl-glucose side chains and a higher content of alanine), and is arranged in a lattice structure in which the fibrils are suspended in a viscous hyaluronan solution. This structure is lost with age and in disease by a process known as syneresis in which the hyaluronan molecules are degraded to smaller moieties and the collagen fibrils coagmentate to form larger fibrils, becoming visible as ‘floaters’. Some type VI and type IX collagens are also present and play structural roles in gel formation. In addition, a hybrid molecule composed of type V/XI α chains has been detected in vitreous in trace amounts and has been implicated in vitrous fibril formation as for cornea. In cartilage, collagen type II fibrils exist in two forms, thin (~16 nm) and thick (~40 nm), and collagen type XI is found exclusively in the thin fibrils (Fig. 4-56). Presumably collagen type XI has a similar role in the vitreous in regulating fibril diameter.

A recent member of the small leucine-rich repeat (SLR) extracellular matrix protein family, termed opticin, has been identified in vitreous and also in ligament, skin and retina. It also has a role in regulating fibril thickness in the vitreous similar to collagen type IX and V/XIα proteoglycans. Importantly, opticin has anti-angiogenic properties.

The normal vitreous in the young adult has a distinct architecture (see Ch. 1, p. 37). The cortex has a higher concentration of collagen and hyaluronan than the central vitreous and, in addition, the cortex contains other GAGs such as chondroitin sulphate, which may be important in vitreo/retinal apposition.

In the central vitreous gel, hyaluronan is essentially the sole GAG. Hyaluronan occurs as stiff, open coil disaccharide chains, which in solution become entangled at concentrations above 300 μg/mL and thus add support to the gel matrix. Hyaluronan concentrations in human vitreous vary between 100 and 400 μg/mL and the molecule binds to type IX collagen, which acts as a proteoglycan to bind the hyaluronan to the collagen fibrils. However, hyaluronan is not essential for maintaining the gel structure of the vitreous. Therefore, some long-range interactions mediated via type IX collagen are necessary. Type IX collagen is a small non-fibrillar type of collagen that contains several non-collagenous domains; these act as the proteoglycan bridges. In addition, the chondroitin sulphate proteoglycan is present in vitreous at a concentration of approximately equal ratio with type IX collagen. Traces of chondroitin sulphate also comprise part of the proteoglycan versican in the vitreous with a molecular weight of 2×10^4 to 4×10^4 Da. Hyaluronan is highly polydisperse (variable molecular sizes) and its breakdown products can activate innate immune cells

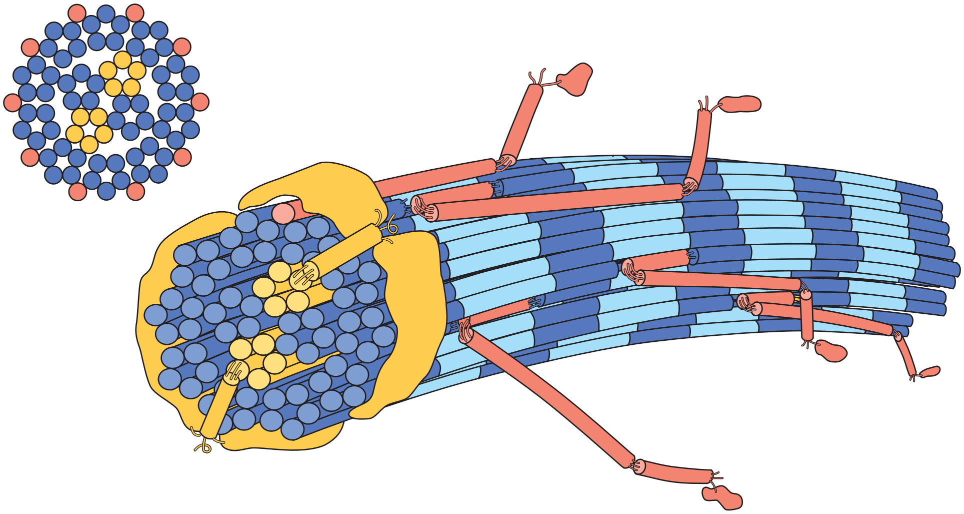

FIGURE 4-56 Schematic representation of the 10 + 4 structure of a type II collagen microfibril, two pairs of collagen type XI forming a central core surrounded by 10 type II microfibrils. Vitreous collagen has a very similar arrangement of type XI/IX/II assembly with fewer cross-links. (From Kadler et al., 2008.)

(see Ch. 7, pp. 388–389). However, as indicated above, depolymerization or even loss of hyaluronan does not of itself destroy the vitreous gel structure.

Vitreous cells

The vitreous contains a single monolayer of cells (hyalocytes), which line the adult vitreous cortex and are responsible for production of hyaluronan in the gel. However, there is no regeneration of collagen in the vitreous and thus there is no reconstitution of the gel after syneresis.

Hyalocytes are of two types: fibrocyte-like and macrophage-like. The latter are equivalent to the resident myeloid cells present in most non-CNS tissues. The role of hyalocytes in health and disease remains obscure. However, they do increase in number with age and have been implicated in macular hole formation (see Ch. 9, p. 513).

PHYSICOCHEMICAL PROPERTIES OF THE VITREOUS GEL

The viscoelasticity of the vitreous protects the retina during eye movement and deformations of the globe

The vitreous gel is non-compressible but highly viscoelastic. Thus it responds to deformations of the globe by altering its shape to comply with external forces, but permits rapid restoration of global architecture. In this respect it behaves as a shock absorber, similar to synovial fluid, which also has a very high content of hyaluronan. These properties of the vitreous are the result of its matrix structure, particularly its content of high molecular weight hyaluronan. This molecule has a very large hydrodynamic volume and at the concentrations present in the human vitreous completely fills the interfibrillar spaces. Deformability of the vitreous has mostly been measured experimentally *in vitro*. However, recent MRI imaging has allowed *in vivo* measurements to be made which reveal that even on eye movement there is considerable deformability of the vitreous gel, which may have relevance for traction events on the retina and risk of retinal detachment (Fig. 4-57). With age, this shock-absorbing property declines as syneresis takes place (see p. 242).

Vitreous retards bulk flow of fluid and diffusion of small molecules

The flow of fluid through solutions of GAGs is variably retarded, depending on the nature and molecular weight of the GAG, and is greatest with hyaluronan. Flow of aqueous from the posterior chamber towards the retina is therefore slower in young eyes with formed vitreous gel than in older eyes where the vitreous gel has undergone liquefaction. In addition,

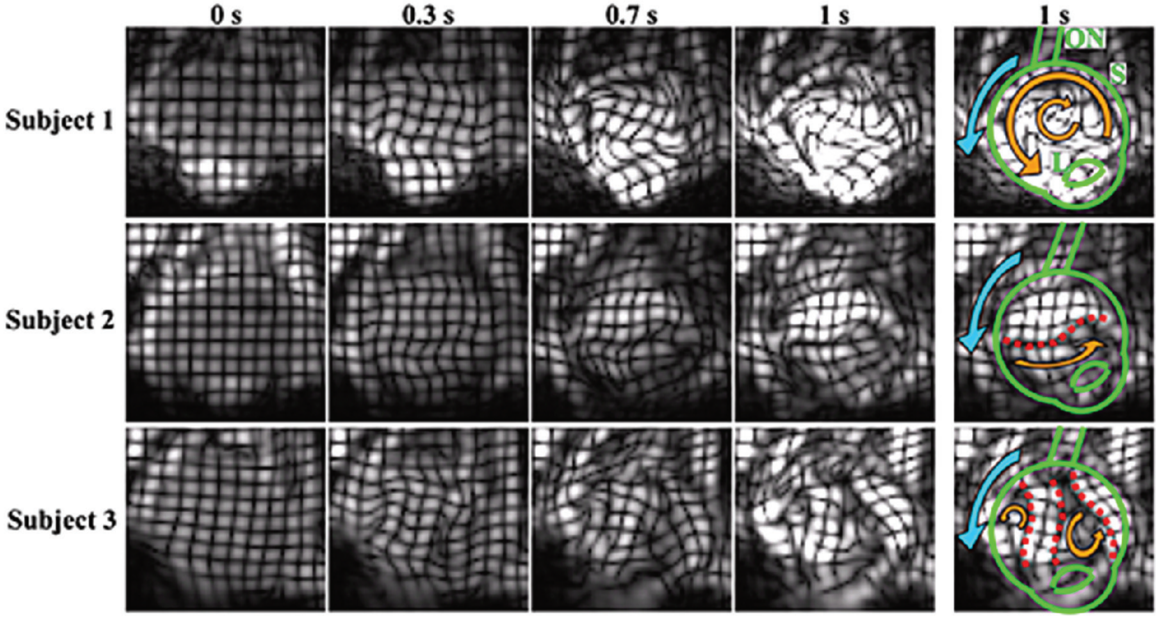

FIGURE 4-57 The technique of complementary spatial modulation of magnetization was used to show the deformation of the vitreous gel during adduction movement of the eye. These changes are shown in frames 1, 5, 10 and 15 and are overlaid with sketches demonstrating the deformation in frames 5, 10 and 15. Differences in the deformations were observed for individual subjects as shown with subject one showing homogeneous viscoelasticity while strong 'whirling' movements were observed in subject 3. (From Piccirelli et al., 2012.)

diffusion of small molecules such as glucose is retarded by the hyaluronan in the vitreous. Electrolytes are more affected in their transvitreal transport by electrostatic interactions with hyaluronan, which is a polymeric polyelectrolyte (see p. 187).

The vitreous is important for maintenance of lens clarity (sooner or later the lens develops cataract after vitrectomy) and this has been attributed to the greater access of oxygen and thus the increased risk of oxidation events due to reactive oxygen species (Fig. 4-58). Transport of fluid and electrolytes in a posterior direction across the retina is an important mechanism in the process of retinal apposition to the RPE (see next section).

The retina

THE NEURAL RETINA IS HIGHLY ORGANIZED IN LAYERS

The retina has two components, the neural retina and the RPE (see Ch. 1, p. 38).

Metabolic function in the retina

The retina's metabolism correlates with its blood supply: the outer retina, comprising the photoreceptors and the RPE, has a high metabolic activity and receives most of its blood supply from the choroid, while the metabolism of the inner retina is supplied by the retinal circulation and is much less demanding on high energy supplies.

Glucose metabolism in the retina

Despite having the highest rate of aerobic glucose consumption of any tissue, a large proportion of the glucose utilized in the retina is converted to lactate. Lactic acid production, oxygen utilization and glucose consumption are also highest in the presence of CO_2/bicarbonate buffering systems, suggesting a role for carbonic anhydrases in the retina. Most of the glucose utilization in the retina is taken by the photoreceptors (>80%). Sodium-coupled moncoarboxylate transporters also occur in the retina, in astrocytes and Müller cells, where amongst other duties, they function in

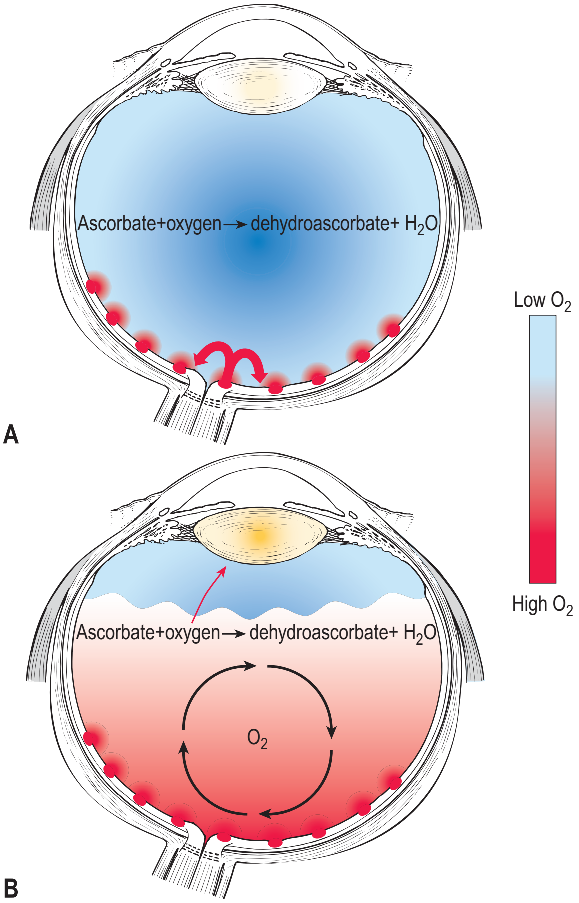
Ascorbate+oxygen→ dehydroascorbate+ H_2O
A
Low O_2
High O_2
Ascorbate+oxygen→ dehydroascorbate+ H_2O
O_2
B

FIGURE 4-58 The distribution of oxygen (**A**) in the normal eye and (**B**) after degeneration or removal of the vitreous body. Oxygen normally diffuses from the retinal vessels into the vitreous but most of the oxygen has been consumed by retinal tissue and the concentration declines towards the central region of the gel. In the degenerate vitreous or the vitrectomized eye, more oxygen enters the fluid-filled vitreous cavity and combines with ascorbate as well as diffusing to the lens where it is relatively toxic and can promote nuclear cataract through oxidative damage. (From Beebe et al., 2011.)

'lactate shuttling' between glycolytic and oxidative pathways of glucose metabolism.

The retina can also metabolize other substrates for (ATP) energy stores such as glutamate, glutamic acid, malate and succinate. Retinal glucose is used to produce glutathione via the pentose phosphate pathway, which can be upregulated under conditions of oxidative stress. However, NADPH for glutathione stores is also produced by other non-pentose-dependent systems involving malate and isocitrate. Glucose, glutathione and oxygen are all required for generation of electrical activity, including the ATPase-dependent 'dark currents' (see Ch. 5, Box 5-6, p. 288) in the retina.

The retina is regarded as an insulin-independent tissue, i.e. glucose enters retinal cells by transport mechanisms that are regulated directly by the extracellular concentration of glucose rather than indirectly by insulin. Glucose transport occurs by facilitated diffusion via GLUT 1 and GLUT 3 transporter proteins, similar to glucose transport in the brain. GLUT 1 and GLUT 3 are present in endothelial cells of the blood–retinal barrier, where much of the transport occurs. The insulin-linked GLUT 2 and recently GLUT 4 have also been identified in retinal tissue. Photoreceptors respond to insulin via a retina-specific insulin receptor, which is similar to brain insulin receptor in that it exists in a 'tonic' state of activity and does not change in conditions of fasting or excess glucose. The insulin receptor phosphorylates cyclic GMP-gated channels, which is central to phototransduction (see below).

Most of the retinal glucose metabolism is dealt with by retinal neurones. However, Müller cells are also likely to be involved in facilitating glucose metabolism in the retina, as energy stores, as well as 'managing' lactate production, since they contain high levels of glycogen, especially in species that lack a retinal blood supply. In addition, lactate released by Müller cells can be metabolized by photoreceptors. Photoreceptors in turn release glutamate, which is taken up and metabolized by Müller cells ('lactate shuttling').

Protein metabolism in the retina

Many of the neurotransmitters required for normal retinal cell function occur as free amino acids in the retina (see p. 265). Most of them are generated during glucose metabolism in the citric acid cycle; in addition, taurine, which is not a neurotransmitter but appears to be essential for, and is avidly taken up by, photoreceptor cells, is the most abundant amino acid in the retina although its receptor/transporter has not been identified. Taurine also regulates voltage-gated channels in ganglion cells. Taurine is not incorporated into proteins but has other functions. Interestingly, taurine may have a protective role for the retina in diabetes. Glutamate is neurotoxic and is converted by the retina to glutamine by glutamine transferase (synthase) localized to the Müller cells. Transport of amino acids is now known to require specific amino acid transporters, for which there are several ‘systems’ (systems y, b, B, b° and more). Some are linked to Na$^+$ transport while others are independent of Na$^+$ intake. Rapid uptake of amino acids is essential not only for the supply of neurotransmitters but also for arginine uptake, which is necessary for the synthesis of nitric oxide, the major regulator of endothelial cell function.

Protein synthesis, as studied by methods such as leucine incorporation, is most active in the photoreceptors during such processes as photoreceptor renewal (see Box 4-17).

In addition to retina-specific proteins (see below), several proteins common to many tissues are present.

BOX 4-17 PHOTORECEPTOR RENEWAL

Photoreceptor renewal occurs by synthesis of new protein-rich membrane at the outer limiting membrane, with shedding and phagocytosis of the outer segment tips by the RPE (**A**). Disks are formed by the evagination of the plasma membrane ⊛ at the junction between the inner and outer segments, while rhodopsin and other proteins synthesized in the ER and Golgi apparatus are transported in vesicles for fusion to the newly formed plasma membrane ⊛. Photoreceptor renewal is similar in rods and cones; phagocytosis of the receptor tips occurs in a diurnal manner. The mechanism of phagocytosis is unclear but involves a membrane glycoprotein CD36, which is also involved in the uptake of apoptotic neutrophils and oxidized low-density lipoprotein by haematopoietic cells such as macrophages. In addition, rod outer segment (ROS) phagocytosis is associated with induction of cyclo-oxygenase-2 (COX-2), an enzyme involved in prostaglandin synthesis. Recent studies have also shown the importance of the cilia in the functioning of the photoreceptor, since genetic mutations in a number of genes controlling cilia proteins produce forms of retinal degeneration now known as ciliopathies (**B**).

There are four distinct compartments in photoreceptor primary cilia, indicating known proteins that define their respective extent, are (1) distal cilium or axoneme (Axo; green); (2) connecting cilium/transition zone (CC/TZ; orange); (3) basal body (BB; purple); and (4) periciliary complex or ciliary pocket (PCC/CP; red). These compartments serve discrete functions in the cilium (see inset in Figure).

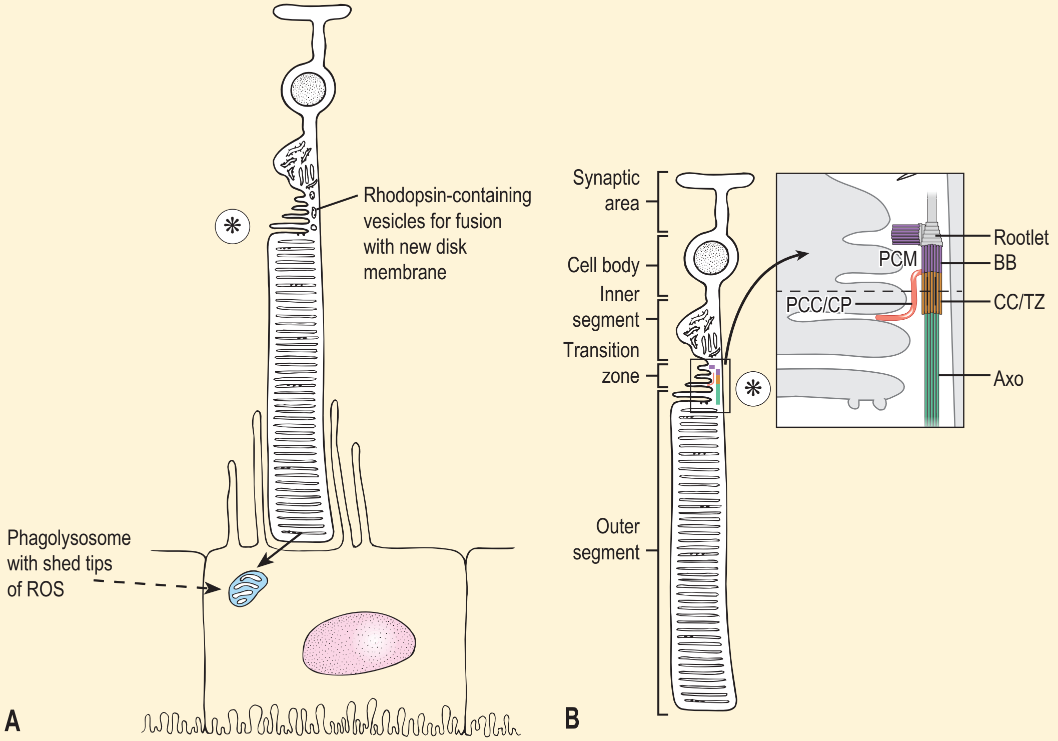
⊛
Rhodopsin-containing vesicles for fusion with new disk membrane
Phagolysosome with shed tips of ROS
A
Synaptic area
Cell body
Inner segment
Transition zone
Outer segment
⊛
PCM
PCC/CP
Rootlet
BB
CC/TZ
Axo
B

(Part B reproduced from Rachel and Swaroop, 2012.)

For instance, laminin is present in vessel structures, fibronectin in the interphotoreceptor space, and matrix proteoglycans are widely distributed throughout the retina. Tenascin-C is present in the extracellular matrix and is thought to play a role in preventing myelination of retinal neurones. However, tenascin-C knockout mice show no myelination of their retina. Several growth factors are present in the retina, such as insulin-like growth factor 1 (IGF-1) and acidic and basic fibroblast growth factor (bFGF). Basic FGF is not only present in basement membranes of vessels but is also distributed in such regions as the photoreceptor layer, where it may play a trophic role in outer segment renewal.

The retina contains a high content of lipid (20%)

The predominant lipids in the retina are the phospholipids, phosphatidylcholine and phosphatidylethanolamine (in total around 80%). There is also a high content of polyunsaturated fatty acids in the retina, especially in the outer segments, some containing more than six double bonds (known as 'supraenes'). This renders the retina particularly susceptible to oxidative damage (see p. 193). Disk membranes of photoreceptors are rich in phosphatidylethanolamine while the plasma membrane has higher levels of cholesterol. This is relevant to the activity of rhodopsin because it is inhibited in the presence of the sterol. Thus the older disk membranes at the tip of the photoreceptor (see Box 4-17) have lower levels of cholesterol than the fresh, newly formed disks, allowing easier activation of rhodopsin at the photoreceptor tip.

Lipid metabolism is varied and complex in the retina; thus, in addition to synthetic activities in microsomes, exchange of bases between different lipid species occurs, while frequent acylation–deacylation reactions also occur. Lipids are continually undergoing degradation via phospholipases and modifications including decarboxylation and methylation by the appropriate enzymes.

BLOOD FLOW IN THE RETINA

Blood flow is determined by a balance between the perfusion pressure and the resistance in the blood vessels. Blood flow in the retina is autoregulated. In general, this means that the retinal vessel calibre varies with the cardiac output to ensure that blood flow is kept constant, although it has also been interpreted as meaning that the blood flow is varied according to the nutritional demands of the tissue. In the adult, the blood flow in the retina is maintained constant over a range of ocular perfusion pressures from 45 to 145 mmHg and this is mostly achieved by changes in blood vessel diameter, as shown in mice deficient in the endothelial nitrous oxide synthase.

Retinal blood flow comprises about 5% of the total flow to the eye, the majority passing through the choroid.

The blood–retinal barrier regulates the passage of molecules into the retina

The blood–retinal barrier is maintained by tight junctions that exist between the endothelial cells of the retinal vessels and similar tight junctions in the RPE (see pp. 254–258). Thus, the retinal vessels and the RPE are impermeable to the passage of molecules greater than 20 000–30 000 Da, and small molecules such as glucose, amino acids and ascorbate are transported by facilitated diffusion (mostly GLUT 1 at both sites, but also GLUT 3 in the retinal vessels). Similar transporters are present for amino acids including arginine, and both classes of transporters are present in smooth muscle cells and pericytes ensuring rapid transport and uptake of these moieties into the retinal tissue. Arginine for instance is central to the production of NO by endothelial nitric oxide synthase (eNOS), a critical regulatory element in retinal blood flow (see above). Endothelial cells principally metabolize glucose by anaerobic glycolysis.

Although the retina is considered an insulin-independent tissue, the endothelial cells and pericytes possess high-affinity receptors for insulin, IGF-1 and IGF-2. The role of these receptors in the regulation of glucose transport in the retina is not clear because their effect is delayed for some hours, suggesting that they stimulate protein synthesis and the production of new transporters rather than recruit existing transporters, as occurs in insulin-dependent tissues such as muscle.

Retinal blood flow may be partly under autonomic control

Retinal vessels possess all four types of high-affinity adrenergic receptors, although in low numbers. In addition there is indirect evidence that, despite well-recognized mechanisms of autoregulation, some degree of autonomic control exists in humans. How this might occur in the absence of nerve fibres is unclear, but it is possible that autonomic nerves in the choroid supply vessels at the optic nerve head. The retinal vascular bed may be one of the few systems to lack perivascular mast cells, an important response element for catecholamines.

Retinal blood flow is responsive to hyperoxia (vasoconstriction) and hypercapnia (vasodilatation), the latter via the prostaglandins PGD_2 and PGE_2. Other mediators of changes in vessel diameter include the eicosanoid, PGI_2, endothelin and nitric oxide. PGI_2 and endothelin have been detected in retinal vessels and presumably are released under appropriate conditions. Nitric oxide is released via endothelial nitric oxide synthase (see Ch. 7, p. 385) and provides basal levels of vascular dilatation. Retinal illumination induces release of nitric oxide but autoregulation in retinal vessels is not significantly affected by nitric oxide. Contractile activity in the retinal vessels is attributed to the pericytes, whose role therefore may be to regulate blood flow. Their early loss in diabetes may account for the increase in blood flow that occurs in the retina in diabetes, and may contribute to the development of retinopathy.

PHOTORECEPTORS

Photoreceptors are specialized for reception of visual stimuli and have unique characteristics (see Ch. 1).

Metabolism and turnover

Photoreceptors are some of the most highly metabolic cells in the body, utilizing glucose both aerobically and anaerobically. Photoreceptor outer segments lie in apposition to the RPE in the interphotoreceptor matrix between the apical microvilli of the RPE cell. Extensive protein and lipid synthesis ensures a continuous turnover of new outer segment membrane at the junction with the inner segment; the tips of the outer segment containing the ‘oldest’ disks are phagocytosed as small packets of about 200 disks by the RPE cell, a process that occurs in a diurnal manner just after light onset (Fig. 4-59). Three RPE cell surface receptors have been identified which are involved in the phagocytosis of outer segments, namely CD36, the tyrosine kinase MerTK and the integrin $\alpha_v\beta_5$. Loss of these proteins may lead to retinal degenerations (MerTK: human retinitis pigmentosa and rat RCS disease; $\alpha_v\beta_5$:RPE lipofuscin accumulation). Diurnal regulation of phagocytosis is under control of the secreted glycoprotein milk fat globulin-EGF8. Activation of MerTK may be under autocrine regulation of its naturel ligand, Gas6, which is also expressed in RPE cells or via ‘tubby’ proteins which are expressed in photoreceptors and generate an ‘eat me’ signal similar to that which occurs in apoptotic cells.

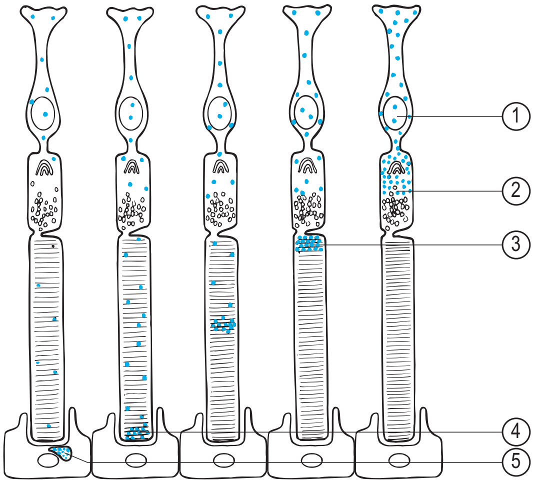

FIGURE 4-59 Photoreceptor renewal and points where damage may occur: during (1) transcription; (2) post-translational modification; (3) incorporation into disk membrane; (4) disk shedding; (5) phagocytosis by RPE.

Complete renewal of the rod outer segment takes about 9–10 days. In contrast, although cone outer segments are phagocytosed in a similar manner, the process appears to be more random and occurs when darkness occurs; cone membranes and their integral proteins are much more stable and long-lasting.

Insertion of rhodopsin into the disk plasma membrane follows a well-defined pathway from the inner segment RER to the outer segment plasma membrane infolding (see Box 4-18). Glycosylation of rhodopsin takes place through combined co-translational and post-translational events in a classic lipid-carrier mechanism using Dol-P-P-GlcNAc (Fig. 4-60), which can be inhibited by tunicamycin. Acylation of rhodopsin also occurs in the membrane via palmitic acid.

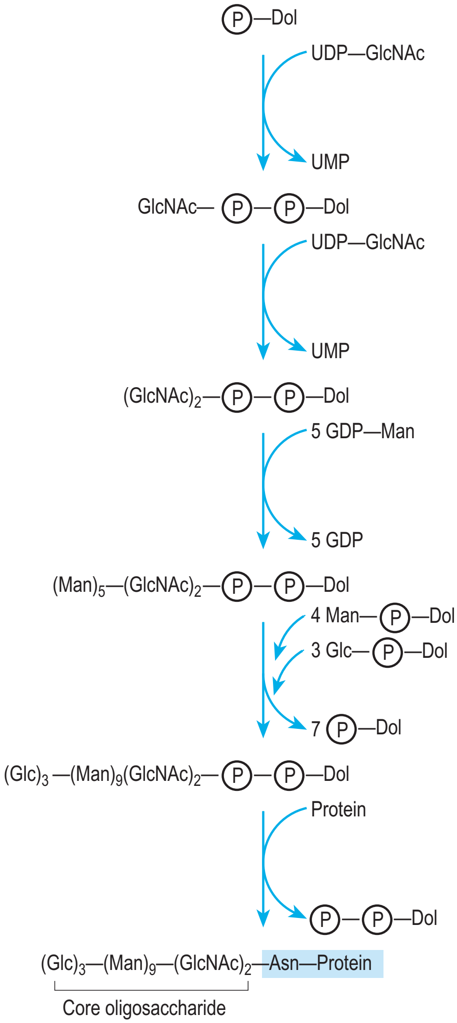

FIGURE 4-60 Glycosylation and acetylation of rhodopsin.

Lipids in photoreceptors are replaced both by membrane turnover and by molecular replacement. The abundant stores of phosphatidylcholine are synthesized from large intracellular pools of free choline and phosphorylation by ATP; the activated choline then reacts with 1,2-diacylglycerol to form phosphatidylcholine. Similar mechanisms operate for the synthesis of phosphatidylserine and phosphatidylethanolamine; all three phospholipids are synthesized in the RER but are transported to the newly forming outer segment membrane by different mechanisms. The role of the high concentrations of docosahexanate phospholipids in photoreceptors has recently been identified as enhancing the function of many of the proteins involved in the visual transduction cascade, such as GDP-bound transducin (see below).

Several chemical reactions are associated with disk shedding, although the specific stimulus and its site of origin (i.e. the photoreceptor or the RPE cell) are not known. The circadian light–dark rhythm of shedding is under the control of melatonin- and/or 5-methoxytryptophol-synthesizing enzymes taht are predominantly active in photoreceptors (Fig. 4-61). Melatonin synthesized and released in the retina does not enter the circulation but acts locally. The retina has widely distributed melatonin receptors, both the MT1 and MT2 (Mel1a and Mel1b) receptors being found in the photoreceptors and in inner retinal amacrine cells. A third receptor (Mel1c) has been detected in non-human species. In contrast to melatonin, dopamine appears to counter-regulate retinal circadian rhythms by suppressing melatonin synthesis through the D2/4 receptor, while melatonin works on the MT2 receptor in the reverse direction. Dopamine may also influence a second circadian regulated system involving melanopsin, a photopigment present in inner retinal ganglion cells, and possibly does so through activating the MT2 receptor (see p. 267).

The mammalian retina also appears to possess an autonomous melatonin-responsive circadian oscillator independent of central (suprachiasmatic nucleus) control and rhodopsin and cone opsin synthesis are in phase with this rhythmic oscillation. However, several other compounds have an effect on disk shedding, such as excitatory amino acids, glutamine and aspartate, while certain divalent ions are also essential (Ca^{2+}, Mn^{2+}). Recent studies in mice lacking the D4 dopamine receptor have shown that this membrane protein may regulate several of the controlling mediators of disk shedding apart from melatonin, such as a light-sensitive pool of cAMP.

Phosphoinositide metabolism is considerably greater than phosphatidylcholine or phosphatidylethanolamine metabolism in the photoreceptor but its precise role in phototransduction is unclear (see p. 258). Cytidine triphosphate is also a product of light transduction and is linked to phosphatidylinositol formation.

BOX 4-18 RHODOPSIN SYNTHESIS

Insertion of rhodopsin into the outer segment plasma membrane is facilitated by the lack of a signal peptide, which permits integration of opsin into the lipid bilayer by co-translational coupling of glycosylation and asymmetric insertion via specific insertion sequences in the protein.

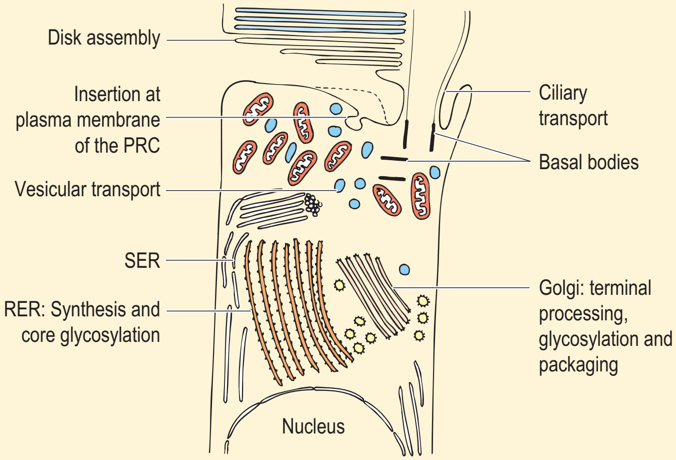
Disk assembly
Insertion at plasma membrane of the PRC
Vesicular transport
SER
RER: Synthesis and core glycosylation
Ciliary transport
Basal bodies
Golgi: terminal processing, glycosylation and packaging
Nucleus

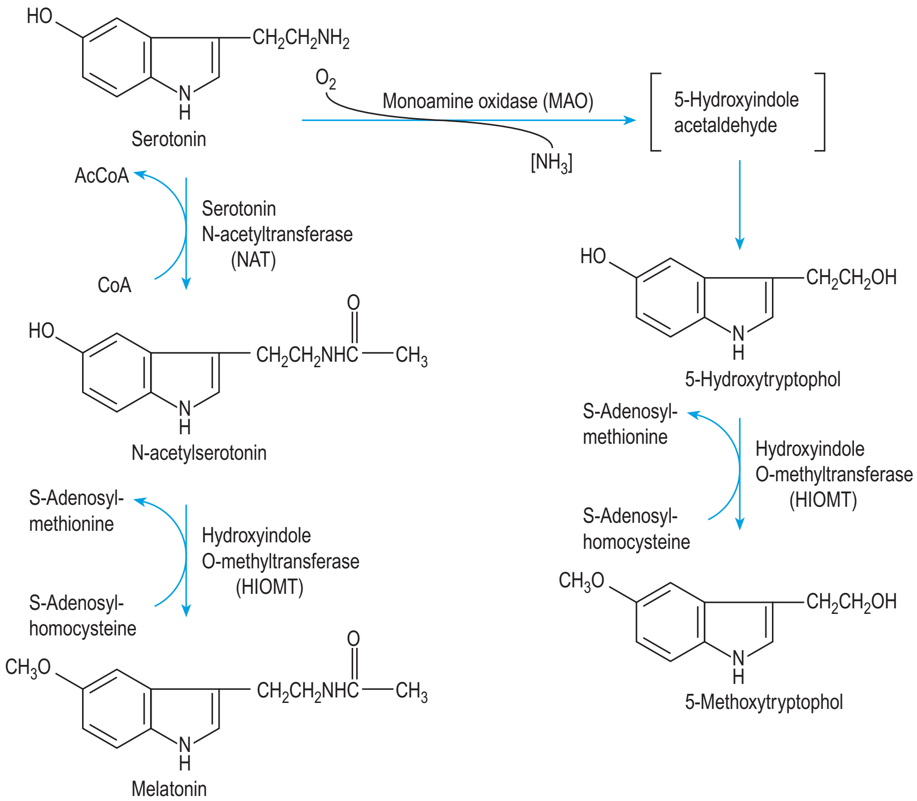
HO
CH2CH2NH2
N
H
Serotonin
O2
Monoamine oxidase (MAO)
[NH3]
5-Hydroxyindole acetaldehyde
AcCoA
Serotonin N-acetyltransferase (NAT)
CoA
HO
O
CH2CH2NHC—CH3
N
H
N-acetylserotonin
S-Adenosyl-methionine
Hydroxyindole O-methyltransferase (HIOMT)
S-Adenosyl-homocysteine
CH3O
O
CH2CH2NHC—CH3
N
H
Melatonin
HO
CH2CH2OH
N
H
5-Hydroxytryptophol
S-Adenosyl-methionine
Hydroxyindole O-methyltransferase (HIOMT)
S-Adenosyl-homocysteine
CH3O
CH2CH2OH
N
H
5-Methoxytryptophol

FIGURE 4-61 Regulation of disk shedding. Disk shedding may be under the control of melatonin-synthesizing enzymes such as serotonin N-acetyltransferase (NAT), which is present in the pineal gland and retina.

Photoreceptor cell-specific proteins

The highly differentiated visual cell contains many unique proteins, including integral membrane proteins, membrane-associated proteins and cytosolic proteins, of which more than 30 alone are involved in the phototransduction cascade. Some of these are shown in Table 4-8. However, there are many more proteins, and many of them have been evolutionarily conserved. For instance, 57 genes and their respective proteins were identified and selected in a recent study for analysis and categorized in three sets: photoreceptor cell type-specific, process-specific (e.g. phototransduction itself, the retinoid cycle and developmental regulation) and a specific functional role (e.g. opsin, G protein, etc.).

The protein composition of the disk membrane is predominantly rhodopsin (90%), whereas the plasma membrane has a wider range of cell-specific proteins and less rhodopsin (50%). Rhodopsin is the visual receptor protein; proteins such as peripherin/RDS (Fig. 4-62) and the spectrin-like protein Rom-1 have structural functions in the maintenance of the photoreceptor shape, similar to those of spectrin/ankyrin-like proteins in red cells. Both are members of the tetraspanin family of proteins, and may also have a role in membrane fusion important for rod disk generation.

Remarkably different mutations in the RDS protein (Fig. 4-63) produce different clinical types of retinal degeneration such as autosomal dominant retinitis

TABLE 4-8 **Photoreceptor proteins**

Integral membrane proteins	Peripheral/cytosolic proteins
Rhodopsin	Arrestin (48 kDa protein, S antigen)
cGMP channel	
$Na^+/Ca^{2+}-K^+$ exchanger	Transducin
Glucose transporter	Phosphodiesterase
Guanylate cyclase	Phosducin
Peripherin/RDS	Rhodopsin kinase
Rom-1	Guanylate cyclase
ABCR/rim protein	
Retinal dehydrogenase	

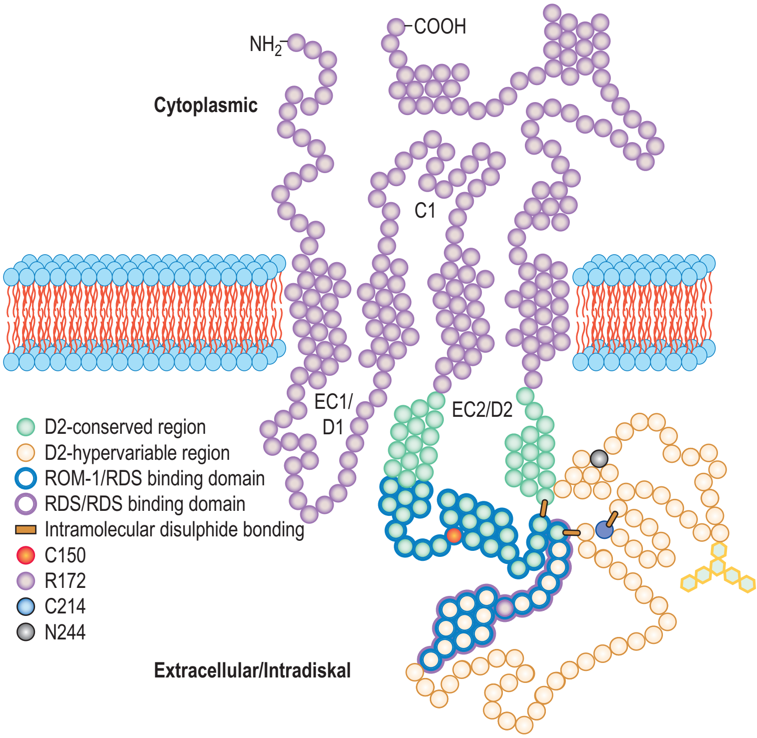

FIGURE 4-62 Structure of the RDS/peripherin protein. Structural features of the RDS protein. RDS contains four transmembrane domains, two intradiskal loops (D1 and D2), cytoplasmic N- and C-termini, and a small cytoplasmic loop (C1). (From Chakraborty et al., 2013.)

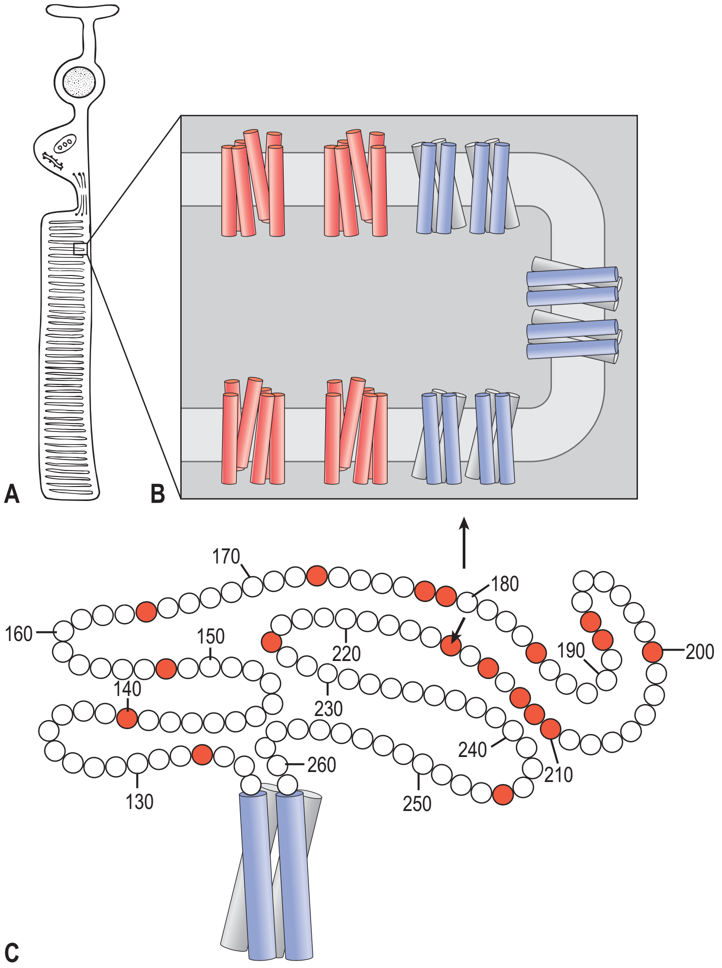

FIGURE 4-63 The rim region of photoreceptor disk shown in **A** are expanded in **B** to reveal the 4-span (tetraspanin) arrangement of the RDS/peripherin molecule with the intradiscal portion of the molecule shown in **C**. Specific amino acids locating mutations responsible for forms of retinitis pigmentosa are identified in red. (From Vos et al., 2010.)

pigmentosa and macular dystrophy (see Ch. 9, p. 514). The Na$^+$/Ca^{2+} exchanger facilitates ionic transport during phototransduction and maintains Ca^{2+} homeostasis. In addition, the spectrin-like protein Rom-1 appears to be linked to a second ion-channel protein of 67 kDa. Stacking of rod and cone disks is a central organizational requirement for ordered photoreceptor function and a set of glutamic acid-rich proteins (GARPs) interact with RDS/Rom to both structure stacking of disks and participate in ion-channel functions.

In contrast to the integral membrane proteins, many of the peripheral membrane and cytosolic proteins are not exclusive to the photoreceptor and are intimately involved in the light amplification cascade itself (see below). Transducin is a member of the family of G proteins and is composed of three chains (α, β, γ), which dissociate during the light response.

Certain other proteins are located in the photoreceptors as well as in other areas of the retina; these include the IGF-1 receptor, IGF-1-binding protein, and FGF, both of which have been implicated in the induction of proliferative diabetic retinopathy (see Ch. 9, p. 504). In addition, there is a glucose transporter (GLUT 1) for control of intracellular glucose levels.

Melanopsin comes of age

Because it was recognized that certain light-sensitive processes, such as those entrained in a circadian fashion, were likely to reside at a different site from

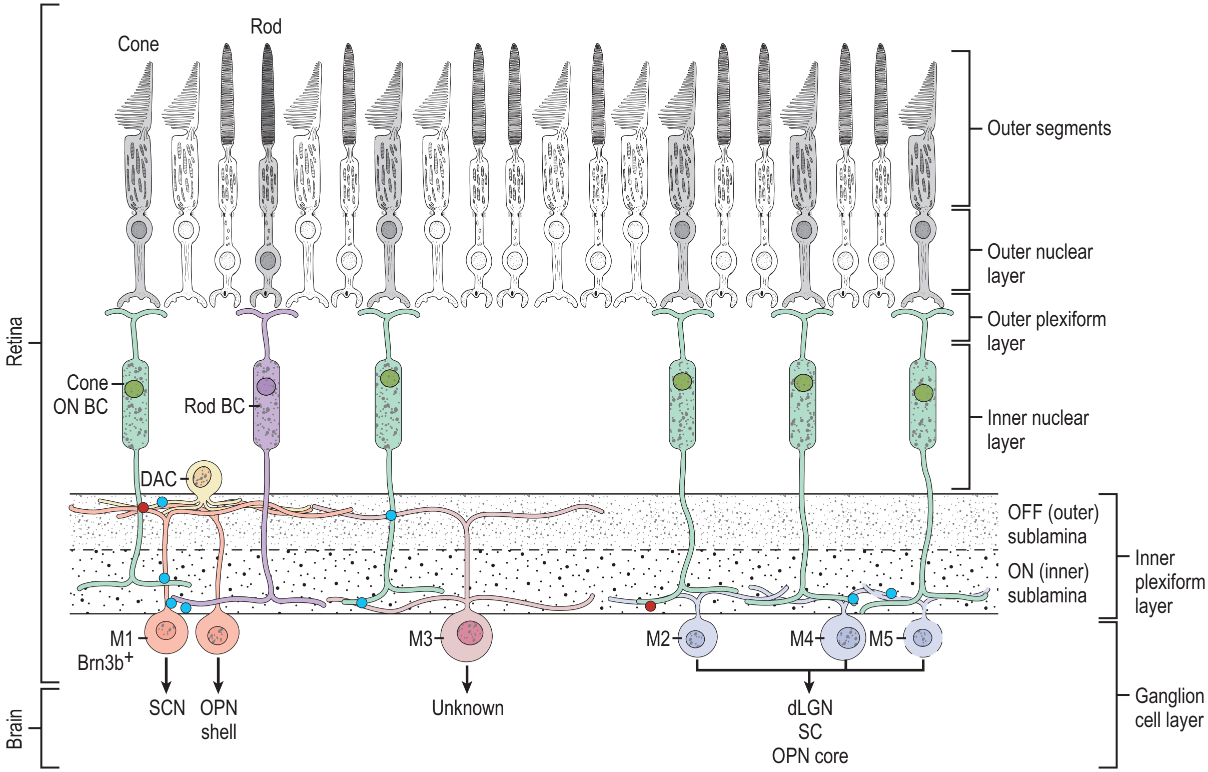
Retina
Brain
Cone
Rod
Outer segments
Outer nuclear layer
Outer plexiform layer
Inner nuclear layer
Cone ON BC
Rod BC
DAC
OFF (outer) sublamina
ON (inner) sublamina
Inner plexiform layer
M1
Brn3b+
M3
M2
M4
M5
Ganglion cell layer
SCN
OPN shell
Unknown
dLGN
SC
OPN core

FIGURE 4-64 Five distinct morphological types of melanopsin-expressing ipRGC exist, as shown here, and project to different parts of the brain such as the lateral geniculate nucleus (LGN), the suprachiasmatic nucleus (SCN) and the olivary pretectal nucleus (OPN) (see Chs 1 and 5). (From Schmidt et al., 2011.)

visual photoreception, a search for other photoreceptors was initiated. Early studies indicated that the likely site was intraocular and probably retinal and in due course a photopigment, melanopsin, was detected in specialized retinal ganglion cells with large receptive fields (see Ch. 5, p. 306). These are now known as intrinsically photosensitive retinal ganglion cells (ipRGCs) of which there are five types with different functions, including entrainment of the circadian clock, sleep and pupillary responses to light (Fig. 4-64).

Photoreceptors are easily damaged but readily regenerate

The extensive metabolism and rapid turnover of photoreceptor outer segments render them highly susceptible to damage. Damage may occur at any level from synthesis of new membrane to phagocytosis of the outer segment tips (see Fig. 4-59).

Photoreceptors also degenerate when they are separated from the RPE, as in retinal detachment, or when there is a subretinal collection of fluid; photoreceptors are lost in inflammatory and metabolic retinal diseases and are probably highly susceptible to free radical damage. Specific photoreceptor loss occurs in various forms of retinal degeneration: inherited forms of these disorders are collectively known as retinitis pigmentosa and for many of them mutations in retinal specific proteins have been described (see p. 261 and Ch. 3, p. 151). To date, more than 150 gene mutations have been associated with various retinal degenerations.

Damage to the retina also occurs as part of normal physiology via both light- and oxygen-induced mechanisms. Around one-third of the mutations associated with retinal degenerations occur in proteins associated with the photoreceptor cilium, the connecting structure between the inner and outer segments (see Ch. 1, p. 41). These are known as ciliopathies, in which the Wnt signalling pathway is implicated (see eFig. 4-5). The *Wnt/Drosophila wingless* family of genes express highly conserved secreted glycoproteins which spread across tissues to reach their targets where they interact with the products of two sets of genes (*Frizzled* and genes for low-density lipoprotein-related proteins, LRPs). In the fruit fly, *Drosophila*, a species that is frequently used to study molecular genetics, Wnt signalling has been shown to be central to normal photoreceptor development as well as to central nervous system development, probably through regulation of the expression of pro-survival factors such as Dickkopf (*Dkk3*) in cells. In fact this mechanism highlights the fact that photoreceptor damage in retinal degenerations and other conditions is mediated via apoptosis, which may occur by both caspase-dependent and caspase-independent mechanisms. Light-induced retinal damage varies with the intensity, wavelength, duration, cyclical nature and previous antioxidant status.

For instance, it has been shown that cyclical light is less injurious to the retina than constant illumination of equivalent power, an effect associated with higher levels of ascorbic acid, vitamin E and glutathione, and lower levels of 22:6(n-3) fatty acids in the tissue. RPE65, the retinoid cycle enzyme and an intrinsic retinal pigment epithelium protein with a critical amino acid residue at position 450, regulates the rate of rhodopsin synthesis through the action of two proteins, elongation of very long fatty acids-like 1 (ELOVL1) and fatty acid transport protein 4 (FTAP4), both of which have very long fatty acid synthase activity. This demonstrates the complex biochemical machinery which is in place to protect photoreceptors from light damage. The death signal for photoreceptors involves induction of the pro-apoptotic transcription factor AP-1 such that inhibition of regeneration of rhodopsin or suppression of AP-1 can prevent light-induced damage to photoreceptors (Fig. 4-65).

Extensive antioxidant enzyme systems are present in the retina and, under conditions of stress such as light injury, levels of glutathione peroxidase and glutathione *S*-transferase are markedly raised. In

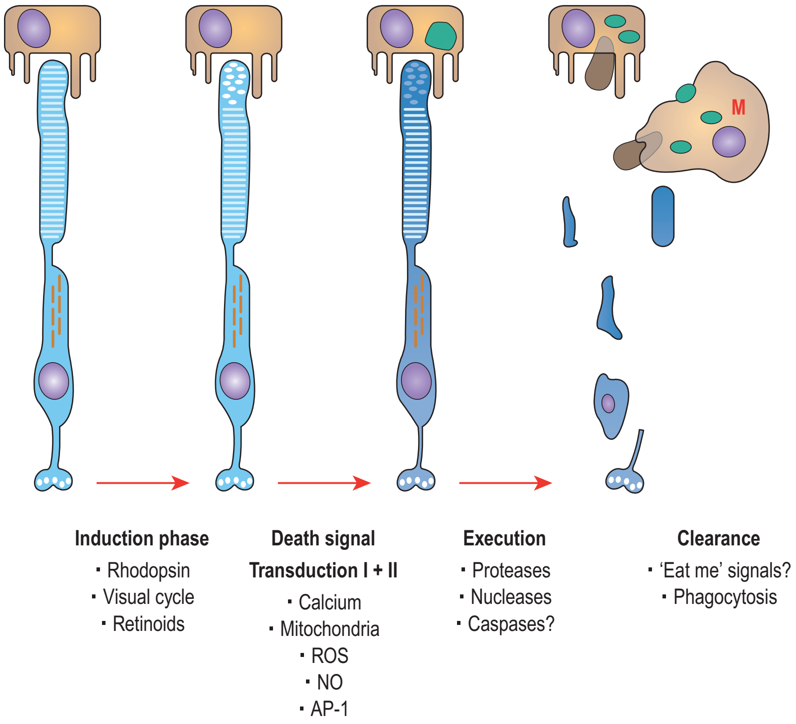

FIGURE 4-65 Schematic drawing of major components which contribute to light-induced photoreceptor apoptosis. Induction phase: rhodopsin is essential in this process. Death signal transduction: requires Ca^{2+} and transcription factor AP-1. Execution: mechanism is not clear, and may require caspases. Clearance: signals inducing phagocytosis by RPE and, in acute light damage, macrophages, are not known. (From Wenzel et al., 2005, with permission from Elsevier.)

addition, photoreceptor expression of glutathione peroxidase4 (Gpx4) is absolutely required for normal photoreceptor development. Superoxide dismutase (the Cu^{2+}/Zn^{2+} form) is also present in significant amounts in several layers of the retina, but levels of catalase are low. Despite this, treatment with either of these protective agents as well as other antioxidant nutrients, *N*-acetylcysteine and thioredoxin can prevent retinal degeneration in experimental models. Melatonin also has important antioxidant activity. In addition, vitamin E is of major importance in reducing photoreceptor damage, mainly by its role in inhibiting lipid peroxidation at different stages in the process of free radical damage (see Box 4-16).

Iron in the form of Fe^{2+} ions can interact with hydrogen peroxide to produce high levels of hydroxyl radicals (see Box 4-8, p. 194) and cause retinal injury. Local release of iron may occur in diseases involving retinal haemorrhage and when intraocular foreign bodies are present. Ceruloplasmin, which contains 95% of the copper in the serum, mediates iron metabolism via ferroxidase, which converts ferrous to ferric iron before it is delivered to serum transferrin after efflux from the cell. Patients with the recessive disorder leading to altered ceruloplasmin blood levels have severe progressive retinal degeneration. Iron homeostasis is essential to normal retinal function and the five main regulatory proteins are present in the RPE cell (Fig. 4-66).

The macular region of the retina is particularly susceptible to light damage. Interestingly, this region contains additional yellow pigments lutein and zeaxanthin and a further pigment mesozeaxanthin, which may be transported to this region from the blood and whose function is thought to be related to reducing glare from short-wavelength blue light. These pigments, which belong to the family of carotenoids (tetraterpenoids with 40 carbon isoprene backbone) (Fig. 4-67), also have an antioxidant effect, particularly at low levels of oxygenation. They are referred to as xanthophyll macular pigments and can be measured non-invasively in the retina using Raman spectroscopy. Sustained levels of these pigments are considered protective against macular degeneration.

Finally, melanin in the RPE layer is a very effective free radical scavenger, particularly in the reduced form, where it may act as a trap for stray light-induced free radicals emanating from the photoreceptors, but the role of melanin is more likely to be related to protection of the RPE cells themselves because these are terminally differentiated cells. Melanin may act as a antioxidant due to its ability to bind metal ions such as Fe^{2+}, a process well known to occur in fungi. Melanin granules in RPE cells appear to be connected to the lysosomal enzyme system in RPE cells and loss of melanin is associated with age-related macular degeneration.

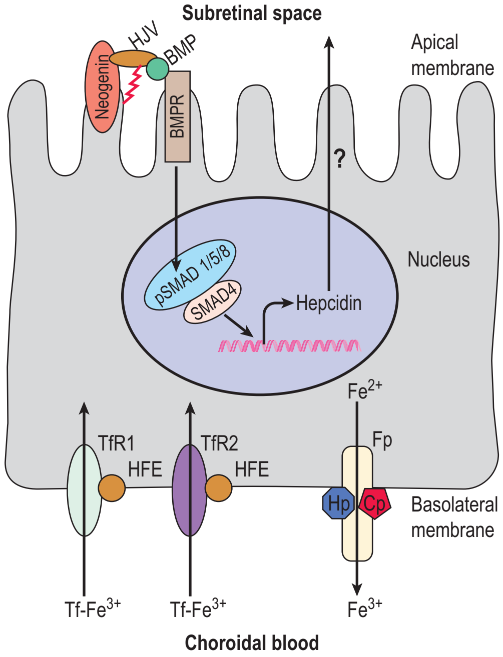

FIGURE 4-66 Expression pattern of haemochromatosis genes in retinal pigment epithelium (RPE). TfR, transferrin receptor; HFE, HLA-like protein involved in iron homeostasis; Fp, ferroprotein; Hp, hephastin; Cp, ceruloplasmin. (From Gnana-Prakasam et al., 2010.)

The interphotoreceptor matrix is the biological glue for retinal adhesion

The interphotoreceptor matrix extends from the outer limiting membrane of the retina (see Ch. 1, p. 41) to the surface of the RPE cell. It is an extremely narrow space, almost a potential space, but contains some unique and physiologically important molecules. These include interphotoreceptor retinol-binding protein (IRBP; accounts for at least 70% of the

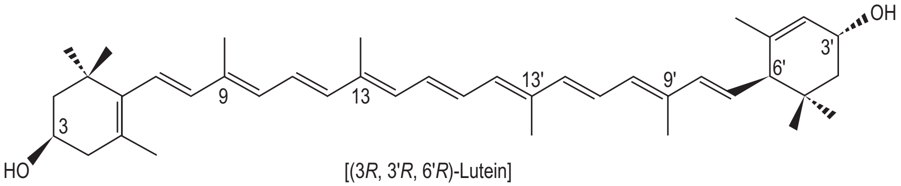

[(3*R*, 3'*R*, 6'*R*)-Lutein]

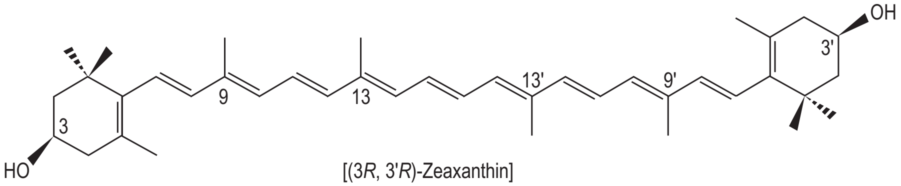

[(3*R*, 3'*R*)-Zeaxanthin]

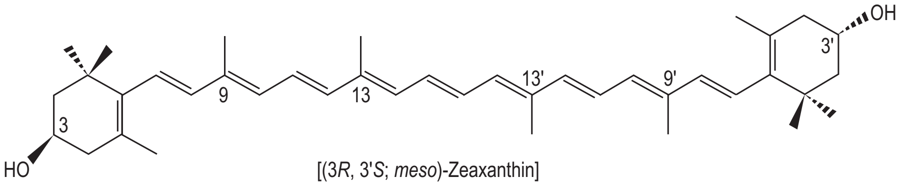

[(3*R*, 3'*S*; *meso*)-Zeaxanthin]

FIGURE 4-67 The chemical structure of macular pigments. (From Abdel-Aal et al., 2013.)

interphotoreceptor matrix protein), which transports retinoids between the RPE cell and the photoreceptor, and several species of proteoglycan, which provide a coating for the photoreceptor outer segment. Interestingly, the 6-sulphated (DeltaDi6S) chondroitin species appears to be more prominent around cones than around rods. IRBP appears to be essential for photoreceptor survival but is not required for the visual cycle. In addition, there are several cell surface proteins that form part of the glycocalyx of the cell but may also play a part in maintaining retinal apposition, which is essentially maintained by the bulk flow of fluid across the retina. These include fibronectin, intercellular adhesion molecule 1 (ICAM-1) (see Ch. 7) and the CD44 antigen, also known as the hyaluronan receptor. There is also a second hyaluronan-binding protein, SPARC (sialo-protein associated with rods and cones), which may also be important in providing a scaffold to the matrix.

The interphotoreceptor matrix of the cone is compartmentally separated from that of the rod by an insoluble matrix sheath containing its own specific proteoglycans, which presumably play a role in the regulation of their different forms of visual excitation. Chondroitin-6-sulphate appears to be the major proteoglycan–GAG in the cone matrix sheath, while a similar but less well-defined sheath around rod outer segments appears to be composed of sialyl conjugates. It has been suggested that these proteoglycan sheaths around both rods and cones are important in retinal–RPE adhesion, for instance via hyaluronon–CD44 interactions on the apical microvilli as well as via SPARC and SPARCAN through hyaluronan-mediated motility receptor (RHAMM) domains.

The interphotoreceptor matrix contains other GAGs, including non-sulphated chondroitin and hyaluronan (about 14% of total). These are present as proteoglycans and are probably mostly synthesized by the RPE cell. There are also a number of proteolytic enzymes and their inhibitors, including matrix metalloproteinases and tissue inhibitor of matrix metalloproteinases (TIMPs) mostly as RPE cell surface bound molecules to mannose-6-phosphate receptors. In addition, αβ crystallins have been identified in the IPM derived from the RPE in an exosome-mediated mechanism.

THE RETINAL PIGMENT EPITHELIUM

The RPE is a pluripotent cell

The retinal pigment epithelium (RPE) is a multifunctional pluripotent cell, which befits its embryological origins. It expresses many of the proteins considered characteristic of other cell types. Thus it possesses several cytokeratins characteristic of epithelial cells, yet also contains vimentin, which is a mesenchymal cell protein. One group of microfilaments, namely actin, myosin, α actinin and vimentin, is organized as a ring (or belt) around the cell and inserts into typical zonulae occludens; a second type of microfilament bundle, containing actin, myosin, fodrin and vimentin, occurs in the apical processes and may be involved in photoreceptor renewal.

Under conditions of stress, RPE cells may express proteins more typical of macrophages and other myeloid cells. These include receptors for the Fc portion of immunoglobulin, the CD68 molecule and inducible nitric oxide synthase (iNOS) (see Ch. 7, p. 385). RPE cells also express the leucocyte marker CD36, which is involved in ROS phagocytosis (see p. 254) as well as $\alpha_v\beta_5$ integrin and the tyrosine kinase MerTK. In addition, RPE cells contain high quantities of phosphatidylcholine and phosphatidylinositol with high levels of saturated fatty acids and a high content of arachidonic acid. This may explain to some extent their ready ability to generate prostaglandins with immunosuppressive properties. The high content of cholesterol in RPE membranes indicates a low plasma membrane fluidity compared, for instance, with rod outer segment membranes, whose cholesterol content is low.

Turnover in normal RPE is similar to that in endothelial cells (i.e. very slow or nil) since it is a post-mitotic cell, but under certain circumstances RPE cells can proliferate and contribute to pathological processes as in retinal detachment (see Ch. 9, p. 492). When appropriately stimulated, RPE cells can synthesize and secrete growth factors such as FGF, IGF-1 and interleukin-1, which most likely have a role in the normal physiology of the retina. In addition, RPE cells are an important inducible source of vascular endothelial growth factor (VEGF). The production of growth factors by the RPE is not fully understood teleologically and it may play a part in regulating other tissues, such as the choroid and sclera, and thus indirectly have a role in the development of myopia (see eFig. 4-5). However, in pathological situations this may contribute to conditions such as diabetic retinopathy and subretinal neovascularization. To offset this risk, RPE cells constitutionally secrete a specific anti-angiogenic protein, pigment epithelium-derived factor (PEDF), as well as secreting thrombospondin-1 into the pericellular matrix.

The role of the RPE as the second site of the blood–retinal barrier (see Ch. 1, p. 40) is based on normal tight junctions containing the tight junction-specific proteins claudin-3 and -19 and is completed with a ring of cytoskeletal actin–myosin (see above); since the RPE also has a major role in pumping fluid across the retina from the subretinal space to the choroid, the tight junctions are integrated with a series of membrane transporters including Cl^- and HCO_3^- channels, under control of ATPases and carbonic anhydrase, as well as GLUTs 1 and 3. Thus bidirectional transport of various metabolites occurs, but the main bulk flow of fluid from the retinal side to the choriocapillaris is mostly by CFTR, the Ca^{2+}-dependent Cl^- channel, which is one of the family of ABC transporters (mutations in which are responsible for certain forms of retinal degeneration) rather than by aquaporin 1. Bestrophin 1, the gene responsible for Best's disease, may also be involved in either Cl^- or HCO_3^- transport in the RPE. The RPE cell is therefore a highly polarized epithelium, organized for transport, with its apical microvilli in apposition to the photoreceptor cell and its basal infoldings towards the choroid (Fig. 4-68).

RPE cells also express proteins on their apical surface that would be basolaterally expressed in other epithelia. These include N-CAM (a cell adhesion molecule) and EMMPRIN (extracellular matrix metalloproteinase inducer), which are likely to be involved in photoreceptor adhesion and phagocytosis, respectively. In addition, the phagocytosis receptor $\alpha_v\beta_5$ integrin is expressed on the apical surface with the Na^+-K^+ ATPases involved in transport. Other apical membrane-associated proteins include ezrin, which is associated with long apical microvilli, radixin and moesin/yurt, some of which are essential to apical basal polarization and are dependent on a range of signalling molecules such as a lipid phosphatase (PTEN), a serine-threonine kinase, and a transmembrane protein (Crumbs). Reversed apical polarization occurs postnatally and is the result of suppressed decoding of specific basolateral signals.

The RPE sits on a prominent basement membrane, Bruch's membrane, composed of five discrete layers,

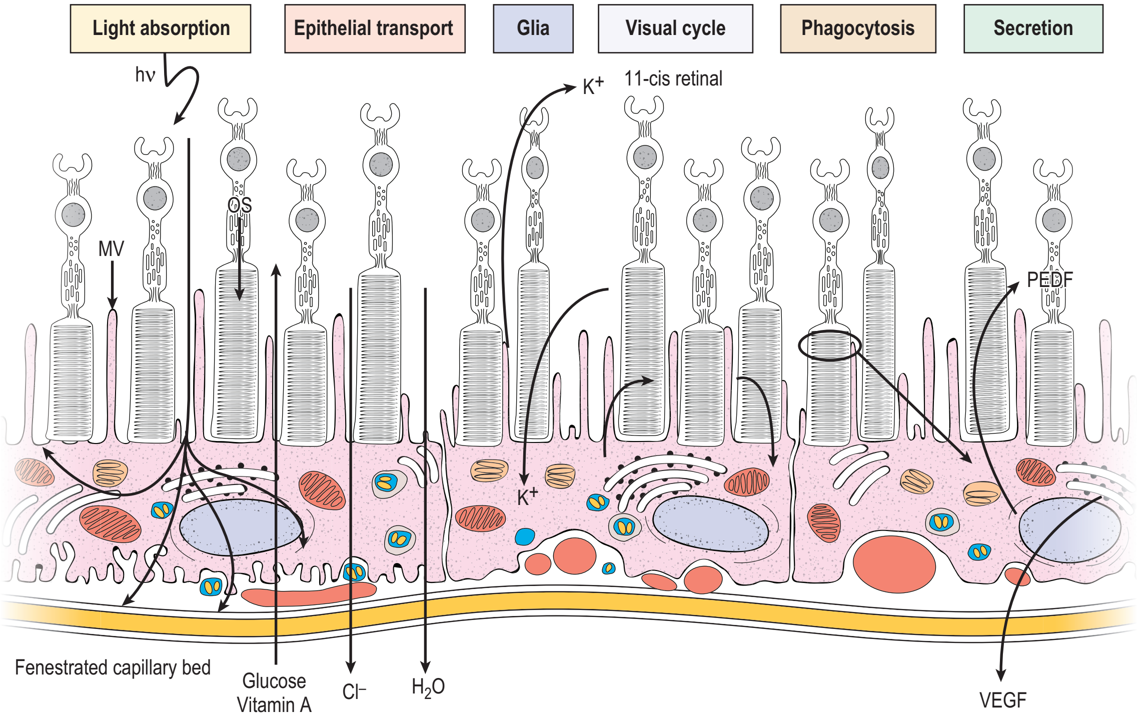

FIGURE 4-68 Summary of retinal pigment epithelium (RPE) functions. PEDF, pigment epithelium-derived growth factor; VEGF, vascular epithelium growth factor. (From Strauss, 2005.)

part of which incorporates the basement membranes of the RPE and the choriocapillaris, respectively. In addition to proteoglycans and matrix proteins typical of any basement membrane, Bruch's membrane contains hyaluronan and chondroitin sulphate plus types I, III, VI and VII collagen, and elastin.

Photoreceptor function is critically dependent on a healthy RPE layer

The RPE has multiple functions (Table 4-9) but its essential role is to maintain the physiology of the photoreceptors. Thus it removes 'spent' photoreceptor tips in the diurnal process of receptor renewal and participates in 11-*cis* retinol recycling. Indeed, it has been known for over 100 years that contact with the RPE was necessary for the bleached retina to regain its 'visual purple'.

TABLE 4-9 **Functions of the RPE**
Photoreceptor renewal
Retinal attachment
Interphotoreceptor matrix production
Transport of water and metabolites
Retinoid metabolism
Blood–retinal barrier
Immunoregulation
Free radical scavenging

As stated above, during the process of photoreceptor disk renewal the outer segment tips are shed in a diurnal manner and removed by the RPE cells in a short burst of phagocytic activity. Cone outer segments are similarly removed by the RPE cells but the process is considerably slower. Phagocytosed outer segment tips are digested in the extensive RPE phagolysosomal system, a process that continues throughout life. Solubilized waste material is then transported across the extensive basal infoldings of the cell into the choriocapillaris. It is not surprising therefore that, with age, there is accumulation of lysosomal bodies and lipofuscin pigment, which may reflect the declining ability of the RPE cell to handle large amounts of relatively indigestible material. This may be linked to reduced melanin production with age. This partly digested rod outer segment 'retinoid' material, some of which is known as A2E (see below), has been implicated in the development of age-related macular degeneration.

Although some retinoid material will be incorporated in the phagocytosed outer segment tip, this material is not used for regeneration of bleached rhodopsin. Instead this occurs in the cytosol of the RPE cell. The conversion of 11-*cis* retinal to the all-*trans* retinal during phototransduction is accompanied by release of the chromophore into the interphotoreceptor matrix (Fig. 4-69; and see below). Regeneration of the 11-*cis* retinal from the all-*trans* retinal takes place only in the RPE, since retinol isomerase is unique to this cell. Therefore, the all-*trans* retinal must somehow enter the RPE cell by a means other than outer segment tip phagocytosis. A slightly different process occurs for cone chromophore regeneration. In cones, after phototransduction, 11-*cis* retinal is reduced to all-*trans* retinal, which is then transported to Müller cells, where it is converted to 11-*cis* retinol and transported back to cones, where it is reconverted back to 11-*cis* retinal (Fig. 4-69). This last step can only occur in cones.

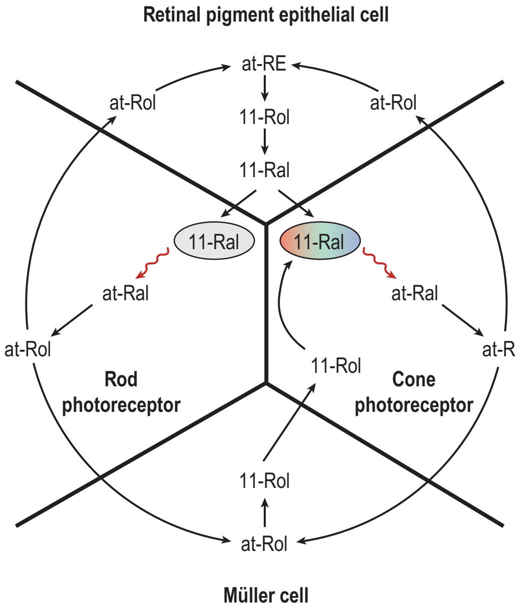

FIGURE 4-69 Reactions of rod and cone visual cycles. The areas separated by solid lines represent cellular compartments of a retinal pigment epithelial cell, rod and cone photoreceptor cells, and Müller cells. The ovals surrounding 11-Ral represent rod (grey fill) and cone (tricoluor fill) visual pigments. Photoisomerization reactions are shown in red. All other chemical reactions are catalysed by enzymes. Retinoids are chaperoned by retinoid-binding proteins (not shown) during intercellular and intracellular diffusion. Diffusion of at-Rol from rod photoreceptor cells to Müller cells has not been demonstrated. at-RE, all-*trans* retinyl esters; at-Ral, all-*trans* retinal; at-Rol, all-*trans* retinol; 11-Ral, 11-*cis* retinal; 11-Rol, 11-*cis* retinol. (From Saari, 2012.)

How the transport of the chromophores in the interphotoreceptor matrix is achieved is not yet clear, but it is probably more subtle than the proposed mechanism of retinoid shuttling in which IRBP 'transports' the retinoids between the two cells; rather, it is dependent upon the 'buffering' effect of IRBP, in which low-affinity binding of retinoid to IRBP permits its release at the appropriate site depending on the local concentration and amount of chromophore bleach. This notion is indirectly supported by evidence from the IRBP knockout mouse; this mouse has a fully functional neurochemical visual cycle but an accelerated dark adaptation response. In contrast, the exclusive localization of the high-affinity binding protein cellular retinol-binding protein (CRBP) to the RPE cell will help to drive the 'flow' of all-*trans* retinol in this direction and assist in the development of concentration gradients. Docosahexanoic acid, an important photoreceptor fatty acid, in contradistinction to palmitic acid, induces a rapid and specific release of 11-*cis* retinal from one of the two retinoid binding sites on IRBP.

For rods most of this activity takes place in the dark. However, IRBP also functions in transporting 11-*cis* retinol to cones which can take place in light conditions. There are two binding sites for the chromophore on the IRBP molecules which may also have functional significance in terms of vitamin A transport.

The RPE is polarized with tight junctions, but transport is bidirectional

The transport function of the RPE is bidirectional in that, as shown above, retinoid transport, as well as glucose transport, is achieved through polarized distribution of specific receptors (Fig. 4-68). In contrast, the bulk of transport occurs from the retina to the choriocapillaris, particularly for digested outer segment material. There is also a significant bulk flow of fluid from the retina to the choroid, probably via 'solute drag' through mechanisms such as an Na^+/K^+ ATPase pump located in the apical plasma membrane of the RPE and by transport of non-ionic solutes such as amino acids and glucose. Interestingly, RPE cells do not appear to express the water transporters aquaporin 3–5, unlike Müller cells and astrocytes. However, active HCO_3^- transport appears to be the major ion linked to fluid transport and is mediated by a carbonic anhydrase-regulated system (see section on ciliary body, p. 217). There are six isoforms of carbonic anhydrase in the RPE cell. The exchange of ions (and bulk water) establishes an electrical potential across the RPE, positive on the retinal side, of 5–15 mV, and also maintains steady pH regulation in the face of high lactic acid production by the outer segments of the photoreceptor cell. However, as indicated above, the Ca^{2+}-activated Cl^- channel, CFTR, and bestrophin membrane proteins are currently considered strong players in trans-RPE fluid flow.

The net rate of fluid transport across the RPE is about 4–6 $\mu L/cm^2$ per hour. Transretinal fluid flow has been proposed as a major mechanism for maintaining retinal apposition; indeed, it has been suggested that clinical RPE detachments may result from a breakdown of the transport mechanisms for fluid across the RPE as a result of focal damage.

Thus any process that impairs retinol transport into the RPE from central stores in the liver will affect vision; because vitamin A requirements are supplied through the diet, any condition affecting this, such as protein–calorie malnutrition or one of the malabsorption syndromes, may produce visual symptoms. One of the earliest symptoms of severe malnutrition, endemic in developing countries, is night blindness caused by lack of vitamin A.

The RPE also contains α and β_2 adrenergic receptors, plus enzymes of the cytochrome P_{450} drug-metabolizing system. In addition, melanin, a major constituent of RPE cells is effective in drug detoxification. The melanin content of RPE cells decreases with age.

The chemistry of the visual response

The main function of the retina is to convert light energy into an interpretable signal for cortical cells in the brain. This process begins with photochemical events in the photoreceptor.

PHOTOCHEMICAL REACTIONS IN THE RETINA

Rhodopsin, vitamin A and photoreceptor turnover

Rhodopsin, a 348 amino acid long major integral cell membrane protein in rod photoreceptor outer segments, is synthesized in the ER and Golgi apparatus of the inner segments and transported in protein-rich vesicles to the outer segment, where fusion occurs with the newly formed disk membranes in the periciliary ridge complex (see ciliopathies). It has three glycosylation sites containing various branched combinations of *N*-acetylglucosamine and mannose residues. Opsin is a seven-turn (α helix) membrane-spanning protein containing a serine- and threonine-rich cytosol-exposed C-terminus, which is variably phosphorylated, and an intradisk N-terminus (Fig. 4-10). This structure follows the general pattern for G protein types of membrane receptor such as the adrenergic and muscarinic receptors which induce cell signalling on binding of their specific ligand by activating adenyl cyclase to raise intracellular concentrations of the second messenger cAMP (see pp. 243–245 and Ch. 6, pp. 172 and 258). The ligand for rhodopsin, 11-*cis* retinal, is already bound in the dark and dissociates when activated by a photon of light.

Rhodopsin behaves like a genuine receptor, but with differences: in the resting state (i.e. in the dark), Na^+ channels in the rod outer segment plasma membrane are held 'open' by cGMP, synthesized by guanylate cyclase. This provides the electrochemical basis for the relative depolarization of the photoreceptor outer segment compared with other cells (–57 versus –78 mV). On stimulation of rhodopsin with light, transdisk membrane signalling occurs via sequential activation of other membrane-bound proteins, transducin and phosphodiesterase, to lower the cytosolic concentration of cGMP, i.e. it acts like a second messenger in reverse (see below). This has the effect of closing the leaky Na^+ channels in the plasma membrane and causing a relative hyperpolarization (to –87 mV), thereby generating the electrical response (see below).

Activation of rhodopsin is achieved via isomerization of retinol, a vitamin A compound that lies 'nested' between the first and last transmembrane loops of the rhodopsin molecule, with its long axis in the plane of the membrane (Fig. 4-10). Modelling studies show the relationship between the chromophore pocket and the opsin molecule (Fig. 4-70). In this reaction, the tail of 11-*cis* retinal on conversion to all-*trans* retinal becomes elongated and more perpendicularly disposed to the isoprene retinal ring structure; this has the effect of generating greater interaction between the retinal and its binding sites to the side-chains of amino acids Lys206 and Lys296, thereby heightening the energy state of the rhodopsin molecule. All-*trans* retinal becomes converted to all-*trans* retinol, which does not fit within the rhodopsin transmembrane loops and it is during this stage that rhodopsin converts through its various intermediaries to opsin, the changes being demonstrable by colour bleaching (Fig. 4-70). The chromophore thus detaches from the bleached opsin and diffuses away into the interphotoreceptor matrix to be taken up by the RPE cell; in the dark it undergoes isomerization within the RPE cell (see Figs 4-69 and 4-70) to 11-*cis* retinol, which binds to cellular retinal binding protein (CRALBP) and becomes converted to 11-*cis* retinal-CRALBP, which is transported to the cell membrane and is transferred to interphotoreceptor retinal binding protein (IRBP) on which it is shuttled back to the photoreceptor to reattach to rhodopsin and recommence the cycle.

The process of 11-*cis* retinal binding to opsin occurs by the formation of a Schiff base to Lys296, releasing a single molecule of water. Interestingly, 'ordered' water molecules are also included in the overall conformation of rhodopsin as tightly bound 'prosthetic' structural moieties giving cohesion to the overall structure of the molecule (see Fig. 4-10). It is possible that these water molecules are held to the structure through electrochemical bonds that become disrupted when light converts 11-*cis* retinal to all-*trans* retinal, thus creating space for the chromophore to detach.

There are three cone opsins in humans which have around 50% homology to rhodopsin and to each other apart from the red and green opsins which are 95% homologous to each other (Fig. 4-71). The release of all-*trans* retinal from cone pigments is much faster than from rhodopsin and even binding of *cis*-retinal is less stable, possibly related to the amount of water binding.

The conversion of 11-*cis* retinal to all-*trans* retinal and then to all-*trans* retinol is the fundamental chemical reaction to take place during the visual impulse, and the chemical reaction has been known for many years. Laser flash photolysis studies have chemically identified the intermediates that occur in the breakdown of vitamin A (see Figs 4-69 and 4-70). Normal vision depends on a plentiful supply of vitamin A, which must be provided exogenously because humans cannot synthesize it. Dietary vitamin A comes from β-carotene in plants, while animal sources provide the chromophore as a retinyl ester linked to fatty acids. Thus, visual deterioration, especially at low illumination (night blindness), is an early sign of malnutrition and malabsorption syndromes, but sensitive electrophysiological tests may be required to reveal its full extent. Retinoids are stored in the RPE as non-toxic retinyl esters; retinoic acid, which can also be synthesized during retinoid metabolism, has many functions in cell biology and has recently been shown to have a role in immune tolerance, but free retinoic acid is toxic to cells and thus has to be generated from retinyl esters, similar to visual retinoids. In the retina, all-*trans* retinol to 11-*cis* retinol isomerization is regulated by palmitoylation under the control of the RPE-specific protein RPE-65 (see Fig. 4-69) whose role serves to switch off the visual cycle in the dark.

Phototransduction: the conversion of light energy to an electrochemical response

The conversion of the energy stored in a single photon of light to an electrical response is possible because of the extensive amplification of the molecular cascade involved in closure of the Na^+ channels (see eBox 4-5).

These channels are kept open by cGMP (termed cyclic nucleotide gated channels, CNGP), which acts as a second messenger in this system.

In the dark, open Na^+ channels allow Na^+ to exit and Ca^{2+} to enter and maintain a relative depolarization. This is linked to a Na^+/Ca^{2+} exchanger, which maintains a steady intracellular concentration of Ca^{2+}. In the light, the Na^+ channels are closed, leading to a relative hyperpolarization and a decline in the intracellular Ca^{2+} concentration. Closure of the Na^+ channels is achieved by hydrolysis of cGMP by phosphodiesterase (PDE), one of the responses that is greatly amplified in the cascade. Signal amplification is achieved by activation of many molecules of

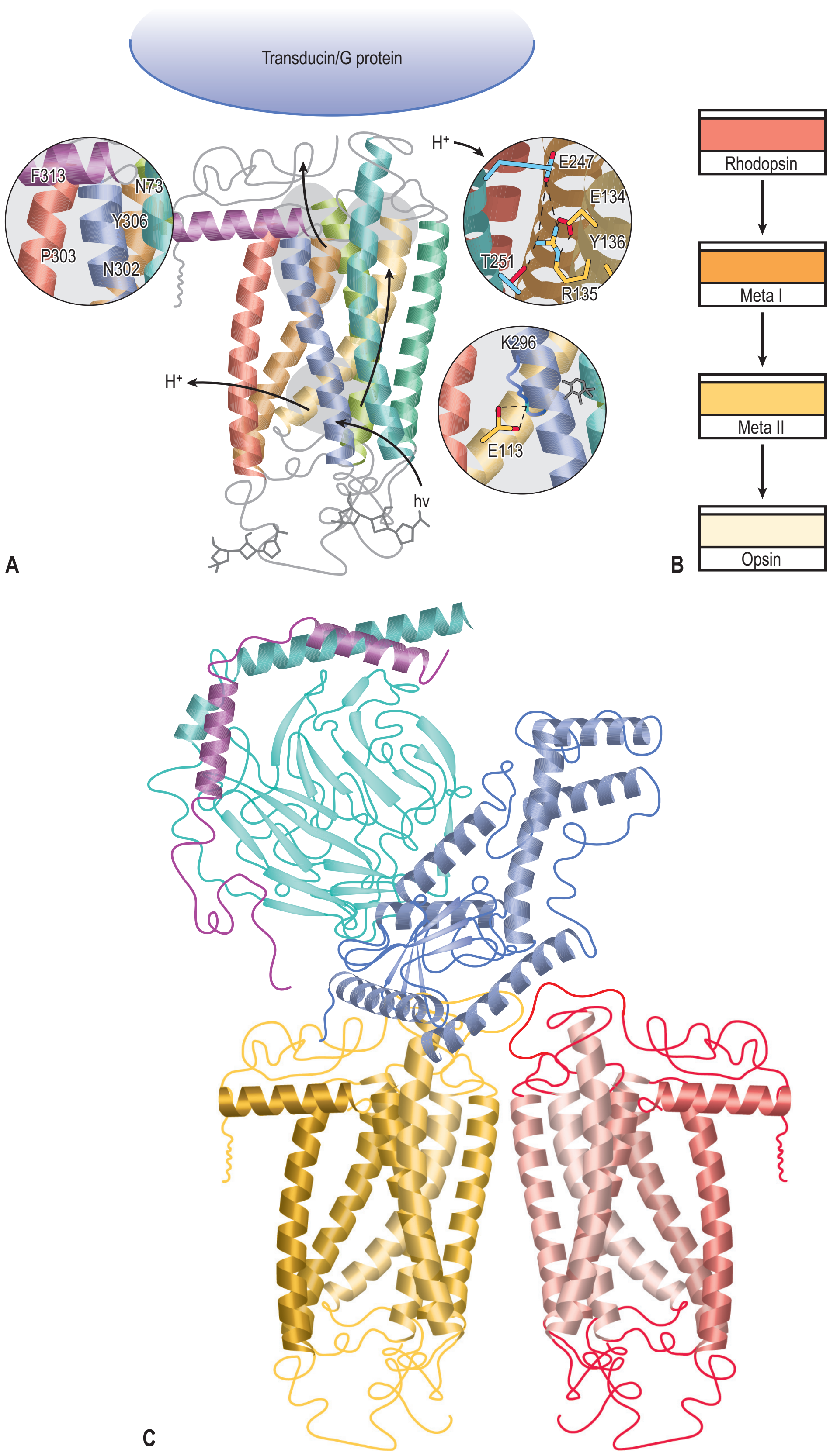
Transducin/G protein
F313
N73
Y306
P303
N302
H+
E247
E134
Y136
T251
R135
K296
E113
H+
hν
A
Rhodopsin
Meta I
Meta II
Opsin
B
C

transducin by one molecule of metarhodopsin (R*), probably in the region of 300× (Fig. 4-72 and eBox 4-5).

In this sense, activated rhodopsin acts like an enzyme since it is available to activate further molecules of transducin until the light is switched off by the combined action of recoverin (rhodopsin kinase) and arrestin (also known as S antigen). Similarly, around 600 molecules of cGMP are hydrolysed by a single molecule of PDE and, indeed, the limitation of its activity is determined by the availability of cGMP. There is a further three-fold amplification of channel opening by one cGMP molecule.

Phototransduction is a biological cascade

The mechanism of cGMP hydrolysis in phototransduction has been extensively investigated; the sequence of events in this response is outlined in Figure 4-72.

Certain enzymes play important roles in this process, including guanylate cyclase and cGMP phosphodiesterase. Guanylate cyclase activation restores cGMP levels with the help of recoverin (see Fig. 4-72), while the enzyme R* is deactivated to the phosphorylated form R–P and bound by arrestin (S antigen). During these changes there is considerable redistribution of photoreceptor molecules in the shift from light to dark (Fig. 4-73). Amplification is therefore a function of the time interval at each of the stages of processing.

Does phosphoinositide metabolism have a role in phototransduction?

From the above it would appear that much of the transduction mechanism has been deduced; however, the role and mechanism of Ca^{2+} influx are unclear. Ca^{2+} mobilization is the product of a major second messenger system in many cells, and is activated via receptor-mediated activation of phospholipase C (PLC) and G (PLG) proteins. This results in the hydrolysis of phosphatidylinositol 4,5-bisphosphate (PIP_2) to form inositol triphosphate (IP_3) and diacylglyceroal (DAG). IP_3 induces mobilization of calcium stores, while DAG activates protein kinase C. Two major calcium-binding proteins, calmodulin and GCAP1 and 2 (calcium-dependent modulator of guanylate cyclase), are involved in phototransduction. Both PKC and PLC regulate the translocation of arrestin from the inner to the outer segments in the dark.

In the invertebrate retina, light induces hydrolysis of PIP_2, with production of IP_3 and DAG, but its role in the transduction cascade is unclear. However, there is evidence that an integral transmembrane phosphatidylinositol transfer protein prevents retinal degeneration in *Drosophila* (the rdg B protein). The evidence in the vertebrate retina is even less convincing. It is possible that this mechanism may have some other function related to membrane conductance and mobilization of ion stores, but unrelated to transmission of an electric impulse. However, there is little doubt that phosphoinositide metabolism and turnover is a major pathway in the photoreceptor cell.

Renewal of photoreceptors is associated with accumulation of lipofuscin pigment in the RPE cells

Estimates suggest that RPE cells phagocytose about 2000–4000 disks per day. To cope with this phagocytic load, the RPE cell has an extensive lysosomal enzyme system that can digest about 50% of its load within 1–2 hours (important in terms of cyclic rod outer segment tip shedding). Uptake of shed rod outer segment tips is receptor mediated, through the $\alpha_v\beta_5$ integrin receptor signalling, the receptor tyrosine kinase Mer (Mertk) and CD36. CD36 recognizes oxidized phosphatidylcholine, a product of aged outer segment tips, and may be involved in the engulfment process rather than initial binding. Interestingly, the $\alpha_v\beta_5$ receptor is also involved in retinal adhesion to

FIGURE 4-70 Activation of rhodopsin by light. (**A**) Rhodopsin absorbs a photon of light, which leads to isomerization of 11-*cis* retinal to all-*trans* retinal. Transmission of this signal to the cytoplasmic surface triggers nucleotide exchange on the heterotrimeric G protein transducin. Transducin dissociates and activates the downstream signalling events. (**B**) Spectral changes in rhodopsin upon activation as detected by photography. Once rhodopsin in its dark 11-*cis* retinal-bound state (**A**) is exposed to light, it immediately goes through a series of photointermediate states, including metarhodopsin I (Meta I) (**B**), and eventually progressing to the Rho* (metarhodopsin II (Meta II)) activated state (**C**). Upon treatment with hydroxylamine, the chromophore is hydrolysed, resulting in a largely colorless solution. (**C**) Shows a model of the G protein rhodopsin complex. (From Palczewski, 2012.)

eBox 4-5

Phototransduction cascade and channel opening

Activation of a single molecule of rhodopsin generates an amplification cascade that leads to the opening of many channels, and the induction of a change in the resting potential of the photoreceptor.

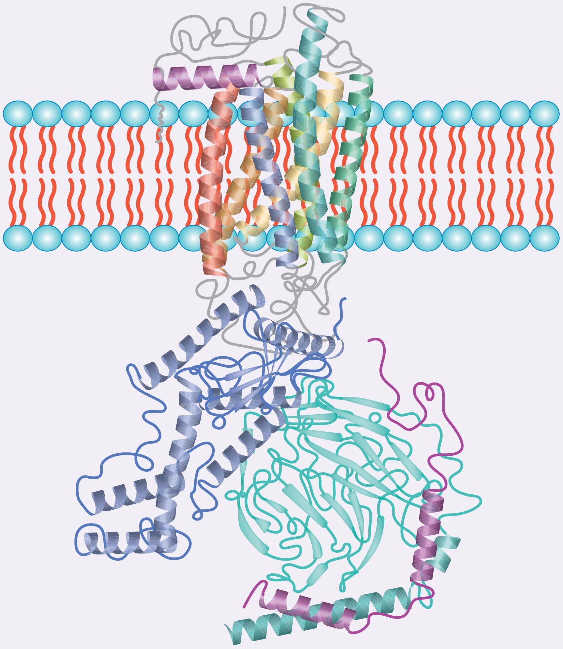

Type 2 rhodopsin (Rh) (rainbow coloured) embedded in a lipid bilayer (heads red and tails blue) with transducin (GDP) below. $G_{\tau\alpha}$ is coloured red, $G_{\tau\beta}$ blue and $G_{\tau\gamma}$ yellow. There is a bound GDP molecule in the $G_{\tau\alpha}$-subunit and a bound retinal (black) in the rhodopsin. The N-terminus of rhodopsin is red and the C-terminus blue. Anchoring of transducin to the membrane has been drawn in black. PDE, phosphodiesterase.

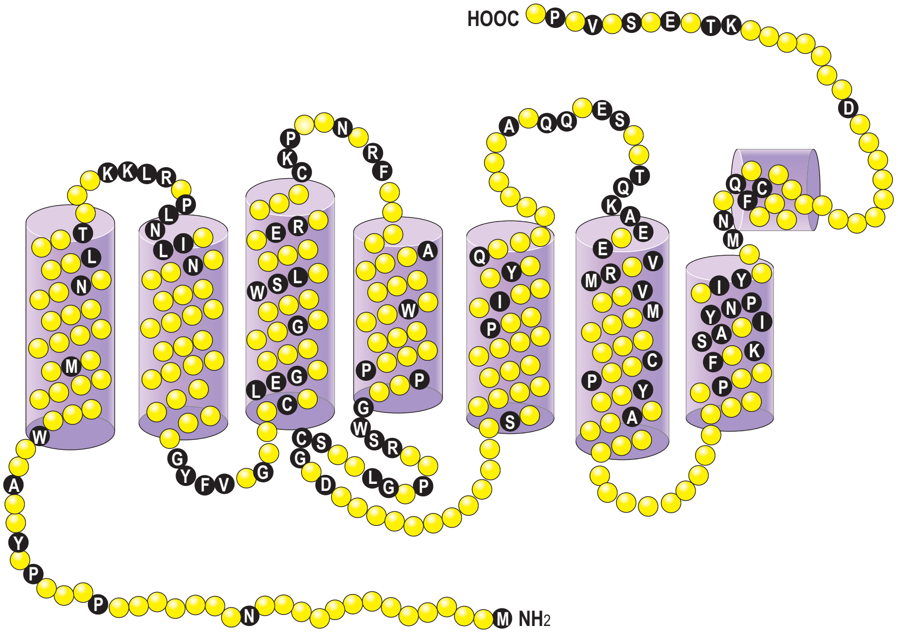

FIGURE 4-71 Sequence identity among the human rod and cone opsins. Opsin secondary structure is represented in two dimensions based on alignment with the known structure of bovine rhodopsin. Residues that are identical between the human rod and human red, green and blue cone opsins are denoted as black circles with the conserved amino acid residues identified. (From Tang et al., 2013.)

the RPE layer, functioning in a different diurnal cycle for phagocytosis.

With age, accumulation of retinoid material and lipofuscin occurs in the RPE. Free all-*trans* retinal, produced after the photoactivation of rhodopsin by light, is considered to bind to phosphatidylethanolamine (PE), forming N-retinylidene-PE (NRPE) from which N-retinylidene-N-retinylphosphatidylethanolamine (A2PE) is generated by the binding of a second molecule of all-*trans* retinal. Phospholipase D then cleaves PE from A2PE to generate A2E, which has been identified as a potential culprit in the lipofuscin responsible for age-related macular degeneration (AMD). This process is light dependent and thus retinal light damage is thought to be a major factor in development of AMD. However, recently, 11-*cis* retinal has also been shown to generate A2E, which raises a question mark concerning this model of light damage to the retina.

Photoreceptors are easily damaged, particularly by excess light radiation, and several free radical scavenging mechanisms are in place to minimize the damage. In addition, vitamin E has a significant role as a free radical scavenger (see Box 4-19).

RPE lysosomes contain a battery of enzymes capable of degrading complex lipid–glycoprotein aggregates

Phagocytosis of rod outer segment tips takes place in stages. First, there is formation of the primary phagolysosome by fusion of the endosome with the lysosome. This may then fuse with lipofuscin granules or with melanosomes to form melanolipofuscin bodies. With age, and experimentally in vitamin E deficiency, there is a greater shift towards fusion with lipofuscin granules. Eventually these bodies are degraded and their products are transported out of the cell. However, the accumulation of insoluble material between the basement membrane and the RPE cell has been attributed to a breakdown in the capacity of the cell to deal with this load. In addition, accumulation of lipofuscin is associated with reduced phagocytic capacity of RPE cells.

SYNAPTIC EVENTS BETWEEN PHOTORECEPTORS AND CELLS OF THE INNER NUCLEAR LAYER

Bipolar cells of the inner nuclear layer (INL) synapse either with rod or with cone photoreceptors but not with both. In addition, they synapse with other cells

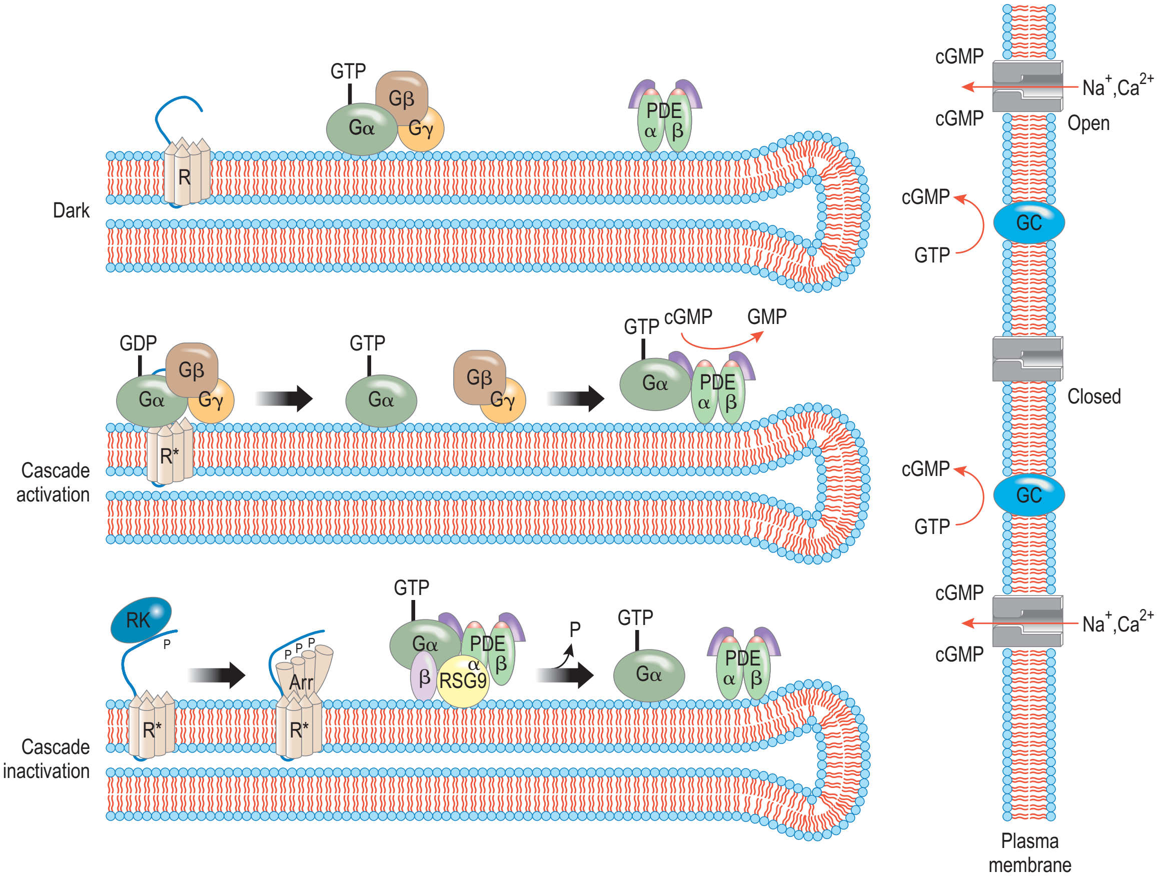
Dark
R
GTP
Gα
Gβ
Gγ
PDE
α
β
cGMP
cGMP
Na+,Ca2+
Open
cGMP
GTP
GC
Cascade
activation
GDP
Gα
Gβ
Gγ
R*
GTP
Gα
Gβ
Gγ
GTP
cGMP
GMP
Gα
PDE
α
β
Closed
cGMP
GTP
GC
Cascade
inactivation
RK
P
R*
P P P
Arr
R*
GTP
Gα
β
RSG9
PDE
α
β
P
GTP
Gα
PDE
α
β
cGMP
Na+,Ca2+
cGMP
Plasma
membrane

FIGURE 4-72 Schematic of phototransduction cascade activation and inactivation. The upper disk illustrates inactive rhodopsin (R), transducin (Gα, Gβ and Gγ subunits), and PDE (α, β and γ subunits), in the dark. The reactions in the middle disk illustrate light-induced transducin and PDE activation. The reactions in the lower disk represent R* inactivation via phosphorylation by rhodopsin kinase (RK) followed by arrestin (Arr) binding and transducin/PDE inactivation by RGS9–Gαβ-PDEαβ complex. (From Burns et al., 2005, with permission from Elsevier.)

in the INL such as horizontal cells. There are around 10 different types of cone bipolar cells but only one type of rod bipolar. Several types of neurotransmitter may be involved (Table 4-10). In the resting state, the dark currents produced by the open cation channels in the outer segment (see Ch. 5, Box 5-6, p. 288) are accompanied by high levels of neurotransmitter release at the synaptic junction with the bipolar cells. In the light, there is a decrease in transmitter release, which alters the transmission of electric potentials in the bipolar cells. However, the situation is more complex than this.

Two types of bipolar cell

Synaptic transmission at neuromuscular junctions is mediated by acetylcholine, released from the nerve ending when the action potential wave arrives. Binding of acetylcholine to its receptor on the muscle initiates a second message, which causes a depolarization in the cell owing to opening of the voltage-gated Na^+ channels.

At the synaptic junction between the photoreceptor cell and the bipolar cell, a different type of reaction occurs. First, the photoreceptor does not transmit an action potential wave, but presents a graded hyperpolarization response, which depends on the intensity of the light stimulus. In the dark, while the photoreceptor is in a relatively depolarized state, the neurotransmitter glutamate (see Table 4-10) is released from the presynaptic photoreceptor terminal and binds to a retina-specific metabotropic receptor (see next section for definitions) on one type of bipolar cell (the

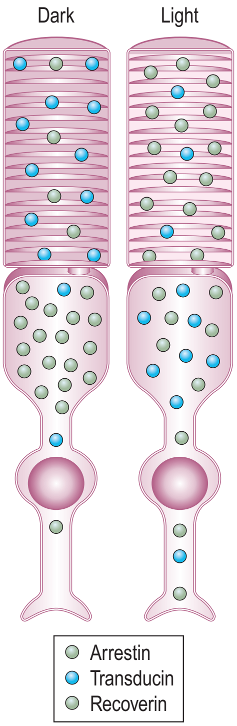
Dark
Light
Arrestin
Transducin
Recoverin

FIGURE 4-73 Schematic representation of transducin, arrestin and recoverin subcellular distribution in the dark-adapted and light-adapted rod. (From Burns et al., 2005, with permission from Elsevier.)

BOX 4-19 VITAMIN E AS A FREE RADICAL SCAVENGER

Vitamin E limits free radical damage at several stages in lipid peroxidation.

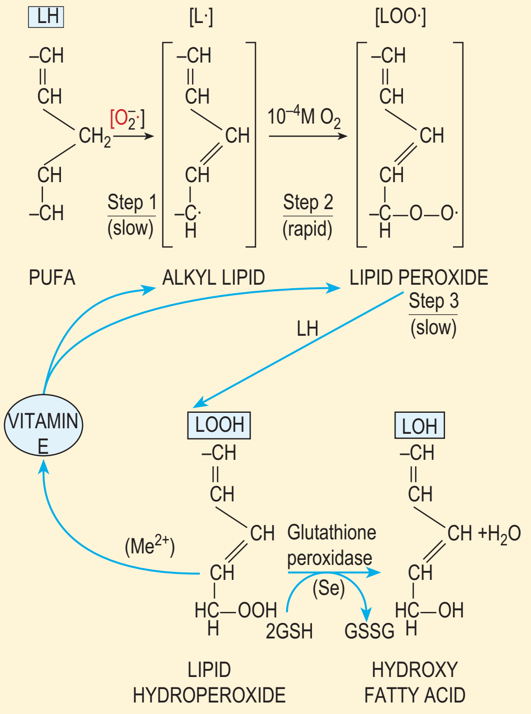
LH
[L·]
[LOO·]
–CH
CH
CH₂
[O₂⁻·]
Step 1
(slow)
10⁻⁴M O₂
Step 2
(rapid)
–C·
H
–C–O–O·
PUFA
ALKYL LIPID
LIPID PEROXIDE
Step 3
(slow)
LH
VITAMIN E
LOOH
LOH
(Me²⁺)
Glutathione peroxidase
(Se)
HC–OOH
2GSH
GSSG
CH +H₂O
HC–OH
LIPID HYDROPEROXIDE
HYDROXY FATTY ACID

Regeneration of vitamin E under light stress requires ascorbic acid. In experimental vitamin E deficiency there is a marked accumulation of lipofuscin in the RPE, which may be relevant to a similar accumulation of lipofuscin in age-related macular degeneration. PUFA, polyunsaturated fatty acid.

ON-bipolar), which produces hyperpolarization and keeps the bipolar Na^+ channel closed. When light activates the photoreceptor, glutamate release ceases and the intracellular concentration of cGMP rises, thus opening the Na^+ channels and causing depolarization of the cell membrane, allowing the ON-bipolar cell to transmit a signal.

In contrast, glutamate binding to the ionotropic receptor on a second type of bipolar cell (the OFF-bipolar; see Ch. 5) has the reverse effect of inducing hyperpolarization. To summarize, therefore, when light activates the photoreceptor and transmits the message to the bipolar cell, ON-bipolar cells are depolarized by reduction in glutamate release and, OFF-bipolar cells are hyperpolarized by an increase in glutamate release. Additionally, glutamate may bind to horizontal cells, where it also produces hyperpolarization.

This type of graded response is achieved by specialized photoreceptor structures termed ribbon synapses which are exquisitely organized to transmit information. Several membrane proteins are involved in the cytoskeletal structure and function of the synapse, such as dystrophin and retinoschisin, which have clinical relevance (mutations in retinoschisin cause retinoschisis, a retinal degeneration involving splitting of the retinal layers) as well as in the inter-synaptic cleft, a space containing other essential matrix components (Fig. 4-74). In both ON- and OFF-bipolar cells, the action of glutamate is considered excitatory, since there is a defined response.

TABLE 4-10 **Retinal neurotransmitters and neuromodulatory peptides**

Agent	Site of action
Neurotransmitters	
Glutamate	Photoreceptor
γ-aminobutyric acid	Horizontal amacrine cells
Glycine	Amacrine cell
Taurine	
Tyrosine derivatives	
Dopamine	Horizontal amacrine cells
Noradrenaline	
Adrenaline	
Tryptophan derivatives	
Serotonin	Photoreceptor
Aspartate	Photoreceptor
β-alanine	
Histidine derivative	
Histamine	
Neuromodulators	
Peptides	
Vasoactive intestinal peptide	Amacrine, ganglion cells
Angiotensin I and II	Amacrine, ganglion cells
Substance P	Amacrine, ganglion cells
Luteinizing hormone releasing hormone and thyrotrophin releasing hormone	Amacrine, ganglion cells
Leu- and metenkephalin	Amacrine cell
β-endorphin	Amacrine, ganglion cells
Somatostatin	Amacrine, ganglion cells
Neurotensin	Amacrine cell
Glucagon	Amacrine cell

Glutamate is unlike acetylcholine in that there is no enzyme that rapidly degrades the neurotransmitter and shuts down the response; instead, the glutamate merely diffuses away or is taken up by nearby glial cells and inactivated.

The differential binding of glutamate to ON- and OFF-bipolar cells via specific receptors (see below) is a direct chemical correlate of the electrophysiological and psychophysical responses that can be obtained from these cells, and is a good demonstration of nature's use of a binary (ON–OFF) system to process complex sensory phenomena. Apparently, the segregation of responses at the bipolar cell level is retained to the level of the visual cortex.

Neurotransmitters and neuromodulators

In contrast to glutamate, other neurotransmitters such as glycine and GABA have a lateral inhibitory action at the synaptic junction and are considered to be neuromodulators, i.e. they modulate the effect of glutamate; neuromodulators have a much longer action, often mediated through second messenger systems in the cell. There are many different types of neurotransmitter, most of which are amino acids or amino acid derivatives, and the majority are active in transmission of electric impulses within the mammalian retina. As indicated above, typical excitatory transmitters include acetylcholine and glutamate, while γ-aminobutyric acid (GABA) and glycine are inhibitory to their function. Several other neuromodulatory peptides exist which have direct or indirect effects on neural responses. Dopamine is probably the best known of these (see below). In addition, recent studies indicate that neuronal nitric oxide synthase may modulate retinal signalling via release of nitric oxide.

Glutamate is the major neurotransmitter in the retina

As indicated above, glutamate (Fig. 4-75) is the major mediator of synaptic transmission between the photoreceptor and the bipolar and horizontal cells. The functional organization of the retina is centred around the specificity of glutamate for four discrete types of glutamate receptor on different types of cells (ON- and OFF-bipolar cells, and two types of horizontal cell). Glutamate receptors are classified as ionotropic (iGluR) and metabotropic (mGluR). iGluR are by far the more ubiquitous and are further grouped as:

- *N*-methyl-D-aspartate (NMDA) selective receptors
- quisqualate receptors
- AMPA/kainate receptors on hyperpolarizing bipolar cells
- APB (sign-inverting glutamate receptor agonist) receptors.

Each of these receptors has specific characteristics related to whether it is ionotropic (directly affecting ion channels, e.g. NMDA) or metabotropic (primarily affecting aspects of neuronal cell metabolism, such as cGMP levels, and secondarily altering ion-channel permeability, e.g. APB receptors) (see Fig. 5-12, p. 295). Synaptic transmission in ganglion cells may also utilize glutamate. However, for some time it has

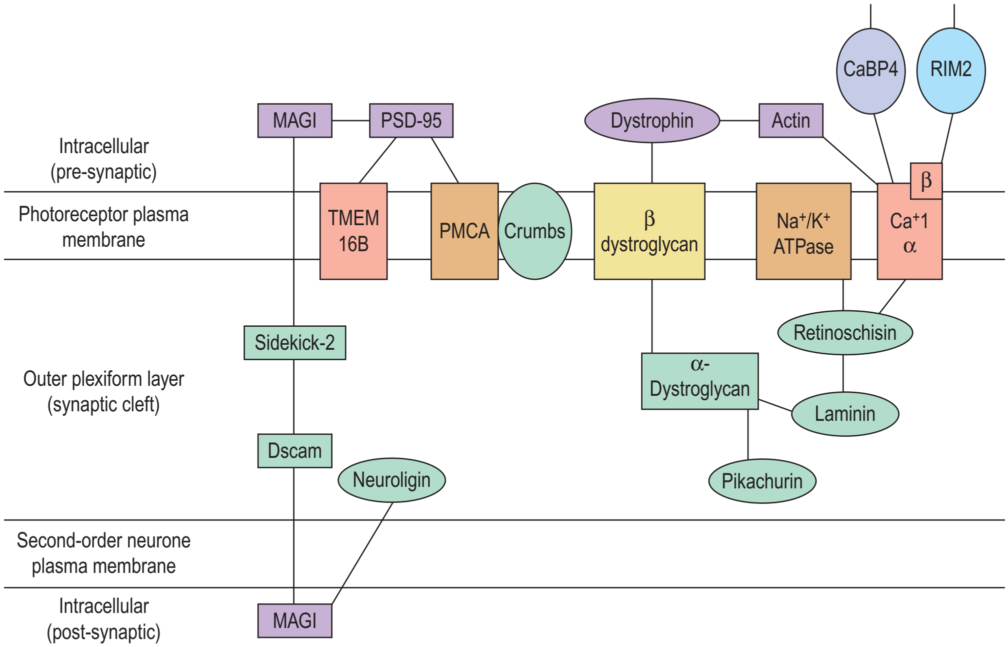
CaBP4
RIM2
MAGI
PSD-95
Dystrophin
Actin
Intracellular
(pre-synaptic)
β
Photoreceptor plasma
membrane
TMEM
16B
PMCA
Crumbs
β
dystroglycan
Na+/K+
ATPase
Ca+1
α
Retinoschisin
Sidekick-2
α-
Dystroglycan
Outer plexiform layer
(synaptic cleft)
Laminin
Dscam
Neuroligin
Pikachurin
Second-order neurone
plasma membrane
Intracellular
(post-synaptic)
MAGI

FIGURE 4-74 Proteins central to the organization of the outer plexifom layer (OPL). Two major nodes of protein–protein interactions occur in the extracellular matrix (ECM) of the OPL ECM within the OPL. The Sidekick-2 trans-synaptic scaffold attaches MAGI proteins in apposed cells, with putative lateral presynaptic interactions to PSD-95. PSD-95 also appears to anchor the CI(Ca) channel. TMEM16B and PMCA extrusion pump at non-ribbon sites in the terminal. Scaffolding interactions between a transmembrane dystroglycan complex and the retina-specific ECM proteins pikachurin and retinoschisin, linked through synaptic laminin and cytoplasmic actin, while also attaching to the extracellular face of transmembrane CaV channels and the Na^+/K^+-ATPase. (From Mercer and Thoreson, 2011.)

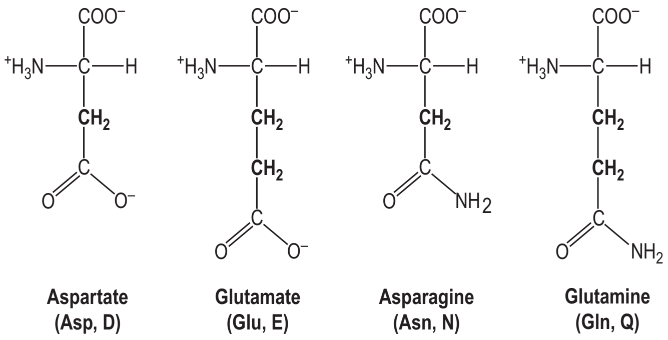
Aspartate
(Asp, D)
Glutamate
(Glu, E)
Asparagine
(Asn, N)
Glutamine
(Gln, Q)

FIGURE 4-75 Chemical structure of the acidic amino acids aspartate, glutamate asparagine and glutamine.

been proposed that small cyclic nucleotides may confer gating activity, and recent studies have confirmed this (Fig. 4-76). The cAMP and cGMP appear to have opposing gating effects in their interactions with horizontal/amacrine cells and ganglion cells in the regulation of Ca^{2+} influx and in addition appear to be modulated themselves by neuronal nitric oxide synthase (nNOS) derived from neighbouring amacrine cells. The precise role of NO in the retina is unclear but may be important in visual adaptation.

Horizontal and amacrine cells

Horizontal and amacrine cells directly modify the rate of electrical firing in bipolar cells by release of

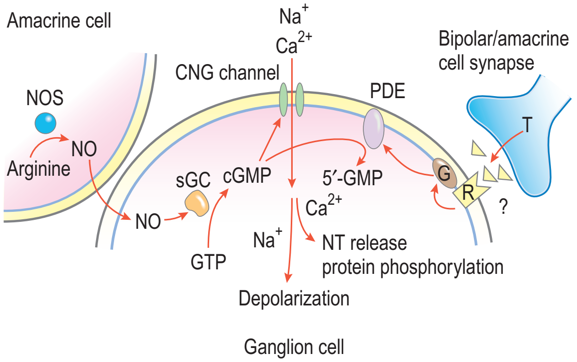

FIGURE 4-76 Schematic representation of pathways that might regulate the activity of the cGMP-gated channel in retinal ganglion cells. The cGMP synthesis could be stimulated by nitric oxide adjacent amacrine cells. Activation of cGMP-gated channels will increase Ca^{2+} influx and enhance Ca^{2+}-driven processes. At the same time, G protein-coupled receptors activated by neurotransmitters from bipolar and amacrine cell terminals may regulate the activity of one or more PDEs which control hydrolysis of cGMP and thus the activity of the cGMP-gated channels. (From Barnstable et al., 2004; modified from Ahmad et al., 1994.)

excitatory (horizontal cells) and inhibitory (horizontal and amacrine cells) neurotransmitters. Glutamate is the major excitatory transmitter of horizontal cells, while GABA, a derivative of glutamate distributed widely throughout the retina, functions particularly in the horizontal and amacrine cells as an inhibitory mediator. Recently, it has been shown that some horizontal cell negative feedback to cones can be mediated via glutamate-gated Ca^{2+} channels. Glycine is also an inhibitory neurotransmitter in amacrine cells, which have a significant role in determining the size of receptive fields of individual ganglion cells (see Ch. 5).

Acetylcholine is a major excitatory neurotransmitter of amacrine cells and often co-localizes in GABAergic starburst cells, suggesting that these cells can act in a complex excitatory/inhibitory manner in relation to ganglion cell receptive fields. Ganglion cells are thought to utilize glutamate as a major transmitter.

Amacrine cells also receive input from the third class of photoreceptor in the retina. Intrinsically photosensitive retinal ganglion cells (ipRGCs, see pp. 249–250) excite a subset of amacrine cells via dopamine receptors and are coupled to a further subset of GABA-responsive amacrine cells via gap junctions. In this way information regarding irradiance, circadian rhythms and pupillary responses is integrated with primary photoreceptor responses.

Dopamine exerts a neuromodulatory effect on retinal function

Dopamine occurs in amacrine cells and inner plexiform layer cells in the inner retinal layers and clearly has a role to play in visual function because the reduced and delayed b-wave electroretinogram responses (see Ch. 5) seen in patients with Parkinson's disease are normalized on treatment with l-DOPA. Dopamine receptors (there are two broad classes and several subclasses) are present on all retinal neurones and therefore dopamine probably has many as-yet undetermined effects, one of which is to uncouple photoreceptor-driven horizontal cell gap junctions. Horizontal cells are probably the source of activity in the surround region in bipolar cell centre-surround receptive fields, suggesting that regulation of spatial contrast sensitivity may be under dopaminergic control.

Dopamine may have many other effects including regulation of A11-type amacrine cells, which couple rod-associated bipolar cells and ganglion cells, and it may even have a paracrine role in altering the level of photoreceptor cytoplasmic cAMP. Dopamine has also been implicated in development in retinal neurones because it has been shown that apomorphine inhibits form-deprivation myopia in experimental animals by a dopamine D_2-receptor mechanism acting within either the retina or the RPE.

Dopamine is inactivated by cellular re-uptake, a process mediated by specific Na$^+$-dependent membrane transporter proteins, which have been cloned. Dopamine may also be antagonized by melatonin, which is synthesized and released by photoreceptors, accounting for the photopic to scotopic transition in rod responsiveness. It is also counteracted by NMDA, which may induce impaired dopamine synthesis. Interestingly, melatonin has neuromodulatory effects on dopamine-sensitive cells as well as on photoreceptors (see p. 261).

It can be seen therefore that, while much remains to be determined concerning the mediators of neural function in the retina and visual cortex, a view is already emerging of differential stimulation and inhibition of function induced by a variety of interacting neuromodulators and transmitters, which correlates with observed psychophysical events.

Conclusion

While the eye has many ‘eye-specific’ biochemical and cell biological functions and roles, many of these are founded on general principles that apply to other systems and, indeed, comparison of ocular and non-ocular processes offers much insight into basic fundamental principles. Where differences exist, they assist in highlighting the properties and functions of molecules and reactions, while introducing us to the specific mechanisms in the eye. In this way, a greater understanding of the general principles is achieved.

FURTHER READING

A list of further reading can be found online.

FURTHER READING

Abdel-Aal el, S.M., Akhtar, H., Zaheer, K., Ali, R., 2013. Dietary sources of lutein and zeaxanthin carotenoids and their role in eye health. Nutrients 5, 1169–1185.

Ambache, N., 1957. Properties of irin, a physiological constituent of the rabbit's iris. J. Physiol. 135, 114.

Andersen, J.L., Kornbluth, S., 2013. The tangled circuitry of metabolism and apoptosis. Mol. Cell 49, 399–410.

Appenzeller-Herzog, C., Hall, M.N., 2012. Bidirectional crosstalk between endoplasmic reticulum stress and mTOR signaling. Trends Cell Biol. 22, 274–282.

Asnacios, A., Hamant, O., 2012. The mechanics behind cell polarity. Trends Cell Biol. 22, 584–591.

Barabino, S., Chen, Y., Chauhan, S., Dana, R., 2012. Ocular surface immunity: homeostatic mechanisms and their disruption in dry eye disease. Prog. Retin. Eye Res. 31, 271–285.

Beebe, D.C., Holekamp, N.M., Siegfried, C., Shui, Y.B., 2011. Vitreoretinal influences on lens function and cataract. Phil. Trans. R. Soc. London B. Biol. Sci. 366, 1293–1300.

Bishop, P.N., 2000. Structural macromolecules and supramolecular organisation of the vitreous gel. Prog. Retin. Eye Res. 19, 323–344.

Bloomfield, S.A., Volgyi, B., 2009. The diverse functional roles and regulation of neuronal gap junctions in the retina. Nat. Rev. Neurosci. 10, 495–506.

Bonanno, J.A., 2012. Molecular mechanisms underlying the corneal endothelial pump. Exp. Eye Res. 95, 2–7.

Borchman, D., Yappert, M.C., 2010. Lipids and the ocular lens. J. Lipid Res. 51, 2473–2488.

Bovolenta, P., Esteve, P., Ruiz, J.M., Cisneros, E., Lopez-Rios, J., 2008. Beyond Wnt inhibition: new functions of secreted Frizzled-related proteins in development and disease. J. Cell Sci. 121, 737–746.

Bowne-Anderson, H., Zanic, M., Kauer, M., Howard, J., 2013. Microtubule dynamic instability: a new model with coupled GTP hydrolysis and multistep catastrophe. Bioessays 35, 452–461.

Boyer, N.P., Higbee, D., Currin, M.B., Blakeley, L.R., Chen, C., Ablonczy, Z., et al., 2012. Lipofuscin and N-retinylidene-N-retinylethanolamine (A2E) accumulate in retinal pigment epithelium in absence of light exposure: their origin is 11-cis-retinal. J. Biol. Chem. 287, 22276–22286.

Bresson, E., Lacroix-Pepin, N., Boucher-Kovalik, S., Chapdelaine, P., Fortier, M.A., 2012. The prostaglandin F synthase activity of the human aldose reductase AKR1B1 brings new lenses to look at pathologic conditions. Front. Pharmacol. 3, 98.

Brieher, W.M., Yap, A.S., 2013. Cadherin junctions and their cytoskeleton(s). Curr. Opin. Cell Biol. 25, 39–46.

Brizzi, M.F., Tarone, G., Defilippi, P., 2012. Extracellular matrix, integrins, and growth factors as tailors of the stem cell niche. Curr. Opin. Cell Biol. 24, 645–651.

Brockerhoff, S.E., 2011. Phosphoinositides and photoreceptors. Mol. Neurobiol. 44, 420–425.

Bron, A.J., Yokoi, N., Gaffney, E.A., Tiffany, J.M., 2011. A solute gradient in the tear meniscus. I. A hypothesis to explain Marx's line. Ocul. Surf. 9, 70–91.

Brooke, M.A., Nitoiu, D., Kelsell, D.P., 2012. Cell-cell connectivity: desmosomes and disease. J. Pathol. 226, 158–171.

Bulow, H.E., Hobert, O., 2006. The molecular diversity of glycosaminoglycans shapes animal development. Annu. Rev. Cell Dev. Biol. 22, 375–407.

Burke, J.M., 2008. Epithelial phenotype and the RPE: is the answer blowing in the Wnt? Prog. Retin. Eye Res. 27, 579–595.

Burridge, K., Wittchen, E.S., 2013. The tension mounts: stress fibers as force-generating mechanotransducers. J. Cell Biol. 200, 9–19.

Butovich, I.A., 2011. Lipidomics of human meibomian gland secretions: chemistry, biophysics, and physiological role of meibomian lipids. Prog. Lipid Res. 50, 278–301.

Candia, O.A., Alvarez, L.J., 2008. Fluid transport phenomena in ocular epithelia. Prog. Retin. Eye Res. 27, 197–212.

Chakraborty, D., Rodgers, K.K., Conley, S.M., Naash, M.I., 2013. Structural characterization of the second intra-discal loop of the photoreceptor tetraspanin RDS. FEBS J. 280, 127–138.

Cheema, A.S., Mozayan, A., Channa, P., 2012. Corneal collagen crosslinking in refractive surgery. Curr. Opin. Ophthalmol. 23, 251–256.

Cipriani, V., Leung, H.T., Plagnol, V., Bunce, C., Khan, J.C., Shahid, H., et al., 2012. Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. Human Mol. Gen. 21, 4138–4150.

Claessen, J.H., Kundrat, L., Ploegh, H.L., 2012. Protein quality control in the ER: balancing the ubiquitin checkbook. Trends Cell Biol. 22, 22–32.

Couchman, J.R., 2010. Transmembrane signaling proteoglycans. Annu. Rev. Cell Dev. Biol. 26, 89–114.

Dagenais, M., Skeldon, A., Saleh, M., 2012. The inflammasome: in memory of Dr. Jurg Tschopp. Cell Death Differ. 19, 5–12.

Dartt, D.A., 2009. Neural regulation of lacrimal gland secretory processes: relevance in dry eye diseases. Prog. Retin. Eye Res. 28, 155.

Denu, J.M., Gottesfeld, J.M., 2012. Minireview series on sirtuins: from biochemistry to health and disease. J. Biol. Chem. 287, 42417.

Desouza, M., Gunning, P.W., Stehn, J.R., 2012. The actin cytoskeleton as a sensor and mediator of apoptosis. Bioarchitecture 2, 75.

Diamond, J.S., 2011. Calcium-permeable AMPA receptors in the retina. Front. Mol. Neurosci. 4, 27.

Domogatskaya, A., Rodin, S., Tryggvason, K., 2012. Functional diversity of laminins. Annu. Rev. Cell Dev. Biol. 28, 523.

Doughty, M.J., 2012. Goblet cells of the normal human bulbar conjunctiva and their assessment by impression cytology sampling. Ocu. Surf. 10, 149.

Edwards, M.M., Lefebvre, O., 2013. Laminins and retinal vascular development. Cell Adh. Migr. 7, 82.

Efron, N., Hollingsworth, J.G., 2008. New perspectives on keratoconus as revealed by corneal confocal microscopy. Clin. Exp. Optom. 91, 34.

Eichmann, A., Simons, M., 2012. VEGF signaling inside vascular endothelial cells and beyond. Curr. Opin. Cell Biol. 24, 188.

Elagouz, M., Stanescu-Segall, D., Jackson, T.L., 2010. Uveal effusion syndrome. Surv. Ophthalmol. 55, 134.

Elhawy, E., Kamthan, G., Dong, C.Q., Danias, J., 2012. Pseudoexfoliation syndrome, a systemic disorder with ocular manifestations. Hum. Genom. 6, 22.

Fanaei, M., Monk, P.N., Partridge, L.J., 2011. The role of tetraspanins in fusion. Biochem. Soc. Trans. 39, 524.

Fischbarg, J., 2012. Water channels and their roles in some ocular tissues. Mol. Aspects Med. 33, 638–641.

Florey, O., Overholtzer, M., 2012. Autophagy proteins in macroendocytic engulfment. Trends Cell Biol. 22, 374.

Freddo, T.F., 2013. A contemporary concept of the blood-aqueous barrier. Prog. Retin. Eye Res. 32, 181.

Giorgio, M., Trinei, M., Migliaccio, E., Pelicci, P.G., 2007. Hydrogen peroxide: a metabolic by-product or a common mediator of ageing signals? Nat. Rev. Mol. Cell Biol. 8, 722.

Gnana-Prakasam, J.P., Martin, P.M., Smith, S.B., Ganapathy, V., 2010. Expression and function of iron-regulatory proteins in retina. IUBMB Life 62, 363.

Gonzalez-Fernandez, F., Bevilacqua, T., Lee, K.I., Chandrashekar, R., Hsu, L., Garlipp, M.A., et al., 2009. Retinol-binding site in interphotoreceptor retinoid-binding protein (IRBP): a novel hydrophobic cavity. Invest. Ophthalmol. Vis. Sci. 50, 5577.

Gouyer, V., Gottrand, F., Desseyn, J.L., 2011. The extraordinarily complex but highly structured organization of intestinal mucus-gel unveiled in multicolor images. PLoS ONE 6, e18761.

Govindarajan, B., Gipson, I.K., 2010. Membrane-tethered mucins have multiple functions on the ocular surface. Exper Eye Res 90, 655.

Hafezi, F., Kanellopoulos, J., Wiltfang, R., Seiler, T., 2007. Corneal collagen crosslinking with riboflavin and ultraviolet A to treat induced keratectasia after laser in situ keratomileusis. J. Cat. Refract. Surg. 33, 2035.

Hardarson, S.H., 2013. Retinal oximetry. Acta Ophthalmol 91, 1. Thesis 2.

Houtkooper, R.H., Pirinen, E., Auwerx, J., 2012. Sirtuins as regulators of metabolism and healthspan. Nat. Rev. Mol. Cell Biol. 13, 225.

Inder, K.L., Davis, M., Hill, M.M., 2013. Ripples in the pond – using a systems approach to decipher the cellular functions of membrane microdomains. Mol. Biosystems 9, 330.

Invergo, B.M., Montanucci, L., Laayouni, H., Bertranpetit, J., 2013. A system-level, molecular evolutionary analysis of mammalian phototransduction. BMC Evol. Biol. 13, 52.

Iozzo, R.V., 1998. Matrix proteoglycans: from molecular design to cellular function. Annu. Rev. Biochem. 67, 609.

Ito, S., Pilat, A., Gerwat, W., Skumatz, C.M., Ito, M., Kiyono, A., et al., 2013. Photoaging of human retinal pigment epithelium is accompanied by oxidative modifications of its eumelanin. Pigment Cell Mel Res 26, 357.

Jaqaman, K., Grinstein, S., 2012. Regulation from within: the cytoskeleton in transmembrane signaling. Trends Cell Biol. 22, 515.

Ji, F., Jung, J., Koharudin, L.M., Gronenborn, A.M., 2013. The human W42R gammaD-crystallin mutant structure provides a link between congenital and age-related cataracts. J. Biol. Chem. 288, 99.

Kadler, K.E., Hill, A., Canty-Laird, E.G., 2008. Collagen fibrillogenesis: fibronectin, integrins, and minor collagens as organizers and nucleators. Curr. Opin. Cell Biol. 20, 495.

Kannan, R., Sreekumar, P.G., Hinton, D.R., 2012. Novel roles for alpha-crystallins in retinal function and disease. Prog. Retin. Eye Res. 31, 576.

Kaushik, S., Cuervo, A.M., 2012. Chaperone-mediated autophagy: a unique way to enter the lysosome world. Trends Cell Biol. 22, 407.

Keenan, T.D., Clark, S.J., Unwin, R.D., Ridge, L.A., Day, A.J., Bishop, P.N., 2012. Mapping the differential distribution of proteoglycan core proteins in the adult human retina, choroid, and sclera. Invest. Ophthalmol. Vis. Sci. 53, 7528.

Kevany, B.M., Palczewski, K., 2010. Phagocytosis of retinal rod and cone photoreceptors. Physiol. (Bethesda) 25, 8.

Kim, C., Ye, F., Ginsberg, M.H., 2011. Regulation of integrin activation. Annu. Rev. Cell Dev. Biol. 27, 321.

Kitagishi, Y., Matsuda, S., 2013. RUFY, Rab and Rap family proteins involved in a regulation of cell polarity and membrane trafficking. Int. J. Mol. Sci. 14, 6487.

Klingberg, F., Hinz, B., White, E.S., 2013. The myofibroblast matrix: implications for tissue repair and fibrosis. J. Pathol. 229, 298.

Koch, K.W., Dell'orco, D., 2013. A calcium-relay mechanism in vertebrate phototransduction. ACS Chem Neurosci 4, 909–917.

Koob, S., Reichert, A.S., 2014. Novel intracellular functions of apolipoproteins: the ApoO protein family as constituents of the Mitofilin/MINOS complex determines cristae morphology in mitochondria. Biol. Chem. 39, 285–296.

Last, J.A., Liliensiek, S.J., Nealey, P.F., Murphy, C.J., 2009. Determining the mechanical properties of human corneal basement membranes with atomic force microscopy. J. Struct. Biol. 167, 19.

Lee, Y.S., Tresguerres, M., Hess, K., Marmorstein, L.Y., Levin, L.R., Buck, J., et al., 2011. Regulation of anterior chamber drainage by bicarbonate-sensitive soluble adenylyl cyclase in the ciliary body. J. Biol. Chem. 286, 41353.

Lee, J.G., Ko, M.K., Kay, E.P., 2012. Endothelial mesenchymal transformation mediated by IL-1beta-induced FGF-2 in corneal endothelial cells. Exp. Eye Res. 95, 35.

Leitinger, B., 2011. Transmembrane collagen receptors. Annu. Rev. Cell Dev. Biol. 27, 265.

Levin, M.H., Verkman, A.S., 2006. Aquaporins and CFTR in ocular epithelial fluid transport. J. Mem. Biol. 210, 105.

Liton, P.B., Gonzalez, P., Epstein, D.L., 2009. The role of proteolytic cellular systems in trabecular meshwork homeostasis. Exp. Eye Res. 88, 724.

Liu, W., Chun, E., Thompson, A.A., Chubukov, P., Xu, F., Katritch, V., et al., 2012. Structural basis for allosteric regulation of GPCRs by sodium ions. Science 337, 232.

Logue, S.E., Cleary, P., Saveljeva, S., Samali, A., 2013. New directions in ER stress-induced cell death. Apoptosis 18, 537.

Lorenzi, M., 2007. The polyol pathway as a mechanism for diabetic retinopathy: attractive, elusive, and resilient. Exper. Diab. Res. 2007, 61038.

Luensmann, D., Jones, L., 2012. Protein deposition on contact lenses: the past, the present, and the future. Cont. Lens Ant. Eye 35, 53.

McBrien, N.A., 2013. Regulation of scleral metabolism in myopia and the role of transforming growth factor-beta. Exper. Eye Res. 114, 128–140.

McCaig, C.D., Rajnicek, A.M., Song, B., Zhao, M., 2005. Controlling cell behavior electrically: current views and future potential. Physiol. Rev. 85, 943.

Malhotra, A., Minja, F.J., Crum, A., Burrowes, D., 2011. Ocular anatomy and cross-sectional imaging of the eye. Semin. Ultrasound CT MR 32, 2.

Mann, G.E., Yudilevich, D.L., Sobrevia, L., 2003. Regulation of amino acid and glucose transporters in endothelial and smooth muscle cells. Physiol. Rev. 83, 183.

Mao, Y., Schwarzbauer, J.E., 2005. Fibronectin fibrillogenesis, a cell-mediated matrix assembly process. Matrix Biol. 24, 389.

Massague, J., 2012. TGFbeta signalling in context. Nat. Rev. Mol. Cell Biol. 13, 616.

Mathias, R.T., Rae, J.L., Baldo, G.J., 1997. Physiological properties of the normal lens. Physiol. Rev. 77, 21.

Mathias, R.T., White, T.W., Gong, X., 2010. Lens gap junctions in growth, differentiation, and homeostasis. Physiol. Rev. 90, 179.

Maurice, D.M., 1957. The structure and transparency of the cornea. J. Physiol. 136, 263.

Maxfield, F.R., van Meer, G., 2010. Cholesterol, the central lipid of mammalian cells. Curr. Opin. Cell Biol. 22, 422.

Meek, K.M., Boote, C., 2009. The use of X-ray scattering techniques to quantify the orientation and distribution of collagen in the corneal stroma. Prog. Retin. Eye Res. 28, 369.

Meek, K.M., Hayes, S., 2013. Corneal cross-linking – a review. Ophthalmic Physiol. Opt. 33, 78.

Mercer, A.J., Thoreson, W.B., 2011. The dynamic architecture of photoreceptor ribbon synapses: cytoskeletal, extracellular matrix, and intramembrane proteins. Vis. Neurosci. 28, 453.

Mimura, T., Yamagami, S., Amano, S., 2013. Corneal endothelial regeneration and tissue engineering. Prog Retinal Eye Res 35, 1–17.

Moreau, K.L., King, J.A., 2012. Protein misfolding and aggregation in cataract disease and prospects for prevention. Trends Mol. Med. 18, 273.

Mosher, D.F., Adams, J.C., 2012. Adhesion-modulating/matricellular ECM protein families: a structural, functional and evolutionary appraisal. Matrix Biol. 31, 155.

Mostowy, S., Cossart, P., 2012. Septins: the fourth component of the cytoskeleton. Nat. Rev. Mol. Cell Biol. 13, 183.

Mueckler, M., Thorens, B., 2013. The SLC2 (GLUT) family of membrane transporters. Mol. Aspects Med. 34, 121.

Muiznieks, L.D., Keeley, F.W., 2012. Molecular assembly and mechanical properties of the extracellular matrix: a fibrous protein perspective. Biochim Biophy Acta 1832, 866–875.

Muller-McNicoll, M., Neugebauer, K.M., 2013. How cells get the message: dynamic assembly and function of mRNA-protein complexes. Nat. Rev. Gen. 14, 275.

Nagaraj, R.H., Linetsky, M., Stitt, A.W., 2012. The pathogenic role of Maillard reaction in the aging eye. Amino Acids 42, 1205.

Neisch, A.L., Fehon, R.G., 2011. Ezrin, Radixin and Moesin: key regulators of membrane-cortex interactions and signaling. Curr. Opin. Cell Biol. 23, 377.

Neuhuber, W., Schrodl, F., 2011. Autonomic control of the eye and the iris. Autonom Neurosci 165, 67.

Nickla, D.L., Wallman, J., 2010. The multifunctional choroid. Prog. Retin. Eye Res. 29, 144.

Niehrs, C., 2012. The complex world of WNT receptor signalling. Nat. Rev. Mol. Cell Biol. 13, 767.

Nightingale, T.D., Cutler, D.F., Cramer, L.P., 2012. Actin coats and rings promote regulated exocytosis. Trends Cell Biol. 22, 329.

Nikitovic, D., et al., 2012. The biology of small leucine-rich proteoglycans in bone pathophysiology. J. Biol. Chem. 287, 33926.

Okada, Y., et al., 2010. Neurotrophic keratopathy; its pathophysiology and treatment. Histol. Histopathol. 25, 771.

Orban, T., Gupta, S., Palczewski, K., Chance, M.R., 2010. Visualizing water molecules in transmembrane proteins using radiolytic labeling methods. Biochemistry 49, 827.

Palczewski, K., 2012. Chemistry and biology of vision. J. Biol. Chem. 287, 1612.

Palfy, M., Remenyi, A., Korcsmaros, T., 2012. Endosomal crosstalk: meeting points for signaling pathways. Trends Cell Biol. 22, 447.

Palkovits, S., et al., 2013. Measurement of retinal oxygen saturation in patients with chronic obstructive pulmonary disease. Invest. Ophthalmol. Vis. Sci. 54, 1008.

Palty, R., Hershfinkel, M., Sekler, I., 2012. Molecular identity and functional properties of the mitochondrial $Na^+/Ca2^+$ exchanger. J. Biol. Chem. 287, 31650.

Panaser, A., Tighe, B.J., 2012. Function of lipids – their fate in contact lens wear: an interpretive review. Cont. Lens Ant. Eye 35, 100.

Parker, R., Wang, J.S., Kefalov, V.J., Crouch, R.K., 2011. Interphotoreceptor retinoid-binding protein as the physiologically relevant carrier of 11-cis-retinol in the cone visual cycle. J. Neurosci. 31, 4714.

Peti, W., Nairn, A.C., Page, R., 2013. Structural basis for protein phosphatase 1 regulation and specificity. FEBS J. 280, 596.

Piccirelli, M., Bergamin, O., Landau, K., Boesiger, P., Luechinger, R., 2012. Vitreous deformation during eye movement. NMR Biomed. 25, 59.

Pickard, G.E., Sollars, P.J., 2012. Intrinsically photosensitive retinal ganglion cells. Rev. Physiol. Biochem. Pharmacol. 162, 59.

Prince, J., Chuck, R.S., 2012. Refractive surgery after Desçemet's stripping endothelial keratoplasty. Curr. Opin. Ophthalmol. 23, 242.

Quinlan, R.A., et al., 2013. Changes in the quaternary structure and function of MjHSP16.5 attributable to deletion of the IXI motif and introduction of the substitution, R107G, in the alpha-crystallin domain. Phil. Trans. R. Soc. Lond B. Biol. Sci. 368, 0120327.

Rabbani, N., Thornalley, P.J., 2012. Glycation research in amino acids: a place to call home. Amino Acids 42, 1087.

Rachel, R.A., Li, T., Swaroop, A., 2012. Photoreceptor sensory cilia and ciliopathies: focus on CEP290, RPGR and their interacting proteins. Cilia 1, 22.

Racioppi, L., Means, A.R., 2012. Calcium/calmodulin-dependent protein kinase kinase 2: roles in signaling and pathophysiology. J. Biol. Chem. 287, 31658.

Rask-Madsen, C., Kahn, C.R., 2012. Tissue-specific insulin signaling, metabolic syndrome, and cardiovascular disease. Arterioscl Thromb Vasc Biol 32, 2052.

Reinoso, R., et al., 2012. Topographical distribution and characterization of epithelial cells and intraepithelial lymphocytes in the human ocular mucosa. Mucosal. Immunol. 5, 455.

Rhee, I., Davidson, D., Souza, C.M., Vacher, J., Veillette, A., 2013. Macrophage fusion is controlled by the cytoplasmic protein tyrosine phosphatase PTP-PEST/PTPN12. Mol. Cell. Biol. 33, 2458–2469.

Richdale, K., et al., 2013. Quantification of age-related and per diopter accommodative changes of the lens and ciliary muscle in the emmetropic human eye. Invest. Ophthalmol. Vis. Sci. 54, 1095.

Ritch, R., 2008. Exfoliation syndrome: beyond glaucoma. Arch. Ophthalmol. 126, 859.

Rizzuto, R., De Stefani, D., Raffaello, A., Mammucari, C., 2012. Mitochondria as sensors and regulators of calcium signalling. Nat. Rev. Mol. Cell Biol. 13, 566.

Rousselle, P., Beck, K., 2013. Laminin 332 processing impacts cellular behavior. Cell Adh. Migr. 7, 122.

Saari, J.C., 2012. Vitamin A metabolism in rod and cone visual cycles. Annu. Rev. Nutr. 32, 125–145.

Sahin, E., DePinho, R.A., 2012. Axis of ageing: telomeres, p53 and mitochondria. Nat. Rev. Mol. Cell Biol. 13, 397.

Sasai, Y., 2013. Cytosystems dynamics in self-organization of tissue architecture. Nature 493, 318.

Schaffler, A., Buechler, C., 2012. CTRP family: linking immunity to metabolism. Trends Endocrinol. Metab. 23, 194.

Schlotzer-Schrehardt, U., Kruse, F.E., 2005. Identification and characterization of limbal stem cells. Exp. Eye Res. 81, 247.

Schmidt, T.M., Chen, S.K., Hattar, S., 2011. Intrinsically photosensitive retinal ganglion cells: many subtypes, diverse functions. Trends Neurosci. 34, 572.

Sebastian, C., Satterstrom, F.K., Haigis, M.C., Mostoslavsky, R., 2012. From sirtuin biology to human diseases: an update. J. Biol. Chem. 287, 42444.

Sharma, K.K., Santhoshkumar, P., 2009. Lens aging: effects of crystallins. Biochim. Biophys. Acta 1790, 1095.

Shin, K., Fogg, V.C., Margolis, B., 2006. Tight junctions and cell polarity. Annu. Rev. Cell Dev. Biol. 22, 207.

Shoulders, M.D., Raines, R.T., 2009. Collagen structure and stability. Annu. Rev. Biochem. 78, 929.

Singh, P., Carraher, C., Schwarzbauer, J.E., 2010. Assembly of fibronectin extracellular matrix. Annu. Rev. Cell Dev. Biol. 26, 397.

Sit, A.J., McLaren, J.W., 2011. Measurement of episcleral venous pressure. Exp. Eye Res. 93, 291.

Soboloff, J., Rothberg, B.S., Madesh, M., Gill, D.L., 2012. STIM proteins: dynamic calcium signal transducers. Nat. Rev. Mol. Cell Biol. 13, 549.

Song, B., Zhao, M., Forrester, J., McCaig, C., 2004. Nerve regeneration and wound healing are stimulated and directed by an endogenous electrical field in vivo. J. Cell Sci. 117, 4681.

Song, S., et al., 2009. Functions of the intermediate filament cytoskeleton in the eye lens. J. Clin. Invest. 119, 1837.

Song, M.S., Salmena, L., Pandolfi, P.P., 2012. The functions and regulation of the PTEN tumour suppressor. Nat. Rev. Mol. Cell Biol. 13, 283.

Sterling, P., 2013. Some principles of retinal design: the Proctor lecture. Invest. Ophthalmol. Vis. Sci. 54, 2267.

Strauss, O., 2005. The retinal pigment epithelium in visual function. Physiol. Rev. 85, 845.

Stringer, J.M., Barrand, S., Western, P., 2013. Fine-tuning evolution: germ-line epigenetics and inheritance. Reproduction 146, R37.

Su, X., Ohi, R., Pellman, D., 2012. Move in for the kill: motile microtubule regulators. Trends Cell Biol. 22, 567.

Subczynski, W.K., Raguz, M., Widomska, J., Mainali, L., Konovalov, A., 2012. Functions of cholesterol and the cholesterol bilayer domain specific to the fiber-cell plasma membrane of the eye lens. J. Memb. Biol. 245, 51.

Summers, J.A., 2013. The choroid as a sclera growth regulator. Exp. Eye Res. 114, 120–127.

Suresh, S., 2007. Biomechanics and biophysics of cancer cells. Acta Biomat 3, 413.

Swamynathan, S.K., 2013. Ocular surface development and gene expression. J. Ophthalmol. 103947.

Swaroop, A., Kim, D., Forrest, D., 2010. Transcriptional regulation of photoreceptor development and homeostasis in the mammalian retina. Nat. Rev. Neurosci. 11, 563.

Tang, P.H., Kono, M., Koutalos, Y., Ablonczy, Z., Crouch, R.K., 2013. New insights into retinoid metabolism and cycling within the retina. Prog. Retin. Eye Res. 32, 48.

Teng, P.Y., Wanek, J., Blair, N.P., Shahidi, M., 2013. Inner retinal oxygen extraction fraction in rat. Invest. Ophthalmol. Vis. Sci. 54, 647.

Thevenin, A.F., et al., 2013. Proteins and mechanisms regulating gap-junction assembly, internalization, and degradation. Physiology (Bethesda) 28, 93.

Thoreson, W.B., Mercer, A.J., Cork, K.M., Szalewski, R.J., 2013. Lateral mobility of L-type calcium channels in synaptic terminals of retinal bipolar cells. Mol. Vis. 19, 16.

Toivola, D.M., Strnad, P., Habtezion, A., Omary, M.B., 2010. Intermediate filaments take the heat as stress proteins. Trends Cell Biol. 20, 79.

Ujiie, H., Shibaki, A., Nishie, W., Shimizu, H., 2010. What's new in bullous pemphigoid. J. Dermatol. 37, 194.

van Reeuwijk, J., Arts, H.H., Roepman, R., 2011. Scrutinizing ciliopathies by unraveling ciliary interaction networks. Hum. Mol. Gen. 20, R149.

Varma, S.D., Kovtun, S., Hegde, K.R., 2011. Role of ultraviolet irradiation and oxidative stress in cataract formation—medical prevention by nutritional antioxidants and metabolic agonists. Eye Cont. Lens 37, 233.

Vos, W.L., et al., 2010. Expression and structural characterization of peripherin/RDS, a membrane protein implicated in photoreceptor outer segment morphology. Eur. Biophys. J. 39, 679.

Wang, J., et al., 2011. Ultra-high resolution optical coherence tomography for imaging the anterior segment of the eye. Ophthal. Surg. Lasers Imag. 42, S15.

Wenzel, A., Grimm, C., Samardzija, M., Reme, C.E., 2005. Molecular mechanisms of light-induced photoreceptor apoptosis and neuroprotection for retinal degeneration. Prog. Retin. Eye Res. 24, 275.

Wright, A.F., Chakarova, C.K., Abd El-Aziz, M.M., Bhattacharya, S.S., 2010. Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. Nat. Rev. Genet. 11, 273.

Yamada, K.M., Miyamoto, S., 1995. Integrin transmembrane signaling and cytoskeletal control. Curr. Opin. Cell Biol. 7, 681.

Yamashiro, S., Gokhin, D.S., Kimura, S., Nowak, R.B., Fowler, V.M., 2012. Tropomodulins: pointed-end capping proteins that regulate actin filament architecture in diverse cell types. Cytoskeleton (Hoboken) 69, 337.

Young, T.L., Metlapally, R., Shay, A.E., 2007. Complex trait genetics of refractive error. Arch. Ophthalmol. 125, 38.

Yu, F.X., Guan, K.L., 2013. The Hippo pathway: regulators and regulations. Genes Dev. 27, 355.

Yucel, Y.H., et al., 2009. Identification of lymphatics in the ciliary body of the human eye: a novel 'uveolymphatic' outflow pathway. Exp. Eye Res. 89, 810.

Zheng, Z., et al., 2012. Sirtuin 1-mediated cellular metabolic memory of high glucose via the LKB1/AMPK/ROS pathway and therapeutic effects of metformin. Diabetes 61, 217.

Zhou, L., Beuerman, R.W., 2012. Tear analysis in ocular surface diseases. Prog. Retin. Eye Res. 31, 527–550.

Zhou, L., Zhao, S.Z., Koh, S.K., Chen, L., Vaz, C., Tanavde, V., et al., 2012. In-depth analysis of the human tear proteome. J. Proteomics 75, 3877–3885.

Chapter 5

Physiology of vision and the visual system

- Introduction
- Light detection and dark adaptation
- Visual acuity and contrast sensitivity
- Electrophysiology of the visual system
- Colour vision
- Visual perception
- Division of labour in the visual system: parcellation and the human connectome
- Physiology of ocular movement
- Conclusion

Introduction

Vision in all its aspects could arguably be described as the most important physiological function for survival. Vision encompasses detection of luminance (and, we now understand, irradiance), contrast sensitivity (visual acuity), discrimination of texture, colour, depth and motion disparities and integrates them in to what we describe as perception. Most of these functions are located in the higher cortical centres where the retinal 'sensation' (image) is converted into our personal view (percept) of the outside world. What is seen (conceived) may not be the same as what is perceived, or even detected, and the latter may be extensively edited through input from other non-visual centres, especially memory and previous visual experience.

Zeki's notion (1992) of the brain constructing an image of the world by segregating the component parts via cortical regions that are, for instance, directionally selective (motion detectors), orientation selective, hue discriminators (colour) and that assess depth (stereopsis) has now been extended using functional magnetic resonance imaging (fMRI) to show how the segregated parts are integrated to provide a final but individual-specific image. Indeed fMRI reveals which specific areas of the brain are activated when we visually register different 'objects' (Fig. 5-1). At face value it might be thought that spatial resolution would be most important to survival but in fact it has been shown that colour is what we see best.

There are further subtleties to the business of seeing. For instance the manner in which we detect shape/form depends on more than activation of orientation-selective neurones. Remarkably, the recognition of faces and the recognition of the expression on a face are processed separately, as has been demonstrated in patients with damage to highly selective regions of the brain (Fig. 5-1). The recognition of texture is akin to a form of visual 'touch'. This chapter provides a rather simplified and brief overview of aspects of the complex sensory and psychophysical responses to visual stimuli.

DO I HAVE GOOD VISION?

A certain level of good vision is required for many daily activities, some of which may have a legal requirement, such as driving. However, 'good vision' is a variable measure and depends on the set standard. Perhaps it is more valuable to have a concept of the limits of our visual capabilities. Vision can be considered in two ways: the optical requirements to achieve an image (i.e. refraction of light by the eye to focus the image on the retina, also known as physiological optics) and the neural processing of visual stimuli by the retina and the brain. The visual process is initiated by the detection of a light signal by photoreceptor cells in the outer retina. Photoreceptor cells convert light energy to an electric stimulus, which is then

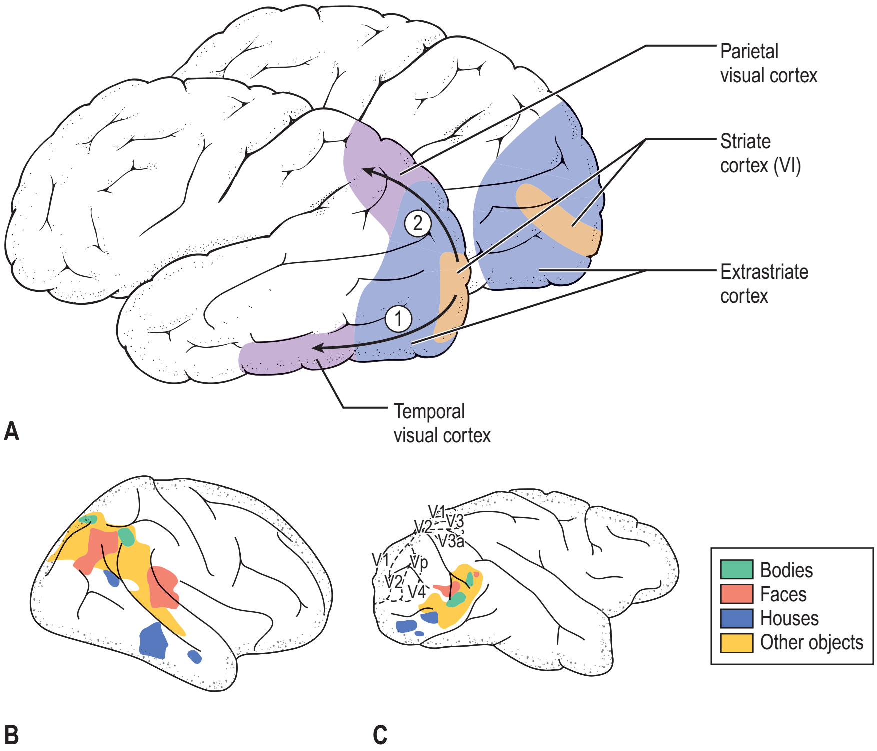

FIGURE 5-1 Typical locations of category-selective regions in the human ventral visual cortex. (**A**) The location of visual regions in the human cortex, including the primary visual cortex (area V1 in the striate cortex) and the extrastriate cortex in the occipital lobe, and the traditional distinction into two visual cortical pathways that start in area V1 and extend into the temporal lobe (the ventral 'what' or 'object-vision' pathway (1)) or into the parietal lobe (the dorsal 'where' pathway (2)). (**B**,**C**) Ventral pathway regions in one individual that were activated significantly for selectivity of bodies, faces, houses or other objects. In addition, the yellow areas represent the regions that, in a group of people (n = 9), activated significantly in the contrast: intact objects > scrambled objects. All data were processed using SPM5 (Wellcome Department of Cognitive Neurology, London). (From Op de Beeck et al., 2008.)

transmitted to the bipolar cells and onwards to the ganglion cells in the retina (see Ch. 1, p. 46). The information is further transmitted in the axons of these cells (the optic nerves which, after 50% crossover in the optic chiasma, become the optic tracts) to the visual thalamic organ, the lateral geniculate nucleus (LGN). Synaptic contact with neurones in the LGN that project to the cerebral cortex permits onward transmission of the signals via optic radiation to the visual or striate cortex (V1), where they interact with many other neuronal connections from visual cortical cells in the prestriate cortex (V3–V5), and where parcelling out and processing of the signals takes place to build up the final perceived visual image. Input is also received by the visual cortex from many other areas, particularly those controlling general motor function and eye movement, cerebellar and spatial sense, memory and many other functions located in prefrontal cortex. This produces some of what is known as 'top down' modulation of visual responses whereby signals received and interpreted in the visual cortex can be influenced by input to the final image (the perceived image) from other areas, such as visual area 4 (V4) which combines elements of object recognition with visual attention (see eFigs 5-1 and 5-8).

Additional content available at https://expertconsult.inkling.com/.

In biophysical terms a photoreceptor is capable of detecting a single photon of light (see Ch. 4, p. 261), but in practice what are the limits of detection of a visual stimulus? This depends on the nature of the stimulus and the nature of the ambient conditions in which it is presented. Sensing light is a function of all regions of the retina but the foveal region is

The brain is constantly shifting its position to meet the requirements of changes in behaviour. It continuously adapts its processing machinery to behavioural demands. Information is therefore transformed, modulated and rechannelled through different neural cortical wiring circuits which have been revealed in series of novel experiments in various animal models which can assess small changes in activity occurring at low frequency. Harris and Thiele suggest that processes involved in selective attention are similar to those involved in state changes; these are summarized in eFig. 5-1 (Harris and Thiele, 2011). They include increased activity of cortical neuromodulatory afferents (red (cholinergic), blue (serotonergic) and green (noradrenergic) arrows) which causes a general desynchronization and reduction in spontaneous fluctuation, but may lack the spatial selectivity to desynchronize the patch of cortex representing the attended stimulus. Focused glutamatergic inputs arising from feedback connections could provide this specificity (yellow arrows), causing enhanced desynchronization and sensory responses in the regions of cortex activated by the object of attention. The yellow circle in the visual display indicates the focus of attention, which affects processing in thalamic and cortical areas at specific locations (indicated by the yellow patches). The distorted replication of the visual world in the different areas illustrates (approximately) the known retinotopic organization of these different areas.

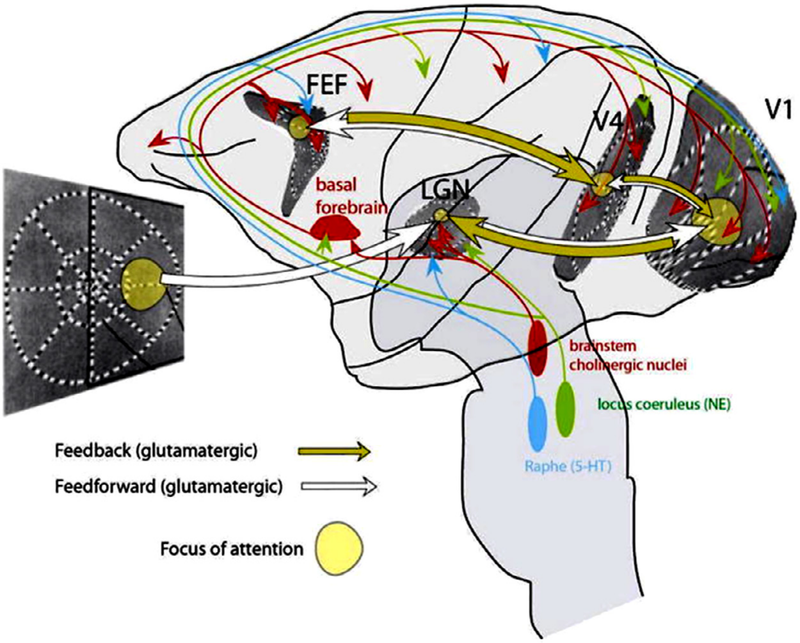

eFIGURE 5-1 Processes involved in attention are thought to be similar to changes occurring in state changes in cortical activity for specific functions. Neuromodulatory effects are identified by the red, blue and green arrows and are explained in the text. (From Harris and Thiele, 2011.)

specialized for high spatial resolution (visual acuity) and colour detection, served by the small 'midget', slow-transmitting ganglion cells (the parvocellular or P system). In contrast, luminance and motion detection are served by the large, fast-transmitting ganglion cells (the magnocellular or M system) that dominate the remaining retina and thus incorporate the entire visual field. According to Barlow's single neurone doctrine, it should take only one neurone to detect a visual stimulus (Barlow, 1972); however, the question is whether the signal received from the one rod in a thousand stimulated in the dark is sufficient to activate this neurone. Psychophysical studies have shown that both the luminance and colour thresholds for vision are different by orders of magnitude for P neurones between monkeys and humans, suggesting that more than one neurone is involved in detecting a light stimulus. In fact, continuous pooling of information occurs both in the excitatory and inhibitory neuronal activity that is present at all times and that, after a visual stimulus, the changes in the response rate of many neurones are 'sampled' by the brain until they reach a certain threshold level, at which point they register and the stimulus is 'recognized' (Hurlbert and Derrington, 1993). This perhaps explains how we can sometimes look at an object and yet not 'see' it; furthermore, these psychophysical considerations are highly relevant to methods for testing vision, for instance with regard to setting luminance thresholds for studies of visual fields using small transient targets.

Whether or not we have good vision at any moment in time depends on our level of awareness, consciousness and attention to visual stimuli which have many properties such as depth, shape, form, colour texture and more besides, each with its own rate of detection/discrimination.

Flicker can be used to determine limits of vision

Detection of a stationary target or spot depends on the size and brightness of the spot relative to the background. The limits of detectability of the target are therefore determined by the spatial resolution and the anatomical relationships between stimulated receptors (see below). Spatial resolution is highest at the fovea and declines sharply towards the peripheral retina; this is clearly demonstrated by the detection threshold at different eccentricities in the visual field.

The threshold for spatial resolution is, however, considerably higher than that for detecting light; this latter parameter can be measured by flicker detection, which is the ability to detect two stimuli separated in time. This function is normally subserved by rod photoreceptors, while spatial resolution is subserved by cones, with some input from rods.

The critical flicker fusion (CFF) frequency test may be a useful predictor of cataract surgery outcome in cases of co-morbidity of lens opacity and macular disease because the CFF (see below) is relatively unaffected by image degradation due to cataract but would be affected by foveal disease. In addition, as a neurophysiological test, it has been used in the early detection of hepatic encephalopathy.

Motion detection is also a feature of rod vision

It is clear, therefore, that the ability to detect a standard small bright spot in specified regions of the visual field is in fact a much more complex task than would at first appear. Not only does it depend on the absolute brightness of the stimulus but also on the background on which the stimulus is presented and thus on contrast. It also depends on whether the target is moving or stationary and, if stationary, for how long the target is presented. Its detection depends on the density of photoreceptors and thus the region of retina stimulated. If it is a moving target it will stimulate different cortical neurones depending on which direction it is moving. This functional segregation of visual input is retained at several levels within the cortex before construction of the final visual image.

SENSING COLOUR

Cone photoreceptors are built to sense colour, which they do through cone opsin proteins. There are three types in humans: long (L, red), medium (M, green) and short (S, blue) wavelength cones, each with its own specific opsin (see p. 302). Early colour matching experiments in which the colour of a test stimulus is matched by adding together stimuli composed of the three primary colours verified that three colours were all that was required to detect the full spectrum of white light (see below, Colorimetry, p. 302). This is known as the trichromatic theory of colour vision. Each colour has properties such as hue and chromaticity.

There are many hues but only three primary colours

Hue is an idealized term for the colour produced by light of a single wavelength. In spite of having only three cone photoreceptors, we are able to distinguish many hues of colour, e.g. lilac and violet. It is therefore clear that any single colour is recognized by a mixture of the three primaries and that there must be overlap in the spectral sensitivity for each primary colour (see below). Theoretically it should be possible to produce light of a single wavelength using a narrow slit on a device such as a monochrometer, but photoreceptor sensitivity is also subject to the intensity of the light, and narrow wavebands of this degree of selectivity are not sufficiently intense to produce a stimulus.

The hue-discrimination curve (Fig. 5-2) describes the physiological limits at which a shift in wavelength can be discriminated as a change in colour. It was derived empirically by using fixed amplitude selective wavelength stimuli (1.5 cycles/300 nm), which found that there were two peaks of discrimination, one in the yellow/orange and a second in the blue/violet range. Monochromatic light therefore is not a practical reality; most colours are in fact tints, i.e. they are unsaturated hues, the degree of unsaturation being determined by the amount of additional white light they require to match them to a hue.

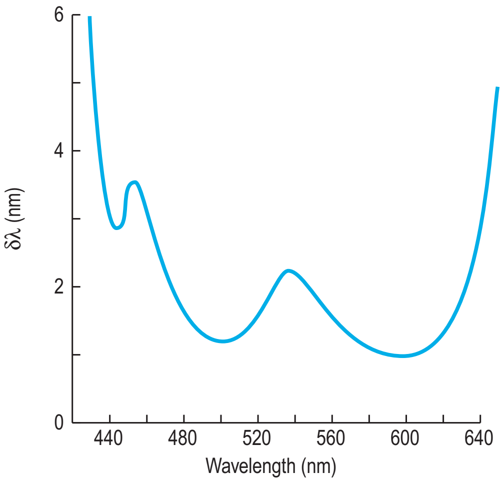

FIGURE 5-2 Hue-discrimination curve comparing wavelength discrimination (*y*-axis) with changing wavelength (*x*-axis). Discrimination of hues varies for any given wavelength, being best at 455 and 535 nm.

Chromaticity is semiquantified 'colouredness'

Chromaticity refers to 'colouredness' and depends on hue, saturation and intensity of light (luminosity). Indeed, hue itself is not independent of the luminosity of the stimulus and chromatic shifts occur as the intensity increases until all hues appear yellow–white (the Bezold–Brucke phenomenon) or as the intensity decreases, when all hues appear achromatic (the Purkinje shift; see below). Any colour can thus be matched by a mixture of the three primary colours plus or minus a proportion of white light to account for unsaturation; these are formally described in the chromaticity chart and can be determined at different levels of lower or higher colour metrics (discrimination) (Fig. 5-3). The International Commission on Illumination (CIE) has developed standard colour 'observers' (colorimeters) which have proved valuable in defining 'true' colours but do not fully take into account the minimally perceptible differences in

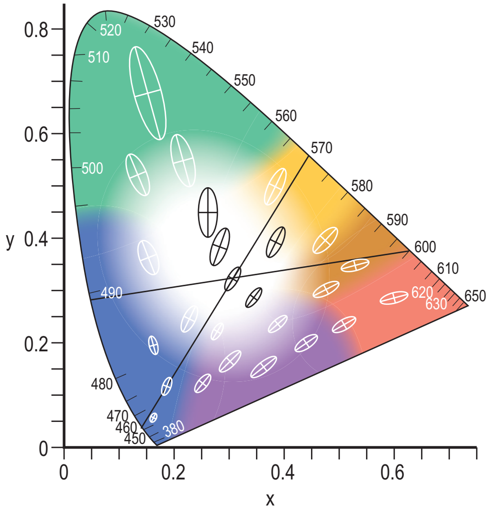

FIGURE 5-3 Lower versus higher colour metrics. The chromaticity diagram is long established in the field of colour discrimination. It represents the laws of colour mixing in terms of (x,y). This is the domain of lower colour metrics. The straight lines through x = 0.305, y = 0.323 indicate that this white colour can be obtained, for instance, by mixing, in suitable quantities, light of wavelengths 570 and 465 nm or of wavelengths 600 and 489 nm. The ellipses, drawn here at 10 times their true size, are contours of just noticeably different colours from their central colour. The description of the differences in size, shape and orientation of these 'JND' ellipses is the domain of higher colour metrics. (From Vos, 2006.)

colour discrimination (just noticeable differences, JND) based on aspects of photoreception affected by ‘background noise’.

However, the chromaticity chart has true practical value, for instance in colour-mixing techniques used routinely in computer programs for producing different colours digitally for image creation and other purposes, and have been developed into a computerized colour vision test. Clinically, colour vision can be tested using hue-discrimination techniques (e.g. the Farnsworth–Munsell 100 hue test) and normal values vary with age (affected by JND effects), peak ability occurring around the age of 19 years. Some effect by rods on cone vision has also been shown by rod function studies (‘background noise’).

SHAPE, FORM AND DEPTH PERCEPTION (AND MORE) HELP TO ‘SHAPE’ VISION

The discrimination of shape and form is highly developed in primate vision and the cortical localizations which define these functions are now well established. The appreciation of the complexity and sophistication of this aspect of visual perception has in part developed from the realization that the brain recognizes and can categorize objects according to shape irrespective of the angle or distance from which they are viewed, the ambient lighting conditions, or other factors.

Shape processing is achieved by specialized orientation-sensitive cells in the visual cortex, but the extra dimension of form recognition, as in recognition of facial features, requires additional processing. Studies of patients with specific visual defects such as prosopagnosia (inability to recognize familiar faces), however, are powerful indicators of the localization of visual functional sites. Even this apparently specific defect can be subdivided into a perceptual form and an associative form, the latter arising when the patient can perceive the image but cannot draw on visual memory sufficient to ‘remember’ the face. These separate functions have been ascribed to different regions of the brain, although associative face recognition may also involve other senses such as voice recognition (Fig. 5-4).

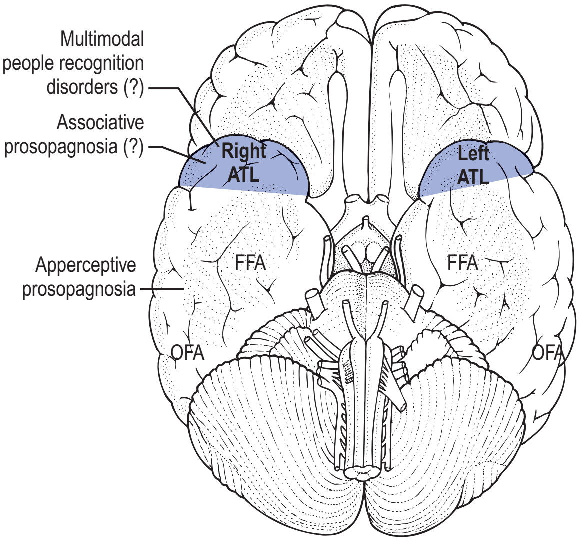
Multimodal people recognition disorders (?)
Associative prosopagnosia (?)
Right ATL
Left ATL
Apperceptive prosopagnosia
FFA
FFA
OFA
OFA

FIGURE 5-4 Processing information for features (of the face) involves a specific neural cortical network, the occipital face area (OFA), the face fusiform area (FFA) and the anterior temporal lobe (ATL), which have a right hemisphere dominance. This region underlies neurological perceptual defects such as prosopagnosia (lesion in the posterior region) as well as person recognition disorders (lesions in the anterior temporal lobe). (From Gainotti, 2013.)

Much is also known about depth perception. For instance it has been shown that there are specific cortical cells responsive only to disparate but simultaneous orientations of an object, presented to non-corresponding regions of the retina. In addition, clues on the nature of 3D structures can be obtained from motion detection (structure-from-motion). However, although no specific ‘depth appreciation’ cortical region has been identified, the lateral occipital cortex is a favoured area showing much activity. The appreciation of depth is more than simply stereopsis and is built up from many other cues (see below). An understanding of how these modify perception can only come after a description and appreciation of the different types of visual stimuli that separately induce discrete responses in the brain.

Light detection and dark adaptation

WHAT ARE THE LIMITS OF DETECTABLE LIGHT?

As in all biological systems, there is no precise answer to the above question. Light energy comes in quanta (small packets) and it has been estimated that between 50 and 150 quanta of light are required to strike the cornea for a discrete signal to be detected. Of these quanta, only about 10% actually reach the photoreceptors. The detection of this stimulus is not simply a function of photoreceptor stimulation but is subject to 'dark light' (effectively background noise in rhodopsin photoisomerization) and is also dependent on higher neural function, and the concept of a visual threshold is more or less a statistical function dependent on how large the stimulus has to be to reach a level of recognition. This is well recognized by anyone who performs a visual field test using an automated visual field analyser.

Thresholds and the frequency of seeing

A distinction must be made between the theoretical estimate of the number of quanta required to produce an electric stimulus in a patch-clamped photoreceptor cell and the psychophysical conversion of the light stimulus to a perceived sensation. The latter depends on a defined measure, termed the 'frequency of seeing', which is the number of times a repetitively presented minimal stimulus is detected, and is a probability function that varies between and within individual observers.

The former is theoretically a single photon of light. However, there is considerable 'noise' in the system owing, for instance, to random opening and closing of ion channels as a result of thermal isomerization of rhodopsin, or to scatter from background and/or stray light energy from the stimulus itself. These effects can account for up to 1000 quanta/degree, which is well above the absolute threshold for light stimulation. It is thought that some of this is 'smoothed' by coupling between photoreceptors.

What is the minimal stimulus for vision?

Even when theoretical biophysical considerations such as signal-to-noise ratio are taken into account, this deceptively simple question depends on many factors such as background illumination, spatial frequency, summation, wavelength, dark adaptation and optical qualities of the image-gathering system. The specific conditions have to be stated, therefore, before this quantity can be expressed.

In addition, consideration of whether a single rod can detect a single photon of light *in vivo* has to take into account the different routes that a rod can take to stimulate a ganglion cell and convert this into a behavioural response (see pp. 258–262). Furthermore, the minimal stimulus size for a cone is also an important measure to define with physiological and clinical relevance.

DARK ADAPTATION CURVE AND RETINAL SENSITIVITY

The minimum visual stimulus varies depending on ambient light conditions, i.e. whether the stimulus is viewed in the dark or under normal/bright light conditions. In the dark, the eye becomes progressively more sensitive to light stimulation until the light threshold reaches a minimum after about 30 minutes. This is demonstrated in the dark adaptation curve (Box 5-1), which has two components: an early one resulting from increases in the cone sensitivity and a second

BOX 5-1 DARK ADAPTATION

The normal dark adaptation curve (a) varies if the conditions are varied: with a very small central white target, rods fail to become stimulated at all and the curve flattens out (b). If the cones are first light adapted by weakly stimulating them to maximum sensitivity or by adapting subjects to red light before placing them in the dark, the cone component can be 'lost' (c); subjects without cone vision also have no cone component (rod monochromats).

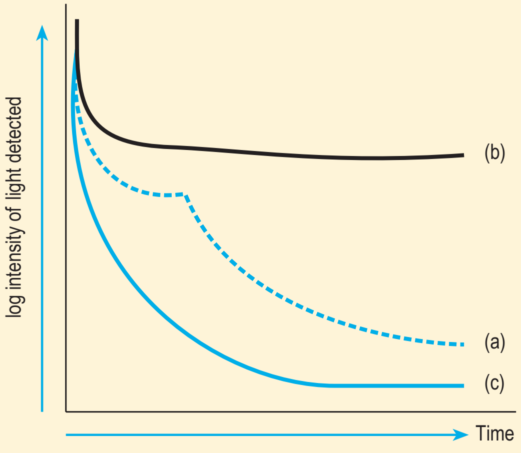

Figure outlining dark/light adaptation responses: (a) mixed rod and cone response of physiological dark adaptation, (b) pure cone response, (c) pure rod response.
(Figure courtesy of H. Dawson.)

produced by increases in rod sensitivity. There is also a light adaptation curve for cones in which the sensitivity to light varies as the luminosity increases or decreases within a wide range of high ambient illumination (see below). Thus the shape of the curve can be varied by altering the conditions.

Light and dark adaptation are the psychophysical correlates of visual pigment bleaching and regeneration, and can be measured by reflection densitometry. This technique is based on the assumption that light reflected from the unbleached retina will contain lower amounts of 500 nm (peak sensitivity for rods) light than that reflected from the bleached retina, since there will be considerable absorption of 500 nm light by the dark-adapted retina. Reflection densitometry studies permit an evaluation of the photosensitivity of the retina, i.e. the rate at which bleaching takes place for a given intensity of illumination. It has been estimated that the normal retina absorbs 50% of the quanta of light striking the retina but, as discussed above, this is not necessarily associated with a perceived visual stimulus because the absorption by a single rod of a photon of light can have at least three outcomes.

Regeneration of rhodopsin after dark adaptation is slow, taking 30 minutes for completion, with a half-time in humans of 5 minutes. This varies significantly between species. Clearly, the sensitivity of the retina in any individual will depend on the total amount of rhodopsin, and this relationship has been delineated in the Dowling–Rushton equation:

$$\log(z)/A = aB$$

where A is the threshold in complete dark adaptation, B is the fraction of bleached rhodopsin and a is a constant of proportionality. This sort of mathematical relationship has been used to estimate the rhodopsin content of the retinas of patients with certain forms of retinal disease, such as Oguchi's disease, fundus albipunctatus and especially vitamin A deficiency. However, it is important to realize that receptor sensitivity and rhodopsin content are not equivalent and that sensitivity to light is markedly reduced after partial bleaching, long before there is a reduction in rhodopsin content. This is clear in the isolated retina where photosensitivity is permanently reduced even after full recovery in the dark. These changes reflect the level of rhodopsin intermediates (metarhodopsin I and II) in the retina, which remain after bleaching (see Fig. 4-70, p. 259).

These effects are important in the determination of photosensitivity of discrete regions of the retina where it has been shown that reduced sensitivity can be detected in regions of the retina not exposed to point sources of light. Although this has been attributed to light scatter, there are probably other mechanisms operative here, particularly related to convergence of neural input (see below).

What does adaptation mean at a molecular level? There is considerable evidence to show that dark adaptation and regeneration of rhodopsin are dependent on the local concentration of 11-*cis* retinal, and the limiting factor for recovery after a large bleach is the rate at which 11-*cis* retinal is delivered to opsin in the bleached photoreceptors (Fig. 5-5) despite some more recent evidence that some of the retinoid conversion steps occur in the Müller cells also (see Ch. 4, Fig. 4-69). Thus, because a healthy retinal pigment epithelium is central to this process, age-related decline in dark adaptation can be explained on this basis. It is also dependent on termination of the

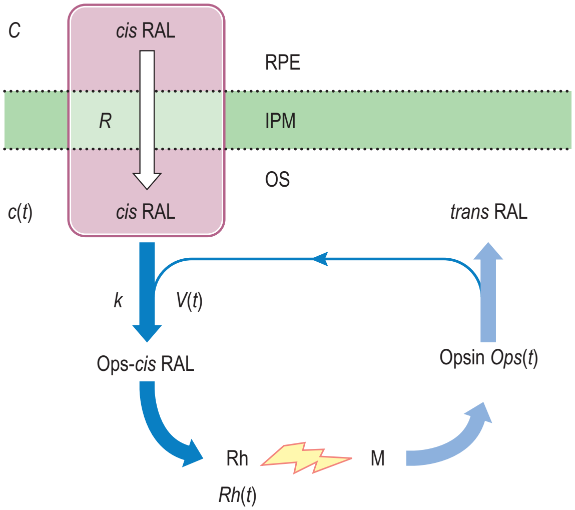

FIGURE 5-5 Schematic of the MLP rate-limited model. Removal of photoproduct and regeneration of visual pigment is rate-limited by the delivery of 11-*cis* retinal (*cis* RAL) from the retinal pigmented epithelium (RPE) to opsin in the outer segment (OS). IPM, interphotoreceptor matrix; Rh, rhodopsin; M, metarhodopsin. (From Lamb and Pugh, 2004, with permission from Elsevier.)

photon-induced signal brought about by efficient phosphorylation of the enzyme rhodopsin (Ch. 4, p. 262) via rhodopsin kinase, with the subsequent docking of arrestin to the complex, so that free opsin can be made available to bind more 11-*cis* retinal and respond to a new photon. Absence of rhodopsin kinase (also known as G protein-coupled receptor kinase 1, Grk1) or of arrestin, underlies the pathology of stationary night blindness (Oguchi's disease) while absence of the equivalent cone opsin kinase (Grk7) causes enhanced S cone syndrome.

Psychophysical evaluation of rhodopsin bleaching has been experimentally tested in humans by comparing a range of dark adaptation curves to different background light levels with the amplitude of the a wave of the electroretinogram, also known as the rod current (see p. 288) since both are desensitized to varying degrees by the amount of ambient light. It appears that rhodopsin regeneration and a-wave recovery rates match well.

Cones also regulate their sensitivity in photopic conditions, but it is much more difficult to saturate this response, i.e. cones still adapt at high intensities of steady illumination and the recovery time is very short (100 ms compared to 20–30 minutes for rods). This is probably furnished by Müller cell-derived 11-*cis* retinal, also under the control of RPE 65 in the Müller cell (see Ch. 4, Fig. 4-69).

In summary, adaptation is exactly what it means: that the retina rapidly adapts to changes in background illumination such that it can respond to increasingly strong or weak stimuli. However, the dynamic range of responses (normally a range from zero to a few hundred impulses per second) over which it functions at any specific level of illumination remains the same and the intensity of the response when it makes one is also the same. Put simply, the retina adapts rapidly to new lighting conditions when there is plenty of light about but slowly when light levels are low.

Melatonin and circadian rhythms

The circadian clock is a process whereby genes regulating various functions such as the sleep–wake cycle, body temperature, immune cell function and behaviour are expressed in a rhythmical manner. At least 11 core clock genes have been discovered, including a set of period genes (*PER* 1,2,3) and clock genes (*CLOCK*). These genes have multiple downstream effects on other important regulatory transcription factors such as POP$\alpha/\beta/\gamma$ important in immune cell function and genes involved in the synthesis of melatonin in the synthesis of melatonin. The light/dark (sleep/wake) cycle, generated by pacemaker cells in the suprachiasmatic nuclei, drives the production of the pineal gland secretory product, melatonin. Melatonin may also be produced at other sites, including the retina and bone marrow. It is synthesized from tryptophan via serotonin in two steps involving the enzymes serotonin-*N*-acetyltransferase (NAT) and hydroxyindole-*O*-methyltransferase (HIOMT) (see Ch. 4, p. 248). Melatonin provides information to the organisms to permit organization of various physiological functions and, because it can adapt to night length, it can promote a seasonal (photoperiod) as well as a diurnal rhythmicity (Fig. 5-6). Apart from its obvious physiological functions, such as sleep–wake patterns, melatonin influences immune diurnal variations in innate immune defence functions such as antioxidation, glucose regulation, blood coagulation enzyme systems and ocular functions such as control of aqueous secretion.

Melatonin is a methoxyindole, synthesized and secreted principally by the pineal gland at night under normal environmental conditions and binds to two receptors (M1 and 2). The endogenous rhythm of secretion is generated by the suprachiasmatic nuclei and entrained to the light/dark cycle. Light is able to either suppress or synchronize melatonin production according to the light schedule. The nyctohemeral rhythm of this hormone can be determined by repeated measurement of plasma or saliva melatonin or urine sulphatoxymelatonin, the main hepatic metabolite.

The primary physiological function of melatonin, whose secretion adjusts to night length, is to convey information concerning the daily cycle of light and darkness to body physiology. This information is used for the organization of functions, which respond to changes in the photoperiod (seasons). There is still, however, only limited evidence for seasonal rhythmicity of physiological functions in humans related to possible alteration of the melatonin message in temperate areas under field conditions, although there is

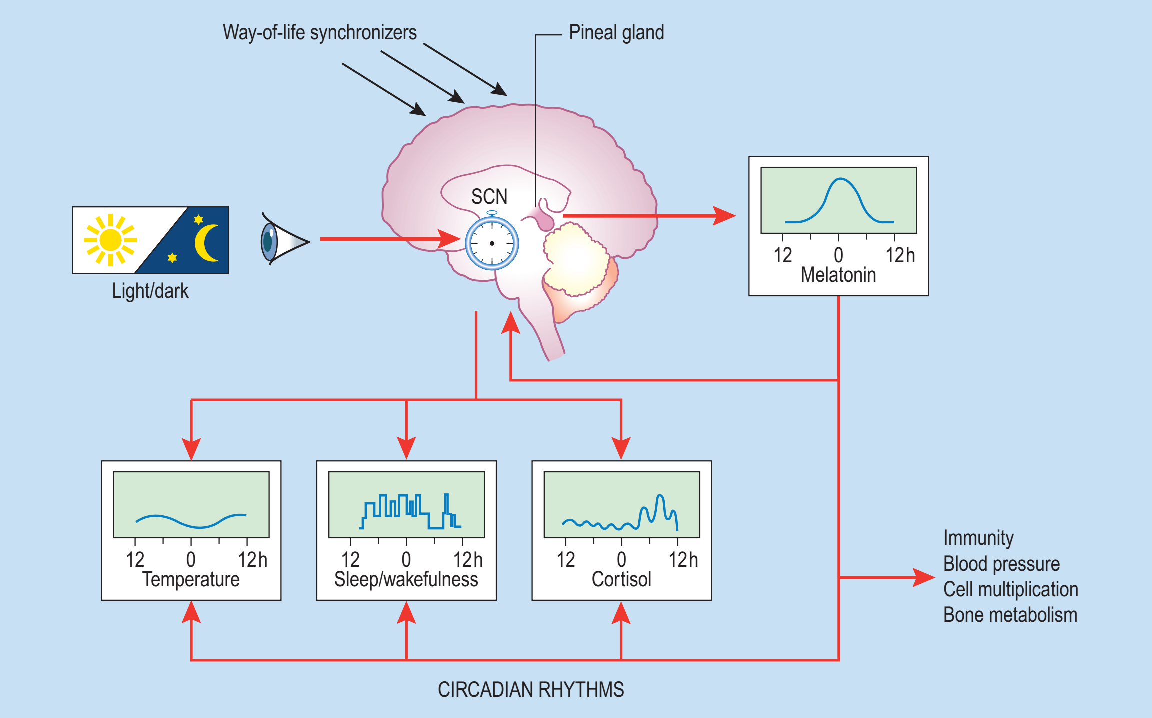

FIGURE 5-6 Melatonin acts as an endogenous synchronizer. (From Claustrat et al., 2005, with permission from Elsevier.)

a reported link between seasonal affective disorder, clinical depression and its control with novel antidepressant drugs based on melatonin receptor stimulation, such as agomelatine. Major clinical depressive illness has also been linked to markedly reduced function of the core clock genes in the brain in a recently reported post-mortem microarray analysis.

The daily melatonin secretion, which is a very robust biochemical signal of night, can be used for the organization of circadian rhythms. Although functions of this hormone in humans are mainly based on correlative observations, there is some evidence that melatonin stabilizes and strengthens the coupling of circadian rhythms, especially of core temperature and sleep–wake rhythms. As the regulating system of melatonin secretion is complex, following central and autonomic pathways, there are many pathophysiological situations where the melatonin secretion can be disturbed. The resulting alteration could increase predisposition to disease, add to the severity of symptoms or modify the course and outcome of the disorder.

Melatonin is also produced by photoreceptors where it can act on melatonin receptors (MRs) in an autocrine manner, as well as on MR+ ganglion cells and other retinal neural cells in a paracrine manner. Thus it regulates the activity of photoreceptors, it acts on horizontal cells stimulated by cones to reduce their responsiveness, but heightens ON-bipolar cells and ganglion cells in some species. In this way melatonin is thought to fine-tune visual function, especially in cone cells under varying ambient light conditions (see eFig. 5-2).

Additional content available at https://expertconsult.inkling.com/.

Interestingly MRs are present on many other ocular tissue cells such as ciliary epithelium, RPE cells, lens cells, corneal endothelium and keratocytes, and stromal cells in the sclera and choroid.

ARE TWO SMALL STIMULI EQUIVALENT TO ONE LARGE ONE (SUMMATION)?

The threshold for light detection can be measured arbitrarily by setting certain conditions of stimulus

This occurs via activation of specific receptors on amacrine, horizontal and photoreceptor cells. Wiechman has proposed a working hypothesis for melatonin paracrine signalling in the retina (eFig. 5-2). Melatonin is normally produced by photoreceptors at night, and diffuses to target cells within the retina that have specific receptors on cells such as GABA-ergic and/or dopaminergic amacrine cells which work in a reciprocal manner to some degree since GABA inhibits dopamine release from amacrine cells. A lower rate of dopamine release from amacrine cells results in lower stimulation of D1 receptors on horizontal cells, which in turn leads to increased coupling of horizontal cells. This would result in an increase in receptive field size and increased sensitivity to light. Lower levels of binding of dopamine to D2 receptors on photoreceptor cells induces an increase in melatonin synthesis. Meanwhile melatonin may bind to horizontal cells to directly inhibit the cellular response to D1 receptor binding. Melatonin may also bind to receptors located on the photoreceptor membrane, which could directly increase rod sensitivity to light, and/or regulate synthesis of melatonin.

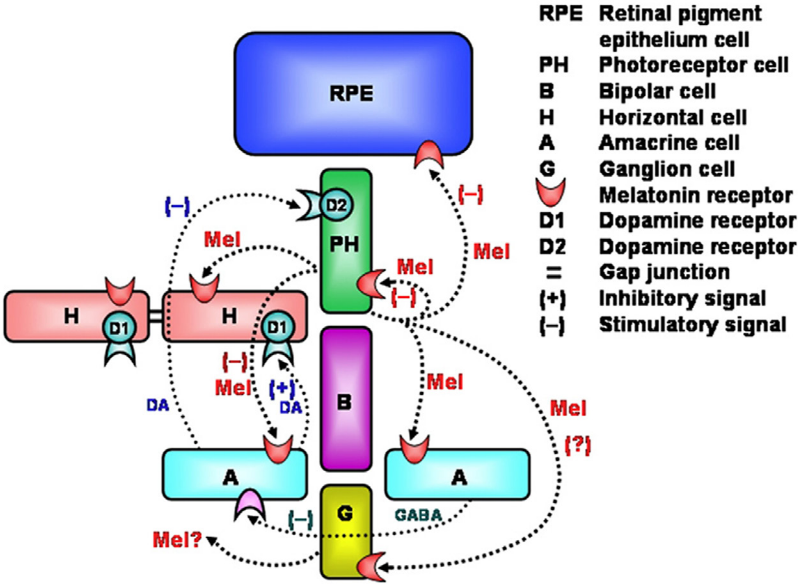
RPE
D2
PH
(–)
Mel
Mel
(–)
Mel
(–)
H
D1
H
D1
(–)
Mel
(+)
DA
DA
B
Mel
Mel
(?)
A
A
G
(–)
GABA
Mel?
RPE Retinal pigment epithelium cell
PH Photoreceptor cell
B Bipolar cell
H Horizontal cell
A Amacrine cell
G Ganglion cell
Melatonin receptor
D1 Dopamine receptor
D2 Dopamine receptor
= Gap junction
(+) Inhibitory signal
(–) Stimulatory signal

eFIGURE 5-2 Diagram outlining mechanism of how melatonin fine-tunes visual function (see text for details). (From Wiechmann et al., 2008.)

size, brightness, pupil size and level of background illumination, and recording how often a subject detects the stimulus. An empirically set level of 'hits' or positive detection responses (e.g. 55%) can then be set and expressed in trolands (Box 5-2). Experimentally it has been estimated that at the limit of light detection in the fully dark-adapted eye, the retina is illuminated to a level of 4.4×10^{-5} trolands, which is equivalent to the stimulation of only 1/5000 rods per second. However, if the light is concentrated on one area it will more readily elicit a response and it therefore becomes less practical to think of light energy in terms of area of retinal illumination; instead the minimum flux in light energy required to induce a detectable response is commonly accepted as the threshold and is around 120 quanta per second or, if the stimulus is instantaneous, between 5 and 15 quanta of light.

From this it is clear that stimulation of a single rod is insufficient to produce a visual sensation (even though an electrical response may occur in terms of a change in hyperpolarization of the cell membrane). Approximately 10–15 rods must be stimulated and the summed response must be collected either at the bipolar cell level or within the ganglion cells to induce a visual sensation. These determinations are approximate as fluctuations occur at all levels from the stimulus itself to the responses in each of the different cell types, and the final analysis is based on probabilities of a response taking place.

BOX 5-2 LIGHT ENERGY

Light energy is measured subjectively by its 'brightness', and luminance can be measured in:
- trolands
- candelas
- luxes.

Specific measures of brightness are as follows:

(a) Intensity of illumination of a surface (L) = intensity of the light source/square of the distance between the source and the surface.

$$L = I/r^2$$

(b) Unit of L = foot-candela or metre-candela
- 1 lux = 1 metre-candela
- 1 phot = 1 cm-candela
- 1 lambert = 1 candela at 1 cm distance for a perfectly diffusing light source on a surface at 1 cm
- 1 troland = a unit of retinal illumination that results when a surface luminance of 1 candela/m^2 is viewed through a pupil area of 1 mm^2
- 1 lumen = one unit of flux C, the spherical illumination from a point source of light of intensity of 1 metre-candela or 1 foot-candela.

Spatial summation

As indicated above, the empirical determination of the absolute threshold of light detection depends on the stimulus size; therefore, spatial summation must be important in setting this threshold. Each ganglion cell has a receptive field in which a light stimulus falling on a point within that field will produce a response. Receptive fields are the result of convergence of several photoreceptors to synapse with one bipolar cell and of several bipolar cells to synapse with a single ganglion cell (see next section).

Some limited general rules have therefore emerged concerning summation. Ricco's law states that the threshold intensity of a stimulus is inversely proportional to the area of the stimulus, provided the total stimulus area is sufficiently small to fit within the receptive field of a single ganglion cell. In terms of quanta, however, the amount of energy is independent of the area. As the receptive field size increases at greater distances from the fovea, Ricco's law also varies in the area in which it can be applied. In overlapping receptive fields, Ricco's law applies only partially in that larger stimuli require more quanta to reach the absolute threshold. This has led to further attempts to formulate equations that would provide a general solution for these phenomena, but in practice no simple solution covers all possibilities and summation is best explained by probability theory (see above).

Temporal summation

When the retina is stimulated in rapid succession by a target, the level of response is the same as when the target is presented continuously for the same total period of time. This is known as temporal summation and is formulated by Bloch's law, which states: the intensity of the threshold stimulus is inversely proportional to the duration of the stimulus. To a degree this is a difficult psychophysical measurement since at very short intervals it is difficult to distinguish different contrast and duration. Bloch's law holds true only for a defined period of time as, if the interval between the stimuli were long, the effect would be rapidly lost. Bloch and others found that, in fact, a 'plateau' effect was observed. However, Broca and Sulzer found that there was a peak in perception and then there was a decay before a plateau effect (see eFig. 5-3).

Additional content available at https://expertconsult.inkling.com/

In practice the peak described by Broca and Sulzer occurred at about 50–100 ms, and beyond this time there is still some degree of summation, known as partial summation, which decays exponentially.

Recently, the question of temporal summation has been revisited in the context of artificial light (which accounts for about 20% of our energy consumption). Using an experimental design in which intrinsic bias from previous learned experience was omitted, Broca's peak was detected and was attributed to a differentiation in duration versus contrast which is eliminated probably by higher neural mechanisms ensuring that the same rapid stimulus is identified with short flashes of light. Bloch's law therefore represents a smoothed-out perception in which the peak of detection/contrast is eliminated by prior learned experience at a subconscious level. If artificial lighting systems were optimally tuned to these temporal summation effects in human vision, for instance by using DC light-emitting diodes, a 20% saving in energy consumption has been estimated, which is not insubstantial.

Detection of minimum stimulus for motion displacement

An extension of these concepts has been developed to evaluate the minimum detectable motion stimulus test, since motion detection is a major function of the magnocellular ganglion cell, i.e. rod-dominated pathway (see below). The test (motion displacement test, MDT) is based on the minimum positional displacement of a standard line stimulus, which is detected as a sensation of motion. Since it is based on a square wave stimulus which oscillates back and forth between two points, it is considered to be the summation of the ON–OFF receptive field responses of the stimulated M cells. The threshold for detection varies with the square root of the stimulus energy and has been described as a new law: namely, the threshold energy displacement law (TED).

Binocular summation

Summation may occur in visual stimuli received by corresponding retinal regions when using both eyes. In practice, mostly because of optical aberrations (see below), the effect is not considered significant. However, it can be demonstrated using wavefront technologies to remove aberrations and indeed it has been shown that such aberrations account for between 5% and 15% of loss in visual discrimination. In a recent study, binocular summation and inhibition, defined as seeing five or more or fewer than five letters on the ETDRS visual acuity chart with both eyes compared with best visual acuity with each eye individually, occurred at a prevalence rate of 21% and 2%, respectively, which has considerable relevance to driving vision. In addition, the effect of amblyopia may be such that the loss of binocular summation has a distinct effect on overall visual acuity.

Visual acuity and contrast sensitivity

VISUAL ACUITY IS NOT SIMPLY A FUNCTION OF CONE ACTIVITY

Vision varies for each individual because of refractive errors and visual physiologists have therefore restricted discussion of normal visual physiology to the emmetropic idealized eye (see eFig. 5-4).

Additional content available at https://expertconsult.inkling.com/.

Visual acuity is a measure of the ability to discriminate two stimuli separated in space. Clinically, this is determined by discriminating letters on a chart, but this task also requires recognition of the form and shape of the letters, processes that involve higher centres of visual perception. Discrimination at a retinal level may therefore be determined by less complex stimuli such as contrast sensitivity gratings. The visual processes that allow discrimination between letters and gratings are fundamentally the same, with finer resolution contrast discrimination at lower luminance levels being provided by some newer test charts, such as the Mars contrast sensitivity charts which are graded in log 0.04 units. Charts that have been customized to test visual acuity in different groups of people, such as the SKILL test (Smith–Kettlewell Institute Low Luminance test), may be a good predictor of eventual development of macular degeneration in older people.

Investigation of potential different effects can be modelled experimentally. For instance the following experiments reported recently by Rieiro and colleagues demonstrate the actual effects of the different stimuli as varying potential outcomes (Rieiro et al., 2012). In eFig. 5-3, (**A**) represents two competing models of temporal vision. Bloch's law postulates a monotonic increase in perceived contrast with increased duration, whereas the Broca–Sulzer effect postulates a peak in perceived contrast with increased duration. In (**B**) subjects fixated on a central cross, and two Gabor patches flashed in succession on opposite sides of the screen. Following stimulus presentation, subjects reported which Gabor had higher contrast. Part (**C**) shows the physical contrasts and stimulus durations used for the comparator and standard stimuli. In the unblocked experiment, all possible combinations were randomized. In the blocked experiment, the different conditions were grouped into four sequential sets of trials or blocks, each with a constant comparator contrast and standard duration and an internally randomized trial sequence. Finally, (**D**) shows the psychometric curve models of the two possible experimental outcomes, colour-coded for different comparator durations. If contrast perception has a peak, as in the Broca–Sulzer effect, the curves will first shift right and then left as stimulus duration increases. If contrast perception follows Bloch's law, the curves will shift monotonically to the right.

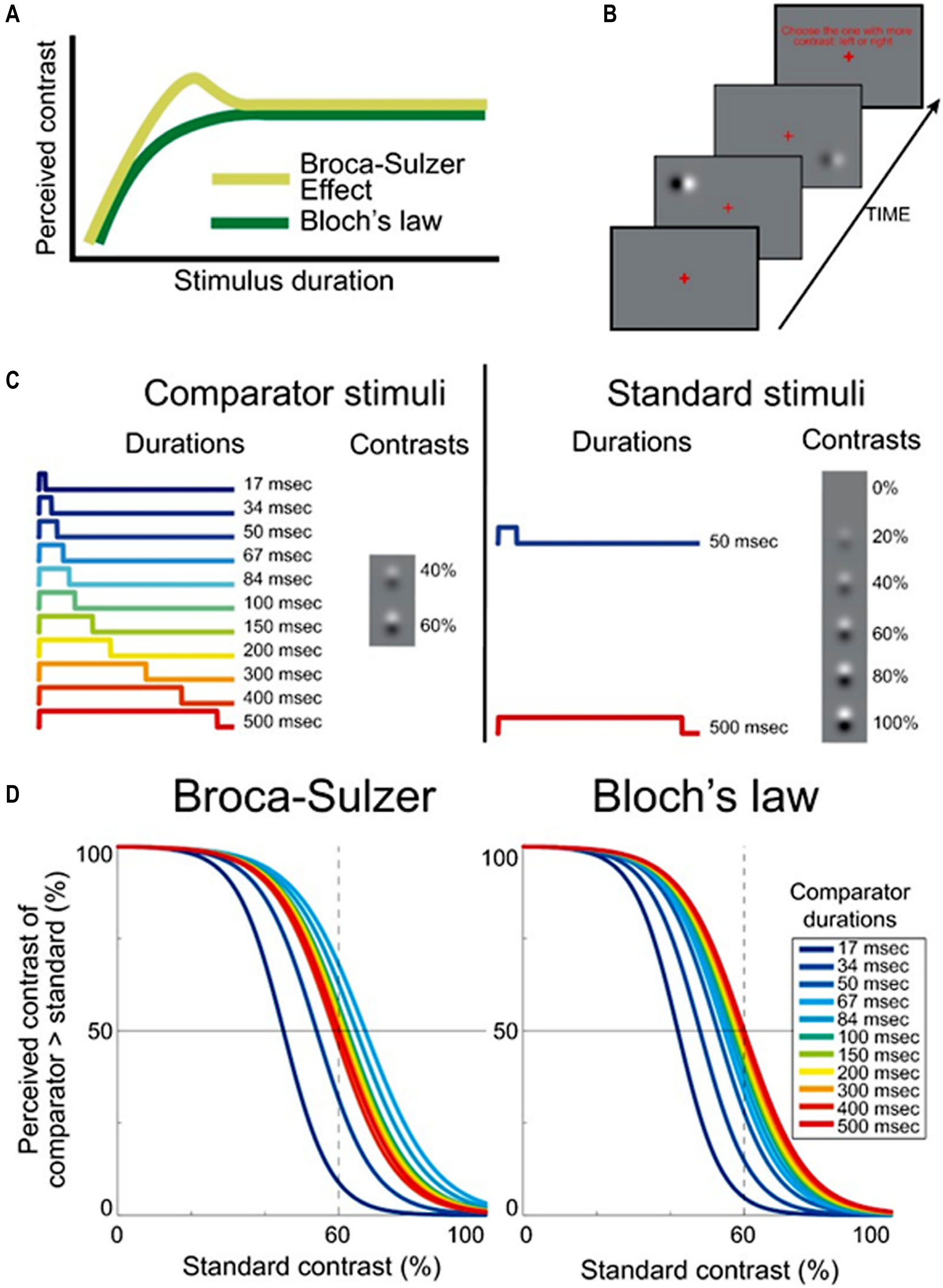

eFIGURE 5-3 Studies of the potential different effects which may occur in temporal summation (see text for details). (From Rieiro et al., 2012.)

The eye has the power to refract (bend) light waves and, as for any lens, this is measured in dioptres. A dioptre (D) is a unit of measurement that describes the strength, or ‘power’, of a lens to bend (refract) light a set amount (degree); the optical power of a lens or curved mirror is equal to the reciprocal of the focal length measured in metres (that is, 1/metres). It is thus a unit of reciprocal length. Thus, a three-dioptre lens brings parallel rays of light to focus at 1/3 metre. The overall refractive power of the eye is around 60D for the normal healthy emmetropic eye and much of its refractive power is attributed to the lens-like (focusing) properties of the cornea (which amount to about 40D), with the remaining 20D due to the ocular lens (eFig. 5-4). The effect of this refractive property of the ocular media (i.e. the tissues through which the light passes) is to focus light rays on the retina, and specifically the fovea, when visual acuity is being measured. Many refractive errors occur in the healthy population, including myopia (short-sightedness), hyperopia (long-sightedness) and astigmatism (non-spherical aberrations of the eye’s refractive power).

The scientific discipline dealing with the optics of the eye is known as optometry. Many textbooks are available which deal with physiological optics as well as optical devices such as spectacle correction, contact lenses and the optics of intraocular lenses, including large treatises dealing with many aspects of physiological optics as well as shorter text books summarizing refraction and refractive errors for those not intending a career in optometry.

The reader is referred to the following texts as examples of (1) a short comprehensive text and (2) a three-volume in-depth treatise:

(1) Hunter DG, West MD. Last-minute optics: a concise review of optics, refraction, and contact lenses. 2010.

(2) von Helmholtz H. Treatise on physiological optics, vols I, II and III. Dover Phoenix Editions; 2005

In addition, several texts are available which deal with the optics of the pseudophakic eye (i.e. the eye with a prosthetic intraocular lens), including official publications of professional bodies such as the American Academy of Optometry and the British College of Optometry.

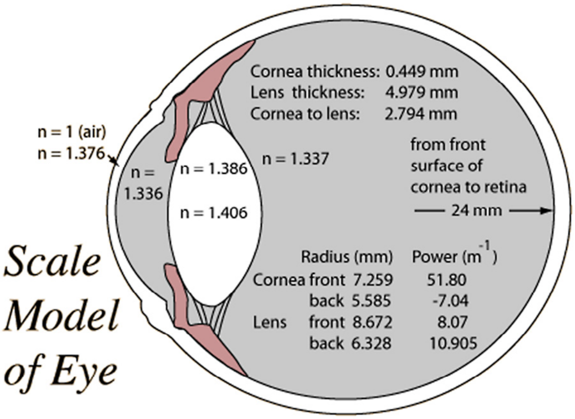

eFIGURE 5-4 Standard dimensions of the eye as related to the model eye. (From Hecht, 1987.)

Theoretically, the resolving power of the eye can be derived from an estimate of the angle subtended by a single photoreceptor (about 1.5 μ or 20 minutes (20′) of arc in the case of cones), as this represents the smallest unit distance separating two individually stimulated photoreceptors. This corresponds to about a pixel on a computer screen when viewed at half a metre. However, it is well recognized that the resolving power of the eye can be as great as 0.5′ of arc, for instance when looking for the gap in a Landholt C target, or 4″ of arc when viewing a thin line on an illuminated background. This hyperacuity, or Vernier acuity, is achieved by the complexities of retinal neuronal synaptic organization and is 5–10× greater than 'standard' visual acuity, but the limits of acuity are still determined to some extent by the retinal photoreceptor mosaic or 'grain'.

The highest discriminatory capacity is subserved by cones, although a certain degree of resolution can be achieved by rods. The level of acuity, however, falls off rapidly the greater the distance from the fovea, such that at 5° from the central fovea visual acuity is only one-quarter of foveal acuity. As rod and cone longitudinal dimensions are not sufficiently different to explain the marked difference in acuity, and as the resolving power of the eye is greater than the theoretical limits based on cell size, other mechanisms must underpin acuity. Visual acuity is affected by the luminance of the test object and the degree of adaptation of the observer; dark adaptation increases both rod and cone acuity and therefore is not affected by the sensitivity of cones *per se*. In contrast, light adaptation increases sensitivity of cones but not rods (see p. 274).

Vernier acuity is used in everyday life, for instance in measuring distance with a ruler or detecting the time on a mechanical clock. Vernier acuity is not present in infancy but reaches its highest level of function around the age of 14. It is absent in strabismic amblyopia but may be present in patients with anisometropic amblyopia. Vernier acuity is different from the recently recognized state of supervision, which has been revealed by the use of adaptive optics. Adaptive optics were developed for use in astronomy to minimize optical aberrations and correct higher-order dynamic aberrations caused by such aspects as angle of viewing and accommodation, as compared with correction of static aberrations such as astigmatism and defocus (see below). In essence, by using a wave form sensor, adaptive optics measures phase aberrations in reflected light produced by the imaging light source. When applied to the eye, for instance in the use of wavefront aberrometers and wavefront-guided vision correction in refractive corneal surgery, adaptive optics can theoretically increase acuity to 'supervision' levels.

LIMITS OF AND LIMITATIONS ON ACUITY

The letters on reading charts such as the Snellen's test type and the ETDRS chart (Box 5-3) have been constructed on the assumption that the average person

BOX 5-3 ASSESSMENT OF VISUAL ACUITY USING STANDARD LETTER CHARTS

Visual acuity in clinical practice is determined as an empirical value based on the assumption that the cone photoreceptor has the ability to discriminate two objects in space subtended by an angle of 1 minute of arc at the nodal point of the eye (**A**). This is measured using a set of charts (optotypes) and standard normal visual acuity equates to the vision of 6/6 or 20/20 (i.e. 1.0 or 100%) when viewing a predetermined standard target (size of optotype letter) at 6 m (UK) or 20 feet (USA). Test conditions describing the ambient illumination, and the illumination of the letters on the chart to provide contrast are also arbitrarily set. The Snellen chart is based on the concept that the smallest spatial target that can be resolved subtends 1 minute of arc at the nodal point of the eye (see above) and although theoretically inaccurate, it serves as a useful parameter. The Snellen chart has rows of letters of decreasing size and is arbitrarily set to produce the standard test at 6 m, although other charts with proportionally smaller letters can be used at shorter distances. The LOGMAR (LOGarithm of the Minimal Angle of Resolution (**B**) is more precisely designed with definitive sizing and spacing of the letters and can provide a more quantitative evaluation of visual acuity. It therefore tends to be the standard for use in clinical trials. As indicated in the text, spatial acuity better than 100% can be achieved, for instance when discriminating the 'offset' of a line or edge (**C**). This is termed Vernier acuity. In addition, visual acuity is modified by such factors as glare and contrast, and indeed can be measured as in the contrast sensitivity test using a sine wave grating as shown in (**D**), where diffraction and aberrations have degraded the contrast of a sinusoidal grating pattern.

BOX 5-3 ASSESSMENT OF VISUAL ACUITY USING STANDARD LETTER CHARTS—cont'd

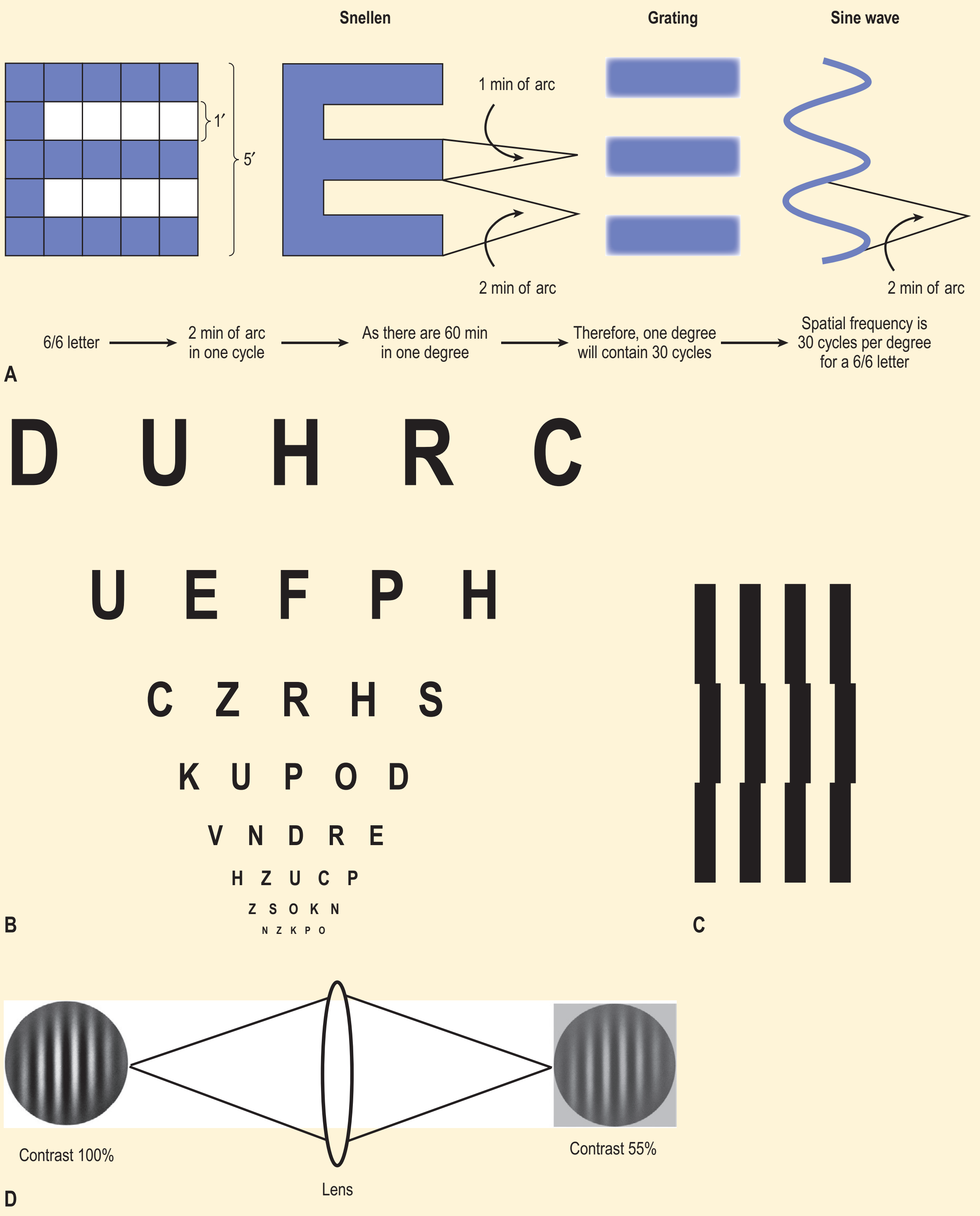
Snellen
Grating
Sine wave
1′
5′
1 min of arc
2 min of arc
2 min of arc
6/6 letter
2 min of arc in one cycle
As there are 60 min in one degree
Therefore, one degree will contain 30 cycles
Spatial frequency is 30 cycles per degree for a 6/6 letter
A
D U H R C
U E F P H
C Z R H S
K U P O D
V N D R E
H Z U C P
Z S O K N
N Z K P O
B
C
Contrast 100%
Lens
Contrast 55%
D

(From Schweigerling, 2000, with permission from Elsevier.)

can resolve two points separated by 1′ of arc. If the limit on acuity is in part determined by the single photoreceptor theory (above) then a one-to-one relationship between the photoreceptor and the nerve cell must exist if there is to be no downstream loss of acuity. For foveal cones such a relationship exists between cone cells, midget bipolar cells and midget ganglion cells (see section on retinal connections, pp. 291–296), but even midget cells have some interconnections with diffuse bipolar and ganglion cells. In spite of these connections, summation of information should not occur for cells subserving the highest levels of acuity and, indeed, is absent from foveal cone cells but is characteristic of rod cells; furthermore, it has been suggested that the improved visual acuity that occurs under conditions of light adaptation is the result of inhibition of these subsidiary connections.

Visual acuity is also limited by the physical behaviour of light, such as diffraction and chromatic/spheric aberration. A single point of light small enough to stimulate a single cone will produce diffraction rings in its traverse through the pupil sufficient to stimulate more than one cone. Similarly, the prismatic separation of white light into its constituent wavelengths will lead to the stimulation of several cones of different types. It is clear therefore that resolution of images must be achieved at a post-receptor level and is in fact a function of the receptive field of each ganglion cell unit. Where there is minimal convergence of information from each receptor, i.e. where the one-to-one relationship between receptor and bipolar cell is maintained, then resolution is at its highest and this occurs at the fovea. However, where there is increasing convergence of information, such as with several parafoveal cones synapsing with one bipolar cell, resolution obviously decreases.

The one-to-one relationship, however, does not adequately explain hyperacuity or Vernier acuity. Diffraction and spheric/chromatic aberrations have ruled out the concept of single unstimulated cones occurring between neighbouring stimulated cells; however, it is likely that discrimination is more a matter of degree than absolute responses, i.e. that resolution is achieved by certain receptors being *less stimulated* than their neighbouring receptors on either side. This is likely to occur with diffraction, where alternating light and dark rings emanate from a point source of light, and with chromatic aberration, where different wavelengths of light are likely to stimulate their respective neighbouring cones to different degrees (Box 5-4). In fact, this fits with the nature of the hyperpolarization response being a graded one in retinal neural cells: as indicated in Chapter 4 (see p. 261), only ganglion cells fire ‘classic’ depolarizing action potentials, while other retinal cells have a graded analogue-type ‘tunable’ electrical response. This differential stimulation from cones is registered with the respective bipolar and ganglion cells and, if combined with a minimal degree of receptor convergence, can explain high levels of discrimination. In this way diffraction could explain, at least partly, the ability to resolve a break in a line subtending an angle of less than 10′ of arc, since partial diffraction lines deriving from the edge of the break would ensure differential stimulation of cone receptors over a very small area. Stimulation of any particular cone is also likely to induce local inhibition (via receptive field mechanisms; see below) in neighbouring cones, thus enhancing resolution and ‘sharpening the image’ further.

These concepts are embedded in the canon of knowledge going back to the time of Rayleigh in the late nineteenth century. However, Rayleigh’s law does not fully explain the real condition where diffraction spreading from a double line is slightly greater than that from a single line for the same total amount of light energy. Information theory, adaptive optics and modern electro-optical devices for generating discrete stimuli might help to explain this anomaly.

These considerations also have a number of implications. In particular, the resolving power of the eye is limited by the distance between two images such that a single cone or set of cones is appreciably less stimulated than the rest; the limit of resolution is therefore not an absolute determinant, but depends on conditions such as light and dark adaptation, background illumination and other factors. Most importantly, it depends on the degree of dendritic connections that occur between the affected cones and the neural cells.

The resolving power of the eye therefore depends on:

- the distance between two objects
- the degree of light and dark adaptation

BOX 5-4 DIFFERENTIAL ACTIVATION OF NEIGHBOURING CONES DETERMINES THE LIMITS OF VISUAL ACUITY

A beam of light interrupted by a small target will produce diffraction rings at its edges – the central rays in each ring will stimulate one cone (2) more than the weaker peripheral rings will stimulate its neighbours (1) and (3). Our ability to detect the break in the light beam is determined by comparing the differential responses in all the cones in the illuminated region and finding two that produce a similar response to correspond to the edges of the target. Visual acuity is therefore a measure of the retina's ability to produce different graded responses and not absolute responses.

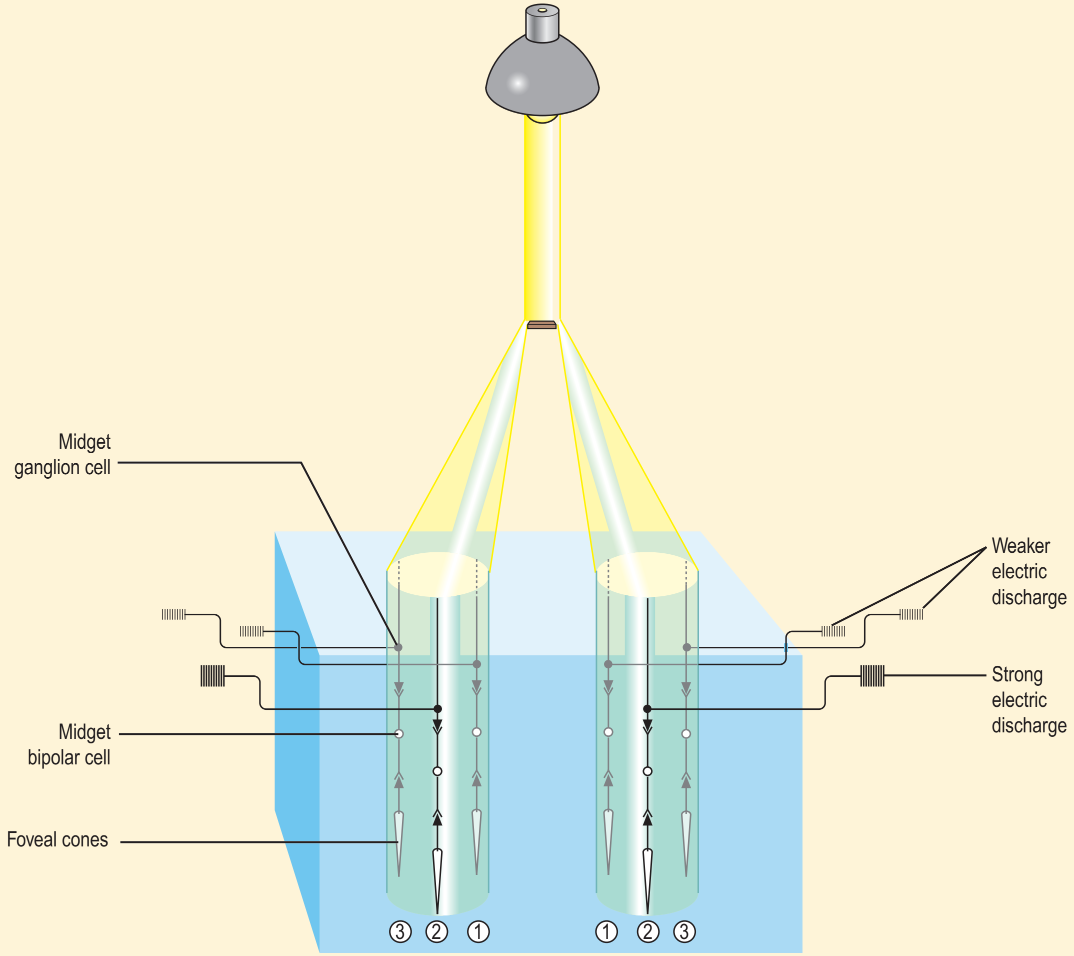

- the background illumination
- the extent of the dendritic connections between the cone and neurones.

CONTRAST SENSITIVITY

Visual acuity is also affected by contrast (sharpness). The finest limits of resolution have been determined by the ability to discriminate a thin white line against a uniform background illumination (0.5′ of arc). The effects of diffraction are such that detection of this line depends on the liminal brightness increment (l.b.i). This increment represents the endpoint at which the differential in brightness between the individual dark/bright oval diffraction rings produced at the edge of the line can be detected; if they are not sufficiently different from the background luminosity, then the line will not be detected. The l.b.i. is determined by the contrast between the light and dark lines, and can be measured quantitatively with a sinusoidal grating (see Box 5-3): a spatial pattern where the average luminance remains the same but the contrast between the light and shaded areas can differ.

The degree of contrast (C) is relative to the background luminance (L) and is described in terms of the maximum (L_{max}) and the minimum (L_{min}) as follows, also known as the Michelson contrast:

$$C = (L_{max} - L_{min})/(L_{max} + L_{min})$$

Alternative measures are the Weber contrast, where the difference between the maximum and the minimum is compared against the background luminance, and RMS contrast, where the mean luminance is a factor of the standard deviation in luminance. The Michelson contrast is used for gratings and a 'threshold' is reached when the target is detected reproducibly. Sensitivity is the inverse of the threshold and thus 'contrast sensitivity' is a measurable quantity, usually described in cycles per degree (the grating frequency) and visual acuity is equivalent to 1/grating frequency. Using this method it has been shown that there is a peak response in the middle range of frequencies (Fig. 5-7).

Contrast sensitivity is therefore set by the limits of the grating frequency and is affected by both the optics of the system and the direction of the grating lines, being most sensitive in vertical or horizontal directions. Remarkably, threshold contrast for many targets sits around 1% independently of target size or brightness, which as Peli observes remains unexplained since originally described by Fechner in 1860 (Peli, 2013).

Contrast sensitivity above threshold is, as for any measure of acuity, affected by luminance. In addition, bar width, length and grating motion all affect sensitivity. In the latter there is likely to be significant cortical processing at this level, as there is for 'line modulation sensitivity', a technique whereby grating bars are composed of wavy lines and the subject is asked to determine whether the line is straight or not. This technique can provide highly sensitive measures of acuity.

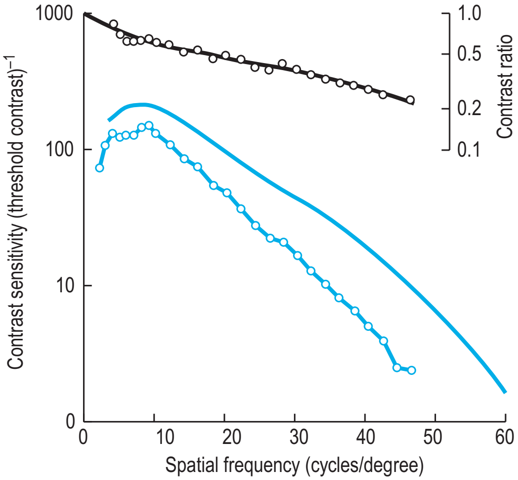

FIGURE 5-7 Contrast sensitivity curve showing peak response at midspatial frequencies.

Contrast sensitivity measurements, while an excellent measure of acuity, are sensitive to phase shifts and grating orientation and for absolute measures of object detection, object contrast is critical. In this context, mesopic low-contrast letter acuity is the most sensitive method for revealing small differences in retinal image 'quality', which influences 'recognition' as opposed to 'detection' of an object. The contrast sensitivity function (CSF) has been arbitrarily measured using a set of five spatial frequencies and has been found to be relatively robust. Age and decreased luminance cause a shift to larger frequencies in the CSF. Glare, which is often a side-effect of refractive surgery, affects the CSF at low rather than high spatial frequencies. For clinical purposes, measurement of the CSF (as opposed to measuring the threshold as in most contrast sensitivity charts) is very time consuming and impractical, but recent developments using customized and selective spatial frequencies and contrasts are allowing tailored CSF tests to be applied to specific conditions, e.g. macular degeneration.

Wavelength also affects contrast sensitivity such that at high spatial frequencies the gratings appear to be of the same colour, whereas at low frequencies (i.e. with coarse gratings), colour differences can be detected. Discrimination is poorest with red–green, however, suggesting that for low frequencies rod–cone interactions are important in achieving best visual acuity. Interestingly, contrast sensitivity appears to induce more electrical signal responses in M cells, generally thought to subserve rod function, than in P cells, which are linked to cone function (see below).

Does the retinotopic arrangement of fibres in the cortex have a bearing on acuity?

The representation of retinal ganglion cells in the LGN and cortex is disproportionately larger for foveal midget cells than for ganglion cells elsewhere in the retina. This produces a ‘cortical magnification factor’ for foveal cones over other cones. However, the magnification factor is not related solely to the reduced convergence of foveal cones on ganglion cells (see below), but also to a disproportionate LGN and cortical representation of neurones served by foveal cones. Current use of fMRI has revealed the retinotopic map of the human visual cortex and demonstrated this magnification factor in technicolour (Fig. 5-8).

BEST-CORRECTED VISUAL ACUITY: EFFECTS OF EXTERNAL FACTORS

Visual acuity is, of course, affected by factors that do not relate directly to the retinal stimulus. These include pupil size, eye movements and binocular viewing.

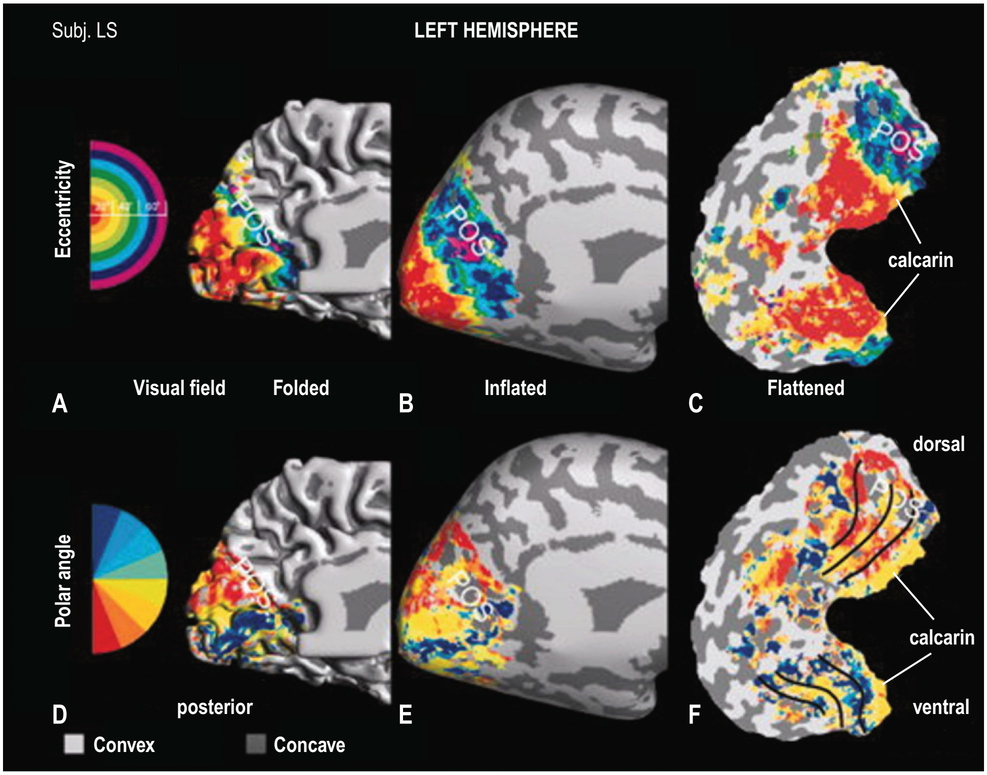

FIGURE 5-8 Dipolar map and isopolar angle maps of human visual areas. These images were prepared using fMRI technology and computer modelling to ‘flatten out’ the visual cortex (**C** and **F**). The top row shows a map of the occipital cortex indicating the retinotopic location of the stimulus as its eccentricity increases from the fovea; the dipole is coded by colour (brown (fovea) -> orange -> blue -> cyan (periphery)) displayed on the original cortical surface (**A**), the unfolded cortical surface (**B**), and the cut and flattened cortical surface (**C**). The bottom row shows the polar angle of the stimulus (blue (upper vertical meridian) -> green (horizontal meridian) -> red (lower vertical meridian)) plotted on the same three surfaces (**D–F**). (From Wu et al., 2012.)

Pupil size and visual acuity testing in infants

The size of the pupil affects the level of visual acuity in that a reduction in pupil size reduces aberrations but increases the effects of diffraction. Below 3 mm, these effects tend to cancel each other out, and visual acuity is independent of pupil size, although wavefront aberrometry reveals that a pupil size of 2.5–3.0 mm produces the best image quality.

The level of visual acuity attained may also have the reverse effect on the size of the pupil. Luminance affects the level of acuity and the size of the pupil is affected by the light level via well-characterized pupil reflexes (Box 5-5). The size of the pupil also indirectly affects the visual acuity by reducing the amount of light entering the eye when the light stimulus is intense, and conversely increasing light capture under dim lighting conditions. This three-way relationship has been used to develop an objective measure of visual acuity, which may be useful in assessing vision in infants and others who are not able to cooperate in standard visual acuity testing. The test uses a high-resolution infrared pupillometry device to show changes in the amplitude of constriction in response to sine-wave gratings presented on a uniformly illuminated test background. As for contrast sensitivity, there is a peak response in the middle range of frequencies and the threshold for response correlates well with contrast sensitivity estimates of acuity. This pupil response is governed by higher visual pathways, being altered in patients with hemianopia but normal pupil light reflexes; indeed, the phenomenon is well recognized by clinical neuro-ophthalmologists. Infrared pupillometry has been shown to be valuable in studies of delayed visual maturation and to be significantly more reliable than the Rosenbaum card method in which subjective comparisons of pupil size are made.

The pupillary response also receives input from the intrinsically photosensitive retinal ganglion cells (ipRGCs) through melanopsin (see Ch. 4, p. 249) and it is possible using chromatic pupillometry (stimulation of pupil responses at different wavelengths) to separate the contributions from rods, cones and ipRGCs. There are diverse types of ipRGCs with several different functions. For instance the sleep/wake circadian rhythm responders connect with the suprachiasmatic nucleus, while the pupillary responsive neurones synapse in the pretectal (olivary) nucleus. In terms of response to light, ipRGCs are considered to register irradiance or radiating light. They may also be responsible for photoallodynia (the photophobia/light aversion response to very bright light).

Eye movements

The concept that the continuous fine eye movements that occur as part of normal viewing are important in ensuring constant stimulation of the photoreceptors to maintain image perception remains popular. Indeed, it has been shown that images received by peripheral

BOX 5-5 PUPILLARY LIGHT REFLEXES

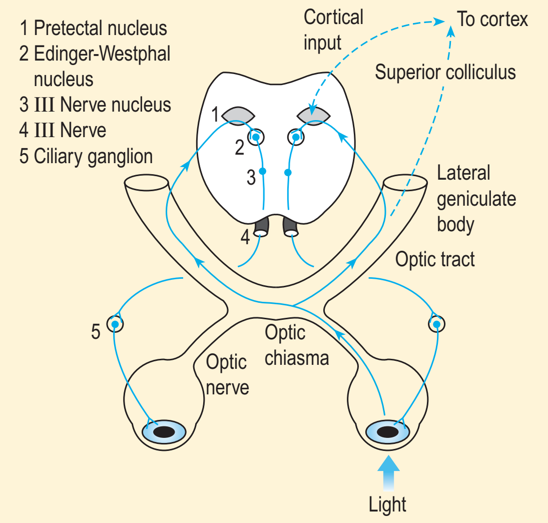

- The afferent response commences in photoreceptors, is transmitted to retinal ganglion cells, enters the optic nerve, decussates at the chiasm, traverses the optic tract and terminates in the pretectal nucleus (bypassing the lateral geniculate nucleus).
- Both crossed (via posterior commissure) and uncrossed fibres pass from pretectal nucleus to Edinger–Westphal nucleus (parasympathetic).
- Parasympathetic fibres pass to the III nerve nucleus and leave the brainstem via the III nerve. Fibres synapse in the ciliary ganglion before supplying sphincter pupillae of iris (constriction) via short ciliary nerves.
- Uniocular light stimulus therefore gives rises to bilateral and symmetrical pupillary constriction.
- Melanopsin signals through ipRGCs (see text) to the suprachiasmatic nucleus (circadian response) and the pretectal nucleus (irradiance response).

receptors fade rapidly if fixation is deliberately maintained in one position – the Troxler phenomenon. Although this was originally considered to be a mechanism for enhancing the central image by inhibiting peripheral images, use of a 'stabilized retinal image' has shown that elimination of these fine movements does not necessarily lead to a reduction in visual acuity. However, these findings were obtained using high-frequency gratings and it is possible that fine eye movements may be important at lower spatial frequencies in improving contrast.

Fine eye movements occur during different visual tasks: for instance, during reading, the fixation time on the target letter is around 200–250 ms and the average saccade is about 8–9 letters. This increases in skim reading but the level of cognition (the perceptual span) is reduced. Useful information is gathered from a region about 3–4 letters to the left of foveal fixation and 8–9 letters to the right.

Binocular viewing and the probability theory of visual perception

Perception is a relative occurrence and depends on many factors to achieve optimal levels (there is a significant element of chance in achieving this optimum which can be expressed as a linear transformation of log odds of frequency and/or probability). Interestingly, determination of some stimuli such as negative (concave) contours versus positive (convex) contours has a greater chance of detection.

It follows therefore that two eyes are better than one, at least in increasing the chances of the highest level of visual processing of the same image.

Electrophysiology of the visual system

The transmission of nerve impulses in retinal receptors and neurones is mediated, as might be expected, by recordable changes in electric potential across the cell membrane (see Ch. 4, p. 161) and is accompanied by electric discharge. The action potential is usually an all-or-nothing event and, in muscle tissue, does not occur in the resting state. However, in neural tissue continuous discharge may be taking place and information is relayed by changes in the frequency or rate of electric discharge in the nerve, an increase in frequency usually representing stimulation and a decrease representing inhibition, thus emphasizing the essential binary nature of biological information systems similar to computers. This applies for all nerves in any system: the character of the received sensation is determined not by the type of nerve but by the site of information relay in the cortex and its subsequent processing in the brain.

In the retina, these general principles hold true for retinal ganglion cells, but in bipolar, horizontal, amacrine and photoreceptor cells the electrical response is more of a tonic or graded response, and the direction of the response can be positive or negative. For instance, it is this graded response that permits spatial discrimination via differential responses to diffraction rings, as described above (see p. 280). However, the graded response in the bipolar cell becomes an ON/OFF response in the ganglion cell. As Ikeda has put it, retinal information is converted from an analogue signal to a digital signal at the final stage of retinal processing, i.e. at the connection between ganglion and bipolar cells (Ikeda, 1993).

THE ELECTRICAL RESPONSE IS INITIATED BY PHOTOTRANSDUCTION

As we have seen (see Ch. 4, when a photon of light strikes the photoreceptor outer segment, conversion of rhodopsin to the activated molecule induces a series of molecular events culminating in an electrical response. Cells, and particularly neurones, normally exist in a 'charged' state in that the inside of the cell is 'negative' with respect to the extracellular environment, creating an electrical potential difference across the cell membrane. This condition is maintained by differential distribution of Na^+ and K^+ ions on either side of the cell membrane. When a neurone is stimulated, there is an initial period of gradually increasing positivity (the generator potential), which culminates in a spike discharge characterized by a rapid depolarization response of the cell. This is achieved by the rapid influx of Na^+ through ion channels that are 'opened'.

In the photoreceptor the reverse situation occurs. Under resting conditions in the dark, the outer segment is maintained in a depolarized state through open ('leaky') Na^+ channels, which permit the influx of sodium ions from the extracellular space. When light stimulates the outer segment, the sodium channels are abruptly closed, stopping the influx of sodium and thereby leading to a reduced level of depolarization, i.e. a relative hyperpolarization (Box 5-6). This is a direct result of rhodopsin isomerization and is mediated by amplification mechanisms involving cyclic guanosine monophosphate (cGMP) (see Ch. 4, p. 262).

The hyperpolarization response is transmitted by a flux in calcium ions along the length of the photoreceptor to the synapse with the bipolar cell (the Ca^{2+} wave), which is then induced to release its transmitter (glutamate). Bipolar cells may then adapt to one of two responses to glutamate, depending on which type of receptor is induced: an ON response, which is a hyperpolarized state, and an OFF response, which is a depolarization response (see below). Indeed, the hyperpolarized state conferred on the bipolar cell is also transmitted to the horizontal cells in the same region. However, the hyperpolarization response of the bipolar cell is not as steep as that of the photoreceptor in the excited state.

BOX 5-6 DARK CURRENTS

Dark currents occur in the resting state (dark adapted eye) owing to 'Na^+-leaking' outer segments.

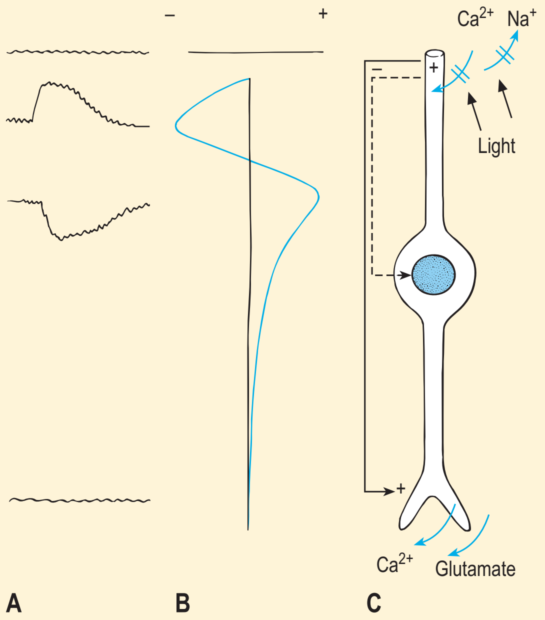

The conversion of light energy to an electric response is dependent on specialized ion channels that tightly control the permeability of the cell membrane to Na^+ and Ca^{2+}. Light stimulation reverses the dark current by closing Na^+ channels in outer segments and releasing Ca^{2+} (and glutamate) at synapses. The cGMP-gated Ca^{2+} channel and the Na^+/Ca^{2+}, K^+ exchanger are located in the plasma membrane of the photoreceptor, not in the disk stack, but are complexed together with peripherin/rds-rom-1, an integral protein of the disk rim.

It will be obvious, therefore, that not only is there a resting potential difference across the photoreceptor cell membrane but there is also a potential difference along the length of the photoreceptor in the dark between the relatively depolarized outer segment tip and the hyperpolarized synaptic region of the cell at its interaction with the bipolar cell. This generates the 'dark currents' in the eye, which are reversed by the photic current on light stimulation when the photoreceptor tip becomes hyperpolarized (see Box 5-6).

ELECTROPHYSIOLOGY OF SINGLE RETINAL CELLS

Early studies in this field concentrated on the large single neurones that could be obtained from invertebrate eyes and showed that typical action potentials could be obtained, usually preceded by a generator potential (Box 5-7). Surrounding neurones were usually inhibited when action potentials occurred in a single nerve.

Later studies of electrode-impaled optic nerves in vertebrates showed that the rate of discharge in certain nerve fibres increased (ON response) when a light stimulus was presented to the eye, while it decreased in other fibres (OFF response). Yet others produced an ON/OFF response. As it was known that there were 150×10^6 photoreceptors but only 1×10^6 optic nerve fibres, it followed that many receptors must feed information into a single neurone, i.e. there must be convergence of signals and some of these must be inhibitory while others are stimulatory. On this basis, the concept of receptive fields was developed and confirmed by direct experimental testing on isolated optic nerve fibres using discrete spots of light to stimulate the retina (Box 5-8). Several phenomena, such as summation, which could previously be inferred only from psychophysical experiments, were directly confirmed. Indeed, summation effects could be compared with the effects of generator potentials in other systems that, in ganglion cells, are called synaptic potentials.

BOX 5-7 GENERATOR POTENTIALS

GENERATOR POTENTIALS

- (**A**) Single neurone responding to single large electric stimulus produces an action potential.
- (**B**) Single neurone responding to several small stimuli produces an action potential.
- (**C**) The retinal photoreceptor-bipolar cell electrical response steps down in quantal rates in a graded fashion.

For instance, three 'summed' small light responses produce a typical ON response in ganglion cell neurone

Generator potentials

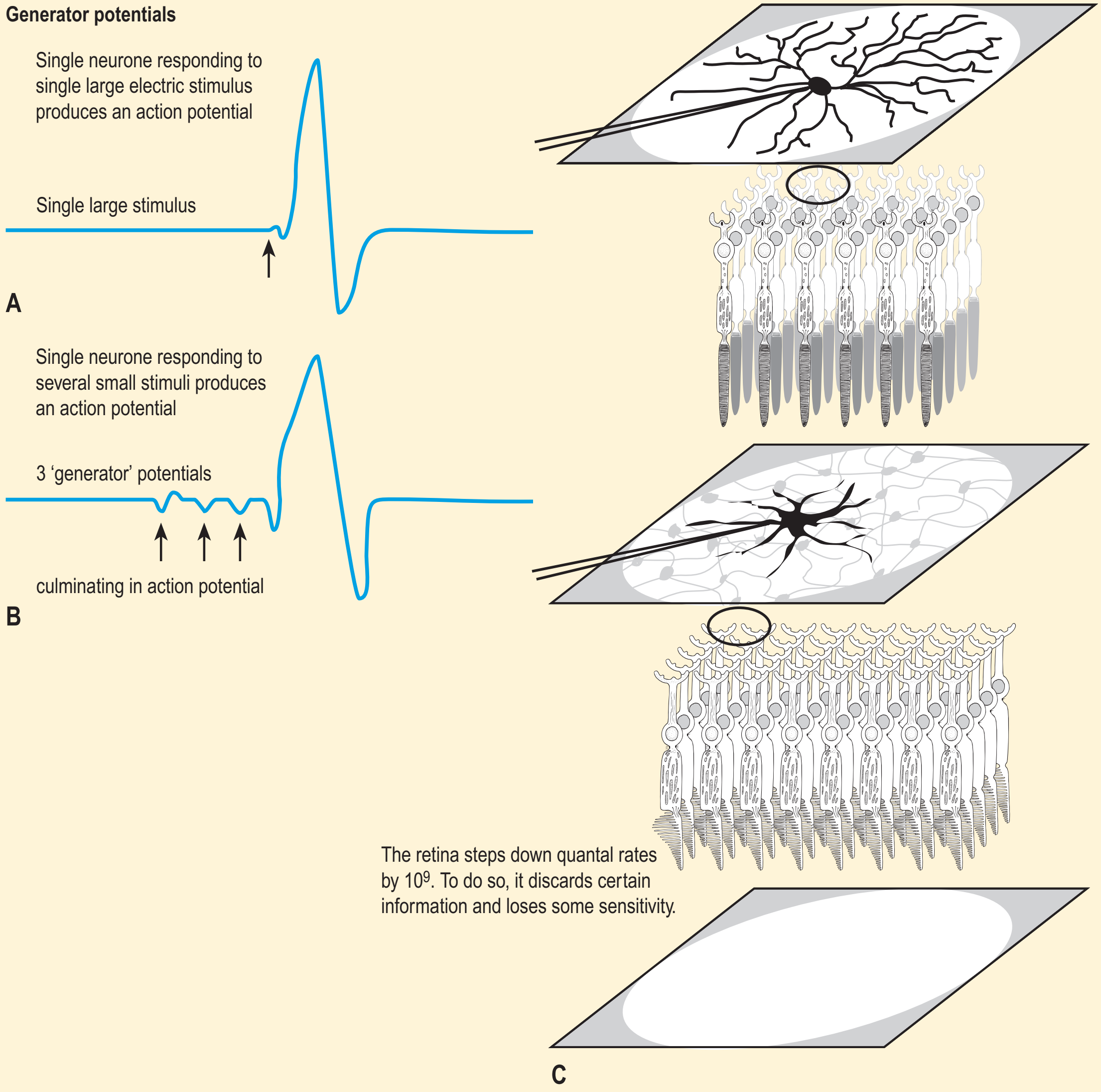
Single neurone responding to single large electric stimulus produces an action potential
Single large stimulus
A
Single neurone responding to several small stimuli produces an action potential
3 'generator' potentials
culminating in action potential
B
C

The retina steps down quantal rates by 10^9. To do so, it discards certain information and loses some sensitivity.

BOX 5-8 ORGANIZATION OF VISUAL INFORMATION INTO DISCRETE RECEPTIVE FIELDS

The receptive field of a retinal ganglion cell neurone is the area of retina covered by that cell in which a light stimulus alters the frequency of discharge (**A**). The size of the receptive field varies from 200 to 600 μm in diameter. Receptive fields may have an ON centre (increased rate of discharge) and an OFF surround (decreased rate of discharge), an OFF centre and an ON surround, a double opponent ON/OFF centre/surround organization, or a pseudodouble opponent centre/surround organization (**B**); in addition, these arrangements occur for detection of dark–light, red–green and yellow–blue. Moreover, the dendritic connections of bipolar cells can overlap, allowing each cone to be contacted by more than one bipolar cell and, where overlap occurs, they share cones (**C**); this, in effect, means that at ganglion cell level, each ganglion cell may have its ON–OFF receptive field but this is based on input from a range of cones allowing some degree of 'colour mixing' in what is essentially a hexagonal array (see Fig. 5-10).

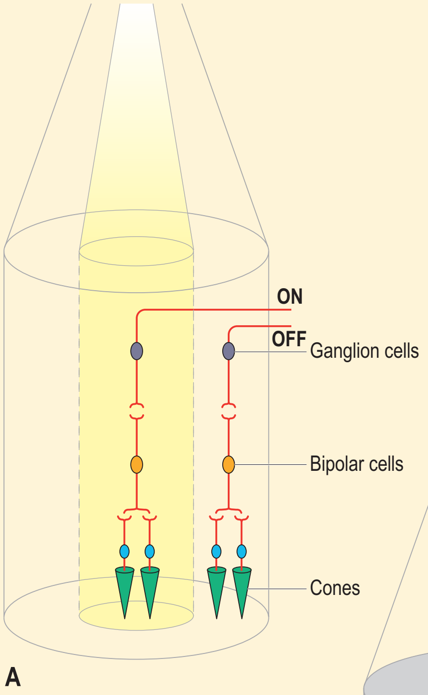

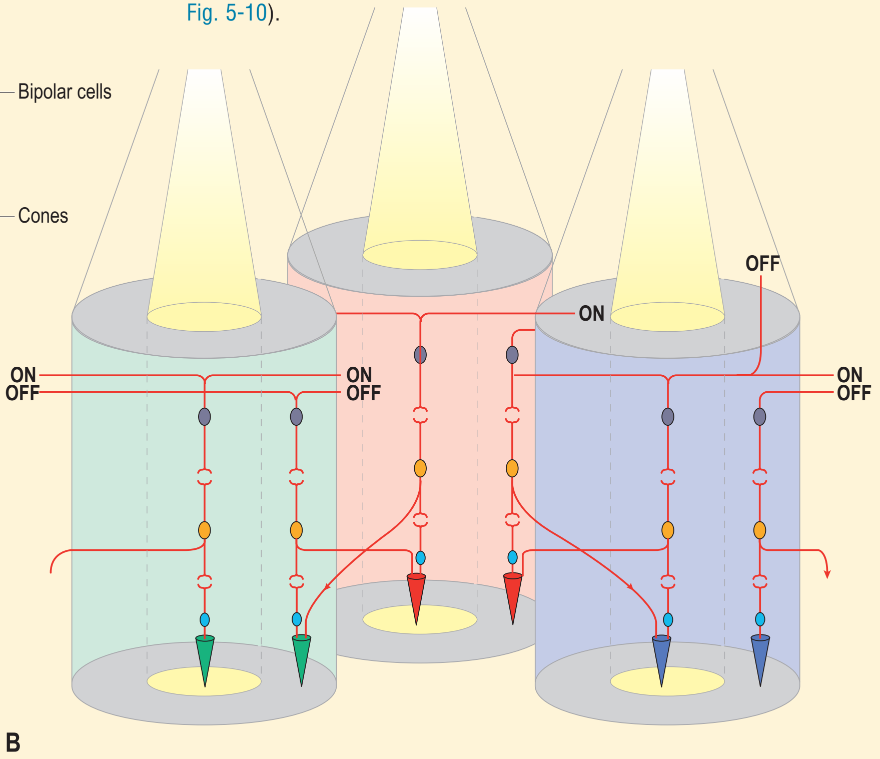

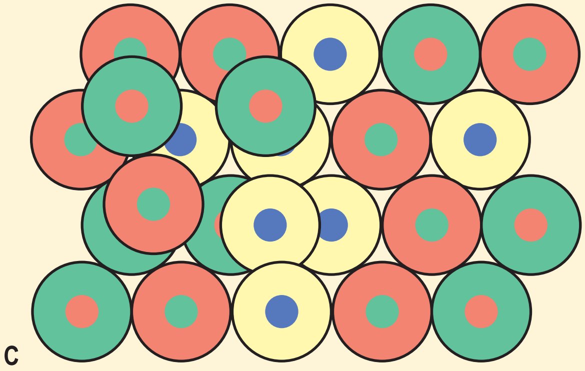

The ON/OFF response in the optic nerve fibres has been shown to correlate with the centre/surround organization of the receptive field and is based on interneuronal interactions causing inhibition in surround cells.

ON/OFF receptive fields apply to cone–bipolar–ganglion cell circuitry while rod cells synapse directly into cone–cone circuits (see below). ON/OFF receptive field organization applies to several different types of signal, including light–dark, blue–yellow and red–green (see below). Amacrine cells and horizontal cells considerably modify the ON/OFF microcircuitry organization and, variably so far, each type of circuit. In addition, the receptive fields of retinal neurones depend on the size of the cell: big ganglion cells (magnocells, M cells) have large receptive fields and small ganglion cells (parvocells, P cells) have small ones. M cells receive information from many amacrine and bipolar cells, producing a high degree of convergence. P cells have small receptive fields receiving information directly from single or only a few bipolar cells (see below).

How does the photochemical change in the receptor produce the spike discharge in the retinal neurone?

Horizontal cells at the receptor–bipolar cell interface and amacrine cells at the bipolar–ganglion cell interface are integrally involved in the organization of the retinal neuronal response to light. Single-cell recordings have shown that the hyperpolarization response at the receptor level is a graded response (see above, discussion of visual acuity), not an ON/OFF response. Similarly, the hyperpolarizing horizontal cell response is a graded response, but in the horizontal cell there is a longer latency and the potential for summation over a wide range. The bipolar cell response is also graded but with a centre/surround effect where the centre hyperpolarizes and the surround depolarizes. In the amacrine cell transient spikes can be observed, especially if correlated with the ON/OFF response, but only in the ganglion cell is a sustained spike discharge observed with a true depolarization occurring for ON, OFF and ON/OFF responses, depending on which type of ganglion cell is studied. This all-or-none response is graded only in the sense that the frequency or rate of discharge that occurs in the ganglion cell varies with the degree of depolarization.

RETINAL CONNECTIONS, CIRCUITRY AND NEUROTRANSMITTERS

Retinal connections

What is the basis of receptive field organization in the retina? As detailed above, detection and processing of the light stimuli by the retina is founded on the retinal receptor/neuronal network comprising the photoreceptors (rods and cones) and the neurones (bipolar cells and ganglion cells). The information finally transmitted to the lateral geniculate nucleus in the brain by the ganglion cells (ON, OFF and ON/OFF cells) is received directly from the bipolar cells but is modulated by horizontal cells and amacrine cells in the plexiform layers in the retina. This is canonically described in a simple arrangement of direct bipolar cell activity and lateral inhibition by horizontal cells and amacrine cells (see Ch. 1 for details). In reality, retinal microcircuitry is more complex than this, underpinned by the fact that each retinal cell comes in several different varieties (Fig. 5-9). For instance, only one type of bipolar cell connects with rod photoreceptors but each of the 12 different types of cone bipolar cells connects with each cone photoreceptor in an integrated manner. In addition, the general arrangements found in most mammalian retinas are complicated by the presence of the fovea in primates, which is characterized by a single type of ganglion cell. A series of excellent reviews of this field has been published by Masland (2011, 2012a,b,c).

Additional content available at https://expertconsult.inkling.com/

There are several retinal microcircuits emanating from the photoreceptors and acting in parallel. As indicated above, each bipolar cell connects all the cones within its branching territory and because of overlap, each cone connects with each of the 12 types of bipolar cells, each of which is transmitting different types of information to the next layer of cells (amacrine, horizontal and ganglion cells) (Fig. 5-10).

Thus, cones from several different types of microcircuits with a range of bipolar cells while rod microcircuitry is minimal. In evolutionary terms, cones developed first, even though there are 20 times more rods than cones in the retina. Cone-bipolar cells synapse in the inner plexiform layer in a highly ordered set of stacks, each with a different number of

Masland's work has shown that the more than 60 different cell types in the retina are functionally organized into three sets which (1) sift the information from photoreceptors into roughly 12 separate channels or streams running alongside each other simultaneously; (2) transfer this information to specific types of ganglion cell; and (3) integrate information from bipolar cells and amacrine cells with ganglion cell output into approximately 20 different coded messages concerning the nature of the visual image which are transmitted to the brain via the ganglion cells (Masland 2011, 2012a,b,c). The very large numbers of amacrine cells are especially interesting and appear to subserve many functions, many of which remain to be discovered but some of which are inferred simply from their positioning and architecture within the retina. For example, most recently, an 'interneurone' which in essence sends a blue OFF signal to the brain has been discovered, further underlying the complexity of neuronal cell integration within the retina.

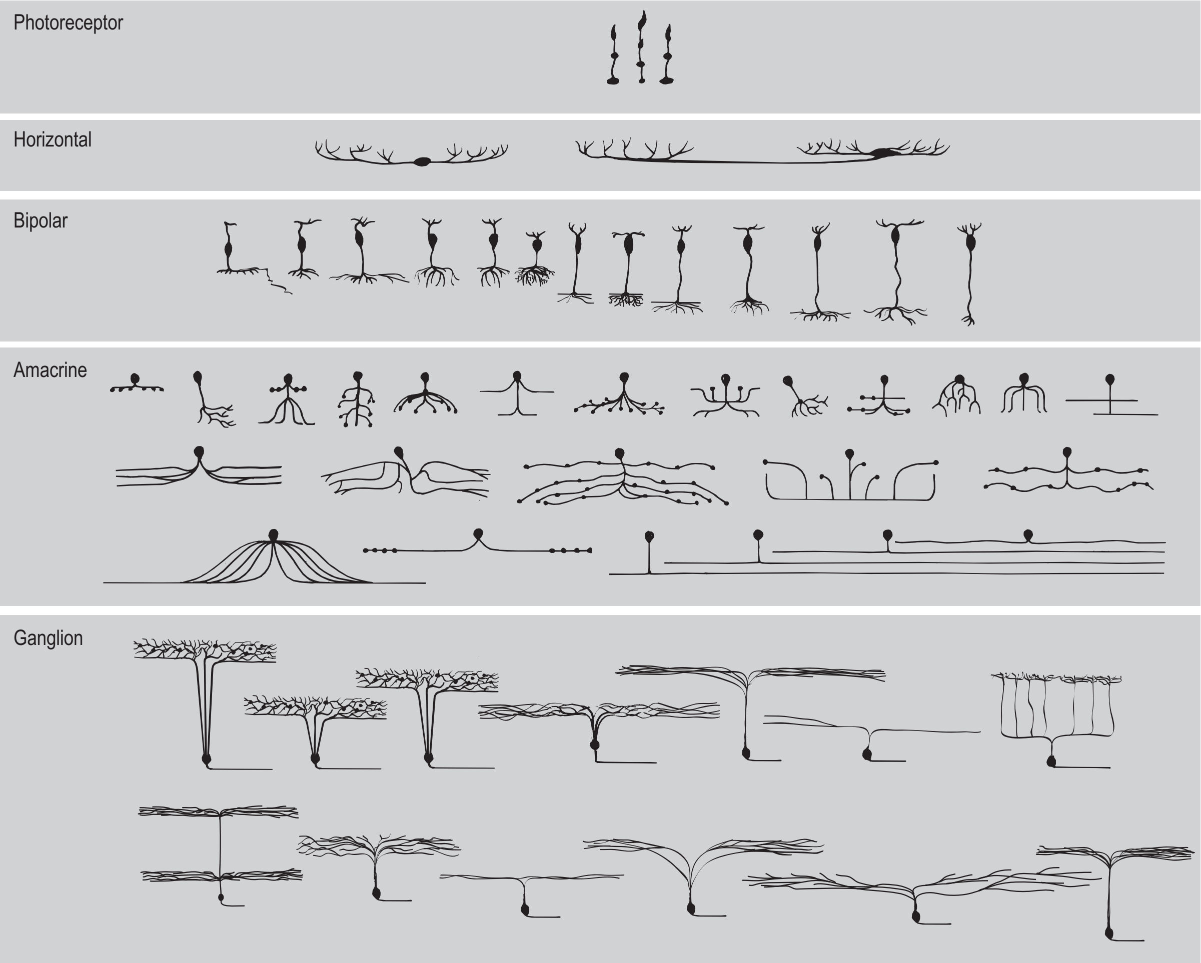

FIGURE 5-9 The major cell types of a typical mammalian retina. (From Masland, 2012.)

connections and with its unique set of ionotropic, metabotropic (mGlur6), glycinergic and GABAergic receptors and calcium-binding proteins reflecting their inhibitory or excitatory output (translated into ON or OFF responses in the ganglion cell (see Box 5-7, Figs 5-9 and 5-10, video 5-1, and additional content online concerning Masland).

There are several different types of ON/OFF bipolar cells, in part determined by the duration of the response (transient versus sustained); in addition, a single bipolar cell for these types of responses (non-chromatic ON/OFF) takes information from more than one cone (Fig. 5-10).

The organization of colour detection is somewhat different. A basic centre surround organization pertains here also. However, wavelength discrimination (see below) requires output from at least two cones to have something to compare against. 'Blue' (short-wavelength) cones make synapses with a single specialized type of cone-bipolar, while the remainder of the cones (long-wavelength red–green cones, comprising around 85%) connect with the several different types of bipolar cells described above (Figs 5-9 and 5-10). The blue bipolar cell registers ON stimuli but until recently it was unclear how blue OFF responses were generated since there is only a single type of blue bipolar cell. It appears that amacrine cells (see below) can convert a blue bipolar ON signal to an OFF signal through its role as an inhibitory interneurone (Fig. 5-11).

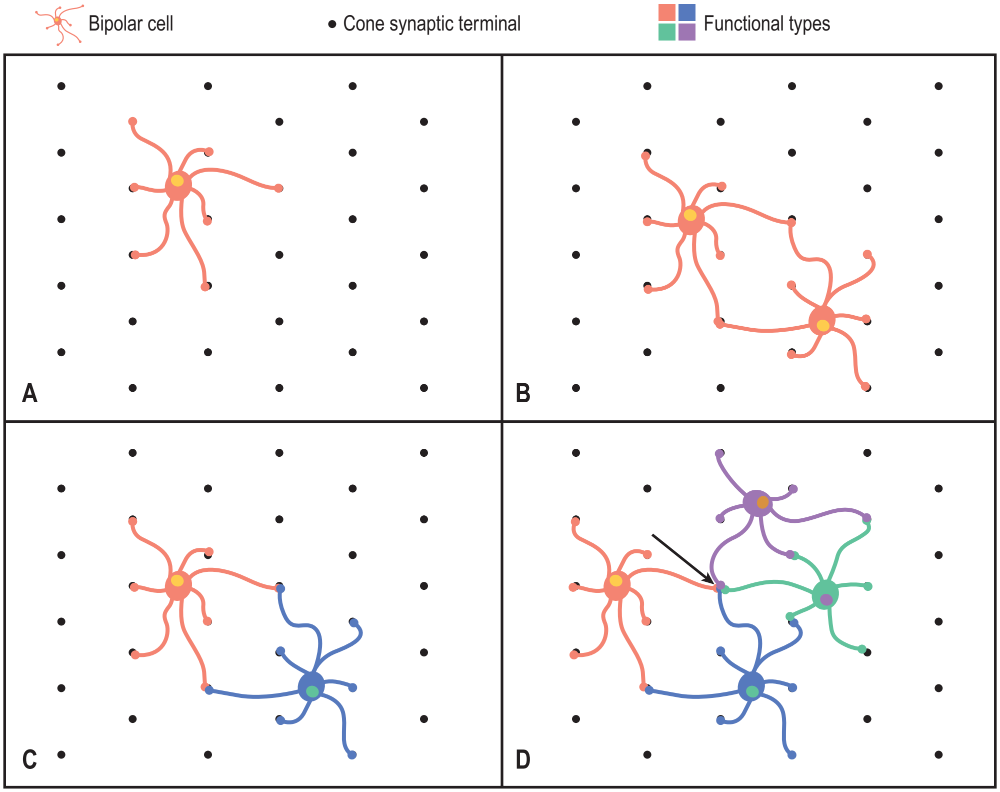

FIGURE 5-10 (**A–D**) Each bipolar cell (represented by the different colours) connects all the cones within the area of its dendritic tree. The retina is organized such that each bipolar cell makes contact with all of the cones within the spread of its dendritic extensions, with the exception of the 'blue cone bipolar' which only connects with blue cones. They are numerically small and the principle remains that each of the different types of bipolar cell that connects with a single cone sends a unique message from that cone to the ganglion (or other) cell in the inner retina. (From Masland, 2012.)

This particular amacrine cell is particularly sensitive to short-wavelength light and uses glycine as its inhibitory neurotransmitter. At the ganglion cell level the blue ON ganglion cells receive input from the blue ON bipolar cell and the red–green OFF bipolar cell and it has been proposed that the blue OFF ganglion cell also receives input from the same amacrine inhibitory cells by sending its dendrites to that stratified layer of the retina. This is less certain, however.

Colour discrimination is therefore made by comparison of light detection by a short-wavelength cone (ON/OFF) with that from a long-wavelength cone (either red or green, ON/OFF), i.e. it is essentially a dichromatic system. In most mammalian retinas, there is no red/green discrimination (i.e. there is true dichromatism with comparison of only one long-wavelength cone with a short-wavelength cone), while in humans there is also red/green discrimination. However, in 5% of humans this is also true dichromatism (red–green colour blindness). These concepts are dealt with in more detail later in the chapter (p. 302).

So where does rod microcircuitry fit into this neuronal organization? Rods detect dim light, while cones detect bright light. Detection of dim light by rods is not to be confused with the OFF response of cone bipolar OFF cells (i.e. the response to the absence of light) but is a positive response by rod cells to very low levels of light. The majority of rods connect indirectly with ganglion cells through cone bipolar cells; large numbers of rods synapse with a single rod bipolar cell, which then connects via gap junctions

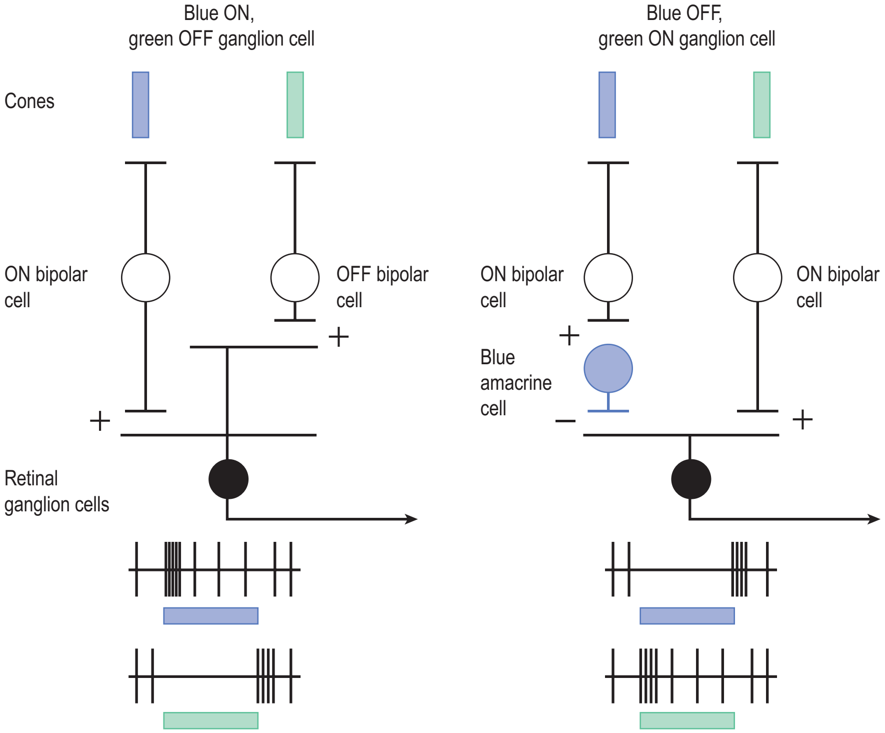

FIGURE 5-11 The physiology of the ON:OFF ganglion cell as shown for the blue ganglion cells. Blue bipolar cells contacting blue-sensitive cones send a signal in response to the brightness of the stimulus (ON stimulus), which synapse directly to a particular ganglion cell (diagram on left of figure). In contrast, if the bipolar cell receives a signal from the amacrine cell, this inhibitory signal generates a response from an OFF bipolar cell. In both cases the signal is complemented by input from an inversely corresponding green cone. The stream of action potentials as the final output to the brain is shown in the lower section of the figure. (From Masland, 2012.)

with a cone bipolar cell through a further particular type of amacrine cell, the AII cell.

In primates, there is further specialization at the ganglion cell level related to the fovea and derived from the system of midget cells: midget ganglion cells connect directly with one cone bipolar cell and through this cell with one cone photoreceptor. Thus there are a huge number of small bipolar cells and midget ganglion cells at the fovea, limited only by the packing of cones in this rod-free area. For instance, midget cells comprise about 70% of the ganglion cells in the monkey fovea. Moreover, in the primate fovea each cone cell connects with an OFF bipolar and an ON bipolar, leading to a greater density of bipolar cells at the fovea than cone photoreceptors. Each midget ganglion cells has a simple centre/surround organization underpinning the excellent spatial resolution of 'supervision'; in addition, it is believed that the midget cell system underpins the dual circuits required for red/green differentiation added to the existing blue cone system (see Box 5-8 and Figs 5-9 and 5-10). At the fovea, the colour opponency theory may not be as strict as suggested above, but is modified by input from horizontal cells: thus short-, middle- and long-wavelength cones may compete against all the cones that contribute to the surround area after modulation by the horizontal cells.

Horizontal cells (Fig. 5-9) in the outer plexiform layer thus provide important feedback for both rod and cone outputs. For cones, this is classically believed to be in the form of contrast and mostly via generating inhibitory information to ganglion cells surrounding the activated cell (similar to the inhibitory effect of an action potential in a bundle of peripheral neurones on the neighbouring unstimulated neurones). This produces the centre/surround

H$^+$ negative feedback onto cone terminals induced by HC polarization

Surround light	HC		V–ATPase at HC		Synaptic cleft [H$^+$]		Cone I_{Ca}		Transmitter release		OFF–BC	ON–BC
OFF	depolarize	→	↑	→	↑	→	↓	→	↓	→	hyperpolarize	depolarize
ON	hyperpolarize	→	↓	→	↓	→	↑	→	↑	→	depolarize	hyperpolarize

A

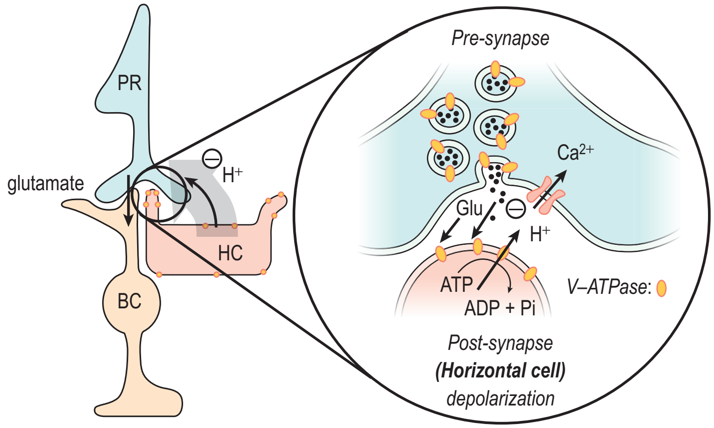

B

FIGURE 5-12 Release of H$^+$ ions exert negative feedback onto the bipolar cells: this is achieved by changing the H$^+$ concentration at the cone photoreceptor:bipolar cell (PR:BC) synapse and is induced by horizontal cell (HC) polarization of BCs via change in H$^+$ concentration change at synaptic clefts of cone terminals induced by HC polarization. (**A**) Shows the direction of change in the various parameters exerted by HC depolarization and hyperpolarization. (**B**) Shows a schematic drawing of H$^+$ negative feedback onto BCs via release of H$^+$ from depolarized HCs illustrating the 'OFF' case in (**A**). H$^+$ release occurs via V–ATPase (a proton pump, turquoise circles in the inset figure) resulting in suppression of glutamate release from cones. V–ATPase is in synaptic vesicles at the cone terminals. (From Hirasawa et al., 2012.)

organization of ganglion cells (see Box 5-8). Some also believe that horizontal cells may mediate their effects based on subtracting the information on the average illumination of the entire retina from that stimulating a restricted set of cones, thus underpinning the mechanism of light adaptation. Rods are separately modified by horizontal cells because of the anatomical location which separates the contact point with the rod far from the rod's contact with the cone bipolar cell.

One biochemical mechanism whereby horizontal cells direct the nature of the ON or OFF response appears to be mediated by acidification at the photoreceptor: bipolar cell synapse, mediated by a membrane-bound proton pump, and suppressing the effects of the glutamate neurotransmitter (Fig. 5-12). In contrast, amacrine cells are much more numerous than horizontal cells (Fig. 5-9) and have a wide range of functions. There are around 30 types of amacrine cells, some modifying ganglion cell response over a wide area while others are much more restricted. In addition, they inhibit, enhance, entrain and refine through the large range of neurotransmitters and receptors discussed above (e.g. dopamine, acetylcholine, glutamate, GABA, glycine, etc.). They thus affect many functions such as centre/surround organization, orientation selectivity, light–dark effects and colour discrimination. Some amacrine cells may have very small arborizations and function entirely within the receptive fields of a wide-field ganglion cell. For instance, starburst amacrine cells have been identified as being ON–OFF direction sensitive, a feature that may involve only a section of the branching arborization in signalling (Fig. 5-13). In addition, other amacrine cells are sensitive to saccade-induced suppression which allows differentiation of true object motion from motion of the entire visual field induced by fine eye movements. Amacrine cells may have paracrine and multiple neurotransmitter secretory functions as well as release nitric oxide.

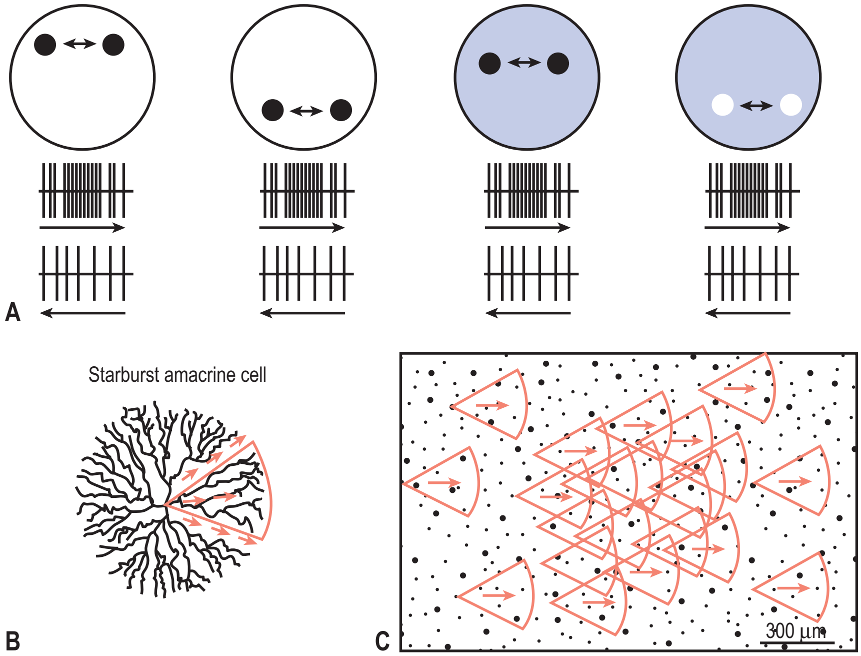

FIGURE 5-13 The ON–OFF direction selectivity of the starburst amacrine cell is the same within all regions of its receptive field (**A**) and can be detected even within a sector of the field (**B**) as shown for small moving targets (**C**). Because the sectors served by each dendrite (represented by the black dots) are smaller than the receptive field, small OFF movements can be detected. (From Masland, 2012.)

Thus some degree of information processing may occur at the inner plexiform/ganglion cell level (e.g. in contrast gain control) before it reaches the visual cortex.

Finally there are about 20 different types of ganglion cell (Fig. 5-10). Originally described as X (slow, tonic) and Y (fast, transient) cells in the cat, and P (parvocellular, midget) and M (magnocellular, parasol) cells in the monkey, several others are now known to be responsible for centre/surround organization and direction/orientation selectivity. Each of the 20 different ganglion cells 'tiles' the retina in its dendritic field so that the specific function of that ganglion cell is represented at that point in the retina (Fig. 5-14).

In addition, a separate rare population of ipRGCs (see Ch. 4, p. 249) connects with neurones in the pretectal and suprachiasmatic nuclei and responds to the level of 'irradiance' and controls pupil light responses as well as sleep/wake cycles (see Box 5-5). It is now known that there are several types of ipRGCs (see eFig. 5-2) which regulate the production of melatonin by the pineal gland and are thus involved in entrainment of circadian rhythms (see p. 277).

From all of the above it is clear that there is considerable diversity in the retinal microcircuitry in which significant information processing occurs before its transmission to the higher centres in the brain.

CLINICAL VISUAL ELECTROPHYSIOLOGY

The electrical potentials which exist in the eye, by virtue of factors such as the dark currents in the retina and the transcorneal epithelial potential difference, can be altered by stimulation with bright flashes of light, producing mass responses of the tissue. These responses represent the resultants of many cellular potentials and the source of the discharges can only be inferred. However, extensive studies have located the source of retinal electric responses at different levels in the retina (Fig. 5-15). There are established internationally agreed standards describing normal responses set by the International Society for Clinical Electrophysiology of Vision (ISCEV).

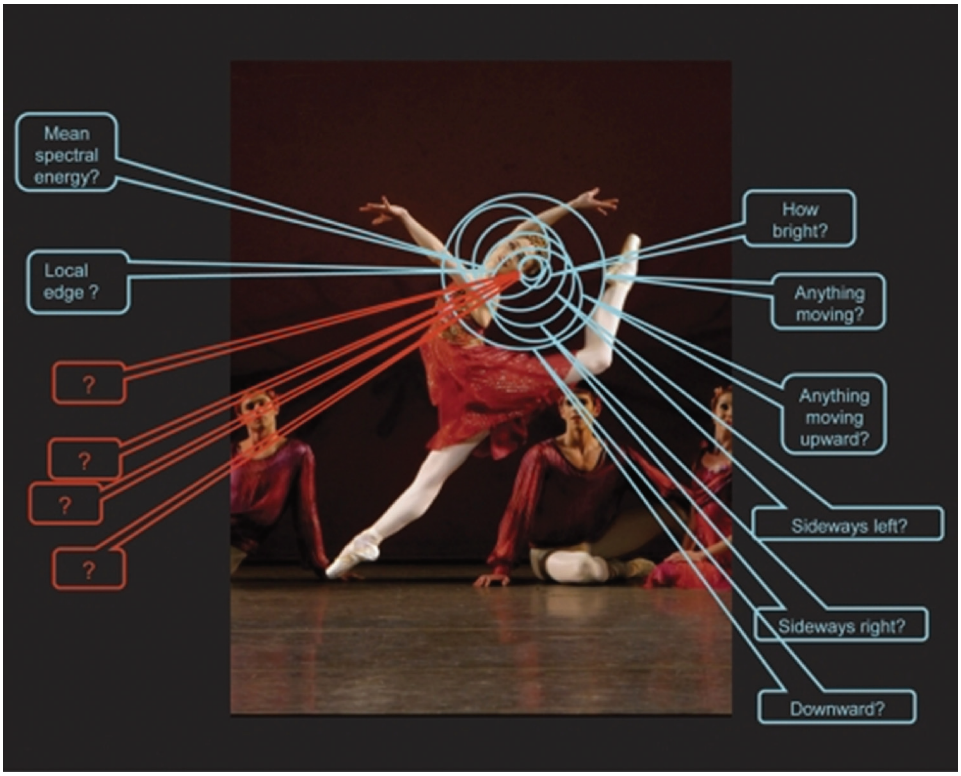
Mean spectral energy?
Local edge ?
?
?
?
?
How bright?
Anything moving?
Anything moving upward?
Sideways left?
Sideways right?
Downward?

FIGURE 5-14 The numerous components that contribute to an image are managed by the retina on a point-by-point basis. Twenty types of ganglion cells provide a unique inner retinal surface mosaic so that each is represented at any point in the retina corresponding to a point in space which is the object being viewed. In the above image, a small area on the dancer's head is assessed for various features, e.g. motion, direction, tilt, spectral composition (colour), tone, texture and more. (From Masland, 2011.)

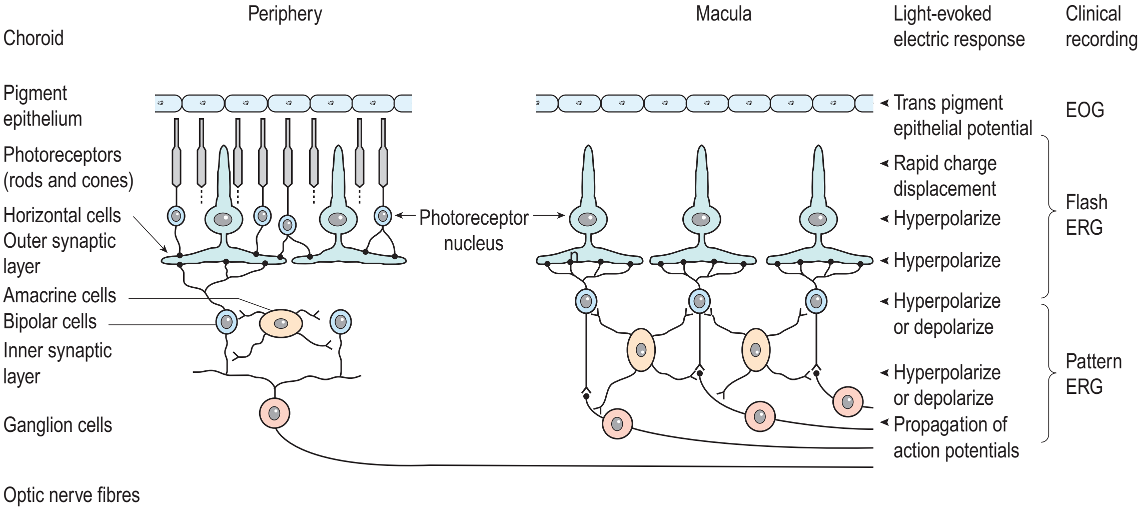
Periphery
Macula
Light-evoked electric response
Clinical recording
Choroid
Pigment epithelium
Photoreceptors (rods and cones)
Horizontal cells
Outer synaptic layer
Amacrine cells
Bipolar cells
Inner synaptic layer
Ganglion cells
Optic nerve fibres
Photoreceptor nucleus
Trans pigment epithelial potential
Rapid charge displacement
Hyperpolarize
Hyperpolarize
Hyperpolarize or depolarize
Hyperpolarize or depolarize
Propagation of action potentials
EOG
Flash ERG
Pattern ERG

FIGURE 5-15 Source of electrophysiological 'traces' from retinal layers. The electro-oculogram lies at the RPE layer, while the ERG is retinal. The pattern ERG represents the response after processing in the bipolar cell layer. (Courtesy of A.M. Halliday.)

Resting potential and the electro-oculogram

Since the eye acts as a dipole, it possesses a measurable resting potential, which is generated at the interface between the retinal pigment epithelium (RPE) and the photoreceptors (the resting retinal potential) and is about 60 mV in height. At a molecular level, the electro-oculogram is representative of the transretinal pigment epithelial potential differences generated by separation of ionic gradients across the RPE by genes such as the bestrophin gene (see Ch. 4, p. 254), and maintained by tight junctions. Similar transepithelial potential differences occur across all non-leaking epithelial layers. As the eye becomes light adapted there is a steady rise in this potential, which is recordable as a reversible potential on horizontal eye movement and is known as the light rise (Box 5-9). Ratios less than 1.5 are abnormal, while a ratio >2.0 is regarded as normal. This effect is the result of an extracellular flow of current caused by changes in the potassium concentration in the interphotoreceptor matrix. The EOG is measured as a ratio of the light peak (i.e. peak amplitude when light adapted) to dark trough (i.e. lowest amplitude measured in the dark). The electro-oculogram (EOG) is lost in conditions that disrupt the RPE–photoreceptor relationship, such as retinal detachment.

The electroretinogram

The electroretinogram (ERG) is superimposed on the electro-oculogram and is the cumulative electrical response to a light stimulus from all the retinal elements. It is affected not only by the intensity and duration of the stimulus but also by the stimulus wavelength and pattern, and the level of light–dark adaptation of the retina as for the sensory response itself.

The ERG has several components (Fig. 5-16). Under mesopic conditions, the early receptor potential is barely detectable and becomes apparent only with an extremely high-intensity flash stimulus in a deeply dark-adapted eye. It originates from the photochemical reactions in the rod outer segments on stimulation by light and is dependent on the density of rods and high levels of unbleached rhodopsin. The early receptor potential may therefore be detectable in eyes where the inner retina has been destroyed but the outer retina is substantially intact, e.g. central retinal artery occlusion. However, it is not normally recorded because of the above physiological constraints.

BOX 5-9 THE ELECTRO-OCULOGRAM

The electro-oculogram (EOG) is a record of the electrical dipole occurring between the front and the back of the eye and reversing in current direction when the eye moves from side to side. The height of the potential difference increases in conditions of bright illumination.

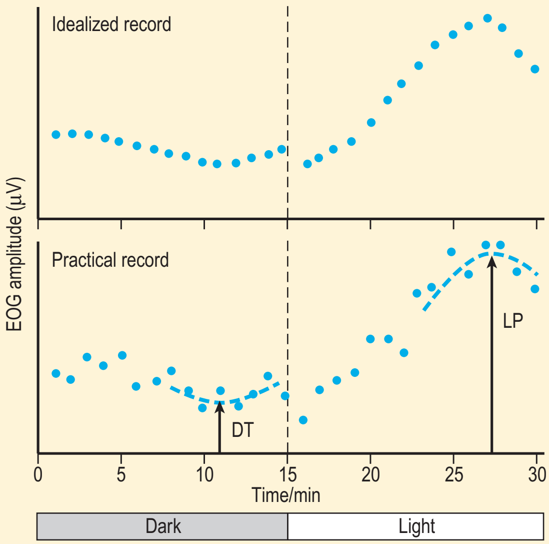

An EOG recording – vertical lines represent the alternating dipole as eyes are moved from left (L) to right (R). The 'light rise' is seen as an increase in the height of the vertical lines as the light is switched 'on'. The electro-oculogram (EOG) is measured in microvolts as a 'light rise', i.e. an increase in amplitude as the light stimulus is increased. The standard curve is shown in the upper two boxes on the left at low light levels and on the right as the light is increased. Actual EOG values are shown in the bottom two boxes. Idealized (*upper*) and practical (*lower*) EOG amplitude versus time curves for a healthy subject. The actual EOG is estimated from the points measured as in the dashed line. DT, dark trough; LP, light peak.

The negative 'a' wave is generated by hyperpolarization in the photoreceptors' inner segments (Granit's PIII component), the a1 component coming from the cones and the a2 from the rods. In contrast, the 'b' wave (Granit's PII component) is believed to come from the bipolar cells either directly or indirectly via signal spread to the Müller cells; b1 is generated by

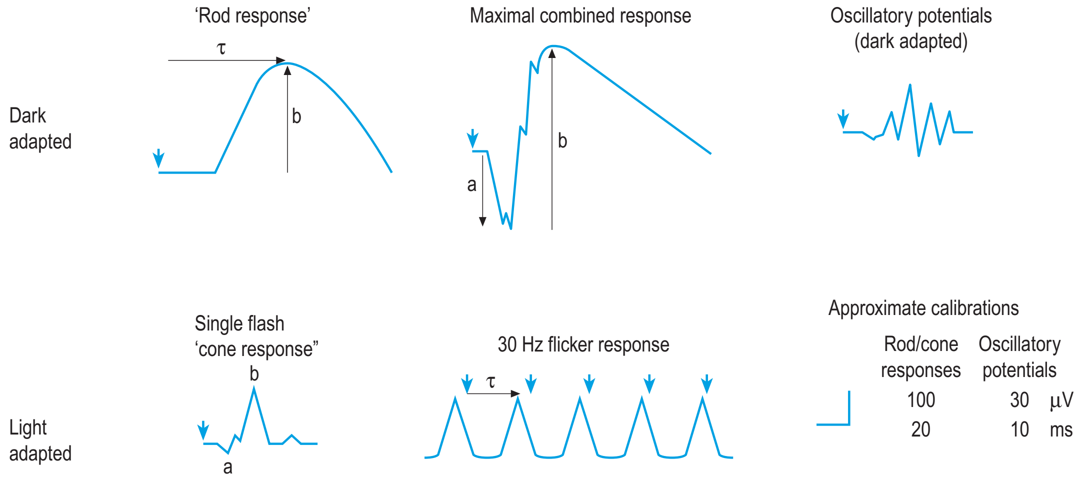

FIGURE 5-16 Diagram of the five basic ERG responses defined by the standard. These waveforms are exemplary only and are not intended to indicate minimum, maximum or even average values. Large arrowheads indicate the stimulus flash. Dotted arrows exemplify how to measure time to peak (τ, implicit time), a-wave amplitude and b-wave amplitude.

cone-dominated and b2 by rod-dominated bipolar cells. The b wave is lost in certain retinal vascular conditions, such as central retinal vein occlusion.

Oscillatory potentials are thought to be generated by amacrine cells, while the slow rising 'c' wave (Granit's PI) depends on an intact pigment epithelium. However, the electro-oculogram provides a more effective estimate of the integrity of the pigment epithelium. Oscillatory potentials are lost in patients with diabetes.

The ERG as described above is in essence a response to luminous intense stimuli. However, the pattern ERG (PERG), which is the ERG response to a reversing checkerboard of black and white squares of equivalent luminance, is thought to represent the electric response to spatial contrast, probably from ON-centre ganglion cells.

The multifocal ERG

The multifocal ERG records discrete electrophysiological responses from specified regions of the retina. An array of hexagonal elements in a changing frame is used to stimulate the retina rather that a full field flash of light and each element has a 50% chance of stimulating the defined retinal area (Fig. 5-17). Each stimulus is randomly illuminated over time, producing a continuous ERG response which is correlated with the ON and OFF phases of each patch and so the final result is not truly an ERG response from a discrete retinal region: instead, the response is interpreted as a statistical/mathematical resolved factor and produces a typical waveform (see eFig. 5-5).

Additional content available at https://expertconsult.inkling.com/.

The multifocal ERG can be evaluated in conjunction with complementary function and imaging techniques which assess the visual field: for instance, microperimetry is a technique in which the visual sensitivity to discrete stimuli of a specific area of the retina, identified by scanning laser ophthalmoscopy, is correlated with optical coherence tomographic imaging of the corresponding retinal region.

The visual evoked potential

The visual evoked potential (VEP) records electric activity from the occipital cortex following presentation of a light stimulus to the retina and represents a limited electroencephalogram (EEG). Recordings are taken from a set of six electrodes placed around both left and right occipital cortices, each producing a discrete wave pattern of different amplitudes (Fig. 5-18).

Several types of VEP can be induced, including flash, flash-pattern, pattern-onset, pattern-offset and

The multifocal ERG has greatly improved the detection of functional abnormalities in the retina as well as precise localization of such abnormalities to specific regions of the retina. In a standard examination protocol several data sets are taken, including: an ISCEV standard ERG (scotopic and photopic), the multifocal ERG, a pattern ERG (both transient and steady state) as well as the visual evoked potential using a large check and a small check stimulus (see eFig. 5-5). In addition, visual acuity and contrast sensitivity measurements are taken. Using this battery of tests, differentiation between generalized retinal dysfunction, disorders of the macula, optic nerve dysfunction and 'non-organic' disorders can be reliably made.

(Courtesy, Dr. Fred Chen, Lions Eye, University of Western Australia.)

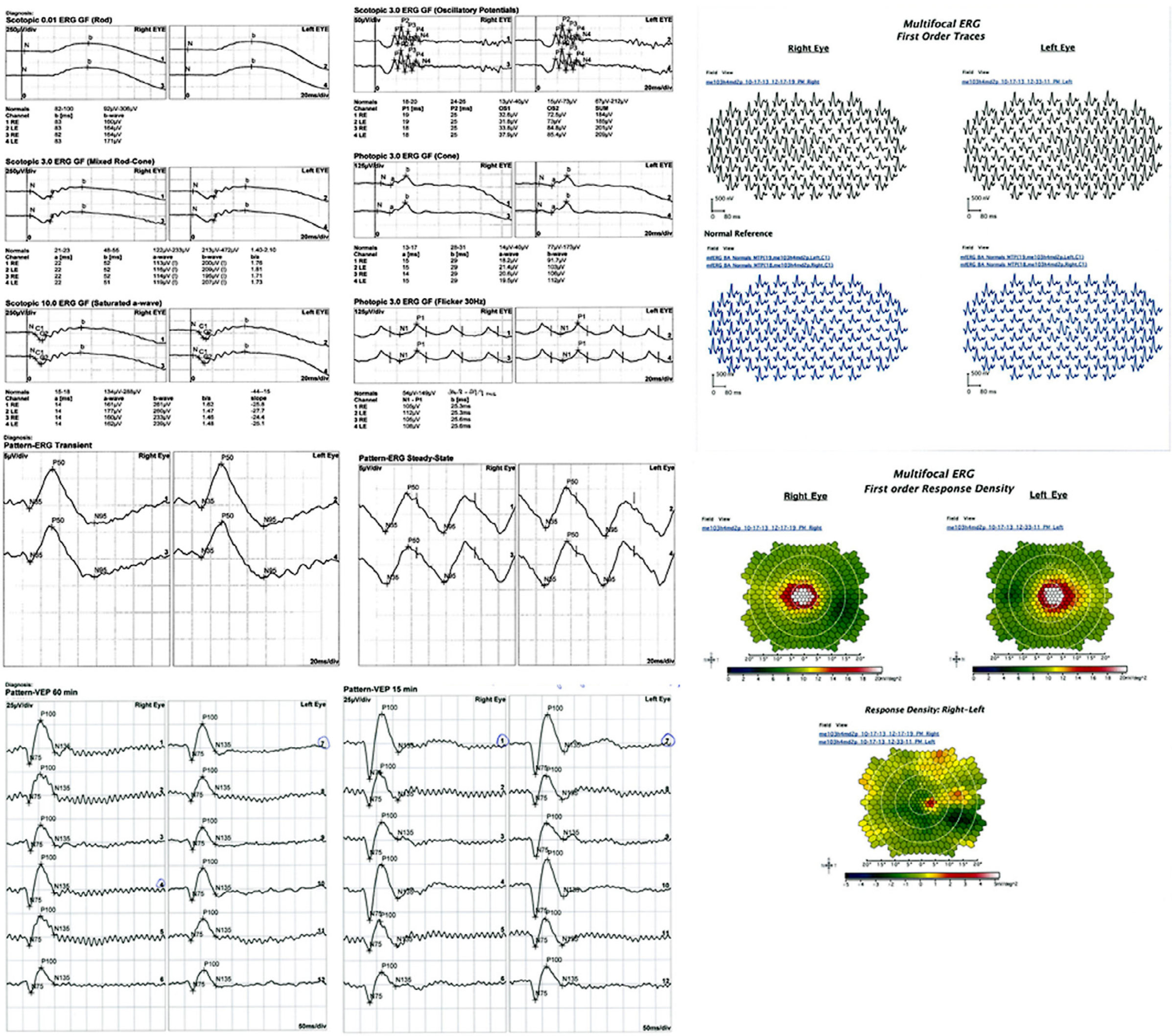

eFIGURE 5-5 Typical set of information obtained from electrophysiological evaluation of retinal and optic nerve function.

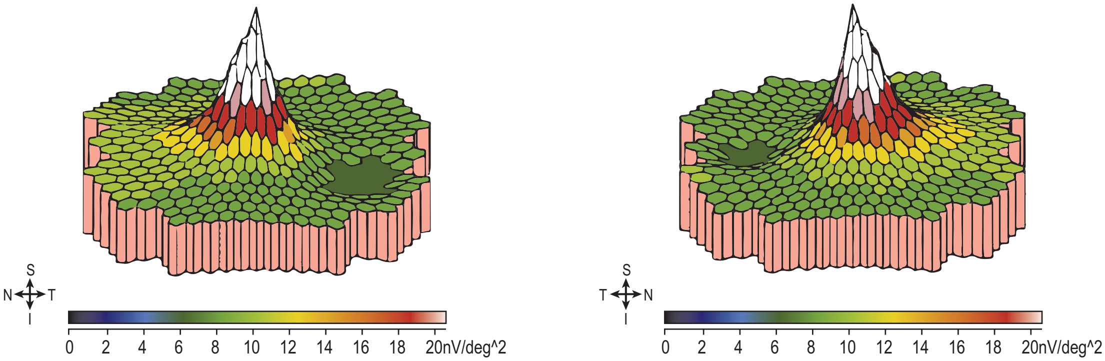

Response density: subject – normal mean (*StDev*)

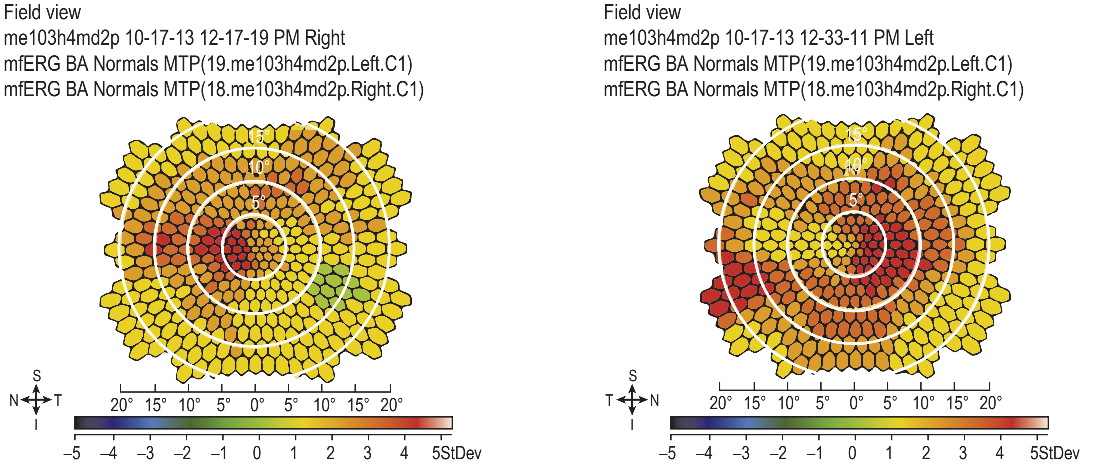

FIGURE 5-17 Normal multifocal ERG record showing the characteristic 'hill of vision' and the hexagonal arrays representing each retinal region analysed. (Courtesy Dr. Fred Chen, Lions Eye Institute, University of Western Australia.)

pattern-reversal. Considerable individual variability in the wave pattern is seen with the flash VEP, which is composed of two phases: the evoked potential and the after discharge.

Variations also occur in the amplitude of response, depending on the level of dark adaptation. A pattern-flash VEP is evoked when a black-and-white checkerboard stimulus is presented (Fig. 5-19). The amplitude of this response is considerably better correlated with the visual acuity. However, a significant electric interference in this response is caused by switching the stimulus on and off (the onset/offset response). The flash-evoked potential has three components – N1, P1; N2, P2; and N3, P3 – while the pattern-evoked potential is essentially monophasic (see Figs 5-18 and 5-19). The flash VEP is considered to arise from area V2 of the cortex with its retinal origins arising in the entire papillomacular bundle, while the pattern response is considered to arise in V1 plus ganglion cell receptive fields corresponding to large checks (M cells) and small checks (P cells).

Presentation of a pattern-reversed black-and-white equiluminant checkerboard can overcome the onset–offset problem if the pattern is reversed at an appropriately short interval, as the effects tend to cancel each other out. Indeed the stimulus can be set to produce pattern-onset, pattern-offset and pattern-reversal

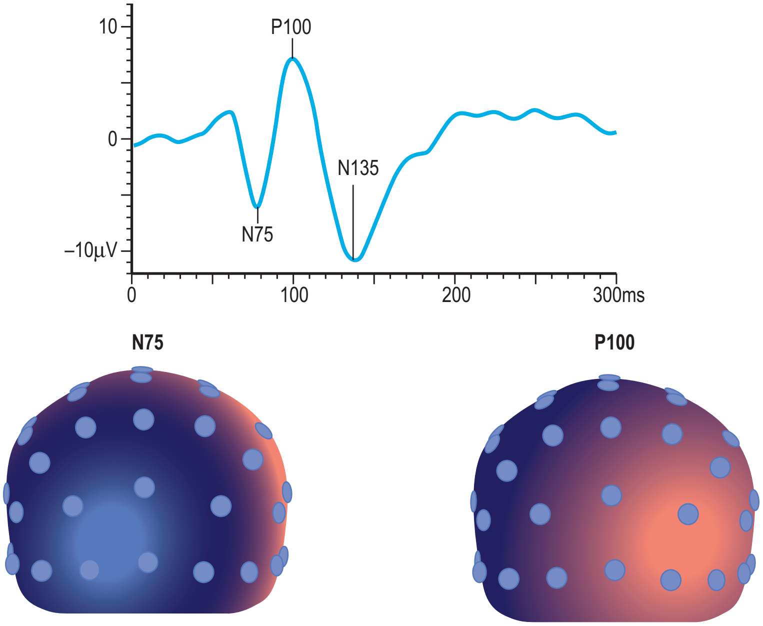

FIGURE 5-18 Diagram demonstrating the visual evoked response showing the potential map for the location of the electrical output from the scalp electrodes. (From Kevin Whittingstall: http://fizz.phys.dal.ca/~medbiophys/kevinp.htm.)

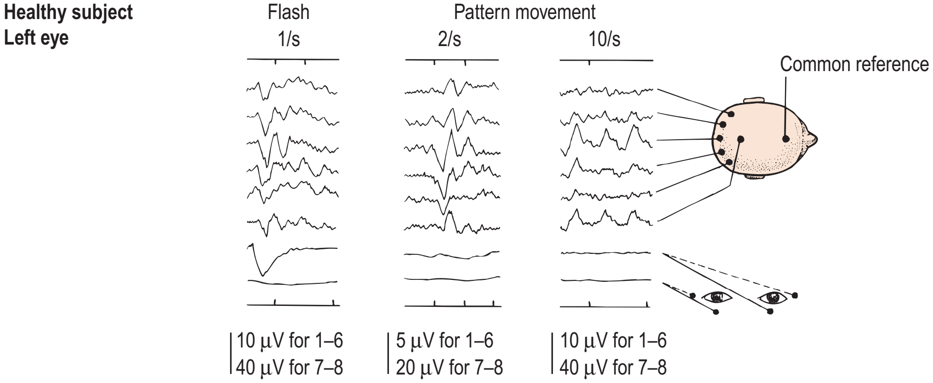

FIGURE 5-19 Flash and pattern-averaged VEP responses from each of six electrodes placed over the occipital region of the skull. Pattern reversals were performed at two different rates (middle and right panels). The two lower tracings represent simultaneous ERG recordings. (Courtesy of Dr Ikeda.)

VEPs, each of which has a characteristic set of wave patterns depending on the conditions (Fig. 5-19). Considerable ingenuity has been developed in techniques for studying half-field stimulation, macular vision and the effects of age, right versus left eye, etc. However, clinically, this test has greatest applicability in assessing the function of the optic nerve by measuring the latency of the response, and in assessing the integrity of foveal vision by evaluation of the pattern and amplitude of the wave forms.

FLICKER

A spoor or beam of light can flicker so fast that the sensation of flicker is lost. The point at which this happens is known as the critical fusion frequency (CFF) and the brightness of the steady-state light is the same as the mean brightness of the flickering light midway through its cycle. Brighter stimuli have a higher CFF and rods have less ability to achieve fusion than cones (see below).

The fact that fusion occurs at all indicates that there must be some degree of persistence of sensation after the stimulus has ceased, but clearly this is more effective at lower levels of illumination than at higher ones. Fusion is thought to be the result of an after-effect, so that a succeeding stimulus can fall on the retina while it is still responding to the first stimulus. However, it is more probably related to light adaptation in which the response to light is accelerated at higher intensities. In terms of flicker measured by an ERG response, the CFF is seen as a smoothing out of the electric response, which is set at a higher level in millivolts. Presumably it would be possible to induce a second ERG response by presenting a superimposed flash on this new level of illumination.

Subjectively, fusion occurs at 60 cycles/s, but the ERG CFF occurs at 25 cycles/s. Therefore, by setting the flicker cycle at 25 cycles/s, one can obtain, through the flicker ERG, a measure of cone function in isolation (see below). Interestingly, direct electrophysiological studies on optic nerve fibres have shown that ganglion cells with a high spike discharge rate also have a high threshold for CFF and vice versa.

Colour vision

COLORIMETRY AND COLOUR DISCRIMINATION

Colorimetry is a measure of visual function at the photoreceptor level. In contrast, colour discrimination is a cortical function related to perception and the later stages of visual processing. Colorimetry or the measurement of colour is based on techniques of colour matching, which have a long history going back to the days of Newton, Helmholtz and particularly Maxwell in the mid-nineteenth century.

For further information on Maxwell, see additional content available at https://expertconsult.inkling.com/.

Maxwell is credited with asserting that all vision is colour vision; in a sense, as will be seen later, this is probably a valid perspective. Standards for measuring colour have been established for certain reference conditions of illumination based on the assumption that only a small area of the central fovea is illuminated by the test stimulus; a standard V (λ) or visible wavelength curve was adopted by the Commission Internationale de l'Eclairage (CIE) in 1964. A colour (C) is specified in terms of the three primary colours by the equation:

$$C = r \cdot (R) + g \cdot (G) + b \cdot (B)$$

and is measured with colour-matching instruments such as the flicker photometer or with spectrophotometers fitted, for instance, with arrays of wavelength-selective photodiode detectors.

DIFFERENT COLOURS HAVE DIFFERENT LUMINOSITY

Spectral sensitivity curves

In the dark-adapted state, light of different wavelengths appears variably bright with a peak luminosity at about 500 nm in the blue–green region, i.e. for lights of equal energy, blue–green appears brightest in the dark (Fig. 5-20). Under photopic conditions, however, peak brightness occurs around 555 nm in the yellow–green. Brightness curves of different wavelengths like this are determined under photopic conditions by flicker photometry. This phenomenon, in which short wavelengths become brighter compared with long wavelengths as luminance is reduced, is known as the Purkinje shift, and begins under mesopic conditions when cone function is still active. Experimentally, the flicker ERG is useful in studying the Purkinje shift because rod and cone responses can be distinguished by setting the flicker rate above 25 cycles/s, which rod photoreceptors cannot detect (see

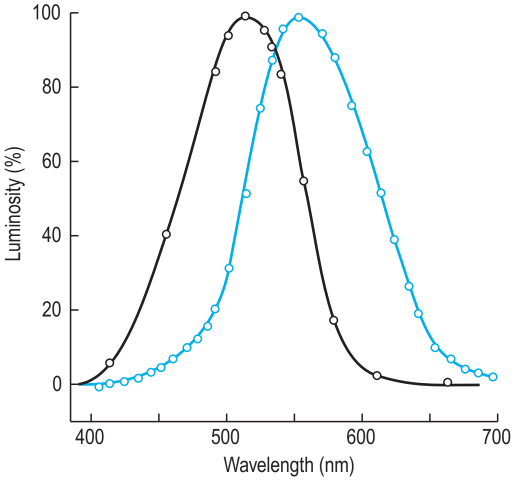

FIGURE 5-20 The Purkinje shift showing differences in luminosity observed at different wavelengths.

Described as ‘the man who changed the world’, James Clerk Maxwell is known as the most influential physicist of the nineteenth century, ranking with Isaac Newton and Albert Einstein. He developed one of the major unifying theories of physics, namely that of electromagnetism, demonstrating that light energy was one component of this unifying theory of electricity, magnetism and optics. Maxwell was born in Edinburgh and studied at Edinburgh University and after further training at Cambridge became professor of Natural Philosophy at University of Aberdeen, before moving to University College London where he developed most of his seminal work. It was while in London that he completed much of his work on colour and produced the first ‘light-fast’ colour photograph. He also completed his work on electromagnetism.

above). Purkinje shifts are therefore not detectable in guinea-pigs (pure rod) or squirrels (pure cone), but can be detected in cats, which are rod-dominated.

Although cone and rod responses can be differentiated using the flicker ERG, it is also possible to differentiate long (L) and middle (M) wavelength responses using appropriate flicker frequencies. In addition, using mixed luminance (rod) and chromatic (cone) flicker frequencies around 12 Hz and appropriate colour backgrounds for the stimuli, it has been possible to separate out the luminance and colour electrophysiological responses detectable at a retinal (ERG) level rather than at a post-signal processing (cortical /perceptual) level.

Photochromatic interval

The Purkinje shift underlies the photochromatic interval, which is a measure of the difference in 'brightness' between the absolute threshold at which light of any wavelength is just detected and the brightness at which it appears coloured. Clearly, this interval is vanishingly small at the red end of the visible spectrum and is maximal at about 570 nm.

Cone thresholds

As we have seen above, thresholds are an artificial concept that depend on a critical number of 'hits' on photoreceptors by quanta of light. In practice, under defined conditions, light thresholds are measurable and have been well characterized for rods. Cone thresholds can also be measured, for instance by using only the early part of the dark adaptation curve or by using very small bright flash stimuli, which only impinge on the central fovea. By choosing suitable conditions of light adaptation (e.g. by adapting with blue light to desensitize the rods), it is possible to measure cone thresholds with different wavelengths of light. Using wavefront adaptive optics technology which minimizes chromatic and other aberrations, the absolute threshold for cone vision has been determined at 203 ± 38 photons at the cornea and is considerably greater than previous estimates. Using these methods cone-specific spectral curves can be produced.

COLOUR DETECTION REQUIRES MORE THAN ONE TYPE OF PHOTORECEPTOR

Photoreceptors respond to stimuli by changes in the frequency of electric discharge. Indeed, this is true for all neuronal impulses. Since photoreceptors respond to both luminosity and wavelength, a retina that has a single type of photoreceptor (such as a rod) will not be able to distinguish one stimulus from the other under different circumstances. Wavelength discrimination therefore requires a panel of photoreceptors (minimum of two) with peak responsiveness at specific wavelengths independent of their responses to changing luminosity. In theory, the more variety in receptor type with specific spectral sensitivities, the greater the ability to discriminate wavelength, as for any single wavelength stimulus it is the pattern of discharges from the entire panel of receptors that determines this discriminatory ability. Probably through evolutionary constraints, two types of cone photoreceptor provide sufficient discriminatory ability for survival of the species and the red–green medium (M) to long (L) wavelength separation is an additional component restricted to primates (see above under Retinal connections).

Trichromatic theory of colour vision

A specific colour or hue is therefore detected by the summation of responses from a mixture of receptors, and the contribution from each of the three primary photoreceptor types can be deduced from spectral mixing curves (Box 5-10). Indeed, such colour mixing phenomena are the result of 'confusion', i.e. our inability to differentiate sufficiently narrow wavelengths. This is a reflection of the physiological limits on wavelength discrimination set by our having only three cone photoreceptors. (Certain species of fish have four cone photoreceptors.)

However, hue discrimination is more than a retinal-processing phenomenon. Experiments testing the ability of humans to detect the four unique hues (red, green, yellow and blue) using three chromatic mechanisms in P cells tuned to detect L/M and S-L/M show that hue discrimination requires higher-order colour perceptual mechanisms.

Psychophysical evidence for three cone photoreceptors

An intrinsic property of sensory receptors is to adapt and this is well demonstrated by rod (dark adaptation) and cones (light adaptation). Experiments using cone thresholds and light adaptation techniques have confirmed that there are three types of cone that respond differently to the same wavelength of light. These are

BOX 5-10 SPECTRAL MIXING CURVES

Mixing the three primary colours will produce any secondary colour or hue. However, a specific quantity of the primary colour is required to produce each hue, and this amount is determined by spectral mixing curves.

(**A**) Graph showing relative amount of each primary colour required to match any specific wavelength.

(**B**) Actual frequency of discharge in retinal neurones for each primary colour at any specific wavelength.

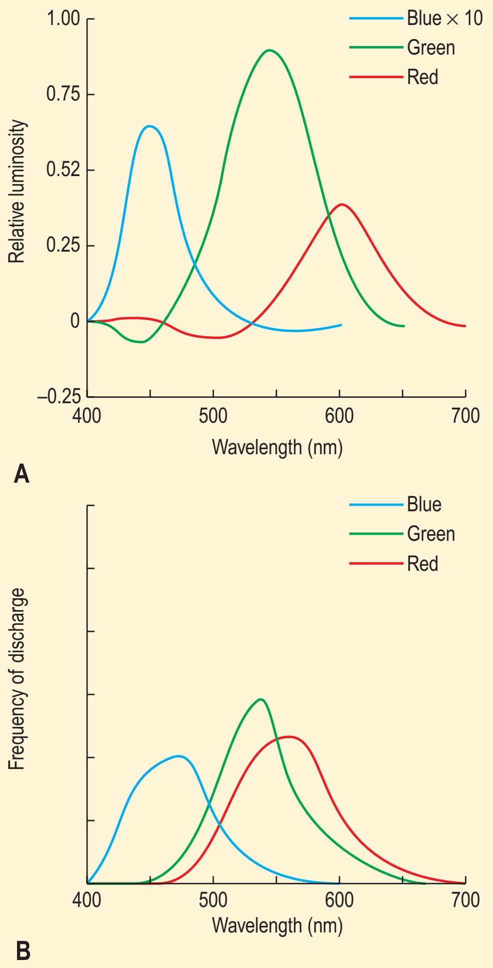

reflected in the spectral sensitivity curves for the three photoreceptors, which form the basis of the chromaticity chart (see pp. 272 and 275). Light adaptation is accompanied by several molecular events: (1) changes in intracellular calcium level regulating guanylate cyclase activity in some species, duration of opsin activation and channel opening; (2) phosphodiesterase activity; and (3) a slower mechanism, possibly involving interaction between dopamine release and melatonin (see Ch. 4, pp. 247–248). Studies of colour-blind individuals have also provided confirmatory evidence of the trichromatic theory.

Similar results can be obtained using a technique known as the liminal brightness increment (l.b.i.), in which the amount of additional light required to produce a detectable difference in the brightness of a target against a changing background luminance is determined. The technique can be highly discriminatory by measuring, for instance, the liminal brightness increment in a blue central target against a green background. The studies confirmed that there are indeed single receptors that peak in the red and green regions, but the blue spectral sensitivity curve is more complex and there are probably three components for the blue mechanism. Whether the influence of rod mechanisms on these tests can be completely eliminated is not clear. Techniques combining spectral reflectance densitometry and adaptive optics in healthy volunteers have demonstrated the unique arrangement of S, M and L photoreceptors. Interestingly, despite normal colour vision, the variation in density of L and M cones, in particular, varies considerably between individuals (Fig. 5-21). It has been suggested that this may represent a compromise between the competing requirements for spatial versus colour vision, and this is reflected in the patchy distribution of the L/M cones at the fovea and their different distribution in the peripheral retina.

Molecular evidence for three receptors

Just as rhodopsin represents the molecular receptor for light energy at the level of the photon, so there are cone pigments that are sensitive to photons of light generated within specific wavebands. The amino acid sequences of these proteins are known and there are surprisingly few differences in the three cone opsins, especially in the transmembrane regions of the proteins that are important for retinal binding (see Ch. 4, eBox 4-5) (Nathans et al., 1986). These differences must, however, account for peak wavelength sensitivity of each of the three opsins, thus revealing the extraordinarily fine spectral tuning that occurs at

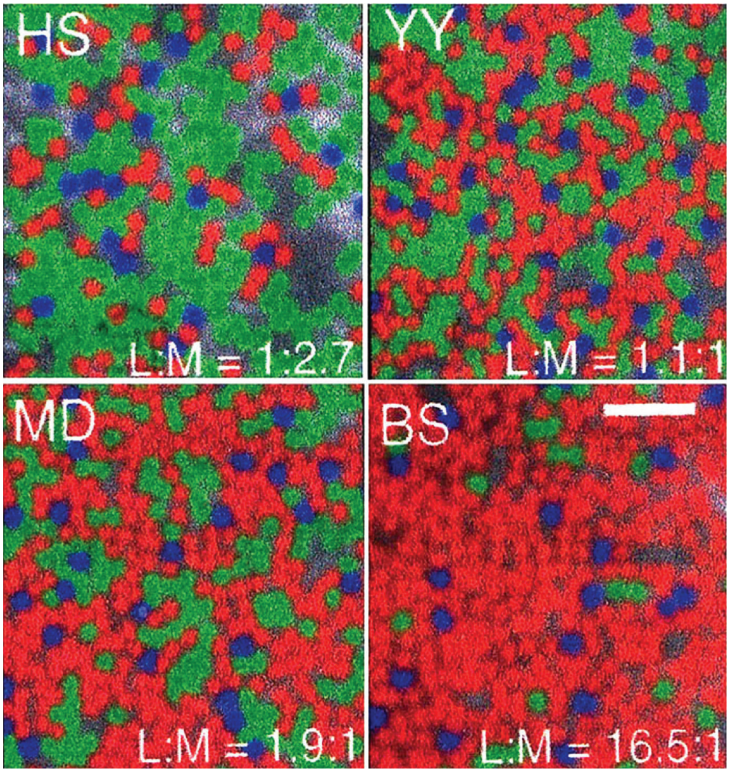

FIGURE 5-21 Images of long-wavelength (L, red), medium-wavelength (M, green) and short-wavelength (S, blue) cones as seen using adaptive optics in three healthy individuals. There is a constant ratio of blue to red and green cones but considerable variation exists between red and green cones between individuals. (From Webvision: Color vision, by Peter Gouras: http://retina.umh.es/webvision/color.html.)

a molecular level. For instance, a threonine at position 65 correlates with the 'red' opsin, while isoleucine is present at the same position in the 'green' opsin. Although the mechanism of light-induced activation of rhodopsin and cone opsins is in principle similarly based on the conversion of 11-*cis* retinal to all-*trans* retinal (see Ch. 4, p. 256), the differences in detail are highly physiologically significant. For instance, although the light response by cones is 100 times less sensitive, the cone opsin responses are several-fold faster, capturing something in the order of 500 photons of light per second. This is related to the availability of Ca^{2+} ions, regulated by a guanylate cyclase-activating protein. In addition, about 10% of cone opsin may be in the apo-state (i.e. lacking any binding to retinol), but still retains a sufficient level of activity to weakly activate transducin, in what is termed 'dark' noise (see Box 5-6).

Studies using microspectrophotometric techniques and infrared photography have shown that the distribution of the photopigments in the short-wave, middle-wave, and long-wave sensitive cones (Fig. 5-21) is not interdependent as might have been expected on the basis of receptive field analyses (see below), but is random or clumped at least for the long- and middle-wavelength receptors. In the case of short-wave receptors, some degree of organization has been observed from immunohistochemistry data, which shows that there are different P ganglion cells for spatial, chromatic and other functions (see section on Retinal connections). In the foveal region, blue-sensitive cones are by far the most infrequent while, in humans at least, psychophysical studies suggest that long- to middle-waveband cones exist in a ratio of 2 : 1 (Fig. 5-21).

CONVERGENCE, YOUNG–HELMHOLTZ AND HERING

Responses between photoreceptors and neural cells

Considerable processing of information occurs in the retina between the photoreceptor and the ganglion cell. In the magnocellular pathway, which deals with light and motion detection, M ganglion cells have large receptive fields and many rod photoreceptors feed information indirectly onto a single ganglion cell (see above). In contrast, the parvocellular pathway deals with spatial and colour vision, P cells have small receptive fields, and a single cone cell may have sole input to a single bipolar cell (see Ch. 1, pp. 45–47). At a retinal level, convergence in the case of colour vision is non-existent (although there is convergence for colour in the cortex).

In spite of this, sensory perception of colour does not correlate directly with stimulation of a wavelength-specific cone. As we have seen, a red-specific bipolar cell responds simply by producing a change in the firing rate in its nerve terminal, which of itself cannot be distinguished from a similar change in a green-specific bipolar cell, i.e. the bipolar cell cannot distinguish wavelengths although its receptor can. In addition, as there is considerable overlap in the spectral sensitivity of the receptors (see Fig. 5-20 and Box 5-10), some degree of confusion must exist in the initial response, which is smoothed out during processing.

Ganglion cell responses and opponent colour theory

Smoothing out this confusion is achieved by the colour-opponent mechanism based on the receptive field organization of the ganglion cells (see Ch. 4, p. 264). In this scheme, there are three colour-opponent arrangements: a reciprocal ON-centre red/OFF-centre green, in which the bipolar cell receives stimulatory or inhibitory input from a single red or green cone; a similar ON-centre blue/OFF-centre yellow; and a third ON-centre white/OFF-centre black in which the bipolar cell receives input from all three cones and in which colour mixing occurs. This has the effect of greatly refining the spectral sensitivity of the ganglion cell responses; it allows the perception of hues and 'unsaturated' colours and accommodates the Young–Helmholtz trichromatic theory of colour vision with Hering's colour-opponent scheme. It also goes some way towards explaining the various colour anomalies found in humans.

Studies have shown that blue-ON/yellow-OFF responses arise from a distinctive bistratified ganglion cell type derived from a dual excitatory cone bipolar cell input: an ON-bipolar cell receiving input only from S cones and an OFF-bipolar cell contacting L and M cone inputs. The mechanism of red–green opponency is still unknown. However, a recently proposed theory that incorporates a contribution of 'white' from rods stimulated under light conditions to the cone input allows a unified concept of how vision is integrated through simultaneous rod and cone input to allow subtle visual perception, such as hue discrimination, motion detection, orientation selectivity and others.

This last concept has been further modified with the discovery of the melanopsin-containing ganglion cells. Although these light-sensitive cells are considered to sense 'irradiance' (see Ch. 4, p. 249), it has been shown that they also can sense colour if the threshold is set sufficiently high. Thus while the trichromatic theory still holds true at the fovea, with three sets of cones detecting L+ (L+M), (L-M)+M, and (S-(L+M)) ON–OFF colour opponent receptive fields, in the periphery there is a fourth receptor, generating a tetrachromatic mode of colour vision detection, but whether this modifies perception of colours as tested in colour matching experiments is not known and the most accepted notion at present is that this fourth receptor influences visibility rather than specific wavelength detection (Fig. 5-22).

Colour constancy

We use colour to detect and recognize objects. Most of the light we detect is light reflected from objects, and their colour depends mainly on the surface properties of the object and not on the illuminating light. The wavelength of the reflected light clearly varies with the lighting conditions. In spite of this, the colour of an object remains the same, a phenomenon known as colour constancy, which is a function of higher visual processing in the cortex.

Factors determining the perception of colour (colour sensitivity) as well as other perceptions (lightness of an object, contrast sensitivity, sensitivity of depth perception) are influenced by several factors including attention, fixation and eye movement. It has been shown that we tend to take the brighter parts of an object much more into account when we make colour or brightness matches.

Additional content available at https://expertconsult.inkling.com/.

Both the illumination and the light reflecting from an object, as well as the composition of the scene, determine the amount of light reaching the eye. However, only the reflectance is constant and is therefore the most important factor for vision. When we look at an object we interrogate the scene both with eye movements and attention and where we look strongly influences our judgement of object properties such as lightness and texture. Our natural tendency is to fixate objects which are brightest and we fix even more closely on the brightest parts of an object and use this level of lightness to determine its colour (as in colour matching tests; eFig. 5-6). For instance in the figure shown opposite, Toscani and colleagues tested an individual's colour matching abilities using the set-up shown: the orange paper cone (**A**) is variably 'bright' from one side to the other and was one of a series of real objects used to match spectroradiometrically with a corresponding image on a cathode ray tube monitor (Toscani et al., 2013). Several real objects were used, such as a green wool ball, a green wool cylinder (same wool), green candle, red candle, yellow candle, and orange paper cone as shown in (**B**), while (**C**) shows the spectroradiometric data transformed to RGB (red–green–blue) values. The reason for this is that the brightest or 'lightest', i.e. most immunodominant, part of an object gives us the greatest information regarding the object's reflectance and, because it is constant, we rely more heavily on this for visual judgements. A recent review of colour vision, perception and colour constancy has been published by Conway (2009).

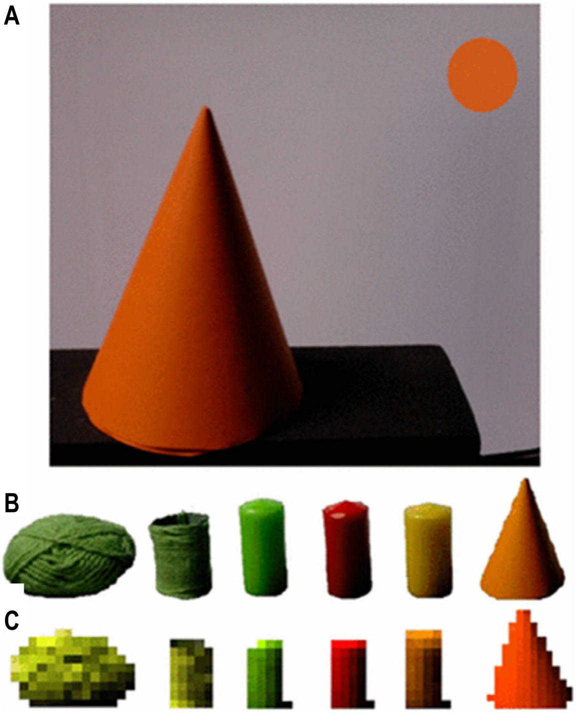

eFIGURE 5-6 Matching set-up for generating visual information (see text for explanation). (From Toscani et al., 2013.)

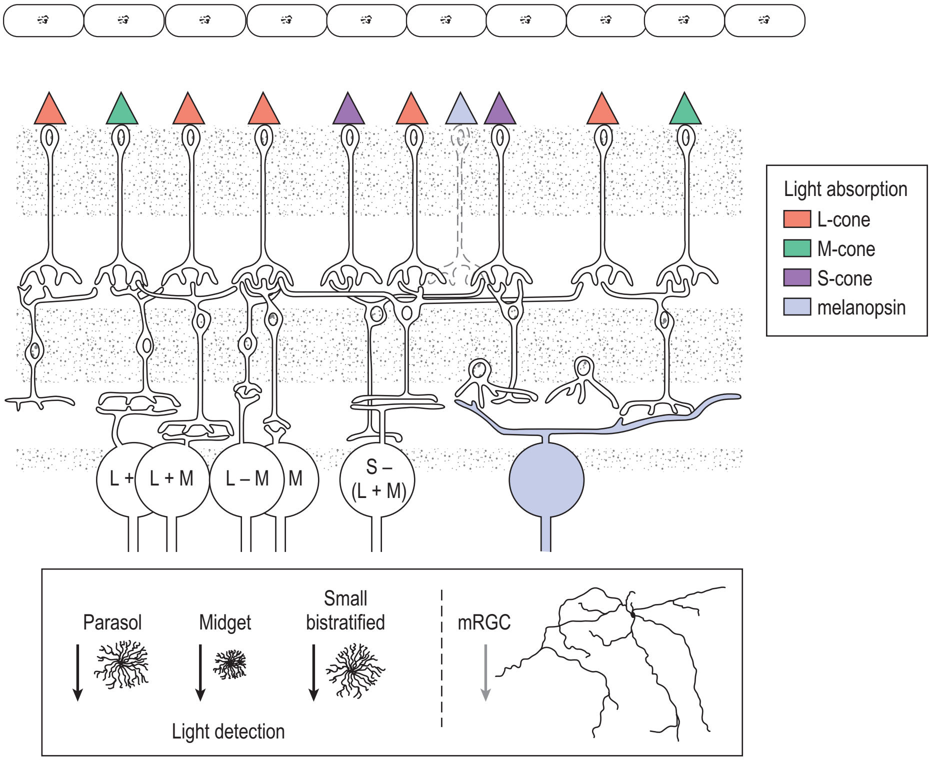

FIGURE 5-22 Schematic diagram of the retina showing the three light-sensing cones: in photopic conditions, rod vision is minimal (as Maxwell claimed: 'most vision is colour vision'). Detection of light is mediated by multiple types of ganglion cell serving the three main channels of composite light (L+M, L-M, and S-(L+M)). In addition, there is a small population of newly discovered ganglion cells containing a fourth photopigment, melanopsin, and most recently melanopsin-containing cones have also been discovered, but the nature of their contribution to vision is as yet unknown. (From Horiguchi et al., 2013.)

Remarkably, colour constancy remains stable during a normal life span despite changes in peripheral vision and lens transparency to yellow light. Cortical compensatory mechanisms are considered to allow this.

COLOUR BLINDNESS

Some of the defined colour vision defects can be explained in simple terms of loss of one or other specific type of receptor. However, in practice the situation is often more complex, involving not the loss of one particular receptor but the production of combination genes as the result, for instance, of aberrant crossing over in meiosis (see Ch. 3, p. 131); these produce proteins that are intermediate in their spectral sensitivities, thus reducing the range of responsiveness of the protein.

Monochromatism

Rod monochromatism occurs in about 1 in 30 000 of the population; such individuals have true achromatic vision, low visual acuity (0.1–0.3), find high-intensity lights uncomfortable, display nystagmus, and may have some signs of macular dystrophy. These patients have morphologically normal cones in their outer retina but their functional status is unclear. It has been suggested that they have a single type of blue cone.

Cone monochromatism is very rare (1 in 100 000). These individuals have normal visual acuity but cannot discriminate coloured lights of equal luminosity. Apparently, cone monochromats possess all three types of cones, indicating that the defect occurs in cortical processing, probably in area V4.

Dichromatism

Dichromatism occurs when the affected individual matches all colours with mixtures of two primaries. Therefore, the range of secondary colours is restricted. Protanopes are missing the red wavelength, deuteranopes the green, and tritanopes the blue. Mixing of the two colours will produce a sensation of white at certain specificities, which for protanopes is 495 nm and for deuteranopes is 500 nm. The dichromat cannot distinguish the large range of non-spectral hues from spectral hues as the trichromat can, leading to a much narrower range of colour detection by the isocolour charts.

Anomalous trichromatism

Red–green colour deficiency occurs in around 10% of males and is X-linked (see Ch. 3, p. 137). Due to their close relationship on the chromosome, unequal recombination events can readily occur, thus removing the green (deuteranopia) or the red (protanopia) genes. Defects in S cone opsin are much less frequent.

Red–green colour deficiency is rarely absolute and in effect such individuals are 'anomalous trichromats' in that they use different proportions of the three primaries to match colour. This is due to the fact that with one single L gene and several M isogenes (mostly single gene polymorphism at Ser180Ala on the red opsin gene), all of which have considerable homology, there is considerable scope for generation of hybrid genes. Thus, protans use more red, deutans more green and tritans more blue. The colour-anomalous individual differs from the trichromat and the dichromat in that he or she will not accept those matches that the other two agree on. This is the common form of 'colour blindness', occurring in 10% of the male population.

Achromatopsia

Colour blindness may also be the result of defects in cortical processing (area V4). Congenital (rare) or acquired lesions in the lingual or fusiform gyrus are associated with cerebral achromatopsia (also accompanied by prosopagnosia – a failure to recognize familiar faces, i.e. from memory). Similarly, cortical lesions in the superior temporal sulcus (V5) can produce defects in the ability to detect motion (motion blindness or akinetopsia; see p. 319). Isolated defects in form vision have not so far been detected, possibly because they involve more than one cortical area, e.g. V3 and V4, plus connections to other cortical regions involved in psychophysical attributes, and texture analysis.

Visual perception

Visual perception is the end product of the processing or reinterpretation by the cortex of sensory responses made by the retina to visual stimuli. However, a strict separation of cortical and retinal events does not occur as some degree of processing takes place in the retina and, conversely, certain processes such as instantaneous parallax (see below) occur so quickly that it is difficult to believe they occur exclusively at a cortical level. In addition, perception should not be considered solely as the end product of the processing of sensory information. Instead, it is part of the 'action–perception cycle', in which perception modifies activity, which then modifies perception in a continuous cyclic pattern, the boundaries of which become indistinct (Wexler and Boxtel, 2005). Motor activity (head and eye positioning, etc.), therefore, is central to perception such as stereopsis (see below) and colour (see above).

As shown above, there are many different types of visual stimuli, each of which may produce one or more different psychophysical perceptual responses. Sensory perception occurring at an elemental level encompasses visual stimuli such as luminosity, flicker, colour and form, because it involves simple processing of features such as points and lines or wavelengths, but even such an apparently fundamental function as visual acuity determination involves higher levels of processing because it is more than simply a point-to-point projection of the retina on the cortex.

Evidence for higher integrative activity at the cortical level comes from illusions such as the Schrödinger staircase and Rubin's vase as well as more specific illusions that involve both space and time such as the tilt illusions (Box 5-11). A clear example of the role of cortical function in visual perception is the phenomenon of colour constancy in which large variations in the chromaticity of an object, induced for instance by changing the wavelength of the illuminating light of the object, do not alter the perceived colour of the object: a yellow banana remains yellow even when illuminated by a green light (see above).

Our understanding of these processes has been greatly advanced by the careful analysis of both the stimuli and the responses, and has allowed specific functions to be attributed to discrete regions of the cortex.

MONOCULAR VERSUS BINOCULAR VISION

Positioning objects in space

Most of the primary visual sensory responses are monocular and are not changed by binocular viewing. Images of objects are projected onto definite positions in space (spatial perception) and each retina has its own delimited visual field. However, the position of an object in space is not an absolute entity but is related to the position of the observer and of other objects. The relative position of objects can be determined only if the retinal sensors are composed of discrete units that have precise 'markers' for localization. This, indeed, is the basis of the visual field.

In spite of this, objects appear fixed in position even when the observer changes position; simple ray diagrams demonstrate that a new set of retinal receptors must be stimulated every time the relative position of the object to the observer changes, but the observer does not experience the sensation of motion. This 'image stabilization' is achieved by compensatory psychophysical events at the cortical level. Recently, however, it has been shown that a proportion of this neural processing takes place at the retinal level through selective inhibition of ganglion cell firing (see above).

The existence of such mechanisms can be deduced in part from experiments showing that accurate localization of objects does not occur with all forms of eye or head movement. For instance, if the eye of an alert individual is forcibly moved using a surgical instrument such as a squint hook, the image is falsely projected to an incorrect position as if the eye has not moved. Spatial perception on normal eye movement must therefore be integrated with, if not controlled by, higher centres within the brain, such as the frontal cortical eye fields, which influence motor discharge in the ocular muscles. The corollary, of course, is that the proprioceptive stretch receptors in the ocular muscles

BOX 5-11 VISUAL ILLUSIONS

Visual illusions such as the Schrödinger staircase (**A**) and Rubin's vase (**B**) occur in a cyclical manner in which each of the perceptions regularly alternates. Such illusions have a periodicity and are in this sense time-dependent. In addition they can be modified spatially, as in the tilt illusion (**C**) in which a vertically oriented tilted image occupying a surround regions of vertical stripes will generate the illusion that the orthogonally vertical surround stripes are tilted in the opposite direction. This illusion can be overcome by adaptation using an image after-effect induced by gazing for 30 seconds or more at stripes tilted anti-parallel to the original image and then quickly fixating on the tilted image once more. The periodicity of the fluctuations is also alterable with drugs.

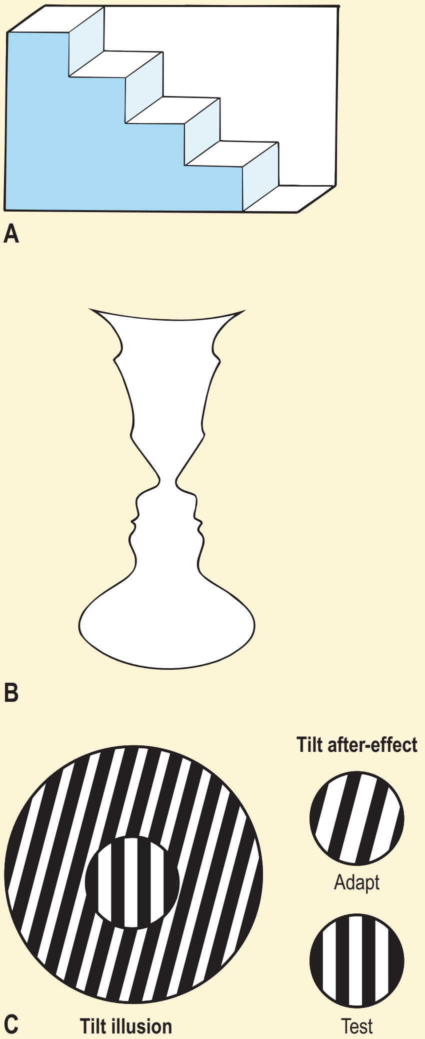

(**A**) Schrödinger's structure; (**B**) Rubin's vase; (**C**) the tilt illusion.

do not have a role in determining eye position as was previously thought, but that their probable role is simply to coordinate muscle tension in opposing muscles at a local 'axon reflex' level. This, however, may not hold true for all situations.

For instance, a problem arises in analysis of images that are perceived during slow visual tracking of a moving object. Despite the fact that no compensation is made for movements of the eyes during tracking, the changing position of the object is accurately observed and followed. This indicates that the higher centres are receiving a continuous flow of information from centres controlling ocular muscle movement, which is assimilated into the total information concerning object positioning; the actual adjustment of speed of eye movement to permit accurate tracking is achieved through visual input, which is ignored by the perceptual process. This means that having initiated a tracking movement, sequential images are interpreted on angular velocity assumptions determined by this initial response and ignored by the higher integrative centres. These assumptions may, of course, be inaccurate, especially if the velocity of the moving target changes. Thus, any induced errors require repetitive readjustments of the tracking response.

Conclusions derived from experiments such as these are greatly influenced by the design of the experiment. It has been observed that the perception of heading (i.e. the direction taken by the observer under conditions of radial retinal image flow, or optic flow) in a situation where a moving target is also fixated can be achieved accurately only if extraretinal information concerning the position of the eyes is available, i.e. via proprioceptors from extraocular and head and neck muscles. Under special circumstances, such information, termed structure from motion (SFM) information, can be integrated with purely retinal image information. In addition, the resultant perception can vary significantly depending on whether the object or the observer is moving, even if the relative disparity in motion between them is the same. This sort of information has direct relevance to clinical problems, as in the condition of oscillopsia, where image stabilization is lost and the patient experiences 'retinal slip' (see next section). In such patients it is not clear whether the defect lies in the well-established vestibulo-ocular reflexes or whether loss of a putative cervico-ocular reflex, via neck proprioceptors, is contributory.

Thus spatial constancy, i.e. maintaining a 'stationary' percept of moving objects despite their repeated and sudden changes in retinal stimulation, is a fundamental visual perception. Spatial constancy also has a memory component. When an illuminated object is viewed in the dark, the eyes will, after an initial 'searching' response, adopt a position close to fixation (approximately within 2° of the target) when the illumination is switched off (Fig. 5-23). This has been attributed to some element of positional sense from an extraretinal source, such as head–eye position in space and locomotion (egocentric versus allocentric signals). Similarly, the constant drift of the eyes towards fixation in the dark has been attributed to a similar mechanism (see next section). 'Place' cells occur in the hypothalamus and thalamus of the brain and correct visual input to active locomotor and possibly navigational (optic flow, see below) input.

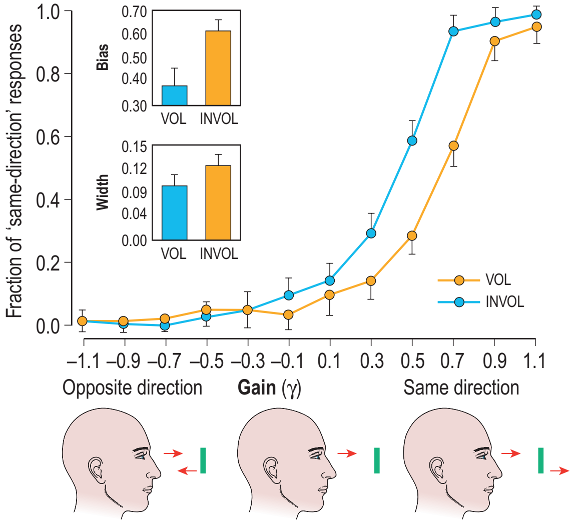

FIGURE 5-23 Results of the voluntary (VOL) and involuntary (INVOL) eye movement studies. The curves show responses averaged over subjects, and over all simulated distances (there was no significant effect of distance). Mean bias and width were calculated by fitting the data of individual subjects with the logistic curves and averaging the parameters thus obtained. Bars indicate between-subjects standard errors. (From Wexler, 2003, with permission from Blackwell Publishing.)

Measuring by eye

We frequently use our vision to measure things by eye, e.g. to line up objects in a row or to determine the distance between two points. Simplistically, it might be thought that, for example, the distance between two objects should correlate with a defined distance between stimulated receptors in the retina, or that the length of a line will be determined by stimulation of a fixed number of retinal receptors. However, the psychophysical basis of these abilities is not as simple as it might appear. For instance, two lines of the same length running in different directions may stimulate a different number of receptors as the direction of lines will be distorted by the curvature of the eye and its position on eye movement. In addition, errors will be introduced on head movement since the head rolls further than the eye in the socket (some of this disparity is compensated for by input from the semicircular canals). Attempts have been made to adjust for image distortion due to eye curvature and movement using after-images, which obviously must reflect closely the actual retinal stimulation versus the 'real' direction of the object, but even this is not sufficient to account for the fact that we perceive the object in its true position and not in its distorted position. This is nicely demonstrated by Aubert's phenomenon (Box 5-12) in which the speed of a moving object appears slower when the object is tracked by eye movements than when its motion is detected with the eyes remaining in one position, i.e. the object is not visually tracked.

A compensatory mechanism based in the higher centres must therefore be in place; it must be highly developed because the accuracy of measuring by eye is extremely high: for detecting the orientation of horizontal lines the error rate is 0.2% and for vertical lines

BOX 5-12 AUBERT'S PHENOMENON

A vertical bright light viewed in a completely dark room will tilt to the left if the head is slowly tilted to the right (**B**). If the head is tilted suddenly or if the line is viewed in the light, the line appears upright in its normal position (**C**). Thus information on retinal position is fed via the semicircular canals to the object-positioning centre; (**A**) resting position.

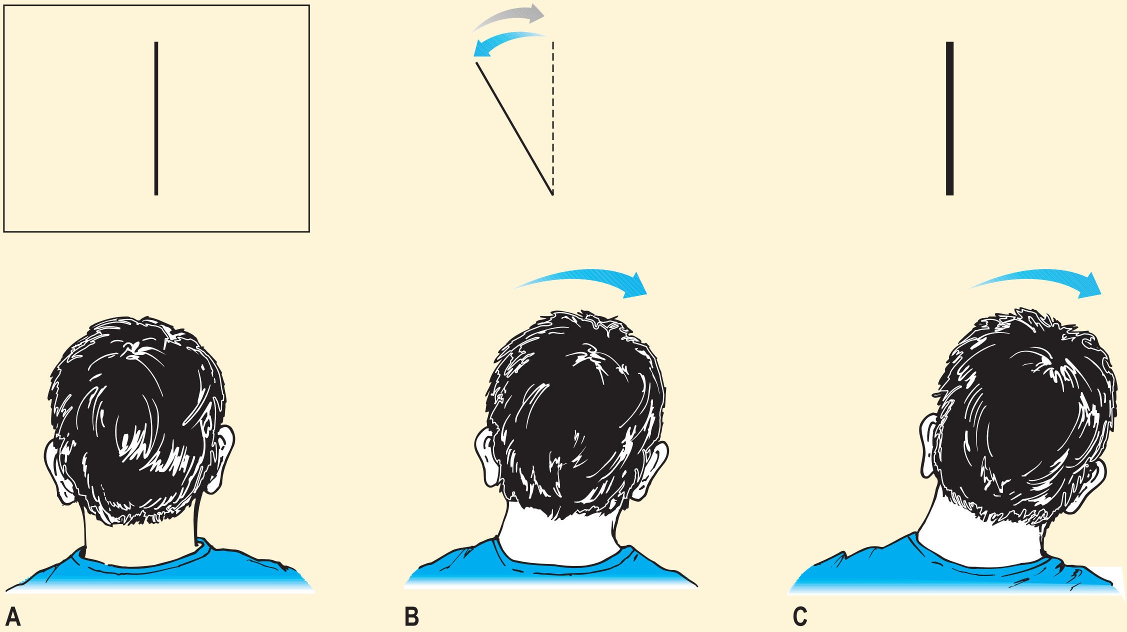

(**A**) Resting position; (**B**) slow head tilt to right, object appears to move to left; (**C**) fast head tilt to right, object stays upright.

it is 1.0%, as determined with a perceived direction test. The greater accuracy for estimating vertical and horizontal lines over oblique lines is not the result of preferred direction of eye movements or of the numbers of retinal receptors stimulated but is a function of cortical activity.

Considerable evidence has now accrued to show that the visual cortex contains specific cells that are responsive to the orientation of lines. Indeed, the ability to 'parse visual scenes for the orientation of purely spatial cues' has been shown to be a fundamental property of even the simple insect brain. Vertical and horizontal orientation detectors occur as simple cortical cells (see below) and are distinct from motion detectors. Orientation detectors are considered essential to the analysis of form.

The patterns of objects also greatly influence our ability to make visual measurements. For instance, two angles can be accurately compared if we can fixate each of them directly and if the sides are parallel. In addition there are numerous examples of optical illusions where objects to be compared appear to be different in length or area/size if one of them has been altered by the addition of other visual cues that are interpreted by the higher visual centres in one particular direction (see Box 5-11).

Therefore, both patterns and directions are important in visually estimating object dimensions: the retina uses the horizontal and vertical meridians as *x* and *y* axes to provide coordinates and thereby to pinpoint objects in space. As these axes are subject to displacement by, for instance, eye movement, it is important that a psychophysical compensatory mechanism exists that will reinterpret the position of the co-ordinates to project the real position and direction of the object in space.

Are two eyes better than one?

As most 'objects' are three-dimensional, it is clearly important that depth perception is achievable by the visual system. The major advantage obtained by binocular viewing is that it permits depth perception or stereopsis. This occurs in the cortex and depends on the fusion of images from each eye. However, depth perception is a complex event and information from many sources is used to achieve it.

One mechanism for depth perception could be related to the convergence of the eyes. Fixation of an object with both eyes requires a variable degree of convergence depending on the distance of the object, and information derived from this can be utilized to determine object distance. Indeed it has been suggested that variations in convergence can lead to three-dimensional illusions, although these may be caused by other mechanisms.

Is it possible to perceive depth with one eye only? Determination of object position using *x* and *y* axes as described above provides two-dimensional information only. Geometrically, it should not be possible to obtain three-dimensional information using one eye. However, certain cues, mostly built upon previous experience, indicate that some sense of depth is possible, e.g. by comparing the relative size of objects (for example, a person and a house), the blue colour of distant mountains (although they should be yellow – they are darker blue in comparison with the background light blue of the sky), overlapping edges of objects, effects of light and dark shading, effects of texture, and parallax on movement of the observer's head. It has also been suggested that the sensory feedback from the ciliary muscles on accommodation might provide some information centrally regarding depth (similar to the effects of convergence when both eyes are used), but this is unlikely.

When both eyes are used to observe an object in the straight ahead (primary) position, the image is still perceived as one, even though the image of that object must appear positionally but symmetrically different to each eye individually – indeed the image can be treated as if it were projected from a single centrally placed eye (the 'cyclopean' eye). When a second object is presented in the primary position, but closer to the observer than the original object, the second object is seen double when the original object is fixated (Fig. 5-24). Diplopia in this position is described as heteronymous; when the second object is distant from the first, the induced diplopia is described as homonymous. However, this form of diplopia is rarely appreciated because we normally do not attempt to fixate more than one object in the primary position when viewing with both eyes. When we aim at a target using a second object to line up fixation, as with the sights

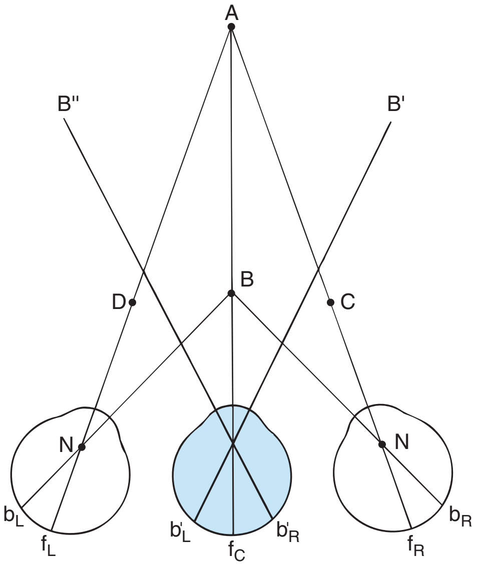

FIGURE 5-24 The 'cyclopean eye' (shaded) is the abstract notion representing fusion of the two images from left and right. Image A is fused ($f_L + f_R$ at f_C) but image B cannot be simultaneously viewed without diplopia (heteronymous) ($b_L + b_R$ at b'_L and b'_R) due to the projected images at B′ and B″, respectively. Homonymous diplopia would occur if B was distant from A. Using either eye alone for alignment of the object would entail point C aligning with A (right eye) and point D aligning with A for the left eye. N, nodal point. (Courtesy of H. Dawson.)

of a rifle, we normally do so with one eye only (see Fig. 5-24).

A similar form of diplopia can be induced by a divergent squint, or by placing a base-out prism in front of one eye; both of these conditions have the effect of re-siting the projection of the image to a 'false' position from its normal cyclopean position. False projection in a squinting eye can lead to the development of a 'false' macula, which in fact is a cortical event because the anatomy of the retina remains the same. The development of a false macula reflects the plasticity of the cortex and indicates that projection of the eye through the nodal point is an innate mechanism. A similar 'pseudofovea' may develop in the presence of hemianopia.

The fact that we rarely experience double vision, even when rapidly and simultaneously fixating many objects in a scene, indicates that the images from both eyes are merged or fused; this is not simply a reduplication of information from both eyes but an actual psychophysical event that is used to provide depth information or stereopsis.

STEREOPSIS AND DEPTH PERCEPTION

The fusion of images to create the perception of depth requires certain conditions: first, the images from each eye must have corresponding points on both retinas. If all points on a sizeable object were exactly corresponding, however, this would merely lead to a reduplication of information and it is likely that only one of these points would register (this would be analogous to allelic dominance in chromosomal gene duplication in which only one gene is expressed transcriptionally while its partner is not; see Ch. 3, p. 137). Single-cell measurements have shown, however, that impulses from both retinas induce electrical activity in cortical neurones.

The second requirement for image fusion is that a certain proportion of points are non-corresponding, and represent the differences between the two images on the retina, which may be smaller than the width of a single cone, and are known as binocular disparities.

Additional content available at https://expertconsult.inkling.com/.

It is the integration of information from corresponding and disparate points that induces the perception of depth. Both position and phase disparity in the corresponding receptive fields are important in the detection of depth cues (Fig. 5-25).

The horopter

Psychophysical and theoretical investigation of the projected points in space where an object is seen as a single image have a long history going back to the ancient studies of Ptolemy and the work of Ibn al-Haytham in the eleventh century. The corresponding regions in the retina include both the horizontal and vertical meridia and equidistant points from each of these meridia. Single vision can be achieved only when the images of the object are projected from each eye to the same point in space.

Corresponding points can be charted as a 'horopter', in which specific points on the retina project to definite single points in space – within the field of binocular single vision (Box 5-13). The vertical horopter has a backwards tilt that passes through the fixation point and a point near the feet of the observer

The images registered through each eye are different (see open access link to http://www.vision3d.com/stereo.html) and the brain processes these images, making note particularly of the differences (disparities). Simultaneous registration of the images is thought to be necessary for the images to be fused, although there is evidence that each image is registered differently in time, i.e. each eye takes a 'shot' of the image with a very brief time interval between the registrations. Stereopsis can be measured with a variety of devices, one of the oldest being the stereoscope, which is essentially an instrument constructed from a pair of base-out prisms, whose position can be manipulated. The stereoscope has a long history but continues to be modernized, one of the latest being the View-Master® which is now fully within the digital age and includes a video-based version (see http://gajitz.com/view-master-grows-up-modern-stereoscope-video-viewer/).

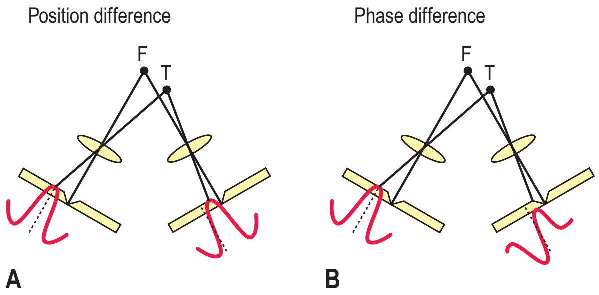

FIGURE 5-25 Depth perception based on binocular disparities. The fovea of each eye fixates point F; because object T is closer to the observer than F, the image of T falls at a different retinal location in the two eyes. The dotted line marks the equivalent retinal location in the two eyes. Neurones with receptive fields in both eyes could detect this disparity in two ways. (**A**) Position difference: the right eye receptive field is an exact copy of the left eye receptive field, but in a different retinal location. (**B**) Phase difference: the envelope enclosing the right receptive field profile sits in the same position as for the left receptive field but, within the envelope, the right receptive field has a different structure, responding best to white light on the right-hand side. When tested with a bright bar, both of these mechanisms produce a maximal response to a stimulus with a disparity equal to that of T. (From Cumming, 1997, with permission from Elsevier.)

and is the result of a shear in binocular retinal correspondence. Thus the vertical horopter takes the form of a cylinder which may be reversed (see Box 5-13). The true (empirical/actual) horopter is strictly limited to an area of about 3° from fixation, as determined experimentally. A special form of horopter is one based solely on corresponding points, defined as a circle of projected points in space passing through the fixation point and the nodal point of the eye (theoretical horopter). This is truly applicable only in the horizontal meridian, as the vertical meridians are not exactly parallel. Horizontally placed horopters can form a ‘stack of slices’, producing a longitudinal horopter named so as to reflect the vertical lines of longitude on the globe of the earth. In practical terms the longitudinal horopter is not a circle but has a well-defined shape, approximately representing the field of binocular single vision (BSV) (see Box 5-13).

The field of BSV is an important parameter, not simply from a physiological standpoint but also for socioeconomic reasons. The normal visual field of each eye is approximately elliptical, with a considerable degree of overlap (Fig. 5-26), and the overlapping fields of each eye represent the field of BSV in which full stereopsis is assumed to occur in the context of horopter-related corresponding points on the retina. In the UK, a certain minimum field of BSV is required to qualify for a driver’s licence and is defined as a BSV field of 20° above and below the horizontal meridian and 60° to either side of the vertical meridian (Fig. 5-26). Measurement of visual fields can be performed by many techniques: currently, static automated visual fields are the normal practice, although kinetic and flicker-based fields are also highly informative. In addition, microperimetry is a development of visual field testing, which, when combined with scanning-laser ophthalmoscopy, permits fine retinal mapping and analysis of discrete regions of retinal function or dysfunction.

Additional content available at https://expertconsult.inkling.com/.

Thus it has considerable value for detecting retina-derived as opposed to visual pathway associated field defects (see Ch. 1, p. 92). In clinical practice, quite different visual field tests are performed when assessing either of these central causes of visual field defects.

The horizontal horopter is also defined for specific fixation points and therefore certain degrees of convergence; clearly this will change with the distance from the observer. At about 2 m from the observer the horopter is approximately a straight line, while it is concave to the face within this distance and convex beyond (see Box 5-13).

Measuring stereopsis

The measurement of stereopsis involves at least two parameters: the degree of convergence required to fuse images from two slightly dissimilar objects, and the limits of dissimilarity between two objects at which the two images can be fused (stereoacuity). The former is relatively easy to measure with an instrument incorporating two base-out prisms, known as a stereoscope. In the stereoscope, two slightly dissimilar but symmetrical images are presented to each eye and the angle of convergence at which the sensation of depth is achieved is recorded. In practice, true stereopsis is not measured by this method because light and shade (i.e. monocular cues) provide considerable amounts of image disparity to the same object viewed by each eye in turn. Other tests that remove the monocular cues but use a camouflaged object include the

BOX 5-13 CORRESPONDING POINTS OF FIXATION ON THE RETINA CONSTITUTE THE 'HOROPTER' FIELD OF VISION

(**A**) The horopter circle. The theoretical horopter circle is shown as the dark inner sphere, while the actual horopter for an emmetropic individual is shown by the dashed line. The horopter has been known for centuries. However, in 1818 Vieth-Muller calculated that the horizontal line of binocular fixation was a circle which passed through the centre (nodal point) of the lenses.

(**B**) The binocular single vision horopter field. In practice the 'empirical' horopter is larger than the theoretical horopter since there are many more areas both forward and rear of the circle where single vision can be obtained: these are the inner and outer limits of binocular single vision and correspond to Panum's area.

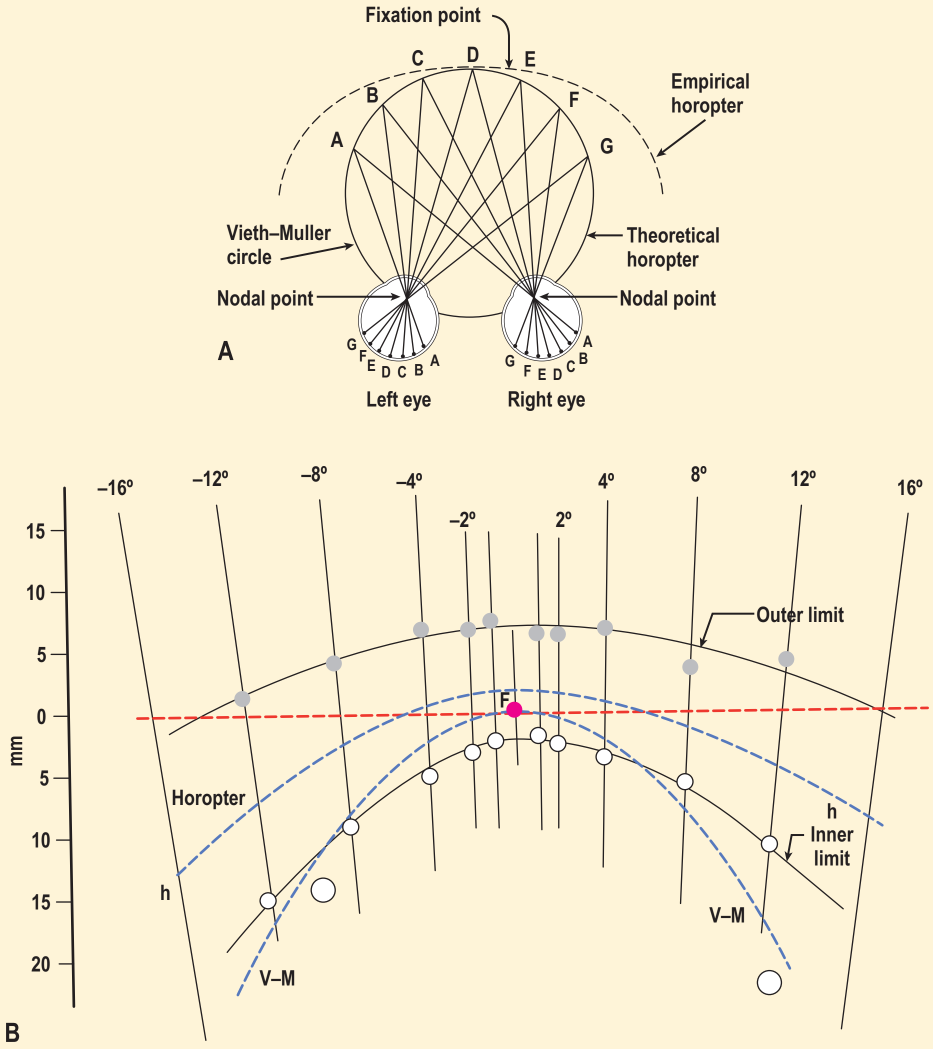
Fixation point
C
D
E
B
F
A
G
Empirical horopter
Vieth–Muller circle
Theoretical horopter
Nodal point
Nodal point
A
G F E D C B A
G F E D C B A
Left eye
Right eye
–16º
–12º
–8º
–4º
–2º
2º
4º
8º
12º
16º
15
10
5
0
5
10
15
20
mm
Outer limit
F
Horopter
h
h
Inner limit
V–M
V–M
B

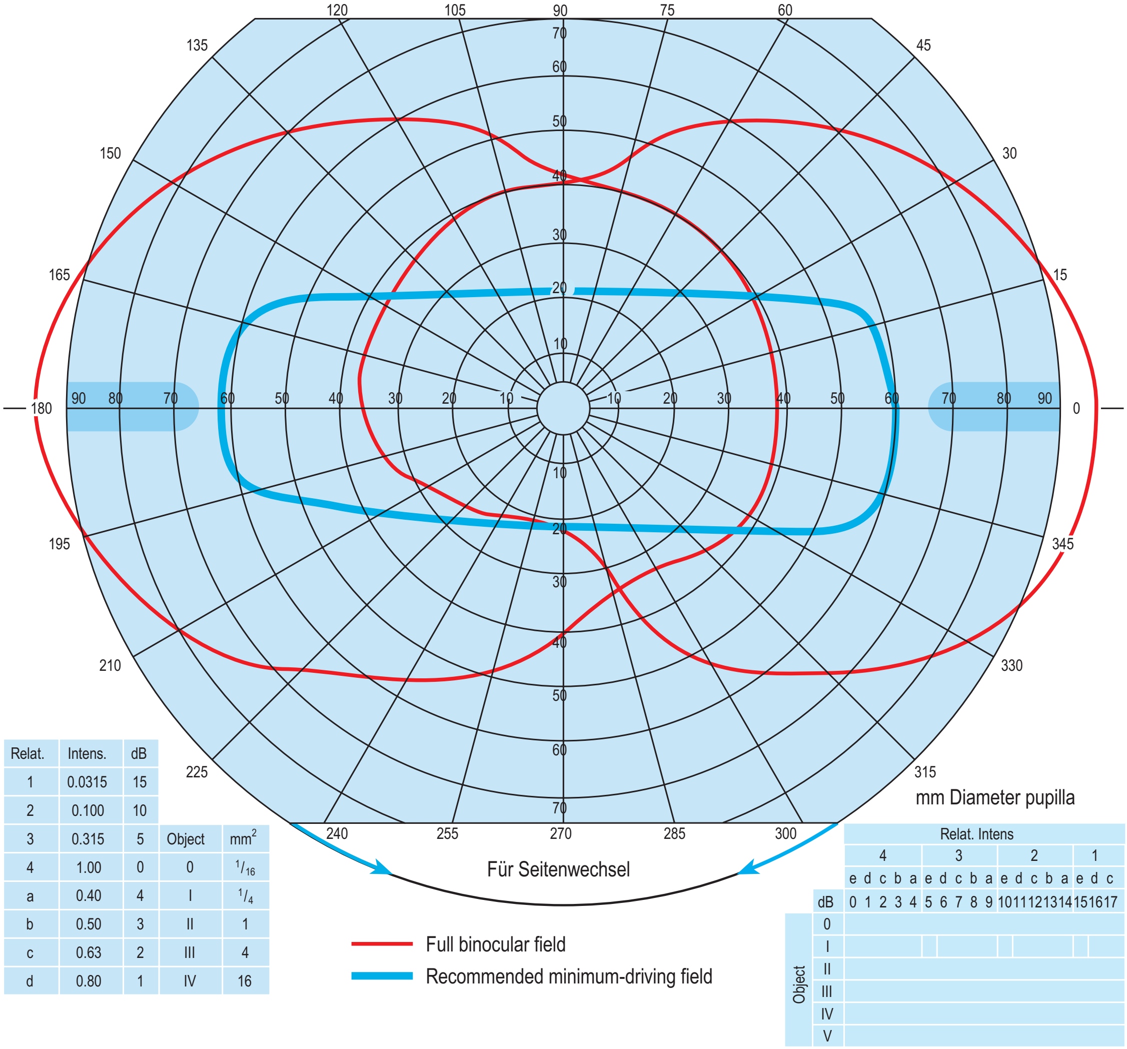

FIGURE 5-26 The field of vision for each eye separately, the field of binocular vision (BSV) and the minimum visual field required for driving licence purposes in the UK are shown. (Courtesy of H. Dawson.)

random-dot stereogram, the random-dot E-test and the Frisby test, in which elements that are non-resolvable monocularly are presented in a random pattern at different disparities and the ability to perceive depth and form in the objects is assessed.

Stereoacuity can be measured as instantaneous parallax, which is the difference in binocular parallax of both objects (Fig. 5-27). The limits of stereoacuity are in the region of 4′ of arc (range 1.6–24′), which is equivalent to an image disparity less than the diameter of a cone. Instantaneous parallax is lost at a distance of about 450 m but varies with the measurement technique. Conversely, within a certain range of distances, stereoacuity is improved the further away the object is from the observer, although theoretically this should not be so. In this case, the improvement is attributed to greater differences in monocular cues such as relative object size, to which the eye is more sensitive, than instantaneous parallax. Stereopsis also improves with duration of the stimulus, which is not the result of small searching or 'image-refreshing' movements of the eye but reflects the minimum time required for

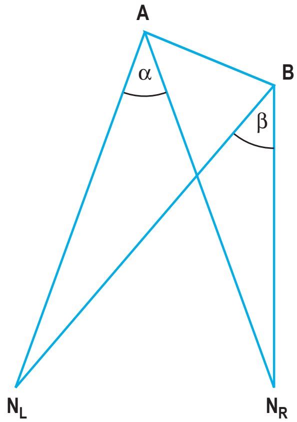

FIGURE 5-27 Images A and B cannot be fused; instead they induce ‘instantaneous parallax’.

neural processing of the stimuli. Indeed, similar but disparate images can be perceived in three dimensions if they are presented to each eye in sequence, but the time interval has to be small (less than 5′ of arc).

Image disparity and stereopsis

True stereopsis is dependent on disparities between the two images received by each eye, and therefore a certain number of points must fall on disparate points on the retina. It is also essential that these disparate points are fused like the corresponding points. The position of corresponding versus disparate points can be assessed by determining the actual differences in the stereoscopic projection of a point in space from the separate projections of the point made by each eye (Nouin’s technique).

There is a limit to the fusional capacity of projected images, which is a circumscribed area known as Panum’s area (see Box 5-13). It has been shown that it is possible to fuse greater disparities in the horizontal meridian than in the vertical, and therefore Panum’s area forms an ellipse. The size of this area varies between individuals, while the threshold disparity that can be fused is greatest at the horopter. The extent of Panum’s area is reduced by small ‘normal’ disjunctive eye movements, which can be compensated for by using stabilized retinal images. Double images can be induced outside Panum’s area and can be used effectively to estimate depth.

Fusion of disparate images to produce stereopsis tends to invalidate the notion of the cyclopean eye in which single vision is produced by fusion of corresponding points. However, the cyclopean eye is of value in providing a baseline on which an estimate of the degree of neural processing involved in the fusion of disparate images can be made. Studies of neural circuitry in stereopsis have thus shown that it is possible to perceive depth without monocular cues, for instance by using random-dot stereograms and other more novel tests.

In these tests, 9 × 10 picture elements composed of dots, some of which correspond while others are symmetrically disparate, are presented in duplicate to each eye. These studies also reveal that discrete contours or edges are not essential for three-dimensional vision, although the contribution from ‘texture analysis’ is not clear (see below). Random-dot stereograms are also not quantitative.

The level at which processing for stereopsis occurs has been questioned on many occasions because it is such an instantaneous response and is difficult to separate from a retinal ‘sensation’. Stereopsis is also sensitive to certain optical effects such as aniseikonia; horizontally it is affected by as little as a 0.25% change in image size, while vertical magnification disparities are simply transferred to the horizontal meridian of the fellow eye. The Pulfrich phenomenon is an optical illusion based on similar processing events (Box 5-14). However, true disparity selective cortical neurones have been detected in V1, although the extent to which these neurones are simply ‘rivalrous’ (see next section) or stereopsis-inducing may depend on the degree of further cortical processing which appears to take place in the middle temporal (MT, V5) area of the visual cortex, an area associated with motion detection (see below). Furthermore, disparity matching appears to be a two-dimensional and not merely a one-dimensional process, involving fusion of images in a vertical as well as horizontal disparity.

RETINAL RIVALRY AND OCULAR DOMINANCE

Retinal rivalry is in essence a term that describes simultaneous perception by each eye individually without fusion of the images. This can be demonstrated as, for instance, when the letters F and L are viewed by each

BOX 5-14 THE PULFRICH PHENOMENON

The illusion of depth can be demonstrated by viewing a swinging, luminous pendulum through both eyes, one of which is covered by a red filter and the other by a green filter. It is thought to be the result of disparate images occurring during the movement of the pendulum stimulating corresponding points at fractionally different times.

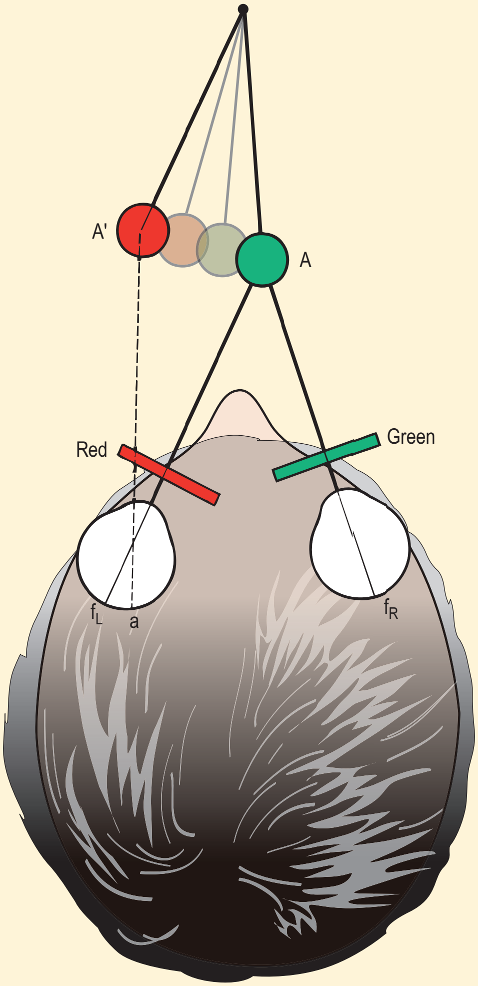

eye separately to produce the letter E. This phenomenon also has a periodicity to it, which is involuntary. It is related to but different from binocular rivalry as illustrated by the Schrödinger staircase or Rubin's vase (see Box 5-11). Both are described as 'bistable phenomena' and are processed at retinal and cortical levels, respectively. Recent fMRI studies confirm that activation of the sensation of monocular rivalry involves recruitment of whole-brain networks as well as visual area V3 (see below), particularly for complex objects such as faces and houses (Fig. 5-28). Retinal rivalry can also be affected by saccades and extraretinal eye movement signals.

Ocular dominance refers to the preferential use of one eye when performing monocular activities. This can be demonstrated electrophysiologically and is not necessarily related to handedness. However, whether there is evidence for true 'cortical' dominance for preferential use of one eye has not been established. In a perfectly equally sighted individual, input from one retina will mirror exactly that from the other retina.

Certain involuntary events take place that bear on retinal rivalry and ocular dominance. For instance, if the same image is presented to each eye at different levels of brightness, then the image in one eye may be suppressed (ocular dominance). Or the binocular image may appear less bright than the same image when viewed monocularly (e.g. with a uniocular cataract).

COLOUR PROCESSING

The perception of colour is a complex cortical event that is dependent on input from several sources. The use of mondrians (coloured patterns produced with variably illuminated narrow-waveband light) has shown that the predominance of a given waveband reflected from a surface does not alone determine its colour, but that its colour also depends on the wavelength composition of the light reflected from its surround. Mondrians have been used to demonstrate the phenomenon of colour constancy but their relative artificiality has been challenged: for instance, demonstration of other phenomena, such as the AMBEGUJAS phenomenon, in which perceived colours can change dramatically depending on the three-dimensional surface of wavelength reflectance. Instead, well-defined real colour scenes have been devised that contain cues for the intrinsic surface colours and the recovery of the light source. These studies show that colour constancy is a real phenomenon but is not as absolute as previously thought and depends on input from local and global contrast. This depends on several factors, including spatial

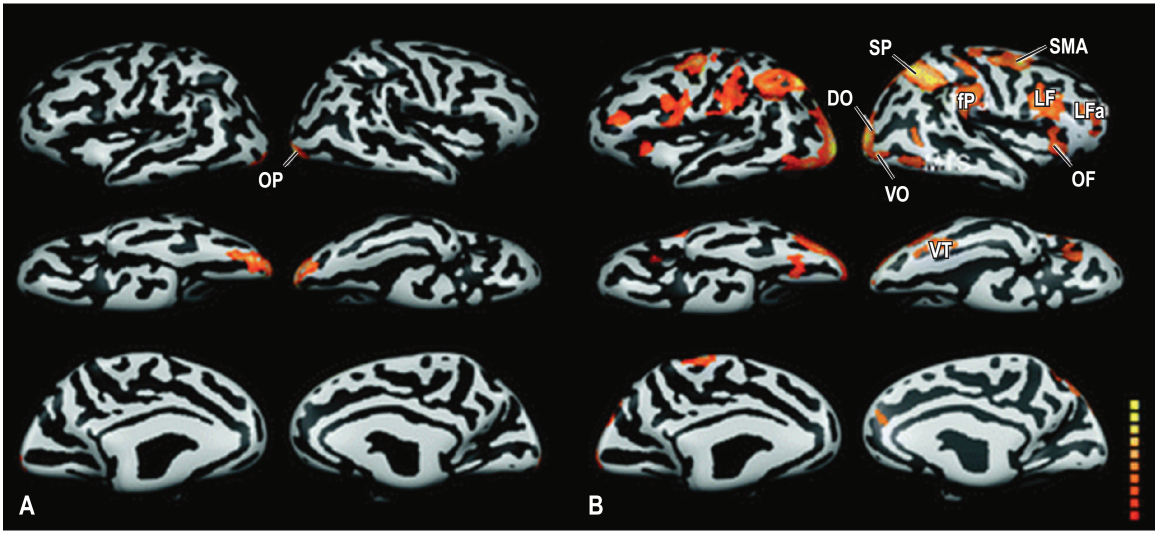

FIGURE 5-28 Monocular rivalry as demonstrated by fMRI using a ‘passive’ stimulation protocol and complex objects such as faces and houses. (**A**) Shows cortical activation for the non-rivalrous control stimulus as a limited region of activation in the occipital pole and includes areas V1 and V2. The lateral, ventral and medial views of the inflated brain are shown (left and right hemisphere). (**B**) Shows cortical activation for passive viewing of monocular rivalry with grating stimuli. Cortical sites included dorso-occipital (DO), ventral-occipital (VO), ventro-temporal (VT), medial temporal sulcus (MTS), superior parietal (SP), temporal-parietal junction (TPJ), supplementary motor area (SMA), lateral prefrontal (LF) and anterior lateral prefrontal (LFa) and orbital frontal (OF). (From Mendola and Buckthought, 2013.)

configuration and scale, and context. Texture, as in the AMBEGUJAS phenomenon, appears to the most important (Fig. 5-29).

It is clear, therefore, that the reflectance of light is central to the perception of colour. Although the amount of light reflected from a surface may vary, the brain constructs an image that fits the reflectance, which is a constant physical attribute of the object. The brain assesses the ‘lightness’ or ‘darkness’ of a surface compared with the surround, for each of the three predominant wavelengths in turn, and this permits it to assign a colour to the surface.

Colour is achieved, therefore, by a comparison of the reflected intensities of lights from one surface with those of surrounding surfaces for lights of different wavebands, followed by a comparison of the comparisons.

SHAPE DETECTION

The detection of form and shape also presents a problem when we consider exactly what we mean by shape. As described above, specialized cells exist that act as edge detectors for vertical and horizontal lines. However, the shapes we perceive are much more complex than can be simply broken down to a series of discrete lines on *x* and *y* axes. For example, most shapes in the natural world are curvilinear and solid, and require significant processing in the cortex. This has led psychophysicists to develop mathematical algorithms and a ‘shape index’ to describe these shapes and thus provide insight into how the brain might compute the information. These effects can be demonstrated using specially oriented stimuli and show that global orientation detection is not simply the result of input to the primary visual cortex (V1). Second-order orientation detection may therefore also exist as ‘collector units’ for first-order V1 stimuli. Such collector units may also be affected by brightness and texture.

CONTRIBUTION OF TEXTURE ANALYSIS AND MOTION DETECTION TO DEPTH PERCEPTION

Depth perception is, of course, not only about locating objects in space but also about perceiving solid shapes. In fact, some of the early work on depth perception involved studies of visual function in aircraft pilots,

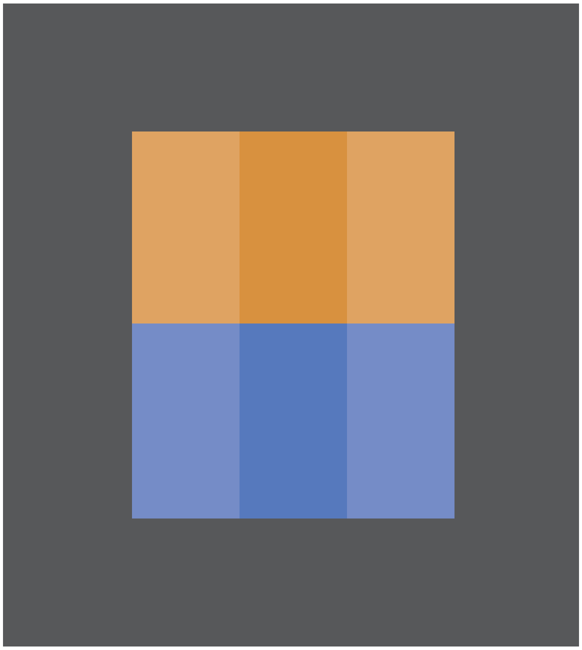

FIGURE 5-29 AMBEGUJAS is a recently described visual perceptual phenomenon in which shapes can be discerned from shading of colours. In this image a 3D illusion is created by the darker-shaded colour strip passing through the centre of the blue and orange rectangles, giving the impression of a box. It is, however, ambiguous as to both shape and colour and frequently returns to a flat surface with a grey central strip. (Bergstrom SS, 2011. The AMBEGUJAS phenomenon and colour constancy. Perception 40: 30–38. http://www.ncbi.nlm.nih.gov/pubmed/15460510, Fig 4, Pion Ltd, London. http://www.pion.co.uk; http://www.perceptionweb.com.)

particularly on take-off and landing. They revealed that motion detection and texture analysis were more relevant for the detection of solid structures, probably the most important aspect of visual perception (Figs 5-30 and 5-31). Many other sources of information combine to produce this effect, including binocular viewing, parallax, illumination and shading, and edge detection. For instance, a random set of dots may assume shape if a group of dots within the set ‘moved’ in relation to the remaining dots – an image would thus ‘pop out’ of the page. Similarly, texture and lustre are attributes of an object that relate to discontinuities in colour and/or luminance coming from the edges of the object, and significantly affect perception of its shape.

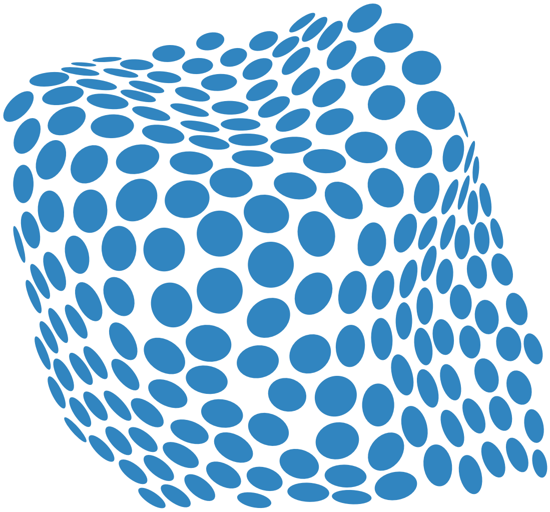

FIGURE 5-30 A pattern of optical texture that is perceptually interpreted as a smoothly curved three-dimensional surface. (From Todd et al., 2005, with permission from Elsevier.)

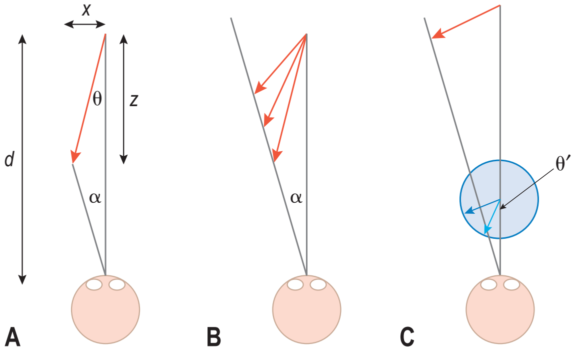

FIGURE 5-31 Schematic of stimulus geometry. (**A**) Observer and typical target trajectory (solid arrow) at an angle θ to straight ahead are illustrated. Visual direction of the moving target at the end of its motion is given by angle α. (**B**) Many different trajectories correspond to a single visual direction. (**C**) Location of the pointer (grey circle). If observers used θ to set the pointer, they would respond as shown by the black arrow. The grey arrow shows a typical response if observers used α. (From Harris and Drga, 2005.)

Most recently it has also been shown that the detection of a shape also depends on previous experience/memory and whether the observer is ‘expecting’ to see the shape. It is clear therefore that this complex response is built up from multiple inputs and that the search for a shape-detecting centre may prove elusive.

Division of labour in the visual system: parcellation and the human connectome

Topographic mapping of brain function has advanced considerably in the last 5 years and this applies particularly to vision, specific visual functions and the connections between visual areas and other areas which influence visual perception. Zeki's early conceptual framework for what he describes as the division of labour in the visual cortex has now been elaborated to reveal the numerous interconnections between different regions of the brain and how both bottom-up and top-down processing occur almost continuously (Fig. 5-32). In this context, it is now well recognized that the dorsal (occipito-parietal) and ventral (occipito-ventral) parcel out bottom-up and top-down processing, respectively. Studies like this are contributing to a major project under construction at the present time aimed at mapping the human neuronal connections in what is termed the 'human connectome' (http://www.humanconnectomeproject.org/).

According to current ideas concerning construction of the visual image, anatomically discrete areas of cortex subserving different functions are reconstituted through their extensive connections to produce the perception of a final integrated image, for instance during complex psychophysical process such as those involved in ocular dominance (see Fig. 5-28). Thus an image should not be regarded as being 'impressed' on the retina like the film in a camera, which is then codified to make it 'understood' by the cortex; rather we should appreciate that the processes of sensing (seeing) and cognition (understanding) are not separate but totally integrated. It is likely that, as additional sensory information is added to the image, a higher level of cognition and thus perception can be achieved.

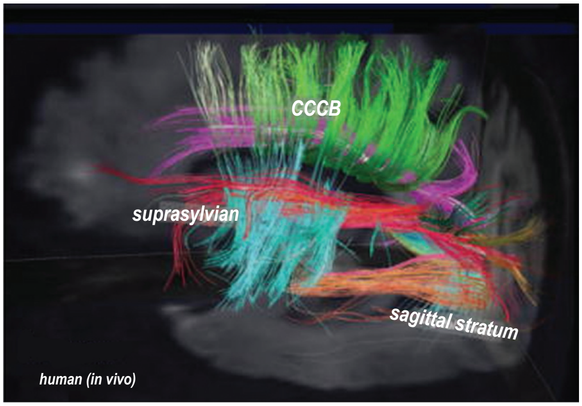

FIGURE 5-32 A cerebral grid structure of the human brain showing the plan of the fibre pathways; image taken using diffusion magnetic resonance imaging. (From Wedeen et al., 2012.)

IMAGING STUDIES

Much of the early information came from work in animals but advances in imaging the human brain have in many ways now taken precedence. 'Parcellation' in the human visual cortex has been shown by positron emission tomography (PET), magnetoencephalography and functional magnetic resonance imaging (fMRI). Several developments in MRI from initial diffusion tensor imaging (DTI) (imaging of molecules as they move) to newer techniques such as diffuse functional MRI, diffusion-weighted imaging and diffusion spectrum imaging have provided the tools for the human connectome project. Importantly, advances in spatial resolution of fMRI images such that active regions in specific gyri can be determined without interference from image signal of irrelevant neighbouring sites has allowed localization of ventral occipito-temporal pathways involved in, for instance, seeing words (Fig. 5-33). Similar fMRI studies of V4 for coloured stimuli and V5 for moving targets have been done.

In humans, therefore, it can be said that there is a 'colour centre' outside the visual cortex; i.e. that there is functional specialization as in the monkey. Another intriguing observation is that the effect of coloured stimuli is lateralized in humans and is not necessarily related to handedness or ocular dominance. Similarly, three different 'types' of motion detection have been located to discrete areas of the visual cortex.

THE MAGNOCELLULAR AND PARVOCELLULAR PATHWAYS SUBSERVE DIFFERENT FUNCTIONS

The striate cortex (area V1) contains the entire map of the retina in a highly ordered and predictable distribution. It receives this information via the LGN, where the neuronal organization is also highly ordered (see Ch. 1, p. 93–95). At first sight it would seem, therefore, that retinal images should be truly represented in the visual cortex on the basis of a point-to-point topographical representation. In a very limited sense

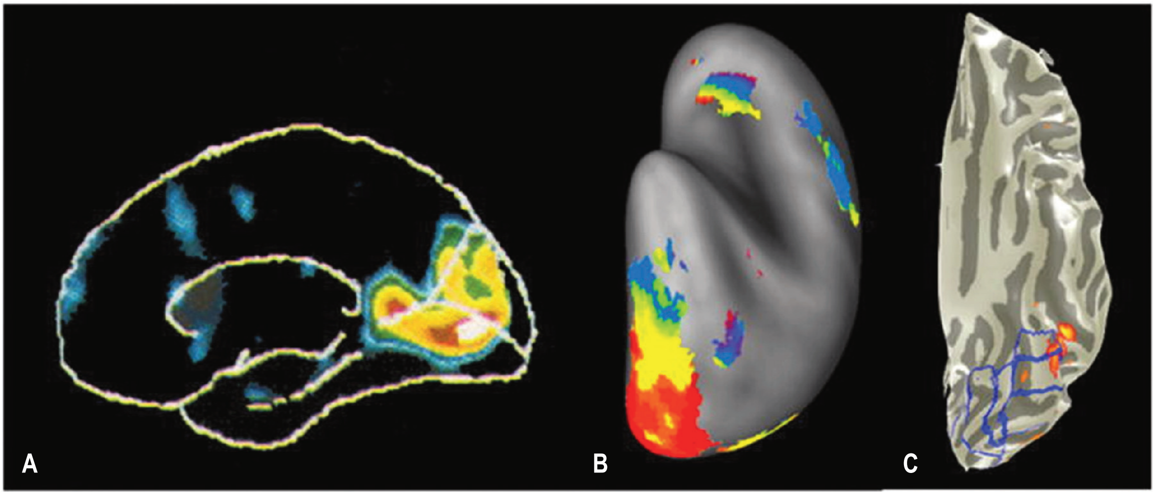

FIGURE 5-33 Neuroimaging technology has greatly improved to allow high levels of spatial resolution from positron emission tomography (PET) (**A**), to fMRI on 3D surfaces (**B**), to fMRI data (**C**) (allowing words to be differentiated from phase-scrambled words) coordinated with visual field maps (shown in the blue outlines) in individual subjects (right, ventral view) using retinotopic mapping procedures. The word-related activation is near the foveal representation of ventral occipital maps (VO-1 and VO-2); V2 and V3 are also outlined. (From Wandell, 2011.)

this is true and may even apply to functional differentiations associated with certain neuronal cell types. Thus, the parvocellular (slow) fibres carry information concerning foveal and parafoveal activity such as spatial discrimination and colour, while the magnocellular (fast) fibres act as transmitters of light detection. This explains in part the high sensitivity that we have for light and motion detection, which are served via the peripheral retina, while contrast and colour detection are slower processes.

However, psychophysical phenomena such as colour and spatial constancy indicate that simple representation of images on the visual cortex in a retinotopic fashion is insufficient to explain the resultant perception. Even the briefest consideration of wavelength discrimination, which is an intrinsically colourless event, despite the fact that there are three discrete receptors, would reveal this truth; thus a pillar box appears red in most conditions of illumination even though the actual wavelength of the reflected light from the pillar box will vary greatly depending on the light source (but see discussion of colour constancy above). The perception of colour, and indeed of any visual stimulus, is the result of input from many other cortical sources in addition to the primary visual cortex. A further unexplained problem in studying cortical and indeed lower levels of activity is the constant high level of neuronal 'noise'. It appears that even in the absence of specific stimulation, there is a significant level of endogenous neuronal activity, which is now presumed to modify output (i.e. perception in the case of vision). This is considered to be one form of top-down modulation of output.

THE STRIATE CORTEX AND THE PRESTRIATE CORTEX SHUFFLE INFORMATION BETWEEN THEM IN THE BUILD-UP TO A PERCEIVED IMAGE

The striate cortex (area V1) is connected to the prestriate cortex (areas V3–V8) directly and also via area V2 (Fig. 5-34). Each of these areas has one or more specific functions. For example, all cells in area V5 respond to motion in the visual system, and are directionally selective (i.e. each cell responds to motion in only one direction); none of these cells, however, is specific for colour. This function is subserved by cells in area V4, in which some of the cells act as wavelength discriminators, but some of these cells also respond to orientation of lines and are involved in shape (form) detection. Other studies have, however, shown that contour information can also be derived from motion detection and that this activity takes place in the primary visual cortex. Cells in V3 and V3A

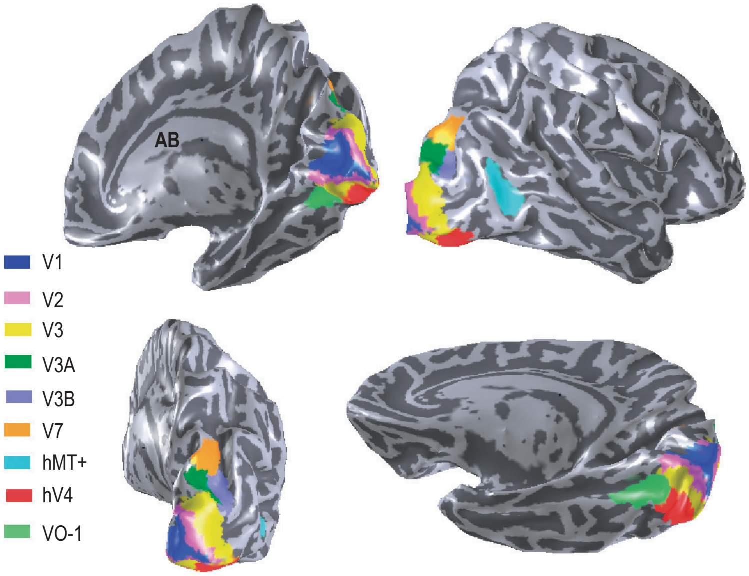

FIGURE 5-34 The locations of nine hemifield maps in the human visual cortex. The maps are shown for one typical subject (AB). (From Wandell et al., 2005, with permission from the Royal Society of London.)

are also selective for form but are indifferent to changes in wavelength.

It is clear, therefore, that colour, orientation, motion, stereoacuity, texture, etc., are all processed separately in areas V3–V8. As areas V3–V8 receive their information from V1, V1 (and V2) must also be functionally specialized. It has been suggested that there may be a population of cells that respond to more than one stimulus, such as texture and motion, when these are tested separately. However, this is not the common response.

V2 is likewise organized into areas with thin stripes (for colour detection) and thick stripes (for motion detection) separated by interstripes. Form-selective detectors are present in both the thick and the thin stripes. This form of organization of the visual, and indeed the entire, grey matter into discrete columns of cells responding to specific stimuli has long been known (>50 years) but its functional significance is not clear.

PARCELLING OUT THE PROCESSING IN V1

The concept of functional/anatomical segregation of visual stimuli into components such as colour, motion and orientation detection, depth perception, and other features is now well established (but see below). The cortical sites of other visual tasks, such as texture analysis and shape recognition, are not so easily located. Still others, such as face recognition, involve regions outside the visual and prestriate cortex, including sites that store memory.

Despite our lack of knowledge, it is still remarkable that such a level of segregation occurs from the retinal ganglion cell input, through the LGN to V1, V2 and V3–V8 in the cortex. Segregation may have developed as a result of the different requirements for generating form, colour and motion (e.g. colour compares input from one part of the visual field to another) but topography for colour may be less important. In contrast, precise topographic localization is important for form analysis and motion detection; in the latter, however, this is assessed only transiently.

Our perception of the external environment may therefore depend on a system of circuits rather like combined and serial parallel processing in computers where there is 'multistage integration', as Zeki describes it, with feedforward and feedback control (now described as bottom-up and top-down processing) (Zeki, 1990). In line with this concept, perception and comprehension of the visual world occur simultaneously and continuous processing of information, both past and present, is ongoing.

IS THE VISUAL CORTEX ORGANIZED FOR HIERARCHICAL NEURAL PROCESSING OR FOR FUNCTIONAL SPECIALIZATION?

The above outline of the organization of visual information at the cortical level may be oversimplistic. In reality, organization of information reception and integration has been considered in two ways, both of which are probably contributory: (1) hierarchical processing of information through the different regions (e.g. sequentially from retinal to LGN to V1, then to V2, and then simultaneously or differentially to V3, V4, V5, etc.); (2) functional specialization in the form of precise topographical localization of aspects of vision to discrete areas of the visual cortex. One of the important recent observations from fMRI work is that the concept of precise retinotopic mapping being restricted to V1 (i.e. representation of the fovea and peripheral retinal regions to precise sites on the striate cortex) may not be accurate. Areas previously considered non-retinotopic such as V4 for colour and V5 for motion also have retinotopic representation although less exact than in V1. Not only is there region-specific cortical representation for instance of colour, but eccentricity maps (distance from fovea) and polar angular maps (angle from the horizontal meridian) cross each other in their cortical representations, allowing a form of mapping of 'visual space' on the cortex (Figs 5-35 and 5-36).

It has now become clear that perception of colour, motion, form, texture and stereopsis all have varying levels of the following attributes: (1) functional specialization in discrete cortical areas; (2) retinotopic representation within each of those areas; and (3) processing in 'streams', e.g. in a colour stream where neural processing for wavelength discrimination takes place at several levels from the retina, to the LGN to V1 and onwards to the specific cortical region in V4. Similar organization underpins motion detection but less is known about stereopsis or texture appreciation. Stereopsis and depth perception are particularly interesting since this requires 'binocular neurones' which will merge signals from right and left eyes: while V1 is clearly important, it is now realized that extrastriate input, e.g. from V2 and V4 as well as other areas, is important. Broadly, dorsal visual pathways control cross-correlation between signals while

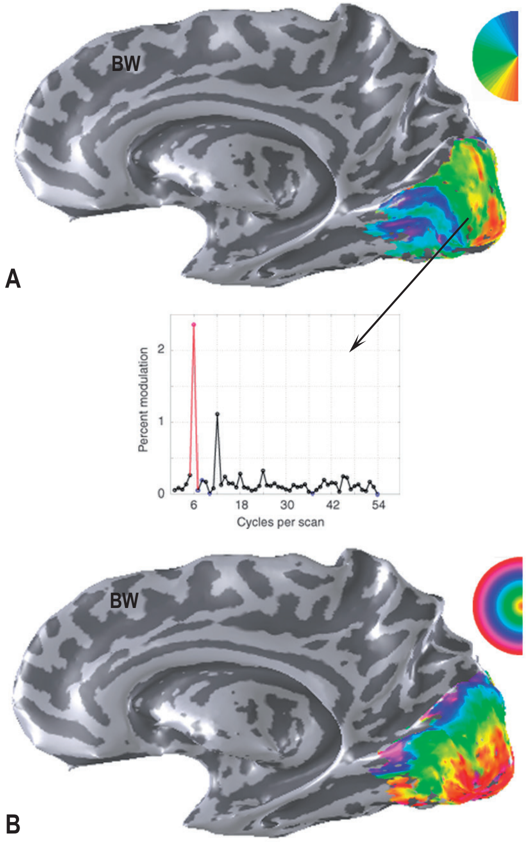

FIGURE 5-35 Angular and eccentricity maps near the calcarine cortex. Maps were measured using (**A**) rotating wedges and (**B**) expanding rings comprising contrast-reversing dartboard patterns. The stimuli extended over the central 20° of the visual field and completed six cycles during each experimental scan. The colour overlay indicates the visual field angle (**A**) or eccentricity (**B**) that produces the most powerful response at each cortical location. For clarity, only responses near the calcarine cortex are shown. The graph plots the response amplitude as a function of temporal frequency as measured in a 3 mm radius disc located in the calcarine (see arrow). The response is significantly greater at the stimulus repetition frequency (six cycles per scan, shown in red) than other temporal frequencies. The secondary peaks at integer multiples of the stimulus frequency are expected and are also significant. The graph is included in the image to provide the reader with an assessment of the reliability of the responses. The stimulus-driven responses shown here are substantially above the statistical threshold ($P <$ 0.001, uncorrected). (From Wandell et al., 2005, with permission from the Royal Society of London.)

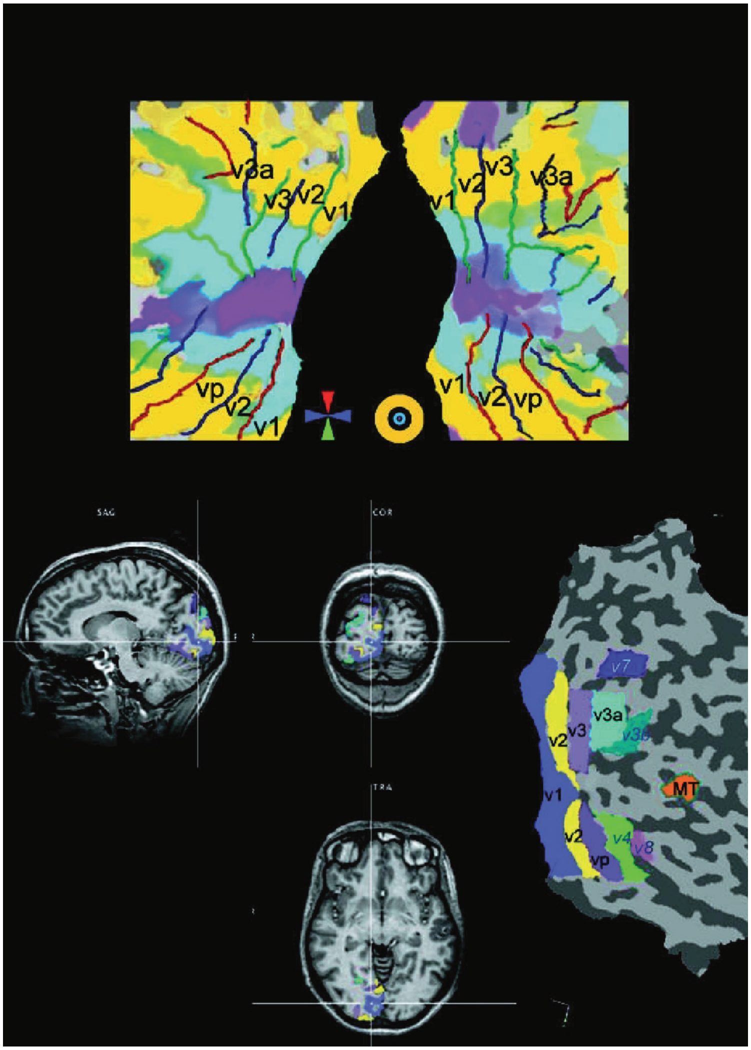

FIGURE 5-36 Early and mid-level visual areas. *Top*: Superposition of eccentricity and polar angle maps. Yellow, blue and pink bands indicate eccentricity maps; lines indicate centres of upper, lower and horizontal representations (see icons). Note that meridian lines cross all eccentricities orthogonally. *Bottom*: Visual areas on a flattened representation and on the brain volume. Visual area names under consensus are denoted in black, and areas under debate are marked in blue italics. (From Grill-Spector and Malach, 2004, with permission from the publisher of Annual Reviews.)

ventral pathways address the problem of making multiple signal matches.

There is further sophistication in perceptual pathways. Regions anterodorsal and anteroventral to the striate cortex are not only specialized to detect aspects of vision such as form or colour but also to detect specific objects such as faces, tools, words and even places (Fig. 5-37). Despite this, there is still a strong hierarchical organization for information processing. For instance the receptive fields for the same stimulus are smallest in V1 and increase progressively through V2, to V3A and V4. To add further complexity to the final percept, there is evidence that top-down processing occurs, for instance activity induced in visual areas V1 and V2 by increased ‘expectation’ or attention to a region even in the absence of a specific visual object. Even input from emotional stimuli or stereotypical events can modify the visual areas as seen on fMRI.

As fMRI advances with ever better technology, including diffusion tensor MRI, detailed information about the organization of areas such as the intraparietal suture (IPS) have been revealed.

Additional content available at https://expertconsult.inkling.com/.

Functional connectivity analyses also showed increased interaction between the IPS and prefrontal

Tasks which involve attention as well as the maintenance and re-configuration of information in working memory (WM) are centred on the IPS (eFig. 5-7). WM tasks are closely integrated with what is termed the 'contralateral visual space' (i.e. fMRI mappable areas on the contralateral visual cortex) while both sides of the IPS are used for processing this information. In a recent study, Bray and colleagues performed visuotopic mapping in a group of volunteers and showed that a briefly flashed target preferentially engaged the contralateral IPS but when the target was 'mentally' rotated around a circle (i.e. manipulating spatial information) both IPS were activated, and activation was most marked in region IPS1 in most but not all individuals (Bray et al., 2013).

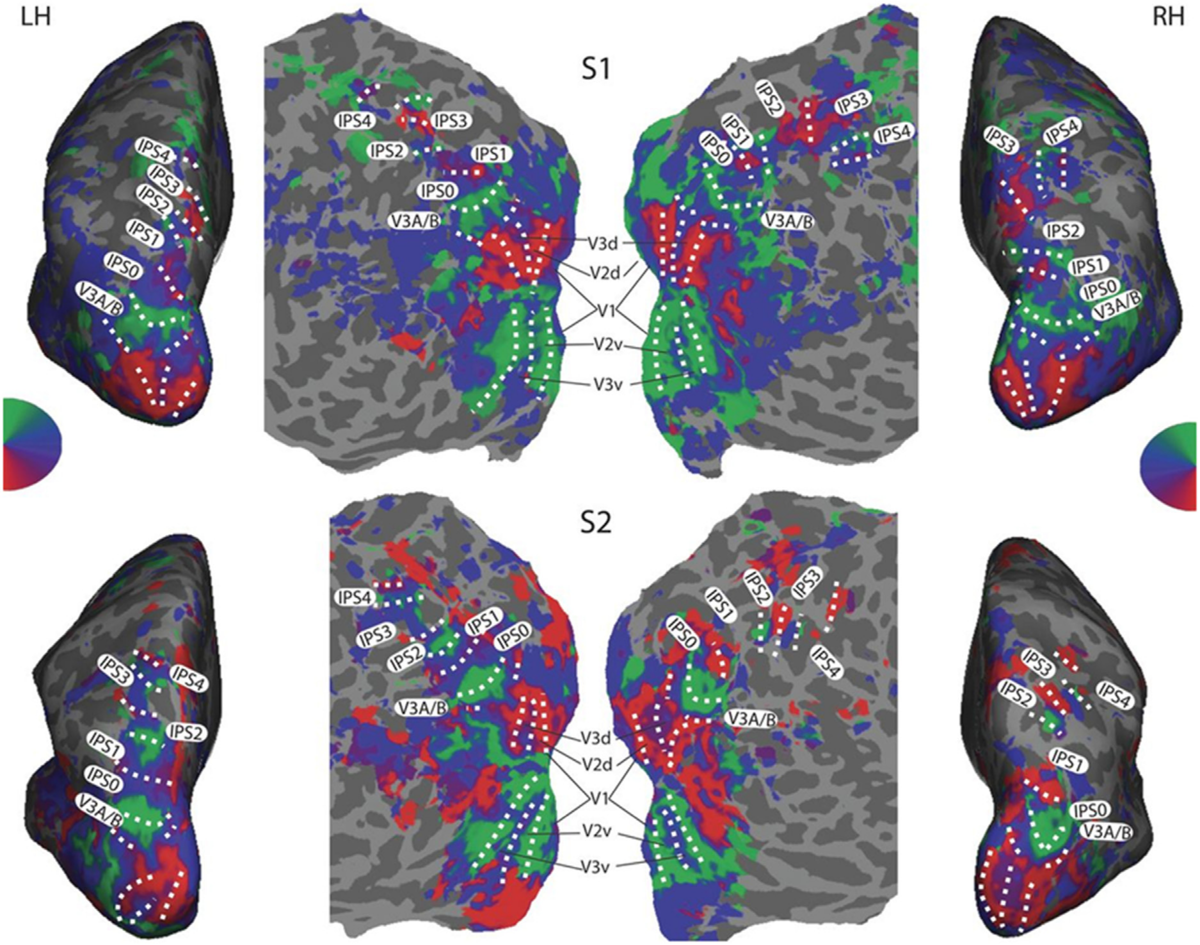

eFIGURE 5-7 The intraparietal sulcus (IPS) shown in its various regions (IPS0-4) in various views of inflated images of the cortex. Visual cortex areas 1–4 are also shown. Dashed white lines represent borders between adjacent areas. (From Bray et al., 2013.)

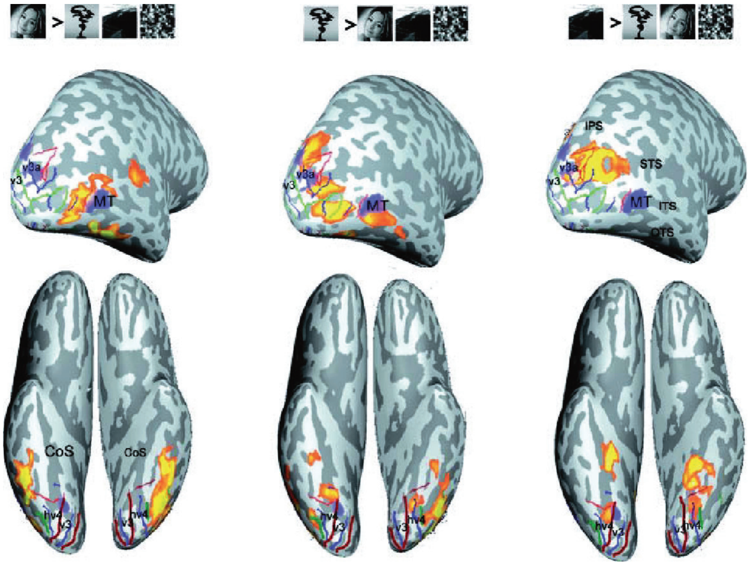

FIGURE 5-37 Face-, object-, and place-selective regions in the human brain displayed on an inflated surface representation of the same subject as in Fig. 5-36. Icons indicate the comparisons done in the statistical tests. *Left*: areas responding more strongly to faces than objects, places, or textures. *Centre*: areas responding more strongly to objects than faces, places, or textures. *Right*: areas responding more strongly to places (scenes) than faces, objects, or textures. Yellow and orange indicate statistical significance: $P < 10^{-12} < P < 10^{-6}$. Coloured lines indicate borders of retinotopic visual areas. Blue indicates area hMT1, defined as a region in the posterior bank of the inferotemporal sulcus that responds more strongly to moving versus stationary low-contrast gratings (with $P < 10^{-6}$). (From Grill-Spector and Malach, 2004, with permission from the publisher of Annual Reviews.)

regions during manipulation, as well as interhemispheric interactions. Two control tasks demonstrated that covert attention shifts and non-spatial manipulation (arithmetic) engaged patterns of IPS activation and connectivity that were distinct from WM manipulation. These findings add to our understanding of the role of IPS in spatial WM maintenance and manipulation. This type of study reveals the visuotopic arrangement of the fibres as they pass through the classically described ‘optic radiation’ (see Ch. 1, p. 95) and will be of considerable value in identifying very specific neurological defects as well as informing us on basic neurophysiology.

Area V4 remains a bit of a puzzle. While many functions have been attributed to it, precisely how this region modulates other visual areas such as colour or feature processing are unclear. It has recently been proposed that the role of V4 is to facilitate neuronal traffic into sets of domain-based networks (such as colour and motion) in two directions, i.e. it probably acts as a hub for the convergence of top-down and bottom-up information (Roe at al., 2012).

Physiology of ocular movement

Many of the aspects of the visual response described above would not be possible without the coordinated movement of the eyes; indeed, eye motion is a fundamental feature of ocular and visual physiology since eyes in the alert state are never at rest. Eye movements are paired even when they move in different directions, as in convergence responses. Neural control of paired eye movements occurs at several levels, as for any neuromuscular event, i.e. at a reflex/subcortical level and via cortical control. The anatomy of the ocular muscles and the innervations of the ocular muscles via the cranial nerves and brainstem nucleus have been reviewed in Chapter 1 (see pp. 68–77).

TYPES OF MOVEMENT

Uniocular eye movements

Each eye can be moved in the direction of action of the ocular muscles (Box 5-15), which are usually described around a centre of rotation of the globe placed about 14 mm behind the cornea. Rotation is either in a vertical (z axis) or a horizontal (x axis) plane, otherwise known as Listing's plane. Torsional motion of the eye occurs around a median or vertical plane through the midline of the skull; this movement can also be described in most circumstances in reference to the retinal horizon (the x,y plane). Rolling movements of the eye occur along an anteroposterior axis, while intermediate movements between any of these axes are possible.

BOX 5-15 EVALUATION OF EXTRAOCULAR MUSCLE FUNCTION

The direction of action of the extraocular muscles is complex and should be considered in three dimensions. The diagram indicates the action of the muscles when they are to be tested clinically for their function, as explained below.

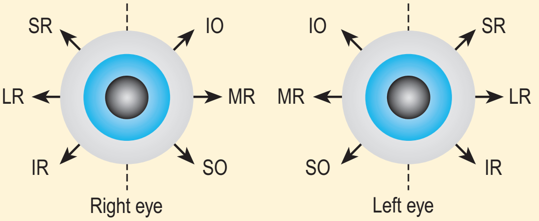

For instance, the action of the SO muscle is to move the eye down (depression) and out (divergence). However, depression of the eye can also be induced by the IR. Therefore, in order to test the depressor (downward) function of the SO, the simplest way is to test this function when the eye is in a position where the other depressor (the IR) cannot act, i.e. when the eye is adducted. Adducting the eye in effect 'shortens' the IR (the other depressor of the eye, thereby compromising its role as a depressor – muscles are less efficient if their muscle belly is shortened). The SO is thus the only depressor in the adducted position and vice versa for IR. IR, inferior rectus; SO, superior oblique; IO, inferior oblique; LR, lateral rectus; MR, medial rectus; SR, superior rectus.

Binocular eye movements

The extent of movement of one eye is equal and symmetric to the other (Hering's rule); in conjugate movements the eyes move in parallel while in dysjunctive movements (convergence and divergence) they move in opposite directions. In the fusion-free or physiological position of rest (not the primary position of gaze, as this requires fixation on a target) the eyes are slightly divergent.

Conjugate movements require reciprocal innervation of the muscles, which can therefore be described as conjugate pairs of muscles for each direction of gaze (see Box 5-15). This is limited in that the excursion of each muscle is usually greater than that of the pair. The effect is to produce a field of binocular single vision, a parameter of great practical significance with respect to standards of normal visual function for the purposes of vehicle license regulations (see Fig. 5-26).

Convergence movements require the combined action of both medial recti; the extent of movement is limited by the near point (5–10 cm from the eyes; this is not affected by age unlike the near point of accommodation) and the far point of convergence (determined in the position of rest as the projected intersection). The converging power of the eye is measured by the metre angle, which depends on the interpupillary distance and can be assessed using graded prisms. From this, the amplitude of convergence can be calculated, which is the difference between the converging power of the eye for the near and far points of convergence. The fusional drive can add a component to the converging power of the eye – the fusion supplement. This has practical significance, for instance, when estimating the effects of accommodating intraocular lenses on depth of field (see eFig. 5-4).

Conjugate movements of the eye may be in the form of short sharp movements (saccades) or continuous tracking movements (smooth pursuit). Even when under apparent steady fixation, there are small conjugate movements (microsaccades). Voluntary gaze or 'search' movements (i.e. directed towards non-defined targets) are under higher cortical control (see below). Experimental studies have shown that visual input is required for saccadic movements and is linked to image latency. Tracking movements, however, require visual input plus object speed to be no greater than 30–40° per second and to match that of the eye movement.

Saccades

- Rapid voluntary relocation of fixation
- Under supranuclear contralateral control
- Latency of 100 ms
- Velocity of 800–1000°/s.

Pursuit

- Slower tracking movements
- Under supranuclear ipsilateral cortical control
- Latency of 150 ms
- Velocity of 30–50°/s.

During saccades, there is selective suppression of motion detection over other stimuli, suggesting that saccades suppress only the magnocellular pathway. It is also important to consider the concept of optic flow, in which an object moving in relation to a static observer generates a pattern of relative motion in the retinal image. The control of eye movements under these conditions may be difficult to analyse, particularly if the observer is tracking a slowly moving object against a faster moving background.

The situation becomes even more complex if the observer is also moving. Here the visuomotor and vestibulomotor interact with each other to stabilize the image (Box 5-16).

Optic flow is based on what are described as ‘short latency’ eye movements generated by the sensation of

BOX 5-16 THE IMAGE WE SEE IS DESTABILIZED WHEN WE ARE MOVING BECAUSE OF OPTIC FLOW

Optic flow is a form of visual streaming which occurs as we are moving continuously in one direction. It occurs because the image of the same object(s) are constantly changing with regards to which area of the retina they stimulate. An object of interest is fixed by our gaze and is usually tracked as we go forward but the eye movement used for this purpose interferes with the flow of information (optic flow) generated by objects in the background which inform us where we are going and allow us to navigate properly. Accordingly, the object of interest loses ‘focus’ and becomes blurred (see figure). Both the visuomotor and vestibulomotor systems are at play here and attempt to compensate for the instabilities of retinal images, which typically vary as a function of retinal location and differ for each eye.

(**A**) Blurring of a natural scene image on the retina during simulated leftward self-motion. Near objects blur more than far objects. Without a compensatory eye movement, only objects at optical infinity (for example, mountains on the horizon) remain stable on the retina. (**B**) Idealized patterns of corresponding planar optic flow (in panel (**A**)). Small black arrows indicate the relative displacement of three groups of targets, located at different isovergence distances, during the first 100 ms of constant velocity (50 cm s^{-1}) translation to the left (large black arrow). Shaded areas indicate the angular displacements of the targets at different distances and horizontal eccentricities, as seen from the right eye. (**C**) White circles indicate two horizontal spatial locations of image blurring on the retina; cross indicates forward movement direction. (**D**) Idealized patterns of corresponding planar optic flow (in panel (**C**)). Shades from red to blue illustrate transition from the fovea to the retinal periphery. Note that these components represent the flow velocity perpendicular to the optic axis.

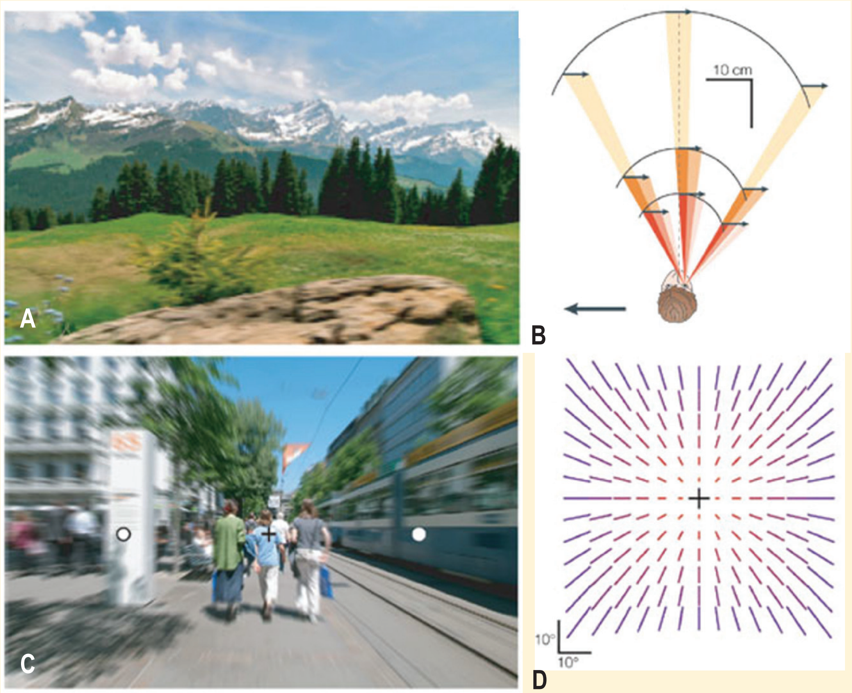

a changing image or scene, experienced by the observer while moving so that correct navigational or directional movement can be made, e.g. 'heading'.

CONTROL OF EYE MOVEMENT

The eye muscles in the primary position of gaze are in a state of tonic activity. Each muscle, however, is activated when the eye moves in its field of action and is inhibited in the opposite direction. The final pathway for neuronal control of eye movement occurs via the cranial nerves (see Ch. 1, p. 69), which are the motor neurone equivalent of the spinal nerves subserving reflex responses. As for any muscle, however, the ocular muscles are under both reflex and 'higher centre' control, with the frontal cortex regulating voluntary activity and the occipital cortex and superior colliculus serving as coordinating centres. In addition, there are numerous interneurones and connections with other pathways at the cortical level, e.g. via the paramedian pontine reticular formation (PPRF), and at the reflex level, e.g. the vestibulo-ocular reflex and the cervico-ocular reflex (see above section in relation to optic flow). The generation of horizontal and vertical saccades (gaze) and the fine-tuning of eye movements involve the integrated supranuclear network within the midbrain (PPRF and rostral interstitial nucleus of medial longitudinal fasciculus (riMLF)) and brainstem (vestibulo-ocular and cervico-ocular reflexes), which will be discussed in more detail below.

The fixation reflex

The ability to fixate a bright light is a basic reflex that is evident within a few days of birth, but the binocular reflex involving conjugate eye movements and a sustained response takes several months to be fully developed. Foveal fixation is the endpoint of the searching movement of the muscles and may be considered the point of peak activity in the nerve/muscle response. The nerve response can therefore be said to be 'tuned' to foveal fixation. In addition, the very small fine eye movements (microsaccades) that occur with sustained foveal fixation are the result of reflex attempts by the oculomotor centre to achieve the best perceived image, as this falls off rapidly unless a new set of cones is stimulated (Box 5-17).

The fixation reflex can be demonstrated easily by testing for optokinetic nystagmus, where either the stationary subject views a moving scene or a moving subject views a stationary scene. The nystagmus has a slow phase when the eyes follow the target and a fast flick when they readjust to the new target position. The optokinetic nystagmus response in humans requires an intact cortex, although there may be a subcortical pathway via the superior colliculus especially for the 'involuntary' searching component of the response (see Box 5-17). Lesions of the cerebral cortex, for example in the temporal lobe, are associated with defects in the optokinetic nystagmus response. The nystagmus is preserved in parietal lobe lesions, and this test is therefore clinically useful in localizing lesions.

Microsaccades are probably as important in achieving a strong fixation reflex as saccades are in searching or exploratory eye movements.

Oculovestibular reflexes are eye movement responses to positional changes in the relationship between the head and the trunk

The vestibular apparatus has structures that convey static head/trunk positional reflex information (i.e. when the subject is not in motion) – the utricle and saccule – and kinetic positional information under conditions of head/trunk acceleration and deceleration – the semicircular canals. In the utricle and saccule, stimulation of the receptor may occur simply on changing position of the head with respect to gravity, but the ampullae in the semicircular canals are stimulated via inertial forces in the endolymph surrounding the hair cells (viscous drag). The semicircular canals are arranged so that they act in synergistic pairs on each side of the head in the x, y and z axes. Vertical and torsional movements involve all four vertical canals. The vestibule ocular apparatus is thus of great importance to navigational movements, heading and optic flow (see above section). Three-dimensional analysis of ocular movements allows the possible location of defects to, for instance, a single semicircular canal. The techniques are based on mathematical models containing information on rotation vectors, reference frames, coordinate systems and Listing's law, and use magnetic search coils in preference to video-based systems.

BOX 5-17 SACCADES ARE GENERATED IN DIFFERENT AREAS OF THE BRAIN

Saccades describe a fundamental feature of our eyes, i.e. they are never at rest (even when we are sleeping). Even when we fix our gaze on an object, our eyes still continually make small movements termed microsaccades. There is also a tendency for the eyes to ‘drift’, especially if we are not fixing strongly on an object, and ocular tremor adds a third intrinsic movement to the eyes. These movements are essential to continued perception of an image since without repeated retinal stimulation the image fades. Saccadic eye movements are generated in several areas of the brain, including the lateral intraparietal area, the frontal eye fields, the supplementary eye fields and the dorsolateral prefrontal cortex (see figure). The neural connections are integrated with input from the basal ganglia (blue in the figure) while the substantia nigra generate inhibitory impulses which prevent unwanted eye movements as well as control the onset of saccades. The brainstem (shown in green in (**A**) and in higher magnification in (**B**)) contains a complex feedback and feedforward circuitry involving ‘excitatory burst’ and ‘omnipause’ neurones in the superior colliculus which control the overall regulation of the microsaccades through inhibitory and activating impulses.

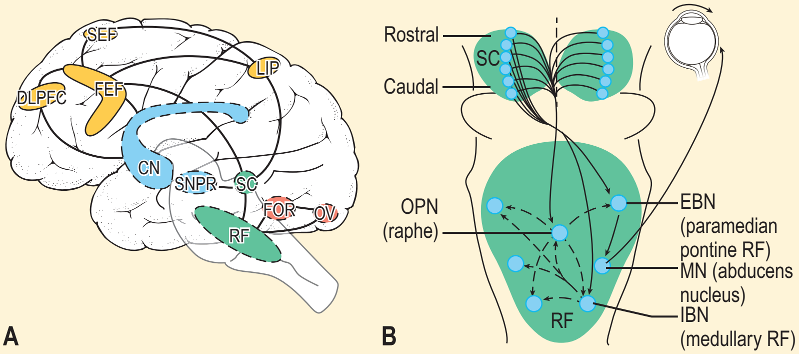

Listing’s law states that, when the head is fixed, the primary position of the eye is such that there is a restricted degree of orientation that can be reached by a single rotation about an axis in Listing’s plane (see above) (nine positions of gaze; Box 5-15 and Figs 5-38 and 5-39). Listing’s law applies during fixation, saccades, smooth pursuit and vergence movements but not during sleep or during vestibulo-ocular reflexes.

The oculovestibular reflex can be demonstrated by the ability of a rotating observer to maintain fixation on a stationary target by reflex movement of the eyes at the same angular rotation (up to 300°/s) as the observer in the opposite direction. In this way there is stabilization of the retinal image. The reflex can also occur in the dark but is less accurate in its predicted excursion.

Rolling eye movements are due to oculovestibular and oculocervical reflexes

Compensatory eye movements during tilting of the head towards the shoulders initially involve the semicircular canals but, if the movement is sustained, static (utricle and saccule) responses participate. However, this compensatory movement of the eyes also involves information about neck position from proprioceptors in the neck (oculocervical reflex). Lateral movement of the head about a vertical axis will induce predominantly oculocervical reflexes, while movement of the head in the median plane with the eyes fixated produces predominantly oculovestibular reflexes (doll’s head movement). Doll’s head movements are an important clinical sign to test for intact brainstem reflexes in cases of cortical damage and loss of supranuclear control.

The midbrain is a coordinating centre for reflex eye movement and connects input from multiple sources

Voluntary eye movements (saccades) are initiated in the contralateral motor strip of the frontal cortex (see Ch. 1, p. 72) and pass down to the midbrain via the anterior limb of the internal capsule to synapse in the horizontal gaze centre within the PPRF (see Box 5-17; Fig. 5-40). Neurones then pass to the ipsilateral VI nerve and interneurones cross to the opposite

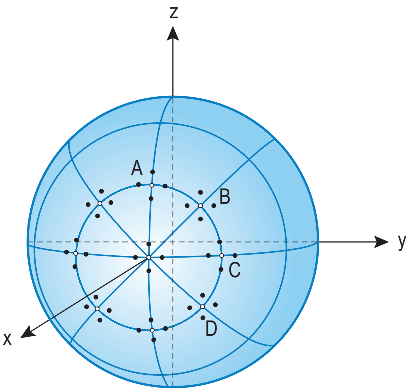

FIGURE 5-38 Primary position and Listing's plane. There is a unique orientation of the eye, called 'primary position' or 'primary gaze direction' (direction parallel to the *x*-axis), such that pure vertical and pure horizontal movements that move the eye or gaze line from primary to secondary positions do not change ocular torsion (eye rotations along the respective meridians through A or through C). Similarly, any movement that rotates the eye/gaze line from primary to tertiary positions on oblique meridian planes does not change torsion (e.g. movements along the meridians through B to D). The axes of single rotations that move the eye from primary to secondary or tertiary positions lie all in one plane, called Listing's plane (the plane containing the *y* and *z*axes). Tertiary positions cannot be reached from secondary positions by any combination of horizontal and vertical ocular rotations (a torsional component is also needed; see half-angle rule). (From Angelaki and Hess, 2004, with permission from Blackwell Publishing.)

medial longitudinal fasciculus to subserve the contralateral III nerve. Within the PPRF are burst cells, which have a high but transient rate of discharge (1000 Hz/s) and, when fired, generate the saccade. Normally the burst cells are continuously inhibited by pause cells, until this inhibition is released by discharge from neurones from the frontal eye fields. Burst-pause cells, and other types of cells such as pacemaking cells, are characteristic of different types of neurones in other areas of the brain such as the cerebellum.

Once the saccade has been generated, eye position and fixation are maintained via the tonic neural integrators, also situated within the PPRF. The PPRF also receives inputs from vestibular nuclei, cerebellum, basal ganglia and cervical proprioceptors, giving rise to fine accurate control of gaze, and their contributions will be discussed below.

The vertical gaze centre is located in the reticular medial longitudinal fasciculus (RMLF), opposite the superior colliculus and above the level of the III nerve nucleus. Unlike the horizontal gaze centre, the vertical gaze centre has no identifiable cortical control, neurones from which cross to the III and IV nerves nuclei to subserve vertical gaze. The medial longitudinal fasciculus, as already mentioned, carries fibres of conjugate horizontal eye movement (involving the VI and III nerves) and also signals for holding vertical eye position, vertical smooth pursuit and vertical vestibulo-ocular reflexes.

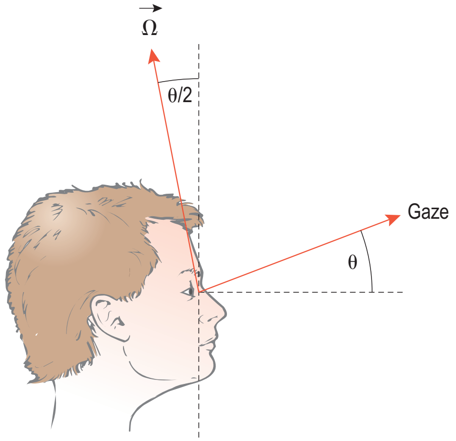

FIGURE 5-39 According to Listing's law, the axis of rotation of the eye (Ω) is neither head-fixed nor eye-fixed, but rotates in the same direction as gaze through half the gaze angle (θ/2; the half-angle rule). Thus, at eccentric eye positions, during a horizontal saccade or pursuit eye movement, the axis of rotation of the eye is not purely horizontal (head-vertical dashed line) but also has a torsional component (head-horizontal dashed line). (From Angelaki and Hess, 2004, with permission from Blackwell Publishing.)

Loss of supranuclear control by lesions affecting the midbrain and brainstem can give rise to a variety of clinical features, the commonest being involvement of the medial longitudinal fasciculus in multiple sclerosis, giving rise to abnormal horizontal saccades (internuclear ophthalmoplegia) (Box 5-18).

The superior colliculus is involved in both perception and eye movement control

A small number of fast (M) fibres relay from the retina to the superior colliculus, and thence to the pulvinar and finally the cortex. These fibres bypass the LGN and are described as the extrageniculostriate pathway. The fibres have crossed chiasmal representation, like optic tract fibres, and synapse in the superior

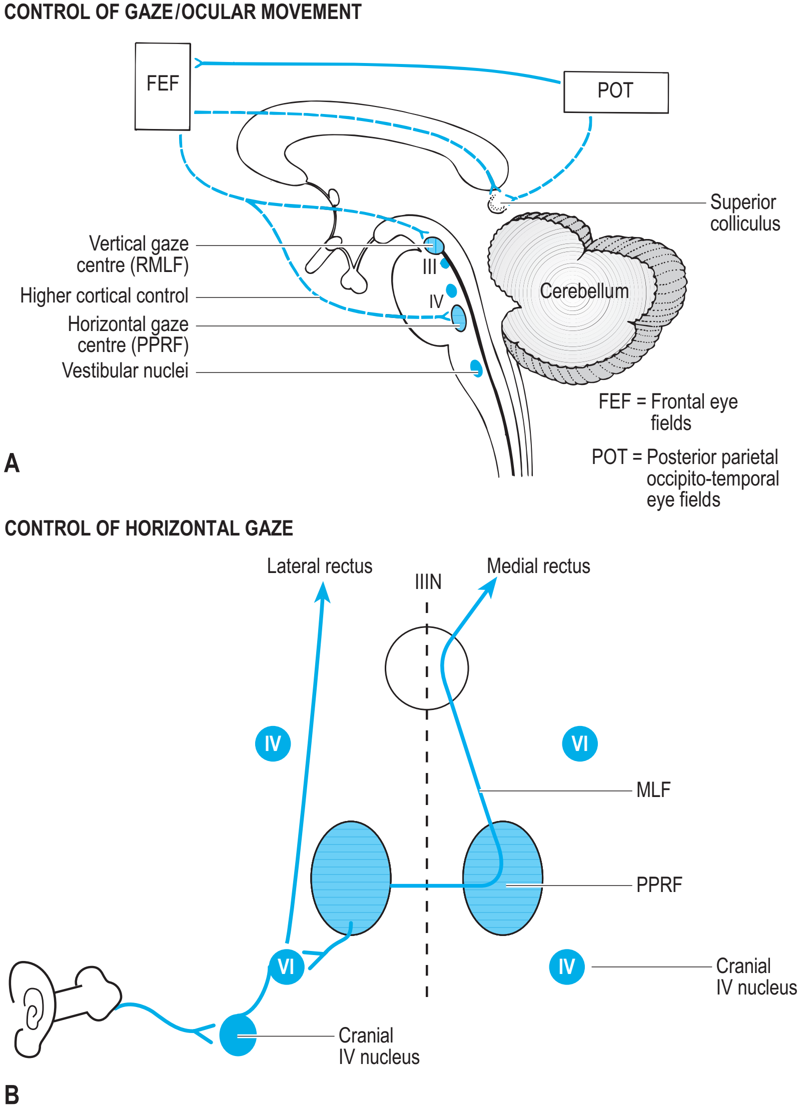
CONTROL OF GAZE/OCULAR MOVEMENT
FEF
POT
Superior colliculus
Vertical gaze centre (RMLF)
Higher cortical control
Horizontal gaze centre (PPRF)
Vestibular nuclei
III
IV
Cerebellum
FEF = Frontal eye fields
POT = Posterior parietal occipito-temporal eye fields
A
CONTROL OF HORIZONTAL GAZE
Lateral rectus
IIIN
Medial rectus
IV
VI
MLF
PPRF
VI
IV
Cranial IV nucleus
Cranial IV nucleus
B

FIGURE 5-40 Outline diagrams for integrated control of ocular movements: (**A**) 'higher centre' regulation; (**B**) brainstem, nuclear control. RMLF, reticular medial longitudinal fasciculus; PPRF, paramedian pontine reticular formation; MLF, medial longitudinal fasciculus.

colliculus in retinotopic structured layers, as occurs in fibres to the LGN. The organization, however, is less clearly demarcated, being in broad superficial and deep categories of fibres.

The function of these fibres is not entirely clear. There is evidence for neuronal delay to and from the visual cortex via the posterior pulvinar system of the thalamus and also to the pretectal region where the pupillary fibres relay (Fig. 5-41). The cells also have a receptive field organization and a preference for motion detection (a 'movement field'), particularly the rapid 'reflex' locking-on eye movement that occurs in the initiation of tracking a moving target or in automatic scanning during reading. In certain patients with occipital cortex lesions, 'blindsight' (the patient can detect motion or adjudge orientation without perceptually 'seeing' the object) may be present through preservation of this extrageniculostriate pathway.

Cells in the superficial layers of the pulvinar respond to visual input, while those in the deeper

BOX 5-18 fMRI IMAGING AND DETECTION OF DISCRETE BRAINSTEM LESIONS

fMRI imaging of patients with gaze palsies and similar oculomotor lesions are revealing discrete lesions in the brainstem. For instance, lesions in the PPRF or the VI nerve nucleus cause lateral gaze palsies, while internuclear ophthalmoplegia is caused by a lesion of the MLF, and the one-and-a-half syndrome is caused by lesions at both sites. Tiny infarcts in the region of the pons correlating with the gaze palsies have been described (reviewed by Bae et al., 2013).

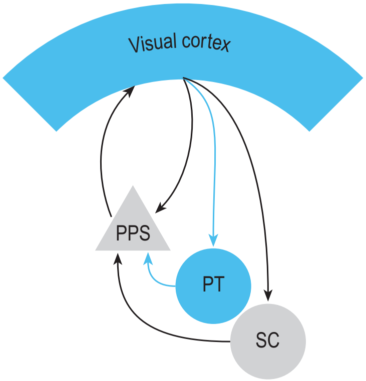

FIGURE 5-41 The extrageniculostriate pathway. Neuronal information to the superior colliculus (SC) is relayed to the posterior pulvinar system (PPS) and on to the visual cortex. From there it relays back to the PPS, the SC and the pretectal nucleus (PT), which completes the reverberating loop with the PPS.

layers respond to motion stimuli, although the cells in both layers are in register with each other. The pulvinar also receives many other subcortical inputs and acts as an ‘early processing centre’, receiving feedback and feedforward information from the cortex and the retina.

Recent studies in rats as well as fMRI studies in humans have suggested that connections through this pathway also occur with the amygdala and several other regions associated with ‘phobic’ stimuli and the experience of visually induced ‘fear’ perceptions induced by danger (Fig. 5-42).

Lesions in the pulvinar may thus affect such diverse functions as pattern recognition, eye movement and cerebellar integration in visual responses (see below). The colliculus–pulvinar–cortex relay system, therefore, is a priming system for ocular movement and for reducing errors in the localization response by linking visual and saccadic activity, possibly in response to danger.

Cortical centres regulate complex eye movements

Voluntary saccadic gaze movements are initiated by centres in the frontal cortex. Most of the fibres cross the midline in the anterior limb of the internal capsule to end in the gaze centres for motor neurone control of eye movement, but some pass to the ipsilateral superior colliculus where they inhibit automatic gaze responses (see above). In addition, cortical efferents remove tonic inhibitory impulses from collicular output to the ocular muscles, which are present between saccades, as if to ‘free them up’ for full-excursion eye movements.

Cortical voluntary saccades are ‘tuned’ as for automatic saccades, controlled by the superior colliculus; in the cortex, tuning is broad and appears to depend on recruitment of a precise number of neurones rather than a selected number of highly tuned neurones responsive to motion and visual activity. Single-cell recording studies have also shown that there are cells that exhibit presaccadic activity, while others respond only to the visual stimuli. A third group appears to show complex responses and may be involved in the integration of the response through a direct connection with the PPRF. The frontal eye fields also have an oculomotor loop to the substantia nigra in the basal ganglia, which contains high levels of dopamine. Loss of cells in the substantia nigra and a consequent decrease in dopamine concentration is a characteristic feature of Parkinson’s disease. This can be tested using the ‘anti-saccadic task’ test, in which the urge to fix on objects in the peripheral field is voluntarily suppressed through frontal eye field and superior collicular activity. Patients with this disorder characteristically have difficulties with voluntary gaze movements and the anti-saccadic task.

Tracking, smooth pursuit eye movements are under cortical control through relay of object position information from the occipital cortex to the posterior parietal cortex (motor cortex of smooth pursuit, posterior parietal occipito-temporal regions) and thence to the PPRF. Here the information is integrated with retinal information on object velocity via the optic tract and

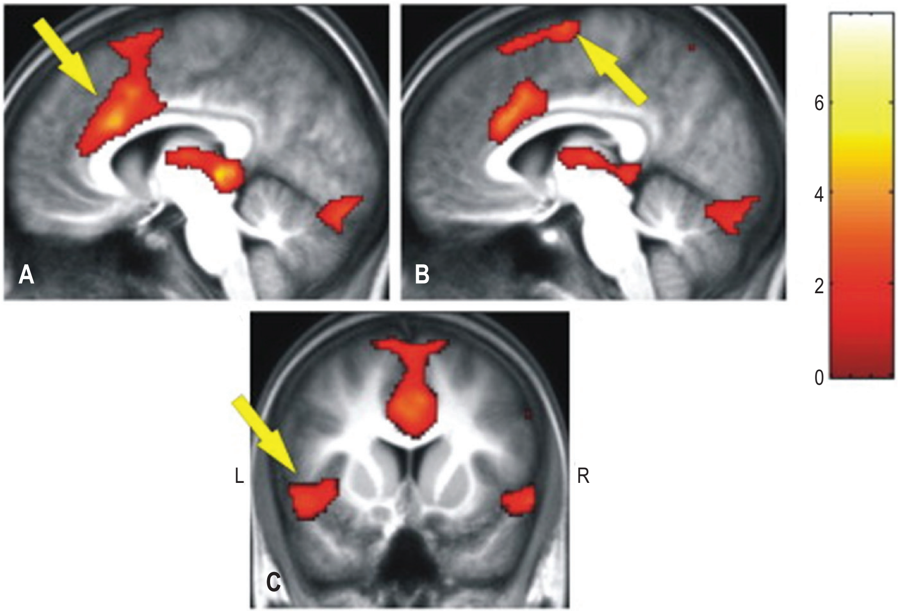

FIGURE 5-42 BOLD MRI activation maps. Patients with arachnophobia show much stronger signals compared with the controls in (**A**) the anterior cingulate cortex, (**B**) the supplementary motor cortex and (**C**) the insula for the contrast 'Spider > Neutral' (colour bar indicates the t-value) (From Goossens et al., 2007.)

with impulses from the head and neck on observer position, before being forwarded to the conjugate gaze centre for horizontal eye movement. In addition, recent studies have revealed that the frontal eye fields have control over pursuits as well as saccades.

The temporal cortex, as a centre for relay of motion detection information (see above), might be expected to be involved in the control of eye movement. Thus, lesions in this area affect saccades to moving targets but not to stationary ones, while smooth pursuit movements are also impaired.

Aside from obvious visual responses to objects and other specific stimuli, the process of entraining visual attention in deciding what is to be seen is extremely important. Clearly, higher cortical centres are at play here. The role of eye movements in initiating visual attention is paramount and requires integration of information from various regions of the brain, particularly the frontal eye fields (Fig. 5-43). Attention can be both overt (rapid fixating, saccadic eye movement) and covert (selective, 'out of the corner of the eye', not involving eye movement) and each can also be voluntary or involuntary (top-down or bottom-up, respectively). Parieto-frontal and superior collicular pathways are important in determining how attention is directed (see Fig. 5-42).

THE CEREBELLUM

Afferent input from the extraocular muscles (the stretch fibres and proprioceptors) is carried in the trigeminal nerve to synapse with cells in the granular cell layer (the Purkinje cell layer; see Ch. 1). There are two, and in some species three, different types of proprioceptor: muscle spindles, Golgi tendon organs and palisade endings, each restricted to the orbital, global and marginal layers of the ocular muscle, respectively. However, some direct cerebellar afferent input is also visual via slit-like, narrow, vertical receptive fields.

The bulk of information connecting the cerebellum with the visual system is transferred via two-way traffic with brainstem centres. The oculomotor cerebellar centre, located in lobules VI and VII, produces saccade-type movements for which Purkinje cells are essential. Input is derived via the PPRF, the vestibular nucleus and the mesencephalic reticular formation.

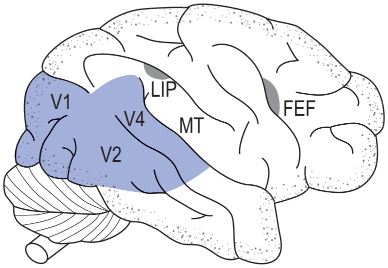

FIGURE 5-43 Schematic diagram of primate brain. Areas of the visual cortex which are activated by visual attention are V1, V2 and V4. The lateral intraparietal areas (LIP) and the frontal eye field (FEF) determine attention responses in conjunction with the control of eye movements, also served by the superior colliculus (not visible from this view). MT, middle temporal area. (From Bisley, 2011.)

The output from the cerebellum predominantly concerns positional sense, and some of it is inhibitory/regulatory. Positional information applies not only to the position of the observer with relation to the object but also to the velocity of the eye movement with relation to the target and the position of the head. Most of this is derived from the vestibular apparatus and not from visual or proprioceptive input.

This combined input also contributes to tracking movements. Experimental data have shown that cells in lobules VI and VII respond with bursts of activity during smooth pursuit movement when the eye is not actually fixating on a target (burst-pause cells).

Compensatory eye movements during movements of the head (e.g. during walking or running) are mediated mostly by the vestibulo-ocular reflex and less so by the cervico-ocular reflex through neck proprioceptors. The effect of these reflexes is to stabilize the retinal image by preventing 'retinal slip,' but the control is imperfect. The perceptual system can cope with a certain amount of retinal slip, but if this is too great (more than 5°/s), symptoms of oscillopsia appear. The cerebellum may contribute to control of the vestibulo-ocular reflex via input from the retina to the flocculus. It has been suggested that the cerebellum is the seat of a control mechanism for integration of information on spatial displacement during eye movement.

Currently, it is considered that there are two processes occurring simultaneously in vestibule-ocular control of posture and 'body sway': L namely, a fast-responding and short-lived response to visual cues such as parallax and a slower developing response to self-motion as it occurs during spontaneous displacement.

Ocular movements during natural activity

Most of the information regarding eye movement has come from studies that were designed to evaluate a particular movement, e.g. saccades or smooth pursuit movements. Investigation of eye movements during normal activities using infrared eye tracking devices have revealed the complex pattern of eye movements involved in performing day-to-day tasks, and has shown how extensively higher cortical information guides eye movements. Most recently, eye tracking devices have been developed which allow differentiation of microsaccades from even finer movements, termed ocular microtremors (OMTs) (Fig. 5-44). This has shown that microsaccades rather than OMTs underpin image persistence (prevent 'perceptual fade') during fixation (see p. 329).

Neural versus mechanical control of eye movement?

A continuing issue in studies of gaze control is to understand how much regulation of muscle function is mediated via neural control and how much can be attributed to mechanical effects of the muscles. Gaze control involves rotational three-dimensional movements (which obey Listing's law when the head is fixed but do not when the head is moving), plus other movements such as optokinetic nystagmus and those generated via the vestibulo-ocular reflex, to achieve image stabilization. Not all movements under control of the vestibulo-ocular reflex fail to obey Listing's law, such as the rotational movement that occurs in response to head movement, but the majority do. Movements which do not obey Listing's law, i.e. gaze shifts when the head is not restrained, obey Donder's law: for each position of gaze there is only one three-dimensional orientation (torsional movement).

Anatomical studies have suggested that the surrounding muscle sheath with the check ligaments can act like a pulley, allowing fine changes in the pulling direction related to the degree of torsional rotation. The data from these studies are also consistent with Listing's law and have formed the basis of some aspects of management of patients after strabismus surgery

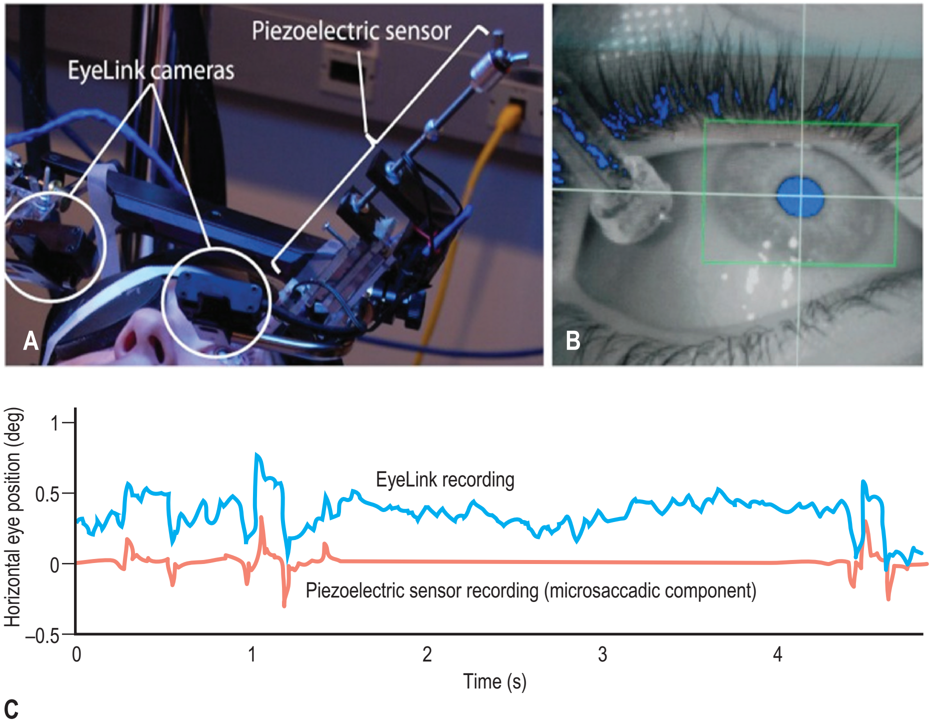

FIGURE 5-44 A custom-built set-up for recording simultaneous eye movement. In (**A**) a piezoelectric sensor is mounted to the EyeLink II helmet, while the image in (**B**) is a close-up of the sensor on the eye. The subject's pupil (blue) was tracked easily and data collected (**C**), showing microsaccade detection as small changes in eye position in degrees. (From McCamy et al., 2013.)

using adjustable sutures. However, the major, if not the sole, control of muscle behaviour rests with neural elements at both nuclear and supranuclear levels, with considerable regulation coming from higher centres such as the frontal eye fields and the cerebro-cerebellar network (see above). Moreover, some of the neurophysiological control would be consistent with a pulley mechanism while others such as saccades, which also obey Listing's law, are not. What is clear is that both mechanical and neural mechanisms regulate motor activity, but what remains to be determined is how the three-dimensional perception of space regulates the final motor command (top-down control) as well as the role of visual attention.

Conclusion

Visual neurophysiology continues to be the cornerstone of psychophysical neurophysiology but it is under increasing challenge from advances in imaging such as diffusion tensor MRI as well as very recent information on high-resolution mapping of the human brain (The Human Connectome Project http://www.humanconnectomeproject.org/). Increasingly, our understanding of what constitutes a visual image becomes more sophisticated. Research in this field is extremely active since there are many areas of uncertainty and, possibly more importantly, many potential applications of this knowledge to clinical medicine.

The primate visual system is a highly complex arrangement for analysing information concerning the external world derived from a wide array of possible signals, all of which are captured by the retinal sensory receptors. It is remarkable to consider that around 30% of all sensory information to the brain comes via the visual system. This information is integrated with input from many other sensory systems and stored information from past experience (memory).

The final image and its interpretation (perception) are extensively edited by the brain to ensure normality and 'constancy' wherever possible. However, this image is, of course, unique to each individual, despite the fact that we ascribe common definitions to familiar objects. Continuing research into the psychophysics of vision will fuel our thirst for knowledge into what makes us tick.

FURTHER READING

A full reading list is available online at https://expert-consult.inkling.com/).

FURTHER READING

Angelaki, D.E., Hess, B.J., 2005. Self-motion-induced eye movements: effects on visual acuity and navigation. Nat. Rev. Neurosci. 6, 966–976.

Anton-Erxleben, K., Carrasco, M., 2013. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. Nat. Rev. Neurosci. 14, 188–200.

Bae, Y.J., Kim, J.H., Choi, B.S., Jung, C., Kim, E., 2013. Brainstem pathways for horizontal eye movement: pathologic correlation with MR imaging. Radiographics 33, 47–59.

Barlow, H.B., 1972. Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1, 371–394.

Bergstrom, S.S., 2004. The AMBEGUJAS phenomenon and colour constancy. Perception 33, 831–835.

Bergstrom, S.S., Gustafsson, K.A., 2011. Concave and convex phases in ambiguous figures showing colour shifts. Mach's figure and the AMBEGUJAS phenomenon. Perception 40, 30–38.

Bertamini, M., Wagemans, J., 2013. Processing convexity and concavity along a 2-D contour: figure-ground, structural shape, and attention. Psychon. Bull. Rev. 20, 191–207.

Bisley, J.W., 2011. The neural basis of visual attention. J. Physiol. 589, 49–57.

Blake, R., Wilson, H., 2011. Binocular vision. Vision Res. 51, 754–770.

Bloomfield, S.A., Volgyi, B., 2009. The diverse functional roles and regulation of neuronal gap junctions in the retina. Nat. Rev. Neurosci. 10, 495–506.

Boynton, G.M., 2002. Color vision: how the cortex represents color. Curr. Biol. 12, R838–R840.

Braunstein, R.E., Sparrow, J.R., 2005. A blue-blocking intraocular lens should be used in cataract surgery. Arch. Ophthalmol. 123, 547–549.

Bray, S., Almas, R., Arnold, A.E., Iaria, G., Macqueen, G., 2013. Intraparietal sulcus activity and functional connectivity supporting spatial working memory manipulation. Cereb. Cortex [Epub ahead of print].

Bridge, H., 2011. Mapping the visual brain: how and why. Eye (Lond) 25, 291–296.

Cangiano, L., Dell'Orco, D., 2013. Detecting single photons: a supramolecular matter? FEBS Lett. 587, 1–4.

Collin, S.P., 2008. A web-based archive for topographic maps of retinal cell distribution in vertebrates. Clin. Exp. Optom. 91, 85–95.

Conway, B.R., 2009. Color vision, cones, and color-coding in the cortex. Neuroscientist 15, 274–290.

Cropper, S.J., Derrington, A.M., 1996. Rapid colour-specific detection of motion in human vision. Nature 379, 72–74.

D'Angelo, E., Mazzarello, P., Prestori, F., Mapelli, J., Solinas, S., Lombardo, P., et al., 2011. The cerebellar network: from structure to function and dynamics. Brain Res. Rev. 66, 5–15.

Demb, J.B., Singer, J.H., 2012. Intrinsic properties and functional circuitry of the AII amacrine cell. Vis. Neurosci. 29, 51–60.

Field, G.D., Rieke, F., 2002. Mechanisms regulating variability of the single photon responses of mammalian rod photoreceptors. Neuron 35, 733–747.

Field, G.D., Rieke, F., 2002. Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity. Neuron 34, 773–785.

Freeman, E.D., Sterzer, P., Driver, J., 2012. fMRI correlates of subjective reversals in ambiguous structure-from-motion. J. Vis. 12, 35.

Gainotti, G., 2013. Is the right anterior temporal variant of prosopagnosia a form of 'associative prosopagnosia' or a form of 'multimodal person recognition disorder'? Neuropsychol. Rev. 23, 99–110.

Gilbert, C.D., Li, W., 2013. Top-down influences on visual processing. Nat. Rev. Neurosci. 14, 350–363.

Goossens, L., Schruers, K., Peeters, R., Griez, E., Sunaert, S., 2007. Visual presentation of phobic stimuli: amygdala activation via an extrageniculostriate pathway? Psychiatry Res. 155, 113–120.

Grimes, W.N., 2012. Amacrine cell-mediated input to bipolar cells: variations on a common mechanistic theme. Vis. Neurosci. 29, 41–49.

Grossberg, S., Yazdanbakhsh, A., Cao, Y., Swaminathan, G., 2008. How does binocular rivalry emerge from cortical mechanisms of 3-D vision? Vision Res. 48, 2232–2250.

Harris, K.D., Thiele, A., 2011. Cortical state and attention. Nat. Rev. Neurosci. 12, 509–523.

Hartveit, E., Veruki, M.L., 2012. Electrical synapses between AII amacrine cells in the retina: Function and modulation. Brain Res. 1487, 160–172.

Hecht, E., 1987. Optics, second ed. Addison Wesley, Reading, Massachusetts, USA.

Hirasawa, H., Yamada, M., Kaneko, A., 2012. Acidification of the synaptic cleft of cone photoreceptor terminal controls the amount of transmitter release, thereby forming the receptive field surround in the vertebrate retina. J. Physiol. Sci. 62, 359–375.

Horiguchi, H., Winawer, J., Dougherty, R.F., Wandell, B.A., 2013. Human trichromacy revisited. Proc. Natl Acad. Sci. U.S.A. 110, E260–E269.

Koch, K.W., Dell'orco, D., 2013. A calcium-relay mechanism in vertebrate phototransduction. ACS Chem. Neurosci. 4, 909–917.

Koenig, D., Hofer, H., 2011. The absolute threshold of cone vision. J. Vis. 11, pii: 21.

Kravitz, D.J., Saleem, K.S., Baker, C.I., Mishkin, M., 2011. A new neural framework for visuospatial processing. Nat. Rev. Neurosci. 12, 217–230.

Kremers, J., Rodrigues, A.R., Silveira, L.C., da Silva Filho, M., 2010. Flicker ERGs representing chromaticity and luminance signals. Invest. Ophthalmol. Vis. Sci. 51, 577–587.

La Morgia, C., Ross-Cisneros, F.N., Sadun, A.A., Hannibal, J., Munarini, A., Mantovani, V., et al., 2010. Melanopsin retinal ganglion cells are resistant to neurodegeneration in mitochondrial optic neuropathies. Brain 133, 2426–2438.

McCamy, M.B., Collins, N., Otero-Millan, J., Al-Kalbani, M., Macknik, S.L., Coakley, D., et al., 2013. Simultaneous recordings of ocular microtremor and microsaccades with a piezoelectric sensor and a video-oculography system. PeerJ 1, e14.

Marmor, M.F., Brigell, M.G., McCulloch, D.L., Westall, C.A., Bach, M., 2011. ISCEV standard for clinical electro-oculography (2010 update). Doc. Ophthalmol. 122, 1–7.

Martinez-Conde, S., Macknik, S.L., Troncoso, X.G., Hubel, D.H., 2009. Microsaccades: a neurophysiological analysis. Trends Neurosci. 32, 463–475.

Martinez-Conde, S., Otero-Millan, J., Macknik, S.L., 2013. The impact of microsaccades on vision: towards a unified theory of saccadic function. Nat. Rev. Neurosci. 14, 83–96.

Masland, R.H., 2011. Cell populations of the retina: the Proctor lecture. Invest. Ophthalmol. Vis. Sci. 52, 4581–4591.

Masland, R.H., 2012. Another blue neuron in the retina. Nat. Neurosci. 15, 930–931.

Masland, R.H., 2012. The neuronal organization of the retina. Neuron 76, 266–280.

Masland, R.H., 2012. The tasks of amacrine cells. Vis. Neurosci. 29, 3–9.

Mendola, J.D., Buckthought, A., 2013. fMRI investigation of monocular pattern rivalry. J. Cogn. Neurosci. 25, 62.

Noudoost, B., Moore, T., 2011. Control of visual cortical signals by prefrontal dopamine. Nature 474, 372–375.

Op de Beeck, H.P., Haushofer, J., Kanwisher, N.G., 2008. Interpreting fMRI data: maps, modules and dimensions. Nat. Rev. Neurosci. 9, 123–135.

Palczewski, K., 2012. Chemistry and biology of vision. J. Biol. Chem. 287, 1612–1619.

Parker, A.J., 2007. Binocular depth perception and the cerebral cortex. Nat. Rev. Neurosci. 8, 379–391.

Peelen, M.V., Downing, P.E., 2007. The neural basis of visual body perception. Nat. Rev. Neurosci. 8, 636–648.

Pelli, D.G., Bex, P., 2013. Measuring contrast sensitivity. Vision Res. 90, 10–14.

Piccolino, M., Wade, N.J., 2008. Galileo Galilei's vision of the senses. Trends Neurosci. 31, 585–590.

Provis, J.M., Dubis, A.M., Maddess, T., Carroll, J., 2013. Adaptation of the central retina for high acuity vision: cones, the fovea and the avascular zone. Prog. Retin. Eye Res. 35, 63–81.

Reuter, T., 2011. Fifty years of dark adaptation 1961–2011. Vision Res. 51, 2243–2262.

Rieiro, H., Martinez-Conde, S., Danielson, A.P., Pardo-Vazquez, J.L., Srivastava, N., Macknik, S.L., 2012. Optimizing the temporal dynamics of light to human perception. Proc. Natl Acad. Sci. U.S.A. 109, 19828–19833.

Roe, A.W., Chelazzi, L., Connor, C.E., Conway, B.R., Fujita, I., Gallant, J.L., et al., 2012. Toward a unified theory of visual area V4. Neuron 74, 12–29.

Rossion, B., Hanseeuw, B., Dricot, L., 2012. Defining face perception areas in the human brain: a large-scale factorial fMRI face localizer analysis. Brain Cogn. 79, 138–157.

Salomon, R.G., Hong, L., Hollyfield, J.G., 2011. Discovery of carboxyethylpyrroles (CEPs): critical insights into AMD, autism, cancer, and wound healing from basic research on the chemistry of oxidized phospholipids. Chem. Res. Toxicol. 24, 1803–1816.

Sand, A., Schmidt, T.M., Kofuji, P., 2012. Diverse types of ganglion cell photoreceptors in the mammalian retina. Prog. Retin. Eye Res. 31, 287–302.

Sasaki, Y., Nanez, J.E., Watanabe, T., 2010. Advances in visual perceptual learning and plasticity. Nat. Rev. Neurosci. 11, 53–60.

Schmidt, T.M., Do, M.T., Dacey, D., Lucas, R., Hattar, S., Matynia, A., 2011. Melanopsin-positive intrinsically photosensitive retinal ganglion cells: from form to function. J. Neurosci. 31, 16094–16101.

Schwartz, O., Hsu, A., Dayan, P., 2007. Space and time in visual context. Nat. Rev. Neurosci. 8, 522–535.

Solomon, S.G., Lennie, P., 2007. The machinery of colour vision. Nat. Rev. Neurosci. 8, 276–286.

Spillmann, L., 2009. Phenomenology and neurophysiological correlations: two approaches to perception research. Vision Res. 49, 1507–1521.

Sterling, P., 2013. Some principles of retinal design: the Proctor lecture. Invest. Ophthalmol. Vis. Sci. 54, 2267–2275.

Sung, C.H., Chuang, J.Z., 2010. The cell biology of vision. J. Cell Biol. 190, 953–960.

Tanimoto, N., Brombas, A., Muller, F., Seeliger, M.W., 2012. HCN1 channels significantly shape retinal photoresponses. Adv. Exp. Med. Biol. 723, 807–812.

Taylor, W.R., Smith, R.G., 2012. The role of starburst amacrine cells in visual signal processing. Vis. Neurosci. 29, 73–81.

Temple, S.E., 2011. Why different regions of the retina have different spectral sensitivities: a review of mechanisms and functional significance of intraretinal variability in spectral sensitivity in vertebrates. Vis. Neurosci. 28, 281–293.

Toscani, M., Valsecchi, M., Gegenfurtner, K.R., 2013. Optimal sampling of visual information for lightness judgments. Proc. Natl Acad. Sci. U.S.A. 110, 11163–11168.

Ts'o, D.Y., Zarella, M., Burkitt, G., 2009. Whither the hypercolumn? J. Physiol. 587, 2791–2805.

van Pelt, S., Fries, P., 2013. Visual stimulus eccentricity affects human gamma peak frequency. Neuroimage 78C, 439–447.

Vaney, D.I., Sivyer, B., Taylor, W.R., 2012. Direction selectivity in the retina: symmetry and asymmetry in structure and function. Nat. Rev. Neurosci. 13, 194–208.

Verdon-Roe, G.M., Westcott, M.C., Viswanathan, A.C., Fitzke, F.W., Garway-Heath, D.F., 2006. Exploration of the psychophysics of a motion displacement hyperacuity stimulus. Invest. Ophthalmol. Vis. Sci. 47, 4847–4855.

Vos, J.J., 2006. From lower to higher colour metrics: a historical account. Clin. Exp. Optom. 89, 348–360.

Wandell, B.A., 2011. The neurobiological basis of seeing words. Ann. N. Y. Acad. Sci. 1224, 63–80.

Wandell, B.A., Smirnakis, S.M., 2009. Plasticity and stability of visual field maps in adult primary visual cortex. Nat. Rev. Neurosci. 10, 873–884.

Watson, A.B., Ahumada, A.J., 2011. Blur clarified: a review and synthesis of blur discrimination. J. Vis. 11, pii: 10.

Wedeen, V.J., Rosene, D.L., Wang, R., Dai, G., Mortazavi, F., Hagmann, P., et al., 2012. The geometric structure of the brain fiber pathways. Science 335, 1628–1634.

Wei, W., Feller, M.B., 2011. Organization and development of direction-selective circuits in the retina. Trends Neurosci. 34, 638–645.

Westheimer, G., 2009. The third dimension in the primary visual cortex. J. Physiol. 587, 2807–2816.

Westheimer, G., 2009. Visual acuity: information theory, retinal image structure and resolution thresholds. Prog. Retin. Eye Res. 28, 178–186.

Westheimer, G., 2011. Three-dimensional displays and stereo vision. Proc. Biol. Sci. 278, 2241–2248.

Westheimer, G., 2012. Optical superresolution and visual hyperacuity. Prog. Retin. Eye Res. 31, 467–480.

Wiechmann, A.F., Summers, J.A., 2008. Circadian rhythms in the eye: the physiological significance of melatonin receptors in ocular tissues. Prog Ret Eye Res 27, 137–160.

Wilson, C.R., Gaffan, D., Browning, P.G., Baxter, M.G., 2010. Functional localization within the prefrontal cortex: missing the forest for the trees? Trends Neurosci. 33, 533–540.

Wise, S.P., 2008. Forward frontal fields: phylogeny and fundamental function. Trends Neurosci. 31, 599–608.

Wong, A.M., 2004. Listing's law: clinical significance and implications for neural control. Surv. Ophthalmol. 49, 563–575.

Wu, J., Yan, T., Zhang, Z., Jin, F., Guo, Q., 2012. Retinotopic mapping of the peripheral visual field to human visual cortex by functional magnetic resonance imaging. Hum. Brain Mapp. 33, 1727–1740.

Zeki, S., 2011. The mind's eye. Cerebrum 2011, 4.

Zhang, H., Maloney, L.T., 2012. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. Front. Neurosci. 6, 1.

Chapter **6**

General and ocular pharmacology

- Introduction
- Pharmacokinetics: drug trafficking in the body
- Pharmacodynamics: drug handling by the body
- Drug–receptor interactions
- Ocular pharmacology: drug handling by cells and tissues of the eye
- Delivery methods of ocular medication
- Reconstituting the tear film
- Ocular drugs and the autonomic nervous system
- Clinical control of intraocular pressure exploits the adrenergic system
- The histaminergic system: histamine is released from conjunctival mast cells during allergic reactions
- Eicosanoids affect multiple ocular functions
- Serotonin: a potent neurotransmitter
- Glucocorticoids
- Immunosuppressant agents: combating ocular inflammatory disease
- Local anaesthetics: an integral part of ophthalmic examination and surgery
- Ocular toxicity from systemic administration of drugs

Introduction

This chapter covers the basic principles of clinical pharmacology, with particular reference to drugs used in the management of ophthalmic disorders, methods of ocular drug delivery, and the interactions of drugs and the eye. Although it is generally thought that the medical management of ocular disease is mainly administered through topical therapeutic agents, many systemic drugs and agents are also used. These include diuretics for the control of intraocular pressure, immunosuppressants for control of intraocular inflammatory conditions and antimicrobials for control of infection. Therefore, basic pharmacological principles (pharmacokinetics and pharmacodynamics) are important to ophthalmologists. The basic pharmacology of systemic therapy, including receptor–drug interactions, is reviewed before discussing the more specific topical and systemic therapies in the treatment of ophthalmic disorders.

Pharmacokinetics: drug trafficking in the body

BASIC CONCEPTS

Pharmacokinetics is the mathematical study of the time-course of drug absorption, distribution, metabolism and excretion. One of the simplest parameters to consider when discussing the pharmacokinetics of any drug is the biological *half-life* of a drug ($t_{1/2}$), which is the time taken for the plasma drug concentration to fall by half after administration. A more accurate method of assessing the efficiency of drug elimination is the estimation of drug clearance from the circulation. For example, after intravenous administration, the $t_{1/2}$ of the drug may be calculated from its plasma concentration–time curve. This simplistic model is based on a single compartment model that states that, following intravenous administration, the distribution of a drug assumes a uniform concentration throughout all compartments (intracellular and extracellular); the elimination by both metabolism and renal excretion is also assumed to be directly proportional to the drug concentration. If this is the case, the volume of distribution may be calculated as:

$$V_d = \text{dose}/C_0$$

where C_0 is the estimated (plasma) concentration at the time of injection.

The apparent volume of distribution is defined as the volume of fluid required to contain the total amount of drug in the body at the same concentration as that present in plasma. Drugs in general may be confined to the plasma compartment because they are either too large to cross the capillary wall or are highly protein-bound (e.g. heparin and warfarin, respectively). Drugs may also be distributed only to the extracellular compartment because they have low lipid solubility (e.g. gentamicin) or distributed throughout all aqueous compartments if they are lipid-soluble. When both the volume of distribution and renal clearance of a drug are known, the $t_{1/2}$ may be estimated. However, when considering a single compartment model, no account is taken of distribution into tissue compartments and metabolism within tissue compartments. When repeated injections of the drug are given, the plasma concentration becomes a function of the rate of both elimination and administration, and the plasma concentration will equilibrate when these two parameters are equal. A steady-state plasma concentration is therefore reached, which in practice may be established after an interval of about three or four plasma half-life lives of the drug given. To reach a steady-state concentration more quickly, loading doses of the drug are often given (for example, antibiotics and warfarin) (Fig. 6-1). In clinical practice the $t_{1/2}$ is important because it will determine the frequency of administration of a drug. If continued intravenous boluses are given, there are frequently large peaks and troughs in drug concentration, which can lead to a greater incidence of toxic side-effects.

Bioavailability describes the amount of oral dose that reaches the systemic circulation and becomes available to the site of drug action. However, this blanket term is not sufficiently precise because rapidly absorbed drugs will reach a much higher plasma concentration than those absorbed slowly and, similarly, rapid elimination would also theoretically lead to low bioavailability. Bioavailability of a drug is measured as the area under the curve of log plasma concentration against time for both intravenous and oral administration, although the reliability of quantitative drug assessment may be variable for several reasons. For instance, the degree of bioavailability may be altered by incomplete absorption of the drug or destruction of the drug by *first-pass metabolism* before the drug reaches the plasma compartment, irrespective of the rate of absorption from the gastrointestinal tract. Drugs instilled into the eye are absorbed from the nasal and nasopharyngeal mucosae directly into the systemic circulation. As such, they escape first-pass metabolism and have a high bioavailability. Thus, topically administered agents can give rise to quite marked systemic effects (see below).

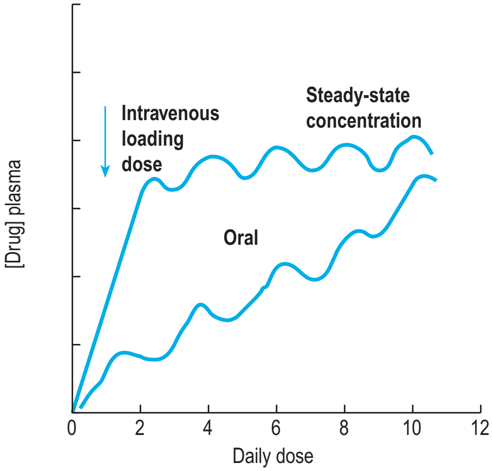

FIGURE 6-1 Plasma concentration of drugs after oral and intravenous administration.

Drug kinetics can be described as first-order (linear) or zero-order (non-linear, saturation) kinetics (Fig. 6-2). First-order kinetics describes a process where rate is proportional to the amount of drug present and can be defined by linear differential equations. Zero-order kinetics occurs when drug dynamics show *saturation* at high drug concentrations. Saturation may occur, for example, when the capacity of drug-metabolizing liver enzymes is surpassed, leading to unmetabolized drug in the circulation for longer periods. The duration of action of a drug that exhibits saturation kinetics is more dependent on the administering dose than in drugs that exhibit first-order kinetics. Also, there is no direct relationship between drug dose and steady-state plasma concentration in zero-order kinetics, which may explain sudden unexpected drug toxicity in a number of clinical settings. For this reason, close drug monitoring is required (e.g. phenytoin).

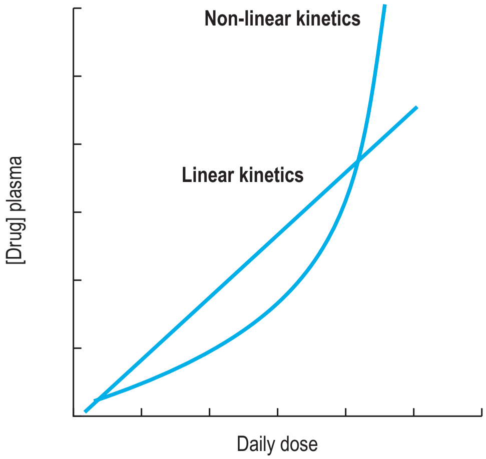

FIGURE 6-2 First-order and zero-order pharmacokinetics.

DRUG ABSORPTION

For a drug to reach the site at which it produces its effect, it must first be absorbed from its site of administration. In general, drug penetration of cell membranes increases with lipid solubility. Drugs can cross cell membranes by diffusing directly through lipid, by diffusion through aqueous pores that traverse the lipid, by utilizing carrier molecules, or by pinocytosis by the cell.

The rate of passage (diffusion) through a cell membrane can be predicted by Fick's law

Non-polar substances dissolve freely in lipid and therefore penetrate the cell by diffusion. The lipid solubility, degree of ionization and molecular size of the drug will determine its diffusion coefficient. The rate is determined by Fick's law, which states that the rate at which drugs cross a biological membrane is directly proportional to the concentration gradient across the membrane and the diffusion coefficient, and inversely proportional to the cell membrane thickness (see below):

$$\text{Rate of diffusion} = KA(x_1 - x_2)/D$$

where K is the diffusion constant, A is the diffusion area, $x_1 - x_2$ is the concentration difference between plasma and intracellular compartments, and D is the thickness of the membrane.

Active transport

As for active transport of ions (see Ch. 4, p. 161), active transport of large polar drugs requires energy-dependent carrier-mediated mechanisms. These transport systems may be disrupted by inhibiting enzyme-dependent carriers or by blocking the carrier mechanism with structurally similar drugs (analogues). Facilitated transport is carrier-mediated transport that does not require energy because it does not proceed against a concentration gradient, for example glucose transport into erythrocytes.

DRUG ABSORPTION IS DEPENDENT ON A DRUG'S LIPID SOLUBILITY

Drug properties

The absorption of a drug depends on its lipid solubility and inversely on its polarity or degree of ionization. An important factor in the degree of penetration of a drug through membranes is that many drugs are weak acids or weak bases. The more the drug is in its un-ionized form, the more likely it is to be lipid-soluble and transferred by passive diffusion through the membrane. For a weak acid or base the pK_a value will determine the degree of ionization, as described by the Henderson–Hasselbalch equation.

For a weak acid the ionizing reaction is:

$$\text{pH} = \text{p}K_a + \log\{[\text{A}^-]/[\text{HA}]\}$$

and for a weak base it is:

$$\text{pH} = \text{p}K_a + \log\{[\text{B}]/[\text{BH}^+]\}$$

The pK_a is a measure of the relative strength (degree of ionization) of a weak acid or base (the pK_a of a drug is that point at which the compound is 50% ionized). The lipid solubility of the uncharged species also depends on the chemical nature of the drug. For example, although streptomycin and the related aminoglycosides are uncharged, the high percentage of hydrogen-bonding groups within these molecules renders them hydrophilic.

The molecular shape, as well as the charge distribution, of the drug molecule determines which membrane pores it may traverse. Small polar molecules, such as urea, readily traverse small aqueous pores in the membrane, thus accounting for the high permeability of cell membranes to these substances. Most drugs, however, are too large to pass through these pores.

The rate of drug absorption varies with the route of administration

The main routes of administration (besides intravenous) are oral, sublingual, rectal, topical (e.g. skin, conjunctival fornix), subcutaneous and intramuscular.

Absorption of orally administered drugs is affected by gastric pH, rate of emptying of gastric contents, presence of food, and surface area of absorptive mucosa (in disorders such as Crohn's disease the absorptive surface area may be reduced). It is important to consider drug interactions during multiple drug therapy and their effect on drug absorption. For example, in migraine, gastric emptying is delayed, reducing the absorption of analgesics such as aspirin and paracetamol. This may be overcome with the adjunctive use of parenteral metoclopramide, which increases the rate of gastric emptying. The presence of food is generally unimportant, except in the case of tetracyclines (used to treat certain forms of external eye disease commonly associated with acne rosacea), which form insoluble salts with magnesium and calcium. Some drugs will be inactivated within the gut lumen (e.g. benzylpenicillin and insulin). Most malabsorption syndromes do not affect drug absorption, but in other disorders, such as congestive cardiac failure, drug absorption may be impaired because of the secondary gastrointestinal mucosal oedema. Factors affecting bioavailability are shown in Box 6-1.

DRUG DISTRIBUTION

Once a drug is absorbed, it has the potential to penetrate most compartments of the body so the distribution of the drug depends largely on the route of administration. Intravenous and intramuscular administration of a drug result in high drug availability. Buccal (sublingual) absorption of a drug is used to reduce the extent of first-pass metabolism by the liver (as with topically applied eye drops), which invariably occurs with orally administered drugs that reach the liver via the portal circulation. The distribution of the drug and its ability to penetrate cells is also dependent on its physicochemical properties and thus the extent of binding to tissue proteins or cell membrane receptors (Box 6-2).

Once the drug has reached the systemic circulation, it may become bound to circulating proteins, commonly *albumin* or *α_1-acid glycoproteins* (for basic drugs). Protein-bound drugs are restricted in their distribution into tissues, which reduces the availability of the free drug for pharmacological effect. This, in turn, depends on the affinity of the protein for the drug. High levels of protein binding may occur with acidic drugs that are bound to albumin. If a drug is less than 90% bound to plasma proteins, changes in the plasma protein concentration make little difference to the overall amount of unbound drug in the circulation. In cases of drugs with a high binding affinity for protein, a decrease in the steady-state total concentration of the drug, and as a consequence a comparative increase in the clearance of the increased amount of free drug, would occur if for any reason protein binding was impaired. For instance, impaired plasma protein binding in the case of phenytoin results in peaks of unbound active drug in the plasma.

Basic drugs are bound in varying degrees to α_1-acid glycoprotein, a protein that increases in concentration in certain pathological conditions such as acute inflammation. Under such conditions, the binding of basic drugs (e.g. propranolol) may be increased, thus reducing their effect.

BOX 6-1 FACTORS ALTERING ABSORPTION AND BIOAVAILABILITY FROM GUT

- Gut motility
- Intestinal pH, mucus, bile salts
- Enterohepatic circulation
- Exercise
- Reduced absorptive area
- Reduced intestinal blood flow
- Intestinal microflora that may metabolize some drugs

BOX 6-2 FACTORS AFFECTING DRUG DISTRIBUTION

- Physicochemical properties of drug
- Binding to plasma proteins
- Binding to tissue proteins
- Relative blood flow to different tissues

Systemic factors that alter protein binding include hypoalbuminaemia (plasma concentration of albumin less than 25 g/L), renal failure, competition by other highly protein-bound drugs, and changes occurring during the last trimester of pregnancy, such as the diluting effect of the increased plasma volume. In general, competition for binding by other drugs is the major factor affecting distribution because adequate compensatory mechanisms can be initiated to counteract the other causes. The distribution of a drug, as mentioned above, may also be regulated by the binding of a drug to tissue proteins, a process regulated by the abundance of binding sites, affinity constants and the binding of a drug to its receptor. This in turn may give rise to a desired or undesired effect, although receptor numbers are unlikely to be high enough to alter the distribution of the drug appreciably.

In the eye, both the blood–retinal barrier and the blood–aqueous barrier (see Ch. 1, p. 30) limit the distribution of drugs. Tight junctions between the retinal pigment epithelium (RPE) and endothelium of the retinal vessel endothelium give rise to a relatively impermeable barrier to water solutes and larger molecules. Under normal conditions only lipid-soluble drugs will move between the blood and retina. Similarly, the apical membranes of the non-pigmented ciliary body epithelium and the capillary endothelium of the iris are bound by tight junctions, raising a barrier to all but lipid-soluble drugs (see Ch. 1, p. 28).

DRUGS ARE METABOLIZED TO FACILITATE CLEARANCE

The metabolism of most drugs occurs almost entirely in the liver, enzymatically altering the drug to increase its water solubility in preparation for excretion, and simultaneously making the compound metabolically and pharmacologically active or inactive. Metabolism can affect the drug in various ways. The first is activation of the parent drug, which may itself be inactive (known as a *prodrug*). The drug may be metabolized to form active metabolites, as for example in the case of diamorphine and diazepam. However, metabolism of a drug can also produce toxic metabolites that may persist in the circulation for longer than the parent molecule and thus restrict the continued use of such drugs (e.g. lidocaine). In general, most drug metabolism modifies the drug so that it can then be excreted, usually in the urine but sometimes in bile.

Drug metabolism takes place in two stages: phase I (oxidation) and phase II (conjugation). Phase I reactions are carried out by a heterogeneous group of microsomal enzymes called cytochrome P_{450}, of which various forms exist (Box 6-3).

This enzyme system exists in abundance in the liver but is also found in some peripheral organs, including the eye. It is important to note that the activity of these enzymes can be induced by other drugs (e.g. phenytoin and carbamazepine), which accounts for many well-recognized drug interactions. Conjugation reactions appear not to be affected by enzyme-inducing drugs to the same extent as oxidative metabolism. There are also other oxidative enzymes that are not part of the cytochrome P_{450} system but which are involved in drug metabolism; these include xanthine oxidase (e.g. purine metabolism), alcohol dehydrogenase and monoamine oxidase (e.g. catecholamine metabolism). Many drugs also affect the function of microsomal enzyme systems (Box 6-4).

The metabolism of a drug is dependent on several factors. Oxidative metabolism is affected by age. Premature babies metabolize poorly and, similarly in the elderly, oxidative drug metabolism is reduced because of reduced liver size. Smoking may induce certain liver enzymes, necessitating an increased dose of drug for a required effect. Alcohol, on the other hand,

BOX 6-3 DRUG-METABOLIZING SYSTEMS

PHASE I METABOLISM: OXIDATION

Cytochrome P_{450} (microsomal)
- aromatic hydroxylation
- aliphatic hydroxylation
- *N*-deamination
- *N*-dealkylation
- *S*-oxidation
- desulphuration

PHASE II METABOLISM: CONJUGATION

Conjugation occurs with:
- glucuronic acid
- glycine
- glutamine
- sulphate
- acetate

BOX 6-4 DRUGS ACTING AS MICROSOMAL ENZYME MANIPULATORS

ENZYME INDUCERS

- barbiturates
- phenytoin
- phenothiazines
- rifampicin
- griseofulvin
- nicotine

ENZYME INHIBITORS

- isoniazid
- chloramphenicol
- metronidazole
- warfarin
- carbon monoxide

inhibits drug metabolism, particularly during and after binge drinking. Both severe liver disease and poor nutrition may markedly impair drug metabolism.

METABOLIZED DRUGS ARE EXCRETED IN URINE AND BILE

Renal excretion

Drugs differ greatly in their excretion via the kidney. Some drugs are cleared in a single transit through the kidney, while others are poorly cleared. This difference is dependent on the kidney's ability to handle the drug and on physicochemical properties of the drug. Certain drugs may be filtered through the glomerulus (depending on molecular weight), others are actively secreted by the tubules, and still others passively diffuse across the tubular epithelia (reabsorption).

Glomerular filtration. Glomerular excretion of a drug is only possible when drugs are not bound to plasma proteins, and the drug is of a molecular size that can be filtered freely, irrespective of charge (less than 20 000 Da). The clearance of the drug is therefore related to the unbound fraction of the drug and is dependent on the glomerular filtration rate. Glomerular filtration at most removes only about 20% of drug reaching the kidney; the remaining drug passes to the capillaries lining the tubules. Most drugs are presented to the kidney in a carrier form and thus their clearance is slow. Carrier-mediated transport can increase the clearance of the drug even if it is highly bound. Benzylpenicillin is 80% bound and cleared very slowly by glomerular filtration but is almost completely removed by tubular secretion.

Secretion. Drugs may be actively secreted by the tubules into the glomerular filtrate. Different secretory systems, which are relatively non-selective carrier systems, exist for acidic, basic and neutral drug transport into the tubular lumen. Competition for active sites occurs, so that the secretion of drugs can be blocked by other drugs, for example secretion of penicillins is blocked by probenecid, thus decreasing the excretion of the drug.

Reabsorption. If the tubules were freely permeable to drugs, the drug concentration in the filtrate would be similar to that in plasma. Drug reabsorption occurs mainly as a result of the enormous reabsorption of water achieved by the nephron. This builds up a concentration gradient that drives the drug back into the plasma. Thus drugs that are highly lipid-soluble are excreted slowly and the reabsorption of drugs, especially those that are weak acids or bases, depends on the drug being in its un-ionized form. Therefore, altering the pH of the urine will increase the elimination of the drug, so that acidic drugs are eliminated more quickly in alkaline urine. This manipulation of urine pH is utilized clinically for the treatment of aspirin overdoses, where forced alkaline diuresis is an effective way of increasing the excretion of the acidic drug. Measurement of glomerular filtration can be achieved with the use of agents that are completely filtered and neither secreted nor reabsorbed, for example inulin. Conversely, the clearance of a drug that is completely secreted in one transit through the kidney will correspond to the renal plasma flow, for example *p*-aminohippuric acid.

Biliary excretion

Liver cells also possess transport systems similar to those in the renal tubules, which can transfer drug metabolites from blood to bile. Conjugated drugs (particularly with glucuronate) are concentrated in the bile and delivered to the intestine, where the conjugate may be hydrolysed, releasing the active drug (Fig. 6-3). The drug may then be reabsorbed and its duration of action prolonged (enterohepatic circulation).

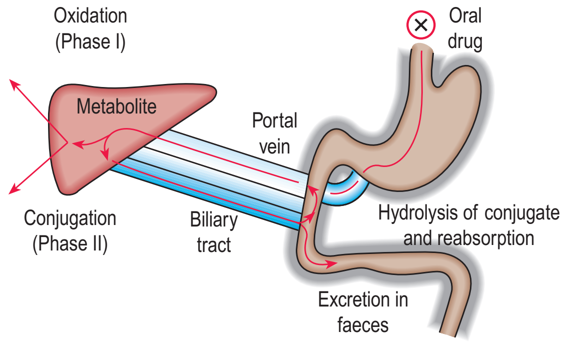

FIGURE 6-3 Drug metabolism: enterohepatic circulation.

This is particularly important for digoxin, which is excreted in the bile in an unconjugated form, and for morphine, which is transported as a glucuronide. Some drugs, for example rifampicin, are excreted into the bile and will be excreted unchanged in the faeces.

Pharmacodynamics: drug handling by the body

Pharmacodynamics considers the effects of a drug and the relationship between drug concentration and response. The drug effect is usually initiated by its binding to a cellular membrane receptor, which is specific for the drug, as is the case with autonomic nervous system neurotransmitters, or by non-specific mechanisms, where specific cell membrane receptors do not exist but drug action is dependent solely on its physical properties (i.e. lipid solubility), or by its ability to inhibit specific biochemical enzymes. The effects of enzyme inhibition may be direct, for example by blocking sodium/potassium ATPase pumps or closing ion channels, or indirect, for example by acting on calcium channels. Some drugs act on intracytoplasmic receptors or cell nucleus receptors. The variety of drug actions is presented in Box 6-5.

Additional content available at https://expertconsult.inkling.com/

Drug–receptor interactions

Drugs bind to target molecules on the cell, which then initiate their pharmacological effect. Receptors are proteins that, in general, are situated on the cell surface and are able to bind ligands (e.g. neurotransmitters and hormones) (see Box 6-5). The recognition of a ligand by a receptor is analogous to antibody–antigen binding, that is the lock and key principle (see Ch. 7, p. 397). The molecular mechanisms involved in this receptor–effector pathway are known as *transduction* mechanisms, where the receptor is linked either directly or indirectly to the effector system altering cell function. Several forms of receptor–effector linkage are recognized, and include:

- direct regulation of membrane permeability to ions
- regulation of cell function via an intracellular second messenger
- regulation of cell function by regulating DNA transcription and protein synthesis.

BOX 6-5 RECEPTORS

Receptors are unique cell surface proteins capable of binding specific associated substances (ligands). Receptors may be divided into subtypes, depending on the type of agonist or antagonist associated with them. Often receptors can be downregulated or upregulated in different disease states or with chronic drug usage.

A ligand is a specific compound (drug or natural substance) that binds to a receptor.

Agonists are substances that, when bound to a receptor, produce a response that may result in stimulation or inhibition of cell function.

Antagonists are substances that prevent receptor activation.

A partial agonist is a ligand that possesses both antagonist and agonist properties. As such, its maximal effect is less than that of a pure agonist. However, when occupying a receptor it prevents the actions of other agonists, which are therefore less efficacious.

ION CHANNELS

Receptors can be linked directly with ion channels, which function only when the receptor is occupied by an agonist. This is the fastest type of receptor response, as, for example, when a neurotransmitter acts on a nerve ending. The excitatory neurotransmitters acetylcholine and glutamate cause a direct increase in both sodium and potassium permeability, which results in depolarization of the cell (see Ch. 5, p. 287). An equilibrium between open and closed ion channels exists and random fluctuations in conductance occur when

Drugs and receptors

A drug that acts on a receptor may act as an *agonist* or *antagonist*, depending on the response that is elicited. In eFigure 6-1A, drugs a and b are both agonists, but drug b is less potent. Drug c, on the other hand, acts as a partial agonist. Antagonists can be either competitive or non-competitive. If the antagonist is displaced by increasing concentrations of the drug (agonist), then competitive inhibition is present. Non-competitive antagonism describes the situation where the antagonist blocks the action of the agonist without competing with the receptor, which may not be overcome by increasing the concentration of the agonist, so shifting the curve to the right and depressing the maximal response (eFig. 6-1B).

Drugs often give a graded dose–response curve, where increasing the drug concentration will increase the drug effect. This graded response is seen with drugs that are not permanently bound to the receptor. Thus the *efficacy* of a drug is defined as the maximal response it can give, whereas *potency* describes the amount of drug required to give the desired response. Thus some drugs may be efficacious but not potent, requiring large doses to give an effect. However, if the drug is irreversibly bound to the receptor, its effect will continue well after the elimination of the drug from the bloodstream. Alternatively, the efficacy of a drug may diminish with time, an effect known as *tolerance*, which is thought to be the result of downregulation of specific drug receptors. Some drugs also produce active metabolites that continue to give a pharmacological effect well beyond the half-life of the parent drug. All these effects may complicate the study of pharmacokinetics.

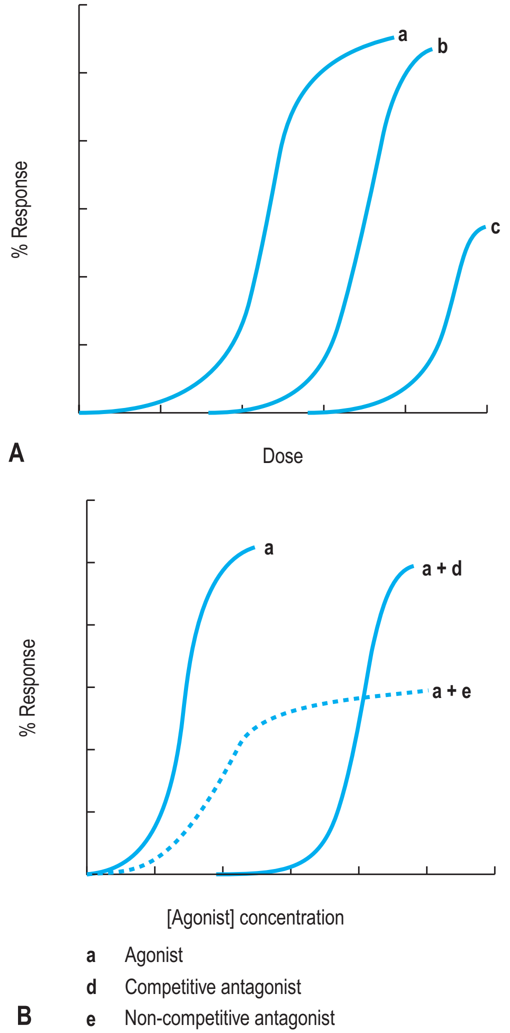

eFIGURE 6-1 (**A**) Potency of receptor stimulation by agonists and partial agonists. (**B**) The effects of agonist action by competitive and non-competitive antagonists.

the ion channels are opened (or closed) after receptor stimulation. The duration of this response can be measured and is referred to as the *mean channel lifetime*.

SECOND MESSENGERS

Receptor binding on extracellular receptors alters intracellular functions by activating a secondary messenger, for example the activation of the enzyme *adenyl cyclase* and calcium ion influx. Activation of these secondary messengers regulates various intracellular activities. In many cases they generate protein kinase activation and phosphorylation of membrane proteins (see Ch. 4, p. 235). The consequence of any change in protein conformation of the cell membrane may lead to opening of ion channels (e.g. sodium/potassium or calcium channels). Adenyl cyclase may also be activated by G proteins, which are discussed below.

Cyclic adenosine monophosphate (cAMP)

Cyclic AMP is a nucleotide synthesized within the cell from adenosine triphosphate (ATP) by the enzyme *adenyl cyclase*, and is inactivated enzymatically to 5′AMP by hydroxylation with a group of enzymes called *phosphodiesterases*. The following components are required to generate a secondary messenger: a receptor, which faces outwards from the cell; a regulatory protein (G protein), which faces inward towards the cytoplasm; adenyl cyclase; and a cAMP-dependent protein kinase. Inhibition of the secondary messenger activation can be induced with cholera toxin, which binds specifically to the G protein in the intestinal mucosa and thus prevents hydrolysis of another membrane nucleotide, guanine triphosphate (GTP), and adenyl cyclase (Box 6-6).

BOX 6-6 GPROTEINS REGULATE SECOND MESSENGER ACTIVITY

Gproteins are so called because of their affiliation with the guanine nucleotides GTP and GDP. They consist of three subunits, which catalyse GTP conversion to GDP. In the resting (empty receptor) state, the Gprotein–GDP complex is associated with the receptor. The presumed conformational change in the cell membrane when the receptor is occupied acts to increase the affinity of the neighbouring Gprotein for GTP and to promote the binding of the Gprotein–GTP complex to the effector site. This complex can activate ion channels, adenyl cyclase and other secondary messengers, as shown in the figure below.

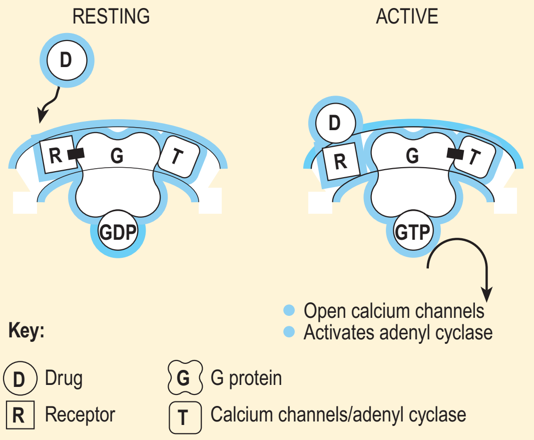

CALCIUM IS A MAJOR MEDIATOR OF CELLULAR ACTIVITY

Virtually all the calcium in the body is in the skeleton as hydroxyapatite. Intracellular calcium only constitutes 1% of total body calcium. The plasma concentration of calcium is 2.5 mmol/L; 50% is in its ionized form and 50% is bound to proteins or complexed with anions. A rise in the intracellular free calcium concentration can occur as a response to hormones and transmitters, by a net influx of calcium ions into the cell. It may also occur as a result of a release of sequestered calcium without a net influx into the cell. The intracellular calcium concentration is 10^{-7} mol/L, most of which is protein bound, and small changes in free calcium concentration may affect many intracellular processes. Intracellular calcium is normally regulated by an ATP-active transport channel. Accompanying the increase in intracellular calcium level, there is also an increase in the rate of degradation of minor membrane phospholipids (phosphatidylinositols) by phospholipase C, which in turn alters the membrane permeability and control of membrane phosphorylation, either directly, via its effect on protein kinases, or indirectly, through activation of adenyl cyclase. At the same time a phospholipid flip-flop mechanism occurs within the cell membrane,

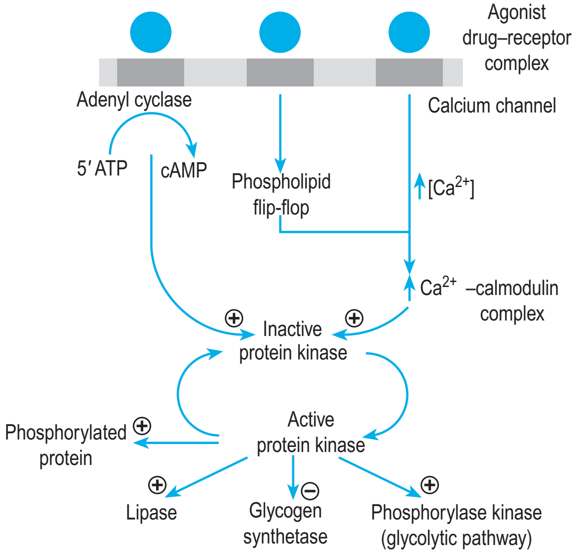

FIGURE 6-4 Activation of membrane-associated enzymes by calcium–calmodulin secondary messengers.

where methylation of phospholipids converting phosphatidylethanolamine to phosphatidylcholine takes place. This further increases calcium permeability, and secondarily regulates adenyl cyclase activation. Intracellular effects of calcium are controlled by an intracellular acidic protein, calmodulin. The calcium–calmodulin complex can thus activate many enzyme systems, including protein kinases, adenyl cyclase, phosphodiesterases and calcium-dependent ATPase (Fig. 6-4). The calcium–calmodulin complex is inactivated via the binding of trifluoperazine (an antipsychotic agent), which can be used to study the mechanism of action of certain therapeutic agents.

REGULATION OF PROTEIN SYNTHESIS

Steroid hormones, because they are lipophilic, can diffuse across the cell membrane and bind to cytoplasmic receptors and nuclear chromatin. Certain regions of the DNA sequence show a high affinity for steroid–receptor complexes, and binding of these complexes generates an increase in RNA polymerase activity and generation of messenger RNA (mRNA) and thus production of proteins (see Ch. 3, p. 133). One steroid may result in the production of several different species of mRNA within each cell, which may explain the great diversity of steroid actions.

Ligand–receptor affinity determines selectivity

The affinity of a ligand for its receptor is measured by the amount of ligand required to achieve half-maximal binding (EC_{50}). The binding force of the receptor–ligand complex can also be measured by the length of time taken for dissociation of the complex to occur. Low-affinity agonists include those required to potentiate a rapid response (e.g. neurotransmitters), and therefore concentrations of the ligand close to the receptor during stimulation must be high. The *selectivity* of a ligand for receptors is a ratio of the EC_{50} for the receptors being compared. If the ligand (or drug) is more selective for receptor A than for receptor B, it will achieve its clinical effect by stimulation of receptor A at lower doses, but will still stimulate receptor B if sufficient drug is given. Therefore, clinically, it is important to establish the EC_{50} ratio of a given drug; ratios lower than 50 are considered unsafe for clinical use if stimulation of other receptors is to be avoided.

Genetic control of drug handling influences response

Defining the genetic make-up within populations or, more pertinently, the individual has assisted in understanding responses to drugs. The hope is that with such understanding we may not only increase our knowledge of disease pathogenesis but also develop tailored therapies for the individual, whilst reducing adverse effects.

There are historical examples of clear distinctions in the individual's capability to metabolize drugs. This is seen with acetylation of drugs. N-acetylation may be either fast or slow, depending on the amount of enzyme present, which is controlled by a single recessive gene associated with low hepatic acetyltransferase activity (e.g. in isoniazid metabolism). Patients with low acetyltransferase activity are known as *slow acetylators*. Another genetic variation in the rate of drug metabolism is seen with suxamethonium (a depolarizing neuromuscular blocker used in general anaesthesia). About 1 in 3000 individuals fails to inactivate suxamethonium by hydrolysis (pseudocholinesterase), which is the result of a recessive gene that gives rise, in homozygotes, to an abnormal cholinesterase with a much lower substrate affinity. Patients who have a hereditary erythrocyte glucose-6-phosphate dehydrogenase enzyme deficiency may develop a haemolytic anaemia when treated with a number of drugs, including chloroquine, vitamin K, acetylsalicylic acid (aspirin) and probenecid.

***P450* gene.** The drug-metabolizing cytochrome P450 enzyme genes have been identified and cloned to include a large family of mono-oxygenases. Genes encoding P450 enzymes are polymorphic and within three families. As a result there are in broadly four functional outcomes: *fast* metabolizers who inherit multiple gene copies; *normal* metabolizers who have two normal alleles; *intermediate* metabolizers who have one deficient allele; and *poor* metabolizers who lack the functional enzyme.

The genes encoding families of enzymes are called *cyp* and are given numbers associated with a specific group of enzymes within the gene family, subfamily and then specific gene itself: e.g. *CYP27A1* relates to group 27, subfamily A and gene 1.

In the eye *CYP1B1* is a gene that encodes for an enzyme localized to the endoplasmic reticulum that is involved in metabolizing aromatic hydrocarbons and steroids. In addition to metabolism, within the eye mutations are linked with congenital glaucoma as well as rare but significant retinal dystrophies such as Bietti crystalline dystrophy. The aryl hydrocarbons have implications in ageing and cell disfunction secondary to smoking and cataract formation or age-related degeneration (see Ch. 9, p. 504) (Box 6-7).

BOX 6-7 GENETIC CONTROL OF IMMUNOSUPPRESSIVES DICTATES DRUGS RESPONSE

Azathioprine is a commonly used steroid-sparing immunosuppressant for inflammatory disorders, including for the eye in thyroid eye disease and uveitis. It is a purine analogue, whose active metabolite component, 6-mercaptopurine's conversion to cytotoxic 6-thioguanine analogues inhibits proliferation of cells. Most important is that it carries a significant potential for adverse effects. These have been reduced with improved understanding of pharmacogenetics. 6-mercaptopurine is metabolized via thiopurine methyltransferase (TPMT). Individuals (up to 1:1000 patients) who inherit two non-functional alleles for the polymorphic *TPMT* gene will have dramatic life-threatening myelosuppression. Importantly, up to 15% of the population are heterozygotes and will have increased incidence of myelosuppression. Genetic testing is now available as well as phenotyping by measuring TPMT activity in red blood cells.

Ocular pharmacology: drug handling by cells and tissues of the eye

This section describes the pharmacokinetics, pharmacodynamics and modes of drug delivery (excluding systemic administration) that are used to treat ocular surface and intraocular conditions.

MECHANISMS OF OCULAR DRUG ABSORPTION

There are several methods of administering ocular medications, including extraocular routes (topical) via either conjunctival/episcleral absorption (non-corneal) or transcorneal absorption, and direct intraocular administration of drugs. The non-corneal route of absorption may be significant for drugs that do not penetrate the cornea well. Corneal absorption still represents the major route of absorption for most ocular medication.

FACTORS INFLUENCING DELIVERY OF DRUGS TO THE EYE

Drugs can be administered in many different topical forms, including solutions, gels and ointments. The efficacy of treatment is usually dependent on intraocular penetration, which in turn is dependent on the permeability of the drug across the cornea, and the anatomical and physiological influences of the local environment, including lacrimation, tear drainage and the composition of the precorneal tear film.

Routes of administration

Conjunctiva. Topical administration into the inferior fornix of the conjunctiva is by far the most common route of ocular drug delivery. Both lacrimation and blinking profoundly influence the residence time of fluid in the fornix. Therefore, the efficacy of such delivery systems depends on the anatomy and physiology of the lids, the precorneal tear film, and the health of the conjunctiva, cornea and lacrimal system.

The conjunctival sac has a capacity of approximately 15–30 μL (dependent on blinking) and the natural tear film volume is 7–8 μL. The tears turn over at approximately 16% per minute during a normal blink rate of 15–20 blinks per minute. Most solution applicators deliver between 50 and 100 μL per drop, so a substantial amount of drug will be lost through overspill on administration. The turnover of tears is also highly dependent on environmental conditions, particularly temperature and humidity. The epithelium of the conjunctiva is continuous with that of the cornea and epidermis of the lids (see Ch. 1, p. 85) and contains goblet cells, which produce mucus and are integral to the stability of the tear film (see Ch. 4, p. 198). Drug absorption through the conjunctiva therefore requires transport firstly through the epithelium. In the subconjunctival stroma, which is a highly vascular conjunctiva owing to the rich superficial venous plexus and lid margin vessels, drugs may be absorbed in significant concentrations into the circulation. Also, after administration into the inferior fornix, drugs drain directly through the nasolacrimal duct into the nose, where measurable systemic absorption of drugs via the nasal and nasopharyngeal mucosa occurs. Restricting the entry of a topically applied ophthalmic dose into the nasal cavity by nasolacrimal occlusion for 5 min, or by making appropriate alterations to the vehicle (i.e. from solution to ointment) increases the residence time of the drug in the fornix, and increases ocular absorption.

Precorneal tear film and cornea

Tears are considered to act as a buffering system for many substances. The pH of normal tears varies between 6.5 and 7.6, while many drug delivery systems are often formulated at pH of less than 7; the return to physiological pH after drug instillation is, however, more likely to be a function of increased tear turnover than the result of a buffering effect. The precorneal tear film is composed of an outer lipid layer (mixed lipids), a middle aqueous layer (including proteins) and a deeper mucin layer (glycoprotein) (see Ch. 4, p. 200). The mucin layer contributes to the stability of the tear film, as well as promoting adherence of the tears to the lipophilic corneal and conjunctival epithelium. Any alteration in the components of the tear film will result in instability of the tear film and a reduced conjunctival residence time of the drug. At the same time alteration in the pH of the tear film may affect the ionization of the drug and thus its diffusion capacity. In spite of the extensive losses to the exterior and to systemic absorption, topical (conjunctival) administration of drugs achieves acceptable intraocular levels, mainly because of the very high concentrations that are administered.

The pH of the tear fluid is important in drug formulation because of the physiological homeostatic mechanisms. Drug penetration may be enhanced by changing the degree of drug ionization and enhancing the product stabilization over a range of pH changes. The epithelium of the cornea represents the most important barrier to intraocular transport of drugs via this route. First, the stratified cellular epithelium is bound by desmosomes between the lateral borders of the superficial cells. Second, the corneal epithelium is hydrophobic (as are all cell membranes), and so will allow only lipid-soluble drugs to pass through. In addition, Bowman's membrane, an acellular collagenous sheet (10 μm thick) between the basement membrane of the epithelium and the stroma of the cornea, acts as a further barrier to the penetration of drugs. In contrast, the stroma, which accounts for 90% of the corneal substance and its ground substance (glycosaminoglycans and water), permits ionized water-soluble drugs to pass more efficiently than lipid-soluble drugs. Finally, transport across the single-layer endothelium of the cornea is relatively free because it contains gap junctions that permit good penetration of most drugs into the aqueous humour. Many topical eye medications are weak bases, for example tropicamide, cyclopentolate and atropine, and exist in both ionized and un-ionized forms within the pH range of the tear film (pH 7.4). Altering the solution pH of timolol (pK_a 9.2) from 6.2 to 7.5 increases its corneal penetration and systemic absorption. The partition coefficient (ratio of concentrations in the two compartments) may therefore be increased by raising the pH of the water phase, rendering the drug non-ionized and more lipid-soluble. The factors affecting topical drug absorption are summarized in Box 6-8.

BOX 6-8 FACTORS INFLUENCING TOPICAL DRUG ABSORPTION

- Environmental conditions – temperature and humidity
- Volume of drug application
- Drug formulation – pH, preservative, vehicle type
- Blink rate
- Stability of tear film
- Absorption through conjunctival vessels and nasal mucosa
- Corneal epithelium and stroma
- Nasolacrimal drainage of tears

Delivery methods of ocular medication

RESIDENCE IN THE CONJUNCTIVAL SAC

The bioavailability of ocular medication depends on the precorneal fluid dynamics, drug binding to tear proteins, conjunctival drug absorption, systemic drug absorption, resistance to corneal penetration, drug binding to melanin and intraocular drug metabolism. Both the absorption and the efficacy of the drug can be increased by altering the formulation of the drug and/or by changing the local conditions.

To increase the residence time of the drug in the inferior fornix, and thus its delivery to the corneal epithelium, attempts at reducing the instilled volume of drug and increasing the viscosity of the solution have been made. For instance, polymers that increase solution viscosity include polyvinyl alcohol, hydroxypropylcellulose and other cellulose derivatives. However, increasing the viscosity of the delivery solution only produces a modest gain in ocular drug absorption, particularly for lipid-soluble drugs.

The residence time for many drugs is especially reduced by their pH, tonicity and the direct effect of certain drugs on lacrimation, which all affect the residence time of drugs in the lower fornix. For example, formulation at acidic or alkaline pH is irritant to the eye and increases both lacrimation and the blink rate, and clearance of the drug. Attempts to reduce the tonicity of the drug solution by using dilute buffers (e.g. phosphate buffers) to prevent stinging and lacrimation, and therefore increase transit time (reduce clearance), have also been employed. Drugs that have a direct pharmacological action on the lacrimal gland, increasing lacrimation and subsequently altering the precorneal fluid dynamics, include muscarinic agonists. Some drugs are also bound to the tear proteins (albumin, globulins and lysozyme), reducing the concentration of the free drug available for absorption.

The corneal epithelium presents a considerably greater barrier to hydrophilic than to lipophilic drugs (10 : 1). Corneal epithelial permeability, however, may be increased during ocular inflammation so that some drugs, for example dexamethasone, are more rapidly absorbed across the corneal epithelium into the eye. Preservatives such as benzalkonium chloride have also been shown to enhance the ocular absorption of drugs. Benzalkonium chloride and other cationic surfactants increase the ocular absorption of drugs by increasing corneal permeability (by compromising corneal integrity), depending on the molecular size and lipophilicity of the drug. Increased absorption can be obtained with pilocarpine, prednisolone and homatropine. For a drug to penetrate optimally, it must be able to exist in both ionized and un-ionized forms. Drugs will be buffered by the precorneal tear film and any alteration in the pH will change the ratio of ionized to un-ionized forms of the drug, dependent on the pK_a of the drug. In general a drug that exists in a purely ionized form will not penetrate the cornea unless the cornea has been damaged. Once absorbed into the eye, drugs may be bound to melanin within the pigment epithelium of the iris and the ciliary body, which may in turn reduce the bioavailability of the drug and also retard its clearance, leading to increased and prolonged drug levels. Similarly, after penetrating into the eye, drugs may be rendered inactive by intraocular metabolism. Enzymes that participate in ocular drug metabolism include those involved in inactivating neurotransmitters, for example monoamine oxidase, catechol *O*-methyltransferase, esterases, cytochrome P_{450}, and other enzymes including ketone transferase, glucuronidase and aldose reductase within the lens (Box 6-9). The majority of enzyme activity is microsomal, although some enzymes, for example esterases, are both cytosolic and extracellular. Within the anterior segment, the corneal endothelium, the non-pigmented cells of the iris, and the ciliary body are metabolically most active, so drugs that are good

BOX 6-9 ENZYMES INVOLVED IN DRUG METABOLISM IN THE EYE

- *Ketone reductase* – this is a cytosolic enzyme dependent on NADPH. It is thought to play a key role in the metabolism of timolol and analogues of propranolol. Found in corneal epithelium, lens, iris and ciliary body
- *Esterases* – important in the activation of ester pro-drugs. Both acetyl- and butyrylcholinesterases are found in rabbits. They are widely distributed throughout the anterior segment
- *Classic phase I and II oxidizing and conjugating enzymes* – cytochrome P_{450} reductase, demethylase, sulphatase and glucuronidase

substrates for these enzymes may suffer substantial degradation during absorption.

DRUG VEHICLES AFFECT DRUG DELIVERY

Several topical drug delivery systems are used in ophthalmology (Box 6-10) and there are a plethora of advances in the pipeline and in early-phase development such as nanoparticles and viral vectors for gene delivery. Nanoparticles are colloidal drug carrier systems in nano- and micro-size range. The advantages are overcoming lipid solubility, improved clinical targeting and depot longer-acting delivery. Similarly there is increasing development of viral delivered genes to replace mutated or loss of gene function in degenerative disease or as a drug depot to release the active component at the tissue site (see Ch. 3, p. 147)

Solutions

Solutions are a common mode of delivery because they cause less blurring of vision than ointments. They are easily administered and achieve high intraocular concentrations if applied regularly. They do, however, possess a short contact time and are quickly washed away at a rate proportional to the volume instilled. Polyvinyl alcohol or methylcellulose added to the solution increases the viscosity and/or lowers the surface tension, and will thus prolong contact time. Biologically active drug compounds that are sparingly soluble in water are often formulated as suspensions. Ophthalmic suspensions, particularly steroids, are thought to be useful delivery systems because it is assumed that the drug particles persist in the conjunctival sac and give rise to a sustained-release effect. Suspensions tend to form precipitates and thus need to be resuspended in the mixing bottle before application.

BOX 6-10 DRUG DELIVERY VEHICLES

- Solutions, colloids, emulsions, suspensions
- Ointments
- Slow-release preparations
- New ophthalmic delivery systems
- Particulates
- Liposomes
- Intravitreal preparations

Semisolids (ointments)

Ointments consist of any one or a combination of hydrocarbons, mineral oils, lanolin and polymers such as polyvinyl alcohol, carbopol and methylcellulose. Drugs applied by this method provide an increase in the duration of action because of reduced dilution, reduced drainage and prolonged corneal contact time. Although these preparations melt at the temperature of the ocular tissue and disperse within the tear film, they are still retained longer than other ophthalmic preparations. They do, however, give rise to blurring of the vision and an increased incidence of contact dermatitis, related commonly to the preservative within the preparation.

Slow-release preparations

The problem of short residence times has been addressed by the development of ingenious vehicle supports. For instance, controlled release of ocular medications can be achieved with conjunctival inserts or with hydrophilic soft contact lenses. Slow-release preparations allow the constant release of drug while minimizing the drainage rate of the drug.

Ocular inserts. Controlled-release delivery systems deliver a bioactive agent to the target site at a controlled concentration over a desired time course. Ocular inserts are flexible, elliptical devices, consisting of three layers. The two outer coats of ethylene vinyl acetate enclose an inner coat of drug/alginate mix. The Ocusert with pilocarpine relies on the solubility properties of pilocarpine-free base, which exhibits both hydrophilic and lipophilic properties. Because the drug is miscible in both aqueous and organic solvent media, it will permeate the hydrophobic controlling membranes by diffusion through the pores. Erodible systems such as Lacrisert® contain drug within carboxymethylcellulose wafers or polyvinyl alcohol disks or rods. They are manufactured principally for the treatment of dry eyes.

Collagen shields. Collagen is thought to be a suitable carrier of drugs. The three polypeptide α-helical chains are held together by cross-linking between proline and hydroxyproline, which account for 30% of the amino acid content of the molecule. Because of the ability to control the amount of cross-linking in the collagen subunits by exposure to ultraviolet light during manufacture, the time taken to dissolve when placed on the cornea can be altered. Also, collagen acts as an ion exchanger and is semipermeable, facilitating controlled release of drugs. Thus, the collagen bandage shields prolong contact between drug and cornea. Drugs can be incorporated into the collagen matrix, absorbed on to the shield during rehydration, or applied topically over a shield when in the eye. As the shield is erodible, release of the drug occurs gradually into the tear film, maintaining higher concentrations.

Soft contact lens. In this case the polymer of the contact lens is hydrophilic and thus water-soluble drugs are absorbed into the lens. The lens is hydrated once placed on to the cornea and so releases the drug until equilibrium is reached between drug concentration in the contact lens and in the conjunctival sac.

A primary concern with all ocular inserts is comfort, while many other criteria determine their usability (Box 6-11).

Intravitreal inserts. Development of inserts directly applied to the vitreous cavity has gained increasing impetus following the advent of successful trial evidence supporting intravitreal drug administration for macular degeneration and vascular occlusions and CMV viral retinitis (see p. 471). Such inserts include, for example the following:

Ozurdex® is a dexamethasone intravitreal implant inserted directly by injection into the vitreous cavity, using a Novadur® polymer system of poly D,L-lactide-co-glycoside (PLGA) which is biodegradable over months.

Retisert™ is a controlled delivery system of pellets containing the steroid fluocinolone coated in polyvinyl alcohol and silicon laminate which is directly inserted to the vitreous and sutured into the sclera for stability.

Ilivien® is a directly injectable-free floating non-biodegradable insert containing the steroid fluocinolone acetonide in a polymide cylindrical tube.

BOX 6-11 CRITERIA FOR SUCCESSFUL OCULAR INSERTS

- Comfortable and easy handling and insertion
- No expulsion of insert during wear
- No interference with vision
- Negligible interference with oxygen access to cornea
- Reproducibility of kinetics of drug release
- Sterility

ADVANCED OCULAR DELIVERY SYSTEMS

New ophthalmic delivery system (NODS)

NODS is a method of administering a drug as a single unit volume within a water-soluble preservative-free form. The system is easily administered and may offer a significant improvement in bioavailability over drops. Essentially the device consists of a water-insoluble drug-loaded flange attached to the end of a water-soluble handle by a soluble membrane film. The flange end is placed into the inferior fornix and the membrane rapidly dissolves, releasing the flange, which then hydrates within the inferior fornix and releases the drug.

Particulates

Microspheres and nanoparticles represent promising particulate polymeric drug delivery systems for ophthalmic medications. These systems may avoid the potential disadvantages of other delivery systems, which include discomfort and difficulty of use with inserts, blurring of vision with viscous solutions, and instability of liposomes (see below). They may not represent real particles (capsular wall containing an aqueous or solid core) but are matrix-type structures of lipid base and drug. The particles are formed by polymerization (with ultraviolet light), during or after which the drug may be added. This leads to covalent drug binding to the polymer, so that when the drugs are absorbed into the polymer matrix they form a solid–solution matrix. The binding of the drug depends on its physicochemical properties as well as the nature of the polymer (polybutylcyanoacrylate). Smaller particles are better tolerated by patients than larger ones, and also have increased drug absorption and much slower elimination rates.

Liposomes

Liposomes are vesicles composed of lipid membranes enclosing an aqueous volume. They form spontaneously when a mixture of phospholipids is agitated in an aqueous medium to disperse the two phases, and are composed of phospholipids, including lecithin, phosphatidylserine and phosphatidylglycerol. They therefore share the properties of the bilayer of an outer cell membrane. The drugs can be trapped in either the lipid or aqueous phase. They provide the possibility of controlled and selective drug delivery and thus increased ocular bioavailability, although liposomes carry lipophilic drugs more readily than hydrophilic ones. The advantage of liposomes is that they are easily prepared, non-irritant, and do not cause any blurring of vision. Altering their surface charge or binding specific ligands to them increases their adherence to cells and subsequent endocytosis of the liposome–drug complex.

Ocular iontophoresis

Iontophoresis is a method of drug delivery that utilizes an electric current to drive a polar drug across a semipermeable membrane; this may be achieved by either a cathode or an anode, depending on whether the molecule is negatively or positively charged. Applications in clinical ophthalmology have yet to be determined but a possibility may be the iontophoretic application of antibiotics in bacterial keratitis, enhancing tissue penetration of the agent. Corneal iontophoresis of gentamicin and aprofloxacin has been successful in treating experimental pseudomonas keratitis. Trans-scleral iontophoresis allows direct drug penetration into the vitreous but has a major disadvantage of discomfort and may produce small areas of retinal necrosis at sites of application.

INTRACAMERAL AND INTRAVITREAL ADMINISTRATION

Despite the considerable ingenuity applied to the development of topical drug preparations, the treatment of many ocular disorders is hampered by poor penetration into the eye. Systemic drug administration also does not guarantee high intraocular drug levels, in part as a result of the integrity of the blood–retinal and blood–aqueous barriers. For instance, the treatment of bacterial endophthalmitis is often inadequate unless vitrectomy and intravitreal antibiotics are used. The recommended dose of intravitreal antibiotics is based on doses that are not toxic to the rabbit or primate retina, but little is known about dose–response in the human vitreous. Experimentally, intravenous administration of the newer cephalosporins produces an incremental increase in the vitreous levels in inflamed eyes (but not in normal eyes). In inflammatory conditions, water-soluble antibiotics (penicillins) have a prolonged half-life as the retinal pump mechanism is damaged in such eyes. However, the commonly used aminoglycosides are more rapidly cleared from inflamed eyes, despite low lipid solubility, because they are eliminated through the aqueous circulation. To achieve continual maximal therapeutic concentrations in the vitreous, repeated injections may be required. This may lead to complications of lens injury, haemorrhage, infection and retinal injury, and methods to reduce the toxicity of this form of treatment have included development of liposome-encapsulated drugs, which may be administered directly into the vitreous, or subcutaneously/subconjunctivally, acting as controlled-release formulations. Similar approaches are being sought both in humans and experimental models with other agents, for example ganciclovir for cytomegalovirus retinitis (see eBox 6-1), ciclosporin A for endogenous uveitis and encapsulated antifibroblastic drugs (5-fluorouracil) for proliferative vitreoretinopathy.

Intravitreal delivery of drugs has revolutionized care for vascular retinal disorders

Understanding that in neovascular (wet) age-related macular degeneration creating choroidal neovascular membranes or diabetic macular oedema is in part driven by pro-angiogenic growth factor, vascular endothelial growth factor (VEGF) has led to development and implementation of anti-VEGF therapies in clinical practice. There are numerous intravitreal injections of drug available to neutralize VEGF activity (Fig. 6-5). In common, the VEGF inhibitors act to bind to soluble VEGF and block activity by preventing binding to VEGF receptors. VEGF actions are mediated principally via receptors VEGF-R1 and VEGF-R2. *Ranibizumab* is a monoclonal recombinant human Fab VEGF; recombinant human FabV2 is a humanized Fab fragment that binds and neutralizes VEGF and *aptamers* (i.e. *pegaptanib sodium*) are nucleic acid ligands that are specific for given proteins and act as decoys, neutralizing their function. Another decoy mechanism is that of high-affinity VEGF blocker that

eBox 6-1

Treatment of cytomegalovirus (CMV) retinitis

Ganciclovir was the first agent found to demonstrate therapeutic activity against CMV retinitis (see Ch. 8, p. 473). At least 80% of patients respond to initial treatment with parenteral ganciclovir. Complications following parenteral administration include thrombocytopenia and leucopenia.

Zidovudine (AZT) has no direct antiviral action against CMV but may act by diminishing human immunodeficiency virus (HIV) enhancement of CMV infection and by improving immune function. However, AZT also causes bone marrow suppression, making combined therapy incompatible. Human recombinant granulocyte–macrophage colony-stimulating factor stimulates proliferation, differentiation and chemotaxis of neutrophils and, in combination with anti-CMV and HIV treatment, may prevent neutropenic episodes.

Cidofovir is a DNA polymerase chain inhibitor that can be used at maintenance dose once a week.

Foscarnet may also be used to treat CMV infection, but up to 30% of patients develop renal toxicity. Despite this, it is safer for administration in combination with AZT.

Control of CMV retinitis may occur with intravitreal injections of either ganciclovir or foscarnet. Injections are not directly retinotoxic even though intravitreal levels of these agents are within the toxic range for the drug. Intravitreal injections of liposome-encapsulated drugs reduce the requirement for repeated injection, without any evidence of retinal toxicity.

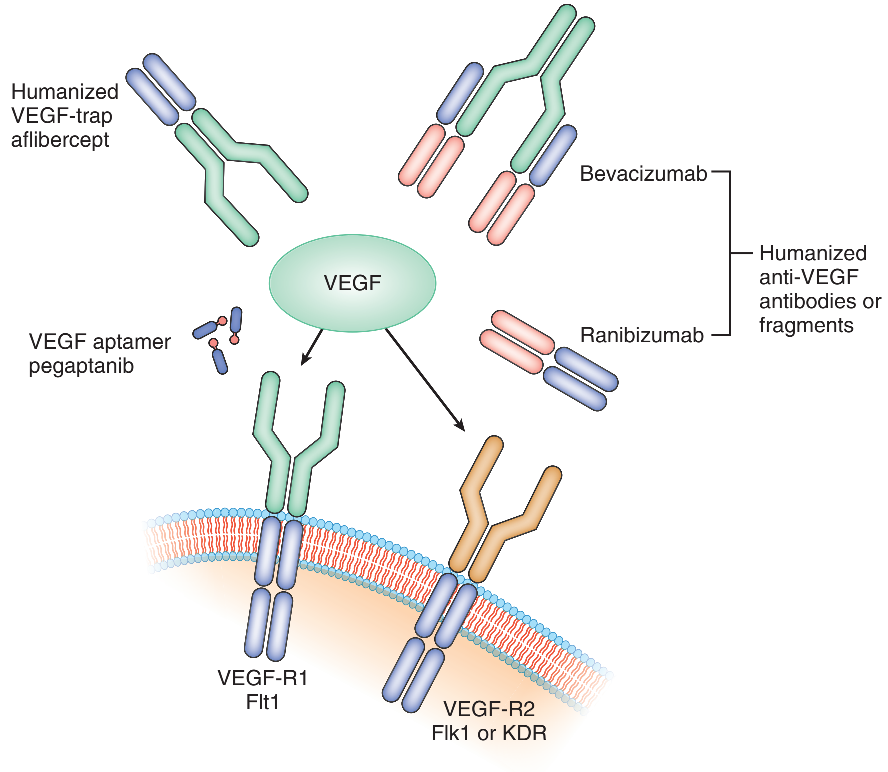

FIGURE 6-5 Intravitreal anti-VEGF therapy. Advent of current biologics to inhibit VEGF activity are shown. They target one of the VEGF family of proteins – VEGF A (there are VEGF A–D). VEGF binds to numerous receptors, including VEGF-R1-R3 and neuropilin receptor. However, the greater receptor affinity for major pro-angiogenic VEGF A actions is via VEGF-R1 and VEGF-R2, as depicted in this schematic.

is a fusion of VEGF receptor 1 to a human immunoglobulin G constant region (aflibercept) (Box 6-12).

DRUGS ADMINISTERED SYSTEMICALLY ALSO PENETRATE THE EYE

The preceding sections focus on the factors that affect the efficacy of topical or direct ophthalmic therapy with minimal systemic side-effects. However, drugs for ocular conditions may be administered systemically to achieve sufficient drug concentration in ocular tissue, although in humans the intraocular drug levels that are reached are largely unknown for most substances.

Commonly used drugs in ophthalmology include the carbonic anhydrase inhibitors (acetazolamide and dichlorphenamide, see below), which are

BOX 6-12 VEGF

VEGF belongs to a platelet-derived growth factor family which includes VEGF A–D.

VEGF A is a product of the *vegfA* gene encoding multiple isoforms identified by their amino acid number (e.g. $VEGF_{165}$).

Alternative splicing may result in two protein families from the same gene, which for the *vegfA* gene, for example, are denoted $VEGF_{165}$ or $VEGF_{165}$b. The latter is anti-angiogenic and is important for tissue homeostasis.

There are a number of VEGF receptors, which include the major pro-angiogenic receptors VEGF-R1 (Flt1), VEGF-R2 (KDR) mediating via tyrosine kinase's VEGF A effect.

VEGF-R3 (Flt4) mediates lymphangiogenesis through VEGF C and D, while the fourth receptor, neuropilin receptor (NRP1 receptor), which binds VEGF A, is important for cell survival and axonal guidance.

administered orally or intravenously to reduce intraocular pressure. Various animal studies have also demonstrated the ability of systemic antibiotics to reach intraocular infections in concentrations that are bactericidal to certain pathogens. For example, ciprofloxacin penetrates the aqueous humour following oral administration. Similarly, both non-steroidal anti-inflammatory drugs and steroids penetrate the eye when given orally.

Conversely, drugs applied topically may also reach the systemic circulation and affect the contralateral eye. This has been recorded with timolol therapy of chronic open-angled glaucoma with unilateral intraocular pressure rise, which resulted in a significant drop in intraocular pressure in both treated and untreated eyes. In experimental models, a drug effect on the contralateral eye has also been shown with apraclonidine. This phenomenon may occur with a drug possessing a long systemic half-life or unique tissue-binding characteristics within the eye. If supplied in too high a dose or concentration, systemic side-effects may occur, as with 10% phenylephrine drops.

TOPICAL MEDICATIONS AND PRESERVATIVES

Ophthalmic solutions and ointments must be sterile, and a wide variety of preservatives are used for this purpose. Most are toxic to the precorneal tear film and epithelium, impeding epithelial healing and disrupting the tear film. The commonly used preservatives are benzalkonium chloride, thiomersal, chlorbutol and organomercuric compounds. Benzalkonium chloride is a surfactant preservative (cationic preservative) which attains its bactericidal activity by attaching to the bacterial cell wall, increasing permeability and eventually rupturing the cell wall. While these preservatives and other cationic surfactants compromise the corneal integrity, they have also been shown to enhance the ocular absorption of drugs. Benzalkonium chloride is most effective at an alkaline pH (approximately pH 8.0) but is inactivated by the presence of soaps and salts, for example magnesium and calcium. As such, some contact lens solutions also combine ethylenediaminetetra-acetic acid (EDTA; a chelating agent) to overcome this problem. Severe toxicity may result from direct cellular damage or from a hypersensitivity reaction to components of the drug, and give rise to papillary conjunctivitis, punctate keratitis and corneal oedema. Chlorbutol also reduces oxygen utilization of the cornea (see Ch. 4, p. 203) and may result in epithelial desquamation. Mercurial compounds include phenylmercuric acetate and thiomersal. Hypersensitivity to these compounds is the most dramatic and common complication of preservatives in 10% of patients, and mercurial deposits may develop in corneal tissues. All these preservatives are absorbed to various degrees by soft contact lenses.

Reconstituting the tear film

With increased knowledge of the anatomy and physiology of the precorneal tear film (Ch. 1, p. 89 and Ch. 4, p. 203), tear substitutes have been generated to provide symptomatic relief by artificially reconstituting individual tear film components. Ocular surface disease results from abnormalities in one, but generally more, of the tear film components. *Aqueous* deficiency is observed classically in keratoconjunctivitis sicca, as occurs in, for example, Sjögren's syndrome. Instillation of artificial tears is the mainstay of treatment in conjunction with reducing drainage by occluding the lacrimal puncta. *Mucin* deficiency occurs in conditions that affect goblet cell function, for example cicatricial conjunctival disease and hypovitaminosis A, the effect of which is an unstable tear film because the surface tension is reduced and so is insufficient to maintain the aqueous layer. Therefore, areas of non-wetting occur and if left untreated corneal ulceration and scarring can result. Finally, *lipid* abnormalities are associated with chronic inflammation of the meibomian glands (see Ch. 1, p. 82) and, again, result in an unstable tear film and inadequate wetting of the cornea.

TEAR SUBSTITUTES

Artificial tears or ocular lubricants are generally formulated as solutions consisting of inorganic ions (0.9% NaCl) and polymers to increase wettability and retention time, i.e. they additionally act to replace one of the actions of the mucin component of the tears. Commonly used polymers include polyvinyl alcohol and semisynthetic celluloses. Methylcellulose, hydroxypropylmethylcellulose and hydroxycellulose are still widely used, although polyvinyl alcohols have additional surfactant properties to further stabilize the tear film for longer periods. Recently, hyaluronic acid has been used, which has much greater retention times than celluloses or polyvinyl alcohols and improves tear film stability. Polyacrylic acid (carbomer) is a gel that is hydrophilic, helping it to form a stable tear film, increasing retention time and thus enabling it to be used less frequently.

Artificial tears are now more commonly available without preservatives. Preservatives affect corneal epithelial stability and thus ability to maintain the precorneal tear film.

MUCOLYTICS

Acetylcysteine, which is a derivative of the amino acid L-cysteine, has been used as a mucolytic agent, dissolving the mucus threads that occur in keratoconjunctivitis sicca, but it does not appear to effect corneal wetting so it has to be used in conjunction with other artificial tear preparations.

Ocular drugs and the autonomic nervous system

Parasympathetic and sympathetic divisions of the autonomic nervous system supply both ocular and extraocular tissues vital for normal ocular function (see Ch. 1, p. 64). Agents that influence neurotransmission in the autonomic system are extremely important in the diagnosis and management of many ophthalmic disorders.

THE PARASYMPATHETIC SYSTEM

Acetylcholine is the major neurotransmitter of the parasympathetic system. It is formed enzymatically from choline and acetyl coenzyme A in the nerve endings (Fig. 6-6). Acetylcholine acts on two types of receptors: muscarinic, which are situated at the effector organ (postganglionic), and nicotinic, which are situated at ganglion synapses and also at neuromuscular junctions. In the eye such receptors are found in the motor endplates of extraocular muscles and levator palpebrae superioris, sympathetic and parasympathetic ganglia, iris sphincter, ciliary body and lacrimal glands. The nicotinic–acetylcholine receptor complex consists of two pairs of polypeptide chains and an additional single polypeptide chain (250 000 Da), providing a hydrophilic channel through which ions can traverse the lipid bilayer. This is known as a *ligand-gated ion channel*. As the acetylcholine binds to the receptor in the synaptic cleft the configuration of the polypeptide chains alters, allowing an influx of sodium and potassium ions down the concentration gradient, thus depolarizing the motor endplate. Cobratoxin, bungarotoxin and tubocurarine (active ingredient of curare) block acetylcholine receptors by preventing the opening of these ion channels. Following the release of acetylcholine, it is largely metabolized by hydrolysis, either within the synapses by acetylcholinesterase, or it is absorbed into plasma and then hydrolysed by butyrylcholinesterase. Drugs can act as agonists or antagonists at either muscarinic or nicotinic receptors.

Parasympathomimetics mimic the action of acetylcholine

Parasympathomimetics are a group of drugs that act either by directly stimulating the muscarinic receptor, for example pilocarpine, or by inhibiting the enzyme acetylcholinesterase, which hydrolyses the acetylcholine in the synapse.

Pilocarpine is used in the treatment of chronic open-angle glaucoma, facilitating aqueous drainage via its miotic action on the iris and contraction of the longitudinal muscle of the ciliary body. This draws on the scleral spur and opens the uveotrabecular meshwork. Parasympathetic stimulation has been shown experimentally to reduce aqueous outflow resistance with concomitant breakdown of the blood–aqueous barrier. Pilocarpine thus not only reduces outflow resistance but also reduces the rate of aqueous secretion. In addition, pilocarpine blocks the uveoscleral drainage route of aqueous (see Ch. 4, p. 35) so that drainage is confined to the canal of Schlemm. The ocular hypotensive effect of pilocarpine is further achieved by a reduction in blood flow within the ciliary body, which in turn reduces aqueous secretion (Box 6-13).

The side-effects of direct muscarinic agonists include conjunctival toxicity, iris cysts, cataracts and systemic absorption, which can give rise to muscarinic

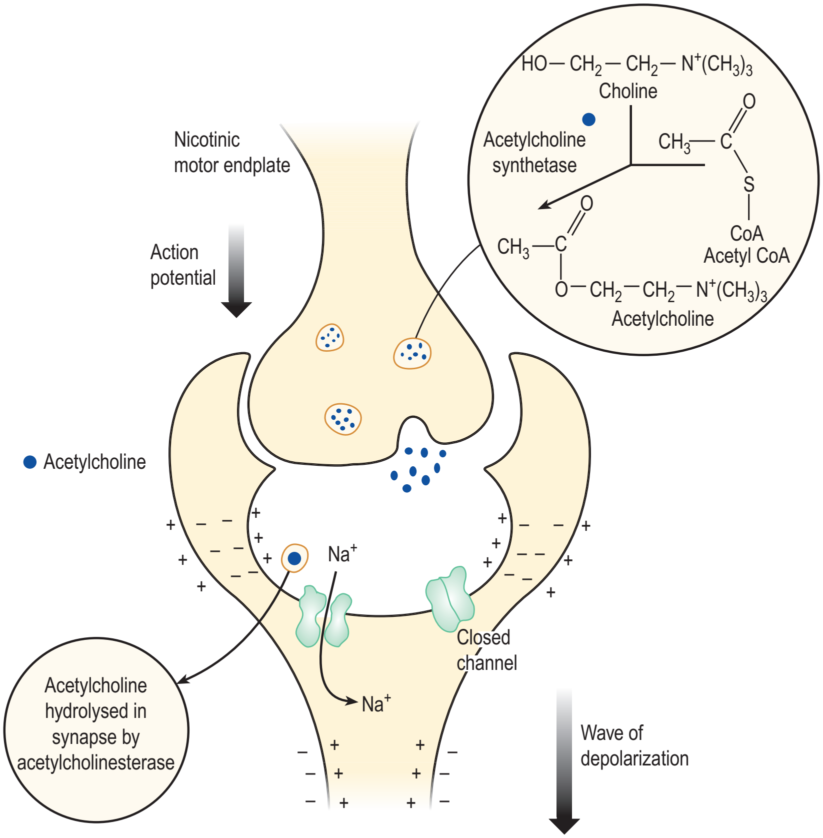
Nicotinic
motor endplate
Action
potential
● Acetylcholine
$HO-CH_2-CH_2-N^+(CH_3)_3$
Choline
Acetylcholine
synthetase
$CH_3-C(=O)-S-CoA$
Acetyl CoA
$CH_3-C(=O)-O-CH_2-CH_2-N^+(CH_3)_3$
Acetylcholine
Na^+
Na^+
Closed
channel
Acetylcholine
hydrolysed in
synapse by
acetylcholinesterase
Wave of
depolarization

FIGURE 6-6 The parasympathetic system: motor endplate.

stimulation resulting in sweating, salivation, vomiting and bradycardia.

Parasympathetic antagonists affect the pupil and ciliary muscle separately and are exploited therapeutically

These drugs block the effect of acetylcholine at muscarinic receptor sites. Their principal effects are those of mydriasis (dilatation of pupil) and cycloplegia (inhibiting ciliary body muscle contraction), but additional effects also include reduction of lacrimal secretions (Table 6-1).

They are used clinically in the management of iritis (preventing adherence of the iris to the anterior capsule of the lens called posterior synechiae), for cycloplegic refraction, to facilitate routine fundal

BOX 6-13 USING PARASYMPATHOMIMETICS IN DIAGNOSIS

Edrophonium (a competitive inhibitor of acetylcholinesterase) is used in the diagnosis of ocular or systemic myasthenia gravis (tensilon test). After intravenous administration, any improvement in ptosis or diplopia confirms the diagnosis. Longer-acting anticholinesterases (neostigmine) can be used to maintain some neuromuscular function. Immunosuppressive therapies are also used to treat myasthenia because they impair the autoimmune response generated by the acetylcholine receptor antibodies.

Low concentrations of pilocarpine are used in the diagnosis of Adie's tonic pupil, by manifesting the miotic response of the supersensitive iris sphincter caused by loss of postganglionic nerve fibres in this condition.

examination and for the provocation test in narrow-angle glaucoma. The side-effects of this group of drugs are seen more often with the longer-acting drugs, for example atropine sulphate. Systemic absorption occurs, particularly with atropine, via conjunctival vessels and via nasal and nasopharyngeal mucosal absorption. Systemic side-effects include dry mouth, facial flushing, sweating and tachycardia. Ocular side-effects include hypersensitivity reactions and conjunctival hyperaemia, blurring of vision and photophobia, a transient rise in intraocular pressure and, occasionally, precipitation of acute-angled glaucoma.

TABLE 6-1 **Antimuscarinic agents**

Drug	**Maximal mydriasis/ cycloplegia (min)**	**Duration**
Atropine sulphate	40	7–10 days
Homatropine	40	1–2 days
Cyclopentolate	30	12–24 h
Tropicamide	20	3–4 h

THE SYMPATHETIC SYSTEM

Acetylcholine, adrenaline (epinephrine) and noradrenaline (norepinephrine) all act as neurotransmitters within the sympathetic autonomic nervous system (Fig. 6-7). Acetylcholine is the neurotransmitter at autonomic ganglia, which includes the sympathetic nervous system. Adrenaline and noradrenaline are generated from hydroxylation of tyrosine to form DOPA, and further modified enzymatically to form dopamine, and finally adrenaline and noradrenaline (Fig. 6-8). Once released from the nerve ending they act on several well-defined subtypes of receptors. The action of the neurotransmitter is terminated by enzyme degradation with monoamine oxidases and catechol *O*-methyltransferase, and also by active reuptake into the nerve endings (Fig. 6-9). Understanding of sympathetic outflow facilitates the diagnosis of Horner syndrome.

Additional content available at https://expertconsult.inkling.com/.

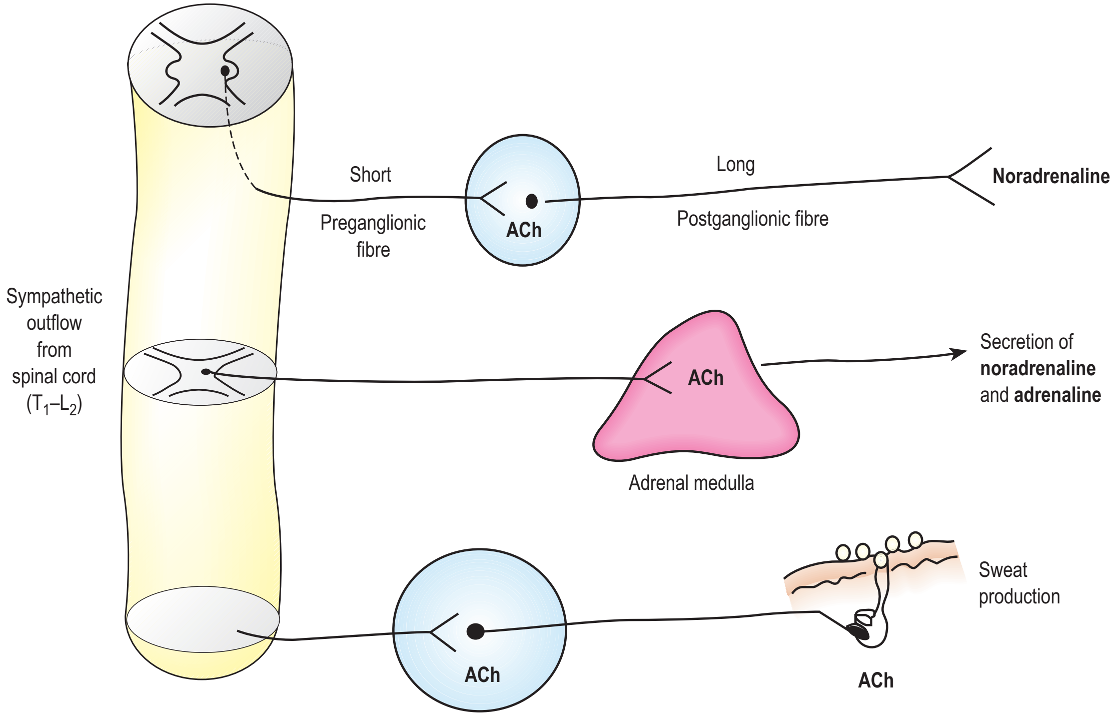
Short
Preganglionic fibre
ACh
Long
Postganglionic fibre
Noradrenaline
Sympathetic outflow from spinal cord $(T_1–L_2)$
ACh
Secretion of noradrenaline and adrenaline
Adrenal medulla
Sweat production
ACh
ACh

FIGURE 6-7 The sympathetic nervous system. ACh, acetylcholine.

Horner syndrome

Horner syndrome is important to recognize given that is can indicate life-threatening disease, for example Pancoast syndrome (apical lung tumour) or dissecting aortic aneurysm. The signs are a result of interference with sympathetic nervous system and anatomical diagnosis of location of pathology can be assisted with response to pharmacological block of sympathetic outflow.

CLASSICAL TRIAD OF SYMPTOMS

- Partial ptosis (loss of sympathetic innervation of Müller's muscle)
- Anhydrosis (decreased sweating affecting side of face)
- Miosis (loss of sympathetic tone to iris).

MAKING A DIAGNOSIS

Even with the advent of medical imaging to precisely locate the lesion within the sympathetic outflow pathway, diagnosis may be confirmed and then localized via pharmacological manipulation. The site depends on where the sympathetic outflow is perturbed.

- First order neurones – central lesions involving the medulla oblongata where autonomic centres reside (e.g. as seen in lateral medullary syndrome)
- Second order neurones – where preganglionic fibres are involved with compression of sympathetic chain (Pancoast syndrome)
- Third-order neurones – where postganglionic lesions at level of internal carotid artery (and within cavernous sinus) are affected.

THE USE OF DROPS

- Cocaine prevents the reuptake of noradrenaline, resulting in pupil dilatation. So in Horner syndrome with the loss of sympathetic drive, there is a failure of cocaine-induced pupil dilatation.
- Hydroxyamphetamine 1% will cause neurotransmitter release from the postganglionic fibres (third-order neurones). Therefore, in Horner syndrome affecting the preganglionic fibres, amphetamine will cause normal pupil dilatation, whereas if there is disruption of postganglionic fibres there will be no effect as there will be no noradrenaline to release.
- Both apraclonidine and phenylephrine drops will dilate postganglionic Horner syndrome due to denervation hypersensitivity.

The control of intraocular pressure is facilitated by similar and opposite effects of adrenergic receptors (Box 6-14)

The control of intraocular pressure (IOP) and aqueous production and outflow has already been discussed (see Ch. 4, p. 222). It is well documented that certain α-agonists cause lowering of the IOP, for example clonidine, which is principally mediated by an α_2-agonist action. On the other hand α_1-agonists (e.g. phenylephrine) cause a rise in IOP with concomitant pupillary dilatation. As α_2-agonists produce an immediate fall in IOP, their effect is likely to be central, mediated by stimulation of the medullopontine sympathetic centre. Conversely, α_1-hypertensive effects are thought to be muscular in origin because the effect is antagonized by muscle relaxants.

β-receptors are the predominant receptor on ciliary epithelium, stimulation of which will reduce IOP experimentally. However, β-blockers (e.g. timolol) also reduce IOP. In humans, stimulation of β-receptors does not cause a reduction in aqueous secretion. Indeed adrenaline (a combined α- and β-agonist) causes an initial rise in IOP that is also associated with an initial rise in aqueous secretion. This effect is mediated by concomitant α_1-receptor stimulation, the hypotensive effect being a purely α_2-agonist action. Experimental evidence suggests that direct application of cAMP also increases aqueous drainage, and adenylate cyclase, which is also activated by β_2-agonists, produces a similar response. Certainly a rise in aqueous cAMP levels correlates well with reduction of IOP, which may be mediated by both α- and β-agonists, but as yet the exact mechanism is unknown (Box 6-15).

Additional content available at https://expertconsult.inkling.com/.

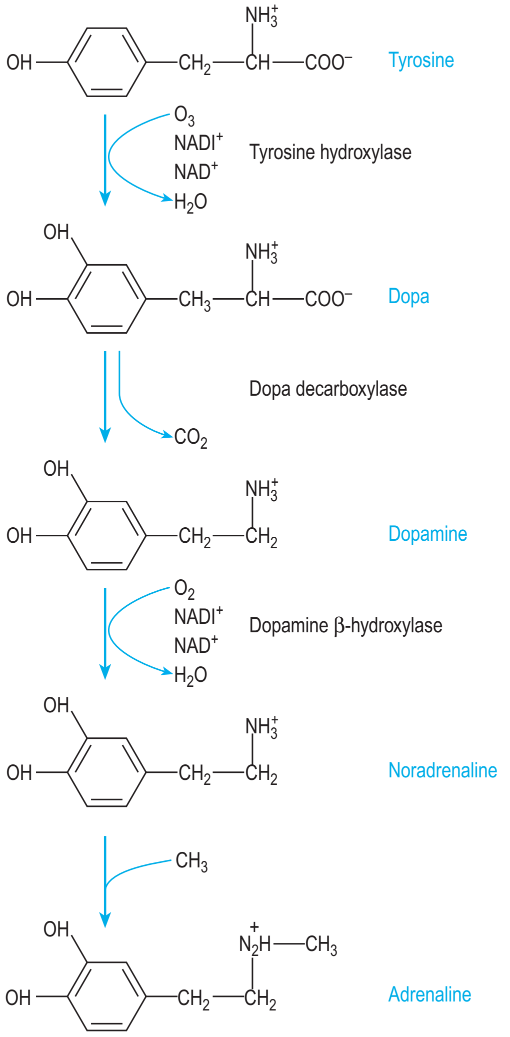

FIGURE 6-8 Synthesis of noradrenaline (norepinephrine) and adrenaline (epinephrine).

BOX 6-14 ADRENERGIC RECEPTORS

Adrenergic receptors are classes of G-coupled protein receptors of two types, α and β; further subtypes exist within each group, classified according to the elicited response and distinguished by selective agonists and antagonists.

- α_1-receptors mediate excitatory responses: principally smooth muscle contraction.
- α_2-receptors are located mostly on the presynaptic nerve endings and are inhibitory because, when stimulated, they prevent further release of the neurotransmitter from the presynaptic terminal (Fig. 6-9).
- β_1-receptors situated in the heart are excitatory and give rise to a positive inotropic and chronotropic response.
- β_2-receptors can mediate inhibitory responses, relaxing the smooth muscle of blood vessels and bronchi.
- β_3-receptors are located in adipose tissue and related to thermogenesis and lipolysis.

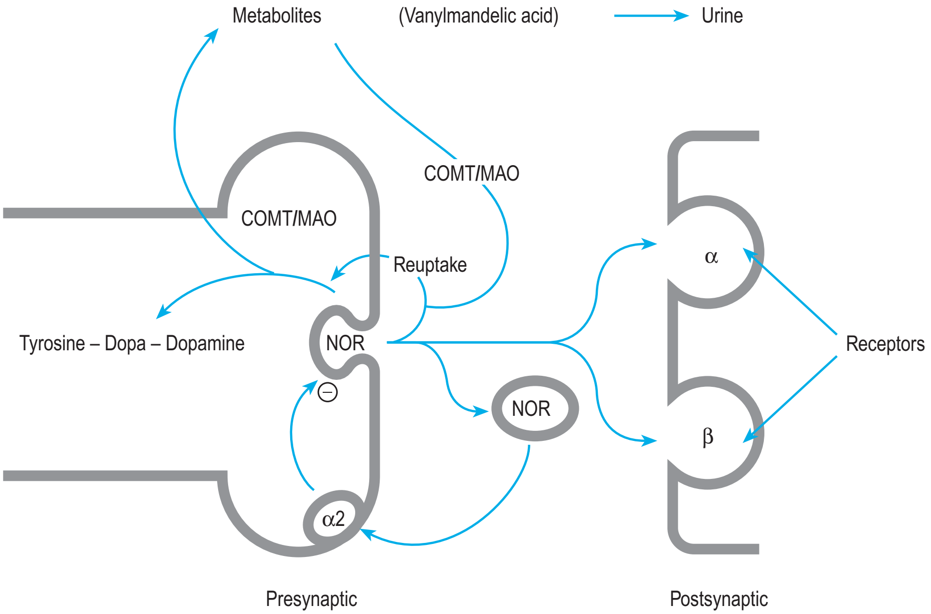

COMT	Catechol *O*-methyltransferase
MAO	Monoamine oxidase
NOR	Noradrenaline

FIGURE 6-9 Adrenergic receptors and metabolism of adrenergic neurotransmitters.

BOX 6-15 ADRENERGIC ANTAGONISTS

NON-SELECTIVE β-BLOCKERS (β_1 AND β_2)

- Carteolol
- Levobunolol
- Timolol

SELECTIVE β_1-BLOCKERS

- Betaxolol

Clinical control of intraocular pressure exploits the adrenergic system

Control of IOP is the main aim of therapy in primary open-angle glaucoma, a disorder that contributes greatly to the demands on eye health care systems. IOP control may be achieved surgically or medically and numerous topical therapies based on the pharmacology of the autonomic nervous system are in use, as described above.

However, medical treatment of glaucoma with respect to control of IOP may also be achieved by systemic or local inhibition of carbonic anhydrase, an enzyme central to the formation of aqueous (see Ch. 4, p. 222), enhancing outflow pathways of aqueous flow or protecting the neurones (Box 6-16). The common drugs used topically are summarized in Box 6-17.

The ocular effects of adrenergic agonists include mydriasis, slight ciliary muscle relaxation (probably not relevant in humans), increased formation and increased outflow of aqueous humour (although a stimulation will decrease aqueous humour production), contraction of Müller's muscle and constriction of conjunctival and episcleral vessels (Table 6-2).

Additional content available at https://expertconsult.inkling.com/.

The histaminergic system: histamine is released from conjunctival mast cells during allergic reactions

Histamine is synthesized and stored in most tissues. It is derived from the amino acid histidine and excreted

ADRENERGIC AGONISTS

- *Adrenaline* stimulates both α- and β-receptors. Adrenergic receptors are found in cell membranes of iris dilator muscle, Müller's muscle, ciliary process epithelium and smooth muscle of ocular blood vessels. Phenylephrine is a synthetic sympathomimetic that acts directly on α-receptors. Mydriasis occurs within 1 hour and lasts 4–6 hours. The mydriasis of phenylephrine can be overcome by the powerful parasympathetic light reflex during fundoscopy. Thus, for maintained pupillary dilatation, a combination of phenylephrine and an antimuscarinic, for example cyclopentolate, is used. Systemic absorption is common with phenylephrine and care must be taken when using it in the elderly, in patients with hypertension, or in neonates and children.
- *Hydroxyamphetamine* acts by releasing noradrenaline from the nerve terminals. It is therefore used in the diagnosis of postganglionic Horner syndrome.
- *Cocaine*, whose membrane-stabilizing effect gives rise to its local anaesthetic properties, also prevents the reuptake of noradrenaline into the nerve endings, and is used in the diagnosis of pre- or postganglionic Horner syndrome in cases of anisocoria.
- *Apraclonidine hydrochloride* is a selective α-adrenergic agonist acting mainly on α_2-receptors. It may be applied topically for the prevention and management of raised IOP after anterior segment laser treatment. It lowers IOP by decreasing aqueous humour formation. It does not penetrate the blood–aqueous barrier easily and therefore has minimal systemic side-effects, unlike clonidine.
- *Brimonidine* is a selective α_2-receptor antagonist which over longer periods can reduce IOP without significant cardiovascular or pulmonary function effects.
- *β_2-adrenergic agonists* also lower IOP, but their effect is mainly to increase the uveoscleral outflow, and they are often used in combination with β-blocking agents.

Commercially available L-adrenaline (Eppy 1%) can produce local irritation, allergy and adrenochrome deposits in the subconjunctiva. These adrenochrome deposits are most likely to be oxidative products of adrenaline and are most common in the palpebral conjunctiva. Dipivalyl adrenaline (Propine® 0.1%) is also used in the treatment of glaucoma. It is a lipophilic prodrug of adrenaline, and can penetrate into the anterior chamber, where it is hydrolysed into the active agent. If adrenaline or its prodrug analogues are used to treat aphakic glaucoma, there is a high incidence of cystoid macular oedema. This macular toxicity has been reported to occur in 20–30% of aphakic eyes undergoing adrenaline treatment. The mechanism of cystoid macular oedema is unknown, but the angiographic features are identical to those of other conditions causing macular oedema, such as Irvine–Gass syndrome (vitreous–wound adherence following cataract surgery). Experimental evidence has shown that the high incidence of cystoid macular oedema in aphakic eyes is the result of higher concentrations of adrenaline in the retina of aphakic eyes (eTable 6-1).

ADRENERGIC ANTAGONISTS

Since the introduction of timolol maleate, topical β-blockers have become the most important therapeutic agents in the medical treatment of glaucoma. The lowering of IOP is a response to a blockade of β-adrenergic receptors in the ciliary epithelium, which produces up to 50% reduction in aqueous production. The site of action probably resides within the ciliary body, but it is not known whether adrenergic antagonists specifically affect ciliary body perfusion or the pumping mechanism of the ciliary epithelium. β-blockers certainly interact with β-adrenergic receptor-coupled adenyl cyclase of the ciliary epithelium. Clinically, there are now many β-blockers available, which vary according to their β-receptor selectivity. However, because of the opposing pharmacological actions of adrenaline and β-blockers, these drugs do not have an additive effect in lowering IOP. Timolol maleate is the most commonly used topical β-blocker and is non-selective, acting on both β_1- and β_2-receptors. Timolol binds reversibly with β-receptors and can reduce IOP in normal human eyes by reducing aqueous production by 15–48%. Betaxolol hydrochloride is a lipid-soluble β_1-antagonist that is relatively cardioselective. It reduces IOP by decreasing the production of aqueous at the ciliary body, with no effect on aqueous outflow. Carteolol hydrochloride is a β-blocker that possesses intrinsic sympathomimetic activity, and is thought to be beneficial in glaucoma because it not only lowers IOP but also increases optic nerve head perfusion. Its intrinsic sympathomimetic activity does not appear to confer any protection from systemic side-effects of non-selective β-blockade. New β-blockers are frequently under investigation for their effectiveness in reducing IOP and preventing the progression of glaucomatous visual field loss. The ideal β-blocker would block β-receptors in the eye without having any effect on systemic β-receptors.

Systemic side-effects of agonist agents include hypotensive episodes, cardiac arrhythmia, headache and anxiety. The systemic effects of β-blockade include bronchospasm, bradycardia, syncopal attacks and central nervous system depression. It is still not confirmed whether systemic absorption of β-blockers raises levels of serum triglycerides and cholesterol, as with systemic β-blockade, which is of particular importance when treating patients with ischaemic heart disease or more widespread vascular disease.

eTABLE 6-1 Sympathomimetic drug receptor profile/action

Sympathomimetic	Receptor/action
Noradrenaline	α and β
Adrenaline	α and β
Phenylephrine	α (non-selective)
Brimonidine	Selective α_2-agonist
Apraclonidine	α_2 partial agonist
Hydroxyamphetamine	Releases noradrenaline from nerve endings; inhibits monoamine oxidases
Cocaine	Inhibits uptake of noradrenaline at nerve endings

Carbonic anhydrase in the eye is a target enzyme for drug action

Aqueous humour is secreted actively into the posterior chamber by the non-pigmented epithelium of ciliary processes, which in turn is dependent on the active transport of sodium using the sodium/potassium ATPase pump (see Ch. 4, p. 257). Aqueous secretion can be decreased by inhibiting bicarbonate formation, an essential component in aqueous production. The production of bicarbonate is catalysed by the ubiquitous enzyme carbonic anhydrase, which exists in at least six isoforms. Carbonic anhydrase catalyses the hydration of carbon dioxide to bicarbonate, which then dissociates to form hydrogen ions and bicarbonate. The subtypes of carbonic anhydrase enzymes located in ocular tissues include carbonic anhydrase isoenzyme II, which is found in the ciliary body, and isoenzyme IV, which is thought to be a membrane-bound fraction of the enzyme found in the apical region of the RPE cells. Inhibition of carbonic anhydrase reduces aqueous production. This may be a direct effect of enzyme inhibition or secondary to the altered intracellular pH and blockade of the sodium/potassium ATPase ion channel. The effect of acetazolamide therapy is perhaps somewhat surprising. Carbonic anhydrase has a high turnover and the drug must not only reach the active site but also remain in high enough concentrations to have a sufficient duration of action. The drug binds avidly to red blood cell carbonic anhydrase and, when saturation is reached, the drug may effectively distribute to other tissue-binding sites, e.g. ciliary processes and proximal tubule of the kidney. Carbonic anhydrase inhibitors are useful in selected cases of glaucoma. However, these drugs are usually sulphonamide derivatives and can have severe systemic side-effects, including potassium depletion, dermatitis, renal stones (the incidence is said to be 11 times higher than normal), acidosis and, most commonly, fatigue and paraesthesia of the extremities, which can make up to 50% of patients intolerant to its long-term use. Acetazolamide is the most frequently used agent, and recently a new sustained-release preparation has become available which is said to reduce the incidence of systemic side-effects. Dorzolamide is a topical carbonic anhydrase inhibitor used to lower IOP in patients resistant or intolerant to β-blockers. It is irritant to the corneal and conjunctival surface and systemic absorption may rarely give rise to sulphonamide-like side-effects.

BOX 6-16 PROTECTING NERVE FIBRES WITHOUT LOWERING IOP: NEUROPROTECTION

- *Glutamate-receptor blocking agents* – reduce excessive calcium influx and ATP synthesis inhibiting excitatory cell death. Agents may act at different sites (receptors) inducing amino acid-induced excitotoxicity.
- *NMDA (N-methyl D-aspartate) receptor inhibition* – NMDA is a postsynaptic ligand-gated ion channel requiring NMDA and glycine for activation. Antagonists can either be competitive or non-competitive, both of which ultimately reduce glutamate secretion and reduce excitotoxicity.
- *Nitric oxide synthase inhibition*
- *Antioxidants*

BOX 6-17 ACTIONS OF TOPICALLY APPLIED DRUGS THAT LOWER IOP

- *β-blockers* (timolol, betaxolol, levobunolol, carteolol) decrease aqueous production by up to 50% and thus lower IOP, even in normal eyes. Their precise action remains unknown but they are likely to affect either vascular perfusion of the ciliary body or adenyl cyclase of the ciliary epithelium.
- *Parasympthomimetics* (pilocarpine) lowers IOP by facilitating aqueous outflow via its direct action on the scleral spur and ciliary body muscle.
- *$\alpha_2$adrenostimulatants* (brimonidine and aproclonidine) stimulate receptors in the trabecular meshwork (increasing both intracellular and aqueous cAMP levels), which increases the facility of outflow. Stimulation of the α-receptors on blood vessels supplying the ciliary body causes vasoconstriction and reduced blood flow, and consequently reduced aqueous production.
- *Prostaglandin analogue* (latanoprost, travoprost) is a prostaglandin $F_{2\alpha}$ analogue and a prostanoid FP receptor (Gprotein-coupled receptor) agonist, which reduces IOP by increasing the outflow of aqueous humour, mainly ureoscleral outflow.
- Carbonic anhydrase inhibitors (systemic – acetazolamide – or topically – brinzolamide, dorzolamide).

TABLE 6-2 **Profile of adrenergic receptors in the eye**

	α	β
Iris dilator	+++	+/–
Iris sphincter	+/–	+
Ciliary process epithelium	+/–	+++
Conjunctival blood vessels	+	+
Müller's muscle	+	–

BOX 6-18 HISTAMINE RECEPTOR ANTAGONISTS

H_1 – diphenhydramine, loratadine, cetirizine
H_2 – ranitidine, cimetidine
H_3 – ciproxifan
H_4 – thioperamide

in the urine after being enzymatically degraded in the liver by histaminase. Histamine is a modulator of the inflammatory response, particularly in allergic type I hypersensitivity reactions (see Ch. 7, p. 445). Histamine also plays an integral role in neurotransmission, for example regulating gastric acid secretion. Four common histamine receptors (7 transmembrane Gprotein-coupled receptors) are described, H_1–H_4 on the basis of the structure of the specific antagonist that binds to them (Box 6-18).

H_1-receptors are found in abundance in human bronchial muscle and at many other sites, including the CNS. The histamine receptors in these tissues have the same affinity for histamine as for the histamine-competitive antagonist cetirizine. H_2-receptors are found in the stomach, heart and uterus. The receptors involved have a common affinity for the competitive histamine antagonist cimetidine. Activation of H_1-receptors results in an increase in intracellular calcium concentration, and activation of H_2-receptors results in stimulation of adenyl cyclase and second messenger production (Box 6-19).

Mast cells are an abundant source of histamine. Histamine release is mediated by allergen-induced immunoglobulin E (IgE) hypersensitivity responses. Control of this allergic response can be obtained by preventing mast cell degranulation at the mucous membranes with mast cell stabilizing agents. This is of particular relevance when considering the management of allergic eye disease (Box 6-20).

ANTIHISTAMINES

H_1 antihistamines inhibit histamine-induced contraction of smooth bronchial muscles and increased

BOX 6-19 ACTIONS OF HISTAMINE

H_1 ACTIONS

- Increases vascular permeability and vasodilatation of arterioles
- Arteriolar dilatation of superficial skin vessels (axon reflex)
- Capillary dilatation and oedema of dermis
- Smooth muscle contraction/bronchospasm
- Increased mucus secretion
- Central nervous system depressant

H_2 ACTIONS

- Increased pepsin and acid production
- Increased myocardial stroke volume

H_3 ACTIONS

- Receptors associated with neural tissue at presynaptic sites, which, when stimulated, inhibit histamine release, the significance of which is unclear

H_4 ACTIONS

- Mediates mast cell chemotaxis and calcium mobilization in mast cells

BOX 6-20 TREATING ALLERGIC EYE DISEASE

TOPICAL ANTIHISTAMINES

- Azelastine
- Levocarbastine

MAST CELL STABILIZERS

- Sodium cromoglicate
- Lodoxamide

vascular permeability caused by histamine. Some H_1 antagonists have pronounced central nervous system side-effects, including drowsiness, but also have the benefit of being antiemetics. Recently introduced H_1 antihistamines have markedly reduced sedative or anticholinergic action and do not so readily cross the blood–brain barrier when administered systemically.

H_2 antagonists are effective in the pharmacological control of gastric acid secretion by decreasing basal and food-stimulated acid secretion (up to 90%). The two drugs commonly used are cimetidine and ranitidine, which are structural analogues of histamine. Cimetidine also inhibits cytochrome P_{450} and decreases the metabolism of drugs (e.g. anticoagulants, phenytoin and aminophylline), thus potentiating their effects. Cimetidine binds to androgen receptors, sometimes leading to gynaecomastia and reduced sexual function.

SODIUM CROMOGLICATE

Disodium cromoglicate is used in the treatment of allergic hypersensitivity reactions and is administered topically, either into the conjunctival fornix or by inhalation to the bronchial mucosa. It acts by inhibiting the release of histamine and slow-releasing substance of anaphylaxis (SRS-A) from the mast cells within the mucosa, by stabilizing mast cell membranes. It must therefore be administered before mast cell priming with IgE and allergen because it will have no effect once the mast cells have degranulated. Lodoxamide has also been developed as a topical application to prevent mast cell degranulation, alongside increasing numbers of topical choices in this class of agent.

Eicosanoids affect multiple ocular functions

Eicosanoids are not found preformed in tissues but are generated *de novo* from cellular phospholipids after a wide range of stimuli. They are important mediators of the inflammatory response, and the non-steroidal anti-inflammatory drugs (NSAIDs) owe some of their activity, especially prostaglandin $F_{2\alpha}$, to the inhibition of the synthesis of eicosanoids. The name prostaglandin derives from reports that semen contained a substance that contracted the uterus and was thought to be derived from the prostate gland. However, the first prostaglandins were described in the eye, although not named as such.

The principal eicosanoids are the prostaglandins, thromboxanes and leukotrienes (see Ch. 7, p. 394). The main source of these substances is from the 20-carbon unsaturated fatty acid arachidonic acid, which is found esterified in the phospholipids of cell membranes. Figure 6-10 describes the generation of these molecules, where the initial rate-limiting step is the generation of arachidonate by phospholipase A_2 or C. Stimuli that liberate these enzymes include thrombin in platelets, C5a from complement and bradykinin, as well as general cell damage. The free arachidonic acid is then metabolized via two pathways mediated by the enzymes cyclo-oxygenase, generating

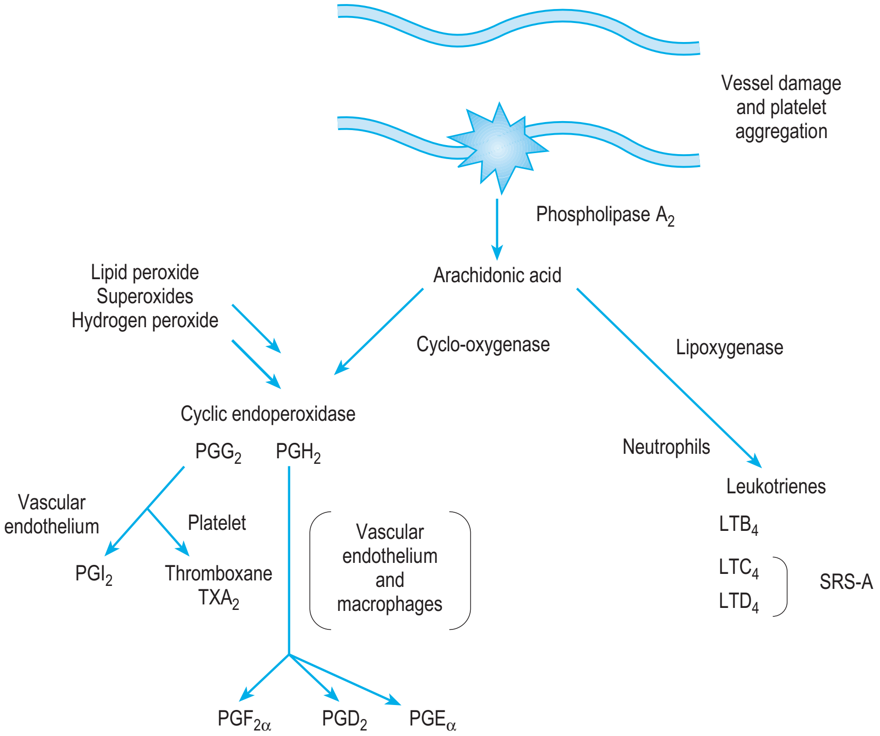
Vessel damage and platelet aggregation
Phospholipase A_2
Arachidonic acid
Lipid peroxide
Superoxides
Hydrogen peroxide
Cyclo-oxygenase
Lipoxygenase
Cyclic endoperoxidase
PGG_2
PGH_2
Neutrophils
Leukotrienes
Vascular endothelium
Platelet
PGI_2
Thromboxane TXA_2
Vascular endothelium and macrophages
LTB_4
LTC_4
LTD_4
SRS-A
$PGF_{2\alpha}$
PGD_2
PGE_{α}

FIGURE 6-10 Synthesis of eicosanoids.

prostaglandins, and lipoxygenase, generating leukotrienes (see Ch. 7, p. 394). The anti-inflammatory action of the NSAIDs is mainly the result of the inhibition of cyclo-oxygenase, and thus prostaglandin synthesis (Box 6-21).

Each prostaglandin appears to act on specific receptors which have yet to be fully characterized. However, receptor antagonists are gradually being developed for each eicosanoid. A recent example is misoprostol, which reduces gastric acid secretion and gastric erosion by NSAIDs, by the antagonism of prostaglandin E_2 class 1 receptors.

Aspirin, among other compounds, inactivates cyclo-oxygenase and reduces inflammation and pain. Other systemically administered NSAIDs, such as indometacin and flurbiprofen, are used in an attempt to reduce the inflammatory response in uveitis and scleritis. Indometacin is used also to reduce cystoid macular oedema, but its effects have yet to be substantiated. Both ibuprofen and diclofenac are now available as topical applications; their principal use is to

BOX 6-21 BIOLOGICAL ACTIONS OF EICOSANOIDS

PROSTAGLANDINS

- Prostaglandin I_2 – vasodilatation, decreases platelet adhesion
- Prostaglandin $F_{2\alpha}$ – bronchial smooth muscle contraction
- Prostaglandin E_2 – vasodilatation, bronchodilatation, uterine contraction, pyretic, stimulates release of pituitary hormones, adrenal cortex steroids and insulin from pancreas, macrophage activation

THROMBOXANE

- Thromboxane A_2 – vasoconstriction, platelet aggregation, bronchoconstriction

LEUKOTRIENES

- Leukotriene B_4 – aggregation of neutrophils, chemotactic, stimulation of phospholipase A_2
- Leukotriene C_4
- Leukotriene D_4 – contraction of smooth muscle, bronchoconstriction, vasoconstriction, leukotrienes C_4 and D_4 together form SRS-A.

prevent perioperative miosis during cataract surgery as well as to reduce postoperative inflammation.

More recently, prostaglandin analogues have been studied for use in primary open-angle glaucoma, e.g. prostaglandin $F_{2\alpha}$.

In an attempt to increase the efficacy of NSAIDs and reduce, in particular, gastric mucosal erosions, newer NSAIDs are being generated which are specific for one of the isoenzymes of cyclo-oxygenase, cyclo-oxygenase 2 (COX-2). COX-2 specificity is said to confer a more specific inhibition of inflammatory prostaglandin synthesis, without impairing synthesis of gastroprotecting prostaglandins, e.g. prostaglandin I_2, which is synthesized by COX-1. Conventional NSAIDs are strong inhibitors of COX-1 and less active against COX-2.

Serotonin: a potent neurotransmitter

Neurotransmitters are integral to retinal and cortical function, and drugs that modulate their function may alter visual perception. In addition, some neuromodulatory drugs are used to treat conditions with visual symptoms, such as migraine.

5-Hydroxytryptamine (5-HT; serotonin) is biosynthesized in a manner similar to noradrenaline from the precursor amino acid tryptophan, which is taken up into the nerve endings and converted by tryptophan hydroxylase to 5-hydroxytryptophan, and then decarboxylated to serotonin. Serotonin is degraded by oxidative deamination by the action of a group of enzymes called *monoamine oxidases* to form the aldehydes 5-hydroxyindoleacetaldehyde and 5-hydroxyindoleacetic acid (5-HIAA), analogous to noradrenaline metabolism. The aldehydes are then excreted in the urine and may be detected in hypersecretory conditions such as carcinoid syndrome.

SEROTONIN RECEPTOR SUBCLASSES MEDIATE DIFFERENT EFFECTS

Serotonin is a neurotransmitter that is widely distributed throughout the body, particularly in platelets, mucosa of the gastrointestinal tract and neurones of both the central and peripheral nervous system. There are four main types of serotonin receptors, designated 5-HT_1, 5-HT_2, 5-HT_3 and 5-HT_4. Although 5-HT is a simple molecule, it has a wide variety of effects in the cardiovascular, gastrointestinal, respiratory and central nervous systems.

5-HT_1-receptors can be further divided into subtypes 5-$HT_{1a\text{–}d}$. They are differentiated by the kinetics of agonist and antagonist binding to the receptor, and by their regional distribution within the central nervous system. 5-HT_{1a} has been found in the raphe nucleus and the hippocampus, activation of which causes hypotensive attacks and behavioural changes. This has been studied with particular reference to the physiology of anxiety. 5-HT_{1b} is found in rodents and is thought to have an inhibitory effect on the release of 5-HT, similar to the 5-HT_{1d}-receptor in humans. 5-HT_{1c} has been localized to the choroidal plexus, but no specific agonist or antagonist has yet been identified. The 5-HT_2-receptor is located in the hippocampus, frontal cortex and spinal cord. These receptors are also located on smooth muscle in the bronchus and blood vessels. They have a direct excitatory effect on smooth muscle. 5-HT_3-receptors are found in the nerves along the gastrointestinal tract and peripheral nervous system. 5-HT exerts an excitatory effect through these receptors, particularly excitation of nociceptive nerve endings. They are also found in the limbic and cortical areas of the brain, and are thought to play an important role in the development of anxiety and psychotic states. Recently 5-HT_4-receptors have been discovered in the CNS but their clinical significance remains unclear.

5-HT ANTAGONISTS

Selective 5-HT_2-receptor antagonists, which are used for the prevention of migraine attacks (e.g. methysergide and pizotifen), probably act by inhibiting the release of serotonin at the aura stage of the attack. During this stage the activation of 5-HT neurones is at its greatest, and thus antagonism of the receptors prevents the sequelae of vascular smooth muscle contraction, local inflammation and nociceptor stimulation. Selective 5-HT_3-receptor antagonism (e.g. ondansetron) reduces the 5-HT nociceptive stimulation, and thus the headache of migraine.

Glucocorticoids

Steroids are applied topically to suppress the inflammatory reaction of many conditions, commonly anterior uveitis, postoperative inflammation and corneal graft rejection. They may also be administered subconjunctivally or systemically in more severe intraocular inflammatory conditions. The mechanism of their anti-inflammatory actions is mediated by the drug's effects on both the number and function of lymphocytes, polymorphonuclear leucocytes and macrophages, and on vascular permeability. They also affect inflammatory mediators by inhibiting phospholipase A_2, prostaglandin, thromboxane and leukotrienes; and they also inhibit histamine release. Despite the pleotropic and what at first seems significantly powerful anti-inflammatory effects, steroids have many ocular side-effects, including cataract formation, reactivation of viral keratitis or increased incidence of bacterial infection, and steroid-induced rises in IOP. The latter side-effect is dependent on the duration and strength (potency) of individual steroid preparations as well as genetic predisposition.. The rise in IOP is thought to occur via the accumulation of glycosaminoglycans and water in the trabecular meshwork, reducing aqueous outflow. Thirty per cent of normal subjects have a hypertensive response when challenged with corticosteroids. This can be characterized further into poor responders, moderate responders (heterozygous responders) and strong responders (homozygous responders). If the individual also has open-angled glaucoma, a greater overall response is seen. It is therefore important that IOP is monitored at regular intervals during the course of prolonged steroid therapy.

The anti-inflammatory potency of any particular steroid is dependent on its ability to penetrate the cornea. Increasing the steroid concentration results in higher intraocular concentrations, but this may also be achieved with different formulations of the same parent steroid that increase the contact time of the topically applied drug. Steroids such as prednisolone phosphate are hydrophilic and therefore penetrate the corneal epithelium poorly; in contrast, the acetate forms of both dexamethasone and prednisolone give rise to a much greater intraocular concentration. To date, it is still not established which concentrations of steroids are desirable for low-grade intraocular inflammatory conditions. If this could be established, ocular side-effects might be kept to a minimum.

Immunosuppressant agents: combating ocular inflammatory disease

Controlling immune responses, particularly T-cell responses (Ch. 7, p. 395), that mediate allograft rejection, endogenous uveitis, associated or not with other autoimmune conditions, and chronic allergic eye disease may necessitate systemic therapy with immunosuppressants. In addition to steroids, there is now a significant armamentarium of agents used in ophthalmology to treat such conditions. Such agents have been generated as a result of our increased understanding of T-cell biology, in particular the interactions of T cells with specific antigens and antigen-presenting cells, activation of T cells and the effects of mediators of inflammation, such as cytokines.

Commonly used agents in addition to the traditional use of steroids for suppressing inflammatory responses (Box 6-22), include ciclosporin A, purine antagonists and cytotoxic agents such as methotrexate, and occasionally alkylating agents such as cyclophosphamide. Newer generations of similarly acting agents are now used, especially in prevention of solid organ allograft rejection, including tacrolimus (FK 506) and mycophenolate mofetil (CellCept®). The mechanism of action of such agents and its effect on T-cell activation are represented in Figure 6-11.

Although successful, immunosuppressive therapy is limited by its relative non-specificity, the refractiveness of some patients to therapy and the high incidence of side-effects (Box 6-23). To reduce unwanted

BOX 6-22 IMMUNOSUPPRESSANTS AND THEIR MODE OF ACTION

- *Corticosteroids* – act on cytosolic receptors and block transcription of cytokine genes (e.g. interleukins 1, 2, 3 and 5, tumour necrosis factor-α and interferon-γ)
- *Ciclosporin* – acts via inhibition of calcineurin and NFAT, resulting in inhibition of interleukin-2 production
- *Tacrolimus* – inhibits interleukin-2; mechanism similar to ciclosporin
- *Azathioprine* – inhibits purine synthesis, blocking RNA and DNA synthesis
- *Methotrexate* – folic acid antagonist, inhibiting dihydrofolate reductase and suppressing DNA synthesis
- *CellCept* – blocks *de novo* pathway of purine synthesis, which is selective for lymphocytes

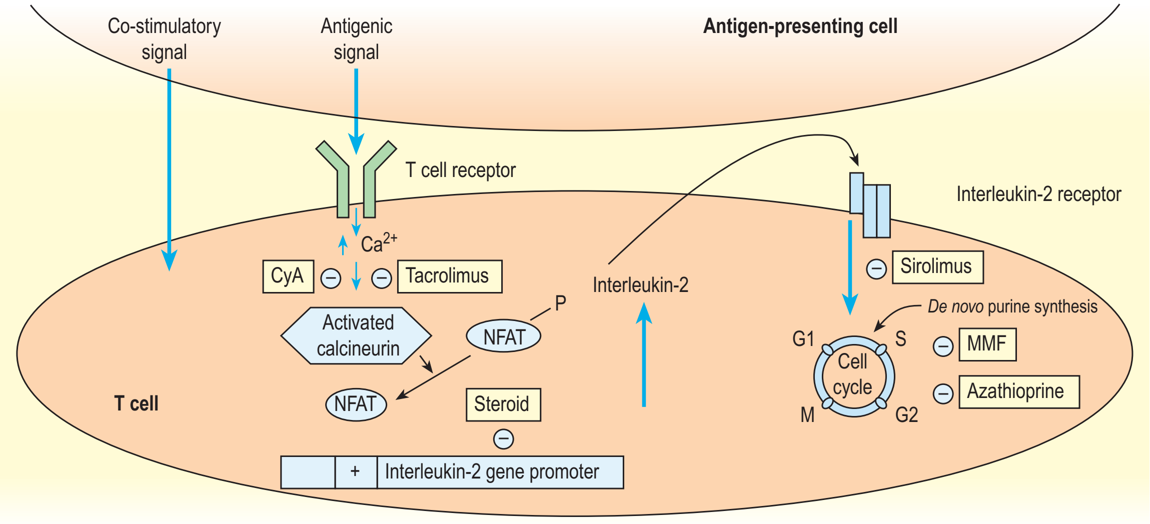

FIGURE 6-11 Stages of T-cell activation: multiple targets for immunosuppressive agents. Signal 1: stimulation of T-cell receptor (TCR) results in calcineurin activation, a process inhibited by ciclosporin (CyA) and tacrolimus. Calcineurin dephosphorylates nuclear factor of activated T cells (NFAT), enabling it to enter the nucleus and bind to interleukin-2 promoter. Corticosteroids inhibit cytokine gene transcription in lymphocytes and antigen-presenting cells by several mechanisms. Signal 2: co-stimulatory signals are necessary to optimize T-cell interleukin-2 gene transcription, prevent T-cell anergy and inhibit T-cell apoptosis. Experimental agents but not current immunosuppressive agents interrupt these intracellular signals. Signal 3: interleukin-2 receptor stimulation induces the cell to enter the cell cycle and proliferate. Signal 3 may be blocked by interleukin-2 receptor antibodies or sirolimus, which inhibits the second messenger signals induced by interleukin-2 receptor ligation. Following progression into the cell cycle, azathioprine and mycophenolate mofetil (MMF) interrupt DNA replication by inhibiting purine synthesis. (From Denton et al., 1999, with permission from Elsevier.)

BOX 6-23 COMMON SIDE-EFFECTS OF IMMUNOSUPPRESSANTS

- *Corticosteroids* – osteoporosis, hypertension, glucose intolerance, altered habitus
- *Ciclosporin and tacrolimus* – nephrotoxicity, hypertension, hyperlipidaemia, glucose intolerance, hirsutism and gingival hyperplasia
- *Azathioprine/CellCept* – bone marrow suppression, diarrhoea and gastrointestinal upset

systemic effects, some agents, for example ciclosporin A and tacrolimus, are currently being formulated for topical delivery to treat chronic allergic disease and corneal allograft rejection.

BIOLOGICS

The current expertise in molecular biological and engineering technology has permitted over the past decade the generation of specific molecules that can be engineered to specifically target receptors, membrane proteins or soluble proteins. With reference to eye diseases, this has resulted in newer treatments for ocular inflammatory disorders and retinal choroidal angiogenesis such as ‘wet’ age-related macular degeneration, as discussed earlier. For example, one of the principal mediators of T-cell responses in autoimmune disorders, such as rheumatoid arthritis, inflammatory bowel disease and posterior uveitis, is tumour necrosis factor-α (TNF-α). Engineering either specific antibodies or immunoadhesins (fusion proteins of their receptors bound to a human immunoglobulin tail) can recognize membrane-bound and soluble TNF-α and neutralize its activity. TNF-α binds two receptors (p55-TNFR1 or p75-TNFR2). The commercially available biologics include *infliximab*, a chimeric antibody (human with mouse-derived variable region recognizing TNF-α) or Humira® (a humanized monoclonal antibody) that inhibits TNF-α by neutralizing both membrane-bound and soluble TNF, and *etanercept*, which is a fusion protein of p75 receptor that successfully binds TNF, thereby preventing further binding (Fig. 6-12). There are now many targets that can be specifically inhibited with monoclonal antibody or other biological therapy (Fig. 6-13).

Local anaesthetics: an integral part of ophthalmic examination and surgery

Local anaesthetics consist biochemically of an aromatic residue linked to an amide or basic side-chain. As such, local anaesthetics are both hydrophobic (aromatic residue) and hydrophilic (amide group), and tend to accumulate at aqueous–non-aqueous interfaces. Because the aromatic residue and side-chain in some local anaesthetics are linked by esters, they are susceptible to metabolic hydrolysis. These compounds are usually inactivated in the liver and plasma by non-specific esterases. As amide links are more stable, this group of anaesthetics has a longer $t_{1/2}$.

LOCAL ANAESTHETICS AFFECT EXCITABLE MEMBRANES

Local anaesthetics block the initiation and propagation of action potentials by preventing the voltage-dependent increase in sodium conductance via a direct action on sodium channel function and, to a lesser degree, by stabilizing membranes. The action of local anaesthetics is strongly pH dependent. Under alkaline conditions the proportion of the local anaesthetic that is ionized is low but it is also lipid-soluble. This allows the anaesthetic to penetrate the myelin of

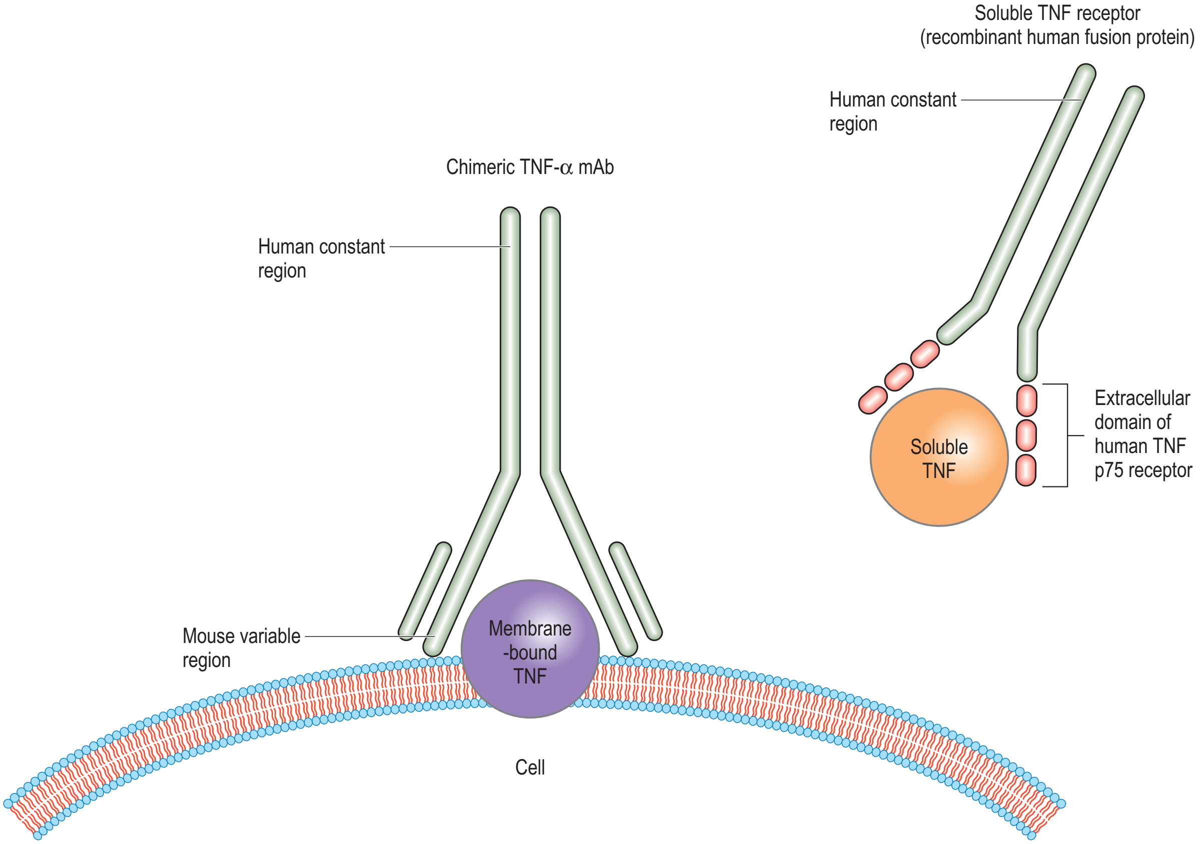
Soluble TNF receptor
(recombinant human fusion protein)
Human constant region
Chimeric TNF-α mAb
Human constant region
Extracellular domain of human TNF p75 receptor
Soluble TNF
Mouse variable region
Membrane -bound TNF
Cell

FIGURE 6-12 Schematic representation of anti-TNF agents infliximab and the fusion protein etanercept. (From Cochrane and Dick, 2007, with permission from Springer Science and Business Media.)

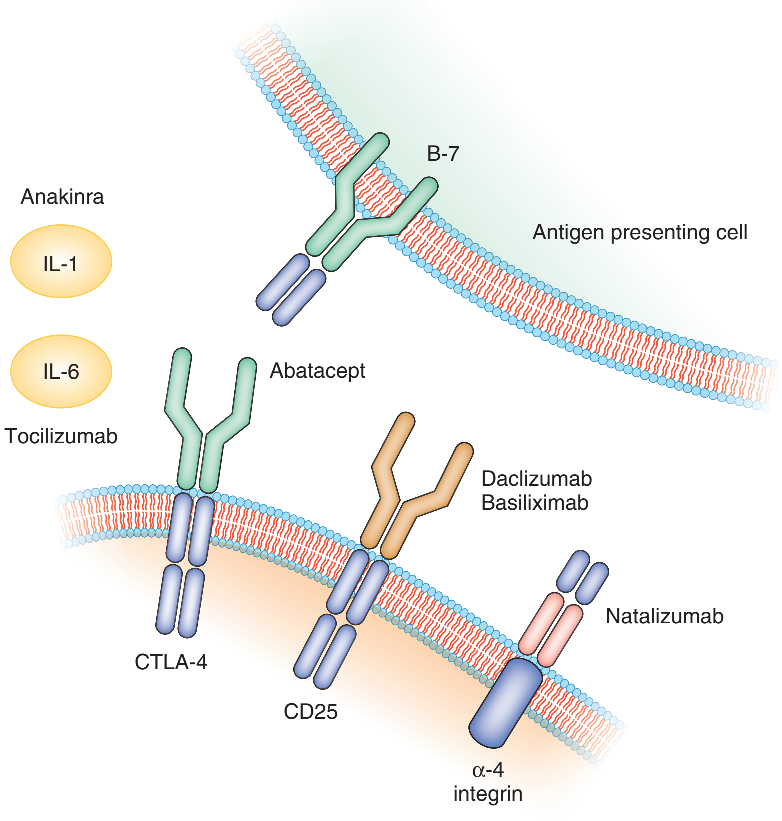

FIGURE 6-13 Biologics target key mediators of inflammation. Chimeric or humanized monoclonal antibodies or fusion proteins can deliver potent inhibition of cell response and inflammation. The schematic shows some currently available targets and biological therapies that have been generated and are being brought into practice for the treatment of inflammatory eye diseases.

the medullated nerve fibres. However, once inside the myelin sheath, the ionized cationic form of the drug is the active species, giving rise to the local anaesthetic effect. If the local anaesthetic is unable to penetrate the myelin sheath, its action can be mediated only via the node of Ranvier at several sites along the nerve. However, in unmyelinated fibres only a short length of fibre needs to be functionally interrupted to induce anaesthesia. In general, local anaesthetics block small-diameter nerves (myelinated more than unmyelinated) more readily, but practically it is not possible to produce a block of pain sensation without affecting other modalities, and there is no inherent difference between the susceptibility of motor and sensory nerves to local anaesthetics.

Local anaesthetics used in ophthalmology are usually aromatic residues linked to tertiary amide chains. These agents are more stable in acidic solutions where they are in their cationic form. However, when applied to the conjunctival sac and orbital tissues (pH 7.4), only 15% will be in the non-cationic form and lipid soluble, allowing the drug to penetrate myelin. Topical application of the anaesthetic blocks parasympathetic and sympathetic fibres first, followed by sensory (pain and temperature), and finally motor (large, myelinated) nerves.

During infiltrative regional anaesthesia for surgery (e.g. peribulbar or subtenons anaesthesia for intraocular surgery) a 1 : 200 000 concentration of adrenaline may be added to constrict blood vessels and retard vascular absorption and hydrolysis of the agent. Adrenaline is inactivated by heat, so may have reduced potency in precombined formulations. In addition, the pH of the local anaesthetic may be altered when mixing with an adrenaline solution which itself is acidic. This results in a reduced amount of the non-cationic form that is able to penetrate medullated nerve fibres.

LOCAL ANAESTHETICS HAVE BOTH LOCAL AND SYSTEMIC SIDE-EFFECTS

Topically applied anaesthetic agents are well recognized as inhibitors of wound healing. They disrupt tight junctions between cells and interfere with corneal epithelial metabolism and, ultimately, the repair of corneal epithelial wounds. Systemic effects include numbness and tingling, dizziness, slurred speech and aggressive behaviour (Box 6-24). Central nervous system toxicity may ultimately lead to convulsions with respiratory and myocardial depression. In the UK it has been recommended by the Royal College of Ophthalmologists that intravenous access (an infusion line) is available and monitoring of both heart rate and oxygen saturation is performed during infiltrative regional local anaesthesia.

Ocular toxicity from systemic administration of drugs

Ocular side-effects from the systemic administration of drugs are well recognized (see above in relation to glaucoma treatment and the use of topical steroids). Access to the globe by systemically administered drugs is restricted by the blood–retinal and blood–aqueous barriers, as stated above. In the absence of ocular inflammation, penetration of the drug into the eye is a function of the drug's physicochemical properties. Drugs or their active metabolites may accumulate in the eye, particularly within the melanin of the uveal tract, the cornea (because of its differential solubility characteristics) and the lens.

BOX 6-24 MAXIMAL SAFE DOSES OF REGIONAL LOCAL ANAESTHETICS

Lidocaine

- 10–15 mL 2% solution (200 mg)
- 20–25 mL 2% solution with adrenaline, 1:200 000 (500 mg)

Bupivacaine

- 15–20 mL 0.75% solution (150 mg)

UVEAL TRACT

Some drugs possess a high affinity for binding to melanin, from which they are only slowly released. A good example of this is chloroquine, which is known to concentrate in the melanin of the RPE and persist in this tissue for prolonged periods. If chloroquine is taken in large enough doses and for a long duration, the drug–melanin complex will result in retinal toxicity, although the exact mechanism remains unknown. The mechanism of drug binding to melanin is complex and involves electrostatic (van der Waals') forces as well as possible cation exchange, which may displace free radicals from the melanin and, in turn, give rise to retinal toxicity. However, binding of the drug alone is not sufficient to give rise to retinal toxicity. For instance, β-blockers and benzodiazepines are irreversibly bound to melanin, although the amount of bound drug does not correlate with the damage to the RPE or uveal tract. Indeed, phospholipid metabolism within the RPE is still normal after 6 months of chloroquine treatment. In the case of chloroquine, the retinopathy is seen usually in patients receiving more than 100 g total dose of drug or in patients taking the drug for more than 1 year. In general there is a low incidence of retinal toxicity from hydroxychloroquine, and some centres regard this drug as safe and not requiring monitoring of patients.

PHOTOSENSITIZATION

Photosensitizing agents absorb visible and ultraviolet radiation and, as a result, generate free radicals (see Ch. 4, p. 193). These photosensitizing agents may become bound to macromolecules in the cornea, lens and retina. Amiodarone, phenothiazines and psoralens are well-known examples of photosensitizing agents. The cornea, lens and retina may also act as drug depots. For example, once the therapeutic agent has circulated through the uveal tract into the aqueous it can rapidly penetrate the corneal endothelium and deposit in the stroma or, if lipophilic, accumulate in the corneal epithelium.

Oculotoxic drugs

Examples of oculotoxic drugs include steroids and ethambutol. Long-term steroid treatment is well recognized as a cause of both cataracts and glaucoma. Cataract formation correlates well with daily dosage and prednisolone-induced cataracts can be seen in most patients receiving 15 mg prednisolone daily. Topical application of steroids causes glaucoma by decreasing the aqueous humour outflow. The rise in IOP is related to the anti-inflammatory strength of the steroid used and the genetic disposition of the individual.

FURTHER READING

(A full reading list is available online at https://expertconsult.inkling.com/).

FURTHER READING

Cochrane, S., Dick, A.D., 2007. Tumor necrosis factor alpha-targeted therapies in uveitis. In: Pleyer, U., Foster, C.S. (Eds.), Essentials in ophthalmology. Uveitis and immunological disorders, Springer, Berlin-Heidelberg, pp. 177–192.

Denton, D.M., Magee, C.C., Sayegh, M.H., 1999. Immunosuppressive strategies in transplantation, Lancet 353, 1084.

Gaudana, R., Ananthula, H.K., Parenky, A., Mitra, A.K., 2010. Ocular drug delivery. AAPS J. 12, 348–360.

Ingelman-Sundberg, M., 2004. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past present and future. Trends Pharmacol Sci. 25, 193–200.

McGhee, C.N.J., 1992. Pharmacokinetics of ophthalmic corticosteroids. Br. J. Ophthalmol. 76, 681–684.

Chapter 7

Immunology

- Introduction
- Overview of the immune system
- Cells and molecules of the immune system
- Initial response of the host to injury (the innate immune response)
- Development of adaptive immunity and immunological memory
- Effector mechanisms
- Organization of the immune system
- Antigen recognition
- The major histocompatibility system
- T-cell activation
- B-cell activation
- Immunological tolerance and autoimmunity
- Allergy and immediate hypersensitivity
- Organ and tissue transplantation
- Tumours induce immune responses
- The eye and the immune system
- Conclusion

Introduction

Immunology is the study of host defence mechanisms. Even the simplest organisms have the ability to mount a variety of specific and non-specific responses to invasion or attack by foreign organisms. Where the host and pathogen meet and interact is where the immune response is initiated and the nature of both players determines the outcome. Thus humans can respond to pathogens in many different ways, depending on their genetic makeup, the type of foreign organism (e.g. virus, bacterium, fungus, etc.) and prevailing conditions or the setting in which the host is under threat.

The immune response is even determined by the nature of the tissue invaded. For instance, the eye (and the brain) respond when under foreign attack, but under certain circumstances the predicted response does not occur but is rather modulated by the tissue; this is called ‘immune privilege’ and is related to the special microenvironment and immunoregulatory mechanisms operating in these tissues.

Most of this chapter outlines the basic principles of immune responses, with reference to the eye as appropriate. At a very simple level, however, it must be remembered that the organism has very effective means to prevent host and pathogen meeting each other in the first place: namely, barriers such as the skin and mucous membranes with their surface coverings (including that most important layer of mucus present on mucosal surface tissues).

Overview of the immune system

Immunity is defined as the ability of the host to protect itself against a foreign organism or pathogen. To do this it requires an *immune system* comprising the cells and molecules used in the host’s defence. For unicellular hosts this may simply mean certain molecules on the cell surface that enable it to recognize foreign organisms. However, for higher-order hosts the immune system is a highly organized network of tissues, cells and molecules.

Hosts defend themselves by mounting an *immune response* involving the activation and recruitment of the cells and molecules of the immune system. Currently, the immune system is considered in terms of its specificity: the *innate (natural or native) immune system* and the *acquired (adaptive) immune system*. In its first line of defence against attack, the host uses the innate immune system because this is rapidly mobilized and is not dependent on previous exposure to the foreign invader, i.e. does not involve 'memory'. This form of response is considered relatively *non-specific* in that the same sort of response occurs to most foreign organisms and even to injury itself. It involves activation of cells (such as macrophages, see p. 377) and molecules (such as complement, see p. 405) which are present in the tissues and fluids of all higher organisms.

Innate immunity, however, has some degree of specificity in that innate immune cells such as macrophages 'sense' pathogens through receptors that recognize generic molecular 'patterns' which characterize different types of pathogens (pathogen-associated molecular patterns or PAMPs). Thus there are receptors for broad groups of molecules present in different organisms such as viral DNA, fungal carbohydrates and bacterial endotoxins (see Ch. 8, p. 464). However, as stated above immune responses can be initiated not only by pathogens, but simply by injury, or by molecules such as plant pollens (causing allergies), and especially by 'alloantigens' (those molecules that differentiate individuals within a species and cause rejection of donor-unrelated transplants). Accordingly, Matzinger (2012) developed Janeway's PAMP hypothesis to suggest that innate immunity has evolved to allow the host to sense 'danger' or damage to tissues and includes anything that threatens host or tissue viability, namely pathogens of all varieties, sterile trauma such as chemical or thermal injuries, non-self proteins such as those present in allografts, and even 'self' proteins if these have become 'abnormal' (as in the case of cancer cell proteins, prion proteins, and even self proteins (auto-antigens).

Additional content available at https://expertconsult.inkling.com/.

Accordingly the molecular patterns associated with danger signals and tissue damage are known as danger- or damage-associated molecular patterns (DAMPs) and include PAMPs. Thus bacterial and viral products such single- or double-stranded RNA, endotoxin, complement fragments, reactive oxygen species, heat shock proteins, high mobility group box (HMGB) proteins, and many other molecules (sometimes called 'alarmins') can act as DAMPs.

The host-defence immune system thus includes:

- physicochemical barriers such as the skin, eyelids, tears (see Ch. 4)
- molecules normally present in body fluids such as blood, tears and aqueous humour (e.g. complement, lysozyme, antiproteases). Antibacterial defensins are also on this list of ocular surface proteins
- phagocytic and cytotoxic cells such as polymorphonuclear leucocytes, macrophages, eosinophilic granulocytes, natural killer (NK) cells
- molecules released by cells responding to attack and acting on other cells (cytokines), such as interleukins and tumour necrosis factor-α (TNF-α) (see below) and complement.

As indicated above, the innate immune system does not have 'memory' and so if, after successfully repelling pathogen invasion and clearing the pathogen from the system, a second attack by the same pathogen occurred, the organism would have to go through the same process again. However, innate myeloid immune cells and specifically one subset of specialized myeloid cells, termed dendritic cells (DCs), after devouring and digesting a pathogen, has the potential to present titbits of the devoured material to specialized immune cells, termed T cells, which once activated in the throes of battle with the pathogens (inflammation), become specialized not only to assist in clearing the pathogen (effector T cells) but also to become long-lived memory cells which can respond rapidly and with minimal fuss to a repeat challenge by that same pathogen. This is acquired immunity, which is exquisitely specific and not only involves T cells which can kill pathogens and infected cells but also can activate other immune cells, namely B cells, to produce very high levels of pathogen-specific killer molecules (antibodies) which are the basis of vaccination protocols now used against many common pathogens.

The acquired immune system includes:

- specific immune systems associated with barrier surfaces (the skin immune system and the mucosa-associated immune system, MALT) where memory T cells reside and have 'first call' on invading pathogens
- a constant source of naive T and B lymphocytes with potential to modulate their receptors to specifically recognize foreign organisms and molecules (antigens)

There has been a recent explosion of knowledge concerning the innate immune system (see below). After the initial discovery of the phagocytic leucocyte (macrophage) by the Russian zoologist Metchnikoff at the beginning of the 20th century, much of immunology was devoted to immunochemistry of antigens and antibodies and then, in the latter half of the 20th century, to developments in cellular immunology, involving cells of the adaptive immune system such as T cells and B cells. Specificity for pathogen recognition by innate immune cells was considered non-existent until Janeway and others identified receptors in the fruit fly *Drosophila* which were important not only for wing patterning in development but also for recognition of pathogens, such as fungi. His earlier proposal that there were classes of receptors recognizing broad groups of pathogens such as viruses, fungi, bacteria, including mycobacteria, and parasites was shown to be true by the discovery of Toll receptors. Other classes of pathogen-associated molecular patterns were later discovered. However, as Matzinger pointed out, pathogens were not the only agents which activated the innate immune system. Allergens also initiate immune responses, as do allografts, where infectious pathogens were considered not to play a part. Her concept of molecules being either 'dangerous' or 'harmless' has changed the thinking of immunologists in many ways. In addition, she has developed her concepts to consider that different tissues respond to the same insult immunologically in different ways: for instance, the skin and the liver might produce a modulated immune response to the same antigen. In one sense this is a rewriting of the notion of immune privilege, which has been long accepted as a modified immune response taking place in the eye (see p. 457).

- molecules that specifically counteract foreign antigens (antibodies); these proteins are known as immunoglobulins and there are five types (see p. 396)
- non-specific molecules (e.g. cytokines) released by antigen-specific cells (e.g. lymphocytes) which will recruit further myeloid cells to sustain the response as needed.

Specific immunity is described as *humoral* when antibodies (derived from B lymphocytes; see below) and complement are involved in removing the antigen or as *cell-mediated* (cellular) when T lymphocytes and macrophages are involved.

Immunity after infection is normally termed *active immunity* in that the host has responded actively to the stimulus. However, immunity may be transferred *passively* by antibodies or cells. Vaccination procedures that involve the administration of antibodies are termed *passive immunization*, while those that involve inducing a response to the antigen or even the attenuated live organism are termed *active immunization* (see Ch. 8).

The development of acquired immunity involves a number of discrete phases, including:

- an afferent phase in which the foreign antigen is transported from the site of entry and presented by specialized antigen-presenting cells (APCs) to the lymphocytes in the lymphoid tissue (see below)
- a phase of T-cell activation in which T cells are transformed from a resting to an active state
- an effector stage in which T cells induce other cells, such as B cells and macrophages, to remove the antigen. If the antigen remains intracellular, as with virus-infected cells, the T cells themselves attack the infected cell (cytotoxic T cells).

This is known as the *primary immune response* and is accompanied by the appearance of antigen-specific T cells and antibody-secreting B cells. As indicated above, the acquired immune system evolved to provide a memory-based rapid reaction force to heighten immune defence. Thus, on second exposure to the same pathogen, antigen-specific memory T and B cells are recruited much sooner and more efficiently, such that antibody levels are considerably higher than on the first exposure. This is known as the *secondary immune response*. In addition, the type (isotype; see below) of antibody in secondary immune responses is different.

Throughout the development of the acquired immune response there are several checkpoints that prevent a runaway overwhelming inflammatory reaction. Most of these regulatory mechanisms are in place to prevent adaptive immune responses to self-antigens but they also participate in downregulatory responses to foreign antigens and help to restore homeostasis.

Innate immunity with its early warning, rapid-response system provides a reliable means of protecting the host against most extracellular organisms (pathogens) and is a property of every living being. It might be asked, therefore, why the acquired immune system has evolved? In part this is due to the remarkable ability of organisms to evade the immune system. Many pathogens reside and hide within the host cell, as in the case of protozoa such as *Toxoplasma*, obligate parasites such as *Chlamydia* or, more frequently, viruses. Viruses use the host cell machinery for survival and replication by incorporating their genome, or at least part of it, into the DNA of the host cell. This may lead to viral latency or persistence, but may also allow the immune system to recognize the infected cells by the expression of the foreign antigen on the surface of the cell in addition to the self molecules (also termed antigens because they can induce an immune response). It is the recognition of the foreign antigen in the context of self-antigen (i.e. together with self-antigens, the peptide–major histocompatibility complex (MHC); see p. 424) that led to the evolution of the acquired immune system. However, many pathogens continue to evade the immune system by subverting the function of the innate immune cells in which they reside or by becoming 'latent' as in the case of herpes simplex-infected neurones.

The innate immune system is the initial primary responder to pathogens or other challenges. The acquired immune system has developed a considerable degree of sophistication, based on memory of previous pathogen encounters, with an increasingly recognized range of different cell types: T and B lymphocytes; T and B cell subsets (e.g. T helper (Th), T cytotoxic (Tc), T and B regulatory (Treg and Breg) cells; and even subsets of these (e.g. Th0, Th1, Th2, Th17) (see below), each designed for specific cellular functions. In addition, T and B cells have evolved in fundamentally different directions – T cells to deal with surface-bound antigen (usually cell-associated) and B cells to deal with soluble (extracellular) antigen. The sophistication and specificity of the acquired immune system involving T cells has thus been harnessed to assist the innate immune system in dealing more efficiently with extracellular organisms, for instance via B cells.

Certain basic concepts about immune mechanisms can therefore be derived from the above considerations:

- extracellular foreign antigen is normally cleared by the innate immune system, with some assistance from B-cell activity
- intracellular foreign antigen may evade the immune system and remain latent during the lifetime of the host unless (1) it kills the cell, is released and generates a second innate immune response to remove the pathogen; (2) it is expressed on the surface of the cell in the context of self-antigen and is recognized by the acquired immune system, which kills the infected cells and clears the pathogen (applies to many viruses, e.g. influenza); (3) the host immune system is compromised, in which case the intracellular pathogen proliferates freely, killing many cells and ultimately the host
- all cellular defence mechanisms involve interactions of cell surface molecules (receptors) with complementary molecules (ligands).

So far we have focused on *immunity*, defined as the response of the host to foreign organism. This suggests that all foreign organisms are pathogens but this is not so. Indeed the great majority of microbes and other living organisms are not pathogenic, as evidenced by the very large numbers of *commensal* organisms which colonize several regions of the body, particularly the GI tract (the microbiome) (Box 7-1), and indeed are required for normal health (homeostasis).

This lack of immune responsiveness to foreign antigens can be described as 'tolerance' of foreign organisms and indeed the immunological obverse of 'immunity' is 'tolerance'. Tolerance as a defining characteristic can be applied to any antigen, but is especially applied to self-antigens, since clearly we do not normally react to our own tissues. It has been realized that tolerance to self-antigens, and indeed to any antigen, is an active process in which certain T cells (T regulatory cells [Tregs]) are critical for its normal function. Many effector immune responses to infectious foreign antigens are controlled by the induction of a specific Tregs to that organism. Tregs are part of our normal circulating T-cell population and when Treg function goes awry, autoimmune diseases can ensue; diseases such as various forms of dry eye disease (Sjögren syndrome, see Ch. 9), rheumatoid arthritis and some forms of uveitis are considered autoimmune diseases. Indeed there is a growing recognition that both the innate and the acquired immune response co-evolved with microbes, each exerting counter-regulation on the other.

To summarize, acquired immune responses have certain features that are inherent to them:

- specificity – based on certain features/determinants (known as epitopes) of the antigenic structure
- non-responsiveness to many antigens (prototypically, self-antigens), also known as tolerance, the loss of which underpins autoimmune disease
- diversity – around 10^9 individual epitopes can be distinguished – this is known as the lymphocyte 'repertoire' and represents the range of potential foreign antigens to which immune responses can be generated
- memory – the secondary immune response, specific to that antigen and present in both B and T cells
- specialization – the immune response is customized to different microbes
- ability to downregulate – the immune response is strictly regulated/limited in magnitude and time through specific and non-specific mechanisms.

Cells and molecules of the immune system

The cells of the immune system are mostly specialized to participate in processes belonging either to the innate or the acquired immune systems (Box 7-2). Some cells are central to both and the distinction between innate and adaptive immune cells is becoming blurred with the discovery of cells such as innate lymphoid cells (see p. 382 and Video 7-1).

BOX 7-1 THE MICROBIOME

The total number of separate genomes from microorganisms colonizing a specific site (known as an environmental 'niche') constitutes a 'microbiome' for that site. There are 10 times more microbial cells than human cells in the body and although it does not weigh much more that 500 grams, its importance to normal physiological processes is such that some regard it as a separate and essential 'organ'. Most of the organisms that constitute the microbiome cannot be cultured and have been identified by genetic mapping techniques. The most important role for the microbiome may be in immunoregulation and prevention of disease including autoimmune disease, diabetes, atherosclerosis, cancer and other diseases associated with chronic inflammation and obesity. It may also have a role in psychiatric disease through its effects on neurotransmitter production.

High-throughput sequencing has revealed substantial intra-individual microbiome variation at different anatomical sites, and inter-individual variation at the same anatomical sites (see figure). However, the microbiome at specific anatomical sites can remain remarkably stable over time in each individual. The figure indicates the relative proportion of sequences determined at the taxonomic phylum level at eight anatomical sites. Certain features, such as the presence (+) or absence (–) of *Helicobacter pylori*, can lead to permanent and marked perturbations in community composition. (adapted from Costello et al., 2012).

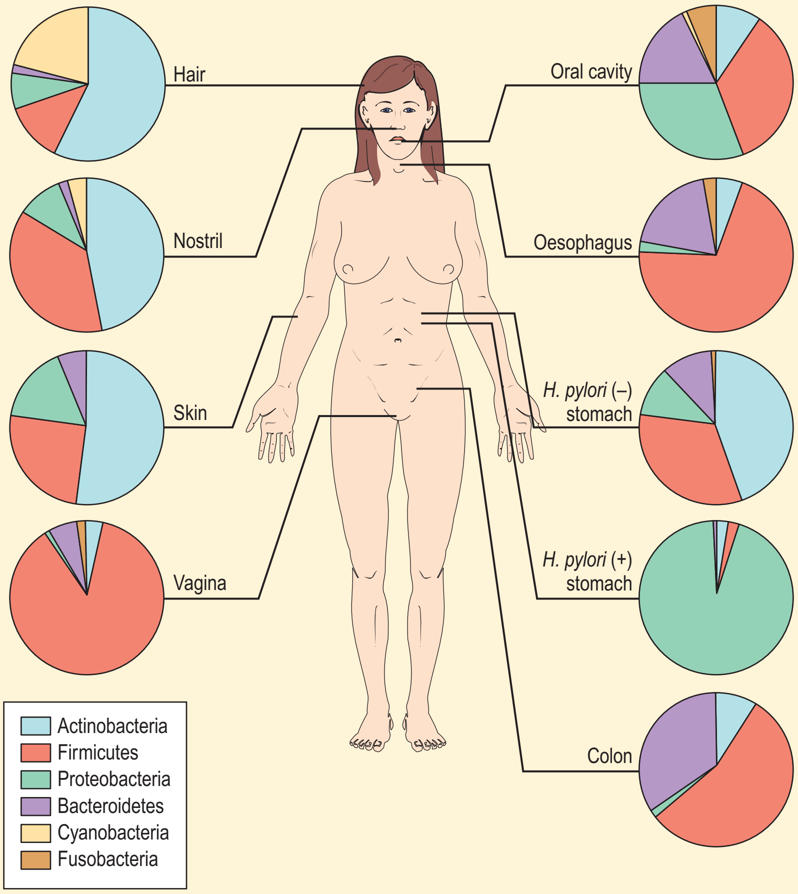

BOX 7-2 CELLS AND TISSUES OF THE INNATE AND ACQUIRED IMMUNE SYSTEMS

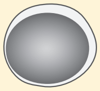	Mature T or B lymphocyte	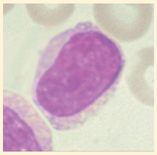	6–9 µm; round or slightly indented nucleus; sparse cytoplasm; few granules; few mitochondria
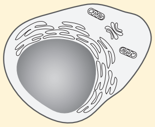	Plasma cell	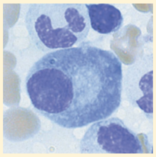	5–30 µm; round or oval nucleus; abundant cytoplasm; no granules; abundant endoplasmic reticulum
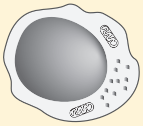	NK cell	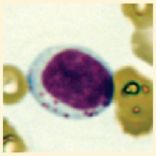	10–12 µm; round nucleus; abundant cytoplasm; many granules; scattered mitochondria
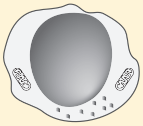	NKT cell	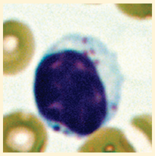	Phenotypically similar to NK cell
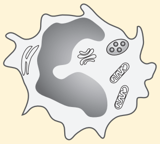	Monocyte	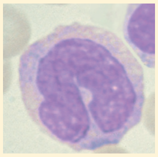	12–20 µm; round, oval, notched, or horseshoe-shaped nucleus; abundant cytoplasm; abundant granules; well-developed Golgi apparatus; abundant mitochondria
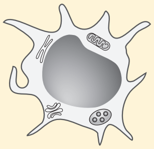	Macrophage	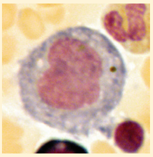	15–80 µm; elongated, indented, or oval nucleus; abundant cytoplasm; many granules and vacuoles; few mitochondria; abundant lysosomes
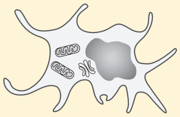	Dendritic cell	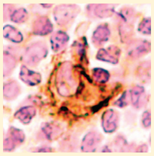	Irregularly shaped cell and nucleus; many cellular processes; few intracellular organelles; prominent mitochondria

Lytic granules
Phagosomes
Mitochondria
Golgi apparatus
Smooth endoplasmic reticulum
Rough endoplasmic reticulum
Granules

BOX 7-2 CELLS AND TISSUES OF THE INNATE AND ACQUIRED IMMUNE SYSTEMS—cont'd

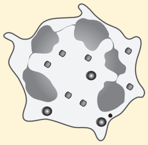	Neutrophil	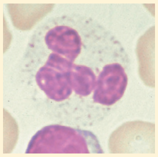	10–15 μm; 2–5 distinct nuclear lobes; abundant cytoplasm; numerous granules; few mitochondria; abundance of glycogen
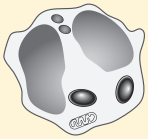	Basophil	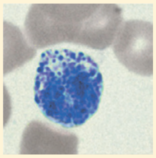	14–16 μm; 2–3 nuclear lobes; abundant cytoplasm; large, coarse granules; many ribosomes and mitochondria; abundance of glycogen
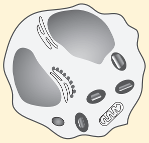	Eosinophil	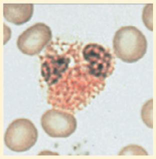	12–17 μm; 2–3 nuclear lobes; abundant cytoplasm; large oval granules containing elongated crystalloid and smaller granules; extensive smooth endoplasmic reticulum; few mitochondria
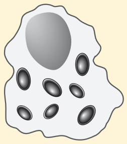	Mast cell	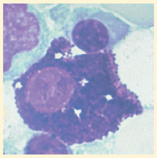	14–16 μm; non-segmented nucleus; abundant cytoplasm; many large granules; few mitochondria
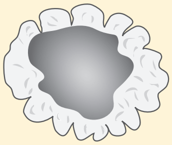	Megakaryocyte	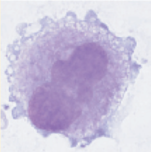	35–160 μm; irregularly shaped nucleus; abundant cytoplasm; fine granules
	Platelet	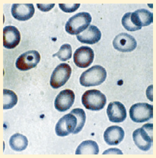	1.5–3.5 μm; non-nucleated; granular cytoplasm
	Erythrocyte	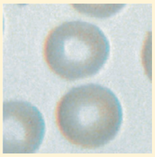	7.2 μm; non-nucleated; no organelles

(From Mak and Saunders, 2006, with permission from Elsevier.)

THE MYELOID SYSTEM AND INNATE IMMUNITY

The cells directly involved in bacterial killing and removal of damaged host tissue at the site of entry of the foreign organism are the cells of the myeloid system. Some of these cells are recruited to the site of injury during acute inflammation. Their defining feature is the presence of cytoplasmic organelles (lysosomes) containing a battery of hydrolytic enzymes for both intracellular and extracellular killing.

Neutrophilic granulocytes

Neutrophilic granulocytes (polymorphonuclear leucocytes) are the most common white cell in the circulation and are attracted to sites of inflammation by chemotaxis (see p. 387). These are fully differentiated cells with no capacity for proliferation. They are primary scavengers, causing much of their effect via release of free radicals and proteases from their numerous cytoplasmic granules and lysosomes. These include defensins, lysozyme, lactoferrin and oxidative enzymes (e.g. NADPH-dependent oxidases, myeloperoxidases and catalase), which are also present in ocular fluids such as tears. The half-life of neutrophils is 1–2 days.

This limited 'clearing-the-bugs' role has been revised recently with many new functions being ascribed to neutrophils. PMNs have been shown to use a remarkable mechanism for 'trapping' bacteria and other 'dangerous' material by extruding NETS (neutrophil extracellular traps), composed predominantly of nuclear DNA material, during the process of cell death. NETS are very 'sticky' and powerfully prevent bacterial dissemination. However, they can also have deleterious effects by trapping metastatic cells and protecting them from immune surveillance (see eFig. 7-2).

Neutrophils may also participate in the adaptive immune response through release of cytokines and extracellular vesicles which mediate intercellular communication.

Myeloid mononuclear cells

Monocytes, macrophages and dendritic cells are cells of the mononuclear phagocyte system. Like neutrophils, these cells are derived from haematopoietic bone marrow stem cells and differentiate into a variety of tissue macrophages (histiocytes) with specific functions (Box 7-3).

Macrophages have many functions as part of the innate immune system: they phagocytose dead and damaged cells and organisms in inflammatory exudates; they release cytokines of various sorts, which may activate other cells such as lymphocytes and eosinophils; and they are involved in the acquired immune system as antigen-presenting cells (APCs), as effector cells involved in the process of cell lysis, and by removing antibody-coated (opsonized) cells and particulate material.

Macrophages are, however, not simple cells. There are several varieties of macrophages, some with 'housekeeping' roles as resident cells whose job is to remove dead and dying cells during the normal tissue cell turnover. Other macrophages are recruited to sites of inflammation, differentiating from circulating blood monocytes, and are highly active in tissue destruction. These cells are sometimes referred to as classically activated 'M1' macrophages with a strongly pro-inflammatory function, while resident and other macrophages are considered to be alternatively activated 'M2' macrophages and function more in healing processes including fibrosis and new vessel formation (eFig. 7-1).

Additional content available at https://expertconsult.inkling.com/.

Macrophages, like B cells, usually capture antigen (e.g. 'opsonized' bacteria) for processing and presentation to T cells via surface receptors such as the complement receptor and the Fc receptor. The Fc receptor on macrophages binds the Fc portion of antibody–antigen complexes while B cells express antibody (immunoglobulin) on their surface and use the Fab portion of the antibody to capture antigen. In both cases, preformed antibody to that antigen is needed and so the immune system must have already been exposed to antigen. It is clear therefore that macrophages and B cells are only effective as APCs in the later stages of an ongoing immune response.

In contrast, the most potent APC in the immune system is the dendritic cell (DC). The DC is the key cell in linking the adaptive and innate immune systems. While the function of most myeloid cells is to engage in innate immune defence functions by marshalling forces against infectious and non-infectious danger signals, the DC takes this one step further: after phagocytosing, killing and digesting an invasive

The functional subtyping of macrophages is of considerable importance to the understanding of disease processes including many ocular diseases such as uveitis, and macular degeneration. The initial classification into M1 classically activated macrophages and M2 alternatively activated macrophages has developed further into a realization that there is greater plasticity in macrophage function than was once thought (eFig. 7-1). Classically activated macrophages remain major players in innate immune defence with the job of removing dangerous foreign and endogenous antigens, clearing necrotic and damaged tissue and setting the scene for repair of tissue. This is conducted by alternatively activated macrophages involved in such processes as generating new blood vessels to restore blood supply to the tissues and produce factors which rein in the excesses of the inflammatory response. Meanwhile a third type of macrophage differentiates to mediate other aspects of wound healing and repair, i.e. to promote maturation of blood vessels and fibrosis, while a fourth hybrid type of macrophage has features of a regulatory (anti-inflammatory) macrophage, producing the anti-inflammatory cytokine IL-10 if it encounters bacterial factors such as lipopolysaccharide.

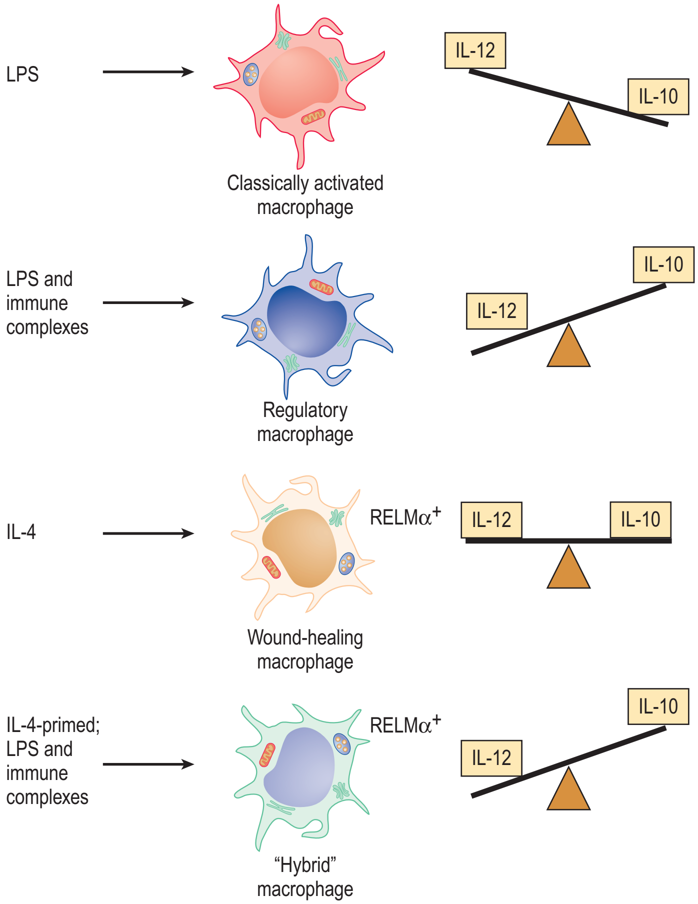

eFIGURE 7-1 Macrophages are classically considered inflammatory cells whose function is to participate in inflammation and get rid of pathogens. However, they are now recognized to be quite heterogeneous in their function: (1) inflammatory macrophages produce interleukin-12 (IL-12) and little IL-10; (2) regulatory macrophages produce IL-10 and little IL-12; (3) alternatively activated macrophages treated with IL-4 ('wound-healing' macrophages) produce little IL-10 or IL-12 but express resistin-like molecule-α (RELMα); (4) lipopolysaccharide or immune complex treatment of the alternatively activated macrophage produces a mixed macrophage (an ELMa+ wound healing-like macrophage) but also produces high levels of IL-10, like a regulatory macrophage. (From Mosser, 2008.)

BOX 7-3 GENERATION OF MACROPHAGES AND DENDRITIC CELLS FROM MYELOID AND LYMPHOID PRECURSORS

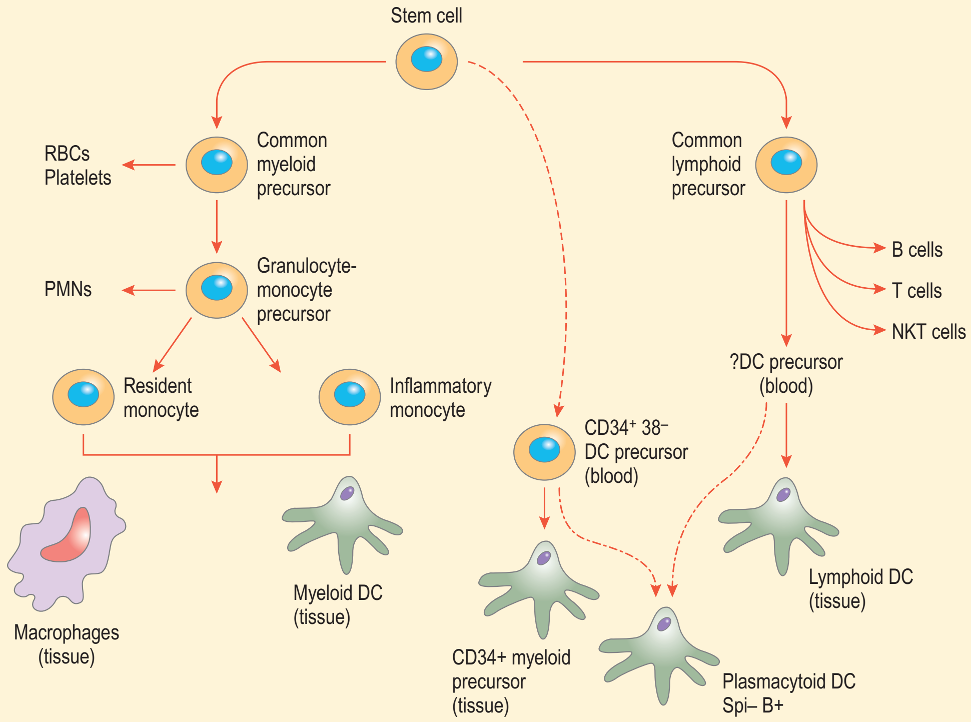
Stem cell
RBCs
Platelets
Common
myeloid
precursor
Common
lymphoid
precursor
B cells
T cells
NKT cells
PMNs
Granulocyte-
monocyte
precursor
?DC precursor
(blood)
Resident
monocyte
Inflammatory
monocyte
CD34$^+$ 38$^-$
DC precursor
(blood)
Macrophages
(tissue)
Myeloid DC
(tissue)
Lymphoid DC
(tissue)
CD34+ myeloid
precursor
(tissue)
Plasmacytoid DC
Spi– B+

Note that dendritic cells (DCs) can be derived from both sources. NKT, natural killer T cells; PMNs, polymorphonuclear cells; RBCs, red blood cells.

microorganism the DC, unlike the macrophage or the B cell, or even 'non-professional' APCs, can present the processed antigen to naive T cells and induce an entirely new adaptive immune response with all its features, including memory (see above section). There are several subsets of DCs but, in essence, there are two major groupings: resident, non-migratory DCs which are located in the secondary lymphoid tissues such as the spleen and the lymph node and migratory DCs which arise in the bone marrow and migrate to the tissues where they are available as sentinels to register pathogens and other danger signals and transport antigen to the secondary lymphoid tissues where they activate the adaptive immune system (T and B cells). Tissues at the front line of attack, e.g. lungs, skin, gut mucosa and conjunctiva, are rich in DCs (see eFig 7-2).

Additional content available at https://expertconsult.inkling.com

Tissue-resident DCs constantly traffic from the bone marrow to the tissues and then to the secondary lymphoid organs. If they do not detect 'dangerous' antigens, but only self-antigens, they do not activate the adaptive immune system and in fact maintain immune tolerance by sustaining the population of endogenous Tregs, i.e. it seems that, although DCs have the powerful potential to rapidly activate adaptive immunity, e.g. in the case of a viral infection, their homeostatic (and probably main function) is to maintain tolerance.

Other granulocytes

Mast cells/basophils/eosinophils are all part of the granulocyte series of cells. Basophils (life span: 2–3 days) are the circulating equivalent of mast cells (life span: weeks), which occur only in the tissues. However, mast cells and eosinophils may have

Dendritic cell (DC) subsets

Several types of dendritic cell have been identified on the basis of cell surface markers (CD markers, see p. 381). Two main subsets are recognized: plasmacytoid DC and conventional DC (eFig. 7-2). Both are derived from precursors in the bone marrow, the myeloid-derived precursor cells (MDP), which also generates precursor cells for the macrophage lineage. Such precursors circulate in the blood and seed the tissues in homeostatic conditions where they provide a surveillance function for pathogens and danger signals. They also proliferate and respond rapidly to danger and can become inflammatory DCs. Some of these are derived from monocytes which can become either tissue inflammatory macrophages (see above) or become Tip DCs (TNF-α, NO-producing DCs) which are active in antigen presentation and generation of the immune adaptive immune response.

Studies on DC lineage and phenotype have mostly been performed in the mouse and human. DC subsets and markers are defined by several different cell surface markers which are variably expressed on different DC subtypes (www.RnDsystems.com). For a summary of how DCs control adaptive immunity, see review by Miriam Merad and colleagues and a summary in the accompanying poster on www.biolegend.com (Merad et al., 2008).

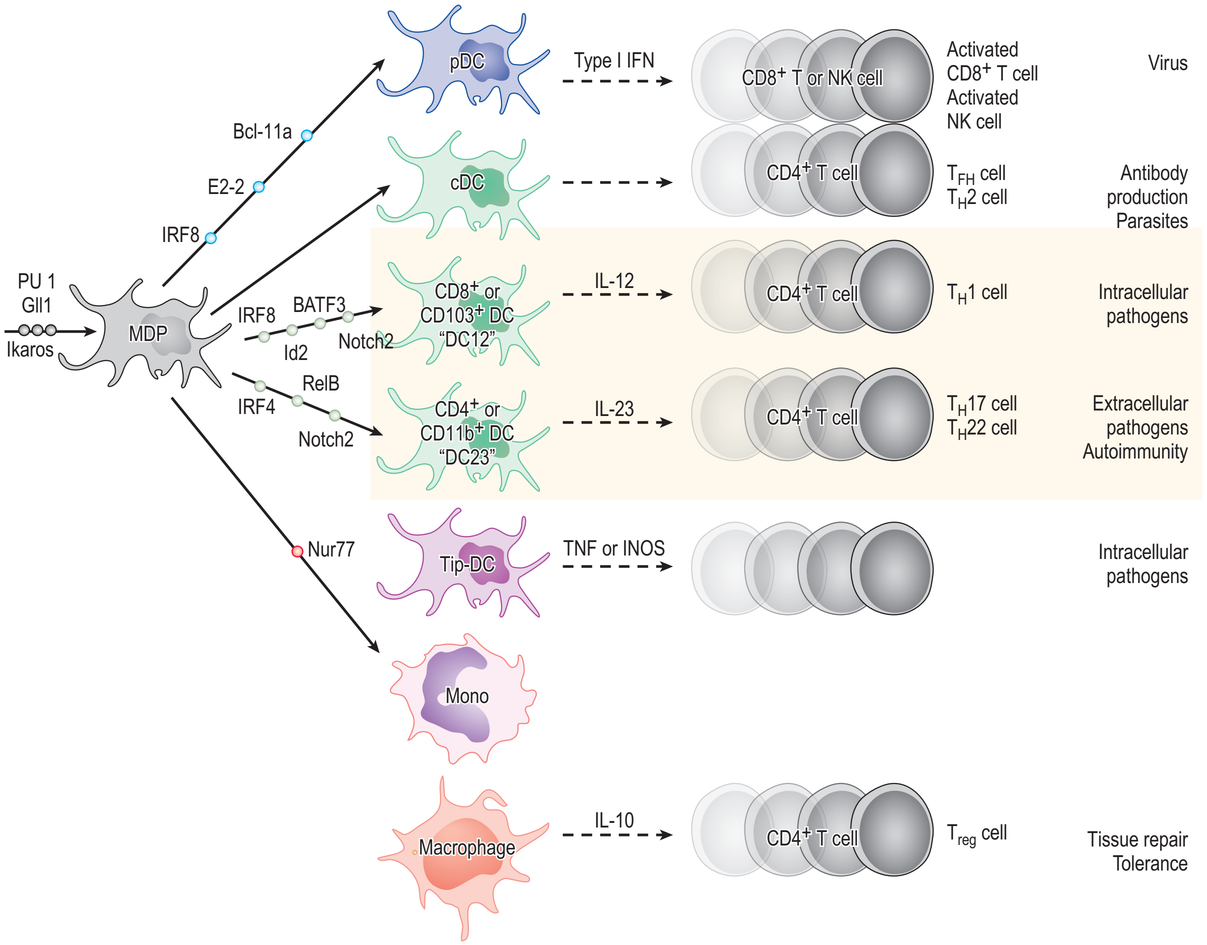
pDC
Type I IFN
CD8+ T or NK cell
Activated
CD8+ T cell
Activated
NK cell
Virus
Bcl-11a
E2-2
IRF8
cDC
CD4+ T cell
TFH cell
TH2 cell
Antibody
production
Parasites
PU 1
Gli1
Ikaros
MDP
IRF8
BATF3
Id2
Notch2
CD8+ or
CD103+ DC
"DC12"
IL-12
CD4+ T cell
TH1 cell
Intracellular
pathogens
IRF4
RelB
Notch2
CD4+ or
CD11b+ DC
"DC23"
IL-23
CD4+ T cell
TH17 cell
TH22 cell
Extracellular
pathogens
Autoimmunity
Nur77
Tip-DC
TNF or INOS
Intracellular
pathogens
Mono
Macrophage
IL-10
CD4+ T cell
Treg cell
Tissue repair
Tolerance

eFIGURE 7-2 Myeloid-derived progenitor (MDP) cells in the bone marrow produce plasmacytoid dendritic cells (pDCs) as well as several types of 'conventional' DC (cDC). CD8$^+$ cDCs mostly reside in the secondary lymphoid tissue, while CD4$^+$ cDCs reside in the tissues and form the major population of migratory DCs. MDP also generate monocytes (Mono), which seed the tissues and become tissue-resident macrophages but can also convert to a form of inflammatory DC (TipDC). Each DC subset 'specializes' in a specific immunological function. (From Satpathy et al., 2012.)

separate lineages and so their ontogeny is not yet decided.

Two types of mast cell are described: the connective tissue mast cells and the mucosal mast cells on the basis of their granule proteases, their susceptibility to degranulation by immunoglobulin E (IgE) and their requirement for different maturation signals (see below). Mucosal mast cells line epithelial and connective tissue mast cells line endothelial (blood vessel) barriers and are thus important innate immune cells, contributing to the initial response to external pathogens or endogenous danger signals by releasing mediators and increasing vascular permeability (Fig. 7-1). Mast cells have a further important role of engineering the contraction of the innate immune response.

Eosinophils account for about 0.1% of circulating leucocytes. Their numbers are elevated in chronic allergic disorders, in both the circulation and the tissues. They are also produced in large numbers in response to parasite and helminth infections and contain a panel of particularly potent parasite-specific proteases. Like mast cells they have high-affinity IgE receptors, and are probably effectors of tissue damage in allergic disease, including asthma and chronic allergic conjunctival disorders such as atopic keratoconjunctivitis and vernal keratoconjunctivitis (see p. 445). Interestingly, eosinophils are major components of allergic disease, and have been reported to be prominent in fungal antigen-driven allergy. In this respect, eosinophils, like mast cells and basophils, are classical biomarkers of type 2 immune responses (see p. 418) and are associated with alternatively activated (M2) macrophages. As such they promote not only immune homeostasis but also metabolic homeostasis by promoting glucose tolerance and preventing fat deposition.

Natural killer cells are an infrequent but important constituent of innate immune cells, although they are technically lymphoid cells. Their main function is to provide defence against virally infected cells and tumour cells which they perform in an antigen-independent manner. Classically they are considered not to generate immunological memory, which is the

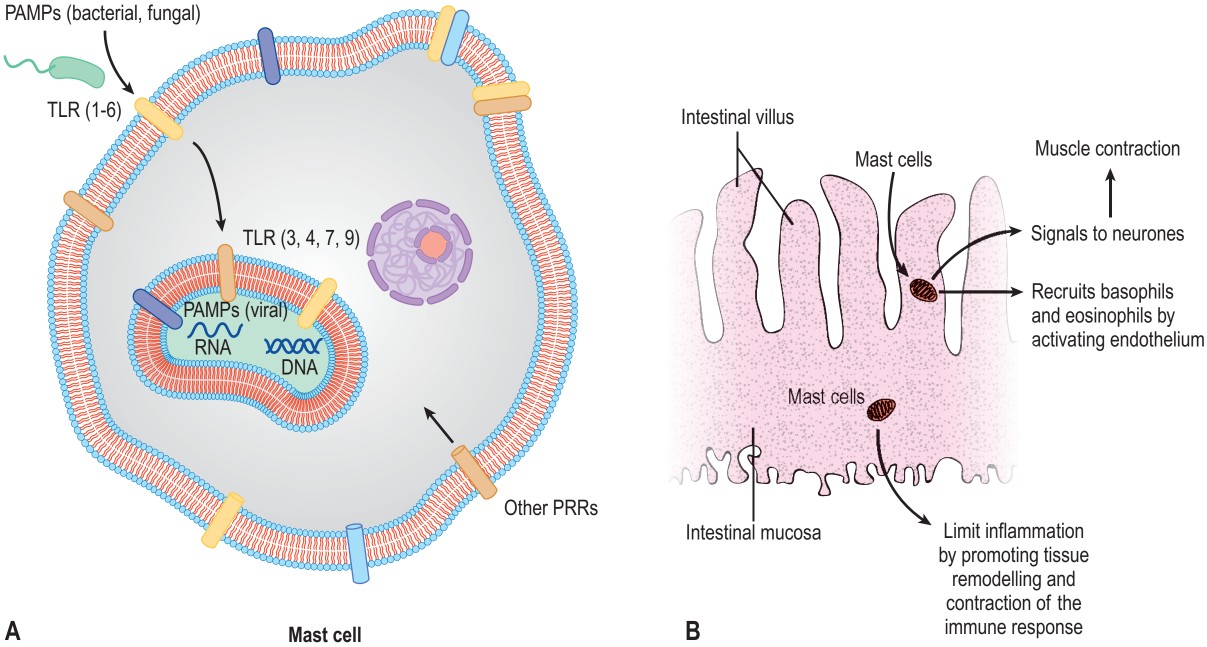
PAMPs (bacterial, fungal)
TLR (1-6)
TLR (3, 4, 7, 9)
PAMPs (viral)
RNA
DNA
Other PRRs
A
Mast cell
Intestinal villus
Mast cells
Muscle contraction
Signals to neurones
Recruits basophils and eosinophils by activating endothelium
Mast cells
Intestinal mucosa
Limit inflammation by promoting tissue remodelling and contraction of the immune response
B

FIGURE 7-1 Pathogen-recognition receptors (PRRs) occur on many cell types including mast cells, which use them to regulate innate immune responses. In (**A**) the different types of PRR on mast cells are shown as well as other unique receptors which drive immune responses by the mast cells determined by a level of tissue specificity which act in several ways including cell chemotaxis, vascular permeability increase and mucin production. In (**B**) location of mast cells in lining of the gut is demonstrated. PAMPs, pathogen-associated molecular patterns. (From St John and Abraham, 2013.)

domain of the adaptive immune system, and so they are not contributors to long-term protective immunity. They express the surface antigen CD56 at both high and low levels. In humans they express the KIR ('killer' inhibitory receptor) family of receptors and in mice a range of Ly-49 antigens which allow subsets of NK cells to be identified and which vary between strains of mice. NK cells use inhibitory receptors to prevent them responding to healthy cells which express MHC class I ('self') antigens (see p. 424) but respond to stressed cells which may have low levels of MHC class I (absent self-antigen) and kill such cells using cytotoxic mechanisms.

NKT cells are a further set of innate immune cells of lymphoid origin which also express a single invariant T-cell receptor which binds CD1 antigens (combined with glycolipid antigens, see p. 412), and thus like NK cells they are essentially involved in innate immune responses to microbial lipid, particularly from Gram-negative microorganisms expressing the glycolipid, glucuronyl ceramides.

THE LYMPHOID SYSTEM AND ACQUIRED IMMUNITY

The two most important features of the acquired immune system are exquisite antigen specificity and immunological memory. These are properties of lymphocytes. In contrast to cells of the myeloid system, which remove debris and organisms by mechanisms that have limited specificity (thus macrophages will phagocytose broad ranges of organisms using 'pattern recognition receptors' (PRRs) to recognize PAMPs; see p. 389), each clone of lymphoid cell responds to a single antigen. T cells respond to antigen by proliferating and releasing cytokines, while B cells respond by maturing to plasma cells and producing antibodies.

T cells

T cells (for thymus-derived) are lymphoid mononuclear cells that recognize antigen in conjunction with self-antigen. T helper (Th) cells respond to antigen in association with MHC class II self-antigen, while T cytotoxic (Tc) cells respond to antigen combined with MHC class I antigen (see p. 424). T cells release cytokines, which are required for T-cell and B-cell proliferation and differentiation, and also for innate immune cell activation. T lymphocytes express surface markers (molecules detectable by specific monoclonal antibodies) characteristic of their phenotype. Thus Th cells are described as CD4$^+$ cells and Tc cells are known as CD8$^+$ cells (Box 7-4).

This general categorization of T cells (CD4 and CD8, T cells) was found to be insufficient to explain the numerous different functions of T cells. After many years of controversy over whether a T-cell subset existed which suppressed (regulated) the immune response, the discovery of a T cell which was required to prevent spontaneous autoimmune disease indicated that regulatory T cells (Tregs) occurred in normal healthy individuals.

T cells, particularly CD4$^+$ T cells, are defined by the cytokines they produce and the transcription factors they utilize. Cytokines are multifunctional short-acting, short-range mediators of cellular activities, released by T cells and other immune and non-immune cells (see p. 399). Cytokines are distinguished from other mediators as the molecules of 'intercellular communication'. In addition to CD4$^+$ Tregs, CD4$^+$ T cells now include several subsets such as Th1, Th2, Th17, IL22-secreting CD4$^+$ T cells and it is likely that this T cell specialization will be increasingly recognized (Fig. 7-2). Subsets of Tc cells and B cells (including B regulatory cells) with defined roles in pathogenesis of disease have been described.

Further subsets of T cells occur, such as the γδT cells (which possess a T-cell receptor (TCR) with a γ–δ dimer, rather than the αβ TCR present on conventional T cells) and NKT cells which combine properties of innate immune NK cells but possess an αβ T-cell receptor.

B cells

B cells are mononuclear lymphoid cells that are specialized for the secretion of antibody. There are five types (isotypes) of antibody (IgG, IgA, IgM, IgD and IgE) and B cells require 'help' from T cells to function.

During a primary immune response, IgM antibody is initially produced by activated B cells. In secondary immune responses, the B cells switch to producing IgG (isotype switching), often with higher binding capacity (affinity) for the antigen, a process termed affinity maturation. During allergic immune responses, a further isotype switch occurs from IgG to IgE. IgA forms part of the mucosal immune system, being

BOX 7-4 WHAT ARE CD NUMBERS?

The letters CD mean cluster of differentiation; this term refers to a molecule that has a defined structure, that can be recognized by a group or cluster of monoclonal antibodies, and that identifies a specific lineage or differentiation stage in the cell. CD numbers are continually being defined and the 10th Human Leukocyte Differentiation Antigen (HLDA10) workshop took place in conjunction with the Australasian Society of Immunology in December 2014. At HLDA9 there were 363 CD labelled genes/molecules.

Further details can be obtained through the following link: www.hcdm.org

Some of the better-known CD numbers are shown below.

CD No.	Cell/molecule	Function/role
1	Thymocytes, Langerhans' cells	
3	T cells	
4	T helper cells	Adhesion
8	T cytotoxic cells	
11a	Leucocytes (αL chain of LFA-1)	Adhesion
11b	Leucocytes (αM chain of LFA-1)	Adhesion
11c	Leucocytes (αX chain of LFA-1)	Adhesion
19	B cells	
25	Activated lymphocytes, IL-2 receptor and macrophages	
28	T cells	Co-stimulation
40	Many cells	Co-stimulation
44	T and B cells	Activated memory cells
45	Common leucocyte antigen	
45RA	Naive T cells	
45RO	Memory T cells	
56	NK cells	
62E,L,P	E, L and P selectins	Adhesion
68	Macrophages	
69	Many cells	Short-term activation
95	Many cells	Fas
120	TNF-α receptor	
142	Tissue factor	Coagulation
152	Cytotoxic T lymphocyte antigen-4	Immunoregulation
169	Sialoadhesin	Macrophages
183	CXCR3	Chemokine
284	Toll-like receptor 4	Myeloid cells
332	FGF receptor	growth factor
363	SIP receptor	lymphocyte homing

FGF, fibroblast growth factor receptor; LFA, leucocyte functional adhesion molecule; IL, interleukin; NK, natural killer; TNF, tumour necrosis factor; SIP, sphingosine-1-phosphate.

present in large amounts in surface secretory fluids including tears. IgA is the most abundant immunoglobulin in the immune system. IgD is present in low amounts in the circulation.

B cells recognize and bind antigen via surface immunoglobulin (sIg) which is in effect the B-cell receptor. Antigen binds to sIg in the afferent phase of the secondary immune response; in contrast, antigen binds to the secreted form of immunoglobulin (antibody) in the effector phase of the response. Secondary afferent (cognitive) interactions are antigen-specific, but effector functions are not (see below).

B-cell subsets also exist such as marginal zone B cells and B1 cells, which like γδ T cells and NKT cells lack diversity of αβ T cells.

BLURRING THE MARGINS BETWEEN INNATE AND ADAPTIVE IMMUNITY

Some cells do not fit into this simple scheme separating innate from adaptive immunity. These includes natural killer (NK) cells, NKT cells and lymphoid tissue inducer (LTi) cells, all of which belong to the recently rebranded set of innate lymphoid cells (ILCs). γδT cells and some of the subsets of B cells described above also probably belong here.

Natural killer cells

Natural killer (NK) cells are circulating granulocytes but are part of the lymphoid system and are particularly effective against tumour cells and virus-infected cells. Previously called null cells because they lack any of the specific lymphocyte markers including a T-cell receptor, they recognize and kill virus-infected cells when activated for instance by the cytokine IL-2. Similarly, they have receptors for antibody allowing them to kill antibody-coated cells. NK cells are characterized by inhibitory receptors which bind MHC class I, thus preventing them from killing uninfected healthy cells. In contrast, NKT cells are specialized lymphoid cells that only express a single type of TCR which responds to lipid antigen presented via the molecule CD1 on antigen-presenting cells (thus they are known as invariant NKT cells).

Lymphoid tissue inducer cells

LTis have characteristics of both NK cells and lymphocytes, but are specialized to produce lymphotoxin-α

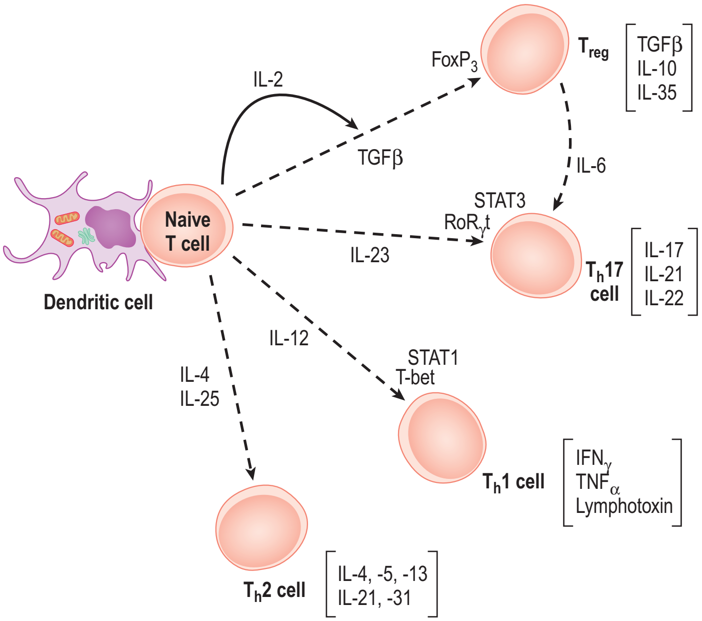

FIGURE 7-2 T-cell differentiation. The dendritic cell activates naive T cells by presentation of antigen. T cells then differentiate into antigen-specific cells with differing properties depending on the cytokines they also experience in the environment where they are activated by the dendritic cell. Four main types of T cell develop: Th1, Th2, Th17 and T-regulatory cells.

and -β (LTα, LTβ) which are required for the development of lymph nodes and other secondary lymphoid tissue.

Innate lymphoid cells

ILCs lack both αβ and γδ T-cell receptors, a feature which could include them in the myeloid system, but have lymphocyte characteristics. Thus NK cells and LTis can be regarded as one type of ILC. However, three further subsets of ILC have recently been identified which secrete cytokines normally associated with CD4 T cells, such as IFN-γ (Th1-like), IL-4 (Th2-like) and IL-22/IL-17 (TH17-like).

Initial response of the host to injury (the innate immune response)

The acute inflammatory response is the host's initial reaction to challenge. During this response, invasive pathogens are removed by cells of the innate immune system brought to the site of injury by changes in tissue components such as blood vessels and extracellular matrix. Meanwhile antigen from degraded microorganisms is transported to lymphoid tissues to activate the acquired immune system.

THE ACUTE INFLAMMATORY RESPONSE

The acute inflammatory response goes through three phases:

- tissue damage and the acute early response
- the delayed cellular response and phagocytosis
- resolution of the inflammation and tissue remodelling.

The acute early phase has several components:

- tissue damage and release of mediators
- vascular changes
- leucocyte activation and adhesion
- leucocyte emigration.

Tissue damage and the release of mediators

The response to tissue injury (physical, chemical or mediated by microorganisms) is immediate. Reactions occur at several levels, both locally and systemically. Immediate local reactions include the release of tissue factors and chemoattractants (chemokines, see p. 404) from damaged tissue and microorganisms. Vessels are also damaged, inducing venous stasis and the leakage of plasma components; platelet and leucocyte activation with intravascular clotting occurs; plasma/serum transudation and exudation lead to tissue fibrin deposition and activation of serum components such as complement (see p. 405).

There are several classes of inflammatory mediators derived from both inflammatory and damaged tissue cells:

- vasoactive amines (e.g. histamine and serotonin)
- cytokines and chemokines
- lipids (e.g. prostaglandins, thromboxane and leukotrienes)
- free radicals (see Ch. 4, Box 4.4-8, p. 193)
- neuropeptides (e.g. substance P, vasoactive intestinal peptide)
- endothelium-derived mediators (endothelin, nitric oxide, prostacyclin, platelet-activating factor, etc.)
- plasma-derived mediators (e.g. complement, kinins and clotting cascade peptides)
- leucocyte-derived mediators (e.g. granule proteases, phospholipase A_2)
- bacterial products (e.g. endotoxin, proteases and chemotactic factors including formylated peptides).

During the first 20 min to 48 h, there is a progressive increase in polymorphonuclear cell infiltration, initiated by innate γδT cell secretion of IL-17, which tend to accumulate around foci of injury or bacterial proliferation (a process known as 'swarming'). Degranulation of these cells leads to high tissue levels of several proteases, cytokines and cationic proteins. Neutrophils contain some of the most powerful antibacterial agents, including defensins, which have similarity to defensins from other species including plants and insects (Table 7-1). Defensins, and the related molecule cathelicidin (LL37), are leucocyte-derived low molecular weight mediators of non-oxidative bacterial killing and are also produced by epithelial cells such as gut and conjunctival mucosal cells. Defensins also direct antiviral activity working on both the virion and the host cell. Among the proteases released by neutrophils during acute inflammation are enzymes that degrade the extracellular matrix, which are also released by activated or injured tissue parenchymal cells and are involved in tissue remodelling. More than 24 of these zinc-dependent endopeptidases, known as matrix metalloproteinases (MMPs) (Table 7-2), have been identified in humans, and include the transmembrane proteins that contain disintegration and metalloprotease domains (ADAMTs) (Fig. 7-3). The MMPs self-activate and cross-activate each other in a cascade-like fashion, thus permitting maximal tissue degradation as required. Their effects are counteracted by naturally occurring inhibitors, termed tissue inhibitors of matrix metalloproteinases (TIMPs). There are four types of TIMP: 1, 2, 3 and 4, TIMP-4 only occurring in the mouse. MMPs are also inhibited by recognized anti-proteases such as α_2-macroglobulin, tissue factor pathway inhibitor 2, and a more recently described plasma membrane inhibitor, RECK (reversion-inducing cysteine-rich protein with Kazal motifs) which appears to be downregulated in cancer cells, thus allowing them greater metastatic potential.

TABLE 7-1 **Neutrophil antibacterial agents**

Class	Agent
Free radicals/gases	Hydrogen peroxide Hypochlorite Chloramine OH radical Nitric oxide
Enzyme	Proteinase 3 Collagenase Elastase Azurocidin Cathepsin G β-glucuronidase Myeloperoxidase Lysozyme
Peptide	Defensin β-lysin Vasoactive intestinal peptide
Ion binders	Lactoferrin Calprotectin

However, much of the cell and tissue damage is mediated by free radicals, particularly hydrogen peroxide (H_2O_2) and superoxide anions (see Ch. 4, p. 193), which are released as part of the respiratory burst (Box 7-5). Interestingly flavonoids and adenosine, both important molecules in ocular physiology, inhibit the respiratory burst.

Free radicals may also combine with reactive nitrogen species (nitric oxide) released from inflammatory cells (Box 7-6).

TABLE 7-2 Matrix metalloproteinases (MMPs) and their substrates

MMP	Interstitial collagens	Basement membrane	Elastin	Other proteins
Collagenases				
MMP-1	Types III, I, II, VII, X	Laminin		
		Entactin		
MMP-8	Types I, III, II	Fibronectin		L-selectin
MMP-13	Types II, I, III	(±) Proteoglycan		
Stromelysin				
MMP-3		Fibronectin	±	EGF-like proteins
		Laminin		
MMP-10		Entactin	±	Plasminogen
		Proteoglycan		
Stromelysin-like				
MMP-7		Fibronectin	+	Plasminogen
MMP-12		Laminin		↓
		Entactin	++	Angiostatin;
		Proteoglycan		α_1-antitrypsin
Gelatinases				
MMP-2	Types I, VII, X, XI	IV/V	++	
		Fibronectin		
MMP-9		Laminin	++	
		Entactin		
		Proteoglycan		
Membrane-type				
MMP-14	Types I, III, II	Fibronectin		
MMP-15		Laminin		
MMP-16		Entactin		
MMP-17		Proteoglycan		
MMP-24				
Furin-recognition site				
MMP-11				

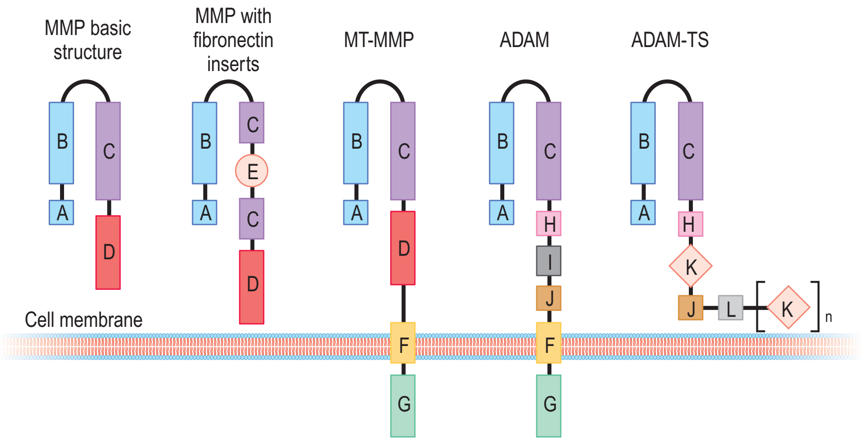

FIGURE 7-3 Domain structure of MMPs also known as metzincin proteases. Each of the domains is shown as rectangles identified by letters, as follows: *A* signal peptide, *B* prodomain, *C* catalytic domain, *D* hemopexin-like domain, *E* fibronectin type II insert, *F* transmembrane domain, *G* cytoplasmic tail, *H* disintegrin domain, *I* cysteine-rich domain, *J* EGF-like domain, *K* thrombospondin type I-like repeat, *L* spacer region. (From Klein and Bischoff, 2011.)

BOX 7-5 THE NEUTROPHIL RESPIRATORY BURST

The neutrophil respiratory burst describes the activation of neutrophils and their utilization of oxygen during the inflammatory cascade. The stimulus for this response is the release of mediators from injured tissue cells. Tissue damage is caused by free radical release and tissue proteases (see also Ch. 4).

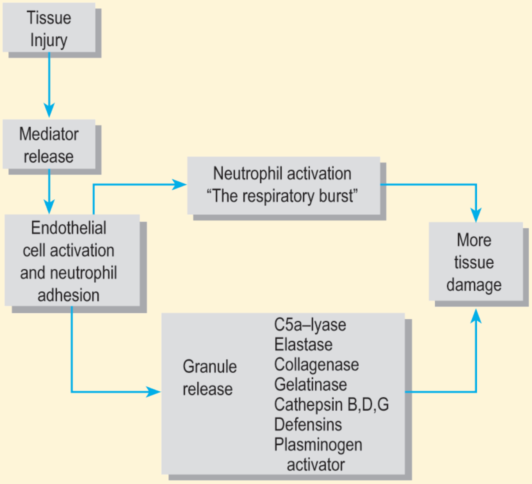
Tissue Injury
Mediator release
Neutrophil activation "The respiratory burst"
Endothelial cell activation and neutrophil adhesion
More tissue damage
Granule release
C5a–lyase
Elastase
Collagenase
Gelatinase
Cathepsin B,D,G
Defensins
Plasminogen activator

The central component of the respiratory burst is H_2O_2, which is metabolized through several pathways, some of which cause further tissue damage (e.g. superoxide and chloramine) while others reduce it to water. Tissue damage is therefore dependent on the levels of reduced glutathione (GSH) in the milieu. In addition, during the respiratory burst, immune cells activate a $Na^{(+)}/H^{(+)}$ membrane exchanger protein which is required to control intracellular pH and cell volume. NADP, nicotinamide adenine dinucleotide phosphate; NADPH, reduced NADP; PKC, protein kinase C.

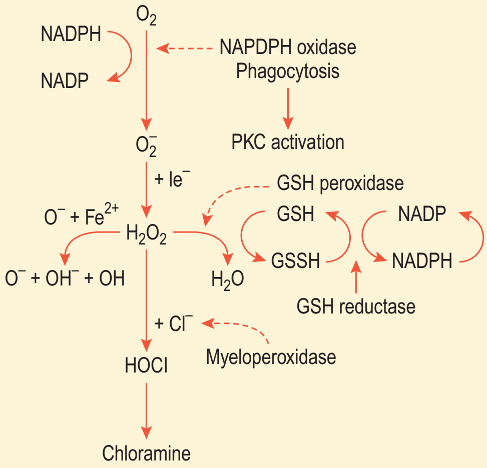
O_2
NADPH
NADP
NAPDPH oxidase
Phagocytosis
O_2^-
PKC activation
+ le$^-$
GSH peroxidase
O^- + Fe^{2+}
H_2O_2
GSH
NADP
GSSH
NADPH
O^- + OH^- + OH
H_2O
GSH reductase
+ Cl^-
Myeloperoxidase
HOCl
Chloramine

Hypochlorous acid (HOCl) is a short-lived, highly reactive oxidant that is lipophilic and membrane permeant. It binds proteins and renders them more susceptible to proteases. Chlorinated proteins are more immunogenic and may provide a link between the innate and the acquired immune systems by acting as 'DAMPs'.

BOX 7-6 NITRIC OXIDE

Nitric oxide was originally described as endothelium relaxing factor because it was found to be the agent released by the endothelium that was responsible for inducing autocrine vasodilatation in response to insult. Nitric oxide is a gas produced by the activity of the enzyme nitric oxide synthase (NOS) on interaction of the amino acid arginine with oxygen:

$$O_2 + \text{L-arginine} \xrightarrow[\text{NOS}]{} \text{Citrulline} + NO$$

There are at least three isoforms of nitric oxide synthase, endothelium-derived (eNOS), neuronal (nNOS) and inducible (iNOS). The iNOS is released from inflammatory and other cells, particularly macrophages, and is involved in both immunoregulation and tissue damage through its interaction with superoxide radicals released from activated neutrophils. In its latter role it may also function as an antibacterial agent by damaging bacterial cell membranes. The prostanoids are co-released with NO through induction of cyclo-oxygenase-2.

NO interacts with superoxide anions to produce per-oxynitrite, which is believed to be directly involved in membrane damage (see equation below). NO has direct effects on many other proteins such as the important zinc finger regions of many enzymes where it nitrosylates free cysteine SH groups with ejection of the Zn moiety from the configured protein. These damaging effects to parenchymal cells may also be the basis for T-cell apoptosis (cell death), so leading to downregulation of the immune response.

$$O_2^{\bullet} + NO^{\bullet} \rightarrow ONOO^- + H^{\bullet} \rightleftharpoons ONOOH \rightarrow HO^{\bullet} + NO_2^+ \\ \rightarrow NO_3^- + H^+$$

Thus NO produced constitutively in small amounts has a physiological role involving guanyl cyclase and increases in cGMP, while in large amounts it may be cytotoxic, producing depression of mitochondrial respiration, metal enzyme damage with consumption of both Zn and Fe ions and DNA damage.

Vascular changes

The immediate cause of inflammation is release of plasma into the extravascular space and the instantaneous coagulation of proteins with activation of inflammatory mediators. Plasma release (vascular leakage) is caused by changes to the blood vessels induced by the inflammatory mediators, particularly reactive nitrogen species (RNS), especially in conditions such as the ischaemia–reperfusion damage of stroke and vascular occlusion. The immediate response of the vascular endothelium is to undergo retraction and this is associated with transient vasoconstriction. The major vascular response, however, is vasodilatation mediated initially by nitric oxide (NO). The initial vasoconstriction is mediated by several locally released compounds, particularly endothelin, which is released by pericytes and smooth muscle cells to act on the endothelium. The later vasodilatation is also mediated by locally released factors, in this case the gas nitric oxide. NO is synthesized by specific enzymes from the amino acid arginine and has widespread physiological and pathological effects, some of which are related to its role as a free radical (see Box 7-6). This is accompanied by an increased blood flow, the opening of capillary channels and the leakage of plasma into the extracellular space. This in turn leads to an increase in the tissue osmotic pressure, thus attracting further fluid build-up in the tissues (oedema). In response to this, there is an increase in lymphatic drainage from the injured site, thus reducing the tissue swelling and at the same time increasing the flow of antigenic material to the draining lymph nodes (see below). These vascular changes vary in degree with the severity of the tissue injury and protective measures are often in place such as the expression of caveolin-1 in vessel walls which mitigates the increased vascular permeability.

The vascular endothelium also undergoes significant functional and morphological changes during the inflammatory response. Whereas the normal endothelium presents a non-adhesive surface to circulating cells such as platelets and leucocytes, during inflammation the endothelium becomes much more adhesive, an effect achieved by the expression of specific adhesion molecules on its surface. There are three major classes of adhesion molecule, the selectins, the integrins and the cell adhesion molecules (CAMs), each with different functions (see below) (Fig. 7-4 and Table 7-3). In addition, the endothelium may undergo a marked morphological transformation, changing from a flat resting cell to a large protruding cell with multiple cytoplasmic organelles. These inflammation-associated vascular changes occur almost exclusively in the postcapillary venule and, in this respect, this region resembles the high endothelial venule cells in the lymph node, which are specialized for leucocyte adhesion (see below).

Leucocyte activation and adhesion to the endothelium

Neutrophils are attracted to the site of inflammation through a series of discrete events that occur during margination and extravasation of the cell from the vessel. These involve rolling, loose attachment, firm adhesion and then extravasation/migration of the cell through the intercellular junction. Each of these steps is mediated by the reciprocal expression of adhesion molecules and their respective ligands on the surface of leucocytes and the endothelium (see below) (Table 7-3). During the later stages of the response (24–72 h) when other inflammatory cells are involved (monocytes and lymphocytes), similar adhesion mechanisms are involved but with different sets of molecules. Thus, the coordinated expression of adhesion molecules appears to regulate the nature of the inflammatory cell exudate.

Adhesion of leucocytes to the endothelium thus involves a series of molecular events

- Selectin–ligand (S–L) interactions occur during the initial rolling phase of leucocyte endothelial cell interactions. These are initially low-strength interactions and are enhanced by the upregulation of selectins on the endothelium by inflammatory cytokines such as interleukin-1 (IL-1) and TNF-α or by contact with an activated T cell.
- Leucocyte activation by chemokines is mediated in part through upregulation of specific chemokine receptors (see later) which induces polarization and firm adhesion of the cell to the endothelium.
- Integrin–CAM interactions induce spreading of the leucocyte on the endothelial cell surface and prevent detachment of the leucocyte.

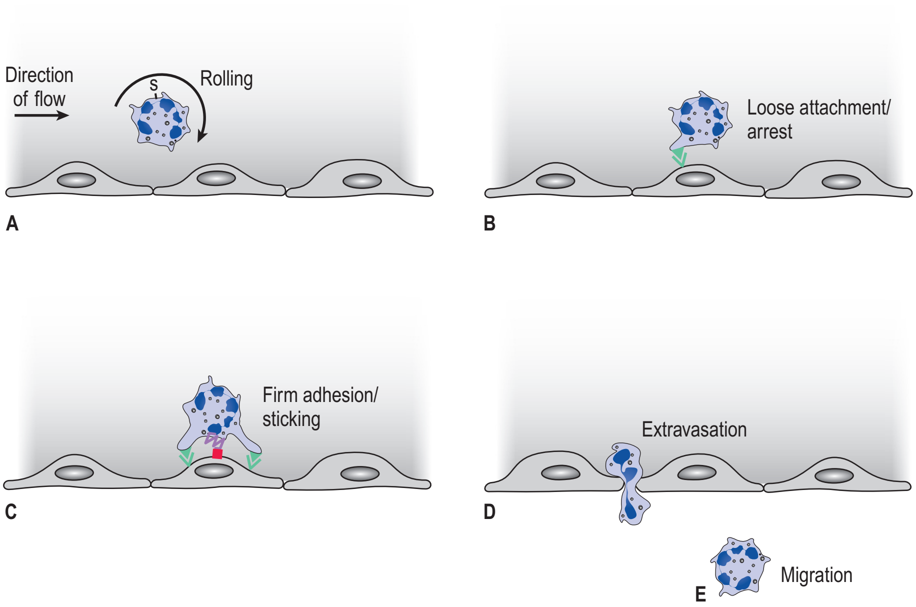

FIGURE 7-4 Sequence of events in leucocyte adhesion: rolling (**A**), loose attachment (**B**), firm adhesion (**C**), extravasation (**D**), leading eventually to migration (**E**) of the cells within the tissue towards the site of inflammation.

- Extravasation of leucocytes through the endothelium is mediated by expression of PECAM-1 (CD31) on both the leucocyte and the endothelium, possibly through a 'zipper' mechanism in which disassembly of intercellular tight (occludin) and adherens junctions occurs with the expression of several junctional adhesion molecules (JAMs; see Table 7-3). CD99, expressed on both leucocytes and endothelium, is also important in transendothelial migration of monocytes.
- Migration of leucocytes through the tissues is the final stage and is induced through binding of chemokines selective for each cell type which activate signalling pathways in the translocating cell activating the motor machinery (actin–myosin cytoskeleton) to propel the cell forward up the chemotactic gradient (Fig. 7-5).

Certain molecules are specific for each type of leucocyte–endothelial cell interaction; for instance, E- and P-selectins mediate the attachment of polymorphonuclear leucocytes to endothelial cells, while vascular cell adhesion molecule (VCAM) preferentially mediates T lymphocyte–endothelium binding. Both of these interactions have been reported in inflammatory tissue in the eye from cases of sympathetic ophthalmia, a form of autoimmune posterior uveitis.

Leucocyte migration into the tissues and chemotaxis

Many of the mediators released in the earliest stages of tissue injury are attractants for inflammatory cells. Both thrombin and the cleavage product fibrinopeptide B from fibrin lead to leucocyte chemotaxis from the onset of vessel leakage and fibrin formation. Prokaryotic peptides released from bacteria, such as formyl-methionine-leucine-phenylalanine, are powerful neutrophil and monocyte chemotactic agents. Activated complement components (see p. 405) have an important role as chemoattractant agents in the neutrophil/monocyte response. Other important chemoattractants include interleukin-8 (IL-8), a cytokine (chemokine) released from tissue cells including the retinal pigment epithelium (RPE), tumour necrosis factor-α (TNF-α) and platelet release compounds such as platelet-activating factor (PAF), transforming

TABLE 7-3 **Endothelial adhesion molecules and their ligands on leucocytes**

	Endothelium	Leucocyte
Loose attachment	P-selectin	PSGL-1, PTX3
	E-selectin	PSGL-1, ESL, CD44
	PSGL, GlyCAM	L-selectin
Slow rolling	ICAM-1	LFA-1/PSGL-1
	E-selectin	PSGL-1, ESL, CD44
Arrest/firm adhesion	ICAM-1	LFA-1
	VCAM-1	VLA-4
	Hyaluronan	CD44
Crawling	ICAM-1	MAC-1
Extravasation	ICAM-1, ICAM-2	LFA-1, MAC-1
	VCAM-1	VLA-4
	CD99	CD99
	PECAM-1	PECAM-1
	JAM-A, -B, -C	LFA-1, VLA-4, MAC-1
	CD99L2	?

The cell behavioural sequence of loose attachment, slow rolling, arrest and firm adhesion, crawling along the inner vascular wall, and finally extravasating into the tissues is utilized by neutrophils which preferentially if not exclusively leave from the postcapillary venules.

? indicates unknown or opposing data. CD99L2, CD99 antigen-like protein 2; ESL1, E-selectin ligand 1 (also known as GLG1); GlyCAM, glycosylation-dependent cell adhesion molecule; ICAM, intercellular adhesion molecule; JAM, junctional adhesion molecule; LFA-1, lymphocyte function-associated antigen 1; PECAM-1, platelet/endothelial cell adhesion molecule 1; PSGL-1, P-selectin glycoprotein ligand 1; VCAM-1, vascular cell adhesion protein 1; VE-cadherin, vascular endothelial cadherin; VLA-4, very late antigen 4.

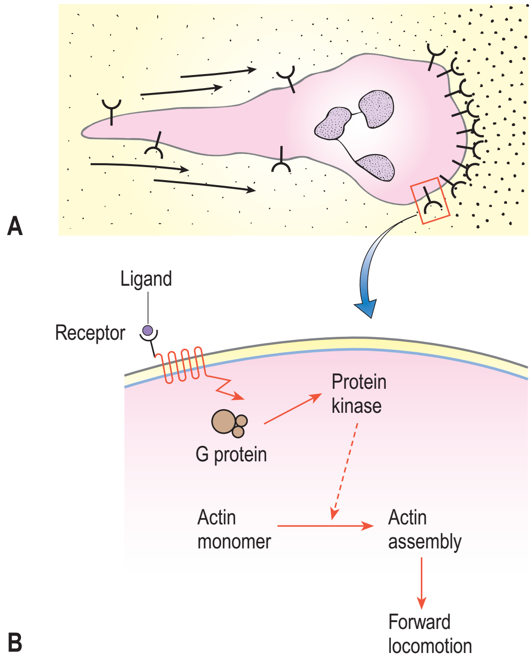

FIGURE 7-5 Neutrophil migration up a chemical gradient. (**A**) Cells sense the concentration gradient and polarize towards it with a wide leading edge and a trailing tail. Receptors for the chemoattractant cluster in the membrane of the leading edge, increasing the signal at this end of the cell. (**B**) Diagram of the seven-coil transmembrane receptor (the same basic structure as rhodopsin) that mediates the intracellular signal via Gprotein links. This activates a protein, tyrosine kinase, to initiate actin assembly and induce forward movement of the cell.

growth factor-β (TGF-β), platelet-derived growth factor, platelet-derived endothelial cell growth factor, and many others. Lipid mediators, such as the leukotrienes, are also important neutrophil chemoattractants, while certain cytokines, such as monocyte chemotactic protein (MCP) and the macrophage inflammatory proteins (MIP-α, MIP-β and the chemokines; see below) are selective for mononuclear cell chemotaxis. Cytokines that induce inflammatory cell migration are known as chemokines (inducing movement through chemotaxis) and receptors for different chemokines are present on different cells, thus regulating not only the numbers of cell that migrate into the tissue but also the type of cells and thus the quality of the inflammatory exudate (see p. 375).

Cells such as neutrophils and monocytes ‘sense’ a chemical gradient of these attractants and migrate up the gradient by using specific cell surface receptors clustered preferentially towards the leading, polarized edge of the cell (Fig. 7-5). These receptors (e.g. the C5a receptor) are composed of seven transmembrane segments (in a manner similar to the transmembrane spanning segments of rhodopsin; see Ch. 4, p. 261.e1) that possess a cytoplasmic connection to a Gprotein-linked second messenger system. This activates the intracellular machinery (actin–myosin motor) required for forward movement. Recent studies have also shown that activation of G protein receptors occurs in waves, thus enhancing the overall effect of the gradient.

Phagocytosis and removal of damaged tissue and microorganisms

Recovery from inflammation requires the removal of dead microorganisms and necrotic tissue by phagocytic cells (polymorphonuclear leucocytes and macrophages). Even in the absence of microorganisms, altered (damaged) self-proteins are recognized by cells of the innate immune system and phagocytosed. In the eye this is classically seen with lens-induced uveitis, in which denatured lens crystallin proteins are released into the anterior chamber in cases of traumatic and hypermature cataract. In the latter circumstance, engorged macrophages may block the outflow channels and produce a 'phacolytic glaucoma' (see Ch. 9, p. 512).

Phagocytosis, particularly of bacteria, is facilitated by certain molecules of the innate immune system known as opsonins, which are present in plasma and bind to the surface of microorganisms when they are released into the extracellular space. One of the complement components, C3b (see below), acts as an opsonin. Interestingly, different actin cytoskeletal structures are constructed to phagocytose different types of particle, e.g. IgG-coated versus complement-coated particles. In addition, the role of microbial pathogen receptors such as Toll-like receptors (see next section) on maturation of the phagosome is unclear but appears not to play a significant part.

Molecules of the acquired immune system also promote phagocytosis, particularly antibodies which bind avidly to the foreign antigen by specific interaction of their antigen-binding site (the Fab portion of the molecule) but are non-specifically removed by binding of their Fc portion to the phagocyte cell surface (see section on antibodies below, p. 396). Cells that express high levels of surface receptors for C3b and Fc are termed 'professional phagocytes', and include neutrophils and macrophages. Studies in genetically targeted knockout mice suggest that the Fc pathway for phagocytic activation is the major one.

Activation of innate immune cells by microorganisms

Invading microorganisms release factors that attract leucocytes to the site of inflammation and also induce those same leucocytes to engage in attempted removal of the offending agent. This requires activation of the immune cells. How is this activation of the innate immune response achieved? As indicated in the introduction, this was previously considered to occur by non-specific pathogen–host cell interaction, i.e. there was one general innate immune mechanism whereby the body reacted to many different types of microorganisms. However, it has long been recognized that responses to different organisms vary greatly, some being virulent or lethal while others are harmless, and this is partly due to how different classes of microorganisms activate the initial innate immune cells they encounter. This they do via broadly specific ligands (PAMPs) which bind to pathogen-recognition receptors (PRRs) (see p. 391). There are several classes of both membrane and soluble PRRs, including the Toll-like receptors (TLR) (Table 7-4), carbohydrate-binding (C)-type lectins, retinoic acid-inducible gene-1 (RIG-1) helicases, nucleotide-binding oligomerization domain protein (NOD)-like receptors (NLRs), the scavenger receptors and some soluble PAMPS such as the collectins (including complement proteins) and acute phase proteins. The components (PAMPs, DAMPs) recognized by the PRRs include pathogen cell wall material, bacterial DNA and proteins, viral DNA and RNA, and lipoproteins from microbes as well as endogenous ligands such as self-DNA (if it is in the wrong place, such as the cytoplasm) and particulate material such as uric acid and cholesterol crystals.

Lipopolysaccharide is a classic microbial product that complexes with a serum protein lipopolysaccharide-binding protein and binds to the microbial co-receptor CD14. This then complexes with TLR4 and initiates the signalling cascade (most involving the adapter protein myeloid differentiation factor 88 (Myd88) via several intermediaries transmitted on a nuclear factor-κB (NF-κB) core to activate the macrophage or dendritic cell (Fig. 7-6) and subsequently primes T cells through presentation of processed antigen in the presence of the necessary cytokines such as IL-12, IL-23 and IL-17.

TLRs are not so much involved in the phagocytosis and clearance of microorganisms (see section above) as in the maturation of antigen-presenting cells for induction of the adaptive immune response through T-cell activation; for this, different TLRs

TABLE 7-4 **Toll-like receptors – mediators of inflammation with relative selectivity for different classes of microorganisms**

TLR	Localization	Pathogen-derived agonists	Endogenous agonists	Synthetic agonists
TLR1 and TLR2	Extracellular	Bacteria: peptidoglycan, lipoproteins, LTA Fungi: zymosan	–	Pam$_3$Cys
TLR2 and TLR6	Extracellular	Bacteria: lipoproteins	Veriscan	MALP2
TLR3	Intracellular	Viruses: dsRNA	mRNA	PolyI:C
TLR4	Extracellular	Bacteria: LPS Viruses: RSV fusion protein Fungi: mannan Protozoa: glycoinositolphospholipids	Saturated fatty acids, β-defensins, oxLDL*, amyloid-β*	Lipid A derivatives
TLR5	Extracellular	Bacteria: flagellin	–	–
TLR7 and TLR8	Intracellular	Viruses: ssRNA	Self-RNA	Imiquimod, R-848
TLR9	Intracellular	Bacteria: CpG DNA Viruses: CpG DNA Protozoa: CpG DNA, haemozoin	Self-DNA	CpG-ODNs
TLR11	Extracellular	Uropathogenic bacteria Protozoa: profilin-like molecule	–	–

CpG-ODNs, CpG-containing oligodeoxynucleotides; dsRNA, double-stranded RNA; LPS, lipopolysaccharide; LTA, lipoteichoic acid; MALP2, mycoplasma macrophage-activating lipopeptide 2; oxLDL, oxidized low-density lipoprotein; polyI:C, polyinosinic-polycytidylic acid; RSV, respiratory syncytial virus; ssRNA, single-stranded RNA. *Amyloid-β and oxLDL bind to CD36 and a TLR4–TLR6 heterodimer. *(From Mills, 2011.)*

activate different types of dendritic cell (see eFig. 7-2). In contrast, macrophages also express other PRRs, including scavenging receptors, which are primarily involved in phagocytosis. Two main classes of scavenging receptors occur in humans: SR-A, which binds molecules on ageing damaged (e.g. oxidized) cells, such as oxidized low-density lipoprotein, as well as some microorganisms: and SR-B, which occurs on other cells as well as macrophages and binds many different types of microorganisms through polyanionic interactions. TLRs are, for the most part, cell surface receptors, although some, such as TLR3 and TLR9, bind intracellular molecules such as viral genomic messenger RNA and nucleotide degradation products such as CpG molecules. In addition, a further set of molecules deals with intracellular pathogen products, the NOD 1 and NOD 2 proteins (nucleotide-binding oligomerization domain family), which like the TLRs have a leucine-rich repeat domain. These proteins are also known as CARD (caspase recruitment domain) proteins and bind unique peptidoglycan bacterial molecules. They are also involved in autoimmunity, especially the inflammatory bowel disease Crohn's disease, which is linked to certain types of ocular inflammation. Activation of an 'inflammasome' complex (see Fig. 7-6) and caspase-1 leads to production of IL-1 and IL-18, which drive the inflammatory process. Interestingly, these molecules have recently been implicated in the pathogenesis of age-related macular degeneration.

NK, NKT and γδT cells also express PRRs and mediate innate immunity

NK cells recognize self and non-self ligands on virus-infected cells, tumour cells and host cells that have increased levels of stress proteins. MHC class I and MHC class I-like molecules (see below) are involved in these interactions. Using receptors such as Ly49H, natural cytotoxicity receptors and a further receptor termed NKG2D, NK cells recognize a range of virus-infected cells such as cytomegalovirus, myxoviruses and influenza viruses. NKG2D is a receptor for tumour proteins and the MHC class I molecules that they recognize in combination are known as MICA and MICB.

NKT cells, in contrast, express the T-cell receptor (see p. 429) but unlike normal T cells they only express

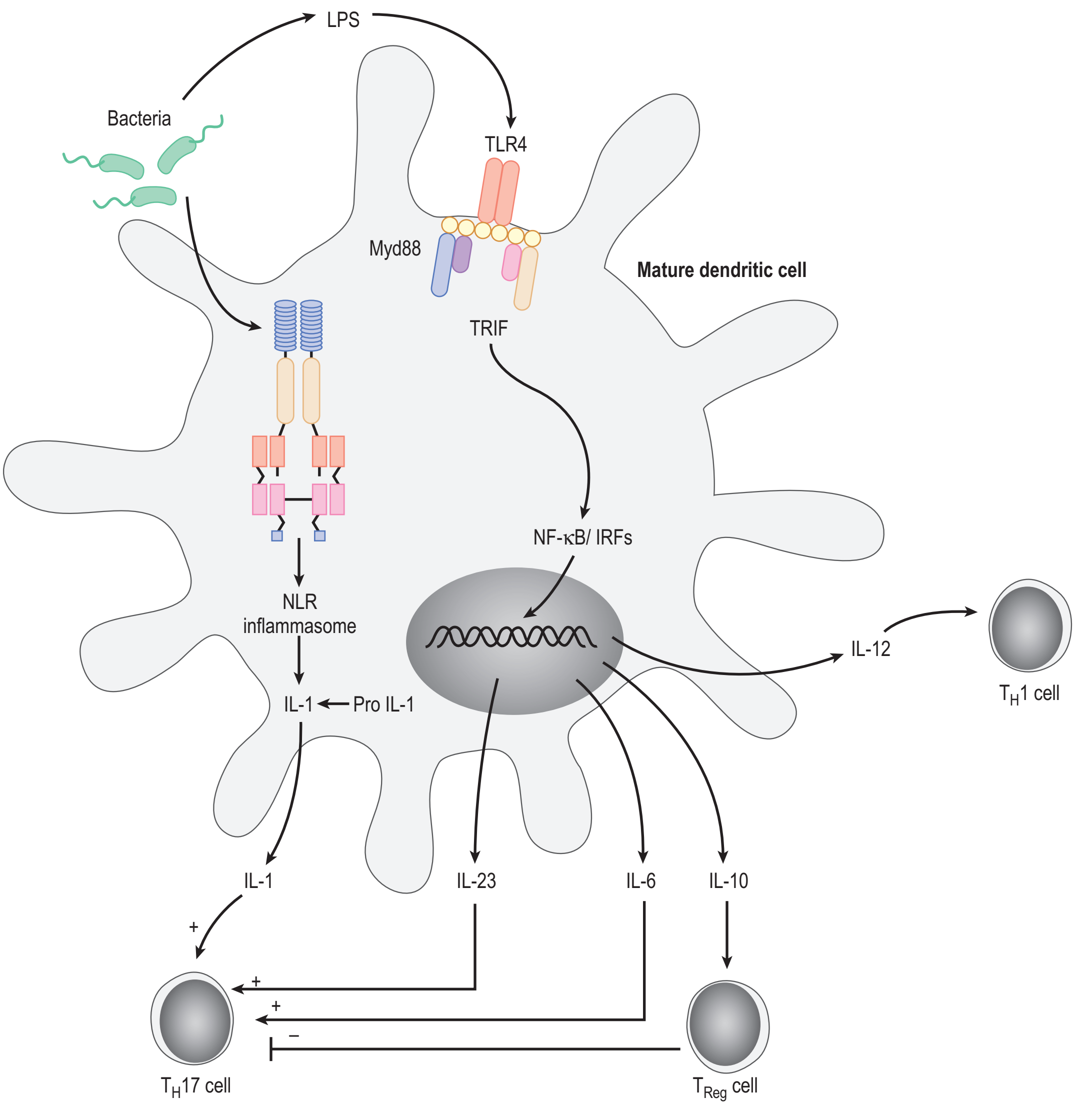

FIGURE 7-6 Toll-like receptors (TLRs) act on myeloid and other cells by activation of transcription factors but they do not do this directly. Instead they do so through linker proteins (adaptors), which fit them to their particular pathway. All TLRs except TLR3 use the adaptor Myd88 while other TLRs use more than one adaptor, such as TRIF, TIRAP and TRAF. These generate cytokines through two main pathways, one involving NFκB, which leads to transcription of many pro-inflammatory cytokines and the other IRF3, which predominantly leads to production of INF-α. Both signalling pathways also lead to inflammasome activation with production of the central pro-inflammatory cytokines IL-1β and IL-18. Ultimately, these molecules assist in the activation of T cells, during antigen presentation. (From Mills, 2011.)

one form of the receptor and use this to bind glycolipids on a range of organisms such as *Leishmania donovani*, Gram-negative glycosyl ceramides and similar molecules on organisms such as *Plasmodium* and *Trypanosoma*. Two important ocular microorganisms, *Pseudomonas* and *Staphylococcus*, may also be detected by NKT cells. NKT cells use a non-classical MHC class I molecule, CD1d, to mediate these interactions. The cytolytic function of NK and NKT cells occurs via the release of granule contents such as perforin, granzyme and other proteoglycans (see later under cytotoxic T cells).

γδT cells colonize the skin and gut epithelium and recognize small alkylamines and pyrophosphomonoesters on microbes and tumour cells. The latter are a major component of mycobacteria. In addition, they recognize heat-shock proteins and have been implicated in inflammatory diseases such as Behçet's disease and the ocular inflammation associated with it.

Effector cells in the inflammation response

Much of the tissue damage in the early stages of the inflammatory response is caused by release of tissue-degrading enzymes such as MMPs (see Table 7-2) from professional phagocytes and as such they are considered to be important as non-specific effector cells in the inflammatory response. Activation of complement during the early phase of the response also provides the materials for non-antibody-dependent complement-mediated lysis of tissue cells, via the membrane attack complex (MAC) of complement and other cells such as NK cells. Macrophages, recruited by T-cell cytokines, also play a major role as effector cells in the adaptive immune response, in addition to antigen-specific cytolytic roles of T cytotoxic (Tc) cells and NK cells.

Lymphocytes also play a significant part in the initial acute inflammatory response. Even in 'sterile wounds', in which defence against microorganisms does not play a major part, lymphocytes participate in the overall response. Lymphocytes enter into the site of inflammation at the same time as the monocytes, and act as bystanders ready for activation by APCs primed with antigen. This may be from degraded organisms or denatured tissue proteins ('altered self'). In most cases, no detectable adaptive immune response occurs locally unless there is activation of tissue-resident memory T cells (see p. 420), specific for a previous pathogen or antigen.

RESOLUTION OF INFLAMMATION

Resolution of inflammation occurs when the foreign antigens have been completely destroyed and removed ('cleared'), and tissue architecture is restored. Macrophages are critically important in the resolution phase of an inflammatory reaction. Effector pro-inflammatory M1 macrophages (see p. 377 and eFigs 7-1 and 7-2) which are involved in clearance of microorganisms and in the process generate considerable collateral damage ('friendly fire') are replaced by alternatively activated macrophages which promote healing (M2 macrophages) either by *de novo* recruitment or by re-programming of inflammatory macrophages. Such cells express high levels of arginase-1, characteristic of inhibitory immune cells, such as myeloid suppressor cells (MSCs) with high potency in inhibiting effector T cells.

Resolution of inflammation involves fibrosis, which is also mediated via alternatively activated M2 macrophages and regulatory macrophages through mediators derived from other cells in the inflammatory milieu (Fig. 7-7). In addition, new vessel formation (angiogenesis), restoration of epithelial surfaces by cell migration and proliferation (see Ch. 4, p. 211, cornea), and remodelling of the extracellular matrix by initial deposition of sulphated glycosaminoglycans and hyaluronan deposition, induced by fibroblasts, represent stages in a precisely orchestrated wound-healing process. For instance, angiogenesis is initiated in quiescent endothelial cells by the expression of protease activity on the cell surface, and release of growth factors from surrounding inflammatory cells such as fibroblast growth factor, platelet-derived growth factor platelet-derived endothelial cell growth factor and vascular endothelial growth factor-A (VEGF-A). VEGF is a major player in the overall wound-healing response: it initiates vascular leakage at the onset of inflammation and promotes angiogenesis in the later stages. The VEGF-C isoform of VEGF activates a specific endothelial cell receptor (VEGF-R3) for induction of lymphangiogenesis, essential for transport of soluble and cell-associated antigen to the draining lymph node.

VEGF induction is itself under the control of hypoxia-inducible factor (HIF1α and β), a transcription factor generated by cells in hypoxic tissue, as occurs in a wound, but is also produced by both M1 and M2 macrophages in the absence of hypoxia. However, VEGF-A attracts 'immunosuppressive' cells such as CCR2$^+$ M2 macrophages which secrete the cytokine CCL2, MSCs, and Tregs and Bregs, all of which promote further angiogenesis. Thus wound healing, acute inflammation, innate immunity and adaptive immunity are all part of a coordinated

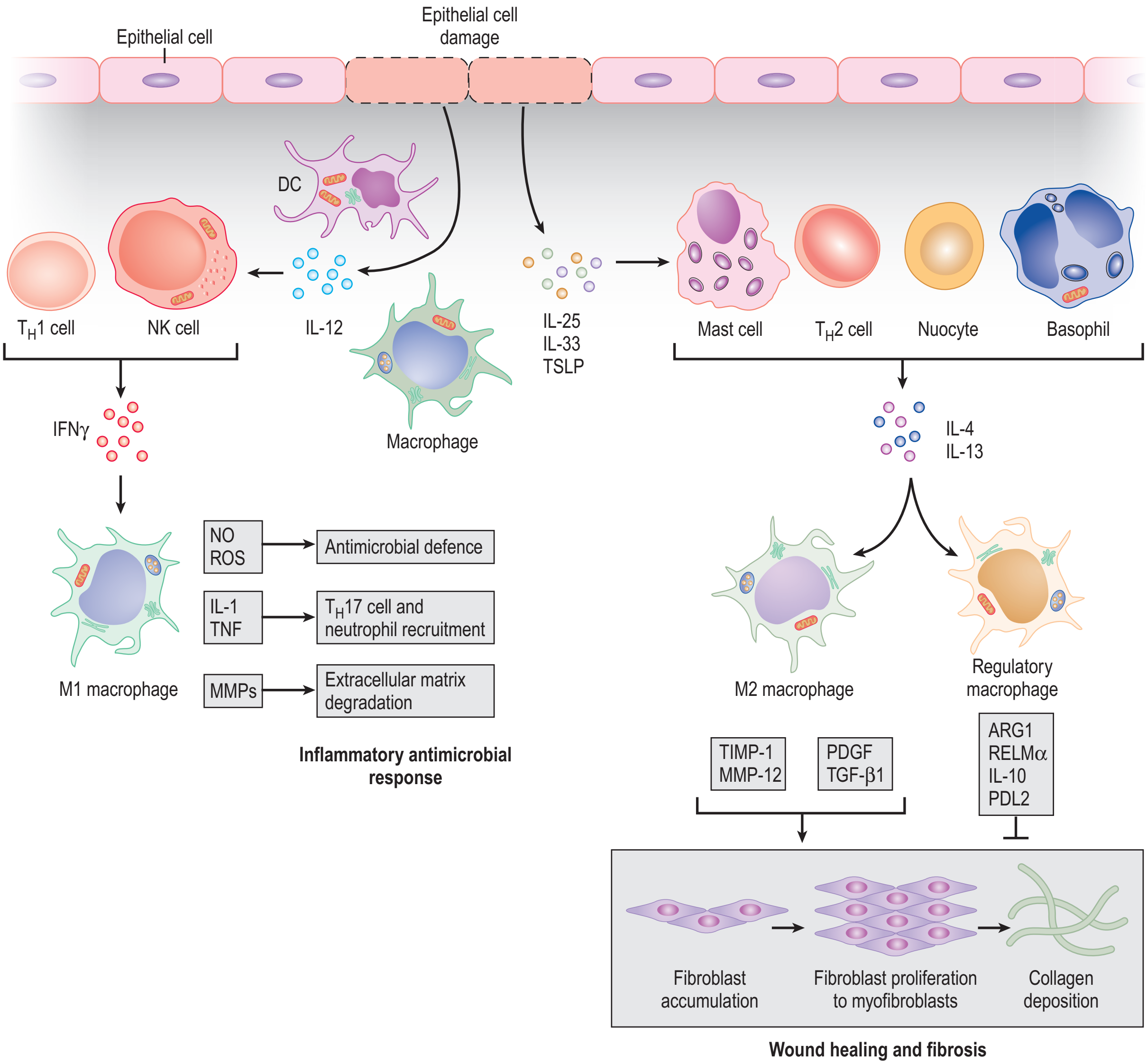
Epithelial cell
Epithelial cell damage
DC
IL-12
T_H1 cell
NK cell
IFNγ
Macrophage
IL-25
IL-33
TSLP
Mast cell
T_H2 cell
Nuocyte
Basophil
IL-4
IL-13
M1 macrophage
NO
ROS
Antimicrobial defence
IL-1
TNF
T_H17 cell and neutrophil recruitment
MMPs
Extracellular matrix degradation
Inflammatory antimicrobial response
M2 macrophage
Regulatory macrophage
TIMP-1
MMP-12
PDGF
TGF-β1
ARG1
RELMα
IL-10
PDL2
Fibroblast accumulation
Fibroblast proliferation to myofibroblasts
Collagen deposition
Wound healing and fibrosis

FIGURE 7-7 The initial stages the inflammatory response involves many early mediators including clotting factors, fibrinolytic products, cytokines such as IFN-γ which programme macrophages towards a pro-inflammatory phenotype (M1). Th2-type cytokines (IL-4, IL-13), released from other cells such as mast cells, generate alternatively activated macrophages as well as regulatory macrophages which promote healing (fibroblasts and collagen deposition) as well as new vessel formation (From Murray and Wynn, 2011.)

response by the organism to remove the offending microorganism and restore tissue homeostasis.

CHRONIC INFLAMMATION

When the foreign antigen is not completely removed, the inflammatory response enters a chronic state characterized by mononuclear inflammatory cells such as monocytes and lymphocytes, often arranged in granulomas (see Ch. 9, p. 502). Failure to remove the foreign antigen may occur because of an inadequate initial response or because the antigen effectively evades the immune system. Intracellular bacteria that evoke phagocytic destruction are a prominent stimulus for induction of chronic inflammation. Persistence of innate immune cells such as neutrophils is a feature of many chronic inflammations and signals a failure to clear the foreign antigen.

However, even if the organisms are cleared, chronic inflammation can ensue in the absence of active processes which lead to the production of specialized pro-resolving mediators (SPMs) such as omega-3 fatty acid-derived lipoxins (see Ch. 4, p. 197), and other more recently described molecules such as resolvins, protectins and maresins. These mediators inhibit leucocyte migrations and suppress effector innate and adaptive immune cells.

In chronic inflammatory disorders, the acquired immune response also participates, but it too appears to be insufficient to remove the foreign antigen. This may be because the antigen has 'fooled' the immune system by parasitizing the inflammatory cells, as in parasitic disease including the worldwide blinding disease *Chlamydia trachomatis* (see Ch. 8). In autoimmune disease (see p. 444), 'altered self' antigen is continually present, acting as a danger signal, and induces persistent inflammation. In certain circumstances a low-grade lymphocytic activation occurs. Such lymphocytes may release cytokines which induce fibroblast activity, such as TGF-β and connective tissue-activating peptides (CTAP-1 to CTAP-6). CTAPs are low molecular weight compounds released from leucocyte and platelet granules during inflammation and which themselves undergo partial degradation to produce other pro-inflammatory peptides such as neutrophil-activating peptides (NAP-1 and -2), thereby sustaining the inflammatory response. If this response is excessive, subepithelial fibrosis may occur and produce conditions such as benign mucous membrane pemphigoid and subretinal fibrosis, both extremely debilitating and blinding diseases. CTAPs have been implicated in the fibrosis of Graves' ophthalmopathy through activation of the insulin-like growth factor receptor on orbital fibroblasts.

In many chronic inflammations the balance of competing mediators and cells may decide the outcome of disease. For instance in hepatitis caused by HepB or HepC virus, IL-22 produced by Th17cells or by iLCs, or even by specialized IL-22-producing T cells, can have either a protective role promoting fibrosis or a pro-inflammatory role leading to liver failure. Thus in chronic inflammation the distinction between innate and adaptive immune responses becomes blurred and such disorders manifest as a mixture of low-grade inflammatory activity with partial attempts at healing (fibrosis). This is well demonstrated in the subretinal neovascular membranes of chronic posterior uveitis.

THE SYSTEMIC RESPONSE TO ACUTE INFLAMMATION: THE ACUTE-PHASE REACTION

Although the acute inflammatory response is initiated at the site of tissue injury, systemic effects are produced in proportion to the level of tissue damage and virulence of the organisms. These effects are mediated primarily by cytokines acting in this situation at a distance, and are known as the acute-phase response. These cytokines include the 'alarm' cytokines, IL-1 and TNF-α, released mainly through 'danger' signals, from macrophages activated by mast cell and platelet degranulation and/or directly by bacterial products such as endotoxin, peptidoglycan and degraded nucleotides through their PRRs (see above). During the acute-phase response, adhesion molecule expression on vascular endothelium is induced and initiates further rounds of inflammatory cell accumulation and cytokine release. In addition, changes in vascular tone are caused by release of low molecular weight metabolic products including the prostaglandins PGI_2, PGE_2, PGD_2 and $PGF_{2\alpha}$ (vasodilatation), thromboxane A_2 (vasoconstriction), and leukotrienes C_4, D_4 and E_4 (smooth muscle contraction).

The effect of massive systemic cytokine release (such as occurs in the 'cytokine storm' of acute sepsis) is to induce a fever response by the direct action of IL-1 and IL-6 on the hypothalamic temperature control system, and, second, to induce hepatocyte gene transcription for several 'acute-phase reactants' such as C-reactive protein, serum amyloid components A and P, α_1-glycoprotein, C3 and collectins (see section on complement below) as well as mannan-binding proteins, and haptoglobin. Fibrinogen, α_2-macroglobulin and α_1-antitrypsin are also synthesized.

While many plasma proteins rise in concentration, others such as albumin and transferrin fall. Clinically, the acute-phase response manifests as an elevated erythrocyte sedimentation rate (ESR) caused by more rapid settling of IL-6-mediated, fibrinogen-mediated rouleaux formation of red blood cells.

Many of the acute-phase proteins enhance existing innate defence mechanisms such as C-reactive protein, which acts as an opsonin and binds complement. Others act to inhibit the effect of inflammatory cytokines such as serum amyloid protein and IL-1. The acute-phase reaction has been described as innate immune 'brinkmanship' in that both the pathogen and the host are stressed but the overall outcome aims to limit systemic sepsis while avoiding disseminated intravascular coagulation (DIC), shock, organ failure and death.

Development of adaptive immunity and immunological memory

Normally, initiation of the adaptive immune response does not take place at the site of injury or penetration by foreign organisms. Instead, antigen is taken up by APCs at the site of inflammation and transported to the regional lymph nodes and/or spleen where it is presented to T and B cells. T and B cells specific for that antigen respond by undergoing a series of activation steps including cytokine production and proliferation (clonal expansion). Such T cells are known as effector T cells which pass into the blood circulation and migrate ('home') back to the site of injury where CD8 T cells (Tc cells) kill infected cells, and CD4 T cells (Th cells) attract other pro-inflammatory phagocytic cells such as activated M1 macrophages. In the lymph node, antigen is also presented to B cells, which differentiate to become antibody-producing plasma cells, some of which migrate to the bone marrow but most of which remain in the draining lymph node follicles and germinal centres, producing high levels of specific antibody, which is released into the circulation (see below).

ANTIGEN RECOGNITION IS MADE POSSIBLE BY ANTIGEN-PRESENTING CELLS

Foreign antigen is *presented* to T cells by three types of 'professional' antigen-presenting cell: macrophages, B cells and dendritic cells (DCs). However, antigen is *recognized* by specific T cells only after it has been processed and made presentable in an appropriate form to the T cell. Some antigens are recognized by T cells without processing, but they are very unusual. Also antigen can be presented by 'non-professional' APCs such as endothelium but the T cells may not become activated (see p. 441, discussion on tolerance).

There are differences in the type of antigen that each of the three cells can present. Macrophages and B cells usually recognize antigen through the immunoglobulin (Ig) molecule. Thus, macrophages and B cells can initiate an immune response only if the host has already been exposed to that antigen and has the capacity to mount a 'memory' response in the form of IgG. In contrast, DCs can process and present antigen to resting, naive T cells, i.e. cells that have not previously 'seen' the antigen. Accordingly, DCs are considered to be the cells which initiate immune responses to new antigens, while macrophages and B cells may be important in sustaining the response while antigen persists in the tissue.

From the earliest stages of an inflammatory response, DCs at the site of injury (generated from extravasated, circulating monocytes see p. 377) start to migrate in large numbers from the subepithelial layers into afferent lymphatics to the regional lymph nodes. During this phase, they prepare the antigen by combining it with MHC class I and II molecules so that it can be presented as a complex to T cells. T cells will respond only if they possess the specific receptor (the T-cell receptor, TCR) for that antigen and if the antigen is sufficiently immunogenic, i.e. has the capacity to activate innate immune cells via PRRs (see p. 389). In the normal course of events it is likely that the great majority of processed antigens never get as far as initiating a perceptible T-cell response.

T CELLS RESPOND TO ANTIGEN BY CLONAL EXPANSION

If the antigen is presented to the specific T cell that recognizes it, in a suitable form and in the presence of the correct co-stimulatory signals, the T cell responds by clonally proliferating, i.e. it rapidly divides, producing many daughter cells, which are all exactly the same in their recognition of that antigen alone. This is a very dramatic response and accounts for the enlarged lymph nodes seen for instance after a viral infection. It has been estimated that up to 20% of lymphocytes in an enlarged lymph node are specific for that virus.

Some of the expanded T cells (Th cells) migrate to the B-cell follicles in the lymph nodes (see p. 414) and release a range of cytokines that 'help' B cells to clonally expand, also in an antigen-specific manner. However, the majority of activated T cells enter the circulation and home to the site of injury where they assist in mounting the antigen-specific effector response that will eliminate the foreign antigen.

T AND B CELLS PARTICIPATE IN THE EFFECTOR RESPONSE

Effector responses are those that actually mediate the immune response. Activated Th cells release cytokines that activate other cells in addition to B cells. These include:

- cytotoxic T cells that recognize intracellular foreign antigen when it is presented on the surface of tissue cells complexed with MHC class I antigen (see below)
- macrophages that, when activated, remove foreign antigen and perpetuate the immune response by engaging in local antigen presentation at the site of inflammation (if this goes awry, conditions such as lepromatous leprosy and sarcoidosis can develop)
- B cells that are stimulated to full differentiation as plasma cells with considerable local antibody production (see below); soluble antibody is then available to form immune complexes and participate in further local antigen presentation and antibody-mediated cytotoxic reactions via NK cells (see below).

HOW DOES THE ORGANISM DEAL WITH INTRACELLULAR ANTIGEN?

In the above scenario, we considered how adaptive immunity is triggered when the foreign antigen is extracellular and is a target for phagocytosis. However, what happens if the antigen has already invaded and infected cells, and is persisting as a viable organism intracellularly? This particularly applies to organisms such as viruses and protozoa which infect many cell types and even bacteria such as mycobacteria which evade the immune system by 'hiding' inside cells. Such cells may be killed by NK cells, or by sensitized Tc cells recognizing viral peptide on MHC class I surface antigen as part of a memory response. The apoptotic and dying infected cells containing virus or other microorganisms are then 'cleared' (phagocytosed) by macrophages and DCs at the site of injury which, after trafficking to the draining lymph node, further activate antigen-specific T cells through presentation of viral antigen (peptide) in conjunction with 'self' MHC class I. Thus Tc can be directly activated on appropriately conditioned DCs without obligatory help from Th cells.

Effector mechanisms

The innate immune system has a range of effector cells that remove damaged tissue and dead microorganisms (DCs, macrophages, NK cells, γδ T cells). The adaptive immune system also has a variety of effector mechanisms, which it uses to rid the host of specific foreign antigen. These include antibodies and cells, but the adaptive immune system also utilizes non-specific mechanisms that are activated by antigen-specific cells and molecules, including complement and cytokines.

ANTIBODIES

Antibodies are distributed in the endoplasmic reticulum, the Golgi apparatus and the surface of B cells, monocytes, mast cells and NK cells, and in secretory fluids. Each antibody binds uniquely to a single antigen, which is usually a short sequence of an immunogenic molecule and is normally defined as an antigenic epitope. Most antigens are proteins, but carbohydrates and lipid antigens also occur.

All five antibody isotypes have similar basic structure

Antibodies are Y-shaped molecules composed of light and heavy chains (Fig. 7-8). Two identical heavy (H) chains linked by disulphide bridges to two identical light (L) chains, either a κ chain or a λ chain, form the basic structure of antibodies. Each chain is composed of a series of repeating homologous units, about 110 amino acids in length, comprising discrete immunoglobulin domains. Many other molecules adopt a similar folded structure and are classed together in the immunoglobulin superfamily. Differences exist in the precise geometry of the molecules, as shown by crystallography, which has significance for antigen binding. There are five immunoglobulin isotypes (Box 7-7).

Special features of H and L chains

Although not directly involved in antigen binding, the framework region determines the folding of the

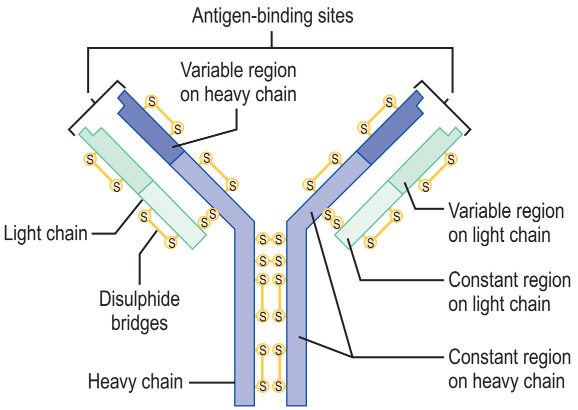
Antigen-binding sites
Variable region on heavy chain
Light chain
Disulphide bridges
Heavy chain
Variable region on light chain
Constant region on light chain
Constant region on heavy chain

FIGURE 7-8 Cartoon of generic immunoglobulin structure showing the light chains, heavy chains and the variable and constant regions with the linking disulphide bridges. (Figure adapted from http://www.emc.maricopa.edu/faculty/farabee/biobk/BioBookIMMUN.html. Image used taken from Purves et al., *Life: The Science of Biology, 4th Edition, by Sinauer Associates* www.sinauer.com) and WH Freeman (www.whfreeman.com), used with permission.)

molecules and thus the amount of complementarity-determining region (CDR) that is presented on the surface of the variable sections of the molecule for interaction with antigen (the antibody-binding site) (Fig. 7-8).

The secretory forms of IgM, IgA and IgD have C-terminal extensions (tail pieces) which allow multimer formation by attachment to the J chain, while N-linked oligosaccharides bound via asparagine residues contribute greatly to the differences in the overall conformation. Complement binds to the constant regions of the immunoglobulin molecules.

Limited proteolysis of antibody produces fragments: papain attacks the hinge region, producing single Fab fragments which bind one antigen molecule; pepsin attacks the second CH segment, producing $(Fab)_2$ fragments which can bind two antigen molecules (eFig. 7-3).

Additional content available at https://expertconsult.inkling.com/.

Immunoglobulin isotypes not only vary on the basis of the number of immunoglobulin chains which make the definitive molecule but also the ends of the heavy chains show considerable and conserved amino acid variations. Similarly, the light chains exist in two isotypes, κ and λ, which vary in proportion (60 : 40 in humans, 95 : 5 in mouse). In addition, membrane-bound immunoglobulin (mIg) is known as the B-cell receptor and contains a transmembrane region and an intracellular cytoplasmic domain.

Antibodies produced in response to the initial encounter with antigen are of the IgM or IgD isotype, while later antibody production (particularly that in response to rechallenge by the same antigen) is of the IgG, IgA or IgE isotypes. This is known as *isotype switching* and is regulated by the enzyme cytidine deaminase, activated through CD40L–CD40 interactions.

Antigen–antibody binding

The specific binding site on the antigen is known as the epitope (see p. 421). Antibody binding to carbohydrate or lipid antigens is strictly dependent on the structure of the antigen. Binding to protein antigens may, however, depend on either the linear sequence of amino acids or the three-dimensional conformation of the molecule. This can lead to the appearance of cryptic epitopes or neoepitopes in protein molecules, which are uncovered after changes to the three-dimensional conformation by partial hydrolysis.

Epitopes on protein antigens can be overlapping or non-overlapping. Overlapping epitopes lead to competition for binding; in such circumstances antibodies to an overlapping region might sterically inhibit presentation of a peptide to T cells. This, however, would occur only with MHC class II–peptide complexes where the binding site for the TCR embedded in the MHC groove overlaps with a binding site for antibody. This mechanism has been suggested for the inhibition of experimental uveitis with an antigen-specific monoclonal antibody. Antibodies may also compete by allosteric mechanisms in which binding of the antibody alters the conformation of the molecule so that it does not bind a second antibody.

B cells arise from precursors in the bone marrow as immature B cells expressing IgM, and Igα and β, and go through a series of transitional stages to become naive mature B cells in the spleen at which stage they express IgD. A small proportion of these cells become marginal zone B cells in the spleen and short-lived IgM-producing plasma cells, while the majority are available for interaction with specialized B-cell follicle (follicular) DCs to initiate antigen-specific immunoglobulin production through the B-cell receptor (BCR). The BCR, similar to the TCR, signals through

There are three components or segments to the CDR (CDR-1, -2 and -3) and the third (CDR-3) in both heavy and light chains is the most variable (Fig. 7-8). The C regions on the heavy chains are globular structures attached to the binding region by a flexible rod-like hinge portion of the molecule (see Fig. 7-8 and eFig. 7-3). The hinge region has both rigidity (conferred by proline residues at the top of the CH rod) and flexibility as the result of a large number of glycine residues. The last CH domain of immunoglobulin has a transmembrane and cytoplasmic portion, involved in intracellular signalling. Each domain is composed of two layers of β-pleated sheets with three or four strands of anti-parallel polypeptides. Therapeutic monoclonal antibodies are used extensively for many diseases including eye diseases such as age-related macular degeneration (AMD, see Ch. 9, p. 513) in which a humanized (eFig 7-3, lower panel) monoclonal antivascular endothelial growth factor (anti-VEGF) antibody is administered to patients with wet AMD by successive injections of the antibody into the vitreous chamber of the eye.

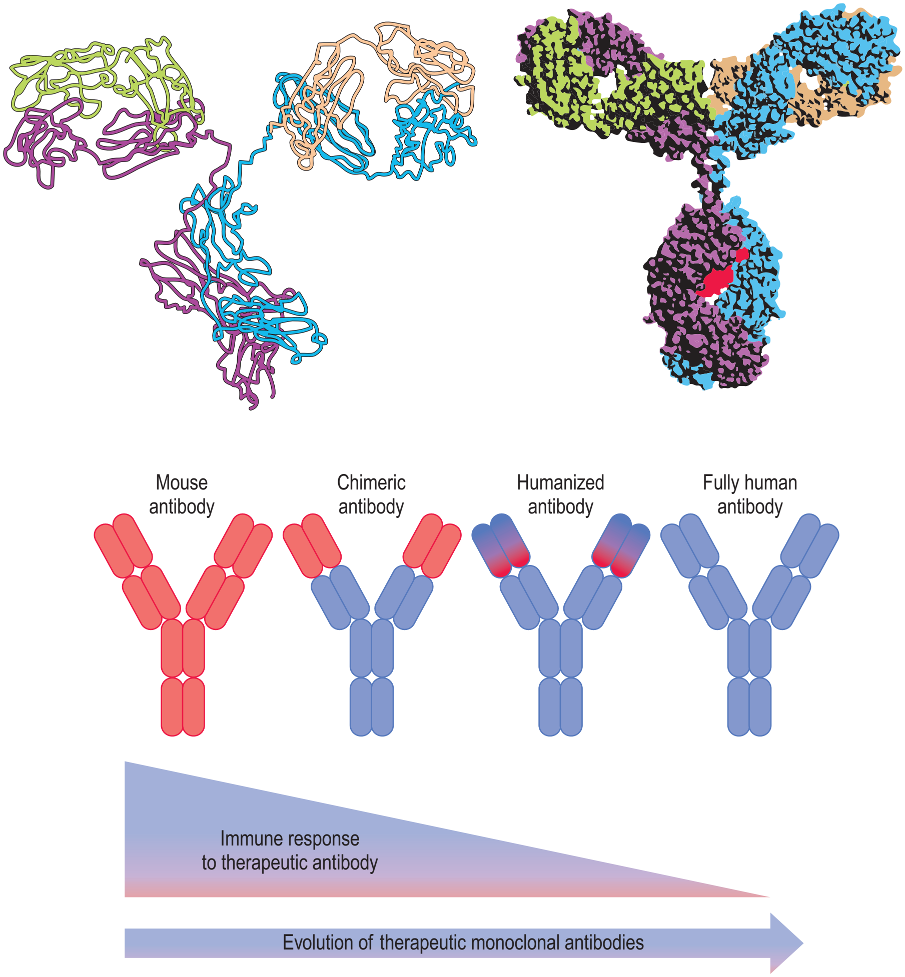
Mouse antibody
Chimeric antibody
Humanized antibody
Fully human antibody
Immune response to therapeutic antibody
Evolution of therapeutic monoclonal antibodies

eFIGURE 7-3 Antibody structure. The globular domains of the Fab and the Fc fragments are shown in the panel on top right and the anti-parallel pleated sheets on the top left. Development of engineered monoclonal fully humanized antibodies is shown on lower panel. http://www.dsch.univ.trieste.it/~benedetti/antibody_catalysis.htm http://dict.space.4goo.net/dict?q=antibody. http://www.chiscientific.cn/service.aspx?ID=30

BOX 7-7 ISOTYPES OF IMMUNOGLOBULIN

Five isotypes of immunoglobulin exist, denoted by their heavy chain (α, γ, δ, ε, μ); there are two light chains (κ, λ), which occur in a ratio of 60:40 in humans. The Fc and C3 binding regions are responsible for the effector functions of antibodies (see below). Three complementarity determining regions (CDRs) are nested within framework regions, in the V and hyper-V regions of the H and L chains, and receive contributions from both chains. Somatic mutations in the CDRs are responsible for the enormous antibody diversity and affinity maturation that occurs on repeated exposure to antigen.

IgG

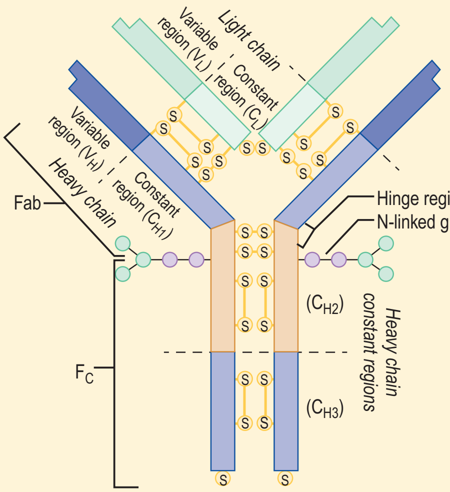
Light chain
Variable region (V_L)
Constant region (C_L)
Variable region (V_H)
Constant region (C_{H1})
Heavy chain
Fab
Hinge region
N-linked glycan
S
(C_{H2})
(C_{H3})
Heavy chain constant regions
F_C

IgM Pentamer

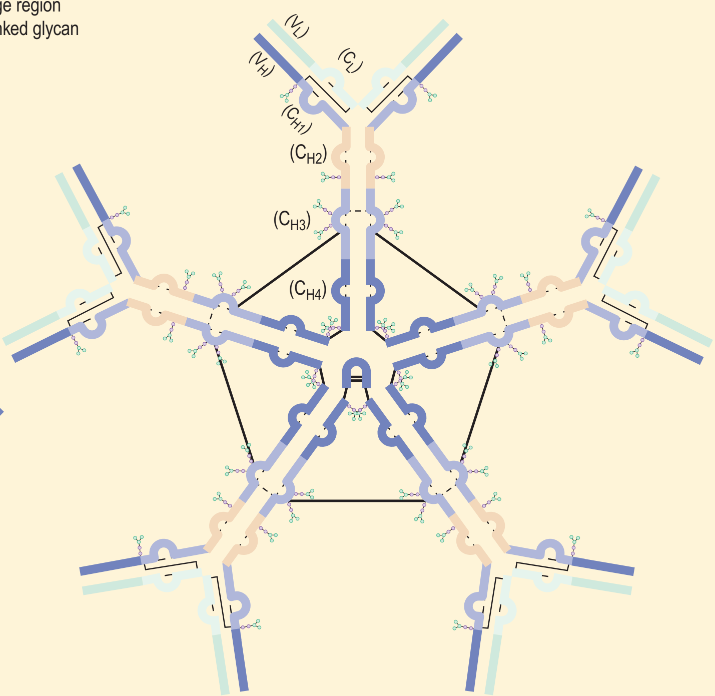
(V_L)
(V_H)
(C_L)
(C_{H1})
(C_{H2})
(C_{H3})
(C_{H4})

IgA Dimer

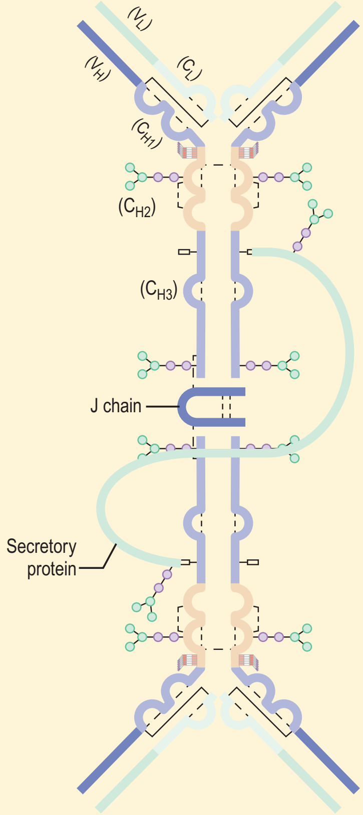
(V_L)
(V_H)
(C_L)
(C_{H1})
(C_{H2})
(C_{H3})
J chain
Secretory protein

IgD

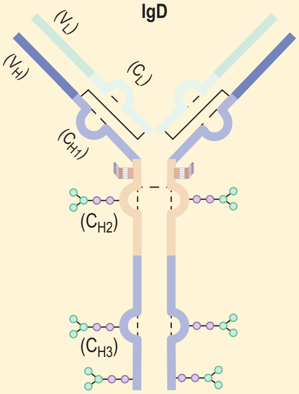
(V_L)
(V_H)
(C_L)
(C_{H1})
(C_{H2})
(C_{H3})

IgE

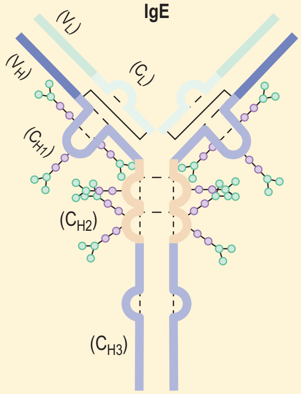
(V_L)
(V_H)
(C_L)
(C_{H1})
(C_{H2})
(C_{H3})

cytoplasmic molecules such as Syk and Lyn, as well as PI3 kinase and SHIP to generate plasmablasts in the B-cell follicle. These cells leave the secondary lymphoid tissue to return to the bone marrow and other tissues, where they function as long-lived antibody-producing plasma cells. Isotype switching, i.e. from IgM to antigen-specific IgG, E and A, occurs while in the germinal centre of the B-cell follicle.

As indicated above, antigen specificity is determined by complementarity between the epitope on the antigen and the CDR. However, this is not exclusive and other neighbouring regions on the antigen (paratopes) and on the IgG molecule outside the hypervariable region may influence the final antigen specificity and avidity. In addition, antibodies as proteins may have other properties relating to their non-ligand (antigen) binding sites, for instance the recombination of the V_H and V_L chains may fortuitously produce a site that binds ADP or acts in a catalytic fashion. Such antibodies may therefore have additional functions that may be more closely ascribed to innate rather than adaptive immunity. Such antibodies are termed superantibodies.

Monoclonal and polyclonal antibodies

Any single antigen, especially large proteins, may have multiple antibody-binding regions (epitopes). Antibody responses may therefore be polyclonal, oligoclonal or monoclonal depending on the immunodominance of the antigenic determinants. Selection of cells from an immunized mouse under special conditions and fusion of those cells with an immortalized cell line is a powerful technique for producing large quantities of antibody to a specific antigenic epitope.

Monoclonal antibodies (Mabs) have revolutionized diagnostic techniques in medicine today, particularly using flow cytometry in which antibodies specific to cell proteins are 'tagged' with fluorescent markers which identify even very rare cell types using narrow-wavelength lasers. Mabs are also widely used for therapy of diseases such as cancer and autoimmunity, particularly in ophthalmology where anti-TNF-α therapy is used to treat uveitis and anti-VEGF therapy is used to block blood vessel growth 'wet' AMD. Some of these Mabs are chimeric antibodies, i.e. molecules which combine human and mouse antibody segments, while others are fully 'humanized' (see eFig. 7-3). Fragments of antibodies are sometimes as effective as the intact antibody and due to their reduced size they possibly have better penetration into tissues as therapeutic agents (Box 7-8).

CYTOKINES ARE THE EFFECTOR ELEMENTS RELEASED BY CELLS DURING INNATE AND ADAPTIVE IMMUNE RESPONSES

Almost all of the biological effects of T cells are mediated by cytokines. More importantly, T cells alter the characteristics of an immune response by releasing different cocktails of cytokines. In a feed-forward mechanism dictated by the cytokine milieu of the tissues, naive T cells differentiate into one or more of the several different T-cell types (Fig. 7-2).

Cytokines, in the broadest sense, are produced by a wide variety of leucocytes and tissue cells, particularly monocytes/macrophages, epithelial cells and fibroblasts. The response of tissues to invasion by viruses and bacteria is to produce cytokines; for example, virus-infected cells activate NK cells, which can be induced to release cytokines by innate recognition of viral double-stranded RNA (e.g. IL-1, IL18); bacterial lipid (endotoxin) is recognized by CD14 and TLR4 on monocytes and by complement, all leading to pro-inflammatory cytokine release, e.g. IL-12, and IL-23.

What makes a cytokine a cytokine?

Cytokines have the following properties:

- they are secreted by cells in response to a specific stimulus.
- they are short-lived and short-range molecules, acting on cells within their neighbourhood.
- they may have effects at a distance if they are liberated into the circulation in sufficient concentration.
- they are effective at very low concentration.
- they may secondarily induce cytokine release by the target cell.
- they may act upon many different cell types (pleiotropism) and may have multiple different effects on the same target cell.
- they may be redundant, may induce cytokine synthesis themselves and may alter the effects of other cytokines.

BOX 7-8 ANTIBODY ENGINEERING FOR THERAPEUTIC USE

Monoclonal antibodies are produced by the fusion of antibody-secreting B cells with an immortalized myeloma cell line, engineered in such a way that only the B-cell/myeloma fusion cells survive, thus producing an immortalized antibody-producing cell. Some monoclonal antibodies have been produced continuously for decades.

The antibodies themselves can be engineered to produce fragments such as Fab fragments or single chain variable fragments (scFv) or even antibodies which combine a double set of variable fragments (diabodies) (see figure below). In this way the antibodies can be made more selective and potent.

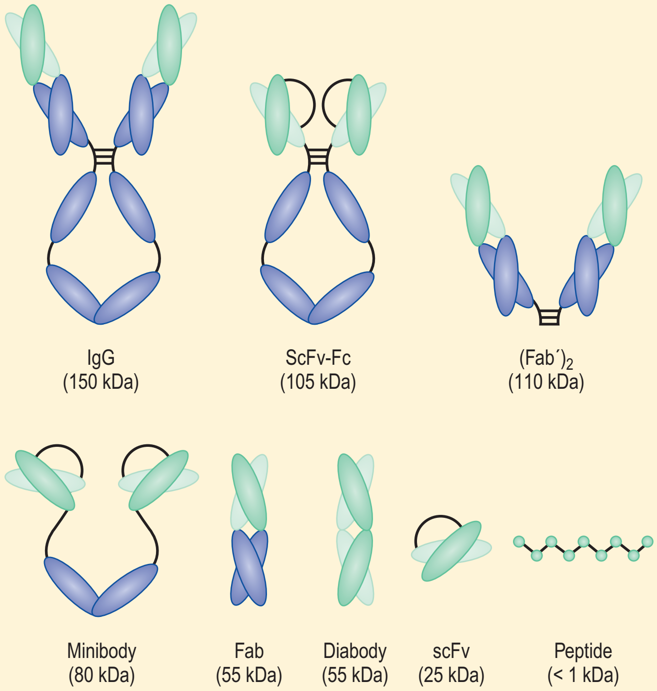

Cytokines have three general effects:

- they regulate the innate immune response.
- they regulate the adaptive immune response.
- they regulate the growth and differentiation of haematopoietic cells.

Cytokines now include several groups of molecules, such as the interleukins, growth factors, colony-stimulating factors, transforming growth factors, interferons, tumour necrosis factors, chemokines and monokines. Considerable functional overlap exists between these groups. At least 39 interleukins have now been described, each with a range of actions. IL-1 is produced by almost all nucleated cells, including ocular cells such as the RPE, and is central to the initiation of most inflammatory and immune responses. Many of the more recently described interleukins have homology to IL-1, including IL-18, IL-33, IL-35 and IL-39. Some of these have anti-inflammatory effects, unlike IL-1 and IL-18 which are secreted on activation of the inflammasome (Fig. 7-9). IL-1 has extensive homology to fibroblast growth factor and may also be implicated in angiogenesis.

IL-2 is the major T-cell growth factor and initiates release of cytokines from the cells upon which it acts (Fig. 7-2).

Cytokines function in a vast interconnected network of agonist/antagonistic feed-forward and feedback inhibitory and stimulatory loops that ensures fine

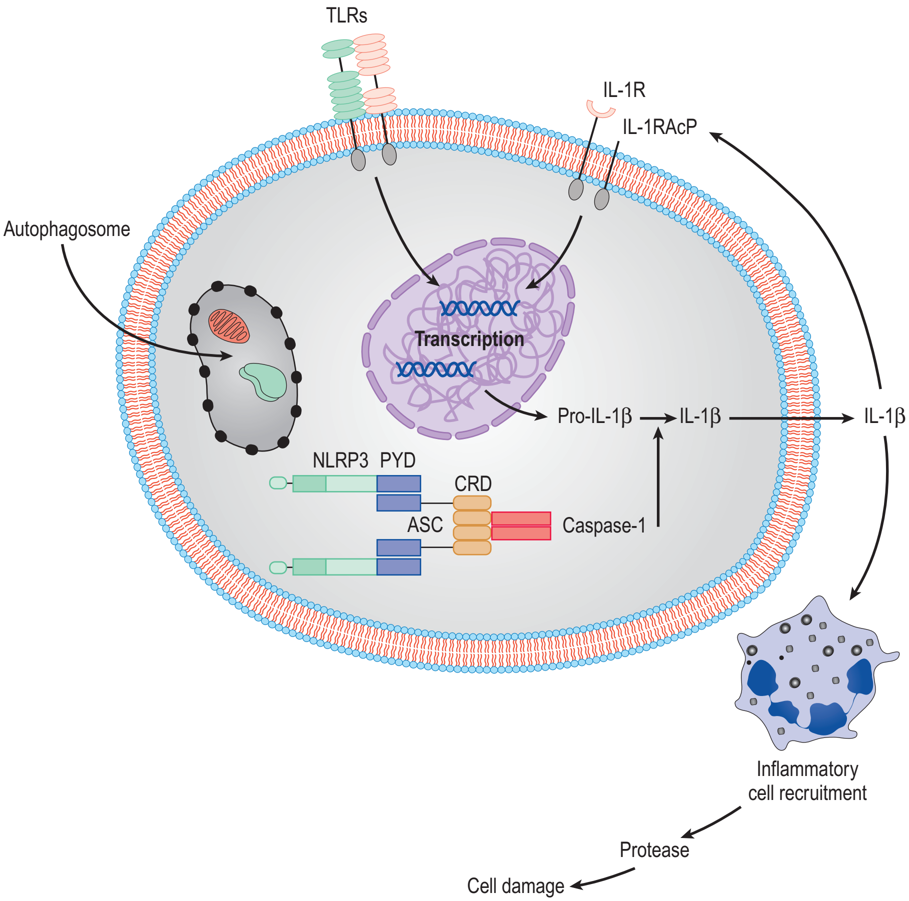

FIGURE 7-9 IL-1β is transcribed through a process involving activation of PRRs such as Toll receptors or via the IL-1 receptor itself, leading to activation of the inflammasome and induction of caspase-1 to generate IL-1β from its precursor pro-IL-1β. The process of autophagy regulates inflammasome activation and IL-1 production. (From van de Veerdonk and Netea, 2013.)

control over the immune response, mostly to avoid the excessive collateral damage that would occur in an over-robust response to a pathogen.

Cytokines involved in specific immune reactions

The interleukins, interferons and tumour necrosis factors are central to the immune response, and the character of the immune response is determined by the set of cytokines released. Activation of a naive T cell (a lymphoblast that has recently been presented with antigen) by different cytokines (in the presence of IL-2) released from the APC drives the T cell to differentiate into a Th1, Th2, Th17/Th22 or Treg cell (see Fig. 7-2). Th1 cells secrete interferon-γ (IFN-γ), in response to IL-12 secreted by macrophages or DCs during the initial innate response to antigen. Th1 cells are involved in delayed-type hypersensitivity responses and tissue damage associated with granuloma formation (see Ch. 9, p. 502). This is achieved by activation of macrophages and NK cells, which release reactive oxygen and nitrogen intermediates (free radicals). In addition, release of IL-2 by Th1 cells activates cytotoxic T cells. In contrast, Th2 cells release IL-4, IL-5, IL-6, IL-10 and IL-13, activate B cells and induce antibody production. IL-5 also stimulates eosinophils, which are the effector cells in allergy-associated tissue damage, while IgE avidly binds to mast cells and causes release of the mediators of immediate hypersensitivity. In mucosal tissues such as the gut lining, IgA is produced with the help of Th2 cytokines.

This simple outline, describing the generation of either Th1 or Th2 responses is considerably modified during the event. For instance, IL-1, IL-2 and IL-4 can also activate macrophages directly, while IFN-γ is involved in the production of $IgG_{2\alpha}$ (the only immunoglobulin controlled by this cytokine). In contrast, activation of a naive T cell in the presence of the cytokine TGF-β leads to the generation of Treg cells which suppress the immune response, unless IL-6 is also present in the milieu, in which case the strongly pro-inflammatory Th17 cell differentiates (see Fig. 7-2) driven by IL-23 production from the APC. IL-6 is produced by many tissue cells, including ocular cells such as the RPE, especially under stress.

Cytokines involved in lymphomyeloid cell maturation

The bone marrow is the powerhouse of our immune defence system. Several cytokines are involved in the growth and maturation of lymphomyeloid cell populations from stem cell precursors in the bone marrow. These include the colony stimulation factors, granulocyte colony-stimulating factor (G-CSF), macrophage colony-stimulating factor (M-CSF) and granulocyte–macrophage colony-stimulating factor (GM-CSF). In addition, several other cytokines with pleiotropic effects, including TNF-α and IL-1, have important roles in the maturation of these cells.

Cytokine receptors and cytokine receptor antagonists

T-cell activation is antigen specific but the cytokines released are not. However, cells targeted by cytokines require the appropriate receptor for that cytokine to respond (Fig. 7.10). In addition, the targeted cells require a cell-signalling mechanism to mediate the response.

Cytokines utilize a common cell-signalling mechanism involving a cytosolic protein NF-κB, which regulates genes encoding cytokines, their receptors and several other genes involved in the acute inflammatory response. NF-κB is released from its inhibitor (IκB) when the cell is stimulated by cytokine and enters the nucleus bound to the transcription factors p65/p50. These initiate the changes in gene transcription with changes in function.

Five families of cytokine receptors are described based on structural motifs in the proteins: (1) the immunoglobulin superfamily (IL-1 and c-kit, IL-1Ra, IL-18, IL-33, IL-35, IL-37, IL-39); (2) a two cysteine/WSXWS or type 1 receptor family (binds IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-11, IL-13, IL-15, IL25, IL27, IL31, GM-CSF, G-CSF); (3) a type II receptor (binds IFNs, IL-10, IL-9, IL-20, IL-22, IL-24, IL-26); (4) a TNF receptor (part of a larger family of receptors involved in apoptosis including Fas–Fas ligand (FasL) mechanisms, TRADD (TNF receptor-associated death domain), TRAF (TNF receptor-associated factor) and CD40; see eBox 7-1); and (5) a seven transmembrane helix (chemokine) receptor family (see p. 404). Some cytokines have more than one receptor and some of the structural motifs are shared between the receptors: for instance, the common γ chain of the IL-2 and IL-15 receptors. Binding of cytokine to the receptor initiates signal transduction pathways such as the Janus family kinases (Jaks) and the signal transducer and activator of transcription (STATs) proteins which act upstream of NF-κB (Fig. 7-11). Several proteins are included in the Jak–STAT families and these intracellular proteins are prominent targets for drug discovery programmes in attempts to control immune responses.

Cytokines, when released, do not have a free rein in mediating their activities. Cytokine receptor antagonists are well recognized, the prototype in this field being IL-1Ra. This protein is a naturally occurring competitive binding protein for the IL-1 receptor but fails to induce any of the signal transduction events of IL-1. This is important because the uncontrolled activity of IL-1 in large amounts produces severe side-effects similar to the acute-phase response. IL-1Ra recognizes the separate receptors for IL-1 on T and B cells. IL-1 can also bind to a 'decoy' receptor, IL-1RII, which fails to transmit the signal.

Receptors may be of low or high affinity in their ability to bind ligands and this encroaches somewhat on receptor specificity, allowing certain cytokines to compete for the same receptor. In addition, for certain multichain receptors, such as the IL-2R, one of the chains may be shared by several cytokines (the common γ chain) and so assembly of the appropriate receptor on the membrane may depend on the precise cytokine milieu presented to the cell. The local concentration of any particular cytokine will therefore

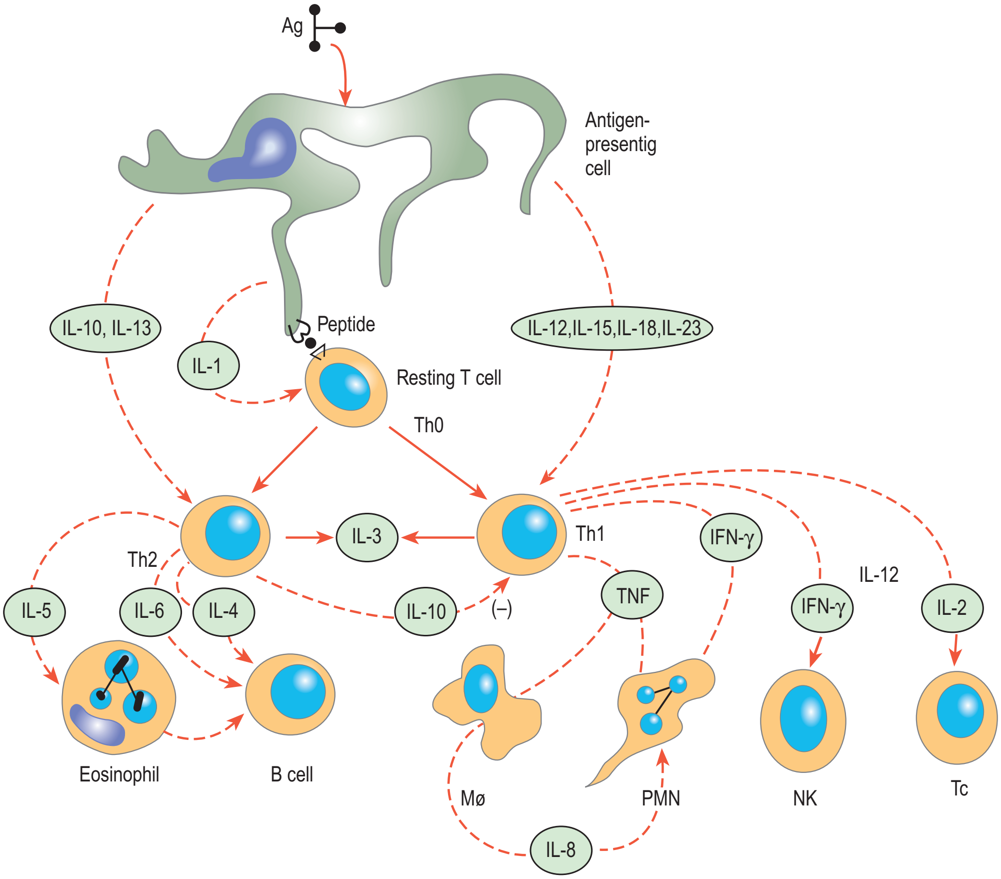

FIGURE 7-10 Cytokine networks induced by immune responses. Ag, antigen; IL-1, interleukin-1; IFN, Interferon; Mϕ, macrophage; NK, natural killer cell; PMN, polymorphonuclear leucocyte; Tc, T cytotoxic cell; T_h, T helper cell; TNF, tumour necrosis factor; ● peptide. (Courtesy of A. Abbas.)

have an influence on the final cell response. This cytokine redundancy is seen for instance with GM-CSF, IL-3 and IL-5. GM-CSF and IL-3 have widespread effects, whereas IL-5 is more restricted. By competing for the same receptor on different cell types, the amount and receptor-affinity of any one cytokine can determine the nature of the cellular response.

Some cytokines play a bigger role in inflammation than others

Lessons learned from treating patients with autoimmune inflammatory disease point towards the importance of certain selected cytokines as major effectors of inflammation. For instance, by selectively targeting TNF-α, remarkable recoveries in function can be achieved in patients blinded by uveitis or crippled with rheumatoid arthritis. Several other cytokines have primary roles in disease: for instance, IL-1 (and its family members) is involved in virtually every form of inflammation, while IL-2 is very important for T-cell function. Effector T cells can make their own IL-2, but Tregs, which are essential for controlling the inflammatory response, are unable to synthesize IL-2 and depend on IL-2 from other sources, including T effector cells (thus T effector cells initiate their own regulation).

TGF-β is also a major regulator of T-cell differentiation (see Fig. 7-2) as well as being much more generally involved in several cell biological processes such as fibrosis and angiogenesis. In addition, it is constitutively present in the anterior chamber where it mediates aspects of immune privilege (see p. 457).

Additional content available at https://expertconsult.inkling.com/.

The number of cytokines being discovered continues to increase and many of these are being shown to play important roles in inflammation and its regulation. The following sections provide brief information on specific cytokines.

TUMOUR NECROSIS FACTOR

Certain cytokines have a prominent role in tissue damage, one of which is TNF. TNF exists in both soluble and membrane-bound forms and can induce a variety of responses in cells, including activation of polymorphonuclear leucocytes, induction of MHC antigens and adhesion molecule expression, and prostaglandin synthesis. Two forms of TNF exist, TNF-α and TNF-β, the latter also known as lymphotoxin; both forms bind to the same cell receptor.

TNF-α is pro-inflammatory, pro-coagulant, cytotoxic and antiviral, and modulates haematopoiesis. Its inflammatory effects act on both the acute-phase response and locally on cells at sites of inflammation. As such it is a mediator of endotoxic shock. TNF-α signals through two receptors (p55 and p75), which are members of a larger TNF receptor family involved in several aspects of the immune response (eBox 7-1). Many are associated with cell death such as Fas-FasL (apoptosis), also important in immune privilege. Each receptor forms a (p55 or p75) homotrimer before ligand binding. TNF-α is produced as a pro-protein with a long signalling sequence. When cleaved off, the molecule trimerizes and binds to the trimeric receptor. TNF receptors can be cleaved from the cell surface and persist in the extracellular space and the bloodstream, where they can act as competitive inhibitors of TNF itself. Humanized fusion proteins using TNF have been used in this way to prevent disease.

TNF is a major mediator of experimental and clinical uveitis, and thus blockade of TNF-α ameliorates disease and preserves sight.

TRANCE

TRANCE (TNF-related activation-induced cytokine, also known as osteoprotegrin ligand, or OPGL, and receptor activator for NF-κB ligand, or RANKL) was defined initially as an activator of osteoclasts in bone for bone turnover and of innate immune cells, macrophages and dendritic cells for the immune response. Deficiency of TRANCE leads to the severe bone disease osteopetrosis. It is now also recognized as a major ligand for the promotion of immunological tolerance on medullary epithelial cells (mTECs) in the thymus where it binds the receptor RANK and sustains tolerance to self-antigens during adult life once it is established in the fetus.

INTERLEUKIN-1

Interleukins constitute a subclass of cytokines: their name is derived from their ability to effect communication between leucocytes. However, they are derived from a range of cells and have effects on many cell types, including leucocytes. IL-1 is released on cell death by many cells as part of the injury response and then amplifies this response by signalling via the inflammasome to release more IL-1 and IL-18. In macrophages and endothelial cells (its selective targets) it induces adhesion molecule expression and also promotes prostaglandin synthesis. IL-1 also activates bone cells and accelerates bone turnover, and it induces marrow stromal cells to produce G-CSF and IL-3. Two forms of IL-1 exist, IL-1α and IL-1β, each with its own receptor, but receptor usage is not highly restricted. Thus, IL-1α and IL-1β have broadly similar effects on cells. IL-1 is extremely potent and is counteracted by the IL-1 receptor antagonist, which plays a major role in regulating IL-1 activity.

IL-1 activates cells within the eye, particularly the RPE and the retinal vascular endothelium. It has been implicated in the pathogenesis of various forms of uveitis and has been shown to be uveitogenic experimentally and to have significantly greater damaging effects than TNF-α.

IL-1α and IL-1β are produced as pro-proteins without a secretory signal sequence and must be digested by caspase-1 (IL-1β) or other proteases (IL-1α) before they are activated. IL-1 (and IL-18) signal through Myd88 and the inflammasome.

INTERLEUKIN-2

IL-2 is the major cytokine involved in T-cell-mediated responses, both Th1 and Th2, and also induces NK cell activity, in both of which it may act in an autocrine manner. IL-2 also initiates activation of CD8 T cells. Activated T cells thus express the IL-2 receptor (CD25), and it has, for instance, been detected on circulating T cells as well as lymphocytes from intraocular samples from patients with endogenous uveitis. Interestingly Treg cells express the highest level of the IL-2 receptor and IL-2 is essential for Treg induction even though Tregs cannot produce this cytokine. Tregs express and require the transcription factor FoxP3 and, conventionally, Tregs are described as $CD4^+CD25^+FoxP3^+$ T cells.

THE INTERFERONS

IFN-α and IFN-β are members of the type I interferons, of which there are many, while IFN-γ is the only type II interferon, a classification based on structural differences. IFN-α was the first cytokine to be identified, sequenced,

Continued

cloned and introduced to clinical therapeutics. It is used in the treatment of multiple sclerosis, hepatitis C and certain forms of retinal vasculitis. It is produced by many cells in response to virus infection and in this regard is important in promoting viral clearance. The main source of constitutive IFN-α is a rare population of dendritic cells, the plasmacytoid dendritic cells (pDCs). Large quantities of this cytokine are produced by pDCs in response to virus infection, leading to maturation of myeloid dendritic cells with strong induction of antiviral cytolytic activity and B-cell isotype switching. IFN-α can also be produced by activation of DCs via the TNF receptor but with a gradual kinetic as opposed to the massive antiviral response. INF-α produced by TNF receptor stimulation may be immunoregulatory. IFN-β is produced mainly by fibroblasts.

Two interleukins (IL-28 and IL-29) have been described which have IFN-αβ-like activity and can induce antiviral activity in the absence of IFN. These cytokines have now been rebranded as IFN-λ, i.e. type III interferons, and like the type 1 interferons they act via the Jak-STAT pathway.

IFN-γ is released from virally infected cells and has potent antiviral activity. It is also a major pro-inflammatory cytokine and induces MHC and other antigens on cell surfaces, thus promoting adaptive immunity via antigen presentation. It is the signature cytokine for CD4$^+$ Th1 cells. It has been shown to induce MHC class II on cells in the retina (which does not normally have many MHC class II$^+$ cells); these may be tissue-specific APCs. It also induces the innate immune response by activating macrophages to produce cytokines such as TNF-α, IL-1β, IL-12, IL-18 and IL-23, and also to engage in cytotoxicity through release of reactive oxygen species and nitric oxide. Like TNF-α and IL-1, IFN-γ, has potent uveitogenic activity. In addition, high levels of IFN-γ have been detected in aqueous samples from patients with acquired immune deficiency syndrome (AIDS) retinitis.

Production of interferons is under the control of interferon-regulating factors (IRFs), which bind to interferon-stimulated regulatory elements (ISRE) on the promoters of the IFN genes.

TRANSFORMING GROWTH FACTOR-β

TGF-β is an important immunosuppressive cytokine. There are at least six types of TGF-β, and many cell types elaborate this mediator, including cells within the eye such as ciliary body epithelium and RPE. TGF-β has been suggested to account for part of the immunosuppressive activity normally found within the eye (see p. 457). TGF-β2 appears to be the main isotype found within the eye and is normally secreted in a latent form; however, it is readily converted to the active form by enzymes such as plasmin, which would normally be present in an inflammatory exudate.

TGF-β promotes T-cell differentiation as well as angiogenesis and chemotaxis of several leucocyte types.

TGF-βs have widespread effects on cell adhesion, differentiation, proliferation, migration, maturation, activation and regulation, both within and on cells outside the immune system. *In vitro* studies are greatly affected by small changes in the TGF concentrations, and thus many effects require *in vivo* investigation. TGF-βs are essential, however, because deletion of the TGF-β gene in mice is lethal for the embryo. Overall, TGF-β1 and TGF-β2 are anti-inflammatory and immunoregulatory, but TGF-β1 is pro-fibrotic. TGF-βs are produced by leucocytes and by parenchymal cells in the central nervous system, kidney and eye. TGF-β receptors comprise three chains, of which chains I and II bind to form a high-affinity receptor while III binds either of the other two chains in a regulatory, non-signalling role.

INTERLEUKIN-4

Interleukin-4, -5, and -13 are prototypical cytokines inducting Th2-type responses and are thus major players in allergic diseases including asthma and allergic skin and mucous membrane disorders such as allergic conjunctivitis. In addition, IL-4 and IL-13 are important inducers of alternative macrophage (M2) activation promoting wound healing, fibrosis and angiogenesis.

INTERLEUKIN-6

IL-6 has pro-inflammatory and fever-inducing activity, but recent evidence suggests that its major role may be to limit tissue damage. It has multiple and wide-ranging effects, and participates in the acute-phase reaction (fever) as well as in haematopoiesis. IL-6 is produced by many cell types, both immune and non-immune, including ocular cells such as RPE cells. IL-6 may be instrumental in promoting a Th2-type response with preferential activation of B cells. However, experimentally, IL-6 appears to have a marked uveitogenic effect, similar in severity to that of endotoxin. In addition, IL-6 has been detected in the aqueous of patients with endogenous and postsurgical uveitis. IL-6 combines with TGF-β to promote induction of Th17 effector cells, Th17.

Like many cytokines, IL6 and the IL6 receptor (IL6R) can exist in soluble form and engage in ‘trans-signalling’. IL-6, IL-11, IL-30 and IL-31 are grouped together due to similarities in their receptor and signalling pathways. In addition, the immunosuppressive cytokine IL-35 combines the components of IL-6 and IL-12 (http://www.sciencedirect.com/science/article/pii/S1359610112000160).

INTERLEUKIN-8

IL-8 is regarded as a chemokine for neutrophils but also has some level of chemotactic properties for monocytes and lymphocytes. IL-8 is released by immune and non-immune cells but is somewhat more restricted than IL-6. RPE cells, corneal endothelium and stromal cells release IL-8 after appropriate stimulation. IL-8 is less uveitogenic than IL-1, which appears to be the main cytokine in this respect. IL-8 binds to the chemokine receptors CXCR1 and CXCR2.

INTERLEUKIN-10

IL-10 is an 18 kDa cytokine produced by many cells. It is the prototype cytokine of the IL-10 family of cytokines, which consists of nine members: IL-10, IL-19, IL-20, IL-22, IL-24, IL-26 which are also related to IL-28 and IL-29. The IL-10 family are predominantly an 'immunosuppressive' cytokine family and are involved for instance in immunological privilege of the eye and other organs. The effects of IL-10 can override those of many pro-inflammatory cytokines. IL-10 does this by inhibiting synthesis or secretion of TNF-α, IL-1, chemokines and IL-12 by macrophages. It also reduces MHC class II expression on APCs, thus downregulating specific and innate immune responses. Certain viruses such as Epstein–Barr virus release an alternative form of IL-10, raising the possibility of virus-induced immune suppression. The heterodimer IL-10 receptor is expressed mainly on haematopoietic cells.

While IL-10 is mostly produced by all leucocytes, IL-19, IL-20 and IL-24 are produced by myeloid and epithelial cells and are involved in antibacterial responses, wound healing and tissue remodelling. In contrast, IL-22 and IL-26 have similar functions but are produced by T cells, NK and NKT cells, as well as innate immune cells such as those in the gut. IL-28 and IL-29 are predominantly antiviral.

INTERLEUKINS 12, 17, 23 AND 27

IL-12 rose to prominence as the major mediator and inducer of IFN-γ-producing CD4$^+$ Th1 responses during the time that IL-2 was recognized as having a regulatory (inducer of Tregs) as well as a T-cell-activating role. IL-12 is produced by APCs, particularly mature dendritic cells and activated macrophages, and is strongly pro-inflammatory both for Th1 cells and NK cells. IL-12 synergizes with IL-18 to induce IFN-γ production by activated T cells and NK cells. IL-12 is a heterodimer with p35 and p40 chains. It shares the p40 chain with IL-23, in which the p40 chain combines with a separate p19 chain to form the IL-23 heterodimer. IL-23 is the main pro-inflammatory cytokine, while the p35 chain of IL-12 may have some regulatory activity. IL-23 is responsible for induction of IL-17-producing CD4$^+$ Th17 cells, which are major players in the pathogenesis of autoimmune disease as well as adaptive immunity to many infectious agents, particularly fungi. Both IL-1 and IL-17 play major roles in recruiting neutrophils to sites of infection, e.g. to *Staphylococcus aureus* skin infections. They may also be involved in tumour immunity. In contrast, IL-27, with p35-like and a p40-like chain, induces the IL-12Rβ chain on Th1 cells during the initiation of Th1 responses, but its main function is to put a brake on inflammatory responses (eFig. 7-4). IL-12, IL-18, IL-23 and IL-27 induce both primary and memory adaptive T-cell responses in a coordinated series of reciprocal interactions between the dendritic cells and T cells.

INTERLEUKINS 15 AND 21

IL-2, IL-4, IL-7, IL-9, IL-15 and IL-21 share the common γ-chain of the IL-2 receptor. NK cells are absolutely dependent on IL-15. Both IL-15 and IL-12 synergize to activate NK cells. IL-15 also activates γδ T cells. IL-15 has similar functions to IL-2 on T cells. However, there are striking differences from IL-2; for instance, IL-15 is not produced by T cells but by bone marrow stromal cells. The IL-15 receptor has a unique α chain but uses the IL-2R β and γδ chains. IL-21 also activates NK cells and induces T-cell activation but unlike IL-2 and IL-15 it inhibits B-cell proliferation.

INTERLEUKIN-18

Both IL-1 and IL-18 are released from pro-peptides in APCs on activation of the inflammasome and caspase-1. IL-18 has similar pro-inflammatory functions to IL-1 and IL-12, and induces production of IFN-γ and TNF-α in macrophages. Interestingly, IL-18-deficient mice show marked susceptibility to bacterial infections but have no impairment of response to challenge with ocular antigens. In contrast, IL-18 may have an anti-angiogenic role in diseases such as age-related macular degeneration.

INTERLEUKIN-32

IL-32 is a pro-inflammatory and a pro-apoptotic cytokine involved in many autoimmune inflammatory conditions and also in viral and mycobacterial infections and cancer in humans. It has not been found in rodents as yet, and there is no clear evidence of a specific receptor, although it does appear to signal via integrins.

Continued

INTERLEUKINS 33 AND 36

IL-33 and IL-36 are members of the IL-1 family and, like IL-32, have both intracellular and extracellular effects. IL-33 is released on cell form death from nucleus disintegration and mediates both Th2 and anti-Th1/Th17 immunosuppressive effects. IL-36 is associated with a psoriasis-like disease.

INTERLEUKIN-34

IL-34 binds to the CSF-1R (M-CSFR) and appears to be necessary for Langerhans cell and microglia development, differentiation and activation, and may be important in the pathogenesis of neurodegenerative disorders.

INTERLEUKIN-37

IL-37 is the newest member of the IL-1 family and is an immunosuppressive cytokine. Like IL-1a and IL-33, IL-37 binds to the nucleus in a receptor-independent manner to exert its anti-inflammatory effects.

INTERLEUKIN-38

IL-38 binds to the IL-36 receptor (also a member of the IL-1 family) and inhibits IL-17 and IL-22 effects.

eBox 7-1

TNF receptor superfamily

The tumour necrosis factor receptor (TNFR) superfamily (there are about 50 soluble and membrane-bound members) is a very important class of signal transduction molecules in the immune system. Included in this family are the TNF receptors 1 and 2 (TNFR1 and 2), the low-affinity nerve growth factor receptor (NGFR), CD40, CD30, Fas, and others indicated in the figure. Members of the TNF receptor family are generally transmembrane proteins, but many of them can also be secreted as soluble molecules, derived either by proteolytic cleavage from the membrane or by differential mRNA processing. The TNFR family is characterized by the presence of cysteine-rich motifs of 40 amino acids in the extracellular domain that are involved in ligand binding. The ligands for these receptors are type II transmembrane proteins with a 'jelly-roll' β-sandwich structure, many of which can also be secreted.

Some TNF receptor family members are associated with additional signal transduction molecules called TNFR-associated factors (TRAFs). These proteins contain zinc-binding domains that are thought to mediate the binding of the protein to DNA, leading to transcriptional activation. Signal transduction is initiated by stimulation of oligomerized receptor complexes upon ligand binding. The ligands, such as TNF, are often found in multimeric form, and this multivalency enhances the induction of signalling. Other members of the TNFR family (most notably, the apoptosis-inducing molecule Fas) contain death domains (DDs) in their cytoplasmic tails. For example, the DD sequence allows the interaction of the membrane-bound Fas molecule with the signal-transducing protein FADD (Fas-associated death domain), which in turn delivers signals causing the cell to initiate apoptosis. More recently it has been discovered that several proteins containing the DD sequence are not involved in apoptosis at all but actually promote cell survival. In fact, depending on downstream events, engagement of TNFR1 (which contains a DD sequence) can lead either to cell death or to cell survival.

The functional consequences of ligand engagement by members of the TNF receptor superfamily are very diverse. Fas binding induces apoptosis, which is important in the maintenance of immune self-tolerance. CD27, CD30, 41BB, and CD40 signalling enhance the survival, proliferation and activation of B or T lymphocytes, often playing critical roles as co-stimulators or signal modulators. Signalling by other TNFR family members results in the activation of NF-κB in macrophages and the induction of an inflammatory response. The TNF receptor superfamily is thus central to many aspects of both innate and acquired immunity and plays fundamental roles in both cell death and survival. Different TNFR members are involved in different diseases (see figure). A greater understanding of this receptor family, as well as its corresponding family of ligands, holds great potential for the creation of novel drugs and therapeutics that could be used to control many aspects of autoimmunity and other immunopathologies.

eBox 7-1

TNF receptor superfamily—cont'd

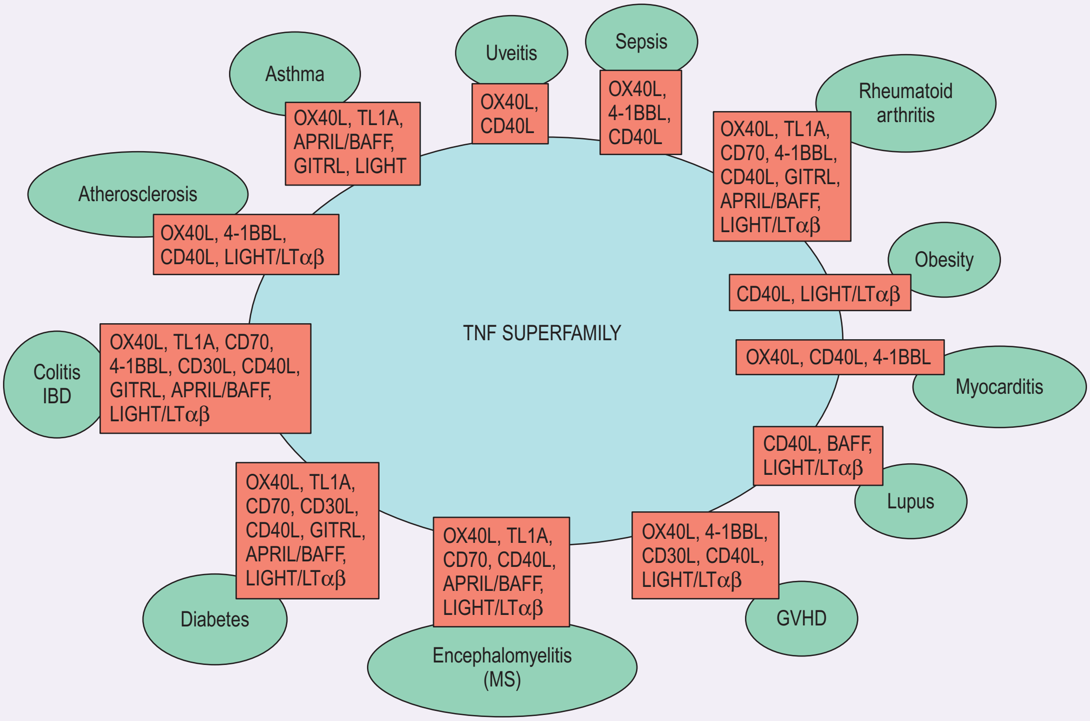
TNF SUPERFAMILY
Asthma
OX40L, TL1A, APRIL/BAFF, GITRL, LIGHT
Uveitis
OX40L, CD40L
Sepsis
OX40L, 4-1BBL, CD40L
Rheumatoid arthritis
OX40L, TL1A, CD70, 4-1BBL, CD40L, GITRL, APRIL/BAFF, LIGHT/LTαβ
Atherosclerosis
OX40L, 4-1BBL, CD40L, LIGHT/LTαβ
Obesity
CD40L, LIGHT/LTαβ
Colitis IBD
OX40L, TL1A, CD70, 4-1BBL, CD30L, CD40L, GITRL, APRIL/BAFF, LIGHT/LTαβ
OX40L, CD40L, 4-1BBL
Myocarditis
CD40L, BAFF, LIGHT/LTαβ
Lupus
OX40L, TL1A, CD70, CD30L, CD40L, GITRL, APRIL/BAFF, LIGHT/LTαβ
Diabetes
OX40L, TL1A, CD70, CD40L, APRIL/BAFF, LIGHT/LTαβ
Encephalomyelitis (MS)
OX40L, 4-1BBL, CD30L, CD40L, LIGHT/LTαβ
GVHD

Legend eBox 7.1 Figure: TNF family molecules are implicated in driving inflammatory and autoimmune disease. The figure depicts the TNFSF and TNFRSF molecules that control disease in experimental models in the mouse as a diagram of the TNFR superfamily with identification of the autoimmune diseases associated with specific molecules. Some molecules, such as OX40/OX40L, are associated generally with organ-specific diseases (such as myocarditis and uveitis) but not with systemic autoimmune diseases such as lupus (Croft et al., 2012).

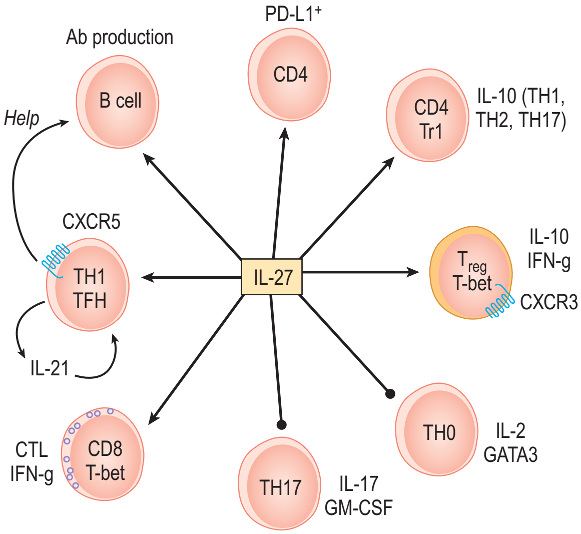
Ab production
B cell
PD-L1+
CD4
CD4 Tr1
IL-10 (TH1, TH2, TH17)
Help
CXCR5
TH1 TFH
IL-21
IL-27
Treg T-bet
IL-10 IFN-g
CXCR3
CTL IFN-g
CD8 T-bet
TH17
IL-17 GM-CSF
TH0
IL-2 GATA3

eFIGURE 7-4 IL-27 is produced by innate immune cells and influences the behaviour of adaptive immune cells.

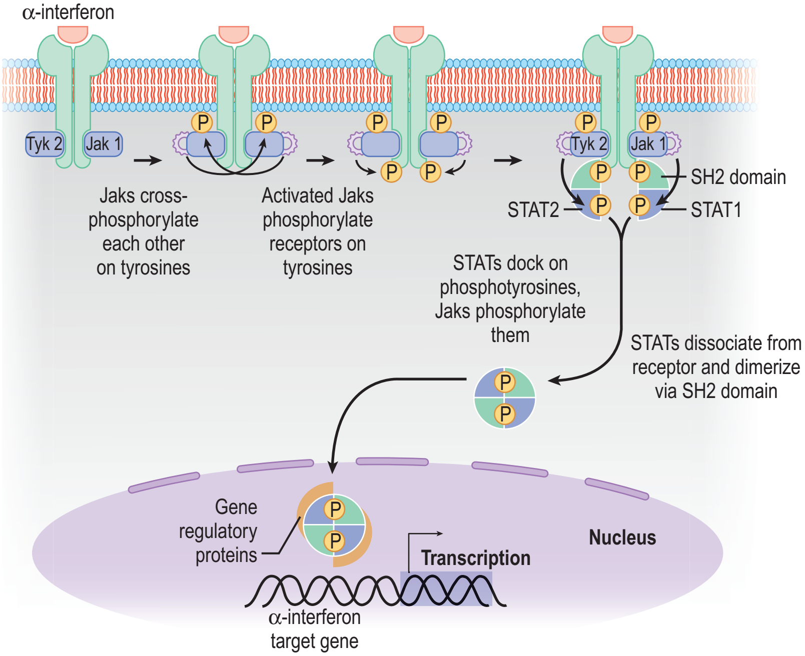
α-interferon
Tyk 2
Jak 1
P
Jaks cross-phosphorylate each other on tyrosines
Activated Jaks phosphorylate receptors on tyrosines
SH2 domain
STAT2
STAT1
STATs dock on phosphotyrosines, Jaks phosphorylate them
STATs dissociate from receptor and dimerize via SH2 domain
Gene regulatory proteins
Transcription
Nucleus
α-interferon target gene

FIGURE 7-11 The Jak-STAT pathway is activated by many cytokines, in this example by the cytokine interferon-α. (From http://www.motifolio.com/5111157.html.)

Chemokines and chemokine receptors

Cytokines that are specific for induction of leucocyte migration to the site of inflammation are termed chemokines and include some interleukins as well as other cytokines with additional functions. Chemokines are small peptides (usually 8–15 kDa) and are classified in two main subsets based on a particular amino acid sequence involving two cysteine residues, –C–C– chemokines and –C–X–C– chemokines, in which the latter contains an intervening non-cysteine residue. However, many chemokines were discovered before the current nomenclature came into being, and in many cases the older terminology is still in use (Table 7-5). Some chemokines do not belong to either category, having either only one C residue or having additional intervening amino acids, but they are unusual.

Regulation of inflammatory cell traffic to sites of inflammation is determined by the set of chemokines released in the tissues and the expression of specific chemokine receptors on different cell types. There is considerable redundancy in the system but, despite this, there is temporal regulation of specific leucocyte recruitment depending on which chemokines and chemokine receptors are active. Chemokines are released by tissue cells at the time of injury or antigen challenge and, due to their low molecular weight, percolate and diffuse through the tissues to line blood vessels and lymphatics, thereby promoting leucocytes to find a pathway to the site of inflammation or, in the case of the lymph node, to the site of T-cell activation. For instance, chemokines such as IL-8 (CXCL8) specifically attract neutrophils through the receptor CXCR1, while CCL19 and 21 attract T cells and dendritic cells using CCR7. In addition there are several other receptors and chemokine-binding agents which act as ‘decoys’ by binding the chemokine but failing to signal, thus effectively ‘confusing’ the cell and preventing effective migration. Such receptors include D6 and DARC (Fig. 7-12).

TABLE 7-5 **Nomenclature for chemokines and their receptors**

Chemokine	Old name	Receptor
CXC family	α family	
CXCL1	Groα	CXCR2
CXCL2	Groβ	CXCR2
CXCL5	ENA-78	CXCR2
CXCL8	IL-8	CXCR1, CXCR2
CXCL9	HuMIG	CXCR3
CXCL10	IP-10	CXCR3
CXCL11	ITAC	CXCR3
CXCL12	SDF-1	CXCR4
CXCL16	SR-PSOX	CXCR6
CX3CL1	Fractalkine	CX3CR1
CC family	**β family**	
CCL2	MCP-1	CCR2
CCL3	MIP-1α	CCR1, CCR5
CCL4	MIP-1β	CCR5
CCL5	RANTES	CCR1, CCR3, CCR5
CCL7	MCP-3	CCR1, CCR2
CCL19	ELC	CCR7
CCL21	SLC	CCR7
CCL28	MEC	CCR10

(From Lalor et al., 2007.)

COMPLEMENT

The term 'complement' was coined to describe an activity in sera required for antibody-mediated lysis of bacteria but that was lost after heating to 56°C. Antibody itself was heat-stable and retained the ability to agglutinate the bacteria but could not kill them. This additional activity therefore 'complemented' antibody-mediated cytotoxicity.

What is complement and what does it do?

Complement is a property of serum derived from sequential zymogen activation of a series of plasma proteins (a zymogen is an enzyme that is activated by a second enzyme, which itself has been activated by proteolytic cleavage, i.e. an enzyme cascade as occurs in kinin formation, coagulation and visual transduction).

Complement is a major component of innate immunity, i.e. the initial response to attack by pathogens, and in this sense can be regarded as a soluble pattern recognition receptor (PRR) as opposed to the membrane-bound PRRs such as the TLRs (see p. 389). Two enzyme cascades exist, one initiated by antibody combining with antigen, the classical pathway, and a second initiated directly by bacterial surface components, the alternative pathway (Box 7-9). Thus, even at this level, a distinction between innate and acquired immunity exists. The classical complement pathway is in fact the major effector mechanism for humoral immunity (see above).

A third mechanism for complement activation has been described. This involves members of a family of molecules called lectins (lectins are non-antibody, non-enzyme carbohydrate-binding proteins, as in the selectin adhesion molecules; see above). Lectins that contain collagen-like domains are known as collectins; these molecules are of considerable importance in innate immune mechanisms against microorganisms. Collectins bind to the same receptor as the C1q component of complement (see below) and are thus able to activate the classical pathway.

A major lectin in this pathway is mannan-binding lectin (MBL), which is linked to serine proteases MASP-1 and MASP-2. These enzymes are responsible for cleaving C4 and C2, respectively.

Complement has the following effects:

- it is involved in the initiation of the acute inflammatory response by release of certain peptides that act as chemotactic factors and induce vasodilatation with increased permeability (anaphylatoxin).
- it mediates antibody-dependent cytolysis by polymerizing on cell surfaces to form pores in the cell membrane.
- it solubilizes and removes immune complexes from the circulation.
- it induces phagocytosis by acting as an opsonin.

Complement proteins in the normal circulation are inactive and are maintained in this state by an elaborate system of inhibitors that not only inhibit activation of the various enzyme systems but also limit the response once activated.

The central axis of the complement pathway is the conversion of C3, activated by C3 convertases, to C5 convertase by the binding of C3b (Box 7-9). This leads to the sequential addition of a series of complement

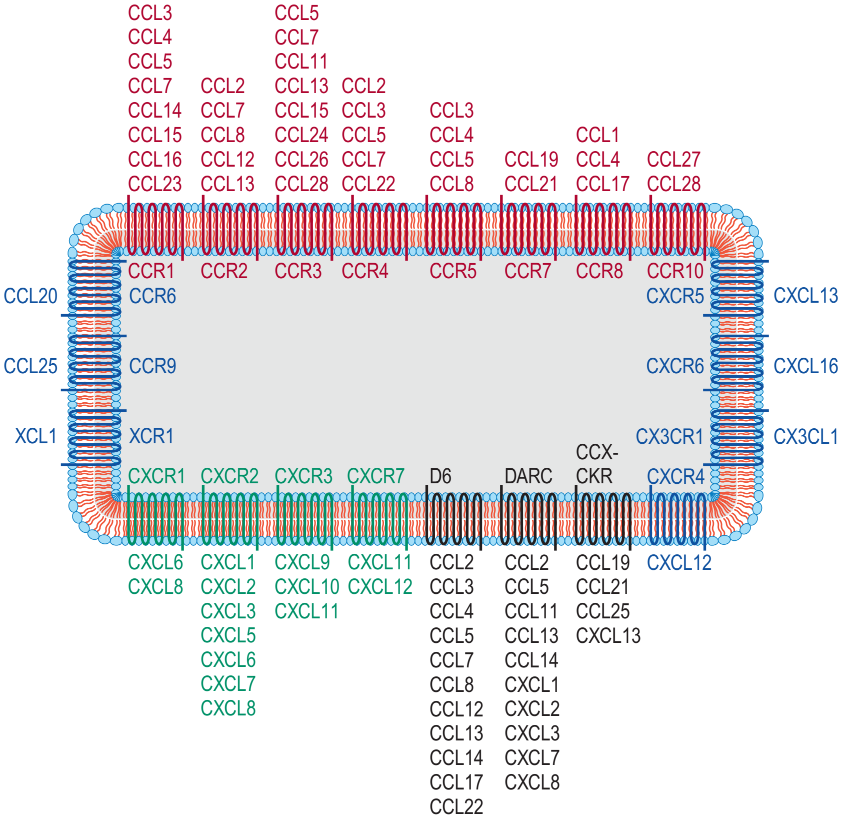
CCL3
CCL4
CCL5
CCL7
CCL14
CCL15
CCL16
CCL23
CCR1
CCL2
CCL7
CCL8
CCL12
CCL13
CCR2
CCL5
CCL7
CCL11
CCL13
CCL15
CCL24
CCL26
CCL28
CCR3
CCL2
CCL3
CCL5
CCL7
CCL22
CCR4
CCL3
CCL4
CCL5
CCL8
CCR5
CCL19
CCL21
CCR7
CCL1
CCL4
CCL17
CCR8
CCL27
CCL28
CCR10
CCL20
CCR6
CCL25
CCR9
XCL1
XCR1
CXCR5
CXCL13
CXCR6
CXCL16
CX3CR1
CX3CL1
CXCR1
CXCL6
CXCL8
CXCR2
CXCL1
CXCL2
CXCL3
CXCL5
CXCL6
CXCL7
CXCL8
CXCR3
CXCL9
CXCL10
CXCL11
CXCR7
CXCL11
CXCL12
D6
CCL2
CCL3
CCL4
CCL5
CCL7
CCL8
CCL12
CCL13
CCL14
CCL17
CCL22
DARC
CCL2
CCL5
CCL11
CCL13
CCL14
CXCL1
CXCL2
CXCL3
CXCL7
CXCL8
CCX-
CKR
CCL19
CCL21
CCL25
CXCL13
CXCR4
CXCL12

FIGURE 7-12 Chemokines and chemokine receptors bind through multiple overlapping, and probably redundant, ligand–receptor interactions. (From Lazennec and Richmond, 2010.)

proteins that result in the membrane attack complex (MAC). Complement proteins are synthesized in the liver and in mononuclear phagocytes. In addition, there is some local production by parenchymal tissue cells, as occurs in the retina.

The classical complement cascade

Activation of C1 is induced by binding to IgM or IgG, but only if the immunoglobulin has bound antigen in the form of an immune complex. Free immunoglobulin does not activate complement. Binding occurs to the C_H3 domain of IgM or the C_H2 domain of IgG, and requires at least two immunoglobulin molecules. Thus a single molecule of IgM, which is a pentamer, is able to 'fix' complement, while several molecules of IgG, usually aggregated together, are required to achieve the same. Alternatively, cell surface-bound IgG can suffice. IgM is therefore known as complement-fixing antibody and it is this antibody that plays the major role in the initial response to pathogens as part of innate immunity.

C1 can be activated by antibody-independent mechanisms including contact with retroviruses, mycoplasma, or even polyelectrolytes such as DNA and heparin. These presumably act in a non-specific manner by virtue of their charge. Importantly, the acute-phase proteins, C-reactive protein (CRP) and serum amyloid protein (SAP), can also bind complement non-specifically. C1 is composed of three molecules: C1q, C1r and C1s. C1q is a collagen-like molecule with a triple-helix conformation, while C1r and C1s are serine proteases. The molecular complex comprises a tetramer of C1s and C1r with six or more C1q molecules (Fig. 7-13).

BOX 7-9 INNATE IMMUNITY: NEW INSIGHTS IN COMPLEMENT AND COLLECTINS

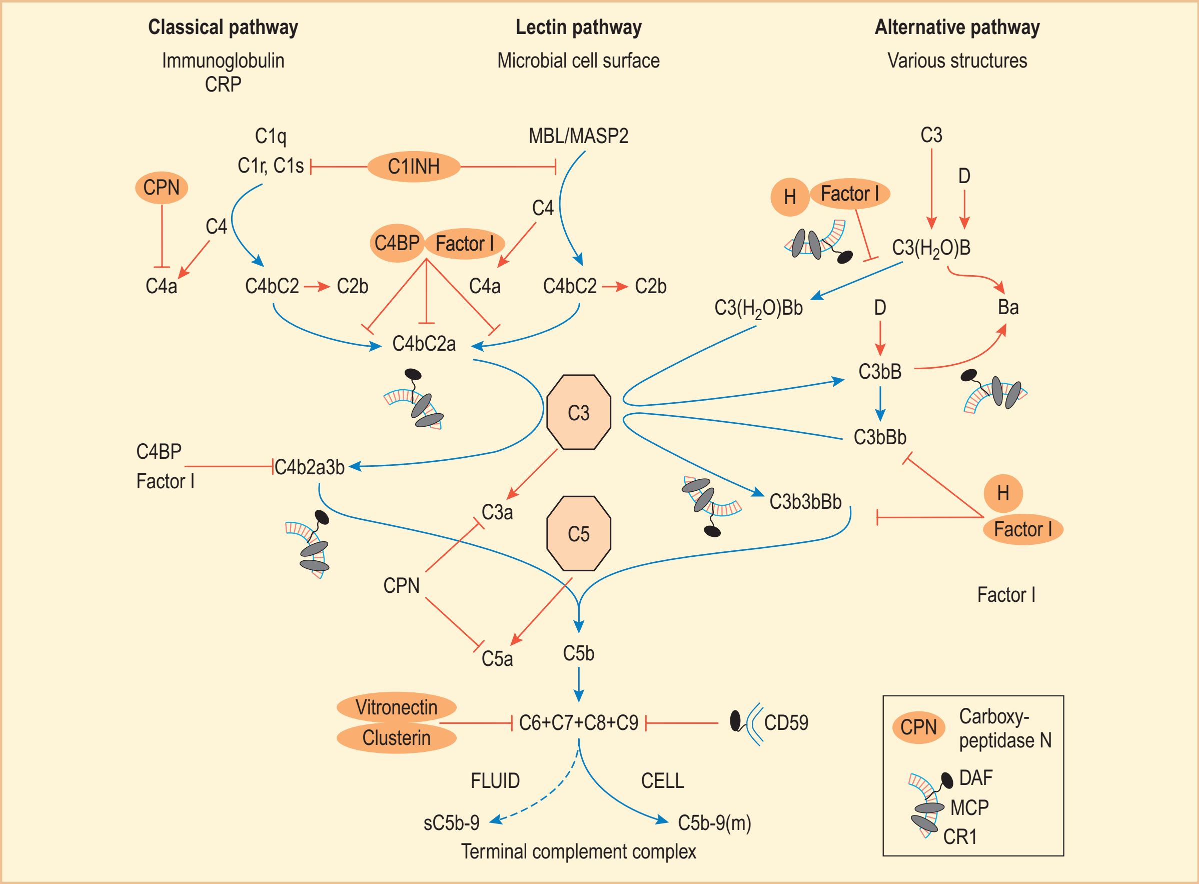
Classical pathway
Immunoglobulin
CRP
Lectin pathway
Microbial cell surface
Alternative pathway
Various structures
C1q
C1r, C1s
C1INH
MBL/MASP2
C3
D
CPN
C4
C4
H
Factor I
C4BP
Factor I
C3(H2O)B
C4a
C4bC2
C2b
C4a
C4bC2
C2b
C3(H2O)Bb
D
Ba
C4bC2a
C3bB
C3
C3bBb
C4BP
Factor I
C4b2a3b
C3a
C5
C3b3bBb
H
Factor I
CPN
Factor I
C5a
C5b
Vitronectin
Clusterin
C6+C7+C8+C9
CD59
CPN
Carboxy-peptidase N
DAF
MCP
CR1
FLUID
CELL
sC5b-9
C5b-9(m)
Terminal complement complex

When two or more globular heads of C1q bind IgM or cell-surface bound IgG, C1r is cleaved to C1r^{-}, which then cleaves C1s to C1s^{-}; this activates C4 to C2 to undergo partial proteolysis and bind to form C4b2a^{-}. C4b is unstable because it contains an internal thioester bond and is rapidly inactivated by binding a water molecule to form iC4b. If, however, C4b is formed in close contact with a cell membrane it can covalently bind to the surface and remain in an active state. In contrast, C2 is a single-chain molecule that binds to surface-bound C4a in the presence of Mg^{2+} ions. C2b, produced by C1s^{-} -induced partial proteolysis of C2, diffuses away while the C2a binds to C4b.

The complex of C4b2a^{-} contains the C3 convertase activity. The C4b component binds to the C3 molecule and brings the C2a moiety into close contact with C3, which it cleaves to C3a and C3b. C3 is a 195 kDa αβ heterodimer with an internal thioester bond similar to C4.

When partially cleaved, this molecule is also unstable and is rapidly inactivated to iC3b. However, if in contact with a cell membrane, it binds covalently in conjunction with C4b2a to form C4b2a3b^{-}, i.e. C5 convertase (see below).

C3a, C4a and C5a are small cationic peptides also known as anaphylatoxins that bind to specific receptors on basophils and mast cells. C3 cleavage products also have a role in antibody production by interacting with follicular dendritic cells (FDC) in the germinal centre (see p. 414).

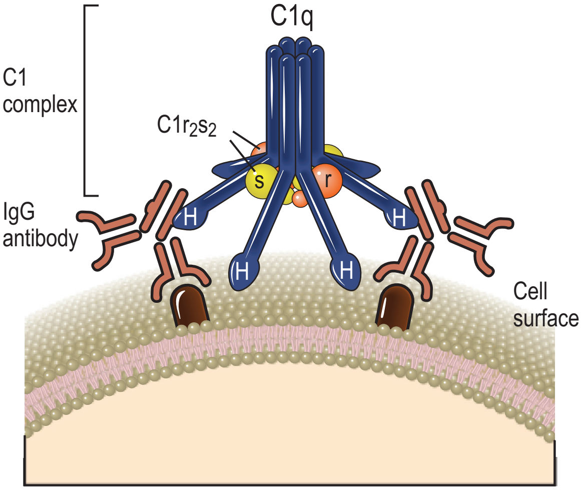

FIGURE 7-13 Diagrammatic representation of the generation of the C1 complex comprising a tetramer of C1s and C1r with six molecules of C1q. (Figure 12-10 from Abbas, A.K. Cellular and Moleclar immunology 7th Edition, p. 381, Elsevier, 2012)

The alternative pathway

The alternative pathway is triggered by low levels of C3b, which are spontaneously produced *in vivo* by proteolysis, and by C3(H_2O), which is also formed by spontaneous hydrolysis of C3. On normal cells C3b is rapidly deactivated to iC3b by innate regulatory mechanisms (see below), but on foreign surfaces C3b can remain active. C3b binds to factor B on the surface of the cells and is converted by factor D, a serine protease, to C3bBb. This contains the C3 convertase activity and requires a further protein, properdin, to protect it from proteolysis.

Deposition of C3b on foreign surfaces such as bacteria leads to further production of C3b, i.e. a positive feedback amplification loop occurs, which helps to eradicate foreign particles rapidly. C3bBb is combined to form C3bBb3b, which represents the alternative form of C5 convertase (see Box 7-9).

Collectin activation of complement

Collectins include several well-characterized proteins, such as conglutinin and MBL, which directly bind to the C1q receptor on cell surfaces and initiate such phenomena as C4-mediated red cell lysis. MBL directly activates the $C1r_2C1s_2$ tetramer in the absence of C1q. This may be mediated in association with a serine protease (MBP-associated serine protease, or MASP; see above). Collectins also have direct opsonin activity (see below, macular degeneration).

Cytolysis and the membrane attack complex

The membrane attack complex (MAC) is formed by a set of complement proteins inserting themselves into the lipid bilayer and is possible because certain proteins within the complex have a lipophilic core. C5 binds loosely to C5 convertase on the cell membrane, and is split into C5a, which diffuses away, and C5b, which complexes with C6 and then C7. C5b,6,7 is highly lipophilic and burrows into the cell membrane. There it acts as a receptor for C8, an $\alpha\beta\gamma$ trimer whose γ chain is also lipophilic and similarly inserts into the bilayer. The C5b,6,7,8 complex is weakly cytolytic but becomes considerably more so when it binds C9, a serum protein that polymerizes to form the MAC with 12–15 C9 molecules per C5–9 complex. This forms a 'pore' in the cell membrane, similar to the perforin pore of cytotoxic T cells and NK cells. The pore renders the cell permeable to small ions but not to proteins, and is therefore thought to cause cell death by osmotic effects. It is also possible that the large influx of Ca^{2+} ions poisons the cell.

The effects of the MAC on the cell are dose dependent. Sublethal doses of MAC may 'activate' the cell and induce a protective response against further attack and even cell proliferation. Genes activated in this manner are known as RGCs (response genes to complement) and several have been described.

Regulation of complement activation

Complement activation is an extremely powerful cytolytic mechanism that can be rapidly activated as a first line of defence against invasion by foreign organisms. It is also extremely effective in memory B-cell responses and antibody-dependent cytotoxicity. It can also swiftly remove potentially toxic immune complexes from the circulation. However, it is potentially extremely hazardous if randomly and uncontrollably activated, and there are therefore several inhibitory mechanisms in place to regulate this system (Box 7-10).

Activation of complement can be induced by many cell types via receptors that exist on their cell surfaces (see Boxes 7-9 and 7-10). Furthermore, some of the biological effects of complement are produced by the cleavage products of complement activation. For instance, C3a and C5a are potent chemoattractants and anaphylatoxins, and mediate early-phase responses in acute inflammation (see p. 382). Anaphylatoxins act directly via complement receptors on granulocytes, macrophages and mast cells to induce degranulation and the release of vasoactive mediators (Table 7-6).

BOX 7-10 REGULATION OF COMPLEMENT ACTIVATION

Complement proteins comprise around 15% of total serum proteins, much of which is synthesized in the liver, although some local production occurs in the tissues. Inhibitory activities exist for both fluid phase-activated complement and for cell surface-bound complement. Regulation occurs at all stages of activation.

Soluble inhibitors include: factor H, factor HL1 and properdin (alternative pathway); carboxypeptidase N, C4-binding protein, C1q and C1INH (classical pathway); CFHR1, clusterin and vitronectin (terminal complex formation). Surface-bound inhibitors include: CR1, CR2, CR3, CR4, CR-Ig, CD46, CD55 and CD59. Most of these are expressed on leucocytes, particularly myeloid cells, while some, such as CD55 and 59, are expressed on some tissue cells such as renal and retinal pigment epithelium. CD59 is particularly widely expressed.

Some receptors for complement effector proteins also modulate the overall response, such as C3aR and C5aR, C5L2, C1qR and SIGNR (the last on dendritic cells and microglia). Cell surface sialic acid preferentially binds factor H to factor B on material such as mucin, important on many mucosal surfaces including the conjunctiva.

Complement activation is usually incomplete: implications for age-related macular degeneration

As shown in Box 7-9, the multistep complement activation cascade requires the correct conditions for activation of each step and, because of instability in many of the molecular intermediates, it frequently fails to proceed to formation of the full membrane attack complex (MAC), particularly if activated via the collectins. In addition, the process can be blocked by a range of inhibitors at many stages. What happens to these intermediates? Recently, it has been suggested that partial activation of complement may have a physiological role and that not all aspects of complement activation are harmful or even beneficial to the

TABLE 7-6 **Receptors for fragments of C3**

Receptor	Structure	Ligands	Cell distribution	Function
Type 1 complement receptor (CR1, CD35)	160-250 kDa; multiple CCPRs	C3b > C4b > iC3b	Mononuclear phagocytes, neutrophils, B and T cells, erythrocytes, eosinophils, fDCs	Phagocytosis Clearance of immune complexes Promotes dissociation of C3 convertases by acting as cofactor for cleavage of C3b, C4b
Type 2 complement receptor (CR2, CD21)	145 kDa; multiple CCPRs	C3d, C3dg > iC3b	B lymphocytes, fDCs, nasopharyngeal epithelium	Co-receptor for B-cell activation Trapping of antigens in germinal centres Receptor for EBV
Type 3 complement receptor (CR3, Mac-1, CD11bCD18)	Integrin, with 165-kDa α chain and 95-kDa β_2 chain	iC3b, ICAM-1; also binds microbes	Mononuclear phagocytes, neutrophils, NK cells	Phagocytosis Leukocyte adhesion to endothelium (via ICAM-1)
Type 4 complement receptor (CR4, p150,95, CD11cCD18)	Integrin, with 150-kDa α chain and 95-kDa β_2 chain	iC3b	Mononuclear phagocytes, neutrophils, NK cells	Phagocytosis, cell adhesion?

CCPRs, complement control protein repeats; EBV, Epstein–Barr virus; fDCs, follicular dendritic cells; ICAM-1, intercellular adhesion molecule 1.

(Table 12-8 from Abbas, A.K. Cellular and Molecular Immunology 7th Edition, p. 284, Elsevier, 2012)

host overall. For instance, partially activated complement molecules may not cause lysis of the target cell; they may coat the cell and promote apoptosis; this is then followed by the silent (i.e. non-inflammatory) removal of the cell debris by scavenger macrophages. Such a mechanism may be occurring at some level of activity during the normal housekeeping actions of the resident macrophages. Indeed, it is likely to take place at sites where there is minimal cell turnover but where removal of cell debris is important, e.g. at the retinal pigment epithelium. Such a fine homeostatic mechanism may be susceptible to dysfunction: indeed, mutations in the complement inhibitory protein, complement factor H, have been associated with a higher than normal risk of age-related macular degeneration. Factor H provides a check on the complement cascade at two critical points, C3 and C5 induction by their immediate precursors.

In contrast, defective complement activation may fail to properly clear opsonized pathogenic organisms, and continued cell damage may occur to the detriment of the host (Fig. 7-14).

CELLULAR MECHANISMS OF TISSUE DAMAGE

Tissue damage in cell-mediated immune reactions may be induced by a variety of cell types including

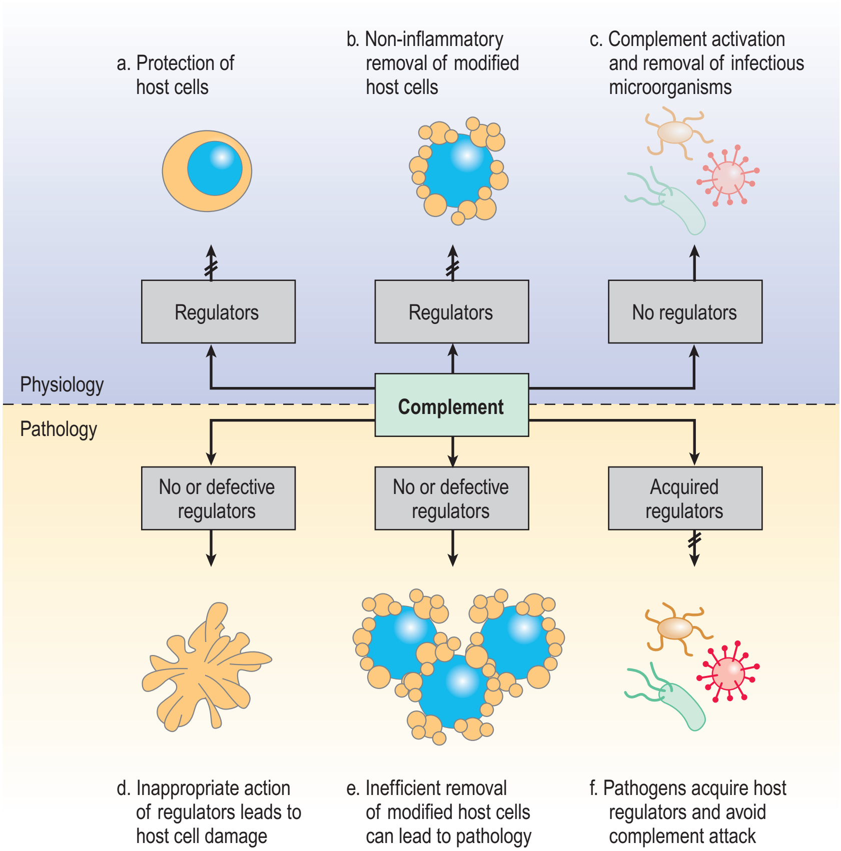
a. Protection of host cells
b. Non-inflammatory removal of modified host cells
c. Complement activation and removal of infectious microorganisms
Regulators
Regulators
No regulators
Physiology
Complement
Pathology
No or defective regulators
No or defective regulators
Acquired regulators
d. Inappropriate action of regulators leads to host cell damage
e. Inefficient removal of modified host cells can lead to pathology
f. Pathogens acquire host regulators and avoid complement attack

FIGURE 7-14 Complement is a double-edged sword for the host. It can either have beneficial effects as in the upper panel where it can protect viable cells (a), remove aged or dying cells (b) or remove infectious organisms (c). If these mechanisms fail, healthy cells can be damaged (d), dead cell debris can accumulate and cause inflammation (e) or pathogens can use complement to evade the immune system (f). In the eye, mechanism (d) is thought to account for diseases such as AMD due to genetic faulty complement proteins such as complement factor H. (From Zipfel and Skerka, 2009.)

macrophages, cytotoxic T cells and NK cells. In addition, macrophages are involved in the clearance of cell debris in acute inflammatory reactions and in antibody-dependent cytotoxicity.

The delayed-type hypersensitivity reaction is the hallmark of cell-mediated immunity

The delayed-type hypersensitivity reaction is a Th1-mediated reaction to foreign and/or autoantigen and is characterized by the presence of granuloma in the tissues. These accumulations of cells contain a central core of macrophages around a vessel with T cells in the surrounding area. Fibrinoid necrotic material may be present in the centre with giant cells (fused macrophages) and epithelioid cells. Such lesions are typical of reactions to mycobacteria and also occur in less well-defined diseases such as sarcoidosis. Similar microgranulomas are typical of sympathetic ophthalmia (Fig. 7-15) and indeed of several chronic posterior uveitis syndromes.

Such granulomas contain many types of T, B and macrophage-like cells, and also have a high content of dendritic cells. It is possible that these cell collections represent small extralymphatic lymphoid follicles where extensive antigen presentation is in progress. Once the antigen has been removed, the granuloma subsides.

Inflammatory (M1-like) macrophages cause tissue damage: alternatively activated (M2-like) macrophages are more likely to promote healing

Macrophages cause tissue damage by release of reactive intermediate metabolic products and tissue proteases, express high levels of adhesion molecules and chemokine receptors such as CCR5 and CCR2 and are generally recruited *de novo* from circulating monocytes (see p. 404). However, not all macrophages behave in this manner. In the resting state, tissues contain resident macrophages that, in certain tissues, may have an immunoregulatory role (such as alveolar macrophages in the lung). In central nervous system tissue, including the retina, specialized resident tissue macrophages occur (microglia) (see Ch. 1, p. 52). These cells may be induced to express MHC class II antigen during inflammation but are still more likely to promote tissue homeostasis. Such cells express the mannose receptor, arginase and the fractalkine receptor (see p. 404).

Alternatively, activated macrophages are more likely to occur in parasitic infections under control by cytokines such as IL-3, IL-5 and IL-13 and promote a predominantly eosinophilic response. In addition, there is more likely to be fibrosis than tissue damage with angiogenesis.

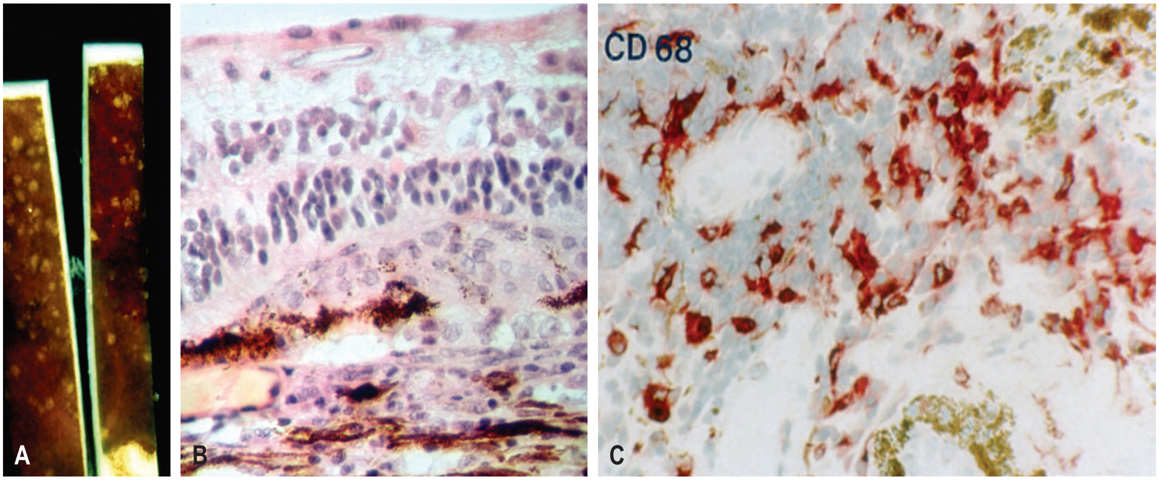

FIGURE 7-15 (**A**) Macroscopic view of a case of sympathetic ophthalmia showing Dalen–Fuchs granulomatous lesions as white 'excrescences'; (**B**) macroscopic view of Dalen–Fuchs granuloma; (**C**) immunohistochemical view of activated (CD68+) myeloid cells in granuloma surrounding an occluded vessel. (Courtesy Prof. W.R. Lee.)

Do cytotoxic T and NK cells induce cellular damage by making holes in the cell membrane?

Killing by cytotoxic T cells is a multistep process. Initial recognition and binding of a target cell (e.g. an infected or mutated cell) is followed by damage to the cell membrane. This lethal insult induces apoptosis with DNA fragmentation and lysis of the cell. The cytotoxic cell then disengages to attack another target cell.

Membrane damage takes the form of pore formation similar to that induced by the complement membrane attack complex (MAC; see above). In this case, however, it is induced by perforin, a protein released from lysosomes of antigen-activated Tc cells that polymerizes in the cell membrane to form a leaky pore (Fig. 7-16). Perforin is released in association with a granule proteoglycan and serine proteases (granzymes), which are also likely to be involved in the lytic process. However, cell death is not automatically caused by osmotic lysis. Instead prelytic DNA breakup occurs as a result of activation of the apoptosis genes and killing occurs via caspase-3 and -7. Therefore there are important differences between MAC-mediated cell lysis and Tc/NK cell killing.

Alternative mechanisms have also been shown for T-cell cytotoxicity. CD8$^+$ T cells may recognize non-peptide (e.g. lipid) antigen in the context of CD1. Thus cells containing bacteria are directly recognized by CD8$^+$ T cells and the bacterial lipid induces the release of a novel enzyme, granulysin, which directly kills the intracellular bacteria.

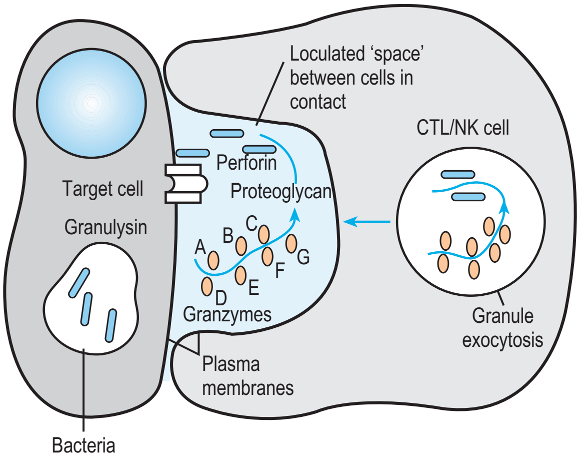

FIGURE 7-16 Perforin killing.

Organization of the immune system

The immune system is designed to provide cells that can circulate freely through the tissues and organs of the body in such a way that they are readily available to mount a defence against foreign organisms at short notice. Immune cells are thus highly motile, normally quiescent cells travelling to and from the lymphoid organs; they can be readily activated if required.

Centralized antigen recognition mechanisms provide the most efficient means of rapid response because all the necessary requirements for cell activation can be concentrated at one site. This takes place in the lymphoid organs (primary and secondary lymphoid organs, SLOs). Some cells carry afferent information concerning possible breaches in the body's defences to the central lymphoid tissues (particularly dendritic cells and other APCs), while effector cells (T and B cells) remove the invading organism (or arrange for this to be done by other cells such as inflammatory macrophages) and restore tissue homeostasis. Trafficking of cells to and from the tissues to the lymphoid organs requires specific receptors on the circulating leucocytes and the vascular endothelium in each tissue (chemokine and adhesion molecule receptors).

Tertiary outposts of local antigen presentation can be set up in sites where normal lymphoid tissues do not exist (such as the brain and eye) or where persistent antigen generates a chronic inflammatory response (frequent sites are liver and lung). Granulomas, the hallmark of chronic inflammation in the tissues usually forming at sites of small blood venules, represent one type of lymphoid cell accumulation where an antigen processing and presentation 'factory' might develop, but true tertiary lymphoid organs (TLOs) containing discrete T- and B-cell areas with a lymphatic as well as a blood circulation, may also develop *de novo* at sites of persistent chronic inflammation.

FUNCTIONAL ANATOMY OF LYMPHOID ORGANS

During development, the primary sources of lymphomyeloid cells are the bone marrow and thymus (primary lymphoid organs). Ultimately, all cells derive from stem cells in the bone marrow. Stem cells are

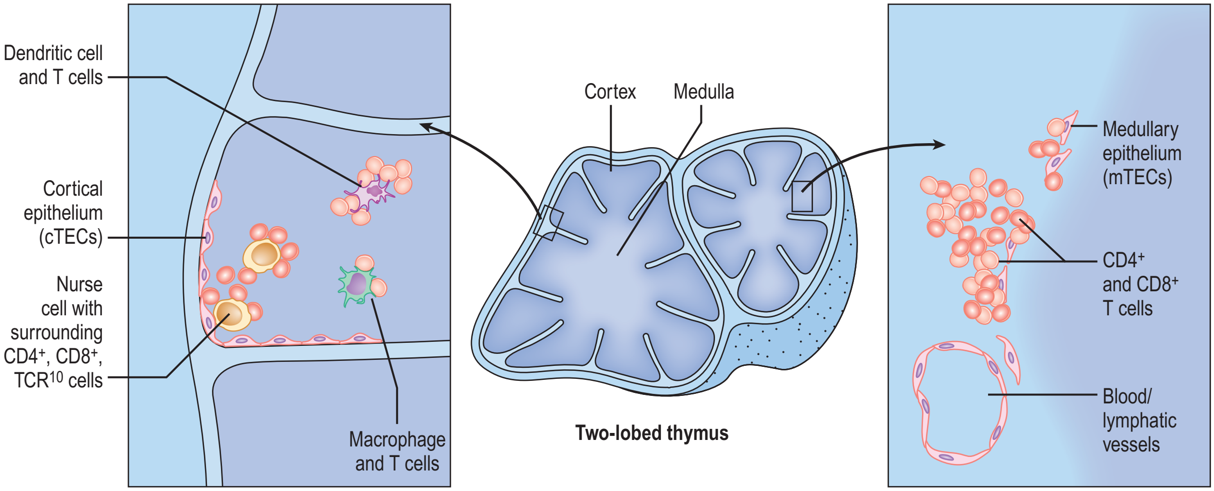
Dendritic cell and T cells
Cortical epithelium (cTECs)
Nurse cell with surrounding CD4+, CD8+, TCR10 cells
Macrophage and T cells
Cortex
Medulla
Two-lobed thymus
Medullary epithelium (mTECs)
CD4+ and CD8+ T cells
Blood/ lymphatic vessels

FIGURE 7-17 The bi-lobed thymus. The central region shows the cortex and the medulla, while the cortical thymic epithelial cells (cTECs) and medullary medullary thymic epithelial cells (mTECs) are shown to left and right in greater detail.

poorly characterized and their existence is based mainly on evidence for cell differentiation, often from *in vitro* cell culture studies. Haematopoietic stem cells have few lineage markers (i.e. they are Lin^-) that give rise to common myeloid progenitors and common lymphoid progenitors. These can be induced to give rise to T, B and NK/NKT cells (via IL-7) or myeloid cells (macrophage/dendritic cells/neutrophils (M-CSF, GM-CSF, G-CSF)). B cells mature in the peripheral/secondary lymphoid tissues, particularly Peyer's patches (lymphoid tissue structures in the wall of the small intestine). T-cell precursors from the bone marrow colonize the thymus, where they undergo selection and lineage differentiation before being distributed to the SLOs.

Bone marrow stem cells produce all blood cells

Blood cells – including red cells and platelets, granulocytes, monocytes and dendritic cells – are produced in the bone marrow and released directly into the circulation. B cells are also released directly into the circulation and circulate between the lymphoid organs and blood.

The marrow is a loose spongy stromal network whose cells, together with local macrophages, release growth factors (cytokines, see above) that initiate differentiation of each cell type. IL-3, G-CSF, GM-CSF and M-CSF are particularly important. IL-1 and IL-6, released by stromal marrow cells, also participate in T-cell maturation, while IL-7 promotes B-cell development. Stromal marrow cells are therefore essential to the survival of the host.

The thymus regulates T-cell development and maturation

T-cell precursors enter the thymic cortex (Fig. 7-17). Here they interact with cortical thymic epithelial cells (cTECs or 'nurse' cells), which encapsulate them in large numbers. At this stage they are still immature cells expressing cell surface markers for Th and Tc cells (CD3^+, CD4^+ and CD8^+: the so-called double-positive cell; see Box 7-4 for explanation of CD numbers). After some time they are released from the nurse/cTEC cells and migrate through the cortex to the medulla, making contact with macrophages and dendritic cells as they go. Many T cells die during this process (clonal deletion and/or apoptosis) but selected cells differentiate to express one or other T-cell phenotypic marker (CD4 or CD8). Maturation in the medulla also involves contact with the medullary thymic epithelial cells (mTECs) cells and thymic dendritic cells, which express high levels of MHC class II antigen. The T cells then enter the blood vessels and migrate to the lymph nodes and spleen.

cTECs and mTECs arise from a common TEC progenitor probably located at the interface between the cortex and medulla. mTECs are important in generation of Tregs as well as deletion of tissue-specific autoreactive T cells through the autoimmune regulator (*AIRE*) gene. Defects in this gene causes a condition in humans known as the autoimmune polyendocrinopathy candidiasis ectodermal dystrophy syndrome (APECED) which includes a low-grade form of uveitis due to the failure to delete T cells reactive against IRBP (see p. 459). The medulla of the thymus contains several characteristic whorled bodies (Hassall's corpuscles) (Fig. 7-17), which are probably remnants of 'used' nurse epithelial cells. In the adult, the thymus involutes but some T-cell maturation continues into adult life, both in the thymic remnant and extrathymically.

Lymph nodes are designed for antigen trapping

Lymphocytes and APCs from the tissues enter the cortex of the lymph node via the afferent lymphatics (Fig. 7-18) where they present antigen to T and B cells.

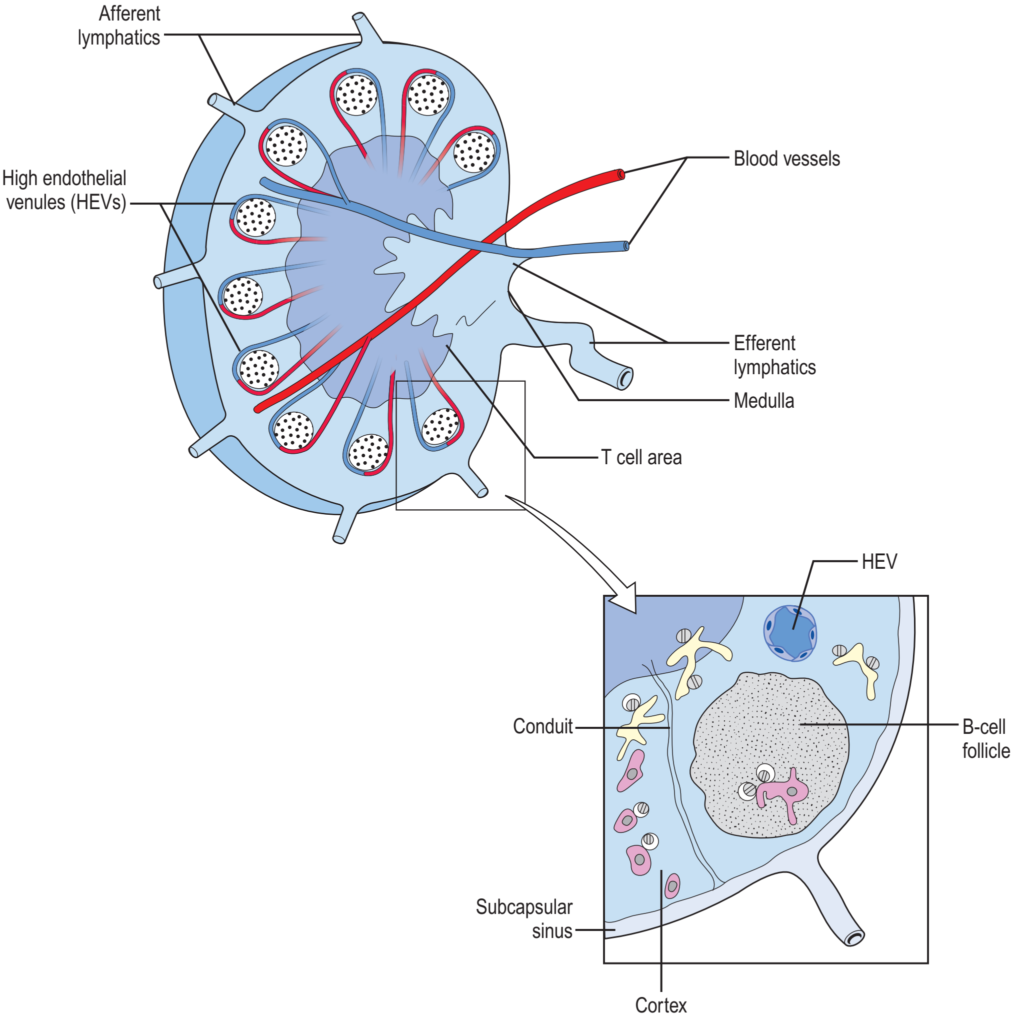
Afferent lymphatics
High endothelial venules (HEVs)
Blood vessels
Efferent lymphatics
Medulla
T cell area
HEV
Conduit
B-cell follicle
Subcapsular sinus
Cortex

FIGURE 7-18 Diagram of the lymph node. The T-cell area surrounds the B-cell follicle in the cortex; high endothelial venules (HEVs) are present in the T-cell area, mostly at the junction with the medulla. Conduits allow trafficking antigen-presenting cells to migrate into the node after entry through the subcapsular sinus.

Follicular dendritic cells (fDCs) in germinal centres of the lymph nodes and the spleen present antigen–antibody complexes to B cells in the B-cell areas, while conventional dendritic cells (cDCs) present antigenic peptides to T cells in the T-cell area. cDCs comprise two main types: migratory DCs (CD8α-) carrying antigen to the lymph node and resident, 'lymphoid' CD8α^{+} DCs which may sample soluble antigen in the afferent lymph or receive antigen from migratory DCs. The eye (and the brain) connect directly with cervical lymph nodes (see Chs 1 and 4), but lymphocytes and APCs from these tissues also find their way to the spleen through the aqueous veins.

Recent studies have indicated that lymph nodes are organized in chains, the more peripheral node receiving antigen from the tissues and activating (or not as appropriate; see section on tolerance, p. 441) T cells, which leave (egress) via the blood vessels and home to the site of inflammation to clear antigen-infected cells. Naive T cells (i.e. non-activated T cells) meanwhile egress through the efferent lymphatic and traffic to the next more central lymph node in the chain until they reach the thoracic duct and re-enter the circulation to begin another cycle of migration.

Antigen is carried in soluble form to the lymph node in blind-ended vascular lymphatics from the tissues, and also in cell-associated form inside APCs (dendritic cells and macrophages). APCs migrating from the tissues to the initial peripheral lymph nodes do not leave the node but present antigen to the LN T and B cells. However, unactivated T and B cells circulate between the lymphatics and the bloodstream as indicated while activated T cells specifically home to sites of inflammation.

The spleen receives antigen from all sources (lymphatics and blood)

The spleen is a central secondary lymphoid organ at the interface between the blood and lymphatic circulations. The bloodstream communicates with the lymphatic system at the thoracic duct where recirculating lymphocytes enter. The spleen receives antigen-primed APCs from all regions including those that may have bypassed their regional lymph node. For instance, APCs that encounter antigen within the anterior chamber of the eye may pass through the trabecular meshwork and aqueous veins (see Ch. 1, p. 24) to enter the conjunctival veins and eventually drain into the lungs and spleen. However, the spleen has no afferent lymphatics, i.e. it only receives lymphocytes from the bloodstream which are overwhelmingly naive, unactivated T cells.

The spleen is organized like a lymph node with the addition of the red pulp (red cell area) and its venous sinuses (Fig. 7-19). T-cell and B-cell areas are separated, with the T-cell areas (white pulp) being particularly large and aggregated around an arteriole (periarteriolar lymphoid sheath, or PALS) and surrounded by well-defined macrophage/dendritic cell zones (marginal zone).

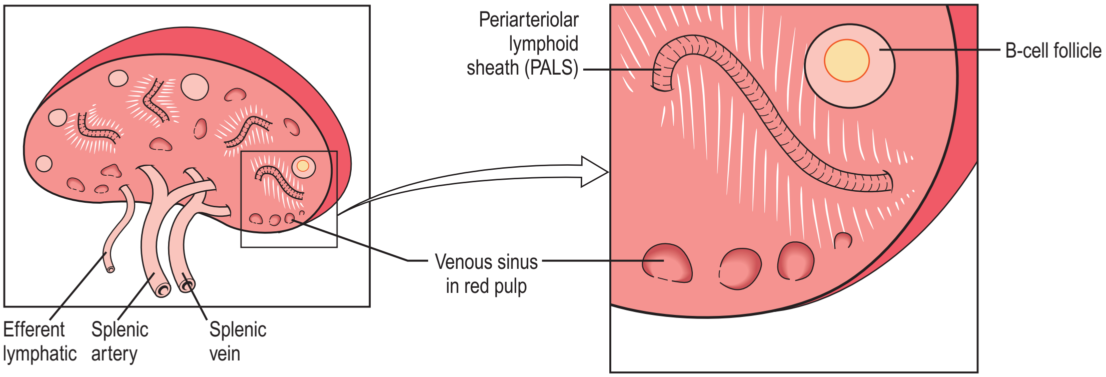

FIGURE 7-19 Diagram of the spleen. The white pulp contains the T-cell area surrounded by the B-cell follicles; the red pulp is situated closer to the medulla and contains the venous sinuses with large numbers of red cells and macrophages. There are no afferent lymphatics to the spleen, all cells entering into the spleen stroma via the bloodstream near the periarteriolar lymphatic sheath and leaving via efferent lymphatics.

Tertiary lymphoid organs

TLOs (see introduction to this section, p. 412) develop around sites of persistent chronic inflammation, spurred on by lymphoid tissue inducer (LTi) cells, similar to those which establish the development of primary and secondary lymphoid tissues which secrete lymphotoxin β (LTb), the essential cytokine for differentiation of the stromal cells which permit the development of the lymphoid organs. Chemokines such as CXCL13 and cytokines such as IL-17 derived from the inflammatory process provide the conditions for the further differentiation of lymphoid vessels and the T- and B-cell separation (Fig. 7-20).

THE MUCOSAL IMMUNE SYSTEM

Initial contact of the host with exogenous antigen takes place at surfaces such as skin and mucous membranes. The skin provides an effective physical barrier to microorganismal contact unless penetration is achieved, as with insect vectors and trauma. However, the mucous membranes are more easily breached by microorganisms and, not surprisingly, considerable lymphoid tissue is concentrated at these surfaces to deal with new antigens as they arrive. This is termed the mucosa-associated lymphoid tissue (MALT) and is specialized to some degree for each tissue, e.g. bronchus-associated lymphoid tissue (BALT), gut-associated lymphoid tissue (GALT) and conjunctiva-associated lymphoid tissue (CALT). In essence it comprises intraepithelial lymphocytes and APCs (especially dendritic cells) and aggregations of lymphoid tissue in the mucosal layers. These are best exemplified by the Peyer's patches of the gastrointestinal mucosa and the tonsils in the oropharynx. Regional specialization occurs in various compartments of the mucosal immune system with specific homing receptors ('addressins', adhesion molecules) directing movement of cells to different mucosal sites (see below) and like TLOs, are induced by innate lymphoid cells (ILCs) (see p. 382), such as LTis as well as a second set of ILCs which are NK-like IL-22 and (possibly) IL-17-secreting cells. In the lamina propria of the gut, $CD4^+$ Th17 cells as well as $CD4^+$ Treg cells abound, and are important cells balancing the response of the host to commensal versus pathogenic microorganisms in the microbiome (see p. 374). Similar immune defence arrangements are likely to occur in all mucosal tissues but are most developed in the gut due to the constant exposure to microorganisms. Disturbance in this regulation accounts for many types of inflammatory bowel disease (IBD), some of which are complicated by systemic involvement including the eye, as in HLA-B27-associated acute anterior uveitis.

As indicated below, MALT is specially adapted for antigen capture and presentation to specific lymphocytes that 'home' to that tissue. In addition, exposure of the host to antigen via the MALT frequently leads to tolerance to that antigen by mechanisms that are as yet poorly understood but involve Tregs (see p. 380).

THE IMMUNE SYSTEM AS A POLICE FORCE

The immune system can be regarded as a surveillance system constantly checking for intruders. Cells of the innate immune system, such as neutrophils and monocytes, undertake part of this function but at a relatively non-specific level. The adaptive immune system has cells that mediate this function in a much more targeted way. The afferent limb of the immune system uses dendritic cells in this role, while the messengers sent out to do the 'dirty work' are the T and B lymphocytes.

Dendritic cells are the major surveillance cells of the afferent lymphoid system

Dendritic cells are derived from stem cell precursors in the bone marrow and proliferate/differentiate under the influence of cytokines such as GM-CSF, stem cell factor (Flt3 ligand) and TNF-α. They leave the bone marrow and circulate in various shapes and sizes (one of which is as very large 'veiled cells', like large floating jellyfish) in the bloodstream before entering the tissues, where they remain for a few days to weeks. In the skin, conjunctiva and peripheral corneal epithelium, they can be identified as Langerhans' cells; in the stroma of these tissues, they occur as migratory cDCs. They interact closely/intimately with tissue cells such as keratinocytes and epithelial cells. In the uveal tract, there is a rich network of dendritic cells in association with tissue macrophages (see Ch. 1, p. 28).

Dendritic cells from the tissues constitutively migrate in the afferent lymph to the lymph nodes and

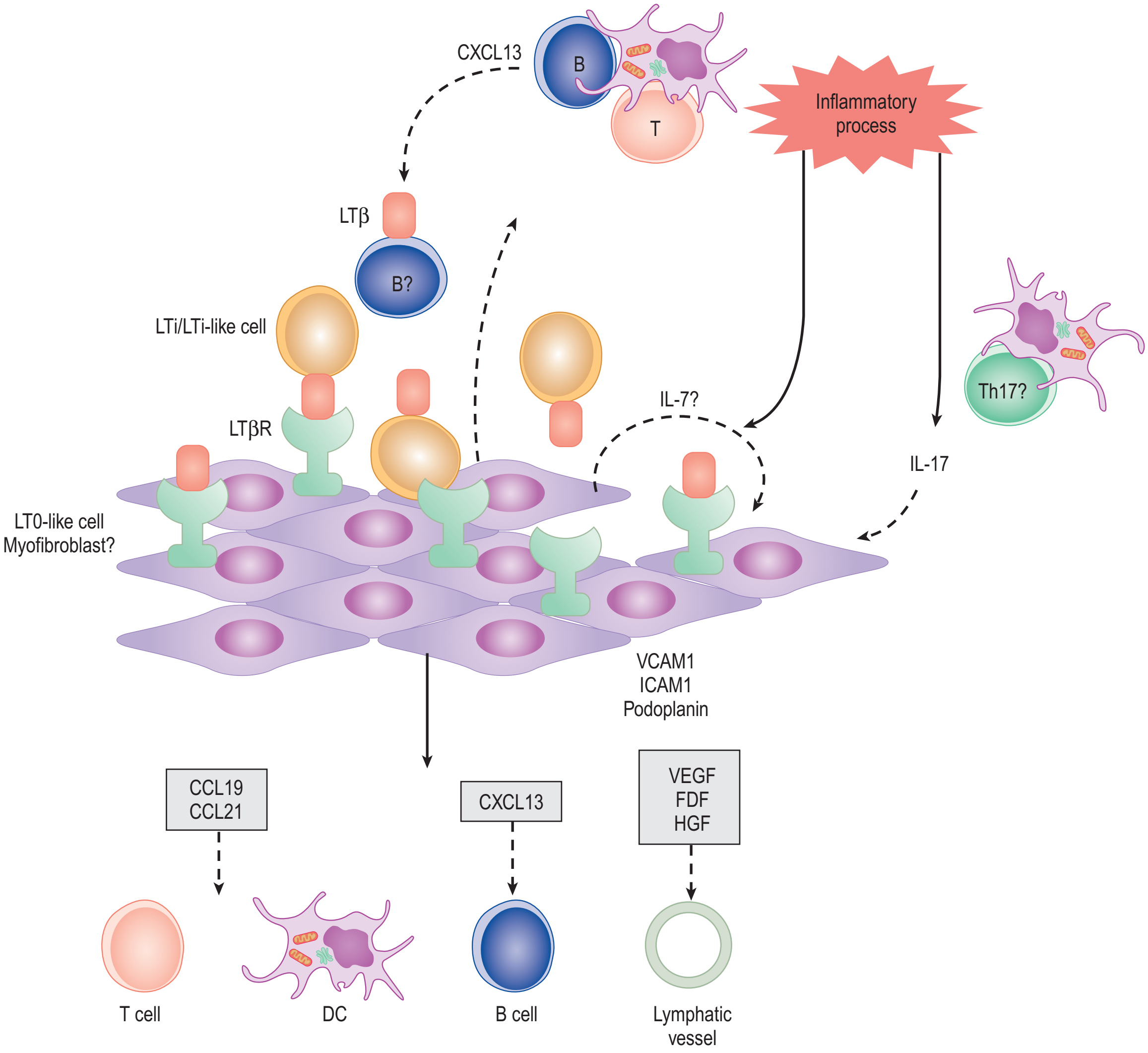
CXCL13
B
T
Inflammatory process
LTβ
B?
LTi/LTi-like cell
LTβR
IL-7?
Th17?
IL-17
LT0-like cell
Myofibroblast?
VCAM1
ICAM1
Podoplanin
CCL19
CCL21
CXCL13
VEGF
FDF
HGF
T cell
DC
B cell
Lymphatic vessel

A **TLO formation**

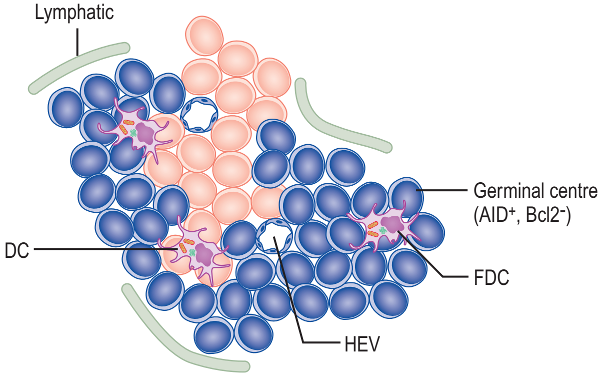
Lymphatic
Germinal centre
(AID⁺, Bcl2⁻)
DC
FDC
HEV

B **TLO structure**

FIGURE 7-20 Tertiary lymphoid structures develop during chronic inflammation and take on a form similar to secondary lymphoid structures such as the lymph node and Peyer's patches. In a sense they are outposts of the lymphoid defence system but often do not have the full characteristics of the structured lymph node. However, they can develop germinal centres and afferent and efferent lymphatics. The granuloma has similarities but is different since it contains a central core of antigen with myeloid cells and a large cuff of T cells with some B cells. (From Neyt et al., 2012.)

spleen (migratory CD8α-cDC) in response to appropriate stimulation (e.g. TNF-α); they interact with T cells in the T-cell areas of these tissues (interdigitating dendritic cells). In contrast, cDCs which have directly arrived in the SLOs from the bloodstream without passing through other tissues, become resident CD8α-cDCs and take up antigen from other APCs which have reached lymphoid organs and undergone apoptosis. Other specialized dendritic cells that possess Fc receptors and serve as antigen traps because they have the ability to bind immune complexes occupy the follicles in lymphoid tissue where they present antigen to B cells (follicular dendritic cells, fDCs). Dendritic cells do not leave the peripheral lymphoid tissue; they are short-lived cells with a half-life of a few days and their numbers are replenished from the bone marrow.

A third set of DCs, termed plasmacytoid DCs (pDCs, so named because they had some morphological similarities to plasma cells), arrive in the SLO from the blood and are programmed to constitutively produce INF-α, which is massively increased in response to viral infections. They may also be involved in tolerance induction (see p. 441).

Trafficking of cells to and from the lymphoid system depends on specific cell surface adhesion molecules

In contrast to DCs, T and B cells recirculate through the lymphoid organs many times. They do this by binding to specific regions of the lymphoid vasculature, the high endothelial venules (HEVs) (see Fig. 7-18), which are specialized for lymphocyte capture by the expression of specific adhesion molecules on their surface. About 25% of the lymphocytes passing through a lymph node will leave the bloodstream and any given naive lymphocyte can traverse each lymph node at least once a day. However, in practice, lymphocyte recirculation follows some patterns: B cells rarely recirculate, γδ T cells tend to exist in skin, and most other T lymphocytes recirculate preferentially through their lymph node of origin. If they fail to meet cognate antigen (antigen specific to that T cell's T-cell receptor), they remain unactivated naive cells and pass on to the next lymph node in the chain before finally reaching the thoracic duct. Activated lymphocytes, on the other hand, egress into the bloodstream and 'home' to (more likely are 'captured' by) HEV-like structures which develop in non-lymphoid vascular endothelium at sites of inflammation (e.g. in retinal vessels during the active phase of retinal vasculitis; see Ch. 9). Activated T cells can circulate many hundreds or thousands of times before they are 'captured' (it has been estimated to take at least 16 hours from time of entry into the circulation to time of entry into the retina in the mouse). Receptors for specific adhesion molecules are reciprocally expressed on activated lymphocytes (and indeed on all classes of inflammatory cells during acute inflammation; see above), which assist in directing them to sites of tissue injury.

Specific adhesion molecules are expressed on HEV in lymph nodes, where they are involved in physiological lymphocyte recirculation. These molecules are therefore sometimes referred to as 'addressins' (one address only!). For each addressin there is a corresponding lymphocyte 'homing receptor'. Circulation through typical lymph nodes is mediated by L-selectin on lymphocytes, which binds to the adhesion molecule GlyCAM-1 on the HEV. In Peyer's patch lymph nodes, HEVs express MAdCAM (mucosal addressin cell adhesion molecule) and VCAM-1 (see Table 7-3), which are the ligands for $\alpha_4\beta_7$ and $\alpha_4\beta_1$ integrins of lymphocytic microvilli. It is likely that each tissue has specific addressins that direct the circulation of surveillance lymphocytes through that tissue; in particular, most mucosal lymphoid tissue (MALT, CALT; see above) has addressins that ensure adequate trafficking of lymphocytes through that tissue during health and disease.

Some cytokines and chemokines facilitate the interaction of T cells with activated endothelium, such as macrophage inflammatory protein-1β (CCL4), which has been shown, for instance, to be particularly effective in mediating $CD8^+$ T-cell adhesion to the endothelium (see section on chemokines above). In addition, certain adhesion molecules are involved not only in endothelial cell interactions but also in mediating the close cell–cell contact required during antigen presentation (see below), particularly ICAM-1/LFA-1, ICAM-1/LFA-3, CD44 and its ligand cell surface hyaluronan, plus several others (see Table 7-3). In addition, chemokines and their receptors are intimately involved in the recirculation and trafficking of T and B lymphocytes and of APCs. In general, however, it is now clear that adhesion of leucocytes, either to

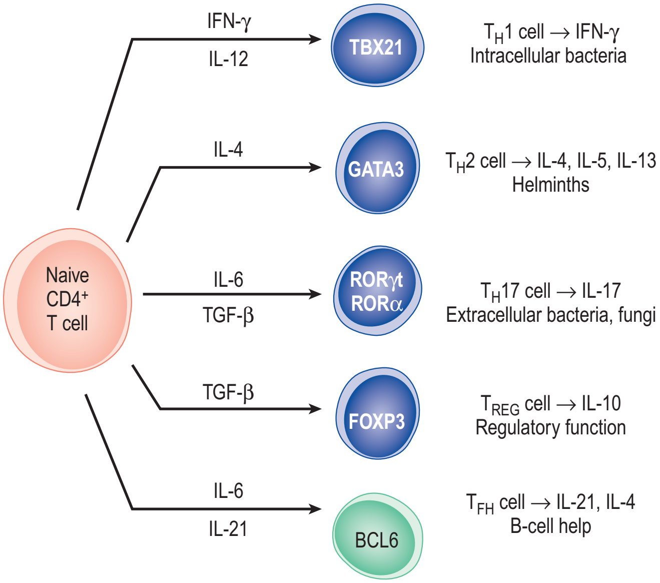

FIGURE 7-21 The local microenvironment provides signals which permit differentiation of 'newborn' naive T cells emerging from the thymus. The cytokines signal via specific receptors and activate transcription factors specific for each cell type which sets in motion a protein synthesis programme leading to production of factors specific to that newly differentiated T cell. BCL6, B-cell lymphoma 6; FOXP3, forkhead box protein 3; GATA3, GATA-binding factor 3; RORα, retinoid-related orphan receptor-α; RORγt, retinoid-related orphan receptor-γt; TBX21, T-box transcription factor TBX21; T_{FH}, follicular helper T; TGF-β, transforming growth factor-β; T_H1, T-helper-1; T_H2, T-helper-2; T_H17, T-helper-17; T_{REG}, regulatory T. (From Craft, 2012.)

each other or to the endothelium, during normal recirculation or as part of inflammation, is a complex molecular and tissue-specific process.

What turns a lymphocyte on?

Most circulating lymphocytes are resting. Most are 'naive' cells (i.e. have not been activated by interaction with antigen) and have recently been released from the thymus or have passaged through the peripheral lymphoid tissues. Others are memory T and B cells that can be very long-lived. Memory cells arise as a residual population after most of the antigen-specific T cells have died once their effector duties are completed (i.e. getting rid of the pathogen). Both CD4⁺ and CD8⁺ T memory cells occur, some residing long term in the tissues (e.g. the lung parenchyma) where they can rapidly respond to re-challenge with the antigen (e.g. influenza virus). These cells are known as effector memory T cells. Other memory cells recirculate through the lymphoid tissues and are known as central memory cells.

The main stimulus that activates T cells (both naive and memory cells) is antigen presented in the form of peptide by professional APCs in the lymphoid tissue. Activation of the T cell induces it to release cytokines, particularly IL-2, which initiates the process of clonal expansion. Depending on the local conditions and the cocktail or 'panel' of cytokines produced by the T cells and by stromal cells in the lymph nodes, further functional diversification is achieved by driving T-cell clonal expansion towards one of several subsets (Fig. 7-21).

Th1 cells, expanded by IL-2 and IFN-γ, induce a delayed-type hypersensitivity (DTH) response (type IV hypersensitivity in the classic nomenclature) (Table 7-7), which involves considerable macrophage activation, granuloma formation and tissue damage. It should be noted, however, that memory T cells do not release IFN-γ, while release of the chemokine RANTES (CCL5, see Table 7-5) by tissue cells appears to be important in the recruitment of memory T cells and macrophages to the site of granuloma formation. IL-12, IL-23 and IL-27 produced by macrophages and dendritic cells also play a major role in directing Th1/Th17 T-cell responses.

In contrast, Th2 cells, expanded by IL-2 and IL-4, provide B-cell 'help' and lead to the production of antibodies. In addition, IL-4 inhibits Th1 cell activation and the delayed-type hypersensitivity reaction. IL-10 and IL-13 produced locally in tissues and lymph nodes may condition APCs to induce a Th2 response. IL-2 and TGF-β combine to promote Treg formation, while IL-6 and TGF-β promote Th17 T cells.

TABLE 7-7 **Immunopathology of tissue reactions**

Hypersensitivity response type	Immune effector process in tissue	Type of mechanism
I	Allergic reaction	Humoral (IgE)
II	Cytotoxicity	Humoral (IgG/M)*
III	Complement	Humoral (Ag/Ig)
IV	Macrophages	Cellular (T cells)†

*Cytotoxicity in this process is antibody-dependent.
†Type IV hypersensitivity is the classically described delayed hypersensitivity response, mediated by macrophages and memory T cells.

Th1/Th17 (DTH) and Th2 cells (allergy, see p. 445) are directed by particular sets of cytokines which determine the nature of the pathological event. Classically, four types are described (types I, II, III and IV hypersensitivity; see Table 7-7) but mixed responses also occur. Th1/Th17 and Th2 cells may initiate reciprocal downregulation of either response via regulated cytokine release. Chemokines and chemokine receptor expression also play a major part in determining whether a Th1/Th17 or a Th2 response is induced (see p. 404).

Where are memory cells found?

T effector cells that exit the lymph node lose their expression of L-selectin and cannot therefore recirculate through the lymph node. Most of them home to the site of inflammation to promote removal of the antigen and then die *in situ*; as indicated above, a small number of these cells have the potential to survive as memory cells, some of which re-enter the lymph–blood circuit as central memory cells while others remain long term in the tissues as effector memory cells. In contrast, memory B cells do not recirculate but continue to produce antibody of different isotype, with increasing affinity for the antigen, and release it into the blood via the efferent lymphatics. Memory cells for the mucosal lymphoid tissues have been shown to express a specific addressin, LPAM-1 (lymphocyte Peyer's patch adhesion molecule-1), or $\alpha_4\beta_7$ integrin which binds to MAdCAM-1 in the GALT (see p. 416). IL-7, an essential cytokine for ontogeny of T cells in mice, is also required for induction of memory cells.

What turns a lymphocyte off?

Most immune reactions do not persist, and it has been presumed that this is the result of effective removal of antigen. Indeed, the corollary is also probably true: that most chronic immunological diseases are the result of persistence of antigen in the tissues in a form that activates lymphocytes, i.e. on the surface of APCs. However, it is now recognized that switching off an immune response is an active process, involving a process known as activation-induced cell death (AICD) as well as a subset of T cells, i.e. T regulatory cells (Tregs).

T regulatory cells. Lymphocytes that are not exposed to their cognate antigen in the appropriate form (i.e. on the surface of APCs and in the presence of co-stimulatory support mechanisms) enter a state of anergy and eventually die out. The process of cell death or apoptosis during lymphocyte ontogeny is well established and is associated with the expression of specific 'suicide' or 'death' genes in the cell (see Box 4.4, p. 168).

Tregs also switch off inflammatory responses. Tregs are the most frequent cell type in the circulation and in the secondary lymphoid organs. Antigen-specific Tregs are induced as part of all-adaptive immune responses to exogenous and endogenous (including self) antigens but these expand with a delayed kinetic compared to effector cells, allowing T effectors cells to finish clearing damage-inducing agents (DAMPs, see p. 389). Tregs are characterized by expression of high levels of the IL-2 receptor (CD25) and transcription factor FoxP3 which is essential for their function. Tregs are generated as in the thymus (natural or thymic Tregs) while inducible (iTrgs) are also generated in the periphery from naive CD4$^+$ T cells after presentation of antigen by APCs. Antigen-specific iTregs are more effective than thymic or natural Tregs but are relatively non-specific in their action (i.e. although they are induced by specific antigen, they can suppress T cells which have been activated by other antigens), i.e. they have a broad-spectrum effect. IL-10 and TGF-β are important immunosuppressive cytokines secreted by Tregs. They are present in mice and humans and currently there is considerable interest in developing methods to induce them to high levels so as to prevent a range of autoimmune diseases.

CD4$^+$ CD25$^+$ T regulatory cells are only one set of 'suppressive' T cells that exist; there are also other T cells that produce high levels of IL-10 (Tr1 cells) and TGF-β (Th3) cells, some CD8 T regulatory cells, which may be involved in aspects of ocular immune privilege and require NK T cells to mediate their activity (see p. 390), as well as some recently described B regulatory cells (Bregs).

Antigen recognition

The immune system recognizes antigen through molecular interaction with specific molecules, three of which are antibody (B-cell receptor, sIg), the TCR, and the MHC molecule. MHC molecules are highly polymorphic: i.e. although basically similar in structure, there are many different types (alleles) that occur in each individual, generated by small differences in nucleotide sequences in the MHC genes. Each TCR, antibody and MHC molecule has varying degrees of ligand-binding specificity, the MHC molecule being required to bind to a large variety of antigens, while antibody is specific for only one antigen. The interactions between antigen and antibody, and antigen and the TCR, have, however, many similarities.

The great majority of immune responses are T-cell-dependent in that the final effector response, including B-cell responses, is achieved through an initial interaction between a resting T cell and an APC (Fig. 7-22).

THE APC MAKES THE ANTIGEN RECOGNIZABLE

The function of the APC is to capture antigen and to process it into a form that can be recognized by the T cell. The immature (veiled) dendritic cell with its numerous cell surface receptors is ideally designed for this (Fig. 7-23). This it does by partially digesting the antigen into short peptides and complexing the peptide with MHC molecules. Classically, intracellular antigens are complexed with MHC class I molecules for presentation to cytotoxic (CD8$^+$) T cells, while extracellular antigens are complexed with MHC class II molecules for presentation to helper (CD4$^+$) T cells. However, 'cross-presentation' of antigen from either the extracellular or intracellular pathway to pathway can also occur, i.e. both intracellular and extracellular antigens can be presented on either MHC class 1 or MHC class II (see below).

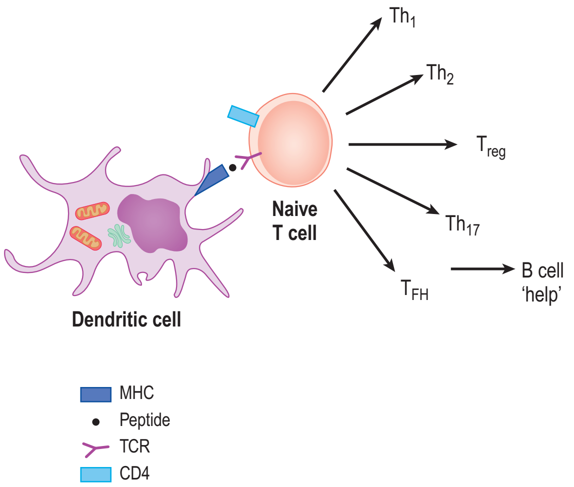

FIGURE 7-22 Diagram of interaction between dendritic cells presenting peptide on the MHC molecule to a naive T cell prior to T-cell differentiation.

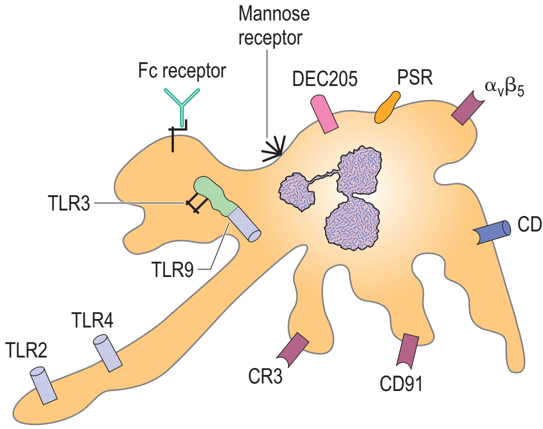

FIGURE 7-23 Diagram of immature dendritic cell showing some cell surface receptors for innate immune responses. DEC205 (CD205); PSR, phosphatidylserine receptor; $\alpha_v\beta_5$ integrin; TLR, Toll-like receptor.

Each APC can present hundreds of thousands of MHC–peptide complexes on its surface. In addition, some non-protein lipid antigens, such as those in mycobacterial cell walls, can be presented atypically on MHC class I-like molecules, the CD1 receptor. Most of the antigens expressed on an APC surface are self-antigens, which promote tolerance (see later). Immunogenic peptides occur at a level of 10–100 copies per cell.

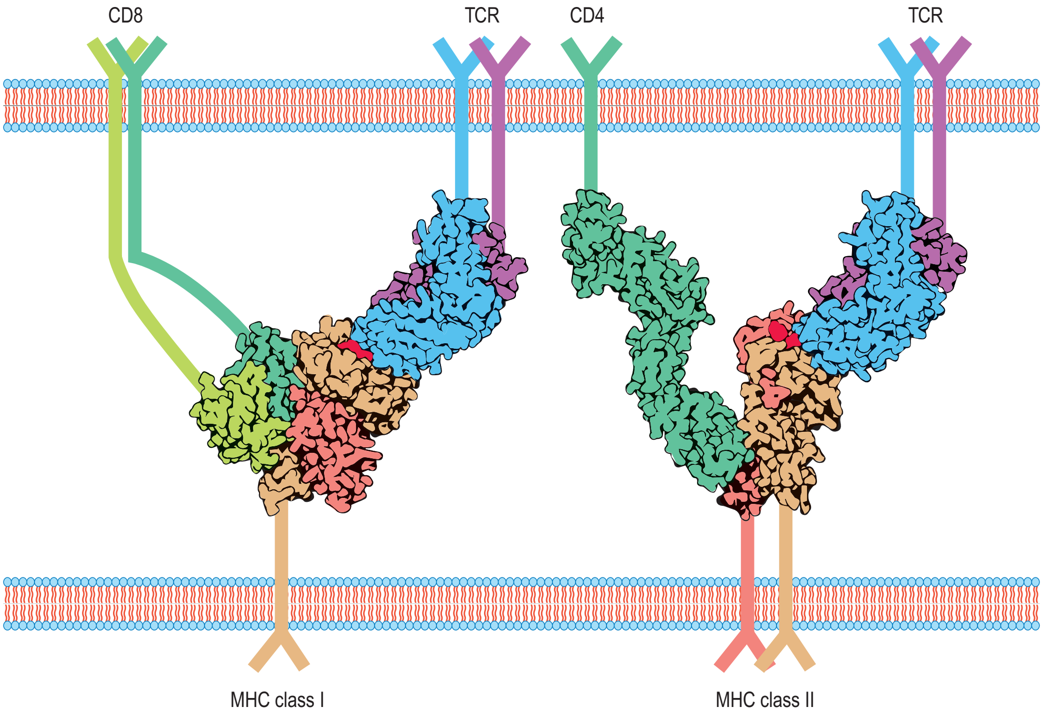

FIGURE 7-24 Models of antigen presentation. The antigen-presenting cell membrane (lower grey band) houses the MHC molecules and presents peptide (small red shapes) embedded in the MHC 'groove' to the T cell: the composite MHC–peptide ligand is bound by the T cell receptor (TCR) while the entire structure is stabilized by CD8 molecules on cytotoxic CD8 T cells and CD4 molecules on CD4 helper T cells. (Image courtesy of: http://www.rcsb.org/pdb/101/motm.do?momID=63 From the World Wide Protein Data bank.)

Processing by APCs is under tight cellular control

Intracellular antigens that are likely to be processed and bound to MHC class I mostly comprise host cytoplasmic and nuclear proteins, digested during normal 'housekeeping' cellular repair work mediated through autophagy. These do not normally induce immunogenic responses. However, intracellular foreign material such as viruses in infected cells is similarly processed via an ATP-dependent structure (the proteasome) and bound to the MHC class I molecule in the endoplasmic reticulum. The proteasome is a constituent of all cells and accepts ubiquinated peptides and proteins for degradation (see Ch. 4, eFig. 4-3); in APCs, the 'immunoproteasome' is specialized to degrade antigens into a set of overlapping peptides. For this purpose it contains two MHC-related proteins, LMP-1 and LMP-2. The peptide is bound specifically to a groove formed by the α_1 and α_2 helical domains of the MHC class I molecule, and is anchored at specific sites to the β-pleated sheets that form the floor of the groove (Fig. 7-24 and eFig. 7-5).

Additional content available at https://expertconsult.inkling.com/.

The CD8 and CD4 molecules help to stabilize the TCR:MHC peptide interaction in the MHC class I and MHC class II molecules, respectively. Defects in antigen processing can be associated with disease: for instance, uveitis in certain patients with the joint disease ankylosing spondylitis and severe rheumatoid arthritis is linked to abnormality in the genes coding for the proteasome and in *LMP-2* genes involved in MHC class I processing.

Additional content available at https://expertconsult.inkling.com/

This illustrates how each step in the antigen-processing pathway (eFig. 7-5) is essential in ensuring correct presentation of 'normal' non-immunogenic host peptides rather than peptides that may induce inflammatory disease. This also applies to the other

The proteasome is a general protein processing machine which degrades exogenous and endogenous molecules to peptides and transfers them to other organelles for further processing, in lysosomes finally to amino acids (see Ch. 4, eFig. 4-3). In antigen-presenting cells the proteasome prepares peptides for presentation via MHC molecules to T cells but this process can be disrupted, for instance by infectious organisms such as cytomegalovirus (CMV), which produces proteins that prevent proper preparation of the peptide. This can occur at several stages, for instance by the viral proteins pp65 and US3-10. These act at various stages either prior to entry of the peptide into the endoplasmic reticulum or before entry of the composite MHC–peptide ligand into the secretory machinery of the cell for ultimate expression on the cell surface (eFig. 7-5).

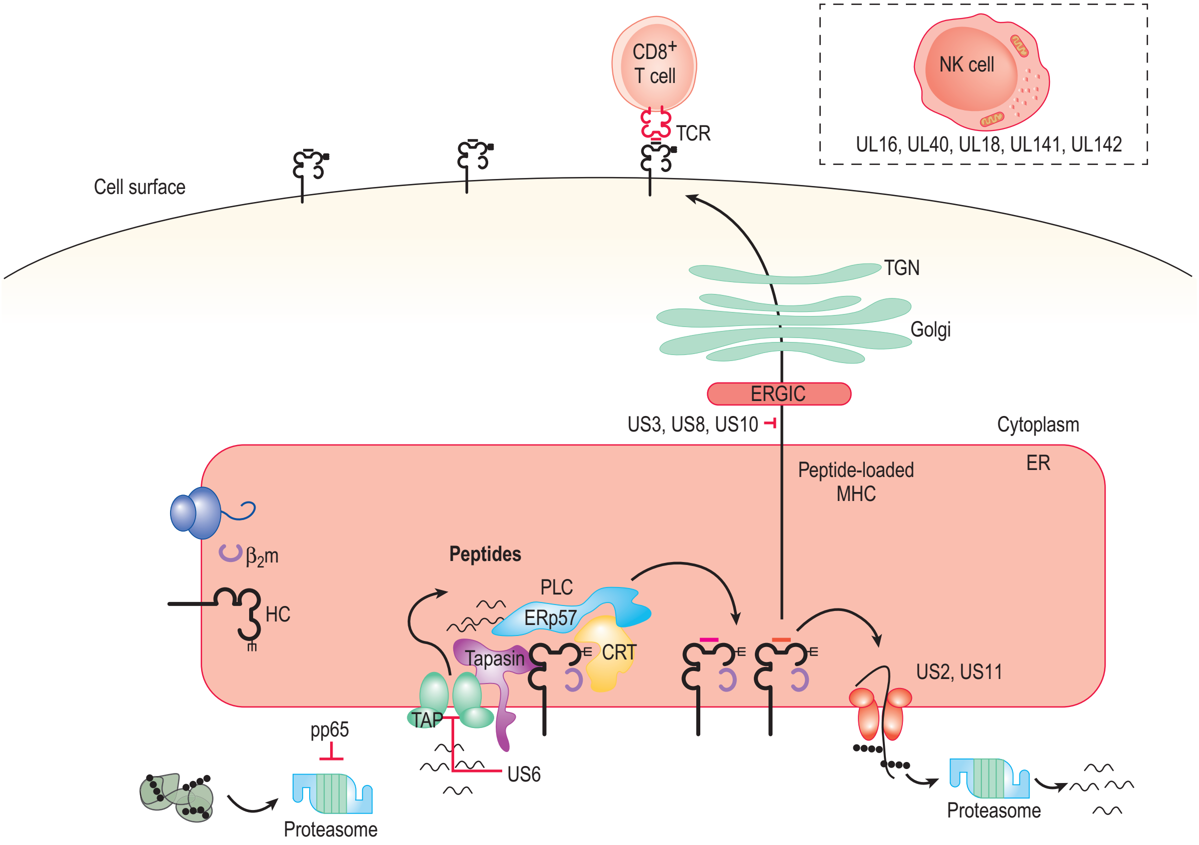
CD8+ T cell
TCR
NK cell
UL16, UL40, UL18, UL141, UL142
Cell surface
TGN
Golgi
ERGIC
US3, US8, US10
Cytoplasm
ER
Peptide-loaded MHC
β_2m
HC
Peptides
PLC
ERp57
Tapasin
CRT
US2, US11
TAP
pp65
US6
Proteasome
Proteasome

eFIGURE 7-5 The proteasome (see Ch. 4) is a 'peptide processing factory' which degrades proteins by ubiquitination and prepares the peptides for incorporation onto the MHC molecules for ultimate presentation on the cell surface. CMV interferes with class I MHC antigen presentation by generating 'immunoevasins' to block CD8 T-cell cytotoxicity. The immunoevasins are the U-labelled proteins and pp65. TCR, T-cell receptor; PLC, peptide-loading complex; TAP, transporter associated with antigen presentation; CRT, calreticulin.

genes that regulate the transport of peptides during their intracellular pathway, such as *TAP1* and *TAP2* (eFig. 7-5).

Loading of peptides via TAP on to the MHC class I molecule is regulated by a further membrane protein, tapasin, while loading of peptide on to the MHC class II molecule is 'chaperoned' by the invariant chains.

Although the processing and presentation of intracellular and extracellular antigens have many similarities in detail, there are some fundamental differences. For instance, the term APC refers to cells that express high levels of MHC class II antigen and 'professionally' process extracellular antigen. The term is thus restricted to a few cell types such as macrophages, B cells and dendritic cells. In contrast, the presentation of intracellular antigen to $CD8^+$ T cells on MHC class I has until recently been more difficult to explain. Most cells of the body express low levels of MHC class I and this can be upregulated in cells exposed to cytokines such as IL-1 and IFN-γ. However, effective antigen presentation requires co-stimulatory molecules, which are not normally present on healthy parenchymal cells. When antigen is presented in the absence of these molecules, the effect is to tolerize the T cell to the antigen.

This is fine for self-antigens but defeats the purpose when intracellular foreign antigen has to be dealt with. Ideally, foreign antigen should be presented when co-stimulatory molecules are also expressed in order to ensure an immunogenic, pro-inflammatory response mediated by activated T cells. This is achieved by cross-presentation. Tissue cells infected by virus (or other pathogens) are killed and APCs, especially dendritic cells, phagocytose dead or dying infected tissue cells; the viral or bacterial antigens are then processed within the dendritic cells for presentation on dendritic cell MHC class I antigen in the presence of co-stimulatory molecules (see below). Receptors involved in uptake of dead and dying cells (apoptotic cells) include $\alpha_V\beta_5$ integrin, CD91 and CD36 (scavenger receptors), CR3 (if coated with complement) and phosphatidylserine (see Ch. 4, p. 168, apoptosis pathways). Cross-presentation involves degradation of polypeptides in the proteasome (Fig. 7-25) and may be a major route for activation of CD8 T cells, as well as activation of CD4 T cells on MHC class II antigens. In this way $CD8^+$ cytotoxic T cells can be directly activated without requiring T-cell help and are therefore armed to induce killing in other infected tissue cells. Thus, dendritic cells deal with both extracellular and intracellular foreign antigen while the response to self-antigen is minimized.

Making the antigen presentable: the MHC molecule as candy wrapper

Presentation of the peptide by the MHC molecule to the TCR requires binding of the peptide to the groove in the MHC molecule and has some degree of specificity. Thus, only certain peptides will bind to each MHC allotype, of which there are many (see below). The specificity of the interaction between the MHC molecule and the peptide, however, is orders of magnitude less than that for the TCR, which is in turn considerably less specific than antigen–antibody binding. Preferential binding of certain peptides to different MHC molecules, however, is well established both for peptide length and sequence.

Additional content available at https://expertconsult.inkling.com/.

Differences also occur in binding of the TCR to MHC class I and II molecules. For instance, only a single MHC class I complex is required to present peptide to the $CD8^+$ T cell but two or more MHC class II–peptide complexes combine with two or more TCRs to activate $CD4^+$ T cells in the immunological synapse (see below). Binding of peptide to the TCR has many resemblances to antigen–antibody binding in terms of specificity but peptide binding affinities vary greatly for the TCR (see section on T-cell activation, p. 427). In the SLO, T cells repeatedly 'sample' peptides presented on MHC molecules on APCs by making brief interactions which become increasingly prolonged as the TCR : MHC complex 'finds' a perfect fit.

In terms of the peptide, this is recognized as 'immunodominance'. For any given antigen, only one in 2000 peptides is likely to have sufficient binding affinity to make that perfect fit and initiate an immune response. Moreover, it is unclear what represents optimal affinity of the peptide for the receptor: both too little and too much can prevent T-cell activation. This variable peptide affinity is reflected in the TCR, which can bind reciprocally many different peptides: thus, specificity is much less rigorous for the T cell than the B cell. These many factors, i.e. immunodominance of peptides, TCR degeneracy with regard to

The size of the peptide is much more restrictive for MHC class I than for MHC class II. Class I molecules will only bind peptides that are 9 or 10 amino acids long, while class II molecules will bind those of any size, but usually between 16 mer and 30 mer. Class I peptides are anchored to a deep pocket in the groove at the second residue. The C-terminal end of the peptide is also bound to a shallower pocket, normally at the ninth residue. This leaves the peptide essentially free in the middle, apart from some less strong side-pocket binding; as a result, the peptide is usually arched in the middle with its amino acid residues projecting outwards to the TCR (Fig. 7-24). It is these exposed residues that determine the specificity of the reaction with the TCR.

In contrast, peptides in the class II groove overlie the sides of the molecule. Binding to the class II molecule, however, is restricted to the same number of residues as for class I (i.e. 9 or 10), except that they occupy the central portion of the peptide. Anchoring at the second residue is the strongest, frequently mediated by the non-polar residue proline. A further difference is that anchoring of the peptide to the class II molecule occurs at positions 2, 4, 6 and 9, with greater side-pocket interaction. This has the effect of straightening the peptide to take up a twisted, linear conformation.

These differences in class I and II peptide binding can be accounted for by specific amino acid sequences. In addition, the affinity of a peptide for a particular MHC allele is dependent on this sequence and is determined by how well each peptide fits into the pocket in the groove. Clearly, peptides composed of amino acids with large or highly charged side chains will have different requirements for binding than those with smaller amino acids, which might fit easily into the pocket.

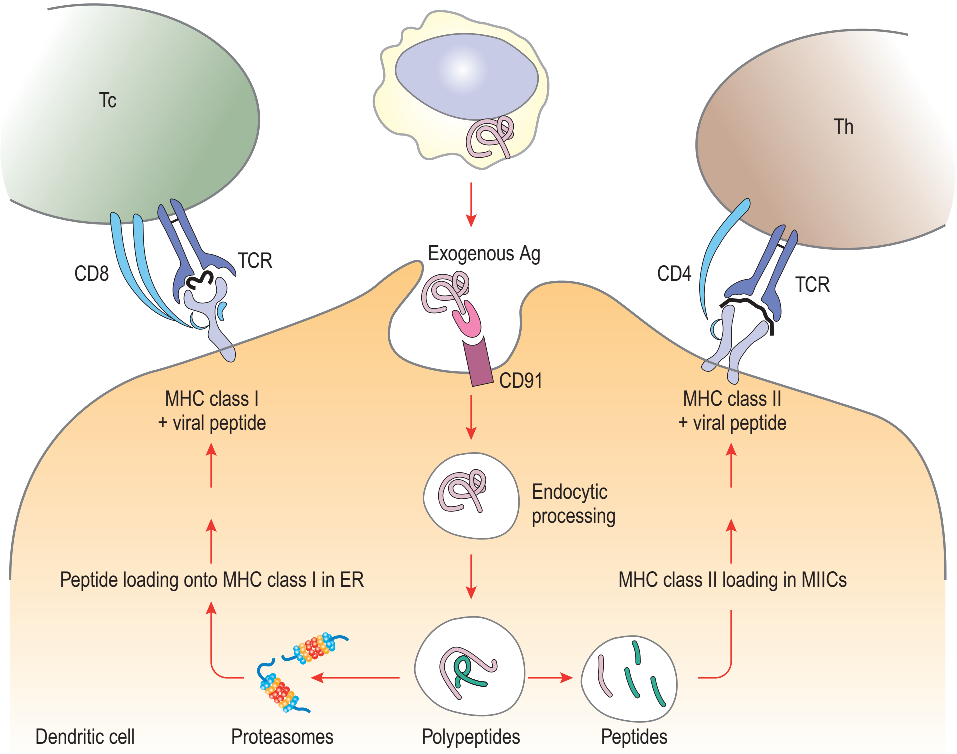

FIGURE 7-25 Cross-presentation. In the upper centre of the figure an endogenously produced antigen (Ag) that eventually ends up in cellular debris has been captured by a scavenger receptor (CD91) on a dendritic cell. The antigen is internalized into the endocytic system and polypeptide fragments are produced. Some of these fragments remain in the exogenous processing pathway (right) and are degraded to peptides that are loaded on to MHC class II in MIICs. These peptides are presented to $CD4^+$ T cells. However, some of the polypeptide fragments are released from the endosomes into the cytosol (left), where they are taken up by proteasomes and enter the endogenous processing pathway. These peptides are loaded on to MHC class I in the endoplasmic reticulum (ER) and are cross-presented to $CD8^+$ T cells. (From Mak and Saunders 2006, with permission from Elsevier.)

specificity and the need for repeated TCR activation as well as cross-linking more than one TCR in the immunological synapse before cell signalling can occur, all play a part in determining whether an immune response takes place and indeed what type of T cell is induced.

The major histocompatibility system

As indicated above, most antigens initiate immune responses by being degraded inside the APC into small peptides, which are then complexed with MHC molecules and presented on the surface of the APC to T cells. So what are MHC molecules? MHC molecules were discovered by Peter Medawar during his studies on rejection of transplants between individuals of the same species but different genetic make-up, and are now known to determine not only the susceptibility to graft rejection but also to infectious and autoimmune diseases generally.

WHAT ARE MHC ANTIGENS AND WHERE ARE THEY FOUND?

The MHC gene cluster on human chromosome 6 is a region of highly polymorphic genes whose products are expressed on a variety of cells. Class I genes are termed HLA-A, -B and -C, while those of class II are known as HLA-D. Class III genes lie between the centromeric class II genes and the telomeric class I genes. MHC genes differ from typical germline genes because they are polymorphic and responsible for some of the traits that distinguish one individual from

BOX 7-11 EXTENDED MAP OF THE MHC GENES

The map of the MHC genes is constantly being revised as new information is added (see figure). The HLA-D genes are located towards the centromere and are separated from the class I genes by the complement protein genes (class III). Each of these regions is known to contain multiple genes, which are variably present within each haplotype. For instance the DR4 haplotype contains DRB1, DRB7, DRB8, DRB4 and DRB9; in contrast, the DR1 haplotype contains DRB1, DRB6 and DRB9.

Several associated genes are located close to the MHC genes. These include the TAP1 and TAP2 genes, the DMA and DMB genes, and the LMP-2 and LMP-7 proteasome genes, all of which are involved in antigen processing (see p. 422e1). The TNF genes are also located in the class III region, as are the RAGE (receptor for advanced glycation end products) genes.

The human MHC gene map has been extended at both the class I region and the class II region to include several additional genes such as histone genes, genes for zinc finger proteins and the important regulatory gene retinoid X receptor gene (RXRB).

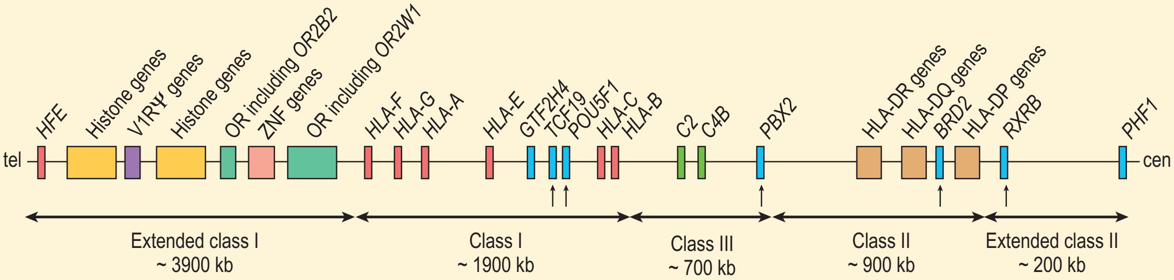

another. Most germline genes are by definition nonpolymorphic (i.e. identical) within a species. The MHC gene cluster has now been extended to include more than 500 genes (Box 7-11) and has been fully characterized (see eFig. 7-6).

Additional content available at https://expertconsult.inkling.com/.

As indicated above, MHC class I genes were discovered during studies in inbred mice of the genetic control of transplantation rejection phenomena. There are now many HLA class I genes, some with novel functions: for instance, HLA-G genes are expressed in the placenta and may have a role in protecting the fetus from attack by maternal NK cells. MHC class II genes were discovered later in analysis of mixed leucocyte reactions (MLR) in mice, a test that is the basis of tissue typing. The polymorphisms (genetic differences) in MHC gene products that characterize each individual are the antigens responsible for inducing graft rejection, and in humans were originally discovered in pregnancy sera and blood transfusion samples by careful documentation of antibody responses for each allotype (allotype = different type of individual). Currently, direct gene sequencing techniques are used to identify specific alleles (allele = specific genetic difference belonging to one individual) and have led to subtype identification even within alleles.

Association of allotypes with disease is now performed in genome-wide association studies (GWAS) (see Ch. 3, pp. 143–145), which includes the MHC genes. The long-standing association, for instance, of certain HLA-B27 haplotypes with ankylosing spondylitis and uveitis has now been extended to include some of the genes involved in trimming of the processed peptides for binding to the HLA-B27 peptide groove: these are the *ERAP1* and *ERAP2* (endoplasmic reticulum aminopeptidase) genes, which show haplotype susceptibility at least for the spondylitis component of the disease. Different MHC protein alleles differ in their ability to bind and present different antigenic determinants, and this is probably the basis of each individual's susceptibility to disease. Studies of the association of disease susceptibility and specific HLA alleles have revealed many interesting linkages (see eFig. 7-6).

Additional content available at https://expertconsult.inkling.com/.

Most disease associations have, however, been detected in MHC class I genes. In particular, The gene map of the extended major histocompatibility complex (xMHC) is shown from telomere (left) to centromere (right) on the short arm of chromosome 6 I (eFig. 7-6). The five colour-coded sub-regions making up the xMHC span about 7.6 Mb and are defined as: the extended class I sub-region (green block; *HIST1H2AA* to *MOG*; 3.9 Mb), the classical class I sub-region (yellow block; *C6orf40* to *MICB*; 1.9 Mb), the classical class III sub-region (orange block; *PPIP9* to *NOTCH4*; 0.7 Mb), the classical class II sub-region (blue block; *C6orf10* to *HCG24*; 0.9 Mb) and the extended class II sub-region (pink block; *COL11A2* to *RPL12P1*; 0.2 Mb). Regions that flank the extended class I and II sub-regions are shown as grey blocks. Insets denote the hypervariable RCCX and DRB regions. Numbers and positions of tRNA genes are represented by indigo bars, the length of which is proportional to the gene number between other loci. Vertical lines connect the two main groupings of tRNA genes of 1.6 Mb and 0.5 Mb of the sequence (separated by 0.6 Mb). Circles to the left of each locus indicate disease status, polymorphism, immune status and paralogy as described in the text. The gene map of the xMHC is also available as a poster and online (http://www.nature.com/nrg/journal/v5/n12/ poster/MHCmap).

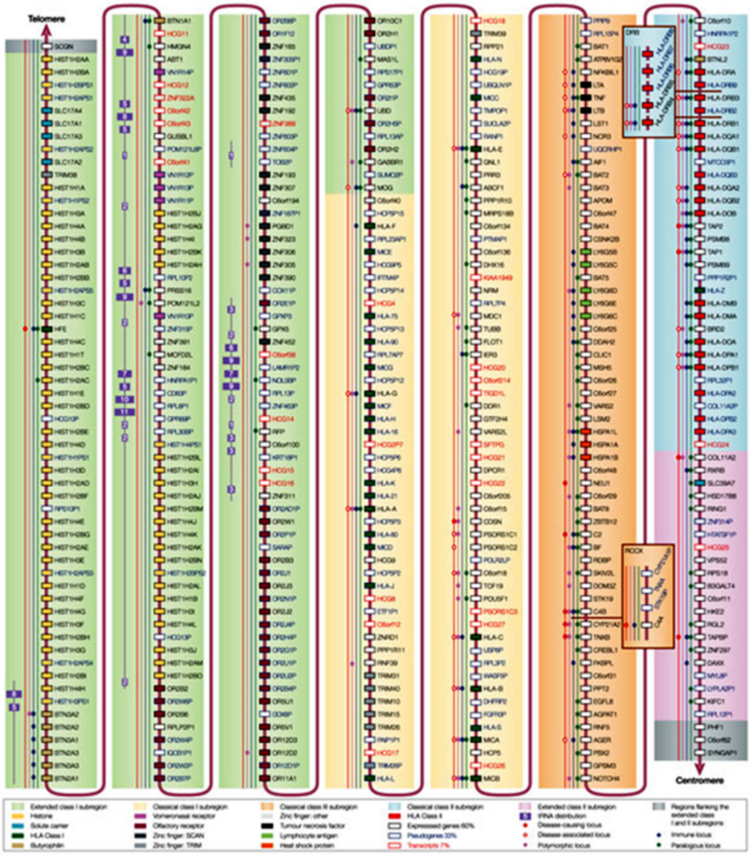

Nature Reviews | Genetics

eFIGURE 7-6 A map of the extended MHC (xMHC) gene region in man. The different regions are colour coded as in the key. Some specific regions are expanded in the inset boxes. (From Horton et al., 2004.)

BOX 7-12 TRACKING MAP FOR MITOCHONDRIAL DNA

A haplotype is a set of allelic differences (i.e. genes) inherited from a single parent, and has been used in anthropological studies to track the migration of humans over hundreds of thousands of years from the first migration of *Homo sapiens* out of Africa (see below a tracking map for mitochondrial DNA – further information is available through the link given).

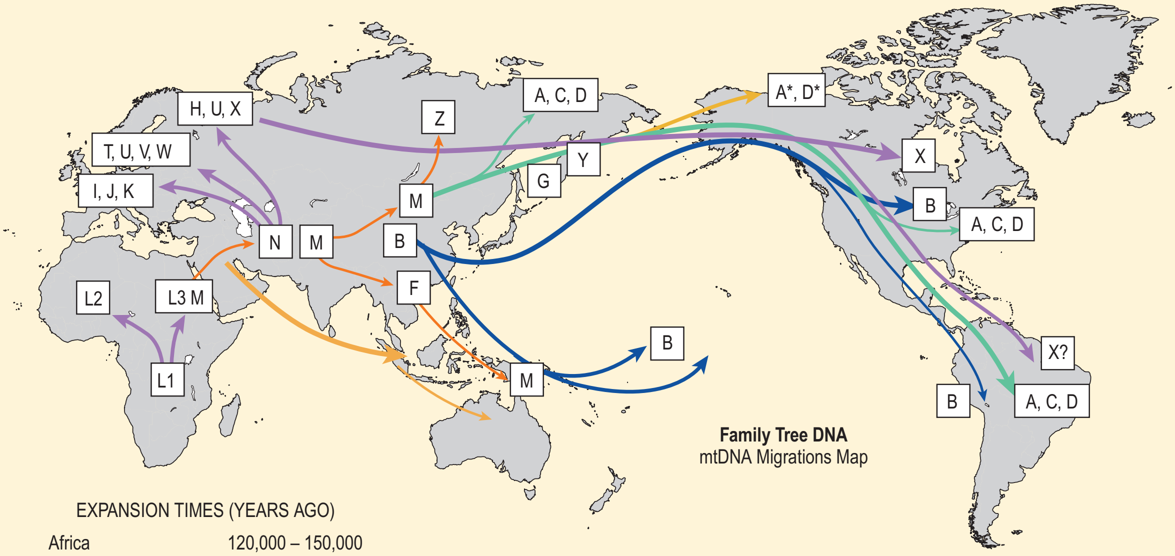

EXPANSION TIMES (YEARS AGO)

Africa	120,000 – 150,000
Out of Africa	55,000 – 75,000
Asia	40,000 – 70,000
Australia/ PNG	40,000 – 60,000
Europe	35,000 – 50,000
Americas	15,000 – 35,000
Na-Dene/ Esk/ Aleuts	8,000 – 10,000

http://mathildasanthropologyblog.wordpress.com/2008/06/16/mitochondrial-dna-haplotypes-for-dummies/

the HLA-B27/ankylosing spondylitis/enteric infection/uveitis link is long established. In addition the HLA-B51, but not B52, is linked with Behçet's disease in the Middle East and Japan. Perhaps the strongest known association of any human disease (>90%) is between HLA-A29 and a rare form of endogenous posterior uveitis, birdshot retinochoroidopathy. MHC mapping has greatly informed disease susceptibility among different population groups; and indeed haplotype mapping has allowed the migration of humans during the thousands of years that *Homo sapiens* has populated the earth to be tracked in a similar way (Box 7-12).

ORGANIZATION OF THE MHC GENES IN THE GENOME

The MHC genes are located on chromosome 6 (see Box 7-11), while the non-polymorphic component of the class I molecule β_2-microglobulin is on chromosome 15. The MHC region is very large, occupying about 4000 kilobases – as large as the entire genome of certain bacteria.

The class II genes, placed near the centromere, are commonly involved in crossing-over during meiosis. Unlike non-polymorphic genes, both alleles are expressed. Furthermore, there are two or three functioning β chain genes in class II, each of which can combine with the α chain. This allows some class II alleles to be expressed in more than one allelic form on the same cell. Thus, a heterozygous individual expresses six different class I alleles (two HLA-A, HLA-B and HLA-C from each parent) as separate molecules on each cell. In contrast, for class II, many more than six heterodimers can be inherited (commonly there are 10–20 different class II genes per cell) and this permits a large range of peptides to be bound on each cell. Furthermore, each MHC molecule can bind many different peptides with differing affinities (see above).

As for other immune genes that involve recombination events such as the immunoglobulin and TCR genes, there is coordinate expression of the HLA genes, e.g. class I on chromosome 6 and β_2-microglobulin on chromosome 13 must be simultaneously activated, and equally so the class II genes during transcription of HLA-DP, -DQ and -DR.

REGULATION AND TRANSCRIPTION OF THE MHC GENES

Regulatory elements that lie 5′ upstream of the MHC class I β_2-microglobulin and MHC class II genes, plus other essential genes such as the invariant chain, tapasin genes, *ERAP1* and *ERAP2* genes and several others, coordinate the expression of MHC genes. These elements include the S, X, Y box and the complex is known as the class II transactivator (CIITA) gene, which regulates MHC class I and II. Mutations in these genes can cause immunodeficiency syndromes.

Cytokines modulate the rate of constitutive transcription and expression of MHC genes. Class I expression is increased by α, β and γ interferons, and induction varies with cell type. IFN-γ also alters transcription of class II: in macrophages, endothelial cells and some parenchymal cells the change is upwards, but in B cells it is decreased. In contrast, IL-4 has the opposite effect on class II expression of B cells. Some cells, such as neuronal cells, do not respond at all to IFN-γ. The interactive responses of immune and tissue cells profoundly affect the final nature and character of the immune response and can convert a tissue-destructive process to a protective response or, depending on the tissue, to an allergic response (see below).

The genetic control of MHC expression and function is also controlled by other genes such as the *TAP1* and *TAP2* genes, which regulate the entry of peptides into the endoplasmic reticulum, the invariant chain gene that regulates stability of the class II complex, the *ERAP* genes that perform final trimming of the peptide in the endoplasmic reticulum and the calnexin gene that regulates the binding of the peptide to MHC class I.

T-cell activation

Any molecule, large or small, can act as an antigen but only macromolecules can activate lymphocytes and act as immunogenes. Small molecules may activate lymphocytes if they are bound to a larger molecule; in this situation the small molecule is called a hapten.

Therefore, although peptides may initiate an immune response, there are several constraints on the induction of a response. For instance, there is a minimum size of peptide for an effective response, which in the case of MHC class I responses is precisely nine amino acids, and between 12 and 30 for class II responses. More importantly, the precise interaction between the T-cell receptor (TCR) and the MHC–peptide complex is under the control of genes regulating both sides of this interaction. The TCR genes have a greater role in determining the specificity of the response than the MHC genes.

THE TCR AND ANTIGEN BINDING

Processing of antigens to peptides in the APC may potentially yield a large number of immunogenic peptides. However, only a small number of these peptides will activate lymphocytes and only one will be specific for a particular cell. This is determined partly by the nature of the antigen but more importantly by the 'TCR repertoire', i.e. the potential range of different TCRs that exist to deal with very large numbers of possible antigens. The immune system has developed a mechanism to deal with this problem, which involves the use of multiple germline genes, each of which undergoes somatic rearrangement on challenge with antigen. The same basic mechanism is used by B cells in their production of antigen-specific antibodies (see below). This is known as genetic recombination and is under the control of enzymes (recombinases), products of specific genes known as recombination activation genes (*RAG-1* and *RAG-2*) which are expressed only in lymphocytes.

TCRs are members of the immunoglobulin gene superfamily (Box 7-13) and are dimeric proteins composed of an α and a β chain which possess V (variable), D (diverse) and C (constant) regions as for

BOX 7-13 IMMUNOGLOBULIN SUPERFAMILY OF PROTEINS

The immunoglobulin superfamily is characterized by the common structural motif, the immunoglobulin domain. This domain may occur singly in small molecules such as Thy-1 or multiply in both chains of dimers such as the T-cell receptor.

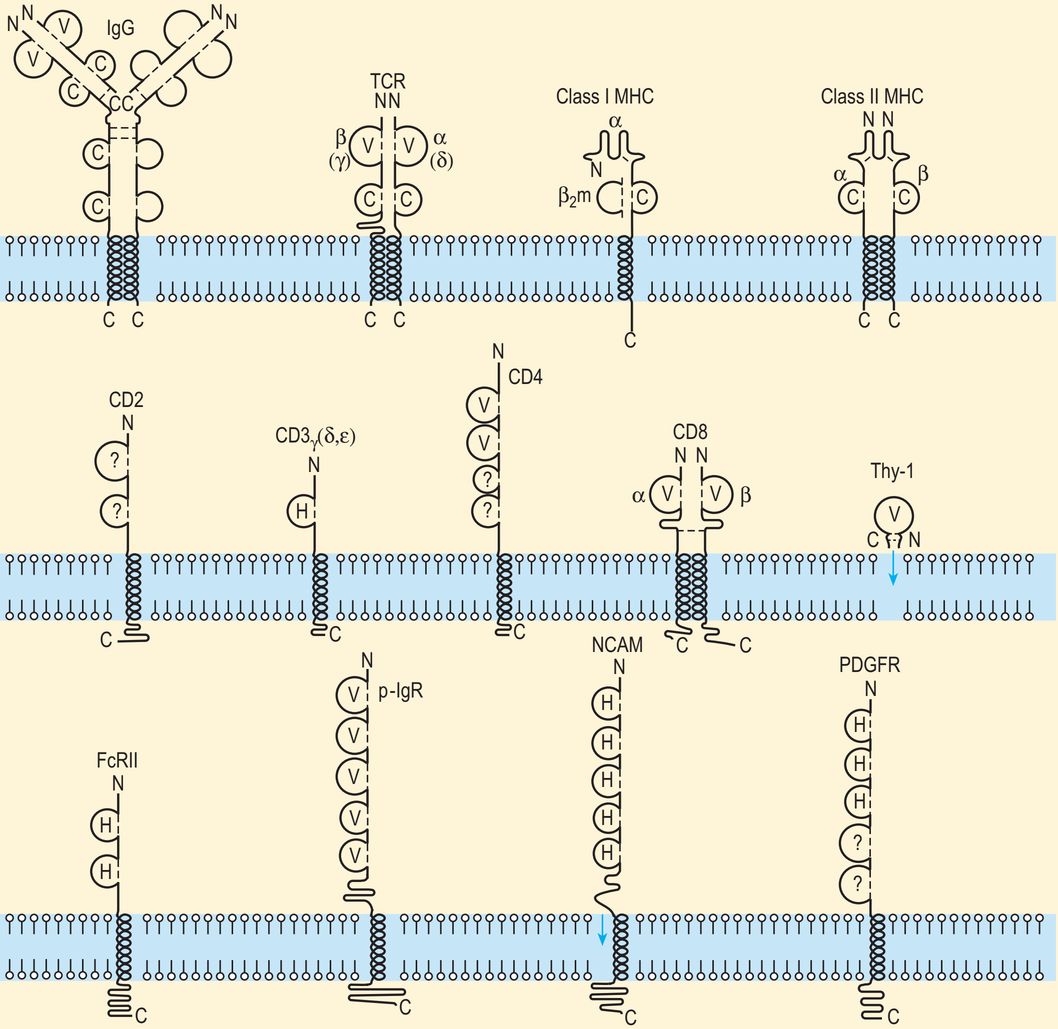
IgG
N
N
V
V
C
C
CC
C
C
C C
TCR
NN
β
(γ)
V
V
α
(δ)
C
C
C C
Class I MHC
α
N
β_2m
C
C
Class II MHC
N N
α
β
C
C
C C
CD2
N
?
?
C
CD3$_\gamma$(δ,ε)
N
H
C
CD4
N
V
V
?
?
C
CD8
N N
α
V
V
β
C
C
Thy-1
V
C
N
FcRII
N
H
H
C
p-IgR
N
V
V
V
V
V
C
NCAM
N
H
H
H
H
H
C
PDGFR
N
H
H
H
?
?
C

immunoglobulin molecules (Figs 7-26 and 7-27). The V region contains the antigen-binding site and interacts with the peptide–MHC complex. At least 20 families of Vα chains are recognized and a similar number of Vβ families are also known. Several V genes (2–10) exist within each family. A single antigen binds to a single TCR and stimulates antigen-specific clonal expansion. However, as indicated above, several antigens have the potential to bind to the same TCR, each with different affinities, unlike antigen–antibody interactions.

Databases holding germline gene sequences as a reference resource have been set up, such as the International Immunogenetics information system (IMGT) linked through the *Ensembl* database.

Additional content available at https://expertconsult.inkling.com/.

Information on specific TCR and immunoglobulin gene segment data can be obtained through the http://www.ensembl.org/index.html database which gives full instructions on how to use the protein 'gene build' programs. Gene building for TCRs and immunoglobulins is somewhat specialized due to the recombination events which take place between specific gene segments. This can be investigated using the link to the International Immunogenetics information system (IMGT) (http://www.imgt.org/) which can identify the locus on the genome showing the gene segment frequency for the specific Ig sequence of interest, e.g. the human IGH (immunoglobulin heavy chain) locus with specificity for herpes simplex virus type 1 (as of July 2013 there were 4 IGHV group, 4 IGHJ group, and 4 IGHD group TCR gene segments on the IGH locus).

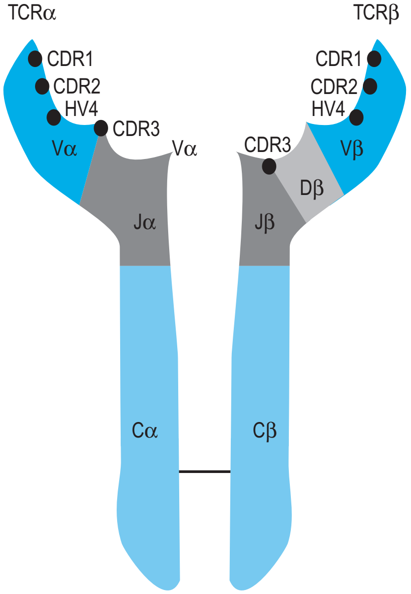

FIGURE 7-26 Correspondence of TCR hypervariable regions to TCR gene segments. In this schematic example, the areas of the TCRα and TCRβ proteins derived from the indicated gene segments are shown in different colours. The CDR1, CDR2 and HV4 hypervariable regions are clustered in the variable domains of each chain, while the CDR3 region encompasses the VJ joint in the TCRα chain and the DJ joint in the TCRβ chain. (From Mak and Saunders 2006, with permission from Elsevier.)

At present there are many allelic variations in all four gene segments, C, D, J and V, for each of the α, β, γ and δ TCR chains, totalling over 200 genes. There are many thousands of possible combinations of the V, J and C genes, which increases the diversity of TCRαβ-binding capacity enormously. A single V gene encodes a sequence in the TCR, a 'complementarity determining region' (CDR) (Fig. 7-26), which binds to that region on the antigenic protein corresponding to the peptide–MHC complex that binds to the TCR. This region of the antigenic protein is known as an epitope. Each protein may contain several epitopes that interact with different TCRs and produce different responses *in vivo* (similar epitopes exist for CDRs on antibodies; see below). Some epitope may even induce anergy in a specific clone of T cells. TCR gene loci to a wide range of antigens from many species can be identified using the database programs.

Immunization with a multideterminant antigen (i.e. an antigen that has several epitopes, which accounts for most protein antigens) will therefore lead to a polyclonal T-cell response in which several T-cell clones are variably activated by each determinant depending on its immunogenicity. Usually one or two epitopes on a molecule are immunodominant.

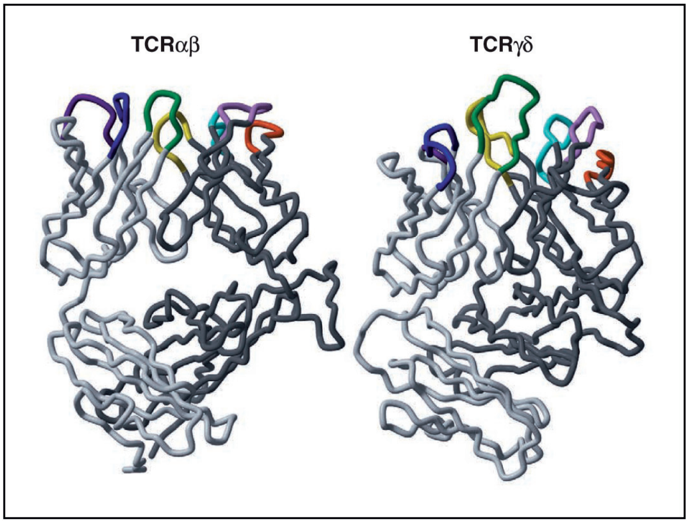

FIGURE 7-27 Structures of TCRαβ and TCRγδ. Crystal structures showing the carbon backbone of TCRαβ and TCRγδ proteins; TCRα and δ chains are light grey, TCRβ and γ chains are dark grey. For the α and δ chains, CDR1 is dark blue, CD2 is magenta and CDR3 is green. For the β and γ chains CDR1 is turquoise, CDR2 is pink, CDR3 is yellow and HV4 is orange. (From Mak and Saunders 2006, with permission from Elsevier.)

Epitope mapping of proteins is therefore possible. It has been shown that only about 30% of most proteins contain sequences that are recognized by lymphocytes. For certain autoantigens, such as retinal S antigen (implicated in the pathogenesis of uveitis, see below), various parts of the molecule have been mapped as regions which bind antibody, regions which are immunogenic (stimulates T cells *in vitro*), and regions which are pathogenic *in vivo* (induces autoimmune uveoretinitis in experimental animals) (peptide band on amino acids from 280 to 364). These sites are different from the rhodopsin-binding sites on S antigen, also known as arrestin (Fig. 7-28), and may represent cryptic epitopes, i.e. epitopes that are revealed only when the molecule is partly degraded.

Conversely, preferential usage of certain clones of T cells with particular Vα and/or Vβ sequences is recognized in certain diseases, such as rheumatoid arthritis and multiple sclerosis, and has also been described in non-infectious uveitis in humans, where the (?auto)antigen is suspected but not known. Varicella zoster and *Toxoplasma*-specific Vα and Vβ TCRs have been identified in samples from patients with infectious uveitis. Thus, investigation of TCR sequences on these cells may provide clues to the nature of possible autoantigens.

Regulation of TCR gene expression is under the control of several ‘gene enhancers’ such as activating transcription factor (ATF), cAMP-responsive element binding protein (CREB), T-cell factor (Tcf-1) and lymphoid enhancer factor (Lef-1) as well as specific transcription factors for each T-cell subset. For instance, GATA-3 regulates Th2 cells, while FoxP3 regulates Treg cells. Transcription factors specific for each T cell subset are shown in Figure 7-21.

The right time and the right place for T-cell activation

One of the long-standing conundrums in T-cell activation has been how to explain the logistics of the interaction between a T cell and an APC. As discussed

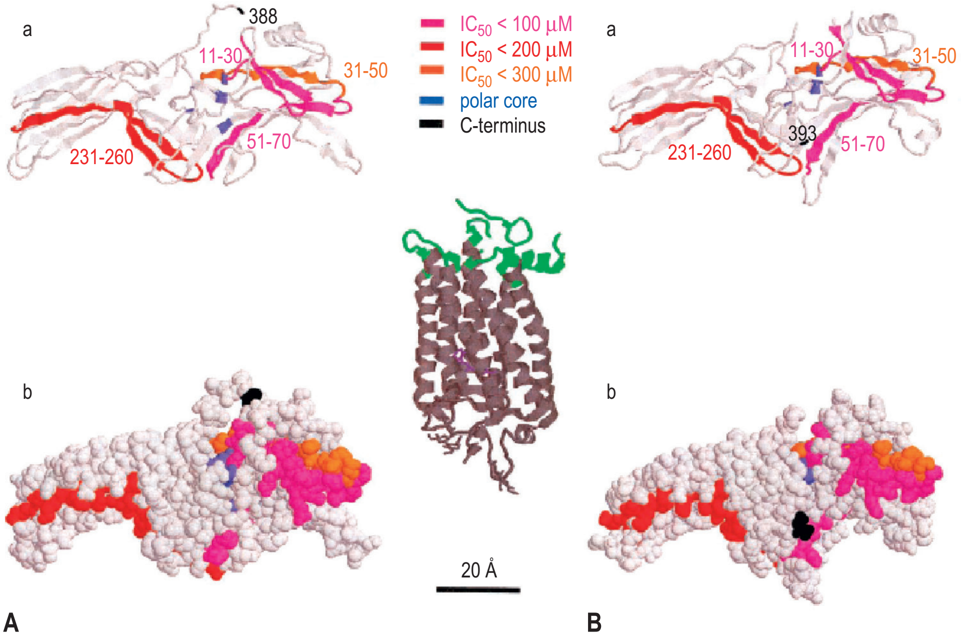

FIGURE 7-28 Rhodopsin–arrestin interaction: binding sites of arrestin (also known as S antigen) from peptide competition. In each case, molecule B of the unit cell is shown in (a) ribbon drawing and (b) space-filling model. The picture in the centre represents the rhodopsin structure in the ground state with the cytoplasmic loops in green and the retinal in purple. (From Pulvermüller et al. 2000, with permission from the American Society for Biochemistry and Molecular Biology.)

Binding of CD21 to C3d-tagged antigens allows the coreceptor to cluster with the antigen receptor coligation of the co-receptor allows receptor-associated kinases to phosphorylate CD19 phosphorylated CD19 binds Src-family tyrosine kinases (e.g. Lyn) and PI3-kinase PI3-kinase initiates a signalling pathway involving the GEF protein Vav above, antigen traffics either as soluble antigen in the lymph or cell-bound in dendritic cells to the secondary lymphoid tissues (SLOs), usually the peripheral draining lymph node. *In vivo* two-photon studies have shown that T cells entering the lymph node make repeated short interactions with several APCs, and when a T cell decides that one particular antigenic peptide is the specific antigen for that receptor (based on its affinity for binding, i.e. how strongly it is bound), it engages in sustained contact (several hours to 2 days) until, when sufficiently activated (as determined by expression of activation markers such as CD69), it detaches from the APC. If this process takes place with the support of additional molecules which provide co-stimulation (co-stimulatory molecules), the T cell will undergo intense activation, proliferate (clonal expansion) and secrete cytokines. If there is no co-stimulation, then the T cell may proliferate somewhat, but does not become active for effector function but for tolerance (see p. 441). Once the clone of T cells has expanded sufficiently, they will leave (egress) the lymph node essentially *en masse* under the control of a specific receptor ligand interaction (sphingosine-1 receptor) to enter the bloodstream. As more cell-associated antigen arrives from the tissues to activate further clones of T-cells, T-cell egress from the lymph node occurs in waves. As indicated above (see p. 414), unactivated passenger naive T cells and possibly 'tolerized' T cells pass through the lymph node via the efferent lymphatics to the next node in the chain and on to the thoracic duct and the spleen.

Circulating activated effector T cells make the rounds many hundreds of times before entering the site of tissue inflammation since the blood vessels at these sites need to be 'softened up', i.e. start to express appropriate adhesion molecules and chemokines. Once the T cells enter the tissues, they are probably further activated by engaging with antigen, or they may simply continue on their proliferative programme, secrete cytokines, kill infected cells via MHC class I mechanisms (Tc cells, Th killer cells), and recruit inflammatory monocytes/macrophages (M1-like) to clear infected tissues (macrophages), and later alternatively activated macrophages (M2-like) to restore tissue integrity (see eFig. 7-1).

Later, T cells entering the tissues have been programmed as Tregs and probably function *in situ* to promote resolution of the inflammation but it is also likely that Treg activity in the lymph node is equally important by preventing further activation/proliferation of T cells. This question has not been completely answered as yet nor has the question of whether antigen needs to be completely cleared from the tissues before inflammation can subside, and how this plays out probably determines whether the inflammation persists as chronic disease or whether the infection becomes 'latent' (i.e. ignored by the immune system).

What are γδT cells and what is their role?

γδ T cells are T cells which express γ and δ IgG family chains, almost identical to α/β chains of conventional T cells, but they have very limited diversity and recognize a broad range of antigens including self- and non-self-antigens such as heat-shock proteins and lipid antigens; they are, in essence, innate immune cells, becoming involved very early in the inflammatory process. Thus, they populate sites of pathogen entry such as the skin (first line of defence) but also circulate in the blood and respond to 'stress' induced by pathogens or endogenous stimuli (DAMPs). γδ T cells do not require classical MHC molecules to recognize antigen but they can respond to unprocessed peptide and other antigens. γδ T cells expand 2–10% in response to alkylamines derived from microbes and edible plants and may represent a link between innate and acquired immunity.

γδ T cells are also implicated in autoimmunity; they appear to recognize heat-shock proteins (proteins expressed in 'stressed cells' and highly conserved across species), which have been suggested to play a role in autoimmune diseases such as rheumatoid arthritis and uveitis associated with Behçet's disease. γδ T cells have been reported to be essential in the induction of experimental autoimmune diabetes and are early participants in models of autoimmune uveitis. High levels of γδ T cells have also been cultured from the vitreous in a case of sympathetic ophthalmia. Interestingly, as part of the immune response to tumours, γδ T cells have been associated with better survival from choroidal malignant melanoma.

Superantigens

Some diseases induced by organisms are so rapid in their onset and catastrophic in their manifestations that it is difficult to explain their pathogenesis in terms of an adequate innate or acquired immune response. Examples include the toxic shock syndrome, meningococcal meningitis and leptospirosis. These disorders are caused by superantigens derived from bacteria (staphylococci, streptococci, mycobacteria, *Clostridia* and many other organisms release superantigens), viruses (e.g. rabies and ebola viruses) and retroviruses (e.g. mouse mammary tumour virus and human immunodeficiency virus). Superantigens are generated by genetic elements known as 'pathogenicity islands', which are phage-related (a phage = a parasitic virus for bacteria which uses the bacteria's metabolic machinery to replicate) genes (phage-related chromosomal islands, PCRIs) which they use for rapid transduction.

Staphylococcal enterotoxins (SE) are a good example of superantigens and are the commonest cause of food poisoning. There are more than 20 SEs, SEA–SEE. They have to bind to the MHC class II antigen to be presented to the T cell but they can bind many polymorphic MHC class II molecules. Superantigens do not require processing by APCs because they activate the T cell by binding to the side of the TCR β chain. They are, therefore, quite promiscuous in that they can polyclonally activate several species of T cell. In superantigen infection, about 10% of the T cells are activated, whereas in immune responses to regular antigens a very restricted set of T cells are activated. Indeed, few antigen-specific T cells can be detected either in the circulation or even in the lesion. The result is the very rapid production of cytokines from activated T cells as well as T-cell-induced macrophages, causing a 'cytokine storm'.

Polyclonal activation in superantigen infection may involve autoreactive T cells and thereby initiate an autoimmune disease, which may persist even if recovery from the infection occurs. Studies of preferential Vβ TCR gene usage by lymphocytes infiltrating the tissues may shed light on the nature of the autoantigens in conditions such as uveitis and rheumatoid arthritis and help to identify sequence similarities between autoantigens and superantigens.

Antigens can be presented by other MHC and MHC-like molecules

Antigens presented through the canonical MHC class I and II routes are exclusively small peptides. However, it is well known that other molecules such as sugars and lipids can evoke T-cell and antibody responses. MHC class Ib molecules include human leucocyte antigen (HLA) -E, -F, -G and in the mouse molecules such as Qa-1. Although these molecules have some diversity, the range of peptides they can present is restricted, e.g. Qa-1 presents only a small set of microbial peptides, while HLA-G is involved in maternofetal antigen presentation.

The MHC-like molecule CD1, which does not contain an α_2-microglobulin protein, has a highly hydrophobic groove with two deep pockets that neatly binds the fatty acid chains of glycolipid antigens while the sugar moieties bind the TCR (Fig. 7-29). Recent studies suggest that both innate and adaptive immunity may be involved in age-related macular degeneration (AMD) and that modification of self-antigens by products of lipid peroxidation such as carboxyethylpyrrole (CEP) may act as autoantigens in causing disease. Lipidated antigens may be presented via CD1 molecules; for instance, CD1d binds NK T cells via αGal-ceramide, while CD1c binds γδ T cells via an unknown antigen.

CO-STIMULATION: PRESENTATION OF ANTIGEN REQUIRES 'HELP' FROM OTHER MOLECULES

As discussed in the previous sections, MHC : peptide complexes act as ligands for binding to the TCR. The outcome of ligand binding is dependent on several factors: in particular, the cytokine milieu decides which type of T cell is induced (see Figs 7-2, 7-22 and 7-23). The cytokine milieu is itself dependent on whether the antigen is presented by an APC which has been activated by a DAMP (see p. 389). For instance, activation of a TLR or NALP3 with consequent induction of the inflammasome will release pro-inflammatory cytokines such as IL-1β, IL-12, IL-23 and TNF-α, all of which provide a suitable environment for activation of a Th1 or Th17 $CD4^+$ T cell; in contrast, if IL-10 or TGF-β are released, then Tregs are more likely to be induced. In both cases the TCR is triggered but different transcriptional pathways are induced. Activation of Th1 or Th17 cells via the TCR not only requires the appropriate cytokine milieu but also requires several TCRs to be simultaneously engaged in a membrane structure known as the immunological synapse. This comprises the aggregated TCRs as well as several surrounding co-stimulatory molecules. In contrast, less is

A

CD1	Lipid antigen		Source
CD1b	Ganglioside GM1	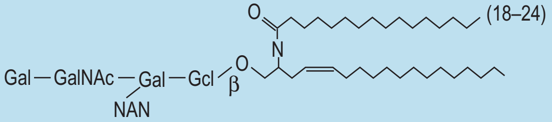 O (18–24) N Gal—GalNAc—Gal—Gcl O β NAN	Self
	Glucose monomycolates	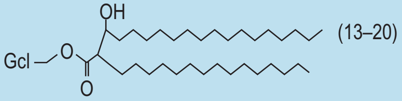 OH (13–20) Gcl O O	Mycobacteria
CD1c	Hexosyl-1-phosphoisoprenoids	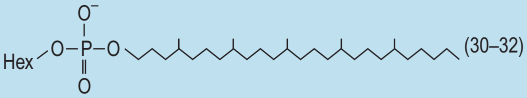 O⁻ Hex O–P–O O (30–32)	Mycobacteria
CD1d	α galactosyl ceramide	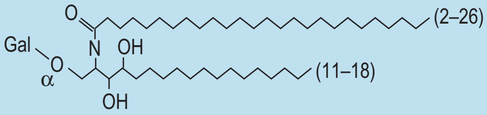 O (2–26) Gal N OH O α (11–18) OH	Marine sponge

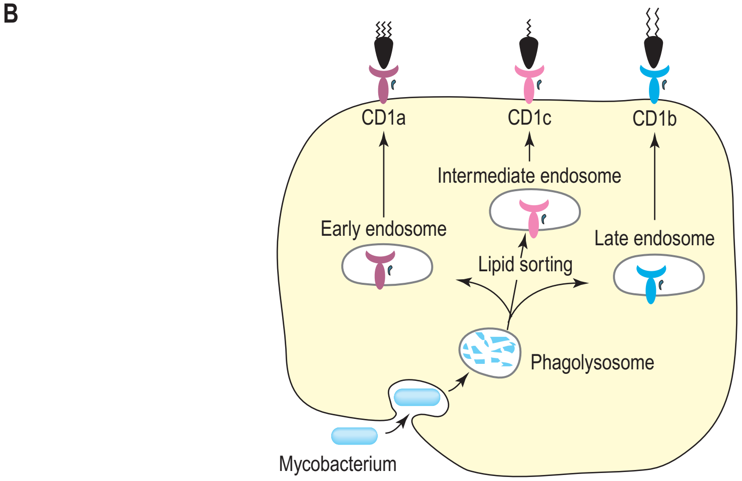
B
CD1a
CD1c
CD1b
Intermediate endosome
Early endosome
Late endosome
Lipid sorting
Phagolysosome
Mycobacterium

FIGURE 7-29 Lipid antigen presentation by CD1 molecules. (**A**) Structures of lipid antigens presented by CD1 molecules. The precise structure of the CD1a antigen is unknown. (**B**) Sorting of mycobacterial lipids. At the bottom left of the figure a mycobacterium has been phagocytosed and degraded. The mycobacterial lipids are sorted by structure into different endosomal compartments where they are loaded on to different CD1 molecules for presentation to T cells. (From Mak and Saunders 2006, with permission from Elsevier.)

known about induction of Tregs, or even engagement of naive T cells to induce anergy (i.e. non-responsiveness), but several 'co-inhibitory' molecules have now been discovered.

Co-stimulatory molecules

Activation of the T cell is achieved by the clustering of several ligand–receptor pairs between it and the APC. Clustering of these ligands at the point of contact between the cells helps to strengthen and prolong the contact between the cells and thus facilitate the presentation of peptide to the TCR (Fig. 7-30). Mobilization of the various sets of molecules to the site of TCR–pMHC interaction occurs within localized membranous patches (lipid rafts, see Ch. 4, pp. 161–162) and the various ligand–receptor pairs participate in the formation of the immunological synapse (Fig. 7-30). Some of the molecules involved either in T-cell

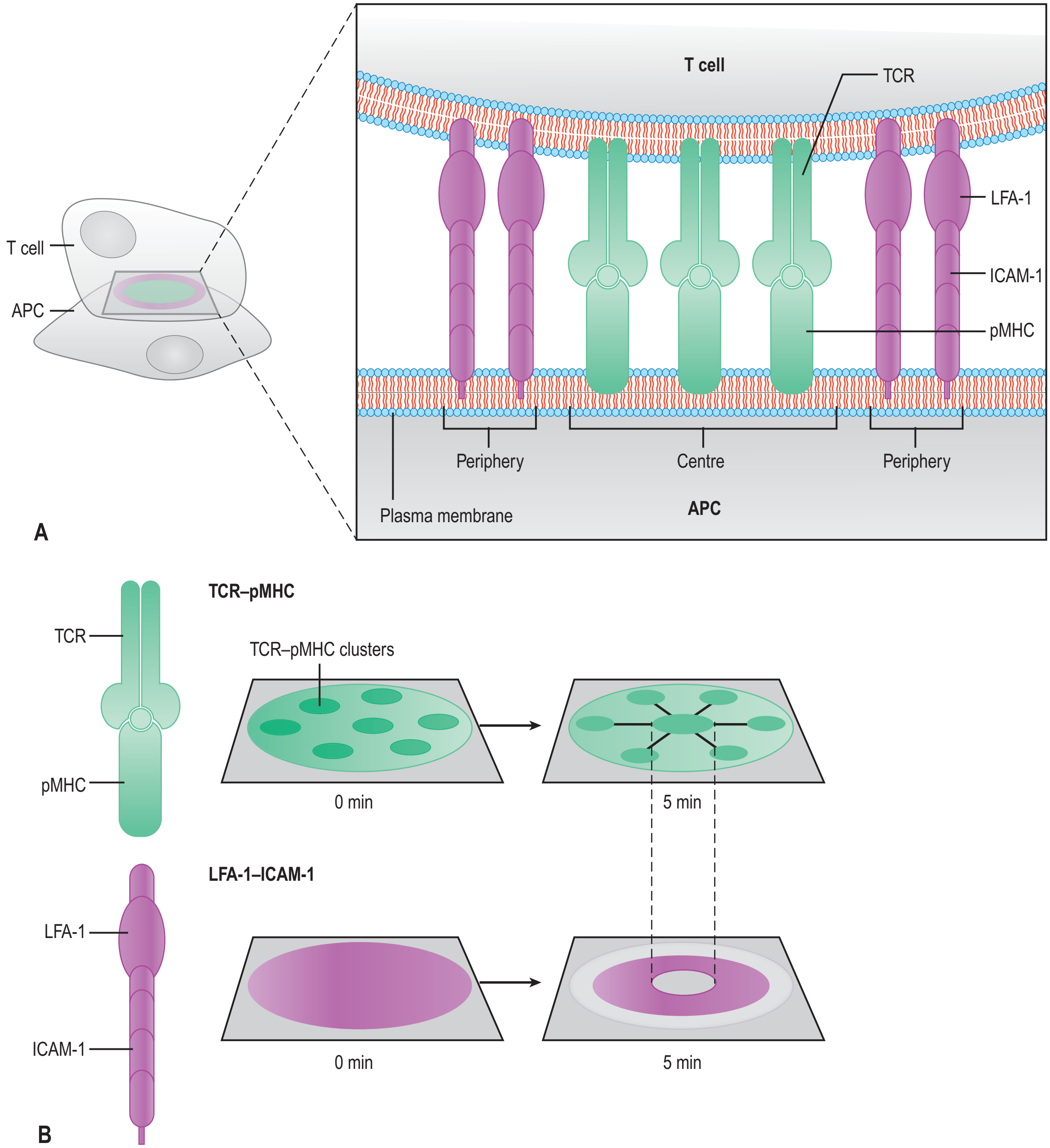
T cell
APC
T cell
TCR
LFA-1
ICAM-1
pMHC
Periphery
Centre
Periphery
Plasma membrane
APC
A
TCR–pMHC
TCR
pMHC
TCR–pMHC clusters
0 min
5 min
LFA-1–ICAM-1
LFA-1
ICAM-1
0 min
5 min
B

FIGURE 7-30 The immunological synapse. This term has been used to describe the binding between the T cell and the antigen-presenting cell because of its structural similarity to the neurological synapse. There is a centre containing the ligand receptor pair (MHC peptide:TCR) surrounded by a ring of adhesion and other accessory molecules which are progressively recruited to the synapse as initial non-T-cell activating binding is converted to a firm prolonged contact necessary to activate the T cell. (**A**) A side and enhanced view of the synapse. (**B**) An *en face* view. Over a period of about 5 min, many TCRs are recruited to the centre of the synapse while further adhesion and accessory molecules firm up the outer ring. The synapse is generated in a membrane lipid raft (see text). (From Manz and Groves, 2010.)

co-stimulation or co-inhibition including adhesion molecules such as ICAM-1/LFA-1 (see Table 7.3), co-stimulatory molecular pairs such as CD40-CD40L, ICOS-ICSOL and CD80-CD28, as well as inhibitory molecular pairs such as PD1-PDL1, and CD28-CTLA4 (reviewed by Chen et.al., 2013).

Two important ligand–receptor pairs during CD4$^+$ T-cell activation (not shown in Fig. 7-3) are the B7 : CD28 interaction and CD40 : CD40 ligand (CD40L). Interaction between these molecules is a prerequisite for T-cell activation and, indeed, in the absence of this 'second signal' presentation of peptide to the TCR is likely to lead to anergy rather than stimulation. B7 is a 45 kDa to 60 kDa cell surface glycoprotein expressed on activated B cells, dendritic cells and macrophages, which binds to CD28, a 44 kDa homodimeric molecule expressed constitutively on the T cell, and leads to the proliferation of, and IL-2 secretion by, T cells. B7 can also bind to a second receptor, cytotoxic T-lymphocyte antigen 4 (CTLA-4), which is slowly induced in T cells and is an inhibitory molecule with greater affinity for B7; as such this ligand pair is involved in downregulation of T-cell activation in the later stages of the response. CTLA-4 is also expressed at high level in Tregs, thus mediating part of their suppressive function.

CD40–CD40L interactions are also very important in T-cell activation, and in reciprocal dendritic cell conditioning there is a two-way signalling interaction via this ligand pair. CD40 is present on APCs and can be upregulated. In addition, it is present on many non-APCs such as endothelial cells. CD40L is only present on activated T cells and even then transiently, so that the time of CD40L expression in part regulates the overall duration of the immune response. CD40L and a further molecule, Lag3 (CD223), may be involved in the concept of APC 'licensing', in relation to activation of cytotoxic T cells (see later, T-cell activation).

Several other accessory molecules are involved in the T-cell/APC interaction leading to the immunological synapse to maximize the T-cell/APC contact. Activation of the T cell leads to the expression of integrin adhesion molecules on the cell surface, including LFA-1 and VLA-4 ($\alpha_4\beta_1$) (see Table 7-3). Reciprocal expression on cells promotes binding. VLA-4 also binds to extracellular matrix molecules, as does CD44 – the hyaluronan receptor. Some molecules provide co-stimulation for specific T-cell activities, such as OX40 and ICOS, while others such as PDL–PDL1 are negative regulators of T-cell activation.

The process of antigen presentation for the generation of an active immune response (immunity) therefore involves multiple molecular interactions and requires co-stimulatory activity for induction of positive adaptive immunity and co-inhibitory molecular interaction for downregulation and anergy induction. It is not clear, however, whether all of the above interactions are required to initiate T-cell clonal expansion or whether immunogenicity is a function not just of peptide/TCR specificity but also depends on which and how many of the accessory interactions are entrained.

Getting the message across: cell signalling through the immunological synapse

Cells activated through cell surface receptors alter cell function through second messenger systems in the cytoplasm. These are likewise linked to events in the nucleus that mediate protein transcription and ultimately to changes in cell function (see cell signalling, Ch. 4, p. 159). The immunological synapse (also known as the supramolecular activation complex, SMAC) provides this function for the T cell. The initial contact between the T cell and the APC mediated through reciprocal LFA-1 and ICAM-1 interactions and also probably by CD45, stabilized by the CD4 and CD8 molecules in the appropriate T cells, induces changes in the actin cytoskeleton that draw several molecules on both sides in lipid rafts into a stable adhesive complex (Fig. 7-31).

The effect is dependent on the signalling molecule VAV. In the centre of the synapse lies the CD3/CD4 (or CD8)/CD45/TCR–pMHC complex surrounded by an inner ring composed of CD2–LFA-3 paired molecules, itself surrounded by an outer ring of ICAM-1–LFA-1 molecules linked to talin and thus to the cytoskeleton (Fig. 7-31). The TCR is linked to its second messenger system through a complex of transmembrane dimeric proteins, the CD3 complex which links intracellularly to cytoplasmic vesicles via an adapter protein, linker for activation of T cells (LAT), and to a signal transduction complex, which belongs to a group of molecules known as immunoreceptor

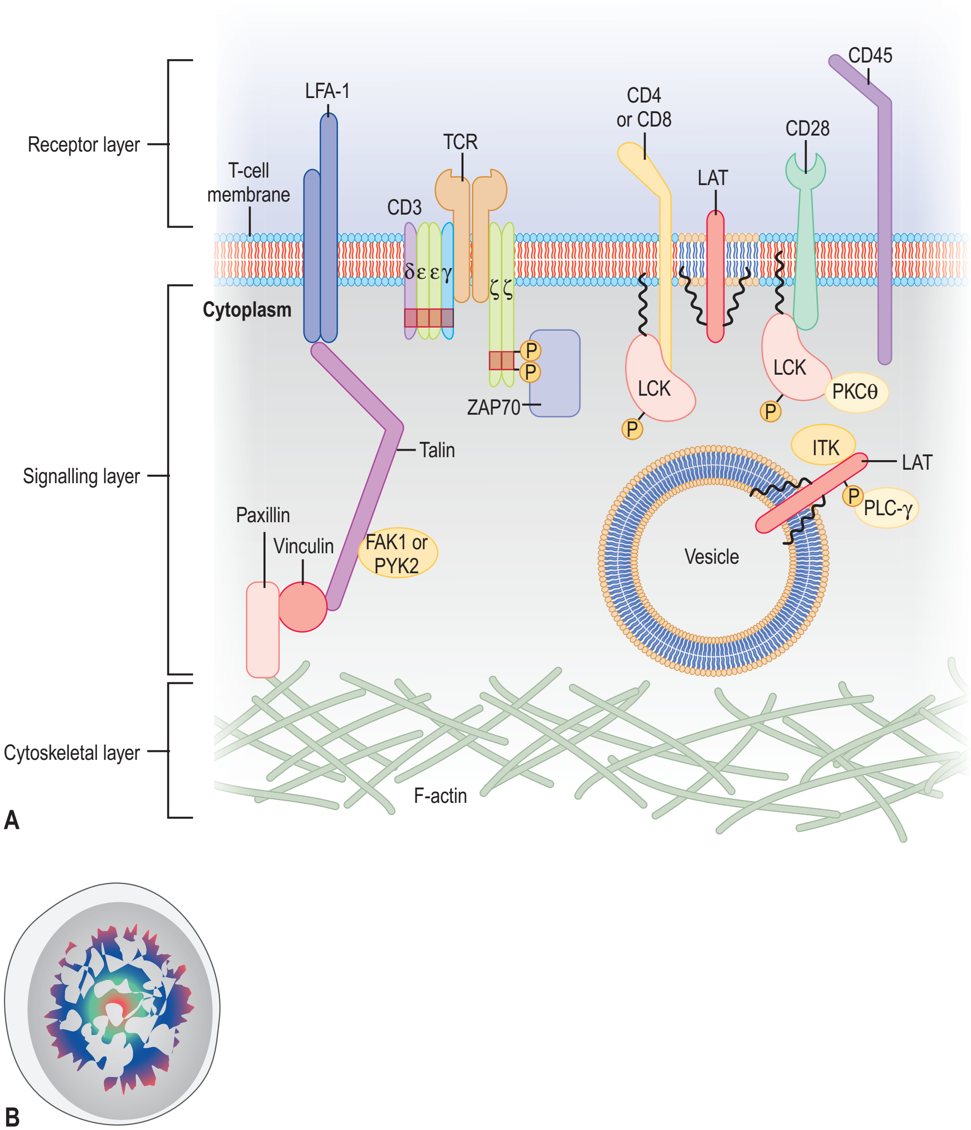
Receptor layer
T-cell membrane
Cytoplasm
Signalling layer
Cytoskeletal layer
A
LFA-1
CD3
TCR
δεεγ
ζζ
P
P
ZAP70
CD4 or CD8
LAT
CD28
CD45
LCK
P
LCK
PKCθ
P
ITK
LAT
P
PLC-γ
Vesicle
Talin
Paxillin
Vinculin
FAK1 or PYK2
F-actin
B

FIGURE 7-31 Molecular details of the immunological synapses on the T-cell receptor side. The TCR linked to the adhesion molecule LFA-1, the CD4 T-cell-specific anchoring molecule CD4, the accessory activation molecules CD28 and CD45, activates the signalling machinery (VAV, ZAP-70, LCK, PKCθ, and LAT) as well as the T-cell cytoskeletal motor (vinculin, F-actin, etc.) which sets the T cell on a cytokine secretion programme, and migration towards chemotactic stimuli. The side view is shown in (**A**) and the *en face* view on the T cell in (**B**). (Dustin and Depoil, 2011.)

tyrosine-based activation motifs (ITAMs). The T-cell ITAM (known as $p59^{fyn}$) acts in conjunction with a second molecule ($p56^{lck}$)(LCK) as the major tyrosine phosphorylating mechanism in T cells.

The four TCR–CD3 complexes may each initiate discrete signalling events, the summation of which produces an effective stimulus for T-cell activation (similar to events in neural cells; see Ch. 4). A progressive accumulation of signal thus develops through the sustained presentation of antigen mediated by recruitment of co-stimulatory and accessory molecules into the membrane surrounding the immunological synapse. In contrast, short-lived contacts between APCs and T cells may fail to initiate responses in the T cells and probably do not even involve the TCR. Downstream signalling from the Src family of tyrosine kinases is another set of protein tyrosine kinases with docking sites that bind SH-2 molecules. One of these is a protein known as ZAP-70 (associated with the ζ protein of CD3) (Fig. 7-31), which plays a major role in sustained TCR signalling. In fact, the interactions between the various signalling molecules through the TCR are a good example of how signalling networks function.

Action of drugs on cell activation

Certain immunosuppressive drugs, used clinically in transplantation and autoimmune diseases, are known to act at various stages in T-cell activation. These include steroids, ciclosporin A, tacrolimus, rapamycin and mycophenolate mofetil. Ciclosporin, tacrolimus and mycophenolate mofeteil are particularly relevant to ophthalmology because they are used in a variety of conditions such as sight-threatening uveitis, severe scleritis and corneal graft rejection.

Ciclosporin and tacrolimus specifically inhibit the transcription of the *IL-2* gene in CD4 T cells by binding to intracellular proteins (immunophilins) that subsequently bind a Ca^{2+}-regulated phosphatase, calcineurin. This enzyme is activated by Ca^{2+} influx during T-cell activation and is required for assembly of the nuclear transcription factor (NF-AT) involved in IL-2 secretion. Ciclosporin may also act via induction of TGF-β release, a cytokine involved both in Treg cell function and in induction of Th17 cells (see Fig. 7-2).

B-cell activation

Until now we have considered only cellular interactions with antigens. However, the first evidence for the existence of an adaptive immune system came through the discovery of antibodies, circulating proteins in the globulin fraction of the serum. Free antibodies are produced by plasma cells, the fully differentiated version of the B cell. B cells are derived from stem cell precursors in the bone marrow and are released into the circulation as $CD45^+$ $sIgM^+$ $CD19^+$ $CD20^+$ MHC class II^+ immature cells which migrate to the B-cell follicle regions of the SLOs. B cells comprise 5–10% of the circulating lymphocyte pool. B cells are important not only as antibody producers but also as APCs in their own right.

ANTIGENS, B CELLS, AND T CELLS: WHICH DOES WHAT?

Antigen recognition by B cells

The rules governing the size of peptide antigen required to induce an immune response apply to T and B cells. However, most T cells are restricted in that they can respond only to peptides. B cells can also respond to carbohydrate and glycolipid antigens, producing T-independent B-cell responses. This occurs for instance with blood group antigens.

B-cell responses to peptide antigens require T-cell help, usually provided by cytokines such as IL-2, IL-4, IL-5 and IL-6 released from activated Th2 cells. B cells acquire antigen from specialized follicular dendritic cells (fDCs) which present antigen as immune complexes bound to Fc and C3 receptors generating an activated germinal centre B cell. This cell undergoes isotype switching (see pp. 396–398), somatic hypermutation and clonal expansion (proliferation) to produce memory B cells and long-lived plasma cells, each of which secretes a specific monoclonal antibody. B cells also present processed antigen on MHC class II molecules to T follicular helper cells which reciprocally assist in the presentation of antigen to B cells by fDCs. This is mediated by co-stimulation via CD40–CD40L interactions. This form of T-cell and B-cell recognition of the same peptide antigen is known as mature B-cell (plasma cell)-linked recognition. Other co-receptor molecules are involved in the formation of the contact site of the B–T-cell interface, e.g. CD30 : CD30L and B-lymphocyte stimulator (Bly5) and its receptor TAC1.

B cells arriving from the bone marrow to the secondary lymphoid tissues traverse the T-cell area on their way to the B-cell follicles. However, if they meet an antigen-specific Tfh cell, they become trapped in the marginal T-cell zone and start proliferating and producing antibody (a primary lymphoid focus). Some of these B cells mature to antibody-producing plasma cells, which leave the T-cell area and return to the bloodstream and ultimately the bone marrow. In contrast, the majority enter the B-cell area where they form a B-cell follicle, and ultimately form the B-cell germinal centre, an antibody-producing mini-factory, where extensive isotype switching takes place.

The B cells then migrate to the medulla of the SLO, and out into the circulation where they home to the bone marrow and differentiate into mature antibody-secreting plasma cells. They may also populate sites of inflammation and develop into plasma cells. Three checkpoints exist therefore for B-cell expansion: the first is at the stage of T-cell help, regulated by CD40L; the second is at the stage of selection in the germinal centre by the follicular dendritic cell delivering anti-apoptosis signals; and the third is at the stage of migration of the mature B cell (plasma cell) to bone marrow and tissues generally. However, most B cells remain in the follicle/germinal centre and secrete antibody directly into the bloodstream via the efferent lymphatics and thoracic duct.

B cells as antigen-presenting cells

Endocytosis of antigen via its sIg receptor activates the B cell to process and present peptide fragments in association with MHC class II on its surface (see above). B cells are efficient APCs but can be activated to present antigen only via the sIg receptor. Therefore, in this regard they act as memory cells.

Although B cells are activated in an antigen-specific manner and their effector molecules (antibodies) are also antigen specific, they induce their effects in an antigen-non-specific way, predominantly via complement. B-cell responses are also greatly enhanced by activation of the B-cell co-receptor complex (CD19 : CD21 : CD81). CD21 is the receptor for complement component C3d, thus permitting antibody–antigen–complement interaction.

B-CELL DIFFERENTIATION

As indicated above, B cells are derived from stem cell precursors in the bone marrow (Fig. 7-32). Immature B cells home to the SLOs. However, if they do not encounter specific antigen, they undergo apoptosis and are removed within 3–4 days. In contrast, when activated, they develop into lymphoblasts and ultimately plasma cells (see above). Some activated cells develop into long-lived memory cells.

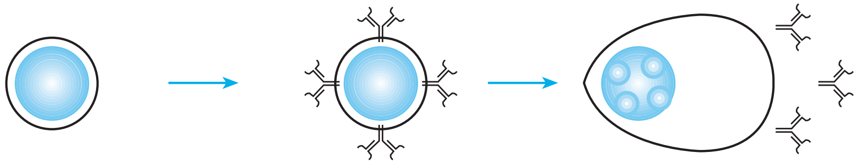

STAGE OF MATURATION	Stem cell →	Pre-B cell →	Immature B cell →	Mature B cell →	Activated B cell →	Antibody-secreting cell
PATTERN OF IMMUNOGLOBULIN PRODUCTION	None	Cytoplasmic μ only	Membrane IgM (κ or λ light chain)	Membrane IgM, IgD	Low rate Ig secretion; heavy chain isotype switching	High rate Ig secretion, reduced membrane Ig
IMMUNOGLOBULIN DNA, RNA	?	Rearranged H chain gene (VDJ-C_μ); μ mRNA	Rearranged μ and κ or λ genes; μ and κ or λ mRNA	Alternate splicing of VDJ RNA (primary transcript) to form C_μ and δC mRNA	Alternate splicing of VDJ RNA to membrane or secreted Ig mRNA; H chain switching by deletion of C region exons or alternate splicing	Predominantly secretory form of H chain mRNA

FIGURE 7-32 B-cell maturation. (Courtesy of A. Abbas.)

Pre-B cells lack membrane-bound IgM and immature B cells may fail to respond to antigen despite the presence of sIgM, owing to lack of accessory molecules. Anergy or tolerance may instead be induced under these conditions. Immature and mature B cells express surface IgM and IgD, whereas activated B cells express IgG after heavy-chain isotype switching. At this stage most of the immunoglobulin is still membrane bound. After some rounds of activation, the cell switches to high secretion of antibody and becomes a plasma cell.

Antibody generation during B-cell ontogeny

The immune system utilizes the same general mechanism to generate an almost infinite range of specific antibodies as it does for TCRs, via recombination of the heavy and light chains of the C (constant) regions, and somatic mutations in the V (variable) regions of the respective antibody and TCR molecules under control of the *RAG* genes (see p. 427). The genes encoding antibody structure are located on chromosomes 14, 2 and 22 in humans (Fig. 7-33).

In memory B cells particularly, somatic mutations that occur during progression of the B cell through isotype switching to the plasma cells permit the antibody to be 'shaped' to fit the antigen better. This is known as 'affinity maturation' of the B cell and its immunoglobulin molecules, a process that describes the stronger affinity of antibodies for an antigen on repeated exposure (Fig. 7-34). Thus, antibodies produced after repeated exposure to the antigen have a much higher binding capacity for the antigen and are more effective in dealing with its removal. This is the basis of various immunization protocols. Isotype switching is under the control of cytokines and varies for different immunoglobulins; thus IL-4 induces switching to IgG1 and IgE while IFN-γ induces IgG3 and IgG2a and TGF-β induces IgG2b and IgA switching.

Under normal bone marrow conditions of continual cell proliferation and maturation, with unique mutations being produced in each cell at each cell division, a very large range of possible permutations of antigen specificity is met within a short time but initially with low affinity. When the correct match between a particular antigen and receptor occurs, antigen-specific clonal expansion ensues. However, this happens only rarely; most cells expressing a particular sIg do not meet the corresponding antigen and are eliminated.

Somatic mutations occur in the CDR of the V genes

Antigen-binding sites are located in the complementarity determining regions (CDRs) of the V segments of the H and L chains (see eFig. 7-3, particularly in the hypervariable region). However, some overlapping binding to the conserved regions of these proteins may also occur. Mutations that increase the affinity of the antigen–antibody reaction occur in the hypervariable regions and become increasingly important with each exposure to the antigen (affinity maturation). Mutations occur in all three CDRs in both chains and may also occur in some intervening regions of the encoded sequence. Antibody diversity is therefore due to several factors (Box 7-14).

Some microbial antigens (superantigens) have the ability to induce a polyclonal response in B cells in the absence of T cells, similar to T-cell superantigens (see p. 431). Superantigens bind to the majority of VH gene families, especially VH3$^+$ IgM. They also bind to conserved regions of the antibody outside the CDR at FR3 (framework region 3). Antibodies are very important to immunity and protection from microorganisms, especially for memory responses to viruses. In such situations the paracrystalline repeated structural epitopes on the virus surface may induce a B-cell response without T-cell help.

Genetic control of antibody production

In addition to the normal TATA boxes of V region promoters, immunoglobulin genes contain regions modulated by trans-acting nuclear factors that regulate the promoters and enhancers. Trans-acting nuclear factors are DNA-binding proteins, some of which are specific to B cells. Others occur in a number of cells that have to respond rapidly and quantitatively, e.g. NF-κB (also involved in IL-2 transcription) and NF-AT (the target for ciclosporin and FK506).

Immunological tolerance and autoimmunity

Many diseases involve the immune system in the absence of clear evidence for a direct causation by a

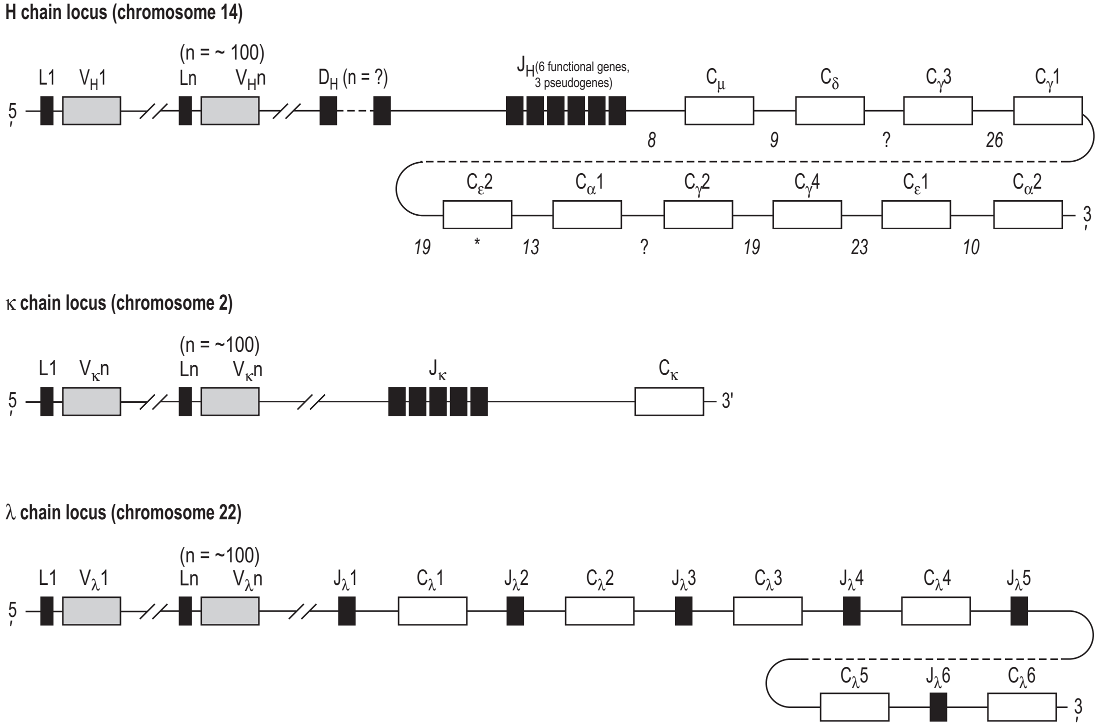

FIGURE 7-33 The heavy (H) and light (L) chain genes are located in the sequence in which they are transcribed. There are over 100 V genes, each of around 300 base pairs in length and grouped in six or seven families on the basis of > 80% homology. Each V gene is preceded by a leader sequence characteristic of secreted and transmembrane proteins. There are an unknown number of D genes and five or six J genes. The J and D genes code for the C-terminal region of the V gene including the third complementarity-determining region (CDR) (see below). Each C gene is composed of several exons represented by the single box. The D_H gene is the main genetic determinant of D–H–J diversity. The amino acid composition of the IgH CD3 region is the driving force for selection. Hydrophilic/aromatic amino acid groups are more likely to be selected than hydrophobic groups.

Recombination of the gene segments to generate a mature immunoglobulin is achieved by excision of the intervening DNA and ligation of the immunoglobulin genes. This ‘looping out’ is facilitated by specific enzymes (recombinases) that act during the earliest stages of B-cell maturation. Segments of DNA, 3′ of each V gene and 5′ of each J gene in the non-coding intervening DNA of the light chain and similarly in the *VDJ* genes of the heavy chain, are excised and re-annealed. Recombinases themselves are under genetic control via recombination activation genes (*RAG* genes).

Recombination occurs in a precise sequence. V_H genes join to the DJ gene, at which point the cell is committed to becoming a B cell. The VDJ sequence then binds to C_μ and a poly-AAA tail is added. The L chain follows the same sequence and the expressed L chain then assembles with the ‘μ only’ H chain in the endoplasmic reticulum. When this pre-B cell expresses the IgM on its cell surface, it becomes an immature B cell.

The μ chain regulates somatic rearrangement by allelic exclusion (activation of genes on one set of chromosomes suppresses activation of the other chromosome) and by initiating L chain rearrangement (the κ chain is activated first; if this is non-productive, λ chain genes are activated). (Courtesy of A Abbas.)

foreign infective organism. In many of these diseases, antibodies and T-cell-mediated immune responses to ‘self-antigens’ can be demonstrated and these disorders are considered to be ‘autoimmune diseases’. The notion that immune responses to self-antigens are unusual implies that immunological non-responsiveness to self-antigens is the normal situation; this state is known as *tolerance*. Therefore, autoimmune diseases are considered to be the result of a breakdown in the normal immunological machinery that permits tolerance to self-antigens, resulting in pathology, i.e. actual tissue damage. The detection of B (antibodies) or T-cell

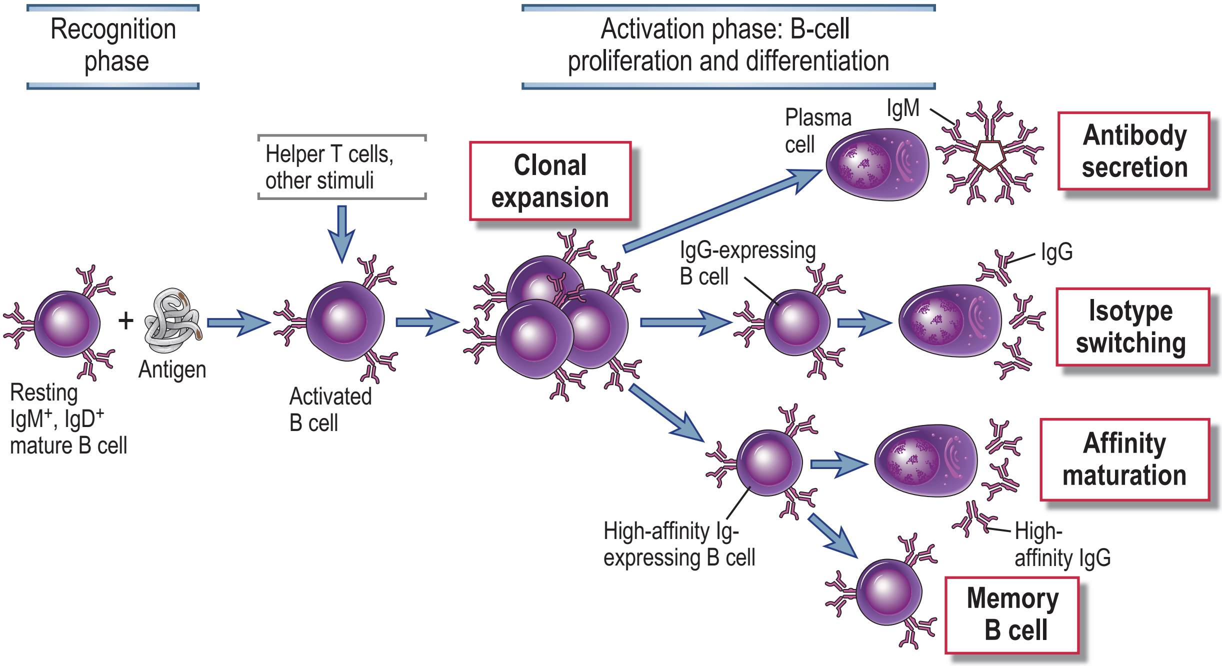
Recognition phase
Activation phase: B-cell proliferation and differentiation
Helper T cells, other stimuli
Clonal expansion
Plasma cell
IgM
Antibody secretion
IgG-expressing B cell
IgG
Isotype switching
Affinity maturation
High-affinity Ig-expressing B cell
High-affinity IgG
Memory B cell
+
Antigen
Resting IgM+, IgD+ mature B cell
Activated B cell

FIGURE 7-34 Abbas, A.K. Figure 11-1 Cellular and Molecular Immunology

BOX 7-14 FACTORS CONTROLLING ANTIBODY DIVERSITY

- Multiple germline genes
- Somatic combinatorial diversity
- Junctional diversity
- H and L chains contributing to the antibody combining site
- Somatic mutations

In the mouse it has been estimated that the potential antibody repertoire due to recombination mechanisms is 10^9–10^{11} (i.e. before antigen has initiated a response).

responses in the absence of pathology is not sufficient for the diagnosis of autoimmune disease since these occur in normal individuals also and are a measure of normal regulated autoimmunity.

Thus, in a broader sense, the immune system has evolved mechanisms that downregulate (switch off) the immune response after it has been activated because it would clearly be autodestructive to have a continuing inflammatory response. The mechanisms that invoke tolerance and those that switch off immune responses may be the same.

WHAT IS TOLERANCE?

Tolerance may be defined as antigen-induced inhibition of the development, growth, or differentiation of antigen-specific lymphocytes, i.e. adaptive immunity. Tolerance has the following properties:

- it is antigen-specific – individuals who are tolerant to one antigen are not necessarily tolerant to all antigens or even a second antigen
- tolerance to autoantigen is acquired during development – immature lymphocytes develop tolerance more easily than adult ones
- maintenance of tolerance requires persistence of (auto)antigen throughout the life of the individual
- tolerance to foreign antigens can be induced if the conditions are right.

SITES OF TOLERANCE INDUCTION

Tolerance classically has been described to occur in two main 'sites': centrally in the thymus (central tolerance) and peripherally in the secondary lymphoid organs (SLOs) such as the spleen and the lymph nodes. According to the thymic deletion concept of

Burnett (see below), most of the autoreactive T cells which would be generated by random re-arrangements of the TCR permitting reactivity to self-antigens by chance, are deleted as part of central tolerance. However this is not a fail-safe mechanism and some autoreactive T cells ‘escape’ to the periphery where they are also deleted, anergized or regulated (see below): this is peripheral tolerance.

Recently, as an extension of Matzinger’s concept that the organism does not truly distinguish between self- and non-self-antigens but between immunogenic and non-immunogenic antigens (the Danger hypothesis, see p. 389) and that self-antigens themselves can become ‘dangerous’ under appropriate conditions, the notion has arisen that the immune response may be modified by the tissues and the ‘self’ antigens (or cells and mediators) they contain (tissue-based tolerance).

Additional content available at https://expertconsult.inkling.com/.

Thus, immune responses to the same antigen or microorganism may differ depending on the tissues: this in fact is not a new concept but an elaboration of the notion of *immune privilege* (see below), which is particularly well expressed in the eye and the brain and is a means whereby ocular and brain tissues modify (downregulate/dampen) immune and inflammatory responses generally.

The same cellular mechanisms are utilized to induce central, peripheral and tissue-based tolerance.

MECHANISMS OF TOLERANCE INDUCTION

Tolerance (better described as immunological non-responsiveness, although this is rather unwieldy) is considered under various forms, e.g. neonatal versus adult, central versus peripheral, innate versus acquired, etc. In fact, tolerance is always acquired; the differences arise when the acquisition of tolerance occurs in the thymus during development (central) or in the peripheral lymphoid tissues during adulthood (peripheral).

Several mechanisms of tolerance have been suggested.

Clonal deletion

This mechanism was originally proposed by Burnett to explain the antigen-induced destruction of self-reactive lymphocytes that occurs in the thymus during development (negative selection) in relation to tissue-specific antigens expressed in the thymus under the control of genes such as *Aire*. This is a form of activation-induced cell death (AICD) that is a characteristic of peripheral tolerance as well as central tolerance.

It has been postulated that removal of all autoreactive cells occurs in the thymus, both for the large pool of thymic antigens and for tissue-specific antigens, which are transported to the thymus on circulating dendritic cells. This has been shown for certain brain (myelin basic protein) and retinal (interphotoreceptor binding protein and retinal S antigen) proteins under the control of *Aire* expressed in thymic medullary epithelial cells (Fig. 7-17).

However, this mechanism does not account for the induction of tolerance to all tissue-specific (non-thymic) antigens, and peripheral tolerance mechanisms are also necessary, some of which also include some level of T-cell deletion.

Anergy

Anergy describes T- and B-cell non-responsiveness to specific antigens but in cells with the capability to respond to non-specific mitogenic stimuli, i.e. the cells are in a sort of ‘suspended animation’. Anergy in lymphocytes is thought to arise via lack of co-stimulation required for antigen presentation to Th1/Th2/Th17 cells (the B7 : CD28 and the CD40 : CD40L interactions are considered particularly important; see pp. 432–434) and has been proposed to account for the non-responsiveness of lymphocytes to peripheral autoantigens. It is therefore considered to be a major mechanism of tolerance induction in the adult.

However, it is likely that clonal deletion and anergy involve the same processes. Thymic education of lymphocytes involves a discrete series of events in arming the T cell to develop towards a $CD4^+$ or a $CD8^+$ phenotype (see Fig. 7-17) and antigen specificity. Each of these steps requires more than one signal and, at any point, lack of a particular signal might induce anergy. If anergy persists, the cell is driven down a pathway towards apoptosis (i.e. deletion). Similar mechanisms exist in the bone marrow and in the germinal centres of lymph nodes and spleen. Apoptotic B cells are removed from germinal centres by the well-recognized ‘tingible body’ macrophages.

Matzinger has suggested a third mechanism for immunological tolerance. In addition to the two canonical mechanisms of central (thymic) and peripheral tolerance, she has suggested that the immune response is modulated by the tissue in which the reaction takes place. Thus, for instance, certain tissues constitutively secrete mediators which influence the nature (e.g. Th1 vs Th2), the strength (a cytokine storm or a chronic low-level inflammation), or even the outcome (complete healing of the tissue with resolution of the inflammation or permanent tissue damage with continued dysfunction). For instance, mediators such as immunosuppressive cytokines, TGF-β and vasoactive intestinal peptide (VIP) are common to both the eye and the gut. The ocular cells also secrete α melanocyte stimulating hormone (αMSH), while the gut produces thymic stromal lymphopoietin (TSLP), which ultimately promotes IgA production. IgA is the immunoglobulin generated by mucosal tissue to provide mucosal immunity. Matzinger (2011) has gone as far as to suggest that the tissues are 'in control' of the immune response (see Further reading).

Thus the same mechanisms for removal of autoreactive cells, involving anergy and cell death, probably occur in the periphery and the central lymphoid tissues, although there is a fine balance between deletion and anergy in the induction of tolerance.

Active suppression by specific lymphocytes: T-regulatory cells

Experimental studies in animals have shown that tolerance, which can be induced to autoantigens and foreign antigens alike using specific techniques (see below), can be adoptively transferred via an infusion of 'tolerized' lymphocytes to naive animals that have not previously been exposed to the antigen. This suggested that tolerance is mediated by cells that inhibit or regulate the immune response, probably at the level of antigen presentation ('suppressor or regulatory cells'). Such regulatory cells have been difficult to identify, mainly because of their lack of responsiveness to specific antigen *in vitro* (anergy). However, seminal work by Sakaguchi in Japan showed that depleting mice of a subset of $CD4^+$ T cells expressing high levels of the IL-2 receptor (CD25) lead to the development of spontaneous autoimmune disease with severe pathology particularly affecting the bowel (inflammatory bowel disease). These cells, known as T regulatory cells (Tregs, see p. 380), are defined by the transcription factor FoxP3 (see Figs 7-2, 7-21 and 7-22), and are required for prevention of autoimmune disease in humans as well as mice. Tregs are decreased in various forms of ocular inflammation such as Vogt–Kayanagi–Harada (VKH) disease. Other Tregs have also been discovered and the important mediators of immune suppression are inhibitory cytokines. Such cells include IL-10-producing Tr1 cells, the TGF-β-producing Th3 cells, $CD8^+$ suppressor T cells which mediate their action through macrophages and NKT cells, B regulatory cells, and γδ T cells. It is not clear how Tregs induce the suppression of T effector cells but the effect is believed to be via dendritic cells and not directly through Treg : T effector interactions (Fig. 7-35).

T-CELL TOLERANCE IS DIFFERENT FROM B-CELL TOLERANCE

The Th cell is the critical control element for adaptive immune responses. Therefore inhibition of Th responses should lead to inhibition of all self-reactivity.

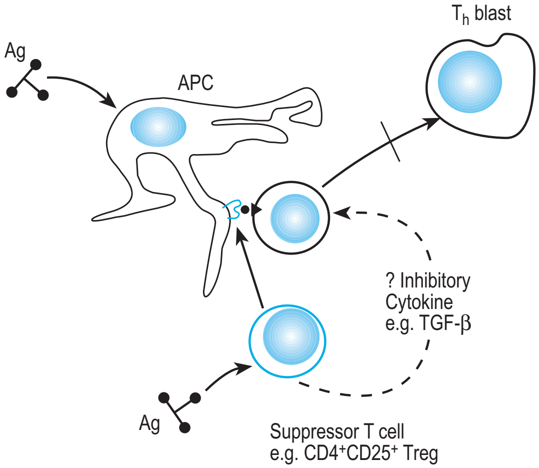

FIGURE 7-35 Mode of action of regulatory cells.

T cells can be rendered unresponsive by very low doses of antigen, while B cells usually require larger doses. Most self-antigen T-cell tolerance is induced in the thymus by clonal anergy/deletion and is therefore antigen-specific and MHC restricted. In addition, natural/thymic (nTregs, tTregs, both terms used reciprocally) are generated in the thymus and released to the periphery. In the periphery, Tregs constitute the main regulator of autoimmunity (see above) and include both natural / thymic (n/t) Tregs and antigen-specific induced (iTregs) derived from naive T cells during the progression of all adaptive immune responses (i.e. to both foreign and auto-antigens). Anergic T cells and Tregs may also be induced by presentation of antigen by aberrantly expressed MHC class II antigen on parenchymal cells in the absence of co-stimulation.

B-cell tolerance occurs particularly to antigens that are 'T-independent'. This includes carbohydrate and glycolipid antigens such as the ABO blood groups, and thus has direct clinical relevance. Tolerance is induced by anergy/deletion of antigen-specific B cells in the marrow by mechanisms similar to those described for T cells in the thymus. Studies using transgenic mice in which genes for neoantigens and the corresponding antibody were inserted into the genome have elegantly demonstrated that B-cell anergy in the marrow occurs by maturation arrest and failure of these autoreactive B cells to enter the peripheral lymphoid organs. After some time these cells underwent apoptosis (i.e. were deleted). Interestingly, some of these B cells could render the autoantigen non-anergy-inducing by switching the κ/λ chain. B cells that had undergone 'receptor editing' of this nature appeared in the periphery (i.e. were not deleted).

FAILED TOLERANCE

A failure to develop or maintain tolerance to autoantigens leads to the development of autoimmune disease. To some degree, tolerance is incomplete and 'natural autoimmunity' to autoantigens is the norm. This has been shown for most autoantigens including several ocular/retinal antigens. However, there are immunological mechanisms in place that inhibit excessive expression of natural autoimmunity.

Autoimmune disease is the dysfunction or damage of tissue caused by immune responses to autoantigens

Autoimmune diseases take many forms, from organ and even cell-specific antibody-mediated diseases such as myasthenia gravis (where the antigens are located at the neuromuscular junction) to widespread systemic diseases such as systemic lupus erythematosus, where the antigen is distributed in all tissues (DNA).

Tissue damage in autoimmune disease can be induced by any of the accepted forms of immunopathological mechanisms (types I–IV; Table 7-7). However, as for most immune mechanisms, autoimmune diseases are usually initiated by CD4$^+$ T cells. Several mechanisms have been proposed to account for this process:

- Molecular mimicry between foreign and autoantigen sequence homologies – as might be expected with the wide range of antigenic peptides occurring in infective organisms – occur with predictable frequency. Processing of foreign antigenic peptide might thus lead to activation of autoreactive T cells if the foreign antigen is 'mistaken' for self. This has been shown for several retinal antigens that have amino acid sequences similar to bacterial and viral antigens, including Gram-negative bacteria such as *Escherichia coli* and parasites such as *Onchocerca*, which causes endemic blindness in certain regions of Africa ('river blindness').
- Bystander activation – the initiation of autoimmune disease may also occur by bystander activation, for instance during infection. Upregulation of co-stimulatory molecules on APCs in the lymphoid tissues in the vicinity of anergized but not deleted autoreactive T and/or B cells may lead to autoimmune disease. Anergy may require expression of CTLA-4, which, if absent, may also allow self-autoreactivity.
- Failure of Treg activity – tolerance is mediated by Tregs (see above) that homeostatically inhibit autoreactive T cells, possibly by direct cell contact or release of cytokine. Tregs constitute a major proportion of circulating CD4$^+$ T cells. Reduced Treg activity is a feature of most autoimmune diseases.
- Failure of deletion (activation induced self-death) – T-cell deletion and apoptosis is mediated by cell surface molecules such as Fas–FasL and TNF-receptor apoptosis inducing ligand (TRAIL) – death receptor (DR4/DR5). Failure of the apoptotic machinery might lead to autoimmunity. TRAIL and Fas have both been shown to play a role in immune privilege.
- Polyclonal B-cell activation – certain compounds such as endotoxin and bacterial glycolipids can activate B cells directly, either to produce cytotoxic antibody or to act as APCs and thus present autoantigen to responding T cells.
- Superantigen – simultaneous activation of several subsets of T cell by superantigens, which do not require processing because they link the T cell and MHC antigen directly, may also lead to activation of autoreactive T cells. Preferential usage of certain TCRs by superantigens may enhance the risk of autoimmune responses.

Parasitic infections persist by giving false signals to the immune system

Diseases caused by parasites are of major medical and economic importance worldwide. Parasites invade the tissues and enter into a symbiotic relationship, sometimes within the cytoplasm of the cell, particularly macrophages. Extracellular parasitic infections are even more frequent, such as those involving nematodes. Examples include filariasis, schistosomiasis and toxocariasis, the first and last of which can infect the eye. Ocular toxocariasis causes blindness in children, presenting as one cause of the white (cat's eye) pupillary reflex.

Parasites present the immune system with enormous problems, partly because they go through a life cycle in which they present different antigens at different times; the host therefore has difficulty in generating an adequate response and may enter a chronic state of inflammation. The parasite achieves this by manipulating, evading or diverting the immune system. Although initially a Th1 response may be induced after infection by the parasite, the response is later diverted to become Th2-mediated with secretion of IL-2, IL-4 and IL-10. Several explanations have been proposed for this change, including defects in the presentation of parasite antigen, inappropriate production of regulatory cytokines such as IL-10 and IL-13, or the induction of a form of tolerance via prolonged binding of the antigen to the TCR. However, rather than being an anergic or impaired type of immune response, the switch from Th1 to Th2 may be quite pronounced. In some respects it is similar to the response in allergic disease because the Th1/Th2 immune response system becomes similarly dysregulated and can lead to excessive mast cell infiltration and eosinophilia, particularly with helminth infestations. Cytokines play a significant role in protection from parasites. In many infections, induction of a strong INF-γ-mediated Th1 response through APC production of IL-12, IL-23 and IL-27 eliminates most of the parasites. This occurs in primary toxoplasmic infections, but some parasites escape detection inside circulating monocytes, which then 'hide' in sites of immune privilege such as the eye, brain and liver.

The parasite thus remains undetected until it kills its host cell, and the newly exposed antigens are released into the tissues to induce an immune response. Toxoplasmic retinochoroiditis is thought to occur by this mechanism. This condition is a relatively common cause of posterior uveitis, characterized by one or more large chorioretinal scars that undergo spontaneous reactivation, producing recurrent attacks of uveitis as the organisms spread.

Allergy and immediate hypersensitivity

The immune system's attempts to rid itself of harmful foreign elements can sometimes be so exaggerated or inappropriate that the host is damaged in the process. Such a situation arises in type I hypersensitivity (see Table 7-7) involving IgE responses to foreign antigens (allergens), thereby causing allergy. However, not all allergic responses are IgE mediated.

ATOPY, ASTHMA AND ALLERGIC EYE DISEASE

Allergic reactions are usually effective, short-lived immune responses which clear the allergen. Allergic disease is where this fails.

Allergic reactions have the following features:

- allergen-specific Th2 cell activation and production of IgE by B cells
- binding of IgE to mast cells with mast cell degranulation
- very wide range of 'allergens' including animal dander, chemicals (haptens), house dust mite proteins, commensals including fungal proteins, plant and other environmental proteins
- symptoms of immediate hypersensitivity due to cytokines.

Allergic disease occurs when the allergic reaction (i.e. the immune response) is exaggerated and the symptoms are the result of this exaggerated reaction. This can be very severe and acute, even life-threatening (as in acute asthmatic attacks). The allergen may have been cleared or may persist and induce the development of chronic disease (e.g. eczema, atopy, chronic asthmatic wheeze).

The prototype allergic disease is asthma, in which reversible obstruction of the airways occurs. This is often the acute manifestation of a chronic inflammatory process in the mucous membranes and skin. Asthma frequently coincides with chronic dermatitis (eczema), rhinitis and sinusitis, and conjunctivitis. When there is a strong genetic predisposition to synthesize IgE specific for certain external antigens, the disease is known as atopy. However, patients with asthma are not necessarily atopic. Other forms of asthma in non-atopic individuals are recognized (intrinsic versus extrinsic asthma). Similarly, all patients with severe allergic conjunctivitis are not always atopic. These concepts are important clinically as seen by the selective efficacy of monoclonal antibody anti-IgE therapy in the treatment of refractory asthma.

Allergic conjunctivitis induced by specific antigens such as pollens or house dust mite may be chronic

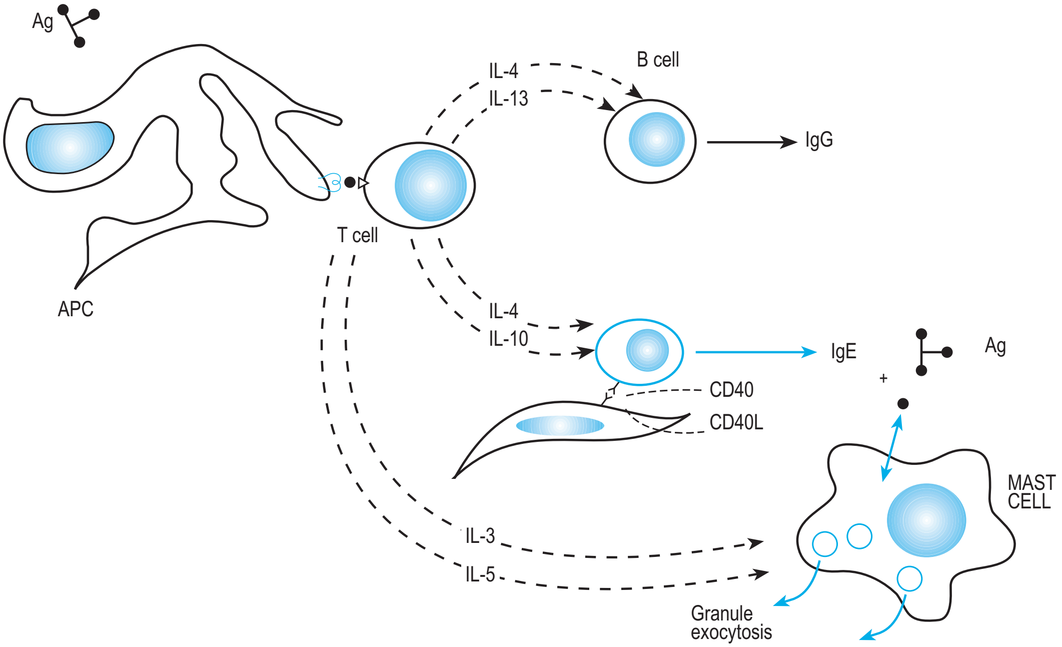

FIGURE 7-36 Mast cell activation is dependent on antigen-specific activation of T cells (Ag) which, in concert with concomitant engagement of the CD40 membrane protein by the CD40 ligand (CD40L), induces an isotype switch in the presence of Th2-type cytokines to direct the B cell to produce IgE. IgE–antigen complexes are then available to activate the mast cell.

(perennial allergic conjunctivitis) or intermittent (seasonal allergic conjunctivitis), and usually produces mild symptoms. Chronic conjunctivitis in patients with atopy is more damaging, often with involvement of the cornea (atopic keratoconjunctivitis; see next section). A distinct clinical entity with massive follicular conjunctivitis and corneal opacification is known as vernal keratoconjunctivitis. In addition, contact lens wear is associated with a significant amount of allergic eye disease, producing giant papillary conjunctivitis.

In spite of the clinical differentiation between various forms of allergy, it is likely that the underlying mechanisms are similar. Activation of Th2 cells is mediated by migratory dendritic cells which capture allergen, activate Th cells in the LN to become Th2 cells which secrete IL-4, IL-5, IL-13 and Tfh cells (T follicular helper cells which activate B cells in the B-cell follicle). Mast cells have long been recognized as effector cells in allergic disease. However, mast cells are dependent on IL-3 and IL-5 production by T cells, which is ultimately induced by antigen presentation to T cells (Fig. 7-36). In this regard, conjunctival dendritic cells have been considered to be major contributors to allergen capture and presentation to T cells locally in the tissues. In severe chronic asthma and ocular surface (conjunctival) disease there is massive thickening of the subepithelial mucosa with fibrosis and infiltration of other pro-inflammatory granulocytes such as eosinophils, which are important in tissue damage.

Mast cells and mast cell degranulation

Mast cells are derived from precursors in the marrow and mature into one of two phenotypes depending on the microenvironment in which they reside. Stem (c-kit) cell factor is an important mast cell growth factor. In humans, mucosal mast cells contain tryptase (M_T), histamine and heparin in their secretory granules, while connective tissue mast cells also contain chymase (M_{TC}). In the normal conjunctiva and choroid, M_{TC} cells predominate. However, in allergic conjunctivitis, M_T cells appear to increase in number in association with the expression of adhesion molecules

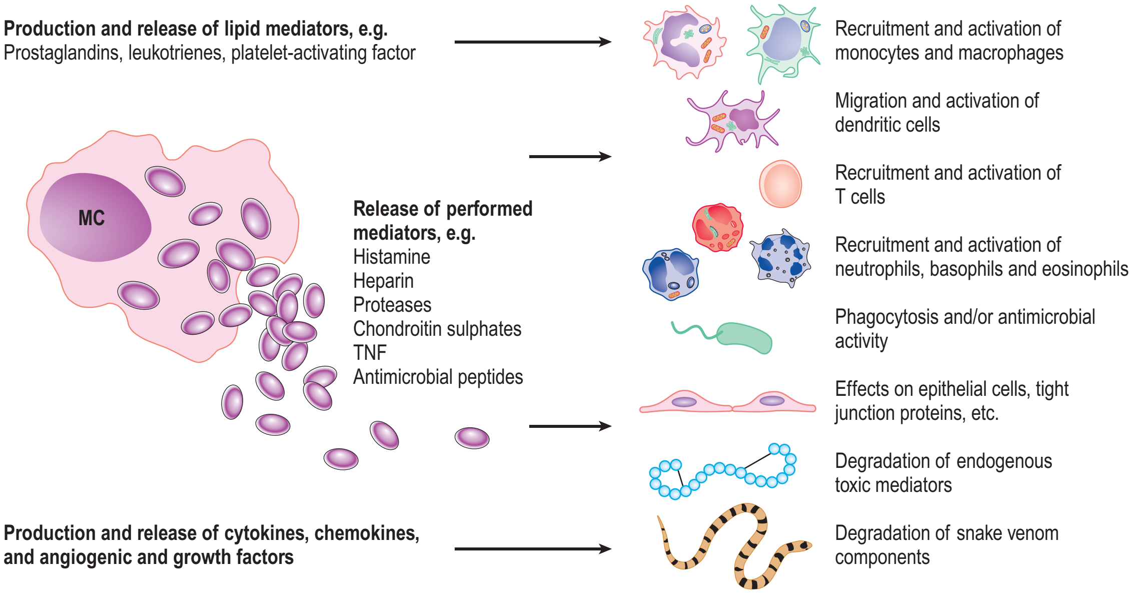
Production and release of lipid mediators, e.g.
Prostaglandins, leukotrienes, platelet-activating factor
MC
Release of performed mediators, e.g.
Histamine
Heparin
Proteases
Chondroitin sulphates
TNF
Antimicrobial peptides
Production and release of cytokines, chemokines, and angiogenic and growth factors
Recruitment and activation of monocytes and macrophages
Migration and activation of dendritic cells
Recruitment and activation of T cells
Recruitment and activation of neutrophils, basophils and eosinophils
Phagocytosis and/or antimicrobial activity
Effects on epithelial cells, tight junction proteins, etc.
Degradation of endogenous toxic mediators
Degradation of snake venom components

FIGURE 7-37 Mast cells have preformed granules ready to be released instantly on activation. They can also produce massive amounts of mediators *de novo* and are thus very potent cells with diverse effects on a wide range of cells as indicated in the figure. (From Metz and Maurer, 2007.)

such as E-selectin and ICAM-1. These changes are likely to be important in the pathogenesis of disease (see below). The significance of mast cell heterogeneity is not yet clear but it would appear that M_T are involved in the active stages of inflammation.

Mast cell degranulation is mediated both by IgE or by non-IgE mechanisms (see below). The role of antigen-specific IgE in this response can be demonstrated *in vivo* experimentally by the passive cutaneous anaphylaxis test. However, IgE is neither sufficient nor essential to induce mast cell-related allergic disease. Also involved are chemokines, mediating not only leucocyte recruitment but also mast cell activation through mediators such as eotaxin 1 (CCL11) and MIP-1α.

Mast cell degranulation leads to the release of preformed mediators including histamine and serotonin, which bind to a variety of receptors and induce second signalling events with release of a great variety of secondary agents depending on the cell type. This includes $NO^{\bullet}$, prostacyclin, smooth muscle relaxants and many others. Several receptors exist for histamine (H_1, H_2, H_3, etc.), which can be distinguished by pharmacological agents (Fig. 7-37). Mast cells also synthesize and release newly formed materials after stimulation. These include prostaglandin D_2 (vasodilator and bronchoconstrictor), leukotrienes (LTC_4, LTD_4, LTE_4, previously known as slow-releasing substance A; cause prolonged bronchoconstriction), and platelet-activating factor (PAF).

Cytokine release by mast cells is the major inflammatory pathway. TNF-α, IL-1, IL-4, IL-5, IL-6, IL-13 and several colony-stimulating factors are released and help to recruit cells to the site of antigen exposure on the mucous membrane.

IMMUNOGLOBULIN AND HELPER T CELLS

There are two types of IgE receptor: FcεRI, which is present in mast cells, basophils and activated eosinophils, and FcεRII (CD23), a low-affinity receptor present on many cell types including B, T and dendritic cells, monocytes and some thymic epithelial cells. FcεRII is important for initial antigen capture by dendritic cells.

IgE-mediated degranulation of mast cells occurs via high-affinity receptors (FcεRI). Signalling via the IgE

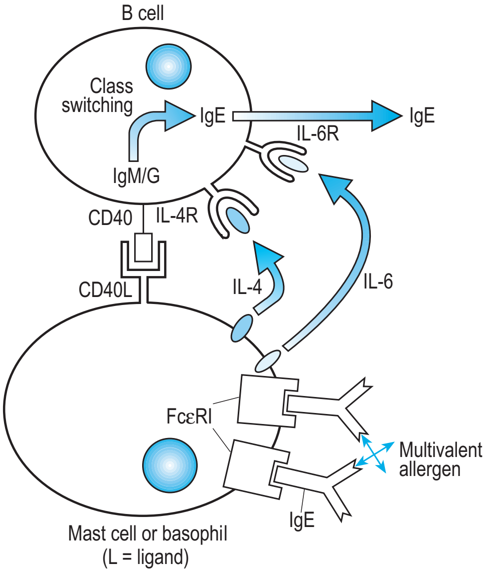

FIGURE 7-38 Mast cells can themselves initiate isotype switching via CD40 because they express the CD40L antigen when they combine with allergen–IgE complexes. In this way the allergic response is amplified and perpetuated.

receptor is inducible only by IgE–antigen complexes and requires local production of IgE by infiltrating B cells.

Isotype switching in B cells from IgG to IgE requires binding of the cell surface antigen CD40 on the B cell to a ligand on the T cell (gp39). This can also be induced by mast cells that express the CD40 ligand and promote local production of IgE in the tissue under cytokine stimulation (Fig. 7-38). Thus the disease can be perpetuated locally and enter a phase of chronicity, as occurs in asthma.

MECHANISM OF DISEASE PRODUCTION IN ALLERGIC DISEASE

Exposure of atopic individuals to allergen is accompanied by an immediate response mediated by mast cell degranulation and a later response occurring after 4–6 hours. The later response is induced by cytokine release from mast cells, which causes recruitment of M_T cells and basophils to the tissue and the expression of adhesion molecules which promote eosinophil accumulation. In addition, release of the chemokine eotaxin by mast cells is probably the main stimulus for eosinophil recruitment. Basophils are circulating granulocytes that are similar in many respects to tissue mast cells in their expression of FcεRI and release of cytokines. However, they are considered to be of a different lineage in the bone marrow (see pp. 376 and 378). T cells are not involved in this stage of the disease.

Eosinophils are strictly regulated because of their potential toxicity. In the absence of infection, very few eosinophils are produced. Even in the presence of an eosinophil, entry to the tissues is tightly controlled by eotaxins (via CCR3) and by other chemokines such as MCP-3, MCP-4 and RANTES.

Release of granule content by eosinophils, particularly erythropoietin (EPO), eosinophil major basic protein (MBP) and eosinophil cationic protein (ECP), causes much of the tissue damage. Eosinophils are attracted to the site of inflammation by specific adhesion molecules, particularly VCAM-1, on the endothelial cells. Induction of VCAM expression on endothelial cells is mediated by cytokine release from other inflammatory cells such as macrophages and T cells, and thus T cells are intricately involved in destructive tissue damage during the progression of allergy from a reaction to a disease. In addition, neutrophils are attracted to the site as a result of upregulated adhesion molecule expression, particularly of E-selectin and ICAM-1 in the acute stage. Neutrophils are themselves responsible for considerable tissue injury in certain forms of allergic disease and characterize a subtype of refractory asthma. The risk of superadded infection in these cases is unclear.

In the later stages of the disease, increased numbers of T cells may be recruited as part of the continuing inflammatory response because of persistent adhesion molecule expression on endothelial cells, especially VCAM. T cells are then available to interact with APCs, particularly activated B cells, causing further release of IL-3, IL-5, IL-13 and GM-CSF, which perpetuates the disease.

MAST CELLS ALSO HAVE A PROTECTIVE ROLE

Mast cells may also have a role in the initiation of the immune response as effectors of innate immunity mediating adaptive immunity, particularly to organisms inducing Th2 responses such as helminths and ticks (Fig. 7-39). Mast cells express MHC class I and

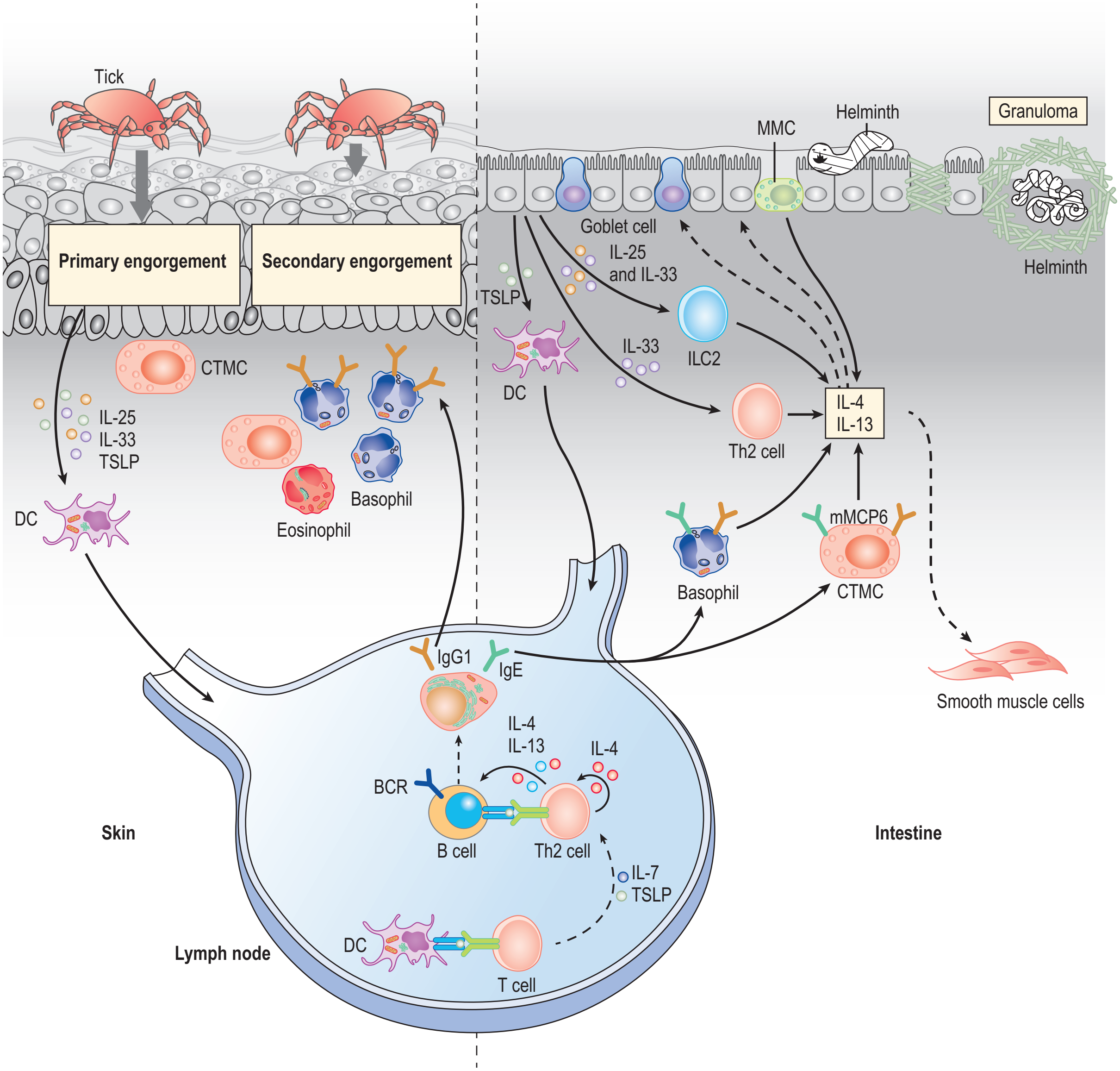
Tick
Primary engorgement
Secondary engorgement
CTMC
IL-25
IL-33
TSLP
DC
Basophil
Eosinophil
Skin
Lymph node
IgG1
IgE
IL-4
IL-13
IL-4
BCR
B cell
Th2 cell
IL-7
TSLP
DC
T cell
MMC
Helminth
Granuloma
Goblet cell
IL-25
and IL-33
TSLP
DC
IL-33
ILC2
Th2 cell
IL-4
IL-13
Helminth
mMCP6
Basophil
CTMC
Smooth muscle cells
Intestine

FIGURE 7-39 Parasites, like any antigen, are taken up by phagocytes such as dendritic cells, which, selectively, through particular Toll-like receptors, activate signals which polarize T helper cells towards Th2 through activation of STAT5. IL-4, IL-5 and IL-13 produced by the DCs themselves or by neighbouring tissue and other cells (e.g. mast cells) help to drive the response down a Th2 route. The process is amplified by continued production of Th2 cytokine, the formation of a parasite-enclosed granuloma and other cytokines such as IL-33, released from damaged and dying cells. (From Voehringer, 2013.)

class II antigens, and presentation of exogenous helminth antigen to T cells leads to activation reciprocally of the mast cell with release of chemokines and cytokines. These direct the overall response towards a Th2-type response.

Organ and tissue transplantation

Transplants or grafts are described as syngeneic when they are between genetically identical individuals, allogeneic when they are between individuals of the same species and xenogeneic when they are between individuals of different species. Autologous grafts refer to the transplantation of tissue within the same individual, as for instance when the cornea from one eye is transplanted to the fellow eye. Allogeneic and xenogeneic grafts are normally not accepted unless the immune system is suppressed (usually by immunosuppressant drugs) and rejection is described as acute, chronic or accelerated.

The discovery of HLA antigens came about as a result of studies that attempted to explain how organ and tissue grafts, especially skin grafts, between members of the same species were rejected. Only in the case of genetically identical individuals were grafts accepted. For instance, in skin grafts, not only was there rejection of the initial graft between two non-identical individuals but also subsequent attempts to graft donor skin between the same two individuals led to an even more rapid rate of rejection, indicating that there was some sort of memory response to the antigens in the original graft.

Although potentially any antigen in the donor graft could act as a target for rejection ('danger' signal similar to an altered autoantigen), the strongest candidate antigens are those where there are the greatest frequency of polymorphisms (differences between individuals or alloantigens) and top of the list are MHC antigens. Cells with the strongest expression of MHC antigens are leucocytes and so rejection of grafts is considered to be the result of the presence of alloantigens on 'passenger leucocytes' in the donor graft. For solid organs and bone marrow grafts, tissue matching on the basis of both blood group and HLA antigens are therefore routinely performed and the closest 'match' attempted in order to ensure graft survival. However, even in these circumstances immunosuppressive drugs are often necessary and usually for the lifetime of the patient.

Since in any one individual leucocytes express a large number of MHC alloantigens (see p. 425), a wide range of allopeptides will be presented to the host, thus leading to a polyclonal T-cell expansion. Damage to the graft is the result of alloantigen-specific $CD8^+$ cytotoxic T cells recognizing MHC antigens on the donor tissue. The number of mismatches on the MHC antigens between the donor and host determines the rapidity and acuteness of the graft rejection.

There are two main mechanisms whereby MHC antigens in the donor graft induce graft rejection. Passenger leucocytes in the donor graft can act as APCs and present antigen directly to host T cells (direct allorecognition) or passenger leucocytes can be killed by host immune cells, the dead cell material phagocytosed by host APCs, and the donor MHC antigens processed and presented on host APCs to host T cells (indirect allorecognition). Grafts are therefore more likely to be accepted if the MHC antigens between donor and host are closely matched and this is clearly the case in identical twins.

Donor non-MHC antigens (minor antigens) can also generate a host immune response and induce graft rejection, but the intensity of the response is considerably less. Mismatching at the minor alloantigen loci is a cause of chronic rejection, while accelerated rejection occurs predominantly between xenogeneic grafts (where differences exist in non-protein antigens such as carbohydrate antigens) or in individuals who have been sensitized to the donor, e.g. via previous blood transfusion. Disparities in MHC genetic loci have classically been determined by the mixed leucocyte response (MLR), an *in vitro* test in which leucocytes derived from one individual (donor) are mixed with those from another individual (host) and an immune response determined by the proliferation of the host leucocytes (the donor leucocytes are irradiated to prevent them from proliferating so that any response is the result of antigen presentation to host T cells). If the number of MHC disparities between the host and the donor is large, there is a correspondingly high proliferative response. However, molecular typing of genetic loci is now more common.

Most donor solid organ, vascularized and bone marrow transplants induce a strong early direct alloimmune response which requires long-term therapeutic immunosuppression to be controlled if the graft is to survive. Graft rejection is mediated by $CD8^+$ T cytotoxic T cells in this circumstance. Simultaneously, as donor leucocytes die, an indirect alloimmune response develops which is slower but more insidious and is mediated by host $CD4^+$ T cells activating innate immune cells such as macrophages and other cells. Many types of chronic rejection, for instance of kidney allografts, are the result of the indirect route of antigen presentation and $CD4^+$ T-cell activity.

Corneal transplantation differs from other forms of skin and solid organ transplantation in the number of donor present in the graft, since there are no leukocyte-rich blood vessels in the central cornea, and even leukocytes that are resident in the corneal stroma are generally not equipped to activate the host T cells directly (i.e. in the quiescent cornea they do not express high levels of MHC antigen).

Additional content available at https://expertconsult.inkling.com/.

Rejection of corneal grafts is thus mainly through indirect allorecognition. However, if rejection does occur, regrafting with a second donor who expresses some MHC antigens that are similar to the first donor will induce a strong $CD4^+$ T-cell memory response and cause rapid rejection of the second graft. For this reason alone there is a strong case for matching corneal grafts, although this practice is not currently routine due to the assumed high success rate of unmatched corneal grafts.

The role of other cells in the immune response to allografts is increasingly recognized. Th cells, both Th1 and Th17 cells, almost certainly have a role via CD40–CD40L interactions and may perhaps involve cross-presentation of antigen. In addition, ‘danger signals’ which activate cells of the innate immune system such as macrophages play a significant role. Surgical technique influences the acceptance or failure of grafts, both vascularized grafts, such as kidneys, and even more so corneal grafts. Donor-specific anti-HLA antibodies have the potential to activate complement, for instance in living-donor kidney allografts. In apparently accepted clear corneal grafts, rejection can be triggered by an unrelated event such as herpes simplex virus infection in the graft or even by a simple procedure such as corneal suture removal. Presumably, activation of innate immune cells such as macrophages and dendritic cells via the inflammasome and antigen presentation with co-stimulation is sufficient to lead to activation of previously primed or memory T cells leading to rejection.

Tregs also play an important role in the overall immune response to a foreign allograft. In the initial inflammation associated with the surgical trauma, both naive T cells and Tregs extravasate into the tissues. In addition, APCs migrating to the draining lymph node activate both Tregs and non-Tregs which home with different kinetics to the graft. The balance between T-effector cells (Th1/Th17) and Tregs determines the survival of the graft.

Finally, rejection of xenografts (the first ever corneal graft performed was a xenograft) is mediated by preformed antibody and complement and generates hyperacute rejection, induced through failure of complement regulatory protein to act across species barriers.

Tumours induce immune responses

Clinicopathological studies have long indicated that tumours induce some form of immune response in the host. This is based mainly on observations of T-cell, macrophage and NK-cell infiltration of tumours independent of whether there is any inflammatory response or tissue necrosis. Experimental studies of chemically induced tumours have also shown that transplantation of resected tumours to the original host, or to a host previously sensitized to tumour antigens, leads to rapid rejection of the tumour, but not when the tumour is transplanted to a syngeneic naive host. Rejection is tumour-specific. Thus, tumours possess specific antigens and these induce an MHC class I-restricted T-cell cytotoxic response.

The concept that cancer cell mutations occur very frequently but are held in check from developing into full-blown tumours is long-standing and indeed NK cells are specially equipped to kill and delete malignantly transformed cells as well as infected cells (see pp. 389 and 390). Antigen-specific cytotoxic T cells (mostly $CD8^+$ but also $CD4^+$ killer cells) may also play this role.

Tumour antigens may be tumour-specific, i.e. only expressed on that tumour, or they may be

Corneal grafting is unusual. Successful corneal allografts between unrelated donors and recipients have been performed since the early part of the 20th century, long before HLA antigens were discovered. This was attributed to immune privilege of the eye, and specifically to the lack of 'passenger' leucocytes in the normal cornea. However, it is now accepted that the apparent success rate of corneal graft applies only to 'low-risk' recipients, namely patients with opaque corneas without significant chronic inflammation or infiltration of the cornea with abnormal blood vessels. Work from Australia has re-examined the question of the privileged status of the cornea in protection against rejection when used as an unmatched allograft and without concomitant immunosuppression. In fact the 5-year survival of unmatched corneas in high-risk recipients, i.e. those patients who have ongoing low-grade inflammation or blood vessel infiltration as in herpes simplex keratitis (see Ch. 9, p. 495), is less than for renal allografts, where steady improvements in survival of allografts has continued (Williams and Koster, 2007). Recent experimental studies of high-risk graft rejection have shown that sensitization, even to a single corneal antigen, in high-risk grafting (particularly in regraft procedures where the first graft failed) can be highly immunogenic and leave a strong immunological memory. To reduce the risk of rejection due to memory T cells sensitized to the antigens present in the first graft, HLA matching of all grafts should be performed so that a second graft has the fewest HLA matches to the first donor and thus maximizes the chances of acceptance of the regraft (Vitova et al., 2013).

tumour-associated, in which case they are also expressed on normal cells. Most tumour antigens are identified in transplantation-type experiments such as that described above in which cloned cytotoxic T cells are generated. This allows identification of the tumour-derived peptides and the genes that regulate their expression in the tumour. Tumour antigens are believed to be recognizable by cytotoxic T cells in this manner, owing to antigenic mutation that converts the normal self-protein, which would induce tolerance, to a non-self-protein, which is recognized as foreign and induces immunity. Thus some tumour antigens are products of normal cells such as tyrosinase in melanoma cells. However, many tumour antigens are produced by oncogenes, genes which are implicated in the cell cycle and in cell differentiation. These include *p21 ras, HER-2/neu*, the very important *p53* gene and others not normally expressed on cells, such as the MAGE series of proteins expressed by melanoma cells including cells of the choroid. Several viral genes may also be expressed in tumours, including those for the SV40 T antigen, the human papillomavirus E6 gene and the EBV antigen. Certain B-cell tumours are indirectly linked to EBV genes, particularly in the presence of immunodeficiency, and appear to involve the translocation of the *myc* oncogene to the immunoglobulin locus. Retroviruses also have the potential to induce malignant transformation of normal cells including the *src*, the *myc* and the *k-ras* gene. The human T-cell lymphotropic virus-I gene is also implicated in certain aggressive T-cell tumours and interestingly is also involved in some forms of intraocular inflammation (uveitis).

Certain tumour antigens are recognized by antibodies as well as by cytotoxic T-cell lines. These include the oncofetal antigens, carcinoembryonic antigen and α-fetoprotein. Other tumour-associated antigens are linked to tumours such as the surface glycoprotein MUC-1 in breast cancer, S-100 in neural crest cell tumours and malignant melanoma, and cytokeratins in epithelial cell tumours.

Immune cells infiltrating tumour are considered to be effector cells against the tumour at least in some instances. Indeed the process of immune surveillance in which T cells and antigen-presenting cells migrate through the tissues detecting altered self-antigens applies particularly to tumours and is necessary for the initiation of the anti-tumour response. Tumour-infiltrating lymphocytes (TIL) include both $CD4^+$ and $CD8^+$ cytotoxic cells and the former probably supply essential cytokines to the latter to promote tumour-killing ability. Some tumours aberrantly express MHC class II antigens plus co-stimulatory molecules; they may directly present tumour antigens to $CD4^+$ T cells and initiate the immune response *in situ*.

NK cells may also be important in killing tumours (as well as removing individual transformed cells as they arise), particularly those that have been induced by viruses. IL-2-activated NK (lymphokine-activated killer, LAK) cells have a markedly enhanced ability to lyse tumour cells and they are in trials as immunotherapeutic agents. Macrophages are also important in tumour killing, usually via release of TNF-α, a cytokine that was first identified by its ability to induce necrosis in tumours. Tumour cells appear to be unable to synthesize superoxide dismutase, which is required to protect cells from the TNF-α-induced release of cytotoxic superoxide free radical.

Despite the variety of mechanisms for tumour killing, many tumours evade death. Tumours utilize a variety of strategies: e.g. they downregulate MHC class I, thus inhibiting cytotoxic T-cell killing; they may not express co-stimulatory molecules necessary for T-cell activation; they may secrete immunosuppressive cytokines such as TGF-β; tolerance to tumour cells may occur if the tumour antigen is expressed during the neonatal period; antigenic modulation of the tumour may occur if the antigen binds non-complement binding antibody; or the tumour antigen may be prevented from gaining access to the immune system by a dense glycocalyx on the cell surface. The immune response to tumours is more likely to be modulated by any or some of these mechanisms, thus explaining why tumour immunity is imperfect and cancer remains a major cause of death.

Tumours may evade the immune system by promoting tolerance rather than immunity. In particular, Tregs comprise a significant component of tumour infiltrating lymphocytes (TILs) and prevent development of an effective antigen-specific anti-tumour response. In addition, the recently identified Gr1+Lys6C-myeloid derived suppressor cell (MDSC) was first isolated from within tumour tissue and is now recognized to have a significant immunosuppressive effect, not only against tumours but also against any immune response. MDSCs are thus being proposed as therapy for autoimmune and other immune-mediated disease. In contrast, other cell-based therapies are in the vanguard of 'cell vaccination' therapies for cancer treatments: these are based on the use of activated, pro-inflammatory dendritic cells loaded with tumour-specific antigens to enhance strong anti-tumour responses and have been introduced for certain forms of prostate cancer. The main difficulty with this approach is identifying peptides specific for the tumour that are sufficiently immunogenic.

The eye and the immune system

The eye participates in all aspects of immune responses like any other tissue, but the immune response is modulated by the cells and tissues of the eye. In this respect, the eye (and the brain) are regarded as 'immunologically privileged', a concept which has now been broadened to be included as a third mechanism of tolerance induction. Both the innate and acquired immune systems function in ocular defence mechanisms.

THE INNATE IMMUNE SYSTEM AND THE EYE

Reference has already been made to the several physical and chemical barriers to ocular infection included in the blink reflex, the lids, and the components in the tears such as lysozyme, lactoferrin and complement (see also Ch. 4, p. 198). Lysozyme is effective against Gram-negative bacteria and certain fungi but is ineffective against Gram-positive organisms such as *Staphylococcus aureus*. Lactoferrin and transferrin, however, are more effective in defence against Gram-positive bacteria because they bind iron, an essential cofactor for eukaryotic as well as prokaryotic cell growth. In addition, tears have specific anti-adhesive properties for bacteria and therefore inhibit bacterial attachment and invasion of the ocular surface, and are incidentally important for prevention of contact lens contamination.

Tears also contain polymorphonuclear leucocytes (PMNs), which increase in number when the lids are closed for prolonged periods, e.g. during sleep. The anti-adhesive properties of tears extend to these cells, ensuring that they pass through the lacrimal passages and do not penetrate the corneal surface. Many of these leucocytes pass directly from the conjunctival vessels through the epithelial layers into the tear film. PMNs contain numerous antibacterial protein enzymes, including proteinase 3, myeloperoxidase (which generates free radicals), calprotectin, β-lysin and the cathepsins.

Tear lipid also has an antibacterial effect. This applies to both short- and long-chain fatty acids, the former affecting surface properties of the bacterial cell membrane and the latter having a direct effect on metabolism. In addition cationic peptides, such as the defensins, as well as surfactant protein-D are part of the conjunctiva-associated lymphoid system (CALT, see p. 416).

The cells of the ocular surface also express, or can be induced to express, many of the innate immunoreceptors such as TLRs and are involved in many of the cornea and conjunctival innate immune responses to infectious agents (see Ch. 8, p. 469). More importantly, the conjunctiva contains numerous dendritic cells (similar to Langerhans' cells in the skin) which act as APCs in the draining lymph nodes in the afferent limb of the immune response. It is thus possible to become sensitized to environmental antigens and allergens via the conjunctiva (Fig. 7-40).

THE ADAPTIVE IMMUNE SYSTEM AND THE FIRST LINE OF DEFENCE IN THE EYE

Tears contain immunoglobulins such as IgA (see Ch. 4) and occasional specific immune cells such as

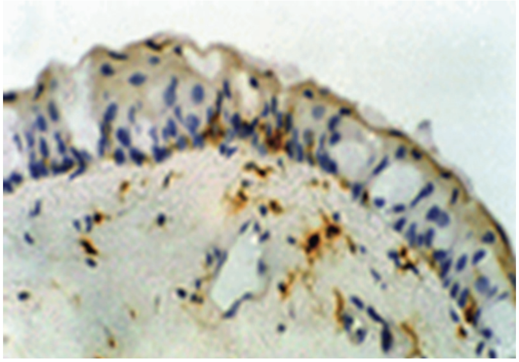

FIGURE 7-40 Conjunctival DCs in the rat. Both intraepithelial and stromal DCs are shown expressing MHC class II (immunoperoxidase, dark brown stain). DCs, dendritic cells; MHC, major histocompatibility complex.

lymphocytes. IgA is produced by B cells in the lacrimal gland and secreted as sIgA in the tears. In addition, the lacrimal gland produces other immunosuppressive cytokines such as TGF-β. The TGF-β-deficient mouse exhibits extensive ocular surface pathology, indicating the importance of this cytokine in tears.

Conjunctival sensitization to environmental antigens is best demonstrated by the allergic (Th2) type response. Common antigens include pollens, house dust mite and animal dander (particularly cat dander). Although the effect is produced locally, the initial sensitization is a systemic one via activation of Th2 cells in the draining lymph nodes (see pp. 414–415). This results in local conjunctival mast cells becoming loaded with antigen-specific IgE, which renders these cells acutely sensitive to re-exposure to antigen. The chemokine CCL2 appears to be partly involved in this process of mast cell degranulation in giant papillary conjunctivitis, while the receptor CCR7 is active in directing conjunctival dendritic cell migration during this process.

If the sensitization is long-lived and recurrently activated with frequent mast cell degranulation episodes, it can lead to chronic inflammation. Thus seasonal allergic conjunctivitis can progress to perennial conjunctivitis and, in atopic individuals, may manifest as severe atopic conjunctivitis and/or vernal keratoconjunctivitis. Much of the damage in vernal keratoconjunctivitis is thought to be induced by eosinophils, because eosinophil cationic protein levels in tears are greatly increased in both atopic and vernal keratoconjunctivitis (Fig. 7-41). In addition, different types of allergic conjunctivitis are associated with different patterns of cytokine production and Th-cell patterns. Th2-like profiles are linked to vernal keratoconjunctivitis, while Th1 is associated with atopic keratoconjunctivitis. There may also be a defect in histaminase function in allergic eye disease causing prolonged histamine effects after mast cell degranulation.

The corneal surface also contains populations of intraepithelial and stromal leucocytes, some of which have the characteristics of dendritic APCs (see Ch. 1, p. 18). The role of these cells in protective immune responses is not clear but there is evidence that, like the conjunctival cells, they can capture antigen and transport it to the draining lymph node. In addition, they are likely to respond to cytokines produced by surrounding epithelial cells in response to TLR engagement by common invading microorganisms such as herpes viruses and bacterial pathogens. This may be a direct effect of the pathogen; for instance, the obligate intracellular parasite *Chlamydia*, which is the cause of the worldwide blinding disease trachoma, signals through TLR2 inside epithelial cells (Fig. 7-42). Indeed the possibility of regulating host–pathogen interactions has been tested in an experimental model of *Pseudomonas* keratitis in which 'silencing' of expression of TLR9 was achieved using small interfering RNA for this receptor.

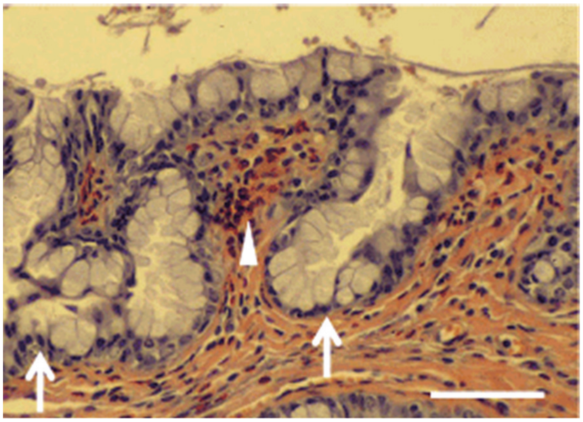

FIGURE 7-41 Histopathology in allergic conjunctivitis in an experimental model. Note the pseudotubular structure formations in conjunctival epithelium (arrows) and infiltration of eosinophilic cells in the subepithelial stroma (arrowhead). Bar = 100 μm. (From Hara et al., 2012.)

PROGRESSIVE OCULAR SURFACE DISEASE

Certain other ocular surface diseases occur that do not appear to be allergic in nature but are considered to be immune-mediated if not autoimmune in their pathogenesis. These include cicatrizing conjunctival disorders, various forms of 'melting' corneal disorders and a number of scleral and orbital inflammations.

Cicatrizing disease of the conjunctiva

Subconjunctival fibrosis occurs in certain rare disorders such as benign mucous membrane pemphigoid and the Stevens–Johnson syndrome. Both are considered to be autoimmune in nature by virtue of the detection of antibodies to basement membrane components such as integrins. Pemphigoid is characterized

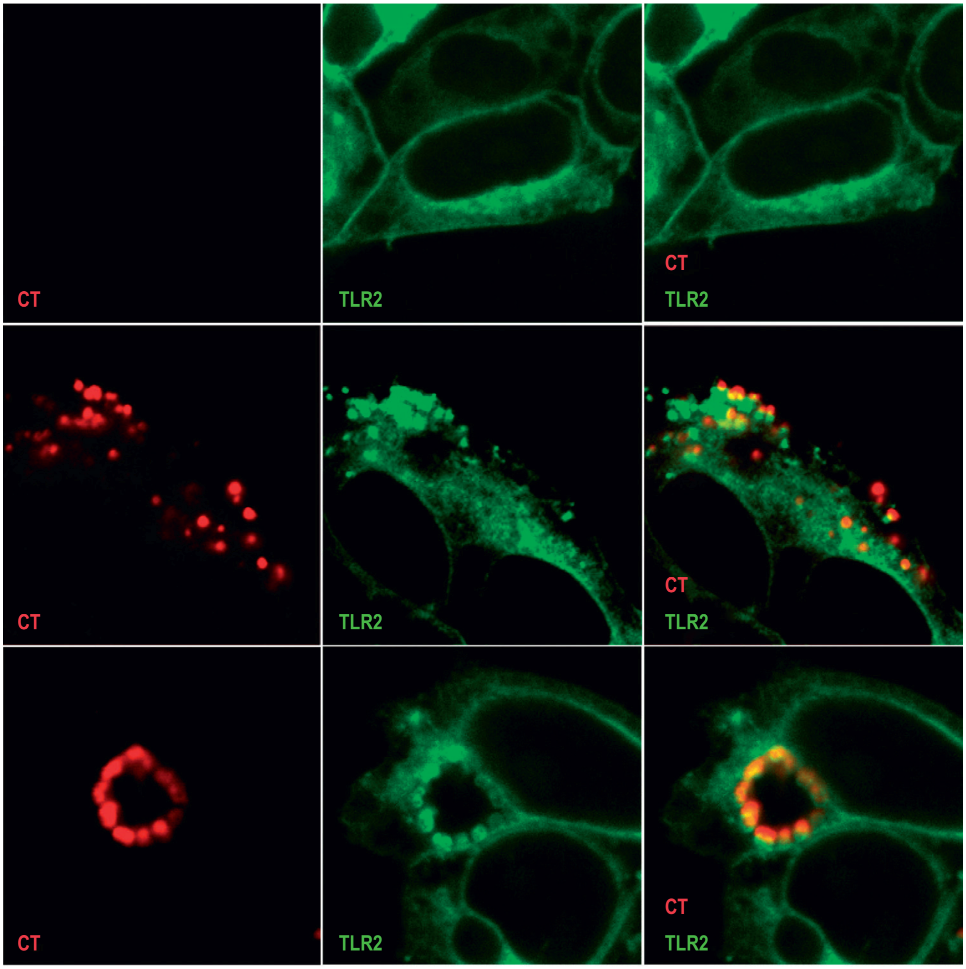

FIGURE 7-42 TLR2 co-localizes with intracellular *Chlamydia trachomatis*. HEK293 cells stably expressing CFP-tagged TLR2 (green) were infected with *C. trachomatis*. Uninfected cells (top row) and infected cells at 16 hours (middle row) and 24 hours (bottom row) post-infection were stained using a monoclonal antibody against *Chlamydia* lipopolysaccharide. (From O'Connell et al., 2006, with permission from the American Society for Biochemistry and Molecular Biology.)

by progressive cicatrization of the conjunctiva stroma leading to severe shallowing of the fornices (see Ch. 9, p. 501). In this condition, specific autoantibodies against conjunctival epithelial β_4 integrin, a component of the hemidesmosome (see Ch. 4, eBox 4-2) have been identified. This is in contrast to bullous pemphigus, a skin disorder that does not affect the conjunctiva but is characterized by widespread areas of epithelial detachment, and in which antibodies against other proteins such as desmoglein and plectin may be detected. Dysregulation of TGF-β in conjunctival cells has been reported in ocular pemphigoid.

The Stevens–Johnson syndrome has similar appearances but is much more acute, although self-limiting. This condition is normally associated with drug administration in which the drug is considered to act as a hapten. The severity of the condition is limited in its effects by the degree and duration of exposure to the drug.

Keratitis and 'melting' corneal ulcers

Many forms of keratitis are considered to be immune-mediated, including postherpetic disciform keratitis (see Ch. 9, p. 501) in which residual herpes simplex virus antigen may play a role. In addition, certain debilitating corneal diseases characterized by peripheral corneal thinning and ulceration, sometimes leading to perforation, may be autoimmune or at least immune-mediated, mainly because they represent 'vasculitic' complications of 'classic' autoimmune diseases such as rheumatoid arthritis and also because of the absence of overt infection. In this condition, reductions in Tregs cells have been reported as well as increases in B cells and anticorneal epithelial antibodies. Some forms of peripheral ulcer such as Mooren's ulcer have been linked to a cornea-associated antigen, calgranulin, also found in peripheral blood neutrophils and filarial nematodes. These disorders occur just distal to the source of corneal epithelial stem cells at the limbus, which suggests a defect at this level.

Inflammatory disorders of the orbit and sclera

While many inflammatory conditions of the sclera (scleritis) are considered to be autoimmune, or at least immune-mediated, it is important to consider infectious aetiologies such as syphilis and other bacterial causes. Spontaneous (autoimmune) inflammatory disorders of the episcleral tissue and the sclera (episcleritis and scleritis) represent a type IV immunopathological disease with close association with rheumatoid arthritis and a similar pathogenesis. Scleritis occurs in rheumatoid arthritis patients who have vasculitis. Rare histological studies have shown activated CD4$^+$ T cells in the lesion in a perivascular location, in addition to macrophages. CD8$^+$ T cells have also been demonstrated, but in fewer numbers. There is a prominent vasculitic component to the disease with extensive necrosis of the scleral layers. Typical granulomatous lesions are a feature of this condition (nodular scleritis). However, not all of these lesions are T-cell-dominated, because B-cell 'follicles' have been identified in some cases, suggesting the development of tertiary lymphoid structures (see p. 417). In addition, increased matrix metalloproteinase activity has been detected in these lesions (e.g. collagenase and stromelysin) (Fig. 7-43). Although the antigen(s) for scleral inflammation has not been identified, it is presumed to be a component of the extracellular matrix such as dermatan sulphate proteoglycan or type 1 collagen. Scleritis most commonly affects the anterior sclera but if it affects the posterior sclera it is more difficult to diagnose. In addition, it may be mistaken for a less well-defined group of orbital inflammatory disorders known as pseudotumour of the orbit, for which the aetiology remains obscure but which responds to systemic steroid therapy. A specific form of pseudotumour known as orbital myositis, in which an acute inflammatory swelling of a single ocular muscle occurs, is particularly responsive to steroid therapy. The autoantigen in this disorder is assumed to be a component of the ocular muscle.

Swelling of the orbital muscles also occurs in dysthyroid eye disease, causing proptosis and exophthalmos, but in this condition all four muscles are variably involved. This disorder is closely linked to Graves' disease of the thyroid in which thyroid autoantigens such as thyroglobulin and the thyroid-stimulating hormone receptor are implicated. Patients with dysthyroid eye disease have circulating lymphocytes that react with ocular muscle cell membrane antigens. However, thyroglobulin does not appear to be the important antigen for ocular muscle damage, and some other antigen such as the thyroid-stimulating hormone receptor (TSHr) may be involved. Models of thyroid ophthalmopathy using T cells sensitized to the thyroid-stimulating hormone receptor have been reported. Immunological studies have shown that the T-cell infiltrate is almost exclusively Th1 in type, with secretion of IL-2 and IFN-γ.

Inflammatory disease of the lacrimal gland may be primary (autoimmune), as in Sjögren syndrome, or secondary, as in sarcoidosis, in which the aetiology is unknown. In both disorders there is a deficiency of tear secretion that produces a secondary keratoconjunctivitis (keratoconjunctivitis sicca or the dry eye syndrome), common in the elderly. Primary Sjögren syndrome involves other secretory glands such as the salivary glands and is characterized by specific autoantibodies against ribonucleoproteins (antiRo and antiLa) whose role in the pathogenesis of the condition is not clear. Biopsies of salivary gland tissue have shown a predominant T-cell infiltrate, but with little evidence of T-cell activation (as evidenced by the lack of IL-2 receptor (CD25) expression). In contrast, conjunctival biopsies from patients with Sjögren syndrome have shown significant T-cell infiltrates with activation

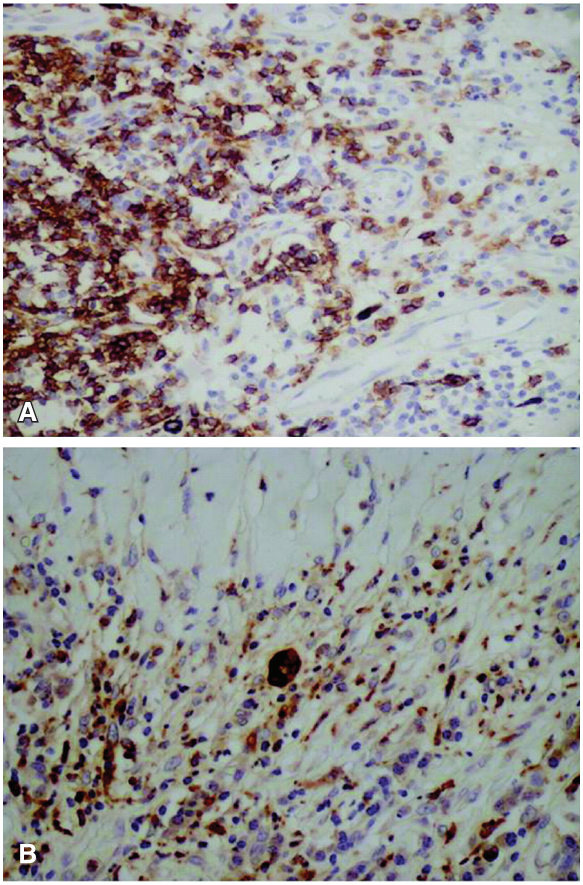

FIGURE 7-43 Immunopathology of scleritis. (**A**) Autoimmune scleritis. Infiltration of CD20-positive B cells in the autoimmune group (original magnification ×100). (**B**) Necrotizing scleritis. Large numbers of CD68-positive macrophages in the idiopathic necrotizing scleritis groups (original magnification ×80). (From Usui et al., 2008.)

markers. In addition, extensive adhesion molecule expression has been detected in the lacrimal gland, not only on endothelial cells but also on the acinar epithelial cells. Both VCAM-1 and E-selectin are upregulated on the endothelium, indicating that this chronic disease is in a state of persistent activation. Eventually these glands undergo involutionary atrophy. Recent extensive microarray studies have shown that mammalian chitinases, innate immune products of macrophages with specificity against chitin-containing pathogens are highly active in severe acinar gland destruction of Sjögren syndrome (Fig. 7-44).

The eye as a privileged site: what does this mean for the survival of corneal grafts?

The eye has been considered immunologically 'different' since the first corneal graft in a human was performed more than 100 years ago, and shown not to be rejected. However, the eye as an immunologically privileged site was not formally recognized until 1945 when Peter Medawar in his seminal studies on alloreactivity in rejection of skin grafts, showed that skins grafts placed in the anterior chamber of the eye or the brain were not rejected: i.e. these tissues demonstrated *tolerance* (see p. 441) to foreign antigens. The basis for this phenomenon was attributed to the avascularity of the cornea and/or to the lack of lymphatic drainage for intraocular structures. However, alloreactive T cells specific for corneal antigens can be detected in the circulation and appear to be generated in the secondary lymphoid tissues (SLO), indicating that despite the apparent lack of ocular lymphatics, foreign antigens in the cornea (and indeed from other ocular structures such as the retina) are clearly transported to the SLO. Most recently, antigen tracking to the submandibular draining lymph node has been detected using green fluorescent-labelled proteins applied to the corneal surface (Fig. 7-45).

Despite this, immunological 'privilege' (IP) in the eye is a real phenomenon and a property of the intraocular compartments, because antigens and cells including some xenogeneic tumour cells appear to be well tolerated in the anterior chamber of the eye. In addition, after acceptance of grafts in the eye, second grafts from the same individual are accepted at sites where they would previously have been rejected, e.g. the skin. This indicates that the tolerance induced by the graft placed in the eye is a systemic phenomenon which was previously considered a 'deviation' from normal immune responses; it was thus termed anterior chamber-associated immune deviation (ACAID). Interestingly, ACAID requires an intact spleen and is mediated by a range of immunosuppressive factors in the eye, particularly TGF-β, as well as by immuosuppressive ligand receptor pairs such as Fas–FasL, PD1–PDL1 and TRAIL–D5 (see p. 433). Importantly, Tregs (both CD4$^+$ and CD8$^+$ Tregs) play a significant role in sustaining IP, both of which are generated in the SLOs, particularly the spleen.

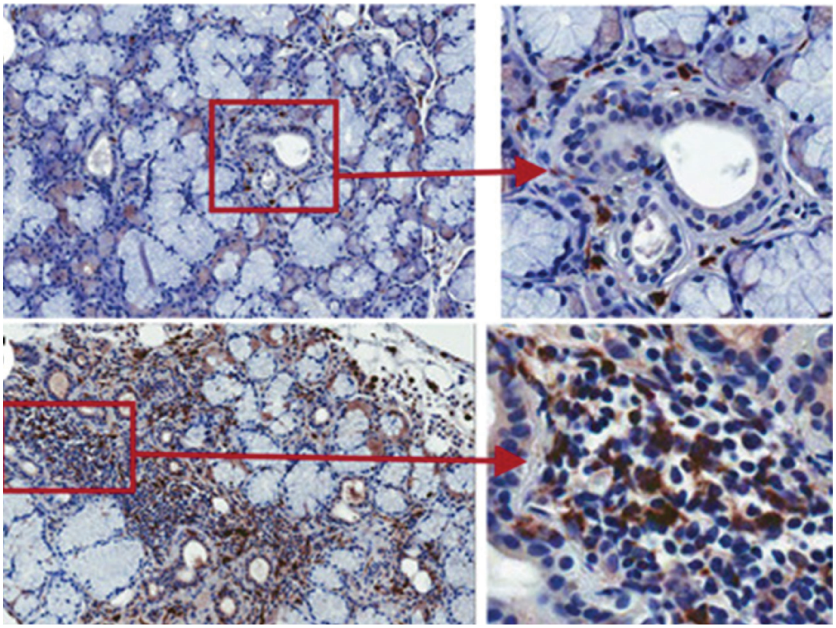

FIGURE 7-44 Upper panels: normal mouse salivary gland has only a few resident macrophages (immunoperoxidase staining for CD68) (original magnification ×20, ×40 right panel). Lower panels: diseased salivary gland is heavily infiltrated with CD68$^+$ macrophages (magnification also ×20, ×40). (From Greenwell-Wild et al., 2011.)

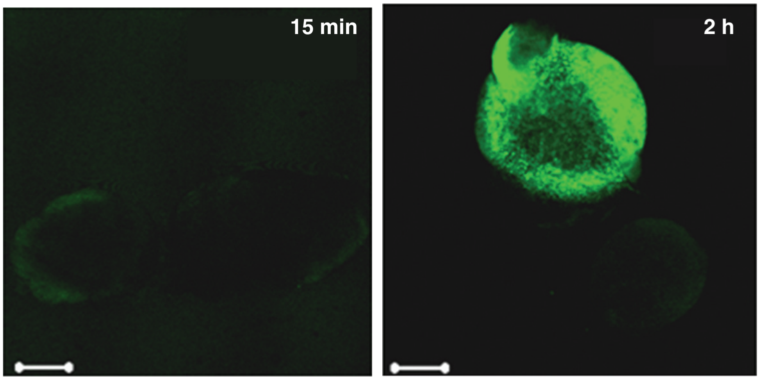

FIGURE 7-45 Immunofluorescence of submandibular draining lymph node from mouse after application of a green fluorescent-labelled protein to the surface of the mouse cornea. Two hours after application, the protein was readily detectable in the lymph node. (From Dang et al., 2013.)

Additional content available at https://expertconsult.inkling.com/.

Ocular IP (and its secondary phenomenon ACAID) is a manifestation of *tissue-based immune tolerance* (see online text referring to Matzinger in this Chapter). Most compartments and tissues of the eye have tissue-based 'tolerizing' properties, as do many other tissues such as the brain, testis, pancreas and several acinar tissues. This contrasts with tissues closer to the external environment, where there is greater likelihood of encountering pathogenic foreign antigens, and the default response in these tissues (e.g. skin, conjunctiva, lung and gut) is more likely to be an inflammatory response aimed at clearing the pathogen. However, this response is still modulated by powerful regulatory mechanisms, particularly by Tregs in the GI tract, the absence of which leads to severe autoimmune disease.

Recently it has been recognized that ocular immune privilege comes with risks. While it is important for the preservation of sight that immune responses within the eye do not get out of control when inflammation occurs, those modified immune defences which the eye has may not be enough to prevent an exaggerated response. Thus, certain organisms can proliferate or reactivate with less of a trigger than if they were activated in other sites. For instance, toxoplasmosis is a common cause of ocular inflammation but is still relatively infrequent given that almost a third of the population have been infected, mainly through eating infected meat. This indicates that for most of the population toxoplasmosis is controlled and indeed in the initial infection in the gut, the organism is rapidly cleared and cleared completely. Some organisms (tachyzoites) escape into the bloodstream carried within infected monocytes and find their way to sites of immune privilege where they can lie dormant and never cause a reaction. However, if the immune defences are diminished, as occurs in HIV-positive patients or patients taking immunosuppressant drugs, the ***Toxoplasma*** organisms can escape immune control and reactivate, causing a severe retina-destructive inflammation. Similar pathogenesis occurs with other organisms such as cytomegalovirus and *Mycobacterium tuberculosis* (Forrester and Xu, 2012).

Tissue-based immune tolerance (ACAID-like mechanisms) may extend to innate immunity and have implications for the ability of the eye to counteract microorganisms. Thus the eye appears to be the preferred site for certain parasites such as *Toxoplasma* and *Toxocara* (see Ch. 8) and intravenously injected fungi such as *Candida* may find a 'tolerant' environment within the retina and vitreous with disastrous effects on vision. In addition, certain viruses such as cytomegalovirus and herpes simplex can proliferate unchecked within the retina. While there is no direct evidence that this is a result of the less than optimal ('privileged') immune microenvironment within the eye, it is a possibility that remains to be tested. It seems therefore that IP may come at a cost.

Despite these properties of the intraocular compartments, bacterial infection during intraocular surgery is remarkably infrequent. This has been attributed in part to the direct bacteriostatic properties of the aqueous, which have been shown to inhibit bacterial growth *in vitro*. The nature of this activity is not known but several antibacterial proteins are present in the aqueous, including complement, immunoglobulin, defensins and β-lysin. In addition, it is dependent on the size of the bacterial inoculum and the virulence of the organism (see Ch. 8, p. 473).

Intraocular inflammation

A common cause of visual impairment is intraocular inflammation, which includes the many forms of uveitis as well as conditions such as retinitis, and even optic neuritis. Intraocular inflammation may be exogenous, in which the organism is clearly evident, as in bacterial endophthalmitis (see Ch. 9, p. 494) or cytomegalovirus retinitis, or there may be no obvious infectious agent. Most such cases are described as idiopathic or non-infectious uveitis and may affect the anterior segment (iridocyclitis, inflammation of the iris and ciliary body) or posterior segment (posterior uveitis). Posterior uveitis may take many clinical forms such as multifocal choroiditis and retinal vasculitis and formal definition and classification of the numerous entities is work in progress.

Additional content available at https://expertconsult.inkling.com/.

Non-infectious intraocular inflammation is thought to be autoimmune or at least immune-mediated. Around 50% of cases of acute anterior uveitis have a strong association with HLA-B27 alloantigen and are linked to ankylosing spondylitis and low-grade Gram-negative enteric infection with organisms such as *Yersinia, E. coli* and *Klebsiella*. Posterior uveitis is much less closely linked to MHC class I antigens, except for certain well-defined syndromes such as birdshot retinochoroidopathy (HLA-A29) and Behçet's retinal vasculitis (HLA-B51 in oriental and Middle Eastern people). Vogt–Koyanagi–Harada disease has been closely linked to HLA-DR4 (subtype DRB1 0405; see Box 7-11) and there is a similar association with sympathetic ophthalmia.

Posterior uveitis, however, has close similarity to certain experimental $CD4^+$ T-cell-mediated uveoretinal inflammations (experimental autoimmune uveoretinitis, EAU) in which the autoantigens have been well defined. Most of these are derived from the outer retinal layers and include the visual protein rhodopsin. The relationship between EAU and non-infectious posterior uveitis is, however, tantalizingly tenuous because patients with these diseases do not have significantly raised levels of antibodies to retinal antigens, although some patients manifest T-cell responses to these antigens. The most compelling evidence has recently come from a transgenic humanized mouse model in which mice demonstrated autoimmunity to human retinal S antigen, the commonest autoantigen in uveitis patients.

Clinical studies of aqueous and vitreous samples have also been relatively uninformative to date, although in certain diseases such as Fuch's heterochromic cyclitis high levels of CD8 T cells have been found. In addition, high concentrations of IL-6 and IL-8 have been detected in ocular fluid samples from patients with uveitis, and Fas–FasL interactions have been shown to be active in patients with acute anterior uveitis.

A strong hint that idiopathic non-infectious posterior uveitis is immune-mediated is its clinical response to immunosuppressive agents such as ciclosporin A and, currently, several immunological approaches, including anti-TNF-α humanized monoclonal antibodies, and oral tolerization schedules, are being evaluated for their efficacy in this condition. Other

Classification of the many forms of uveitis/intraocular inflammation as diagnostic entities has lagged behind many other diseases, mainly because of difficulties in reaching agreement about what constitutes each entity. For instance, diseases such as presumed ocular histoplasmosis syndrome are accepted entities in the USA but the same clinical disease in Europe is termed punctate inner choroidopathy. In addition, many of the conditions present with creamy-white lesions on ophthalmoscopic examination, but these ‘white dot’ syndromes comprise many different entities, both infectious and non-infectious. Accordingly, a large international consortium of experts has gathered together many of the different clinical examples and has been cataloguing and characterizing the different uveitis syndromes, both in terms of general levels of activity/severity and risk of threat to sight as well as the phenotypic characteristics which define each disease. This effort is named the Standardization of Uveitis Nomenclature (SUN) project and the group is continuing to publish its findings. These will be of great value to clinicians generally, as well as to the conduct of clinical trials of new drugs (Trusko et al., 2013).

immunosuppressive agents are now being used in a range of ocular inflammatory diseases including mycophenolate mofetil and FK506 (see Ch. 6, p. 364).

The role of the immune system in ageing and degeneration in the eye

The recent understanding of the innate immune system, its inducers and receptors and its close link with the adaptive immune system, has revealed the role of immune processes in several ocular and non-ocular disease which were previously considered genetic, developmental, metabolic, age-related/degenerative and even cancer. This includes many conditions such as diabetes and its complications, obesity, neurodegenerations such as Alzheimer's and classical genetic disease such as retinitis pigmentosa. Inflammation induced via PAMPs, DAMPS and PRRs (see p. 391) with activation of inflammasomes and secretion of IL-1 and its many family members is now known to play a significant part in the pathogenesis of many of these diseases, initiated by the original abnormality, be it developmental, viral induced, genetic or other. For instance, diabetic retinopathy is now recognized to be the result of leucocyte activation, in part via upregulation of the CCR5 receptor in response to high glucose levels. Succinate, generated in excess via the glycolytic pathway, is a central ligand for inducing inflammation in diabetes, ageing, cancer and obesity.

In the eye the most prominent example of the involvement of innate immune mechanisms and inflammation is in age-related macular degeneration (AMD). The risk of developing AMD, the commonest cause of blindness in developed nations, is directly linked to mutation in a number of innate immune genes, particularly complement factor H (see p. 405) and the risk increases with accumulation of mutations.

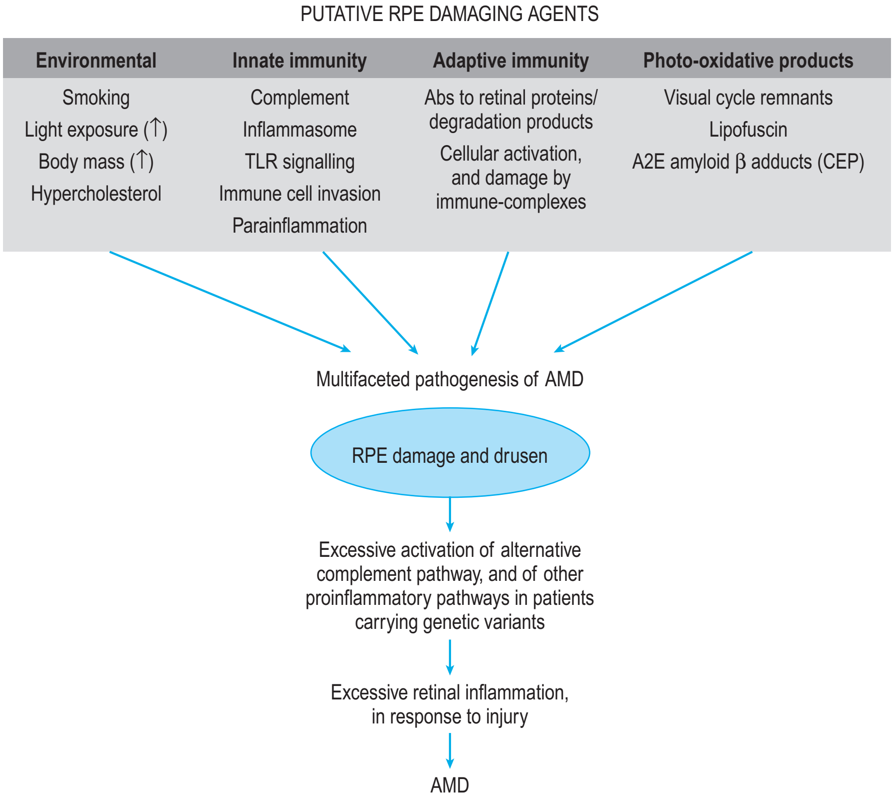
PUTATIVE RPE DAMAGING AGENTS
Environmental
Smoking
Light exposure (↑)
Body mass (↑)
Hypercholesterol
Innate immunity
Complement
Inflammasome
TLR signalling
Immune cell invasion
Parainflammation
Adaptive immunity
Abs to retinal proteins/ degradation products
Cellular activation, and damage by immune-complexes
Photo-oxidative products
Visual cycle remnants
Lipofuscin
A2E amyloid β adducts (CEP)
Multifaceted pathogenesis of AMD
RPE damage and drusen
Excessive activation of alternative complement pathway, and of other proinflammatory pathways in patients carrying genetic variants
Excessive retinal inflammation, in response to injury
AMD

FIGURE 7-46 Flow diagram outlining the main checkpoints in the development of age-related macular degeneration (AMD). (From Whitcup et al., 2013.)

The identification of associations of genetic mutations in complement proteins with AMD has opened a rich area of research into the possible role of inflammation in both wet and dry AMD. Wet AMD is characterized by the growth of abnormal vessels from the choroid through Bruch's membrane into the subretinal space where bleeding and vessel leakage cause an acute loss of vision if the lesion is at the fovea. In dry AMD there is progressive atrophy of the retinal pigment epithelium, with loss of photoreceptors and progressive choroidal avascularity. In both conditions macrophages have been implicated in wet AMD, activated macrophages release VEGF, which promotes new vessel formation, while in dry AMD, macrophages are involved in progressive cell death of the RPE layer. Macrophages are switched on by many mediators, in particular complement components (see p. 405), and complement inhibitors are in place to prevent excessive complement activation. Mutations in complement proteins, such as complement factor H (CFH), CFH-related proteins 1 and 3, factor B/C2, C3 and factor I, are linked to increased risk of AMD and for the inhibitory proteins evidence has been produced to show that proteins such as complement factor H are less effective in controlling C activation if they contain one or more of the mutations. The pathogenetic role of complement factors in AMD is, however, complex. Partial activation of complement can produce factors which promote removal of tissue debris and thus prevent accumulation of material which is presumed to lead to overt AMD such as drusen. However, defective inhibitors clearly promote overactivation of complement. Although most of the complement components are produced in the liver, several can be generated locally within the retina and other parts of the eye, and thus locally generated inflammatory responses are possible given the correct triggers.

Additional content available at https://expertconsult.inkling.com/.

Macrophages and dendritic cells are directly implicated in the pathology of AMD, classical inflammatory macrophages in the early stages and alternative (angiogenic) macrophages when active choroidal neovascularization occurs in the wet form of the disease. In the atrophic form of AMD, inflammasome products IL-1 and IL-18 appear to be important in causing progressive atrophy of the RPE, while in the wet form of the disease IL-18 has been proposed to be anti-angiogenic. However, there are many different aspects to AMD pathogenesis which involve more than complement and induction of the inflammasome and AMD truly reflects a multifactorial disease (Fig. 7-46).

Much is still to be learned concerning the role of the innate immune system, and possibly the adaptive immune system in AMD pathogenesis, as well as in other diseases such as glaucoma and intraocular tumours, not to mention more obvious inflammatory diseases such as uveitis, but a rich harvest of potential therapeutic targets using immune modulators has yet to be reaped.

Conclusion

The eye and its several tissues may be involved in any of the immune responses described in this chapter, either as a primary target of attack (e.g. in disciform herpetic keratitis or toxoplasmic choroiditis) or as part of a generalized immune disorder such as in Wegener's granulomatosis or sarcoidosis. The pathological processes and the mechanisms of initiation of immune responses are fundamentally similar from tissue to tissue. However, as stated in the introduction, each tissue has its unique microenvironment and this undoubtedly plays a part in the final expression of the immune response.

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

FURTHER READING

Anderson, G., Beischlag, T.V., Vinciguerra, M., Mazzoccoli, G., 2013. The circadian clock circuitry and the AHR signaling pathway in physiology and pathology. Biochem. Pharmacol. 85, 1405–1416.

Bajénoff, M., 2012. Stromal cells control soluble material and cellular transport in lymph nodes. Front Immunol. 3, 304.

Cho, I., Blaser, M.J., 2012. The human microbiome: at the interface of health and disease. Nat. Rev. Genet. 13, 260–270.

Chong, A.S., Alegre, M.L., 2012. The impact of infection and tissue damage in solid-organ transplantation. Nat. Rev. Immunol. 12, 459–471.

Craft, J.E., 2012. Follicular helper T cells in immunity and systemic autoimmunity. Nat. Rev. Rheumatol. 8, 337–347.

Croft, M., Duan, W., Choi, H., Eun, S.Y., Madireddi, S., Mehta, A., 2012. TNF superfamily in inflammatory disease: translating basic insights. Trends Immunol. 33, 144–152.

Croker, B.A., Lawson, B.R., Rutschmann, S., Berger, M., Eidenschenk, C., Blasius, A.L., et al., 2008. Inflammation and autoimmunity caused by a SHP1 mutation depend on IL-1, MyD88, and a microbial trigger. Proc. Natl Acad. Sci. U.S.A. 105, 15028–15033.

Dang, Z., Kuffova, L., Liu, L., Forrester, J.V., 2013. Soluble antigen traffics rapidly and selectively from the corneal surface to the eye draining lymph node and activates T cells when codelivered with CpG oligonucleotides. J. Leukoc. Biol. 95, 431–440.

Dustin, M.L., Depoil, D., 2011. New insights into the T cell synapse from single molecule techniques. Nat. Rev. Immunol. 11, 672–684.

Engelhardt, B., Ransohoff, R.M., 2012. Capture, crawl, cross: the T cell code to breach the blood-brain barriers. Trends Immunol. 33, 579–589.

Fallarino, F., Grohmann, U., Puccetti, P., 2012. Indoleamine 2,3-dioxygenase: from catalyst to signaling function. Eur. J. Immunol. 42, 1932–1937.

Forrester, J.V., Xu, H., 2012. Good news-bad news: the Yin and Yang of immune privilege in the eye. Front Immunol. 3, 338.

Geering, B., Stoeckle, C., Conus, S., Simon, H.U., 2013. Living and dying for inflammation: neutrophils, eosinophils, basophils. Trends Immunol. 34, 398–409.

Ginhoux, F., Merad, M., 2011. [Microglia arise from extra-embryonic yolk sac primitive progenitors]. Med. Sci. (Paris) 27, 719–724.

Greenwell-Wild, T., Moutsopoulos, N.M., Gliozzi, M., Kapsogeorgou, E., Rangel, Z., Munson, P.J., et al., 2011. Chitinases in the salivary glands and circulation of patients with Sjogren's syndrome: macrophage harbingers of disease severity. Arthritis Rheum. 63, 3103–3115.

Greter, M., Lelios, I., Pelczar, P., Hoeffel, G., Price, J., Leboeuf, M., et al., 2012. Stroma-derived interleukin-34 controls the development and maintenance of Langerhans cells and the maintenance of microglia. Immunity 37, 1050–1060.

Hara, Y., Shoji, J., Hori, S., Ishimori, A., Kato, H., Inada, N., et al., 2012. Evaluation of eosinophilic inflammation in a novel murine atopic keratoconjunctivitis model induced by crude Dermatophagoides farinae antigen. Allergol. Int. 61, 331–338.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., et al., 2004. Gene map of the extended human MHC. Nat. Rev. Genet. 5, 889–899.

Hunter, C.A., Kastelein, R., 2012. Interleukin-27: balancing protective and pathological immunity. Immunity 37, 960–969.

Joffre, O.P., Segura, E., Savina, A., Amigorena, S., 2012. Cross-presentation by dendritic cells. Nat. Rev. Immunol. 12, 557–569.

Kim, M., Johnston, M.G., Gupta, N., Moore, S., Yucel, Y.H., 2011. A model to measure lymphatic drainage from the eye. Exp. Eye Res. 93, 586–591.

Kiss, E.A., Diefenbach, A., 2012. Role of the aryl hydrocarbon receptor in controlling maintenance and functional programs of RORgammat(+) innate lymphoid cells and intraepithelial lymphocytes. Front Immunol. 3, 124.

Klein, T., Bischoff, R., 2011. Physiology and pathophysiology of matrix metalloproteases. Amino Acids 41, 271–290.

Kolaczkowska, E., Kubes, P., 2013. Neutrophil recruitment and function in health and inflammation. Nat. Rev. Immunol. 13, 159–175.

Kumar, P., Rajasekaran, K., Palmer, J.M., Thakar, M.S., Malarkannan, S., 2013. IL-22: An evolutionary missing-link authenticating the role of the immune system in tissue regeneration. J. Cancer 4, 57–65.

Lalor, P.F., Faint, J., Aarbodem, Y., Hubscher, S.G., Adams, D.H., 2007. The role of cytokines and chemokines in the development of steatohepatitis. Semin. Liver Dis. 27, 173–193.

Lazennec, G., Richmond, A., 2010. Chemokines and chemokine receptors: new insights into cancer-related inflammation. Trends Mol. Med. 16, 133–144.

Manz, B.N., Groves, J.T., 2010. Spatial organization and signal transduction at intercellular junctions. Nat. Rev. Mol. Cell Biol. 11, 342–352.

Matzinger, P., Kamala, T., 2011. Tissue-based class control: the other side of tolerance. Nat. Rev. Immunol. 11, 221–230.

Merad, M., Ginhoux, F., Collin, M., 2008. Origin, homeostasis and function of Langerhans cells and other langerin-expressing dendritic cells. Nat. Rev. Immunol. 8, 935–947.

Metz, M., Maurer, M., 2007. Mast cells – key effector cells in immune responses. Trends Immunol. 28, 234–241.

Mills, K.H., 2011. TLR-dependent T cell activation in autoimmunity. Nat. Rev. Immunol. 11, 807–822.

Minton, K., 2011. Granulocytes: a weighty role for eosinophils. Nat. Rev. Immunol. 11, 299.

Mjosberg, J., Bernink, J., Peters, C., Spits, H., 2012. Transcriptional control of innate lymphoid cells. Eur. J. Immunol. 42, 1916–1923.

Mosser, D.M., Edwards, J.P., 2008. Exploring the full spectrum of macrophage activation. Nat. Rev. Immunol. 8, 958–969.

Murray, P.J., Wynn, T.A., 2011. Protective and pathogenic functions of macrophage subsets. Nat. Rev. Immunol. 11, 723–737.

Neyt, K., Perros, F., GeurtsvanKessel, C.H., Hammad, H., Lambrecht, B.N., 2012. Tertiary lymphoid organs in infection and autoimmunity. Trends Immunol. 33, 297–305.

Ng, D., Gommerman, J.L., 2013. The regulation of immune responses by DC derived type I IFN. Front Immunol. 4, 94.

Quintana, F.J., 2013. The aryl hydrocarbon receptor: a molecular pathway for the environmental control of the immune response. Immunology 138, 183–189.

Rask-Madsen, C., Kahn, C.R., 2012. Tissue-specific insulin signaling, metabolic syndrome, and cardiovascular disease. Arterioscler. Thromb. Vasc. Biol. 32, 2052–2059.

St John, A.L., Abraham, S.N., 2013. Innate immunity and its regulation by mast cells. J. Immunol. 190, 4458–4463.

Satpathy, A.T., Wu, X., Albring, J.C., Murphy, K.I.M., 2012. Re(de)fining the dendritic cell lineage. Nat. Immunol. 13, 1145–1154.

Swain, S.L., McKinstry, K.K., Strutt, T.M., 2012. Expanding roles for CD4(+) T cells in immunity to viruses. Nat. Rev. Immunol. 12, 136–148.

Timar, C.I., Lorincz, A.M., Ligeti, E., 2013. Changing world of neutrophils. Pflugers Arch. 465, 1521–1533.

Trusko, B., Thorne, J., Jabs, D., Belfort, R., Dick, A., Gangaputra, S., et al., 2013. The Standardization of Uveitis Nomenclature (SUN) project. Development of a clinical evidence base utilizing informatics tools and techniques. Methods Info. Med. 52, 259–265.

Usui, Y., Parikh, J., Goto, H., Rao, N.A., 2008. Immunopathology of necrotising scleritis. Br. J. Ophthalmol. 92, 417–419.

van de Veerdonk, F.L., Netea, M.G., 2013. New insights in the immunobiology of IL-1 family members. Front Immunol. 4, 167.

Vantourout, P., Hayday, A., 2013. Six-of-the-best: unique contributions of gammadelta T cells to immunology. Nat. Rev. Immunol. 13, 88–100.

Veldhoen, M., Brucklacher-Waldert, V., 2012. Dietary influences on intestinal immunity. Nat. Rev. Immunol. 12, 696–708.

Vitova, A., Kuffová, L., Klaska, I.P., Holan, V., Cornall, R.J., Forrester, J.V., 2013. The high-risk corneal regraft model: a justification for tissue matching in humans. Transpl. Int. 26, 453–461.

Voehringer, D., 2013. Protective and pathological roles of mast cells and basophils. Nat. Rev. Immunol. 13, 362–375.

Whitcup, S.M., Sodhi, A., Atkinson, J.P., Holers, V.M., Sinha, D., Rohrer, B., et al., 2013. The role of the immune response in age-related macular degeneration. Int J Inflam 348092.

Williams, K.A., Coster, D.J., 2007. The immunobiology of corneal transplantation. Transplantation 84, 806–813.

Yucel, Y.H., Johnston, M.G., Ly, T., Patel, M., Drake, B., Gümüş, E., et al., 2009. Identification of lymphatics in the ciliary body of the human eye: a novel 'uveolymphatic' outflow pathway. Exp. Eye Res. 89, 810–819.

Ziegler, A., Kentenich, H., Uchanska-Ziegler, B., 2005. Female choice and the MHC. Trends Immunol. 26, 496–502.

Zipfel, P.F., Skerka, C., 2009. Complement regulators and inhibitory proteins. Nat. Rev. Immunol. 9, 729–740.

Chapter 8

Microbial infections of the eye

- Introduction
- Microbes in the environment
- Host defences at the ocular surface
- Innate immunity to microbial infection
- Ocular infections worldwide
- Ocular infections in developing countries
- Conclusion

Introduction

The outcome of microbial infections depends on both the virulence of the infecting organism and the veracity of the host response. The balance of these two entities will determine the severity of disease and the longevity of infection. While in some infections tissue damage is a direct result of cytotoxic activity of microbial products such as exotoxins, the host inflammatory response causes collateral tissue damage associated with antimicrobial activity. Microbial virulence may then be defined as the ability of the organisms to survive in the presence of the host immune response. This chapter will focus on infections of the anterior eye based on exposure to the ocular surface, but will also discuss infections of the posterior eye, including endophthalmitis and ocular toxoplasmosis.

Microbes in the environment

Most of the organisms that cause severe ocular infections are either ubiquitous in the environment or are part of our normal body flora. These microbes are opportunistic pathogens and require breach of the physical barriers of the eye. For example, *Pseudomonas aeruginosa* and *Acanthamoeba* are normal freshwater organisms present in ponds, lakes and household water supplies, including showerheads. Similarly, *Aspergillus* and *Fusarium* moulds, which are plant saprophytes and pathogens, are ubiquitous in the air we breathe, although spore counts are higher in hot and humid areas of the world, and in agricultural regions, especially during harvest seasons. *Aspergillus* is among the most widely distributed organisms worldwide and, except in extremely cold environments, most individuals inhale *Aspergillus* spores on a daily basis. In humid conditions *Aspergillus* is the cause of the green mould found on bread and other foods, and routine examination of the average bathroom will detect *Aspergillus* and *Fusarium* spores, *Pseudomonas*, *Serratia* and *Acanthamoeba* (which feeds on bacteria and fungi). The number of organisms in the environment is generally higher in warmer climates and during the summer in temperate climates.

In addition to the external environment, several organisms are present in the normal body flora of the human conjunctiva (adenovirus), the skin (staphylococci) and the nasopharynx (streptococcus pneumoniae, candida albicans). Most individuals also harbour herpes simplex virus (HSV) and herpes zoster as latent forms in the trigeminal ganglia. HSV is the most common cause of corneal infections in the USA and other industrialized countries and causes resurgent keratitis following viral exit from latency. *Toxoplasma gondii* is also ubiquitous in the environment and humans are infected following ingestion of cat faecal material or infected meat.

The exceptions to pathogens with near ubiquitous distribution are those that are very restricted geographically or which thrive under conditions of poor hygiene. The bacterium that causes trachoma (*Chlamydia trachomatis*) is the major example of an organism that thrives in unsanitary environments, as it is readily transmitted by flies that breed on human waste. *Onchocerca volvulus*, the cause of river blindness, and *Loa loa* (eyeworm) are highly adapted human parasites that require an intermediate insect vector, and are found primarily in Africa. These will be discussed later in this chapter. Table 8-1 lists many of the causes of ocular infection.

Host defences at the ocular surface

PHYSICAL BARRIERS

Blinking is a very effective cleansing mechanism and eyelashes can trap microbes, preventing access to the globe. In addition, the eyelids contain sebaceous glands that secrete lactic acid and fatty acids in a low pH environment, which has a direct inhibitory effect on bacterial replication. Tears also contain antimicrobial compounds including lacritin, lactoferrin, lipocalin lysozyme, β-defensins and other antimicrobial peptides. The tear film also contains secretory immunoglobulin A (IgA), IgG and complement that contribute to protection against bacterial invasion (see Ch. 7). Protection is also afforded by neutrophils in the tears, which increase in numbers following sleep when the eyelids are closed.

TABLE 8-1 **Microbial pathogens of the eye***

Group/phylum	Genus, species	Major site of infection
Viruses	Adenovirus	Conjunctiva
	Herpes simplex	Cornea
	Herpes zoster	Cornea
	CMV	Retina
Obligate intracellular	*Chlamydia*	Conjunctiva, cornea (trachoma)
Gram-negative bacteria	PA	Cornea
	Sm	Cornea
Gram-positive bacteria	SP	Conjunctiva, cornea, vitreous
	Sa	Cornea
	Bacillus cereus	Vitreous
Yeast	*Candida*	Cornea
Mould	*Fusarium*	
	Aspergillus	
Protozoa	*Toxoplasma*	Cornea
	Acanthamoeba	
	Microsporidia	
Helminths	*Toxocara canis*	Retina
Filarial	*Onchocerca volvulus* (river blindness)	Cornea, retina
	Loa loa (eyeworm)	Conjunctiva

*Although a search of case reports will produce a much longer list of organisms, the table includes only the most common causes of ocular infection.

CMV, cytomegalovirus; PA, *Pseudomonas aeruginosa;* Sm, *Serratia marcescens;* SP, *Streptococcus pneumoniae;* Sa, *Staphylococcus aureus.*

EPITHELIUM

The corneal epithelium has a glycocalyx on the apical surface composed of large proteoglycan mucins including MUC1, MUC4 and MUC16 (see Ch. 4). This mucin layer provides a physical barrier that restricts bacterial adherence to corneal epithelial cells in addition to releasing mucins into the tear film (Fig. 8-1). Epithelial cells also form tight junctions that are a very effective barrier against subsequent penetration to the corneal stroma; infections therefore generally occur as a consequence of traumatic injury or alteration of the corneal surface microenvironment. Epithelial cells also secrete antimicrobial peptides including β-defensins, cathelicidin (LL-37) and calprotectin (S100A8/A9), and by degradation of cytokeratin-6A in corneal epithelial cells. These keratin-derived

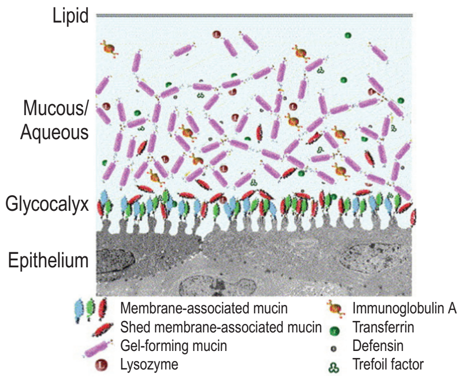

FIGURE 8-1 Ocular surface mucins. The aqueous phase of the tears contains MUC5AC (purple) and other tear components. The glycocalyx has extended membrane-associated MUC1 (blue), MUC4 (green) and MUC16 (red). Extracellular domains are released into the tear film. (From Gipson, 2004.)

antimicrobial peptides (KDAMPs) have a distinct secondary structure and exhibit broad antimicrobial activity.

RESIDENT MACROPHAGES AND DENDRITIC CELLS IN THE CORNEA

The normal mammalian cornea was generally thought to be devoid of immune cells; however, with the advent of improved immunostaining methods, together with examining whole-mount corneas rather than histological sections, an entire network of macrophages and dendritic cells was revealed. Most bone marrow-derived cells in the normal corneal stroma have characteristics of macrophages, with dendritic cells more prominent in the peripheral limbal region of the cornea and in the basement membrane of the epithelium (Bowman's membrane). These cells extend pseudopodia ('periscopes') through to the apical surface, presumably to detect microbes or and microbial products (Fig. 8-2A–D). Nanotubes that appear to connect distant cells can also be detected in the corneal stroma during inflammation (Fig. 8-2D).

PATHOGEN RECOGNITION RECEPTORS AND RECRUITMENT OF NEUTROPHILS

Most nucleated cells are able to recognize and respond to microbial products, although macrophages and dendritic cells are specifically adapted for this purpose. These cells express multiple copies of surface receptors that recognize bacterial and fungal proteins, carbohydrates, lipids and DNA and RNA, and are termed pathogen recognition receptors (PRRs) (see Ch. 7, p. 379). Ligand binding initiates intracellular signalling events that result in production of pro-inflammatory and chemotactic cytokines. These are invariant receptors encoded in the germ line genes, which is in contrast to T and B cells, where receptors are generated following gene rearrangement. Further, whereas mature T and B cells primarily recognize specific peptides (and in some cases well-defined carbohydrates), the invariant receptors associated with innate immunity recognize conserved proteins, lipids and nucleic acids primarily associated with microbes. Though not a focus of the current chapter, these receptors can also recognize endogenous self-antigens, danger-associated molecular patterns (DAMPs), which include silica, uric acid or asbestos crystals, and advanced glycation end products. Among the best characterized receptors are cell surface and endosomal Toll-like receptors, cell surface C-type lectins, and intracellular NOD-like receptors (NLRs).

TOLL-LIKE RECEPTORS (TLR)

TLR family members are single transmembrane receptors that recognize structurally conserved microbial products; further, activation of these receptors leads to production of pro-inflammatory and chemotactic cytokines that mediate recruitment of neutrophils, macrophages and lymphocytes to the site of infection. As shown in Figure 8-3, TLRs are located in cholesterol rich regions of the plasma cell membrane (lipid rafts) and in endosomes, and can recognize lipids, proteins or nucleic acids. Lipid-binding TLRs include TLR2, which forms heterodimers with TLR1 or TLR6 to bind lipopeptides, and TLR4/MD-2, which recognize the lipid A moiety of lipopolysaccharide (LPS). TLR5 and TLR11 recognize proteins, and are activated by bacterial flagellin (TLR5) or uropathogenic *E. coli* or *Toxoplasma gondii* profilin (TLR11). TLR3, TLR7, TLR8 and TLR9 are located on endosomal membranes and bind viral and bacterial nucleic acids. With the exception of TLR3, all TLRs stimulate the cells through the MyD88 common adaptor molecule, leading to NF-κB translocation to the nucleus and expression of genes encoding pro-inflammatory cytokines and chemotactic cytokines (chemokines). TLR3 and TLR4 activate the TRIF pathway, which induces IRF3 transcription and production of type I interferons that mediate antiviral responses. TLR4 activation involves accessory molecules, including lipopolysaccharide (LPS)-binding protein and CD14, which combine to extract single endotoxin molecules from the outer membrane and form monomeric endotoxin, and MD-2, which is the receptor for the lipid A moiety of LPS. CD14 also chaperones TLR4 from the plasma membrane to endosomes in order to activate the TRIF pathway.

NOD-LIKE RECEPTORS

NOD-like receptors (NLRs) comprise an intracellular family of pathogen recognition molecules which also activate the NF-κB complex, leading to expression of pro-inflammatory and chemotactic cytokines. NOD2 has been well characterized and shown to recognize

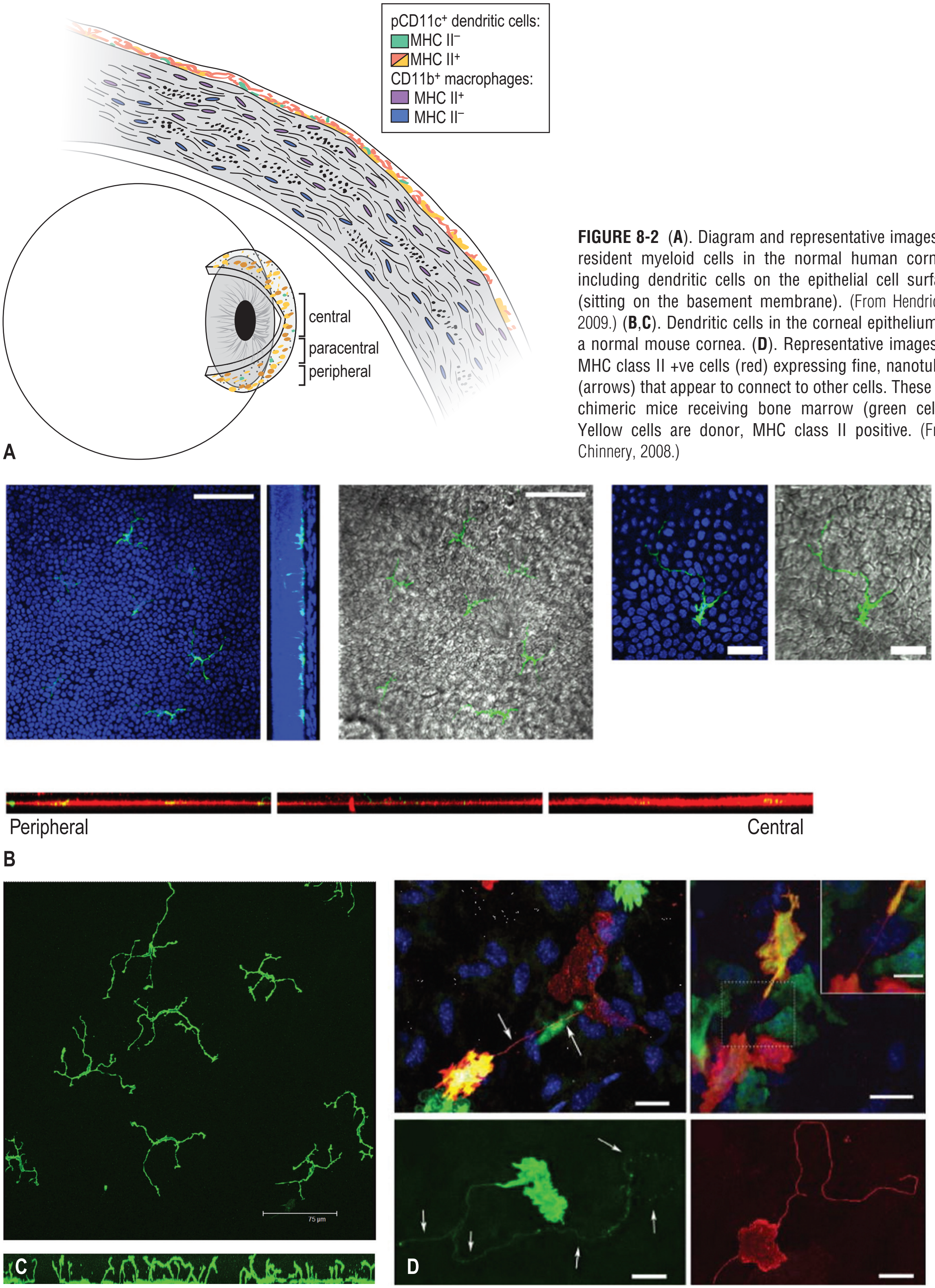
pCD11c⁺ dendritic cells:
🟩 MHC II⁻
🟥 MHC II⁺
CD11b⁺ macrophages:
🟪 MHC II⁺
🟦 MHC II⁻
central
paracentral
peripheral
A
Peripheral
Central
B
75 µm
C
D

FIGURE 8-2 (**A**). Diagram and representative images of resident myeloid cells in the normal human cornea, including dendritic cells on the epithelial cell surface (sitting on the basement membrane). (From Hendricks, 2009.) (**B**,**C**). Dendritic cells in the corneal epithelium of a normal mouse cornea. (**D**). Representative images of MHC class II +ve cells (red) expressing fine, nanotubes (arrows) that appear to connect to other cells. These are chimeric mice receiving bone marrow (green cells). Yellow cells are donor, MHC class II positive. (From Chinnery, 2008.)

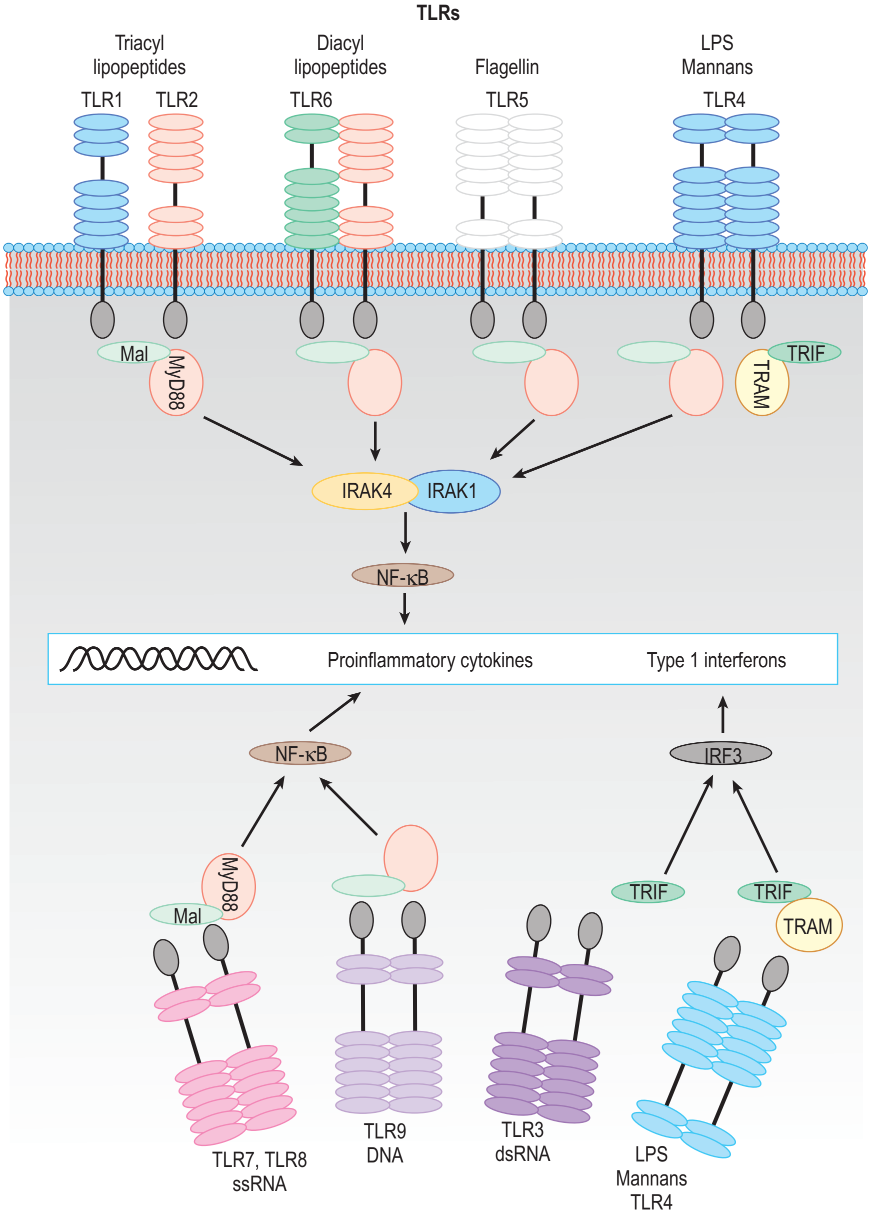
TLRs
Triacyl lipopeptides
Diacyl lipopeptides
Flagellin
LPS Mannans
TLR1
TLR2
TLR6
TLR5
TLR4
Mal
MyD88
TRIF
TRAM
IRAK4
IRAK1
NF-κB
Proinflammatory cytokines
Type 1 interferons
NF-κB
IRF3
MyD88
Mal
TRIF
TRIF
TRAM
TLR7, TLR8
ssRNA
TLR9
DNA
TLR3
dsRNA
LPS
Mannans
TLR4

FIGURE 8-3 Toll-like receptors (TLR) respond to cell surface and endosomal microbial products: ss, single stranded; ds: double stranded. (From O'Neill et al., 2013.)

muramyl dipeptide on degraded bacterial cell wall peptidoglycan, and can therefore respond to invading Gram-negative and Gram-positive bacteria (Box 8-1). However, as Gram-positive bacteria have more peptidoglycan in the cell wall, NOD2 activation occurs following infection by staphylococci or streptococci.

NLRs include NLRP3, which recognizes bacterial toxins and crystals, and NLRC4, which recognizes flagellin of Gram-negative bacteria such as *Pseudomonas aeruginosa* (Fig. 8-4). Once activated, these NLRs form a large, multi-protein complex called an inflammasome, which activates caspase-1 and cleaves IL-1β, IL-18 and IL-33 from the inactive pro-form, to the bioactive, mature form of these cytokines. Prolonged activation of inflammasomes also leads to caspase-1-mediated cell death, termed pyroptosis.

BOX 8-1 NOD2 AND INFLAMMATION

Mutations in NOD2 are associated with susceptibility to autoimmune diseases that include Crohn's disease, which is a common and painful form of inflammatory bowel disease, and Blau's syndrome, which is manifest by multiple autoimmune disorders, including a severe form of uveitis. Although not completely understood, some of the polymorphisms of NOD2 result in hyperresponsiveness to MDP from otherwise harmless commensal bacteria in the intestine and skin.

C-TYPE LECTINS

In addition to bacterial products, host cells recognize fungal cell wall components by activation of C-type lectins on the cell surface (Fig. 8-5). Dectin-1 recognizes β-glucan, whereas Dectin-2 and Dectin-3 recognize α-mannans, and are activated after either clustering of Dectin-1 or heterodimerization of Dectin-2 and Dectin-3, which leads to production of pro-inflammatory and chemotactic cytokines.

NEUTROPHILS

Neutrophils are the first cells to respond to invading microbes. They are the most abundant leucocytes in the blood of normal individuals, comprising ~25% total white cells. Neutrophils in the blood constitutively express receptors for chemokines, specifically

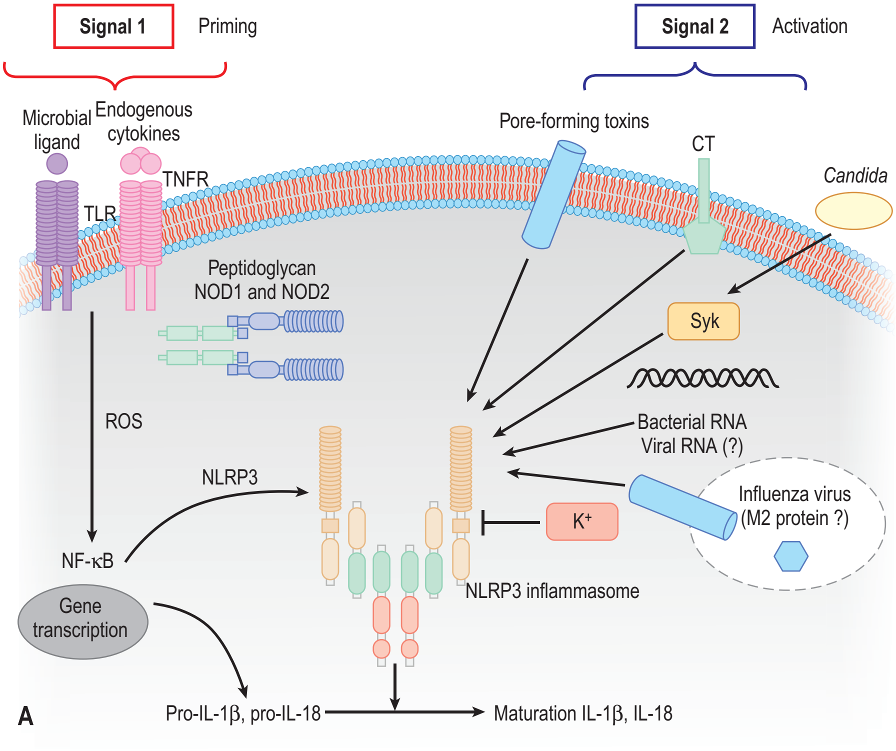
Signal 1
Priming
Signal 2
Activation
Microbial ligand
Endogenous cytokines
TNFR
TLR
Peptidoglycan
NOD1 and NOD2
Pore-forming toxins
CT
Candida
Syk
Bacterial RNA
Viral RNA (?)
Influenza virus
(M2 protein ?)
K⁺
ROS
NLRP3
NF-κB
Gene transcription
NLRP3 inflammasome
Pro-IL-1β, pro-IL-18
Maturation IL-1β, IL-18
A

FIGURE 8-4 Nod-like receptors NLRP3 (A) and NLRC4 (B) activation of IL-1b. Signal 1 TLR activation leads to gene expression of the IL-1 pro-form through NF-κB. Signal 2 activates the NLRP3 or NLRC4 inflammasome complexes that activate caspase 1 and cleavage of IL-1β to the bioactive 17 kDa form that is secreted from the cells. (From Nunez, Nat Imm Rev 2012.)

Continued

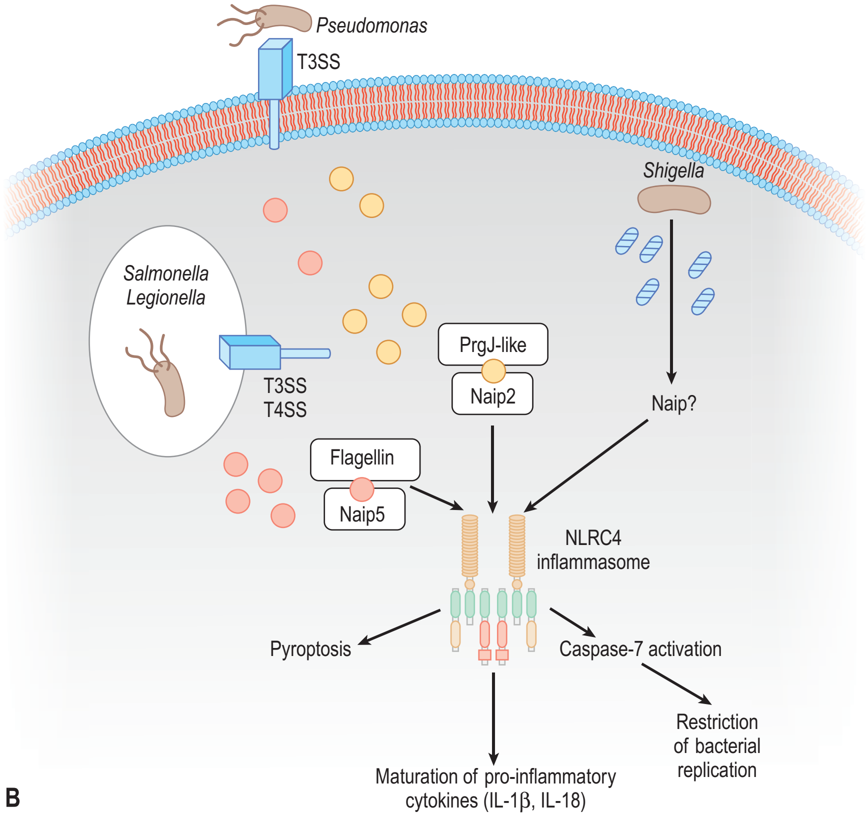
Pseudomonas
T3SS
Shigella
Salmonella
Legionella
T3SS
T4SS
PrgJ-like
Naip2
Naip?
Flagellin
Naip5
NLRC4
inflammasome
Pyroptosis
Caspase-7 activation
Restriction
of bacterial
replication
Maturation of pro-inflammatory
cytokines (IL-1β, IL-18)
B

FIGURE 8-4, cont'd

CXCR1, which binds IL-8/CXCL8, and CXCR2, which binds CXCL1, CXCL2 and CXCL5 (although murine neutrophils express CXCR1, they do not produce IL-8 and primarily respond to CXCL1, 2 and 5 through CXCR2). Neutrophils also constitutively express adhesion molecules that bind to receptors on vascular endothelial cells (Box 8-2).

Neutrophils (and other leucocytes) are recruited from capillaries (including limbal blood vessels) to infected tissues through a sequence of events in which pro-inflammatory cytokines produced at the site of infection induce expression of adhesion molecules on vascular endothelial cells. Selectins mediate tethering, intracellular adhesion molecules (ICAM-1, -2 and VCAM-1) bind to integrins on the neutrophils, and chemokines stimulate their transmigration across the vascular endothelium (Fig. 8-6A). Once in the tissue, migration of neutrophils to the site of infection is also dependent on a chemokine gradient, especially in an avascular tissue such as the cornea, where cells need to migrate from peripheral limbal vessels.

BOX 8-2 CHEMOTACTIC CYTOKINES (CHEMOKINES)

CXC chemokines such as IL-8 are directly (and specifically) chemotactic for neutrophils, whereas pro-inflammatory cytokines such as IL-1α, IL-1β and TNF-α can induce increased expression of vascular cell adhesion molecules on capillary endothelial cells in the limbus. Expression of these adhesion molecules has a critical role in tethering, binding and facilitating neutrophil transmigration into the corneal stroma (Fig. 8.6). This is a general mechanism for extravasation of leucocytes such as lymphocytes, where expression of specific adhesion molecules and chemokines on vascular endothelial cells are recognized by specific receptors on different cell types, thereby coordinating the cellular recruitment to the tissue.

The response of neutrophils to bacteria and yeasts involves phagocytosis, degranulation and neutrophil extracellular trap formation (NETs, Fig. 8-6B). If they

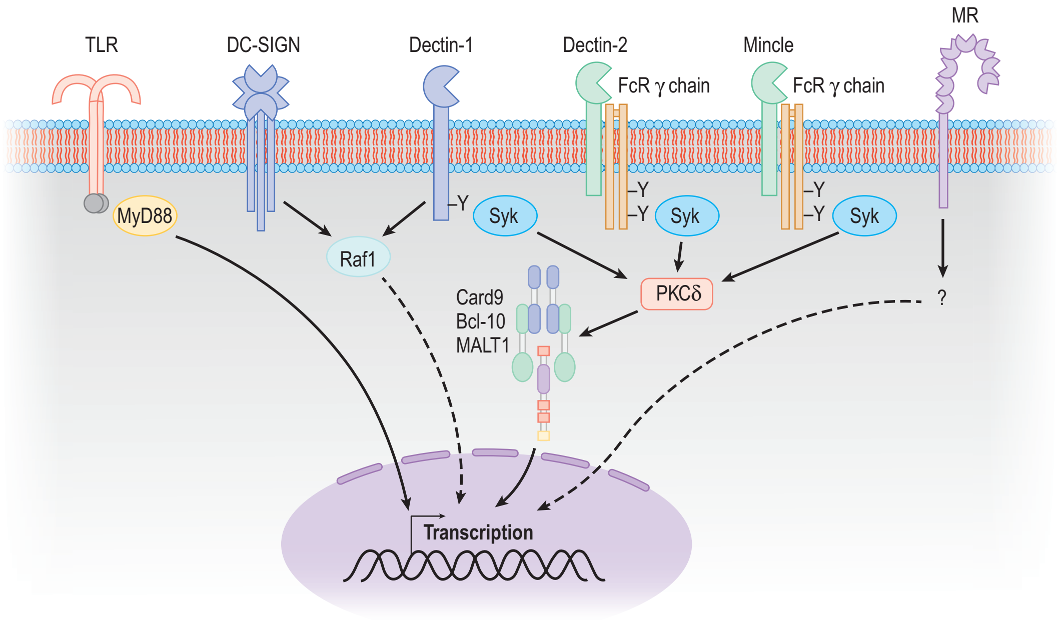

FIGURE 8-5 C-type lectins recognize fungal cell wall carbohydrates and mediate antifungal immunity, including phagocytosis and cell signalling. (From Brown, Nat Imm 2012.)

are unable to ingest the much larger fungal hyphae or *Acanthamoeba*, they can bind to the pathogen surface and release cytotoxic components, including reactive oxygen and nitrogen species, antimicrobial peptides, serine proteases and matrix metalloproteinases. They can also undergo NETosis, releasing these cytotoxic components in the context of a DNA/histone-rich net. Neutrophils also express pathogen recognition molecules and ligand activation stimulates production of pro-inflammatory and chemotactic cytokines that exacerbate neutrophil infiltration and recruit other cells such as macrophages and T cells. The role of macrophages and dendritic cells in the retina and uveal tract is discussed in Chapters 1 and 2. However, these mediators are also cytotoxic, and can contribute to loss of epithelial cells and keratocytes. Further matrix metalloproteinases can degrade the stromal collagen resulting in visual impairment and corneal scarring.

Adaptive immunity to microbial infection

Although most bacterial infections are extracellular and are dealt with rapidly by innate immunity, intracellular pathogens such as mycobacteria, protozoa and especially viruses survive longer and can induce an adaptive immune response. Thus they stimulate T- and B-cell responses which control infection and regulate the severity of infection. For example, CD4 cells play an important regulatory role in herpes simplex keratitis, and CD8 cells regulate herpes latency in trigeminal ganglia. Also, long-term exposure to airborne fungal spores induces systemic T-cell responses that likely regulate the severity of subsequent corneal infection. Toxoplasmosis and onchocerciasis are examples of chronic infections in which adaptive immunity plays an important role in determining disease severity and outcome. Other factors that affect adaptive immunity and increased susceptibility to microbial infections include immunosuppressive drugs and HIV infection (see also Ch. 7, p. 437).

Ocular infections worldwide

CONTACT LENSES

Contact lenses are a major risk factor for microbial keratitis. Approximately 134 million people worldwide wear contact lenses, and lens wear is the most common risk factor for corneal infections in the industrialized world. Long-term contact lenswear inhibits epithelial cell proliferation and migration and suppresses limbal stem cell production of basal corneal

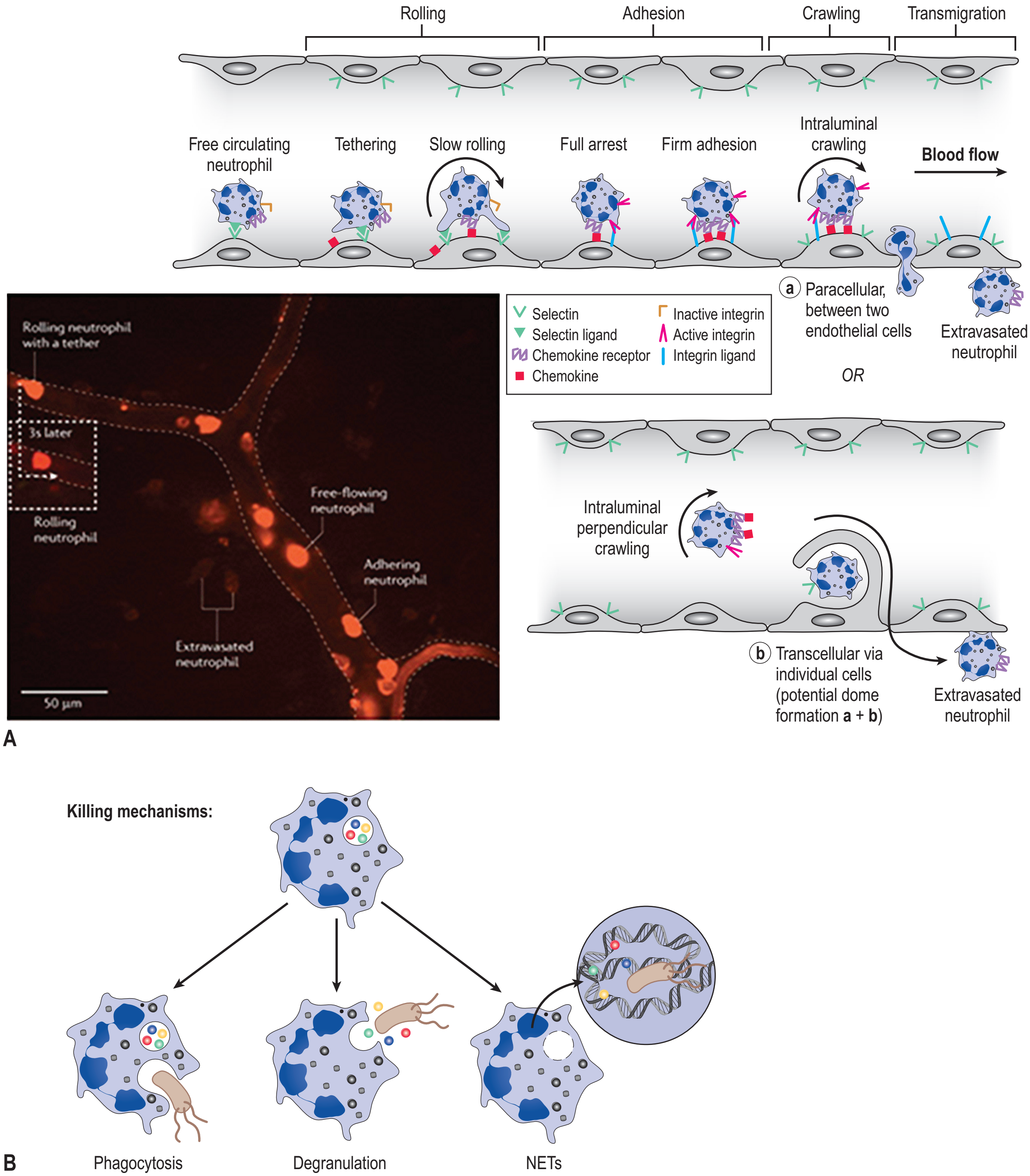
Rolling
Adhesion
Crawling
Transmigration
Free circulating neutrophil
Tethering
Slow rolling
Full arrest
Firm adhesion
Intraluminal crawling
Blood flow
Rolling neutrophil with a tether
3s later
Rolling neutrophil
Free-flowing neutrophil
Adhering neutrophil
Extravasated neutrophil
50 µm
Selectin
Selectin ligand
Chemokine receptor
Chemokine
Inactive integrin
Active integrin
Integrin ligand
a Paracellular, between two endothelial cells
Extravasated neutrophil
OR
Intraluminal perpendicular crawling
b Transcellular via individual cells (potential dome formation a + b)
Extravasated neutrophil
A
Killing mechanisms:
B
Phagocytosis
Degranulation
NETs

FIGURE 8-6 Neutrophil infiltration and activation. (**A**) Multistep process of neutrophil migration from capillaries to sites of infection. (**B**) Antimicrobial activity of neutrophils occurs: phagocytosis, degranulation and release of reactive oxygen and proteolytic enzymes. A third mode of killing is formation of neutrophil extracellular traps (NETs) where neutrophils release DNA, histones and other microbicidal proteins as they undergo necrosis. (From Kubes, Nat Imm Rev 2013.)

epithelial cells. Soft contact lenses, especially extended wear lenses, alter the microenvironment of the ocular surface by reducing the flow and effectiveness of tears, and trapping microbes at the cell surface. Further, poor hygiene in relation to lens cases and lens care solutions facilitates growth of bacteria and fungi, often forming multi-organism biofilms which are more resistant to antibiotics and lens care solutions. In this outbreak, one Lens care solution was found to be ineffective killing clinical and environmental isolates of *Fusarium*, resulting in over 300 cases of keratitis in the USA, Europe and Singapore in 2005/2006. Similarly, the increased incidence of *Acanthamoeba* keratitis in Chicago in 2007 was initially thought to be due to a lens care solution, resulting in withdrawal of the product; however, later findings showed the outbreak was due instead to reduced chlorination of the Chicago river.

VIRAL INFECTIONS OF THE EYE

There are several viruses that infect the eye, including adenovirus, which causes epidemic keratoconjunctivitis (serotypes 3, 7, 8 and 19), and pharyngoconjunctival fever (serotypes 1, 2, 3, 5, 7 and 14). Human papillomavirus causes epithelial proliferation resulting in formation of benign papilloma (warts) on the lids and conjunctiva. Herpes zoster ophthalmicus causes an extremely painful corneal infection as well as extensive involvement of the skin in the same dermatome served by the ophthalmic division of the trigeminal nerve (see Ch. 1, p. 74) and occurs following resurgence of latency from the nerve. However, globally, herpes simplex virus 1 is the most common cause of ocular viral infections.

Herpes stromal keratitis

Herpes simplex virus 1 (HSV-1) is among the most common causes of ocular infections worldwide, as evidenced by high seroprevalence rates in industrialized and developing countries. Infection is most often asymptomatic, but oral and genital lesions are common manifestations of infection. However, HSV-1 can also cause herpes stromal keratitis (HSK) and can infect the eyelids, conjunctiva, cornea, uveal tract and retina. As with oral and genital infection, HSK can occur repeatedly and cause progressive corneal scar formation.

Following primary infection of the corneal epithelium, the virus enters corneal neurones and migrates to the trigeminal ganglia, which provides sensory innervation to the cornea (see Ch. 1, p. 14). HSV-1 then enter a latent state where viral DNA is present in neurones but no infectious virus is produced. Following exposure to ultraviolet light, the virus can be reactivated and axonally transported into the corneal epithelium, which is highly innervated. Immune suppression also leads to reactivation, indicating an essential role of the host immune response to maintain latency (discussed below).

Primary infection. TLR9 expression on corneal epithelial cells is important in the initial activation by HSV-1 infections and can be induced by HSV-1 DNA alone. These cells produce type 1 interferons (IFN-α/β), which inhibit viral replication. Natural killer cells are also recruited to the corneal stroma and produce IFN-γ and tumour necrosis factor α (TNF-α), which activate macrophages. $CD4^+$ Th1, Th17 and T regulatory (Treg) cells also play an important role in limiting the primary response, with the suppressive activity of Tregs balancing the pro-inflammatory and pro-angiogenic activity of Th17 cells.

Angiogenesis and lymphangiogenesis. Blood and lymph vessel formation of the normally avascular cornea is a characteristic feature of HSK, and is important to initiate an adaptive immune response to the virus as lymphatic vessels transport viral antigens to draining nodes, and blood vessels transport mature HSV-1-specific T cells to the cornea. However, angiogenesis also impairs visual acuity, and neovascularization is tightly regulated by selective production of vascular endothelial cell growth factors and receptor antagonists.

Latency. In the course of primary corneal infection, the virus enters the axons of sensory neurones and is transported in a retrograde manner to the cell bodies, which are located in the trigeminal ganglia, where viral DNA is inserted into the nucleus. This episomal stage of the virus is maintained by expression of HSV-1 latency-associated transcripts. HSV-1 specific $CD8^+$ T cells are found in close association with infected neurones and can form immunological synapses in which the T-cell receptor and CD8 molecules are in the same proximity as class I receptors on the neurones (Fig. 8-7). Although IFN-γ is important, release of perforin in CD8 cell lytic granules is essential to maintain latency, which occurs without killing the neurones.

Viral retinitis

Viral retinitis is an important cause of blindness and visual impairment, especially in AIDS patients. The most common causes of retinitis include herpes simplex virus, varicella zoster virus and human cytomegalovirus (CMV). Less common but significant causes include the lymphocytic choriomeningitis virus (LCMV), Epstein–Barr virus, rubella, measles, West Nile virus, dengue and the chikungunya virus, which is considered an emerging pathogen.

CMV is a double-stranded DNA virus in the herpes family that infects an estimated 40–70% of the world's population. Primary CMV infection is controlled by CD8$^+$ T cells, and as with all herpes viruses, the virus remains persistent over the lifetime of the host. CMV is mostly pathogenic following congenital infection, and can cause chorioretinitis, hydrocephalus and microcephaly. LCMV is a single-stranded RNA arenavirus, and humans are infected

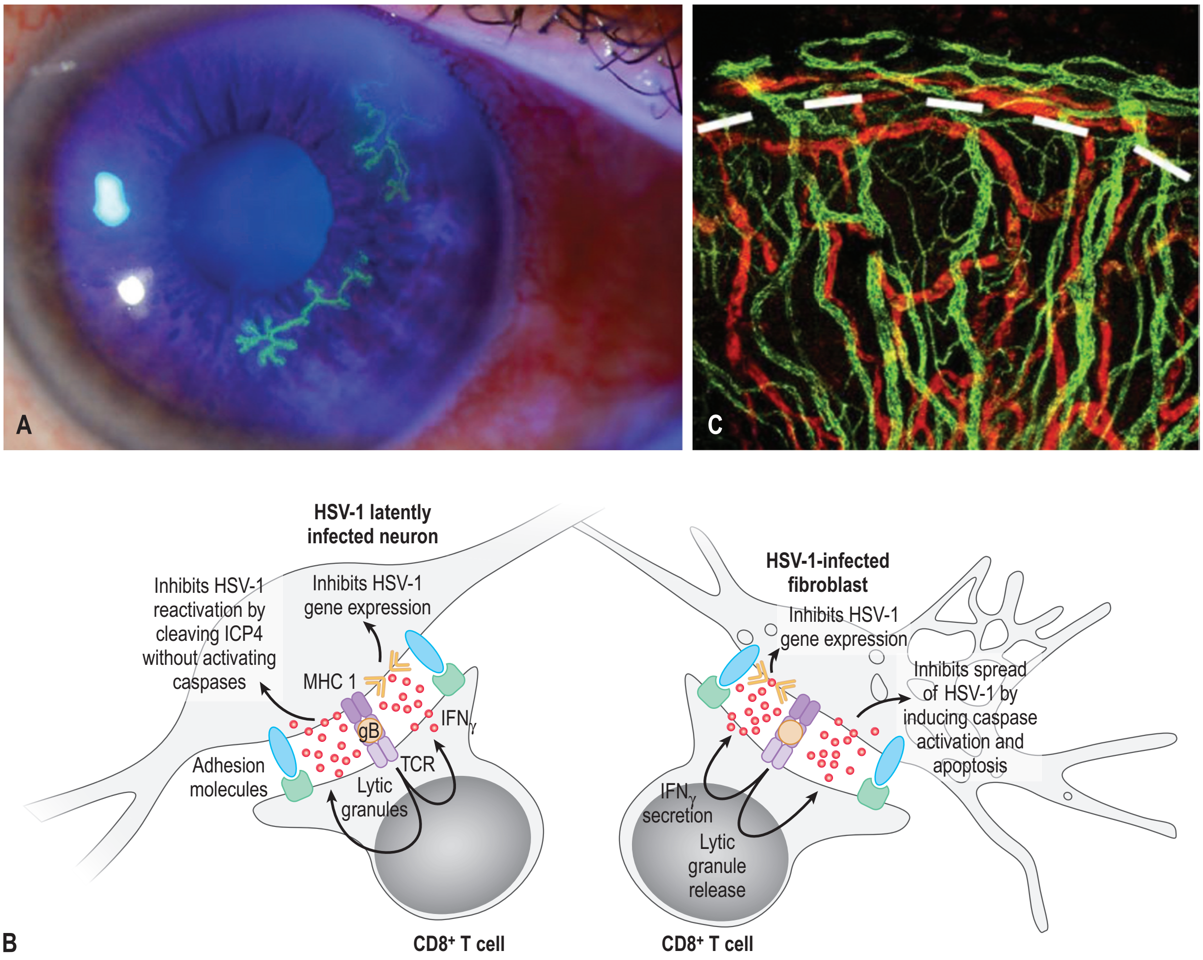
A
C
HSV-1 latently infected neuron
Inhibits HSV-1 reactivation by cleaving ICP4 without activating caspases
Inhibits HSV-1 gene expression
MHC 1
gB
IFN$_{\gamma}$
TCR
Adhesion molecules
Lytic granules
HSV-1-infected fibroblast
Inhibits HSV-1 gene expression
Inhibits spread of HSV-1 by inducing caspase activation and apoptosis
IFN$_{\gamma}$ secretion
Lytic granule release
B
CD8$^+$ T cell
CD8$^+$ T cell

FIGURE 8-7 Herpes simplex keratitis. (**A**) Herpes simplex keratitis (HSK) showing characteristic dendritic-shaped lesion. HSK is also characterized by pronounced corneal angiogenesis. (**B**) HSV-specific CD8$^+$ T cells, interactions with HSV-1 latently infected neurones and lytically infected fibroblasts. IFN-γ and lytic granules inhibit HSV-1 reactivation from latency through a non-lytic mechanism; therefore virus emerges only when T-cell function is compromised. However, lytic granules induce apoptotic cell death of infected fibroblasts. (From Hendricks, Science 2008.). (**C**) Confocal image of representative cornea from animal latently infected with HSV-1 showing LYVE-1+ lymphatic vessels (red) and CD31+ blood vessels (green) at day 30 post-infection. Limbus is above and cornea is below the dashed line; Bar, 500 μm. (From Wuest and Carr, J Exp Med 2010)

by inhalation or ingestion of particles contaminated with mouse faeces, urine or saliva. Experimental models of CMV and LCMV retinitis in which the virus is injected into the vitreous show rapid dissemination to the lymph nodes and spleen, production of virus-specific T cells and invasion of the retina by cytotoxic T cells, which can cause significant retinal disease.

HIV patients are at higher risk for CMV, and the advent of highly active antiretroviral therapy (HAART) has reduced the incidence of CMV retinitis; however, a long-term study of CMV retinitis patients with AIDS taking HAART showed that these patients still have an increased risk for progression of retinitis with complications and visual loss.

BACTERIAL INFECTIONS OF THE EYE

Ocular surface

Bacteria causing conjunctivitis and blepharitis include *Streptococcus pneumoniae* and *Staphylococcus aureus*. These infections cause inflammation and irritation but do not affect vision and are relatively easily treated. In contrast, bacterial infections of the cornea are a major cause of blindness and visual impairment worldwide. Once bacteria gain entry to the corneal stroma they can replicate quickly and induce a pronounced neutrophil infiltrate that contributes to vision loss.

***Pseudomonas aeruginosa* virulence factors.** *P. aeruginosa* is the most common cause of bacterial keratitis worldwide. These Gram-negative bacteria are ubiquitous in fresh water, and contact lens wear and corneal injury are the major risk factors. Although *P. aeruginosa* induces a rapid host response in the cornea, most clinical isolates express virulence factors that counter this response and facilitate bacterial survival. Although these bacteria have a broad spectrum of virulence factors, *P. aeruginosa* produces exotoxins as part of the type III secretion system (T3SS) where they play an essential role in bacterial survival *in vivo*. As shown in Figure 8-8, the T3SS includes a needle-like organelle that injects exotoxins directly into host cells. The needle structure protrudes from the surface of the bacterium and forms a pore in the host cell membrane through which ExoS, ExoT or ExoU are injected. ExoU causes rapid cell lysis and severe corneal disease due to its phospholipase activity, whereas ExoS and ExoT are related proteins that have ADP ribosyl transferase and GTPase activating protein (GAP) enzymatic activity that inhibit cell migration and phagocytosis by blocking cytoskeleton activity. Animal models show that expression of specific T3SS exotoxins and the needle structure is essential for bacterial survival and replication in the cornea. Although almost all clinical and environmental isolates express either ExoS or ExoU, they do not produce both exotoxins. Strains expressing ExoU but not ExoS have been designated cytotoxic strains, whereas ExoS-expressing strains are considered invasive as they can replicate in corneal epithelial cells without causing rapid lysis. ExoS and ExoT faciliate bacterial survival in neutrophils by inhibiting production of reactive oxygen species.

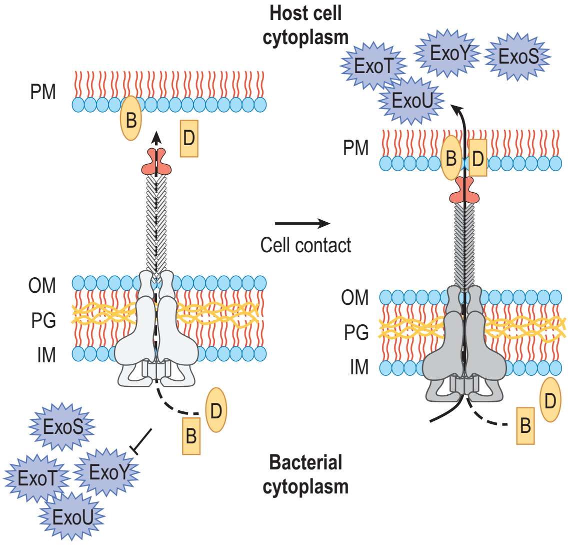

FIGURE 8-8 *Pseudomonas aeruginosa* type III secretion system (T3SS). The T3SS spans the bacterial cell envelope including the inner membrane, (IM), outer membrane (OM) and peptidoglycan layer (PG). Following contact with the host cell plasma membrane (PM), the translocator proteins PopB (B) and PopD (D) form a pore-forming complex with the needle tip (PcrV, red), and the effector proteins are exported through the bacterial cell envelope and the needle complex, and into the cytosol of the targeted cell. To date, four effector proteins have been described in *P. aeruginosa*, ExoS, ExoT, ExoY, and ExoU, although the exact complement of effectors varies among strains and clinical and environmental isolates. Effector proteins are either directly cytolytic (ExoU) or modulate cell function to facilitate bacterial survival in the host cell. (Diagram by A. Rietch, from Pearlman et al., 2013.)

The host response to *P. aeruginosa* in the cornea has been extensively studied using murine models, and although adaptive immunity can be demonstrated at later stages of infection, innate immune responses are essential to regulate the rapid growth of these bacteria in the cornea. Experimental evidence supports the concept that *P. aeruginosa* LPS and flagellin activate TLR4 and TLR5 on resident macrophages, which produce CXC chemokines and pro-inflammatory cytokines that mediate neutrophil recruitment from limbal vessels to the corneal stroma. Inhibition or blockade of TLR4/5, intracellular signalling pathways, IL-1R, or neutralization of IL-1α/β or CXC chemokines impairs neutrophil infiltration and function, allowing the bacteria to survive, which results in more severe corneal disease. Expression of these mediators is also elevated in corneal exudates from infected individuals. Suppressor molecules such as members of the TLR family, neuropeptides such as vasoactive intestinal protein (VIP) and even high concentrations of flagellin also have a suppressive effect on the host inflammatory response, although the mechanisms have yet to be fully characterized. Ultimately, a therapeutic approach involving targeted anti-inflammatory rather than corticosteroids may have potential application when given together with antibiotics.

***Streptococcus pneumoniae*.** *Streptococcus pneumoniae* is a common cause of bacterial keratitis, especially in developing countries. As a normal commensal in the lungs and upper respiratory tract, *S. pneumoniae* is frequently isolated from the conjunctival sac of healthy individuals, and corneal infection is likely associated with traumatic injury. *S. pneumoniae* produces several exotoxins; however, the most potent is pneumolysin, which can directly form pores in the host cell membrane, leading to rapid leakage and cell death. Pneumolysin is released as a monomer but forms a multimer on the host cell membrane, leading to pore formation and rapid cell death. Although some serotypes are more commonly found in clinical isolates, there is no apparent correlation between the serotype expressed and severity of infection.

***Staphylococcus aureus*.** *Staphylococcus aureus* causes less severe disease than *Streptococcus* or *Pseudomonas*, and is associated with peripheral ulcer formation. However, antibiotic-resistant isolates, including meticillin-resistant *Staphylococcus aureus* (MRSA), make treating this infection more challenging. *S. aureus* also produces several virulence factors, including haemolysins α, β and γ, which facilitate bacterial survival *in vivo*.

Contact lens-related corneal inflammation. Other organisms, including *Serratia marcescens* and coagulase-negative staphylococci such as *S. epidermidis*, can cause a mild corneal inflammation associated with contact lens wear, although secretion of the serralysin toxin may also contribute to disease severity. In many cases of contact lens-related inflammation such as contact lens-associated red eye (CLARE) and contact lens peripheral ulcers (CLPU) bacteria cannot be cultured, although corneal infiltrates that are most likely neutrophil-rich are present. As CLARE and CLPU can be replicated in experimental models using TLR agonists such as LPS, it is likely that products of dead and degenerating bacteria activate TLRs on resident corneal cells, leading to neutrophil recruitment to the stroma.

Bacterial endophthalmitis

Bacterial endophthalmitis is characterized by a pronounced inflammatory response in the vitreous, resulting in extensive tissue damage and vision loss. Bacteria can invade the posterior segment of the eye from exogenous infection, following ocular surgery (most commonly cataract) or after traumatic or penetrating injury. Endophthalmitis can also be initiated haematogenously via the retinal vasculature from infection elsewhere in the body. Recent studies also indicate that diabetes mellitus can increase the likelihood of bacterial endophthalmitis due to increased permeability of the blood–retinal barrier.

Coagulase-negative staphylococci including *S. epidermidis* cause the majority of postoperative endophthalmitis, although streptococci, *S. aureus* and enterococci such as *Enterococcus faecalis* and *Propionibacterium acnes* can also cause infection. Staphylococci, streptococci, *E. coli* and *Klebsiella pneumoniae* are among the most common causes of exogenous endophthalmitis, whereas *Bacillus cereus* is more likely to be isolated from cases of post-traumatic endophthalmitis. Animal models show studies indicate that bacterial cell wall components activate TLRs on Müller cells, microglia and macrophages in the neural retina and the inner retinal layer, in addition to hyalocytes in the vitreous. Although most infections respond to antibiotics, retinal damage and visual impairment occur as a result of bacterial toxin production and cellular infiltration.

FUNGAL INFECTIONS OF THE EYE

Yeast and moulds can infect the cornea, causing pronounced long-lasting infections that are very difficult to treat. *Candida albicans* is the most common yeast causing corneal infections, most often following cataract surgery or corneal and endothelial transplants. *Candida* is also associated with therapeutic contact lenses, steroid use or immunosuppressive disease and corneal surgery. In these dimorphic organisms it is the yeast stage that infects the cornea and then germinates to form pseudohyphae in the corneal stroma.

However, worldwide, it is moulds rather than yeast that are the predominant cause of fungal keratitis, with ocular trauma being the major risk factor. As moulds such as *Aspergillus* and *Fusarium* species are plant saprophytes and grow abundantly on crops, the highest incidence and prevalence of fungal keratitis is in agricultural regions, especially during harvest season when there is an abundance of airborne spores. On a global scale, fungal keratitis accounts for ~65% of all corneal ulcers and fungi are more common than bacteria as the cause of ocular infections. *Fusarium solani* and *F. oxysporum* are the main aetiological agents of fungal keratitis followed by *Aspergillus species* (*A. flavus, A. fumigatus*), while *Curvularia, Alternaria* and *Penicillium* species are less common causes.

Corneal infection with *Aspergillus* and *Fusarium*

Spores, or conidia, are produced in large numbers on conidiophores attached by stalks to live or decaying plants (Fig. 8-9), Conidia are dispersed by wind, and corneal injury by airborne plant material or by insects carrying plant material can introduce multiple spores into the corneal stroma (in rural India, whiplash injury by a cow tail is a not uncommon cause of infection). Once in the corneal stroma, conidia germinate and hyphae can penetrate throughout the corneal stroma. The hyphal tips contain multiple proteases, including collagenases that facilitate migration of the hyphae throughout the corneal stroma. If not treated in time, the hyphae will also penetrate into the anterior chamber, where they stimulate a pronounced neutrophil infiltrate seen clinically as a hypopyon, which limits further penetration to the posterior segment. Although rare, fungal endophthalmitis is very difficult to treat and often requires removal of the entire globe.

Contact lens-associated fungal keratitis. In contrast to rural agricultural areas, contact lens wear is the primary risk factor in industrialized countries. The impact of this disease on the industrialized world was illustrated by several hundred cases in the USA, Western Europe and Singapore during the 2005/2006 outbreak of contact lens-associated fungal keratitis. *Fusarium solani* and *F. oxysporum* were the cause of this outbreak, which was resolved in large part by full-thickness keratoplasty, and occasionally by removal of the entire globe. In contrast to trauma-induced fungal keratitis in which conidia directly cause infection, contact lens-associated fungal keratitis is more likely initiated by hyphae. Lenses and lens cases tested during this outbreak not only carried fungal spores but also contained *Fusarium* in the form of biofilms. Fungal biofilms are complexes of hyphae and extracellular carbohydrates that have increased resistance to antimicrobial agents and are an important cause of infection associated with implants and catheters. Following contact with the ocular surface, hyphae in the biofilm can penetrate into the cornea stroma through minor epithelial abrasions where they establish infection.

Treatment of fungal keratitis. In contrast to antibiotics used to treat bacterial keratitis, antimycotic agents are much less effective as hyphae can penetrate into the deeper stroma, thereby limiting accessibility. Topical natamycin or voriconazole are effective if given early after infection, but in general fungal keratitis is notoriously difficult to treat, especially after the hyphae penetrate deeper stromal layers. Therefore, full-thickness keratoplasty is required; however, transplant failure is not uncommon if the fungus has not been completely cleared (high-risk grafts, see Ch. 4), Even in milder cases that respond to antifungal treatment, resolution is associated with fibrosis, resulting in corneal opacity and visual impairment.

Pathogenesis of fungal keratitis. Although *Aspergillus* and *Fusarium* produce mycotoxins that are important in plant disease, there is no direct evidence that they contribute to the pathogenesis of keratitis. Most

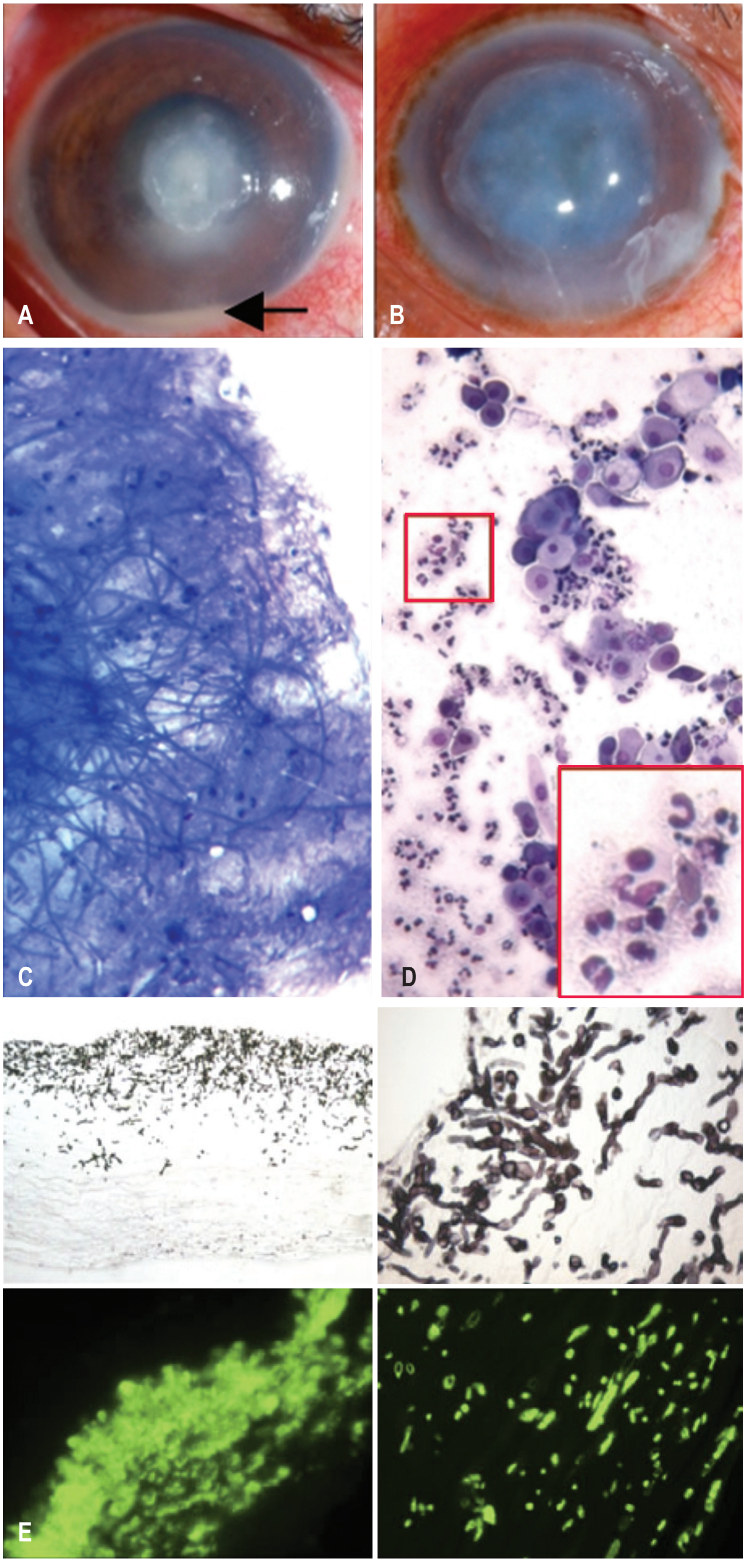

FIGURE 8-9 Fungal keratitis caused by *Fusarium* and *Aspergillus* moulds. *Aspergillus flavus* (**A**) and *Fusarium solani* (**B**) conidiophores, which contain multiple conidia (spores) on their stalks, can penetrate the corneal stroma following ocular trauma with plant material. Conidia germinate and form hyphae that can be detected in corneal ulcer smears after Giemsa staining (**C**) *Aspergillus*. Neutrophils are the predominant cell type in corneal ulcers (**D**). Hyphae penetrate the corneal stroma and express the cell wall component β-glucan (**E**) upper panels: silver stain of *Aspergillus*-infected cornea post-transplant. Lower panels stained with antibody to β-glucan. (Parts C–E from Karthikeyan et al., 2011.)

evidence points to the role of the inflammatory response as a major factor in causing tissue damage. The host response to fungal pathogens involves recognition of cell wall components, including β-glucan and α-mannose, which bind to the C-type lectins Dectin-1 and Dectin-2 on the host cell membrane (β-glucan also binds to CD18 on neutrophils). Figure 8-9 shows β-glucan expression in an infected post-transplant cornea. Activation leads, through a distinct signalling pathway, to the production of pro-inflammatory and chemotactic cytokines. However, these responses also activate IFN-γ and IL-17 producing T cells (Th1 and Th17 cells), in addition to IL-17 production by neutrophils. IL-17 activates receptors on epithelial cells and fibroblasts to produce cytokines and exacerbates the host response.

Corticosteroids. Treatment options for inflammation are limited to corticosteroids and non-steroidal anti-inflammatory agents such as ciclosporin. Steroid use is particularly risky, Because if given before the infection is controlled, there will be unrestricted hyphal growth and tissue destruction. Thus, it will be important to develop more targeted and effective anti-inflammatory agents. Other approaches will be to target essential pathways of survival such as antioxidant and iron scavenging pathways, which are effective in controlling disease in animal models of disease.

PROTOZOAN CAUSES OF OCULAR DISEASE

Acanthamoeba keratitis

Acanthamoeba castellani and other species is an emerging cause of corneal infections, causing severe pain and visual impairment and forming a characteristic lesion (Fig. 8-10). Infection is often associated with contact lens wear and occurs when the cornea is exposed to the relatively high concentration of free-living amoebae that are present in water supplies. The increased incidence of keratitis was noted in relation to changes in water treatment such as in an outbreak in Chicago in 2007. The US Centers for Disease Control and Prevention (CDC) continues to monitor cases of *Acanthamoeba* keratitis.

Acanthamoeba is a free-living protozoan that exhibits a biphasic life cycle as a vegetative trophozoite and a physiologically inert cyst form. Trophozoites are the active stage that exist as free-living amoebae and feed on bacteria and fungi. As noted below, some species of ingested bacteria survive, notably *Legionella* species, and replicate in the trophozoites, which can behave as 'Trojan horses' when they infect the cornea, releasing bacteria into the tissue. Under adverse physiological conditions, the trophozoites encyst and the cyst wall is extremely resistant to environmental agents, irradiation and the host immune system. These infections are therefore extremely difficult to treat and require the use of highly cytotoxic drugs such as biguanides and chlorhexidine (propamidine isothionate, chlorhexidine digluconate and polyhexamethyl biguanide) given over months. These drugs are extremely

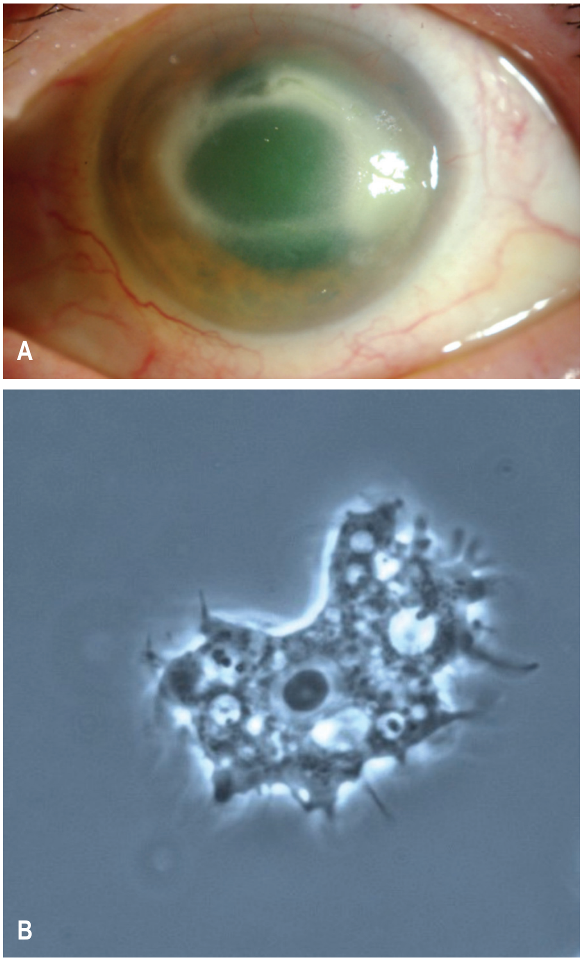

FIGURE 8-10 (**A**) *Acanthamoeba* keratitis showing characteristic radial neuropathy. (**B**) *Acanthamoeba* trophozoite. (Part A from Tu and Joslin, 2012.)

cytotoxic, and can cause severe ocular surface damage. Also, they are not always effective and the patients still require corneal transplantation. *Acanthamoeba* keratitis is most prevalent among contact lens wearers, although it can also be introduced into the corneal stroma by ocular trauma. *Acanthamoeba* has been recovered from soil, air, chlorinated swimming pools, hot tubs, tap water and contact lens solutions. Wearing contact lenses when showering or swimming in fresh water ponds increases the risk of infection. Although the underlying conditions for infection are not known, it likely involves a combination of having trophozoites trapped under the contact lens.

***Acanthamoeba* virulence factors.** The first step in the pathogenesis of infection is adhesion to the ocular surface, which is mediated by carbohydrate–lectin interactions. *Acanthamoeba* expresses a 400 kDa transmembrane mannose-binding protein (MBP) composed of 130 kDa subunits and has characteristics typical of a cell surface receptor. MBP is associated with virulence as it is expressed at higher levels by pathogenic strains, and immunization of experimental animals results in protection from infection. A second virulence factor is the mannose-induced protein (MIP) 133, which is a serine protease with collagenase activity and cytopathic effect on corneal epithelial cells.

Host response to *Acanthamoeba*. There are very few studies on the host response to this pathogen; serum antibodies to *Acanthamoeba* antigens are commonly detected in uninfected individuals, which is consistent with environmental exposure by inhalation or ingestion. In animal models IgA in the tears is protective and trophozoites can be killed *in vitro* by neutrophils or macrophages. Both cell types have been detected by analysis of post-transplant corneas; however, as these corneas are examined only after long-term treatment prior to keratoplasty, it is difficult to determine which cells are killing the trophozoites during infection. It seems reasonable to postulate that activation of pathogen recognition receptors stimulates infiltration of these cells to the cornea. These have not been identified, but could be a result of collagen breakdown products such as Pro-Gly-Pro tripeptides, which activate neutrophil chemokine receptors directly. In addition, release of endogenous bacteria could trigger this pathway and induce an inflammatory response. This concept does not address the chronic infection seen in this disease, which most likely relates to the inability to kill *Acanthamoeba* cysts.

***Acanthamoeba* as a 'Trojan horse'.** As some bacteria are not killed by *Acanthamoeba*, trophozoites inadvertently act as a host of intracellular bacteria that are also associated with human disease. For example, *Acanthamoeba* is a natural host of *Legionella spp.* (mostly *L. pneumophila*), which is associated with legionnaires disease; however, *Acanthamoeba* keratitis clinical isolates also harbour *Pseudomonas* (mostly *P. aeruginosa*), *mycobacteria* and *Chlamydia*, and *L. pneumophila and P. aeruginosa* survival in *Acanthamoeba* is dependent on expression of bacterial virulence factors.

Toxoplasmosis

Toxoplasma gondii is the most common cause of infectious retinochoroiditis worldwide, causing up to 50% of all posterior uveitis cases in some countries in the industrialized world, with higher levels in developing countries. *T. gondii* is an obligate intracellular protozoan parasite and is among the most ubiquitous pathogens worldwide, with >70% of the global population seropositive. This parasite has been found in almost all mammals and can cause infections in domesticated farm animals (especially pigs) and in pets. Cats are the definitive hosts and are the only mammals that support the sexual stage of this parasite (Fig. 8-11). Transmission therefore occurs by ingesting the oocysts in food that has been contaminated with cat faeces, although individuals can also be infected following ingestion of tissue cysts (bradyzoites) in undercooked meat or by maternal transmission during pregnancy (often in the third trimester). In humans and other intermediate hosts, the tachyzoite form can invade and undergo rapid asexual replication in any nucleated cell. Following oral infection, this occurs in enterocytes, where they rapidly replicate and cause cell death. *T. fondii* can cross the epithelial barrier to infect circulating monocytes, and are disseminated to mutiple organs including the brain and the retina. Despite the presence of a functional innate immune response, the parasites evade the immune system in these immune-privileged sites but remain semi-dormant as viable bradyzoites within sequestered tissue cysts. However,

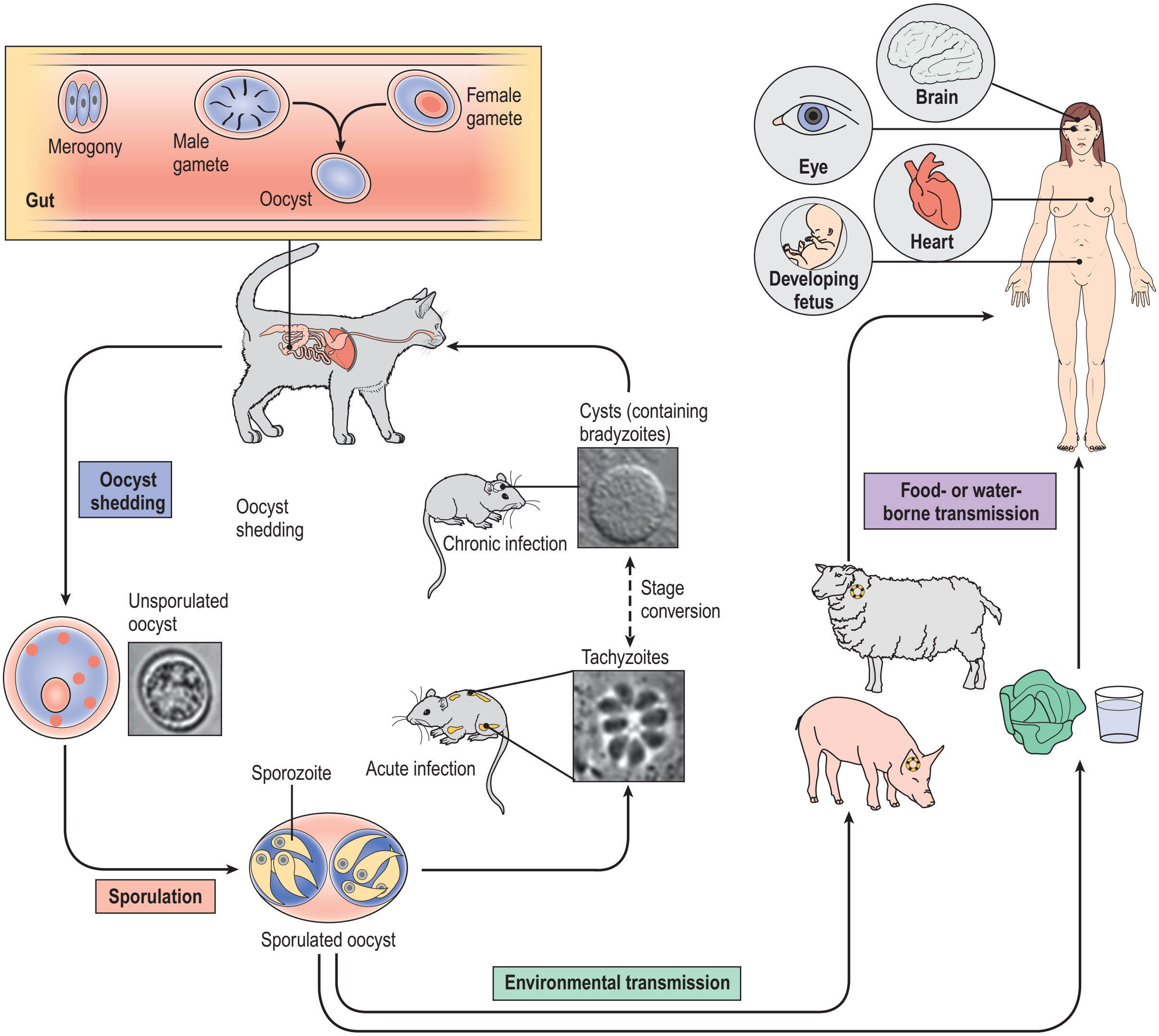
Merogony
Male gamete
Female gamete
Oocyst
Gut
Oocyst shedding
Oocyst shedding
Cysts (containing bradyzoites)
Chronic infection
Stage conversion
Tachyzoites
Unsporulated oocyst
Acute infection
Sporozoite
Sporulation
Sporulated oocyst
Environmental transmission
Brain
Eye
Heart
Developing fetus
Food- or water-borne transmission

FIGURE 8-11 The complex life cycle of *Toxoplasma gondii*. Humans are accidental hosts and represent a dead end to continuing the life cycle. Cats are the only definitive host permissive for sexual replication in which male and female gametes are formed in the cat enterocytes. After fusion of the *T. gondii* gametes to form oocytes, these survive in the environment and are ingested by intermediate hosts, including rodents, sheep, pigs and humans. Fast-replicating tachyzoites and slow-replicating bradyzoites are generated in epithelial cells, and can cause inflammation and tissue damage, including in the eye. (From Hunter and Sibley, 2011.)

in immune compromised individuals the parasites can again replicate and cause multifocal brain and retinal lesions.

Host response. Most of our current understanding of the control of *T. gondii* infection comes from murine models. *T. gondii* proteins activate TLR2, TLR4, and TLR11 on monocytes and dendritic cells to produce IL-12, which stimulate natural killer (NK) cells to produce IFN-γ, resulting in activation of CD4$^+$ and CD8$^+$ T cells. However, there is also an IFN-γ-independent mechanism of parasite killing which is mediated by CD40–CD40L-dependent autophagy (see Ch. 7).

Ocular toxoplasmosis. *T. gondii* induces an inflammatory response in the uveal tract and in the retina, causing uveitis and retinochoroiditis, although specific mediators of inflammation have yet to be identified. *In vitro*, the parasites can replicate in Müller cells, retinal epithelial cells and retinal endothelial cells. These cells then produce pro-inflammatory and chemotactic cytokines that mediate recruitment of macrophages and lymphocytes to the retina. Animal models of toxoplasmosis have shown that depletion of both CD4$^+$ and CD8$^+$ T cell subsets, but not either one alone, results in recurrence of the disease, implying that both cell types are required.

In immunocompetent individuals, *Toxoplasma* uveitis is self-limiting; however, recurrent lesions are associated with development of retinochoroiditis and loss of vision. It is likely that reactivation of tissue cysts is the basis for recurrence, with the stimulus likely to lead to the failure of the host response to contain the infection.

HELMINTH CAUSES OF OCULAR INFECTION

Helminths are a rare but well recognized cause of ocular disease, the most ubiquitous being toxocariasis. Other less well recognized or unidentified helminths may cause an unusual retinal infection known as diffuse unilateral subacute neuroretinitis in which the worms are found in the fundus and in the subretinal space, causing a pigmentary retinopathy and low-grade retinal inflammation with progressive loss of visual acuity.

TOXOCARIASIS

Toxocara canis and *Toxocara cati* are nematodes transmitted by dogs and cats, respectively, where the adult worms are found in the intestines. Eggs are secreted in the faeces and are ingested by the secondary hosts, which include pigs and rodents. Larvae migrate through the intestine and can be carried to any part of the body via the bloodstream. The cycle is completed when the dog or cat eats the secondary host. Human infection occurs following ingestion of contaminated food, although humans are a dead end in terms of the worm's life cycle. However, the larvae still migrate through the intestine and are disseminated, to multiple organs including the eye.

Toxocara that appears as a large inferior white intraocular vitreoretinal mass, although the live worm can also migrate to the anterior chamber. Treatment of ocular disease is with oral tiabendazole or albendazole.

Ocular infections in developing countries

ONCHOCERCIASIS (RIVER BLINDNESS)

The filarial nematode *Onchocerca volvulus* is the causative organism of onchocerciasis, also known as river blindness, which has caused devastating blindness throughout sub-Saharan Africa, Central and South America, and in the Arabian peninsula. The 'river' connection relates to the *Simulium* blackfly vector that breeds in fast-flowing rivers. Although extensive control efforts have been underway since the 1970s, several foci of disease remain, primarily in West and Central Africa. Blindness occurs as a result of anterior and posterior segment ocular inflammation, and manifestations include sclerosing keratitis, chorioretinitis and optic neuritis.

Life cycle and parasite burden

First-stage larvae (microfilariae, L1) in infected individuals are ingested during the female blackfly's blood meal, and migrate through the gut and thorax to the salivary gland, having undergone two moults to the third-stage larvae (L3). During the second blood meal, the L3 enter the bloodstream, develop into stage four larvae (L4), and then into adult males and females that can live for over 10 years. Adult worms are sequestered in collagenous nodules and females produce millions of microfilariae during their lifetime and can survive in the skin for up to 30 months; therefore individuals can be very heavily infected. Microfilariae are readily detected in the anterior chamber by slit lamp microscopy. Fundus examination can also detect microfilariae in the retina.

Pathogenesis

Given the large parasite burden in infected individuals, a strong immune response might be expected; however, live microfilariae appear to cause minimal tissue damage either to the skin or the eyes (Fig. 8-12). As with other chronic parasitic infections in which the host and parasite have co-evolved, the infected host generates a predominantly immunosuppressive response that involves T regulatory cells, IL-10 and TGF-β, which facilitates the longevity of adult

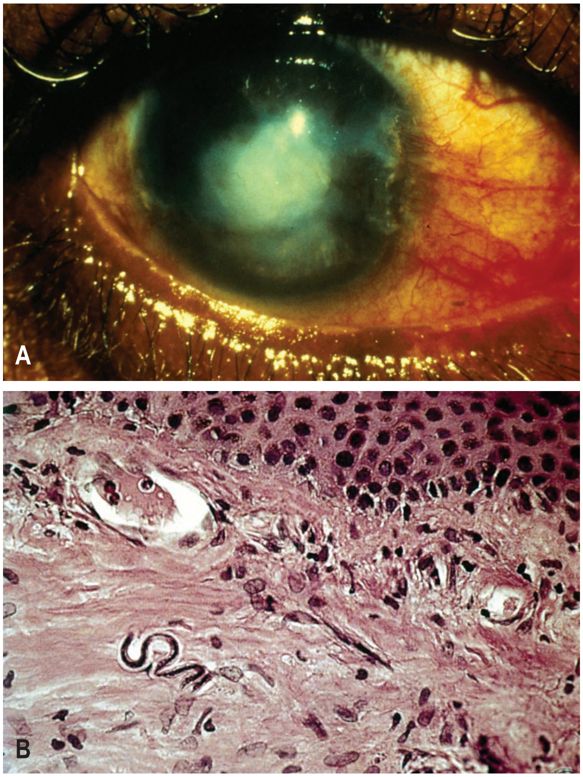

FIGURE 8-12 (**A**) Late stages of ocular onchocerciasis (river blindness) caused by the parasitic nematode *Onchocerca volvulus*. (**B**) A microfilaria in the corneal stroma not associated with an inflammatory response, whereas cellular infiltrates and corneal blood vessels are present at other sites. (From Taylor and Pearlman (Duane's Ophthalmology).)

and larval worms and continuation of the life cycle. In marked contrast to the suppressive response to live *O. volvulus*, when the larvae die either by natural attrition or following chemotherapy, infected individuals experience tissue damage, which manifests as onchodermatitis and stromal keratitis. Epidemiological and experimental data indicate that almost all immunopathology associated with onchocerciasis is directly or indirectly related to the local death of microfilariae.

Endosymbiotic *Wolbachia* bacteria

Although immunopathology has long been associated with dead and degenerating microfilariae, the finding that *O. volvulus* and other filariae harbour intracellular *Rickettsia*-like bacteria led to an increased understanding of the disease. *Wolbachia* are ubiquitous in the ovaries of insects and other arthropods, where they are transmitted transovarially. As filariae have obligate developmental stages in insects (*O. volvulus* in blackflies, *Wuchereria* and *Briugia* species that cause lymphatic filariasis in mosquitoes) this likely explains why they are the only group of nematodes that harbour *Wolbachia*. Using antibodies to the major *Wolbachia* surface protein (WSP), *Wolbachia* were found to be abundant in the nematode hypodermis in the gravid uterus (Fig. 8-13) and also in microfilariae. Thus, larvae migrating through the skin and the eye also harbour *Wolbachia*, which are released following microfilaria death. Several lines of evidence support the concept that *Wolbachia* mediate the immunopathology associated with ocular onchocerciasis: (1) *Wolbachia* organisms and DNA are found in the blood of filarial infected individuals following chemotherapy; (2) *O. volvulus* from individuals depleted of *Wolbachia* by antibiotic treatment do not induce corneal inflammation; and (3) isolated *Wolbachia* induce corneal inflammation through TLR2 activation, which is the same as soluble extracts from the whole worms. *Wolbachia* are also detected in microfilariae in the skin of infected individuals. Analysis of the *Wolbachia* genome showed they do not have the LPS synthase gene, and therefore cannot produce LPS. A synthetic *Wolbachia* lipopeptide was able to induce corneal inflammation that is dependent on TLR2 and TLR6. These findings therefore indicate that the early inflammatory response in river blindness is mediated by activation of TLR2/6 by *Wolbachia* lipopeptides.

Onchocerciasis control

In 1974 the World Health Organization launched the Onchocerciasis Control Programme (OCP), which was based on aerial spraying of insecticides combined with chemotherapy. Although diethylcarbamazine had serious side-effects due to rapid death of the worms, ivermectin (Mectizan®) was found to kill the microfilariae (though not the adults). After repeated annual treatments, transmission was significantly reduced in the initial target area in West Africa, and as ivermectin was provided free by Merck, Inc. Onchocerciasis control was expanded as the African Programme for Onchocerciasis Control (APOC) and a similar ivermectin distribution programme was initiated in the Americas and in the Arabian peninsula. These programmes have met with considerable success and improved the lives of millions of individuals.

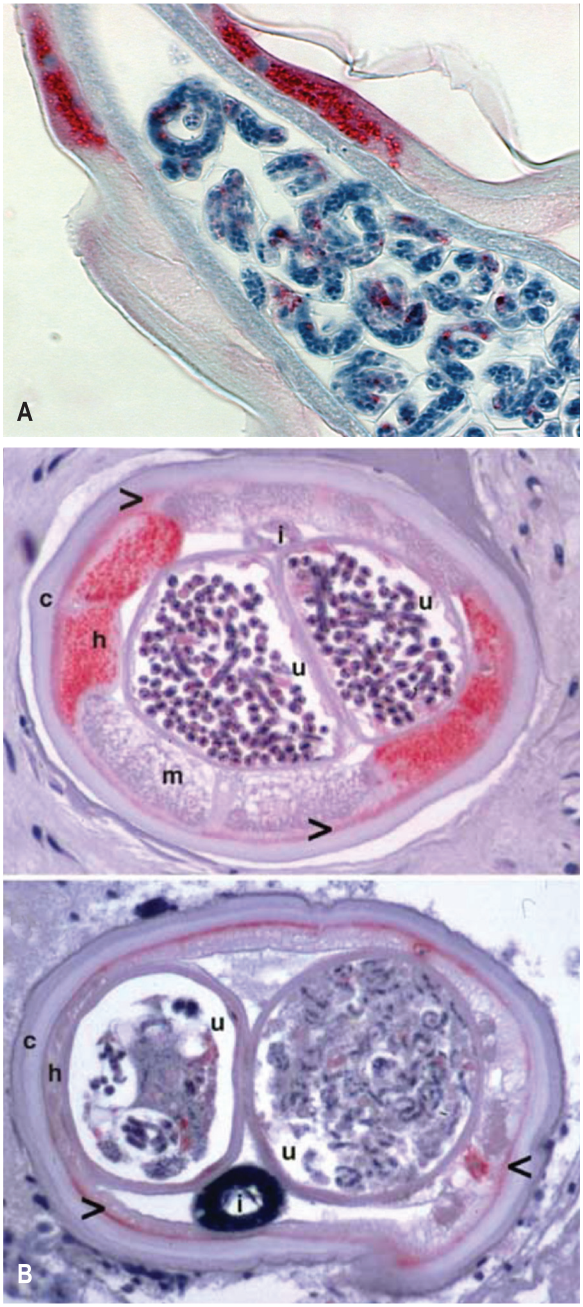

FIGURE 8-13 (**A**) Endosymbiotic *Wolbachia* bacteria in the hypodermis and embryos of an adult female worm (sections immunostained with antibody to the *Wolbachia* surface protein. (**B**) Cross-section of adult female worms following treatment with ivermectin alone (upper panel) or with doxycycline (lower panel). (Part A Photomicrograph by A. Hise, from Pearlman et al., Science 2002. Part B from Hoerauf, Lancet, 2001.)

Remaining challenges are to maintain this level of reduced transmission and to examine the effectiveness of additional therapeutic agents such as albendazole and antibiotics that target *Wolbachia*.

In targeting *Wolbachia*, studies showed that doxycycline and related antibiotics reduced not only the microfilarial load but also the adult worm burden, effectively curing the infection. The caveat has been that in contrast to ivermectin, which is given annually, doxycycline requires multiple doses over several weeks. The challenge remains to discover more potent antibiotics that can be administered *en masse*. The Bill and Melinda Gates Foundation has invested in this approach to develop new antibiotics (the Anti-Wolbachia Consortium, A-Wol.com), and results from this project are likely to have important implications for this disease.

LOA LOA (LOIASIS, EYEWORM)

Loa loa is a filarial nematode with a similar life cycle to *O. volvulus* except the adults migrate through subcutaneous tissues and the microfilariae are in the blood. The intermediate vectors are large *Chrysops* flies rather than the skin and transmission is limited to regions of western Africa. Disease manifestations are due to the migrating adults that can live for up to 17 years in a human host and can move through subcutaneous tissues at an estimated 1 cm per minute. Adult worms also migrate through the conjunctiva (not the cornea), where they can be surgically removed (Fig. 8-14). Localized inflammation, known as Calabar swellings, may occur in subcutaneous tissues of the skin but they do not cause major tissue damage. *Loa loa* does not harbour *Wolbachia* and infection is treated with corticosteroids and diethylcarbamazine.

TRACHOMA

Trachoma is the most prevalent microbial cause of blindness worldwide, with ~1.3 million people blind and an estimated 1.8 million with impaired vision. Figure 8-15 shows the global distribution of trachoma, which is endemic in some 50 countries, especially in sub-Saharan Africa. In Ethiopia and Sudan an estimated 50% of children and 9% of adults have active trachoma.

Trachoma is caused by *Chlamydia trachomatis*, specifically subgroups (serovars) A, B, Ba and C. These

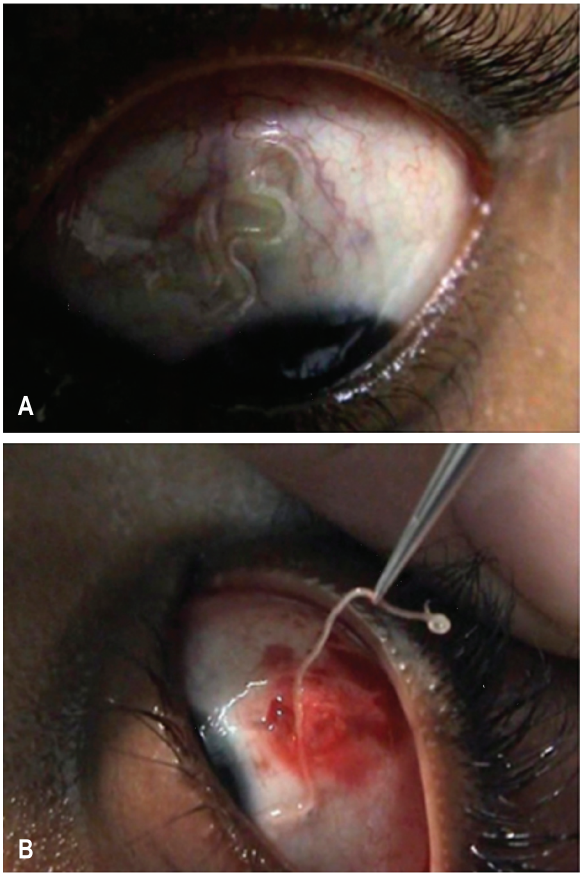

FIGURE 8-14 Loiasis (eyeworm). Adult *Loa loa* migrating across the conjunctiva (**A**), where they can be surgically removed (**B**). (From Shah and Saldana, 2010.)

bacteria replicate in conjunctival epithelial cells as reticulate bodies that do not appear to have a cell wall. They also form spore-like elementary bodies which do have cell walls, and this is the stage that is transmitted, either person to person or by flies that are attracted to the eyes.

The predominant disease manifestations are shown in Figure 8-15. Recurrent infection of the conjunctival epithelium results in inflammation and development of characteristic papillary conjunctivitis, which eventually cause conjunctival scarring. Contraction of scar tissue causes entropion (rolling of the eyelids inward). The eyelashes then scratch the cornea (trichiasis), which can result in corneal opacification and blindness. The inflammatory response has not been well characterized, although animal models show an important role for IFN-γ.

The World Health Organization initiated the Global Alliance for the Elimination of Blinding Trachoma (GET 2020), which is based on the *SAFE* strategy to control transmission by means of *S*urgery for inturned eyelids, *A*ntibiotic use, with at least two oral treatments of azithromycin (Zithromax®, donated by Pfizer), *f*acial cleanliness to prevent person-to-person transmission and fly- (*Musca sorbens*) to-person transmission, and, lastly, *e*nvironmental changes to increase access to clean water and encourage handwashing and the use of latrines, which would also reduce the number of these flies.

Trachoma in Australia

Trachoma has long been prevalent in the indigenous population of Australia, and a 2008 survey reported that children in 60% of outback communities have blinding disease. However, the Australian government has made a major commitment to reduce the prevalence by using the SAFE strategy and making improvements in housing and health care.

Conclusion

In summary, microbial infections of the eye are a major cause of blindness and visual impairment worldwide. Infections can affect the anterior and posterior segments and disease pathogenesis is the net result of both microbial virulence and the host response. Pathogens have evolved multiple strategies for survival in the eye, and the host is not always discriminating in its ability to kill the microbes without causing tissue damage. However, the ocular surface barriers block invasion by environmental microbes under normal physiological conditions, and infection occurs primarily when the ocular surface is compromised by trauma or extended contact lens wear.

Future treatments for these blinding eye infections will be based on an increased understanding of the pathogenesis of disease which will also identify new, more specific targets for immune intervention compared with corticosteroids. Potential strategies include blocking pathogen recognition molecules or neutrophil chemokines which could be given together with effective antibiotics. However, many of these

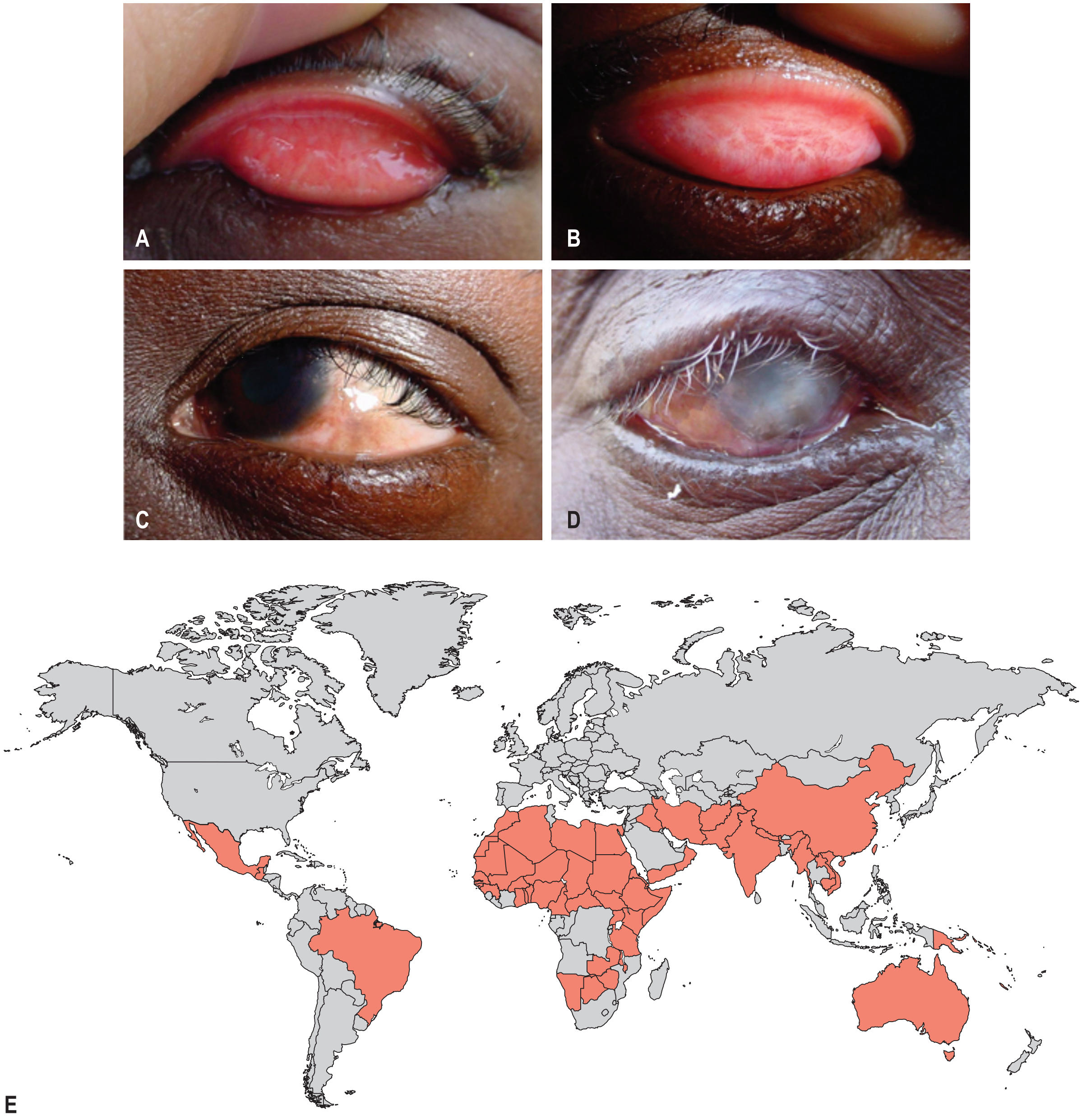

FIGURE 8-15 Clinical features of trachoma: (**A**) active trachoma in a child, characterized by a mixed papillary (TI) and follicular response (TF); (**B**) tarsal conjunctival scarring (TS); (**C**) entropion and trichiasis (TT); and (**D**) blinding corneal opacity with entropion and trichiasis (TT). (**E**) Prevalence of trachoma worldwide. (From Burton and Mabey, 2009.)

microbes, bacteria in particular, have become resistant to antibiotics, and there is a pressing need to develop new classes of antimicrobial agents. To this end, antimicrobial peptides have some potential as they are biological rather than pharmacologic agents, their specificity is well understood and they can be readily synthesized as a native sequence or with substituted amino acids that may improve efficacy. Targeting microbial scavenging for essential metals such as iron and zinc may also prove a feasible approach.

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

FURTHER READING

Bowler, G.S., Shah, A.N., Bye, L.A., Saldana, M., 2011. Ocular loiasis in London 2008–2009: a case series. Eye 25, 389–391.

Burton, M.J., Mabey, D.C., 2009. The global burden of trachoma: a review. PLoS. Negl. Trop. Dis. 3, e460.

Franchi, L., Munoz-Planillo, R., Nunez, G., 2012. Sensing and reacting to microbes through the inflammasomes. Nat. Immunol. 13, 325–332.

Gipson, I.K., 2004. Distribution of mucins at the ocular surface. Exp. Eye Res. 78, 379–388.

Hardison, S.E., Brown, G.D., 2012. C-type lectin receptors orchestrate antifungal immunity. Nat. Immunol. 13, 817–822.

Hunter, C.A., Sibley, L.D., 2012. Modulation of innate immunity by *Toxoplasma gondii* virulence effectors. Nat. Rev. Microbiol. 10, 766–778.

Karthikeyan, R.S., Leal, S.M., Jr., Prajna, N.V., Dharmalingam, K., Geiser, D.M., Pearlman, E., et al., 2011. Expression of innate and adaptive immune mediators in human corneal tissue infected with *Aspergillus* or *Fusarium*. J. Infect. Dis. 204, 942–950.

Knickelbein, J.E., Khanna, K.M., Yee, M.B., Baty, C.J., Kinchington, P.R., Hendricks, R.L., 2008. Noncytotoxic lytic granule-mediated $CD8^+$ T cell inhibition of HSV-1 reactivation from neuronal latency. Science 322, 268–271.

Knickelbein, J.E., Watkins, S.C., McMenamin, P.G., Hendricks, R.L., 2009. Stratification of antigen-presenting cells within the normal cornea. Ophthalmol. Eye Dis. 1, 45–54.

Kolaczkowska, E., Kubes, P., 2013. Neutrophil recruitment and function in health and inflammation. Nat. Rev. Immunol. 13, 159–175.

Leal, S.M., Jr., Pearlman, E., 2012. The role of cytokines and pathogen recognition molecules in fungal keratitis - Insights from human disease and animal models. Cytokine 58, 107–111.

O'Neill, L.A., Golenbock, D., Bowie, A.G., 2013. 2013. The history of Toll-like receptors – redefining innate immunity. Nat. Rev. Immunol. 13, 453–460.

Pearlman, E., Sun, Y., Roy, S., Karmakar, M., Hise, A.G., Szczotka-Flynn, L., et al., 2013. Host defense at the ocular surface. Int. Rev. Immunol. 32, 4–18.

Saint Andre, A., Blackwell, N.M., Hall, L.R., Hoerauf, A., Brattig, N.W., Volkmann, L., et al., 2002. The role of endosymbiotic *Wolbachia* bacteria in the pathogenesis of river blindness. Science 295, 1892–1895.

Shah, A.N., Saldana, M., 2010. Images in clinical medicine. Ocular loiasis. N. Engl. J. Med. 363, e16.

Tu, E.Y., Joslin, C.E., 2012. *Microsporidia* and *Acanthamoeba*: the role of emerging corneal pathogens. Eye (Lond) 26, 222–227.

Wuest, T.R., Carr, D.J., 2010. VEGF-A expression by HSV-1-infected cells drives corneal lymphangiogenesis. J. Exp. Med. 207, 101–115.

Chapter 9

Pathology

- Introduction
- Cell and tissue damage
- Healing and repair
- Inflammation
- Metabolic disease
- Ageing, degeneration and dystrophies
- Neoplasia

Introduction

A knowledge of pathology is the key to good clinical practice, and the deeper this knowledge the better the practice. Less well recognized is that a knowledge of pathology also underpins much of the work of the ophthalmic and vision scientist and offers wide scope for research ideas. This chapter describes basic aspects of disease processes with reference to specific entities relevant in ophthalmology.

Cell and tissue damage

Cells and tissues die as part of normal physiological homeostasis and different tissues have different susceptibility to death. Indeed, different processes occur as cells die and are ultimately cleared away. Similarly, tissues have different capability for regeneration and this applies to all of the particular ocular tissues.

AGENTS OR CONDITIONS THAT CAN CAUSE DIRECT DAMAGE TO OCULAR CELLS AND TISSUES

There are a wide variety of noxious agents, and processes or deficiencies capable of causing damage to the eye and some of these are described here.

Reduced oxygen supply

Central retinal artery occlusion causes reduced oxygen flow to the retina, which becomes opaque due to oedema.

Physical agents

- *Mechanical trauma* is the commonest form of ocular injury. Blunt trauma describes mechanical injury in which the globe remains intact. Anteroposterior deformation separates the delicate attachments between the intraocular structures, e.g. the zonular fibres, rupture of which causes lens dislocation. Separation of the attachment of the ciliary muscle to the scleral spur leads to collapse of the trabecular meshwork and secondary glaucoma (angle recession glaucoma). Penetrating and perforating injuries are caused by a wide variety of weapons, tools, sporting equipment and domestic utensils, and almost inevitably lead to severe disorganization within the ocular structures.
- *Extremes of temperature* such as cryotherapy are used surgically to induce an adhesive scar by freezing and thawing in the peripheral retina during the prophylaxis or treatment of retinal detachment. Cryoablation is also used to destroy the ciliary body in cases of intractable glaucoma, the aim being to suppress aqueous formation.
- *Light exposure* – excess exposure to ultraviolet light may lead to an overproduction of free radicals by photons (e.g. hydrogen peroxide, superoxides) with insufficient free radical scavengers (e.g. vitamin A, superoxide dismutase and glutathione transferase) with resultant tissue damage. For example, damage may occur to the corneal epithelium in snow blindness. Photon bombardment of the photoreceptors at the

macula over six or more decades has been suggested as a contributing factor in age-related macular degeneration. Lasers used in ophthalmology cause tissue damage dependent on the wavelength, the site of absorption and the quantity of energy released. In panretinal photocoagulation, which is used to treat diabetic vasoproliferative retinopathy, individual burns destroy localized patches of the outer third of the retina and the underlying RPE cells. Reactive proliferation of RPE at the edge of the burn leads to clinically visible pigmentation around a white circle, which is the result of a 'scar' formed by glial cells. Transpupillary thermotherapy (TTT) uses infrared light to heat (to 40°C) choroidal melanomas and bring about tumour cell necrosis. Several types of laser are used in ophthalmology to treat various eye diseases, and newer lasers such as the nanolaser are continuing to be developed (see Table 9-1).

- *Ionizing radiation* – each of the three types of ionizing radiation in common use in medical practice has applications in ophthalmology:
  - charged particles (electrons and α-particles)
  - uncharged particles (neutrons)
  - electromagnetic radiation (X-rays and γ-rays).
- Proton beam therapy is now an accepted choice for the external treatment of ocular melanomas, while γ-emitters such as ^{106}Ru and ^{60}Co can be applied topically to the sclera over a melanoma in the form of a radioactive plaque. X-rays and γ-rays are used for the external beam treatment of retinoblastoma.

TABLE 9-1 Lasers used in ophthalmology

Type of laser	Wavelength (nm)	Application
Argon (CW)	Green (457, 488, 514, 610)	Ablation of RPE/ outer retina (in diabetic retinopathy)
YAG/Nd	Pulse 1064	Disruption of lens capsular membranes; destruction of ciliary body (glaucoma)
Excimer (CW)	193	Radial keratometry Remodelling of corneal surface; refractive keratometry

CW, continuous wave; YAG/Nd, yttrium-aluminium-garnet/ neodymium.

The unit (gray; Gy) is a measure of the amount of energy absorbed in the target tissue. For melanomas, a dose of up to 110 Gy is required at the base of the tumour if a ^{106}Ru plaque is applied, while 40–60 Gy external radiation is required for the treatment of a retinoblastoma. With such high doses, cataract is inevitable if the lens is not shielded. Radiotherapy in and around the eye and orbit can have short- and long-term side-effects (Box 9-1).

Chemical agents

Toxic chemicals of any sort will damage the delicate ocular tissue when applied externally. Acid and alkali burns are encountered most commonly in clinical practice, but detergents can also cause significant damage.

- *Acid burns* – acids (hydrochloric, nitric and sulphuric) and acidic fluids coagulate proteins, so that diffusion through the cornea and the sclera into the eye may be limited. Necrosis of epithelial and stromal cells leads to reactive fibrosis in the conjunctiva with eyelid distortion (entropion and ectropion). Corneal scarring requires keratoplasty to restore vision.
- *Alkali burns* – alkalis, such as ammonia and sodium hydroxide, pass through tissue easily and the high pH is sufficient to destroy the cells of the lens, uveal tract and retina as the alkaline fluid diffuses through the vitreous as far as the optic nerve.

Toxins

Toxins are generally proteins derived from bacteria, plants and animals (e.g. snakes). Some have been purified and have value as therapy such as botulinum toxin. However, most toxins from organisms cause severe tissue damage, such as those from *Pseudomonas aeruginosa* which may cause a severe lytic keratitis (see Ch. 8). Tobacco and alcohol abuse may lead to toxic-nutritional optic neuropathy with damage to the papillomacular bundle. Similarly, many drugs used to treat

BOX 9-1 COMPLICATIONS OF IRRADIATION THERAPY: EFFECTS ON CELLS AND TISSUES

Radiation from any source has a profound effect on nuclear DNA, with fracture of chromosomes, dislocation and translocations. The consequence is tissue destruction and suppression of cell division, which is essential for the ablation of tumour tissue. The side-effects for non-tumour tissues are:

- Endarteritis – infiltration of the vessel wall by inflammatory cells and proliferation of spindle cells (myofibroblasts, intimal fibroplasia) within the internal elastic lamina. There is endothelial cell swelling and fibrin deposition. Narrowing of the vessel lumen contributes to tumour destruction by ischaemic necrosis (see figure). A long-term secondary complication of radiation arteritis is dilation of the capillary bed (telangiectasia). Radiation vasculopathy causes leakage of plasma constituents in retinal vessel walls.
- Loss of hair, teeth and glandular tissue. Irradiation of the orbit damages lacrimal gland tissue and leads to a dry eye.
- Massive necrosis of normal tissue when the radiation dose is excessive.
- An increased risk of mutations, with malformation in offspring and the induction of a second malignant tumour.

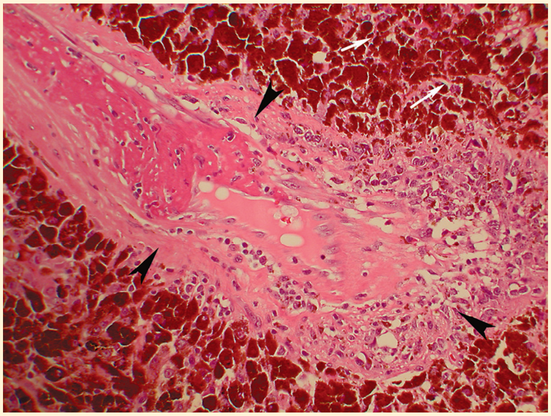

The figure shows an irradiated melanoma with radiation endarteritis (black arrowheads) and surrounding pigment-laden melanophages resulting from tumour necrosis (white arrows).

systemic disease can have toxic side-effects for the retina, e.g. ethambutol for tuberculosis.

Viruses

Viruses that commonly cause ocular damage include herpes simplex virus and herpes zoster virus (see under Inflammation, p. 492).

Dysregulated immunity

Several autoimmune diseases, such as rheumatoid arthritis and multiple sclerosis, affect the eye and will be discussed under Inflammation. Those with reduced immunity due to human immunodeficiency virus may develop retinitis due to cytomegalovirus.

Nutritional deficiencies

Vitamin A deficiency leads to night blindness and later xerophthalmia and keratomalacia. Vitamin B_{12} deficiency is implicated in tobacco/alcohol optic neuropathy.

Genetic abnormalities

It is not surprising that errors occur during complex genetic activities. Germ cells and proliferating somatic cells (including stem cells) are susceptible to such errors. When abnormalities arise in germ cells the abnormality may be transmitted to the next generation. Abnormalities in somatic cells may result in cell death or malignant transformation. Genetic abnormalities may involve whole chromosomes, parts of chromosomes, gene clusters or single genes.

Chromosomal abnormalities

- Polypoidy occurs when chromosomal numbers are increased by an exact multiple of normal. This may be seen in hypertrophied muscle cells and ageing cells but does not usually occur in gametogenesis as it is incompatible with life.
- Aneuploidy occurs when the number of chromosomes is increased, usually by one (trisomy). This may occur in Down syndrome.
- Structural abnormalities may arise during replication and may result in duplication and deletion of gene clusters of single genes.

Single gene defects. These are caused by defects in single genes. They usually affect the structure of the

BOX 9-2 NEUROFIBROMATOSIS TYPE 1 (NF1) AND THE EYE

This is an autosomal dominant condition with almost complete penetrance due to a mutation in the *NF-1* gene on chromosome 17. It is characterized by café-au-lait spots and neurofibromas on the skin. Ocular abnormalities include the following:

- *Eyelids* – café-au-lait spots and neurofibromas can affect the lids. Plexiform neurofibromas may cause cosmetic and functional problems.
- *Orbit* – proptosis may occur because of optic nerve gliomas or orbital plexiform neurofibromas. Bony defects may occur in the orbital floor or roof.
- *Anterior segment* – Lisch nodules, melanocytic hamartomatous lesions of the iris are a defining feature of NF1. Enlarged corneal nerves and conjunctival, episcleral and scleral involvement by neurofibromas may occur.
- *Glaucoma* – the major cause of glaucoma is infiltration of the angle by a neurofibroma obstructing aqueous flow. Goniodysgenesis may also occur.
- *Retina and choroid* – choroidal hamartomas consist of a mixture of melanocytic and neuronal elements. Hamartomas of the retina and retinal pigment epithelium may also occur.
- *Optic nerve* – optic pathway gliomas are the cause of most visual morbidity associated with NF1. They occur in 15% of patients and are usually juvenile pilocytic astrocytomas. They can involve the optic nerve, chiasm, optic tract and hypothalamus.

resulting protein product and can have wide-ranging effects. They can be inherited as an autosomal dominant (e.g. neurofibromatosis type 1; Box 9-2), autosomal recessive (e.g. Wilson's disease (Box 9-3), Tay–Sachs disease) or sex-linked (e.g. some forms of retinitis pigmentosa) disorder. Whilst most genetic disorders occur in nuclear DNA, it should be remembered that mitochondria also contain DNA in chromosomes, which are self-replicating and encode enzymes involved in oxidative phosphorylation. When mutations occur in mitochondrial chromosomes, the defect is passed on via the ovum because the sperm does not have cytoplasmic constituents; thus mitochondrial disorders are maternally transmitted (see Ch. 3, p. 138 and Box 9-4).

BOX 9-3 WILSON'S DISEASE

This is an autosomal recessive conditions due to mutations in the Wilson's disease protein (*ATP7B*) gene which encodes a cation transport enzyme that transports copper into bile and incorporates it into ceruloplasmin. Abnormal transportation of copper results in its accumulation, principally in the liver and brain. This leads to cirrhosis and neurological and psychiatric symptoms.

The ophthalmic findings, which rarely cause significant visual loss, are:

- *Kayser–Fleischer rings* – these are rings in the cornea caused by deposition of copper in Desçemet's membrane.
- *Sunflower cataracts* – due to brown or green pigmentation beneath the anterior lens capsule.

MECHANISM OF CELL DEATH

Cells can die through a variety of processes, some of which cause more 'bystander' tissue damage than others. The nature of the injurious agent (e.g. if it is an organism, its virulence) or set of conditions (e.g. if it is a nutritional deficiency, then the level of deficiency) determines the outcome for the cell or tissue. Mild injury may be reversible. Severe injury will result in death either by necrosis or by apoptosis.

Necrosis

Necrosis describes death of a cell or group of cells, typically following severe hypoxia, physical or chemical injury. It is always pathological. There is a rapid depletion of intracellular energy systems. The cell membranes and intracellular organelles are disrupted and typically there is an inflammatory reaction.

Apoptosis

Unlike necrosis, apoptosis, or programmed cell death, is important in both health and disease. It is a rapid process usually affecting single cells within a population of healthy cells. Apoptosis is an active process in which proteins are cleaved by caspase enzymes which break up the nuclear scaffold and cytoskeleton (see Ch. 4, p. 157). The caspases then activate DNAses which degrade nuclear DNA. The next phase involves early phagocytic recognition with expression of phosphatidylserine in the outer layers of the cell membrane due to flipping of phospholipid from the inner layers.

BOX 9-4 MITOCHONDRIAL DISORDERS AND THE EYE

Inherited mitochondrial diseases are rare and mainly affect the central nervous and musculoskeletal system. Abnormal mitochondria are found in the periphery of skeletal muscle fibres and have a characteristic ‘ragged-red’ appearance with Gomori trichrome stain as shown in the figure below.

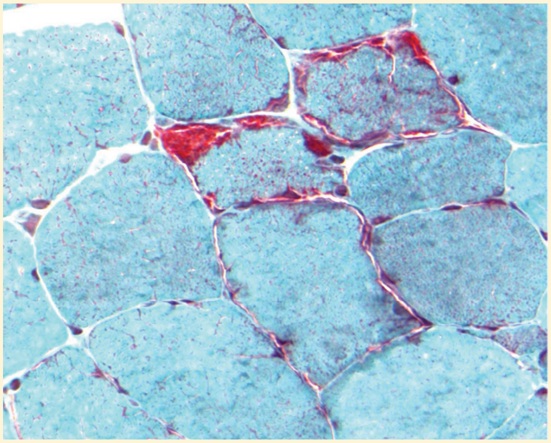

There are significant ophthalmic signs and symptoms in the following conditions:

- *Leber’s hereditary optic atrophy* – point mutations occur in mitochondrial DNA resulting in abnormal ATPase 6. Young males lose vision because of demyelination in the optic nerve; the papillomacular bundle is most severely affected.
- *Chronic progressive external ophthalmoplegia (CPEO)* – this is a slowly progressive, bilateral symmetrical ocular muscle dystrophy that starts in late childhood or early adulthood. Ptosis, external ophthalmoplegia and occasionally a pigmentary retinopathy occur.
- *Kearns–Sayre syndrome* – initial signs and symptoms are similar to CPEO but pigmentary retinopathy is more common with the greatest effect on the macula. Cardiac conduction abnormalities present after the development of ophthalmoplegia.
- *MELAS syndrome (mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes)* – onset is usually in childhood with muscle weakness, recurrent headaches, vomiting and seizures. Stroke-like episodes involve temporary weakness on one side of the body. Eye signs include external ophthalmoplegia, atypical pigmentary retinopathy and nuclear cataract.

The molecular events occurring in apoptosis are outlined in Figure 9-1.

Sublethal cell injury

Various injurious agents will cause reversible cell damage without cell death. This may result in the following cellular changes:

Hydropic swelling. This occurs when cells are incapable of maintaining ionic and fluid homeostasis. This leads to a loss of function of plasma membrane energy-dependent ion pumps. The cell and organelles become swollen, often with cytoplasmic vacuolation. Blunt trauma may cause widespread retinal oedema (commotio retinae), which may be caused by transient spasm of the retinal vessels, producing ischaemia and damage to endothelial cells with leakage into the tissue. However, it is equally likely to result from transient interruption of axoplasmic flow in the ganglion cell processes.

Atrophy. This is a simple decrease in cell size or number resulting in shrinkage of affected tissues and organs. Blunt trauma may also cause shearing of photoreceptors, leading to atrophy of the photoreceptors and focal hypertrophy of the retinal pigment epithelium (RPE) (see below) which extends into the retina (pseudoretinitis pigmentosa).

Lesser degrees of damage may allow the cell time to adapt so that it is able to withstand injury. Mechanisms of adaptation include:

Hypertrophy. This occurs when the cells increase in size, often as a result of increase in functional demand. This can occur in the RPE following injury (see above).

Metaplasia. This is a change from one type of differentiated tissue to another. In dry eye there is a reduction in tear production and this results in a change of the normal conjunctival epithelium with goblet cells to stratified squamous type.

Healing and repair

The eye, like any other tissue, responds to injury by an initial acute inflammatory response followed by vascularization and wound closure in a fibroblastic

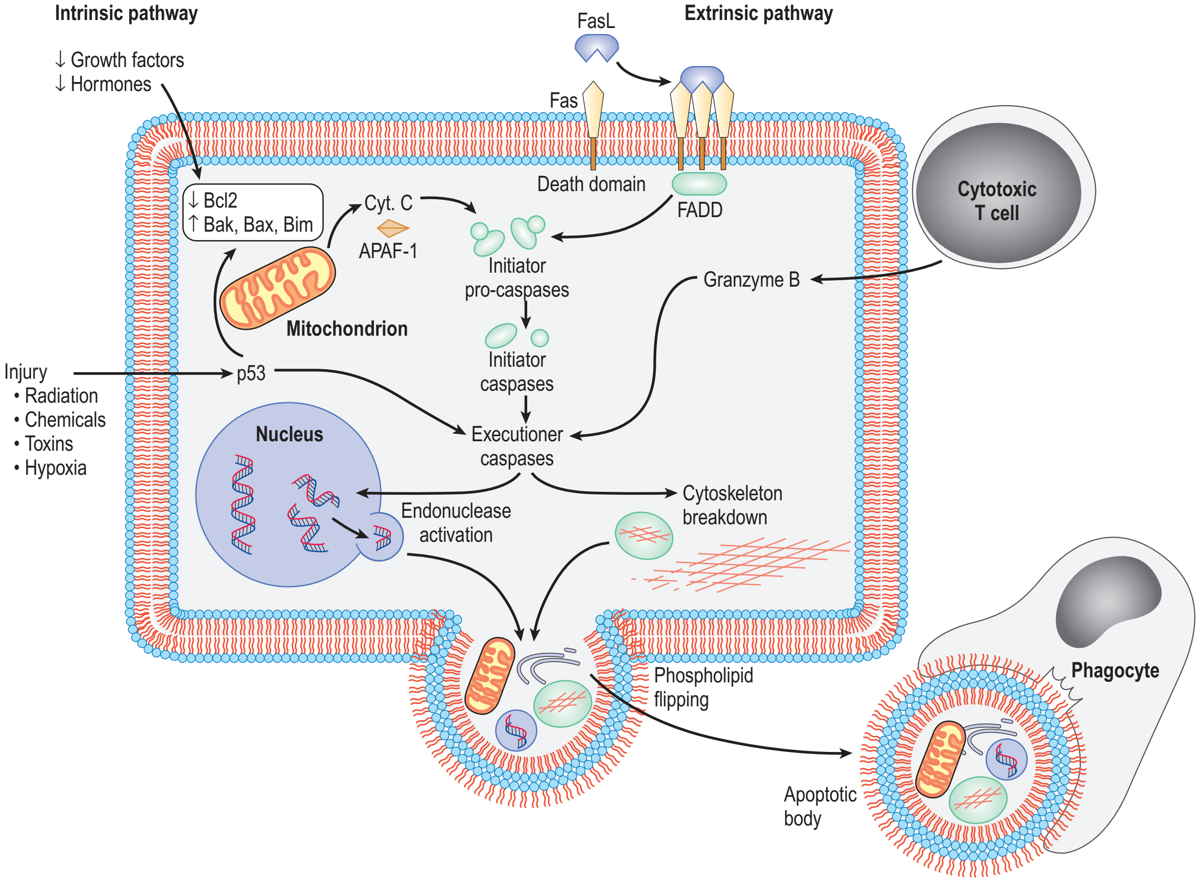
Intrinsic pathway
↓ Growth factors
↓ Hormones
FasL
Extrinsic pathway
Fas
Death domain
FADD
↓ Bcl2
↑ Bak, Bax, Bim
Cyt. C
APAF-1
Initiator
pro-caspases
Mitochondrion
Cytotoxic
T cell
Granzyme B
Initiator
caspases
Injury
• Radiation
• Chemicals
• Toxins
• Hypoxia
p53
Nucleus
Executioner
caspases
Endonuclease
activation
Cytoskeleton
breakdown
Phospholipid
flipping
Phagocyte
Apoptotic
body

FIGURE 9-1 Mechanisms of apoptosis. Apoptosis can be induced by intrinsic or extrinsic pathways. Forms of cell injury or reduction in growth factors or hormones can activate the intrinsic or mitochondrial pathway either by directly influencing the balance of anti-apoptotic proteins (Bcl2 family) and pro-apoptotic proteins (Bak, Bax and Bim) or by inducing activation of p53, which acts on these proteins. An excess of pro-apoptotic proteins increases mitochondrial permeability and allows leakage of other pro-apoptotic proteins (e.g. cytochrome C) into the cytoplasm. Cytochrome binds apoptosis activating factor-1 (APAF-1) and this complex can cleave initiator caspases. The extrinsic pathway is initiated by engagement of cell surface death receptors with the appropriate ligand. Cell surface death receptors include Fas and Type1 TNF receptor. Fas ligand (FasL) cross-links several receptors, allowing the death domains to come in contact, forming a binding site for Fas-associated death domain (FADD). This in turn can active pro-caspases. The execution phase is then carried out by a cascade of further caspases. Cytotoxic T cells can directly activate executioner caspases via granzyme B. Executioner caspases also activate endonucleases, which break up nuclear chromatin. The cell breakdown products are packaged into the apoptotic bodies. Flipping of membrane phospholipids allows rapid phagocytic recognition of apoptotic bodies, avoiding an inflammatory response.

scarring response. These responses may fundamentally alter the architecture and specific function in the tissue and can have serious effects on vision. For example, the wound response of the retina to retinal hole formation and detachment is to stimulate glial proliferation, termed proliferative vitreoretinopathy (PVR). This proliferation does not, however, restore retinal and choroidal integrity but has the opposite effect by contracting and shortening the retina, to the point that retinal reattachment may be surgically impossible. Similar serious consequences result from scarring in other ocular tissues.

Healing in the cornea usually leads to a corneal opacity (scar)

The corneal epithelium regenerates at the limbus (limbal stem cells, see Ch. 4, p. 211) and spreads rapidly across the cornea. Bowman's layer does not regenerate. Stromal keratocytes transform into fibroblasts to heal stromal wounds. Transparency is lost because the collagen fibres are not aligned properly (see Ch. 4, p. 203). Desçemet's membrane does not regenerate. The corneal endothelium fills in defects by sliding and in so doing deposits secondary layers in Desçemet's membrane: the membrane is elastic and there is often recoil at the edge of a deficit.

Iris

The presence of fibrinolysins in the aqueous inhibits fibrin clot formation, and scar tissue does not appear in the iris stroma; this, plus the continuing flow of aqueous from posterior to anterior chamber of the eye, leads to the persisting patency of defects made by iridectomy and iridotomy procedures (with beneficial clinical results). Reactive proliferation of the iris pigment epithelium may occur in response to trauma.

Lens

The lens epithelium responds to some forms of trauma by undergoing fibrous metaplasia, but in general the lens does not heal in the true sense but rather responds to trauma by becoming opaque (cataract).

Retina

Damaged nerve cells are replaced by glial cells (gliosis), which are derived from perivascular astrocytes and Müller cells. In the RPE there is proliferation and metaplasia to fibrous tissue. A combination of glial cells and metaplastic RPE cells is found in preretinal membranes, particularly when such cells migrate via a retinal hole. In cases of retinal detachment, this can lead to extensive gliosis/fibrosis with a fixed, inoperable retina: this condition is termed proliferative vitreoretinopathy (PVR) and is the leading cause of failure in retinal detachment surgery. Currently there is no effective preventive treatment.

Choroid

The melanocytes of the choroid do not proliferate in response to trauma; scar tissue in the choroid is derived from scleral fibroblasts.

Sclera

Scars are formed by proliferation of episcleral fibroblasts.

Optic nerve

Trauma is followed by axonal loss and demyelination with reactive proliferation of glial cells and connective tissue cells.

Inflammation

WHAT HAPPENS IN ACUTE INFLAMMATION?

Inflammation is the dynamic process by which living tissues react to injury (Fig. 9-2). The injurious agent may be physical, chemical, infective or immunological. The classic signs of acute inflammation are redness, heat, swelling, pain and loss of function. These signs can be explained by changes occurring at the microscopic level. Hyperaemia is associated with microvascular changes and accounts for redness and heat. Following injury there is an initial period of vasoconstriction, followed by capillary and then arteriolar dilatation. This is due to the direct effect of injury on the vessels as well as release of chemical mediators from damaged cells. Exudation then occurs, accounting for swelling. This is the increased passage of protein-rich fluid through the vessel wall into the interstitial tissue. This increase in fluid results in dilution of toxins and provides protective antibodies, fibrin and various factors to promote healing. Passage of protein is mediated by direct endothelial injury and chemical mediators, including histamine, bradykinins and leukotriene. Fluid movement is due to increased filtration pressure in part due to hyperaemia and loss of proteins from the capillaries. Following exudation, leucocytes (neutrophils and monocytes) migrate to the site of injury. These cells come in contact with the vascular wall as the flow decreases (margination of neutrophils). They then migrate through the vessel wall by expression, initially of selectins (on endothelial cells), and later of integrins on neutrophils (see Ch. 7, p. 386). Following transendothelial migration and extravasation, the subsequent movement of leucocytes is controlled by chemotaxis. The cells move in response to an increasing concentration gradient of the chemotactic agents. Chemotactic agents are released from other leucocytes (cytokines), complement components (C3a), arachidonic acid derivatives (leukotrienes) or pathogenic bacteria. Once within the tissues, leucocytes clear the injurious agent by

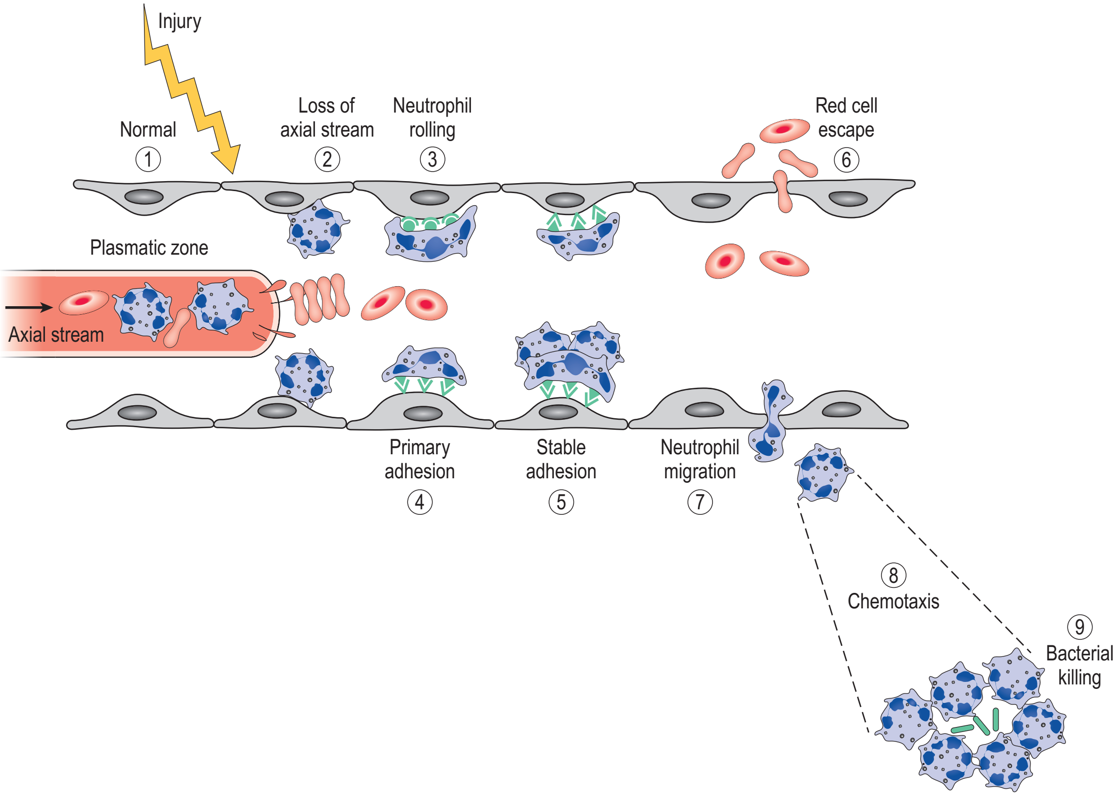

FIGURE 9-2 Acute inflammation. In the normal situation the cellular constituents of the blood move in a central column known as the axial stream with the plasma constituents peripherally (1). Following injury, vascular changes lead to exudation of fluid and slowing of blood flow with loss of the axial stream of cells and margination of neutrophils (2). Neutrophil rolling then occurs due to endothelial selectins recognizing carbohydrate groups on the neutrophils (3). Following this loose adhesion, ICAM-1 and VCAM-1 expressed on the endothelial cells form a primary adhesion by binding integrins expressed on the neutrophils (4). A more stable adhesion forms with aggregation of neutrophils (5). Red cells begin to escape passively through gaps between endothelial cells (6). Transendothelial migration and extravasation of neutrophils occurs due to integrin adhesion to endothelial PECAM-1 and ICAM-1 (7). There is enzymatic degradation of the basement membrane and neutrophils then travel along a chemotactic gradient to the site of injury (8). At the site of injury neutrophils will engulf opsonized bacteria and kill them either by oxygen-dependent formation of free radicals or by oxygen-independent release of lysozyme (9).

phagocytosis. This process involves opsonization of bacteria by complement components before engulfment within the leucocyte. Once within the leucocytes the lysosomes fuse with the phagosome and bacteria may be killed by oxygen-dependent formation of free radicals or by activation of lysosomal enzymes. Various chemical mediators are also released from inflammatory cells or are present within plasma. Those released from inflammatory cells include histamine, serotonin, prostaglandins and leukotrienes as well as a range of cytokines. These mediators may result in vasodilatation (serotonin, prostaglandins), increased vascular permeability (histamine, leukotrienes) and lymphocyte proliferation and macrophage activation (cytokines). Those present in plasma are interrelated cascade systems, including the clotting cascade, fibrinolysis, complement system and bradykinin system. The outcome of acute inflammation depends on various factors, including the infecting organism and the extent of tissue necrosis. Inflammation may resolve, suppurate, repair with organization and scarring, or progress to chronic inflammation.

Inflammation in the eye has special features which are characteristic of the tissue involved. For instance, in the cornea a small traumatic erosion may resolve without scarring. However, bacterial keratitis may lead to formation of a corneal abscess that may ultimately heal with extensive scarring. Ongoing chronic inflammation may occur with herpes simplex stromal keratitis. The 'privileged' status of the eye modifies its response to trauma and infection and thus modifies the overall inflammatory response.

HOW DOES INFECTION CAUSE INFLAMMATION?

Infection occurs when microorganisms invade the sterile internal body tissues and then multiply. The local reaction to infections is usually inflammatory and is evoked by cell damage and death. There are many classes of organisms which cause inflammation (see Ch. 8) and each does so in different ways. Some of the pathology induced by these organisms is briefly described here.

Bacteria

Intraocular bacterial infections are rare but highly destructive. The pathogenicity of bacteria varies according to the species (see Ch. 8, p. 473) and this is reflected in the nature of the immune response to the organism.

Pyogenic infection. Pyogenic bacterial infection of the intraocular compartment is termed endophthalmitis and may be exogenous or metastatic. Panophthalmitis (in which the infection involves the whole of the ocular and periocular tissues) describes a rapid and devastating tissue destruction, which may be complete by 48 hours (Fig. 9-3). Gram-positive cocci are among the commoner organisms that produce a purulent infection when introduced into the eye either by accidental or surgical trauma or by blood spread. The vitreous is an ideal medium for bacterial proliferation, in that the sites for ingress of inflammatory cells (e.g. retinal vessels, the optic disk and pars plana) are some distance from the proliferating pathogens. (Therapy may also be less than optimal because the barrier imposed by the endothelium of the retinal blood vessels impedes the diffusion of antibiotics; see Ch. 6, p. 341.)

Neutrophils are the predominant cell type in metastatic endophthalmitis, i.e. when bacteria enter the eye via the bloodstream, as in bacterial endocarditis or after contamination of an indwelling intravenous catheter. During bacterial killing, neutrophils may release lytic enzymes that can lead to the total destruction of all layers of the retina.

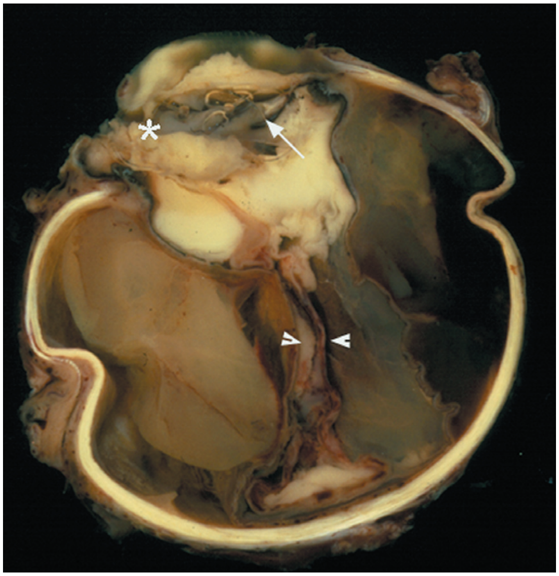

FIGURE 9-3 Enucleation was required after a cataract operation when an intraocular lens implant (arrow) was complicated by a dehiscence of the wound (*). Gram-positive cocci were found in the vitreous abscess. The retina (arrowheads) was detached by haemorrhage.

The ocular coats normally provide an excellent barrier to invasion by bacteria. However, if the cornea is in some way compromised, as for instance during the inappropriate use of steroids in corneal ulceration, the prolonged and incorrect use of contact lenses or inadequate wound closure after intraocular surgery, direct bacterial entry through the cornea may occur.

Several factors increase the risk of endophthalmitis (Box 9-5).

Chronic infection. The mycobacteria (tuberculosis, leprosy), the actinomycetes (skin and lung infection) and the spirochaetes (syphilis, yaws) are responsible for chronic destruction of tissue. *Borrelia burgdorferi*, a spirochaete transmitted by ticks, may cause arthritis, neurological disease and conjunctivitis (Lyme disease). *Bartonella* spp., which causes cat scratch disease, has also been implicated as a cause of neuroretinitis. Pathogens that produce a chronic infection induce humoral and cellular responses in the host (see Ch. 7); the latter

BOX 9-5 FACTORS THAT INCREASE THE RISK OF ENDOPHTHALMITIS

- The major factor in non-accidental post-surgical endophthalmitis is active host infection, e.g. bacterial conjunctivitis.
- The major factor in accidental post-injury endophthalmitis is retention of infected foreign material, especially organic material.
- Poor surgical technique, inadequate wound closure and leaking blebs, e.g. after drainage procedures, can cause endophthalmitis.
- Contaminated intraocular lenses are a common cause of late low-grade bacterial endophthalmitis.
- Debility, chronic corneal ulceration and immunosuppression are rare causes of endophthalmitis.
- Contact lens wear with poor hygiene remains a cause of severe intraocular infection if corneal ulceration occurs.

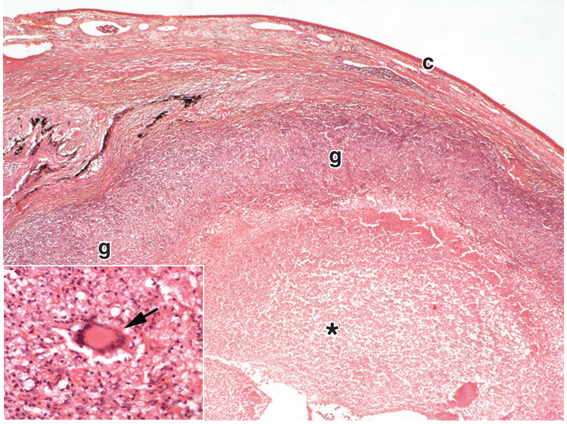

FIGURE 9-4 A tuberculoma involving the anterior chamber. The inflamed cornea (c) is anterior. There is caseous necrosis (*) filling the anterior chamber, which is surrounded by granulomatous inflammation (g). The granulomatous inflammation is shown on higher power in the insert, where epithelioid macrophages have fused to form a multinucleated giant cell (arrow).

is characterized by macrophages, lymphocytes and plasma cells. Many of the pathogenic organisms, such as *Mycobacterium tuberculosis*, have the capacity to survive within the host macrophage. In the case of *M. tuberculosis*, macrophages attempt to limit the spread of the organism by accumulating around the dead and dying (necrotic) cells killed by the organism. The macrophages then become more elongated, taking on an epithelial morphology (epithelioid macrophages). This collection of epithelioid macrophages is known as a granuloma (Fig. 9-4). Fusion with neighbouring macrophages forms a characteristic multinucleate giant cell (Langhans cell). The central mass of dead tissue within the granuloma appears cheese-like macroscopically: hence, the terms 'caseous necrosis,' 'caseating' and 'caseation'.

Viruses

The following viruses most commonly cause ocular disease: herpes simplex virus, herpes zoster virus (varicella group) and cytomegalovirus.

Herpes simplex keratitis. Type 1 herpes simplex virus causes superficial corneal ulceration which is finger-like or dendritic within the epithelium. Primary herpes simplex infection usually occurs through the oral mucosa, the lips or the skin of the face. This is followed by transneural and, subsequently, latent viral infection in the neurones of the trigeminal ganglion and the sympathetic ganglia. The virus is morphologically detectable by ultrastructural investigation within the corneal cells (keratocytes and epithelial cells) and nerves (Fig. 9-5). A variety of possible triggers (ultraviolet light, cold) reactivate the virus, which migrates along the sensory nerves to the corneal epithelium. The virus then invades, replicates and spreads within the epithelium. Effective antiviral therapy has dramatically reduced the complications of herpes simplex keratitis. These include chronic fibrosis and scarring (disciform keratitis) with stromal vascularization and persisting chronic inflammation: leakage of lipid-rich plasma into the corneal stroma leads to pale yellow deposits (secondary lipid keratopathy).

Herpes zoster ophthalmicus. Herpes zoster virus infects the ganglia and branches of sensory nerves such as the trigeminal nerve. When the virus is reactivated, it replicates and produces vesicle formation in the skin in the distribution of the affected nerve or its branches. The eyelids, conjunctiva, cornea and uveal tract are involved in the inflammatory process. A lymphocytic infiltrate appears around the long and short ciliary nerves and is present in the choroid and ciliary body.

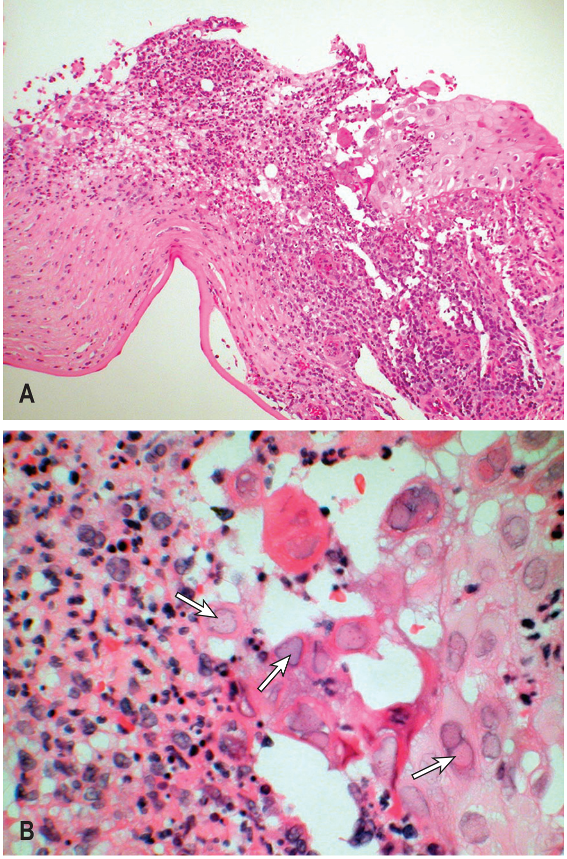

FIGURE 9-5 (**A**) Herpes simplex ulcer in the cornea. (**B**) The epithelial cells contain intranuclear inclusion of herpes viral particles (arrows).

BOX 9-6 POLYMERASE CHAIN REACTION AND THE EYE

Polymerase chain reaction (PCR) involves the selective amplification of specific segments of DNA. The basic reaction involves repetitive cycles of DNA synthesis. Each cycle consists of three steps:

- *Denaturation* – the first step involves denaturation of the target nucleic acid, which renders it single stranded
- *Annealing* – denaturation is followed by annealing of synthetic oligonucleotide primers specifically designed to hybridize to the target nucleic acid region
- *Extension* – the third step involves extension from the annealed primer catalysed by a DNA polymerase enzyme. In a typical PCR analysis 20–40 cycles are carried out with successive products becoming templates for subsequent cycles such that there is exponential amplification of the target region.

The basic technique may be adapted for various applications.

1. Detection of Genetic Mutations and Deletions (e.g. Leber's Hereditary Optic Atrophy)

- Direct sequencing of the PCR product
- Detection of single-strand conformation polymorphisms (SSCP)
- Detection of restriction fragment length polymorphisms (RFLP).

2. Detection of Pathogens

PCR is useful for detecting various pathogens, particularly viral pathogens (e.g. hepatitis B, cytomegalovirus, herpes simplex virus).

3. Detection of Changes in Gene Expression

Reverse transcription PCR is based on the comparison of the amount of PCR product generated with the amount produced from a known concentration or copy number of control amplification targets in the same reaction. This technique can be useful for studying clonality in lymphoma.

Acute retinal necrosis. Acute retinal necrosis has only been recognized as a disorder in the past three decades. The condition can be unilateral or bilateral and usually occurs in immunocompetent individuals. Acute retinal necrosis is usually caused by infection with herpes simplex or varicella zoster virus. Morphological distinction of these viruses can be difficult in retinal biopsies, and polymerase chain reaction (Box 9-6) or *in situ* hybridization (Box 9-7, Fig. 9-6) may be more helpful in providing an accurate diagnosis. Enucleation specimens show sectorial or massive haemorrhagic retinal necrosis, associated with vitreous exudation and choroidal inflammatory cell infiltration. Intranuclear viral inclusion bodies can be seen by light microscopy. Ultrastructural examination reveals viral particles within retinal neurones and within the RPE and vascular endothelium. The virion appears as a central electron-dense core with a surrounding layer (the capsomere) and an outer envelope. The dimension of the infective viral particle is 190–220 nm (see Ch. 8).

Progressive outer retinal necrosis. In immunocompromised individuals, herpes simplex and herpes zoster viruses may cause destruction of the outer

BOX 9-7 *IN SITU* HYBRIDIZATION

In this technique single-stranded complementary nucleic acid sequences can join with specific DNA or RNA sequences in cells or tissues and these hybridization sites can be identified by the addition of a fluorescent *in situ* hybridization (FISH) or enzyme-labelled probe.

- FISH is utilized to detect chromosomal gains, losses and translocations in neoplasms (particularly lymphoma).
- *In situ* hybridization may be used for assessing clonality in lymphoid proliferations and detecting virus nucleic acids in infections with Epstein–Barr virus or cytomegalovirus.

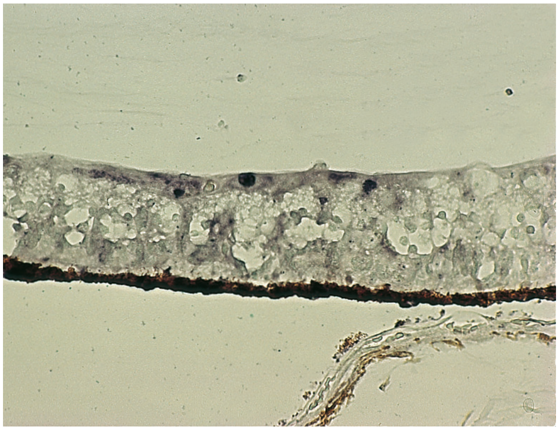

FIGURE 9-6 *In situ* hybridization reveals cytomegalovirus in the large cells in the inner retina in a case of retinitis.

retina without the accompanying vitritis, retinal vasculitis or papillitis usually associated with acute retinal necrosis.

Cytomegalovirus retinitis. Before antiretroviral therapy became available, cytomegalovirus retinitis was a common ocular infection in individuals with acquired immunodeficiency syndrome (AIDS). It is characterized by progressive areas of retinal necrosis, usually without haemorrhage. Infected cells are often enlarged with a characteristic owl-eye inclusion body (Fig. 9-7).

Chlamydia

Chlamydia are a genus of bacteria that are considered to be unique in the sense that they are obligate intracellular microorganisms (see Ch. 8). On a global scale, ocular infection by *Chlamydia trachomatis* (see Ch. 8) is one of the commoner causes of human blindness. The organism is spread by direct contact and is also insect-borne. In an environment with poor hygiene the disease flourishes but the Chlamydiaceae respond to broad-spectrum antibiotics, particularly tetracyclines, and as living standards have improved, the incidence of blindness as the result of trachoma has declined (Box 9-8). Chlamydial conjunctival infection in developed countries, for instance, is much less sight-threatening than in less developed nations and is associated with genitourinary infection.

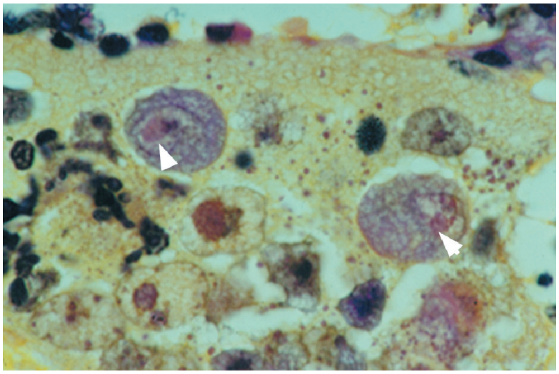

FIGURE 9-7 In cytomegalovirus infection the necrotic retina contains enlarged cells in which characteristic owl-eye intranuclear inclusions are present (arrowheads) (phloxine tartrazine stain).

Fungi

(see Ch. 8) In Europe the most important ocular fungal pathogens include:

- *Candida* spp.
- *Aspergillus* spp.
- *Mucor* spp.

Aspergillus and *Candida* spp. may directly invade the cornea and anterior segment in postoperative infections or can be blood-borne, particularly in drug addicts. In the latter individuals the infection may present as a vitreous abscess.

Mucormycosis is the result of a blood-borne infection and occurs in patients with poorly controlled diabetes or who are immunocompromised. The fungus has a predilection for the lumen of blood vessels, which are occluded by secondary thrombosis: the

BOX 9-8 TRACHOMA

Trachoma is described in four stages:

- In stage I there is epithelial infection with early lymphoid hyperplasia and polymorphs within the conjunctival stroma, which is oedematous.
- Stage II is often subgrouped into type A, in which the lymphoid follicular reaction predominates, or type B, in which there is fibrosis with the formation of papillae. The latter reaction probably represents the effect of secondary bacterial infection. The papillae are formed by fibrovascular proliferation within the thickened and inflamed stroma. The cornea is involved at this stage with ingrowth of a fibrovascular pannus onto the superior corneal periphery.
- Stage III is characterized by fibrous replacement of the inflammatory tissue.
- In stage IV there is contraction within the palpebral conjunctival stroma so that there is internal deformation of the lids (entropion) and trichiasis, which leads to abrasion of the cornea by the lashes. Suppression of tear production is the result of inflammation and fibrosis within the lacrimal gland and its ductular system. Secondary changes occur in the conjunctival epithelium, e.g. stratification and loss of goblet cells, and these also impair tear film stability.

organism parasitizes the ophthalmic artery and its branches, leading to necrosis of the orbital tissues, the nose and the eye (Fig. 9-8).

Some fungal infections are more prevalent in certain regions of the world: e.g. in the USA, where coccidioidomycosis, cryptococcosis, histoplasmosis and blastomycosis are all reported as ocular pathogens.

Protozoal and metazoal infections

(see Ch. 8) Several organisms from this group cause ocular pathology.

***Toxoplasma gondii*.** This is the commonest protozoal parasite to infect the eye. Congenital infection may occur when a woman becomes infected for the first time during pregnancy. Infection causes a classical tetrad of clinical features (meningoencephalitis, hydrocephalus, intracranial calcification and retinochoroiditis). Disease severity depends on in which trimester the infection occurs. Congenital ocular toxoplasmosis is a recurring and progressive disease because of the persistence of the parasite as bradyzoites within tissue cysts, which can reactivate. Foci of reactivation can be seen as an irregular area with associated vitreous haze at the border of a retinochoroidal scar.

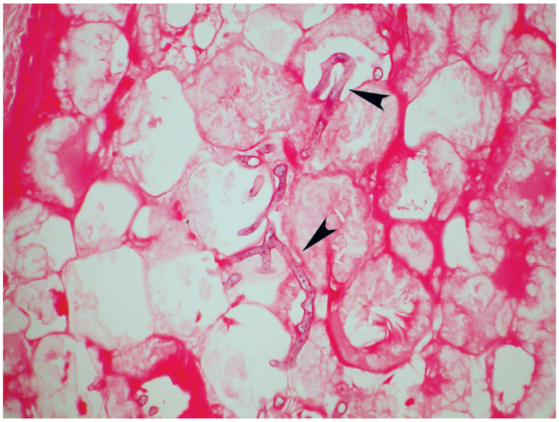

FIGURE 9-8 *Mucor* is the largest of the pathogenic fungi and consists of broad, non-septate branching hyphae (arrowheads). In this case the hyphae are present within infracted orbital fat.

Acquired infection is usually from cysts in undercooked meat or from drinking water containing oocysts from soil contaminated with cat faeces. Acquired infection is commonly asymptomatic, although some individuals will develop a 'flu-like illness with lymphadenopathy. Retinochoroiditis may occur as a result of acute acquired infection, but in contrast with congenital infection there is no pre-existing scar. *Toxoplasma* retinochoroiditis is less common in immunocompromised patients than cytomegalovirus retinitis but can cause extensive retinal necrosis. Individual parasites (tachyzoites) and cysts (containing bradyzoites) can be identified in paraffin sections (Fig. 9-9).

***Acanthamoeba* spp.** The incidence of acanthamoebal keratitis initially increased with greater use of soft contact lenses, but this form of keratitis it is now less common because of increased realization of the importance of hygiene in the handling of the contact lenses. *Acanthamoeba* is a free-living protozoan parasite and the main source of contamination is the fluid in the contact lens case and scales on taps. These loci often contain bacteria that provide nutrition for the protozoa. In the cornea, the acanthamoebae phagocytose

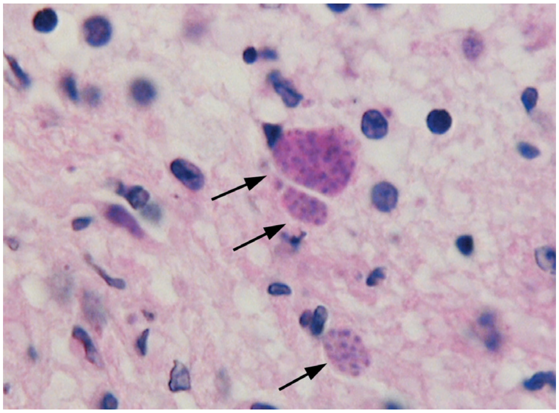

FIGURE 9-9 *Toxoplasma* cysts (arrows) in a retina in which the normal architecture has been destroyed.

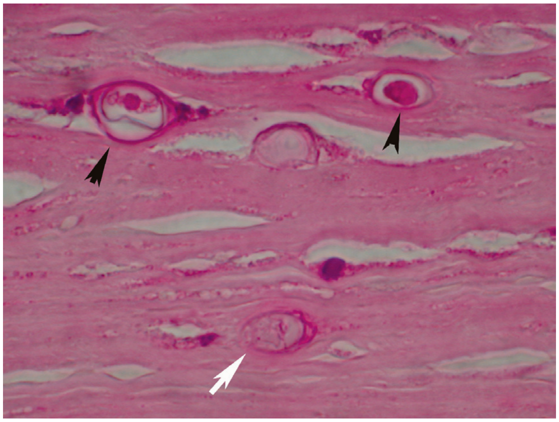

FIGURE 9-10 *Acanthamoeba* cysts (black arrowheads) and a trophozoite (white arrow) lying between the corneal stromal lamellae in a soft contact lens wearer (PAS stain).

remnants of dying keratocytes and polymorphs (Fig. 9-10). The organism can be difficult to identify without the use of immunohistochemistry. Acanthamoebal keratitis is painful, but does not normally penetrate the cornea, remaining as a superficial keratitis; it responds only slowly to appropriate therapy.

Toxocara canis. The adult worm of *Toxocara canis*, which can be several centimetres long, lives in the intestinal tract of the puppy and the eggs are passed out with the faeces. Infants may ingest the eggs, which release second-stage larvae in the stomach. The larvae of *Toxocara* can pass easily through the body tissues and the living organisms do not elicit an inflammatory response. The pathological consequences are usually related to death of the organisms, which heralds the onset of an immune response. The most sensitive organs appear to be the liver, lung, brain and eyes where both a delayed-type and immediate-type hypersensitivity response may occur. In the eye there are three possible outcomes:

- At the posterior pole, a slowly developing low-grade fibrous reaction within the retina can produce a tumour that resembles a retinoblastoma.
- In the mid-periphery, a more rapid active inflammatory reaction, which is characterized by the presence of numerous eosinophils, is followed by exudation into the retina and subretinal space, and then secondary retinal detachment.
- Inflammation and fibrosis occur in the vitreous base over the pars plana, and can induce a form of posterior uveitis, easily mistaken for pars planitis.

Other parasites

In equatorial climates, infection by helminths such as *Wuchereria* spp., *Loa loa*, *Ascaris* and several other nematode larvae causes retinal damage and inflammation when microfilariae migrate into the retina and vitreous. Remarkably, the level of inflammation can be relatively low grade and there develops a mottled pigmentary retinopathy (diffuse unilateral neuroretinitis, DUSN).

INFLAMMATION MAY OCCUR WITHOUT INFECTION

Many forms of intra- and extraocular inflammation occur without evidence of infection and in these cases autoimmune, allergic or otherwise termed ‘immune-meditated’ inflammatory mechanisms are envisaged.

GRANULOMATOUS INFLAMMATION

Macrophages accumulate often with multinucleate giant cells to form granulomas around a focus of tissue destruction. In addition to infectious diseases mentioned above, it is also a feature of a number of non-infectious diseases. Immunologically, these reactions often represent a delayed-type hypersensitivity response (DTH response; see Ch. 7) mediated by T-helper type 1/17 cells in which macrophages represent the effector cell. These granulomatous reactions include the following types.

Reactions to endogenous materials

Products of plasma, blood or cell breakdown can induce a giant cell granulomatous reaction when released into tissue. The commonest eyelid granuloma is the chalazion, which is a reaction to a blocked meibomian gland duct (Fig. 9-11). Rupture of a cyst (either an epidermal inclusion cyst or a dermoid cyst) releases keratin, which is irritant and induces a chronic granulomatous reaction. Red cells and plasma in the extracellular matrix provoke an inflammatory reaction and the cellular response depends on the presence of fibrinolysins (plasmin and plasminogen activator) in the tissue fluids. In the anterior chamber, fibrin is diluted by the aqueous and is rapidly dissolved by fibrinolysins present in the aqueous. Thus, the cellular response to red cells is restricted to migration of macrophages from the iris vessels into the anterior chamber. A similar reaction occurs in the vitreous after bleeding from a tear in the retina or from torn preretinal vessels in vasoproliferative retinopathy. Similar haemogranulomas also occur within the orbit after trauma followed by bleeding (traumatic blood cysts).

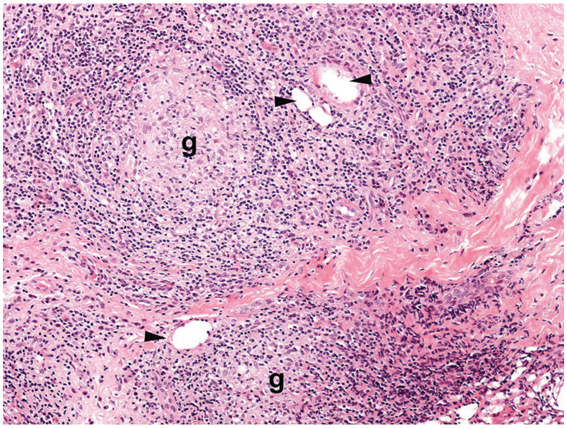

FIGURE 9-11 In a chalazion, blockage of the meibomian gland results in lipid release in the form of fat globules and this stimulates the formation of a granuloma (g). The lipid is dissolved during processing, leaving empty 'fat' spaces (arrowheads). The granulomatous foci are surrounded by lymphocytes and plasma cells. Recurrence is common in chalazia, because the inflammatory reaction and the subsequent fibrosis interfere with drainage from the adjacent glands in the tarsal plate. Sebaceous gland carcinoma is an important entity in the differential diagnosis of recurrent chalazion.

Reactions to exogenous non-biological materials

Implantation in tissue of vegetable or organic matter such as wood excites a similar cellular response to that of sutures derived from cotton or synthetic materials. Synthetic fibres or fragments of plant are seen in polarized light as birefringent particles surrounded by macrophages and lymphocytes. Metallic fragments are slowly dissolved in tissue fluids, but elements such as iron are toxic to the retina, which undergoes neuronal loss, so that metallic foreign bodies in the vitreous or retina are especially dangerous. Brass contains copper and tin, and the reaction to copper ions is pyogenic for reasons that are as yet unknown. On occasion, similar reactions may be seen to materials used in ophthalmic surgery (Box 9-9).

NON-GRANULOMATOUS INFLAMMATION

Lymphocytes and plasma cells are found in a number of conditions in which the aetiology and pathogenesis are unknown. Many clinical forms of anterior and posterior uveitis, e.g. Behçet's disease, are characterized by diffuse and intense lymphocyte and plasma cell infiltration as well as a prominent neutrophil infiltration. Lymphocytic perivasculitis is a feature of demyelinating disease (multiple sclerosis) in the optic nerve and of retinal vasculitis. In endocrine exophthalmos there may be focal clusters of lymphocytes (lymphorrhages) within the extraocular muscles (see p. 505).

INFLAMMATION DUE TO AUTOIMMUNE DISEASE

The basic mechanisms of autoimmune disease pathophysiology have been dealt with elsewhere (see Ch. 7, p. 439). As indicated above, many inflammatory diseases of the eye and adnexae occur in the absence of identifiable causative organisms. Some are associated with generalized connective tissue disease or diseases with a recognized autoimmune aetiology. Others are restricted to the eye and periocular tissues. The following section describes some of these disorders.

Sjögren syndrome

Sjögren syndrome is a disorder of the lacrimal gland, acinar glands of the conjunctiva (glands of Wolfring

BOX 9-9 MATERIALS USED IN OPHTHALMIC SURGERY

Plastic encircling bands made from silicone are used to indent the sclera in detachment surgery, in part because they produce little inflammatory reaction. However, there is always a surrounding fibrous capsule. In contrast a giant cell granulomatous reaction occurs around sutures. Particles of glass or the plastic used for intraocular lenses (polymethylmethacrylate; PMMA) do not stimulate a marked inflammatory reaction: the membranes on the posterior lens capsule are derived from metaplastic lens epithelial cells, not inflammatory cells. The silicone plastic plates and tubes (Molteno tubes and setons) that are designed to drain aqueous in advanced neovascular glaucoma do not excite an inflammatory response, although like the encircling bands the drainage orifices can be blocked by a fibrous capsule. Of the viscous fluids instilled to replace ocular fluids, hyaluronic acid (Healon®) is inert, while silicon oil employed in retinal detachment surgery stimulates a low-grade macrophage reaction after the oil becomes emulsified (see figure).

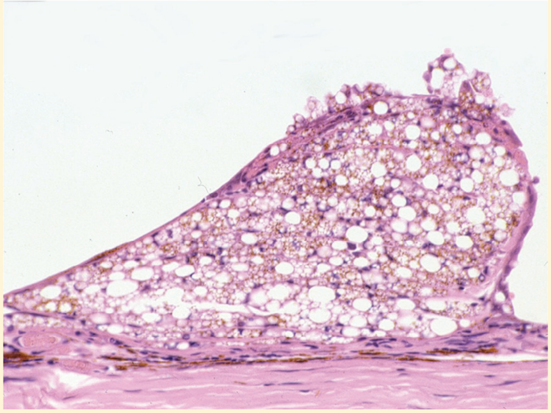

This is a nodule on the retinal pigment epithelium in an eye treated with intravitreal silicone oil for retinal detachment. This nodule consists of macrophages that have phagocytosed emulsified oil globules.

and Krause, see Ch. 1, p. 84) and salivary glands. The secretory acinar tissue of the lacrimal gland is destroyed by a lymphocytic infiltrate, with formation of lymphoepithelial lesions. Impaired secretion of saliva and tears leads to a dry mouth and dry eyes, respectively, and there is an associated loss of goblet cells in the conjunctival epithelium with squamous metaplasia of the surface epithelium. Patients with primary Sjögren syndrome have specific antibodies (anti-Rho, La).

Rheumatoid eye disease

Destruction of ocular tissue in rheumatoid arthritis is T cell-mediated but also includes an immune complex-mediated vasculitis. Impairment of blood flow to the anterior segment causes a necrotizing scleritis and peripheral corneal ulceration. Spontaneous central corneal ulceration (often without an inflammatory cell infiltration) occurs in rheumatoid disease (corneal melt), a phenomenon that is promoted by release of metalloproteinases. If the destruction of collagen in the sclera is severe, the tissue undergoes fibrinoid necrosis with a granulomatous reaction around the necrotic sclera. Thinning of the sclera leads to exposure of the underlying uveal tract (scleromalacia perforans) but ciliary body prolapse and perforation are uncommon. In some patients the inflammatory process is slower and is accompanied by a reactive fibrosis with massive thickening of the sclera (brawny scleritis). A localized inflammatory reaction at the posterior pole of the eye causes macular oedema and the mass may simulate a malignant melanoma (posterior nodular scleritis) with or without an exudative retinal detachment.

Other ocular surface disease

Autoimmune disease is seen in the skin of the eyelid in bullous diseases such as pemphigus and mucous membrane pemphigoid, which are extremely debilitating diseases and very difficult to treat. In the conjunctiva, autoimmune responses against basement membrane components and the attachments of the epithelium lead to severe stromal fibrosis. Immunofluorescence studies performed on tissue submitted in Michel's transport medium show linear deposition of immunoglobulin G and sometimes C3 along the basement membrane (Fig. 9-12). Detachment of epithelium from the basement membrane leads to inadequate protection and predisposes to secondary inflammation and the exudation of fibrin. In the fornix, a fibrinous exudate provides a scaffold for fibroblastic migration; the subsequent scarring process leads to adhesions between the eyelids and the globe (symblepharon).

Lens-induced uveitis

A cataractous lens contains proteins that are the breakdown products of the primary soluble crystallins and

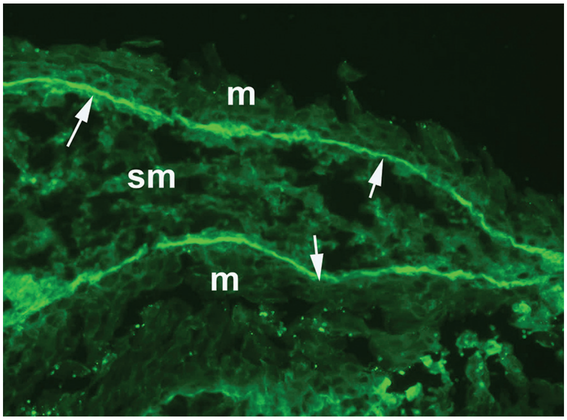

FIGURE 9-12 In ocular cicatricial pemphigoid, immunofluorescence studies may show linear staining for IgG in the basement membrane (arrows) in the conjunctiva. There is only background staining in the mucosa (m) and submucosa (sm).

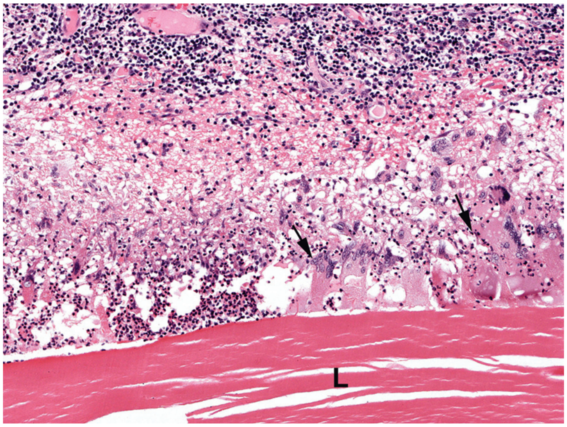

FIGURE 9-13 In lens-induced uveitis, lens cortical matter (L) is attacked by a giant cell granulomatous reaction (arrows). There is a surrounding lymphocytic infiltrate.

other proteins (see Ch. 5). Leakage of lens protein into the anterior chamber, either spontaneously or as a result of trauma, may induce a massive giant cell granulomatous reaction. Macrophages and lymphocytes enter the anterior chamber from dilated blood vessels in the iris and ciliary body, and may enter the lens cortex itself directly through a rupture in the lens capsule. There are commonly significant numbers of neutrophils and eosinophils in the inflammatory infiltrate. The inflammatory cells then pass directly through the epithelium of the pars plana during the associated cyclitis (Fig. 9-13).

Leakage of lens protein into the anterior chamber does not automatically induce a prominent inflammatory response. Sometimes lens protein is relatively inert and induces an uncomplicated macrophage response. This may be associated with a rise in pressure caused by outflow obstruction by engorged macrophages (phakolytic glaucoma). If the rupture is acute, the outflow system is blocked by lens matter (lens particle glaucoma).

Sympathetic ophthalmia

A bilateral granulomatous inflammation of the choroid, ciliary body and iris (panuveitis) can occur after injury to one eye; the injury usually includes uveal incarceration within the sclera. Subsequent sensitization to previously sequestered ocular antigens leads to posterior uveitis mediated by major histocompatibility complex class II-restricted $CD4^+$ T cells, which affect both the exciting eye and contralateral sympathizing eye. This immune response is directed against ocular self-antigens. The uveal tract becomes considerably thickened by an inflammatory infiltrate that includes collections of macrophages (Fig. 9-14). The inflammatory process also involves the retinal pigment epithelium with the accumulation of macrophages (essentially small granulomas) at this site (Dalen–Fuchs nodules).

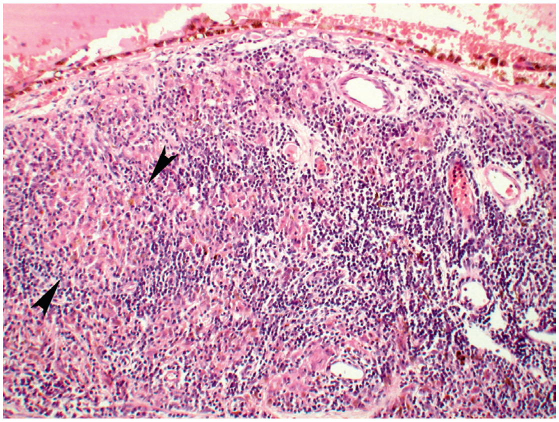

FIGURE 9-14 In sympathetic ophthalmia there is expansion of the choroid by a chronic inflammatory infiltrate including granulomas (arrowheads). A fine dusting of melanin pigment within macrophages is characteristic. The choriocapillaris is spared and the retinal pigment epithelium contains infiltrating macrophages.

Vasculitis

The common feature of vasculitis is inflammation of a vessel wall often accompanied by necrosis. Apart from infection, the majority of the vasculitides are immune-mediated. These are classified according to the size of the vessel principally affected. Temporal arteritis and Takayasu's arteritis are the main large-vessel vasculitides. Those involving medium-sized vessels include classical polyarteritis nodosa and Kawasaki disease. The small-vessel vasculitides include Wegener's granulomatosis, Churg–Strauss syndrome, microscopic polyarteritis and Henoch–Schönlein purpura.

Temporal (giant cell) arteritis. This systemic condition involves the cerebral arteries, the ophthalmic arteries, the posterior ciliary branches and the central retinal arteries. Diagnosis is important as this is a preventable cause of blindness most commonly due to anterior ischaemic optic neuropathy. In susceptible arteries the local dendritic cells recruit and activate $CD4^+$ T cells, which direct the activity of effector macrophages. At an early stage these $CD4^+$ T cells are polarized into Th1 and Th17 cells. The latter are most important in the early stages. The cytokines produce activated macrophages (including IL-1β and IL-6), giant cells and smooth muscle cells, leading to vascular remodelling.

Microscopic examination shows extensive inflammatory cell infiltration, with lymphocytes, plasma cells, eosinophils and macrophages in the media (Fig. 9-15). Inflammation may be patchy with skip lesions. Multinucleated giant cells are usually located near to the fragmented internal elastic lamina. The lumen may be obliterated by a fibrin thrombus, which becomes organized and recanalized. The adventitia and the small periarterial vessels are often involved in the chronic inflammatory process. Steroid therapy does not always totally suppress the inflammatory process and prevent occlusion of the arteries with subsequent blindness in the contralateral eye.

Takayasu's disease. This condition resembles temporal arteritis but the major vessels arising from the arch of the aorta are involved. Unlike temporal arteritis, Takayasu's disease usually occurs in patients under 50 years of age.

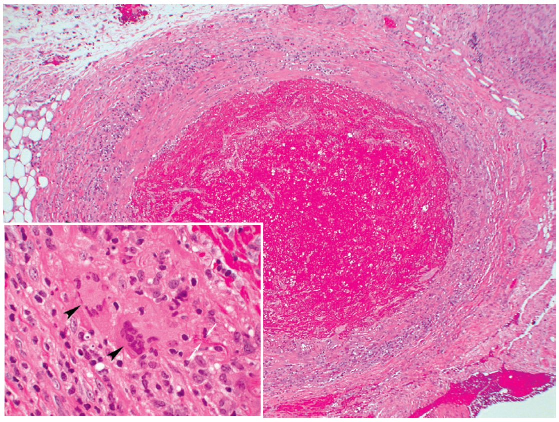

FIGURE 9-15 In temporal arteritis the lumen may be occluded by a thrombus. There is intimal proliferation and inflammatory cell infiltration of the media with multinucleated giant cells (black arrows and inset) in relation to the internal elastic lamina.

Polyarteritis nodosa. In polyarteritis nodosa medium- and small-sized arteries are focally involved in an inflammatory process, which leads to fibrinoid necrosis of the vessel wall and thrombosis. This leads to ischaemia and focal infarction in the heart, central nervous system, kidney and muscle. Polyarteritis nodosa may unusually involve the ophthalmic artery, the central retinal artery and, in addition, the retinal and choroidal blood vessels; subretinal exudation of plasma leads to retinal detachment in this condition.

Wegener's granulomatosis. Wegener's granulomatosis is a small-vessel vasculitis with necrosis and granulomatous inflammation. Ocular manifestations include scleritis, corneoscleral ulceration or an orbital mass. These ocular findings may be part of the generalized systemic disease or its limited form. Generalized disease classically presents with renal, lung, upper respiratory tract and paranasal sinus involvement. The limited form manifests upper respiratory and lung disease without kidney involvement. Serology for circulating antineutrophil cytoplasmic antibodies (c-ANCA) is positive in over 90% of patients with generalized Wegener's granulomatosis but in only 60% of those with the limited form.

Systemic lupus erythematosus

In this multisystem autoimmune disease, antibodies to blood constituents such as leucocyte DNA are responsible for anaemia, thromocytopenia and leukopenia. Phospholipid antibody levels are raised and the consequent disturbances of coagulation explain the occlusive vasculopathy and haemorrhagic diathesis. In addition, DNA–antiDNA antibody complexes are formed in response to the release of DNA from dying cells, and these initiate a type III hypersensitivity reaction. Small-vessel vasculitis may occur in association with systemic lupus erythematosus and is responsible for tissue damage in the heart, lungs, kidney, brain and skin.

Ocular disease is rare but appears as a 'lupus retinopathy' characterized by the presence of retinal microinfarcts; in more severe forms there is occlusive disease of the central retinal artery and vein with haemorrhagic infarction. Choroidopathy is uncommon as a clinical manifestation, but may be apparent on fluorescein angiography.

Metabolic disease

The eyes and extraocular tissues are involved in a number of systemic metabolic diseases and the ophthalmic complications may often be serious.

DIABETES

Diabetic retinopathy is predominantly a microvascular disease in which capillary occlusion and retinal ischaemia are the major features (see Degenerative vascular disease). The fundamental abnormality in the smaller vessels is multilayering of the basement membrane and degeneration of the endothelial cells and the pericytes induced by hyperglycaemia. The thickened, occluded capillary in turn leads to retinal hypoxia which becomes progressively worse with increasingly large areas of retinal tissue involved. This alters the balance of various growth factors, including an increase in vascular endothelial growth factor (VEGF) and placental growth factor (PlGF), which are both angiogenic, with reduction in pigment epithelium-derived growth factor (PEDF) which is anti-angiogenic. Untreated retinal ischaemia then leads to proliferative diabetic retinopathy (PDR), which is the immediate cause of blindness.

In addition to vascular changes (see Video 9-1), diabetes may affect other tissues in the eye, with histological features that include:

- vacuolation of the iris pigment epithelium
- thickening of the basement membranes of the ciliary processes
- cataract (there are no specific features).

THYROID EYE DISEASE

Endocrine exophthalmos may present with unilateral (15%) or bilateral proptosis and limitation of ocular movement in the absence of other clinical signs of hyperthyroidism. In addition to the clinical signs, the diagnosis is usually made by abnormally high values of triiodothyronine (T_3) and thyroxine (T_4), and low values of thyroid-stimulating hormone. The demonstration of uniform swelling of the extraocular muscles on computed tomography or orbital ultrasonography confirms endocrine exophthalmos, and is present in a high proportion of cases (up to 85%) in the absence of overt proptosis. This is an immune-mediated disease for which orbital fibroblasts appear to be the main target. In endocrine exophthalmos they have increased numbers of insulin-like growth factor 1 receptors. Binding of these receptors can increase extracellular matrix production, activate T cells and macrophages and, through cytokine production, initiates and propagates an immune cascade.

This correlates with the findings in extraocular muscle of perivascular lymphocytic infiltration (lymphorrhage) with accumulation of glycosaminoglycans within and around muscle fibres (Fig. 9-16). As the disease progresses, there is replacement fibrosis between the muscle fibres. Study of the nerves at the orbital apex reveals loss of larger axons in the motor nerves; this is attributed to compression by the swollen muscle and explains the limitation of movement (ophthalmoplegia).

DISORDERS OF AMINO ACID METABOLISM

Homocystinuria

The biochemical abnormality in this disease is a reduction in levels of cystathione β-synthetase. Patients suffering from this condition are at surgical and anaesthetic risk because of a tendency to thromboembolic disease, which can be fatal. Dislocation of the lens (inferiorly and somewhat posteriorly) is the result of an acquired metabolic abnormality of the zonular fibres. Histological examination of the zonular fibres

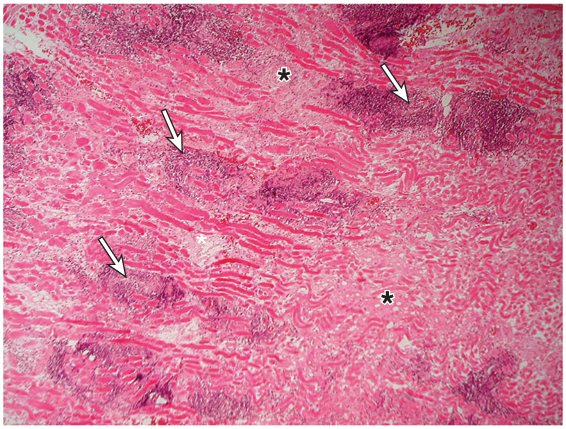

FIGURE 9-16 Endocrine exophthalmos. The striated muscle fibres in the extraocular muscle are separated by lipid and mucopolysaccharide (*). Throughout the muscle there are collections of lymphocytes (arrows).

reveals deposition of a thick band of periodic acid–Schiff-positive material on the inner surface of the ciliary processes and the pars plana.

Cystinosis

The biochemical disturbance occurs in the lysosomal membrane transport of cystine, which is continuously released in the lysosomes in the degradation of protein. In cystinosis the amino acid is trapped within the lysosome because of a defect in the transport system. The ocular manifestations – the accumulation of birefringent cystine crystals in conjunctiva, cornea, choroid, pigment epithelium and retina – have long been recognized. Alcohol fixation (100%) is necessary for histological verification of the presence of cystine crystals in a conjunctival biopsy.

Ageing, degeneration and dystrophies

Ageing can be difficult to define but in general there is overlap in the conditions to be described here since a gradual decline in tissue function (degeneration) is often accompanied by failing nutritive resource to the tissues (dystrophy) as part of the ageing process.

AGEING

There are two main theories of ageing which need not be mutually exclusive. The first suggests that we are genetically programmed to age and that this will differ between individuals. These intrinsic controlling factors include expression of ageing genes in mitochondria and loss of cells' ability to divide due to sequential telomeric shortening during life. Telomeres are a region of repetitive nucleotide sequence at the end of the chromosome arms which protects the chromosome from progressive shortening which naturally occurs during cell division. Telomeres consist of approximately 8000 base pairs at birth and around 3000 base pairs as people age. Cells can normally only divide between 50 and 70 times, with the telomeres getting progressively shorter before the cells become senescent, die or sustain genetic damage that causes neoplastic transformation. The second theory suggests that damage is due to the additive effects of sublethal injury by a variety of injurious agents, e.g. diet, smoking, UV light, etc. This wear and tear is caused by the production of tissue free radicals induced by damage from these extrinsic factors. Ageing changes occur in all the structures of the eye and cause varying effects. The effects of ageing and the eye are summarized in Box 9-10.

DEGENERATION

Degeneration of tissues in pathological terms commonly involves connective tissue components such as collagen, elastin and proteoglycans, but probably represents progressive failure, dysfunction and death of parenchymal tissue cells which are responsible for the continuous production of these elements. Degenerations are described in terms of their composition and their location.

Composition/type of degeneration

There are many types of degeneration which vary in terms of their biochemical composition, probably as a reflection of the initial insult which triggered the process.

Degeneration can affect many tissues (systemic) or be localized to one site, for instance in relation to a previous injury or insult (the withered arm of birth injury is an example). The eye can be the preferred site for certain degenerative processes.

Hyalinization. This describes the replacement of normal cells by an acellular, almost transparent, matrix

BOX 9-10 AGEING AND THE EYE

Ageing changes occur in all structures of the eye with variable effects on vision.

Eyelid and Lacrimal System

Skin thinning and laxity

- Redundant skin folds and wrinkles
- Ectropian or entropian
- Blepharoptosis and brow prolapse
- Orbital fat prolapse
- Tearing (due to punctual eversion)

Atrophy of lacrimal gland

- Dry eye

Cornea

- Changes in curvature, loss of lustre and sensitivity
- Arcus senilis
- Hassall–Henle bodies at periphery
- Krukenberg's spindle (uveal pigment in endothelium)
- Decrease in corneal endothelium

Trabecular Meshwork and Uvea

- Increased pigmentation of trabecular meshwork
- Increased resistance to aqueous outflow
- Decreased pupil size and reactivity
- Fibrosis of ciliary muscle (contributing to presbyopia)

Lens

- Increased absorption of blue light due to accumulation of yellow pigments
- Nuclear sclerosis (contributing to presbyopia)
- Loss of elasticity of zonules (contributing to presbyopia)

Vitreous

- Condensation of vitreous gel
- Liquefaction of vitreous gel with lacunae
- Detachment of posterior vitreous face

Retina and Retinal Pigment Epithelium

- Neuronal cell loss and degeneration
- Thickening of basement membrane
- Atrophy of retinal pigment epithelium with decrease in melanin and increase in lipofuscin
- Age-related macular degeneration (see Box 9-14)

(which consists of collagens and glycoproteins) that imparts a glassy appearance to the tissues on microscopy. Hyalinization is typically seen in the eye and kidney in the walls of small blood vessels in ageing, benign hypertension and diabetes. Leakage of plasma into the vessel wall, owing to breakdown of the normal endothelial barrier, is thought to be one cause of hyalinization.

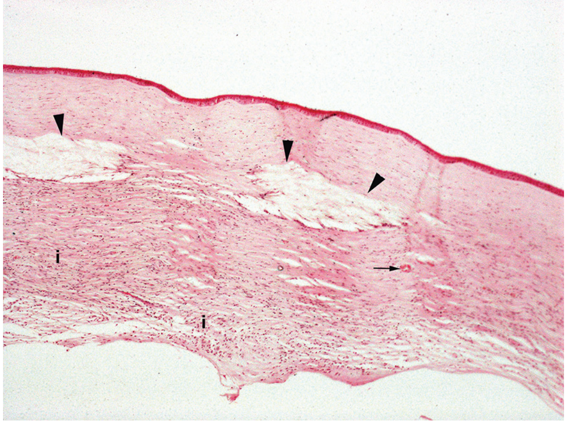

FIGURE 9-17 At the end stage of corneal inflammatory disease the stroma is invaded by blood vessels (arrow) that leak lipid in the form of cholesterol crystals (arrowheads). There is an intense inflammatory infiltrate (i).

Fatty degeneration. The most innocuous form of fatty infiltration in tissue is seen in the peripheral corneal stroma as part of the normal ageing process and is described as arcus senilis. After prolonged inflammation followed by corneal vascularization, plasma lipids leak from the blood vessels and are deposited in the stroma (Fig. 9-17).

Deposition of fat (neutral lipids and cholesterol) in the intima of medium and large muscular arteries (atheroma) is followed by thrombosis (see above under vascular disease). Lipids within clumps of macrophages in the dermis of the eyelid (xanthelasma) are usually a feature of ageing, but hypercholesterolaemia must be excluded.

Elastic fibre degeneration. Elastic fibres can be visualized with special stains (such as orcein). They appear as fine strands in tissues such as skin and blood vessels, and zonule of the lens. The constituent protein, elastin, is arranged in coils, imparting elasticity to the strand (see Ch. 4). 'Elastotic degeneration' in skin is frequently the result of chronic sun exposure, which induces defective fibroblast function and an altered elastic matrix, which has poor elasticity. Reduced skin elasticity is also seen in pseudoxanthoma elasticum. In this condition ruptures in Bruch's membrane expose the choroid (angioid streaks).

Pinguecula and pterygium. These are types of degeneration specific to the ocular surface. In the conjunctiva, deposition of elastic-like material causes thickening and formation of nodules on the bulbar conjunctiva (pinguecula). In individuals exposed to a hot, dry, dusty environment, foci of elastotic degeneration form at the limbus in the interpalpebral fissure and encroach on the cornea as a wing-shaped wedge (pterygium) (Fig. 9-18). Various changes, including dysplasia and carcinoma, may occur in the epithelium overlying the elastotic tissue. Pterygium is considered to be due to high levels of UV light damage and is especially common in certain groups of individuals such as surfers, and those living in islands in the Pacific Ocean.

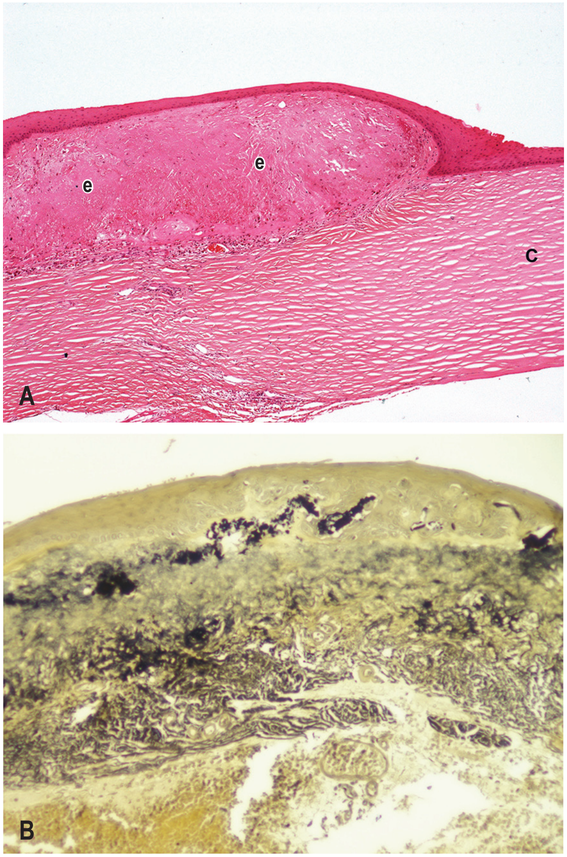

FIGURE 9-18 (A) A pterygium encroaching on the cornea (C). It consists of dense deposition of degenerating elastin (e). **(B)** The degenerate elastin stains black using a special stain (Elastic Van Gieson).

Calcification. Calcium is deposited in both normal and diseased tissue as hydroxyapatite crystals [$Ca_{10}(PO_4)_6(OH)_2$)]. In hypercalcaemic states such as hyperparathyroidism, hypervitaminosis D and excessive bone resorption from skeletal metastases, calcium is deposited in normal tissue such as the kidney and the conjunctiva; this process is called metastatic calcification.

By contrast, calcium can be deposited in hyalinized connective tissue (blood vessels) or necrotic tissue (such as post-tuberculous scars in lung, atheromatous plaques, necrotic tumour tissue in a retinoblastoma) in a normocalcaemic state; this is referred to as dystrophic calcification. Calcification of ocular tissues also occurs in the end-stage phthisical eye usually within the metaplastic fibrous tissue derived from the retinal pigment epithelium. This may sometimes ossify.

Amyloid

Amyloid is an insoluble protein deposited in tissues, particularly around blood vessels and in basement membranes. In haematoxylin and eosin-stained sections amyloid has a homogeneous pink appearance; staining with Congo red, followed by examination in polarized light, reveals apple green birefringence (Fig. 9-19). The composition of amyloid varies and the effects depend on whether it is systemic or localized.

Systemic amyloid deposition

- Associated with monoclonal plasma cell proliferation, e.g. myeloma, Waldenström's macroglobulinaemia. The amyloid is light-chain-derived (AL) from fragments of immunoglobulin.
- Associated with chronic inflammation, e.g. rheumatoid arthritis; genetically inherited familial Mediterranean fever. The amyloid is derived from serum AA protein (AA), an acute-phase reactant in many inflammatory conditions.

Localized amyloid deposition

- Amyloid derived from polypeptide hormones may be deposited in endocrine tumours, e.g. medullary carcinoma of the thyroid.
- Amyloid derived from prealbumin may be deposited in the heart, brain and joints in the elderly. Cerebral deposits of amyloid are

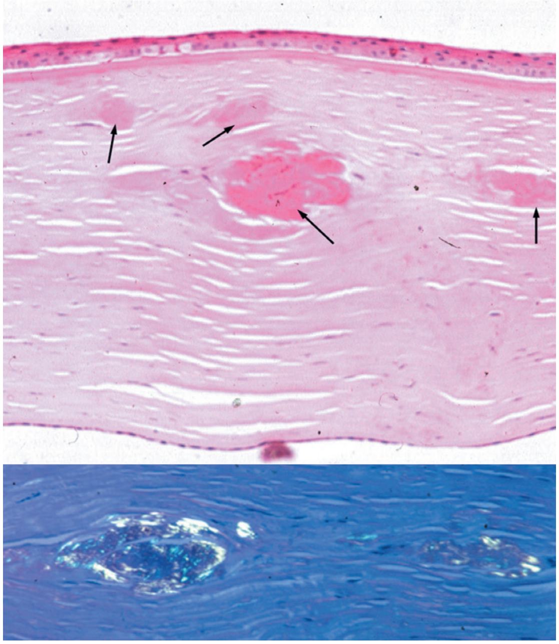

FIGURE 9-19 Upper panel: in lattice dystrophy of the cornea, stromal deposits (arrows) have the staining characteristics of amyloid and exhibit apple green birefringence when a Congo red-stained section is observed in polarized light (lower panel).

important in Alzheimer's disease. Amyloid protein is also a component of drusen in macular degeneration.

Amyloid may be observed as a solitary nodule within the eyelid, the orbit, or in the conjunctiva in the absence of systemic disease. Amyloid is deposited in the choroid and vitreous in systemic amyloidosis. In the cornea, amyloid deposition (Fig. 9-19) is the characteristic feature of lattice dystrophy.

The exfoliation syndrome (pseudoexfoliation syndrome)

Pseudoexfoliation syndrome is an ageing-related systemic disease in which granular amorphous eosinophilic material accumulates throughout the body. In the eye, this material forms fluffy white deposits on the surface of the lens, the ciliary processes, the iris surface and the inner surface of the trabecular meshwork. Involvement of the outflow system leads to secondary open-angle glaucoma. The true pathogenesis of the exfoliation syndrome is not fully understood although it is thought to have a genetic component due to increased prevalence in certain populations (e.g. Scandinavians).

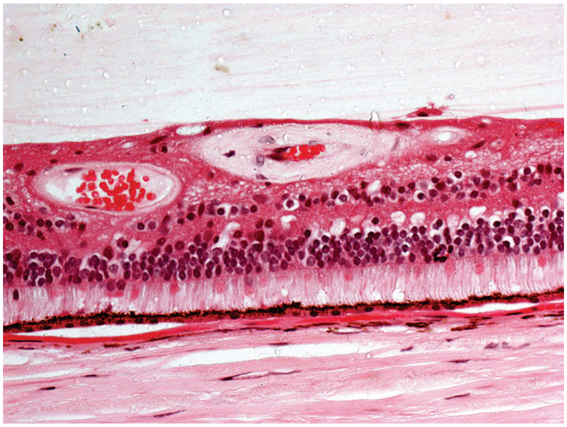

FIGURE 9-20 A hyalinized retinal vessel with a thickened wall and narrowed lumen in a case of long-standing hypertension.

Degenerative vascular disease

Degenerative vascular disease (atheroma, atherosclerosis) is a major cause of morbidity and mortality in industrialized nations, and is also increased in metabolic disease such as diabetes. The degenerative process in the vessel wall can take on various forms as described above:

Hyalinization. In ageing, hypertension and diabetes, the walls of arterioles and venules become thickened by deposition of collagen (hyalinization) with a loss of the normal smooth muscle layer (Fig. 9-20). Consequently there is a reduced capacity to respond to metabolic demand while narrowing of the lumen reduces the overall perfusion of the tissues. Hypertension can be superimposed on both diabetic and senile degenerative vasculopathy, adding a contractile (vasospastic) component to the disease.

Vaso-occlusive disease. Retinal vaso-occlusive disease may be macro- or microvascular. Macrovascular occlusions involve the central retinal artery or vein, or its major branches. Microvascular occlusion occurs at the arteriolar level, causing non-perfusion of tissue and ischaemia.

Microvascular occlusion. The following abnormalities may be found in conditions in which focal capillary occlusion causes patchy interference with blood flow within the retina such as hypertension, diabetes, AIDS, radiation vasculopathy and the vasculitides:

- *Microinfarction* – fluffy white swellings (cotton-wool spots) in the retina on ophthalmoscopy are representations of the swollen ends of interrupted axons. Build-up of axoplasmic flow occurs at the edge of the area previously supplied by the occluded vessel (Fig. 9-21).
- *Hard exudates* – underperfusion of the vascular bed and damage to the endothelium of the deep capillaries leads to plasma leakage into the outer plexiform layer. Clinically this exudation is yellow and well circumscribed. Histologically, 'hard' exudates are eosinophilic masses, and these contain foamy macrophages with lipid in the cytoplasm (Fig. 9-22).
- *Microaneurysms* – another effect of ischaemia on the capillary is weakening of the wall by necrosis of the supporting cell (the pericyte) in diabetes and the endothelial cell in central retinal vein occlusion. The ensuing small bulges or blowouts in the capillary wall are referred to as microaneurysms. With time, microaneurysms become filled by basement membrane deposits and consequently may disappear on fluorescein angiography.
- *Haemorrhage* – breakdown of the vessel wall leads to leakage of red cells and can take several forms in the retina:
  - flame haemorrhages follow rupture of a small arteriole so that blood tracks into the nerve fibre layer
  - dot haemorrhages follow rupture of capillaries in the outer plexiform layer; these are smaller and more circumscribed than flame haemorrhages
  - blot haemorrhages are larger than dot haemorrhages, and represent bleeding from capillaries with tracking between the photoreceptors and the RPE.

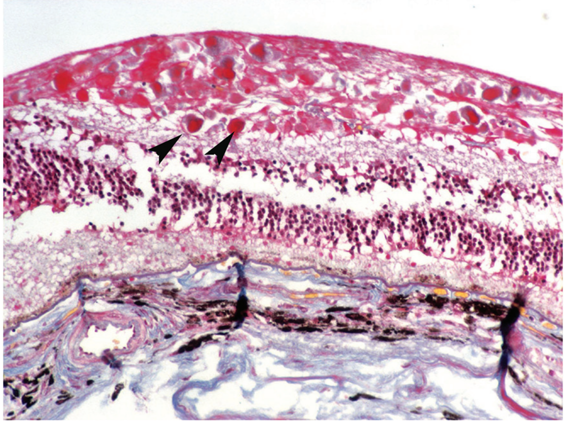

FIGURE 9-21 A microinfarct in the retina is seen as a swollen sector of disrupted axons. Smudgy eosinophilic structures (cytoid bodies) represent the swollen ends of axons (arrowheads). The infarct mainly involves the nerve fibre layer (Bodian stain).

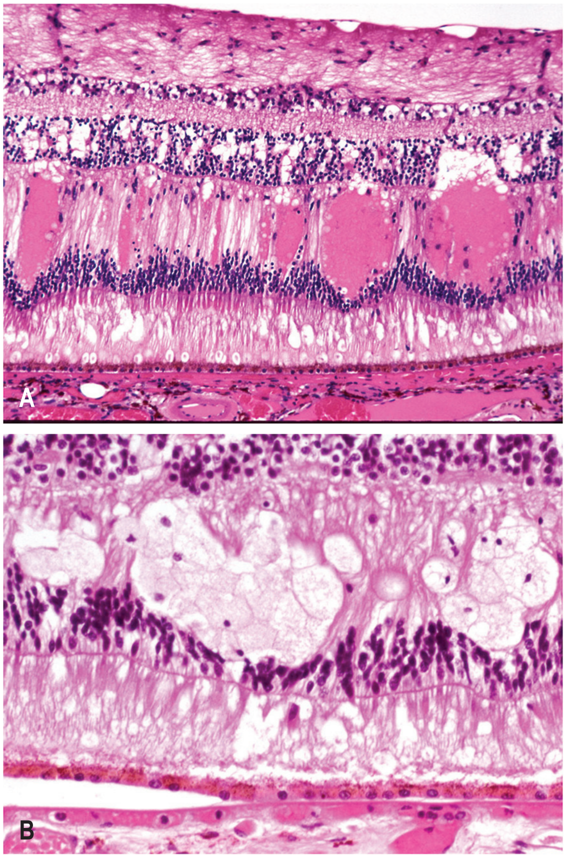

FIGURE 9-22 Hard exudates are the result of leakage of plasma through the capillary endothelium. The exudates occur in the outer plexiform layer and initially consist of a proteinaceous exudate (**A**). Later, foamy macrophages (**B**) are attracted to the deposit in an attempt to remove it.

- *Neovascularization* – newly formed vessels grow from the venous side of the capillary bed within an area of arteriolar non-perfusion; this change represents a response to ischaemia within the retina. These vessels leak on fluorescein angiography and they occur in an eye which will progress to vasoproliferative retinopathy. The new vessels arise in the prevenular capillaries and in the walls of hyalinized venules, and proliferate within and on the surface of the retina. If the vitreous is detached, the fibrovascular tissue grows on the inner surface of the retina; the membrane contracts, leading to retinal detachment. Fibrovascular proliferation within the attached vitreous leads to haemorrhage and further formation of traction bands. The diffusion of vasoformative factors from the vitreous through the posterior and anterior chambers induces blood vessel formation on the iris surface (rubeosis iridis) and the inner surface of the trabecular meshwork, sealing off the angle and causing secondary (neovascular) glaucoma. Vasoproliferation also occurs in retinopathy of prematurity (Box 9-11), which is also intricately associated with relative retinal ischaemia.

BOX 9-11 RETINOPATHY OF PREMATURITY

In the premature infant vascularization of the retina is incomplete. Normally the blood vessels grow from the disk toward the periphery during intrauterine life, and the process is not complete until term, particularly at the temporal periphery, furthest from the disk. The extension of the normal vascular bed appears to be a response to the relative hypoxia of the proliferating neural cells. Migration of blood vessels does not occur while the premature infant is maintained in an atmosphere of high oxygen tension, possibly because the neural tissue is adequately oxygenated and the drive for normal vascularization is lost. Excessive proliferation of blood vessels (retinopathy of prematurity) occurs when the infant is returned to an atmosphere containing a normal partial pressure of oxygen. The peripheral non-vascularized retina is now ischaemic and the neovascular outgrowths from the peripheral vessels proliferate rapidly and in a disorganized manner within the retina and vitreous. The process may result in bilateral retinal detachment in the worst cases, but many cases of retinopathy of prematurity regress without permanent damage. Non-invasive techniques for the measurement of blood oxygen levels and careful control of the oxygen levels in the incubator can reduce the incidence of retinopathy of prematurity. Where necessary, retinopathy of prematurity can be treated by laser photocoagulation of the peripheral non-vascularized retina.

Macrovascular occlusion. Macrovascular occlusion refers to obstruction of vessels of diameters equal to or greater than a medium-sized arteriole. It includes thrombotic occlusion of the central retinal artery, in association with systemic conditions such as hypertension and atherosclerosis.

- *Hypertension* – the classic textbook appearances of the retina in accelerated (malignant) hypertension, i.e. haemorrhage, exudates and papilloedema, are not normally seen with antihypertensive therapy. In mild cases of hypertensive retinopathy the vessels are described as ‘copper-wire’ or ‘silver-wire’ as a result of hyalinization. In more advanced disease, narrowing of the blood column followed spasm of the vessels, produces ischaemic damage to the endothelium distal to the constriction. Swelling and degeneration of the endothelium is followed by leakage of fibrin into the vessel wall and further narrowing of the lumen. Fibrinoid necrosis, characteristic of hypertensive renal vasculopathy, may be found in the walls of choroidal and retinal vessels in uncontrolled hypertension. If the choriocapillaris is occluded by fibrinoid necrosis, there is exudation beneath a necrotic retinal pigment epithelium, leaving small areas of depigmentation (Elschnig’s spots).
- *Central retinal artery occlusion* – the central retinal artery is an end-artery, and obstruction of flow leads immediately to blindness. The normally transparent retina becomes opaque, preventing transmission of the red reflex created by the choroidal vasculature except at the macula where the choriocapillaris is visible (the classic cherry-red spot). Central retinal artery occlusion may be the result of thrombosis in a degenerate central retinal artery but is more often the result of an embolus, typically from a mural thrombus on the endocardium after a myocardial infarction, an atheromatous plaque in the carotid artery, or the heart valves in subacute bacterial endocarditis. After total infarction, none of the inner retinal tissue survives and vasoformative

factors are not released; thus, rubeotic (neovascular) glaucoma occurs as a complication in fewer than 5% of cases, unlike central retinal vein occlusion in which up to 50% of cases may progress to glaucoma. Emboli may originate from many sources (Box 9-12).

- *Central retinal vein occlusion* – the characteristic difference between central retinal artery and vein occlusion on fundoscopy is the presence of extensive haemorrhages within the retina in the latter. There may also be some recovery of vision in venous occlusion, unlike retinal occlusion, but in a significant number of patients, rubeotic (neovascular) glaucoma develops within 3 months. Preretinal neovascularization and glaucoma are most frequent where there is extensive retinal ischaemia and may be prevented by peripheral retinal photoablation if the ischaemia is identified by fluorescein angiography. Frank neovascular glaucoma is usually not amenable to treatment, which is usually aimed at palliation of symptoms. Enucleation of the eye may be necessary ultimately, to relieve intractable pain. Central retinal vein occlusion is also seen in a younger age group, where it may be extremely difficult to differentiate from retinal vasculitis (see above); in some women blood hyperviscosity as a result of oral contraception may be implicated. This form of central retinal vein occlusion has a better prognosis if it is the result of retinal vasculitis, and steroids or ciclosporin A will prevent its progression to neovascular glaucoma. The pathogenesis is not entirely clear (Box 9-13).

Cataract

Cataract is almost a normal part of the ageing process, which starts with the hardening of the lens, which prevents accommodation and near vision for reading (presbyopia, see Ch. 1, p. 36). Secondary cataract also occurs after any insult to the lens (see Ch. 4).

Cataract associated with ageing. The lens crystallins (see Chs 1 and 4) break down to albuminoids, partly as an age-related process and partly in response to exposure to light, particularly of ultraviolet/blue wavelength. The amino acids (e.g. tyrosine) that are released are converted to adrenaline and melanin, so

BOX 9-12 EMBOLISM

An embolus is any abnormal mass of matter carried in the bloodstream and large enough to occlude a vessel. The various types of embolism are listed below:

- *Thrombotic* – thrombus formation in the leg and pelvic veins is the principal cause of pulmonary embolism and intraoperative and postoperative death. In the retinal or choroidal vessels an embolus can originate from a thrombus on the mitral and aortic valves or from ulcerating atheromatous plaques in the aorta or carotid arteries. Another source may be a mural thrombus in the left ventricle.
- *Air embolism* occurs when negative pressure in the neck veins follows thyroid surgery or when fluid or air is forced into the venous circulation during a blood transfusion. Frothing of the blood in the right ventricle interferes with ventricular pumping and is fatal.
- *Tumour emboli* are usually small and not visible in the retinal circulation; larger metastases occur in the choroid.
- *Fat and marrow embolization* occurring after severe trauma to the limbs and trunk is accompanied by multiple fractures. Purpuric spots are seen on the upper thorax, and small haemorrhages are found in the retina. A severe form may rarely occur as Purtscher's retinopathy in which florid embolization of multiple small vessels occurs.
- *Emboli from atheromatous aortic or carotid plaques* may consist of cholesterol/calcified tissue/fibrin/platelets, and can be seen migrating through the retinal circulation by ophthalmoscopy or videofluorescein angiography.
- *Septic emboli* were described in the retina by Roth in 1905, when subacute bacterial endocarditis was common. The typical Roth's spot has a white centre and red surround, and is thought to be the result of vascular damage from an impacted mass of white cells and bacteria in a retinal arteriole. The similarity between this appearance and the deposit of leukaemic cells or a simple infarct surrounded by red cells in a thrombocytopenic immunosuppressed patient has broadened the definition of Roth's spots in contemporary ophthalmology.
- *Amniotic fluid embolism* is a complication of parturition, particularly when manipulation of the fetus is required. Release of amniotic fluid, vernix, hairs and fetal squames into the maternal circulation is commonly fatal.

BOX 9-13 PATHOGENESIS OF CENTRAL RETINAL VEIN OCCLUSION

The radius of the central vein in the lamina cribrosa of the optic disk is about 50% of that in the prelaminar part of the disk and smaller than that in the retrolaminar region. Furthermore, this narrowing of the vein within the lamina cribrosa becomes more variable with age. This is an unusual configuration (in most venous drainage systems the tributaries have a smaller radius than trunk vessels) and it can be explained by the necessity to maintain a high pressure in the retinal capillary bed against an intraocular pressure of 10–20 mmHg. The resistance provided by venous narrowing in the lamina cribrosa is disadvantageous in that flow is markedly increased in the narrowed segment and the resultant turbulence predisposes to thrombosis, particularly if there is an increase in blood viscosity (e.g. hyperglobulinaemia, polycythaemia) or in intraocular pressure. While systemic factors are important, an anatomical explanation has attractions because central retinal vein occlusion is almost always unilateral.

Other important systemic factors in central retinal vein thrombosis include abrupt falls in systemic blood pressure and hence the pressure in the central artery. If these are coupled with a diseased retinal vasculature (which predisposes to underperfusion and stasis in the vascular bed) the criteria for Virchow's triad are fulfilled.

that lens pigmentation progresses from yellow to brown (brunescent cataract) to black (cataracta nigra) (see Ch. 4, p. 228).

Secondary cataract. The biochemical requirements for the maintenance of transparency of the tissue are discussed in Chapter 4. Any metabolic disturbance, such as diabetes or hypocalcaemia, may potentially alter this microenvironment and can lead to lens opacification. The epithelial cells in the lens are also particularly sensitive to ionizing radiation and mechanical trauma; breakdown of transport mechanisms in the membranes of the lens fibre cells and the epithelium promotes ionic imbalance and fluid inflow, causing disorganization of the lens proteins and loss of transparency.

Glaucoma

Although malformation of the outflow system can lead to congenital or juvenile glaucoma (see Ch. 2), the most common forms of glaucoma are the result of degenerative disease and may be subclassified into primary or secondary types.

Primary open-angle glaucoma. Primary open-angle glaucoma (POAG) is a disease that increases in incidence with age and is one of the most common causes of blindness. In some cases the disease has a genetic basis – the myocilin gene (*MYOC*, formerly known as the trabecular meshwork-induced glucocorticoid response gene, or *TIGR*), located on chromosome 1, encodes the protein myocilin, which in turn is involved in the contractile function of trabecular meshwork cells. The optineurin (*OPTN*) gene, located on the short arm of chromosome 10, has also been implicated. OPTN has a role in exocytosis and Golgi ribbon formation. However, the exact pathogenesis remains elusive.

In POAG the raised intraocular pressure is attributable to abnormal resistance of the outflow system; however, to date, no significant morphological abnormality has been demonstrated within the outflow system. Nonetheless in POAG, obstruction to aqueous outflow develops progressively and the intraocular pressure gradually rises from the normal value of 18–23 mmHg to 25–35 mmHg.

The slow, progressive rise in intraocular pressure may be accompanied by occlusive disease in the posterior ciliary arteries, so that ischaemic optic atrophy may contribute to visual loss. Damage to the prelaminar optic nerve fibres may therefore be compounded by pressure-induced ischaemia in the capillary bed of the optic disk or to direct mechanical pressure preventing axoplasmic flow (see Ch. 1, p. 61) in the axons passing through the lamina cribrosa, across which a pressure differential builds up. Nerve fibre bundles passing into the optic nerve head above or below the horizontal line on the temporal side of the disk are selectively damaged and the prelaminar part of the nerve becomes atrophic (Fig. 9-23). Clinically, defects occur in the visual field (arcuate scotoma), but fibres from the macula, the papillomacular bundle, are spared. As the atrophy progresses, the cup in the optic nerve head is enlarged more extensively in the vertical plane than in the horizontal.

Primary closed-angle glaucoma. Primary closed-angle glaucoma is also the result of degenerative

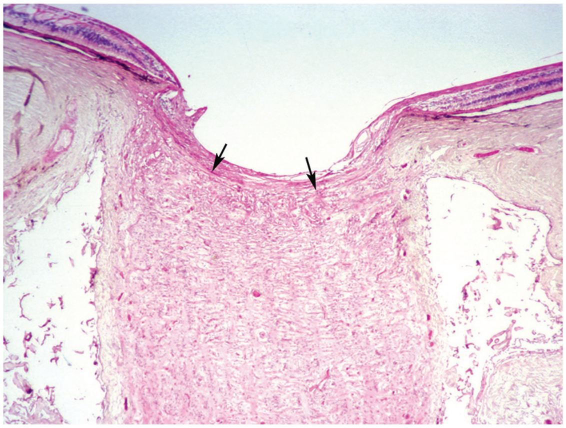

FIGURE 9-23 The optic nerve is atrophic in advanced primary open-angle glaucoma and the lamina cribrosa (arrowheads) is bowed posteriorly so that the optic disc is cupped.

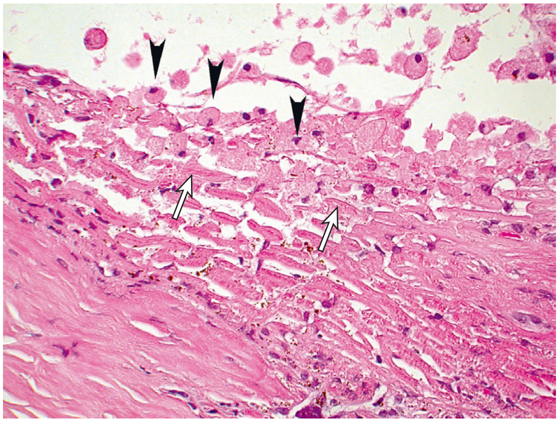

FIGURE 9-24 In phakolytic glaucoma swollen macrophages that have phagocytosed lens matter (black arrowheads) clog the trabecular meshwork (white arrows).

disease. With age, the lens thickens anteroposteriorly and in constitutively small eyes, such as in long-sighted individuals, the anterior surface of the lens displaces the pupillary part of the iris anteriorly, causing the anterior chamber to become shallow; as the angles become narrower, pressure builds up behind the iris and pushes the peripheral iris towards the trabecular meshwork. This initiates a vicious circle and pressure rises in the posterior chamber to reach levels of 40–80 mmHg. Because the pressure increase is acute, the effect on the prelaminar nerve fibres is to block axoplasmic flow so that the optic disk swells (papilloedema). Interestingly, in Asian populations a form of chronic angle closure glaucoma occurs which presents much less acutely and often painlessly, but with severe retinal nerve fibre layer damage and blindness.

Secondary glaucoma. Secondary glaucoma is also of two types: secondary open-angle and secondary closed-angle. In secondary open-angle glaucoma the angle is obstructed by cells in inflammation (uveitis), haemorrhage or tumour cell infiltration, or by lens matter when a degenerate lens capsule ruptures, flooding the anterior chamber with cortical lens matter and macrophages which have phagocytosed the lens cell fragments (Fig. 9-24).

In secondary closed-angle glaucoma, the chamber angle may be closed mechanically by anterior displacement of the lens, for instance by an intraocular tumour (e.g. uveal melanoma or retinoblastoma). A more common form of secondary closed-angle glaucoma occurs in uveitis, in which fibrin initiates adhesion formation between the peripheral iris and the trabecular meshwork (anterior synechiae), or the pupillary iris and the lens (posterior synechiae). Both can produce a rise in pressure because of obstruction to flow at the pupil (iris bombé) or in the angle. A special form of secondary glaucoma is neovascular (rubeotic) glaucoma caused by fibrovascular proliferation in the chamber angle; this is most commonly secondary to retinal ischaemia (see earlier). Rubeotic fibrovascular proliferation produces adhesions between the iris and the trabecular meshwork, which lead to a painful high-pressure glaucoma that is particularly resistant to therapy.

Congenital glaucoma in infants and children is the result of malformation of the chamber angle and failure of development in the trabecular meshwork.

Age-related macular degeneration and disciform degeneration of the macula

Age-related macular degeneration (ARMD in Box 9-14) is a disease leading to severe visual loss and blindness in the elderly population. Its pathogenesis is poorly understood and is probably multifactorial. It appears to involve a complex interaction of metabolic, functional, genetic and environmental factors.

Abnormalities are seen in the photoreceptors, retinal pigment epithelium (RPE), Bruch's membrane and choriocapillaries. The impairment of RPE cell function is an early and crucial event in the molecular pathways leading to clinically relevant AMD changes. As the RPE degenerates, there is irreversible degeneration of photoreceptors. The pathogenesis is described in Box 9-14 and pathological features in Figure 9-25.

DYSTROPHIES

In contrast to degeneration, dystrophies may occur at any age because they represent a disturbance of normal cellular functions (*dys* – altered; *trophy* – nutrition). Dystrophies encompass a range of disorders in which functional and morphological abnormalities appear in cells at various stages in life.

Retinal dystrophies

There are many disorders in which visual loss is the result of photoreceptor degeneration associated with patchy atrophy and proliferation in the RPE. A detailed account is outwith the scope of this chapter and the reader is referred to the recommended reading list available at https://expertconsult.inkling.com/.

A dystrophy may initially involve only the peripheral retina and progress later toward the macula, or may primarily involve the macula – the heredomacular degenerations. Over 100 gene mutations have been linked to one or more forms of retinal dystrophy. Interestingly, the same genetic defect can be associated with two or more phenotypically different forms of dystrophy, suggesting epigenetic effects on the disease manifestation (see Ch. 3, p. 150) (an updated list of the genetic abnormalities can be obtained on www.ncbi.nlm.nih.gov/omim).

Three examples of retinal dystrophies are provided here to demonstrate aspects of the pathology.

BOX 9-14 PATHOGENESIS OF AGE-RELATED MACULAR DEGENERATION

The pathogenesis of AMD is considered to be multifactorial, involving complex interactions of genetic, metabolic and environmental factors. At least four processes contribute to the disease:

GENETICS

- Familial component to AMD
- A single nucleotide polymorphism, common variant (Y402H) of complement factor H increases the risk of AMD (7× in homozygotes; 2–3× in heterozygotes)
- Other genes identified as potentially having a role include *ABCA4, ELOVL4, FIBL-6, APOE, SOD2.*

IMPAIRMENT OF RPE CELL FUNCTION

- RPE cells have an important metabolic and supportive function for the retina and photoreceptors
- Ageing RPE shows phagocytic and metabolic insufficiency with accumulation of lipofuscin
- Constituents of liposfuscin may contribute to free radical formation with damage to proteins, lipids and DNA and promotion of local inflammation.

SUB-RPE DEPOSIT FORMATION

- Sub-RPE deposits are commonly found in AMD
  - hard drusen –hyaline and well circumscribed; occur between the RPE basement membrane and Bruch's membrane
  - soft drusen – granular and vesicular and often more extensive; occur in same location as hard drusen
  - basal linear deposit occurs between the RPE basement membrane and RPE cell cytoplasm
- Local inflammation, in particular complement activation, may contribute to these deposits
- Photoreceptor cell death occurs overlying these deposits.

CHOROIDAL NEOVASCULARIZATION

- Increased production of extracellular matrix due to local inflammatory reaction may cause local ischaemia
- Inflammatory cells, including neutrophils and macrophages, may produce pro-angiogenic molecules such as VEGF, resulting in an imbalance in pro-angiogenic and anti-angiogenic activity.

Retinitis pigmentosa. This group of diseases mostly affects individuals in early adult life (for details of the genetic aspects see Ch. 3; p. 151). The first symptoms are night blindness and a progressive reduction in visual field from the periphery toward the posterior pole. At the end stage, retinal function is restricted to the central macular region ('tunnel vision'). Cataract is a common late complication. Fundoscopy reveals retinal atrophy, opacification and narrowing (hyalinization) of the retinal vessels, and mixed fine and coarse strands of pigmentation ('bone spicules'); the macular region is spared until the final stages.

On microscopy of advanced disease, the outer nuclear layer at the fovea appears as a single layer of cells with markedly stunted photoreceptors. Towards

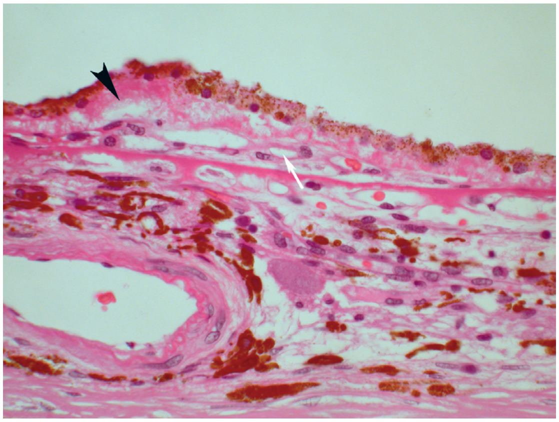

FIGURE 9-25 In age-related macular degeneration the retinal pigment epithelium overlies a basal linear deposit (arrowhead), which attracts blood vessels (arrow); this precedes the florid fibrovascular proliferation seen in disciform degeneration of the macula. The overlying retina is detached by artefact.

the periphery, the outer nuclear layer vanishes and is replaced by Müller cells, which fuse with the RPE. The RPE cells react by proliferation and migration into the retina to become distributed around the hyalinized vessels (Fig. 9-26), hence the 'bone spicules' seen on fundoscopy.

Recent reports have shown that autosomal dominant forms of retinitis pigmentosa are associated with mutations in the gene coding for rhodopsin, the rod photoreceptor pigment, which is located on the long arm of chromosome 3q and in the peripherin gene on chromosome 6p; other abnormalities which may implicate a single abnormal amino acid have been found on chromosomes 7p, 7q, 8 and 19q. If the rhodopsin molecule is abnormal, it is not difficult to appreciate that the normal process of disk replacement will be disturbed and that this will lead to photoreceptor atrophy. Other dominantly inherited forms have been associated with genes mapping to the long arm of chromosome 8 and mutations were shown in the peripherin gene (photoreceptor cell-specific glycoprotein), which is located on the short arm of chromosome 6. The loci for the X-linked forms of retinitis pigmentosa have also been identified on the short arm of the X chromosome (Xp11 and Xp21). The availability of genetic probes for the investigation of retinal degenerations has led to a massive increase in the volume of information in these disorders.

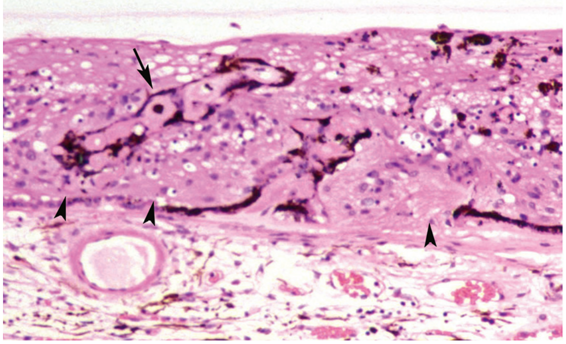

FIGURE 9-26 In retinitis pigmentosa there is extensive atrophy of the photoreceptor layer, which is replaced by glial cells (arrowheads). The retinal pigment epithelium migrates into the retina and forms clusters around hyalinized blood vessels (arrow).

Vitelliform dystrophy (Best disease). This is an autosomal dominant heredomacular degeneration in which there is loss of central visual acuity associated with a disk of yellow tissue at the macula. Histologically there is a massive accumulation of lipofuscin in the RPE cells in association with atrophy of the photoreceptor layer of the retina. Mutations in the *BEST1* and *PRPH2* genes cause vitelliform macular dystrophy. *BEST1* enodes bestrophin, a protein channel which controls chloride ion movement. *PRPH2* encodes peripherin 2, which is essential for the normal function of photoreceptor cells in the retina (Fig. 9-27).

Stargardt's disease (fundus flavimaculatus). In this condition there is atrophy of the macula in association with the appearance of small yellow flecks. At the end stage the outer layer of the retina is lost and the pigment epithelium is absent, so the gliotic retina fuses with Bruch's membrane. At the earliest stages the RPE is enlarged by accumulation of lipofuscin and melanin. The majority of cases of Stargardt's disease are caused by mutations in the *ABCA4* gene which normally transports potentially toxic substances out of the photoreceptor cells (Fig. 9-28).

Corneal dystrophies

This term includes a large group of inherited conditions that cause bilateral, slowly progressive, corneal

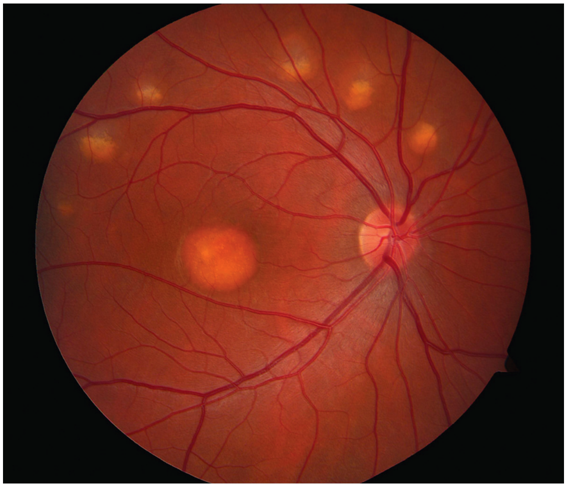

FIGURE 9-27 Fundus in Best disease. There is a classic 'egg-yolk' lesion with satellite lesions. (From Hart Moss, M.D, Eyewiki.aao.org/Best_Disease.)

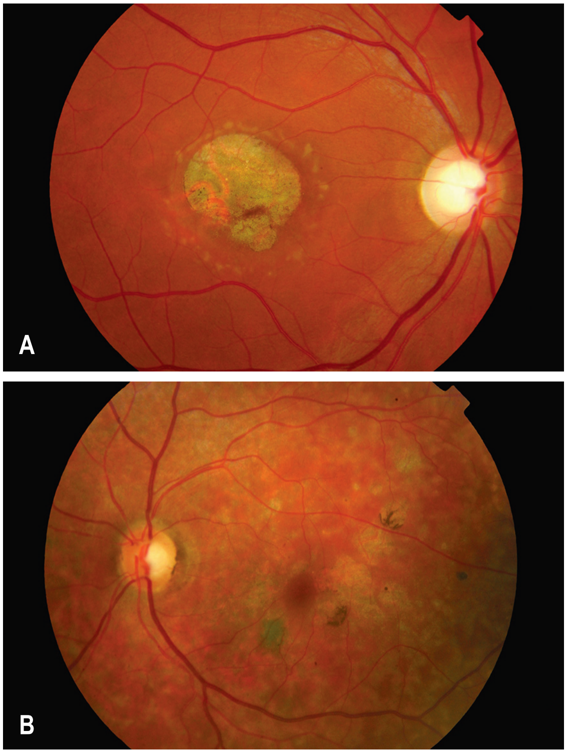

FIGURE 9-28 Fundus in Stargardt's disease (fundus flavimaculatus). There are yellow flecks in the fundus with a beaten-metal appearance at the macula owing to RPE atrophy. (Figure courtesy of Noemi Lois.)

opacification and occur in the second, third and fourth decades of life. The conditions are classified as epithelial, stromal or endothelial, although the stromal diseases may involve the other layers.

Epithelial dystrophies. Superficial dystrophies involve the epithelium, which is unable to maintain normal replication and adhesion to Bowman's layer. Degeneration of cells with cyst formation leads to an unstable epithelium in Cogan's microcystic dystrophy. Separation of cells with invagination of neighbouring cells leads to the formation of loops of basement membrane, which is a diagnostic feature of Meesman's dystrophy. These changes may also be non-specific and the diagnosis often depends on a strong family history and the clinical appearance in bilateral symmetrical disease.

Corneal dystrophies of Bowman's layer. Reis–Bücklers dystrophy is an autosomal dominant dystrophy that results in fine reticular opacities in the superficial cornea in early adult life. The histological features are not specific, with nodules of fibrous tissue between Bowman's layer and the epithelium. Electron microscopy shows characteristic electron-dense rods. Thiel–Behnke dystrophy is a histologically identical Bowman's layer dystrophy that has a honeycomb pattern on clinical examination and presents in older patients. In Thiel–Behnke dystrophy, electron microscopy shows curly fibres within the superficial fibrous nodules.

Stromal dystrophies. Although the opacities are the result of deposits of abnormal material in the stroma, there may be extensions into Bowman's layer and the endothelium may be involved. Many of the stromal dystrophies share a common genetic abnormality (Box 9-15).

Lattice dystrophy. Inherited in an autosomal dominant manner, this dystrophy is characterized clinically by fine lines criss-crossing the stroma. Microscopy shows the deposits to consist of amyloid (Fig. 9-19). Secondary non-specific amyloid deposition is sometimes seen in the cornea at the end stage of post-inflammatory scarring and fibrosis. Lattice dystrophy commonly recurs in a graft.

Macular dystrophy. In this autosomal recessive disorder, the corneal opacities take the form of smudgy

BOX 9-15 TRANSFORMING GROWTH FACTOR-β-INDUCED ASSOCIATED CORNEAL DYSTROPHIES

Many of the corneal dystrophies are associated with mutations in the transforming growth factor-β-induced gene (*BIGH1*) situated on chromosome 5q31. This encodes a protein that is expressed on the cell membrane of corneal epithelium and stromal keratocytes and plays a role in adhesion and wound healing. Mutations in this gene result in abnormal folding for the resulting proteins. Accumulation of these proteins forms amyloid or other non-fibrillar deposits. The transforming growth factor-β-induced associated corneal dystrophies include the dystrophies involving Bowman's layer (Reis–Bücklers, Thiel–Behnke) and the stromal dystrophies (granular, lattice and Avellino). All transforming growth factor-β-induced associated corneal dystrophies show autosomal dominant inheritance with complete penetrance.

'snowflake'-like areas; these are predominant in the axial region and cause severe visual impairment. This disease is a localized form of mucopolysaccharidosis. Mucopolysaccharide (acidic glycosaminoglycans) granules accumulate in the cytoplasm of the keratocytes and in the adjacent interlamellar spaces when the cells rupture. The corneal endothelial cells are involved and the material also accumulates within Desçemet's membrane and beneath the epithelium (Fig. 9-29). Macular dystrophy rarely recurs in a graft.

Granular corneal dystrophy. This is an autosomal dominant inherited disorder in which the anterior corneal stroma contains discrete opaque granules within transparent tissue. Histologically, the mid and anterior stroma and Bowman's layer contain non-birefringent hyaline bodies (Fig. 9-30) with positive staining for a keratin-like substance (so-called 'keratinoid'). The endothelium and Desçemet's membrane are not involved. Granular dystrophy occasionally recurs in a graft.

Combined granular–lattice dystrophy (Avellino dystrophy). In the original reports, the combination of stromal amyloid deposits and the classic stromal granular deposits in the superficial cornea was thought to be a feature of individuals of Italian origin. Subsequently, the disease has been found to be more widespread and the spectrum of the disease process more diverse than was originally thought.

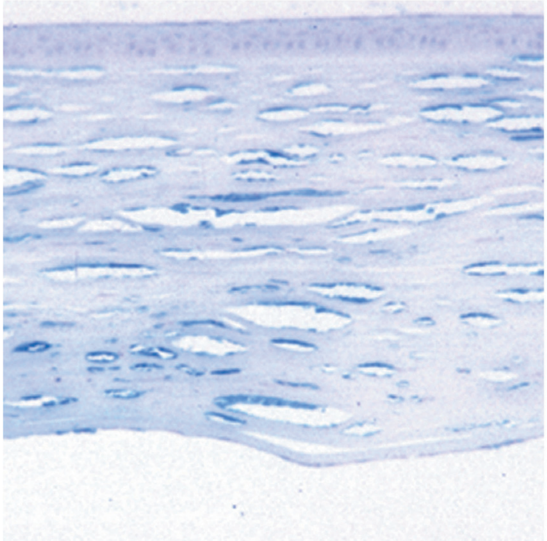

FIGURE 9-29 In macular dystrophy of the cornea, mucopolysaccharide is deposited within the keratocytes and the endothelium (Alcian blue stain).

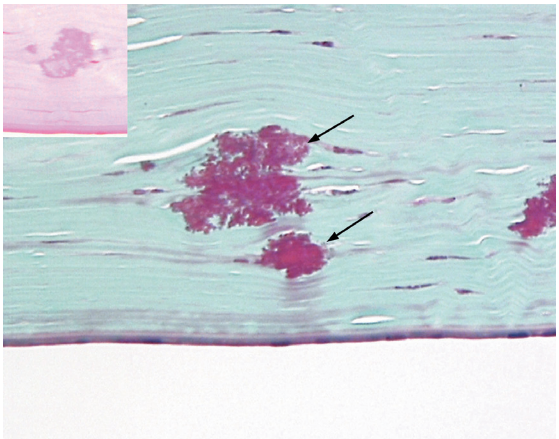

FIGURE 9-30 In granular dystrophy of the cornea, the hyaline amorphous deposits (keratinoid) stain strongly with the Masson stain and weakly in a haematoxylin and eosin-stained section (see inset).

Endothelial dystrophies. This is a group of disorders characterized by corneal oedema with opacification, occurring relatively early in life in the absence of pre-existing inflammation, glaucoma or identifiable systemic metabolic disorders. The clinical patterns are more important in the classification than the pathological findings, which often overlap.

Congenital hereditary endothelial dystrophy. A cloudy cornea in childhood or early adult life may occur in this dystrophy, which can be autosomal recessive or autosomal dominant. The autosomal recessive form is linked to mutations in *SLC4A11*. The endothelium is abnormal and may be attenuated and vacuolated, but the characteristic feature is seen in Desçemet's membrane, which shows fine lamination because of the deposit of an abnormal layer of collagen at the ultrastructural level.

Iridocorneal endothelial syndrome. This disorder is non-familial, unilateral and occurs in adults. The corneal endothelium, as studied by *in vivo* specular microscopy, reveals areas of degenerate endothelial cells, which have a bright halo around a dark spot; these areas may be surrounded by endothelial cells of normal appearance. The late outcome in the so-called iridocorneal endothelial syndrome is corneal decompensation and oedema and/or glaucoma. This corneal endothelial abnormality is seen in association with several conditions:

- progressive atrophy of the iris stroma (essential iris atrophy)
- glaucoma due to endothelial sliding across the trabecular meshwork in the presence of a normal iris (Chandler syndrome)
- the presence of an iris naevus (the iris naevus syndrome).

In this syndrome the affected endothelial cells undergo marked changes at the ultrastructural level, such as bleb formation and the acquisition of numerous surface microvilli on the posterior surface.

Posterior polymorphous dystrophy. This rare autosomal dominant disease, which causes bilateral non-progressive, circumscribed or diffuse opacities, is not usually severe enough to require keratoplasty until after the second decade. The condition is caused by mutations in the *VSX1* homeobox gene. Morphologically, in the severe diffuse form of the disease, the posterior corneal surface is lined by stratified cells with prominent desmosomal attachments resembling corneal epithelial cells (Fig. 9-31).

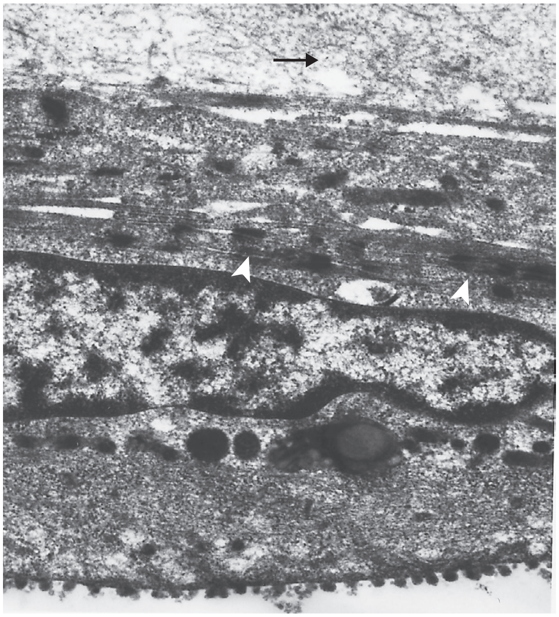

FIGURE 9-31 In posterior polymorphous dystrophy the cornea is lined on its posterior surface by cells that, by electron microscopy, exhibit all the features of epithelial cells with intracytoplasmic filaments and desmosomal attachments (arrowheads). An abnormal collagenous layer (arrow) is deposited on the posterior surface of the original Desçemet's membrane.

Fuchs' endothelial dystrophy. In this common dystrophy, elderly patients are affected, females more than males. The clinical presentation includes bilateral diffuse oedema with cloudiness of the stroma. The abnormalities are restricted to epithelial oedema and thickening of Desçemet's membrane with an obvious reduction in the endothelial cell population. Large nodular excrescences are present on the posterior surface of Desçemet's membrane (Fig. 9-32), and the endothelial cells are of varying size. When excrescences (Hassall–Henle warts) are confined to the far periphery of the normal cornea, the effect on corneal transparency is minimal.

Neoplasia

Ocular tissues, like any other tissue, are subject to metaplastic and dysplastic alterations ultimately leading in some instances to neoplastic growth. Some of the tumours and growths have counterparts in non-ocular tissues such as skin melanoma but the behaviour of a uveal melanoma in terms of life-threatening disease may be quite different from a skin melanoma or even a conjunctival melanoma. This has been attributed to the immune-privileged status of the intraocular compartment (see Ch. 7) and for

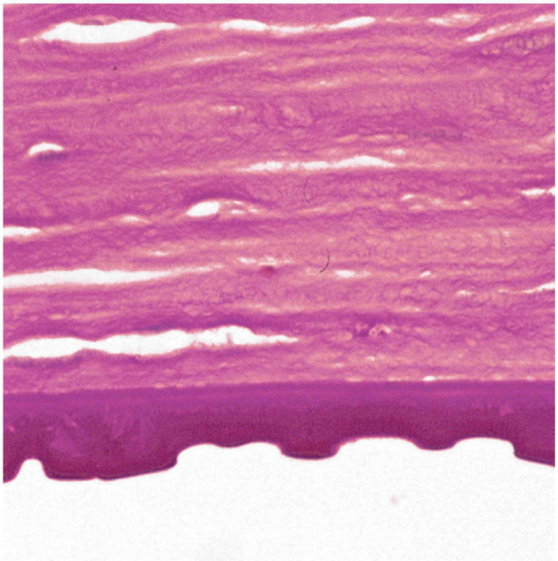

FIGURE 9-32 In Fuchs' dystrophy, Desçemet's membrane is thickened and large excrescences project from the posterior surface (PAS stain).

intraocular tumours at least there is good experimental evidence that the ocular environment modifies the immune response to tumours. In contrast, in other tumours such as squamous cell carcinoma of the lid, there may not be significant difference in the behaviour of periocular tumours compared with similar tumours at other sites.

A neoplasm is a proliferation of cells, the growth of which is progressive, purposeless, regardless of surrounding tissue, not related to the needs of the body and persists after the stimulus that initiated it has been withdrawn.

Neoplasms may be classified clinically, as benign or malignant or according to their histological tissue of origin (Table 9-2). A benign tumour is usually well circumscribed and may be encapsulated. They grow slowly and remain localized at the site of origin. They may affect the host by producing pressure on adjacent structures (e.g. proptosis secondary to pleomorphic adenoma of the lacrimal gland). Malignant tumours have an irregular, ill-defined boundary and are non-encapsulated. They grow rapidly with local and distant spread and produce effects by destroying adjacent structures (e.g. liver metastases from uveal melanoma).

PATHOGENESIS OF NEOPLASIA

Premalignancy

A number of pathological conditions are associated with the development of malignancy. The main categories include malignant transformation of benign tumours, chronic inflammatory conditions and intraepithelial neoplasia.

Benign tumours may undergo malignant transformation. A good example of this is colonic cancer arising from a benign adenoma. This is thought to occur by progressive acquisition of genetic changes. Malignant transformation of benign tumours also occurs in ophthalmic pathology. For example, adenocarcinoma may arise in a long-standing pleomorphic adenoma of the lacrimal gland. Chronic inflammatory conditions, particularly if they are very long-standing, may promote transformation of 'stressed tissues' to malignancy; this link is currently an area of intense research. For instance, in Sjögren syndrome there is chronic lymphocytic infiltration of the lacrimal gland with acinar atrophy later leading to the clinical symptoms of dry eye. Evolution to lymphoma occurs in a significant number of patients with Sjögren syndrome, possibly by the development of monoclonal lymphocytic populations within the lacrimal gland. Intraepithelial neoplasia represents an intermediate stage in the production of cancer. In the skin, excessive exposure to ultraviolet light may lead to development of an actinic or solar keratosis. Clinically, these appear as hyperkeratotic lesions on the face and histological examination reveals premalignant changes in the epidermis. This is seen as an increased mitotic rate, a loss of the normal polarity of maturation from basal cells to squamous cells, and a marked variation in the size and shape of nuclei (pleomorphism) within the epithelium (dysplasia). These histological changes precede invasion through the basement membrane of the epithelium into the underlying tissue, and are therefore designated carcinoma *in situ*.

Carcinogenesis

Both environmental and genetic factors contribute to a cell undergoing malignant change. This should be regarded as a multi-step process. The three major environmental factors that induce tumours

TABLE 9-2 **Histological classification of tumours**

Histological origin	**Benign**	**Malignant**
Epithelial cells		
Surface	Papilloma	Carcinoma (squamous, basal cell, etc.)
Glandular	Adenoma	Adenocarcinoma
Mesenchymal cells		
Adipose	Lipoma	Liposarcoma
Fibrous	Fibroma	Fibrosarcoma
Cartilage	Chondroma	Chondrosarcoma
Bone	Osteoma	Osteosarcoma
Smooth muscle	Leiomyoma	Leiomyosarcoma
Striated muscle	Rhabdomyoma	Rhabdomyosarcoma
Neuroectodermal cells		
Glial cells	Nerve	Glioma
	Ganglioneuroma	Neuroblastoma
Retinal cells		Retinoblastoma
Melanocytes		Melanoma
Meninges	Meningioma	Malignant meningioma
Schwann cells	Neurofibroma	Malignant peripheral nerve sheath tumour
Haemopoietic/lymphoreticular		Leukaemia
		Lymphoma
Germ cells	Benign teratoma	Malignant teratoma
		Dysgerminoma Seminoma

Tumours may arise from any tissue in the body but for convenience these are divided into five groups. Not all malignant tumours have a benign counterpart and similarly there are some types of benign tumour for which malignant counterparts are extremely rare.

are chemicals, radiation and viruses. In chemical carcinogenesis the first step, initiation, involves a short exposure of the cell to a carcinogen. This is followed by promotion, the long-term exposure to a substance that is usually not mutagenic but acts by stimulating cell proliferation (although some compounds can act as both initiators and promoters, so-called complete carcinogens). Ionizing radiation directly damages DNA, especially during cell proliferation, and can result in a range of changes from single gene mutations to major chromosome deletions. Ultraviolet radiation mainly affects the skin-forming pyrimidine dimers that can usually be excised by DNA repair mechanisms. In xeroderma pigmentosum these are deficient and multiple skin tumours occur. Viruses may contribute to the development of some human cancers. Oncogenic viruses may contribute to several conditions in the eye. Conjunctival papillomas and papillomas of the lacrimal passages may be caused by human papillomavirus types 16 and 11, respectively. Epstein–Barr virus contributes to the development of orbital Burkitt's lymphoma and to intraocular diffuse large B-cell lymphoma occurring in the immunosuppressed.

A genetic influence in cancer is now well recognized. Certain syndromes inherited in a Mendelian fashion show a high risk of cancer. Examples of these include xeroderma pigmentosum, an autosomal recessive trait where failure of DNA repair leads to skin cancer, and neurofibromatosis, an autosomal dominant trait characterized by multiple neurofibromas and increased risk of sarcoma, which is the result of a defect of the *NF1* gene on chromosome 17. In Li–Fraumeni syndrome there is a high risk of several types of cancer, such as childhood sarcomas and breast cancer in young women. This is the result of a germline mutation of the *p53* gene.

Oncogenes and tumour suppressor genes

Cellular proto-oncogenes are normal genes that stimulate cell division. Tumour suppressor genes are normal genes that inhibit cell division. Proto-oncogenes and tumour suppressor genes are active during somatic growth, regeneration and repair and the balance between stimulation and inhibition of cell growth is strictly controlled. This balance is permanently lost in cancer cells.

Cellular proto-oncogenes code for a number of proteins involved in cell proliferation, including growth factors, growth factor receptors, signal transducers within the cell cytoplasm and nuclear-regulating proteins. In cancer these normal genes are permanently changed to oncogenes and proliferation is uncontrolled. Proto-oncogenes may become oncogenes by mutation, resulting in the production of a functionally abnormal protein or overexpression.

Tumour suppressor genes are normal genes that switch off cell proliferation. Loss of both copies of a tumour suppressor gene is required for cancer to develop. Loss of the *Rb* gene is important in the development of retinoblastoma (Ch. 3).

Tumour spread and metastases

Malignant tumours spread by several routes:

- local invasion of normal tissue (e.g. basal cell carcinoma)
- lymphatic spread (e.g. squamous carcinoma of the eyelid) or haematogenous spread (e.g. malignant melanoma of the choroid)
- intraepithelial spread (e.g. Pagetoid spread of sebaceous carcinoma of the eyelid)
- dissemination along natural passages (e.g. retinoblastoma extending to subarachnoid space; bronchial carcinoma spreading to pleura; ovarian carcinoma involving peritoneum).

The basic mechanisms of tumour cell invasion involve several mechanisms:

- tumour cells secrete lytic enzymes to breach the basement membrane
- there is a loss of cell–cell adhesion molecules, often accompanied by an increase in cell–matrix adhesion molecules
- increased cell movement allows tumour cells to penetrate further and spread.

HAMARTOMAS

A hamartoma is a tumour-like but non-neoplastic malformation consisting of a mixture of tissues normally found at a particular site. The commonest forms of hamartoma are those composed of blood vessels and those involving melanocytes of the skin.

Haemangiomas

Capillary haemangiomas are a proliferation of small-calibre vascular channels with a lobulated growth pattern. Cavernous haemangiomas consist of large-calibre thick-walled vascular channels with intervening fibrous septae. Both capillary and cavernous haemangiomas may occur in the eyelid, orbit or choroid. Extensive haemangiomas may occur as part of encephalo-trigeminal angiomatosis (Sturge–Weber syndrome). Capillary haemangiomas may regress spontaneously during childhood but cavernous haemangiomas show no tendency for spontaneous regression.

Naevi

The word ‘naevus’ means a birthmark, but most naevi are acquired during childhood and adolescence. Melanocytes are of neural crest origin and migrate through the dermis to reach epithelial cells. A naevus is the result of abnormal migration, proliferation and maturation of these neuroectodermal cells. They may occur in the uvea as well as the skin (Fig. 9-33). In the skin the melanocytes initially form clumps at the junction between the epidermis and dermis. Clinically this appears as a brown macule, referred to as a junctional naevus. With age, the proliferating melanocytes begin to detach from the epithelium and migrate into the dermis, forming a brown papule. When proliferation is found in the dermis as well as the junctional area the naevus is classified as compound. At a later stage the proliferation is wholly in the dermis and is classified as an intradermal naevus (Fig. 9-34).

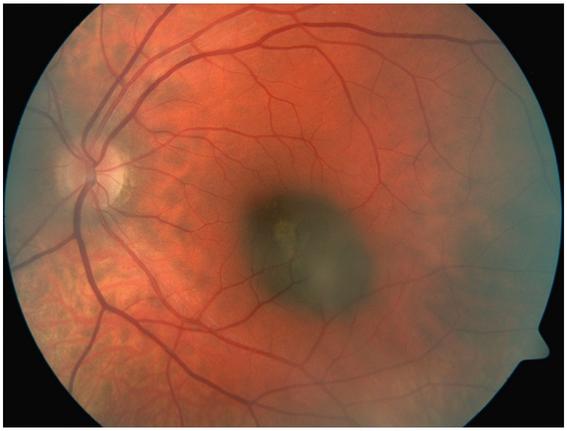

FIGURE 9-33 Fundoscopic image of a choroidal naevus showing a well-circumscribed, slightly elevated, pigmented tumour. Distinguishing features from melanoma are the clearly defined margins and lack of overlying subretinal fluid and orange pigment. Over time there will be minimal or no documented growth. (Image courtesy of Dr Paul Cauchi, Glasgow.)

Naevi, similar to their cutaneous counterpart, occur in the conjunctiva. In the iris and choroid, naevi are seen as static flat brown or black areas. Naevi in any site may occasionally progress to malignant melanoma.

CHORISTOMAS

In contrast with hamartoma, a choristoma is a tumour-like but non-neoplastic malformation consisting of a mixture of tissues not normally present at a particular site.

Dermoid

Epibulbar dermoids are relatively common choristomas. They occur as a nodule (smooth white swellings from which hairs project) on the bulbar conjunctiva in children or at the outer angle of the bony orbit on the skin. Histological examination reveals a mixture of fat, fibrous tissue, hair follicles and sweat glands.

Phakomatous choristoma

This is a rare lesion presenting as a nodule in the eyelid. It is composed of epithelial cells and basement membrane, resembling lens capsular material, set in a dense fibrous stroma.

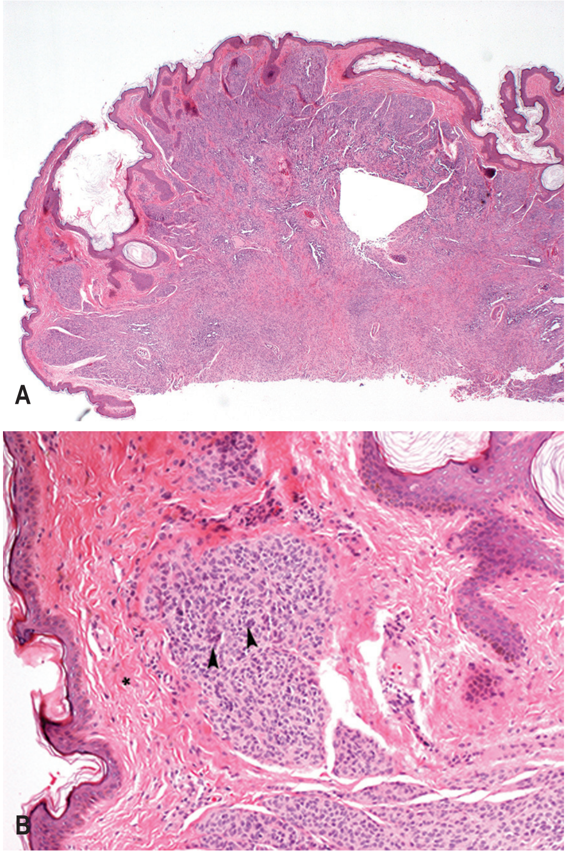

FIGURE 9-34 In an intradermal naevus all the naevus cells lie within the dermis and clinically this often appears as a polypoid or warty nodule. The higher power in (**B**) shows that there is a gap (*) between the naevus cells and the epidermis. The cells have small bland nuclei and are sometimes multinucleated (arrowheads).

TERATOMA

This is a tumour derived from totipotent germ cells. They can occur at any site in the mid-line where germ cells have stopped on their migration to the gonads. Orbital teratomas are rare and occur in neonates. An orbital teratoma causes proptosis and histological examination of the large cystic retroocular mass will reveal tissue derived from the three embryonic germ cell layers such as respiratory or gastrointestinal epithelium, stroma containing fat, cartilage and bone, and neuroectodermal tissues. Most orbital teratomas are benign and surgical removal is curative.

BENIGN EPITHELIAL TUMOURS

Benign tumours of surface epithelium

A papilloma is a benign tumour originating from an epithelial surface. In the eyelid the commonest tumours are basal cell papilloma (seborrhoeic keratosis) and squamous cell papillomas. The former retains the basaloid appearance of the basal cells of the normal epidermis, whereas the latter shows features of squamous differentiation. Benign squamous proliferations may be associated with poxvirus (molluscum contagiosum) or human papillomavirus (viral wart). Conjunctival papillomas can be pedunculated or sessile. The pedunculated papillomas are usually covered by conjunctival epithelium, whereas sessile papillomas commonly show squamous differentiation. Conjunctival papillomas are also commonly associated with human papillomavirus.

Benign tumours of adnexal glands

An adenoma is derived from the ducts and acini of glands. In the eyelid and caruncle these may be derived from sweat glands, pilosebaceous hair follicles and sebaceous glands, the largest being the meibomian gland in the tarsal plate. Sweat gland adenomas are subclassified according to the degree of differentiation towards acinar or ductular structures. Similarly, tumours of hair follicles are classified according to differentiation towards different components of the hair follicle. For example, a pilomatrixoma shows differentiation towards hair matrix. Sebaceous adenomas are proliferations of lipid-laden sebaceous cells and most commonly occur as a yellow mass at the caruncle.

MALIGNANT EPITHELIAL TUMOURS

Basal cell carcinoma

Basal cell carcinoma is the most common malignant tumour in clinical ophthalmology, accounting for more than 90% of malignant eyelid tumours. It usually occurs in Caucasians over 50 years of age and is associated with sunlight exposure. It may also occur in younger patients in association with the Gorlin–Goltz syndrome (basal cell naevus syndrome). Clinically, these tumours may present as nodular lesions, which later may develop a central ulcer with a rolled edge. Morpheiform types present as a scirrhous plaque. Occasionally, basal cell carcinomas are pigmented because of melanin deposition and clinically may be confused with malignant melanoma. Basal cell carcinoma is locally aggressive and adequate surgical excision is the treatment of choice to prevent recurrence and orbital invasion. Orbital invasion may necessitate exenteration.

Four main histological subtypes should be recognized:

- *Nodular/solid basal cell carcinoma* – consists of well-circumscribed and relatively large islands of proliferating basal cells (Fig. 9-35A). Mitotic figures are usually plentiful. At the periphery of the tumour cell islands the cells are arranged as a palisade. Cystic degeneration may occur in this subtype (nodulocystic). Surgical excision should not present problems.
- *Superficial basal cell carcinoma* – is less common than the nodular subtype and presents as a scaly plaque. Histology shows small nests of tumour cells budding from the undersurface of the epidermis only as far as the superficial dermis. There may be substantial gaps between the nests of tumour cells such that complete surgical excision may be difficult.
- *The infiltrative/sclerosing subtype* – this corresponds to the morpheic clinical subtype and is a more aggressive form of basal cell carcinoma where the tumour cells grow in small strands rather than nests and are embedded in a dense fibrous stroma (Fig. 9-35B). Peripheral palisading is much less pronounced than in the other subtypes. Infiltrative basal cell carcinoma does not have a distinct border, making adequate surgical excision difficult.
- *The micronodular subtype* – is also a more aggressive form. The tumour forms small nodular aggregates of basaloid cells and, similar to the infiltrative subtype, subclinical involvement is often significant.

Squamous cell carcinoma

Compared with basal cell carcinoma, the incidence of this form of malignancy is low, representing between 1% and 5% of all eyelid cancers. Risk factors

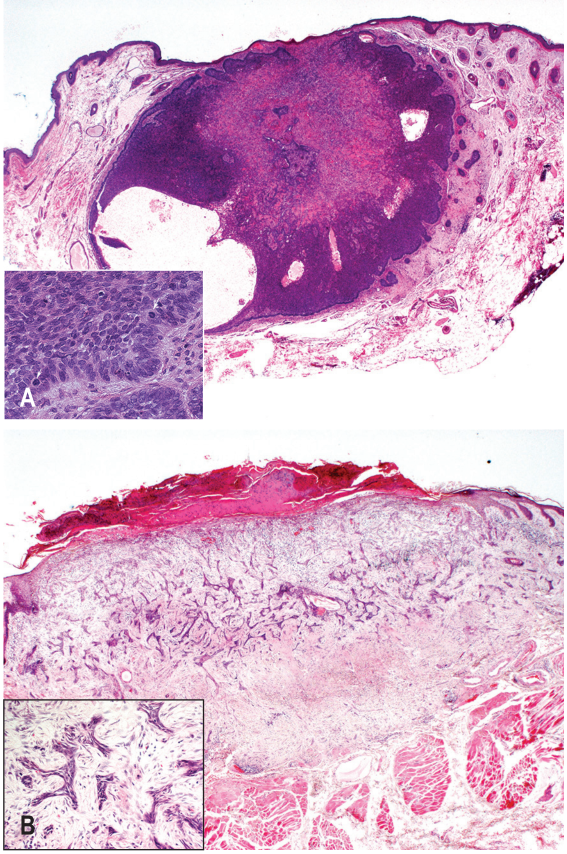

FIGURE 9-35 (**A**) A nodular basal cell carcinoma has a well-defined margin. Cystic degeneration is common. The inset shows the characteristic peripheral palisading at the edge of tumour cell lobules. (**B**) The margins of a sclerosing basal cell carcinoma are poorly defined and small strands of tumour can invade deeply without being clinically apparent. The jagged outline of the strands of tumour cells are shown in the inset. Peripheral palisading of nuclei is not prominent in this type of basal cell carcinoma. The fibroblastic intervening tissue is known as a desmoplastic stroma.

for squamous cell carcinoma include sunlight exposure and immunosuppression. Clinically, squamous cell carcinoma presents as a rapidly growing nodular ulcer or as a papillomatous growth, which in some cases has an overlying keratinous horn. Inadequate primary local excision may be followed by recurrence and orbital invasion. Lymphatic spread may occur to pre-auricular and submandibular lymph nodes according to the site of origin – upper and lower lid, respectively.

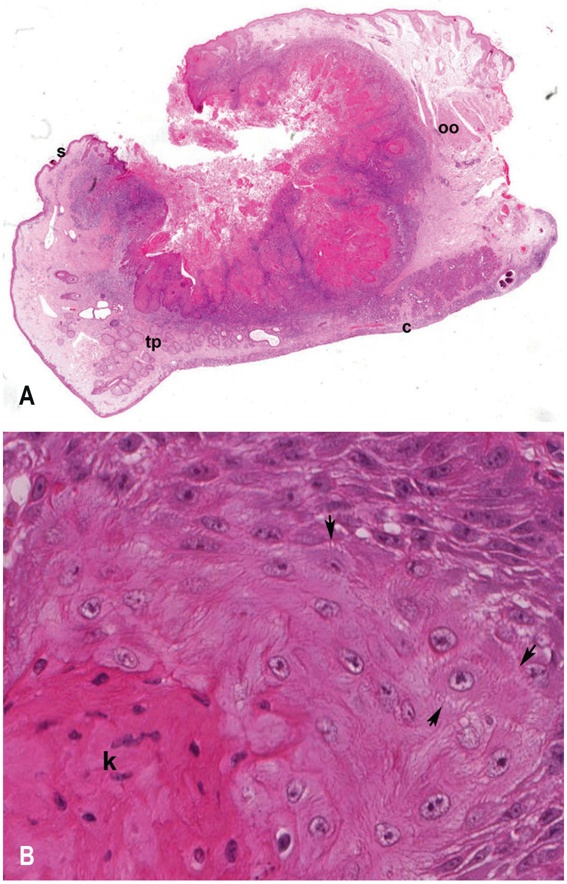

FIGURE 9-36 (**A**) This squamous cell carcinoma of the eyelid infiltrates extensively from the skin surface (s) through orbicularis oculi (oo) and just into the tarsal plate (tp). The conjunctival surface (c) is not involved). (**B**) The cells show keratinization (k) and intercellular bridges (arrows).

Histologically, squamous cell carcinoma may be classified as well, moderately or poorly differentiated. In a well-differentiated tumour the cells have glassy pink cytoplasm and intercellular bridges and keratin pearls may be present (Fig. 9-36). Some of these features are lost in more poorly differentiated tumours, but intercellular bridges can usually still be identified. Rarely, squamous cell carcinoma will adopt a spindle cell morphology and this variant is more aggressive.

In situ and invasive squamous cell carcinoma may also involve the conjunctiva and cornea. The morphology of these tumours is identical to that of the eyelid tumours and they are also associated with sunlight exposure and immunosuppression, particularly AIDS.

Sebaceous gland carcinoma

Sebaceous gland carcinoma accounts for 1–5% of all eyelid cancers. These tumours usually originate in the meibomian gland but may also arise from the gland of Zeis or sebaceous glands of the eyelid skin. Sebaceous gland carcinoma commonly occurs in elderly patients and shows a female preponderance. The clinical appearance of sebaceous gland carcinoma is variable and it may be indistinguishable from squamous cell carcinoma or basal cell carcinoma or may mimic a range of benign conditions, including chalazion and blepharoconjunctivitis.

The variable clinical appearance is related to the different histological growth patterns of this tumour. These tumours may show a nodular or diffuse pattern of growth and may be well, moderately or poorly differentiated. The nodular pattern consists of lobules of tumour cells with foamy or vacuolated cytoplasm (Fig. 9-37A). Diffuse tumours show individual tumour cells spreading within the surface epithelium (Pagetoid spread) and adnexal structures. Stains for fat performed on frozen section can be helpful (Fig. 9-37B). Immunohistochemical staining which is positive for BerEP4, epithelial membrane antigen and androgen receptors is more helpful in small biopsy specimens which cannot be divided.

Sebaceous carcinomas may be seen in Muir–Torre syndrome in association with a visceral malignancy. In these tumours there may be loss of heterozygosity of the chromosomal regions containing the mismatch repair genes *hMSH2* and *hMLH1*. The prognosis is poor compared with most other malignant eyelid tumours but is significantly improved with early diagnosis and surgery.

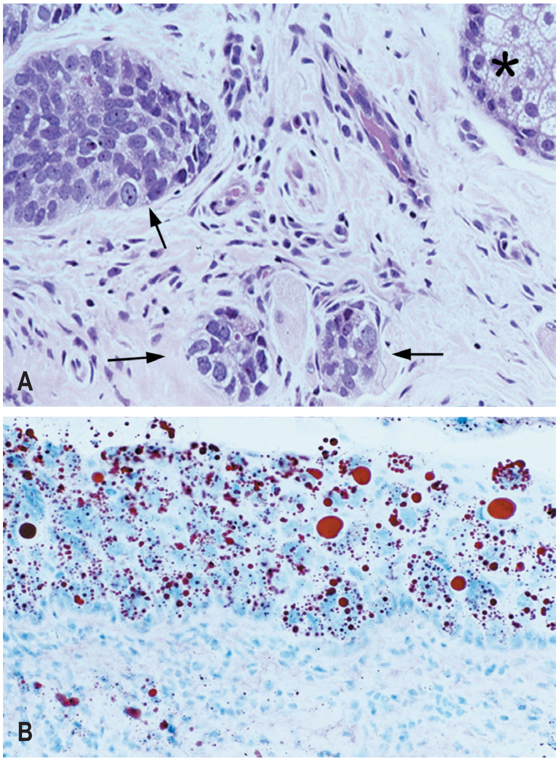

FIGURE 9-37 (**A**) The islands (arrows) of infiltrating sebaceous gland carcinoma contain cells with pale cytoplasm that bear some resemblance to normal sebaceous gland cells (*). (**B**) A stain for fat (Oil red O) reveals malignant cells infiltrating the conjunctival epithelium in Pagetoid spread of a sebaceous carcinoma.

MALIGNANT MELANOMA

Conjunctiva

Malignant melanoma may arise from primary acquired melanosis, a pre-existing naevus or *de novo*. Primary acquired melanosis (also known as conjunctival intraepithelial melanocytic neoplasia) appears as unilateral or bilateral, diffuse flat areas of conjunctival pigmentation in middle-aged to elderly patients. Conjunctival melanoma presents as a raised, pigmented or fleshy conjunctival lesion. Like its cutaneous counterpart (and in contrast with uveal melanoma), it has a tendency to metastasize to regional lymph nodes but may also spread to brain and other organs. The prognosis is worse for tumours thicker than 5 mm and located in the fornix. Complete excision is the treatment of choice. Those arising on a background of primary acquired melanosis may be multifocal and topical chemotherapy with mitomycin C can be helpful.

Uveal melanoma

Malignant melanoma of the uveal tract arises from melanocytes in the iris, ciliary body and choroid, and the relative incidence is roughly in proportion to the volume of tissue in each compartment – 8%, 12% and 80%, respectively. These tumours are almost always unilateral and grow initially as pigmented or non-pigmented plaque-like lesions; the macroscopic appearances are shown in Figure 9-38 (A–D).

Iris melanomas. Iris melanomas are usually slow-growing nodular tumours that may be present for many years. On histology, iris melanomas may consist of small, rather bland, spindle-shaped cells and the diagnosis depends on the identification of surface or stromal invasion. Recurrent iris melanomas often transform into pleomorphic epithelioid tumours. Although they often remain localized for a long period of time, iris melanomas can spread diffusely on the iris surface and around the chamber angle, resulting in secondary glaucoma as a result of infiltration of the trabecular meshwork.

Ciliary body and choroidal melanomas. Ciliary body and choroidal melanomas can grow to a large size (10–20 mm) before recognition. The macroscopic appearances can vary considerably. The tumours may be ovoid, nodular or a classical mushroom shape. This is caused by tumour spread in the subretinal space after breaching Bruch's membrane. Tumours may cause an exudative retinal detachment, or acute angle closure glaucoma due to direct pressure effects or neovascular glaucoma due to production of vasoformative factors by the tumour. Extraocular extension may be identified in relation to collector channels (anteriorly), vortex veins (in the mid-periphery) or short ciliary vessels (posteriorly) and occasionally tumours may present as proptosis due to an orbital mass. Larger tumours may undergo spontaneous necrosis and produce symptoms of endophthalmitis. Uveal melanoma can be simulated clinically by various other entities. Some typical examples are shown in Figure 9-39 (A–D).

On histology, the tumours are classified according to cell type as spindle, epithelioid and mixed (Fig. 9-40). In practice the majority of tumours are of mixed cell type. Vascular patterns may be assessed in melanomas using a periodic acid–Schiff (PAS) stain. There are nine recognized patterns, including parallel, parallel with cross-linking and a network of closed vascular loops. The presence of microscopic intrascleral spread or extraocular extension is also important for staging uveal melanoma. Metastatic spread (most commonly to the liver, *cf* conjunctival melanoma) usually occurs within 2–3 years but has also been recorded up to 40 years later. Immunohistochemistry, which is usually positive for S100, HMB45 and Melan A, can be a useful ancillary technique, particularly in metastatic melanoma.

Treatment includes enucleation, local resection and ionizing radiation in the form of plaque brachytherapy or proton beam. Secondary enucleation following surgery or irradiation may show a surgical coloboma or neovascular glaucoma, cataract and radiation retinopathy in addition to recurrent tumour (see Video 9-2).

Prognostic parameters in uveal melanoma include:

- *age of patient* – the prognosis is worse for older patients
- *tumour size* – larger tumours carry a worse prognosis
- *tumour location* – ciliary body location carries a worse prognosis compared with choroid
- *cell type* – tumours containing an epithelioid cell component carry a poorer prognosis than those composed only of spindle cells
- *vascular patterns* – tumours with a closed loop vascular pattern on Periodic Acid Schiff stain carry a poorer prognosis
- *cytogenetics* – (Box 9-16) loss of heterozygosity of chromosome 3 (monosomy 3), particularly when combined with additional copies of chromosome 8q is strongly associated with death from metastases, whereas aberrations, particularly numerical gain, of chromosome 6p are associated with a more favourable prognosis.

NEURAL TUMOURS

Neurofibroma and schwannoma

These tumours arise within the orbit and are derived from peripheral nerves. Neurofibroma is derived from the endoneurium and schwannomas from the Schwann cells intimately surrounding axons. On histological examination neurofibromas consist of spindle cells with wavy nuclei and collagen, with occasional axons running through the tumour. Neurofibromas,

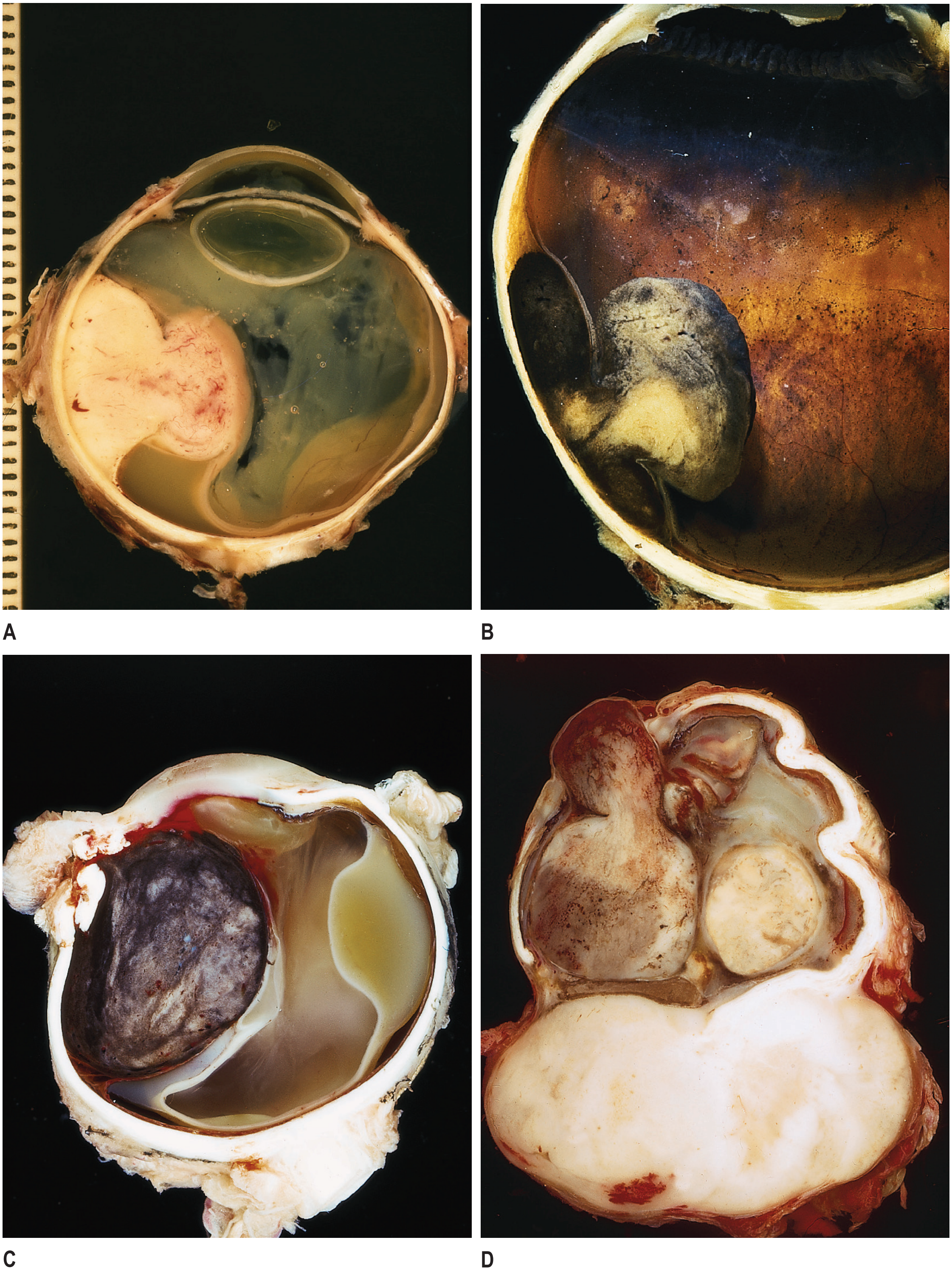

FIGURE 9-38 Various macroscopic appearances of uveal melanoma. (**A**) The majority of tumours are amelanotic and have a mushroom shape. (**B**) A partially pigmented melanoma that has perforated the retina. (**C**) This ovoid black melanoma has leaked fluid into the subretinal space, causing an exudative retinal detachment. An attempt to remove the tumour surgically was abandoned. (**D**) An advanced melanoma perforating the anterior and posterior sclera.

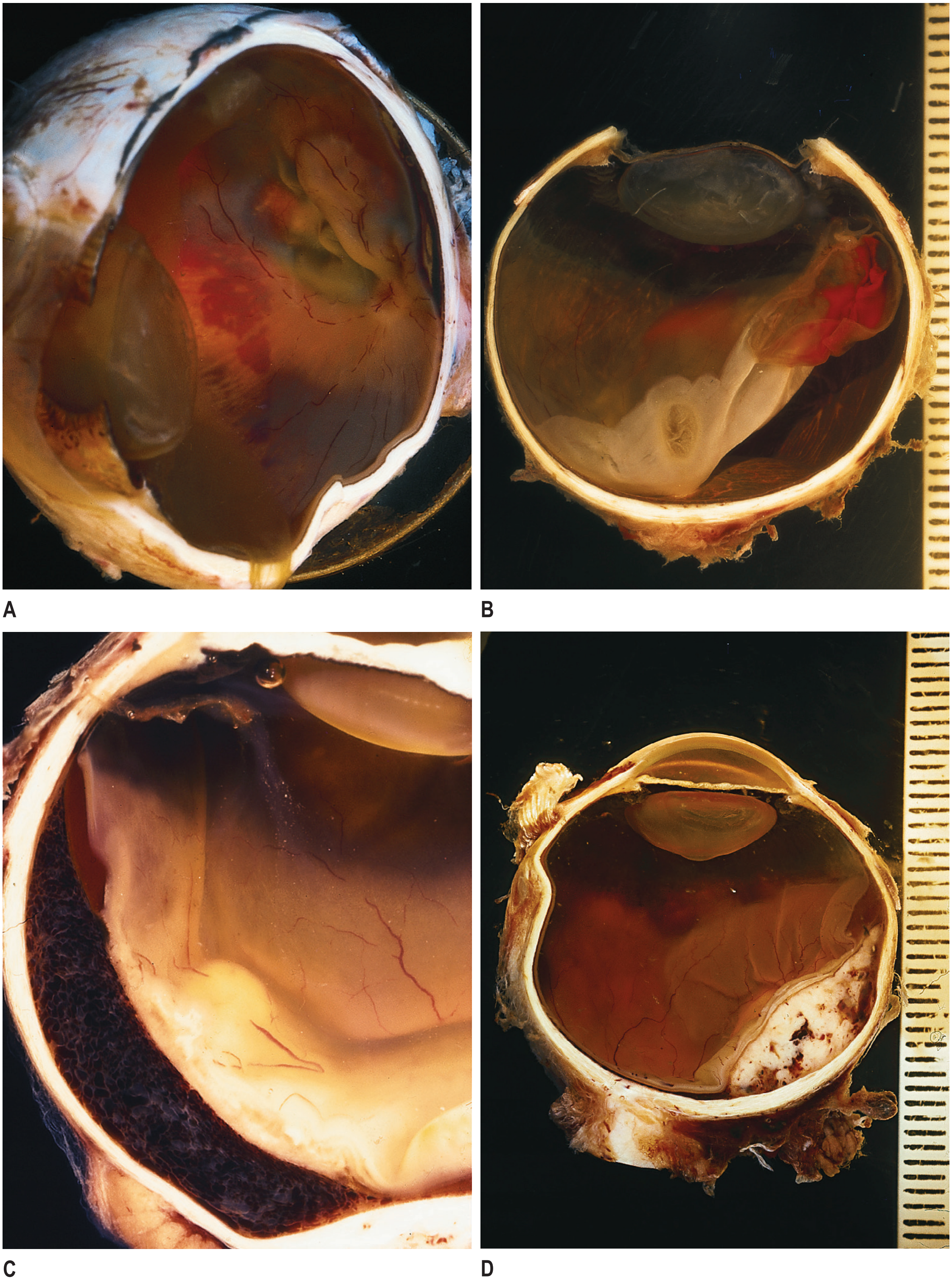

FIGURE 9-39 Diseases simulating uveal melanomas. (**A**) Blood arising from disciform degeneration of the macula. (**B**) Bleeding into a macrocyst arising from a peripheral microcystoid degeneration of the retina. (**C**) Angioma of the choroid. (**D**) Metastatic tumours from breast or lung.

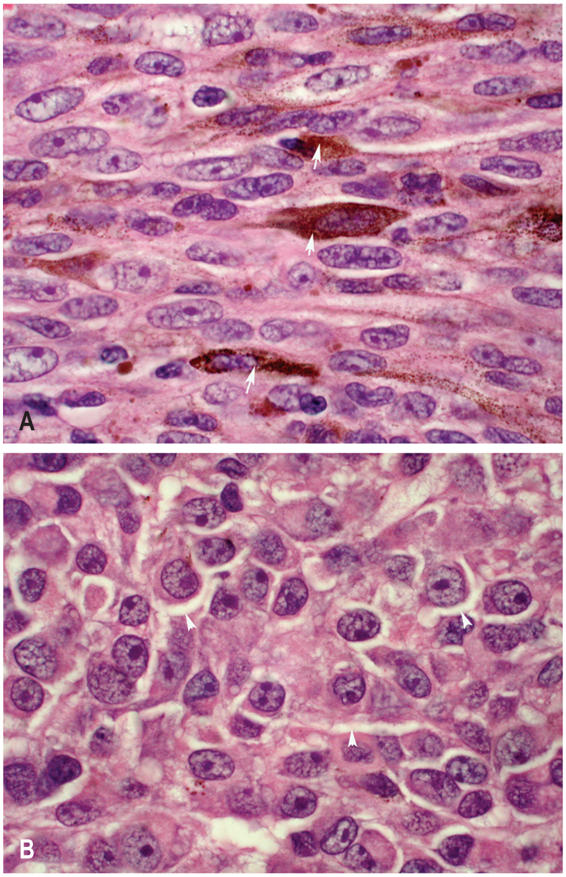

FIGURE 9-40 (**A**) In this malignant uveal melanoma of spindle B type the cytoplasm of the cells contains melanin granules (arrows). (**B**) Epithelioid melanoma cells are larger than spindle B cells and are separated from each other by prominent intercellular spaces (arrowheads).

particularly the plexiform and diffuse subtypes, may be associated with neurofibromatosis type 1. Schwannomas show a palisaded arrangement of spindle cells (Antoni A) and myxoid (Antoni B) areas and, in contrast to neurofibromas, there are occasional axons in the peripheral part of the tumour. Degenerative changes, with thick-walled blood vessels with evidence of previous haemorrhage and atypical nuclei, are relatively common. Occasionally a schwannoma may contain melanin pigment and the differential diagnosis of extraocular extension of a spindle cell melanoma should be considered.

BOX 9-16 TUMOUR CYTOGENETICS

Numerical and structural chromosomal abnormalities occur in tumour cells compared with normal cells. Chromosomal abnormalities within a tumour cell can fall into one of three categories:

- *Primary abnormality* – essential for establishing tumorigenesis, e.g. *Rb1* gene in retinoblastoma
- *Secondary abnormality* – a manifestation of tumour progression and clonal evolution, e.g. loss of *p53* in colonic adenoma–carcinoma sequence
- *Cytogenetic noise* – this is due to genetic instability in any tumour.

CYTOGENETICS IN UVEAL MELANOMA

- Loss of heterozygosity of chromosome 3 (monosomy 3) ± additional copies of chromosome 8q is strongly associated with death from metastases
- Aberrations, particularly numerical gain of chromosome 6p, is associated with a more favourable prognosis.

Malignant peripheral nerve sheath tumours

Malignant peripheral nerve sheath tumours are rare in the orbit. Most arise *de novo* without previous evidence of a neurofibroma or schwannoma. They may be associated with neurofibromatosis type 1.

Retinoblastoma

Retinoblastoma is a malignant tumour of infancy which is lethal if untreated; the incidence is 1 in 20 000 live births. The tumour arises from embryonal retinal cells and can be unilateral or bilateral. The macroscopic appearances are of a smooth-surfaced white mass that may show endophytic growth, into the vitreous or exophytic growth into the subretinal space (Fig. 9-41A–D). Yellowish areas of necrosis or bright white flecks of calcification may be evident within the tumour. On histology the tumour consists of small blue cells with scanty cytoplasm. There is usually a high mitotic rate, with prominent apoptosis and areas of necrosis within the tumour indicating high cell turnover. DNA from necrotic tumour may precipitate on blood vessel walls or as basophilic lakes. Differentiation may be seen in the form of:

- Homer–Wright rosettes – a multilayered circle of nuclei surrounding eosinophilic fibrillar material (Fig. 9-42A)

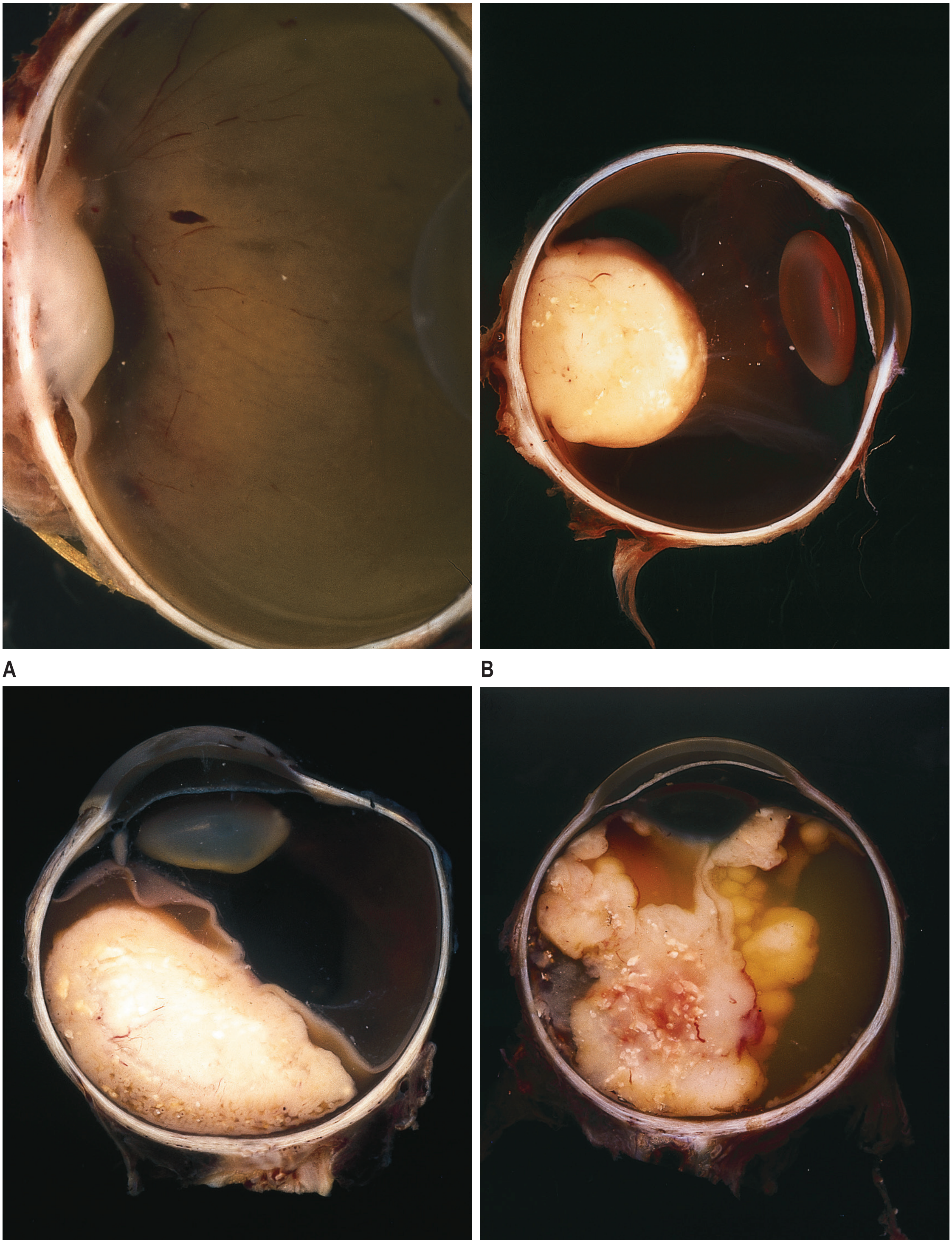

FIGURE 9-41 Macroscopic appearances of retinoblastoma. (**A**) A small retinoblastoma overlying the optic nerve head contains a few flecks of calcium. (**B**) A large retinoblastoma with seedlings in the vitreous. (**C**) An exophytic retinoblastoma detaching the retina. (**D**) A large exophytic retinoblastoma with prominent calcification and funnel-shaped retinal detachment.

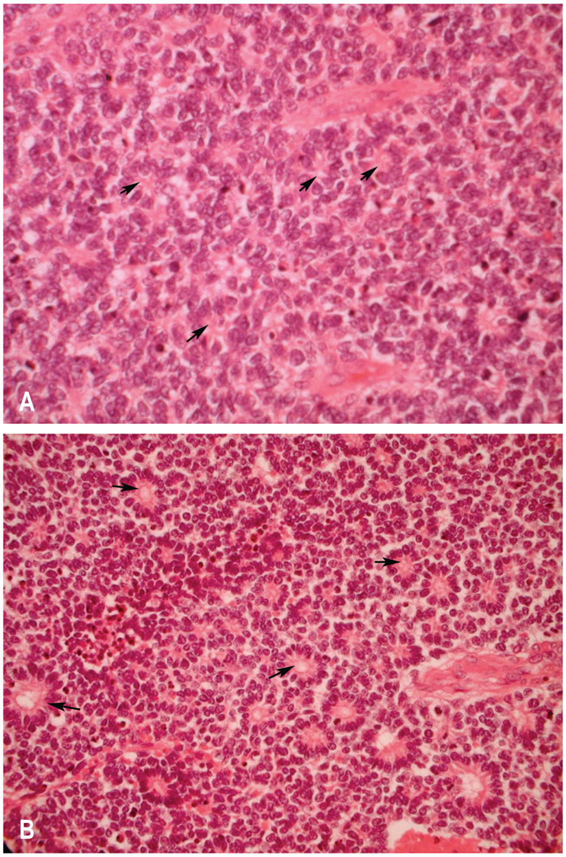

FIGURE 9-42 Differentiation in retinoblastoma is represented by various forms of rosette. (**A**) Homer–Wright rosettes (arrows) consist of a central fibrillary tangle surrounded by cell nuclei which may overlap. (**B**) Flexner–Wintersteiner rosettes are lined internally by a membrane similar to the outer limiting membrane of the normal retina (arrows) and surrounded by a single layer of cells.

- Flexner–Wintersteiner rosettes – a circle of cells limited internally by a continuous membrane (Fig. 9-42B)
- Fleurettes – primitive photoreceptor bodies arranged in 'fleur de lys' shape. These structures are most commonly found in irradiated tumours.

Features of prognostic importance in retinoblastoma include tumour size, degree of differentiation, choroidal invasion and optic nerve invasion. With early diagnosis and modern treatment, including irradiation and chemotherapy, cure rates are in excess of 90%. In untreated cases, death is caused by tumour spread to the brain through the optic nerve or along the meninges; metastatic dissemination is via the bloodstream to the viscera and skeleton.

The genetics of retinoblastoma are discussed in Chapter 3, but it is noteworthy that the abnormal gene carries the risk of a pineal tumour in childhood (trilateral retinoblastoma), soft tissue and osteogenic sarcoma in early adult life and carcinomas in later life.

The differential diagnosis of retinoblastoma includes:

- Coats' disease (Fig. 9-43A)
- Astrocytic hamartoma (Fig. 9-43B)
- Retinopathy of prematurity
- Persistent hyperplastic primary vitreous (Fig. 9-43C)
- Endophthalmitis (Fig. 9-43D)
- *Toxocara* retinitis.

Astrocytic hamartoma

Benign astrocytic tumours occur in the retina as part of the tuberous sclerosis syndrome or as an isolated feature. They consist of astrocytes, which form a matrix conducive to the deposition of calcospherites. The presence of calcification may lead to an erroneous diagnosis of retinoblastoma.

Glioma

Juvenile and adult forms of optic nerve glioma are recognized, the former carrying a good prognosis; the latter are very rare and invariably lethal, being associated with extensive intracranial extension.

Around 50% of gliomas involve the orbital portion of the nerve but the intracranial or chiasmal portions may also be involved. In the orbital portion the tumour may cause proptosis in addition to optic disc swelling and visual loss. Computed tomography or magnetic resonance imaging may be helpful in delineating the location, configuration and extent of the tumour. In over 50% of patients the tumour does not grow but in the remaining cases the tumour does grow and may require surgical intervention. The affected region of the nerve may be excised or if there is extensive tumour with secondary complications, such as exposure keratitis, then the eye may be removed along with the affected segment of the nerve. Optic nerve gliomas have an excellent prognosis following complete surgical excision although vision is usually sacrificed.

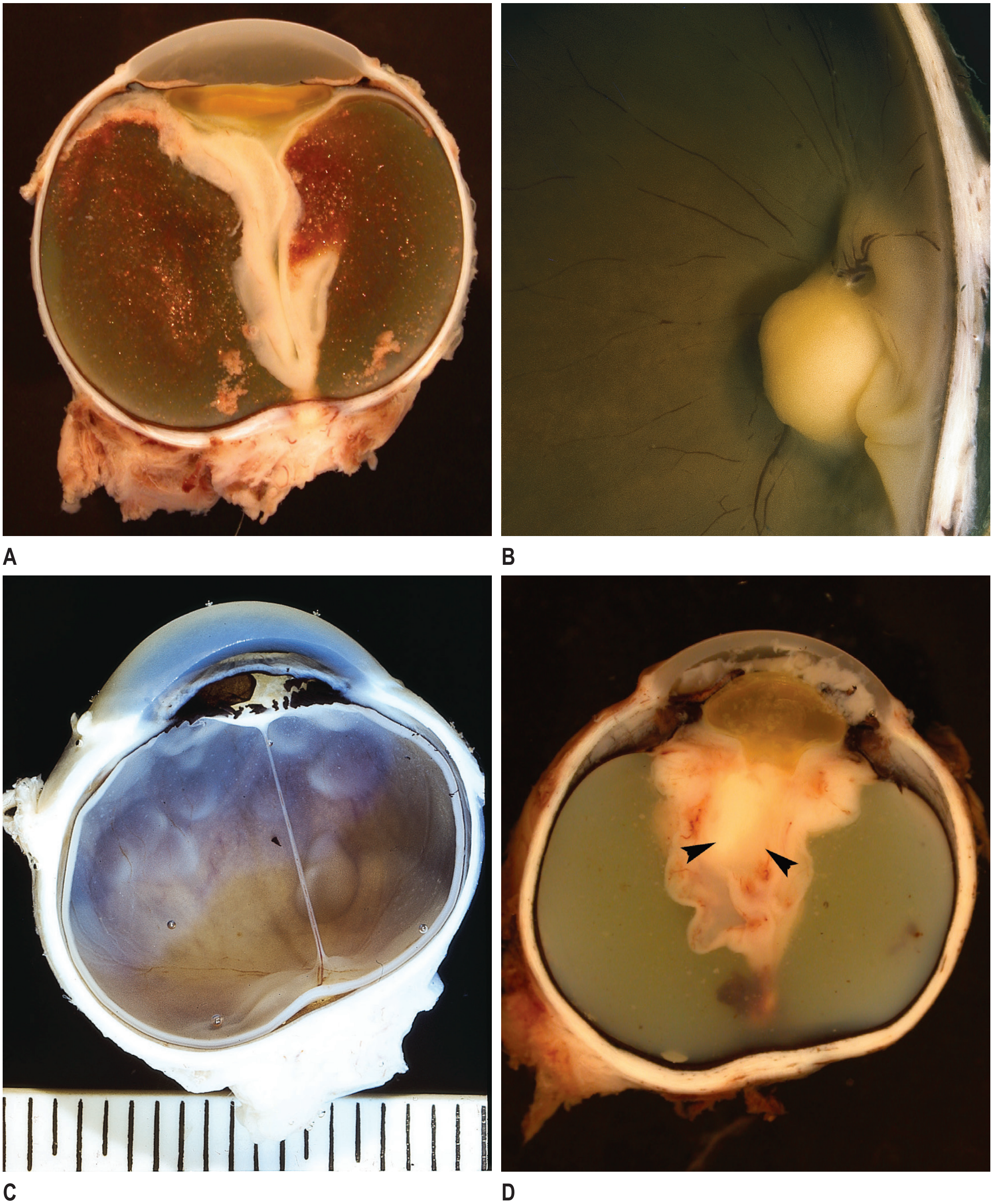

FIGURE 9-43 Diseases simulating retinoblastoma. (**A**) In Coats' disease the abnormal vasculature leaks lipid-rich plasma, and cholesterol crystals are present in the subretinal exudate. (**B**) Astrocytic hamartoma appears as a static round nodule projecting from the retina into the vitreous. (**C**) Persistent hyperplastic vitreous forms a white mass behind the lens and the persistent hyaloid artery passes back to the optic nerve head. (**D**) In metastatic endophthalmitis an abscess fills the vitreous cavity, forming a white mass (arrowheads), and the retina is detached.

Excised tumours show a fusiform swelling of the nerve and the residual nerve may be barely visible within the tumour mass. The histology of these tumours is identical to that of intracranial astrocytomas and the majority are pilocytic, often containing areas of myxoid degeneration and Rosenthal fibres (Fig. 9-44). A potential diagnostic pitfall is that these tumours can induce proliferation of the overlying arachnoid. This hyperplastic tissue may be misdiagnosed as meningioma if the biopsy contains only perineural tissue.

Meningioma

Meningioma of the optic nerve may be primary, arising from the meninges of the optic nerve, or secondary, as the result of extension of an intracranial meningioma. In adults meningiomas of the optic nerve characteristically show indolent growth but in children they may be more aggressive. The tumour ensheaths the optic nerve, which may become atrophic. The histology is similar to intracranial meningioma, with a transitional pattern, sometimes with psammoma bodies, predominating.

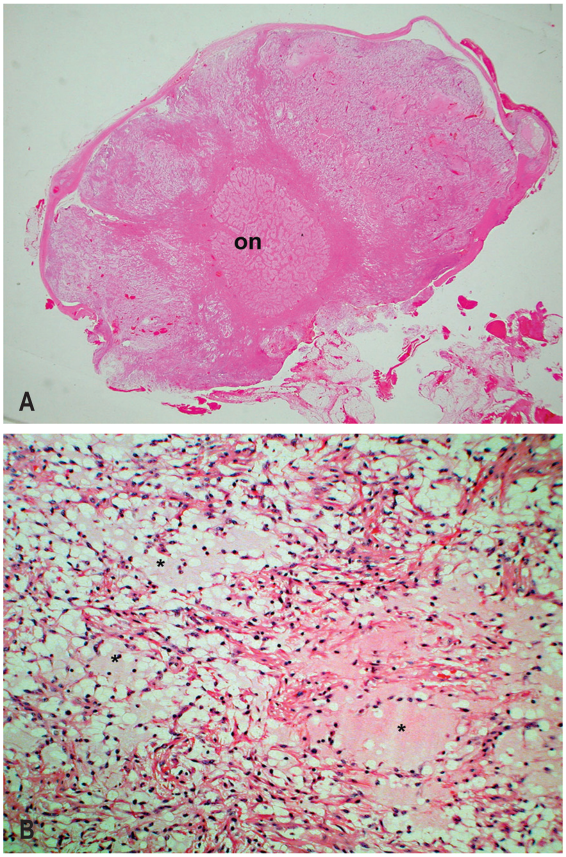

FIGURE 9-44 (**A**) Low-power view of an optic glioma that has infiltrated the nerve columns but has preserved some recognizable architecture in the optic nerve (on). (**B**) At higher magnification the tumour consists of irregularly arranged proliferating astrocytes with numerous small cystic spaces (*).

TUMOURS DERIVED FROM MUSCLE

These rare tumours may affect the eye or the orbit.

Leiomyoma/leiomyosarcoma

A leiomyoma occasionally arises from the smooth muscle of the iris and ciliary body. The malignant counterpart, leiomyosarcoma, is extremely rare.

Rhabdomyoma/rhabdomyosarcoma

Benign tumours of striated muscle are virtually unknown in the eyelid and orbit.

Rhabdomyosarcoma is the most common orbital malignancy of childhood. It generally occurs in the first two decades of life and usually presents with rapidly progressive proptosis and displacement of the eye. If clinical suspicions are high, a prompt biopsy should be performed to confirm the diagnosis and the patient should be treated with a combination of chemotherapy and radiotherapy. With this regimen the survival of children with rhabdomyosarcoma has dramatically improved. On macroscopic examination, these tumours consist of tan-coloured fleshy tissue. On histopathological examination, rhabdomyosarcoma can be divided into three subtypes (embryonal, alveolar and pleomorphic). The embryonal subtype is the most common type in the orbit, consisting of sheets of small ovoid to spindle-shaped cells (Fig. 9-45). Cytoplasmic cross-striations can be seen with difficulty in a small number of cases. Immunohistochemistry for *MyoD1*, a muscle regulatory gene, may be helpful in confirming the diagnosis. Alveolar rhabdomyosarcoma is more common in older children; pleomorphic rhabdomyosarcoma is rare in the orbit and usually occurs in adults.

VASCULAR TUMOURS

These tumours are also rare but occur at all age groups.

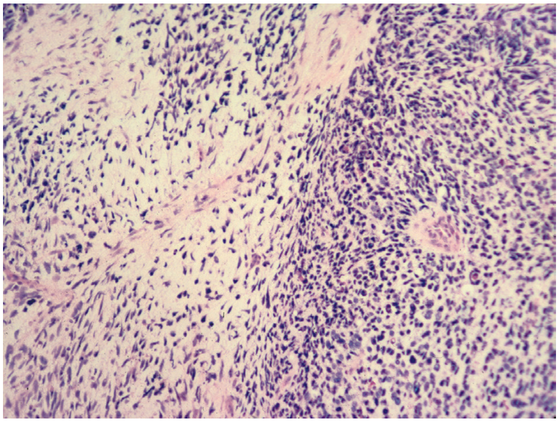

FIGURE 9-45 In embryonal rhabdomyosarcoma, the cells are ovoid to spindle-shaped and there are alternating cellular and myxoid areas.

Haemangiomas

These are described under Hamartomas.

Epithelioid haemangioma

Epithelioid haemangioma, previously named angiolymphoid hyperplasia with eosinophilia, is a benign vascular lesion that may occur on the eyelid skin or occasionally in the orbit. It consists of blood vessels with prominent endothelial cells and accompanying inflammatory cells, including lymphoid follicles and prominent eosinophils.

Kaposi sarcoma

This is a tumour of endothelial cells that may occur as a rapidly growing tumour on the eyelid and conjunctiva. It most commonly occurs in immunocompromised patients, especially those with AIDS, and is caused by infection with herpes virus type 8. Histologically, these tumours consist of malignant spindle cells lining a network of sieve-like spaces containing extravasated red cells.

OTHER CONNECTIVE TISSUE TUMOURS

Tumours can be derived from any of the cellular constituents of connective tissue, including adipose tissue, cartilage and bone. Orbital lipomas are considered to be relatively rare but this may reflect underdiagnosis because of the difficulty of distinguishing a lipoma from excised orbital fat. Primary liposarcoma rarely involves the orbit. Similarly, cartilaginous neoplasms of the orbit are extremely rare. Orbital osteosarcoma is also rare but well recognized as a second primary neoplasm following successfully treated retinoblastoma.

Solitary fibrous tumour, haemangiopericytoma and giant cell angiofibroma are a group of related neoplasms that, compared with other connective tissue tumours, are relatively common in the orbit. Due to a recently identified common chromosomal translocation (inv(12)(q13q13)) which results in a NAB2-STAT6 fusion product these tumours are now regarded as the same entity and the term solitary fibrous tumour is preferred. Histologically, they all consist of spindle cells with a thin-walled branching vascular pattern. Solitary fibrous tumour shows a range of cellularity with the more cellular tumours being designated cellular solitary fibrous tumour (formerly haemangiopericytoma). Some tumours contain stromal giant cells (formerly giant cell angiofibroma). The majority of these tumour follow a benign course but the behaviour is unpredictable. Larger and more cellular tumours are more likely to recur and metastasize.

LYMPHOID TUMOURS

The tissues behind the orbital septum contain neither lymphatics nor lymphoid tissue. Lymphocytes may, however, be found in the conjunctiva, the lacrimal gland and the lacrimal drainage system.

Lymphomas of the ocular adnexa include lesions of the conjunctiva, eyelids, lacrimal gland and orbit. Those situated in the conjunctiva are associated with a lower incidence of systemic disease (20%) compared with those of the orbit (35%), lacrimal gland (40%) or eyelid (67%). Ocular lymphomas may be the first manifestation of disseminated disease and it is essential to undertake a full systemic and haematological examination of all patients presenting with ocular lymphoma. Immunohistochemistry (Fig. 9-46 and Box 9-17), molecular techniques (see Boxes 9-6 and 9-7) and flow cytometry are important for identifying monoclonal populations and subclassifying lymphoma. Flow cytometry is used to isolate and phenotype different cells within a mixed population. Cells in suspension are labelled with fluorescent markers, excited by a laser and counted electronically by passing them through a flow cytometer: this may be undertaken on vitreous samples.

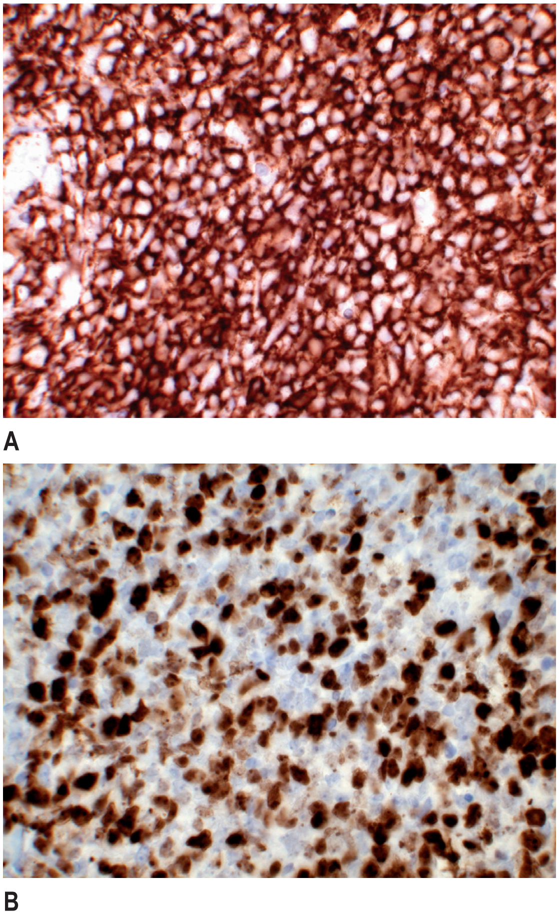

FIGURE 9-46 (**A**) Immunohistochemistry for CD20 shows membranous staining of the cells in this B-cell lymphoma. (**B**) Immunohistochemistry for MIB-1 (proliferating cell nuclear antigen) shows widespread nuclear staining indicating a high proliferation fraction in this high-grade B-cell lymphoma (diffuse large B-cell lymphoma).

BOX 9-17 IMMUNOHISTOCHEMISTRY

Immunohistochemistry is a method of detecting the presence of specific proteins in cells or tissues. It consists of the following steps:

- Antigen retrieval either by proteolytic digestion or microwaving the tissue sections.
- A primary antibody binds to a specific antigen.
- A secondary enzyme-conjugated antibody is then bound to this primary antibody–antigen complex.
- An appropriate substrate and chromagen are added and the enzyme catalyses the formation of a coloured deposit at the sites of antibody–antigen binding.

Immunohistochemistry is important in:

- *Tumour classification* – lymphomas are positive for common leucocyte antigen (CLA), carcinomas are positive for cytokeratins (CK), melanoma is positive for S100 and Melan A.
- *Specific tumour subtyping* – lymphomas may be subtyped into B cell (CD20 positive) and T cell (CD3 positive); carcinoma of the lung is usually TTF-1 positive, carcinoma of the gastrointestinal tract is usually CDX-2 positive.
- *Providing prognostic or therapeutic information* – lymphomas that have a high proliferation index with MIB-1 (see Fig. 9-46) are aggressive. Breast cancers that are oestrogen-receptor positive may be treated with hormonal therapy.
- *Diagnosis of infectious agents* – antibodies are available to various infectious agents such as herpes simplex virus and cytomegalovirus.

With these techniques, the vast majority of lymphoid proliferations can be classified as benign (reactive lymphoid hyperplasia) or malignant (lymphoma) and terms such as ‘pseudolymphoma’ should be avoided.

The most common ocular lymphoproliferative lesions include:

- Benign lymphoid hyperplasia, a similar process to reactive follicular hyperplasia of lymph nodes may form a tumour mass within the conjunctiva or orbit.
- Extranodal marginal zone lymphoma (EMZL) is the most common type of ocular lymphoma and is a low-grade B-cell lymphoma derived from the mucosal associated lymphoid tissue. EMZL usually follows an indolent course but it may recur at other extranodal sites or rarely undergo transformation to high-grade lymphoma.
- Follicular lymphoma is identical to its nodal counterpart and in the majority of cases represents part of a systemic disease (see Video 9-3).
- Diffuse large B-cell lymphoma is less common and around 40% of cases are associated with systemic disease. These lymphomas tend to pursue an aggressive clinical course.
- Primary intraocular lymphoma involves the retina, subretinal space, vitreous and optic nerve (Fig. 9-47). It can occur in conjunction with or independent of primary central nervous system lymphoma. These are rare lymphomas but there

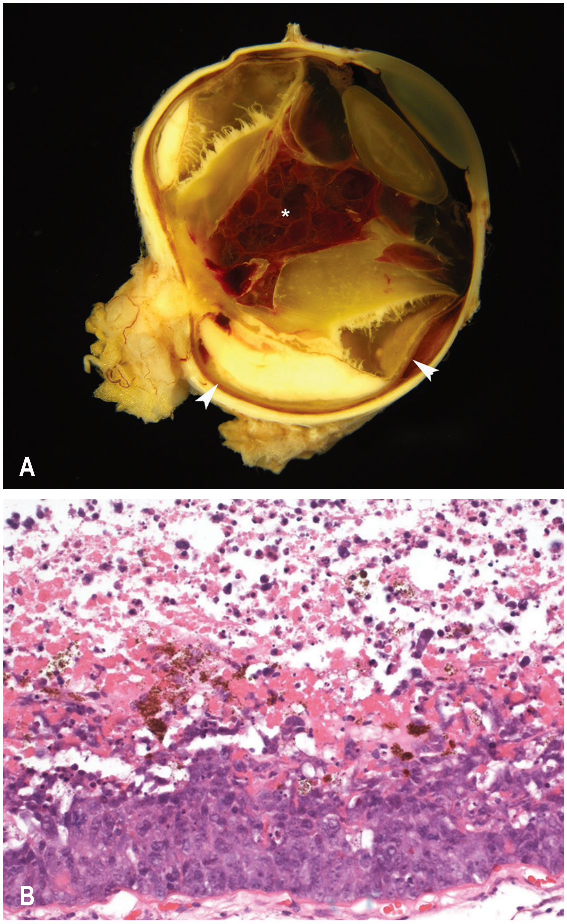

FIGURE 9-47 (**A**) An eye removed for primary intraocular lymphoma. There is haemorrhage and exudate in the vitreous (*). There is creamy white tumour overlying the retina (arrowheads). (**B**) On histology the retina is replaced by pleomorphic large lymphoma cells, with necrosis and apoptosis extending into the vitreous.

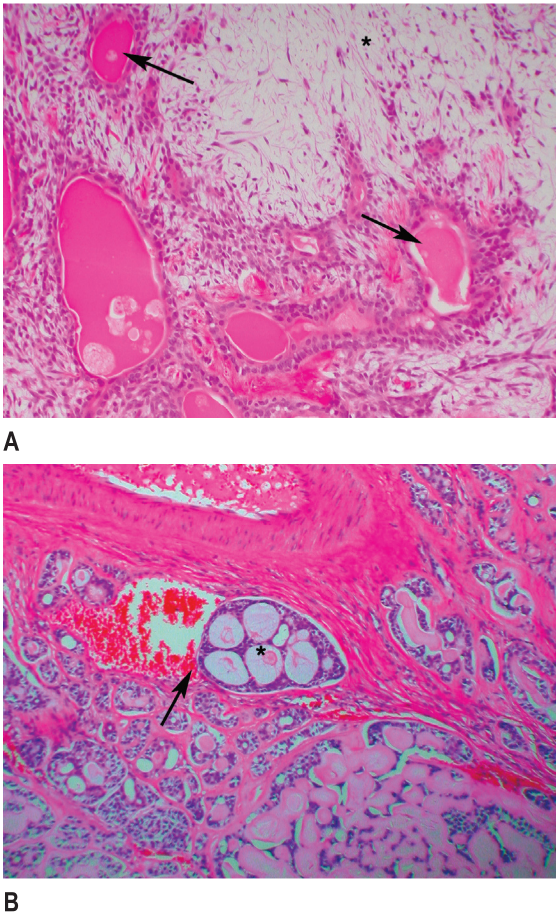

FIGURE 9-48 (**A**) Pleomorphic adenoma consists of benign branching glands (arrows) lying within a myxoid stroma (*). (**B**) Adenoid cystic carcinoma has a Swiss-cheese pattern (*) and may show vascular (arrow) and perineural invasion.

has been a dramatic increase in incidence in recent years. The majority are diffuse large B-cell lymphomas. They usually occur in elderly patients but can also be associated with AIDS. Many other lymphomas may uncommonly involve the ocular region, including mantle cell lymphoma, B-cell chronic lymphocytic leukaemia, Burkitt's lymphoma, peripheral T-cell lymphoma and natural killer cell lymphoma. Ocular Hodgkin's disease is very rare. Leukaemic infiltration of the eye or orbit may also occur. In particular, granulocytic sarcoma can present as an isolated orbital mass in an otherwise healthy child.

LACRIMAL GLAND TUMOURS

Pleomorphic adenoma (benign mixed tumour)

Pleomorphic adenoma is the most common epithelial tumour of the lacrimal gland. It usually occurs in late to middle age but may occur at any age. This tumour is slow growing and pseudoencapsulated with surface bosselations. Histologically, it consists of a mixture of epithelial and mesenchymal elements, including myxoid tissue, cartilage, fat and, rarely, bone (Fig. 9-48A). It is important to completely excise these tumours because if they are 'shelled out' they can recur and there is a small but significant risk of surviving residual tumour undergoing malignant change to a pleomorphic carcinoma (malignant mixed tumour).

Adenoid cystic carcinoma

This is the second most common epithelial neoplasm of the lacrimal gland after pleomorphic adenoma. Although it is usually diagnosed in middle-aged or older patients, it frequently occurs in younger patients as well. The history is shorter than for pleomorphic adenoma and the patient may present with proptosis, numbness, pain and diplopia because invasion of nerves and extraocular muscles may occur early in tumour development. Histologically, these tumours can assume a range of patterns, the most common of which is a cribriform or 'Swiss-cheese' pattern (Fig. 9-48B). The pain and numbness may be explained by the tumour's propensity for perineural infiltration. These are aggressive neoplasms that require radical surgery with supplemental radiotherapy or chemotherapy.

Other malignant epithelial tumours

A small number of epithelial tumours of the lacrimal gland are adenocarcinomas arising *de novo* with no evidence of a pre-existing benign mixed tumour. Mucoepidermoid carcinoma is another rare form of carcinoma that may arise in the lacrimal gland.

Lacrimal sac tumours

Tumours of the lacrimal sac are uncommon and are usually of epithelial origin. Papillomas may show an exophytic, inverted or mixed growth pattern and the epithelium may be of squamous or transitional cell type. Carcinoma of the lacrimal sac may develop within a papilloma or arise *de novo*. These are locally aggressive tumours and if neglected can invade surrounding structures.

METASTATIC TUMOURS

In adults, metastatic tumours most commonly involve the uveal tract (Fig. 9-39D). Orbital involvement occurs about one-tenth as often, and metastases to the eyelid and conjunctiva are rare. Most metastatic tumours are carcinomas and the most common primary sites are usually the breast, prostate, lung or gastrointestinal tract, but metastases from a wide range of primary carcinomas have been described. Immunohistochemistry and novel molecular techniques (see Box 9-17) can be useful in identifying the primary site. In children metastatic disease is usually orbital and uveal involvement is rare. Orbital involvement by neuroblastoma, Ewing's sarcoma, Wilms' tumour and rhabdomyosarcoma may occur.

DISORDERS MISDIAGNOSED AS NEOPLASMS

There are several 'lumps and bumps' which occur around the eye and orbit which can be mistaken for tumours.

Cysts

Eyelid. Simple cysts are common in the eyelid:

- *Sudoriferous cysts* or sweat gland cysts (hidrocystomas) are derived from the ducts of the glands of Moll. The cysts are thin-walled and appear as translucent or bluish swellings at the lid margin. Histologically, they are lined by a double layer of epithelium with an inner layer of cuboidal cells and an outer layer of myoepithelial cells.
- *Epidermoid cysts* may occur secondary to obstruction of the duct of a pilosebaceous follicle or as the result of epithelial inclusion following trauma or surgery. Epidermoid cysts are lined by keratinizing squamous epithelium and are filled with keratin. A foreign body giant cell reaction may be seen in relation to cyst rupture.
- *Dermoid cysts* occur in children as a result of the incarceration of ectoderm between the frontal and maxillary process during embryogenesis. Dermoid cysts contain hairs and pilosebaceous follicles.

Conjunctiva. Cystic lesions are also common in the conjunctiva. Epithelial inclusion cysts are the result of previous trauma with incarceration of conjunctival epithelial nests in the stroma. Lymphatic cysts occur when ectatic lymphatics coalesce. This lymphangiectasia may occur as a hamartomatous malformation or secondary to inflammation.

Orbit. Cysts may also occur in the orbit and computed tomography scans may be helpful in identifying some lesions. Dermoid cysts may also occur in the orbit and should be excised intact because leaked contents may induce a granulomatous inflammatory reaction. A mucocoele is an expansion of the paranasal space secondary to drainage obstruction from chronic sinusitis. It consists of a cystic cavity lined with epithelium, which sometimes contains goblet cells. Haematic cyst is an organizing haematoma, which can occur spontaneously, or following blunt trauma.

Pseudoepitheliomatous hyperplasia

Surface epithelium overlying an inflamed stroma or tumour can be stimulated by various released growth factors. This may result in an exuberant proliferation of the epithelium that can be mistaken for squamous carcinoma by the unwary. This benign reactive process has been termed pseudoepitheliomatous hyperplasia.

Idiopathic orbital inflammation

Idiopathic orbital inflammation (formerly inflammatory pseudotumour) is a non-granulomatous inflammatory process within the orbit for which there is no recognized local cause or any underlying systemic disease. The disease presents as a unilateral or bilateral mass, which clinically may be mistaken for tumour. In biopsy specimens an early lesion shows oedema of orbital tissues and an inflammatory infiltrate composed predominantly of lymphocytes and plasma cells and lymphoid follicles may be present. As the disease progresses, collagen is laid down and the collections of inflammatory cells may be separated by fibrous tissue (Fig. 9-49). Most cases show a dramatic response to corticosteroid therapy unless the lesion has extensive fibrosis. Other immunosuppressive agents, such as azathioprine, or low-dose radiotherapy may be used in patients who fail to respond to steroids. Idiopathic sclerosing inflammation, is a distinct form of orbital inflammation characterized by a slow and relentless fibrosing process, with progressive involvement of orbital structures. It may be part of a multisystem disease with progressive fibrosis at other sites. In both sclerosing and non-sclerosing inflammation, raised serum IgG4 and increased IgG4 positive plasma cells may be found in the tissues in some cases. This IgG4-related disease may be bilateral and affect other sites, but is usually responsive to steroids.

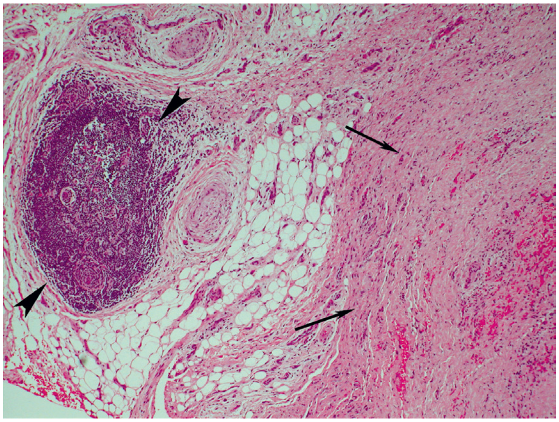

FIGURE 9-49 In idiopathic orbital inflammation the orbital fatty tissue is replaced by fibrous tissue (arrows). There is an accompanying inflammatory cell infiltrate that may include lymphoid follicles (arrowheads).

FURTHER READING

A full reading list is available online at https://expertconsult.inkling.com/.

FURTHER READING

Reid, R., Roberts, F., MacDuff, E., 2011. Pathology illustrated, Elsevier Churchill Livingstone, Edinburgh.
Roberts, F., Thum, C.K., 2014. Lee's Ophthalmic histopathology, Springer, London.
Sehu, K.W., Lee, W.R., 2005. Ophthalmic pathology: an illustrated guide for clinicians, Blackwell Publishing, Oxford.
Yanoff, M., Fine, B.S., 2008. Ocular pathology, fifth ed. Elsevier Mosby-Wolfe, London.

Index

Page numbers followed by ‘f’ indicate figures, ‘t’ indicate tables, and ‘b’ indicate boxes.

A

Abducent nerve, 74
  course of, 74, 76f
  lesions of, 76b
ABO blood groups, 141
Acanthamoeba, 462, 477–478, 498–499, 499f
  host response to, 478
  keratitis, 477–478, 477f
  as ‘Trojan horse’, 478
  virulence factors of, 478
Accommodation, 30
  ciliary muscles, 217–218, 312
  lens zonules, 35
Acetazolamide, 359.e2b
Acetyl coenzyme A, 193
Acetylcholine, 355
  amacrine cells and, 267
  cell depolarization and, 344–345
  depolarization, 263
  inhibition of, 356
  parasympathomimetics and, 355–356
  sympathetic system and, 357, 357f
  tear fluid release and, 202.e1f
Acetylcysteine, 355
Achromatopsia, 308
Acid burns, 487
α_1-acid glycoproteins, 341
Acinar cells, 200–201
Actin, 226
  dynamin interaction, 173, 173f
  stress fibre, 172, 189
Actin-binding proteins, 172–173
Actin-myosin motor, 388–389
Actinomycetes, 494–495
‘Action-perception cycle’, 308
‘Active pulley hypothesis’, 62–63
Acute-phase response, 394
Acute retinal necrosis, 496
ADAMTs (A disintegrin and metalloproteinase with thrombospondin motifs), 182
  host response to injury, 383
Adaptive optics, 280, 304, 305f
Addressins, 418
Adenoid cystic carcinoma, 537
Adenosine monophosphate-activated protein kinase (AMPK), 190
Adenyl cyclase, 345–346, 345f
Adenylate cyclase, 225
Adherens junctions, 189, 189t
Adhesion molecules, 448
Adnexa, 77–92
Adnexal glands, benign tumours of, 523
Adrenergic agonists, 225, 208.e1b, 359.e1
β_2-adrenergic agonists, 359.e1
Adrenergic antagonists, 225, 359b, 359.e1
Adrenergic neurotransmitters, 359f
Adrenergic receptors, 208, 358b, 359f, 208.e1f
  intraocular pressure control, 225, 358
  profile of, 360t
Advanced glycation end (AGE) products, 191, 192f
Affinity chromatography, 145
Ageing, 505, 506b
  cataract associated with, 511–512
α-agonists, 358
β-agonists, 358
Aicardi syndrome, 138
AIDS (acquired immunodeficiency syndrome)
  cytomegalovirus retinitis and, 497
  retinitis and, 403.e2
Albinism, 155
  enzyme defects in, 155f
Albumin, drug binding, 341
Aldose reductase pathway, 191, 191f, 235
  cataract, 238
Alkali burns, 487
Alleles, 130, 139
  linkage analysis of, 140
Allergic eye disease, 445–447
  treatment of, 361b
Allergy, 445–450
  disease production mechanism in, 448
Amacrine cells, 43f, 47f, 49, 266–267
  dopamine effect on, 267
  ON/OFF microcircuitry, 291
  retinal microcircuitry, 265–266, 292f, 295, 296f
Amadori products, 191, 192f
AMBEGUJAS phenomenon, 318–319, 320f
Amino acid(s)
  cataract, 511–512
  metabolism disorders, 504–505
Amino acid transporters, 243–244
Aminoacyl-tRNA synthetases, 133–134
γ-aminobutyric acid (GABA), 265
  amacrine cells and, 266–267
  horizontal cells and, 48–49
Amniotic cavity, 103.e2
Amyloid, 507–508
Anaphase, 131
Anatomical terms of reference, 1, 2f
Anergy, 442–443
Aneuploidy, meiotic, 136.e1
Angioid streaks, 506
Aniridia, 127b
Aniseikonia, 317
Ankylosing spondylitis, 139
Ankyrin-spectrin-actin, 172–173
Anophthalmia, 112b
Anterior chamber angle
  development of, 127, 128f
  fluid transport and, 215–216
  malformation of, 127b
Anterior chamber-associated immune deviation (ACAID), 457
Anterior segment
  antioxidation in, 220–221
  detoxification in, 220–221
    cytochrome P_{450} system in, 220
Antibodies, 396–399
  diversity, 441b
  engineering for therapeutic use, 400b, 400f
  fragments, 399
  generation during B-cell ontogeny, 439, 441f
  genetic control of production, 439
  H chains, 396–397
  L chains, 396–397
  production, 439
  structure, 396, 397f, 397.e1f
Anticholinergic drugs, 217–218
Anticholinesterases, 356b
Antigen(s)
  binding, 427–432, 429f
  endocytosis, 438
  epitopes, 397
  intracellular, 396
  MHC, 424–426, 432
  microbials, 439
  presentation, 432–437
Antigen presentation
  active immune response, 435
  B cells as, 438
  cross-presentation, 421
  HLA, 432
  lipid, 433f
  local, 396, 412
  sugars, 432

Antigen presenting cells (APCs), 395, 422f
  antigen recognizable, 421–424
  lymph nodes, 414–415
  processing, 422–423, 424f
    defects, 422
Antigen recognition, 395, 421–424
  by B cells, 437–438
  MHC molecule, 421
Antigen-antibody binding, 397–399
Antihistamines, 360–361
  actions of, 361b
Antimuscarinic agents, 357t
Antioxidant systems, 220–221
Antioxidants, 360b
'Anti-saccadic task' test, 333
Anti-sense oligonucleotides, 147
Anti-TNF agents, 365–366, 366f
Anti-VEGF therapy, 353f
Apolipoprotein D, 199–200
Apoptosis, 146–147, 489–490, 491f
  and death receptors, 167, 168f
  detection of, 147
  gene regulation of, 147b
  signaling pathways, 168b
Apraclonidine hydrochloride, 359.e1
Aptamers, 352–353
AQPO gene, 233
Aquaporin(s), 233
  expression, 233f
Aquaporin-0, 229, 233, 233f
  modification with age, 237
Aquaporin 1, aqueous humour transport, 207–208
Aquaporin water channels, 224
Aqueous deficiency, 354
Aqueous humour, 218
  blood-aqueous barrier, 218–220
  circadian regulation of, 225
  composition of, 222–223
  dynamics, 222–228
  episcleral circulation of, 227–228
  functions of, 36–37
  glucose, 235
  hydrogen peroxide, 221
  production of, 30–31, 31f
  secretion of, 222–223, 224b, 224f
    and adrenergic receptors, adenylate cyclase with, 225
    and M_3 muscarinic receptors, 223–225
    neural/autonomic control of, 223–225
    and nucleotide receptors, 225
    and nucleotides, 225
  uveoscleral drainage of, 227
Aqueous outflow pathways, 24–26, 225–228
  ciliary muscle action on, 30
  control of, at trabecular meshwork, 226
  'conventional', 24
  development of, 127
  and juxtacanalicular cribriform meshwork, 226
  'non-conventional', 24–26
Arachidonate, 197, 197b, 197f
Arachnoid granulations, 12
Arachnoid mater, 12, 9.e1f
Arachnoid villi, 12
Arcuate eminence, 10t
Arcus senilis, 506
ARP 2/3 complex, 172f, 173
Arrestin, 261, 264f, 430, 430f
Arteriovenous fistulas, in cavernous sinus, 13b
Ascaris spp., 499
Aspartate, and disc shedding, 247, 248f
Aspergillus spp., inflammatory disease, 497
Aspirin, 239–240, 362–363
Asthma, 445–447
Astigmatism, 15b, 15f, 209
Astrocytes, in retina, 49, 50f
Astrocytic hamartoma, 531, 532f
Atopy, 445–447
Atrial natriuretic peptide (ANP), 225
Atrophy, 490
Aubert's phenomenon, 311, 311b, 311f
Autoantigens, 444
Autoimmune disease, 394
  dysregulated, 488
  inflammatory, 500–504
Autoimmune regulator (AIRE) gene, 413–414
Autoimmunity, 439–445
Autonomic nervous system, ocular drugs and, 355–358
Autosomal dominant disorders, 137.e1, 137.e1f
  ocular disease, 152t
Autosomal recessive disorders, 137.e1, 137.e1f
  ocular disease, 152t
Autosomes
  dominant, 137
  recessive, 137, 137.e1, 137.e1f
Avellino dystrophy, 517
Axenfeld-Rieger's anomaly, 120b, 127b
Azathioprine, 347b, 364b

B

B cell(s), 380–381
  activation, 437–439
  affinity maturation, 439
  anergy, 443–444
  antibody production, 439
  differentiation, 438–439
  in effector response, 396
  immune response, 380–381
  isotype switching in, 448
  maturation, 438f
  memory, 420
  ontogeny, 439, 441f
  polyclonal response in, 439
  somatic mutations, 439
  tolerance, 443–444
B cell lymphoma, 535
B7:CD28 interaction, 435
Bacteria
  opsonization of, 492–493
  toxins, 487–488
Bacterial infection(s), 473–475
  chronic, 494–495
  endophthalmitis, 459, 474–475
  inflammatory disease, 494–495
  ocular surface in, 473–474
Bartonella spp., 494–495
Basal cell carcinoma, 523
  infiltrating/sclerosing subtype, 523, 524f
  micronodular subtype, 523
  nodular/solid, 523, 524f
  superficial, 523
Basal cells, proliferation of, 211
Basal lamina, 15–16
Basement membrane
  endothelial cell binding, 189, 189t
  epithelial cell binding, 189, 189t
  structure of, 183–184, 185f
Basic fibroblast growth factor (bFGF), 244–245
Basophilic cytoplasm, 89–91
Basophils, 376f, 378–379
BAX gene, 147b
Bcl-2 gene, 147b
Behçet's disease, 425, 500
Benign mixed tumour, 536–537, 536f
Benzalkonium chloride, 354
Bergmeister's papilla, 118–119
Best disease, 515, 516f
Bestrophins
  and aqueous humour, 228
  and ion transport, in RPE, 255
Betaxolol, 360b
Bezold-Brucke phenomenon, 272–273
Binocular single vision (BSV) field, 314, 315b, 315f–316f
  driving licence requirement, 316f
Binocular viewing, perception, 287
Binocular vision, 309–313
Biochemical pathways, ocular function effects, 189–198
Biologics, 365–366
Bipolar cells, 43f, 45–46, 47f, 263–265, 266f
  blue cone-specific, 46
  colour opponency, 306
  colour perception, 306
  midget, 46, 280–282
  retinal light stimulus detection, 291
  retinal microcircuitry, 265–266, 291, 293f, 295f
Blastocyst, 103.e2
'Blind spot' phenomenon, 59–61

- Blindness
  - chlamydial infection, 497
  - from optic nerve lesion, 99f
- ‘Blindsight’, 332
- Blink rate, 198
- Blink reflex, 80, 82t, 198, 202
- Blinking, 463
- Bloch’s law, 278–279
- β-blockers, 358
  - actions on IOP, 360b
  - topical, 360b
- Blood, inflammatory reactions of, 500
- Blood-aqueous barrier, 218–220
  - drug distribution limitation by, 342, 352
  - iris blood vessels, 216–217
- Blood-ocular barrier, 28
- Blood-retinal barrier, 52, 245
  - drug distribution limitation by, 342, 352
- Blue cone monochromacy, 150–151
- ‘Blue’ sclera, 22b
- Bone marrow
  - lymphomyeloid cell maturation, 402
  - stem cells, 413
- Bones, membranous, 129.e1
- *Borrelia burgdorferi*, 494–495
- Bowman’s layer, 16, 16f, 203f
  - adhesion to basement membrane, 211
  - collagen, 204
  - corneal dystrophies of, 516
  - corneal epithelium integrity, 203
- Bowman’s membrane, 464
  - development of, 119–120
  - drug administration in, 348
- Bradykinin system, 492–493
- Brain, processes involved, 270.e1f
- Brain natriuretic peptide (BNP), 225
- Brainstem, 329, 334–335
  - lesions, 331, 333b
- Brevican, 215
- Brimonidine, 359.e1
- Bronchus-associated lymphoid tissue (BALT), 416
- Bruch’s membrane, 41, 42f, 55, 56f–57f
  - composition, 255–256
  - elastic fibre degeneration, 506
  - morphogenesis, 114
- Bulbar fascia, 62
- Bullous pemphigoid antigen 180, 182
- Buphthalmos, 21b
- Bystander activation, 444

C

- C-type lectins, 467, 469f
- Ca^{2+}/Mg^{2+}ATPase, 208–209, 235
- Cadherins, 189
- Caffeine, 239
- Calcium
  - in cellular activity mediation, 345–346
  - intracellular concentration of, 345–346
- Calcium-calmodulin complex, 345–346, 346f
- Calcium ions
  - intracellular concentration, 259–261
  - mobilization, 261
- Calgranulin, 455–456
- Calpactins, 236
- Calpains, 227, 236
- Canal of Schlemm, 24, 25f
- *Candida*, 459
  - inflammatory disease, 497
- *Candida albicans*, 475
- Candidate gene analysis, 150
- Canthus, 82
- Capsulopupillary vessels, 121–123, 123f
- Carbohydrate metabolism in lens, 235
- Carbonic anhydrase, 359.e2b
  - type II, 224
- Carbonic anhydrase inhibitors, 353–354
- Carcinogenesis, 519–521
- Carotid artery, internal, aneurysm in, 99f
- Carteolol, 360b
- Caruncula lacrimalis, 85–86
- Caspase enzyme activation, 146–147
- Cataract(s), 35b, 511–512
  - ageing associated, 511–512
  - formation of
    - age-related, 238–240, 239f
    - from chaperone function, failure of, 240
    - and insult to lens, 238
  - myotonic dystrophy and, 155–156
  - reduced vision in, 237–238
  - secondary, 512
  - from steroid oculotoxicity, 369
  - types, 238
- Cavernous sinuses, 10, 11f–12f
  - arteriovenous fistulas in, 13b
  - cranial nerve lesions in, 71b, 76b
  - thrombosis in, 13b
- Cavum trigeminale, 12, 12.e1
- CCN (CYR61/CTGF/NOV) proteins, 184f, 185
- CD1, 432
- $CD4^+$ $CD25^+$ T regulatory cells, 421
- $CD4^+$ T cells, 380
  - activation, 435
  - scleritis, 456
- $CD8^+$ T cells, 412
  - scleritis, 456
- CD31, 387
- CD36, 214–215, 246, 255, 261–262
- CD40-CD40L interactions, 435
- CD40:CD40 ligand, 435
- CD44, 435
- Cell(s)
  - endoplasmic reticulum and Golgi apparatus, 164
  - intermediate filaments, 173–174, 175f–176f
  - intracellular matrix, 168b, 169–177
  - microfilaments, 165b, 169–173, 173t
  - mitochondria, 164–169, 167f
  - nucleolus, 169
  - nucleus, 169
  - plasma membrane, 161–164, 162f
  - structure of, 160–161, 161f
- Cell adhesion, 186–188
  - and transmembrane signaling, 188
- Cell adhesion molecules (CAMs), 386
  - integrin, 386
- Cell adhesion proteins, 189
- Cell biology, 146–150
- Cell cycle, 131f
  - control of, 132b, 132f
- Cell damage, agents and conditions causing, 486–489
- Cell death, mechanism of, 489–490
- Cell destiny, control of, 146–150
- Cell division, 130–132, 135–136
- Cell-mediated immunity, 411
- Cell membrane, proteins, glycation of, 191–193
- Cell signalling, immunological synapse, 423–424
- Cellcept, 364b
- Cellular retinal binding protein (CRALBP), 258–259
- Cellular retinol-binding protein (CRBP), 257
- Central retinal artery occlusion, 510–511
- Central retinal vein occlusion, 511, 512b
- Central vitreous, 38
- Centriole, 177
- Cerebellum, 334–336
- Ceroid/lipofuscin, 237
- Ceruloplasmin, 253
- Cervico-ocular reflex, 329
- Chalazion, 83b, 500, 500f
- Chamber angle, malformation of, 127b
- Chandler syndrome, 518
- Chaperone proteins, 159.e2, 159.e2f
  - failure of, and cataracts, 240
- CHARGE syndrome, 108b
- *CHD7* gene, 108b
- Chediak-Higashi syndrome, 155
- Chemical agents, ocular tissue damage, 487
- Chemoattractants, 387–388
- Chemokine receptors, 404, 406f
- Chemokines, 404, 406f
  - leucocyte activation by, 386
  - nomenclature for, 405t
- *Chlamydia*, 372, 454, 455f, 497
- *Chlamydia trachomatis*, 462–463
  - blindness and, 497
- Chlorbutol, 354
- Chloride pump, corneal, 208
- Chlorine channel, Ca^{2+} dependent, 255
- Chloroquine, retinopathy from, 368
- Cholesterol, 194–195, 197f
  - lens fiber cell content, 233, 237f
  - retinal pigment epithelium, 255

Cholinergic receptors, in corneal epithelium, 208.e1, 208.e1f
Chondroitin sulphate, 226
  corneal, 187b, 205
  vitreous, 240–241
Chondroitin sulphate proteoglycans, 21
Choriocapillaris, 55–56, 56f–57f, 221
  development of, 124
Chorioretinopathy, birdshot, 139
Choristomas, 522
Choroid, 55–59, 56f–57f
  ageing changes in, 58b
  angioma, 528f
  biochemical process, 221–222
  blood flow in, 221
  dendritic cells in, 59f
  development of, 124
  embryology, 108–110
  functions of, 221–222
  infarctions in, 59b
  lymphoid function, 222
  macrophages in, 59f
  mast cells in, 58f
  melanoma, 526
  neovascularization, 514b
  nerve supply of, 58–59
  non-vascular, 222
  as paracrine tissue, 215
  pigmentation of, 56
  P_{O2} and P_{CO2} and, 222
  stroma of, 56
  trauma response, 492
  vascular, 56, 56f–57f, 221–222
  venous blood in, 219f, 222
Choroidal fissure, 106–108
  closure, 111
Choroideraemia, 153–154
Chromatic aberrations, 282
Chromaticity, 272–273, 272f
Chromatids, 131
Chromatin, 157–158, 169
Chromophore, 259
  cone, regeneration of, 257, 257f
  and opsin, 258–259, 260f–261f
Chromosomal abnormalities, 488
Chromosomal anomalies, 129.e4
Chromosomes, 130
  crossing over of, 140
  cytogenetic analysis of, 130
  defects and gene mutations in, 135–137
  deletion of, 136, 137b
  division of, 135–136
  insertions in, 137b
  inversion in, 136
  meiosis, 131–132
  mitochondrial, 138, 167–169
  mitosis, 130–131, 131b
  numeric abnormalities in, 136.e1
  structural abnormalities in, 136–137
  translocation, 136
Chronic progressive external ophthalmoplegia (CPEO), 490b
Cicatricial conjunctival disease, 354
Cicatrization, conjunctival, 454–455
Ciclosporin, 364b, 437
Cidofovir, 352.e1b
Ciliary arteries, posterior, 124
Ciliary body, 30–32, 31f
  accommodation in, 30
  antioxidant systems, 220–221
  biochemical processes, 217–221
  blood-aqueous barrier, 219f
    breakdown of, 218–220
  blood flow in, 217, 217f
  blood supply of, 32
  coloboma, 108b
  cytochrome P_{450} system in, 220
  development of, 124–126, 125f
  eicosanoids, 220
  epithelial cells, non-pigmented, 31, 216
  epithelium, 124–125
    aqueous humour secretion, 222–223
  functions of, 217
  hydrogen peroxide, 221
  melanoma, 526
  nerve supply of, 32
  parasympathetic innervation of, 32
  pars plana in, 30
  pars plicata in, 30
  sensory innervation of, 32
  sympathetic innervation of, 32
Ciliary folds, 117, 124–125
Ciliary muscle, 30
  accommodation, 217–218, 218f, 312
  on aqueous outflow, 30
  arrangement of, 32f
  development of, 126
  parasympathetic antagonists, 356–357
Ciliary nerves, 58–59
Ciliary processes, 30, 31f
Ciliary zone, 26
Ciliopathies, 154b
Cilioretinal artery, 52–53
Cimetidine, 361
Circadian rhythms
  entrainment, 296
  melatonin, 247, 276–277
11-*cis* retinal, 275–276, 275f
Citric acid cycle, 209
  in lens, 235
Clathrin-coated vesicles, 165, 165f
Clinoid processes, posterior, 9
Clonal deletion, 442
Cloning, 142, 142f
Cloquet's canal, 123–124
Clotting cascade, 492–493
'Clump cells,' in iris stroma, 27f, 28
Cluster of differentiation (CD), 381b
Clusterin, 201–202
Coat's disease, 532f
Cocaine, 359.e1
Codon deletions and insertion, 137b
Cofilin, 189
Colchicine, 174
Collagen, 180–182, 180f
  in cornea, 16–18
  corneal, 182
    transparency, 204–205
  Desçemet's membrane, 182
  fibrils, arrangement of, 204, 212–213, 214f
  glycation, 191–193, 192f
  iris, 182
  lens, 182
  retinal pigment epithelium, 117–118
  scleral, 21, 22f, 205, 205f
  triple helix of, 178–180, 179b, 179f, 181f
  types of, 180t
  vitreous, 182
Collagen shields, 350–351
Collagenous lamellae, 16–18
Collarette, 26
Collectins, 405
  activation of complement, 408
  innate immunity, 407b, 407f
'Collector units', 319
Colliculus, 333
Coloboma, 108b, 108f, 119b
  axial, 119b
  enucleation for malignant melanoma, 526
Colorimetry, 302
Colour
  constancy, 306–307, 318–319, 322
  discrimination, 302
  hues, 272
    discrimination, 303
  luminosity, 302–303
  perception, 324–325
  primary, 272
  processing, 318–319
  sensing, 271–273
Colour blindness, 307–308
  red-green, 150–151
Colour detection, 292, 294f
  photoreceptor, 303–305
Colour matching tests, 306.e1, 306.e1f
Colour-opponent mechanism, 306
Colour vision, 271, 302–308
  trichromatic theory, 303
'Colouredness', 272–273
Common tendinous ring, 5–6
Complement, 405–410, 492–493
  activated, 387–388
  activation, 405
    regulation, 408–409, 409b, 409t
  cascade, 406–407, 408f
  effects, 405–406
  innate immunity, 407b, 407f
  ocular microbial barrier, 463
  receptors for, 409t
  tears, 463

Complement pathway, 405–406
  alternative, 408
Complementarity-determining region (CDR), 396–397, 429, 429f
  antigen binding sites, 439
  somatic mutations occur, 439
Complementary DNA (cDNA), 146
Cone(s), 44f, 45
  bipolar cells, 46, 262–263, 292, 293f
  colour discrimination, 293
  colour sensing, 271
  discriminatory capacity, 280
  foveal, 280–282, 285
    blue-sensitive, 305
  horizontal cell feedback, 292f, 294–295
  light sensing, 307f
  microcircuits, 291–292, 293f
  monochromatism, 307
  opsins, 304–305
  pedicles, 44f, 45
  photopigments, 305
  psychophysical evidence for, 303–304
  regeneration of, 257, 257f
  resolving power of eye, 282
  retinal light stimulus detection, 291
  thresholds, 303
  types, 303
Congenital hereditary endothelial dystrophy, 518
Congenital malformations, 129.e4
  anterior chamber angle, 127b
  corneal, 120b, 127b
  glaucoma, 127b
  hyaloid system, 124b
  lens, 121b
  nasolacrimal duct, 129.e6b
  neural tube, 112b
  optic nerve head, 119b
  optic vesicle, 112b
  periods of vulnerability, 129.e4
  vitreous system, 124b
Conjunctiva, 83–89, 84f–87f
  allergic responses in, 85b
  biochemical processes, 201–202
  bulbar, 83f, 85–86, 86f–87f
  cicatricial disease of, 354
  cicatrizing disease of, 454–455
  cysts, 537
  dendritic cells, 86f–87f, 88, 453
  drug administration in, 347–348
  epithelium, 201–202
    mucin production, 200–201
  forniceal, 84f–85f, 85
  glands in, 88–89
  goblet cells, 86f–87f, 88, 200–201
  keratin, 202
  lymphocytes in, 88
  lymphoid layer, 202
  malignant melanoma, 525
  mast cells, 359–361
  melanocytes in, 86f–87f, 88
  palpebral, 84f–85f, 85, 88
  papillomas, 519–520
  stroma, 202
  structure of, 86–88
  systemic disease in, 86b
Conjunctiva-associated lymphoid tissue (CALT), 88, 202, 416
Conjunctival sac, 83–85
  drug administration in, 347–348
  drug residence time in, 349–350
Conjunctivitis, allergic, 445–446, 454f
Connective tissue-activating peptides, 394
Connexin, 229
Consanguinity, 137.e1
Contact lenses, 469–471
  associated to fungal keratitis, 475
  related corneal inflammation, 474
  soft, 351
  types, 210–211
  wearing, 210–211, 210f
Contraction furrows, of iris, 26
Contrast sensitivity, 283–285, 284f
Convergence, 305–307, 312, 327
Cornea, 14–20, 198
  ageing changes to, 20b
  Aspergillus and *Fusarium* in, 462
  biochemical processes of, 203–215
  calcification of, 507
  cell turnover and wound healing in, 211–215, 211.e1f
    limbal corneal stem cells in, 211–212, 212f
  chloride pump, 208
  collagen, 204–205
    anchoring fibrils, 205, 205f
    fibres in, 16–18
    turnover, 207
    type V, 204–205, 204f
  combined granular-lattice dystrophy, 517
  components of, 203
  contact lens wearing, 210–211, 210f
  crystallins in, 207
  curvature, 209
  dendritic cells, 207, 453, 464, 465f
  development of, 119–120, 120f
  drug administration in, 348
  endothelial cell pump in, 207, 208b, 208f
  endothelium, 19–20, 20f
    development of, 119–120
    transparency, 207–208
    wound healing, 213
  epithelium, 14–16, 16f, 203f
    adhesion to basement membrane, 211
    cell migration, 211
    transparency, 203–204
    wound healing, 211, 211.e1
  fluid transport, 186f, 203–204
  glucose utilization, 209
  glycosaminoglycans, 16–18, 186, 187b, 205–206
  granular dystrophy, 517, 517f
  hysteresis of, 207
  image formation, 209
  iontophoresis of drugs in, 352
  keratan sulphate, 186
  keratocytes, 207
  lattice dystrophy, 516
  leucocytes, 207
  leucoma, 120b
  macrophages, 19f, 464, 465f
  malformation of, 120b
  and mucin, 200–201
  mucous layer, 198f, 200–201
  nerve supply of, 20, 21b
  neurotransmitters in, 210
  oedema of, 517
  opacification of, 517
  oxidative metabolism of, 209
  oxygen handling by, 209–210
  proteoglycans, 186
    binding maps of, 206, 206f
    and collagen fibrils, 205–206
    linkage sites of, 205
    organization of, 206, 206f
  pyogenic infection of, 494
  refractive surgery, 213
  shape of, 14
  stroma of, 16–18, 18f
    wound healing, 212–213, 214f
  structure of, 14–20, 16f
  swelling pressure, 206–207
  transparency of, 14, 203–209
    surgical wounds to, 213
  vascularization, 214–215
  vitamin A, 215
  in vivo confocal microscopy of, 15–16, 17f
Corneal dystrophies, 515–518
Corneal grafting, 457–459, 458f, 451.e1
Corneal opacity, 491–492
Corneal transplantation, 451
Corneal ulcers
  melting, 455–456
  spontaneous central, 501
Corneoscleral envelope, 14f
Corneoscleral meshwork, 24, 25f
Corrugator supercilii, 80–81, 81f
Cortical activity, neuronal ‘noise’, 322
Cortical centres, complex eye movement regulation, 333–334
‘Cortical magnification factor’, 285, 285f
Cortical processing defect, 307–308
Cortical vitreous, 37f, 38
Corticosteroids, 364b, 477
Cranial cavity, 1
  osteology of, 7–13, 9f, 7.e2f
Cranial fossa
  anterior, 7–8
  middle, 9, 9f
    features on floor of, 10t
  posterior, 7.e2

Cranial nerves, 68–77
  III, 69–71, 73f
  IV, 71–74, 75f
  V, 74–77, 78f
  VI, 74, 76f
  VII, 77
  components of, functional analysis of, 68.e1t
  origin of, 72f
  sensory neurones, 129.e1
Craniofacial abnormalities, 129.e5b
Cranium, 1
C-reactive protein, 395
Cribriform meshwork, 24, 25f
  development of, 127, 128f
Critical flicker fusion, 271
Critical fusion frequency, 301
Crystallins, 121, 123f, 230–232, 230.e1b
  cataract formation, 237
  enzyme activities of, 235–236
  glycation of, 237
  heat-shock protein, model for, 230–231, 231f
  homology of, with other enzymes, 230–231
  micelle, 230–231
  phosphorylation of, 231
  structure of, 232f, 230.e1f
Cuneus gyrus, 95
Curuncula lacrimalis, 82
CXC chemokines, 468b
Cyanate, and cataract, 239–240
Cyclic adenosine monophosphate (cAMP), 177
  in drug-receptor interactions, 345
  in intraocular pressure control, 225, 358
  retinal neurotransmission, 258
Cyclic guanosine monophosphate (cGMP)
  neurotransmitter, 265–266, 267f
  phototransduction, 259
Cyclin-dependent kinases (CDK), 132b
Cyclo-oxygenase 2 (COX-2), specificity of, 363
'Cyclopean' eye, 312–313, 313f, 317
Cyclophosphamide, 364
Cyclopia, 112b
Cycloplegic refraction, antimuscarinic agents, 356–357
CYP1B1 gene, 227, 347
CYP2D6 enzyme, 220
Cyst, 537–538
  dermoid, 500, 537
Cystic fibrosis, 147b, 137.e1
Cystic fibrosis transmembrane conductance regulator (CFTR) gene, 147b, 207
Cystinosis, 505
Cytidine triphosphate, 247
Cytochrome C, 167–168
Cytochrome P_{450} system, 220, 342
  and drug detoxification, 220, 221b, 221f
Cytokine(s), 380
  acute phase response, 394
  adaptive immune responses, 399–404
  inflammation, 403
  innate immune responses, 399–404
  Jak-STAT pathway, 404f
  lymphomyeloid cell maturation, 402
  networks, 403f
  properties, 399–400, 401f
  in specific immune reactions, 401–402
Cytokine receptor(s), 402–403, 403f
  antagonists, 402–403
Cytolysis, 408
Cytomegalovirus (CMV) retinitis, 472–473, 497, 497f
  treatment of, 352.e1b
Cytoplasm, 169
Cytoskeletal fibres, group 3, 174–177
Cytotoxic agents, 364
Cytotoxic T cells, 396, 412
  activation of, 435

D

Dalen-Fuchs nodules, 502
Damage-associated molecular patterns (DAMPs), 371
Dark adaptation, 257, 273–279
  curve, 274–277, 274b, 274f
Dark currents, 262–263, 288, 288b, 288f
  ATPase dependent, 243
Death receptors, 167, 168f
Decorin, 205
β-defensins, ocular infection barrier, 463
Degeneration, 505–514
  composition/type of, 505–507
  elastic fibre, 506
  fatty, 506, 506f
Degenerative vascular disease, 508–511
Delayed-type hypersensitivity, 411
Deletion, 136
  failure, 444
Dendritic cell (DC), 207, 371, 375f, 416–418
  choroidal, 59f
  conjunctiva, 86f–87f, 88, 453, 453f
  conjunctival epithelial, 202
  cornea, 453, 464, 465f
  immature, 421f
  from myeloid and lymphoid precursors, 378b, 378f
  subsets, 378.e1b
  surveillance function, 378.e1b
Depolarization, 263, 265, 287, 344–345
Depressor supercilii, 77–80
Depth perception, 273, 312–317, 314f
  motion detection/texture analysis, 319–320, 320f
Dermatan sulphate, 21, 207, 226
Dermoid, 522
  cysts, 500, 537
Desçemet's membrane, 16f, 18–19, 20f
  collagens, 182, 204
  congenital hereditary endothelial dystrophy, 518
  development of, 119–120
Desmin, 174
Desmocollin, 203–204
Desmoglein, 203–204
Desmosomes, 14–15, 174, 189, 189t, 172.e1, 172.e1f
  spot, 203–204
Diabetes, 504
Diabetes mellitus, 191
  cataract, 238
  retinal blood flow, 246
Diacylglycerol (DAG), 261
Diakinesis, 131
Diaphragma sella, 9, 11f–12f, 12, 12.e1
Dichlorphenamide, 353–354
Dichromatism, 308
Diclofenac, 362–363
Diencephalon, 103
Diffraction, 282
Dilator pupillae muscles, 26, 28, 126
Diplopia
  divergent squint, 313
  heteronymous, 312–313
Diplotene, 131
Direct reverse signalling, 189
Discoidin domain receptors (DDRs), 182
Discrimination, contrast sensitivity, 284–285
Discriminatory capacity, 280
Disease associations, 139
  HLA-linked, 139t
DNA, 132–135
  fragment, 142
  ordering of, 135
DNA laddering, 147
DNA ligase, 142
DNA polymorphisms, 143
DNA probe, 142
Docosahexanate phospholipids, 247
Docosahexanoic acid, 257
Doll's head movements, 330
Donder's law, 335
Dopamine, 265, 267–268
  receptors of, 208
  and regulation of shedding, 247, 248f
Dorsum sellae, 9
Dorzolamide, 359.e2b
Dose-response curve, 344.e1b
Dot-blot hybridization, 146
Double-stranded DNA (dsDNA), 132–133
Dowling-Rushton equation, 275
Down syndrome (trisomy 21), 136.e1, 129.e6b, 136.e1f
DR4 associations, 139

- Drug(s)
  - absorption of, 340, 347
    - ocular, 347
    - topical, 348, 348b
  - acetylation of, 346–347
  - active transport of, 340
  - administration route of, 341, 341b, 347–348
  - agonist, 344b, 344.e1b, 344.e1f
  - antagonist, 344b, 344.e1b, 344.e1f
  - autonomic nervous system and, 355–358
  - biliary excretion of, 343–344
  - bioavailability of, 339
  - on cell activation, 437
  - clearance of, 342–343
  - conjugation, 342
  - contralateral eye effect, 354
  - delivery to eye, 347–348
    - advanced systems, 351–352
    - vehicles, 350b
  - diffusion of, 340
  - distribution of, 341–342, 341b
  - dose-response curve of, 344.e1b
  - efficacy of, 344.e1b
  - excretion of, 343–344
  - first-order kinetics of, 339, 340f
  - glomerular filtration of, 343
  - half-life of, 338–339
  - immunosuppressant, 364–366
  - immunosuppressive, 437
  - inactivation of, 349–350
  - interactions, 341
  - intracameral administration of, 352–353
  - intravitreal administration of, 352–353
  - ligand, 344b
  - lipid solubility of, 340–341
  - liposomes, 351–352
  - metabolism of, 342–343, 342b, 344f
    - in eye, enzymes involved in, 349b
  - as microsomal enzyme manipulators, 343b
  - ocular absorption of, 347
  - ocular iontophoresis, 352
  - oculotoxic, 369
  - oxidation of, 342
  - partial agonist, 344b
  - particulates, 351
  - pH of, 348–349
  - plasma concentration-time curve of, 338–339
  - potency of, 344.e1b
  - preservatives, 354
  - properties of, 340–341
  - protein-bound, 341
  - reabsorption of, 343
  - receptors, 344b
    - interactions, 344–347
  - renal excretion of, 343
  - secretion of, 343
  - semisolids (ointments), 350
  - slow-release preparations of, 350–351
  - solutions, 350
  - steady-state plasma concentration of, 339
  - systemic administration of, 353–354
    - ocular toxicity from, 368–369
  - tolerance, 344.e1b
  - topical medications, 354
  - vehicles for, 350–351, 350b
  - volume of distribution of, 339, 339f
  - zero-order kinetics of, 339, 340f
- Drug distribution limitation, 342
- Dura mater, 9–12, 11f–12f
- Dural folds, 9–10, 11f–12f
- Dural venous sinuses, 10, 11f–12f, 9.e1f
- Dysthyroid eye disease, 456
- Dystrophies, 514–518
- Dystrophin, 189, 264–265, 266f

E

- E selectin, 448
- Ectopis lentis, in Marfan syndrome, 35b
- Ectropion, 83b
- Edrophonium, 356b
- EDTA, 354
- Edwards syndrome (trisomy 18), 136.e1
- Eicosanoid metabolite 12(R) HETE, 208–209
- Eicosanoids
  - biological actions of, 362b
  - ciliary body, 220
  - iris, 220
  - ocular function effects of, 361–363
  - and retinal blood flow, 246
  - synthesis of, 362f
- Elastic fibre degeneration, 506
- Elastin, 182–183, 183f
  - degradation of, 183
- Electro-oculogram, 298, 298b, 298f
- Electroretinogram (ERG), 298–299, 299f
  - flicker, 302
- Ellipsoid, 41–45
- Elongation of very long fatty acids-like 1 (ELOVL1), 252
- Embolism, 511b
- Embryoblast, 103.e2
- Embryology, 103, 104f–105f, 113f, 103.e1f
- Embryonic stem cells, 148
- Embryotoxon, 120b
  - posterior, 127b
- Encephalo-trigeminal angiomatosis, 521
- Endocrine exophthalmos, 500, 504, 505f
- Endocytosis, 165b, 165f
  - in lens, 234
- Endogenous materials, reactions to, 500, 500f
- Endophthalmitis
  - bacterial, 459, 474–475
  - metastatic, 494, 532f
  - risk for, 495b
- Endoplasmic reticulum, 164
  - and exocytosis, 165b
  - signaling, 164
- Endostatin, 214–215
- Endosymbiotic *Wolbachia* bacteria, 481, 482f
- Endothelial dystrophies, 517–518
- Endothelin, 386
- Enterohepatic circulation, 343–344
- Environmental causes, of congenital malformations, 129.e4
- Enzymes, conjugating, 349b
- Eosinophil(s), 376f, 379
- Eosinophil cationic protein (ECP), 448
- Eosinophil major basic protein, 448
- Epidermoid cysts, 537
- Epigenetic development, 110
- Epigenetic inheritance, 158.e1, 158.e1f
- Epigenetics, 135, 141, 141b
- Epigenome, 159
- Epinephrine, 225, 359.e1
  - intraocular pressure, 358
  - sympathetic system, 357, 357f
  - synthesis of, 358f
- Episclera, 21, 22f
- Episcleritis, 456
- Epithelioid haemangioma, 534
- Epithelium
  - corneal
    - drug administration in, 349–350
    - permeability of, 349–350
  - dystrophies of, 516
  - embryology, 108–110
- Epitopes, 373, 397
  - mapping, 430
- Epstein-Barr virus, 519–520
- ER-mitochondria organizing network (ERMIONE), 164–166
- ER stress, 164.e3, 164.e3f
- Erythrocyte, 376f
- Erythropoietin, 448
- Esterases, 349b
- Etanercept, 365–366
- ETDRS chart, 280–282
- Ethambutol, as oculotoxic drug, 369
- Ethmoidal canals, 2
- Ethmoidal sinuses, 4f, 8f, 7.e1
- Exfoliation syndrome, 508
- Exocytosis, 165b, 165f
- Exome sequencing, 146
- Exons, 132–133
- Exosomes, 165b, 165f
- Experimental autoimmune uveoretinitis, 459
- Extracellular matrix (ECM), 178–189
  - glycosaminoglycans, 186, 187b
  - mesenchymal cells binding, 189, 188.e1, 188.e2f
  - molecules, cell adhesion, 186–188
  - non-structural proteins, 184–186
  - proteins, 178–184, 179b, 179f
  - remodelling, 392

Extradural haematomas, 13b
Extrageniculostriate pathway, 331–332, 333f
Extranodal marginal zone lymphoma (EMZL), 535
Extraocular muscles, 64, 65f–66f
  anatomical features of, 67t, 70f
  development of, 129
  fibre classification of, 70t
  insertions of, 69f
  origins of, 68f
Eye
  anatomy of, 1–102
  fibrous coat of, development of, 119–120
  microbial infections of, 462–485
  shape, size and position of, 13–14
  structure of, 13–61, 14f
Eye field, frontal. *see* Frontal eye field
Eye field transcription factors, 110–111, 111f
Eye movements, 286–287
  binocular, 327–329
  compensatory, 335
  conjugate, 327
  control, 329–334
    superior colliculus, 331–333
  convergence, 327
  horizontal, 333–334
  mechanical control, 335–336
  natural activity, 335, 336f
  neural control, 335–336
  physiology, 326–336
  rapid 'reflex' locking-on, 332
  reflex, 330–331, 332f
  regulation of complex, 333–334
  rolling, 330
  rotational, 326
  torsional, 326, 329
  tracking, 310
  types, 326–329
  uniocular, 326
    see also Saccades
  voluntary (VOL) and involuntary (INVOL), 310f
'Eye shine', 56b, 56f
Eye worm, 482, 483f
Eyelashes, 82, 202, 129.e4
  hair follicles, 202
Eyelids, 82–83, 83f
  ageing changes in, 83b
  autoimmune disease of, 501
  blood and nerve supply of, 83, 86f
  cysts, 537
  development of, 129.e4
  granulomatous, 500
  margins of, 82, 83f–85f
  movements of, 83
  muscles of, 77–82
  phakomatous choristoma, 522
Ezrin, 255

F

Face
  development of, 129.e4, 129.e5f
  muscles of, 77–82
Facial defects, congenital, 129.e6b
Facial nerve, 77, 81f
  lesions of, 80b
  reflexes involving, 82t
Facial skeleton, 1
FACIT (fibril-associated collagens with interrupted triple helices) collagens, 180–182, 180t
False projection, 313
Falx cerebri, 12, 12.e1
Family pedigree, 137–138, 137.e1f
FAS ligand, 147b
Fasciculus, medial longitudinal, 69
Fatty acids, polyunsaturated, 195–197
Fatty degeneration, 506
FCO1 gene, 238
Fibrillin, 183, 183f, 185–186
Fibrin, 387–388
Fibrinolysis, 492–493, 500
Fibroblast growth factor (FGF), 120–121, 250
Fibronectin, 179, 179f, 184–185, 188.e1, 184.e1, 188.e2f, 184.e1f
  and fibrin, 184–185
  lens capsule, 234
  retinal, 253–254
  in reverse integrin-mediated cell signaling, 184–185
Fibrosis, 394
Fibrous coat, 14
  development of, 119–120
Fick's law, 340
Filarial nematodes, 463t
First-pass metabolism, 339
Fixation reflex, 329
Flat midget bipolar cells, 46
Fleurettes, 531
Flexner-Wintersteiner rosettes, 531, 531f
Flicker, 301–302
  limits of vision, 271
Flicker photometer, 302
Flow cytometry, 147
Fluorescent resonance energy transfer (FRET), 145
Flurbiprofen, 362–363
Focal adhesion proteins, 189
Follicular lymphoma, 535
Foramen lacerum, 10t
Foramen ovale, 10t
Foramen rotundum, 10t
Foramen spinosum, 10t
Forebrain, 103
Form
  discrimination, 273
  perception, 324–325
Foscarnet, 352.e1b
Fovea
  depression, 115–117
  development of, 115–117
Fovea centralis, of retina, 39
Fragile sites, 130
Fragile X syndrome, 137b
Frameshifts, 137b
Free radical scavenger, vitamin E, 198
Free radicals, 193–194, 193f, 200.e1
  cataract formation, 238–239
  damage, and lens, 239b, 239f
  detoxification of, 193t
  generation of, 164.e3, 164.e3f
  host response to injury, 383
  from hydrogen peroxide, 194b, 194f
Frequency of seeing, 274
Frontal eye field, 96, 97f
Frontal sinuses, 5f, 8f, 7.e1
  development of, 129.e4
Fronto-nasal process, 129.e4
Fuchs' endothelial dystrophy, 518, 519f
Fuch's heterochromic cyclitis, 459
Fundus flavimaculatus, 515, 516f
Fungal infections, 475–477
Fungal keratitis
  contact lens-associated, 475
  pathogenesis of, 475–477, 476f
  treatment of, 475
Fungi, inflammatory disease and, 497–498
Fusion supplement, 327

G

G-banding, 130
G-proteins, in second messenger activity regulation, 345b
Galea aponeurotica, 81
Gametes, 103.e2
  formation of, 131–132
Ganciclovir, 352.e1b
Ganglion cells, in retina, 43f, 46–48, 47f
  diffuse, 48
  midget, 47f, 48
Gap junctions, 189, 172.e1, 172.e1f
  lens semipermeable membranes, 229, 230f, 234
Gap phases, 130–131
Gene(s), 130
  candidate identification, 150
  cloning, 142
  components of, 133f
  decoding DNA, 133–135
  downstream, 110
  finding and tracking, 146
  frequency of, 140
  in ocular development, 109t
  ocular disease-causing, 152t
  regulation of expression, 135
Gene arrays, 143–145, 144f, 145b
Gene augmentation therapy, 147
Gene expression, cascade of, 110
Gene gun, 147

Gene loci, 130
  linkage of, 140–141
  retinal disorders, 153f
Gene mutations, 135–137
  screening for, 150
Gene probes, 146
Gene therapy, 147–148
  vectors in, 148b
Generator potentials, 287–288, 289b, 289f
Genetic abnormalities, 488–489
Genetic linkage, 140–141
Genetic makeup, 132
Genetic mutations, somatic, 439
Genetic polymorphisms, 139
Genetics, 130–156
  clinical, 137–139
  molecular, 132–135, 150–156, 152t
  population, 140–141
Geniculate bodies, lateral. *see* Lateral geniculate bodies
Geniculocalcarine tracts, 93f, 95, 96f
Genome, stability of, 169
Genome-wide association studies (GWAS), 143–145, 145b
Genomics, 143–145, 144f, 159
Giant cell angiofibroma, 534
Giant cell arteritis, 503, 503f
Glaucoma, 26b, 61b, 512–513
  carbonic anhydrase inhibitors for, 359
  closed-angle
    primary, 512–513
    secondary, 513
  congenital, 127b, 513–514
  exfoliation syndrome and, 508
  infantile, 127b
  intraocular pressure control in, 359
  iridocorneal endothelial syndrome, 518
  lens particle, 502
  narrow-angle, 356–357
  neurovascular (rubeotic), 513
  open-angle
    primary, 512, 513f
    secondary, 513
  phakolytic, 502
  rubeotic, 513
  secondary, 513, 513f
  from steroid oculotoxicity, 369
  topically applied drugs for, 360b
Glioma, 531–533, 533f
Glomerular filtration, 343
Glucocorticoids, 363–364
Gluconeogenesis, 190
Glucose
  facilitated diffusion of, 190
  protein glycation, 190f, 191, 192f
  tissue glycation, 189–193
Glucose-6-phosphate dehydrogenase enzyme deficiency, 346–347
Glucose metabolism
  and cellular metabolism, 191
  and lipid metabolism, 193
Glucose transporters (GLUTs), 190, 235, 190.e1, 190.e1f
  in glucose homeostasis, 194.e1f
Glutamate, 243–244, 265t
  cell depolarization, 344–345
  receptors of, 265–266
  retinal neurotransmitter, 265–266, 266f
  retinal receptor binding to bipolar cells, 265
Glutamate-receptor blocking agents, 360b
Glutamine, and disc shedding, 247, 248f
Glutathione
  redox systems involving, 209–210
  in redox systems of lens, 234b, 234f, 236
Glutathione peroxidase, 252–253
Glutathione-S-transferase, 193–194, 195f, 220
Glutathione SH, 237
Glycine, amacrine cells, 265–267
Glycosaminoglycans
  in cornea, 16–18, 120, 186
  extracellular matrix, 186
  in lens, 33
  in sclera, 21, 187b, 205–206
  structure of, 187b, 187f
Goblet cells
  in conjunctiva, 86f–87f, 88, 200–201
  epithelial, 202
Golgi apparatus, 164, 375f
  and exocytosis, 165b
  signaling, 164
Golgi complex, conjunctival epithelium, 202
Goniodysgenesis, 127b
Gorlin-Goltz syndrome, 523
Grafts, 450
  rejection, 450–451
Gram-negative bacteria, 463t
Gram-positive bacteria, 463t
Granules, 375f
Granulocyte-macrophage colony-stimulating factor (GM-CSF), 214, 402
Granulocytes, 378–380
  neutrophilic, 377
Granulomas, 411
Granulomatous inflammation, 499–500
Grating frequency, 284
Graves' ophthalmopathy, fibrosis of, 394
Gray, 487
Growth factor
  corneal, 214
  from retinal pigment epithelium, 255
Guanine triphosphate (GTP), 345, 345b
Guanylate cyclase, 261, 263f
Gut-associated lymphoid tissue (GALT), 416

H

H_1-receptors, 360
H_2 antagonists, 361
Haemangioma, 521, 534
  epithelioid, 534
Haemangiopericytoma, 534
Haematopoietic stem cells, 412–413
Haemoglobin glycation, 191–193, 192f
Haemopoeitic stem cells, in cornea, 207
Haemorrhage, 509
Hallerman-Streiff syndrome, 129.e5b
Hamartomas, 521–522
  astrocytic, 532f
Haplotype, allelic differences, 426b, 426f
Hard exudates, 509, 509f
Hardy-Weinberg equilibrium, 140
Hassall-Henle warts, 518
Hassall's corpuscles, 413–414
Heavy (H) chain, 440f
Helminths-nematodes, 463t
Hemianopia
  contralateral homonymous, 100f, 102f
  ipsilateral nasal, 99f
  pupil size, 286
Hemidesmosomes, 14–15, 174, 189, 189t
  in corneal epithelium, 203
Henderson-Hasselbalch equation, 340
Heparan, 187, 187f
Heparan sulphate, 226
Hereditary causes, of congenital malformations, 129.e4
Hering's colour-opponent scheme, 306
Herpes simplex ophthalmicus, 495
Herpes simplex virus (HSV)
  acute renal necrosis, 496
  keratitis, 462, 495, 496f
Herpes simplex virus 1 (HSV-1), 471
Herpes stromal keratitis, 471–472
  angiogenesis and lymphangiogenesis, 471
  latency of, 471–472, 472f
  primary infection, 471
Heterozygotes, 137, 140
High endothelial venules (HEVs), 418
High-resolution banding, 130
Histamine
  inhibition of, 360–361
  release from mast cells, 360
Histamine receptor antagonists, 360b
Histaminergic system, 359–361
Histidine-rich glycoprotein (HRG), 185
Histones, 135
HLA-A29, 139
HLA-B27, 139, 425
  acute anterior uveitis, 459
Homeobox, 110
Homeostasis, and intermediate filaments, 174, 176f
Homeotic genes, 110
Homer-Wright rosettes, 529, 531f
Homocystinuria, 504–505
Homozygotes, 137

Hordeolum, 83b
Horizontal cells, 266–267
  feedback, 292f, 294–295
  ON/OFF microcircuitry, 291
  in retina, 43f, 47f, 48–49
Horizontal gaze centre, 330–331, 332f
Horner syndrome, 85b, 357.e1b
Horopter, 313–314, 315b, 315f
Host defences, at ocular surface, 463–469
Host response, injury, 382–395
Hox genes, 110, 124
  retinoic acid, 129.e6
Hue-discrimination curve, 272, 272f
Hues, 272
  discrimination, 303
Human genome
  DNA analysis, 142–146
  haploid, 132–133
  mutations within, 137b
Human leucocyte antigens (HLA), 139, 141
  antigen presentation, 432
  disease susceptibility, 425
Human papillomavirus, 519–520
  E6 gene, 451–452
Huntington's disease, 137b
Hyalinization, 505–506, 508, 508f
Hyalocytes, 38, 39f
Hyaloid artery, 104f–105f, 106–108
  development of, 117, 121
Hyaloid canal, 117
Hyaloid fossa, 37, 37f
Hyaloid system
  development of, 121–124
  malformations of, 124b
Hyaluronan, 123, 186
Hyaluronate receptor, 253–254
Hyaluronic acid, 186, 186f, 215
  in aqueous humour, 223, 226f
  in vitreous, 240–241
Hydrogen peroxide, 194, 194b, 194f, 209–210, 200.e1, 200.e1f
  ciliary body, 221
  detoxification by trabecular meshwork, 227
  host response to injury, 383
  and iron, 253
Hydropic swelling, 490
3-Hydroxy-kynurenine, 238–239, 248b
Hydroxyamphetamine, 359.e1
Hydroxycholesterols, 197
Hydroxyeicosatetraenoic acids (HETE), 197, 197b, 197f
Hydroxyoctadecadienoic acids (HODE), 197
5-Hydroxytryptamine (5-HT), 363
5-Hydroxytryptamine (5-HT) antagonists, 363
Hyperacuity, 280, 282
Hypercapnia, and retinal blood flow, 246
Hyperglycaemia, chronic, 191
Hyperopia, 209.e1
Hyperoxia, and retinal blood flow, 246
Hyperparathyroidism, 507
Hyperpolarization response, 263–264, 288, 291
Hypertension, 510
Hypertrophy, 490
Hypervitaminosis D, 507
Hypochlorous acid (HOCl), 385f
Hypophyseal fossa. *see* Pituitary fossa
Hypophysis cerebri, 9
Hypothalamus, 'place' cells, 310
Hypovitaminosis A, 354
Hypoxia-inducible factor, 392–393
Hypoxia-inducible factor-1, 190

I

Ibuprofen, 362–363
Idiopathic orbital inflammation, 538, 538f
Ilivien®, 351
Image disparity, 317, 313.e1f
Image fusion, 313
'Image stabilization', 309
Immediate hypersensitivity, 381
Immune cells, activation of innate, 389–390
Immune complexes, 416–418
Immune response, 370
  acute, 382–392
  adaptive, cytokines, 399–404
  innate, 382–395
    cytokines, 399–404
  interferons, 401–402
  interleukins, 401–402
  primary, 372
  tumours induce, 451–453
Immune system, 370–373, 416–421
  acquired, 370–371
  adaptive, 453–454
  in ageing and degeneration in eye, 460–461
  cells, 373–382, 375b–376b
  effector mechanisms, 396–412
  evolution, 372
  eye, 453–461
  host-defence of, 371
  innate, 370–371
    and eye, 453
  molecules of, 373–382
  mucosal, 416
  organization of, 412–421
  parasitic infections, 444–445
  surveillance, 416
  tissues, 375b–376b, 377–378
Immunity, 370
  acquired, 380–381
    development, 372
  active, 372
  cell-mediated, 411
  innate, 377–380
    collectins, 407b, 407f
    complement, 407b, 407f
    mast cell, 448–450
    mediate, 390–392
  passive, 372
Immunization
  active, 372
  passive, 372
Immunoglobulin(s)
  isotypes of, 397, 398b, 398f
  superfamily of proteins, 428b, 428f
Immunoglobulin A (IgA), tears, 380–381, 453–454, 463
Immunoglobulin E (IgE), 379
  in hypersensitivity response, 360
  mast cell degranulation, 447
Immunoglobulin G (IgG), tears, 463
Immunohistochemistry, 535b
Immunological memory, 395–396
Immunological synapse, 434f, 436f
  cell signalling, 423–424, 435–437
Immunological tolerance, 439–445
  B cell, 443–444
  failed, 444–445
  induction mechanism, 442–443, 442.e1
  T cell, 443–444
Immunology, 370–461
Immunopathology, of tissue reactions, 420t
Immunoreceptor tyrosine-based activation motifs (ITAMs), 435–436
Immunosuppressive agents, 364–366, 459–460
  genetic control of, 347b
  mode of action of, 364b
  side effects of, 365b
In situ hybridization, 146, 496, 497b, 497f
Indometacin, 362–363
Infants, visual acuity testing, 286
Infections
  see also Fungal infections
  barriers, 453
  in developing countries, 480–483
  epithelium, as barrier for, 463–464
  helminth, 480
  inflammation and, 494–499
  innate immunity to, 469
  microbes in environment, 462–463
  microbial, 462–485
  ocular, 469–480
  physical barriers for, 463
  protozoan, 477–480
  pyogenic, 494, 494f
Inflammation, 492–504
  acute, 492–494, 493f
  chemical mediators, 492–493
  chronic, 393–394
  contact lens-related, 474
  due to autoimmune disease, 500–504
  effector cells, 392
  granulomatous, 499–500
  idiopathic orbital, 538, 538f

  - infection causing, 494–499
  - intraocular, 459–460
  - non-granulomatous, 500
  - phagocytosis, 389
  - resolution of, 392–393
  - systemic response, acute, 394–395
  - vascular changes, 386
  - without infection and, 499
- Inflammatory disorders
  - chronic, 394
  - endogenous material reactions, 500, 500f
  - exogenous non-biological materials reactions, 500
  - orbital, 456–457, 458f
  - scleral, 456–457, 458f
  - viruses and, 495–497
- Inflammatory mediators, 383
- Inflammatory response, 462
  - acute, 382–392
    - phases, 382–392
    - tissue damage, 382–383
    - vascular changes, 386
  - chronic, 412
  - stages of, 393f
- Infliximab, 365–366, 367f
- Information processing, hierarchical organization, 325
- Infranuclear lesions, of facial nerve, 80b
- Infratentorial compartment, 12.e1
- Inheritance
  - autosomal, 137–138
  - Mendelian, 138b
  - mitochondrial, 138, 156
  - multifactorial, 138–139
  - ocular disease, 152t
  - X-linked, 137–138
- Injury, host response, 382–395
- Innate immunity, to microbial infection, 469
- Innate lymphoid cells, 382
- Inositol triphosphate (IP_3), 261
- Instantaneous parallax, 316–317, 317f
- Insulin, 190
- Insulin-like growth factor-1 (IGF-1), 244–245
  - receptor, 250
- Integrin, 160, 160f, 182, 188, 188.e1
  - binding to extracellular matrix proteins, 188t
  - receptor, 185f, 189
- Integrin $\alpha_v\beta_5$, 246, 255, 261–262
- Integrins, 386
- Intercellular adhesion molecule 1 (ICAM-1), 253–254
- Interferon α, 404f
- Interferon-regulating factors, 403.e2
- Interferons, 401–402, 403.e1–403.e2
- Interleukin-1, 403.e1
- Interleukin-2, 403.e1
- Interleukin-4, 403.e2
- Interleukin-6, 403.e2–403.e3
- Interleukin-8, 403.e3
- Interleukin-10, 403.e3
- Interleukin-12, 403.e3
- Interleukin-15, 403.e3
- Interleukin-17, 403.e3
- Interleukin-18, 403.e3
- Interleukin-21, 403.e3
- Interleukin-27, 403.e3, 403.e5f
- Interleukin-32, 403.e3
- Interleukin-33, 403.e4
- Interleukin-34, 403.e4
- Interleukin-36, 403.e4
- Interleukin-37, 403.e4
- Interleukin-38, 403.e4
- Interleukins, 214, 401–402
- Intermediate filaments, 173–174
  - classes of, 174
  - distribution of, 175f
  - stress response, of cells, 176f
- Intermembrane space, in mitochondria, 164–166
- Interphotoreceptor binding protein (IRBP), 41–45
- Interphotoreceptor matrix, 41–45
- Interphotoreceptor retinol-binding protein (IRBP), 253–254
  - and retinoid transport, 257
- Intracellular matrix, 165b, 168b, 169–177
- Intraocular pressure, 223f
  - adrenergic receptor regulation, 225
  - aqueous humour dynamics, 222
  - and cAMP levels, 225
  - control of, 358–359
  - and guanylate cyclase, 225
  - M_3 muscarinic receptors and, 223–225
  - and melatonin, 225
  - and nucleotides, 225
- Intraparietal sulcus (IPS), 325.e1f
- Intraretinal space, 108–110
- Intravitreal delivery, of drugs, 352–353
- Intravitreal inserts, 351
- Intravitreal therapy, 353f
- Introns, 132–133
- Invaginating midget bipolar cells, 46
- Inversion, 136
- Ion channels, 344–345
- Ion exchanger, collagen as, 350–351
- Ionizing radiation, 487
- Iontophoresis, ocular, 352
- Iridocorneal endothelial syndrome, 518
- Iridogoniodysgenesis, 127b
- Iris, 26–30, 27f, 216–217
  - anterior border layer of, 26–28, 27f, 29f
  - blood flow in, 216–217, 217f
  - blood supply of, 28
  - capillaries, 28
  - collagens, 182
  - development of, 125f–126f, 126–127
  - dilator pupillae muscle of, 28, 126
  - eicosanoids in, 220
  - epithelium, 126
  - functions of, 216
  - hypoplasia, 120b
  - innervation, 126–127
  - melanomas, 526
  - microscopic anatomy of, 26–28
  - nerve supply of, 30
  - neuromuscular junctions of, 216
  - physiology of, 216
  - pigment epithelium of, 28
  - pigmentation of,
  - smooth muscles of, 126
  - sphincter pupillae muscles, 126
  - stroma of, 28, 29f
  - structure of, 26, 27f
  - transillumination, 155
  - trauma response, 492
- Iris naevus syndrome, 518
- Iritis, antimuscarinic agents for, 356–357
- Iron homeostasis, 253, 253f
- Irradiation therapy, complications of, 488b, 488f
- Irvine-Gass syndrome, 359.e1
- Isoprostanes, 197, 197b, 197f
- Isotype switching, 397

J

- Jak-STAT pathway, 404f
- Jaundice, bilirubin deposition in, 21b

K

- Kaposi sarcoma, 534
- Kayser-Fleischer rings, 489b
- Kearns-Sayre syndrome, 156, 490b
- Keratan sulphate, 186–187, 187f, 226
  - in cornea, 205
- Keratin, 174, 203
  - in conjunctival epithelium, 202
- ‘Keratinoid’, 517
- Keratitis, 455–456
  - *Acanthamoeba*, 477–478, 477f, 498–499
  - disciform, 495
  - herpes simplex virus, 495, 496f
  - herpes stromal, 471–472
  - postherpetic disciform, 455–456
- Keratoblasts, 119–120
- Keratocan, 205
- Keratoconjunctivitis sicca, 354, 456–457
- Keratoconus, 209
- Keratocytes, 16–18, 17f–18f
  - corneal, 207
  - metabolism of, 209
  - specialization, loss of, 212–213
- Ketone reductase, 349b
- Kinetochore engineers chromosome, 174
- Klinefelter syndrome, 136.e1
- Knudson’s two-hit hypothesis, 154–155

L

Lacrimal apparatus, 89–92, 89f
  collecting portion of, 90f, 91–92
Lacrimal component, of orbicularis oculi, 77–80
Lacrimal fossa, 2
Lacrimal gland, 89–91
  adenocarcinoma, 537
  blood supply of, 91
  development of, 129.e4
  histological structure of, 89–91, 90f
  inflammatory disease of, 456–457
  nerve supply of, 91, 91f
  secretion, 199–200
  tumours, 536–537
Lacrimal puncta, 82, 91–92
Lacrimal sac, tumours, 537
Lacrimation, excess, 91
Lacritin, ocular infection barrier, 463
Lactic acid production, in retina, 242–243
Lactoferrin, 200, 209
  ocular infection barrier, 453, 463
Lacus lacrimalis, 82
Lamina cribrosa, 59–61
Lamina fusca, 21, 22f
Lamina vitrea, 55, 56f–57f
Laminin, 179, 179f, 183, 244–245, 208.e2
  of the internal limiting membrane (ILIM), 184
  isoforms of, 183
  structure of, 208.e2f
Lamins, 174
Langerhans' cells, 416
Laser in situ keratomileusis (LASIK), 213
Lasers, in ophthalmology, 487t
Latanoprost, 360b
Lateral geniculate bodies, 93f–94f, 94–95
Lateral geniculate nucleus (LGN), 269–270
Lattice dystrophy, 516
Leber's hereditary optic atrophy, 490b
Leber's hereditary optic neuropathy, 156
Lectins, 405
Leiomyoma/leiomyosarcoma, 533
Lens, 32–35, 34f, 228–240
  ageing in, 237–240
    changes, 36b
  capsule, 33, 106–108, 234
    development of, 121
  changes in, 230f
  collagens, 182
  cytoskeletal proteins of, 232–233
  development of, 120–121, 122f–123f
  epithelium, 33, 34f
    anterior, 121, 123f
    function, 228
  equatorial, 229
  fiber cell organization, 228–230, 229f–230f
  fibres, 33–35, 34f
  free radical damage and, 239b, 239f
  insult, 238
  intra-lens microcirculation, 244b
  lipids, 236, 237f
  malformations of, 121b
  membrane lipids and proteins of, 233
  metabolism in, 235–237
  microcirculation in, 229f–230f, 234
  physiology of, 234–235, 244b
  placode, 104f–105f, 106, 120–121, 122f–123f
  position of, 33
  proteins, 235–236
    carbamylation of, 239–240
  redox systems in, 236–237
  semipermeable membranes of, 234–235
  shape of, 33
  size of, 33
  structure of, 33–35, 34f, 230.e1f
  transparency of, 228–235, 228f
  trauma response, 492
Lens bow, 33, 34f
Lens cavity, 120–121, 122f–123f
Lens fibres
  primary, 108–110, 121, 123f
  secondary, 121, 123f
Lens-induced uveitis, 501–502, 502f
Lens pit, 104f–105f, 106–108
Lens pore, 120–121, 122f–123f
Lens vesicle, 104f–106f, 106–108, 120–121, 122f–123f
Lens zonules, 30, 35, 36f
  accommodation in, 35
  chambers in, 35–37, 37f
Lenticular fossa, 37, 37f
Leptomeninges, 9
Leptotene, 131
Leucocyte
  activation, 386
  adhesion
    to each other, 418–419
    to endothelium, 386–387, 387f, 388t
    sequence of events in, 387f
  migration, 387–389
  tears, 453
Leukotrienes, 361–362, 362b
Levator palpebrae superioris, 81–82
LFA-1, 435
Li-Fraumeni syndrome, 520–521
Lids, 202–203
  closure of, 202
  lacrimal glands in, 202
  movement, 202
    control of, 202
  mucus-secreting glands, 202
Ligand-gated ion channel, 355
Ligand-receptor affinity, 346
Ligand selectivity, 346
Light
  physical behaviour, 282
  reflectance, 319
  sensing, 270–271
Light adaptation, 275
  cone thresholds, 303–304
Light (L) chain, 440f
Light detection, 273–279
  threshold, 277–278
Light energy
  measurement, 278b
  quanta, 273–274
Light exposure, 486–487
Light reflexes, pupillary, 286, 286b, 286f
Light rise, 298, 298b, 298f
Limbus, 21–26, 23f
  anatomical and surgical, 24b
Liminal brightness increment (lbi), 304
Limiting lamina
  anterior. *see* Bowman's layer
  posterior. *see* Desçemet's membrane
Line orientation, 312
Lingual gyrus, 95
Linkage analysis, 140–141
Linkage disequilibrium, 141
Linkage equilibrium, 141
Lipid(s)
  abnormalities, 354
  antigen presentation, 432
  membranes, oxidation of, 197b, 197f
  metabolism, 193
  peroxidation, 193f, 194–198, 197b
  retinal, 245
  structure of, 196b, 196f
  tear, 453
  tear film, 198–199, 199f
Lipid bilayer, 162f, 163–164, 197f
  formation of, 195
Lipid deposition, 506
Lipid phosphatase (PTEN), 255
Lipid rafts, 161–163
Lipocalin, ocular infection barrier, 463
Lipofuscin
  accumulation, 261–262
  and photoreceptor renewal, 261–262
Lipopolysaccharide, 389
Liposomes, 351–352
Lipoxidation, 197
Lisch nodules, 489b
Listing's law, 330, 331f, 335
Loa loa, 482, 499
Local anaesthetics, 366–368
  excitable membrane effects of, 366–368
  regional, maximal safe doses of, 368b
  side-effects of, 368
  topical, 367
Localized amyloid deposition, 507–508, 508f
Lod score, 140–141
Lodoxamide, 361
Loiasis, 482, 483f
Lumican, 205
Luminance
  contrast sensitivity, 284
  visual acuity, 280, 286
Luminosity, 302–303

Lutein, 253, 254f
Lyme disease, 494–495
Lymph nodes, 414f
  for antigen trapping, 414–415
Lymphatic system, 415
Lymphocytes, 419–420, 419f
  activated, 418
  in conjunctiva, 88
  conjunctival intraepithelial, 202
  effector cells in inflammation response, 392
  inflammatory reactions of, 500
  switching off, 420
  thymic education, 442
Lymphocytic perivasculitis, 500
Lymphoid cells, innate, 382
Lymphoid hyperplasia, benign, 535
Lymphoid organs, functional anatomy, 412–416
Lymphoid system, 380–381
  trafficking of cells to, 418–419
Lymphoid tissue inducer cells, 381–382
Lymphokine-activated killer (LAK) cells, 452
Lymphoma, 534
Lymphomyeloid cells, development of primary sources of, 412–413
Lyonization, 138b
Lysosomes, 492–493
Lysozyme, 453
  ocular infection barrier, 463
Lysyl oxidase, 227
Lytic granules, 375f

M

M cells, 291, 305–306
  contrast sensitivity, 284–285
Macrophages, 375f, 377, 411, 377.e1f
  activated, 392
  in cornea, 207, 464
  granulomatous reactions of, 494–495, 495f
  from myeloid and lymphoid precursors, 378b, 378f
  professional phagocytes, 389
Macrovascular occlusion, 510–511
Macula
  development of, 115–117
  false, 313
Macula lutea, 39
Macular degeneration
  age-related, 58b, 460f, 513–514, 514b, 515f
    complement activation, 409–410, 410f
  disciform, 513–514
Macular dystrophy, 516–517, 517f
Macular oedema
  cystoid, 359.e1
  rheumatoid disease, 501
Maculopapillary bundle, 59
Maculopathies, hereditary, 153
Magnetic resonance imaging, functional (fMRI), 321, 325
Magnetoencephalography, 321
Magnocellular pathways, 321–322
Major histocompatibility complex (MHC), 15–16, 372, 424–427
  alleles, 425
  antigen presentation, 423–424, 432, 422.e1f
  antigen recognition, 421
  class II-peptide complexes, 397
  gene cluster, 424–425
  gene map, 425b, 425f, 403.e4b–403.e5b
    extended, 425.e1, 425.e1f
  gene regulation/transcription, 427
  genes, 425b, 426–427
  innate immunity, 390
  peptide, 423.e1
  polymorphisms, 425
Major histocompatibility system, 424–427
Malabsorption syndromes, night blindness, 258–259
Malignancy
  epithelial tumours, 523–525
  lacrimal gland tumours, 536–537
  lymphoid tumours, 534–536, 535f
  peripheral nerve sheath tumours, 529
Malignant melanoma, 521, 525–526
  enucleation, 526
  histological classification, 526, 529f
  treatment, 526
  uveal, 525–526, 527f
Malignant transformation, 519
Malnutrition, night blindness, 258–259
Mammalian target of rapamycin (mTOR), 164, 164.e3, 164.e3f
Mandibulofacial dysostoses, 129.e5b
Mannan-binding lectin (MBL), 405
Marfan syndrome
  ectopis lentis, 35b
  fibrillin mutations, 185–186
Marx's line, 201
Mast cell(s), 376f, 378–379
  activation, 446f–447f, 447
  allergic disease, 447
  in choroid, 58f
  connective tissue, 379, 446–447
  degranulation, 447–448
  histamine release from, 360
  immune response, 448–450, 449f
  isotype switching, 448f
  mucosal, 379, 446–447
  types, 447
Maternal age, 129.e4
Matrilysin, 214
Matrix metalloproteinases (MMPs), 185, 214, 384t, 208.e2
  host response to injury, 383
  structure of, 384f
Matrix sheath, insoluble, 219f, 254
Mature B lymphocyte, 375f
Mature T lymphocyte, 375f
Maxillary nerve, 75, 78f–80f
Maxillary sinus, 4f, 8f, 7.e1
Mean channel lifetime, 344–345
Measuring by eye, 311–312
Megakaryocyte, 376f
Meibomian gland duct, blocked, 500
Meibomian glands, 202, 129.e4
  inflammation of, 354
Meiosis, 131–132
Melanin
  albinism, 155f
  drug binding, 368
  retinal pigment epithelium, 108–110, 253, 256–258
Melanocytes, iris, 126–127
Melanomas, cutaneous, 147b
Melanopsin, 247, 250–251, 251f
Melanosis, primary acquired, 525
Melanosomes, 155
MELAS syndrome, 490b
Melatonin, 221
  circadian rhythms, 267, 276–277
  disc shedding, 247, 248f
  function, 276–277, 277f, 277.e1f
  and intraocular pressure, 225
  production regulation, 296
  synthesis and secretion, 276
Membrane attack complex (MAC), 408
Membranous bones, 129.e1
Memory
  shape detection, 320
  spatial constancy, 310
Memory B cells, 420
Memory T cells, 419
Mendelian randomization, 141
Meningeal arteries, 11f–12f, 12, 7.e2f
Meninges, 9–13, 11f–12f, 9.e1f
Meningitis, 62b
  neck rigidity and, 12.e1
Mercapturic acid pathway, 236
Mesencephalic nucleus, 77
Mesenchyme, 103, 104f–106f, 108–110, 106.e1b
  corneal development, 119–120
  fronto-nasal process, 129.e4
  nasolacrimal duct, 129.e4
  neural crest derived from periocular, 111–112
  periocular, 113f
Mesoderm, paraxial, 112, 114f, 129
Messenger ribonucleoprotein particles (mRNPs), 169, 171f
Messenger RNA (mRNA), 133b, 135
  gene probes, 146
  Northern blotting, 142
Metabolic disease, 504–505
Metalloproteinases, 501
Metaphase, 131–132
Metaplasia, 490

Metastases, 521
  skeletal, 507
Metastatic tumours, 537
Metazoal infection, inflammatory, 498–499
Methotrexate, 364b
Methysergide, 363
Metzincin proteases, 384f
Micro-RNA, pathway of, 136f
Microaneurysms, 509
Microbes, in environment, 462–463
Microbial infections, of eye, 462–485
Microbiome, 159, 374b, 374f
Microfilaments, 169–173, 172.e1b
  actin bundles in, 172
  in exocytosis, 165b, 169–173
  proteins associated with, 172–173, 173t
Microglia, in retina, 50f–51f, 52
Microinfarction, 509, 509f
Microorganisms, activation of innate immune cells, 389–390
Microphthalmia, 112b
Microsomal enzyme system, 342–343, 343b
Microtubules, 174, 177f
Microvascular occlusion, 509–510
Midbrain
  lesions, 331
  reflex eye movement, 330–331, 332f
Midget bipolar cells, 280–282
Midget ganglion cells, 280–282
  microcircuitry, 294
Mimecan, 205
Miosis, 28
Missense mutation, 137b
Mitochondria, 164–169, 375f
  and apoptosis, signaling pathways in, 168b
  membrane organization in, 167f
Mitochondrial disorders, 156, 490b, 490f
Mitochondrial inheritance, 138, 156
Mitochondrion inner membrane organizing system (MNOS), 164
Mitogen activation of cells, signaling of, 164, 166b, 166f
Mitosis, 130–131
  stages of, 131b
Moesin/yurt, 255
Molecular biology, 146
Molecular genetics, 132–135, 150–156, 152t
  ophthalmology and, 150–156
Molecular mimicry, 444
Molecular typing, 450
Monoamine oxidases, 363
Monochromatism, 307
Monoclonal antibodies (Mabs), 399
Monocular rivalry, 317–318, 319f
Monocular vision, 309–313
Monocytes, 375f, 377
  migration of, 387
Monosomy, 136.e1
Mooren's ulcer, 455–456
'Morning glory syndrome', 119b
Morula, 103.e2
Motion detection, 319–320, 320f, 324–325
  rods, 271
Motion displacement, detection of minimum stimulus for, 279
Motor root, of trigeminal nerve, 74
Mould, 463t, 475
'Movement field', 332
Mucins, 200, 201f
  deficiency of, 354
  ocular surface, 463–464, 463f
Mucolytics, 355
Mucormycosis, inflammatory disease, 497–498, 498f
Mucosa-associated lymphoid tissue (MALT), 202, 416
Mucosal immune system, 416
Müller cells, 49–52, 50f–51f
  development of, 114–115
  glucose metabolism, 243
  glycogen stores, 190
'Müller's muscles', 30
Multifactorial traits, continuous or discontinuous, 138–139
Multifocal electroretinogram (ERG), 299, 300f
Muscle fibres, of ciliary body, 30
Muscular dystrophy, 189
Mutation, 136
Mycobacterium tuberculosis, 494–495
Mycophenolate mofetil, 364
Mydriasis, 28, 216
Myeloid-derived progenitor (MDP) cells, 378.e1f
Myeloid mononuclear cells, 377–378
Myeloid system, 377–380
Myeloma, 507
Myocillin, 227
Myocytes
  eyelid, 129.e4
  migration, 129
MyoD gene expression, 129
Myoid, 41–45
Myopia, 215, 209.e1
Myositis, orbital, 456
Myotonic dystrophy, 137b, 155–156

N

Na^+/Ca^{2+} exchanger protein, 249–250, 259–261
NADH, 189
NADPH, 189, 209
Naevi, 521–522, 522f
Na^+/H^+ exchanger protein, 207–208, 208b
Na^+/K^+ATPase, 208–209
  in lens, 235
Nanodiscs, 188.e1, 188.e2f
Nanophthalmia, 112b
Nasal cavity, 1
Nasal mucosa, drug absorption in, 347–348
Nasociliary nerve, 69–71
Nasolacrimal duct, 92, 129.e4
  developmental anomalies of, 129.e6b
Nasolacrimal groove, 129.e4
Nasopharyngeal mucosa, drug absorption in, 347–348
Natural killer (NK) cells, 375f, 381, 390–392
  membrane damage, 412
Natural killer-T (NK-T) cells, 375f, 390–392
N-CAM, 255
Necrosis, 489
Neoplasia, 518–538
  pathogenesis of, 519–521
Neoplasms
  classification of, 519
  misdiagnosis, 537–538
  pathogenesis of, 519–521
Neovascularization
  preretinal, 511
  of retina, 510
Nervus intermedius, 77
Nestin, 148–149
Nettleship-Fells syndrome, 155
Neural cells, photoreceptor responses, 305–306
Neural crest, 104f–105f, 103.e3f
  cell migration and differentiation, 112, 114f
  fate of cells, 111
  mapping of, experimental methods of, 112.e1b
Neural layer, 14f
Neural plate, 103
Neural progenitor cells, 148
Neural tube, 103, 104f–105f
  malformations of, 112b
Neural tumours, 526–533
Neuroblastic layer, 108–110
Neurocan, 215
Neurocranium, 1
Neurocristopathies, 129.e5b
Neurofibroma, 526–529
Neurofibromatosis type 1, 489b
Neuroglia, retinal, 49–52
Neurone, visual stimulus detection, 270–271
Neurospheres, 148–149, 150f
Neurotransmitter(s), 265
  corneal, 210
  retinal, 265t
  serotonin as, 363
Neutrophilic granulocytes, 377
Neutrophils, 376f, 377
  allergic disease, 448
  antibacterial agents, 383t
  bacterial killing, 467–469, 468b

  infiltration and activation of, 470f
  migration, 388f
  professional phagocytes, 389
  pyogenic infection, 494
  recruitment of, 464
  respiratory burst, 383, 385b, 385f
New ophthalmic delivery systems (NODS), 351
NF-1 gene, 520–521
Night blindness, 258–259
  choroideraemia, 153–154
  congenital stationary, 150–151
  in retinitis pigmentosa, 514
Nitric oxide
  host response to injury, 383, 385b, 385f
  and retinal blood flow, 246
Nitric oxide synthase (NOS), inhibition of, 360b
N-methyl D-aspartate (NMDA), 360b
NOD-like receptors (NLRs), 464–467, 467b, 467f–468f
NOD2 gene, 464–466, 467b
Non-granulomatous inflammation, 500
Non-steroidal anti-inflammatory drugs (NSAIDs), 361
Nonsense mutation, 137b
Norepinephrine
  sympathetic system, 357, 357f
  synthesis of, 358f
Norma frontalis, 3f
Norma lateralis, 3f
Norrie's disease, 154, 154t
Northern blotting, 142–143
Notochord, 103.e2
Nouin's technique, 317
Nuclear factor-κB (NF-κB), 389
Nuclear lesions, of facial nerve, 80b
Nuclear pores, 164
Nucleolus, 169
Nucleotide substitution, 137b
Nucleus, 166b, 169, 170f
  intermediate filaments, 173–174
  ligands, receptors for, in, 169
  receptor mediated transport in, 166b, 169
Nutritional deficiencies, 488

O
Object positioning, 309–310
Obligate intracellular, 463t
Oblique muscle, of eye, 67t
Occipitofrontalis, 81, 81f
Ocular albinism, 155
Ocular appendages, 77–92
Ocular dominance, 317–318
Ocular embryology, 103–111
Ocular inserts, 350, 351b
Ocular lubricants, 354–355
Ocular muscles, eye movements, 326, 327b, 327f
Ocular Response Analyzer® (ORA), 207
Ocular surface, biochemical processes, 198–203
Ocular surface disease, 501, 502f
  progressive, 454–461
Ocular toxicity, from systemic administration of drugs, 368–369
Oculocervical reflexes, 330
Oculocutaneous albinism, 155
Oculomotor cerebellar centre, 334–335
Oculomotor nerve, 69–71
  lesions of, 71b
  origin of, 69, 73f
  sensory endings in, 74
Oculotoxic drugs, 369
Oculovestibular reflexes, 329–330
OFF bipolar cells, 292–293, 293f–294f, 306
OFF response, 288–291
Ointments, 350
  ophthalmic, 354
Oligodendroglia, 118–119
Oligonucleotides, anti-sense/triple helix, 147
'OMICS' analysis, 145f
ON bipolar cells, 292–293, 293f–294f, 306
ON/OFF response, 288–291
Onchocerciasis, 480–482
  control of, 481–482
  life cycle and parasite burden, 480
  pathogenesis of, 480–481, 481f
Oncogenes, 521
Opacification, 212–213
Ophthalmia
  sympathetic, 139, 411f
    microgranulomas, 411
  systemic, granulomatous inflammation of, 502, 502f
Ophthalmic artery, 7, 7f
  aneurysm in, 99f
Ophthalmic nerve, 75, 79f
Ophthalmic surgery, materials used in, 501b, 501f
Opponent colour theory, 306
Opsin, 172f, 258, 275–276
  and 11-cis retinal binding to, 259
  and chromophore pocket, 258–259, 260f–261f
  cone, 259, 262f, 304–305
Opsonins, 389
Optic canal, 1, 7, 10t
Optic chiasma, 92–93, 93f–94f
Optic cup, 104f–106f, 106–108
  margin, 112–113
    formation of, 112–113
Optic disk, 39, 40f
  development of, 117f, 118–119
Optic fissure, 106–108
Optic flow, 328–329, 328b, 328f
Optic nerve, 7, 59–61, 60f
  blood supply of, 60f, 61
  development of, 117f, 118–119
  glioma, 531–533
  intracanalicular portion of, 61, 62f
  intracranial portion of, 92, 93f
  intraocular portion of, 59–61, 60f
  lymphoma, 535–536
  malformations of head, 119b
  meningioma, 533
  orbital portion of, 60f, 61
  trauma response, 492
Optic pit, 119b
Optic radiations, 93f, 95, 96f
Optic stalk, 104f–106f, 106, 117, 117f
  development of, 117f
Optic sulci, 103, 104f–105f
Optic tracts, 93–94
  lateral root of, 94
Optic ventricle, 106
Optic vesicle, 103, 104f–106f
  malformations of, 112b
Optical coherence tomography (OCT)-guided femtosecond laser surgery, 213
Optical illusions, 312
  Pulfrich phenomenon, 317, 318b, 318f
Optogenetics, 148b
Optoneurin gene, 227
Ora serrata, 30, 40
Oral cavity, 1
Orbicularis oculi muscle, 77–80, 81f
  development of, 129.e4
Orbit, 1
  anatomy of, 1–102
  blood vessels of, 68, 70f
    anastomosis of, 70t
  blow-out fractures in, 6b, 6f
  contents of, 61–68
  cysts, 537–538
  extraconal space in, 63f
  fibroadipose tissue of, 62–64, 62f
  inflammatory disorders, 456–457, 458f
  intraconal space in, 62–63, 63f
  lipoma, 534
  osteology of, 1–7, 4f–6f
  osteosarcoma, 534
  periosteum of. *see* Periorbita
  walls of, 2–3, 5f
Orbital aperture, 2
Orbital cellulitis, 4b, 4f
Orbital fissure
  inferior, 5f–6f, 6–7
  superior, 5–6, 5f–7f, 10t
Orbital floor, 2, 5f
  fractures in, 7.e1
Orbital margin, 1, 3–5
Orbital muscles, swelling, 456
Orbital myositis, 456
Orbital plate, 7–8
Orbital roof, features of, 2
Orbital septum, 82
Orbital tubercle, lateral, 3

Orbital wall, 2–3, 5f
  lateral, 2–3, 5f
  medial, 2, 5f
Organ transplantation, solid, 451
Organomercuric compounds, 354
Orientation detectors, horizontal and vertical, 312
Osteoprotegerin ligand, 378.e1b
Oxidative stress, glucose-derived, 194
α-Oxoaldehydes, 191
Oxygen flux, 211
Oxygen supply, reduced, 486
Ozurdex®, 351

P

p-21 activated kinases (PAKs), 173
P cells, 291, 305–306
p53 tumour suppressor protein, 147b
P450 gene, in drug metabolism, 347
P_{450}-linked reductases, 193
Pachymeninx, 9
Pachytene, 131
Paired-box genes, 110
Palpebral fissure, 82
Palpebral ligament, 81–82
Palpebral part, of orbicularis oculi, 77–80
Palpebral sulcus, 82
Panretinal photocoagulation, 486–487
Panum's area, 317
Papillae lacrimalis, 91–92
Papilloedema, 62b
Papilloma, 523
Paramedian pontine reticular formation (PPRF), 329, 331, 332f
Paranasal sinuses, 1, 4f, 7, 8f
  development of, 129.e4
Parasitic infection, 444–445
Parasympathetic antagonists, 356–357
Parasympathetic system
  motor endplate of, 356f
  ocular drugs, 355–357
Parasympathomimetics, 355–356, 356b, 360b
Parvocellular pathways, 321–322
Patau syndrome (trisomy 13), 136.e1, 129.e6b
Patellar fossa, 37, 37f
Pathogen-associated molecular patterns (PAMPs), 371
Pathogen-recognition receptors (PRRs), 379f, 390–392, 464
Pathogens, 462, 463t
Pathology, 486–538
  ageing, 505–518
  cell and tissue damage, 486–490
  degeneration, 505–518
  dystrophies, 505–518
  healing and repair, 490–492
  inflammation, 492–504
  metabolic disease, 504–505
  neoplasia, 518–538
Pattern electroretinogram (PERG), 299
Pax-6 gene, 107f, 111, 120–121
Pax genes, 110
Pax2 gene, 111
Paxillin, 189
PECAM-1, 387
Pegaptanib sodium, 352–353
Pemphigoid, 454–455, 501
Pentose phosphate pathway, 190, 190f, 209
  in cornea and lens, 190
Peptidyl transferase, 133–134
Perception, superior colliculus, 331–333, 334f
Perforating/penetrating injuries, 486–487
Perforin, 412
  killing, 412f
Periorbita, 62–64, 62f
Peripheral nerve sheath tumours, malignant, 529
Peripherin gene, 153
Peroxiredoxin, 221
Peroxisome proliferator-activated receptors (PPARs), 169
Persistent hyperplastic primary vitreous, 532f
Peter's anomaly, 120b, 127b
  craniofacial abnormalities, 129.e5b
  infantile glaucoma, 127b
Phagocytes, professional, 389
Phagocytosis
  inflammatory response, 382
  removal of damage tissue and microorganisms, 389
Phagosomes, 375f, 492–493
Phakomatous choristoma, 522
Pharmacology
  autonomic nervous system and, 355–358
  drug-receptor interactions, 344–347
  ocular, 338–369
  pharmacodynamics, 344
  pharmacokinetics, 338–344
  tear film reconstitution, 354–355
Pharyngeal arches, 129, 129.e1, 129.e2f, 129.e6b, 127.e1f
  derivatives, 129.e3t
  neural crest migration pathways, 129.e6
Phosphatidylinositol
  degradation of, 345–346
  in lens, 236–237
Phosphodiesterases, 345
Phosphoinositide metabolism, 247
Phospholipase A_2, 197
Phospholipase C, 345–346
Phospholipids
  flip-flop mechanism of, 345–346
  structure of, 196b
Photo-oxidation, rate of, 239
Photochromatic interval, 303
Photometer, 302
Photoreceptors, 41–45, 44f, 161f, 246–254
  bleached, 275–276
  colour detection, 303–305
  damage, 246f, 251
  death signal for, 252, 252f
  degeneration, 154b
  electrical response, 287–288
  hyperpolarization, 288
  light energy, 273–274
  lipids in, 247
  metabolism, 246–247, 246f
  plasma membrane, 163–164
  potential difference, 288
  proteins in, 249–250, 249t
  renewal of, 244, 244b, 244f
    and lipofuscin accumulation, 261–262
  responses with neural cells, 305–306
  resting potential, 298
  retinal light stimulus detection, 291
  shedding in, regulation of, 247, 248f
  synaptic events with inner nuclear layer cells, 262–268, 265t
  turnover, 246f
  visual stimulus, 270–271
Photosensitivity, retinal, 275
Photosensitization, from systemic drug toxicity, 368–369
Phototransduction, 263f, 261.e1b, 261.e1f
  cascade, 261
  electrical response initiation, 287–288
  and phosphoinositide metabolism, 261
Phthisical eye, 507
Physical agents, 486–487
Physiological cup, 59–61
Pia mater, 11f–12f, 13, 9.e1f
'Pie in the sky' defects, 101f
Pierre Robin syndrome, 129.e5b
Pigment epithelium-derived factor (PEDF), 255
Pilocarpine, 223–225, 355, 356b
  intraocular pressure control, 360b
Pilomatrixoma, 523
Pineal tumour, 531
Pinguecula, 507
Pituitary fossa, 9
Pituitary gland, 9
Pituitary stalk, 9
Pizotifen, 363
'Place' cells, 310
Plasma cells, 375f
  inflammatory reactions of, 500
  monoclonal proliferation, 507
Plasma membrane, 161–164
  lipid domains in, 161, 163f
  proteins in, motion and tension of, 162b, 162f
  tiny domains, 161
Plasminogen activator, 214

Plasminogen activator inhibitor 1 (PAI-1), 159, 185, 191–193
Platelet-activating factor (PAF), 447
Platelets, 376f, 386
Plectin, 174
Pleomorphic adenoma, 536–537, 536f
Plica semilunaris, 85–86
Pluripotential cells, 149f
Point mutation, 137b
Poly-A polymerase, 133–134
Polyarteritis nodosa, 503
Polyclonal antibodies, 399
Polyclonal B-cell activation, 444
Polymerase chain reaction (PCR), 143, 496b
Polymorphisms
  DNA, 143
  genetic, 139
Polymorphonuclear leucocytes (PMNs), 377, 453
Polyploidy, 136.e1
Polyunsaturated fatty acids, in the retina, 245
Pontine nucleus, 77
Positional changes, 329–330, 331f
Positron emission tomography (PET), 321
Posterior pigment epithelium, 28, 29f
Posterior polymorphous dystrophy, 518, 518f
Posterior uveitis, microgranulomas, 411
Potassium channels, in lens, 235
Premalignancy, 519
Preotic region, 129
Presbyopia, 36b
Preservatives, for drugs, 354
Prestriate cortex, 269–270, 322–323
Primary intraocular lymphoma, 535–536, 536f
Primate brain, 335f
Primitive streak, 103.e2
Prochordal plate, 129
Prodrug, 342
Progressive outer retinal necrosis, 496–497
Proliferative vitreal retinopathy, 48b
Proline-arginine-rich and leucine-rich repeat (PRELP), 215
Promoter, 133–135
Prosencephalic vesicle, 103
Prostaglandin(s), 220, 361–362, 362b
Prostaglandin analogue, 360b
Prostaglandin D_2, 447
Prostaglandin E_2, 362b
Prostaglandin $F_{2\alpha}$, 362b
Prostaglandin I_2, 362b
Proteasome, 422.e1, 422.e1f
Protein
  epitope mapping of, 430
  glycation, 190f, 191, 192f
    non-enzymatic, 191
  metabolism, retinal, 231f
  synthesis, 133b
    regulation of, 346–347
Proteoglycans, 186, 186f, 186t, 188f
  fibril-regulating, 186
  organization of, 206b
Proteomics, 143–145
Proteosomes, 147.e1b
Proto-oncogenes, 521
Proton beam therapy, 487
Protozoa, 463t
Protozoal infection, inflammatory, 498–499
Pseudoepitheliomatous hyperplasia, 538
Pseudoexfoliation syndrome, 508
Pseudomonas aeruginosa, 462
  virulence factors, 473–474, 473f
Pterygium, 507, 507f
Pulfrich phenomenon, 317, 318b, 318f
'Pulley sleeve', 62–63
'Pulley suspensions', 62–63
Pulvinar, 332–333
Pupil, 26
  light reflexes, 286, 286b, 286f
  movements of, 28
  parasympathetic antagonists, 356
  size, 286
Pupil zone, 26
Pupillary margin, 26
Pupillary membrane, 121–123
Purine antagonists, 364
Purkinje cells, 334–335
Purkinje shift, 302–303, 302f
Pursuit, 327–329
Pyruvate, 239

Q

Quadrantanopia, contralateral homonymous superior, 101f

R

Radial streaks, of iris, 26
Radiotherapy, 487
Radixin, 255
Ranibizumab, 352–353
Ranitidine, 361
Receptive fields, 288–291, 290b, 290f
β-receptors, 358
Recombinant DNA, 142
Recombinant DNA technology
  cloning, 142f
  Southern blotting, 143f
  X chromosome, 150–151
Recombination rate, 140–141
Recoverin, 261, 264f
Rectus muscle, of orbit, 67t
Red-green differentiation, 294
Reflectance, 319
Regulatory cells, mode of action of, 443f
Release factor, 133–134
Repeat expansions, 137b
Resolving power of eye, 280, 282–283
Resting potential, 298
Restriction endonucleases, 142, 142f
  mapping, 146
Restriction fragment length polymorphism (RFLP), 143, 146
  X chromosome, 150–151
Reticular medial longitudinal fasciculus (RMLF), 331, 332f
Retina, 38–55, 39f
  ageing changes in, 50b
  amacrine cells in, 43f, 47f, 49, 265–267
  association neurones in, 43f, 47f, 48–49
  astrocytes in, 49, 50f
  astrocytic hamartoma, 531, 532f
  biochemical processes, 242–258
  bipolar cells, 43f, 45–46, 47f, 263–265, 266f
    colour opponency, 306
    colour perception, 306
    microcircuitry, 265–266, 291, 293f, 295f
  blood flow in, 245–246
  blood-retinal barrier, 245
    autonomic control of, 245–246
  blood supply of, 40f, 52–55, 53f–54f
  central, 39
  circuitry, 291–296
  coloboma, 108b
  cones in, 44f, 45
  detachment of, 40b, 241, 242f
  dichromatic, 45
  differentiation of layers, 112–113
  dopamine neuromodulation, 267–268
  dysplasia, 129.e6b
  electrophysiology, 287
  ganglion cells in, 43f, 46–48, 47f, 267f
  glycogen stores, 190
  horizontal cells in, 43f, 47f, 48–49, 266–267
    feedback, 292f, 294–295
  inner neurosensory, 38
  inner nuclear layer of, 114–115, 115f
    photoreceptor synaptic events, 262–268, 265t
  interphotoreceptor matrix, 253–254
    retinal adhesion for, 253–254
  light sensing cones, 307f
  lipids in, 245
  lymphoma, 535–536, 536f
  metabolic function, 242
  microcircuitry, 291, 296
  microglia in, 50f–51f, 52
  morphogenesis of, 114–115, 115f
  Müller cells in, 49–52, 50f–51f
  neural, 108–110
    axes in, 113–114, 114f

    derived from neuroepithelium, 112–118
    development of, 117
  neuroglia in, 49–52
  neurosensory, 41, 43f
  neurotransmitters in, 265t
  outer nuclear layer of, 114–115, 115f
  peripheral, 39
    development of, 117
  photochemical reactions in, 258–259
  photoreceptors in, 41–45, 44f
  photosensitivity, 250–251, 251f, 275
  posterior pole of, 39
  protein metabolism, 231f, 243–245
  regions of, 38–40, 40f
  rhodopsin content, 275
  rod cells in, 41–45, 44f
  shedding in, 247
  transient layer of Chievitz, 114, 115f–116f
  trauma response, 492
  trichromatic, 45
  vasculature, development of, 117, 117f
11-cis Retinal, conversion of, 257, 257f, 259, 260f–261f
Retinal artery, central, 40f, 52–55, 53f–54f
  development of, 117
  occlusion, 55b
Retinal capillaries, 53
  networks of, 54f, 55b
Retinal cells
  differentiation of, 148f
  electrophysiology, 288–291
Retinal connections, 291–296, 293f, 299.e1f
Retinal degeneration
  microcystoid, 528f
  routes for intervention for, 151f
Retinal degeneration show (RDS) protein, 153, 249–250, 250f
Retinal disk, 104f–105f, 106, 112–113
Retinal disorders, gene loci, 153f
Retinal dystrophies, 514–515
Retinal ganglion cells
  axons from, 118–119
  colour-opponent mechanism, 306
  embryology of, 118–119
  M cells, 305–306
  organization, 294–295
  P cells, 305–306
  responses, 306
  types, 296, 297f
  visual acuity, 285
Retinal neuroglia, 49–52
Retinal pigment epithelium, 40–41, 42f, 160–161, 161f
  ageing changes in, 41b
  bidirectional transport in, 257–258
  biochemical processes, 254–258
  and blood-retinal barrier, 255
  cell size, shape and structure of, 41, 42f
  development of, 116f, 117–118
  embryology, 108–110
  function of, 256f, 256t
  lipofuscin accumulation, 261–262
  lysosomes of, 262
  in macular degeneration, 513–514
  melanin, 108–110, 253, 256–258
  phagosomal system of, 256–257
  photoreceptor function and, 256–257
  pluripotent cell function, 254–256
  resting potential, 298
  and tight junctions, 257–258
  transport function, 258
  turnover, 255
Retinal rivalry, 317–318
Retinal slip prevention, 335
Retinitis
  acquired immunodeficiency syndrome, 497
  AIDS, 403.e2
  CMV, 459
    treatment of, 352.e1b
  cytomegalovirus, 472–473, 497, 497f
  viral, 472–473
Retinitis pigmentosa, 151–153, 251, 514–515, 515f
  albescens, 153
  autosomal dominant, 153
  autosomal recessive, 153
  X-linked, 151
Retinoblastoma, 154–155, 529–531
  differential diagnosis of, 531, 532f
  histology of, 529–531, 530f
  trilateral, 531
Retinoblastoma *(Rb)* gene, 154–155, 521
Retinoblastoma predisposition gene, 154–155
Retinoic acid, 259
  topical, and wound healing, 215
Retinoic acid receptors, 129.e6
Retinoid
  conjunctival epithelium, 202
  shuttling mechanism, 257
Retinol, isomerization of, 172f, 248, 248b, 248f, 257f, 258–259
Retinopathy
  from chloroquine, 368
  diabetic, 504
  myotonic dystrophy, 155–156
  of prematurity, 510b
  proliferative vitreal, 48b
Retinopathy of prematurity, 118b, 118f
Retinoschisin, 264–265, 266f
Retisert®, 351
Retrolental fibroplasia. *see* Retinopathy of prematurity
Reversion-inducing cysteine-rich protein with Kazal motifs (RECK), 383
Rhabdomyoma/rhabdomyosarcoma, 533, 534f
Rheumatoid arthritis, 455–456
  systemic amyloid deposition, 507
Rheumatoid eye disease, 501
Rhodopsin, 41–45, 249, 249f, 258–259, 261, 261.e1, 261.e1f
  activation of, with light, 258, 260f–261f
  acylation of, 246, 247f
  arrestin interaction, 430f
  bleached, 276
  electrical response, 287
  glycosylation of, 246, 247f
  insertion of, into disc plasma membrane, 244b, 246
  molecule, 153
  in photoreceptor disc, 172f
  regeneration, 275
  retinal content, 275
  synthesis of, 248b, 248f
Ribbon synapses, 264–265
Ribosomal RNA (rRNA), 138
Ribozymes, 147
Ricco's law, 278
Rieger's anomaly, 120b, 127b, 129.e5b
River blindness, 480–482
  see also Onchocerciasis
RNA, non-coding, in gene expression, 135
RNA-binding proteins, in gene expression, 135
RNA polymerase, 133–135
Robertsonian translocation, 136
Rod(s), 41–45, 44f
  bipolar, 46
  horizontal cell feedback, 292f, 294–295
  light photon absorption, 275
  microcircuitry, 293–294
  motion detection, 271
  outer segment tip phagocytosis, 262
  spherules, 44f, 45
  stimulation, 278
Rod monochromatism, 307
Rosenthal fibres, 533, 533f
Rough endoplasmic reticulum (RER), 164
RP3 gene defect, 153–154
RPE65 enzyme, 252
Ru plaque, 487
Rubin vase, 308–309, 309b, 309f

S

Saccades, 327, 330b, 330f
  initiation, 333
  voluntary, 333
Sarcoidosis, 456–457
Schiff base products, 191, 192f
Schlemm's canal, 24, 25f
  aqueous humour drainage, 227–228
  development of, 127, 128f
Schrodinger staircase, 308–309, 309b, 309f
Schwalbe's line, 120b

Schwannoma, 526–529
Sclera, 20–21, 22f
  ageing changes in, 22b
  biochemical processes, 215–216
  collagens, 182, 205
  development of, 119
  fibroblast in, 215
  fluid transport, 186f, 203–204, 215–216
  inflammatory disorders, 456–457, 458f
  matrix, 215
  nanophthalmia, 216
  stroma of, 22f
  trauma response, 492
Scleral canal, 59–61
Scleral spur, 21–22
  development of, 127
Scleral sulcus, 21–22
Scleritis, 456
  immunopathology of, 457f
Sebaceous adenomas, 523
Sebaceous gland carcinoma, 525, 525f
Second messenger systems, 159, 159f, 177
Second messengers, drug-receptor interactions, 345, 345b, 345f
Seeing, frequency, 274
Segmentation genes, 110
Selectins, ligand interactions, 386–387
Selector genes, 110
Selenite, cataract formation, 239–240
Self-organization, of eye, 157, 158f
Sella turcica, 10t
Semicircular canals, 329
Sensory root, of trigeminal nerve, 74
Septins, 177, 178f
Serotonin, 363
  receptors, 363
Shape
  detection, 319
  discrimination, 273
  processing, 273, 273f
Sheddases, 182
Signal amplification, 259–261, 263f
Signalling mechanisms, intracellular, 177–178
Signalling molecules, 157
Signalling networks, 159, 159.e1, 159.e1f, 159.e2b
Signalling pathways
  apoptosis, 168b
  endoplasmic reticulum, 164, 166b
  Golgi apparatus, 164, 166b
Single gene defects, 488–489
Sinus venosus sclerae. *see* Schlemm's canal
Sirtuins, 190
Six3 gene, 111
Sjögren syndrome, 354, 456–457, 500–501
  malignant transformation, 519
Skeletal metastases, 507
SKILL test (Smith-Kettlewell Institute Low Luminance test), 279
Skull
  arrangement of, 1
  chondrocranium, 129.e1
  development of, 129.e1
  features of, 1
  osteology of, 1–13, 3f
  viscerocranium, 129.e1
Slow acetylators, 346–347
Slow-releasing substance of anaphylaxis (SRS-A), 361
Smooth endoplasmic reticulum (SER), 164
Snellen's test type, 280–282
Sodium channels, cell depolarization, 287–288
Sodium coupled monocarboxylate transporters, 242–243
Sodium cromoglycate, 361
Sodium/potassium ATPase pump, 359.e2b
Solitary fibrous tumour, 534
Solutions, ophthalmic, 354
Somitomeres, 112, 114f
Sonic hedgehog, 111
Sorbitol, 191
  pathway, 191f, 235
Southern blotting, 142–143, 143f, 146
SPARC (sialo-protein associated with rods and cones), 185, 234, 253–254
Spatial constancy, 310, 322
Spatial perception, 309
Spatial resolution, 271
Spectral mixing curves, 303, 304b, 304f
Spectral reflectance densitometry, 304
Spectral sensitivity curves, 302–303
Spectrophotometers, 302
Sphenoidal sinus, 4f, 8f, 7.e1
Sphincter pupillae muscle, 26, 27f, 28, 126, 216
Sphingomyelin, lens fibre cell content, 233, 236, 237f
Spina recti lateralis, 2–3
Spinal nucleus, 77
Spinocerebellar ataxia, 137b
Spirochaetes, 494–495
Spleen, 415, 415f
Splice site mutation, 137b
Spliceosomes, 135
Spores, 475
Squamous cell carcinoma, 523–525, 524f
Staphylococcal enterotoxins (SE), 432
Staphylococcus aureus, 474
Stargardt's disease, 515, 516f
Stem cell(s), 412–413
  bone marrow, 438
Stereoacuity, 316–317, 317f
Stereopsis, 312–317, 324–325
  image disparity, 317
  measurement, 314–317
Steroids, 363–364
  for idiopathic orbital inflammation, 538
  as oculotoxic drug, 369
  in protein synthesis regulation, 346
  in trabecular meshwork, 227
Stevens-Johnson syndrome, 455
Streptococcus pneumoniae, 474
Striate cortex (area V1), 322–323, 323f
  processing, 323
Stroma
  ablation of, 213
  atrophy, 518
  dystrophies, 516–517
Stromelysin, 214
Structural proteomics, 145
Structure from motion (SFM) information, 310
Sturge-Weber syndrome, haemangiomas, 521
Subarachnoid cisterns, 12, 9.e1f
Subarachnoid space, 9, 9.e1f
Subbasal plexus, 20
Subdural haematomas, 13b
Subepithelial plexus, 20
Sublethal cell injury, 490
Subretinal space, 38, 108–110
  lymphoma, 535–536
Substantia nigra, 333
Substantia propria, 16–18, 18f
Subtarsal sulcus, 85
Sudoriferous cysts, 537
Sugars, antigen presentation, 432
Sulcus chiasmatica, 9, 10t
Summation, 288–291
  binocular, 279
  spatial, 278
  temporal, 278–279, 279.e2f
Sunflower cataracts, 489b
Superantigens, 431–432, 444
Superior colliculus, 94
  eye movement control/perception, 331–333
Superoxide anions, 193
  host response to injury, 383
Superoxide dismutase, 193–194, 194b, 194f, 252–253
Suprachoroid, 58
Supranuclear lesions, of facial nerve, 80b
Supratentorial compartment, 12.e1
Surface ectoderm, 103–106, 119–120
Surface epithelium, benign tumours of, 523
Sutures, 1
Sweat gland adenoma, 523
Sweat gland cysts, 537
Switch genes, 110
Symblepharon, 501
Sympathetic ophthalmia, 502, 502f
Sympathetic system, 357–358, 357f
Sympathomimetic drug receptor, profile/action of, 359.e2t
Synapse, 172.e1, 172.e1f

Synaptic potentials, 288–291
Synaptic ribbon, 45
Syndecans, 186–188
Synophthalmia, 112b
Systemic amyloid deposition, 507
Systemic lupus erythematosus, 444, 503–504
Systems biology, 159

T

T cell(s), 380
  activation, 427–437
  antigen response, 395–396
  clonal expansion, 395–396
  cytotoxicity, 412
  deletion, 444
  development and maturation, 413–414
  differentiation, 382f, 421f
  effector response, 396
  maturation, 413–414
  memory, 420
  stages of activation, 365f
  suppressive, 421
  tolerance, 443
  trafficking, 412
γδT cell(s), 390–392, 431
T cell receptor (TCR), 380
  antigen binding, 427–432, 429f
  chains, 429
  gene expression, 430
  immunoglobulin gene segment data, 428.e1
  regions, 429f
T helper (Th) cells, 380, 447–448
  response in parasitic infections, 449f
T regulatory cells, 420–421
  active suppression by specific lymphocytes, 443
Tacrolimus, 364b, 437
Takayasu's disease, 503
Talin, 189
Tapetum, 56b, 56f
Tarsal plates, 82
TATA box, 134–135, 439
Taurine, 243–244
Taxol, 174
Tear(s), 198
  antimicrobial compounds, 463
  aqueous component of, 199–200
  artificial, 354–355
  complement, 463
  in drug administration, 348
  IgA, 199–200, 380–381, 453–454, 463
  IgG, 463
  lipid, 453
  mucins, 199–200, 201f, 348
    deficiency, 354
  ocular infection barriers, 453, 463
  pH, 348
  protein concentration of, 200f
  secretion
    by lacrimal gland, 200, 202.e1, 202.e1f
    psychoneuroendocrine control of, 201, 201f, 202.e1
Tear film, 89, 198–201, 198f, 202.e1
  break-up time, 198
  characteristics of, 196b, 198–199
  composition of, 199t
  evaporation, 198
  lipids, and lung surfactant, 198–199, 199f
  precorneal, 348
  reconstitution of, 354–355
  viscosity, 200
Tear meniscus height (TMH), 198
Tear-specific prealbumin, 199–200
Tear substitutes, 354–355
Tectum, 94
Tegmen tympani, 10t
Telodendria, 45
Temperature, extremes of, 486
Temporal arteritis, 503, 503f
Tenascin, 183, 184f, 234, 244–245
Tenon's capsule, 62
Tentorium cerebelli, 12, 12.e1
Teratogenic agents, facial defects, 129.e6
Teratoma, 522
Tertiary lymphoid organs (TLOs), 416, 417f
Texture analysis, 319–320, 320f
Texture perception, 324–325
Thalamus, 'place' cells, 310
Thiomersal, 354
Three receptors, molecular evidence for, 304–305
Thrombosis, in cavernous sinus, 13b
Thrombospondins, 184f, 185, 214–215
Thromboxanes, 361–362, 362b
Thymus
  bi-lobed, 413f
  regulates T-cell development, 413–414
Thyroid eye disease, 504
'Tic douloureux', 77b
Tight junctions, 172.e1, 172.e1f
Timolol, 220, 360b
Tissue damage
  agents and conditions causing, 486–489
  cellular mechanisms of, 410–412
Tissue glycation, 189–193
Tissue inhibitor of matrix metalloproteinases (TIMPs), 185, 254
  host response to injury, 383
Tissue plasminogen activator (tPA), in trabecular meshwork cells, 226–227
Tissue transplantation, 450–451
Tolerance induction
  mechanisms of, 442–443
  sites of, 441–442
Toll-like receptors, 389, 390t, 391f, 464, 466f
Toxocara (toxocariasis), 459, 480
Toxocara canis, 480, 499
Toxocara cati, 480
Toxoplasma, 372, 459
Toxoplasma gondii, 462, 478–479, 498, 499f
Toxoplasmosis, 478–480
  host response to, 479
  life cycle of, 479f
  ocular, 480
Trabeculae, 10
Trabecular meshwork, 24, 25f
  aqueous outflow, 226
  detoxifying enzyme systems in, 227
  metabolism of, 226–227
  prostaglandin synthesis by, 227
Trabecular meshwork inducible glucocorticoid response (TIGR) protein, 227
Trachoma, 202, 454, 482–483, 498b
  in Australia, 483
  clinical features of, 484f
Tracking, 310, 327
TRANCE, 403.e1
Trans-acting nuclear factors, 439
Transcellular pores, formation of, 226
Transcription, 133–134, 134f
  accessibility for, 135
Transcriptome, 145
Transcriptomics, 143–145, 159
Transcytosis, 226
Transducin, 250, 264f
Transduction, 147
  in drug-receptor interactions, 344
Transfection, 147
Transfer RNA (tRNA), 133–134
  mitochondrial inheritance, 138
Transforming growth factor-β (TGF-β), 159, 191, 227, 403.e2
Transforming growth factor-β-induced *(BIGH1)* gene, associated corneal dystrophies, 517b
Translation, 133–134, 134f
Translocation, 136
Transmembrane collagen XVII, 182
Transmembrane proteoglycans, 186, 188f
Trauma
  blunt, 537–538
  haematic cyst, 537–538
  mechanical, 486–487
Treacher Collins syndrome, 129.e5b
Treg activity, failure of, 444
Trichromatism, anomalous, 308
Trigeminal ganglion, 74
Trigeminal impression, 10t
Trigeminal nerve, 20, 74–77, 78f
  development of, 129.e4
  sensory nuclei of, 77
Trigeminal neuralgia, 77b

Triple helix
  collagens, 178–180, 179b, 179f, 181f
  oligonucleotides, 147
Triplet repeats, 137b
  myotonic dystrophy, 155–156
Trisomies, autosomal, 136.e1, 136.e1f
Trisomy 13, 136.e1, 129.e6b
Trisomy 18, 136.e1
Trisomy 21, 136.e1, 129.e6b, 136.e1f
Trochlear fossa, 2
Trochlear nerve, 71–74
  lesions of, 71b
  origin of, 71, 75f
Trophoblast, 103.e2
Troxler phenomenon, 286–287
Tryptophan, 238–239
Tuber cinereum, 9, 92–93
Tuberculoma, 495f
Tuberculum sellae, 9
Tumour(s)
  benign, 519
  choristomas, 522
  connective tissue, 534
  cytogenetics, 529b
  epithelial
    benign, 523
    malignant, 523–525
  hamartomas, 521–522
  histological classification of, 520t
  lacrimal gland, 536–537
  lymphoid, 534–536, 535f
  metastatic, 537
  muscle-derived, 533
  neural, 526–533
  spread and metastases, 521
  teratoma, 522
  vascular, 533–534
Tumour necrosis factor (TNF), 214, 401–402, 403.e1
  superfamily, 403.e4b–403.e5b, 403.e4f–403.e5f
Tumour necrosis factor-α (TNF-α), 147b, 365–366, 367f, 387–388
  acute phase response, 394
  viral infections, 468b
Tumour suppressor genes, 521
  p53, 147b
  retinoblastoma gene, 154–155
TUNEL staining, 147
Tunica vasculosa lentis, 121–123, 123f
  persistent, 124b
Tunics, 13, 14f
'Turkish saddle', 10t
Turner syndrome, 136.e1
'Two-hit hypothesis', 154–155
Tyrosinase, albinism, 155

U

Ubiquitin, 164.e1, 147.e1f, 164.e2f, 147.e1b
  conjugation in lens, 236
Ubiquitin-proteasome system, 227
Ubiquitinylation, 164
Ultraviolet light, cataract formation, 238–239
Unfolded protein response (UPR), 164
Uveal effusion syndrome, 215–216
Uveal melanoma, 525–526, 527f
Uveal meshwork, 24, 25f
Uveal tract, 14f, 26
  biochemical processes, 216–222
  development of, 124–127
  inflammation of, 502
  systemic drug toxicity, 368
Uveitis
  acute anterior, 459
  intraocular, 459.e1
  lens-induced, 501–502, 502f
  posterior, 459
Uveoretinitis, experimental autoimmune, 459

V

Variable number of tandem repeats (VNTR), 143
Varicella zoster virus, in acute retinal necrosis, 496
Vasa hyaloidea propria, 121
Vascular cell adhesion molecule (VCAM), 387
Vascular endothelial growth factor (VEGF), 353b
Vascular endothelial growth factor A (VEGF A), 353b, 392
Vascular endothelial growth factor receptor (VEGFR), 353b
Vascular endothelium
  inflammatory response, 386
  leucocyte adhesion, 386
Vasculitis, 503
Vaso-occlusive disease, 508
Vasodilator-stimulated phosphoprotein (VASP), 189
Vectors
  non-viral, 148b
  viral, 148b
Vernier acuity, 280, 282
Versican, 215
Vertical gaze centre, 331, 332f
Vestibular apparatus, 329
Vestibulo-ocular reflex, 329, 335
Viewing, binocular, 287
Vimentin, 174
Viral infections, 471–473
  herpes stromal keratitis, 471–472
    angiogenesis and lymphangiogenesis, 471
    latency of, 471–472
    primary infection, 471
  retinitis, 472–473
Viral vectors, 148b
Virulence
  microbial, 462
  of *Pseudomonas aeruginosa*, 473–474
Viruses, 463t, 495–497
  inflammation and, 488, 495–497
Viscerocranium, 1
  development of, 129.e1
Vision, 269–271
  binocular, 309–313
  depth perception, 273, 312, 314f
  form discrimination, 273
  limits, 271
  measuring by eye, 311
  minimal stimulus, 274
  monocular, 309–313
  physiology of, 269–337
  shape discrimination, 273
Visual acuity, 279–280, 279.e3f
  assessment, 280b, 281f
  best corrected, 285–287
  binocular summation, 279
  contrast sensitivity, 283–284
  limits of and limitations, 280–283, 283b, 283f
  luminance, 280
  pupil size, 286
  retinal ganglion cells, 285
Visual association areas, secondary, 95, 97f
Visual cortex
  hemifield maps, 323f
  line orientation, 312
  object selective regions, 326f
  organization, 323–326
  primary, 95, 97f
    cytoarchitecture of, 97f
  smooth pursuit eye movements tracking, 333–334
  ventral, 270f
  visual space mapping, 324, 324f–325f
Visual cortical cells, 269–270
Visual electrophysiology, clinical, 296–301, 297f
Visual evoked potential (VEP), 299–301, 301f
Visual field, 270–271
  driving licence requirement, 316f
Visual illusions, 309b, 309f
Visual pathway
  anatomy of, 92–102, 93f
  blood supply of, 98f, 102
  disturbances on, 97
  retinotopic organization of, 97
Visual perception, 308–320
  probability theory, 287
Visual process, 269–270
Visual response, chemistry of, 258–268
Visual stimulus/stimuli
  binocular summation, 279
  differential, 282
  minimal, 274
  size, 277–278

Visual system physiology, 269–337
  division of labour, 321–326, 321f
  electrophysiology, 287–302
  imaging studies, 321, 322f
  ocular movement, 326–336
Visual threshold, 273–274
Vitamin A, 258–259
  conjunctival epithelium, 202
Vitamin B_6, 193, 197
Vitamin E
  as free radical, 264b, 264f
  free radical scavenger activity, 198
  and light damage, 239, 239f
  and photoreceptor damage, 252–253
  regeneration of, 264
Vitelliform dystrophy, 515, 516f
Vitreal collagen, production of, 31–32
Vitreal hyaluronic acid, production of, 31–32
Vitreoretinopathy, familial exudative, 154
Vitreous, 37–38, 37f
  base, 123, 125f
  biochemical processes, 240–242
    and glucose transporters (GLUTs), 243
    glycosaminoglycans, 186
    hyalocytes, 241
    and lens clarity, 242, 243f
    matrix of, 240–241
    physiochemical properties of, 241–242
  collagens, 182, 240–241, 241f
    type IX, 240–241
  development of, 121–124
  gel physiochemical properties of
    bulk flow, retardation of, 241–242
    deformation of globe, protection from, 241
    diffusion of small molecules, retardation of, 241–242
    retina, protection of, 241
  hyalocytes in, 38, 39f
  lymphoma, 535–536
  malformations of, 124b
  persistent hyperplastic primary, 124b
  posterior detachment of, 38b
  primary, 121
  pyogenic infection, 494
  secondary, 123
  tertiary, 123
VLA-4, 435
Vogt-Koyanagi-Harada syndrome, 139

W

Waldenström's macroglobulinaemia, 507
Water
  and rhodopsin conformation, 172f, 259
  transport
    and *AQPO* gene, 233
    in lens, 235
Wavelength
  contrast sensitivity, 284–285
  discrimination, 292, 322–323
WD repeat domain *(WDR36)* gene, 227
Wegener's granulomatosis, 503
Wilson's disease, 489b
Wnt signaling, 251–252, 178.e1, 178.e1f
Wolfram *(WFS1)* gene, 238
Wuchereria spp., 499

X

X chromosome, 150–151
Xanthine, 193
Xanthine oxidase, 221
Xanthophyll macular pigments, 253, 254f
Xenograft rejection, 451
X-linked disorders, 137–138
  mapping of, 150–151
  ocular disease, 152t
  retinitis pigmentosa, 151

Y

Yeast, 463t, 475
Yolk sac, 103.e2
Young-Helmholtz trichromatic theory, 306

Z

ZAP-70, 436–437
Zeaxanthin, 253, 254f
Zidovudine (AZT), 352.e1b
Zonular apparatus. *see* Lens zonules
Zonule(s)
  collagens, 182
  production of, 31–32
  relaxation of, 217–218, 218f
Zonules of Zinn, 35
Zygomatic foramen, 2
Zygomatic nerve, 75
Zygotene, 131
Zygotic genes, 110

Johannes W. Rohen
Chihiro Yokochi
Elke Lütjen-Drecoll

Color Atlas of Anatomy

FOURTH EDITION

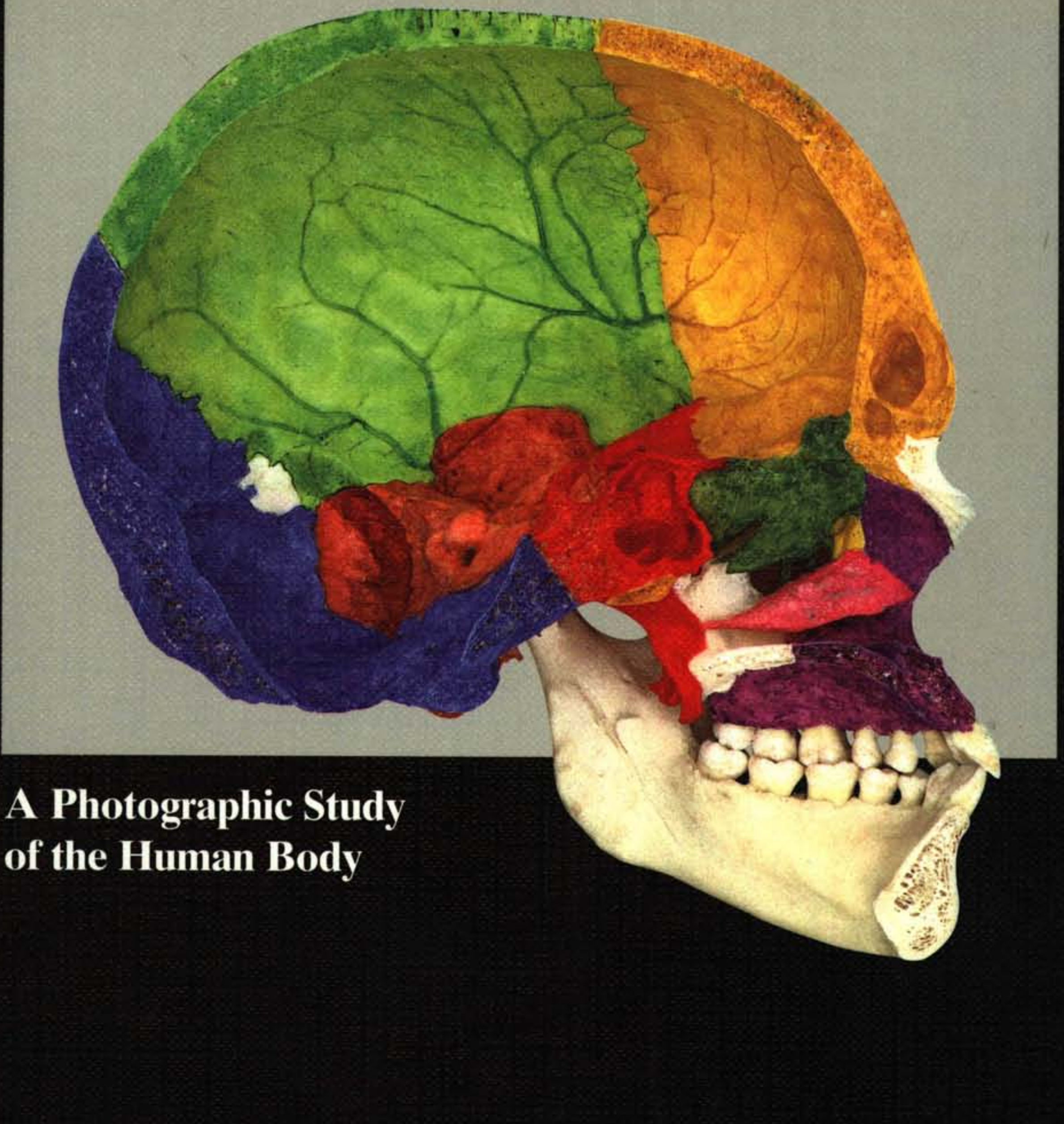

A Photographic Study
of the Human Body

In contrast to most other mammals the human body is adapted for bipedal locomotion. Three general principles in the architecture of the human organism are recognizable:

1. The principle of **segmentation,** which dominates in the trunk. The vertebral column and the thorax consist of relatively equal, segmentally arranged elements.
2. The principle of **bilateral symmetry.** Both sides of the body are separated by a midsagittal plane and resemble each other like image and mirror-image.
3. The principle of **polarity** between the head at one end of the body and the lower extremities at the other. As the center of the information system the head contains the main sensory organs and the brain. The head has a predominantly spherical form while the extremities consist of radially formed skeletal elements, the number of which increases distally.

A. The **skull** consists of two parts: 1. a **cranial part** containing mainly the brain and the sensory organs and 2. a **facial part** which contains the nasal and oral cavity and the chewing apparatus. The cranial cavity is continuous with the vertebral canal which contains the spinal cord.

B. The **thorax** contains the respiratory and circulatory organs (lung, heart, etc.) but also some of the abdominal organs which are located underneath the diaphragm.

C. The **abdominal cavity** contains the organs of metabolism such as the liver, the stomach and the intestinal tract as well as the excretory and genital organs (kidney, uterus, urinary bladder, etc.). The latter are located primarily in the **pelvic cavity** with the exception of the testes.

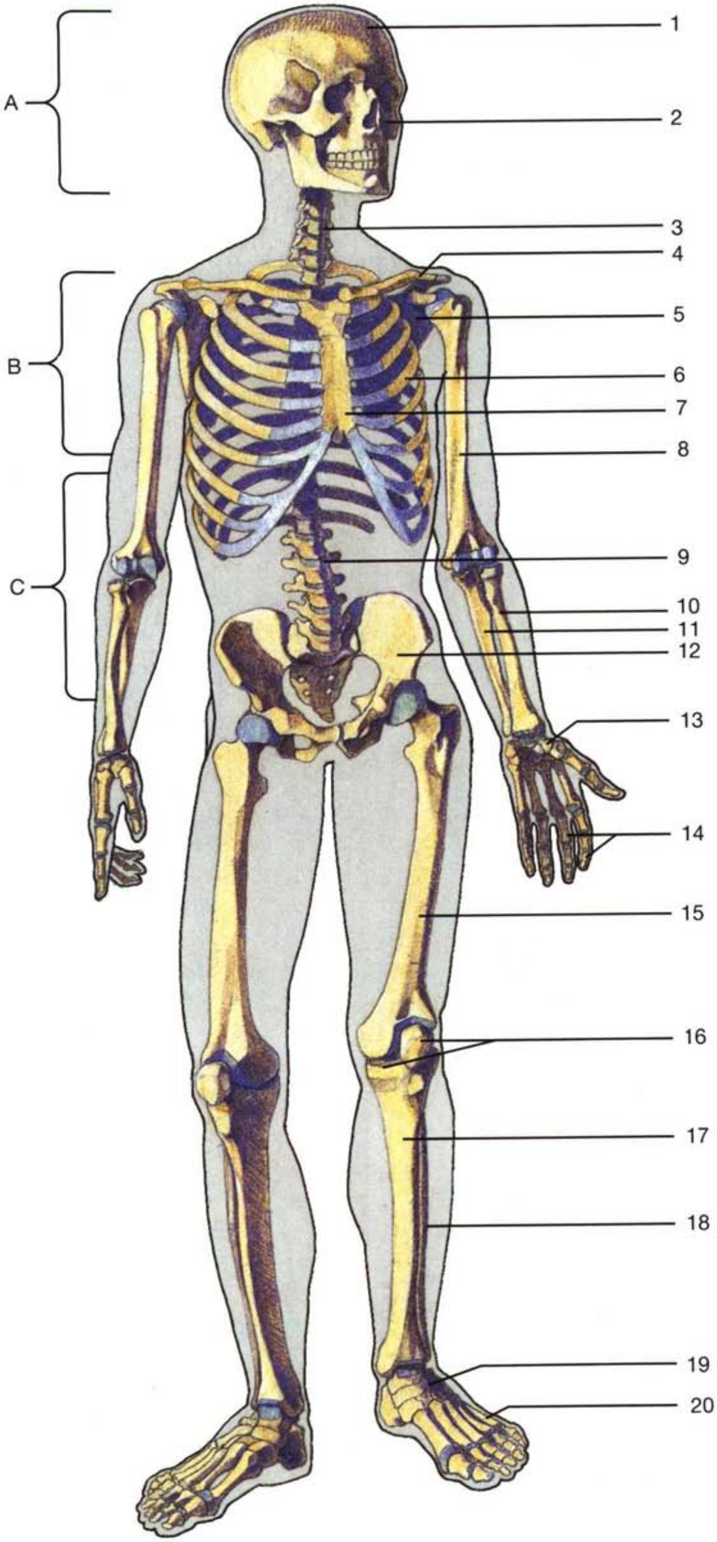

Structure of the human body and the skeleton. Blue = joints.
A Head (caput) B Thorax (thoracic cavity) C Abdominal and pelvic cavities

1	Cranial part	of the skull
2	Facial part	
3	Vertebral column (cervical part)	
4	Clavicle	
5	Scapula	
6	Ribs	
7	Sternum	
8	Arm (humerus)	
9	Vertebral column (lumbar part)	
10	Radius	forearm
11	Ulna	
12	Pelvis	
13	Wrist (carpals)	hand
14	Fingers (phalanges)	
15	Thigh (femur)	
16	Patella and knee joint	
17	Tibia	leg
18	Fibula	
19	Tarsals	foot
20	Metatarsals	

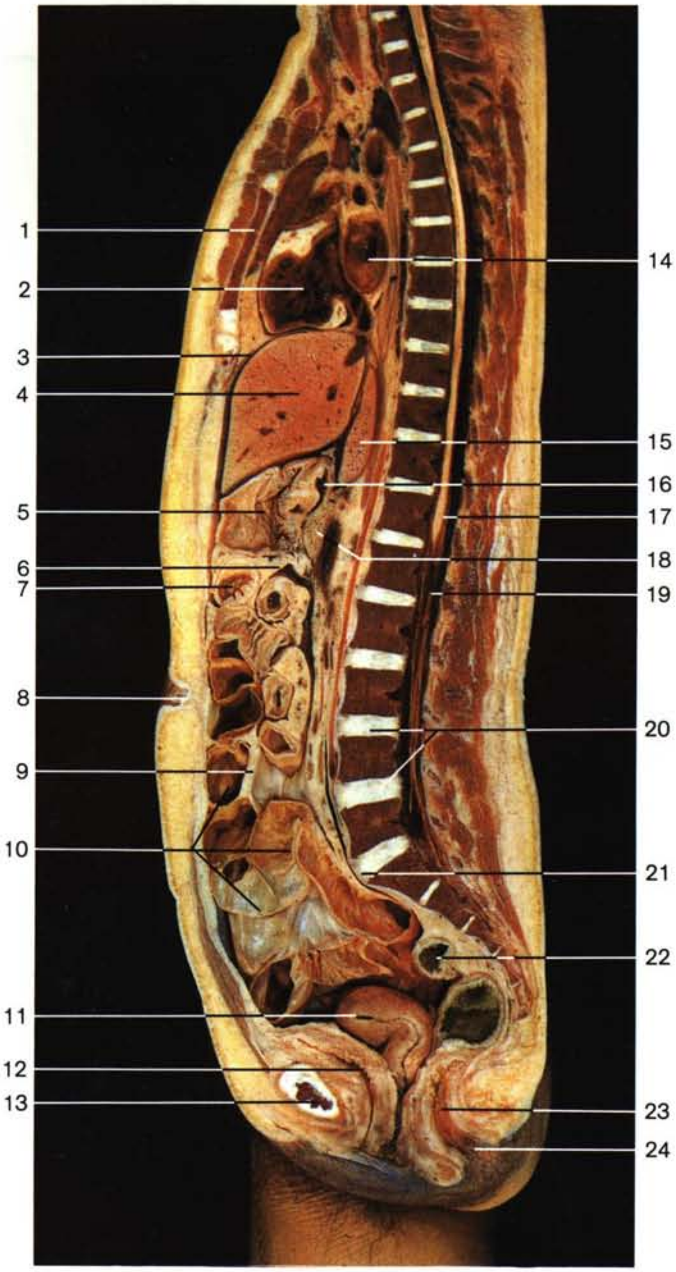
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Median section through the trunk (female).

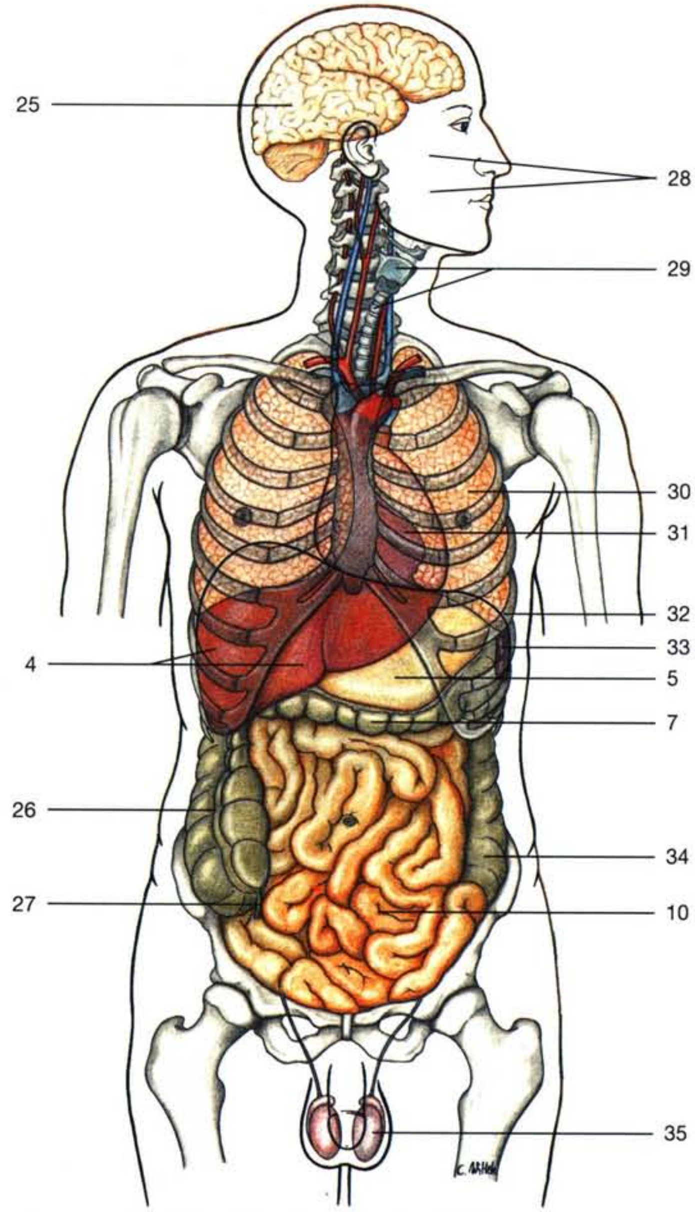
25
28
29
30
31
32
33
4
5
7
26
34
27
10
35

Positon of the inner organs of the human body (anterior aspect)
The main cavities of the body and their contents.

1. Sternum
2. Right ventricle of heart
3. Diaphragma
4. **Liver**
5. **Stomach**
6. Transverse mesocolon
7. Transverse colon
8. Umbilicus
9. Mesentery
10. **Small intestine**
11. Uterus
12. Urinary bladder
13. Pubic symphysis
14. Left atrium of heart
15. Caudate lobe of liver
16. Omental bursa or lesser sac
17. Conus medullaris
18. **Pancreas**
19. Cauda equina
20. Intervertebral discs (lumbar vertebral column)
21. Sacral promontory
22. Sigmoid colon
23. Anal canal
24. Anus
25. Head (neurocranium) with brain
26. Ascending colon
27. Appendix
28. Facial region (viscerocranium) with oral and nasal cavities
29. Trachea and larynx
30. Thorax with the lungs
31. **Heart**
32. Surface projection of the diaphragm
33. **Spleen**
34. Descending colon
35. Testis

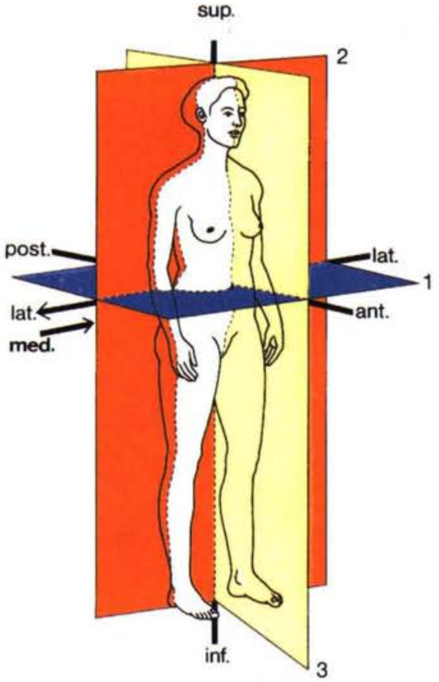
sup.
2
post.
lat.
1
lat.
ant.
med.
inf.
3

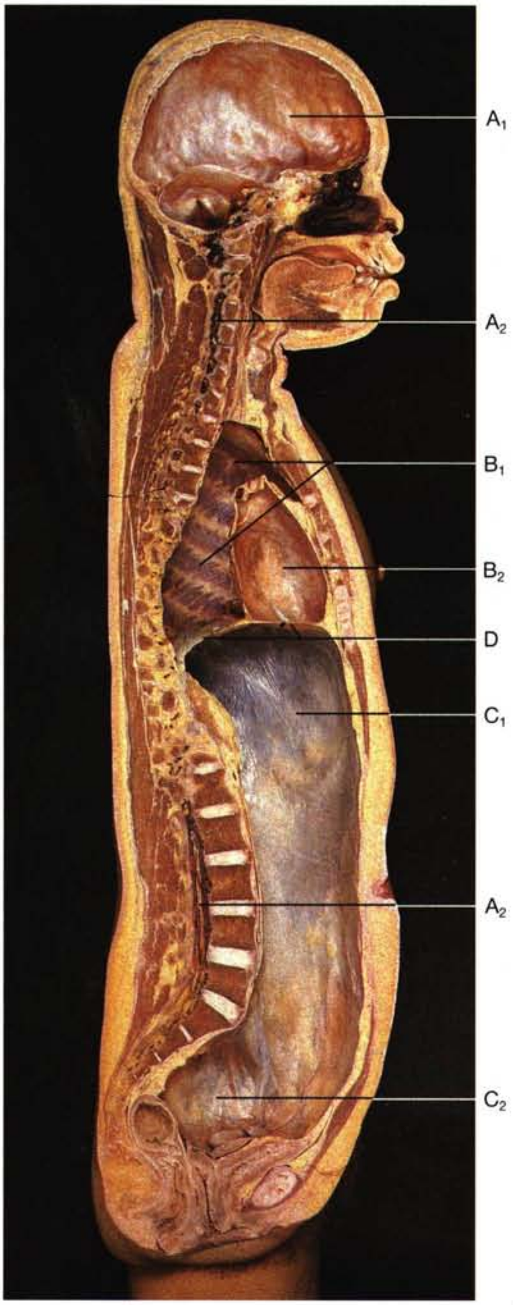

Sagittal section through the human body (female).
Demonstration of the main cavities of the body.
Internal organs are removed.

- A_1 Cranial cavity
- A_2 Vertebral canal
- B_1 Thoracic cavity
- B_2 Pericardial cavity
- C_1 Abdominal cavity
- C_2 Pelvic cavity
- D Diaphragm

◁

Planes of the body

1 Transverse plane
2 Frontal plane
3 Sagittal plane (midsagittal)

Lines of direction

ant. = anterior
inf. = inferior
lat. = lateral
med. = medial
post. = posterior
sup. = superior

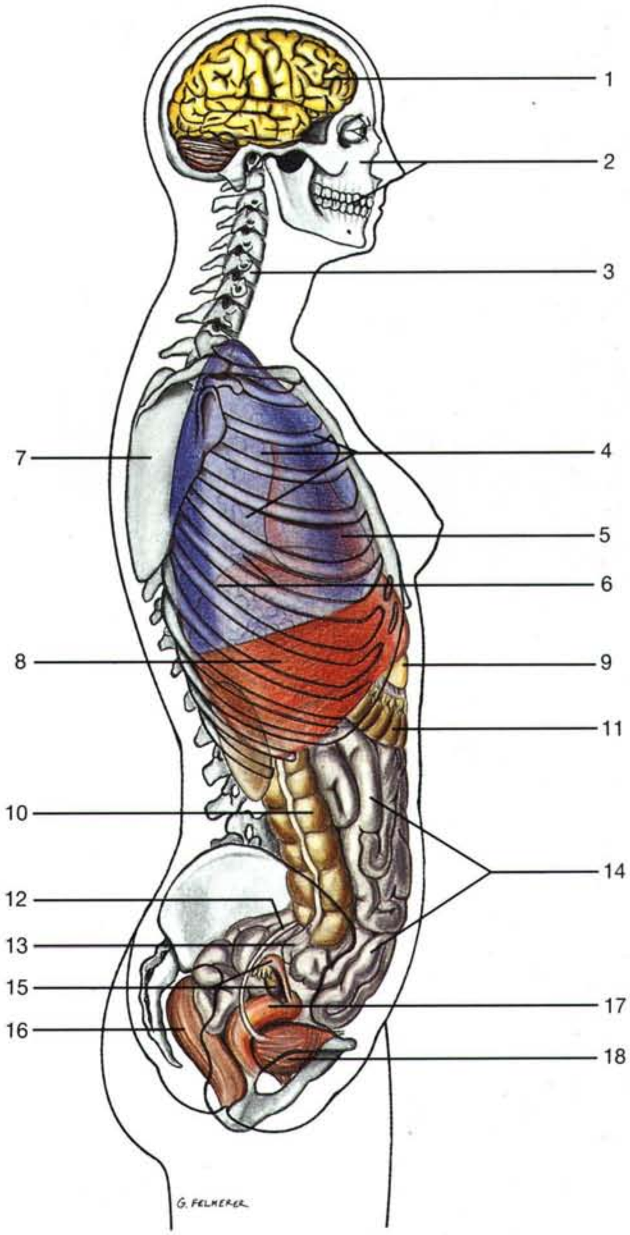

Position of the inner organs of the human body (lateral aspect).
The three main cavities of the body and their contents.

1 Head (neurocranium) with the brain
2 Facial bones with oral and nasal cavities
3 Vertebral column (cervical part)
4 Thorax with the lungs
5 Heart
6 Surface projection of the diaphragm
7 Scapula
8 Liver
9 Stomach
10 Ascending colon
11 Transverse colon
12 Ureter
13 Appendix
14 Small intestine
15 Ovary, uterine tube
16 Rectum
17 Uterus
18 Urinary bladder

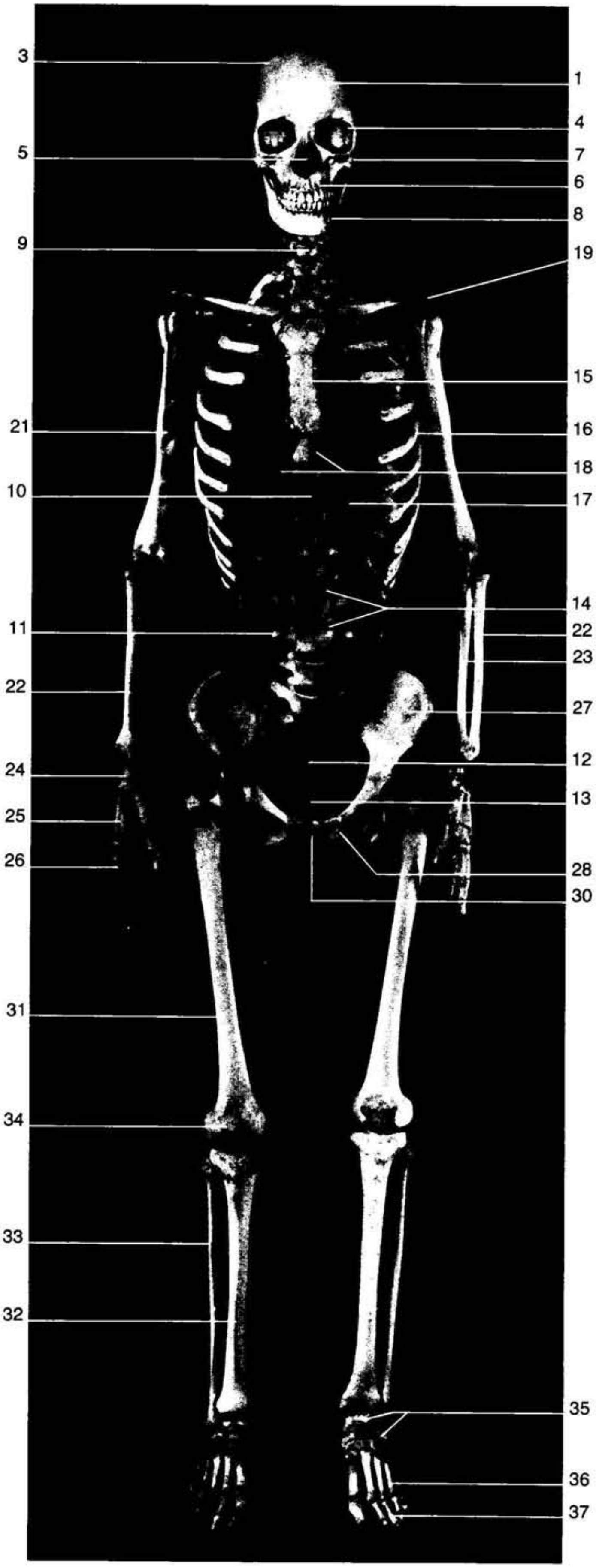
3
1
4
5
7
6
8
9
19
15
21
16
18
10
17
14
11
22
23
22
27
24
12
25
13
26
28
30
31
34
33
32
35
36
37

Skeleton of a female adult (anterior aspect).

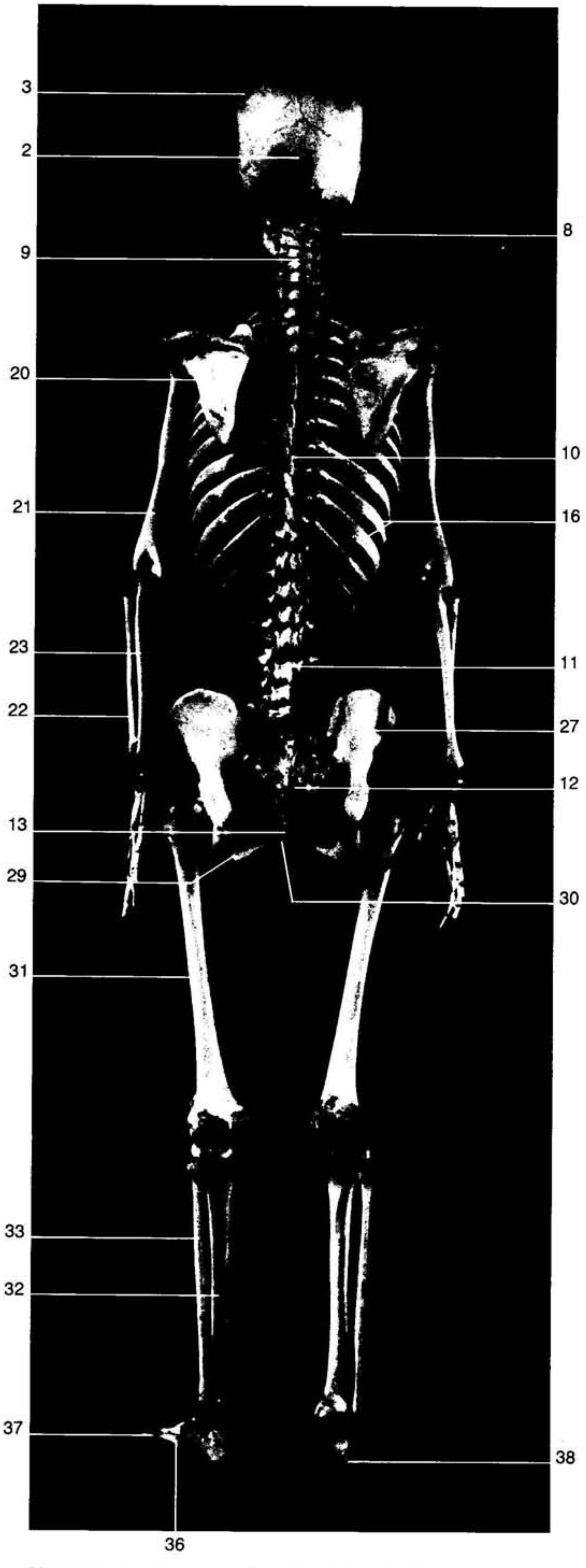
3
2
8
9
20
10
21
16
23
11
22
27
12
13
29
30
31
33
32
37
38
36

Skeleton of a female adult (posterior aspect).

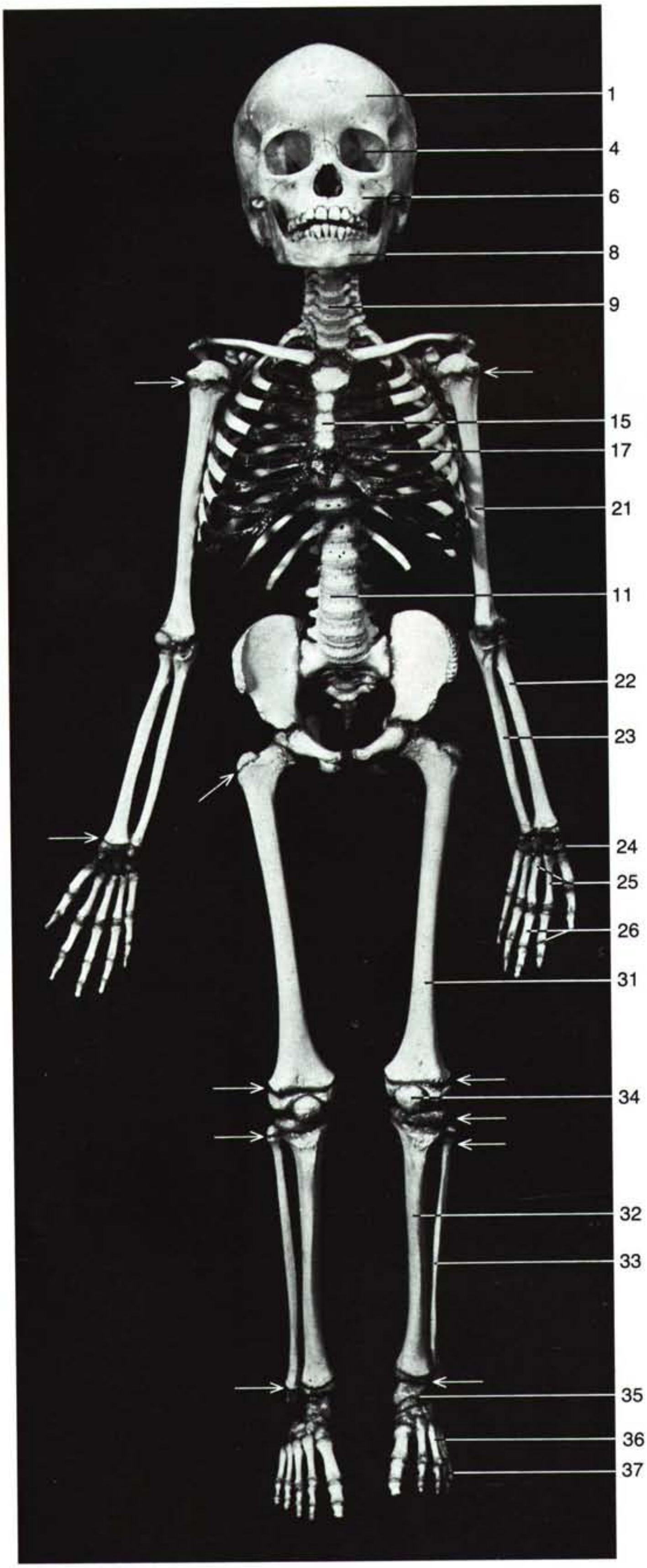

Skeleton of a 5-year-old child (anterior aspect). The zones of the cartilaginous growth plates are seen (arrows). In contrast to the adult, the ribs show a predominantly horizontal position.

Axial skeleton

Head

1 Frontal bone
2 Occipital bone
3 Parietal bone
4 Orbit
5 Nasal cavity
6 Maxilla
7 Zygomatic bone
8 Mandible

Trunk and thorax

Vertebral column

9 Cervical vertebrae
10 Thoracic vertebrae
11 Lumbar vertebrae
12 Sacrum
13 Coccyx
14 Intervertebral discs

Thorax

15 Sternum
16 Ribs
17 Costal cartilage
18 Infrasternal angle

Appendicular skeleton

Upper limb and shoulder girdle

19 Clavicle
20 Scapula
21 Humerus
22 Radius
23 Ulna
24 Carpal bones
25 Metacarpal bones
26 Phalanges of the hand

Lower limb and pelvis

27 Ilium
28 Pubis
29 Ischium
30 Symphysis pubis
31 Femur
32 Tibia
33 Fibula
34 Patella
35 Tarsal bones
36 Metatarsal bones
37 Phalanges of the foot
38 Calcaneus

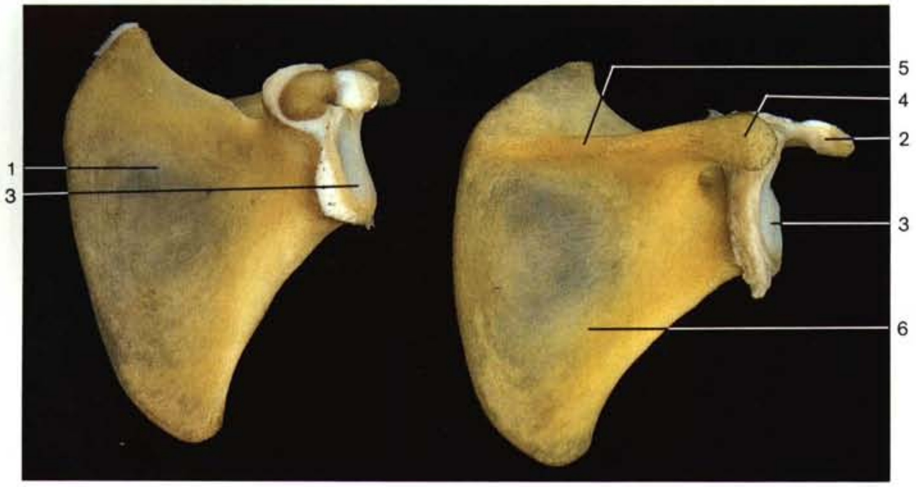

1. Subscapular fossa
2. Coracoid process
3. Glenoid fossa
4. Acromion
5. Spine of scapula
6. Infraspinous fossa

Ossification of the scapula
(left: anterior aspect, right: posterior aspect).

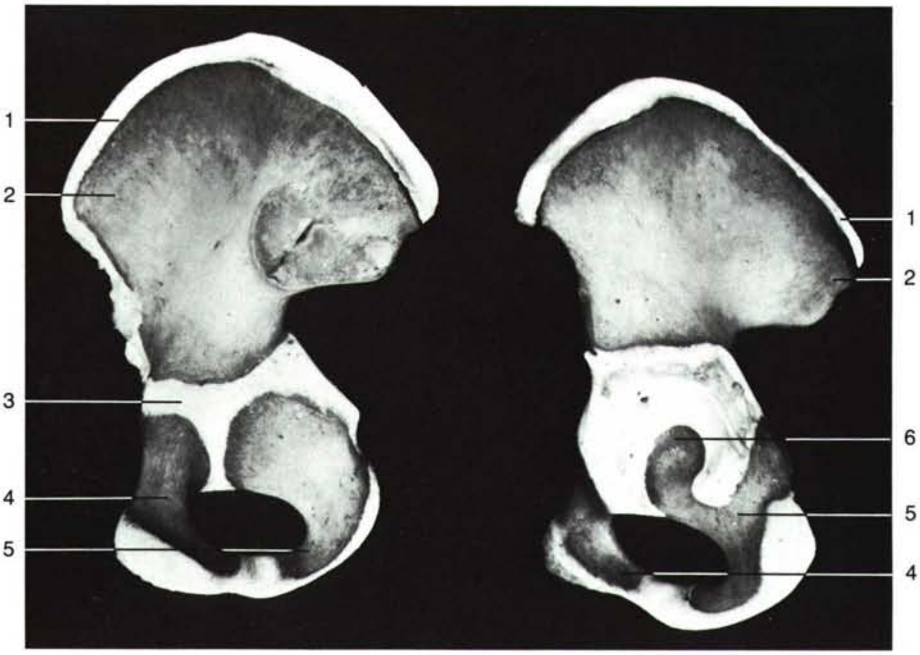

1. Cartilage of the iliac crest
2. Ilium
3. Cartilage
4. Pubis
5. Ischium
6. Acetabulum

Ossification of the hip bone
(left: medial aspect, right: lateral aspect).

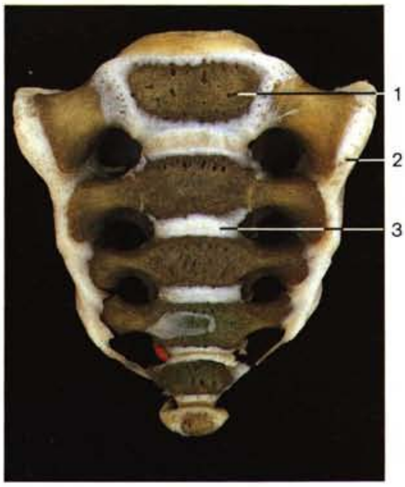

1. Bone tissue (vertebral body)
2. Cartilaginous tissue (lateral epiphysis)
3. Intervertebral discs

Ossification of the sacrum (anterior aspect).
Note the five vertebral bones, which are still separated from each other.

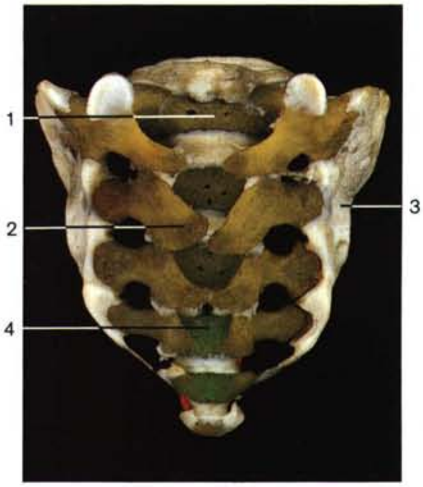

1. Bony tissue (center of ossification)
2. Vertebral arch (not completely united)
3. Cartilaginous tissue (lateral epiphysis)
4. Sacral canal

Ossification of the sacrum
(posterior aspect).

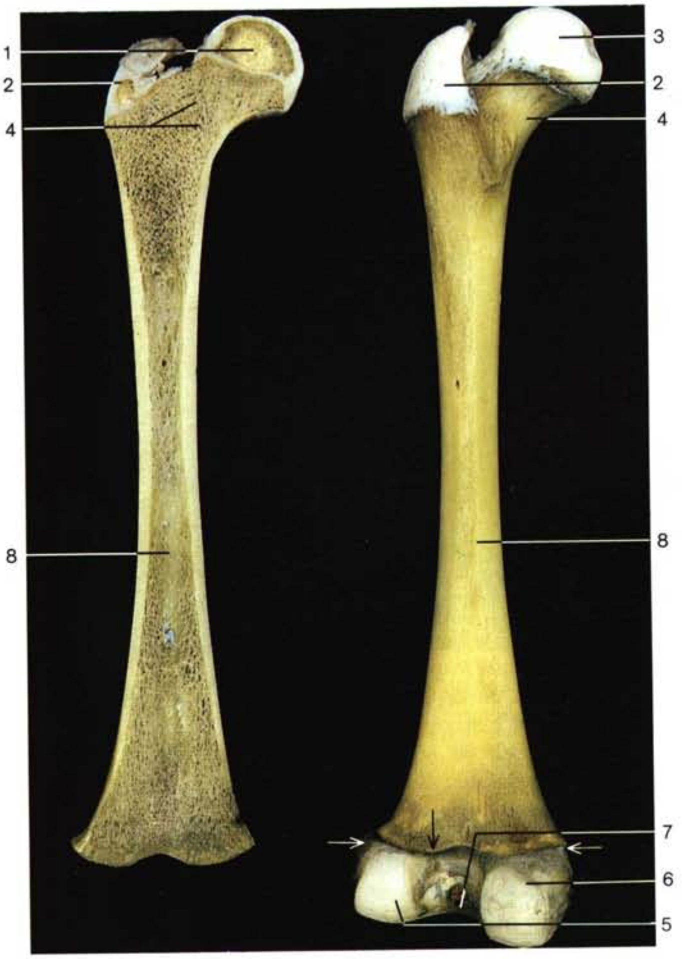

1 Ossification center in the head of the femur
2 Greater trochanter
3 Head of the femur
4 Neck of the femur
5 Lateral condyle
6 Medial condyle
7 Intercondylar notch
8 Diaphysis

Ossification of the femur (left: coronal section, right: posterior view of the femur). Arrows: distal epiphysis.

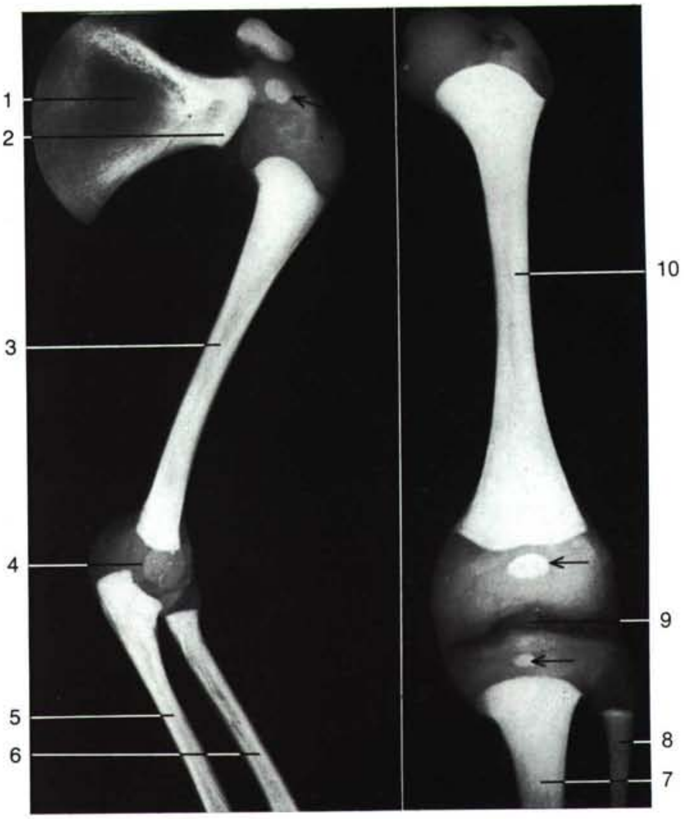

X-ray of the upper and lower limb of a newborn child.
Left: upper limb. Right: lower limb.
Arrows: ossification centers.

1 Scapula
2 Shoulder joint
3 Humerus
4 Elbow joint
5 Ulna
6 Radius
7 Tibia
8 Fibula
9 Knee joint
10 Femur

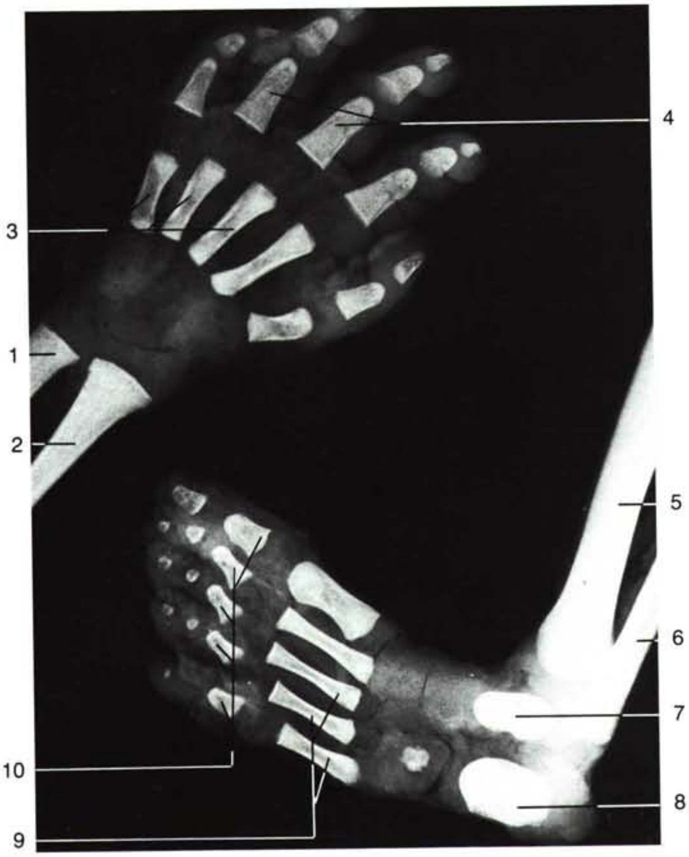

X-ray of hand and foot of a newborn.

1 Ulna
2 Radius
3 Metacarpals
4 Phalanges
5 Tibia
6 Fibula
7 Talus
8 Calcaneus
9 Metatarsals
10 Phalanges

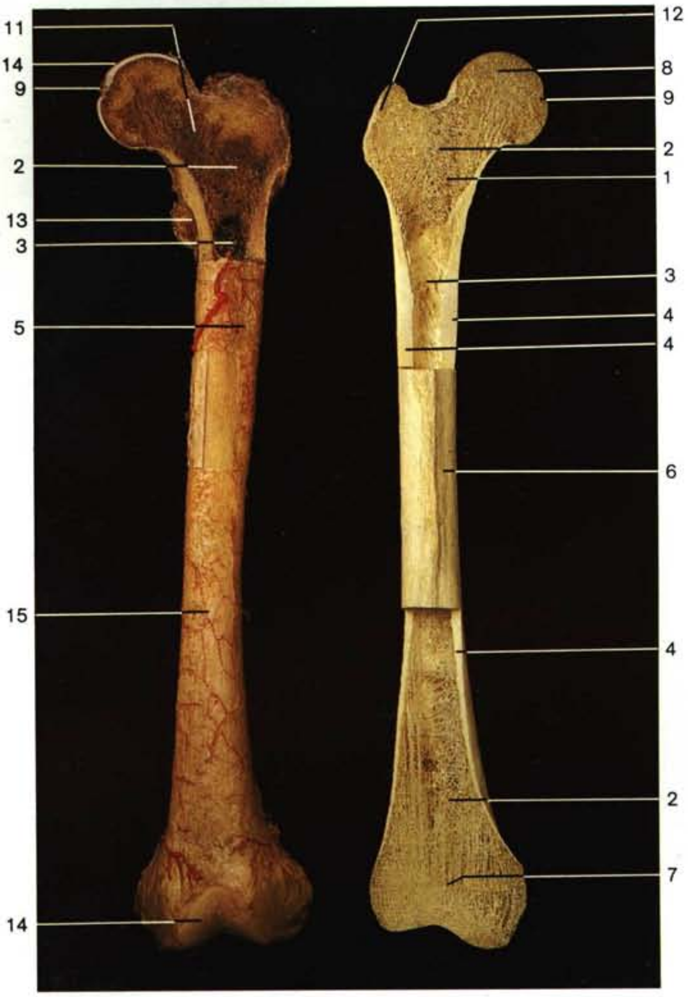

Femur of the adult. Left: the periosteum and the nutrient vessels are preserved. Right: coronal section of the proximal and distal epiphyses to display the spongy bone and the medullary cavity.

1 Metaphysis
2 Spongy bone
3 Medullary cavity in the diaphysis
4 Compact bone
5 Nutrient canal
6 Diaphysis
7 Epiphyseal line (remnants of the epiphyseal plate)
8 Epiphysis (head of the femur)
9 Fovea of head
10 Trabeculae of spongy bone
11 Neck of the femur
12 Greater trochanter
13 Lesser trochanter
14 Articular surface
15 Periosteum
16 Skin
17 Vastus medialis muscle
18 Sartorius muscle
19 Femoral artery and vein
20 Great saphenous vein
21 Gracilis muscle
22 Adductor longus muscle
23 Adductor magnus muscle
24 Semimembranosus muscle
25 Semitendinosus muscle
26 Rectus femoris muscle
27 Vastus lateralis muscle
28 Femur and medullary cavity
29 Vastus intermedius muscle
30 Sciatic nerve
31 Biceps femoris muscle
32 Spongy bone trabeculae containing bone marrow
33 Compact bone
34 Osteon with Haversian lamellae
35 Periosteum
36 Blood vessels and nerves for periosteum and bone

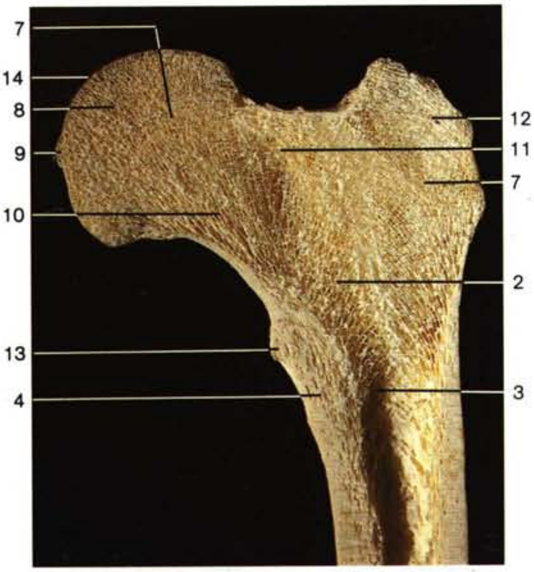

Coronal section through the proximal end of the adult femur, revealing the characteristic trajectorial structure of the spongy bone.

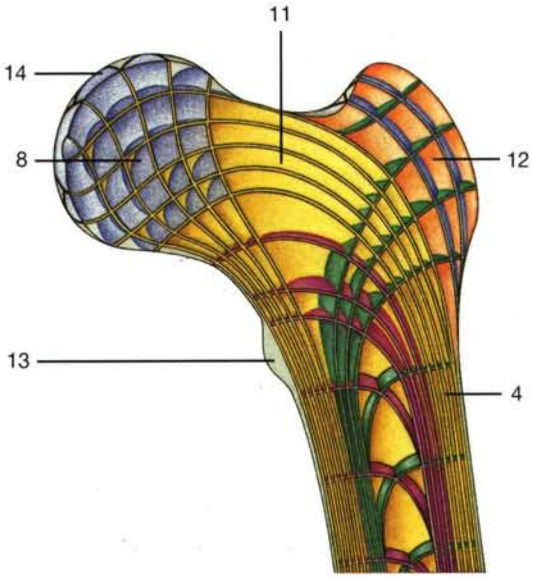

Three dimensional representation on the trajectorial lines of the femoral head (according to B. Kummer).

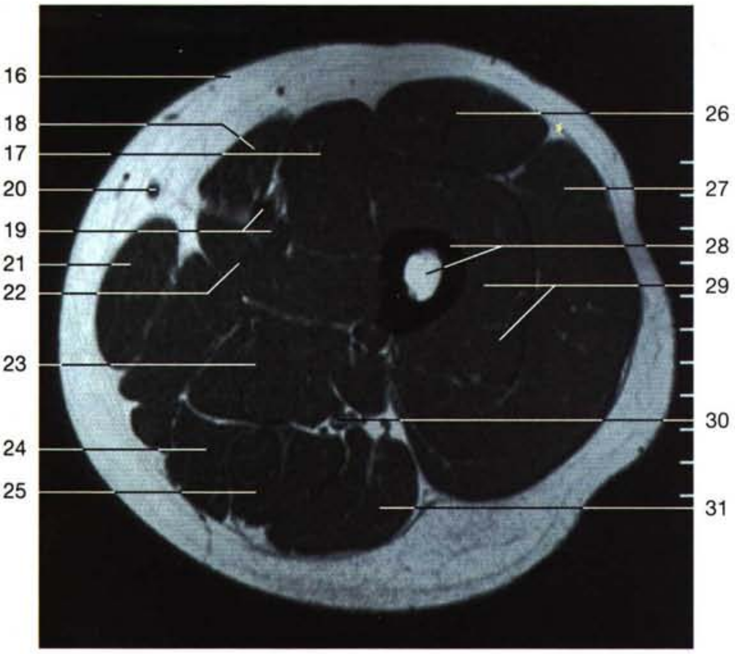

MR-Image of the thigh (axial section through the middle of the left thigh, the same level as the CT-Image).

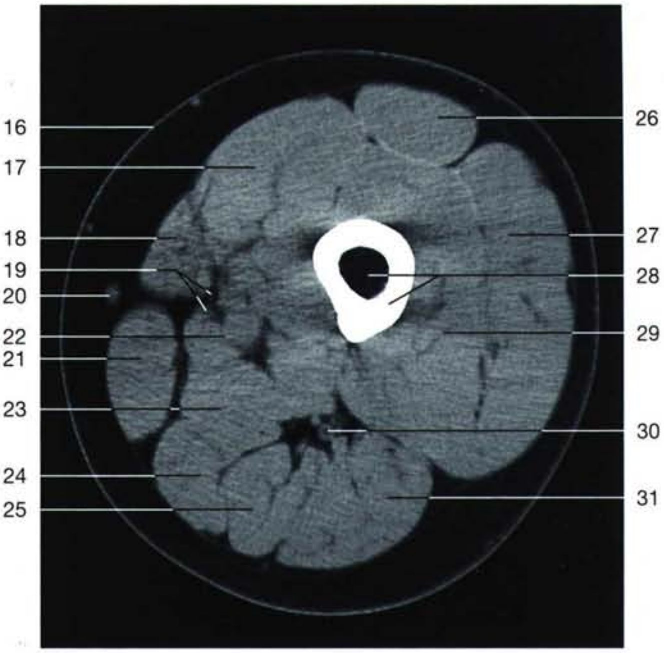

CT-Image of the section through the middle of the left thigh (axial section). Note the differences between the CT- and MR-Image (see text below).

The **bones** of the skeletal system consist of two different parts, the spongy and the compact bone. The spongy bone trabeculae are highly adapted to mechanical forces revealing a trajectorial structure. The intertrabecular spaces are filled with bone marrow, the site of blood formation. The appearance of bones, muscles and soft tisssues is quite different in **CT-** and **MR-Images**. The CT-images relate well to radiographs in that areas of great absorption such as bones are white, and those with little absorption such as fat appear black. In contrast, the intensity of signals in MR-Images, obtained without X-rays but by magnetic forces, is different so that dense areas of bones appear black and soft tissues such as bone marrow and fat appear white (for comparison see above figures).

A highly innervated **periosteum** is an essential structure for bone nutrition, blood supply, growth and bone repair.

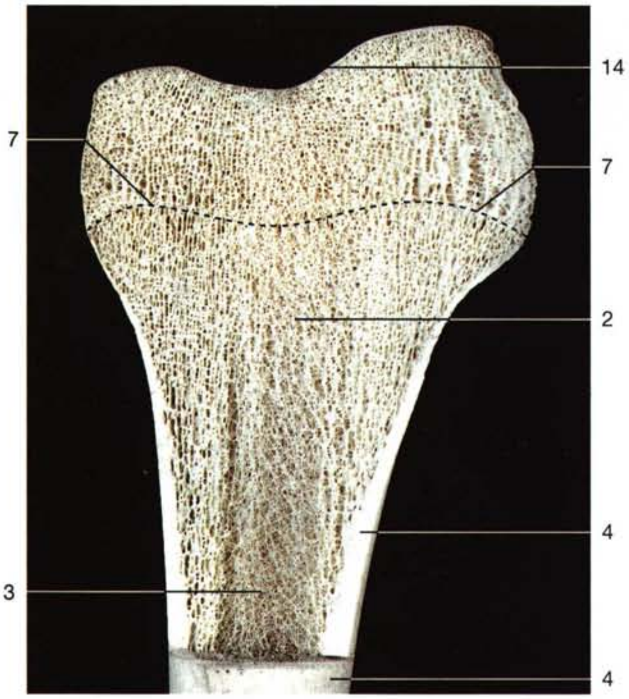

Coronal section through the proximal epiphysis of the adult tibia. Note the zone of dense bone at the site of the former epiphyseal plate (dotted line).

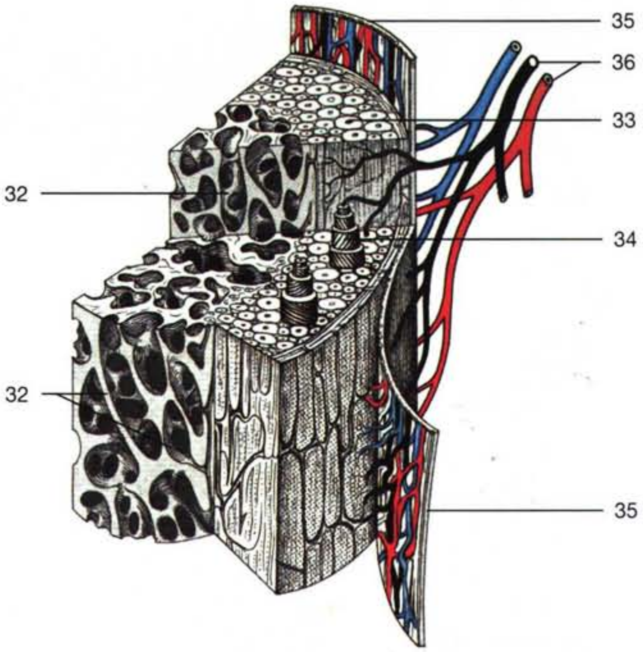

Structure of bones of the skeletal system (after Benninghoff). Note that the compact bone reveals a lamellar structure with Haversian lamellae and canals.

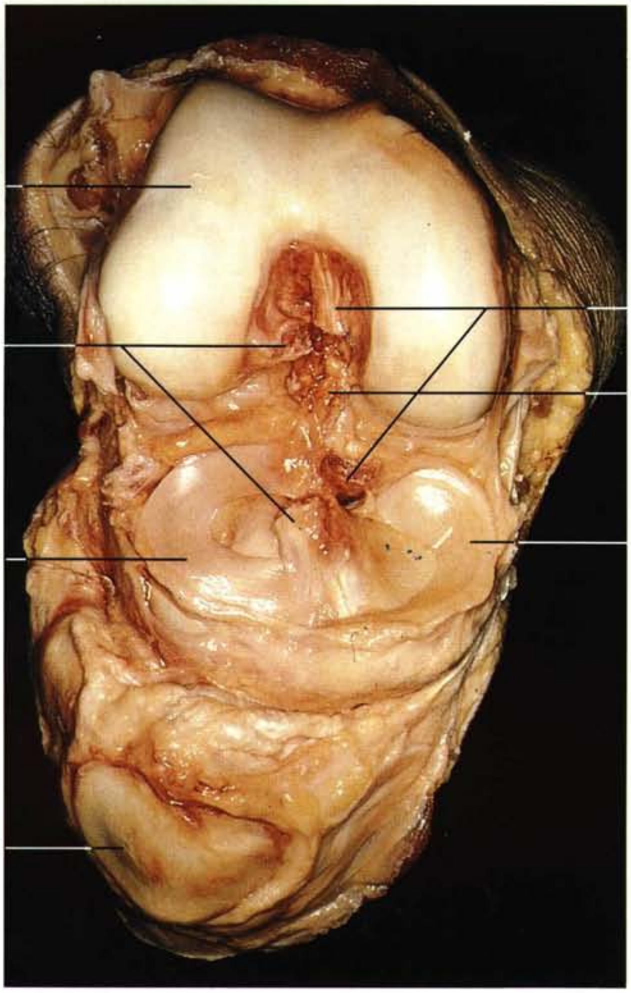

Knee joint. Anterior aspect, showing menisci and cruciate ligaments (cut). Quadriceps tendon cut and patella reflected distally.

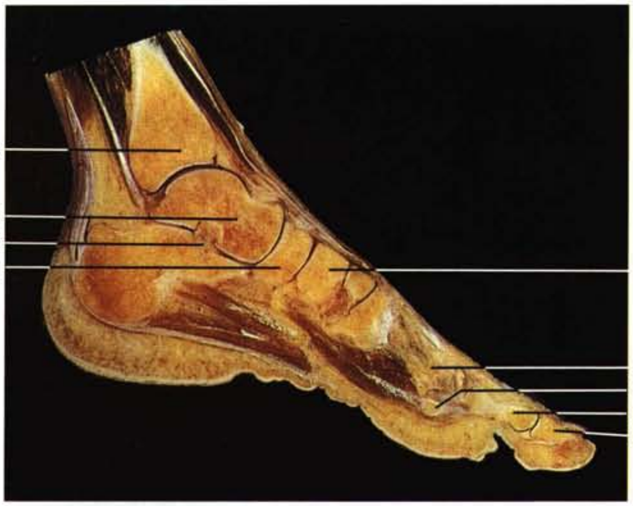

Sagittal section through the lower limb and the foot.

1 Femur
2 Anterior cruciate ligament
3 Lateral meniscus
4 Patella
5 Posterior cruciate ligament
6 Posterior meniscofemoral ligament
7 Medial meniscus
8 Fibular collateral ligament
9 Fibula
10 Tibia
11 Tibial collateral ligament
12 Articular cartilage
13 Articular capsule
14 Tendon of long head of biceps brachii muscle
15 Head of humerus
16 Glenoid labrum
17 Articular cartilage of glenoid fossa

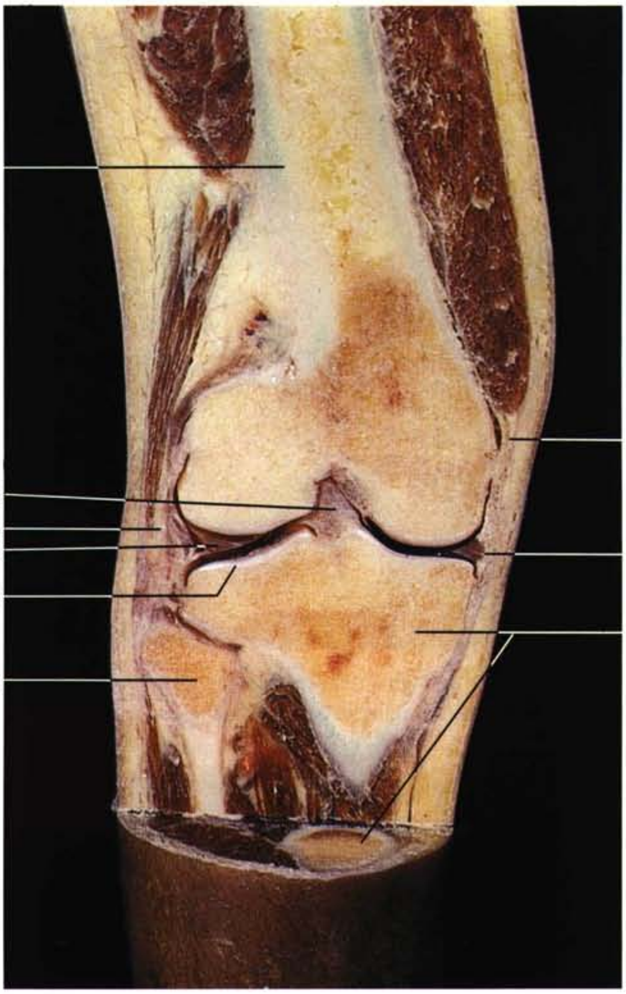

Coronal section through the knee joint. Anterior aspect of the right joint in extension.

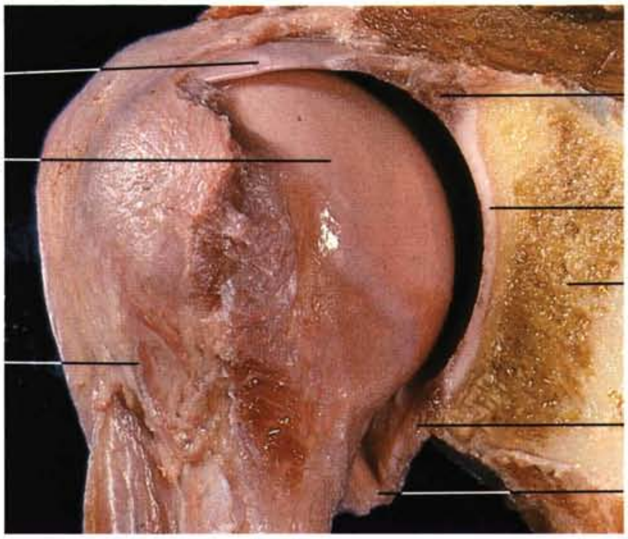

Shoulder joint (anterior view). The anterior part of the articular capsule has been removed.

18 Scapula
19 Talus
20 Interosseous talocalcaneal ligament
21 Navicular bone
22 Medial cuneiform bone
23 First metatarsal bone
24 Proximal phalanx of the hallux (great toe)
25 Distal phalanx of the hallux
26 Sesamoid bone

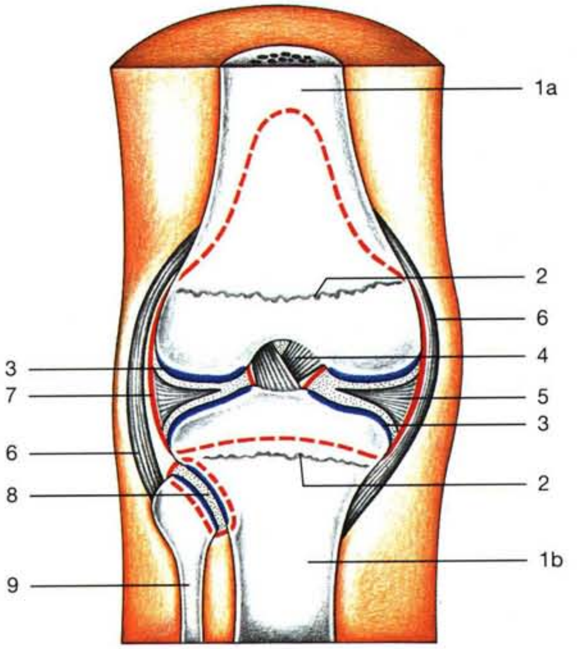

General architecture of a synovial joint with 2 articulating bones and a synovial cavity (right side, anterior view).
Coronal section throught the knee joint.
Red line = Articular capsule with synovial membrane.
Dotted red line = extension of articular capsule (suprapatellar bursa).

1 Articulating bones: a) Femur, b) Tibia
2 Epiphysial line
3 **Articular cartilage**
4 Intraarticular ligaments (e.g. cruciate ligaments)
5 Fibrocartilaginous disk (e.g. meniscus)
6 Collateral ligaments
7 Articular capsule with synovial membrane
8 **Tibiofibular articulation** (example of gliding synovial joint)
9 Fibula

	Movement	Examples
A Fibrous joints		
1 Sutures	No movements	Sutures of the skull
2 Syndesmoses	No movements	Distal tibiofibular joint
3 Gomphosis	No movements	Roots of teeth in alveolar process
B Cartilaginous joints		
1 Synchondroses	No movements	Epiphyseal plates
2 Symphyses	Slight movement	Symphysis pubis intervertebral discs
C Synovial joints		
1 Gliding	Monaxial	Intercarpal joint Intertarsal joint Sacroiliacal joint
2 Hinge	Monaxial	Interphalangeal joint Humeroulnar joint Talocrural joint
3 Pivot	Monaxial	Atlantoaxial joint Radioulnar joint
4 Ellipsoidal	Biaxial	Radiocarpal joint
5 Saddle	Biaxial	Carpometacarpal joint of the thumb
6 Ball-and-socket	Multiaxial	Shoulder and hip joint

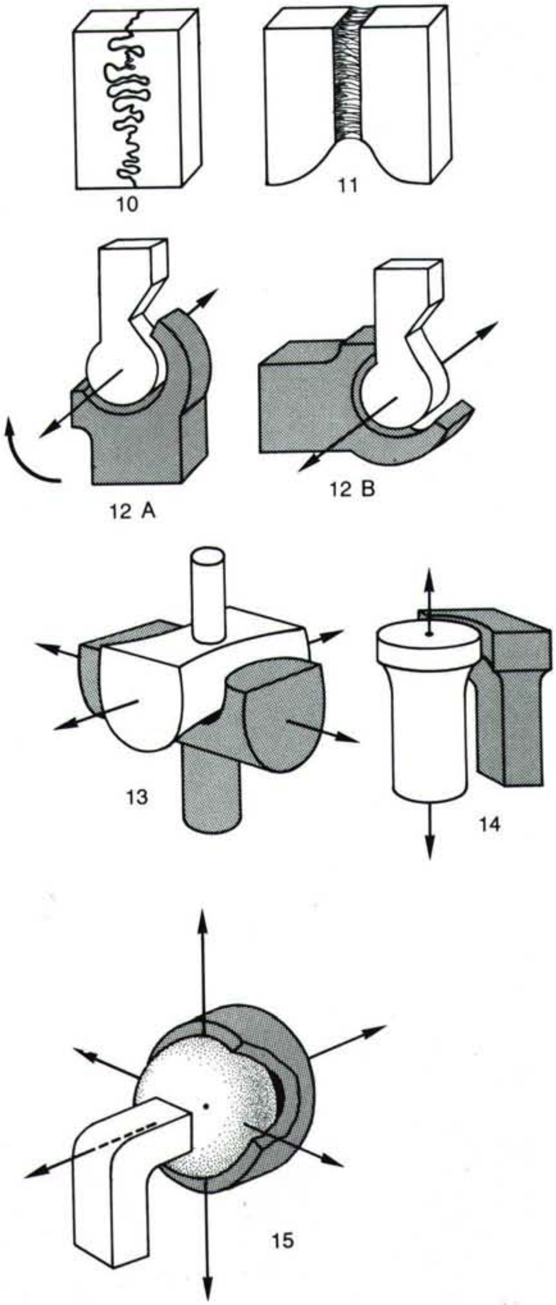

Main types of joints. Arrows: axes of movement.

Fibrous joints (synarthroses)
10 Serrate suture
11 Syndesmosis

Synovial joints (diarthroses)
12 Hinge joints (monaxial) ginglymus
  A Extension
  B Flexion
13 Saddle joint (biaxial)
14 Pivot joint (monaxial, rotation)
15 Ball-and-socket joint (multiaxial)

An articulation or joint is the functional connection between two or more bones. Joints can be divided into two categories depending upon whether the articulating surfaces of the bones are separated by a real cavity (joint cavity) so that they are movable against each other **(synovial joints)** or whether the bones are firmly connected by fibrous or cartilaginous tissue and practically immovable (**fibrous joints, cartilaginous joints,** symphysis, etc.). Synovial joints always possess a joint capsule (with a vascularized synovial membrane), articular cartilages, and a joint cavity. They are grouped according to the degree of movement they permit. A hinge joint (ginglymus) permits movement in only one plane about a single axis **(uniaxial** or **monaxial),** an ellipsoidal joint permits movements in two planes **(biaxial),** and ball-and-socket joints permit a range of movements around several axes **(multiaxial).** The following survey gives a few examples of these types of articulation.

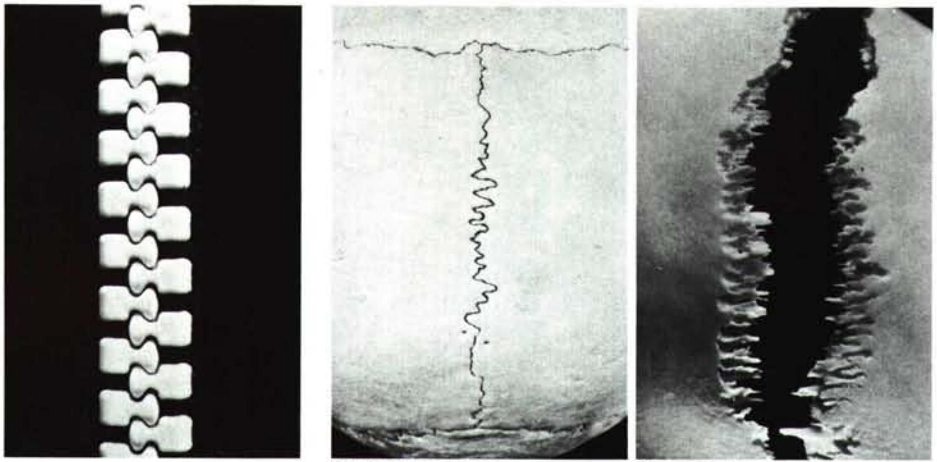
Sagittal suture

Separated

1. Fibrous joints

A Sutures

a) Serrate or dentate suture

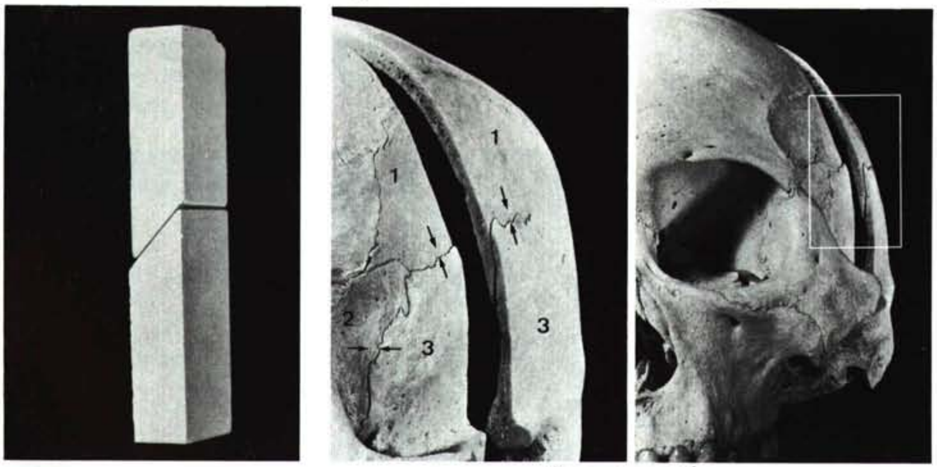

Squamous suture

b) Squamous suture (arrows)

1 Parietal bone
2 Sphenoid bone
3 Squama of temporal bone

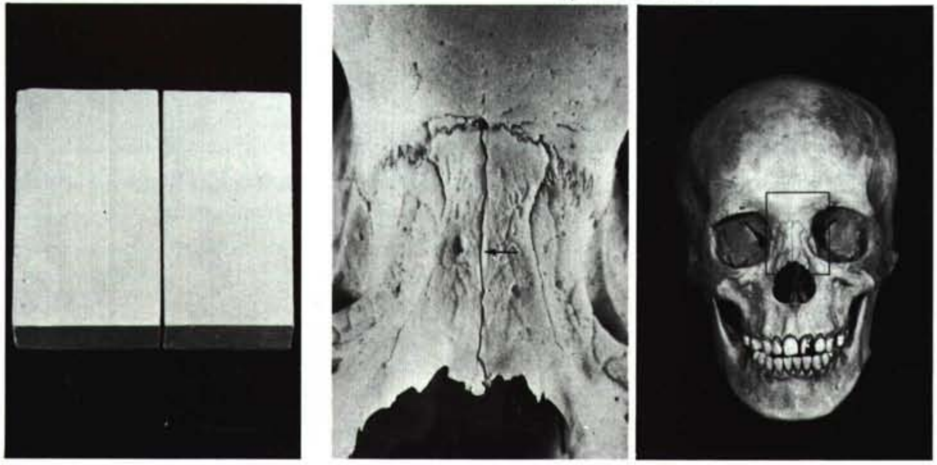
Nasal bones

c) Plane suture (arrow)

B Peg suture
(Gomphosis)

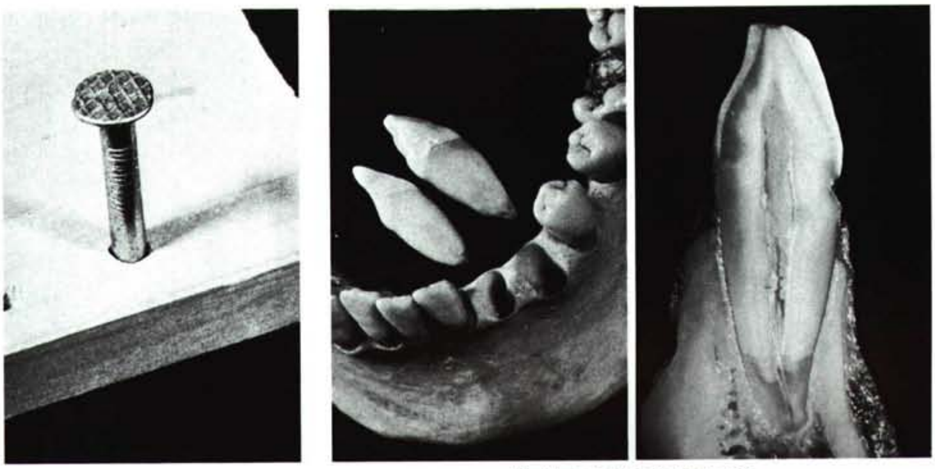

Teeth and alveolar process

2. Cartilaginous joints

a) Symphysis
(fibrocartilage)

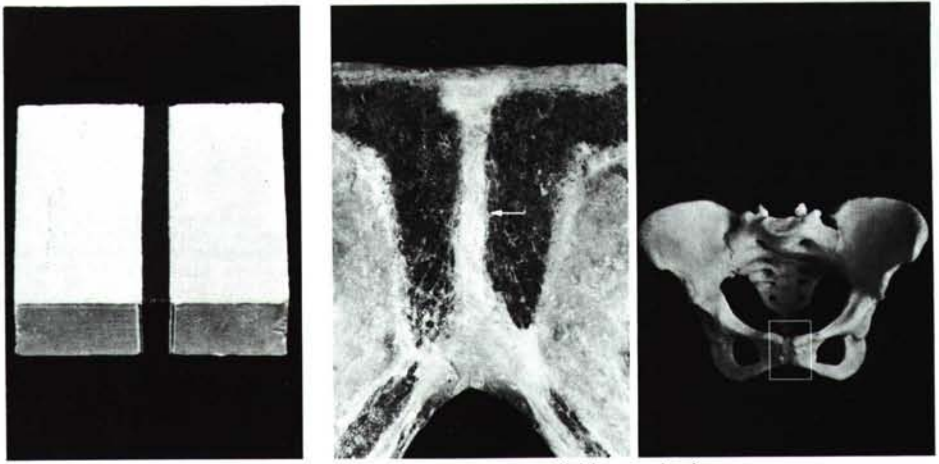

Pubic symphysis

b) Synchondrosis
(hyaline cartilage)

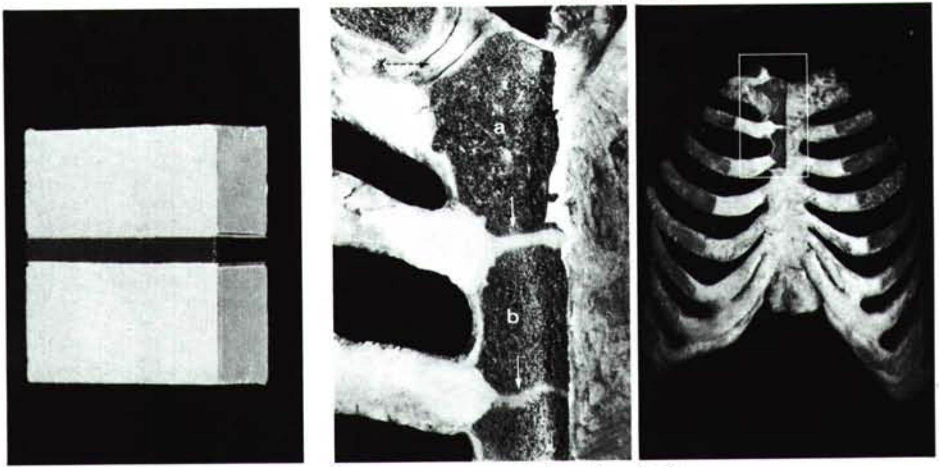

Sternum a = manubrium; b = body

* Articular disc
(sternoclavicular joint)

3. Osseous joints
(synostosis)

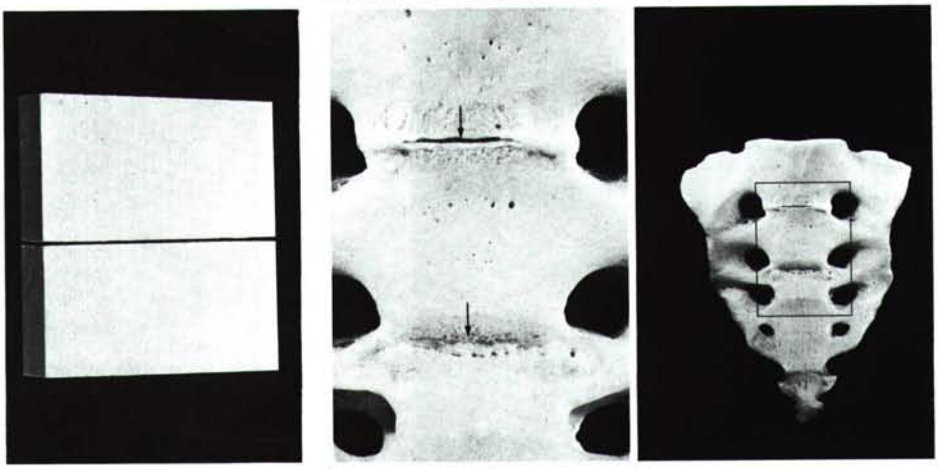

Transverse ridge (arrows)

Sacrum

1. Ball-and-socket joint

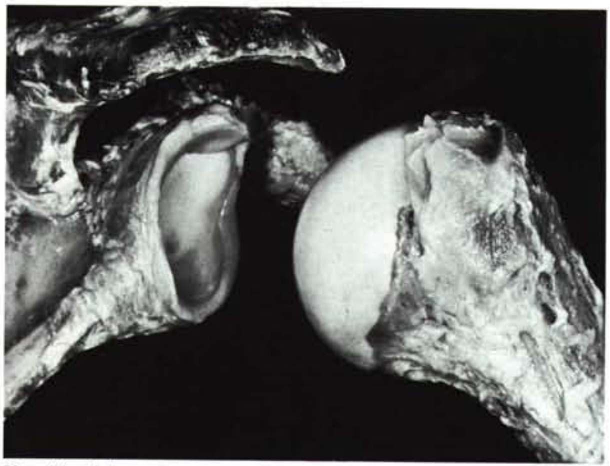

Shoulder joint

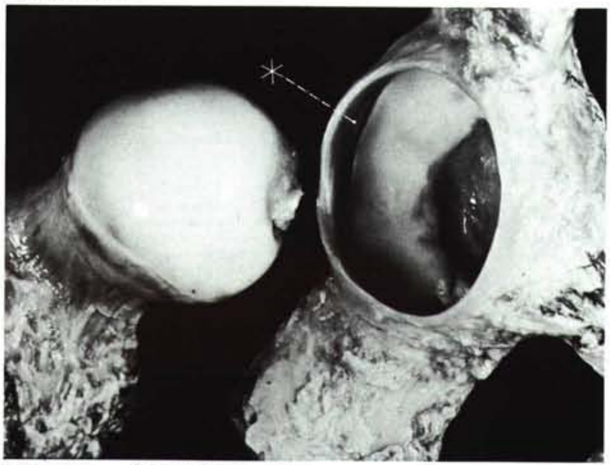

Hip joint *Acetabular labrum

2. Ellipsoid joint

OFFICIAL
Rugby Football
OFFICIAL
SIZE-WEIGHT

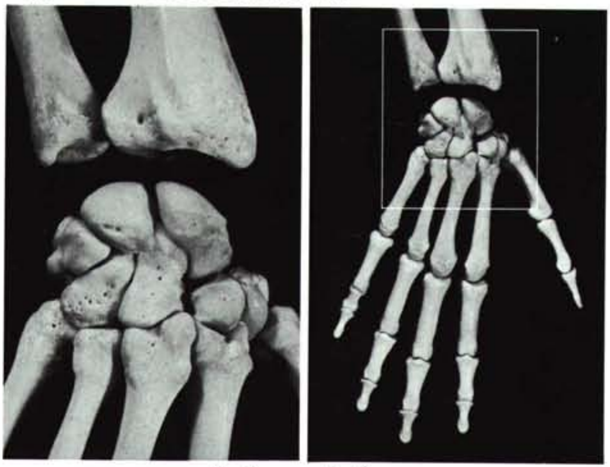

Radiocarpal joint

3. Hinge joint

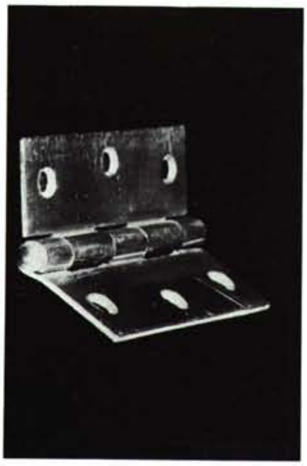

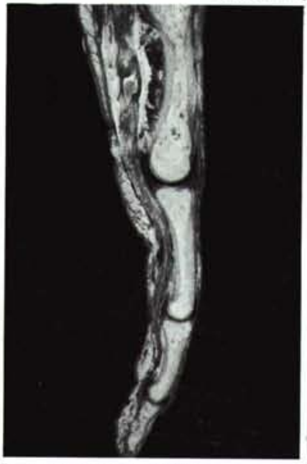

Interphalangeal joint

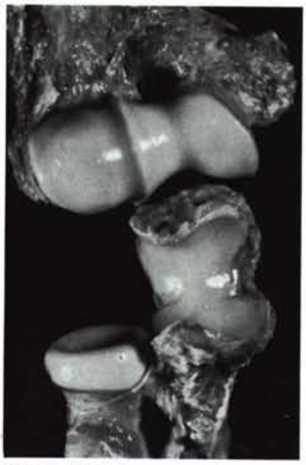

Elbow joint (humeroulnar joint)

4. Pivot joint

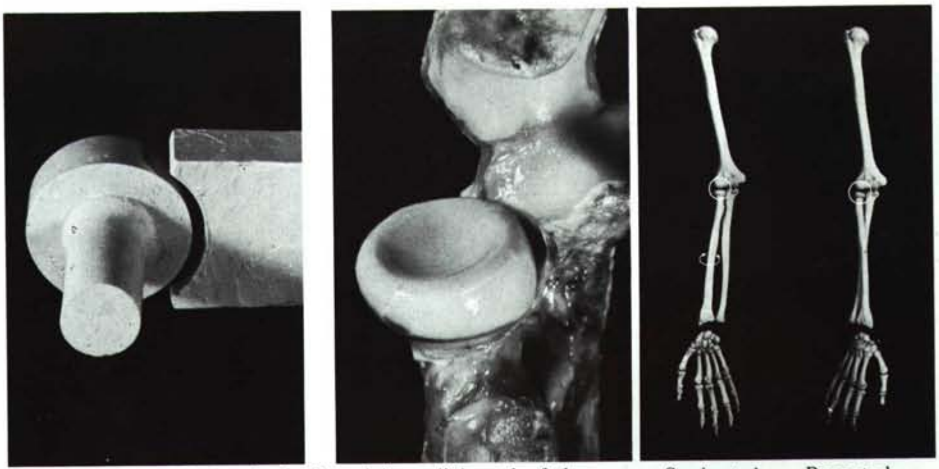
Radioulnar joint; left: head of radius, right: radial notch of ulna

Supinated Pronated

5. Hinge joint

(but a gradually
tilting axis produces
a slight spiral motion)

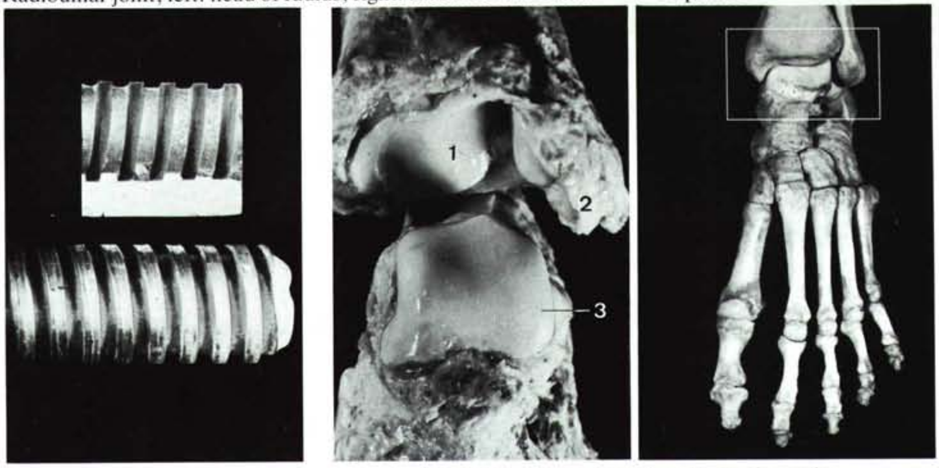

Ankle joint: 1 = tibia; 2 = fibula; 3 = talus

6. Saddle joint

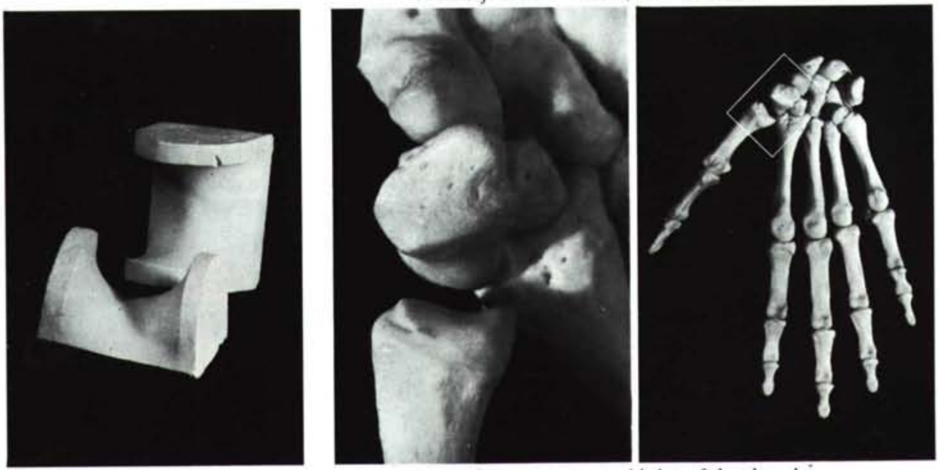
Carpometacarpal joint of the thumb

7. Plane joint

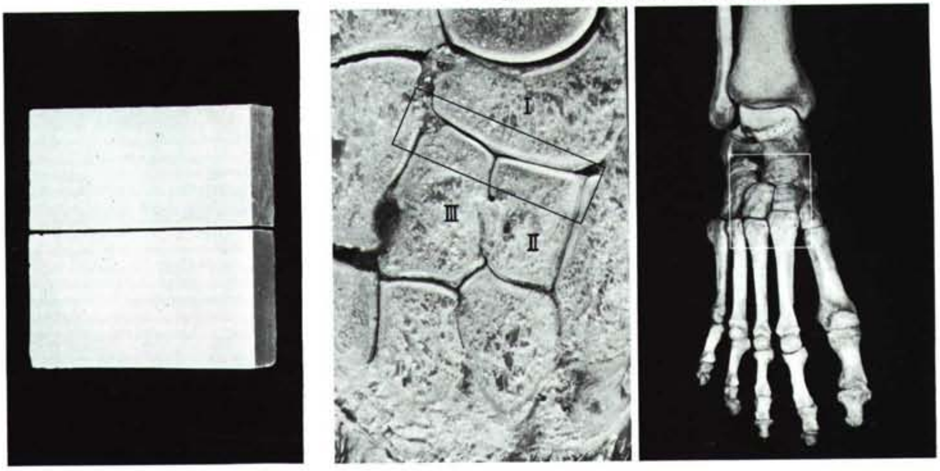

I = navicular; **II** = intermediate cuneiform; **III** = lateral cuneiform

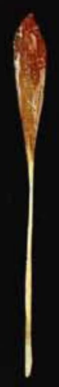

Fusiform
(palmaris longus muscle)

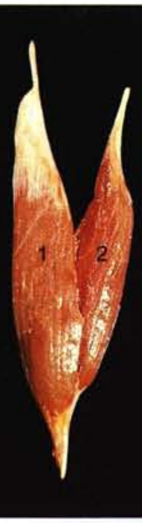

Bicipital
(biceps brachii muscle)

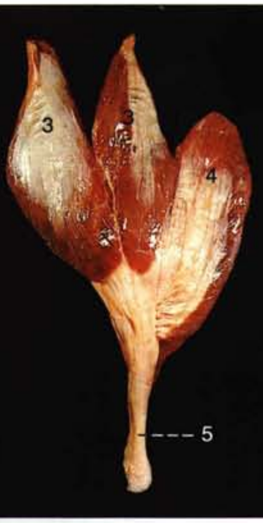

Tricipital
(triceps surae muscle)

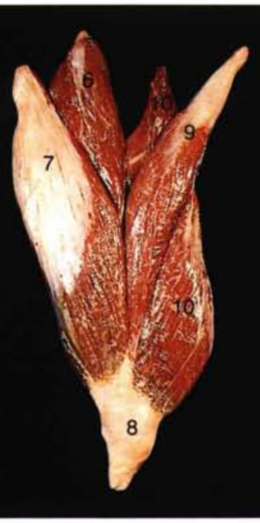

Quadricipital
(quadriceps femoris muscle)

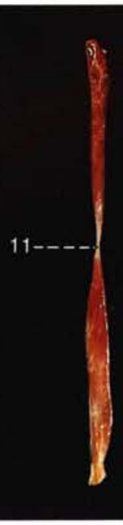

Digastric
(omohyoid muscle)

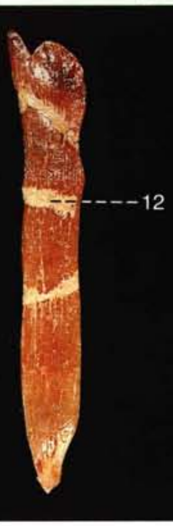

Multiventral
(rectus abdominis muscle)

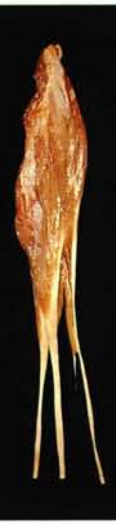

Multicaudal
(flexor prof. muscle)

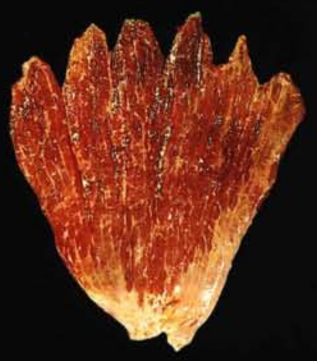

Serrated
(serratus anterior muscle)

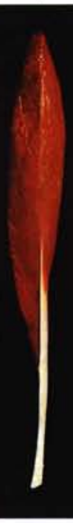

Bipennate
(tibialis anterior muscle)

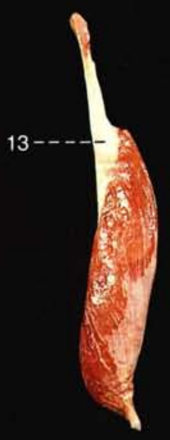

Unipennate
(semimembranous muscle)

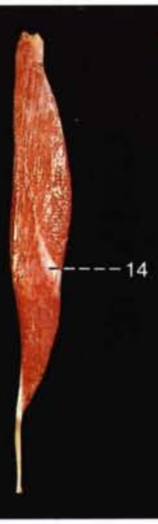

Semitendinous
(semitendinous muscle)

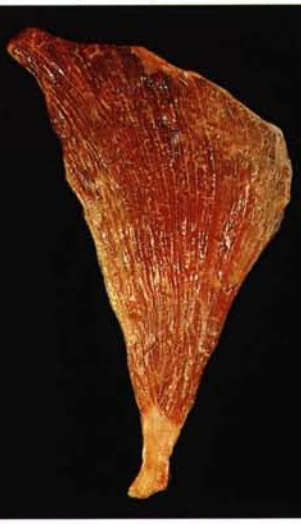

Broad, flat muscle
(latissimus dorsi muscle)

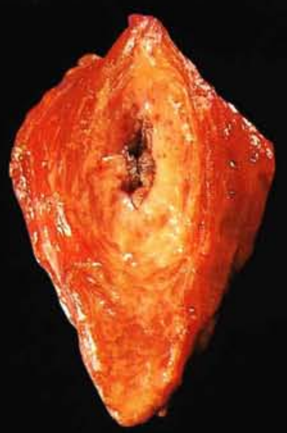

Ring-like
(external anal sphincter muscle)

1 Long head
2 Short head
3 Gastrocnemius muscle (medial head, lateral head)
4 Soleus muscle
5 Achilles tendon
6 Vastus intermedius muscle
7 Vastus lateralis muscle
8 Patella
9 Rectus femoris muscle
10 Vastus medialis muscle
11 Intermediate tendon
12 Tendinous intersection
13 Aponeurosis
14 Tendinous intersection

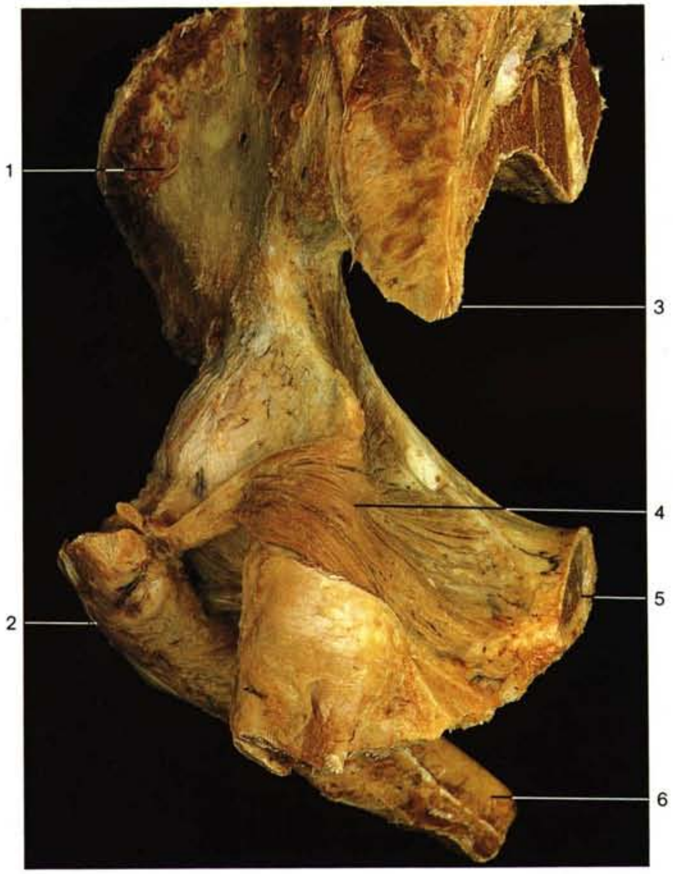

Left half of the pelvis (posterior aspect).
Obturator internus muscle as an example of a muscle, the tendon of which does not act in the direction of the main muscle fibers. Its fibers originate at the internal aspect of the obturator foramen, turn around the posterior rim of the ischium and insert at the greater trochanter of the femur. The ischium thereby serves as a pulley.

1 Ilium
2 Greater trochanter
3 Coccyx
4 Obturator internus muscle
5 Pubis
6 Femur

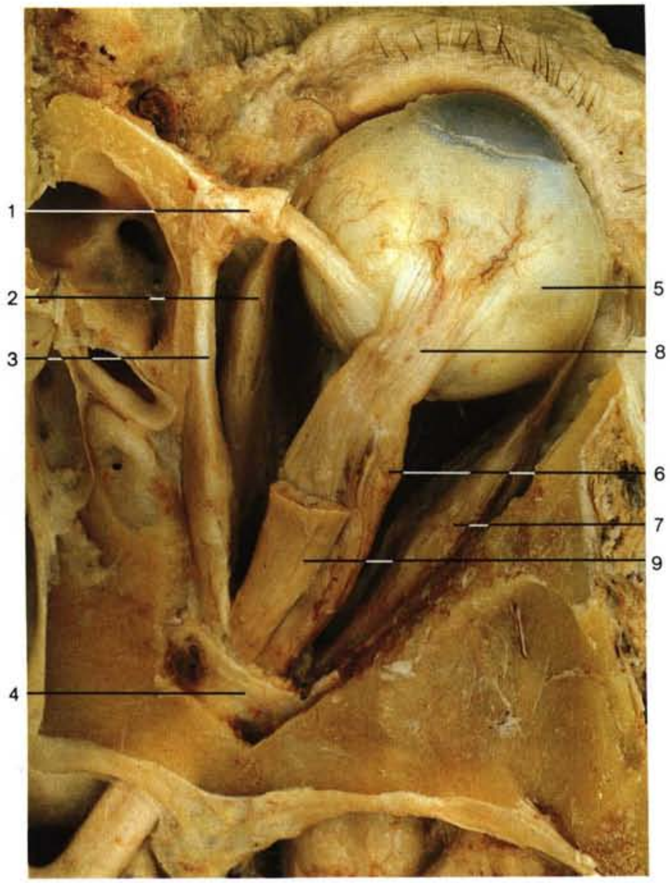

Superior oblique muscle of the eyeball, right eye (superior aspect). The tendon of this muscle bends over the trochlea changing its direction so that it becomes attached to the posterior lateral quadrant of the eyeball.

1 Trochlea
2 Medial rectus muscle
3 Superior oblique muscle
4 Common annular tendon
5 Eyeball
6 Superior rectus muscle
7 Lateral rectus muscle
8 Superior rectus muscle (tendon)
9 Levator palpebrae superioris muscle (divided)

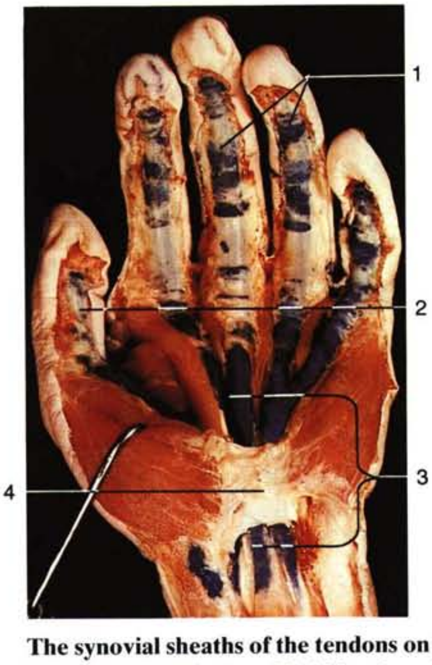

1 Digital synovial sheaths of the tendons of flexor digitorum superficialis and profundus muscles
2 Digital synovial sheaths of the tendon of long flexor pollicis longus muscle
3 Common flexor synovial sheaths of flexor digitorum superficialis and profundus muscles
4 Flexor retinaculum

The synovial sheaths of the tendons on the palmar aspect of the left wrist (colored fluid has been injected).

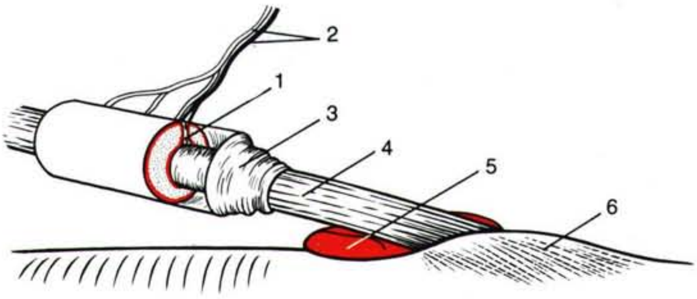

1 Mesotendon
2 Blood vessels
3 Synovial sheath
4 Tendon
5 Synovial bursa
6 Bone (tuberosity)

Structure of a tendon sheath. The synovial membrane which also forms the mesotendon is indicated in red. (Schematic drawing.)

1 Falx cerebri
2 Cerebral hemispheres
3 Tentorium cerebelli
4 Cerebellum
5 Medulla oblongata
6 Spinal cord, cervical enlargement
7 Spinal ganglia
8 Spinal cord, lumbar enlargement
9 Conus medullaris
10 Cauda equina
11 Cervical plexus (formed from ventral rami of C_1–C_4)
12 Brachial plexus (formed from ventral rami of C_5–T_1)
13 Lumbosacral plexus (formed from ventral rami of L_1–S_4)
14 Sympathetic trunk

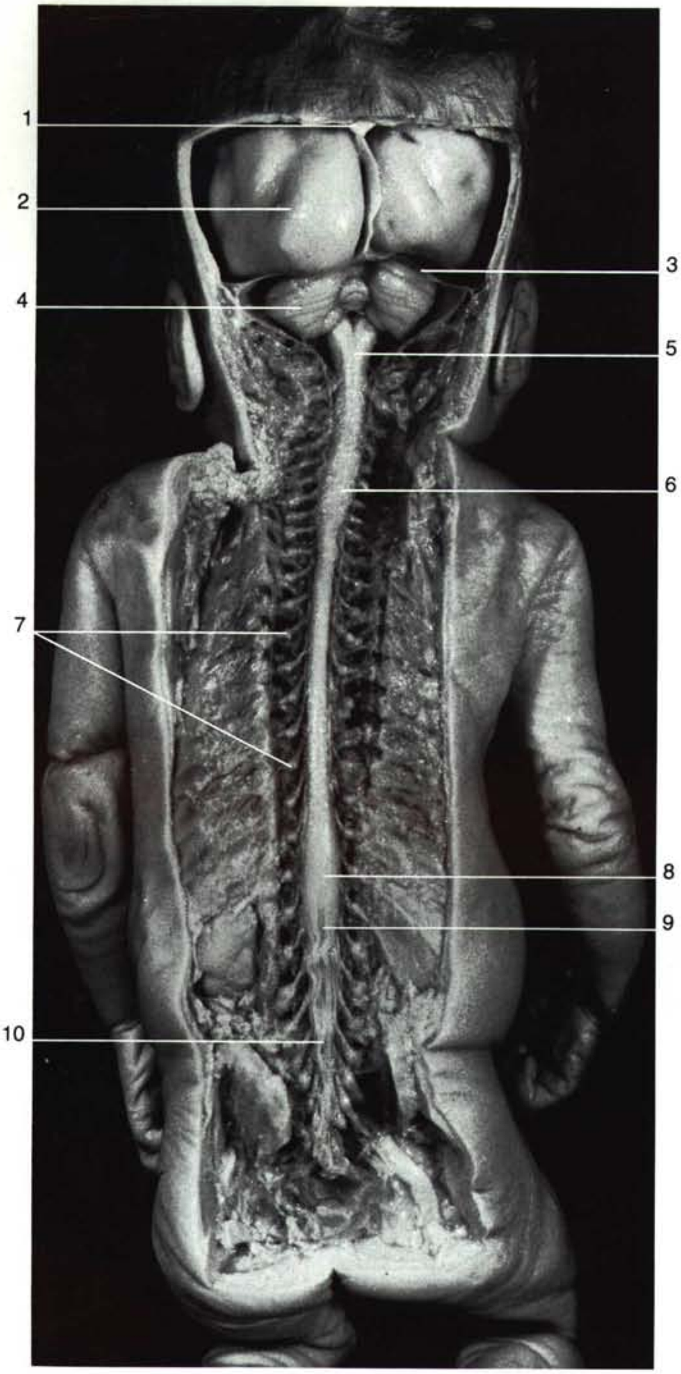

Brain, the **spinal cord** and the **spinal nerves** in the fetus (posterior aspect).

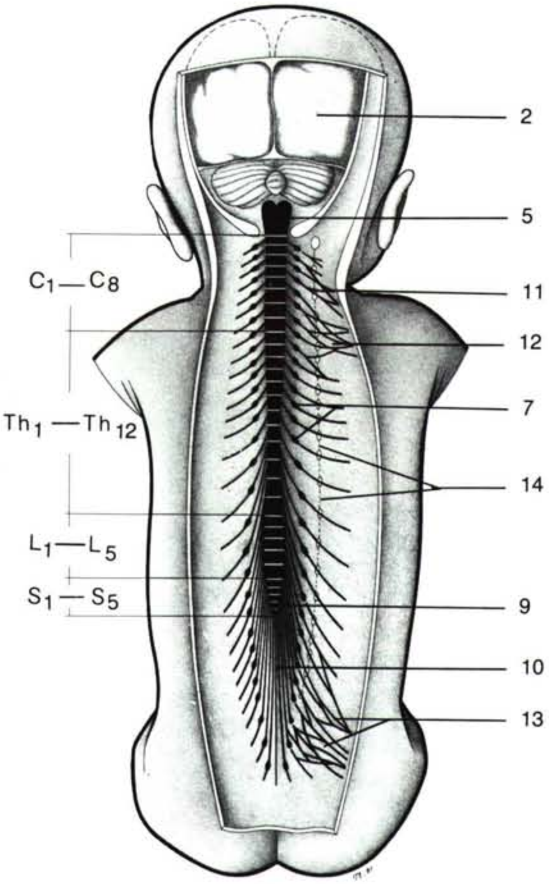

Schematic drawing to illustrate the three main parts of the nervous system in general.

The **nervous system** can be divided into three, functionally distinct parts: 1. the cranial part which comprises the great sensory organs and the brain, 2. the spinal cord which shows a segmental structure and serves predominantly as a reflex-organ, and 3. the autonomic nervous system which controls the unvoluntary functions (subconscious control) of organs and tissues. The autonomic part of the nervous system forms many delicate plexus within the organs. At certain places these plexus contain aggregations of nerve cells (prevertebral and intramural ganglia). The spinal nerves leave the spinal cord at regular intervals, forming the 8 cervical, 12 thoracic, 5 lumbar, 5 sacral and a varying number of coccygeal segments. The ventral rami of the first four cervical spinal nerves (C_1–C_4) form the cervical plexus (for innervation of the anterior neck), the ventral rami of the lower cervical spinal nerves (C_5–T_1) form the brachial plexus which innervates the upper leg, and the ventral rami of the lumbar and sacral spinal nerves form the lumbosacral plexus (L_1–S_4) which innervates the pelvic and genital organs and the lower extremity.

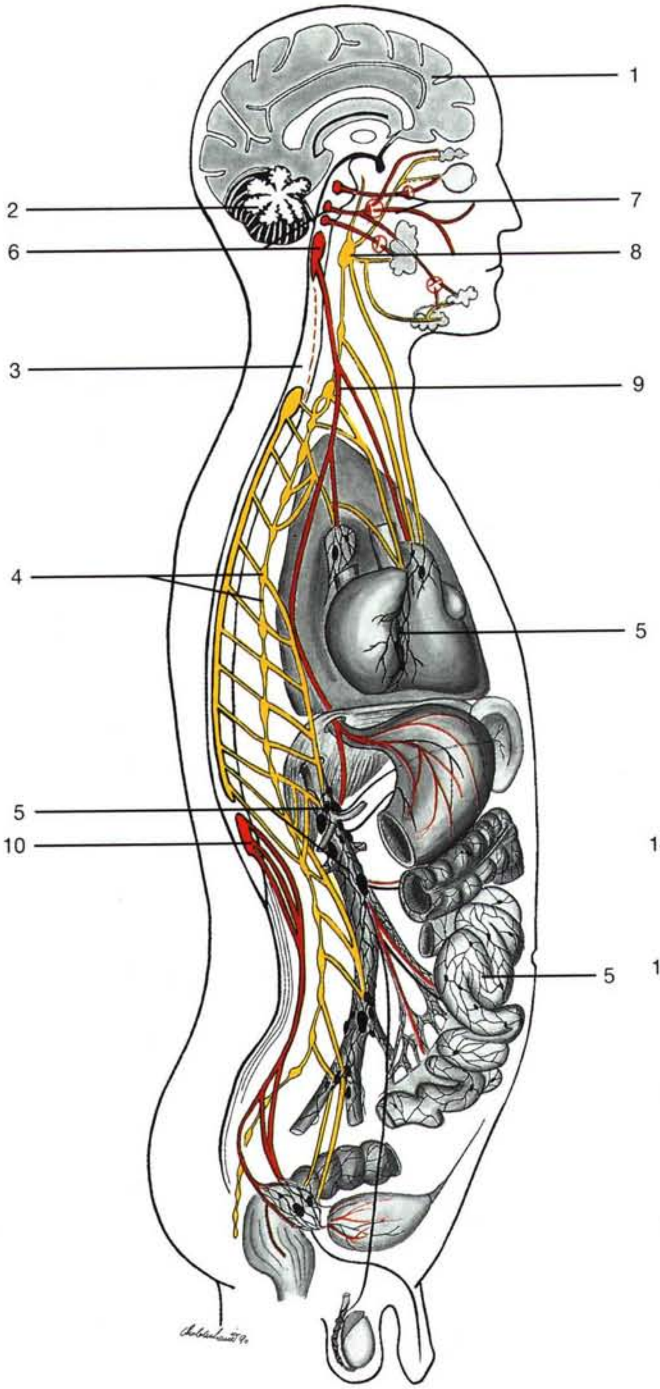

Diagram illustrating the localization of the **three functional portions** of the **nervous system** (brain, spinal cord and autonomic nervous system). Yellow = sympathetic system; red = parasympathetic system.

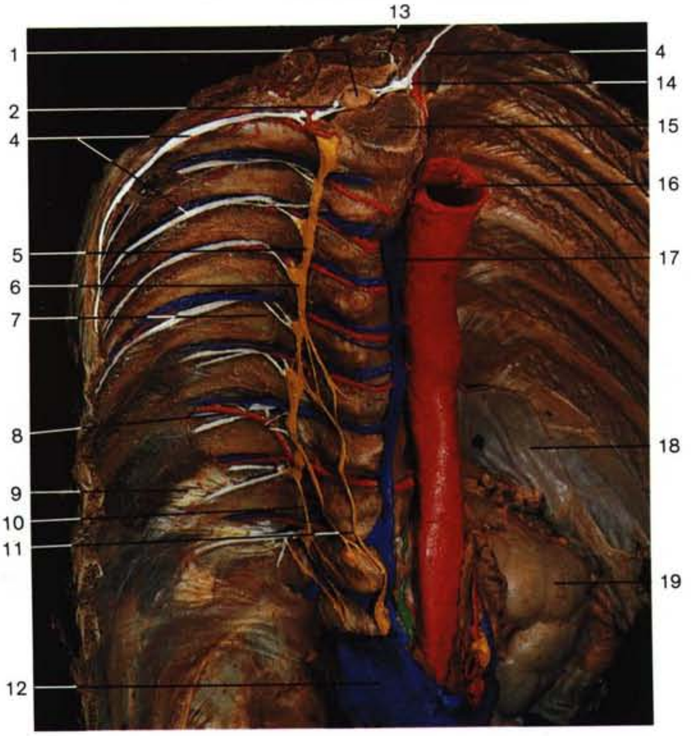

Posterior part of the thorax. Cross-section at the level of the 5th thoracic segment. Spinal nerves and their connections to the sympathetic trunk.

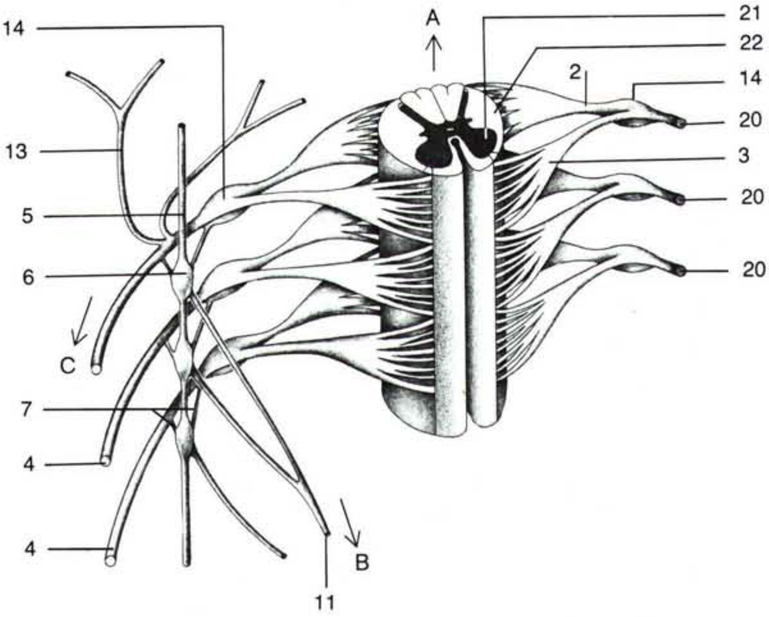

Organization of the spinal cord in structurally equal segments which form the paired spinal nerves. A = connections to the brain; B = connections to the autonomic nervous system; C = connections to the trunk and extremities (intercostal nerves and plexus). (Schematic drawing.)

1 Cerebrum
2 Cerebellum
3 Spinal cord
4 Sympathetic trunk and ganglion
5 Plexus and ganglia of the autonomic nervous system
6 Cranial autonomic system
7 Cranial nerves (n. III and n. VII)
8 Superior cervical ganglion
9 Vagus nerve (n. X)
10 Sacral autonomic system

1 Spinal cord
2 **Dorsal root**
3 **Ventral root**
4 Intercostal nerves
5 Sympathetic trunk
6 Ganglia of the sympathetic trunk
7 Rami communicantes
8 Intercostal artery and vein
9 Subcostalis muscle
10 Lesser splanchnic nerve
11 Greater splanchnic nerve
12 Inferior vena cava
13 Dorsal ramus of spinal nerve
14 **Spinal (dorsal root) ganglion**
15 Body of the vertebra
16 Aorta
17 Azygos vein
18 Diaphragm
19 Left kidney
20 **Spinal nerve**
21 Gray matter of the spinal cord
22 White matter of the spinal cord

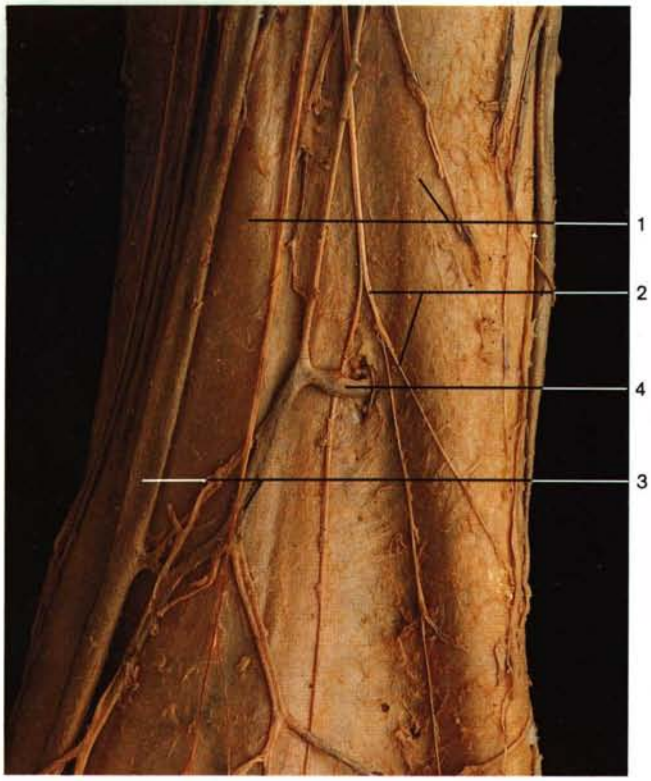

Superficial nerves and vessels of the lower leg, illustrating the structural differences between veins and nerves.

1 Crural fascia (fascia cruris)
2 Cutaneous nerves
3 Superficial cutaneous veins
4 Perforating vein

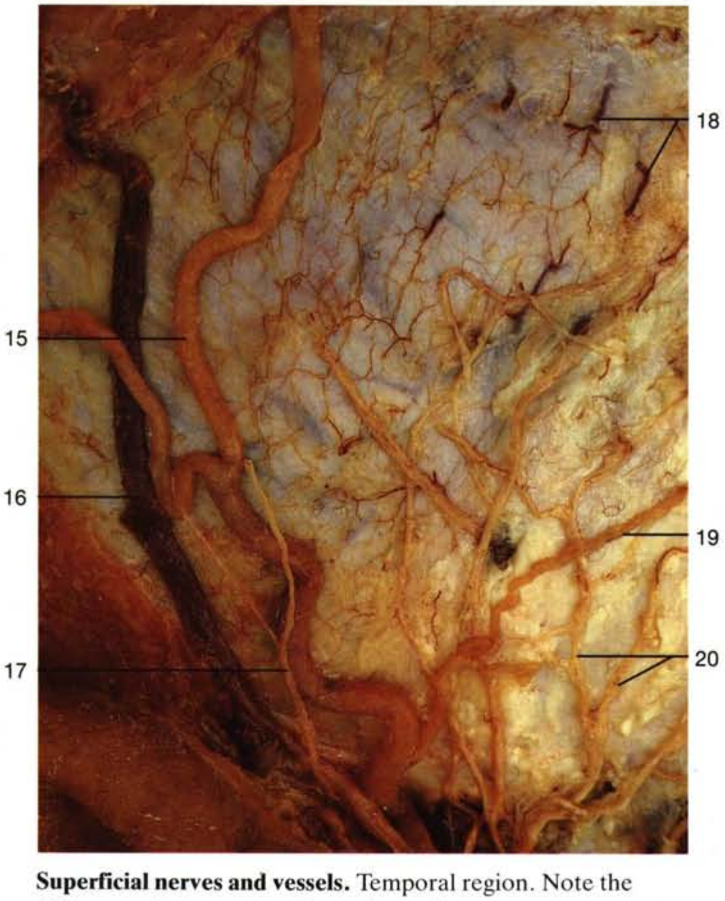

Superficial nerves and vessels. Temporal region. Note the differences between arteries, veins and nerves.

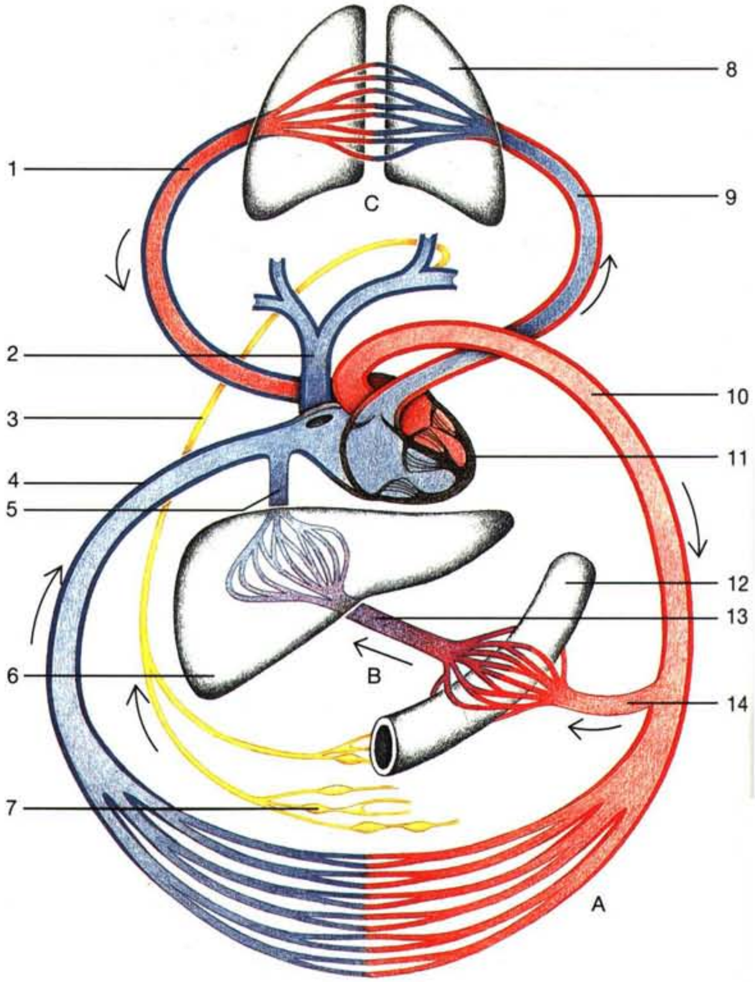

Organization of the circulatory system. Arrows: direction of the blood flow.

Vessel wall in red = Arteries
Vessel wall in blue = Veins
Yellow = Lymphatic vessels

A Systemic circulation
B Hepatic portal circulation
C Pulmonary circulation

1 Pulmonary vein
2 Superior vena cava
3 Thoracic duct
4 Inferior vena cava
5 Hepatic vein
6 Liver
7 Lymph nodes and lymphatic vessels
8 Lung
9 Pulmonary artery
10 Aorta
11 Heart
12 Small intestine with capillary network
13 Portal vein
14 Mesenteric artery
15 Superficial temporal artery
16 Superficial temporal vein
17 Auriculotemporal nerve
18 Perforating veins for subcutaneous fatty tissue
19 Small artery
20 Small nerves (branches of facial nerve)

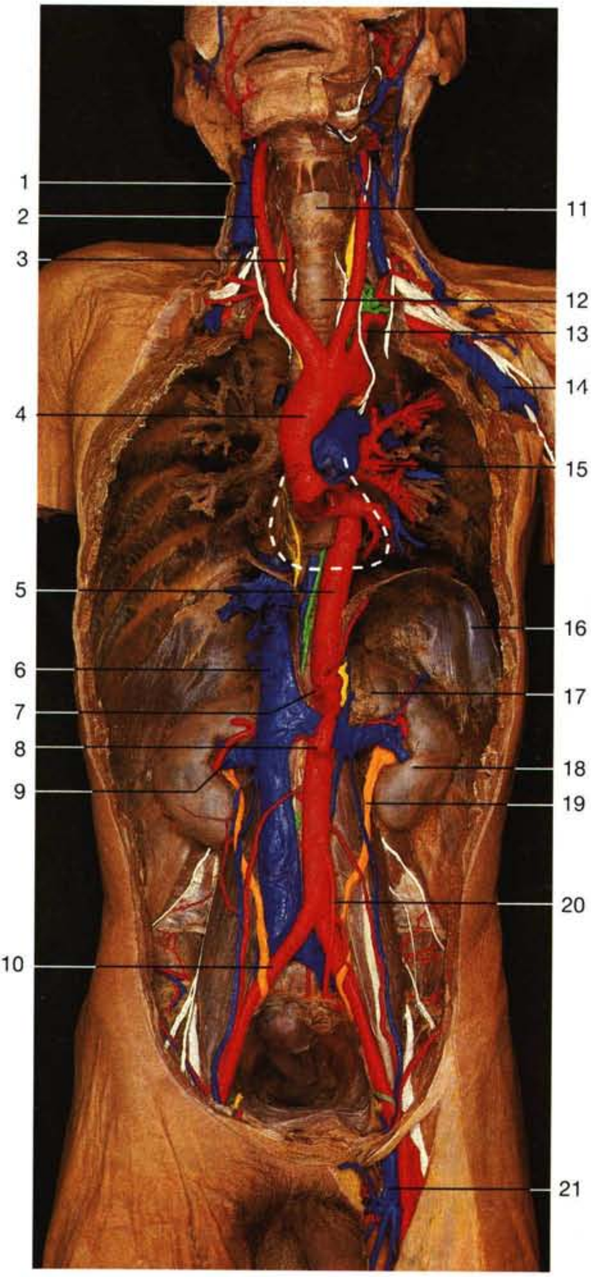

Major vessels of the trunk. The position of the heart is indicated by the dotted line.

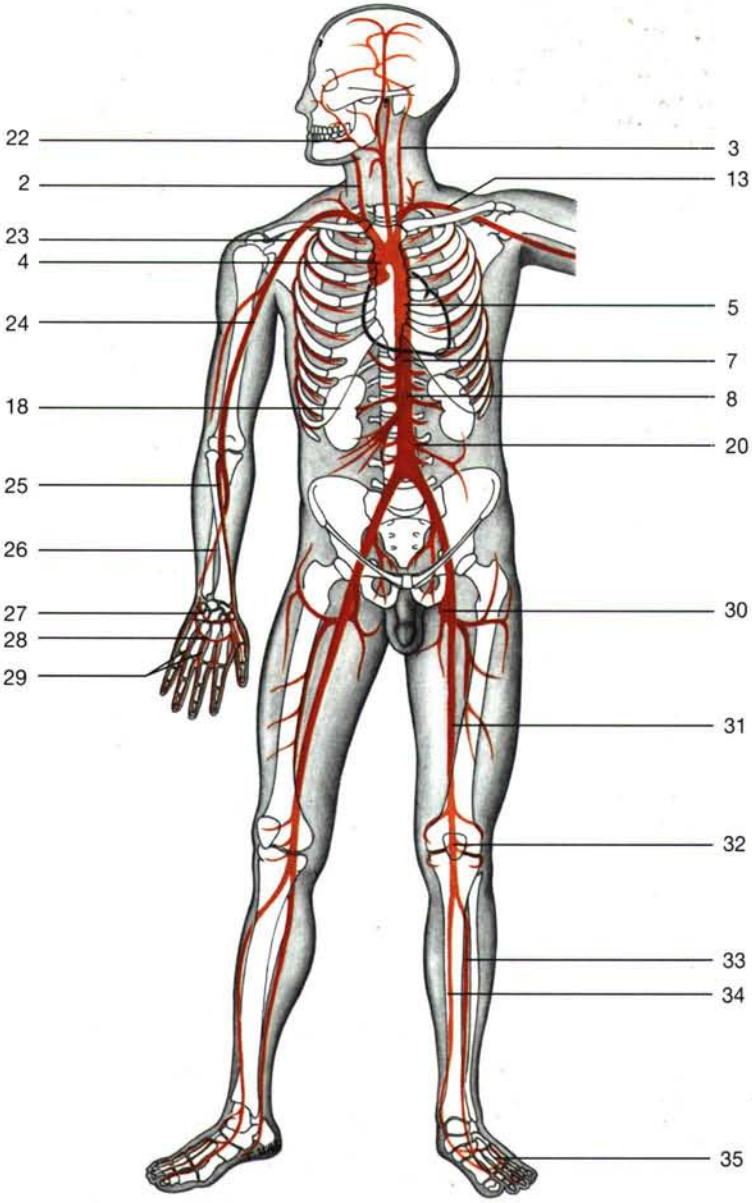

Major arteries of the human body. (Schematic drawing.)

1 Internal jugular vein
2 **Common carotid artery**
3 Vertebral artery
4 **Ascending aorta**
5 **Descending aorta**
6 **Inferior vena cava**
7 Celiac trunk
8 Superior mesenteric artery
9 Renal vein
10 **Common iliac artery**
11 Larynx
12 Trachea
13 Left subclavian artery
14 Left axillary vein
15 **Pulmonary veins**
16 Diaphragm
17 Suprarenal gland

18 Kidney
19 Ureter
20 Inferior mesenteric artery
21 **Femoral vein**
22 Facial artery
23 Axillary artery
24 Brachial artery
25 Radial artery
26 Ulnar artery
27 Deep palmar arch
28 Superficial palmar arch
29 Common palmar digital arteries
30 Profunda femoris artery
31 **Femoral artery**
32 Popliteal artery
33 Anterior tibial artery
34 Posterior tibial artery
35 Plantar arch

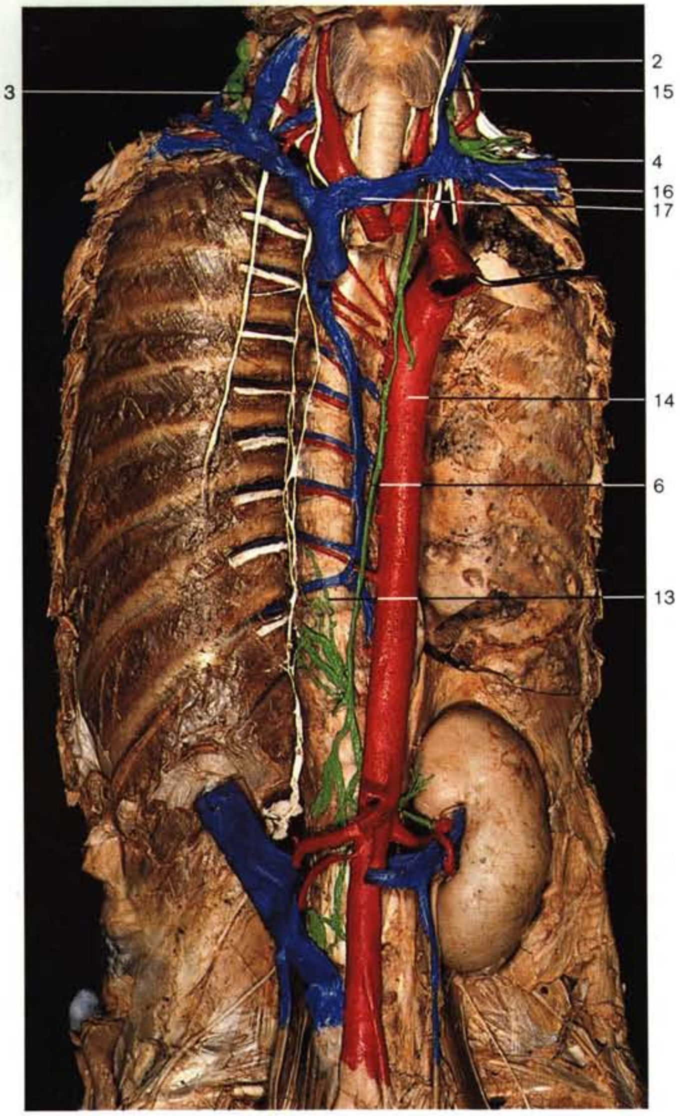

Major lymph vessels of the trunk.

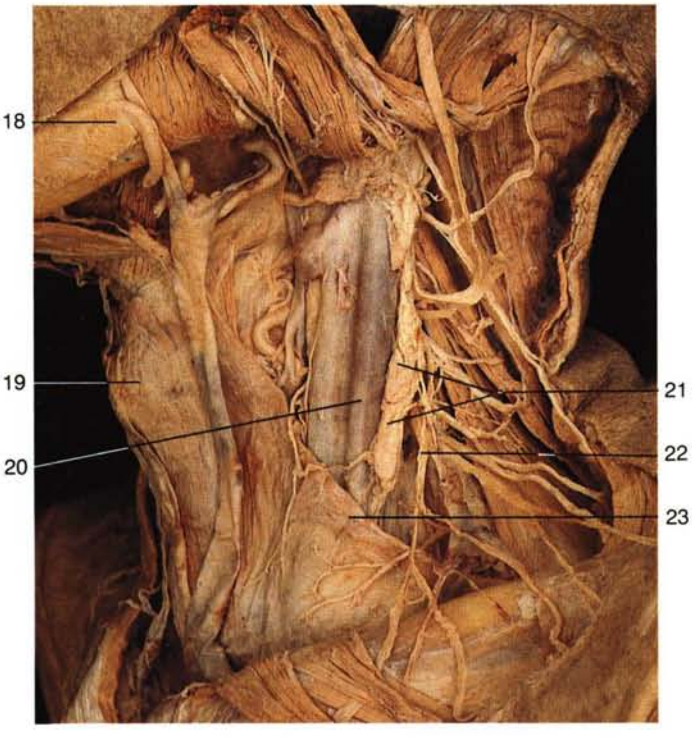

Deep cervical nodes, adjacent to the internal jugular vein.

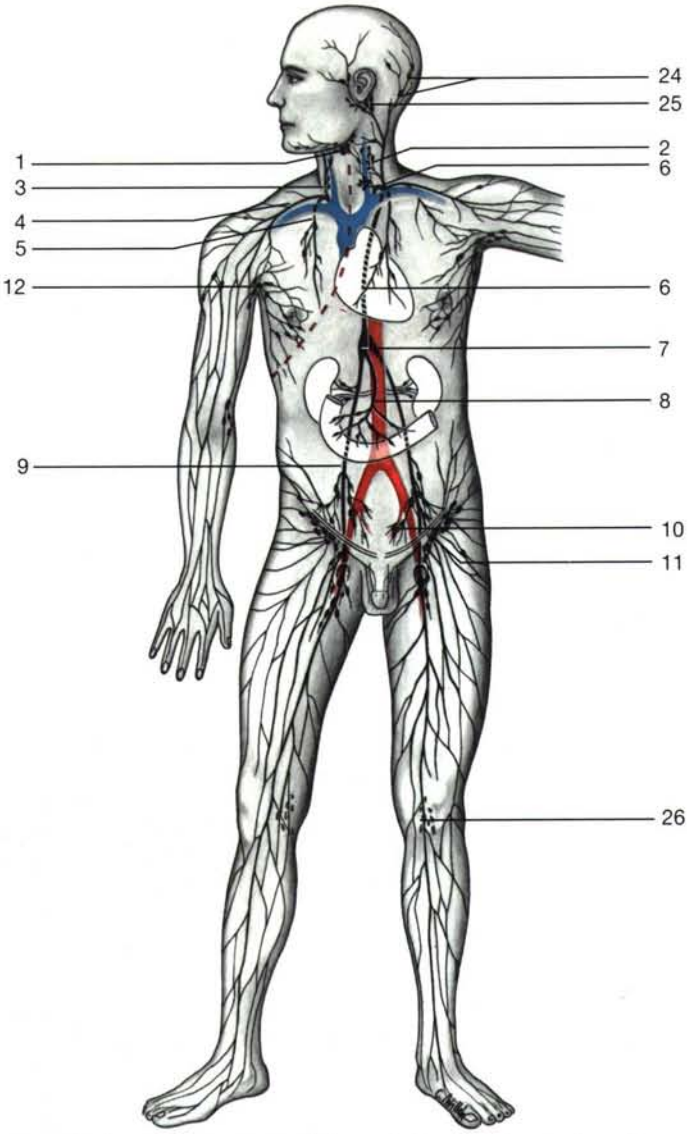

Lymphatic system. Course of the main lymphatic vessels and lymph nodes in the body. Dotted line = border between lymphatic vessels draining towards the right venous angle and towards the left.

1 Submandibular nodes
2 Deep cervical nodes
3 Right jugular trunk
4 Subclavian trunk
5 Right bronchomediastinal trunk
6 Thoracic duct
7 **Cisterna chyli**
8 **Intestinal trunk**
9 Right **lumbar trunk**
10 Internal iliac nodes
11 **Inguinal nodes**
12 **Axillary nodes**
13 Descending trunk
14 Descending aorta
15 Internal jugular vein
16 Subclavian vein
17 Left brachiocephalic vein
18 Mandible
19 Larynx
20 Internal jugular vein
21 **Deep cervical nodes**
22 Cervical nerve plexus
23 Superficial layer of deep fascia
24 Occipital nodes
25 Parotid nodes
26 Popliteal nodes

Lymphatic vessels originate as blind-ending tubes in the tissue spaces (lymph capillaries) and unite to form larger vessels (lymphatics). These resemble veins but have a much thinner wall, more valves and are interrupted by lymph nodes at various intervals. Large groups of lymph nodes are located in the inguinal and axillary regions, deep to the mandible and sternocleidomastoid muscle and within the root of the mesentery of the intestine.

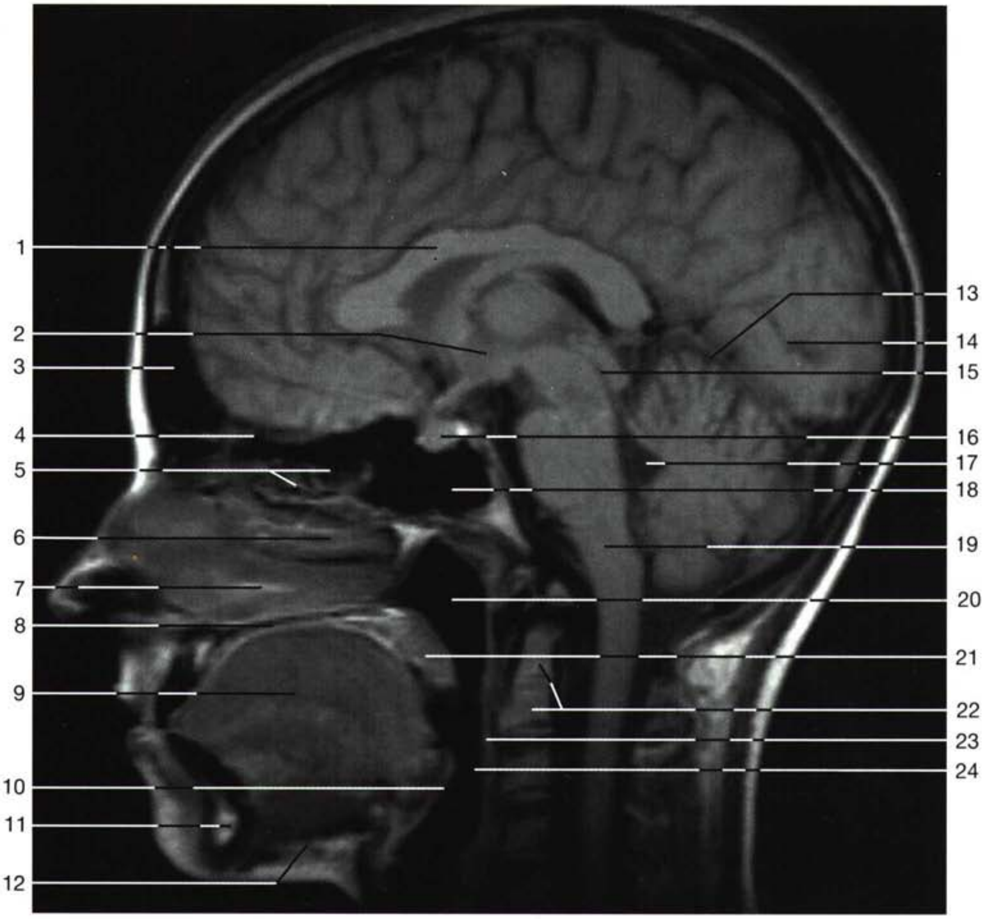

1 Corpus callosum
2 **Hypothalamus**
3 Frontal sinus
4 Cribriform plate
5 Ethmoidal air cells
6 Middle nasal concha
7 Inferior nasal concha
8 Hard palate
9 **Tongue**
10 Epiglottis
11 Mandible
12 Mylohyoid muscle
13 Tentorium of cerebellum
14 Calcarine fissure
15 Cerebral aqueduct
16 **Pituitary gland**
17 Fourth ventricle
18 **Sphenoidal sinus**
19 Medulla oblongata
20 Nasopharynx
21 Uvula
22 Dens of axis
23 Constrictor muscle of pharynx
24 Oral part of pharynx
25 **Cerebrum** (right hemisphere)
26 Calvaria
27 **Cerebellum**

Sagittal section through head and neck (MR-Scan. 23-year-old female, courtesy of PD Dr. A. Heuck, Munich).

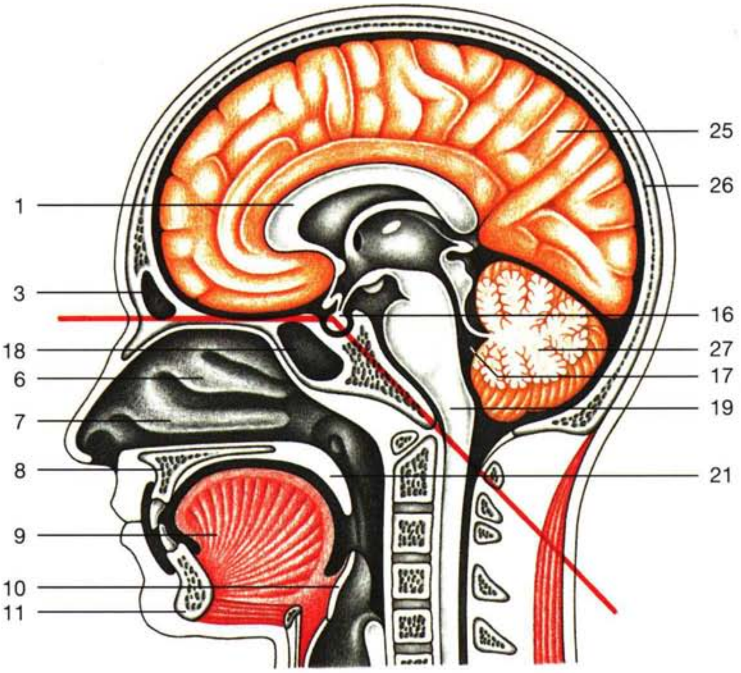

Sagittal section through the head
(schematic drawing). The red line represents the border between the neurocranium and viscerocranium forming the clivus angle. The neural cavity contains the brain; the viscerocranium comprises the orbit, the nasal cavity and the oral cavity arranged one beneath the other.

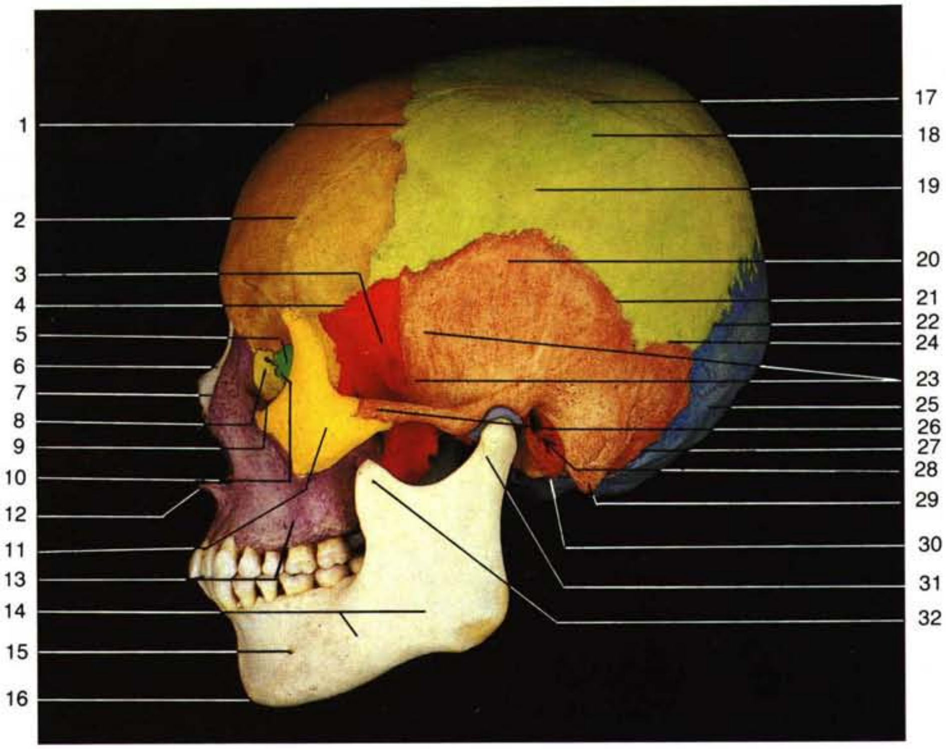

General architecture of the skull (lateral aspect). The different bones are indicated in color (numbers cf. table).

1 Coronal suture
2 **Frontal bone**
3 **Sphenoid bone**
4 Sphenofrontal suture
5 **Ethmoid bone**
6 **Nasal bone**
7 Nasomaxillary suture
8 **Lacrimal bone**
9 Lacrimomaxillary suture
10 Lacrimoethmoid suture
11 **Zygomatic bone**
12 Anterior nasal spine
13 **Maxilla**
14 **Mandible**
15 Mental foramen
16 Mental protuberance
17 Superior temporal line
18 Inferior temporal line
19 **Parietal bone**
20 **Temporal bone**
21 Squamous suture
22 Lambdoid suture
23 Temporal fossa
24 Parietomastoid suture
25 **Occipital bone**
26 Zygomatic arch
27 Occipitomastoid suture
28 External acoustic meatus
29 Mastoid process
30 Tympanic portion of temporal bone
31 Condylar process of mandible
32 Coronoid process of mandible

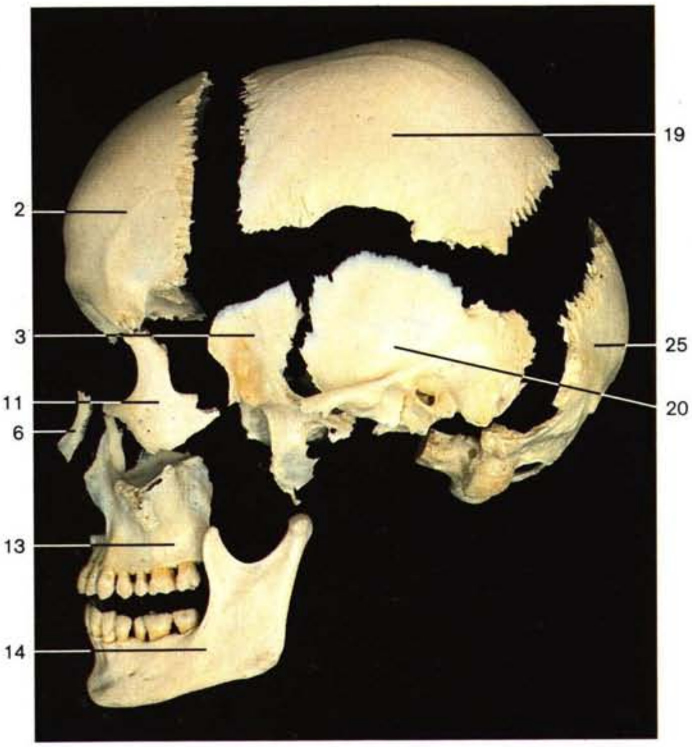

Lateral aspect of the disarticulated skull (palatine bone, lacrimal bone, ethmoid bone and vomer are not depicted).

2	Frontal bone (orange)	Cranial bones
19	Parietal bone (light green)	
3	Greater wing of sphenoid bone (red)	
25	Squama of occipital bone (blue)	
20	Squama of temporal bone (brown)	
5	Ethmoid bone (dark green)	Base of skull
3	Sphenoid bone (red)	
	Temporal bone excluding squama (brown)	
30	Tympanic portion of temporal bone (dark brown)	
	Occipital bone excluding squama (blue)	
6	Nasal bone (white)	Facial bones
8	Lacrimal bone (yellow)	
	Inferior nasal concha	
	Vomer	
11	Zygomatic bone (light yellow)	
	Palatine bone	
13	Maxilla (violet)	
14	Mandible (white)	
	Malleus Incus Stapes } within petrous portion of temporal bone	Auditory ossicles
	Hyoid	

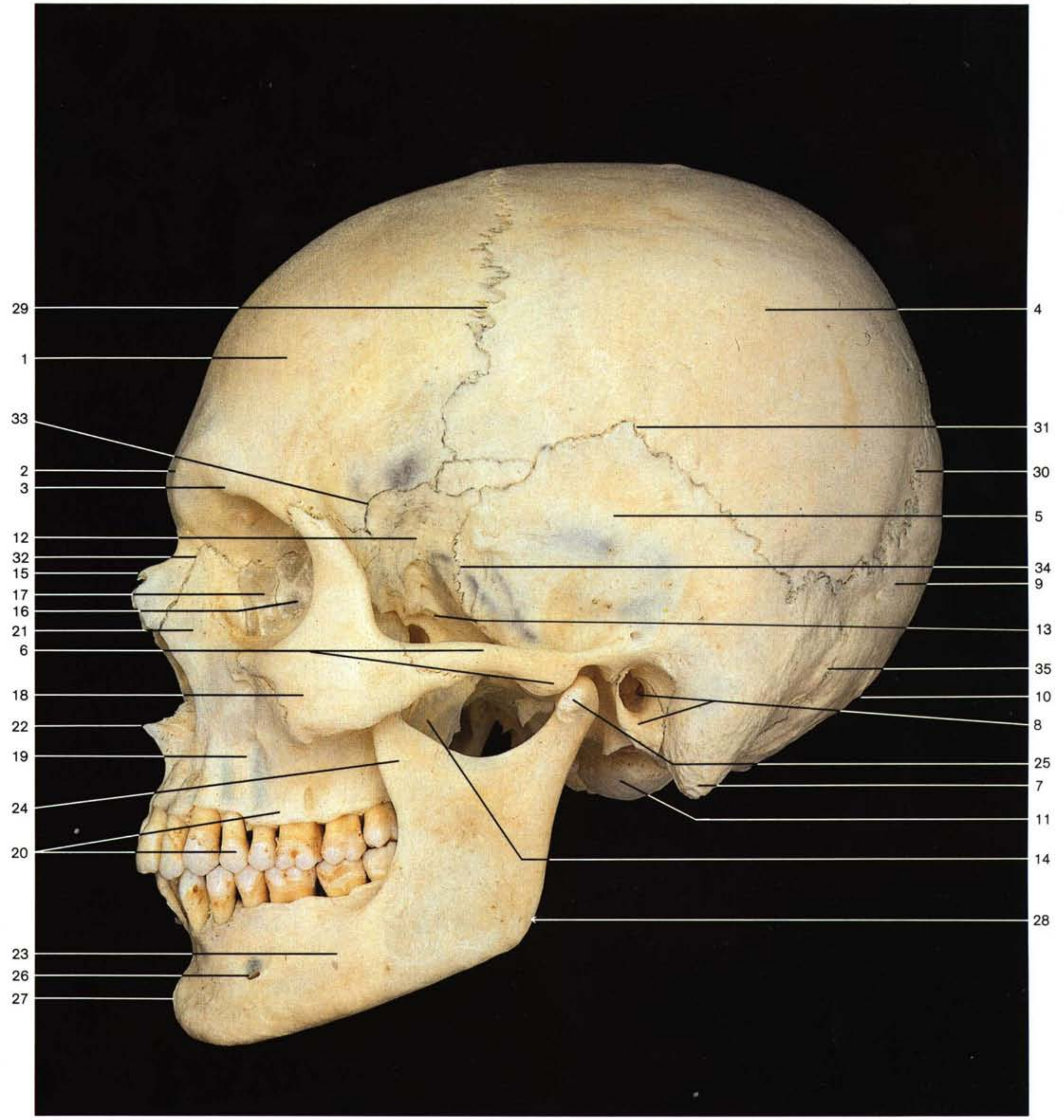

Lateral aspect of the skull.

1 **Frontal bone**
2 Glabella
3 Supraorbital margin
4 **Parietal bone**
5 **Temporal bone** (squamous part)
6 Zygomatic process (articular tubercle)
7 Mastoid process
8 Tympanic part (tympanic plate) and external acoustic meatus
9 **Occipital bone** (squamous part)
10 External occipital protuberance
11 Occipital condyle
12 **Sphenoid bone** (greater wing)
13 Infratemporal crest of sphenoid
14 Pterygoid process (lateral pterygoid plate)
15 **Nasal bone**
16 **Ethmoid bone** (orbital part)
17 **Lacrimal bone**
18 **Zygomatic bone**
19 **Maxilla** (body)
20 Alveolar process and teeth
21 Frontal process
22 Anterior nasal spine
23 **Mandible** (body)
24 Coronoid process
25 Condylar process
26 Mental foramen
27 Mental protuberance
28 Angle of the mandible

Sutures

29 Coronal suture
30 Lambdoid suture
31 Squamous suture
32 Nasomaxillary suture
33 Frontosphenoid suture
34 Sphenosquamosal suture
35 Occipitomastoid suture

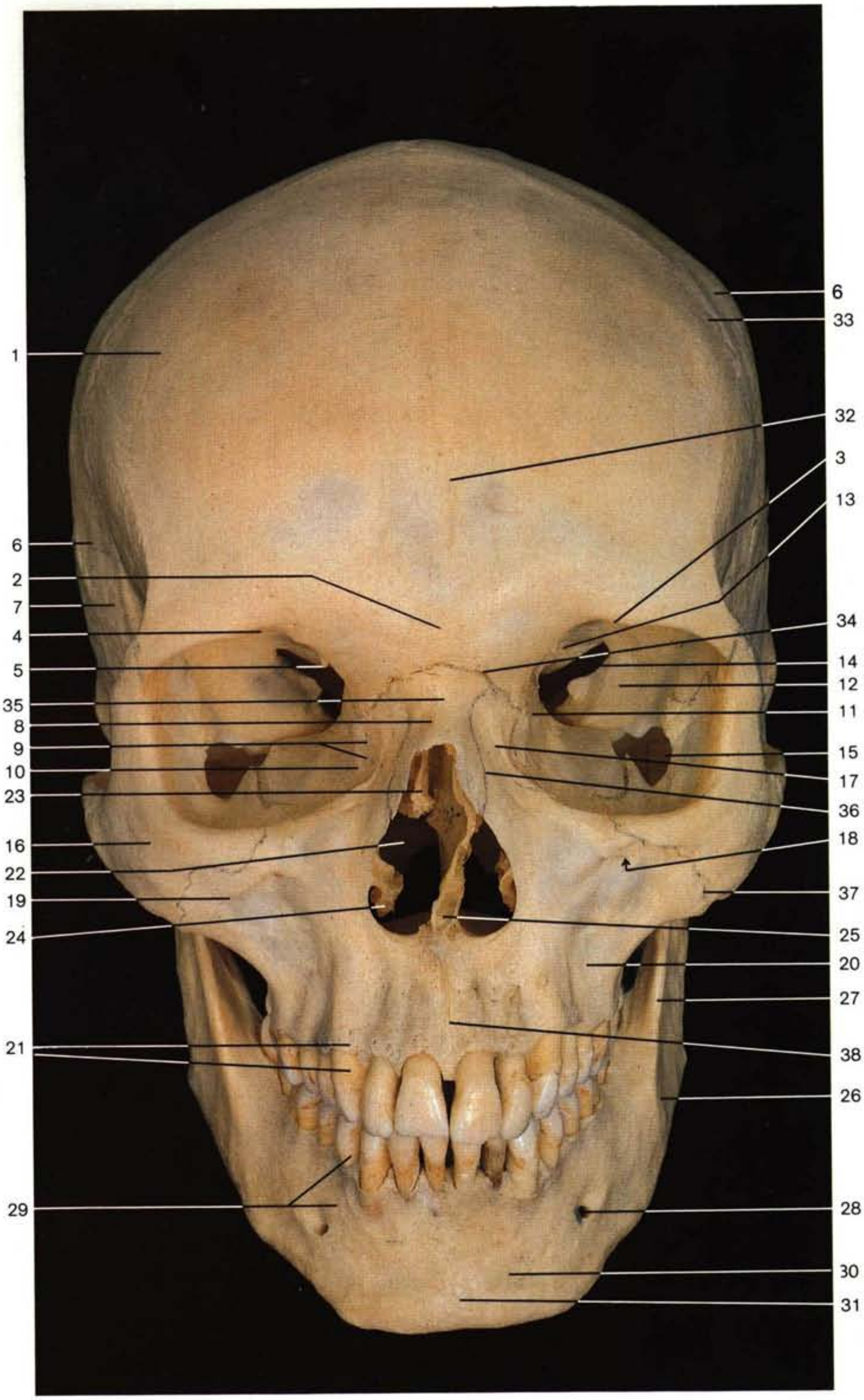

Anterior aspect of the skull.

1 **Frontal bone**
2 Glabella
3 Supraorbital margin
4 Supraorbital notch
5 Trochlear spine
6 **Parietal bone**
7 **Temporal bone**
8 **Nasal bone**

Orbit
9 Lacrimal bone
10 Posterior lacrimal crest
11 Ethmoid bone

Sphenoid bone
12 Greater wing of sphenoid bone
13 Lesser wing of sphenoid bone
14 Superior orbital fissure
15 Inferior orbital fissure
16 **Zygomatic bone**

Maxilla
17 Frontal process
18 Infraorbital foramen
19 Zygomatic process
20 Body of maxilla
21 Alveolar process with teeth

Nasal cavity
22 Anterior nasal aperture
23 Middle nasal concha
24 Inferior nasal concha
25 Nasal septum, vomer

Mandible
26 Body of mandible
27 Ramus of mandible
28 Mental foramen
29 Alveolar part with teeth
30 Base of mandible
31 Mental protuberance

Sutures
32 Frontal suture
33 Coronal suture
34 Frontonasal suture
35 Internasal suture
36 Nasomaxillary suture
37 Zygomaticomaxillary suture
38 Intermaxillary suture

The skull comprises a mosaic of numerous complicated bones which form the cranial cavity protecting the brain **(neurocranium)** and several cavities such as nasal and oral cavities in the facial region. The neurocranium consists of large bony plates which develop directly from the surrounding sheets of connective tissue **(desmocranium).** The bones of the skull base are formed out of cartilaginous tissue **(chondrocranium)** which ossifies secondarily. The **visceral skeleton** which, in fish, gives rise to the gills, has in higher vertebrates been transformed into the bones of the masticatory and auditory apparatus (maxilla, mandible, auditory ossicles and hyoid bone).

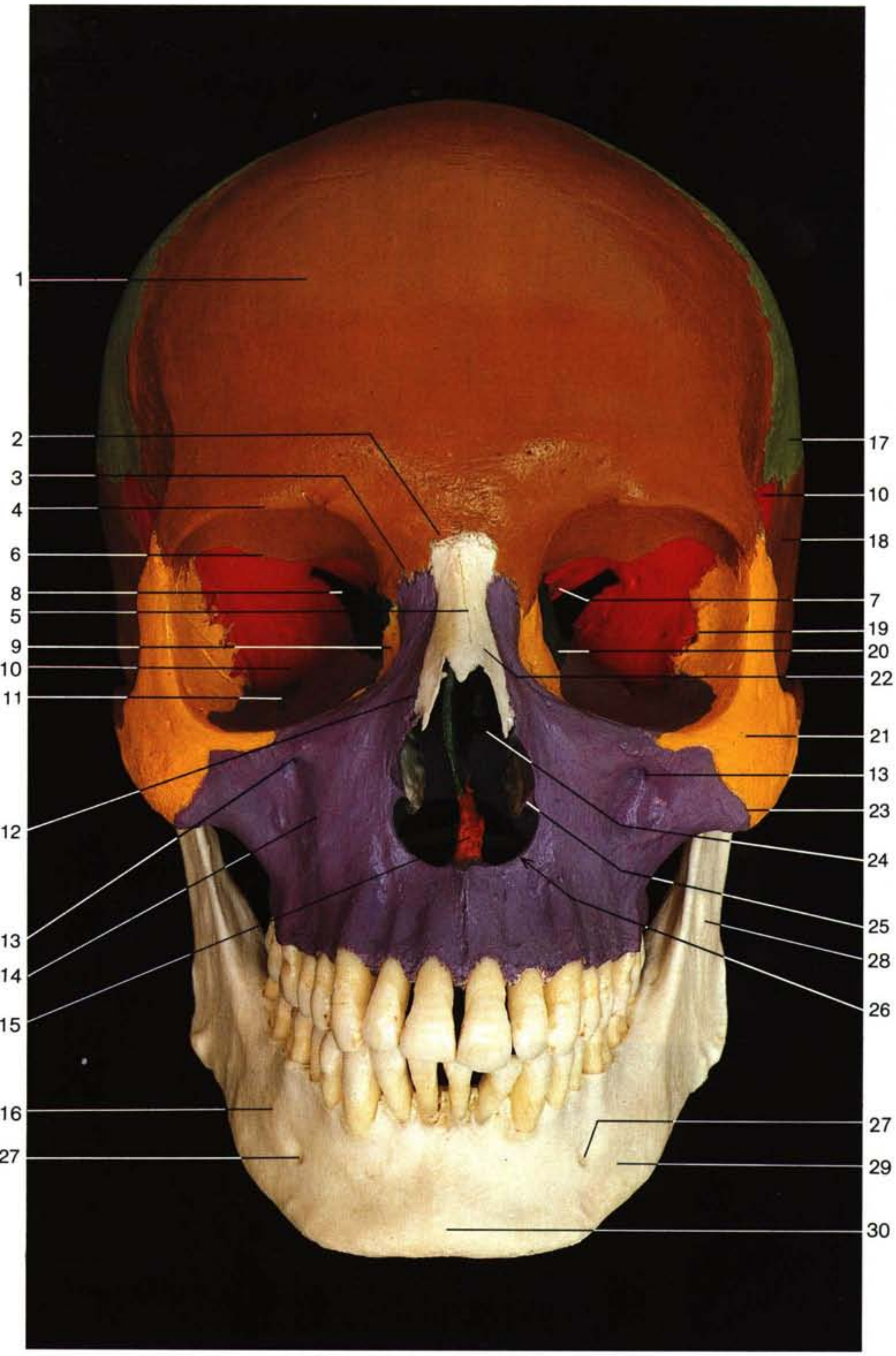

1 **Frontal bone**
2 Frontonasal suture
3 Frontomaxillary suture
4 Supraorbital margin
5 Internasal suture
6 Sphenofrontal suture
7 Optic canal in lesser wing of sphenoid bone
8 Superior orbital fissure
9 **Lacrimal bone**
10 **Sphenoid bone** (greater wing)
11 Inferior orbital fissure
12 Nasomaxillary suture
13 Infraorbital foramen
14 **Maxilla**
15 **Vomer**
16 Body of mandible
17 **Parietal bone**
18 **Temporal bone**
19 Sphenozygomatic suture
20 **Ethmoid bone**
21 **Zygomatic bone**
22 **Nasal bone**
23 Zygomaticomaxillary suture
24 Middle nasal concha
25 Inferior nasal concha
26 Anterior nasal aperture
27 Mental foramen
28 Ramus of mandible
29 Base of mandible
30 Mental protuberance

Bones
Frontal bone (brown)
Parietal bone (light green)
Temporal bone (dark brown)
Sphenoidal bone (red)
Zygomatic bone (yellow)
Ethmoid bone (dark green)
Lacrimal bone (yellow)
Vomer (orange)
Maxilla (violet)
Nasal bone (white)
Mandible (white)

Anterior aspect of the skull (individual bones indicated by color).

The following series of figures are arranged so that the mosaic-like pattern of the skull becomes understandable. It starts with the bones of the **skull base** (sphenoid and occipital) to which the other bones are added step by step. The facial skeleton is built up by the ethmoid bone to which the palatine bone and maxilla are attached laterally; the small nasal and lacrimal bones fill the remaining spaces. Cartilages remain only in the external part of the nose.

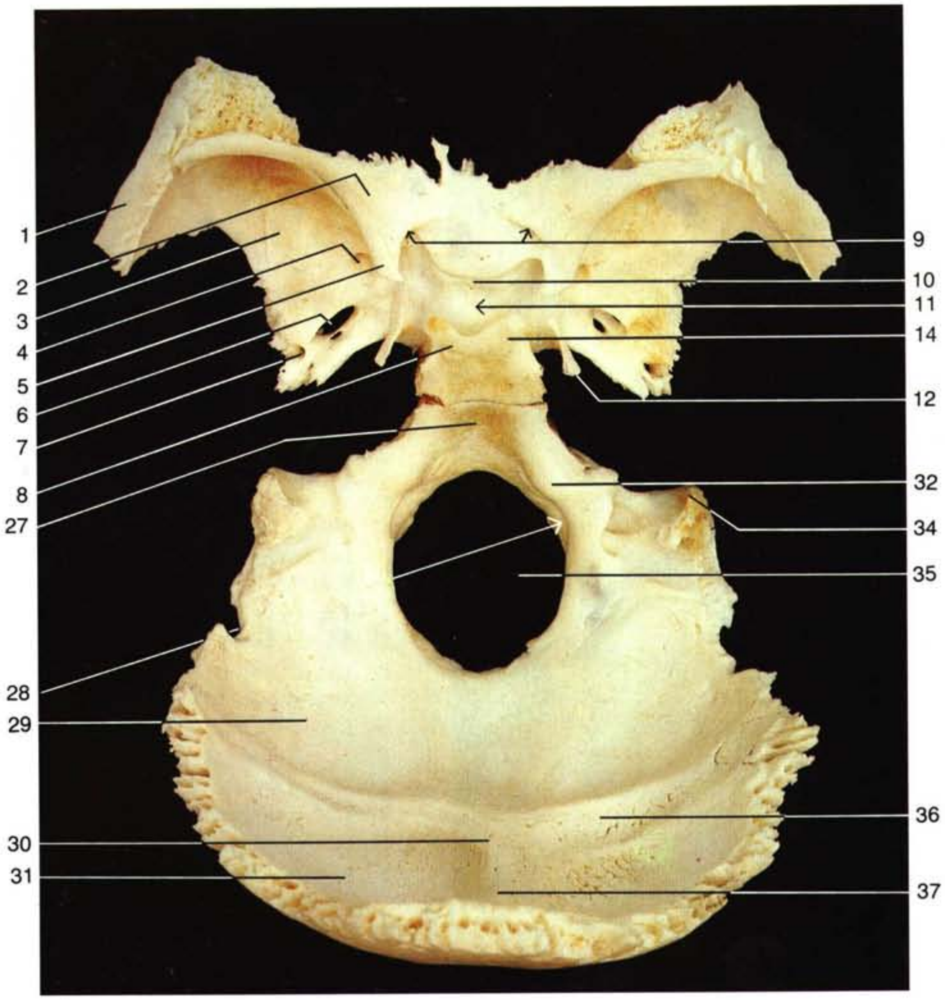
1
2
3
4
5
6
7
8
27
28
29
30
31
9
10
11
14
12
32
34
35
36
37

Sphenoid and occipital bone (from above).

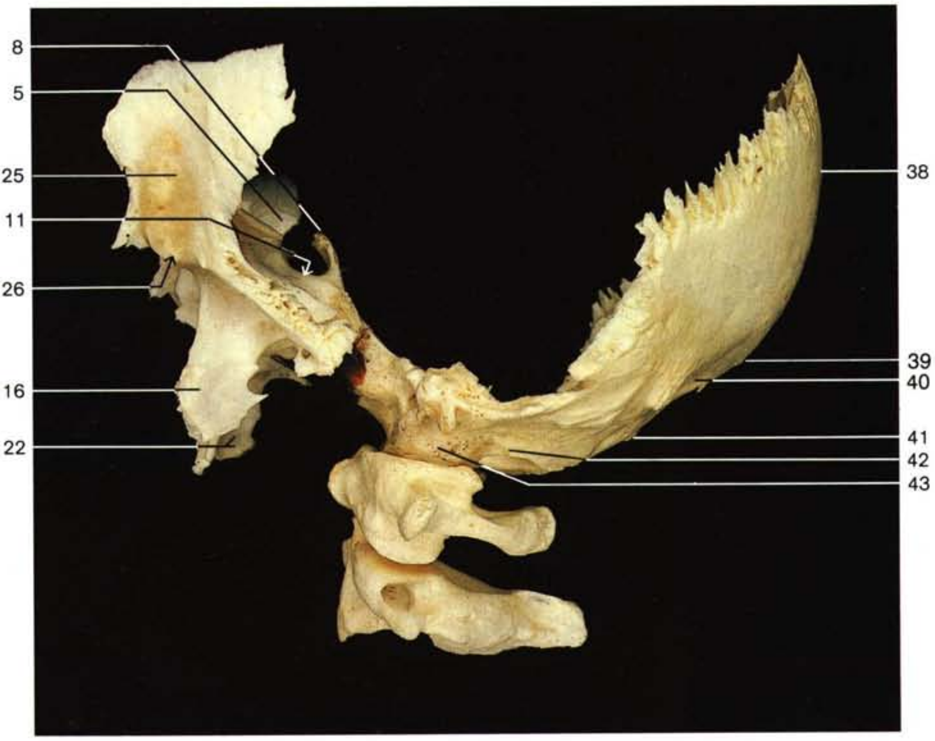
8
5
25
11
26
16
22
38
39
40
41
42
43

Sphenoid and occipital bone in connection with the atlas and axis (1st and 2nd cervical vertebrae) (left lateral view).

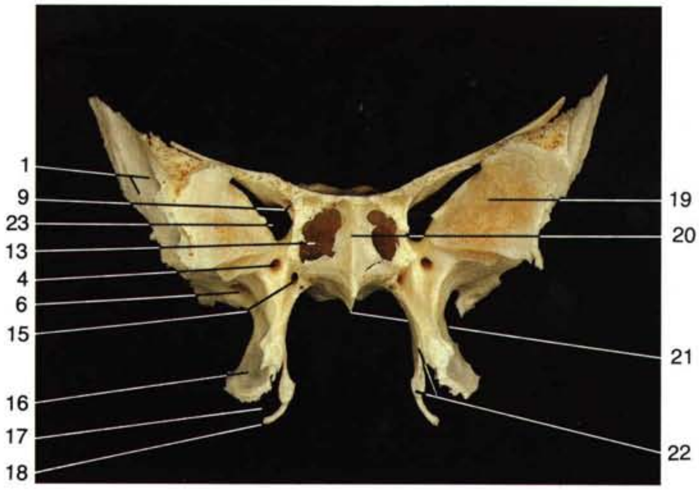

Sphenoid bone (anterior aspect).

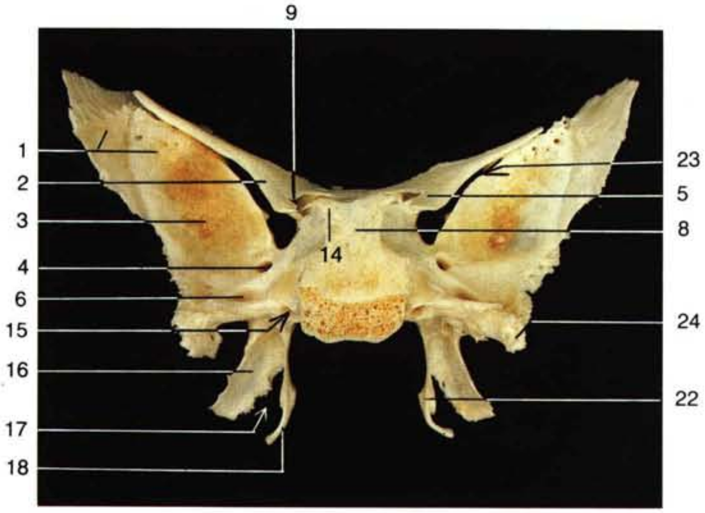

Sphenoid bone (posterior aspect).

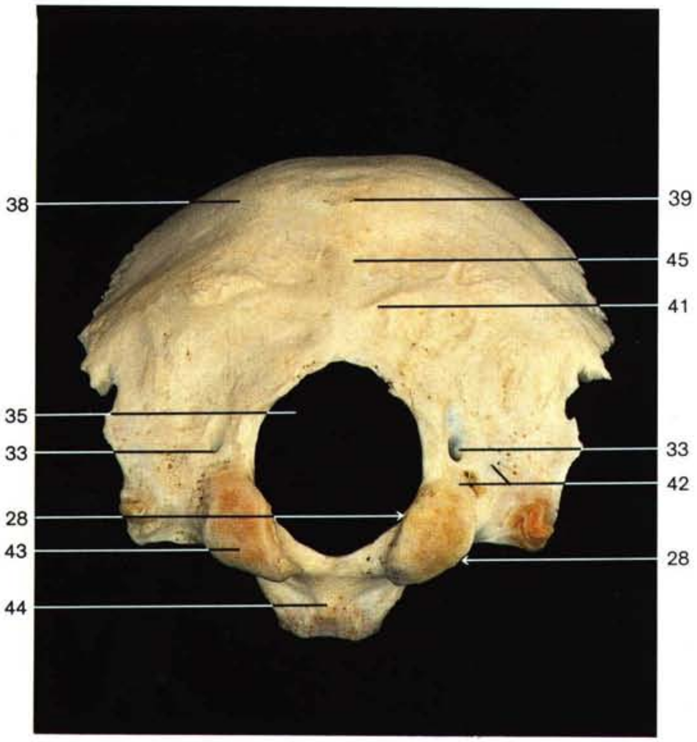

Occipital bone (from below).

Sphenoid bone

1 **Greater wing**
2 **Lesser wing**
3 Cerebral or superior surface of greater wing
4 **Foramen rotundum**
5 Anterior clinoid process
6 Foramen ovale
7 Foramen spinosum
8 Dorsum sellae
9 Optic canal
10 Chiasmatic groove (sulcus chiasmatis)
11 **Hypophysial fossa (sella turcica)**
12 Lingula
13 Opening of sphenoidal sinus
14 Posterior clinoid process
15 **Pterygoid canal**
16 Lateral pterygoid plate of **pterygoid process**
17 Pterygoid notch
18 Pterygoid hamulus
19 Orbital surface of greater wing
20 Sphenoid crest
21 Sphenoid rostrum
22 Medial pterygoid plate
23 Superior orbital fissure
24 Spine of sphenoid
25 Temporal surface of greater wing
26 Infratemporal crest

Occipital bone

27 **Clivus** with basilar part of occipital bone
28 **Hypoglossal canal**
29 Fossa for cerebellar hemisphere
30 Internal occipital protuberance
31 Fossa for cerebral hemisphere
32 Jugular tubercle
33 Condylar canal
34 Jugular process
35 **Foramen magnum**
36 Groove for transverse sinus
37 Groove for superior sagittal sinus
38 **Squamous part** of the occipital bone
39 External occipital protuberance
40 Superior nuchal line
41 Inferior nuchal line
42 Condylar fossa
43 **Condyle**
44 Pharyngeal tubercle
45 External occipital crest

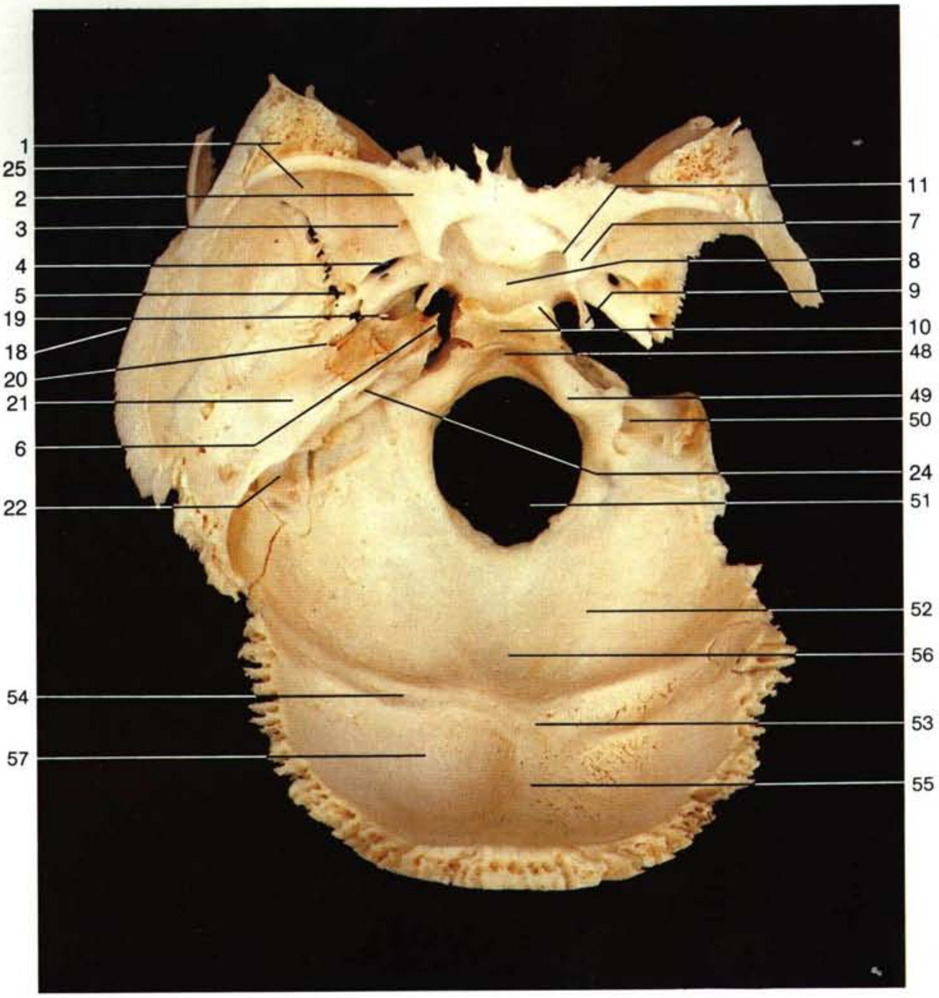

Sphenoid, occipital and left temporal bone (from above). Internal aspect of the base of the skull. The left temporal bone has been added to the preceeding figure.

Sphenoid bone

1	**Greater wing**
2	**Lesser wing**
3	Foramen rotundum
4	**Foramen ovale**
5	Foramen spinosum
6	Foramen lacerum
7	Anterior clinoid process
8	Hypophysial fossa (**sella turcica**)
9	Lingula
10	Dorsum sellae and posterior clinoid process
11	**Optic canal**
12	Sphenoid rostrum
13	Medial pterygoid plate
14	Lateral pterygoid plate
15	Pterygoid hamulus
16	Infratemporal crest
17	Body of the sphenoid

Temporal bone

18	**Squamous part**
19	**Carotid canal**
20	Hiatus of facial canal (for the greater petrosal nerve)
21	Arcuate eminence
22	Groove for the sigmoid sinus
23	Mastoid foramen
24	**Internal acoustic meatus**
25	Zygomatic process
26	Mandibular fossa
27	Petrotympanic fissure
28	Canalis musculotubarius (bony part of auditory tube)
29	**External acoustic meatus**
30	Styloid process (remnant only)
31	Stylomastoid foramen
32	Mastoid canaliculus
33	Jugular fossa
34	**Mastoid process**
35	Mastoid notch

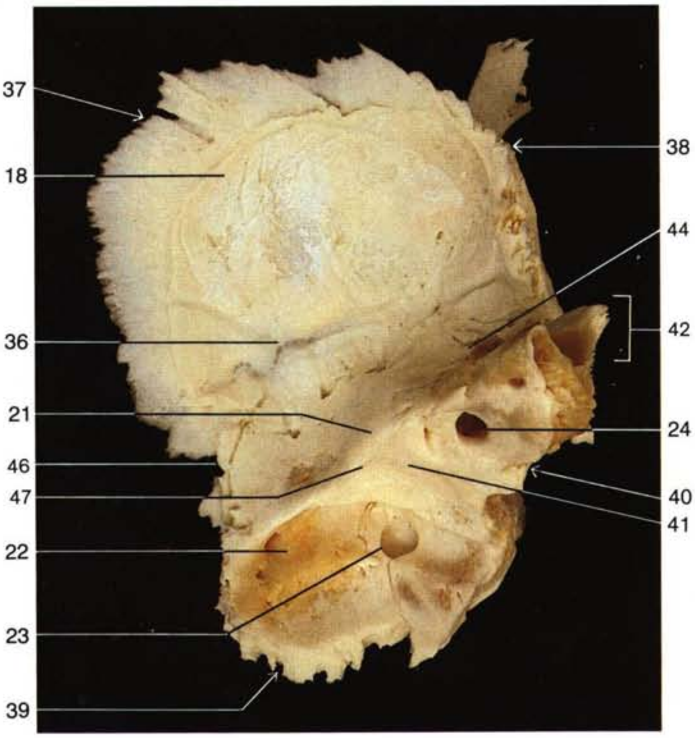

Left temporal bone (medial aspect).

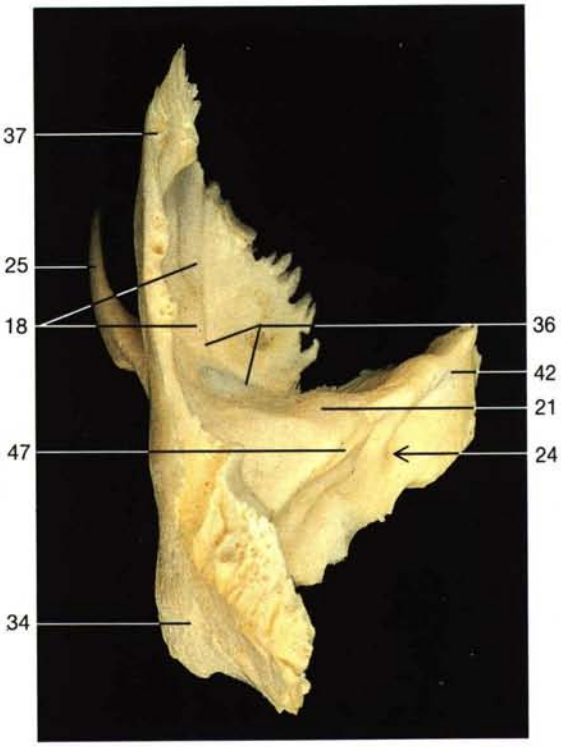

Left temporal bone (from above).

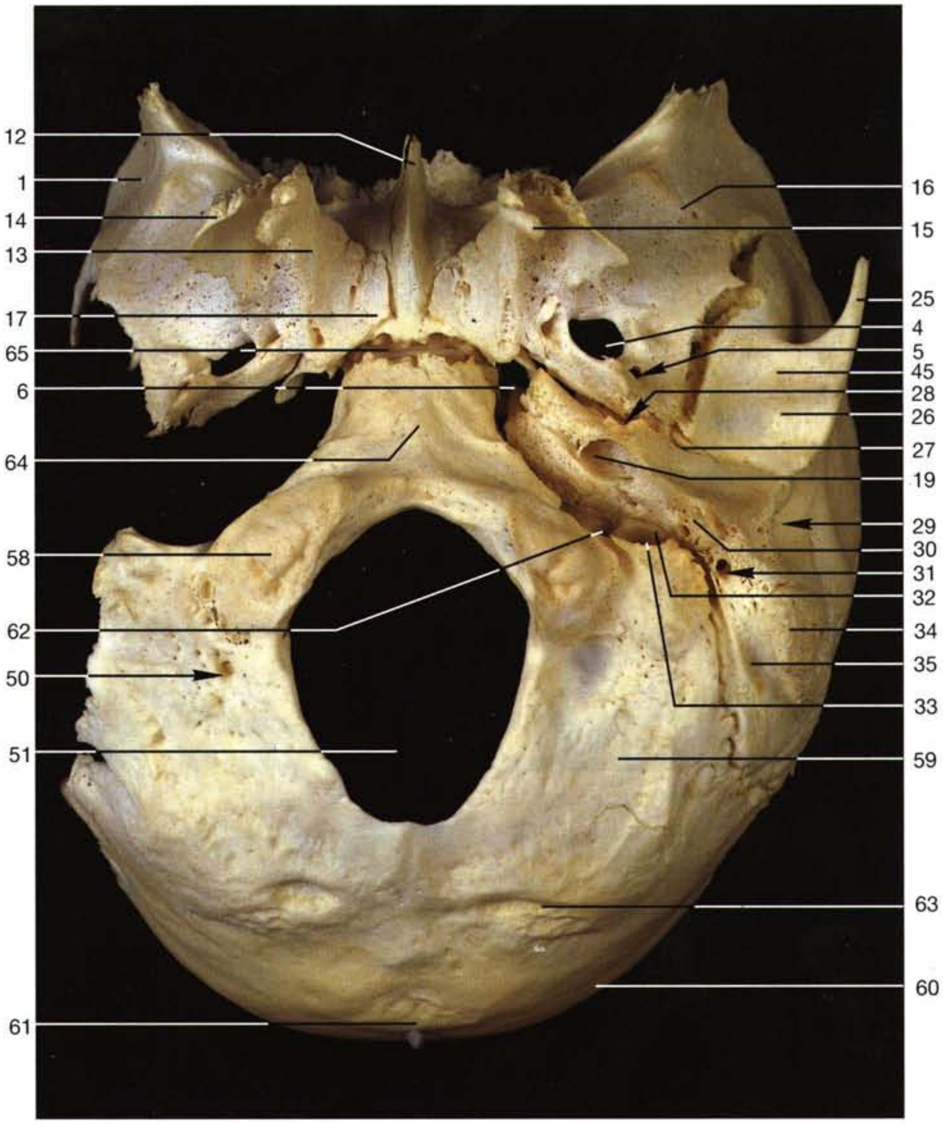
12
1
14
13
17
65
6
64
58
62
50
51
61
16
15
25
4
5
45
28
26
27
19
29
30
31
32
34
35
33
59
63
60

Sphenoid, occipital and left temporal bone.
Base of the skull (external aspect).

36 Groove for middle meningeal vessels
37 Parietal margin
38 Sphenoid margin
39 Occipital margin
40 Cochlear canaliculus
41 Aqueduct of the vestibule
42 Apex of the petrous part
43 **Tympanic part**
44 Trigeminal impression
45 **Articular tubercle**
46 Parietal notch
47 Groove for the superior petrosal sinus

Occipital bone
48 **Clivus**
49 Jugular tubercle
50 Condylar canal
51 **Foramen magnum**
52 Lower part of squamous occipital bone (cerebellar fossa)
53 Internal occipital protuberance
54 Groove for the transverse sinus
55 Groove for the superior sagittal sinus
56 Internal occipital crest
57 Upper part of squamous occipital bone (cerebral fossa)
58 **Condyle**
59 Nuchal plane
60 Superior nuchal line
61 **External occipital protuberance**
62 **Jugular foramen**
63 Inferior nuchal line
64 Pharyngeal tubercle
65 **Sphenooccipital synchondrosis**

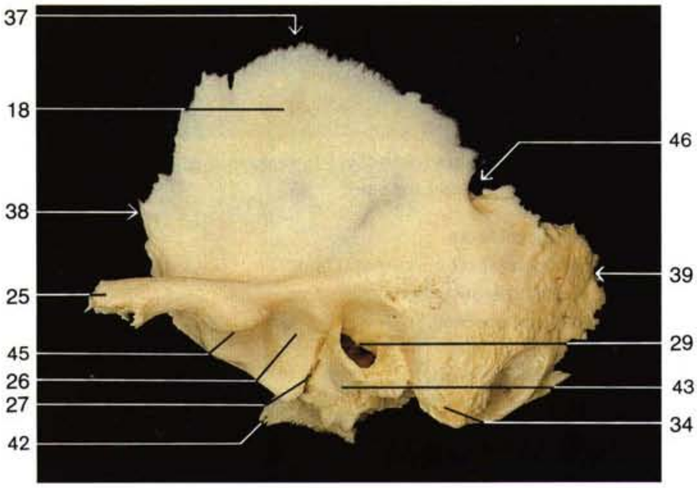
37
18
38
25
45
26
27
42
46
39
29
43
34

Left temporal bone (lateral aspect).

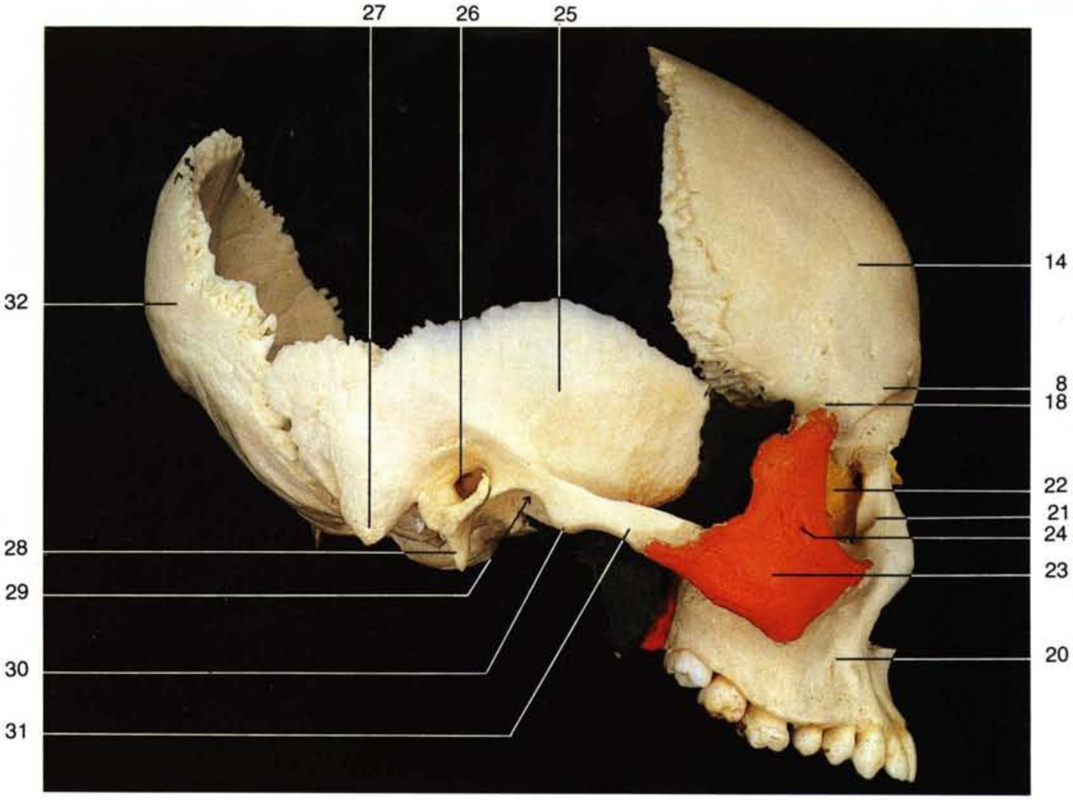

Part of a disarticulated skull (right lateral aspect). The **frontal bone** and the maxilla are connected with the temporal bone by the zygomatic bone (orange). Sphenoid bone (green), palatine bone (red), lacrimal bone (yellow).

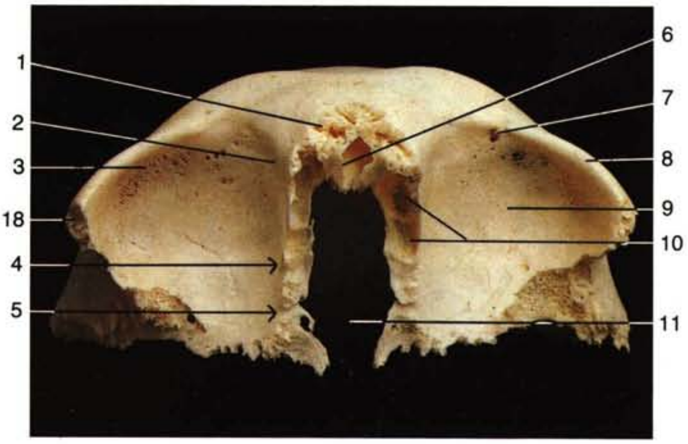

Frontal bone (inferior aspect). The ethmoidal foveolae cover the ethmoidal cavities of the ethmoid bone.

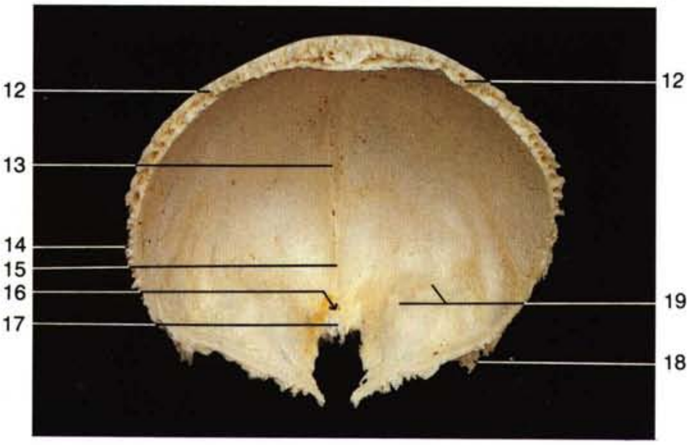

Frontal bone (posterior aspect).

Frontal bone

1 Nasal margin
2 Trochlear fossa
3 Fossa for lacrimal gland
4 Anterior ethmoidal foramen
5 Posterior ethmoidal foramen
6 Nasal spine
7 Supraorbital notch
8 Supraorbital margin
9 **Orbital plate**
10 Roofs of the ethmoidal air cells
11 Ethmoidal notch
12 Parietal margin
13 Groove for superior sagittal sinus
14 **Squamous part** of frontal bone
15 Frontal crest
16 Foramen cecum
17 Nasal spine
18 **Zygomatic process** of frontal bone
19 Juga cerebralia

20 **Maxilla**
21 Frontal process of maxilla
22 **Lacrimal bone** (yellow)
23 **Zygomatic bone** (orange)
24 Zygomaticofacial foramen

Temporal bone

25 Squamous part of temporal bone
26 External acoustic meatus
27 **Mastoid process**
28 **Styloid process**
29 **Mandibular fossa**
30 Articular tubercle
31 Zygomatic process

Occipital bone

32 Squamous part of occipital bone

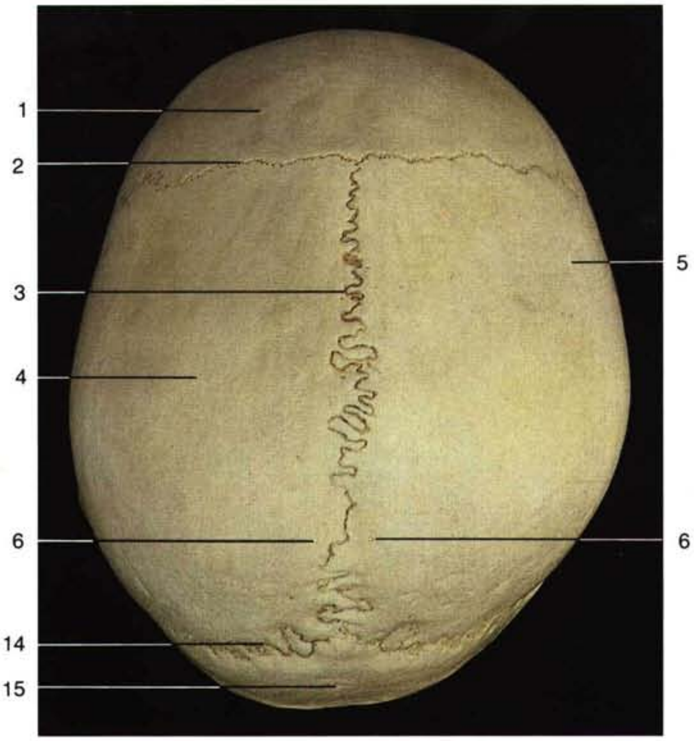

Calvaria (superior aspect).

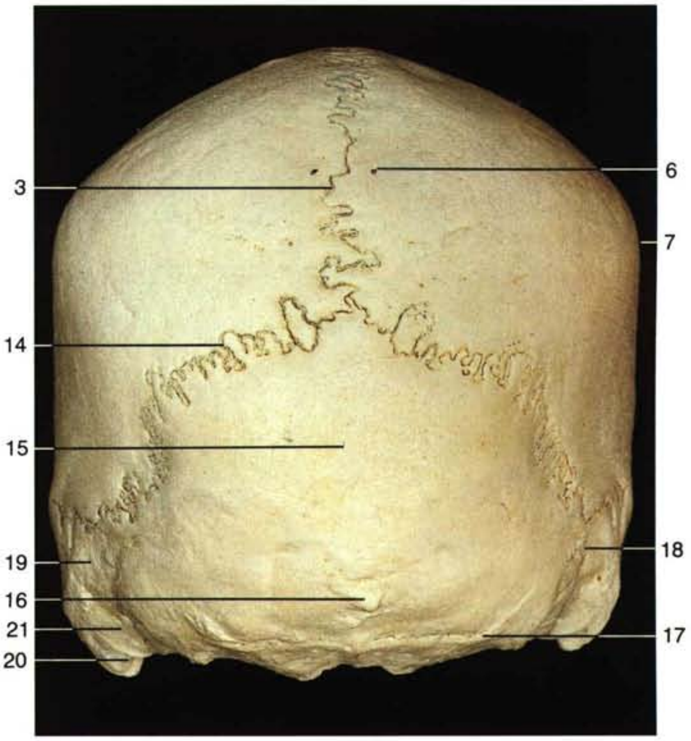

Calvaria (posterior aspect).

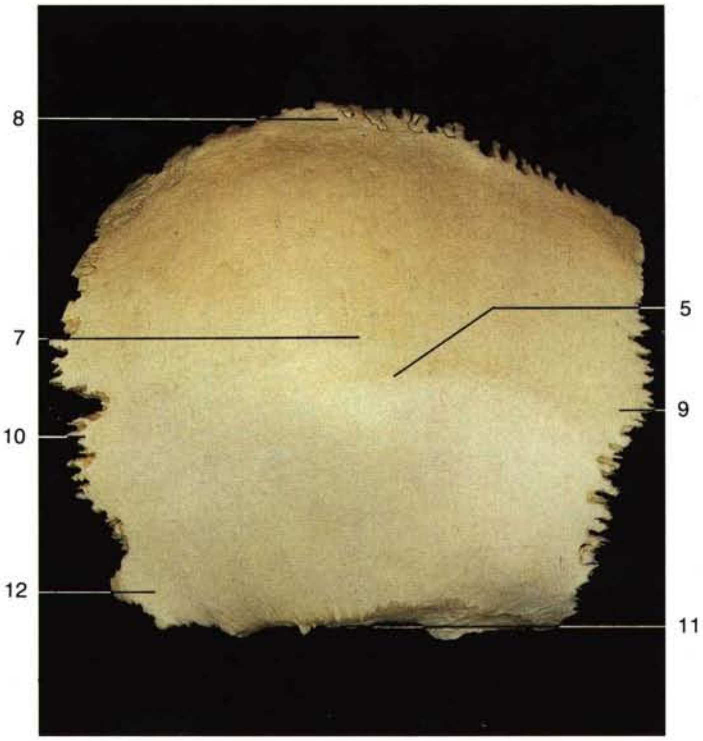

Left parietal bone (external aspect).

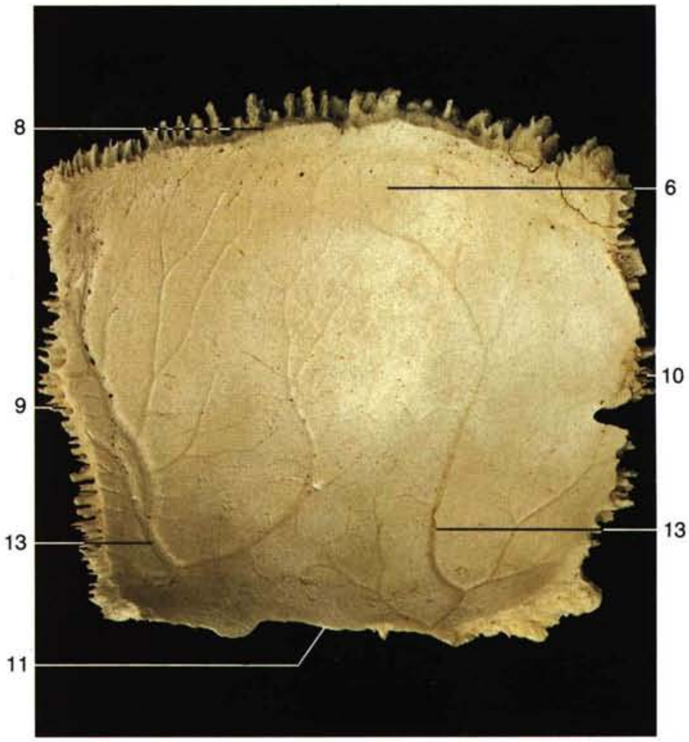

Left parietal bone (internal aspect).

1 **Frontal bone**
2 Coronal suture
3 Sagittal suture
4 **Parietal bone**
5 Superior temporal line
6 Parietal foramen
7 Parietal tuber or eminence
8 Sagittal margin
9 Occipital margin
10 Frontal margin
11 Squamous margin
12 Sphenoidal angle
13 Groove for middle meningeal artery
14 Lambdoid suture
15 **Occipital bone**
16 External occipital protuberance
17 Inferior nuchal line
18 Occipitomastoid suture
19 **Temporal bone**
20 Mastoid process
21 Mastoid notch

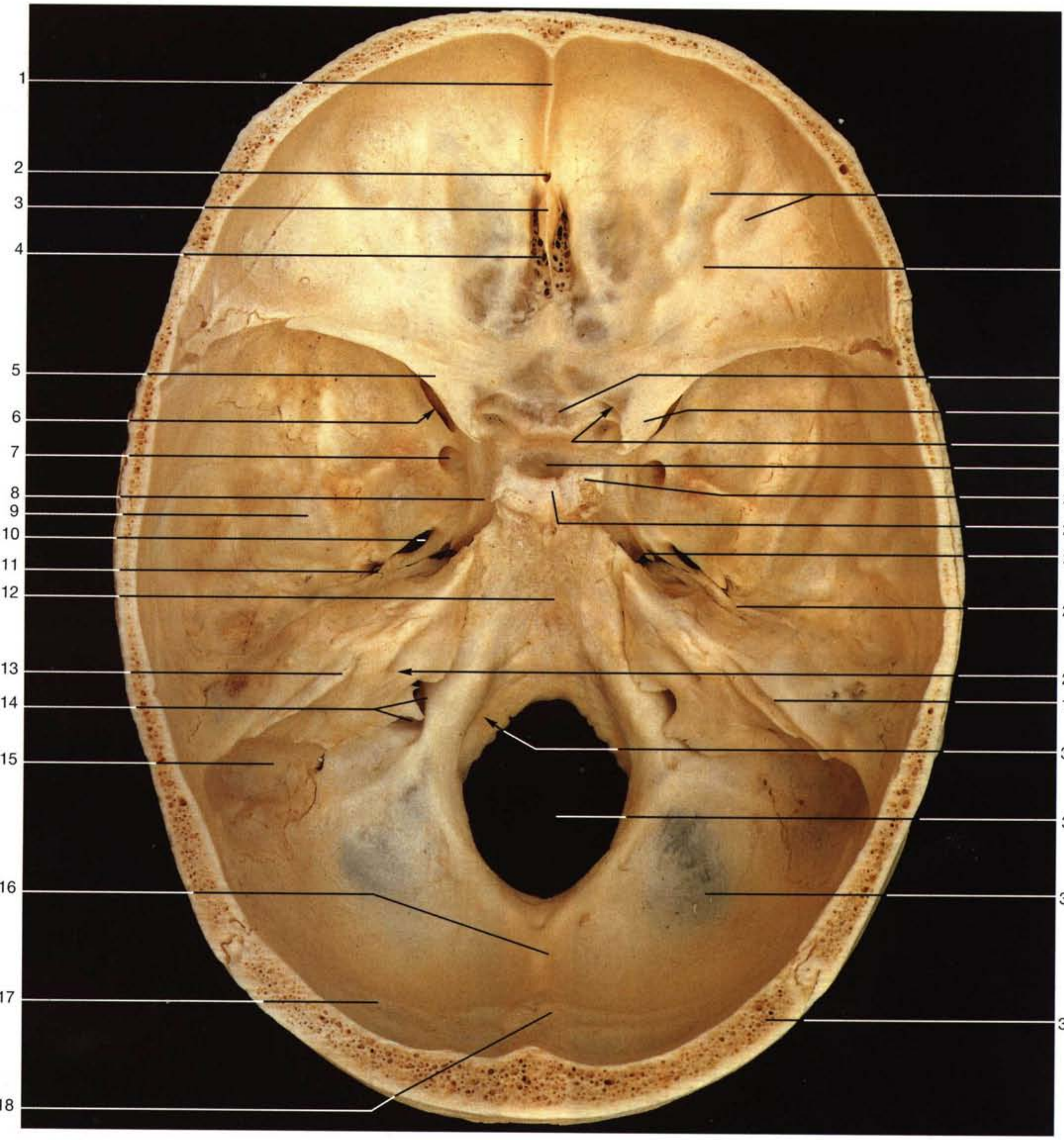

Base of the skull, calvaria removed (internal aspect).

1	Frontal crest	12	**Clivus**	24	**Sella turcica** (hypophysial fossa)
2	Foramen cecum	13	Groove for superior petrosal sinus	25	Posterior clinoid process
3	**Crista galli**	14	**Jugular foramen**	26	Dorsum sellae
4	**Cribriform plate** of ethmoid bone	15	Groove for sigmoid sinus	27	**Foramen lacerum**
5	Lesser wing of sphenoid bone	16	Internal occipital crest	28	Groove for greater petrosal nerve
6	Superior orbital fissure	17	Groove for transverse sinus	29	Internal acoustic meatus
7	**Foramen rotundum**	18	Internal occipital protuberance	30	**Hypoglossal canal**
8	**Carotid sulcus**	19	Marking of brain convolutions	31	**Foramen magnum**
9	Middle cranial fossa	20	Anterior cranial fossa	32	Posterior cranial fossa
10	**Foramen ovale**	21	Chiasmatic sulcus	33	Diploe
11	Foramen spinosum	22	Anterior clinoid process		
		23	**Optic canal**		

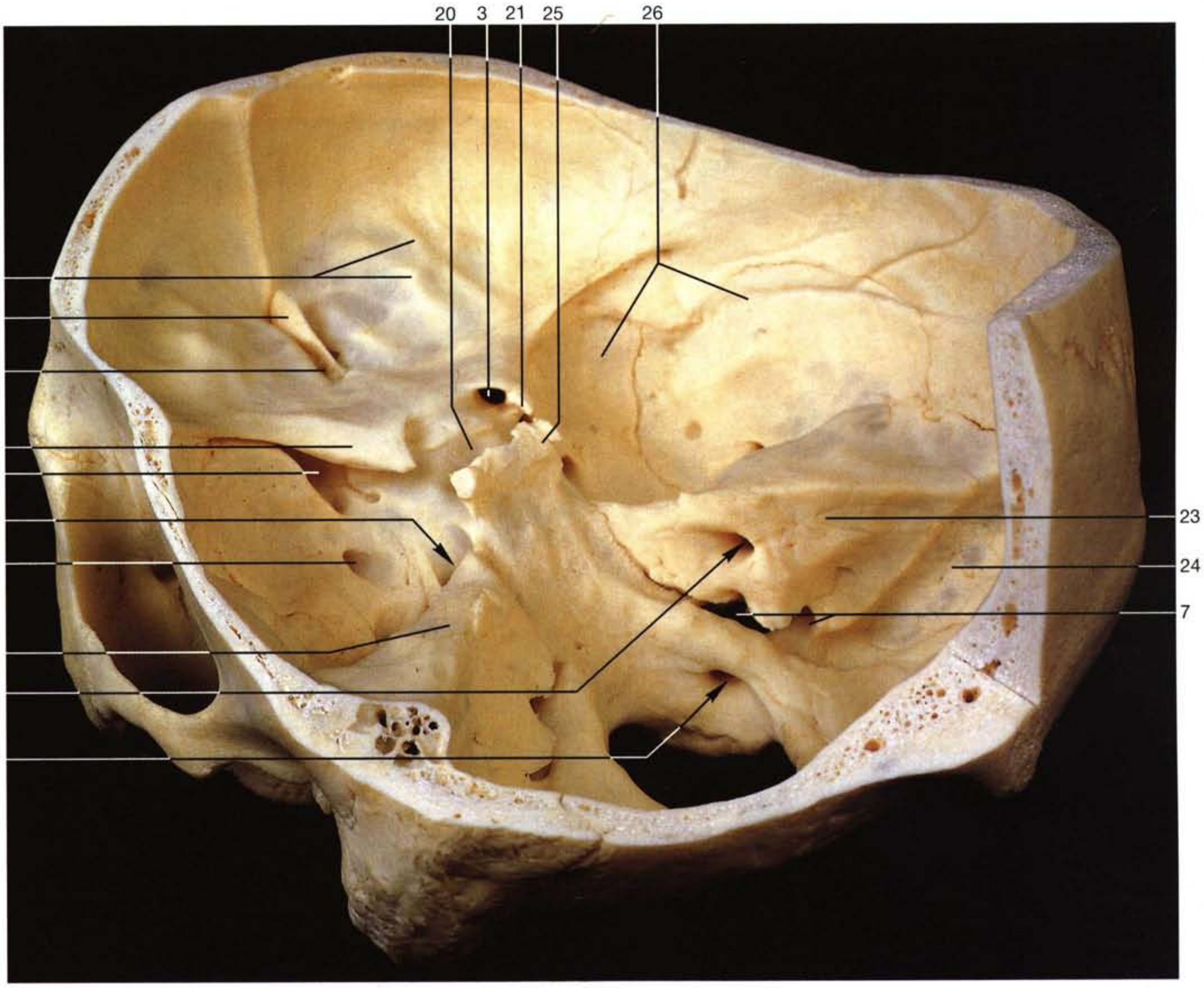

Base of the skull (internal aspect, oblique lateral view from left side).

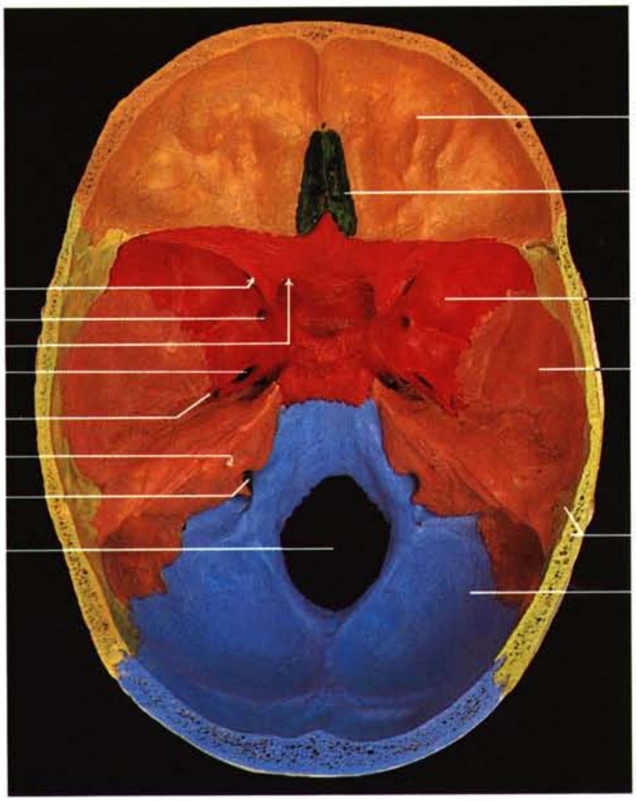

Base of the skull (internal aspect, superior view).
Individual bones indicated by color.

Canals, fissures and foramina of the base of the skull

1 Superior orbital fissure
2 Foramen rotundum
3 Optic canal
4 Foramen ovale
5 Foramen spinosum
6 Internal acoustic meatus
7 Jugular foramen
8 Foramen magnum

Bones

9 Frontal bone (orange)
10 Ethmoid bone (dark green)
11 Sphenoid bone (red)
12 Temporal bone (brown)
13 Parietal bone (yellow-green)
14 Occipital bone (blue)

Details of bones

15 Crista galli
16 Cribriform plate
17 Marking of brain convolutions (frontal bone)
18 Lesser wing of sphenoid bone
19 Foramen lacerum
20 Hypophysial fossa (sella turcica)
21 Anterior clinoid process
22 Depression for trigeminal ganglion
23 Petrous part of temporal bone
24 Groove for sigmoid sinus
25 Dorsum sellae (posterior clinoid process)
26 Greater wing of sphenoid bone, groove for middle meningeal artery
27 Hypoglossal canal

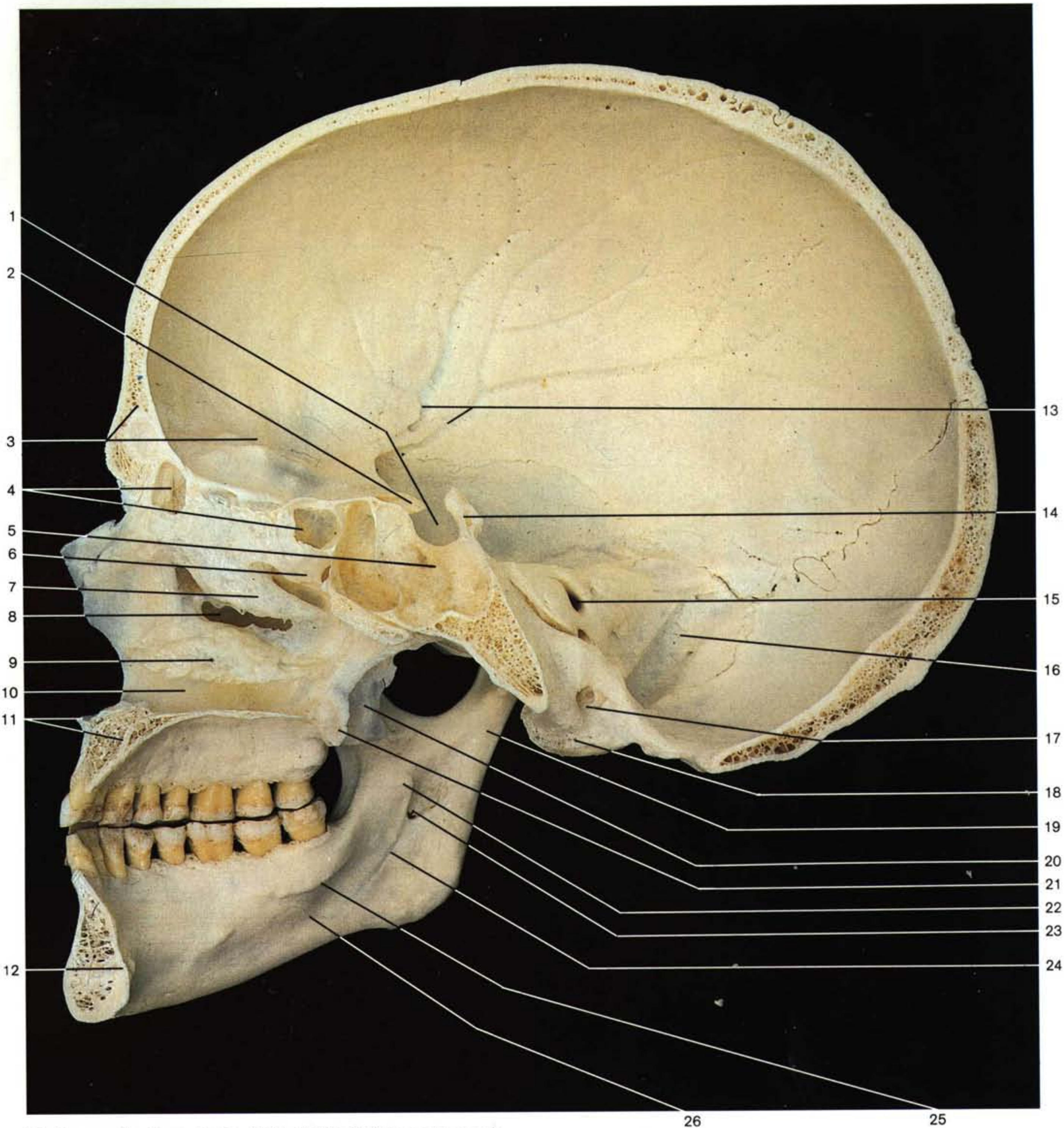

Median section through the skull, right half (internal aspect).

1 **Hypophysial fossa (sella turcica)**
2 Anterior clinoid process
3 **Frontal bone**
4 **Ethmoidal air cells**
5 **Sphenoidal sinus**
6 Superior concha
7 Middle concha
8 **Maxillary hiatus**
9 **Inferior concha**
10 Inferior meatus
11 Anterior nasal spine and maxilla
12 Mental spine or genial tubercle
13 Groove for middle meningeal artery
14 Dorsum sellae
15 Internal acoustic meatus
16 Groove for sigmoid sinus
17 **Hypoglossal canal**
18 **Occipital condyle**
19 Condylar process
20 Lateral pterygoid plate } of **pterygoid process**
21 Medial pterygoid plate } of **pterygoid process**
22 Lingula of mandible
23 **Mandibular foramen**
24 Mylohyoid groove
25 Mylohyoid line
26 Submandibular fovea

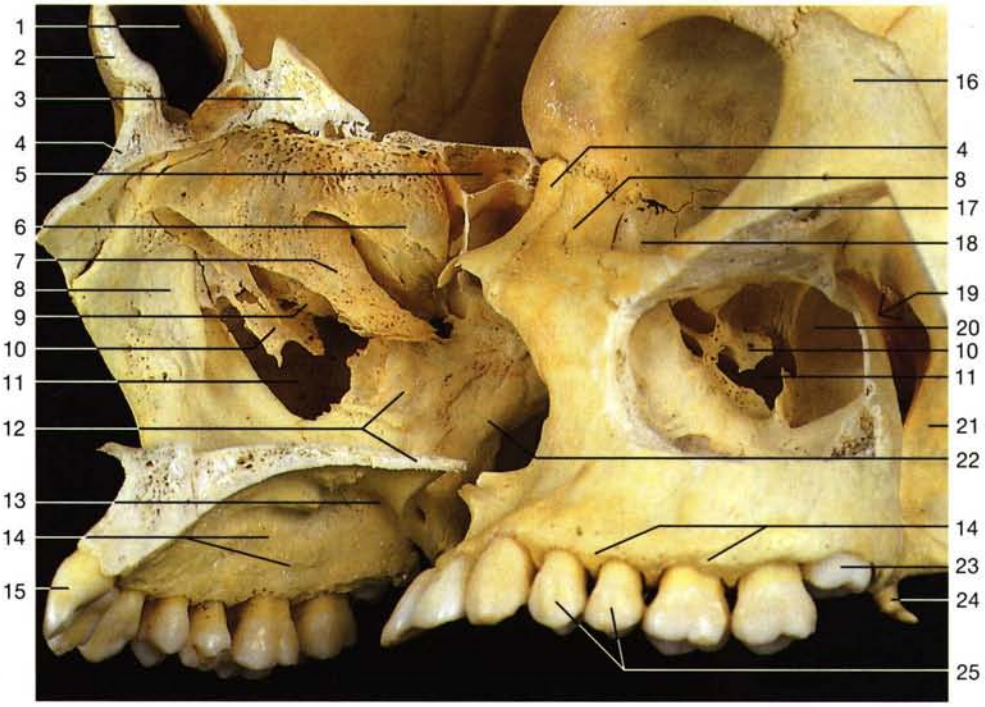

1 Frontal sinus
2 **Frontal bone**
3 Crista galli
4 **Nasal bone**
5 Sphenoidal sinus
6 **Superior concha** } of **ethmoid**
7 **Middle concha** } **bone**
8 Frontal process of maxilla
9 Ethmoidal bulla
10 **Uncinate process**
11 Maxillary hiatus
12 **Palatine bone**
13 Greater palatine foramen
14 **Alveolar process of maxilla**
15 Central incisor
16 Zygomatic bone
17 Ethmoid bone
18 **Lacrimal bone**
19 Pterygopalatine fossa
20 **Maxillary sinus**
21 Lateral pterygoid plate
22 Medial pterygoid plate
23 Third molar tooth
24 Pterygoid hamulus
25 Two premolar teeth

Facial part of the skull (viscerocranium), divided in two halves (lateral and medial aspect). Right inferior concha has been removed to show the maxillary hiatus. Left maxillary sinus opened.

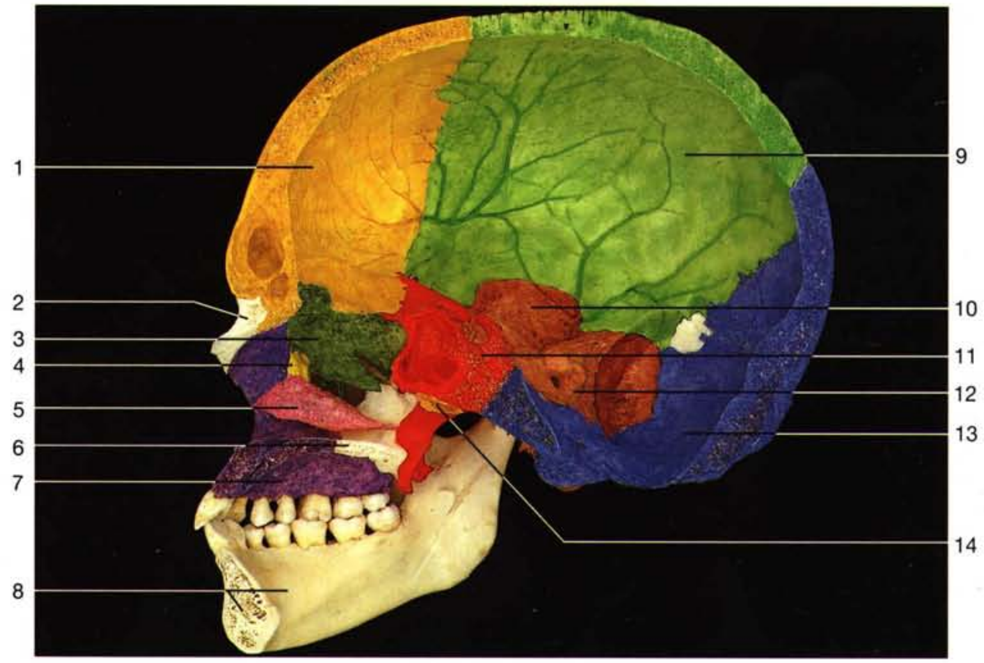

Bones (indicated by colors)
1 Frontal bone (yellow)
2 Nasal bone (white)
3 Ethmoid bone (dark green)
4 Lacrimal bone (yellow)
5 Inferior nasal concha (pink)
6 Palatine bone (white)
7 Maxilla (violet)
8 Mandible (white)
9 Parietal bone (light green)
10 Temporal bone (brown)
11 Sphenoid bone (red)
12 Petrous part of temporal bone (brown)
13 Occipital bone (blue)
14 Ala of vomer (light brown)

Median section through the skull. The nasal septum has been removed. Bones indicated by colors.

Because of the upright posture which the man developed in the course of evolution, the cranial cavity greatly increased in size whereas the facial skeleton decreased. As a result, the base of the skull developed an angulation of about 120° between the clivus and the cribriform plate. The hypophysial fossa containing the pituitary gland lies at the angle formed between these two planes.

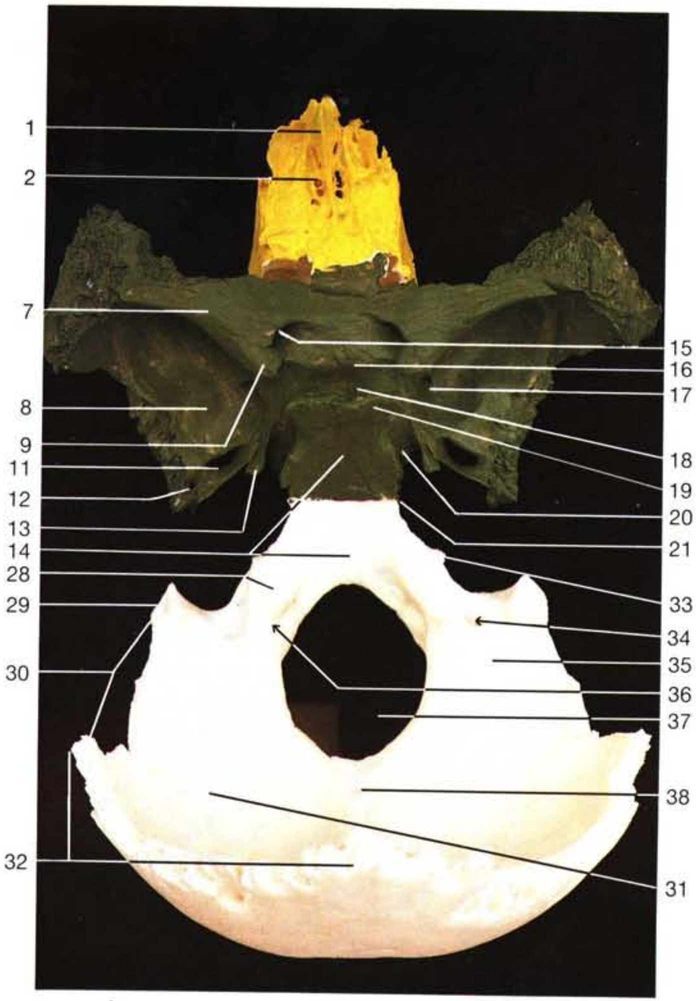

Part of the disarticulated base of the skull.
Ethmoid, sphenoid and occipital bones (from above).
Green = sphenoid bone; yellow = ethmoid bone.

Ethmoid bone

1 Crista galli
2 Cribriform plate
3 Ethmoidal air cells
4 Middle concha
5 Perpendicular plate (part of nasal septum)
6 Orbital plate

Sphenoid bone

7 Lesser wing
8 Greater wing
9 Anterior clinoid process
10 Posterior clinoid process
11 Foramen ovale
12 Foramen spinosum
13 Lingula of the sphenoid
14 Clivus
15 Optic canal
16 Tuberculum sellae
17 Foramen rotundum (right side)
18 Hypophysial fossa (sella turcica)
19 Dorsum sellae
20 Carotid sulcus
21 Sphenooccipital synchondrosis
22 Lateral pterygoid plate
23 Greater wing of sphenoid bone (orbital surface)
24 Greater wing of sphenoid bone (maxillary surface)
25 Foramen rotundum (left side)
26 Superior orbital fissure
27 Infratemporal crest of the greater wing

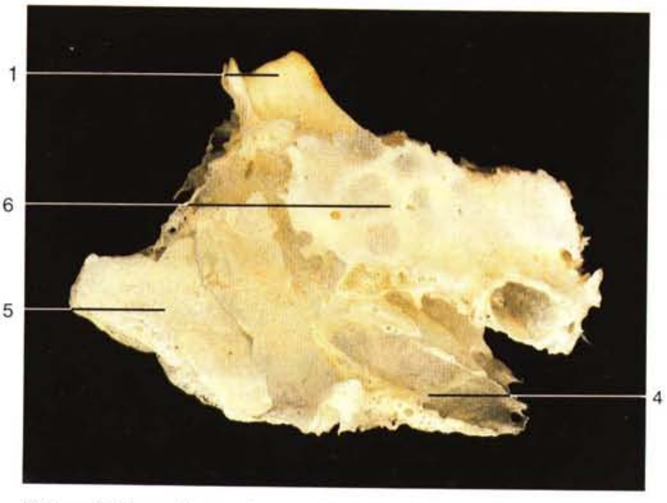

Ethmoid bone (lateral aspect), posterior portion to the right.

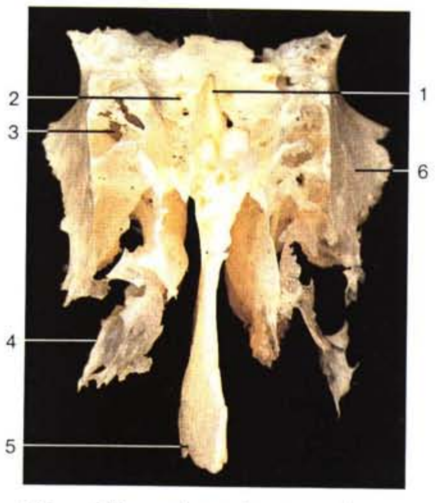

Ethmoid bone (anterior aspect).

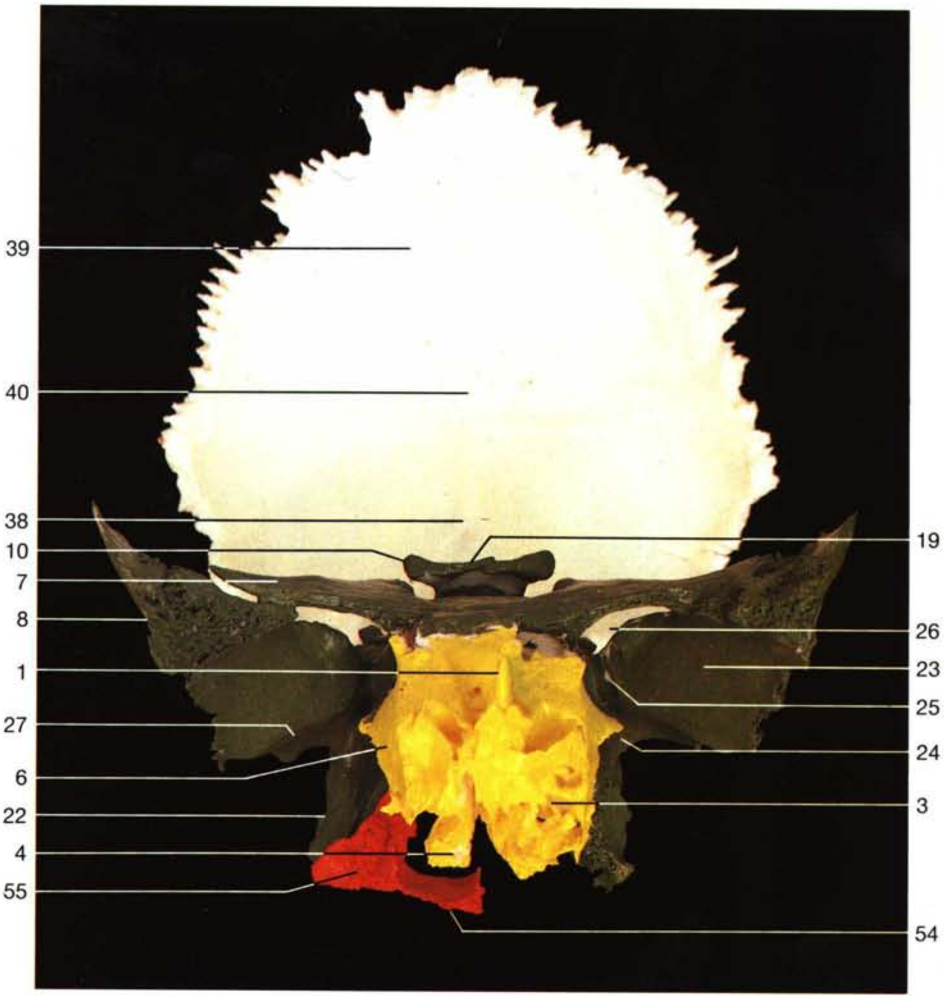

Disarticulated base of the skull (anterior aspect).
Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone.

Occipital bone
28 Jugular tubercle
29 Jugular process
30 Mastoid margin
31 Posterior cranial fossa
32 Lambdoid margin
33 Intrajugular process
34 Condylar canal
35 Lateral part of occipital bone
36 Hypoglossal canal
37 Foramen magnum
38 Internal occipital crest
39 Squamous part of occipital bone
40 Internal occipital protuberance

Maxilla
41 Orbital surface
42 Infraorbital groove
43 Maxillary tuberosity with foramina
44 Frontal process
45 Nasolacrimal groove
46 Infraorbital margin
47 Anterior nasal spine
48 Zygomatic process
49 Alveolar process

Palatine bone
50 Orbital process
51 Sphenopalatine notch
52 Sphenoidal process
53 Perpendicular plate
54 Horizontal plate
55 Pyramidal process

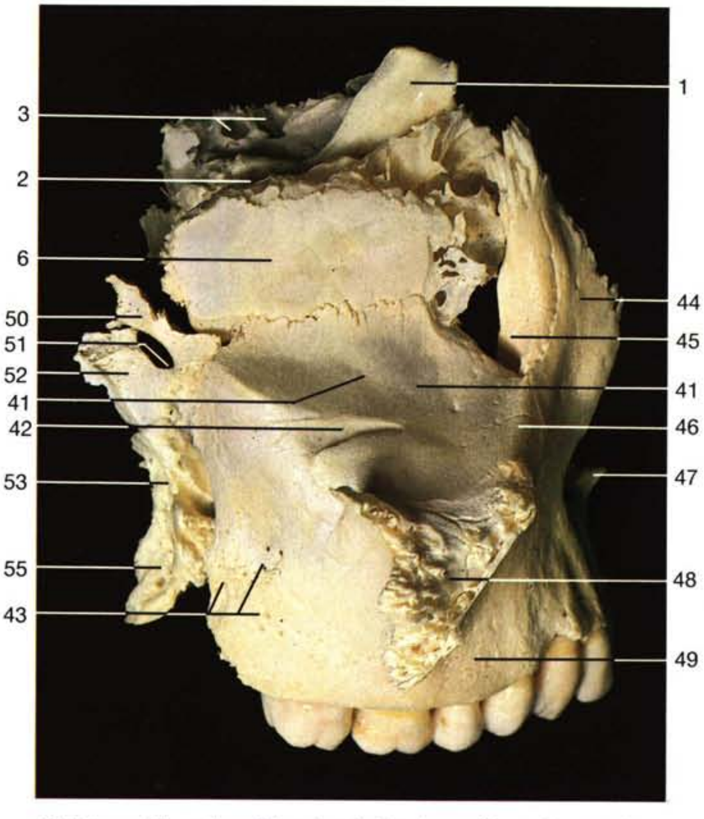

Right maxilla, ethmoid and palatine bone (lateral aspect).

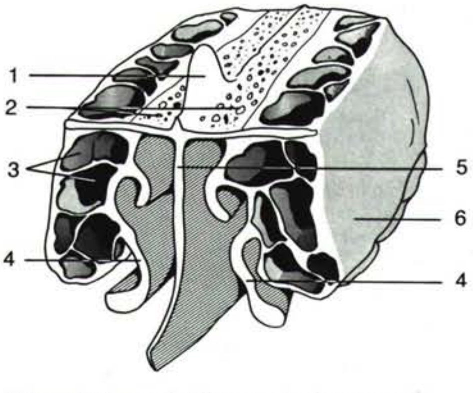

Ethmoid bone (oblique anterior aspect).
(Schematic drawing.)

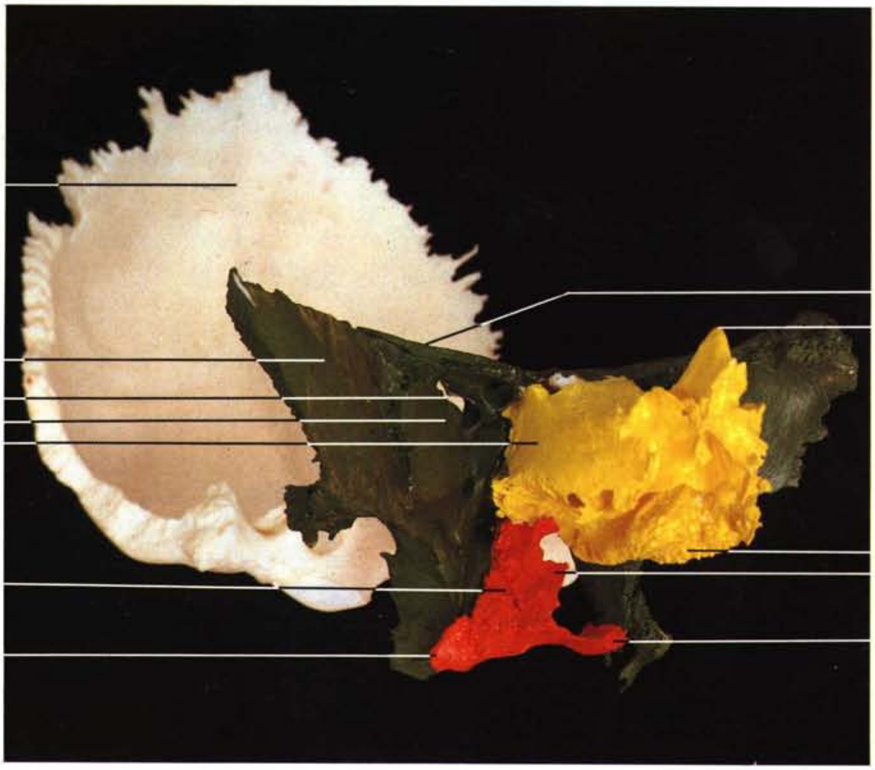

Part of a disarticulated skull base, similar to the foregoing figures, but with palatine bone. Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone.

Ethmoid bone

1 Crista galli
2 Orbital plate
3 Middle concha

Palatine bone

4 Horizontal plate of palatine bone
5 Greater palatine canal
6 Pyramidal process
7 Maxillary process
8 Orbital process
9 Sphenopalatine notch
10 Perpendicular plate of palatine bone
11 Conchal crest
12 Nasal crest
13 Sphenoidal process

Sphenoid bone

14 Greater wing
15 Superior orbital fissure
16 Greater wing (orbital surface)
17 Lesser wing

Occipital bone

18 Squamous part of occipital bone

Maxilla

19 Maxillary tuberosity
20 Frontal process
21 Orbital surface
22 Infraorbital margin
23 Infraorbital groove
24 Zygomatic process
25 Alveolar process

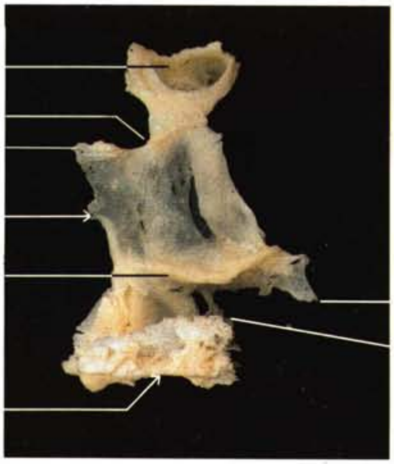

Left palatine bone (medial aspect, posterior aspect to the left).

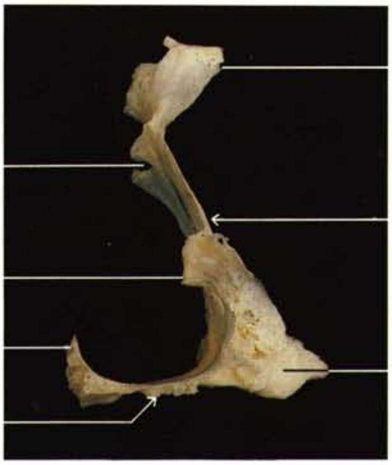

Left palatine bone (anterior aspect).

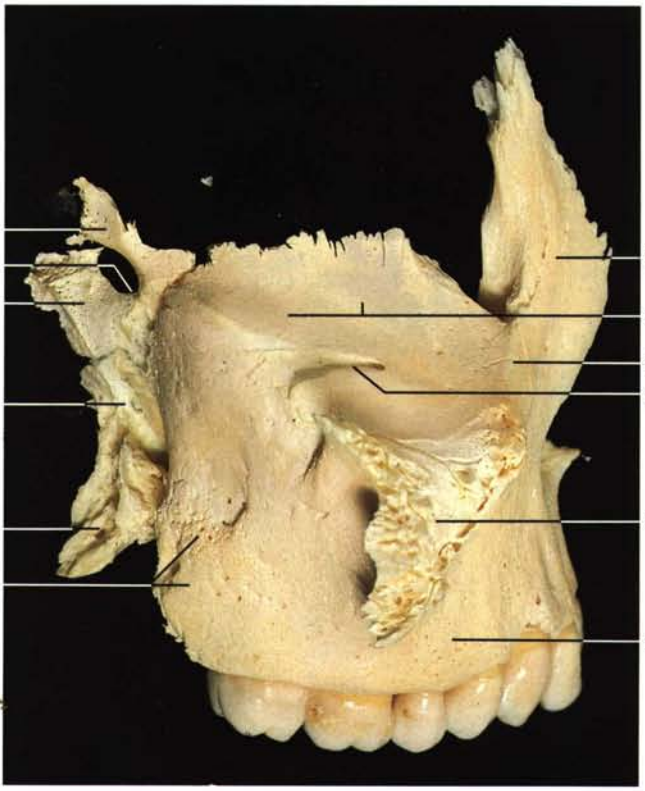

Right maxilla and right palatine bone (lateral aspect).

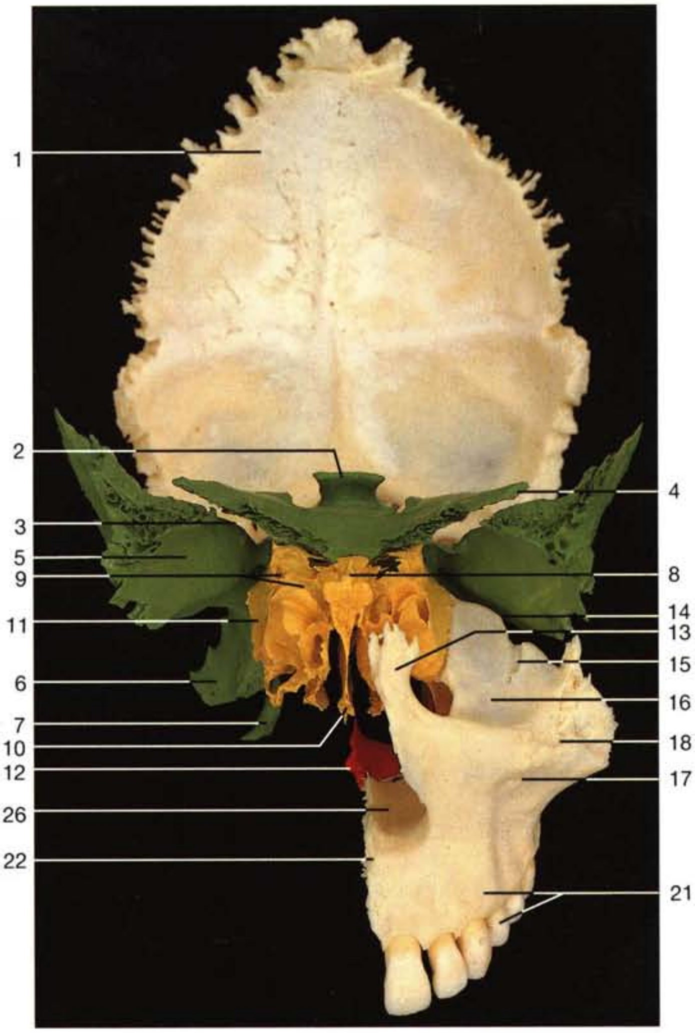

Part of a disarticulated skull.
The left **maxilla** is added to the preceding specimen.

Occipital bone
1 Squamous part

Sphenoid bone
2 Dorsum sellae
3 Superior orbital fissure
4 Lesser wing
5 Greater wing (orbital surface)
6 Lateral pterygoid plate
7 Medial pterygoid plate

Ethmoid bone
8 Crista galli
9 Ethmoidal air cells
10 Perpendicular plate
11 Orbital plate

Palatine bone
12 Horizontal plate (nasal crest)

Maxilla
13 Frontal process
14 Inferior orbital fissure
15 Infraorbital groove
16 Orbital surface
17 Infraorbital foramen
18 Zygomatic process
19 Anterior lacrimal crest
20 Canine fossa
21 Alveolar process with teeth
22 Anterior nasal spine
23 Juga alveolaria (elevations formed by roots of teeth)
24 Lacrimal groove
25 Maxillary tuberosity with alveolar foramina
26 Palatine process of maxilla

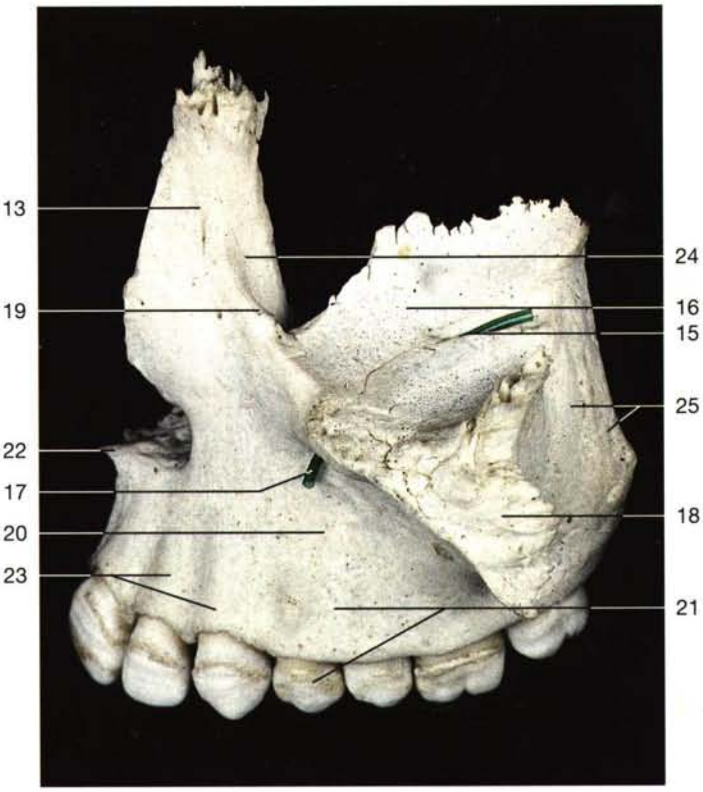

Left maxilla (lateral aspect). Probe = infraorbital canal.

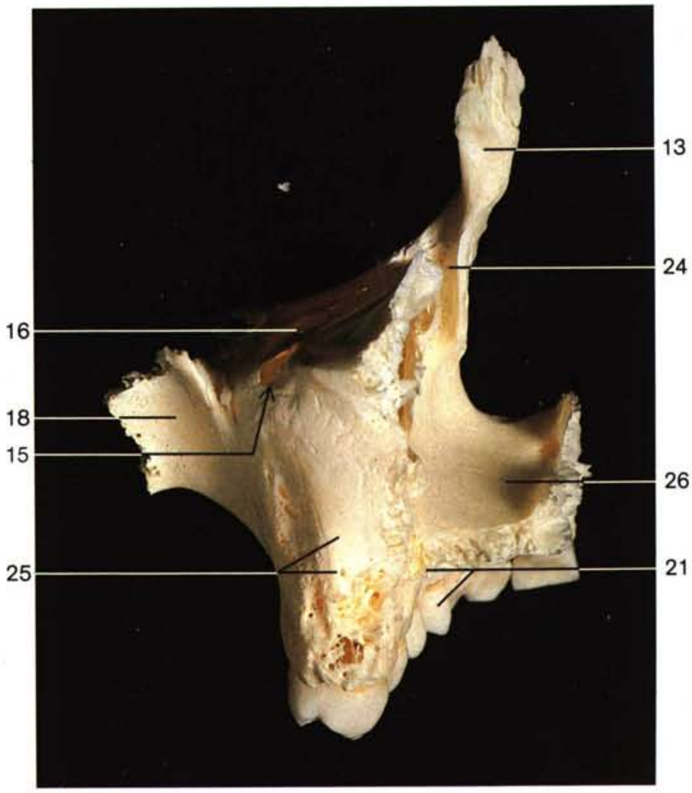

Left maxilla (posterior aspect).

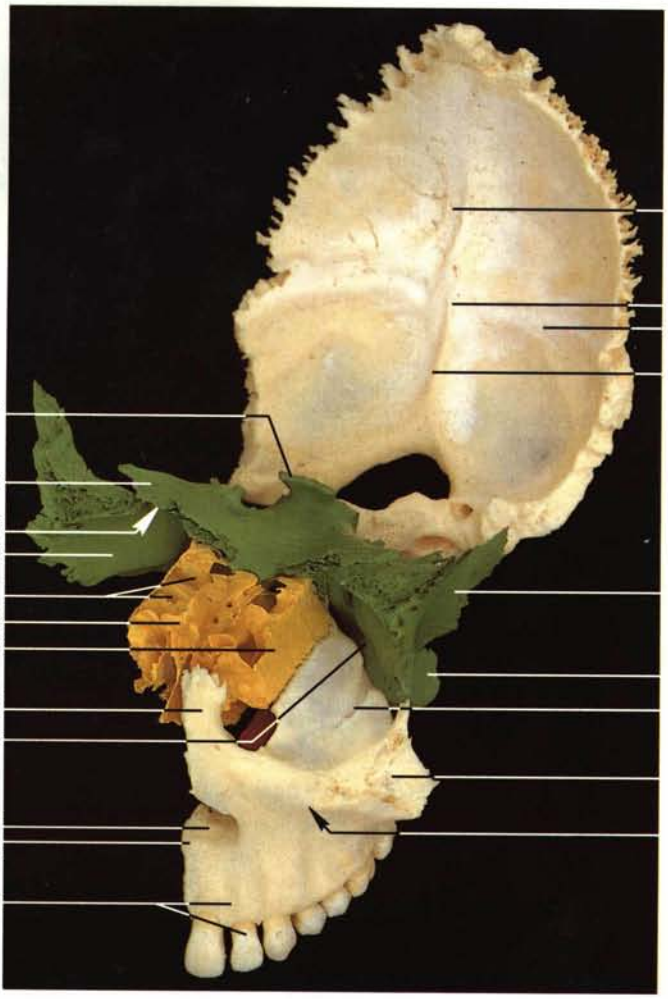

Part of a disarticulated base of skull. The mosaic of the facial bones [sphenoid bone (green), ethmoid bone (yellow), and palatine bone (red)] is seen from the anterior-lateral aspect.

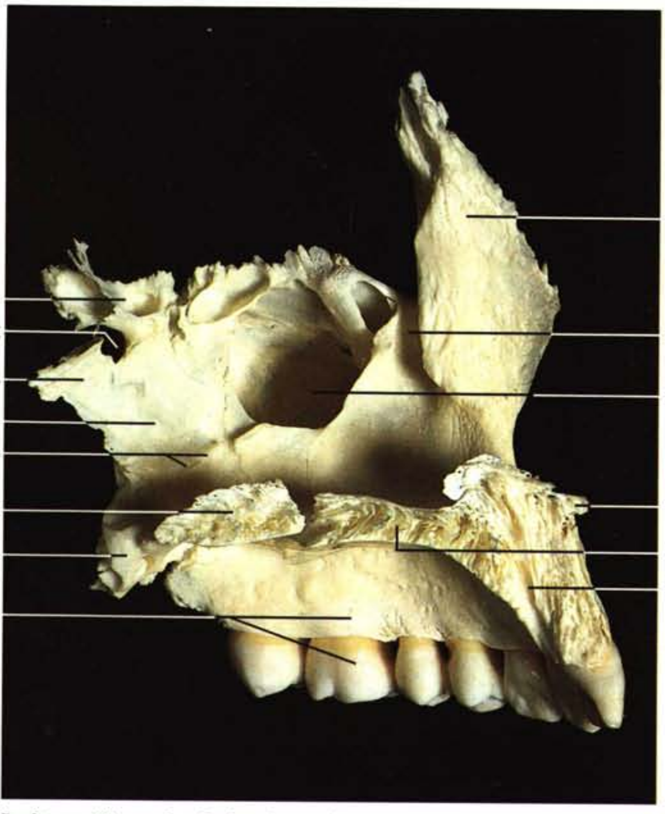

Left maxilla and palatine bone (medial aspect).

Occipital bone

1 Groove for superior sagittal sinus
2 Internal occipital protuberance
3 Groove for transverse sinus
4 Internal occipital crest

Sphenoid bone

5 Greater wing (temporal surface)
6 Lateral pterygoid plate
7 Dorsum sellae
8 Lesser wing
9 Superior orbital fissure
10 Greater wing (orbital surface)

Ethmoid bone

11 Ethmoidal air cells
12 Crista galli
13 Orbital plate

Maxilla

14 Frontal process
15 Inferior orbital fissure
16 Alveolar process with teeth
17 Palatine process
18 Anterior nasal spine
19 Infraorbital groove
20 Zygomatic process
21 Location of infraorbital foramen
22 Middle nasal meatus
23 Inferior nasal meatus
24 Maxillary hiatus (leading to maxillary sinus)
25 Third molar
26 Lacrimal groove
27 Conchal crest
28 Body of maxilla (nasal surface)
29 Nasal crest
30 Incisive canal

Palatine bone

31 Orbital process
32 Sphenopalatine notch
33 Sphenoidal process
34 Perpendicular plate
35 Conchal crest
36 Horizontal plate
37 Pyramidal process

Frontal bone

38 Squamous part
39 Supraorbital foramen
40 Frontal notch
41 Frontal spine

Inferior nasal concha

42 Inferior nasal concha with maxillary process

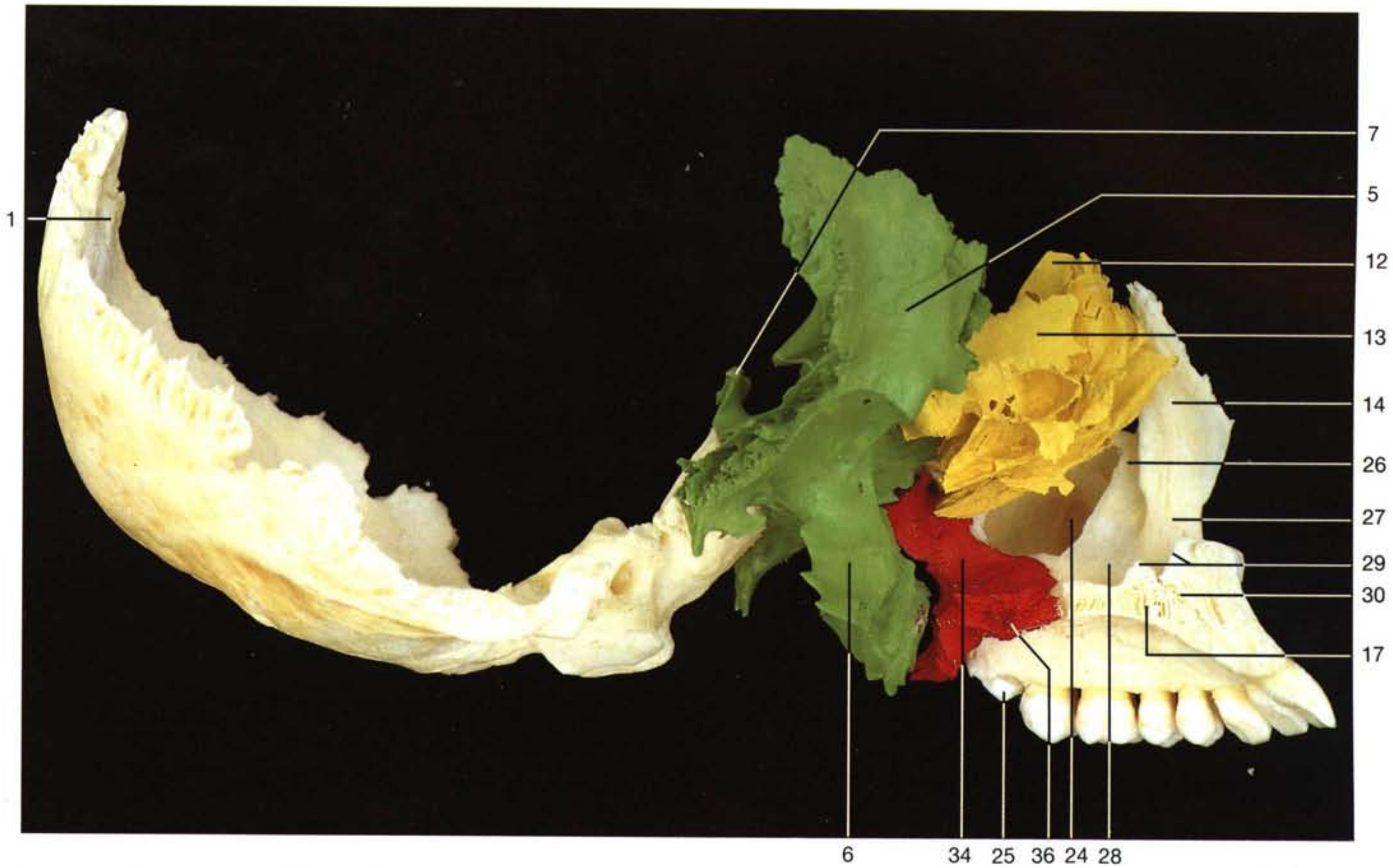
7
5
12
13
14
26
27
29
30
17
1
6
34
25
36
24
28

Part of a disarticulated base of skull (medial aspect). Green = sphenoid bone; yellow = ethmoid bone; red = palatine bone; natural colored = left maxilla.

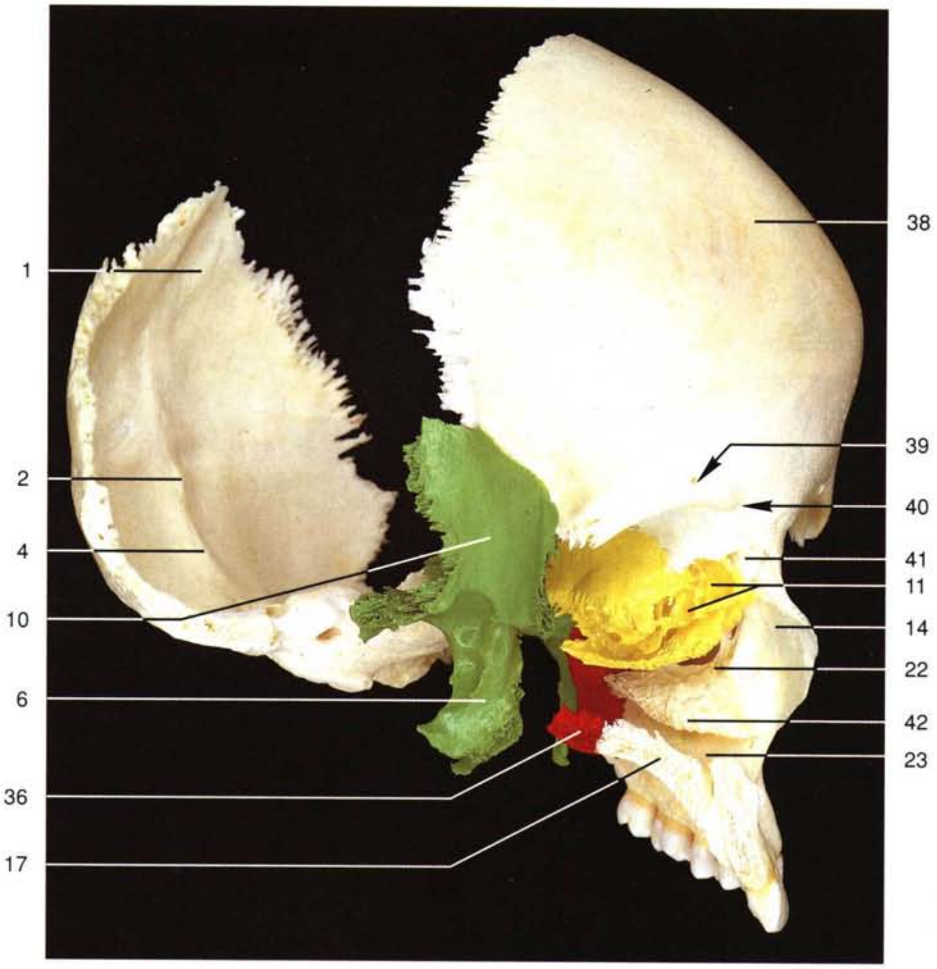
38
1
39
40
2
41
4
11
10
14
22
6
42
23
36
17

Part of a disarticulated base of skull. The same specimen as shown above but with frontal bone (oblique-lateral aspect).

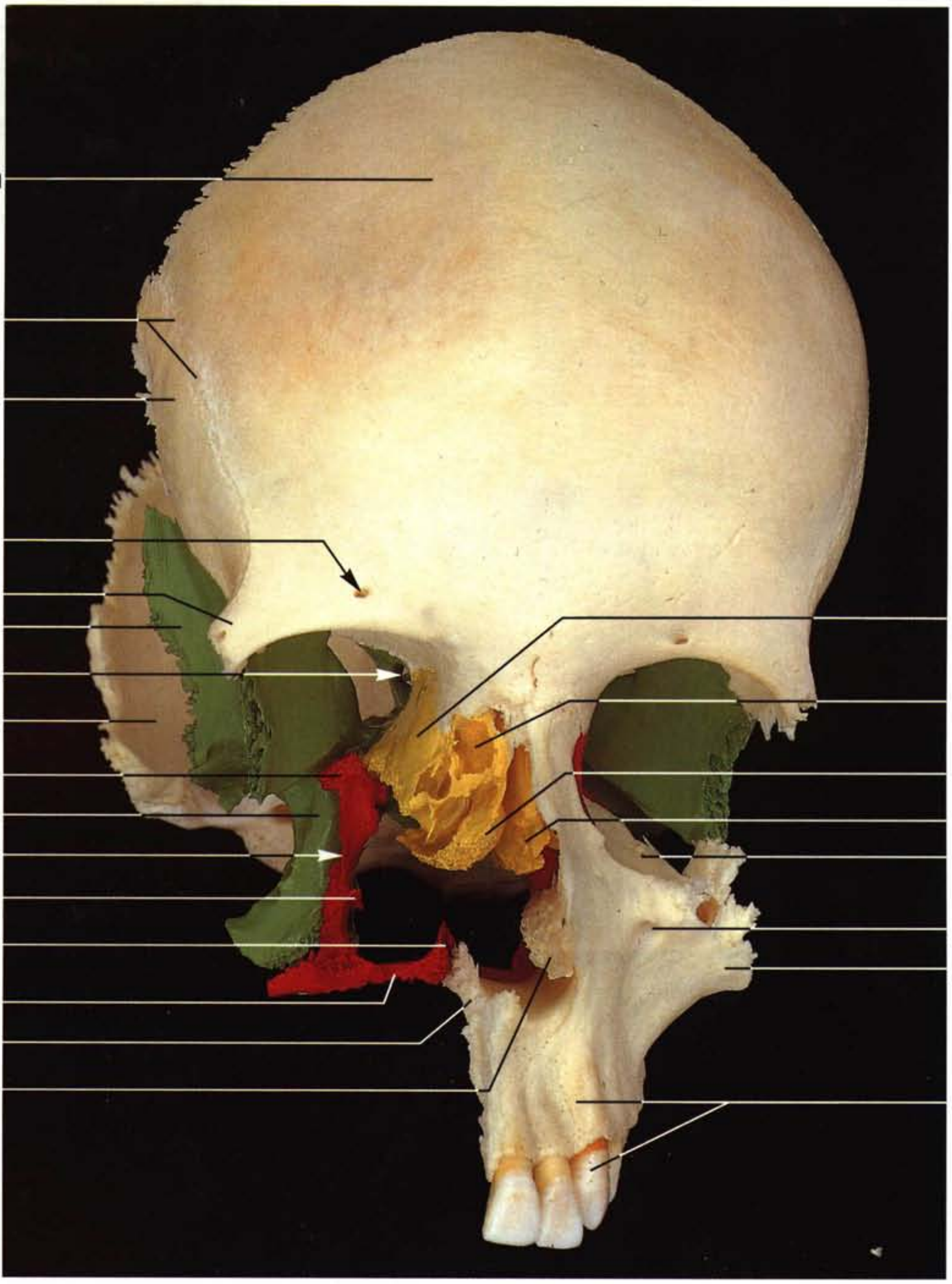

Part of a disarticulated skull, showing the connection of the palatine bone (red) and the maxilla with ethmoid bone (yellow) and sphenoid bone (light green) (anterior aspect).

Frontal bone

1 Squamous part
2 Inferior temporal line
3 Temporal surface
4 Supraorbital foramen
5 Zygomatic process

Occipital bone

6 Squamous part

Sphenoid bone

7 Greater wing (temporal surface)
8 Optic canal within the lesser wing
9 Lateral pterygoid plate

Palatine bone

10 Orbital process
11 Perpendicular plate
12 Conchal crest
13 Nasal crest
14 Horizontal plate

Ethmoid bone

15 Orbital plate
16 Ethmoidal air cell
17 Middle concha
18 Perpendicular plate (part of bony nasal septum)

Maxilla

19 Infraorbital groove
20 Infraorbital foramen
21 Zygomatic process
22 Alveolar process with teeth
23 Palatine process

Left inferior nasal concha

24 Anterior part of inferior concha

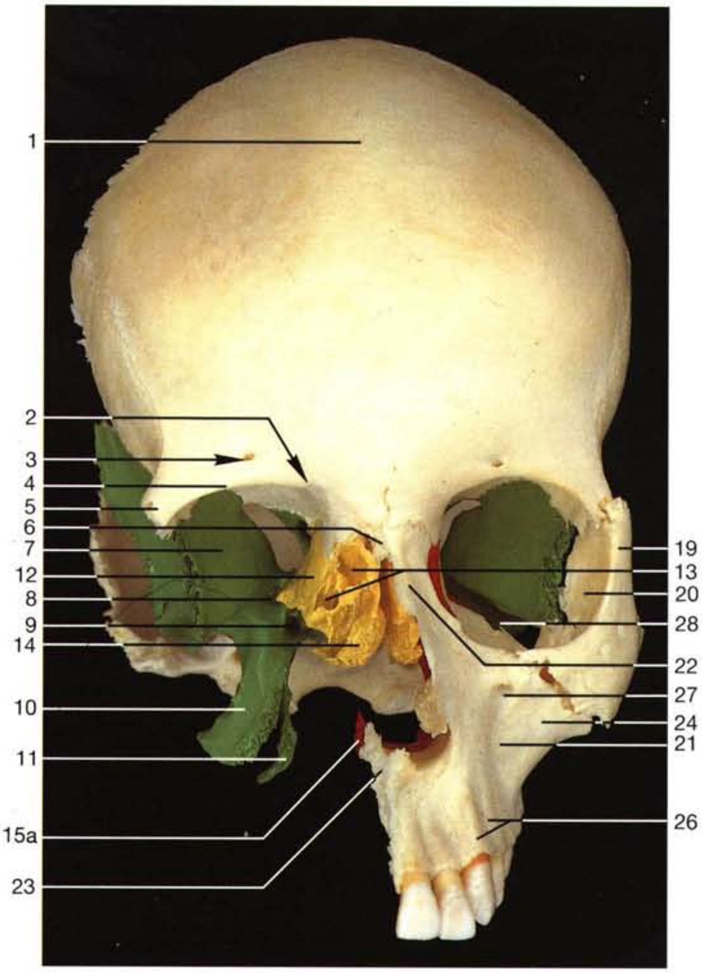

Anterior view of a disarticulated skull, showing the connection of the maxilla with the frontal and zygomatic bones. Yellow = ethmoid bone; red = palatine bone; green = sphenoid bone.

Frontal bone

1 Squamous part
2 Frontal notch
3 Supraorbital foramen
4 Supraorbital margin
5 Zygomatic process
6 Frontal spine

Sphenoid bone

7 Greater wing (orbital surface)
8 Foramen rotundum
9 Pterygoid or Vidian canal
10 Lateral pterygoid plate
11 Medial pterygoid plate

Ethmoid bone

12 Orbital plate
13 Ethmoidal air cells
14 Middle concha

Palatine bone

15 Horizontal plate
15a Nasal crest
16 Pyramidal process
17 Lesser palatine foramen
18 Greater palatine foramen

Zygomatic bone

19 Frontal process
20 Orbital surface

Maxilla

21 Canine fossa
22 Frontal process
23 Palatine process
24 Zygomatic process
25 Alveolar process and teeth
26 Juga alveolaria
27 Infraorbital foramen
28 Infraorbital groove
29 Anterior nasal aperture
30 Anterior nasal spine

Incisive bone

31 Central incisor and incisive bone or premaxilla
32 Incisive fossa

Vomer

33 Ala of the vomer

Sutures and choanae

34 Median palatine suture
35 Transverse palatine suture
36 Choanae

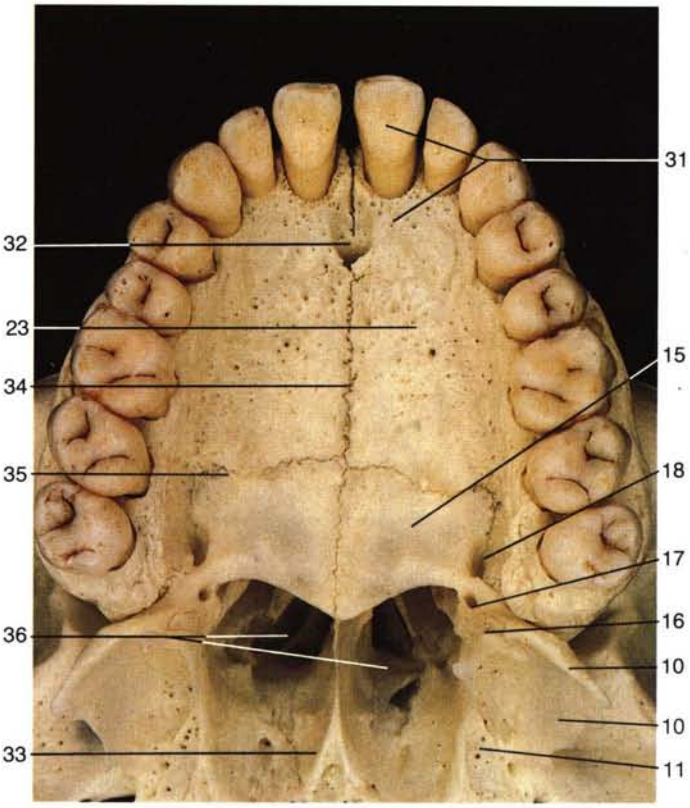

Bony palate and teeth of the maxillae (from below).

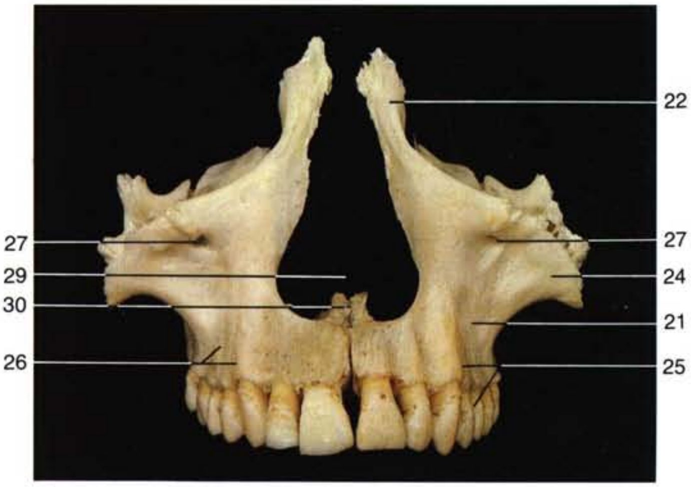

Anterior view of both maxillae, forming the anterior bony aperture of the nose.

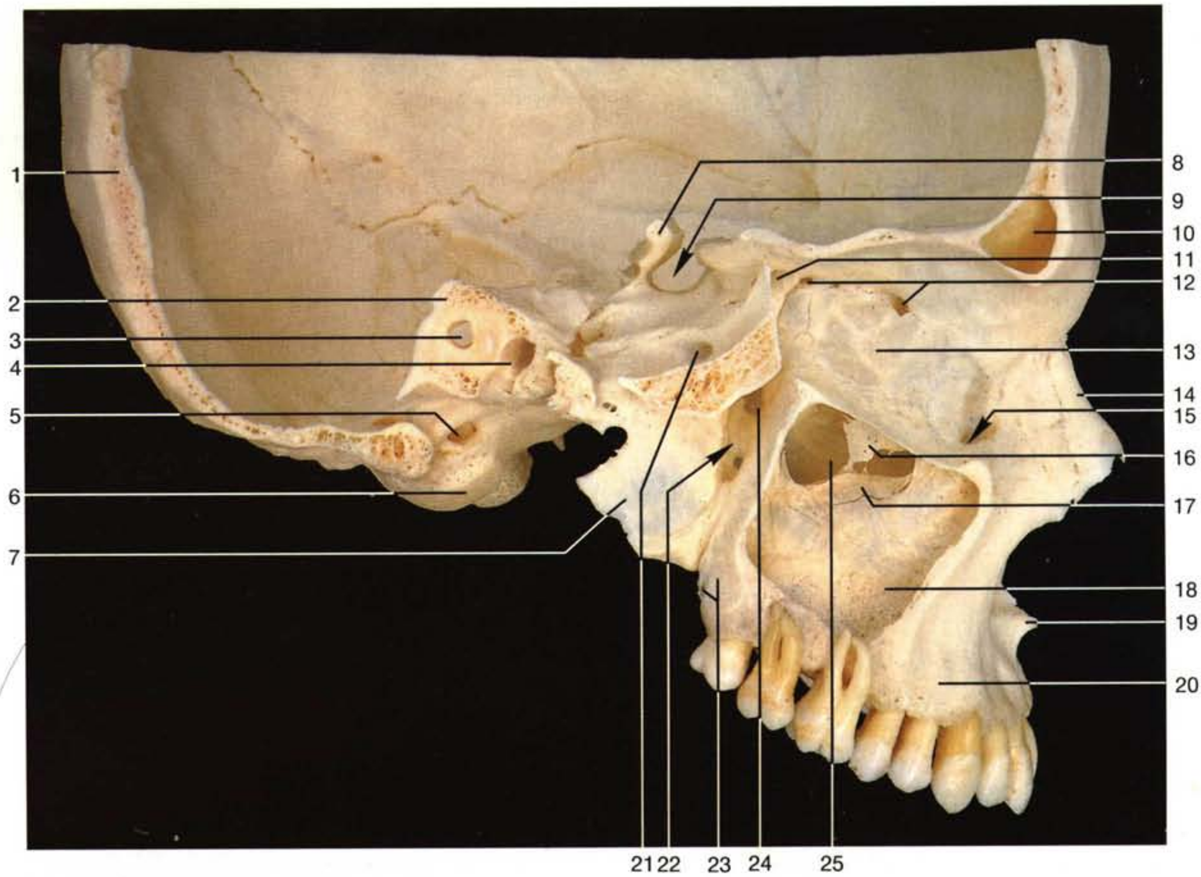

Paramedian section through the skull, right side (lateral aspect). Frontal and maxillary sinus are opened.

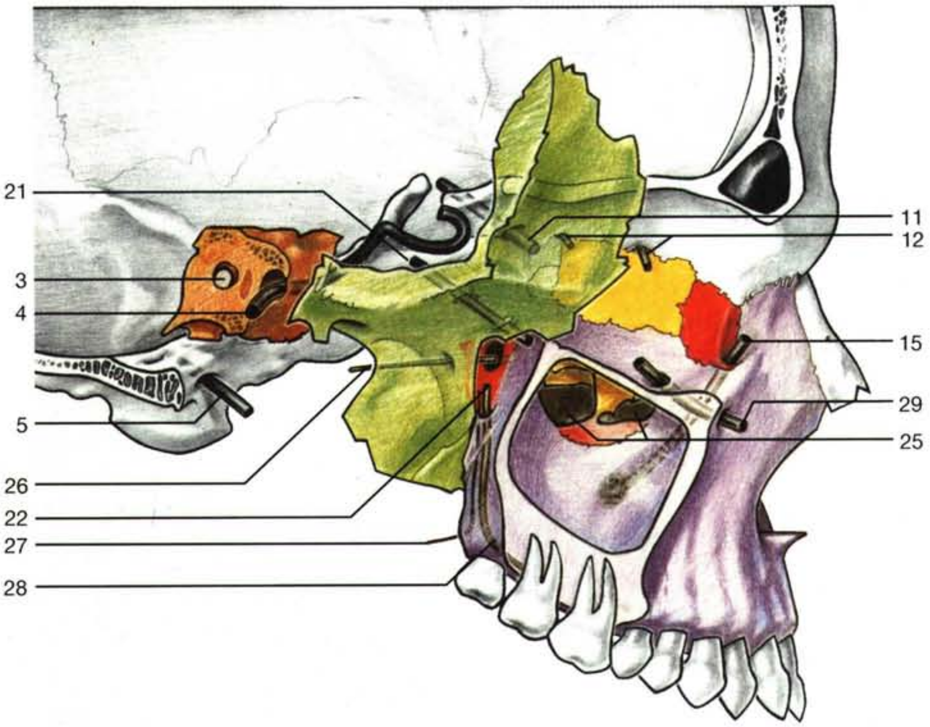

Illustration of canals and foramina connected with the right orbit and pterygopalatine fossa (compare the above figure). The greater wing of sphenoid bone (green) is shown as being transparent. Brown = temporal bone; yellow = ethmoid bone; red = laerimal bone; light red = inferior nasal concha; violet = maxilla; orange = palatine bone.

1 Occipital bone
2 Temporal bone (petrous part)
3 **Internal acoustic meatus**
4 **Carotid canal**
5 **Hypoglossal canal**
6 Occipital condyle
7 Lateral plate of pterygoid process
8 Dorsum of sella turcica
9 Sella turcica
10 Frontal sinus
11 **Optic canal**
12 **Posterior** and **anterior ethmoidal foramen**
13 Orbital plate of ethmoidal bone
14 Nasal bone
15 **Nasolacrimal canal**
16 Uncinate process
17 Inferior nasal concha (maxillary process)
18 Maxillary sinus
19 Anterior nasal spine
20 Alveolar process of maxilla
21 **Foramen rotundum**
22 **Pterygopalatine fossa**
23 Tuberosity of maxilla with **alveolar foramina**
24 **Sphenopalatine foramen**
25 Maxillary hiatus
26 **Pterygoid** or Vidian **canal**
27 **Lesser palatine canal**
28 **Greater palatine canal**
29 **Infraorbital canal**

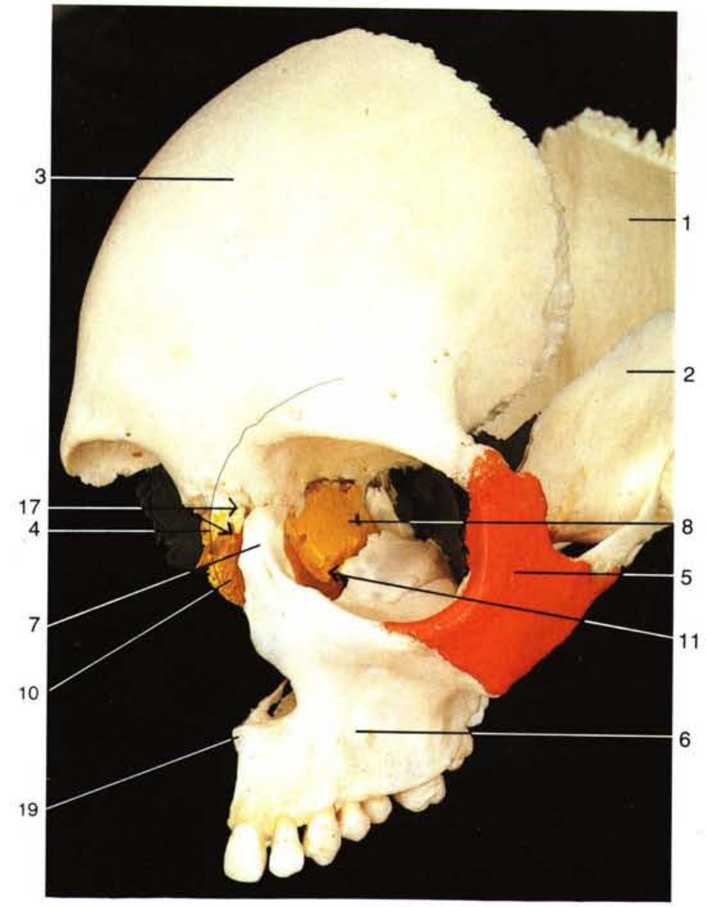

Anterior part of a disarticulated skull.
Orange = zygomatic bone; yellow = ethmoid bone; green = sphenoidal bone. The arrows indicate the locations of the lacrimal bone (11) and the nasal bone (17).

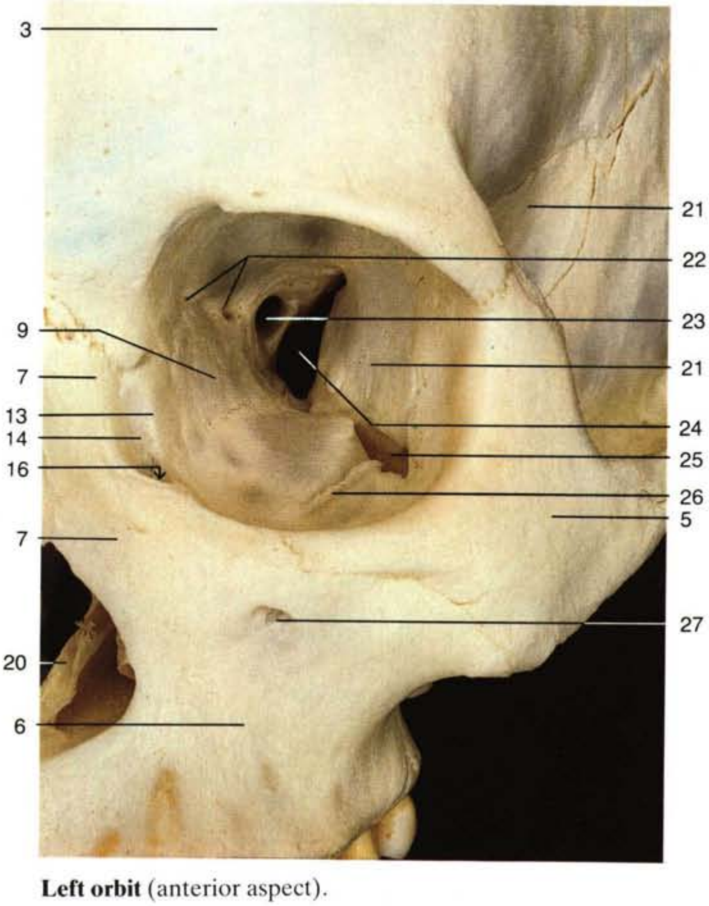

Left orbit (anterior aspect).

1. **Occipital bone**
2. **Temporal bone**
3. **Frontal bone**
4. Nasal spine of frontal bone
5. **Zygomatic bone**
6. **Maxilla**
7. Frontal process of maxilla
8. **Ethmoid bone**
9. Orbital plate of ethmoid bone
10. Perpendicular plate of ethmoid bone
11. Site of **lacrimal bone**
12. Lacrimal groove of lacrimal bone
13. Posterior lacrimal crest
14. Fossa for lacrimal sac
15. Lacrimal hamulus
16. Nasolacrimal canal
17. Site of **nasal bone**
18. Nasal foramina of nasal bone
19. Anterior nasal spine of maxilla
20. **Vomer**
21. Greater wing of sphenoid bone
22. Anterior and posterior ethmoidal foramina
23. Optic canal
24. Superior orbital fissure
25. Inferior orbital fissure
26. Infraorbital groove
27. Infraorbital foramen

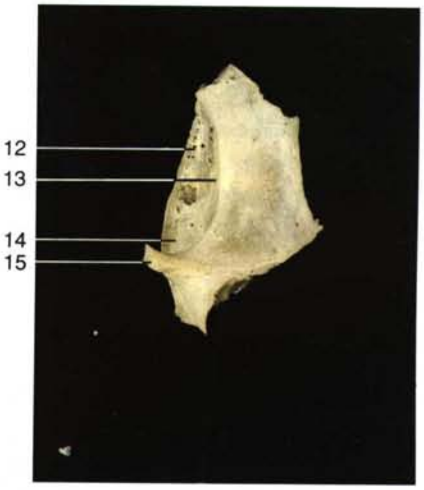

Left lacrimal bone (anterior aspect).

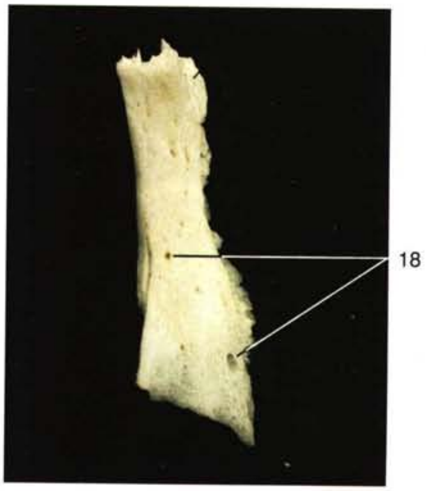

Left nasal bone (anterior aspect).

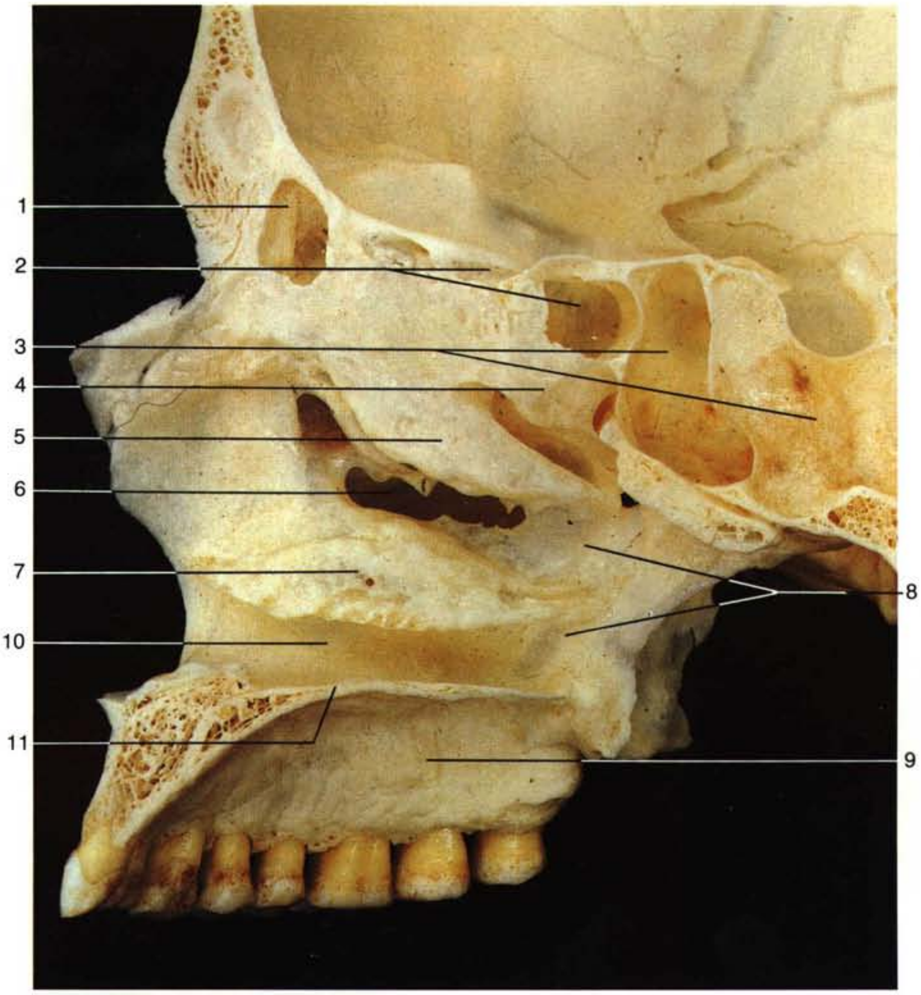

Lateral wall of the nasal cavity. Median section through the skull.

1 Frontal sinus
2 Ethmoidal air cells
3 Sphenoid sinus
4 Superior nasal concha
5 Middle nasal concha
6 Maxillary hiatus
7 Inferior nasal concha
8 Palatine bone
9 Maxilla
10 Inferior meatus
11 Palatine process of the maxilla

▷

To page 49:

Blue = Occipital bone
Light green = Parietal bone
Light brown = Frontal bone
Dark brown = Temporal bone
Red = Sphenoid bone
Dark green = Ethmoid bone
Light blue = Nasal bone
Pink = Inferior concha
Orange = Vomer
Violet = Maxilla
White = Palatine bone
White = Mandible

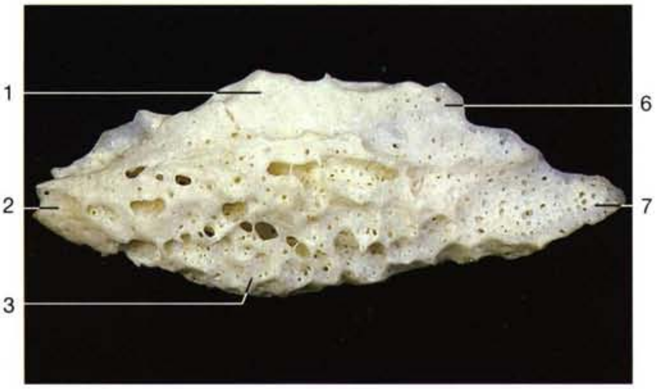

Right inferior nasal concha (medial aspect). Anterior part to the left.

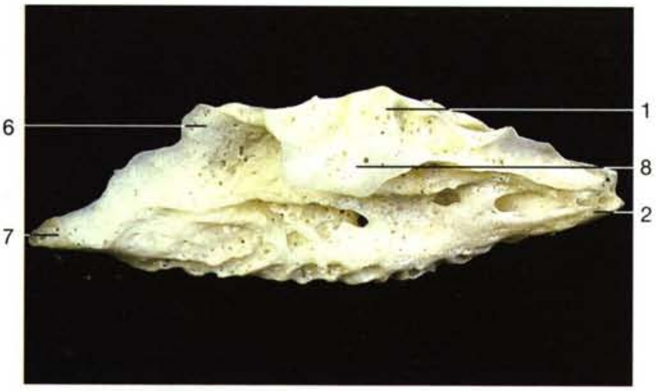

Right inferior nasal concha (lateral aspect). Anterior part to the right.

Inferior Concha and Vomer

1 Ethmoidal process
2 Anterior part of concha
3 Inferior border
4 Ala of vomer
5 Posterior border of nasal septum
6 Lacrimal process
7 Posterior part of concha
8 Maxillary process

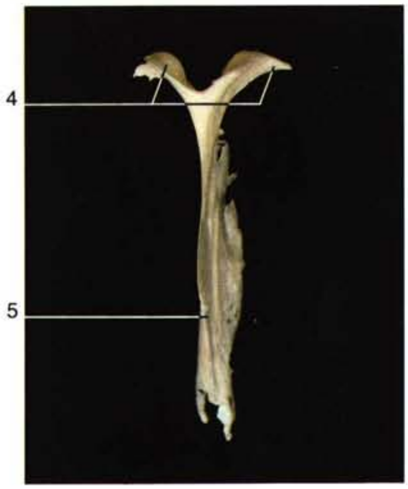

Vomer (posterior aspect).

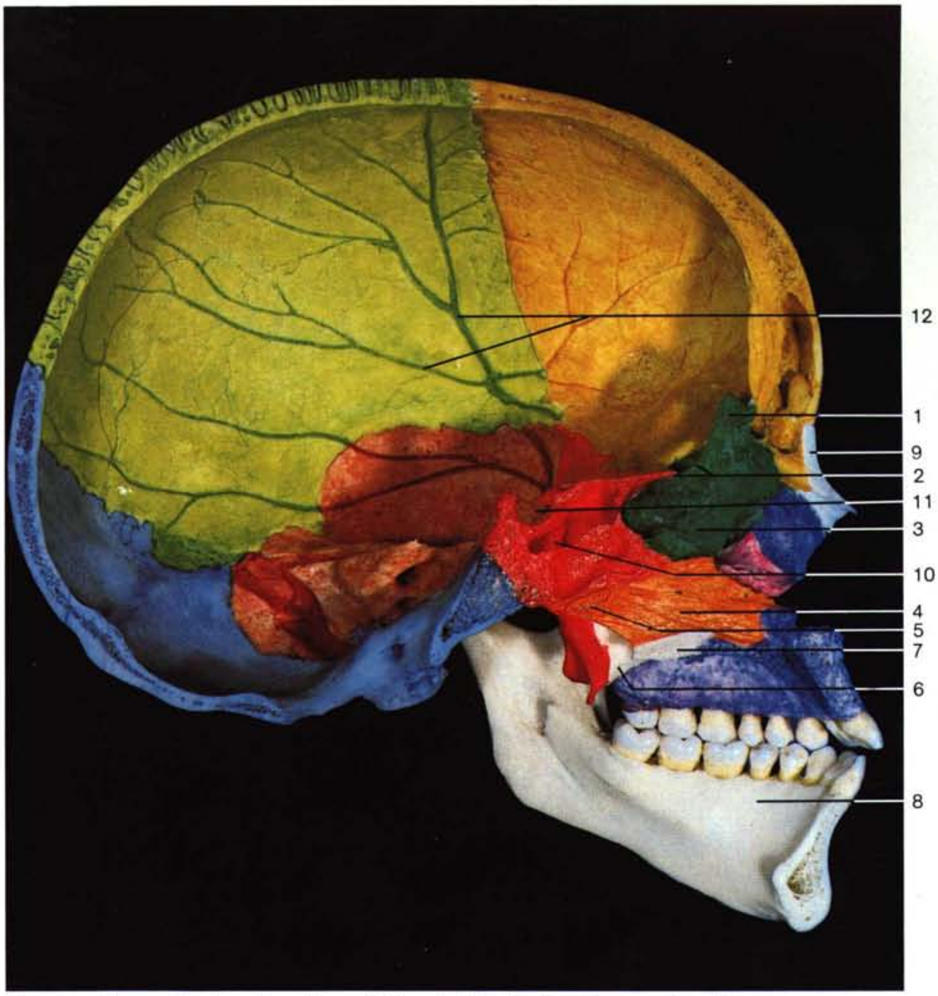

1 Crista galli
2 Cribriform plate of ethmoid bone
3 Perpendicular plate of ethmoid bone
4 Vomer
5 Ala of the vomer
6 Palatine bone (perpendicular process)
7 Palatine bone (horizontal plate)
8 Mandible
9 Nasal bone
10 Sphenoidal sinus
11 Hypophysial fossa (sella turcica)
12 Grooves for the middle meningeal artery

Cartilages of the nose

13 Lateral nasal cartilages
14 Greater alar cartilage
15 Lesser alar cartilages
16 Septal cartilage

17 Location of nasal bone

Paramedian sagittal section through the skull including the nasal septum.

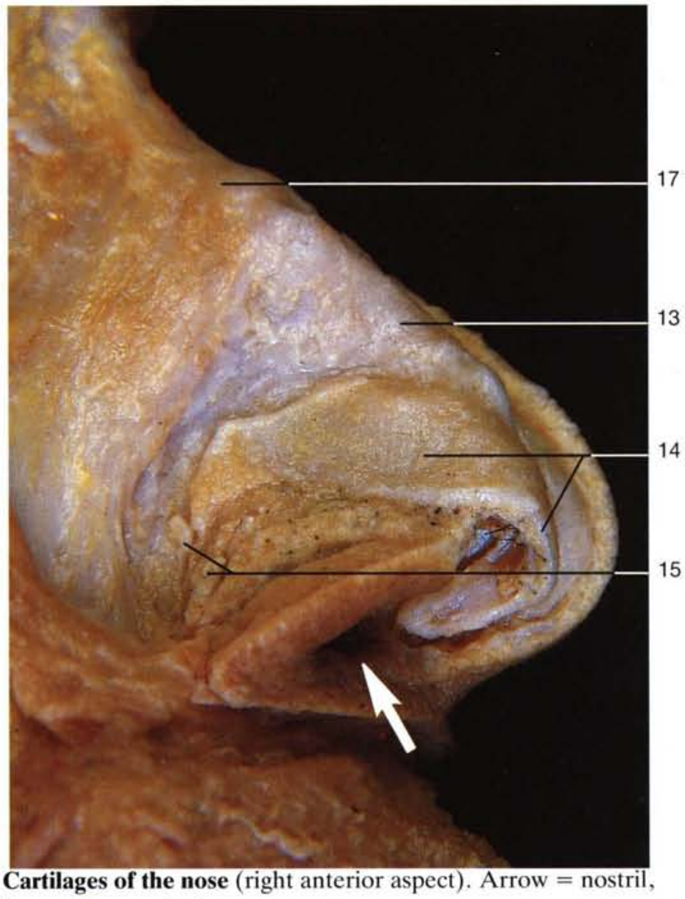

Cartilages of the nose (right anterior aspect). Arrow = nostril, framed by nasal wing.

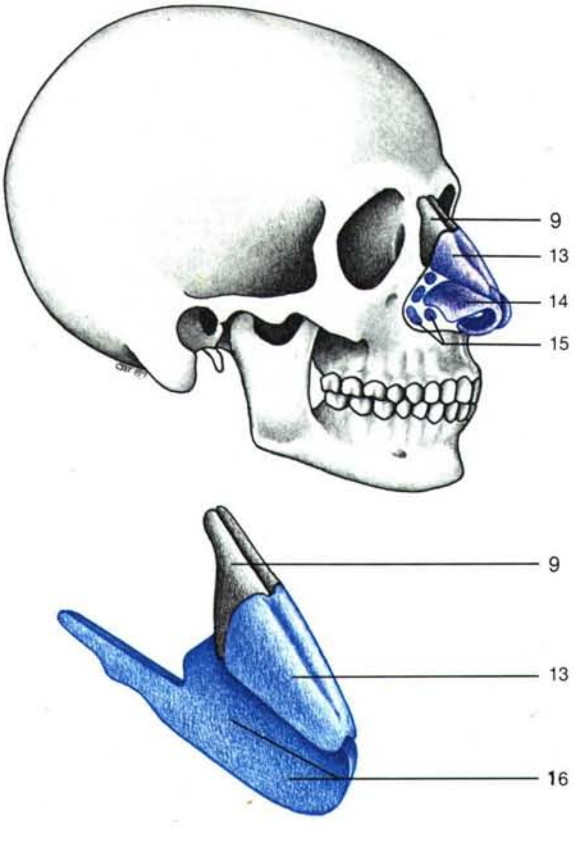

Cartilages of the nose.
Schematic diagram of the external nose.

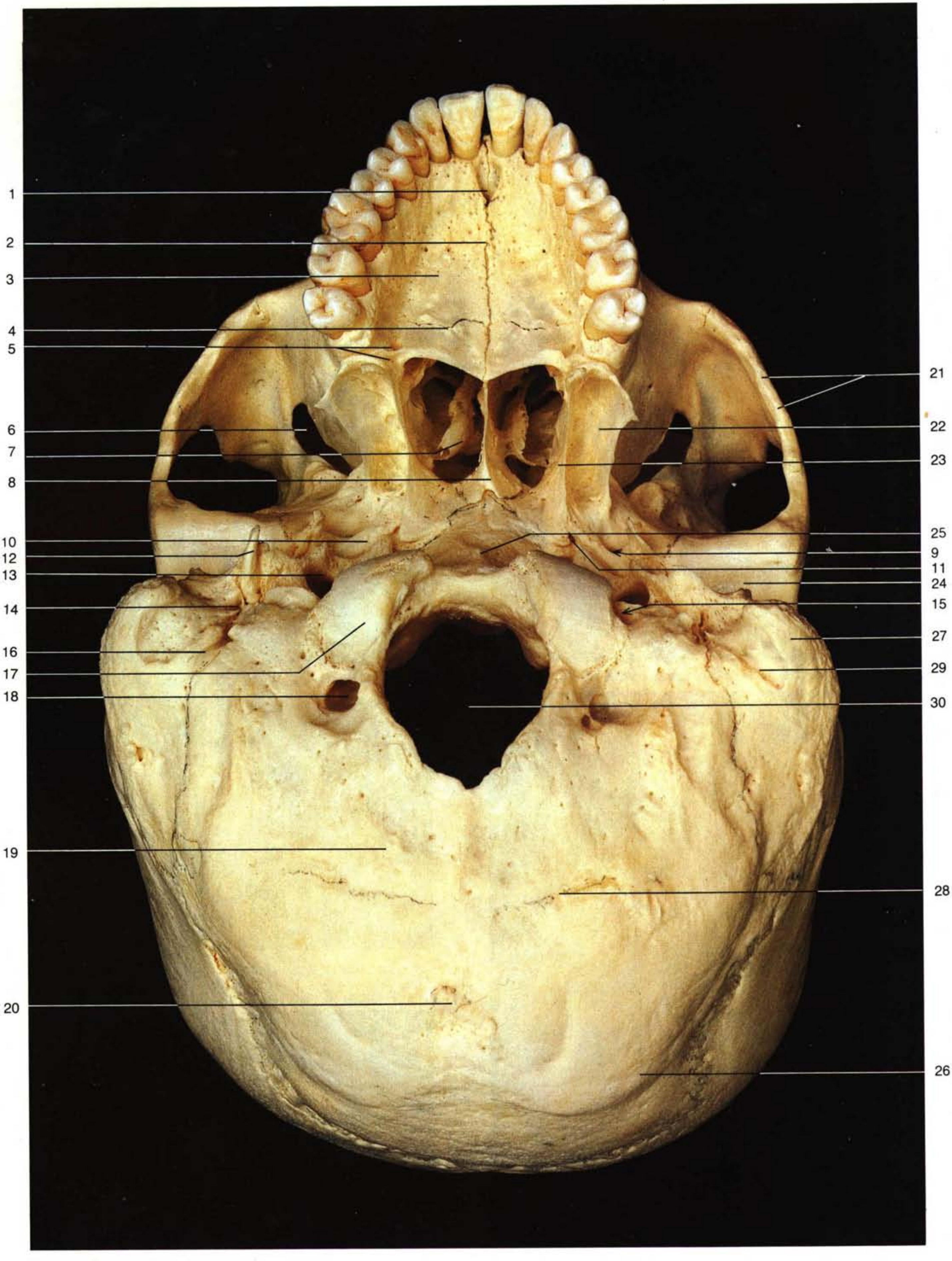
1
2
3
4
5
6
7
8
10
12
13
14
16
17
18
19
20
21
22
23
25
9
11
24
15
27
29
30
28
26

Base of the skull (inferior aspect).

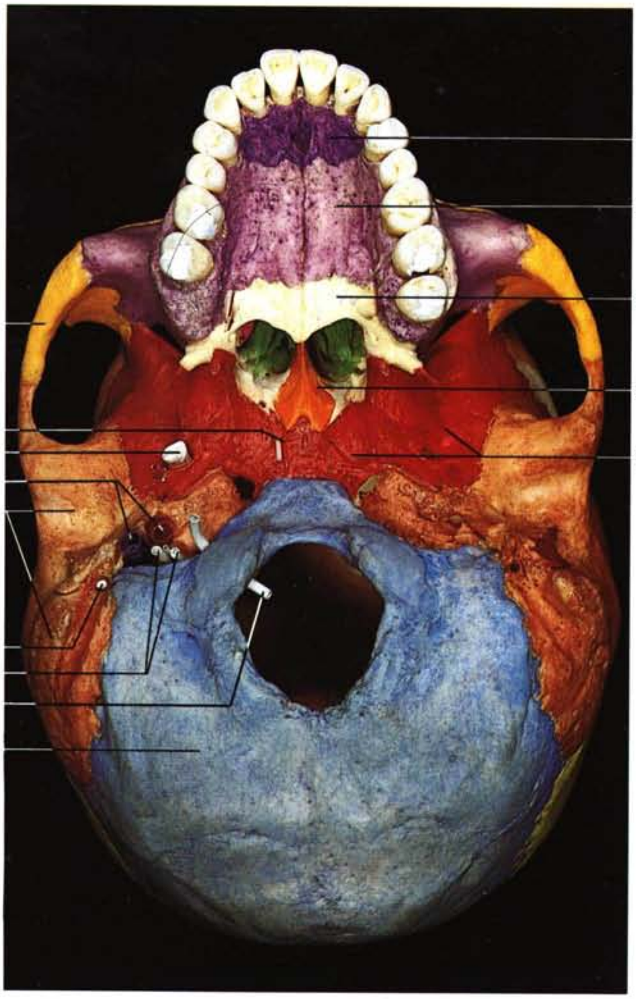

Base of the skull (from below). The individual bones are indicated by different colours.

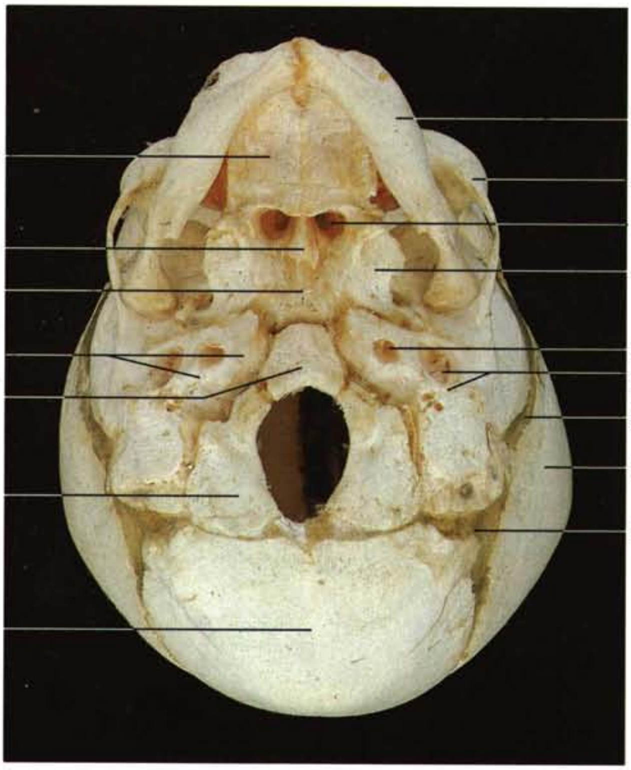

Skull of the newborn (inferior aspect).

A **Pterygoid canal**
B **Foramen ovale**
C Internal carotid artery within **carotid canal** and internal jugular vein within the venous part of jugular foramen
D **Stylomastoid foramen** (facial nerve)
E **Jugular foramen** (glossopharyngeal, vagus and accessory nerves)
F **Hypoglossal canal** (hypoglossal nerve)

1 Incisive canal
2 Median palatine suture
3 Palatine process of maxilla
4 Palatomaxillary suture
5 Greater and lesser palatine foramina
6 Inferior orbital fissure
7 Middle concha (process of ethmoid bone)
8 Vomer
9 Foramen ovale
10 Groove for auditory tube
11 Pterygoid canal
12 Styloid process
13 Carotid canal
14 Stylomastoid foramen
15 Jugular foramen
16 Groove for occipital artery
17 Occipital condyle
18 Condylar canal
19 Nuchal plane
20 External occipital protuberance
21 Zygomatic arch
22 Lateral pterygoid plate
23 Medial pterygoid plate
24 Mandibular fossa
25 Pharyngeal tubercle
26 Superior nuchal line
27 Mastoid process
28 Inferior nuchal line
29 Mastoid notch
30 Foramen magnum
31 **Incisive bone** or **premaxilla** (dark violet)
32 **Maxilla** (violet)
33 **Palatine bone** (white)
34 **Vomer** (orange)
35 **Sphenoid bone** (red)
36 **Zygomatic bone** (yellow)
37 **Temporal bone** (brown)
38 **Occipital bone** (blue)
39 Palatine process of maxilla
40 Vomer
41 Sphenoid bone
42 Petrous part of temporal bone
43 Basilar part }
44 Lateral part } of occipital bone
45 Squamous part }
46 Mandible
47 Zygomatic arch
48 Choana
49 Pterygoid process of sphenoid bone
50 Carotid canal
51 External acoustic meatus (tympanic annulus)
52 Sphenoidal fontanelle
53 Parietal bone
54 Mastoid fontanelle

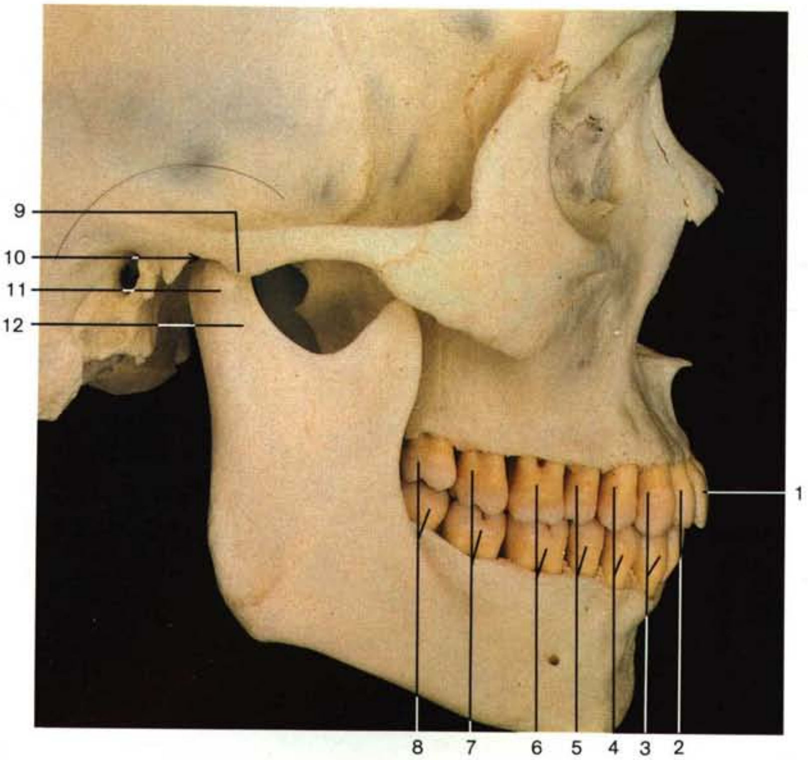

1 Central incisor
2 Lateral incisor
3 Canines
4 First premolars or bicuspids
5 Second premolars or bicuspids
6 First molars
7 Second molars
8 Third molars
9 Articular tubercle
10 Mandibular fossa
11 Head of mandible
12 Condylar process

Normal position of teeth. Dentition in centric occlusion (lateral view).

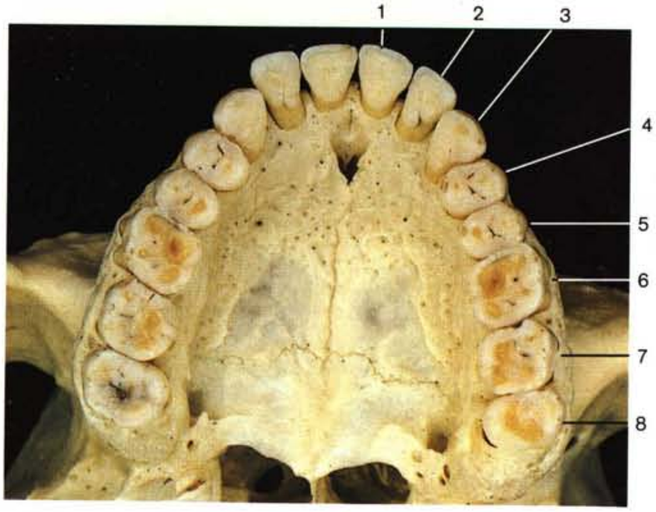

Upper teeth of the adult (inferior aspect).

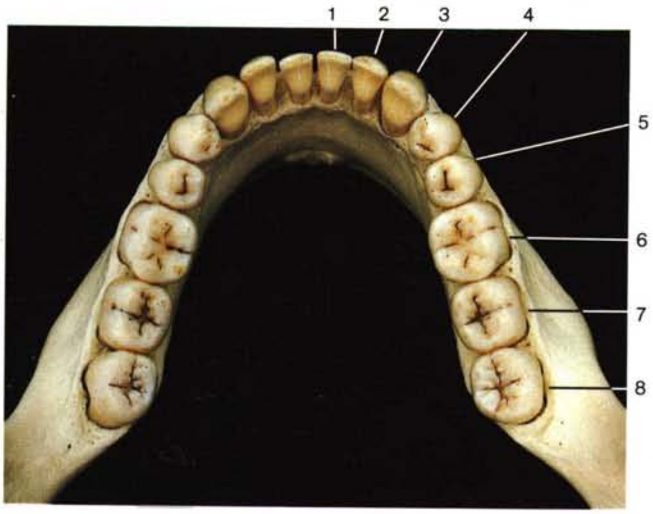

Lower teeth of the adult (superior aspect).

Table of dentition. Eruption of deciduous and permanent teeth (after C. Röse according to A. Kröncke).

Primary dentition (Deciduous teeth)	Maxilla months post partum	Mandible months
1. Central incisor	10.3	8.6
2. Lateral incisor	12.2	14.4
3. Cuspid incisor	19.5	20.1
4. First molar	15.5	16.5
5. Second molar	24.8	24.5

Secondary dentition (Permanent teeth)	years and months ♂	♀	years and months ♂	♀
1. Central incisor	7/8	7/5	6/10	6/7
2. Lateral incisor	8/11	8/6	7/11	7/7
3. Cuspid incisor	12/2	11/7	11/12	10/3
4. First premolar	10/5	10/1	11/3	10/8
5. Second premolar	11/4	11/1	12/0	11/7
6. First molar	6/7	6/6	6/5	6/3
7. Second molar	12/9	12/5	12/3	11/9

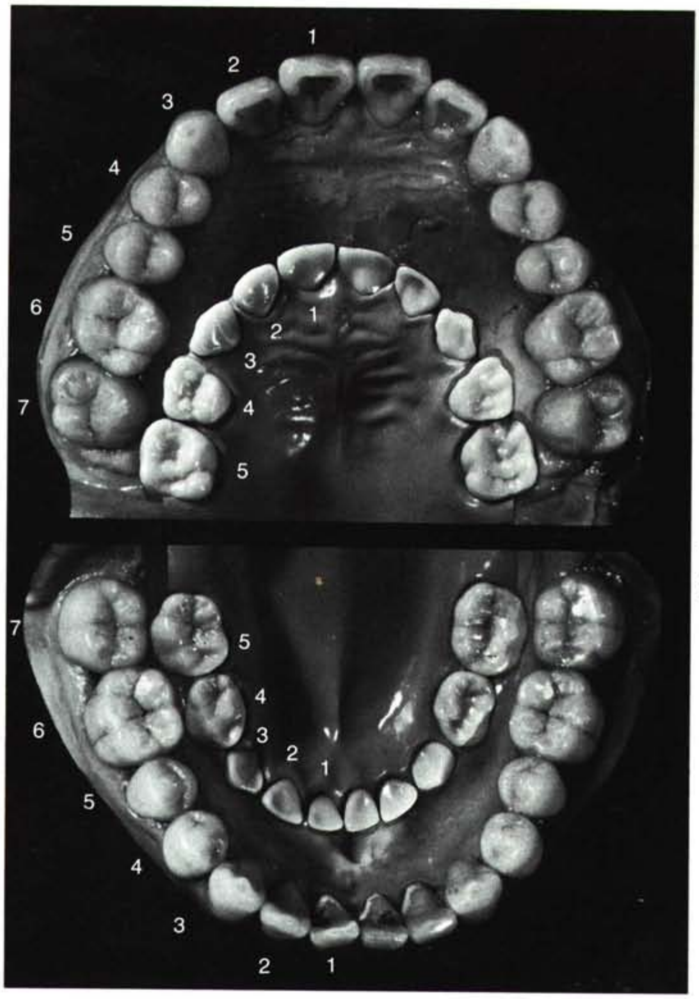

Comparison of the deciduous and permanent teeth. Notice that the breadth of the alveolar arch of the child's mandible and maxilla holding the deciduous teeth is nearly the same as the comparable portion in the jaws of the adult. Note the unerupted third molars.

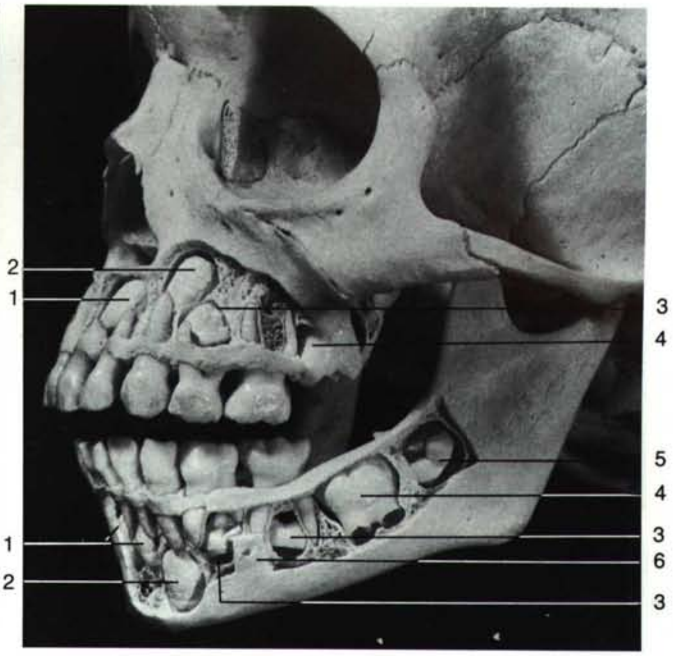

Deciduous teeth in child's skull. The developing crowns of the permanent teeth are displayed in their sockets in the maxilla and mandible.

△

1 Permanent incisors
2 Permanent cuspid (canine)
3 Premolars
4 First permanent molar
5 Second permanent molar
6 Mental foramen

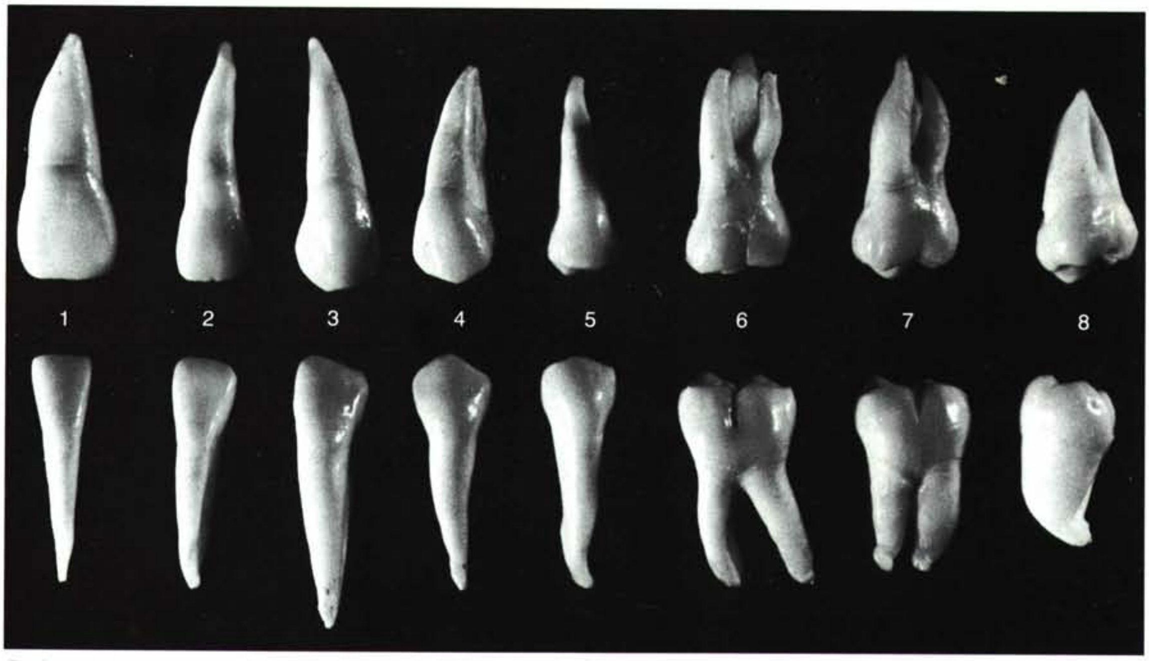

Isolated teeth of the alveolar part of the maxilla (top row) and the mandible (lower row), labial surface of the teeth.

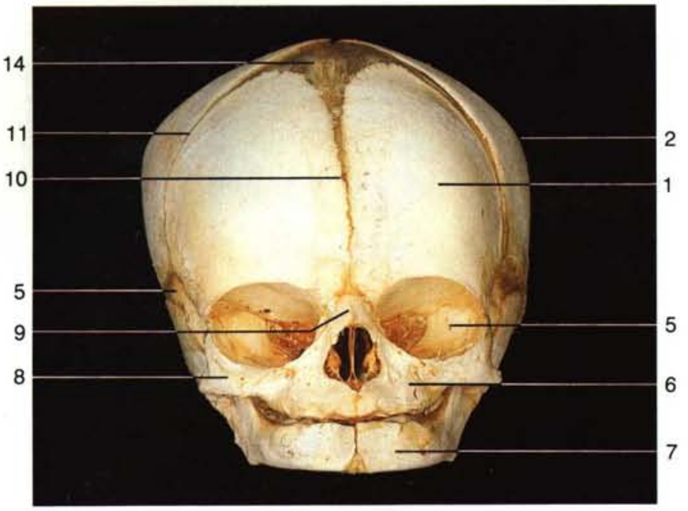

Skull of the newborn (anterior aspect).

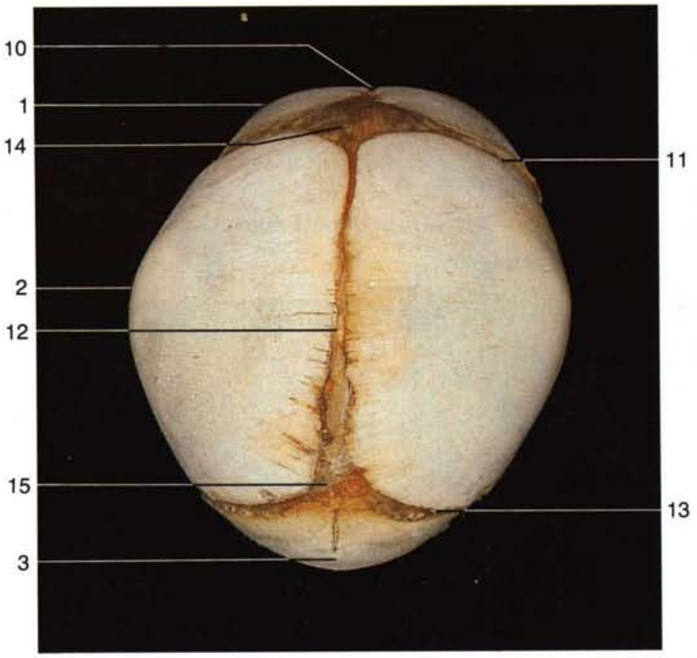

Skull of the newborn (superior aspect). Calvaria.

Cranial skeleton

1 Frontal tuber or eminence
2 Parietal tuber or eminence
3 Occipital tuber or eminence
4 Squamous part of temporal bone
5 Greater wing of sphenoid bone

Facial skeleton

6 Maxilla
7 Mandible
8 Zygomatic bone
9 Nasal bone

Sutures and fontanelles

10 Frontal suture
11 Coronal suture
12 Sagittal suture
13 Lambdoid suture
14 Anterior fontanelle
15 Posterior fontanelle
16 Sphenoidal fontanelle
17 Mastoid fontanelle

Base of the skull

18 Frontal bone
19 Ethmoid bone
20 Sphenoid bone
21 Hypophysial fossa (sella turcica)
22 Dorsum sellae
23 Temporal bone
24 Mastoid fontanelle
25 Occipital bone

In the newborn the facial skeleton, in contrast to the cranial skeleton, appears relatively small. There are no teeth presenting. The bones of the cranium are separated by wide fontanelles.

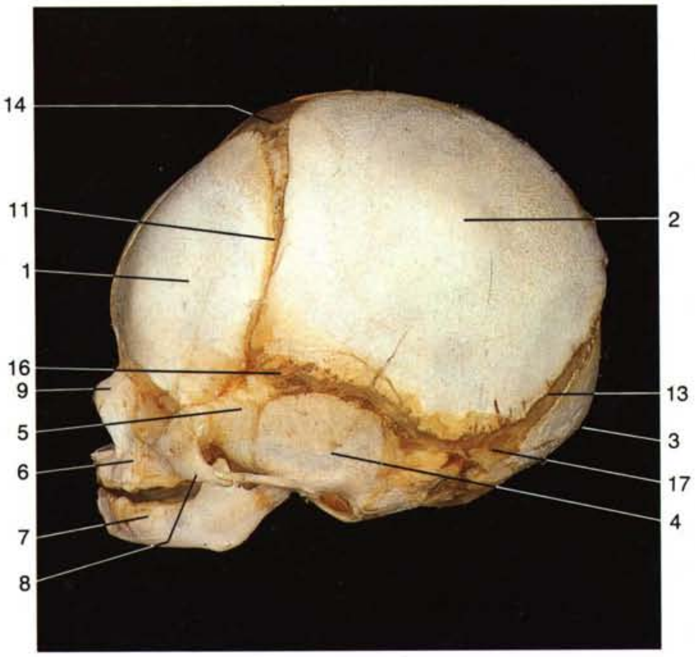

Skull of the newborn (lateral aspect).

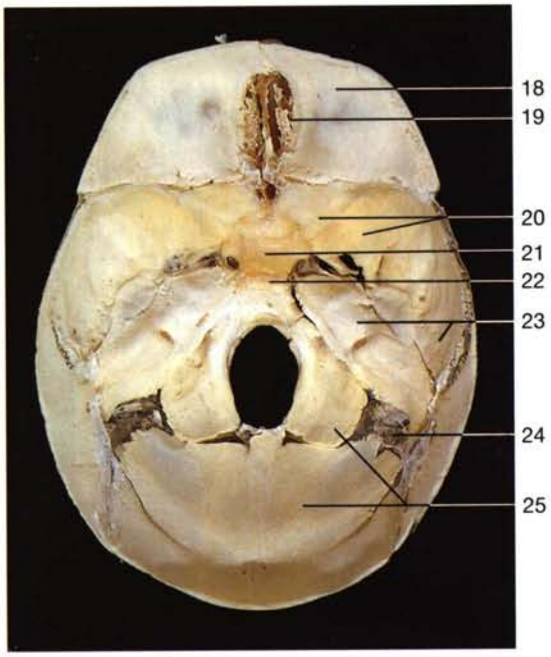

Base of the skull of the newborn (internal aspect).

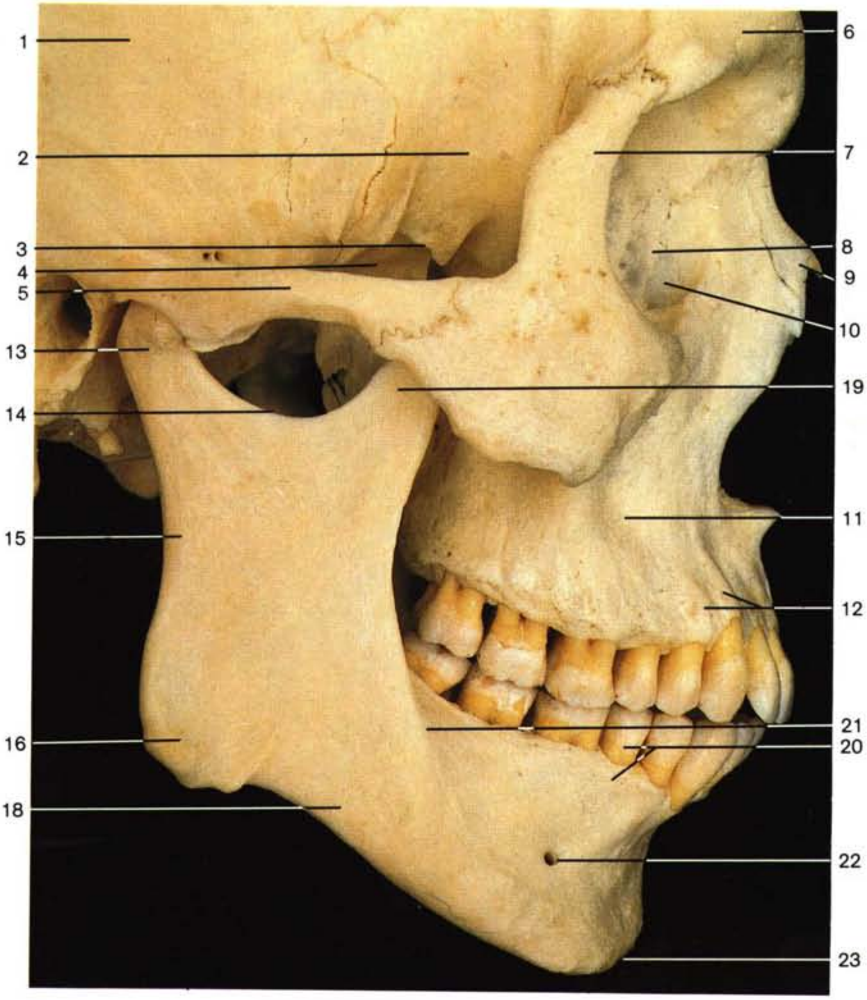

Lateral aspect of the facial bones. Mandible and teeth in the position of occlusion. Upper and lower jaw occluded.

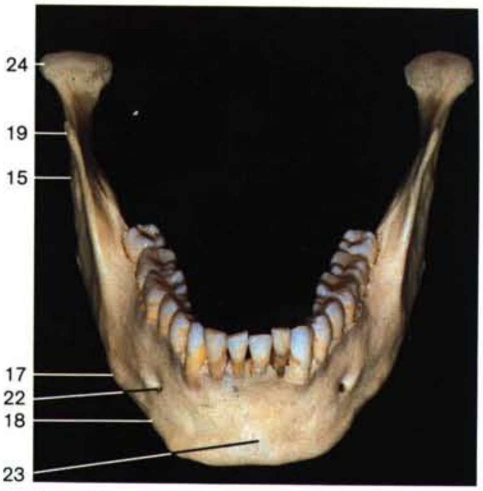

Mandible of the adult (anterior aspect).

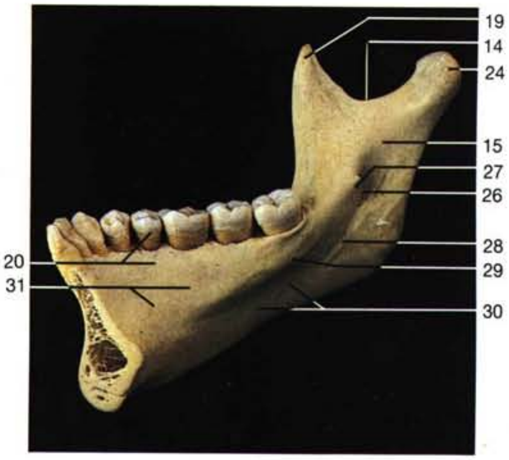

Right half of mandible (medial aspect).

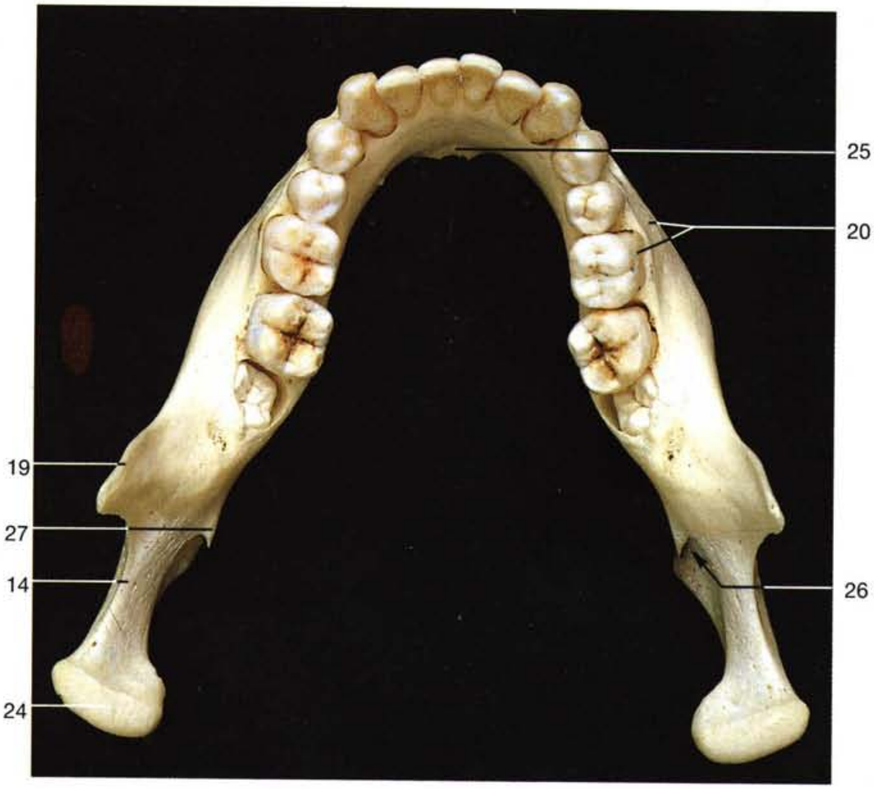

Mandible of the adult (superior aspect).

1 **Temporal bone**
2 Temporal fossa (greater wing of sphenoid bone)
3 Infratemporal crest
4 Infratemporal fossa
5 Zygomatic arch
6 **Frontal bone**

Mandible

13 Condylar process
14 Mandibular notch
15 Ramus of the mandible
16 Masseteric tuberosity
17 Angle of the mandible
18 Body of the mandible
19 Coronoid process
20 Alveolar process including teeth
21 Oblique line
22 Mental foramen
23 Mental protuberance
24 Head of the mandible
25 Genial tubercle or mental spine
26 Mandibular foramen (entrance to mandibular canal)
27 Lingula
28 Mylohyoid sulcus
29 Mylohyoid line
30 Submandibular fossa
31 Sublingual fossa

7 **Zygomatic bone** (frontal process)
8 **Lacrimal bone**
9 **Nasal bone**
10 Lacrimal groove
11 **Maxilla** (canine fossa)
12 Alveolar process of maxilla

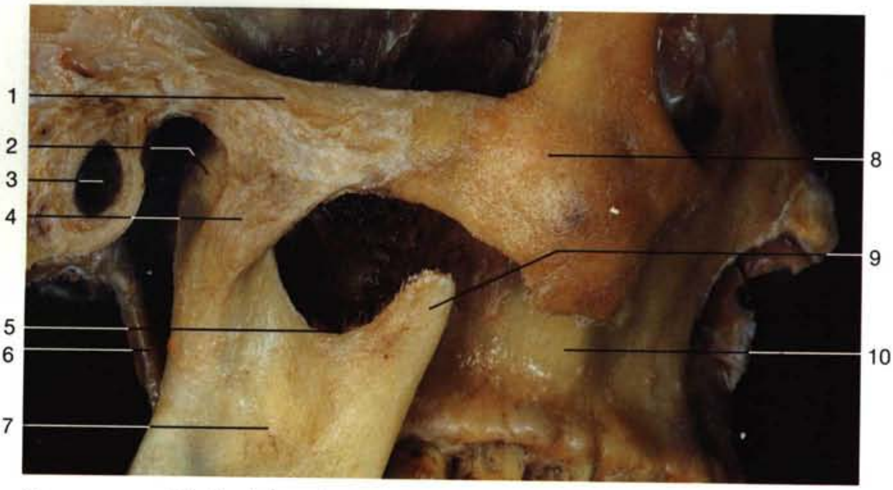

Temporomandibular joint with ligaments.

1 Zygomatic arch
2 Articular capsule
3 External acoustic meatus
4 **Lateral ligament**
5 Mandibular notch
6 **Stylomandibular ligament**
7 Ramus of the mandible
8 Zygomatic bone
9 Coronoid process
10 Maxilla
11 Articular cartilage of condylar process
12 Styloid process
13 Mandibular fossa
14 **Articular disc**
15 Articular tubercle
16 Lateral pterygoid muscle
17 Condylar process of mandible

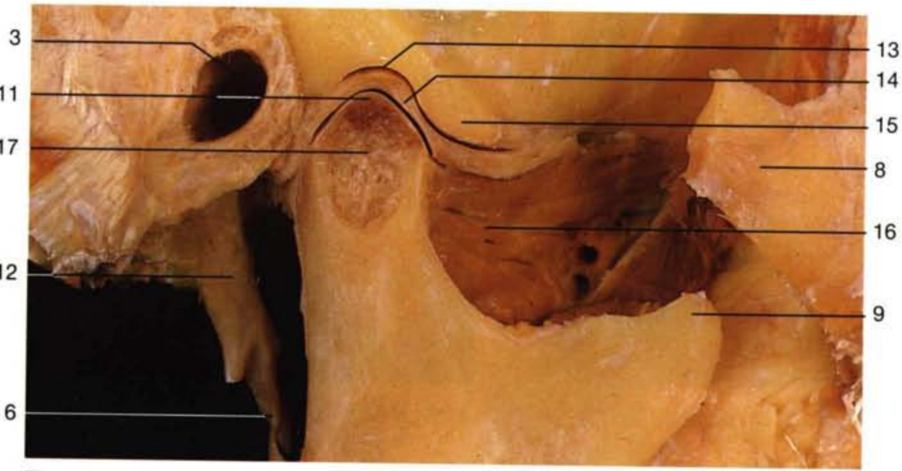

Temporomandibular joint, sagittal section.

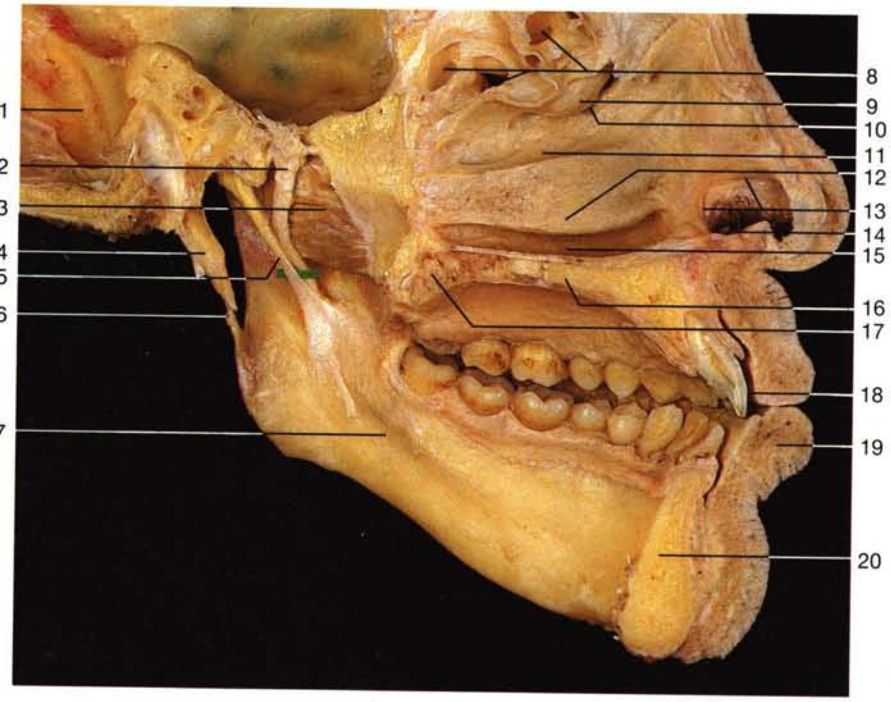

Ligaments of temporomandibular joint. Left half of the head (medial aspect).

1 Groove for sigmoid sinus
2 Mandibular nerve
3 Lateral pterygoid muscle
4 Styloid process
5 **Sphenomandibular ligament**
6 **Stylomandibular ligament**
7 Mylohyoid groove
8 **Ethmoidal air cells**
9 Ethmoidal bulla
10 **Hiatus semilunaris**
11 Middle meatus
12 **Inferior nasal concha**
13 Limen nasi
14 Vestibule with hairs
15 Inferior meatus
16 Hard palate
17 Soft palate
18 Vestibule of oral cavity
19 Lower lip
20 Mandible

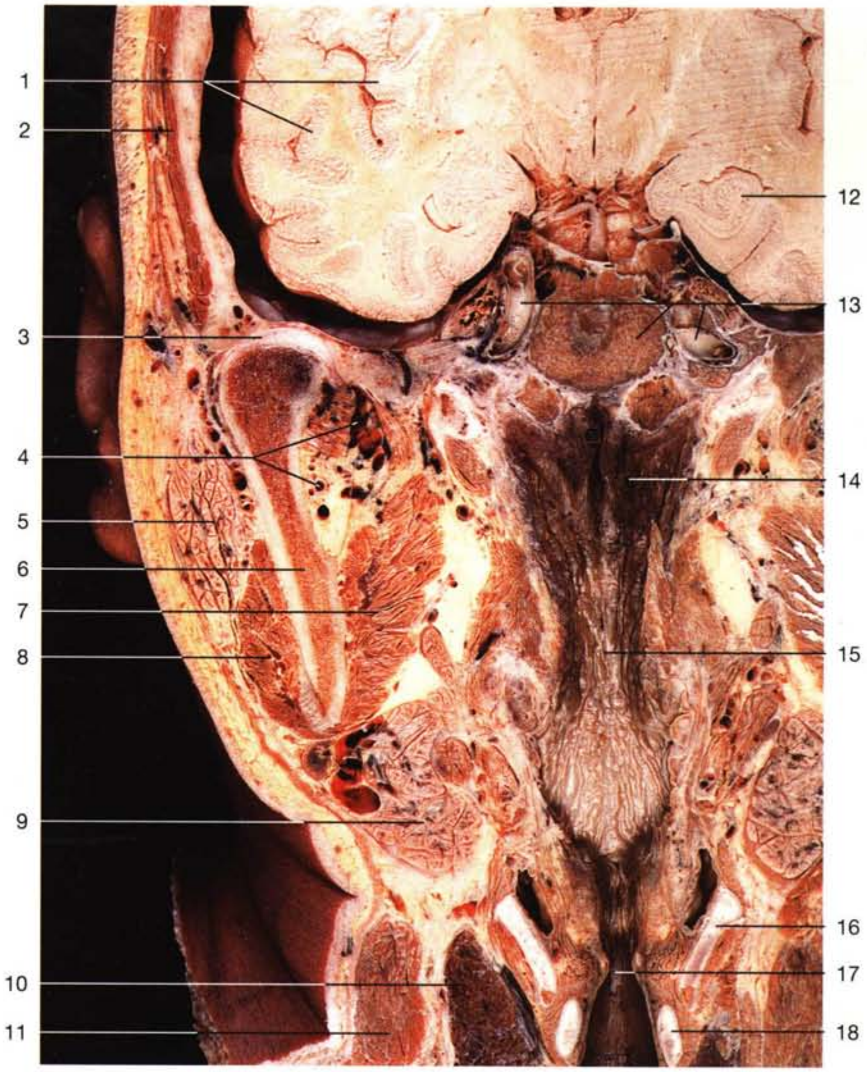

Coronal section through the head at the level of the temporomandibular joint (right side, anterior aspect).

1 Insular lobe and temporal lobe
2 Temporalis muscle
3 **Articular disc of temporomandibular joint**
4 Maxillary artery and pterygoid venous plexus
5 Parotid gland
6 **Mandible**
7 **Medial pterygoid muscle**
8 **Masseter muscle**
9 Submandibular gland
10 Thyroid gland
11 Sternocleidomastoid muscle
12 Hippocampus
13 Internal carotid artery and sphenoid bone
14 Pharyngeal tonsil
15 Pharynx
16 Thyroid cartilage
17 Rima glottidis
18 Cricoid cartilage
19 **Hard palate** and palatine glands
20 **Oral cavity**
21 Upper molar
22 **Oral vestibule**
23 Lower molar
24 Platysma muscle
25 Maxillary sinus
26 Superior longitudinal muscle of tongue
27 **Transverse muscle of tongue**
28 Buccinator muscle
29 Inferior longitudinal muscle of tongue
30 Sublingual gland
31 **Genioglossus muscle**
32 Mastoid process
33 Styloid process
34 **Stylomandibular ligament**
35 Articular capsule
36 **Lateral ligament**
37 Zygomatic arch
38 **Sphenomandibular ligament**
39 Mandibular foramen

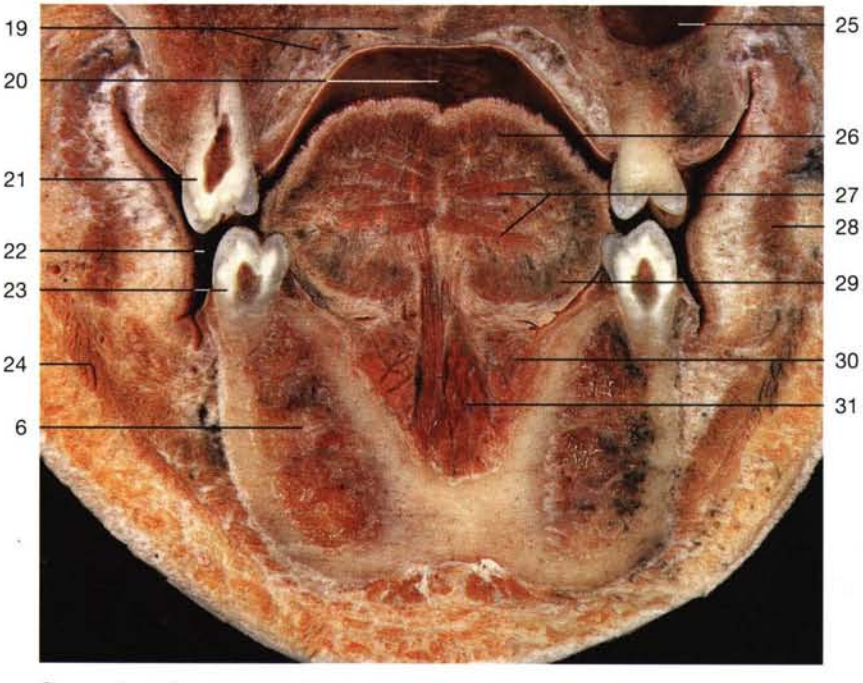

Coronal section through the oral cavity

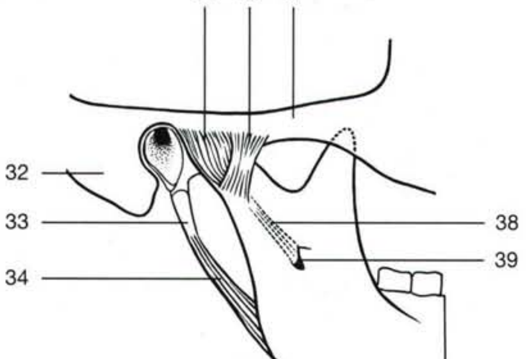

Ligaments related to the **temporomandibular joint**

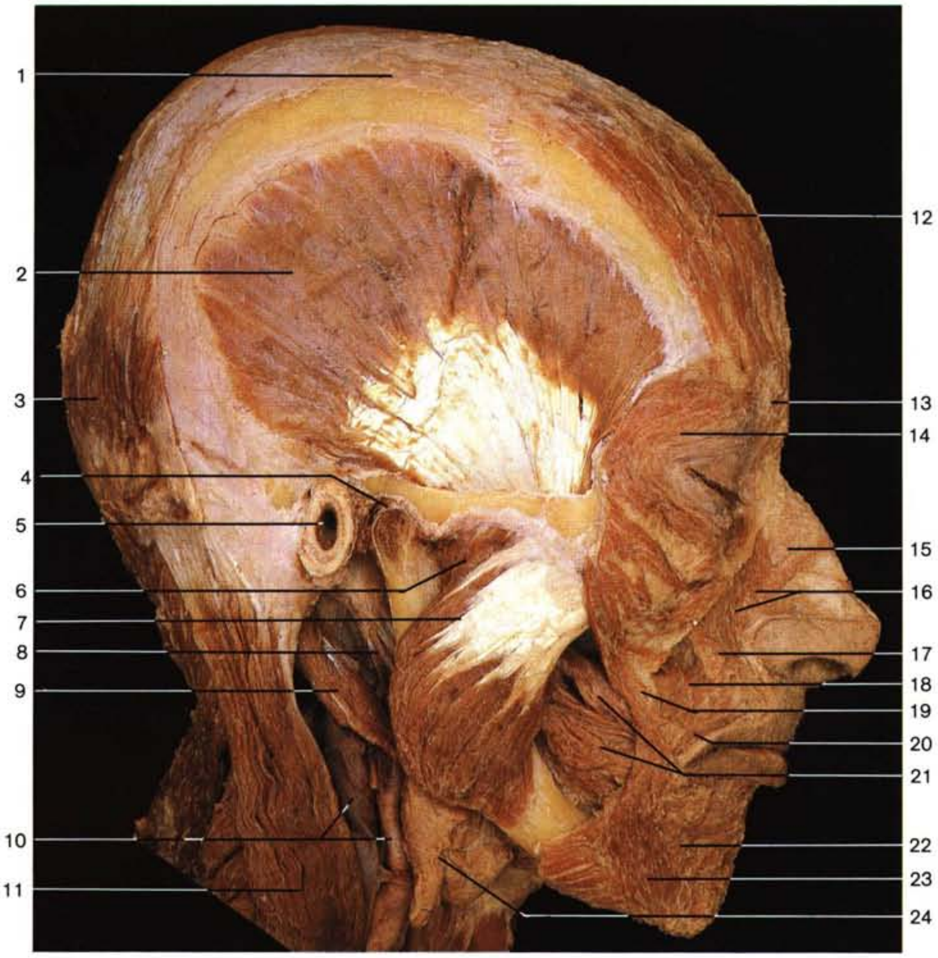

Temporalis and masseter muscles. The temporal fascia has been removed, the temporomandibular joint severed and the zygomatic arch displayed.

1 Galea aponeurotica
2 **Temporalis muscle**
3 Occipital belly of occipitofrontalis muscle
4 **Temporomandibular joint**
5 External acoustic meatus
6 Deep layer of **masseter muscle**
7 Superficial layer of masseter muscle
8 Stylohyoid muscle
9 Posterior belly of **digastric muscle**
10 Internal jugular vein and external carotid artery
11 **Sternocleidomastoid muscle**
12 Frontal belly of occipitofrontalis muscle
13 Depressor supercilii muscle
14 **Orbicularis oculi muscle**
15 Transverse part of nasalis muscle
16 Levator labii superioris alaeque nasi muscle
17 Levator labii superioris muscle
18 Levator anguli oris muscle
19 **Zygomaticus major muscle**
20 **Orbicularis oris muscle**
21 **Buccinator muscle**
22 Depressor labii inferioris muscle
23 Depressor anguli oris muscle
24 **Submandibular gland**
25 Lateral nasal cartilage
26 Greater alar cartilage (lateral part)
27 Lesser alar cartilages
28 Greater alar cartilage (medial part)
29 Infraorbital nerve
30 Anterior belly of digastric muscle
31 Hypoglossal nerve and hyoglossus muscle
32 Superior thyroid artery
33 **Zygomatic arch**
34 Internal carotid artery
35 Common carotid artery

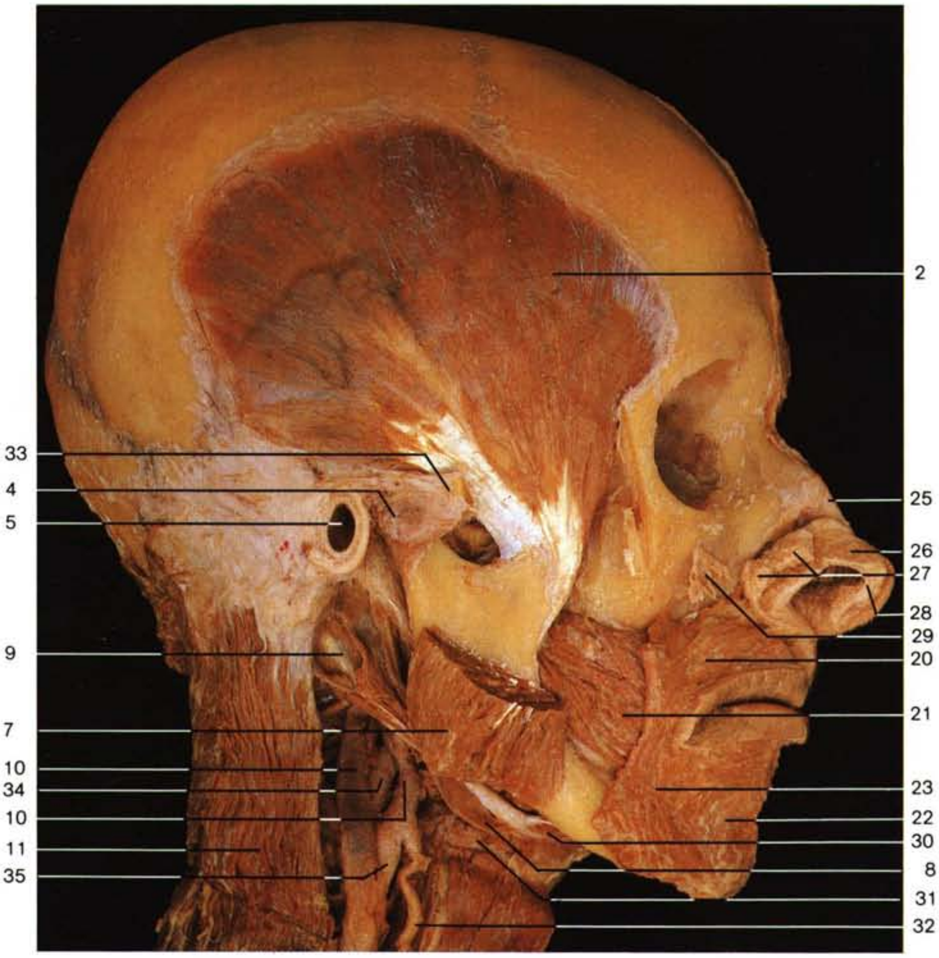

Temporalis muscle and temporomandibular joint. The zygomatic arch and the masseter muscle have been partially severed to display the insertion of the temporalis muscle.

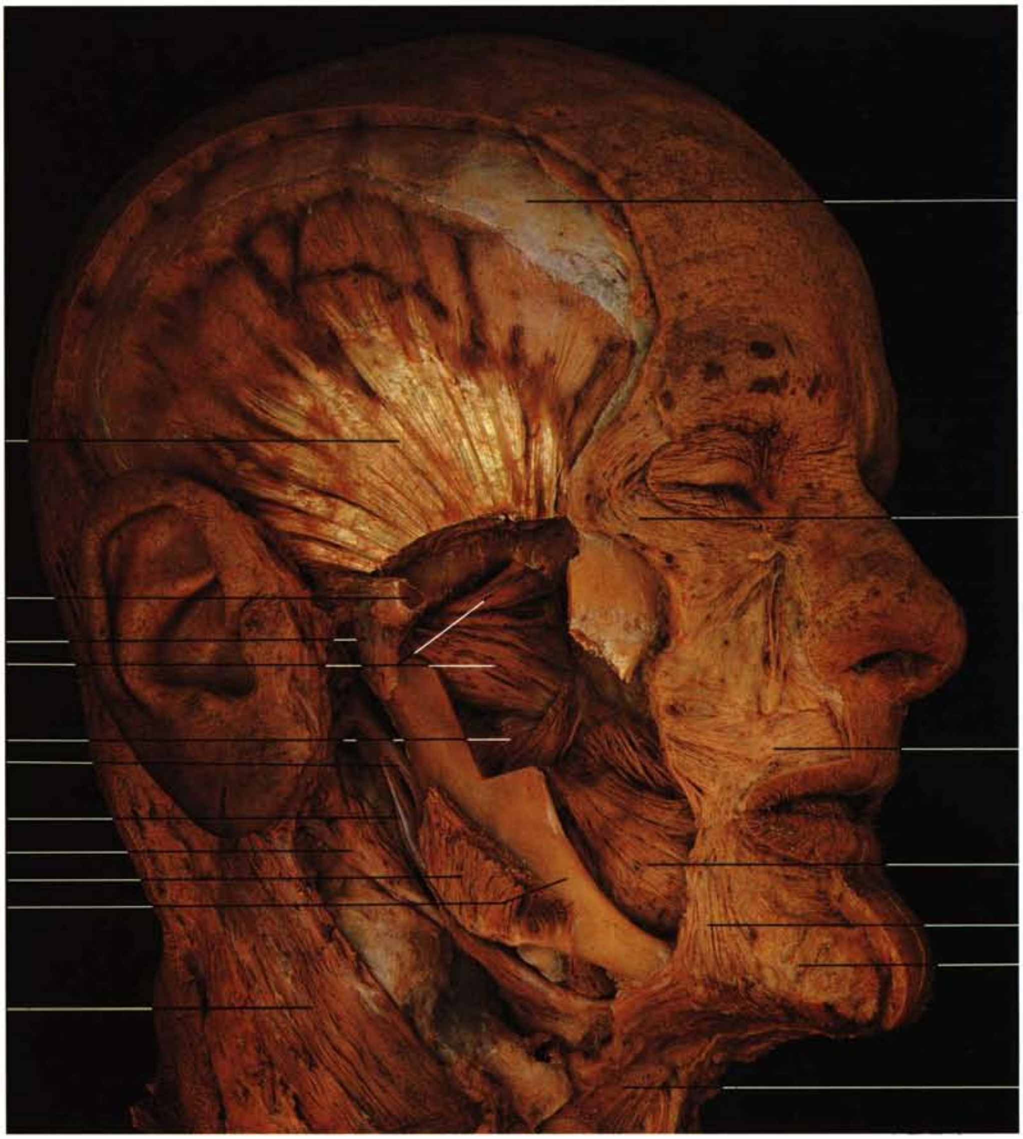

1 Periosteum
2 **Temporalis muscle**
3 Zygomatic arch
4 Articular capsule of temporomandibular joint
5 **Lateral pterygoid muscle** (upper and lower head)
6 **Medial pterygoid muscle**
7 Styloglossus muscle
8 Stylohyoid muscle
9 Posterior belly of digastric muscle
10 **Masseter muscle** (severed)
11 Mandible
12 Sternocleidomastoid muscle
13 Orbicularis oculi muscle
14 Orbicularis oris muscle
15 Buccinator muscle
16 Depressor anguli oris muscle
17 Depressor labii inferioris muscle
18 Platysma muscle
19 **Articular disc of temporomandibular joint**
20 Head of mandible
21 Anterior belly of digastric muscle
22 Mylohyoid muscle
23 Hyoid bone

Medial and lateral pterygoid muscles. A portion of the mandible and the zygomatic arch has been removed revealing the pterygoid region or infratemporal fossa.

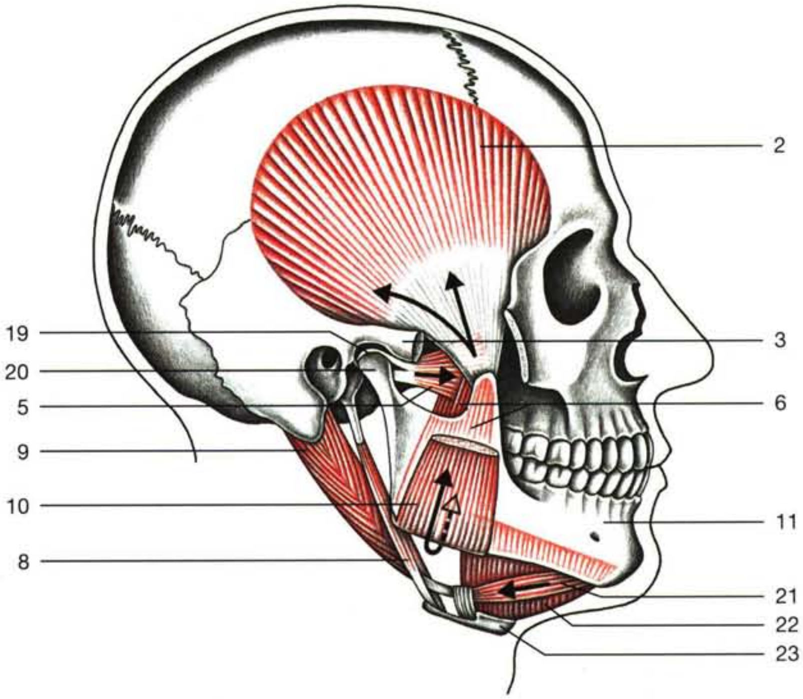

Effect of the muscles of mastication on the temporomandibular joint (arrows).

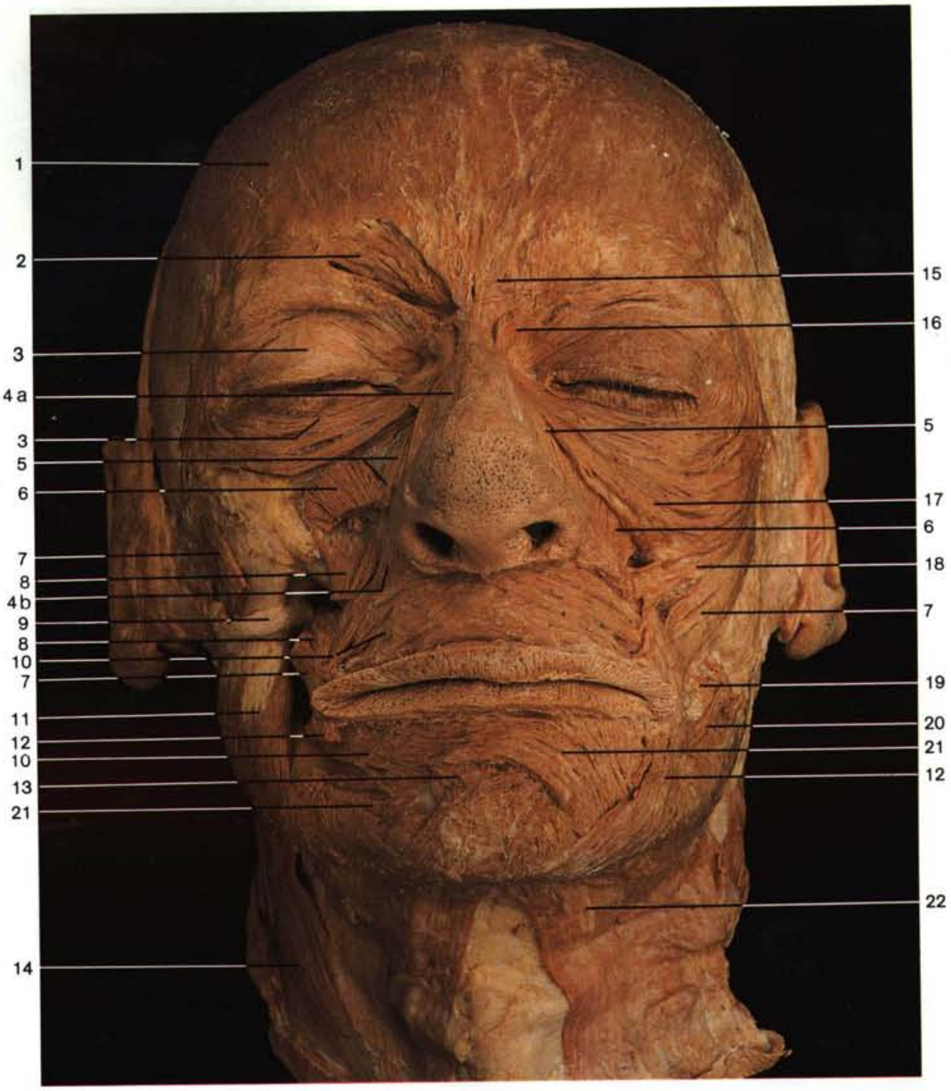

Facial muscles (anterior aspect). Left side: superficial layer; right side: deeper layer.

1 Frontal belly of **occipitofrontalis muscle**
2 **Corrugator supercilii muscle**
3 Palpebral part of orbicularis oculi muscle
4a Transverse part of nasalis muscle
4b Alar part of **nasalis muscle**
5 Levator labii superioris alaeque nasi muscle
6 **Levator labii superioris muscle**
7 **Zygomaticus major muscle**
8 Levator anguli oris muscle
9 Parotid duct
10 **Orbicularis oris muscle**
11 Masseter muscle
12 **Depressor anguli oris muscle**
13 **Mentalis muscle**
14 Sternocleidomastoid muscle
15 **Procerus muscle**
16 **Depressor supercilii muscle**
17 Orbital part of **orbicularis oculi muscle**
18 Zygomaticus minor muscle
19 **Buccinator muscle**
20 Risorius muscle
21 **Depressor labii inferioris muscle**
22 **Platysma muscle**
23 Galea aponeurotica
24 **Temporoparietalis muscle**
25 Occipital belly of occipitofrontalis muscle
26 Parotid gland with fascia
27 Temporal fascia
28 **Orbicularis oculi muscle**
29 Parotid duct, masseter muscle

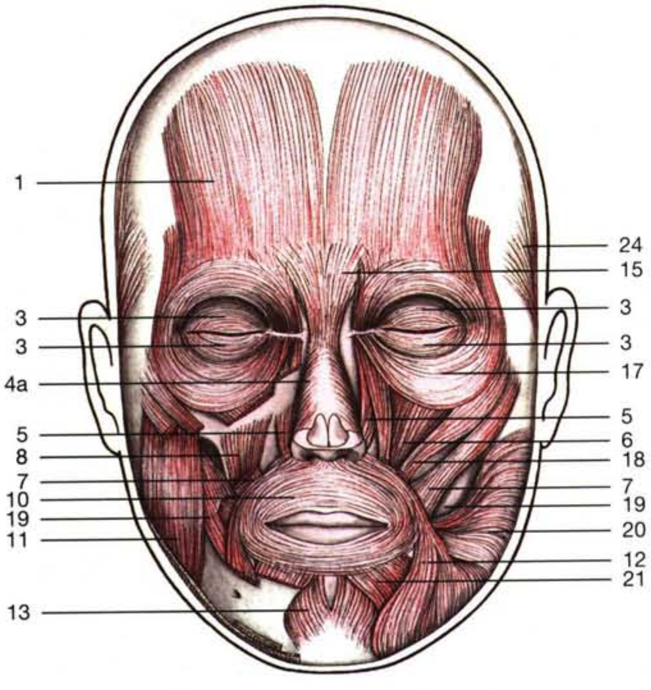

Facial muscles (schematic drawing). Left side: superficial layer; right side: deeper layer.

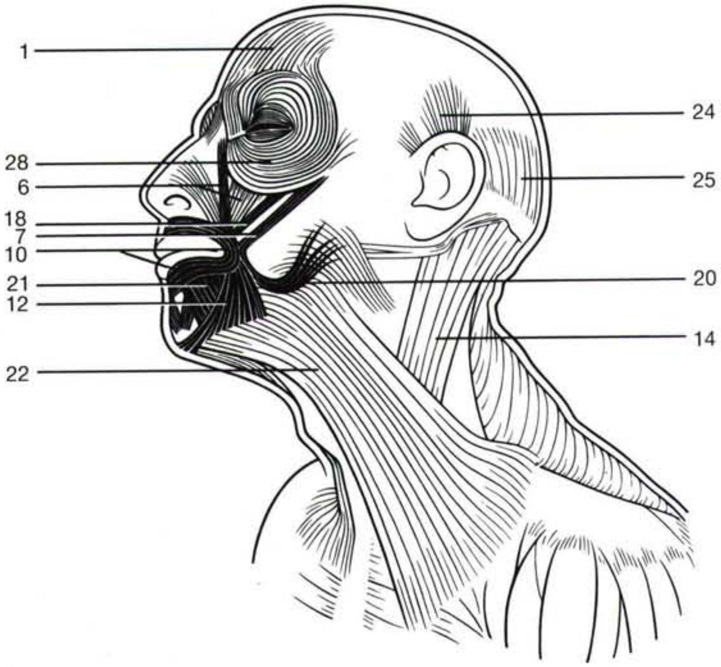

Facial muscles. Sphincter-like muscles surround the orifices of the head. Radially arranged muscles work as their antagonists.

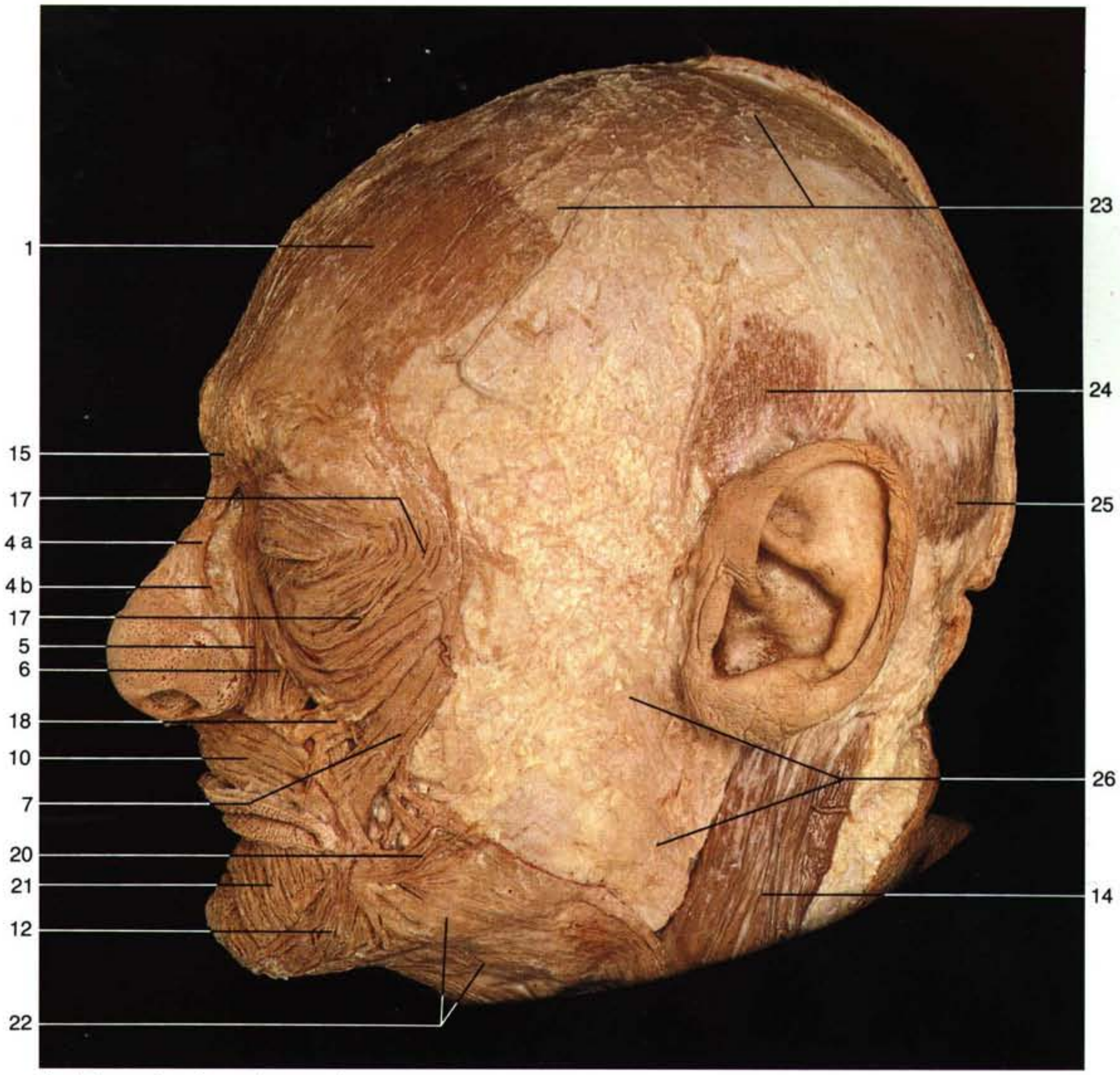
1
15
17
4 a
4 b
17
5
6
18
10
7
20
21
12
22
23
24
25
26
14

Facial muscles (lateral aspect).

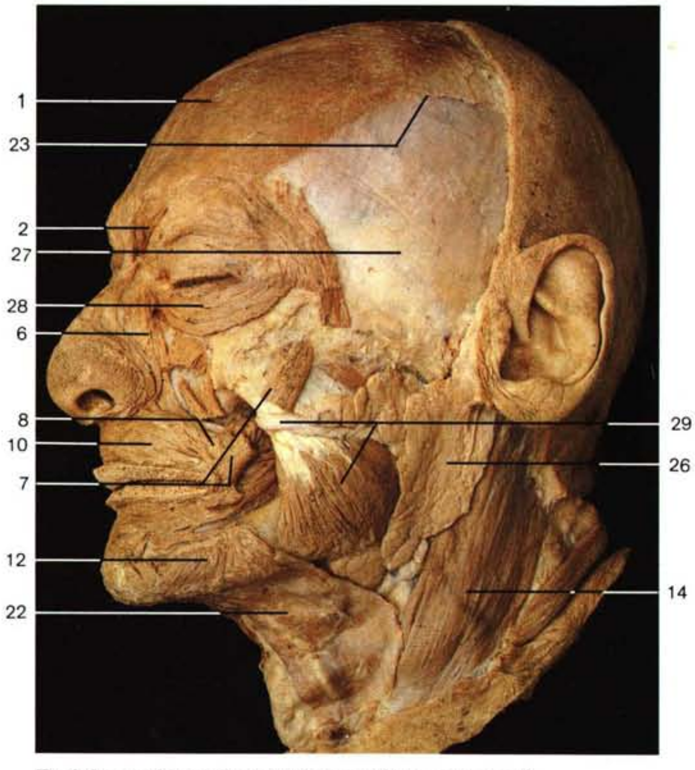
1
23
2
27
28
6
8
10
7
12
22
29
26
14

Facial muscles and parotid gland (lateral aspect).

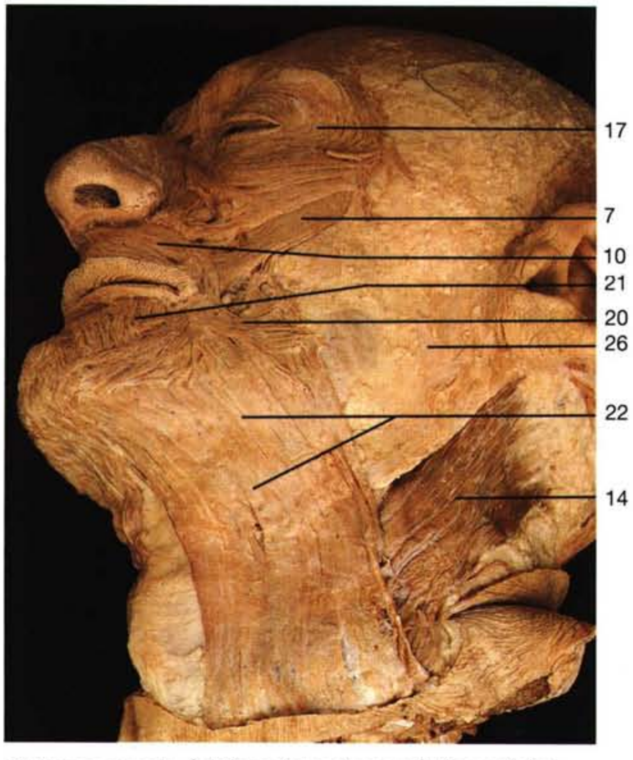
17
7
10
21
20
26
22
14

Platysma muscle (oblique lateral aspect). Superficial lamina of cervical fascia partly removed.

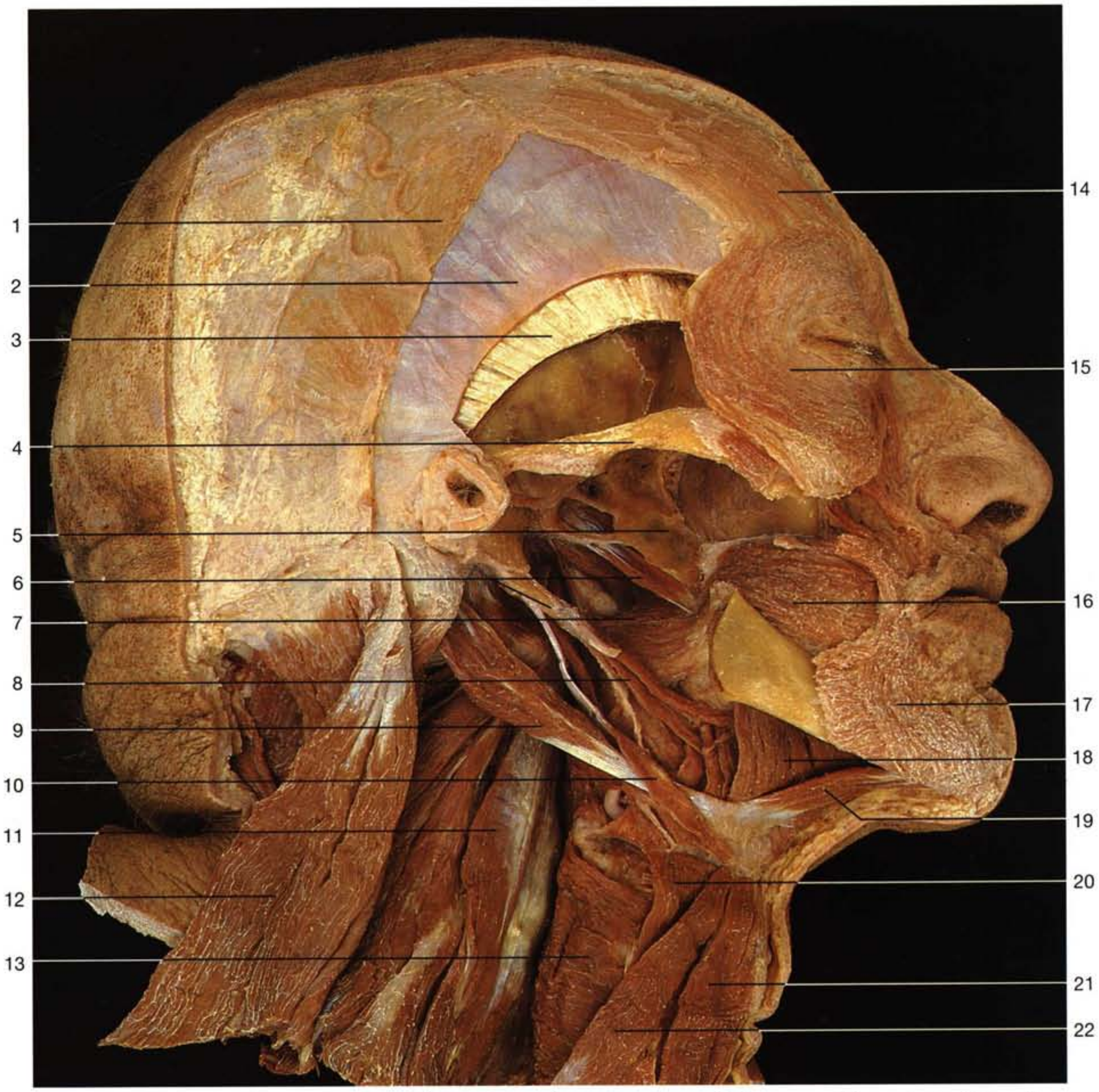

Supra- and infrahyoid muscles, pharynx I (lateral aspect). Ramus of mandible, pterygoid muscles and insertion of temporalis muscle removed.

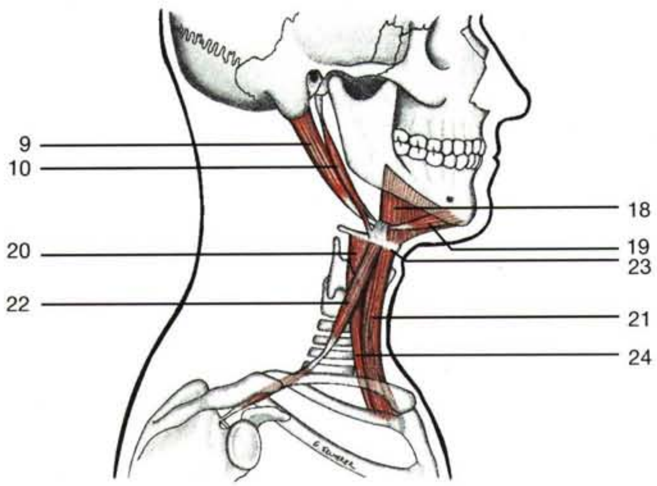

Supra- and infrahyoid muscles. (Schematic diagram.)

1	Galea aponeurotica
2	**Temporal fascia**
3	Tendon of temporalis muscle
4	Zygomatic arch
5	Lateral pterygoid plate
6	Tensor veli palatini muscle (styloid process)
7	Superior constrictor muscle of pharynx
8	**Styloglossus muscle**
9	Posterior belly of digastric muscle
10	**Stylohyoid muscle**
11	Longus capitis muscle
12	Sternocleidomastoid muscle (reflected)
13	Inferior constrictor of pharynx
14	Frontal belly of occipitofrontalis muscle
15	Orbital part of orbicularis oculi muscle
16	**Buccinator muscle**
17	Depressor anguli oris muscle
18	**Mylohyoid muscle**
19	Anterior belly of digastric muscle
20	**Thyrohyoid muscle**
21	**Sternohyoid muscle**
22	**Omohyoid muscle**
23	Hyoid bone
24	Sternothyroid muscle

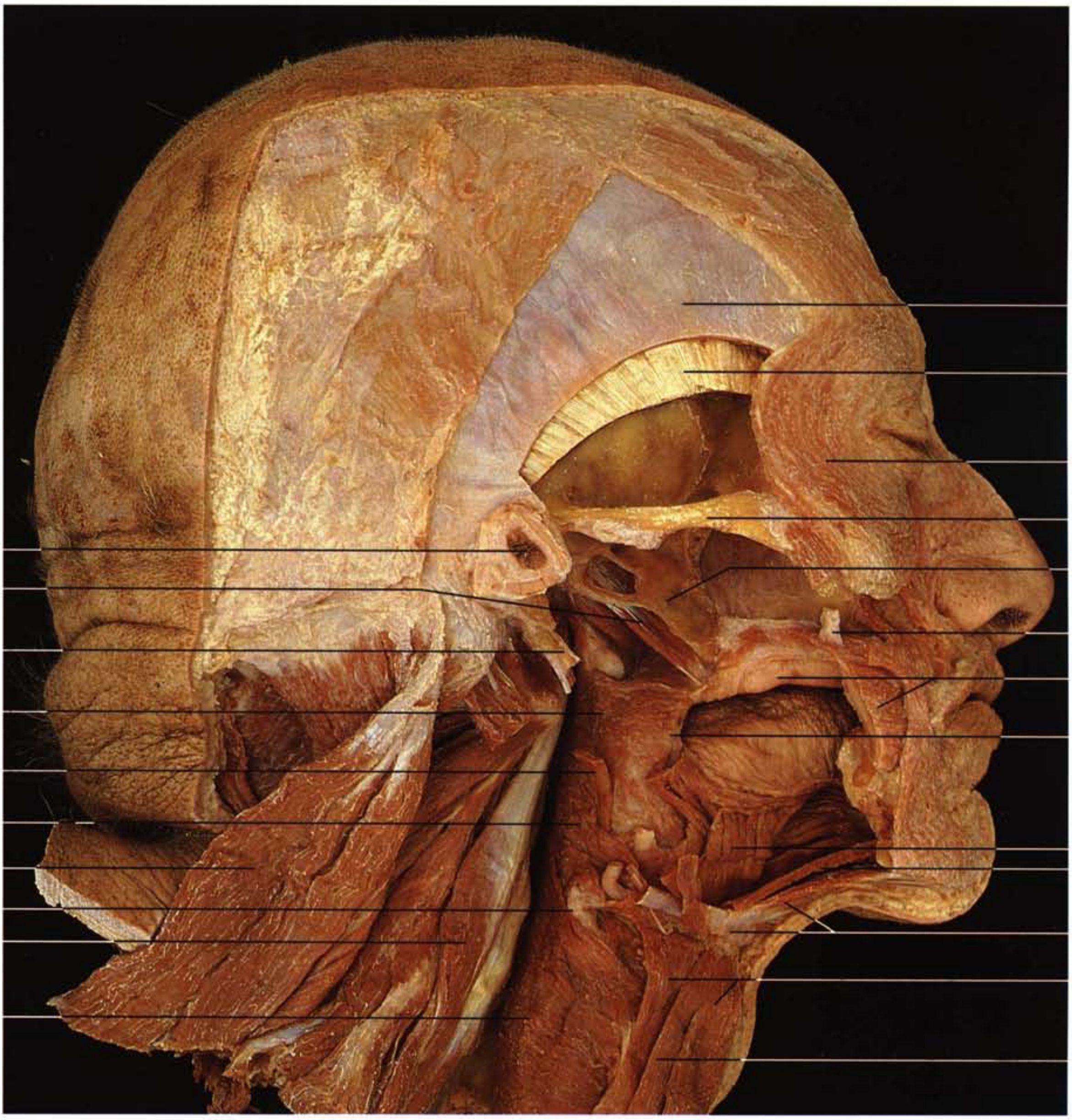

Supra- and infrahyoid muscles, pharynx II. Buccinator muscle removed; oral cavity opened.

1	External acoustic meatus	13	Orbicularis oculi muscle
2	Tensor veli palatini muscle	14	Zygomatic arch
3	Styloid process	15	Lateral pterygoid plate
4	**Superior constrictor muscle of pharynx**	16	Parotid duct
5	Stylopharyngeus muscle (divided)	17	Gingiva of upper jaw (without teeth), buccinator muscle (divided)
6	**Middle constrictor muscle of pharynx**	18	Pterygomandibular raphe
7	Sternocleidomastoid muscle	19	Hyoglossus muscle
8	Greater horn of hyoid bone	20	Mylohyoid muscle
9	Longus capitis	21	Anterior belly of digastric muscle (hyoid bone)
10	**Inferior constrictor muscle of pharynx**	22	Sternohyoid and thyrohyoid muscles
11	Temporal fascia	23	Omohyoid muscle
12	Tendon of temporalis muscle		

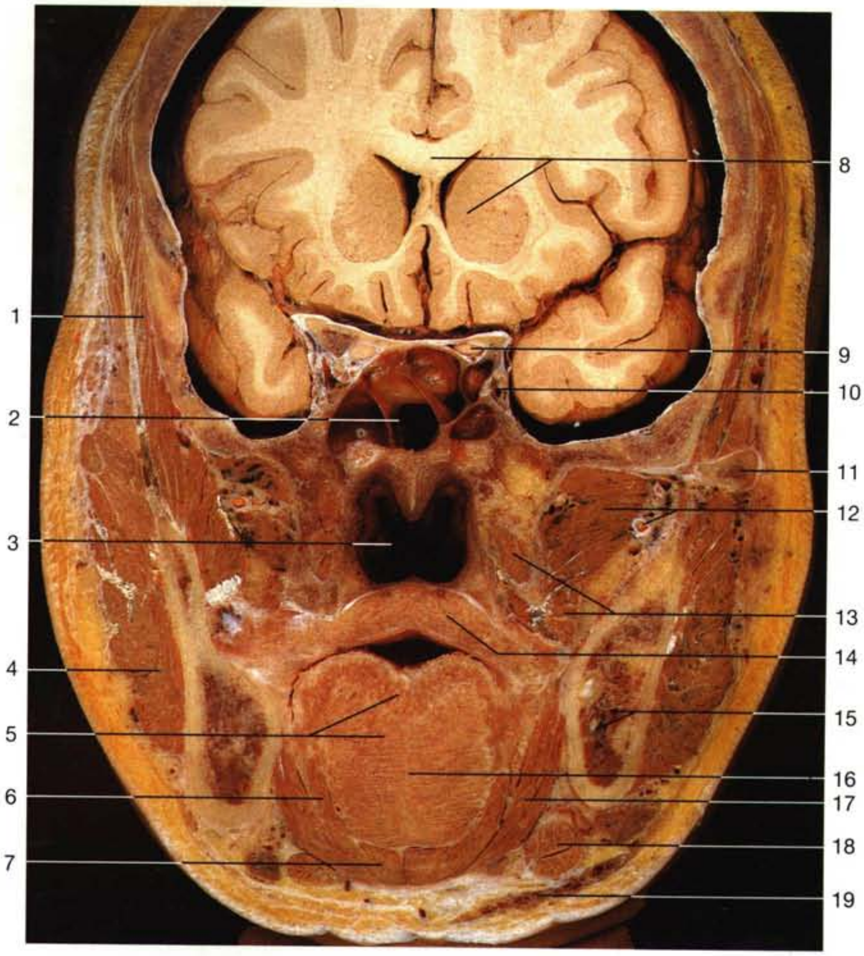

Coronal section through cranial, nasal and oral cavity at the level of sphenoidal sinus.

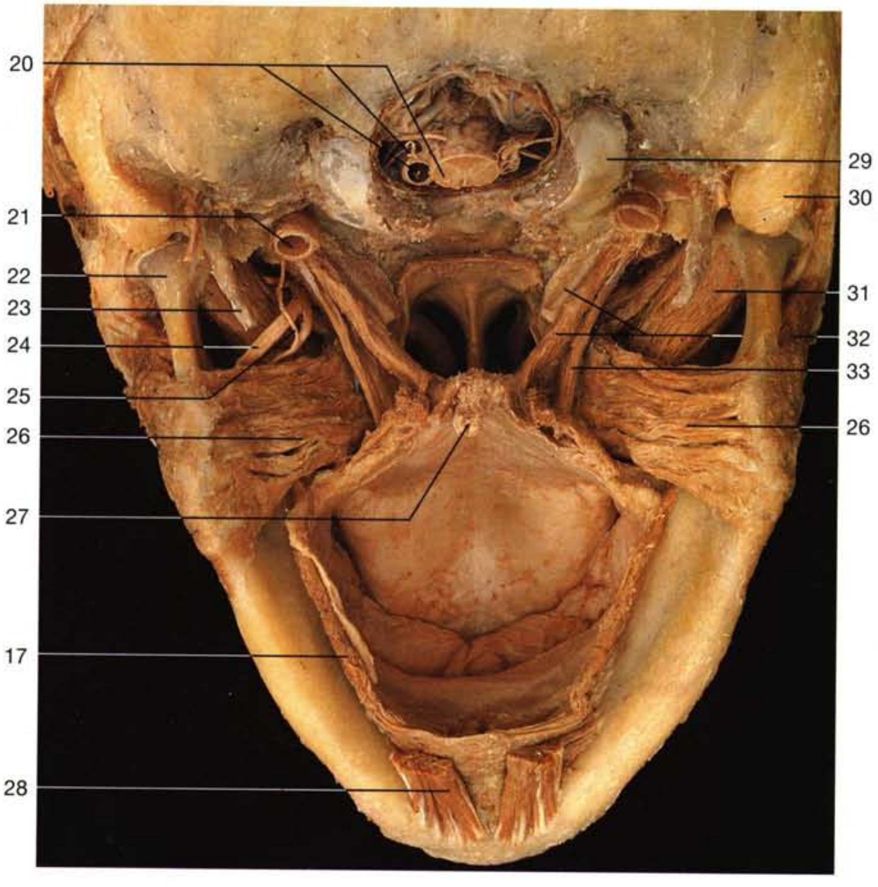

Pterygoid and palatine muscles (posterior aspect).

1 Temporalis muscle
2 **Sphenoidal sinus**
3 **Nasopharynx**
4 Masseter muscle
5 Superior longitudinal, transverse and vertical muscles of tongue
6 Hyoglossus muscle
7 Geniohyoid muscle
8 Corpus callosum (caudate nucleus)
9 Optic nerve
10 Cavernous sinus
11 Zygomatic arch
12 Cross section of lateral pterygoid muscle and maxillary artery
13 Section of medial pterygoid muscle
14 **Soft palate**
15 Mandible and inferior alveolar nerve
16 Septum of the tongue
17 **Mylohyoid muscle**
18 Submandibular gland
19 Platysma muscle
20 Foramen magnum, vertebral artery and spinal cord
21 Internal carotid artery
22 **Head of mandible**
23 Styloid process
24 Inferior alveolar nerve
25 Lingual nerve and chorda tympani nerve
26 **Medial pterygoid muscle**
27 Uvula
28 Anterior belly of digastric muscle (cut)
29 Condyle of occipital bone
30 Mastoid process
31 **Lateral pterygoid muscle**
32 Auditory tube and **levator veli palatini muscle**
33 **Tensor veli palatini muscle**

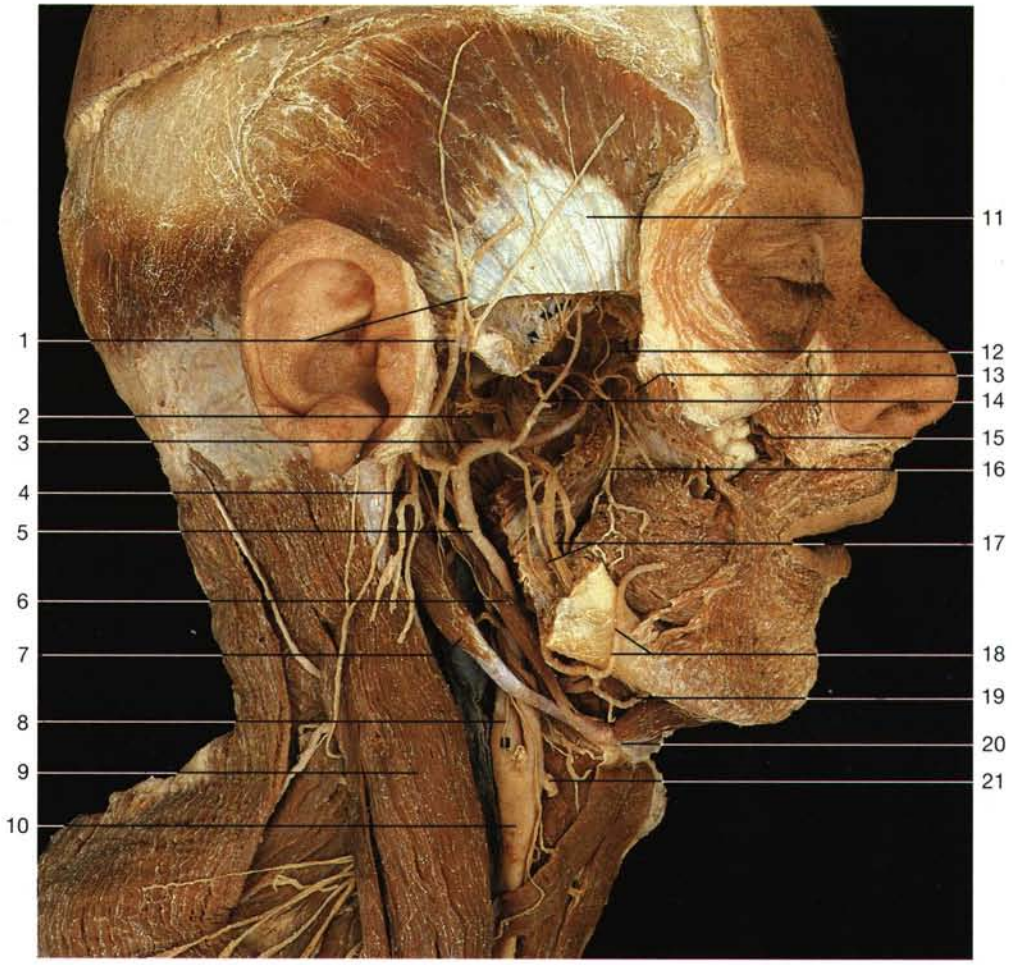

Dissection of maxillary artery. Mandible and lateral pterygoid muscle partly removed.

1 Superficial temporal artery and auriculotemporal nerve
2 Middle meningeal artery
3 **Maxillary artery**
4 Facial nerve (divided and reflected)
5 **External carotid artery**
6 Stylohyoid muscle
7 Digastric muscle (posterior belly)
8 Internal carotid artery and carotid sinus branch of glossopharyngeal nerve
9 Sternocleidomastoid muscle
10 Common carotid artery
11 Temporalis muscle
12 Pterygopalatine fossa
13 Posterior superior alveolar artery
14 Deep temporal artery
15 Infraorbital artery
16 Buccal nerve
17 Inferior alveolar artery and nerve
18 **Facial artery**
19 Submental artery
20 Hyoid bone
21 Superior thyroid artery (divided)

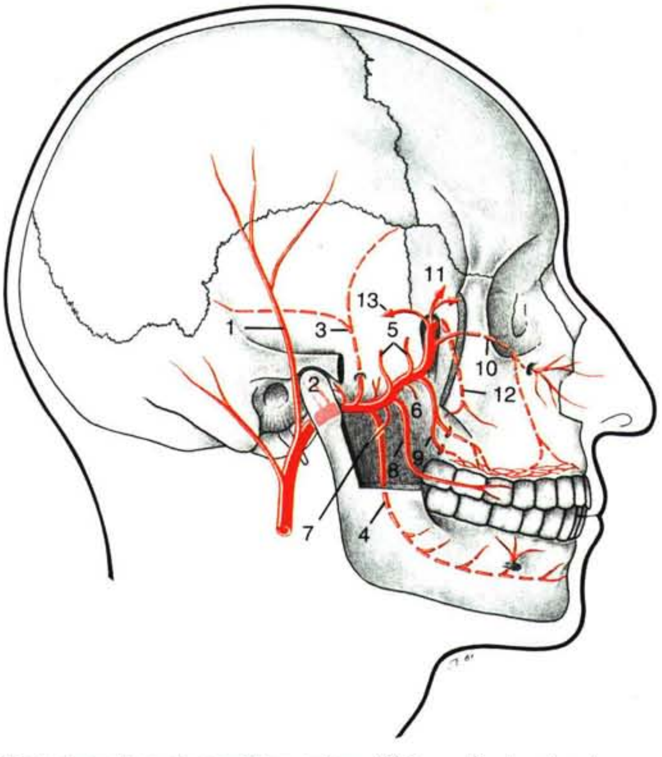

Main branches of maxillary artery. (Schematic drawing.)

1 Superficial temporal artery

Branches of the first part
2 Deep auricular artery and anterior tympanic artery
3 Middle meningeal artery
4 Inferior alveolar artery

Branches of the second part
5 Deep temporal branches
6 Pterygoid branches
7 Masseteric artery
8 Buccal artery

Branches of the third part
9 Posterior superior alveolar artery
10 Infraorbital artery
11 Sphenopalatine artery and branches to the nasal cavity
12 Descending palatine artery
13 Artery of the pterygoid canal

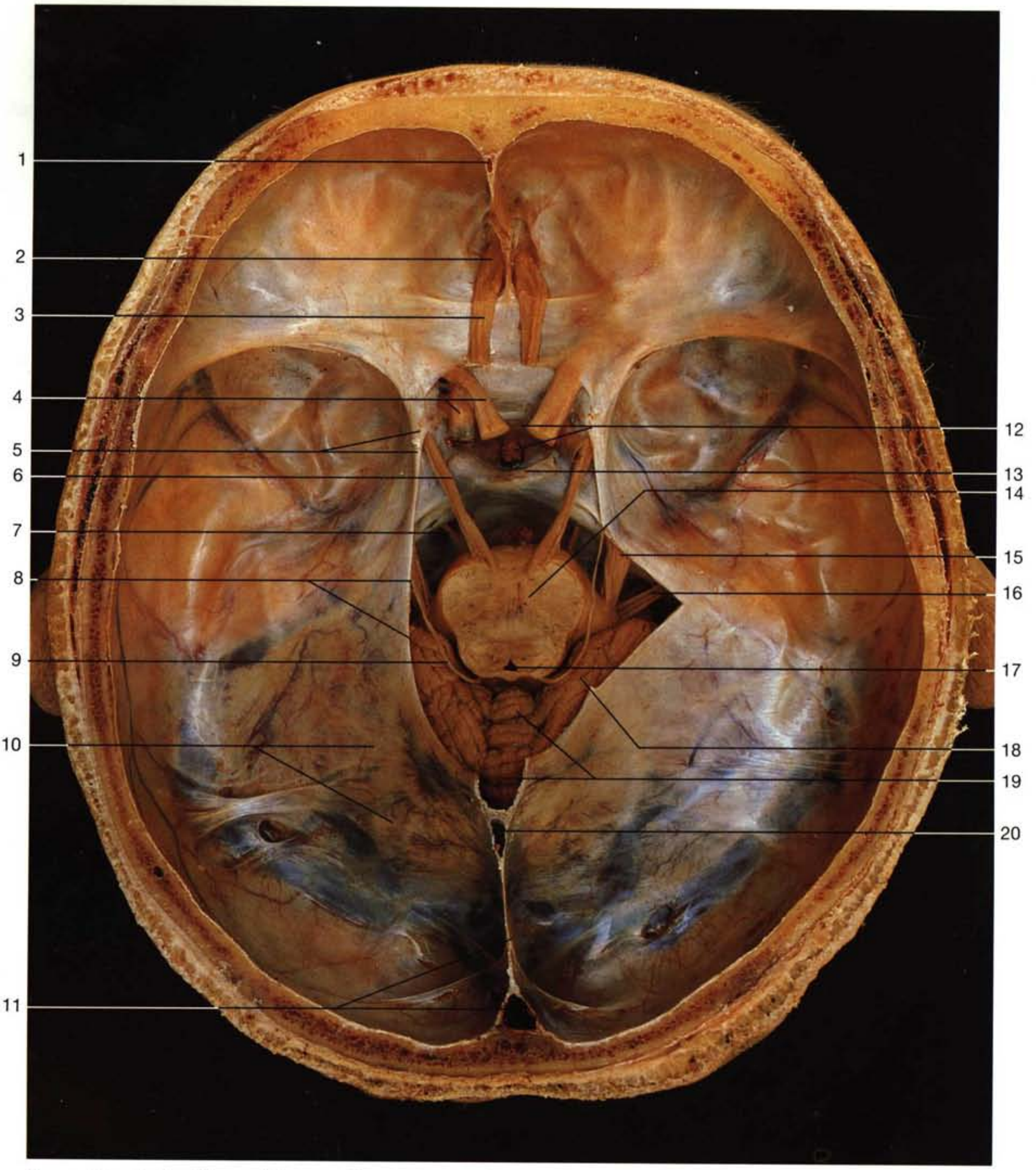

Base of the skull with cranial nerves (internal aspect). Both cerebral hemispheres and upper part of the brain stem removed. Incision on the right tentorium cerebelli to display the cranial nerves of the infratentorial space.

1	Superior sagittal sinus with falx cerebri	12	Hypophysial fossa, infundibulum, and diaphragma sellae
2	Olfactory bulb	13	Dorsum sellae
3	**Olfactory tract**	14	Midbrain (divided)
4	**Optic nerve** and internal carotid artery	15	**Trigeminal nerve** (n. V)
5	Anterior clinoid process and anterior attachment of tentorium cerebelli	16	**Facial nerve** (n. VII), nervus intermedius, and **vestibulocochlear nerve** (n. VIII)
6	**Oculomotor nerve** (n. III)	17	Cerebral aqueduct
7	**Abducens nerve** (n. VI)	18	Right hemisphere of cerebellum
8	Tentorial notch (incisura tentorii)	19	Vermis of cerebellum
9	**Trochlear nerve** (n. IV)	20	Straight sinus
10	Tentorium cerebelli		
11	Falx cerebri and confluence of sinuses		

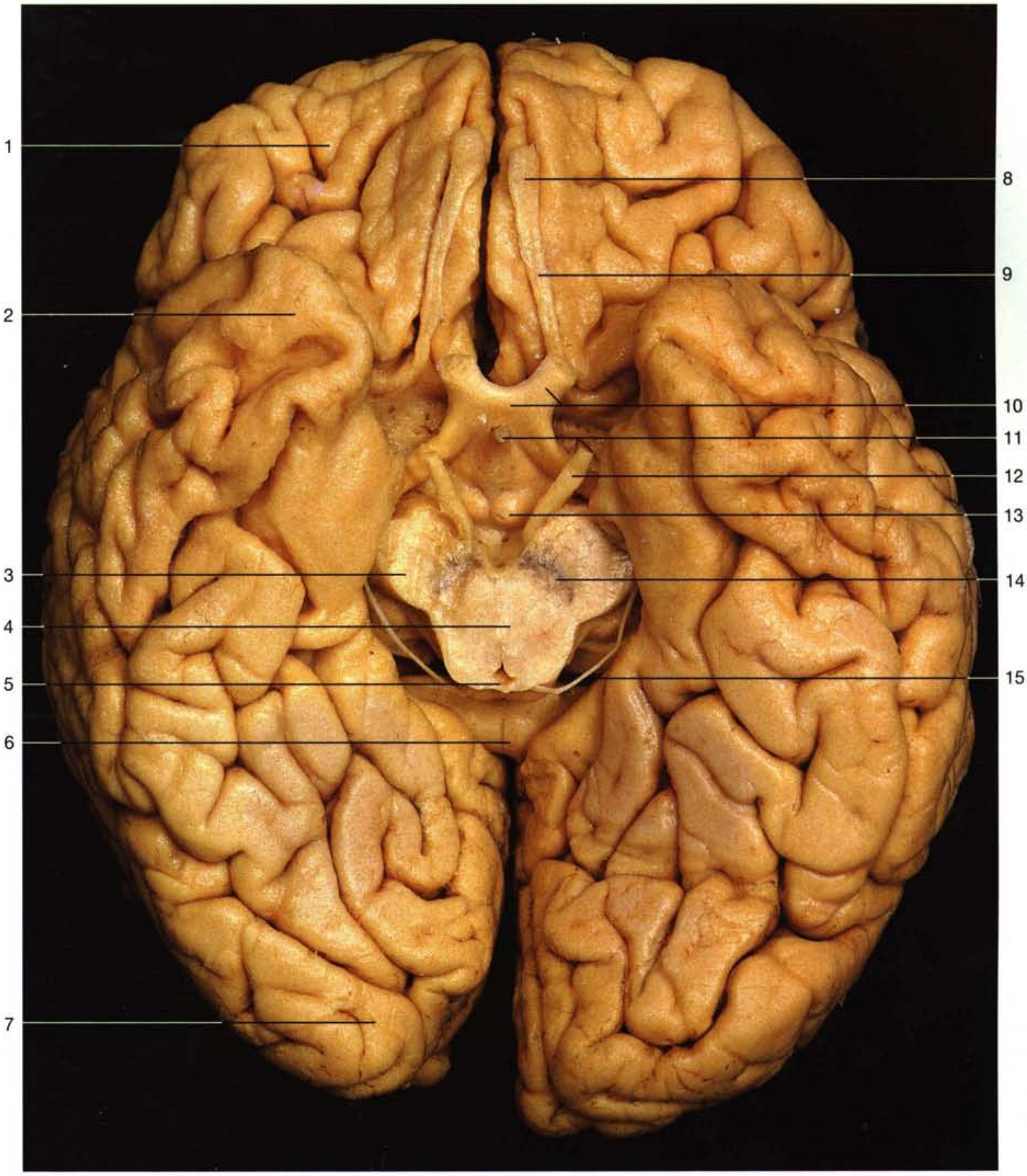

Inferior aspect of the brain with cranial nerves. Midbrain divided.

1 **Frontal lobe**
2 **Temporal lobe**
3 Pedunculus cerebri
4 **Midbrain** (divided)
5 Cerebral aqueduct
6 Splenium of corpus callosum
7 **Occipital lobe**
8 Olfactory bulb
9 Olfactory tract
10 **Optic nerve and optic chiasma**
11 Infundibulum
12 **Oculomotor nerve** (n. III)
13 Mamillary body
14 Substantia nigra
15 **Trochlear nerve** (n. IV)

Cranial nerves

I = Olfactory nerves
II = Optic nerve
III = Oculomotor nerve
IV = Trochlear nerve
V = Trigeminal nerve
VI = Abducens nerve

VII = Facial nerve
VIII = Vestibulocochlear nerve
IX = Glossopharyngeal nerve
X = Vagus nerve
XI = Accessory nerve
XII = Hypoglossal nerve

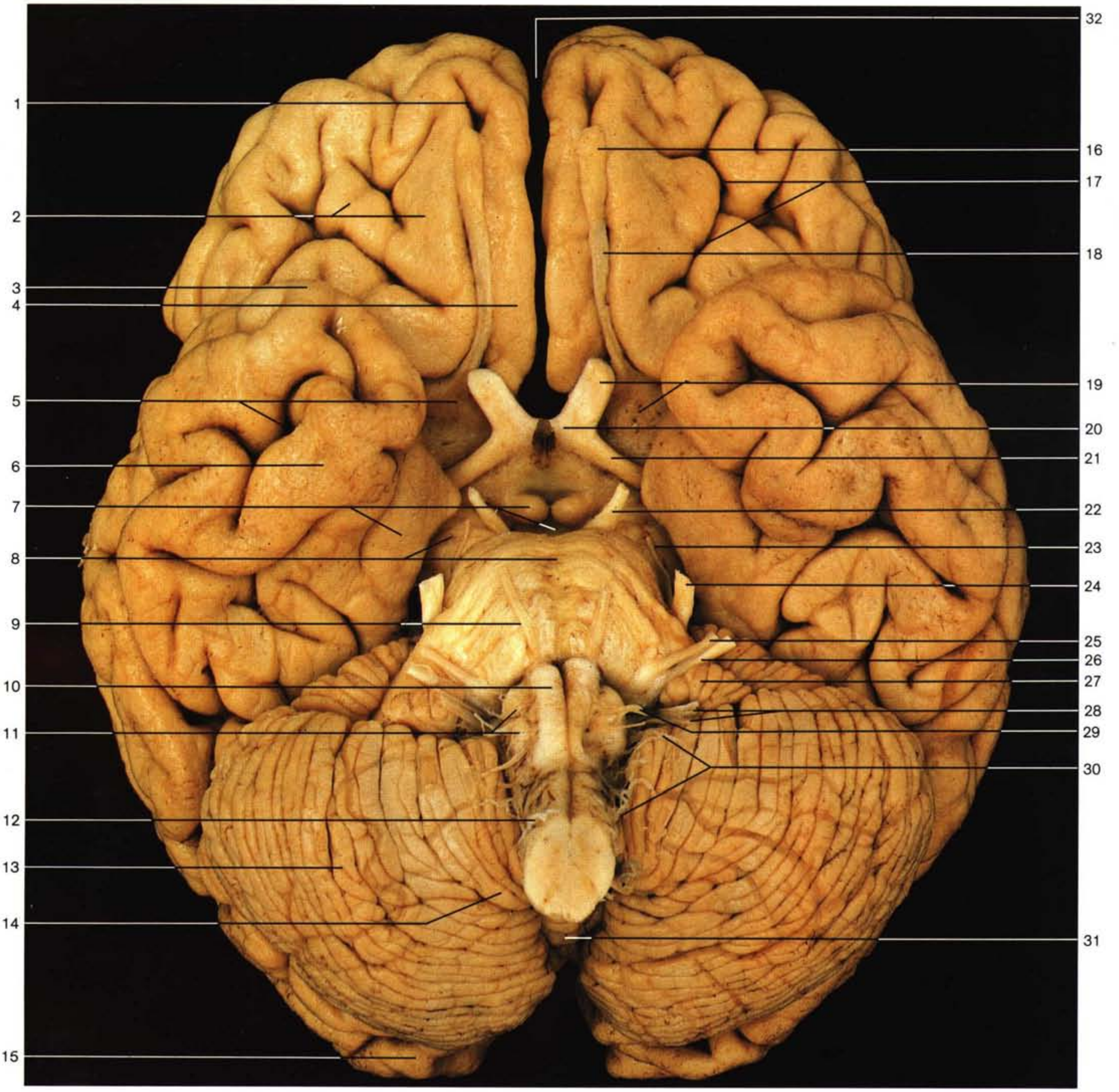

Cranial nerves. Brain (inferior aspect).

1 Olfactory sulcus (termination)
2 Orbital gyri
3 Temporal lobe
4 Straight gyrus
5 Olfactory trigone and inferior temporal sulcus
6 Medial occipitotemporal gyrus
7 Parahippocampal gyrus, mamillary body, and interpeduncular fossa
8 Pons and cerebral peduncle
9 **Abducens nerve** (n. VI)
10 Pyramid
11 Lower part of olive
12 Cervical spinal nerves
13 Cerebellum
14 Tonsil of cerebellum
15 Occipital lobe (posterior pole)
16 Olfactory bulb
17 Orbital sulci of frontal lobe
18 **Olfactory tract**
19 **Optic nerve** (n. II) and anterior perforated substance
20 Optic chiasma
21 Optic tract
22 **Oculomotor nerve** (n. III)
23 **Trochlear nerve** (n. IV)
24 **Trigeminal nerve** (n. V)
25 **Facial nerve** (n. VII)
26 **Vestibulocochlear nerve** (n. VIII)
27 Flocculus of cerebellum
28 **Glossopharyngeal** (n. IX) and **vagus nerve** (n. X)
29 **Hypoglossal nerve** (n. XII)
30 **Accessory nerve** (n. XI)
31 Vermis of cerebellum
32 Longitudinal fissure

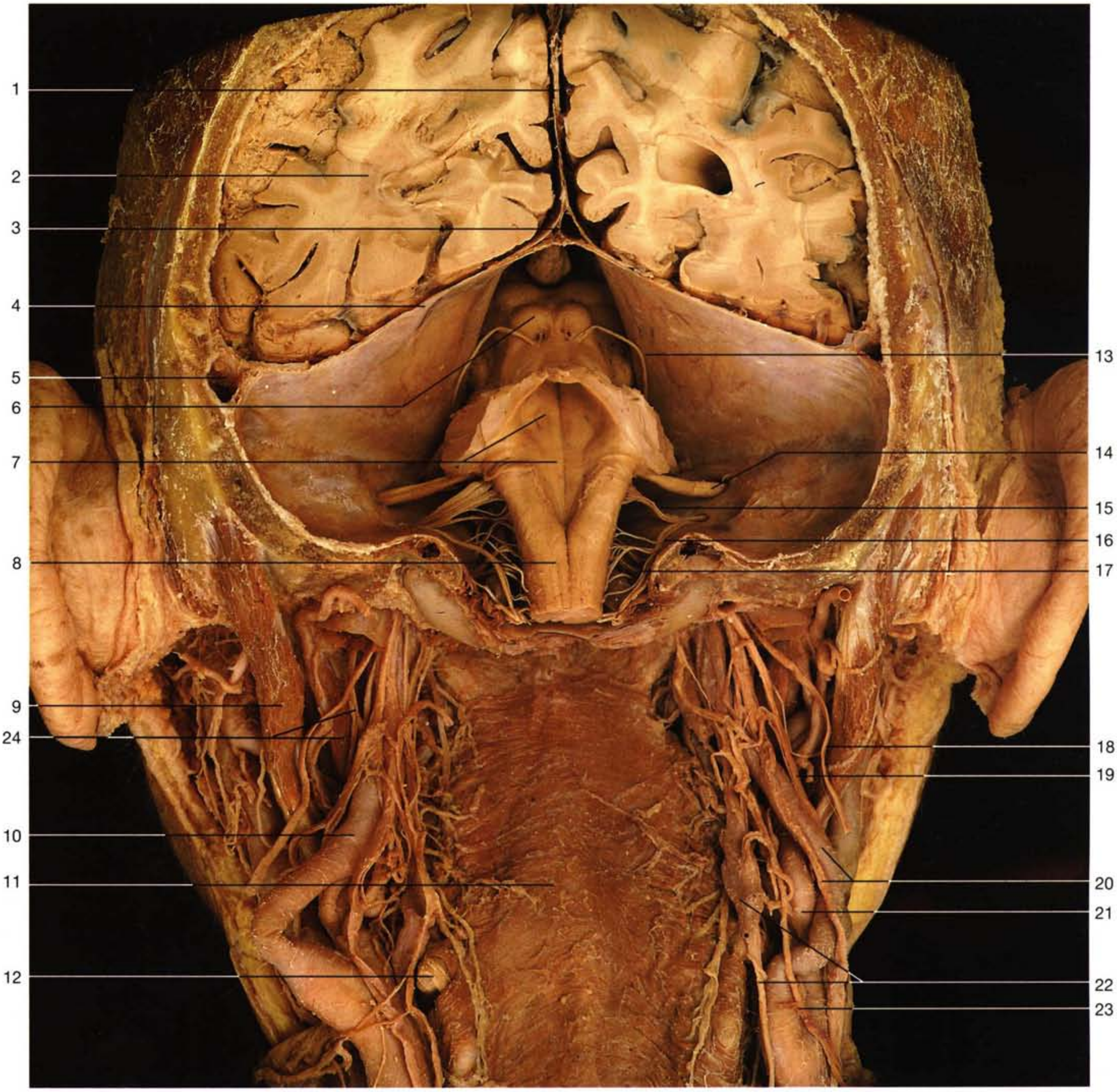

Brain stem and pharynx with cranial nerves (posterior aspect). Cranial cavity opened and cerebellum removed.

1 Falx cerebri
2 Occipital lobe
3 Straight sinus
4 Tentorium cerebelli
5 Transverse sinus
6 Inferior colliculus of midbrain
7 **Rhomboid fossa**
8 Medulla oblongata
9 Posterior belly of digastric muscle
10 Internal carotid artery
11 Pharynx (middle constrictor muscle)
12 Hyoid bone (greater horn)
13 **Trochlear nerve** (n. IV)
14 **Facial nerve** (n. VII), **vestibulocochlear nerve** (n. VIII)
15 **Glossopharyngeal nerve** (n. IX) and vagus nerve (n. X)
16 **Accessory nerve** (intracranial portion) (n. XI)
17 **Hypoglossal nerve** (intracranial portion) (n. XII)
18 **Accessory nerve** (n. XI)
19 **Hypoglossal nerve** (n. XII)
20 **Vagus nerve** (n. X) and internal carotid artery
21 External carotid artery
22 **Sympathetic trunk** and superior cervical ganglion
23 Ansa cervicalis (superior root of hypoglossal nerve)
24 **Glossopharyngeal nerve** (n. IX) and stylopharyngeus muscle

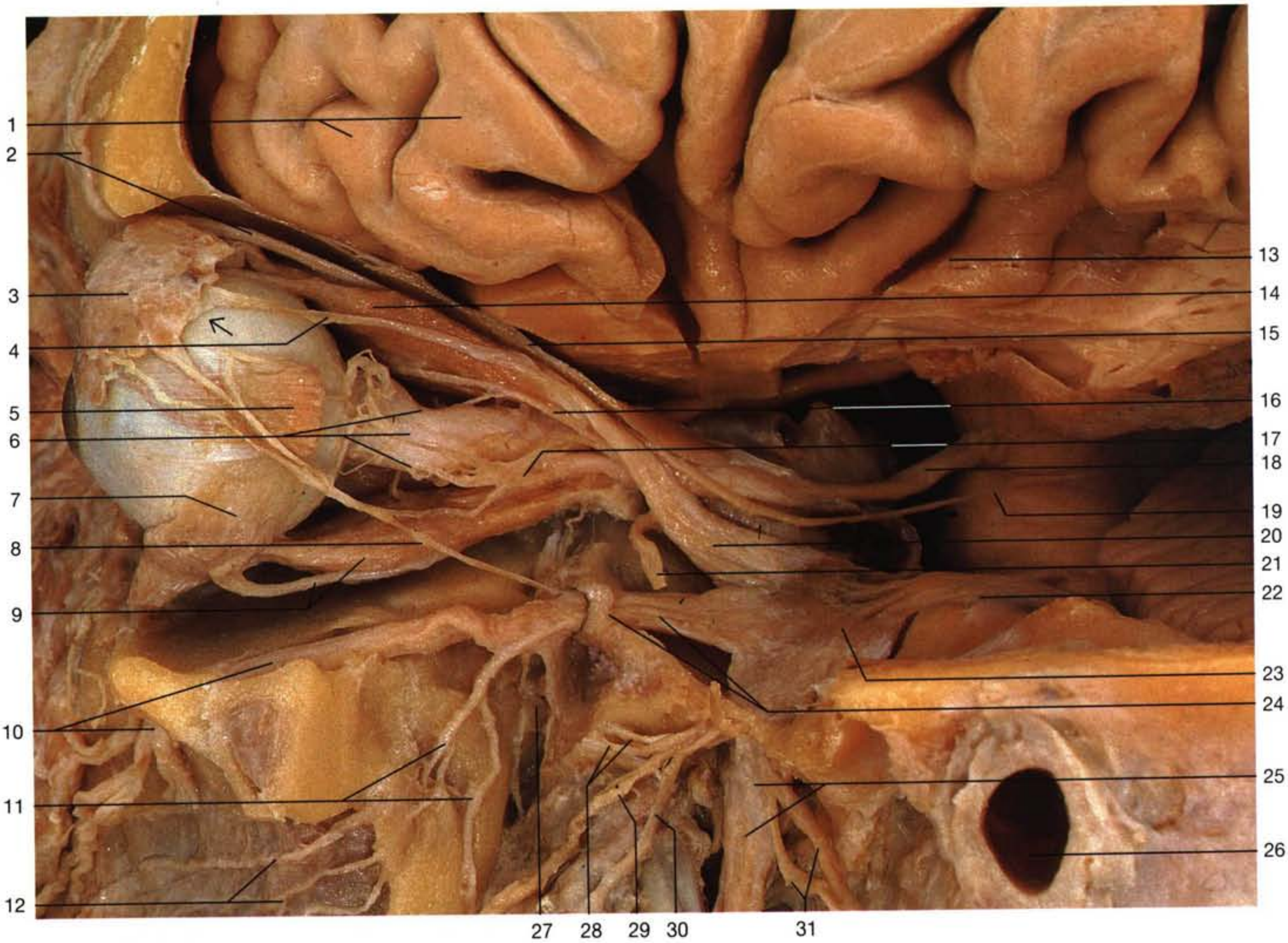

Cranial nerves of the orbit and pterygopalatine fossa. Left orbit (lateral aspect).
Note the zygomaticolacrimal anastomosis (arrow).

1 Frontal lobe
2 Supraorbital nerve
3 Lacrimal gland
4 Lacrimal nerve
5 Lateral rectus muscle (divided)
6 **Optic nerve** and short ciliary nerves
7 Inferior oblique muscle
8 Zygomatic nerve
9 Inferior branch of oculomotor nerve and inferior rectus muscle
10 Infraorbital nerve
11 Posterior superior alveolar nerves
12 Branches of superior alveolar plexus adjacent to mucous membrane of maxillary sinus
13 Central sulcus of insula
14 Superior rectus muscle
15 Periorbita (roof of orbit)
16 Nasociliary nerve
17 Ciliary ganglion
18 **Oculomotor nerve** (n. III)
19 **Trochlear nerve** (n. IV)
20 **Ophthalmic nerve** (n. V_1)
21 **Abducens nerve** (n. VI) (divided)
22 **Trigeminal nerve** (n. V)
23 Trigeminal ganglion
24 **Maxillary nerve** (n. V_2) and foramen rotundum
25 **Mandibular nerve** (n. V_3)
26 External acoustic meatus
27 Pterygopalatine nerves
28 Deep temporal nerves
29 Buccal nerve
30 Masseteric nerve
31 Auriculotemporal nerve
32 Trochlea and superior oblique muscle

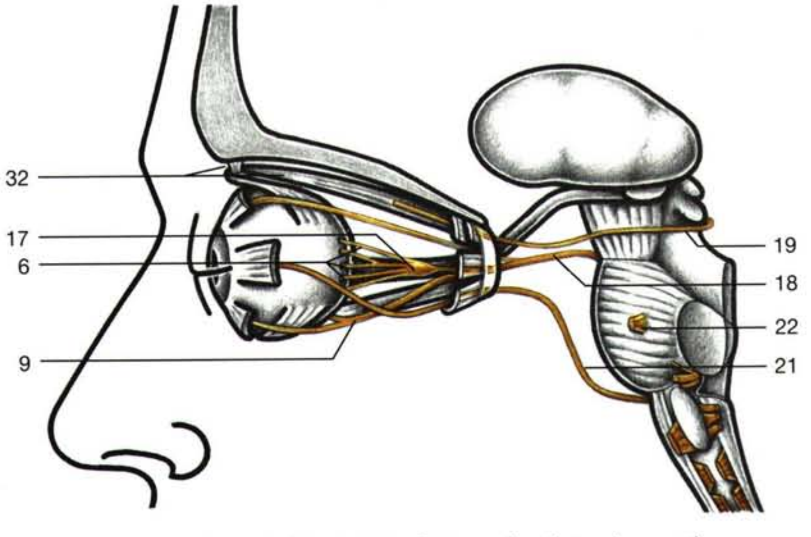

Cranial nerves innervating extraocular muscles (lateral aspect).
(Schematic drawing.)

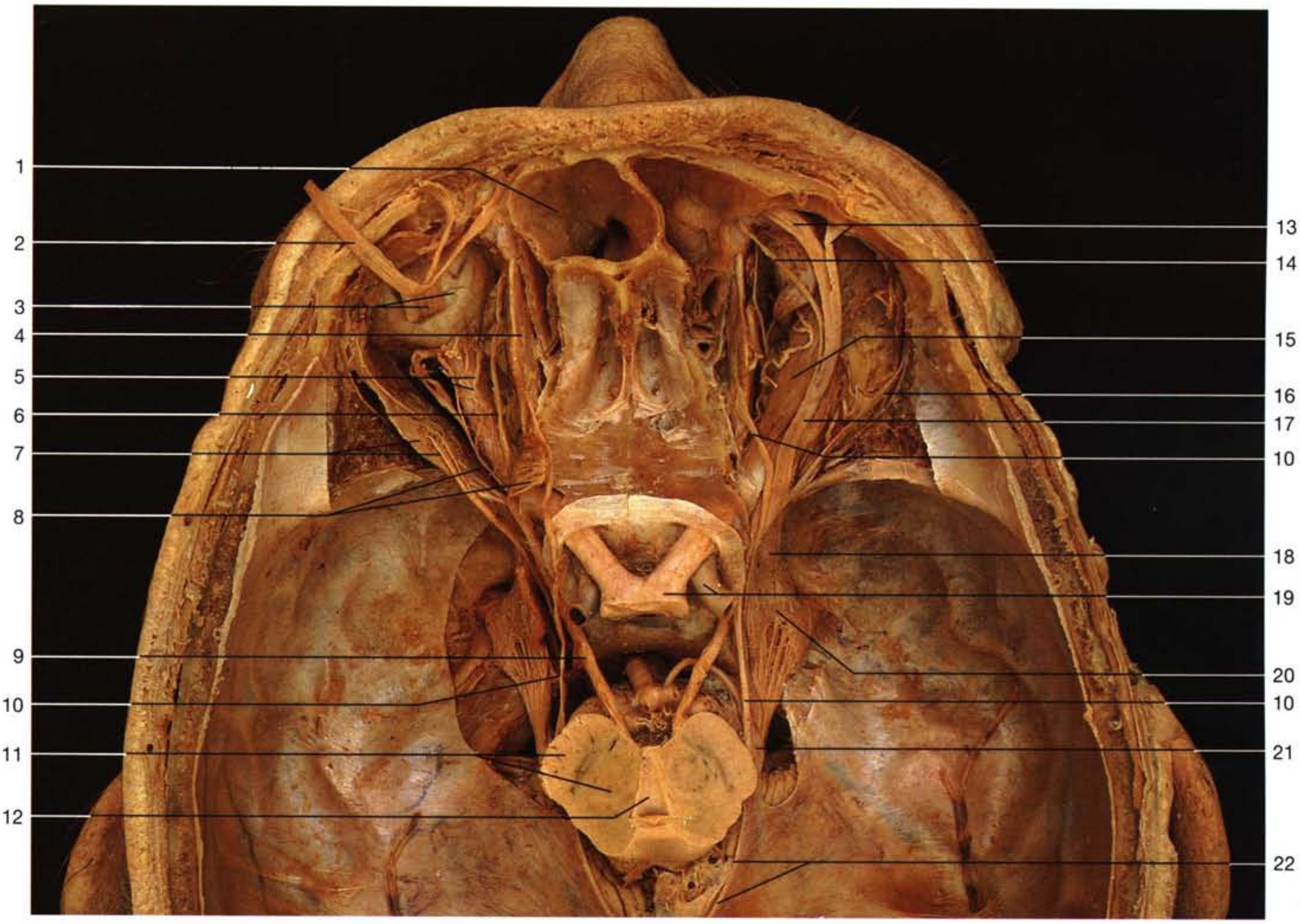

Cranial nerves of the orbit (superior aspect). Right side: superficial layer; left side: middle layer of the orbit (superior rectus muscle and frontal nerve divided and reflected). Tentorium and dura mater partly removed.

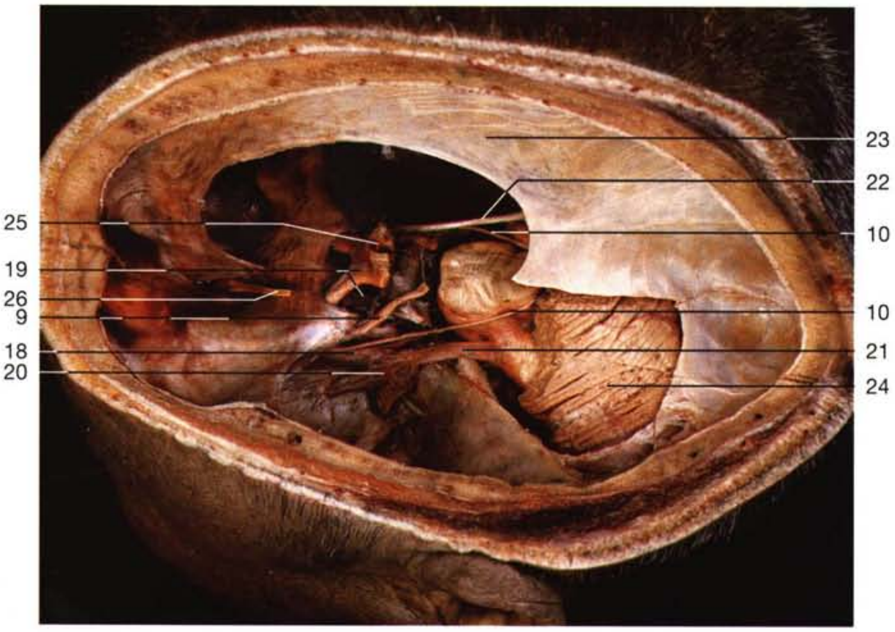

Cranial nerves at the skull base. The brain stem was divided and the tentorium fenestrated. Both hemispheres were removed.

1 Frontal sinus (enlarged)
2 Frontal nerve (divided and reflected)
3 Superior rectus muscle (divided) and eyeball
4 Superior oblique muscle
5 Short ciliary nerves and optic nerve (n. II)
6 Nasociliary nerve
7 **Abducens nerve** (n. VI) and lateral rectus muscle
8 Ciliary ganglion and superior rectus muscle (reflected)
9 **Oculomotor nerve** (n. III)
10 **Trochlear nerve** (n. IV)
11 Crus cerebri and midbrain
12 Inferior wall of the third ventricle connected with cerebral aqueduct
13 Lateral and medial branch of supraorbital nerve
14 Supratrochlear nerve
15 Superior rectus muscle
16 Lacrimal nerve
17 Frontal nerve
18 **Ophthalmic nerve** (n. V1)
19 Optic chiasma and internal carotid artery
20 Trigeminal ganglion
21 **Trigeminal nerve** (n. V)
22 Tentorial notch
23 Falx cerebri
24 Cerebellum
25 Infundibulum
26 Olfactory tract

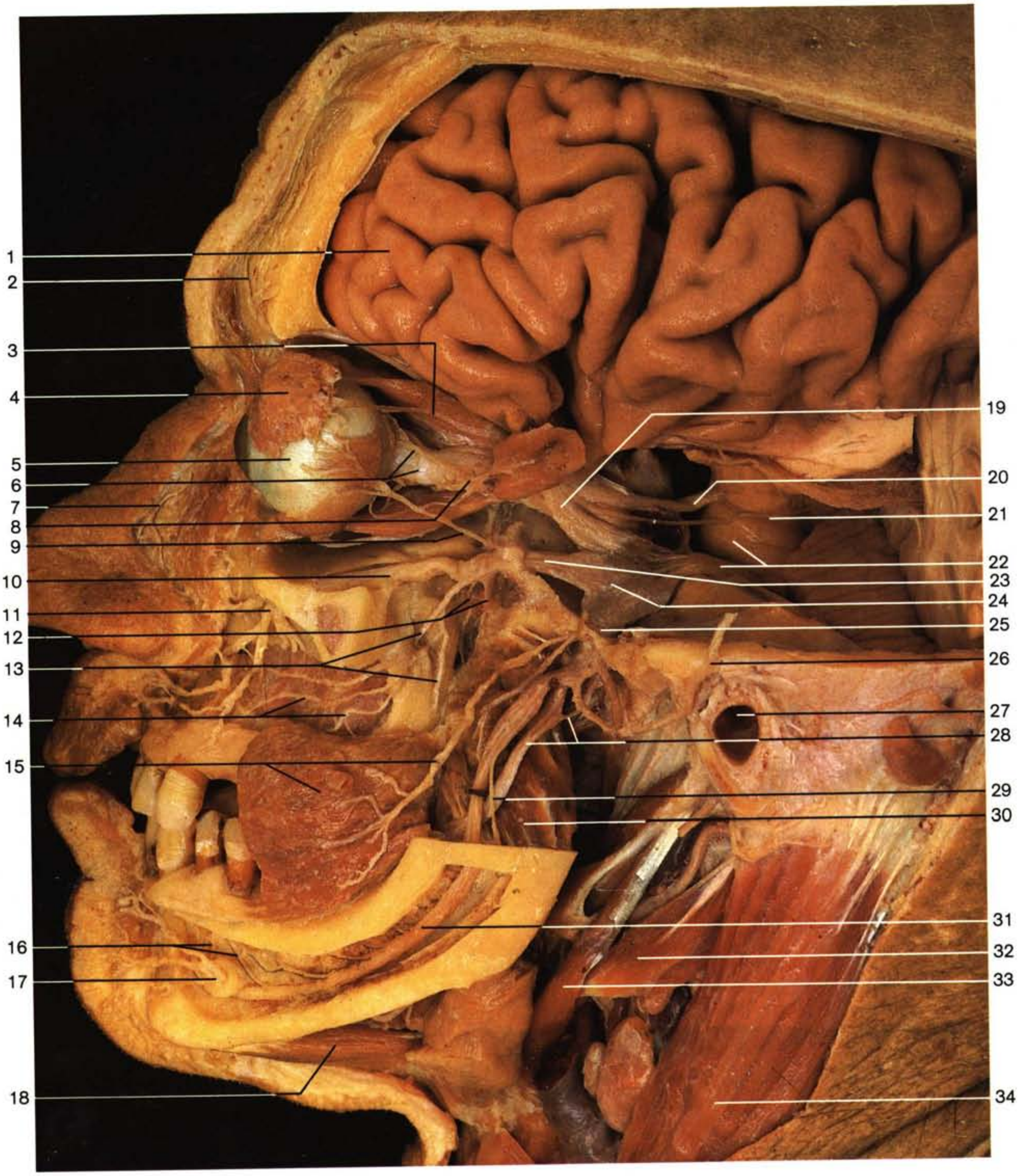

Dissection of the trigeminal nerve in its entirety. Lateral wall of cranial cavity, lateral wall of orbit, zygomatic arch and ramus of the mandible have been removed and the mandibular canal opened.

1 Frontal lobe of cerebrum
2 Supraorbital nerve
3 Lacrimal nerve
4 Lacrimal gland
5 Eyeball
6 **Optic nerve** and short ciliary nerves
7 External nasal branch of anterior ethmoidal nerve
8 Ciliary ganglion
9 Zygomatic nerve
10 Infraorbital nerve
11 Infraorbital foramen and terminal branches of infraorbital nerve
12 Pterygopalatine ganglion and pterygopalatine nerves
13 Posterior superior alveolar nerves
14 Superior dental plexus
15 Buccinator muscle and buccal nerve
16 Inferior dental plexus
17 Mental foramen and mental nerve
18 Anterior belly of digastric muscle
19 **Ophthalmic nerve** (n. V_1)
20 **Oculomotor nerve** (n. III)
21 **Trochlear nerve** (n. IV)
22 Trigeminal nerve and pons
23 **Maxillary nerve** (n. V_2)
24 Trigeminal ganglion
25 **Mandibular nerve** (n. V_3)
26 Auriculotemporal nerve
27 External acoustic meatus (divided)
28 Lingual nerve and chorda tympani
29 Mylohyoid nerve
30 Medial pterygoid muscle
31 Inferior alveolar nerve
32 Posterior belly of digastric muscle
33 Stylohyoid muscle
34 Sternocleidomastoid muscle

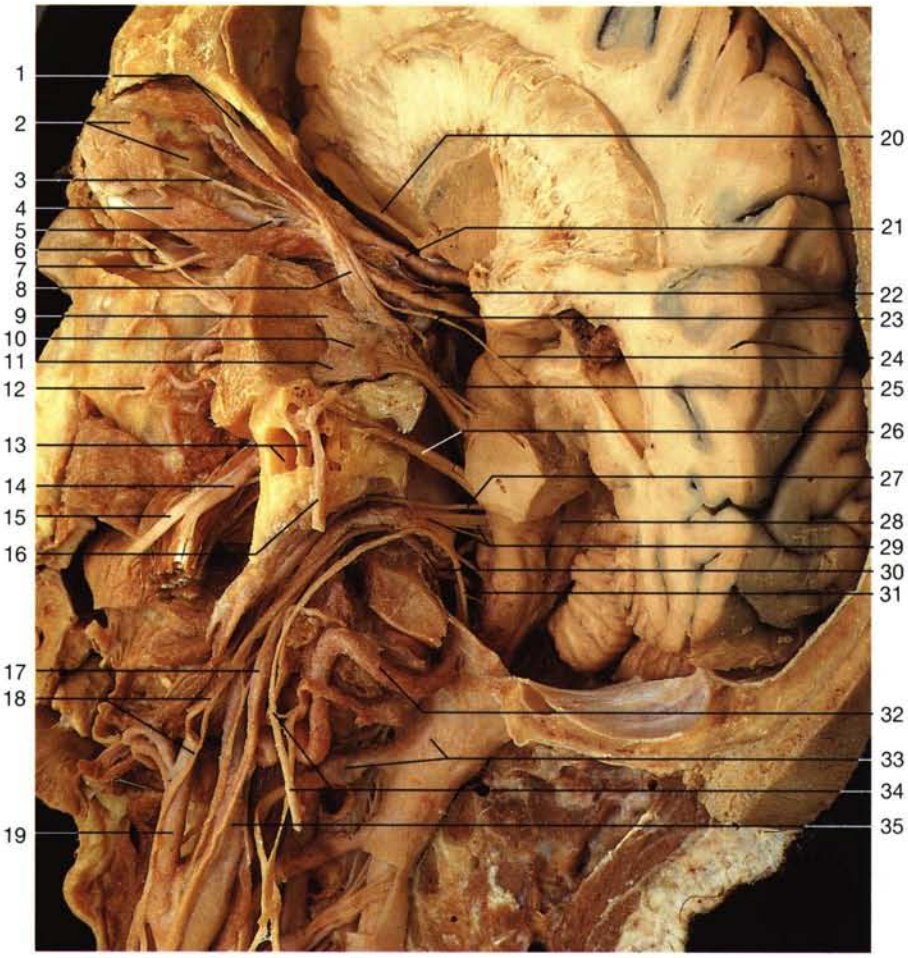

Cranial nerves in connection with the brain stem. Left side (lateral superior aspect). Left half of brain and head partly removed. Notice the location of trigeminal ganglion.

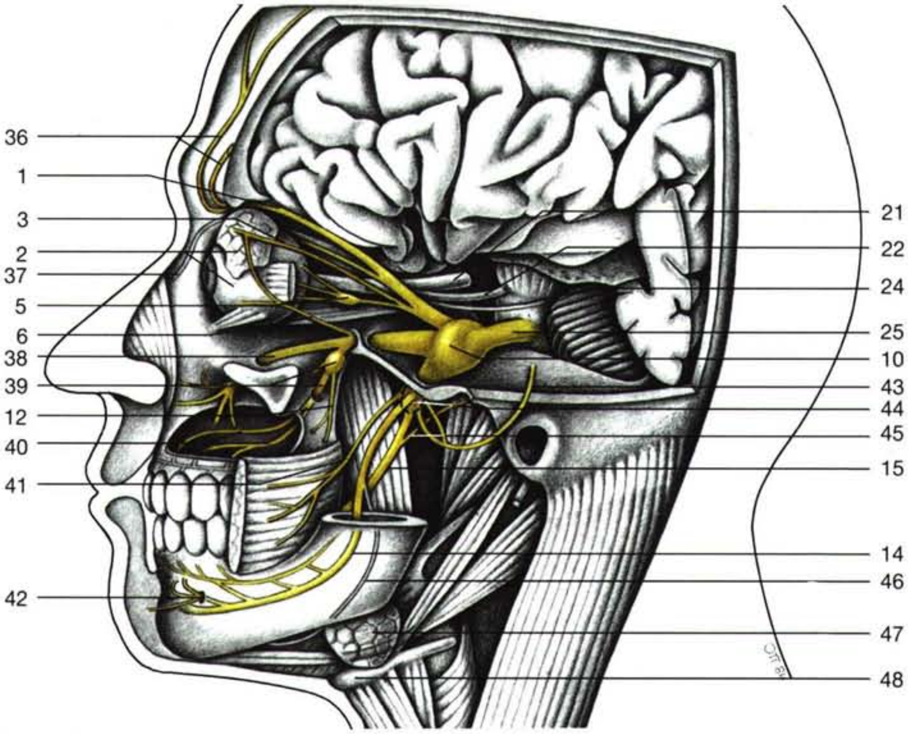

Main branches of trigeminal nerve. (Schematic drawing of figure on opposite page.)

1 Frontal nerve
2 Lacrimal gland and eyeball
3 Lacrimal nerve
4 Lateral rectus muscle
5 **Ciliary ganglion** lateral to optic nerve
6 Zygomatic nerve
7 Inferior branch of oculomotor nerve
8 **Ophthalmic nerve** (n. V_1)
9 **Maxillary nerve** (n. V_2)
10 **Trigeminal ganglion**
11 **Mandibular nerve** (n. V_3)
12 Posterior superior alveolar nerves
13 Tympanic cavity, external acoustic meatus, and tympanic membrane
14 Inferior alveolar nerve
15 Lingual nerve
16 **Facial nerve** (n. VII)
17 **Vagus nerve** (n. X)
18 **Hypoglossal nerve** (n. XII) and superior root of ansa cervicalis
19 External carotid artery
20 **Olfactory tract** (n. I)
21 **Optic nerve** (n. II) (intracranial part)
22 **Oculomotor nerve** (n. III)
23 **Abducens nerve** (n. VI)
24 **Trochlear nerve** (n. IV)
25 **Trigeminal nerve** (n. V)
26 **Vestibulocochlear nerve** (n. VIII) and **facial nerve** (VII)
27 **Glossopharyngeal nerve** (n. IX) (leaving brain stem)
28 Rhomboid fossa
29 **Vagus nerve** (n. X) (leaving brain stem)
30 **Hypoglossal nerve** (n. XII) (leaving medulla oblongata)
31 **Accessory nerve** (n. XI) (ascending from foramen magnum)
32 Vertebral artery
33 Spinal ganglion and dura mater of spinal cord
34 Accessory nerve (n. XI)
35 Internal carotid artery
36 Lateral and medial branch of supraorbital nerve
37 Infratrochlear nerve
38 Infraorbital nerve
39 **Pterygopalatine ganglion** and middle superior alveolar nerve
40 Middle superior alveolar nerves (entering superior dental plexus)
41 Buccal nerve
42 Mental nerve and mental foramen
43 Auriculotemporal nerve
44 Otic ganglion (dotted line)
45 Chorda tympani
46 Mylohyoid nerve
47 Submandibular gland
48 Hyoid bone

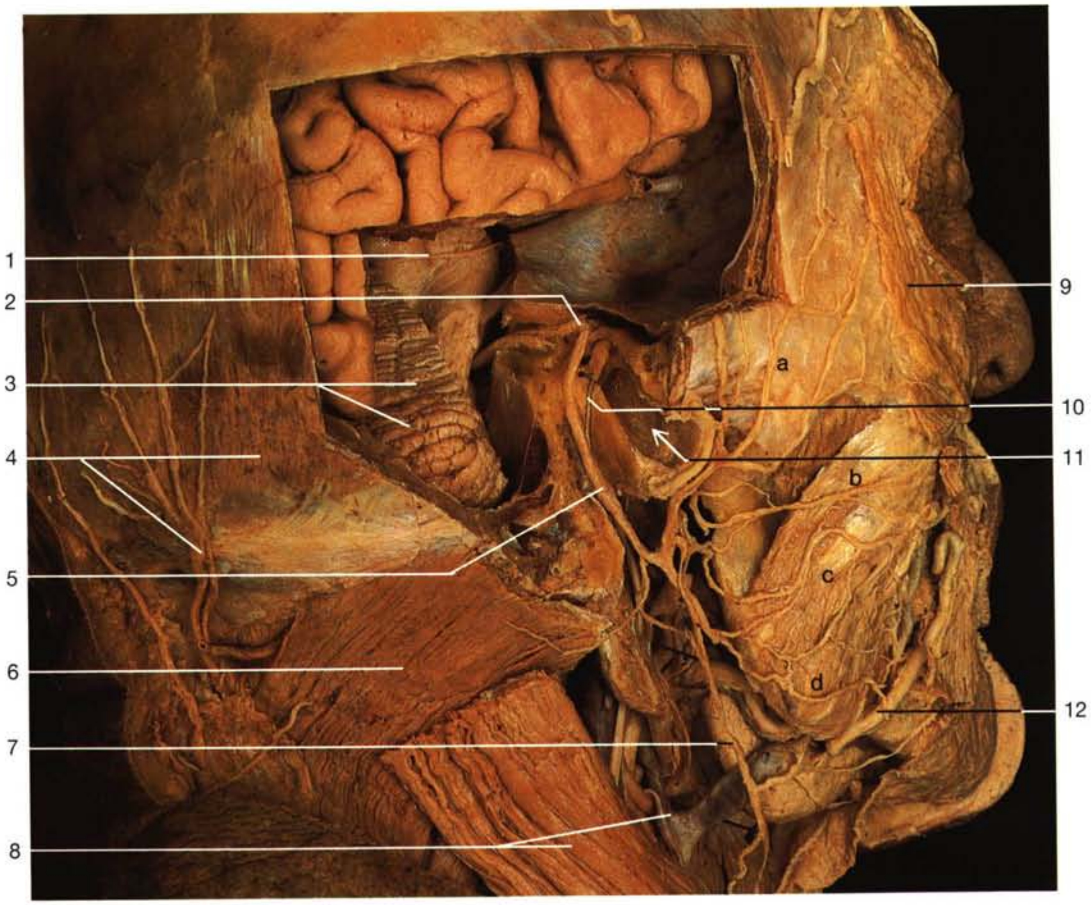

Dissection of facial nerve in its entirety. Cranial cavity fenestrated; temporal lobe partly removed. Facial canal and tympanic cavity opened, posterior wall of external acoustic meatus removed. **Branches of facial nerve:** a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

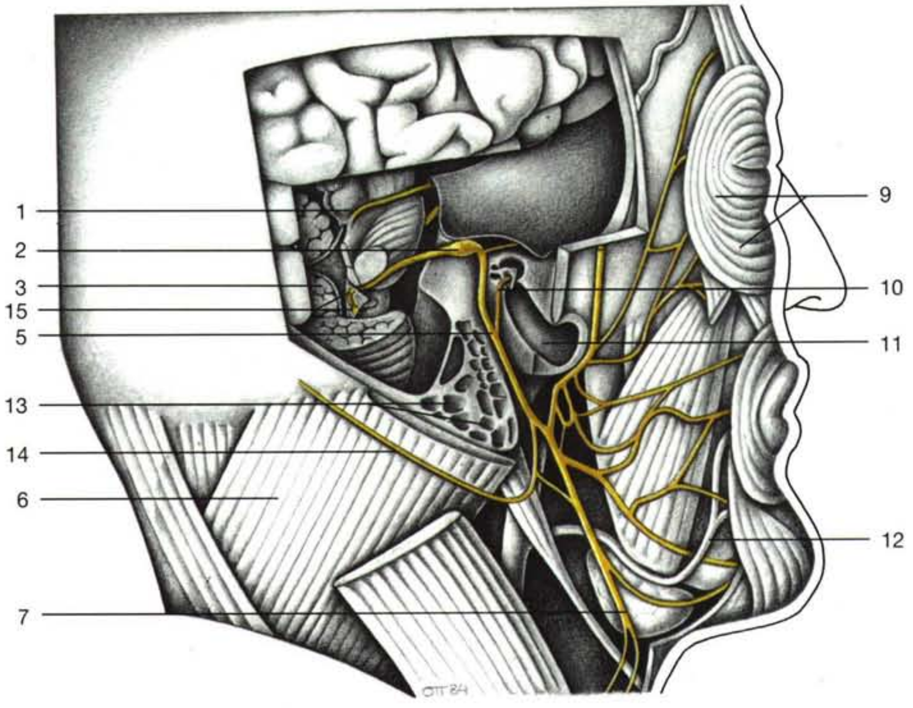

Facial nerve. (Schematic drawing of the dissection above.)

1 Trochlear nerve
2 Facial nerve with geniculate ganglion
3 Cerebellum (right hemisphere)
4 Occipital belly of occipitofrontalis and greater occipital nerve
5 Facial nerve at stylomastoid foramen
6 Splenius capitis muscle
7 Cervical branch of facial nerve
8 Sternocleidomastoid muscle and retromandibular vein
9 Orbicularis oculi muscle
10 Chorda tympani
11 External acoustic meatus
12 Facial artery
13 Mastoid air cells
14 Posterior auricular nerve
15 Nucleus and genu of facial nerve

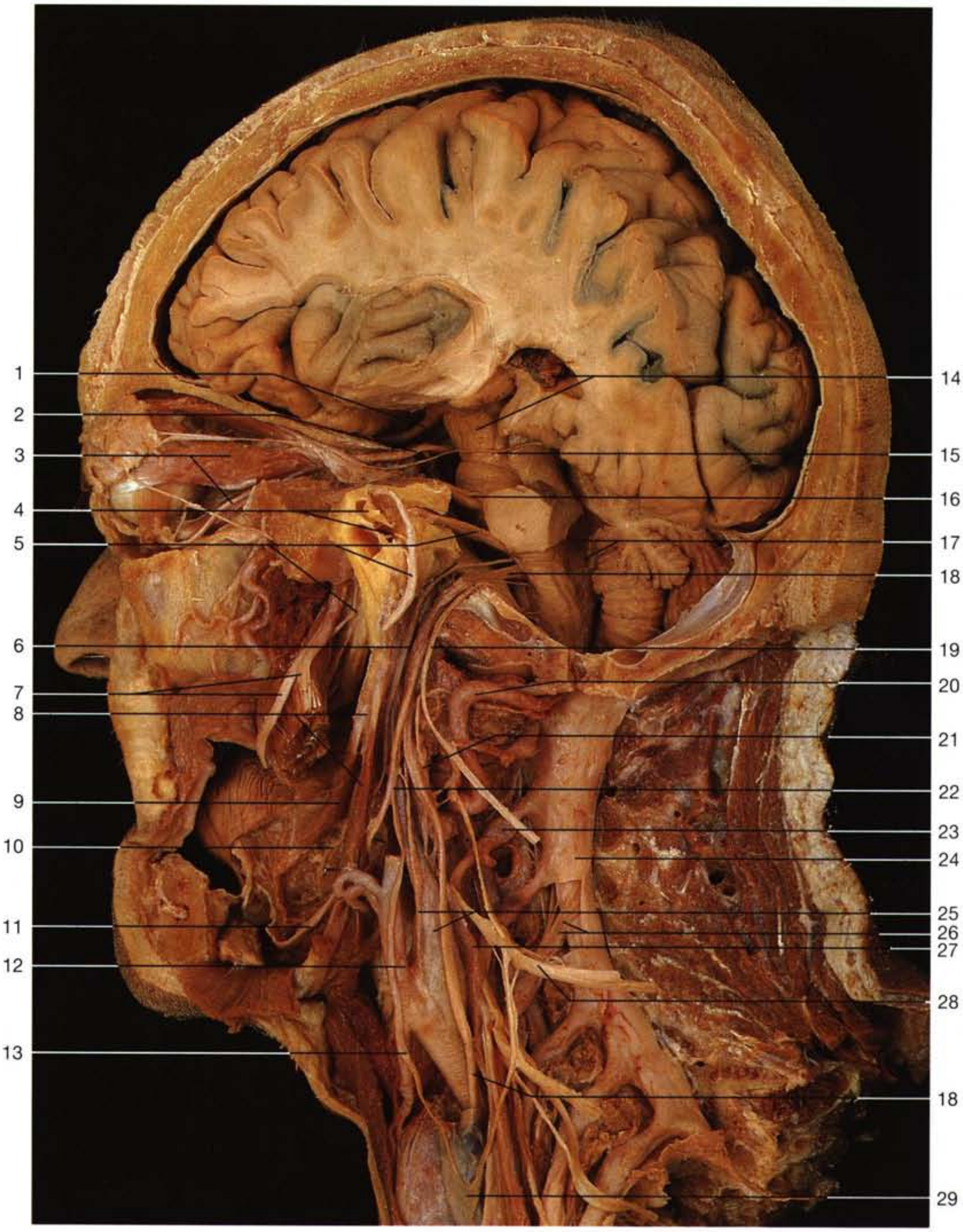

Cranial nerves in connection with the brain stem (oblique-lateral aspect). Lateral portion of the skull, brain, neck and facial structures, lateral wall of orbit and oral cavity have been removed. The tympanic cavity has been opened. The mandible has been divided and the muscles of mastication have been removed.

1 Optic tract
2 **Oculomotor nerve** (n. III)
3 Lateral rectus muscle and inferior branch of oculomotor nerve
4 Malleus and chorda tympani
5 Chorda tympani, **facial nerve** (n. VII), and **vestibulocochlear nerve** (n. VIII)
6 **Glossopharyngeal nerve** (n. XI)
7 Lingual nerve and inferior alveolar nerve
8 Styloid process and stylohyoid muscle
9 Styloglossus muscle
10 Lingual branches of glossopharyngeal nerve
11 Lingual branch of hypoglossal nerve
12 External carotid artery
13 Superior root of ansa cervicalis (branch of hypoglossal nerve)
14 Lateral ventricle with choroid plexus and cerebral peduncle
15 **Trochlear nerve** (n. IV)
16 **Trigeminal nerve** (n. V)
17 Fourth ventricle and rhomboid fossa
18 **Vagus nerve** (n. X)
19 **Accessory nerve** (n. XI)
20 Vertebral artery
21 Superior cervical ganglion
22 **Hypoglossal nerve** (n. XII)
23 Spinal ganglion with dural sheath
24 Dura mater of spinal cord
25 Internal carotid artery and carotid sinus branch of glossopharyngeal nerve
26 Dorsal roots of spinal nerve
27 Sympathetic trunk
28 Branch of cervical plexus (ventral primary ramus of third cervical spinal nerve)
29 Ansa cervicalis

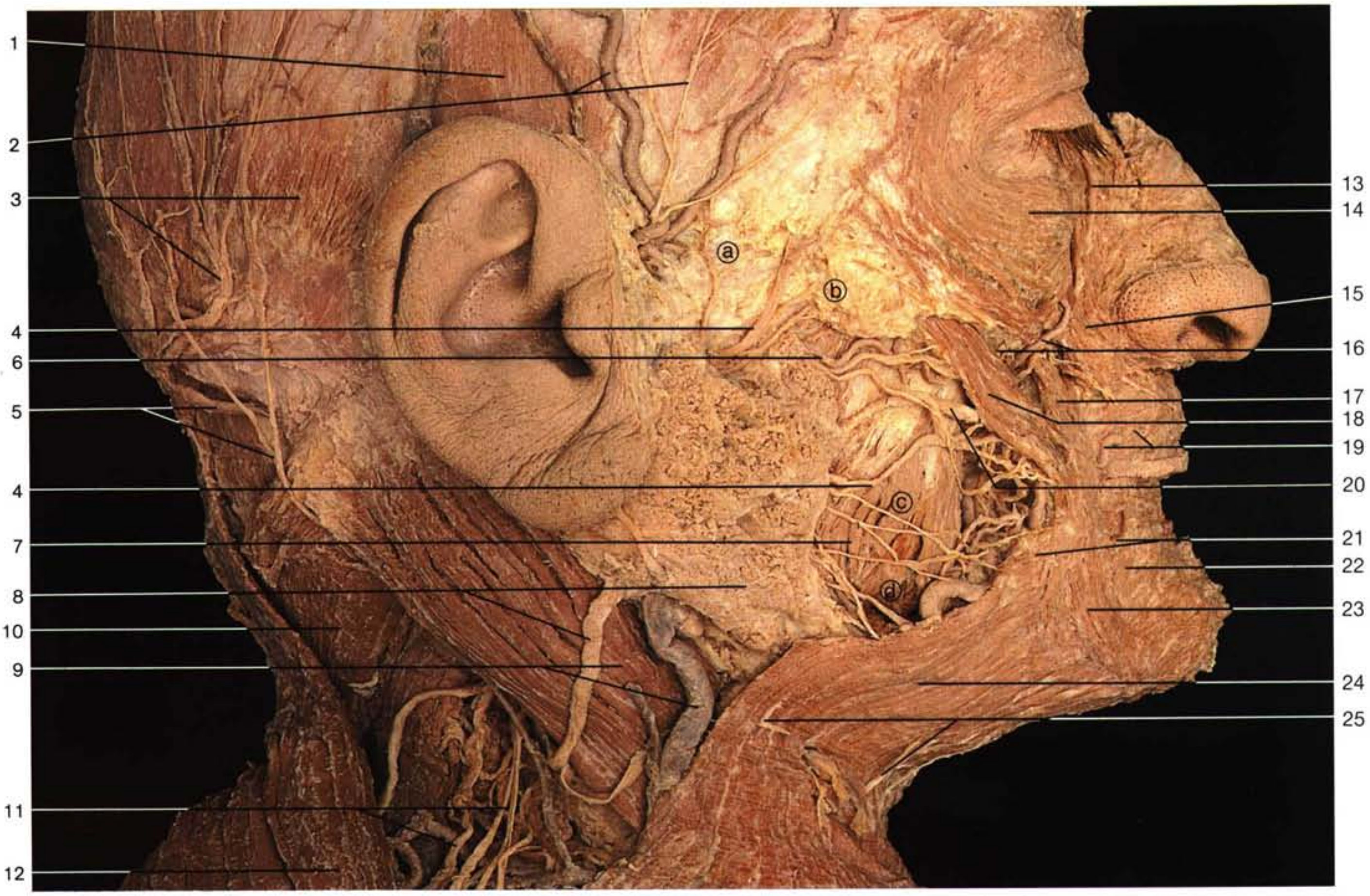

Lateral superficial aspect of the face. Peripheral distribution of facial nerve. a–d = branches of facial nerve: a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

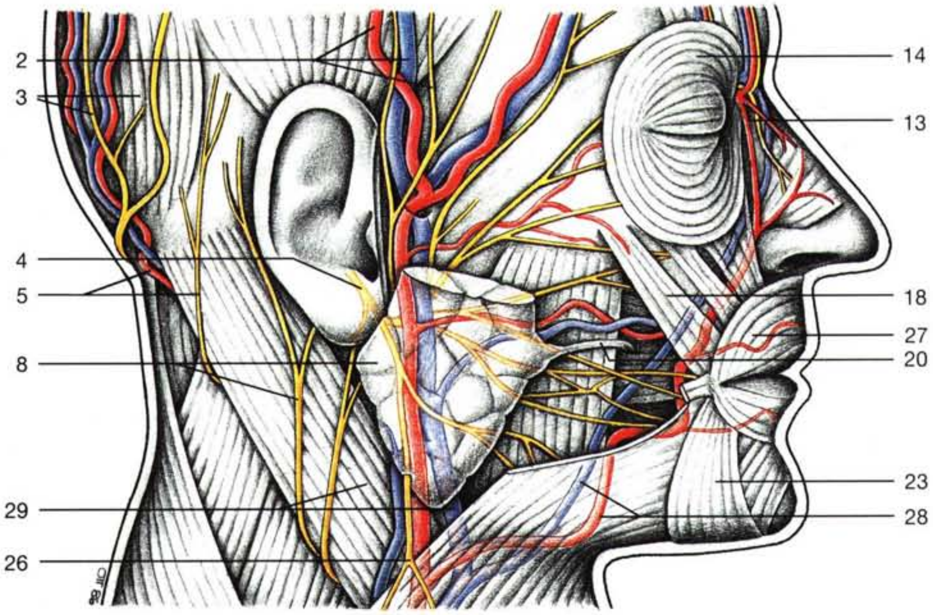

Superficial region of the face. Note the facial plexus within the parotid gland. (Semischematic drawing.)

1 Temporoparietalis muscle
2 Parietal branch of **superficial temporal artery** and **vein** and **auriculotemporal nerve**
3 Occipital belly of occipitofrontalis muscle and greater occipital nerve
4 **Facial nerve**
5 Lesser occipital nerve and **occipital artery**
6 Transverse facial artery
7 Masseter muscle
8 **Parotid gland** and great auricular nerve
9 Sternocleidomastoid muscle and external jugular vein
10 Splenius capitis muscle
11 Branches of cervical plexus
12 Trapezius muscle
13 Angular artery (terminal portion of facial artery)
14 Orbicularis oculi muscle
15 Levator labii superioris alaeque nasi muscle
16 **Facial artery** and zygomaticus minor muscle
17 Levator anguli oris muscle
18 Zygomaticus major muscle
19 Orbicularis oris muscle and superior labial artery
20 Parotid duct
21 Risorius muscle and inferior labial artery

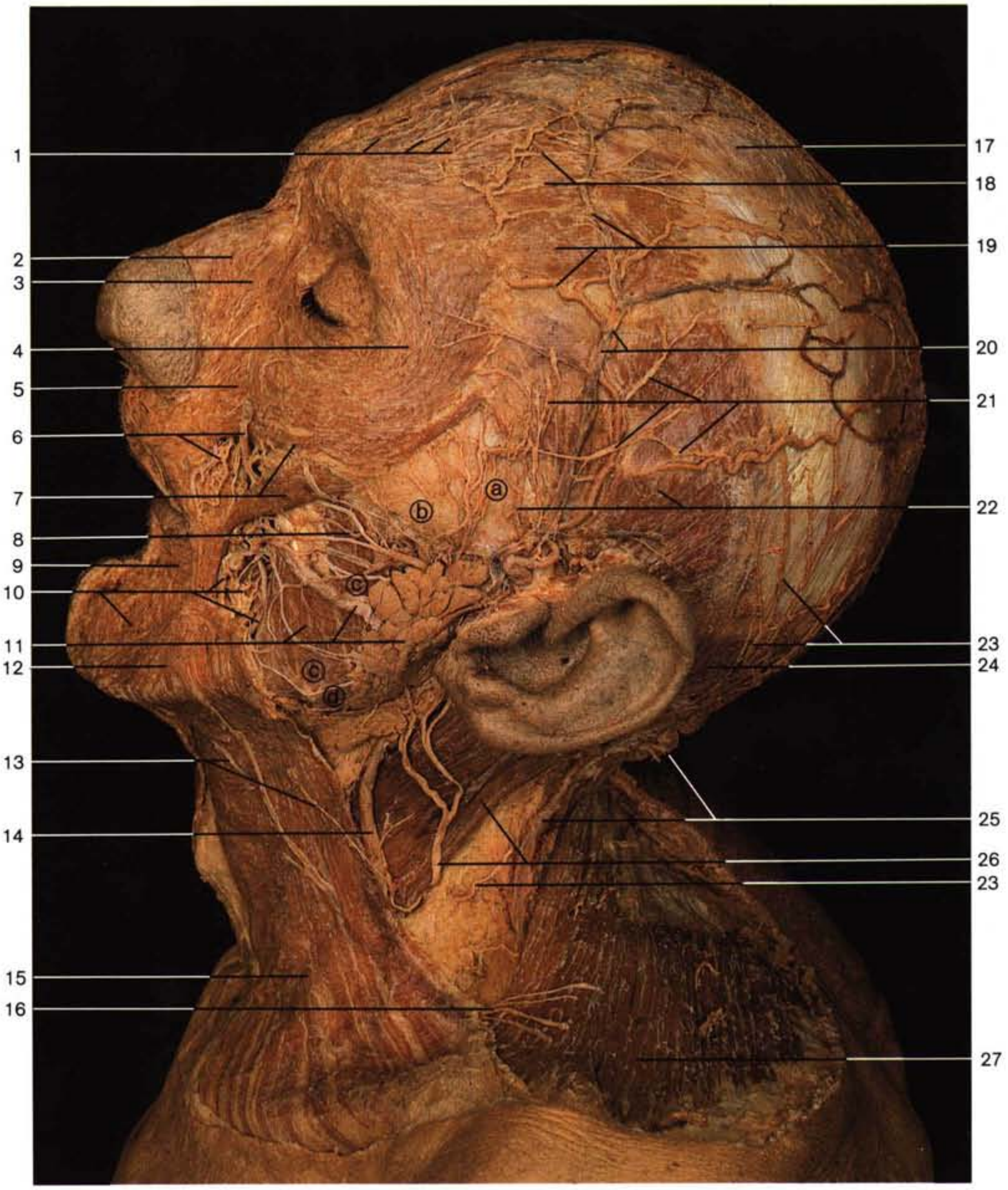

Superficial layer of the head and neck. Parotid fascia removed. Branches of **facial nerve:** a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch.

1 Medial branch of supraorbital nerve
2 Nasalis muscle
3 Levator labii superioris alaeque nasi muscle
4 **Orbicularis oculi muscle**
5 Levator labii superioris muscle
6 **Facial artery and vein**
7 Zygomaticus minor and major muscle
8 Transverse facial artery
9 **Orbicularis oris muscle**
10 Buccal nerves, depressor labii inferioris muscle, **facial artery** and **vein**
11 **Parotid gland** and duct and masseter muscle
12 Depressor anguli oris muscle
13 Transverse cervical nerve
14 External jugular vein
15 **Platysma muscle**
16 Supraclavicular nerve
17 Galea aponeurotica
18 Lateral branches of the supraorbital nerve
19 Frontal belly of occipitofrontalis muscle and branches of superficial temporal vein and artery
20 Superficial temporal artery and vein
21 **Auriculotemporal nerve**
22 Zygomatico-orbital artery and temporoparietalis muscle
23 Lesser occipital nerve
24 Occipital belly of occipitofrontalis muscle
25 Occipital vein and occipital artery
26 **Great auricular nerve** and sternocleidomastoid muscle
27 Trapezius muscle
28 Greater petrosal nerve
29 Geniculate ganglion
30 Chorda tympani
31 Posterior auricular nerve
32 Stylomastoid foramen

◁ **to page 76**
22 Depressor labii inferioris muscle
23 Depressor anguli oris muscle
24 Platysma muscle
25 Terminal branches of transverse cervical nerve
26 Cervical branch of facial nerve
27 Orbicularis oris muscle
28 **Facial artery** and **vein**
29 Sternocleidomastoid muscle and retromandibular vein

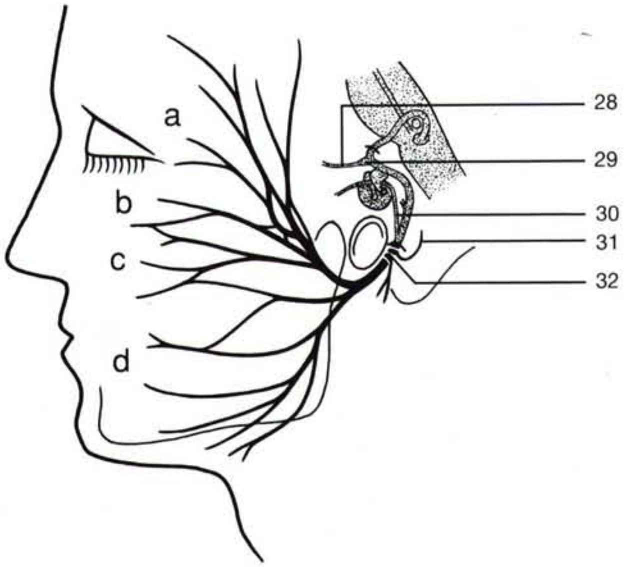

Facial nerve. Main branches. (Schematic diagram.)

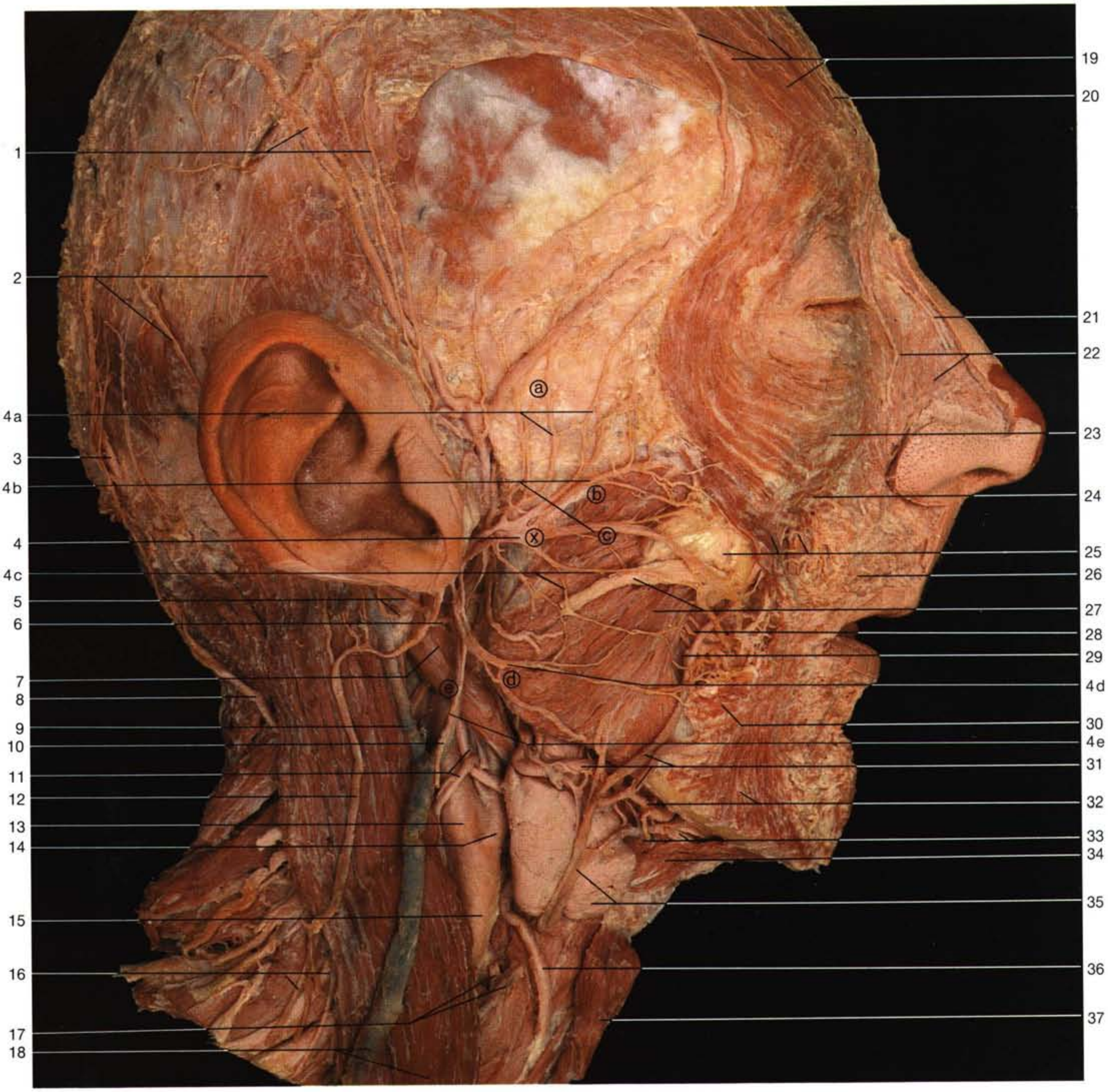

Lateral superficial aspect of the face. The parotid gland has been removed to display the parotid plexus of the facial nerve. a–e = branches of facial nerve: a = temporal branch; b = zygomatic branches; c = buccal branches; d = marginal mandibular branch; e = cervical branch.

1 Superficial temporal artery and auriculotemporal nerve
2 Posterior auricular artery and nerve and temporoparietalis muscle
3 Occipital artery
4 **Facial nerve** (n. VII) (**parotid plexus** [×])
  a Temporal branches
  b Zygomatic branches
  c Buccal branches
  d Marginal mandibular branch
  e Cervical branch
5 Posterior auricular nerve
6 Posterior auricular artery
7 Digastric muscle (posterior belly)
8 **Lesser occipital nerve**
9 Posterior auricular vein
10 Retromandibular vein
11 **Hypoglossal nerve** and sternocleidomastoid artery
12 **Great auricular nerve**
13 Internal carotid artery
14 External carotid artery
15 Common carotid artery
16 Branches of cervical plexus
17 Superior laryngeal artery and vein
18 External jugular vein and sternocleidomastoid muscle
19 Frontal branch of superficial temporal artery, lateral branch of supraorbital nerve, and frontal belly of occipitofrontalis muscle
20 Medial branch of supraorbital nerve
21 Dorsal nasal artery
22 Angular artery and nasalis muscle
23 **Orbicularis oculi muscle**
24 Zygomaticus minor muscle
25 Buccal fat pad, zygomaticus major muscle, and infraorbital nerve
26 **Orbicularis oris muscle**
27 **Parotid duct** and **masseter muscle**
28 Buccal artery and nerve
29 Buccinator muscle
30 Risorius muscle
31 **Facial artery** and **vein**
32 Submental artery and depressor anguli oris muscle
33 Mylohyoid nerve and mylohyoid muscle
34 Digastric muscle (anterior belly)
35 Facial vein and submandibular gland
36 **Superior thyroid artery**
37 Sternohyoid muscle

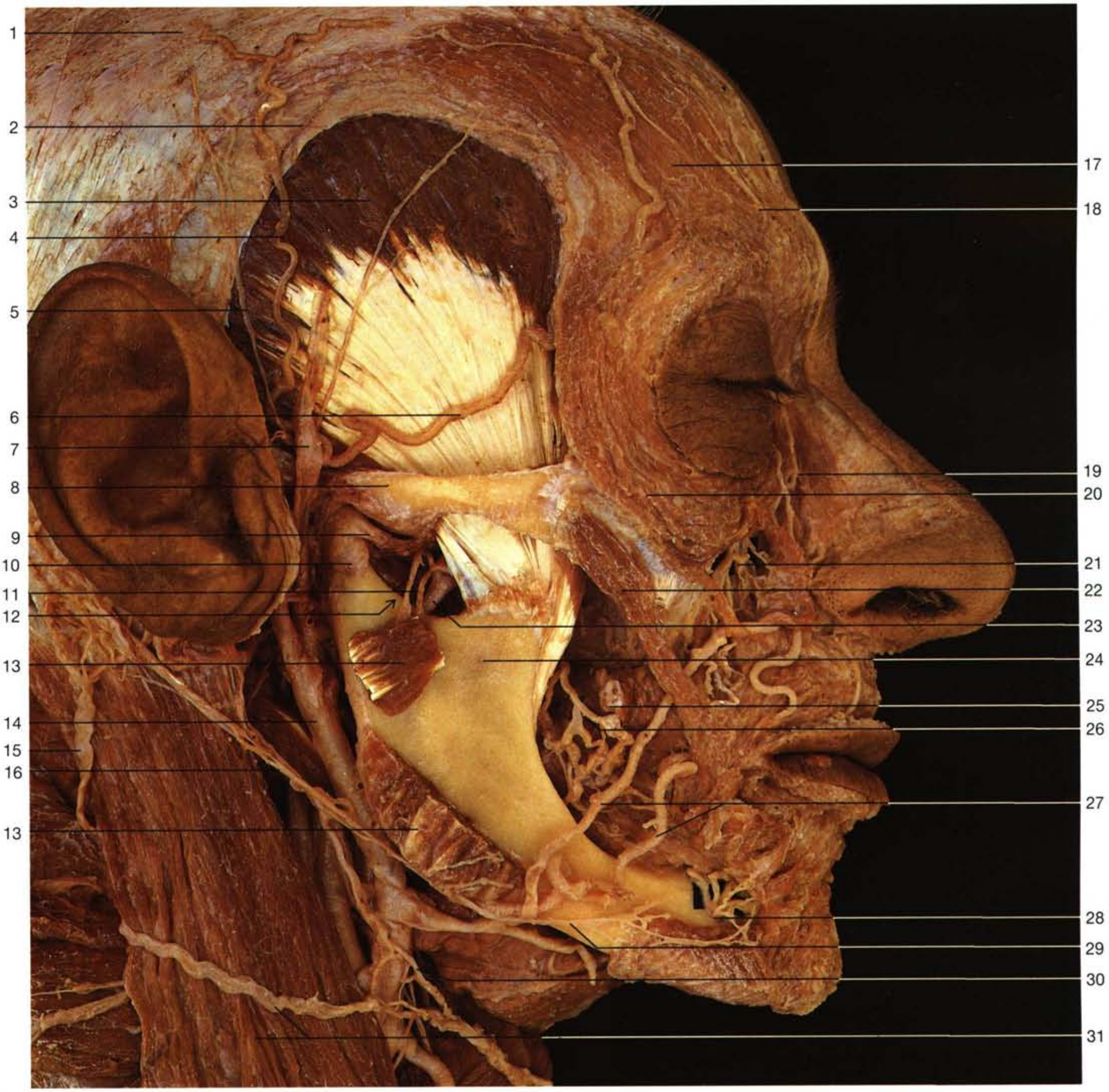

Lateral superficial aspect of the face. Masseter muscle and temporal fascia have been partly removed to display the masseteric artery and nerve.

1 Galea aponeurotica
2 Temporal fascia
3 **Temporalis muscle**
4 Parietal branch of **superficial temporal artery**
5 **Auriculotemporal nerve**
6 Frontal branch of superficial temporal artery
7 Superficial temporal vein
8 **Zygomatic arch**
9 **Articular disc of temporomandibular joint**
10 **Head of mandible**
11 Masseteric artery and nerve
12 Mandibular notch
13 Masseter muscle (divided)
14 **External carotid artery**
15 Great auricular nerve
16 Facial nerve (reflected)
17 Frontal belly of occipitofrontalis muscle
18 Medial branch of supraorbital nerve
19 **Angular artery**
20 Orbicularis oculi muscle
21 **Infraorbital nerve**
22 Zygomaticus major muscle
23 **Maxillar artery**
24 **Coronoid process**
25 Parotid duct (divided)
26 Buccal nerve
27 **Facial artery** and **vein**
28 **Mental nerve**
29 Mandibular branch of facial nerve
30 Cervical branch of facial nerve
31 Transverse cervical nerve (communicating branch with facial nerve) and sternocleidomastoid muscle

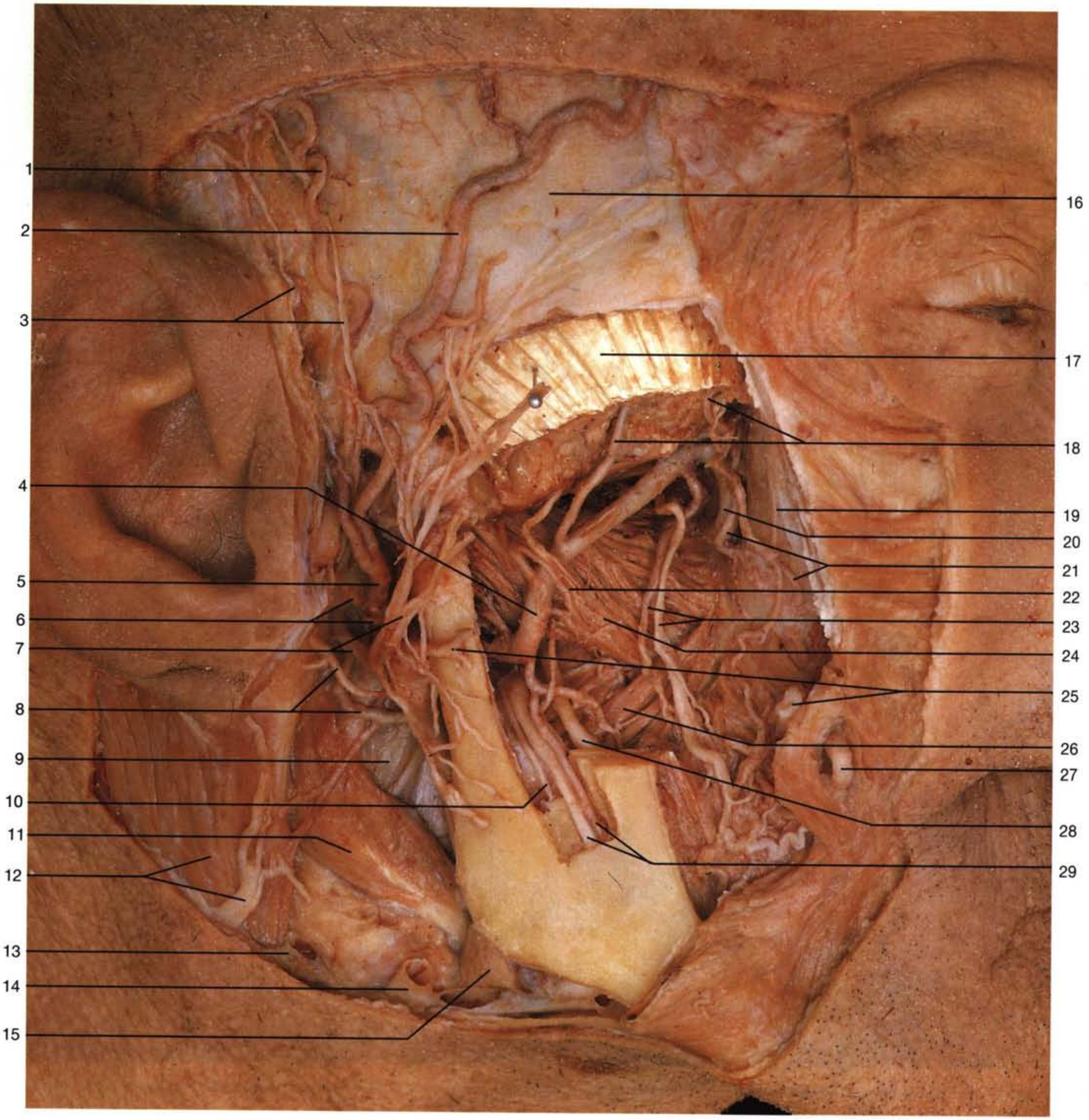

Deep dissection of facial and retromandibular regions. The coronoid process together with the insertions of temporalis muscle have been removed to display the maxillary artery. The upper part of the mandibular canal has been opened.

1 Parietal branch of the superficial temporal artery
2 Frontal branch of the superficial temporal artery
3 Auriculotemporal nerve
4 **Maxillary artery**
5 **Superficial temporal artery**
6 Communicating branches between facial and auriculotemporal nerves
7 **Facial nerve**
8 Posterior auricular artery and anterior auricular branch of superficial temporal artery
9 Internal jugular vein
10 Mylohyoid nerve
11 Posterior belly of digastric muscle
12 Great auricular nerve and sternocleidomastoid
13 External jugular vein
14 Retromandibular vein
15 Submandibular gland
16 Temporal fascia
17 **Temporalis tendon**
18 Deep temporal arteries
19 Posterior superior alveolar nerve
20 Sphenopalatine artery
21 Posterior superior alveolar arteries
22 Masseteric artery and nerve
23 **Buccal nerve and artery**
24 **Lateral pterygoid**
25 Transverse facial artery and parotid duct (divided)
26 Medial pterygoid
27 **Facial artery**
28 **Lingual nerve**
29 **Inferior alveolar artery and nerve** (mandibular canal opened)

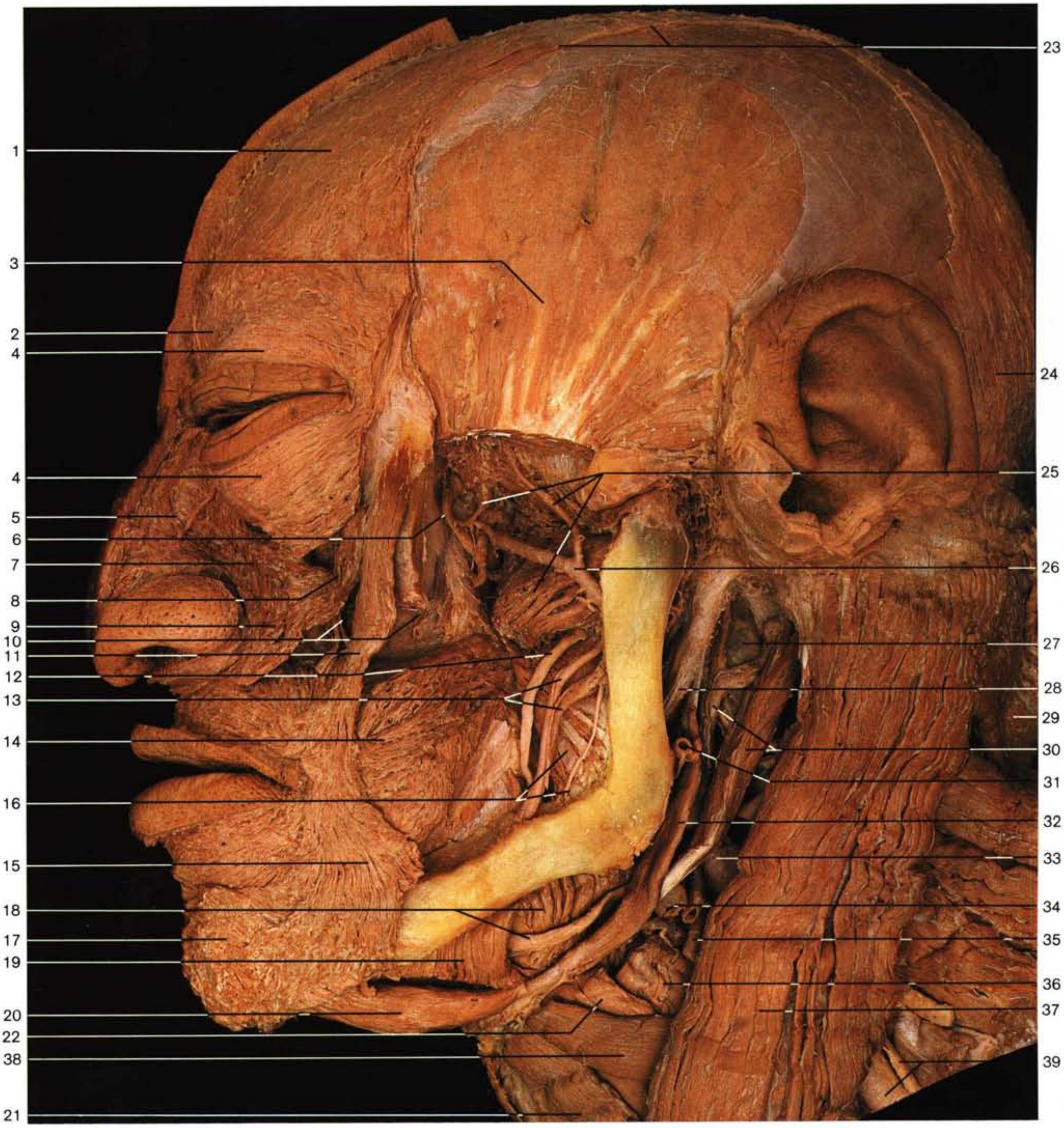

Retromandibular region with maxillary artery and branches of trigeminal nerve (n. V).

1 Frontal belly of occipitofrontalis muscle
2 Depressor supercilii muscle
3 Temporalis muscle
4 **Orbicularis oculi muscle**
5 Nasalis muscle
6 Infraorbital artery
7 Levator labii superioris alaeque nasi muscle
8 Zygomaticus minor muscle
9 Levator labii superioris muscle
10 Infraorbital artery and nerve and posterior superior alveolar artery
11 **Zygomaticus major muscle**
12 **Lingual nerve**
13 **Inferior alveolar artery and nerve**
14 Buccinator muscle
15 Depressor anguli oris muscle
16 **Medial pterygoid muscle** and **mylohyoid nerve**
17 Depressor labii inferioris muscle
18 **Hypoglossal nerve** and **hyoglossus muscle**
19 Mylohyoid muscle
20 Anterior belly of diagastric muscle
21 Sternohyoid muscle
22 Thyrohyoid muscle
23 Galea aponeurotica
24 Occipital belly of occipitofrontalis muscle
25 **Lateral pterygoid muscle** and deep temporal artery
26 **Maxillary artery**
27 Internal jugular vein
28 Styloglossus muscle
29 Splenius capitis muscle
30 Posterior belly of diagastric muscle and occipital artery
31 Superficial temporal artery
32 **Stylohyoid muscle**
33 External carotid artery
34 Retromandibular vein
35 Superior thyroid artery
36 Inferior constrictor muscle of the pharynx
37 Sternocleidomastoid muscle
38 Omohyoid muscle
39 Common carotid artery and vagus nerve

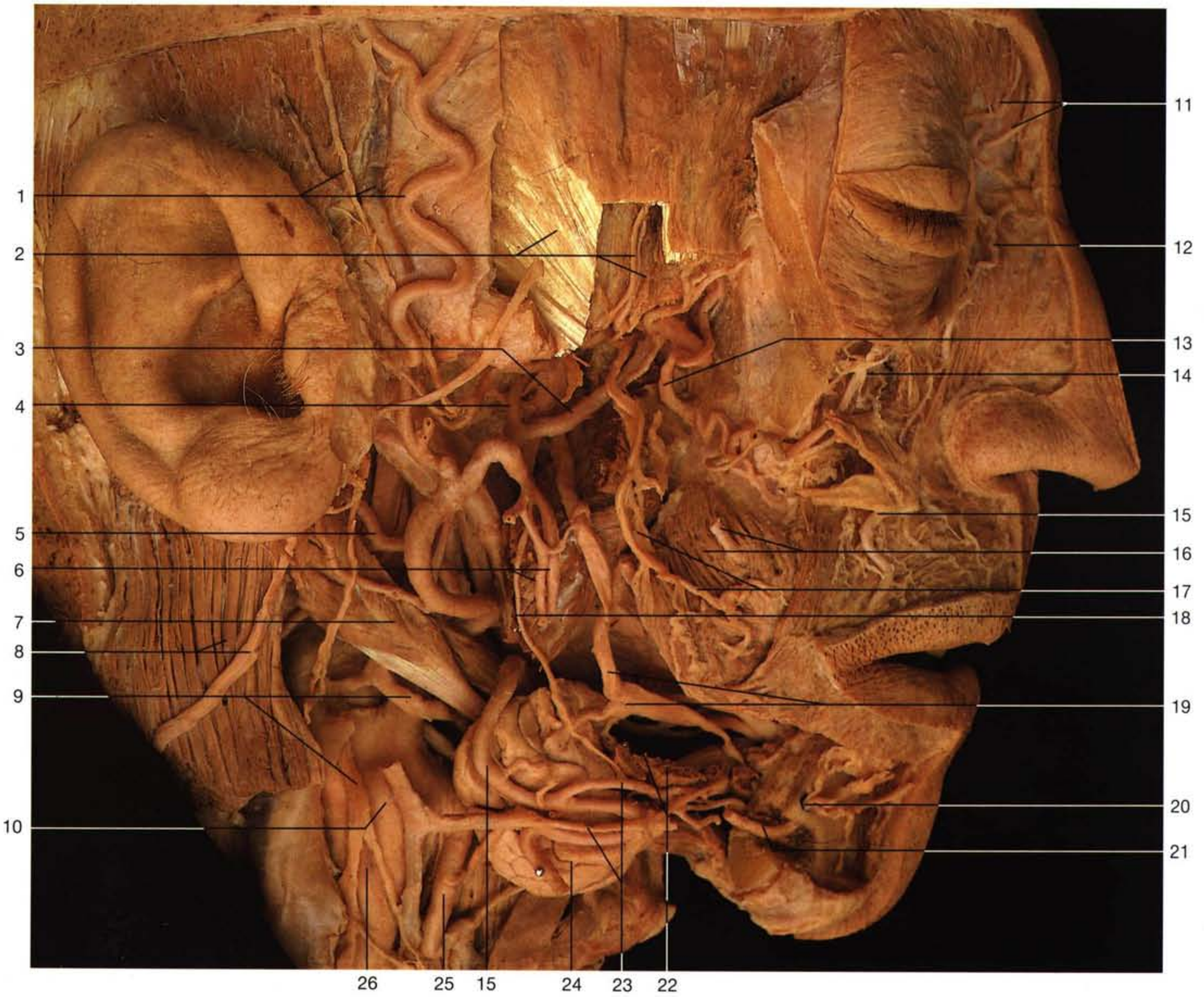

Dissection of deep facial and retromandibular regions after removal of mandible. Pteryoid muscles removed, temporalis muscle fenestrated.

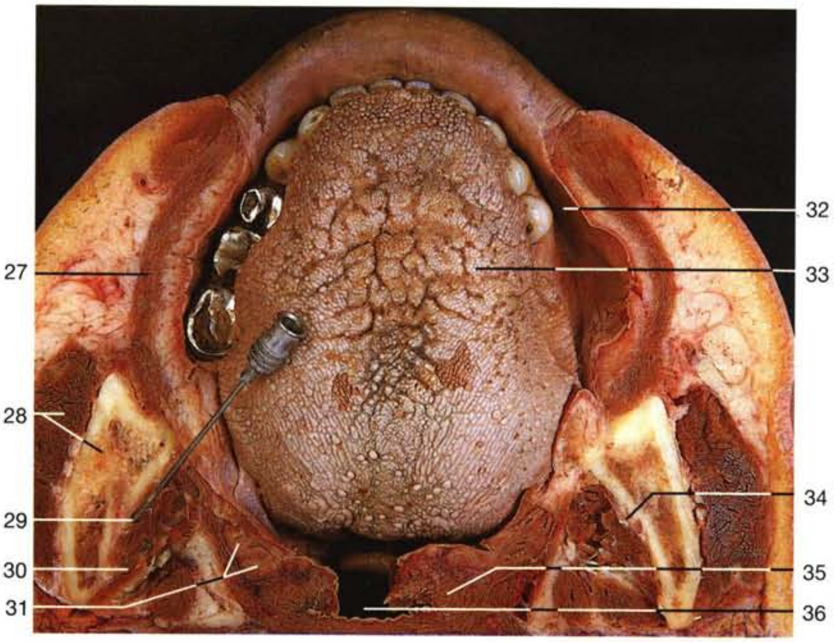

Transverse section through oral cavity and pharynx. The location of inferior alveolar nerve and artery is indicated by a needle.

1 Superficial temporal artery and vein and auriculotemporal nerve
2 Temporalis tendon, deep temporal nerves and artery
3 **Maxillary artery**
4 Middle meningeal artery
5 Occipital artery
6 Inferior alveolar artery and nerve
7 Posterior belly of digastric muscle
8 Great auricular nerve and sternocleidomastoid muscle
9 Hypoglossal nerve and superior root of ansa cervicalis
10 **External carotid artery**
11 Supratrochlear nerve and medial branch of supraorbital artery
12 Angular artery
13 Posterior superior alveolar artery
14 Infraorbital nerve
15 **Facial artery**
16 Parotid duct (divided) and buccinator muscle
17 Buccal artery and nerve

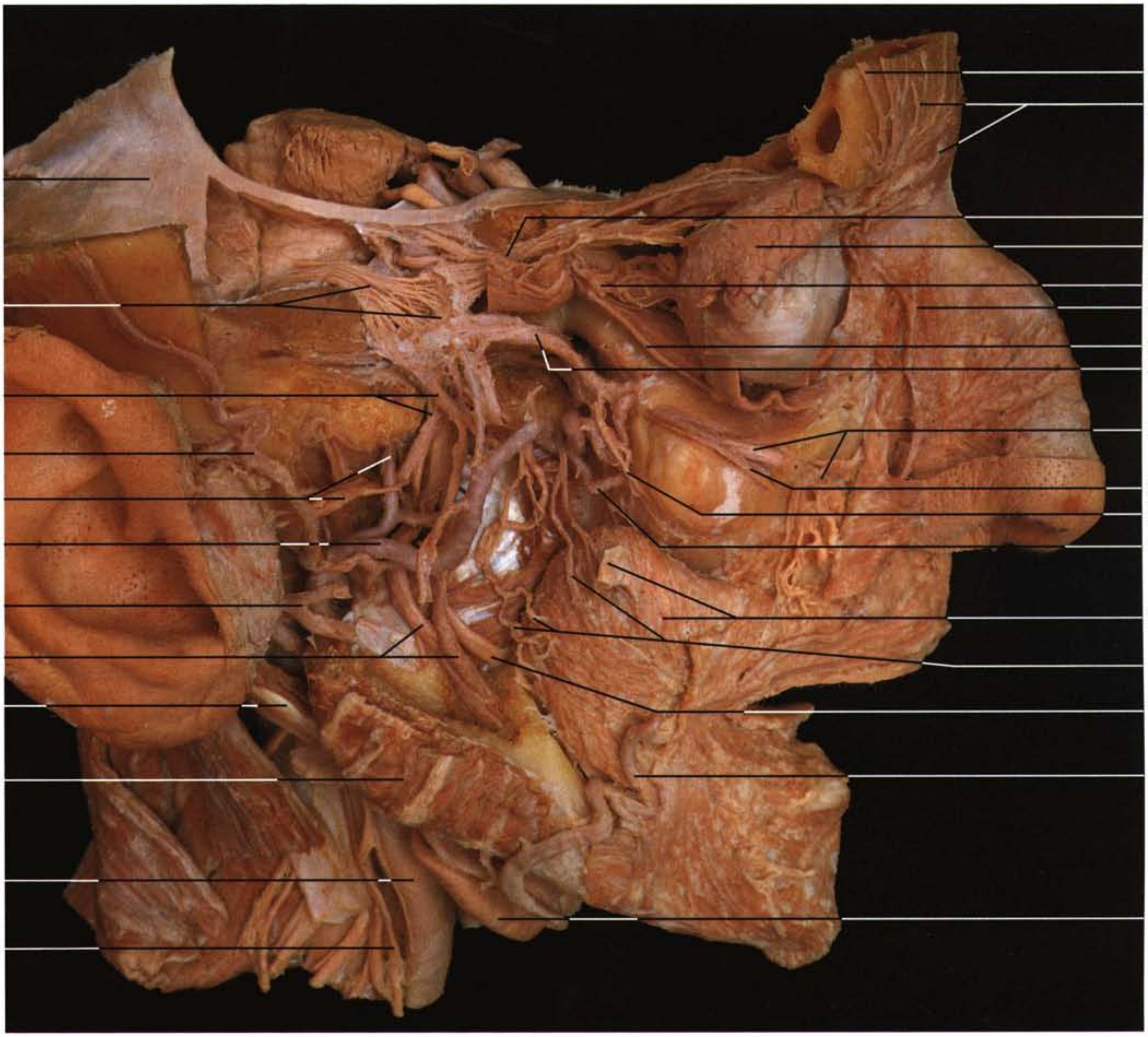

Para- and retropharyngeal regions. The mandible and the lateral wall of the orbit have been removed. The main branches of the trigeminal nerve and its ganglion are displayed.

18	Mylohyoid nerve
19	**Lingual nerve** and **submandibular ganglion**
20	Mental nerve and mental foramen
21	Inferior alveolar nerve
22	Mylohyoid muscle (divided) and hypoglossal nerve
23	Submental artery and vein
24	Submandibular gland
25	Superior thyroid artery
26	Common carotid artery
27	Buccinator muscle
28	Masseter muscle and mandible
29	Entrance of mandibular canal
30	Medial pterygoid muscle
31	Palatine tonsil
32	Oral vestibule
33	Tongue
34	Inferior alveolar nerve, artery and vein
35	Paryngeal constrictor muscle
36	Pharynx
37	Tentorium of cerebellum
38	**Trigeminal nerve** and **ganglion**
39	**Mandibular nerve**
40	Superficial temporal artery
41	Auriculotemporal nerve and middle meningeal artery
42	Facial nerve (divided)
43	Masseter muscle
44	Superior root of ansa cervicalis
45	Lateral branch of supraorbital nerve
46	**Ophthalmic nerve**
47	Lacrimal gland
48	Ciliary ganglion and short ciliary nerves
49	Angular artery
50	Inferior branch of oculomotor nerve
51	**Maxillary nerve**
52	Infraorbital nerve
53	Anterior superior alveolar nerve
54	Posterior superior alveolar nerve

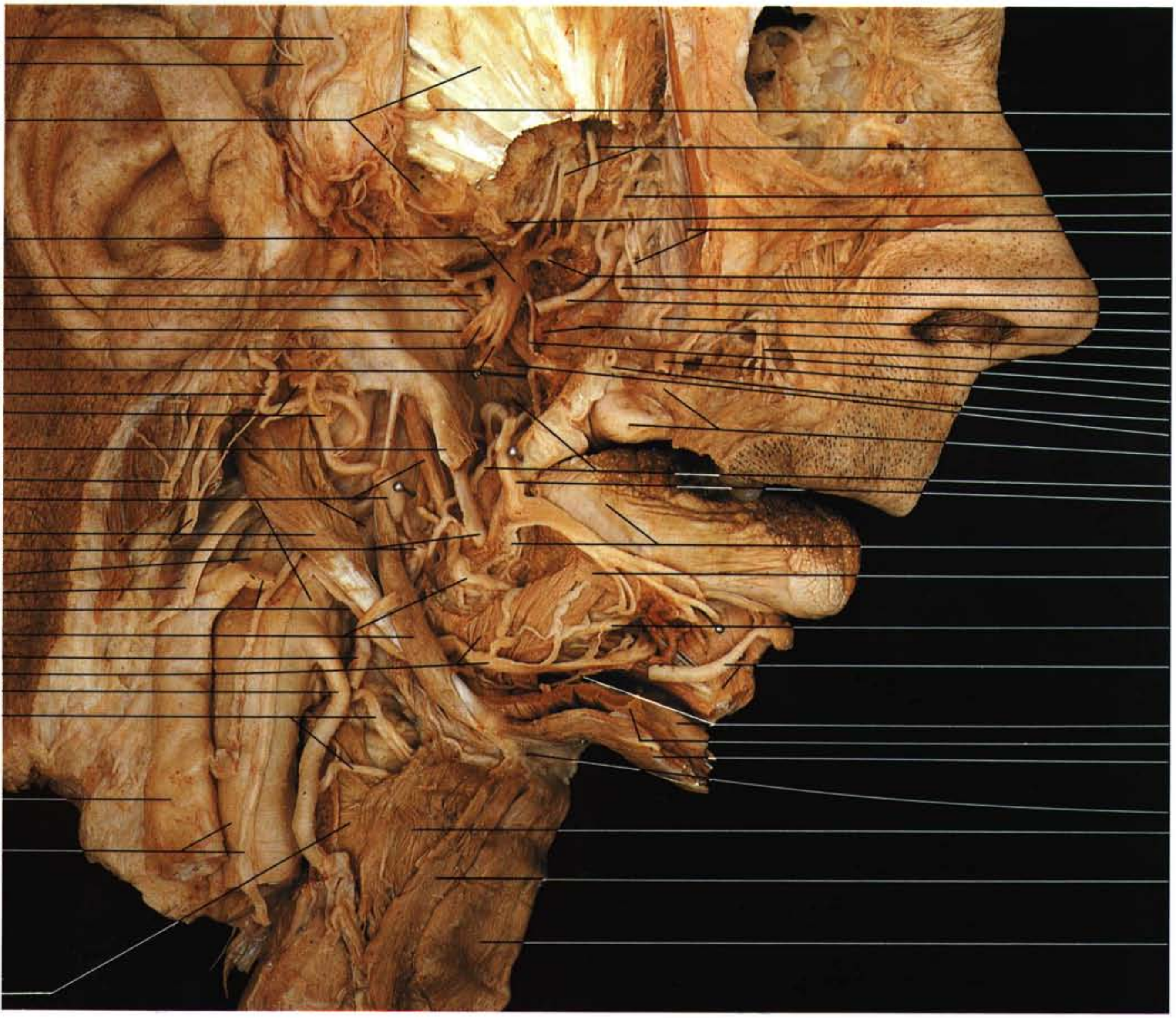

Retropharyngeal and sublingual regions. The mandible has been completely removed.

1 Parietal branch of superficial temporal artery
2 Auriculotemporal nerve
3 Temporalis muscle tendon and zygomatic arch (cut)
4 Mandibular nerve (n. V_3)
5 Middle meningeal artery
6 Chorda tympani
7 Mylohyoid nerve
8 Inferior alveolar nerve
9 **Lingual nerve**
10 Posterior auricular artery
11 Facial nerve
12 Styloglossus muscle (cut)
13 Great auricular nerve
14 Maxillary artery, external carotid artery, and stylopharyngeus muscle
15 Posterior belly of digastric muscle
16 Styloid process and facial artery
17 Vagus nerve (n. X)
18 **Accessory nerve** (n. XI)
19 Hypoglossal nerve (n. XII) and retromandibular vein (cut)
20 Stylohyoid muscle and glossopharyngeal nerve (n. IX)
21 Facial vein
22 Hypoglossal nerve (n. XII) and hyoglossus muscle
23 Superior thyroid artery
24 Superior laryngeal artery and internal laryngeal nerve
25 Internal jugular vein
26 Common carotid artery and superior root of ansa cervicalis
27 Inferior pharyngeal constrictor muscle
28 Middle temporal artery
29 Deep temporal nerves and posterior deep temporal artery
30 Anterior deep temporal artery
31 Masseteric nerve
32 Posterior superior alveolar branches of maxillary artery and nerve
33 Pterygoid branches of mandibular nerve
34 Posterior superior alveolar artery
35 Lateral pterygoid plate and medial pterygoid muscle
36 Infraorbital nerve and artery
37 Buccal nerve
38 Facial artery
39 Parotid duct

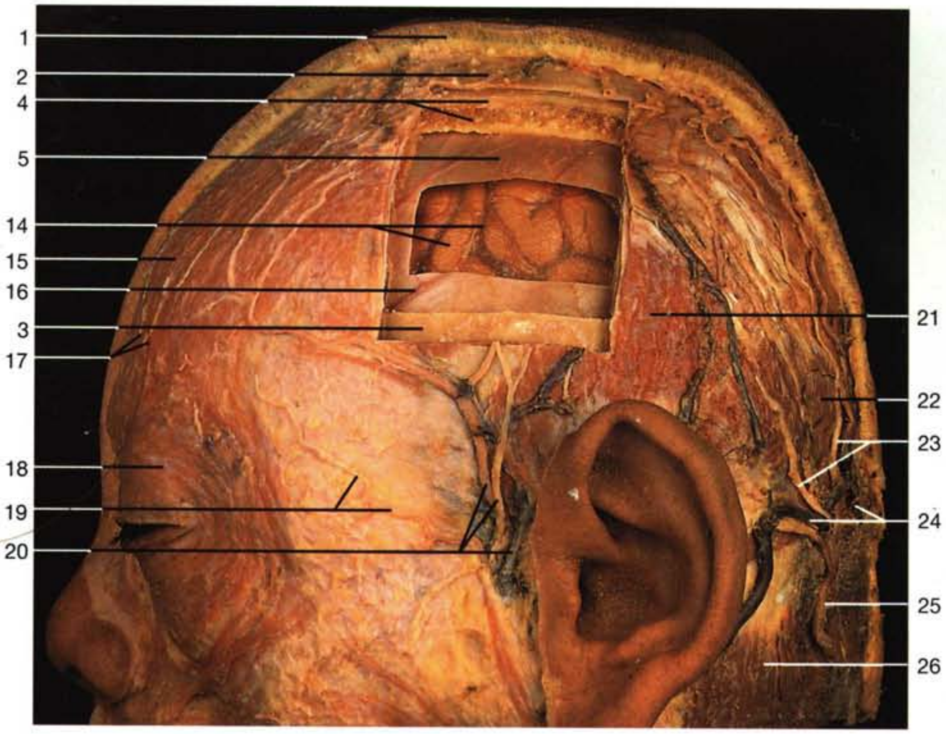

Nerves and blood vessels of the scalp. Scalp and meninges are demonstrated by a series of window-like openings.

1 Skin
2 **Galea aponeurotica**
3 Pericranium (periosteum)
4 Skull with diploe
5 **Dura mater**
6 Subdural space
7 **Arachnoid mater**
8 **Subarachnoid space**
9 Arachnoid granulations
10 Superior sagittal sinus
11 **Pia mater** with cerebral vessels
12 Falx cerebri
13 Cerebral cortex
14 Arachnoid and pia mater with cortical vessels
15 Frontal belly of occipitofrontal muscle
16 Branch of middle meningeal artery
17 Lateral and medial branch of supraorbital nerve
18 Orbicularis oculi muscle
19 Zygomatico-orbital artery
20 **Auriculotemporal nerve, superficial temporal artery** and **vein**
21 Superior auricular muscle
22 Occipital belly of occipitofrontalis muscle
23 Branches of **greater occipital nerve**
24 **Occipital artery** and **vein**
25 Greater occipital nerve
26 Sternocleidomastoid muscle

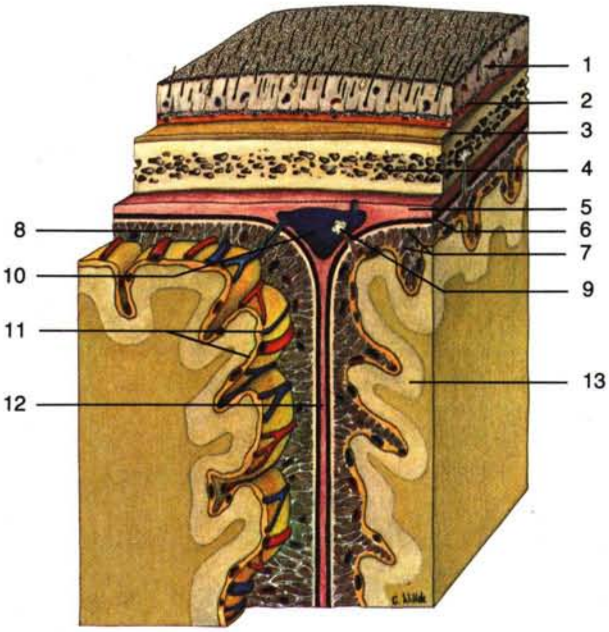

A coronal **section through the vertex of the skull** showing the arrangement of the meninges and vessels of the brain. Together the arachnoid mater and the pia mater form the leptomeninx.

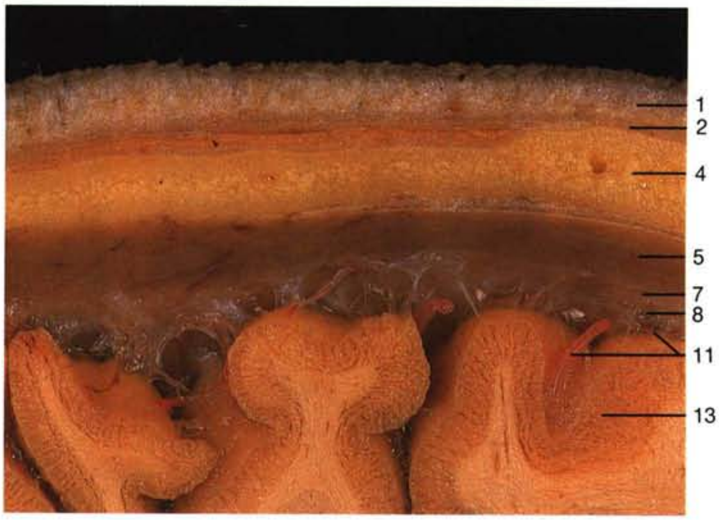

Cross-section of the scalp and the meninges. The subarachnoid space is shown.

40 Levator veli palatini muscle
41 Gingiva and buccinator muscle
42 Ascending pharyngeal artery and superior pharyngeal constrictor muscle (pterygopharyngeal part)
43 Lingual nerve
44 **Submandibular ganglion** and tongue
45 Palatoglossus muscle
46 Deep lingual artery
47 Submandibular duct and genioglossus muscle
48 Geniohyoid nerve and muscle
49 Mylohyoid muscle
50 Anterior belly of digastric muscle
51 Hyoid bone
52 Thyrohyoid muscle
53 Omohyoid muscle (superior belly)
54 Sternohyoid muscle
55 **Maxillary artery**

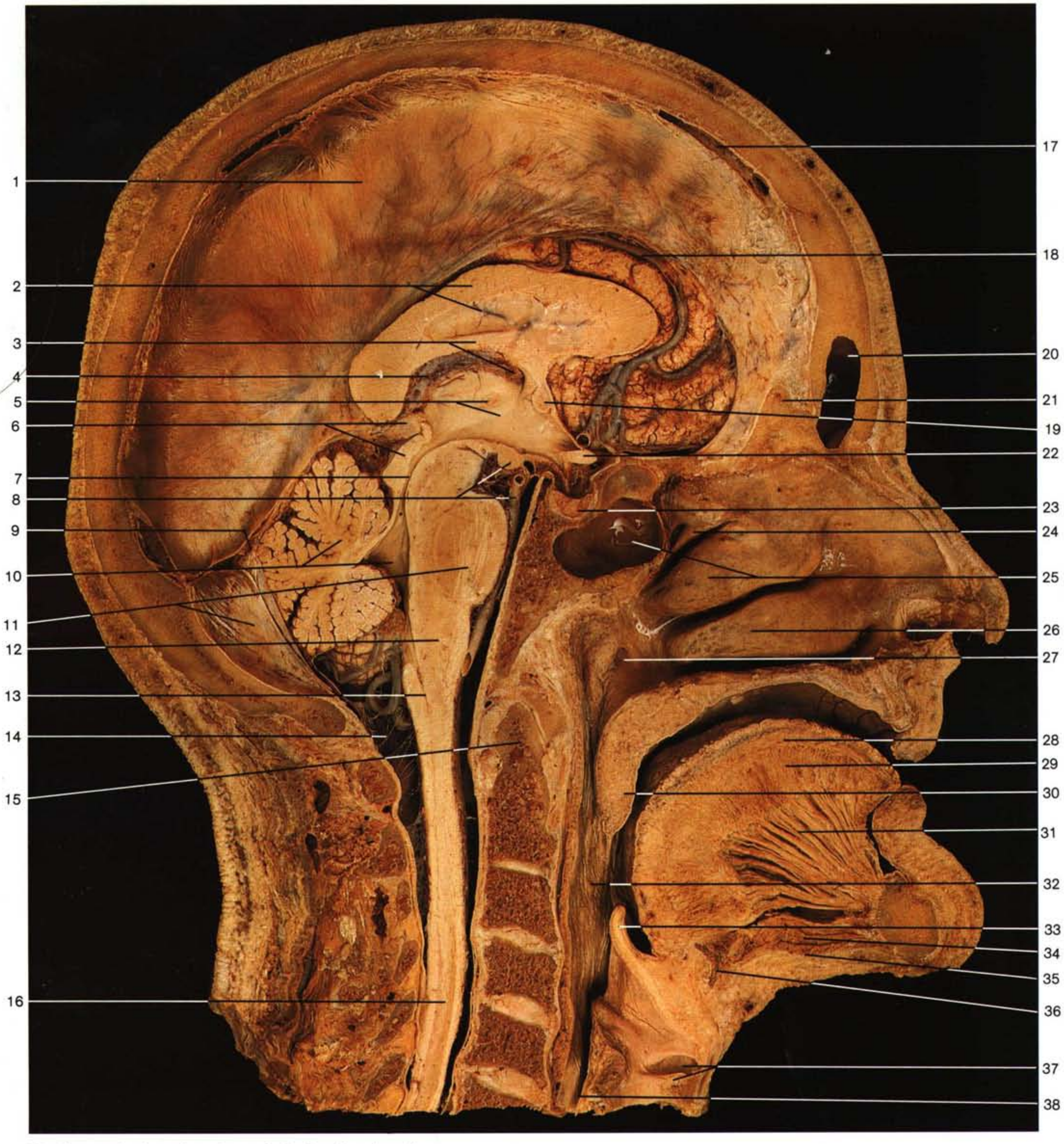

Median sagittal section through the head and neck.

1 Falx cerebri
2 **Corpus callosum** and **septum pellucidum**
3 Interventricular foramen and fornix
4 Choroid plexus of third ventricle and internal cerebral vein
5 **Third ventricle** and interthalamic adhesion
6 **Pineal body** and **colliculi of the midbrain**
7 Cerebral aqueduct
8 Mamillary body and basilar artery
9 Straight sinus
10 **Fourth ventricle** and **cerebellum**
11 Pons and falx cerebelli
12 **Medulla oblongata**
13 Central canal
14 **Cerebellomedullary cistern**
15 Dens of the axis (odontoid process)
16 Spinal cord
17 **Superior sagittal sinus**
18 **Anterior cerebral artery**
19 Anterior commissure
20 **Frontal sinus**
21 Crista galli
22 Optic chiasma
23 Pituitary gland (hypophysis)
24 Superior nasal concha
25 **Middle nasal concha** and **sphenoid sinus**
26 **Inferior nasal concha**
27 **Pharyngeal opening of auditory tube**
28 Superior longitudinal muscle of tongue
29 Vertical muscle of the tongue
30 **Uvula**
31 Genioglossus muscle
32 Pharynx
33 **Epiglottis**
34 Geniohyoid muscle
35 Mylohyoid muscle
36 Hyoid bone
37 Vocal fold and sinus of larynx
38 Esophagus

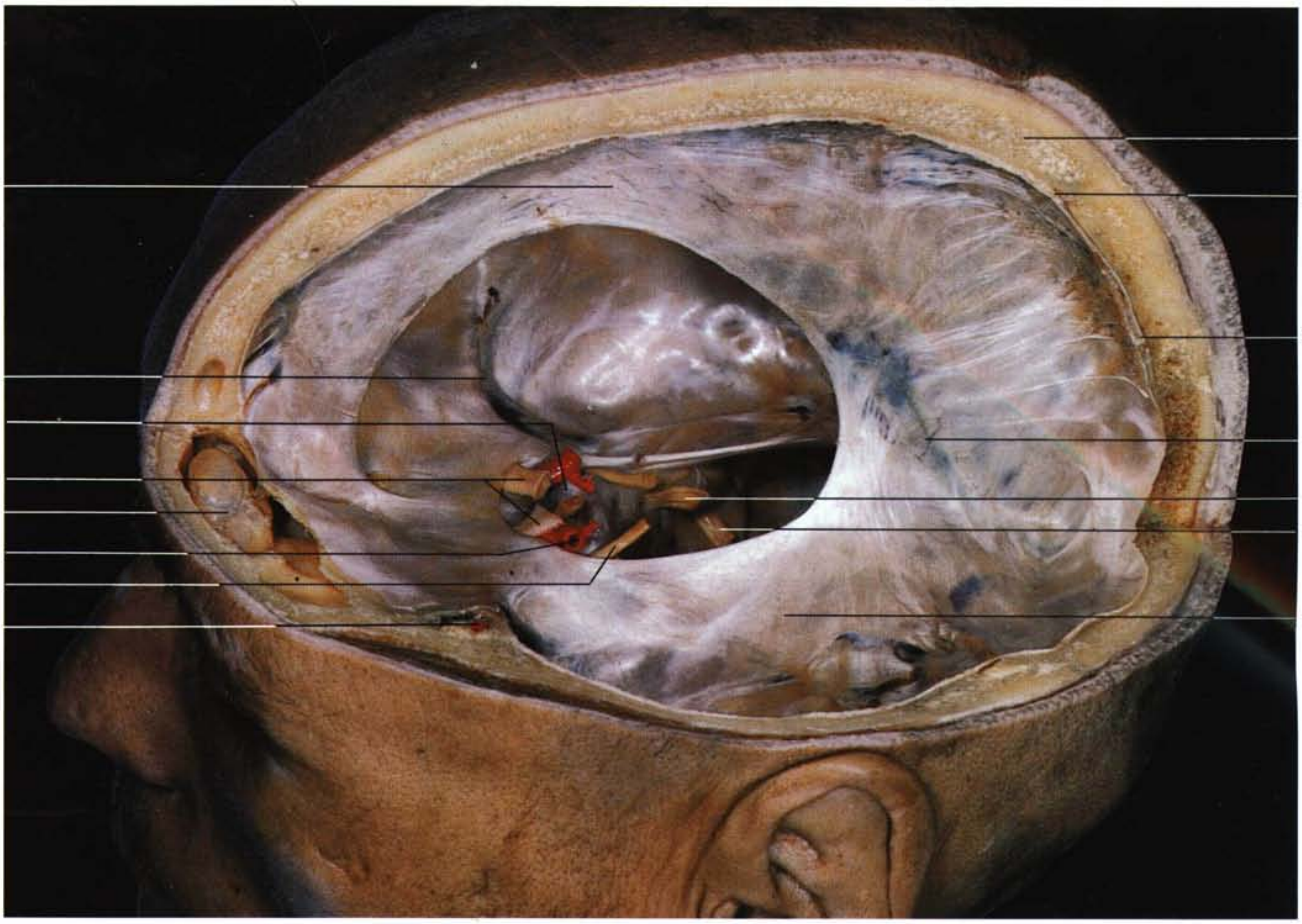

Dura mater and venous sinuses of the dura mater. The brain has been removed (oblique lateral aspect).

1 Falx cerebri
2 Position of middle meningeal artery and vein
3 Internal carotid artery
4 **Optic nerves** (n. II)
5 Frontal sinus
6 **Oculomotor nerve** (n. III)
7 Diploe
8 **Dura mater**
9 Superior sagittal sinus
10 Straight sinus
11 **Trigeminal nerve** (n. V)
12 Facial and vestibulocochlear nerve (n. VII and n. VIII)
13 **Tentorium cerebelli**
14 **Pituitary gland (hypophysis)**
15 Inferior sagittal sinus
16 Sigmoid sinus
17 Confluence of sinuses
18 Inferior petrosal sinus
19 Transverse sinus
20 Superior petrosal sinus
21 **Cavernous** and **intercavernous sinuses**

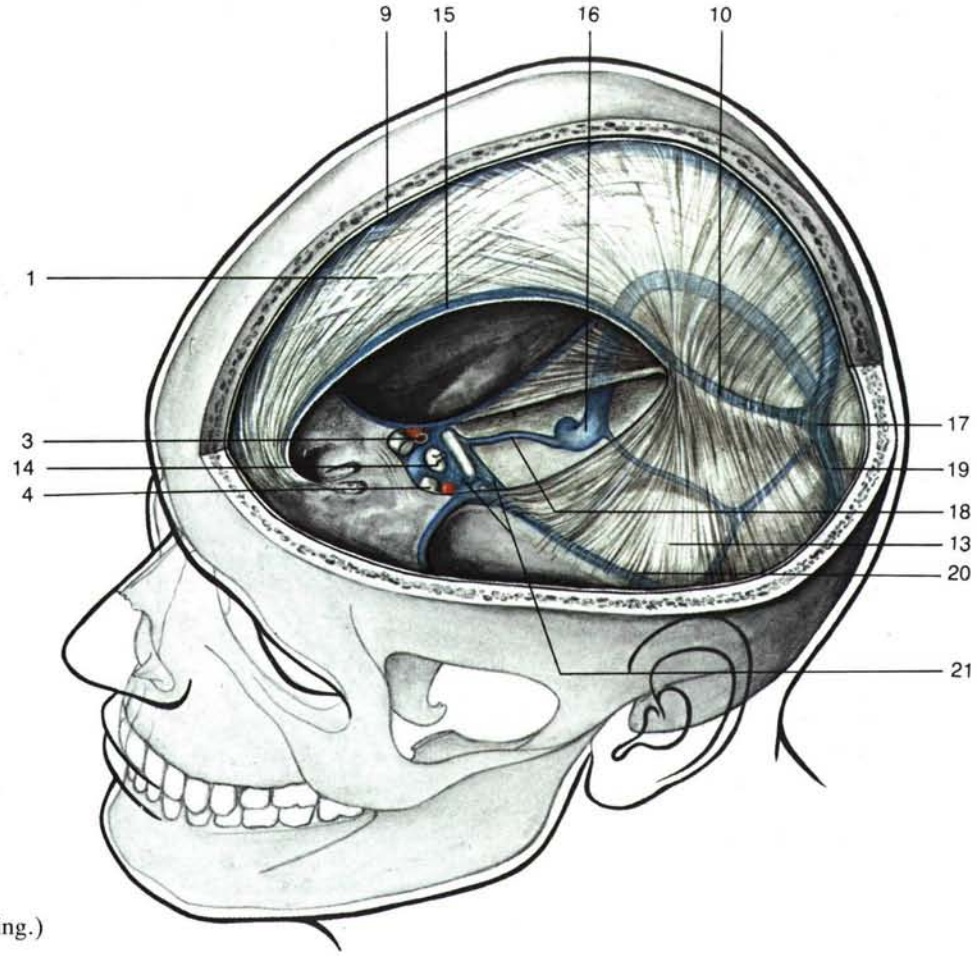

Dura mater and **venous sinuses** (left lateral aspect). (Schematic drawing.)

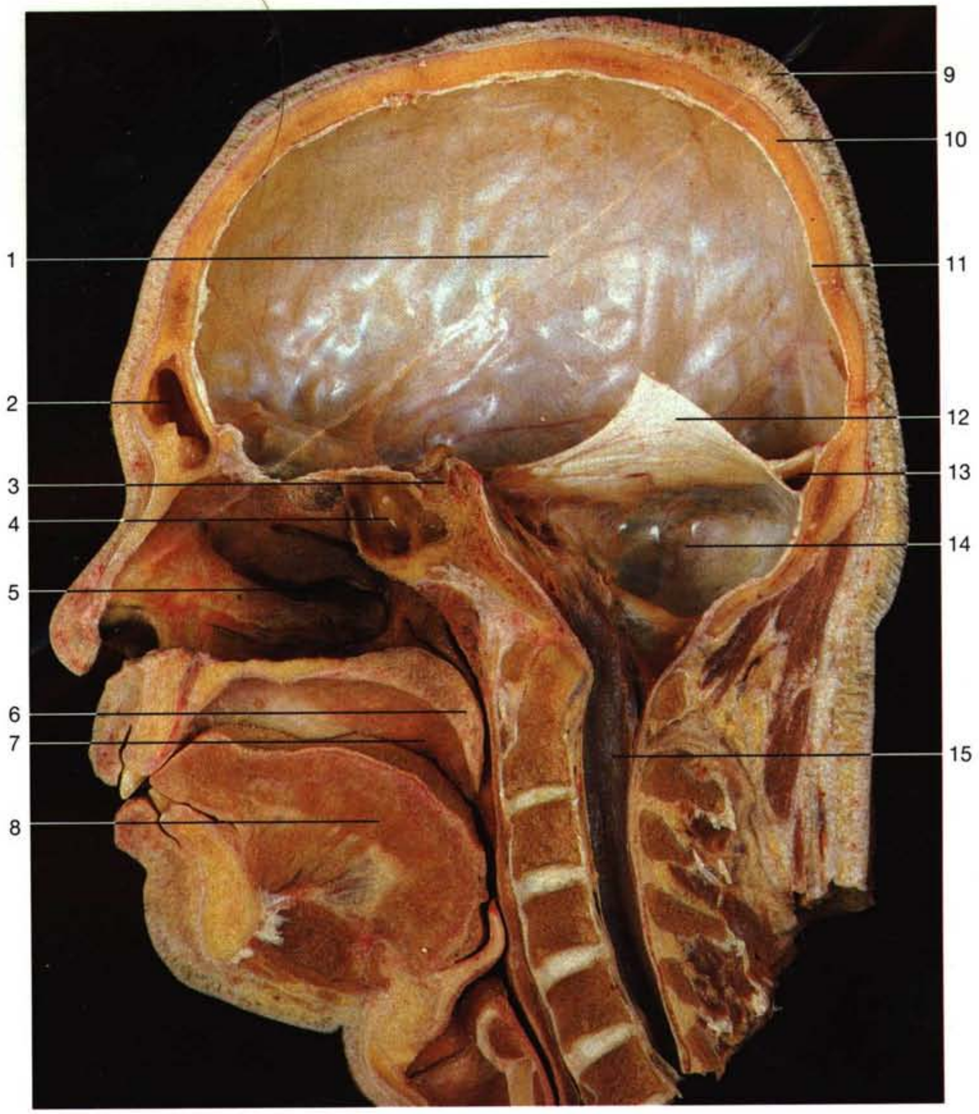

Median section through the head. Demonstration of dura mater covering the cranial cavity. Brain and spinal cord are removed (right half of the head, as seen from medial).

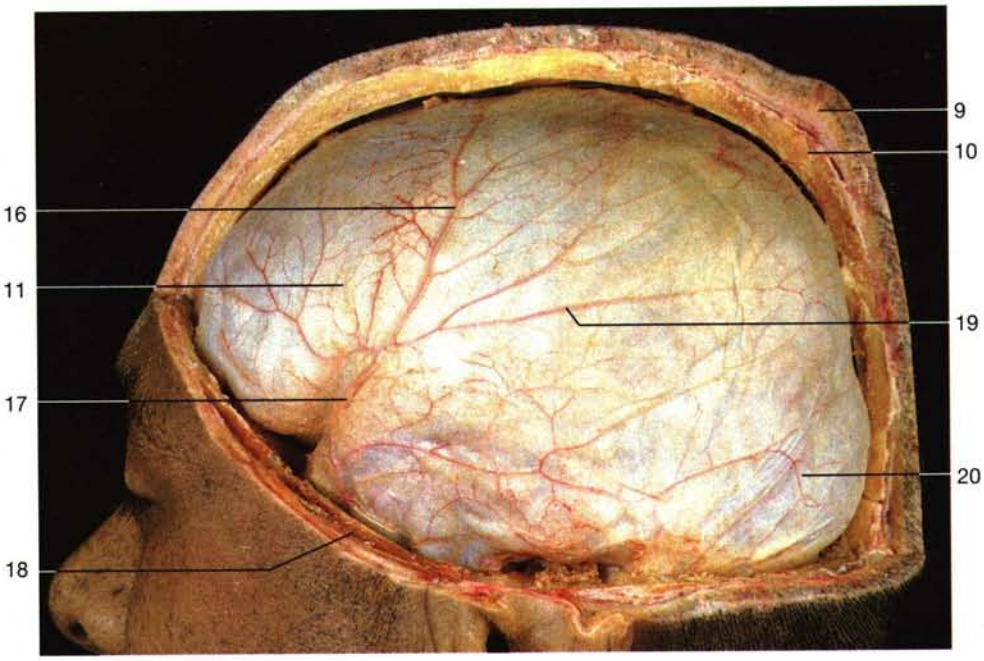

Dissection of dura mater and meningeal vessels. Left half of calvaria removed.

1 **Cranial cavity** with dura mater (the right cerebral hemisphere has been removed)
2 Frontal sinus
3 Hypophysial fossa with pituitary gland
4 Sphenoidal sinus
5 **Nasal cavity**
6 Soft palate (uvula)
7 **Oral cavity**
8 Tongue
9 Skin
10 Calvaria
11 Dura mater
12 Tentorium cerebelli
13 Confluence of sinuses
14 **Infratentorial space** (cerebellum and part of the brain stem have been removed)
15 Vertebral canal
16 Frontal branch of middle meningeal artery and veins
17 Middle meningeal artery
18 **Diploe**
19 Parietal branch of middle meningeal artery and vein
20 Occipital pole of left hemisphere covered with dura mater

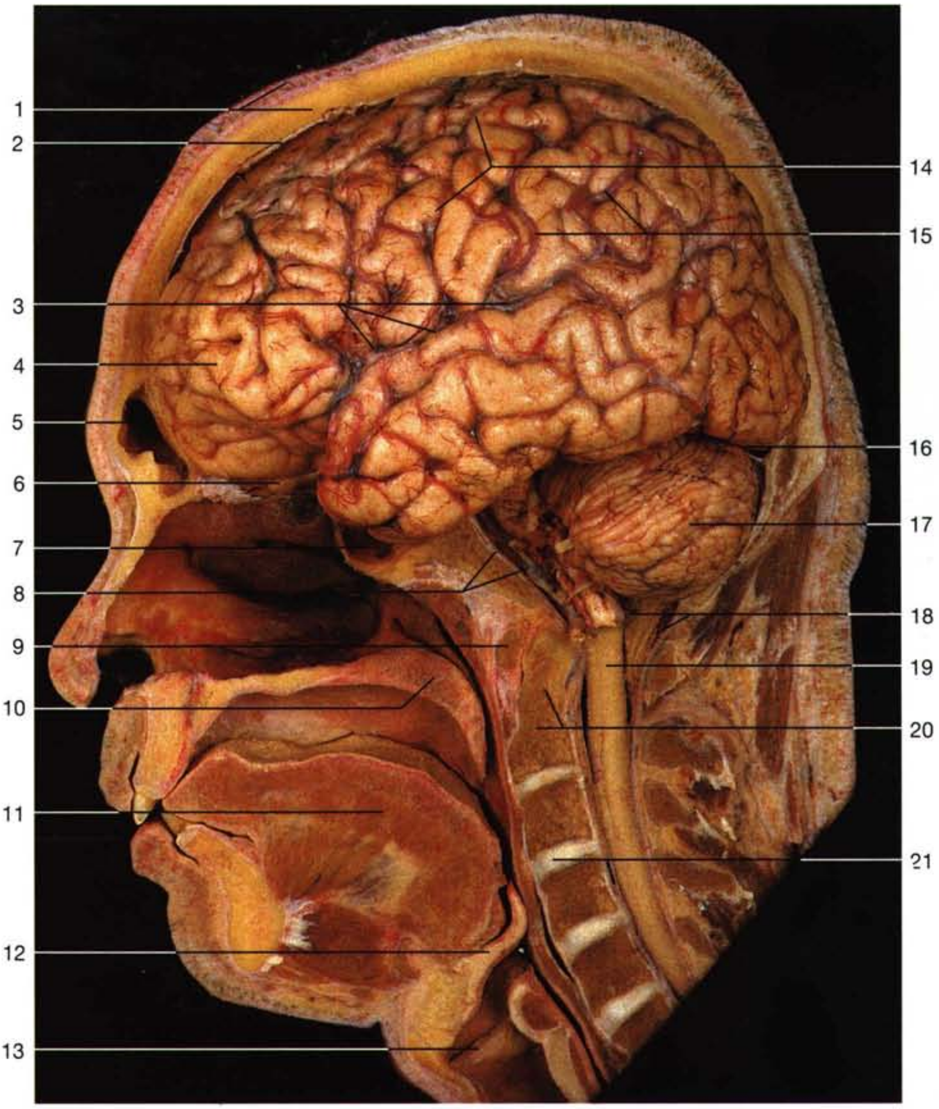

1. Calvaria and skin of the scalp
2. Dura mater (divided)
3. Position of lateral sulcus
4. **Frontal lobe** covered by **arachnoid** and **pia mater**
5. Frontal sinus
6. Olfactory bulb
7. **Sphenoidal sinus**
8. Dura mater on clivus and basilar artery
9. Atlas (anterior arch, divided)
10. **Soft palate**
11. Tongue
12. Epiglottis
13. Vocal fold
14. Position of central sulcus
15. Superior cerebral veins
16. Tentorium (divided)
17. **Cerebellum**
18. Cerebellomedullary cistern
19. Position of foramen magnum and **spinal cord**
20. Dens of axis
21. Intervertebral disc

Dissection of the brain with pia mater and arachnoid in situ. The head is cut in half except for the brain, which is shown in its entirety.

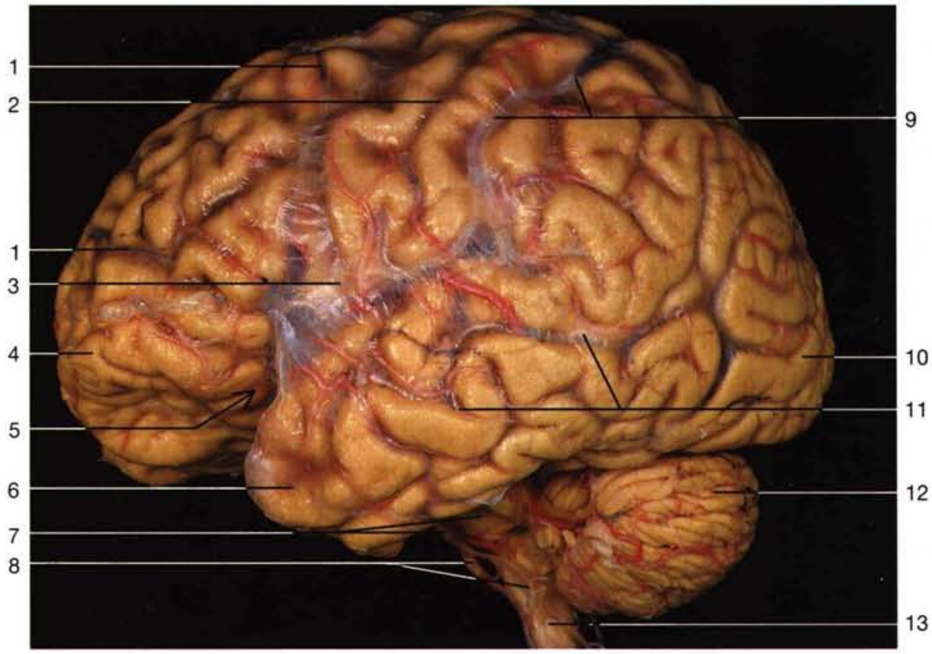

1. Superior cerebral veins
2. Position of central sulcus
3. Position of lateral sulcus and **cistern of lateral cerebral fossa**
4. **Frontal pole**
5. Lateral sulcus (arrow)
6. **Temporal pole**
7. Pons and basilar artery
8. Vertebral arteries
9. Superior anastomotic vein
10. Occipital pole
11. Inferior cerebral veins
12. Hemisphere of **cerebellum**
13. Medulla oblongata

Brain with pia mater and arachnoid. Frontal pole to the left (lateral aspect).

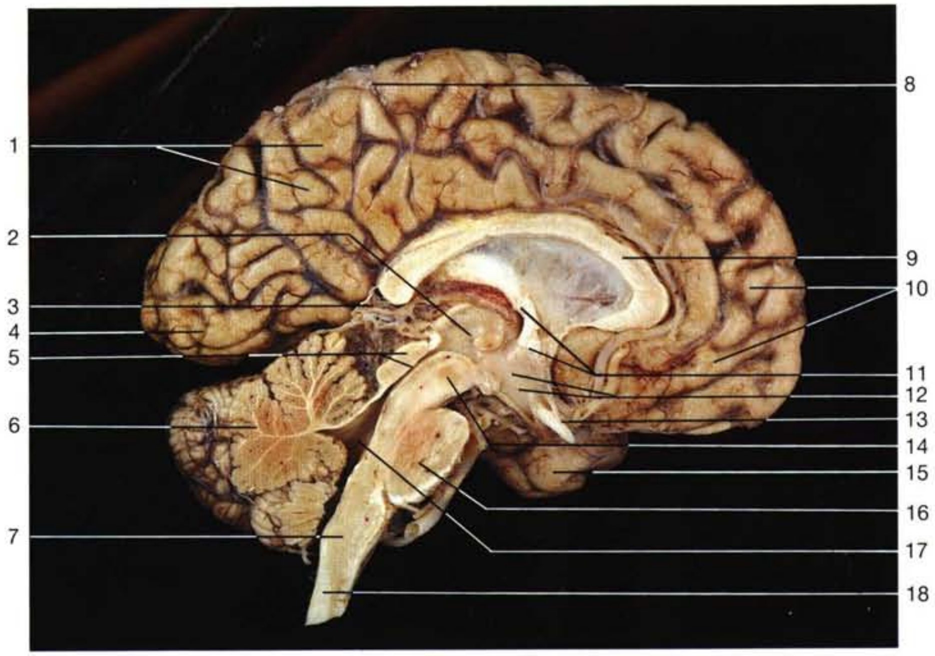
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Brain and brain stem, median section. Frontal pole to the right.

1 Parietal lobe
2 Thalamus, **third ventricle** and intermediate mass
3 Great cerebral vein
4 **Occipital lobe**
5 Colliculi of the **midbrain** and **cerebral aqueduct**
6 Cerebellum
7 **Medulla oblongata**
8 Central sulcus
9 **Corpus callosum**
10 **Frontal lobe**
11 Fornix and anterior commissure
12 **Hypothalamus**
13 Optic chiasma
14 Midbrain
15 Temporal lobe
16 **Pons**
17 **Fourth ventricle**
18 Spinal cord
19 Inferior concha and nasal cavity
20 Alveolar process of maxilla
21 Tongue
22 Dens of axis
23 Oral part of pharynx
24 Alveolar process of mandible
25 Epiglottis

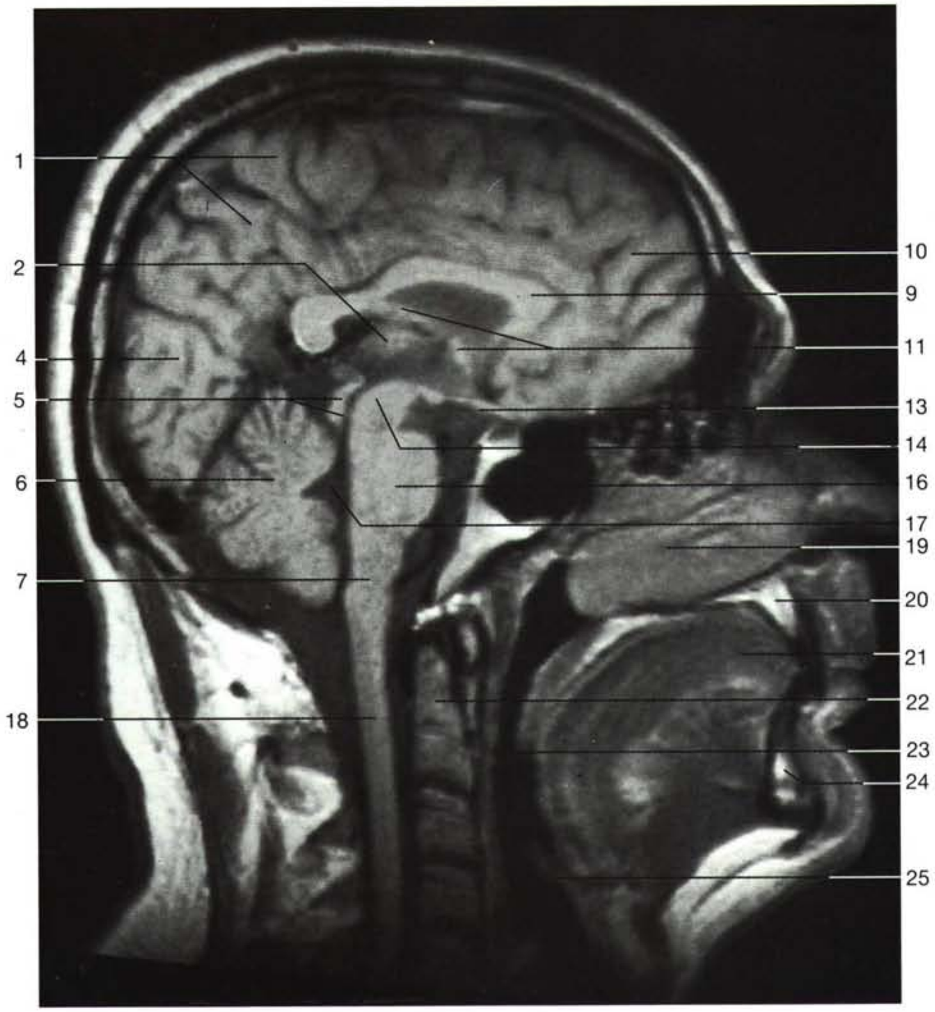
1
2
4
5
6
7
18
10
9
11
13
14
16
17
19
20
21
22
23
24
25

Median section through the head. (MR-Scan, cf. section on opposite page.)

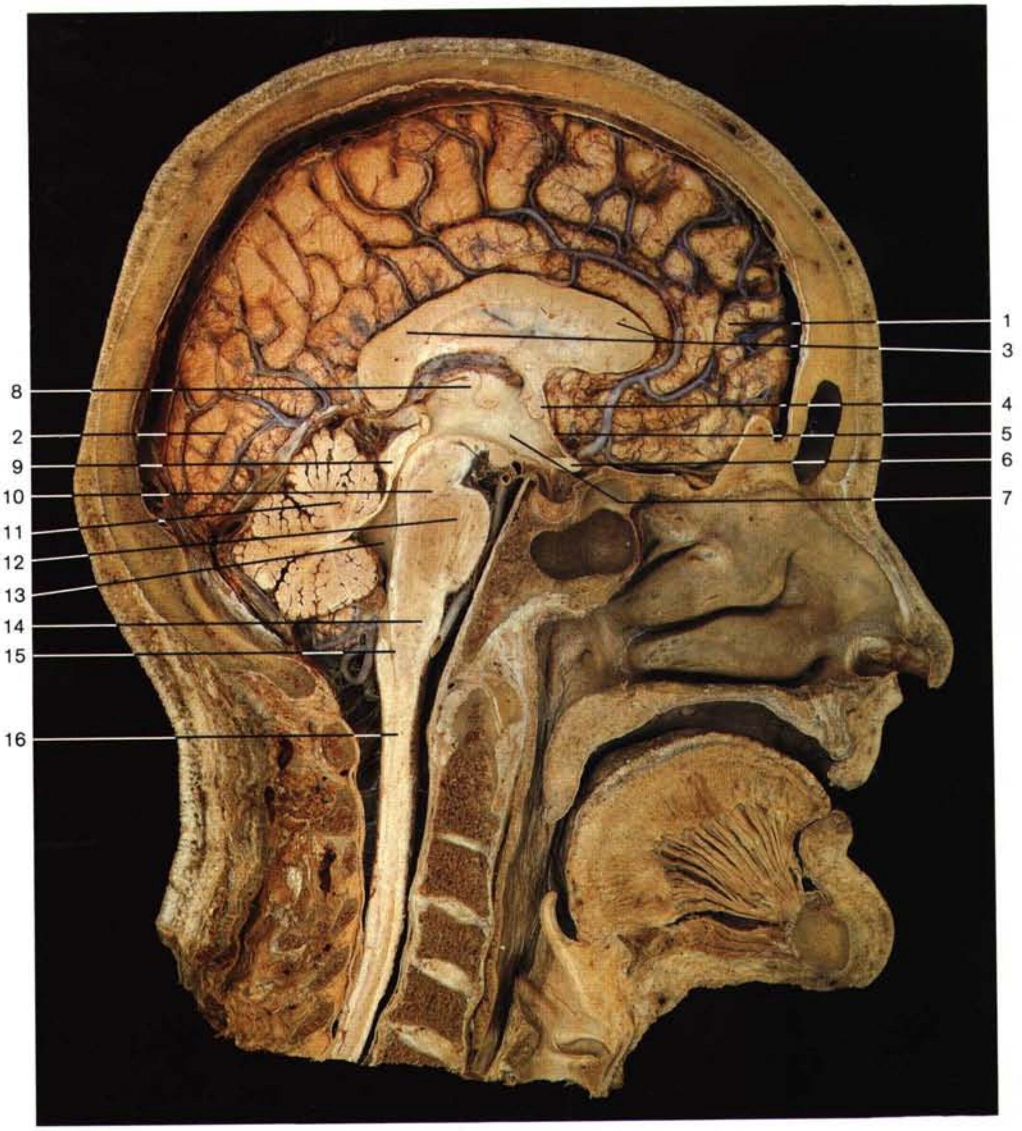

1	Frontal lobe of cerebrum
2	Occipital lobe of cerebrum
3	**Corpus callosum**
4	Anterior commissure
5	**Lamina terminalis**
6	Optic chiasma
7	**Hypothalamus**
8	Thalamus and third ventricle
9	**Colliculi of the midbrain**
10	Midbrain (inferior portion)
11	**Cerebellum**
12	Pons
13	**Fourth ventricle**
14	Medulla oblongata
15	Central canal
16	**Spinal cord**

Median section through the head. Regions of the brain. Falx cerebri removed.

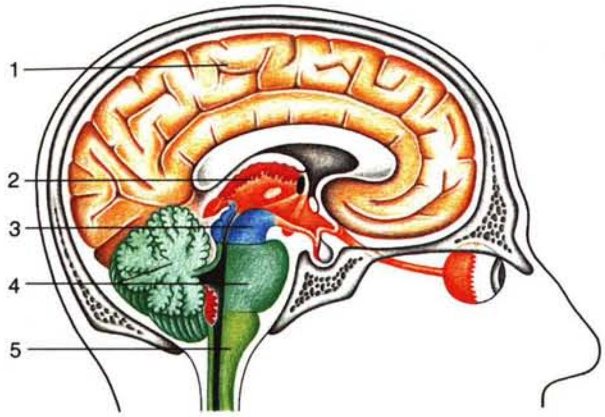

Scheme of brain divisions (cf. table).
Red = choroidal plexus. (Schematic drawing.)
1 Telencephalon (yellow) with lateral ventricles
2 Diencephalon (orange) with III ventricle, optic nerve and retina
3 Mesencephalon (blue) with cerebral aqueduct
4 Metencephalon (green) with IV ventricle
5 Myelencephalon (yellow-green)

I. Prosencephalon (forebrain)	1. Telencephalon (cerebral hemispheres, striatum, etc.) 2. Diencephalon (thalamus, metathalamus, hypothalamus, etc.)
II. Mesencephalon (midbrain)	3. Mesencephalon (colliculi, cerebral peduncles, tegmentum)
III. Rhombencephalon (hindbrain)	4. Metencephalon (pons, cerebellum) 5. Myelencephalon (medulla oblongata)

Main divisions of the brain
I-III = primary brain vesicles; 1-5 = secondary brain vesicles

Midbrain, pons and medulla oblongata are collectively termed the **brain stem.**

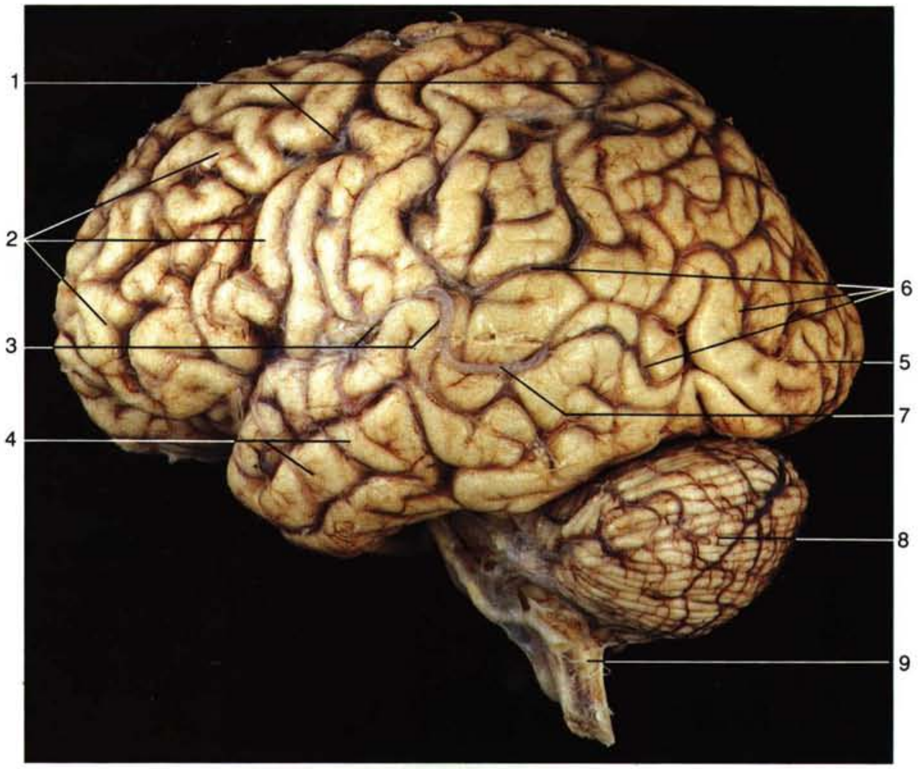

1 Superior cerebral veins and parietal lobe
2 Frontal lobe
3 Superficial middle cerebral vein and **cistern of lateral cerebral fossa**
4 Temporal lobe
5 Occipital lobe
6 Inferior cerebral veins and transverse occipital sulcus
7 Inferior anastomotic vein
8 Cerebellum
9 Medulla oblongata

Brain with pia mater. Cerebral veins (bluish). In the lateral sulcus the cistern of the lateral fossa is recognizable. Frontal lobe to the left.

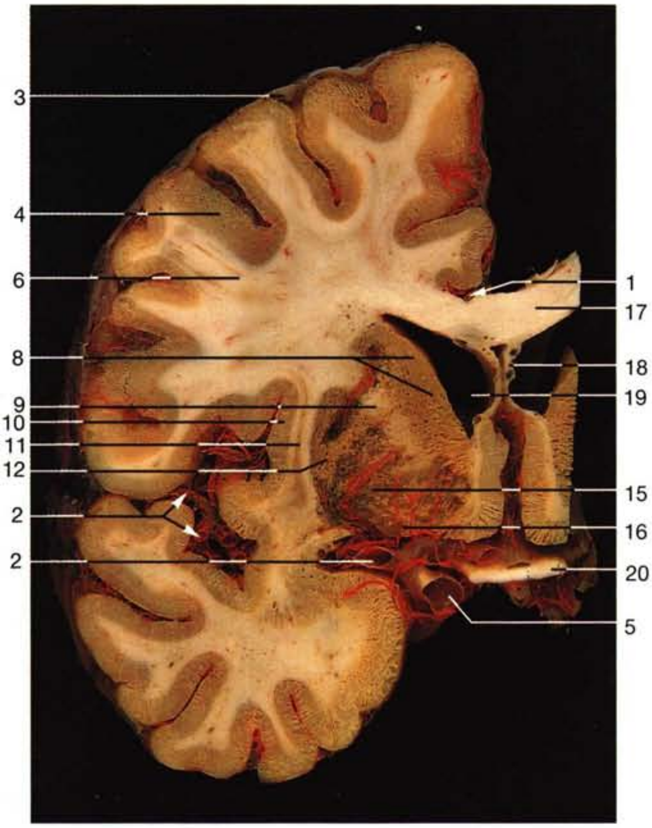

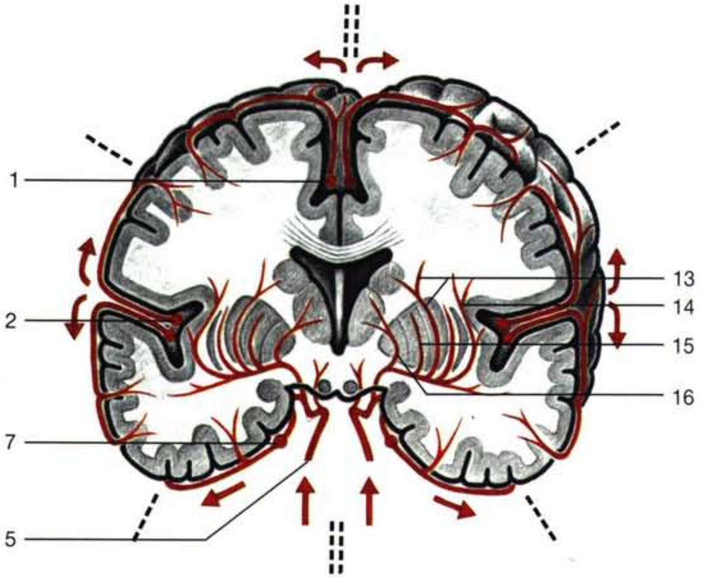

Arteries of the brain. Coronal section. Areas supplied by cortical and central arteries. Dotted lines indicate boundaries of arterial supply areas; arrows: direction of blood flow.

◁

Coronal section through the right hemisphere showing arachnoid, pia mater, and the arterial blood supply (anterior aspect).

1 **Anterior cerebral artery**
2 **Middle cerebral arteries**
3 Arachnoid
4 Cortex
5 **Internal carotid artery**
6 Frontal lobe (white matter)
7 **Posterior cerebral artery**
8 Caudate nucleus
9 Internal capsule
10 Insular lobe
11 Claustrum
12 Putamen
13 Posterior striate branch
14 Insular artery
15 Pallidostriate artery
16 Thalamic artery
17 Corpus callosum
18 Pellucid septum
19 Lateral ventricle
20 Optic chiasm

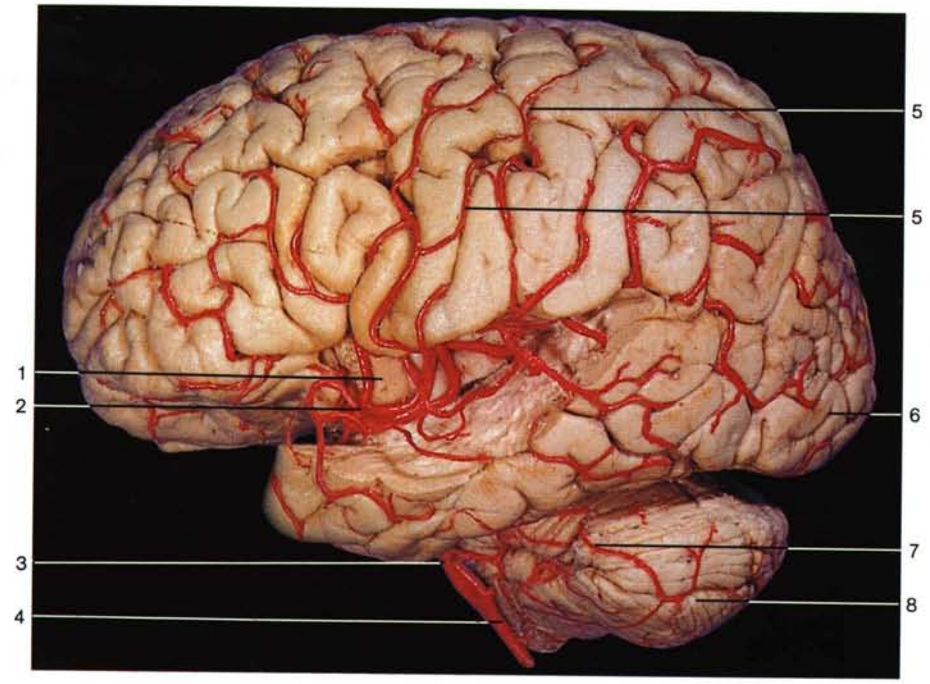

1 Insula
2 **Middle cerebral artery** [2 branches:
a) Parietal branches,
b) Temporal branches]
3 Basilar artery
4 **Vertebral artery**
5 Central sulcus
6 Occipital lobe
7 Superior cerebellar artery
8 Cerebellum
9 **Anterior cerebral artery**
10 Ethmoidal arteries
11 Ophthalmic artery
12 **Internal carotid artery**
13 Posterior communicating artery
14 **Posterior cerebral artery**
15 Anterior inferior cerebellar artery
16 Posterior inferior cerebellar artery

Cerebral arteries. Lateral aspect of the left hemisphere. The upper part of the temporal lobe has been removed to display the insula and cerebral arteries.

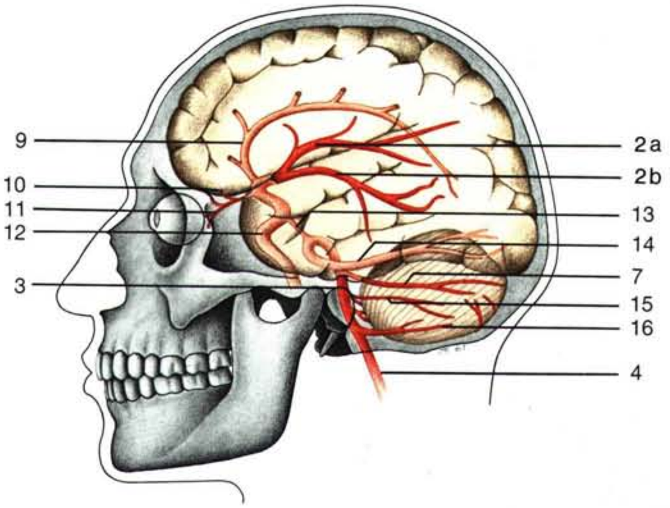

Arteries of the brain.

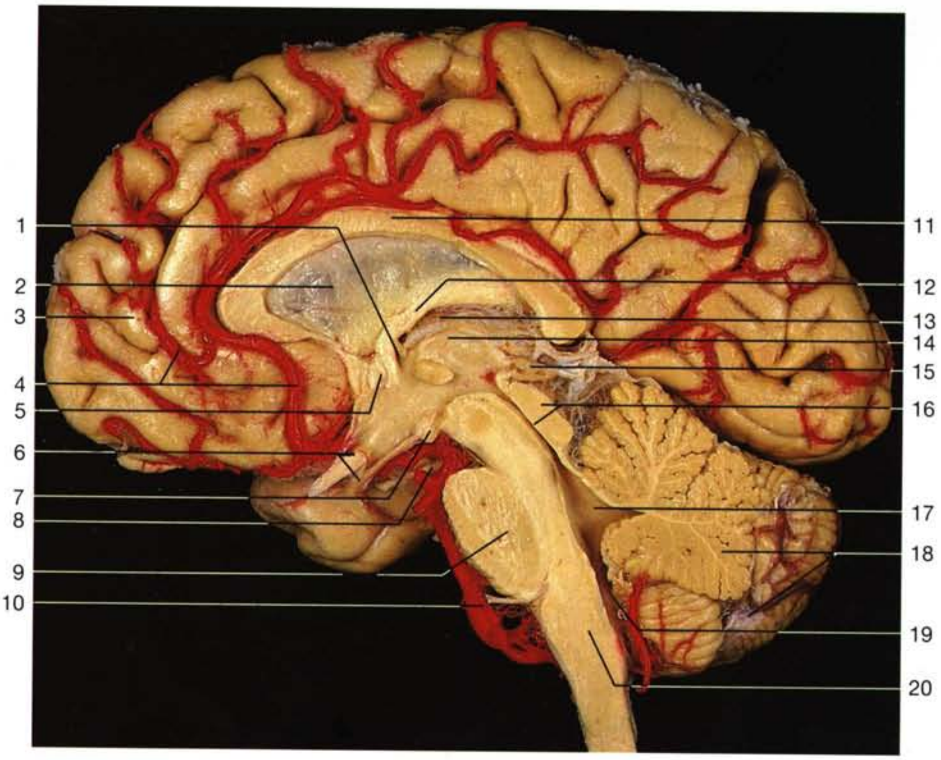

1 Interventricular foramen
2 Septum pellucidum
3 Frontal lobe
4 **Anterior cerebral artery**
5 Anterior commissure
6 Optic chiasma and infundibulum
7 Mamillary body
8 Oculomotor nerve (n. III)
9 Pons
10 **Basilar artery**
11 Corpus callosum
12 **Fornix**
13 **Choroid plexus**
14 **Third ventricle**
15 Pineal body
16 **Tectum** and **cerebral aqueduct**
17 **Fourth ventricle**
18 Cerebellum (arbor vitae, vermis)
19 Median aperture of Magendie
20 Medulla oblongata

Median section through the brain and brain stem. Cerebral arteries injected with red resin.

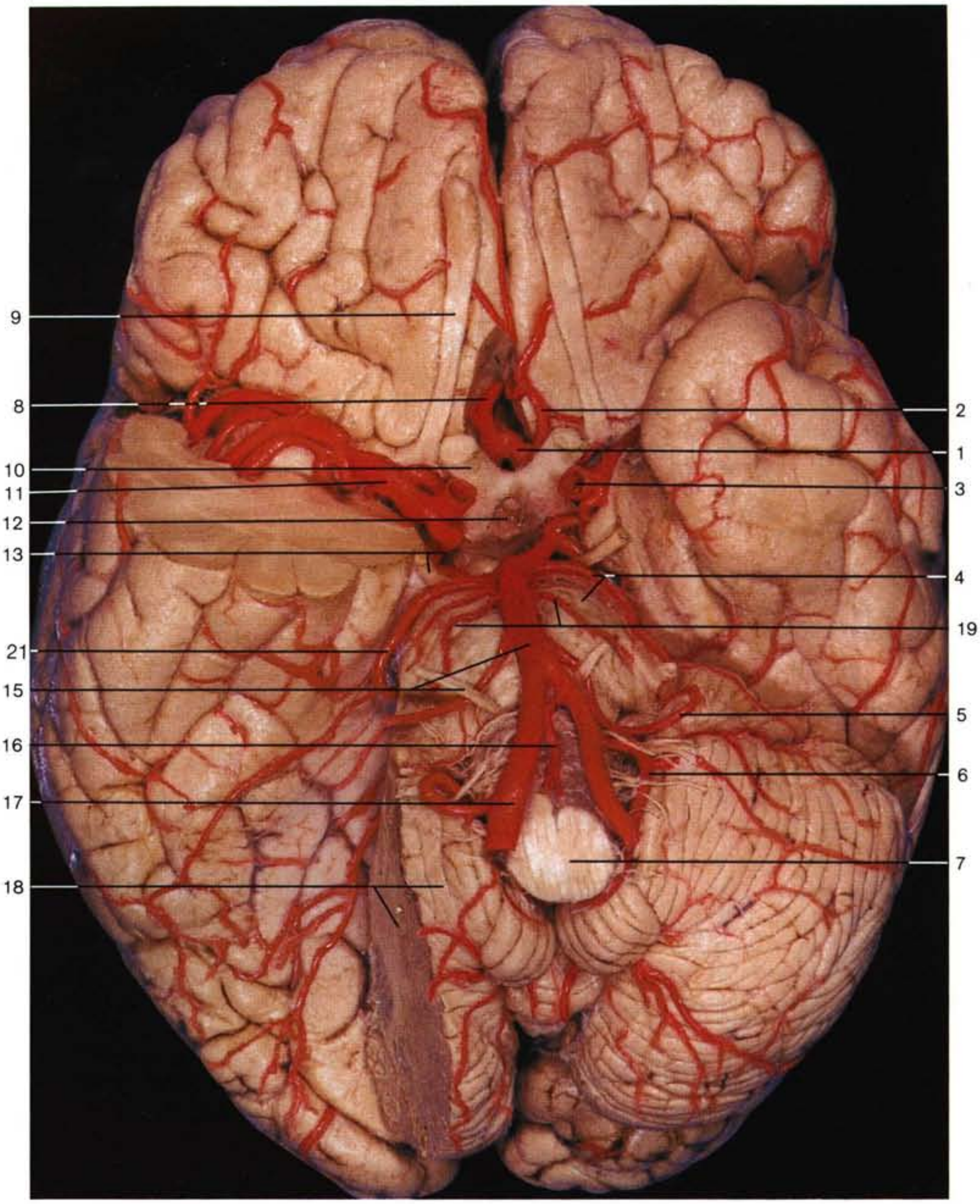

Arteries of the brain (inferior aspect). Frontal pole above; right temporal lobe and cerebellum partly removed.

1 Anterior communicating artery
2 Left **anterior cerebral artery**
3 **Internal carotid artery**
4 Pons and left **superior cerebellar artery**
5 **Anterior inferior cerebellar artery**
6 **Posterior inferior cerebellar artery**
7 Medulla oblongata
8 **Right anterior cerebral artery**
9 Olfactory tract
10 Optic nerve
11 **Middle cerebral artery**
12 Infundibulum
13 Oculomotor nerve and **posterior communicating artery**
14 **Posterior cerebral artery**
15 **Basilar artery** and abducens nerve (n. VI)
16 Anterior spinal artery
17 **Vertebral artery**
18 Cerebellum
19 Labyrinthine arteries
20 Posterior spinal artery
21 Right **superior cerebellar artery**
22 Olfactory bulb

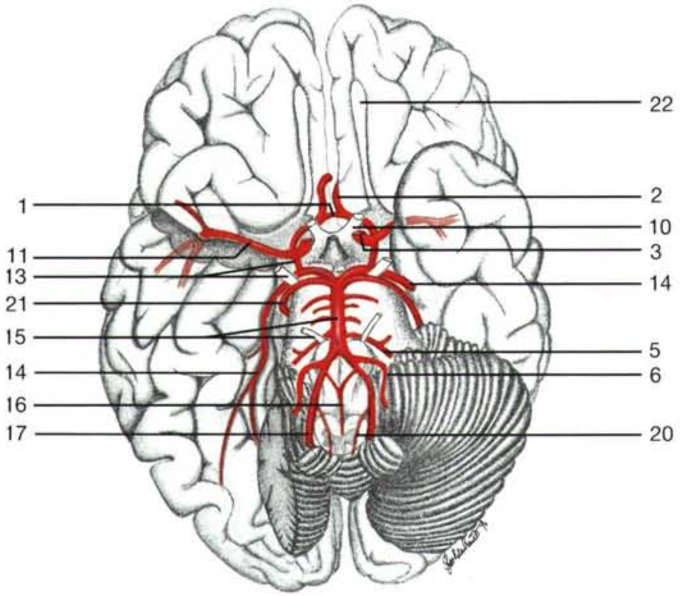

Arteries of the base of the brain, arterial circle of Willis.
(Schematic drawing.)

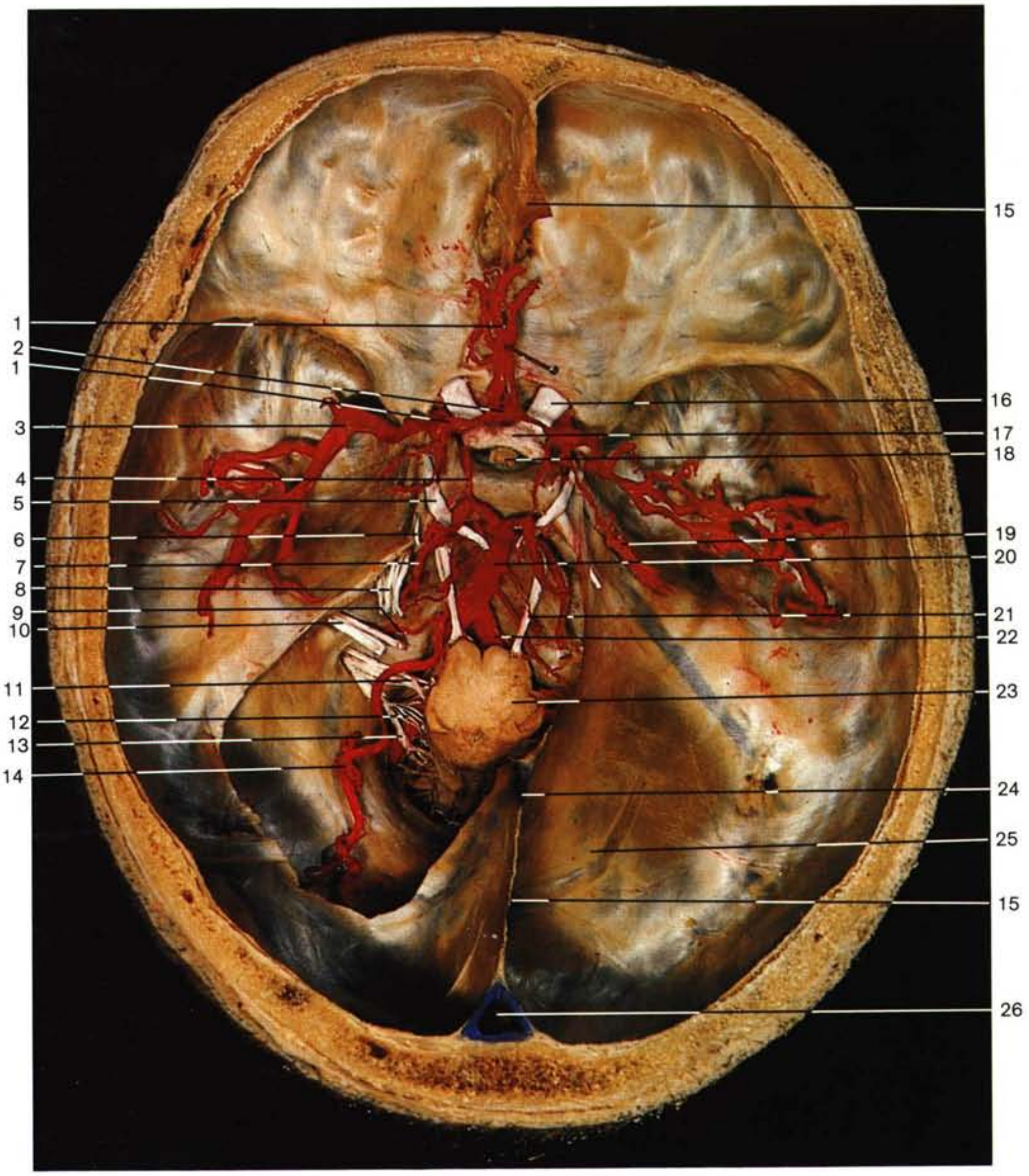

Base of the cranial cavity (internal aspect). Left tentorium has been cut.

1 Anterior cerebral artery
2 Anterior communicating artery
3 **Middle cerebral artery**
4 Posterior communicating artery
5 **Oculomotor nerve**
6 **Trochlear nerve**
7 Posterior cerebral artery
8 **Trigeminal nerve**
9 Internal auditory artery
10 **Facial nerve** and **vestibulocochlear nerve**
11 **Glossopharyngeal nerve** and **vagus nerve**
12 **Hypoglossal nerve**
13 **Accessory nerve**
14 Anterior inferior cerebellar artery
15 Falx cerebri
16 **Optic nerve**
17 Optic chiasma
18 Infundibulum and pituitary gland
19 Anterior choroidal artery and choroid plexus
20 **Basilar artery**
21 **Abducens nerve**
22 Vertebral arteries
23 Medulla oblongata
24 Inferior sagittal sinus
25 Tentorium cerebelli
26 Superior sagittal sinus and confluence of sinuses
27 Anterior spinal artery
28 Internal carotid artery
29 Superior cerebellar artery
30 Anterior inferior cerebellar artery
31 Posterior inferior cerebellar artery
32 Posterior spinal artery
33 Ophthalmic artery

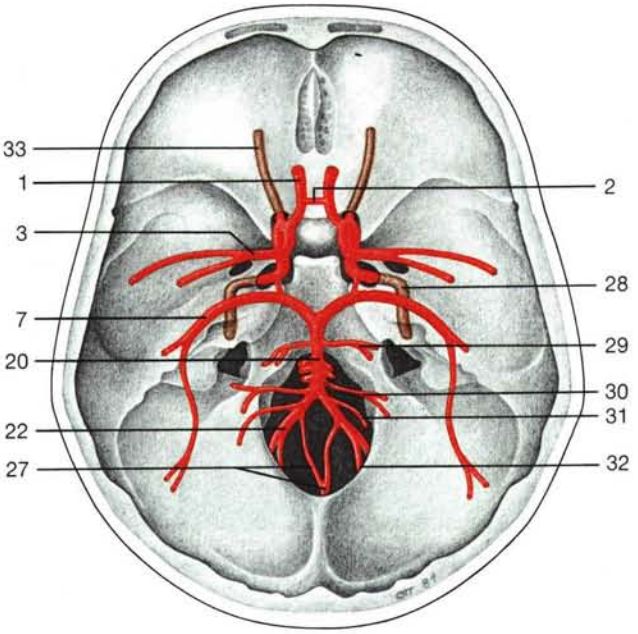

Arterial circle of Willis (superior aspect).
(Schematic drawing.)

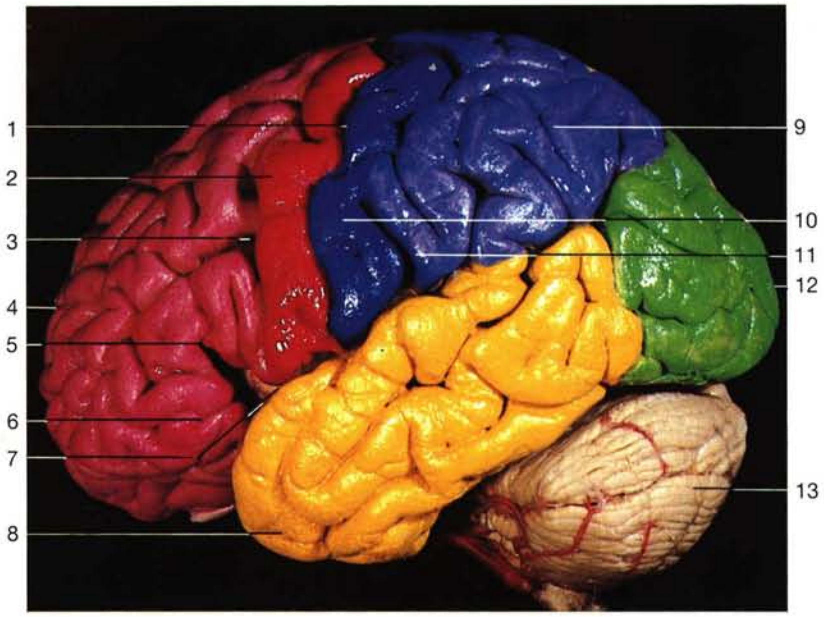

1 Central sulcus
2 Precentral gyrus
3 Precentral sulcus
4 **Frontal lobe**
5 Anterior ascending ramus of lateral sulcus
6 Anterior horizontal ramus of lateral sulcus
7 Lateral sulcus
8 **Temporal lobe**
9 **Parietal lobe**
10 Postcentral gyrus
11 Postcentral sulcus
12 **Occipital lobe**
13 Cerebellum
14 Superior frontal sulcus
15 Middle frontal gyrus
16 Lunate sulcus
17 Longitudinal fissure
18 Arachnoid granulations

Brain, left hemisphere (lateral aspect). Frontal pole to the left.

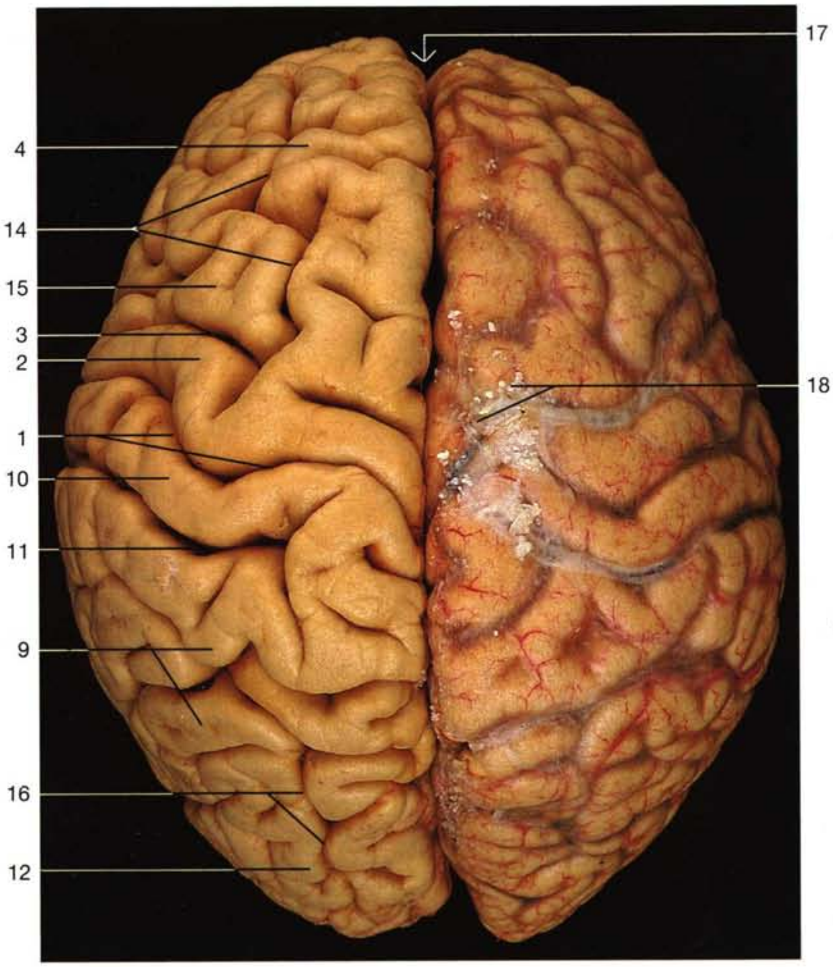

Pink = Frontal lobe
Blue = Parietal lobe
Green = Occipital lobe
Yellow = Temporal lobe
Dark red = Precentral gyrus
Dark blue = Postcentral gyrus

Brain (superior aspect). Right hemisphere with arachnoid and pia mater.

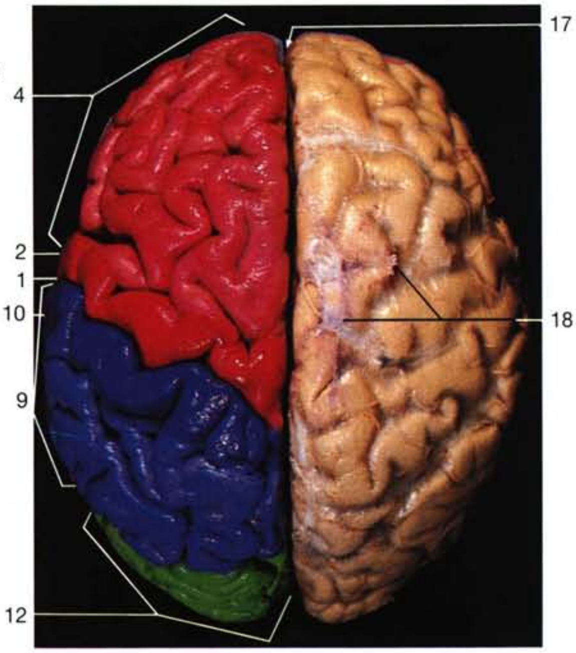

Brain (superior aspect). Lobes of the left hemisphere indicated by color; right hemisphere is covered with arachnoid and pia mater.

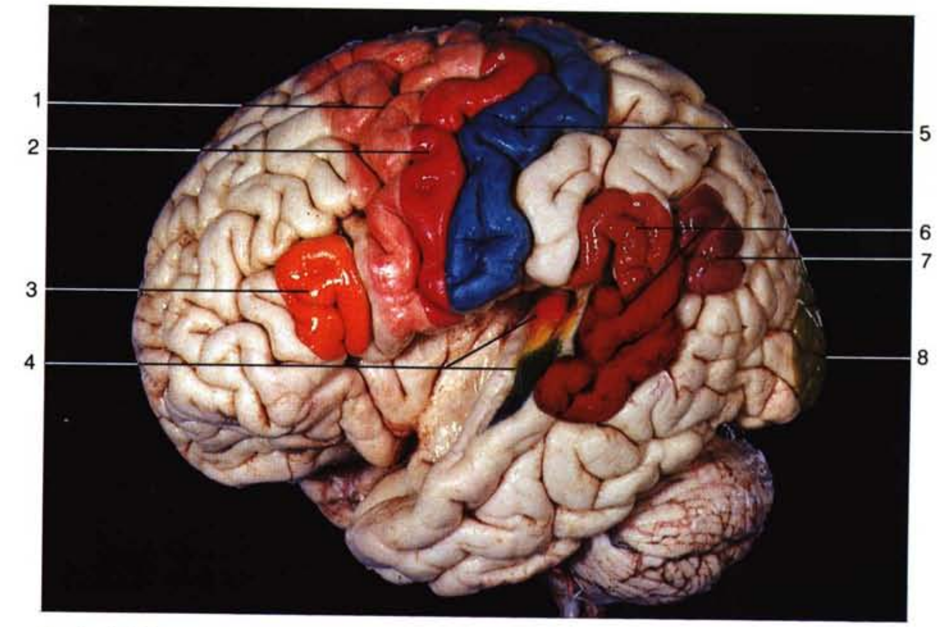

1 Premotor area
2 Somatomotor area
3 Motor speech area of Broca
4 Acoustic area
(red: high tone; blue: low tone)
5 Somatosensory area
6 Sensory speech area of Wernicke
7 Reading comprehension area
8 Visuosensory area

Brain, left hemisphere (lateral aspect). **Main cortical areas** are colored. The lateral sulcus has been opened to display the insula and the inner surface of the temporal lobe.

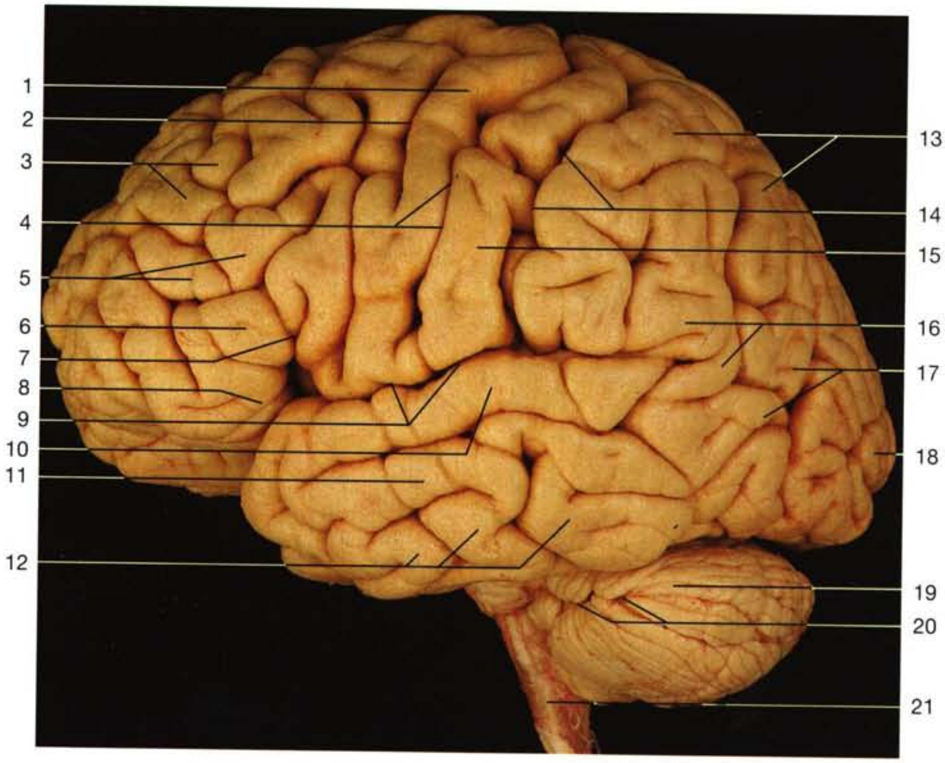

1 Precentral gyrus
2 Precentral sulcus
3 Superior frontal gyrus
4 **Central sulcus**
5 Middle frontal gyrus
6 Inferior frontal gyrus
7 Ascending ramus } of lateral sulcus
8 Horizontal ramus } of lateral sulcus
9 Posterior ramus } of lateral sulcus
10 Superior temporal gyrus
11 Middle temporal gyrus
12 Inferior temporal gyrus
13 Parietal lobule
14 Postcentral sulcus
15 Postcentral gyrus
16 Supramarginal gyrus
17 Angular gyrus
18 **Occipital lobe**
19 **Cerebellum**
20 Horizontal fissure of cerebellum
21 Medulla oblongata

Brain, left hemisphere (lateral aspect). Frontal pole to the left.

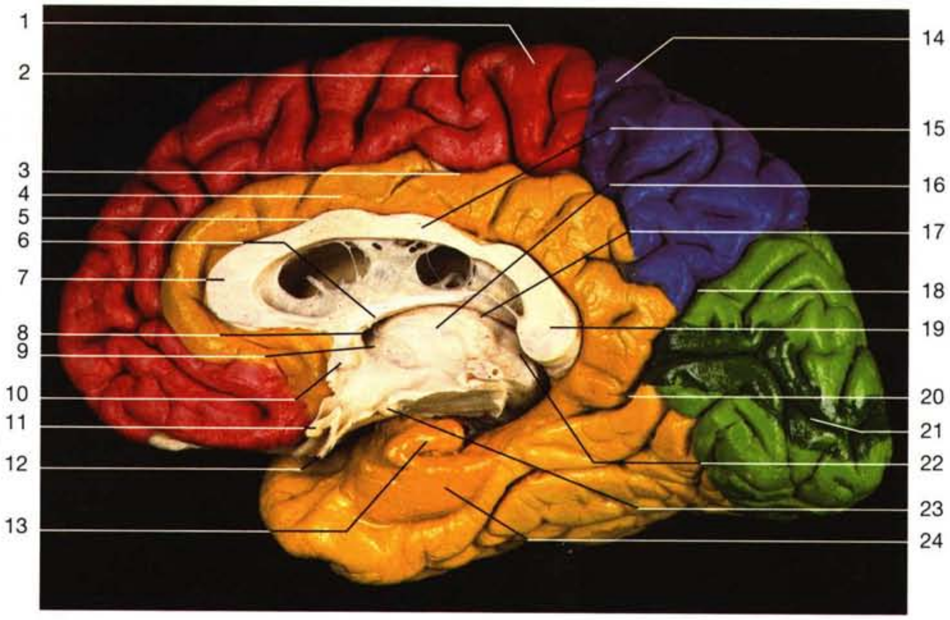

Brain, right hemisphere (medial aspect). Frontal pole to the left (midbrain divided, cerebellum and inferior part of brainstem removed).

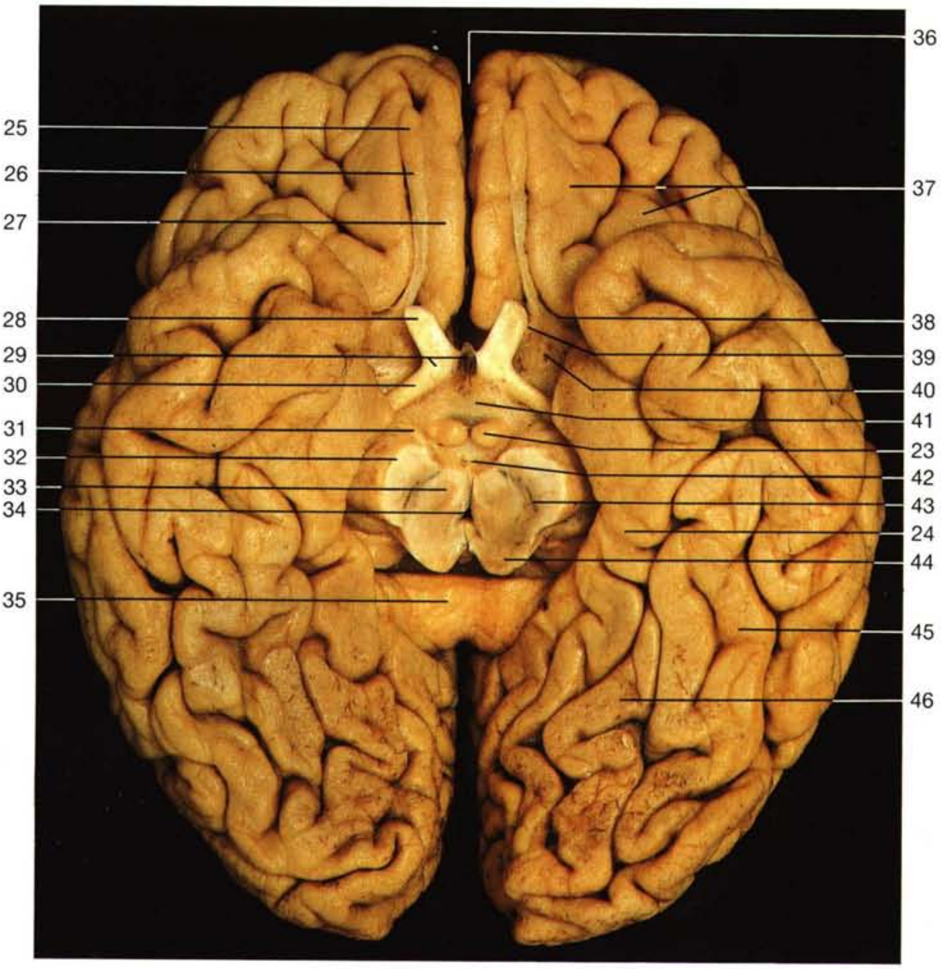

Brain (inferior aspect). Midbrain divided. Cerebellum and inferior part of brain stem removed. Frontal pole at the top.

1 **Precentral gyrus**
2 Precentral sulcus
3 Cingulate sulcus
4 **Cingulate gyrus**
5 Sulcus of corpus callosum
6 **Fornix**
7 Genu of **corpus callosum**
8 Interventricular foramen
9 Intermediate mass
10 Anterior commissure
11 **Optic chiasma**
12 **Infundibulum**
13 Uncus hippocampi
14 **Postcentral gyrus**
15 Body of corpus callosum
16 Third ventricle and thalamus
17 Stria medullaris
18 **Parietooccipital sulcus**
19 Splenium of corpus callosum
20 Communication of calcarine and parietooccipital sulcus
21 Calcarine sulcus
22 **Pineal body**
23 **Mamillary body**
24 **Parahippocampal gyrus**
25 **Olfactory bulb**
26 Olfactory tract
27 Gyrus rectus
28 **Optic nerve**
29 Infundibulum and optic chiasma
30 **Optic tract**
31 Oculomotor nerve
32 Pedunculus cerebri
33 **Red nucleus**
34 **Cerebral aqueduct**
35 Corpus callosum
36 Longitudinal fissure
37 Orbital gyri
38 Lateral root of olfactory tract
39 Medial root of olfactory tract
40 Olfactory tubercle and anterior perforated substance
41 Tuber cinereum
42 Interpeduncular fossa
43 **Substantia nigra**
44 Colliculi of the **midbrain**
45 Lateral occipitotemporal gyrus
46 Medial occipitotemporal gyrus

Pink = Frontal lobe
Blue = Parietal lobe
Green = Occipital lobe
Yellow = Temporal lobe
Dark red = Precentral lobe
Dark blue = Postcentral lobe
Orange = Limbic cortex (cingulate and parahippocampal gyri)

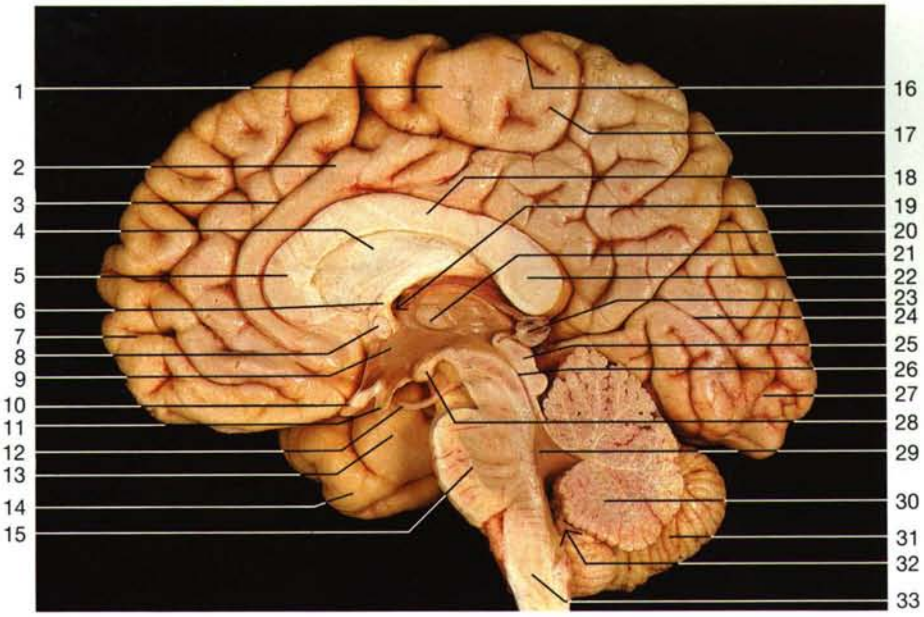

Brain (sagittal section). Frontal pole to the left.

1 Precentral gyrus
2 Cingulate gyrus
3 Cingulate sulcus
4 Septum pellucidum
5 Genu of corpus callosum
6 Fornix
7 **Frontal lobe**
8 Anterior commissure
9 **Hypothalamus**
10 Optic chiasma
11 Infundibulum
12 Oculomotor nerve
13 Uncus
14 **Temporal lobe**
15 Pons
16 Central sulcus
17 Postcentral gyrus
18 Body of corpus callosum
19 Interventricular foramen (arrow)
20 Parietooccipital sulcus
21 Intermediate mass
22 Splenium of corpus callosum
23 Pineal body
24 Calcarine sulcus
25 Colliculi of midbrain
26 Cerebral aqueduct
27 **Occipital lobe**
28 Mamillary body
29 Fourth ventricle
30 Vermis of cerebellum
31 Right hemisphere of cerebellum
32 Median aperture of Magendie (arrow)
33 Medulla oblongata
34 Olfactory tract
35 Optic nerve
36 **Internal carotid artery**
37 Interpeduncular cistern
38 Superior cerebellar artery
39 Anterior inferior cerebellar artery
40 **Vertebral artery**
41 Posterior inferior cerebellar artery
42 **Basilar artery**
43 Trigeminal nerve (n. V)
44 Facial nerve (n. VII)
45 Accessory nerve (n. XI),
hypoglossal nerve (n. XII)
46 Cerebellum

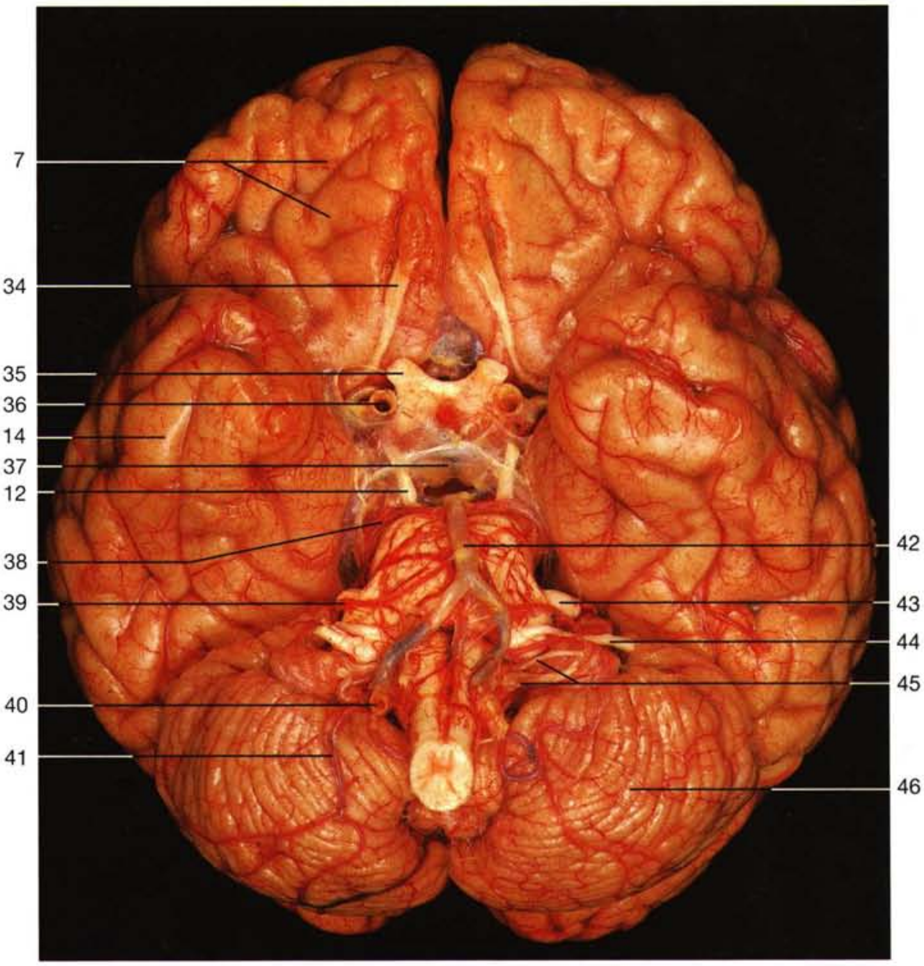

Brain, with pia mater and blood vessels (inferior aspect).

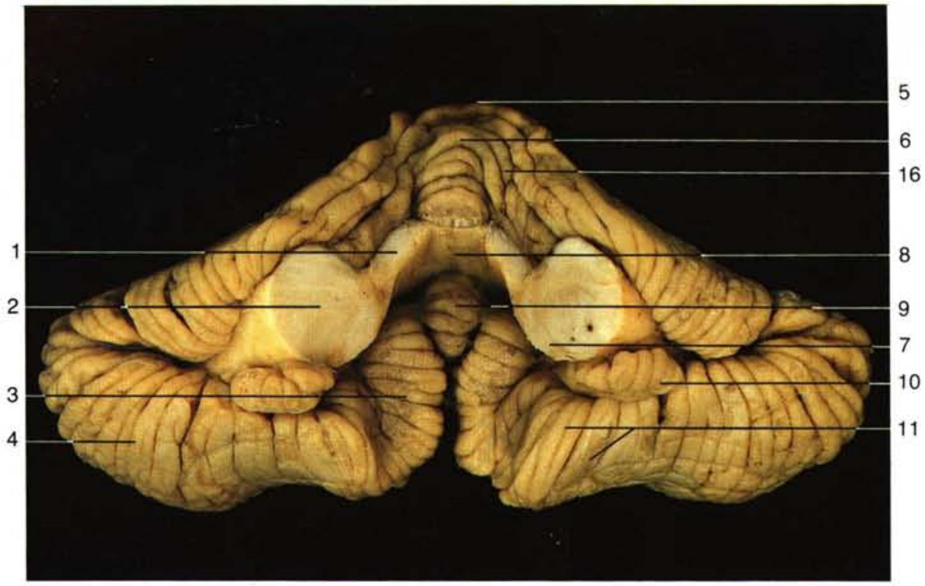

Cerebellum (inferior anterior aspect). The cerebellar peduncles have been severed.

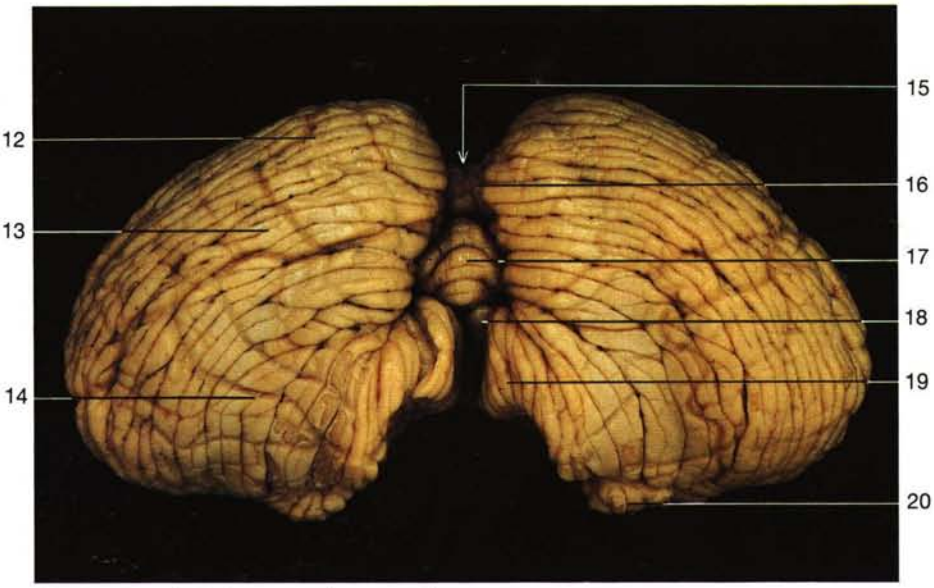

Cerebellum (inferior posterior aspect).

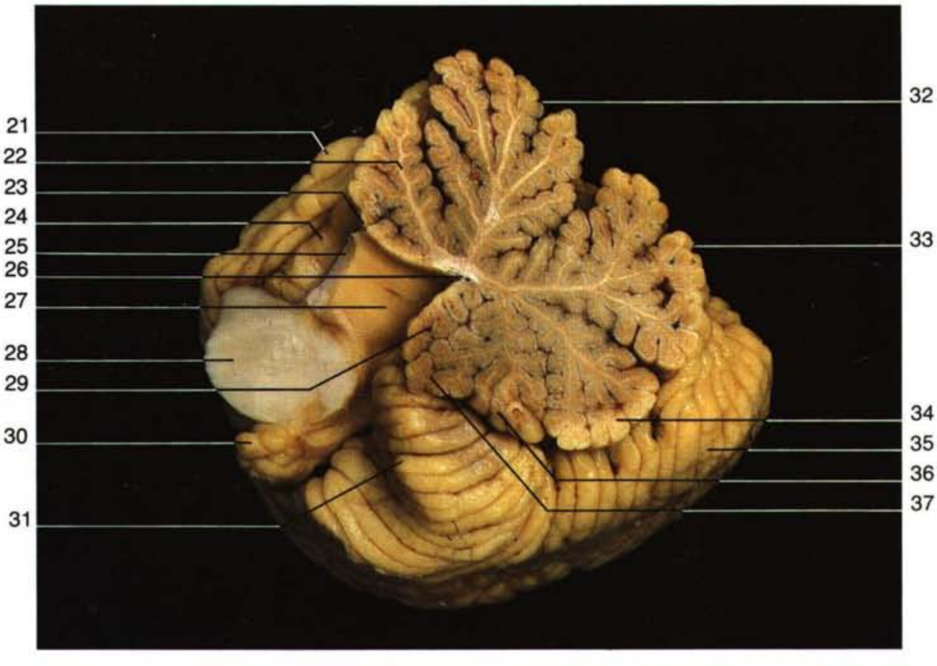

Median section through the cerebellum. Right cerebellar hemisphere and right half of vermis.

1 Superior cerebellar peduncle
2 Middle cerebellar peduncle
3 Cerebellar tonsil
4 Inferior semilunar lobule
5 **Vermis**
6 Central lobule of vermis
7 Inferior cerebellar peduncle
8 Superior medullary velum
9 **Nodule of vermis**
10 **Flocculus**
11 Biventral lobule
12 Left cerebellar hemisphere
13 Inferior semilunar lobule
14 Biventral lobule
15 Vermis of cerebellum
16 Tuber of vermis
17 Pyramid of vermis
18 Uvula of vermis
19 Tonsil of cerebellum
20 Floccule of cerebellum
21 Right cerebellar hemisphere
22 Vermis (central lobule)
23 Cerebellar lingula
24 Ala of central lobule
25 **Superior cerebellar peduncle**
26 Fastigium
27 **Fourth ventricle**
28 **Middle cerebellar peduncle**
29 Nodule of vermis
30 Flocculus of cerebellum
31 Cerebellar tonsil
32 Culmen of vermis
33 Declive of vermis
34 Tuber of vermis
35 Inferior semilunar lobule
36 Pyramis of vermis
37 Uvula of vermis

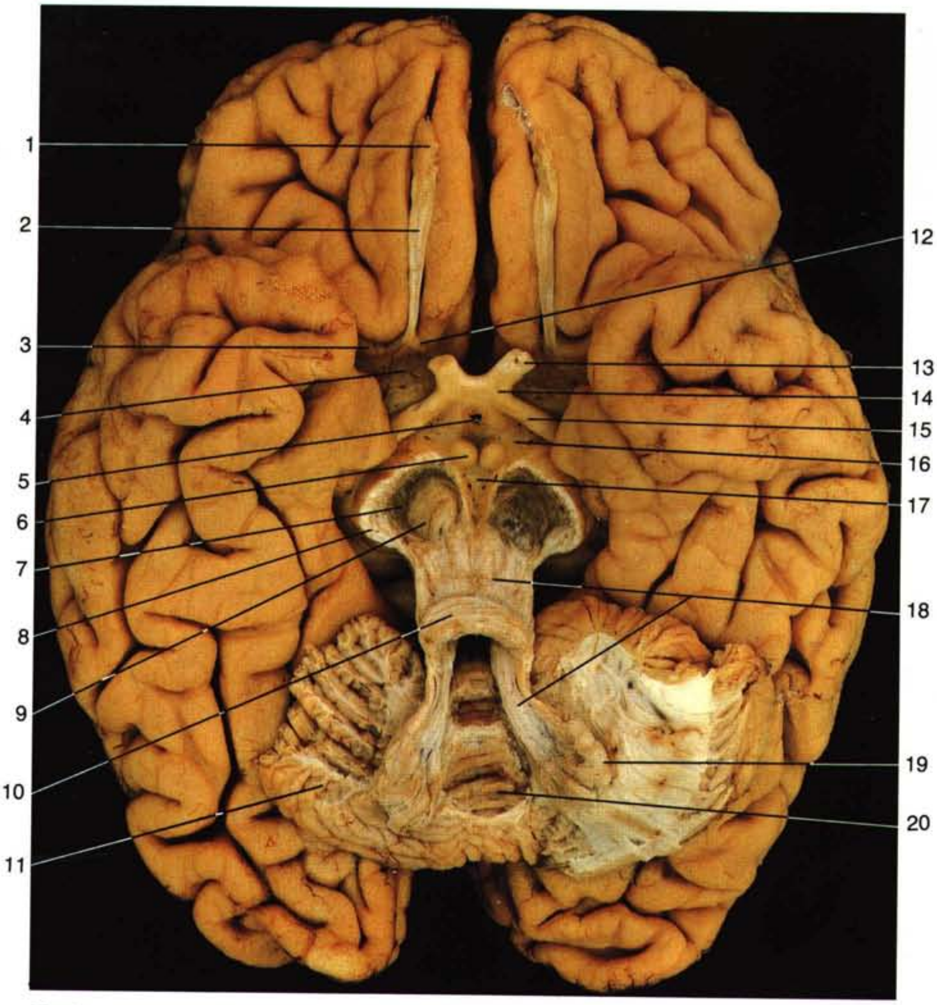

Brain and cerebellum (inferior aspect). Parts of the cerebellum have been removed to display the dentate nucleus and the main pathway to the midbrain (cerebellorubral tract).

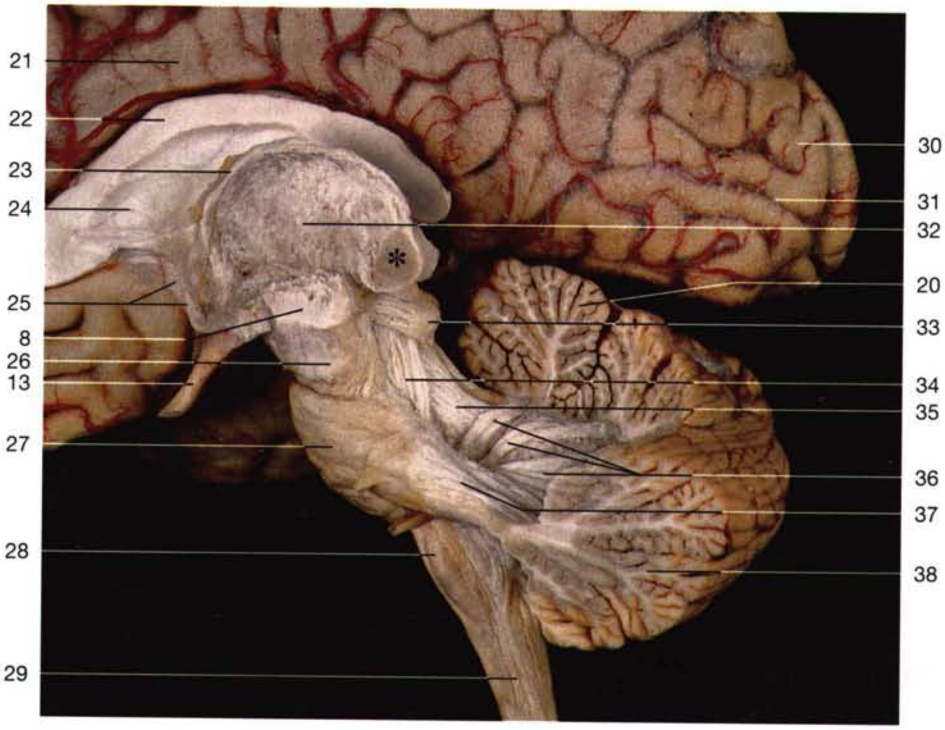

Dissection of the cerebellar peduncles and their connection with midbrain and diencephalon. A small part of pulvinar thalami (*) has been cut to show inferior brachium.

1 Olfactory bulb
2 **Olfactory tract**
3 Lateral olfactory stria
4 Anterior perforated substance
5 Infundibulum (divided)
6 **Mamillary body**
7 Substantia nigra
8 Pedunculus cerebri
9 **Red nucleus**
10 Decussation of superior cerebellar peduncle
11 **Cerebellar hemisphere**
12 Medial olfactory stria
13 **Optic nerve**
14 Optic chiasma
15 Optic tract
16 Posterior perforated substance
17 Interpeduncular fossa
18 **Superior cerebellar peduncle** and cerebellorubral tract
19 **Dentate nucleus**
20 **Vermis of cerebellum**
21 Cingulate gyrus
22 **Corpus callosum**
23 Stria terminalis
24 Septum pellucidum
25 **Columna fornicis**
26 Cerebral peduncle at midbrain level
27 Pons
28 Inferior olive
29 Medulla oblongata with lateral pyramidal tract
30 Occipital lobe
31 Calcarine sulcus
32 **Thalamus**
33 Inferior colliculus with brachium
34 **Medial lemniscus**
35 **Superior cerebellar peduncle**
36 **Inferior cerebellar peduncle**
37 **Middle cerebellar peduncle**
38 Cerebellar hemisphere

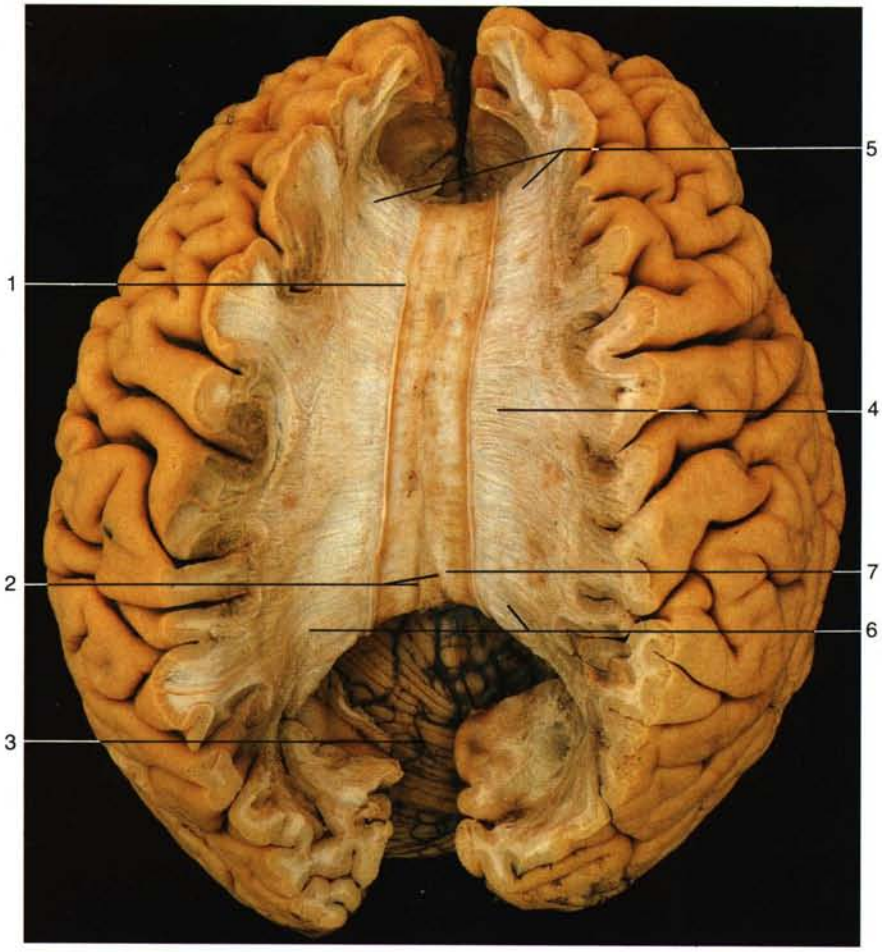

1 Lateral longitudinal stria of indusium griseum
2 Medial longitudinal stria of indusium griseum
3 Cerebellum
4 Radiating fibers of the corpus callosum
5 Forceps minor of corpus callosum
6 Forceps major of corpus callosum
7 Splenium of corpus callosum

Dissection of the brain I. The fiber system of the corpus callosum has been displayed by removing the cortex lying above it. Frontal pole at top.

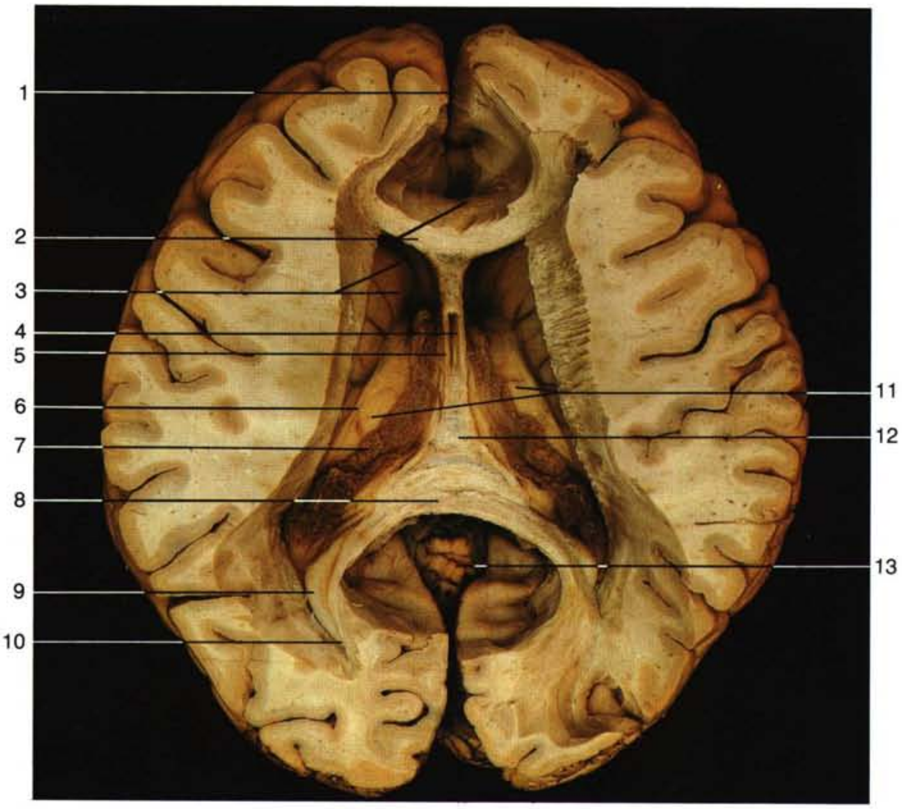

1 Longitudinal cerebral fissure
2 Genu of corpus callosum
3 Head of **caudate nucleus** and anterior horn of lateral ventricle
4 Cavum of septum pellucidum
5 **Septum pellucidum**
6 Stria terminalis
7 **Choroid plexus** of **lateral ventricle**
8 Splenium of **corpus callosum**
9 Calcar avis
10 Posterior horn of lateral ventricle
11 **Thalamus** (lamina affixa)
12 Commissure of fornix
13 Vermis of cerebellum

Dissection of the brain II. The lateral ventricles and subcortical nuclei of the brain are dissected. The corpus callosum has been partly removed. Frontal pole at the top.

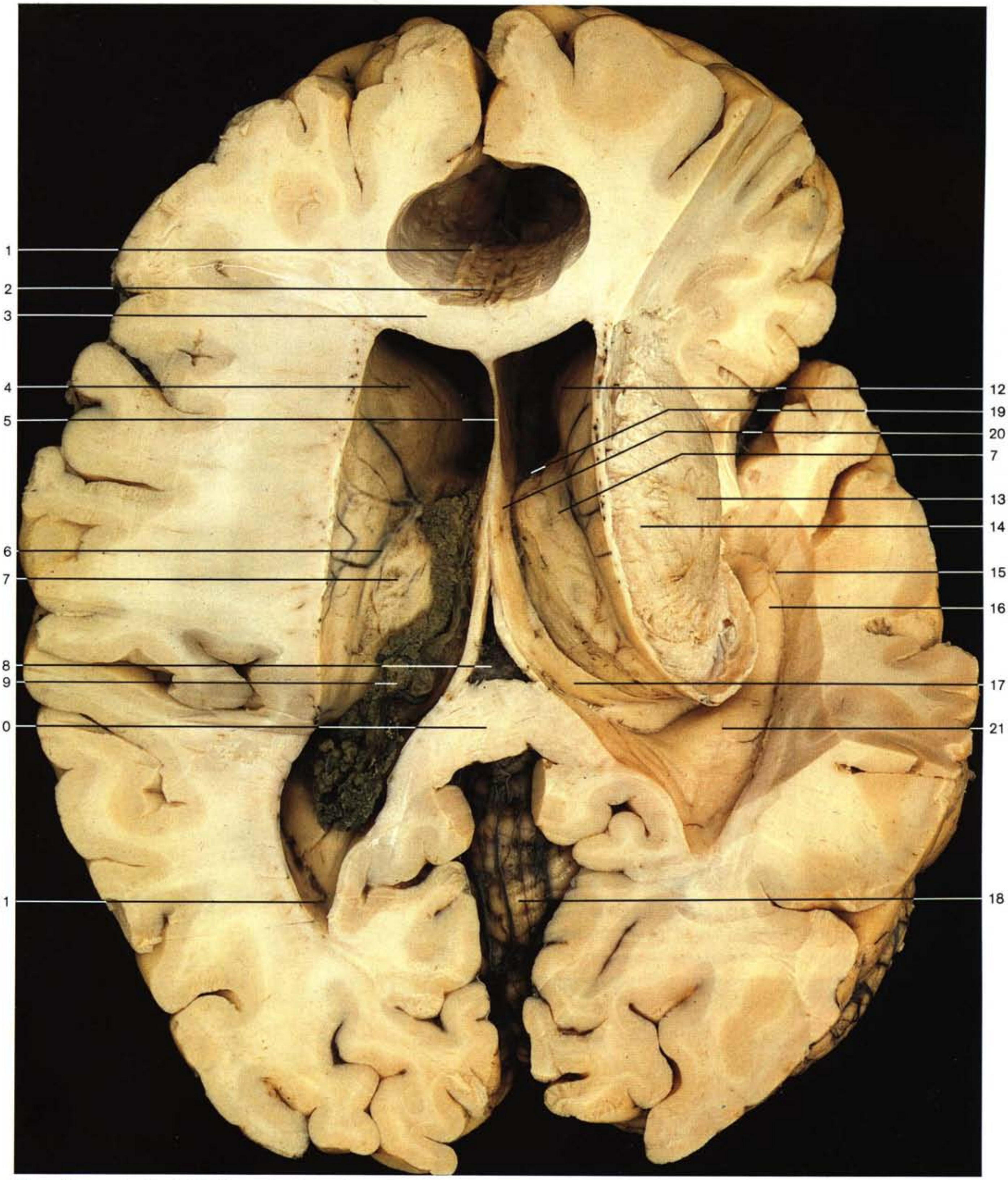

Dissection of the brain III (superior view of lateral ventricle and subcortical nuclei of the brain). Corpus callosum partly removed. At right, the entire lateral ventricle has been opened, the insula with claustrum, extreme and external capsules have been removed, exposing the lentiform nucleus and the internal capsule.

1	Lateral longitudinal stria	9	**Choroid plexus** of **lateral ventricle**	16	**Pes hippocampi**
2	Medial longitudinal stria	10	Splenium of corpus callosum	17	**Crus of fornix**
3	Genu of corpus callosum	11	Posterior horn of lateral ventricle	18	Vermis of cerebellum with arachnoid and pia mater
4	Head of **caudate nucleus**	12	Anterior horn of lateral ventricle (head of caudate nucleus)	19	Interventricular foramen
5	**Septum pellucidum**	13	**Putamen** of **lentiform nucleus**	20	Right column of fornix
6	Stria terminalis	14	**Internal capsule**	21	Collateral eminence
7	**Thalamus** (lamina affixa)	15	Inferior horn of lateral ventricle		
8	Choroid plexus of third ventricle				

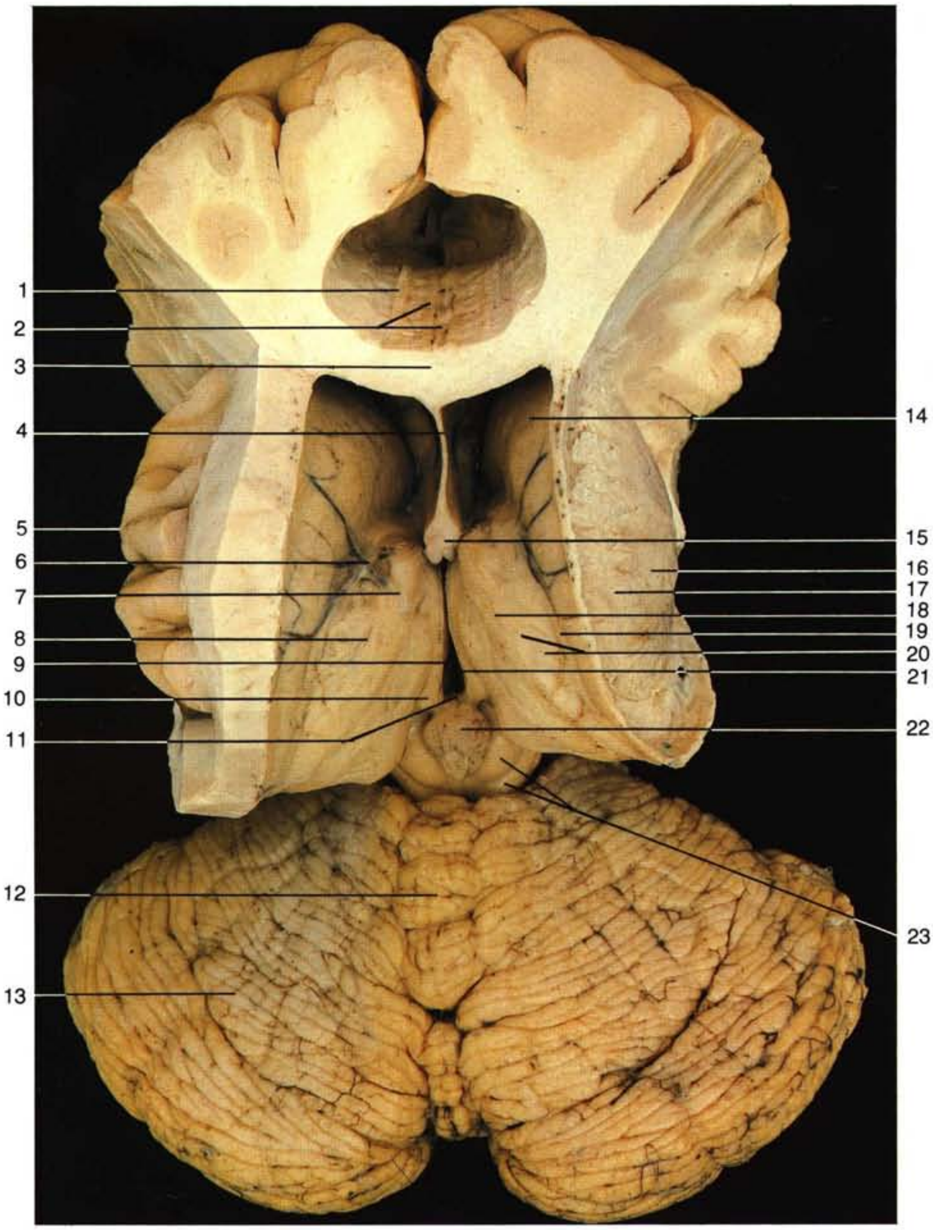

1	Lateral longitudinal stria
2	Medial longitudinal stria
3	**Corpus callosum**
4	Septum pellucidum
5	Insular gyri
6	Thalamostriate vein
7	Anterior tubercle of thalamus
8	**Thalamus**
9	Medullary stria of thalamus
10	Habenular trigone
11	Habenular commissure
12	**Vermis of the cerebellum**
13	Left **hemisphere of cerebellum**
14	**Head of caudate nucleus**
15	**Columns of fornix**
16	Putamen of lentiform nucleus
17	**Internal capsule**
18	Taenia of choroid plexus
19	Stria terminalis and thalamostriate vein
20	Lamina affixa
21	**Third ventricle**
22	**Pineal body**
23	Superior and inferior colliculus of midbrain

Dissection of the brain IVa. Temporal lobe, fornix and the posterior corpus callosum have been removed (this part of the specimen is depicted below). Frontal pole at top (superior aspect).

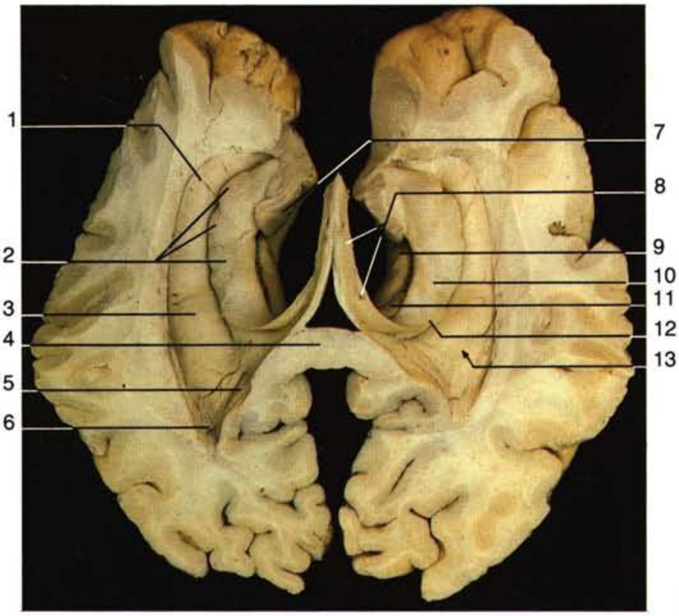

1	Inferior horn of lateral ventricle
2	**Hippocampal digitations**
3	Collateral eminence
4	Splenium of corpus callosum
5	Calcar avis
6	Posterior horn of lateral ventricle
7	**Uncus** of parahippocampal gyrus
8	Body and crus of **fornix**
9	**Parahippocampal gyrus**
10	**Pes hippocampi**
11	Dentate gyrus
12	Hippocampal fimbria
13	Lateral ventricle

Dissection of the brain IVb. Depicted is the portion of the brain removed from the specimen above. **Temporal lobe and limbic system** (superior aspect). Columns of fornix are served.

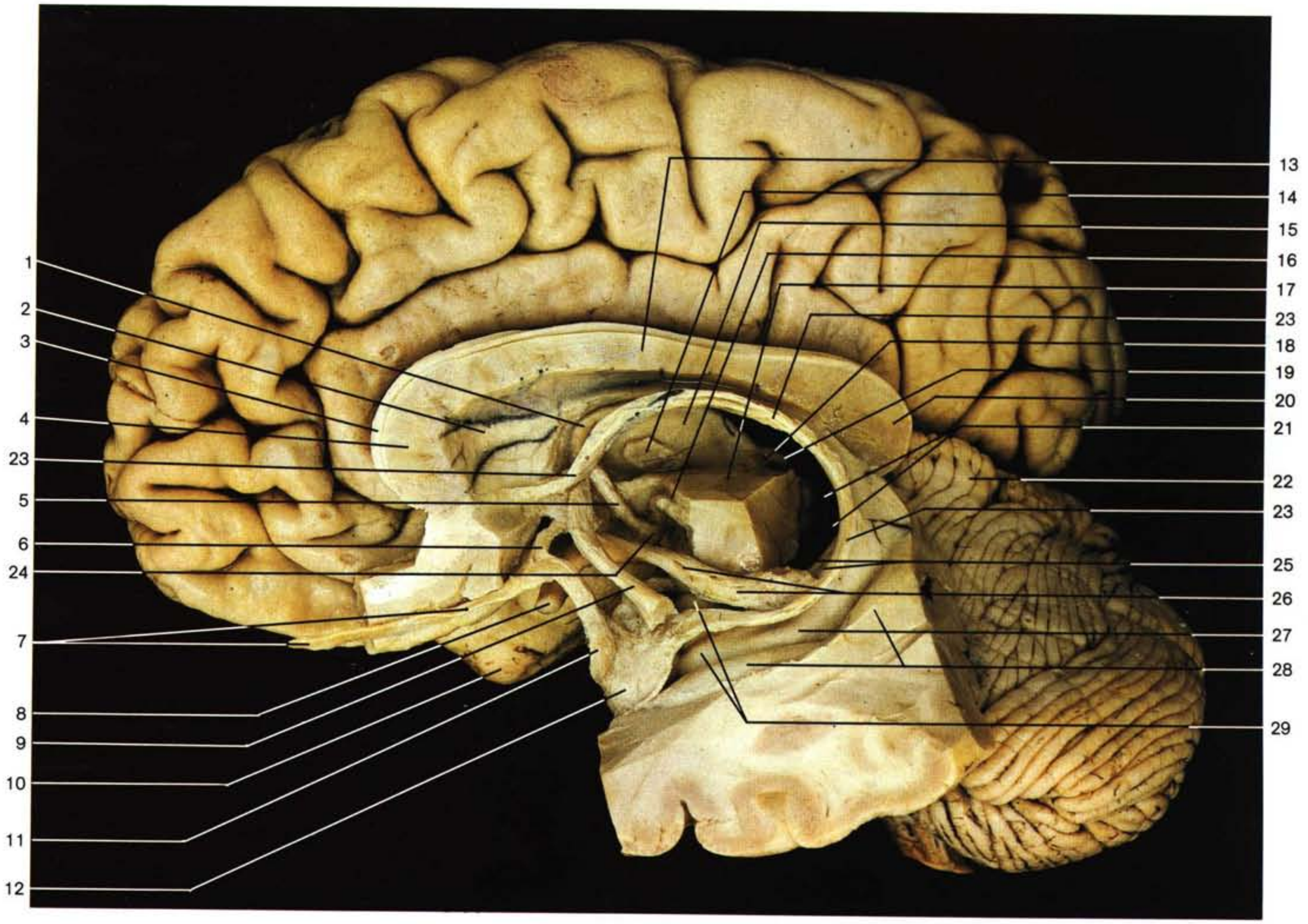

Dissection of the limbic system. Left side, lateral aspect. Corpus callosum has been cut in the median plane. The left thalamus and the left hemisphere have been partly removed.

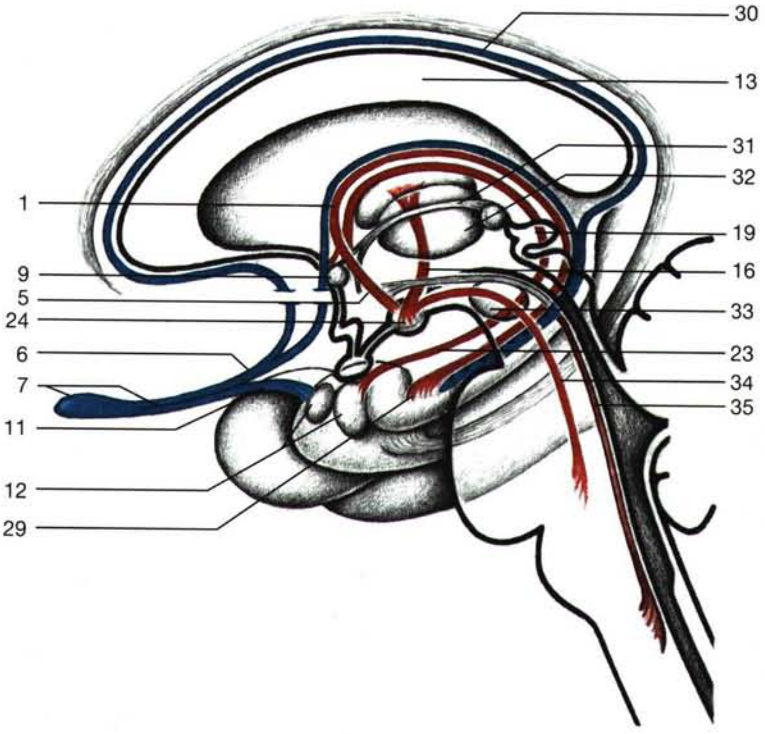

Main pathways of **limbic and olfactory system.** (Schematic drawing.) Blue = afferent pathways; red = efferent pathways.

1 **Body of fornix**
2 Septum pellucidum
3 **Lateral longitudinal stria**
4 Genu of corpus callosum
5 **Column of fornix**
6 Medial olfactory stria
7 **Olfactory bulb** and **olfactory tract**
8 Optic nerve
9 Anterior commissure (left half)
10 Right temporal lobe
11 **Lateral olfactory stria**
12 **Amygdala**
13 Body of **corpus callosum**
14 Interthalamic adhesion
15 Third ventricle and right thalamus
16 **Mamillothalamic tract**
17 Part of the thalamus
18 Habenular commissure
19 Pineal body
20 Splenium of corpus callosum
21 Colliculi of midbrain
22 Vermis of cerebellum
23 **Stria terminalis**
24 **Mamillary body**
25 Fimbria of hippocampus and pes hippocampi
26 Left optic tract and lateral geniculate body
27 Lateral ventricle and **parahippocampal gyrus**
28 Collateral eminence
29 Hippocampal digitations
30 **Supracallosal gyrus** (longitudinal stria)
31 Stria medullaris thalami
32 Thalamus
33 Red nucleus
34 Mamillotegmental tract
35 Dorsal longitudinal fasciculus (Schütz)

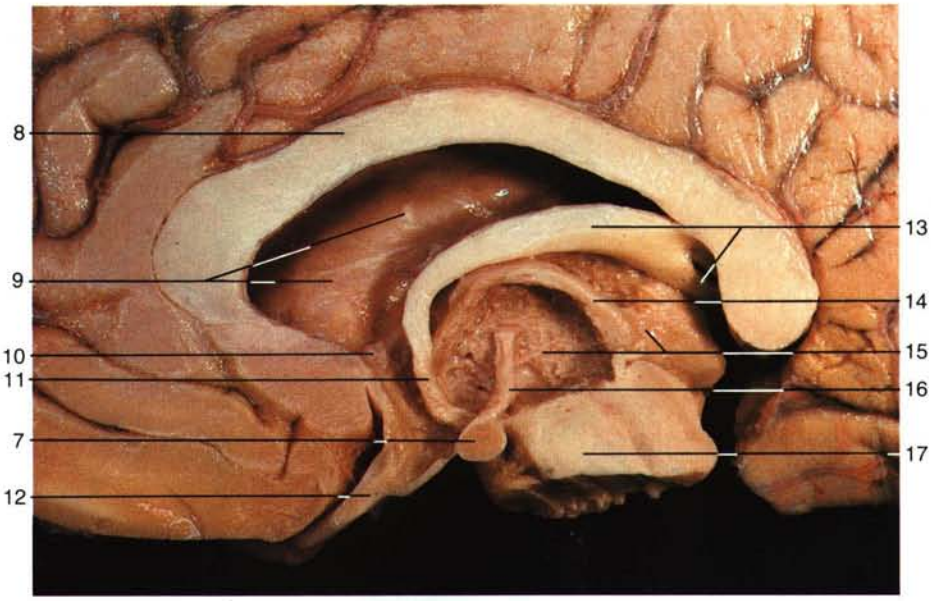

Median section through the diencephalon. Medial part of the thalamus and septum pellucidum have been removed to show the fornix and mamillothalamic tract.

1 Paraventricular nucleus
2 Preoptic nucleus
3 Ventromedial nucleus
4 Supraoptic nucleus
5 Posterior nucleus
6 Dorsomedial nucleus
Hypothalamic nuclei
7 **Mamillary body**
8 **Corpus callosum**
9 Lateral ventricle (showing caudate nucleus)
10 Anterior commissure
11 **Column of fornix**
12 Optic chiasma
13 **Crus of fornix**
14 Medullary stria of thalamus
15 Thalamus and interthalamic adhesion
16 **Mamillothalamic tract** of Vicq d'Azyr
17 Cerebral peduncle
18 **Pineal body**
19 Tectum of midbrain
20 Lamina terminalis

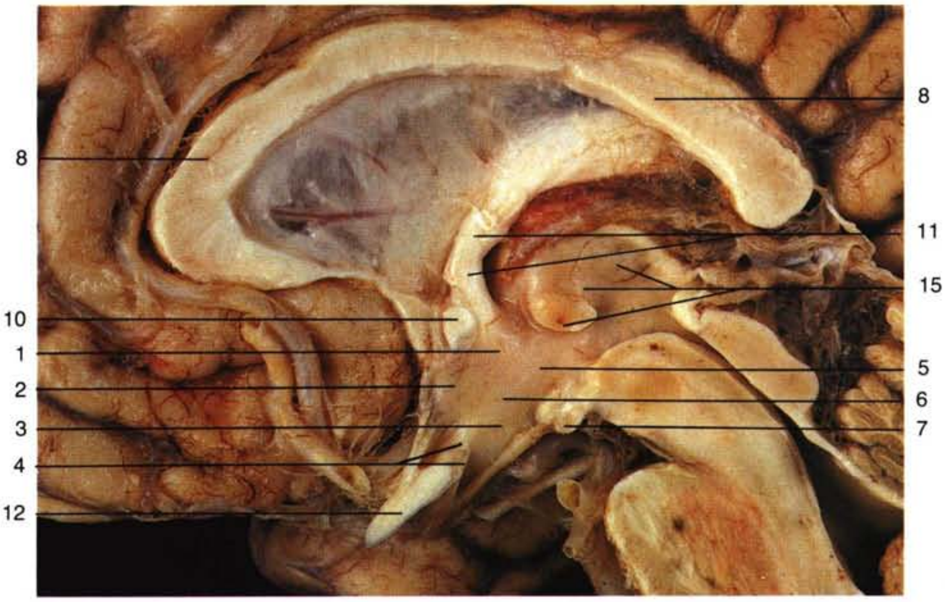

Median section through the diencephalon and midbrain; location of hypothalamic nuclei.

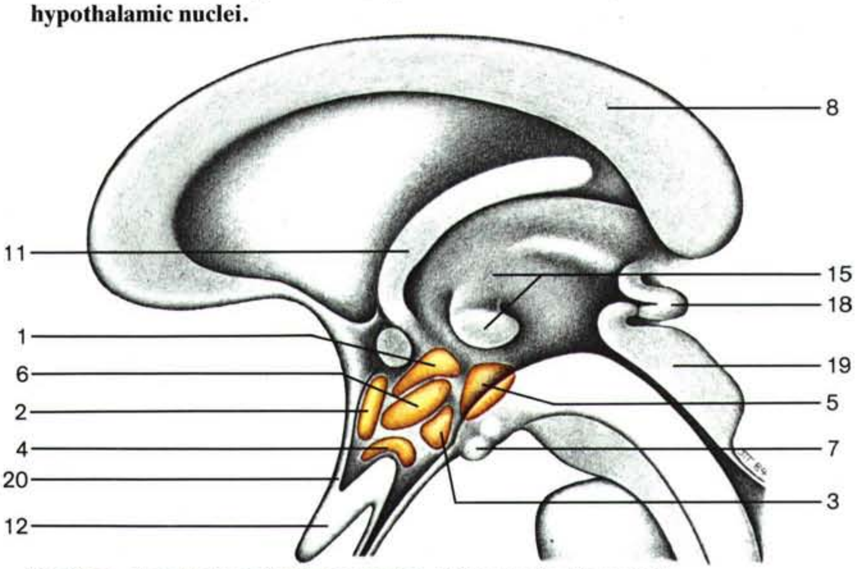

Position of main hypothalamic nuclei. (Schematic diagram.)

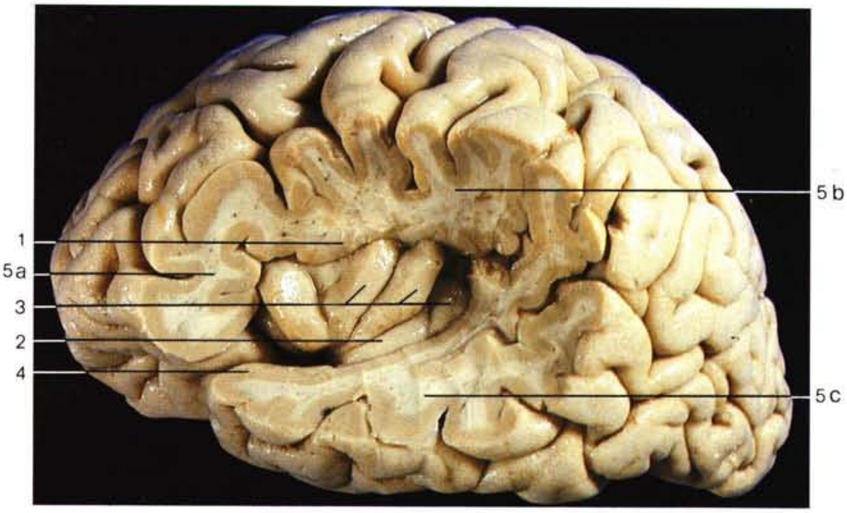

Insula (Reili). The opercula of the frontal, parietal and temporal lobes have been removed to display the insular gyri. Left hemisphere.

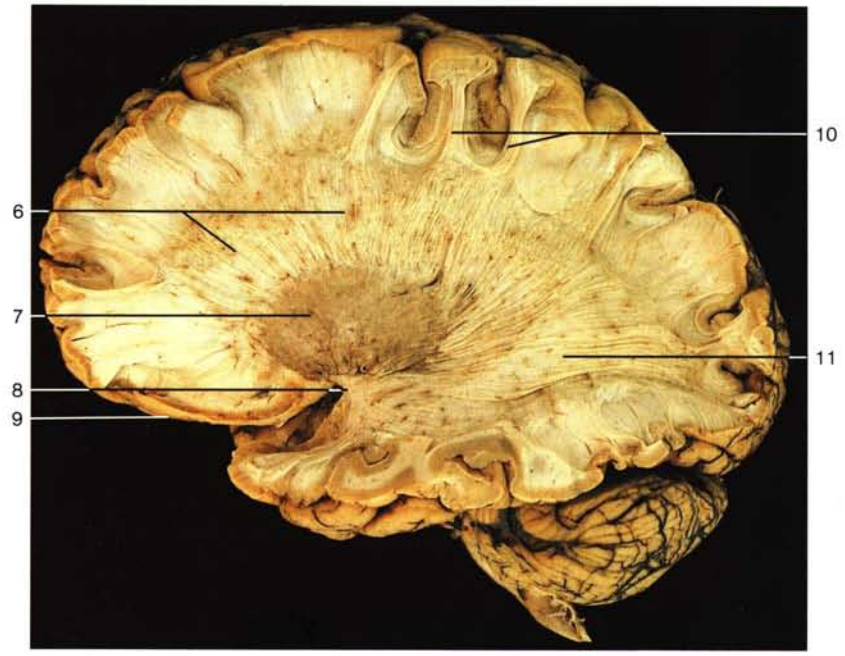

Dissection of the corona radiata, left hemisphere. Frontal pole on the left.

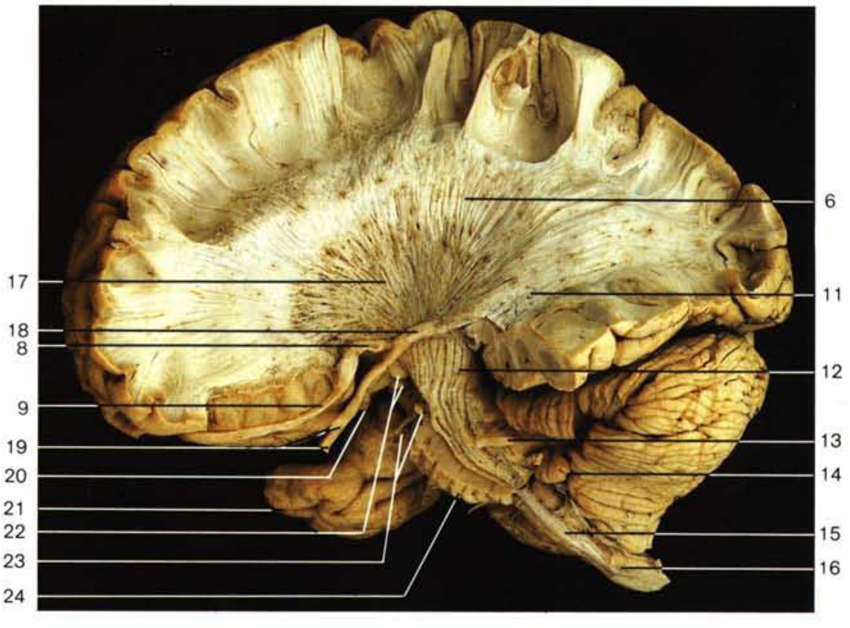

◁ **Corona radiata and internal capsule,** left hemisphere. Lentiform nucleus removed (frontal pole to the left).

1 Circular sulcus of insula
2 Long gyrus of insula
3 Short gyri of insula
4 Limen insulae
5 Opercula (cut)
  a Frontal operculum
  b Frontoparietal operculum
  c Temporal operculum
6 **Corona radiata**
7 **Lentiform nucleus**
8 **Anterior commissure**
9 Olfactory tract
10 Cerebral arcuate fibers
11 Optic radiation
12 Cerebral peduncle
13 Trigeminal nerve (n. V)
14 Flocculus of cerebellum
15 **Pyramidal tract**
16 Decussation of pyramidal tract
17 Internal capsule
18 Optic tract
19 Optic nerve (n. II)
20 Infundibulum
21 Temporal lobe (right side)
22 Mamillary bodies
23 Oculomotor nerve (n. III)
24 Transverse fibers of pons

1 Corona radiata
2 Anterior horn of lateral ventricle
3 Head of **caudate nucleus**
4 **Putamen**
5 Anterior commissure
6 Olfactory tract
7 **Amygdala**
8 Hippocampal digitations
9 **Internal capsule**
10 Calcar avis
11 Posterior horn of lateral ventricle
12 Choroid plexus of lateral ventricle
13 Caudal extremity of caudate nucleus
14 **Thalamus**
15 Cerebral arcuate fibers
16 Globus pallidus (remnants)

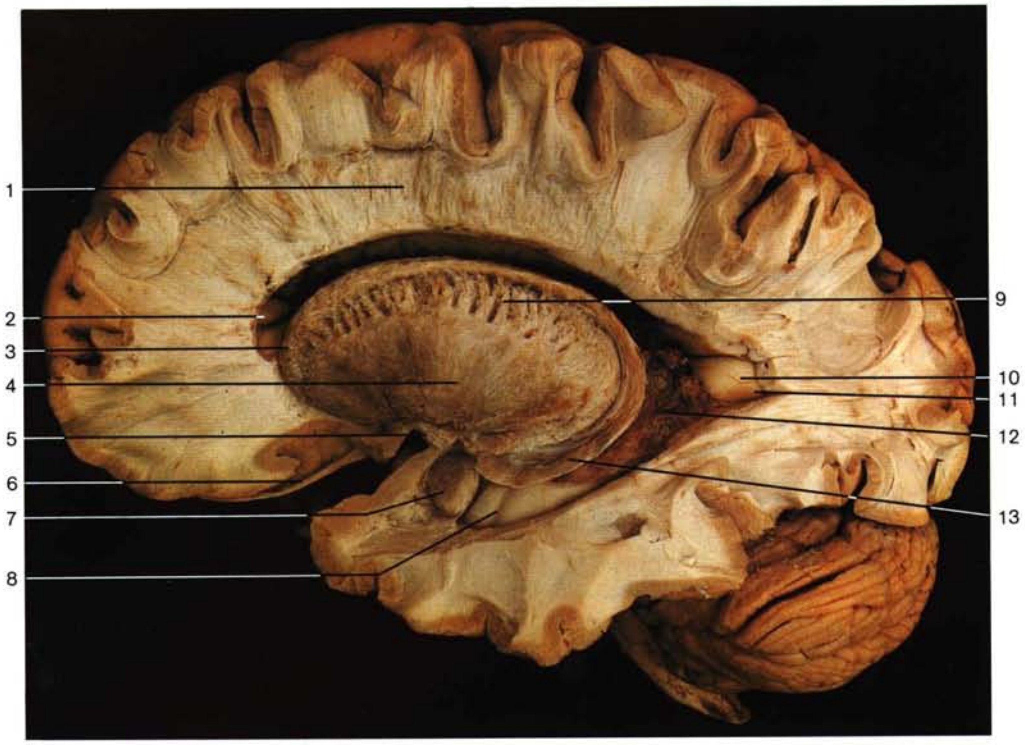

Dissection of the subcortical nuclei and internal capsule, left hemisphere (lateral aspect). Frontal pole to the left. The lateral ventricle has been opened, and the insular gyri and claustrum have been removed, revealing the lentiform nucleus and the internal capsule.

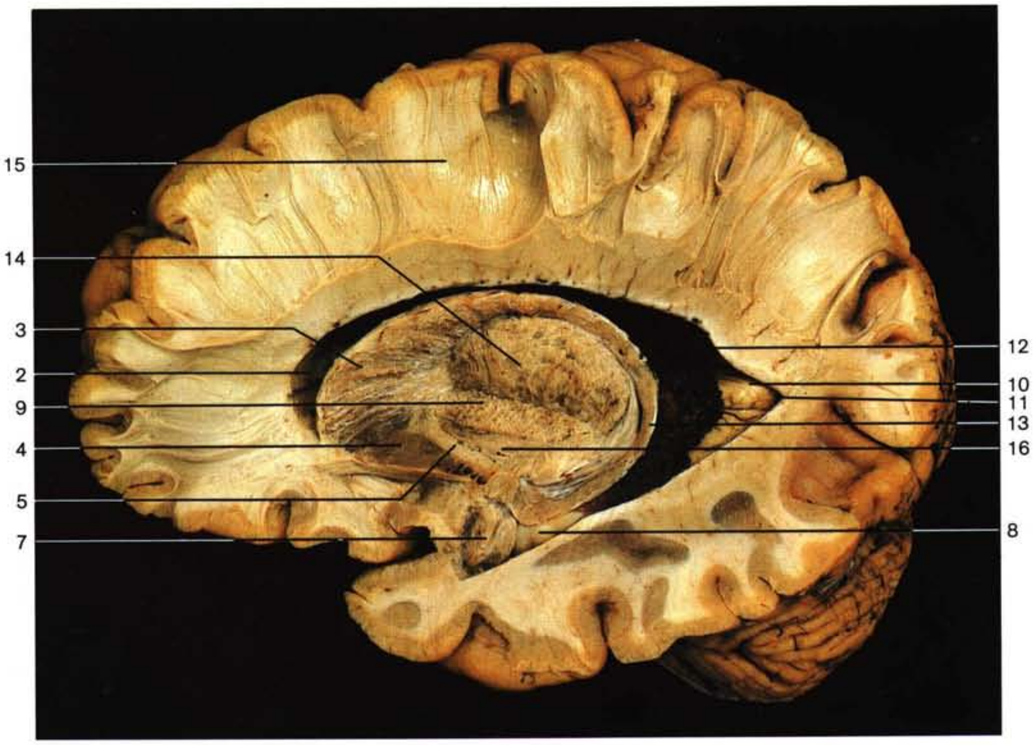

Dissection of the subcortical nuclei (lateral aspect). Lentiform nucleus removed, frontal pole to the left.

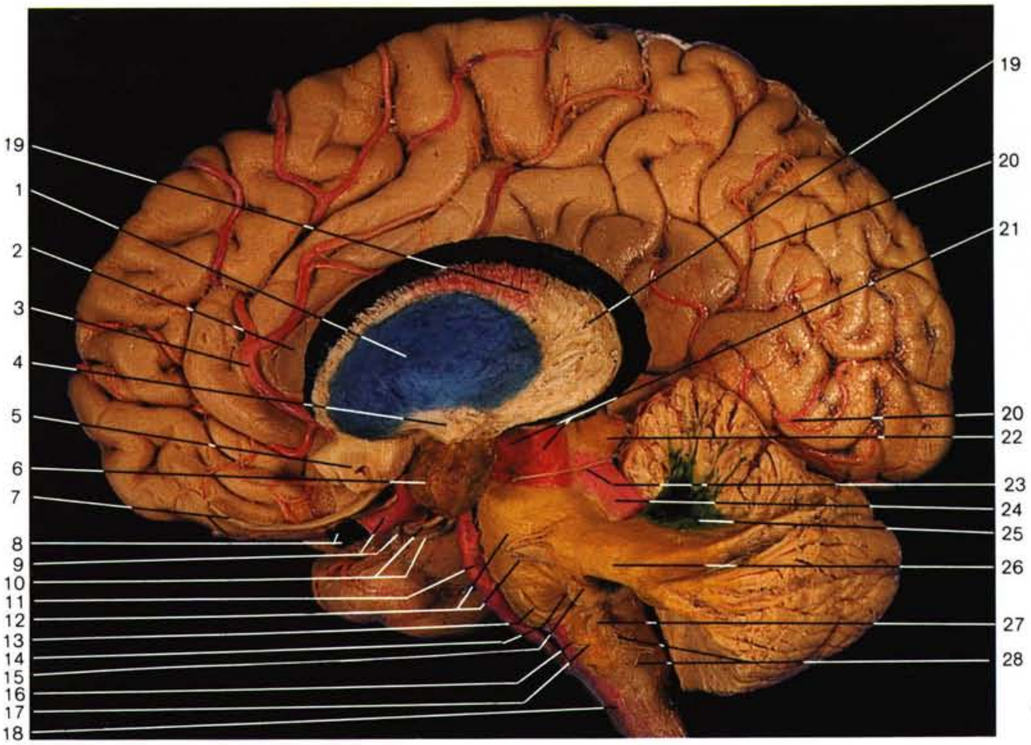

Brain stem and the connections with the cerebellum. Internal capsule (lateral aspect). Red = pyramidal tract; yellow = middle cerebellar peduncle; green = inferior cerebellar peduncle; pink = superior cerebellar peduncle.

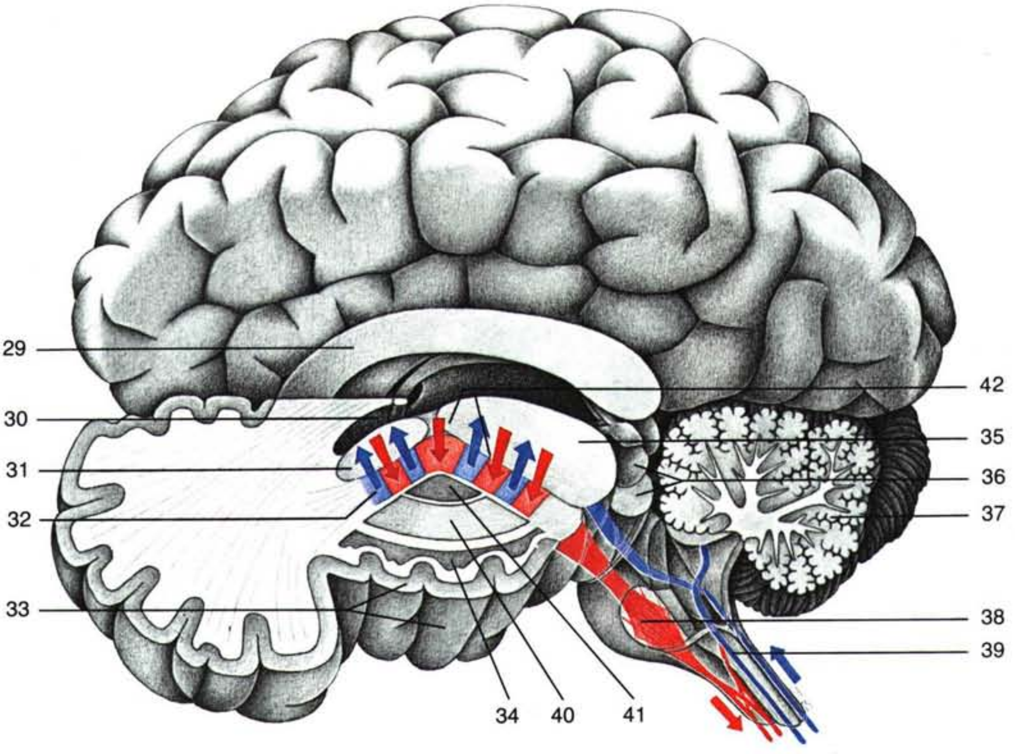

Internal capsule and subcortical nuclei. Left brain, frontal pole to the left, horizontal section. (Semischematic drawing.) Blue = afferent tracts; red = efferent tracts.

1	**Putamen**
2	Genu of corpus callosum
3	Anterior cerebral artery
4	Anterior commissure
5	Subcallosal area
6	Amygdala
7	Olfactory tract
8	Optic nerve (n. II)
9	Internal carotid artery and infundibulum
10	Oculomotor nerve (right and left nerve, n. III)
11	**Basilar artery**
12	Pons and trigeminal nerve (n. V)
13	Abducens nerve (n. VI)
14	Facial nerve (n. VII)
15	Vestibulocochlear nerve (n. VIII)
16	Hypoglossal nerve (n. XII)
17	Olive
18	Pyramidal tract
19	**Internal capsule**
20	Posterior cerebral artery
21	Cerebral peduncle
22	**Colliculi of midbrain**
23	Trochlear nerve (n. IV)
24	Superior cerebellar peduncle
25	Inferior cerebellar peduncle
26	**Middle cerebellar peduncle**
27	Glossopharyngeal nerve (n. IX)
28	Vagus (n. X) and accessory nerves (n. XI)
29	**Corpus callosum**
30	Lateral ventricle (anterior horn)
31	Caudate nucleus
32	Internal capsule (anterior limb)
33	Insula
34	**Claustrum**
35	**Thalamus**
36	Superior and inferior colliculus of midbrain
37	Cerebellum
38	Middle peduncle of cerebellum (efferent tracts)
39	Medulla oblongata (afferent tracts)
40	Putamen of lentiform nucleus
41	**Globus pallidus** of lentiform nucleus
42	Genu and posterior limb of internal capsule

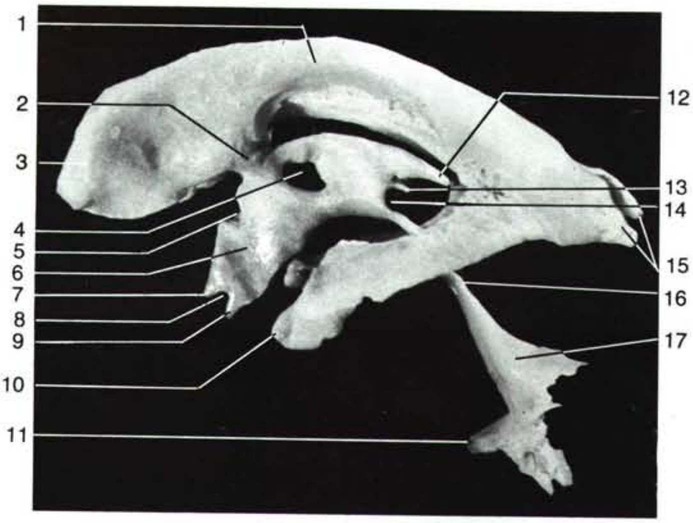

Cast of ventricular cavities of the brain (lateral aspect), frontal pole to the left.

1 Central part of the **lateral ventricle**
2 **Interventricular foramen** of Monro
3 Anterior horn of the lateral ventricle
4 Site of interthalamic adhesion
5 Notch for anterior commissure
6 **Third ventricle**
7 Optic recess
8 Notch for optic chiasma
9 Infundibular recess
10 Inferior horn of lateral ventricle with indentation of amygdaloid body
11 Lateral recess and lateral aperture of Luschka
12 Suprapineal recess
13 Pineal recess
14 Notch for posterior commissure
15 Posterior horn of lateral ventricle
16 **Cerebral aqueduct**
17 **Fourth ventricle**
18 Median aperture of Magendie
19 **Cerebellomedullary cistern**
20 Superior sagittal sinus
21 Inferior sagittal sinus
22 Intervaginal space of optic nerve
23 Arachnoid granulations of Pacchioni
24 Straight sinus

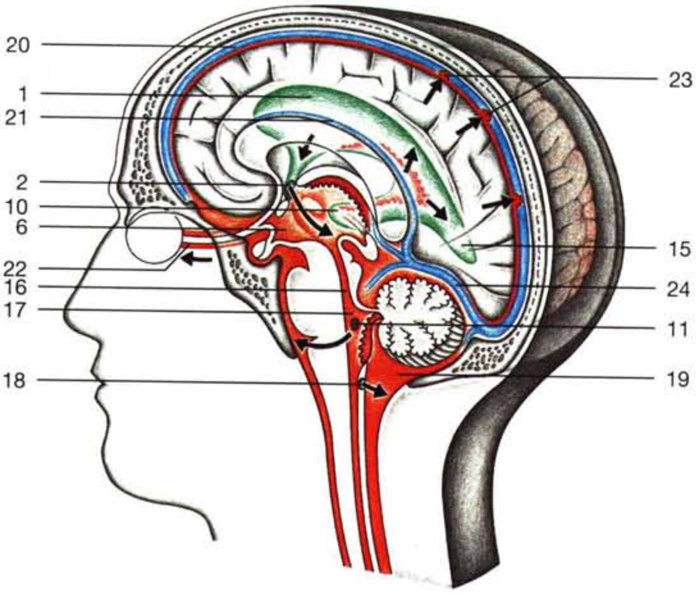

◁ **Position of ventricular cavities.** (Schematic drawing.) The direction of flow of cerebrospinal fluid is indicated by arrows.

Green = Right lateral ventricle
Red = Choroidal plexus

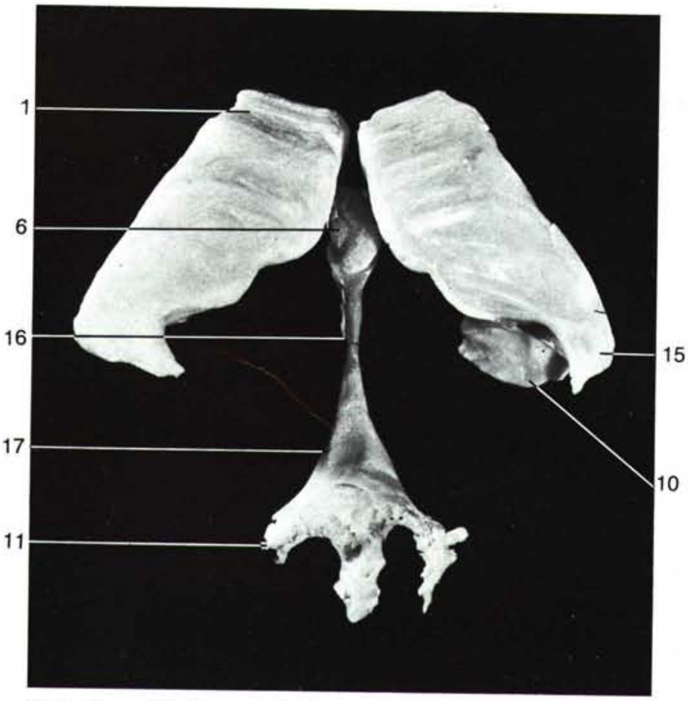

Cast of ventricular cavities (posterior aspect).

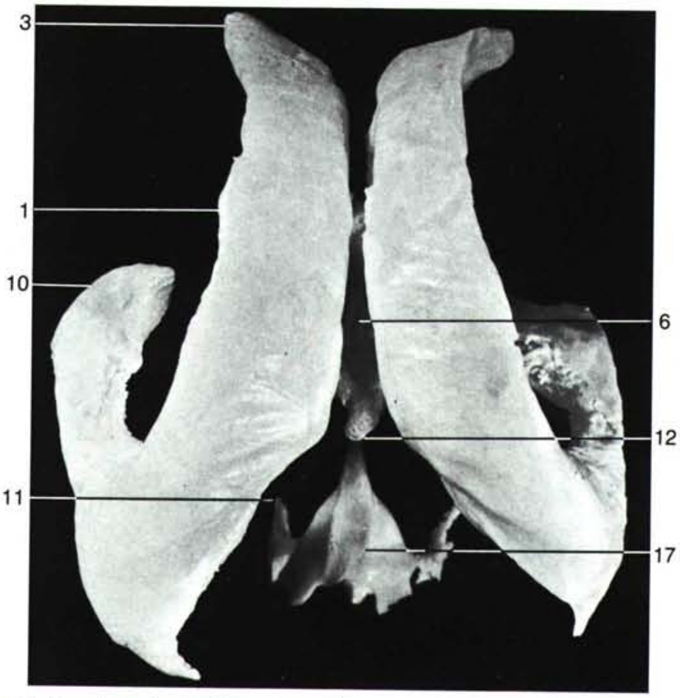

Cast of ventricular cavities (superior aspect). Frontal pole at top.

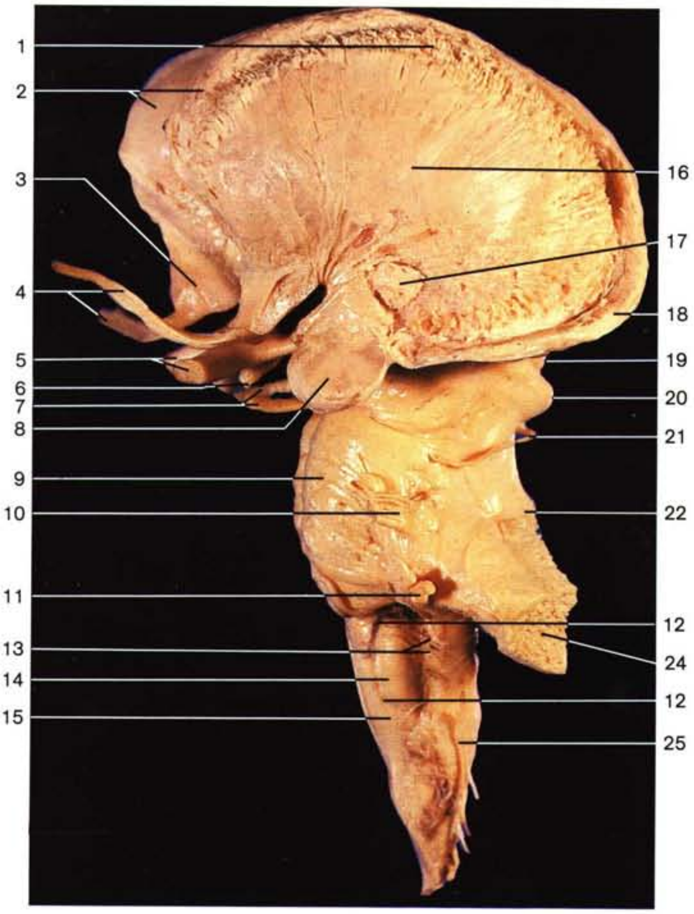

Brain stem (left lateral aspect). Cerebellar peduncles have been severed, cerebellum and cerebral cortex have been removed.

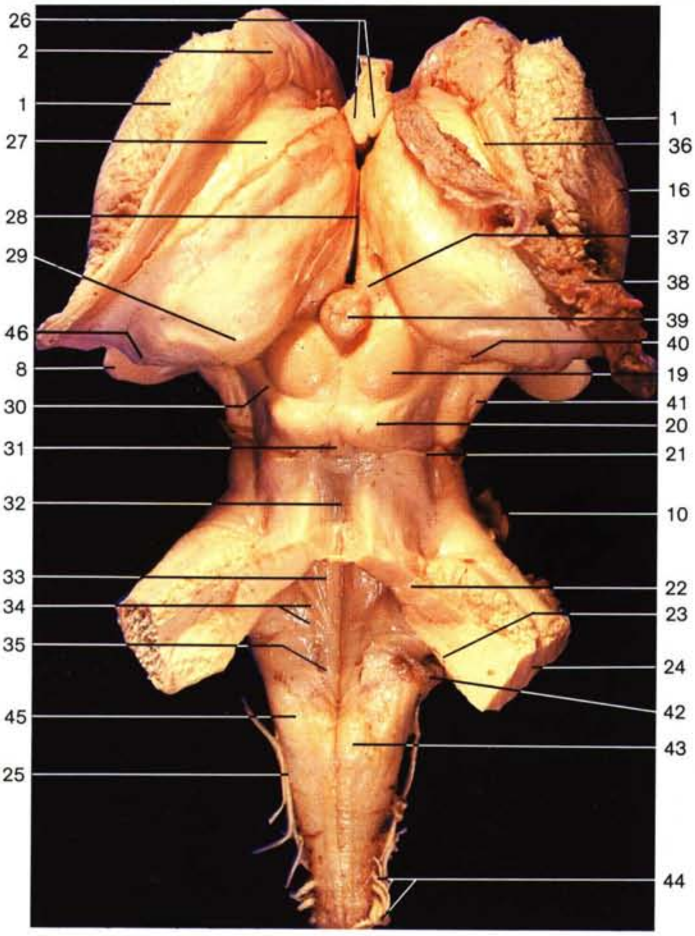

Brain stem (dorsal aspect). Cerebellum removed.

1 **Internal capsule**
2 Head of the caudate nucleus
3 Olfactory trigone
4 Olfactory tracts
5 Optic nerves
6 Infundibulum
7 Oculomotor nerve
8 **Amygdaloid body**
9 **Pons**
10 Trigeminal nerve
11 Facial and vestibulocochlear nerves
12 Hypoglossal nerve
13 Glossopharyngeal and vagus nerves
14 **Olive**
15 Medulla oblongata
16 **Lentiform nucleus**
17 **Anterior commissure**
18 Tail of caudate nucleus
19 Superior colliculus
20 Inferior colliculus
21 Trochlear nerve
22 Superior cerebellar peduncle
23 Inferior cerebellar peduncle
24 **Middle cerebellar peduncle**
25 Accessory nerve (n. XI)
26 Columns of fornix (divided)
27 Lamina affixa
28 **Third ventricle**
29 Pulvinar of thalamus
30 Inferior brachium
31 Frenulum veli
32 Superior medullary velum
33 Facial colliculus
34 Striae medullares and rhomboid fossa
35 Hypoglossal triangle
36 Stria terminalis and thalamostriate vein
37 Habenular trigone
38 Choroid plexus of lateral ventricle
39 **Pineal body**
40 Medial geniculate body
41 Cerebral peduncle
42 Choroid plexus of fourth ventricle
43 Clava
44 Dorsal root of cervical nerve
45 Cuneate tubercle
46 Lateral geniculate body

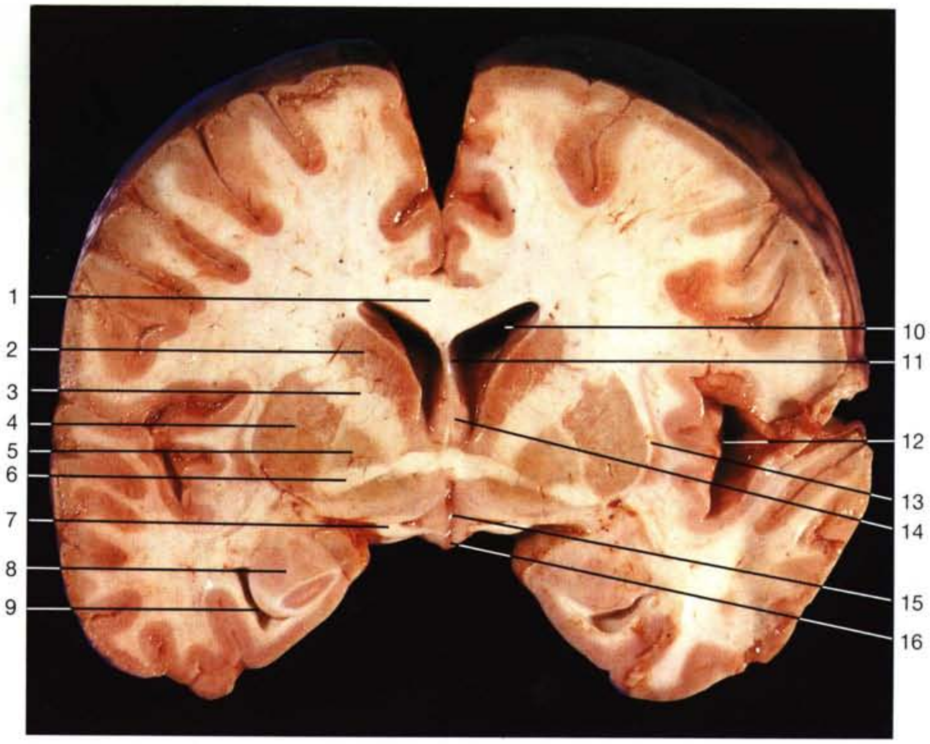

Coronal section through the brain at the level of the anterior commissure. Section 1.

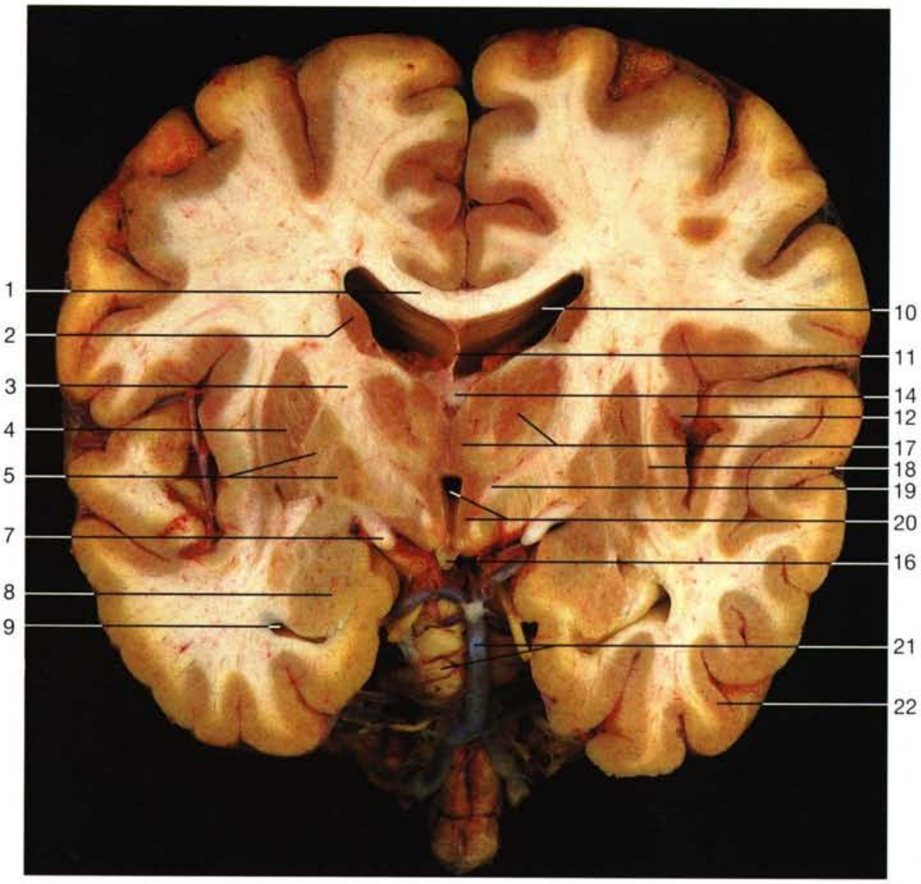

Coronal section through the brain at the level of the third ventricle and the interthalamic adhesion. Section 2.

1 Corpus callosum
2 Head of **caudate nucleus**
3 **Internal capsule**
4 **Putamen**
5 **Globus pallidus**
6 Anterior commissure
7 Optic tract
8 Amygdaloid body
9 Inferior horn of lateral ventricle
10 **Lateral ventricle**
11 Septum pellucidum
12 Lobus insularis (insula)
13 External capsule
14 Column of fornix
15 Optic recess
16 Infundibulum
17 **Thalamus**
18 Claustrum
19 Lenticular ansa
20 **Third ventricle** and **hypothalamus**
21 Basilar artery and pons
22 Cortex of temporal lobe
23 Inferior colliculus
24 Superior colliculus
25 Cerebral aqueduct
26 Red nucleus
27 Substantia nigra
28 Cerebral peduncle
29 Trochlear nerve (n. IV)
30 Gray matter
31 Nucleus of oculomotor nerve (nucleus of Edinger-Westphal)
32 Fibers of oculomotor nerve (n. III)
33 Vermis of cerebellum
34 Fourth ventricle
35 Reticular formation
36 Pons and transverse pontine fibers
37 Emboliform nucleus
38 Dentate nucleus
39 Middle cerebellar peduncle
40 Choroid plexus
41 Hypoglossal nucleus at rhomboid fossa
42 Medial longitudinal fasciculus
43 Trigeminal nerve (n. V.)
44 Inferior olivary nucleus
45 Corticospinal fibers and arcuate fibers
46 Fourth ventricle with choroid plexus
47 Vestibular nuclei
48 Nucleus and tractus solitarius
49 Inferior cerebellar peduncle (restiform body)
50 Reticular formation
51 Medial lemniscus
52 Cuneate nucleus of Burdach
53 Central canal
54 **Pyramidal tract**
55 Flocculus of cerebellum
56 Cerebellar hemisphere with pia mater
57 "Arbor vitae" of cerebellum
58 Nucleus gracilis of Goll
59 Lateral recess of choroid plexus of IVth ventricle
60 Posterior inferior cerebellar artery
61 Choroid plexus of lateral ventricle

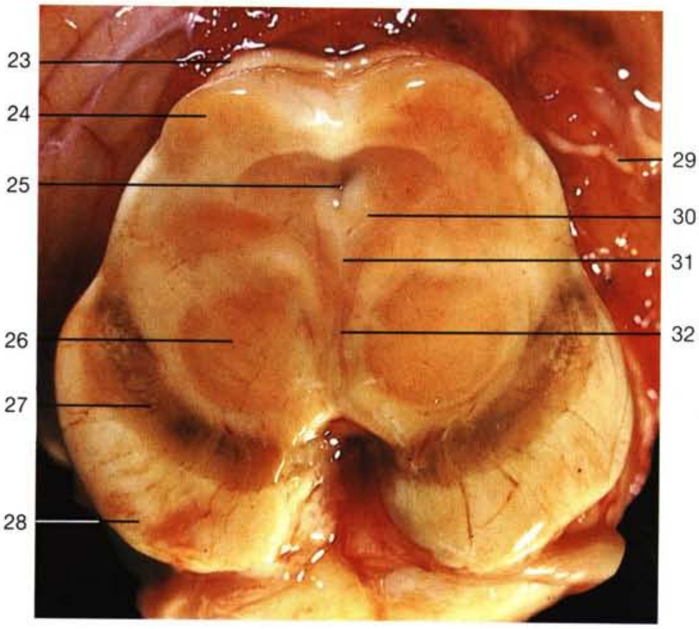
23
24
25
26
27
28
29
30
31
32

Cross-section of the midbrain (mesencephalon) at the level of the superior colliculus (superior aspect). Section 4.

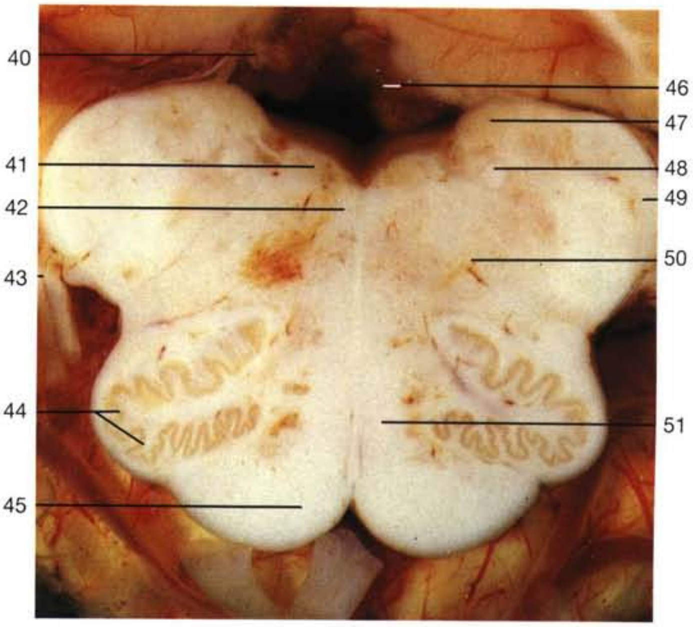
40
41
42
43
44
45
46
47
48
49
50
51

Cross-section of the rhombencephalon at the level of the olive (inferior aspect). Section 6.

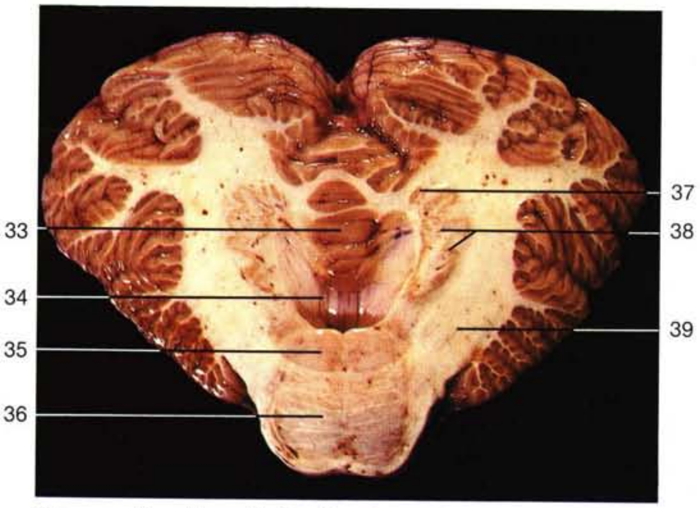
33
34
35
36
37
38
39

Cross-section through the rhombencephalon at the level of pons (inferior aspect). Section 5.

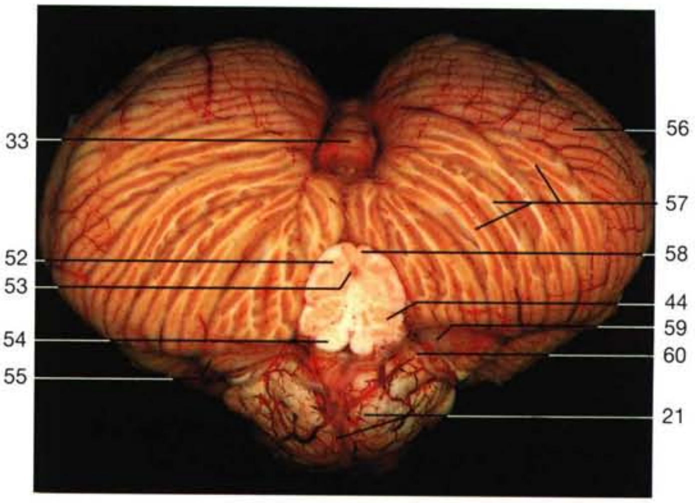
33
52
53
54
55
56
57
58
44
59
60
21

Cross-section through medulla oblongata and cerebellum (inferior aspect). Section 7.

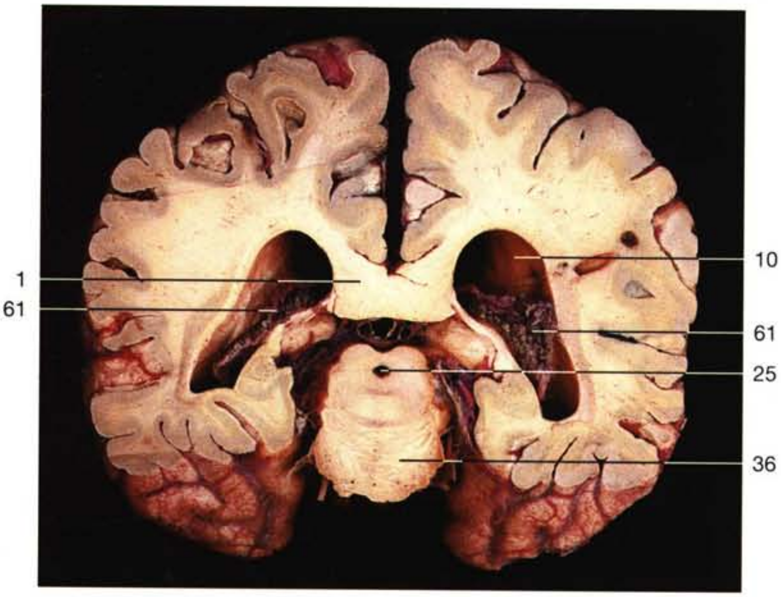
1
61
10
61
25
36

Coronal section at the level of **inferior colliculus** (posterior aspect). Section 3.

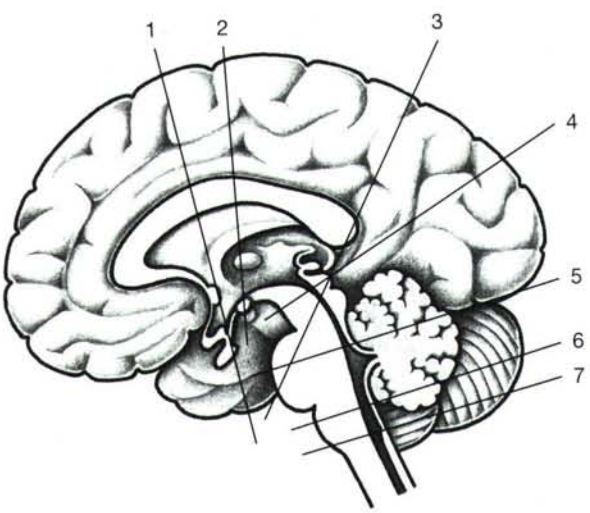

Right half of the brain. Levels of the sections are indicated.

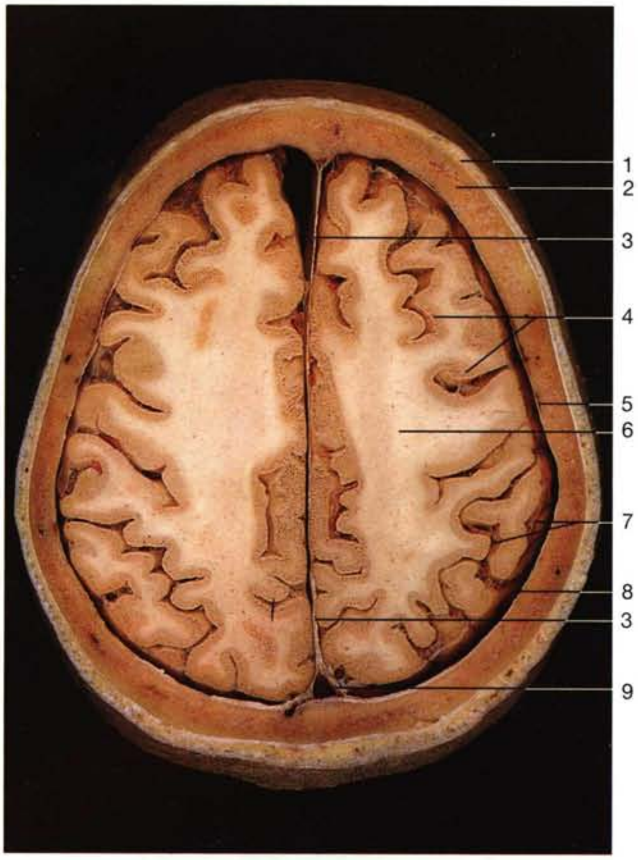

Horizontal section through the head.
Section 1.

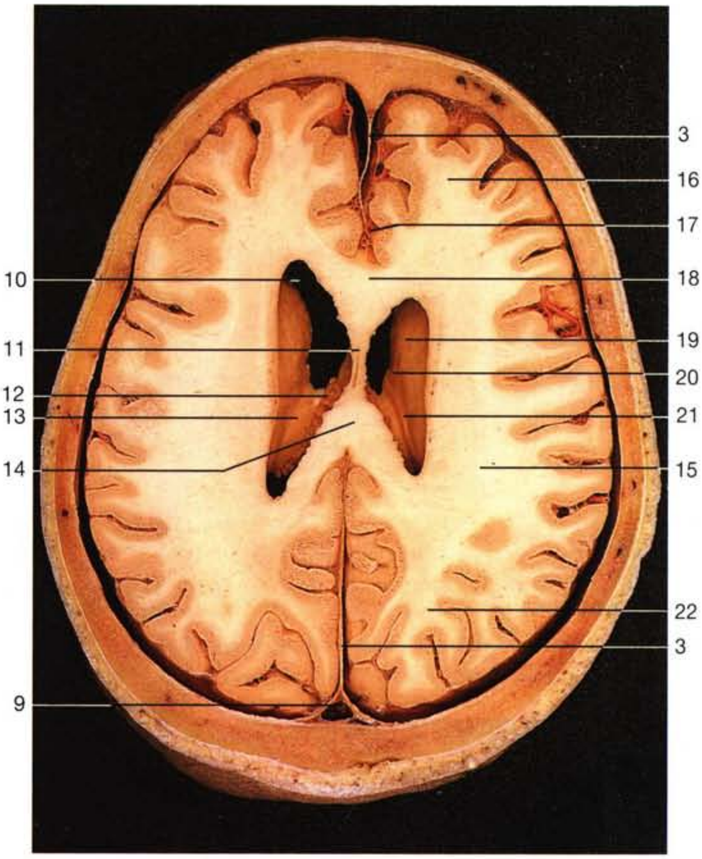

Horizontal section through the head.
Section 2.

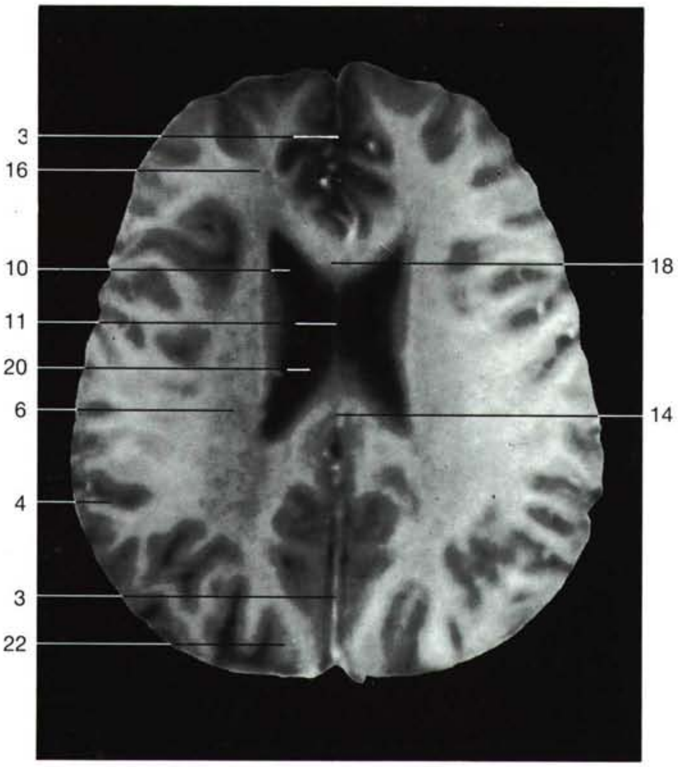

MR-Scan of the human head at the level of section 2.

1 Skin of scalp
2 Calvaria (diploe of the skull)
3 **Falx cerebri**
4 Gray matter of brain (cortex)
5 Dura mater
6 **White matter of brain**
7 Arachnoid and pia mater with vessels
8 Subdural space (slightly expanded due to shrinkage of the brain)
9 Superior sagittal sinus
10 Anterior horn of lateral ventricle
11 **Septum pellucidum**
12 Choroid plexus
13 Thalamus
14 Splenium of corpus callosum
15 Parietal lobe
16 Frontal lobe
17 Anterior cerebral artery
18 Genu of corpus callosum
19 Caudate nucleus
20 **Central part of lateral ventricle**
21 Stria terminalis
22 Occipital lobe

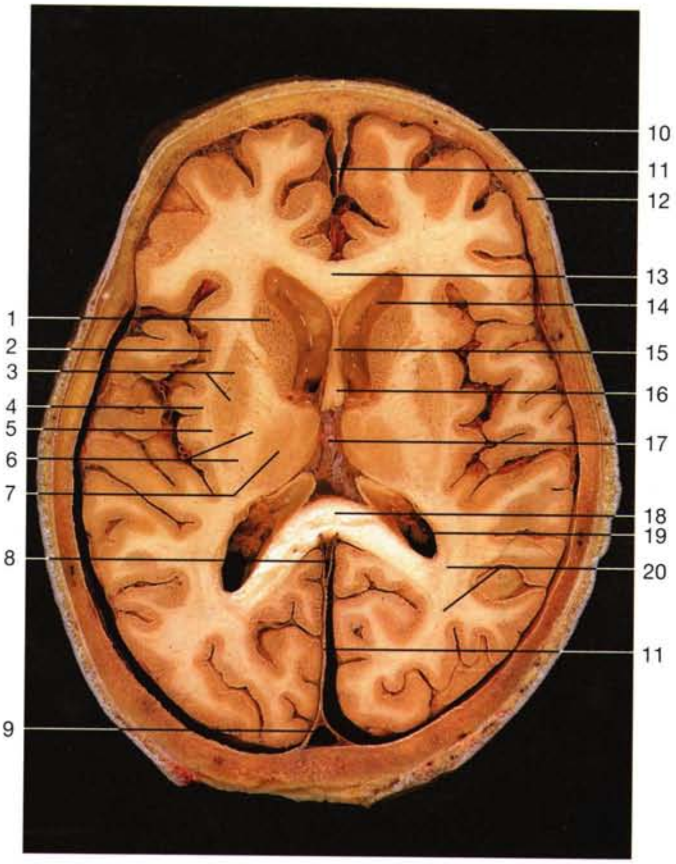

Horizontal section through the head at the level of third ventricle of internal capsule and neighbouring nuclei. Section 3.

1 **Caudate nucleus**
2 Lobus insularis (insula)
3 **Lentiform nucleus**
4 Claustrum
5 External capsule
6 Internal capsule
7 **Thalamus**
8 Straight sinus (sinus rectus)
9 Superior sagittal sinus
10 Skin of scalp
11 Falx cerebri
12 Calvaria (diploe of skull)
13 Genu of corpus callosum
14 Anterior horn of lateral ventricle
15 **Septum pellucidum**
16 Column of fornix
17 Choroid plexus of third ventricle
18 Splenium of corpus callosum
19 Entrance to inferior horn of lateral ventricle with choroid plexus
20 Optic radiation
21 **Third ventricle**

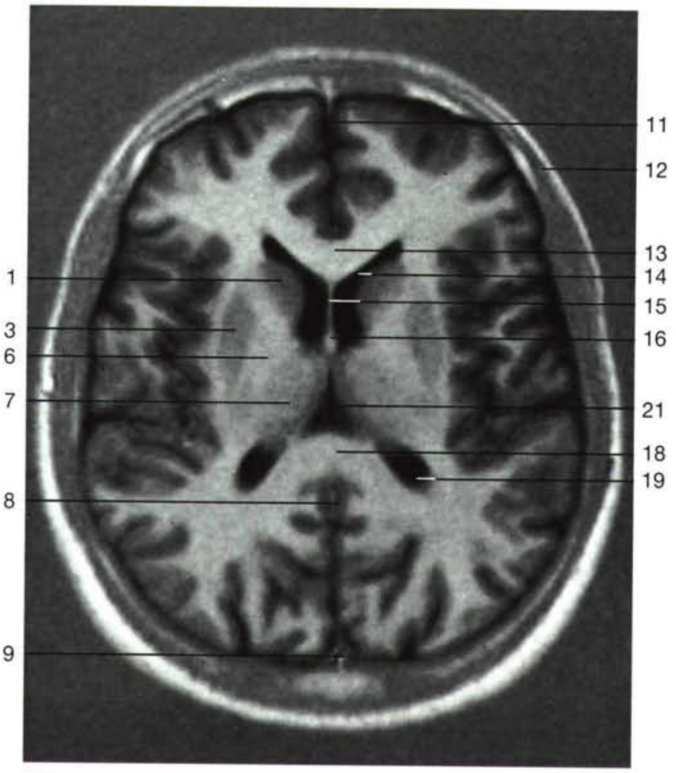

MR-Scan at the corresponding level to the above figure. Section 3.

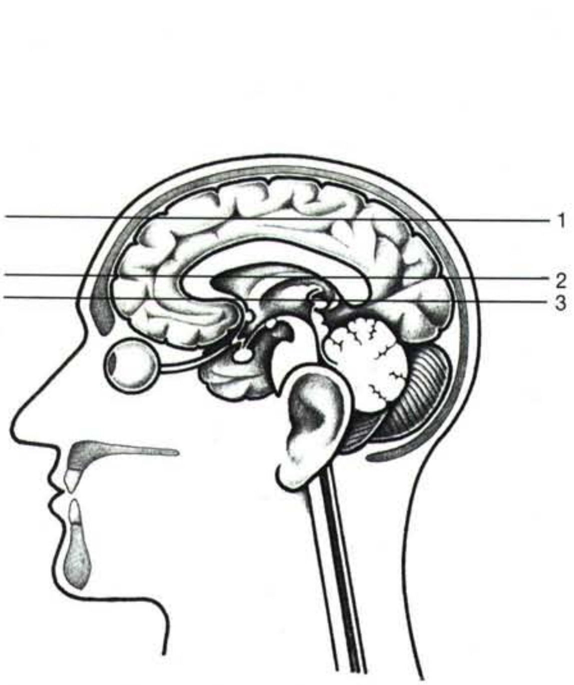

Sagittal section through the head. Levels of the horizontal sections are indicated.

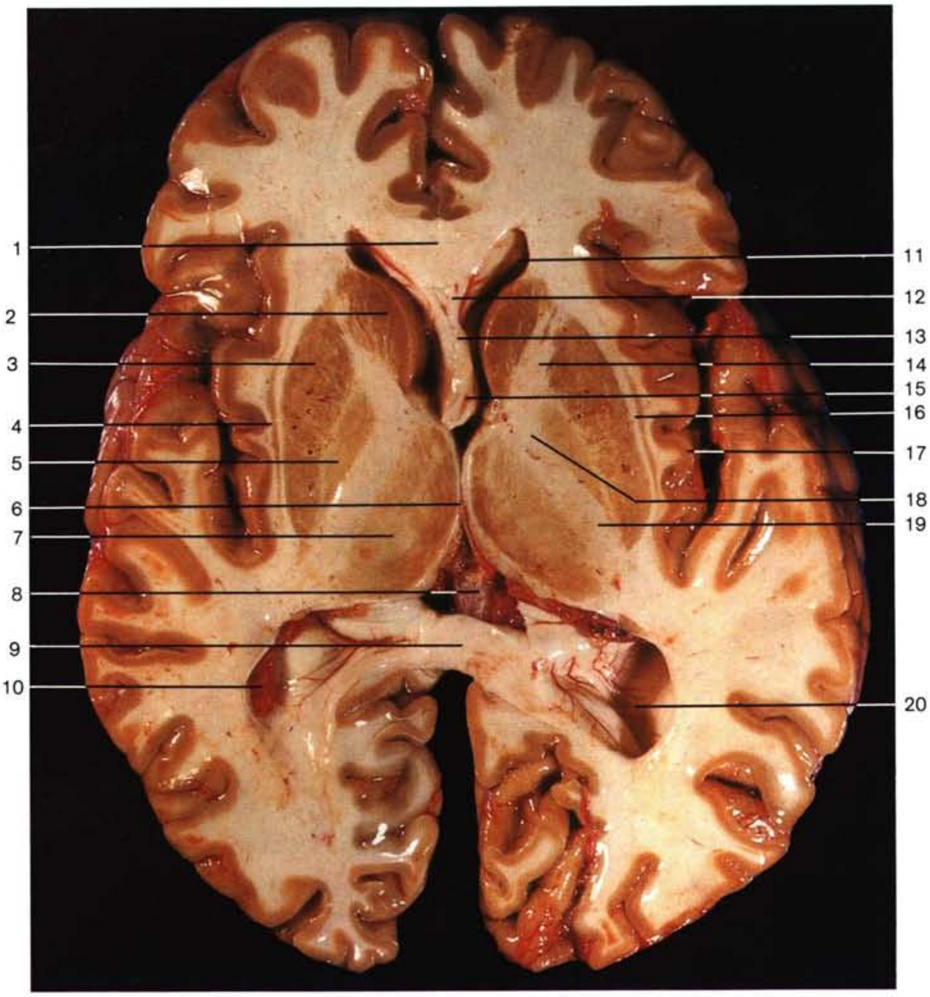

Horizontal section through the brain, showing the subcortical nuclei and internal capsule. Section 1.

1 Genu of corpus callosum
2 **Head of caudate nucleus**
3 **Putamen**
4 **Claustrum**
5 **Globus pallidus**
6 Third ventricle
7 Thalamus
8 **Pineal body**
9 Splenium of corpus callosum
10 Choroid plexus of the lateral ventricle
11 Anterior horn of lateral ventricle
12 Cavity of septum pellucidum
13 **Septum pellucidum**
14 Anterior limb of internal capsule
15 Column of fornix
16 External capsule
17 **Lobus insularis** (insula)
18 Genu of **internal capsule**
19 Posterior limb of internal capsule
20 Posterior horn of lateral ventricle
21 Anterior commissure
22 Optic radiation
23 Falx cerebri
24 **Maxillary sinus**
25 Position of auditory tube
26 Tympanic cavity
27 External acoustic meatus
28 **Medulla oblongata**
29 Fourth ventricle
30 Cerebellum (left hemisphere)
31 Temporomandibular joint
32 Tympanic membrane
33 Base of cochlea
34 Mastoid air cells
35 Sigmoid sinus
36 Vermis of cerebellum
37 Intermediate mass

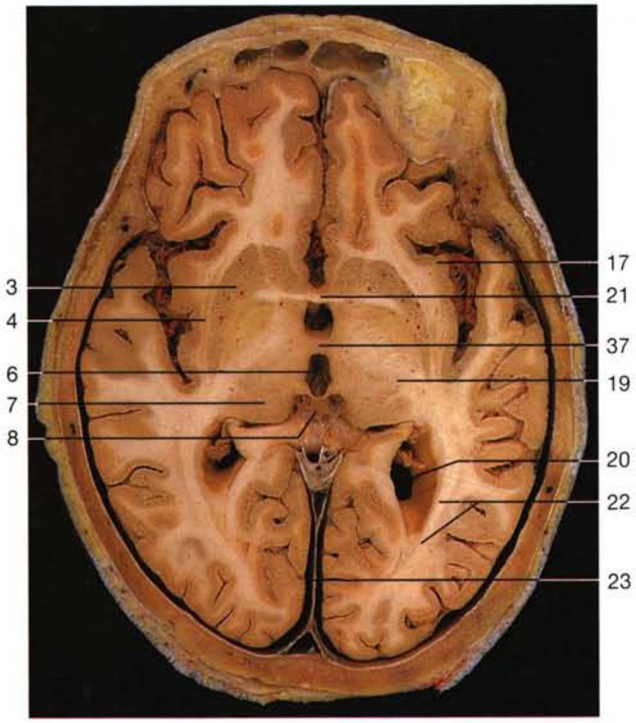

Horizontal section through the head. Section 2.

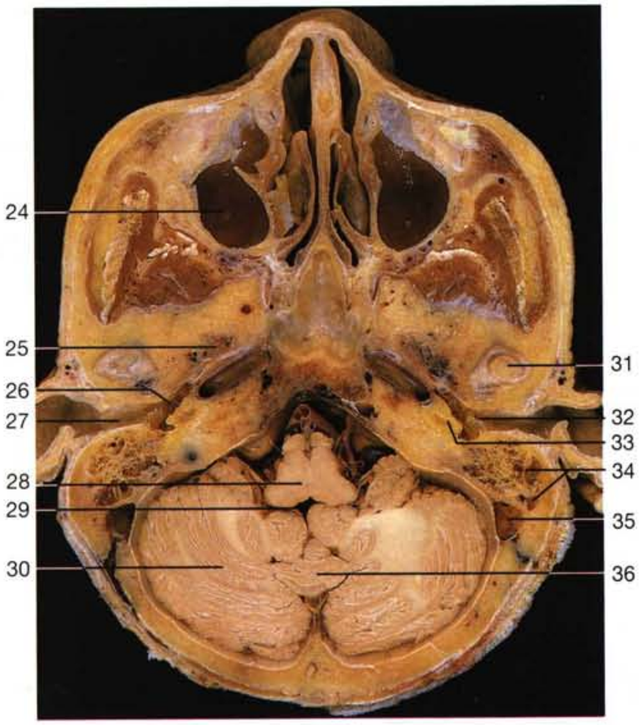

Horizontal section through the head. Section 4.

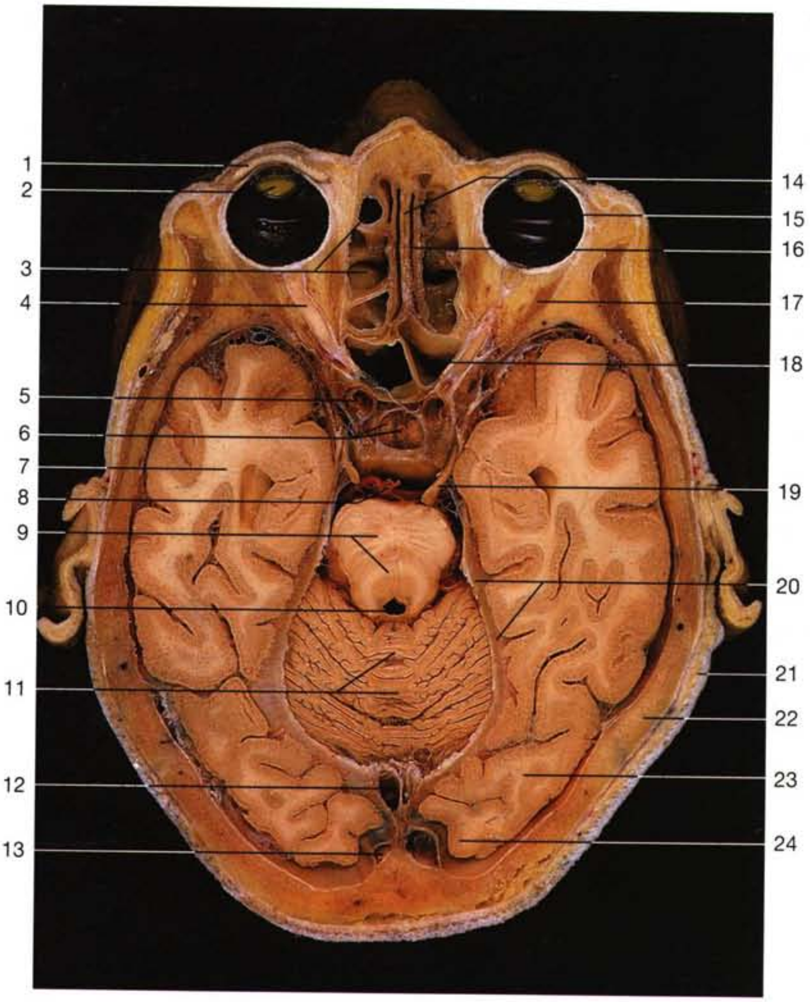

1 Upper lid (tarsal plate)
2 Lens
3 Ethmoidal sinus
4 **Optic nerve** (n. II)
5 Internal carotid artery
6 Infundibulum and **pituitary gland**
7 Temporal lobe
8 Basilar artery
9 Pons (cross section of brain stem)
10 **Cerebral aqueduct** (beginning of fourth ventricle)
11 Vermis of cerebellum
12 Straight sinus
13 Transverse sinus
14 Nasal septum
15 **Eyeball** (sclera)
16 Nasal cavity
17 Lateral rectus muscle
18 Sphenoidal sinus
19 Oculomotor nerve (n. III)
20 **Tentorium of cerebellum**
21 Skin of scalp
22 Calvaria
23 Occipital lobe
24 Striate cortex (visual cortex)

Horizontal section through the head. Section 3.

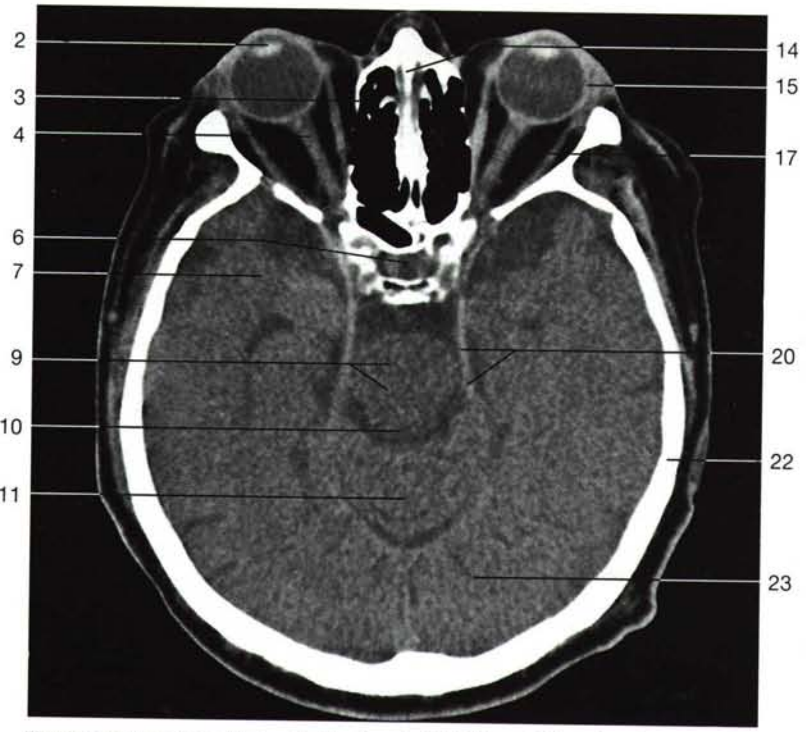

Horizontal section through the head. (CT-Scan.) Section 3.

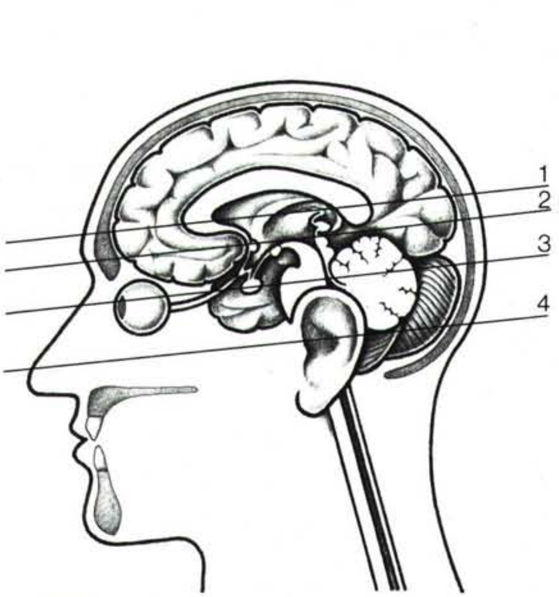

Sagittal section through the head.
Levels of the horizontal sections are indicated.

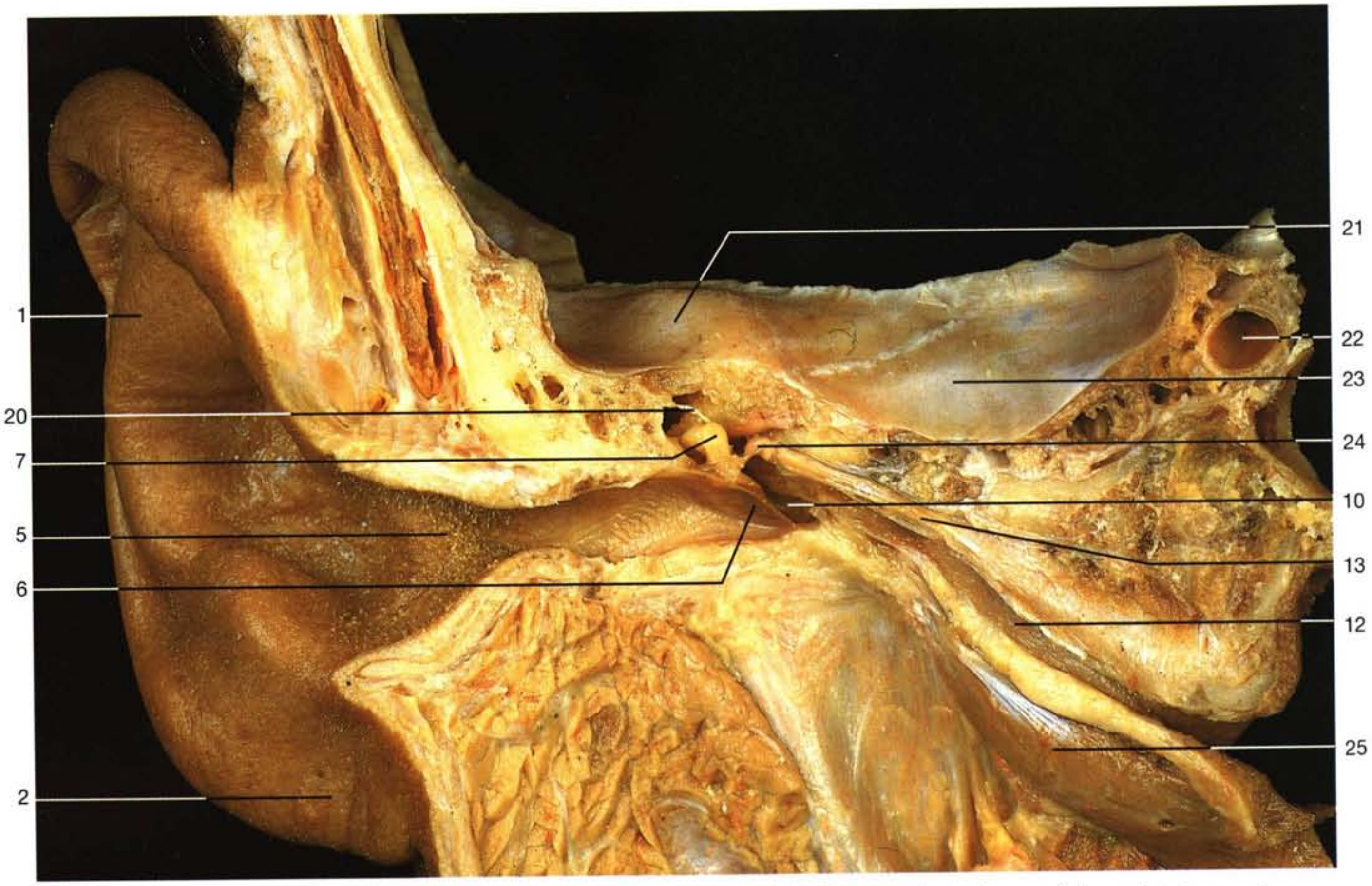

Longitudinal section through the right temporal bone I. The outer and middle ear and auditory ossicles and tube are shown (anterior aspect).

Outer ear
1 Auricle
2 Lobule of auricle
3 Helix
4 Tragus
5 External acoustic meatus

Middle ear
6 Tympanic membrane
7 Malleus
8 Incus
9 Stapes
10 Tympanic cavity
11 Mastoid process
12 Auditory tube
13 Tensor tympani muscle

Inner ear
14 Anterior semicircular duct
15 Posterior semicircular duct
16 Lateral semicircular duct
17 Cochlea
18 Vestibulocochlear nerve
19 Petrous part of the temporal bone

Additional structures
20 Superior ligament of malleus
21 Arcuate eminence
22 Internal carotid artery
23 Anterior surface of pyramid with dura mater
24 Stapes
25 Levator veli palatini muscle

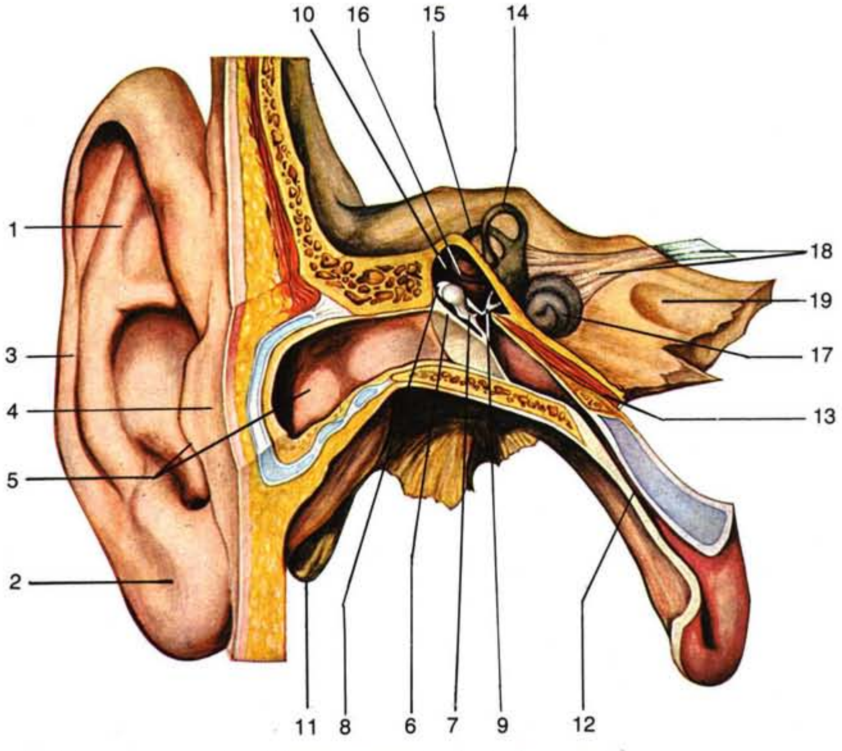

Right **auditory and vestibular apparatus** (anterior aspect). (Schematic drawing.)

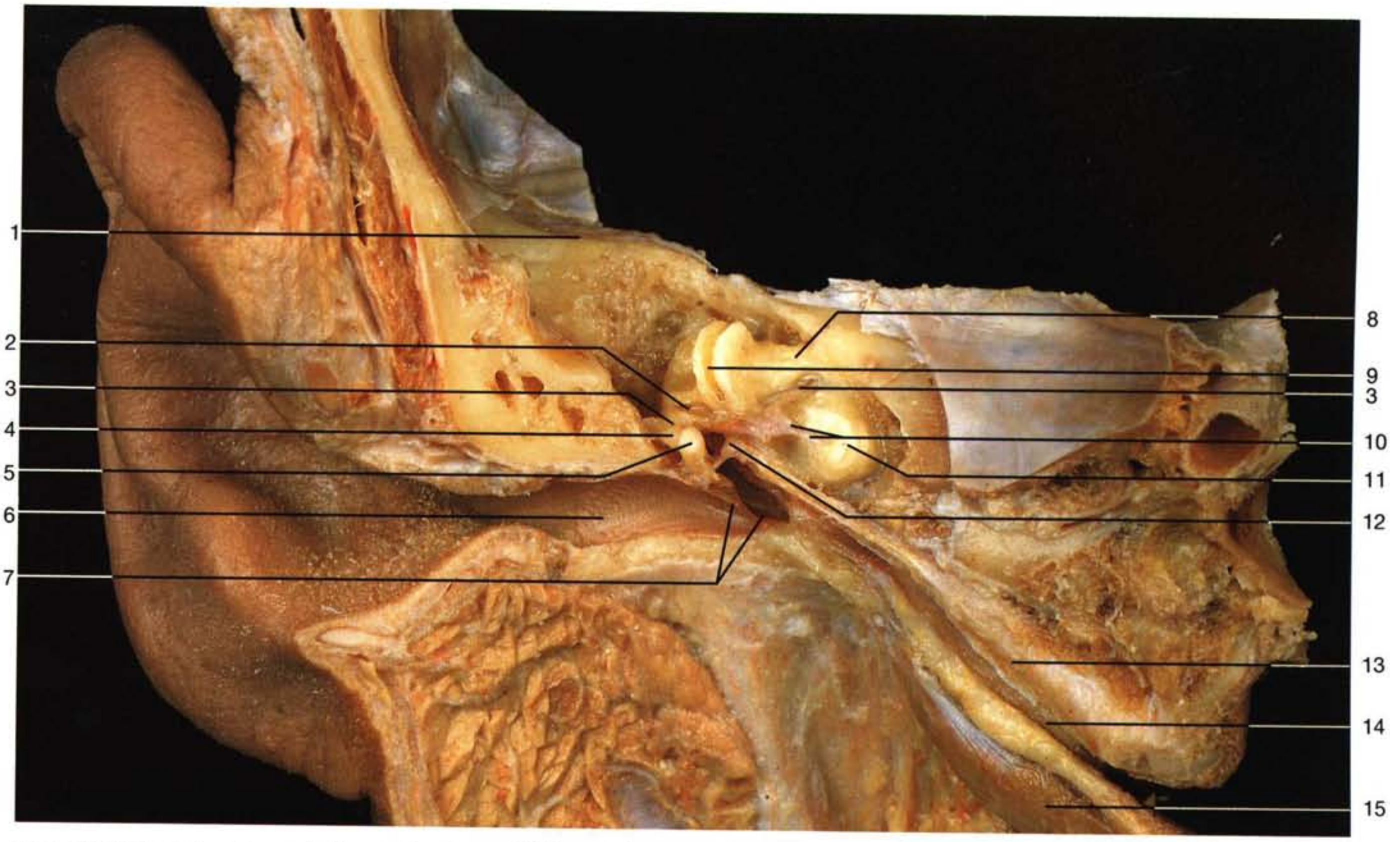

Longitudinal section through the right outer, middle and inner ear II. The cochlea and semicircular canals have been further dissected (anterior aspect).

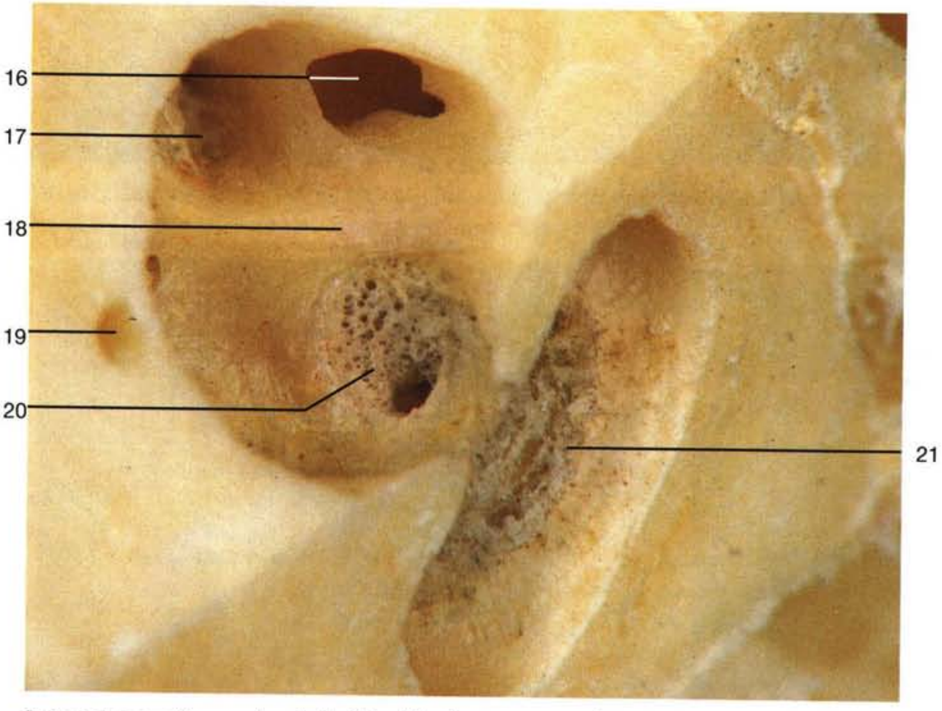

Internal acoustic meatus, left side. The bone was partly removed to show the bottom of the meatus.

1 Roof of tympanic cavity
2 **Lateral osseous semicircular canal**
3 **Facial nerve**
4 Incus
5 **Malleus**
6 External acoustic meatus
7 **Tympanic cavity** and **tympanic membrane**
8 Vestibulochochlear nerve
9 Anterior osseous semicircular canal
10 Geniculate ganglion and greater petrosal nerve
11 **Cochlea**
12 **Stapes**
13 Tensor tympani muscle
14 **Auditory tube**
15 Levator veli palatini muscle
16 Area of facial nerve
17 Superior vestibular area
18 Transverse crest
19 Foramen singulare
20 Foraminous spiral tract (outlet of cochlear part of vestibulocochlear nerve)
21 Base of cochlea

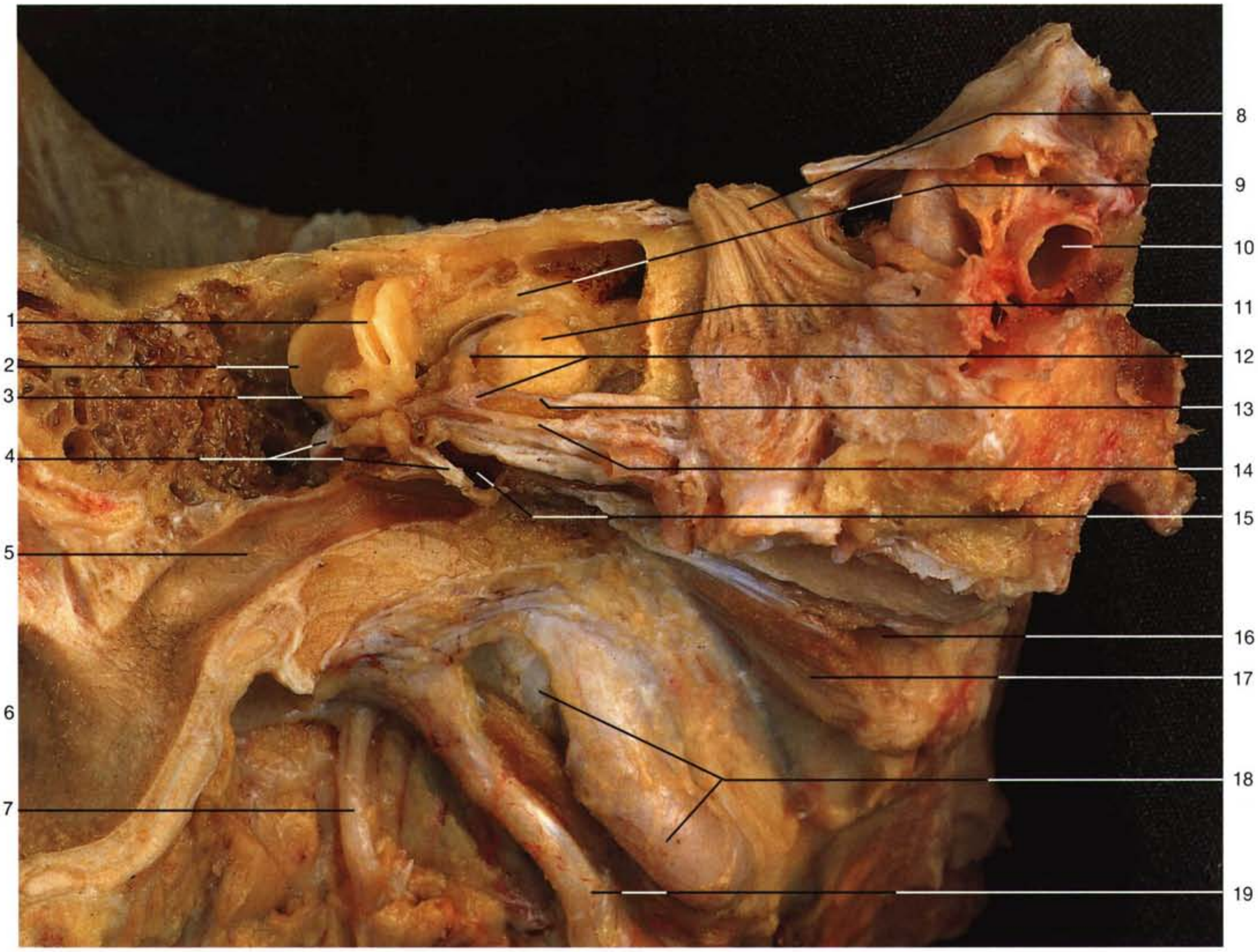

Longitudinal section through the outer, middle and inner ear III. Deeper dissection to display facial nerve and lesser and greater petrosal nerves (anterior aspect).

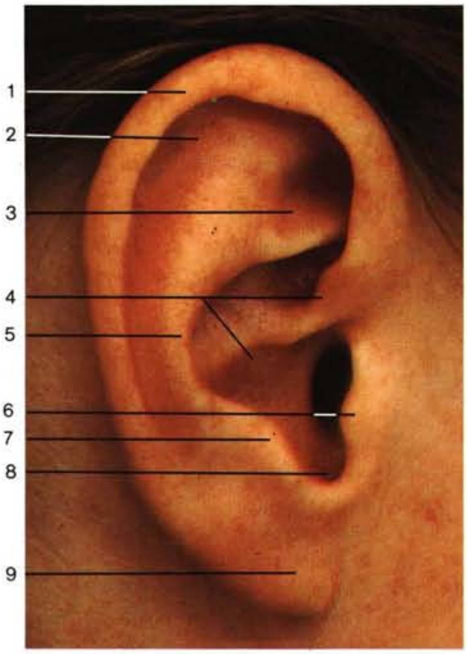

Right auricle (lateral aspect).

1 Helix
2 Scaphoid fossa
3 Triangular fossa
4 Concha
5 Antihelix
6 Tragus
7 Antitragus
8 Intertragic notch
9 Lobule

1 **Anterior osseous semicircular canal** (opened)
2 **Posterior osseous semicircular canal**
3 **Lateral osseous semicircular canal** (opened)
4 Facial nerve and chorda tympani
5 External acoustic meatus
6 Auricle
7 **Facial nerve**
8 Trigeminal nerve
9 Bony base of internal acoustic meatus
10 Internal carotid artery within cavernous sinus
11 **Cochlea**
12 **Facial nerve** with **geniculate ganglion**
13 Greater petrosal nerve
14 Lesser petrosal nerve
15 **Tympanic cavity**
16 Auditory tube
17 Levator veli palatini muscle
18 Internal carotid artery, internal jugular vein
19 Styloid process

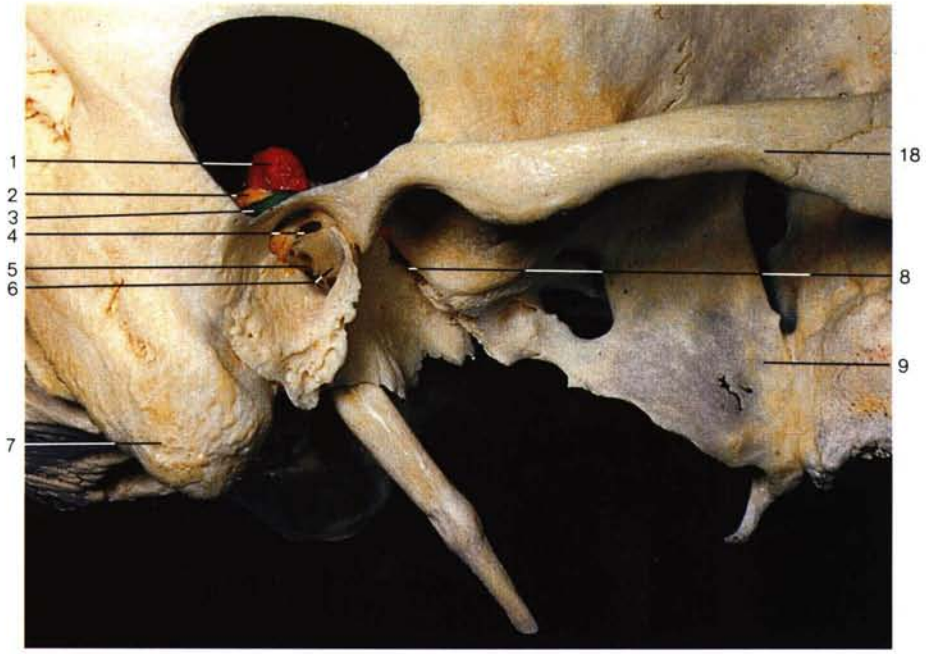

Right temporal bone (lateral aspect). Petrosquamous portion has been partly removed to display the semicircular canals.

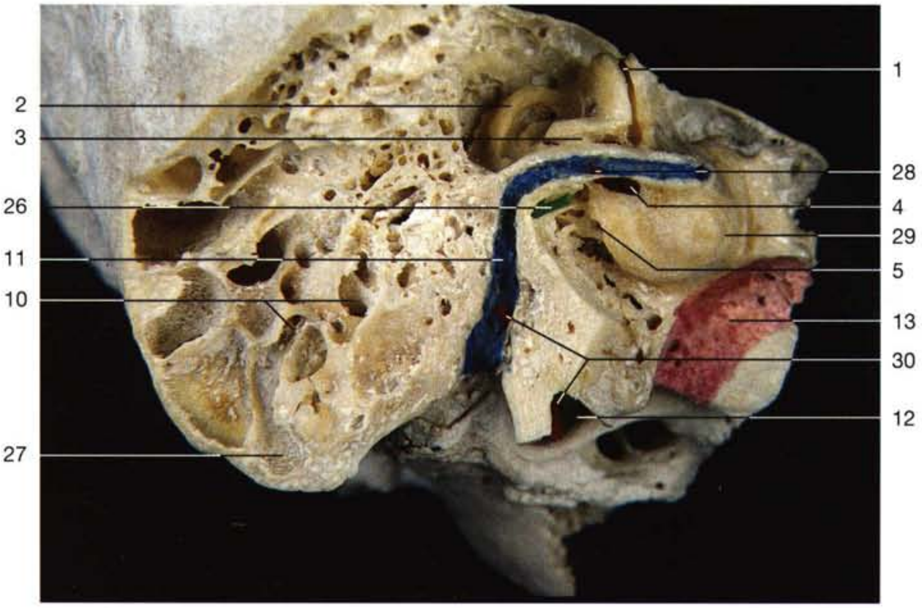

Right temporal bone (lateral aspect). Mastoid air cells and facial canal had been opened. The 3 semicircular canals were dissected.

1 Anterior semicircular canal (red)
2 Posterior semicircular canal (yellow)
3 Lateral or horizontal semicircular canal (green)
4 **Fenestra vestibuli**
5 **Fenestra cochleae**
6 Tympanic cavity
7 Mastoid process
8 **Petrotympanic fissure** (red probe: **chorda tympani**)
9 Lateral pterygoid plate
10 Mastoid air cells
11 **Facial canal** (blue)
12 Foramen ovale
13 **Carotid canal** (red)
14 **Tympanic ring**
15 Petromastoid part of temporal bone
16 Squamous part of temporal bone
17 Squamomastoid suture
18 Zygomatic process of temporal bone
19 Incisure of tympanic ring
20 **Promontory**
21 Apex of cochlea (cupula)
22 Spiral canal of cochlea at base of cochlea
23 Epitympanic recess
24 Auditory ossicles and tympanic cavity
25 Hypotympanic recess
26 Canaliculus chordae tympani (green probe)
27 Mastoid process
28 Canaliculus for stapedius nerve (red)
29 **Cochlea**
30 Canaliculus mastoideus (red probe)

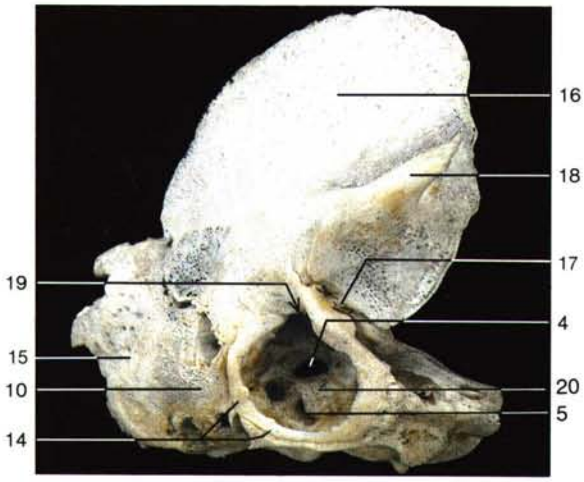

Right temporal bone of the newborn (lateral aspect).

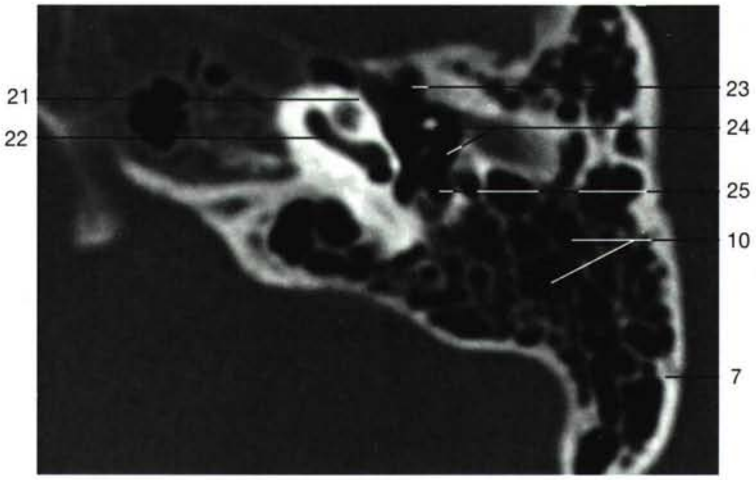

Frontal section through petrous part. (CT-Scan.)

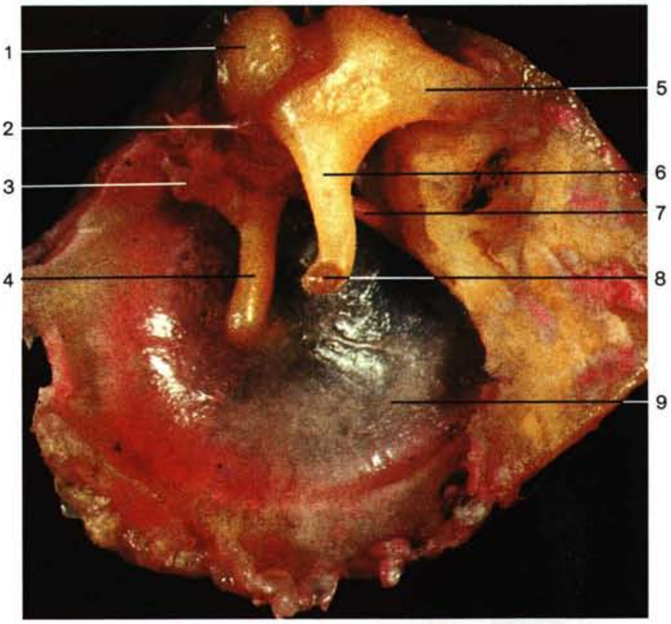

1 Head of malleus
2 Anterior ligament of malleus
3 Tendon of tensor tympani muscle
4 Handle of **malleus**
5 Short crus of incus
6 Long crus of **incus**
7 Chorda tympani
8 Lenticular process
9 **Tympanic membrane**

Tympanic membrane with malleus and incus (internal aspect; right side).

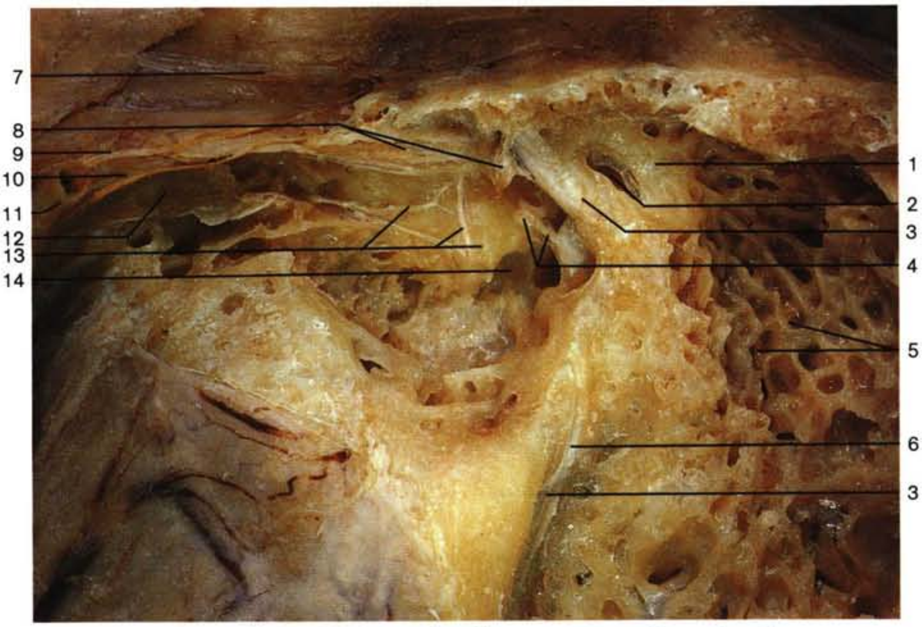

1 Tympanic antrum
2 Lateral semicircular canal (opened)
3 **Facial canal**
4 Stapes with tendon of stapedius
5 Mastoid air cells
6 Chorda tympani (intracranial part)
7 Greater petrosal nerve
8 Tensor tympani muscle (processus cochleariformis)
9 Lesser petrosal nerve
10 Anterior tympanic artery
11 Middle meningeal artery
12 **Auditory tube**
13 **Promontory with tympanic plexus**
14 Fenestra cochleae

Tympanic cavity, medial wall. External auditory meatus and lateral wall of tympanic cavity together with incus. Malleus and tympanic membrane have been removed; mastoid air cells are opened (left side).

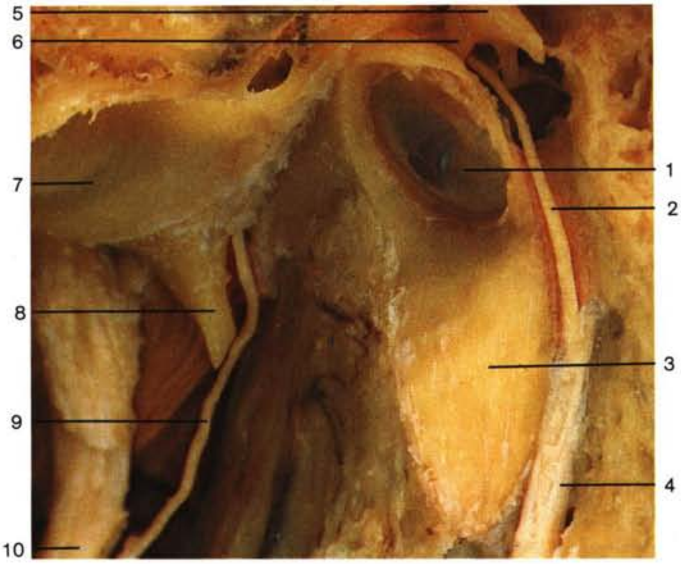

1 **Tympanic membrane**
2 Chorda tympani (intracranial part)
3 Floor of the external acoustic meatus
4 **Facial nerve** and facial canal
5 Incus
6 Head of malleus
7 Mandibular fossa
8 Spine of sphenoid
9 **Chorda tympani** (extracranial part)
10 Styloid process

Tympanic membrane (lateral aspect). External acoustic meatus and facial canal have been opened to expose the chorda tympani (magn. ~1.5×) (left side).

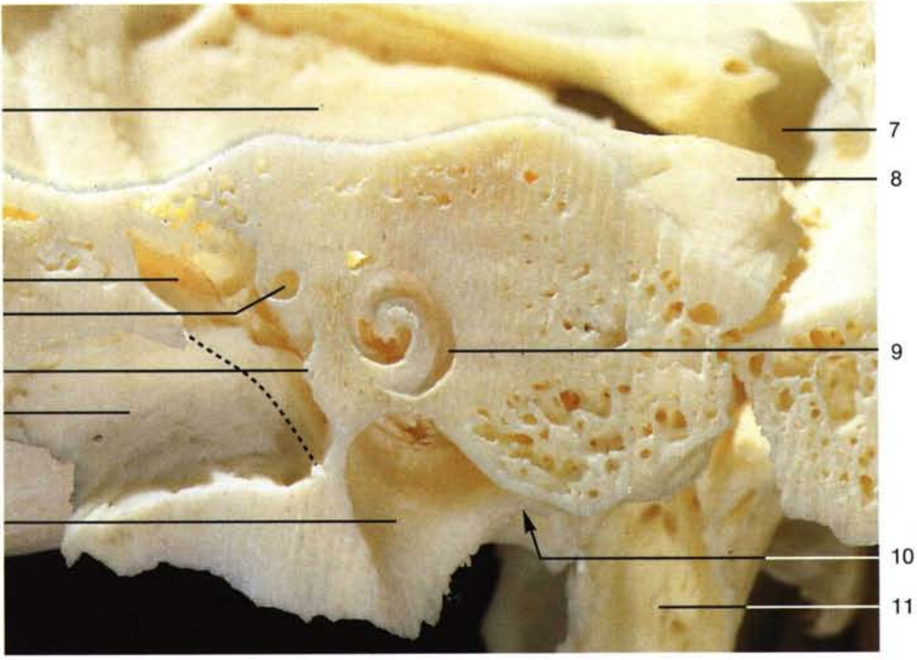

Frontal section through the petrous part of the left temporal bone at the level of the cochlea (posterior aspect). Position of tympanic membrane indicated by dotted line.

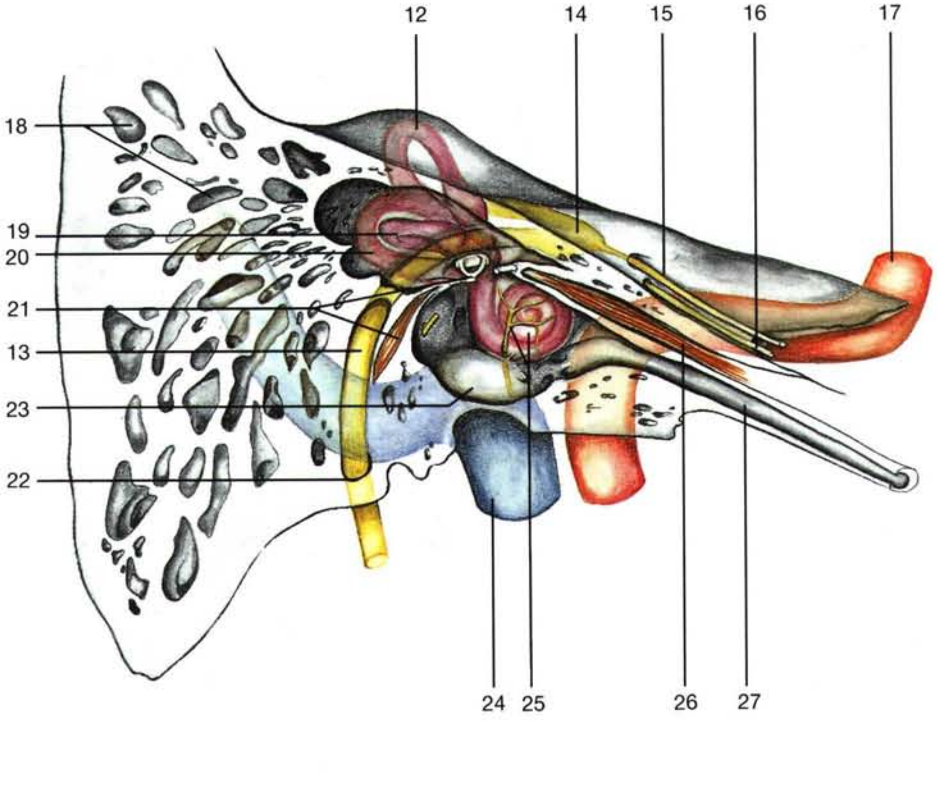

Medial wall of tympanic cavity and its relation to neighboring structures of the inner ear, facial nerve, and blood vessels. (Schematic drawing.) Frontal section through the right temporal bone (anterior aspect).

1 Anterior surface of the pyramid
2 Mastoid antrum
3 Lateral semicircular canal
4 Cochleariform process
5 External acoustic meatus
6 **Jugular fossa**
7 Foramen lacerum
8 Apex of petrous part
9 Position of **cochlea** (modiolus with crista spiralis ossea)
10 Carotid canal
11 Pterygoid process
12 Anterior semicircular duct
13 **Facial nerve**
14 Geniculate ganglion
15 Greater petrosal nerve
16 Lesser petrosal nerve
17 **Internal carotid artery**
18 Mastoid air cells
19 Lateral semicircular duct
20 Posterior semicircular duct
21 Stapes with stapedius muscle
22 Stylomastoid foramen
23 Inferior recess of tympanic cavity (hypotympanon)
24 **Internal jugular vein**
25 Promontory with tympanic plexus (position of cochlea)
26 Tensor muscle of tympanum
27 **Auditory tube**

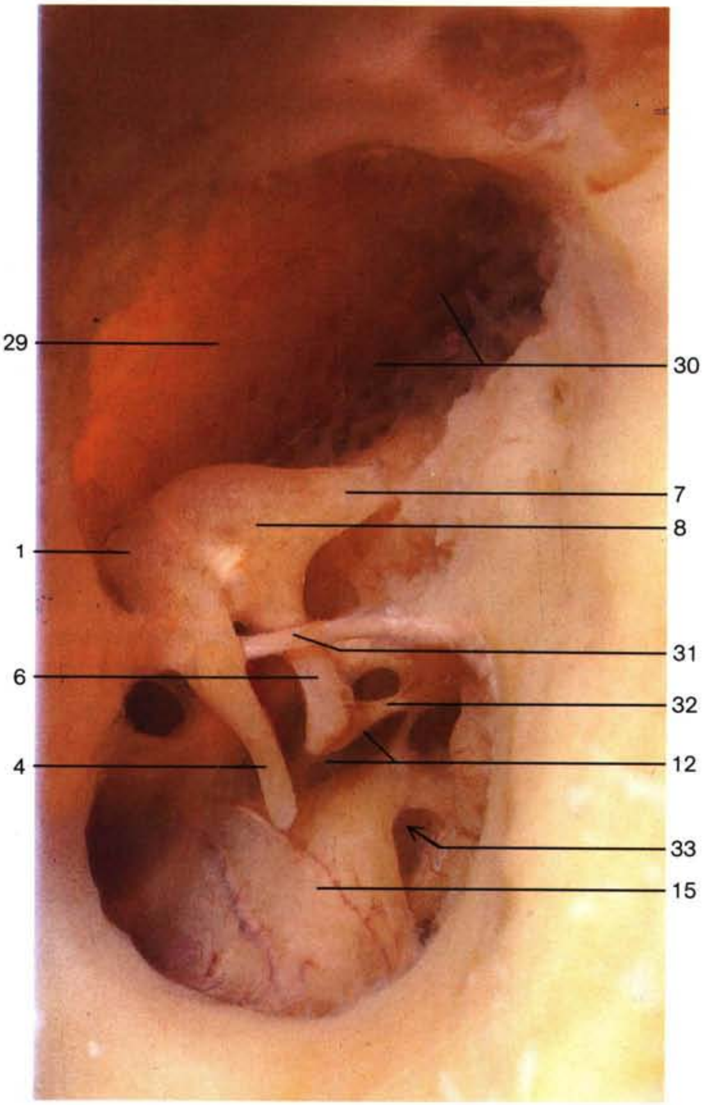

Tympanic cavity with malleus, incus and stapes, left side (lateral aspect). Tympanic membrane removed, mastoid antrum opened.

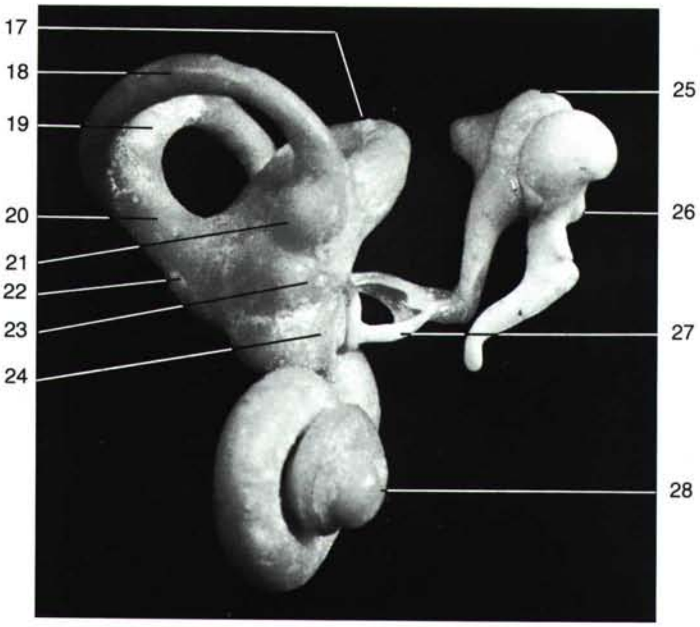

Chain of auditory ossicles in connection with the inner ear, left side (anterior-lateral aspect).

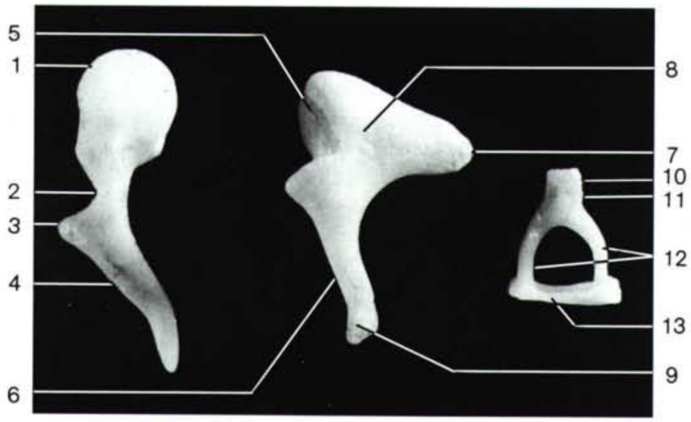

Auditory ossicles (isolated).

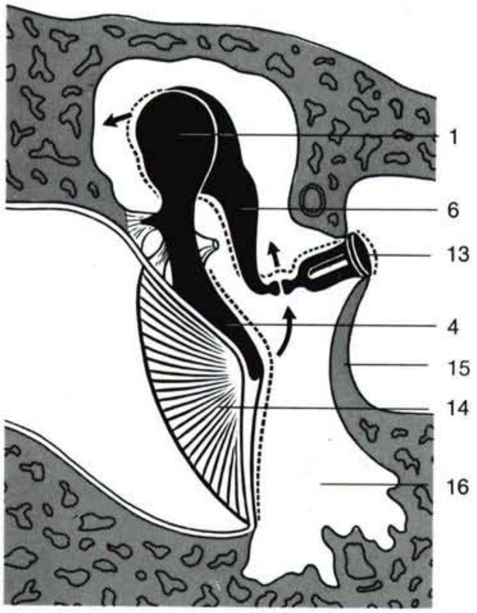

Position and movements of the auditory ossicles. (Schematic diagram.)

Malleus
1 Head
2 Neck
3 Lateral process
4 Handle

Incus
5 Articular facet for malleus
6 Long crus
7 Short crus
8 Body
9 Lenticular process

Stapes
10 Head
11 Neck
12 Anterior and posterior crura
13 Base

Walls of tympanic cavity
14 Tympanic membrane
15 Promontory
16 Hypotympanic recess of tympanic cavity

Internal ear (labyrinth)
17 Lateral semicircular duct
18 Anterior semicircular duct
19 Posterior semicircular duct
20 Common crus
21 Ampulla
22 Beginning of endolymphatic duct
23 Utricular prominence
24 Saccular prominence
25 Incus
26 Malleus
27 Stapes
28 Cochlea

Tympanic cavity
29 Epitympanic recess
30 Mastoid antrum
31 Chorda tympani
32 Tendon of stapedius muscle
33 Round window (fenestra cochleae)

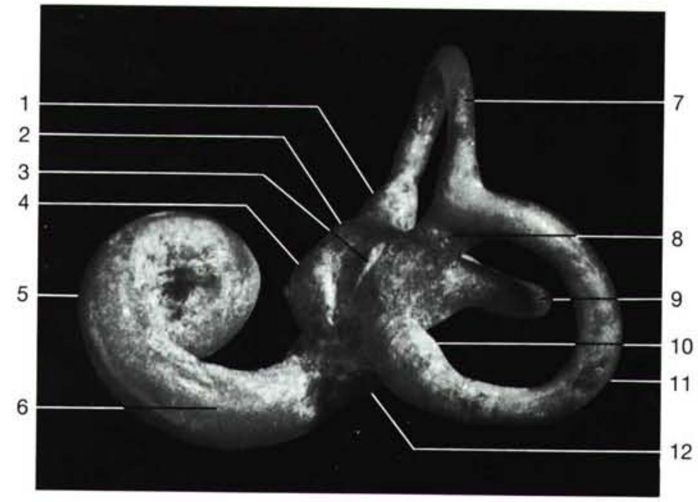

Cast of the right labyrinth (postero-medial aspect).

1 Ampulla (anterior semicircular canal)
2 Elliptical recess
3 Aqueduct of the vestibule
4 Spherical recess
5 **Cochlea**
6 Base of cochlea
7 **Anterior semicircular canal**
8 Crus commune or common limb
9 **Lateral semicircular canal**
10 Posterior bony ampulla
11 **Posterior semicircular canal** (posterior canal)
12 **Fenestra cochleae**
13 Bony ampulla
14 **Fenestra vestibuli**
15 Cupula of cochlea
16 External acoustic meatus
17 Mastoid air cells
18 Tympanic cavity and fenestra cochleae (probe)
19 External acoustic meatus
20 Facial canal
21 Base of cochlea and canalis musculotubarius
22 Malleus and incus
23 Stapes
24 Tympanic membrane
25 Tympanic cavity
26 Aqueduct of cochlea
27 Saccus endolymphaticus
28 Ductus endolymphaticus
29 Macula of utricle
30 Macula of saccule

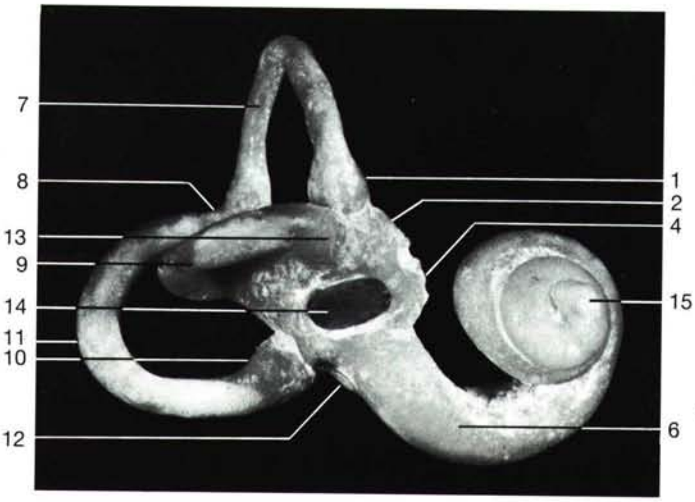

Cast of the right labyrinth (lateral aspect).

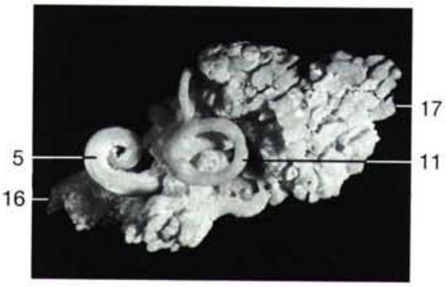

Cast of the labyrinth and mastoid cells. Life size (posterior aspect).

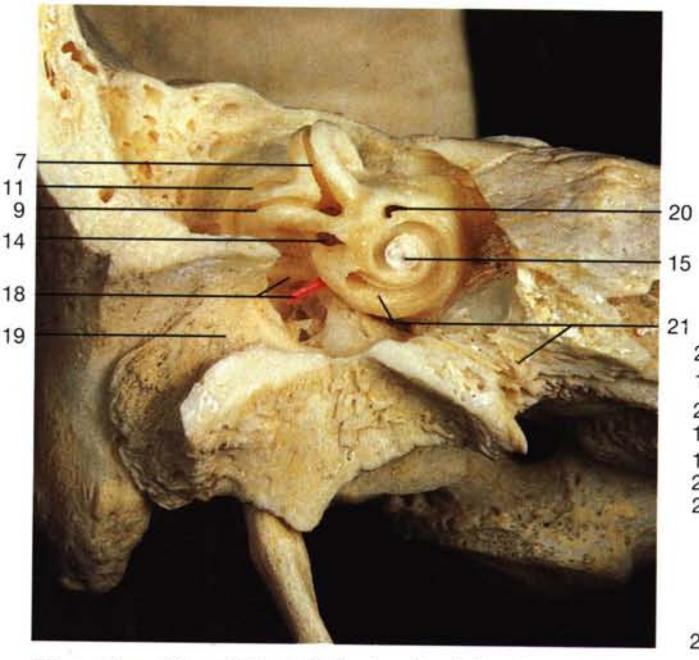

Dissection of bony labyrinth in situ. Semicircular canals and cochlear duct opened.

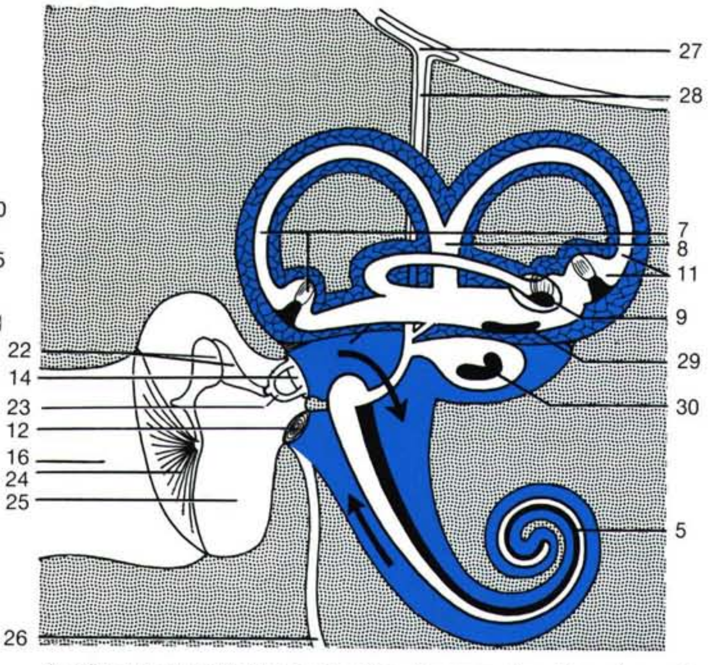

Auditory and vestibular apparatus. Arrows: direction of sound waves; blue = perilymphatic ducts. (Schematic diagram.)

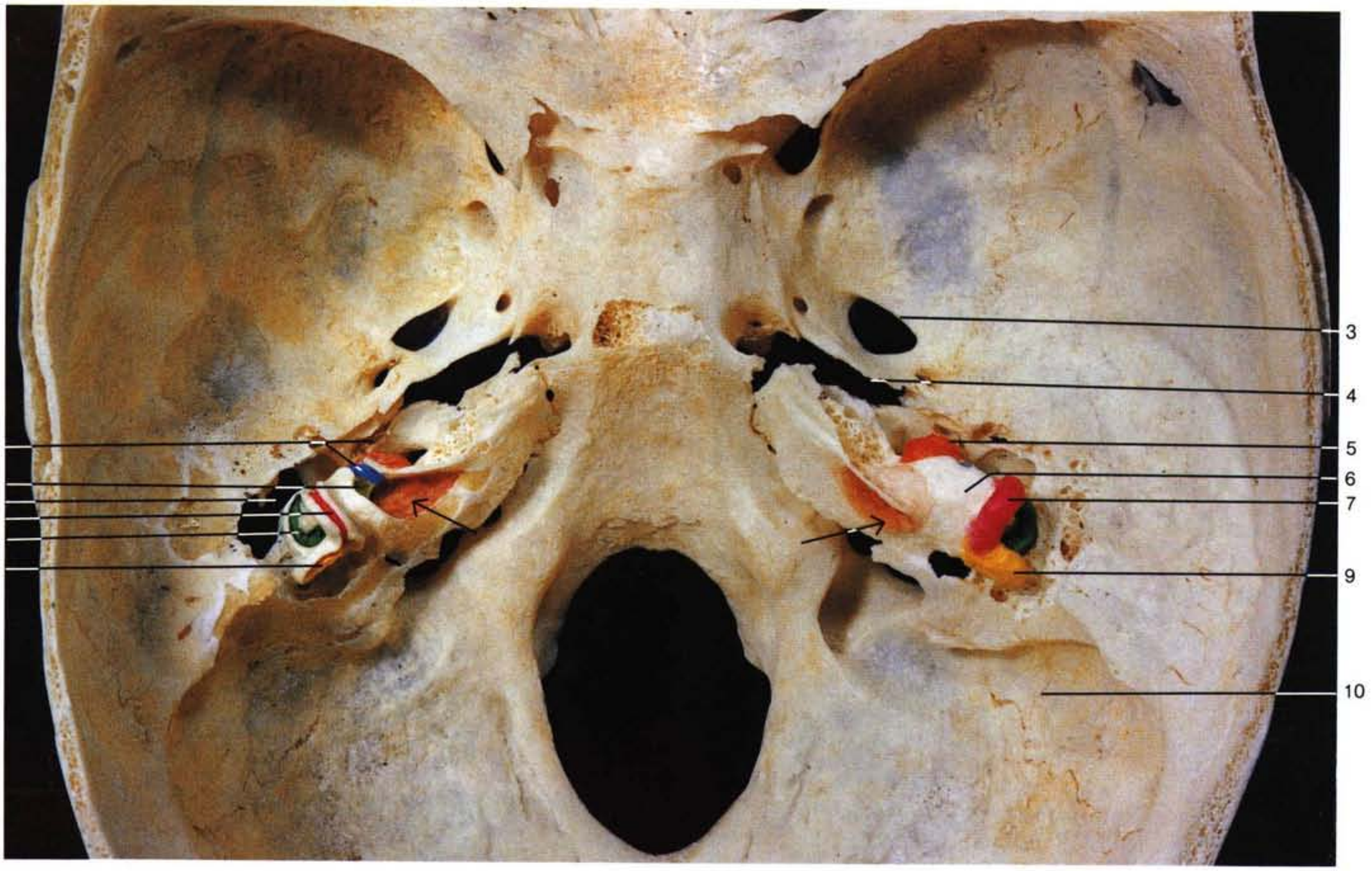

Bony labyrinth, petrous part of the temporal bone (from above). At left: semicircular canals opened, at right closed. Arrows: internal acoustic meatus.

1 Facial canal and semicanal of auditory tube
2 Superior vestibular area
3 Foramen ovale
4 Foramen lacerum
5 Cochlea
6 Vestibule
7 Anterior semicircular canal
8 Lateral semicircular canal
9 Posterior semicircular canal
10 Groove for sigmoid sinus
11 Sigmoid sinus
12 Tympanic cavity
13 Auditory tube
14 Mastoid air cells
15 Facial and vestibulocochlear nerves
16 Temporal fossa
17 Fenestra vestibuli
18 Promontory
19 Zygomatic process
20 Fenestra cochleae
21 Mastoid process

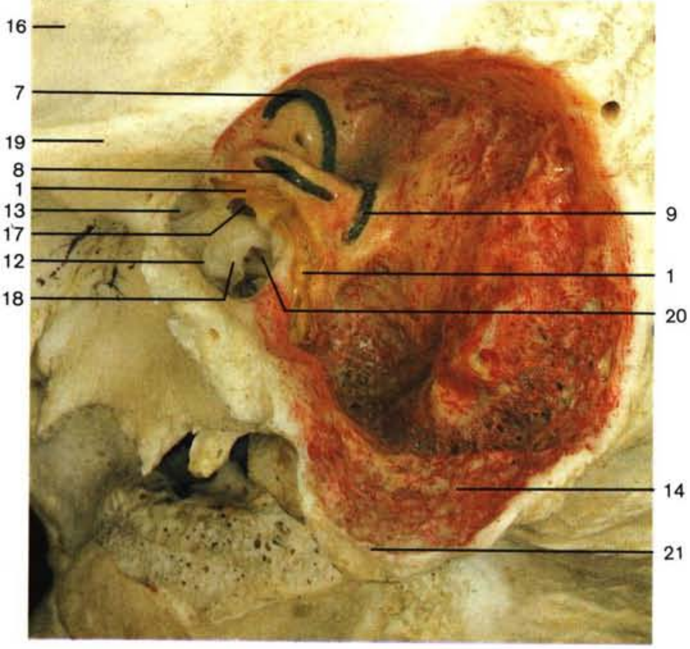

Bony labyrinth (left lateral aspect). Temporal and tympanic bone partly removed, semicircular canals opened.

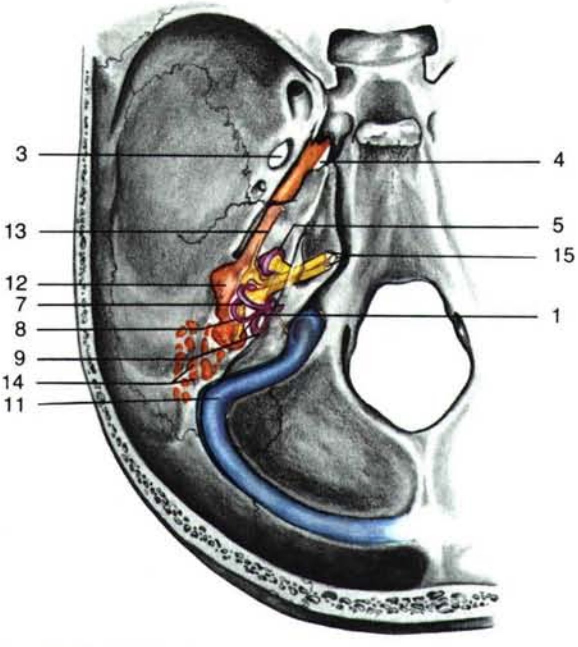

Internal ear. Diagram showing the position of the membranous labyrinth and the tympanic cavity.

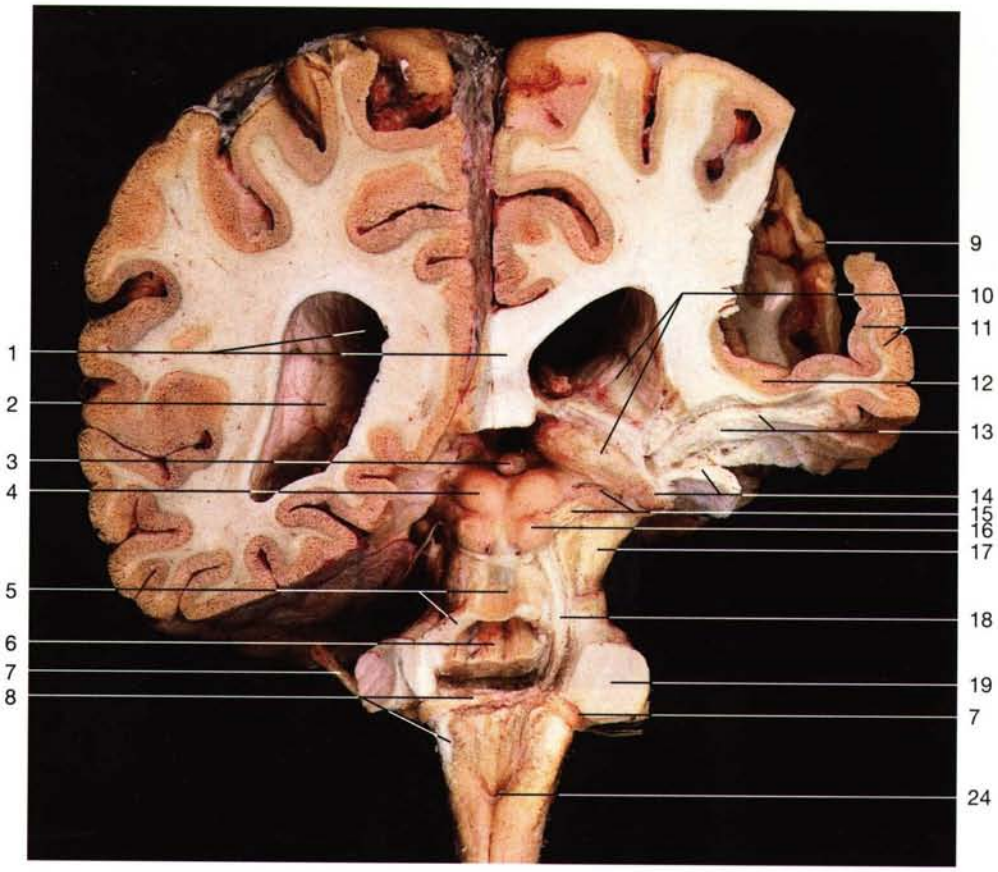

Dissection of the brain stem showing the acoustic pathway. Cerebellum and posterior part of the two hemispheres have been removed (dorsal aspect).

1 Left lateral ventricle and corpus callosum
2 Thalamus
3 Pineal gland (epiphysis)
4 Superior colliculus
5 Anterior cerebellar velum and superior cerebellar peduncle
6 Rhomboid fossa
7 **Vestibulocochlear nerve** (n. VIII)
8 **Dorsal acoustic striae** and inferior cerebellar peduncle
9 Insular lobe
10 Caudate nucleus and thalamus
11 Temporal lobe (superior temporal gyrus) (area of **acoustic centers**)
12 Transverse temporal gyri of Heschl (area of **primary acoustic centers**)
13 Acoustic radiation of internal capsule
14 Lateral geniculate body and optic radiation (cut)
15 **Medial geniculate body** and brachium of inferior colliculus
16 **Inferior colliculus**
17 Cerebral peduncle
18 **Lateral lemniscus**
19 Middle cerebellar peduncle
20 Dorsal (posterior) cochlear nucleus
21 Ventral (anterior) cochlear nucleus
22 Inferior olive with tractus olivocochlearis of Rasmussen (red)
23 Ganglion spirale
24 Obex
25 Frontal lobe
26 Temporal lobe
27 Middle temporal gyrus (area of **tertiary acoustic centers)**
28 Trapezoid body

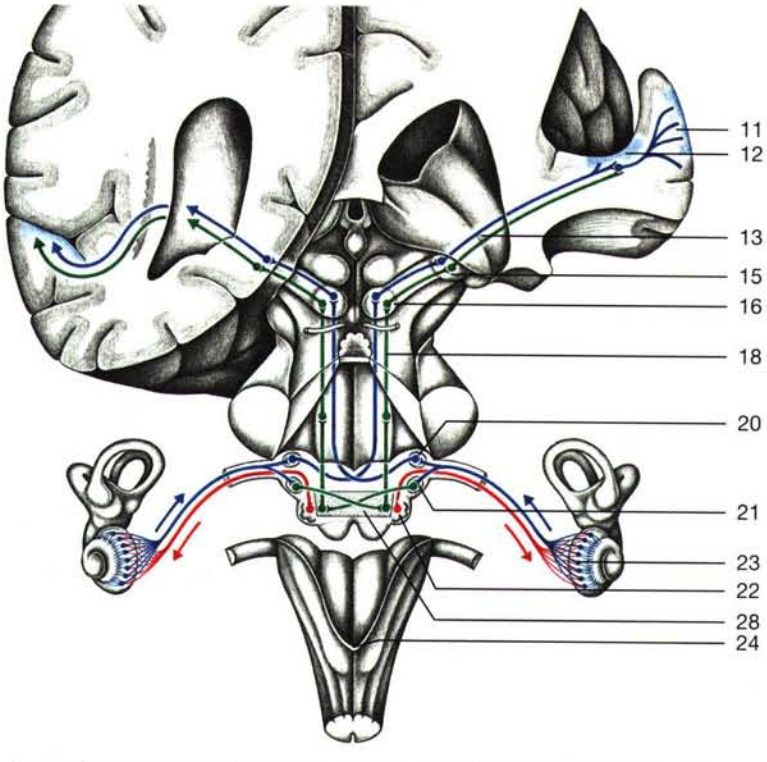

Acoustic pathway (schematic drawing, compare with figure above).

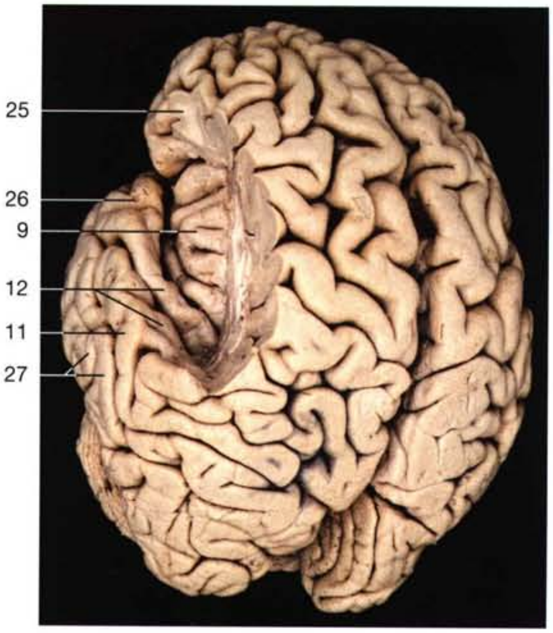

Acoustic areas in the left hemisphere (superior lateral aspect). Parts of the frontal and parietal lobe have been removed.

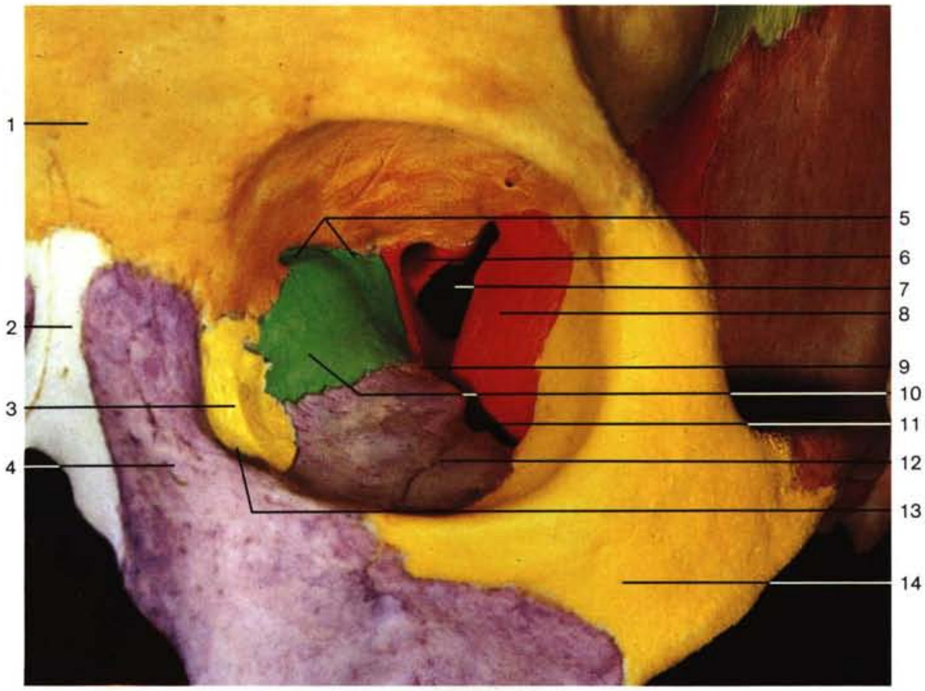

Bones of the left orbit (indicated by different colors).

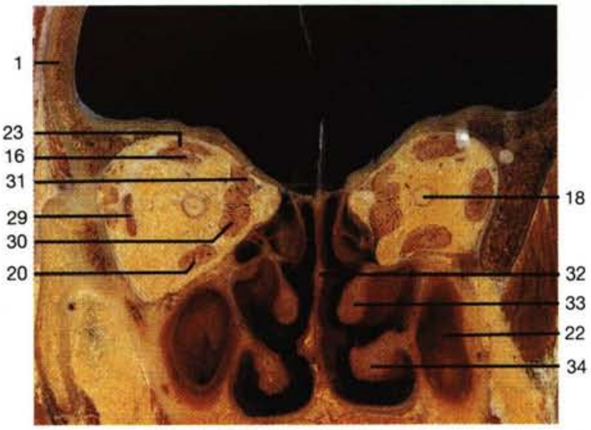

Frontal section through the posterior part of the orbit.

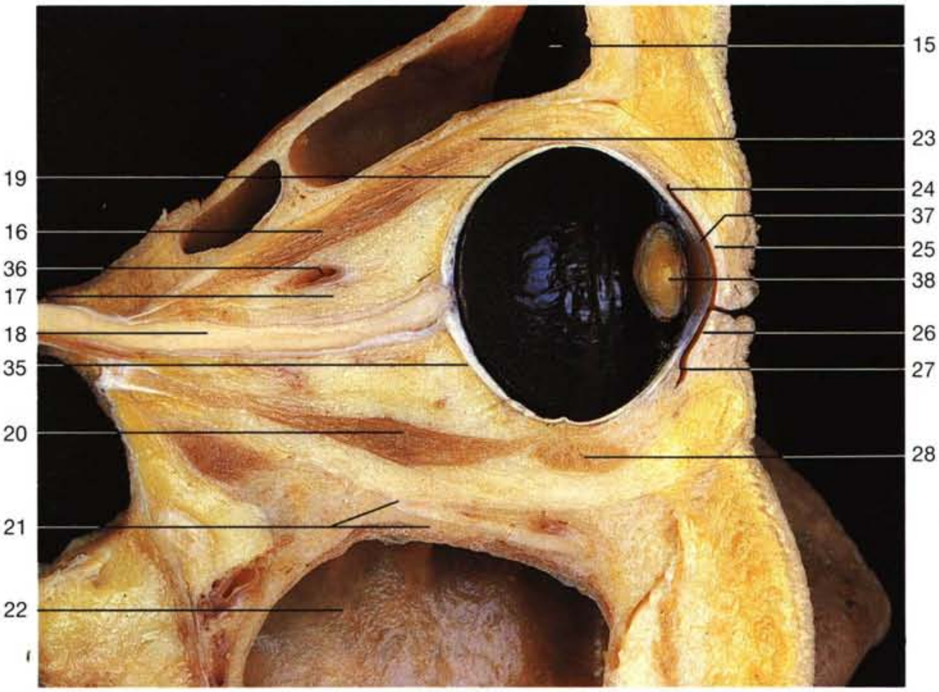

Sagittal section through orbit and eyeball.

1 Frontal bone
2 Nasal bone
3 Lacrimal bone
4 Maxilla (frontal process)
5 Ethmoidal foramina
6 Lesser wing of sphenoid bone and optic canal
7 Superior orbital fissure
8 Greater wing of sphenoid bone
9 Orbital process of palatine bone
10 Orbital plate of ethmoid bone
11 Inferior orbital fissure
12 Infraorbital sulcus
13 Nasolacrimal canal
14 Zygomatic bone
15 Frontal sinus
16 **Superior rectus muscle**
17 Orbital fatty tissue
18 **Optic nerve**
19 **Sclera**
20 **Inferior rectus muscle**
21 Periorbita and maxilla
22 Maxillary sinus
23 Levator palpebrae superioris muscle
24 **Superior conjunctival fornix**
25 Superior tarsal plate
26 Inferior tarsal plate
27 **Inferior conjunctival fornix**
28 **Inferior oblique muscle**
29 **Lateral rectus muscle**
30 **Medial rectus muscle**
31 **Superior oblique muscle**
32 Nasal septum
33 Middle nasal concha
34 Inferior nasal concha
35 Tenon's space
36 Ophthalmic artery
37 **Cornea**
38 **Lens**

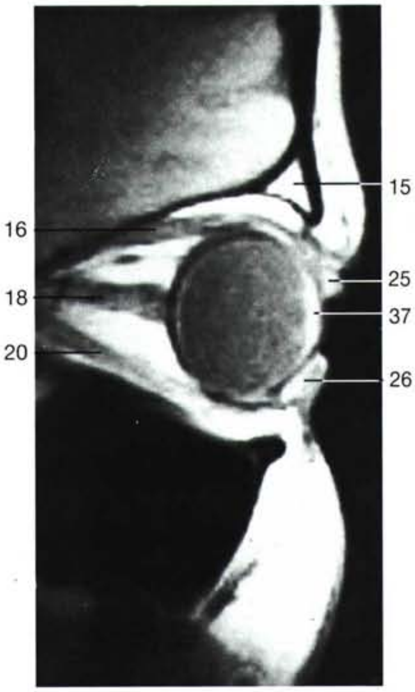

Sagittal section through orbit and eyeball. (MR-Scan.)

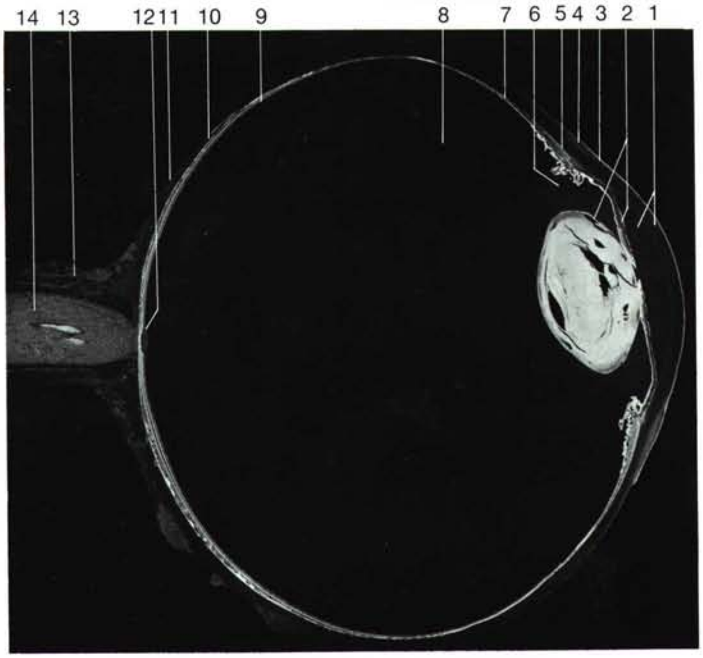

Horizontal section through the human eye (2×).

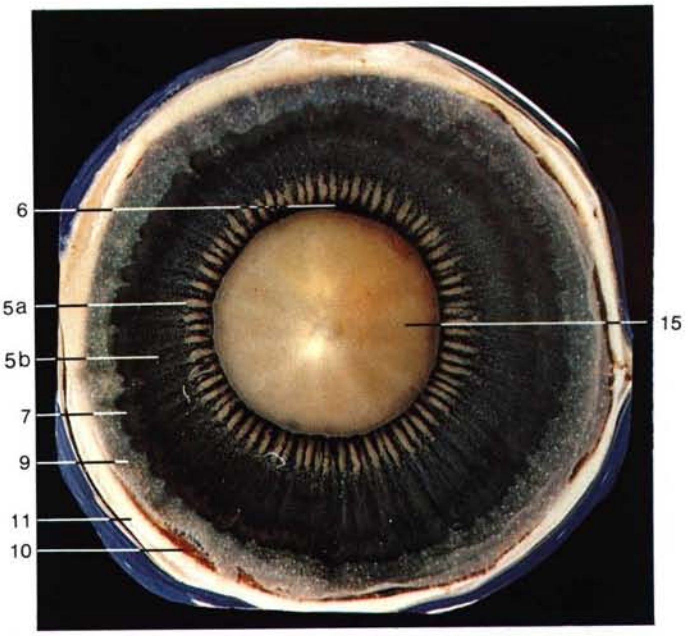

Anterior segment of the eyeball (posterior aspect). The opacity of the lens is an artifact.

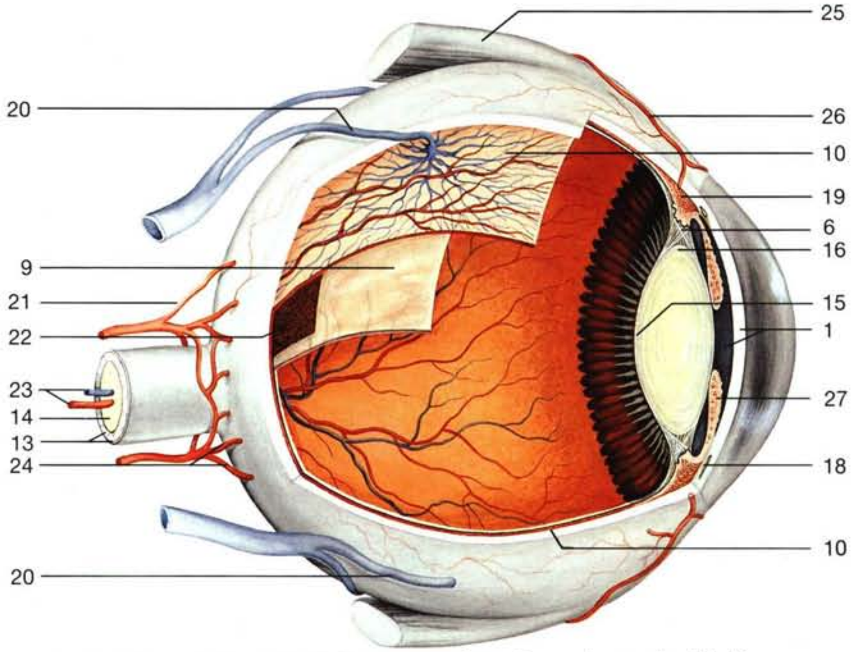

Organization of the eyeball. Demonstration of vascular tunic of bulb. (Schematic drawing.)

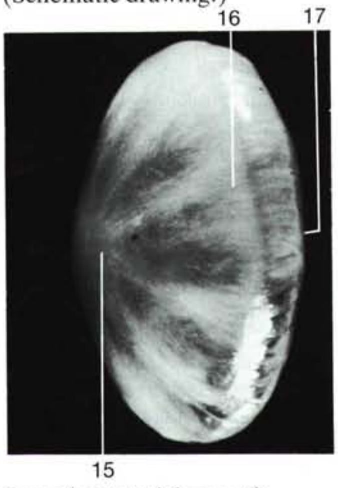

Lens (equatorial aspect), anterior pole to the right.

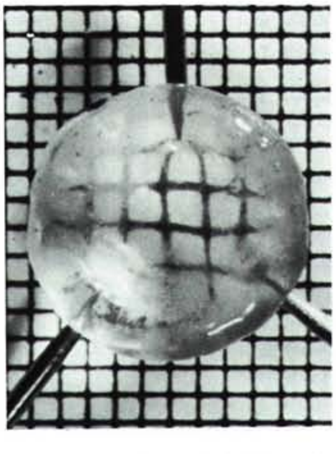

Lens (frontal aspect). Note the magnification effect.

1 **Cornea** and anterior chamber
2 Iris and lens
3 Transitional zone between corneal and conjunctival epithelium
4 Conjunctiva of the eyeball
5 **Ciliary body**
  a Ciliary processes (pars plicata)
  b Ciliary ring (pars plana)
6 Zonular fibers
7 **Ora serrata**
8 Vitreous body
9 **Retina**
10 **Choroid**
11 **Sclera**
12 Optic disc
13 Dura mater and subarachnoid space
14 **Optic nerve** (n. II)
15 Lens (posterior pole)
16 Equator of lens
17 **Lens** (anterior pole)
18 Canal of Schlemm
19 Ciliary muscle
20 Vena vorticosa
21 Long posterior ciliary artery
22 Retinal pigmented epithelium
23 Central retinal artery and vein
24 Short posterior ciliary arteries
25 External ocular muscle
26 Anterior ciliary artery
27 **Iris**

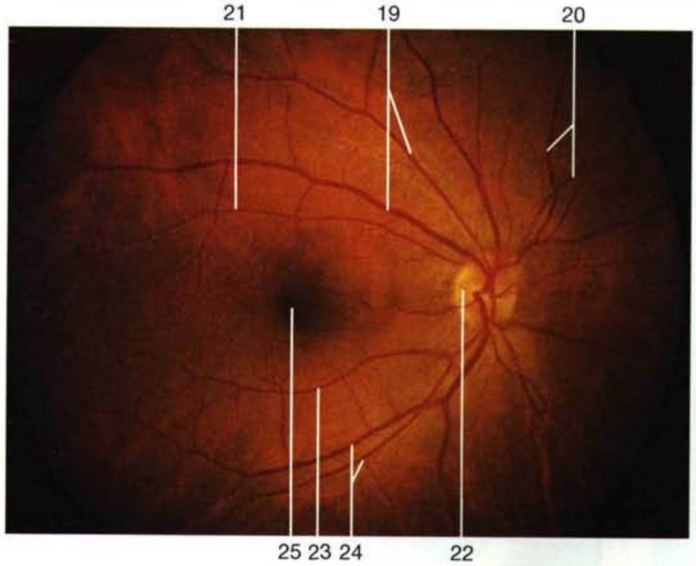

Fundus of a normal right eye (courtesy of Prof. Dr. R. Okamura, Univ. Eye Dept., Kumamoto/Japan). Notice, the arteries are smaller and lighter than the veins.

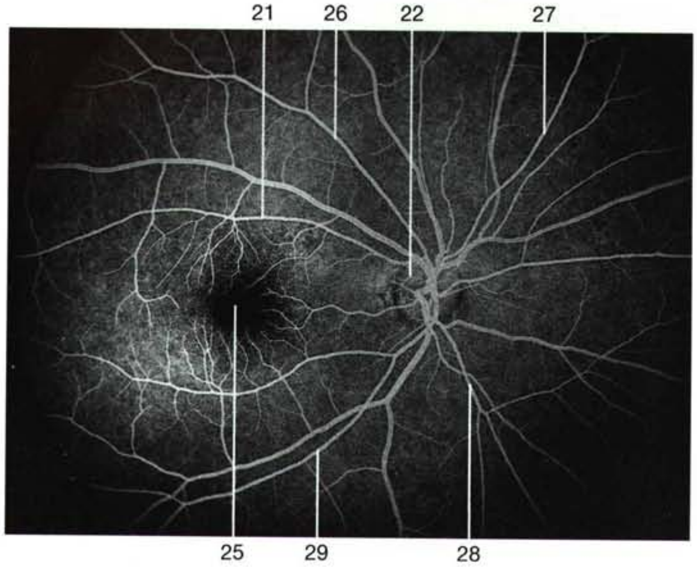

Fluorescent angiography of the right eye; retinal vessels. The same eye as above. (Courtesy of Prof. Dr. R. Okamura, Univ. Eye Dept., Kumamoto/Japan.)

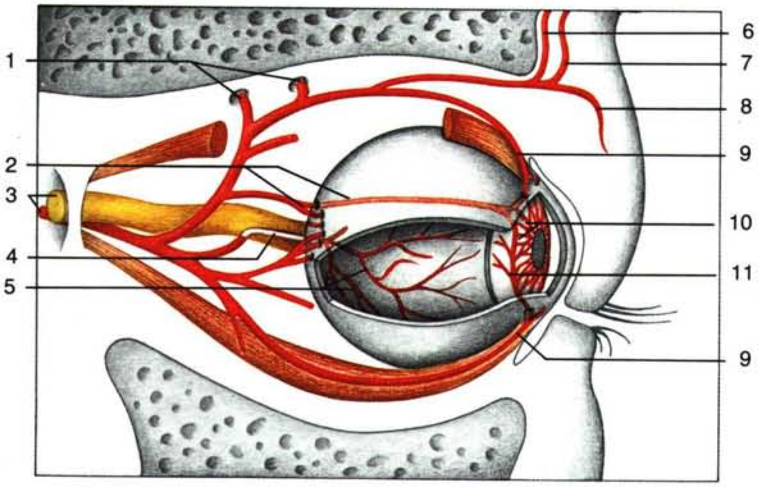

Diagram of the ophthalmic artery and its branches.

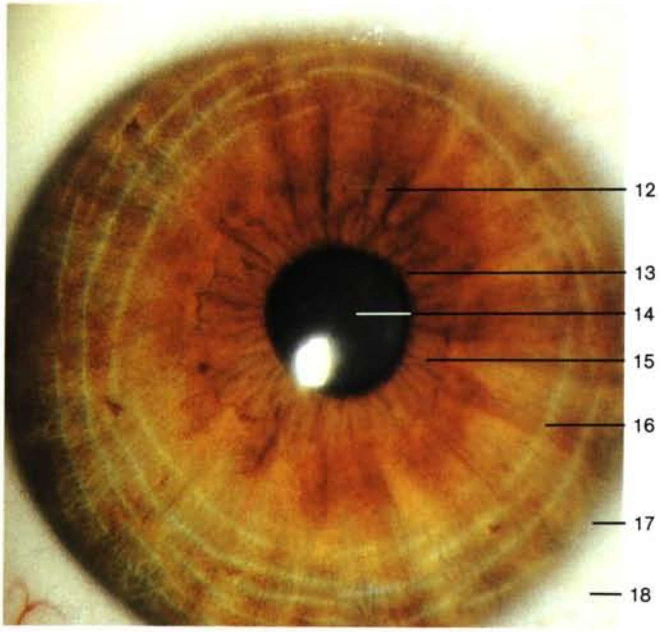

Anterior segment of the human eye (courtesy of Prof. Dr. G.O.H. Naumann, Eye Dept., University of Erlangen, FRG). Note the colored iris and the anterior lens surface.

1 Posterior and anterior ethmoidal arteries
2 Long and short **posterior ciliary arteries**
3 Optic nerve and **ophthalmic artery**
4 Central retinal artery
5 Retinal arteries
6 Supratrochlear artery
7 Supraorbital artery
8 Dorsal nasal artery
9 **Anterior ciliary artery**
10 Iridial arteries
11 Circulus arteriosus major of iris
12 Iridial fold
13 Pupillary margin of iris
14 Anterior pole of lens
15 Lesser circle of **iris**
16 Greater circle of iris
17 Margin of cornea or limbus
18 Sclera
19 Superior temporal artery and vein of retina
20 Superior nasal artery and vein of retina
21 Superior macular artery
22 **Optic disc**
23 Inferior macular artery
24 Inferior temporal artery and vein
25 **Fovea centralis** and **macula lutea**
26 Superior temporal artery
27 Superior nasal artery
28 Inferior nasal artery
29 Inferior temporal artery
of retina

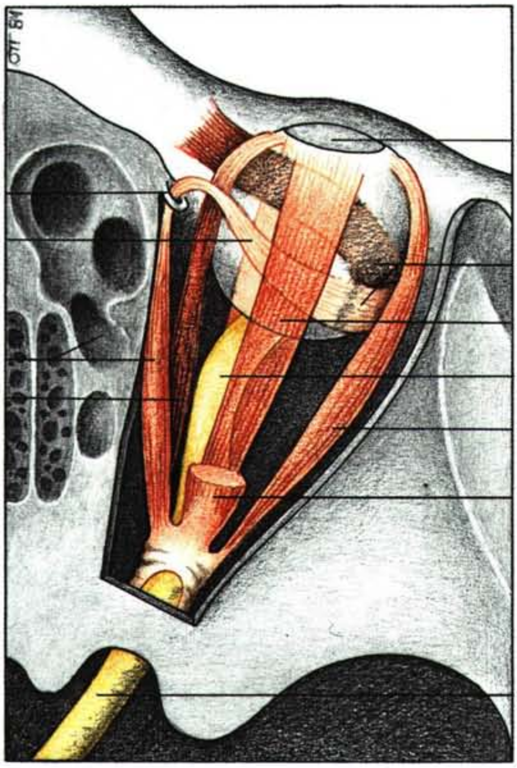

Schematic diagram of the extraocular muscles. Right orbit (from above). Levator palpebrae superioris muscle has been severed.

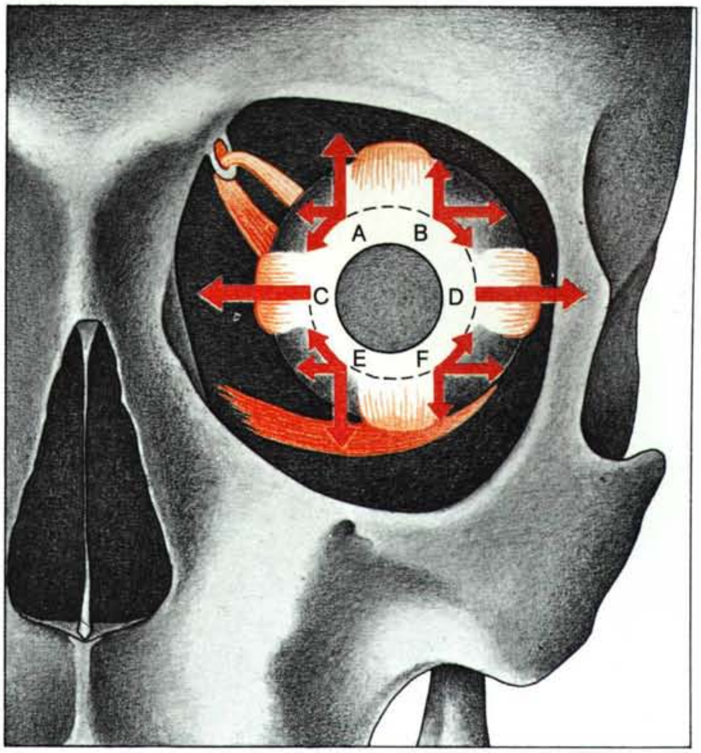

The action of the extraocular muscles. Left orbit (anterior aspect).

A Superior rectus muscle
B Inferior oblique muscle
C Medial rectus muscle
D Lateral rectus muscle
E Inferior rectus muscle
F Superior oblique muscle

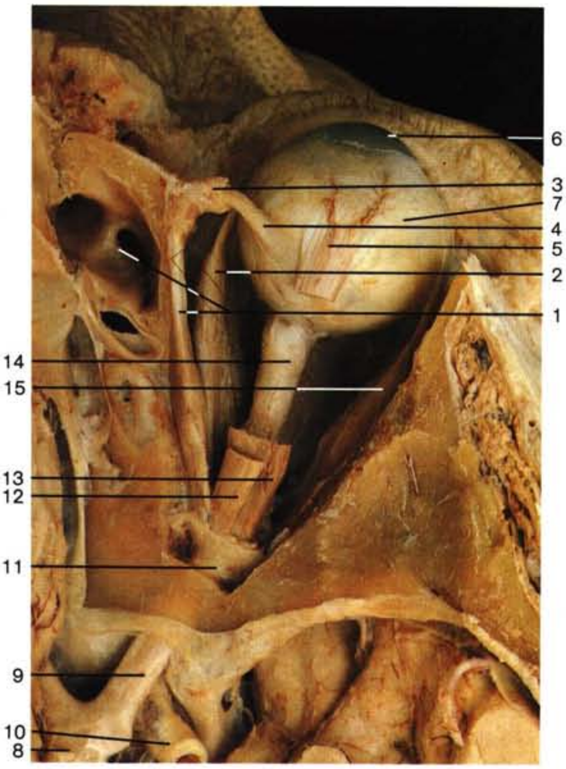

Right orbit with eyeball and extraocular muscles (from above). The roof of the orbit has been removed, the superior rectus muscle and the levator palpebrae superioris muscle have been severed.

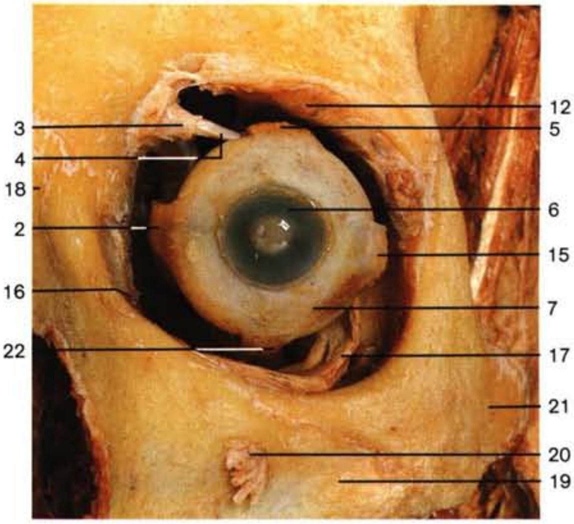

Left orbit with eyeball and extraocular muscles (anterior aspect). Lids, conjunctiva and lacrimal apparatus have been removed.

1 **Superior oblique muscle** and ethmoid air cells
2 **Medial rectus muscle**
3 Trochlea
4 Tendon of superior oblique muscle
5 **Superior rectus muscle**
6 Cornea
7 Eyeball
8 Optic chiasma
9 Optic nerve (intracranial part)
10 Internal carotid artery
11 **Common annular tendon**
12 Levator palpebrae superioris muscle
13 **Superior rectus muscle**
14 Optic nerve (extracranial part)
15 **Lateral rectus muscle**
16 Nasolacrimal duct
17 **Inferior oblique muscle**
18 Nasal bone
19 Maxilla
20 Infraorbital foramen and nerves
21 Zygomatic bone
22 **Inferior rectus muscle**

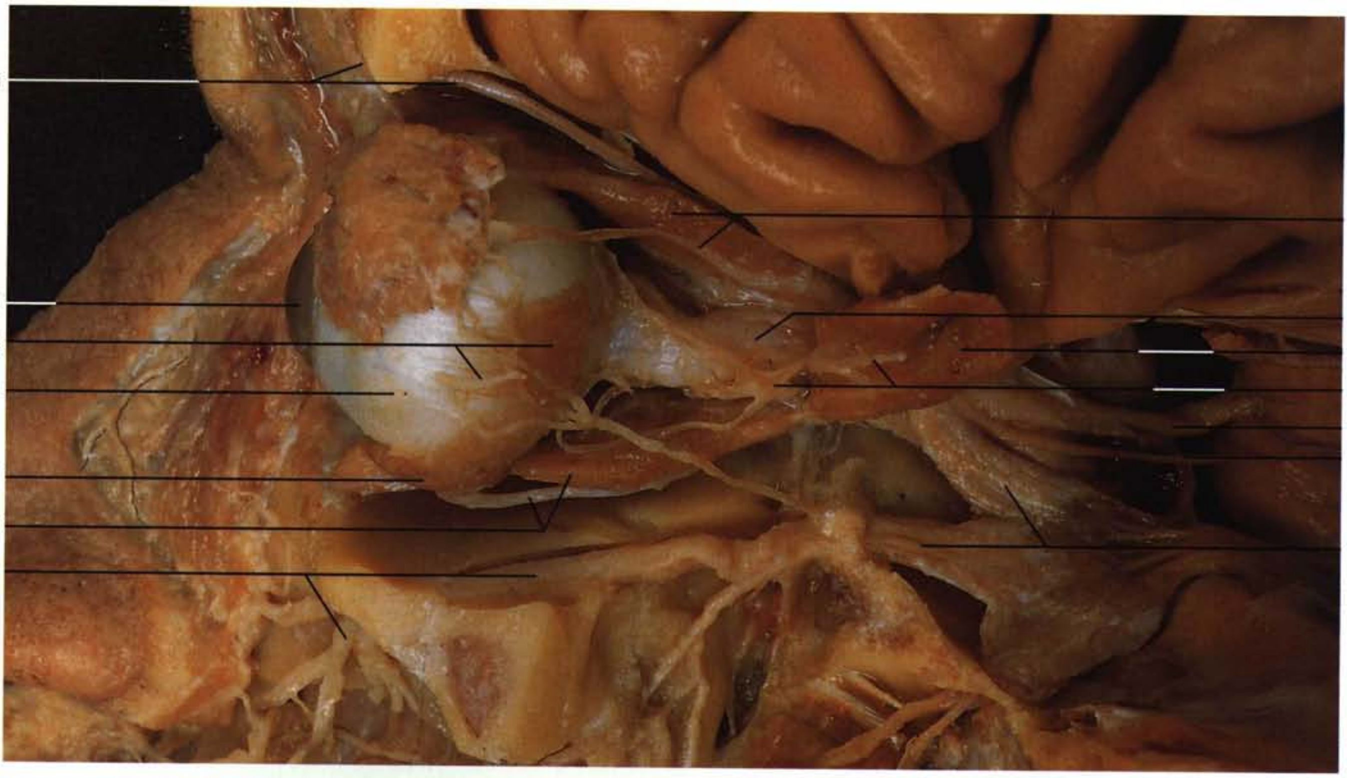

Extraocular muscles and their nerves (lateral aspect of left eye). Lateral rectus divided and reflected.

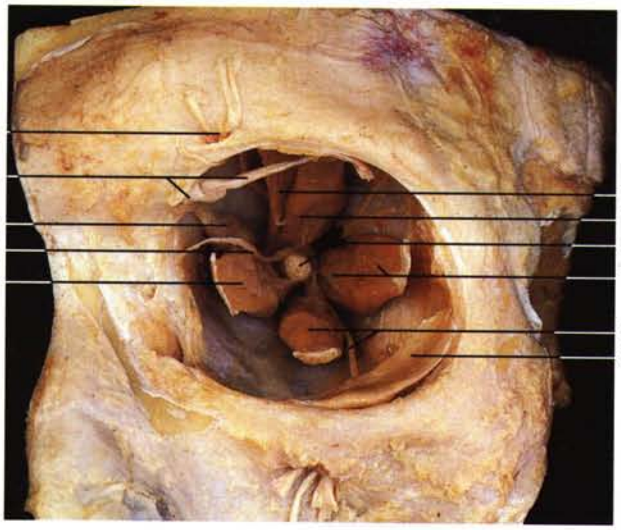

Left orbit with extraocular muscles (anterior aspect). Eyeball removed.

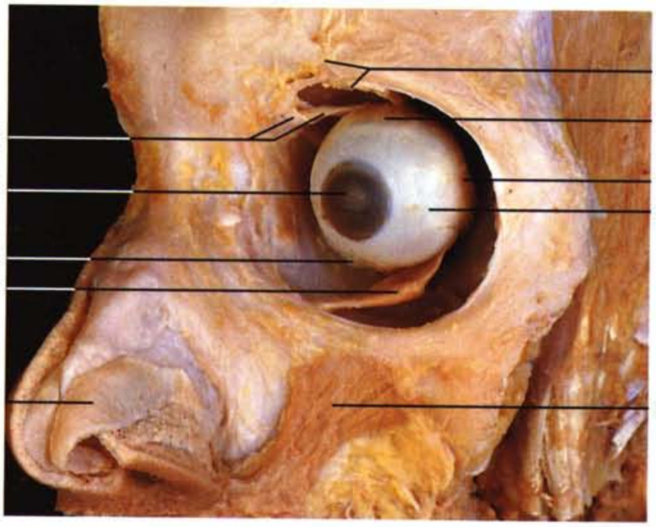

Extraocular eye muscles (anterior-lateral aspect).

1 Supraorbital nerve
2 Cornea
3 Insertion of lateral rectus muscle
4 Eyeball (sclera)
5 **Inferior oblique muscle**
6 **Inferior rectus muscle** and inferior branch of oculomotor nerve
7 Infraorbital nerve
8 Superior rectus muscle and lacrimal nerve
9 Optic nerve
10 **Lateral rectus muscle**
11 Ciliary ganglion and abducens nerve (n. VI)
12 Oculomotor nerve (n. III)
13 Trochlear nerve (n. IV)
14 Ophthalmic nerve (n. V_1) and maxillary nerve (n. V_2)
15 Trochlea and tendon of superior oblique muscle
16 **Superior oblique muscle**
17 **Medial rectus muscle**
18 Levator palpebrae superioris muscle
19 **Superior rectus muscle**
20 **Inferior rectus muscle**
21 Greater alar cartilage
22 Supraorbital nerve and levator palpebrae superioris muscle
23 Levator labii superioris muscle

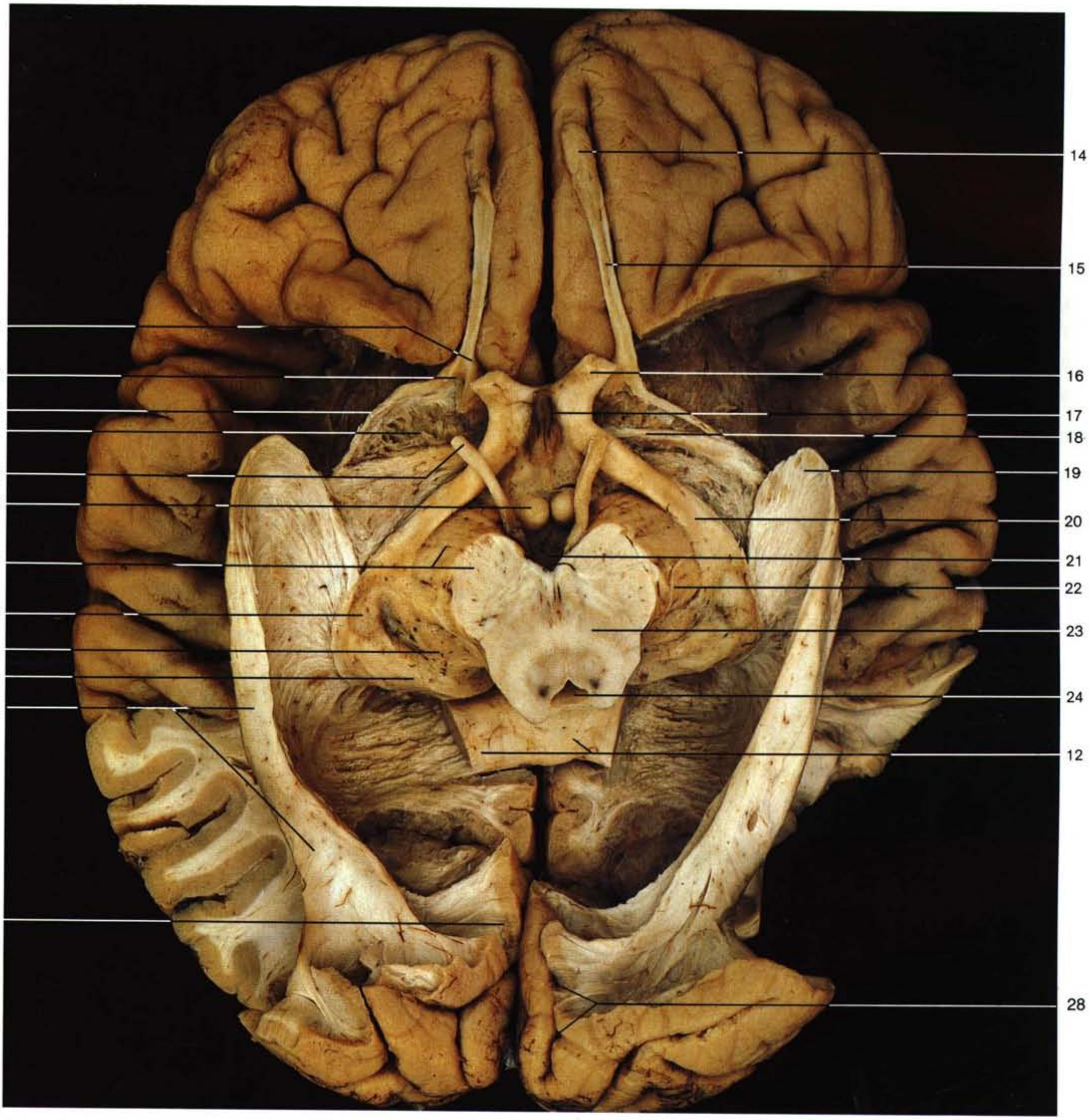

Dissection of the visual pathway (inferior aspect). Frontal pole at top, midbrain divided.

1 Medial olfactory stria
2 Olfactory trigone
3 Lateral olfactory stria
4 Anterior perforated substance
5 Oculomotor nerve (n. III)
6 Mamillary body
7 Cerebral peduncle
8 **Lateral geniculate body**
9 Medial geniculate body
10 **Pulvinar of thalamus**
11 Optic radiation
12 Splenium of the corpus callosum (commissural fibers)
13 Cuneus
14 Olfactory bulb
15 Olfactory tract
16 **Optic nerve** (n. II)
17 Infundibulum
18 Anterior commissure
19 **Genu of optic radiation**
20 **Optic tract**
21 Interpeduncular fossa and posterior perforated substance
22 Trochlear nerve (n. IV)
23 Substantia nigra
24 Cerebral aqueduct
25 **Visual cortex**
26 Line of Gennari
27 Gyrus of striate cortex
28 **Calcarine sulcus**

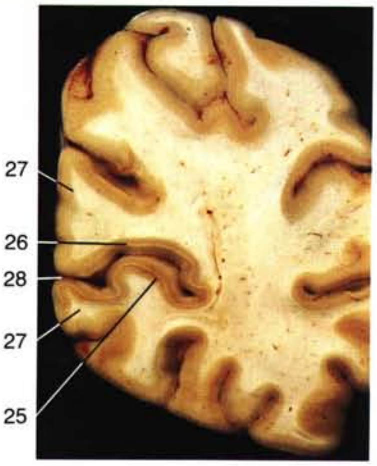

Frontal section of the striate cortex at the level of the striate area in the occipital lobe.

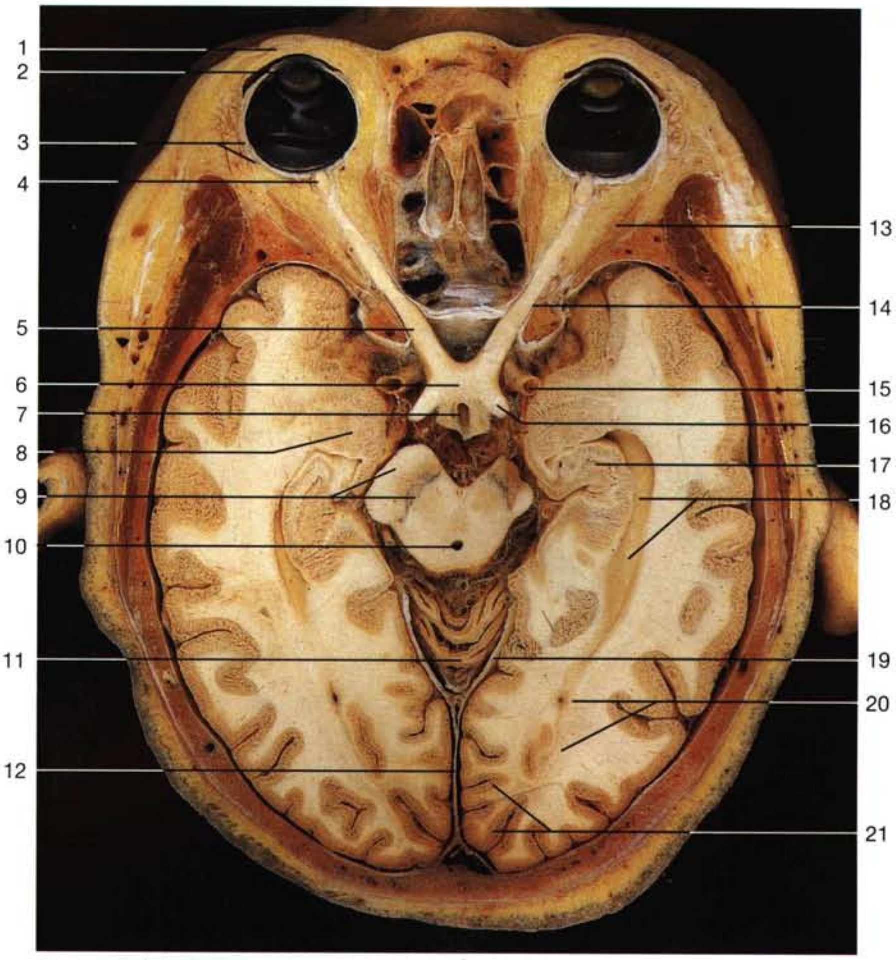

Horizontal section through the head at the level of optic chiasma and striate cortex (superior aspect). Note the relationship of hypothalamic infundibulum to optic chiasma.

1 Upper lid
2 Cornea
3 **Eyeball** (sclera, retina)
4 Head of optic nerve
5 Optic nerve
6 **Optic chiasma**
7 Infundibular recess of hypothalamus
8 Amygdaloid body
9 Substantia nigra and crus cerebri
10 Cerebral aqueduct
11 Vermis of cerebellum
12 Falx cerebri
13 Lateral rectus muscle
14 Optic canal
15 Internal carotid artery
16 **Optic tract**
17 Hippocampus
18 Inferior horn of lateral ventricle
19 Tentorium cerebelli
20 **Optic radiation** of Gratiolet
21 **Visual cortex (area calcarina, striate cortex)**
22 Lens
23 **Eyeball**
24 Ethmoidal cells
25 Optic nerve with dura sheath
26 Cerebral peduncle
27 Aqueduct of mesencephalon
28 Vermis of cerebellum

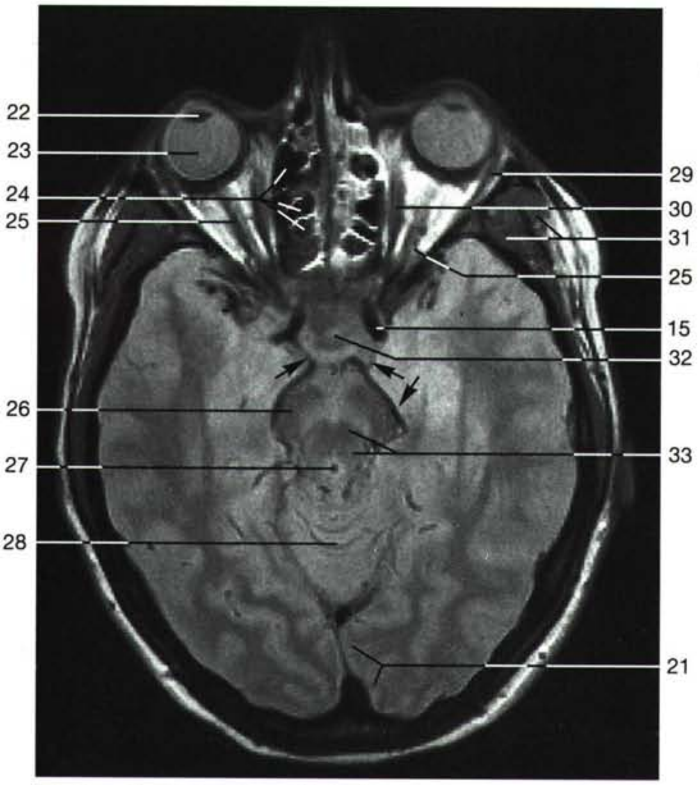

Horizontal section through the human head (MR-Scan, courtesy of Prof. W. J. Huk, Erlangen, FRG). Arrows = branches of arterial circle of Willis.

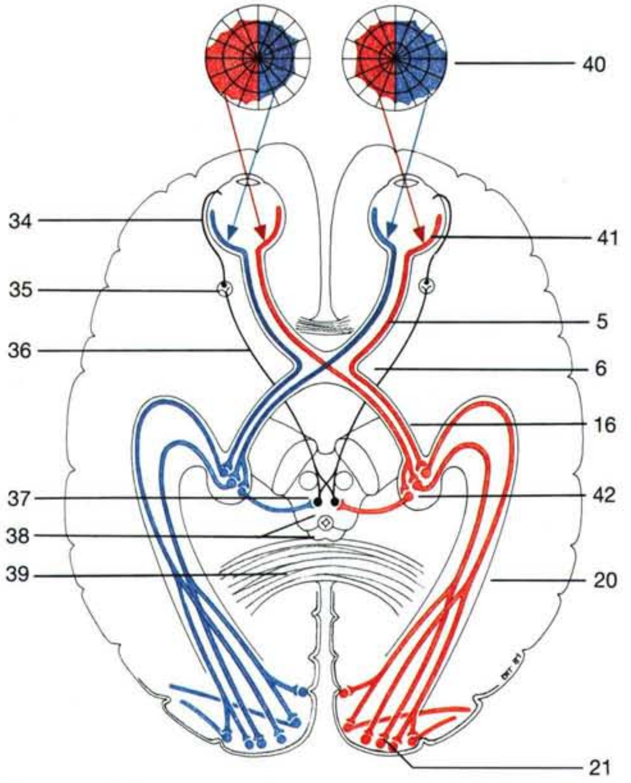

Diagram of the visual pathway and path of the light reflex.

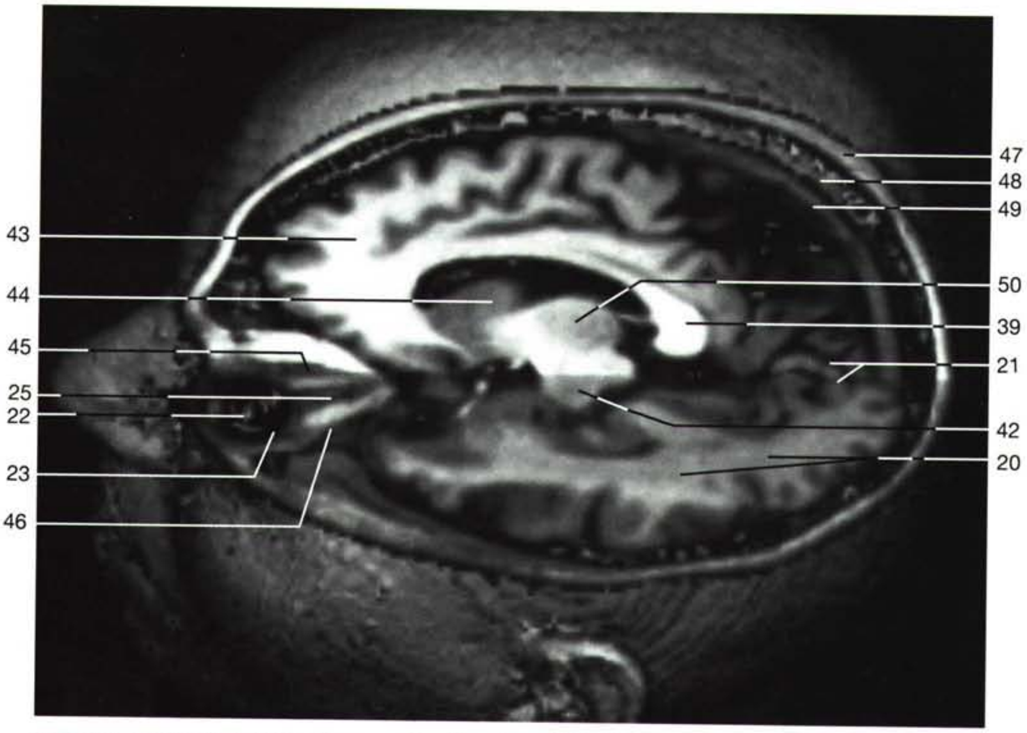

3-D-Reconstruction of the human visual system (MR-Image flash 40°, courtesy of Prof. W. J. Huk, University of Erlangen, FRG).

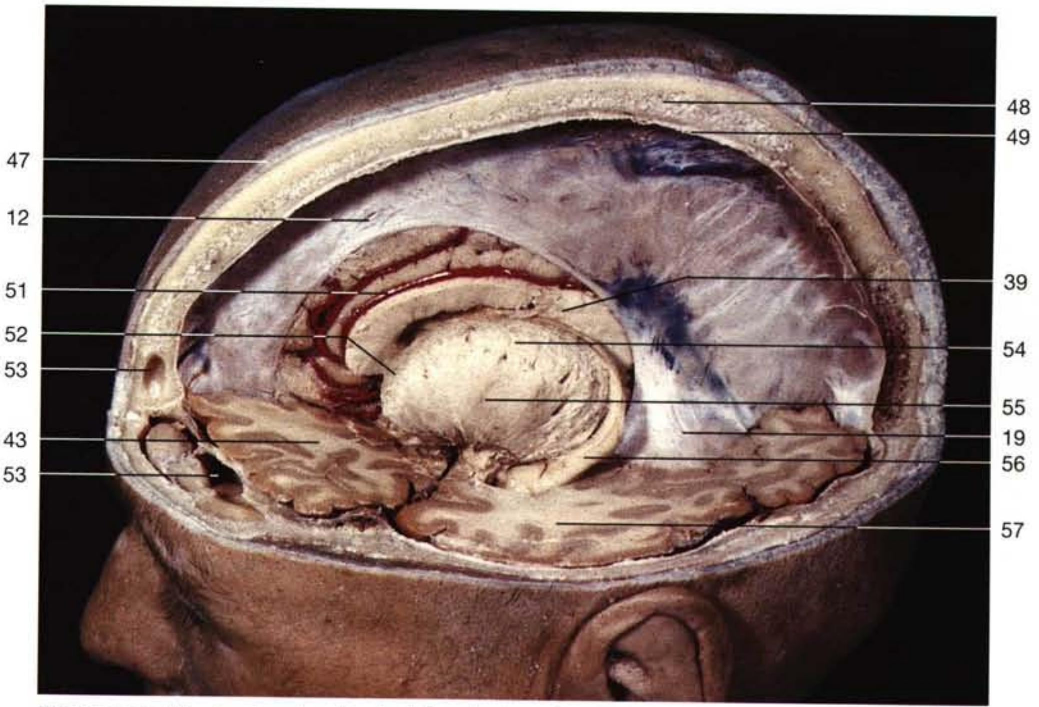

Dissection of brain stem in situ. Left hemisphere has been partly removed (compare with MR-Image above).

29	Lateral rectus muscle
30	Medial rectus muscle
31	Temporalis muscle
32	Hypophysis (pituitary gland)
33	Midbrain
34	Ciliary nerves (long and short)
35	Ciliary ganglion
36	Oculomotor nerve
37	Accessory oculomotor nucleus
38	Colliculi of midbrain
39	Corpus callosum
40	Visual field
41	Retina
42	**Lateral geniculate body**
43	Frontal lobe
44	**Caudate nucleus**
45	Medial rectus muscle
46	Lateral rectus muscle
47	Skin
48	Diploe (skull)
49	Dura mater
50	**Thalamus**
51	**Anterior cerebral artery**
52	Nucleus caudatus
53	Frontal sinus
54	**Internal capsule**
55	Lentiform nucleus (putamen)
56	**Hippocampus**
57	Temporal lobe of left hemisphere

In **binocular vision** the visual field (40) is projected upon portions of both retinae (blue and red in the drawing). In the chiasma the fibers from the two retinal portions are combined to form the left optic tract. The fibers of the two eyes remain separated from each other throughout the entire visual pathway up to their final termination in the calcarine cortex (21). **Injuries** on the optic pathway produce visual defects whose nature depends on the location of the injury. Destruction of one optic nerve produces blindness in the corresponding eye with loss of pupillary light reflex. If lesions of the **chiasma** destroy the crossing fibers of the nasal portions of the retina, both temporal fields of vision are lost (bitemporal hemianopsia). If both lateral angles of the chiasma are compressed, the nondecussating fibers from the temporal retinae are affected resulting in loss of nasal visual fields (binasal hemianopsia). Lesions posterior to the chiasma (i. e., optic tract, lateral geniculate body, optic radiation or visual cortex) result in a loss of the entire opposite field of vision (homonymous hemianopsia).

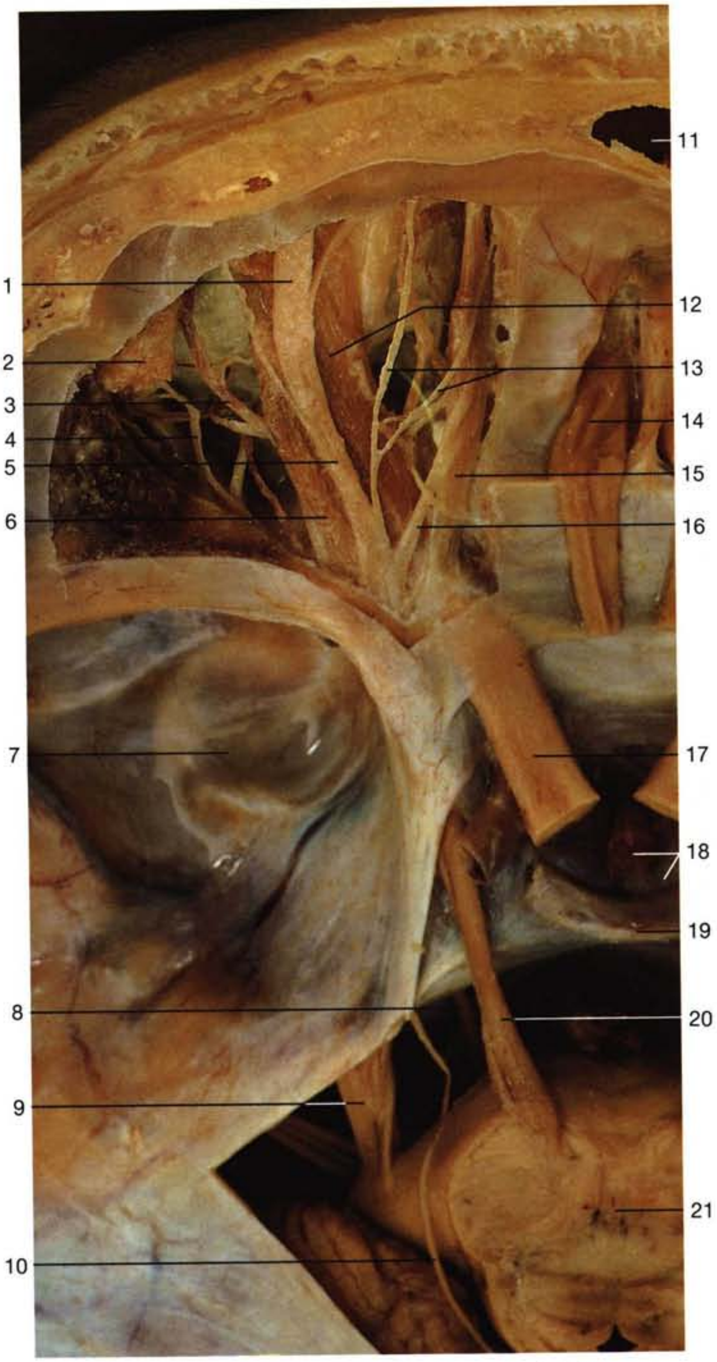

Superficial layer of the left orbit (superior aspect). The roof of the orbit and a portion of the left tentorium have been removed.

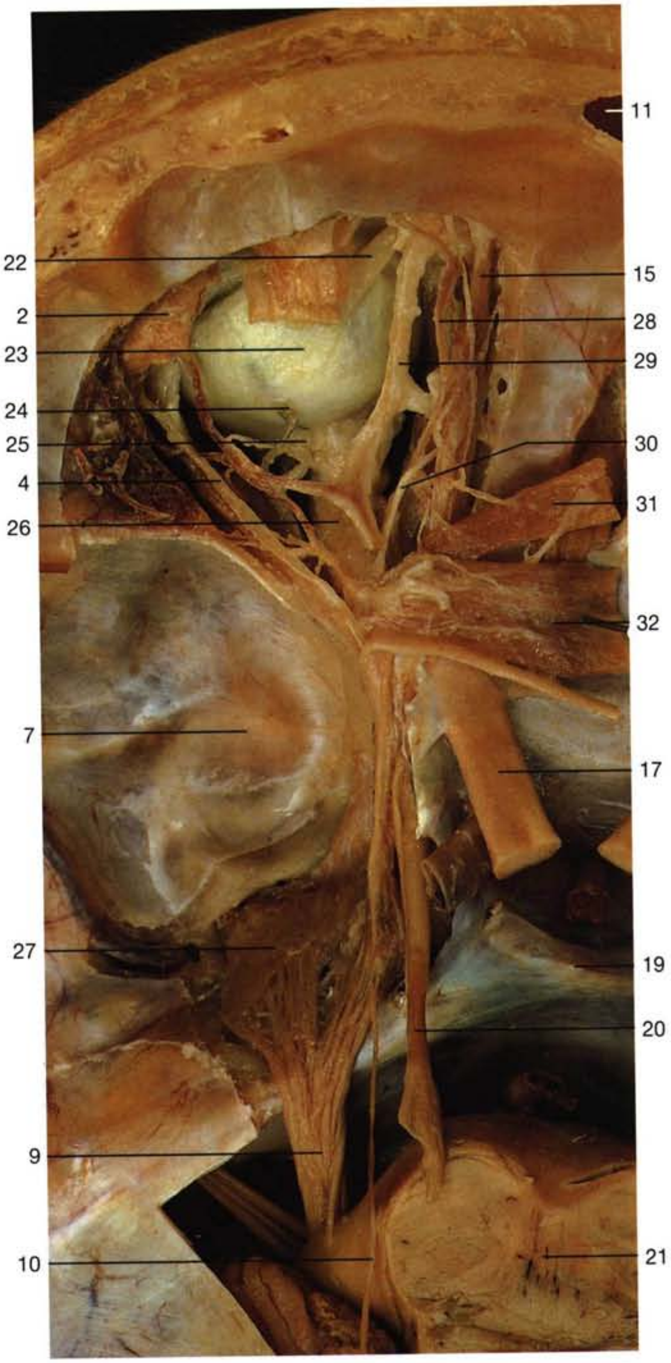

Middle layer of the left orbit (superior aspect). The roof of the orbit has been removed and the superior extraocular muscles have been divided and reflected.

1 Lateral branch of frontal nerve
2 Lacrimal gland
3 Lacrimal vein
4 Lacrimal nerve
5 **Frontal nerve**
6 Superior rectus
7 Middle cranial fossa
8 **Abducens nerve** (n. VI)
9 Trigeminal nerve (n. V)
10 **Trochlear nerve** (intracranial part) (n. IV)
11 Frontal sinus
12 Levator palpebrae superioris muscle
13 Branches of supratrochlear nerve
14 Olfactory bulb
15 Superior oblique muscle
16 **Trochlear nerve** (intraorbital part) (n. IV)
17 **Optic nerve** (intracranial part)
18 Pituitary gland and infundibulum
19 Dorsum sellae
20 **Oculomotor nerve** (n. III)
21 Midbrain
22 Tendon of superior oblique muscle
23 Eyeball
24 Vena vorticosa
25 Short ciliary nerves
26 Optic nerve (extracranial part)
27 **Trigeminal ganglion**

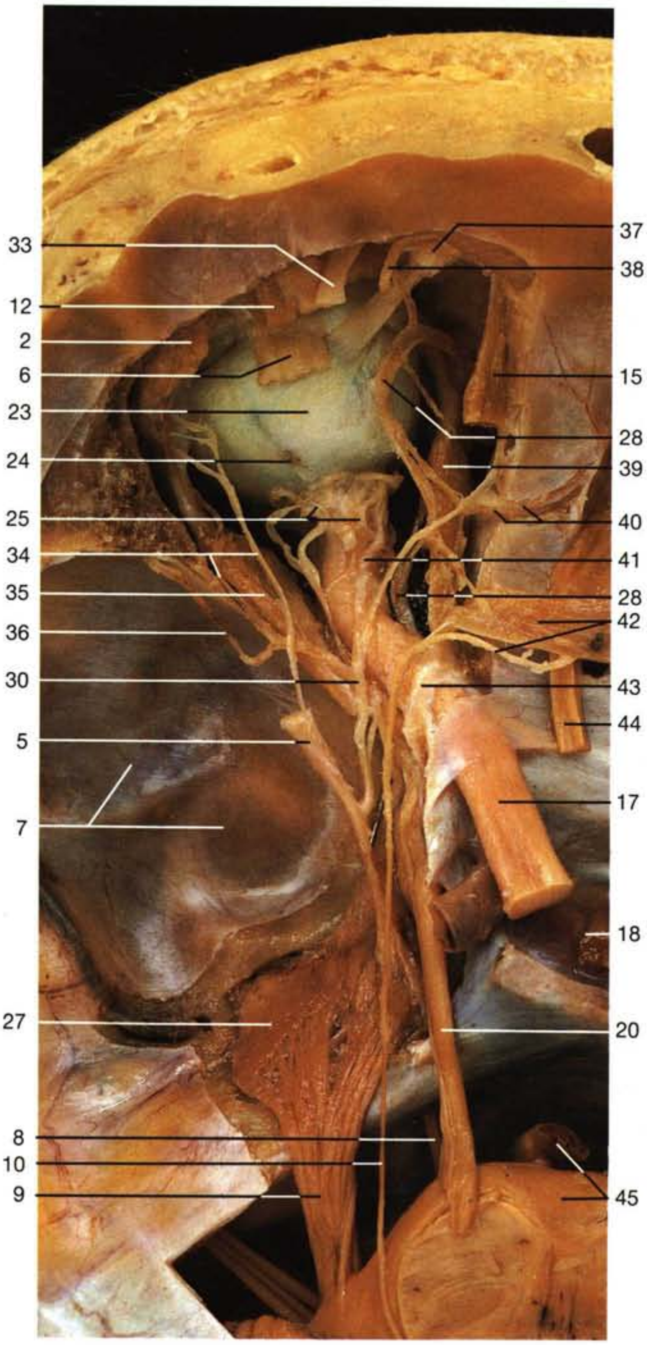
33
12
2
6
23
24
25
34
35
36
30
5
7
27
8
10
9
37
38
15
28
39
40
41
28
42
43
44
17
18
20
45

Middle layer of the left orbit (superior aspect). The roof of the orbit and the superior extraocular muscles have been removed.

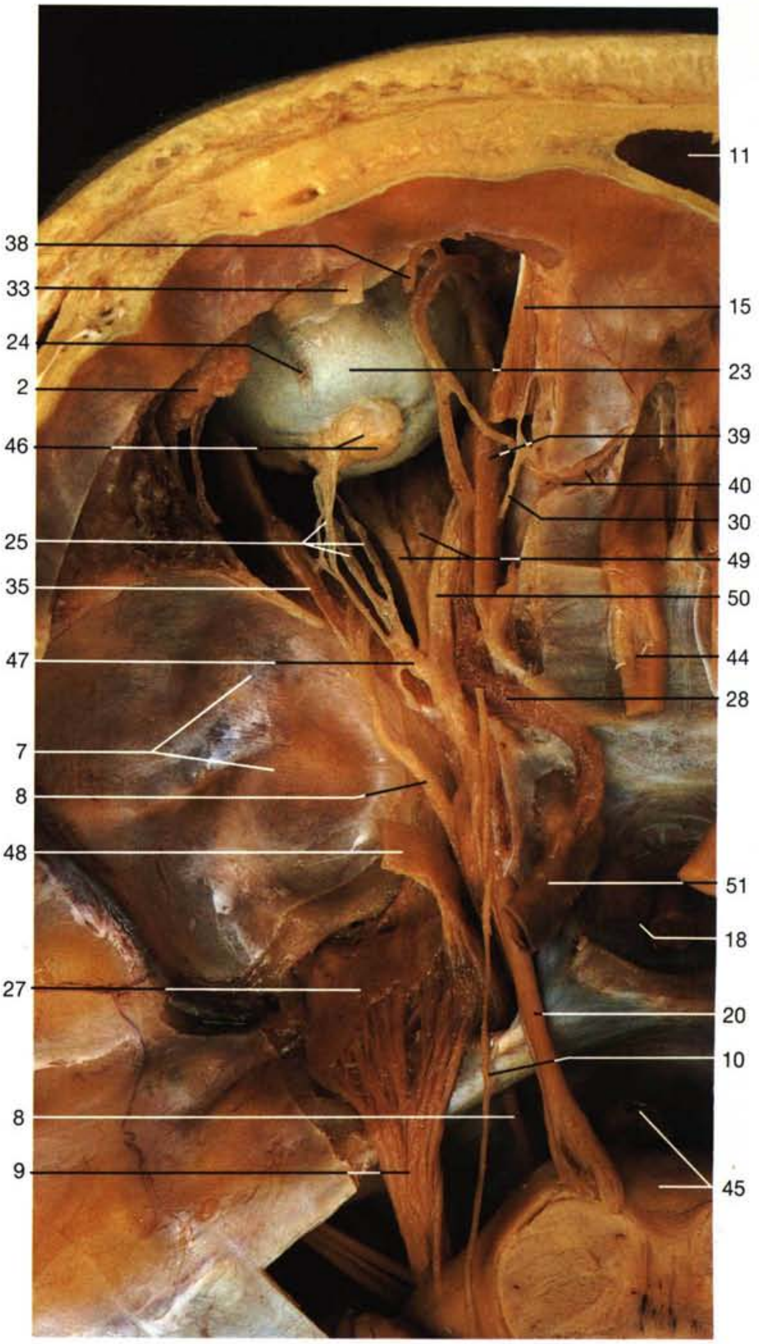
38
33
24
2
46
25
35
47
7
8
48
27
8
9
11
15
23
39
40
30
49
50
44
28
51
18
20
10
45

Deeper layer of the left orbit (superior aspect). The optic nerve has now been removed.

28 **Ophthalmic artery**
29 Superior ophthalmic vein
30 **Nasociliary nerve**
31 Levator palpebrae superioris muscle (reflected)
32 Superior rectus muscle (reflected)
33 Lateral branch of supraorbital nerve
34 Lacrimal nerve and artery
35 Lateral rectus muscle
36 Meningolacrimal artery (anastomosing with middle meningeal artery)
37 Trochlea
38 Medial branch of supraorbital nerve
39 Medial rectus muscle
40 Anterior ethmoidal artery and nerve
41 Long ciliary nerve
42 Superior oblique muscle and trochlear nerve
43 Common tendinous ring
44 Olfactory tract
45 Basilar artery and pons
46 Optic nerve (external sheath of optic nerve, divided)
47 **Ciliary ganglion**
48 Ophthalmic nerve (divided, reflected)
49 Inferior branch of oculomotor nerve and inferior rectus muscle
50 Superior branch of oculomotor nerve
51 Internal carotid artery

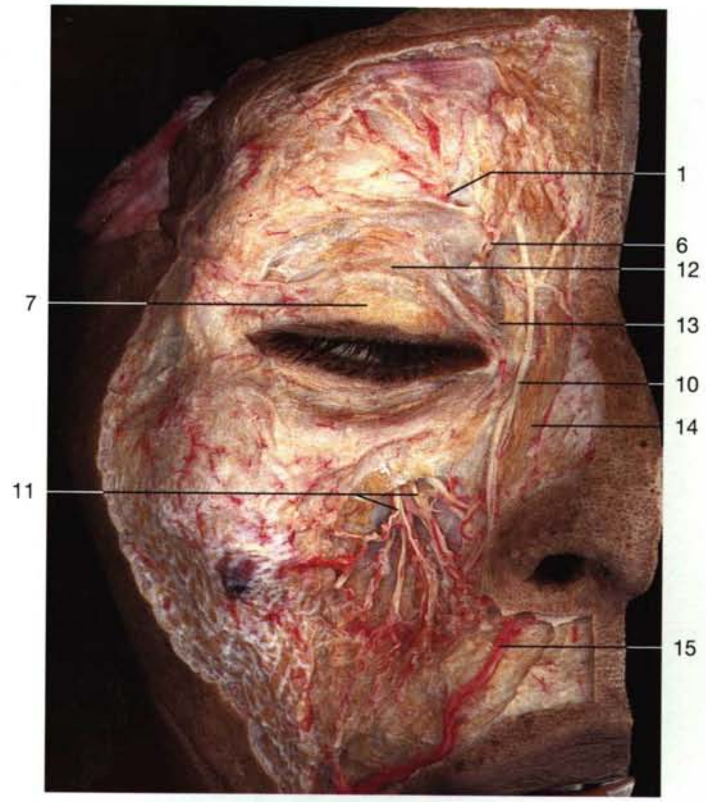

Eyelids (superficial layer, right side).

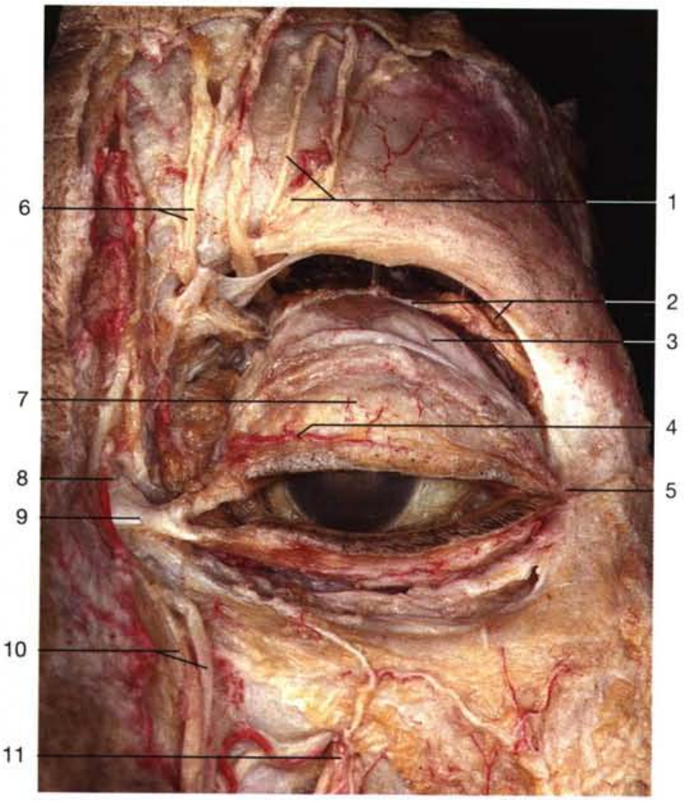

Eyelids (deeper layer, left side).

1 Supraorbital artery, supraorbital nerve, lateral branch
2 **Lacrimal gland**
3 Aponeurosis of levator palpebrae superioris muscle
4 Arterial arch of upper eyelid
5 Lateral palpebral ligament
6 Supratrochlear artery, supratrochlear nerve
7 Upper eyelid, tarsal plate
8 Lacrimal sac
9 **Medial palpebral ligament**
10 Angular artery and vein
11 Infraorbital artery, vein and nerve
12 Orbicularis oculi muscle
13 Infratrochlear nerve
14 Levator labii superioris alaeque nasi muscle
15 Facial artery and vein
16 **Superior lacrimal canaliculus**
17 **Inferior lacrimal canaliculus**
18 Lacrimal papilla and punctum
19 Lacrimal bone
20 **Lacrimal duct**
21 Mucous membrane of nasal cavity
22 Palpebral conjunctiva of lower eyelid
23 Lacrimal sac and superior lacrimal canaliculus
24 Lateral fixation of levator aponeurosis
25 Infraorbital foramen

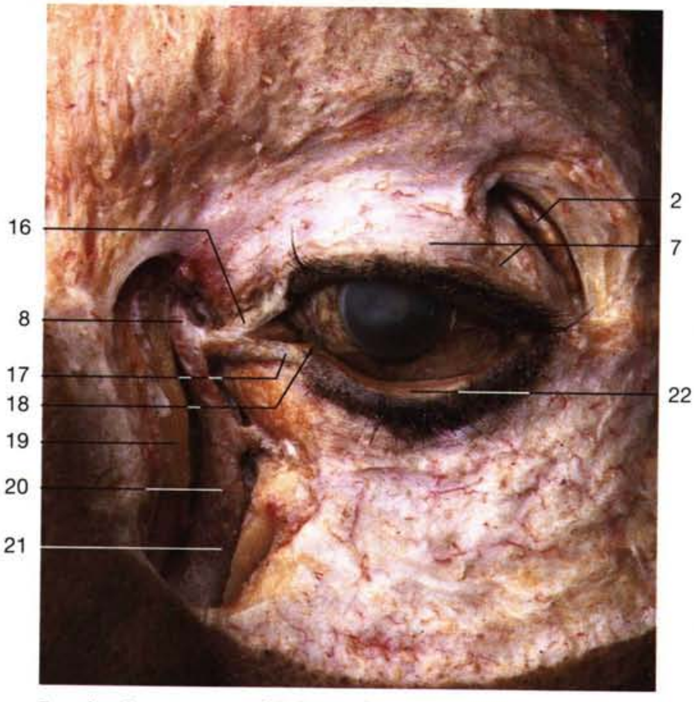

Lacrimal apparatus of left eye (anterior aspect).

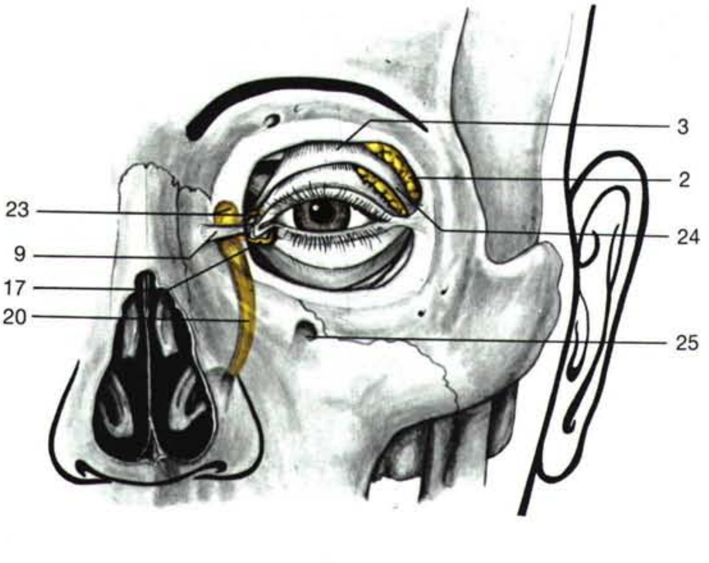

Lacrimal apparatus of left eye (anterior aspect).

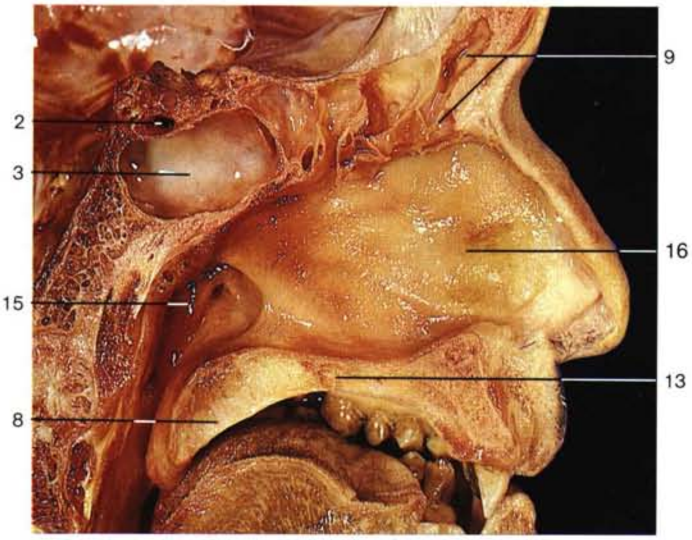

Nasal septum, covered by a mucous membrane.

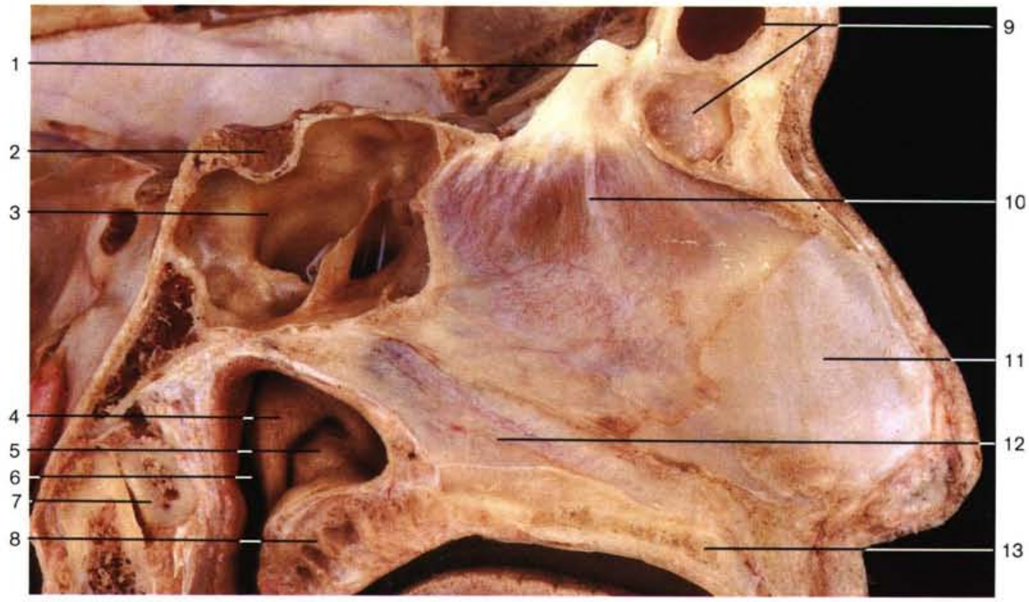

Nasal septum. Mucous membrane removed.

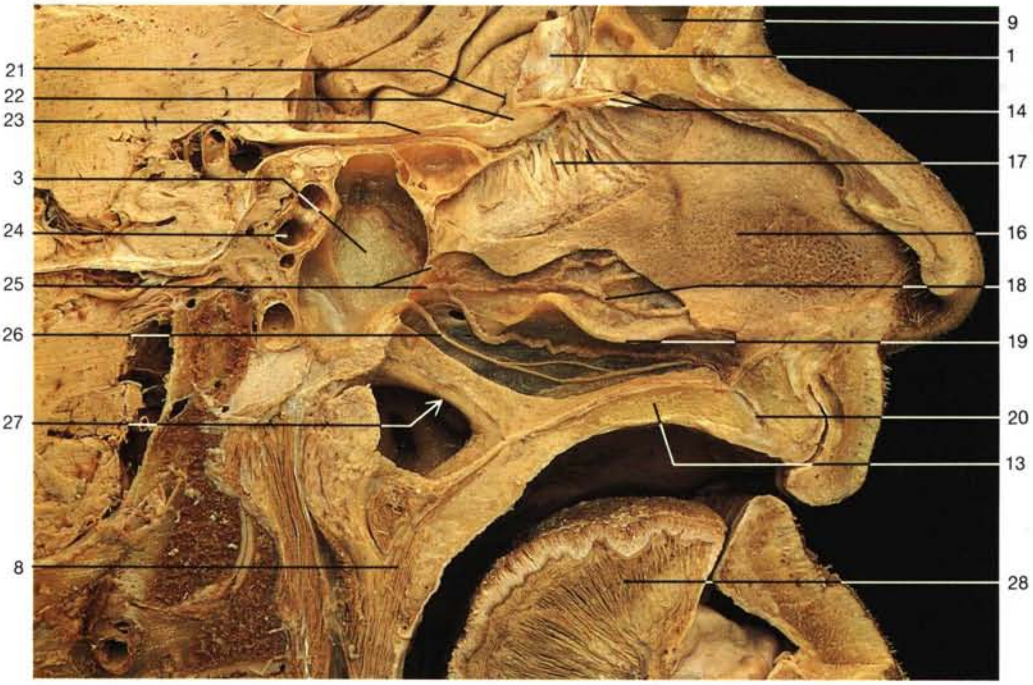

Nasal septum. Dissection of nerves and vessels.

1 Crista galli
2 Pituitary gland and sella turcica
3 **Sphenoidal sinus** (relatively large)
4 Tubal elevation
5 Pharyngeal opening of the **auditory tube**
6 Pharyngeal recess
7 Atlas (anterior arch)
8 Soft palate
9 Frontal sinus
10 Perpendicular plate of ethmoid
11 **Cartilage of nasal septum**
12 Vomer
13 Hard palate
14 Nasal branch of anterior ethmoidal artery and anterior ethmoidal nerve
15 Nasopharynx
16 Nasal septum
17 **Olfactory nerves**
18 Septal artery
19 Crest of nasal septum
20 Incisive canal
21 Anterior ethmoidal artery
22 Olfactory bulb
23 Olfactory tract
24 Internal carotid artery
25 Posterior nasal and septal arteries
26 Nasopalatine nerve
27 **Choana** (arrow)
28 Tongue

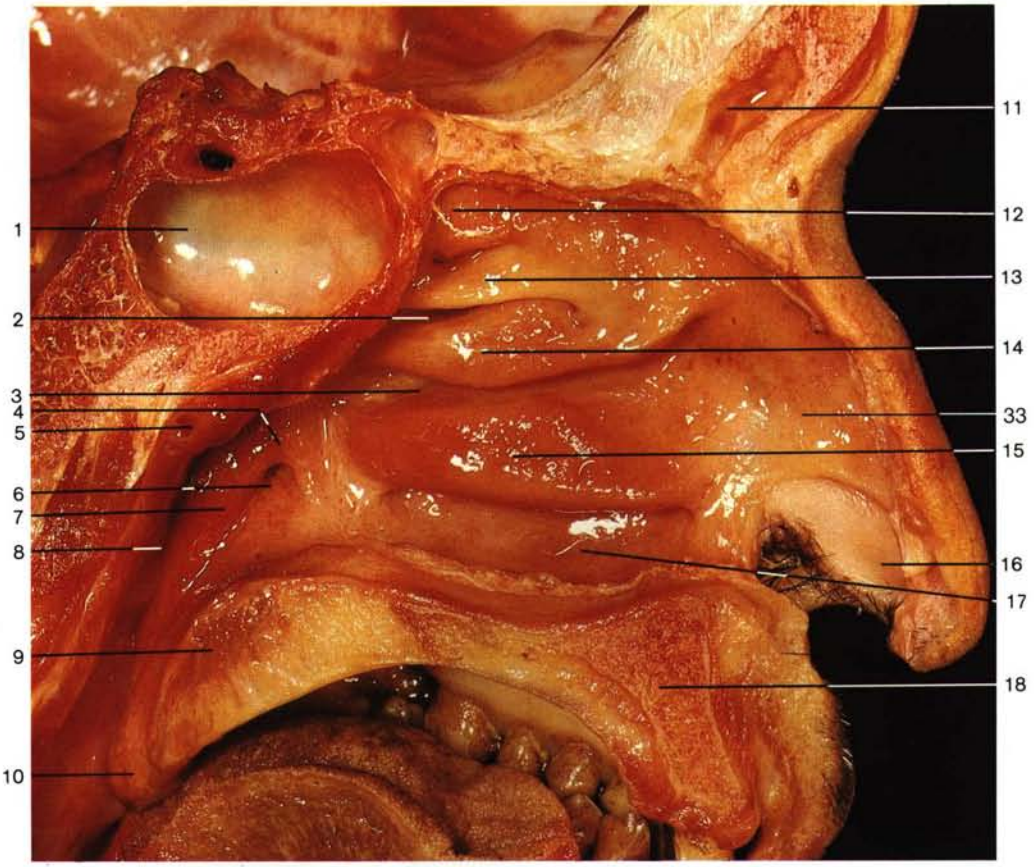

Lateral wall of the nasal cavity, septum removed.

1	Sphenoidal sinus
2	Superior meatus
3	Middle meatus
4	Tubal elevation
5	Pharyngeal tonsil
6	Pharyngeal orifice of **auditory tube**
7	Salpingopharyngeal fold
8	Pharyngeal recess
9	**Soft palate**
10	**Uvula**
11	Frontal sinus
12	Sphenoethmoidal recess
13	**Superior nasal concha**
14	**Middle nasal concha**
15	**Inferior nasal concha**
16	Vestibule
17	Inferior meatus
18	**Hard palate**
19	Grooves for the middle meningeal artery and parietal bone (light green)
20	**Maxillary hiatus**
21	Perpendicular process of palatine bone
22	Openings of ethmoidal air cells
23	Opening of frontal sinus
24	Medial pterygoid plate (red)
25	Horizontal plate of palatine process
26	**Ethmoidal air cells**
27	**Maxillary sinus**
28	Nasal septum
29	Pterygoid hamulus
30	Nasal bone (white)
31	Frontal process of maxilla (violet)
32	Palatine process of maxilla (violet)
33	Nasal atrium

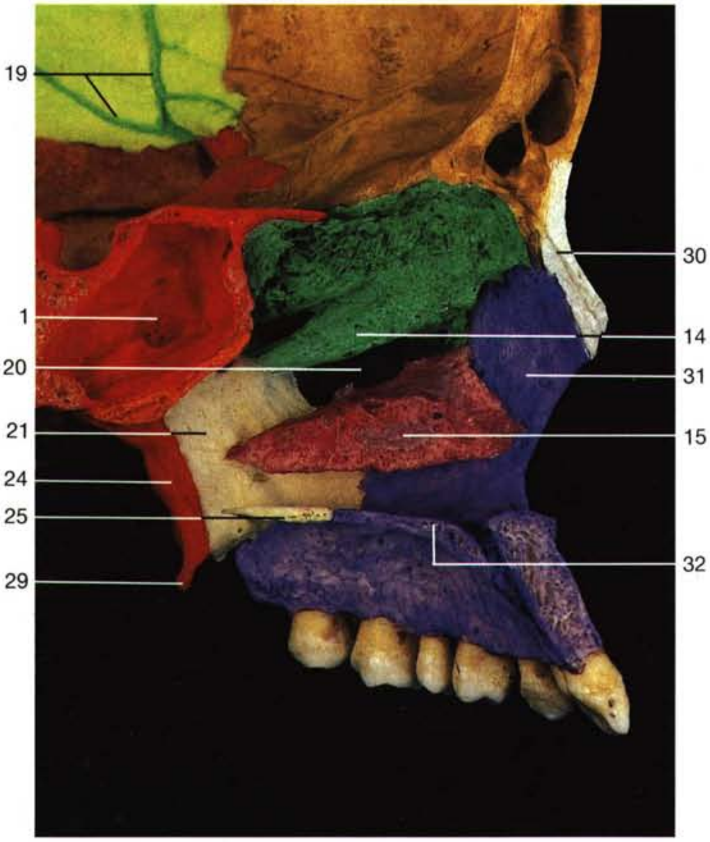

Bones of left nasal cavity, medial aspect.

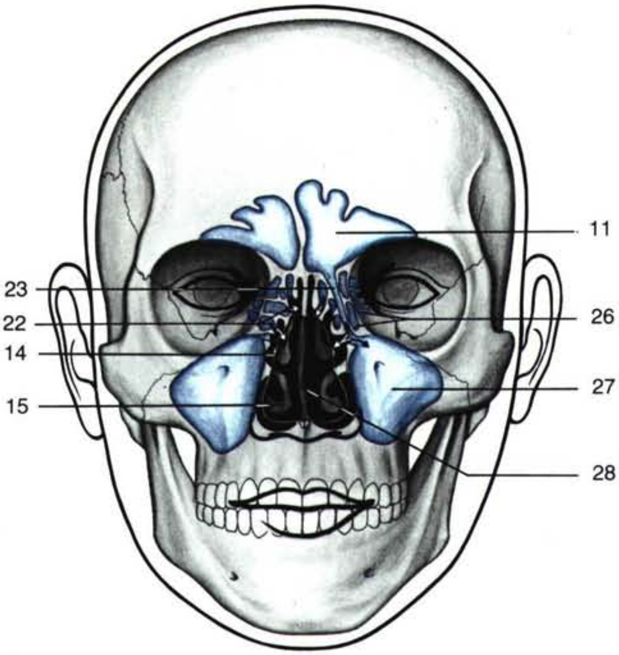

Schematic diagram showing the position of paranasal sinuses, openings indicated by arrows.

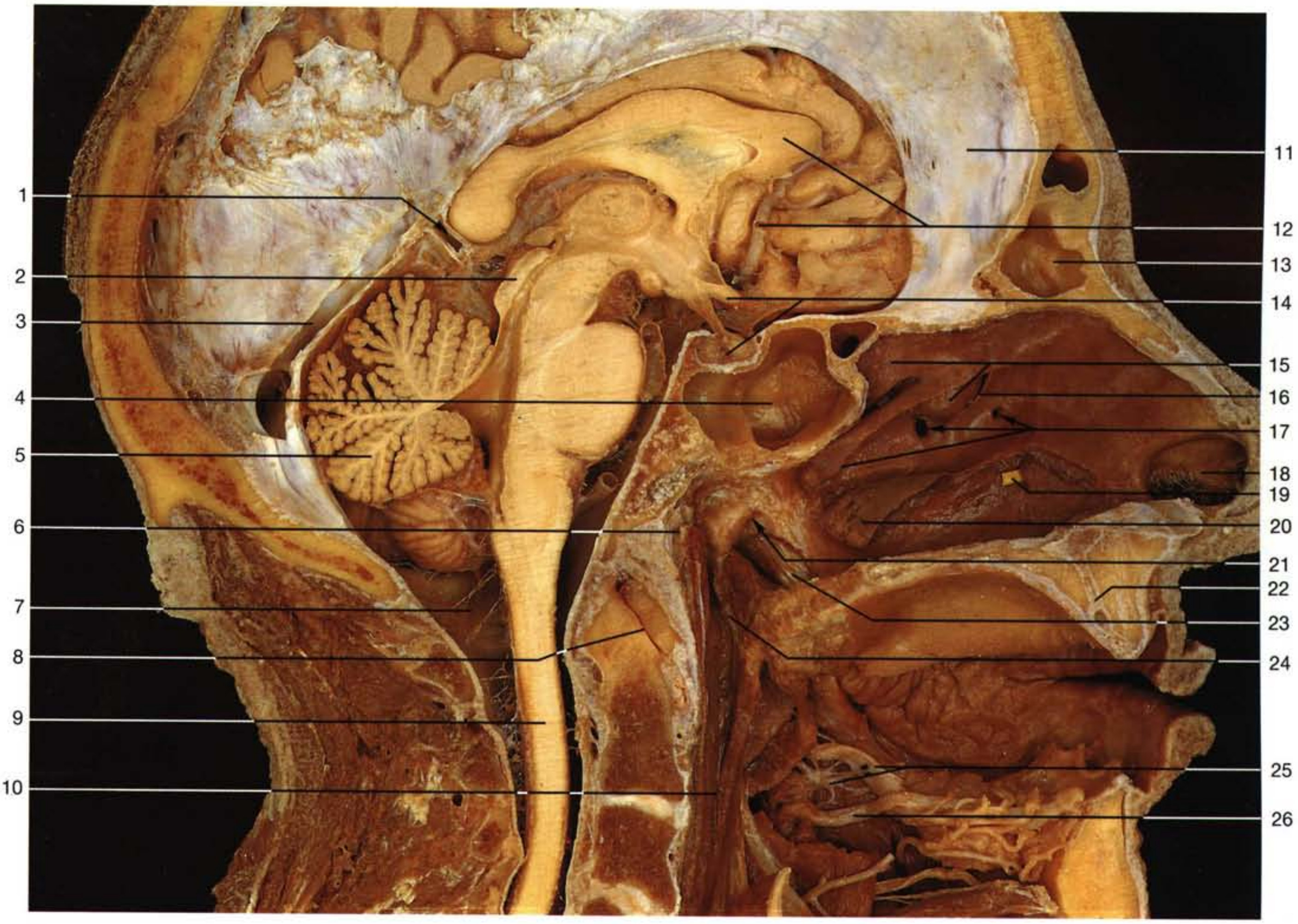

Median section through the head with nasal and oral cavity. The middle and inferior nasal conchae have been partly removed to show the openings of paranasal sinuses.

1 Great cerebral vein (Galen's vein)
2 **Tectum of midbrain**
3 Straight sinus
4 **Sphenoidal sinus**
5 Cerebellum
6 **Pharyngeal tonsil**
7 **Cerebellomedullary cistern**
8 Median atlantoaxial joint
9 Spinal cord
10 Oral part of pharynx
11 Falx cerebri
12 Corpus callosum and anterior cerebral artery
13 **Frontal sinus**
14 Optic chiasm and pituitary gland
15 Superior nasal concha and ethmoidal bulla
16 **Semilunar hiatus**
17 Accessory **openings to maxillary sinus** and cut edge of middle nasal concha
18 Vestibule
19 **Opening of nasolacrimal duct**
20 Inferior nasal concha (cut)
21 **Opening of auditory tube**
22 **Incisive canal**
23 Levator veli palatini muscle
24 Salpingopharyngeal fold
25 Lingual nerve and submandibular ganglion
26 Submandibular duct
27 **Nasofrontal duct**
28 **Nasolacrimal duct**
29 Sphenoethmoidal recess (of Rosenmüller)
30 Salpingopalatine fold

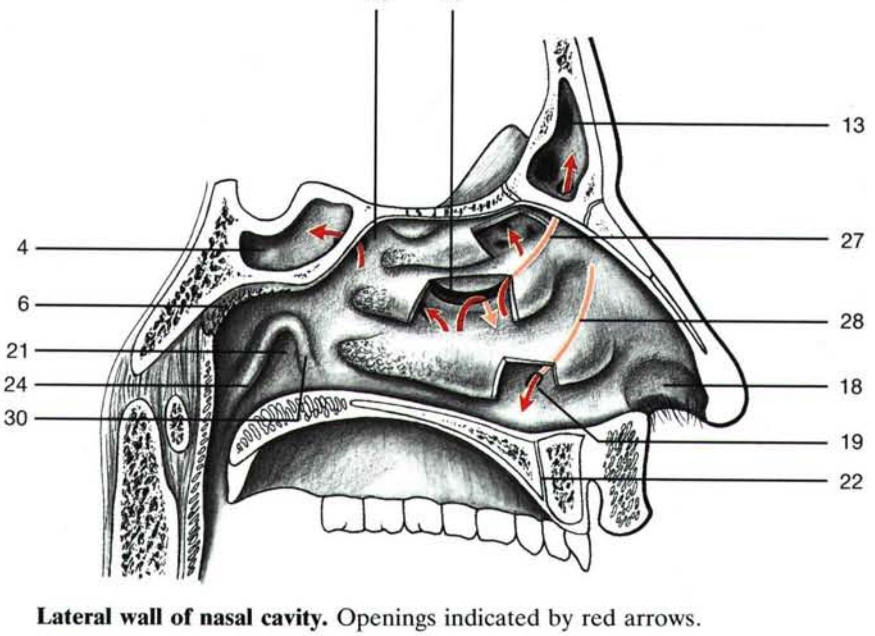

Lateral wall of nasal cavity. Openings indicated by red arrows. (Schematic drawing.)

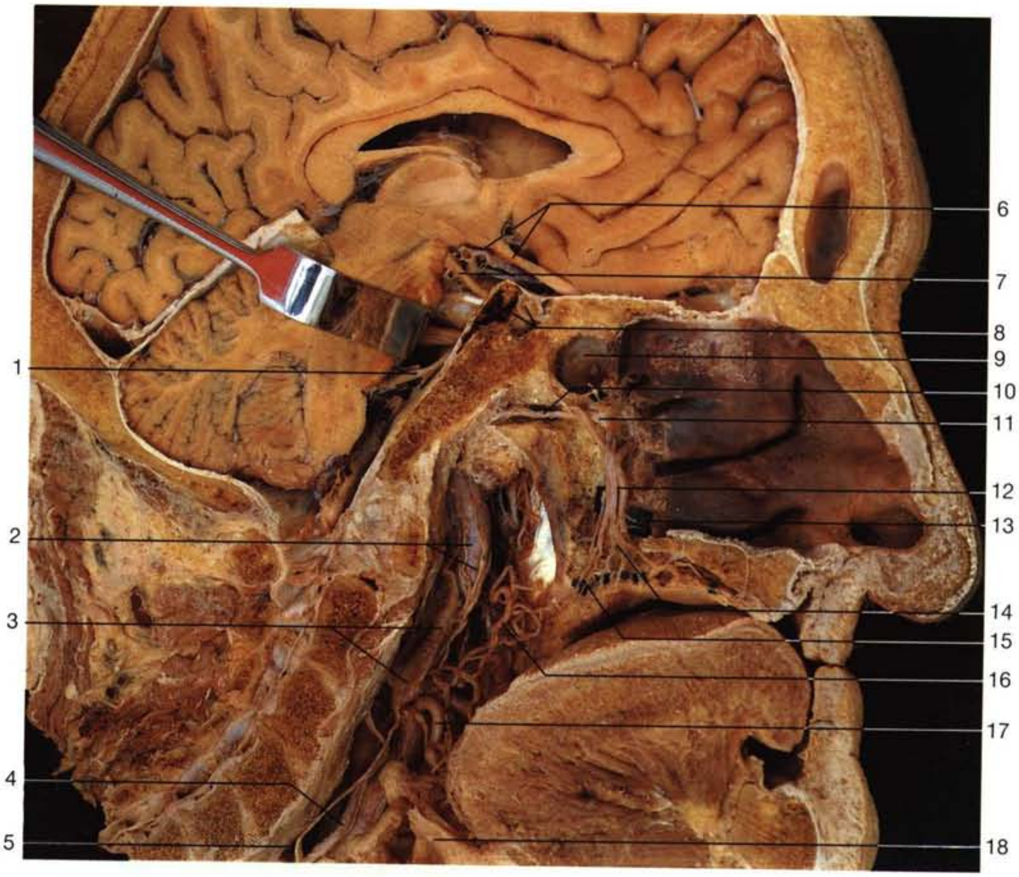

Nerves of the lateral wall of nasal cavity I. Sagittal section through the head. Mucous membranes partly removed, pterygoid canal opened.

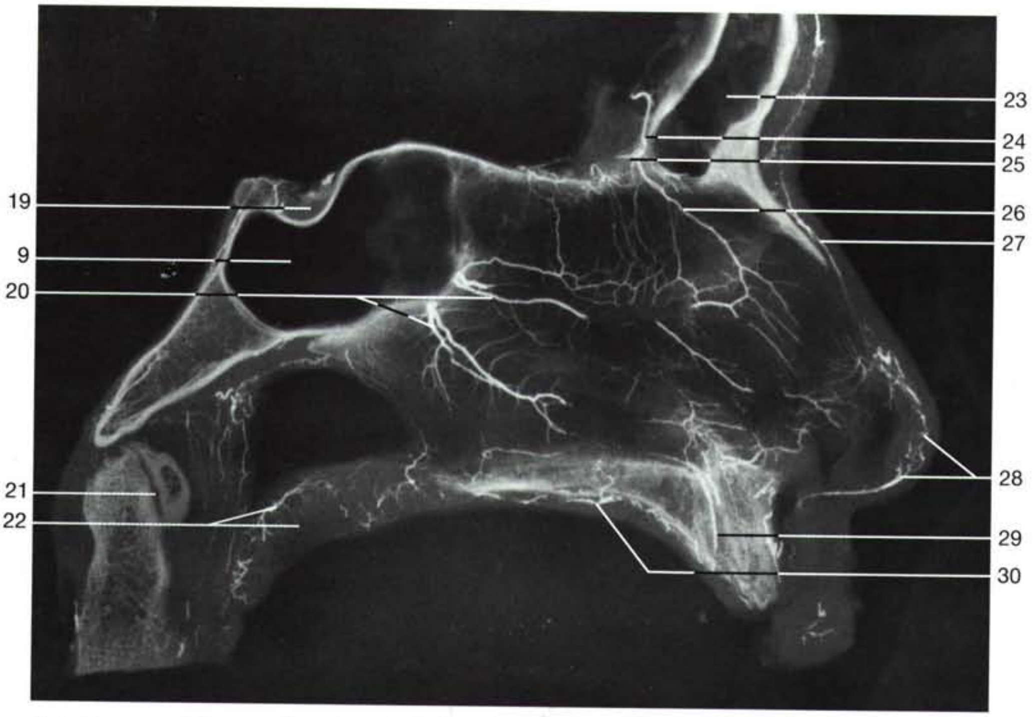

Arteriogram of the nasal septum, left side (lateral aspect).

1 Facial nerve
2 **Internal carotid artery** and **internal carotid plexus**
3 **Superior cervical ganglion**
4 Vagus nerve
5 Sympathetic trunk
6 Optic nerve and ophthalmic artery
7 Oculomotor nerve
8 Internal carotid artery and cavernous sinus
9 Sphenoidal sinus
10 Nerve of the pterygoid canal
11 **Pterygopalatine ganglion**
12 **Descending palatine artery**
13 Lateral inferior posterior nasal branches and lateral posterior nasal and septal arteries
14 Greater palatine nerves and artery
15 Lesser palatine nerves and arteries
16 Branches of ascending pharyngeal artery
17 Lingual artery
18 Epiglottis
19 **Sella turcica**
20 **Posterior lateral nasal and septal arteries** (branches of sphenopalatine artery)
21 Median atlanto-axial joint
22 Soft palate and lesser palatine arteries (branches of descending palatine artery)
23 Frontal sinus
24 Anterior meningeal artery
25 Anterior ethmoidal artery (branch of ophthalmic artery)
26 **Septal branch of anterior ethmoidal artery**
27 Dorsal nasal artery
28 Nasal branches of anterior ethmoidal artery
29 Incisive canal with nasopalatine artery
30 Hard palate and greater palatine artery (branch of descending palatine artery)
31 Tentorium cerebelli
32 Trochlear nerve
33 **Trigeminal nerve** with motor root
34 **Internal carotid plexus**
35 **Lingual nerve with chorda tympani**
36 Medial pterygoid muscle and medial pterygoid plate
37 **Inferior alveolar nerve**
38 **Sympathetic trunk**
39 Oculomotor nerve
40 **Palatine nerves**
41 Tongue
42 **Tringeminal ganglion**
43 Trigeminal nerve (n. V)
44 **Facial nerve** (n. VII)
45 Geniculate ganglion
46 Stylomastoid foramen
47 Medial pterygoid muscle

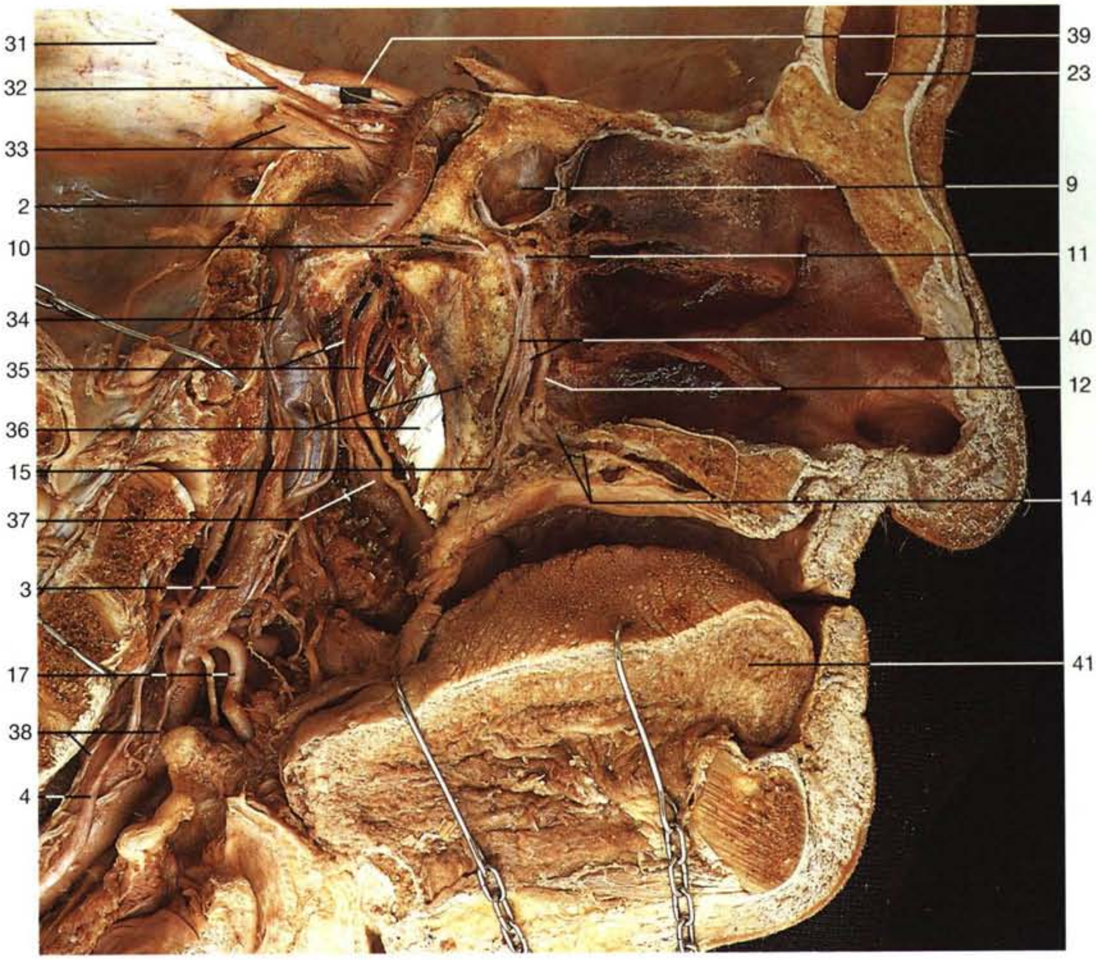
31
32
33
2
10
34
35
36
15
37
3
17
38
4
39
23
9
11
40
12
14
41

Nerves of the lateral wall of nasal cavity II. Carotid canal opened, mucous membranes of pharynx and nasal cavity partly removed.

48 **Greater petrosal nerve**
49 Maxillary nerve
50 Olfactory bulb
51 **Olfactory nerves**
52 Internal nasal branches of anterior ethmoidal nerve
53 Lateral superior posterior nasal branches
54 Lateral inferior posterior nasal branches
55 Incisive canal with nasopalatine nerve
56 **Greater palatine nerve**
57 **Deep petrosal nerve**
58 Mandibular nerve
59 Nasal cavity and inferior nasal concha
60 Opening of auditory tube
61 Tensor veli palatini muscle
62 Levator veli palatini muscle
63 Pharyngeal recess in the nasopharynx
64 Uvula
65 Palatoglossal arch
66 Tonsillar branch of ascending palatine artery
67 **Palatine tonsil**
68 Palatopharyngeal arch

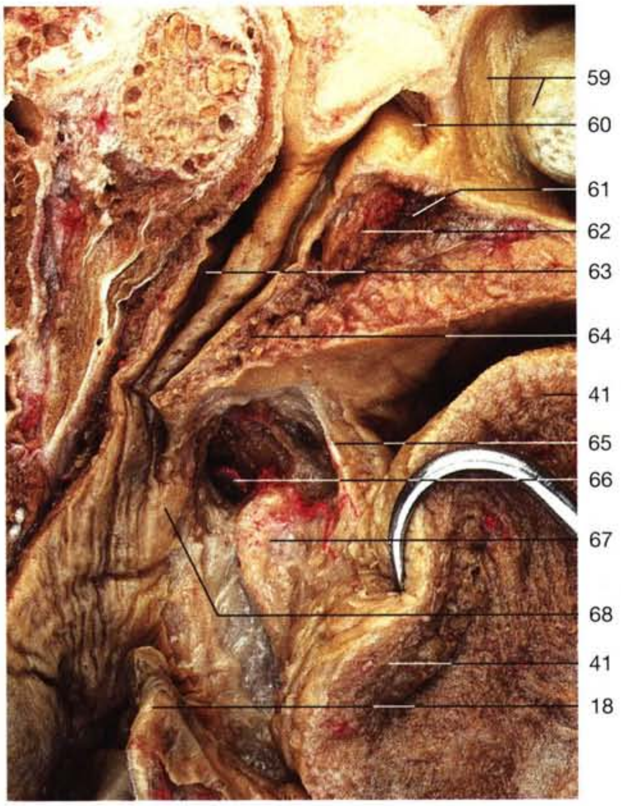
59
60
61
62
63
64
41
65
66
67
68
41
18

Dissection of palatine tonsil located in the lateral wall of the nasopharynx (left side). Root of tongue reflected.

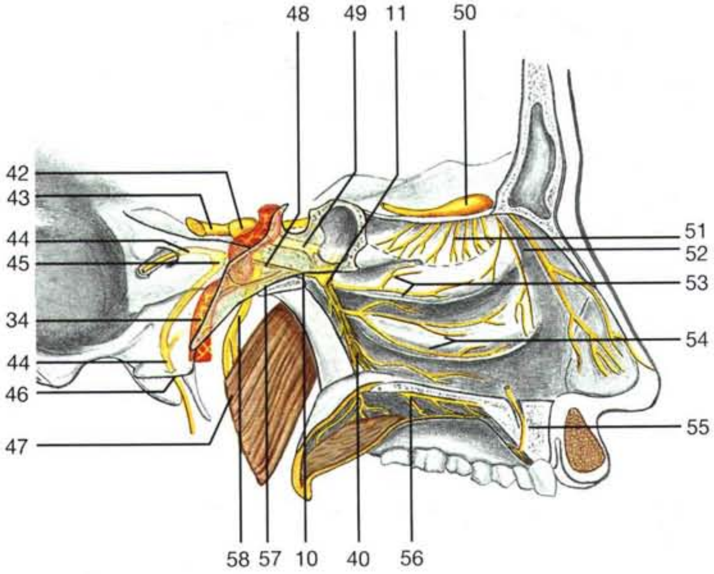
48
49
11
50
42
43
44
45
34
44
46
47
51
52
53
54
55
58
57
10
40
56

Nerves of the lateral wall of nassal cavity. Body of sphenoid bone appears transparent (schematic drawing).

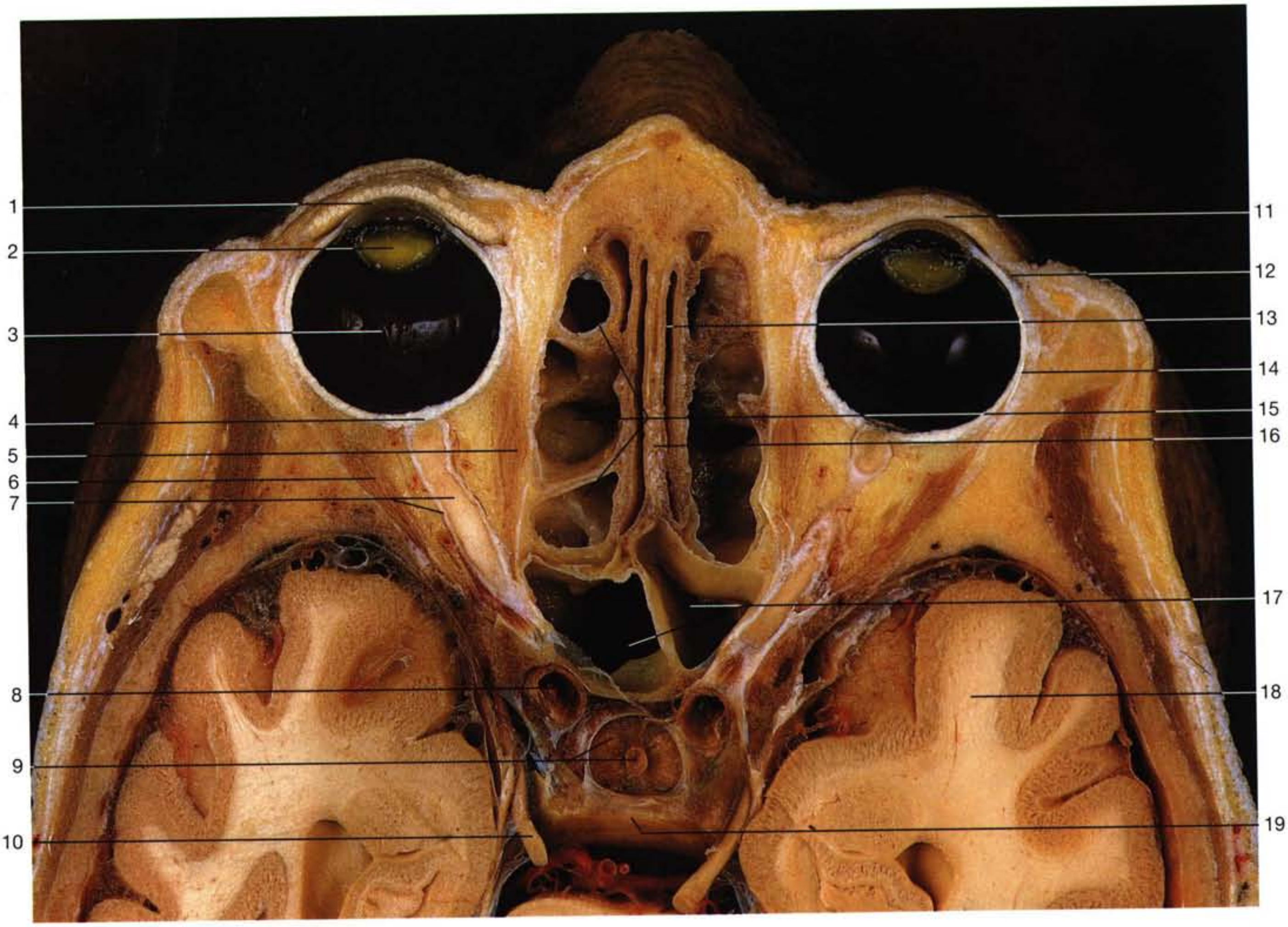

Horizontal section through the nasal cavity, the orbits and temporal lobes of the brain at the level of pituitary gland.

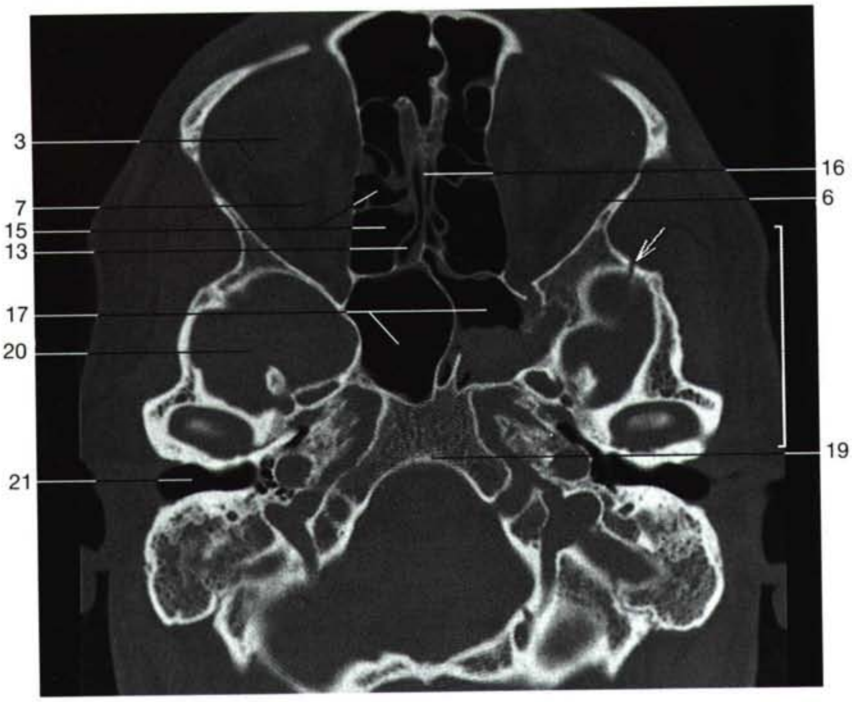

Horizontal section through the head. CT-Scan. Bar = 2 cm.
Arrow: fracture.

1 Cornea
2 Lens
3 Vitreous body (**eyeball**)
4 Head of optic nerve
5 Medial rectus muscle
6 Lateral rectus muscle
7 **Optic nerve** with dural sheath
8 Internal carotid artery
9 **Pituitary gland** and **infundibulum**
10 **Oculomotor nerve**
11 Superior tarsal plate of eye lid
12 Fornix of conjunctiva
13 **Nasal cavity**
14 Sclera
15 Ethmoidal sinus
16 **Nasal septum**
17 **Sphenoidal sinus**
18 Temporal lobe
19 Clivus
20 Middle cranial fossa
21 External acoustic meatus
22 Superior sagittal sinus
23 Falx cerebri
24 Superior rectus and levator palpebrae superioris muscles
25 Eyeball and lacrimal gland
26 Inferior rectus muscle
27 Zygomatic bone
28 **Maxillary sinus**
29 Inferior nasal concha
30 Hard palate
31 Superior longitudinal muscle of tongue
32 Median fibrosus septum of tongue

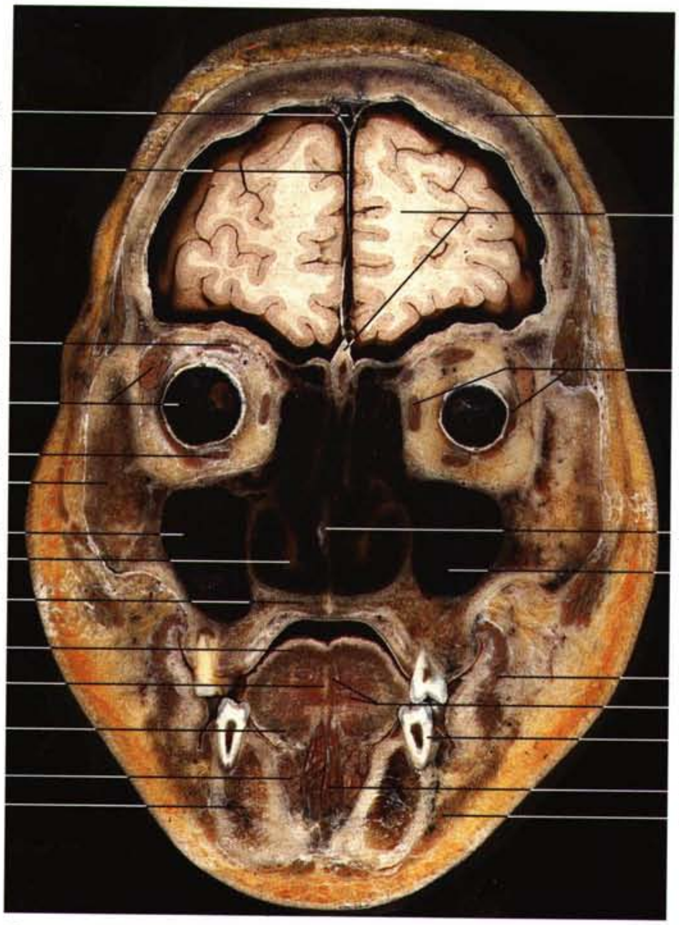
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
16
28
39
40
41
42
43

Coronal section through the head at the level of the second premolar of the mandible.

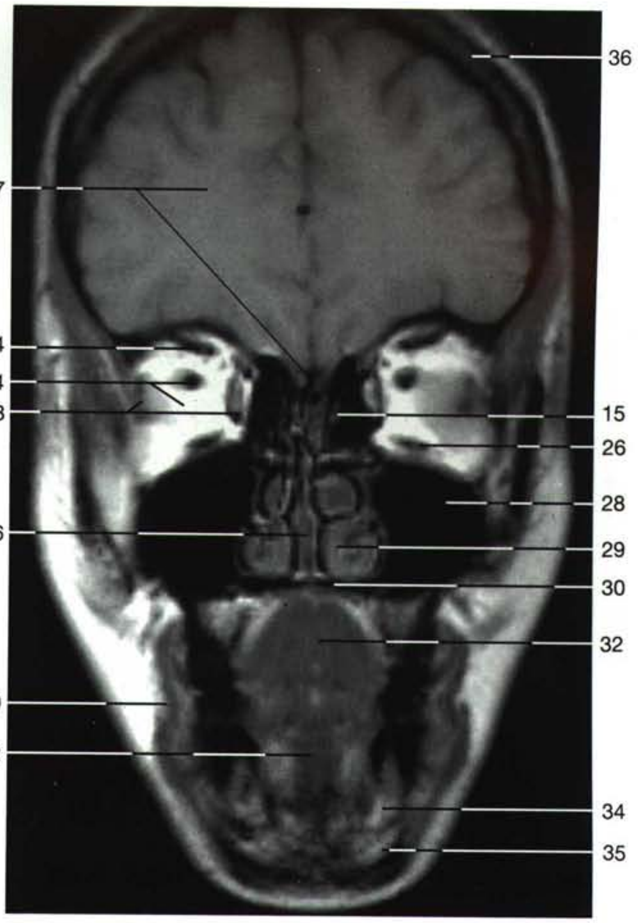
36
37
24
44
38
15
26
28
16
29
30
32
39
42
34
35

Coronal section through the head (MR-Scan, courtesy of Dr. A. Heuck, Munich). Note the situation of the head cavities.

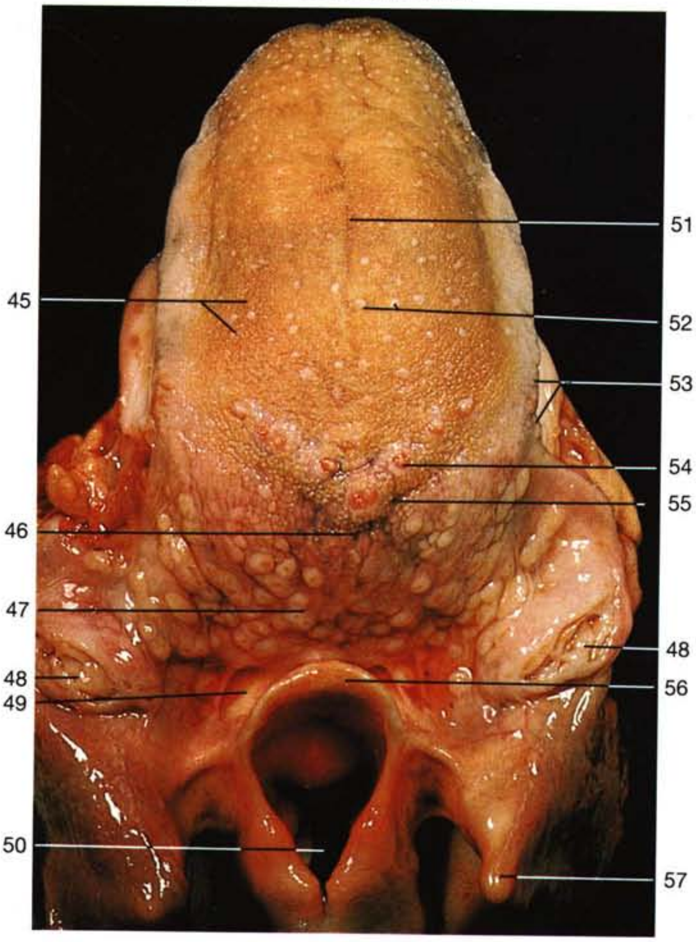
51
45
52
53
54
55
46
47
48
48
56
49
50
57

33 Inferior longitudinal muscle of tongue
34 **Sublingual gland**
35 Mandible
36 Calvaria
37 Frontal lobe of brain and crista galli
38 Lateral and medial eye rectus muscles
39 Buccinator muscle
40 Vertical and transverse muscles of tongue
41 Second premolar of mandible
42 Genioglossus muscle
43 Platysma muscle
44 **Orbit** and **optic nerve**
45 Filiform papillae
46 **Foramen cecum**
47 **Root of tongue** (lingual tonsil)
48 **Palatine tonsil**
49 Vallecula of epiglottis
50 Vestibule of larynx
51 Median sulcus of tongue
52 Fungiform papillae
53 Foliate papillae
54 **Circumvallate papilla**
55 Sulcus terminalis
56 Epiglottis
57 Greater cornu of hyoid bone

Dorsal surface of the tongue and laryngeal inlet.

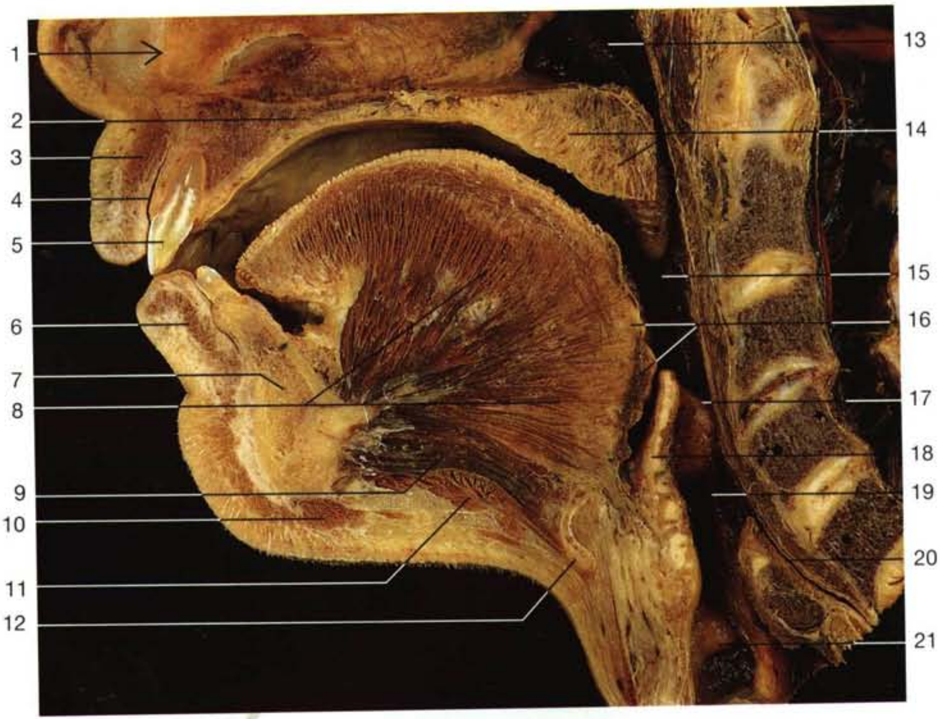

Median sagittal section through the oral cavity and pharynx.

1 Nasal cavity
2 **Hard palate**
3 Upper lip and orbicularis oris muscle
4 Vestibule of oral cavity
5 First incisor
6 Lower lip and orbicularis oris muscle
7 Mandible
8 Genioglossus muscle
9 Geniohyoid muscle
10 Anterior belly of diagastric muscle
11 Mylohyoid muscle
12 **Hyoid bone**
13 **Nasopharynx**
14 **Soft palate** and **uvula**
15 **Oropharynx**
16 Root of tongue and lingual tonsil
17 **Laryngopharynx**
18 Epiglottis
19 Aryepiglottic fold
20 Laryngopharynx continuous with esophagus
21 Larynx

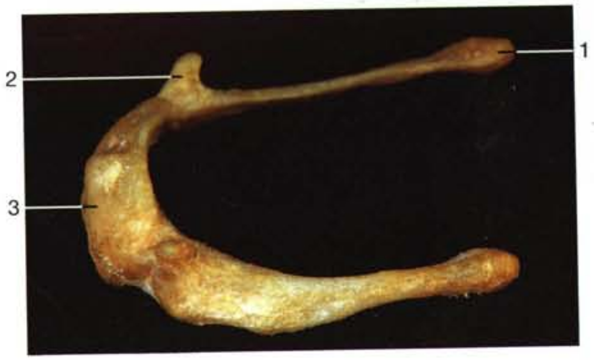

Hyoid bone (oblique lateral aspect).

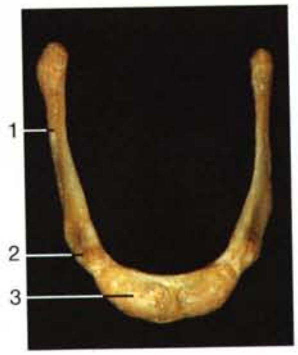

Hyoid bone (anterior aspect).

1 Greater cornu
2 Lesser cornu
3 Body
of hyoid bone

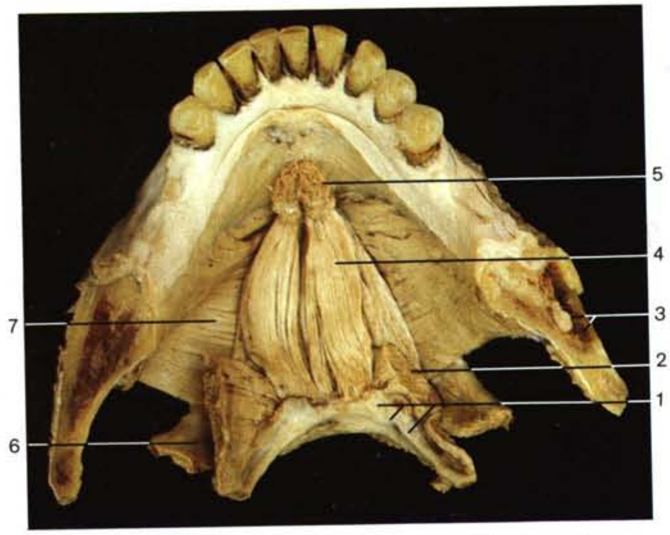

Muscles of the floor of the oral cavity (superior aspect).

1 Lesser cornu and body of hyoid bone
2 Hyoglossus muscle (divided)
3 Ramus of mandible and inferior alveolar nerve
4 **Geniohyoid muscle**
5 Genioglossus muscle (divided)
6 Stylohyoid muscle (divided)

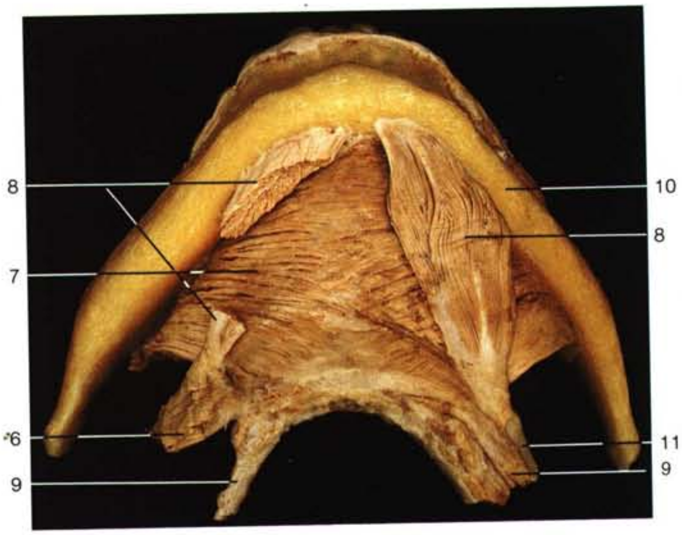

Oral diaphragm, muscles (inferior aspect). Cut on the base.

7 **Mylohyoid muscle**
8 Anterior belly of **digastric muscle**
9 Hyoid bone
10 Mandible
11 Intermediate tendon of digastric muscle

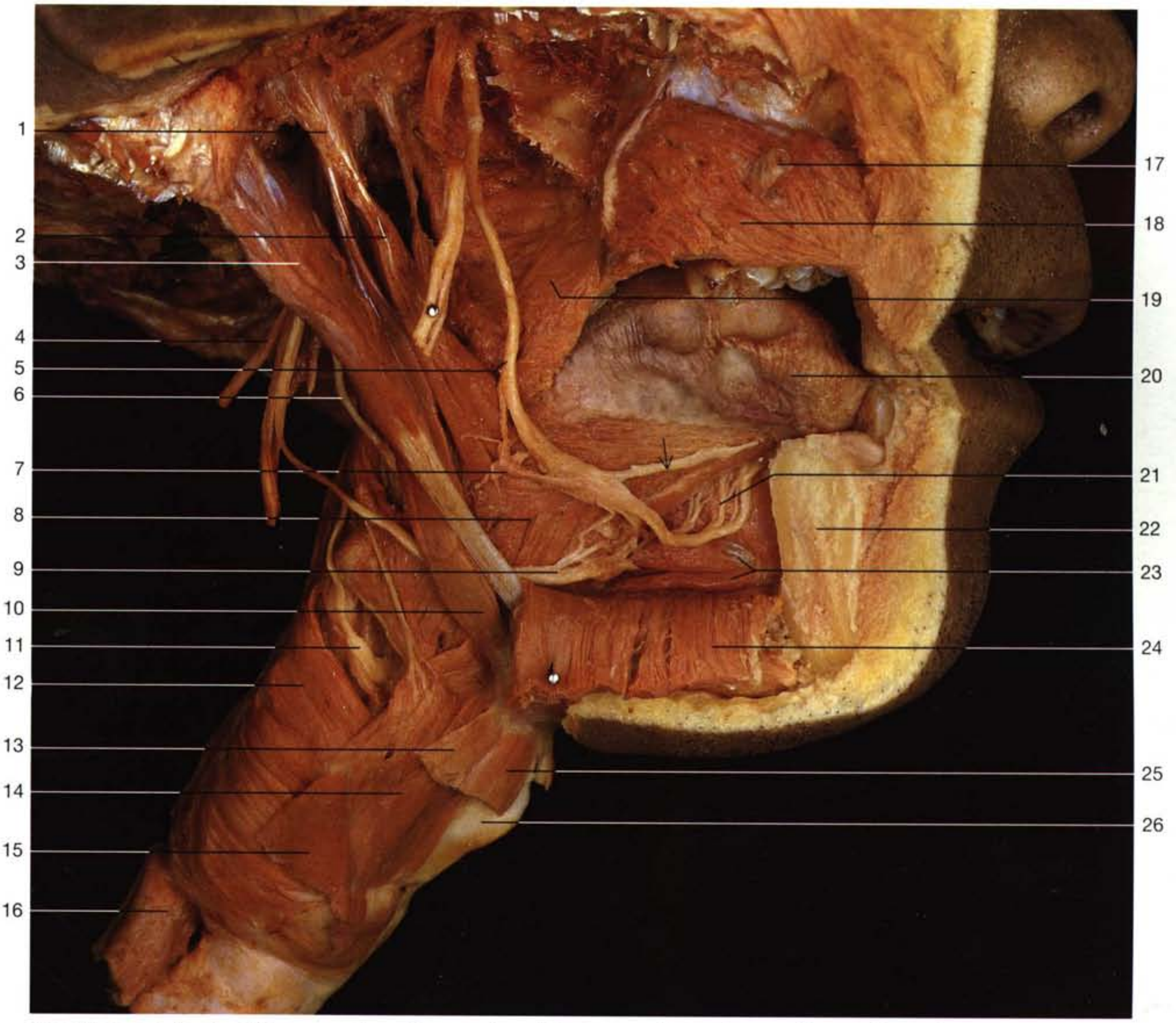

Parapharyngeal and sublingual regions. Innervation of the tongue. Lateral part of face and mandible removed, oral cavity opened. Arrow: submandibular duct.

1 Styloid process
2 Styloglossus muscle
3 Digastric muscle (posterior belly)

4 Vagus nerve (n. X)
5 **Lingual nerve** (n. V_3)
6 **Glossopharyngeal nerve** (n. IX)

7 Submandibular ganglion
8 Hyoglossus muscle
9 **Hypoglossal nerve** (n. XII)
10 Stylohyoid muscle
11 Internal branch of superior laryngeal nerve (branch of vagus nerve, not visible)
12 Middle constrictor muscle of pharynx
13 Omohyoid muscle (divided)
14 Thyrohyoid muscle
15 Sternothyroid muscle
16 Esophagus
17 Parotid duct (divided)
18 Buccinator
19 Superior constrictor muscle of pharynx
20 Tongue
21 Terminal branches of lingual nerve
22 Mandible (divided)
23 Genioglossus and geniohyoid muscles
24 Mylohyoid muscle (divided and reflected)
25 Sternohyoid muscle (divided)
26 Thyroid cartilage
27 Anterior belly of digastric muscle
28 Hyoid bone

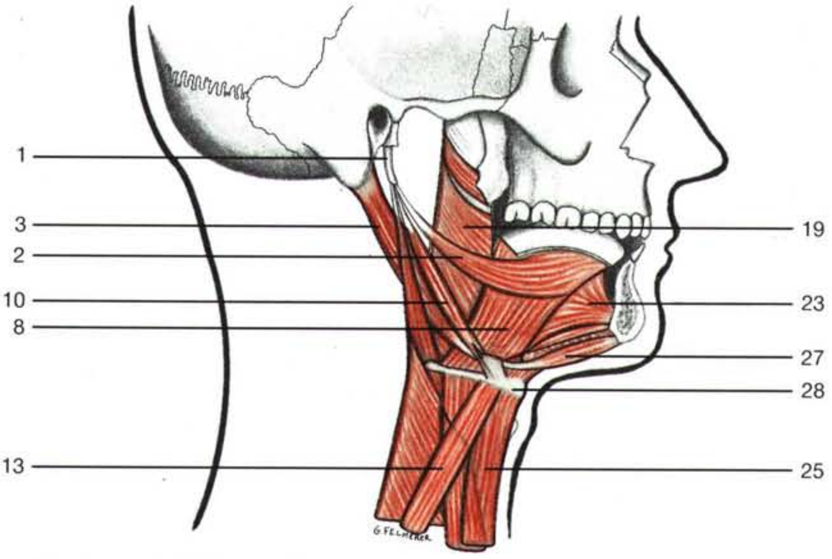

Supra- and infrahyoid muscles and pharynx (schematic drawing).

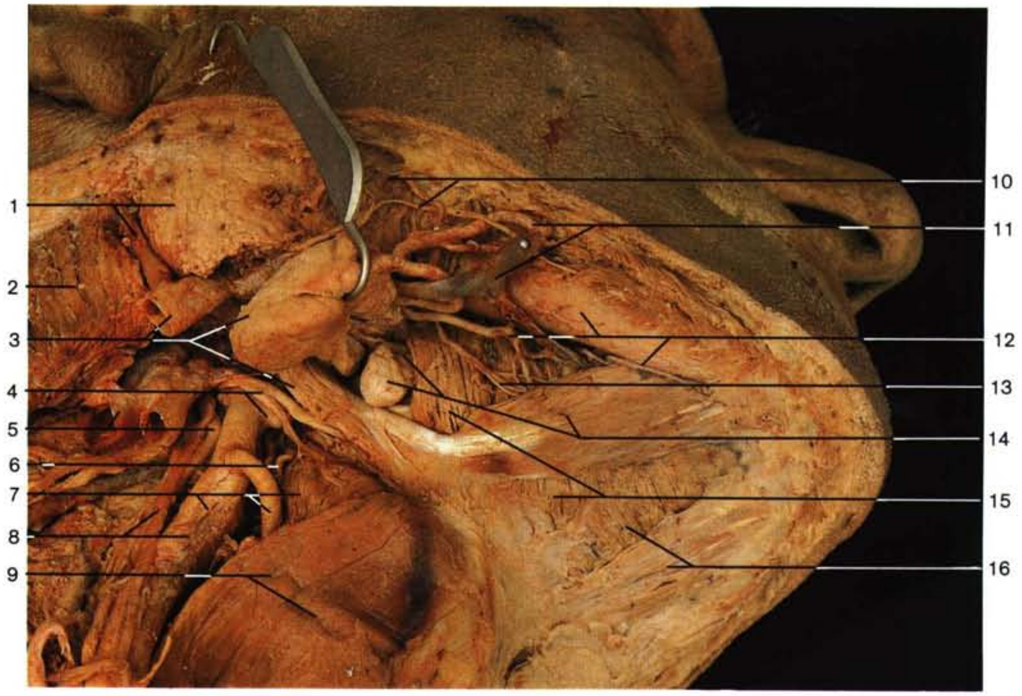

Submandibular triangle, superficial dissection. Right side (inferior aspect). Submandibular gland has been reflected.

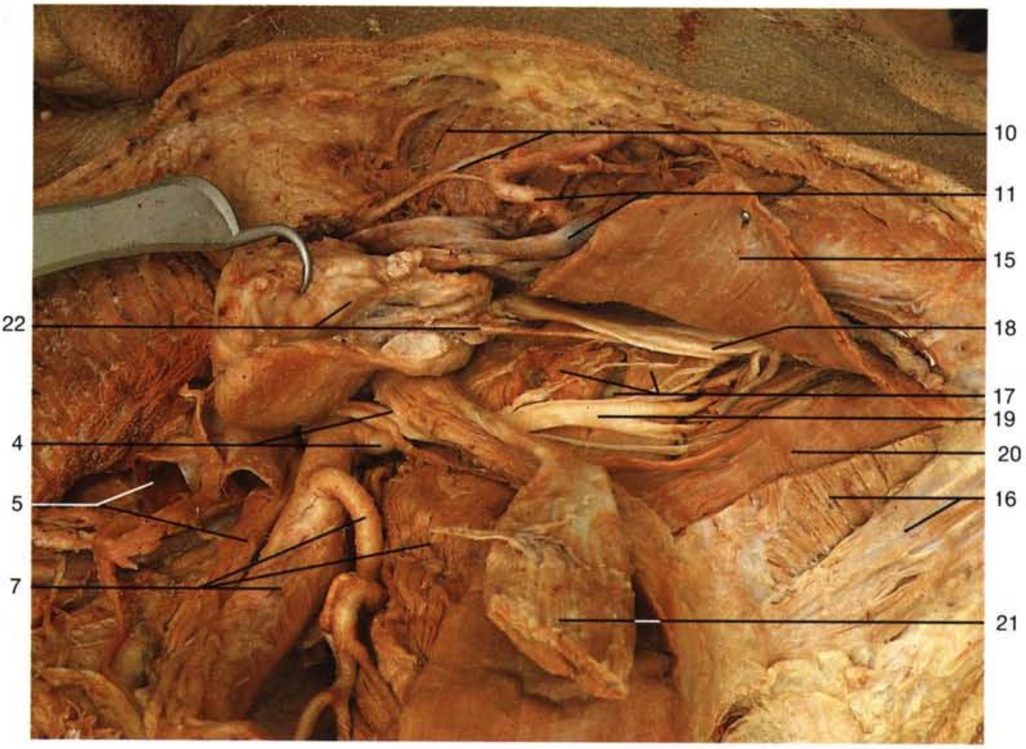

Submandibular triangle, deep dissection. Right side. Mylohyoid muscle has been severed and reflected to display the lingual and hypoglossal nerves.

1 Parotid gland and retromandibular vein
2 Sternocleidomastoid
3 Retromandibular vein, submandibular gland and stylohyoid muscle
4 Hypoglossal nerve and lingual artery
5 Vagus nerve and internal jugular vein
6 Superior laryngeal artery
7 **External carotid artery,** thyrohyoid muscle, and superior thyroid artery
8 Common carotid artery and superior root of ansa cervicalis
9 Omohyoid and sternohyoid
10 Masseter and marginal mandibular branch of facial nerve
11 **Facial artery and vein**
12 Mandible and submental artery and vein
13 Mylohyoid nerve
14 Submandibular duct, sublingual gland, and anterior belly of digastric muscle
15 Mylohyoid (right side)
16 Left mylohyoid and anterior belly of left digastric muscle
17 Hyoglossus muscle and lingual artery
18 **Lingual nerve**
19 **Hypoglossal nerve**
20 Geniohyoid muscle
21 Anterior belly of right digastric muscle
22 Submandibular gland and duct

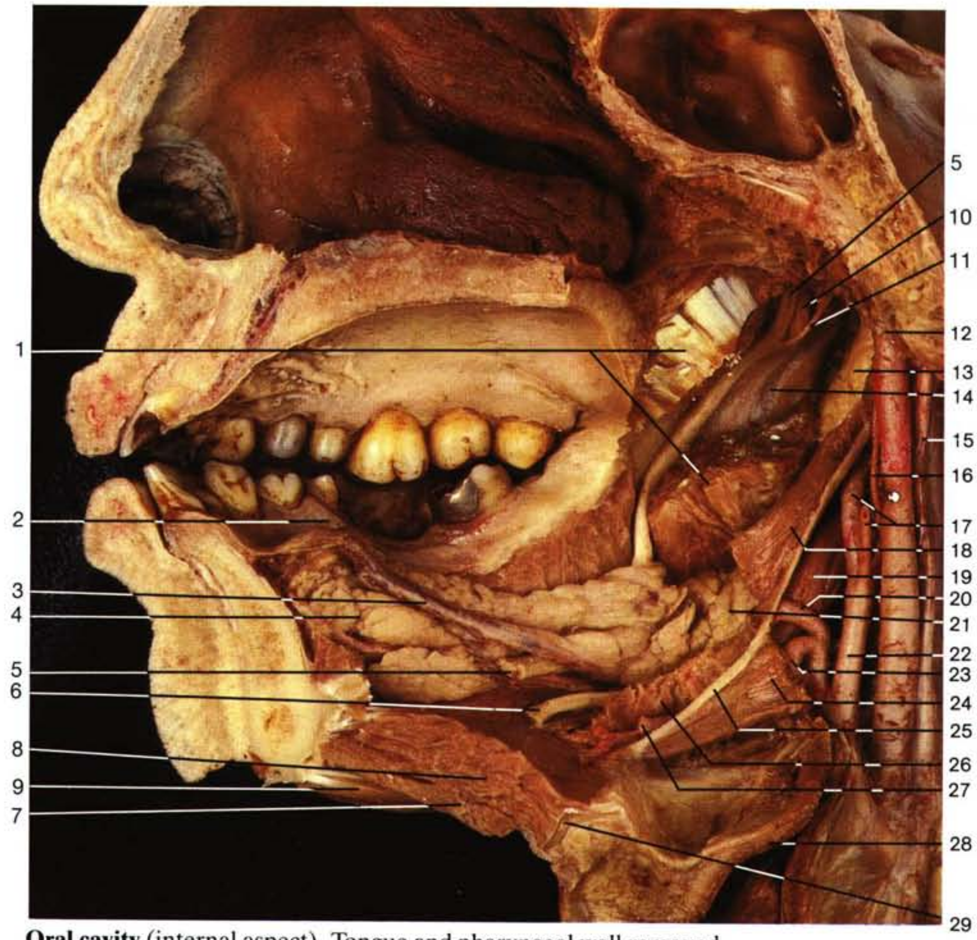

Oral cavity (internal aspect). Tongue and pharyngeal wall removed.

1 Medial pterygoid muscle
2 Sublingual papilla
3 **Submandibular duct**
4 **Sublingual gland**
5 **Lingual nerve**
6 Hypoglossal nerve
7 Mylohyoid muscle
8 Geniohyoid muscle
9 Anterior belly of digastric muscle
10 Inferior alveolar nerve
11 **Chorda tympani**
12 Internal carotid artery
13 Parotid gland
14 Sphenomandibular ligament
15 Vagus nerve
16 Glossopharyngeal nerve
17 Superficial temporal artery and ascending pharyngeal artery
18 Styloglossus muscle
19 Posterior belly of digastric muscle
20 **Facial artery**
21 Submandibular gland
22 **External carotid artery**
23 Lingual artery
24 Middle pharyngeal constrictor muscle
25 **Stylohyoid ligament**
26 Hyoglossus muscle
27 Deep lingual artery
28 Epiglottis
29 Hyoid bone
30 Buccinator muscle
31 Tongue
32 Mandible (divided)
33 **Parotid duct**
34 Masseter muscle
35 Right and left sublingual papilla

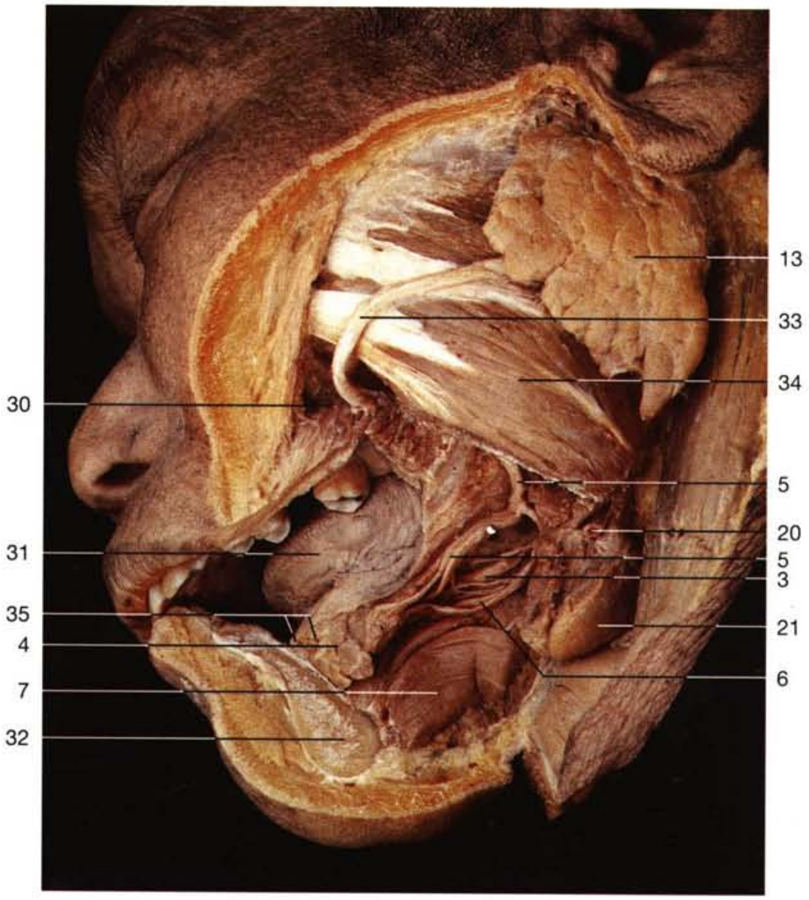

Dissection of major salivary glands. Left mandible and buccinator muscle partly removed to view the oral cavity (inferior lateral aspect).

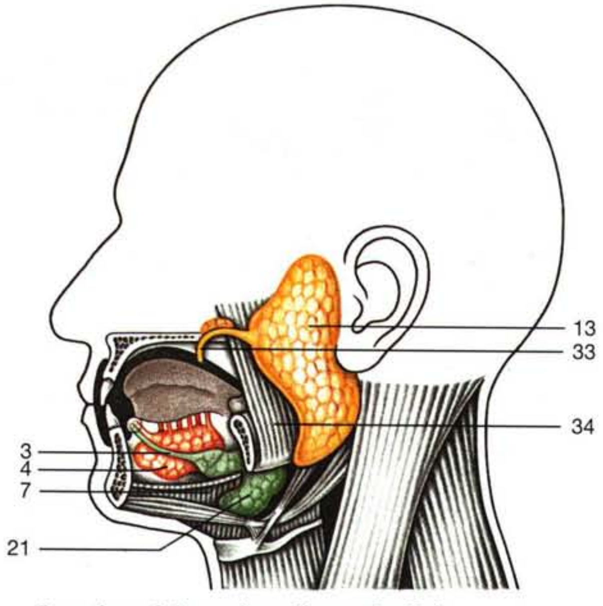

Location of the major salivary glands in relation to the oral cavity.

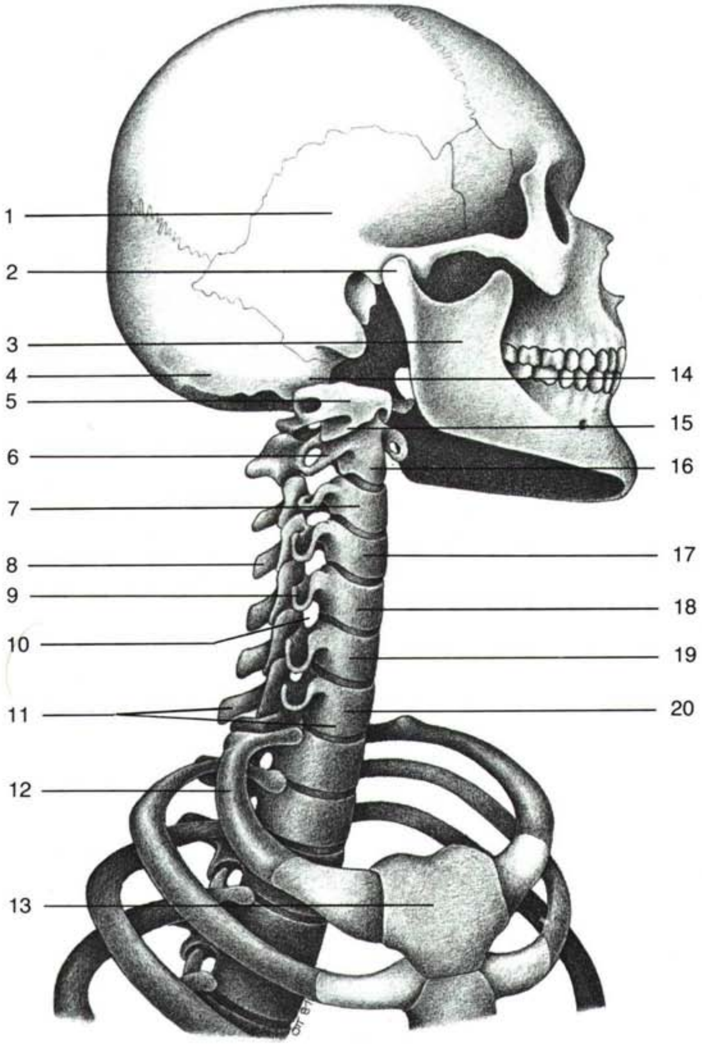

Cervical spine (oblique lateral aspect).

1 Temporal bone
2 Head of mandible
3 Mandible
4 Occipital bone
5 Atlas
6 Axis
7 Third cervical vertebra
8 Spinous process of fourth cervical vertrebra
9 Transverse process of fifth cervical vertebra with groove for spinal nerve
10 Intervertebral foramen
11 Vertebra prominens (C_7)
12 First rib
13 Sternum (manubrium sterni)
14 Atlantooccipital joint
15 Lateral atlantoaxial joint
16 Body of axis
17 Body of fourth cervical vertebra
18 Body of fifth cervical vertebra
19 Body of sixth cervical vertebra
20 Body of seventh cervical vertebra

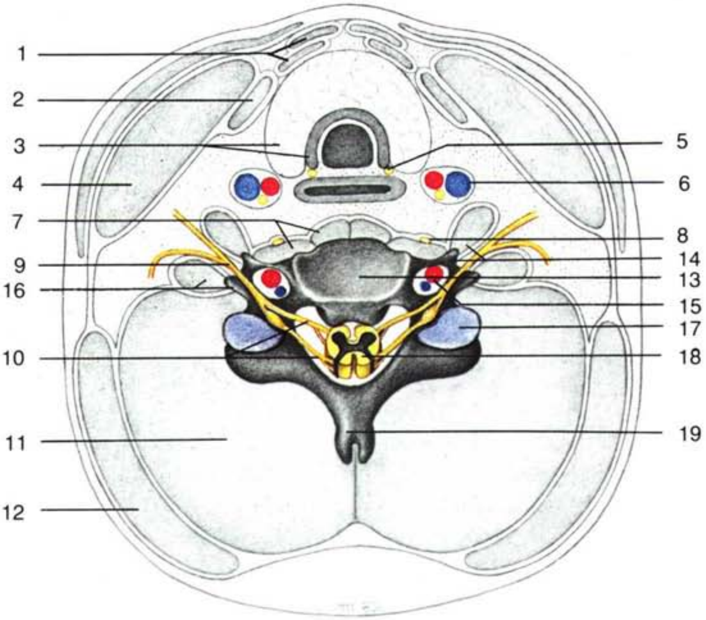

Cervical vertebra and the **organization of the neck.** (Schematic drawing.)

1 Sternohyoid and sternothyroid muscles
2 Omohyoid muscle
3 Thyroid gland and trachea
4 Sternocleidomastoid muscle
5 Recurrent laryngeal nerve
6 Internal jugular vein, common carotid artery and vagus
7 Longus colli and longus capitis muscles
8 Sympathetic trunk
9 Spinal nerve
10 Ventral and dorsal root of spinal nerve
11 True muscles of the neck
12 Trapezius muscle
13 Body of cervical vertebra
14 Anterior tubercle of transverse process and origin of scalenus anterior muscle
15 Vertebral artery and foramen transversarium
16 Posterior tubercle of transverse process and origin of scalenus medius and posterior muscles
17 Superior facet of articular process
18 Spinal cord
19 Spinous process

1 Nasal septum
2 Uvula
3 Genioglossus muscle
4 Mandible
5 Geniohyoid muscle
6 Mylohyoid muscle
7 **Hyoid bone**
8 Thyroid cartilage
9 Manubrium sterni
10 Sphenoidal sinus
11 **Nasopharynx**
12 **Oropharynx**
13 **Epiglottis**
14 **Laryngopharynx**
15 Arytenoid muscle
16 **Vocal fold**
17 Cricoid cartilage
18 **Trachea**
19 Left brachiocephalic vein
20 Thymus
21 Esophagus
22 Occipital lobe
23 Cerebellum and 4th ventricle
24 Medulla oblongata
25 Dens of axis
26 Intervertebral discs of cervical vertebral column

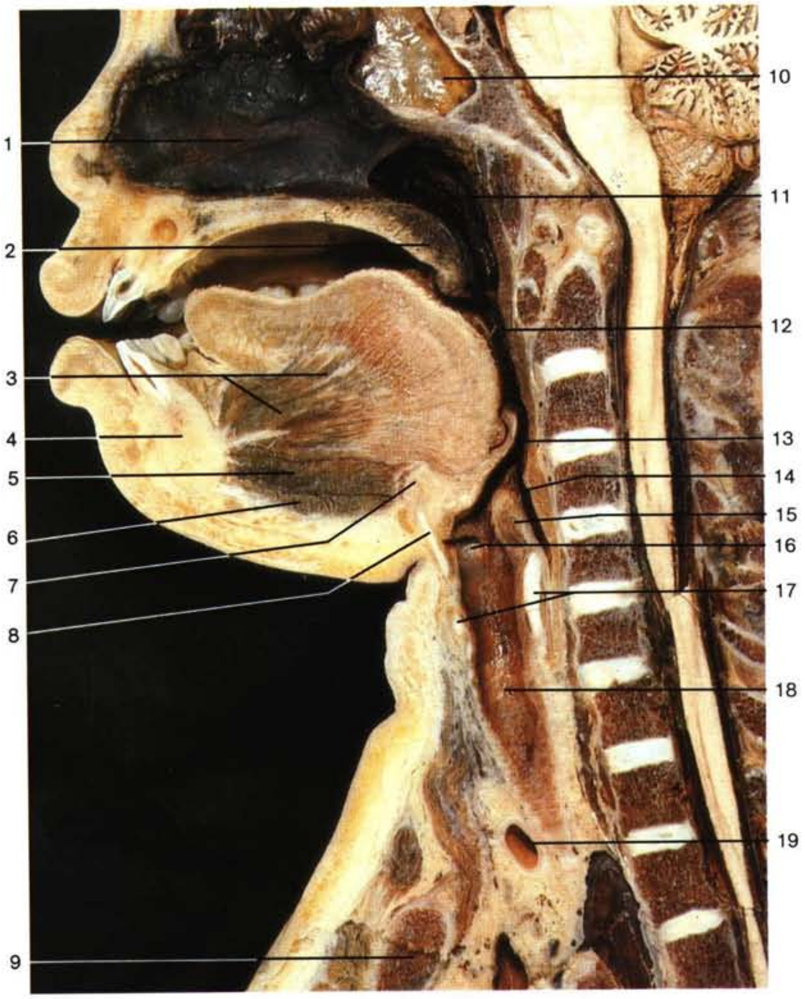

Median section through adult head and neck. Note the low position of the adult larynx when compared with that of the neonate (cf. with figure below).

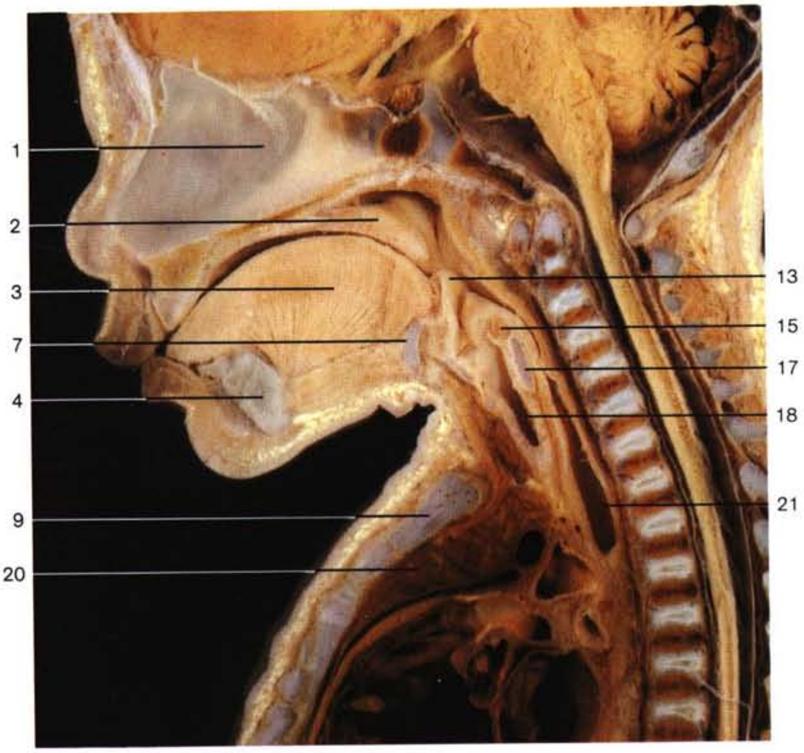

Median section through neonate head and neck. Note the high position of the larynx permitting the epiglottis nearly to reach the uvula (cf. with the figure above).

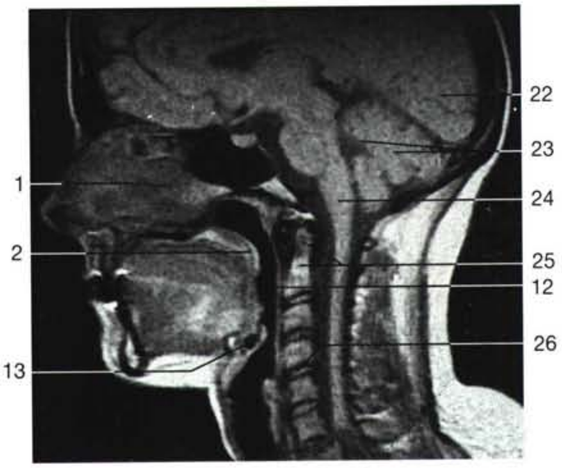

Sagittal section through the head. (MR-Scan.)

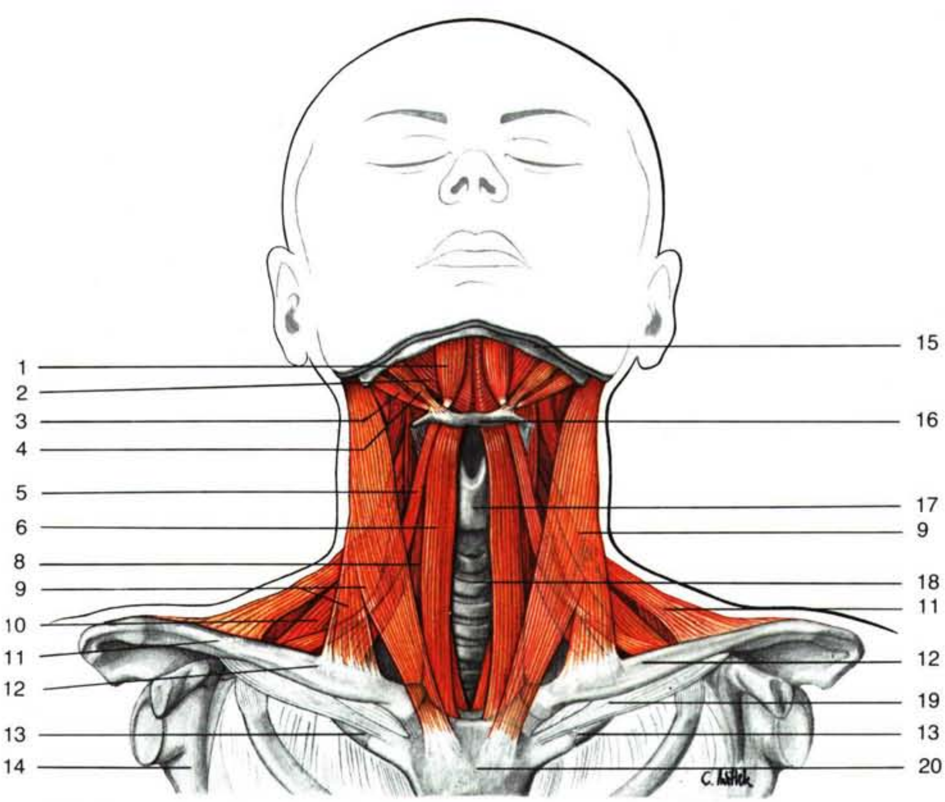

Muscles of the neck (anterior aspect).

Suprahyoid muscles

1 Anterior belly of digastric muscle
2 Mylohyoid muscle
3 Posterior belly of digastric muscle
4 Stylohyoid muscle

Infrahyoid muscles

5 Omohyoid muscle
6 Sternohyoid muscle
7 Thyrohyoid muscle
8 Sternothyroid muscle

Other structures

9 Sternocleidomastoid muscle
10 Scalenus muscles
11 Trapezius muscle
12 Clavicle
13 First rib
14 Scapula
15 Mandible
16 Hyoid bone
17 Larynx (thyroid cartilage)
18 Trachea
19 Subclavius muscle
20 Manubrium sterni
21 Mucous membrane of larynx (conus elasticus)
22 Cricoid cartilage
23 Inferior horn of thyroid cartilage
24 Esophagus
25 Body of cervical vertebra
26 Posterior root ganglion
27 Spinal cord
28 Spinous process
29 Internal jugular vein, common carotid artery, and vagus nerve
30 True muscles of the neck (semispinalis cervicis and capitis muscles)

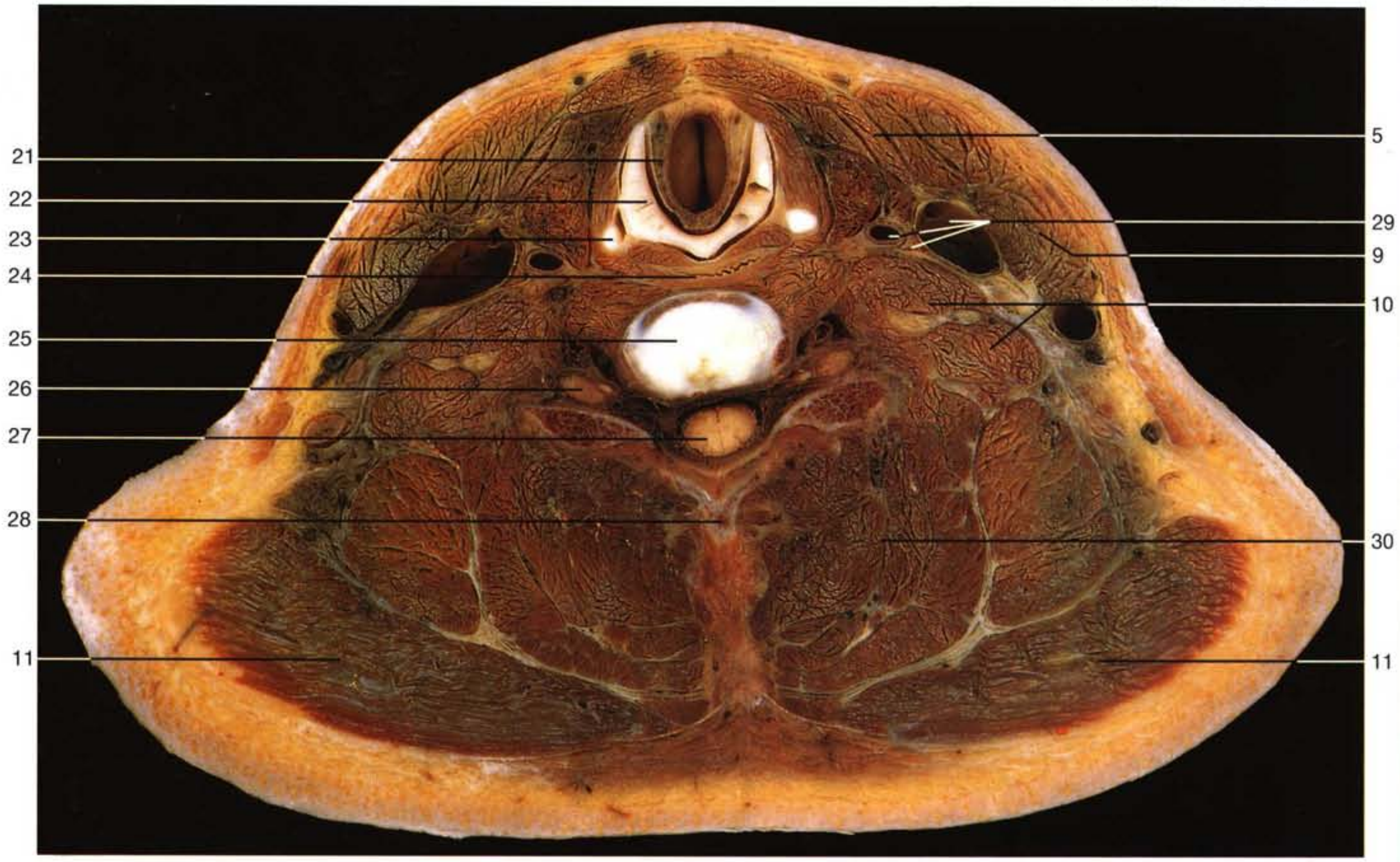

Cross-section of the neck at the level of the intervertebral disc between the 5th and 6th cervical vertebra (inferior aspect).

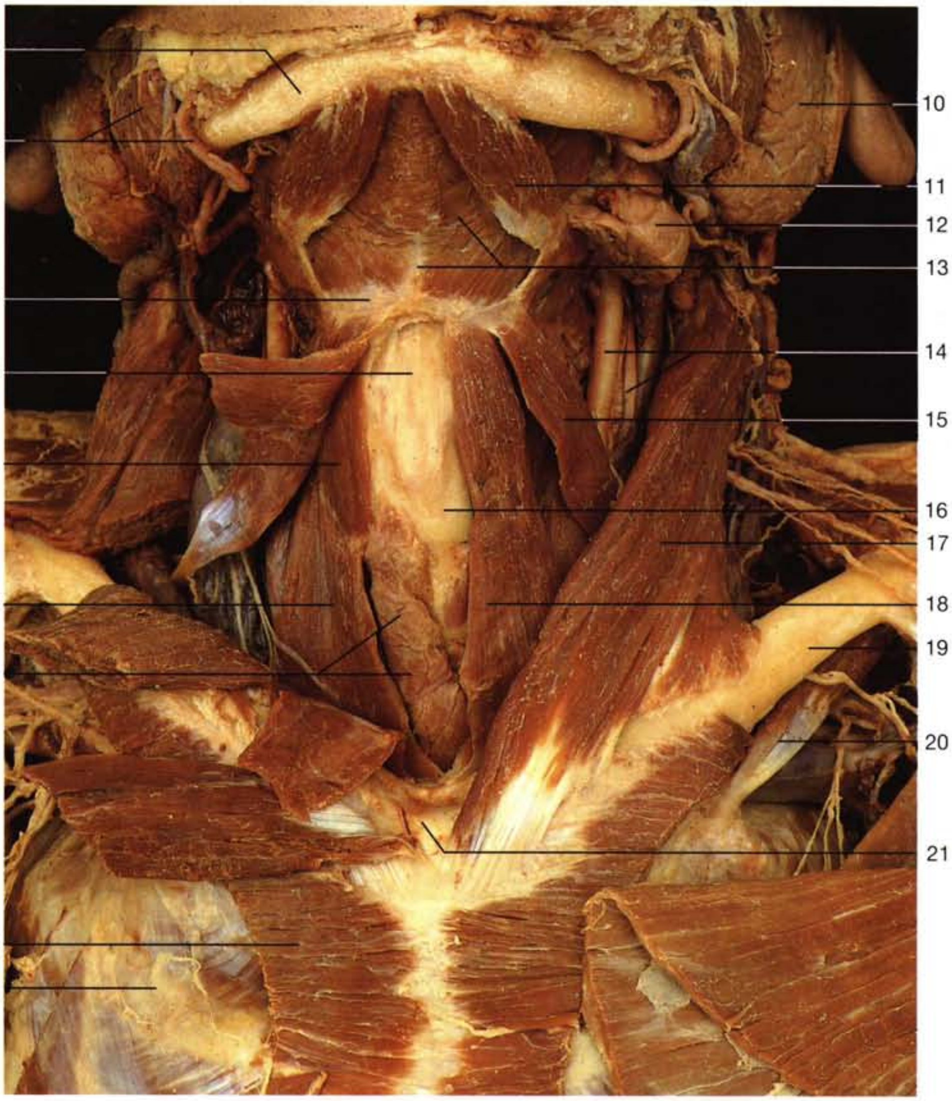

Muscles of the neck (anterior aspect). Sternocleidomastoid and sternohyoid muscles on the right have been divided and reflected.

1 Mandible
2 Masseter muscle and facial artery
3 **Hyoid bone**
4 Median thyrohyoid ligament
5 **Thyrohyoid muscle**
6 **Sternothyroid muscle**
7 Thyroid gland (pyramidal lobe)
8 Pectoralis major muscle
9 Second rib
10 Parotid gland
11 Anterior belly of **digastric muscle**
12 Submandibular gland (divided)
13 **Mylohyoid muscle** and mylohyoid raphe
14 External carotid artery and vagus nerve
15 **Omohyoid muscle**
16 Thyroid cartilage
17 Sternocleidomastoid muscle
18 **Sternohyoid muscle**
19 Clavicle
20 Subclavius
21 Jugular fossa or suprasternal notch

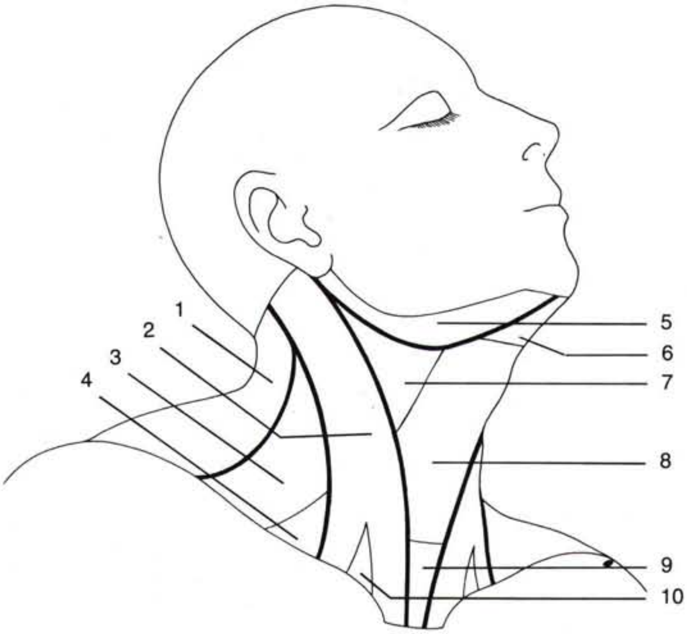

Regions and triangles of the neck.

1	Trapezius	
2	Sternocleidomastoid	
3	Occipital triangle	**Posterior triangle**
4	Supraclavicular triangle	
5	Submandibular triangle	**Anterior triangle**
6	Submental triangle	
7	Carotid triangle	
8	Muscular triangle	
9	Jugular fossa	
10	Lesser supraclavicular fossa	

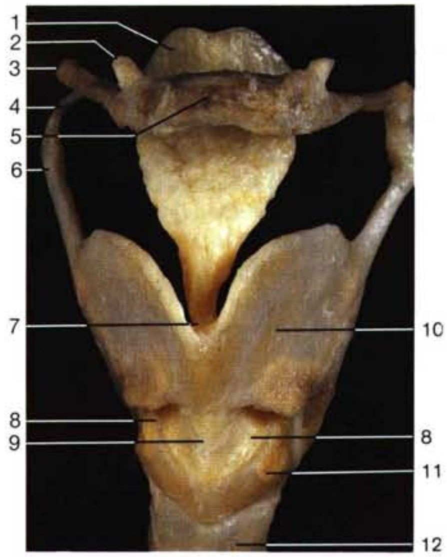

Cartilages of the larynx and the hyoid bone (anterior aspect).

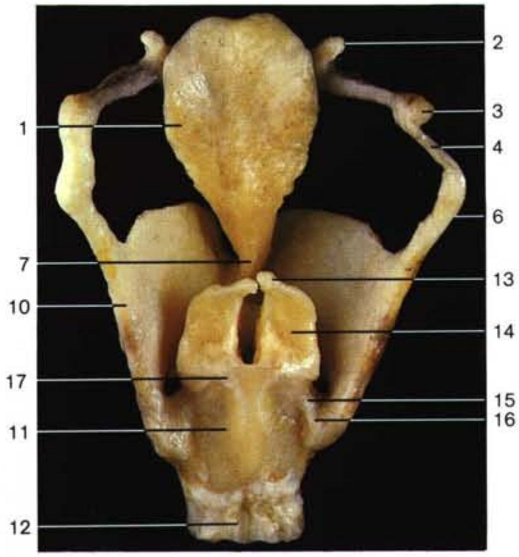

Cartilages of the larynx and the hyoid bone (posterior aspect).

1 **Epiglottis**
2 Lesser cornu of hyoid bone
3 Greater cornu of hyoid bone
4 Lateral thyrohyoid ligament
5 Body of **hyoid bone**
6 Superior cornu of thyroid cartilage
7 Thyroepiglottic ligament
8 Conus elasticus
9 Cricothyroid ligament
10 **Thyroid cartilage**
11 **Cricoid cartilage**
12 Trachea
13 Corniculate cartilage
14 **Arytenoid cartilage**
15 Posterior cricoarytenoid ligament
16 **Cricothyroid joint**
17 **Cricoarytenoid joint**

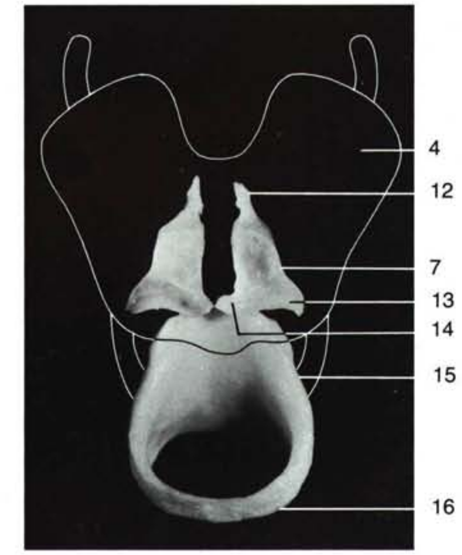

Cartilages of the larynx (anterior aspect). Thyroid cartilage is indicated by the outline.

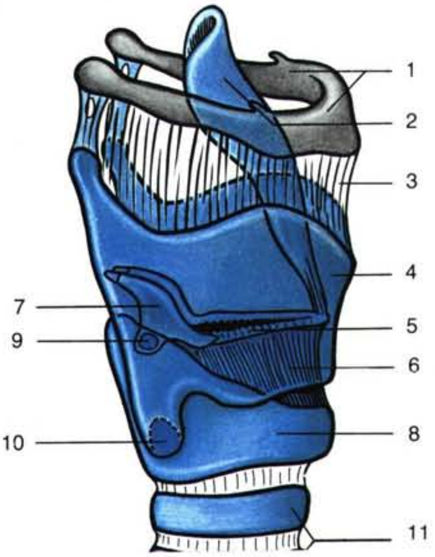

Cartilages and ligaments of the larynx (lateral aspect). (Schematic drawing.)

1 Hyoid bone
2 Epiglottis
3 Thyrohyoid membrane
4 **Thyroid cartilage**
5 **Vocal ligament**
6 Conus elasticus
7 Arytenoid cartilage
8 **Cricoid cartilage**
9 **Cricoarytenoid joint**
10 **Cricothyroid joint**
11 Tracheal cartilages
12 Corniculate cartilage
13 Muscular process of arytenoid cartilage
14 Vocal process of arytenoid cartilage
15 Lamina of cricoid cartilage
16 Arch of cricoid cartilage

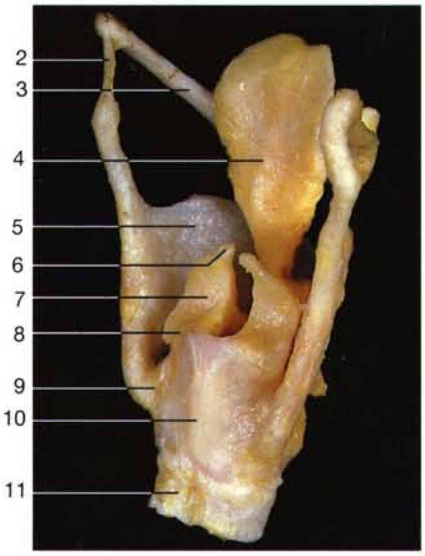

Cartilages of the larynx (oblique-posterior aspect).

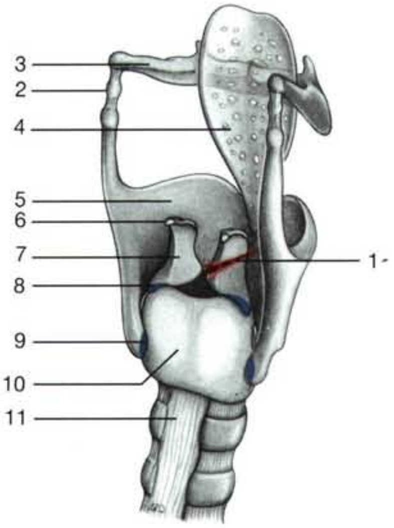

Cartilages of the larynx (oblique-posterior aspect).

1 **Vocal ligament** (red)
2 Lateral thyrohyoid ligament
3 Greater cornu of hyoid bone
4 **Epiglottis**
5 **Thyroid cartilage**
6 Corniculate cartilage
7 **Arytenoid cartilage**
8 **Cricoarytenoid joint**
9 **Cricothyroid joint**
10 **Cricoid cartilage**
11 Trachea

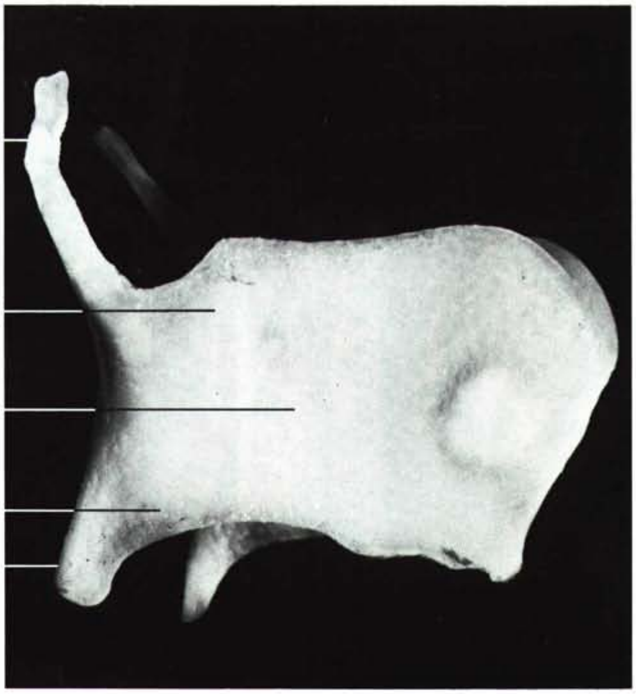

Thyroid cartilage (lateral aspect).

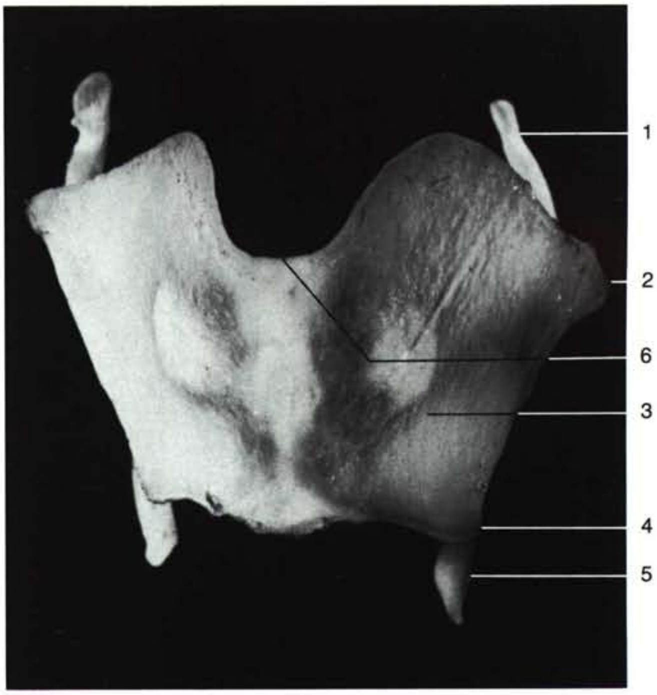

Thyroid cartilage (anterior aspect).

1 Superior cornu
2 Superior thyroid tubercle
3 Lamina of thyroid cartilage
4 Inferior thyroid tubercle
5 Inferior cornu
6 Superior thyroid notch

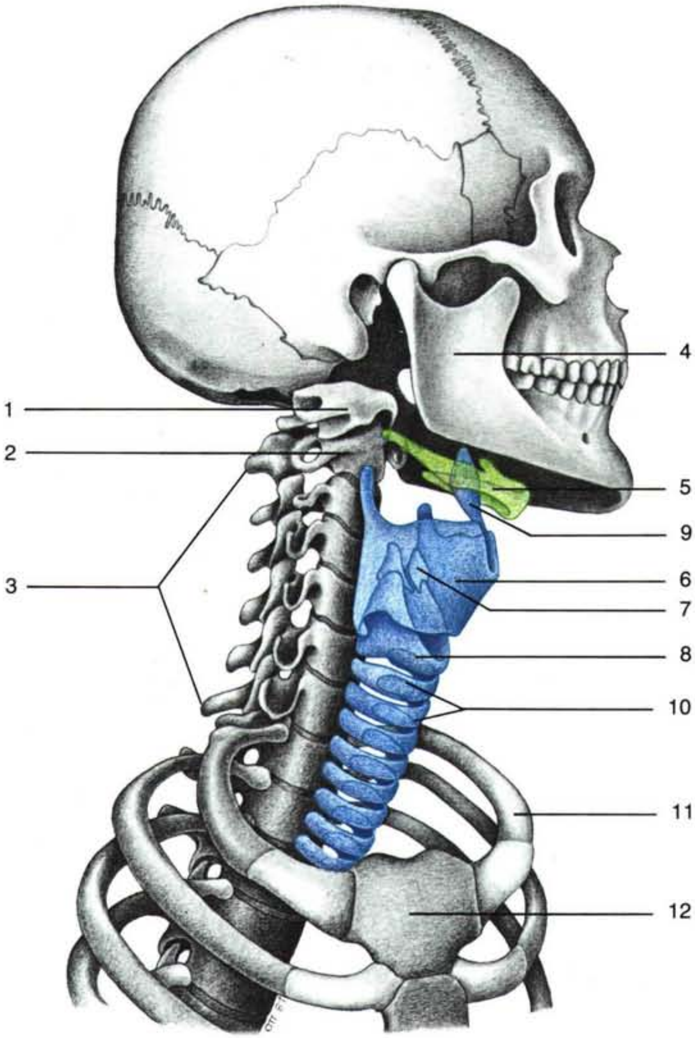

1 Atlas
2 Axis
3 Cervical vertebrae (C_2–C_7)
4 Mandible
5 **Hyoid bone**
6 **Thyroid cartilage**
7 Arytenoid cartilage
8 **Cricoid cartilage**
9 Epiglottis
10 **Tracheal cartilages**
11 First rib
12 Manubrium sterni

Position of the larynx in the neck (oblique lateral aspect). (Schematic drawing.)

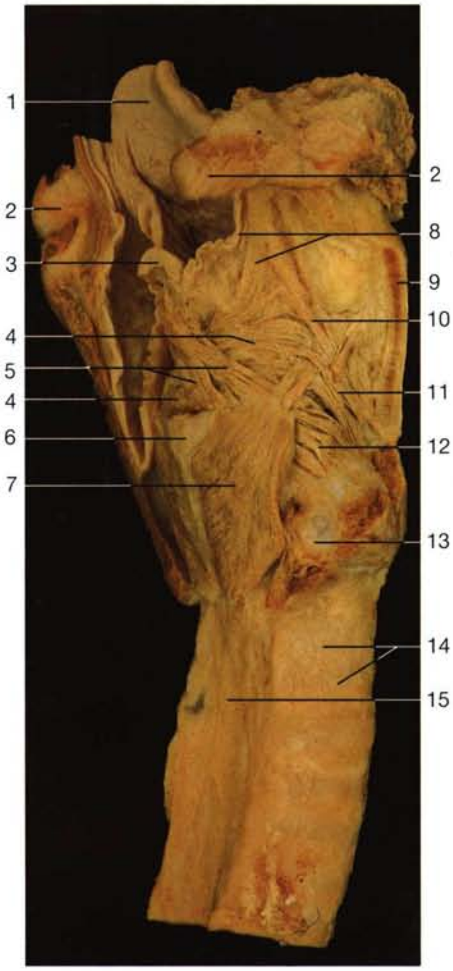

Laryngeal muscles I (lateral aspect). Thyroid cartilage and thyroarytenoid muscle have been partly removed.

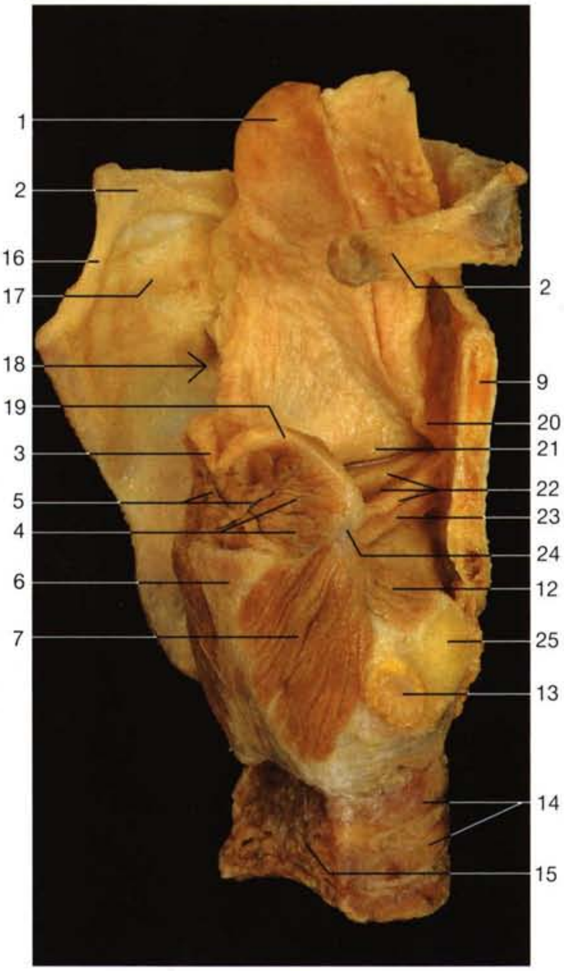

Laryngeal muscles II (lateral aspect). Half of the right side of the thyroid cartilage has been removed.

1 Epiglottis
2 Greater horn of hyoid bone
3 Corniculate cartilage
4 Transverse arytenoid muscle
5 Oblique arytenoid muscle
6 Lamina of cricoid cartilage
7 **Posterior cricoarytenoid muscle**
8 Aryepiglottic fold and muscle
9 Lamina of thyroid cartilage
10 Thyroepiglottic muscle
11 Thyroarytenoid muscle
12 **Lateral cricoarytenoid muscle**
13 Articular facet of thyroid cartilage
14 Cartilages of trachea
15 Membranous part of trachea
16 Lateral thyrohyoid ligament
17 Thyrohyoid membrane
18 Position of piriform recess
19 Arytenoid cartilage
20 Stem of epiglottis
21 Vestibular fold
22 Vocal folds and fissure of glottis
23 **Vocalis muscle**
24 Cricoarytenoid articulation
25 Arch of cricoid cartilage
26 Cricothyroid muscle
27 Cricoid cartilage
28 Piriform recess
29 Superior horn of thyroid cartilage
30 Vocal ligament

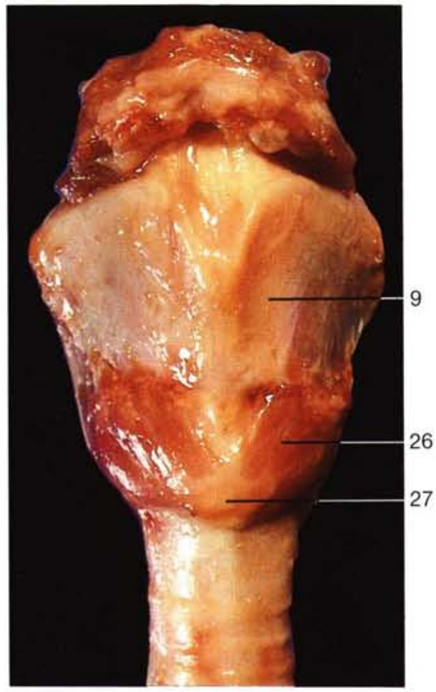

Laryngeal muscles, larynx (anterior aspect).

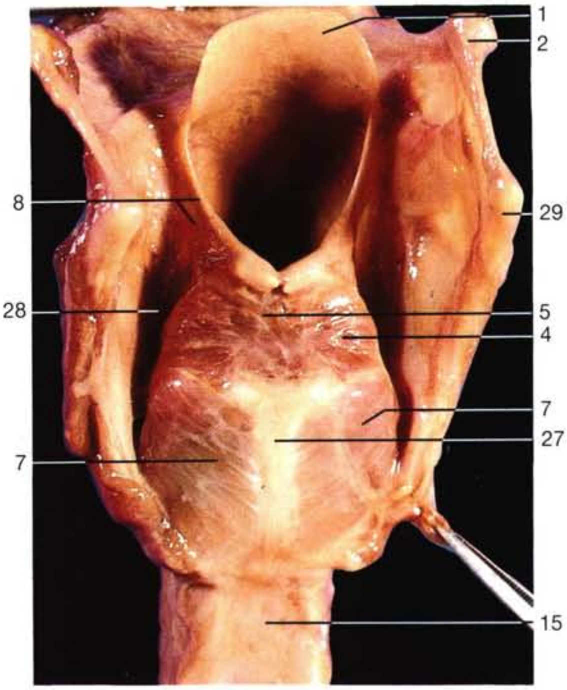

Laryngeal muscles, larynx (posterior aspect).

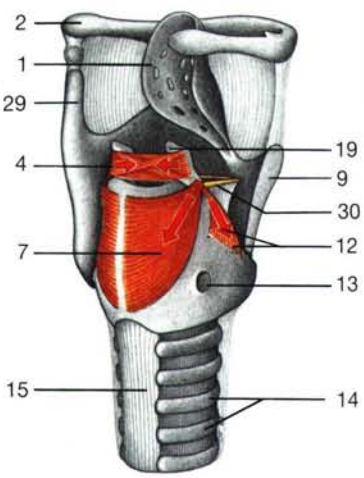

Action of internal muscles of the larynx. (Schematic drawing.)

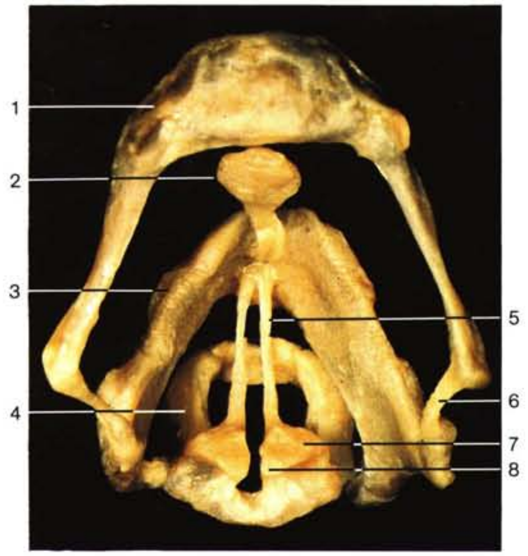

Laryngeal cartilages (superior aspect).

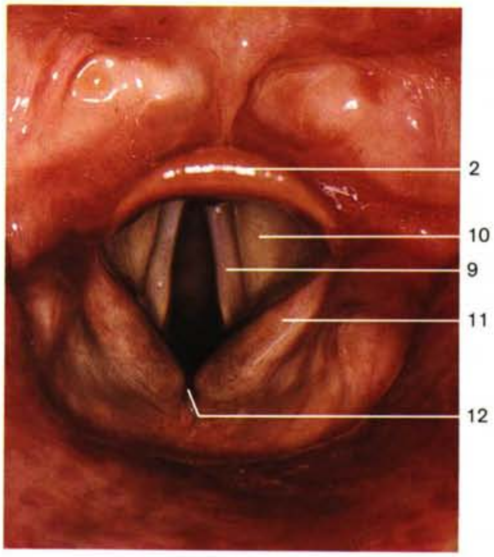

Glottis in vivo (superior aspect).

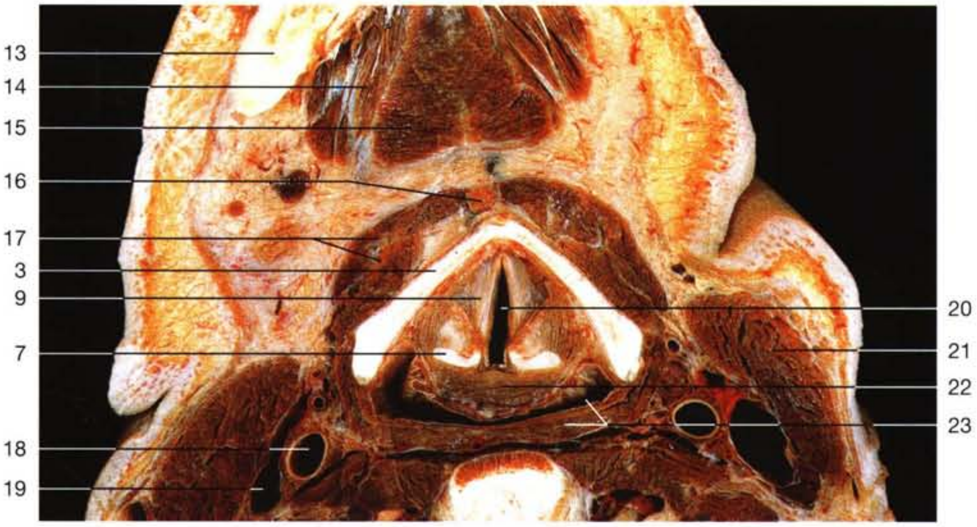

Horizontal section through the larynx at the level of the vocal folds (superior aspect).

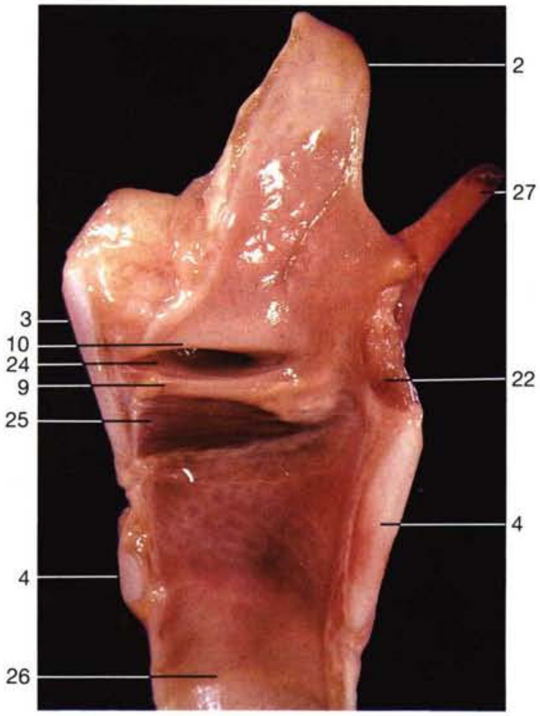

Sagittal section through the larynx.

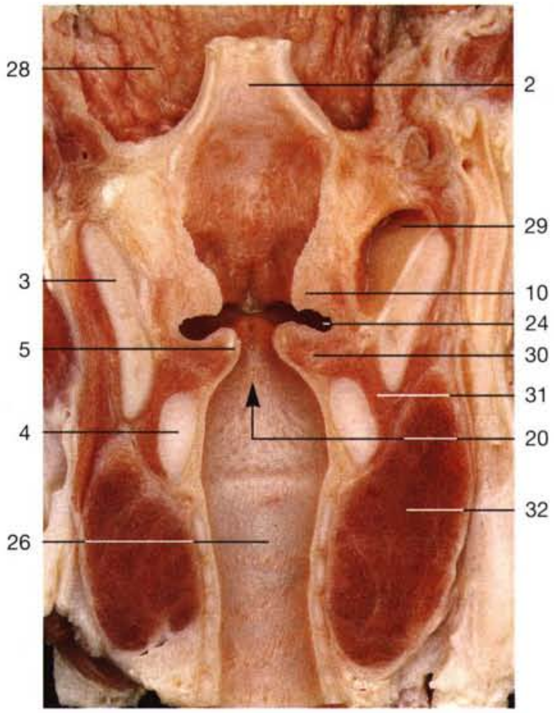

Coronal section through larynx and trachea.

1	**Hyoid bone**
2	Epiglottis
3	**Thyroid cartilage**
4	**Cricoid cartilage**
5	**Vocal ligament**
6	Thyrohyoid ligament
7	**Arytenoid cartilage**
8	Corniculate cartilage
9	**Vocal fold**
10	**Vestibular fold**
11	Aryepiglottic fold
12	Interarytenoid notch
13	Mandible
14	Anterior belly of digastric muscle
15	Mylohyoid muscle
16	Pyramidal process of thyroid gland
17	Sternohyoid and sternothyroid muscles
18	Common carotid artery
19	Internal jugular vein
20	**Rima glottidis**
21	Sternocleidomastoid muscle
22	Transverse arytenoid muscle
23	Pharynx and inferior constrictor muscle
24	Ventricle of larynx
25	**Vocalis muscle**
26	Trachea
27	Superior cornu of thyroid cartilage
28	Root of tongue (lingual tonsil)
29	Piriform recess
30	Vocalis muscle
31	Lateral cricoarytenoid muscle
32	**Thyroid gland**

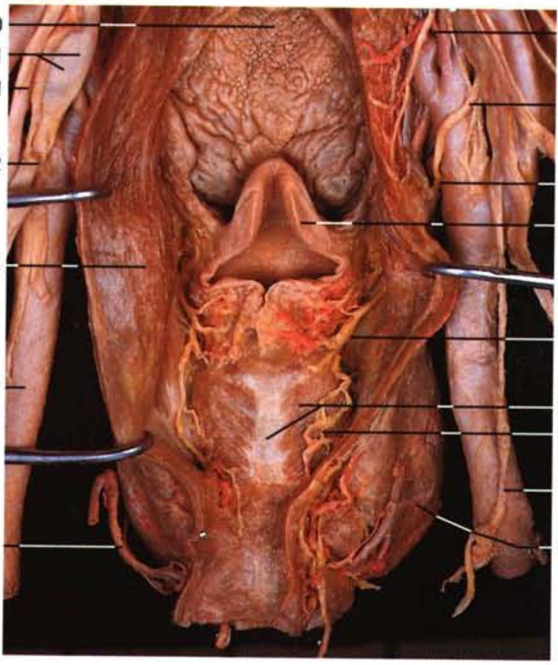

Larynx and its innervation (posterior aspect). Dissection of superior and inferior laryngeal nerves. Pharynx has been opened.

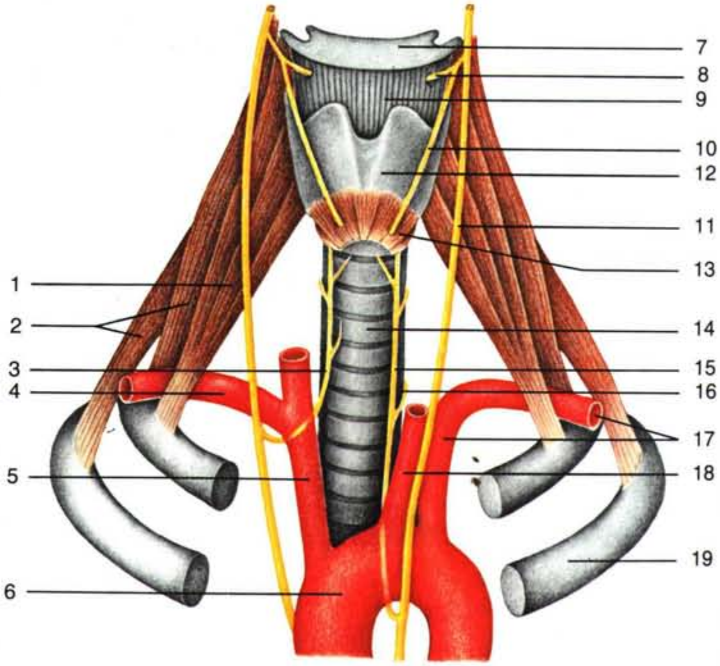

Innervation of the larynx. (Schematic diagram.)

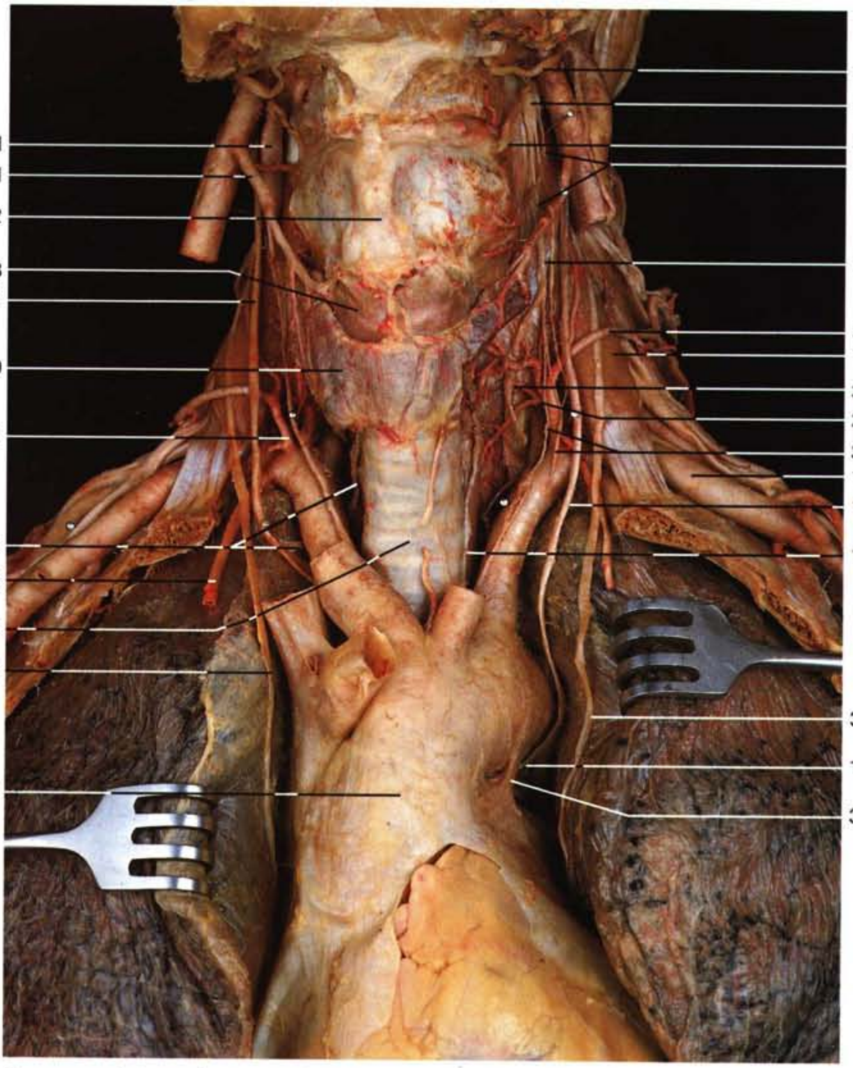

Larynx and thoracic organs (anterior aspect). Dissection of vagus and recurrent laryngeal nerves.

1 Scalenus anterior muscle
2 Scalenus medius and posterior muscles
3 Right **recurrent laryngeal nerve**
4 Right subclavian artery
5 Brachiocephalic trunk
6 Aortic arch
7 Hyoid bone
8 **Internal branch of superior laryngeal nerve**
9 Thyrohyoid membrane
10 **External branch of superior laryngeal nerve**
11 **Vagus nerve**
12 Thyroid cartilage
13 Cricothyroid muscle
14 Trachea
15 **Left recurrent larnygeal nerve**
16 Esophagus
17 Left subclavian artery
18 Left common carotid artery
19 Second rib
20 Tongue
21 Superior cervical ganglion
22 Sympathetic trunk
23 Inferior constrictor muscle of pharynx
24 **Inferior thyroid artery**
25 Glossopharyngeal nerve
26 Superior laryngeal nerve
27 Epiglottis
28 Posterior cricoarytenoid muscle and cricoid cartilage
29 Inferior laryngeal branch of recurrent laryngeal nerve
30 **Thyroid gland**
31 Superior thyroid artery
32 **Thyrocervical trunk**
33 Internal thoracic artery
34 **Phrenic nerve**
35 Hypoglossal nerve
36 Transverse cervical artery
37 Middle cervical ganglion
38 Middle cervical cardiac nerves (branches of sympathetic trunk)
39 Ligamentum arteriosum

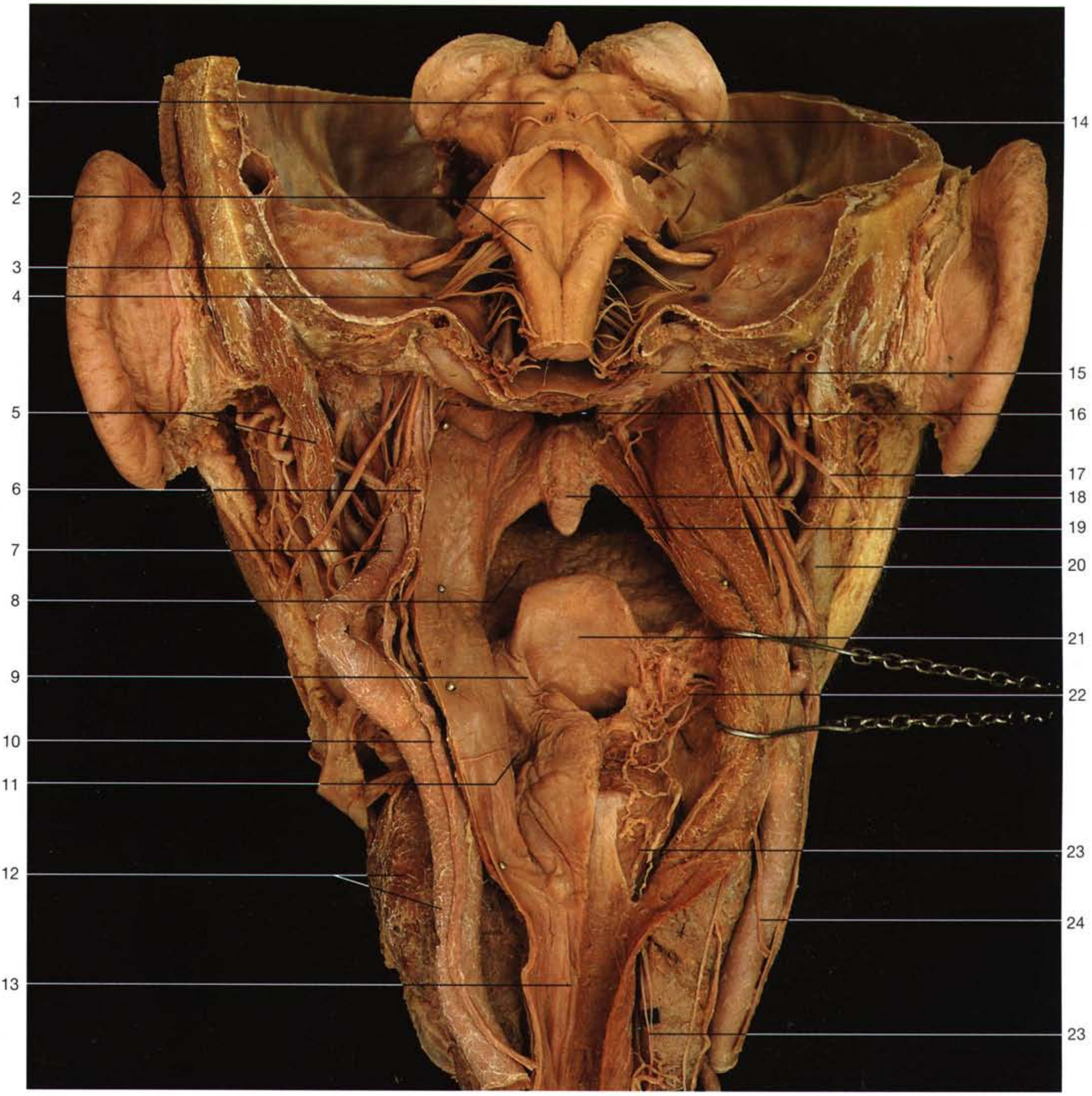

Larynx and oral cavity (posterior aspect). Mucous membrane on the right half of pharynx has been removed.

1 Midbrain (inferior colliculus)
2 Rhomboid fossa and medulla oblongata
3 Vestibulocochlear and facial nerve
4 Glossopharyngeal, vagus and accessory nerve
5 Occipital artery and posterior belly of digastric muscle
6 Superior cervical ganglion
7 **Internal carotid artery**
8 **Oral cavity** (tongue)
9 Aryepiglottic fold
10 Vagus nerve
11 **Piriform recess**
12 Thyroid gland and common carotid artery
13 Esophagus
14 Trochlear nerve
15 Occipital condyle
16 **Nasal cavity** (choana)
17 Accessory nerve
18 **Uvula** and soft palate
19 Palatopharyngeus muscle
20 External carotid artery
21 Epiglottis
22 **Internal branch of superior laryngeal nerve**
23 **Inferior laryngeal nerve**
24 Ansa cervicalis

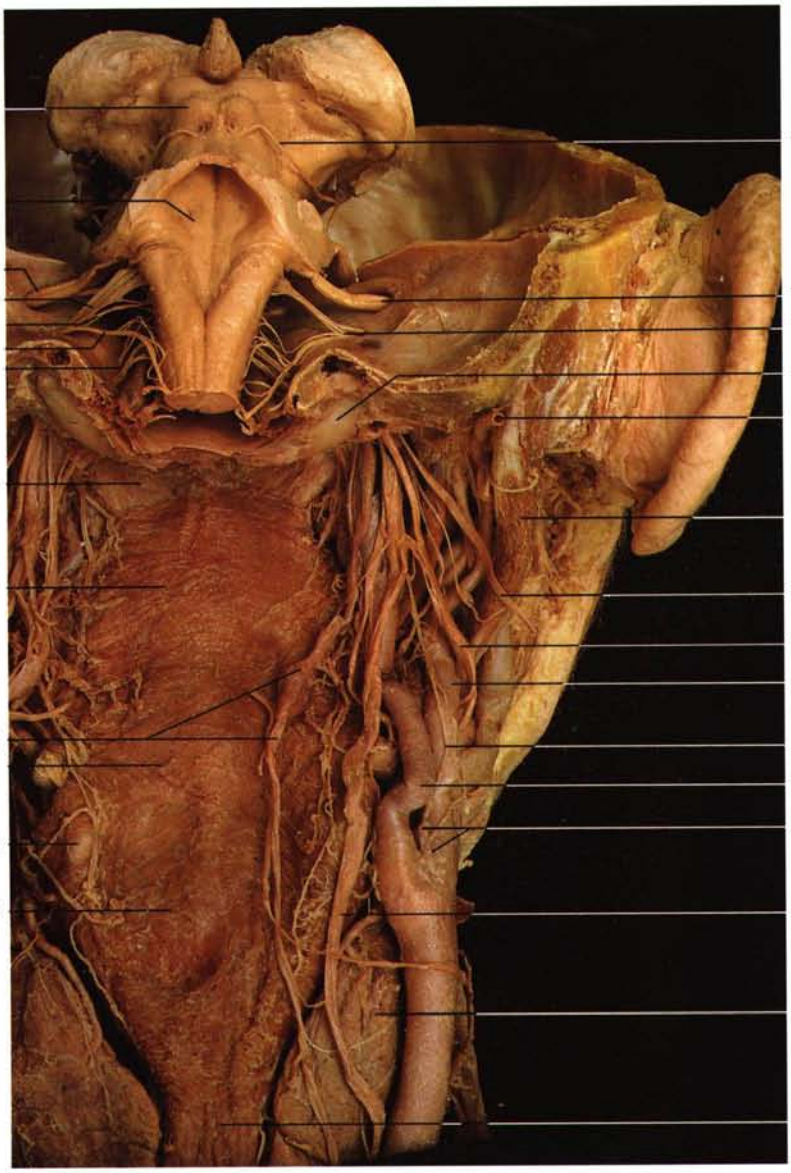

Pharynx and parapharyngeal nerves in connection with brain stem (posterior aspect).

1 Inferior colliculus of midbrain
2 Facial colliculus in **floor of rhomboid fossa**
3 Vestibulocochlear and facial nerves
4 Glossopharyngeal nerve
5 Vagus nerve
6 Accessory nerve
7 Hypoglossal nerve
8 Pharyngobasilar fascia
9 Superior constrictor muscle of pharynx
10 **Sympathetic trunk** and **superior cervical ganglion** (medially displaced)
11 Middle constrictor muscle of pharynx
12 Greater cornu of hyoid bone
13 Inferior constrictor muscle of pharynx
14 Trochlear nerve
15 **Internal acoustic meatus** with facial and vestibulocochlear nerves
16 **Jugular foramen** with glossopharyngeal, vagus and assessory nerves
17 Occipital condyle
18 Occipital artery
19 Posterior belly of digastric muscle
20 **Accessory nerve** (extracranial part)
21 **Hypoglossal nerve** (extracranial part)
22 External carotid artery
23 Carotid sinus nerve
24 **Internal carotid artery**
25 **Carotid sinus and carotid body**
26 Vagus nerve
27 Thyroid gland
28 Esophagus
29 Choanae
30 Medial pterygoid plate
31 Foramen lacerum
32 **Pharyngeal tubercle**
33 Hard palate
34 Greater and lesser palatine foramen
35 Pterygoid hamulus
36 Lateral pterygoid plate
37 Pterygoid canal
38 Foramen ovale
39 Mandibular fossa
40 Carotid canal
41 Styloid process and stylomastoid foramen

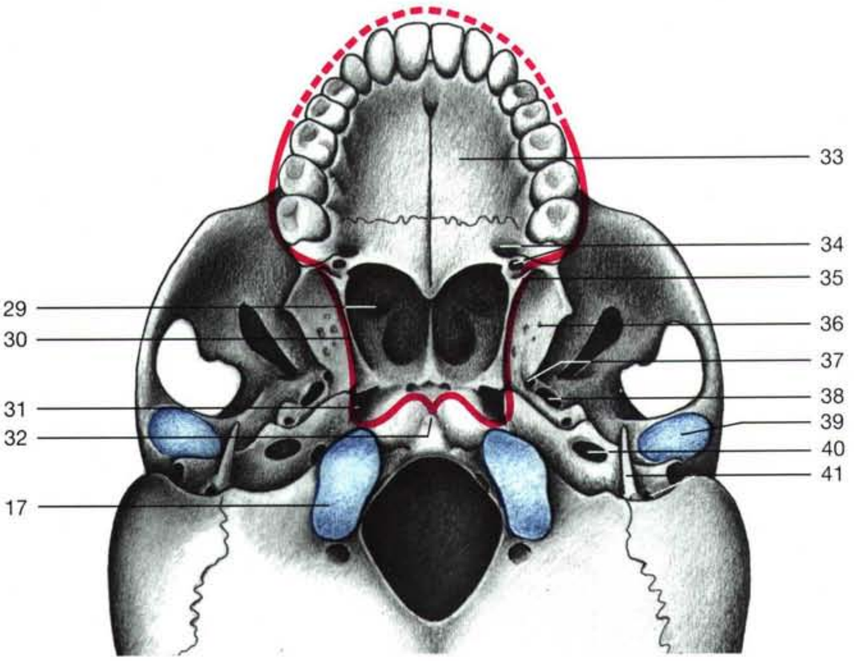

Inferior aspect of the skull. Red line = outline of superior constrictor muscle in continuation with buccinator muscle and orbicularis oris muscle. (Semischematic drawing.)

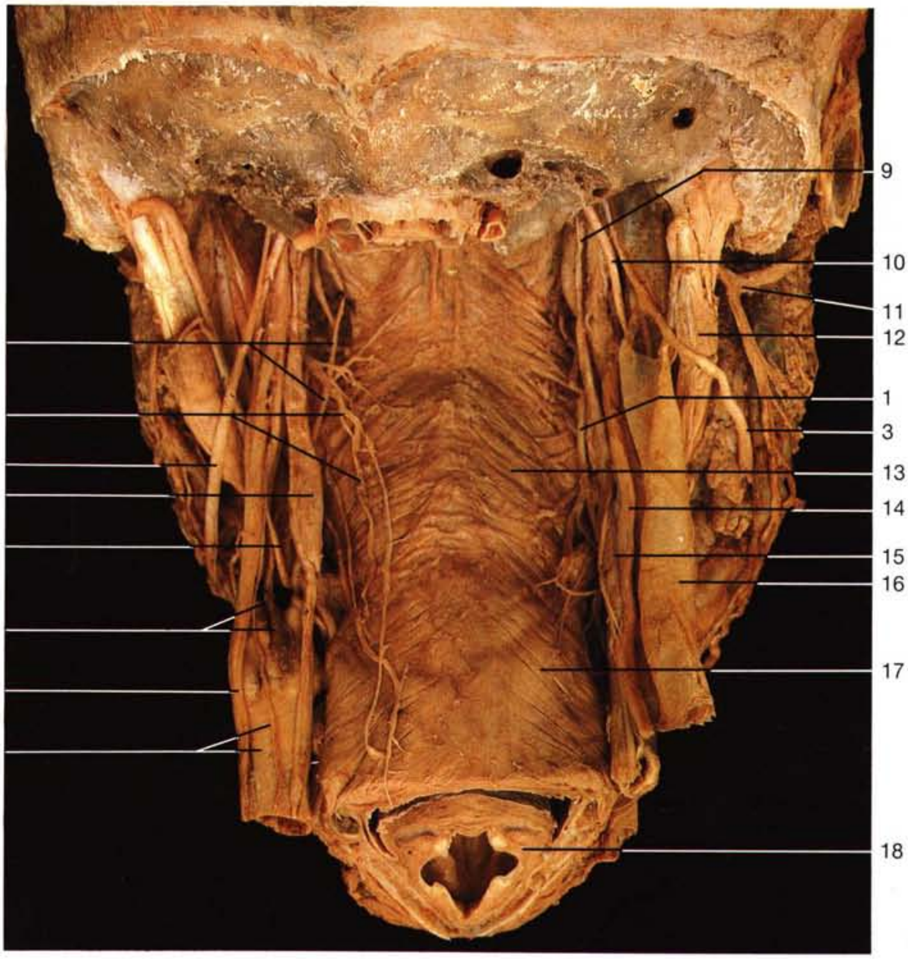

Parapharyngeal nerves and vessels. Dorsal aspect of the pharynx.

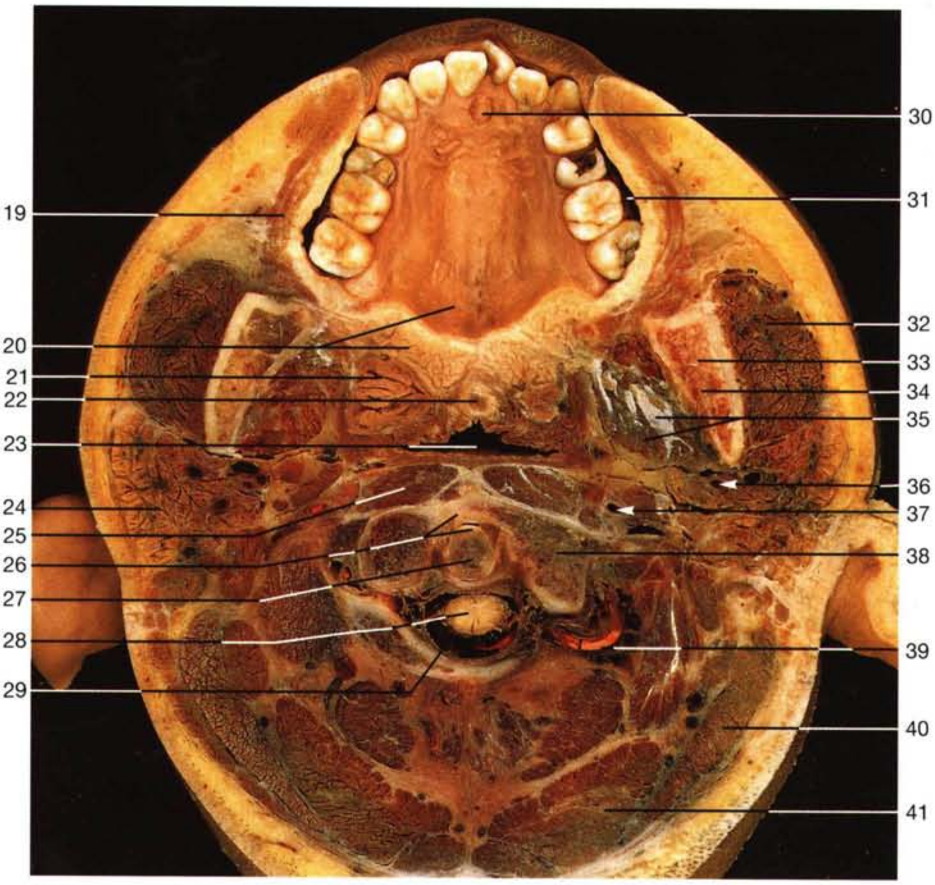

Cross-section of head and neck at the level of the atlas (inferior aspect).

1 Ascending pharyngeal artery
2 Pharyngeal plexus
3 **Accessory nerve**
4 Superior cervical ganglion of sympathetic trunk
5 Superior laryngeal nerve
6 Carotid body and carotid sinus nerve
7 Left **vagus nerve**
8 Common carotid artery and cardiac branch of vagus nerve
9 **Glossopharyngeal nerve**
10 **Hypoglossal nerve**
11 **Facial nerve**
12 Posterior belly of digastric muscle
13 Middle constrictor muscle of pharynx
14 Right vagus nerve
15 **Sympathetic trunk**
16 **Internal jugular vein**
17 Inferior constrictor muscle of pharynx
18 Larynx
19 **Buccinator muscle**
20 Soft palate and palatine glands
21 Palatine tonsil
22 Uvula of palate
23 **Pharynx** (oral part)
24 Parotid gland
25 Longus capitis muscle
26 Medioatlantoaxial joint and anterior arch of atlas
27 **Dens of axis**
28 **Spinal cord**
29 Dura mater
30 Incisive papilla
31 **Oral vestibule**
32 Masseter muscle
33 Mandible
34 Mandibular canal with vessels and nerve
35 Medial pterygoid muscle
36 External carotid artery
37 Internal carotid artery
38 **Atlas**
39 **Vertebral artery**
40 Splenius capitis muscle
41 Semispinalis capitis muscle

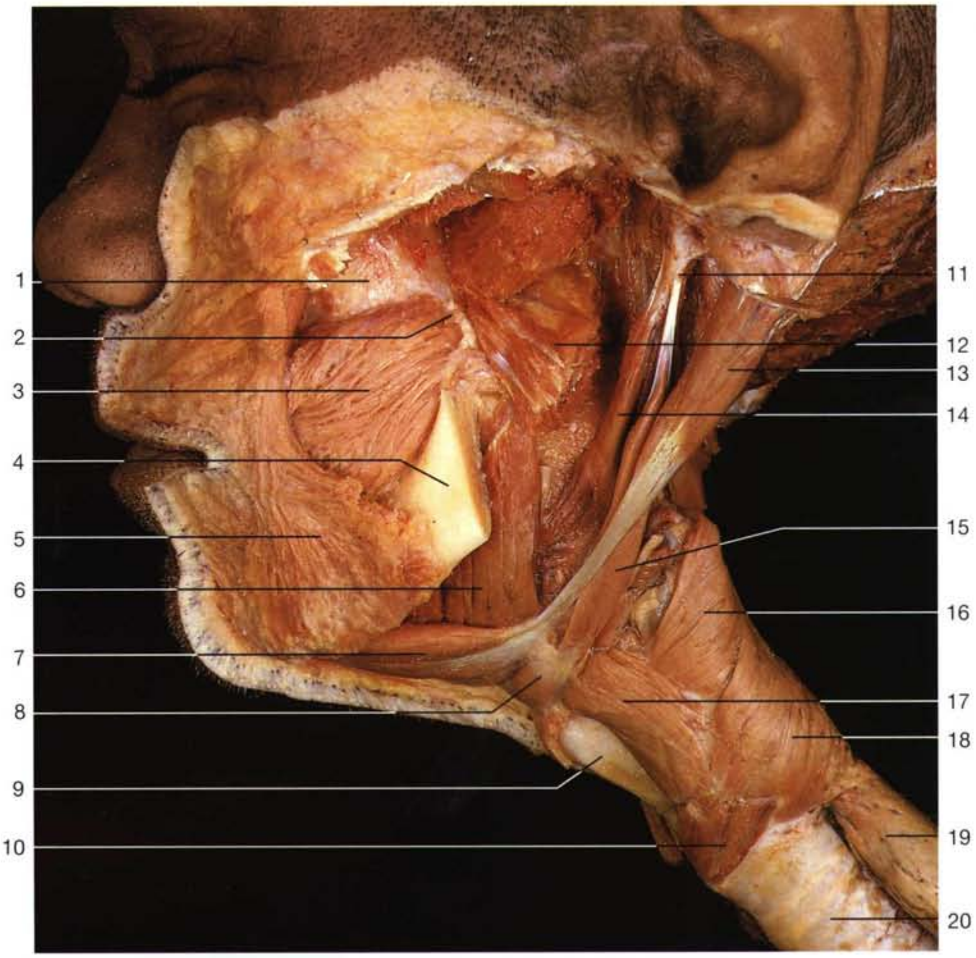

Dissection of pharynx, supra- and infrahyoid muscles I. Mandible partly removed (lateral aspect).

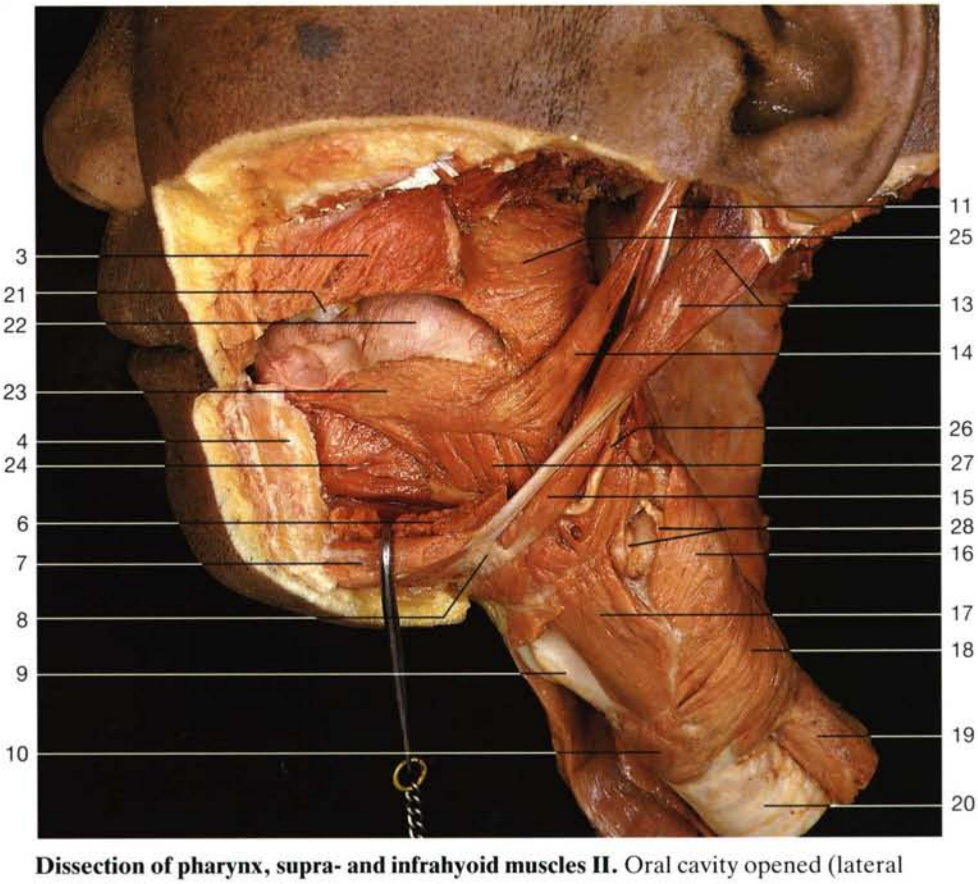

Dissection of pharynx, supra- and infrahyoid muscles II. Oral cavity opened (lateral aspect).

1	Maxilla
2	Pterygomandibular raphe
3	Buccinator muscle
4	Mandible (divided)
5	Depressor anguli oris muscle
6	Mylohyoid muscle
7	Anterior belly of digastric muscle
8	Hyoid bone
9	Thyroid cartilage
10	Cricothyroid muscle
11	Styloid process
12	Medial pterygoid muscle (divided)
13	Posterior belly of digastric muscle
14	Styloglossus muscle
15	Stylohyoid muscle
16	**Thyropharyngeal part of inferior constrictor muscle of pharynx**
17	Thyrohyoid muscle
18	**Cricopharyngeal part of inferior constrictor muscle of pharynx**
19	Esophagus
20	Trachea
21	First molar of maxilla
22	Tongue
23	Inferior longitudinal muscle of tongue
24	Genioglossus muscle
25	**Superior constrictor muscle of pharynx**
26	Hypoglossal nerve
27	Hyoglossus muscle
28	Superior laryngeal nerve and superior laryngeal artery

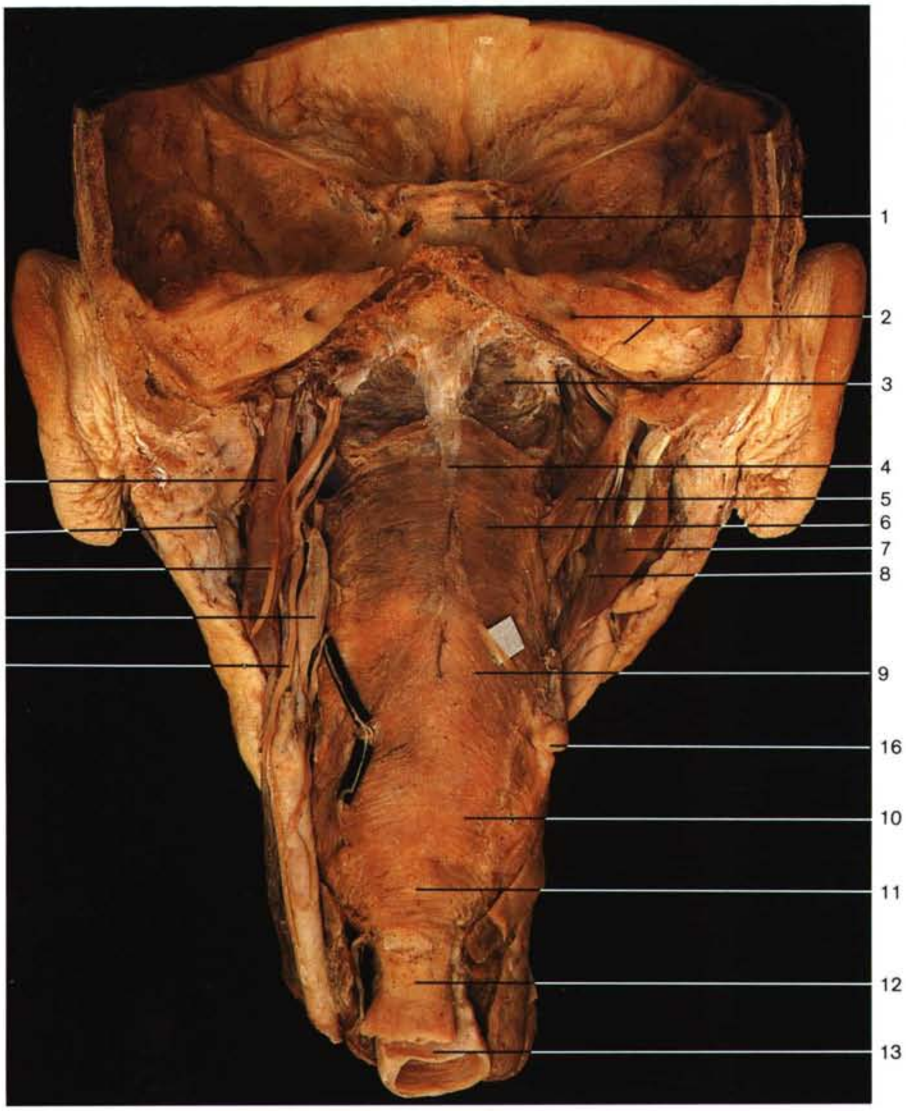

Muscles of the pharynx (posterior aspect).

1 Sella turcica
2 Internal acoustic meatus and petrous part of temporal bone
3 Pharyngobasilar fascia
4 Fibrous raphe of pharynx
5 **Stylopharyngeus muscle**
6 **Superior constrictor muscle of pharynx**
7 Posterior belly of digastric muscle
8 Stylohyoid muscle
9 **Middle constrictor muscle of pharynx**
10 **Inferior constrictor muscle of pharynx**
11 Muscle-free area (Killian's triangle)
12 Esophagus
13 Trachea
14 Thyroid and parathyroid glands
15 Medial pterygoid muscle
16 Greater horn of hyoid bone
17 Internal jugular vein
18 Parotid gland
19 Accessory nerve
20 **Superior cervical ganglion** of sympathetic trunk
21 Vagus nerve
22 Laimer's triangle (area prone to developing diverticula)
23 Orbicularis oculi muscle
24 Nasalis muscle
25 Levator labii superioris and levator labii alaeque nasi muscles
26 Levator anguli oris muscle
27 Orbicularis oris muscle
28 **Buccinator muscle**
29 Depressor labii inferioris muscle
30 **Hyoglossus muscle**
31 Thyrohyoid muscle
32 Thyroid cartilage
33 Cricothyroid muscle
34 Pterygomandibular raphe
35 Tensor veli palatini muscle
36 Levator veli palatini muscle
37 Depressor anguli oris muscle
38 Mentalis muscle
39 Styloglossus muscle

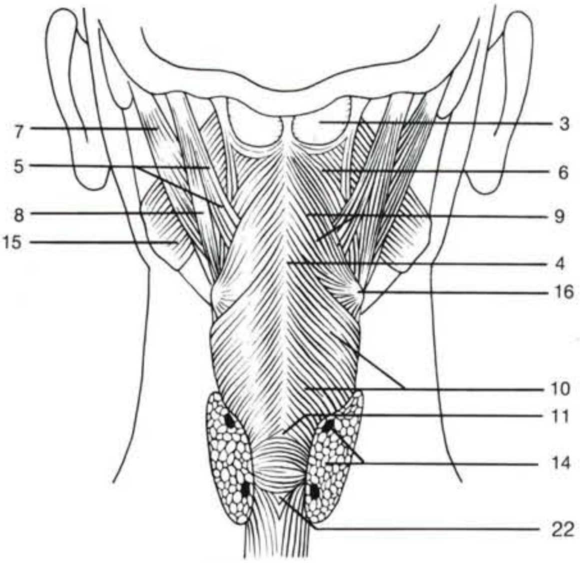

Muscles of the pharynx. (Schematic drawing.)

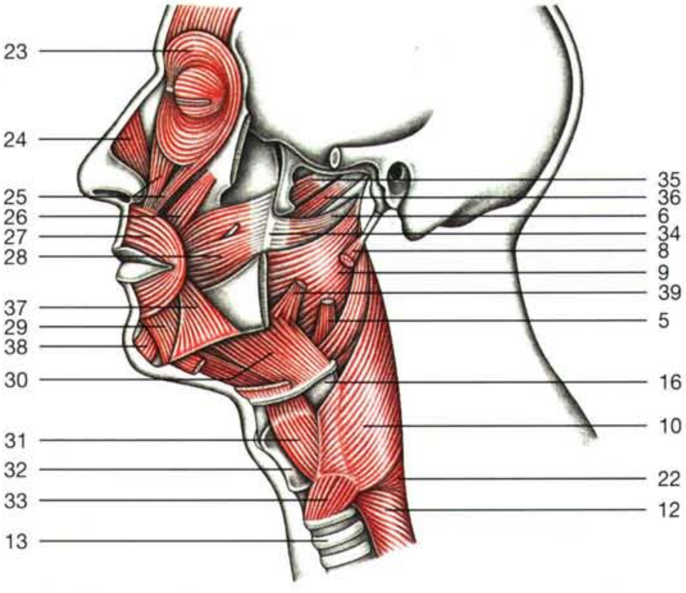

Muscles of pharynx (lateral aspect). (Schematic drawing.)

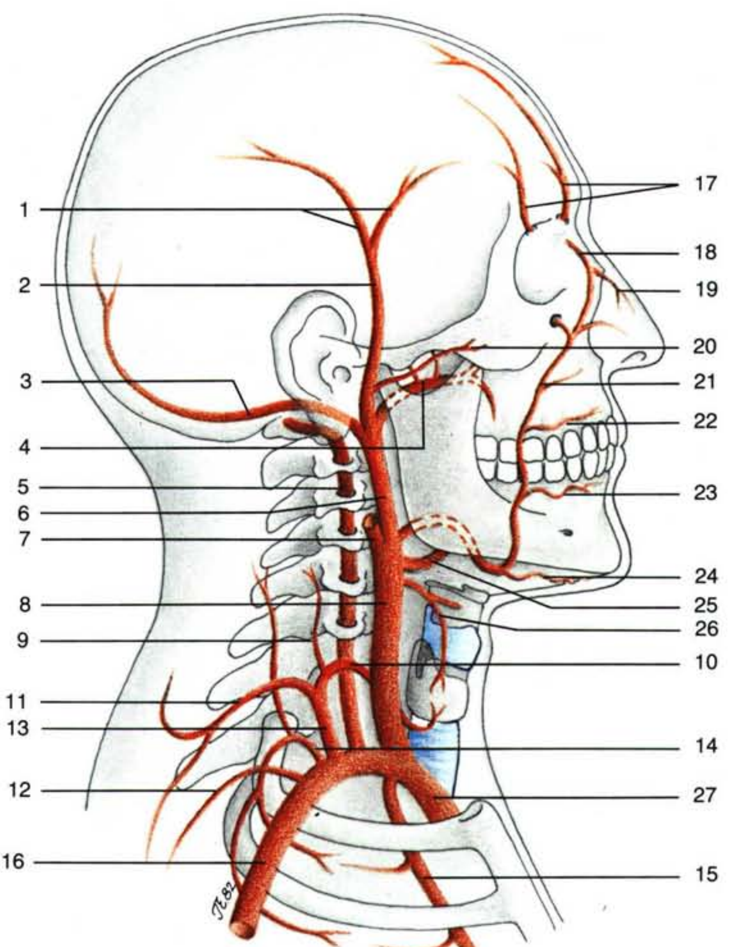

1 Frontal and parietal branches of superficial temporal artery
2 Superficial temporal artery
3 **Occipital artery**
4 **Maxillary artery**
5 **Vertebral artery**
6 **External carotid artery**
7 Internal carotid artery
8 Common carotid artery (divided)
9 Ascending cervical artery
10 Inferior thyroid artery
11 Transverse cervical artery with 2 branches (superficial cervical artery and descending scapular artery)
12 Suprascapular artery
13 **Thyrocervical trunk**
14 Costocervical trunk with 2 branches (deep cervical artery and superior intercostal artery)
15 Internal thoracic artery
16 **Axillary artery**
17 Supraorbital and supratrochlear arteries
18 Angular artery
19 Dorsal nasal artery
20 Transverse facial artery
21 **Facial artery**
22 Superior labial artery
23 Inferior labial artery
24 Submental artery
25 Lingual artery
26 Superior thyroid artery
27 Brachiocephalic trunk

Arteries of head and neck. Diagram of the main branches of external carotid and subclavian artery.

to page 165 ▷

1 Galea aponeurotica
2 Frontal branch } of superficial
3 Parietal branch } temporal artery
4 Superior auricular muscle
5 **Superficial temporal artery and vein**
6 Middle temporal artery
7 Auriculotemporal nerve
8 Branches of facial nerve
9 Facial nerve
10 **External carotid artery** within the retromandibular fossa
11 Posterior belly of digastric muscle
12 Sternocleidomastoid artery
13 Sympathetic trunk and superior cervical ganglion
14 Sternocleidomastoid muscle (divided and reflected)
15 Clavicle (divided)
16 Transverse cervical artery
17 Ascending cervical artery and phrenic nerve
18 Scalenus anterior muscle
19 Suprascapular artery
20 Dorsal scapular artery
21 Brachial plexus and **axillary artery**
22 Thoracoacromial artery
23 Lateral thoracic artery
24 Median nerve (displaced) and pectoralis minor muscle (reflected)
25 Frontal belly of occipitofrontalis muscle
26 Orbital part of orbicularis oculi muscle
27 **Angular artery** and vein
28 **Facial artery**
29 Superior labial artery
30 Zygomaticus major muscle
31 Inferior labial artery
32 Parotid duct
33 Buccal fat pad
34 **Maxillary artery**
35 Masseter muscle
36 Facial artery and mandible
37 Submental artery
38 Anterior belly of digastric muscle

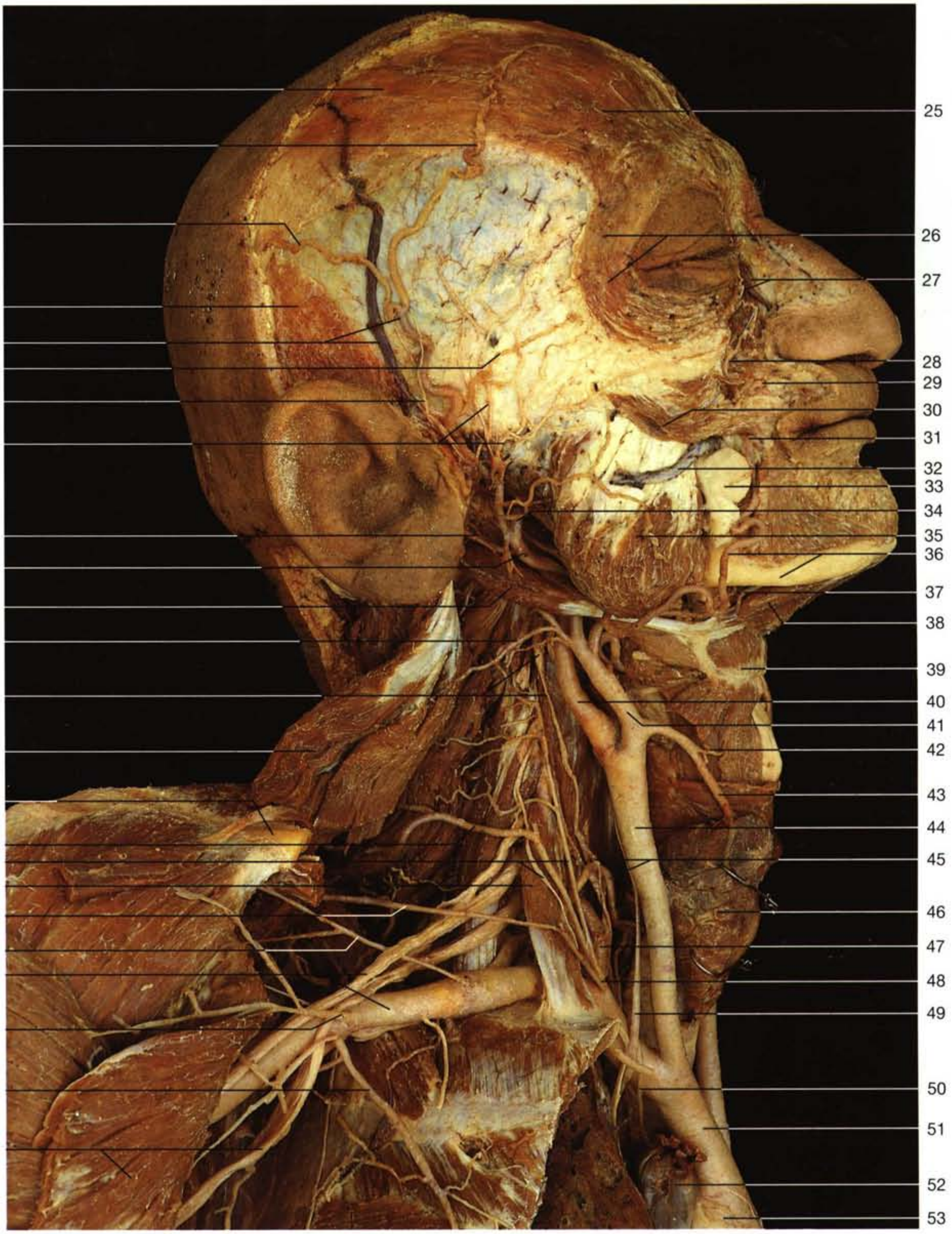

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53

Main branches of head and neck arteries (lateral aspect). Anterior thoracic wall and clavicle partly removed; pectoralis muscles have been reflected to display the subclavian and axillary arteries.

39	Hyoid bone	46	Thyroid gland (right lobe)
40	**Internal carotid artery**	47	**Vertebral artery**
41	**External carotid artery**	48	**Thyrocervical trunk**
42	Superior laryngeal artery	49	Vagus nerve
43	Superior thyroid artery	50	Ansa subclavia of sympathetic trunk
44	**Common carotid artery**	51	**Brachiocephalic trunk**
45	Thyroid ansa of sympathetic trunk and **inferior thyroid artery**	52	Superior vena cava (divided)
		53	Aortic arch

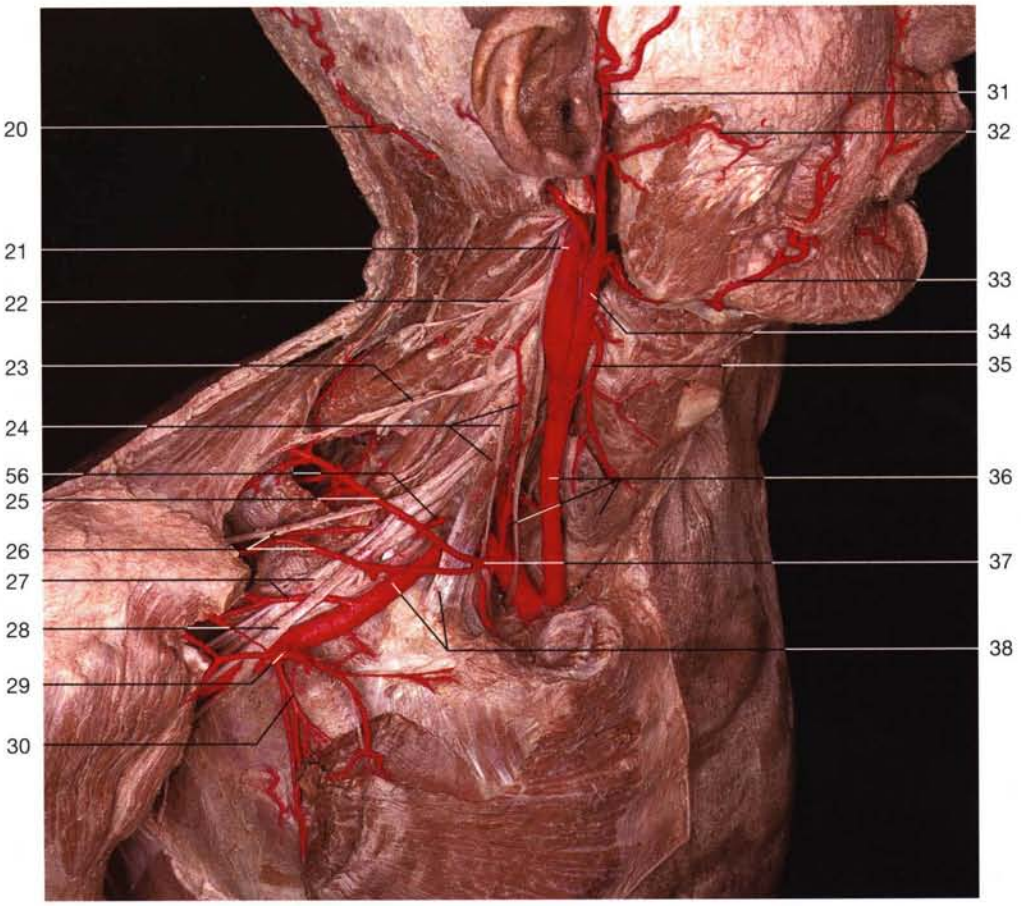

Arteries of head and neck (anterior-lateral aspect). Clavicle, sternocleidomastoid and veins have been partly removed, the arteries were colored.

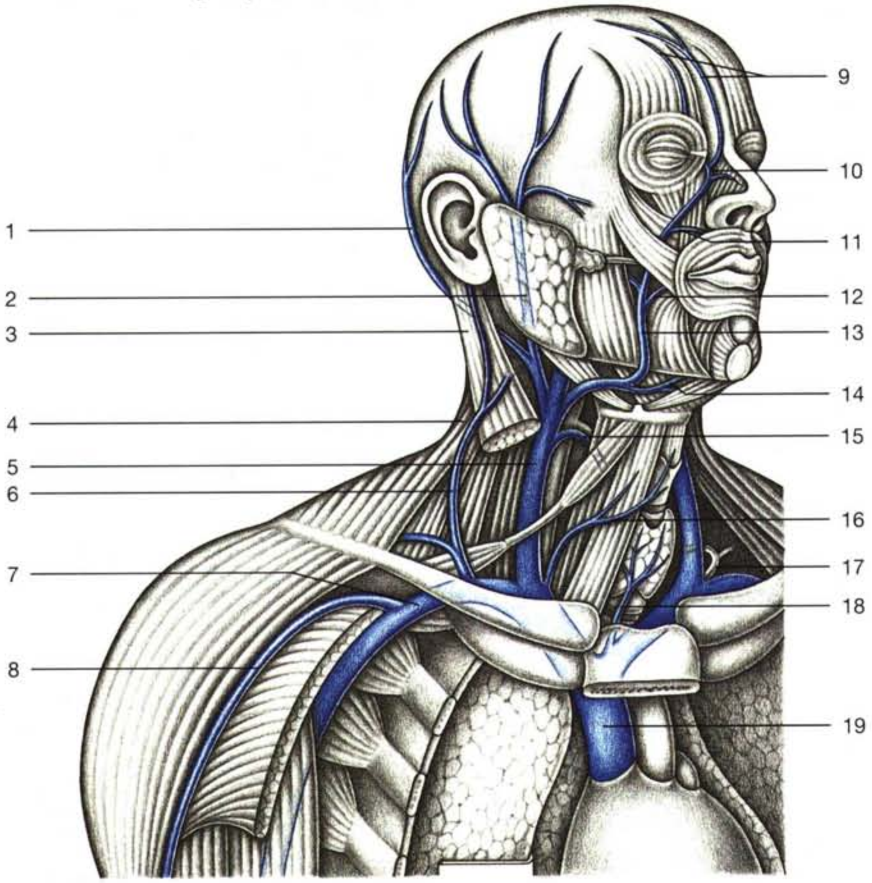

Veins of head and neck. Sternocleidomastoid muscle and anterior thoracic wall partly removed. Note the venous connection with the superior vena cava.

1 Occipital vein
2 Superficial temporal vein
3 Sternocleidomastoid muscle
4 Trapezius muscle
5 **Internal jugular vein**
6 External jugular vein
7 **Subclavian vein**
8 Cephalic vein
9 Supraorbital veins
10 Angular vein
11 Superior labial vein
12 Inferior labial vein
13 **Facial vein**
14 Submental vein
15 Superior thyroid vein
16 Anterior jugular vein
17 Thoracic duct
18 Inferior thyroid vein
19 Superior vena cava
20 Occipital artery
21 **Internal carotid artery**
22 Cervical plexus
23 Supraclavicular nerve
24 Phrenic nerve and ascending cervical artery on scalenus anterior muscle
25 Superficial cervical artery
26 Suprascapular artery and nerve
27 Brachial plexus and anterior circumflex humeral artery
28 Lateral cord of brachial plexus
29 **Thoracoacromial artery**
30 Lateral thoracic artery
31 Superficial temporal artery
32 Transverse facial artery
33 Facial artery
34 External carotid artery
35 Superior thyroid artery
36 **Common carotid artery, vagus nerve** and thyroid gland
37 **Thyrocervical trunk**
38 **Subclavian artery** and scalenus anterior muscle
39 Parotid gland and facial nerve
40 Great auricular nerve
41 External jugular vein
42 Brachial plexus
43 Cephalic vein in deltopectoral groove
44 Axillary vein and artery
45 Right brachiocephalic vein
46 **Superior vena cava**
47 Right lung (reflected)
48 Superficial temporal artery and vein
49 Facial artery and vein
50 Cervical branch of facial nerve and submandibular gland
51 **Internal jugular vein,** common carotid artery and omohyoid muscle
52 Anterior jugular vein and thyroid gland
53 Jugular venous arch
54 Left brachiocephalic vein
55 Pericardium of heart (location of right atrium)
56 Transverse cervical artery

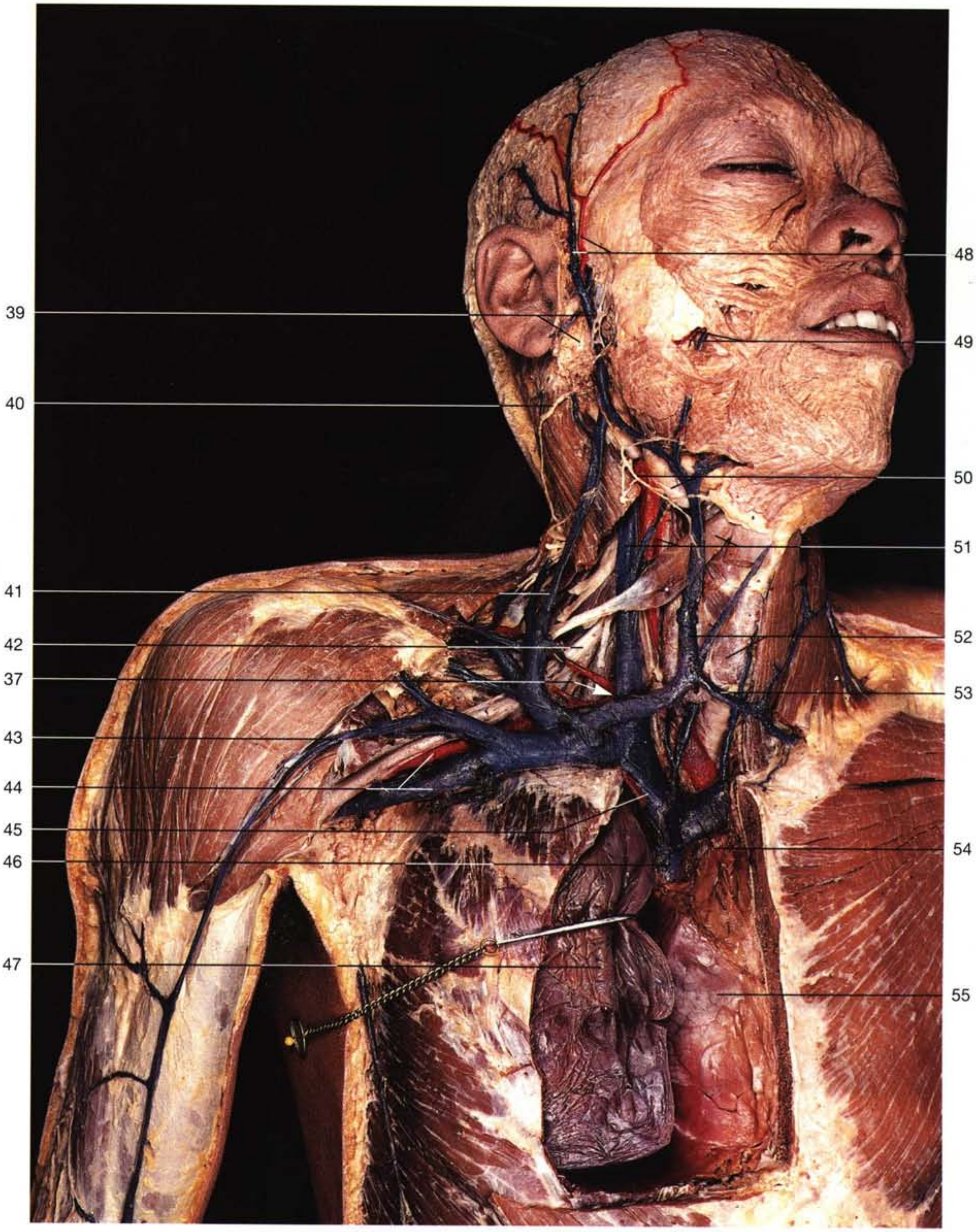

Veins of head and neck (anterior aspect). Part of the thoracic wall, clavicle and sternocleidomastoid muscle have been removed. Veins were colored blue, arteries red.

The **internal jugular vein** is the continuation of the sigmoid sinus which drains most of the venous blood from the brain together with the external cerebrospinal fluid. By joining the subclavian vein it forms the right brachiocephalic vein which continues on the right side directly into the superior vena cava. The common way to introduce the lead from a pacemaker device into the heart is by way of the cephalic vein. On the left side the thoracic duct joins the internal jugular vein at that point where the subclavian vein and the internal jugular vein form the left brachiocephalic vein. Note that the subclavian vein lies in front of the scalenus anterior muscle whereas the subclavian artery together with the plexus brachialis lies posterior to that muscle. The cephalic vein joins the axillary vein by passing into the deltopectoral triangle. The subclavian vein is strongly fixed to the first rib so that it can be punctured with a needle at that point (underneath the sternal end of clavicle) to introduce a catheter (subclavian line).

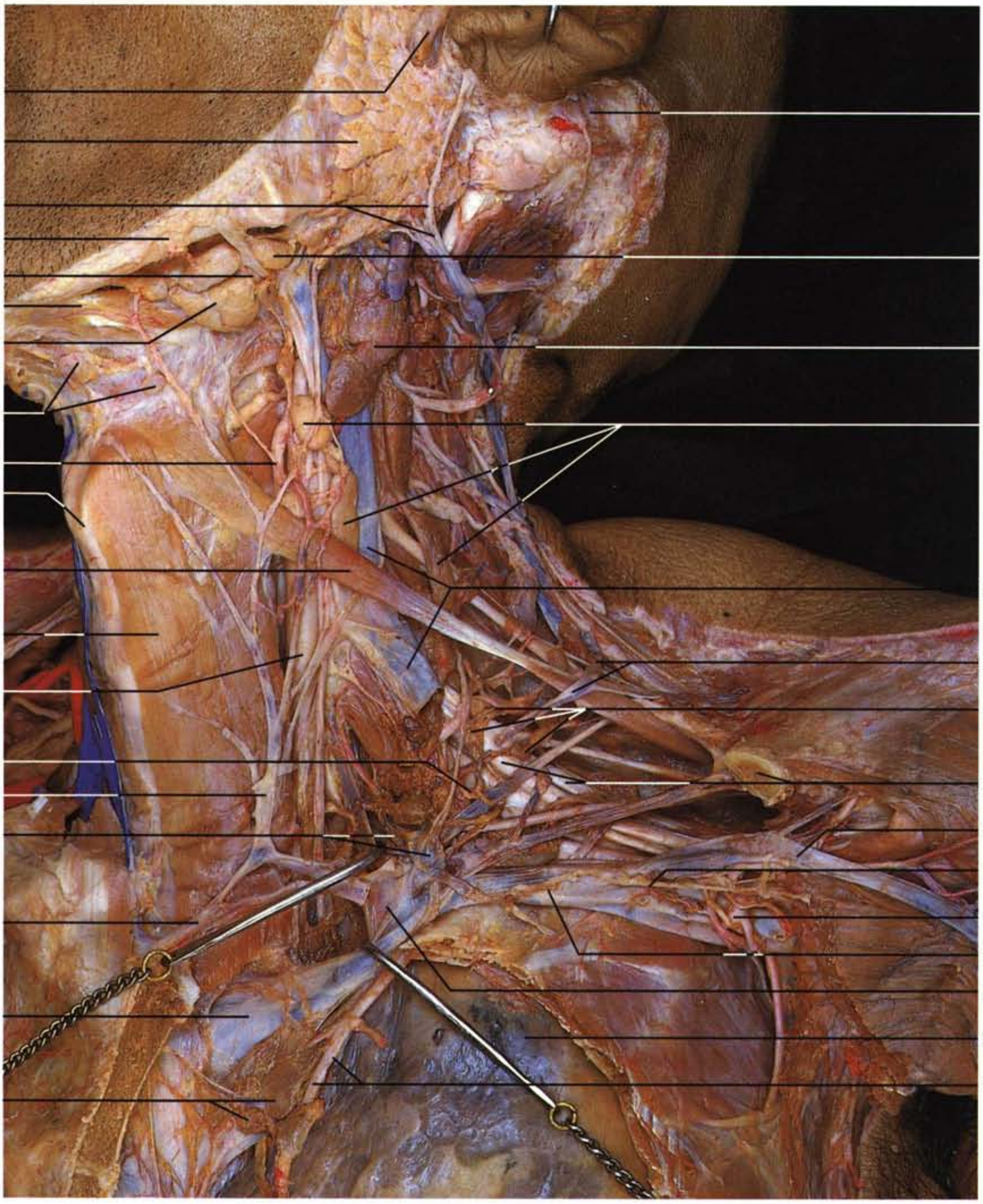

Lymph nodes and lymph vessels of the neck, left side oblique (lateral aspect). The sternocleidomastoid muscle and the left half of the thoracic wall have been removed. Lower part of the internal jugular vein has been cut and laterally displaced to show the thoracic duct.

1 **Superficial parotid lymph node**
2 Parotid gland
3 Great auricular nerve
4 Mandible
5 Facial vein
6 Anterior belly of digastric muscle
7 Submandibular gland
8 **Submental lymph nodes**
9 Superior thyroid artery
10 Thyroid cartilage
11 Omohyoid muscle
12 Sternohyoid muscle
13 Common carotid artery
14 **Supraclavicular lymph nodes**
15 Anterior jugular vein
16 Thoracic duct
17 Jugular venous arch
18 Left brachiocephalic vein
19 **Superior mediastinal lymph nodes**
20 Retroauricular lymph nodes
21 **Submandibular nodes**
22 **Superficial cervical lymph nodes**
23 **Jugulodigastric lymph nodes and jugular trunk**
24 **Internal jugular vein**
25 External jugular vein
26 **Jugulo-omohyoid lymph nodes**
27 Brachial plexus
28 Cephalic vein
29 Subclavian trunk
30 **Infraclavicular lymph nodes**
31 Subclavian vein
32 Lung
33 Internal thoracic artery and vein

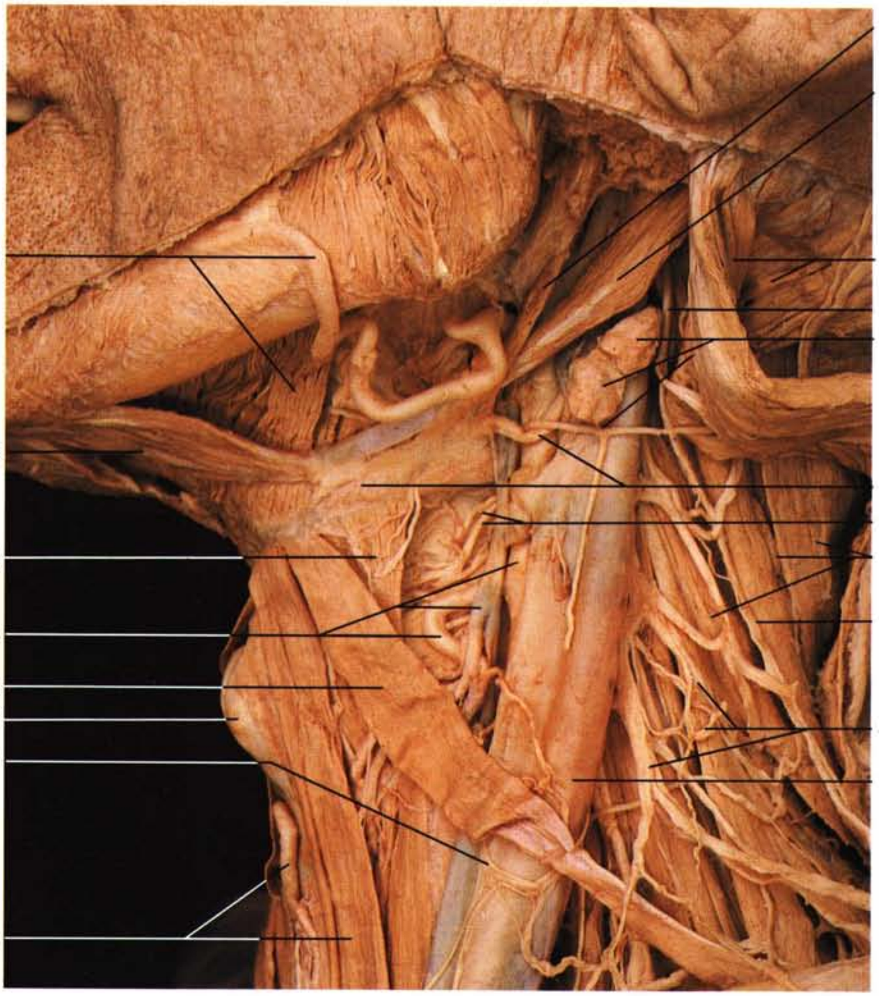

Carotid triangle, left side (lateral aspect). Sternocleidomastoid muscle reflected.

1 Mylohyoid muscle and facial artery
2 Anterior belly of digastric muscle
3 Thyrohyoid
4 **External carotid artery,** superior thyroid artery and vein
5 Omohyoid muscle
6 Thyroid cartilage
7 **Ansa cervicalis**
8 Sternohyoid muscle and superior thyroid artery
9 Stylohyoid muscle
10 Posterior belly of digastric muscle
11 Sternocleidomastoid muscle (reflected)
12 **Superior cervical lymph nodes** and sternocleidomastoid artery
13 Hyoid bone and hypoglossal nerve (n. XII)
14 Splenius capitis and levator scapulae muscles
15 Superior laryngeal artery and internal branch of superior laryngeal nerve
16 **Accessory nerve**
17 Cervical plexus
18 **Internal jugular vein**
19 Facial vein
20 Submental nodes
21 **Thoracic duct**
22 Retroauricular nodes
23 Parotid nodes
24 Occipital nodes
25 **Submandibular nodes**
26 Jugulodigastric nodes } **deep cervical**
27 Jugulo-omohyoid nodes } **nodes**
28 Jugular trunk
29 Subclavian trunk
30 Infraclavicular nodes
31 External jugular vein

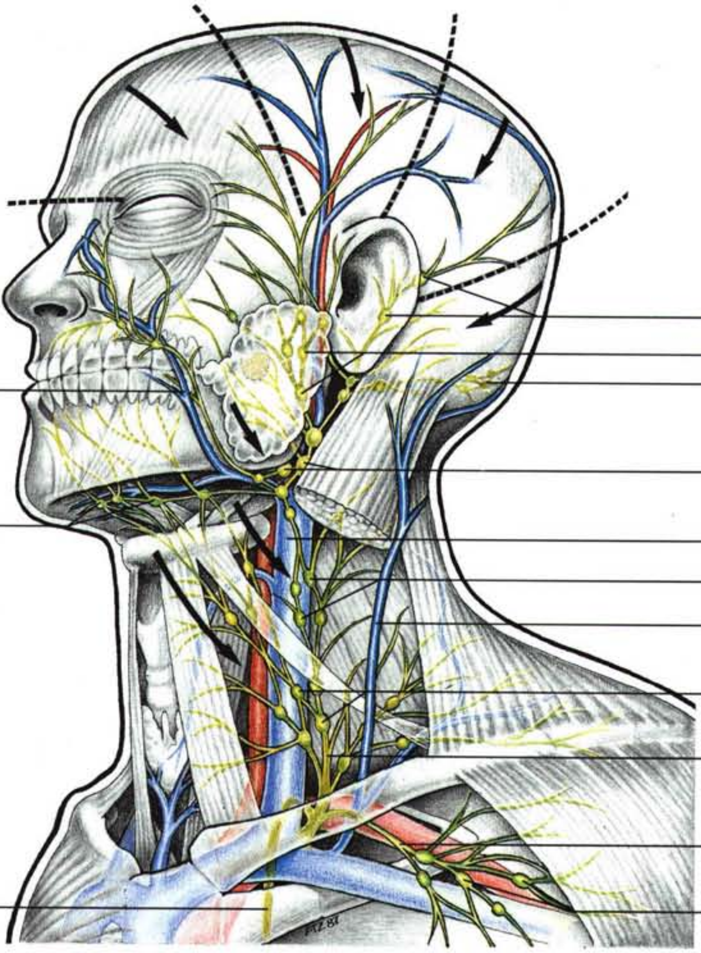

Lymph nodes and veins of head and neck. Dotted lines = border between irrigation areas; arrows: direction of lymph flow.

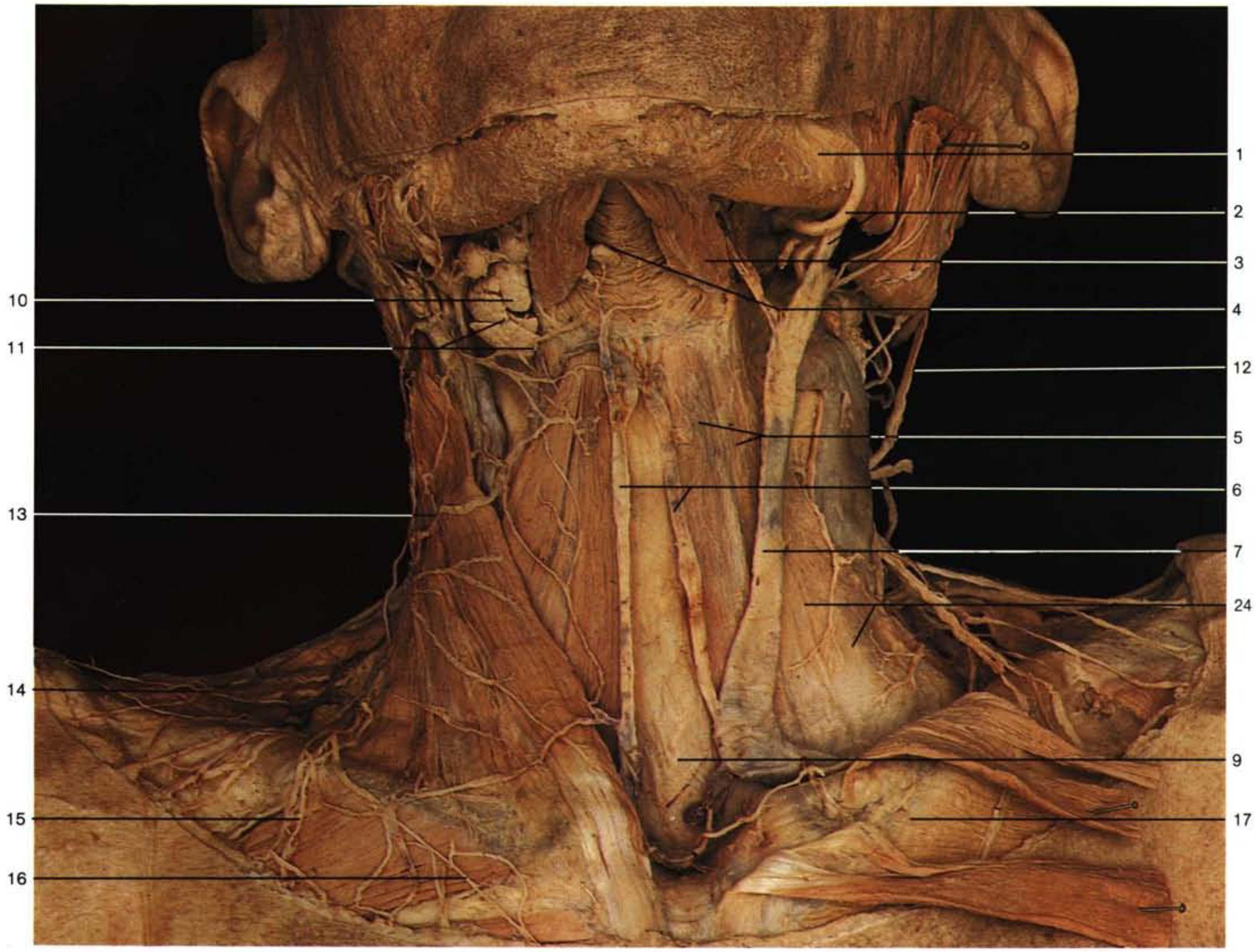

Neck (anterior aspect).
The superficial fascia has been removed.

1 Mandible
2 **Facial artery and vein**
3 Anterior belly of digastric muscle
4 Mylohyoid muscle
5 Infrahyoid muscles (sternohyoid, sternothyroid and omohyoid)
6 **Anterior jugular veins**
7 **External jugular vein**
8 Sternocleidomastoid muscle
9 **Thyroid gland**
10 **Submandibular gland**
11 Cervical branch of facial nerve
12 Great auricular nerve
13 Transverse cervical nerves
14 Lateral supraclavicular nerves
15 Middle supraclavicular nerves
16 Medial supraclavicular nerves
} Cutaneous branches of cervical plexus (12–16)
17 Clavicle
18 Platysma muscle
19 Prevertebral lamina of cervical fascia, covering longus colli muscle
20 Vertebral artery and vein
21 Scalenus muscles
22 Trapezius muscle
23 **Superficial lamina of cervical fascia**
24 **Pretracheal lamina of cervical fascia**
25 **Prevertebral lamina of cervical fascia** with sympathetic trunk
26 **Carotid sheath** with common carotid artery, internal jugular vein and vagus nerve
27 Cervical part of sympathetic trunk
28 Carotid sheath

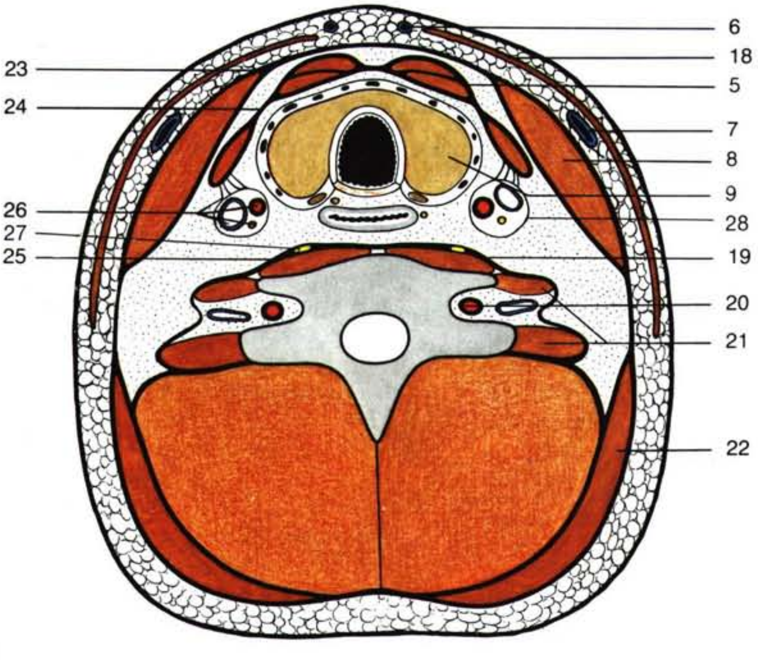

Cross-section of the neck at the level of the thyroid gland.
Notice the postion of the three laminae of cervical fascia (23, 24, 25).

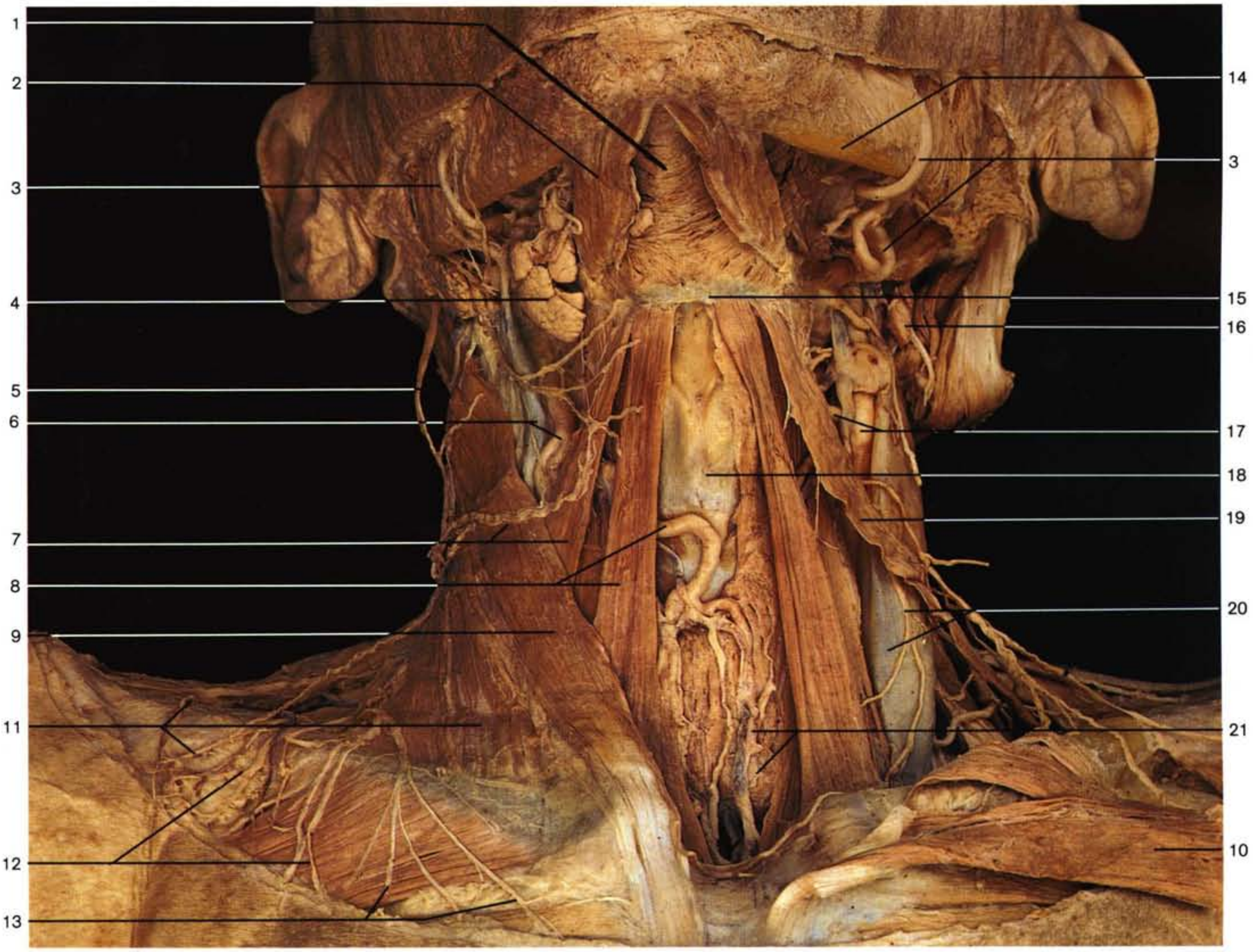

Anterior triangle (anterior aspect). The pretracheal lamina of cervical fascia and left sternocleidomastoid muscle have been removed.

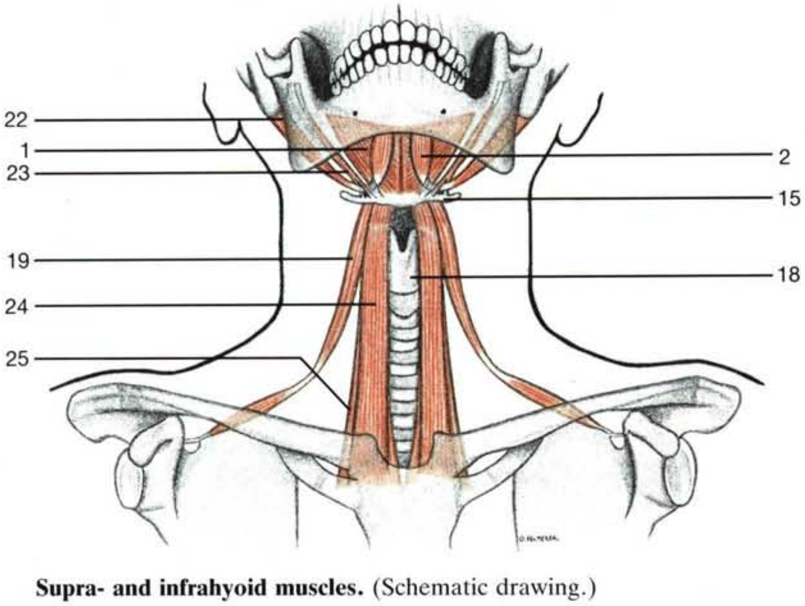

Supra- and infrahyoid muscles. (Schematic drawing.)

1 Mylohyoid muscle
2 Anterior belly of digastric muscle
3 **Facial artery**
4 **Submandibular gland**
5 Great auricular nerve
6 Internal jugular vein and common carotid artery
7 Transverse cervical nerve and omohyoid muscle
8 **Sternohyoid muscle** and **superior thyroid artery**
9 Sternocleidomastoid muscle (sternal head)
10 Left sternocleidomastoid muscle (reflected)
11 Sternocleidomastoid muscle (clavicular head) and lateral supraclavicular nerves
12 Middle **supraclavicular nerves**
13 Medial supraclavicular nerves
14 Mandible
15 **Hyoid bone**
16 Superficial cervical lymph nodes
17 Left superior thyroid artery and external carotid artery
18 **Thyroid cartilage**
19 **Omohyoid muscle** (superior belly)
20 **Internal jugular vein** and branches of ansa cervicalis
21 Thyroid gland and unpaired inferior thyroid vein
22 Posterior belly of digastric muscle
23 **Stylohyoid muscle**
24 **Sternohyoid muscle**
25 **Sternothyroid muscle**

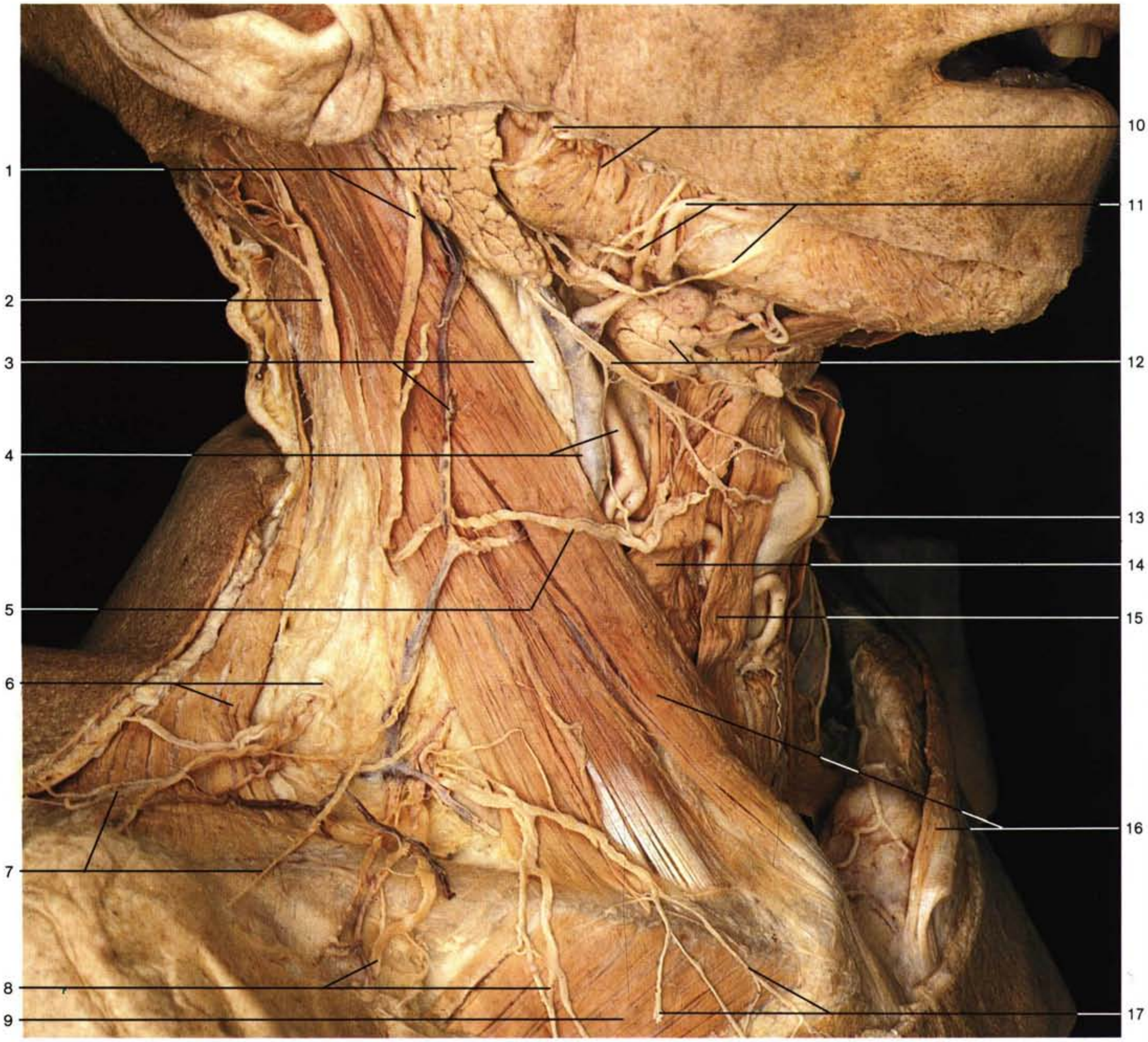

Posterior and carotid triangles (lateral aspect).
Superficial dissection.

1 **Parotid gland** and **great auricular nerve**
2 **Lesser occipital nerve**
3 Internal and external jugular veins
4 Retromandibular vein and external carotid artery
5 **Transverse cervical nerve** with communicating branch to cervical branch of facial nerve
6 Trapezius muscle and superficial lamina of cervical fascia
7 Lateral supraclavicular nerves
8 Middle **supraclavicular nerves**
9 Pectoralis major muscle
10 Buccal branch of facial nerve and masseter
11 **Facial artery and vein** and mandibular branch of facial nerve
12 Cervical branch of facial nerve and submandibular gland
13 Thyroid cartilage
14 Omohyoid muscle
15 Sternohyoid muscle
16 Sternocleidomastoid muscle
17 Medial supraclavicular nerves
18 Mandibular branch of facial nerve
19 Cervical branch of facial nerve with communicating branch to transverse cervical nerve

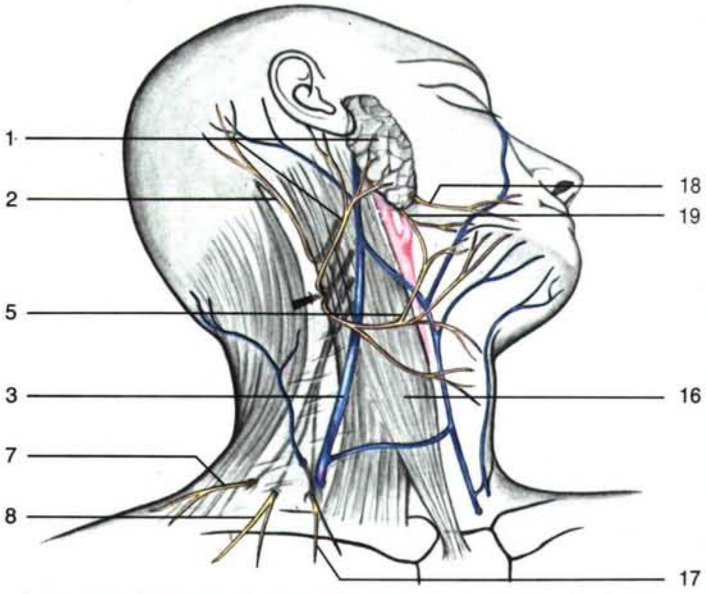

Cutaneous branches of cervical plexus. Erb's point is indicated by an arrowhead. (Schematic diagram.)

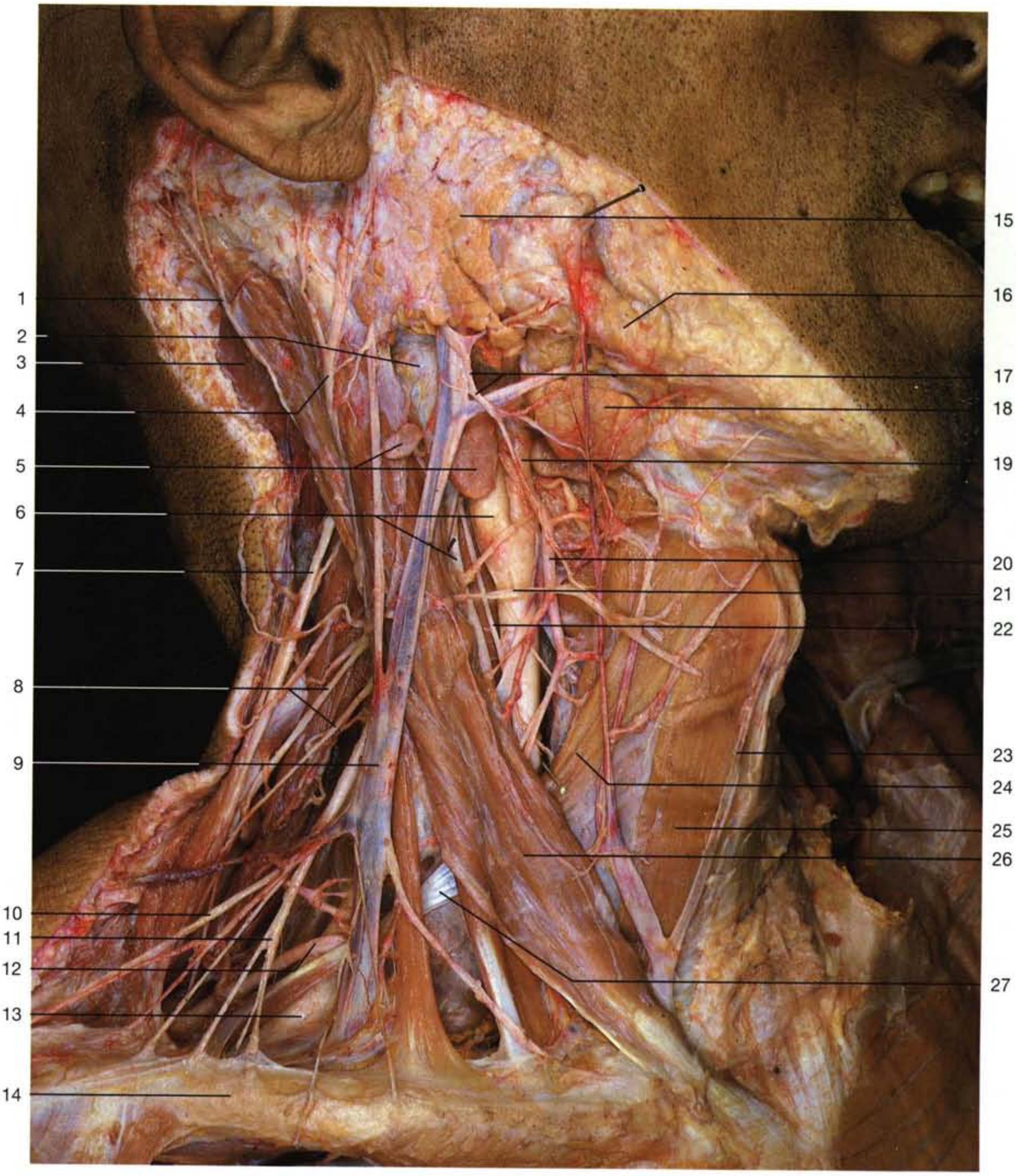

Posterior and carotid triangles (lateral aspect). Superficial dissection. The superficial lamina of cervical fascia has been removed to display the cutaneous branches of the cervical plexus and subcutaneous veins.

1 **Lesser occipital nerve**
2 **Internal jugular vein**
3 Splenius capitis muscle
4 **Great auricular nerve**
5 Submandibular nodes
6 Internal carotid artery and vagus nerve
7 **Accessory nerve**
8 Muscular branches of cervical plexus
9 External jugular vein
10 **Posterior supraclavicular nerves**
11 **Middle supraclavicular nerves**
12 Suprascapular artery
13 Pretracheal lamina of fascia of neck
14 Clavicle
15 Parotid gland
16 Mandible
17 Cervical branch of facial nerve
18 Submandibular gland
19 External carotid artery
20 Superior thyroid artery
21 Transverse cervical nerve
22 Superior root of ansa cervicalis
23 Anterior jugular vein
24 Omohyoid muscle
25 Sternohyoid muscle
26 Sternocleidomastoid muscle
27 Intermediate tendon of omohyoid muscle

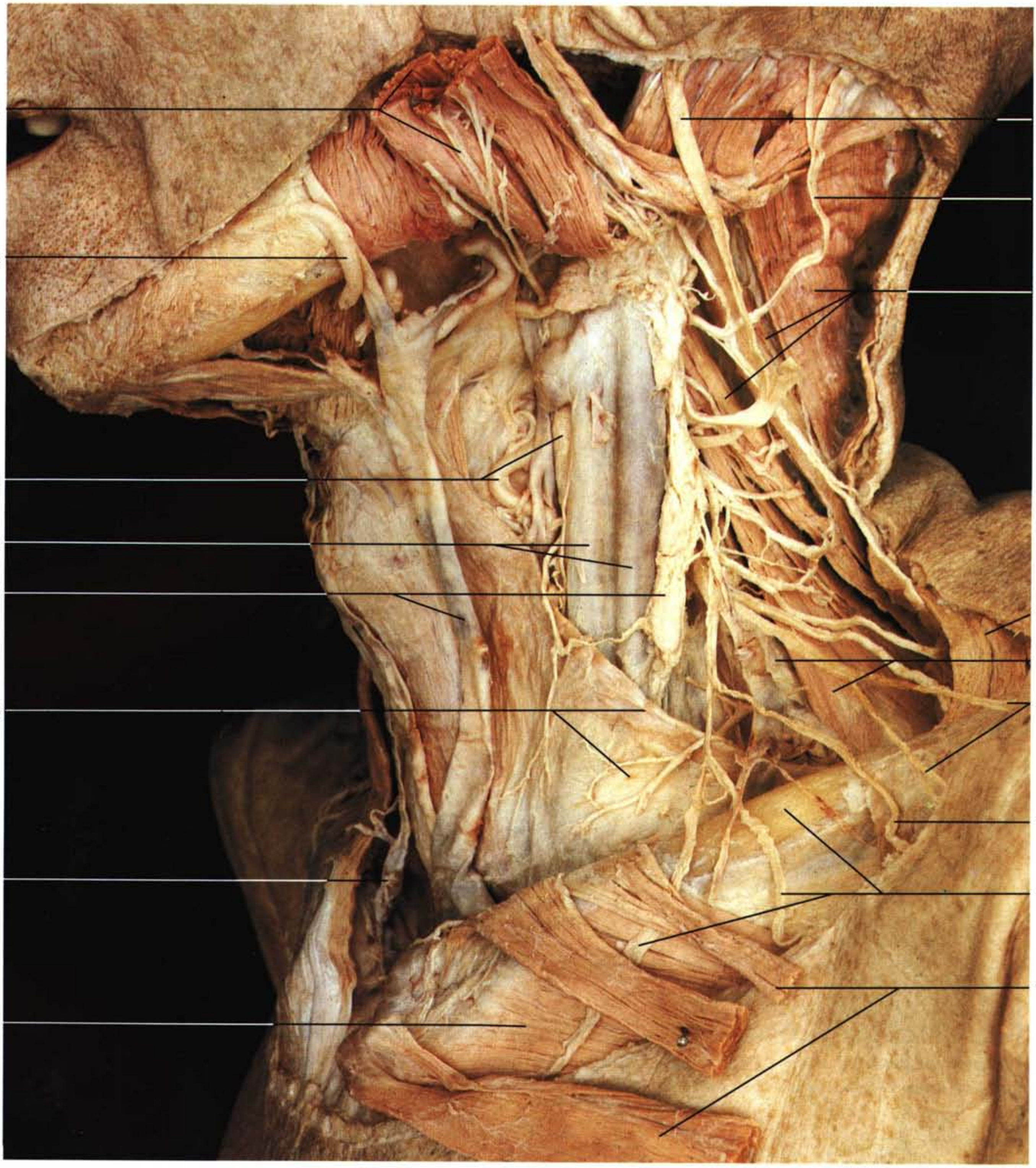

Neck, superficial dissection (lateral aspect). Sternocleidomastoid muscle has been cut and reflected to display the pretracheal lamina of the cervical fascia.

1 Sternocleidomastoid muscle (reflected) and branch of accessory nerve
2 Facial artery
3 External carotid artery and superior thyroid artery
4 **Internal jugular vein**
5 **Deep cervical lymph nodes** and external jugular vein
6 Omohyoid muscle and **pretracheal lamina of cervical fascia**
7 Anterior jugular vein
8 Pectoralis major muscle
9 **Great auricular nerve**
10 **Lesser occipital nerve**
11 Splenius capitis and levator scapulae muscles
12 Trapezius muscle
13 Scalenus medius muscle and brachial plexus
14 Posterior **supraclavicular nerves**
15 Middle supraclavicular nerve
16 Clavicle and anterior supraclavicular nerves
17 Sternocleidomastoid muscle (reflected)

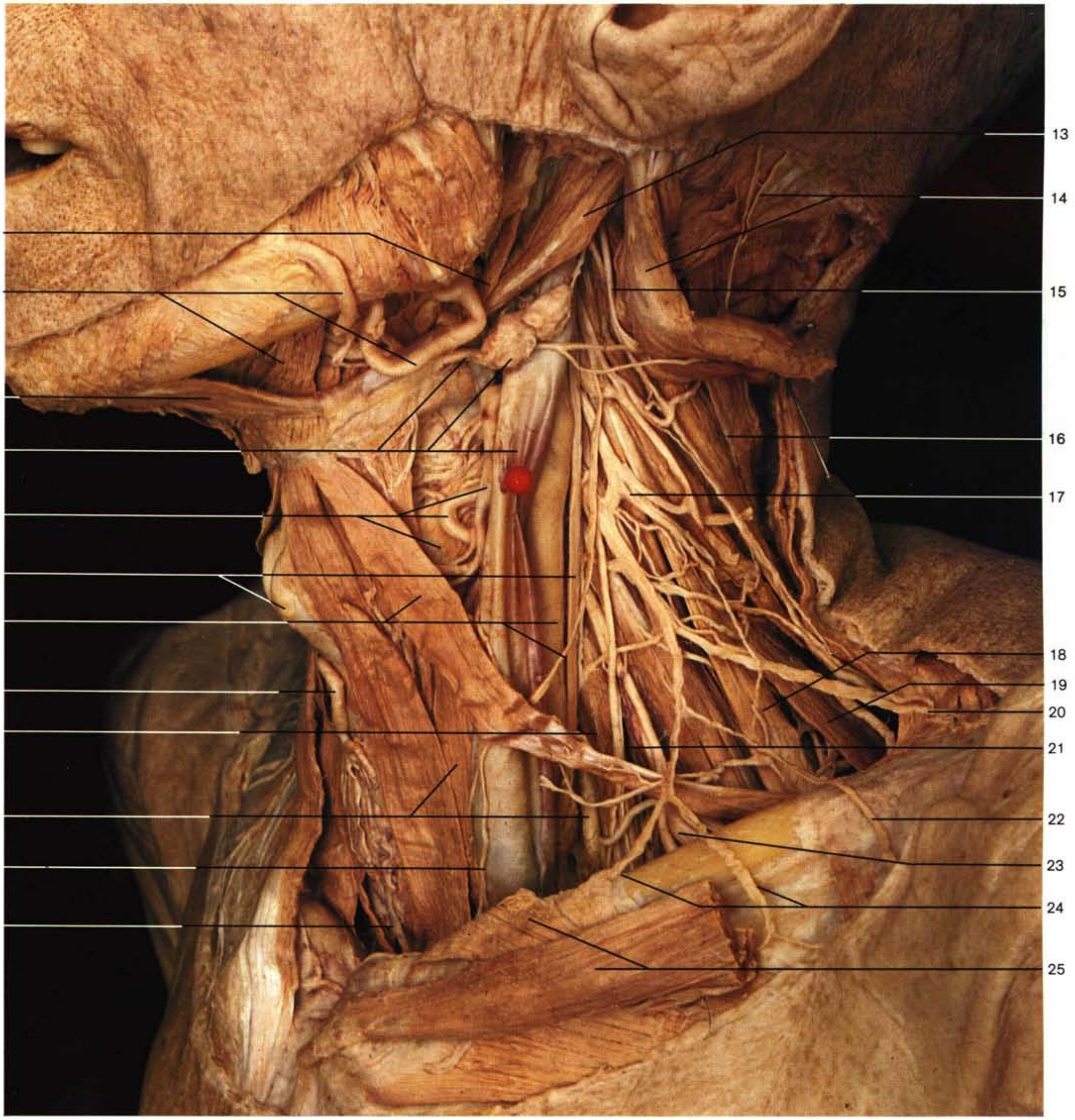

Neck, deep dissection (lateral aspect). The internal jugular vein has been reflected to expose the carotid artery and vagus nerve.

1 Stylohyoid muscle
2 Facial artery and mylohyoid muscle
3 Anterior belly of digastric muscle
4 **Internal jugular vein,** hypoglossal nerve, and superficial cervical lymph nodes
5 **Superior thyroid artery and vein** and inferior pharyngeal constrictor muscle
6 Thyroid cartilage and vagus nerve
7 **Ansa cervicalis,** omohyoid muscle, and common carotid artery
8 Right superior thyroid artery
9 Scalenus anterior muscle
10 Sternothyroid muscle and inferior thyroid artery
11 Muscular branches of ansa cervicalis to the infrahyoid muscles
12 Inferior thyroid vein

13 Posterior belly of digastric muscle
14 Sternocleidomastoid muscle and lesser occipital nerve
15 **Accessory nerve**
16 Splenius capitis muscle
17 **Cervical plexus**
18 Scalenus posterior muscle
19 Levator scapulae muscle
20 Posterior **supraclavicular nerves**
21 Phrenic nerve
22 Middle supraclavicular nerve
23 **Brachial plexus**
24 Anterior supraclavicular nerves
25 Sternocleidomastoid muscle

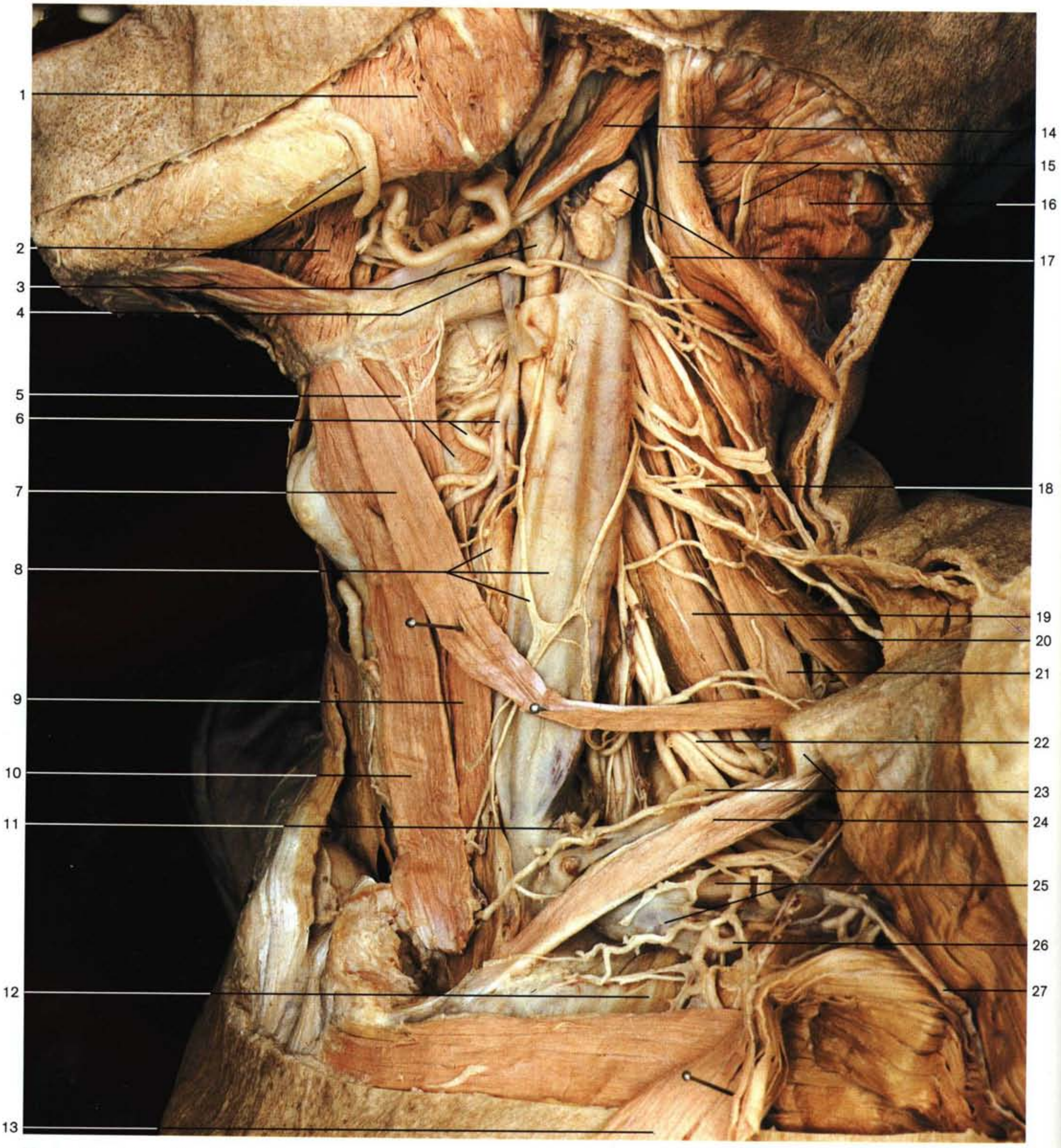

Neck, deeper dissection (lateral aspect). **Ansa cervicalis.** The cervical fascia and the clavicle are partly removed. Ansa cervicalis and infrahyoid muscles are displayed.

1 Masseter muscle
2 Mylohyoid muscle and facial artery
3 **External carotid artery** and anterior belly of digastric muscle
4 **Hypoglossal nerve**
5 Thyrohyoid muscle
6 Superior thyroid artery and vein and inferior pharyngeal constrictor muscle
7 Omohyoid muscle (superior belly)
8 **Ansa cervicalis,** thyroid gland and **internal jugular vein**
9 Sternothyroid muscle
10 Sternohyoid muscle
11 **Thoracic duct**
12 Pectoralis minor muscle
13 Pectoralis major muscle
14 Posterior belly of digastric muscle
15 Sternocleidomastoid muscle and lesser occipital nerve
16 Splenius capitis muscle
17 Superficial cervical lymph nodes and accessory nerve
18 **Cervical plexus**
19 Scalenus medius muscle
20 Levator scapulae muscle
21 Scalenus posterior muscle
22 Brachial plexus
23 Transverse cervical artery and clavicle
24 Subclavius muscle
25 **Subclavian artery and vein**
26 Thoracoacromial artery
27 Cephalic vein

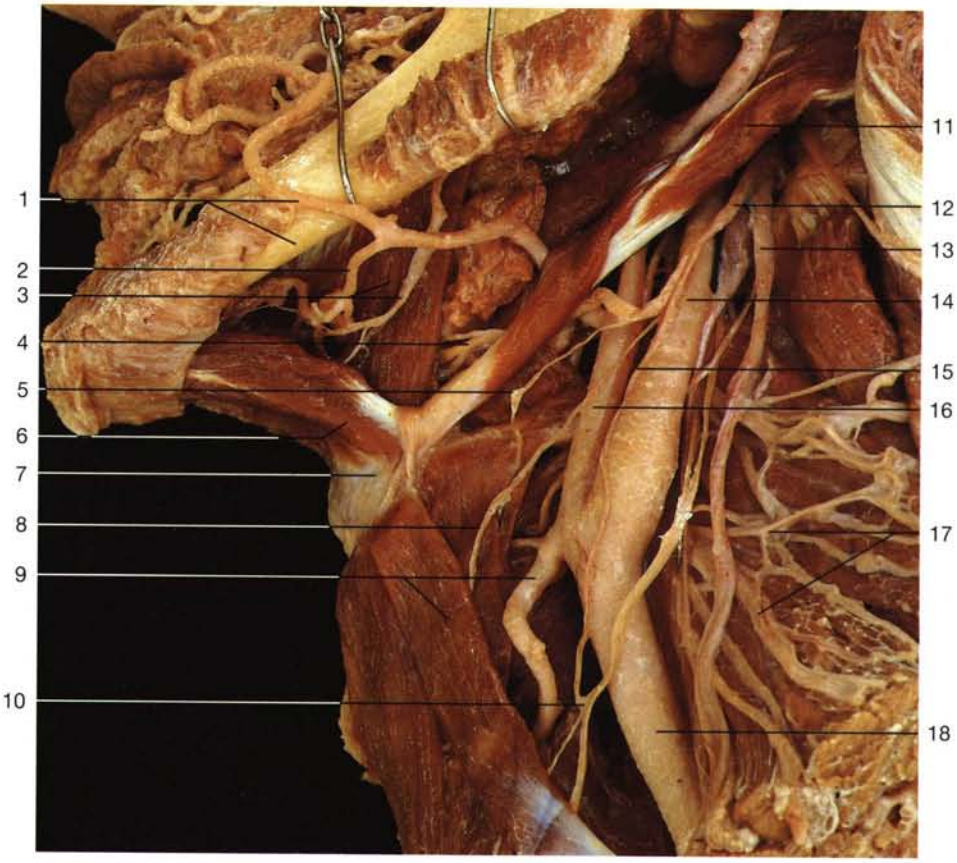

1 **Facial artery** and mandible
2 Submental artery
3 Mylohyoid muscle and nerve
4 **Hypoglossal nerve** (lingual branches)
5 Thyrohyoid branch of hypoglossal nerve (n. XII)
6 Anterior belly of digastric muscle
7 Hyoid bone
8 Omohyoid branch of hypoglossal nerve (n. XII)
9 Omohyoid muscle and superior thyroid artery
10 **Ansa cervicalis**
11 Posterior belly of digastric muscle
12 **Hypoglossal nerve** (n. XII)
13 **Vagus nerve** (n. X)
14 Internal carotid artery
15 Superior root of ansa cervicalis
16 **External carotid artery**
17 Cervical plexus
18 Common carotid artery

Neck, submandibular region (lateral aspect). **Hypoglossal nerve** (n. XII). Mandible slightly elevated.

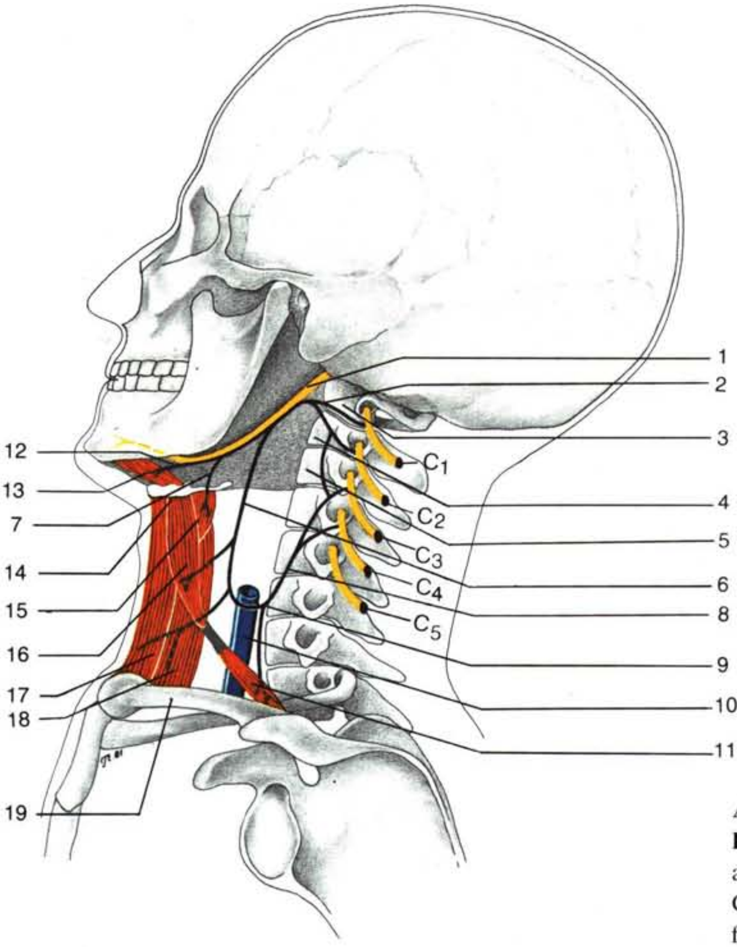

1 **Hypoglossal nerve** (n. XII)
2 Communication from the ventral ramus of the first cervical spinal nerve
3 Atlas
4 Axis
5 Third cervical vertebra
6 Superior root of ansa cervicalis
7 Thyrohyoid branch of hypoglossal nerve
8 Inferior root of ansa cervicalis
9 **Ansa cervicalis**
10 Internal jugular vein
11 Inferior belly of omohyoid muscle
12 Geniohyoid branch of hypoglossal nerve
13 Geniohyoid muscle
14 Hyoid bone
15 Thyrohyoid muscle
16 Superior belly of omohyoid muscle
17 Sternohyoid muscle
18 Sternothyroid muscle
19 Clavicle

Ansa cervicalis.
Innervation of infrahyoid muscles. Cervical plexus and its communication with the hypoglossal nerve. C_1–C_4 = ventral rami of cervical spinal nerves of the first four segments.

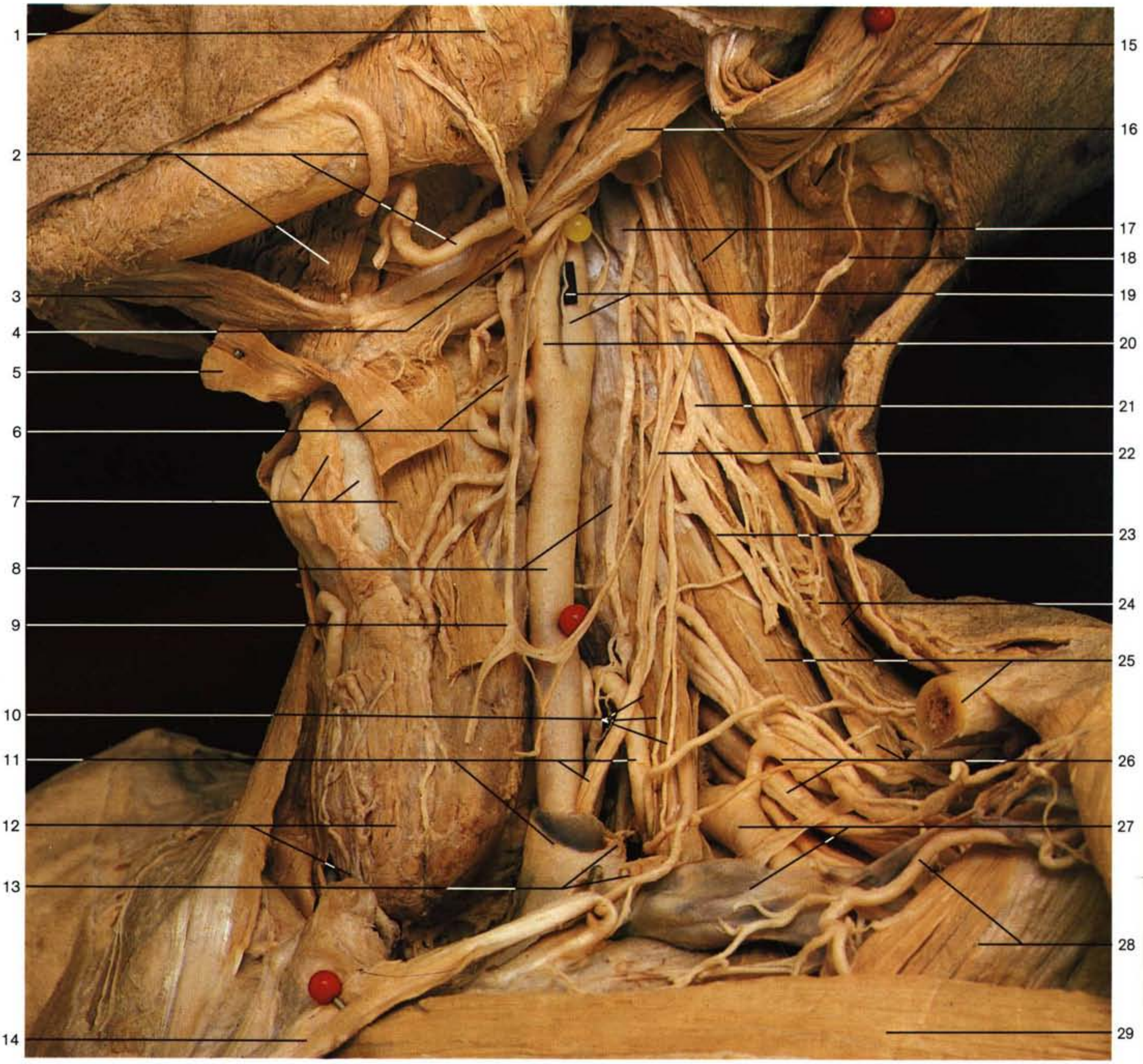

Neck, deep dissection (lateral aspect). Clavicle partly removed to show the slit between the scalenus muscles. Internal jugular vein removed.

1 Masseter muscle
2 Mylohyoid muscle and **facial artery**
3 Anterior belly of digastric muscle
4 **Hypoglossal nerve**
5 Sternohyoid muscle
6 Omohyoid muscle, superior thyroid artery and vein
7 Sternothyroid muscle, thyroid cartilage, and pyramidal lobe of thyroid gland
8 Common carotid artery and sympathetic trunk
9 **Ansa cervicalis**
10 **Phrenic nerve,** ascending cervical artery and anterior scalenus muscle
11 Inferior thyroid artery, vagus nerve and internal jugular vein (cut)
12 **Thyroid gland** and unpaired inferior thyroid venous plexus
13 Thoracic duct and left subclavian trunk
14 Subclavius muscle (reflected)
15 Sternocleidomastoid muscle (reflected)
16 Posterior belly of digastric muscle
17 **Superior cervical ganglion** and splenius muscle
18 **Lesser occipital nerve**
19 Internal carotid artery and branch of the glossopharyngeal nerve to the carotid body
20 **External carotid artery**
21 **Cervical plexus** and **accessory nerve**
22 Inferior root of ansa cervicalis
23 Supraclavicular nerve
24 Levator scapulae muscle
25 Scalenus medius muscle and clavicle
26 Transverse cervical artery, brachial plexus, and scalenus posterior muscle
27 **Subclavian artery and vein**
28 Thoracoacromial artery and pectoralis minor muscle
29 Pectoralis major muscle

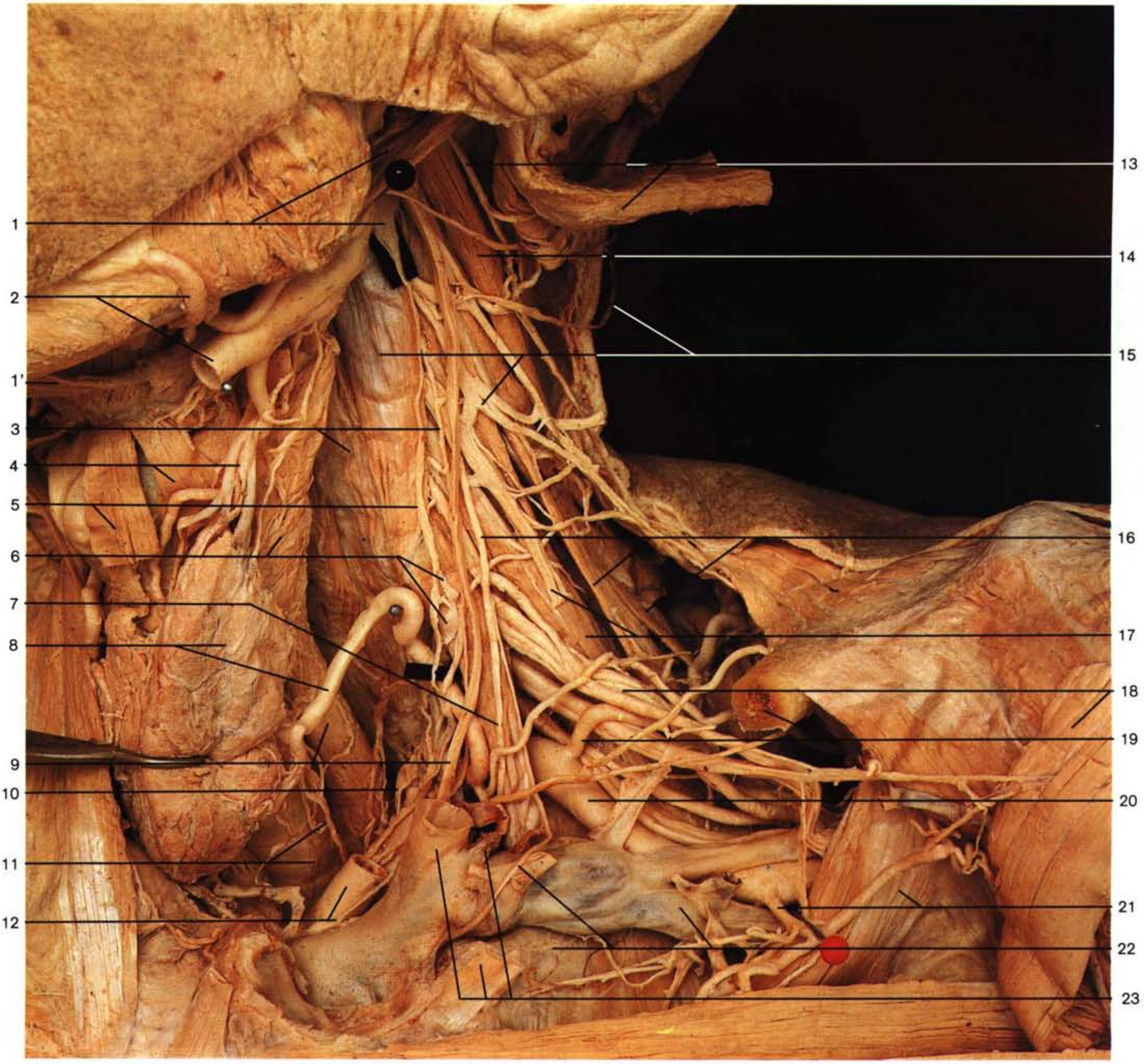

Neck, deepest dissection (anterolateral aspect). Thyroid gland reflected to expose the esophagus and the recurrent laryngeal nerve.

1	**Superior cervical ganglion** of sympathetic trunk and posterior belly of digastric muscle
1'	Anterior belly of digastric muscle
2	**Facial artery** and common carotid artery (reflected anteriorly)
3	Ascending cervical artery and longus colli muscle
4	Omohyoid muscle and superior thyroid artery
5	**Sympathetic trunk** and sternohyoid muscle
6	Middle cervical ganglion and inferior pharyngeal constrictor muscle
7	Scalenus anterior muscle and phrenic nerve
8	**Thyroid gland** and **inferior thyroid artery**
9	Vagus nerve and esophagus
10	**Stellate ganglion**
11	**Recurrent laryngeal nerve** and trachea
12	Common carotid artery and cervical cardiac branch of vagus nerve
13	Sternocleidomastoid muscle and accessory nerve
14	Splenius capitis muscle
15	Lesser occipital nerve, longus capitis muscle and cervical plexus
16	**Phrenic nerve,** scalenus posterior muscle and levator scapulae muscle
17	Supraclavicular nerves and scalenus medius muscle
18	Brachial plexus and pectoralis major muscle (clavicular head)
19	Transverse cervical artery and clavicle
20	**Subclavian artery**
21	**Thoracoacromial artery** and pectoralis minor muscle
22	First rib, accessory phrenic nerve and subclavian vein
23	Internal jugular vein, thoracic duct and subclavius muscle

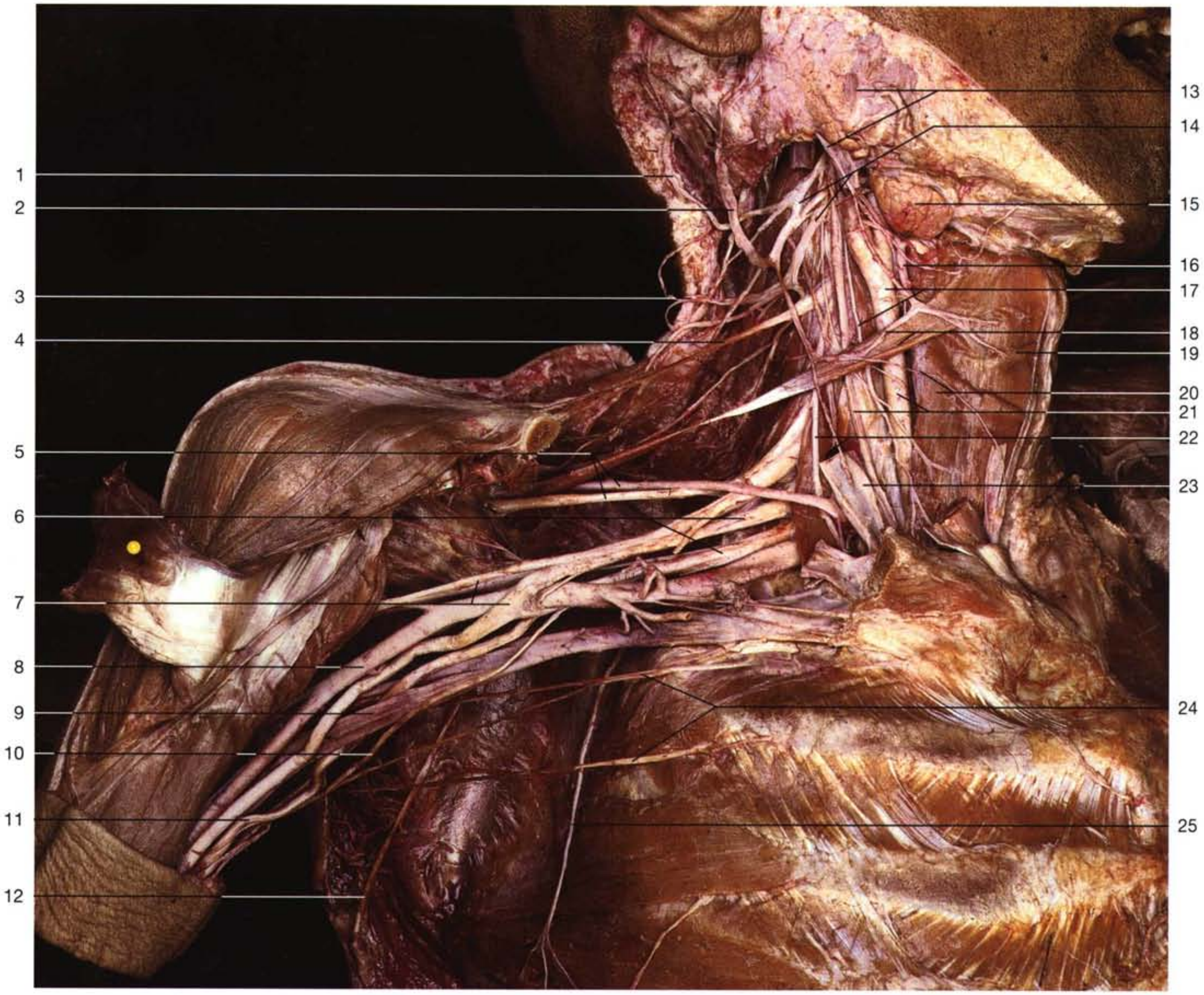

Neck and arm, deepest dissection (anterior-lateral aspect). Cervical and brachial plexus and their relation to the blood vessels are shown. Note the location and content of scalene triangle. Sternocleidomastoid muscle and clavicle have been removed; the internal jugular vein was divided to display the roots of cervical and brachial plexus.

1 Lesser occipital nerve
2 Great auricular nerve
3 Cutaneous branches of cervical plexus
4 Supraclavicular nerve
5 Suprascapular nerve and artery
6 **Brachial plexus**
7 **Median nerve** (with two roots) and musculocutaneous nerve
8 Axillary artery
9 Axillary vein
10 Medial brachial cutaneous nerve
11 **Ulnar nerve**
12 Thoracodorsal nerve
13 Parotid gland and facial nerve (cervical branch)
14 **Cervical plexus**
15 Submandibular gland
16 Superior thyroid artery
17 Common carotid artery dividing in internal and external carotid artery and superior root of ansa cervicalis
18 Omohyoid muscle and cervical branch of facial nerve joining the transverse cervical nerve (C_2, C_3)
19 Sternohyoid muscle
20 Transverse cervical nerve and sternothyroid muscle
21 **Common carotid artery** and **vagus nerve**
22 **Phrenic nerve** and scalenus anterior muscle
23 Internal jugular vein
24 Intercostobrachial nerves
25 Long thoracic nerve

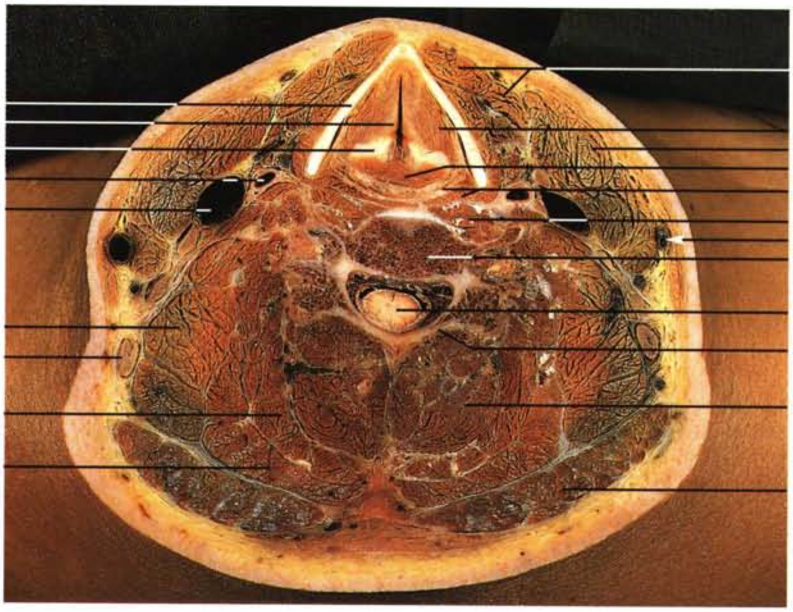

◁ **Horizontal section through the neck** at the level of the fissure of glottis, viewed from above.

1 **Thyroid cartilage**
2 **Vocal fold** and **glottis** (rima glottidis)
3 Arytenoid cartilage
4 Common carotid artery
5 **Internal jugular vein**
6 Infrahyoid muscles
7 Lateral thyroarytenoid muscle
8 Sternocleidomastoid muscle
9 Transverse arytenoid muscle
10 Laryngopharynx and inferior constricor muscle of pharynx
11 Longus colli muscle
12 External jugular vein
13 Body of cervical vertebra (C_5)
14 **Spinal cord**
15 Vertebra arch
16 Deep muscles of neck (semispinalis cervicis muscle)
17 Trapezius muscle
18 Rima glottidis
19 Levator scapulae muscle
20 Lymph node
21 Semispinalis capitis muscle
22 Splenius capitis muscle

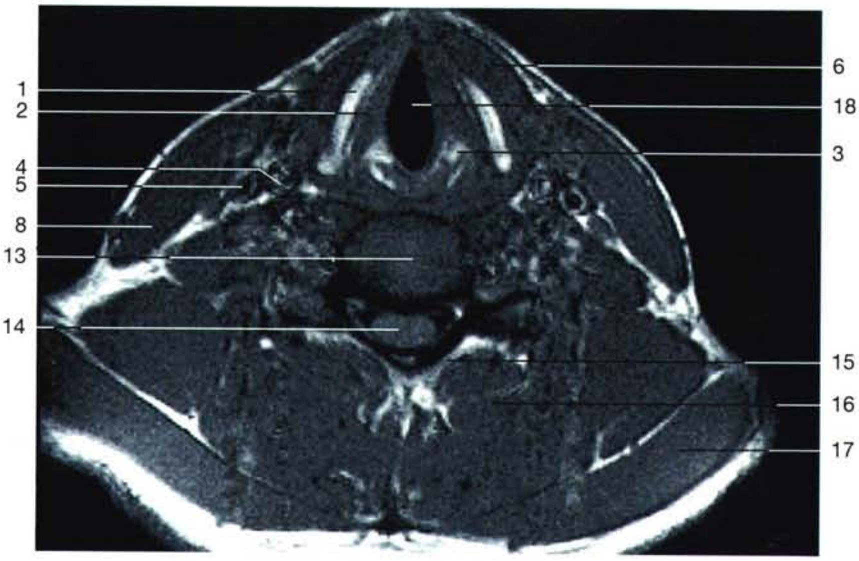

◁ **Section through the neck at the level of larynx.** (MR-Scan.)

1 Hyoid bone
2 Thyroid cartilage
3 **Cervical plexus** (C_1-C_4)
4 **Phrenic nerve** (C_4)
5 Sealenus anterior muscle
6 **Brachial plexus** (C_5-T_1)
7 Scalenus medius and posterior muscles
8 Subclavian artery
9 Subclavian vein
10 Superior vena cava
11 Cricoid cartilage
12 Thyroid gland
13 **Internal jugular vein**
14 **Common carotid artery**
15 Inferior thyroid vein
16 Ascending aorta
17 Descending aorta
18 Second rib

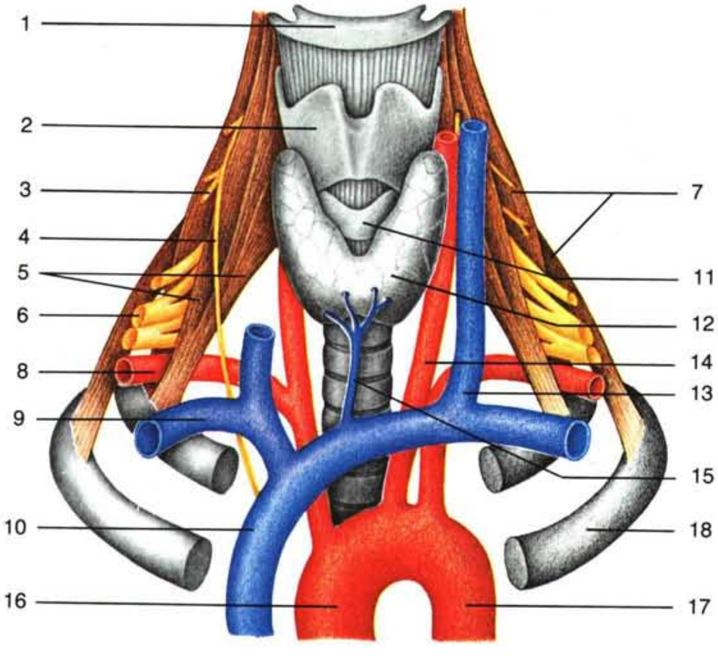

◁ **Scalene triangle, arrangements of blood vessels, and brachial plexus at the lower part of the neck.** (Schematic diagram.)

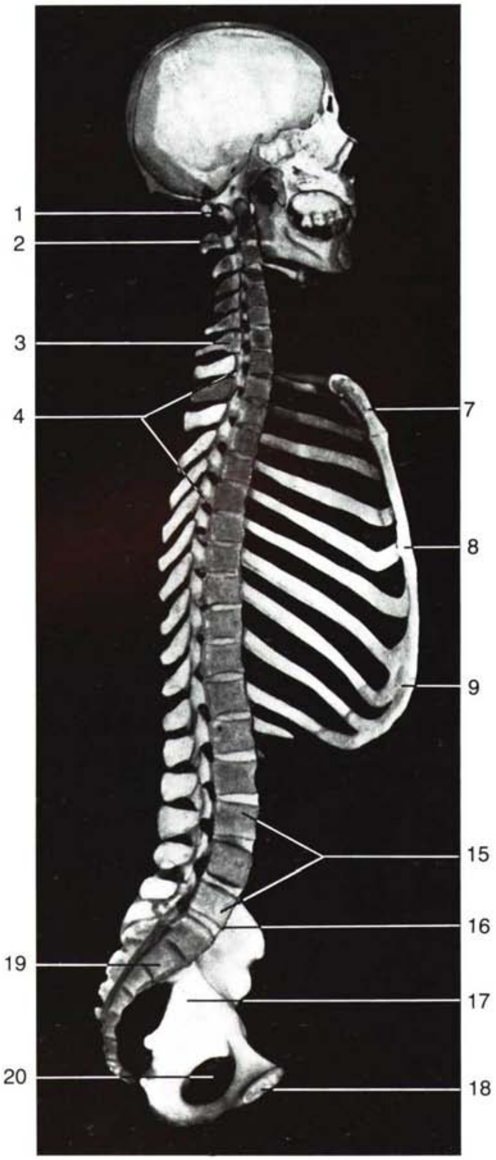

Median sagittal section through the vertebral column, head, and thorax of the adult.

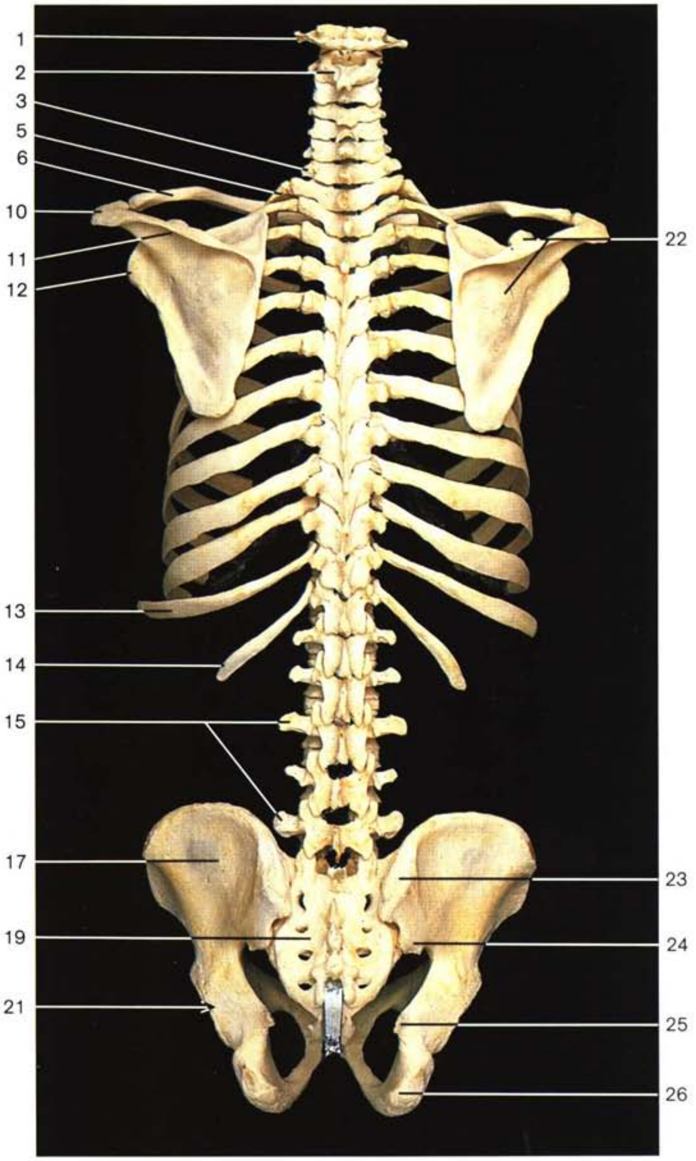

Skeleton of the trunk, vertebral column, thorax, and pelvis (posterior aspect).

1 Atlas
2 Axis
3 Seventh cervical vertebra (vertebra prominens)
4 Vertebral canal
5 First rib
6 Clavicle
7 Manubrium sterni
8 Body of sternum
9 Costal arch
10 Acromion
11 Spine of scapula
12 Glenoid cavity (lateral angle of scapula)
13 Eleventh rib
14 Twelfth rib
15 Lumbar vertebrae
16 Sacral promontory
17 Hip bone
18 Symphysis pubis
19 Sacrum
20 Obturator foramen
21 Acetabulum
22 Scapula with coracoid process
23 Posterior superior iliac spine
24 Posterior inferior iliac spine
25 Ischial spine
26 Ischial tuberosity

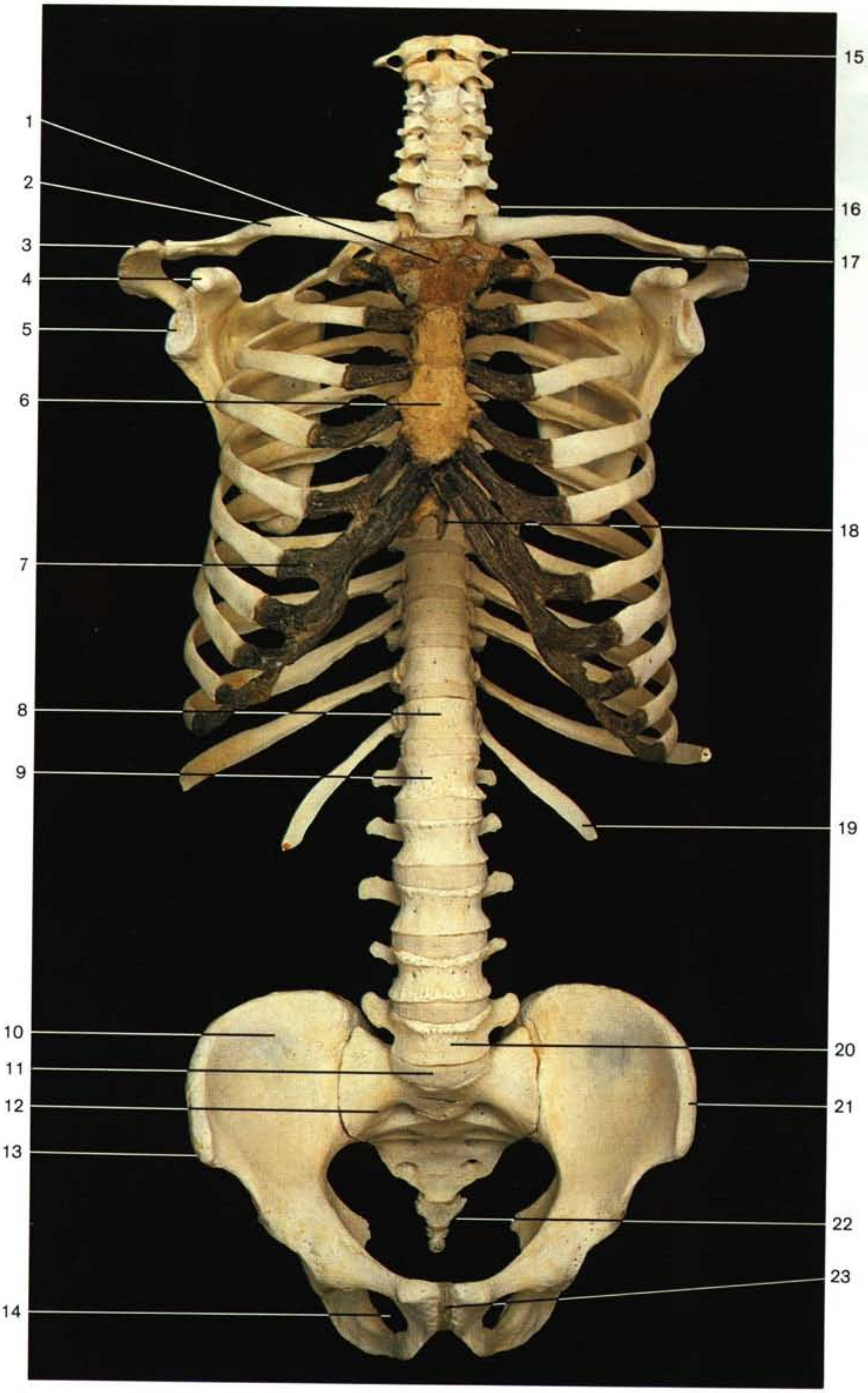

1 Manubrium sterni
2 **Clavicle**
3 Acromion
4 Coracoid process
5 Glenoid cavity
6 Body of **sternum**
7 Costal cartilage
8 Body of the twelfth **thoracic vertebra**
9 Body of the first **lumbar vertebra**
10 Hip bone
11 Sacral promontory
12 **Sacrum**
13 Anterior superior iliac spine
14 Obturator foramen
15 Atlas
16 Seventh **cervical vertebra**
17 First rib
18 Xiphoid process
19 Twelfth rib
20 Body of the fifth lumbar vertebra
21 Iliac crest
22 **Coccyx**
23 Symphysis pubis

Skeleton of the trunk, vertebral column, pelvis, thorax, and shoulder girdle (anterior aspect).

The trunk is divided into segments best visible in the thoracic region, where each segment consists of a pair of ribs connected anteriorly by the sternum and posteriorly by a thoracic vertebra. In the lumbar part of the vertebral column only vestiges of ribs are present which form what appear to be the transverse processes. In cervical vertebrae, remnants of ribs are part of the transverse processes. Each segment also comprises muscles (e.g. intercostal muscles), nerves, and vessels. However, in the cervical and lumbar region the muscular segments fuse with each other forming large muscle plates, for example, the oblique muscles of the abdomen, while vessels and nerves still retain their segmental pattern.

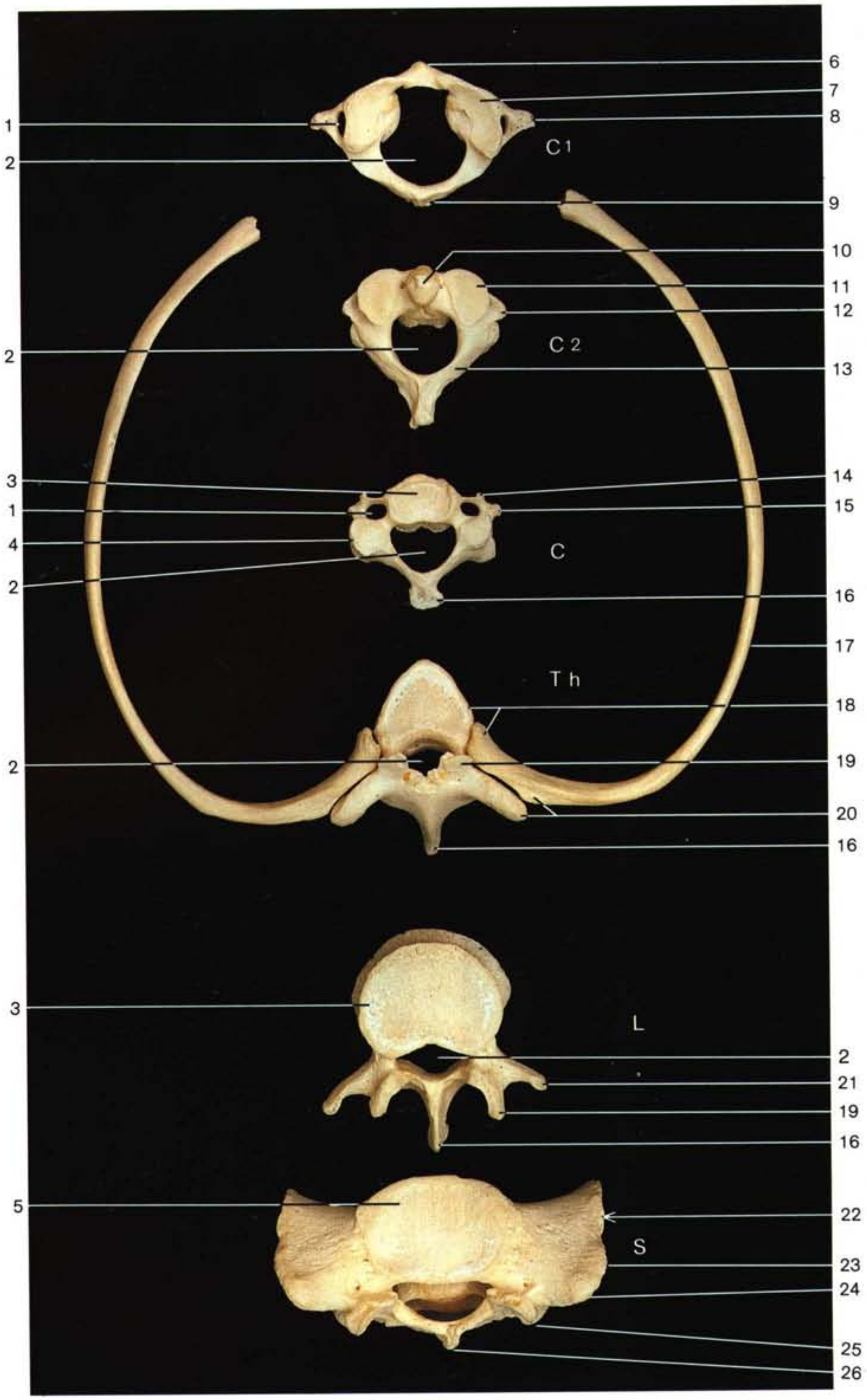
1
2
2
3
1
4
2
2
3
5
6
7
8
C1
9
10
11
12
C 2
13
14
15
C
16
17
Th
18
19
20
16
L
2
21
19
16
22
S
23
24
25
26

Representative vertebrae from each region of the vertebral column (superior aspect). From top to bottom: atlas (C_1), axis (C_2), cervical vertebra (C), thoracic vertebra (Th), lumbar vertebra (L), and sacrum (S).

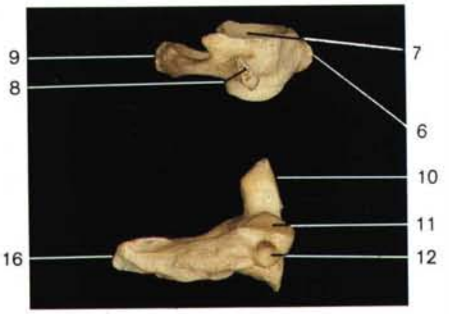
9
8
7
6
10
11
16
12

Atlas and axis

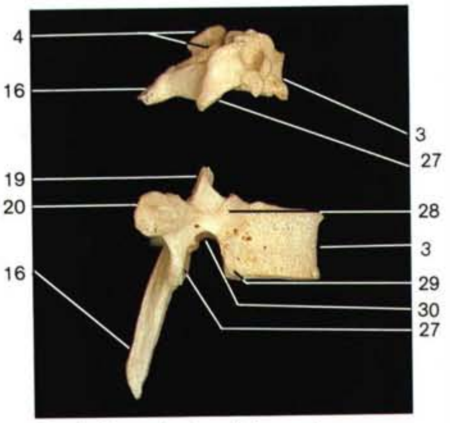
4
16
3
27
19
20
16
28
3
29
30
27

Typical cervical and thoracic vertebrae

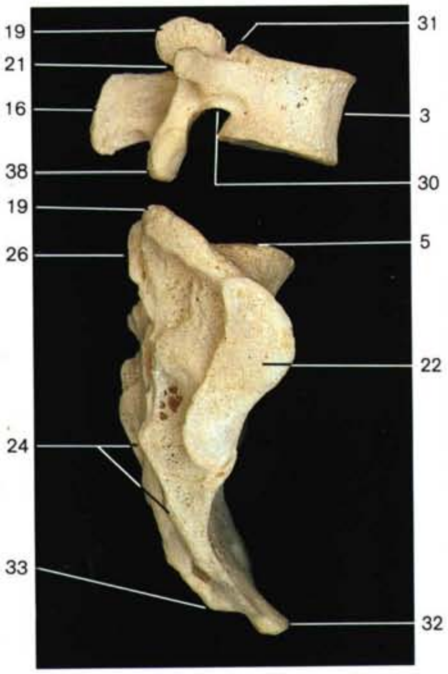
19
21
16
38
31
3
30
19
26
5
22
24
33
32

Typical lumbar vertebra and sacrum

Representative vertebrae from each region of the vertebral column (lateral aspect, ventral surface on the right).

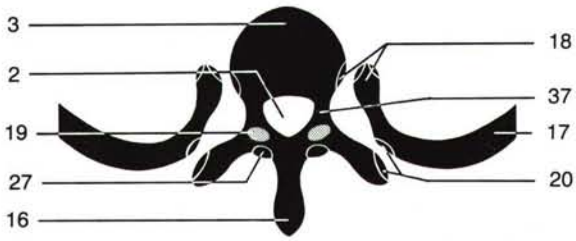
3
2
19
27
16
18
37
17
20

General organization of ribs and vertebrae. (Schematic diagram.)

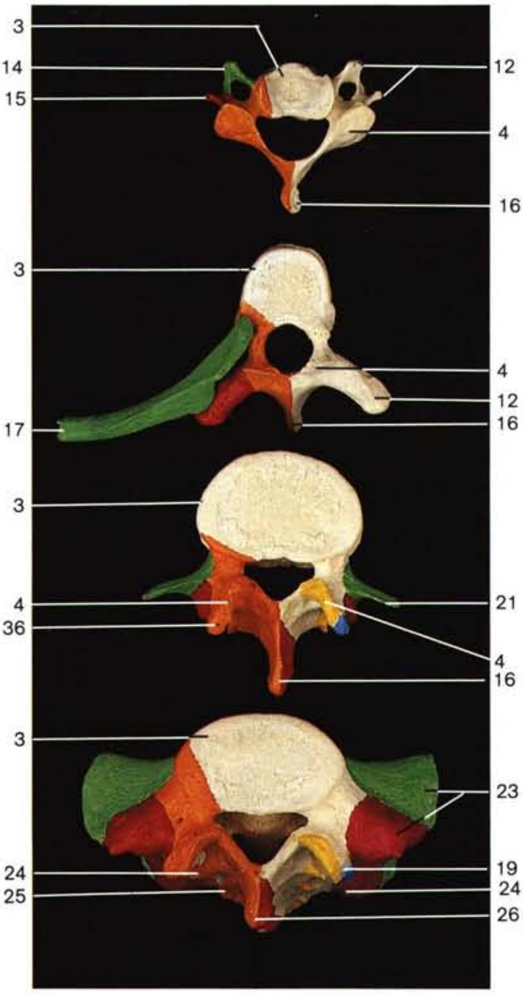

General characteristics of the vertebrae. Typical cervical, thoracic, lumbar vertebrae and sacrum.

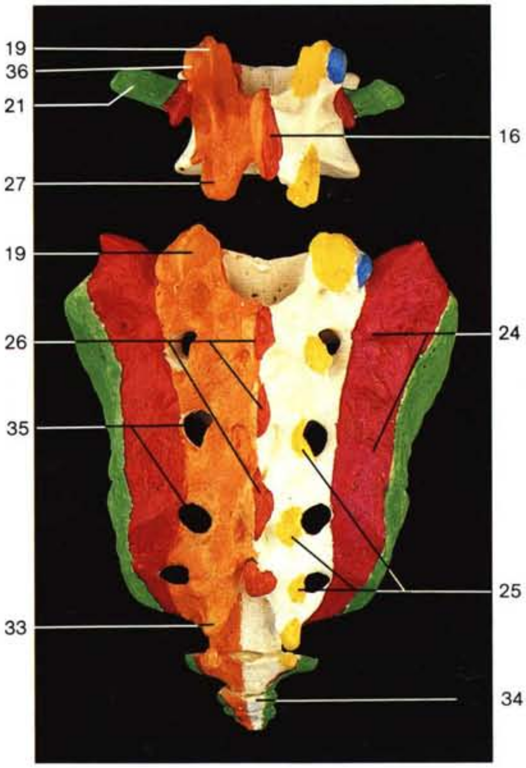

General characteristics of lumbar vertebrae and sacrum (posterior aspect).

1 Foramen transversarium
2 Vertebral foramen
3 Body of vertebra
4 Superior articular facet
5 Base of sacrum
6 Anterior tubercle of atlas
7 Superior articular facet of atlas
8 Transverse process
9 Posterior tubercle of atlas
10 Dens of axis
11 Superior articular surface
12 Transverse process
13 Arch of vertebra
14 Anterior tubercle of transverse process
15 Posterior tubercle of transverse process
16 Spinous process
17 Shaft of rib
18 Body of vertebra and head of rib articulating with each other (costovertebral joint)
19 Superior articular process
20 Transverse process and tubercle of rib articulating with each other (costotransverse joint)
21 Transverse process
22 Auricular surface
23 Lateral part of sacrum
24 Lateral sacral crest
25 Intermediate sacral crest
26 Median sacral crest
27 Inferior articular facet
28 Superior demifacet for head of rib
29 Inferior demifacet for head of rib
30 Inferior vertebral notch
31 Superior vertebral notch
32 Apex of the sacrum
33 Sacral cornu
34 Coccyx
35 Dorsal sacral foramina
36 Mamillary process
37 Pedicle
38 Inferior articular process

Green = Ribs or homologous processes
Red = Muscular processes (transverse and spinous processes)
Orange = Laminae and articular processes
Yellow = Articular facets

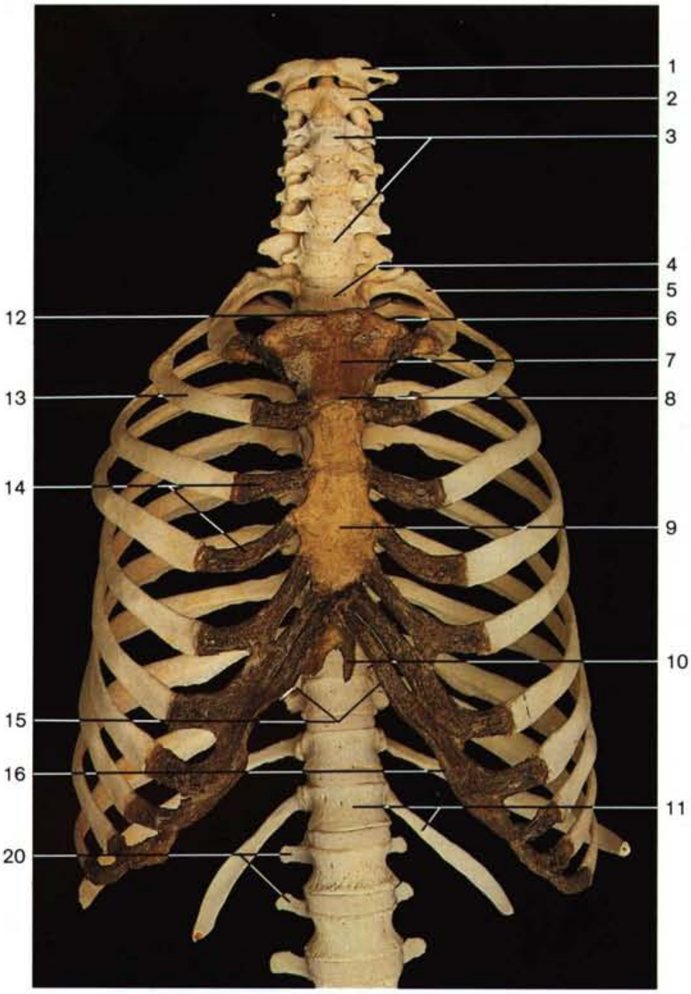

Skeleton of the thorax (anterior aspect).

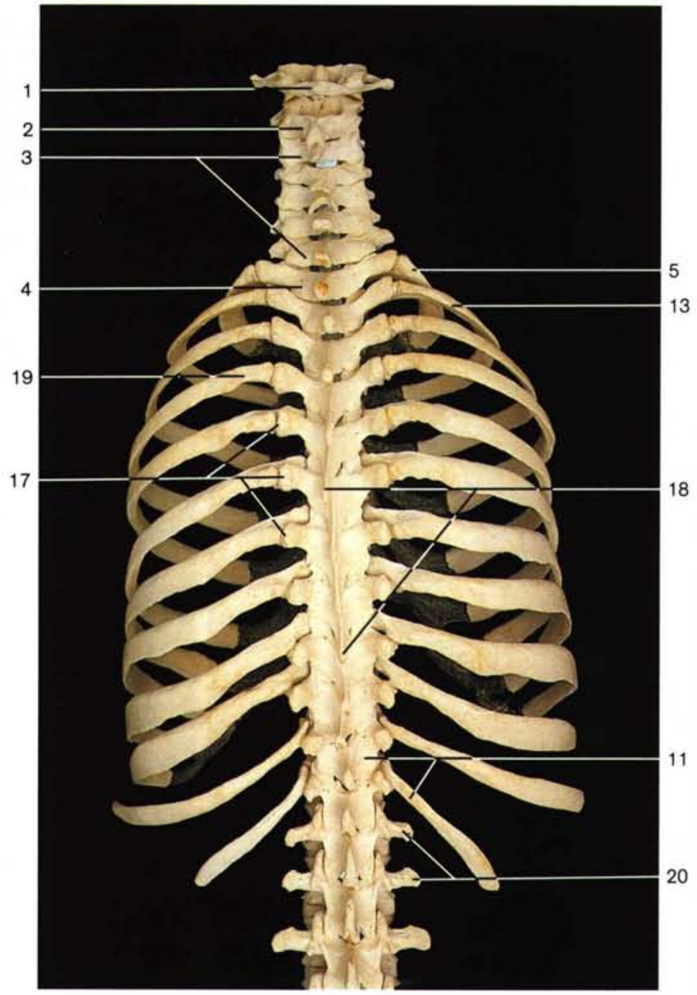

Skeleton of the thorax (posterior aspect).

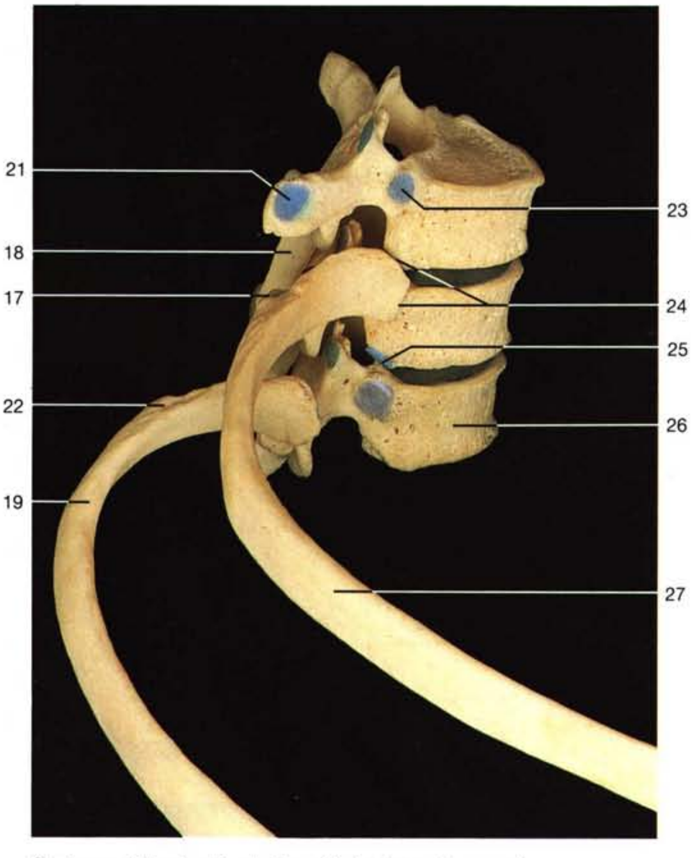

Costovertebral articulation (right lateral aspect).

1 Atlas
2 Axis
3 **Cervical vertebrae**
4 First thoracic vertebra
5 First rib
6 Facet for clavicle and clavicular notch
7 **Manubrium sterni**
8 Sternal angle
9 Body of sternum
10 Xiphoid process
11 Twelfth thoracic vertebra and rib
12 Jugular notch
13 Second rib
14 **Costal cartilages**
15 Infrasternal angle
16 **Costal arch**
17 Costotransverse joints between the transverse processes of thoracic vertebra and the tubercles of the ribs
18 Spinous processes
19 Costal angle
20 Transverse processes of lumbar vertebrae
21 Facet for articulation with rib
22 Tubercle of rib
23 Superior facet for articulation with head of rib
24 Articulation of head of rib with two vertebrae
25 Inferior facet for articulation with head of rib
26 Body of thoracic vertebra
27 Body or shaft of rib

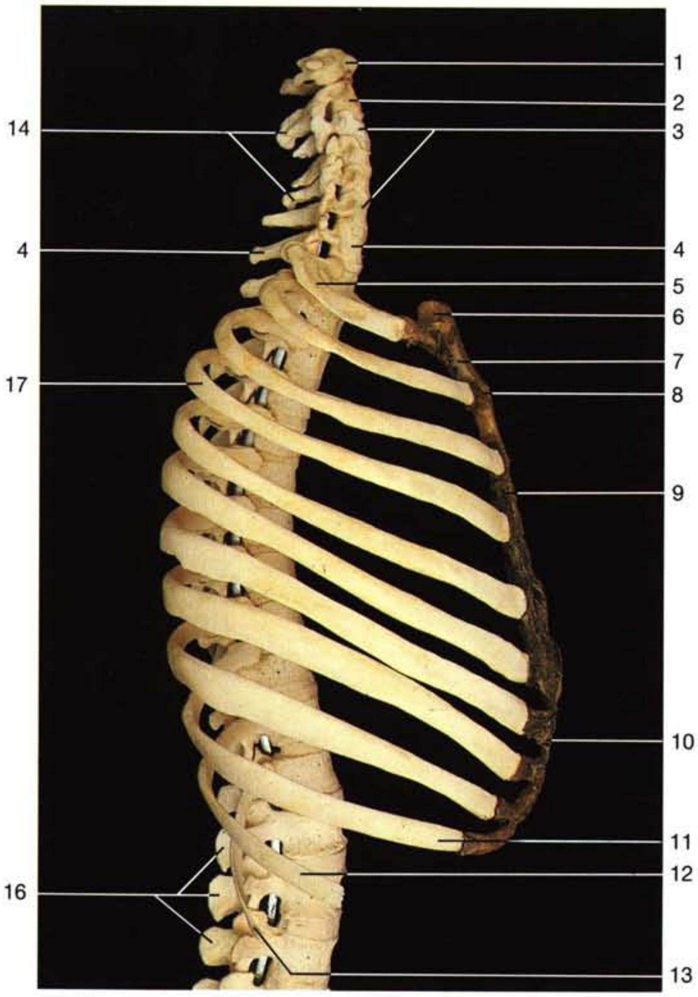
1
2
14
3
4
4
5
6
7
17
8
9
10
11
12
16
13

Skeleton of the thorax (right lateral aspect).

1 Atlas
2 Axis
3 **Cervical vertebrae**
4 Seventh cervical vertebra (vertebra prominens)
5 First rib
6 Facet for clavicle
7 **Manubrium sterni**
8 Sternal angle
9 Body of sternum
10 Costal arch
11 Tenth rib
12 Eleventh rib
13 Twelfth rib
14 Spinous processes of cervical vertebrae
15 Spinous processes of thoracic vertebrae
16 Spinous processes of lumbar vertebrae
17 Costal angle
18 Intervertebral foramina
19 Intervertebral discs
20 **Cervical curvature**
21 **Thoracic curvature**
22 **Lumbar curvature**
23 Sacrum
24 Coccyx

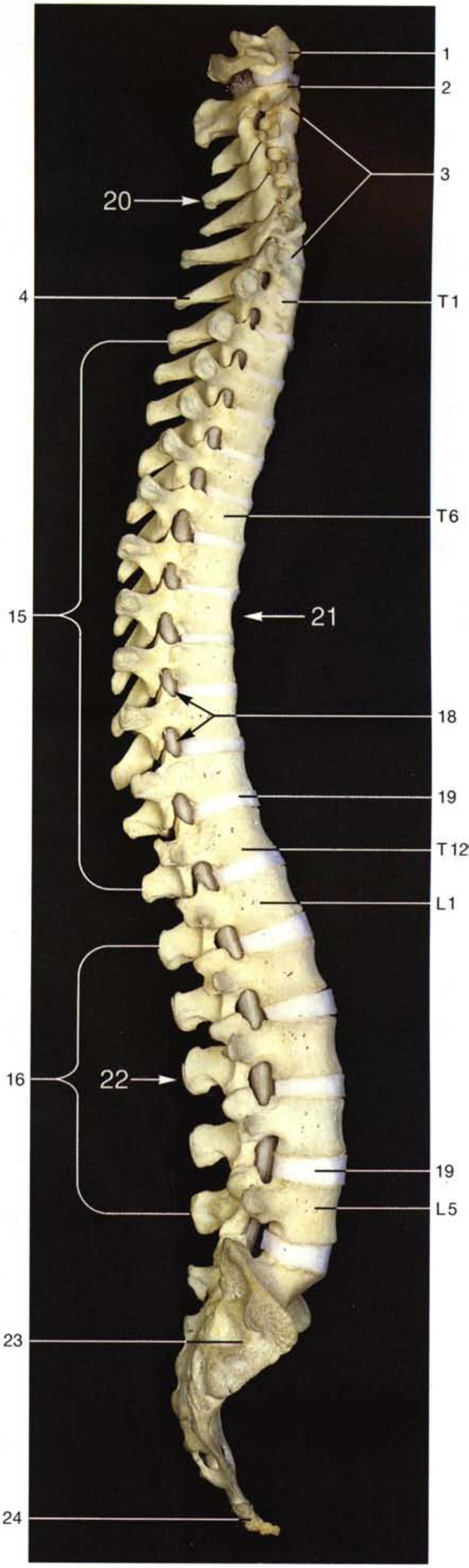
1
2
3
20
4
T1
T6
15
21
18
19
T12
L1
16
22
19
L5
23
24

Vertebral column (right lateral aspect).

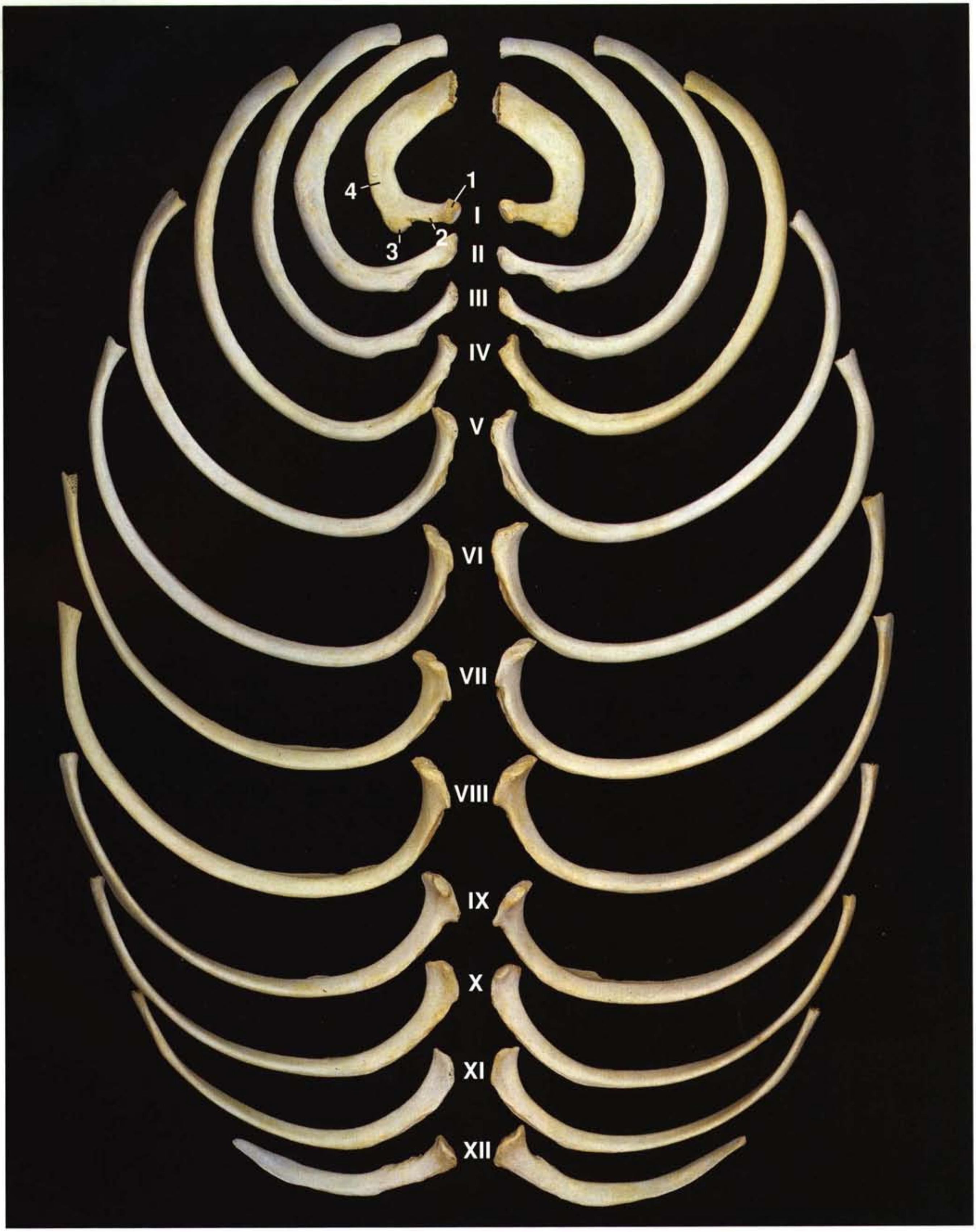

Disarticulated ribs. The twelve ribs arranged in a craniocaudal direction, first rib at top. I–XII = pairs of ribs; 1 = head of rib; 2 = neck of rib; 3 = tubercle of rib; 4 = body of rib.

The head of a typical rib articulates with the bodies of adjacent vertebrae and the intervening intervertebral disc. The first seven ribs are directly connected through the costal cartilages (not depicted) to the sternum (true ribs) while the remaining five ribs are only indirectly connected to the sternum (false ribs, ribs VIII to X) or end freely in the lateral wall of the trunk (floating ribs, ribs XI and XII).

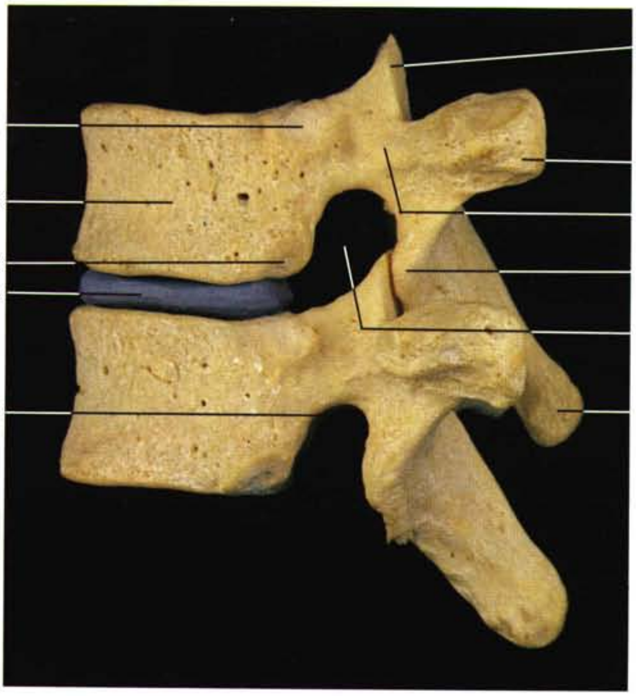

Two **thoracic vertebrae** (left lateral aspect).

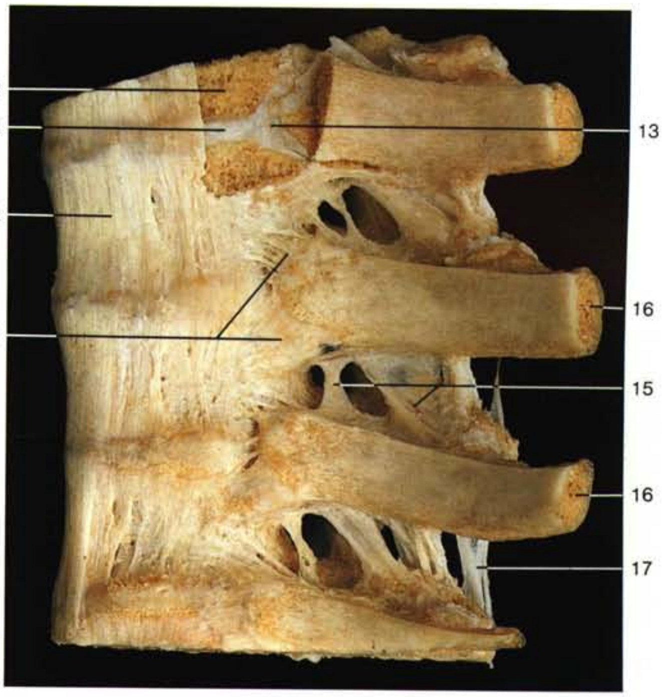

Ligaments of thoracic vertebrae and costovertebral joints (left anterolateral aspect).
In the upper joint, most of the radiate ligament and the anterior part of the head of the rib have been removed to expose the two joint cavities and the interposed intra-articular ligament.

1 Superior demifacet for head of rib
2 Body of vertebra
3 Inferior demifacet for head of rib
4 Intervertebral disc
5 Inferior vertebral notch
6 Superior articular facet and superior articular process
7 Pedicle
8 Transverse process and facet for tubercle of rib
9 Inferior articular process
10 Intervertebral foramen
11 Spinous process
12 Anterior longitudinal ligament
13 Intraarticular ligament
14 Radiate ligament
15 Superior costotransverse ligament
16 Body of rib
17 Intertransverse ligament

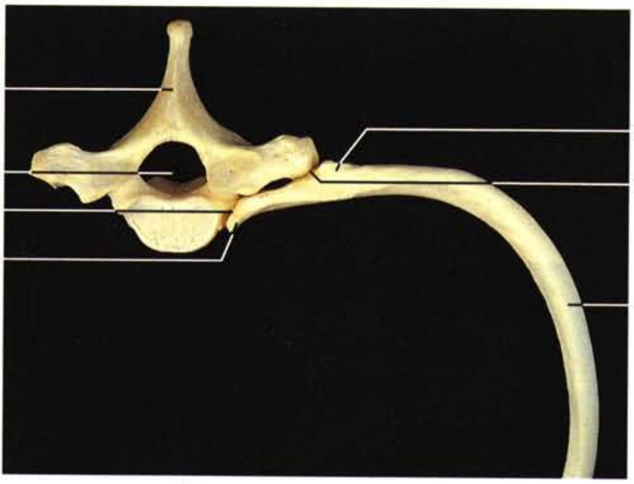

Location of costovertebral joints (superior aspect).

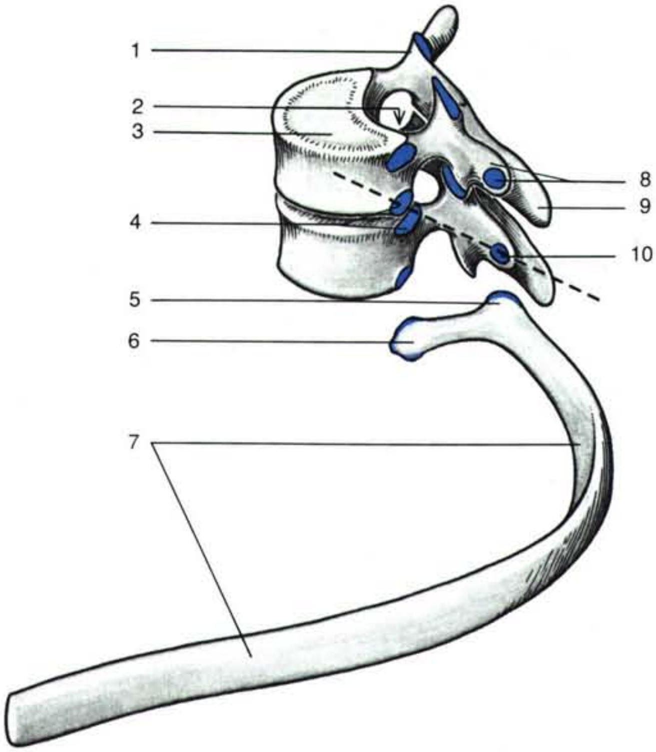

Costovertebral joints. Two thoracic vertebrae with an articulating rib (separated). Axis of movement indicated by dotted line. Blue = articular facets. (Schematic diagram.)

1 Superior articular process
2 Vertebral canal
3 Body of thoracic vertebra
4 Costovertebral joint (articular facets)
5 Tubercle of rib
6 Head of rib
7 Shaft or body of rib
8 Transverse process with articular facet
9 Spinous process
10 Costotransverse joint (articular facets)

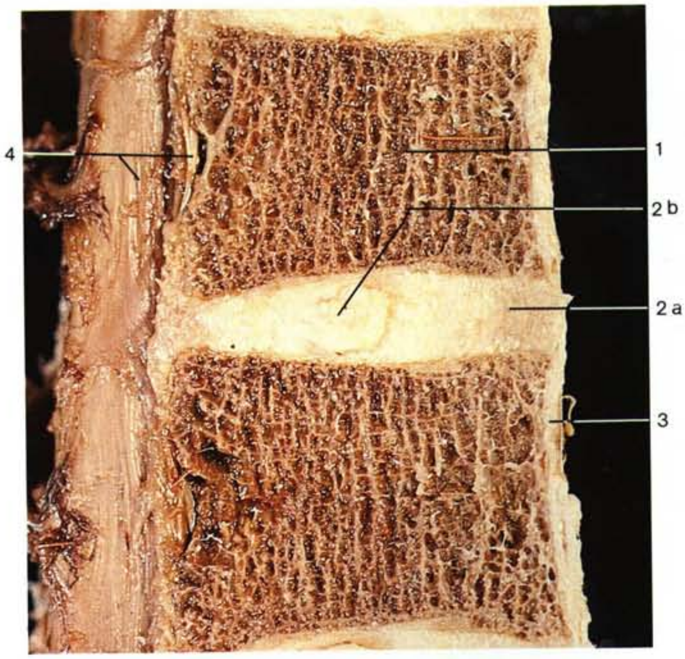

Median-sagittal section of the bodies of the vertebrae showing the **intervertebral discs,** each of which consists of an outer laminated portion and an inner core.

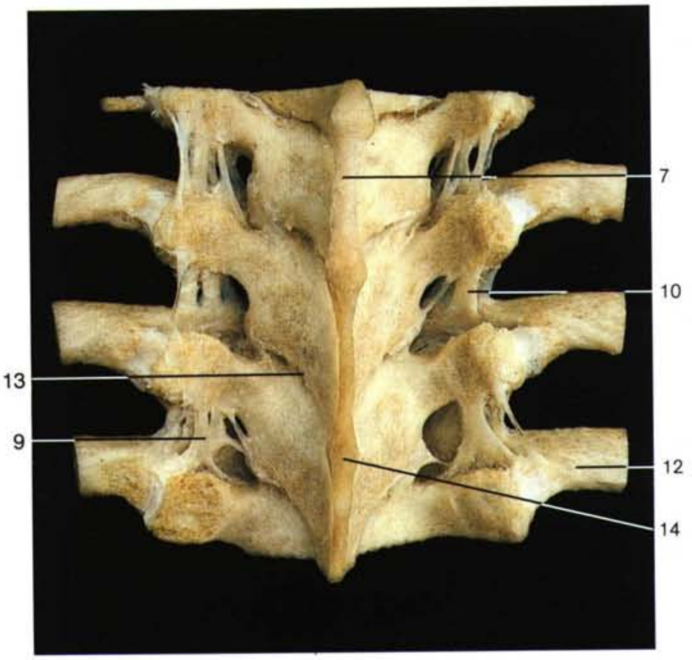

Ligaments of the vertebral column (dorsal aspect).

1 Body of vertebra
2 **Intervertebral disc**
  a Outer portion (anulus fibrosus)
  b Inner core **(nucleus pulposus)**
3 Anterior longitudinal ligament
4 Posterior longitudinal ligament and spinal dura mater
5 Transverse process of lumbar vertebra
6 Sacrum
7 **Supraspinous ligament**
8 Interspinous ligament
9 **Intertransverse ligament**
10 Superior costotransverse ligament
11 Transverse process of thoracic vertebra
12 Rib
13 **Ligamentum flavum**
14 Spinous process
15 Intervertebral foramen

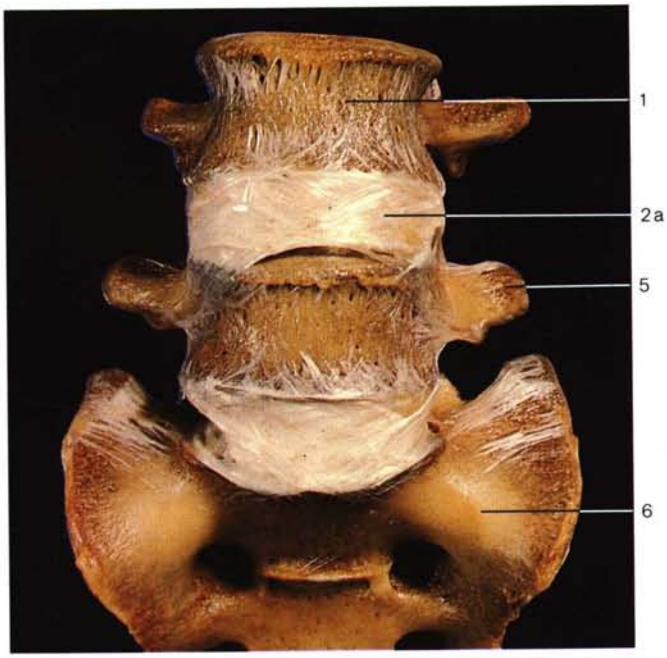

The two caudal lumbar vertebrae and the sacrum with their intervertebral discs (anterior aspect). Anterior longitudinal ligament removed.

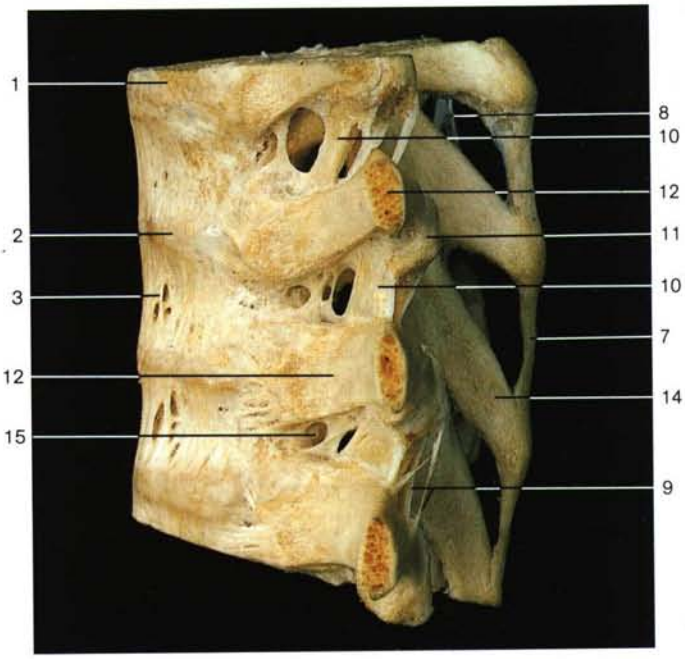

Ligaments of the vertebral column, thoracic part (left lateral aspect).

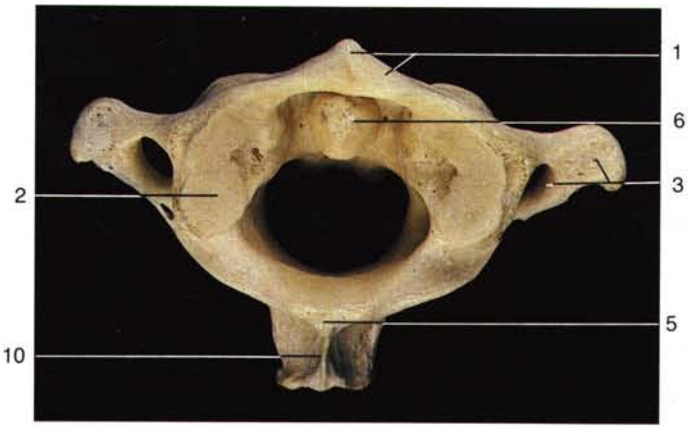

Atlas and axis (from above).

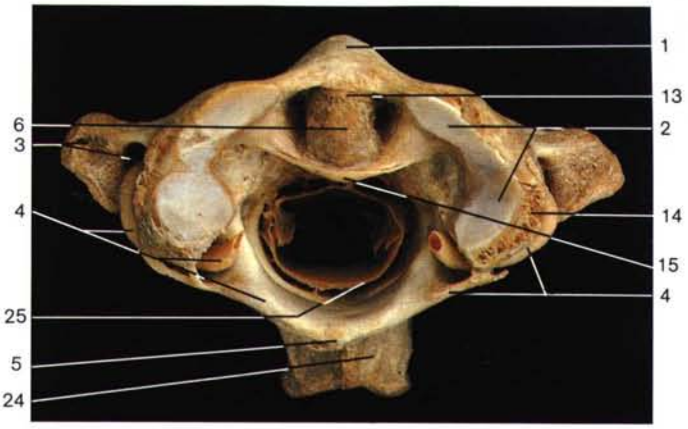

Median atlantoaxial joint and transverse ligament of atlas (from above). Dens of axis partly severed.

1 Anterior arch of **atlas** with anterior tubercle
2 Superior articular facet of atlas
3 **Foramen transversarium** and transverse process
4 Posterior arch of atlas and vertebral artery
5 Posterior tubercle of atlas
6 **Dens of axis**
7 Superior articular surface of axis
8 Body of axis
9 Pedicle and lamina of axis
10 Spinous process
11 Inferior articular process
12 Transverse process and vertebrarterial foramen (foramen transversarium)
13 **Median atlantoaxial joint** (anterior part)
14 Articular capsule of **atlantooccipital joint**
15 Transverse ligament of atlas
16 Occipital bone
17 **Atlantooccipital joint**
18 **Lateral atlantoaxial joint**
19 Third cervical vertebra
20 Superior longitudinal band of cruciform ligament
21 Alar ligaments
22 Transverse ligament of atlas
23 Inferior longitudinal band of cruciform ligament
24 Spinous process of axis
25 Dura mater

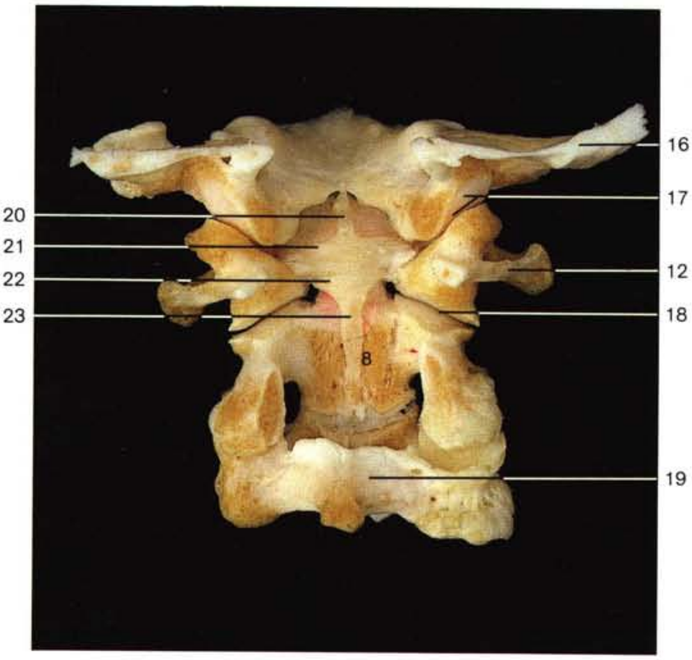

Atlantooccipital and atlantoaxial joints (posterior aspect). Posterior part of occipital bone, posterior arch of atlas and axis have been removed to show the cruciform ligament.

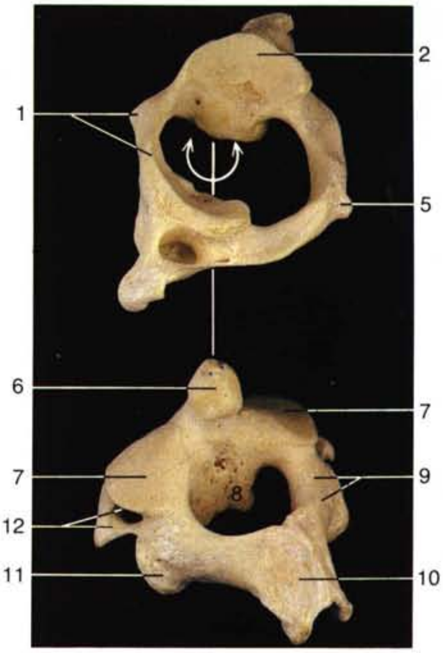

Atlas and axis. Left oblique posterolateral aspect, demonstrating the articulation of the dens of axis with atlas (cf. arrows).

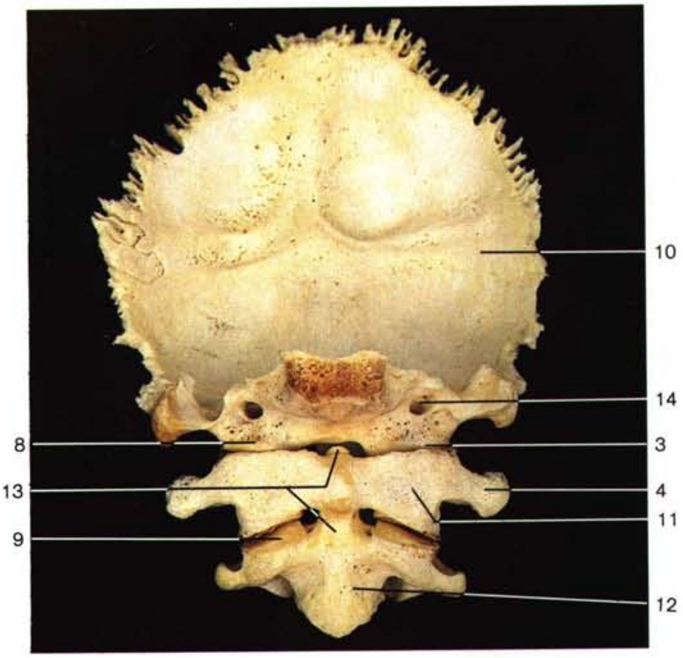

Occipital bone, atlas and axis (anterior aspect).

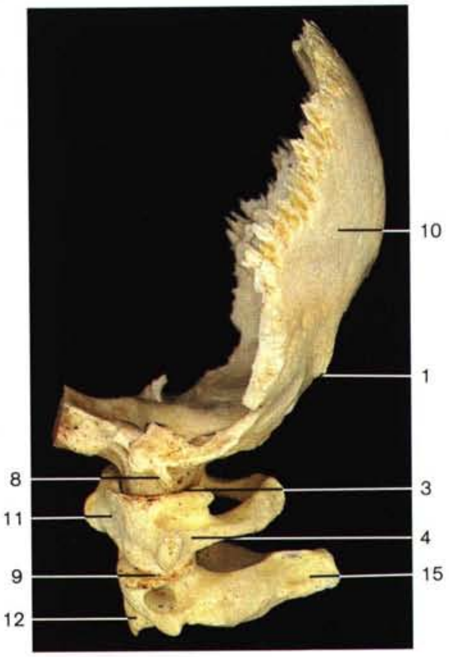

Occipital bone, atlas and axis (left lateral aspect).

1 External occipital protuberance
2 Foramen magnum
3 Atlantooccipital joint
4 Transverse process of atlas
5 Membrana tectoria
6 Posterior longitudinal ligament
7 Spinous process of third cervical vertebra
8 Occipital condyle
9 Lateral atlantoaxial joint
10 Occipital bone
11 Atlas
12 Axis
13 Dens of axis
14 Hypoglossal canal
15 Spinous processes of axis
16 Anterior longitudinal ligament

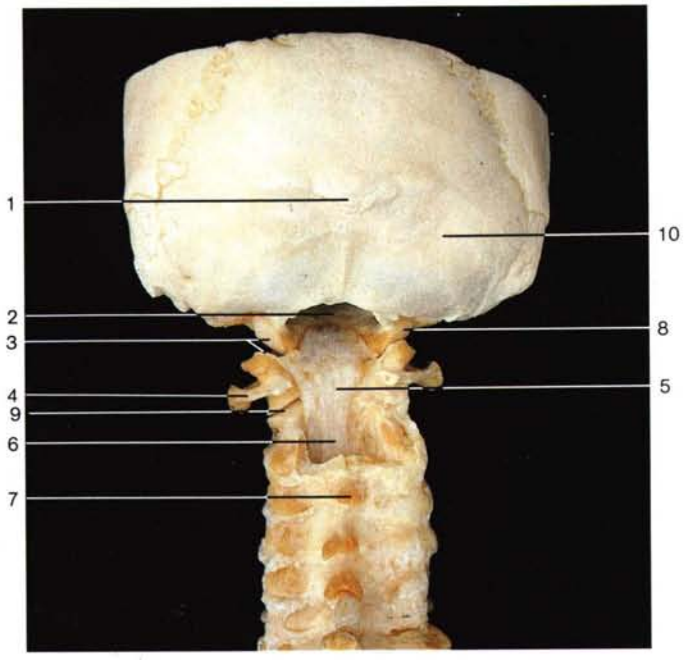

Cervical vertebral column and skull with ligaments (posterior aspect). Posterior arch of atlas and axis removed to show the membrana tectoria.

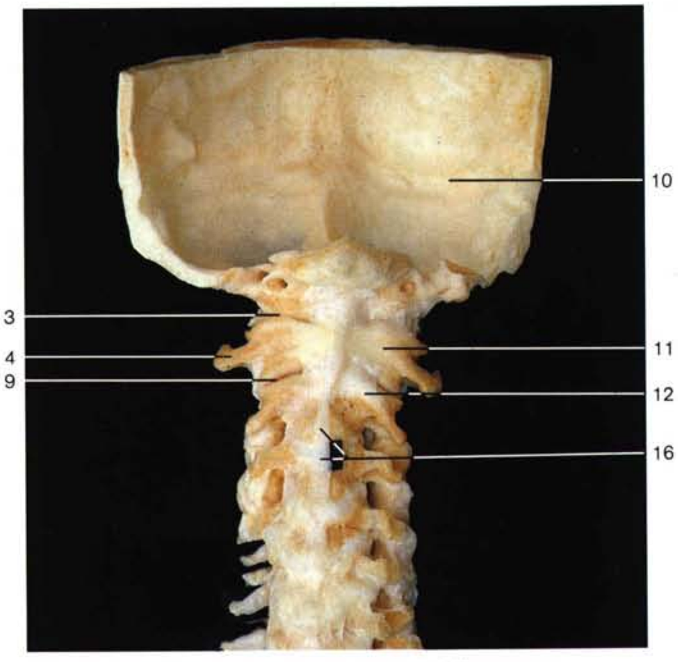

Cervical vertebral column and skull with ligaments (anterior aspect). Anterior part of occipital bone removed.

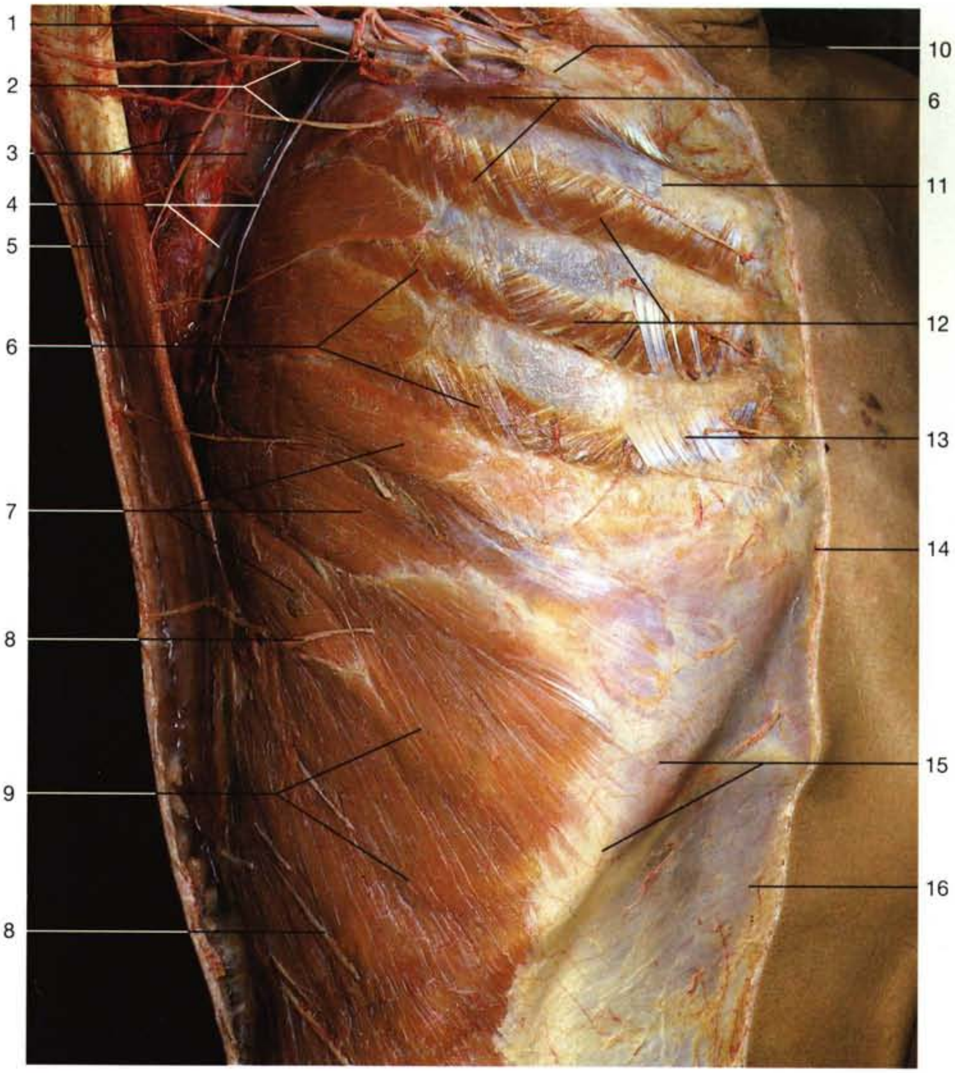

1 Axillary vein
2 Intercostobrachial nerves
3 Subscapularis muscle and thoracodorsal nerve
4 Long thoracic nerve, lateral thoracic artery and vein
5 Latissimus dorsi muscle
6 **External intercostal** muscles
7 **Serratus anterior** muscle
8 Lateral cutaneous branches of intercostal nerves
9 External abdominal oblique muscle
10 Clavicle (divided)
11 Second rib (costo-chondral junction)
12 **Internal intercostal muscles**
13 External intercostal membrane
14 Position of xiphoid process
15 Costal arch or margin
16 Anterior layer of rectus sheath

Muscles of the thorax, superficial layer (lateral aspect). Upper limb elevated. Pectoralis major and minor muscles have been removed.

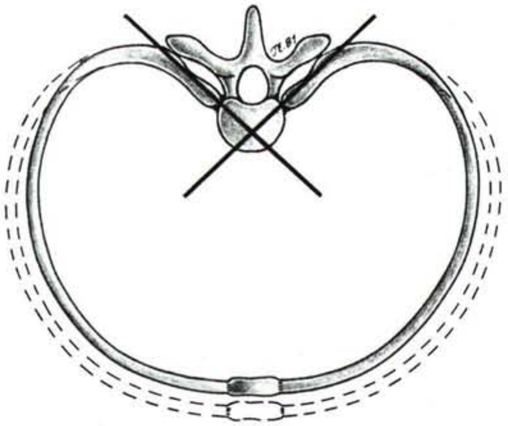

Effect of intercostal muscles on the costovertebral and costotransverse joints. Axes of movement indicated by lines; direction of movements indicated by arrows.

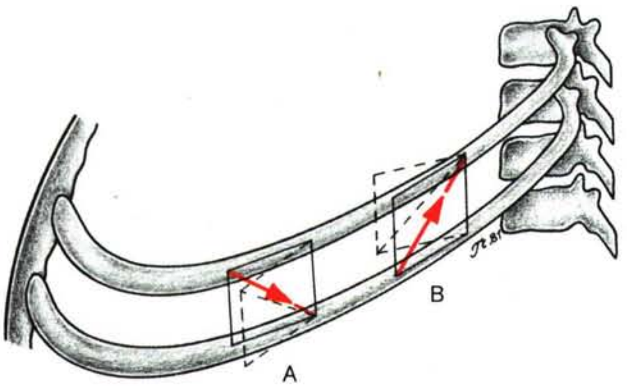

A Action of internal intercostal muscles (expiration)
B Action of external intercostal muscles (inspiration)

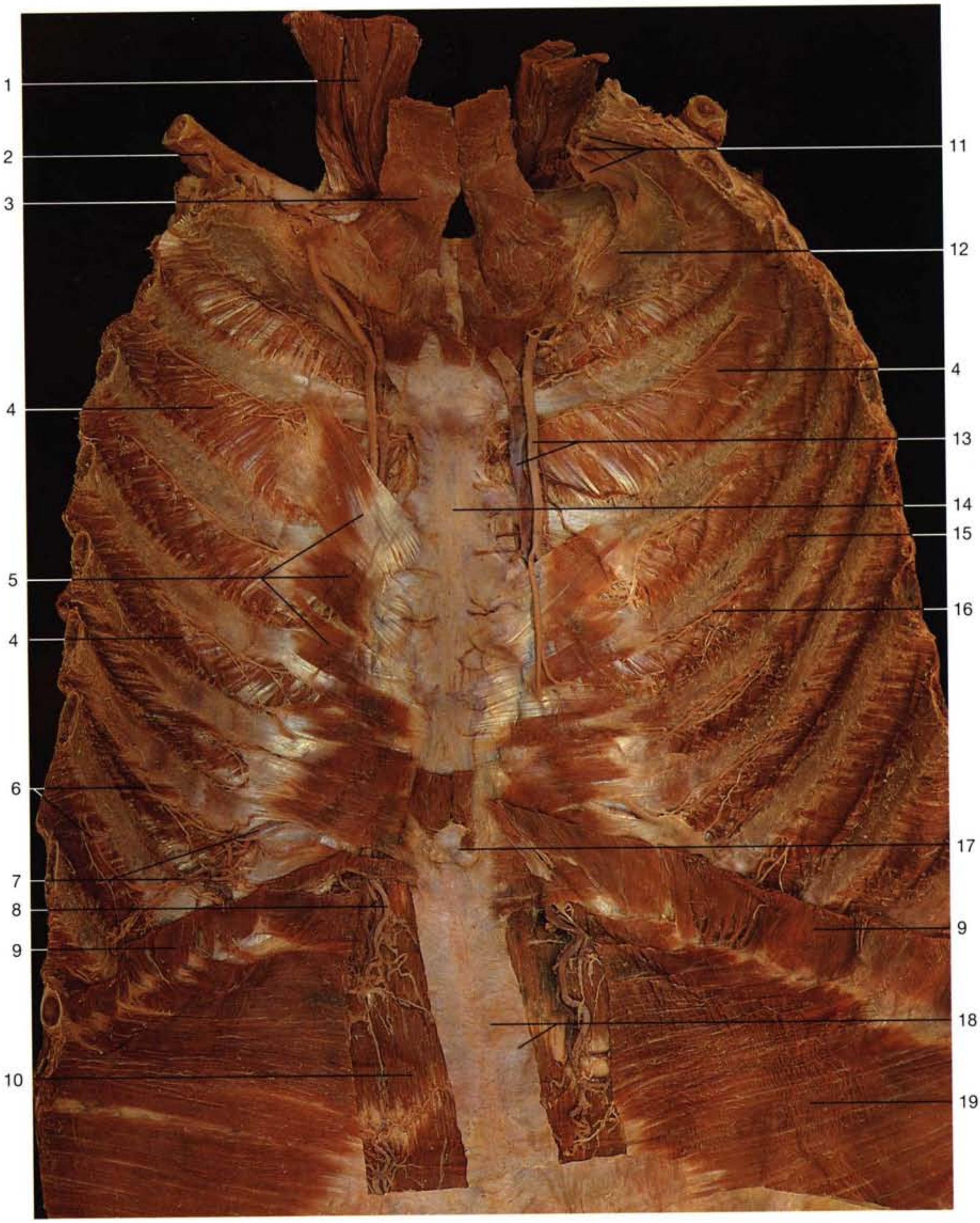

Anterior thoracic wall (posterior aspect). Diaphragm partly removed, posterior layer of rectus sheath fenestrated on both sides.

1 Sternocleidomastoid muscle (divided)
2 Clavicle
3 Sternothyroid muscle
4 **Internal intercostal muscle**
5 **Transversus thoracis muscle**
6 Intercostal arteries and nerves
7 Musculophrenic artery
8 Superior epigastric artery and vein
9 Diaphragm (divided)
10 Rectus abdominis muscle

11 Subclavian artery and brachial plexus
12 First rib
13 Internal thoracic artery and vein
14 Sternum
15 Innermost intercostal muscle
16 Intercostal artery and vein
17 Xiphoid process
18 Linea alba and posterior layer of rectus sheath
19 **Transversus abdominis muscle**

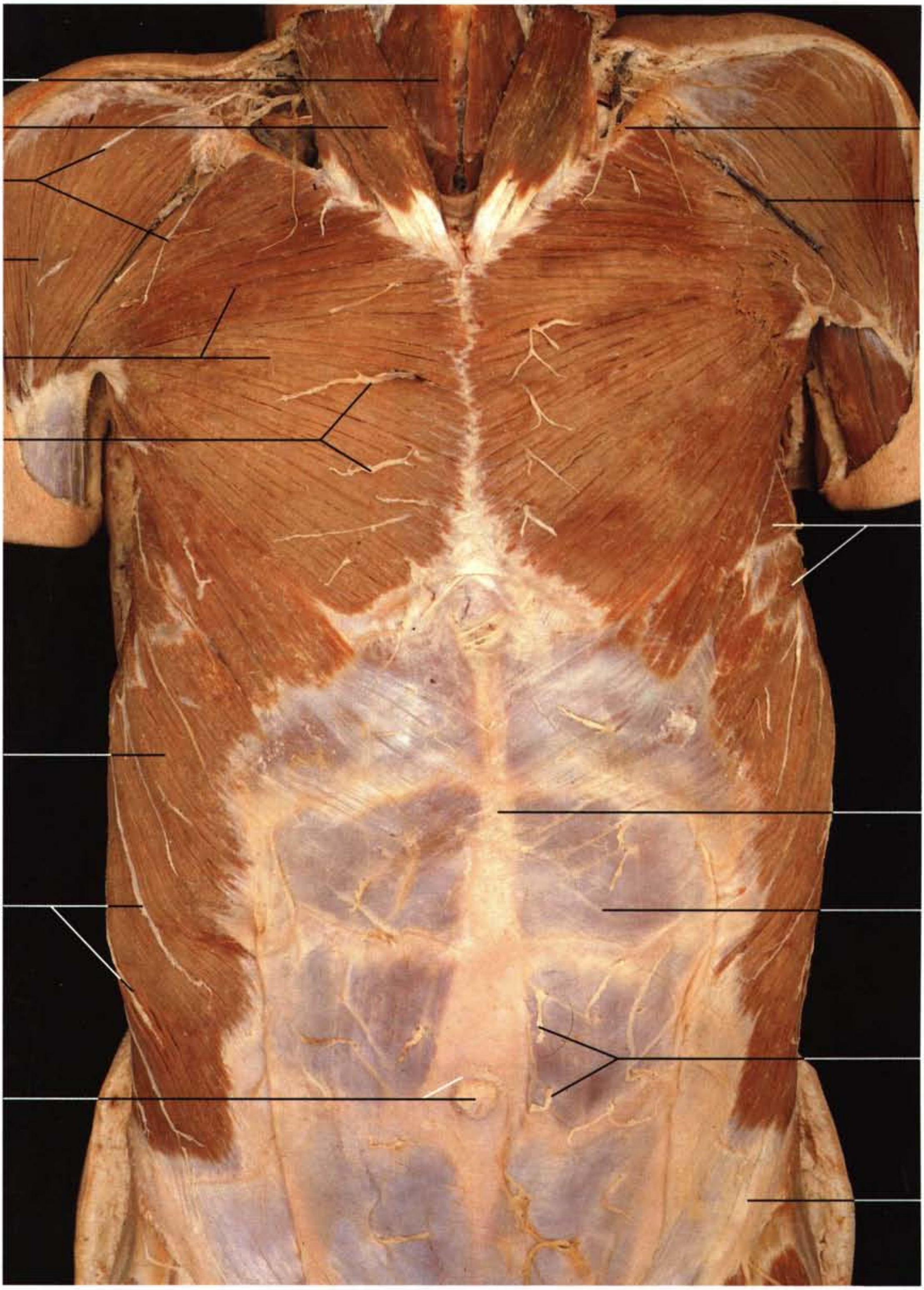

Superficial muscles of the anterior thoracic and abdominal wall. The fascia of pectoralis major muscle and the abdominal wall have been removed; the anterior layer of the sheath of the rectus abdominis muscle is displayed.

1 Sternohyoid muscle
1 Sternocleidomastoid muscle
3 Supraclavicular nerves (branches of cervical plexus)
4 Deltoid muscle
5 **Pectoralis major muscle**
6 **Anterior cutaneous branches of intercostal nerves**
7 External abdominal oblique muscle
8 **Lateral cutaneous branches of intercostal nerves**
9 Umbilicus and umbilical ring
10 Clavicle
11 Cephalic vein
12 Serratus anterior muscle
13 Linea alba
14 **Sheath of rectus abdominis** muscle (anterior layer)
15 Inguinal ligament

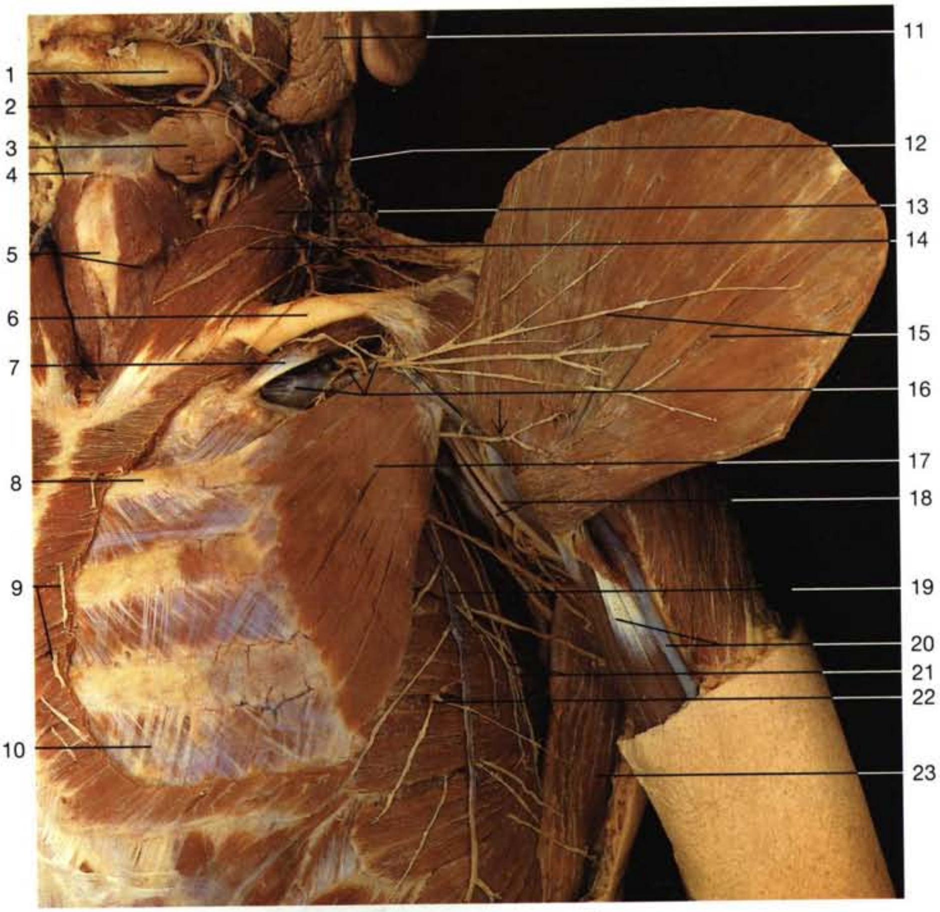

Thoracic wall I (anterior aspect). Left pectoralis major muscle has been divided and reflected. Note the connection of the cephalic vein with the subclavian vein. Arrow: medial pectoral nerve.

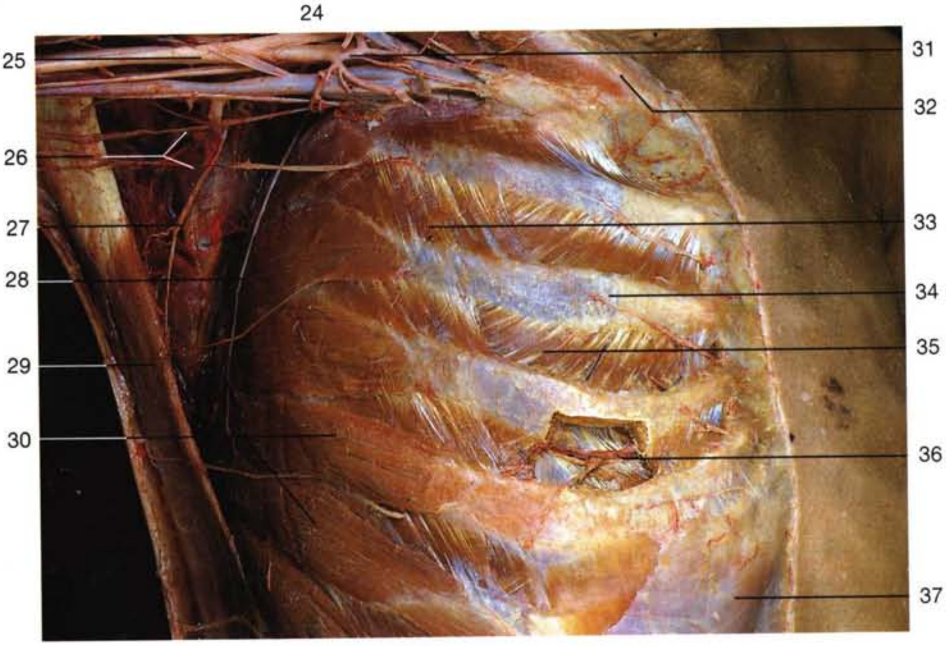

Thoracic wall (lateral aspect). Pectoralis major and minor muscles have been removed. A section of the 4th rib has been cut and removed to display the intercostal vessels and nerve.

1 Mandible
2 Facial artery
3 Submandibular gland
4 Hyoid bone
5 Thyroid cartilage and sternohyoid muscle
6 Clavicle
7 Subclavius muscle
8 Second rib
9 Anterior cutaneous branches of intercostal nerves
10 External intercostal membrane
11 Parotid gland
12 External carotid artery
13 Sternocleidomastoid muscle and cutaneous branches of cervical plexus
14 Supraclavicular nerves
15 Pectoralis major muscle and lateral pectoral nerves
16 Thoracoacromial artery and subclavian vein
17 **Pectoralis minor muscle**
18 **Median and ulnar nerve**
19 Thoracoepigastric vein
20 Cephalic vein and long head of biceps brachii muscle
21 **Lateral thoracic artery and long thoracic nerve**
22 Lateral cutaneous branches of intercostal nerve
23 Latissimus dorsi muscle
24 Median nerve
25 Axillary artery
26 Intercostobrachial nerves
27 **Thoracodorsal nerve**
28 Long thoracic nerve
29 Latissimus dorsi muscle
30 **Serratus anterior muscle**
31 Thoracoacromial artery
32 Clavicle
33 External intercostal muscle
34 Third rib
35 **Internal intercostal muscle**
36 Anterior intercostal artery and vein
37 Costal arch or margin

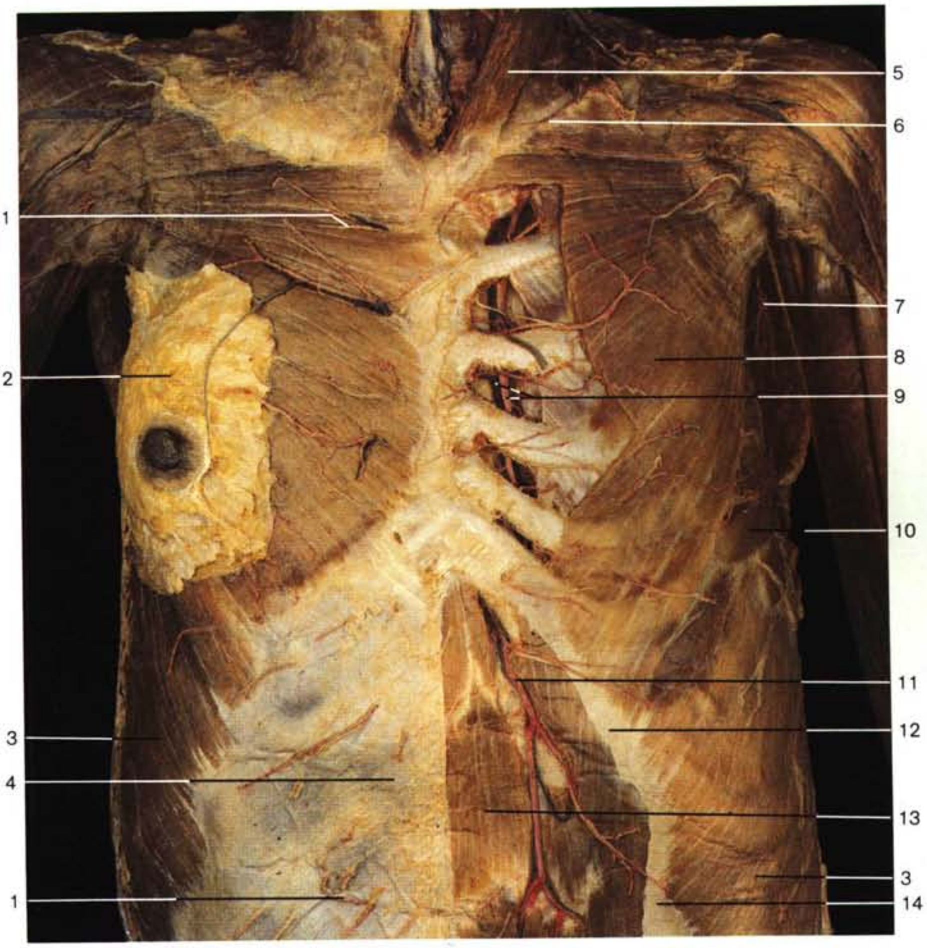

1 Anterior perforating branches of intercostal nerve
2 Mammary gland
3 External abdominal oblique muscle
4 Rectus sheath (anterior layer)
5 Sternocleidomastoid muscle
6 Clavicle
7 Lateral thoracic artery and vein
8 Pectoralis major muscle
9 **Internal thoracic artery and vein**
10 Serratus anterior muscle
11 **Superior epigastric artery and vein**
12 Costal margin
13 Rectus abdominis muscle
14 Cut edge of the anterior layer of the rectus sheath
15 Subclavian artery
16 Highest intercostal artery
17 Internal thoracic artery
18 Musculophrenic artery
19 Superficial epigastric artery
20 Deep circumflex iliac artery
21 Superior epigastric artery
22 Inferior epigastric artery
23 Superficial circumflex iliac artery

Thoracic wall II (anterior aspect). Dissection of the **internal thoracic artery and vein.** Left pectoralis major muscle partly removed. Anterior lamina of the rectus sheath on the left side has been removed.

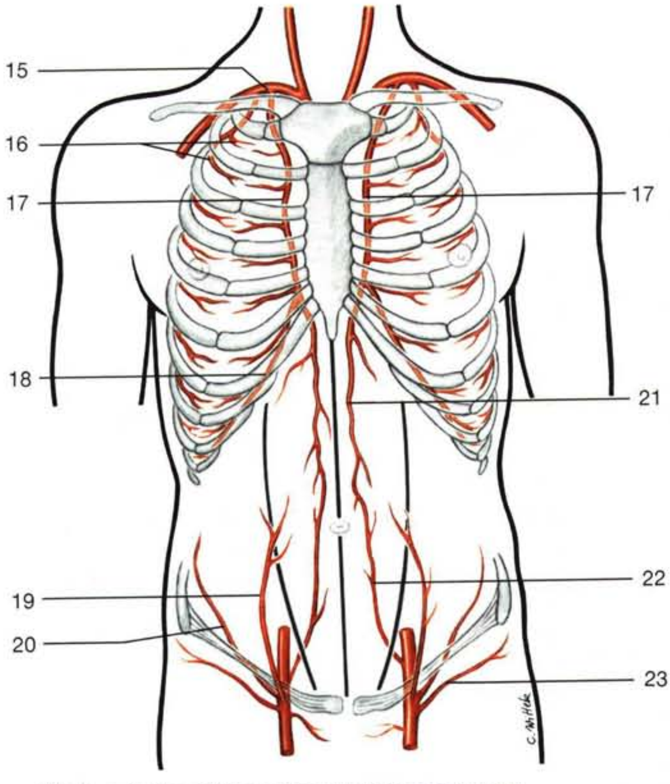

Main arteries of thoracic and abdominal wall.

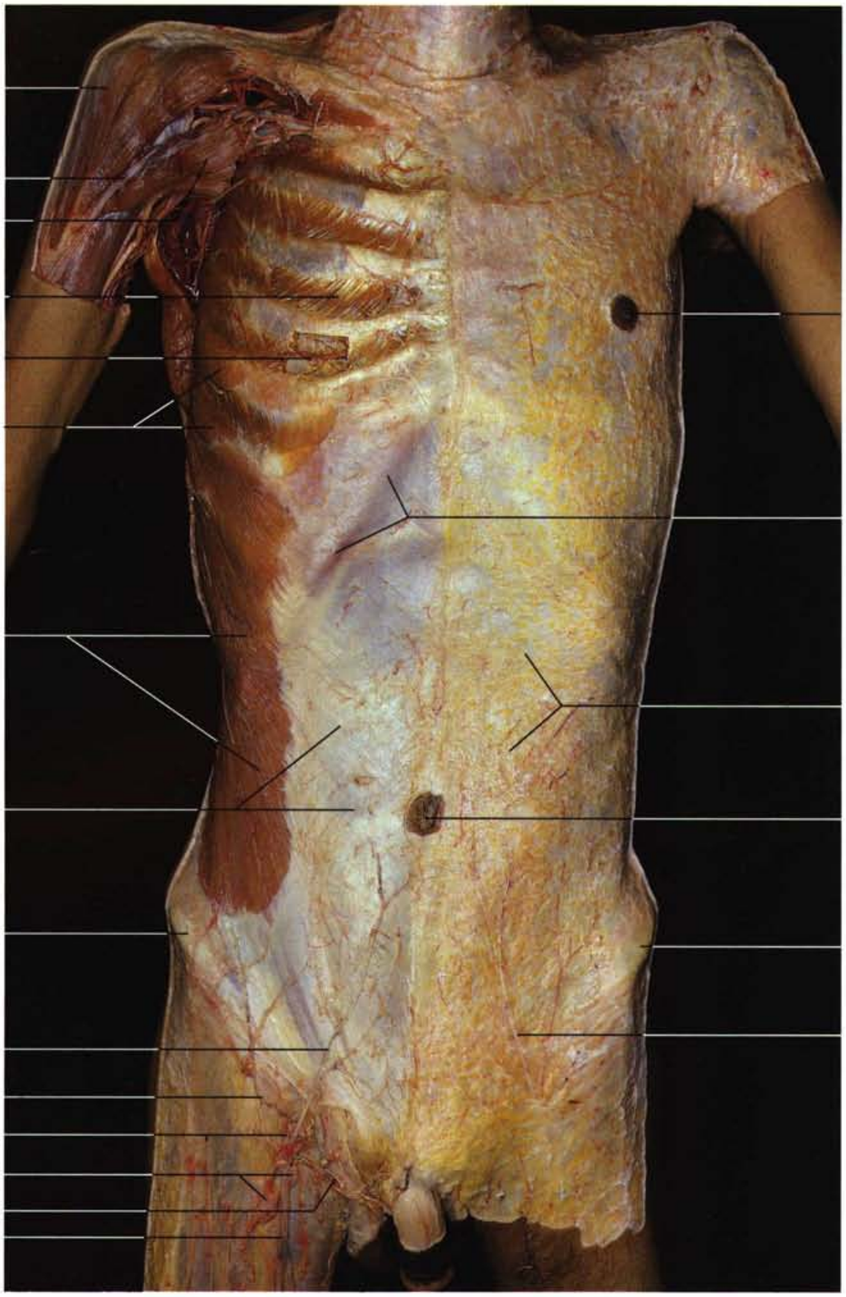

1 Deltoid muscle
2 Cephalic vein
3 Pectoralis major muscle (divided)
4 **Internal intercostal muscle**
5 **Intercostal artery and vein** (intercostal space, fenestrated)
6 Serratus anterior muscle
7 **External abdominal oblique muscle**
8 **Anterior layer of rectus sheath**
9 Iliac crest
10 Superficial epigastric vein
11 Superficial circumflex iliac vein
12 Saphenous opening
13 Superficial inguinal lymph nodes
14 Superficial external pudendal veins
15 Great saphenous vein
16 Nipple
17 Costal margin
18 Subcutaneous fatty tissue
19 Umbilicus
20 Anterior layer of rectus sheath
21 Rectus abdominis muscle
22 Posterior layer of rectus sheath
23 **Internal abdominal oblique muscle**
24 **External abdominal oblique muscle**
25 **Transversus abdominis muscle**
26 Transversal fascia and peritoneum
27 Psoas major muscle
28 Body of lumbar vertebra (L_4)
29 Quadratus lumborum muscle
30 Medial tract of erector spinae muscle
31 Lateral tract of erector spinae muscle (longissimus and iliocostalis muscles)
32 Small intestine
33 Left ureter
34 **Abdominal aorta**
35 **Inferior vena cava**
36 Descending colon
37 Spinous process

Thoracic and abdominal wall I. Right pectoralis major and minor muscles are divided. Muscles of thoracic and abdominal wall on right side are displayed.

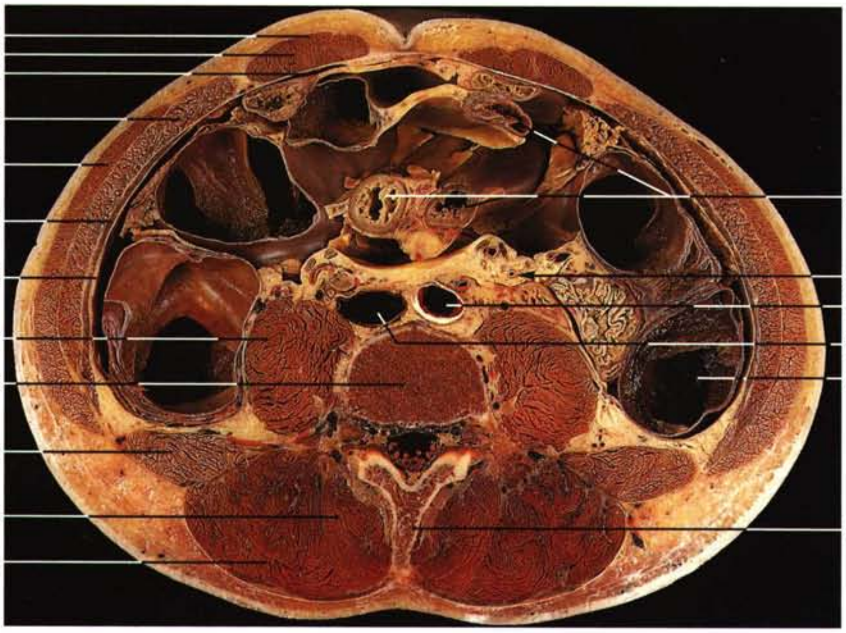

Horizontal section of the trunk at the level of the umbilicus, superior to arcuate line (inferior aspect).

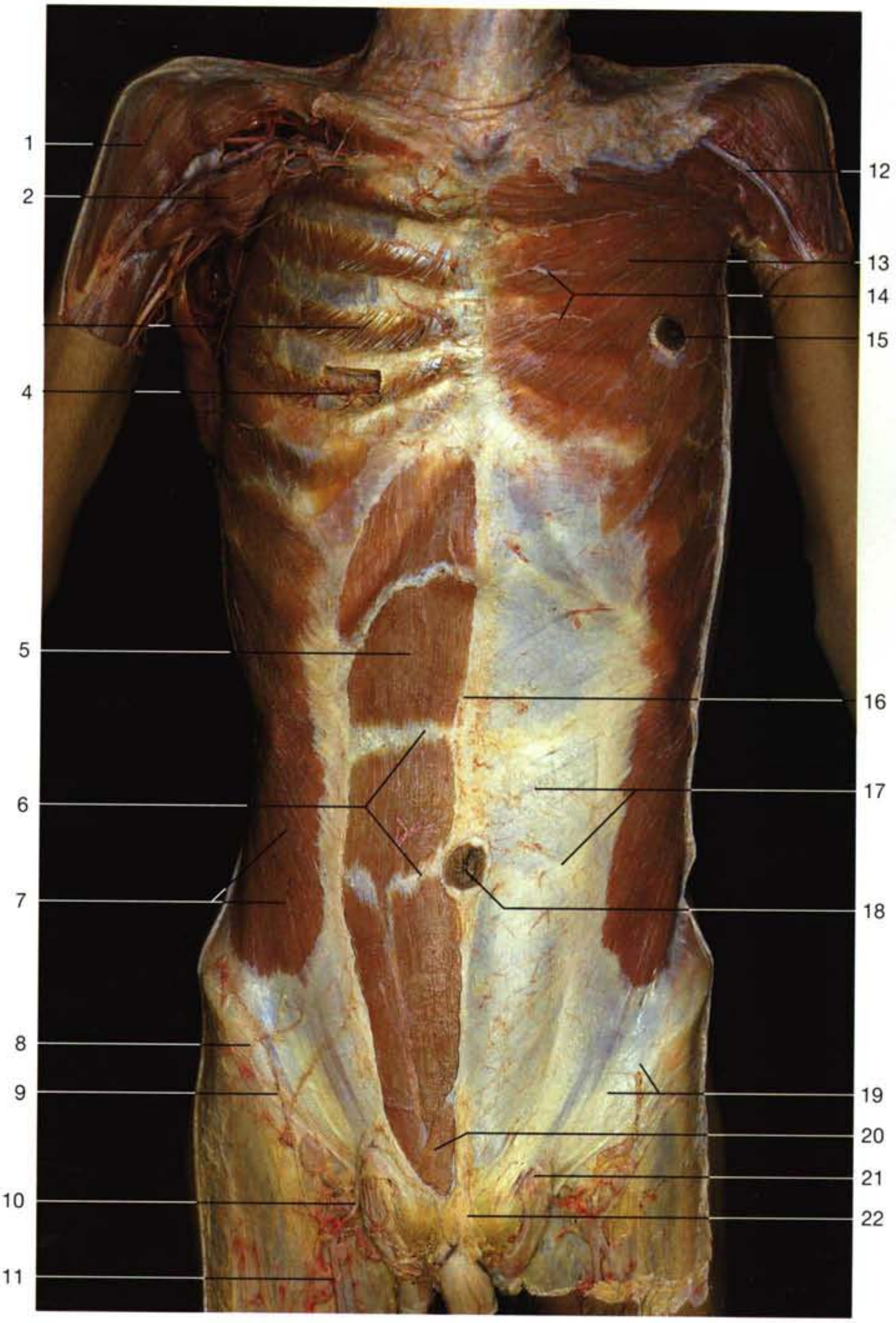

Thoracic and abdominal wall II. Right pectoralis major and minor muscles and anterior layer of rectus sheath have been removed on the right side.

1 Deltoid muscle
2 Pectoralis major muscle (divided)
3 **Internal intercostal muscle**
4 Intercostal artery and vein
5 **Rectus abdominis muscle**
6 Tendinous intersections
7 **External abdominal oblique muscle**
8 Anterior superior iliac spine
9 Superficial circumflex iliac vein
10 Superficial epigastric vein
11 Great saphenous vein
12 Cephalic vein
13 **Pectoralis major muscle**
14 Anterior cutaneous branches of intercostal nerves
15 Nipple
16 Linea alba
17 **Anterior layer of rectus sheath**
18 Umbilicus
19 **Inguinal ligament**
20 Pyramidalis muscle
21 Superficial inguinal ring and spermatic cord
22 Suspensory ligament of penis
23 Longissimus and iliocostalis muscles
24 Multifidus muscle
25 Quadratus lumborum muscle
26 Latissimus dorsi muscle
27 Psoas major muscle
28 Spinous process
29 Body of 1st lumbar vertebra
30 **Transversus abdominis muscle**
31 **Internal abdominal oblique muscle**

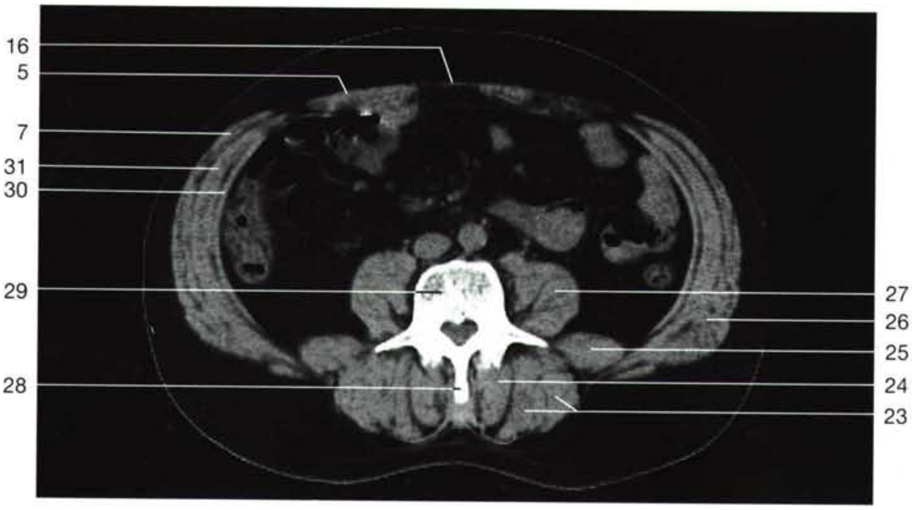

Horizontal section through the body at the level of 4th lumbar vertebra; seen from below. (CT-Scan.)

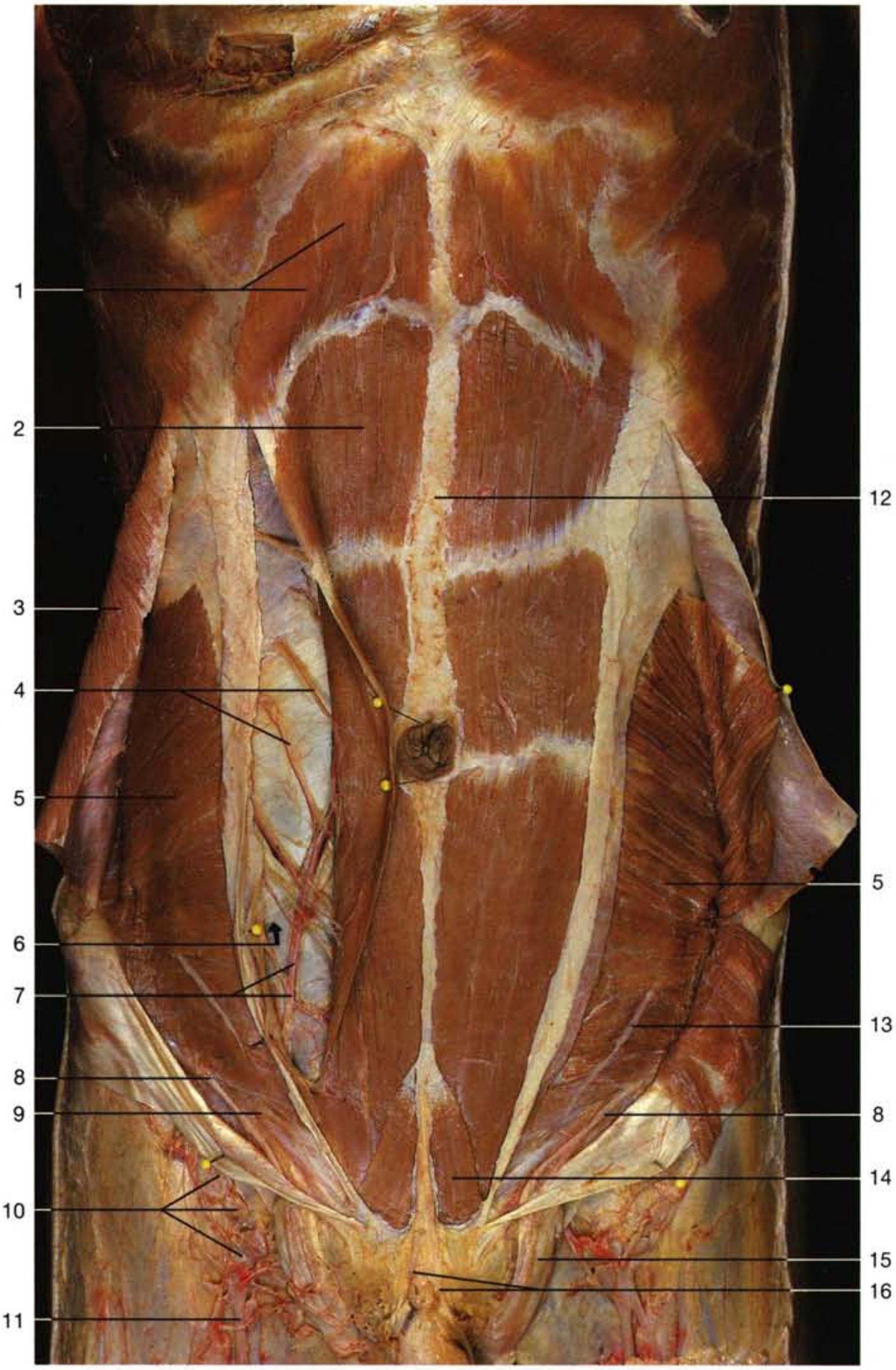

1. Costal margin
2. **Rectus abdominis muscle**
3. External abdominal oblique muscle (reflected)
4. Thoracoabdominal (intercostal) nerves with accompanying vessels
5. Internal abdominal oblique muscle
6. **Arcuate line** (arrow)
7. **Inferior epigastric artery and vein**
8. Ilioinguinal nerve
9. Position of deep inguinal ring
10. Superficial inguinal lymph nodes
11. Great saphenous vein
12. Linea alba
13. Iliohypogastric nerve
14. **Pyramidalis** muscle
15. Spermatic cord
16. Fundiform ligament of penis

Thoracic and abdominal wall III. External abdominal oblique muscle has been divided and reflected on both sides. The right rectus muscle has been reflected medially to display the posterior layer of rectus sheath. Arrow: location of arcuate line.

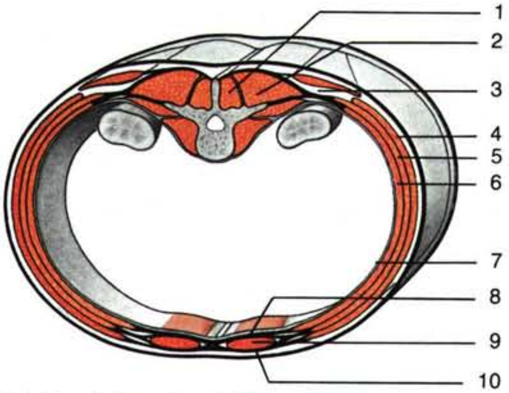

Horizontal-section of the trunk superior to arcuate line. (Schematic drawing.)

1. Medial column of intrinsic muscles of the back
2. Lateral column of erector spinae muscle
3. Thoracolumbar fascia with superficial and deep layer
4. External oblique muscle
5. Internal oblique muscle
6. Transversus abdominis muscle
7. Transversalis fascia
8. Posterior layer of rectus sheath
9. Rectus abdominis muscle
10. Anterior layer of rectus sheath

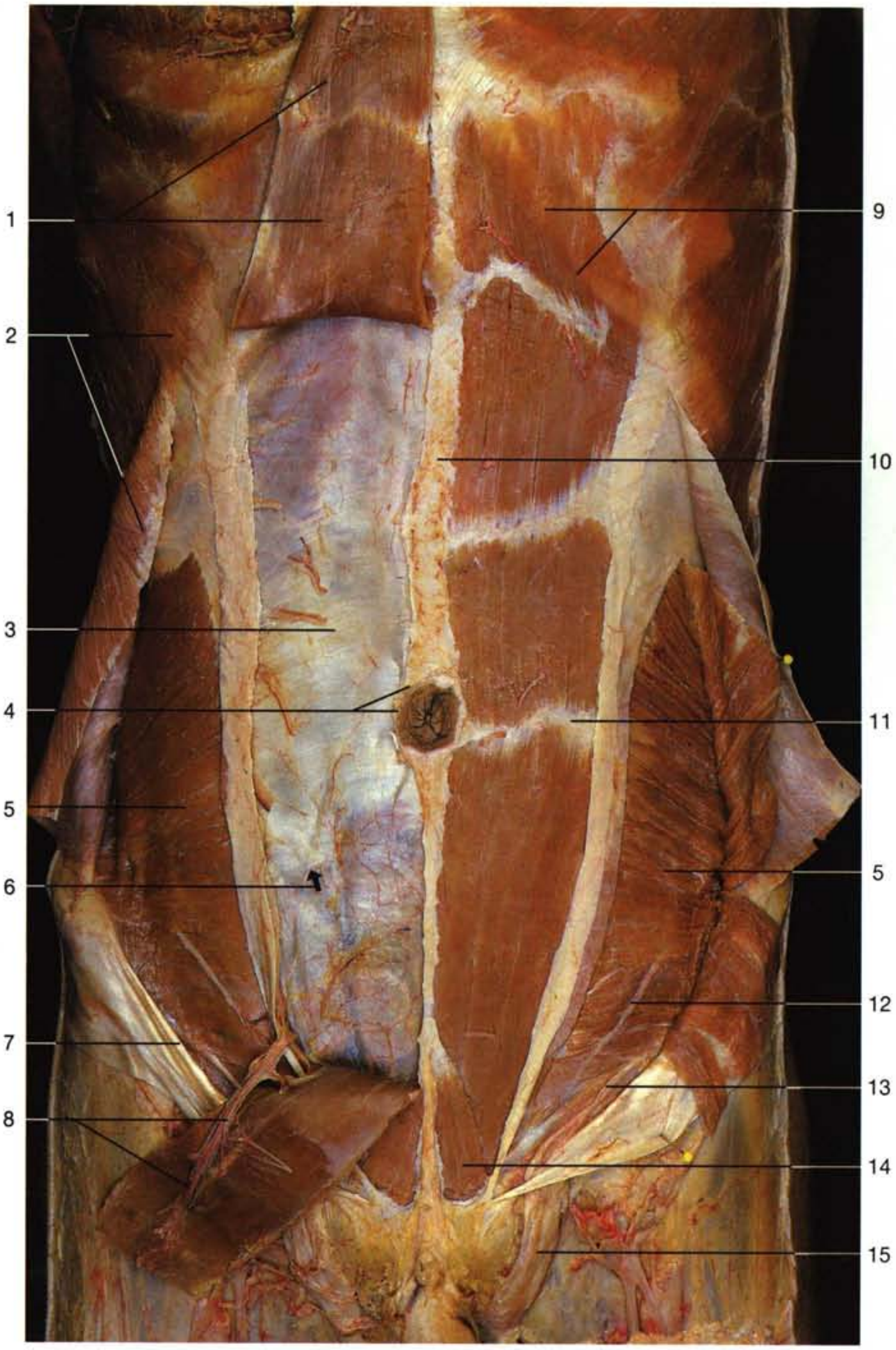

1 **Rectus abdominis muscle** (reflected)
2 External abdominal oblique muscle (divided)
3 Posterior layer of rectus sheath
4 Umbilical ring
5 **Internal abdominal oblique muscle**
6 Arcuate line (arrow)
7 Inguinal ligament
8 **Inferior epigastric artery and vein** and rectus abdominis muscle (divided and reflected)
9 Costal margin
10 Linea alba
11 Tendinous intersection
12 Iliohypogastric nerve
13 **Ilioinguinal nerve**
14 **Pyramidalis** muscle
15 Spermatic cord

Thoracic and abdominal wall IV. External abdominal oblique muscle has been divided and reflected on both sides. The right rectus muscle has been cut and reflected to display the posterior layer of rectus sheath. Arrow: location of arcuate line.

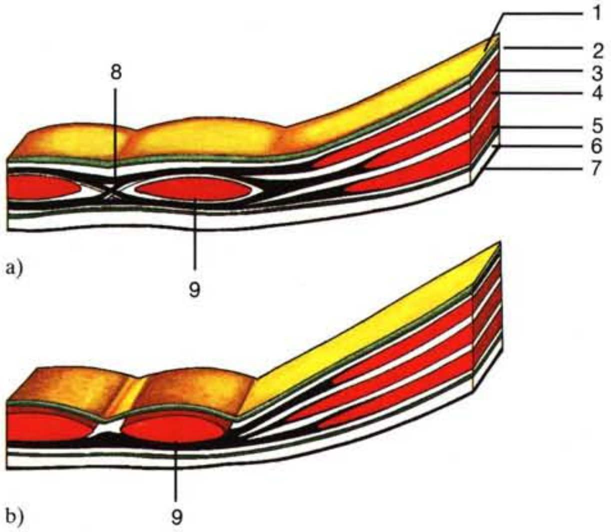

1 Peritoneum
2 Transversalis fascia (green)
3 Transversus abdominis muscle
4 Internal abdominal oblique muscle
5 External abdominal oblique muscle
6 Fascia of external abdominal oblique muscle (green)
7 Skin
8 Linea alba
9 Rectus abdominis muscle

Transverse sections through the abdominal wall superior (a) and inferior (b) to arcuate line.

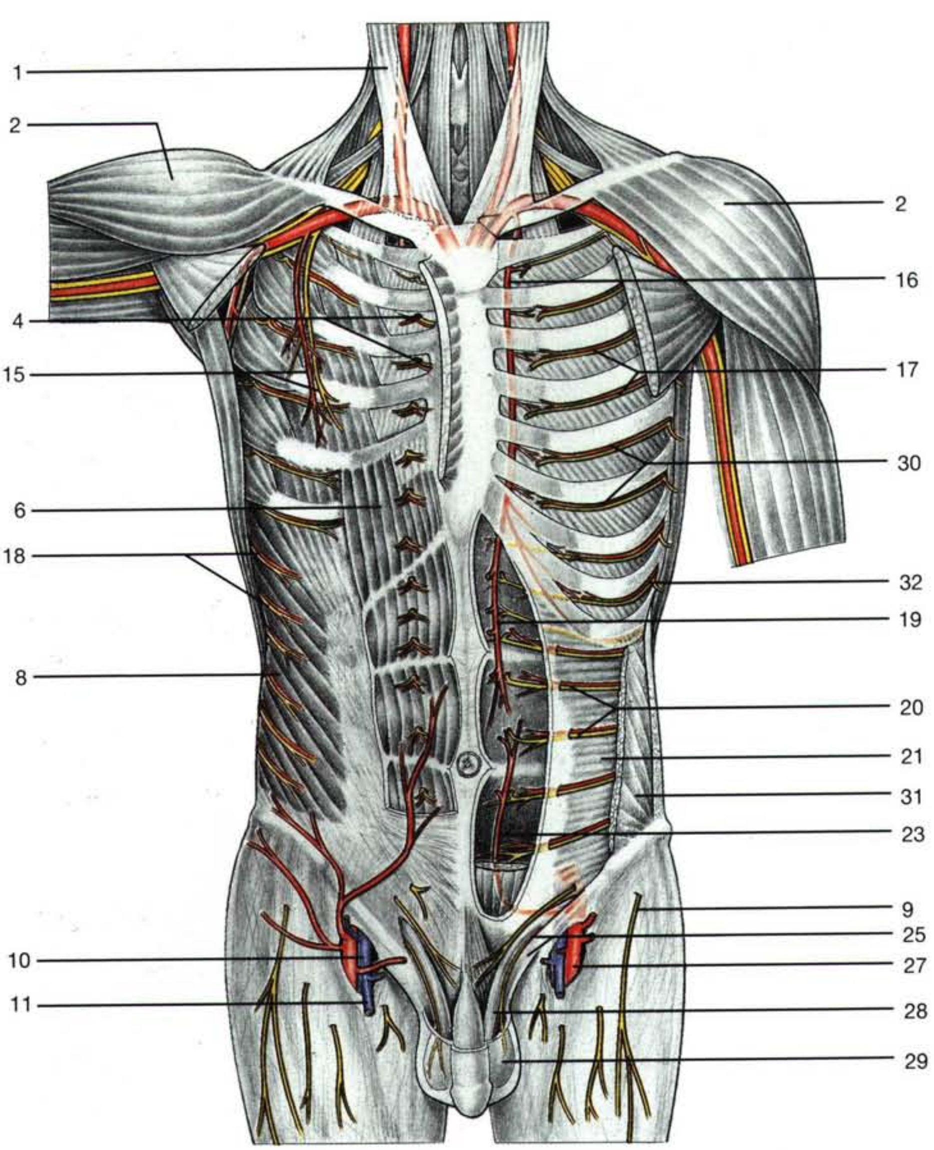

Thoracic and abdominal wall (schematic drawing). Note the segmental organization of the blood vessels and nerves. Right side: superficial layers; left side: deeper layers.

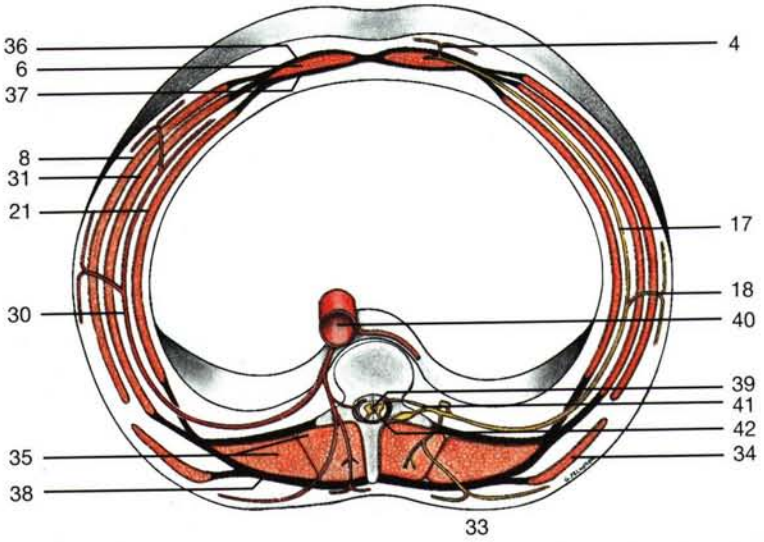

Horizontal section of the abdominal wall (from above) showing the location of the intercostal arteries (left side) and nerves (right side).

1 Sternocleidomastoid muscle
2 Deltoid muscle
3 Pectoralis major muscle
4 **Anterior cutaneous branches of intercostal nerves**
5 Cut edge of anterior layer of rectus sheath
6 **Rectus abdominis muscle**
7 Tendinous intersection
8 External abdominal oblique muscle
9 Lateral femoral cutaneous nerve
10 Femoral vein
11 Great saphenous vein
12 Medial supraclavicular nerves
13 Pectoralis minor muscle (reflected) and medial pectoral nerves
14 Axillary vein
15 Long thoracic nerve and lateral thoracic artery
16 Internal thoracic artery
17 **Intercostal nerves**
18 **Lateral cutaneous branches of intercostal nerves**
19 Superior epigastric artery
20 Thoracoabdominal (intercostal) nerves
21 Transversus abdominis muscle
22 Posterior layer of rectus sheath
23 **Inferior epigastric artery**
24 Lateral femoral cutaneous nerve
25 Inguinal ligament and ilioinguinal nerve
26 Femoral nerve
27 Femoral artery
28 Spermatic cord
29 Testis
30 **Posterior intercostal arteries**
31 Internal abdominal oblique muscle
32 Lateral cutaneous branch of intercostal nerve
33 Dorsal branch of spinal nerve
34 Latissimus dorsi muscle
35 Deep muscles of the back (medial and lateral tract)
36 Anterior layer of rectus sheath
37 Posterior layer of rectus sheath
38 Thoracolumbar fascia
39 Spinal cord
40 Aorta
41 Ventral root ⎫ of spinal
42 Dorsal root ⎭ nerve

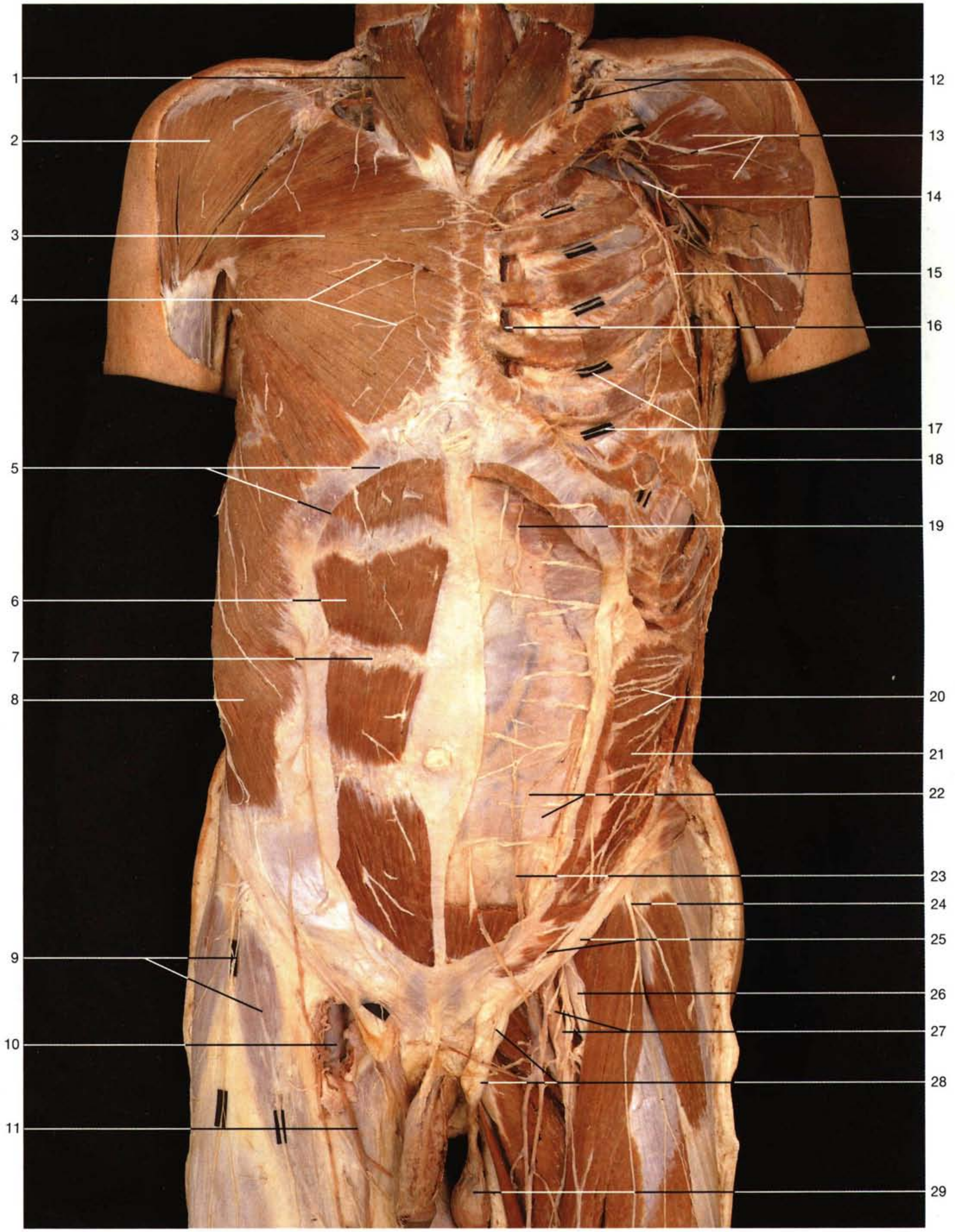
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

Thoracic wall and abdominal wall V. Right side: superficial layers; left side: deeper layers (anterior aspect). Pectoralis major and minor muscles, the external and internal intercostal muscles on the left side have been removed to display the intercostal nerves. The anterior layer of rectus sheath, the left rectus abdominis muscle and the external and internal abdominal oblique muscle have been removed to show the thoracoabdominal nerves within the abdominal wall.

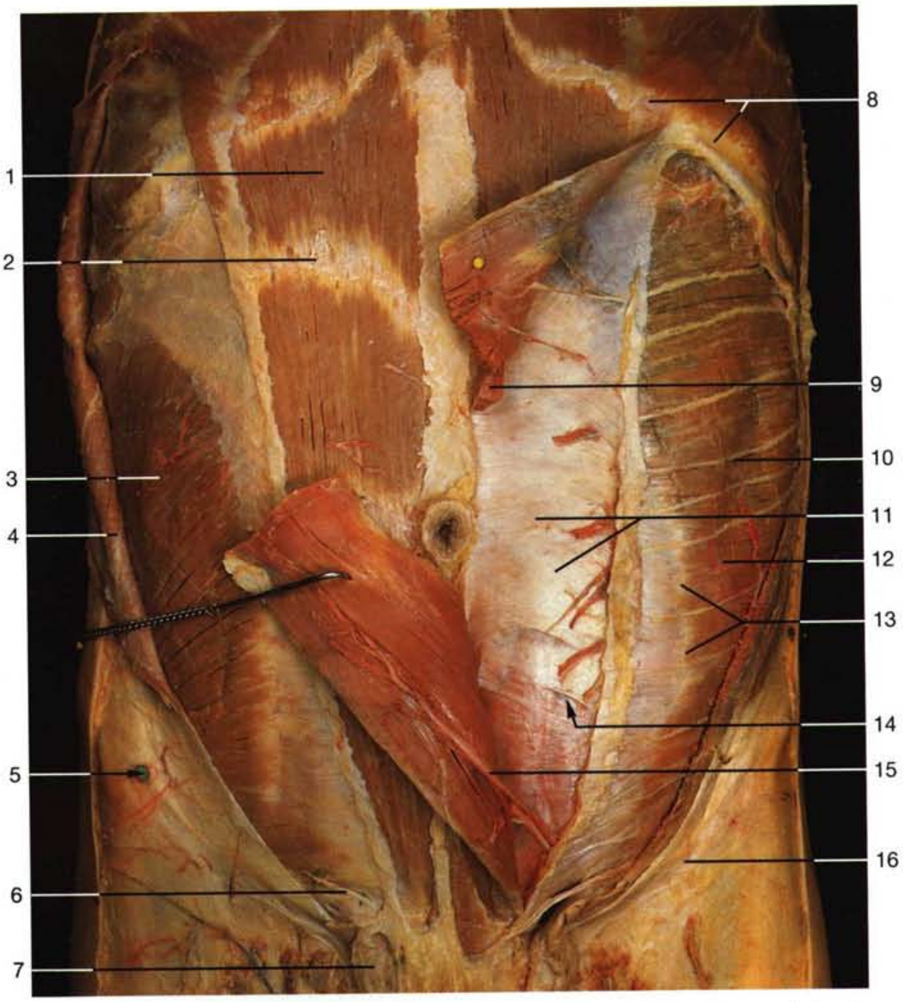

1 **Rectus abdominis muscle**
2 Tendinous intersection
3 **Internal abdominal oblique muscle**
4 External abdominal oblique muscle (reflected)
5 Anterior superior iliac spine
6 Ilioinguinal nerve
7 Spermatic cord
8 Costal margin
9 Superior epigastric artery
10 **Thoracoabdominal** (intercostal) **nerves**
11 Posterior layer of rectus sheath
12 **Transversus abdominis muscle**
13 Semilunar line
14 Arcuate line
15 **Inferior epigastric artery**
16 Inguinal ligament

Abdominal wall with vessels and nerves. The left rectus abdominis muscle has been divided and reflected to display the inferior epigastric vessels. The left internal abdominal oblique muscle has been removed to show the thoracoabdominal nerves.

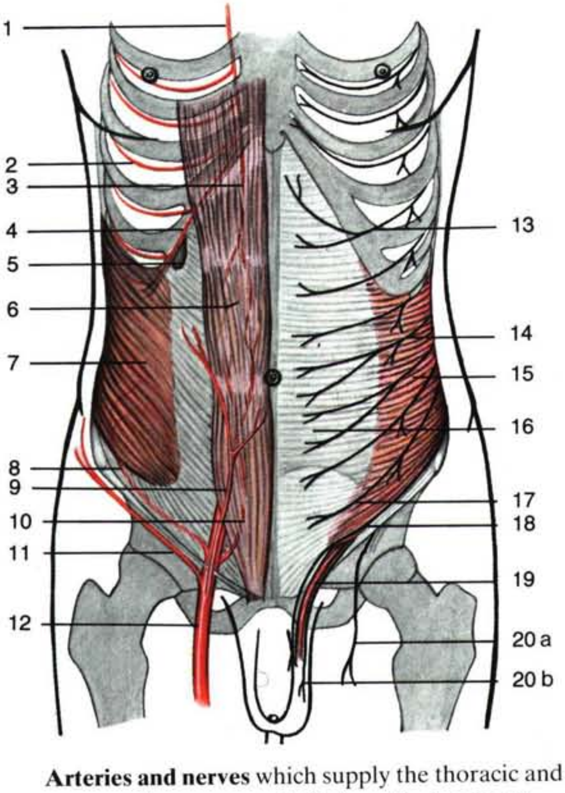

1 Internal thoracic artery
2 Intercostal artery
3 Superior epigastric artery
4 Musculophrenic artery
5 Gallbladder
6 Rectus abdominis muscle
7 External abdominal oblique muscle
8 Deep circumflex iliac artery
9 Superficial epigastric artery
10 Inferior epigastric artery
11 Superficial circumflex iliac artery
12 Femoral artery
13 Intercostal nerve
14 Thoracoabdominal nerve (T_{10})
15 Transversus abdominis muscle
16 Posterior layer of the rectus sheath
17 Iliohypogastric nerve (L_1)
18 Ilioinguinal nerve (L_1)
19 Spermatic cord
20 Genitofemoral nerve (L_1, L_2)
  a Femoral branch
  b Genital branch

Arteries and nerves which supply the thoracic and abdominal wall. Note their segmental arrangement. (Schematic drawing.)

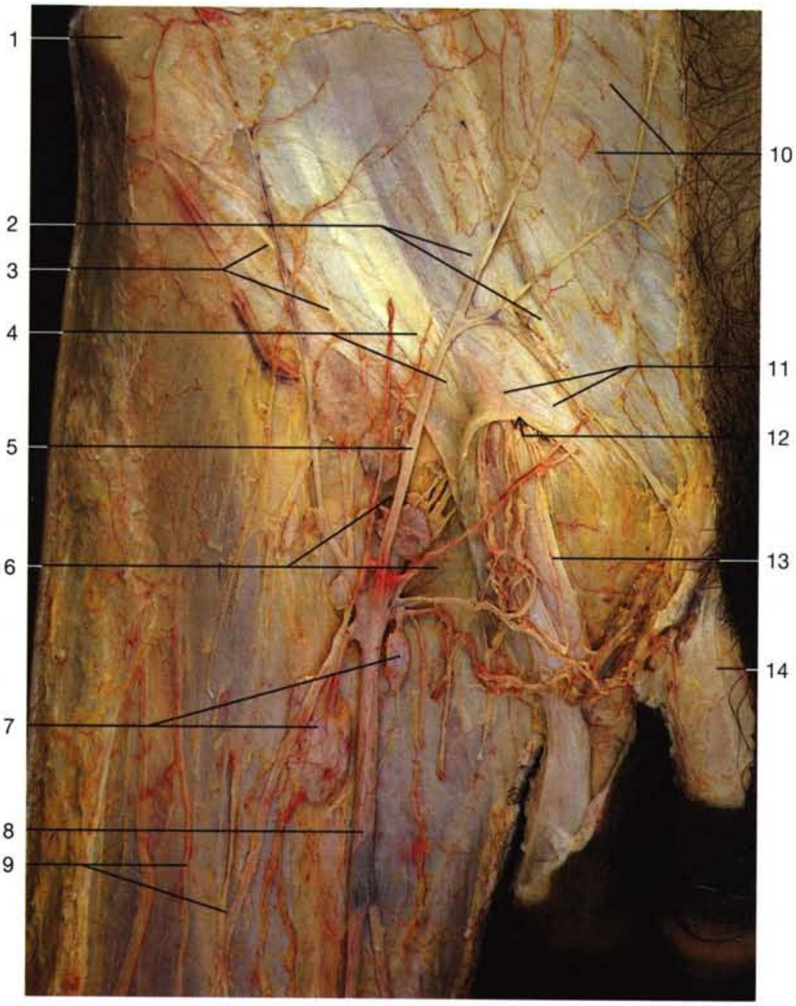

Inguinal canal in the male I. Dissection of superficial layer, right side (anterior aspect).

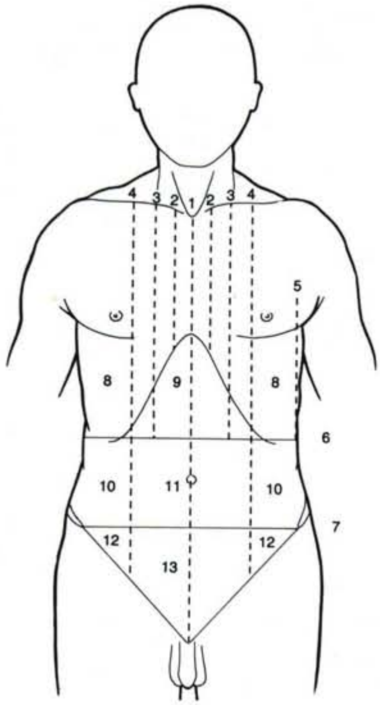

Regions and reference lines
for delineating surface projections.

Reference lines and regions

1. Median line
2. Lateral sternal line
3. Parasternal line
4. Left lateral or midclavicular line
5. Midaxillary line
6. Transpyloric plane
7. Transtubercular plane
8. Hypochondriac region
9. Epigastric region
10. Lumbar region
11. Umbilical region
12. Iliac region
13. Hypogastric region

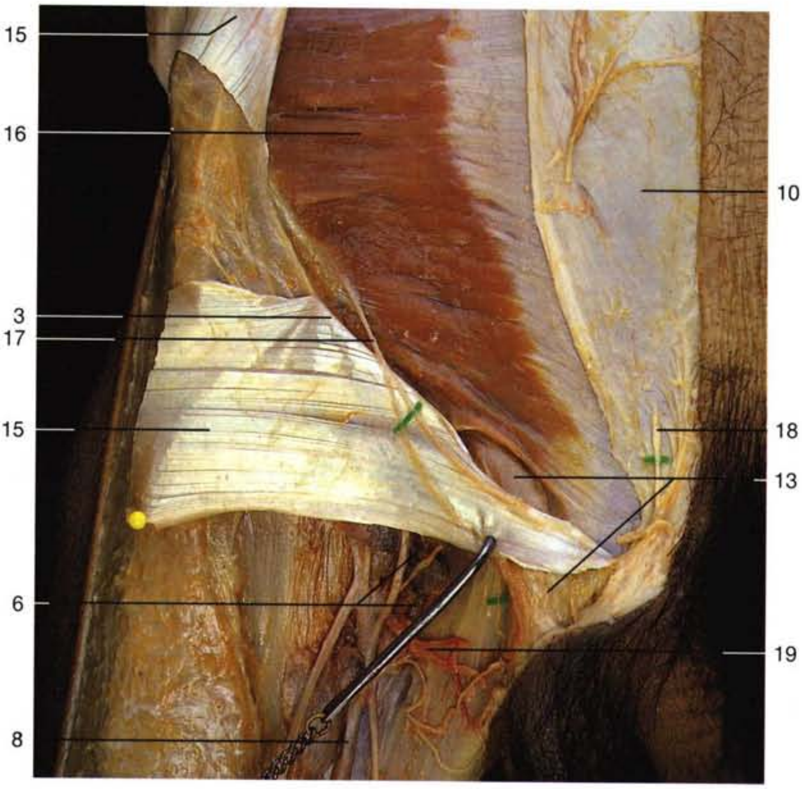

Inguinal canal in the male II, right side (anterior aspect).
The external abdominal oblique muscle has been divided to display the inguinal canal.

1. Anterior superior iliac spine
2. Medial crus of inguinal ring
3. Inguinal ligament
4. Lateral crus of inguinal ring
5. Superficial epigastric vein
6. Saphenous opening
7. **Superficial inguinal lymph nodes**
8. **Great saphenous vein**
9. Anterior cutaneous branches of femoral nerve
10. **Anterior layer of rectus sheath**
11. Intercrural fibers
12. **Superficial inguinal ring**
13. Spermatic cord and genital branch of genitofemoral nerve
14. Penis
15. Aponeurosis of external abdominal oblique muscle (divided and reflected)
16. **Internal abdominal oblique muscle**
17. Ilioinguinal nerve
18. Anterior cutaneous branches of iliohypogastric nerve
19. Superficial external pudendal veins

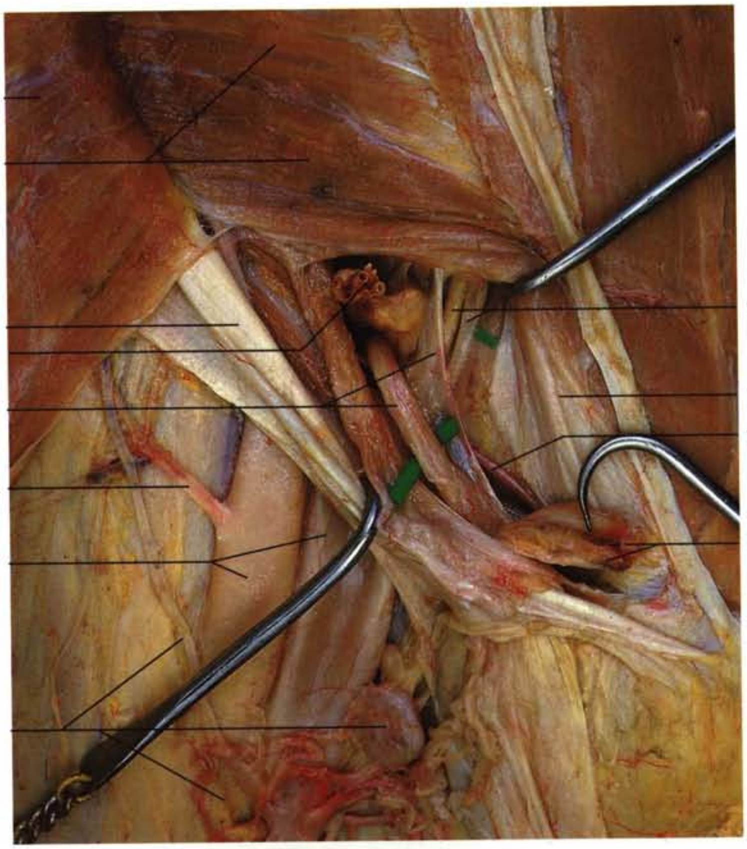

Inguinal canal in the male III. Deep dissection (anterior aspect, right side). Spermatic cord with exception of ductus deferens (probe) has been divided and reflected.

1 Internal abdominal oblique muscle (reflected)
2 Transversus abdominis muscle
3 Inguinal ligament
4 **Spermatic cord** with the exception of the ductus deferens (divided and reflected)
5 **Ductus deferens** and interfoveolar ligament
6 Superficial circumflex iliac artery
7 **Femoral artery and vein**
8 Superficial inguinal lymph nodes and inguinal lymph vessel
9 **Inferior epigastric artery and vein**
10 Falx inguinalis or conjoint tendon (cut)
11 Pubic branch of inferior epigastric artery
12 Superficial inguinal ring
13 Penis
14 External abdominal oblique muscle
15 Anterior superior iliac spine
16 Intercrural fibers
17 Fascia lata and sartorius muscle
18 Saphenous opening and great saphenous vein
19 Deep inguinal ring
20 Skin of scrotum and dartos muscle
21 Cremaster muscle
22 Internal spermatic fascia
23 Ductus deferens
24 Epididymis
25 Peritoneum (blue)
26 Remnant of processus vaginalis
27 Tunica vaginalis testis
28 Rectus abdominis muscle
29 Spermatic cord with ductus deferens covered by external spermatic fascia
30 Anterior layer of rectus sheath
31 Suspensory ligament of penis
32 **Testis and epididymis**
33 Ductus deferens
34 Pampiniform venous plexus and testicular artery
35 Inferior epigastric artery
36 Lateral femoral cutaneous nerve
37 Ilioinguinal nerve
38 **Femoral nerve**
39 Sartorius muscle
40 Deep dorsal vein of penis

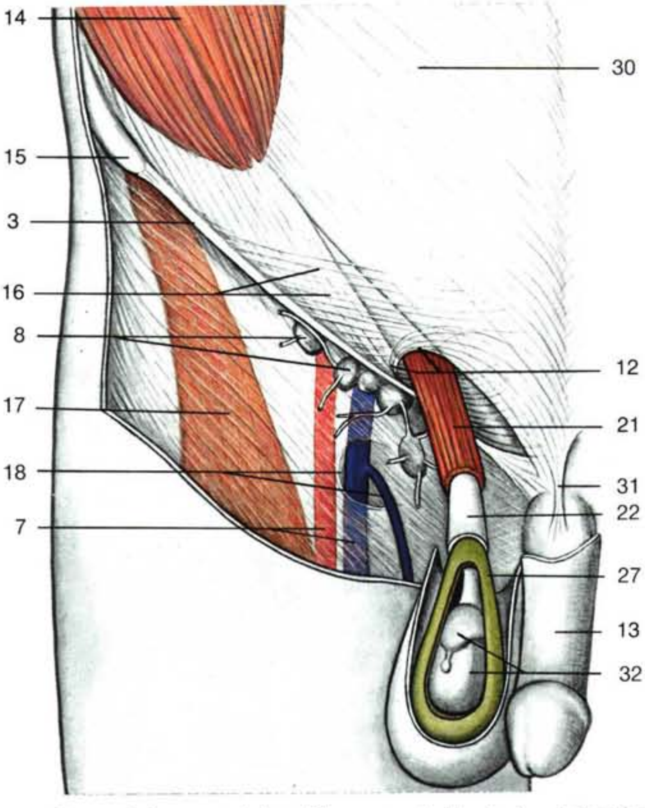

General characteristics of lower part of anterior abdominal wall and inguinal canal. (Schematic drawing.)

Inguinal hernias may either pass through the inguinal canal lateral to the inferior epigastric artery (indirect or lateral inguinal hernias, A and C) or directly penetrate the abdominal wall through the inguinal triangle located medial to the inferior epigastric artery (direct or medial inguinal hernias, B). The lateral hernias can be congenital if the vaginal process remains open (C) or acquired (A) if the hernia develops independently of a patent processus vaginalis.

Femoral hernias generally protrude through the femoral ring below the inguinal ligament. Proper assessment of the site of herniation requires the identification of both the inguinal ligament and the epigastric artery.

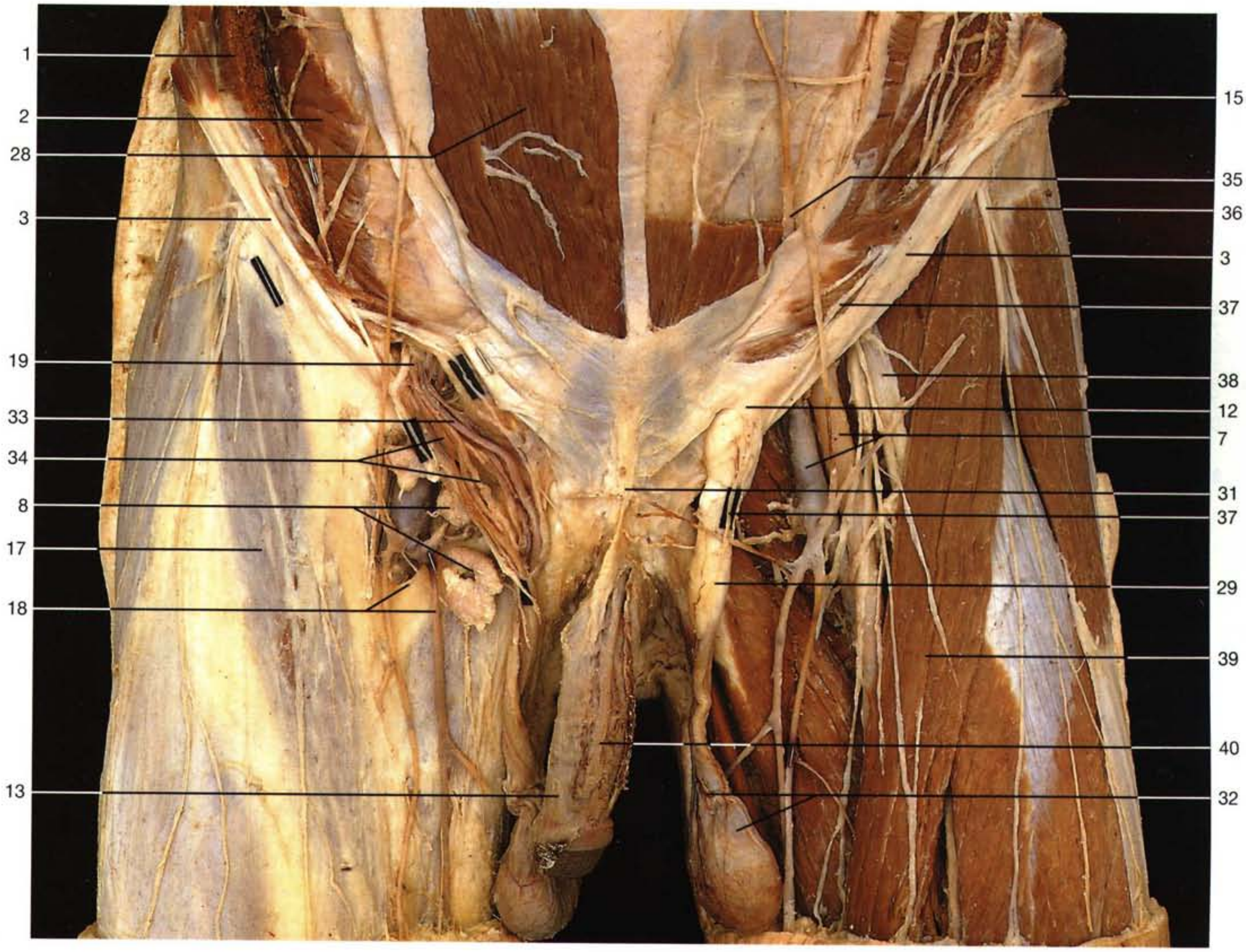
1
2
28
3
19
33
34
8
17
18
13
15
35
36
3
37
38
12
7
31
37
29
39
40
32

Inguinal and femoral region in the male (anterior aspect). On the right, the spermatic cord was dissected to display the ductus deferens and the accompanying vessels and nerves. The fascia lata on the left side has been removed.

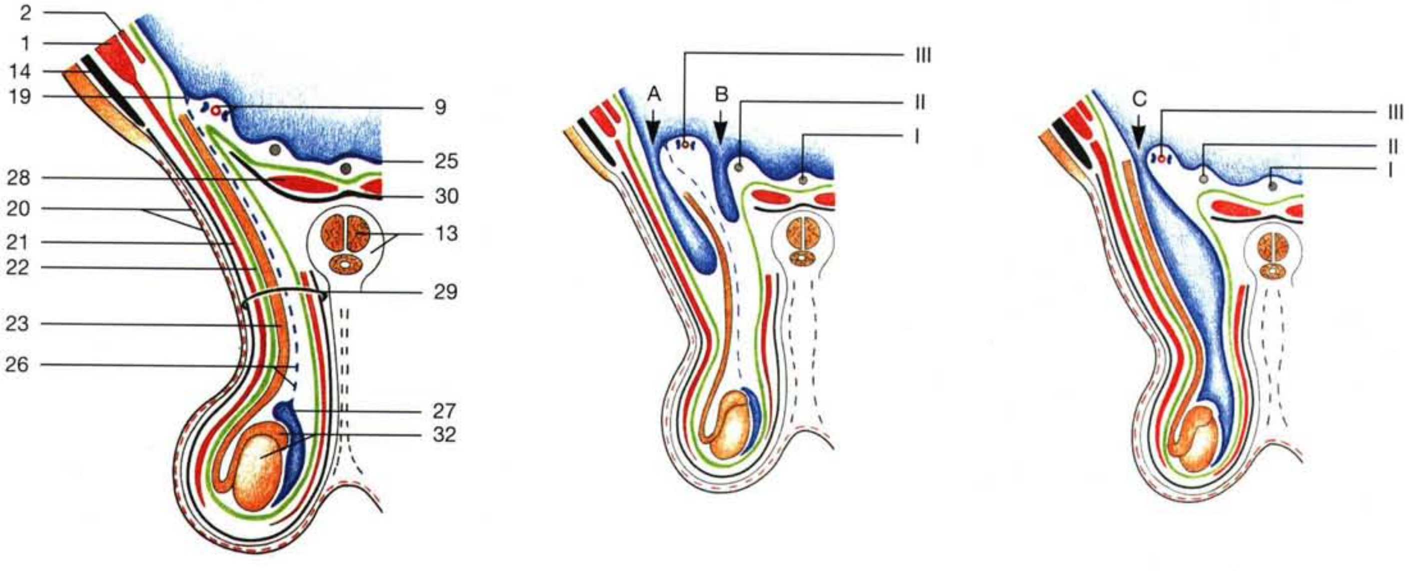
2
1
14
19
9
25
28
30
20
21
22
13
29
23
26
27
32
A
B
III
II
I
C
III
II
I

Layers of spermatic cord and types of hernias. Left: Normal situation; middle: Location of acquired inguinal hernias; A = indirect; B = direct inguinal hernia. Right: congenital indirect inguinal hernia (C); the vaginal process remained open.

I = Median umbilical fold, containing urachus chord.
II = Medial umbilical fold with remnants of umbilical artery and vein.
III = Lateral umbilical fold with inferior epigastric artery and vein.

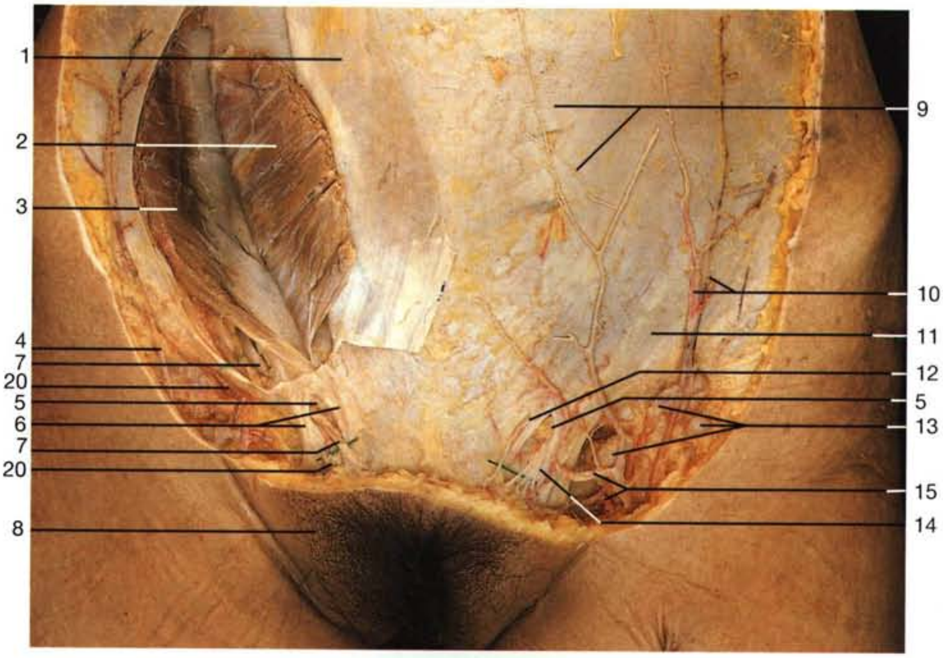

Inguinal region in the female (anterior aspect). Left side: superficial layer; right side: external and internal abdominal oblique muscle divided and reflected.

1 Aponeurosis of external abdominal oblique muscle
2 Internal abdominal oblique muscle (divided and reflected)
3 Transversus abdominis muscle
4 Superficial circumflex iliac artery and vein
5 Superficial inguinal ring with fat pad
6 Medial and lateral crural fibers
7 **Round ligament** (ligamentum teres uteri)
8 Labium majus pudendi
9 Anterior layer of rectus sheath
10 Superficial epigastric artery and vein
11 **Inguinal ligament**
12 Cutaneous branch of ilioinguinal nerve
13 Superficial inguinal lymph nodes
14 Entrance of round ligament into the labium majus
15 External pudendal artery and vein
16 Position of deep inguinal ring
17 **Ilioinguinal nerve**
18 Internal abdominal oblique muscle
19 Pubic branch of inferior epigastric artery
20 Genital branch of genitofemoral nerve
21 Fat pad of inguinal canal
22 Ilioinguinal nerve
23 Sheath of round ligament (inguinal canal)
24 Transversalis fascia

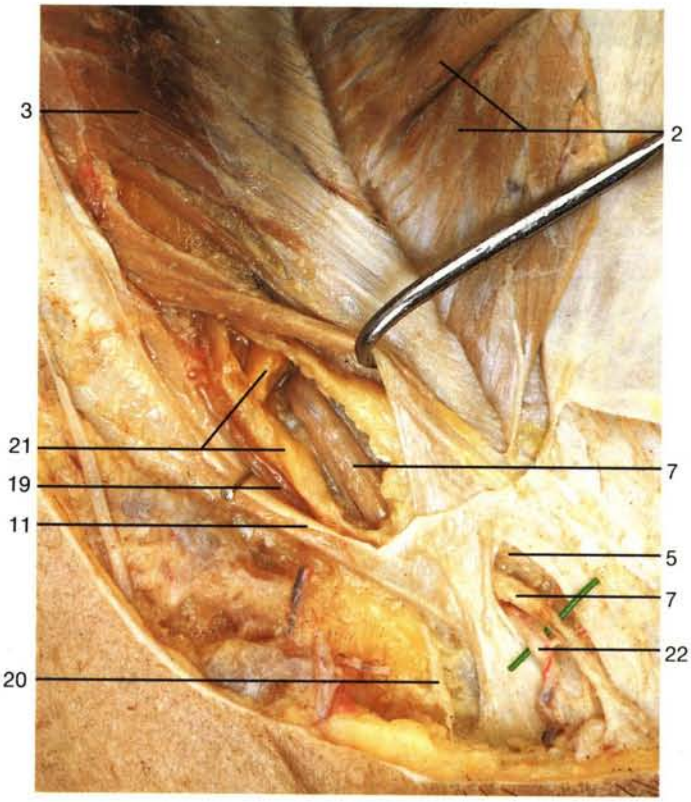

Inguinal canal of the female I (anterior aspect, right side). The external abdominal oblique muscle has been divided and reflected, to display the ilioinguinal nerve and the round ligament.

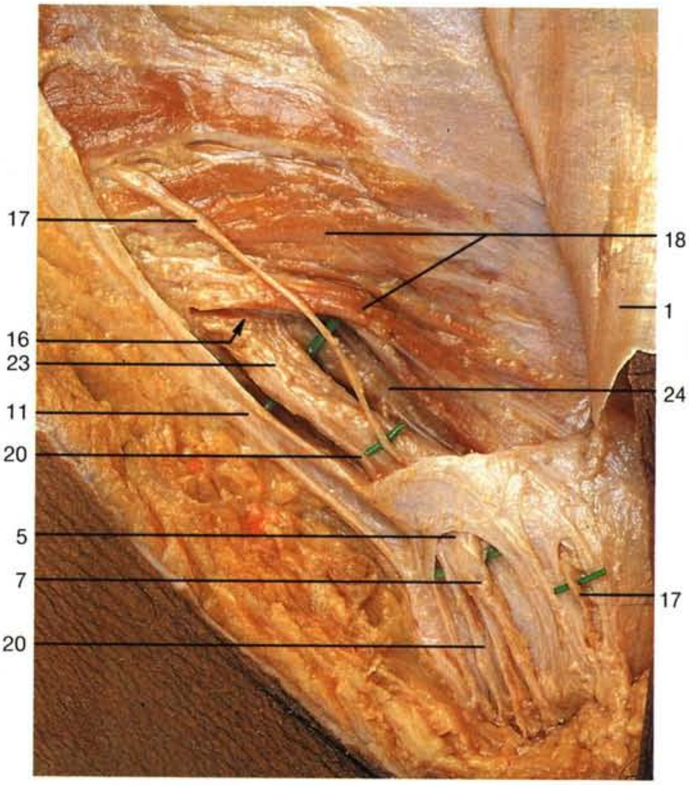

Inguinal canal of the female II (anterior aspect, right side). The external and internal abdominal oblique muscle have been divided and reflected to show the content of the inguinal canal.

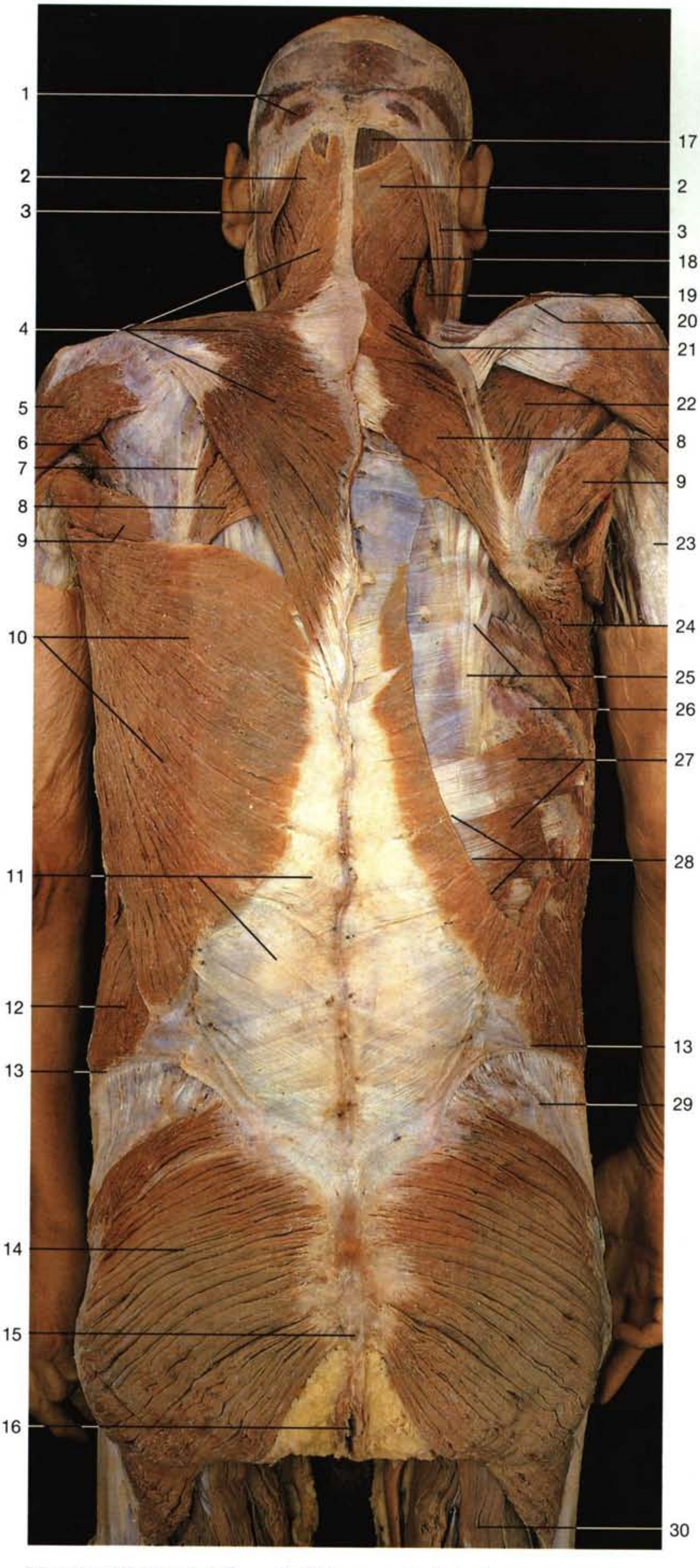

1 Occipital belly of occipitofrontalis muscle
2 Splenius capitis muscle
3 Sternocleidomastoid muscle
4 **Trapezius muscle**
5 Deltoid muscle
6 Teres minor muscle
7 Medial margin of scapula
8 **Rhomboid major muscle**
9 Teres major muscle
10 **Latissimus dorsi muscle**
11 Thoracolumbar fascia
12 External abdominal oblique muscle
13 Iliac crest
14 **Gluteus maximus muscle**
15 Position of last coccygeal vertebra
16 Anus
17 Semispinalis capitis muscle
18 Splenius cervicis muscle
19 Levator scapulae muscle
20 Spine of scapula
21 Rhomboideus minor muscle
22 Infraspinatus muscle
23 Triceps brachii muscle
24 Serratus anterior muscle
25 **Iliocostalis muscle**
26 External intercostal muscle
27 **Serratus posterior inferior muscle**
28 Latissimus dorsi muscle (cut edge)
29 Fascia over gluteus medius muscle
30 Long head of biceps femoris muscle

Muscles of the back I. Superficial layer on the left, deeper layer on the right. Right latissimus dorsi and trapezius muscles are removed.

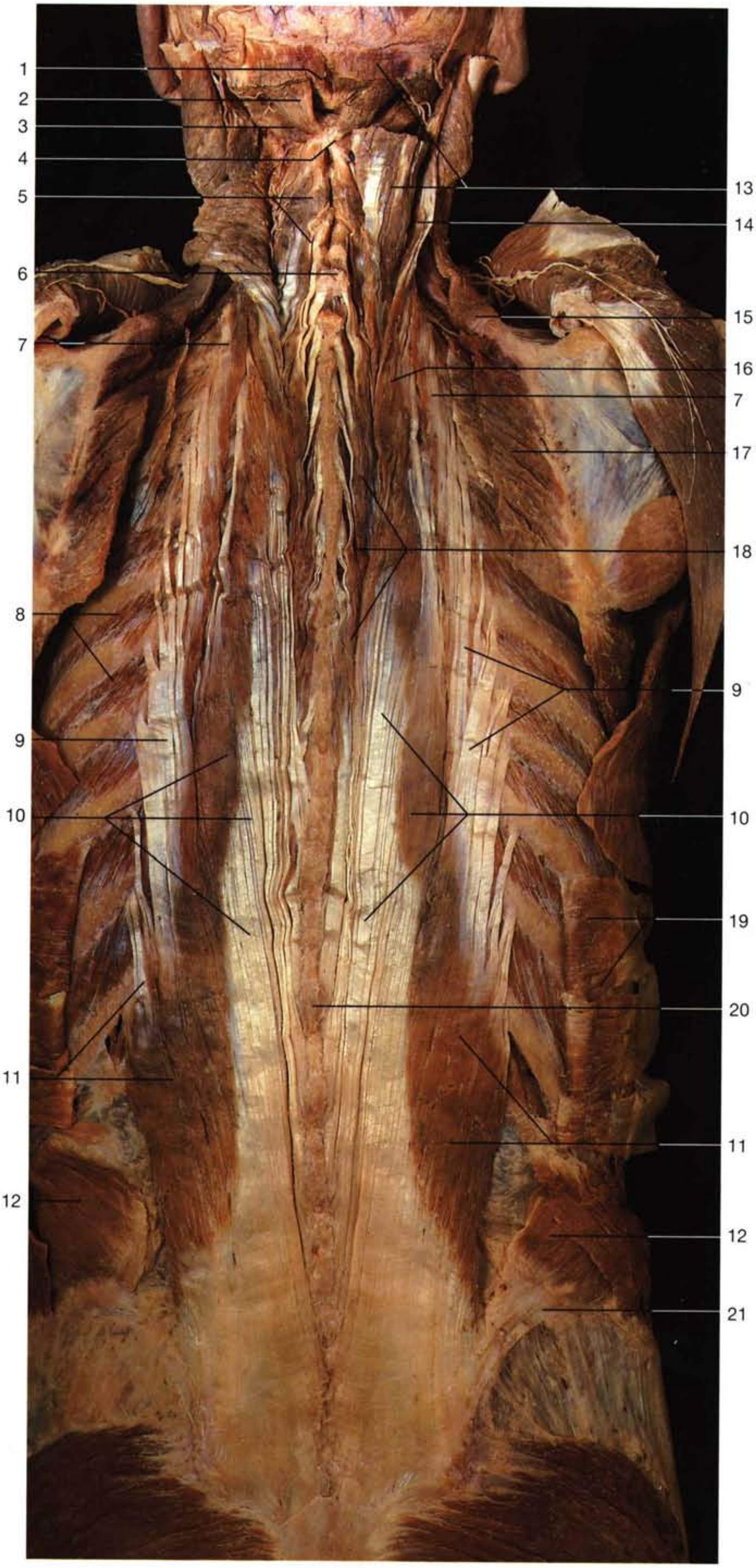

Muscles of the back II. Dissection of deep muscles. Erector spinae muscle.

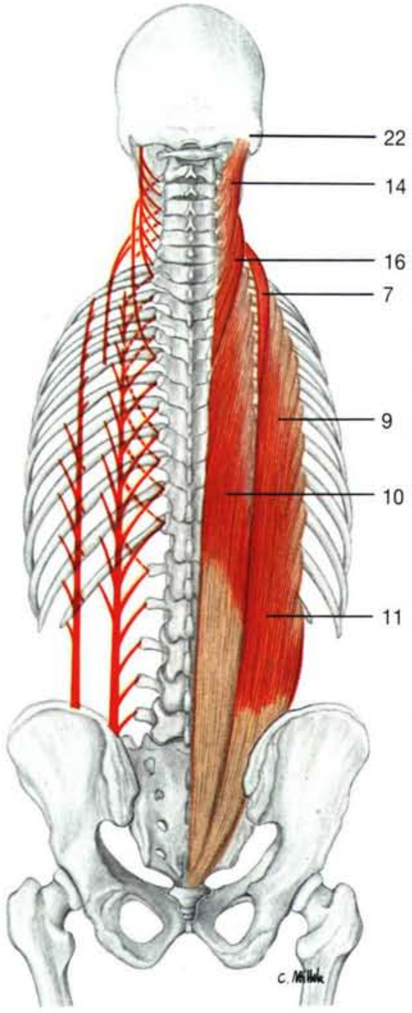

Origin and insertion of iliocostalis and longissimus muscles. (Schematic drawing.)

1 Rectus capitis posterior minor muscle
2 Rectus capitis posterior major muscle
3 Obliquus capitis inferior muscle
4 Spinous process of axis
5 Semispinalis cervicis muscle
6 Spinous process of seventh vertebra
7 **Iliocostalis cervicis muscle**
8 External intercostal muscles
9 **Iliocostalis thoracis muscle**
10 **Longissimus thoracis muscle**
11 **Iliocostalis lumborum muscle**
12 Internal abdominal oblique muscle
13 Semispinalis capitis muscle (divided)
14 **Longissimus capitis muscle**
15 Levator scapulae muscle
16 **Longissimus cervicis muscle**
17 Rhomboid major muscle
18 **Spinalis thoracis muscle**
19 Serratus posterior inferior muscle (reflected)
20 Spinous process of second lumbar vertebra
21 Iliac crest
22 Mastoid process

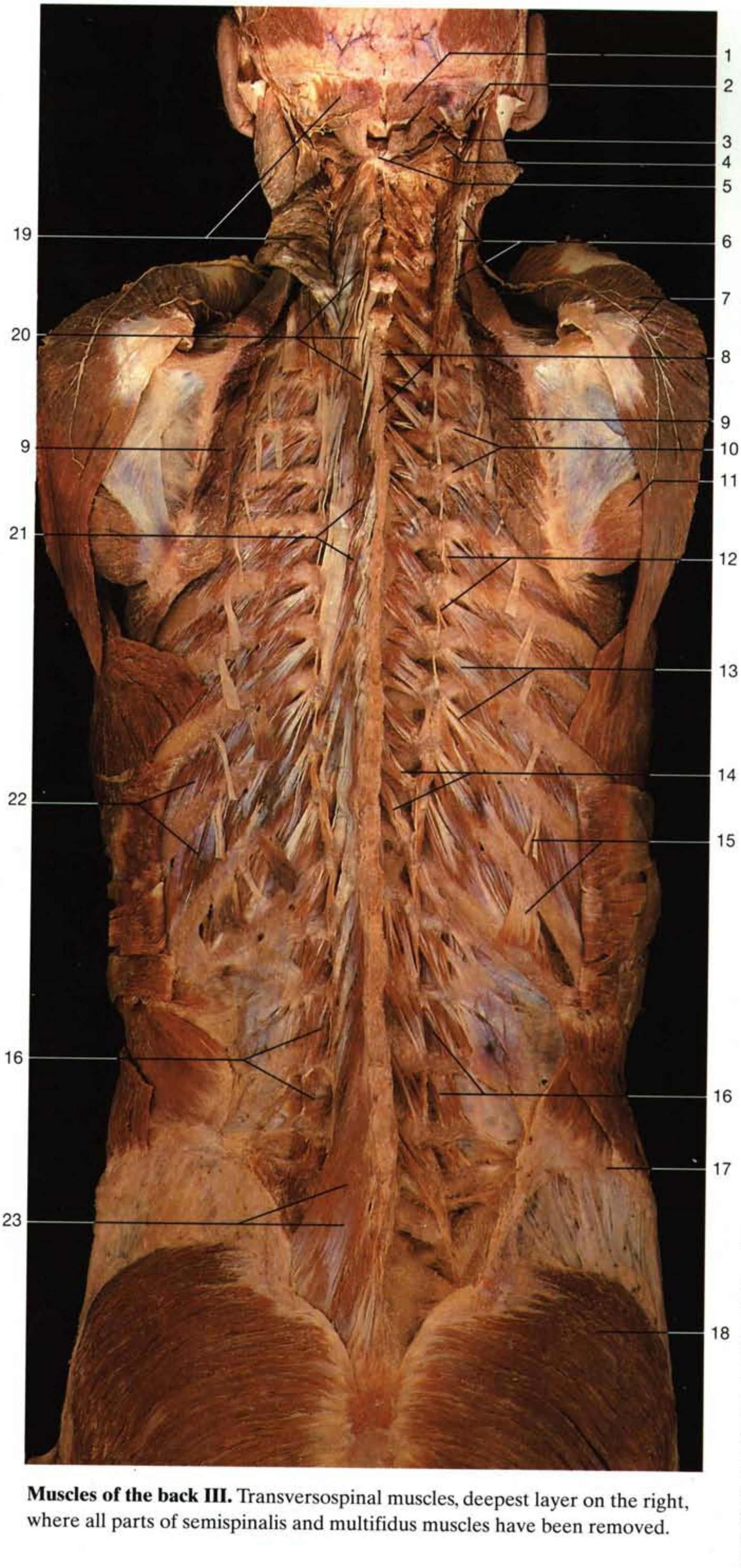

Muscles of the back III. Transversospinal muscles, deepest layer on the right, where all parts of semispinalis and multifidus muscles have been removed.

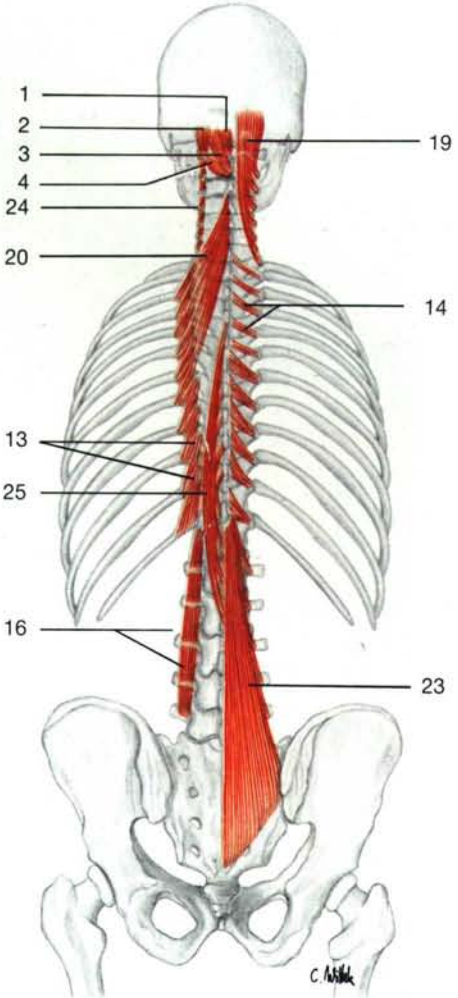

Medial column of intrinsic muscles of the back. Transversospinal and intertransversal system. (Schematic drawing.)

1 Rectus capitis posterior minor muscle
2 Obliquus capitis superior muscle
3 Rectus capitis posterior major muscle
4 Obliquus capitis inferior muscle
5 Spinous process of axis
6 **Longissimus capitis muscle**
7 Trapezius muscle (reflected) and accessory nerve (n. XI)
8 Spinous processes
9 Rhomboid major muscle
10 Transverse processes of thoracic vertebrae
11 Teres major muscle
12 Intertransverse ligaments
13 **Levatores costarum muscles**
14 **Rotatores muscles**
15 Tendons of iliocostalis muscle
16 Intertransversarii lumborum muscle (lateral)
17 Iliac crest
18 Gluteus maximus muscle
19 Semispinalis capitis muscle
20 **Semispinalis cervicis muscle**
21 Semispinalis thoracis muscle
22 **External intercostal muscles**
23 **Multifidus muscle**
24 Intertransversarii posterior cervicis muscle
25 Spinalis thoracis muscle

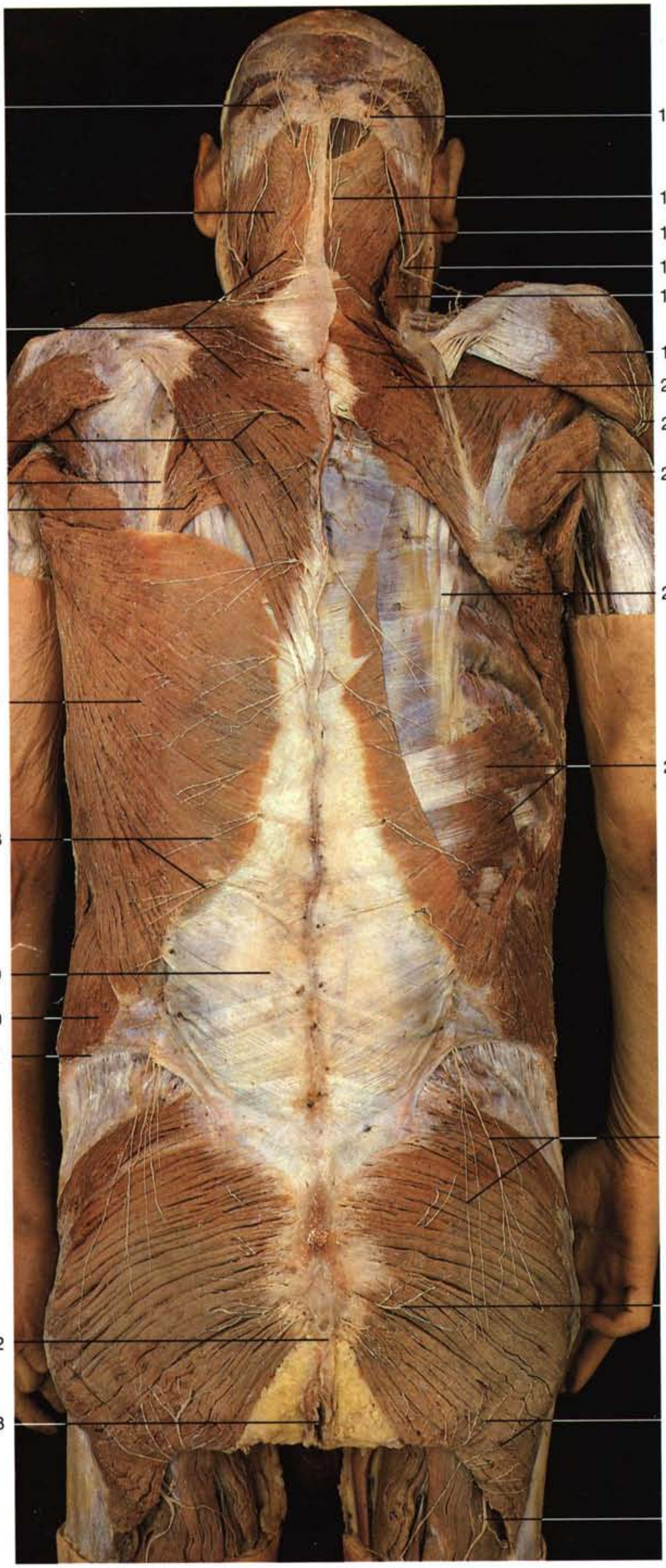

Innervation of the back I. Superficial (left) and deeper (right) layers. Right trapezius and latissimus dorsi muscles removed.

1 Occipital belly of occipitofrontalis muscle
2 Splenius capitis muscle
3 **Trapezius muscle**
4 **Medial cutaneous branches of dorsal rami of spinal nerves**
5 Medial margin of scapula
6 Rhomboid major muscle
7 **Latissimus dorsi muscle**
8 **Lateral cutaneous branches of dorsal rami of spinal nerves**
9 Thoracolumbar fascia
10 External abdominal oblique muscle
11 Iliac crest
12 Last coccygeal vertebra
13 Anus
14 Greater occipital nerve
15 Third occipital nerve
16 Lesser occipital nerve
17 Cutaneous branches of cervical plexus
18 Levator scapulae muscle
19 Deltoid muscle
20 **Rhomboid major and minor muscles**
21 Upper lateral cutaneous nerve of arm (branch of axillary nerve)
22 **Teres major muscle**
23 Iliocostalis thoracis muscle
24 Serratus posterior inferior muscle
25 Superior cluneal nerves
26 Middle cluneal nerves
27 Inferior cluneal nerves
28 Posterior femoral cutaneous nerve

▷

To page 213:

1 **Trapezius muscle**
2 Infraspinatus muscle
3 Left **latissimus dorsi muscle**
4 Thoracolumbar fascia
5 Splenius cervicis muscle
6 Serratus posterior superior muscle
7 **Medial branches of dorsal rami of thoracic spinal nerves**
8 **Lateral branches of dorsal rami of thoracic spinal nerves**
9 Iliocostalis muscle
10 Serratus posterior inferior muscle
11 Latissimus dorsi muscle (reflected)

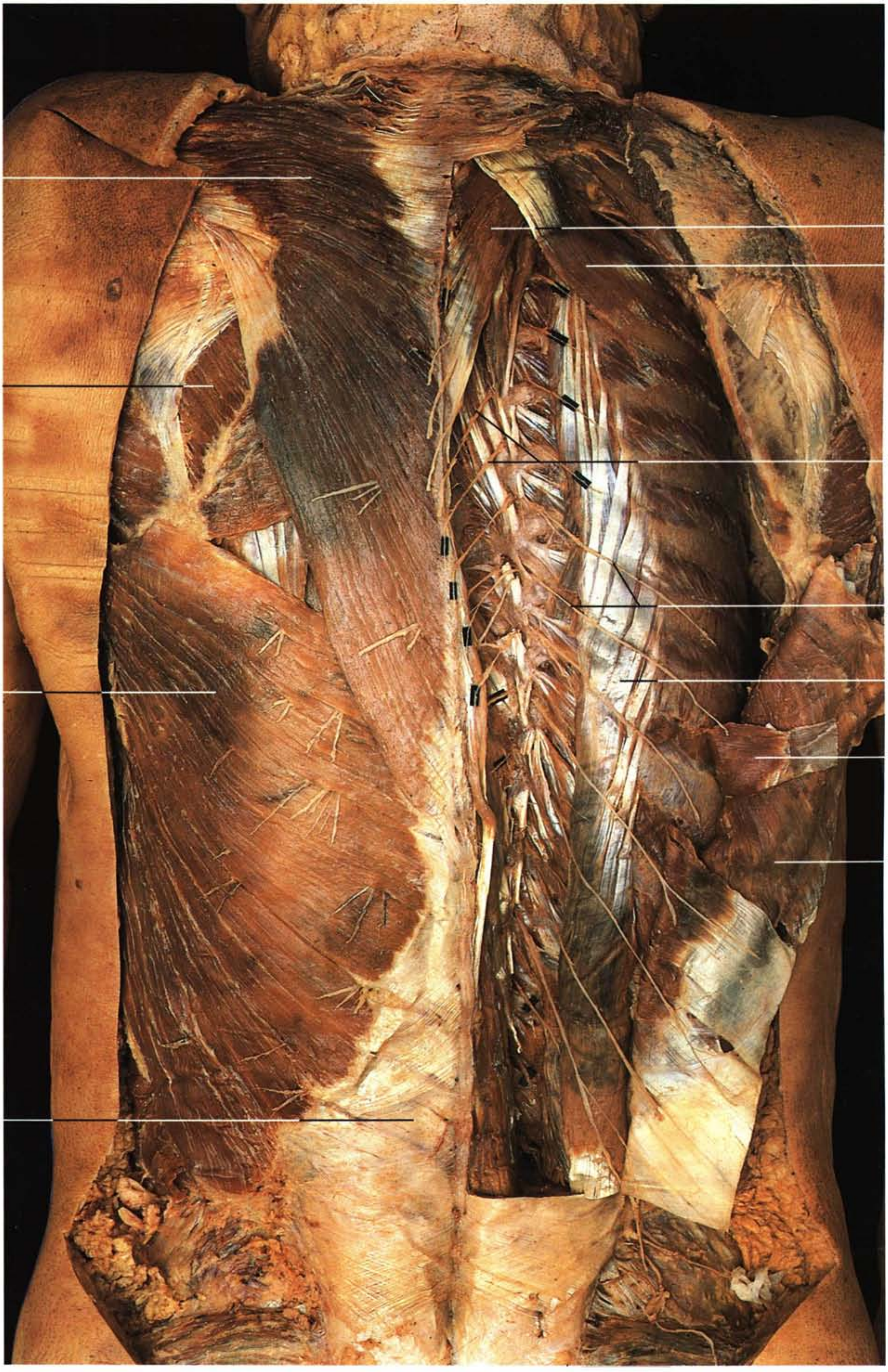

Innervation of the back II. Dissection of the dorsal branches of spinal nerves. On the right, longissimus thoracis muscle has been removed and iliocostalis muscle laterally reflected.

1 Greater occipital nerve (C_2)
2 Suboccipital nerve (C_1)
3 **Medial branches of dorsal rami of spinal nerves**
4 **Lateral branches of dorsal rami of spinal nerves**
5 Superior cluneal nerves (L_1–L_3)
6 Middle cluneal nerves (S_1–S_3)
7 Inferior cluneal nerves (derived from branches of the sacral plexus, ventral rami)
8 Lesser occipital nerve
9 Great auricular nerve
10 Trapezius muscle
11 Deltoid muscle
12 Latissimus dorsi muscle
13 Gluteus maximus muscle
14 External intercostal muscle
15 Internal intercostal muscle
16 Innermost intercostal muscle
17 Dorsal ramus of spinal nerve
18 **Spinal nerve** and **spinal ganglion**
19 Sympathetic trunk with ganglion
20 Intercostal nerve
21 Lateral cutaneous branch } of **intercostal nerve**
22 Anterior cutaneous branch } of **intercostal nerve**

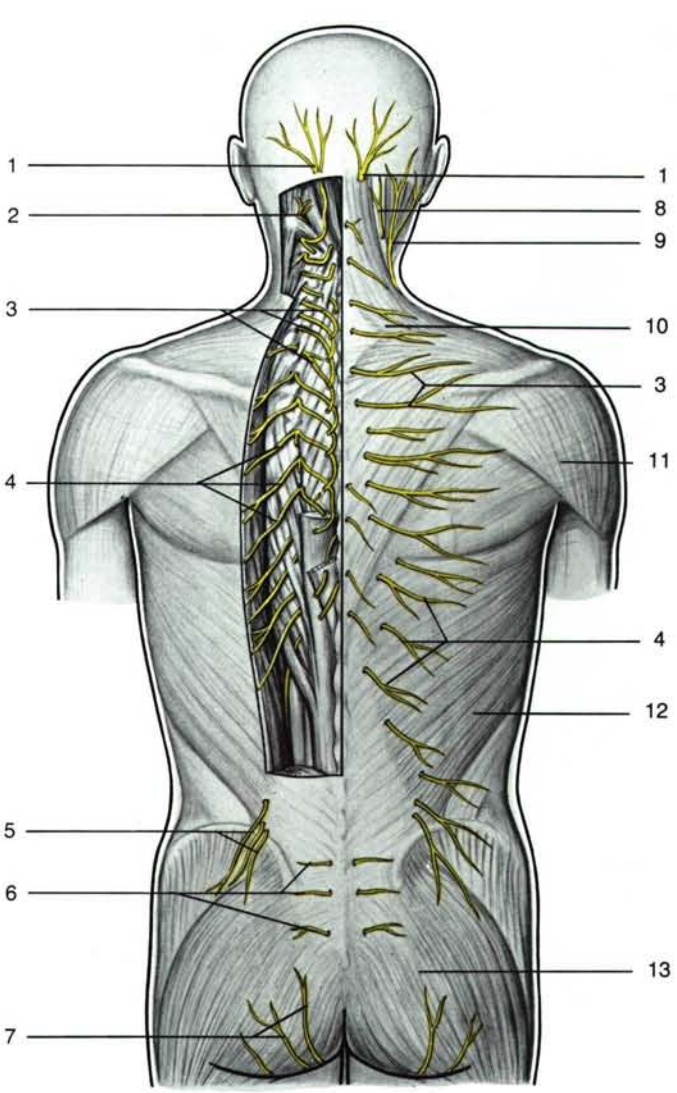

General characteristics of the innervation of the back. Distribution of dorsal branches of spinal nerves. Note the segmental arrangement of the innervation of the dorsal part of the trunk. (Schematic drawing.)

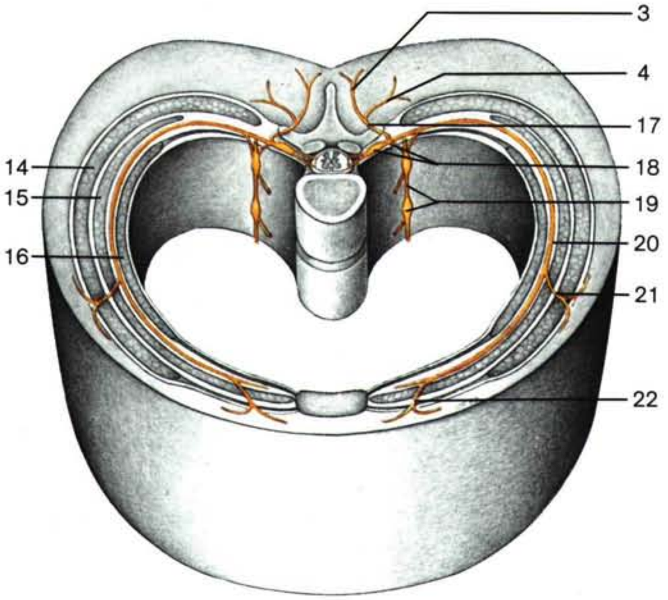

Position and branches of spinal nerves in one segment of thoracic wall. (Schematic drawing.)

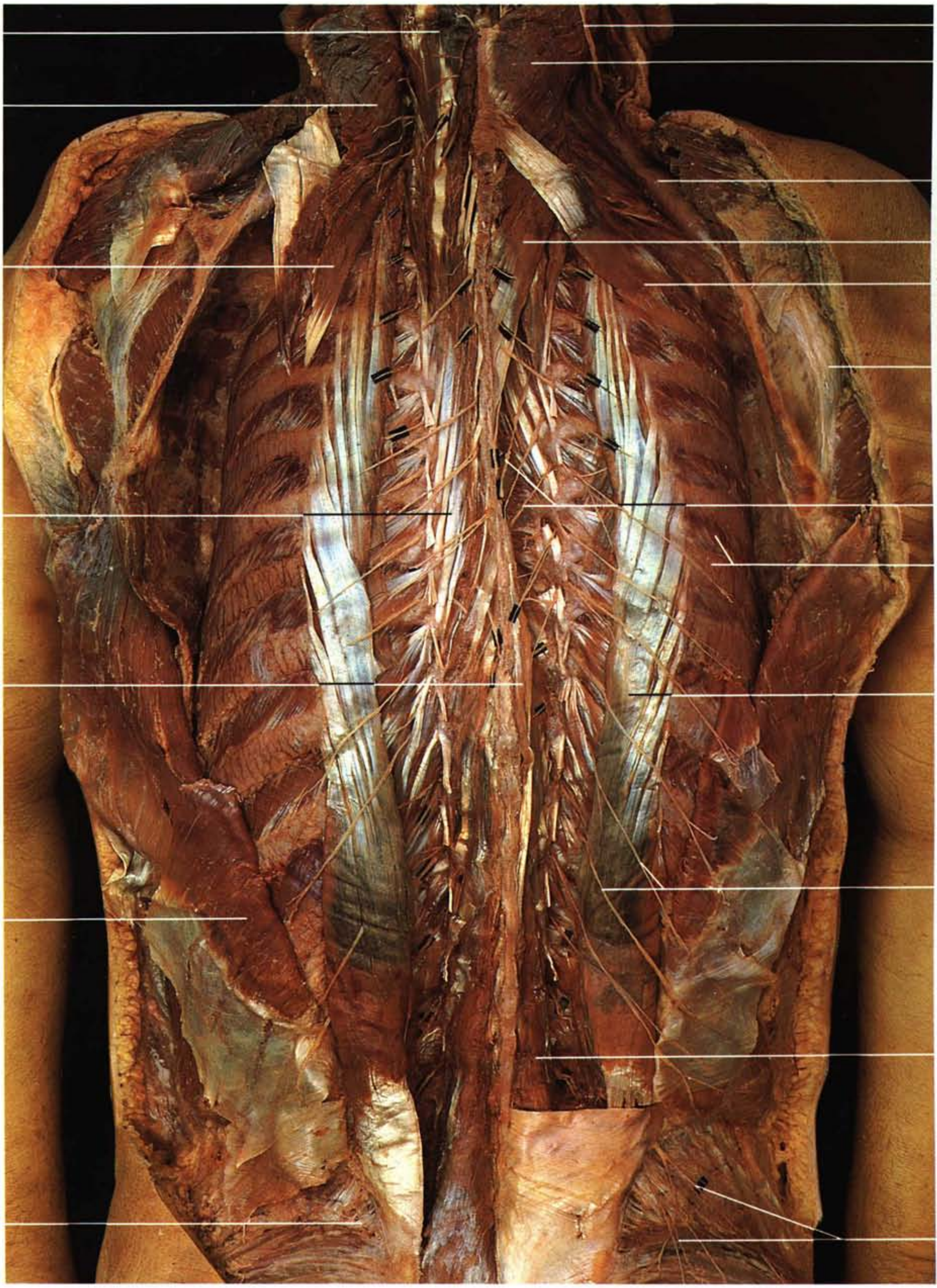

Innervation of the back III. Deeper layer (dorsal aspect).

1 Semispinalis capitis muscle
2 Left splenius capitis muscle (cut and reflected)
3 Left splenius cervicis muscle (cut and reflected)
4 Semispinalis thoracis muscle
5 **Spinalis thoracis muscle**
6 Latissimus dorsi muscle (reflected)
7 Iliac crest
8 Lesser occipital nerve
9 Splenius capitis muscle
10 Levator scapulae muscle
11 **Splenius cervicis muscle**
12 Serratus posterior superior muscle
13 Scapula
14 **Medial branches of dorsal rami of spinal nerves**
15 Rib and external intercostal muscle
16 Iliocostalis thoracis muscle
17 **Lateral branches of dorsal rami of spinal nerves**
18 Multifidus muscle
19 Superior cluneal nerves

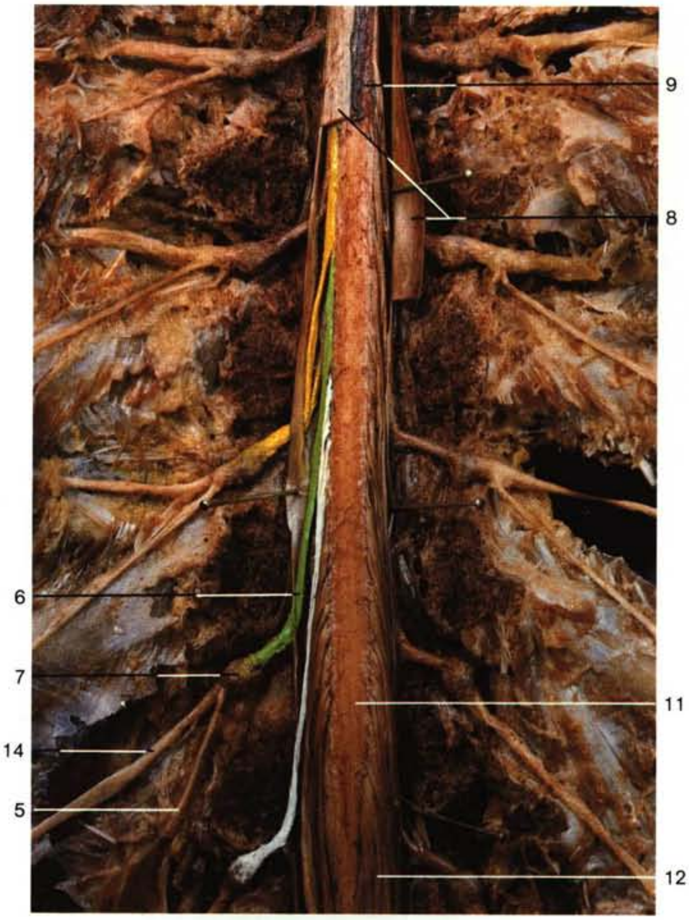

Lumbar portion of spinal cord. Note the relation between the nervous and muscular segments.

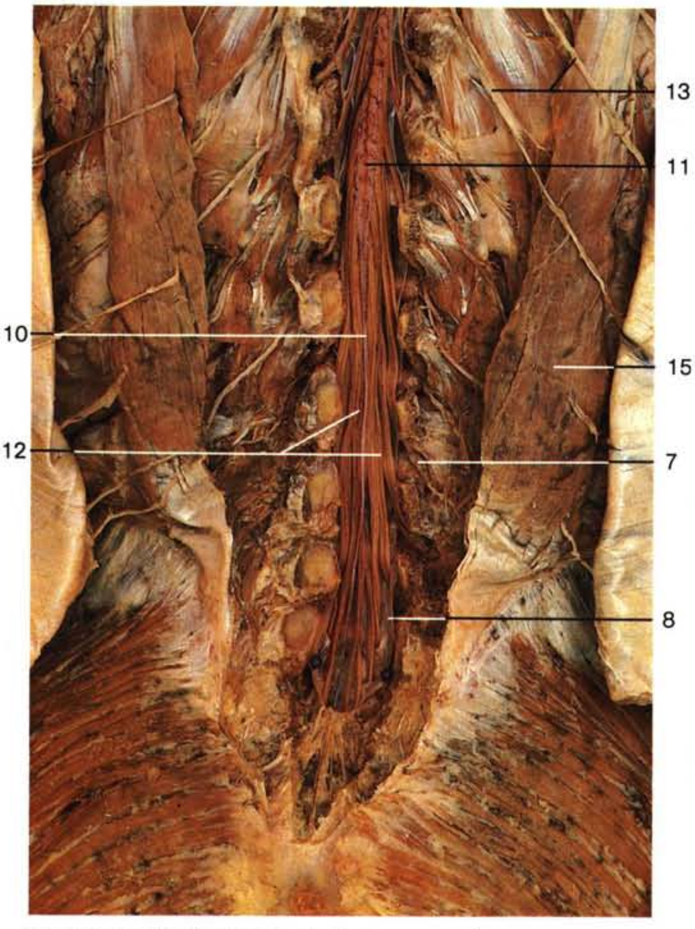

Terminal part of spinal cord. Dura removed.

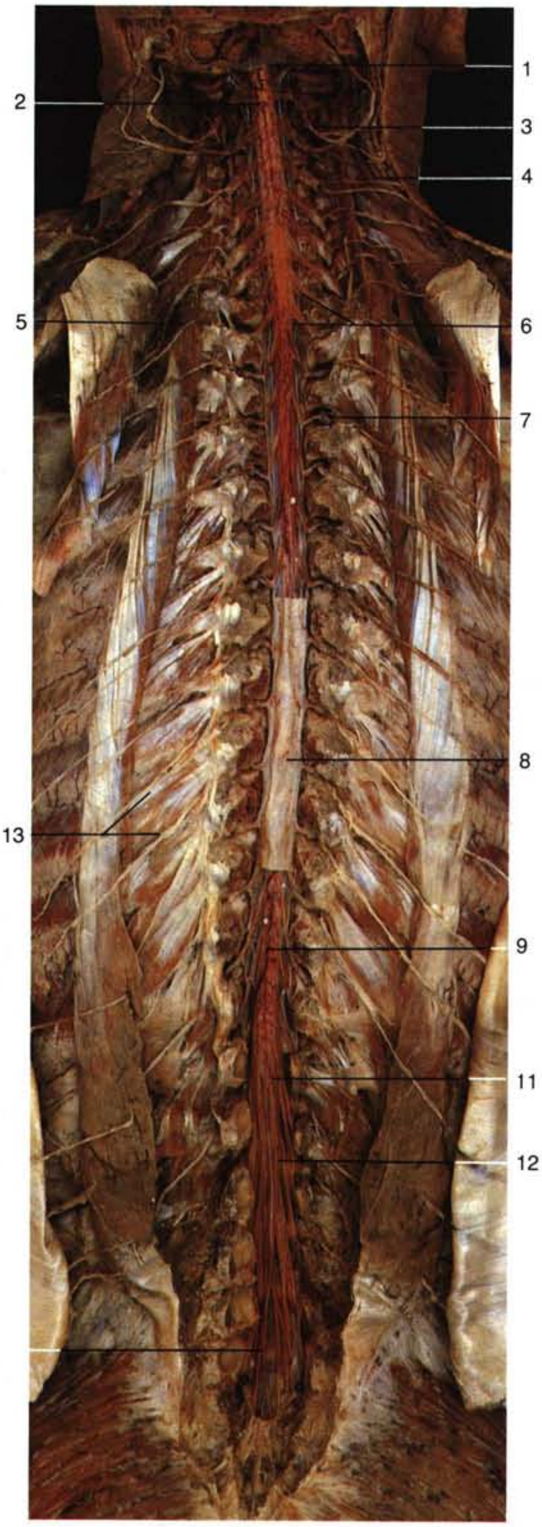

Innervation of the back IV. Spinal cord in the vertebral canal (opened). Longissimus dorsi muscle has been removed and iliocostal muscle reflected.

1 Cerebellomedullary cistern and cerebellum
2 Medulla oblongata
3 Greater occipital nerve (C_2)
4 Third cervical nerve (C_3)
5 Dorsal primary ramus
6 Dorsal roots
7 Spinal ganglion
8 Spinal dura mater
9 Spinal arachnoid mater
10 Filum terminale
11 Conus medullaris
12 Cauda equina
13 Lateral branches of dorsal rami of spinal nerves
14 Ventral ramus of spinal nerve (intercostal nerve)
15 Iliocostalis muscle

1 Arch of vertebra (divided)
2 Spinal nerve with meningeal coverings
3 Dorsal roots of thoracic spinal nerves
4 Spinal cord (thoracic portion)
5 Spinal ganglia with meningeal coverings
6 Pia mater with blood vessels
7 Dura mater (opened)
8 Denticulate ligament
9 Lateral branch of dorsal ramus
10 Dorsal ramus of spinal nerve (dividing into a medial and lateral branch)
11 Medial branch of dorsal ramus of spinal nerve
12 Spinal dura mater
13 Spinal nerves of sacral segments
14 Filum terminale

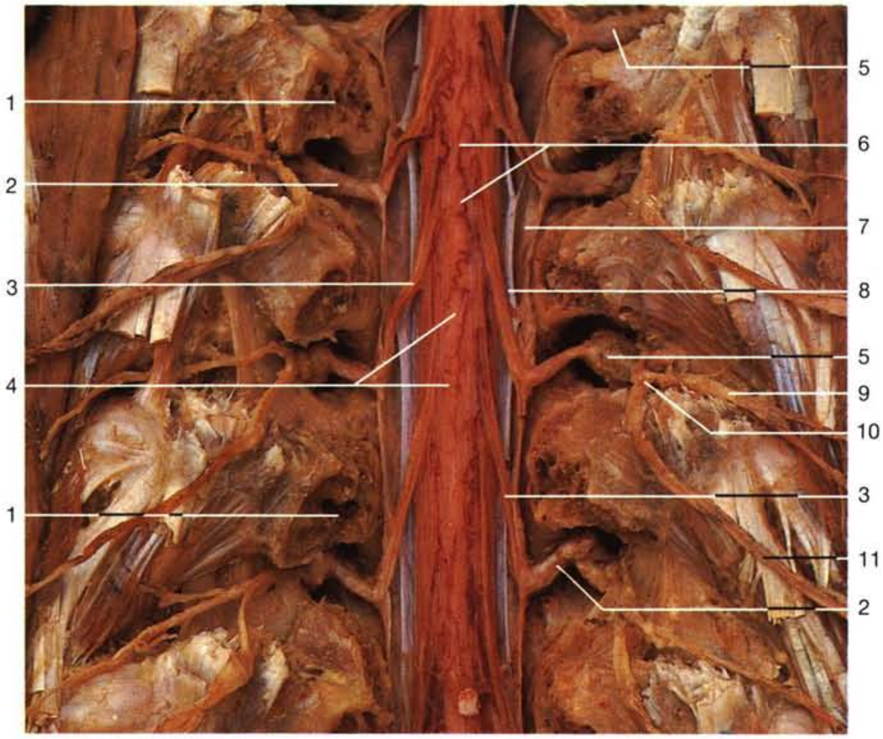

Thoracic portion of spinal cord (dorsal aspect). Vertebral canal and dura mater opened.

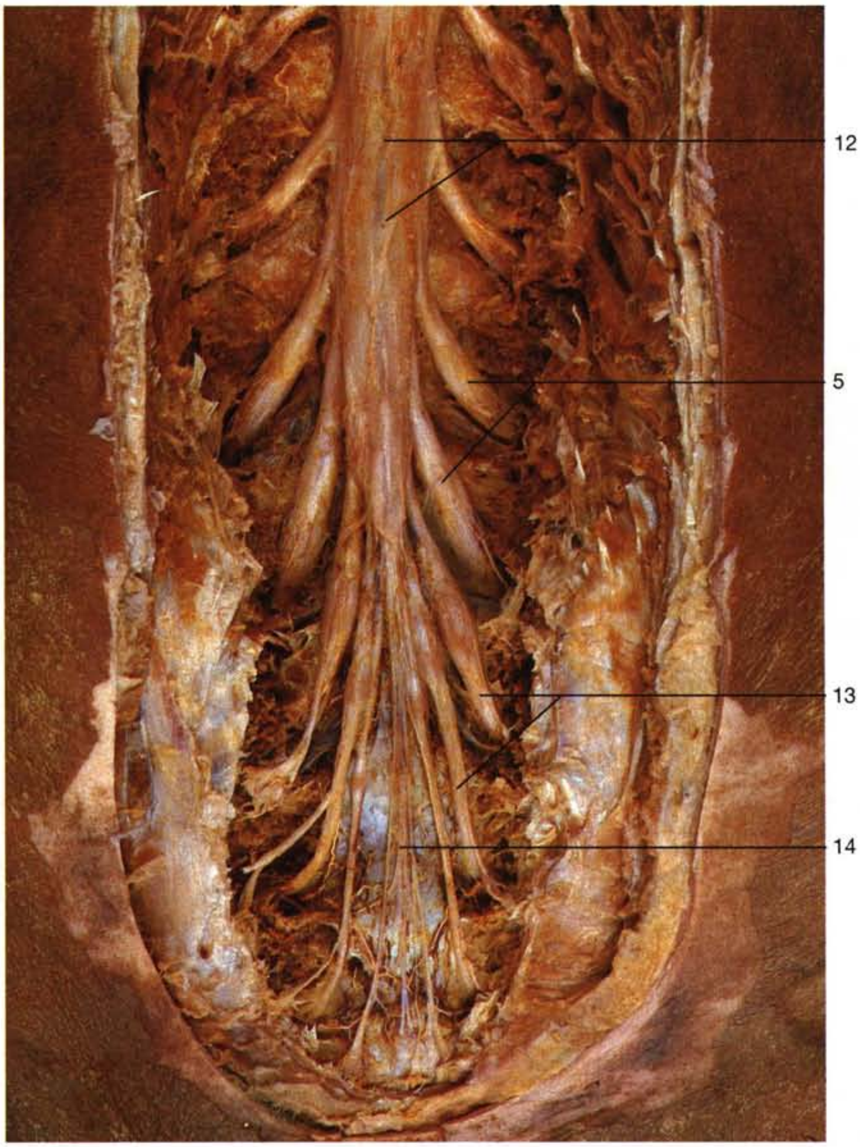

Terminal part of spinal cord with dura mater (dorsal aspect). Dorsal part of sacrum removed.

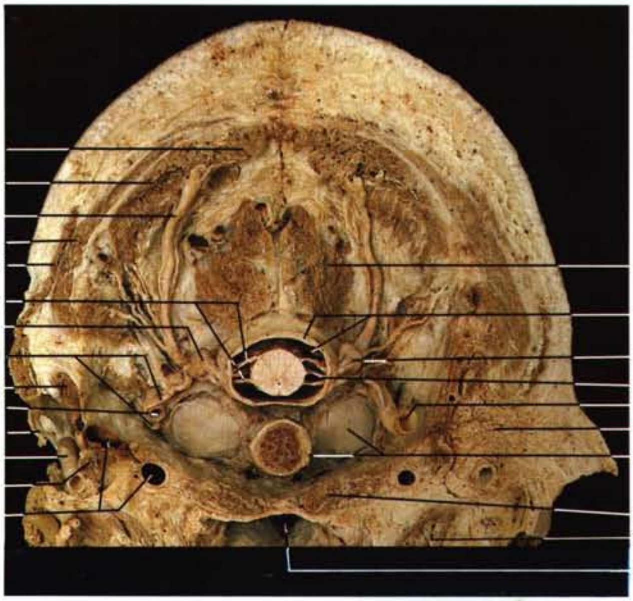

Horizontal section of the neck. Dissection of the second cervical spinal nerve. Posterior surface at top of figure.

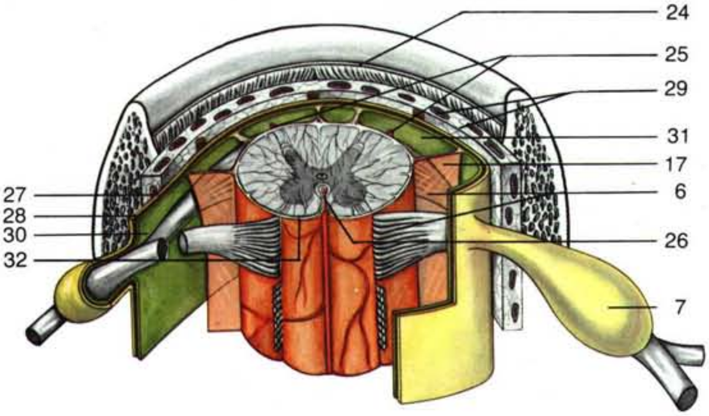

Meningeal coverings of the spinal cord (anterior aspect). (Schematic drawing.)

1 Trapezius muscle
2 Semispinalis capitis muscle
3 **Dorsal ramus of spinal nerve**
4 Sternocleidomastoid muscle
5 Platysma muscle
6 **Dorsal and ventral roots of spinal nerves**
7 **Spinal ganglion**
8 Posterior belly of digastric muscle
9 **Ventral ramus of spinal nerve**
10 Vertebral artery
11 Great auricular nerve
12 Superficial temporal artery
13 Styloid process
14 Internal jugular vein and internal carotid artery
15 Rectus capitis posterior major muscle
16 Dura mater and subarachnoid space
17 Denticulate ligament
18 Vertebral artery
19 Parotid gland
20 Dens of axis (divided) and inferior articular facet of atlas
21 Longus capitis muscle
22 Pharyngeal cavity
23 Medial pterygoid muscle
24 Periosteum of vertebral canal
25 Posterior spinal arteries
26 Anterior spinal artery

Meningeal coverings
27 Dura mater
28 Subdural space
29 Extradural or epidural space with venous plexus and fatty tissue
30 Arachnoid (green)
31 Subarachnoid space
32 Pia mater (pink)
33 Nucleus pulposus
34 Crus of diaphragm
35 Intervertebral disc
36 Body of first lumbar vertebra
37 Spinal cord
38 Conus medullaris
39 Cauda equina
40 Filum terminale
41 Spinous process

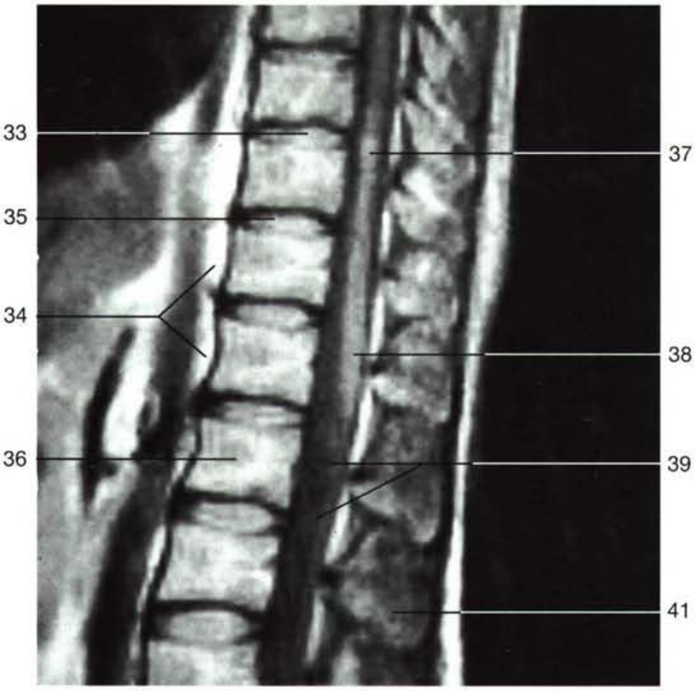

Sagittal section through the vertebral canal, T_9–L_2. (MR-Scan.)

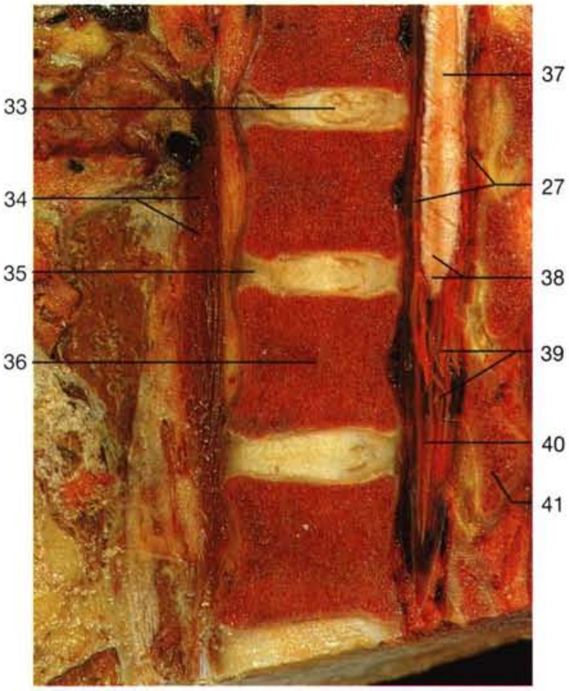

Sagittal section through vertebral canal, T_{12}–L_2. Notice red bone marrow (unfixed).

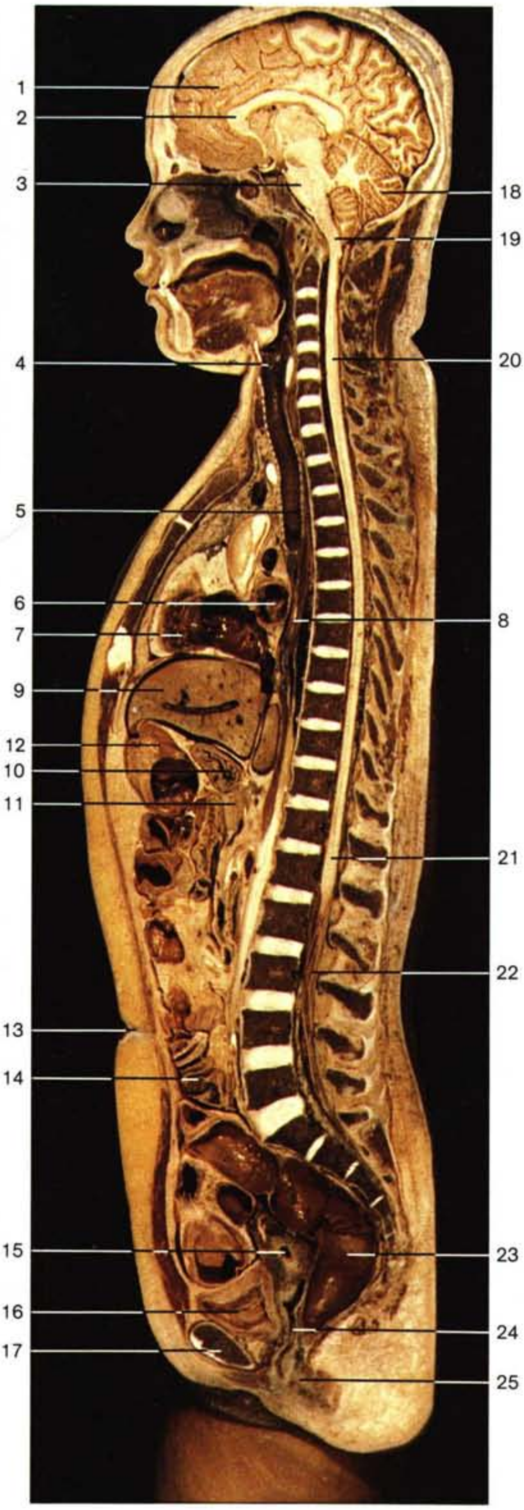
1
2
3
18
19
4
20
5
6
8
7
9
12
10
11
21
22
13
14
15
23
16
24
17
25

Median section of the head and trunk in the adult (female). The conus medullaris of the spinal cord is located at the level of L_1.

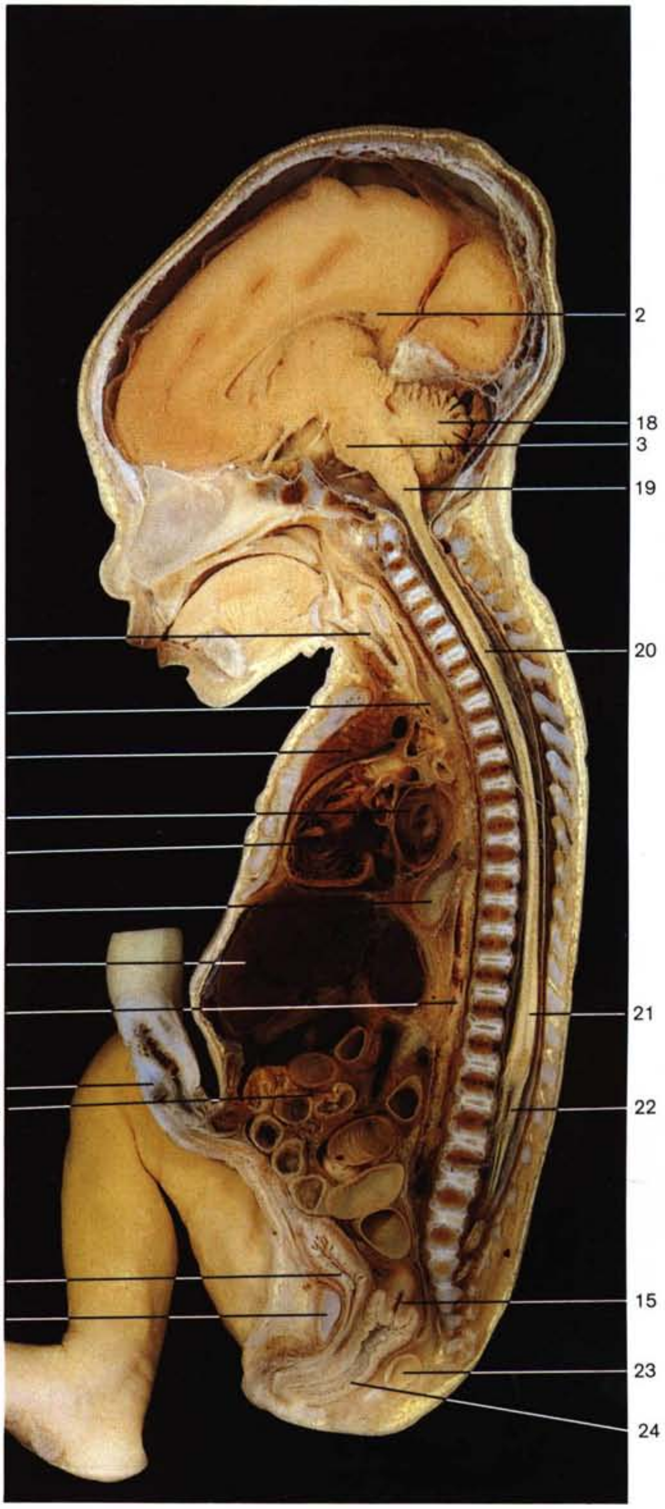
2
18
3
19
4
20
8
29
6
7
26
9
27
21
28
14
22
16
15
17
23
24

Median section of the head and trunk in the neonate. Note that in the neonate the conus medullaris of the spinal cord extends far more caudally than in the adult.

1 Cerebrum
2 Corpus callosum
3 Pons
4 Larynx
5 Trachea
6 Left atrium
7 **Right ventricle**
8 Esophagus
9 **Liver**
10 Stomach
11 Pancreas
12 Transverse colon
13 Umbilicus
14 Small intestine
15 Uterus
16 Urinary bladder
17 Pubic symphysis
18 Cerebellum
19 Medulla oblongata
20 **Spinal cord**
21 **Conus medullaris**
22 **Cauda equina**
23 Rectum
24 Vagina
25 Anus
26 Inferior vena cava
27 Aorta
28 Umbilical cord
29 Thymus

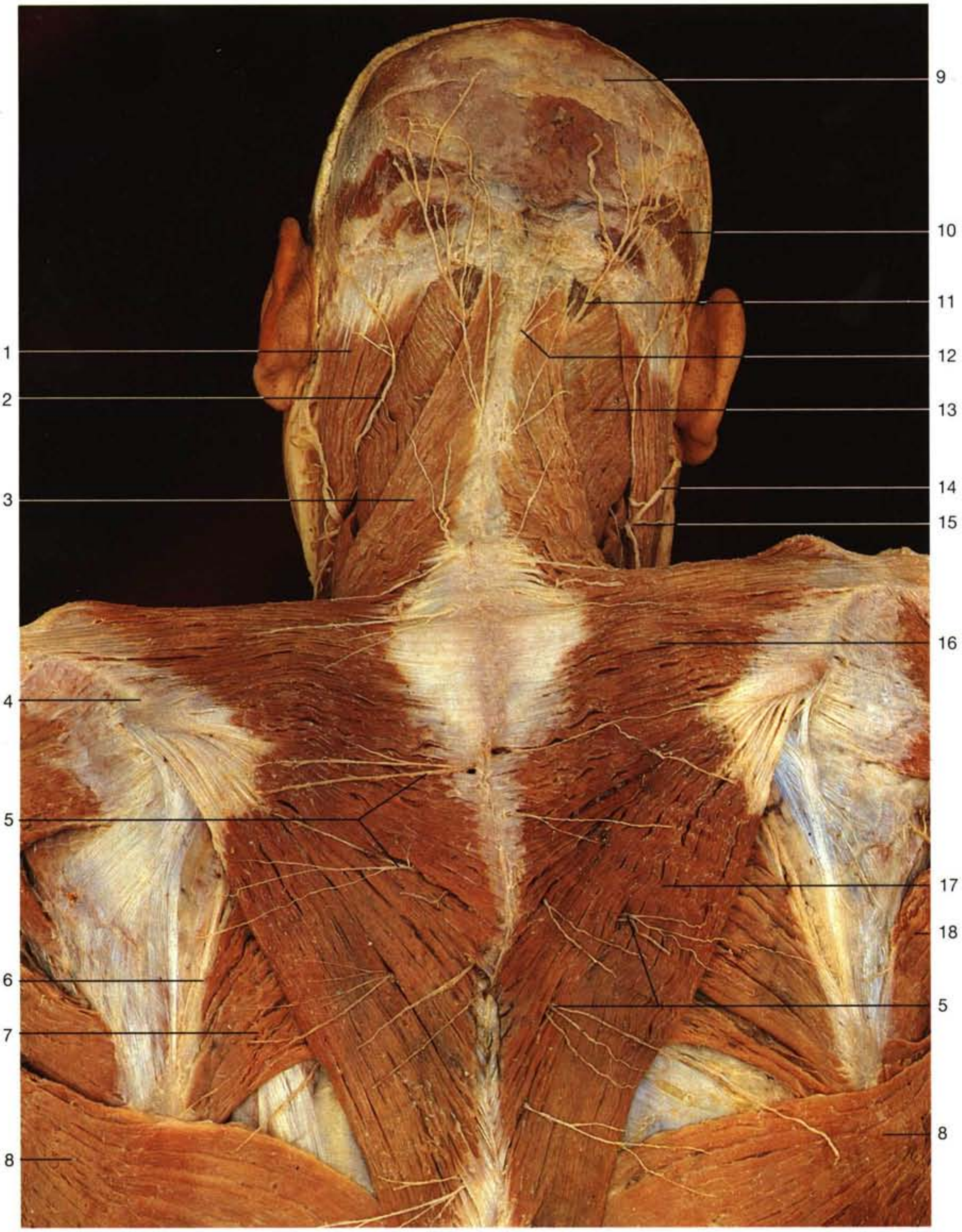

Dorsal aspect of the neck I. Superficial layer. Nuchal region and shoulder.

1	Sternocleidomastoid muscle	10	Occipital belly of occipitofrontalis muscle
2	**Lesser occipital nerve**	11	**Greater occipital nerve**
3	Descending fibers of trapezius muscle	12	Third occipital nerve
4	Spine of scapula	13	Splenius capitis muscle
5	**Medial cutaneous branches of dorsal rami of spinal nerves**	14	**Great auricular nerve**
6	Medial margin of scapula	15	Cutaneous nerves of cervical plexus
7	Rhomboid major muscle	16	Transverse fibers of trapezius muscle
8	Latissimus dorsi muscle	17	Ascending fibers of trapezius muscle
9	Galea aponeurotica	18	Teres major muscle

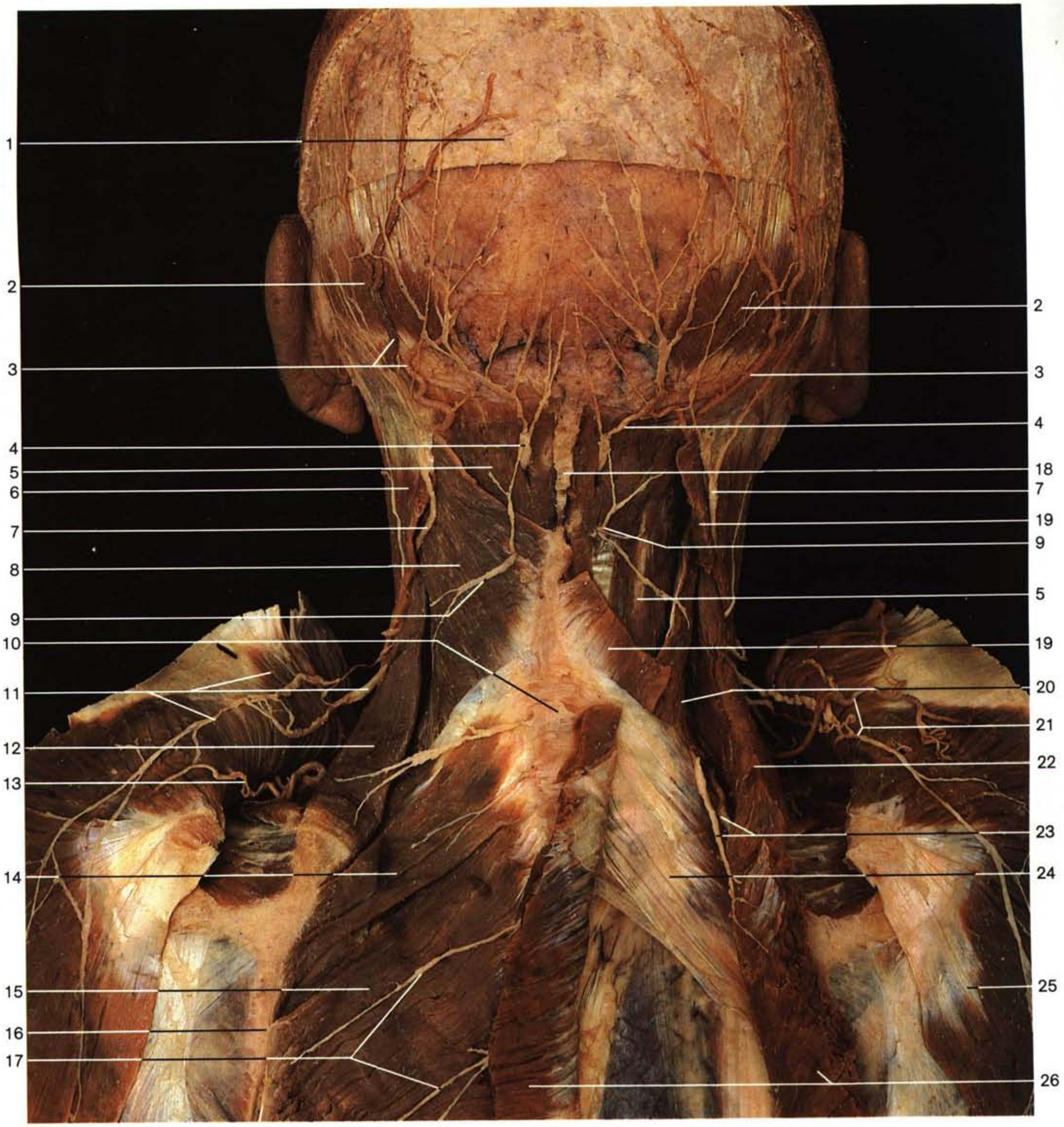

Dorsal aspect of neck II. Deeper layer. The left trapezius muscle has been divided and reflected. On the right, trapezius, rhomboid and splenius muscles have been divided. Right levator scapulae muscle has been slightly reflected.

1 Galea aponeurotica
2 Occipital belly of occipitofrontalis muscle
3 Occipital artery
4 **Greater occipital nerve** (C_2)
5 Semispinalis capitis muscle
6 Sternocleidomastoid muscle
7 Lesser occipital nerve
8 Left splenius capitis muscle
9 Third occipital nerve (C_3)
10 Spinous process of vertebra prominens (C_7)
11 Left trapezius muscle and **accessory nerve**
12 Levator scapulae muscle
13 Superficial branch of transverse cervical artery
14 Thomboid minor muscle
15 Rhomboid major muscle
16 Medial margin of scapula
17 **Medial branches of dorsal rami of spinal nerves**
18 Ligamentum nuchae
19 Splenius capitis muscle (divided)
20 Splenius cervicis muscle
21 Right **accessory nerve** and superficial branch of transverse cervical artery
22 Right levator scapulae muscle
23 Dorsal scapular nerve and deep branch of transverse cervical artery
24 Serratus posterior superior muscle
25 Right trapezius muscle (divided and reflected)
26 Right rhomboid major muscle (divided and reflected)

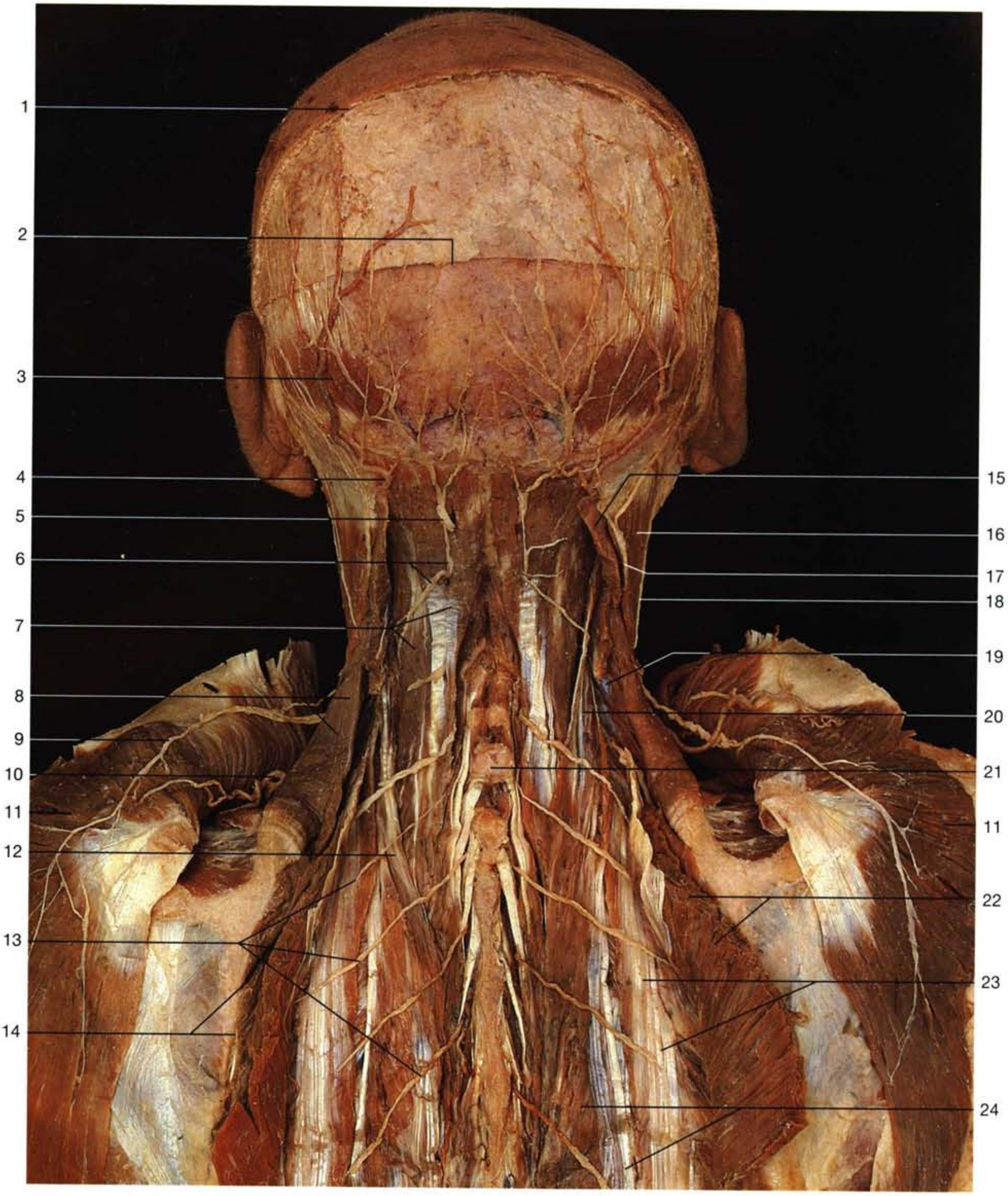

Dorsal aspect of neck III. Deepest layer. Nuchal region. Trapezius muscle and splenius capitis and cervicis muscles have been divided and partly removed or reflected.

1	Skin of scalp
2	Galea aponeurotica
3	Occipital belly of occipitofrontalis muscle
4	**Occipital artery**
5	**Greater occipital nerve**
6	Third occipital nerve
7	Semispinalis capitis muscle
8	Levator scapulae muscle
9	**Accessory nerve** (n. XI)
10	Superficial cervical artery
11	Trapezius muscle (reflected)
12	Longissimus cervicis muscle
13	**Medial cutaneous branches of dorsal rami of spinal nerves**
14	Medial margin of scapula
15	Splenius capitis muscle (divided)
16	Sternocleidomastoid muscle
17	**Lesser occipital nerve**
18	Great auricular nerve
19	Splenius cervicis muscle
20	Longissimus cervicis muscle
21	Spinous process of seventh cervical vertebra (vertebra prominens)
22	Rhomboid muscles (divided)
23	Iliocostalis thoracis muscle
24	Longissimus thoracis muscle

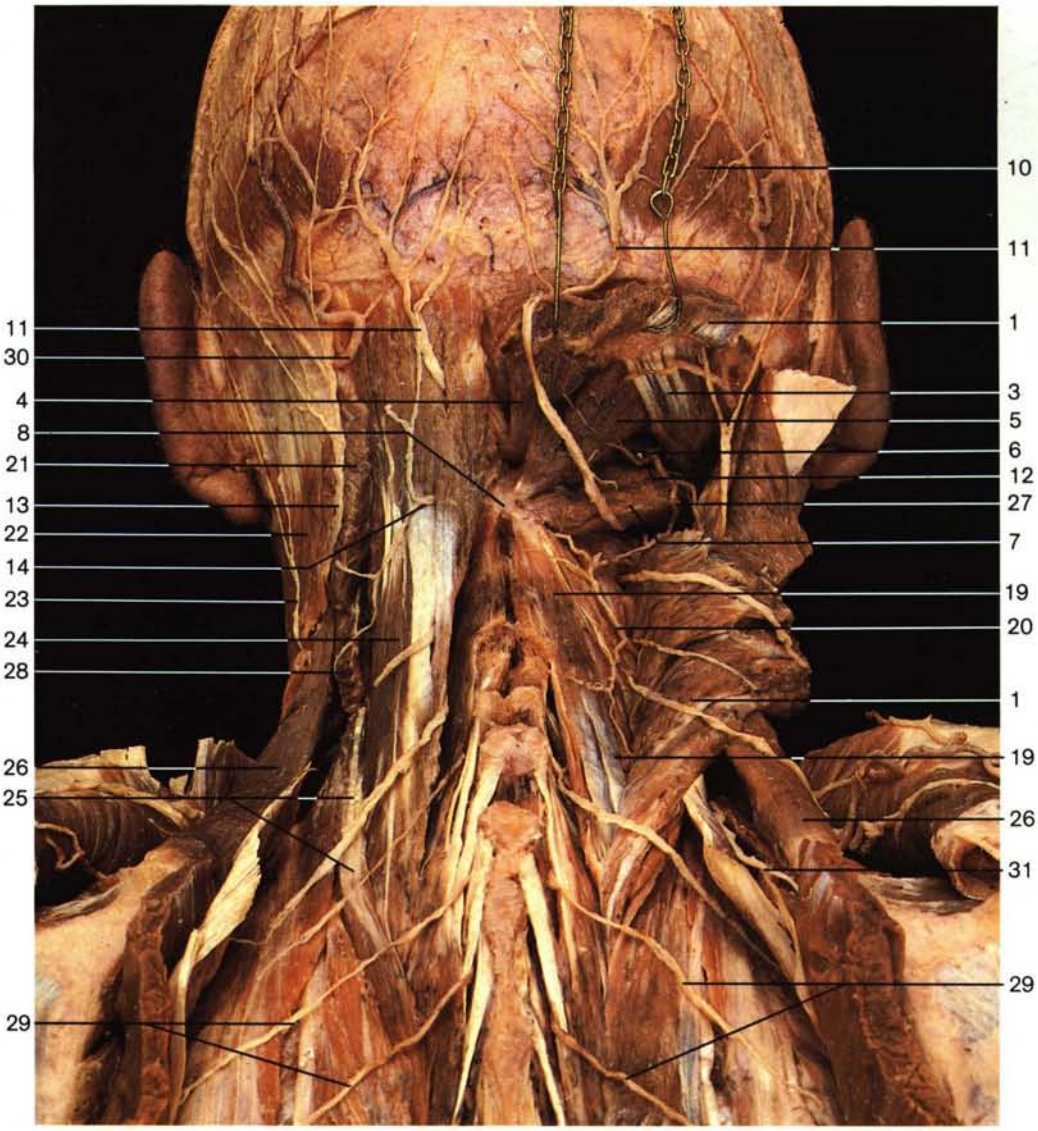

Dorsal aspect of neck IV. Deepest layer, suboccipital triangle; right semispinalis capitis muscle divited and reflected.

1 Semispinalis capitis muscle (divided)
2 External occipital protuberance
3 Obliquus capitis superior muscle
4 Rectus capitis posterior minor muscle
5 Rectus capitis posterior major muscle
6 Vertebral artery
7 Obliquus capitis inferior muscle
8 **Spinous process of axis**
9 Third cervical vertebra
10 Occipital belly of occipitofrontalis muscle
11 **Greater occipital nerve**
12 Suboccipital nerve (C_1)
13 Lesser occipital nerve
14 Third occipital nerve (C_3)
15 Mastoid process and splenius capitis muscle
16 Atlas
17 Axis
18 Spinous process of third cervical vertebra
19 Right semispinalis cervicis muscle
20 Deep cervical artery
21 Left splenius capitis muscle (divided)
22 Left sternocleidomastoid muscle
23 Great auricular nerve
24 Left **semispinalis capitis muscle**
25 Left longissimus cervicis muscle
26 Levator scapulae muscle
27 Muscular branch of vertebral artery
28 Left semispinalis cervicis muscle (divided)
29 **Medial branches of dorsal rami of spinal nerves**
30 **Occipital artery**
31 Dorsal scapular nerve

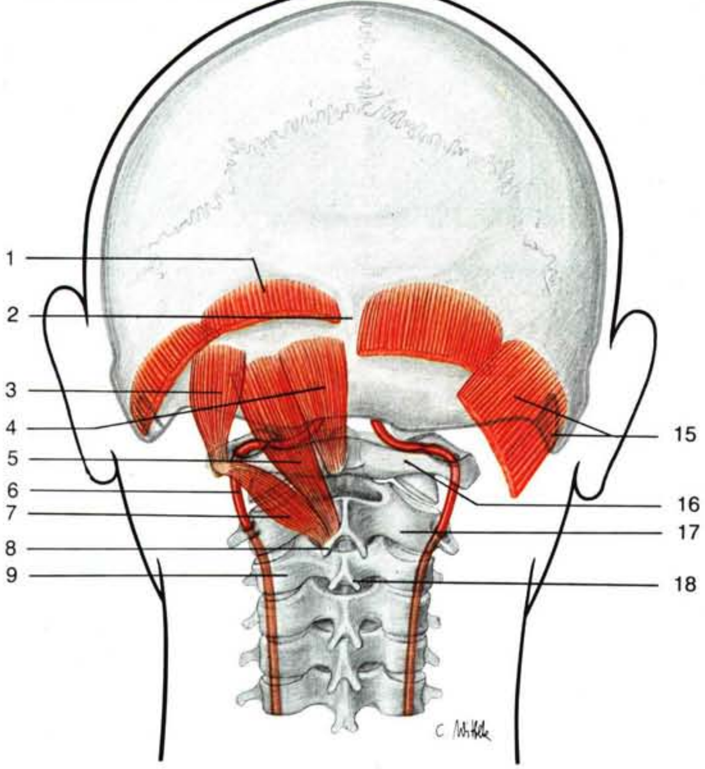

Suboccipital triangle and position of the vertebral artery. (Schematic drawing.)

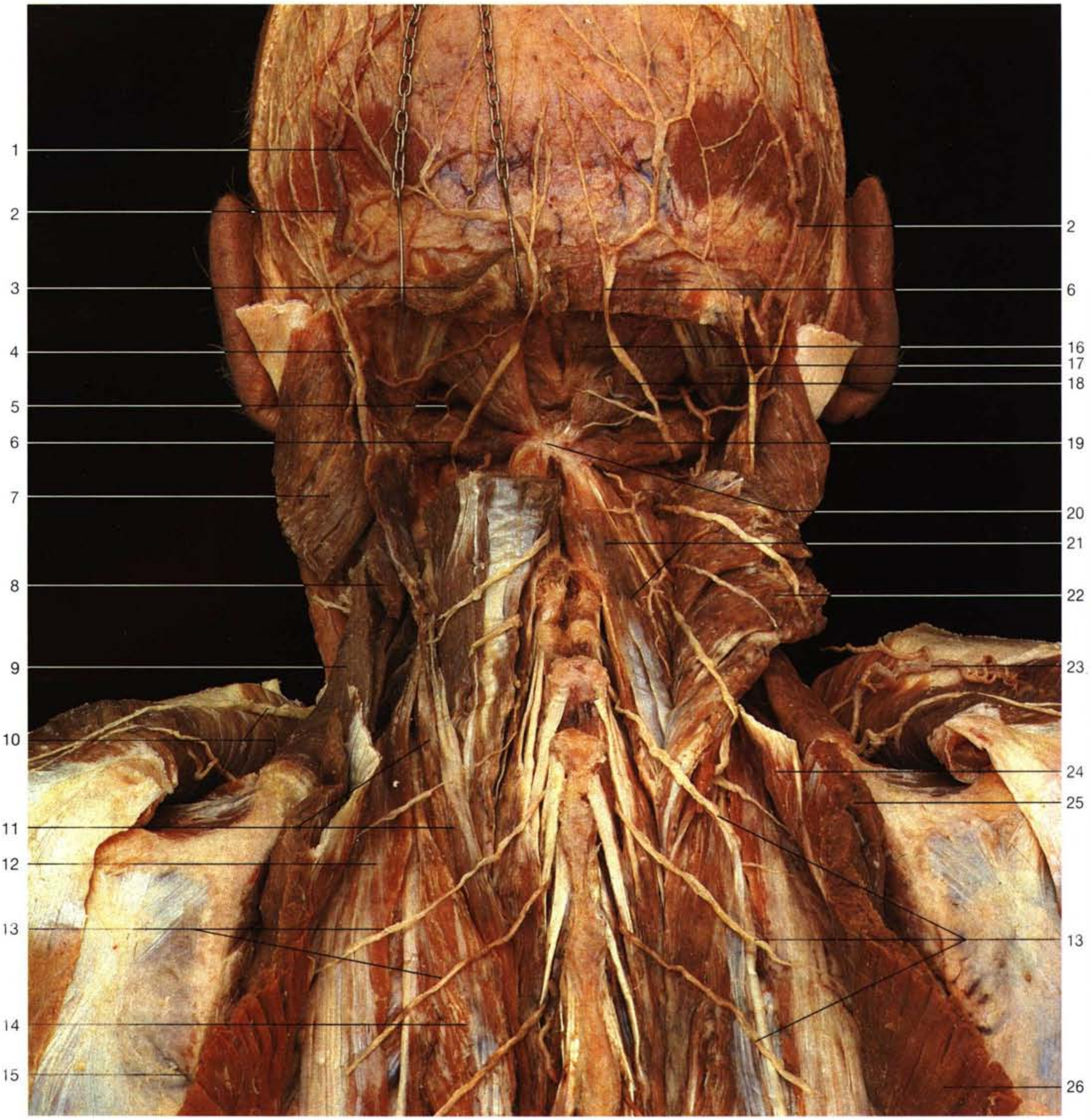

Dorsal aspect of neck V. Nuchal region. Deepest layer. Dissection of suboccipital triangle on both sides.

1 Occipital belly of occipitofrontalis muscle
2 **Occipital artery**
3 Insertion of semispinalis capitis muscle (divided)
4 Lesser occipital nerve (from cervical plexus)
5 Suboccipital nerve (C_1)
6 **Greater occipital nerve** (C_2)
7 Splenius capitis muscle (reflected)
8 Splenius cervicis muscle
9 Levator scapulae muscle
10 Accessory nerve (n. XI), trapezius muscle
11 Longissimus cervicis muscle
12 Iliocostalis cervicis muscle
13 **Medial cutaneus branches of dorsal rami of spinal nerves** (C_7, C_8)
14 Longissimus thoracis muscle
15 Medial margin of scapula
16 **Rectus capitis posterior minor muscle**
17 **Obliquus capitis superior muscle**
18 **Rectus capitis posterior major muscle**
19 **Obliquus capitis inferior muscle**
20 Spinous process of axis
21 Semispinalis cervicis muscle
22 Semispinalis capitis muscle (divided and reflected)
23 Transverse cervical artery (superficial branch)
24 Serratus posterior superior muscle (divided and reflected)
25 Rhomboid minor muscle (divided and reflected)
26 Rhomboid major muscle (divided and reflected)

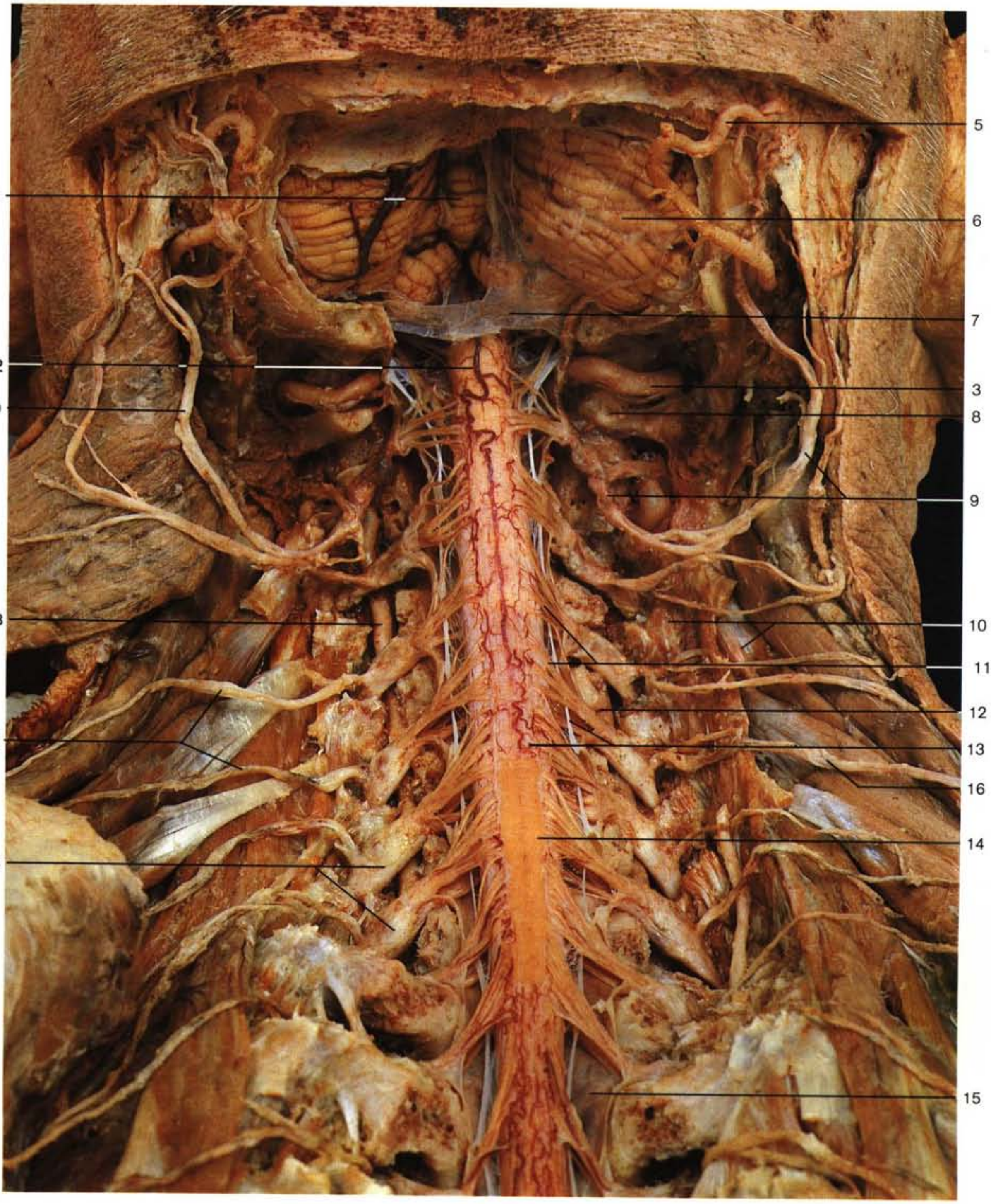

Neck, deepest layer. Spinal cord and medulla oblongata (dorsal aspect). Cranial cavity opened.

1 Vermis of the cerebellum
2 Medulla oblongata and posterior spinal artery
3 Vertebral artery
4 **Spinal ganglion**
5 Occipital artery
6 Cerebellum
7 **Cerebellomedullary cistern**
8 Atlas

9 Greater occipital nerve (C_2)
10 Levator scapulae muscle and intertransverse ligament
11 **Dorsal roots of spinal nerves**
12 Vertebral arch
13 **Denticulate ligament and arachnoid mater**
14 Area where pia mater has been removed
15 Dura mater
16 **Dorsal rami of spinal nerves**

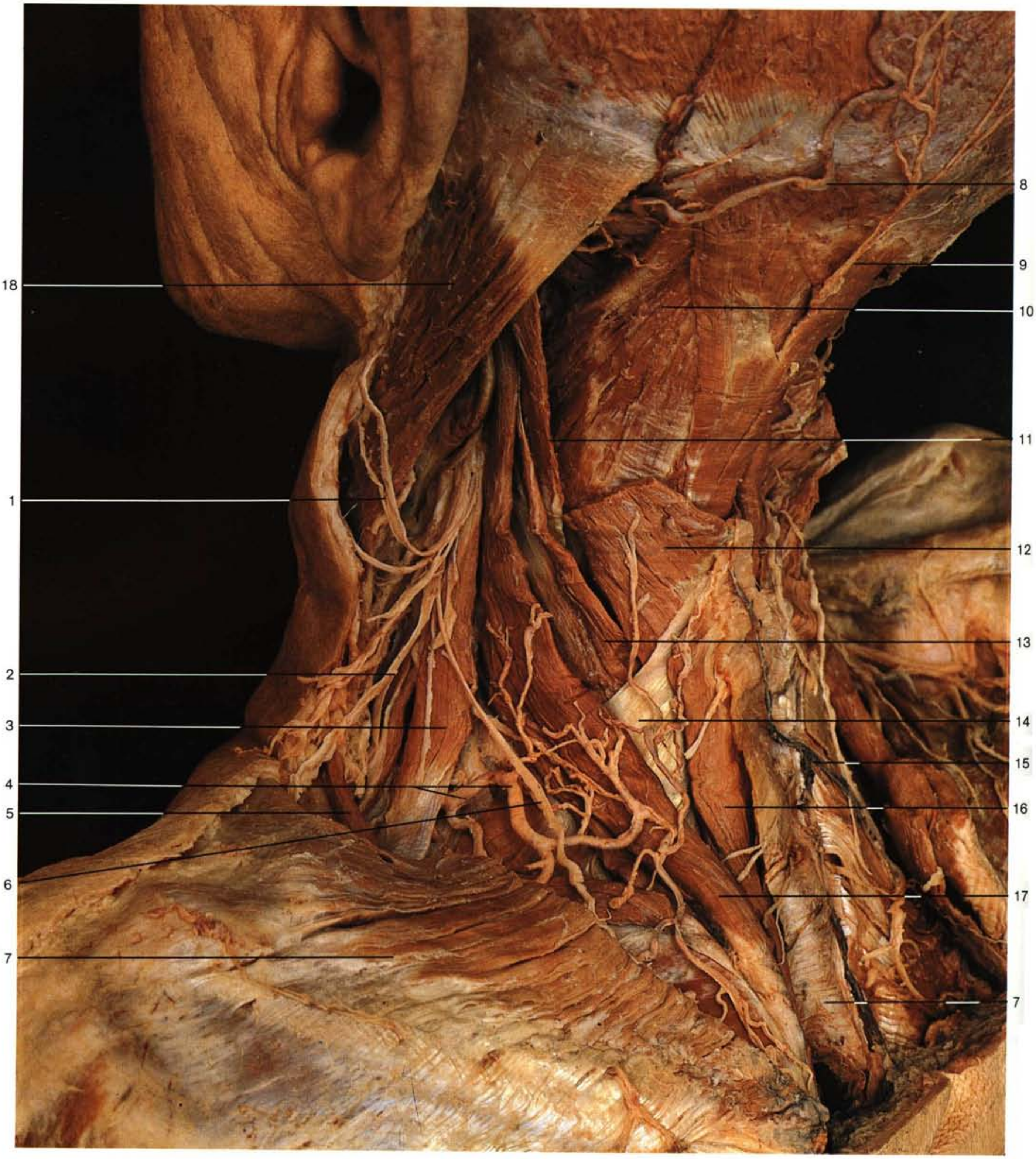

Deeper layer of the dorsal neck (oblique lateral aspect). Trapezius and semispinalis capitis muscles have been divided and partly removed.

1 Great auricular nerve
2 Accessory nerve (n. XI)
3 Scalenus posterior muscle
4 Transverse cervical artery
5 Omohyoid muscle
6 Muscular branch of cervical plexus
7 Trapezius muscle
8 Occipital artery
9 Greater occipital nerve
10 Semispinalis capitis muscle
11 Longissimus capitis muscle
12 Splenius capitis muscle
13 Splenius cervicis muscle
14 Serratus posterior superior muscle
15 Deep cervical vein (variant)
16 Rhomboid minor muscle
17 Levator scapulae muscle
18 Sternocleidomastoid muscle

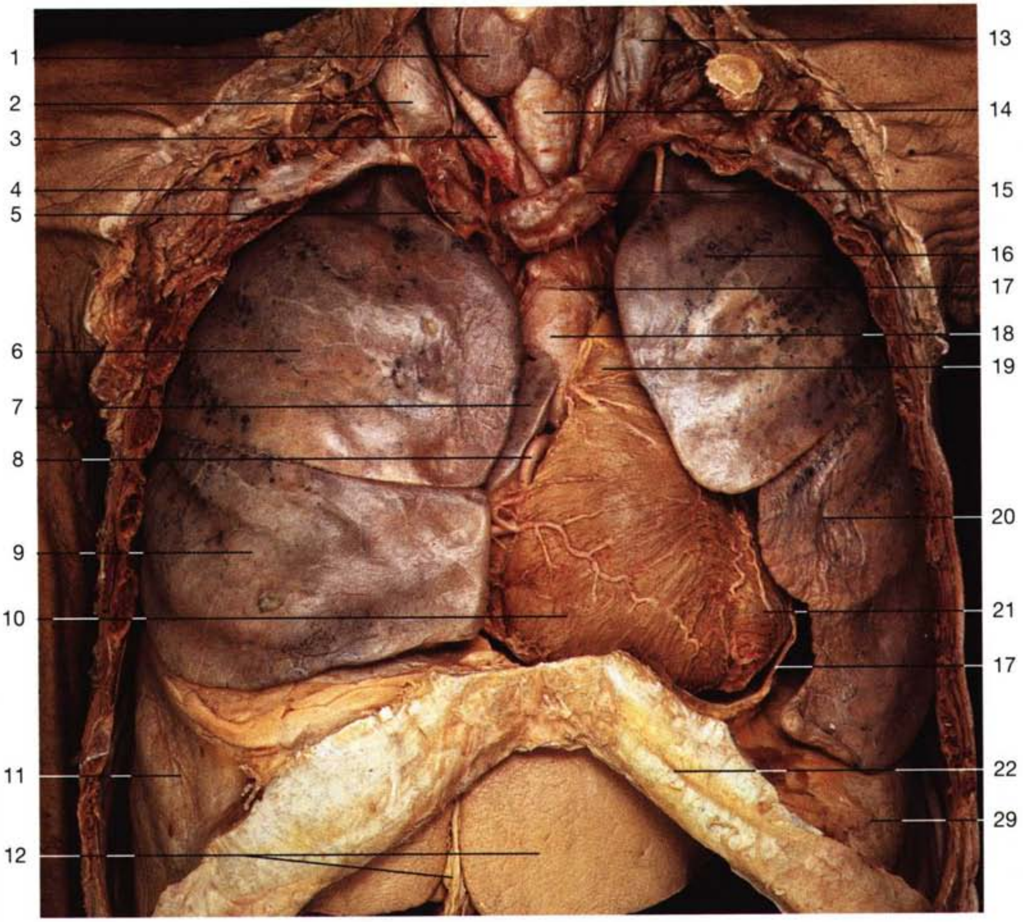

1 Thyroid gland
2 Internal jugular vein
3 Right common carotid artery
4 Right axillary vein
5 Right brachiocephalic vein
6 Superior lobe of right lung
7 Right atrium
8 Right coronary artery
9 Middle lobe of right lung
10 **Right ventricle**
11 Diaphragm
12 Liver (left lobe) and falciform ligament
13 Left internal jugular vein
14 **Trachea**
15 Left brachiocephalic vein
16 Superior lobe of left lung
17 Cut edge of pericardium
18 **Ascending aorta**
19 **Pulmonary trunk**
20 Inferior lobe of left lung
21 **Left ventricle**
22 Costal margin
23 Right pulmonary artery
24 Esophagus
25 Descending aorta
26 Pericardium
27 Aortic valve
28 Thymus
29 Diaphragm

Thoracic organs, heart and lungs in situ (ventral aspect). Anterior thoracic wall, parietal pleura and pericardium had been removed.

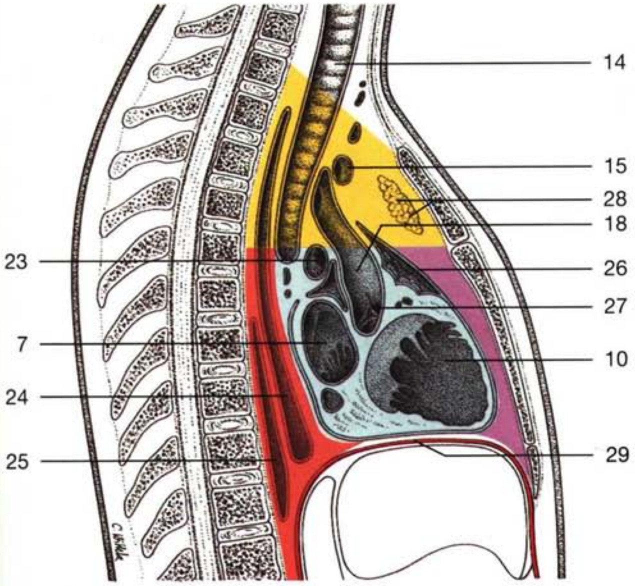

Sagittal section through thoracic cavity. The divisions of the mediastinum are indicated by colors.

Divisions of mediastinum	**Main content**
Superior mediastinum (yellow)	Trachea, brachiocephalic veins, thymus, aortic arch, esophagus, thoracic duct
Middle mediastinum (light blue)	Heart, ascending aorta, pulmonary trunk, pulmonary veins, phrenic nerves
Posterior mediastinum (red)	Esophagus with vagus nerves, descending aorta, thoracic duct, sympathetic trunks
Anterior mediastinum (pink)	Small vessels, connective and fatty tissue, and thymus in a child

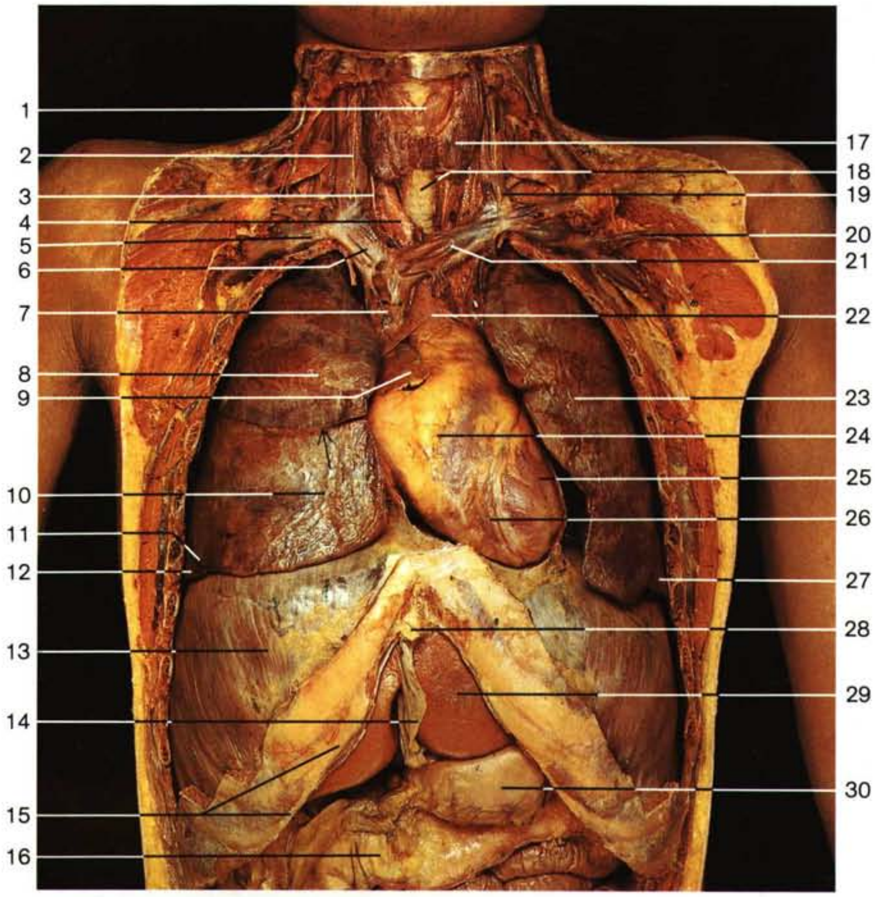
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

Positions of thoracic organs. The anterior thoracic wall has been removed. Arrow: horizontal fissure of right lung.

1 Cricothyroid muscle
2 Right internal jugular vein
3 Vagus nerve
4 Right common carotid artery
5 Right subclavian vein
6 Right brachiocephalic vein
7 Superior vena cava
8 Upper lobe of right lung
9 Right auricle
10 Middle lobe of right **lung**
11 Oblique fissure of right lung
12 Lower lobe of right lung
13 **Diaphragm**
14 Falciform ligament
15 Costal margin
16 Transverse colon
17 Thyroid gland
18 Trachea
19 Left internal jugular vein
20 Left cephalic vein
21 Left brachiocephalic vein
22 Pericardium (cut edge)
23 Upper lobe of left **lung**
24 **Right ventricle**
25 **Left ventricle**
26 Anterior interventricular sulcus
27 Lower lobe of left lung
28 Xiphoid process
29 **Liver**
30 **Stomach**
31 Pectoralis major muscle

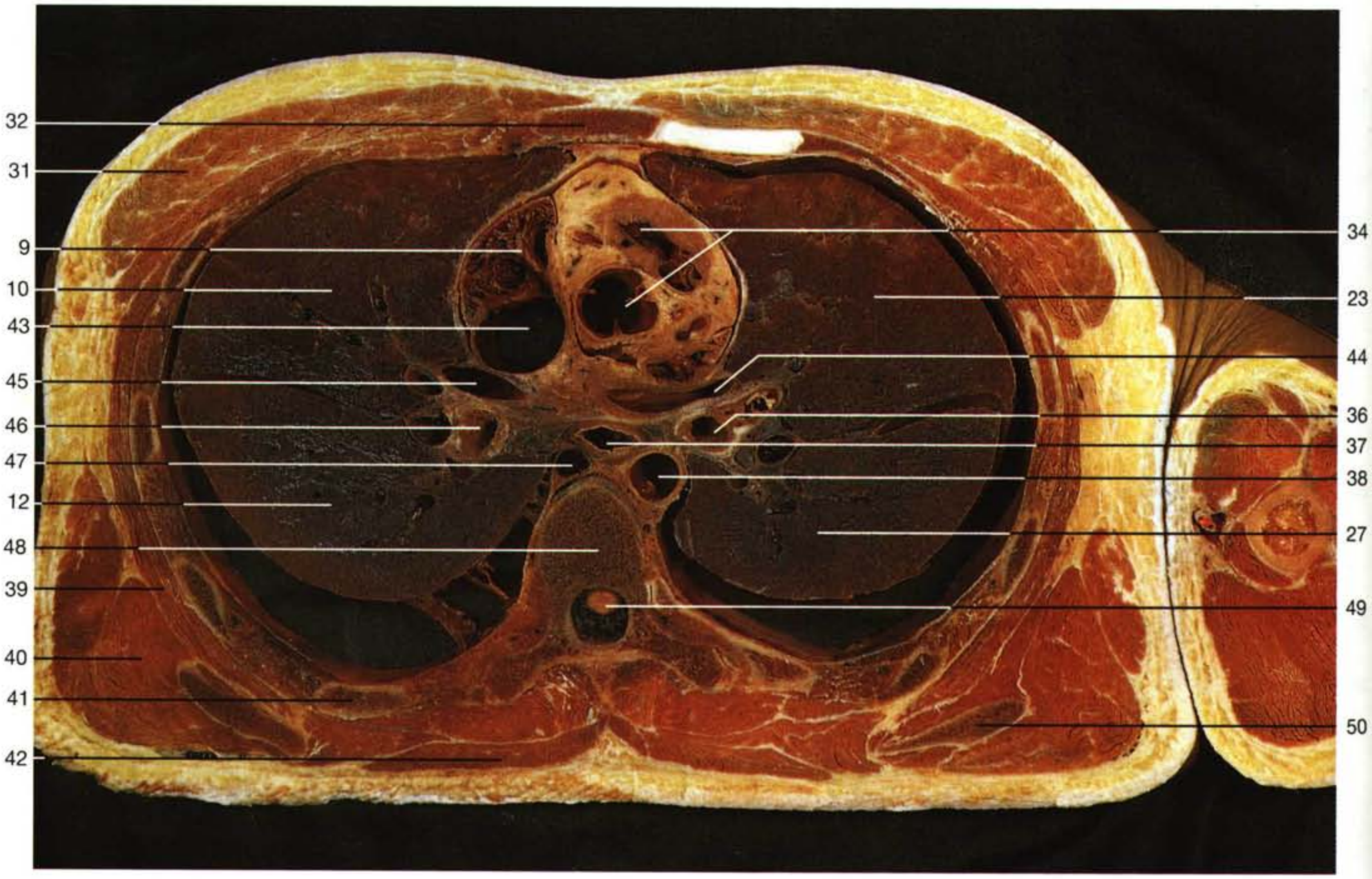
32
31
9
10
43
45
46
47
12
48
39
40
41
42
34
23
44
36
37
38
27
49
50

Horizontal section through the thorax at the level of the 7th thoracic vertebra (from below).

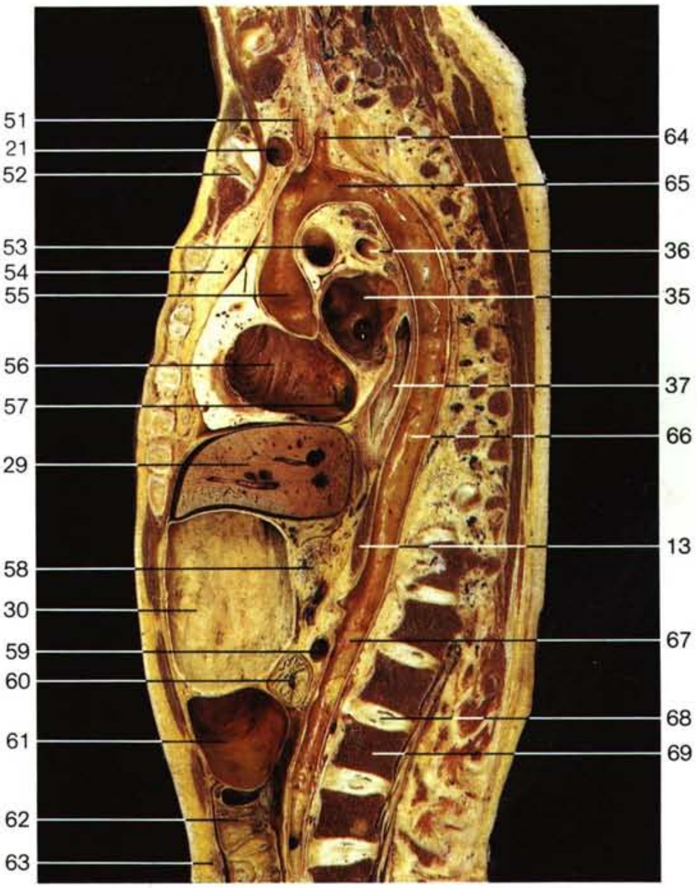
51
21
52
53
54
55
56
57
29
58
30
59
60
61
62
63
64
65
36
35
37
66
13
67
68
69

Sagittal section through the left thorax, 2 cm lateral to the median plane.

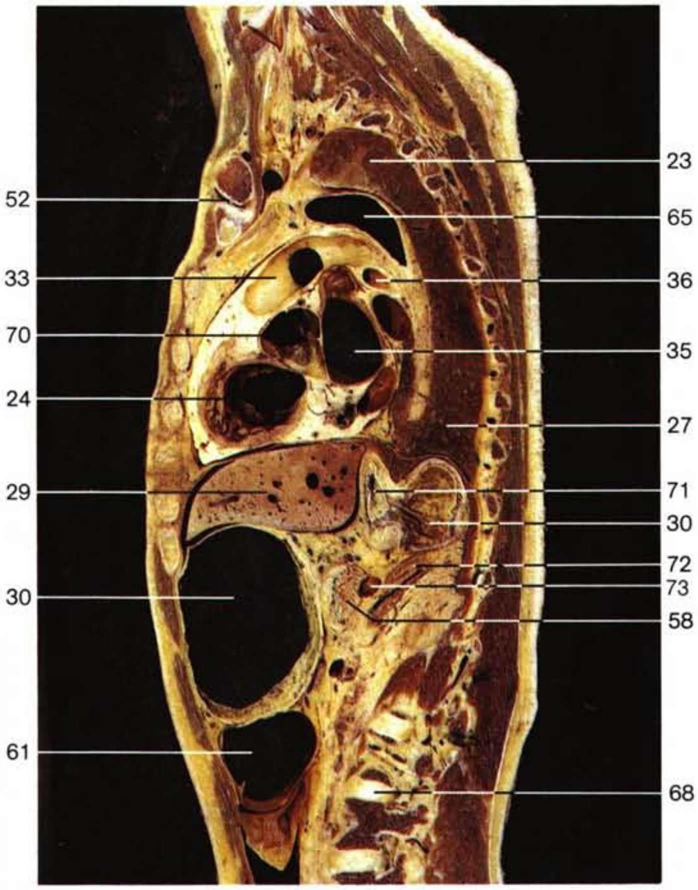
52
33
70
24
29
30
61
23
65
36
35
27
71
30
72
73
58
68

Sagittal section through the left thorax, 3.5 cm lateral to the median plane.

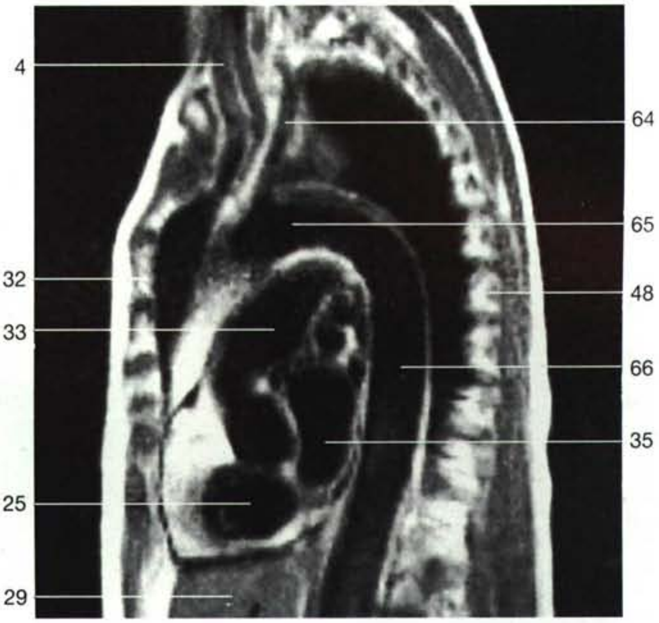
4
32
33
25
29
64
65
48
66
35

Sagittal section through the thorax. MR-Scan.

32	Sternum
33	Pulmonary trunk
34	Left ventricle and bulb of aorta
35	Left atrium
36	Left main bronchus
37	**Esophagus**
38	Descending aorta
39	Serratus anterior muscle
40	Teres major muscle
41	Rib
42	Trapezius muscle
43	Right atrium
44	Left pulmonary vein
45	Right pulmonary vein
46	Right main bronchus
47	Azygos vein
48	Body of vertebra
49	Spinal cord
50	Scapula
51	Left common carotid artery
52	Sternoclavicular articulation with disc
53	Right pulmonary artery
54	Remnants of thymus
55	Aortic bulb
56	Right atrium
57	Entrance of inferior vena cava in right atrium
58	**Pancreas**
59	Left renal vein
60	**Duodenum**
61	Transverse colon (dilated)
62	Small intestine
63	Umbilicus
64	Left subclavian artery
65	Aortic arch
66	**Thoracic aorta**
67	Abdominal aorta
68	Intervertebral disc
69	Body of lumbar vertebra
70	Aortic valve
71	Cardia of stomach
72	Suprarenal gland
73	Splenic vein

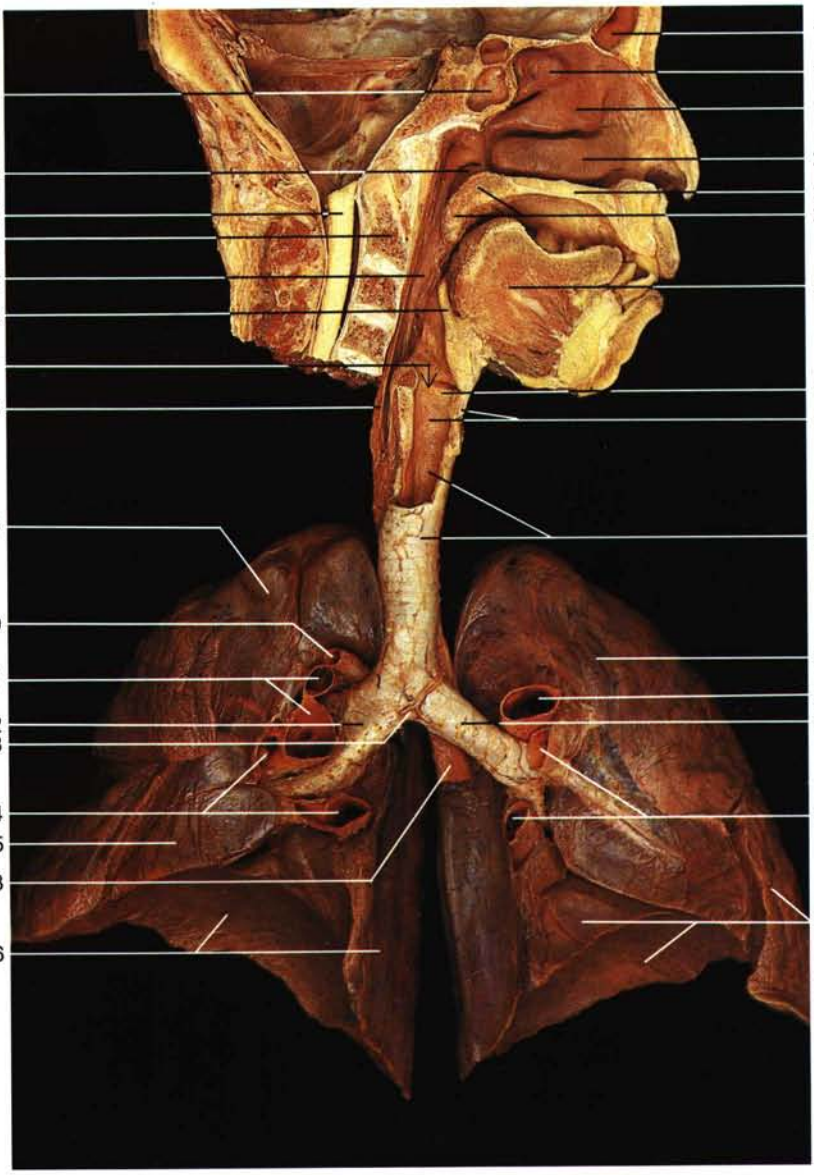

Respiratory system. The lungs have been fixed in expiration and turned laterally. Head bisected and turned laterally.

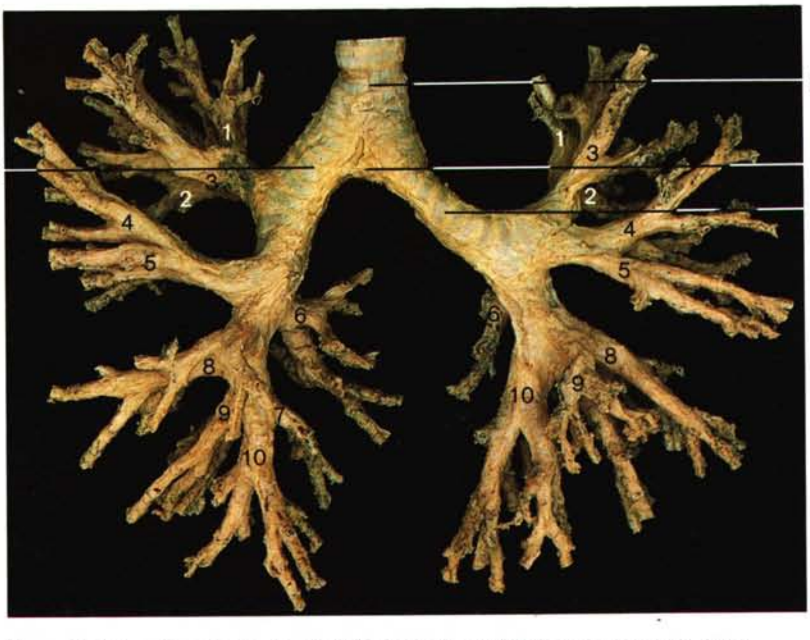

Bronchial tree (ventral aspect). The lung tissue has been removed. The bronchopulmonary segments are numbered 1–10.

1 Sphenoid sinus
2 Pharyngeal opening of auditory tube
3 Spinal cord
4 Dens of axis
5 **Oropharynx** (oropharyngeal isthmus)
6 Epiglottis
7 Entrance of larynx
8 Esophagus
9 Upper lobe of right lung
10 Azygos vein
11 Branches of pulmonary artery
12 **Right main bronchus**
13 **Bifurcation of trachea**
14 Tributaries of right pulmonary veins
15 Middle lobe of right lung
16 Lower lobe of right lung
17 Frontal sinus
18 Superior nasal concha
19 Middle nasal concha
20 Inferior nasal concha
21 Hard palate
22 Soft palate with uvula
23 Tongue
24 Vocal fold
25 Larynx
26 Trachea
27 Upper lobe of left lung
28 Left pulmonary artery
29 **Left main bronchus**
30 Left pulmonary veins
31 Lower lobe of left lung

▷

To page 231:
1 Nasal cavity
2 Pharynx
3 Larynx (thyroid cartilage)
4 Trachea
5 Upper lobe of right lung
6 Bifurcation of trachea
7 Right main bronchus
8 Horizontal fissure of right lung
9 Middle lobe of right lung
10 Oblique fissures of lungs
11 Lower lobe of right lung
12 Clavicle
13 Upper lobe of left lung
14 Left main bronchus
15 Bronchi supplying bronchopulmonary segments
16 Lower lobe of left lung
17 Costal margin
18 Hyoid bone
19 Right superior lobe bronchus
20 Right middle lobe bronchus
21 Right inferior lobe bronchus
22 Left superior lobe bronchus
23 Left inferior lobe bronchus
24 Segmental bronchi
25 Branches of pulmonary arteries
26 Branches of pulmonary veins

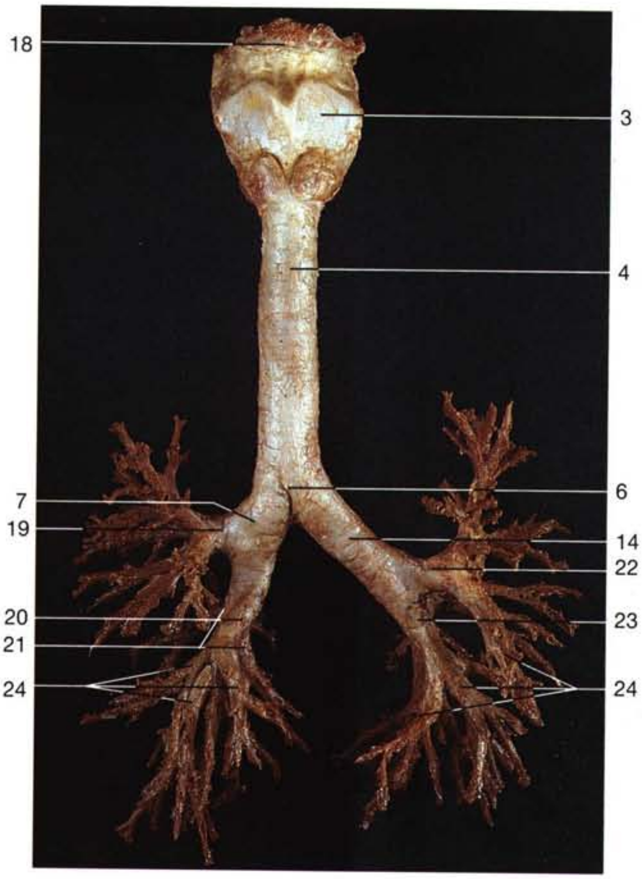
18
3
4
7
19
6
14
22
20
23
21
24
24

Larynx, trachea and bronchial tree (anterior aspect).

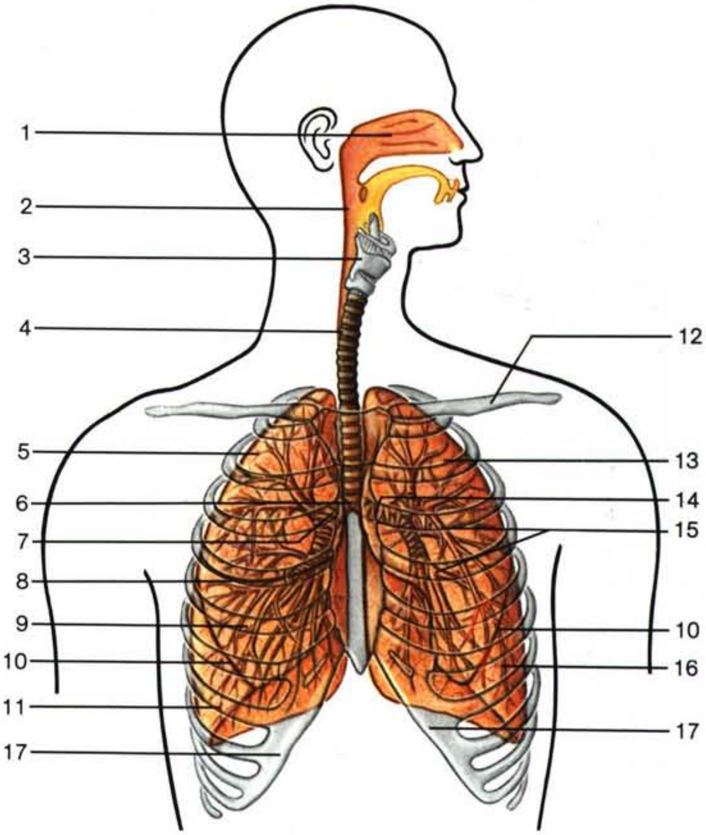
1
2
3
4
12
5
13
6
14
7
15
8
9
10
10
11
16
17
17

Organization and positions of respiratory organs.
(Schematic drawing.)

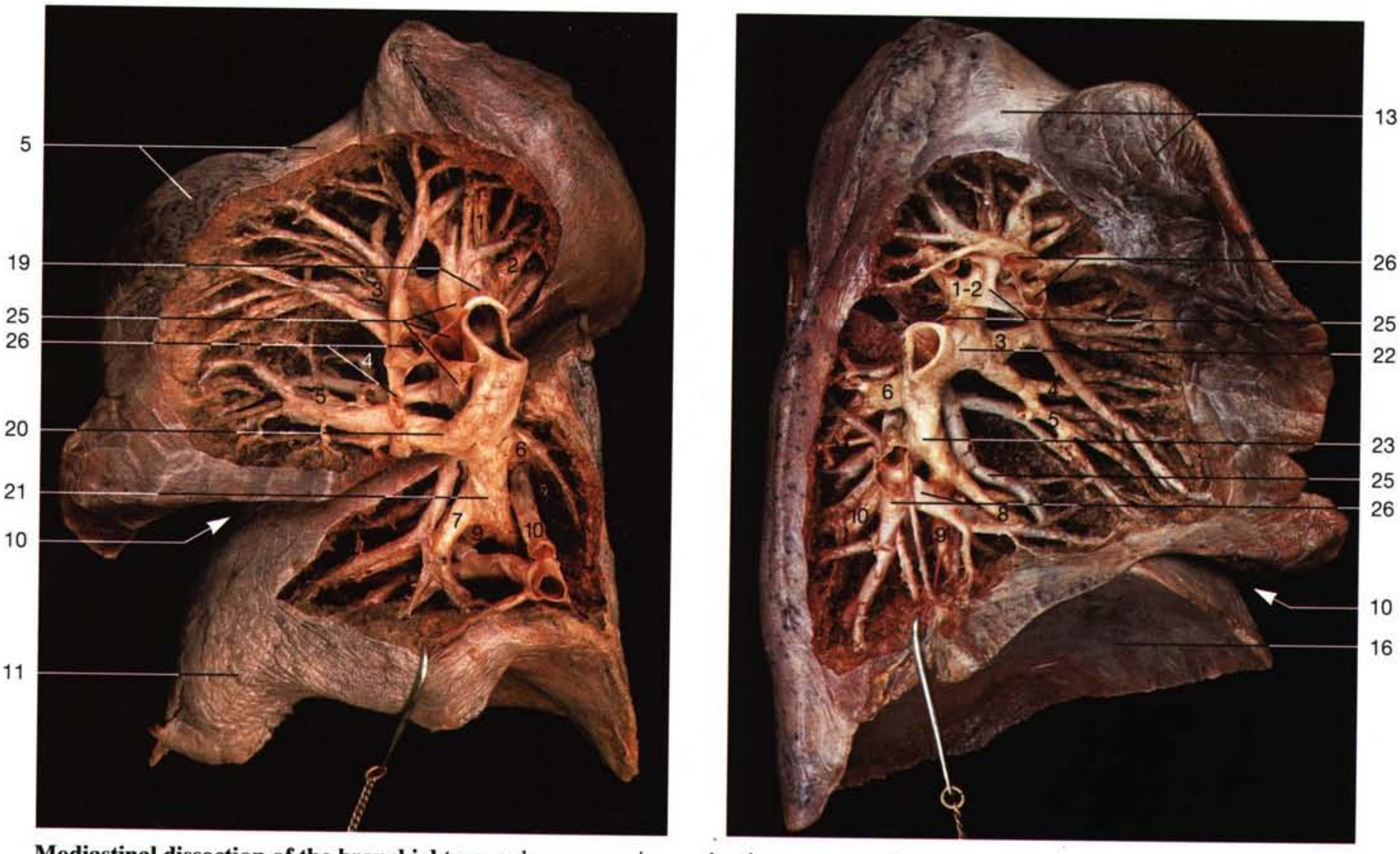
5
19
25
26
20
21
10
11
1
2
3
4
5
6
7
9
10
13
26
1-2
25
3
22
6
4
5
23
25
26
8
10
9
10
16

Mediastinal dissection of the bronchial tree, pulmonary veins, and pulmonary arteries of right lung (left) and left lung (right) (medial aspect). Segmental bronchi are numbered 1–10.

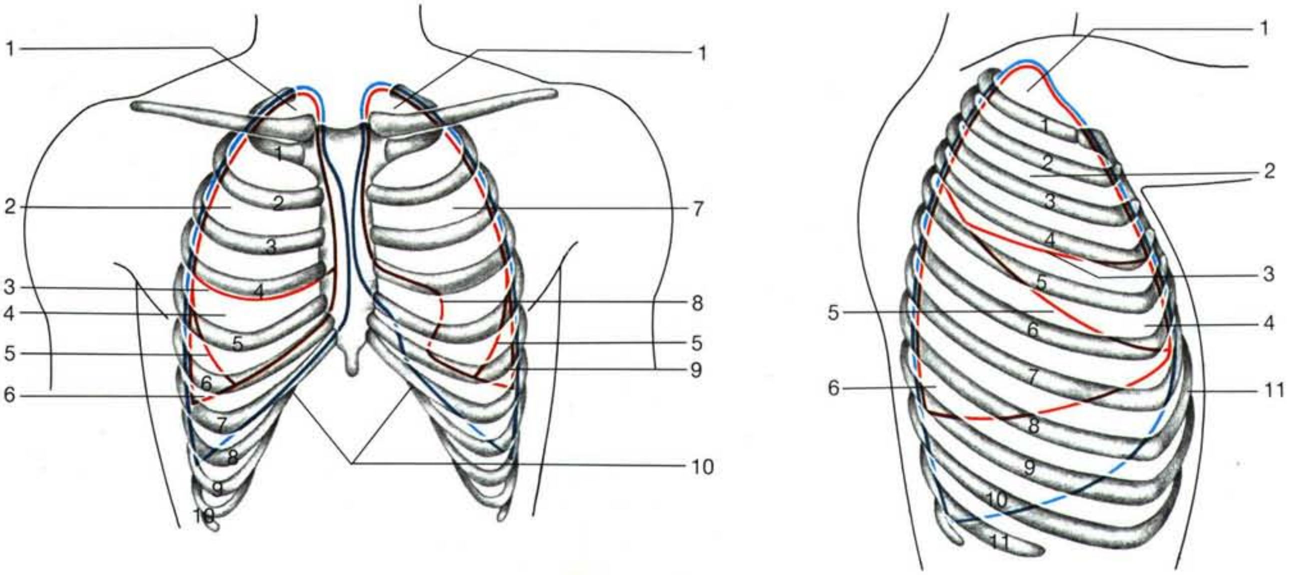

Surface projections of lungs and pleura on the thoracic wall. Left: anterior aspect; right: right-lateral aspect. Red = margins of the lung; blue = margins of pleura. The numbers indicate ribs.

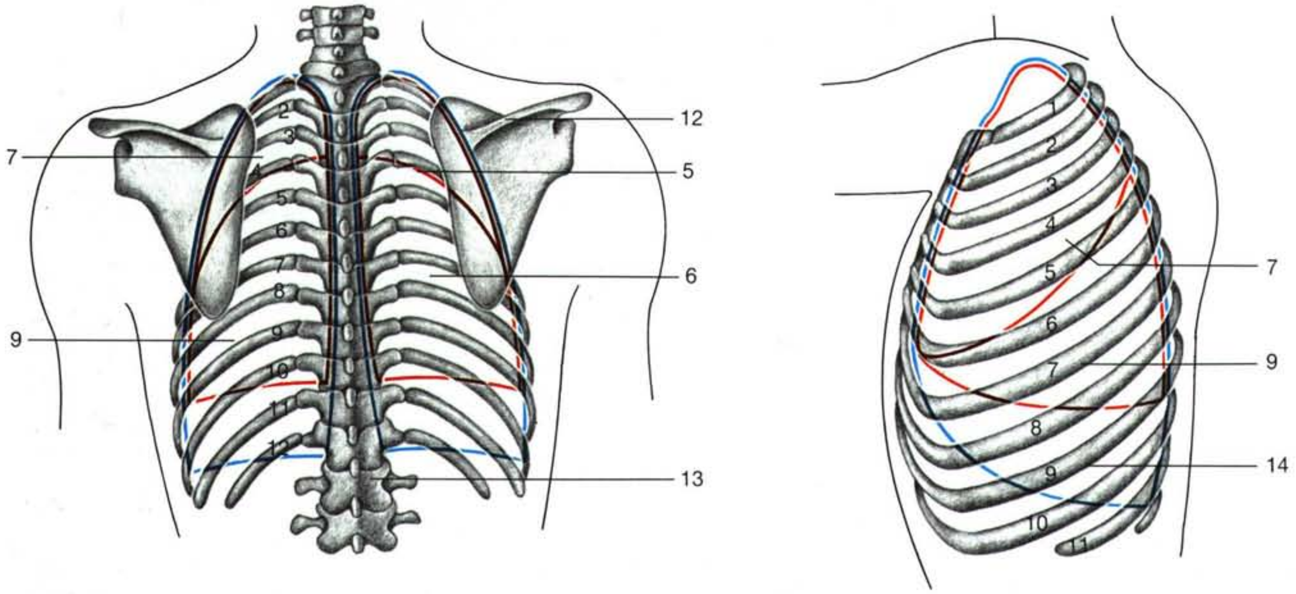

Surface projections of lungs and pleura on thoracic wall. Left: posterior aspect; right: left-lateral aspect. Red = margins of lung; blue = margins of pleura. The number indicate ribs.

1 Apex of lung
2 Upper lobe of right lung
3 Horizontal fissure of right lung
4 Middle lobe of right lung
5 Oblique fissures of lungs
6 Lower lobe of right lung
7 Upper lobe of left lung
8 Cardiac notch of left lung
9 Lower lobe of left lung
10 Infrasternal angle
11 Costal margin
12 Spine of scapula
13 1st lumbar vertebra
14 Space between border of lung and pleura (costodiaphragmatic recess)

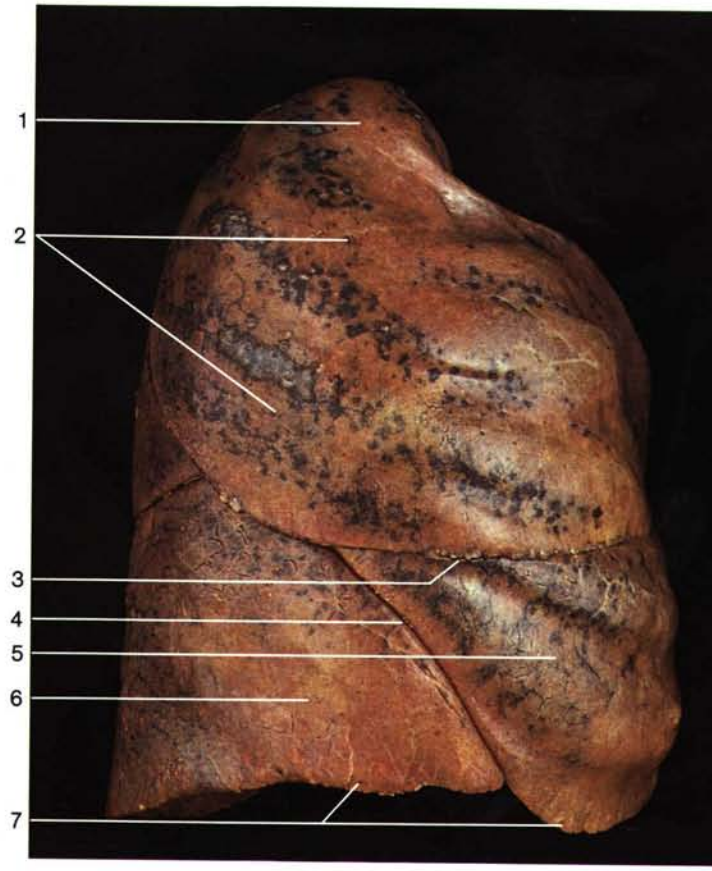

Right lung (lateral aspect).

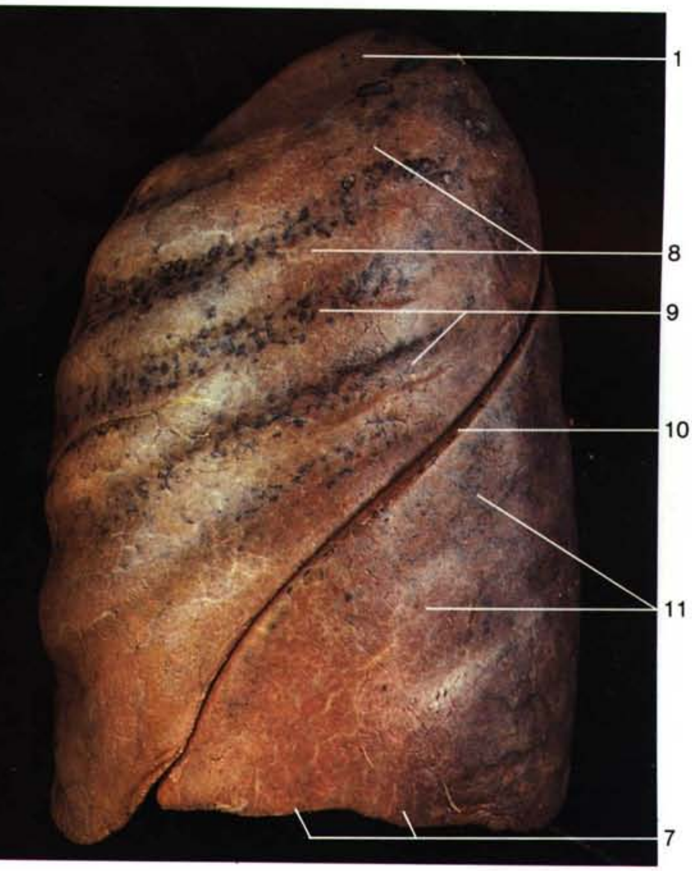

Left lung (lateral aspect).

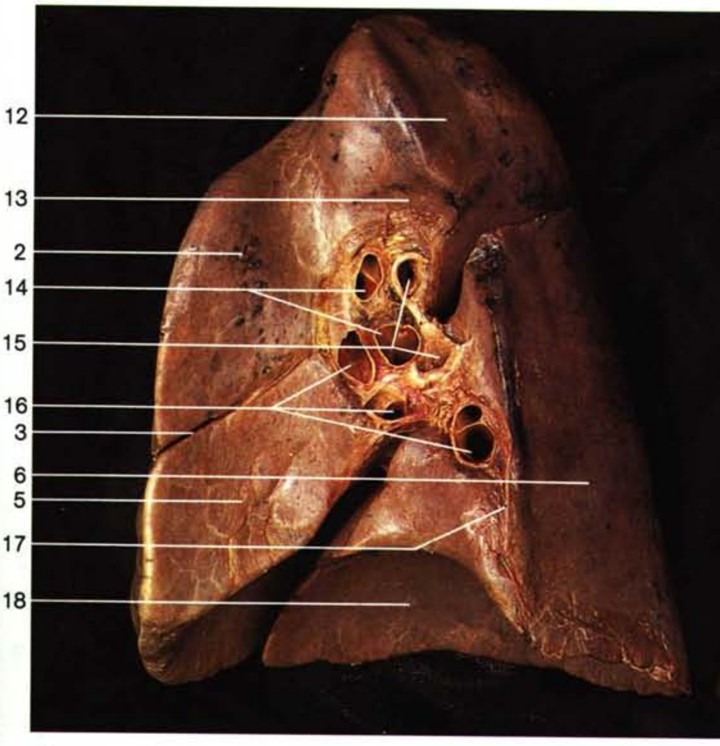

Right lung (medial aspect).

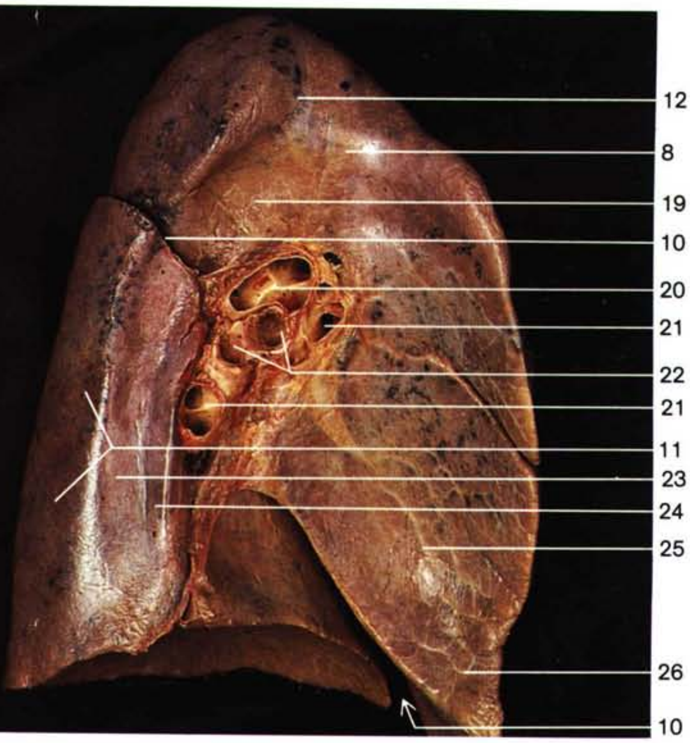

Left lung (medial aspect).

1 Apex of lung
2 Upper lobe of right lung
3 Horizontal fissure of right lung
4 Oblique fissure of right lung
5 Middle lobe of right lung
6 Lower lobe of right lung
7 Inferior border
8 Upper lobe of left lung
9 Impressions of ribs
10 Oblique fissure of left lung
11 Lower lobe of left lung
12 Groove of subclavian artery
13 Groove of azygos arch
14 Branches of right pulmonary artery
15 Bronchi
16 Right pulmonary veins
17 Pulmonary ligament
18 Diaphragmatic surface
19 Groove of aortic arch
20 Left pulmonary artery
21 Branches of left pulmonary veins
22 Left secondary bronchi
23 Groove of thoracic aorta
24 Groove of esophagus
25 Cardiac impression
26 Lingula

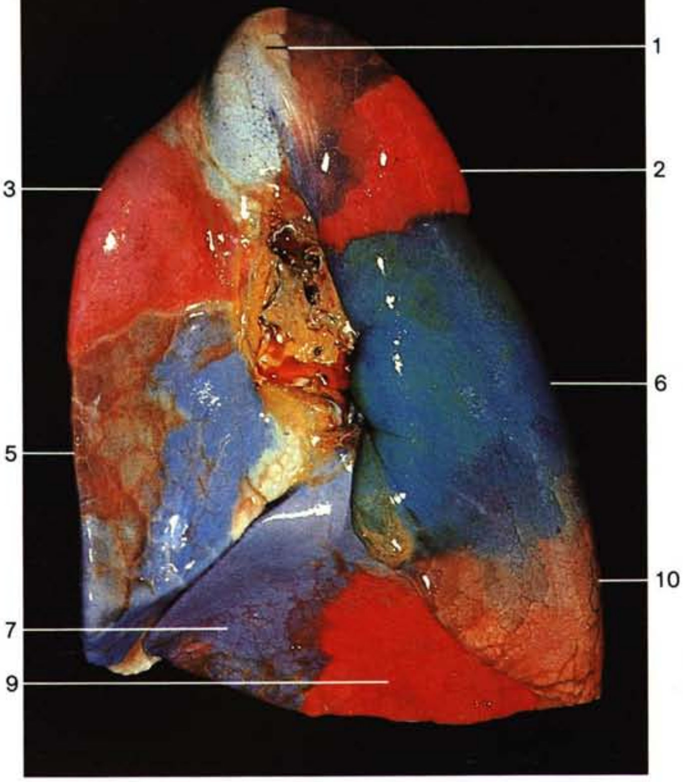

Right lung (medial aspect).

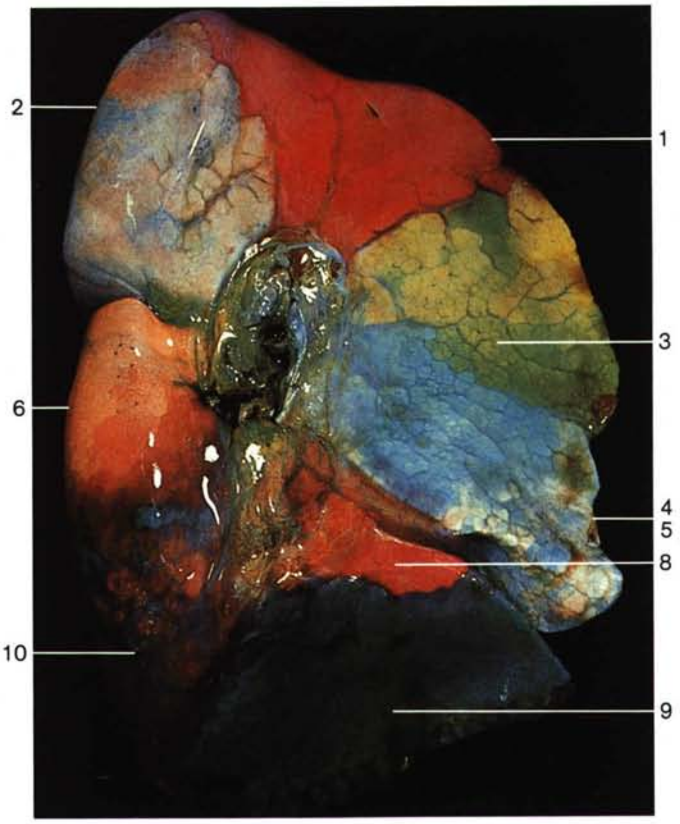
2
1
3
6
4
5
8
10
9

Left lung (medial aspect).

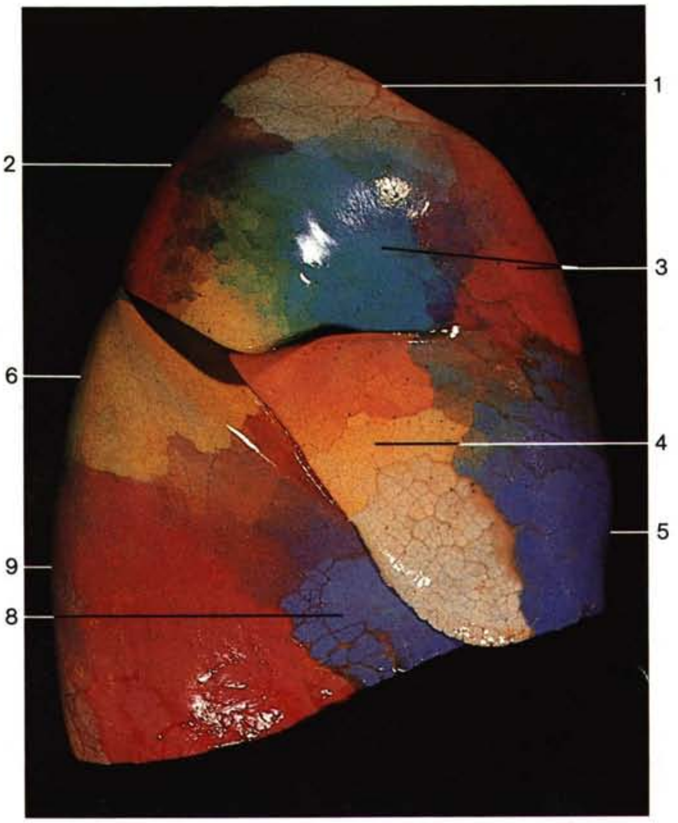

Right lung (lateral aspect).

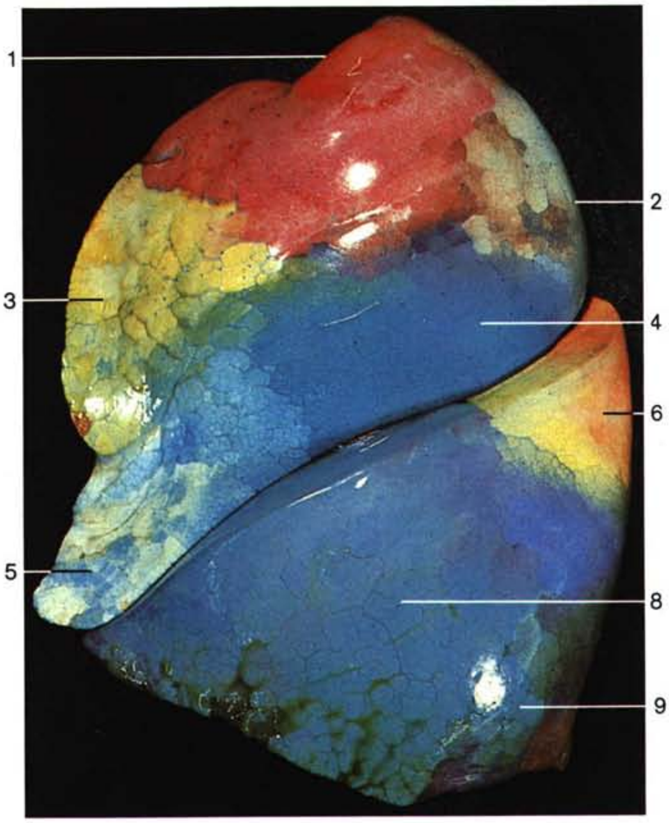

Left lung (lateral aspect).

The bronchopulmonary segments of the lungs are differentiated by the various colors. Notice that there is no segment in the left lung that corresponds to the 7th segment of the right lung. Compare with the schematic drawing on the facing page.

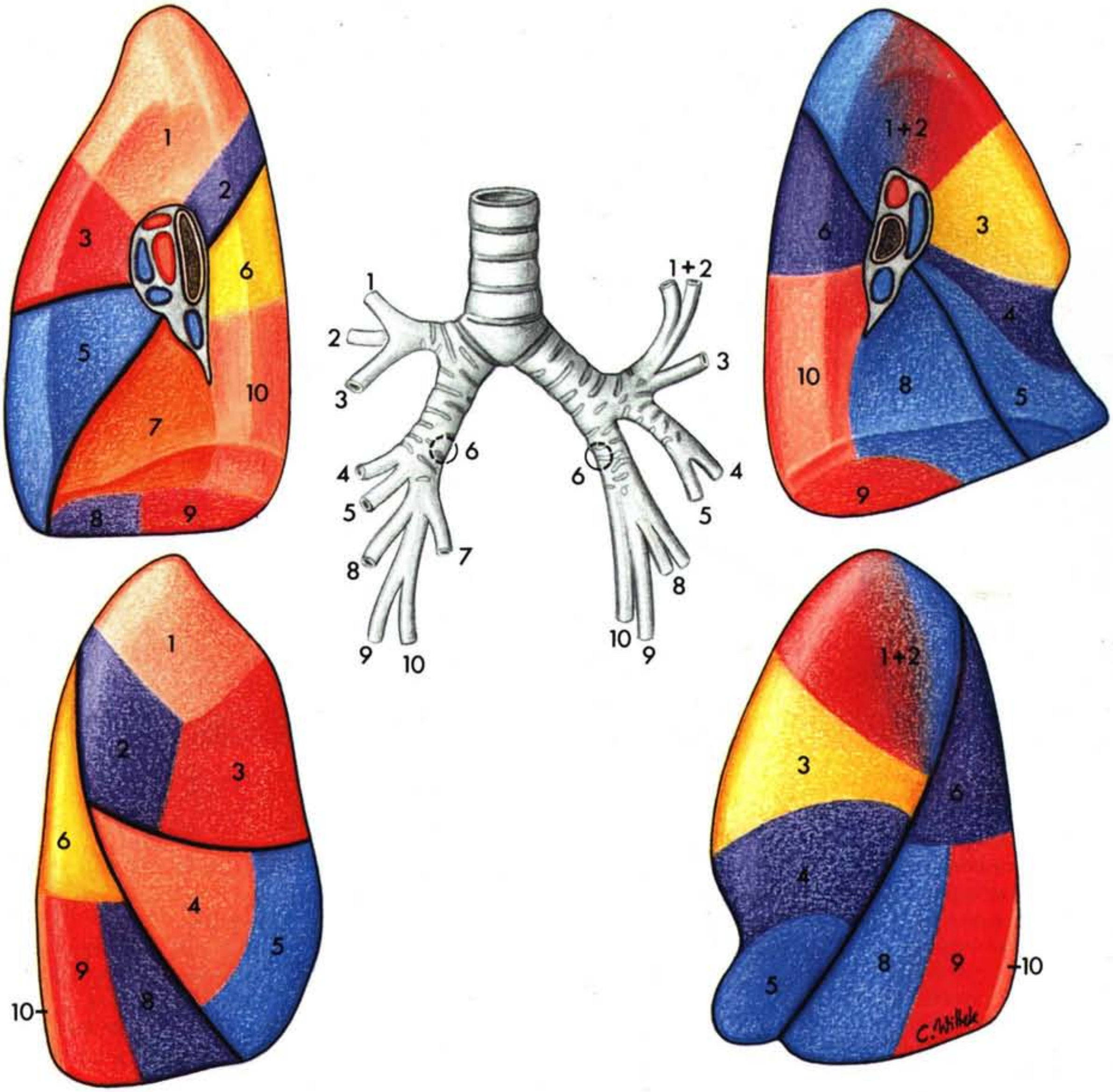

Distribution of bronchopulmonary segments of the lungs and their relation to the bronchial tree (after J. F. Huber).
The bronchopulmonary segments are morphologically and functionally separate independent respiratory units of the lung tissue. Each segment is surrounded by connective tissue which is continuous with the visceral pleura. The segmental bronchi in a segment are central, closely accompanied by branches of the pulmonary arteries whereas the tributaries of the pulmonary veins run **between** the segments. Thus, the veins serve two adjacent segments which drain for the most part into more than one vein. A bronchopulmonary segment is therefore not a complete vascular unit, but segmentation is the result of a specific architecture of the lung vasculature.

	Right lung			**Left lung**		
1	Apical segment	Upper lobe bronchus	1+2	Apico-posterior segment	Superior division	Upper lobe bronchus
2	Posterior segment		3	Anterior segment		
3	Anterior segment					
4	Lateral segment	Middle lobe bronchus	4	Superior lingular segment	Inferior division	
5	Medial segment		5	Inferior lingular segment		
6	Superior (apical) segment	Lower lobe bronchus	6	Superior (apical) segment		Lower lobe bronchus
7	Medial basal segment		7	Absent		
8	Anterior basal segment		8	Anteromedial basal segment		
9	Lateral basal segment		9	Lateral basal segment		
10	Posterior basal segment		10	Posterior basal segment		

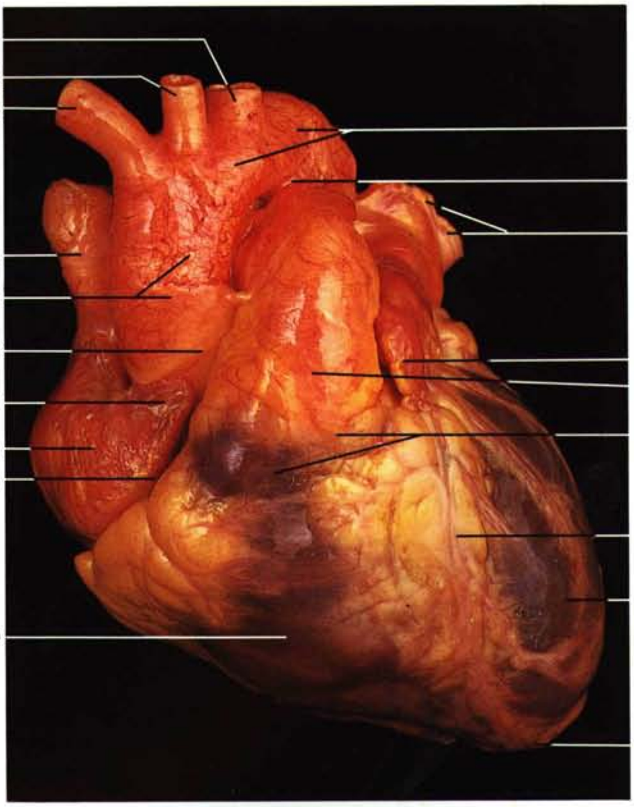

Heart of 30-year-old woman (anterior aspect).

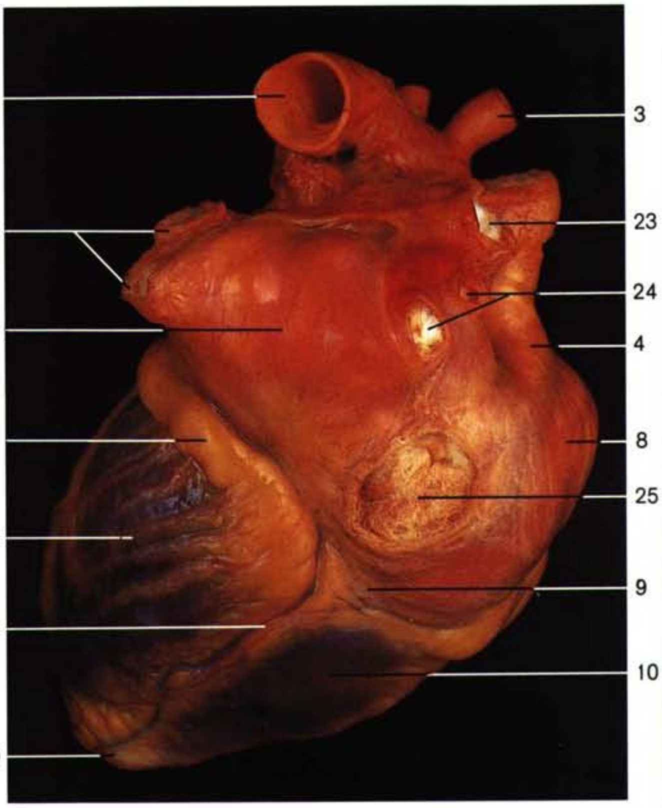

Heart of 30-year-old woman (oblique posterior view).

1 Left subclavian artery
2 Left common carotid artery
3 Brachiocephalic trunk
4 Superior vena cava
5 Ascending aorta
6 Bulb of the aorta
7 Right auricle
8 **Right atrium**

9 Coronary sulcus
10 **Right ventricle**
11 Aortic arch
12 Ligamentum arteriosum
13 Left pulmonary veins
14 Left auricle
15 **Pulmonary trunk**
16 Sinus of pulmonary trunk
17 Anterior interventricular sulcus

18 **Left ventricle**
19 Apex of the heart
20 **Left atrium**
21 Epicardial fat overlying coronary sinus
22 Posterior interventricular sulcus
23 Right pulmonary artery
24 Right pulmonary veins
25 Inferior vena cava

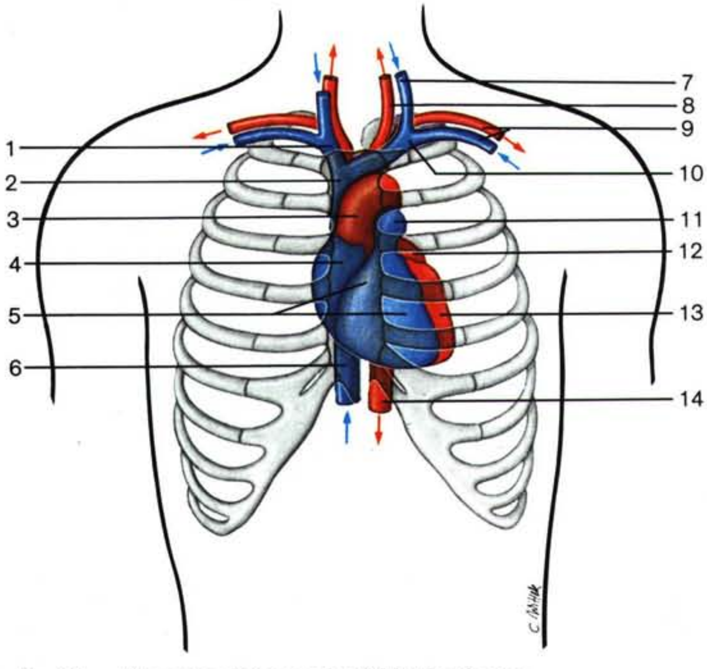

Position of heart and its vessels within the thorax. (Schematic drawing.)

1 Right brachiocephalic vein
2 Superior vena cava
3 Ascending aorta
4 Right atrium
5 Right ventricle
6 Inferior vena cava
7 Left internal jugular vein
8 Left common carotid artery
9 Left axillary artery and vein
10 Left brachiocephalic vein
11 Pulmonary trunk
12 Left auricle
13 Left ventricle
14 Descending aorta

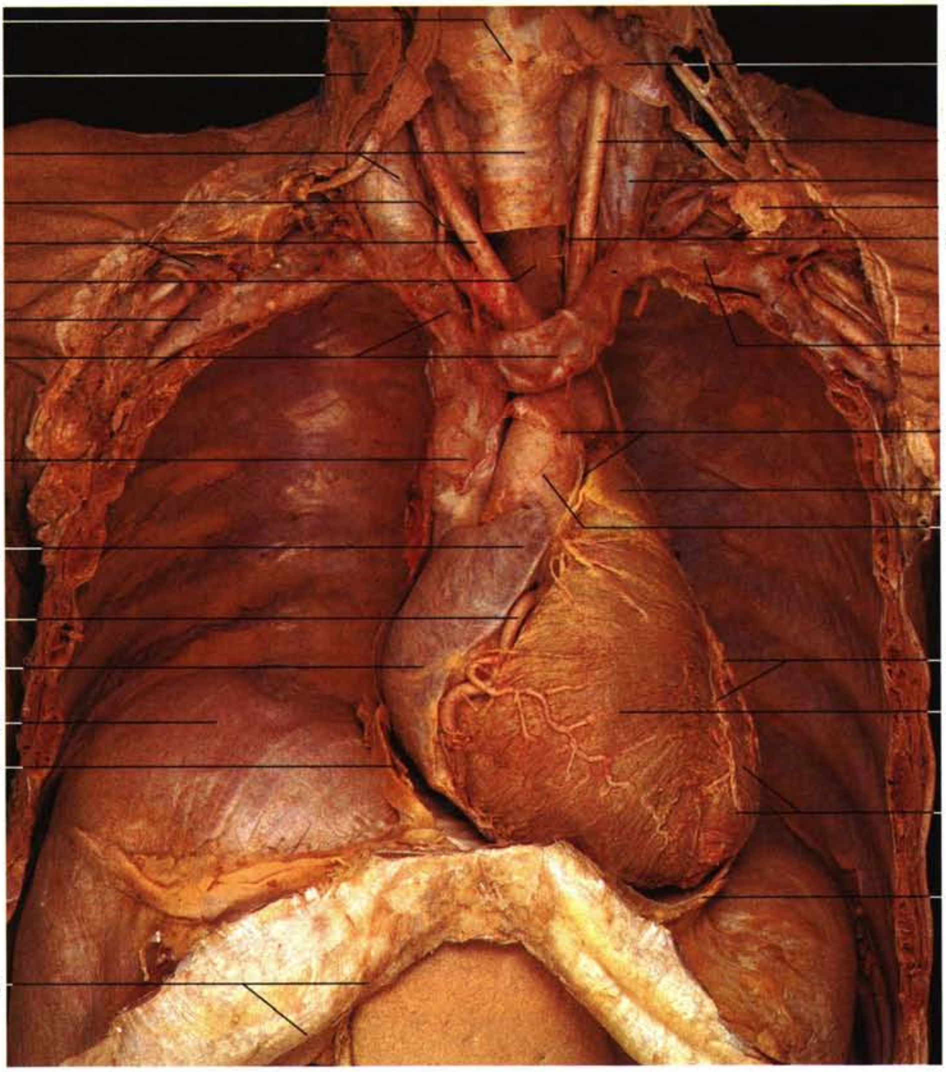

Heart and related vessels in situ (anterior aspect). Anterior thoracic wall, pericardium and epicardium have been removed; trachea divided.

1 Larynx (thyroid cartilage)
2 Sternocleidomastoid muscle (divided)
3 Trachea (divided) and right internal jugular vein
4 Vagus nerve
5 Right common carotid artery and cephalic vein
6 Esophagus
7 Right axillary vein
8 Right and left brachiocephalic veins
9 **Superior vena cava**
10 Right auricle
11 Right **coronary artery**
12 **Right atrium**
13 Diaphragm
14 Pericardium (cut edges)
15 Costal margin
16 Omohyoid muscle
17 Left **common carotid artery**
18 Left internal jugular vein
19 Clavicle (divided)
20 Left recurrent laryngeal nerve
21 **Subclavian vein**
22 Pericardial reflection
23 **Pulmonary trunk**
24 **Ascending aorta**
25 Anterior interventricular sulcus and anterior interventricular branch of left coronary artery
26 **Right ventricle**
27 **Left ventricle**
28 Aortic valve
29 Tricuspid or right atrioventricular valve
30 Inferior vena cava
31 Pulmonary veins
32 Pulmonary valve
33 Left atrioventricular (bicuspid or mitral) valve

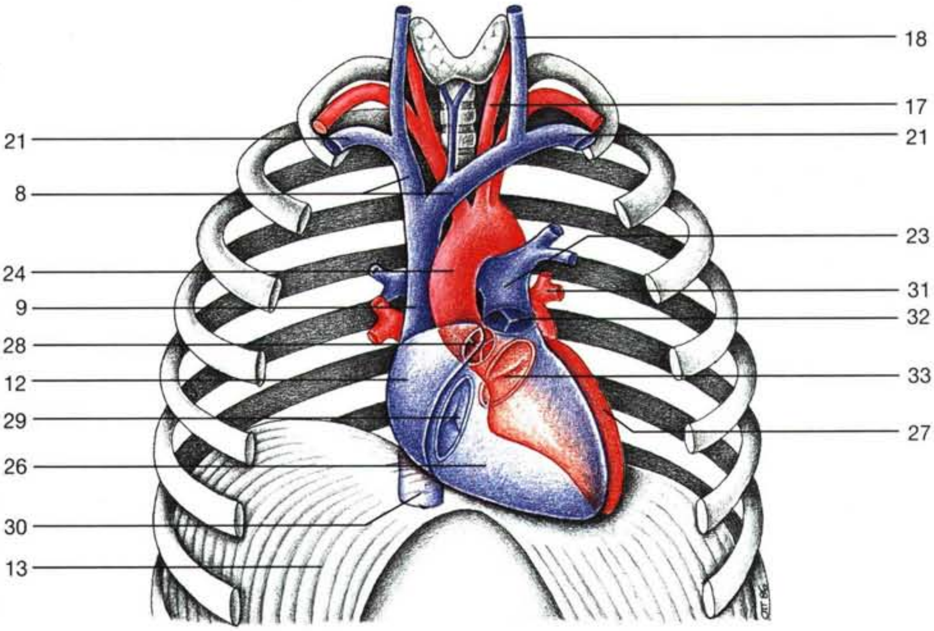

Heart in situ. Position of valves (anterior aspect). (Schematic drawing.)

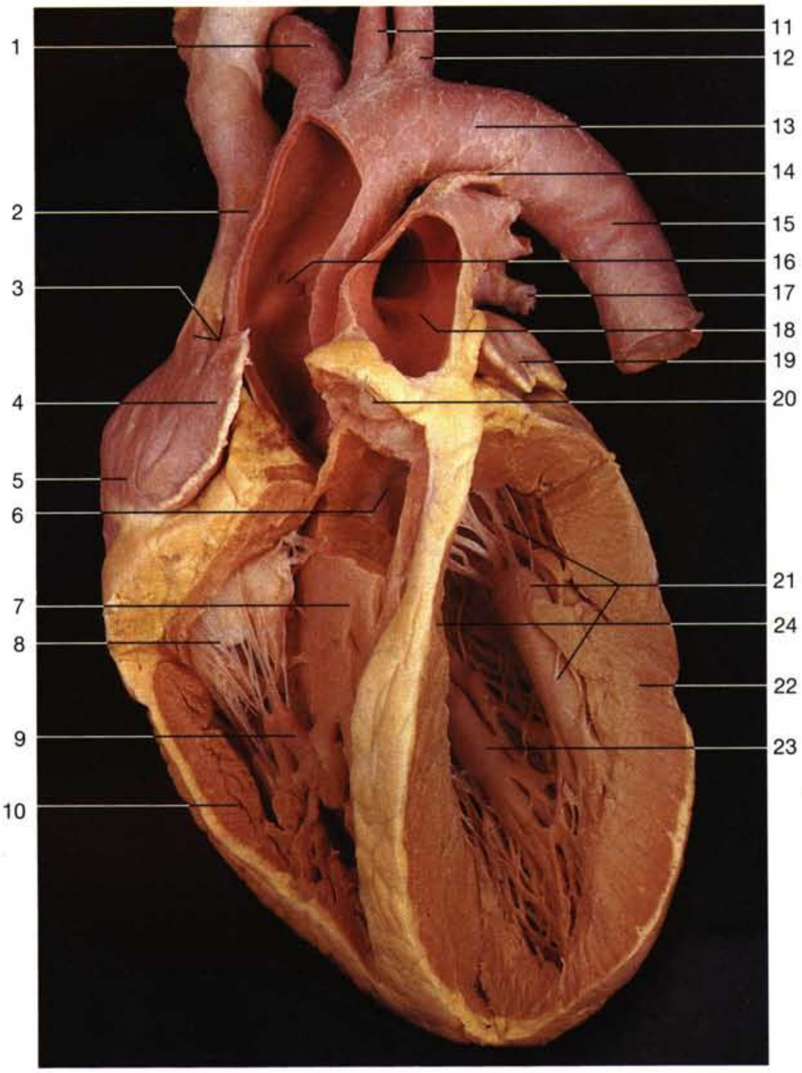

Anterior aspect of the heart. The anterior walls of the ventricles and of the aorta and pulmonary trunk have been fenestrated to show the aortic valve.

1 Brachiocephalic trunk
2 Superior vena cava
3 Sulcus terminalis
4 Right auricle
5 Right atrium
6 **Aortic valve**
7 Conus arteriosus (interventricular septum)
8 **Right atrioventricular (tricuspid) valve**
9 **Anterior papillary muscle**
10 Myocardium of right ventricle
11 Left common carotid artery
12 Left subclavian artery
13 **Aortic arch**
14 Ligamentum arteriosum (remnant of ductus arteriosus)
15 Thoracic aorta (descending aorta)
16 **Ascending aorta**
17 Left pulmonary vein
18 **Pulmonary trunk**
19 Left auricle
20 Pulmonic valve
21 **Anterior papillary muscle** with chordae tendineae
22 Myocardium of left ventricle
23 Posterior papillary muscle
24 Interventricular septum
25 Right and left brachiocephalic veins
26 **Chordae tendineae**
27 Papillary muscles of right ventricle
28 Left atrium
29 Infundibulum
30 Anterior papillary muscle of left ventricle
31 Left atrioventricular (bicuspid or mitral) valve and chordae tendineae
32 Apex of heart

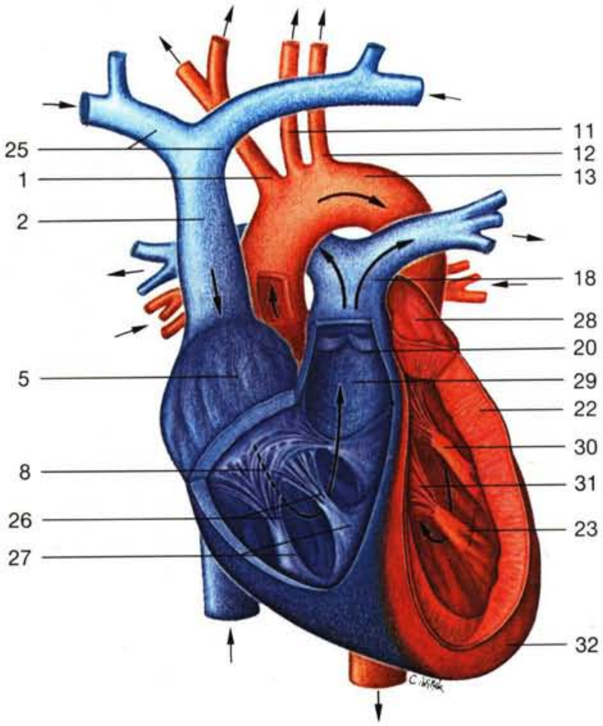

Circulation within the heart. (Schematic drawing.) The arrows indicate the direction of the blood flow through the heart. Blue = right heart; red = left heart.

Heart in situ. Myocardium and coronary arteries (anterior aspect).

1	Internal jugular vein
2	Common carotid artery
3	Brachiocephalic trunk
4	**Ascending aorta**
5	Right lung
6	Right auricle
7	**Right coronary artery**
8	**Myocardium of right ventricle**
9	Diaphragm
10	Costal margin
11	Thyroid gland and internal jugular vein
12	Trachea and left common carotid artery
13	Left brachiocephalic vein
14	Left lung
15	**Pericardium** (cut edge)
16	Pulmonary trunk
17	Anterior interventricular artery
18	**Myocardium of left ventricle**
19	**Muscular vortex** (right ventricle)
20	Posterior interventricular sulcus
21	Anterior interventricular sulcus
22	**Muscular vortex** (left ventricle)
23	Aortic arch
24	Left atrium
25	Coronary sinus
26	Superior vena cava
27	Right pulmonary vein
28	Right atrium
29	Inferior vena cava
30	Coronary sulcus
31	Myocardium of left ventricle
32	Left pulmonary artery
33	Left pulmonary vein
34	Apex of heart

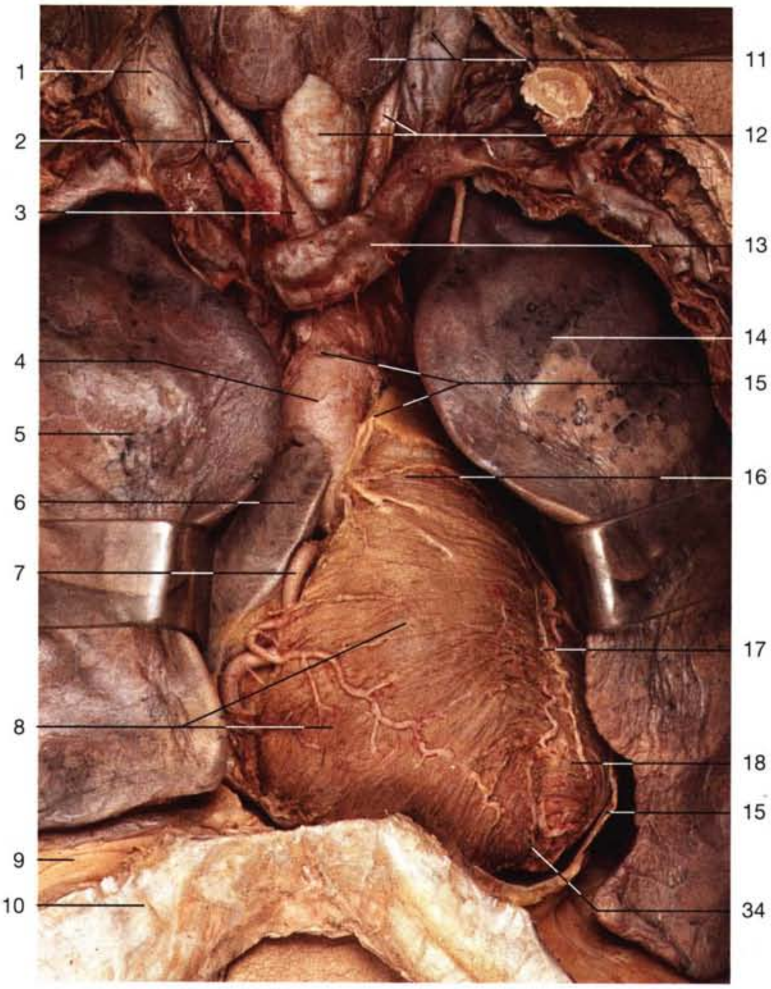

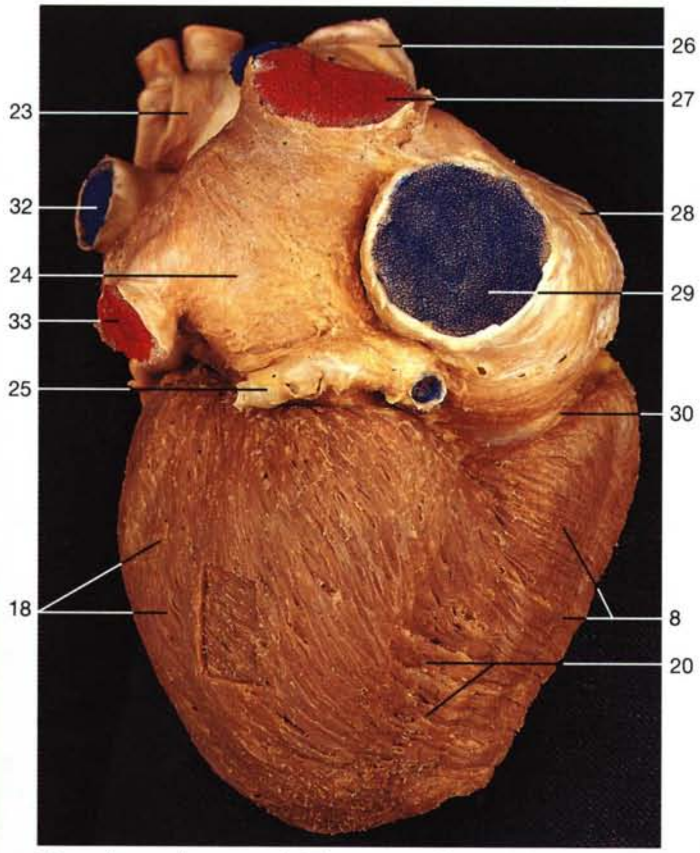

Heart (posterior aspect). The myocardium of the left ventricle has been fenestrated to show the muscle fiber bundles of the deeper layer with their more circular course.

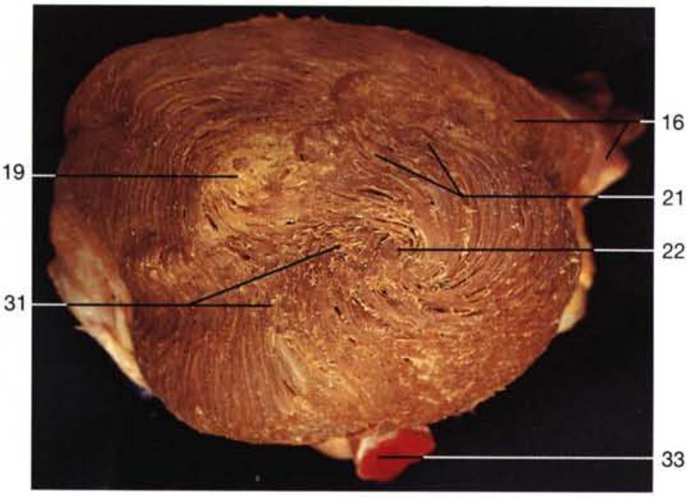

Vortex of cardiac muscle fibers (from below).

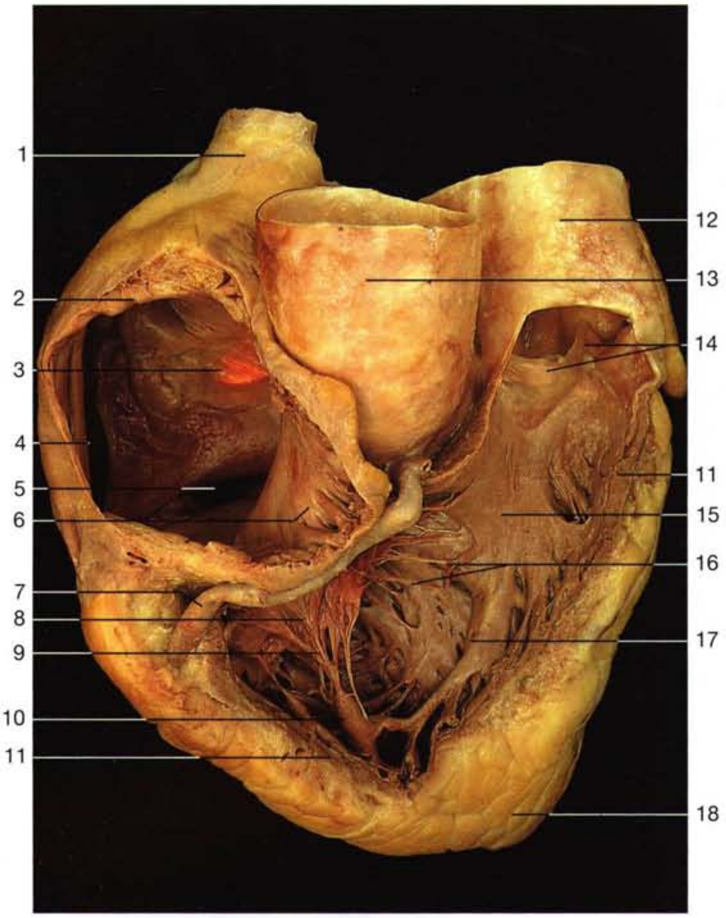

1	Superior vena cava
2	Crista terminalis
3	Fossa ovalis
4	Opening of inferior vena cava
5	Opening of coronary sinus
6	Right auricle
7	Right coronary artery and coronary sulcus
8	**Anterior cusp** of **tricuspid valve**
9	Chordae tendineae
10	**Anterior papillary muscle**
11	Myocardium
12	Pulmonary trunk
13	Ascending aorta
14	**Pulmonic valve**
15	Conus arteriosus (interventricular septum)
16	Septal papillary muscles
17	Septomarginal or moderator band
18	Apex of heart
19	Left auricle
20	**Aortic valve**
21	**Left ventricle**
22	Pulmonary veins
23	Position of fossa ovalis
24	**Left atrium**
25	**Left atrioventricular** (bicuspid or **mitral**) **valve**
26	Coronary sinus
27	Left coronary artery
28	Posterior papillary muscle
29	Left subclavian artery
30	Descending aorta
31	Left pulmonary artery

Right heart (anterior aspect). Anterior wall of right atrium and ventricle removed.

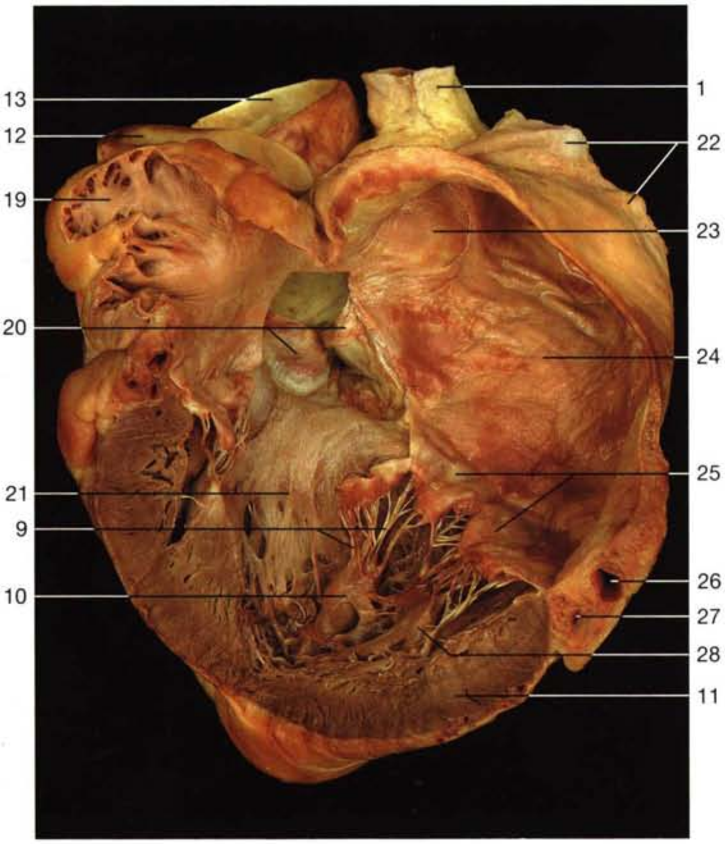

Left heart. Left atrium and ventricle opened. Aortic and mitral valve cut open.

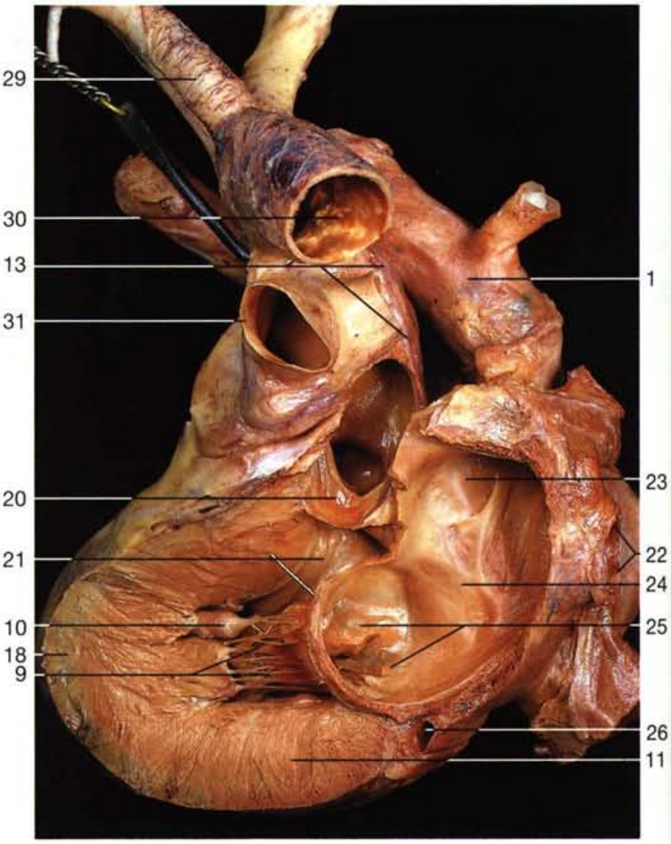

Left heart. Aortic valve cut open. Left atrium and ventricle opened. The atrial wall was cut near the anulus fibrosus of the mitral valve.

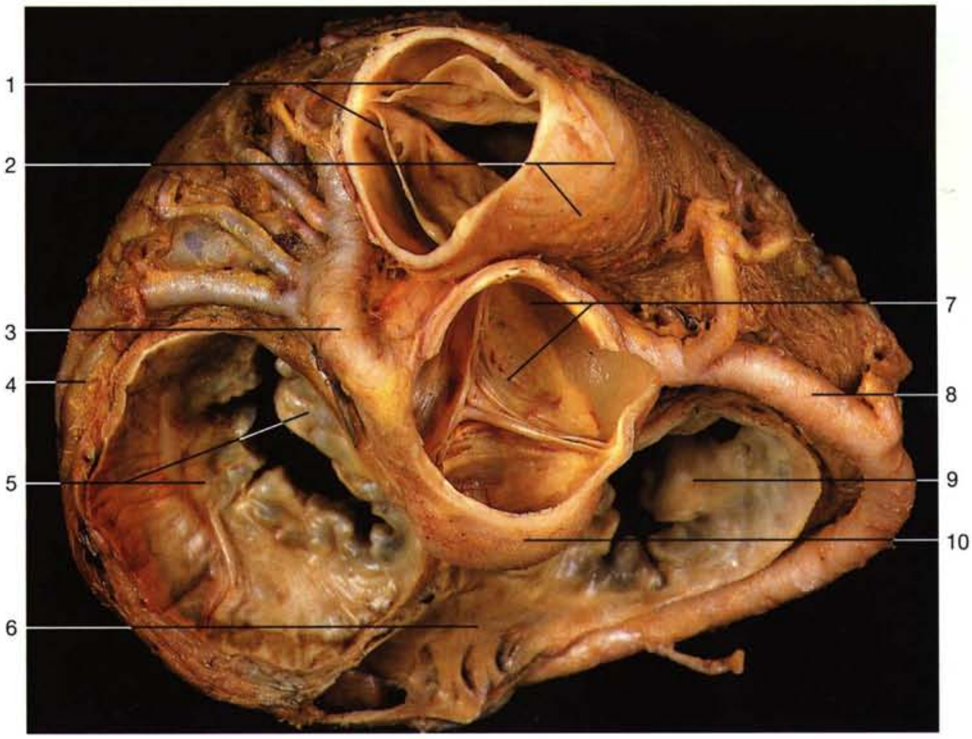

Valves of heart (superior aspect). Left and right atrium removed. Dissection of coronary arteries. Above: anterior wall of the heart.

1 **Pulmonic valve**
2 Sinus of pulmonary trunk
3 **Left coronary artery**
4 Great cardiac vein
5 **Left atrioventricular (mitral) valve**
6 **Coronary sinus**
7 **Aortic valve**
8 **Right coronary artery**
9 **Right atrioventricular (tricuspid) valve**
10 Bulb of aorta
11 Anterior semilunar cusp of pulmonic valve
12 Left semilunar cusp of pulmonic valve
13 Right semilunar cusp of pulmonic valve
14 Left semilunar cusp of aortic valve
15 Right semilunar cusp of aortic valve
16 Posterior semilunar cusp of aortic valve
17 Right atrium
18 Anterior cusp of tricuspid valve
19 **Chordae tendineae**
20 Trabeculae carneae
21 Interventricular septum
22 Septal cusp of tricuspid valve
23 **Anterior papillary muscle**
24 Myocardium of right ventricle

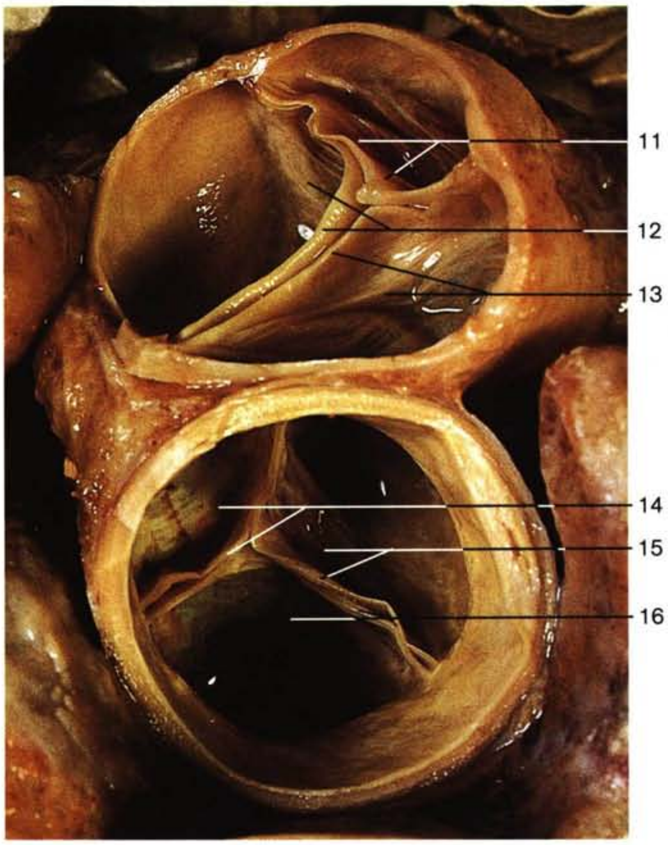

Pulmonic and aortic valves (from above). Anterior wall of the heart at the top. Both valves are closed.

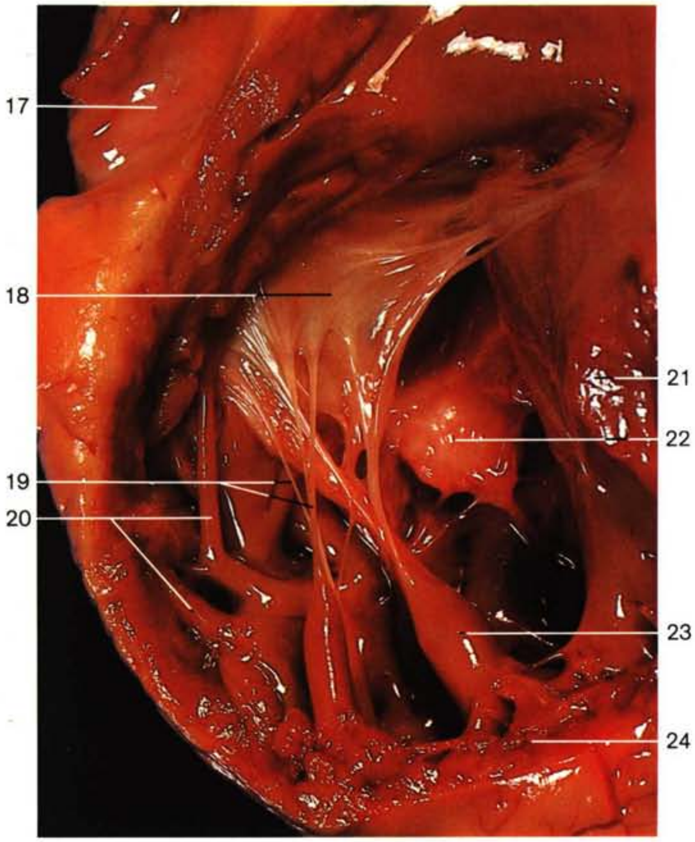

Right atrioventricular (tricuspid) valve (anterior aspect after removal of the anterior wall of the right ventricle).

1 Ascending aorta
2 Superior vena cava
3 Right auricle
4 **Right atrium**
5 Coronary sulcus
6 **Right ventricle**
7 Pulmonary trunk
8 Left auricle
9 Anterior interventricular sulcus
10 **Left ventricle**
11 Right pulmonary artery
12 Sulcus terminalis with sinuatrial node
13 Line indicating plane of position of valves
14 Myocardium of right atrium
15 Inferior vena cava
16 Valve of pulmonary trunk
17 Tricuspid valve
18 Myocardium of right ventricle

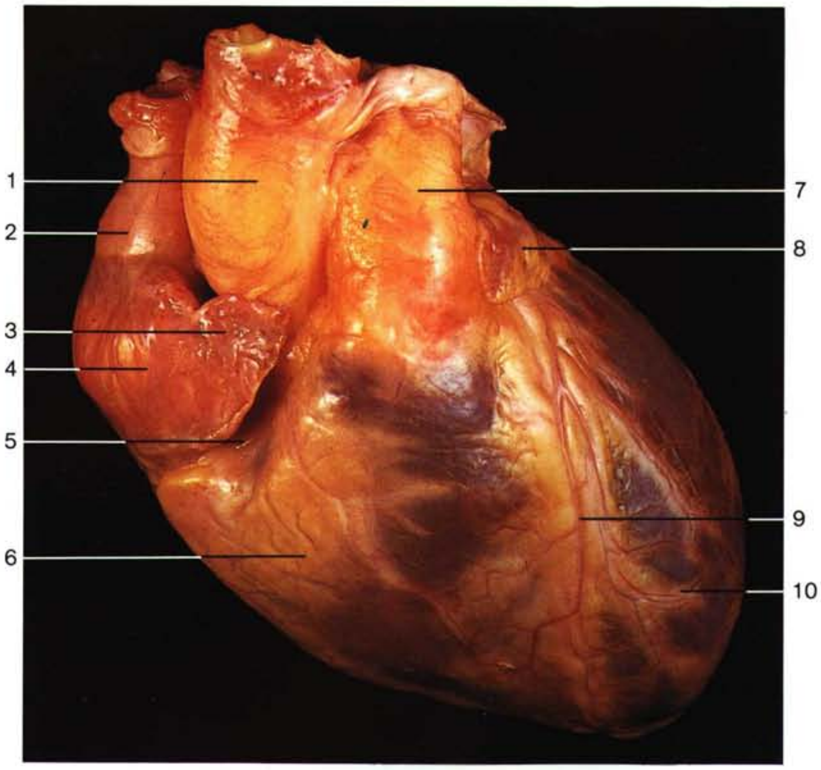

Heart, fixed in **diastole** (anterior aspect). The ventricles are relaxed, atria contracted.

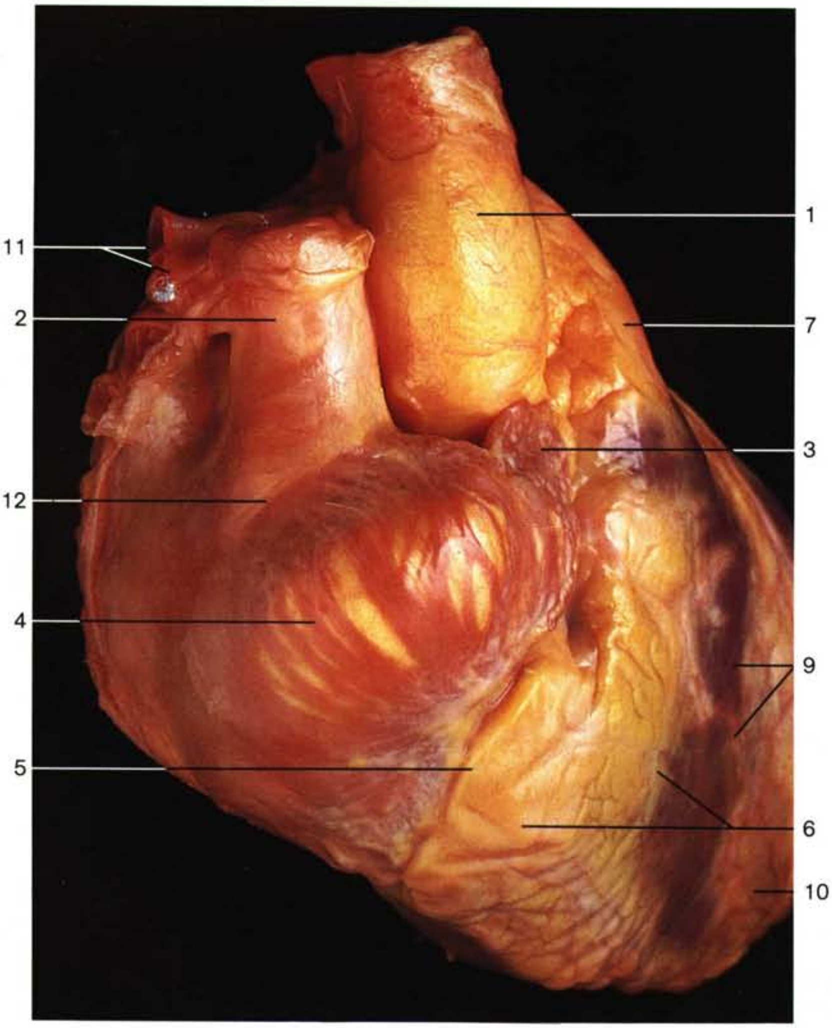

Heart, fixed in **systole** (anterolateral aspect). The ventricles are contracted, atria dilated.

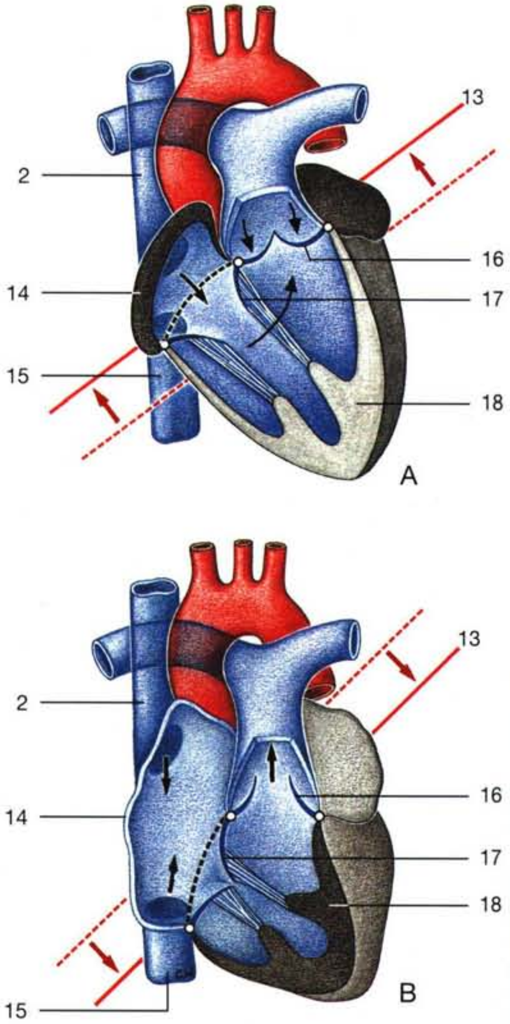

Morphological changes during heart movements. Note the changes in positon of the valves (arrows). Contracted portions of heart are indicated in red.

A. **Diastole,** muscles of the ventricles relaxed; atrioventricular valves open, semilunar valves closed.

B. **Systole,** muscles of ventricles contracted; atrioventricular valves closed, semilunar valves open.

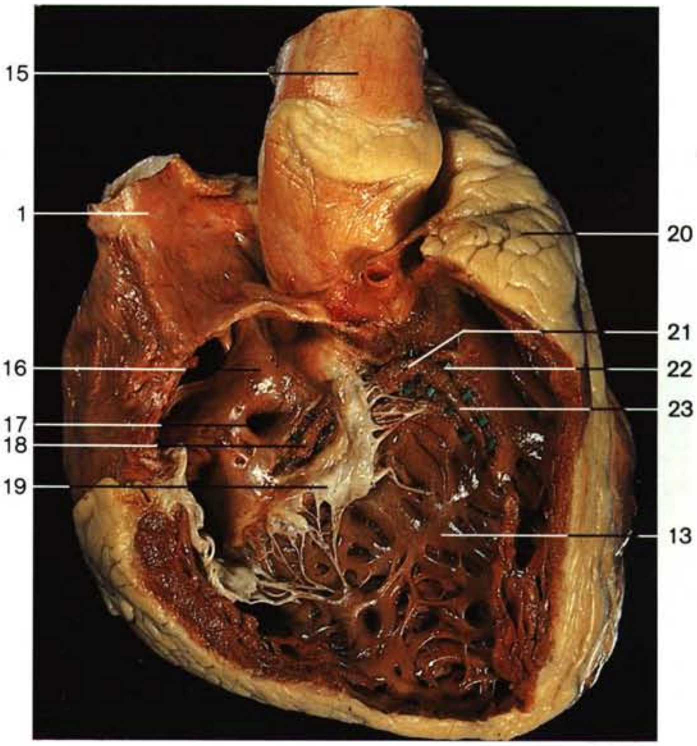

Right ventricle, dissection of **atrioventricular node, atrioventricular bundle (bundle of His)** and right limb or bundle branch (probes).

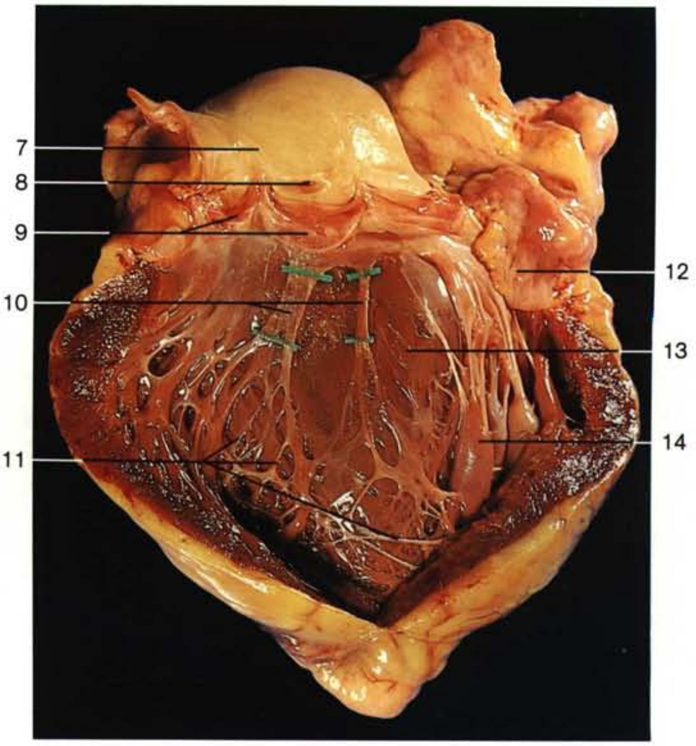

Left ventricle, dissection of the left limb or bundle branch of conducting system (probes).

1 Superior vena cava
2 Sulcus terminalis
3 Bulb of aorta
4 Sinuatrial node (arrows)
5 Muscle fiber bundles of right atrium
6 Coronary sulcus (with right coronary artery)
7 Aortic sinus
8 Entrance to left coronary artery
9 Aortic valve
10 Branches of **left bundle branch**
11 Purkinje fibers
12 Left auricle
13 Interventricular septum
14 Papillary muscles
15 Ascending aorta
16 Right atrium
17 Opening of coronary sinus
18 **Atrioventricular node**
19 Septal cusp of tricuspid valve
20 Pulmonary trunk
21 **Atrioventricular bundle** (bundle of His)
22 Bifurcation of atrioventricular bundle
23 **Right bundle branch**
24 Inferior vena cava
25 Left atrium
26 Left bundle branch
27 Papillary muscles with Purkinje fibers

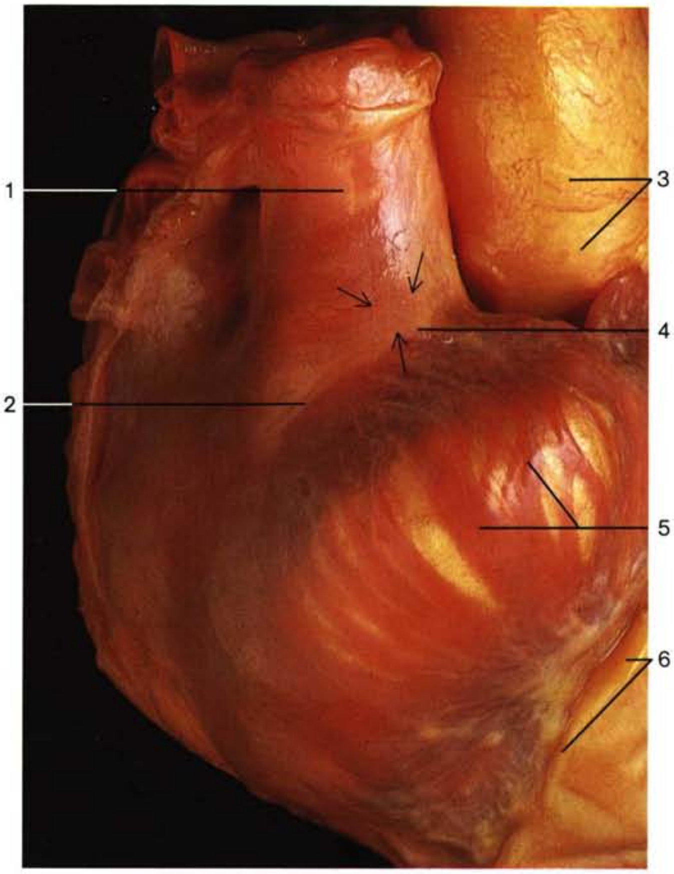

Right atrium, anterior wall, showing the location of the **sinuatrial node** (arrows).

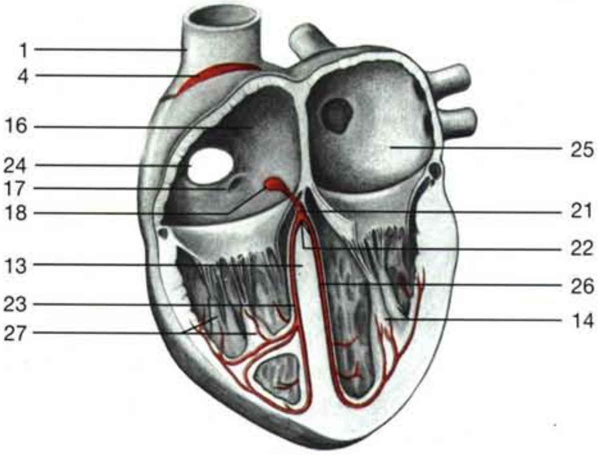

Conducting system of the heart.
(Schematic drawing.)

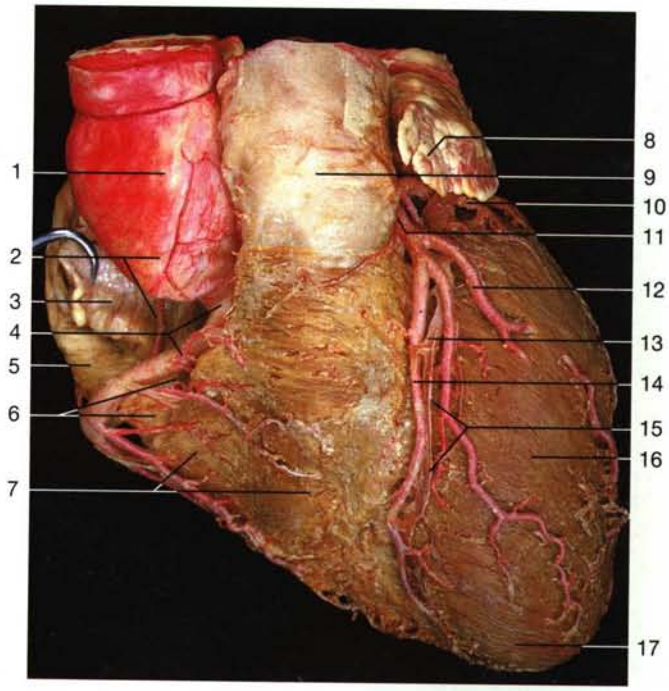

Coronary arteries (anterior aspect). The epicardium and subepicardial fatty tissue have been removed. The arteries have been injected with red resin from the aorta.

1 Ascending aorta
2 Aortic bulb and (in the above specimen) sinuatrial branch of right coronary artery
3 Right auricle
4 **Right coronary artery**
5 Right atrium
6 Coronary sulcus
7 Right ventricle
8 Left auricle
9 Pulmonary trunk

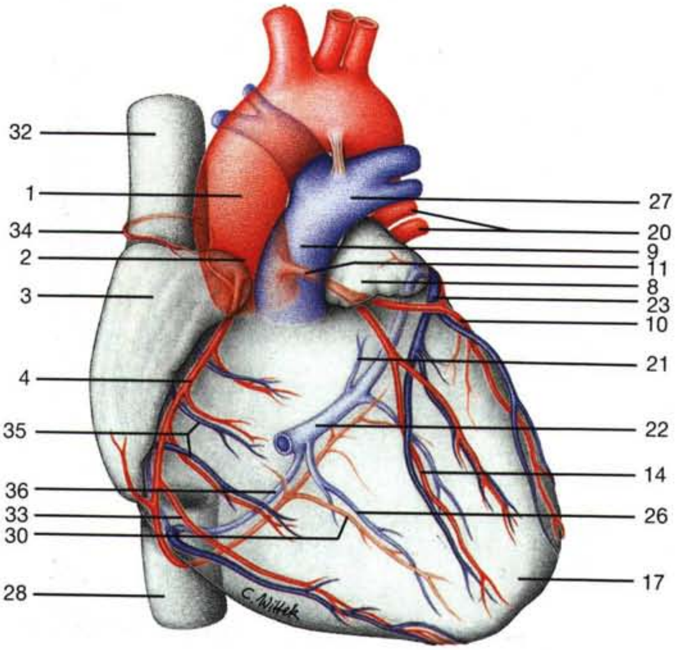

Vessels of the heart. Coronary arteries (red) and veins (blue) of the heart (anterior aspect).

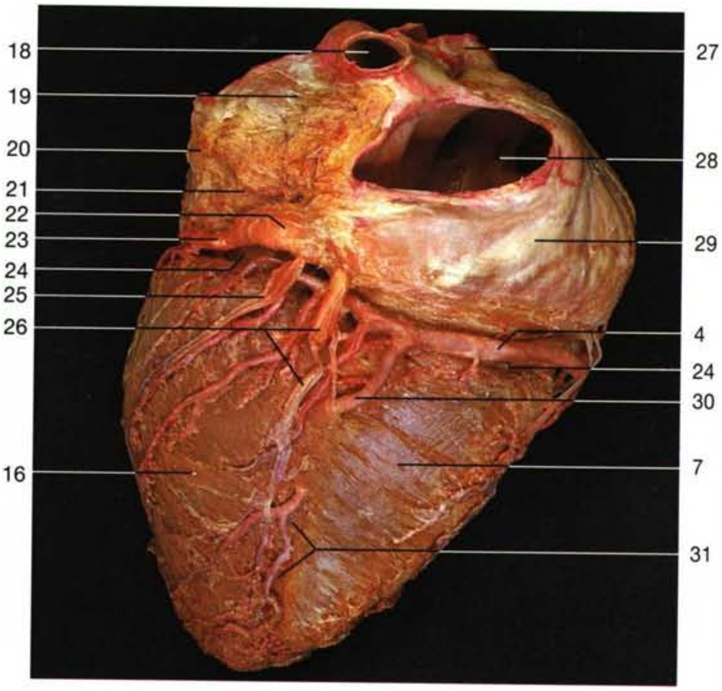

Right coronary artery and veins of the heart (dorsal aspect). The epicardium and subepicardial fatty tissue have been removed.

10 Circumflex branch of left coronary artery
11 **Left coronary artery**
12 Diagonal branch of left artery
13 Great cardiac vein
14 Anterior interventricular artery
15 Anterior interventricular sulcus
16 Left ventricle
17 Apex of heart
18 Right pulmonary vein
19 Left atrium
20 Left pulmonary veins
21 Oblique vein of left atrium (Marshall's vein)
22 **Coronary sinus**
23 **Great cardiac vein**
24 Coronary sulcus (posterior portion)
25 Posterior vein of left ventricle
26 **Middle cardiac vein**
27 Left pulmonary artery
28 Inferior vena cava
29 Right atrium
30 Posterior interventricular branch of right coronary artery
31 Posterior interventricular sulcus
32 Superior vena cava
33 Right marginal branch
34 Branch of sinuatrial node
35 Minimal cardiac veins
36 **Small cardiac vein**

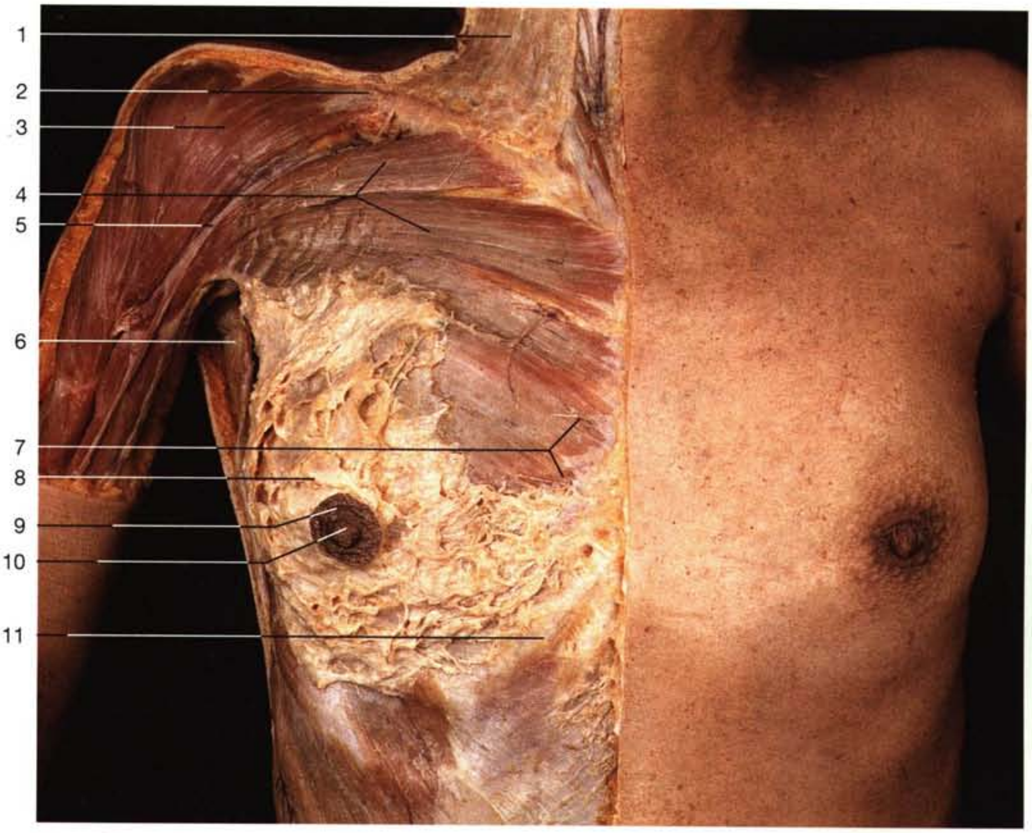

Dissection of mammary gland (anterior aspect).

1 Platysma muscle
2 Clavicle
3 Deltoid muscle
4 Pectoralis major muscle
5 Deltopectoral groove and cephalic vein
6 Latissimus dorsi muscle
7 Medial mamarian branches of intercostal nerves
8 Breast tissue
9 Areola
10 Nipple (papilla)
11 Costal margin
12 Pectoral fascia
13 Mammary gland
14 Serratus anterior muscle
15 Lactiferous sinus
16 Lactiferous ducts

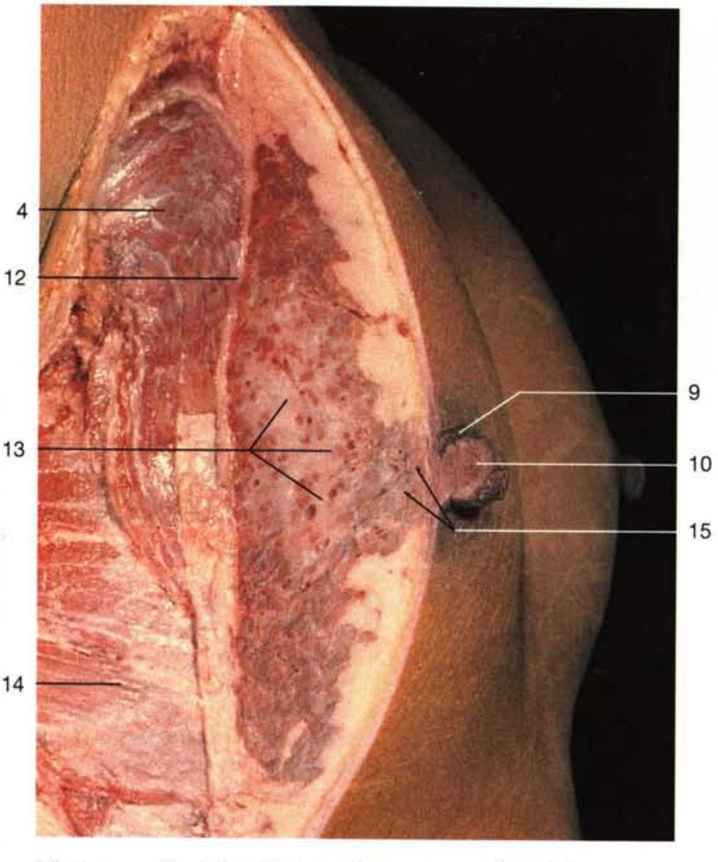

Mammary gland (sagittal section; pregnant female).

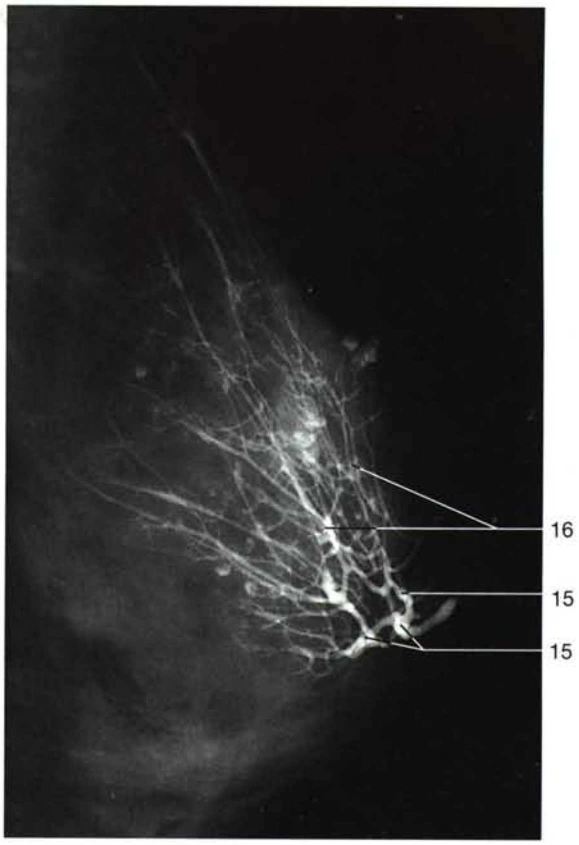

Galactogram of mammary gland (adult female). Notice multiple duct cysts.

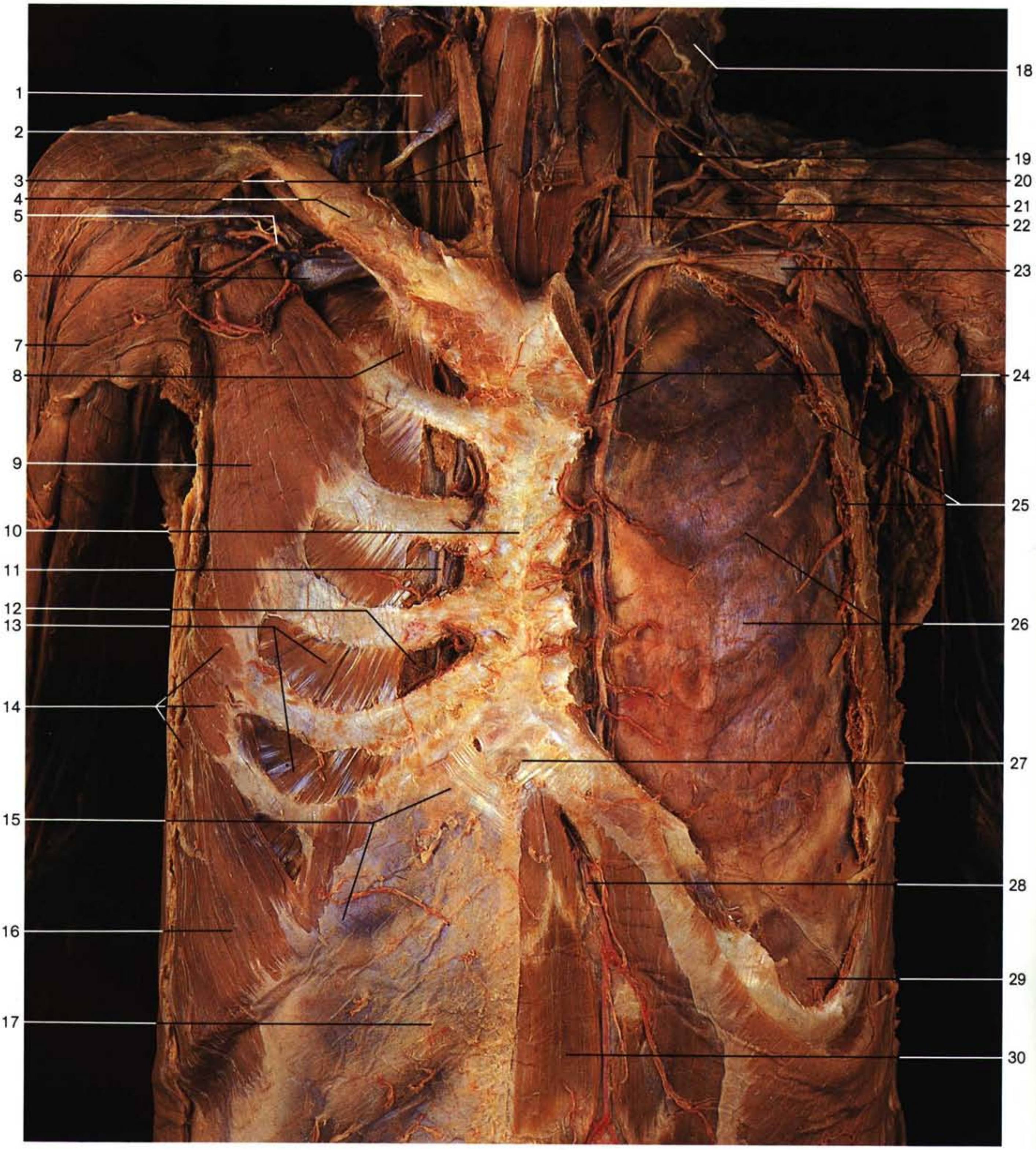

Thoracic wall and organs (ventral aspect). The left clavicle and ribs have been partially removed, and the right intercostal spaces have been opened to show the internal thoracic vein and artery.

1 Right internal jugular vein
2 Omohyoid muscle
3 Sternohyoid muscle and external jugular vein
4 Clavicle
5 Thoracoacromial artery
6 Right **subclavian vein**
7 Pectoralis major muscle
8 External intercostal muscle
9 Pectoralis minor muscle
10 Body of sternum
11 Right internal thoracic artery and vein
12 Fascicles of transversus thoracis muscle
13 **Internal intercostal muscles**
14 Serratus anterior muscle
15 Costal margin
16 External abdominal oblique muscle
17 Anterior sheath of rectus abdominis muscle
18 Sternocleidomastoid muscle
19 Left internal jugular vein
20 Transverse cervical artery
21 Brachial plexus
22 Vagus nerve
23 Left axillary vein
24 Left **internal thoracic artery and vein**
25 Ribs and thoracic wall (cut)
26 Costal pleura
27 Xiphoid process
28 **Superior epigastric artery**
29 Diaphragm
30 Rectus abdominis muscle

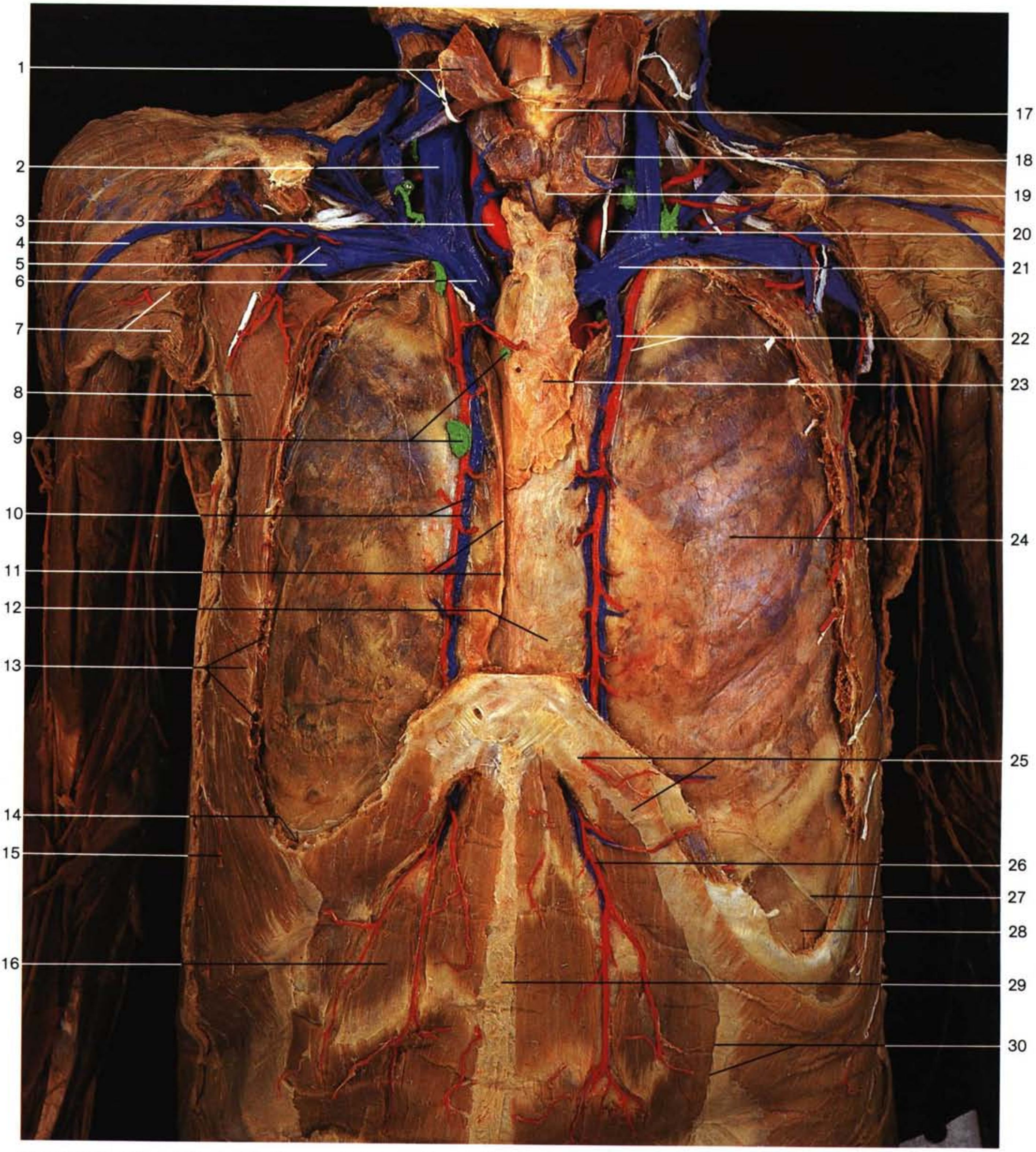

Thoracic organs, anterior mediastinum and pleura. Ribs, clavicle and sternum have been partly removed.
Red = arteries; blue = veins; green = lymph vessels and nodes.

1 Sternothyroid muscle and its nerve (a branch of the ansa cervicalis)
2 Right **internal jugular vein**
3 Right common carotid artery
4 Cephalic vein
5 Right **subclavian vein**
6 **Right brachiocephalic vein**
7 Pectoralis major muscle (divided)
8 Pectoralis minor muscle (divided)
9 Parasternal lymph nodes
10 Internal thoracic artery and vein
11 Anterior margin of costal pleura
12 **Pericardium**
13 5th and 6th ribs (divided) and serratus anterior muscle
14 **Costodiaphragmatic recess**
15 External abdominal oblique muscle
16 Rectus abdominis muscle
17 Larynx (thyroid cartilage)
18 Thyroid gland
19 Trachea
20 Left vagus nerve
21 Left brachiocephalic vein
22 Left **internal thoracic artery and vein**
23 Thymus
24 **Costal pleura**
25 Costal margin
26 **Superior epigastric artery**
27 Margin of costal pleura
28 **Diaphragm**
29 **Linea alba**
30 Cut edge of anterior sheath of rectus abdominis muscle

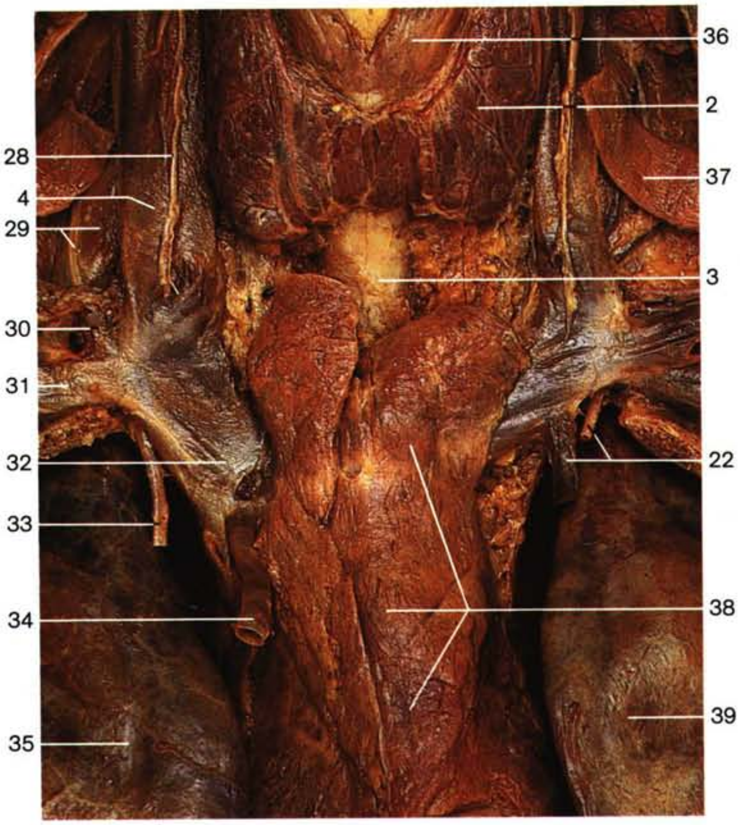

The **thymus** above the heart, showing its position and size.

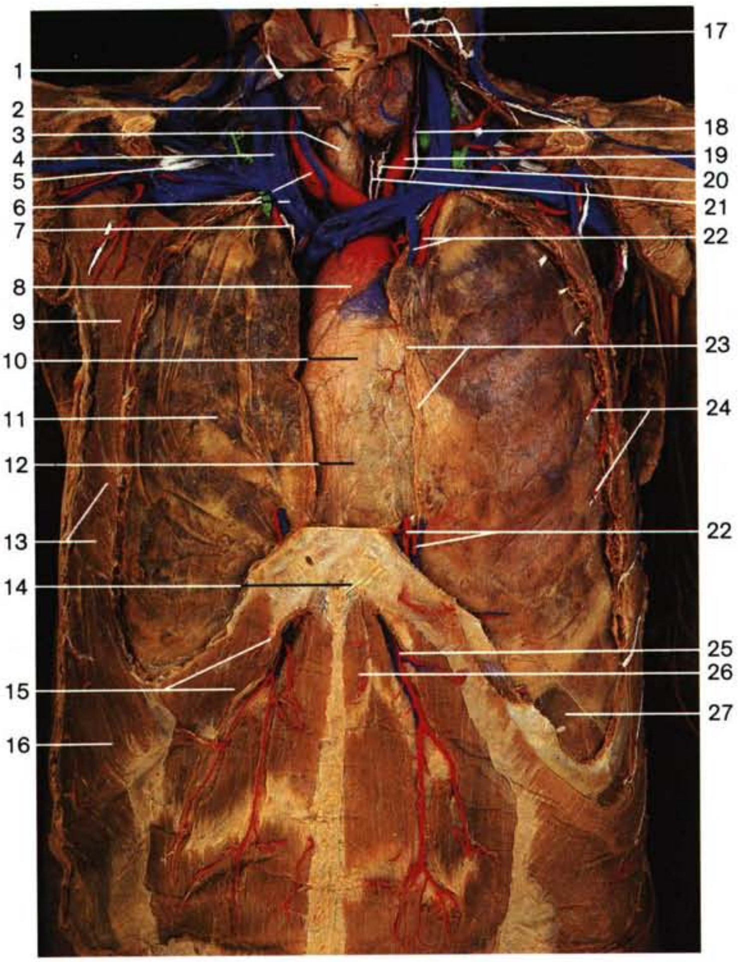

Thoracic organs (ventral aspect). The internal thoracic vessels have been removed, and the anterior margins of the pleura and lungs have been slightly reflected to display the **anterior** and **middle mediastinum,** including the **heart** and **great vessels.**

1 Larynx (thyroid cartilage)
2 Thyroid gland
3 **Trachea**
4 Internal jugular vein
5 Brachial plexus
6 Right brachiocephalic vein and common carotid artery
7 Right phrenic nerve
8 **Ascending aorta**
9 Pectoralis minor muscle (divided)
10 Pulmonary trunk (covered by pericardium)
11 **Costal pleura**
12 Pericardium and **heart**
13 Serratus anterior muscle
14 Xiphoid process
15 Costal margin
16 External abdominal oblique muslce
17 Sternothyroid muscle (divided and reflected)
18 Vagus nerve
19 Left common carotid artery
20 Left sympathetic trunk
21 Left recurrent laryngeal nerve
22 Left internal thoracic artery and vein (divided)
23 Margin of costal pleura
24 Intercostal nerves and vessels
25 Superior epigastric artery
26 Rectus abdominis muscle
27 **Diaphragm**
28 Ansa cervicalis
29 Phrenic nerve and scalenus anterior muscle
30 External jugular vein (divided)
31 Right subclavian vein
32 Right brachiocephalic vein
33 Internal thoracic artery (divided)
34 Internal thoracic vein (divided)
35 Right lung
36 Cricothyroid muscle
37 Omohyoid muscle
38 **Thymus**
39 Left lung

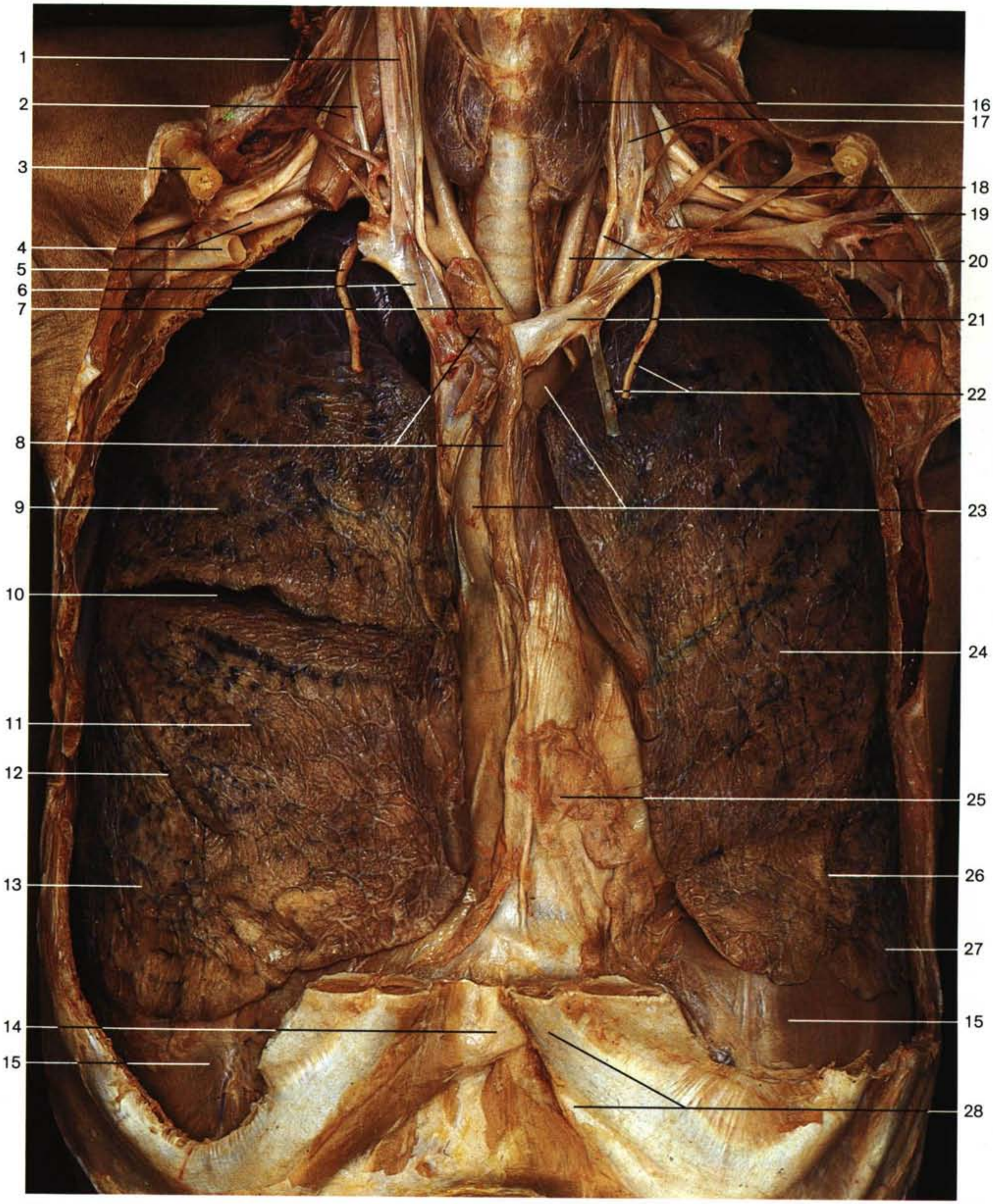

Thoracic organs (ventral aspect). The pleura has been opened and the lungs exposed. Remnants of the thymus and pericardium are seen.

1	Right internal jugular vein	11	Middle lobe of right lung	20	Left common carotid artery and vagus nerve
2	Phrenic nerve and scalenus anterior muscle	12	Oblique fissure of right **lung**	21	**Left brachiocephalic vein**
3	Clavicle (divided)	13	Lower lobe of right lung	22	Internal thoracic artery and vein (divided)
4	Right subclavian artery and vein	14	Xiphoid process	23	**Ascending aorta** and aortic arch
5	Internal thoracic artery	15	**Diaphragm**	24	Upper lobe of left **lung**
6	**Right brachiocephalic vein**	16	**Thyroid gland**	25	**Pericardium**
7	Brachiocephalic trunk	17	Left internal jugular vein	26	Oblique fissure of left lung
8	Thymus (atrophic)	18	Brachial plexus	27	Lower lobe of left lung
9	Upper lobe of right **lung**	19	Left cephalic vein	28	Costal margin
10	Horizontal fissure of right lung (incomplete)				

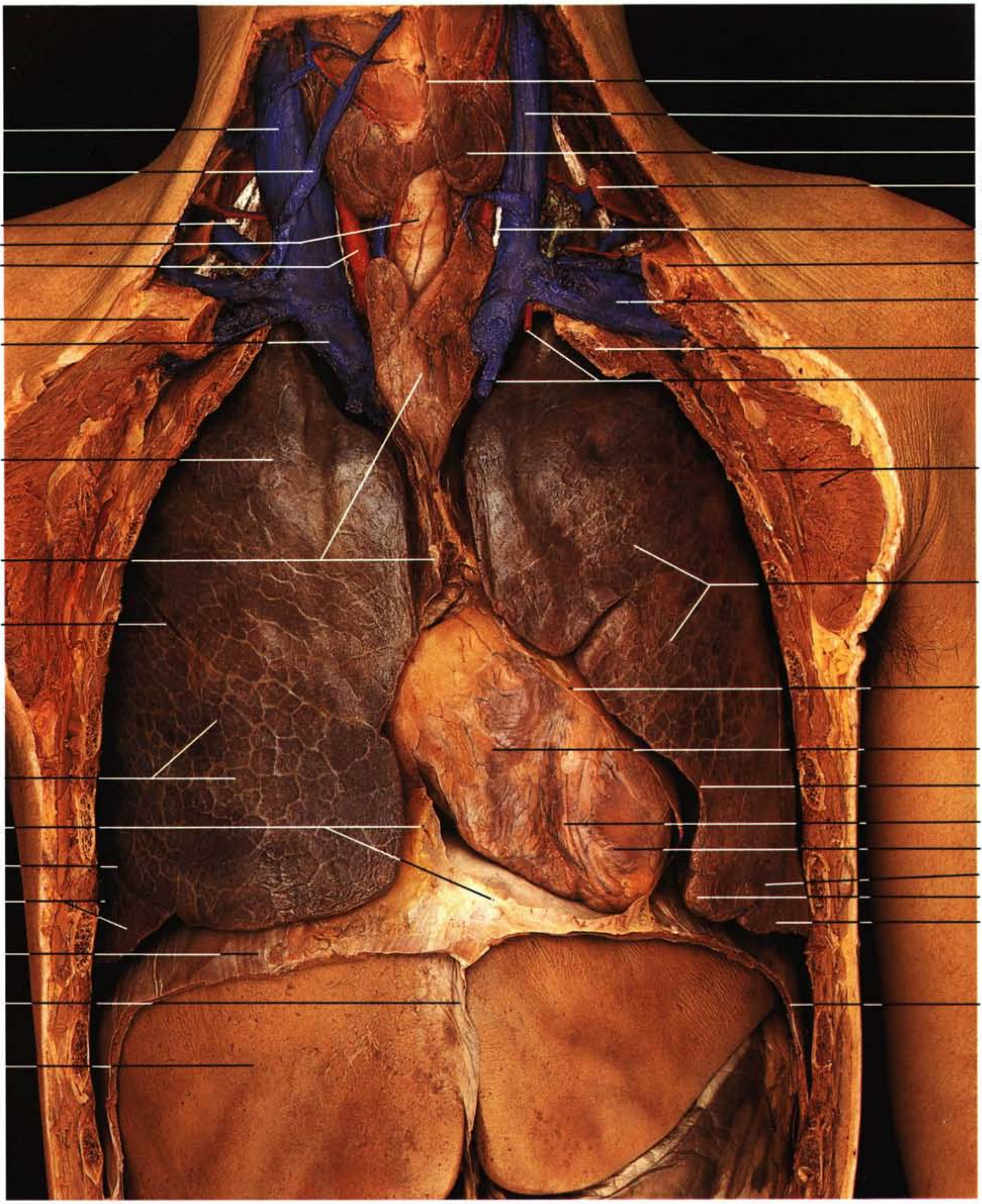

Thoracic organs (ventral aspect). The thoracic wall, costal pleura, pericardium and diaphragm have been partly removed.

1 Internal jugular vein
2 External jugular vein (displaced medially)
3 Brachial plexus
4 Trachea
5 Right common carotid artery
6 Clavicle (divided)
7 Right brachiocephalic vein
8 Upper lobe of right lung
9 **Thymus** (atrophic)
10 Horizontal fissure of right lung
11 Middle lobe of right lung

12 Pericardium (cut edges)
13 Oblique fissure of lung
14 Lower lobe of right lung
15 **Diaphragm**
16 Falciform ligament
17 **Liver**
18 Location of larynx
19 Left internal jugular vein
20 Thyroid gland
21 Omohyoid muscle (divided)
22 Vagus nerve

23 Left subclavian vein
24 1st rib (divided)
25 Internal thoracic artery and vein
26 Pectoralis major and pectoralis minor muscles (cut edges)
27 **Upper lobe of left lung**
28 **Right ventricle**
29 Cardiac notch of left lung
30 Interventricular sulcus of heart
31 **Left ventricle**
32 Lingula
33 Lower lobe of left lung

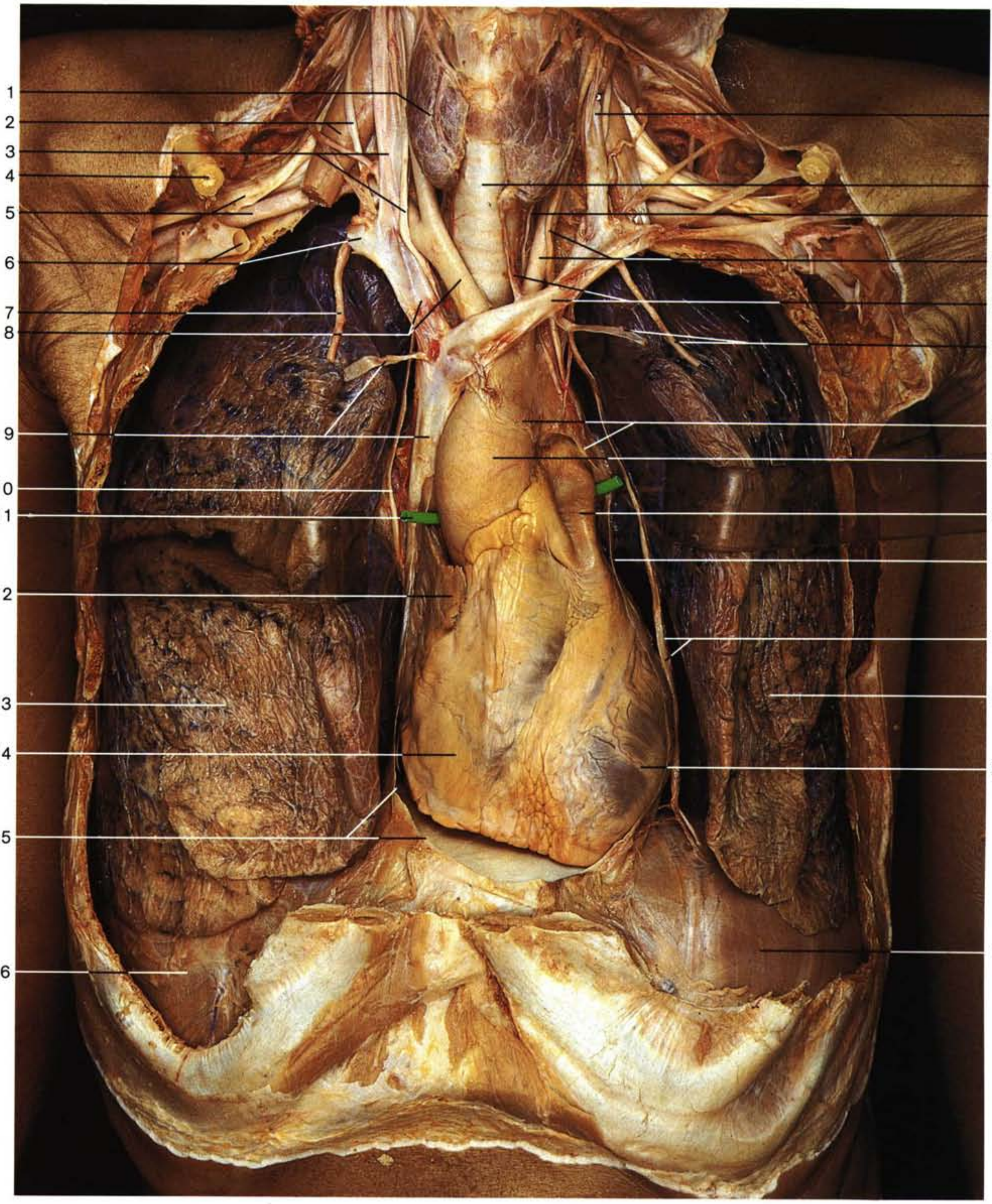

Thoracic organs, position of the heart, middle mediastinum (ventral aspect). The anterior wall of the thorax, the costal pleura and the pericardium have been removed and the lungs slightly reflected.

1 Thyroid gland
2 Phrenic nerve and scalenus anterior muscle
3 Vagus nerve and internal jugular vein
4 Clavicle (divided)
5 Brachial plexus and subclavian artery
6 Subclavian vein
7 Internal thoracic artery
8 Brachiocephalic trunk and right brachiocephalic vein
9 **Superior vena cava** and thymic vein
10 **Right phrenic nerve**
11 Transverse pericardial sinus (probe)
12 Right auricle
13 Middle lobe of right lung
14 **Right ventricle**
15 Cut edge of pericardium
16 Diaphragm
17 Internal jugular vein
18 Trachea
19 Left recurrent laryngeal nerve
20 Left common carotid artery and vagus nerve
21 Left brachiocephalic vein and inferior thyroid vein
22 Left internal thoracic artery and vein (divided)
23 Upper margin of pericardial sac
24 **Ascending aorta**
25 Pulmonary trunk
26 **Left phrenic nerve** and left pericardiacophrenic artery and vein
27 Upper lobe of left lung
28 Left ventricle

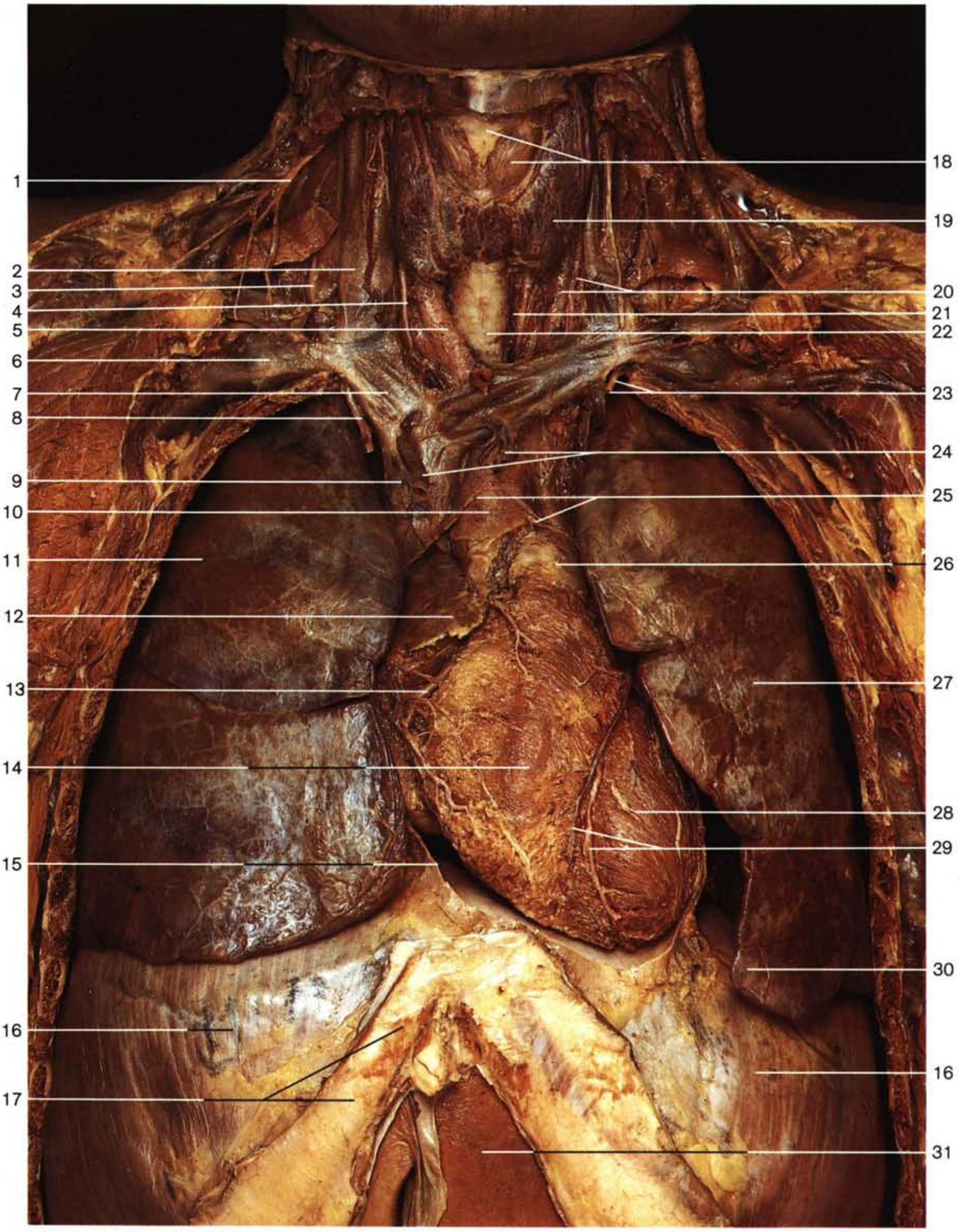

Thoracic organs, position of heart, dissection of coronary vessels in situ (ventral aspect). The anterior wall of thorax, costal pleura and pericardium have been removed.

1 Intermediate supraclavicular nerve
2 Internal jugular vein
3 Right phrenic nerve
4 Right vagus nerve
5 Right common carotid artery
6 Right subclavian vein
7 Right brachiocephalic vein
8 Right internal thoracic artery
9 **Superior vena cava**
10 Ascending aorta
11 Right lung
12 Right atrium
13 **Right coronary artery** and small cardiac vein
14 **Right ventricle**
15 Cut edge of pericardium
17 Costal margin
18 Larynx (cricothyroid muscle and thyroid cartilage)
19 Thyroid gland
20 Left common carotid artery and left vagus nerve
21 Left recurrent laryngeal nerve
22 Trachea
23 Left internal thoracic artery and vein (divided)
24 Thymic veins
25 Margin of pericardial sac
26 Pulmonary trunk
27 Left lung
28 Left ventricle
29 Anterior interventricular artery and vein
30 Lingula
31 Liver

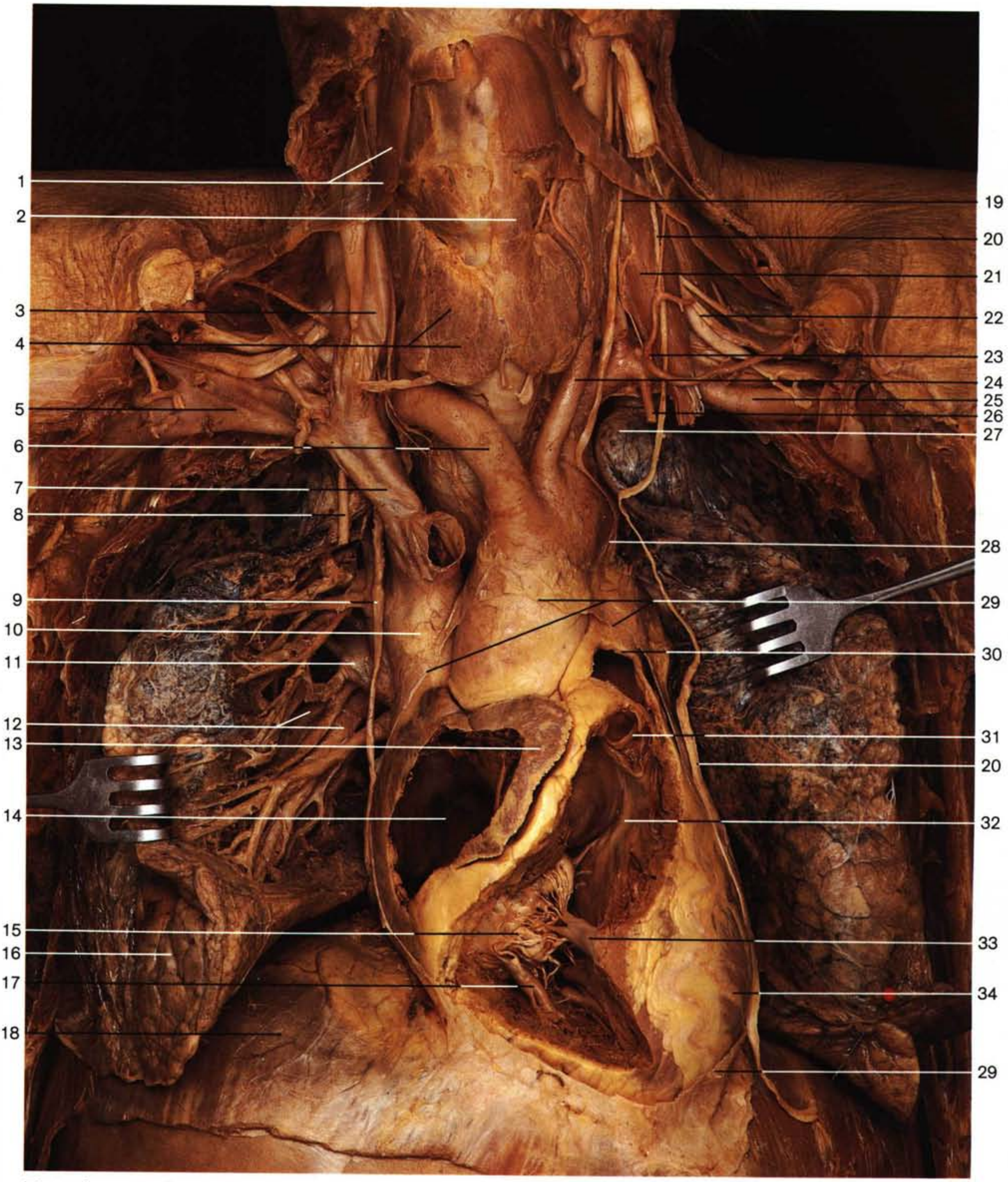

Thoracic organs, heart with valves in situ (ventral aspect). Anterior wall of thorax, pleura and anterior portion of pericardium have been removed. The right atrium and ventricle have been opened to show the right atrioventricular and pulmonary valves.

1 Omohyoid muscle
2 Pyramidal lobe of thyroid gland
3 Internal jugular vein
4 Thyroid gland
5 Right subclavian vein
6 Brachiocephalic trunk
7 **Right brachiocephalic vein**
8 Right internal thoracic artery
9 Right phrenic nerve
10 **Superior vena cava**
11 Pulmonary vein
12 Branches of pulmonary artery
13 Right auricle
14 **Right atrium**
15 **Right atrioventricular (tricuspid) valve**
16 Right lung
17 Posterior papillary muscle
18 Diaphragm
19 Left **vagus nerve**
20 Left **phrenic nerve**
21 Scalenus anterior muscle
22 Brachial plexus
23 Thyrocervical trunk
24 Left common carotid artery
25 Left subclavian artery
26 Left internal thoracic artery
27 Apex of left lung
28 Left recurrent laryngeal nerve
29 Cut edge of **pericardium**
30 Pulmonary trunk (fenestrated)
31 Pulmonic valve
32 Supraventricular crest
33 **Anterior papillary muscle**
34 **Left ventricle**

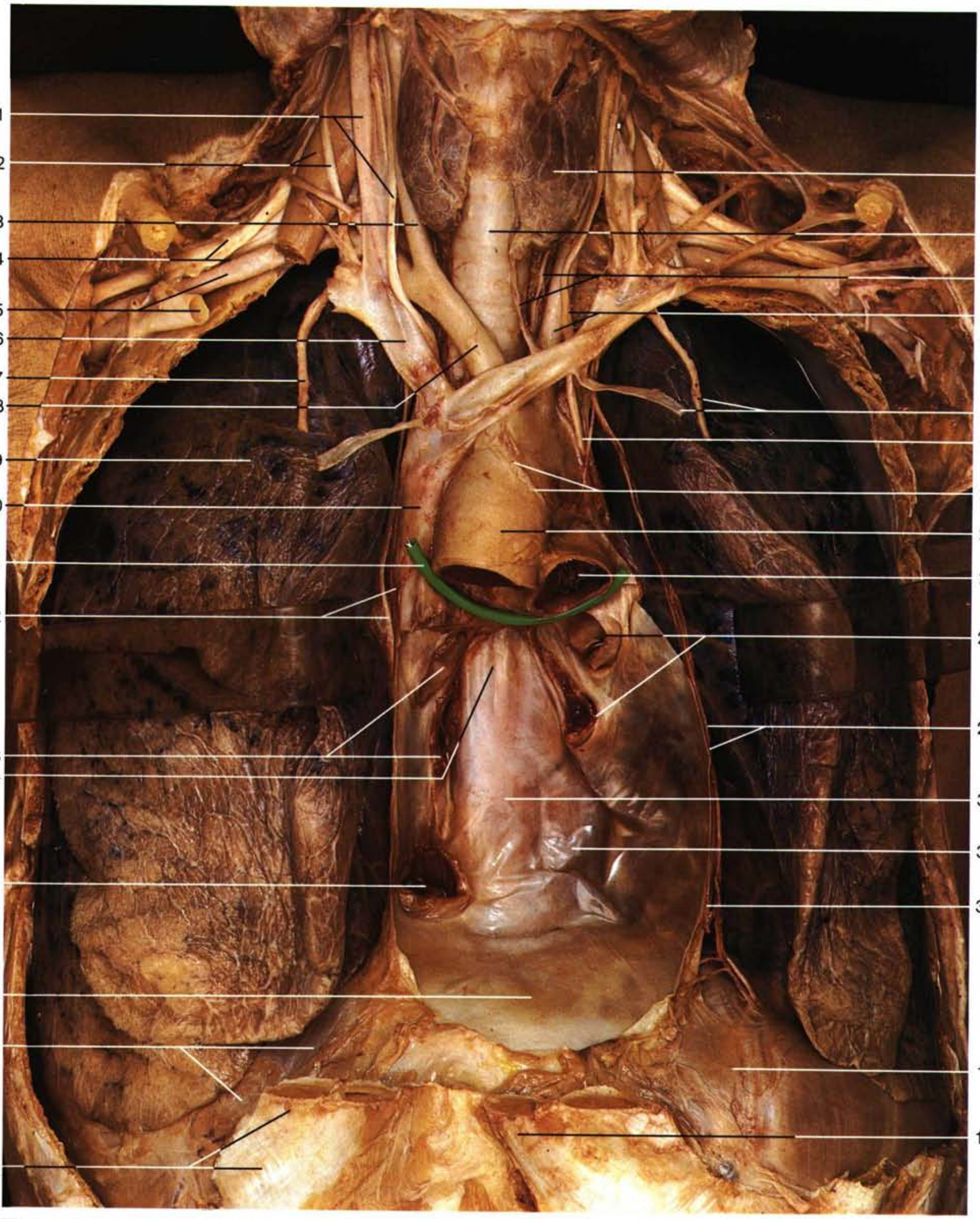

Thoracic organs, pericardium and **mediastinum** (ventral aspect). Anterior wall of thorax and heart have been removed and the lungs slightly reflected. Note probe within transverse pericardial sinus.

1 Right internal jugular vein and right vagus nerve
2 Right phrenic nerve and scalenus anterior muscle
3 Right common carotid artery
4 Brachial plexus
5 Right subclavian artery and vein
6 Right brachiocephalic vein
7 Right internal thoracic artery (divided)
8 **Brachiocephalic trunk**
9 Upper lobe of right lung
10 **Superior vena cava**
11 Transverse pericardial sinus (probe)
12 Right **phrenic nerve** and right pericardiacophrenic artery and vein
13 Right pulmonary veins
14 Oblique sinus of pericardium
15 **Inferior vena cava**
16 Diaphragmatic part of pericardium
17 Diaphragm
18 Costal margin
19 Thyroid gland
20 **Trachea**
21 Left recurrent laryngeal nerve and inferior thyroid vein
22 Left common carotid artery and left vagus nerve
23 Left internal thoracic artery and vein (divided)
24 **Vagus nerve** at aortic arch
25 Cut edge of pericardium
26 **Ascending aorta**
27 Pulmonary trunk (divided)
28 Left pulmonary veins
29 Left phrenic nerve and left pericardiacophrenic artery and vein
30 Contour of esophagus beneath pericardium
31 Contour of aorta beneath pericardium
32 **Pericardium** (cut edge)

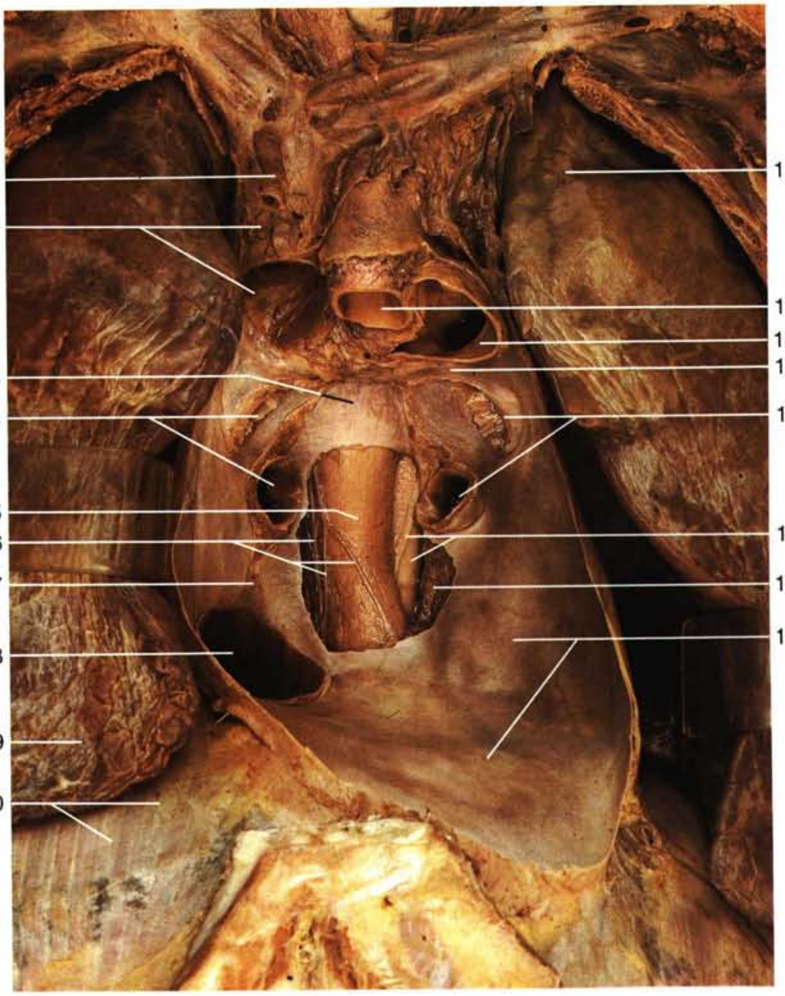
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

Pericardial sac (ventral aspect). The heart has been removed, and the posterior wall of the pericardium has been opened to show the adjacent esophagus and aorta.

1. Internal thoracic vein
2. Superior vena cava
3. Oblique sinus of pericardium
4. Right pulmonary veins
5. Esophagus
6. Branches of right vagus nerve
7. Mesocardium
8. Inferior vena cava
9. Middle lobe of right lung
10. Diaphragm
11. Upper lobe of left lung
12. Ascending aorta
13. **Pulmonary trunk**
14. Transverse pericardial sinus
15. Left pulmonary veins
16. Descending aorta and left vagus nerve
17. Left lung (adjacent to pericardium)
18. **Pericardium**
19. Left subclavian artery
20. Vagus nerve
21. Left recurrent laryngeal nerve
22. Descending aorta
23. Pulmonary artery
24. **Left atrium**
25. **Left ventricle**
26. Coronary sinus
27. Left common carotid artery
28. Brachiocephalic trunk
29. Azygos arch
30. **Right atrium**
31. **Right ventricle**
32. Aortic arch

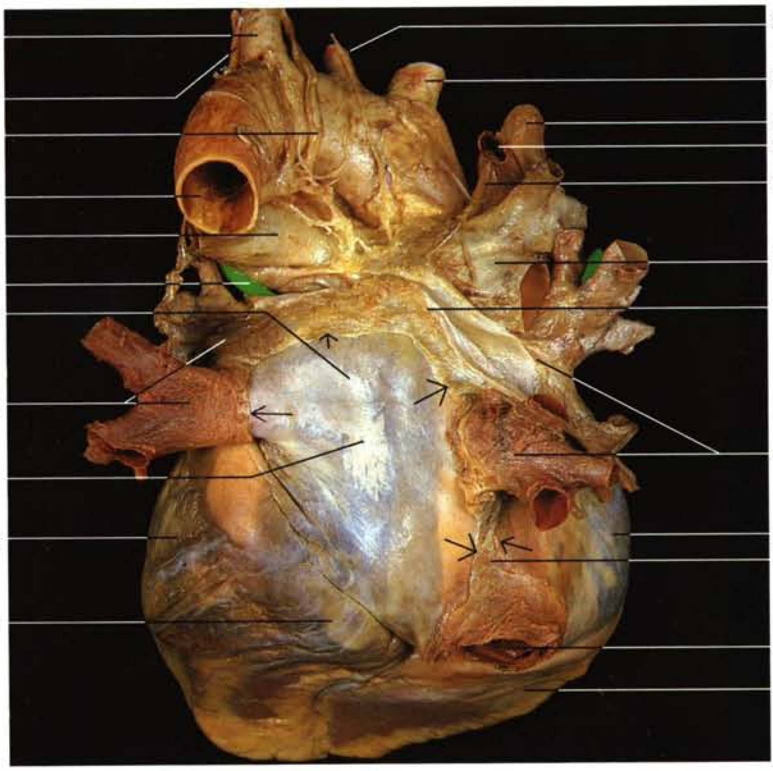
19
20
21
22
23
14
3
15
24
25
26
27
28
2
29
20
23
18
4
30
7
8
31

Heart with epicardium (posterior aspect). Arrows: oblique sinus.

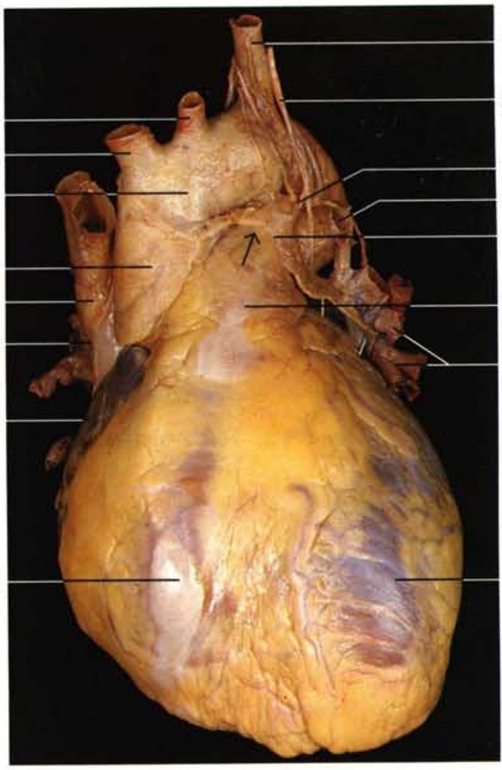
27
28
32
12
2
18
30
31
19
20
21
23
18
13
15
25

Heart with epicardium (anterior aspect). Arrow: pericardial reflection.

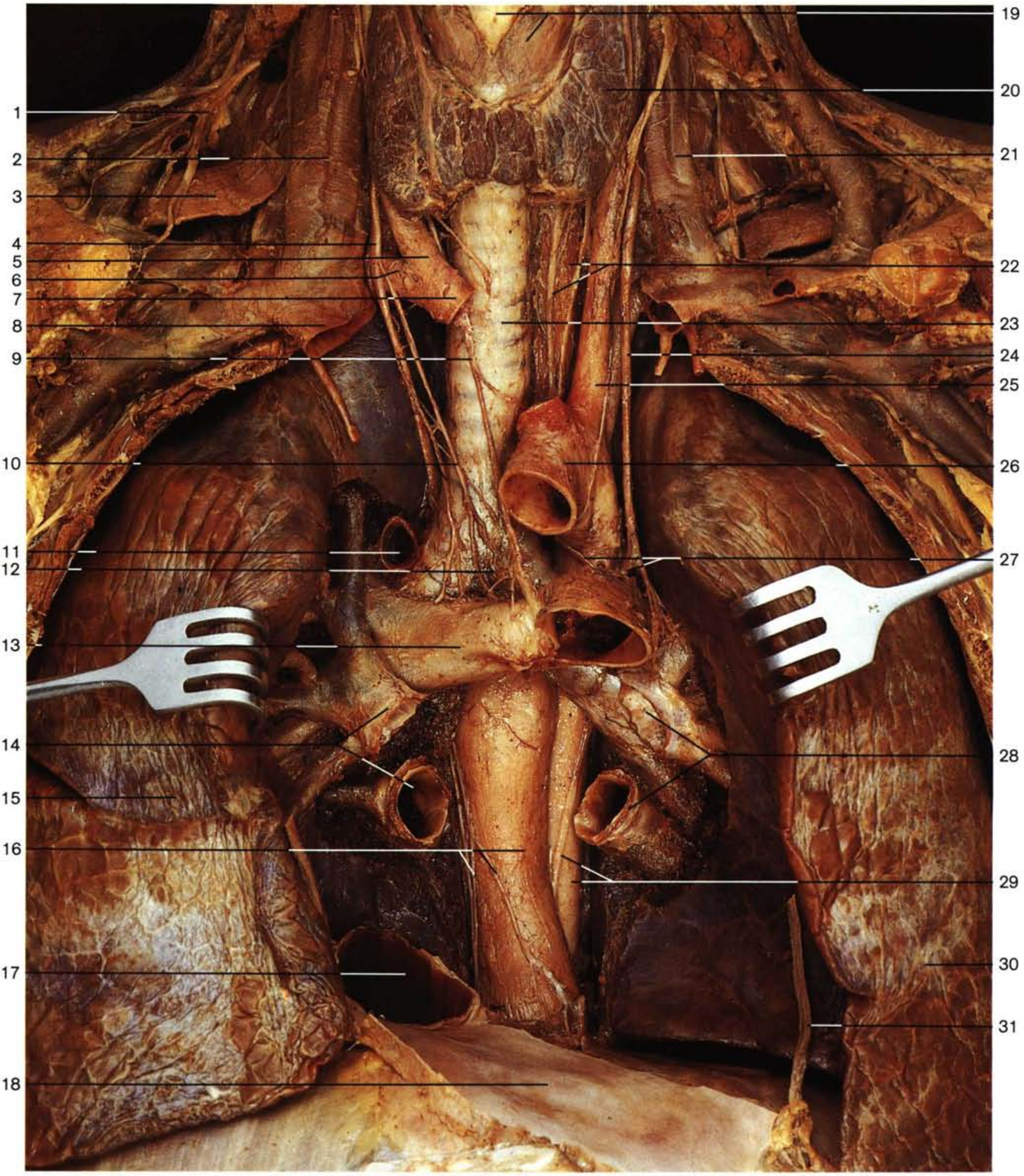

Mediastinal organs after removal of heart and pericardium (ventral aspect). Both lungs have been slightly reflected.

1 Supraclavicular nerves
2 Internal jugular vein
3 Omohyoid muscle
4 **Right vagus nerve**
5 Right common carotid artery
6 Right subclavian artery
7 Brachiocephalic trunk
8 Right brachiocephalic vein
9 Superior cervical cardiac branch of vagus nerve
10 Inferior cervical cardiac branches of vagus nerve
11 Azygos arch (divided)
12 Bifurcation of trachea
13 Right **pulmonary artery**
14 Right **pumonary veins**
15 Right lung
16 **Esophagus** and branches of right vagus nerve
17 Inferior vena cava
18 Pericardium
19 Larynx (thyroid cartilage, cricothyroid muscle)
20 Thyroid gland
21 Internal jugular vein
22 Esophagus and left recurrent laryngeal nerve
23 **Trachea**
24 **Left vagus nerve**
25 Left common carotid artery
26 Aortic arch
27 Left recurrent laryngeal nerve branching off from vagus nerve
28 Left pulmonary veins
29 **Thoracic aorta** and left vagus nerve
30 Left lung
31 Left phrenic nerve (divided)

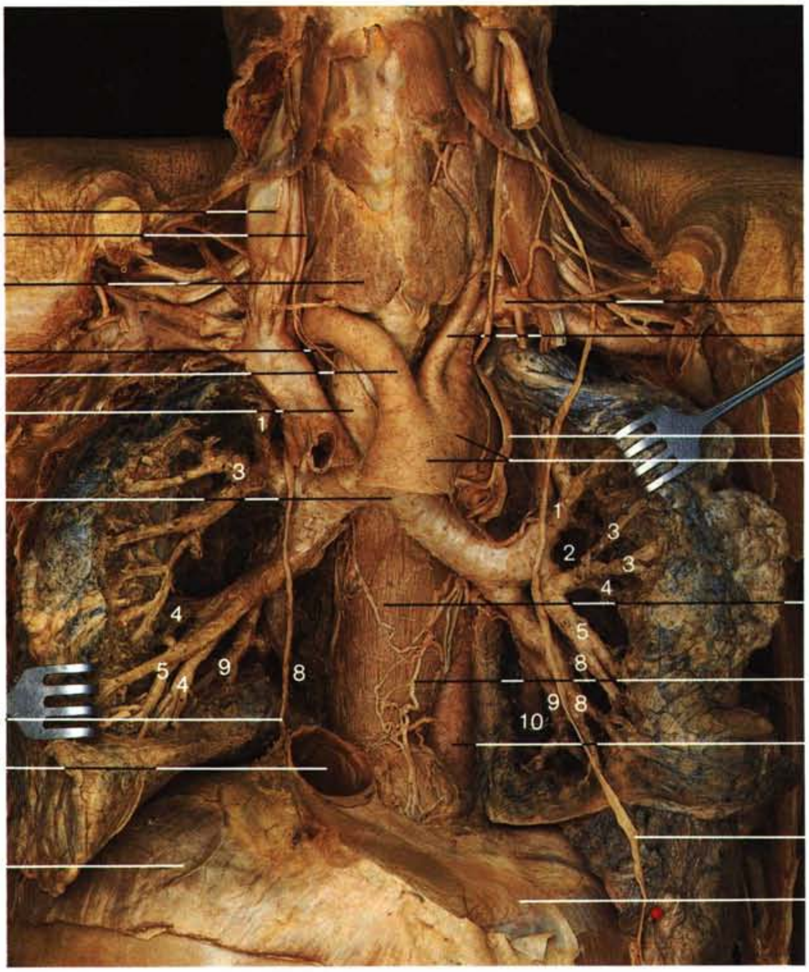
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
1
3
4
5
4
9
8
1
3
2
3
4
5
8
9
8
10

Bronchial tree in situ (ventral aspect). Heart and pericardium have been removed; the bronchi of the bronchopulmonary segments are dissected. 1–10 = numbers of segments.

1 Internal jugular vein
2 Right vagus nerve
3 **Thyroid gland**
4 Right recurrent laryngeal nerve
5 Brachiocephalic trunk
6 Trachea
7 **Bifurcation of trachea**
8 **Right phrenic nerve**
9 Inferior vena cava
10 Diaphragm
11 Left subclavian artery
12 Left common carotid artery
13 Left **vagus nerve**
14 Aortic arch
15 **Esophagus**
16 Esophageal plexus
17 Thoracic aorta
18 **Left phrenic nerve**
19 Pericardium at the central tendon of diaphragm
20 Right pulmonary artery
21 Left pulmonary artery
22 Tracheal lymph nodes
23 Superior tracheobronchial lymph nodes
24 Bronchopulmonary lymph nodes

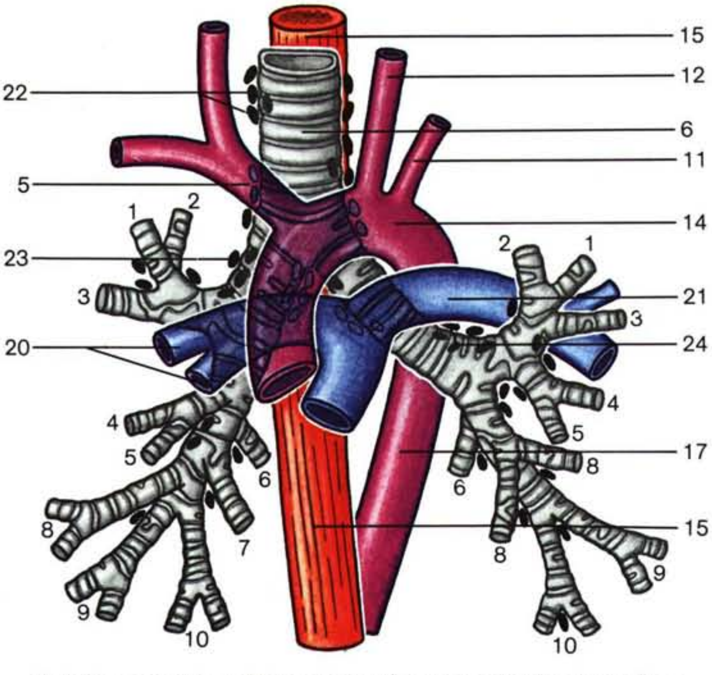
15
12
22
6
11
5
14
1
2
23
2
1
3
21
3
20
24
4
4
5
5
6
17
8
6
8
7
15
8
9
9
10
10

Relation of aorta, pulmonary trunk and esophagus to trachea and bronchial tree. (Schematic drawing.)
1–10 = number of segments (cf. p. 230 and 235).

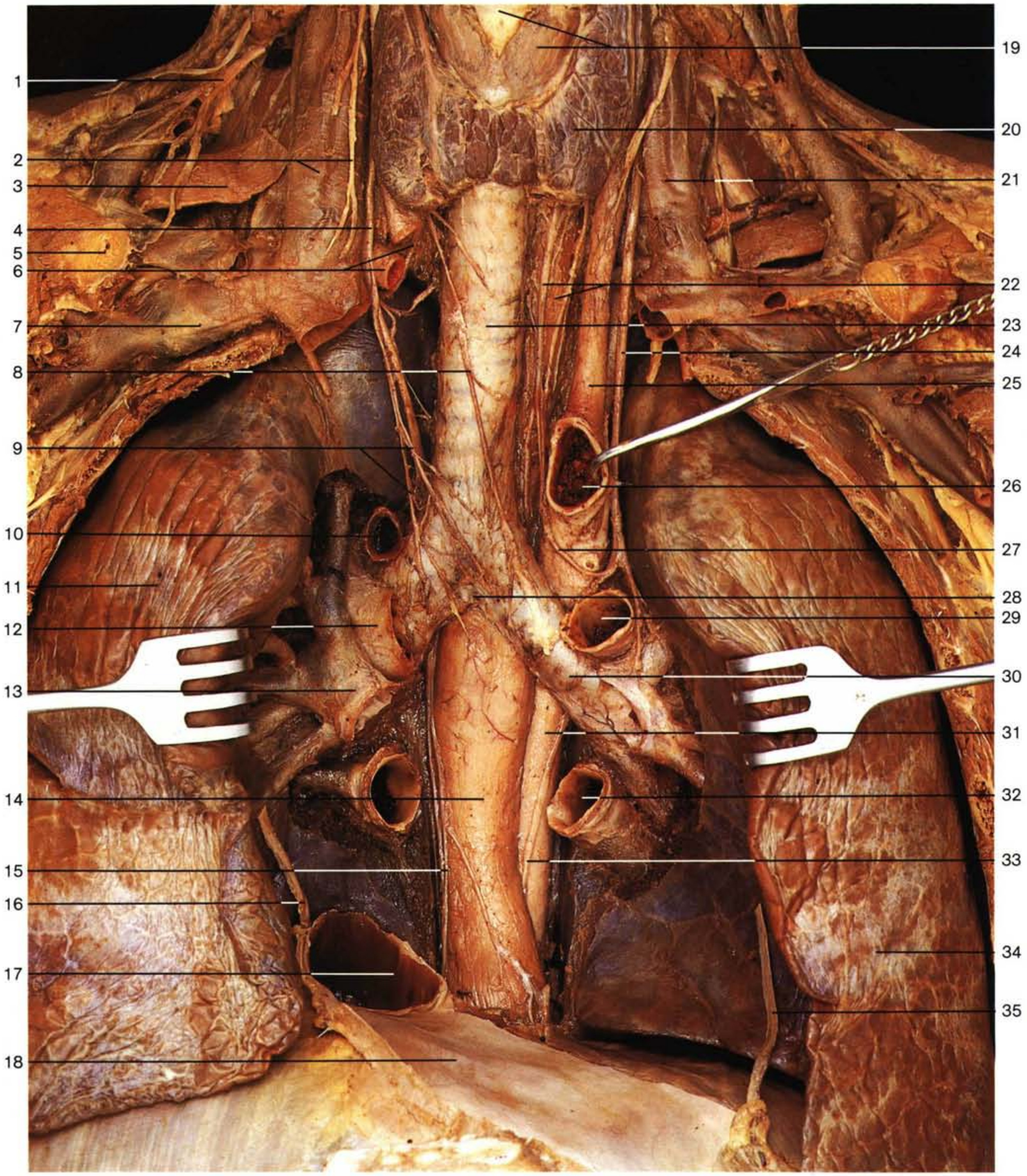

Organs of posterior mediastinum (ventral aspect). The heart with the pericardium has been removed, and the lungs and aortic arch have been slightly reflected to show the vagus nerves and their branches.

1 Supraclavicular nerves
2 Right internal jugular vein with ansa cervicalis
3 Omohyoid muscle
4 **Right vagus nerve**
5 Clavicle
6 Right subclavian artery and recurrent laryngeal nerve
7 Right subclavian vein
8 Superior cervical cardiac branch of vagus nerve
9 Inferior cervical cardiac branch of vagus nerve
10 Azygos arch (divided)
11 **Right lung**
12 **Right pulmonary artery**
13 **Right pulmonary veins**
14 **Esophagus**
15 Esophageal plexus
16 Right phrenic nerve (divided)
17 Inferior vena cava
18 Pericardium covering the diaphragm
19 Larynx (thyroid cartilage and cricothyroid muscle)
20 Thyroid gland
21 Left internal jugular vein
22 Esophagus and left recurrent laryngeal nerve
23 Trachea
24 **Left vagus nerve**
25 Left common carotid artery
26 Aortic arch
27 Left recurrent laryngeal nerve
28 **Bifurcation of trachea**
29 **Left pulmonary artery**
30 Left primary bronchus
31 Descending aorta
32 Left pulmonary veins
33 Branch of left vagus nerve
34 **Left lung**
35 Left phrenic nerve (divided)

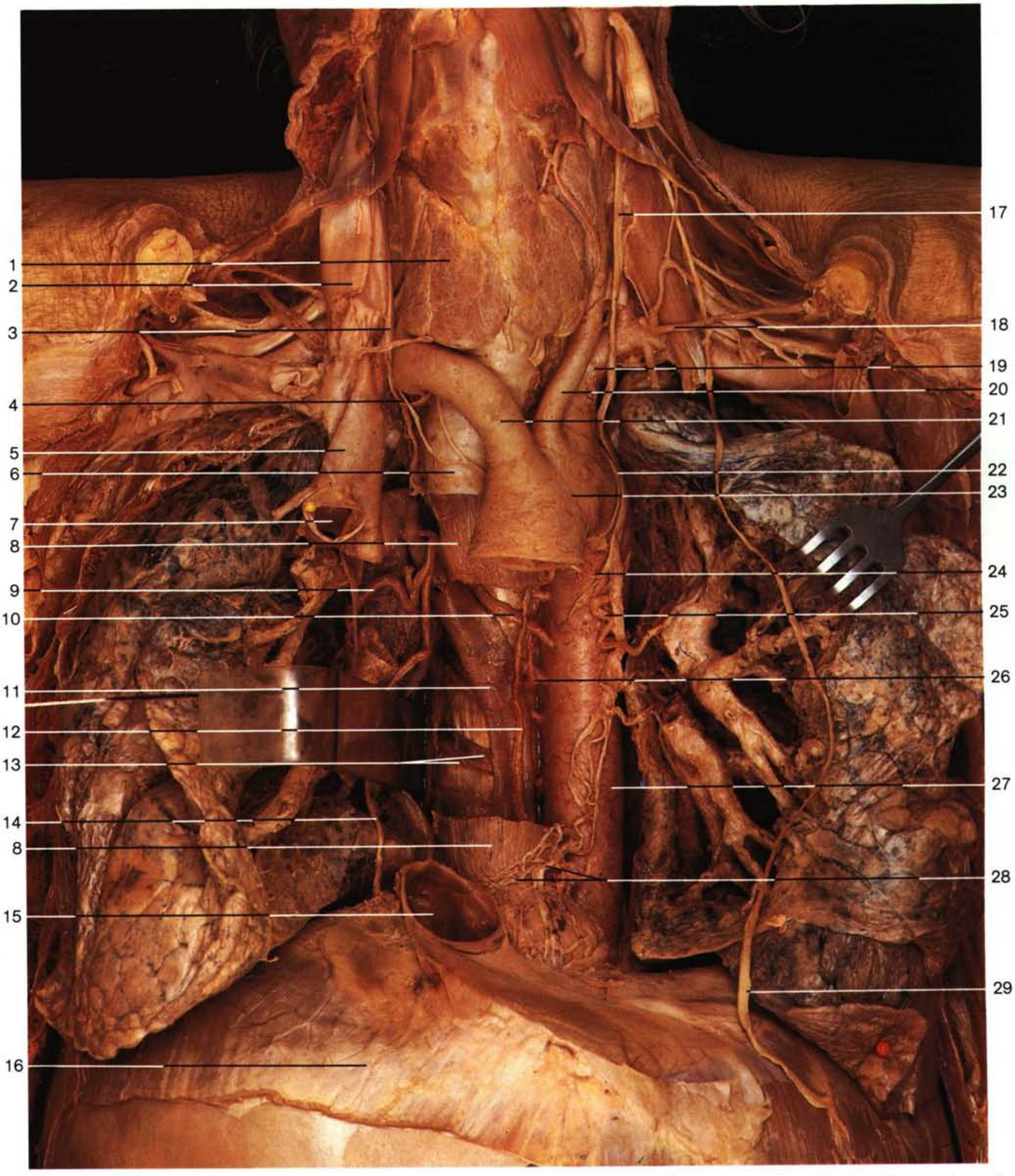

Mediastinal organs (ventral aspect). Heart and distal part of esophagus have been removed to display the vessels and nerves of the posterior mediastinum.

1 Thyroid gland
2 Right internal jugular vein
3 Right vagus nerve
4 Point, where right recurrent laryngeal nerve is branching off the vagus nerve
5 Right brachiocephalic vein
6 Trachea
7 Left brachiocephalic vein (reflected)
8 Esophagus
9 Right bronchial artery
10 Posterior intercostal artery
11 **Azygos vein**
12 **Thoracic duct**
13 Posterior intercostal artery and vein (in front of the vertebral column)
14 Right phrenic nerve
15 Inferior vena cava
16 Diaphragm
17 **Left vagus nerve**
18 Thyrocervical trunk
19 Left subclavian artery
20 Left common carotid artery
21 Brachiocephalic trunk
22 Left vagus nerve
23 **Aortic arch**
24 Left recurrent laryngeal nerve
25 Left bronchial artery
26 Lymph node
27 Thoracic aorta
28 Esophageal plexus
29 Left phrenic nerve

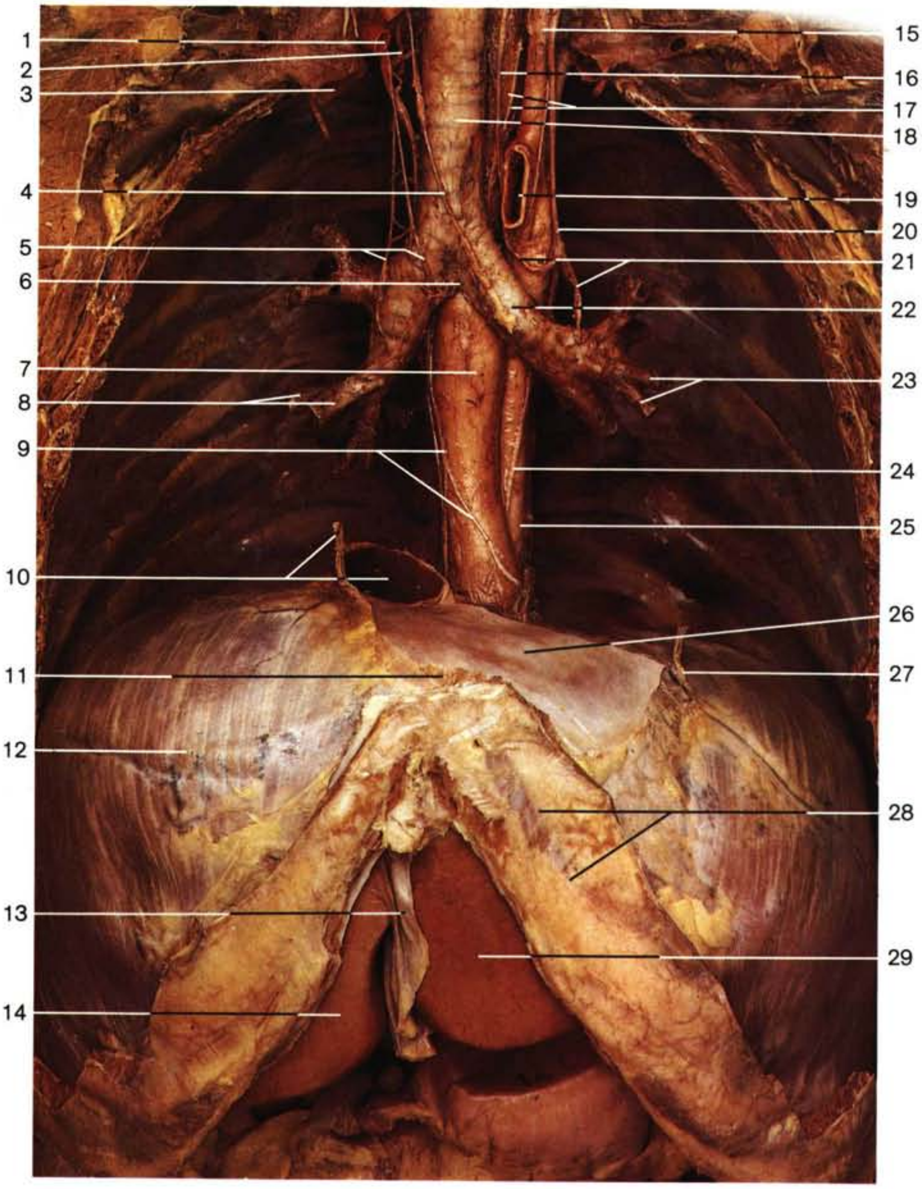

Diaphragm and organs of mediastinum (anterior aspect). Heart and lungs have been removed; the costal margin remains in place. Note the different courses of left and right vagus.

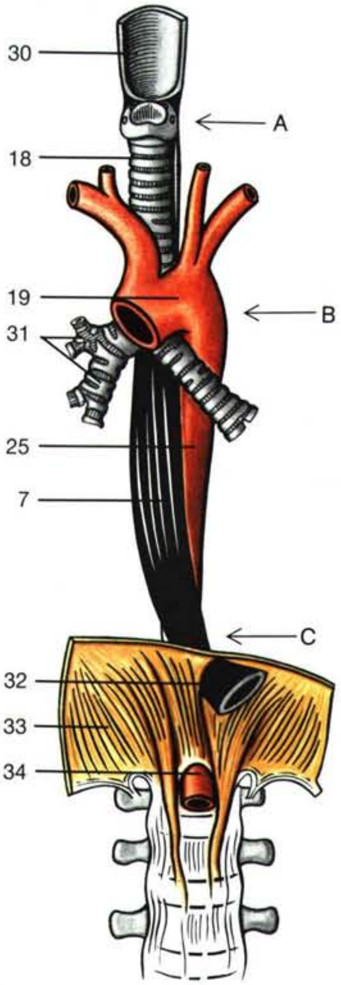

Organs of posterior mediastinum (ventral aspect). (Schematic drawing.) Three regions in which the esophagus is narrowed are shown:

A at the level of the cricoid cartilage;

B at the level of the aortic arch;

C at the level of the diaphragm.

1 Right subclavian artery
2 Right recurrent laryngeal nerve
3 Right brachiocephalic vein
4 Superior cervical cardiac nerve
5 Inferior cervical cardiac nerves and pulmonary branches
6 **Bifurcation of trachea**
7 **Esophagus** (thoracic part)
8 Bronchi of lateral and medial segments of middle lobe
9 Esophageal plexus and branches of right vagus nerve
10 Inferior vena cava and right phrenic nerve (cut)
11 Sternal part of diaphragm
12 Costal part of **diaphragm**
13 Falciform ligament of liver
14 Liver (quadrate lobe)
15 Left common carotid artery
16 Left recurrent laryngeal nerve
17 Esophageal branches of left vagus nerve and esophagus
18 **Trachea**
19 Aortic arch
20 Left vagus nerve
21 Left recurrent laryngeal nerve with inferior cardiac nerve
22 Left primary bronchus
23 Superior and inferior lingular bronchi
24 Esophageal plexus of left vagus nerve
25 **Descending aorta**
26 Central tendon of diaphragm covered with pericardium
27 Left phrenic nerve (divided)
28 Costal margin
29 **Liver,** left lobe
30 Pharynx
31 Secondary bronchi
32 Esophagus (abdominal part)
33 Diaphragm
34 Abdominal aorta

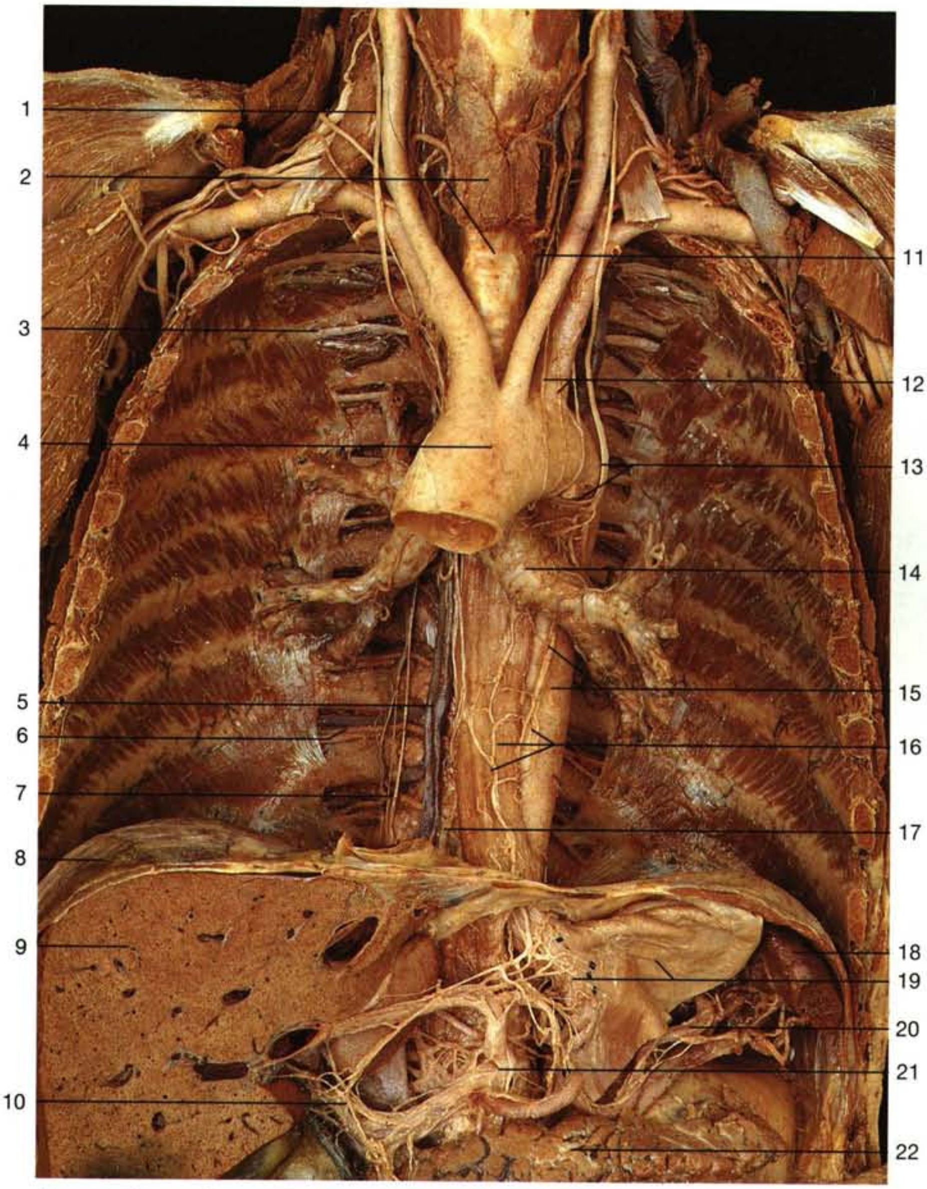

Organs of posterior mediastinum (anterior aspect).

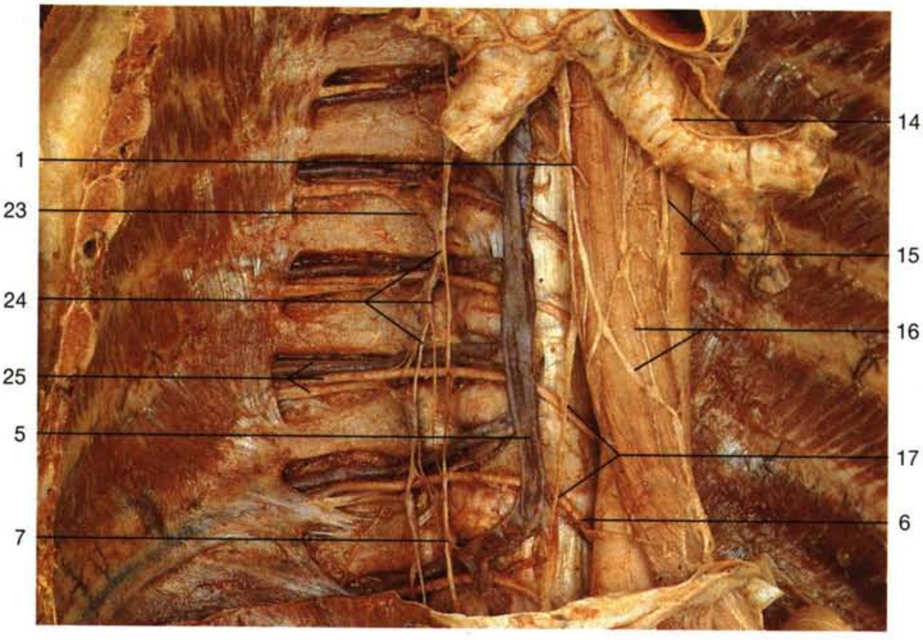

Inferior segment of posterior mediastinum (anterior aspect).

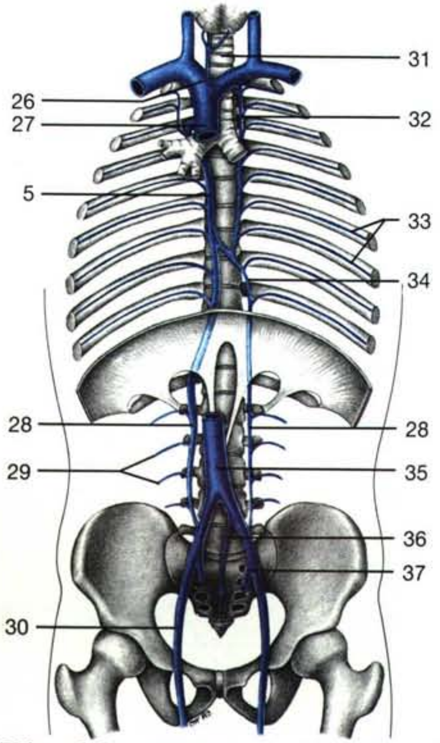

Veins of the posterior wall of thoracic and abdominal cavity. (Schematic drawing.)

1	Right vagus nerve
2	Thyroid gland and trachea
3	Intercostal nerve
4	Aortic arch
5	**Azygos vein**
6	Posterior intercostal artery
7	**Greater splanchnic nerve**
8	Diaphragm
9	Liver
10	Proper hepatic artery and hepatic plexus
11	Left recurrent laryngeal nerve
12	Inferior cervical cardiac nerves
13	Left vagus nerve and left recurrent laryngeal nerve
14	Left primary bronchus
15	Thoracic aorta and left vagus nerve
16	Esophagus and esophageal plexus
17	**Thoracic duct**
18	Spleen
19	Anterior gastric plexus and stomach (divided)
20	Splenic artery and splenic plexus
21	**Celiac trunk** and **celiac plexus**
22	Pancreas
23	Ramus communicans
24	**Sympathetic trunk** and **sympathetic ganglion**
25	**Posterior intercostal vein** and **artery** and **intercostal nerve**
26	Right brachiocephalic vein
27	Superior vena cava
28	Ascending lumbar vein
29	Lumbar veins
30	Right external iliac vein
31	Trachea
32	Accessory hemiazygos vein
33	Posterior intercostal veins
34	Hemiazygos vein
35	Inferior vena cava
36	Median sacral vein
37	Internal iliac vein

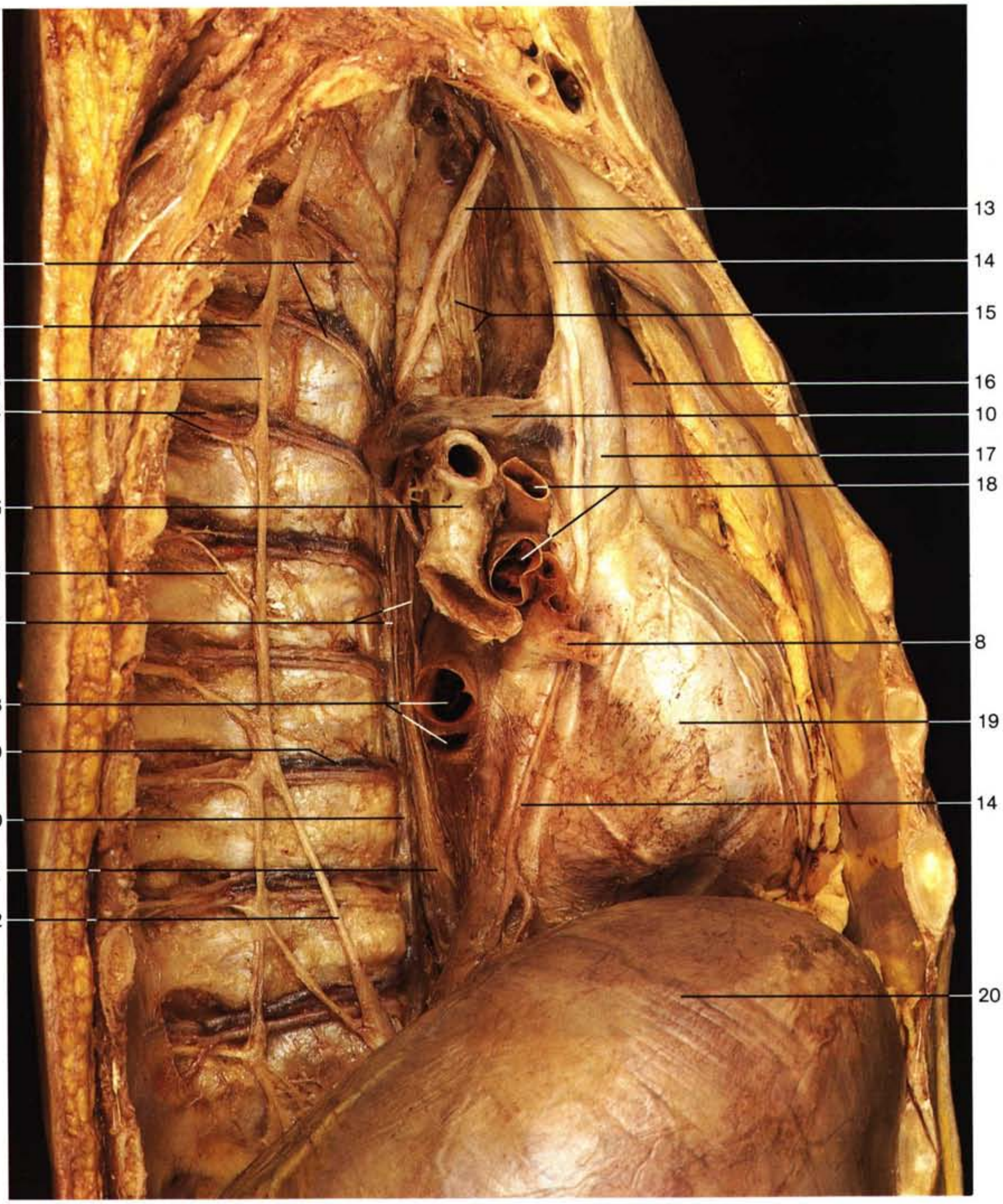

Mediastinal organs (right lateral aspect). Right lung and pleura of right half of the thorax have been removed.

1 Posterior intercostal arteries
2 Ganglion of sympathetic trunk
3 **Sympathetic trunk**
4 Vessels and nerves of the intercostal space (from above: Posterior intercostal vein and artery and intercostal nerve)
5 Right primary bronchus
6 Ramus communicans of sympathetic trunk
7 Esophageal plexus (branches of right vagus nerve)
8 Pulmonary veins
9 Posterior intercostal vein
10 **Azygos vein**
11 Esophagus
12 Greater splanchnic nerve
13 **Right vagus nerve**
14 **Right phrenic nerve**
15 Inferior cervical cardiac branches of vagus nerve
16 **Aortic arch**
17 Superior vena cava
18 Right pulmonary artery
19 **Heart with pericardium**
20 Diaphragm

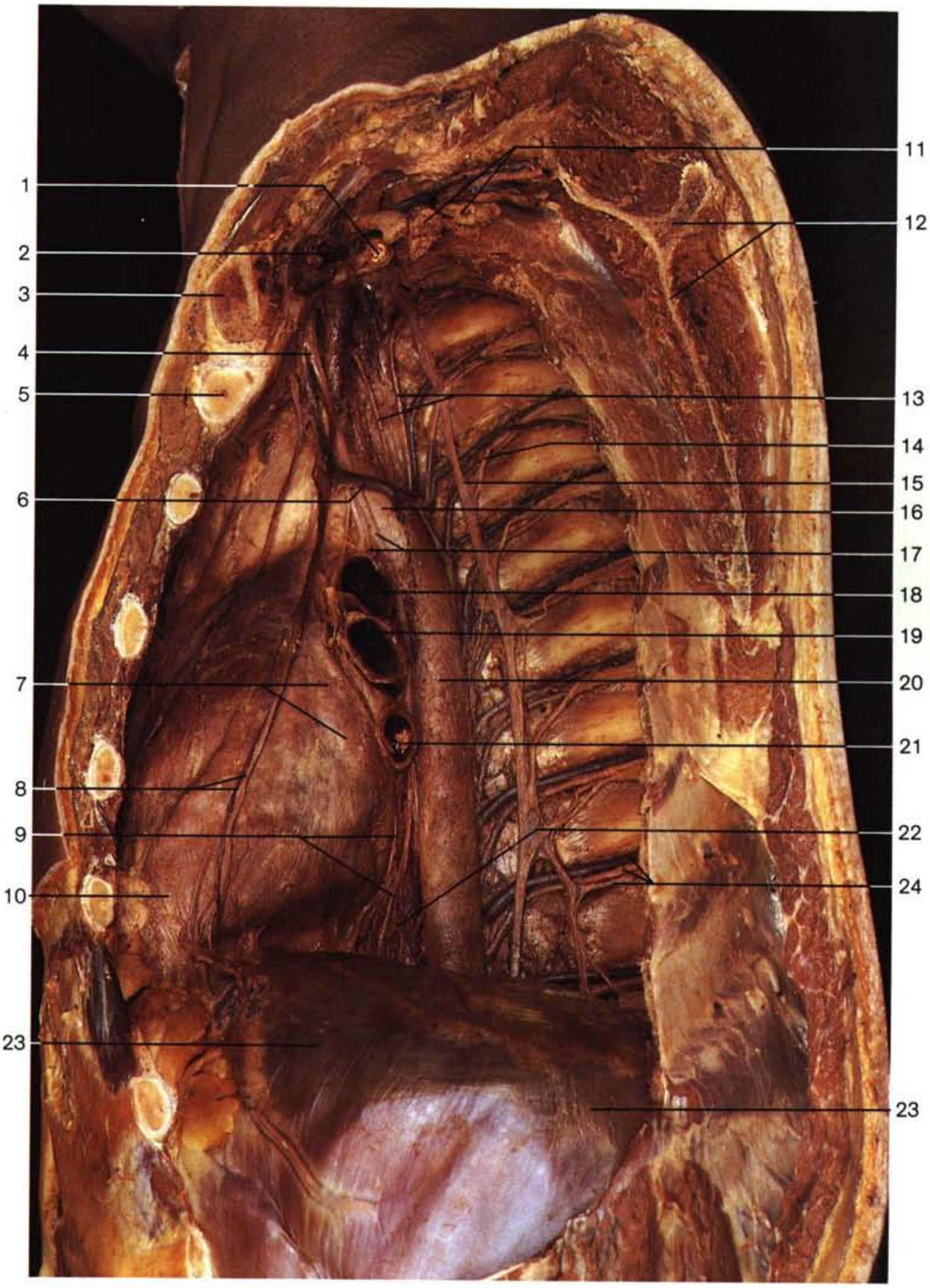

Organs of posterior and superior mediastinum (left lateral aspect).

1 Subclavian artery
2 Subclavian vein
3 Clavicle (divided)
4 **Left vagus nerve**
5 First rib (divided)
6 Left superior intercostal vein
7 **Left atrium with pericardium**
8 **Left phrenic nerve** and pericardiacophrenic artery and vein
9 Esophageal plexus (branches derived from left vagus nerve)
10 **Apex of heart with pericardium**
11 Brachial plexus
12 Scapula (divided)
13 Posterior intercostal arteries
14 Ramus communicans of sympathetic trunk
15 **Sympathetic trunk**
16 Aortic arch
17 Left vagus nerve and left recurrent laryngeal nerve
18 Left pulmonary artery
19 Left primary bronchus
20 **Thoracic aorta**
21 Pulmonary vein
22 Esophagus (thoracic part)
23 **Diaphragm**
24 Posterior intercostal artery and vein and intercostal nerve

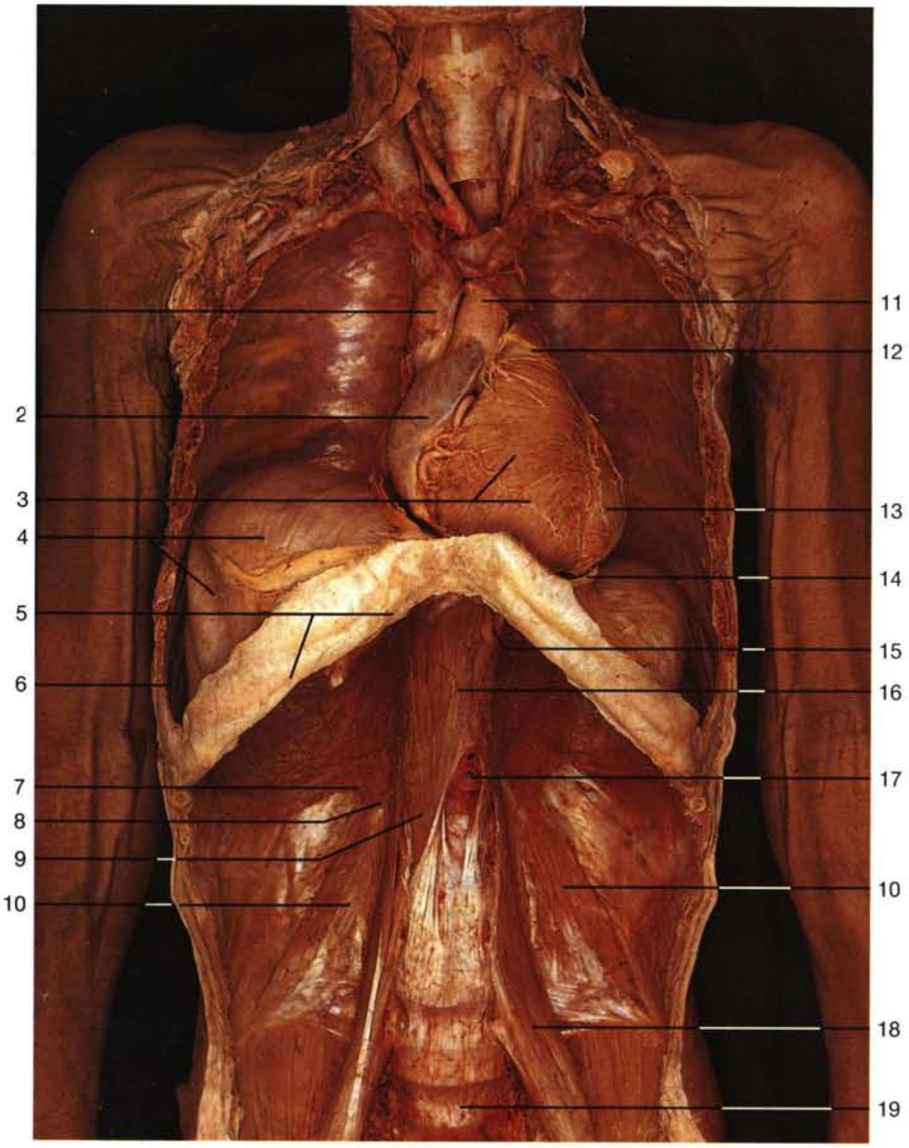

1 Superior vena cava
2 Right atrium
3 Right ventricle
4 **Costal part of diaphragm**
5 Costal margin
6 Position of costodiaphragmatic recess
7 **Lateral arcuate ligament**
8 **Medial arcuate ligament**
9 Right crus of lumbar part of diaphragm
10 Quadratus lumborum muscle
11 Ascending aorta
12 Pulmonary trunk
13 Left ventricle
14 Pericardium, diaphragm
15 Esophageal hiatus and abdominal part of esophagus (cut)
16 **Lumbar part of diaphragm**
17 Aortic hiatus
18 Psoas major muscle
19 Lumbar vertebra

Diaphragm in situ (anterior aspect). Anterior walls of thoracic and abdominal cavities have been removed. Natural position of the heart above the central tendon on the diaphragm is shown.

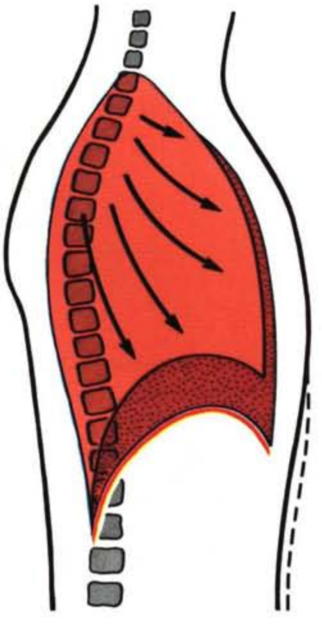

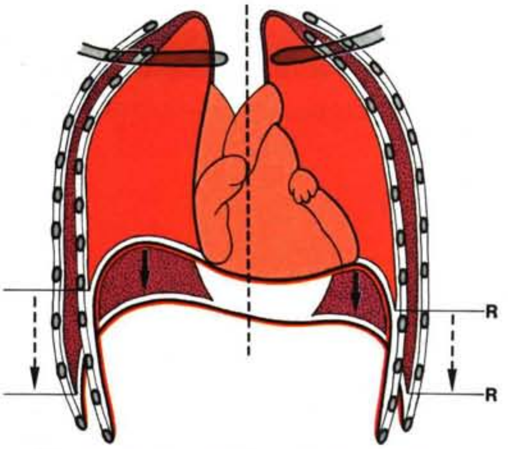

Changes in the position of the diaphragm and thoracic cage during respiration. Left: lateral aspect; right: anterior aspect. During inspiration the diaphragm moves downwards and the lower part of the thoracic cage expands forward and laterally, causing the costodiaphragmatic recess (R) to enlarge (cf. dotted arrows).

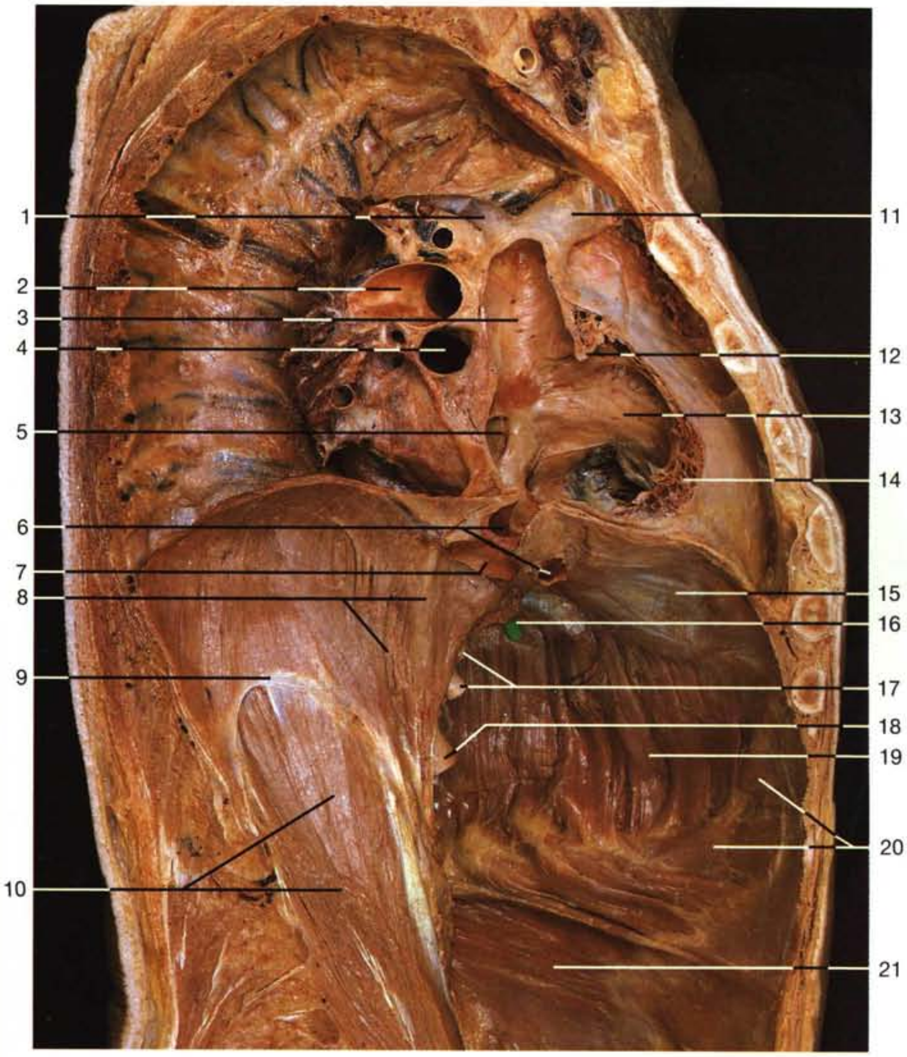

1 Azygos venous arch
2 Right pulmonary artery
3 **Superior vena cava**
4 Right pulmonary vein
5 Fossa ovalis
6 Hepatic veins
7 Inferior vena cava
8 Right crus of lumbar part of **diaphragm**
9 Medial arcuate ligament
10 Psoas major muscle
11 Left brachiocephalic vein
12 Terminal crista
13 **Right atrium**
14 Right auricle
15 Central tendon of diaphragm
16 Esophagus
17 Celiac trunk and superior mesenteric artery
18 Aorta
19 Costal part of diaphragm
20 Costal margin
21 Transversus abdominis muscle

Diaphragm. Paramedian section to the right of the median plane through thoracic and upper abdominal cavity. The plane passes through the superior and inferior vena cava just to the right of the vertebral bodies. Most of the heart remains in situ to the left of this plane (specimen is viewed from the right side).

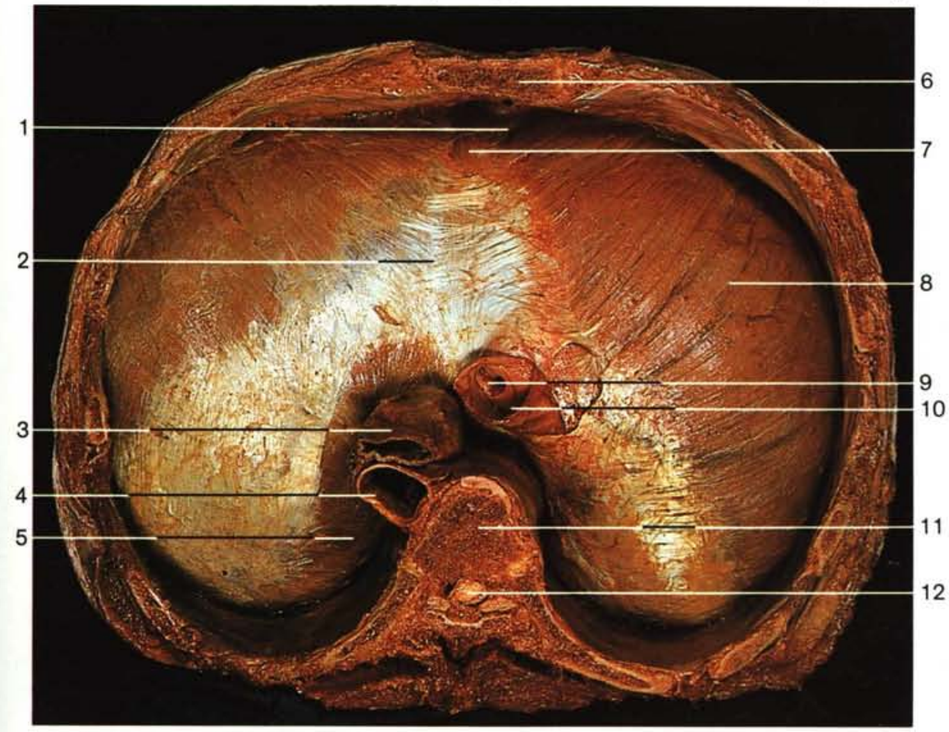

1 Sternocostal triangle
2 Central tendon (from above)
3 Esophagus
4 Aorta
5 Lumbar part of diaphragm
6 Sternum
7 Sternal part of diaphragm
8 Costal part of diaphragm
9 Entrance of hepatic veins
10 Inferior vena cava
11 Body of 9th thoracic vertebra
12 Spinal cord

Diaphragm (superior aspect). The pleura and thoracic wall have been removed.

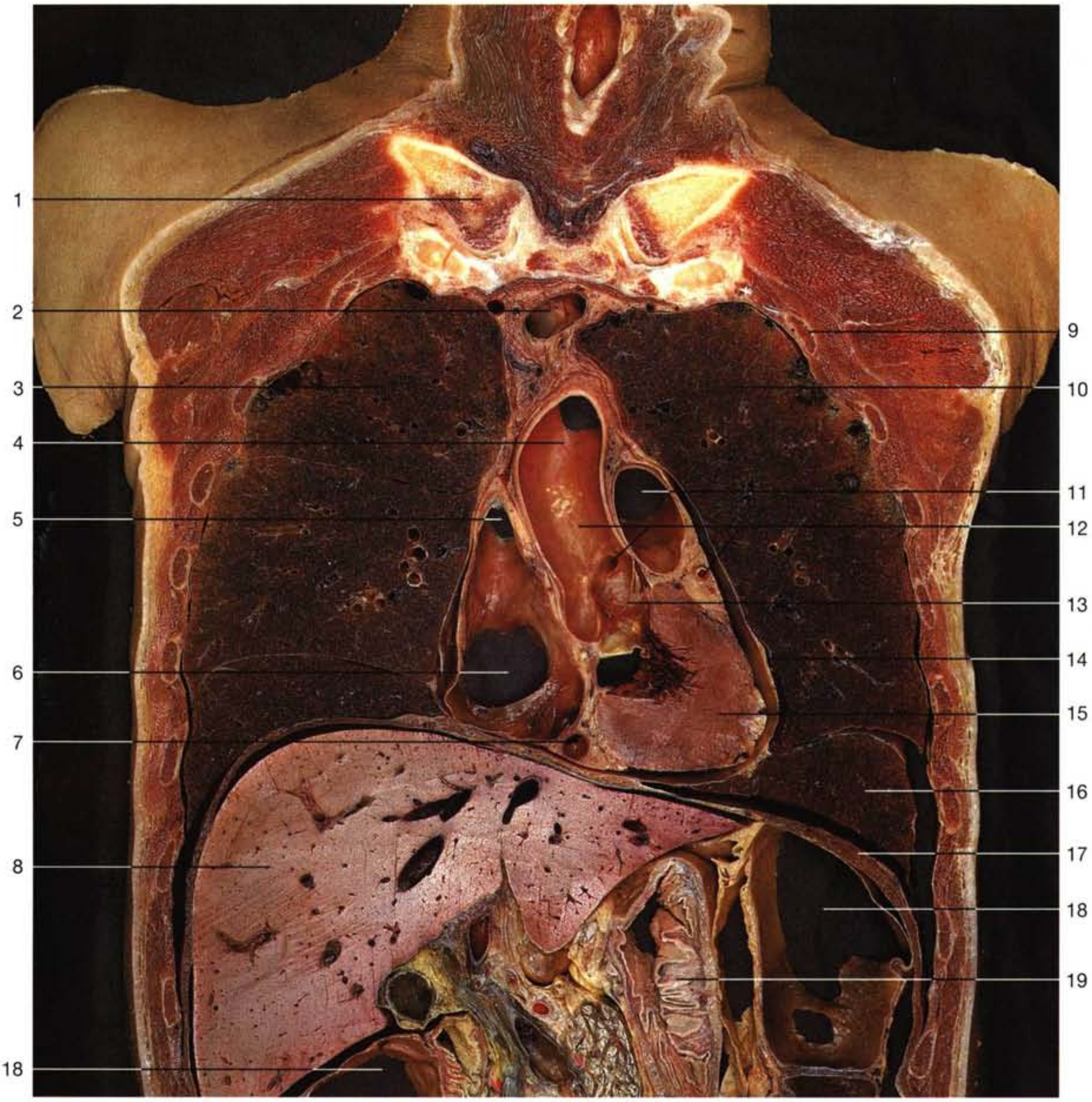

Coronal section through the thorax at the level of ascending aorta (anterior aspect).

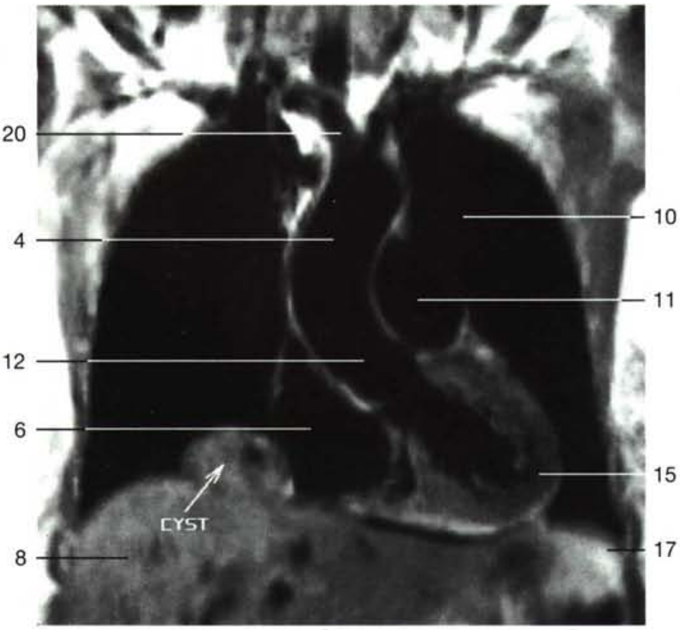

Coronal section through the thorax at the level of ascending aorta. (MR-Scan.)

1	Clavicle
2	Left brachiocephalic vein
3	Superior lobe of right lung
4	Aortic arch
5	Superior vena cava
6	Right atrium (entrance of inferior vena cava)
7	Coronary sinus
8	Liver
9	Second rib
10	Superior lobe of left lung
11	Pulmonary trunk
12	Ascending aorta and left coronary artery
13	Aortic valve
14	Pericardium
15	Myocardium of left ventricle
16	Lower lobe of left lung
17	Diaphragm
18	Colic flexures
19	Stomach
20	Brachiocephalic trunk

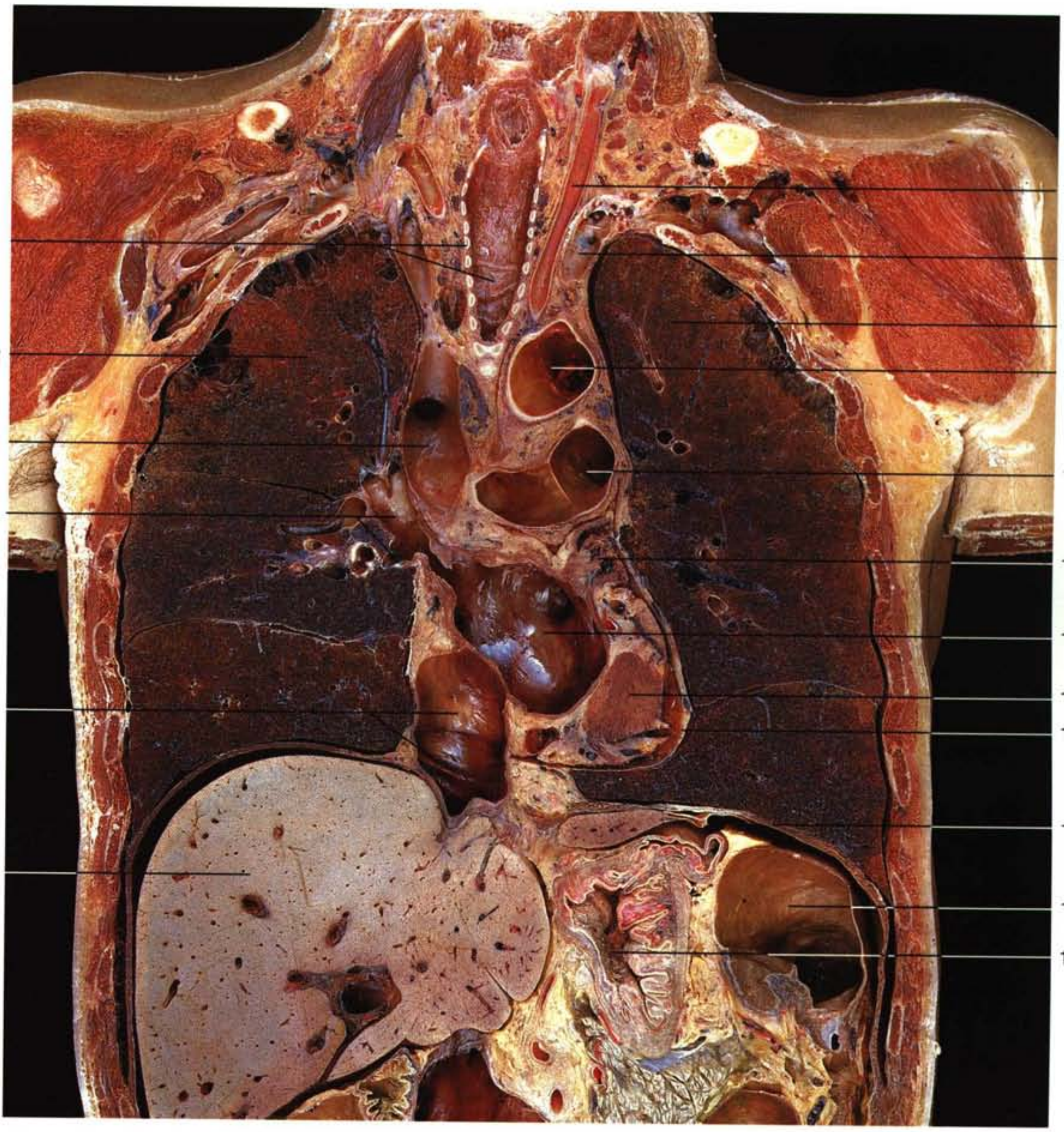

Coronal section through the thorax at the level of superior and inferior vena cava (anterior aspect).

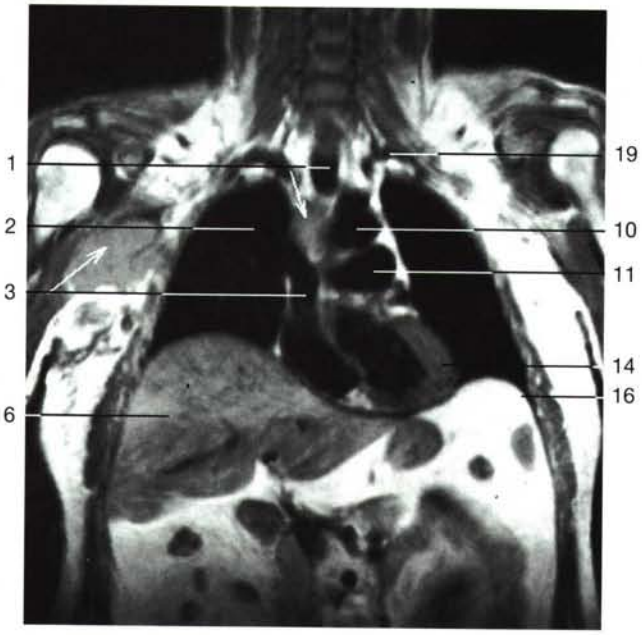

Coronal section through the thorax at the level of superior vena cava. (MR-Scan.)

1 Trachea
2 Upper lobe of right lung
3 **Superior vena cava**
4 Right pulmonary veins
5 Inferior vena cava and right atrium
6 Liver
7 Left common carotid artery
8 Left subclavian vein
9 Upper lobe of left lung
10 Aortic arch
11 Left pulmonary artery
12 Left auricle
13 **Left atrium** with orifices of pulmonary veins
14 Left ventricle (myocardium)
15 Pericardium
16 Diaphragm
17 Left colic flexure
18 Stomach
19 Left subclavian artery

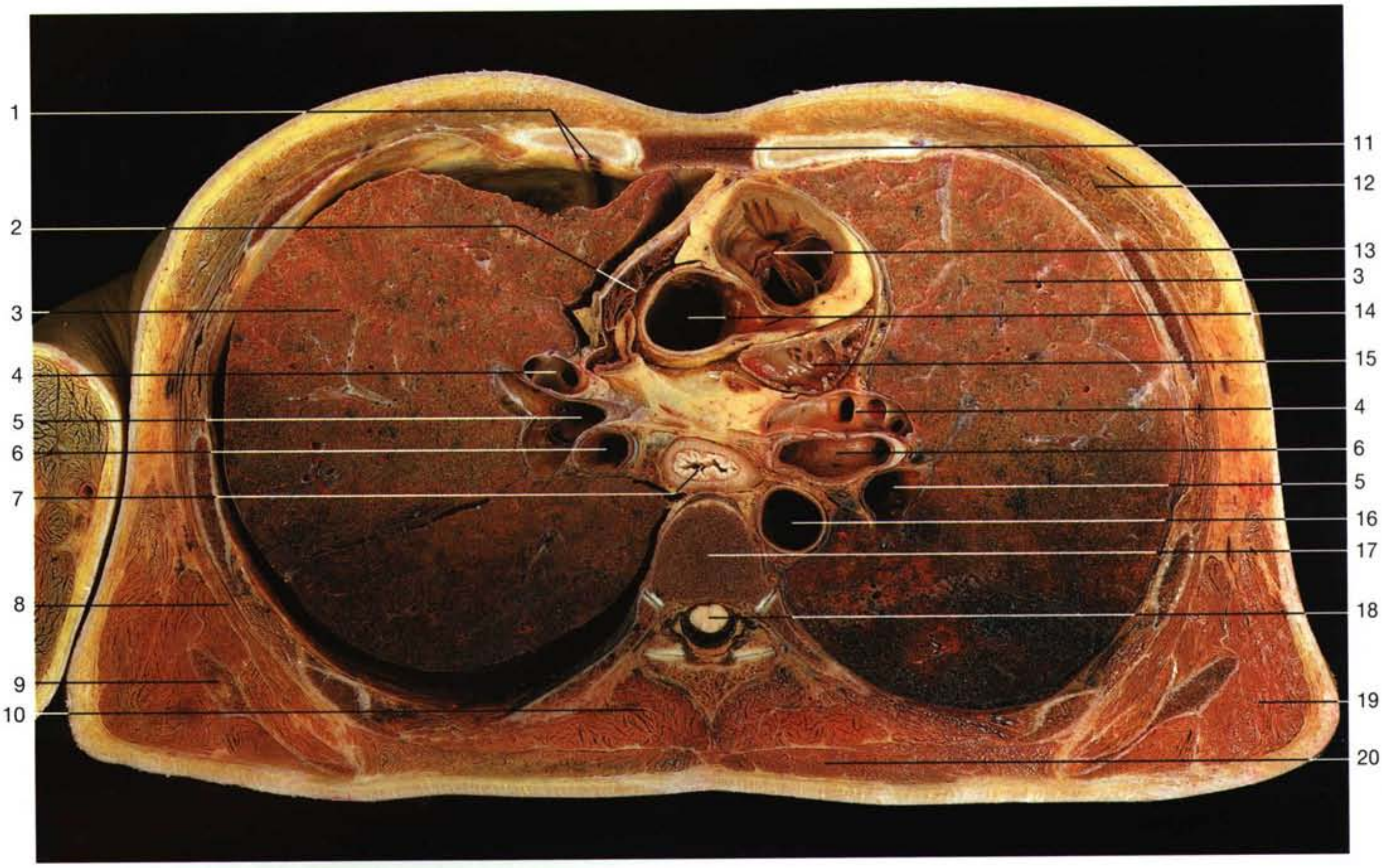

Horizontal section through the thorax at level 1 (from below).

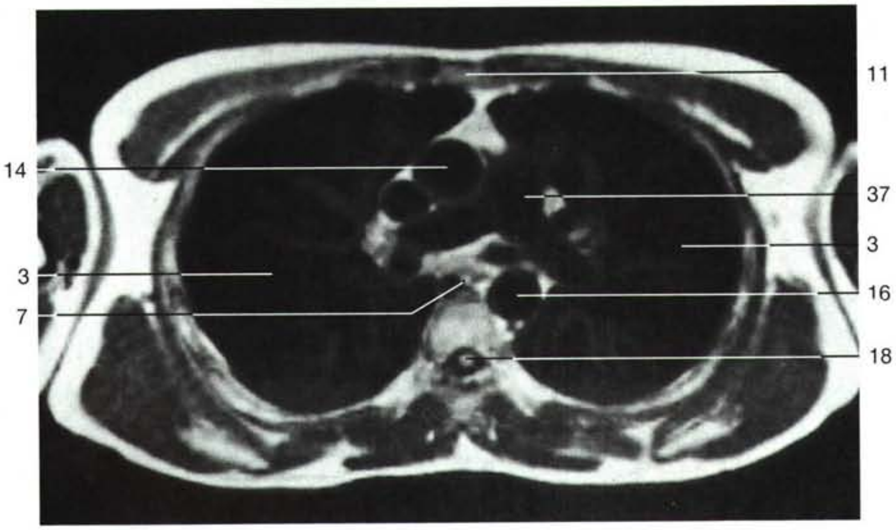

Horizontal section through the thorax at level 1 (from below). (MR-Scan.)

1 Internal thoracic artery and vein
2 Right atrium
3 Lung
4 Pulmonary artery
5 Pulmonary vein
6 Primary bronchus
7 Esophagus
8 Serratus anterior muscle
9 Scapula
10 Longissimus thoracis muscle
11 Sternum
12 Pectoralis major and minor muscles
13 Conus arteriosus (right ventricle), pulmonic valve
14 Ascending aorta and left coronary artery (only in upper figure)
15 **Left atrium**
16 **Descending aorta**
17 Thoracic vertebra
18 Spinal cord
19 Latissimus dorsi muscle
20 Trapezius muscle

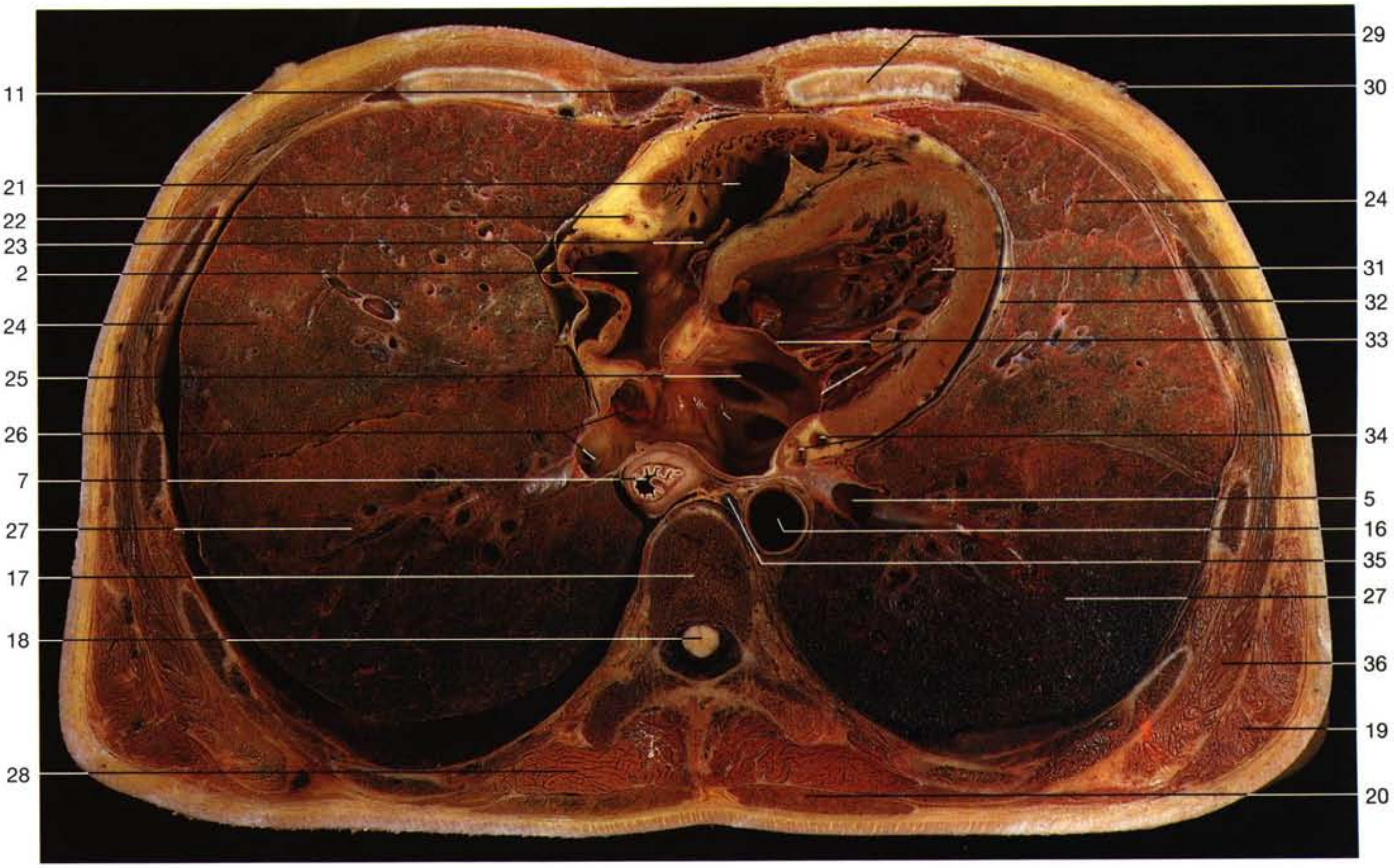
11
21
22
23
2
24
25
26
7
27
17
18
28
29
30
24
31
32
33
34
5
16
35
27
36
19
20

Horizontal section through the thorax at level 2 (from below).

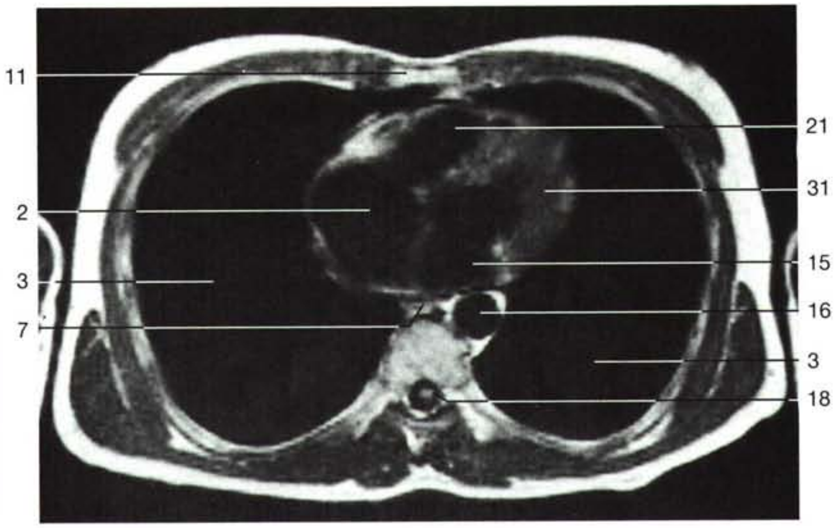
11
2
3
7
21
31
15
16
3
18

Horizontal section through the thorax at level 2 (from below).
(MR-Scan.)

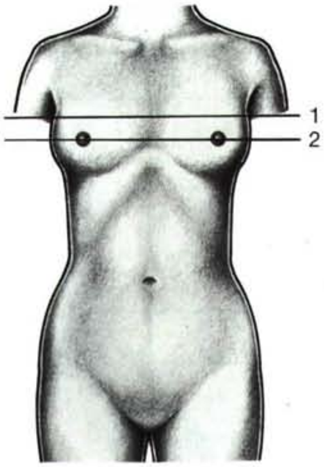

Levels of sections.

21 **Right ventricle**
22 Right coronary artery
23 Right atrioventricular valve
24 Lung (upper lobe)
25 **Left atrium**
26 Pulmonary veins
27 Lung (lower lobe)
28 Erector muscle of spine
29 Third costal cartilage
30 Nipple

31 **Left ventricle**
32 Pericardium
33 Left atrioventricular valve
34 Left coronary artery and coronary sinus
35 Accessory hemiazygos vein
36 Serratus anterior muscle
37 Pulmonary trunk

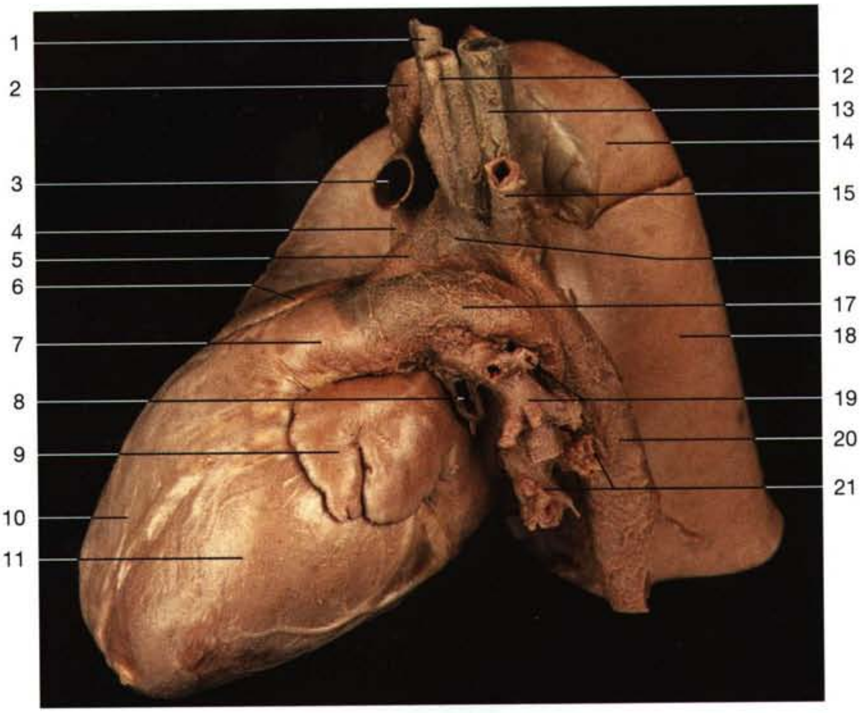

Heart and right lung of the fetus (viewed from left side). The left lung has been removed. Note the ductus arteriosus (Botalli).

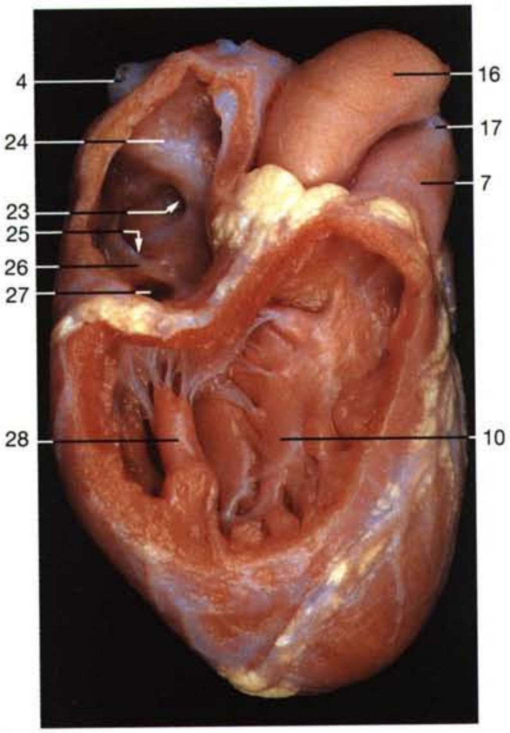

Heart of the fetus (anterior aspect). Right atrium and ventricle opened.

Shunts in the fetal circulation system		
1. Ductus venosus (of Arantius)	between umbilical vein and inferior vena cava	bypass of liver circulation
2. Foramen ovale	between right and left atrium	bypass of pulmonary circulation
3. Ductus arteriosus (Botalli)	between pulmonary trunk and aorta	

1 Right common carotid artery
2 Right brachiocephalic vein
3 Left brachiocephalic vein
4 Superior vena cava
5 **Ascending aorta**
6 Right auricle
7 **Pulmonary trunk**
8 Left primary bronchus
9 Left auricle
10 Right ventricle
11 Left ventricle
12 Left common carotid artery
13 Trachea
14 Superior lobe of right lung
15 Left subclavian artery
16 Aortic arch
17 **Ductus arteriosus (Botalli)**
18 Inferior lobe of right lung
19 Left pulmonary artery with branches to the left lung
20 **Descending aorta**
21 Left pulmonary veins
22 Inferior vena cava
23 **Foramen ovale**
24 Right atrium
25 Opening of inferior vena cava
26 Valve of inferior vena cava (Eustachian valve)
27 Opening of coronary sinus
28 Anterior papillary muscle of right ventricle

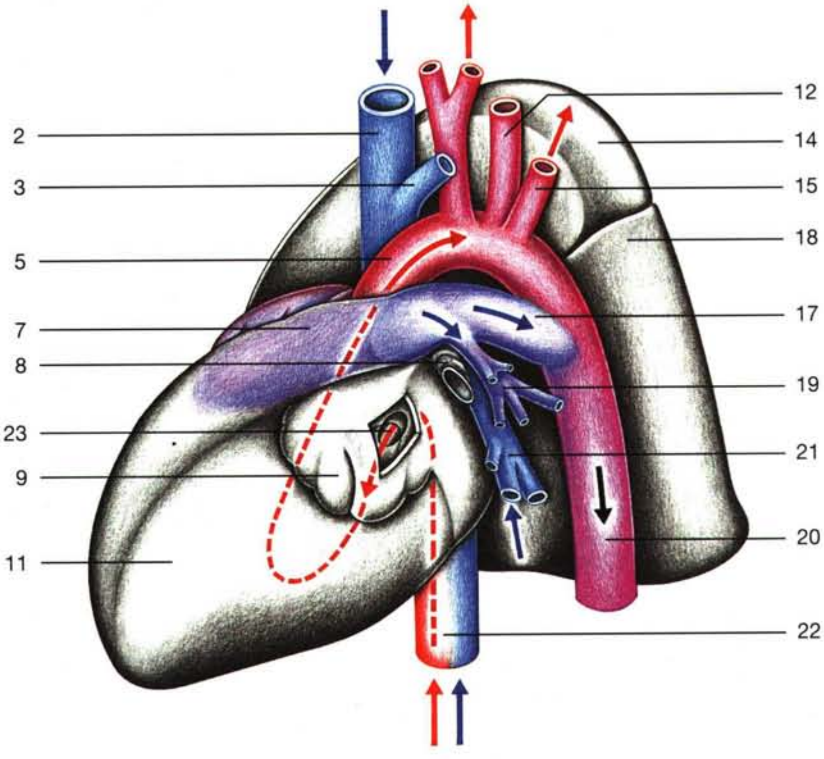

◁ **Heart of the fetus** (schematic drawing). Direction of blood flow indicated by arrows. Note the change in oxygenation of blood after ductus arteriosus entry into aorta.

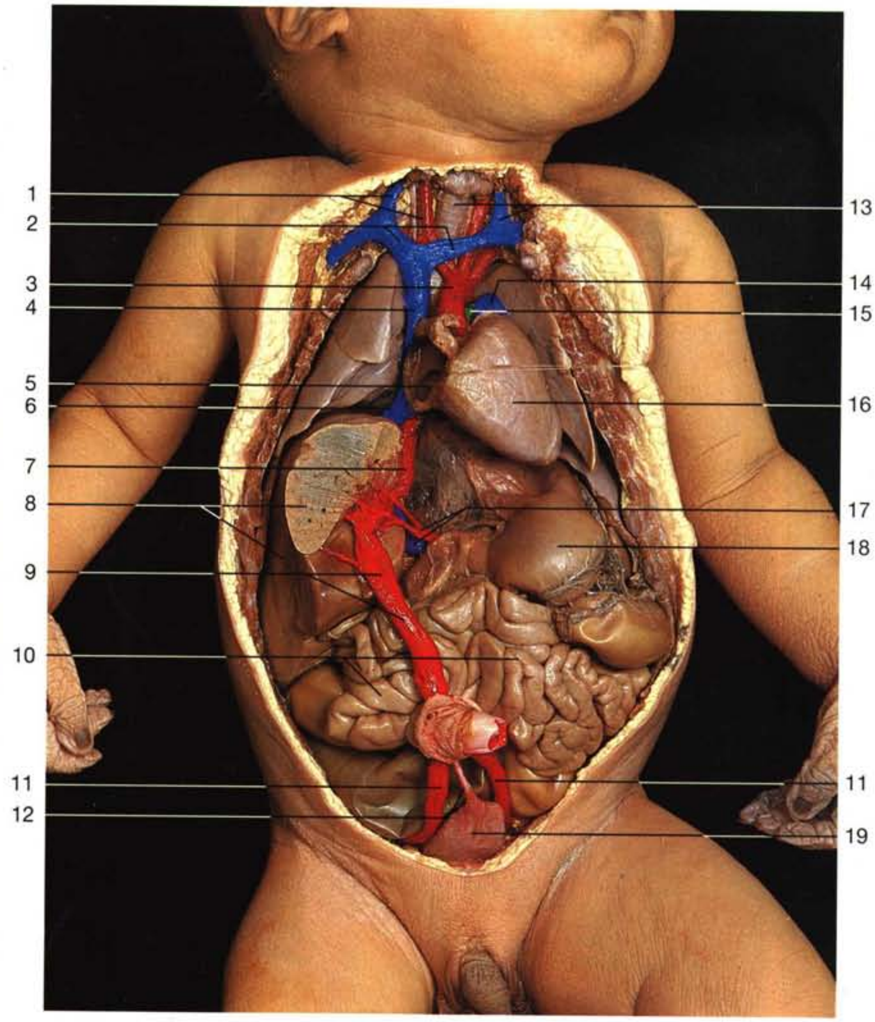

Thoracic and abdominal organs in the newborn (anterior aspect). The right atrium has been opened to show the foramen ovale. The left lobe of the liver has been removed.

1 Internal jugular vein and right common carotid artery
2 Right and left brachiocephalic vein
3 Aortic arch
4 Superior vena cava
5 **Foramen ovale**
6 Inferior vena cava
7 **Ductus venosus**
8 Liver
9 **Umbilical vein**
10 Small intestine
11 **Umbilical artery**
12 Urachus
13 Trachea and left internal jugular vein
14 Left pulmonary artery
15 **Ductus arteriosus** (Botalli)
16 Right ventricle
17 Hepatic arteries (red) and portal vein (blue)
18 Stomach
19 Urinary bladder
20 Portal vein
21 Pulmonary veins
22 Descending aorta
23 Placenta

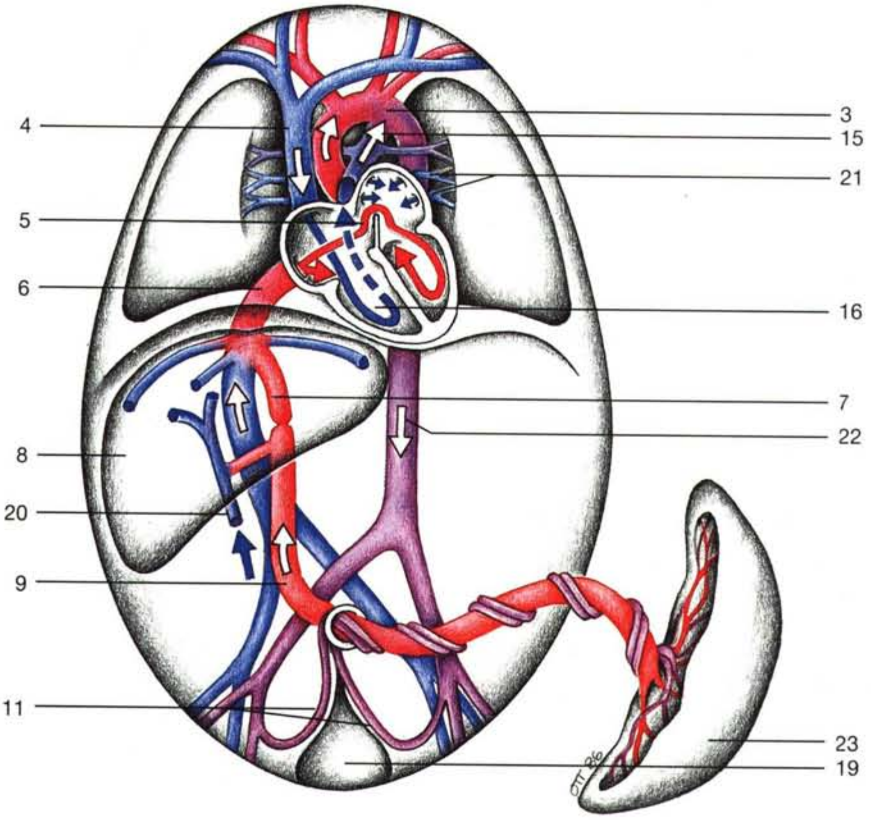

◁ **Fetal circulatory system.** (Schematic drawing.) The oxygen gradient is indicated by color.

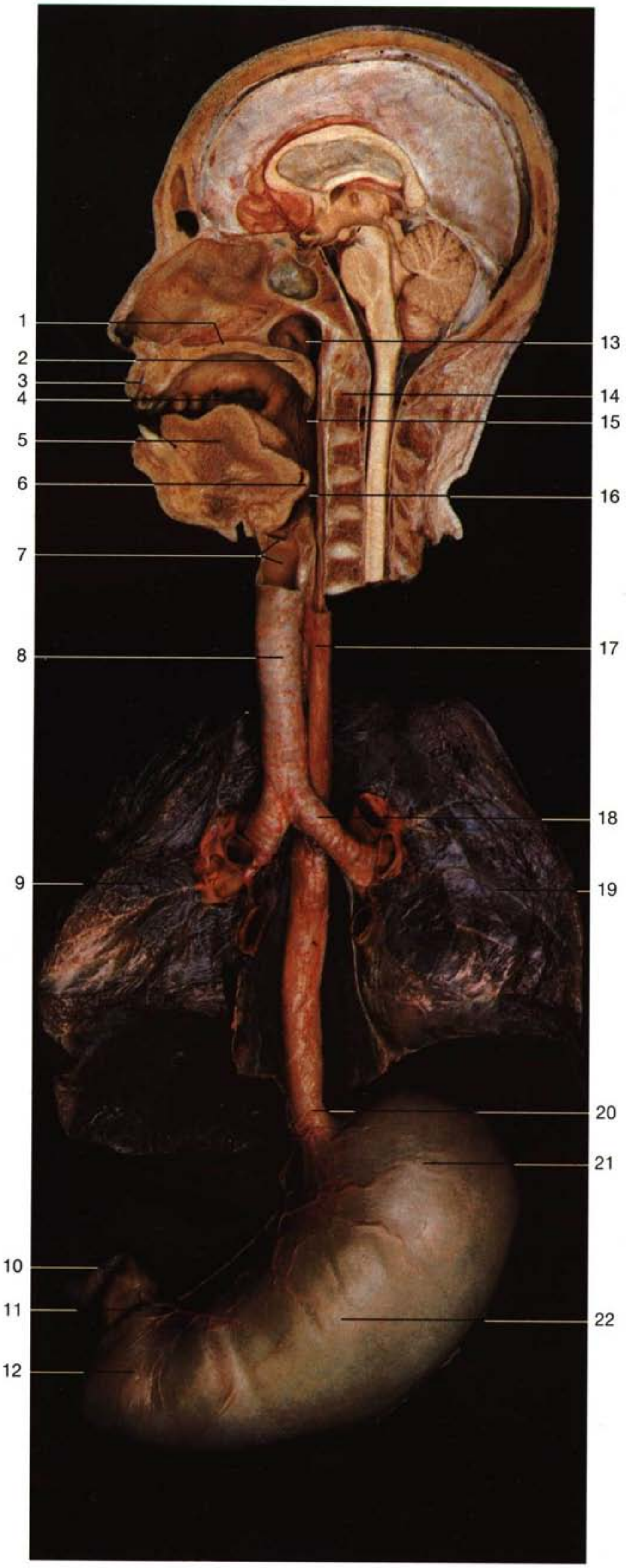

Survey of upper portion of digestive system. Oral cavity, pharynx, esophagus, and stomach.

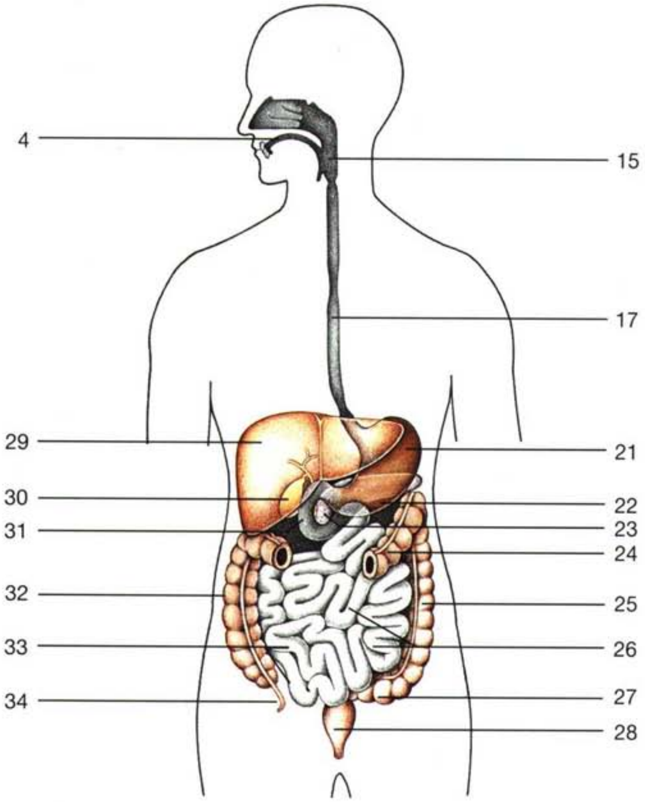

Organization of digestive system. Position of abdominal organs.

1 Hard palate
2 Soft palate with uvula
3 Vestibule of the mouth
4 Oral cavity proper
5 Tongue
6 Epiglottis
7 Vocal ligament and larynx
8 Trachea
9 Right lung
10 Superior part of duodenum
11 Pylorus
12 Pyloric antrum
13 Nasopharynx
14 Dens of axis
15 Oropharynx
16 Laryngopharynx
17 Esophagus (thoracic part)
18 Left primary bronchus
19 Left lung
20 Abdominal part of esophagus and cardia
21 Fundus of stomach
22 Body of stomach
23 Pancreas
24 Transverse colon (divided)
25 Descending colon
26 Jejunum
27 Sigmoid colon
28 Rectum
29 Liver
30 Gall bladder
31 Duodenum
32 Ascending colon
33 Ileum
34 Vermiform appendix

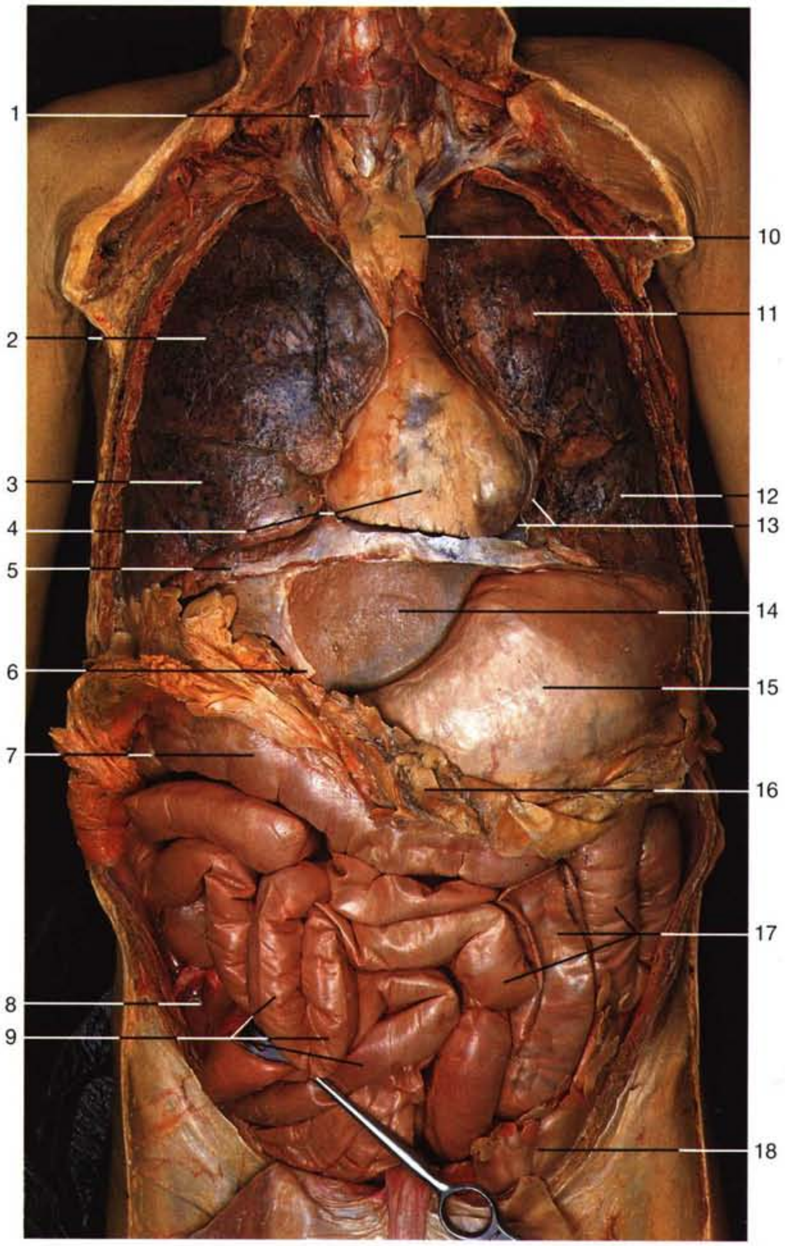

Abdominal organs in situ. The greater omentum has been partly removed or reflected.

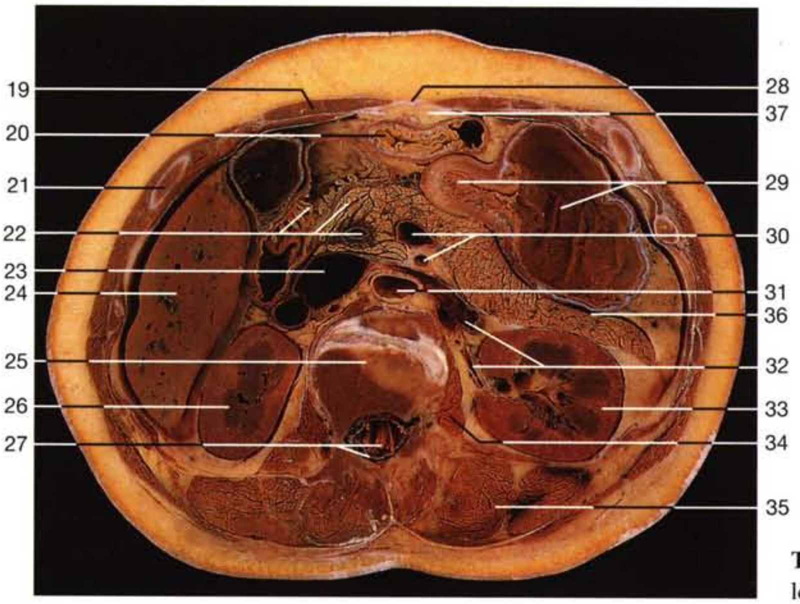

Transverse section through the abdominal cavity at the level of the second lumbar vertebra (from below).

1 Thyroid gland
2 Upper lobe of right lung
3 Middle lobe of right lung
4 Heart
5 Diaphragm
6 Round ligament of liver (ligamentum teres)
7 **Transverse colon**
8 Cecum
9 Small intestine (ileum)
10 Thymus
11 Upper lobe of left lung
12 Lower lobe of left lung
13 Pericardium (cut edge)
14 **Liver** (left lobe)
15 **Stomach**
16 Greater omentum
17 **Small intestine** (jejunum)
18 Sigmoid colon
19 Rectus abdominis muscle
20 Small intestine (section)
21 Rib
22 Common bile duct, duodenum and pancreas
23 Inferior vena cava
24 Liver
25 Body of second lumbar vertebra
26 Right kidney
27 Cauda equina and dura mater
28 Linea alba
29 Stomach and pylorus
30 Superior mesenteric artery and vein
31 Abdominal aorta
32 Left renal artery and vein
33 Left kidney
34 Psoas major muscle
35 Deep muscles of the back
36 Pancreas adjacent to lesser sac (bursa omentalis)
37 Falciform ligament with ligamentum teres

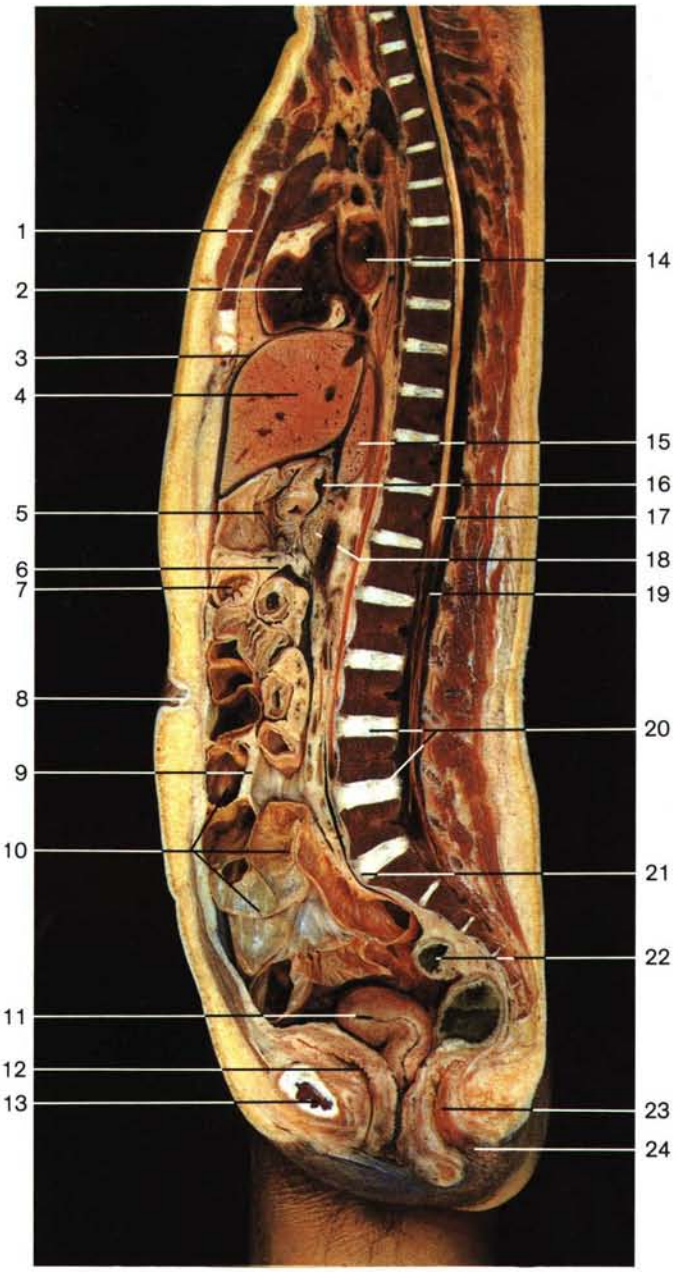
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Midsagittal section through the trunk (female).

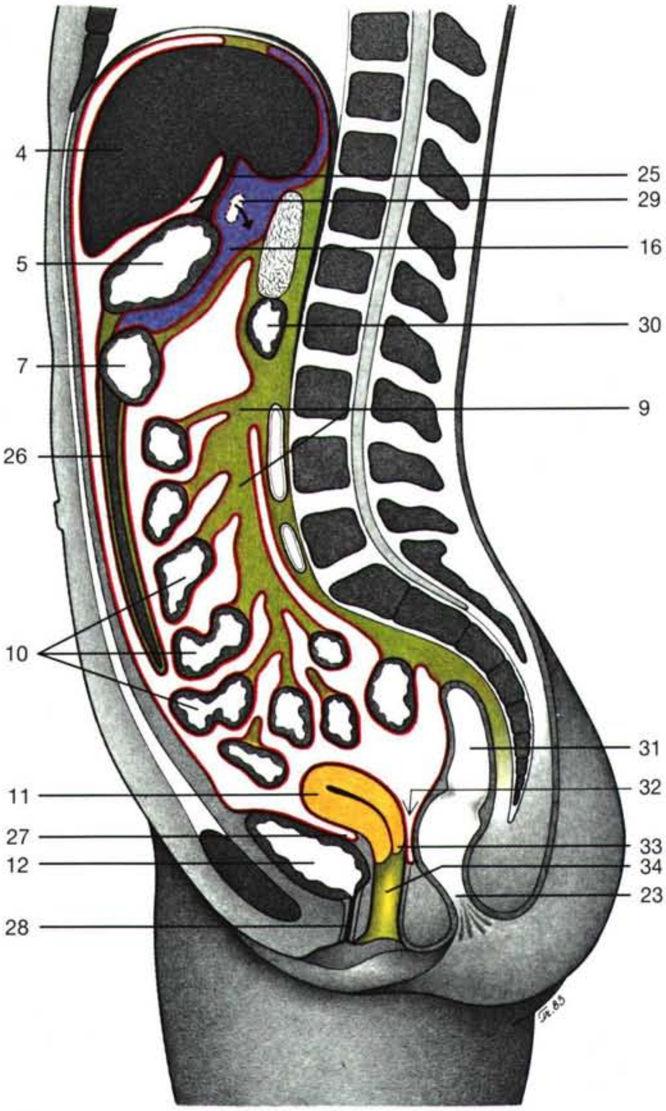
4
5
7
26
10
11
27
12
28
25
29
16
30
9
31
32
33
34
23

Midsagittal section through the tunk (female).
(Schematic drawing.)
Blue = omental bursa; red = peritoneum.

1 Sternum
2 Right ventricle of heart
3 Diaphragm
4 Liver
5 Stomach
6 Transverse mesocolon
7 Transverse colon
8 Umbilicus
9 Mesentery
10 Small intestine
11 Uterus
12 Urinary bladder
13 Pubic symphysis
14 Left atrium of heart
15 Caudate lobe of liver
16 Omental bursa or lesser sac
17 Conus medullaris
18 Pancreas
19 Cauda equina
20 Intervertebral discs (lumbar vertebral column)
21 Sacral promontory
22 Sigmoid colon
23 Anal canal
24 Anus
25 Lesser omentum
26 Greater omentum
27 Vesicouterine pouch
28 Urethra
29 Epiploic (omental) foramen
30 Duodenum
31 Rectum
32 Rectouterine pouch
33 Vaginal part of cervix of uterus
34 Vagina

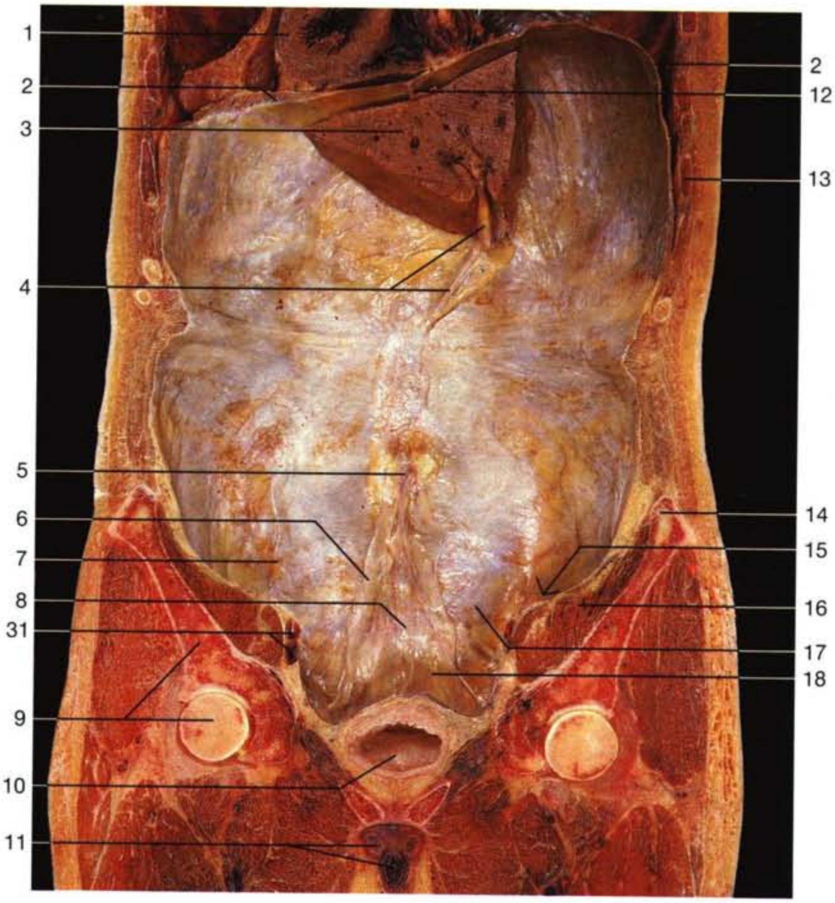

Anterior abdominal wall with pelvic cavity and thigh (frontal section, male) (internal aspect).

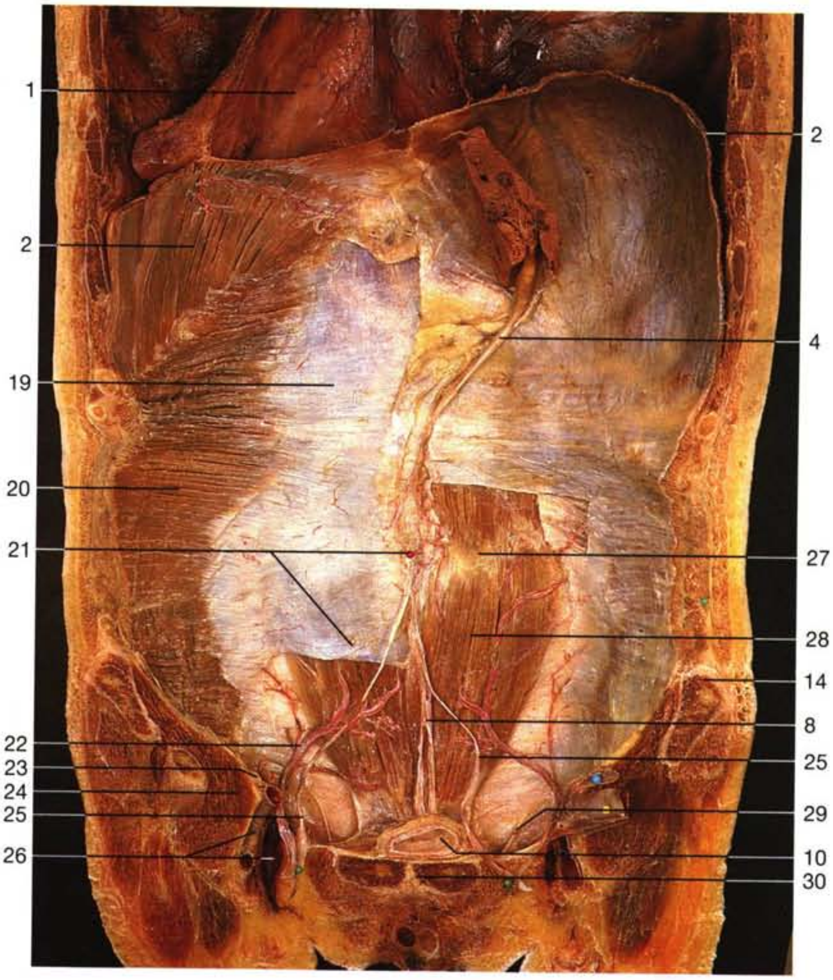

Anterior abdominal wall (male) (internal aspect). The peritoneum and parts of the posterior layer of rectus sheath have been removed. Dissection of inferior epigastric arteries and veins.

1 Left ventricle with pericardium
2 Diaphragm
3 Remnant of liver
4 Ligamentum teres (free margin of falciform ligament)
5 Site of umbilicus
6 **Medial umbilical fold** (containing the obliterated umbilical artery)
7 **Lateral umbilical fold** (containing inferior epigastric artery and vein)
8 **Median umbilical fold** (containing remnant of urachus)
9 Head of femur and pelvic bone
10 Urinary bladder
11 Root of penis
12 Falciform ligament of liver
13 Rib (divided)
14 Iliac crest (divided)
15 Site of **deep inguinal ring** and lateral inguinal fossa
16 Iliopsoas muscle (divided)
17 Medial inguinal fossa
18 Supravesical fossa
19 Posterior layer of rectus sheath
20 Transversus abdominis muscle
21 Umbilicus and arcuate line
22 **Inferior epigastric artery**
23 Femoral nerve
24 Iliopsoas muscle
25 Remnant of umbilical artery
26 Femoral artery and vein
27 Tendinous intersection of rectus abdominis muscle
28 Rectus abdominis muscle
29 Interfoveolar ligament
30 Pubic symphysis (divided)
31 External iliac artery and vein

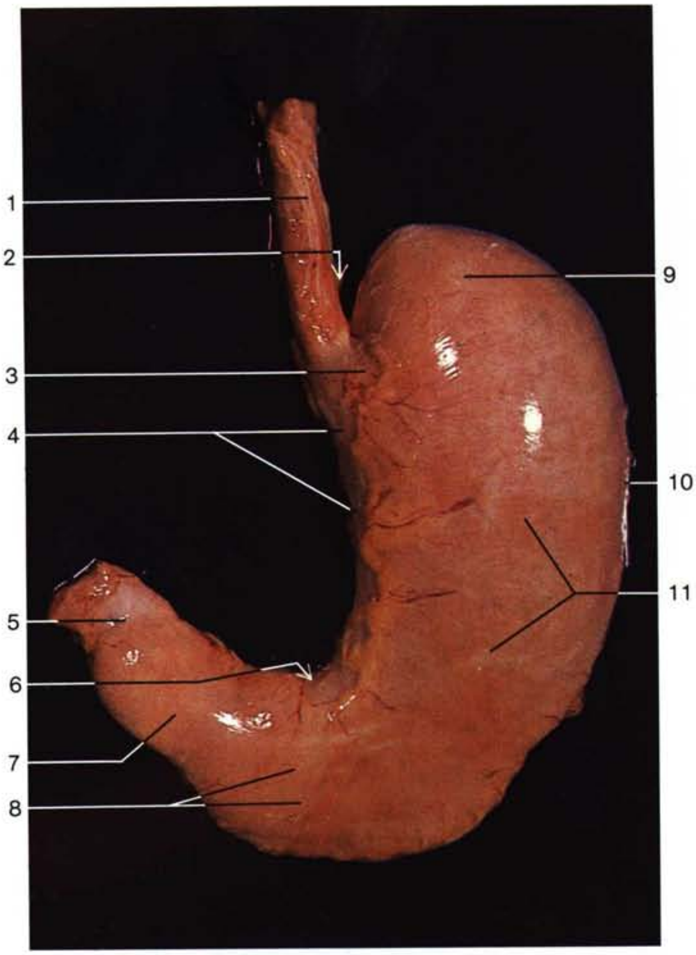

Stomach (ventral aspect).

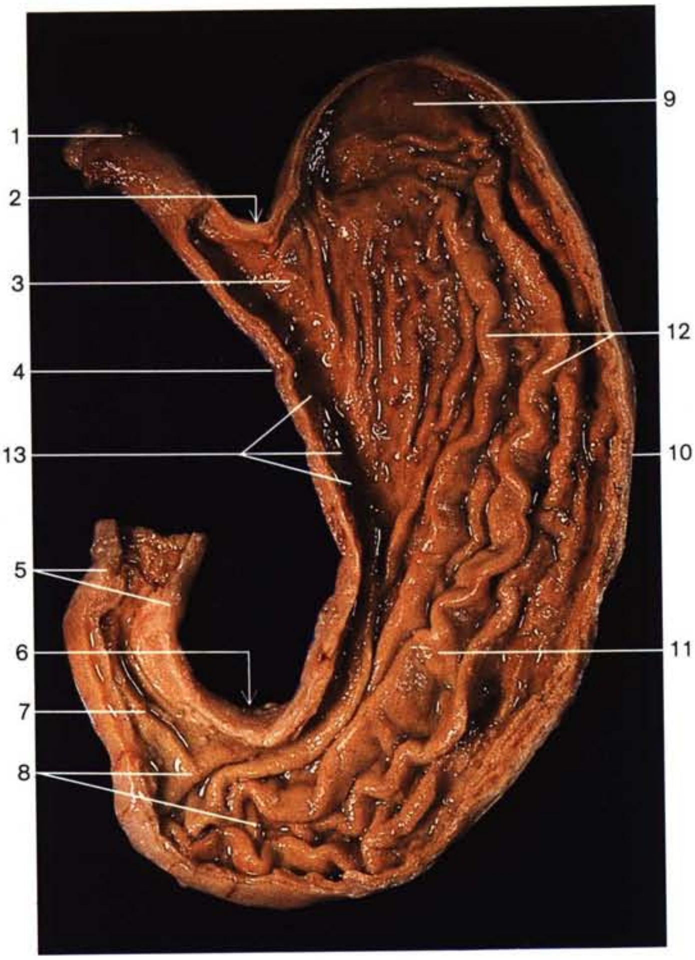

Mucosa of posterior wall of stomach (ventral aspect).

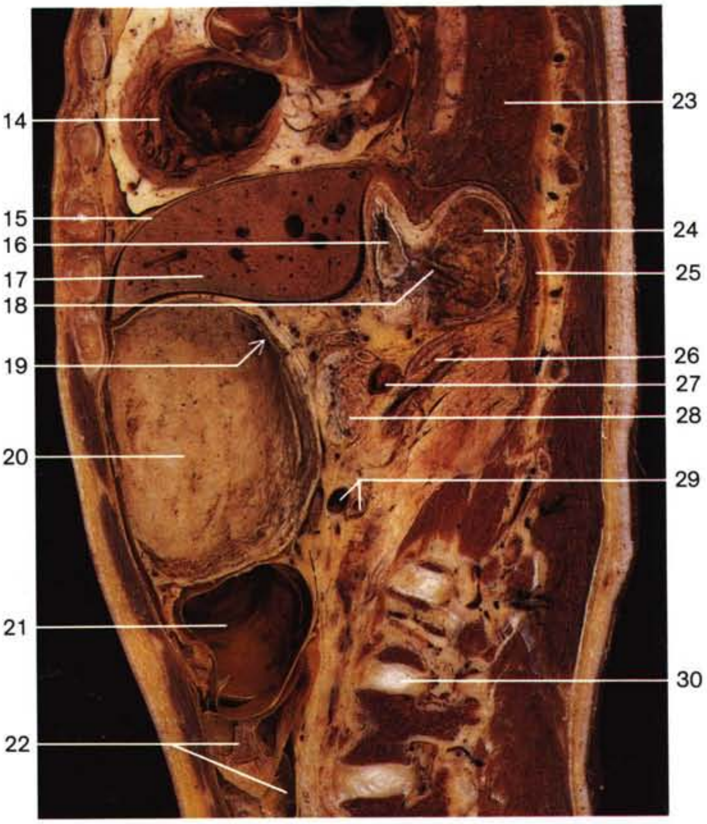

Parasagittal section through upper part of left abdominal cavity 3.5 cm lateral to median plane.

1 Esophagus
2 Cardiac notch
3 **Cardiac part of stomach**
4 Lesser curvature of stomach
5 **Pyloric sphincter**
6 Angular notch (incisura angularis)
7 **Pyloric canal**
8 Pyloric antrum
9 Fundus of stomach
10 Greater curvature of stomach
11 Body of stomach
12 Folds of mucous membrane (gastric rugae)
13 **Gastric canal**
14 Right ventricle of heart
15 Diaphragm (cut edge)
16 Abdominal portion of esophagus
17 **Liver**
18 Cardiac part of stomach (cut edge)
19 Position of pyloric canal
20 Body of stomach
21 Transverse colon
22 Small intestine
23 Lung (cut edge)
24 Fundus of stomach (section)
25 Lumbar portion of diaphragm (cut edge)
26 Suprarenal gland
27 Splenic vein
28 Pancreas
29 Superior mesenteric artery and vein
30 Intervertebral disc

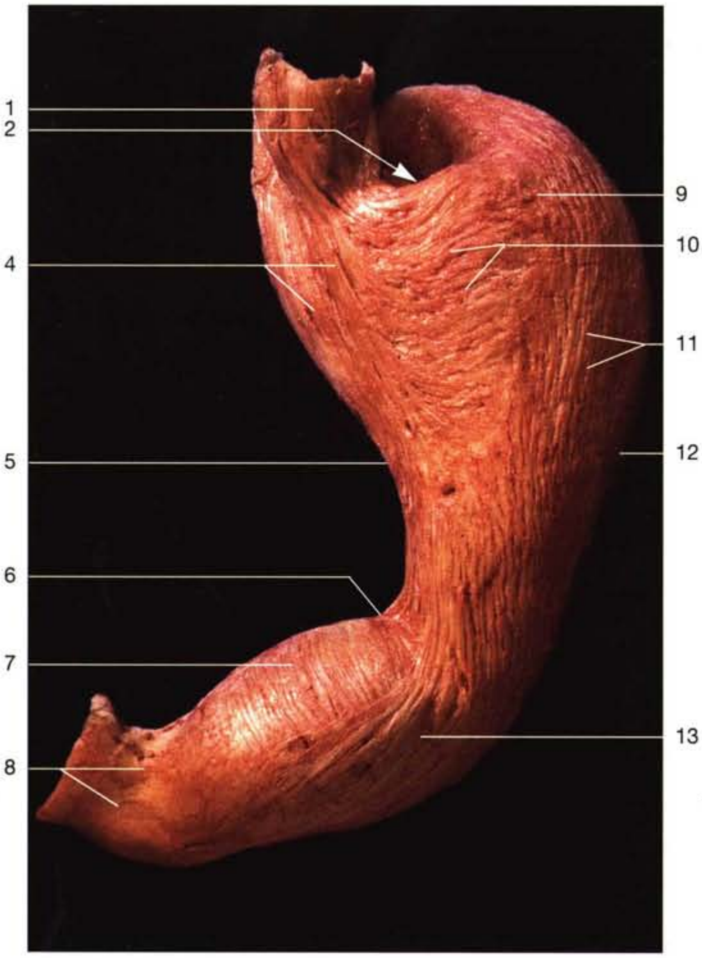

Muscular coat of stomach, outer layer (ventral aspect).

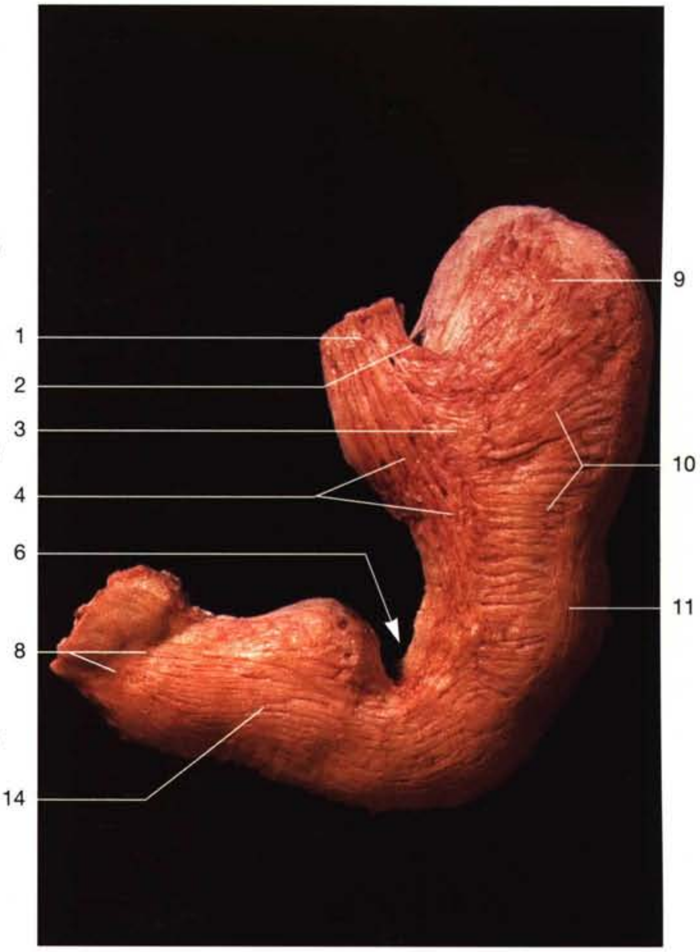

Muscular coat of stomach, middle layer (ventral aspect).

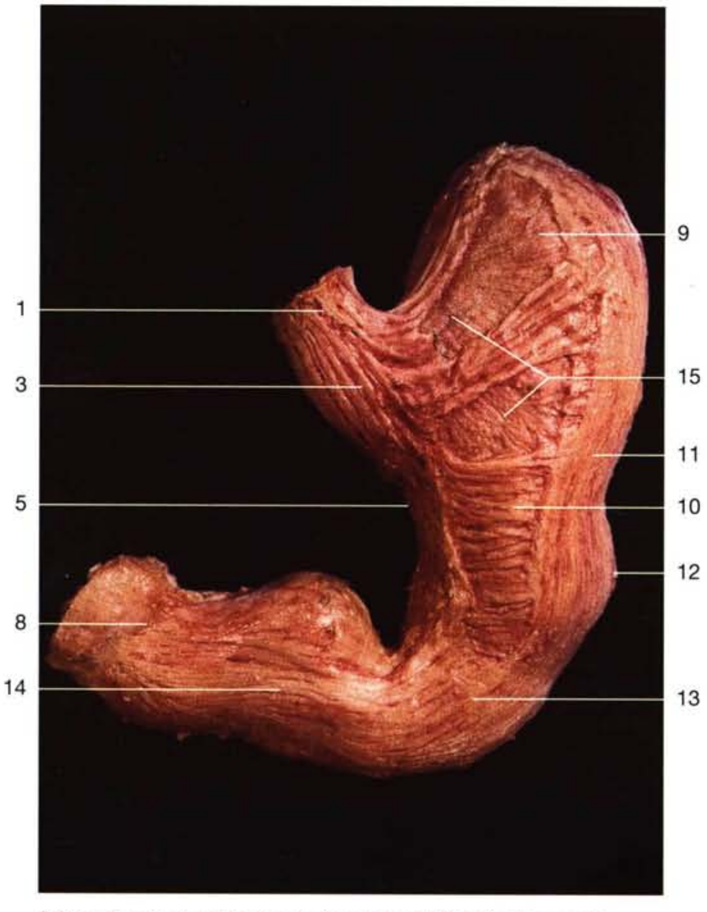

Muscular coat of stomach, inner layer (ventral aspect).

1 Esophagus (abdominal part)
2 Cardiac notch
3 Cardiac part of stomach
4 Longitudinal muscle layer at lesser curvature of stomach
5 Lesser curvature
6 Incisura angularis
7 Circular muscle layer of pyloric part of stomach
8 Pyloric sphincter muscle
9 Fundus of stomach
10 Circular muscle layer of fundus of stomach
11 Longitudinal muscle layer of greater curvature of stomach
12 Greater curvature of stomach
13 Longitudinal muscle layer (transition from body to pyloric part of stomach)
14 Pyloric part of stomach
15 Oblique muscle fibers

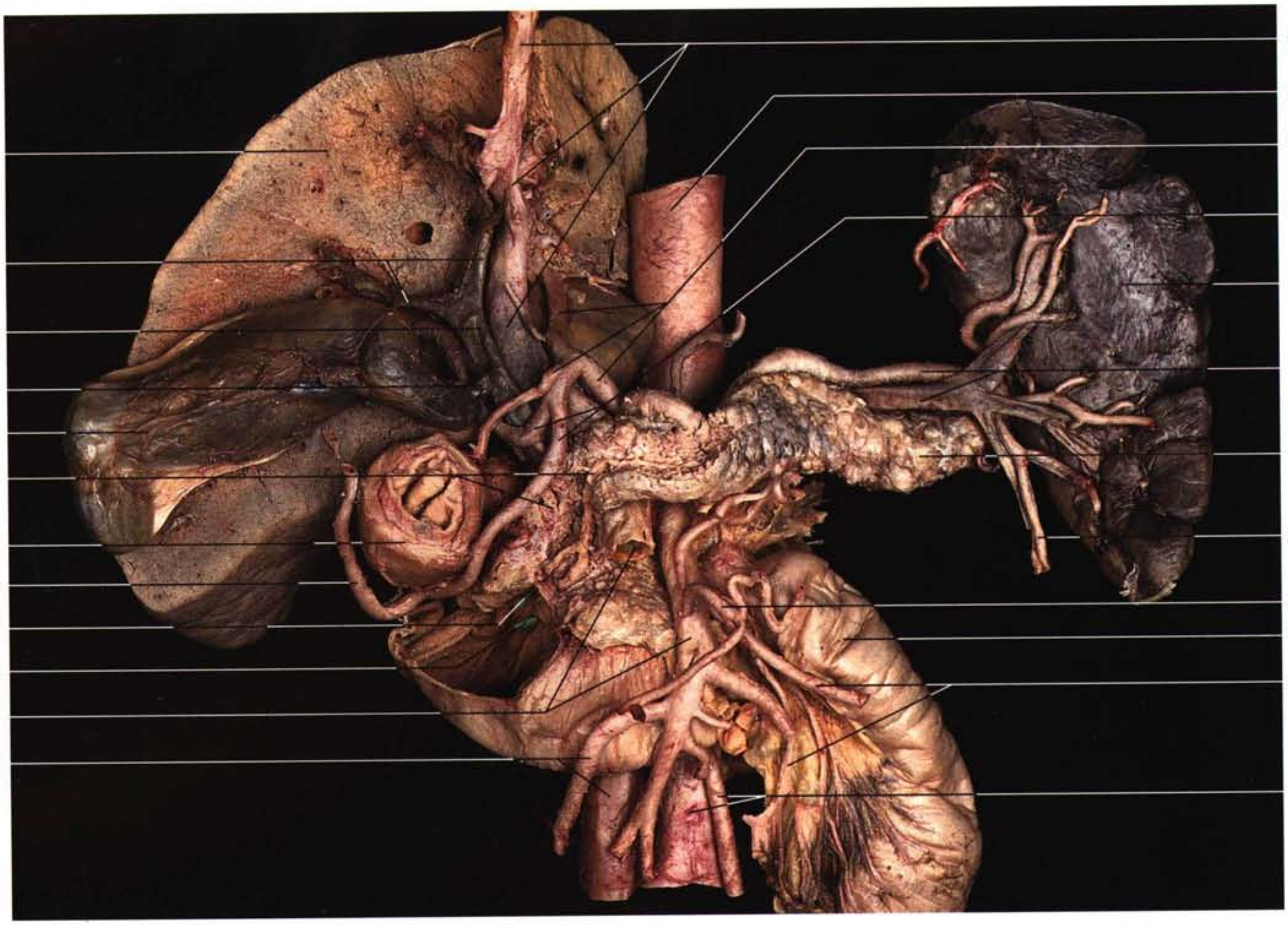

Pancreas with adjacent duodenum, spleen, liver and related arteries. The pancreatic ducts have been partly dissected. The duodenal papillae are indicated by probes. Left half of liver has been removed.

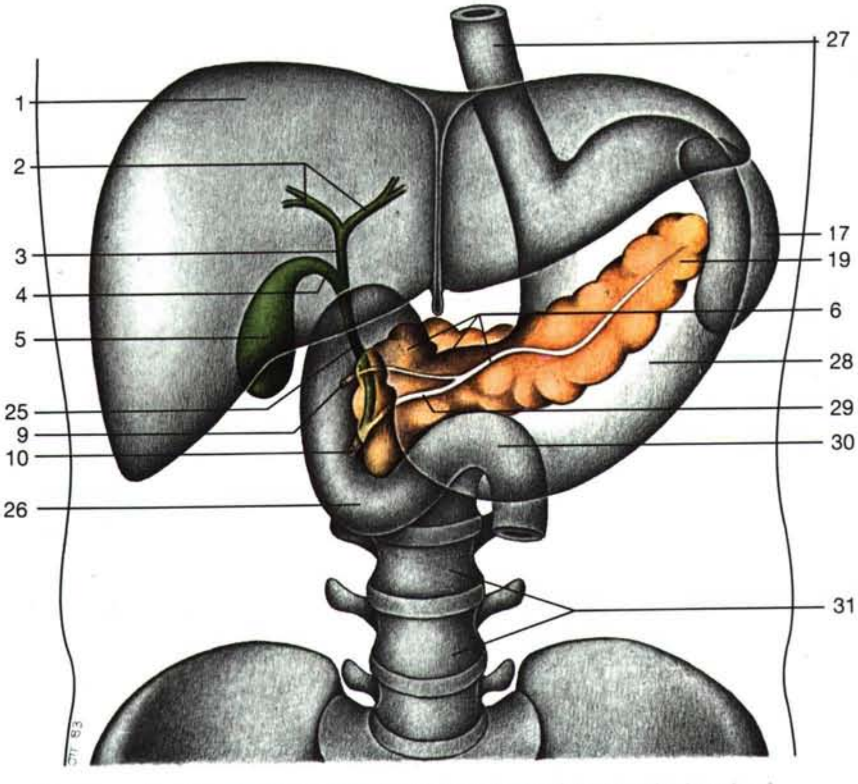

Pancreas, duodenum and extrahepatic bile ducts. (Schematic drawing.)

1 Liver
2 Left and right hepatic ducts
3 Common hepatic duct
4 **Cystic duct**
5 **Gall bladder**
6 **Pancreatic duct** and **head of pancreas**
7 Pylorus (cut)
8 Right gastro-omental (gastroepiploic) artery
9 Minor duodenal papilla (probe)
10 Major duodenal papilla (probe)
11 **Superior mesenteric artery and vein**
12 Inferior vena cava and iliocolic artery
13 **Portal vein,** lig. teres of liver and lig. venosum
14 Aorta
15 Common hepatic artery, left branch of hepatic artery proper, and right gastric artery
16 **Gastroduodenal artery** and right branch of hepatic artery proper
17 Spleen
18 Splenic artery and vein
19 **Tail of pancreas**
20 Left gastro-omental artery
21 Middle colic artery
22 Jejunum
23 Jejunal arteries
24 Aorta and inferior mesenteric artery
25 **Common bile duct**
26 Horizontal part of duodenum
27 Esophagus
28 Stomach
29 **Pancreatic duct**
30 Duodenojejunal flexure
31 Lumbal vertebrae

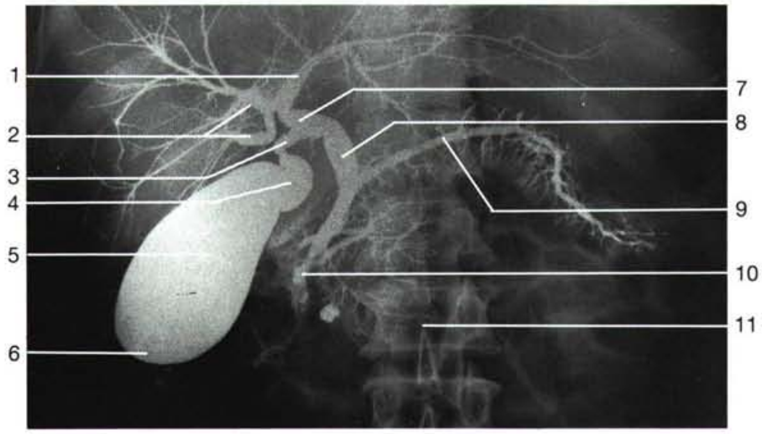

Radiograph of biliary ducts, gallbladder and pancreatic duct (anterior-posterior view).

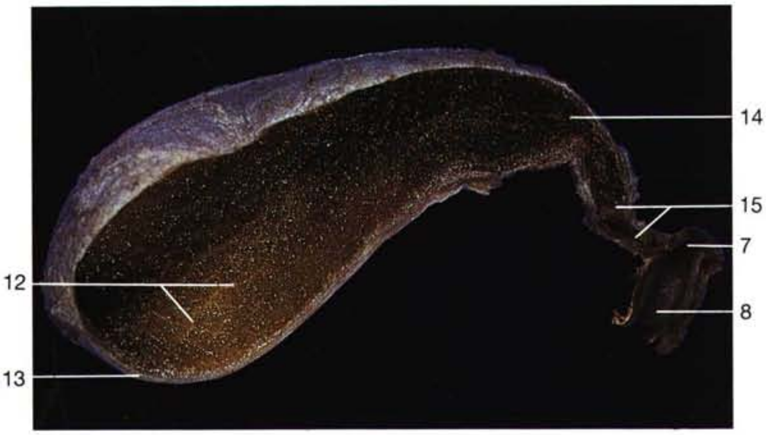

Isolated gallbladder and cystic duct (anterior aspect).
The gallbladder has been opened to display the mucous membrane.

1 Left hepatic duct
2 Right hepatic duct
3 Cystic duct
4 Neck of gallbladder
5 Body of gallbladder
6 Fundus of gallbladder
7 Common hepatic duct
8 **Common bile duct**
9 **Pancreatic duct**
10 **Greater duodenal papilla**
11 Second lumbar vertebra
12 Folds of mucous membrane of gallbladder
13 Muscular coat of gallbladder
14 Neck of gallbladder (opened)
15 Cystic duct with spiral fold
16 **Lesser duodenal papilla**
17 **Accessory pancreatic duct**
18 Uncinate process
19 Plica circularis of duodenum (Kerckrings's fold)
20 Head of pancreas
21 Body of pancreas
22 Tail of pancreas
23 Descending part of duodenum
24 Incisure of pancreas

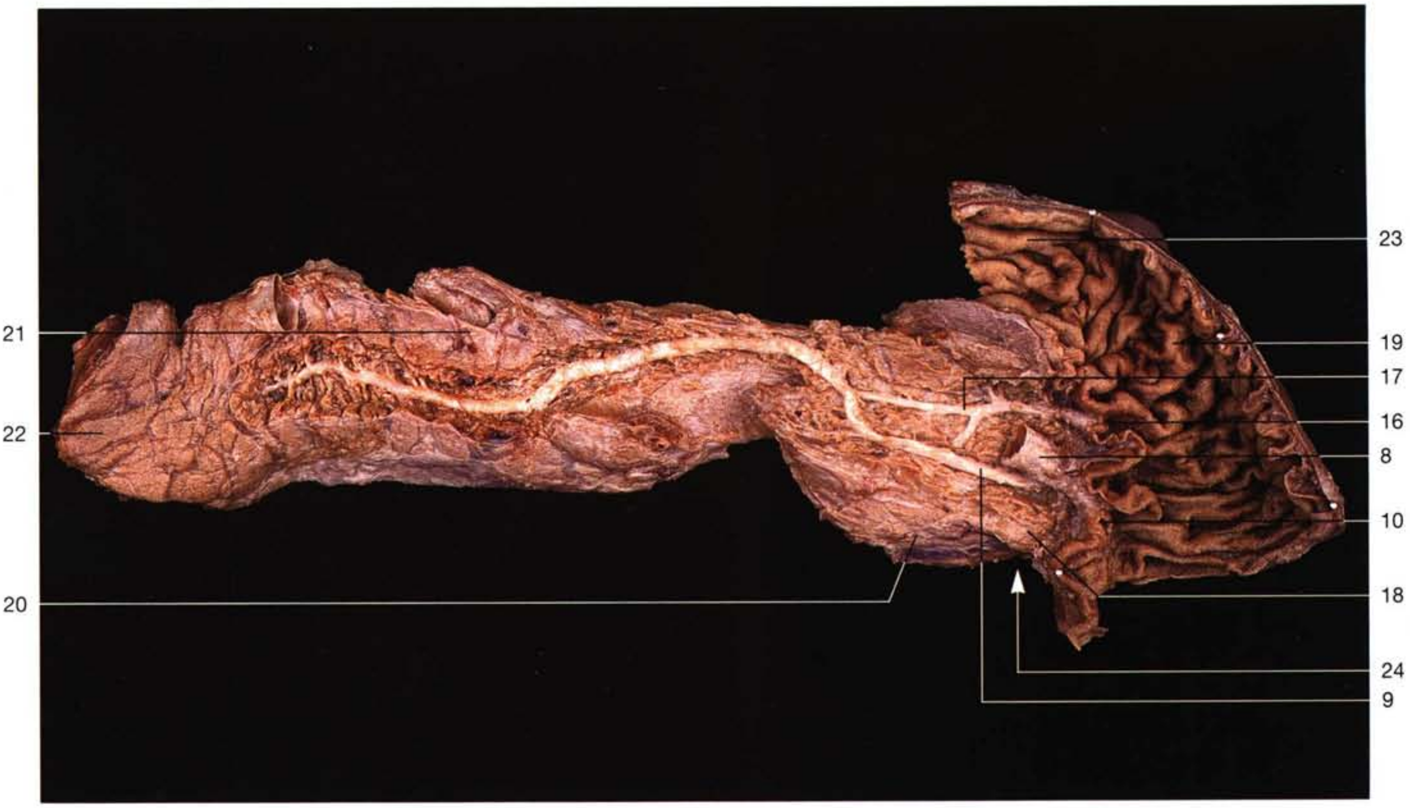

Pancreas with descending part of duodenum (posterior aspect). The duodenum was opened to display the duodenal papillae. Pancreatic duct has been dissected, the common bile duct has been divided. The sphincter of Oddi is shown.

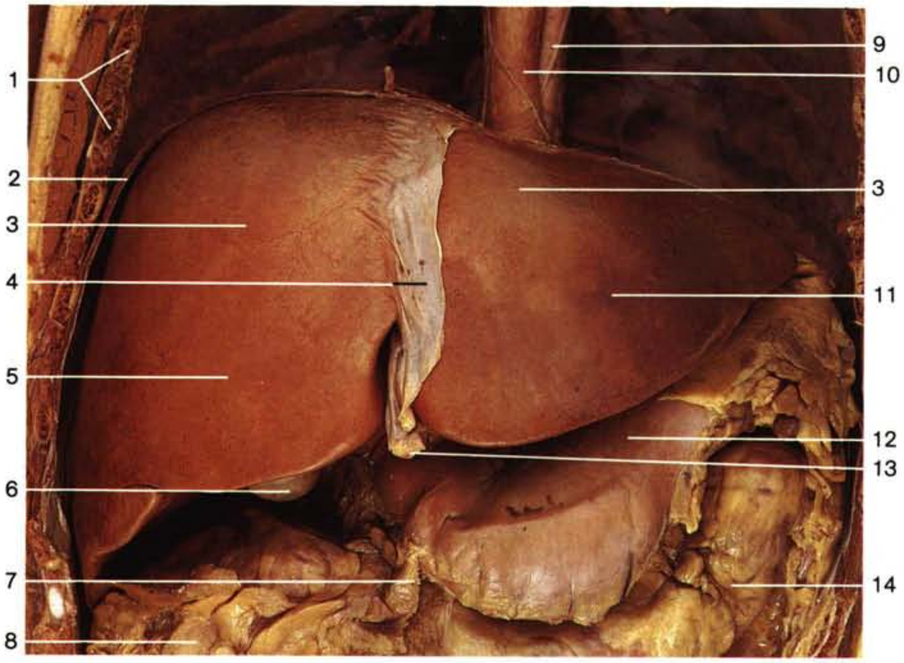

Liver in situ (ventral aspect). Part of the diaphragm has been removed.

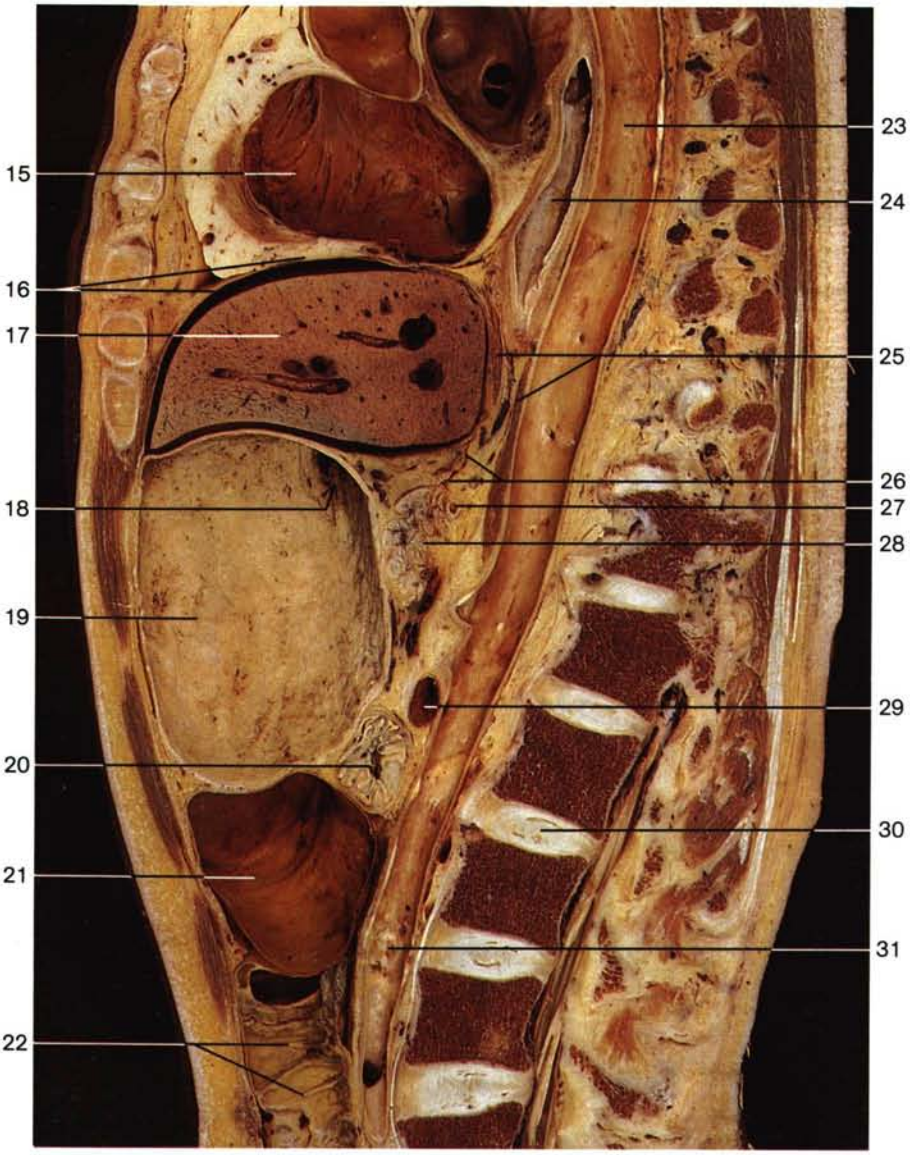

Liver in situ. Parasagittal section through the left side of the abdomen 2 cm lateral to median plane.

1 Ribs (cut edges)
2 Diaphragm
3 Diaphragmatic surface of liver
4 **Falciform ligament of liver**
5 Right lobe of liver
6 Fundus of gallbladder
7 Gastrocolic ligament
8 Greater omentum
9 Aorta
10 Esophagus
11 Left lobe of liver
12 Stomach
13 **Ligamentum teres**
14 Transverse colon
15 Right atrium of heart
16 Central tendon and sternal portion of diaphragm
17 Liver (cut edge)
18 Entrance to duodenum (pylorus)
19 Stomach
20 Duodenum
21 Transverse colon (divided, dilated)
22 Small intestine
23 Thoracic aorta (longitudinally divided)
24 Esophagus (longitudinally divided)
25 Esophageal hiatus of diaphragm
26 **Omental bursa** (lesser sac)
27 Splenic artery
28 Pancreas
29 Left renal vein
30 Intervertebral disc
31 Abdominal aorta (longitudinally divided)

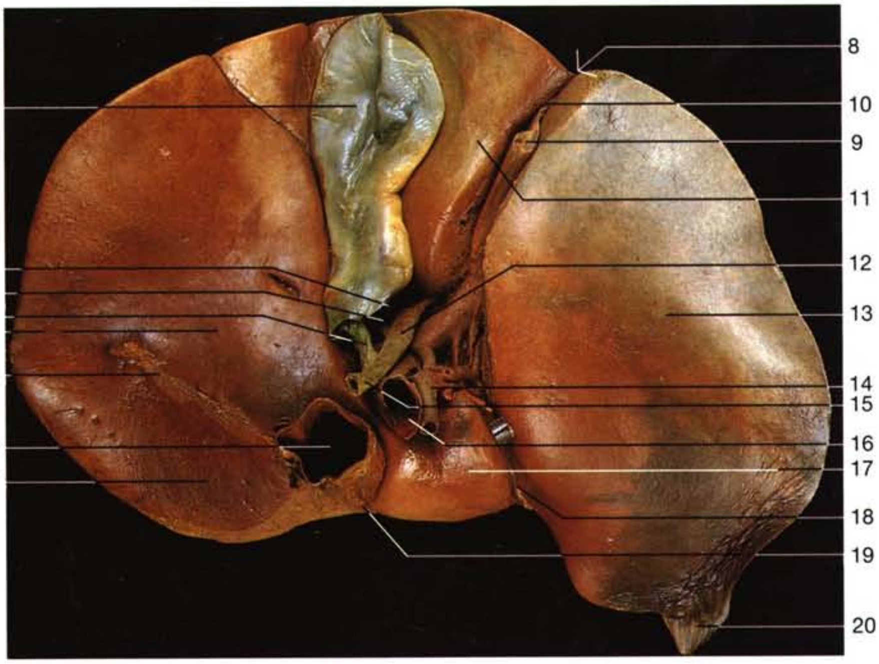

Liver (inferior aspect). Dissection of porta hepatis. Gallbladder partly collapsed. Ventral margin of liver above.

1 Fundus of **gallbladder**
2 Peritoneum (cut edges)
3 Cystic artery
4 **Cystic duct**
5 **Right lobe of liver**
6 Inferior vena cava
7 Bare area of liver
8 Notch for ligamentum teres and falciform ligament
9 Ligamentum teres
10 Falciform ligament of liver
11 **Quadrate lobe of liver**
12 Common hepatic duct
13 **Left lobe of liver**
14 Hepatic artery proper
15 **Common bile duct** } Portal triad
16 **Portal vein**
17 **Caudate lobe of liver**
18 Ligamentum venosum
19 Ligament of inferior vena cava
20 Appendix fibrosa (left triangular ligament)
21 Coronary ligament of liver
22 Hepatic veins
23 Porta hepatis

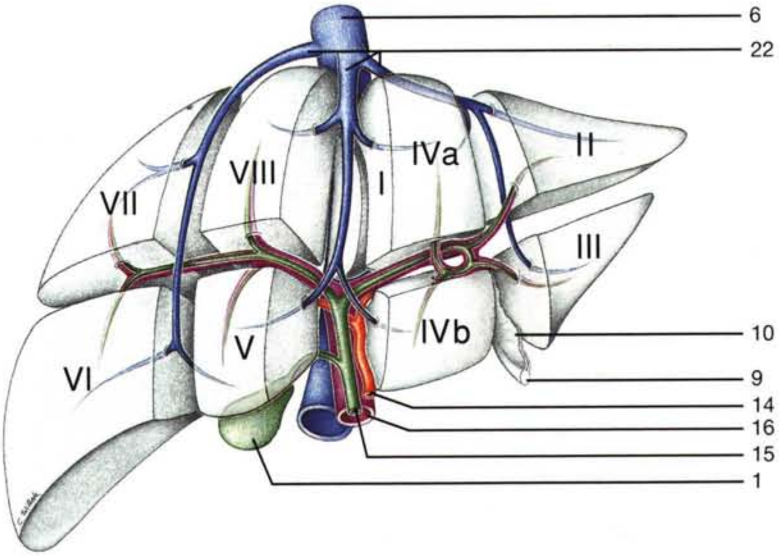

Segmentation of the liver (anterior aspect). Liver segments indicated by roman numerals.

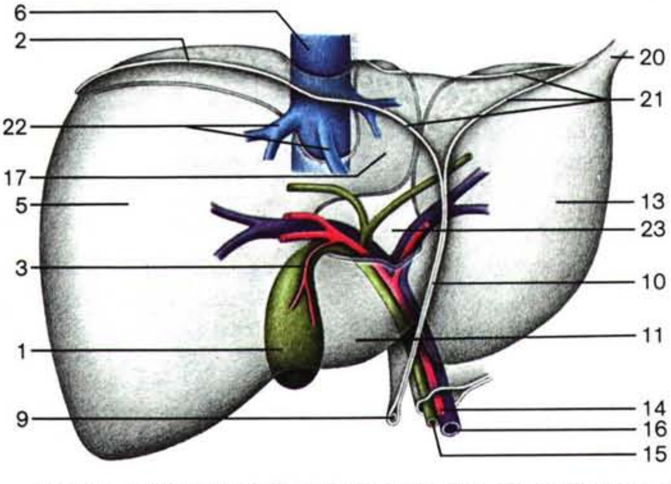

Liver (ventral aspect) (transparent drawing illustrating margins of peritoneal folds).

It should be noted that the anatomical left and right lobes of the liver do not reflect the internal distribution of the hepatic artery, portal vein, and biliary ducts. Using these structures as criteria the left lobe includes both the caudate and quadrate lobes and thus the line dividing the liver into left and right functional lobes passes through the gallbladder and inferior vena cava. The three main hepatic veins drain segments of the liver which have no visible external markings.

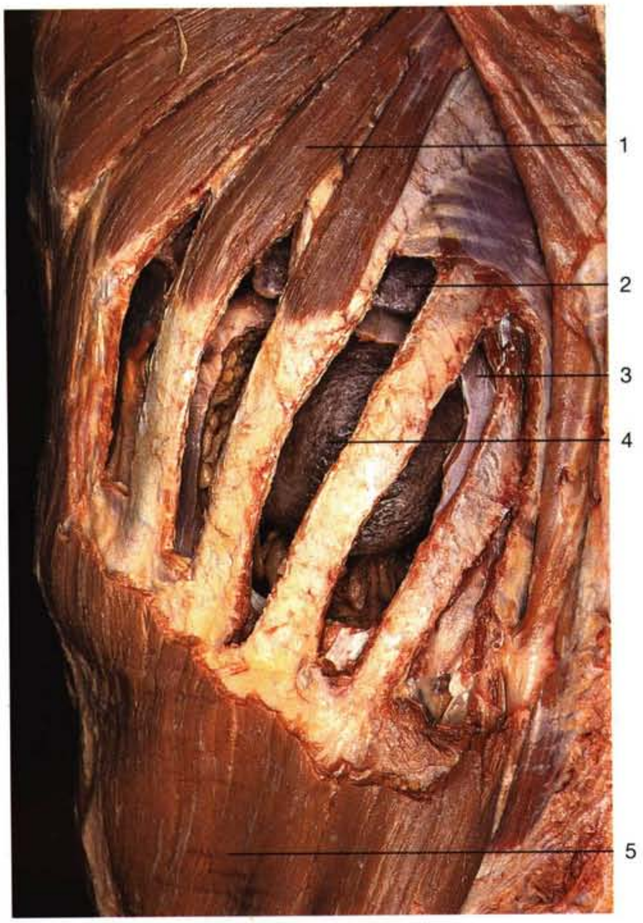

Location of the spleen in situ (left-lateral aspect). Intercostal spaces and diaphragm have been fenestrated.

1 Serratus anterior muscle
2 Left lung
3 Diaphragm
4 Spleen
5 External abdominal oblique muscle
6 Gastrosplenic ligament
7 **Splenic artery**
8 Pancreas tail
9 Superior margin of spleen
10 Anterior border of spleen
11 Inferior vena cava
12 **Hepatic veins**
13 Liver
14 **Portal vein**
15 Paraumbilical veins within falciform ligament
16 **Superior mesenteric vein**
17 Middle colic vein
18 Right colic vein
19 Ascending colon
20 Ileocolic vein
21 Vermiform appendix and appendicular vein
22 Spleen
23 Gastric and esophageal veins
24 **Splenic vein**
25 **Inferior mesenteric vein**
26 Right gastroepiploic vein
27 Ileal veins
28 Descending colon
29 Ileum
30 Sigmoid veins
31 Superior rectal veins

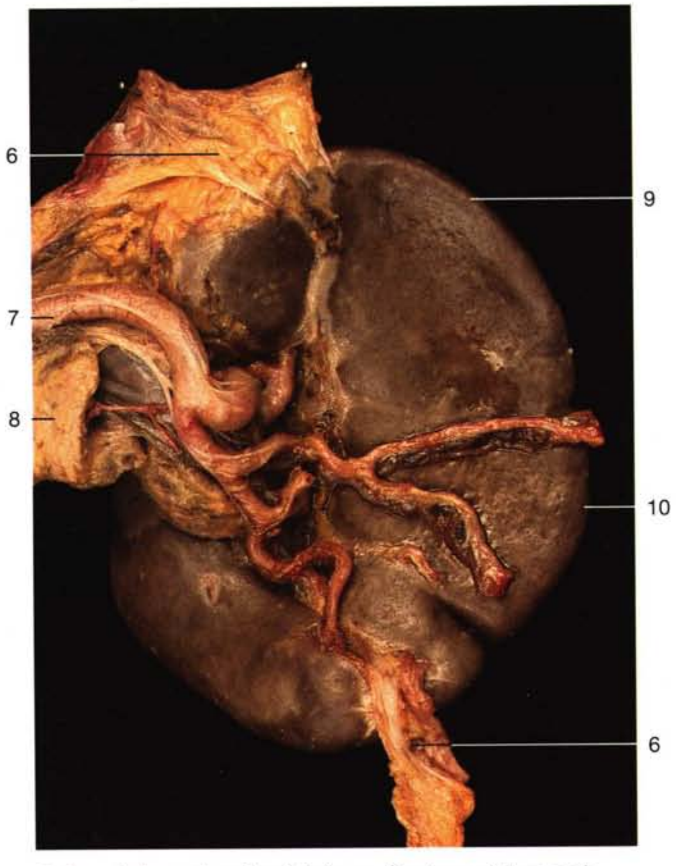

Spleen (visceral surface), hilum of spleen with vessels, nerves and ligaments.

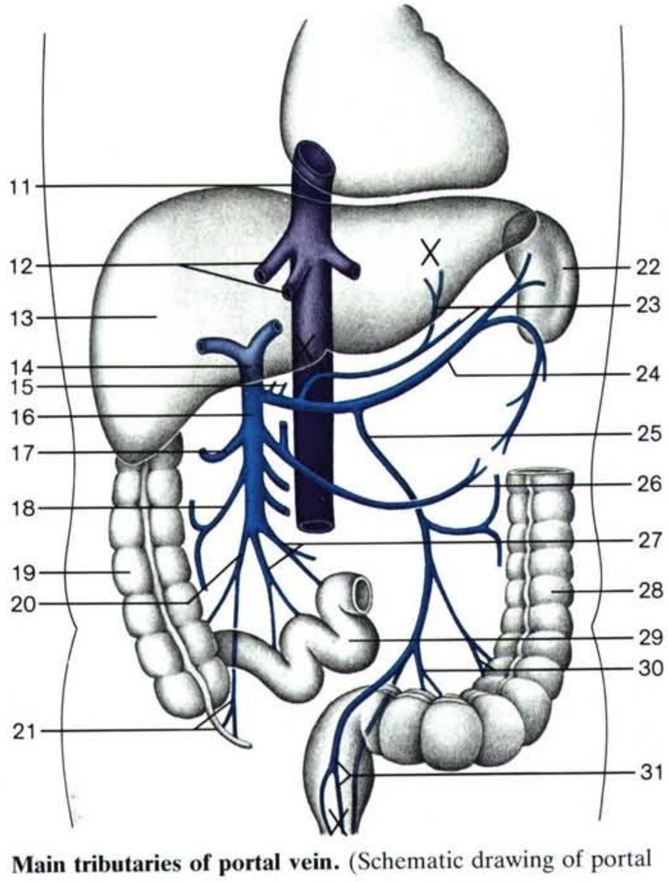

Main tributaries of portal vein. (Schematic drawing of portal circulation.) X = sites of portocaval anastomoses.

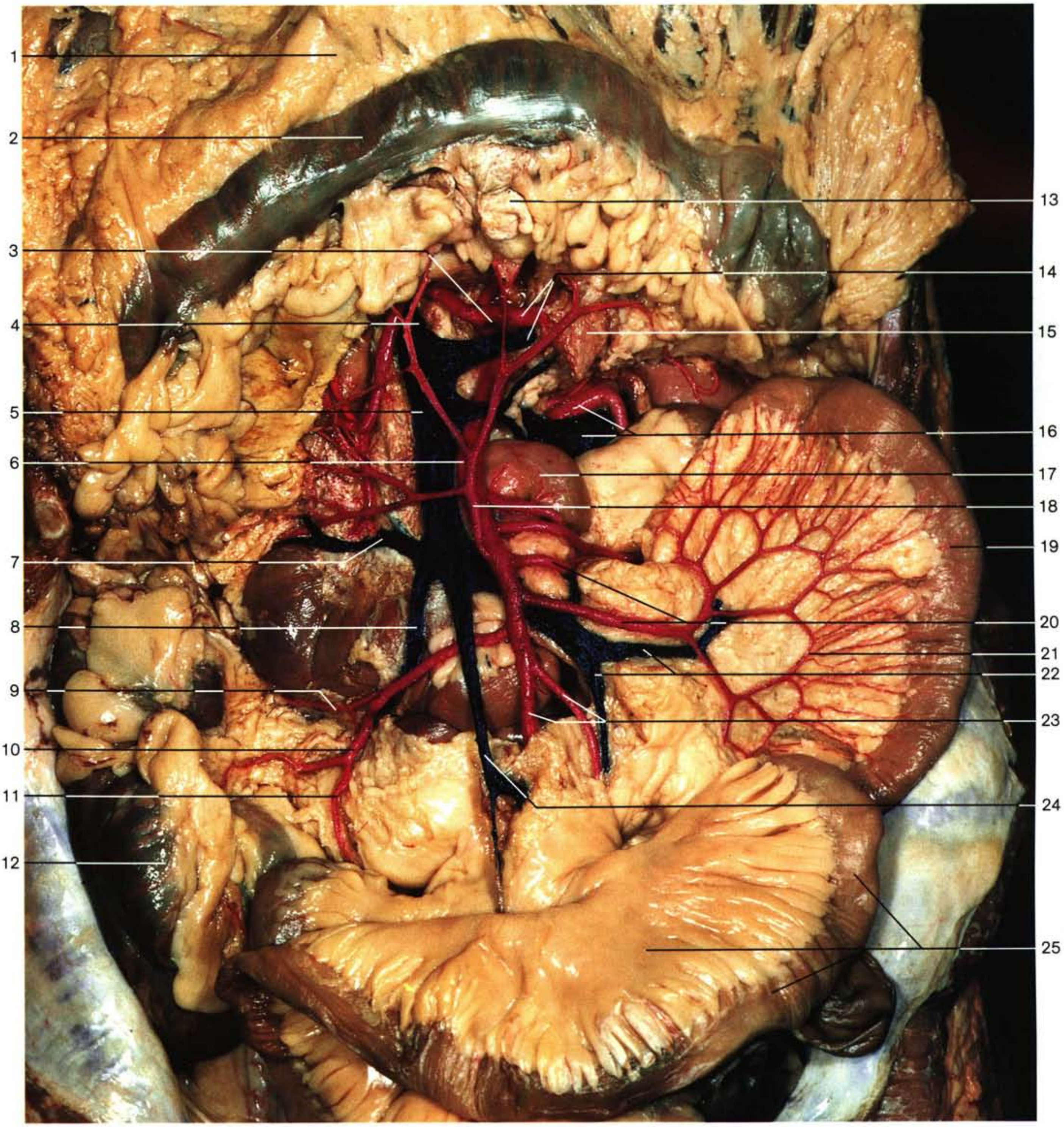

Tributaries of portal vein and branches of superior mesenteric artery (injected with colored solutions). Blue = veins; red = arteries. One layer of peritoneum has been removed to display the vascular arcades of the intestine. Part of the head of the pancreas and the mesocolon have also been removed to show the deeper vessels.

1 Greater omentum (reflected)
2 Transverse colon (raised cranially)
3 Celiac trunk
4 **Portal vein**
5 Superior mesenteric vein
6 **Superior mesenteric artery**
7 Right colic vein
8 Iliocolic vein
9 Right colic artery
10 Iliocolic artery
11 Appendicular artery
12 Cecum
13 Transverse mesocolon
14 **Splenic artery and vein**
15 Pancreas (divided)
16 Renal artery and vein
17 Duodenojejunal flexure
18 Middle colic artery
19 Jejunum
20 Jejunal arteries
21 Arterial arcades to intestine
22 Jejunal veins
23 Ileal arteries
24 Ileal vein
25 Ileum with mesentery

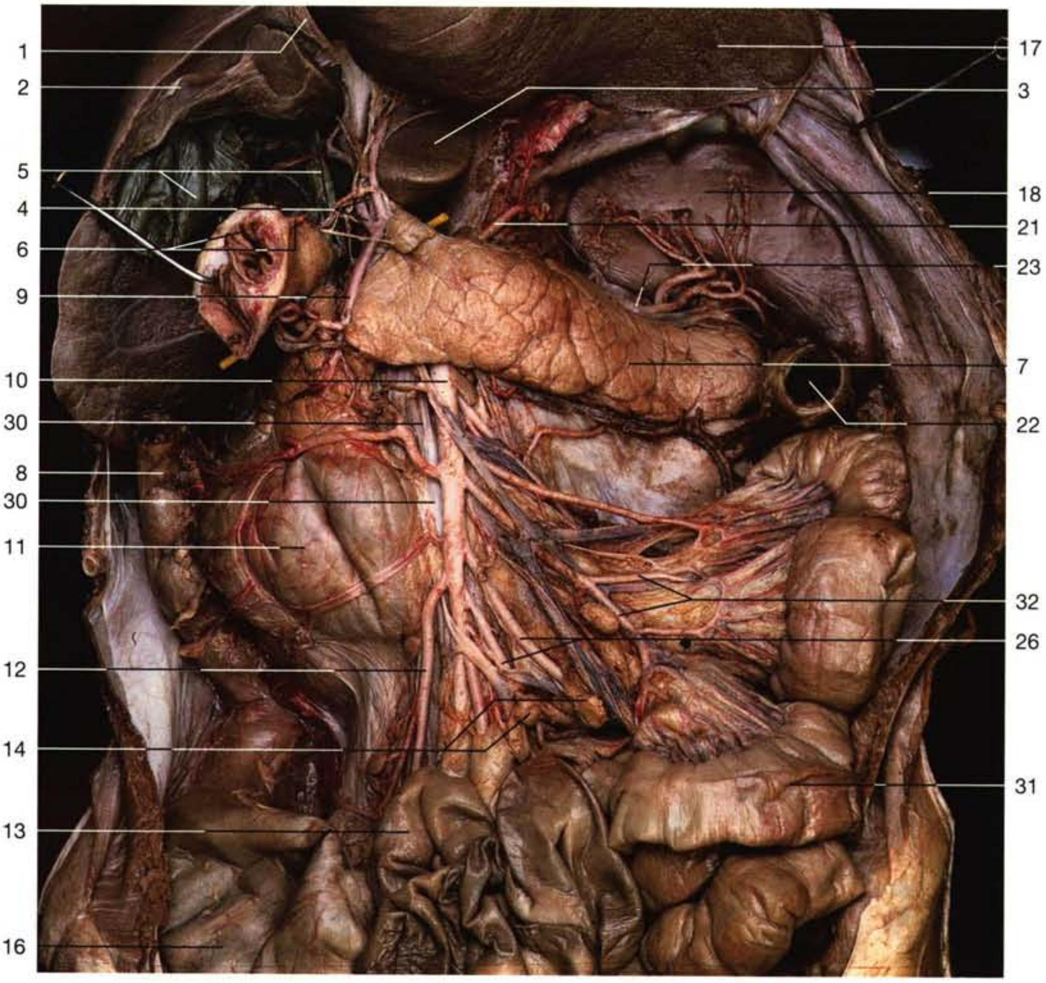

Superior mesenteric artery in relation to pancreas and duodenum. Stomach and transverse colon has been removed and the liver elevated. Note the location of the spleen. A yellow probe is inserted through the omental foramen.

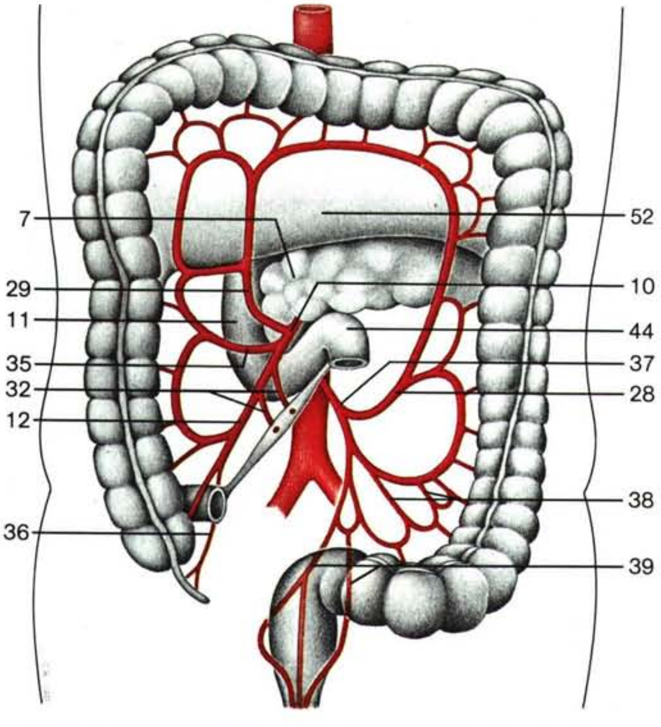

Main branches of superior and inferior mesenteric arteries. (Schematic drawing.)

1	Ligamentum teres
2	Liver
3	Caudate lobe of liver
4	Proper hepatic artery and portal vein
5	Gallbladder and common bile duct
6	Right gastric artery and pylorus
7	**Pancreas**
8	Ascending colon
9	Gastroduodenal artery
10	**Superior mesenteric artery**
11	Duodenum
12	**Iliocolic artery**
13	Terminal part of ileum
14	Mesenteric lymph nodes
15	Right common iliac artery
16	Cecum
17	Left lobe of liver
18	**Spleen**
19	Left gastro-omental (gastroepiploic) artery
20	Stomach
21	Left gastric artery
22	Left colic flexure (cut)
23	**Splenic artery**
24	Right gastro-omental (gastroepiploic) artery
25	Renal artery
26	**Ileal arteries**
27	Left kidney
28	Left colic artery
29	Middle colic artery
30	**Superior mesenteric vein**
31	Jejunum
32	**Jejunal arteries**
33	Inferior vena cava
34	Sigmoid colon
35	Right colic artery
36	Appendicular artery
37	**Inferior mesenteric artery**
38	Sigmoid arteries
39	Superior rectal artery
40	Fundus of gallbladder
41	Common bile duct
42	Portal vein
43	Beginning of jejunum
44	Duodenojejunal flexure
45	Ureter
46	Splenic artery
47	**Celiac trunk**
48	Renal vein
49	Abdominal aorta
50	Descending colon
51	Left common iliac artery
52	Transverse mesocolon

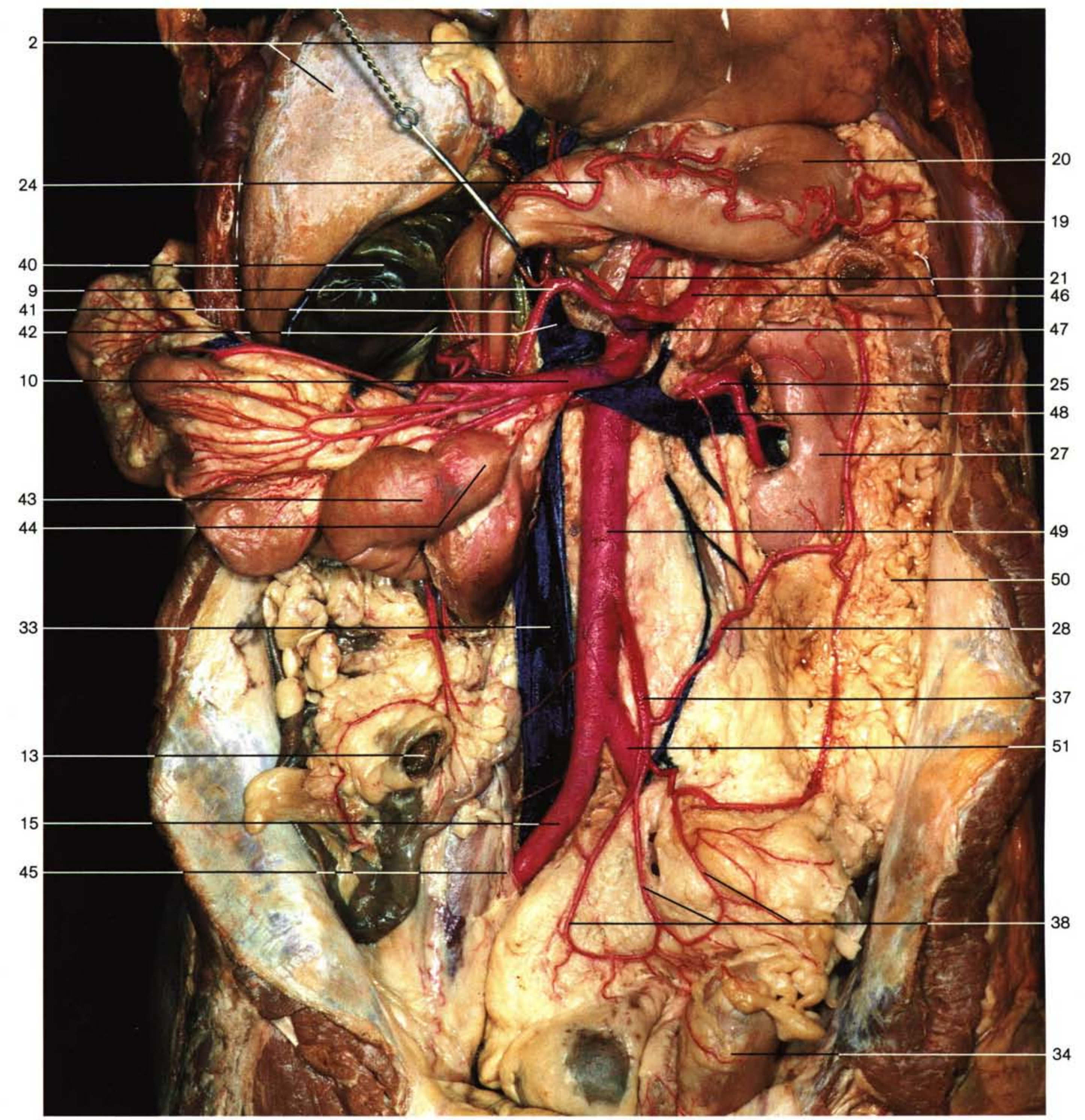

Vessels of abdominal organs; dissection of inferior mesenteric artery and celiac trunk (injected with colored solutions). Blue = veins; red = arteries. The small intestine including the duodenojejunal flexure has been reflected laterally and the stomach and liver have been raised. The peritoneum of the posterior abdominal wall has been removed to display the inferior mesenteric artery and its branches to the colon.

The **superior mesenteric artery** arises at the level of L_1 vertebra branching of the abdominal aorta. It supplies jejunum and ileum as the ascending and transverse colon. At the left colic flexure branches of the middle colic artery (from the superior mesenteric artery) and of the left colic artery (from the inferior mesenteric artery) anastomose (so-called **Riolani's anastomosis**). The arteries unite to form loops or arches called arterial arcades from which straight vessels (vasa recta) arise. The vasa recta do not anastomose within the mesentery. Thus occlusion of these vessels may lead to necrosis of the segment of bowel concerned and ileus. Within the wall of the intestine there are many anastomoses of blood vessels so that blockage of a single vessel is usually not dangerous. The arteries are paralleled by veins which convey blood to the portal vein. Closely related to the venous drainage are the lymph vessels which pass the mesenteric lymph nodes and finally enter the cisterna chyli.

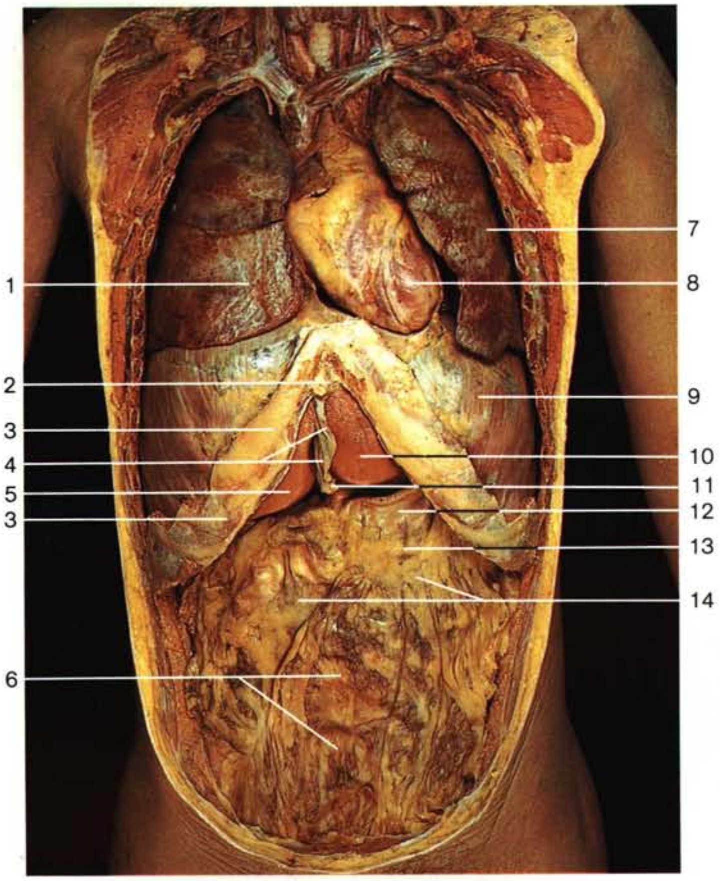

Abdominal organs. The anterior thoracic and abdominal walls have been removed.

1 Middle lobe of right lung
2 Xiphoid process
3 Costal margin
4 Falciform ligament of liver
5 Quadrate lobe of liver
6 Greater omentum
7 Upper lobe of left lung
8 Heart
9 Diaphragm
10 Left lobe of liver
11 Ligamentum teres
12 Stomach
13 Gastrocolic ligament
14 Transverse colon
15 Taenia coli
16 Appendices epiploicae
17 Cecum
18 Taenia coli
19 Ileum
20 Transverse mesocolon
21 Jejunum
22 Sigmoid colon
23 Position of root of mesentery
24 Vermiform appendix
25 Duodenojejunal flexure
26 Mesentery

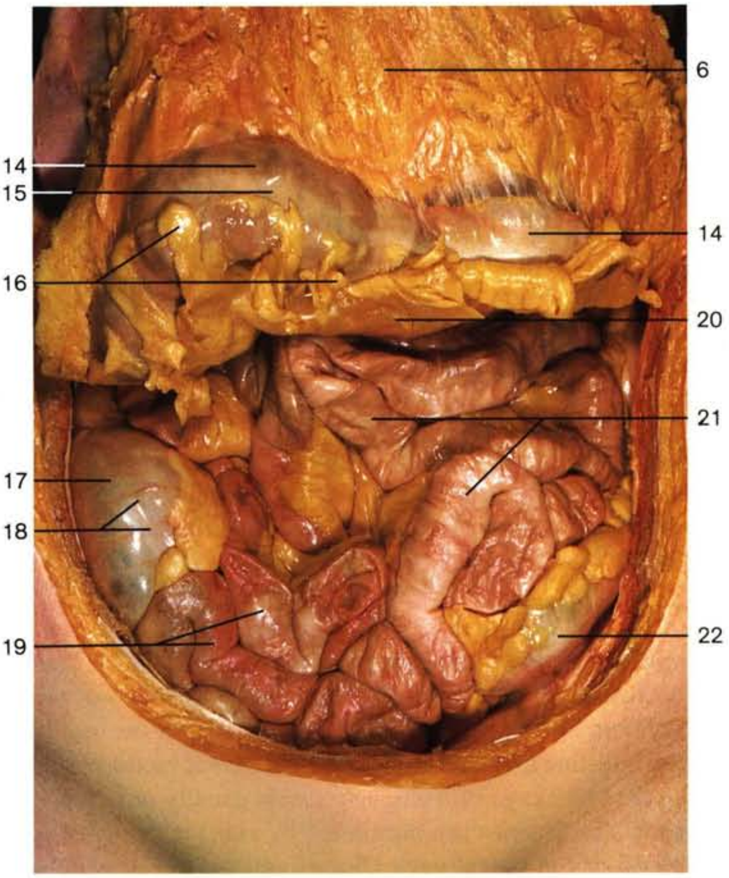

Abdominal organs. The greater omentum which is fixed to the transverse colon has been raised.

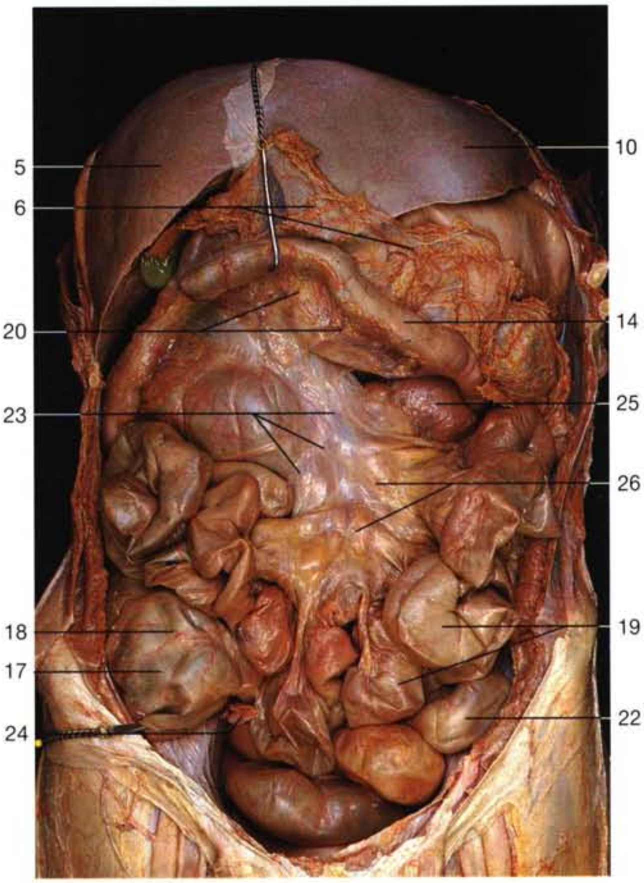

Abdominal organs (anterior aspect). The transverse colon has been reflected.

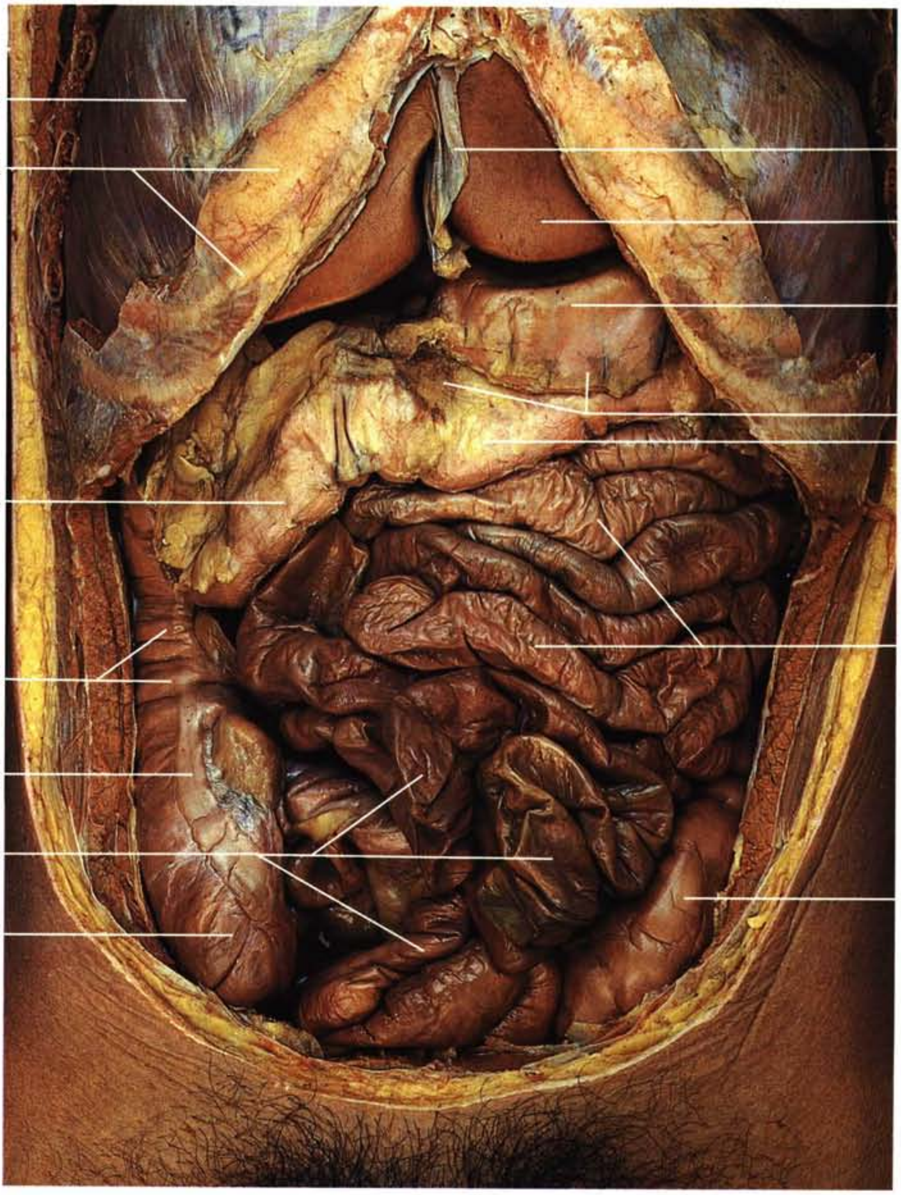

1 Diaphragm
2 Costal margin
3 Transverse colon
4 Ascending colon with haustra
5 Free taenia of cecum
6 **Ileum**
7 **Cecum**
8 **Falciform ligament of liver**
9 Liver
10 Stomach
11 Gastrocolic ligament
12 Jejunum
13 Sigmoid colon
14 **Vermiform appendix**
15 Terminal ileum
16 **Mesoappendix**
17 Mesentery

Abdominal organs in situ. The greater omentum has been removed.

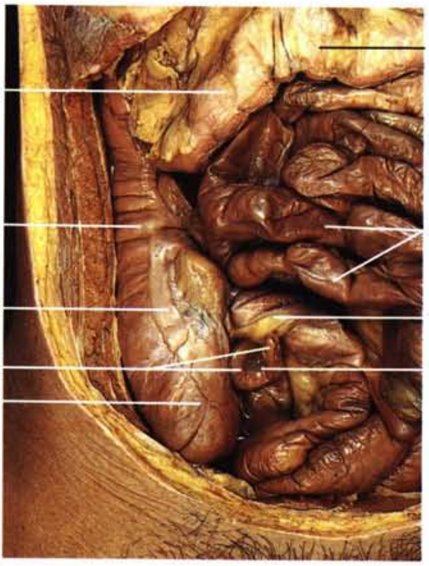

Ascending colon, cecum and vermiform appendix (detail of the preceding figure).

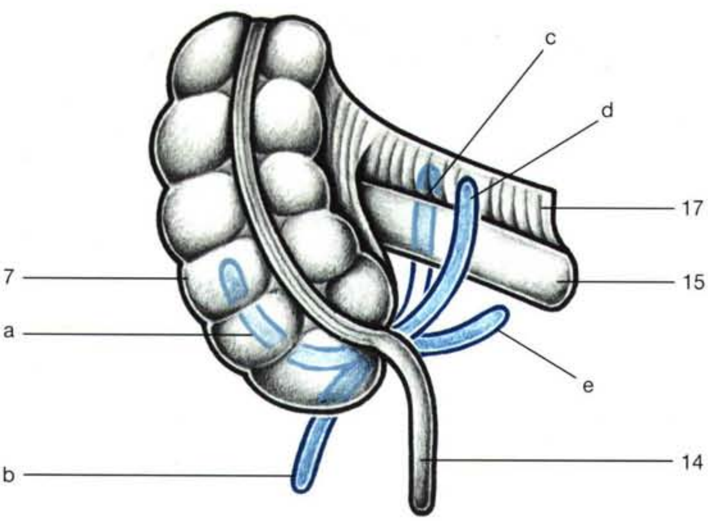

Variations in the position of the vermiform appendix. a = retrocecal; b = paracolic; c = retroileal; d = preileal; e = subcecal.

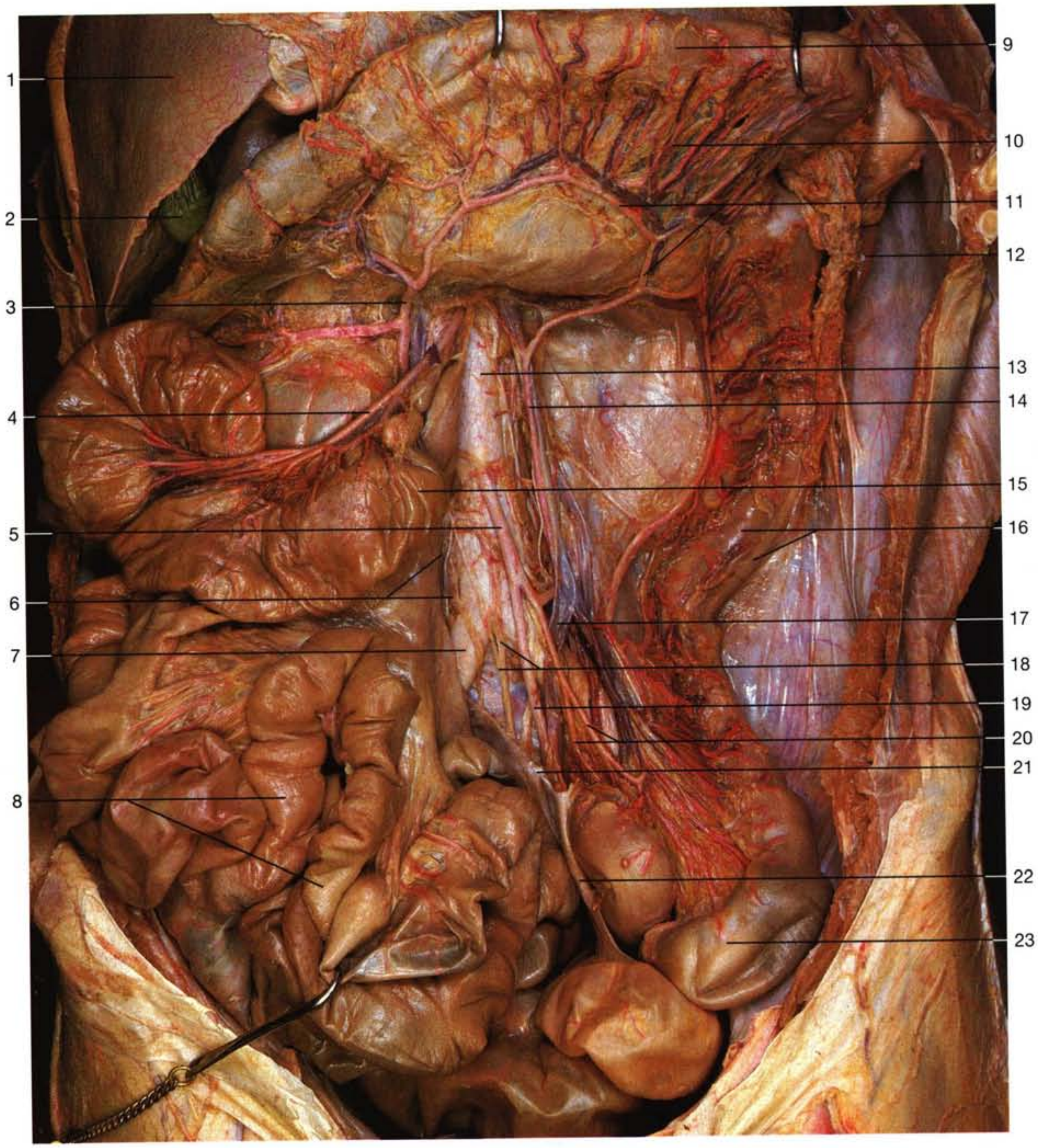

Abdominal organs. Dissection of inferior mesenteric artery and autonomic plexus. The transverse colon with mesocolon has been raised and the small intestine reflected.

1	Liver	12	Spleen
2	Gallbladder	13	**Abdominal aorta**
3	**Middle colic artery**	14	**Left colic artery**
4	Jejunal artery	15	Duodenojejunal flexure
5	**Inferior mesenteric artery**	16	Descending colon (free taenia of colon)
6	Sympathetic nerves and ganglia	17	Inferior mesenteric vein
7	Right common iliac artery	18	Superior hypogastric plexus
8	Small intestine (ileum)	19	**Superior rectal artery**
9	Transverse colon (reflected)	20	Sigmoid arteries
10	Transverse mesocolon	21	Peritoneum (cut edge)
11	Anastomosis between middle and left colic artery	22	Sigmoid mesocolon
		23	Sigmoid colon

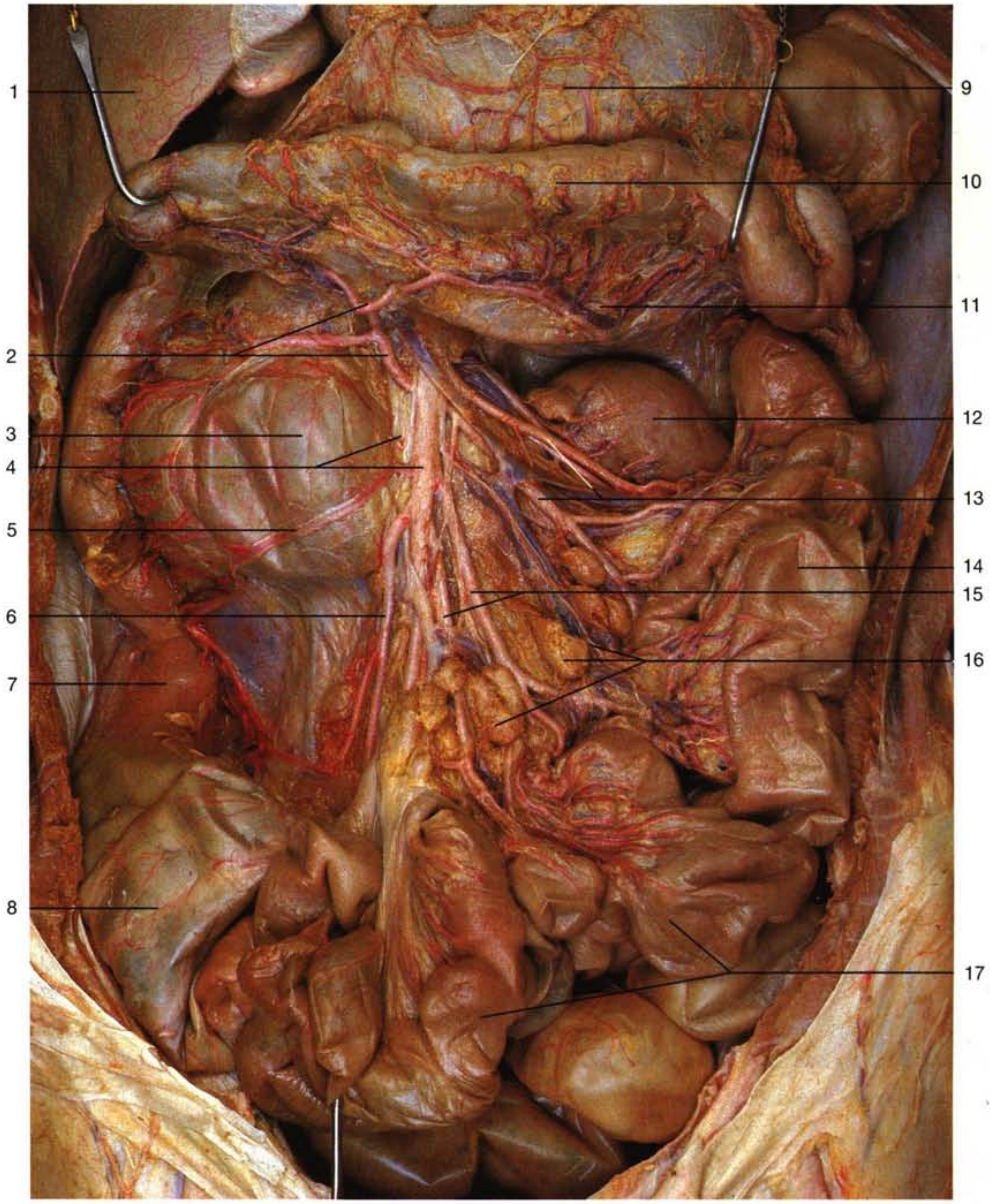

1 Liver
2 Middle colic artery
3 Horizontal part of duodenum (extended)
4 **Superior mesenteric artery and vein**
5 **Right colic artery**
6 **Iliocolic artery**
7 Ascending colon
8 Cecum
9 Greater omentum (reflected)
10 Transverse colon
11 Transverse mesocolon
12 Duodenojejunal flexure
13 **Jejunal arteries**
14 Jejunum
15 **Ileal arteries**
16 **Mesenteric lymph nodes and lymph vessels**
17 Ileum
18 Abdominal aorta
19 Inferior vena cava
20 Stomach
21 Spleen
22 Splenic artery
23 Head of pancreas
24 Superior mesenteric artery

Abdominal organs. Superior mesenteric artery. Mesenteric lymph nodes. Transverse colon reflected.

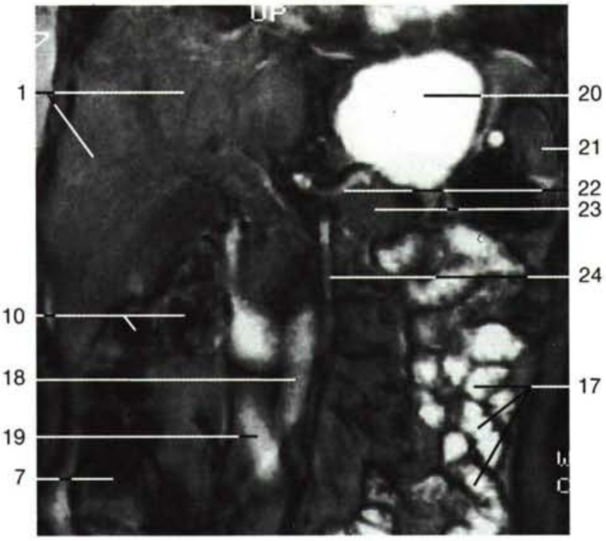

Frontal section through the abdominal cavity. (MR-Scan: the intestinal tract and vessels are filled with a paramagnetic substance [Gadolinium]; courtesy of Dr. W. Rödl, Erlangen.)

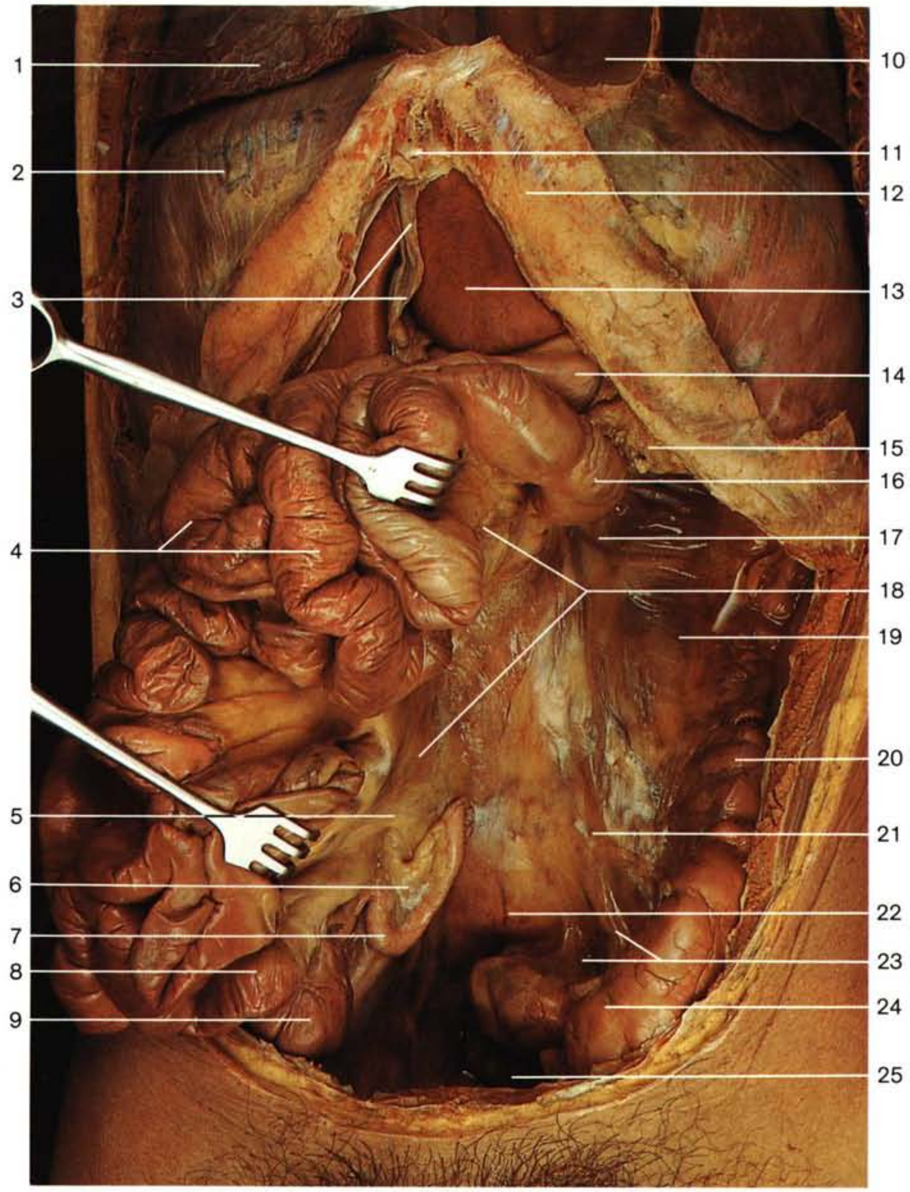

Abdominal cavity. Mesenteries. The small intestine has been reflected laterally to demonstrate the mesentery.

1 Lung
2 Diaphragm
3 Falciform ligament of liver
4 Jejunum
5 Ileocecal fold
6 Mesoappendix
7 Vermiform appendix
8 **Ileocecal junction**
9 **Cecum**
10 Pericardial sac
11 Xiphoid process
12 Costal margin
13 Liver
14 Stomach
15 Transverse colon
16 **Duodenojejunal flexure**
17 Inferior duodenal fold
18 Mesentery
19 Position of left kidney
20 Descending colon
21 Position of left common iliac artery
22 Sacral promontory
23 Sigmoid mesocolon
24 Sigmoid colon
25 Rectum
26 Beginning of jejunum
27 Peritoneum of posterior abdominal wall
28 Transverse mesocolon
29 Superior duodenal fold
30 Superior duodenal recess
31 Retroduodenal recess
32 Free taenia of ascending colon
33 **Ileocecal valve**
34 Frenulum of ileocecal valve
35 Orifice of vermiform appendix (probe)
36 **Ileocolic artery**
37 Vermiform appendix with appendicular artery
38 Ascending colon

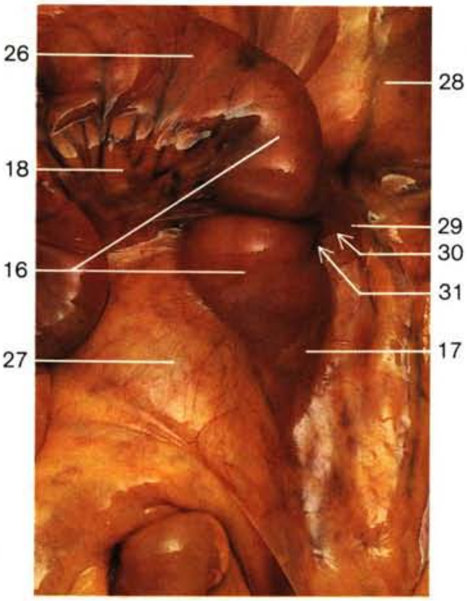

Duodenojejunal flexure (enlargement of preceding figure).

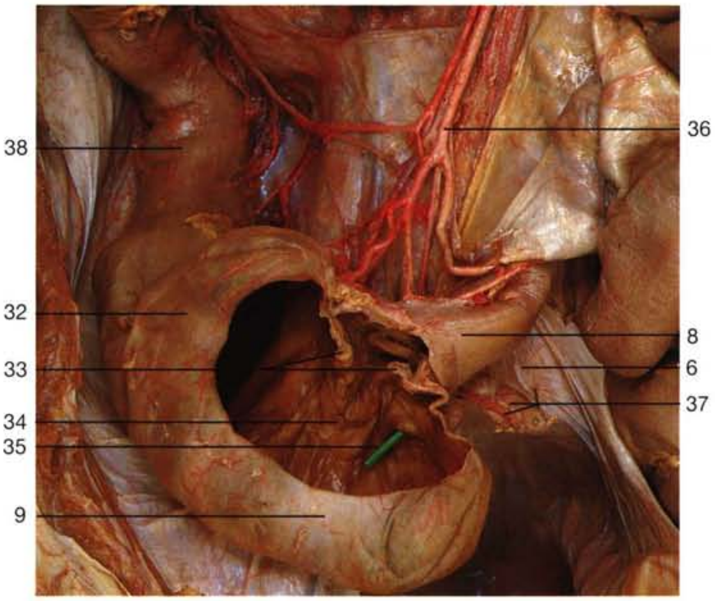

Ileocecal valve (ventral aspect). The cecum and terminal part of the ileum have been opened.

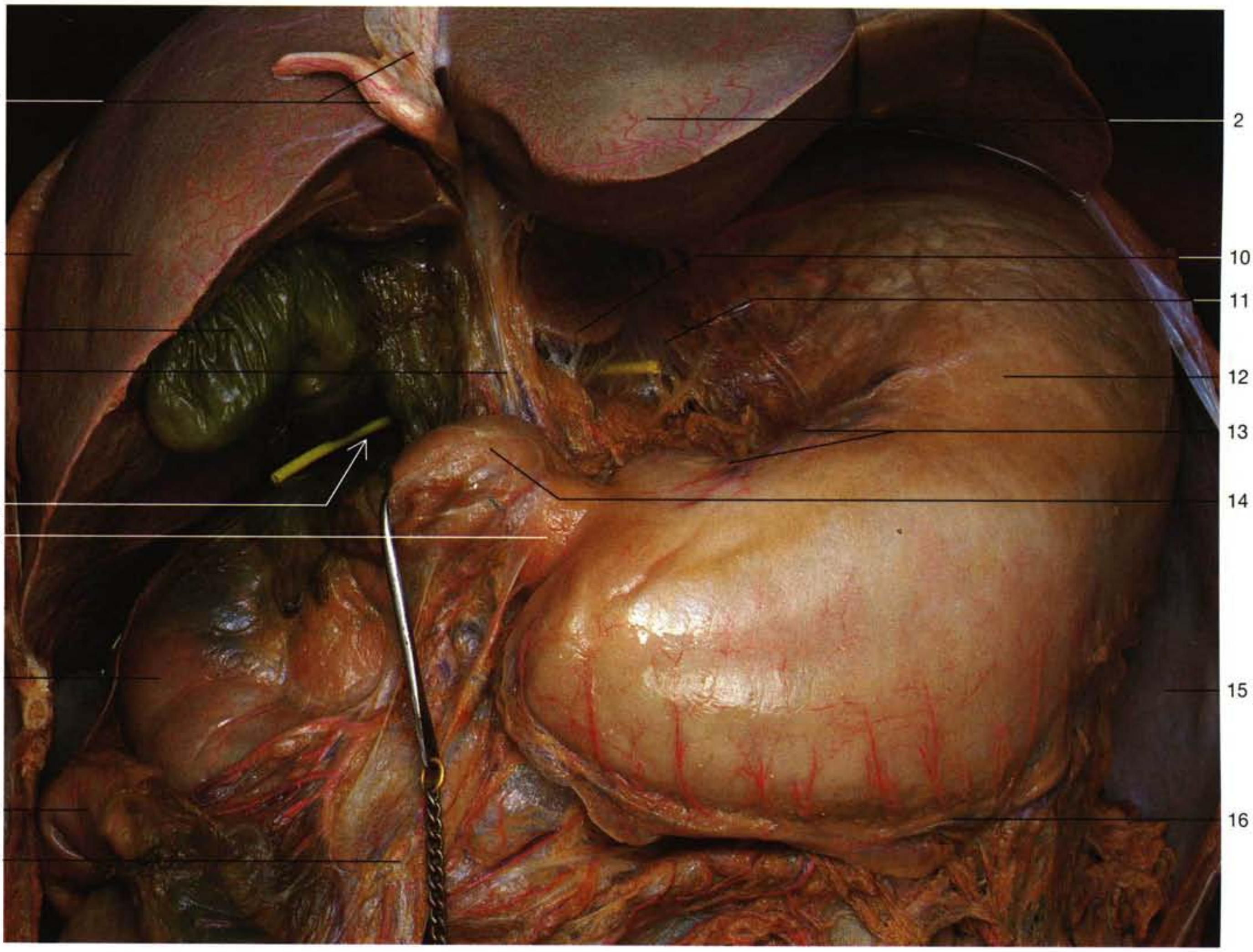

Upper abdominal organs. Thorax and anterior part of diaphragm have been removed and the liver raised to display the lesser omentum. A probe has been inserted into the epiploic foramen and lesser sac.

1 Falciform ligament and ligamentum teres
2 Liver
3 Gallbladder (fundus)
4 Hepatoduodenal ligament
5 **Epiploic foramen** (probe)
6 Pylorus
7 Descending part of duodenum
8 Right colic flexure
9 **Gastrocolic ligament**
10 Caudate lobe of liver (behind lesser omentum)
11 **Lesser omentum**
12 Stomach
13 Lesser curvature of stomach
14 Superior part of duodenum
15 Diaphragm
16 Greater curvature of stomach with gastroepiploic vessels
17 Twelfth thoracic vertebra
18 Right kidney
19 Right suprarenal gland
20 Inferior vena cava
21 Falciform ligament of liver
22 Abdominal aorta
23 Spleen
24 Lienorenal ligament
25 Gastrosplenic ligament
26 Pancreas
27 Lesser sac

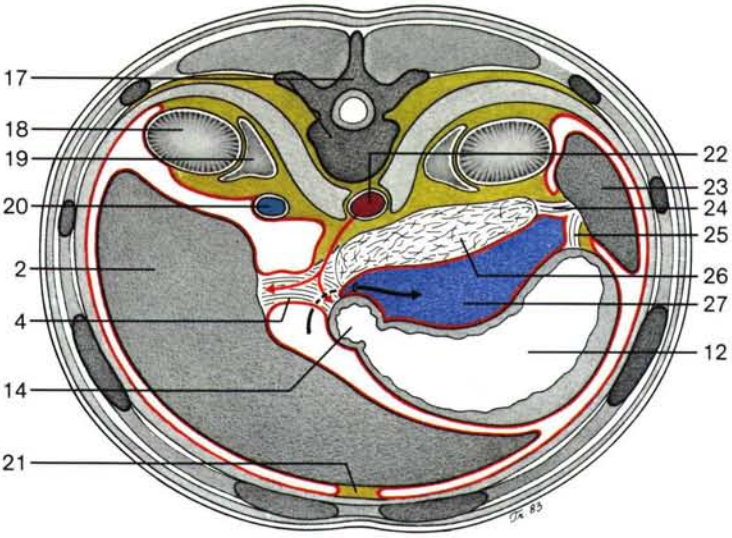

Horizontal section through omental bursa above the level of epiploic foramen (black arrow). Viewed from above. Red arrows: routes of the arterial branches of celiac trunk to liver, stomach, duodenum, and pancreas. (Schematic drawing.)

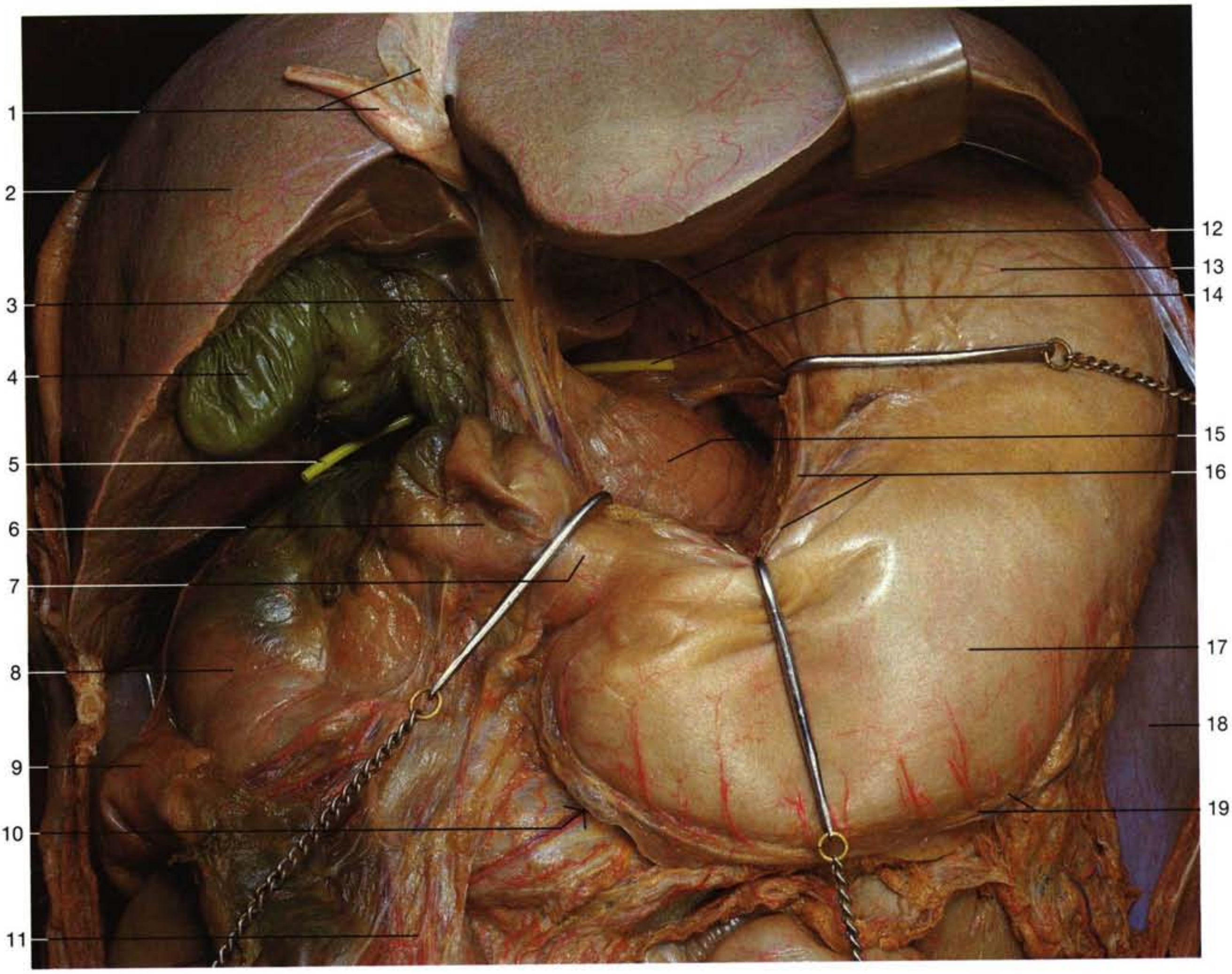

Upper abdominal organs, lesser sac, omental bursa (anterior aspect). Lesser omentum partly removed, liver and stomach slightly reflected.

1 Falciform ligament and ligamentum teres
2 Liver
3 **Hepatoduodenal ligament**
4 Gallbladder
5 Probe within the epiploic foramen
6 Superior part of duodenum
7 Pylorus
8 Descending part of duodenum
9 Right colic flexure
10 Gastrocolic ligament
11 **Greater omentum**
12 Caudate lobe of liver
13 Fundus of stomach
14 Probe at the level of the vestibule of lesser sac
15 **Head of pancreas**
16 Lesser curvature of stomach
17 **Body of stomach**
18 Diaphragm
19 Greater curvature with gastroepiploic vessels
20 Head of pancreas and gastropancreatic fold
21 **Spleen**
22 **Tail of pancreas**
23 Left colic flexure
24 Root of transverse mesocolon
25 Transverse mesocolon
26 Gastrocolic ligament (cut edge)
27 **Transverse colon**
28 Umbilicus
29 Small intestine
30 Lesser omentum
31 Lesser sac (omental bursa)
32 Duodenum
33 Mesentery
34 Sigmoid colon

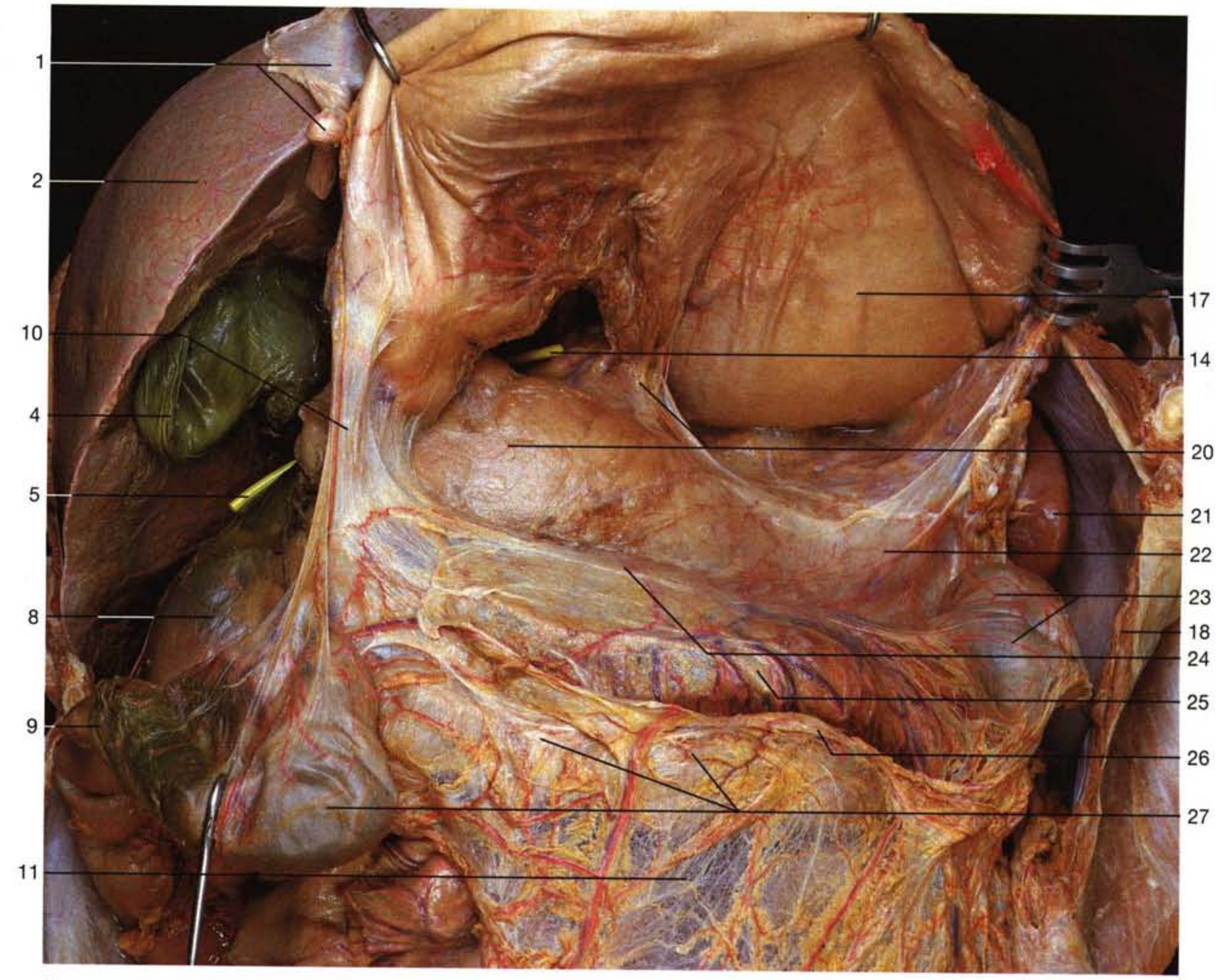
1
2
10
4
5
8
9
11
17
14
20
21
22
23
18
24
25
26
27

Upper abdominal organs, lesser sac, omental bursa (anterior aspect). The gastrocolic ligament has been divided and the whole stomach raised to display the posterior wall of the lesser sac.

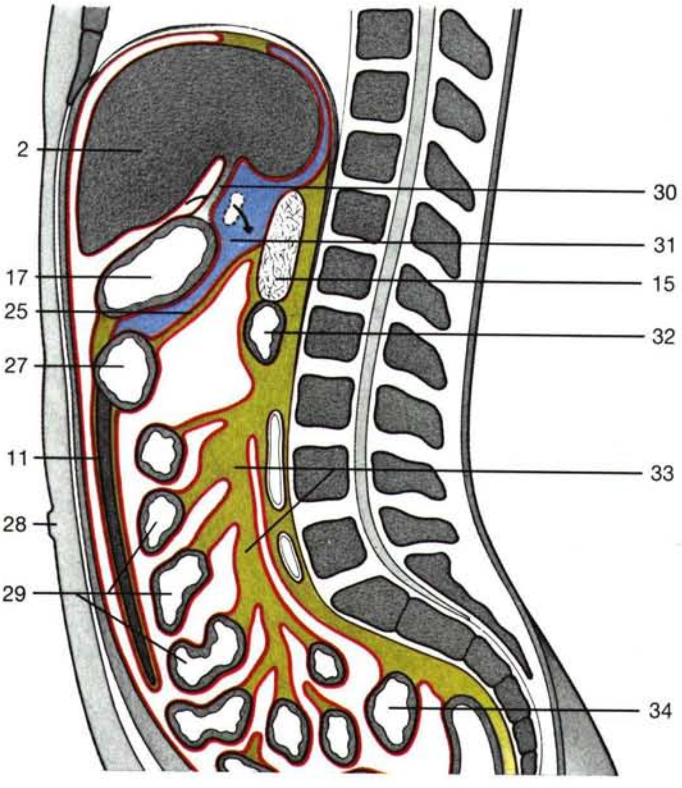
2
17
25
27
11
28
29
30
31
15
32
33
34

Midsagittal section through abdominal cavity demonstrating the site of lesser sac (blue). (Schematic drawing.) The epiploic foramen, entrance to the lesser sac, is indicated by an arrow. Red = peritoneum.

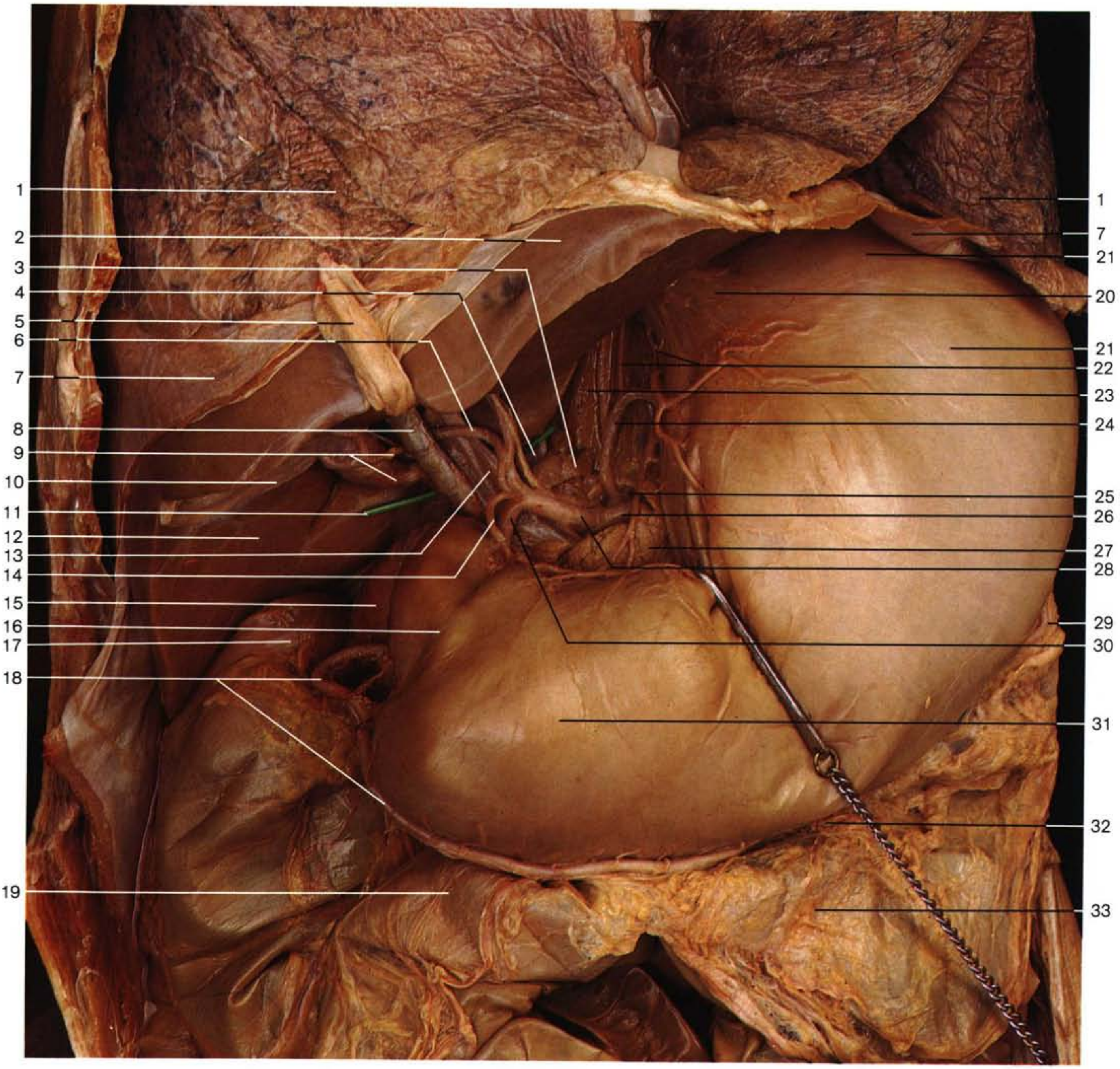
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
1
7
21
20
21
22
23
24
25
26
27
28
29
30
31
32
33

Arteries of upper abdominal organs; dissection of celiac trunk. The lesser omentum has been removed and the lesser curvature of the stomach reflected to display the branches of the celiac trunk.

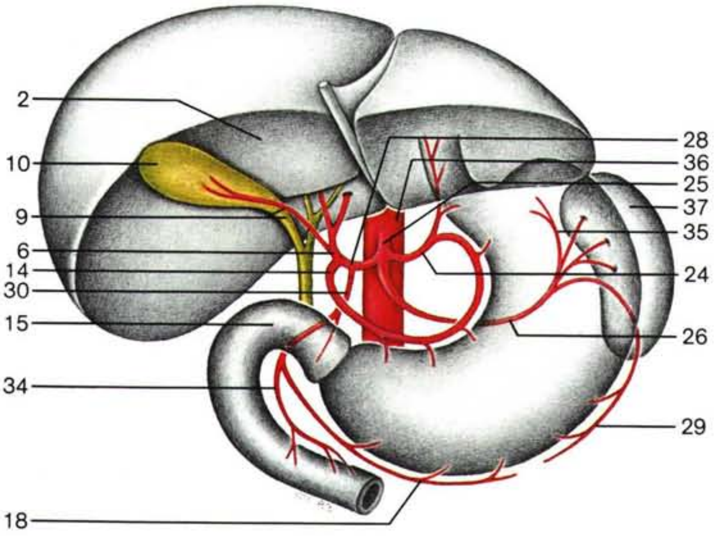
2
10
9
6
14
30
15
34
18
28
36
25
37
35
24
26
29

Branches of celiac trunk.
(Schematic drawing.)

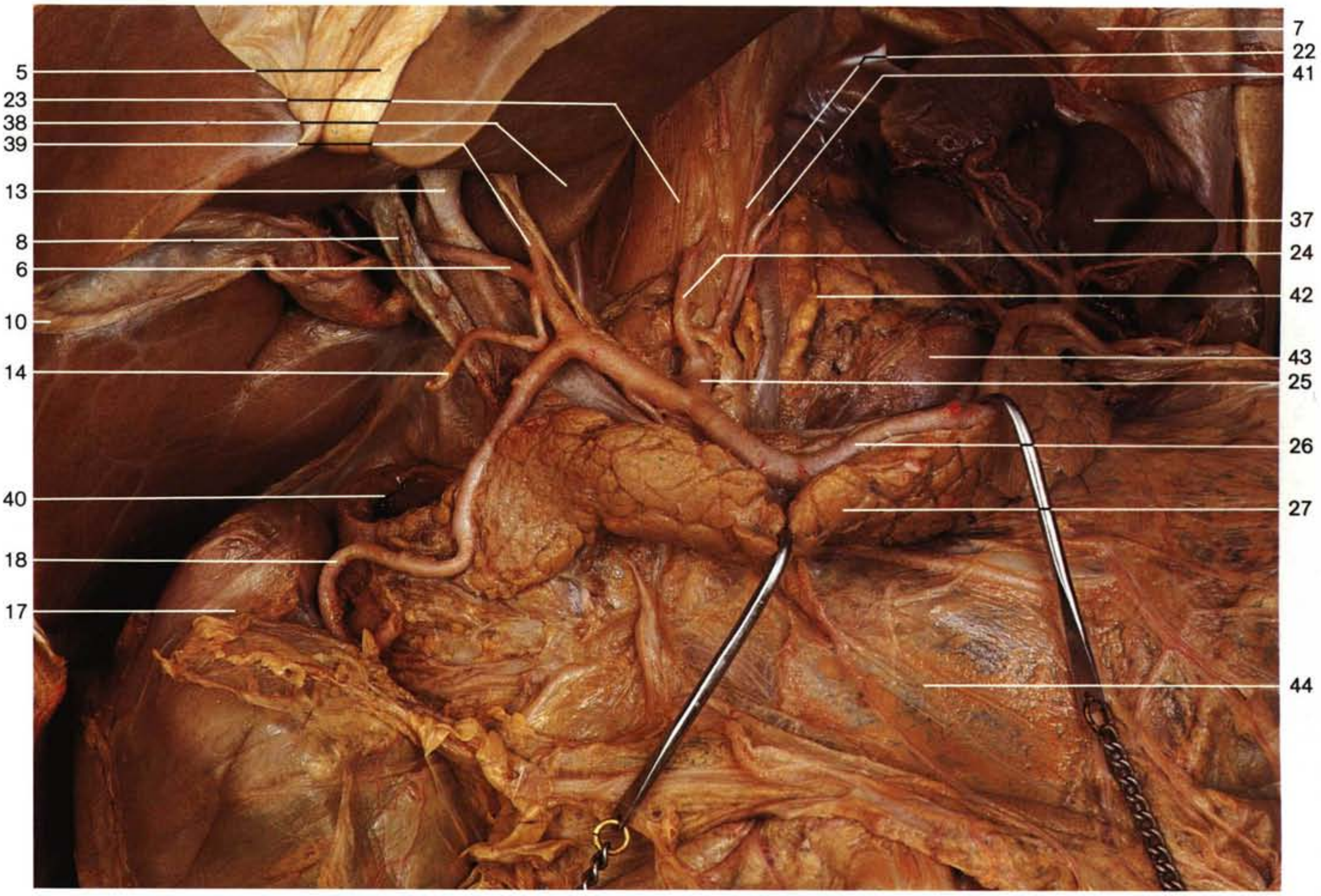

Branches of celiac trunk; blood supply of liver, pancreas and spleen. The stomach, superior part of duodenum and celiac ganglion have been removed to reveal the anterior aspect of the posterior wall of the lesser sac (omental bursa) and the vessels and ducts of the hepatoduodenal ligament. The pancreas has been slightly reflected anteriorly.

1 Lung
2 Liver (visceral surface)
3 Lymph node
4 Inferior vena cava
5 Ligamentum teres (reflected)
6 Right branch of hepatic artery proper
7 Diaphragm
8 **Common hepatic duct** (dilated)
9 Cystic duct and artery
10 Gallbladder
11 Probe in epiploic foramen
12 Right lobe of liver
13 **Portal vein**
14 **Right gastric artery**
15 Duodenum
16 Pylorus
17 Right colic flexure
18 **Right gastro-omental** (gastroepiploic) **artery**
19 Transverse colon
20 Abdominal part of esophagus (cardiac part of stomach)
21 Fundus of stomach
22 Esophageal branches of left gastric artery
23 Lumbar part of diaphragm
24 Left gastric artery
25 **Celiac trunk**
26 **Splenic artery**
27 Pancreas
28 **Common hepatic artery**
29 Left gastro-omental (gastroepiploic) artery
30 **Gastroduodenal artery**
31 Pyloric part of stomach
32 Greater curvature of stomach
33 Gastrocolic ligament
34 Superior pancreaticoduodenal artery
35 Short gastric arteries
36 Aorta
37 Spleen
38 Caudate lobe of liver
39 Left branch of hepatic artery proper
40 Descending part of duodenum (cut)
41 Left inferior phrenic artery
42 Suprarenal gland
43 Kidney
44 Transverse mesocolon

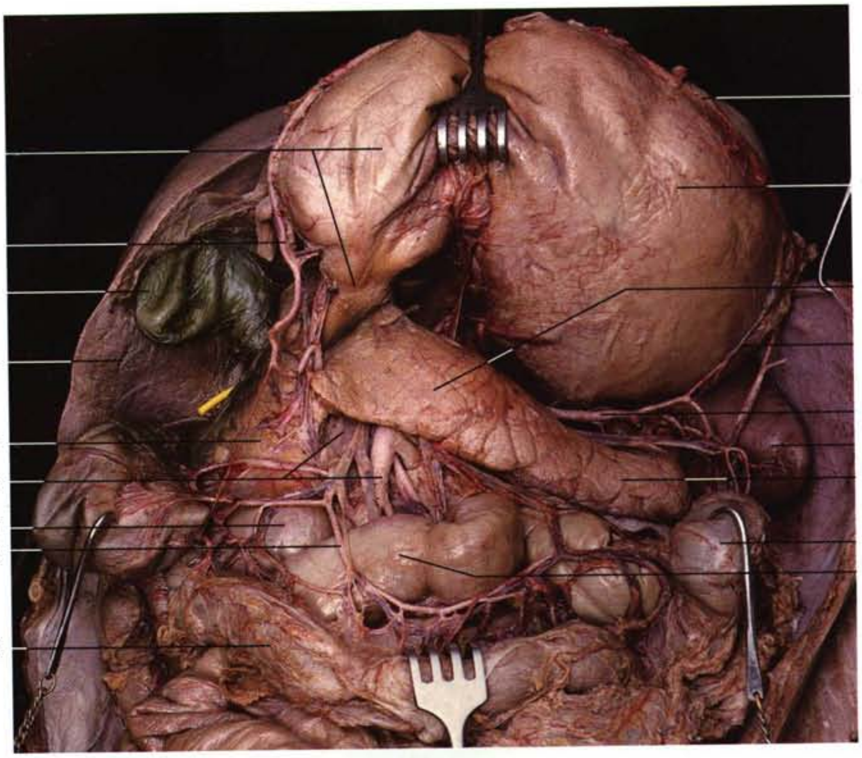

Pancreas and extrahepatic bile ducts in situ (anterior aspect). The gastrocolic ligament has been divided, the transverse colon and the stomach were replaced to display pancreas and superior mesenteric vessels.

1 Stomach (pyloric part) and pylorus
2 Right gastro-omental (gastroepiploic) artery
3 **Fundus of gall bladder**
4 Liver, right lobe
5 Head of pancreas
6 **Superior mesenteric artery** and **vein**
7 Duodenum
8 Middle colic artery
9 Transverse colon
10 Greater curvature of stomach (remnants of gastrocolic ligament)
11 **Body of stomach**
12 **Body of pancreas**
13 Left gastro-omental (gastroepiploic) artery
14 **Splenic artery**
15 **Spleen**
16 Tail of pancreas
17 Left colic flexure
18 Jejunum
19 Lung (inferior lobe of right lung)
20 **Common hepatic duct** and caudate lobe of liver
21 Portal vein and hepatic artery proper
22 Cystic duct and right gastric artery
23 **Common bile duct** and gastroduodenal artery
24 Pylorus (cut)
25 Major and minor duodenal papillas (probes)
26 Duodenum (inferior part)
27 Right kidney, ureter and ovarian artery
28 Iliocolic artery and inferior vena cava
29 Right and left vagal trunk
30 **Esophagus**
31 Diaphragm
32 Cut edge of esophagus, left gastric artery with esophageal branch
33 Celiac trunk
34 Common hepatic artery, main pancreatic duct and **pancreas**
35 Renal vein
36 Duodenojejunal flexure
37 Left kidney
38 Jejunal arteries
39 Abdominal aorta and left ureter
40 Cystic artery
41 Hepatic proper artery
42 **Celiac trunk**
43 Right gastric artery
44 **Common hepatic artery**
45 Gastroduodenal artery
46 **Superior mesenteric artery**
47 Superior posterior pancreaticoduodenal artery
48 Superior anterior pancreaticoduodenal artery
49 Short gastric arteries
50 Left gastric artery
51 Posterior pancreatic branch of splenic artery
52 Inferior pancreaticoduodenal artery

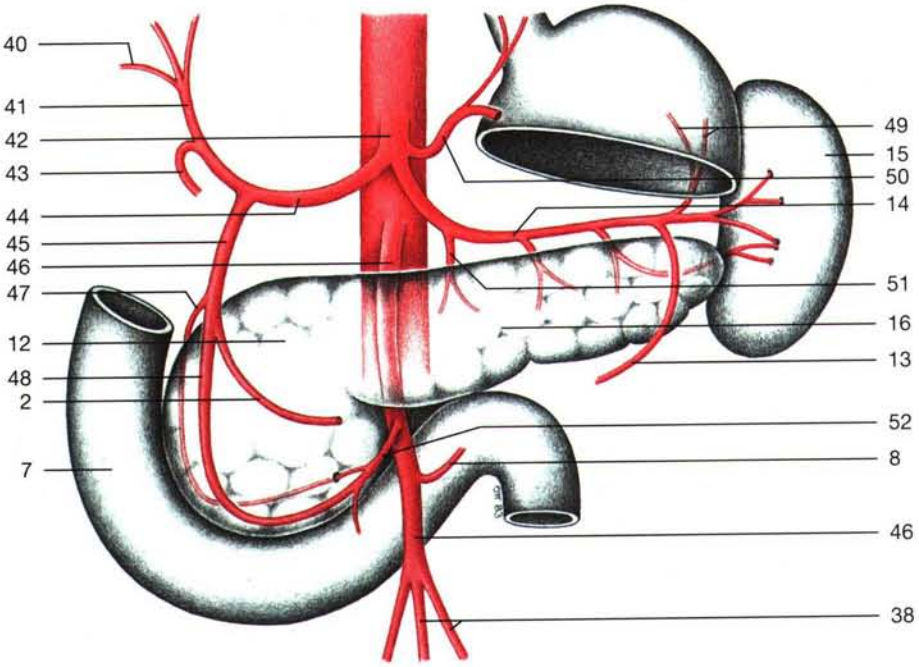

Blood supply of upper abdominal organs (branches of the celiac trunk and superior mesenteric artery). (Schematic drawing.)

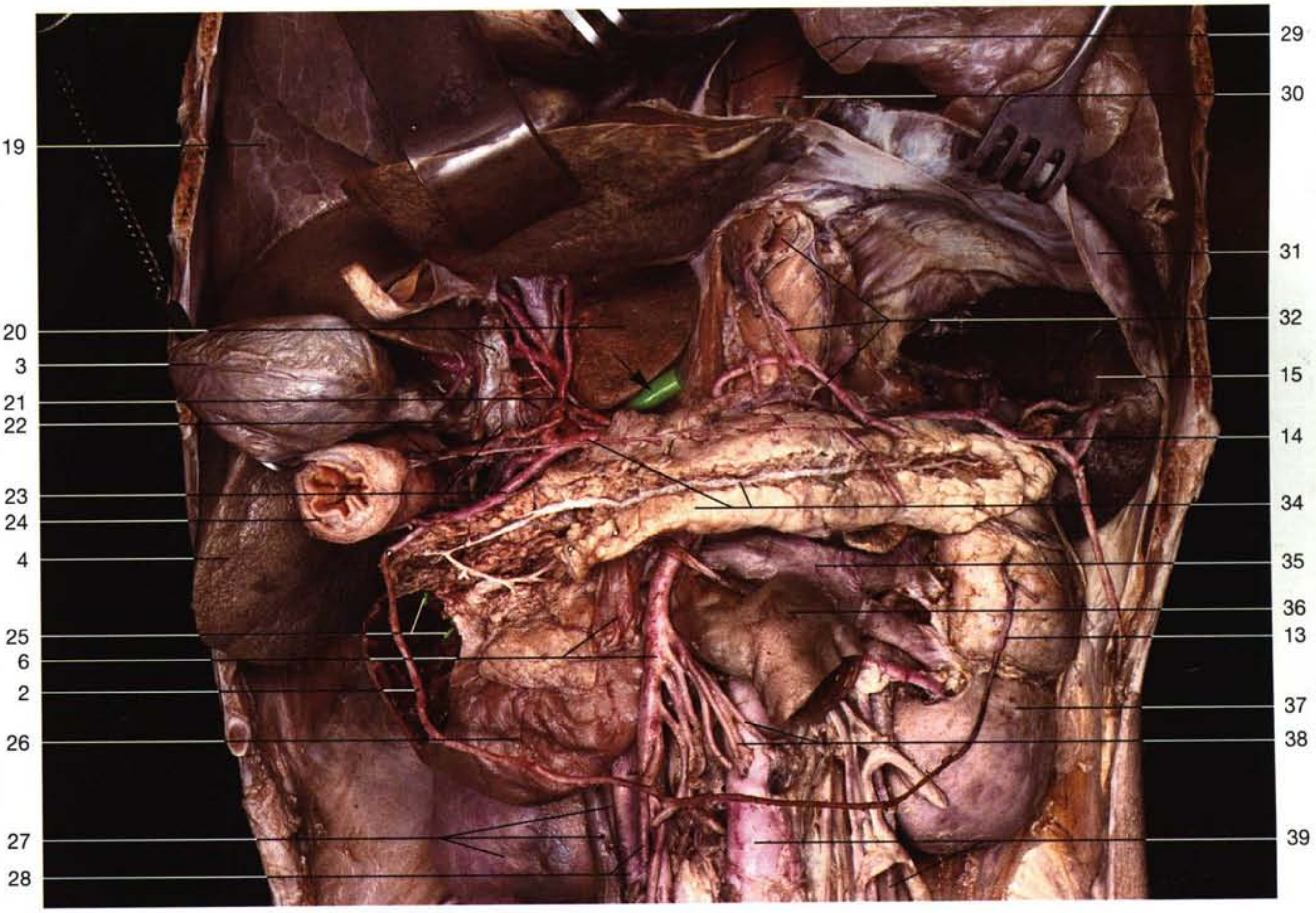
29
30
19
31
20
32
3
15
21
22
14
23
34
24
4
35
36
25
13
6
2
37
26
38
27
39
28

Pancreas, extrahepatic bile ducts, spleen and liver with their vessels in situ. The stomach has been removed, the liver elevated, and the descending part of duodenum opened to display the openings of pancreatic ducts. The pancreatic ducts were dissected. Note the location of superior mesenteric artery and vein between duodenum and pancreas. Arrow: probe within the epiploic foramen.

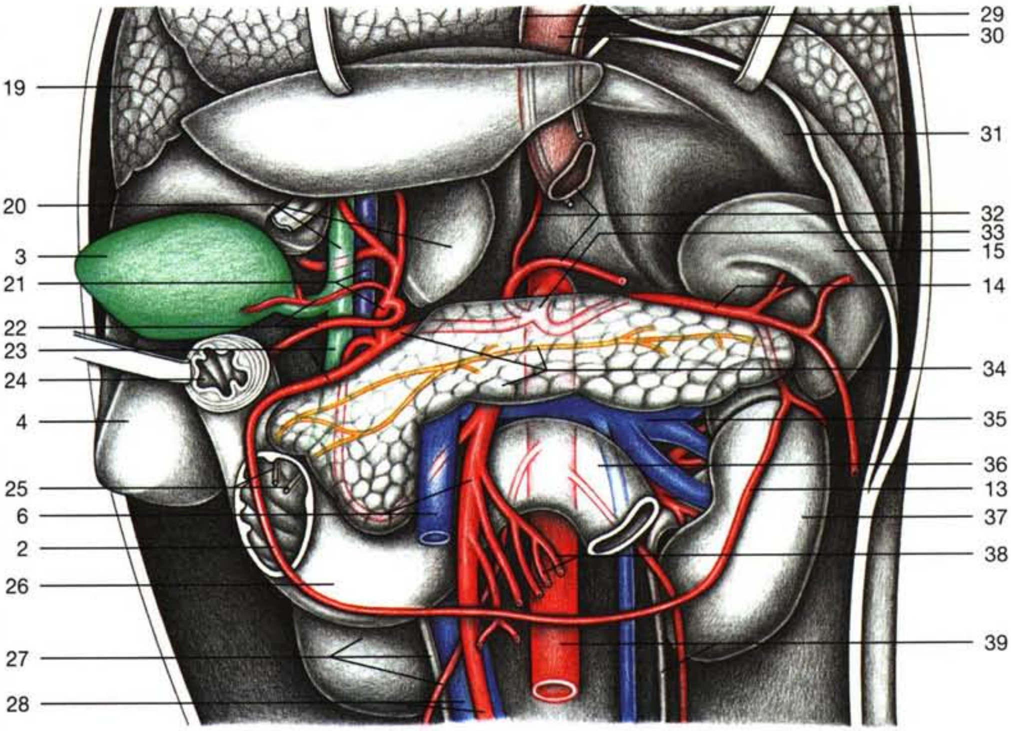
29
30
19
31
20
32
33
3
15
21
14
22
23
34
24
4
35
25
36
13
6
37
2
38
26
27
39
28

Schematic drawing of the specimen depicted above. Note the branching of the celiac trunk (partially covered by the body of the pancreas), and the main arterial supply of liver, spleen, pancreas, stomach, and duodenum.

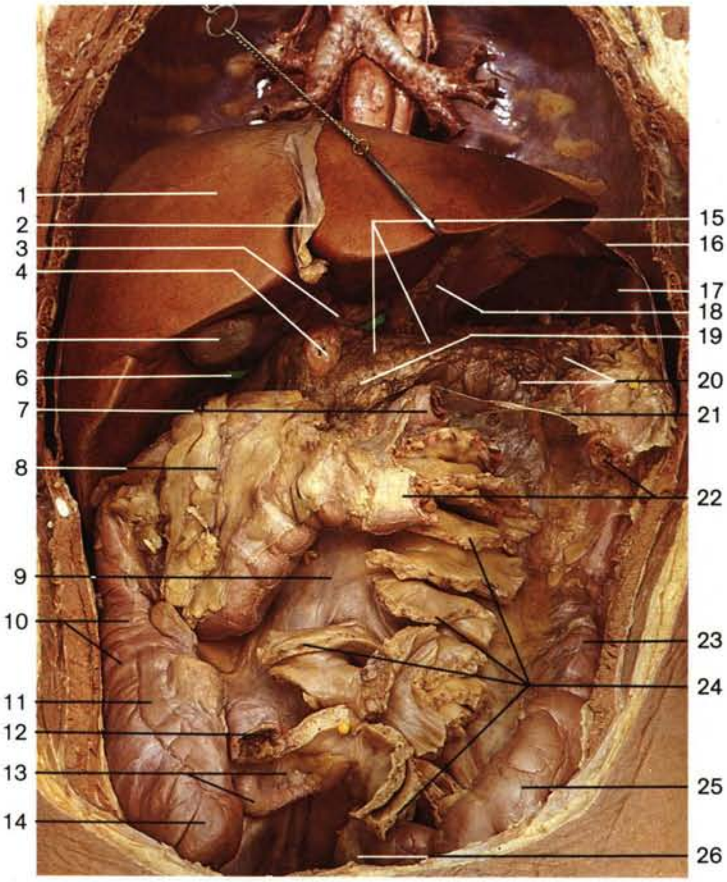

Abdominal cavity after removal of stomach, jejunum, ileum and part of the transverse colon. Liver has been slightly raised.

1 Liver
2 Falciform ligament
3 Hepatoduodenal ligament
4 Pylorus (divided)
5 Gallbladder
6 Probe within the epiploic foramen
7 Duodenojejunal flexure (divided)
8 **Greater omentum**
9 Root of mesentery
10 Ascending colon
11 Free colic taenia
12 End of ileum (divided)
13 Vermiform appendix with mesoappendix
14 Cecum
15 Pancreas and site of lesser sac
16 Diaphragm
17 Spleen
18 Cardia (part of stomach, divided)
19 Head of pancreas
20 Body and tail of pancreas
21 **Transverse mesocolon**
22 Transverse colon (divided)
23 Descending colon
24 Cut edge of mesentery
25 Sigmoid colon
26 Rectum
27 Attachment of bare area of liver
28 Inferior vena cava
29 Kidney
30 Attachment of right colic flexure
31 Root of transverse mesocolon
32 Junction between descending and horizontal parts of duodenum
33 Bare surface for ascending colon
34 **Ileocecal recess**
35 **Retrocecal recess**
36 Root of mesoappendix
37 Superior recess
38 Isthmus (opening) } of **lesser sac (omental bursa)**
39 Splenic recess
40 **Superior duodenal recess**
41 **Inferior duodenal recess**
42 Bare surface for descending colon
43 **Paracolic recesses**
44 **Root of mesentery**
45 Root of mesosigmoid
46 **Intersigmoid recess**
47 Hepatic veins
48 Duodenojejunal flexure
49 Attachment of left colic flexure
50 Esophagus
51 Entrance to lesser sac through the epiploic foramen

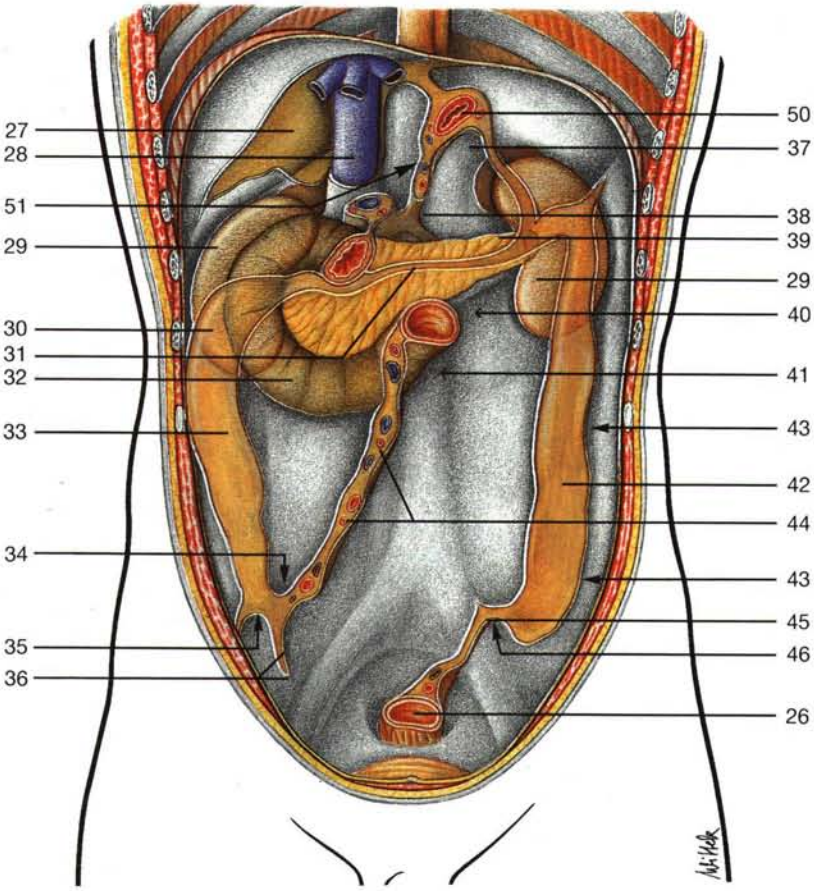

Peritoneal reflections from organs and the position of root of mesentry and peritoneal recesses on the posterior abdominal wall. (Schematic drawing.)

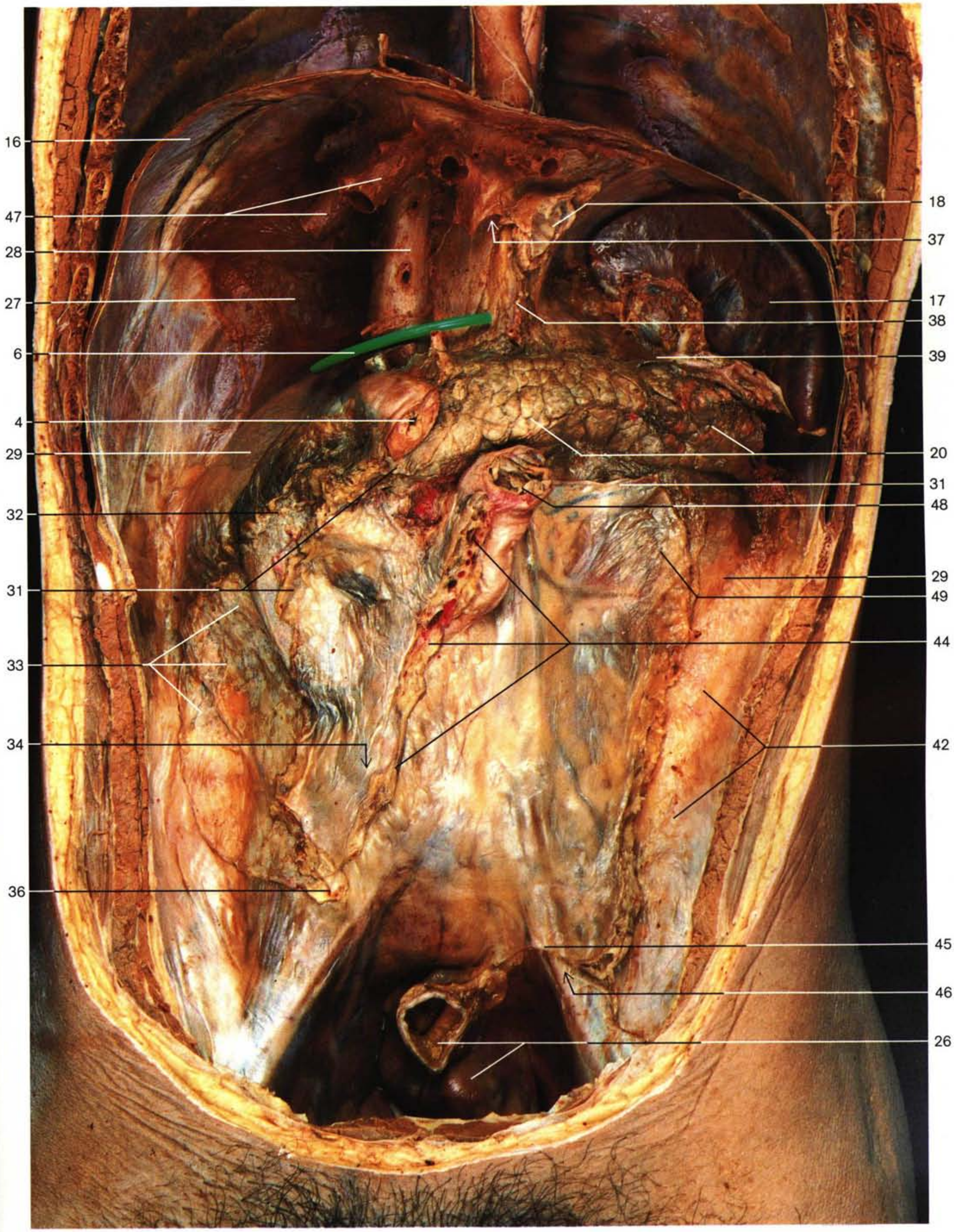
16
47
28
27
6
4
29
32
31
33
34
36
18
37
17
38
39
20
31
48
29
49
44
42
45
46
26

Peritoneal recesses on the posterior abdominal wall. The liver, stomach, jejunum, ileum, and colon have been removed. The duodenum, pancreas, and spleen have been left in place.

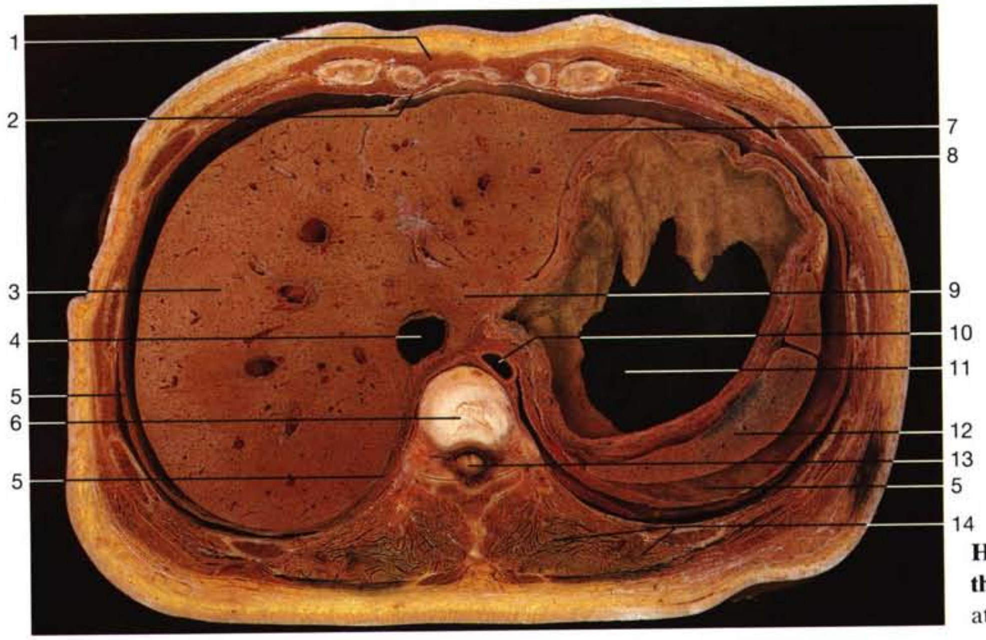

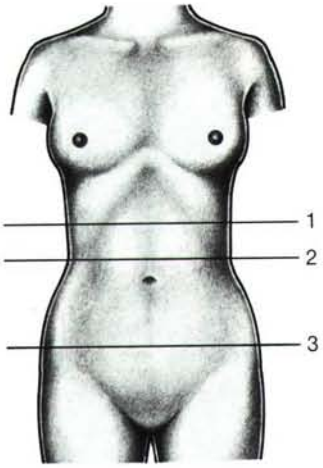

Horizontal section through the abdominal cavity at level 1 (from below).

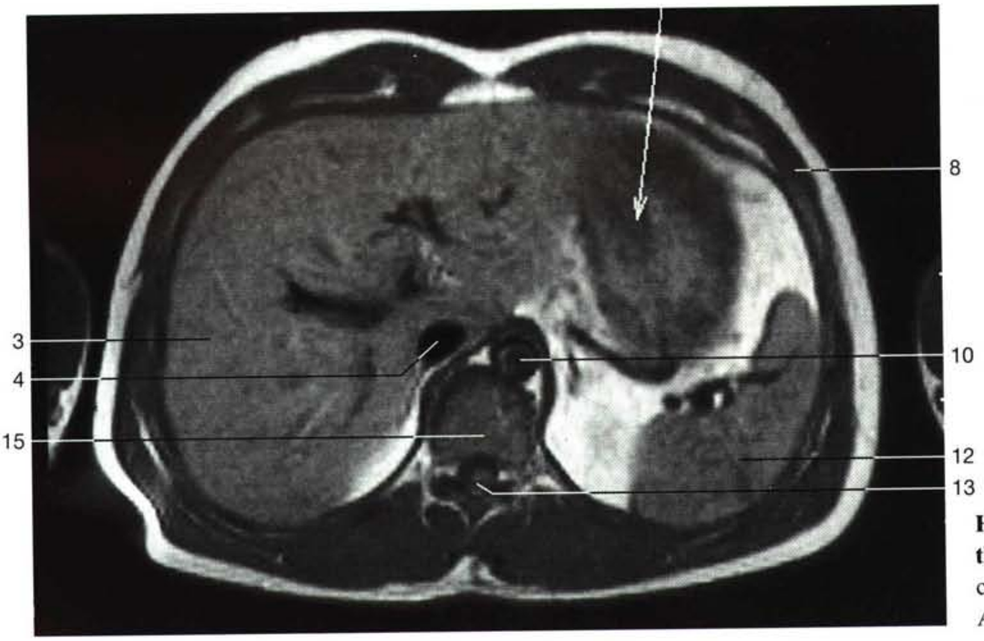

Horizontal section through the body. MR-Scan, corresponding to level 1. Arrow: stomach.

1 Rectus abdominis muscle
2 Falciform ligament
3 **Liver** (right lobe)
4 Inferior vena cava
5 Diaphragm
6 Intervertebral disc
7 Liver (left lobe)
8 Rib
9 Liver (caudate lobe)
10 Abdominal (descending) aorta
11 **Stomach**
12 **Spleen**
13 Spinal cord
14 Longissimus and iliocostalis muscles
15 Body of vertebra
16 Abdominal rectus muscle
17 External abdominal oblique muscle
18 Transverse colon
19 **Head of pancreas**
20 **Greater duodenal papilla**
21 Duodenum
22 Suprarenal gland and ureter
23 **Kidney**
24 Round ligament of liver
25 Superior mesenteric artery and vein
26 Psoas major muscle
27 Descending colon
28 Quadratus lumborum muscle
29 Cauda equina
30 Right renal vein
31 **Small intestine**
32 Iliacus muscle
33 Ilium
34 **Ileocecal valve**
35 Cecum
36 Common iliac artery and vein
37 Gluteus medius muscle
38 Vertebral canal and dura mater

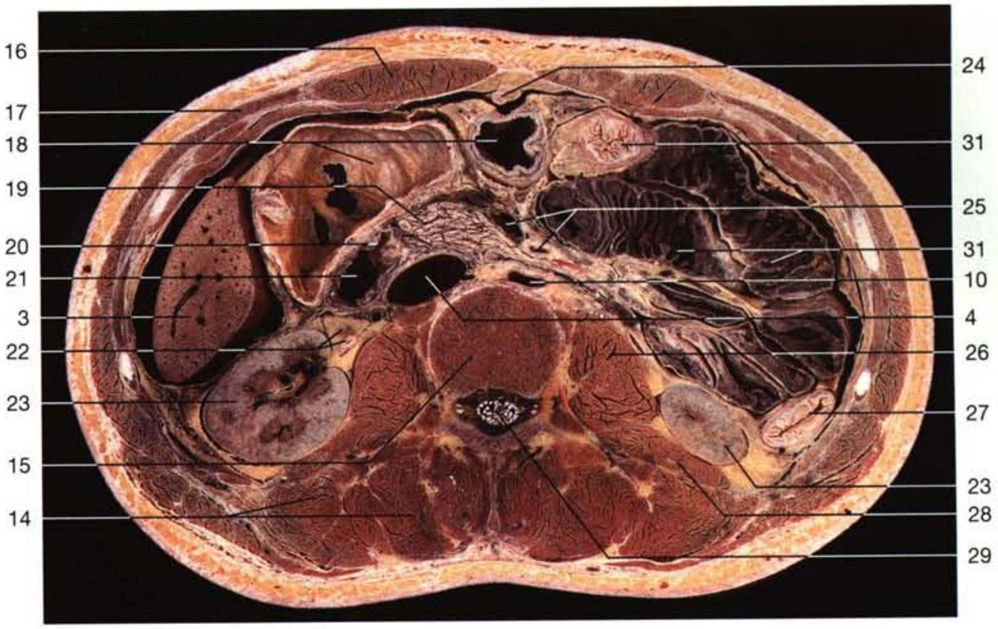
16
17
18
19
20
21
3
22
23
15
14
24
31
25
31
10
4
26
27
23
28
29

Horizontal section through the abdominal cavity at the level of greater duodenal papilla (from below).

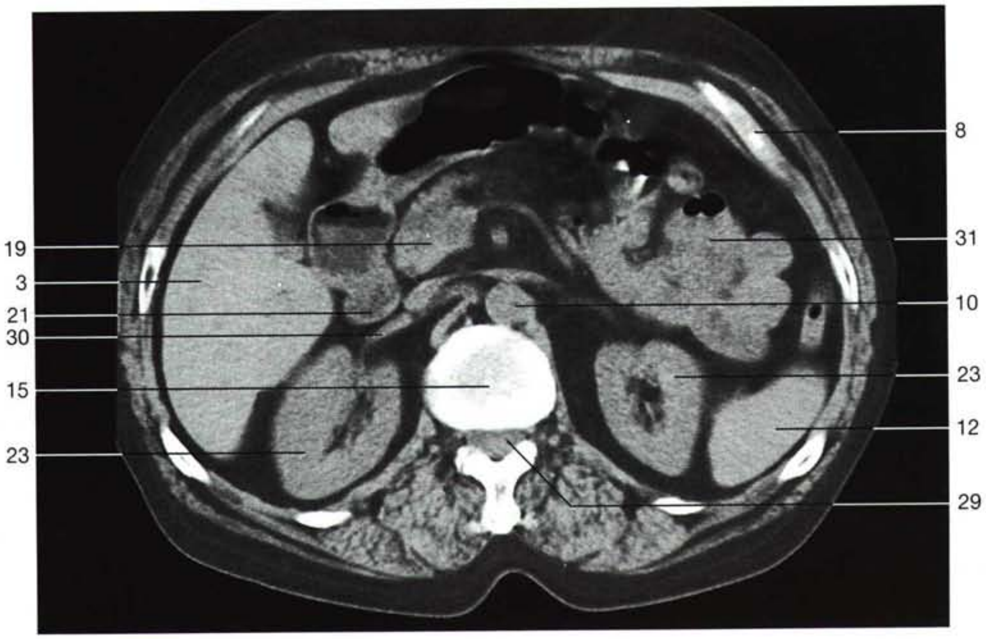
19
3
21
30
15
23
8
31
10
23
12
29

Horizontal section through the body. CT-Scan, corresponding to level 2.

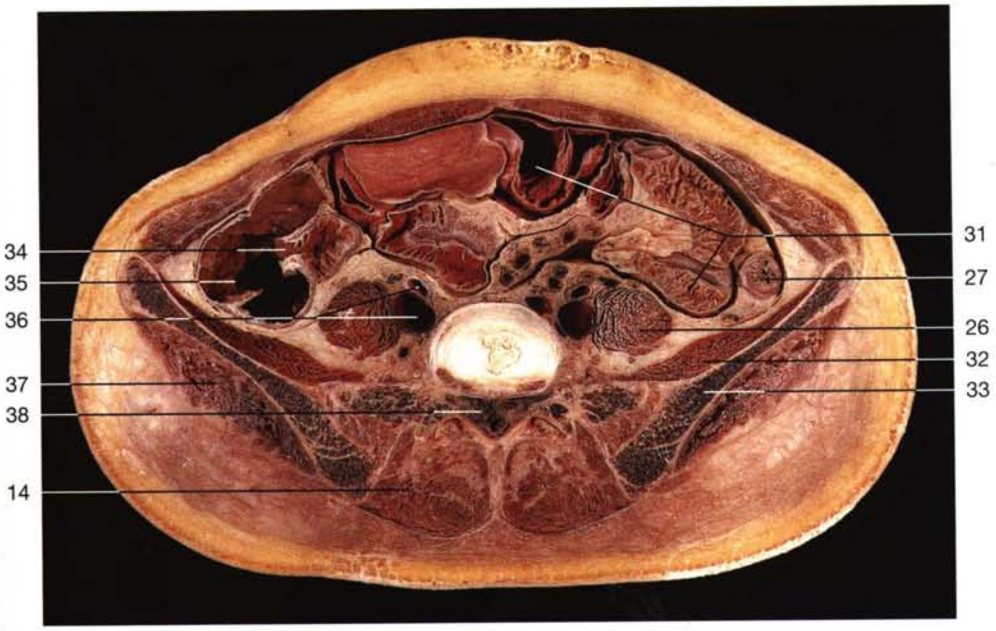
34
35
36
37
38
14
31
27
26
32
33

Horizontal section through the abdominal cavity at level 3 (from below).

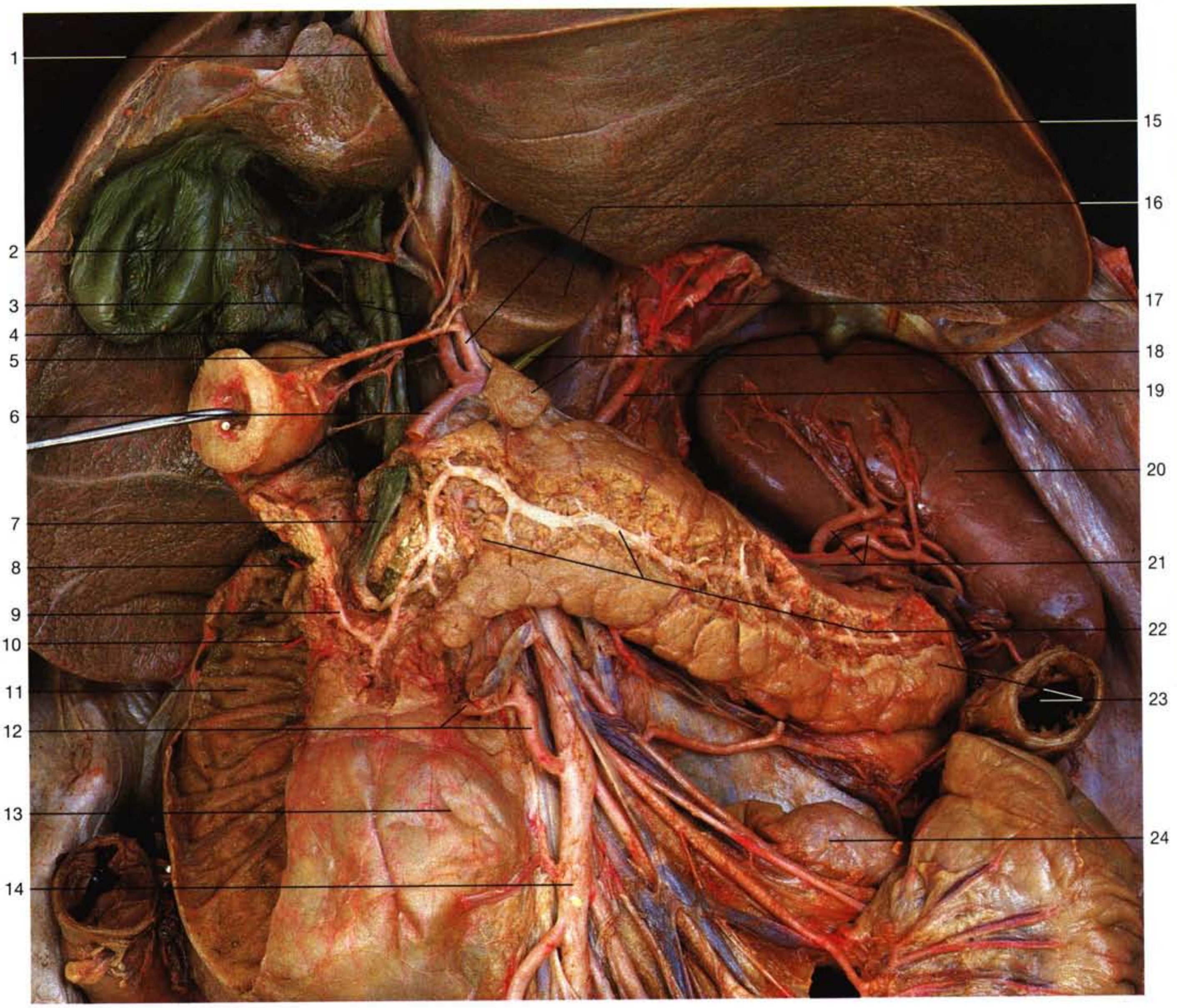

Posterior abdominal wall with duodenum, pancreas and spleen (anterior aspect). Dissection of pancreatic and common bile duct. The stomach has been removed, the liver raised and the duodenum anteriorly opened.

1	Ligamentum teres	13	Horizontal part of duodenum (distended)
2	Gallbladder and cystic artery	14	**Superior mesenteric artery**
3	Common hepatic duct and portal vein	15	Liver (left lobe)
4	Cystic duct	16	Caudate lobe of liver and hepatic artery proper
5	Right gastric artery (pylorus with superior part of duodenum, cut and reflected)	17	Abdominal part of esophagus (cut)
6	Gastroduodenal artery	18	Probe in epiploic foramen and lymph node
7	**Common bile duct**	19	Left gastric artery
8	Probe within the minor duodenal papilla	20	**Spleen**
9	**Accessory pancreatic duct**	21	Splenic vein and branches of splenic artery
10	Probe within the major duodenal papilla	22	**Pancreatic duct and head of pancreas**
11	Descending part of duodenum (opened)	23	Left colic flexure and tail of pancreas
12	Middle colic artery and inferior pancreaticoduodenal artery	24	Duodenojejunal flexure

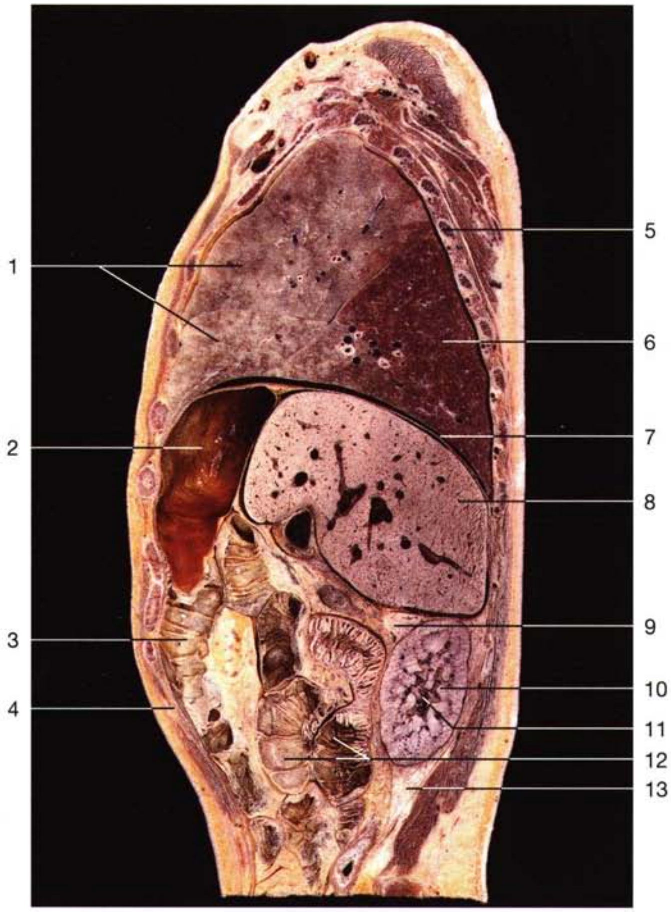

Parasagittal section through the thoracic and abdominal cavities (right side, medial aspect).

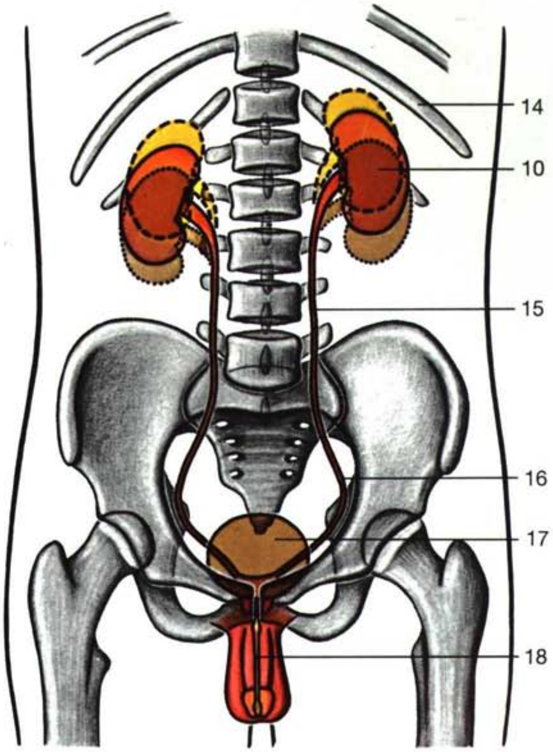

Position of kidneys and urinary system (anterior view). The excursions of the kidneys with the respiratory movements of the diaphragm are indicated. (Schematic drawing.)

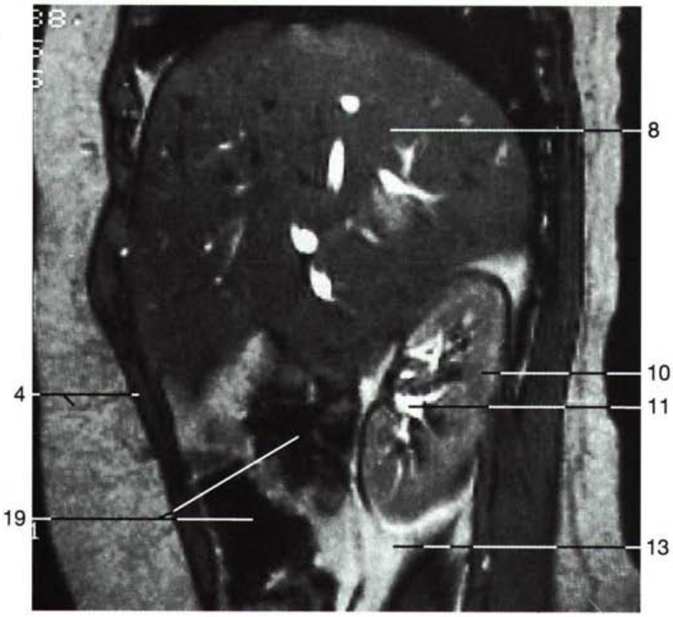

Parasagittal section through the abdominal cavity. (MR-Scan, courtesy of Prof. W. Rödl, Erlangen).

1 Right lung (superior and middle lobes)
2 Transverse colon
3 Jejunum
4 Abdominal wall
5 Fourth rib
6 Right lung (inferior lobe)
7 Diaphragm
8 **Liver**
9 Suprarenal gland
10 **Kidney**
11 Renal pelvis
12 Small intestine
13 Perirenal fat pad
14 Eleventh rib
15 Ureter (abdominal part)
16 Ureter (pelvic part)
17 **Urinary bladder**
18 Urethra
19 Colon

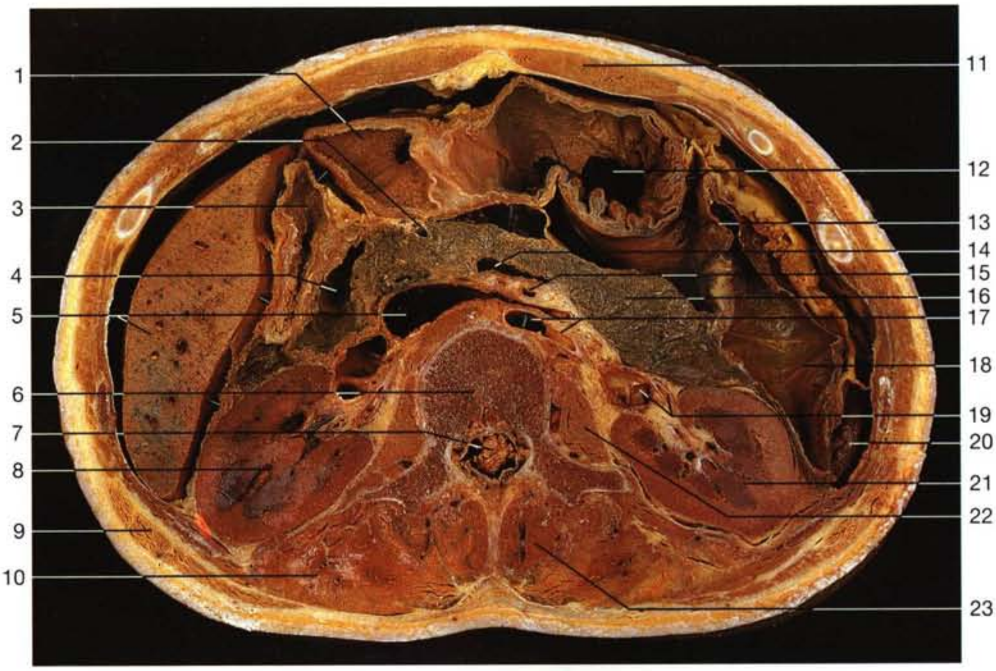

Horizontal section through the abdominal cavity at the level of the first lumbar vertebra (from below).

1	Pyloric antrum
2	Gastroduodenal artery
3	Descending part of duodenum
4	Vestibule of lesser sac
5	Inferior vena cava and liver
6	Body of first lumbar vertebra
7	Cauda equina
8	**Right kidney**
9	Latissimus dorsi muscle
10	Iliocostalis muscle
11	Rectus abdominis muscle
12	Stomach
13	Lesser sac
14	Splenic vein
15	Superior mesenteric artery
16	Pancreas
17	Aorta and left renal artery
18	Transverse colon
19	Renal artery and vein
20	Spleen
21	**Left kidney**
22	Psoas major muscle
23	Multifidus muscle
24	Margin of lung
25	Margin of pleura
26	Renal pelvis
27	**Left ureter**
28	Descending colon
29	Rectum
30	Right suprarenal gland
31	12th rib
32	Ascending colon
33	**Right ureter**
34	Cecum
35	Vermiform appendix
36	Urinary bladder
37	Liver
38	Anterior layer of renal fascia
39	Duodenum
40	Perirenal fatty tissue
41	Posterior layer of renal fascia
42	Abdominal cavity

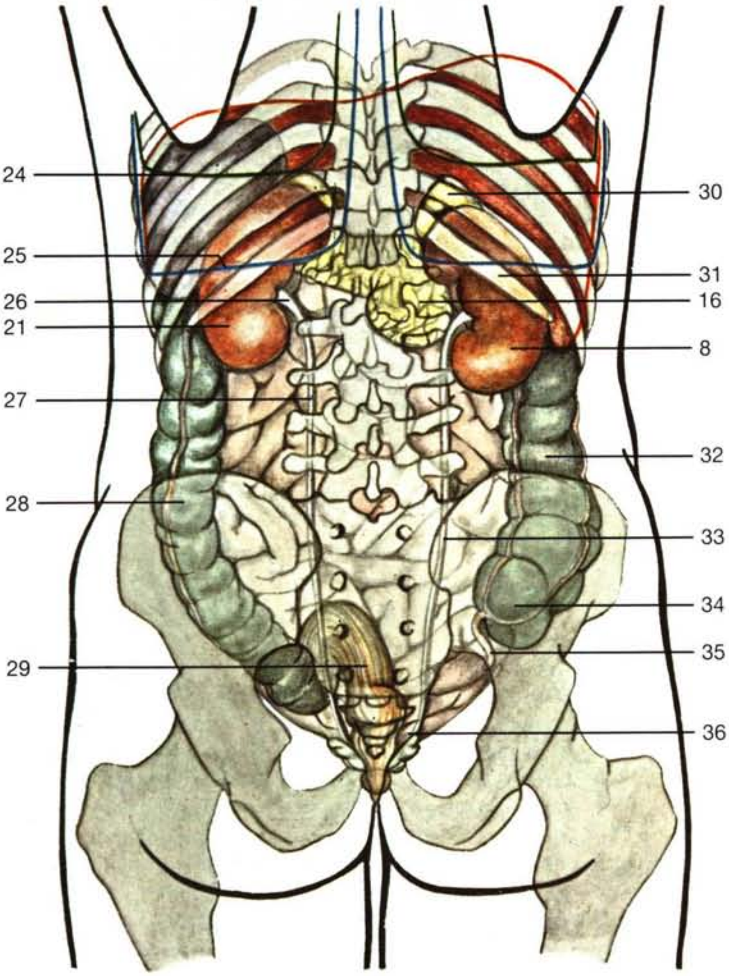

Positions of urinary organs (posterior view). Notice that the upper part of the kidney reaches the level of the margin of pleura and lung.

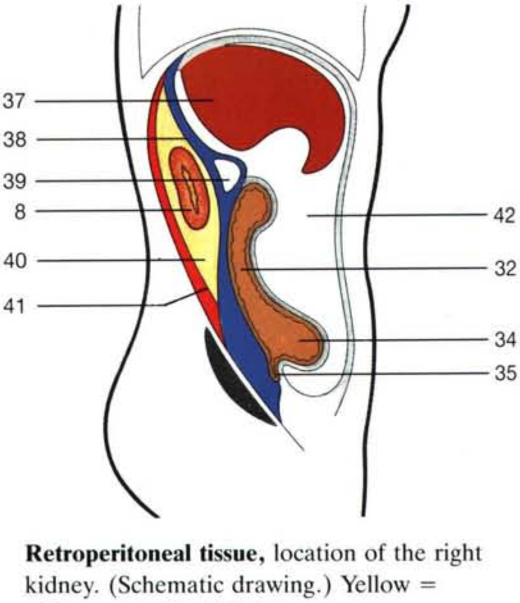

Retroperitoneal tissue, location of the right kidney. (Schematic drawing.) Yellow = adipose capsule of kidney.

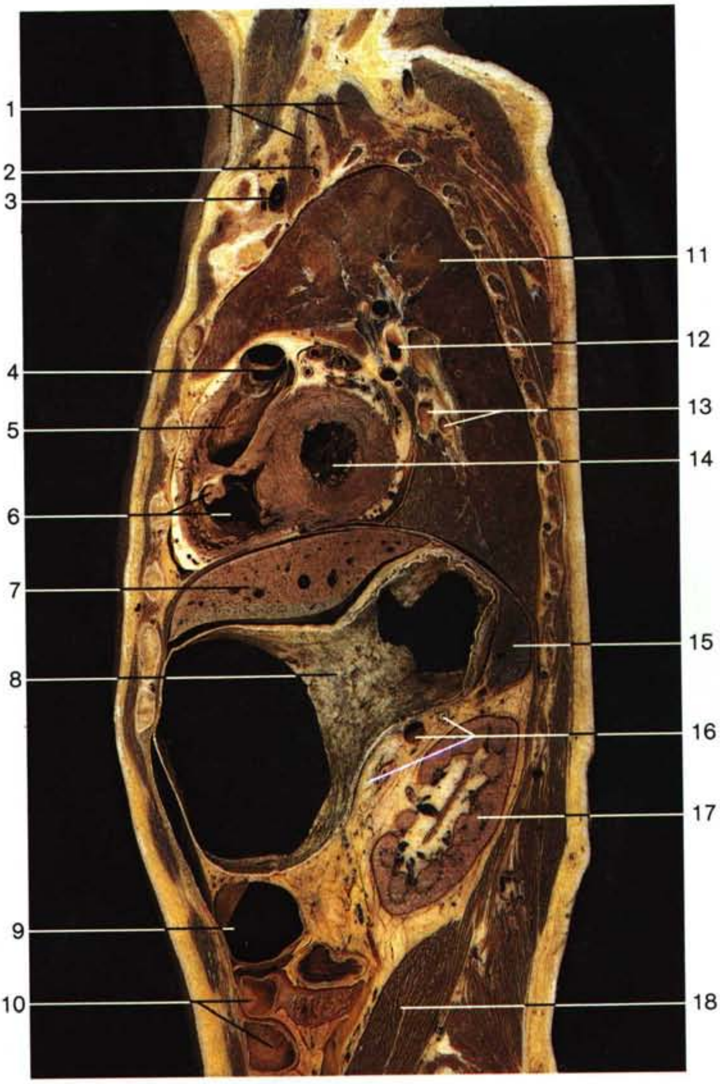

Parasagittal section through the thoracic and abdominal cavities at the level of the left kidney (5.5 cm left of median plane).

1 Scalenus anterior, medius and posterior muscles
2 Left subclavian artery
3 Left subclavian vein
4 Pulmonic valve
5 Arterial cone
6 Right ventricle of heart
7 Liver
8 Stomach
9 Transverse colon
10 Small intestine
11 Left lung
12 Left main bronchus
13 Branches of pulmonary vein
14 Left ventricle of heart
15 Spleen
16 Splenic artery and vein and pancreas
17 **Left kidney**
18 Psoas major muscle
19 Inferior vena cava
20 **Renal vein**
21 Body of 12th thoracic vertebra and vertebral canal
22 **Right kidney**
23 Superior mesenteric artery
24 Superior mesenteric vein
25 Pancreas
26 Abdominal aorta
27 Left psoas major and quadratus lumborum muscles
28 Anterior layer of renal fascia } of Gerota
29 Posterior layer of renal fascia }
30 Perirenal fatty tissue
31 Abdominal cavity
32 Descending and sigmoid colon

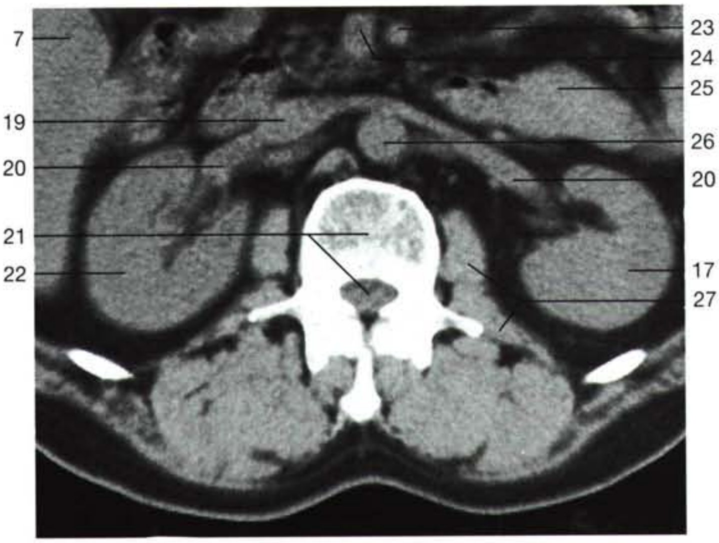

Horizontal section through the retroperitoneal region at level of 12th thoracic vertebra. (CT-Scan, from below.)

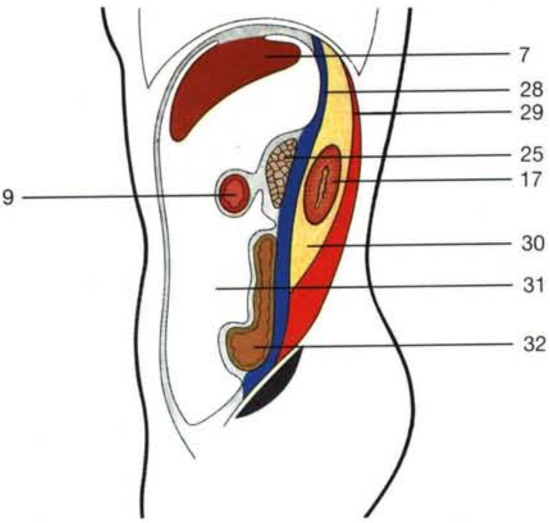

Retroperitoneal tissue, position of left kidney. (Schematic drawing.)

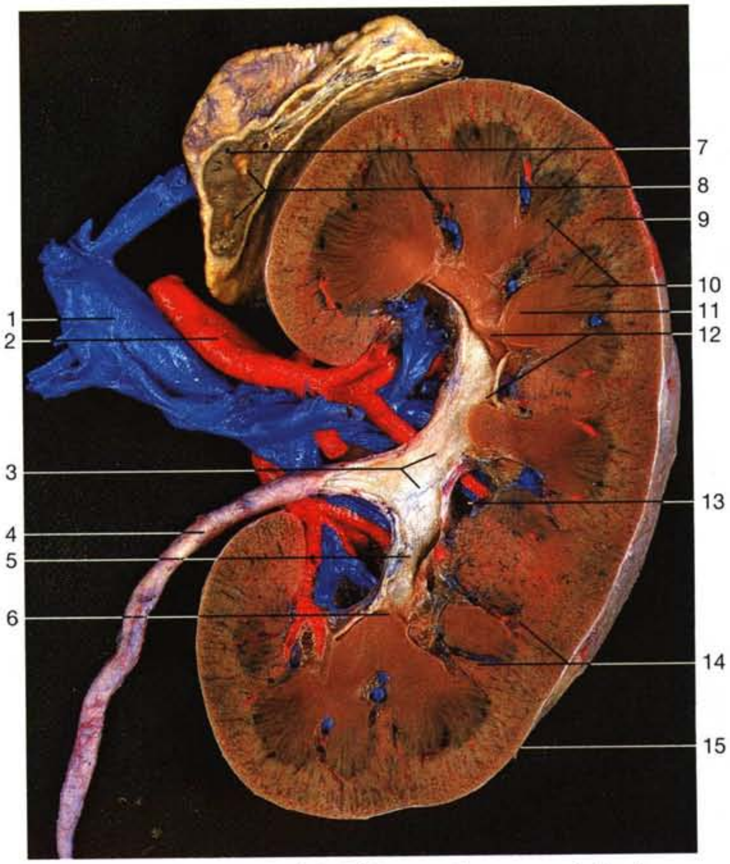

Coronal section through right kidney and suprarenal gland (posterior view). The renal pelvis has been opened and the fatty tissue removed to display the renal vessels.

1 Renal vein
2 Renal artery
3 **Renal pelvis**
4 Abdominal part of ureter
5 **Major renal calyx**
6 Cribriform area of renal papilla
7 Cortex of suprarenal gland
8 Medulla of suprarenal gland
9 **Cortex of kidney**
10 **Medulla of kidney**
11 Renal papilla
12 **Minor renal calyx**
13 Renal sinus
14 Renal columns
15 **Fibrous capsule** of kidney

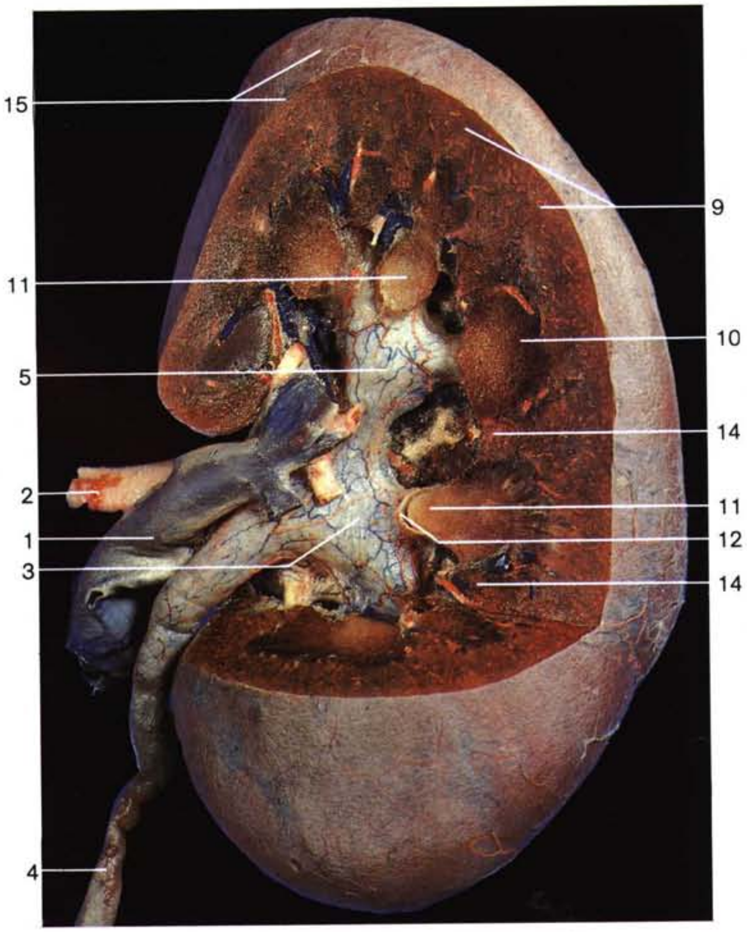

Right kidney (posterior view). Partial coronal section to expose internal aspect of the kidney.

Each kidney can be divided in five segments supplied by individual interlobar arteries considered as end arteries. Thus, obstruction leads to infarcts marking the trace of segment borders. The anterior kidney surface reveals four segments, the posterior only three (No. 1, 4 and 5).

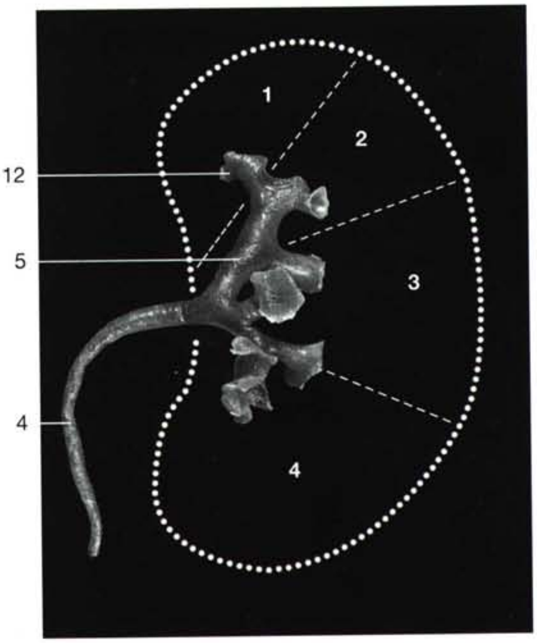

Cast of renal pelvis and calices.
1–4 = Renal segments on anterior surface.

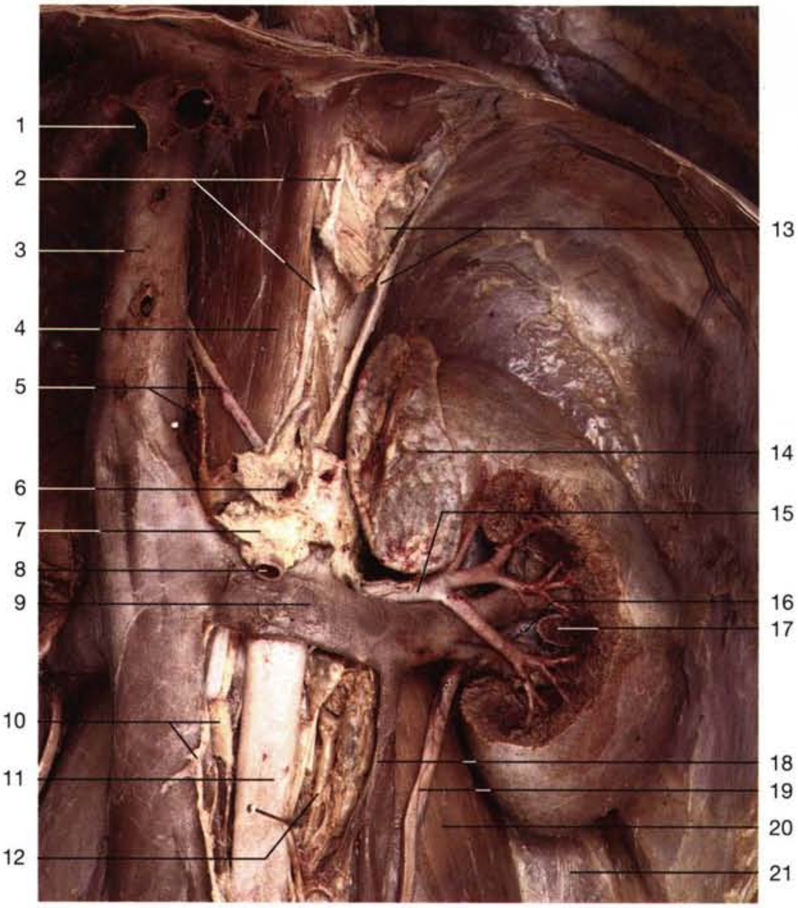

Left kidney and suprarenal gland in situ. The anterior cortical layer of the kidney has been removed to display the renal pelvis and papillae.

1	Hepatic vein
2	Anterior and posterior vagal trunk
3	Inferior vena cava
4	Lumbar part of diaphragm
5	Right greater and lesser splanchnic nerves
6	Celiac trunk
7	Celiac ganglion and plexus
8	Superior mesenteric artery
9	Left renal vein
10	Right sympathetic trunk and ganglion
11	Abdominal aorta
12	Left sympathetic trunk
13	Esophagus (cut), left greater splanchnic nerve
14	**Left suprarenal gland**
15	Left renal artery
16	Renal pelvis
17	Renal papilla with minor calyx
18	Left testicular vein
19	Left ureter
20	Psoas major muscle
21	Quadratus lumborum muscle
22	**Glomerulus**
23	Afferent arteriole of glomerulus
24	**Glomeruli**
25	Radiating cortical artery
26	Subcortical or arcuate artey
27	Subcortical or arcuate vein
28	Interlobular vein
29	Interlobular artery
30	**Interlobar artery and vein**
31	Vessels of renal capsule
32	Efferent arteriole of glomerulus
33	**Vasa recta** of renal medulla
34	Spiral arteries of renal pelvis

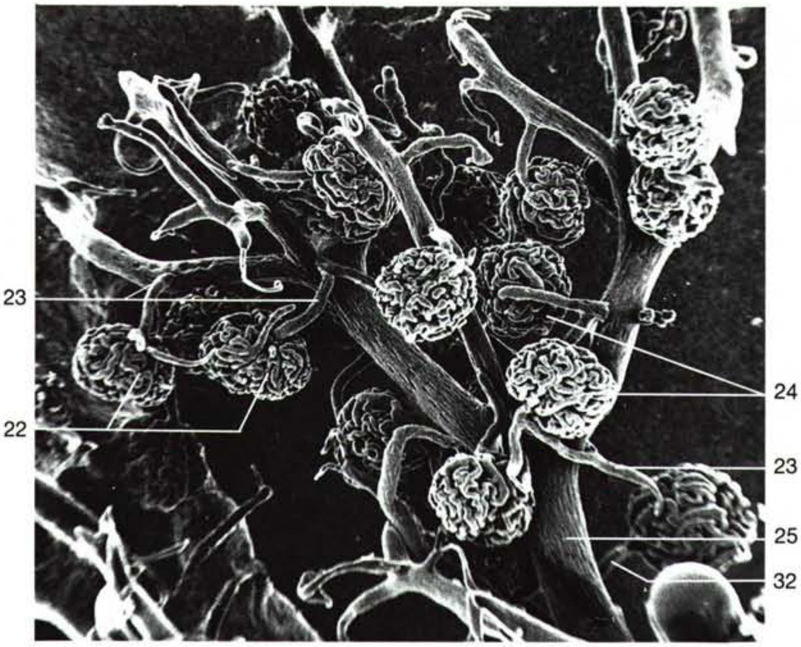

Glomeruli (210×). Scanning electron micrograph showing glomeruli and associated arteries.

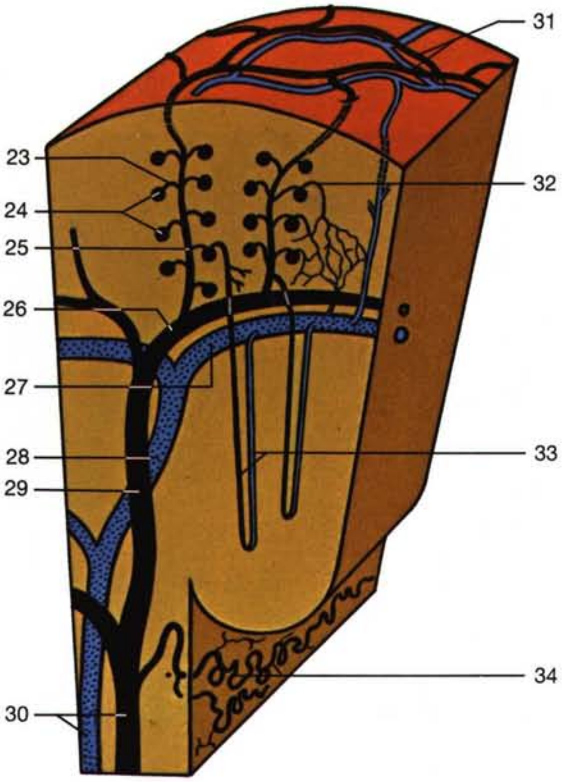

Architecture of vascular system of kidney. (Schematic drawing.)

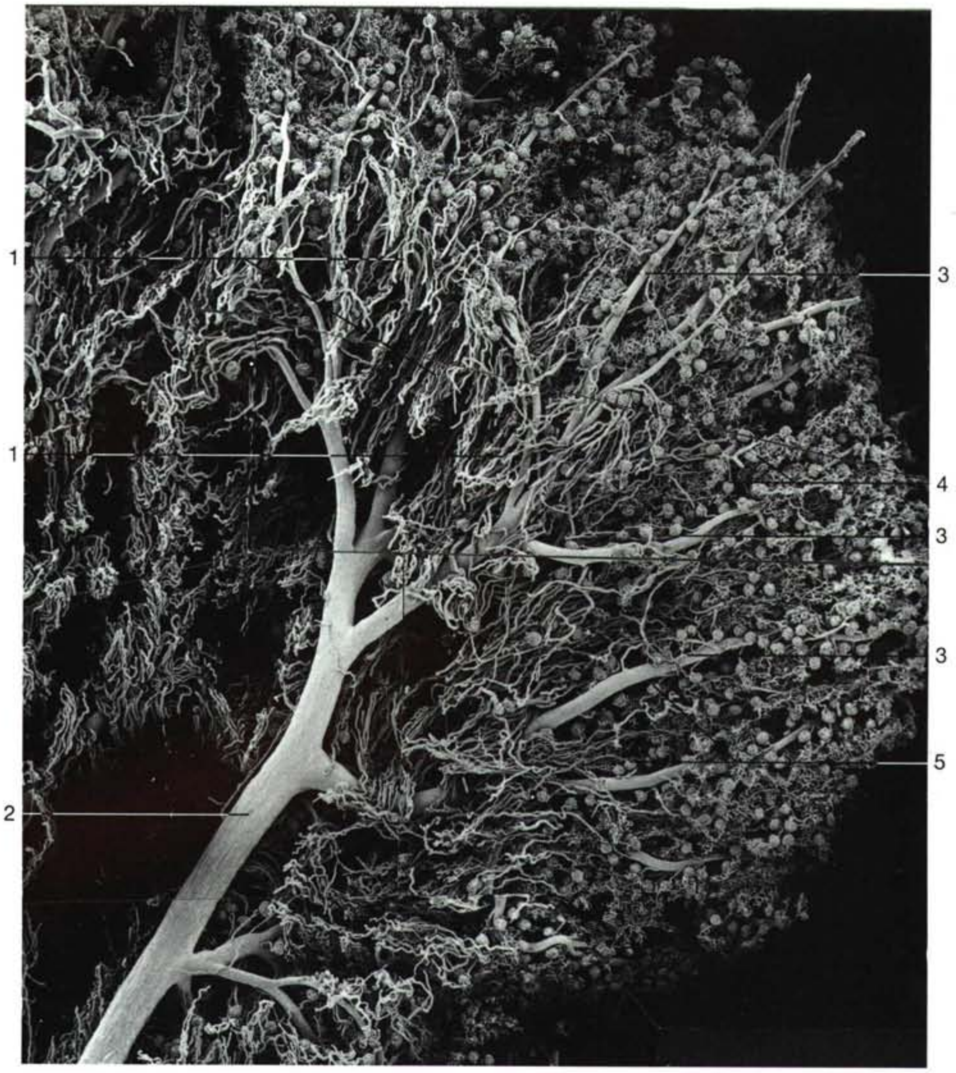

Resin cast of kidney arteries. (Scanning electron micrograph.)

1 Arteriolae rectae of renal medulla
2 Interlobar artery
3 Interlobular arteries
4 Cortical glomeruli
5 Juxtamedullary glomeruli
6 Body of first lumbar vertebra
7 Left renal artery
8 Abdominal aorta with catheter
9 Upper pole of kidney
10 Anterior branch } of renal artery
11 Posterior branch } of renal artery
12 Anterior inferior segmental artery
13 Lower pole of kidney
14 Celiac trunk
15 Superior mesenteric artery
16 Middle colic artery
17 Splenic artery

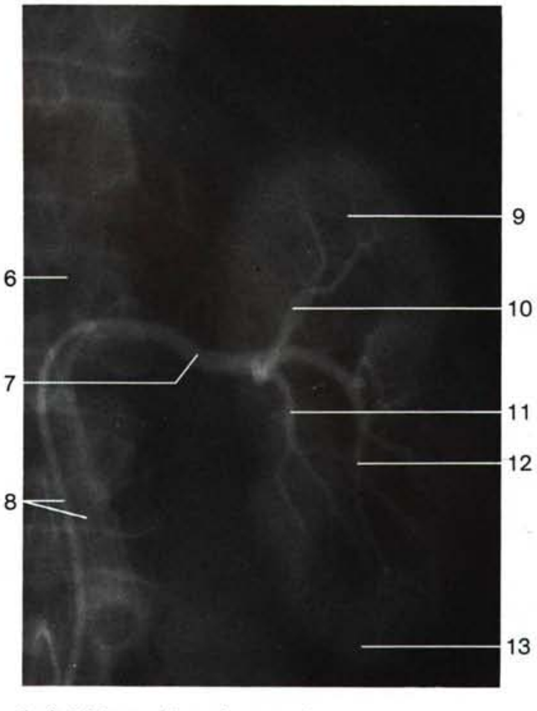

Left kidney. (Arteriogram.)

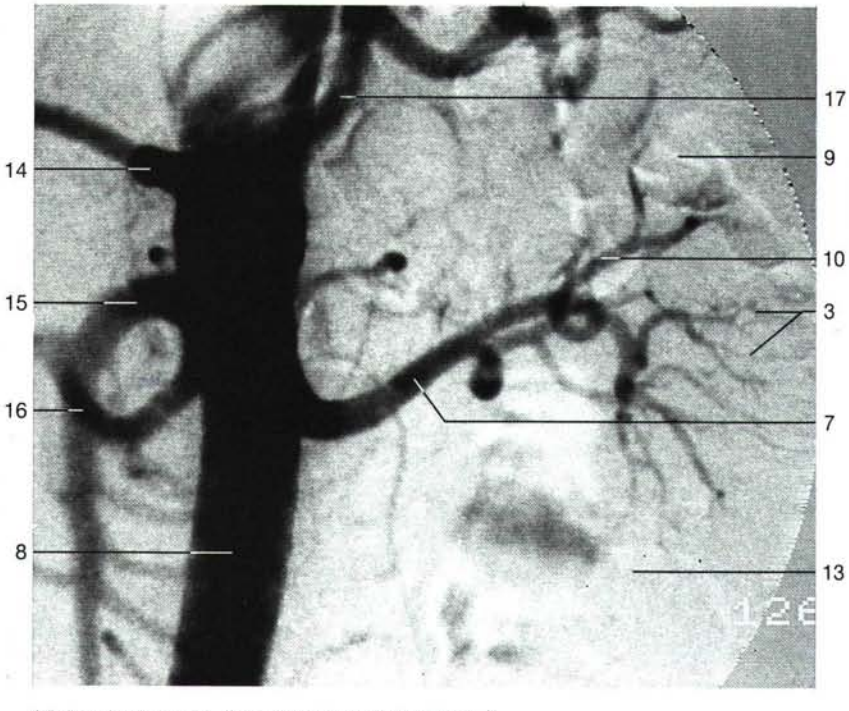

Abdominal aorta. (Subtraction angiograph.)

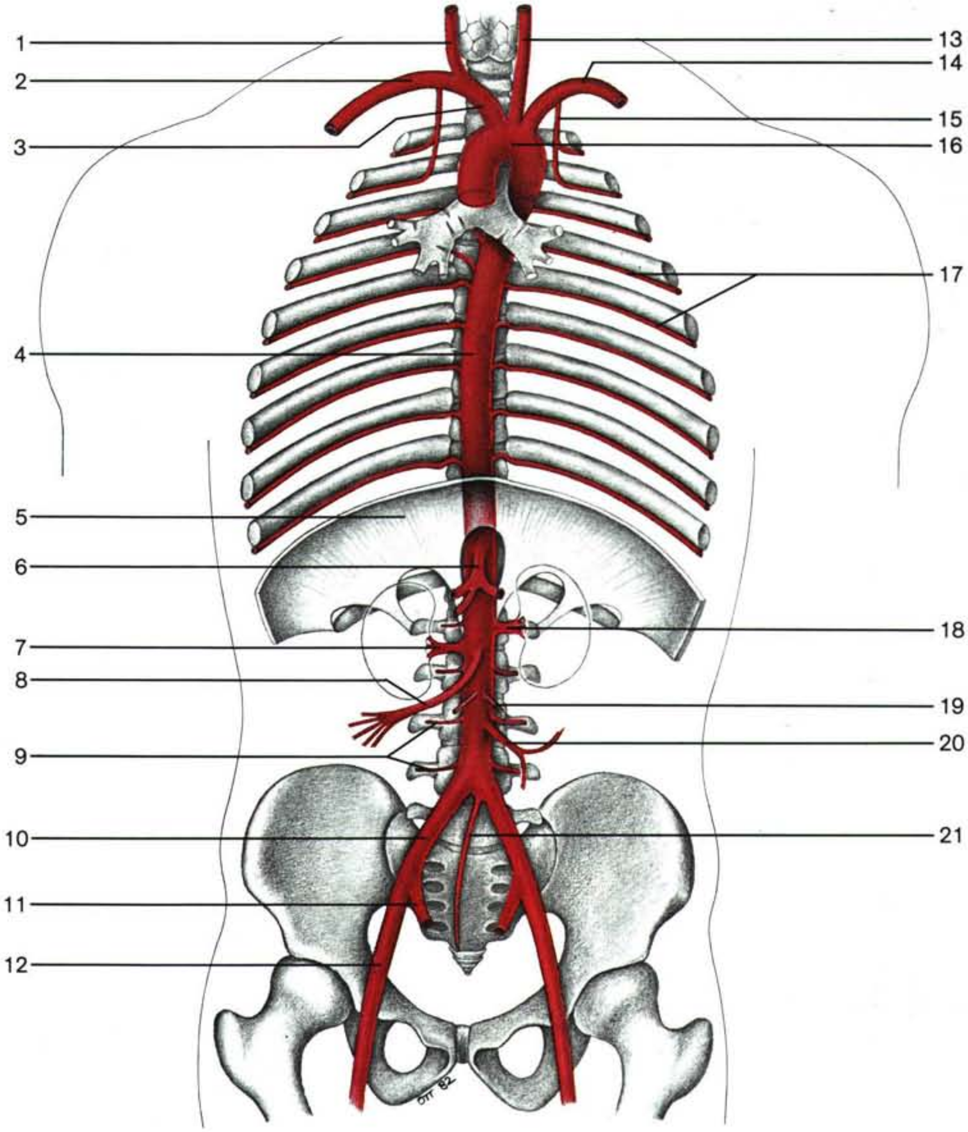

Main branches of descending aorta. (Schematic drawing.)

1	Right common carotid artery
2	Right subclavian artery
3	**Brachiocephalic trunk**
4	Thoracic aorta
5	Diaphragm
6	**Celiac trunk**
7	Right **renal artery**
8	**Superior mesenteric artery**
9	Lumbar arteries
10	Right **common iliac artery**
11	Internal iliac artery
12	**External iliac artery**
13	Left common carotid artery
14	Left subclavian artery
15	Highest intercostal artery
16	Aortic arch
17	**Posterior intercostal arteries**
18	Left renal artery
19	Left testicular (or ovarian) artery
20	Inferior mesenteric artery
21	Median sacral artery
22	**Superior suprarenal artery**
23	Upper capsular artery
24	Anterior branch of renal artery
25	Perforating artery
26	Lower capsular artery
27	Ureter
28	Right inferior phrenic artery
29	Left inferior phrenic artery
30	**Middle suprarenal artery**
31	**Inferior suprarenal artery**
32	Posterior branch of renal artery

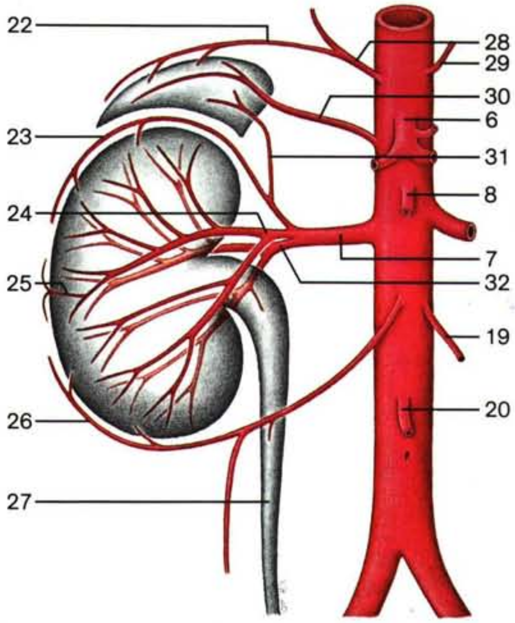

Arteries of kidney and suprarenal gland. (Schematic drawing.)

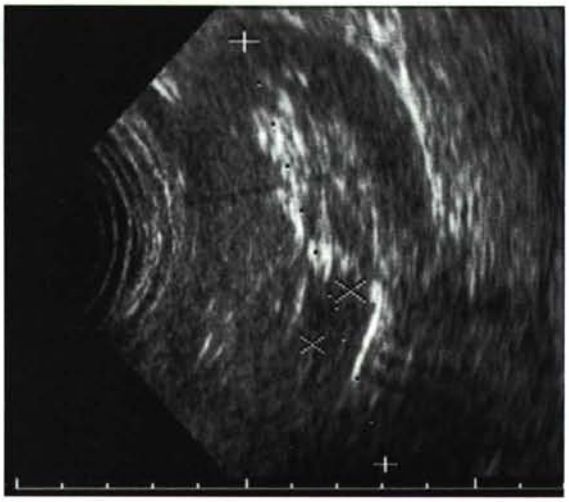

Ultrasound image of the right kidney (upper and lower border of the kidney marked by crosses. × = small cortical cyst).

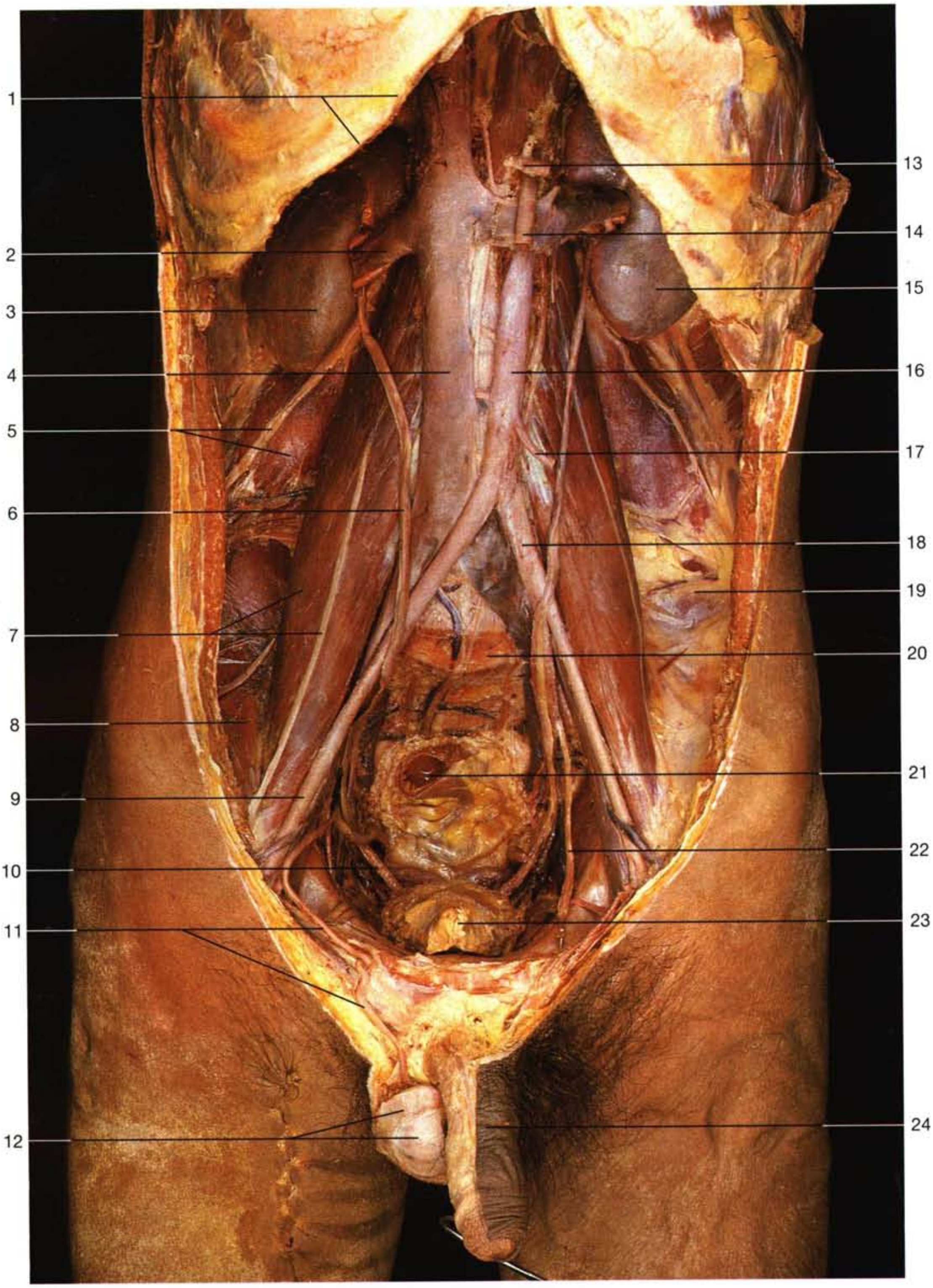

Retroperitoneal organs, urinary system in the male (anterior view). The peritoneum has been removed.

1 Costal arch
2 Right renal vein
3 **Right kidney**
4 Inferior vena cava
5 Iliohypogastric nerve and quadratus lumborum
6 **Ureter** (abdominal part)
7 Psoas major muscle and genitofemoral nerve
8 Iliacus muscle
9 External iliac artery
10 **Ureter** (pelvic part)
11 Ductus deferens
12 Testis and epididymis
13 Celiac trunk
14 Superior mesenteric artery
15 **Left kidney**
16 Abdominal aorta
17 Inferior mesenteric artery
18 Common iliac artery
19 Iliac crest
20 Sacral promontory
21 Rectum (cut)
22 Medial umbilical ligament
23 **Urinary bladder**
24 Penis

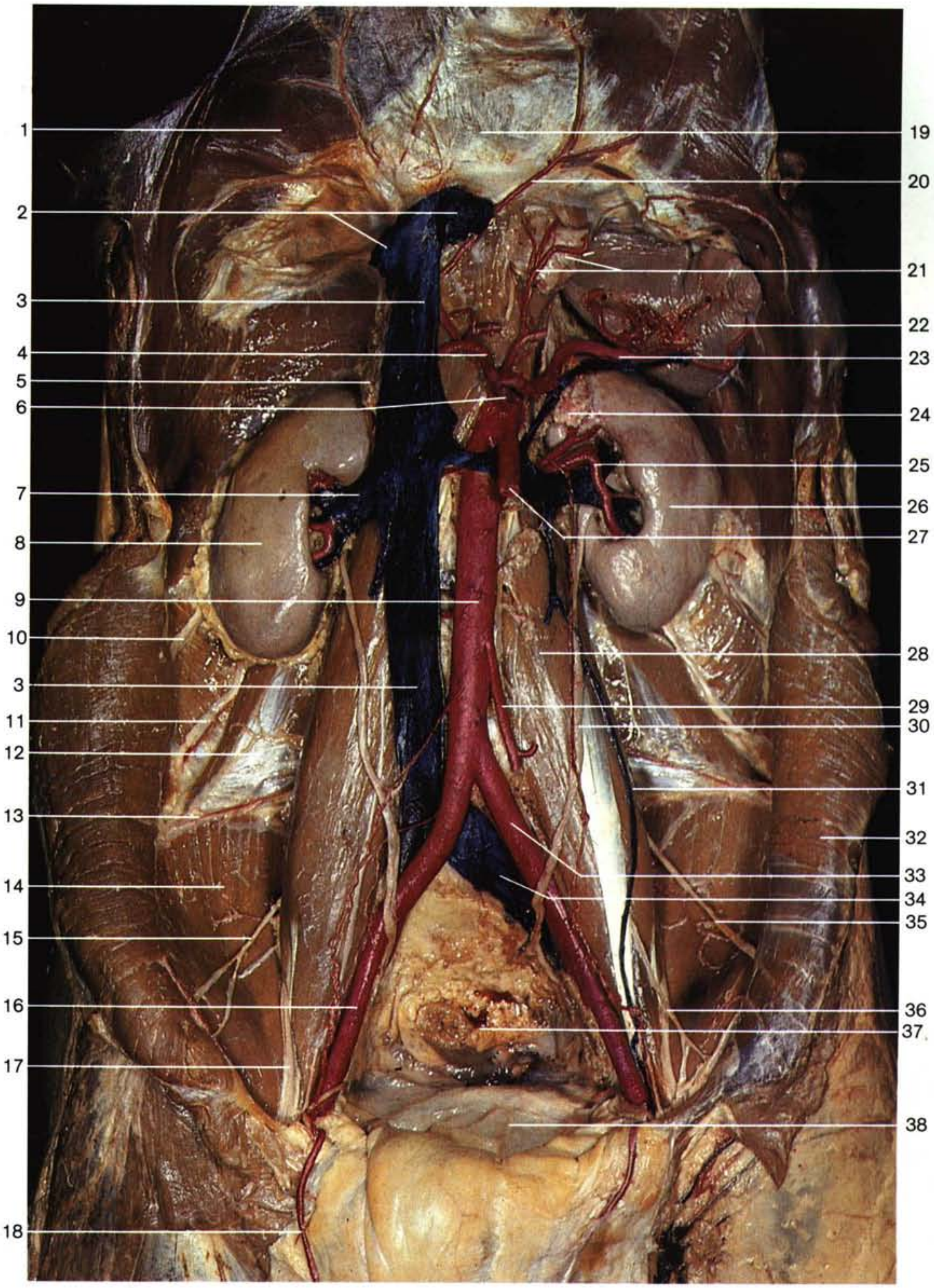

Retroperitoneal organs, urinary system in situ (anterior view). The peritoneum has been removed. Red = arteries; blue = veins.

1 Diaphragm
2 Hepatic veins
3 **Inferior vena cava**
4 Common hepatic artery
5 Right suprarenal gland
6 Celiac trunk
7 **Right renal vein**
8 Right kidney
9 **Abdominal aorta**
10 Subcostal nerve
11 Iliohypogastric nerve
12 Quadratus lumborum muscle
13 Iliac crest
14 Iliacus muscle
15 Right lateral femoral cutaneous nerve
16 **External iliac artery**
17 Femoral nerve
18 Right inferior epigastric artery
19 Central tendon of diaphragm
20 Inferior phrenic artery
21 Cardiac part of stomach and esophageal branches of left gastric artery
22 Spleen
23 Splenic artery
24 Left suprarenal gland
25 **Left renal artery**
26 Left **kidney**
27 Superior mesenteric artery
28 Psoas major muscle
29 Inferior mesenteric artery
30 **Ureter**
31 Testicular artery and vein
32 Transversus abdominis muscle
33 Left common iliac artery
34 Left common iliac vein
35 Lateral femoral cutaneous nerve
36 Genitofemoral nerve
37 Rectum (cut)
38 Urinary bladder

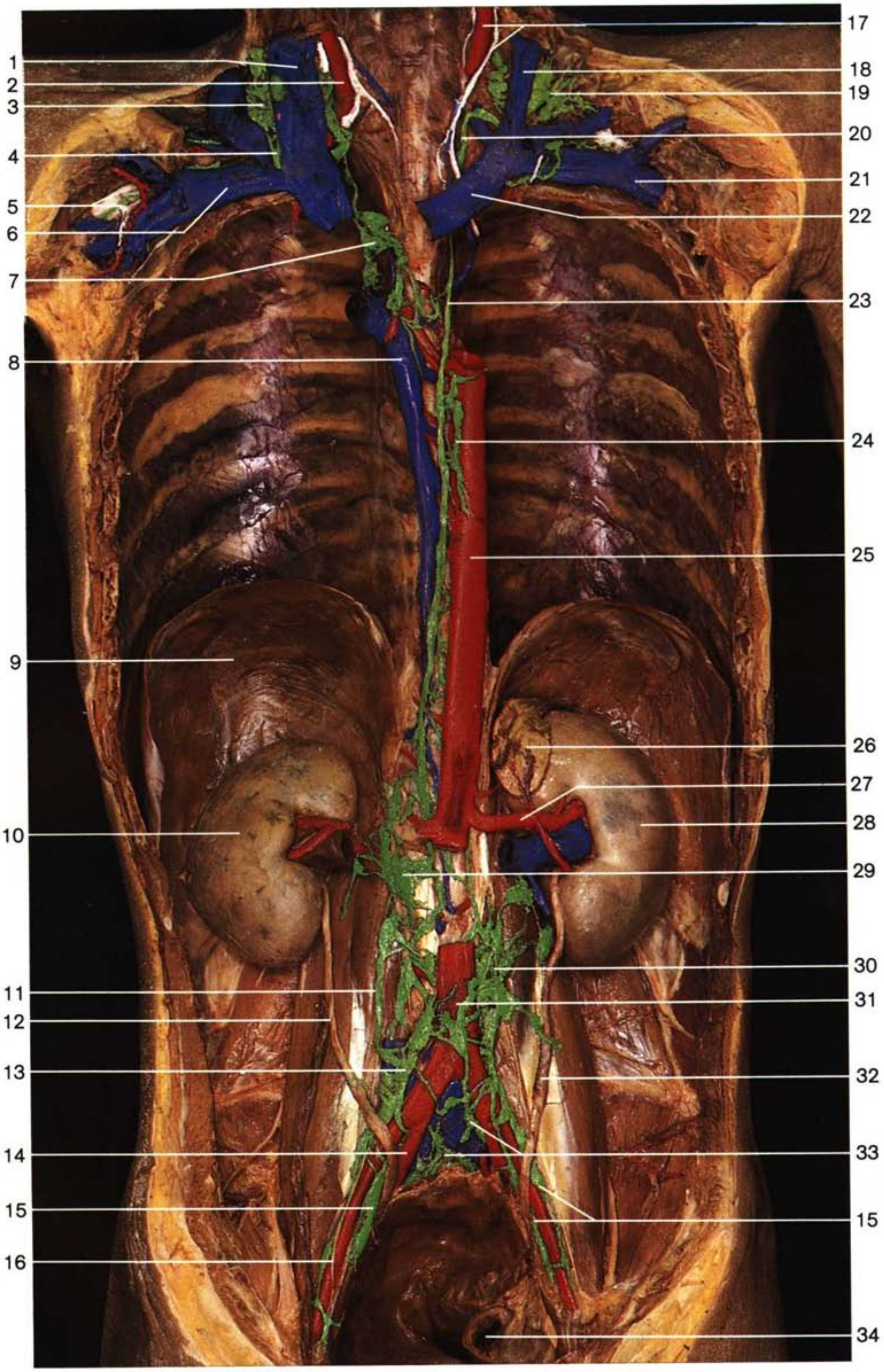

1	Internal jugular vein
2	Right common carotid artery and right vagus nerve
3	Jugulo-omohyoid lymph node
4	Right lymphatic duct
5	Subclavian trunk
6	Right subclavian vein
7	Bronchomediastinal trunk
8	Azygos vein
9	Diaphragm
10	Right kidney
11	Right lumbar trunk
12	Right ureter
13	Common iliac lymph nodes
14	Right internal iliac artery
15	External iliac lymph nodes
16	Right external iliac artery
17	Left common carotid artery and left vagus nerve
18	Internal jugular vein
19	Deep cervical lymph nodes
20	Thoracic duct entering left jugular angle
21	Left subclavian vein
22	Left brachiocephalic vein
23	**Thoracic duct**
24	Mediastinal lymph nodes
25	Thoracic aorta
26	Left suprarenal gland
27	Left renal artery
28	Left kidney
29	Cisterna chyli
30	Lumbar lymph nodes
31	Abdominal aorta
32	Left ureter
33	Sacral lymph nodes
34	Rectum (cut edge)

Lymph vessels and lymph nodes of the posterior wall of thoracic and abdominal cavities (anterior aspect). Green = lymph vessels and nodes; blue = veins; red = arteries; white = nerves.

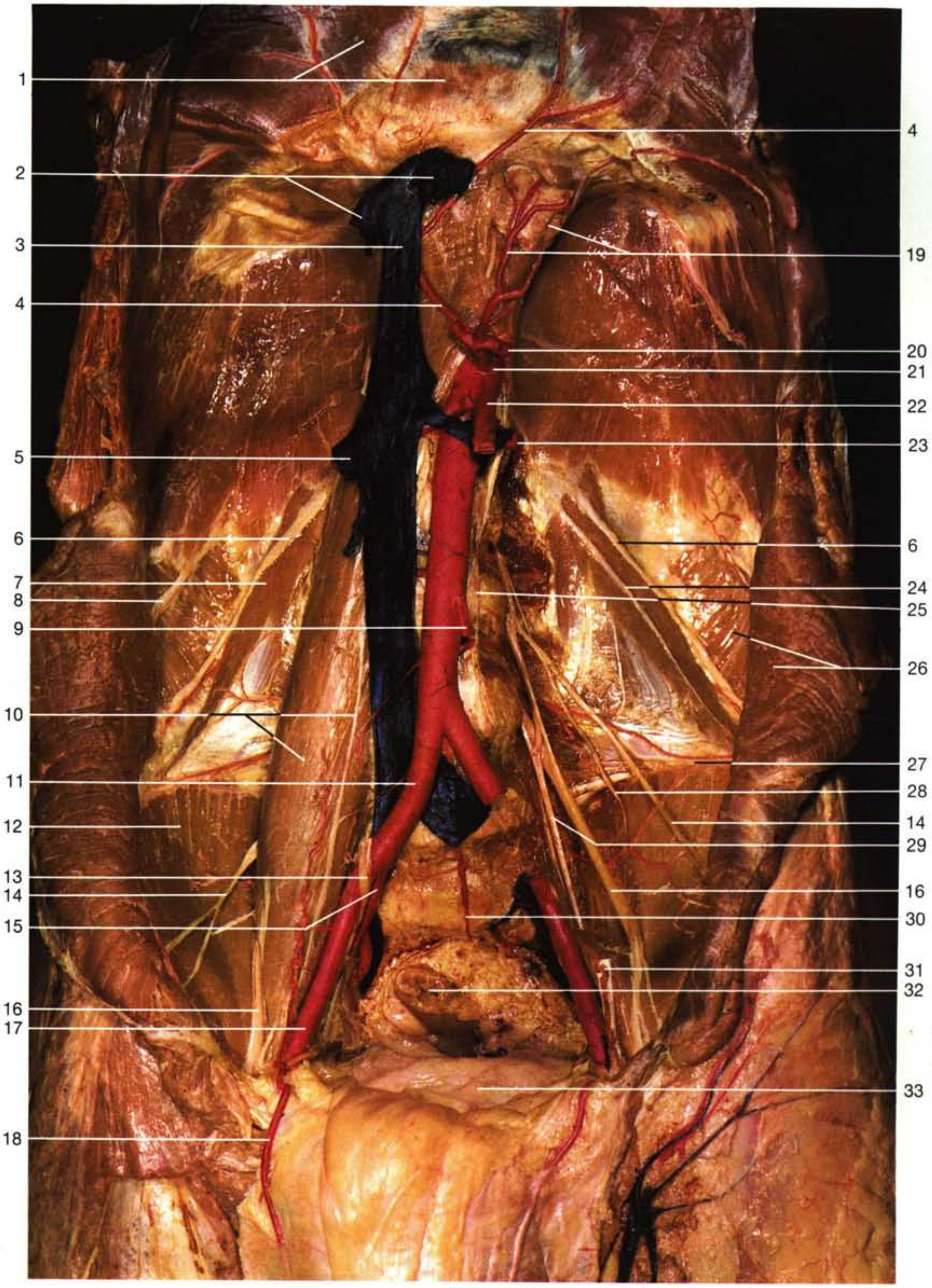

Vessels and nerves of posterior abdominal wall (anterior aspect). Part of the left psoas major muscle has been removed to display the lumbar plexus. Red = arteries, blue = veins.

1 Diaphragm
2 Hepatic veins
3 **Inferior vena cava**
4 Inferior phrenic artery
5 Right renal vein
6 **Iliohypogastric nerve**
7 Quadratus lumborum muscle
8 Subcostal nerve
9 Inferior mesenteric artery
10 Right genitofemoral nerve and psoas major muscle
11 Common iliac artery
12 Iliacus muscle
13 Right ureter (divided)
14 Lateral femoral cutaneous nerve
15 Internal iliac artery
16 **Femoral nerve**
17 **External iliac artery**
18 Inferior epigastric artery
19 Cardiac part of stomach and oesophageal branches of left gastric artery
20 Splenic artery
21 Celiac trunk
22 **Superior mesenteric artery**
23 Left renal artery
24 **Ilioinguinal nerve**
25 Sympathetic trunk
26 Transversus abdominis muscle
27 Iliac crest
28 Left **genitofemoral nerve**
29 Left **obturator nerve**
30 Median sacral artery
31 Psoas major muscle (divided) with supplying artery
32 Rectum (cut)
33 Urinary bladder

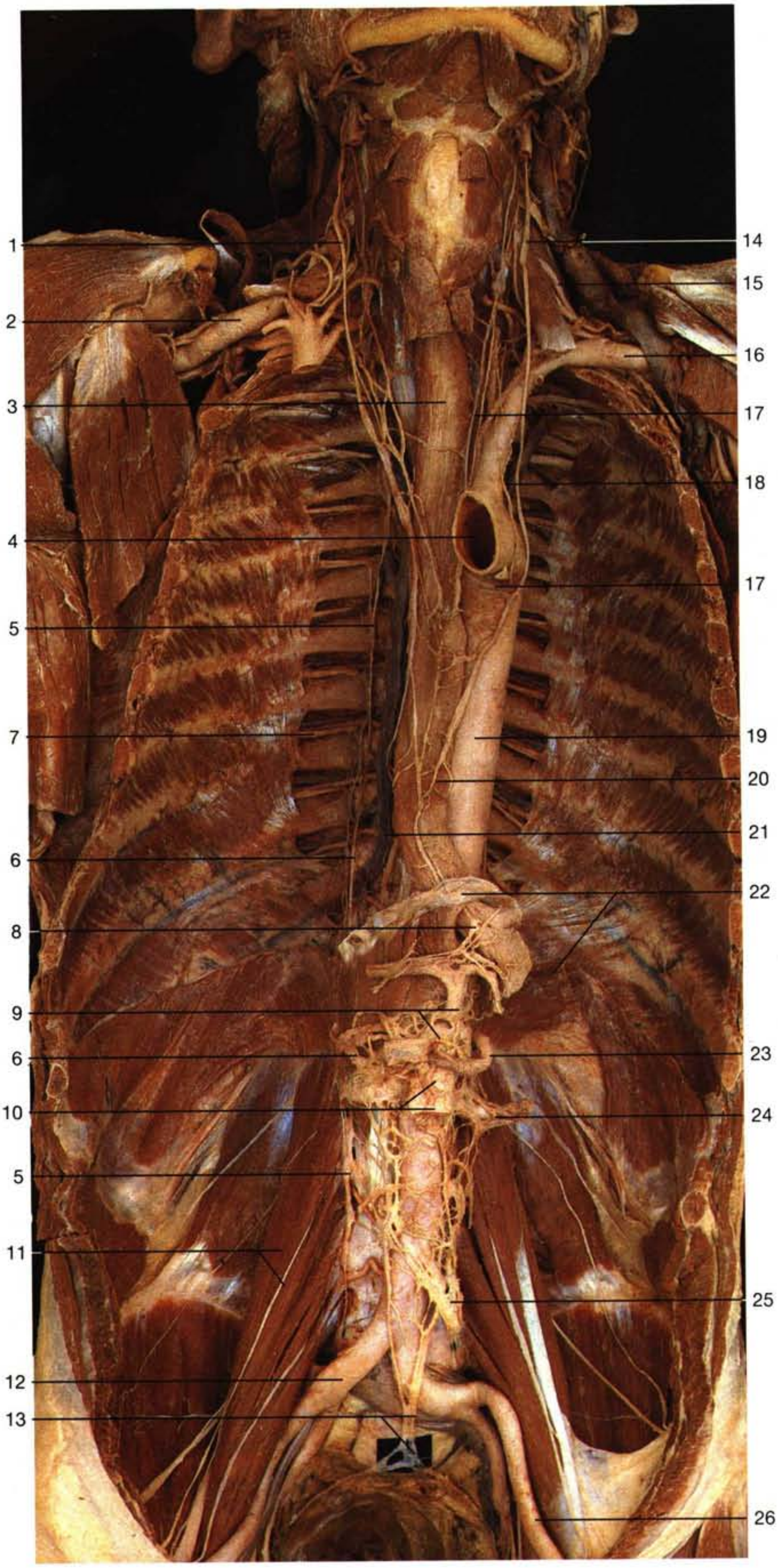
1
2
3
4
5
7
6
8
9
6
10
5
11
12
13
14
15
16
17
18
17
19
20
21
22
23
24
25
26

Posterior wall of thoracic and abdominal cavities with sympathetic trunk, vagus nerve and autonomic ganglia (anterior view). Thoracic and abdominal organs removed, except esophagus and aorta.

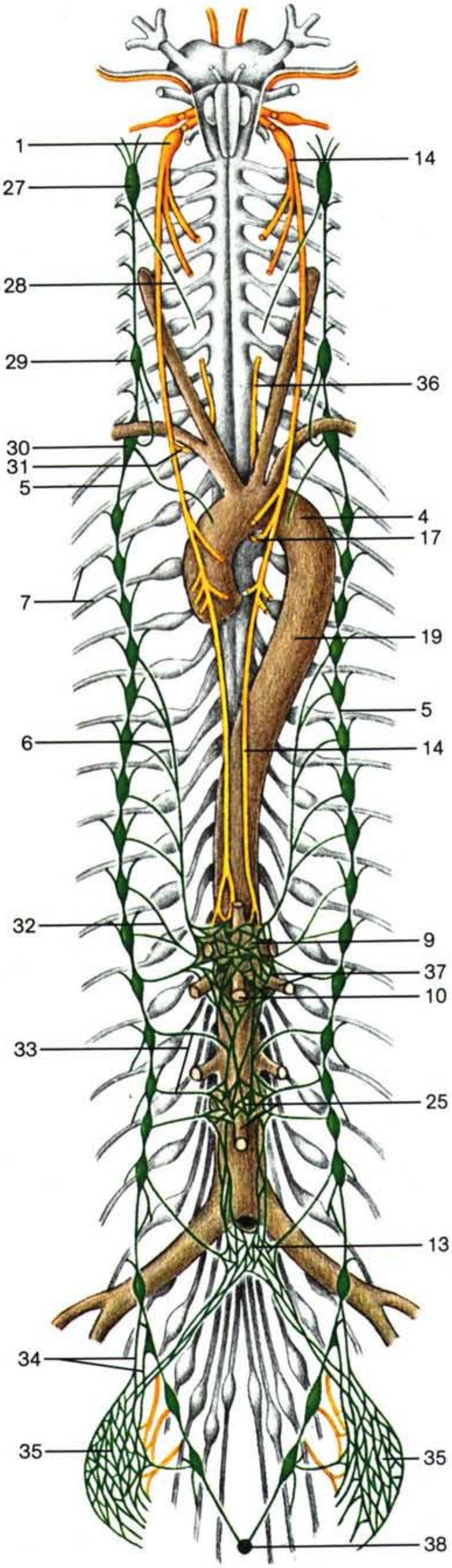
1
27
28
29
30
31
5
7
6
32
33
34
35
14
36
4
17
19
5
14
9
37
10
25
13
35
38

Organization of autonomic nervous system (after Mattuschka). (Schematic drawing.) Yellow = parasympathetic nerves; green = sympathetic nerves.

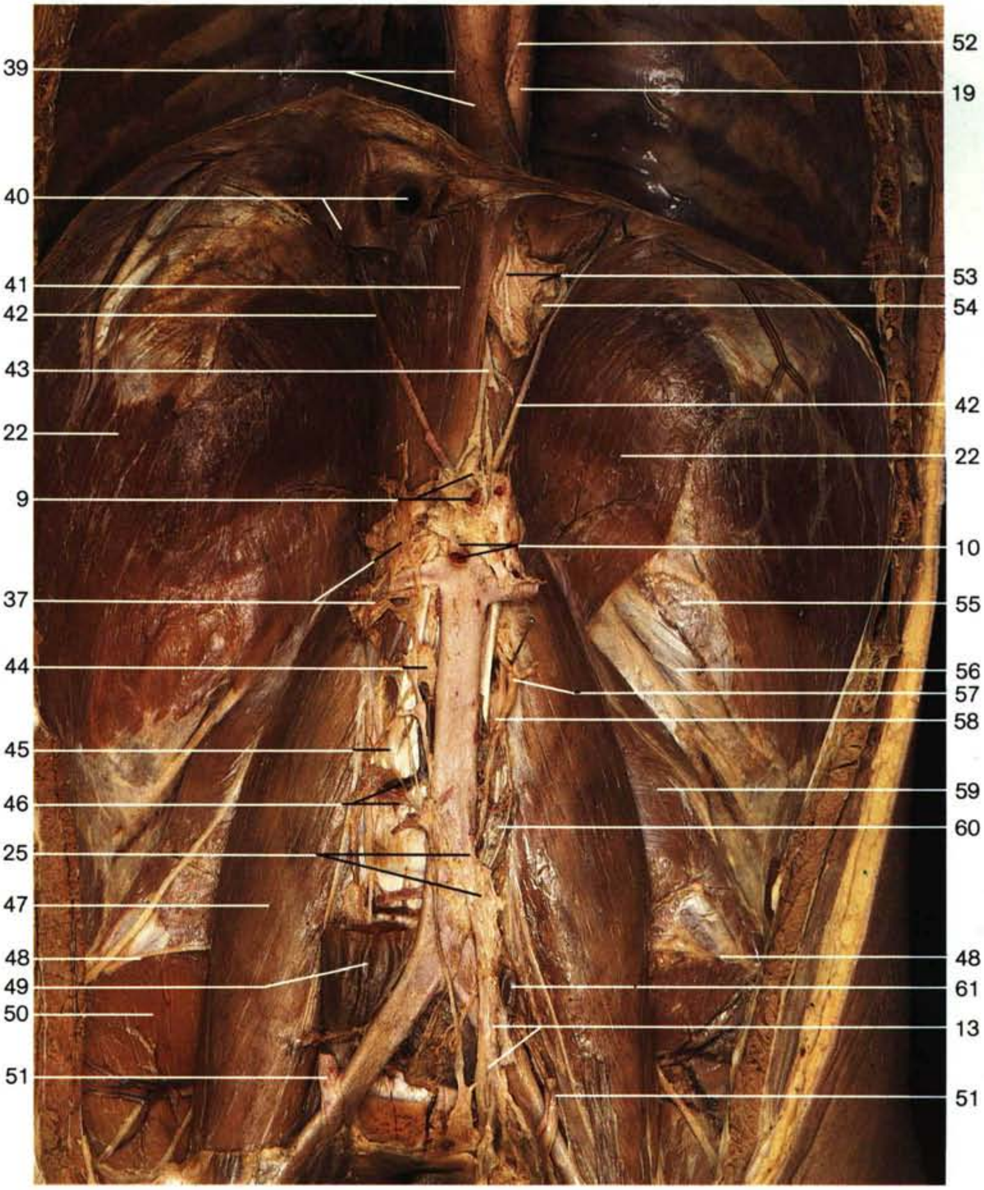

Ganglia and plexus of the autonomic nervous system within the retroperitoneal space (anterior view). The kidneys and the inferior vena cava with its tributaries have been removed (compare pp. 264 and 265).

1 **Right vagus nerve**
2 Right subclavian artery
3 Esophagus
4 Aortic arch
5 **Sympathetic trunk**
6 Greater splanchnic nerve
7 Intercostal nerve
8 Abdominal part of esophagus and vagal trunk
9 **Celiac trunk with celiac ganglion**
10 **Superior mesenteric artery and ganglion**
11 Psoas major muscle and genitofemoral nerve
12 Common iliac artery
13 **Superior hypogastric plexus and ganglion**
14 **Left vagus nerve**
15 Brachial plexus
16 Left subclavian artery
17 Left recurrent laryngeal nerve
18 Inferior cervical cardiac nerve
19 Thoracic aorta
20 Esophageal plexus
21 Azygos vein
22 Diaphragm
23 Splenic artery
24 Left renal artery and plexus
25 **Inferior mesenteric ganglion and artery**
26 Left external iliac artery
27 **Superior cervical ganglion** of sympathetic trunk
28 Superior cardiac branch of sympathetic trunk
29 Middle cervical ganglion of sympathetic trunk
30 **Inferior cervical ganglion** of sympathetic trunk
31 Right recurrent laryngeal nerve
32 Lesser splanchnic nerve
33 Lumbar splanchnic nerves
34 Sacral splanchnic nerves
35 **Inferior hypogastric ganglion and plexus**
36 Left recurrent laryngeal nerve
37 **Aorticorenal plexus** and renal artery
38 Ganglion impar
39 Esophagus with branches of vagus nerve
40 Hepatic veins
41 Right crus of diaphragm
42 Inferior phrenic artery
43 Right vagus nerve entering the celiac ganglion
44 Right lumbar lymph trunk
45 Lumbar part of right sympathetic trunk
46 Lumbar artery and vein
47 **Psoas major muscle**
48 Iliac crest
49 Inferior vena cava
50 **Iliacus muscle**
51 Ureter
52 Left vagus nerve forming the esophageal plexus
53 Left vagus nerve forming the gastric plexus
54 Esophagus continuing into the cardiac part of stomach
55 Lumbocostal triangle
56 Position of 12th rib
57 Left lumbar lymph trunk
58 Ganglion of sympathetic trunk
59 **Quadratus lumborum muscle**
60 Lumbar part of left sympathetic trunk
61 Iliac lymph vessels

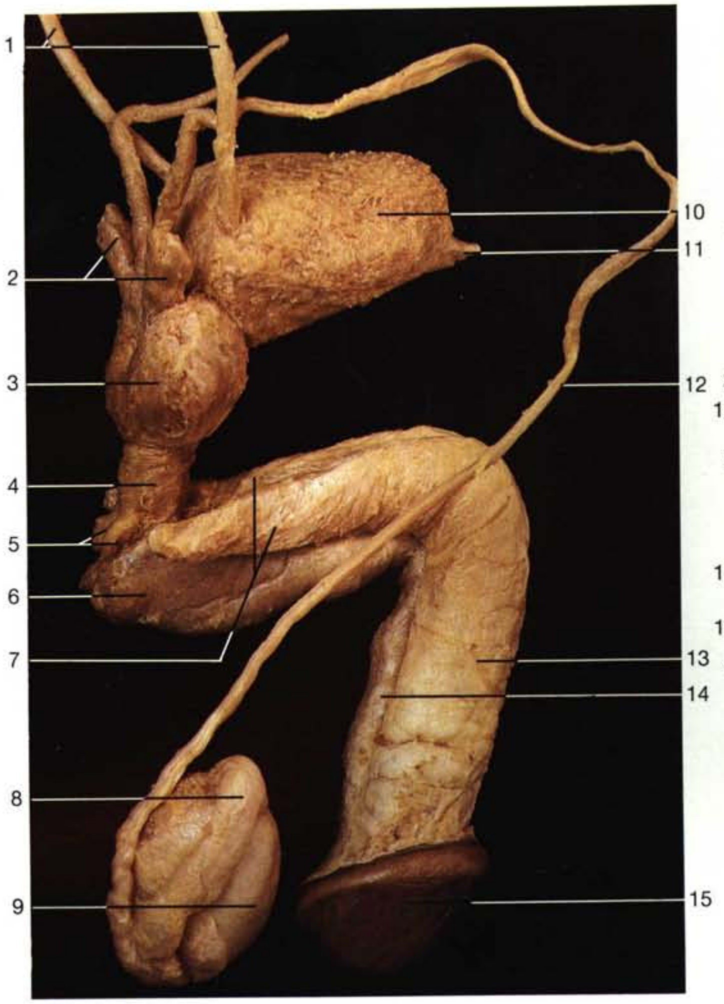
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Male genital organs isolated (right lateral aspect).

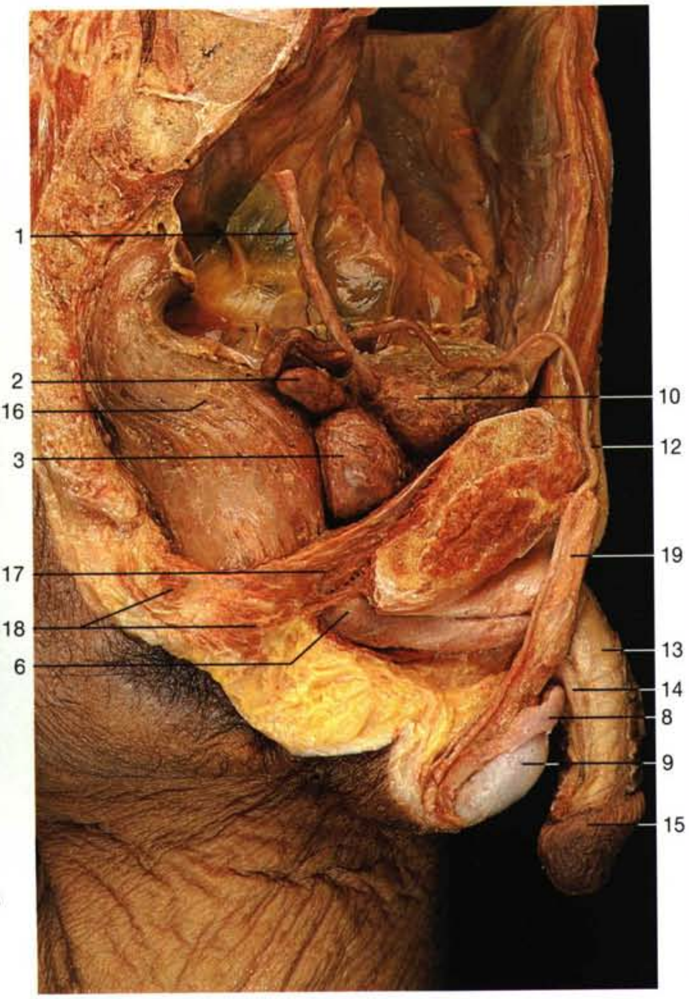
1
2
16
3
17
18
6
10
12
19
13
14
8
9
15

Male genital organs in situ (right lateral aspect).

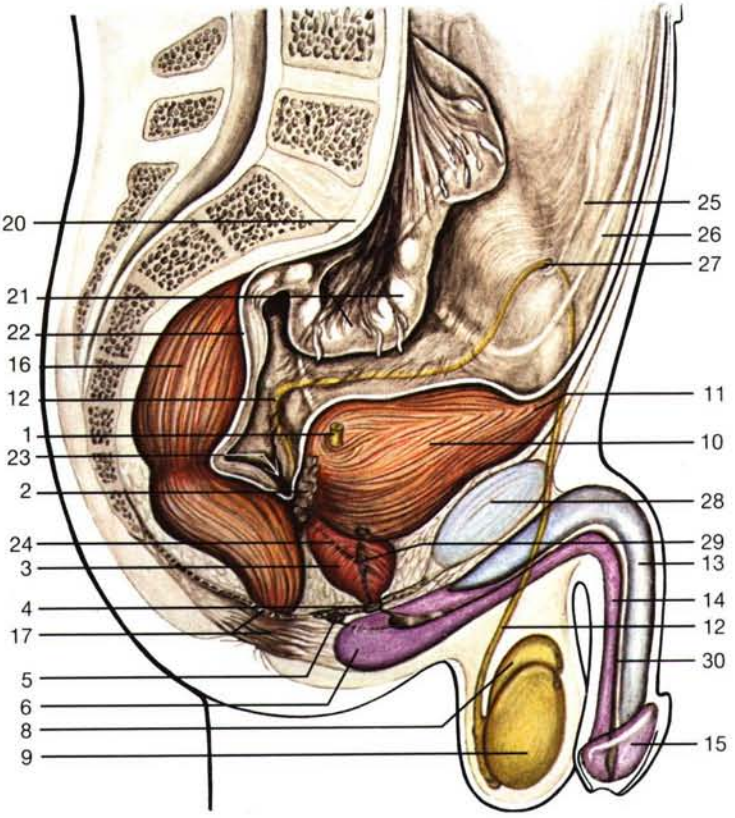
20
21
22
16
12
1
23
2
24
3
4
17
5
6
8
9
25
26
27
11
10
28
29
13
14
12
30
15

Positions of male genital organs (right lateral aspect).
(Schematic drawing.)

1 Ureter
2 **Seminal vesicle**
3 **Prostate gland**
4 Urogenital diaphragm and membranous part of urethra
5 Bulbourethral or Cowper's gland
6 Bulb of penis
7 Left and right crus penis
8 **Epididymis**
9 **Testis**
10 **Urinary bladder**
11 Apex of urinary bladder
12 **Ductus deferens**
13 **Corpus cavernosum of penis**
14 **Corpus spongiosum of penis**
15 Glans penis
16 Ampulla of rectum
17 Levator ani muscle
18 Anal canal and external anal sphincter muscle
19 Spermatic cord (cut)
20 Sacral promontory
21 Sigmoid colon
22 Peritoneum (cut edge)
23 **Rectovesical pouch**
24 Ejaculatory duct
25 Lateral umbilical fold
26 Medial umbilical fold
27 Deep inguinal ring and ductus deferens
28 Pubic symphysis
29 Prostatic part of urethra
30 Spongy urethra

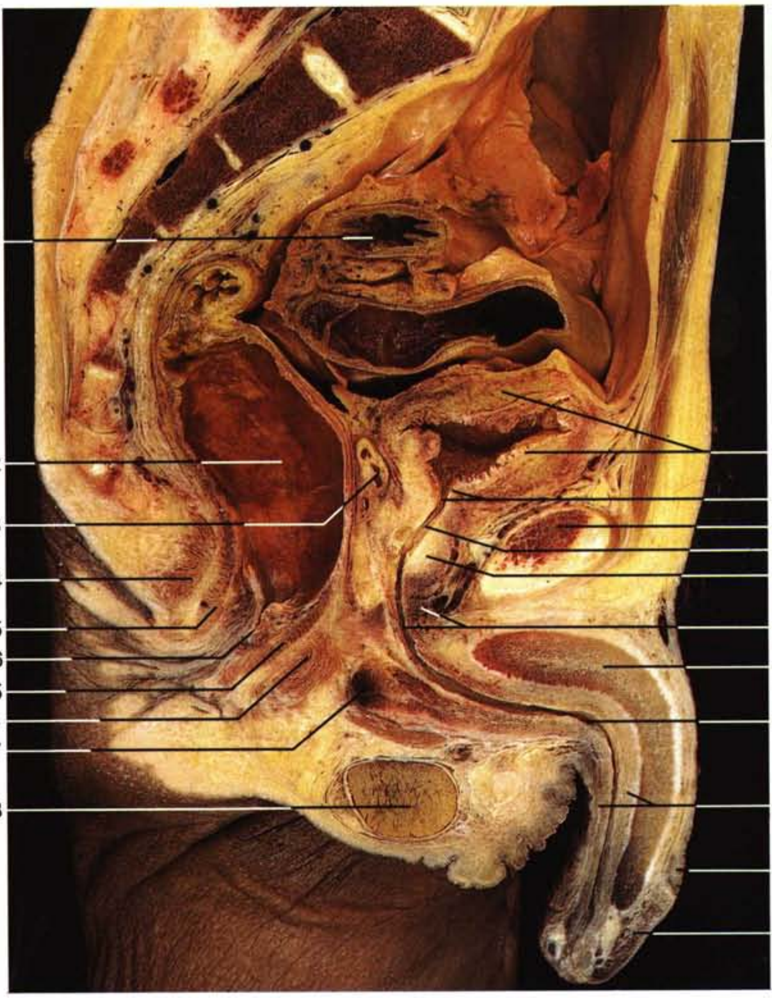

Midsagittal section through the male pelvis.

1 Sigmoid colon
2 **Ampulla of rectum**
3 Ampulla of ductus deferens
4 External anal sphincter muscle
5 Internal anal sphincter muscle
6 **Anal canal**
7 Bulb of penis
8 **Testis** (cut surface)
9 Median umbilical ligament
10 **Urinary bladder**
11 Internal urethral orifice and sphincter muscle (sphincter vesicae)
12 Pubic symphysis
13 Prostatic part of urethra
14 **Prostate gland**
15 Membranous urethra and external urethral sphincter muscle
16 **Corpus cavernosum of penis**
17 **Spongy urethra**
18 Corpus spongiosum of penis
19 Foreskin or prepuce
20 Glans penis
21 Kidney
22 Renal pelvis
23 Abdominal part of ureter
24 Pelvic part of ureter
25 Seminal vesicle
26 Ejaculatory duct
27 Bulbourethral or Cowper's gland
28 **Ductus deferens**
29 **Epididymis**
30 Umbilicus
31 Trigone of bladder and ureteric orifice
32 Navicular fossa of urethra
33 External urethral orifice
34 **Testis**

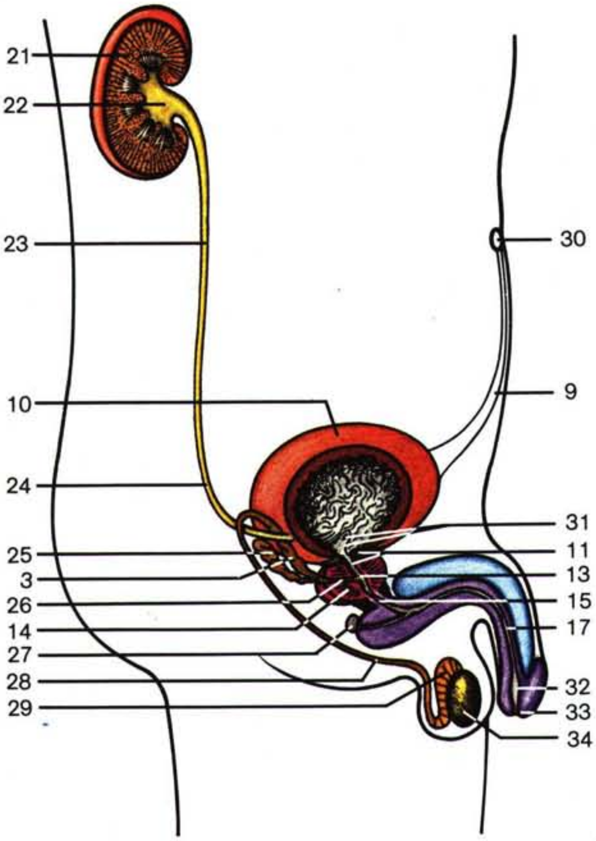

Male urogenital system. (Schematic drawing.)

The **prostate** is located between bladder and urogenital diaphragm. The penis includes the **urethra** and thus serves for both ejaculation and micturition. The internal (involuntary) and external (voluntary) urethral sphincters are widely separated. The **ureter** having crossed the ductus deferens enters the urinary bladder at its base. The peritoneum is reflected off the posterior surface of the bladder onto the rectum thus forming the rectovesical pouch.

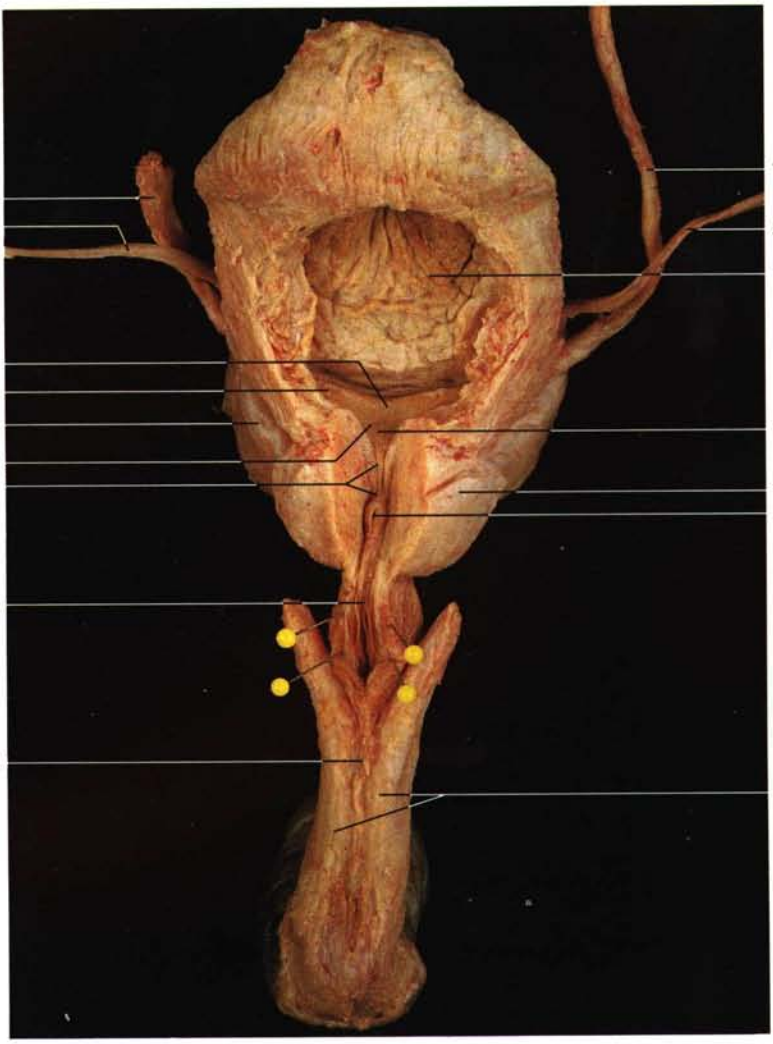

Male urogenital organs; isolated (anterior view). Urinary bladder, prostate and urethra have been opened. The urinary bladder is contracted.

1 Ureter
2 Ductus deferens
3 Interureteric fold
4 Ureteric orifice
5 Seminal vesicle
6 Trigone of bladder
7 Prostatic urethra with seminal colliculus and urethral crest
8 Deep transverse perineal muscle
9 Membranous urethra
10 **Spongy urethra**
11 Mucous membrane of **urinary bladder**
12 Internal urethral orifice and uvula of bladder
13 **Prostate**
14 Prostatic utricle
15 Right and left **corpus cavernosum of penis**
16 Ejaculatory duct
17 **Sphincter urethrae muscle**
18 **Median umbilical fold** with remnant of urachus
19 Medial umbilical fold with remnant of umbilical artery
20 **Urinary bladder**
21 Rectovesical pouch
22 Rectum
23 Sacrum
24 Deep iliac circumflex artery
25 Deep inguinal ring and ductus deferens
26 External iliac artery and vein
27 **Femoral nerve**
28 Obturator nerve
29 Ilium and sacrum
30 Inferior epigastric artery
31 Iliopsoas muscle

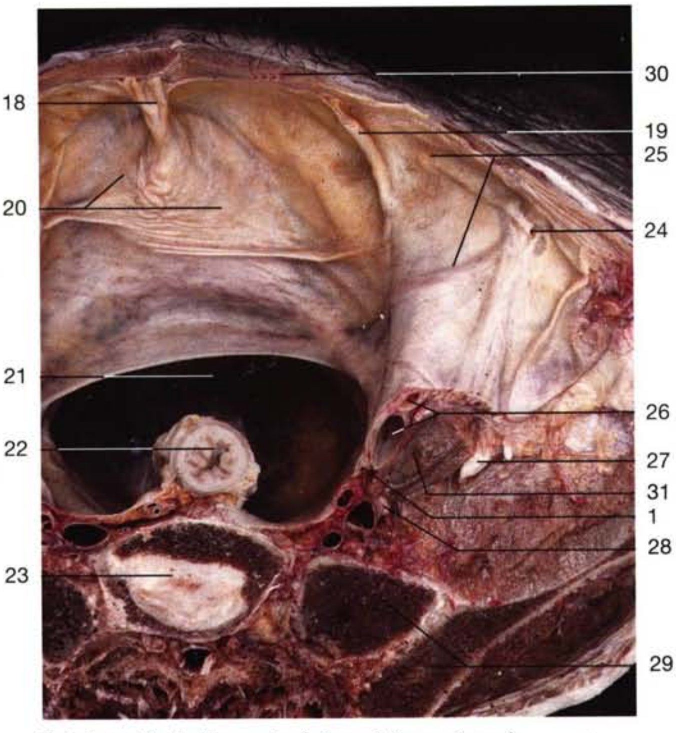

Pelvic cavity in the male (viewed from above).

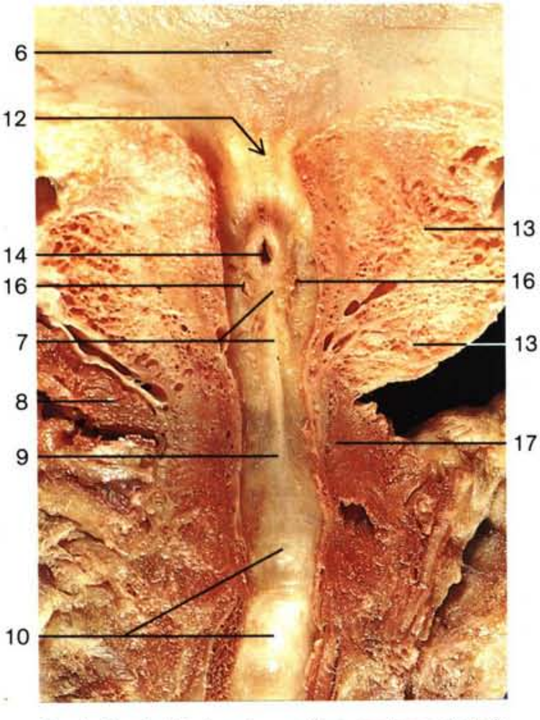

Posterior half of male urethra and prostate in continuity with neck of bladder (anterior view).

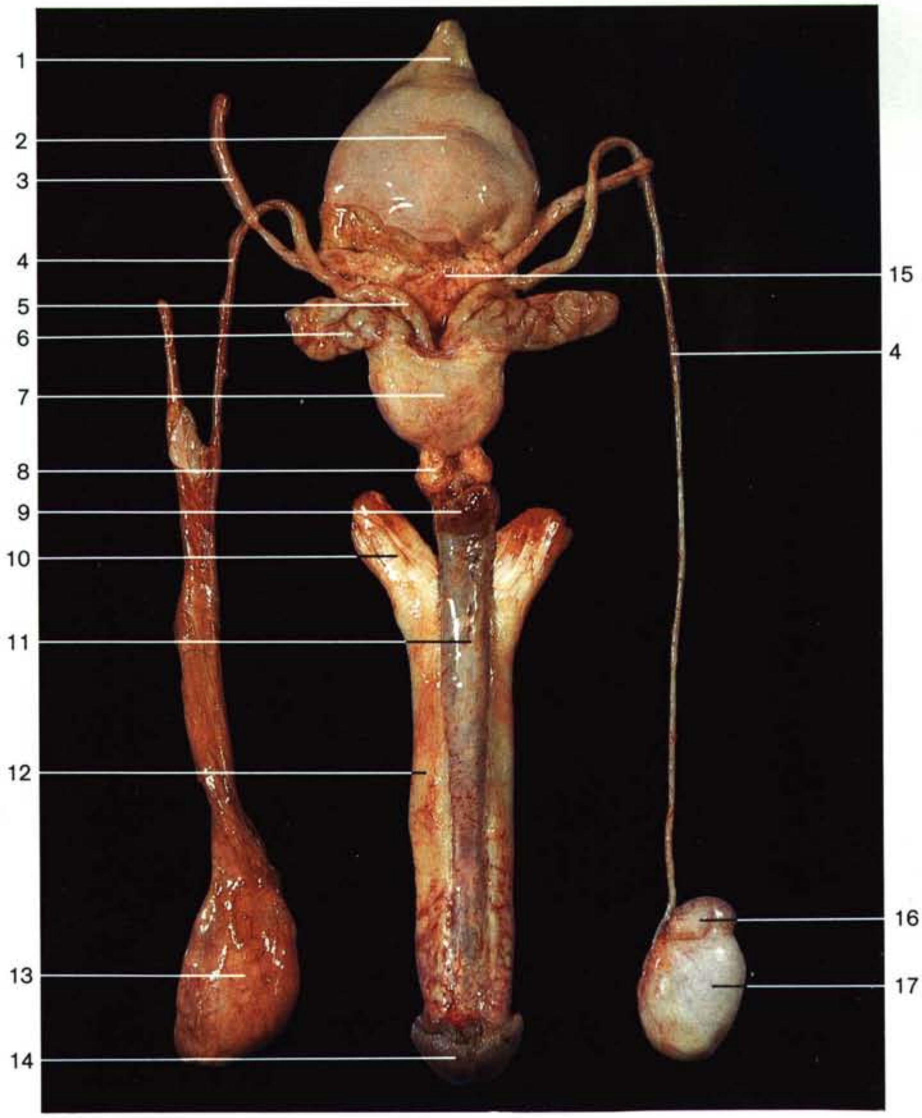

Male genital organs, isolated (posterior view).

1 Apex of urinary bladder with urachus
2 **Urinary bladder**
3 Ureter
4 **Ductus deferens**
5 Ampulla of ductus deferens
6 **Seminal vesicle**
7 **Prostate**
8 **Bulbourethral or Cowper's gland**
9 **Bulb of penis**
10 Crus penis
11 **Corpus spongiosum of penis**
12 **Corpus cavernosum of penis**
13 Testis and epididymis with coverings
14 Glans penis
15 Fundus of bladder
16 **Head of epididymis**
17 **Testis**
18 Mucous membrane of bladder
19 Trigone of bladder
20 Ureteric orifice
21 Internal urethral orifice
22 Seminal colliculus
23 Prostate
24 Prostatic urethra
25 Membranous urethra
26 Spongy (penile) urethra
27 Skin of penis
28 Deep dorsal vein of penis (unpaired)
29 Dorsal artery of penis (paired)
30 **Tunica albuginea of corpora cavernosa**
31 Septum of penis
32 Deep artery of penis
33 Tunica albuginea of corpus spongiosum
34 Deep fascia of penis

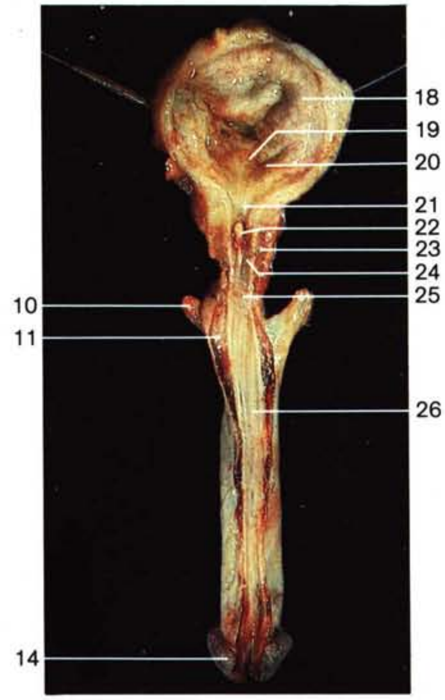

Urinary bladder, urethra and penis (anterior view, open longitudinally).

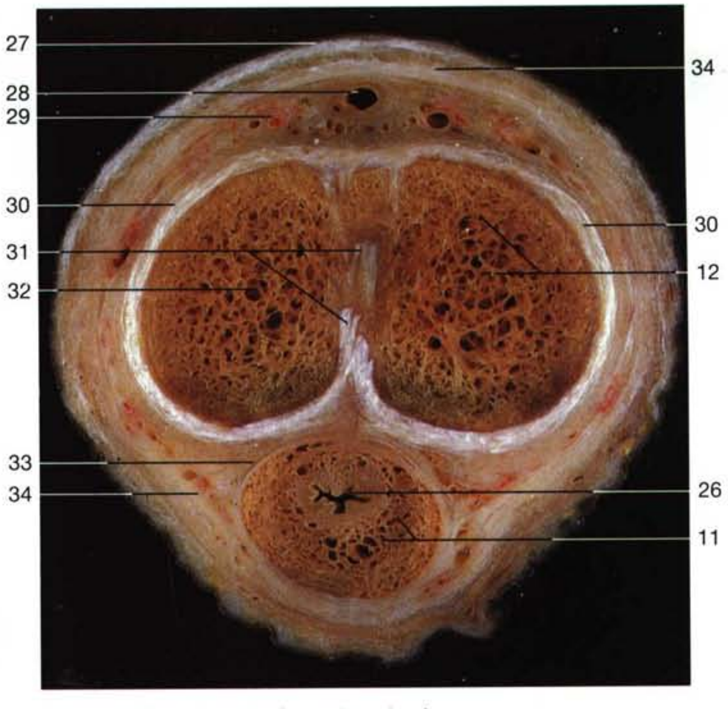

Cross-section of penis (anterior view).

1 **Corpus cavernosum of penis**
2 **Corpus spongiosum of penis**
3 Corona of glans penis
4 **Glans penis**
5 Suspensory ligament of penis
6 Inferior pubic ramus
7 Crus penis
8 **Bulb of penis**
9 Deep dorsal vein of penis
10 Septum pectiniforme
11 Dorsal artery of penis
12 Bulbourethral or Cowper's gland
13 Urinary bladder
14 **Seminal vesicle**
15 Ampulla of ductus deferens
16 Ductus deferens
17 Membranous urethra
18 **Prostate**
19 Ureter

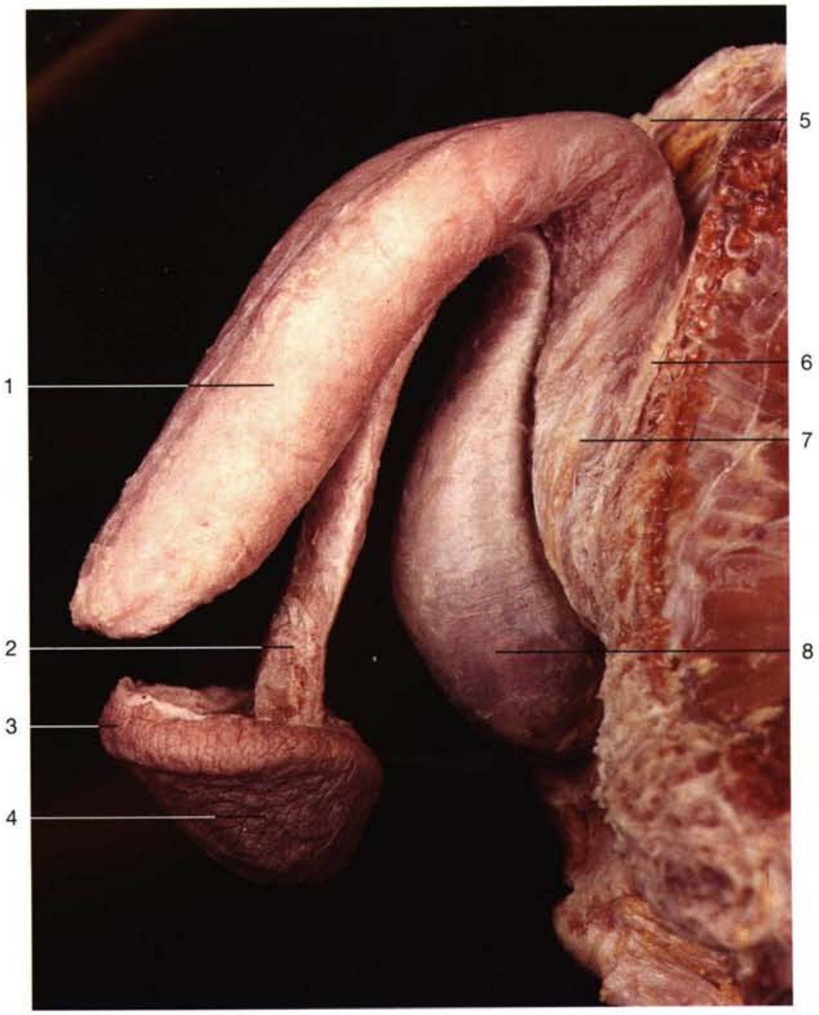

Male external genital organs (lateral view). The corpus spongiosum of the penis with the glans penis has been isolated and reflected.

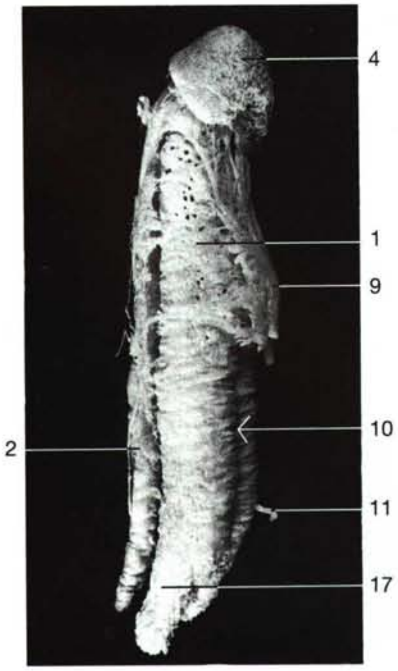

Resin cast of erected penis.

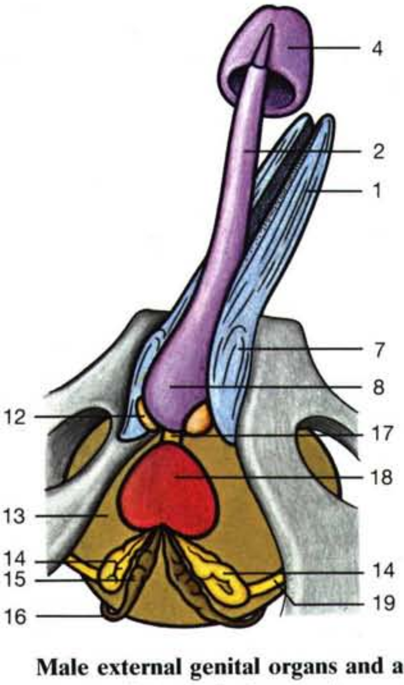

Male external genital organs and accessory glands. (Schematic drawing.)

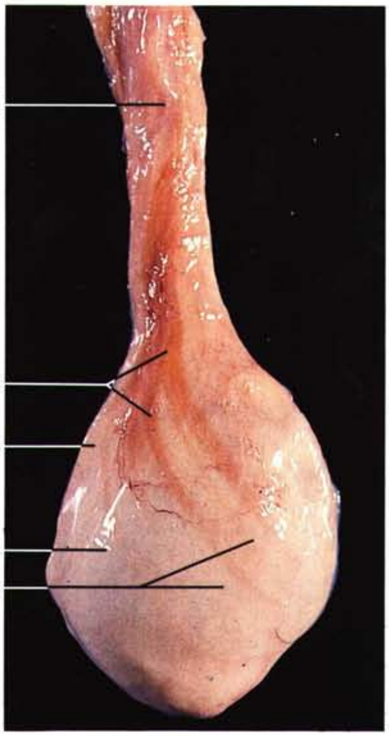

Testis and epididymis with investing layers (lateral view).

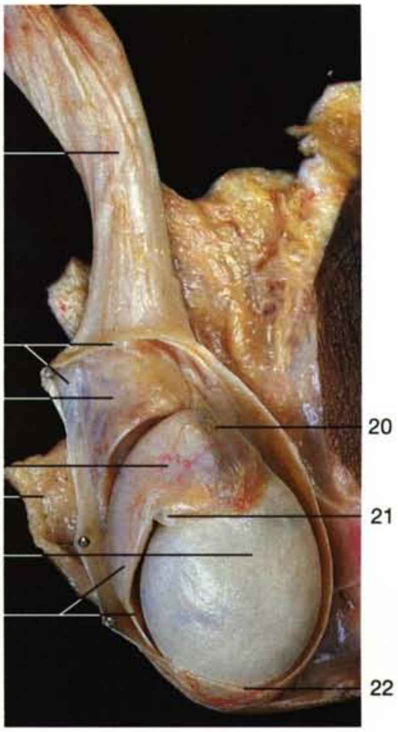

Testis and epididymis (lateral view). The tunica vaginalis has been opened.

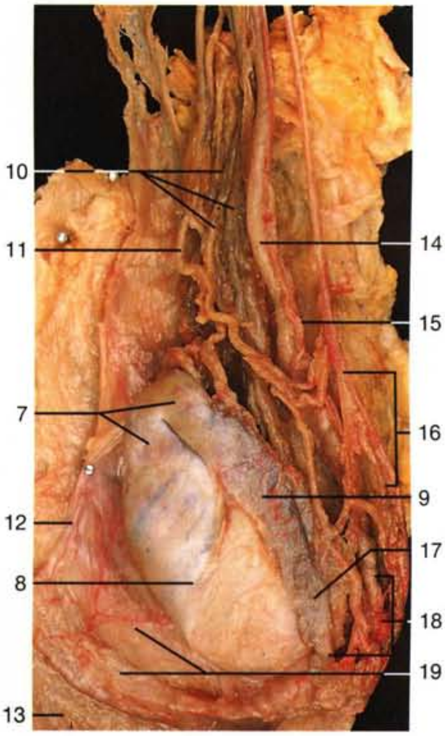

Testis, epididymis and spermatic cord. Dissection of spermatic cord and ductus deferens (left side, posterolateral aspect).

1 Spermatic cord covered with cremasteric fascia
2 **Cremaster muscle**
3 Position of epididymis
4 Internal spermatic fascia
5 Position of testis
6 Internal spermatic fascia with adjacent investing layers of testis (cut surface)
7 **Head of epididymis**
8 **Testis with tunica vaginalis** (visceral layer)
9 Body of epididymis
10 Pampiniform venous plexus (anterior veins)
11 **Testicular artery**
12 Tunica vaginalis (parietal layer, cut edge)
13 Skin and dartos muscle (reflected)
14 **Ductus deferens**
15 Artery of ductus deferens
16 Posterior veins of pampiniform plexus
17 Tail of epididymis
18 Transition of epididymal duct to ductus deferens and venous plexus
19 Parietal layer of tunica vaginalis
20 Appendix of epididymis
21 Appendix of testis
22 Gubernaculum testis

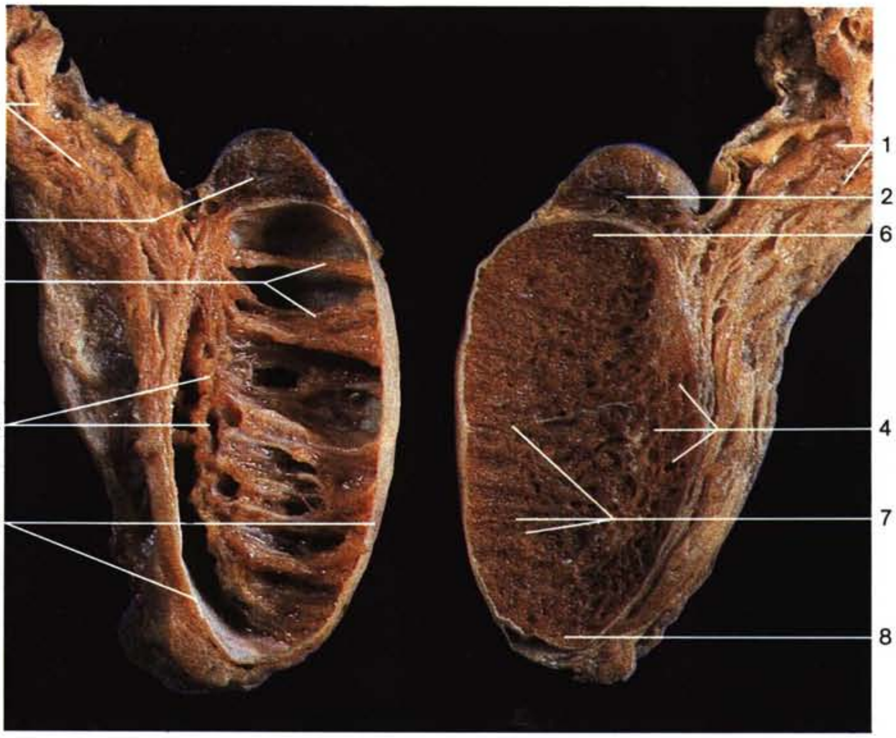

Longitudinal section through testis and epididymis. The left figure shows the testicular septa after removal of the seminiferous tubules.

1 Spermatic cord (cut surface)
2 Head of epididymis (cut surface)
3 Septa of testis
4 Mediastinum testis
5 Tunica albuginea
6 Superior pole of testis
7 Convoluted seminiferous tubules
8 Inferior pole of testis

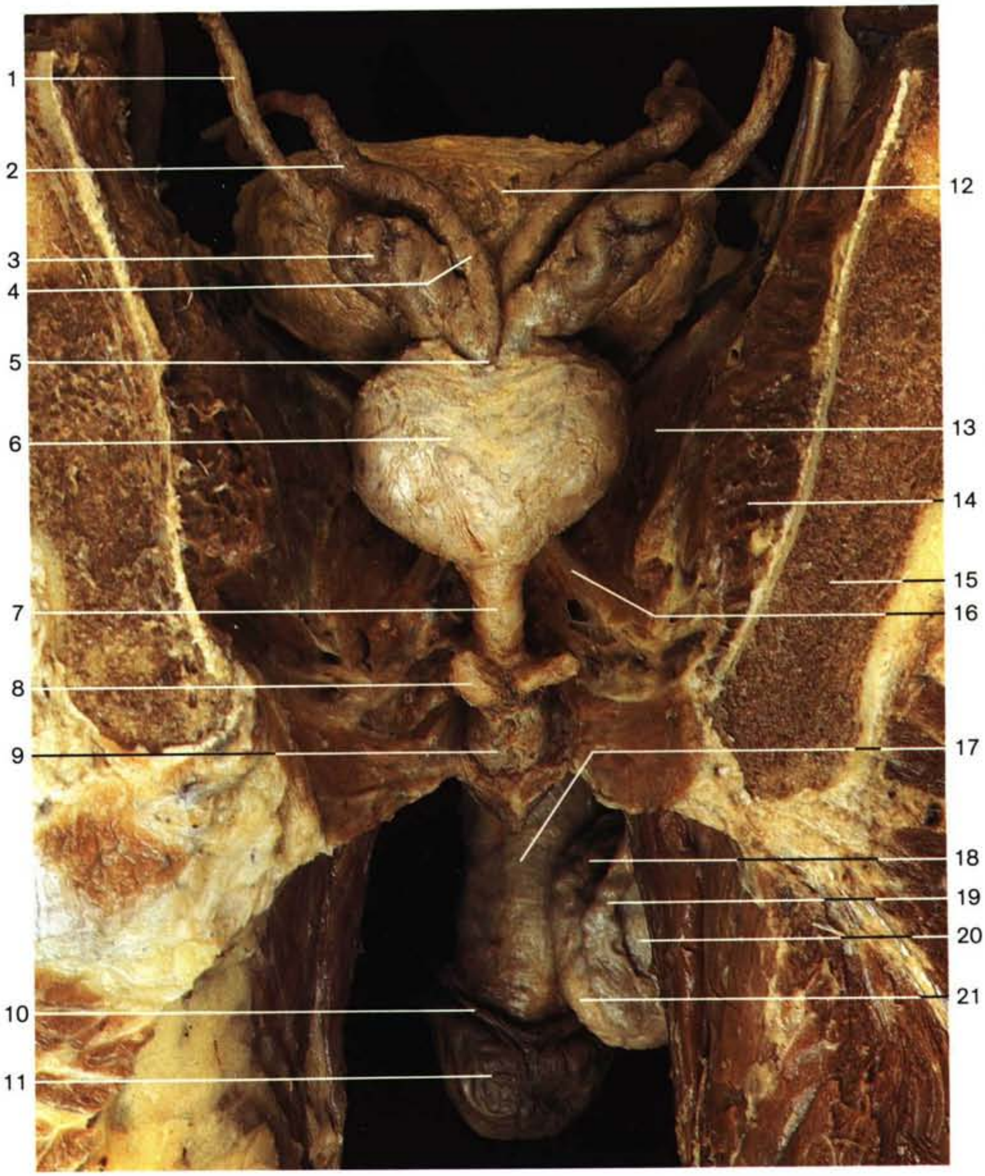

Accessory glands of male genital organs in situ. Coronal section through the pelvic cavity. Posterior aspect of urinary bladder, prostate and seminal vesicles.

1 Ureter
2 **Ductus deferens**
3 **Seminal vesicle**
4 **Ampulla of ductus deferens**
5 Ejaculatory duct (proximal portion)
6 **Prostate**
7 Membranous urethra
8 **Bulbourethral or Cowper's gland**
9 Bulb of penis
10 Penis
11 Glans penis
12 **Urinary bladder**
13 Levator ani muscle
14 Obturator internus muscle
15 Pelvic bone (cut edge)
16 Puboprostatic ligament
17 **Corpus spongiosum of penis**
18 Head of epididymis
19 Beginning of ductus deferens
20 Testis
21 Tail of epididymis
22 Corpus cavernosum of penis
23 Spermatic cord
24 Pectineus and adductor muscles
25 Pubic bone
26 **Prostatic part of urethra** (seminal colliculus)
27 Rectum
28 Sciatic nerve
29 Great saphenous vein
30 Sartorius muscle
31 Femoral artery and vein
32 Rectus femoris muscle
33 Tensor fasciae latae muscle
34 Pectineus muscle
35 Iliopsoas muscle
36 Vastus lateralis muscle
37 Obturator externus muscle
38 Femur
39 Ischial tuberosity
40 Gluteus maximus muscle

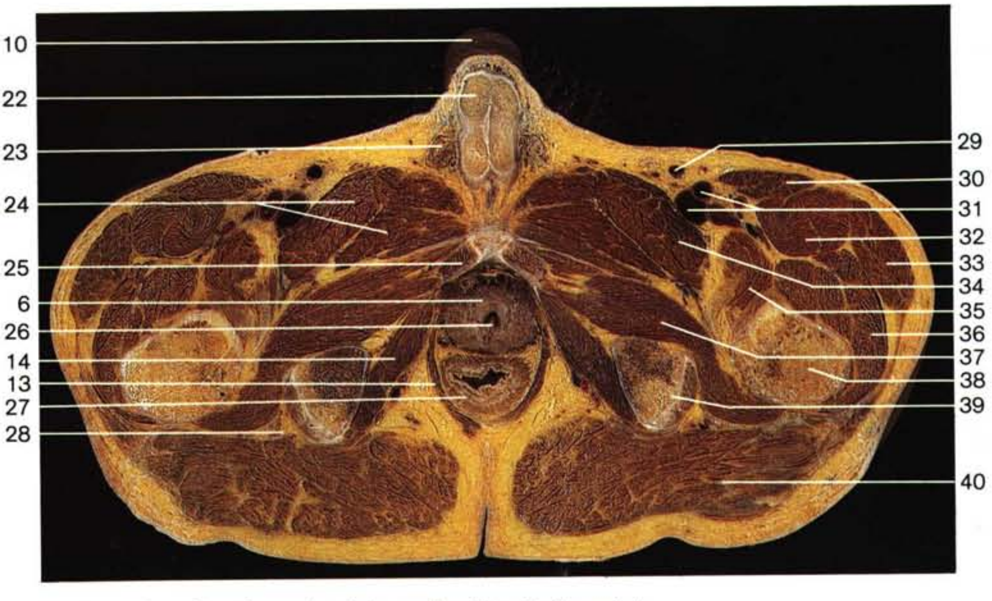

Horizontal section through pelvic cavity at level of prostate.

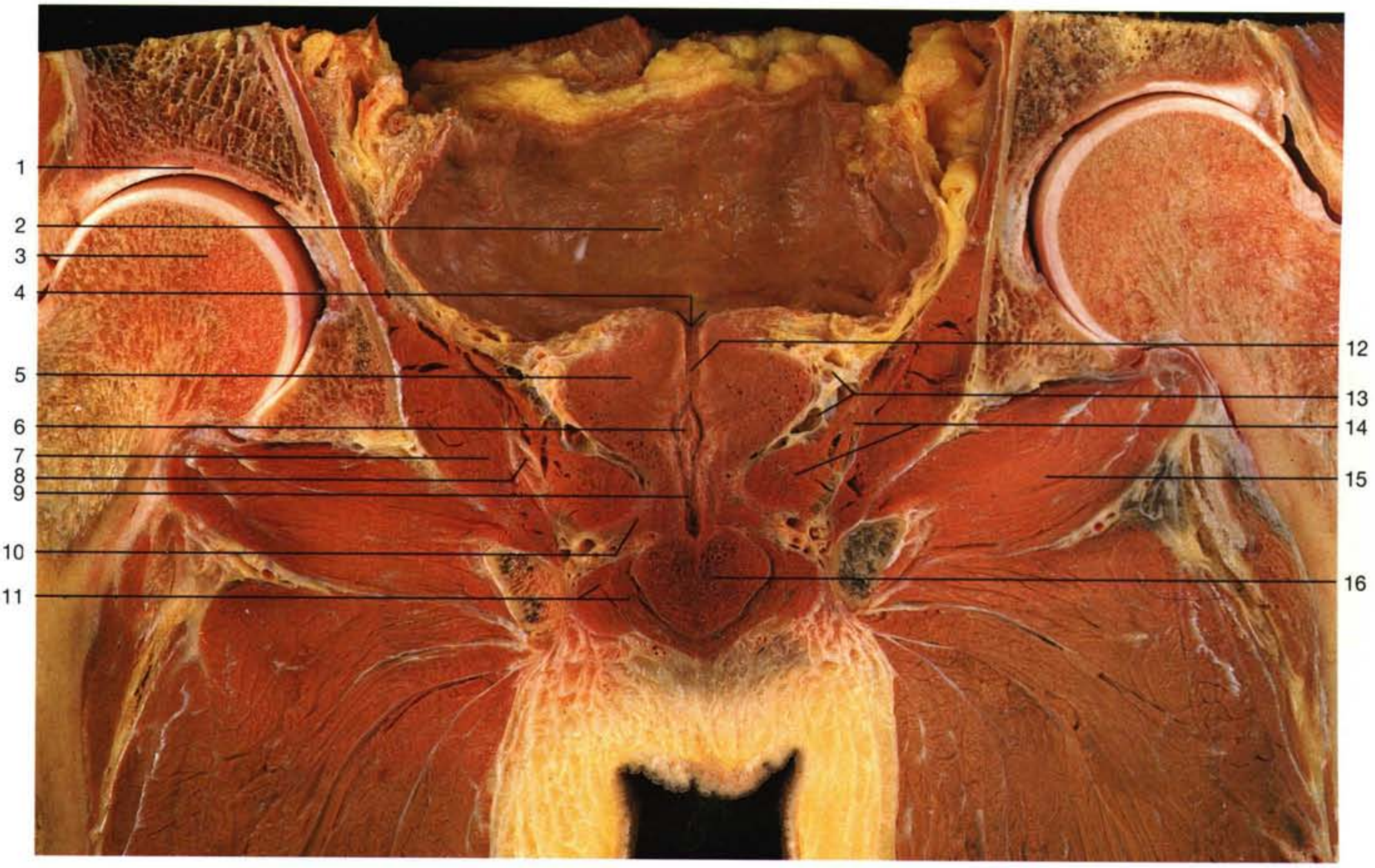

Coronal section through pelvic cavity at the level of prostate and hip joint (anterior aspect).

1 Acetabulum of hip joint
2 **Urinary bladder**
3 Head of femur
4 Internal urethral orifice
5 **Prostate**
6 **Seminal colliculus**
7 Obturator internus muscle
8 Ischiorectal fossa
9 Membranous urethra
10 Deep transverse perineus muscle
11 **Crus penis** and **ischiocavernosus muscle**
12 Prostatic part of urethra
13 Prostatic plexus
14 Levator ani muscle
15 Obturator externus muscle
16 **Bulb of penis**
17 Ampulla of rectum
18 **Anal canal**
19 Seminal vesicle
20 Internal anal sphincter muscle
21 External anal sphincter muscle
22 Anus
23 Psoas major muscle
24 Intervertebral disc
25 Ilium
26 Ligament of the head of the femur
27 Sacral promontory

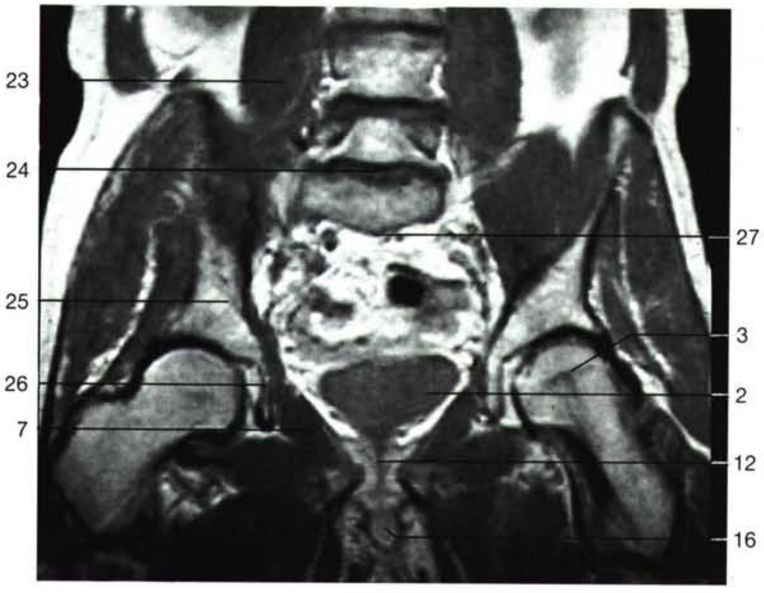

Coronal section through pelvic cavity. (MR-Scan.)

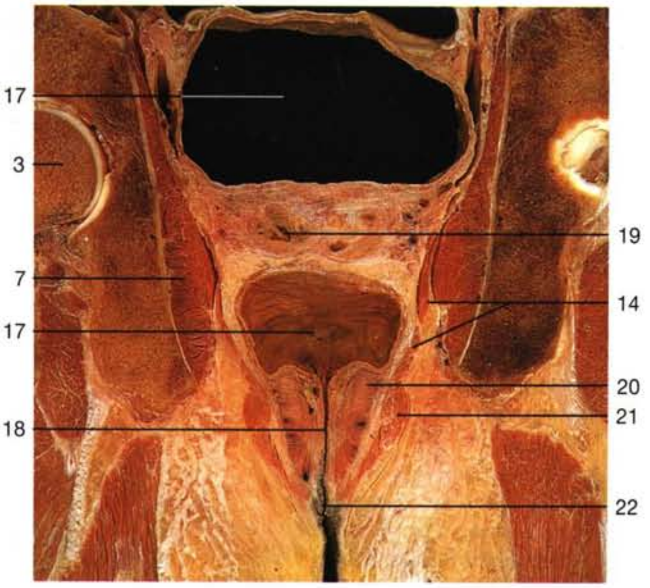

Coronal section through anal canal.

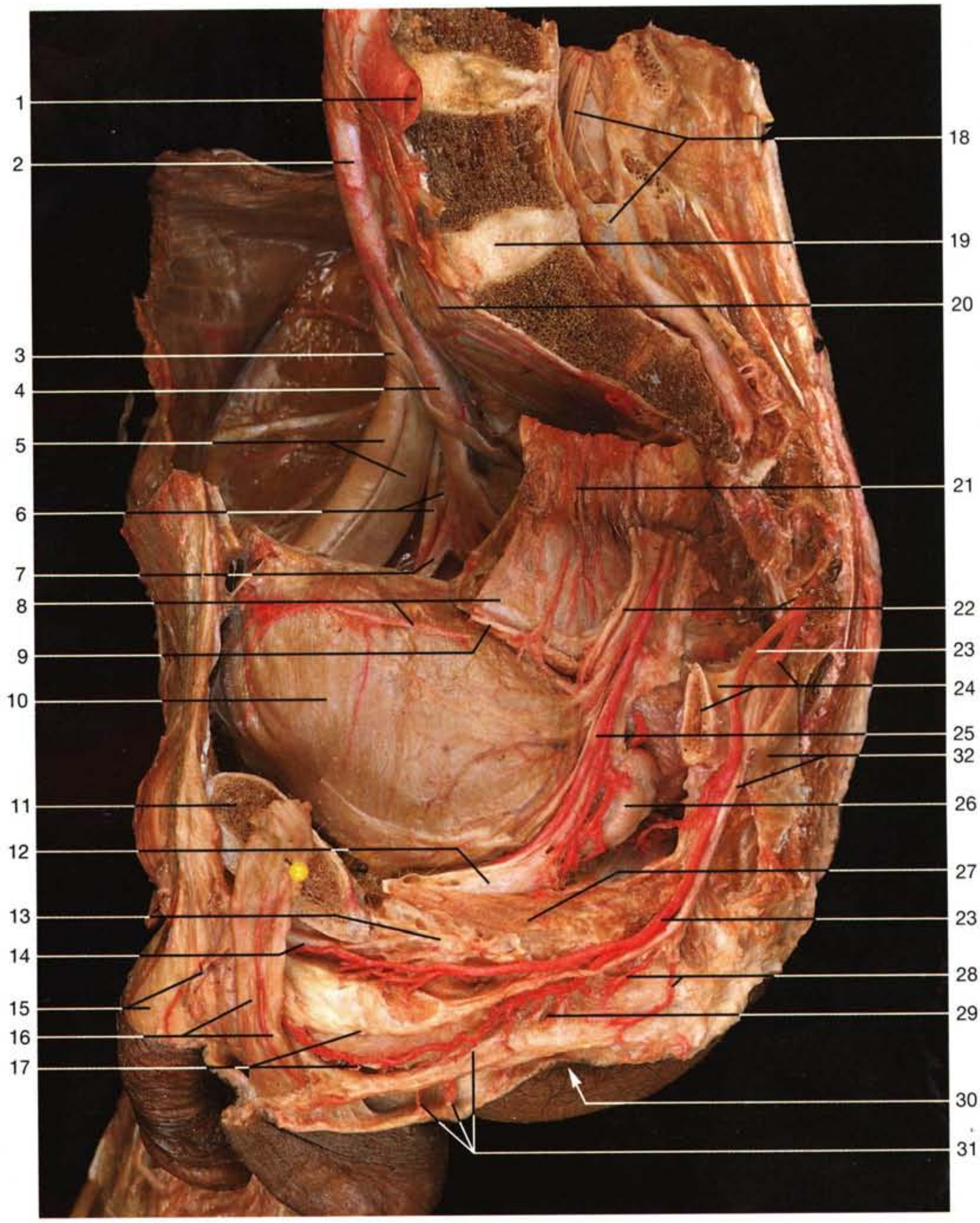

Pelvic cavity in the male (right half of parasagittal section). The arteries have been injected with red resin. The parietal layer of peritoneum has been removed. The urinary bladder is filled to a great extent.

1 Left common iliac artery
2 Right common iliac artery
3 Right **ureter**
4 Right **internal iliac artery**
5 Right external iliac artery and vein
6 Right **obturator artery and nerve**
7 Umbilical artery
8 Sigmoid and superior vesical artery
9 Left ductus deferens
10 **Urinary bladder**
11 Pubic bone (cut)
12 **Prostate**
13 Vesicoprostatic venous plexus
14 Deep dorsal vein of penis and dorsal artery of penis
15 Penis and superficial dorsal vein
16 **Spermatic cord and testicular artery**
17 **Bulb of penis** and deep artery of penis
18 Cauda equina and dura mater (divided)
19 Intervertebral disc between fifth lumbar vertebra and sacrum
20 **Sacral promontory**
21 Mesosigmoid
22 Left ureter
23 Left internal pudendal artery
24 Ischial spine (cut), sacrospinal ligament, inferior gluteal artery
25 Left **inferior vesical artery**
26 **Seminal vesicle**
27 Levator ani muscle
28 Branches of inferior rectal artery
29 Perineal artery
30 Anus
31 Posterior scrotal branches
32 **Pudendal nerve** and sacrotuberal ligament

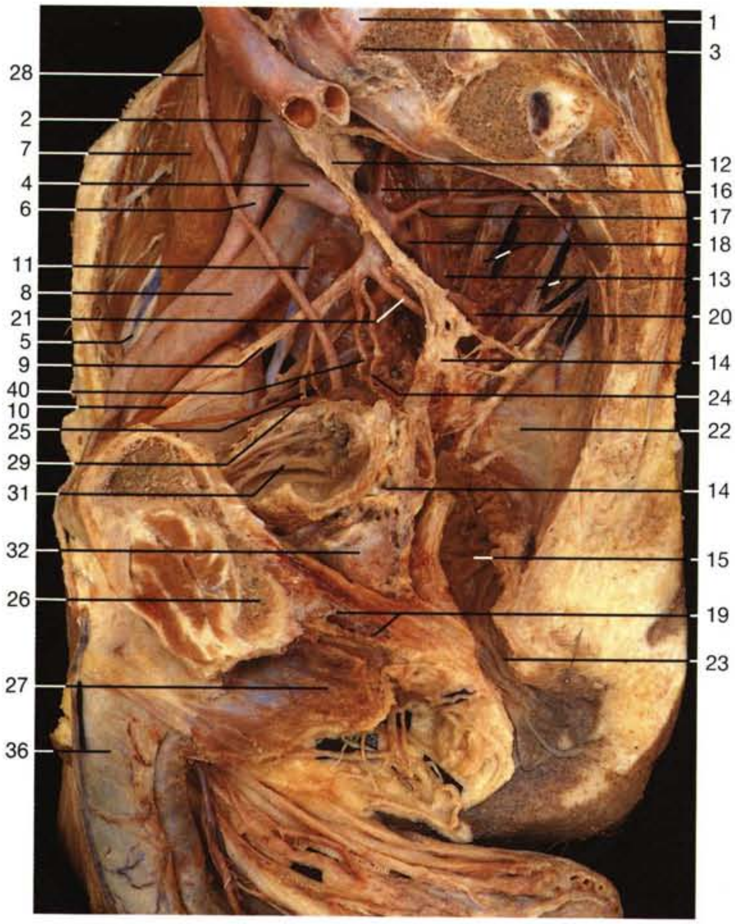

Vessels of pelvic cavity in the male (midsagittal section, right side, medial aspect). Urinary bladder and rectum have been partly removed; the hypogastric plexus is displayed.

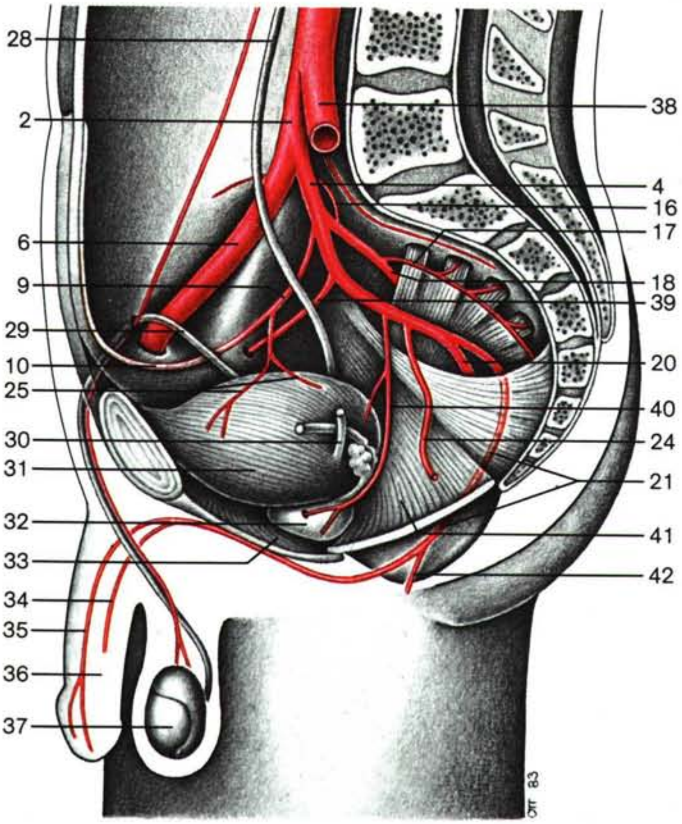

Main branches of internal iliac artery in the male. (Schematic drawing.)

1 Intervertebral disc between 5th lumbar vertebra and sacrum
2 Right common iliac artery
3 Sacral promontory
4 Right **internal iliac artery**
5 Femoral nerve
6 Right external iliac artery
7 Psoas major muscle
8 Right external iliac vein
9 **Umbilical artery**
10 Medial umbilical ligament with obliterated umbilical artery
11 Obturator nerve
12 Superior hypogastric plexus
13 Sacral spinal nerve (of sacral plexus)
14 **Inferior hypogastric plexus**
15 Rectum
16 Iliolumbar artery
17 Lateral sacral artery
18 Superior gluteal artery and sacral spinal nerve
19 Pudendal and coccygeal plexus
20 Inferior gluteal artery
21 **Internal pudendal artery**
22 Coccygeus muscle overlying sacrospinous ligament
23 **Anal canal**
24 Middle rectal artery
25 **Superior vesical artery** and branch to the ductus deferens
26 Pubic bone (cut)
27 Bulbospongiosus muscle
28 Right ureter
29 Right ductus deferens
30 Left ureter
31 **Urinary bladder**
32 **Prostate**
33 Urogenital diaphragm
34 Deep artery of penis
35 Dorsal artery of penis
36 Penis
37 Testis
38 Left common iliac artery
39 **Obturator artery**
40 **Inferior vesical artery**
41 Levator ani muscle
42 Inferior rectal artery

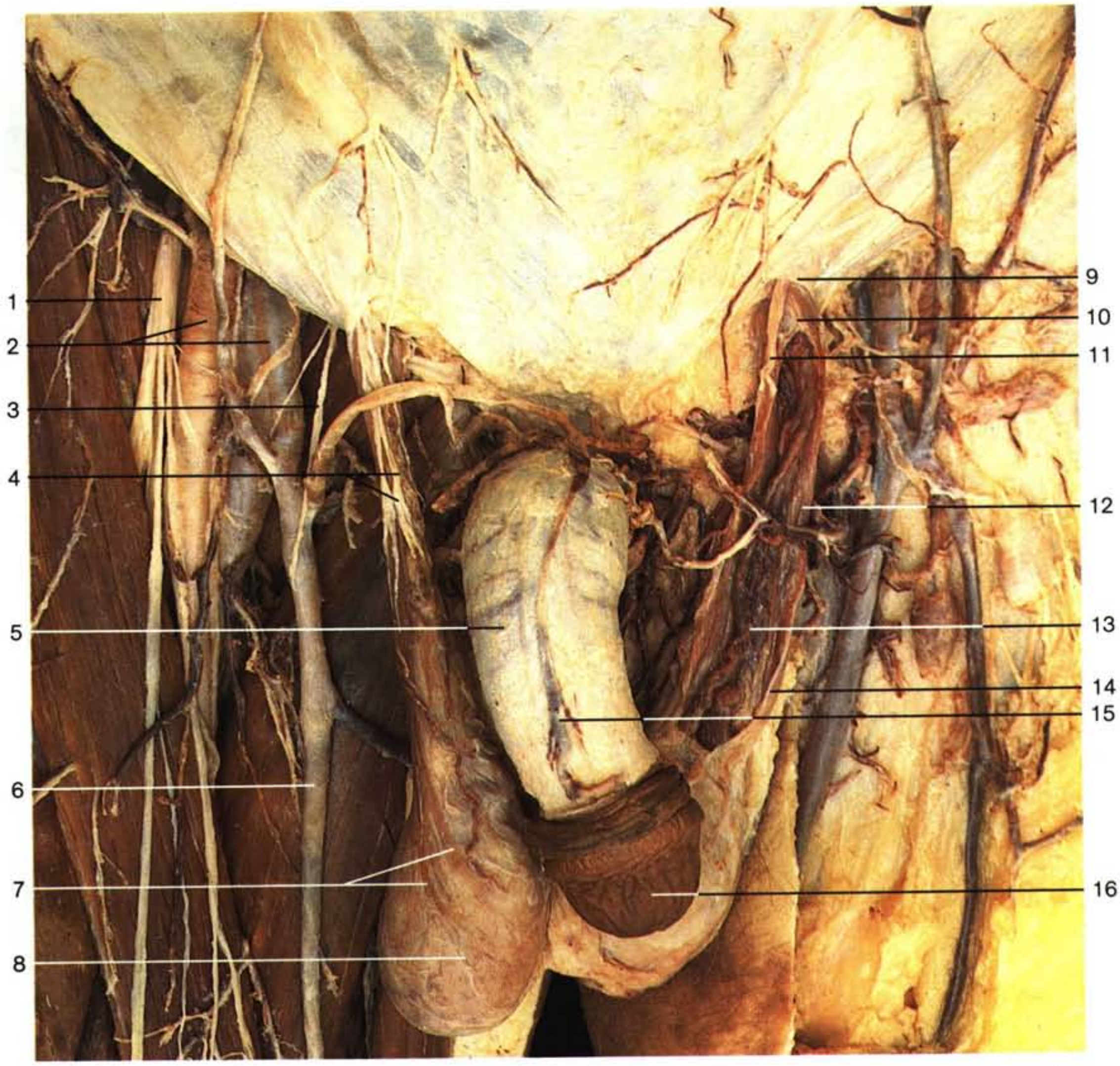
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Male external genital organs with penis, testis, and spermatic cord, superficial layers (anterior view).

1 Femoral nerve
2 Femoral artery and vein
3 Femoral branch of genitofemoral nerve
4 **Spermatic cord** with genital branch of genitofemoral nerve
5 **Penis** with deep fascia
6 Great saphenous vein
7 Cremaster muscle
8 **Testis with cremaster muscle**
9 Superficial inguinal ring
10 Internal spermatic fascia (cut edge)
11 Ilioinguinal nerve
12 Left spermatic cord
13 Pampiniform venous plexus
14 External spermatic fascia
15 Superficial dorsal vein of penis
16 **Glans penis**

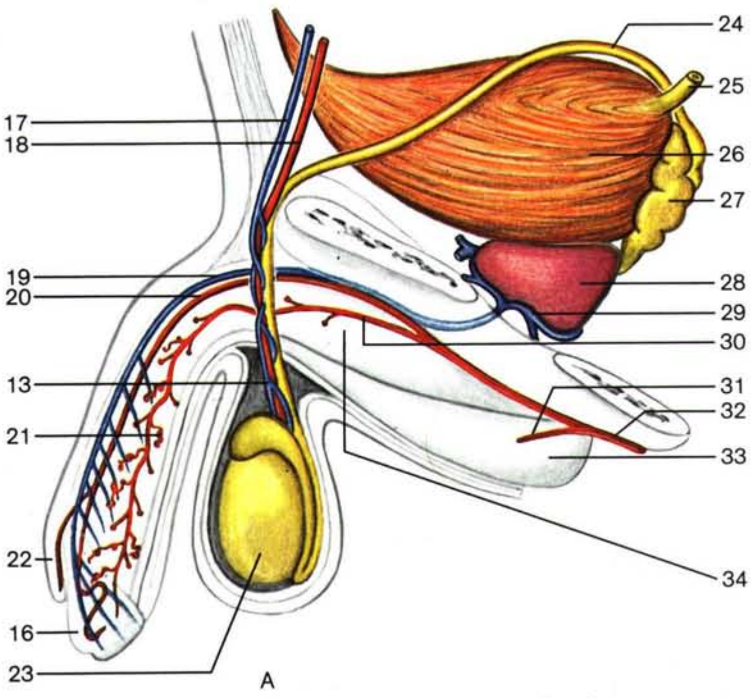
17
18
19
20
13
21
22
16
23
24
25
26
27
28
29
30
31
32
33
34
A

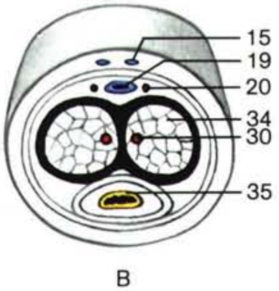
15
19
20
34
30
35
B

Vessels of male genital organs. (Schematic drawing.)
A = lateral aspect; B = cross-section of penis.

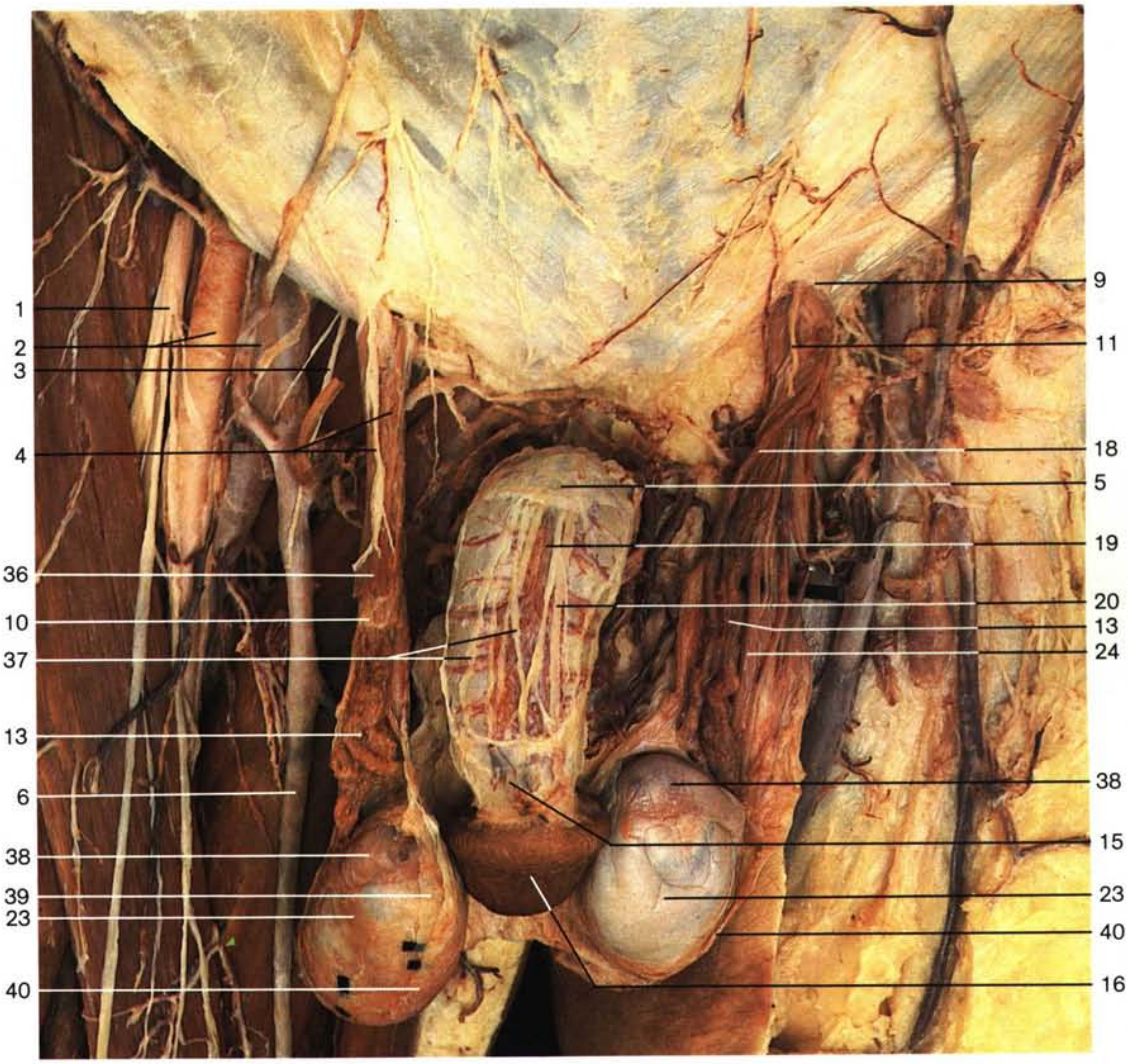

Male external genital organs with penis, testis and spermatic cord, deeper layers (ventral aspect). The deep fascia of the penis has been opened to display the dorsal nerves and vessels.

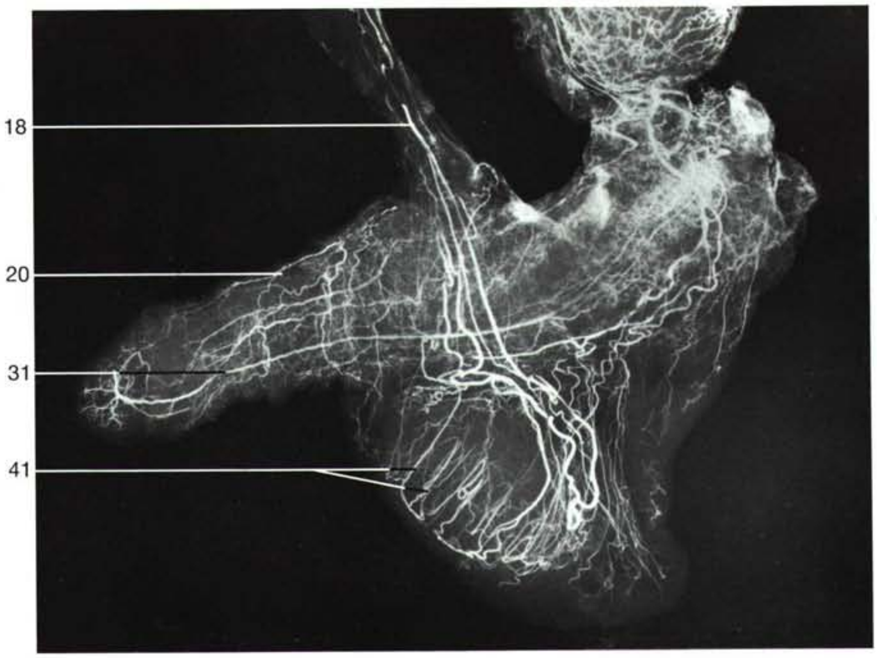

Male genital organs (lateral aspect). (Arteriogram.)

17 Testicular vein
18 Testicular artery
19 Deep dorsal vein of penis
20 **Dorsal artery of penis**
21 **Helicine arteries**
22 Prepuce
23 **Testis** with **tunica albuginea**
24 Ductus deferens
25 Ureter
26 Urinary bladder
27 **Seminal vesicle**
28 **Prostate**
29 Vesicoprostatic venous plexus
30 **Deep artery of penis**
31 Artery of bulb of penis
32 Internal pudendal artery
33 Corpus spongiosum of penis
34 Corpus cavernosum of penis
35 Urethra
36 Cremasteric fascia with cremaster muscle
37 Dorsal nerve of penis
38 **Epididymis**
39 **Tunica vaginalis** (visceral layer)
40 Tunica vaginalis (parietal layer)
41 Testis with vascular loops

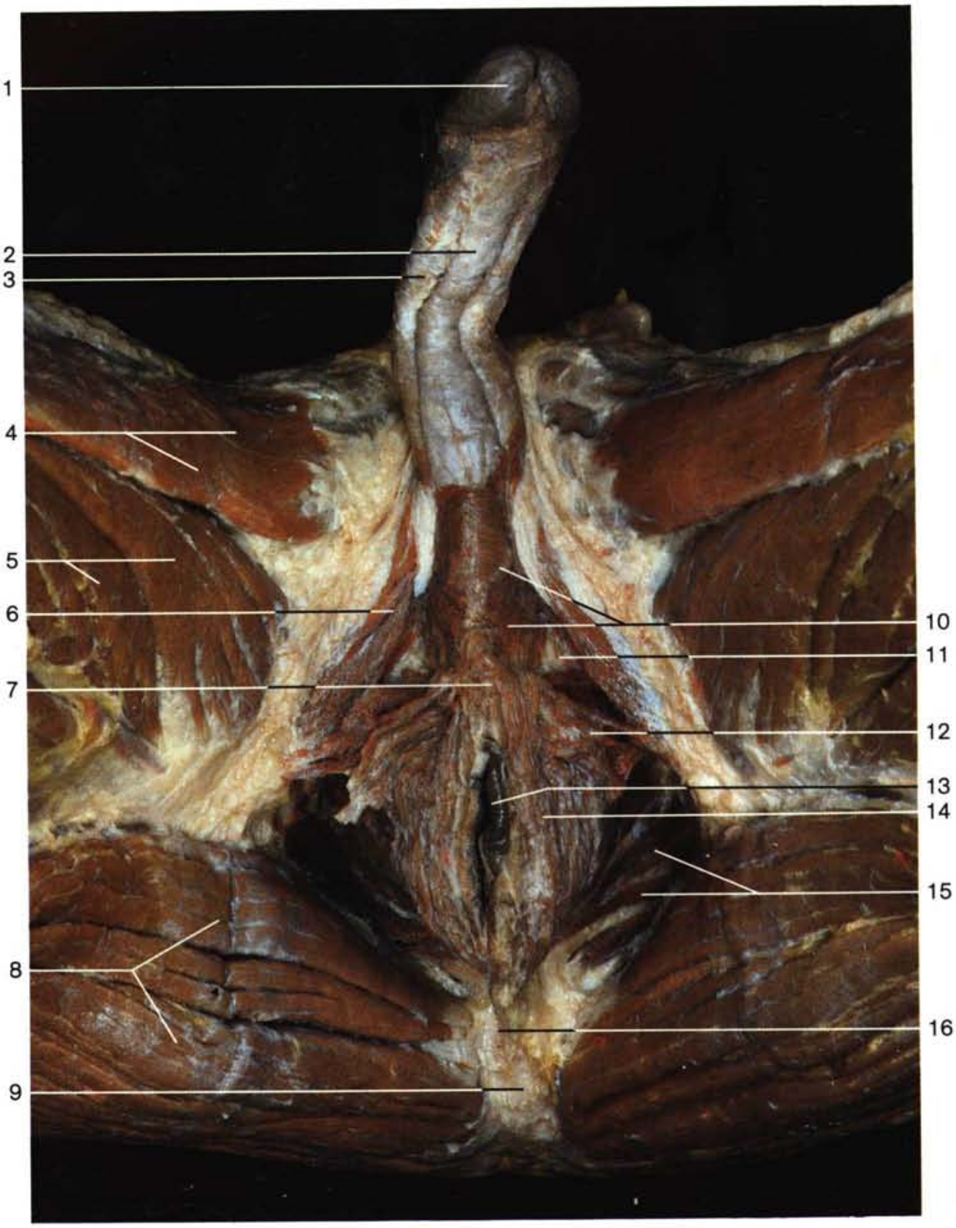

Muscles of urogenital and pelvic diaphragms in the male (from below).

1 Glans penis
2 **Corpus spongiosum of penis**
3 **Corpus cavernosum of penis**
4 Gracilis muscle
5 Adductor muscles
6 Ischiocavernosus muscle overlying crus of penis
7 Perineal body
8 Gluteus maximus muscle
9 Coccyx
10 **Bulbospongiosus** muscle
11 Deep transverse perineus muscle covered by inferior fascia of urogenital diaphragm
12 Superficial transverse perineus muscle
13 Anus
14 **External anal sphincter muscle**
15 **Levator ani muscle**
16 Anococcygeal ligament
17 Obturator internus muscle
18 Urethra
19 **Deep transverse perineus muscle**

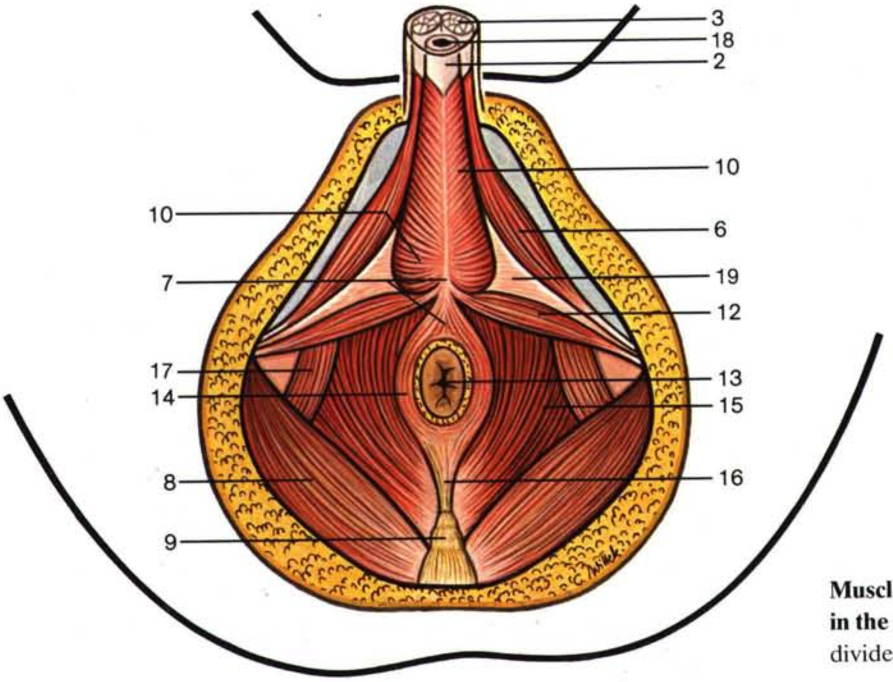

Muscles of urogenital and pelvic diaphragms in the male (from below). The penis has been divided. (Schematic drawing.)

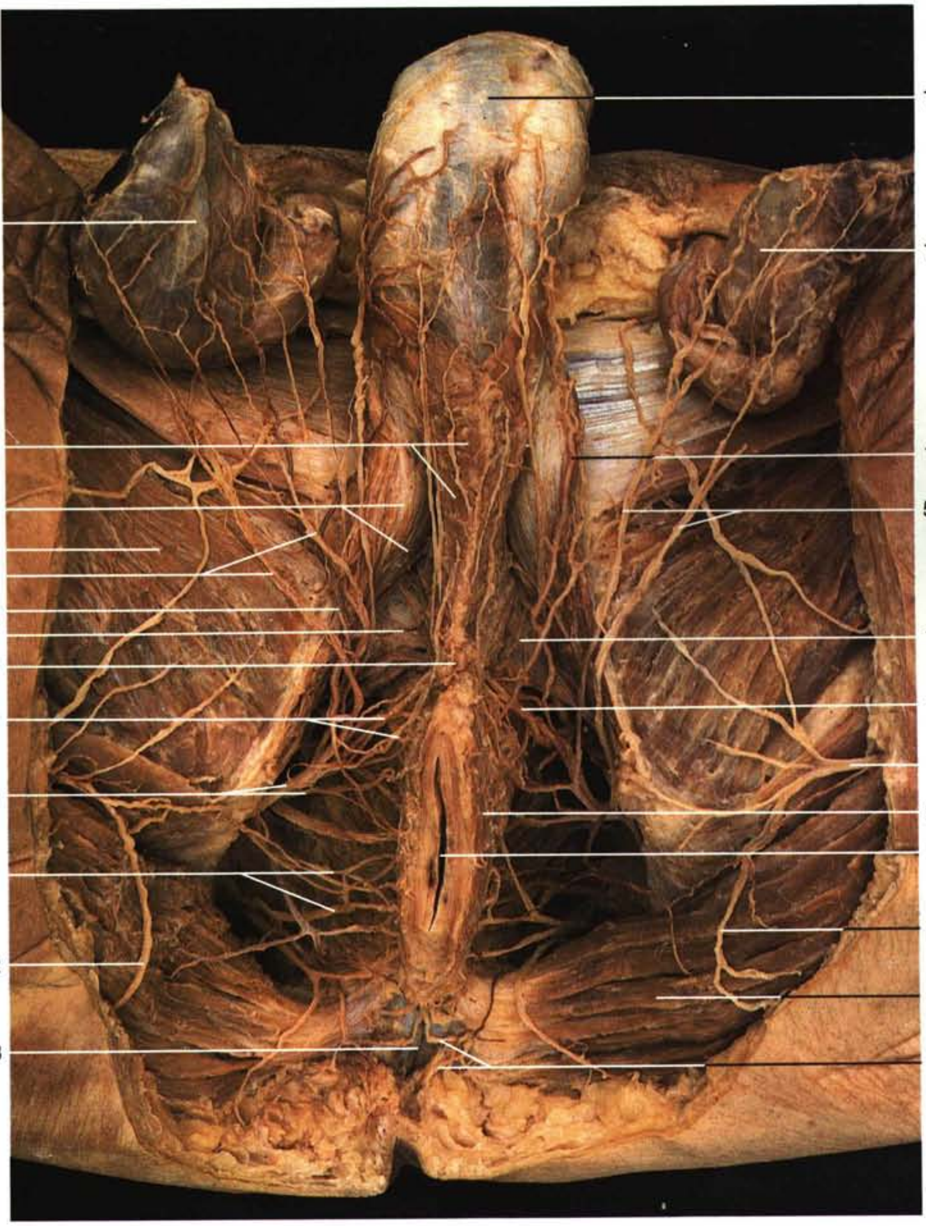

Urogenital diaphragm and external genital organs in the male with vessels and nerves (from below). The testes have been reflected laterally.

1 Right testis (reflected laterally and upward)
2 Bulbospongiosus muscle
3 Ischiocavernosus muscle
4 Adductor magnus muscle
5 **Posterior scrotal nerves** and superficial perineal arteries
6 Posterior scrotal artery and vein
7 Right artery of bulb of penis
8 Perineal body
9 Perineal branches of pudendal nerve
10 **Pudendal nerve** and **internal pudendal artery**
11 **Inferior rectal arteries** and **nerves**
12 Inferior cluneal nerve
13 Coccyx (location)
14 Penis
15 Left testis (reflected laterally)
16 Left posterior scrotal artery
17 **Deep transverse perineal muscle**
18 Left artery of bulb of penis
19 **Posterior femoral cutaneous nerve**
20 External anal sphincter muscle
21 **Anus**
22 Gluteus maximus muscle
23 Anococcygeal nerves
24 Acetabulum (femur removed)
25 Ligament of femoral head
26 Body of ischium (cut)
27 Sciatic nerve
28 Coccygeus muscle
29 **Levator ani muscle**
  a iliococcygeus muscle
  b pubococcygeus muscle
  c puborectalis muscle
30 Prostatic venous plexus
31 Body of pubis
32 Testis

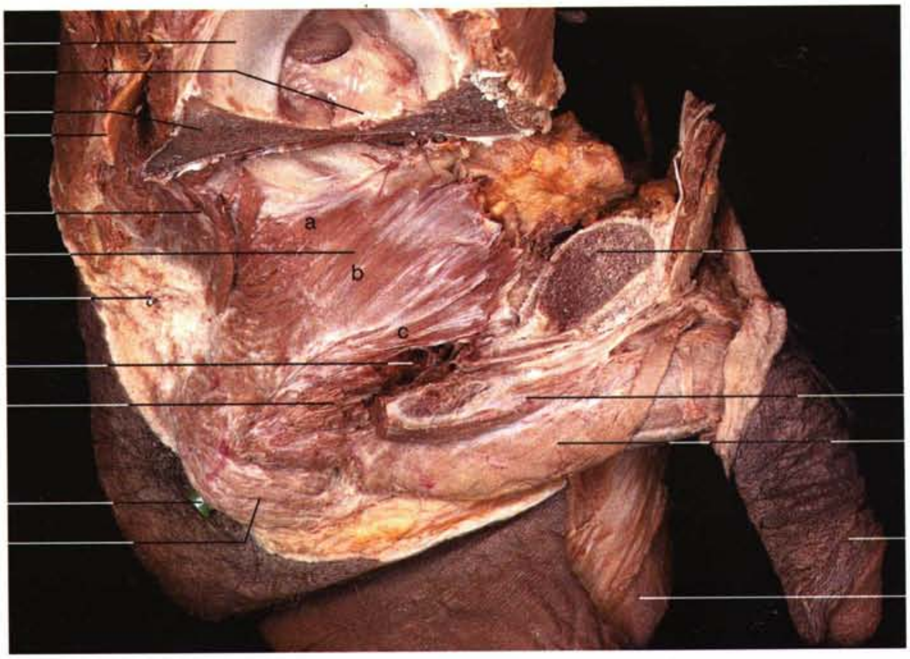

Pelvic diaphragm and external genital organs in the male. The right half of the pelvis including obturator internus muscle and femur had been removed to display the right half of levator ani muscle.

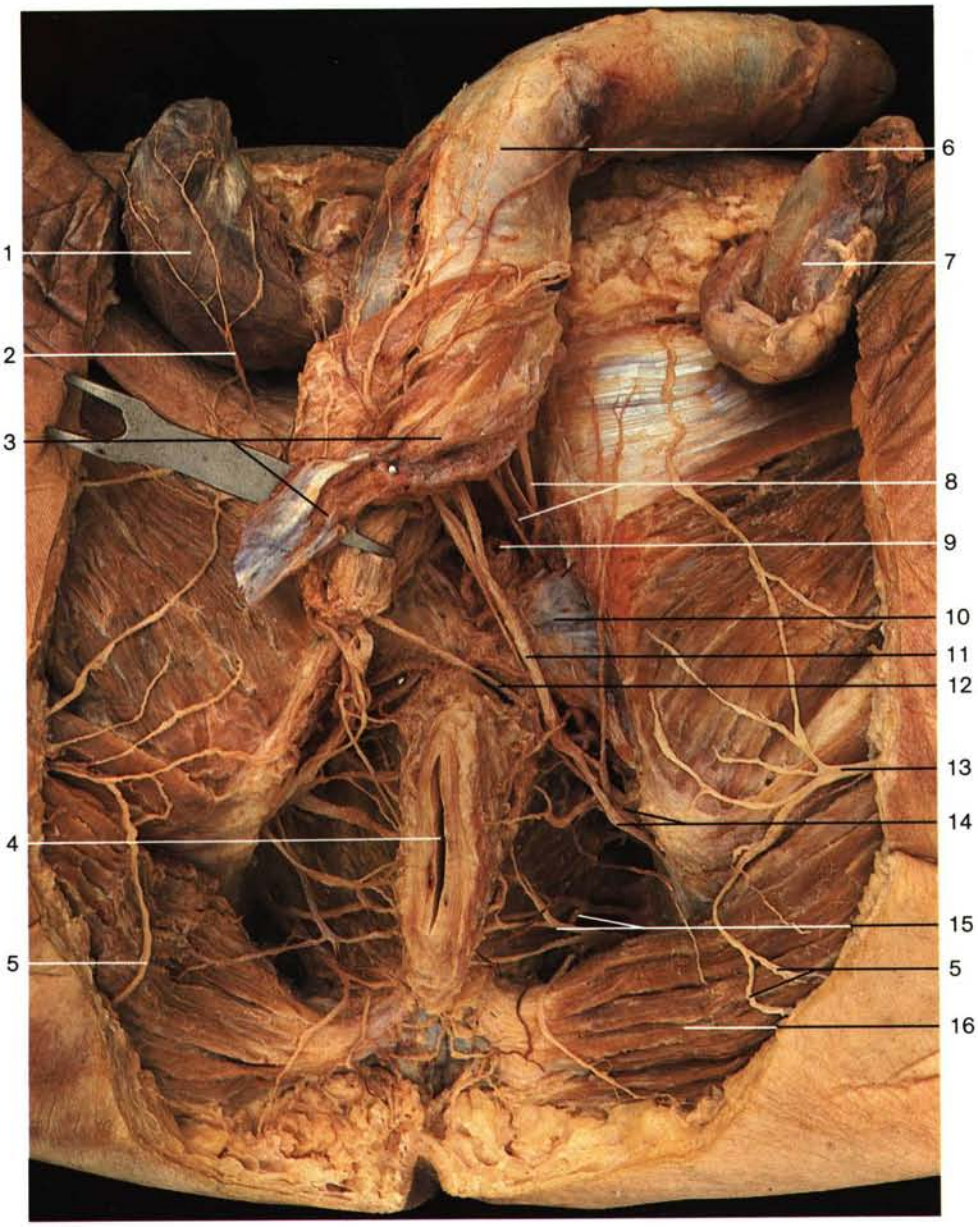

1 Right testis (reflected)
2 Posterior scrotal nerves
3 Left crus penis with ischiocavernosus muscle
4 **Anus**
5 Inferior cluneal nerves
6 **Penis**
7 Left testis (reflected)
8 Dorsal artery and nerve of penis
9 Urethra
10 **Deep transverse perineus muscle**
11 **Perineal branch of pudendal nerve**
12 Artery of bulb of penis (reflected)
13 Branch of posterior femoral cutaneous nerve
14 **Internal pudendal artery** and **pudendal nerve**
15 Inferior rectal arteries and nerves
16 Gluteus maximus muscle
17 Dorsal nerve of penis
18 Posterior femoral cutaneous nerve
19 Perineal branches of pudendal nerve
20 Inferior rectal nerves
21 **Bulbocavernosus muscle** (inside: dorsal artery of penis)
22 Perineal artery
23 **External anal sphincter muscle**
24 Inferior rectal artery and veins

Urogenital diaphragm and external genital organs in the male (from below). The left crus penis has been isolated and reflected laterally together with the bulb of the penis. The urethra has been cut.

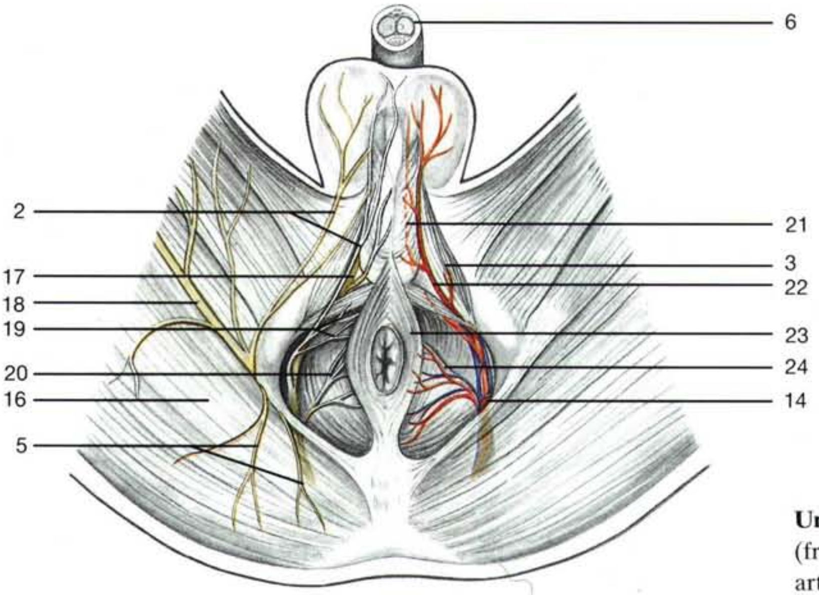

Urogenital and anal region in the male (from below). Right side: nerves; left side: arteries and veins.

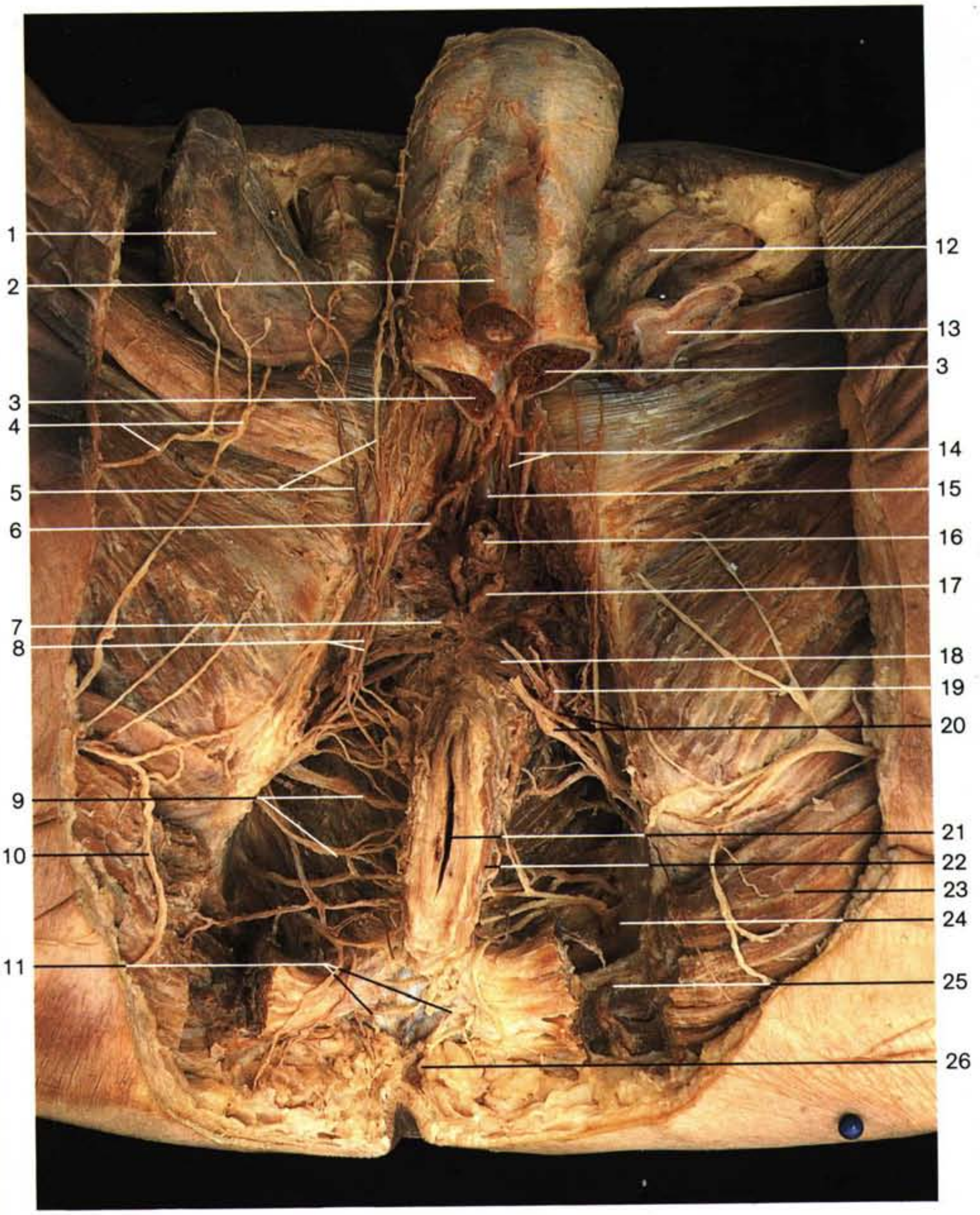

1 Right testis (reflected)
2 **Corpus spongiosum of penis**
3 **Corpus cavernosum of penis**
4 Perineal branch of posterior femoral cutaneous nerve
5 Posterior scrotal arteries and nerves
6 Deep artery of penis
7 Deep transverse perineal muscle
8 Right **perineal nerves**
9 **Inferior rectal nerves**
10 Inferior cluneal nerve
11 Anococcygeal nerves
12 Left spermatic cord
13 Left testis (cut surface)
14 Dorsal artery and nerve of penis
15 Deep dorsal vein of penis
16 **Urethra** (cut)
17 Artery of bulb of penis
18 Superficial transverse perineus muscle
19 Left artery of bulb of penis
20 Perineal branch of pudendal nerve
21 **Anus**
22 **External anal sphincter muscle**
23 Gluteus maximus muscle
24 Internal pudendal artery and pudendal nerve
25 Sacrotuberous ligament
26 Coccyx
27 Urogenital diaphragm (deep transverse perineus muscle)
28 Tendinous center of perineum (perineal body)
29 **Levator ani muscle**
30 Anococcygeal ligament
31 Obturator internus muscle
32 Dorsal artery of penis

Anal and urogenital region in the male (from below). The root of penis has been cut. Dissection of the urogenital diaphragm.

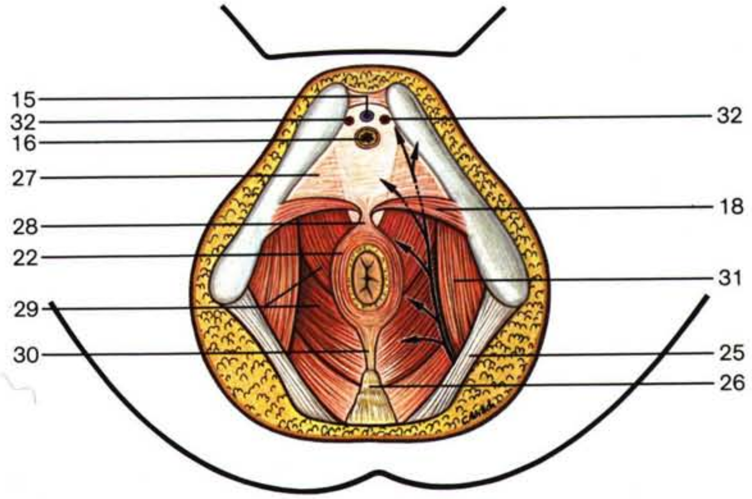

Urogenital and pelvic diaphragms in the male (from below). The penis has been removed. The arrows indicate the course of vessels and nerves. (Schematic drawing.)

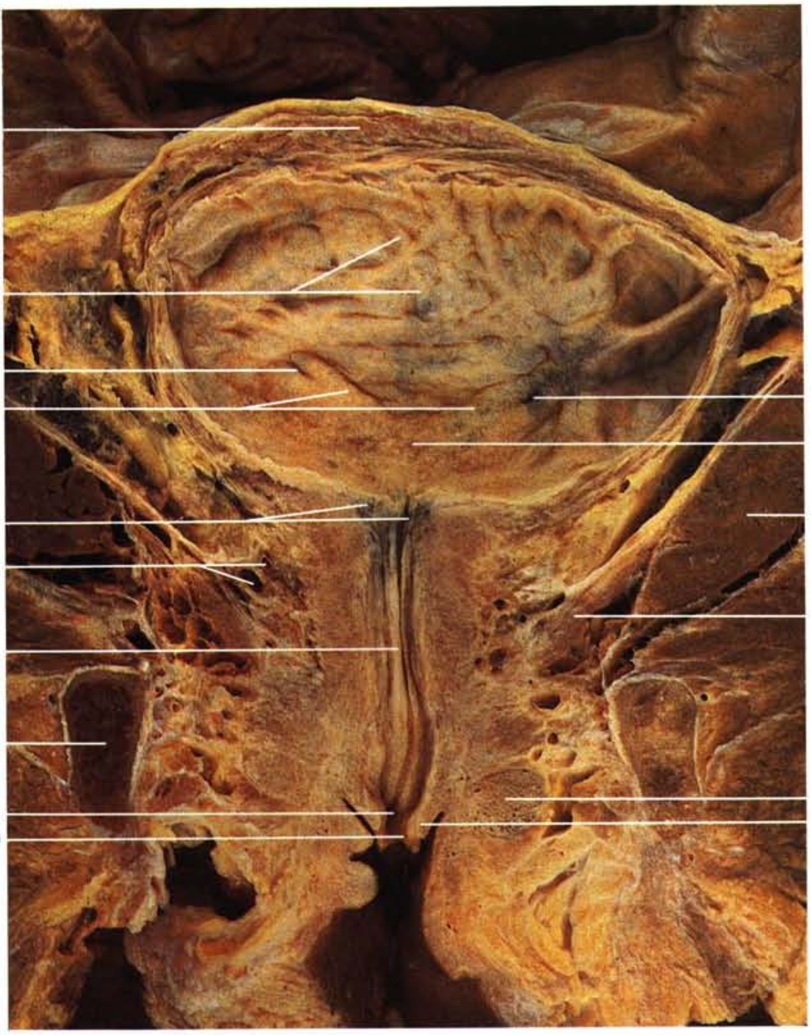

Coronal section through the female urinary bladder and urethra (anterior view).

1 Muscular coat of **urinary bladder**
2 Folds of mucous membrane of urinary bladder
3 Right ureteric orifice
4 Interureteric fold
5 Internal urethral orifice
6 Vesicouterine venous plexus
7 **Urethra**
8 Pubic bone (cut edge)
9 External urethral orifice
10 Vestibule of vagina
11 Left ureteric orifice
12 Trigone of bladder
13 Obturator internus muscle
14 Levator ani muscle
15 Bulb of the vestibule
16 Left labium minus
17 **Uterine tube**
18 Mesosalpinx
19 **Ovary**
20 Sigmoid colon
21 Saphenous opening
22 Urinary bladder
23 Vesicouterine pouch
24 **Fundus of uterus**
25 Rectouterine pouch (of Douglas)
26 Ampulla of rectum
27 Kidney
28 Abdominal part of ureter
29 Pelvic part of ureter
30 Anal canal
31 Perineum (perineal body)
32 Umbilicus
33 Infundibulum of uterine tube
34 Vaginal portion of cervix of uterus
35 Vagina
36 Pubic symphysis
37 Clitoris
38 Deep transverse perineus muscle

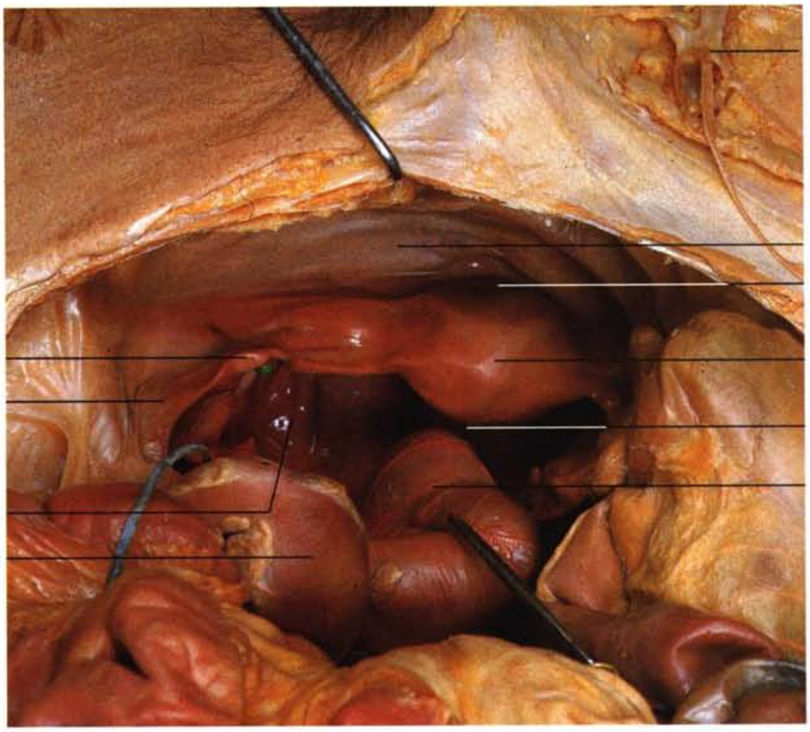

Femal internal genital organs. Pelvic cavity (from above).

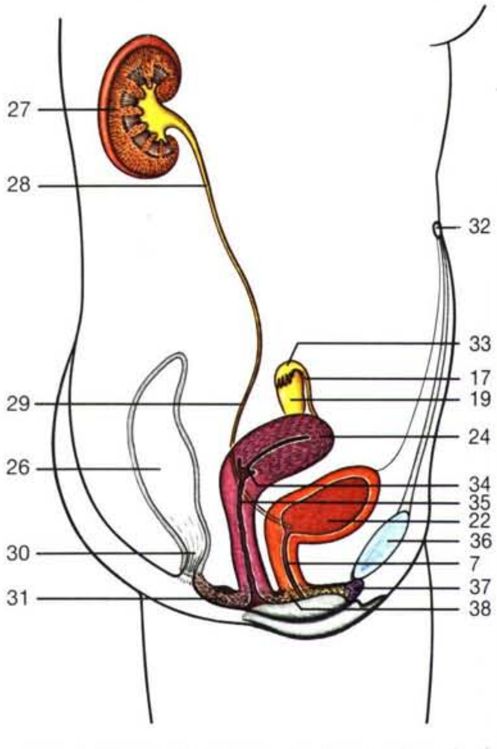

Female urogenital system (midsagittal section). (Schematic drawing.)

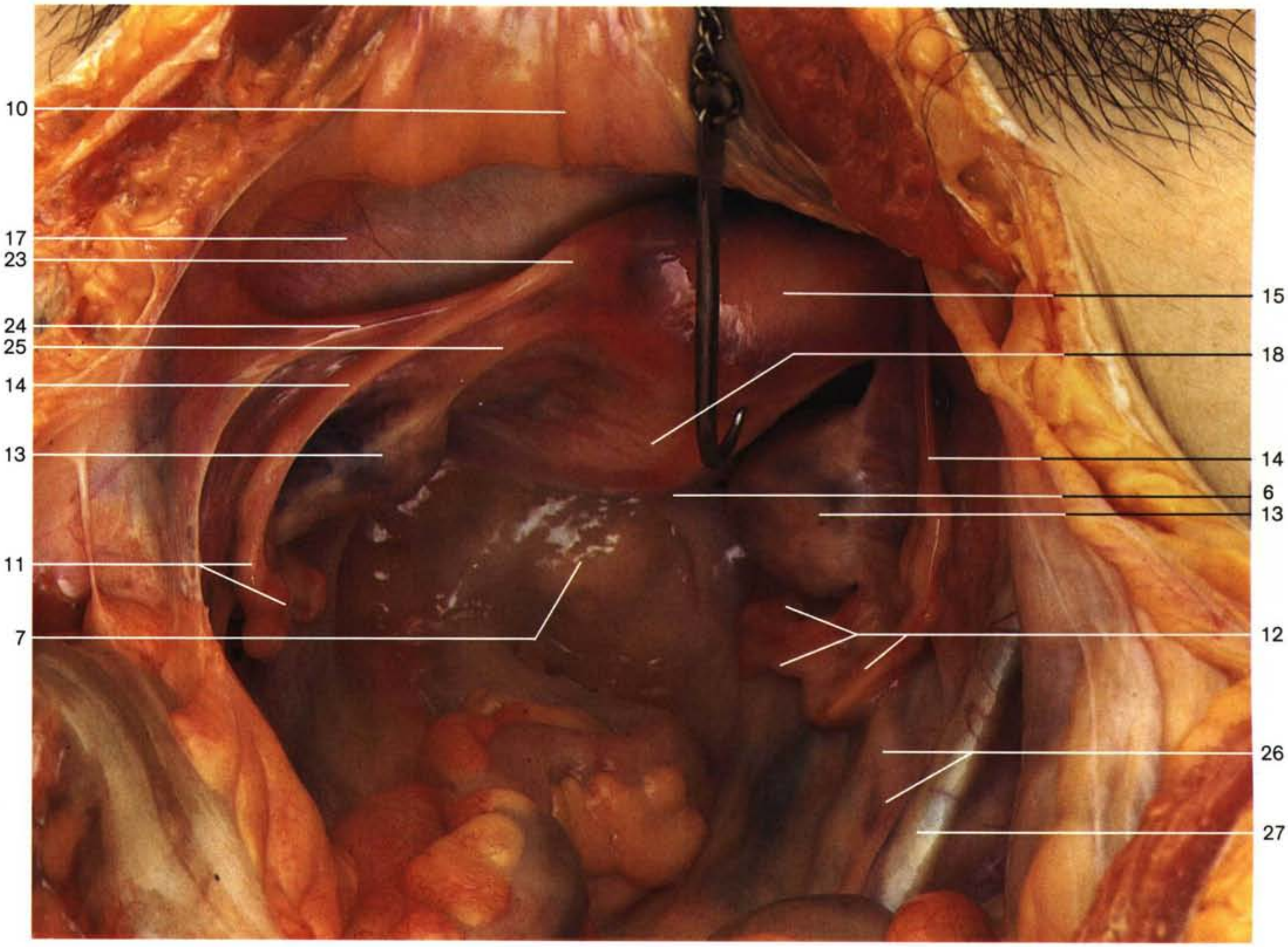

Female internal genital organs. Pelvic cavity, seen from above. The uterus has been reflected to the right.

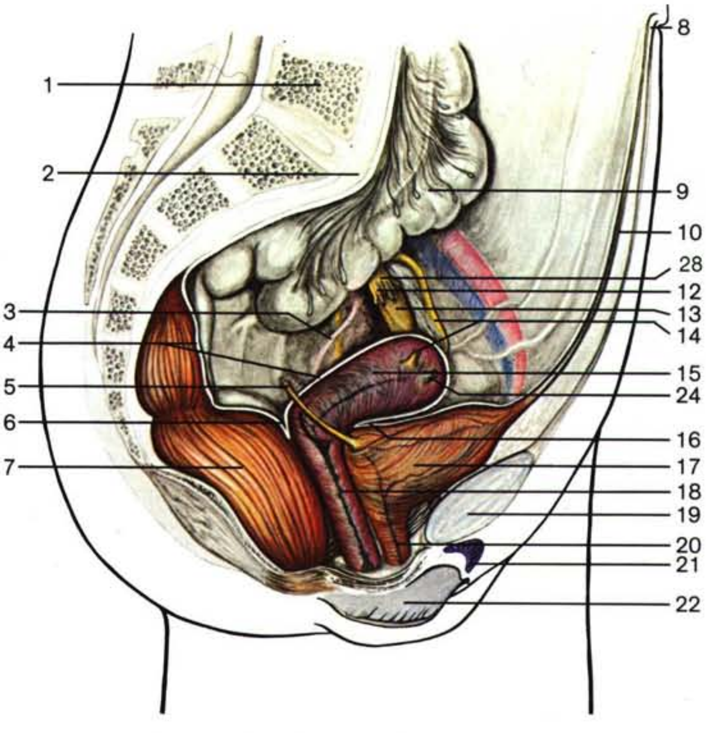

Regional relations of female internal genital organs (medial aspect). (Schematic drawing.)

1 Body of 5th lumbar vertebra
2 Sacral promontory
3 Left ureter
4 Peritoneum (cut edge)
5 Right ureter (divided)
6 **Recto-uterine pouch** (of Douglas)
7 Rectum
8 Umbilicus
9 Sigmoid colon
10 Median umbilical fold with urachus
11 Ampulla of uterine tube
12 Fimbriae of uterine tube
13 **Ovary**
14 **Uterine tube** (isthmus)
15 **Uterus**
16 **Vesico-uterine pouch**
17 Urinary bladder
18 Vagina
19 Pubic symphysis
20 Urethra
21 Clitoris
22 Labium minus
23 Insertion of uterine tube at fundus of uterus
24 Round ligament of uterus
25 Ligament of the ovary
26 Suspensory ligament of ovary
27 Right common iliac artery (covered by peritoneum)
28 Infundibulum of uterine tube

1 Body of 5th lumbar vertebra, **suspensory ligament of ovary** and sacral promontory
2 Ureter
3 Medial umbilical ligament (remnant of umbilical artery) (cut)
4 Infundibulum of uterine tube
5 Ampulla of uterine tube
6 **Ovary**
7 **Uterine artery**
8 **Uterine tube**
9 Rectum
10 Levator ani muscle (pelvic diaphragm – cut edge)
11 External anal sphincter muscle
12 Anus (probe)
13 Internal iliac artery
14 Remnant of urachus (median umbilical ligament)
15 **Uterus**
16 Round ligament of uterus
17 **Urinary bladder**
18 **Vagina**
19 **Clitoris**
20 Labium minus
21 External orifice of urethra (red probe)
22 Vaginal orifice (green probe)
23 Lateral umbilical ligament
24 Inferior epigastric artery
25 Obturator artery, vein and nerve
26 External iliac artery
27 **Recto-uterine pouch** (of Douglas)
28 Recto-uterine fold
29 Vesico-uterine pouch
30 Suspensory ligament of ovary
31 Greater vestibular gland and bulb of the vestibule

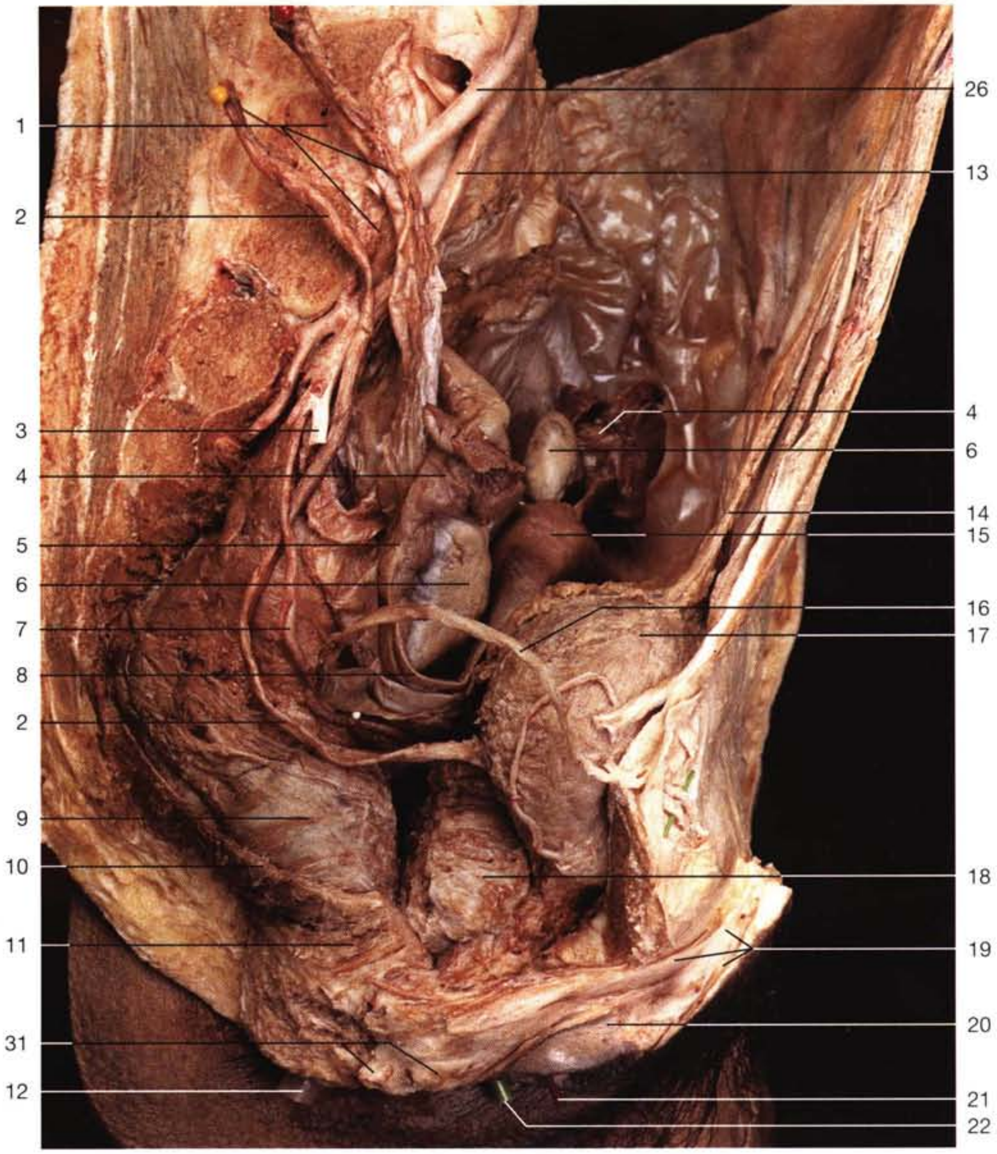

Female internal genital organs in situ. Right half of the pelvis and sacrum have been removed.

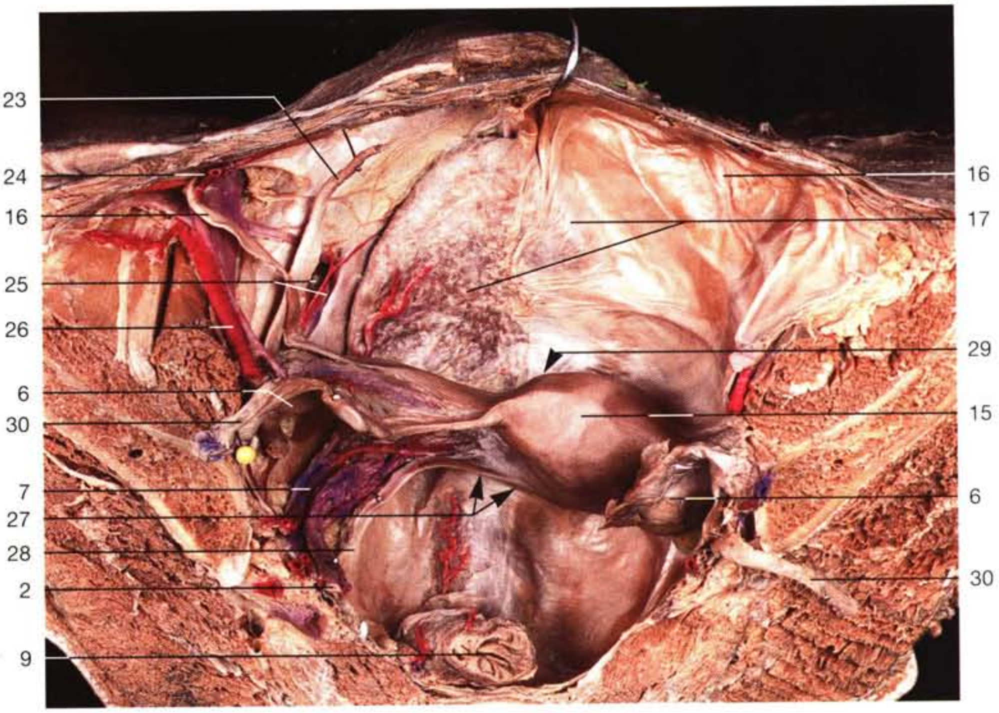

Female internal genital organs in situ (seen from above). The peritoneum at the left half of pelvic cavity has been removed to display uterine tube, vessels, and nerves.

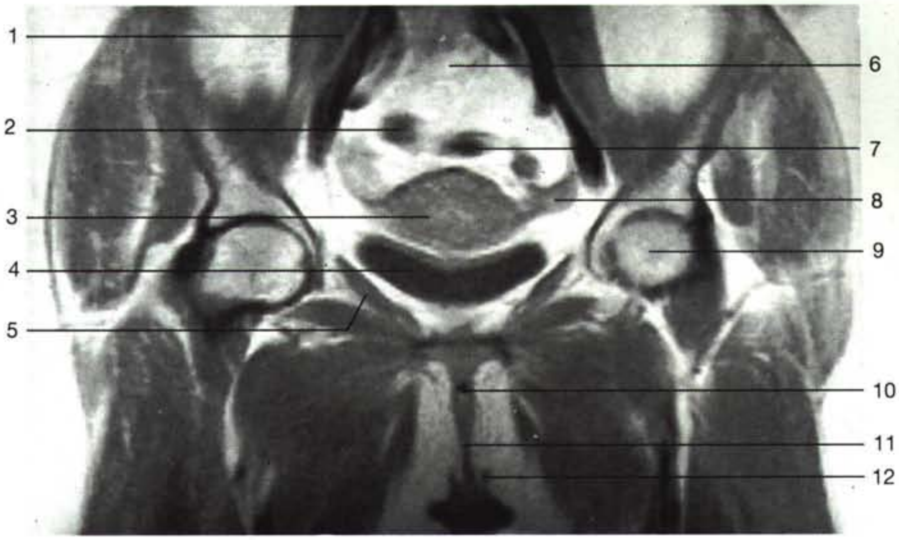

Coronal section through the pelvic cavity of the female. (MR-Scan.)

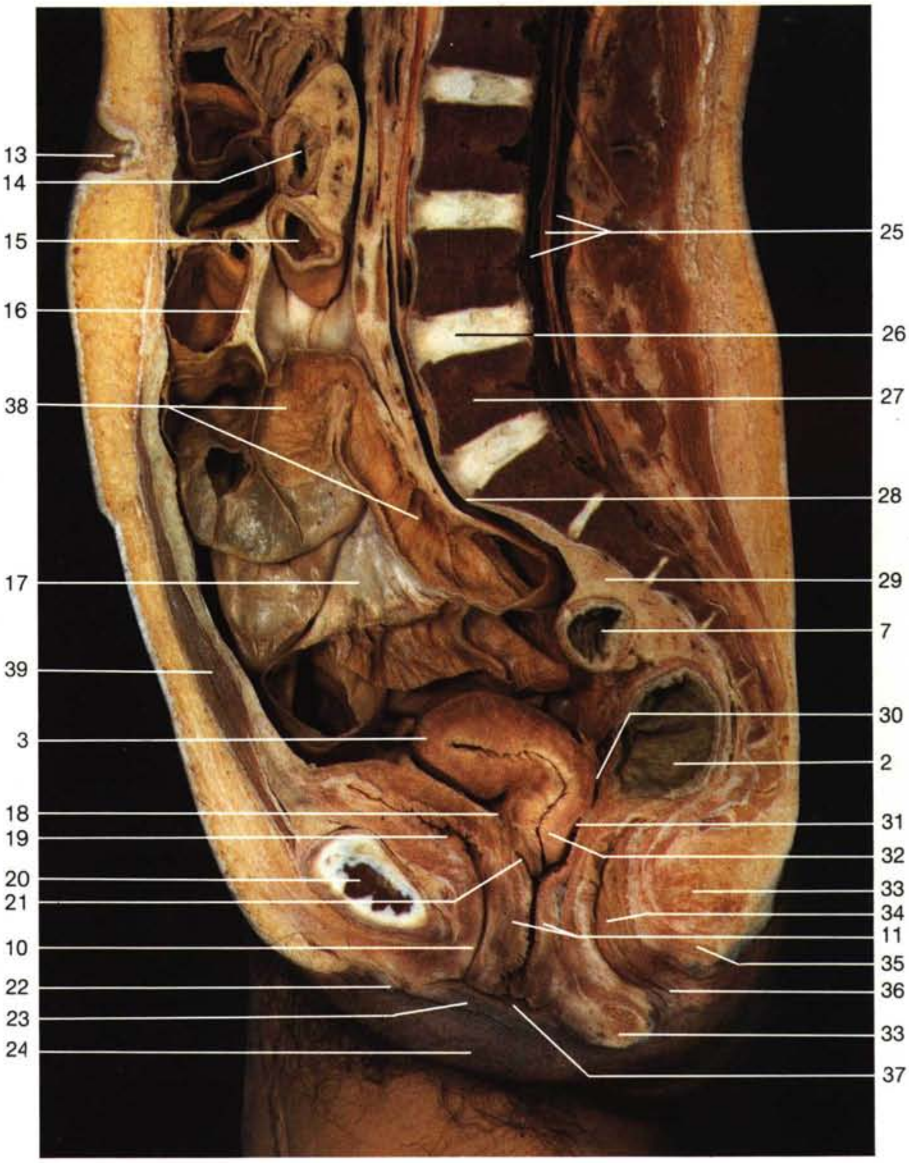

Midsagittal section through the female trunk. The urinary bladder is empty, the position and shape of the uterus are normal.

1 Psoas major muscle
2 Ampulla of rectum
3 **Uterus**
4 **Urinary bladder**
5 Obturator internus muscle
6 Promontory
7 Sigmoid colon
8 **Uterine tube**
9 Head of femur
10 Urethra
11 Vagina
12 Labium minus
13 Umbilicus
14 Duodenum
15 Ascending part of duodenum
16 Root of mesentery
17 Mesentery
18 Vesico-uterine pouch
19 **Urinary bladder** (collapsed)
20 Pubic symphysis
21 Anterior fornix of vagina
22 Clitoris
23 Labium minus
24 Labium majus
25 Vertebral canal with cauda equina
26 Intervertebral disc
27 Body of 5th lumbar vertebra
28 Sacral promontory
29 Mesosigmoid
30 **Recto-uterine pouch** (of Douglas)
31 Posterior fornix of vagina
32 Cervix of uterus
33 External anal sphincter muscle
34 Anal canal
35 Internal anal sphincter muscle
36 Anus
37 Hymen
38 Small intestine
39 Rectus abdominis muscle

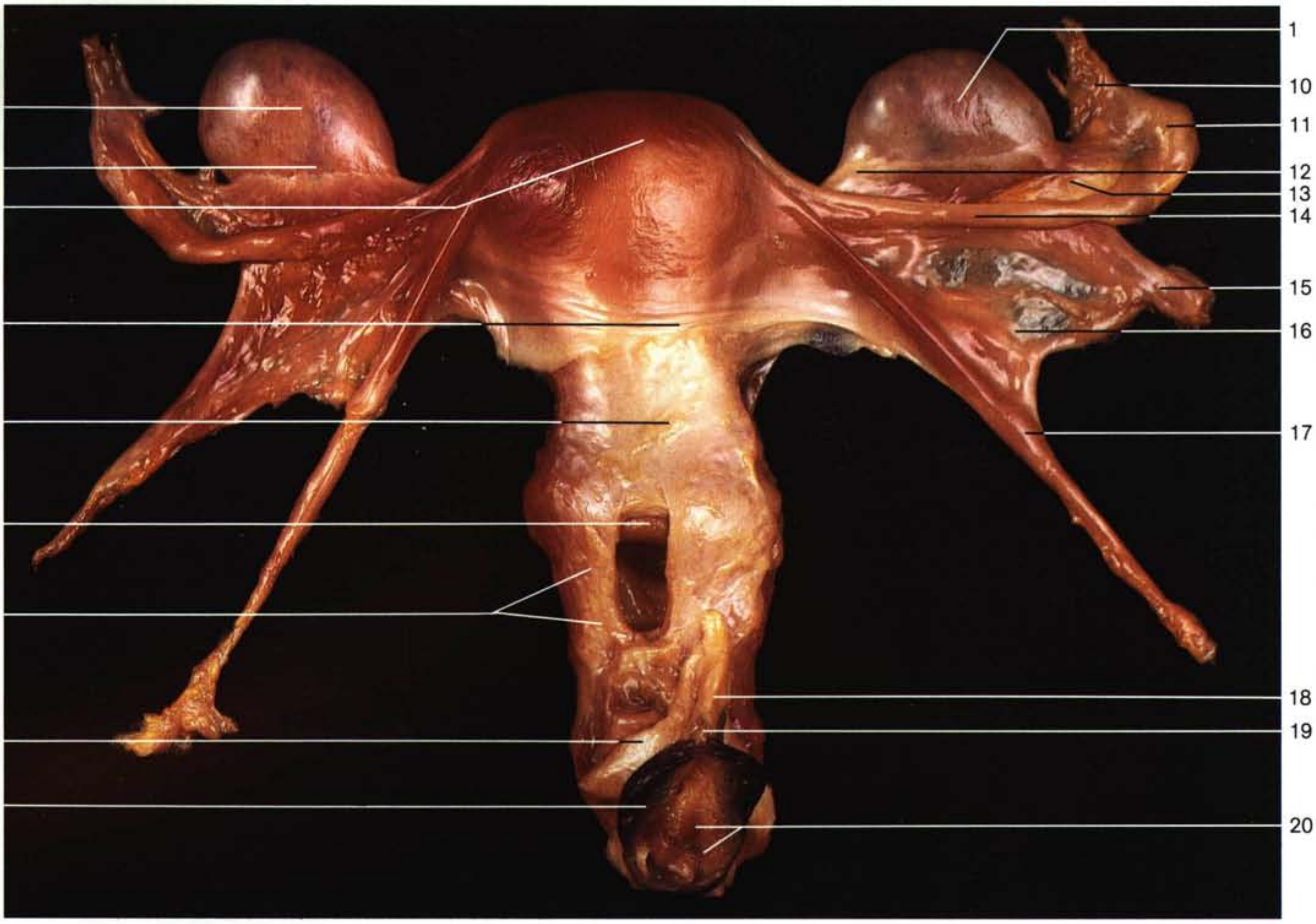

Female genital organs, isolated (anterior view). The anterior wall of the vagina has been opened to display the vaginal portion of the cervix.

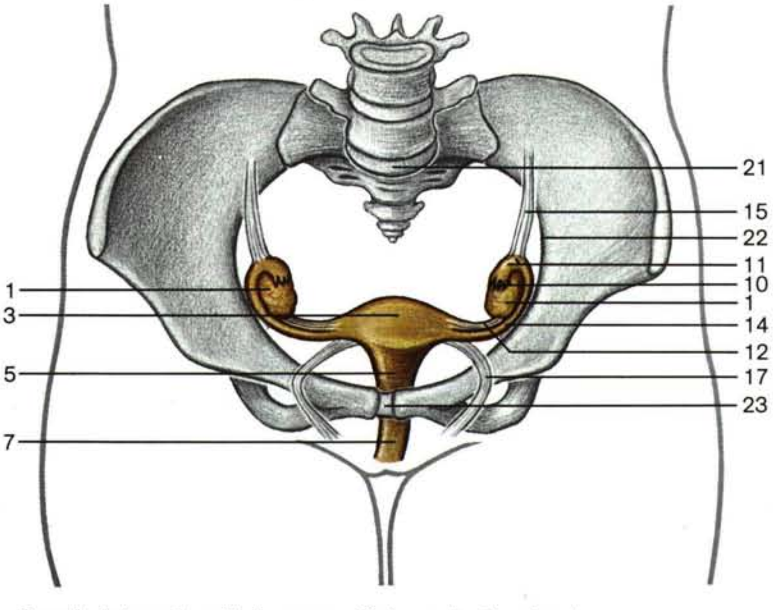

Female internal genital organs. (Schematic drawing.)

1 **Ovary**
2 Mesovarium
3 **Fundus of uterus**
4 Vesico-uterine pouch
5 **Cervix of uterus**
6 Vaginal portion of cervix
7 **Vagina**
8 Crus of clitoris
9 Labium minus
10 Fimbriae of uterine tube
11 Infundibulum of uterine tube
12 Ligament of the ovary
13 Mesosalpinx
14 Uterine tube
15 Suspensory ligament of ovary (caudally displaced)
16 Broad ligament of uterus
17 Round ligament of uterus
18 Corpus cavernosum of clitoris
19 Glans of clitoris
20 **Hymen, vaginal orifice**
21 Promontory
22 Linea terminalis of pelvis
23 Pubic symphysis

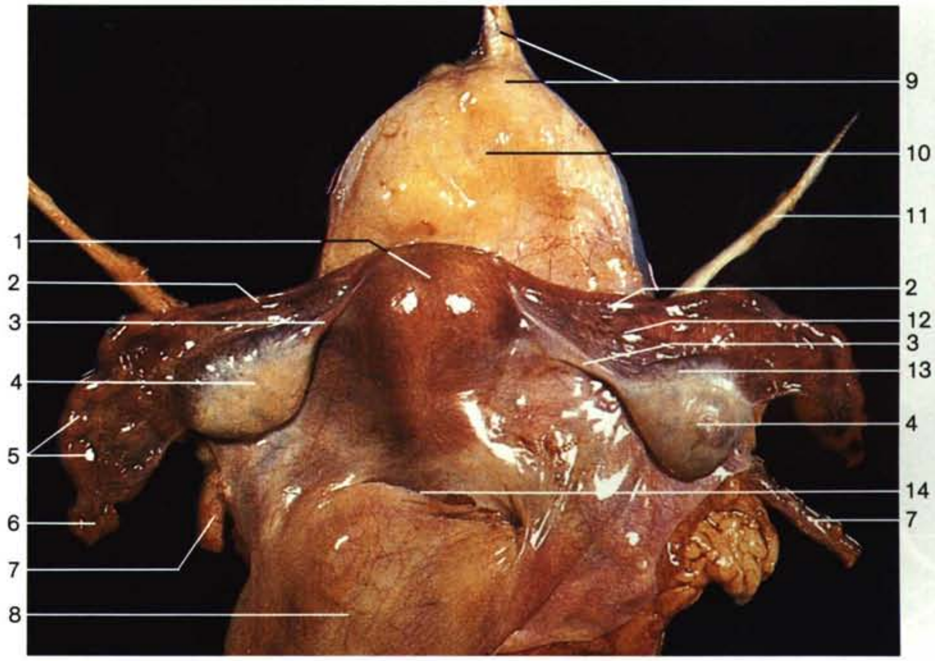

Female internal genital organs, isolated (superior-posterior view).

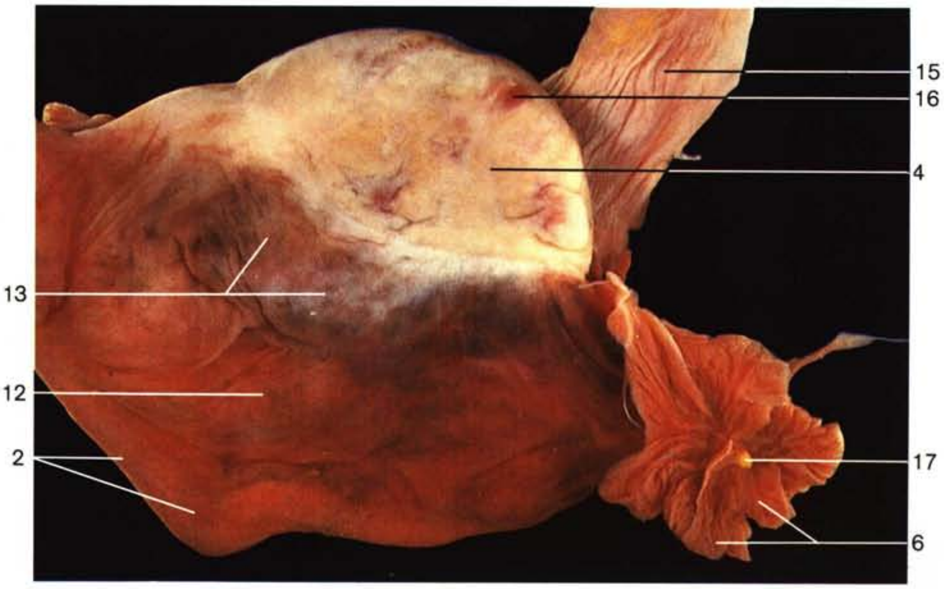

Right ovary and uterine tube, isolated (superior-posterior view). The fimbriae of the uterine tube have been reflected to show the abdominal ostium.

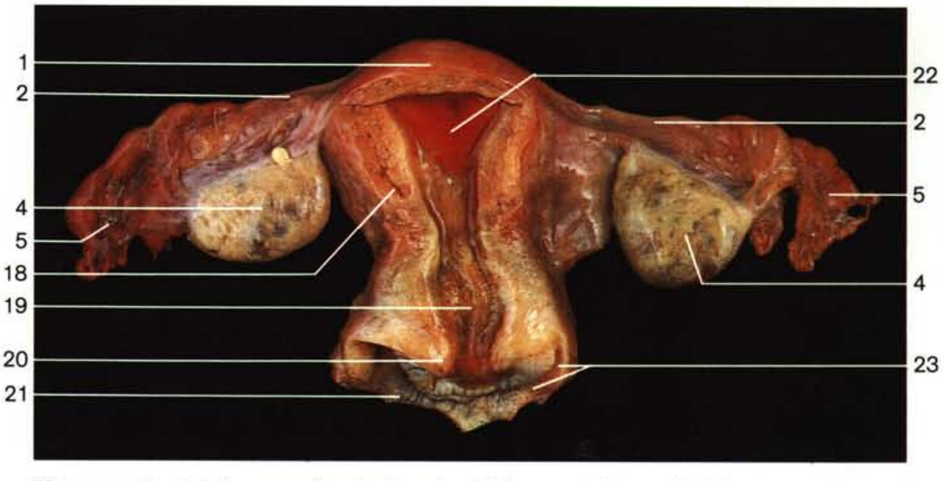

Uterus and related organs (posterior view). The posterior wall of the uterus has been opened.

1 **Fundus of uterus**
2 Uterine tube
3 **Ligament of the ovary**
4 **Ovary**
5 Infundibulum of uterine tube
6 Fimbriae of uterine tube
7 **Ureter**
8 Rectum
9 Apex of urinary bladder and median umbilical ligament
10 Urinary bladder
11 Round ligament of uterus
12 **Mesosalpinx**
13 **Mesovarium**
14 **Rectouterine pouch** (of Douglas)
15 Suspensory ligament of ovary
16 Scarring of ovary (following ovulation)
17 Abdominal opening of uterine tube
18 Body of uterus
19 **Cervical canal**
20 Vaginal portion of cervix of uterus (congestion)
21 Vagina
22 Mucous membrane of uterus
23 Anterior fornix of vagina

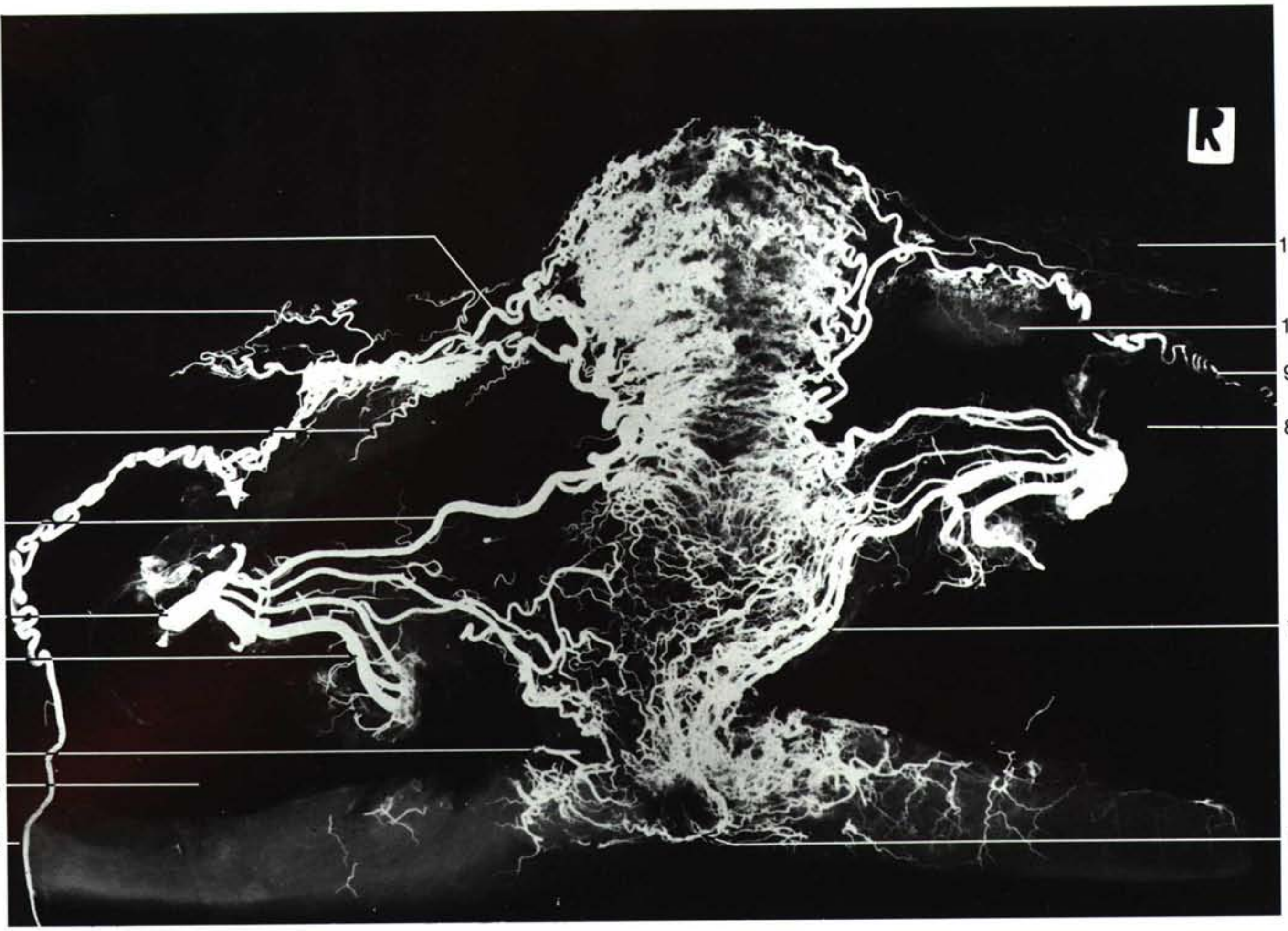

Arteriogram of female genital organs (anterior-posterior view). Notice the helicine arteries supplying uterus and ovary.

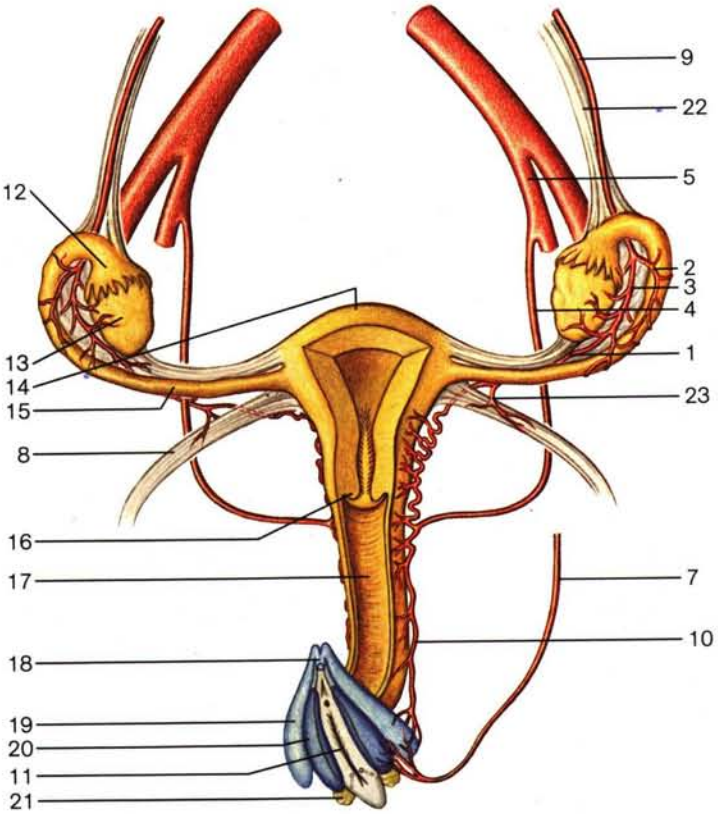

Arteries of female genital organs. (Schematic drawing.)

1 Ovarian branch of uterine artery (anastomoses with ovarian artery)
2 Tubal branch of ovarian artery
3 Ovarian branch of ovarian artery
4 **Uterine artery**
5 Internal iliac artery
6 Inferior gluteal artery
7 **Internal pudendal artery**
8 Round ligament of uterus
9 **Ovarian artery**
10 Vaginal artery
11 Vaginal orifice
12 Infundibulum of uterine tube
13 Ovary
14 Fundus of uterus
15 Uterine tube
16 Vaginal portion of cervix of uterus
17 Vagina
18 Clitoris
19 Corpus cavernosum of clitoris
20 Bulb of vestibule
21 Greater vestibular gland
22 Suspensory ligament of ovary
23 Artery of round ligament

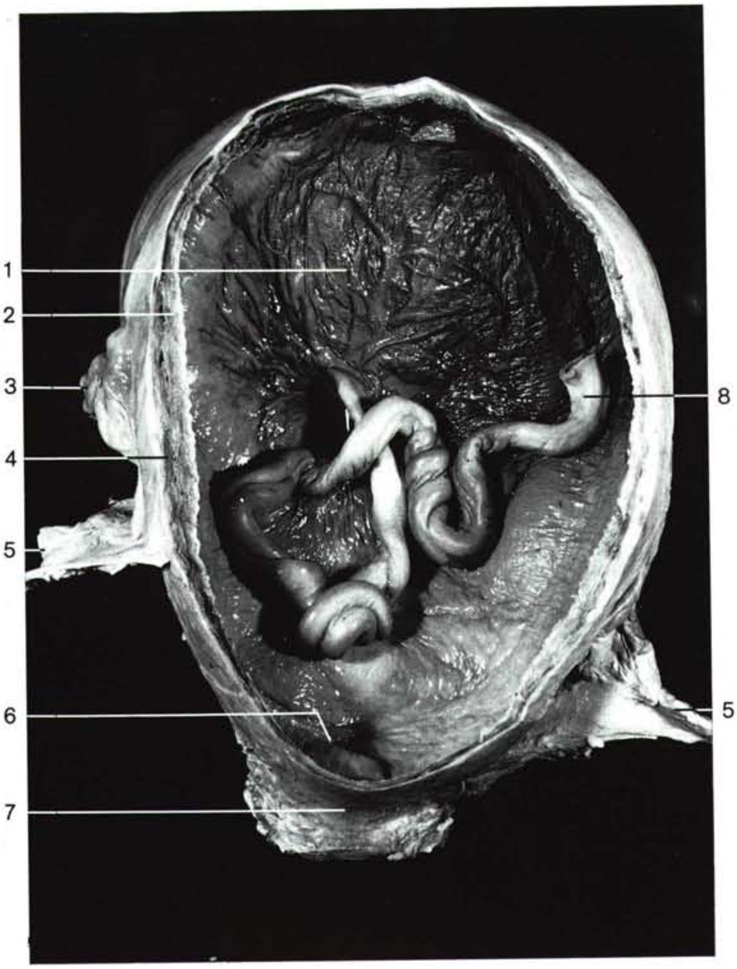

Full term uterus with placenta (anterior view). The anterior wall of the uterus has been removed to show the location of the placenta.

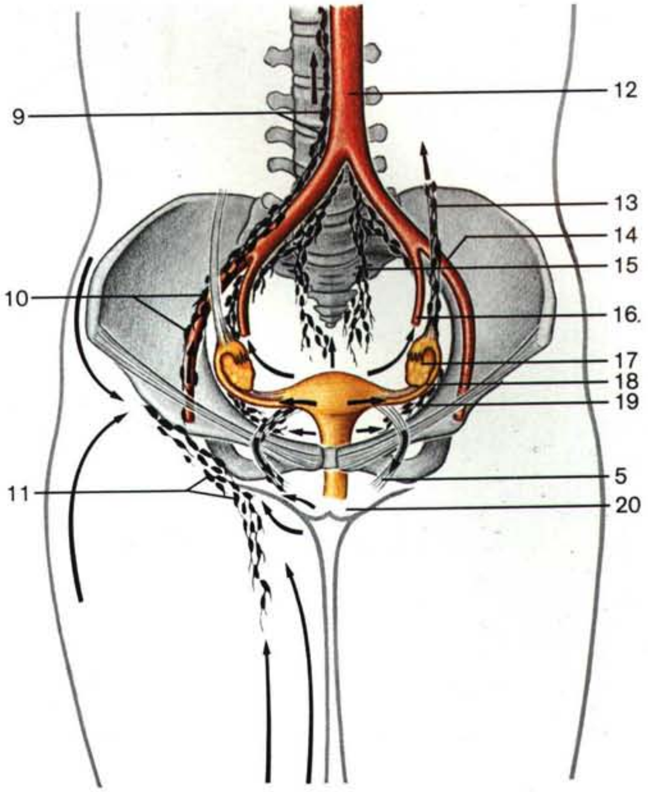

Main drainage routes of lymph vessels of uterus and its adnexa (indicated by arrows). (Schematic drawing.)
Red = arteries; black = lymph vessels and nodes; yellow = internal genital organs.

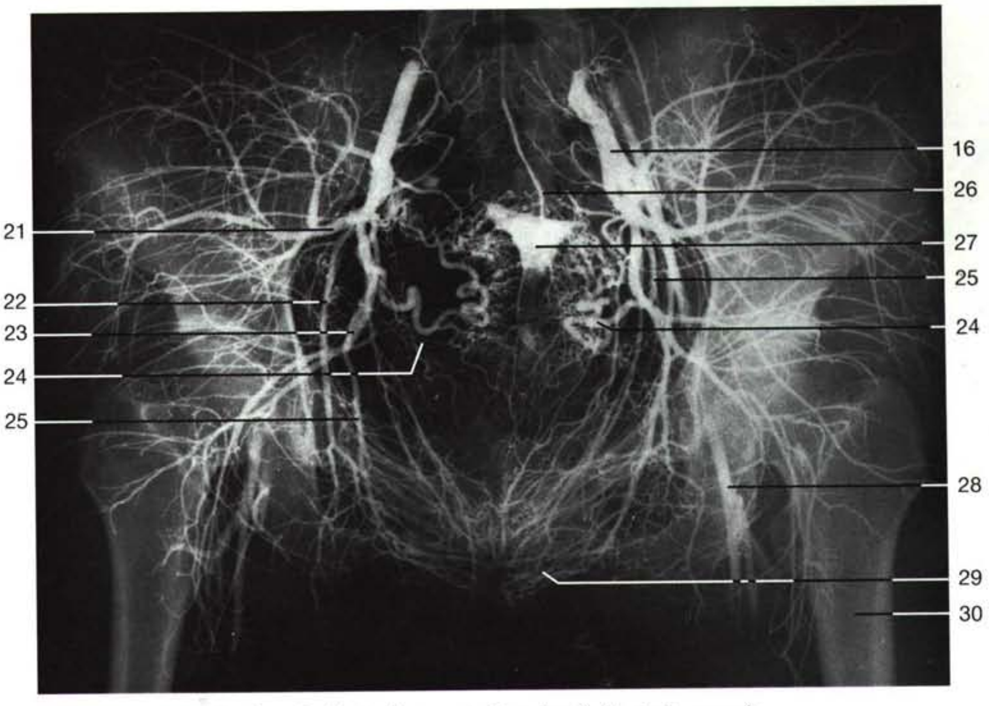

Pelvic vessels in the female (anterior-posterior view). (Arteriogram.)

1 Placenta
2 Amnion and chorion
3 Adnexa of uterus (uterine tube and ovaries)
4 Myometrium
5 Round ligament of uterus
6 Internal orifice of uterus
7 Cervix of uterus
8 **Umbilical cord**
9 Lumbar lymph nodes
10 External iliac lymph nodes
11 Inguinal lymph nodes
12 Abdominal aorta
13 Suspensory ligament of ovary
14 External iliac artery
15 Sacral lymph nodes
16 Internal iliac artery
17 Ovary
18 Uterine tube
19 Internal iliac lymph nodes
20 External genital organs
21 Superior gluteal artery
22 **Obturator artery**
23 Inferior gluteal artery
24 **Uterine artery**
25 **Internal pudendal artery**
26 Middle sacral artery
27 **Uterine cavity**
28 Femoral artery
29 Vessels of labium majus
30 Femur

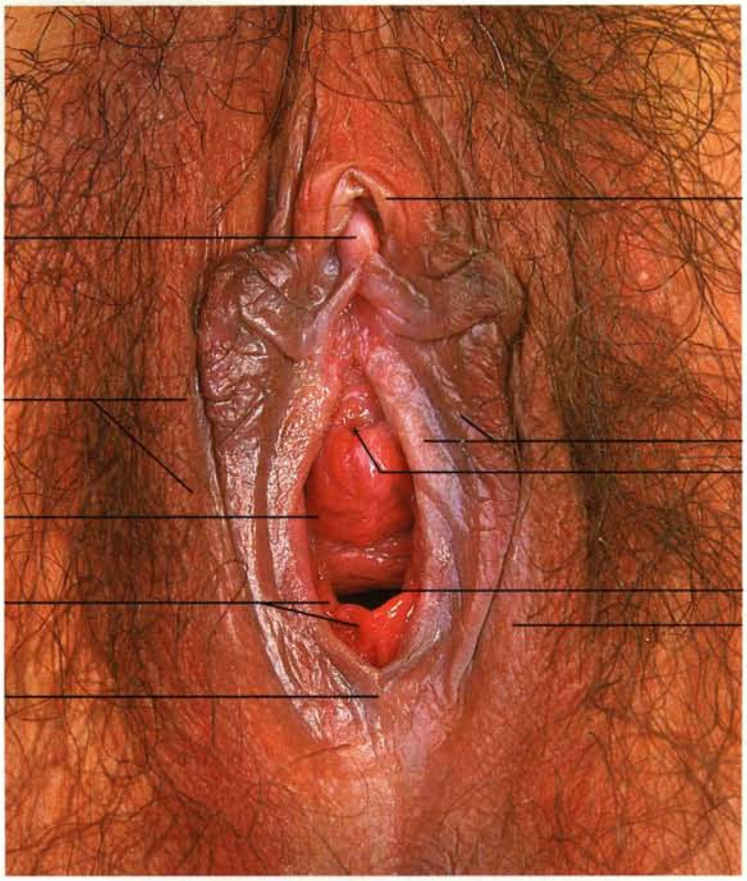

External genital organs in the female (anterior aspect). Labia reflected.

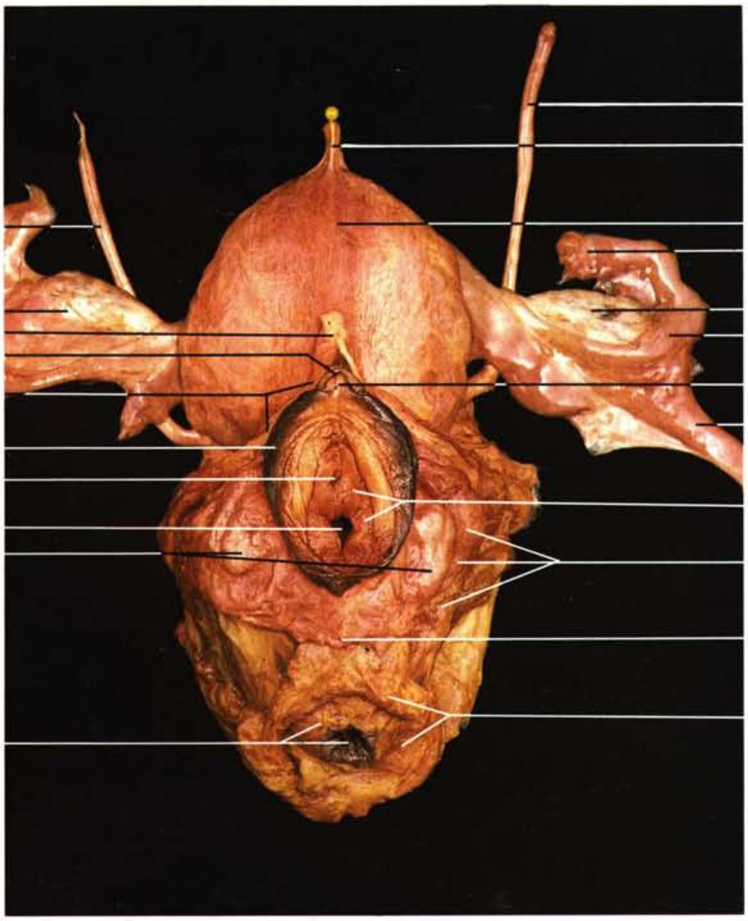

Female external genital organs in relation to internal genital organs and urinary system (isolated, anterior aspect).

1 **Glans clitoris**
2 Labium majus
3 Vestibule of vagina
4 **Hymen**
5 Posterior labial commissure
6 **Body of clitoris**
7 Labium minus
8 **External orifice of urethra**
9 **Vaginal orifice**
10 Ureter
11 Adnexa of uterus
12 Prepuce of clitoris
13 Crus of clitoris
14 Greater vestibular glands
15 Anus and internal anal sphincter muscle
16 Median umbilical ligament containing urachus
17 **Urinary bladder**
18 Infundibulum of uterine tube
19 Ovary
20 Ampulla of uterine tube
21 Suspensory ligament of the ovary
22 Bulbospongiosus muscle and bulb of vestibule
23 Central tendon of perineum (perineal body)
24 External anal sphincter muscle

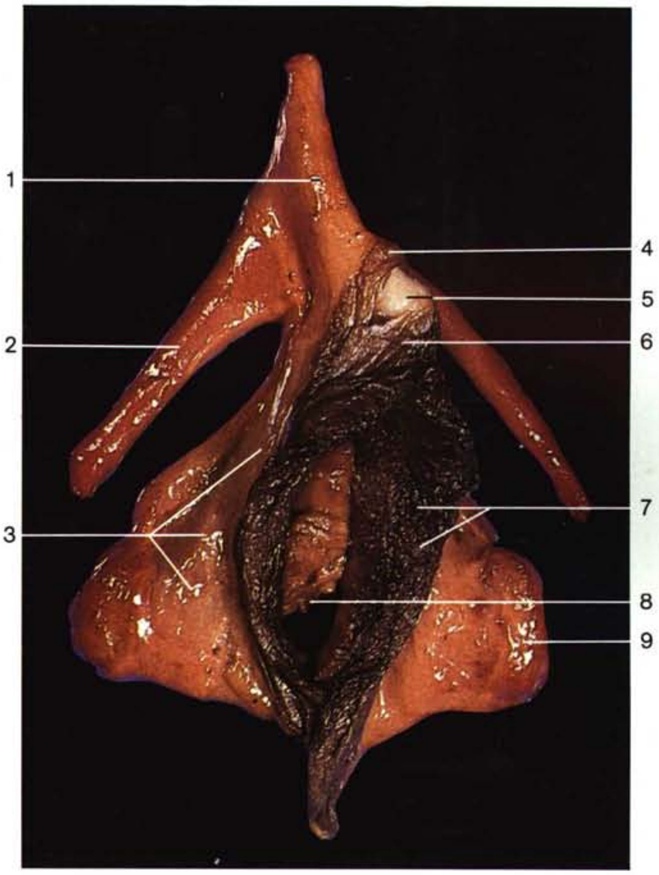

Cavernous tissue of female external genital organs, isolated (anterior aspect).

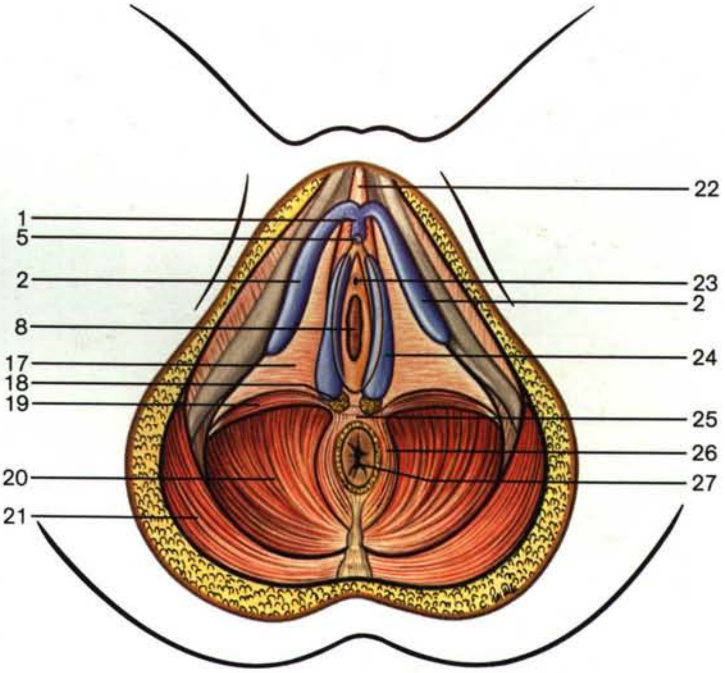

Urogenital and pelvic diaphragms (anterior aspect).
(Schematic drawing.)
Blue = cavernous tissue of clitoris and bulb of vestibule.

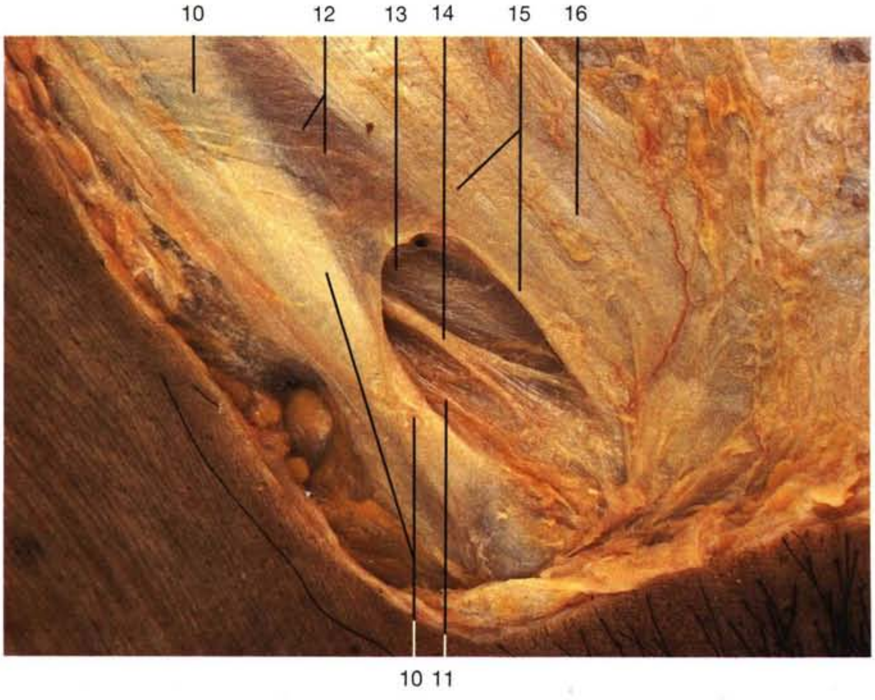

Inguinal canal and round ligament of uterus in situ (right side, ventral aspect).

1 Body of clitoris
2 Crus of clitoris
3 Bulb of vestibule
4 Prepuce of clitoris
5 **Glans of clitoris**
6 Frenulum of clitoris
7 **Labium minus**
8 **Vaginal orifice**
9 **Greater vestibular gland**
10 Lateral crus of superficial inguinal ring
11 Ilioinguinal nerve
12 Intercrural fibers
13 Superficial inguinal ring
14 **Round ligament of uterus**
15 Medial crus of superficial inguinal ring
16 Aponeurosis of external abdominal oblique muscle
17 Deep transverse perineus muscle with fascia
18 Greater vestibular gland
19 Superficial transverse perineus muscle
20 Levator ani muscle
21 Gluteus maximus muscle
22 Suspensory ligament of clitoris
23 External orifice of urethra
24 Bulb of vestibule
25 Perineal body
26 External anal sphincter muscle
27 Anus

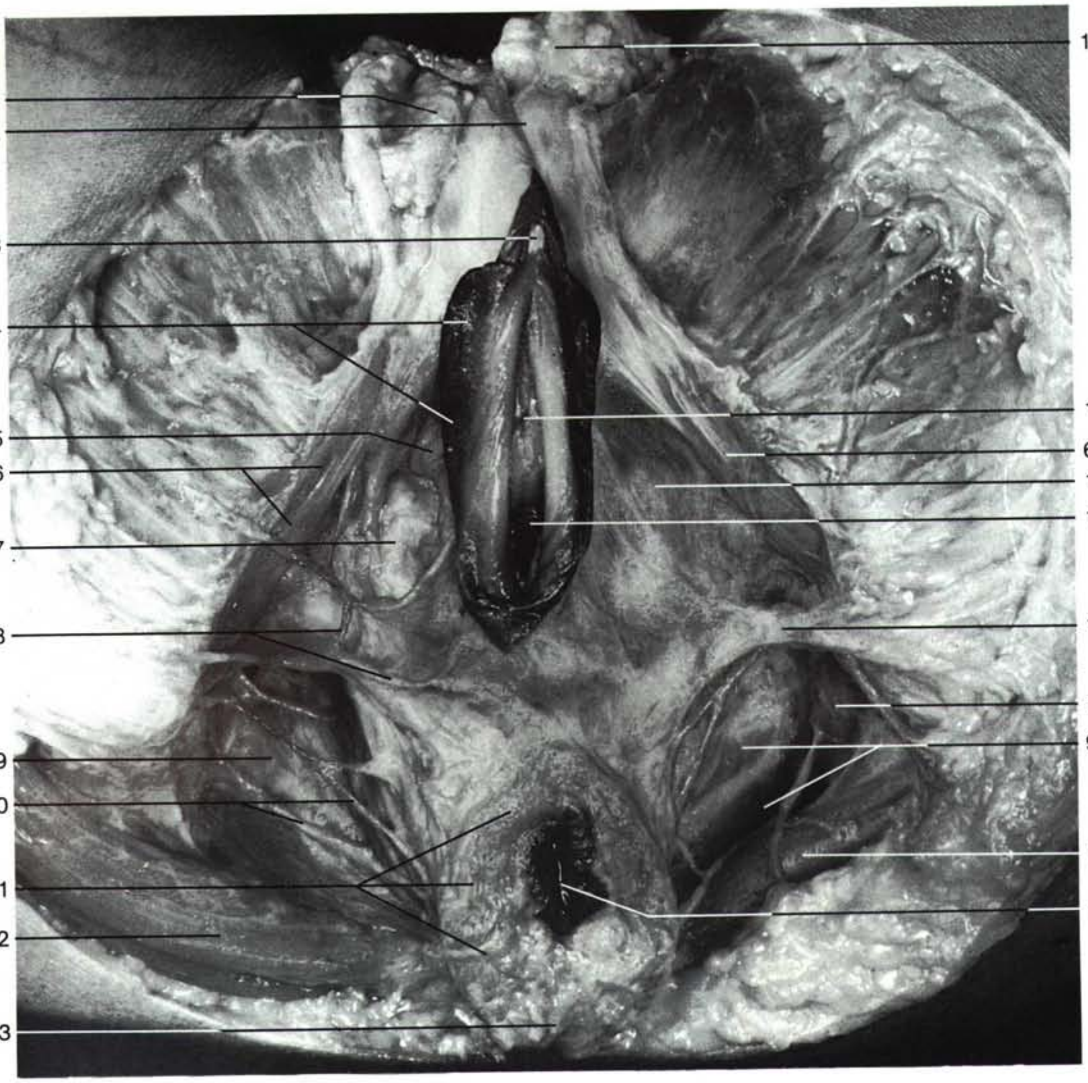

Female urogenital diaphragm and external genital organs, superficial layer (from below).

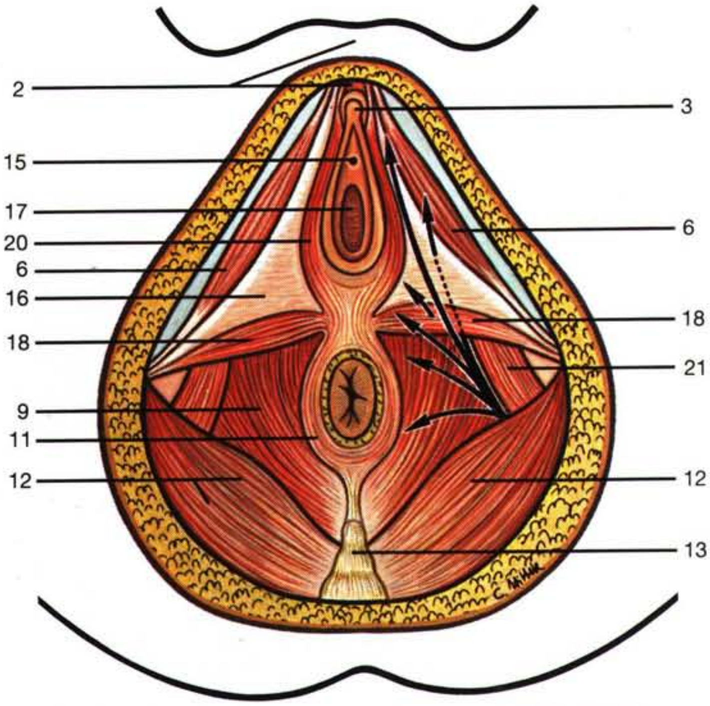

Muscles of pelvic and urogenital diaphragms (from below).
(Schematic drawing.)

1 Fatty tissue encasing round ligament
2 Position of pubic symphysis
3 Clitoris
4 Labium minus
5 Bulb of vestibule
6 Ischiocavernosus muscle
7 **Greater vestibular gland**
8 Perineal branches of pudendal nerve
9 Levator ani muscle
10 Inferior rectal nerves
11 External anal sphincter muscle
12 Gluteus maximus muscle
13 Coccyx
14 Fatty tissue of mons pubis
15 **External orifice of urethra**
16 Urogenital diaphragm with fascia of deep transverse perineus muscle
17 **Vaginal orifice**
18 Superficial transverse perineal muscle
19 **Anus**
20 Bulbospongiosus muscle
21 Obturator internus muscle

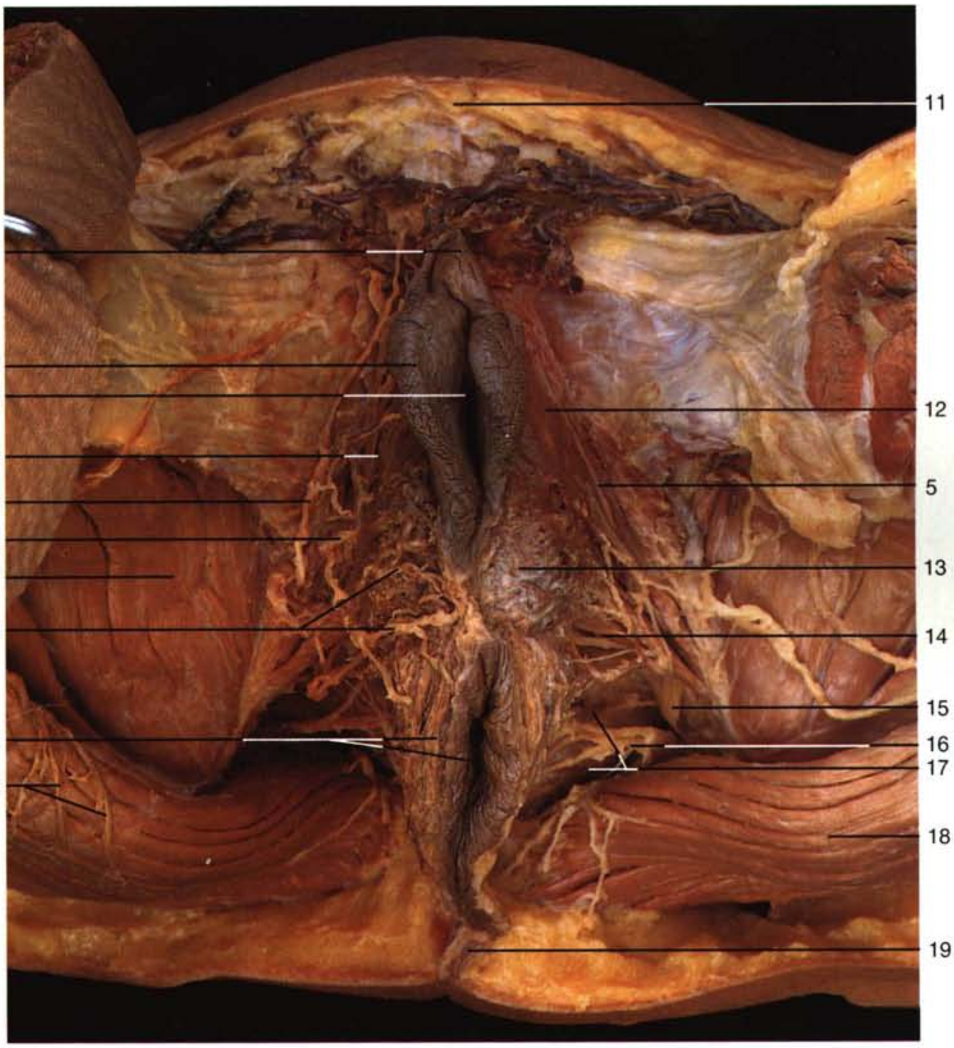

Urogenital diaphragm and external genital organs in the female, superficial layer (from below). On the right side the bulb of vestibule has been removed.

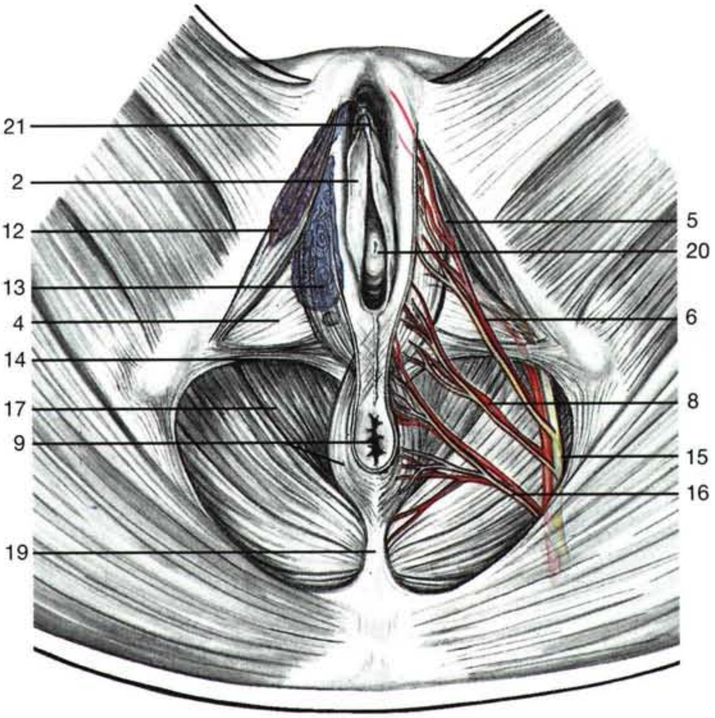

External female genital organs. Position of arteries and nerves; bulb of vestibule in blue. (Schematic drawing.)

1	**Prepuce of clitoris**
2	Labium minus
3	**Vaginal orifice**
4	Deep transverse perineus muscle
5	Dorsal nerve of clitoris
6	Posterior labial nerves
7	Great adductor muscle
8	Perineal branches of pudendal nerve
9	**Anus** and **external anal sphincter muscle**
10	Inferior cluneal nerves
11	Mons pubis
12	Crus of clitoris with ischiocavernosus muscle
13	Bulb of vestibule
14	Superficial transverse perineus muscle
15	**Pudendal nerve** and **internal pudendal artery**
16	Inferior rectal nerves
17	Levator ani muscle
18	Gluteus maximus muscle
19	Anococcygeal ligament
20	External urethral orifice
21	Glans of clitoris

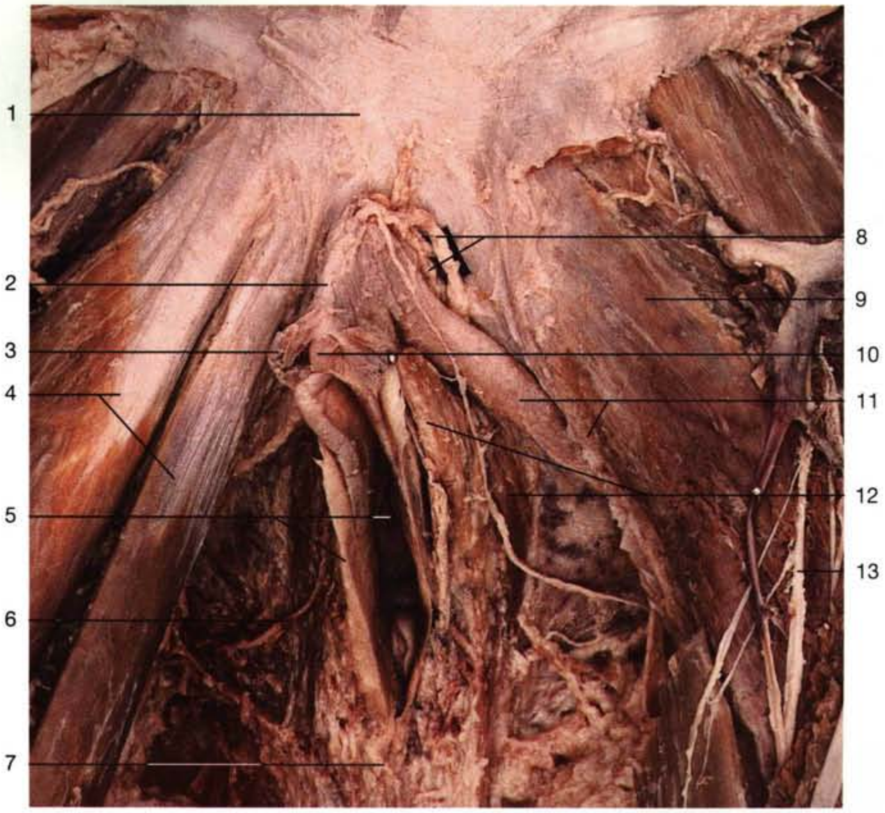

External female genital organs (inferior aspect). The clitoris has been dissected and slightly reflected to the right. The prepuce of clitoris has been divided to display the glans.

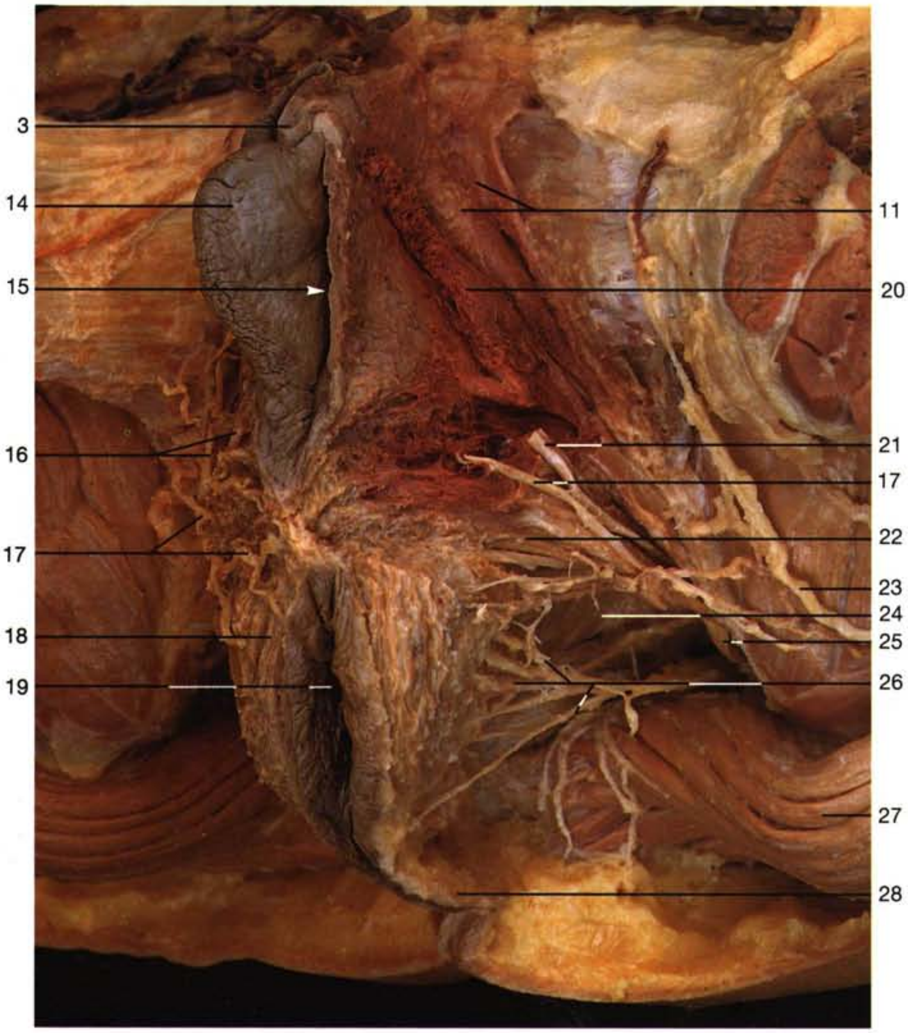

Urogenital diaphragm and external genital organs in the female (lateral inferior view). The bulb of vestibule has partly been removed, the left labium minus was cut away.

1 Position of pubic symphysis
2 Body of clitoris
3 Prepuce of clitoris
4 Adductor longus and gracilis muscles
5 External orifice of vagina and labium minus
6 Posterior labial nerve
7 Perineal body
8 Deep artery of clitoris and dorsal nerve of clitoris
9 Adductor brevis muscle
10 **Glans of clitoris**
11 Crus of clitoris and ischiocavernosus muscle
12 **Bulb of vestibule** and bulbospongiosus muscle
13 Anterior branch of obturator nerve
14 Labium minus
15 **Vaginal orifice**
16 Posterior labial nerves
17 Branches of pudendal nerve
18 External sphincter of anus
19 **Anus**
20 Bulb of vestibule (divided)
21 Dorsal artery of clitoris
22 Superficial transverse perineus muscle
23 Perineal branch of posterior femoral cutaneous nerve
24 Levator ani muscle
25 **Pudendal nerve** and **internal pudendal artery**
26 Inferior rectal nerves
27 Gluteus maximus muscle
28 Anococcygeal ligament

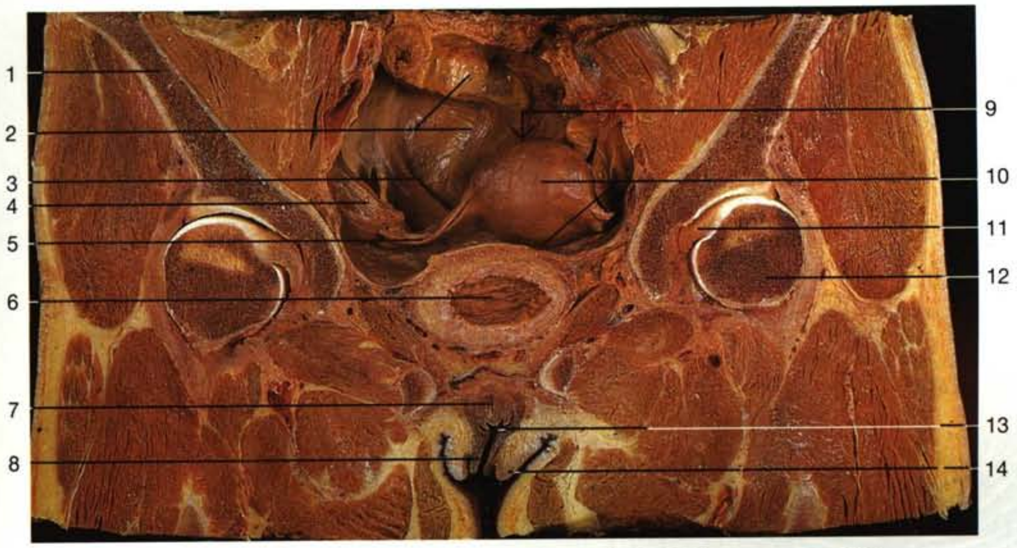

Coronal section through the pelvic cavity of the female (cf. MR-Scan on page 335).

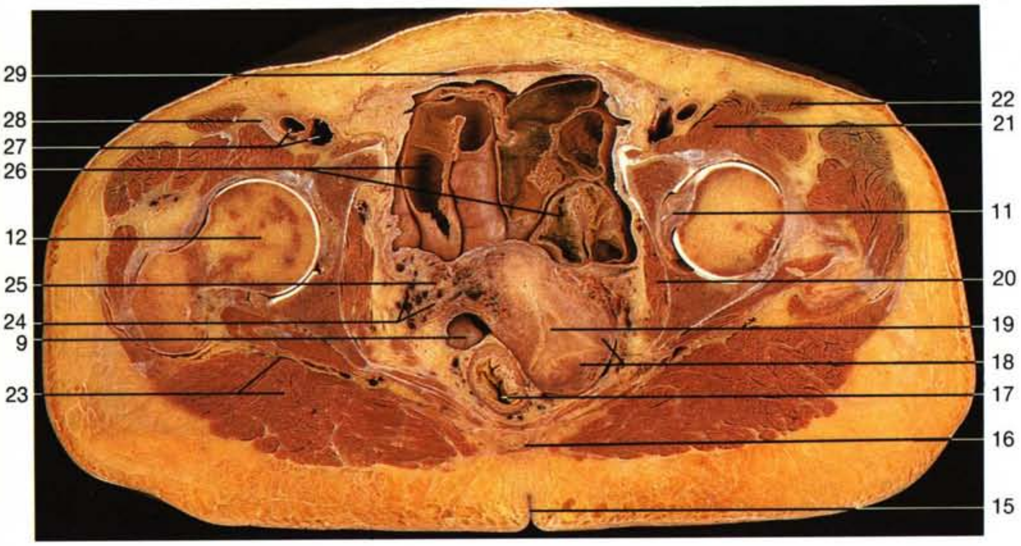

Horizontal section through pelvic cavity at level of uterus (from below). The uterus is retroverted to the left.

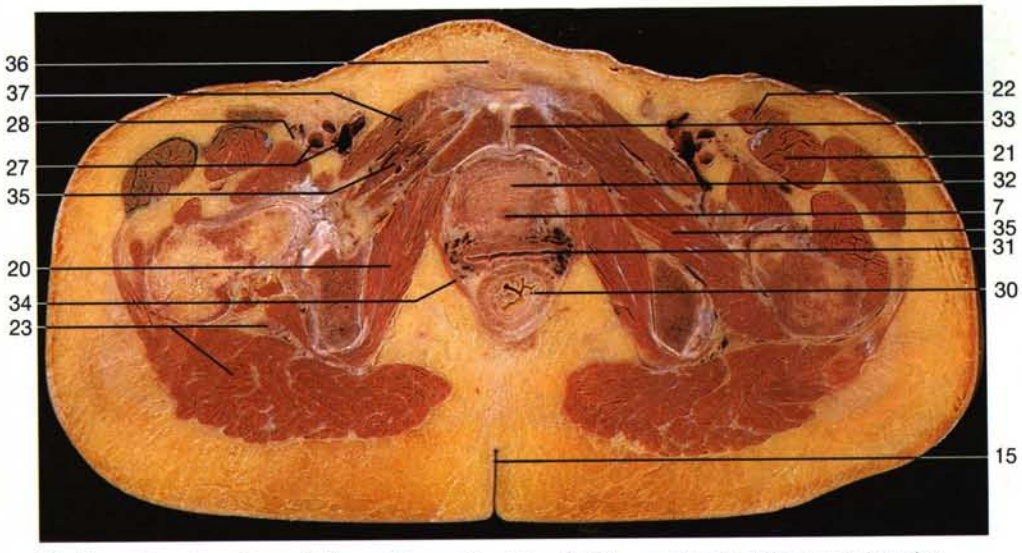

Horizontal section through the pelvic cavity at level of the urethral sphincter and vagina (from below).

1 Ilium
2 Rectum
3 Rectouterine fold
4 **Ovary**
5 **Uterine tube**
6 **Urinary bladder**
7 **Urethra**
8 Labium minus
9 Rectouterine pouch of Douglas
10 **Uterus** (uterovesical pouch)
11 Ligament of the head of the femur
12 Head of femur
13 Vestibule of vagina
14 Labium majus
15 Anal cleft
16 Coccyx
17 Rectum
18 Myometrium of uterus
19 Uterine cavity
20 Obturator internus muscle
21 Iliopsoas muscle
22 Sartorius muscle
23 Sciatic nerve and gluteus maximus muscle
24 Uterine venous plexus
25 Broad ligament
26 Small intestine
27 Femoral artery and vein
28 Femoral nerve
29 Pyramidalis muscle
30 Rectum (anal canal)
31 Vagina
32 Urethral sphincter muscle (base of urinary bladder)
33 Pubic symphysis
34 Levator ani muscle
35 Obturator externus muscle
36 Mons pubis
37 Pectineus muscle

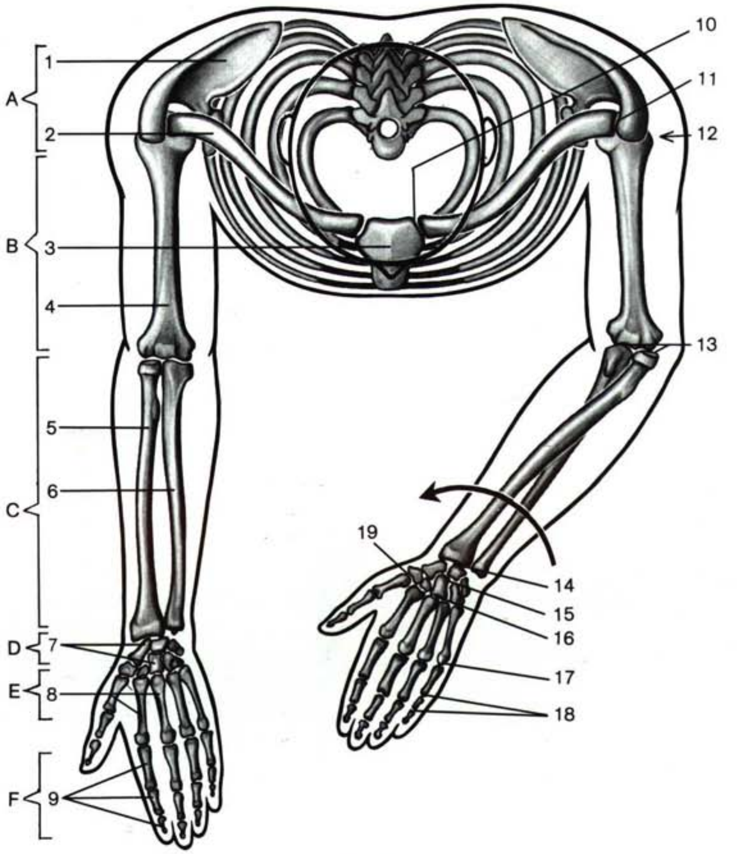

A	Pectoral girdle
B	Arm
C	Forearm
D	Wrist
E	Palm of hand
F	Finger

	Bones
1	Scapula
2	Clavicle
3	Sternum
4	Humerus
5	Radius
6	Ulna
7	Carpal bones
8	Metacarpal bones
9	Phalanges

	Joints
10	Sternoclavicular joint
11	Acromioclavicular joint
12	Shoulder joint
13	Elbow joint
14	Radiocarpal joint
15	Midcarpal joint
16	Carpometacarpal joint
17	Metacarpophalangeal joint
18	Interphalangeal joints of fingers
19	Carpometacarpal joint of thumb

Organization of shoulder girdle and upper limb (superior aspect). The two positions of the forearm essential to manual skills in the human, supination (right arm) and pronation (left arm) are shown.

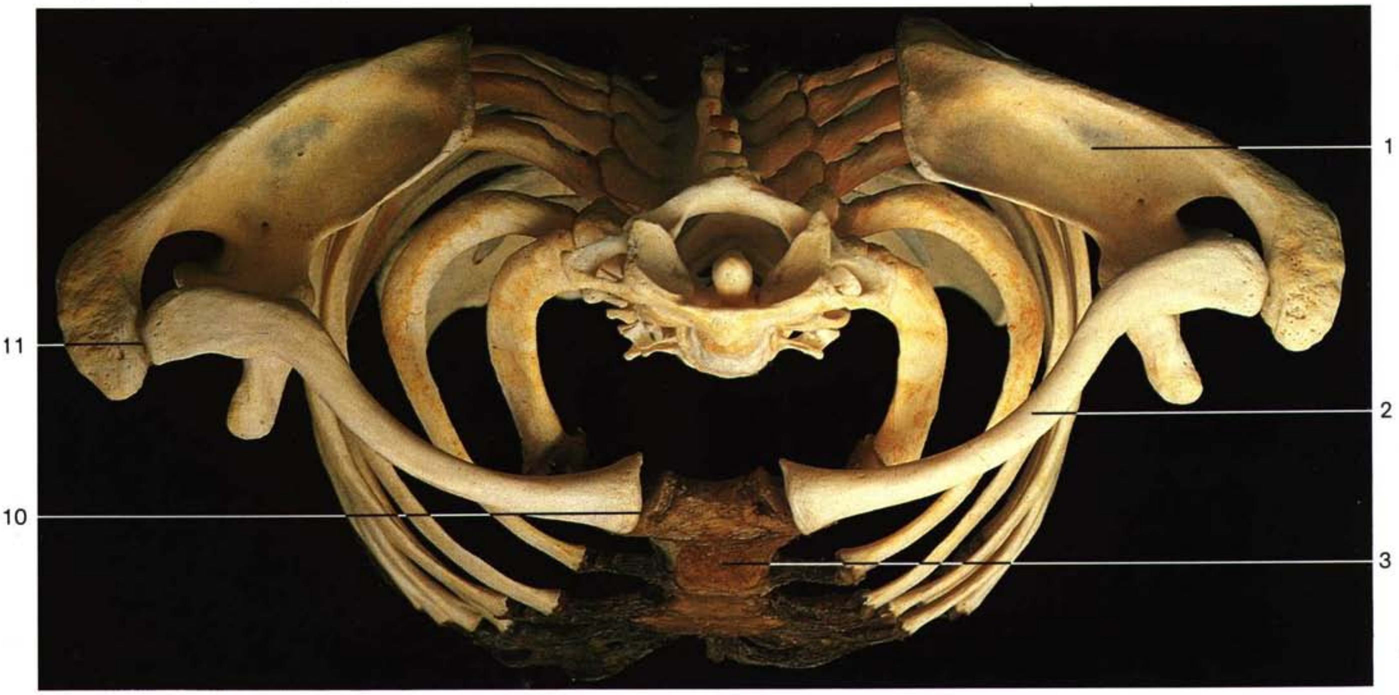

Bones of shoulder girdle articulated with the thorax (superior aspect).

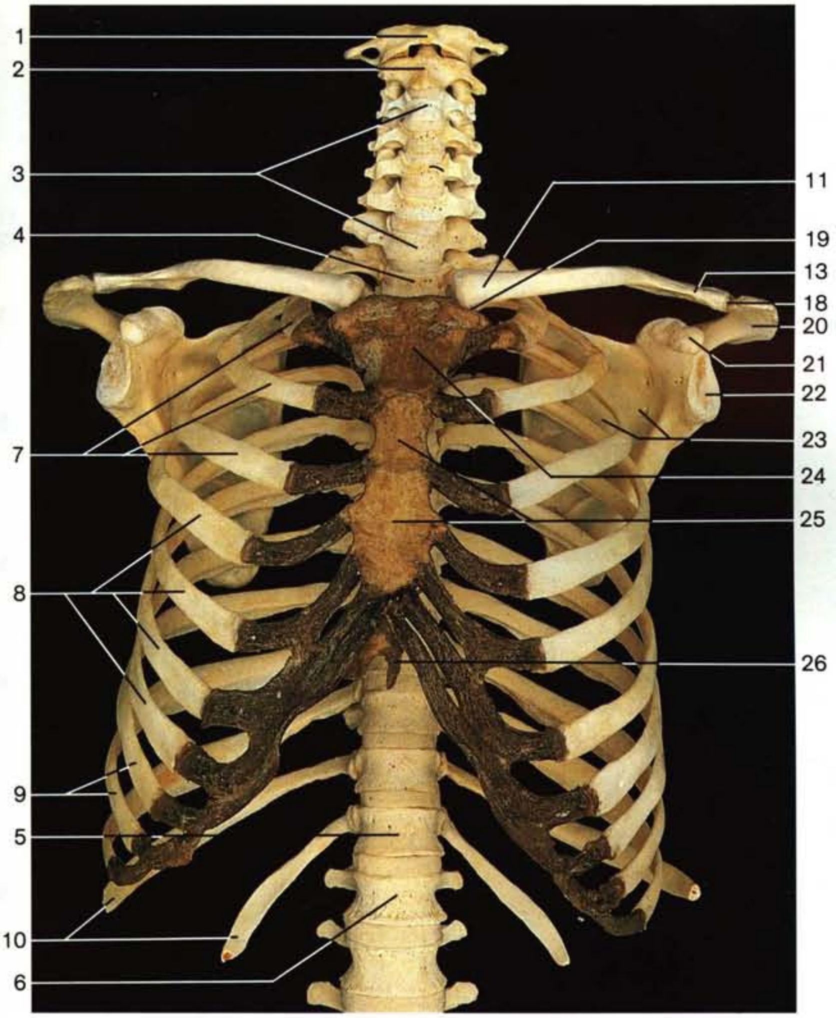

Skeleton of shoulder girdle and thorax (anterior aspect). The cartilaginous parts of the ribs appear dark brown.

Vertebral column

1 Atlas
2 Axis
3 3rd-7th cervical vertebrae
4 1st thoracic vertebra
5 12th thoracic vertebra
6 1st lumbar vertebra

Ribs

7 1st–3rd ribs } True ribs
8 4th–7th ribs
9 8th–10th ribs } False ribs
10 11th and 12th ribs (floating ribs)

Clavicle

11 Sternal end
12 Articular facet for sternum
13 Acromial end
14 Articular facet for acromion
15 Impression for costoclavicular ligament
16 Conoid tubercle
17 Trapezoid line
18 Site of acromioclavicular joint
19 Site of sternoclavicular joint

Scapula

20 Acromion
21 Coracoid process
22 Glenoid cavity
23 Costal surface

Sternum

24 Manubrium
25 Body
26 Xiphoid process

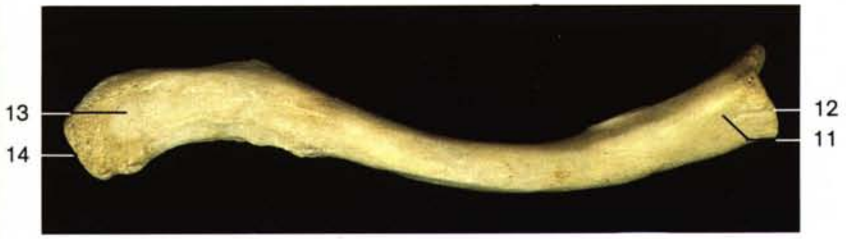

Right clavicle (superior aspect).

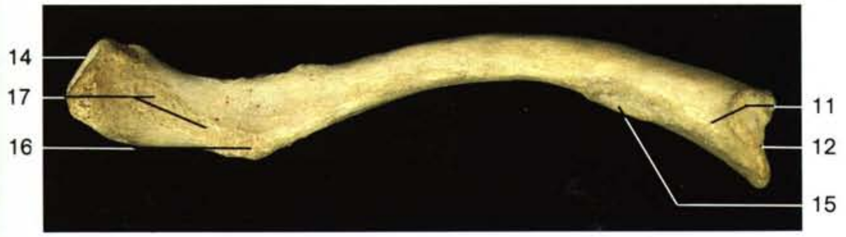

Right clavicle (inferior aspect).

Because of his upright posture, man's upper limb has developed a high degree of mobility. The shoulder girdle is to a great extent movable in the thorax and is connected with the trunk only by the sternoclavicular joint. A further characteristic of man's forearm is the capacity for rotation (i.e., pronation and supination).

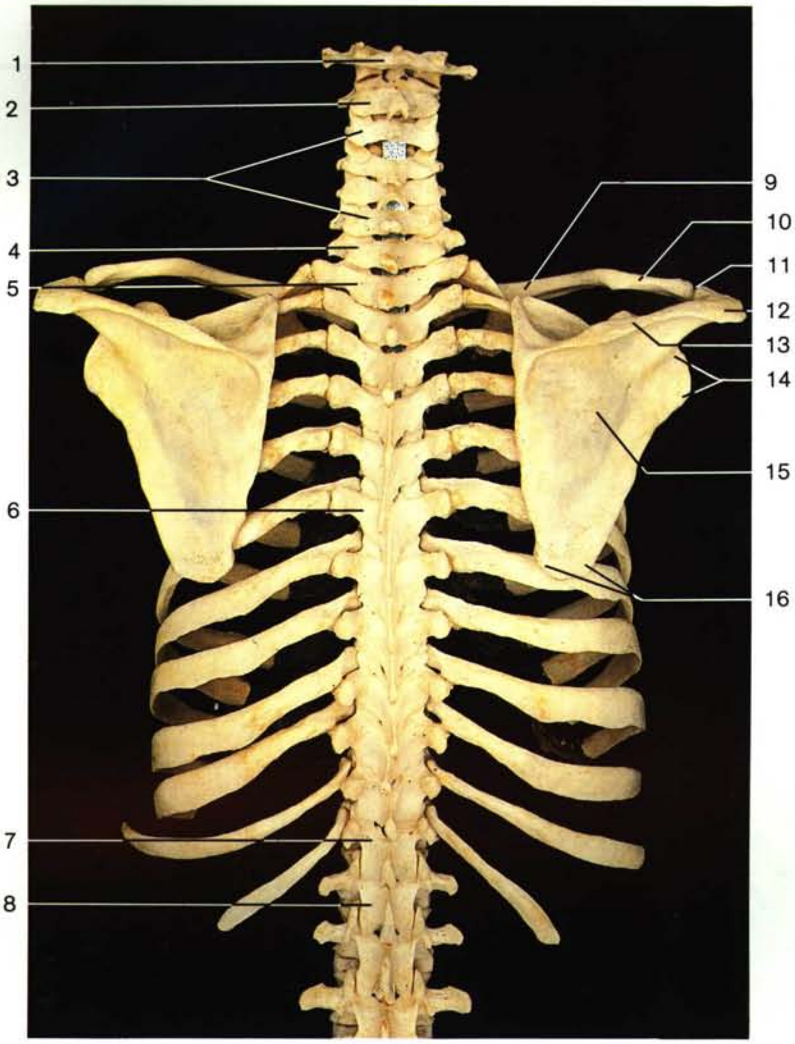

Skeleton of shoulder girdle and thorax (posterior view).

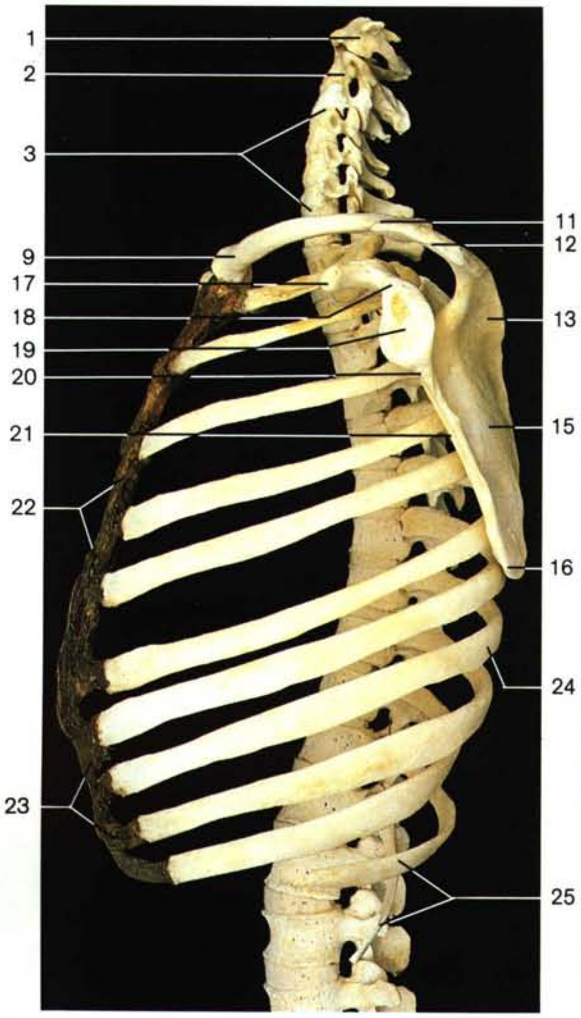

Skeleton of shoulder girdle and thorax (lateral view).

Vertebral column

1 Atlas
2 Axis
3 3rd–6th cervical vertebrae
4 7th vertebra (vertebra prominens)
5 1st thoracic vertebra
6 6th thoracic vertebra
7 12th thoracic vertebra
8 1st lumbar vertebra

Clavicle

9 Sternal end
10 Acromial end
11 Site of acromioclavicular joint

Scapula

12 Acromion
13 Spine of scapula
14 Lateral angle
15 Posterior surface
16 Inferior angle
17 Coracoid process
18 Supraglenoid tubercle
19 Glenoid cavity
20 Infraglenoid tubercle
21 Lateral margin

Thorax

22 Body of sternum
23 Costal arch
24 Angle of ribs
25 Free ribs

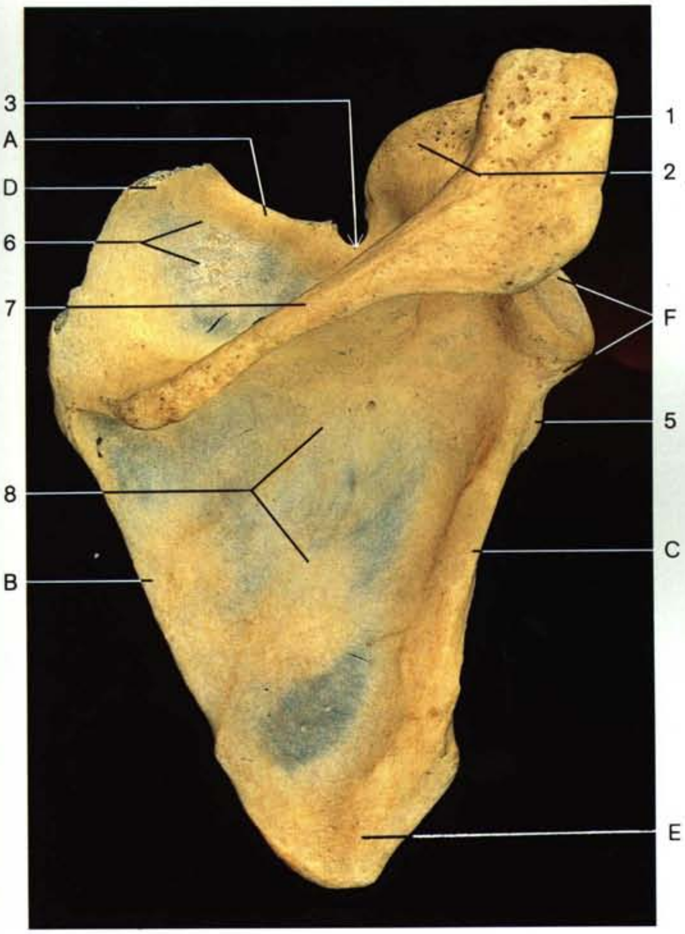

Right scapula (posterior aspect).

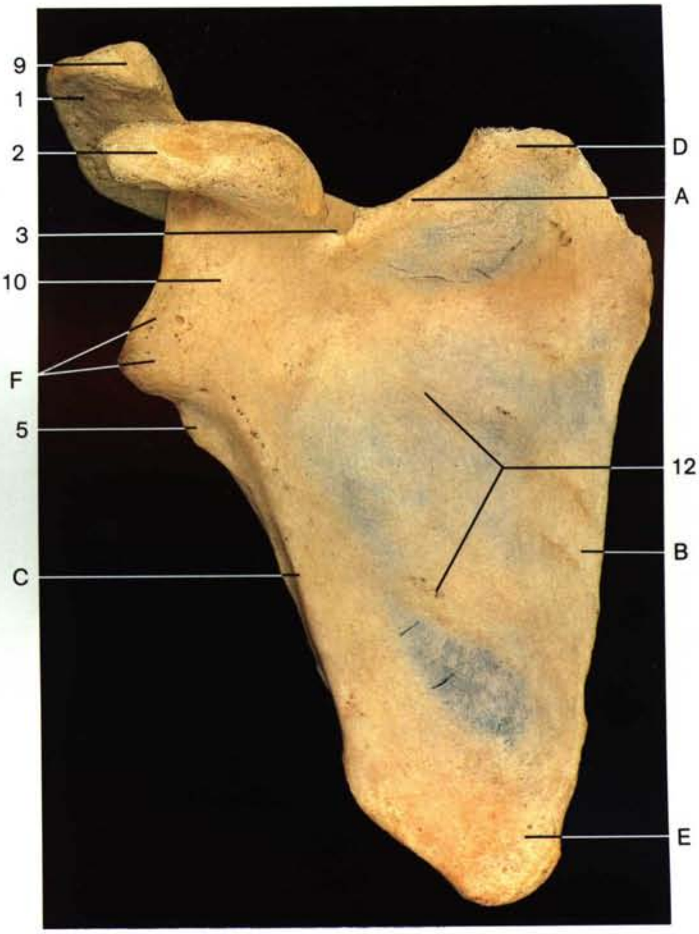

Right scapula (anterior aspect costal surface).

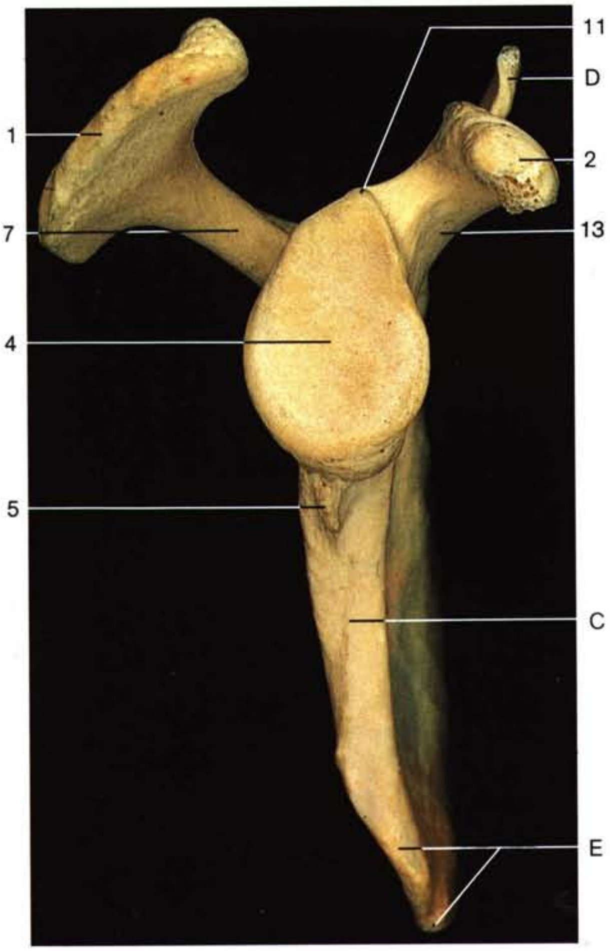

Right scapula (lateral view).

Scapula

A Superior border
B Medial border
C Lateral border
D Superior angle
E Inferior angle
F Lateral angle

1 Acromion
2 Coracoid process
3 Scapular notch
4 Glenoid cavity
5 Infraglenoid tubercle
6 Supraspinous fossa
7 Spine
8 Infraspinous fossa
9 Articular facet for acromion
10 Neck
11 Supraglenoid tubercle
12 Costal (anterior) surface
13 Base of coracoid process

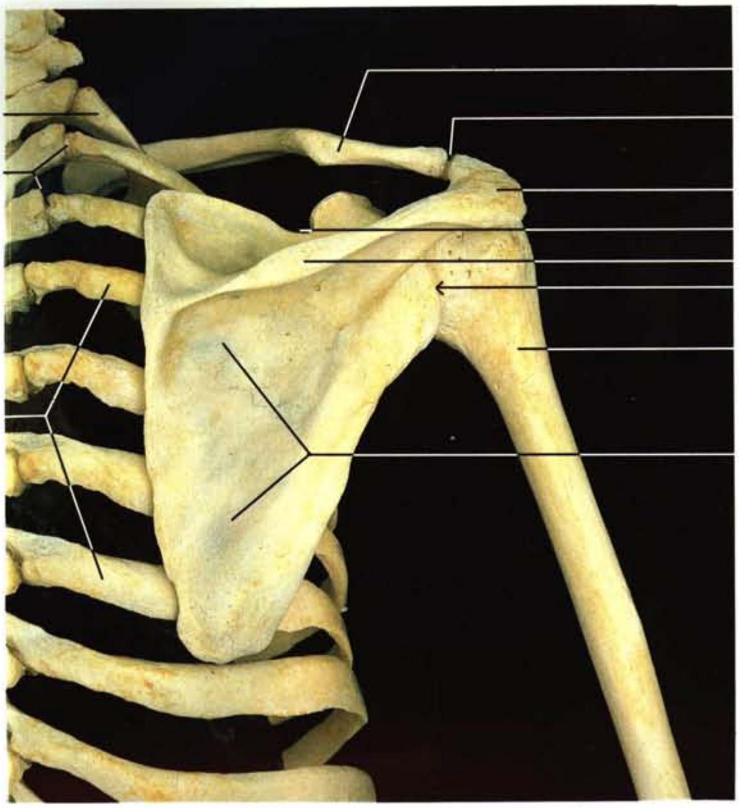

Bones of shoulder joint (posterior aspect).

1 1st rib
2 Position of costotransverse joints
3 4th–7th ribs
4 **Clavicle**
5 Position of acromioclavicular joint
6 Acromion
7 Scapular notch
8 Spine of scapula
9 Head of humerus
10 Glenoid cavity
11 Surgical neck of humerus
12 Posterior surface of **scapula**
13 Coracoid process
14 Infraglenoid tubercle
15 Greater tubercle of **humerus**
16 Anatomical neck of humerus

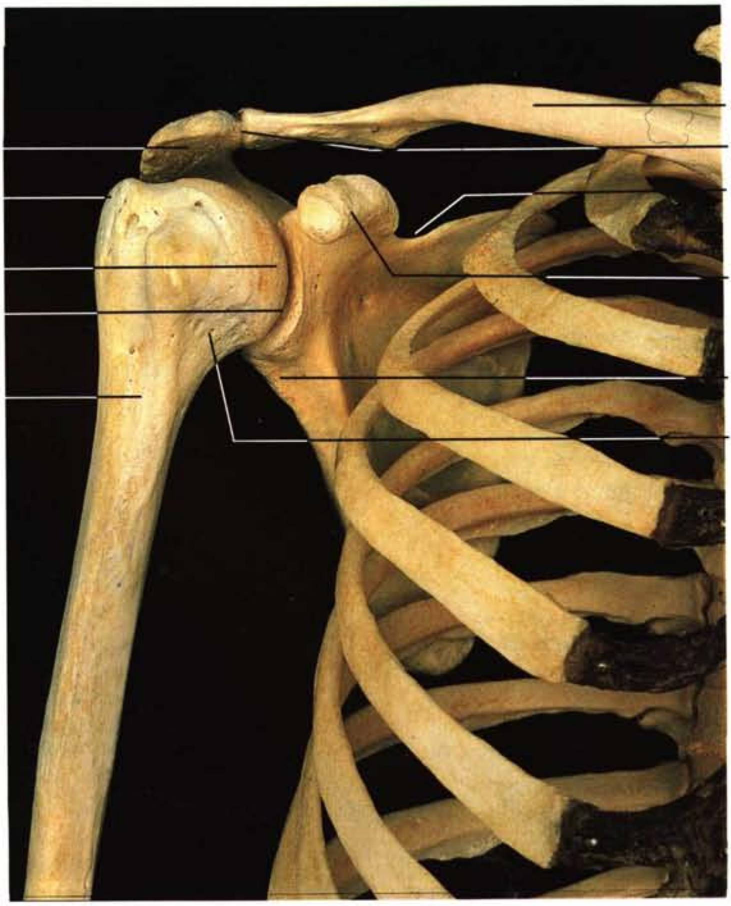

Bones of shoulder joint (anterior aspect).

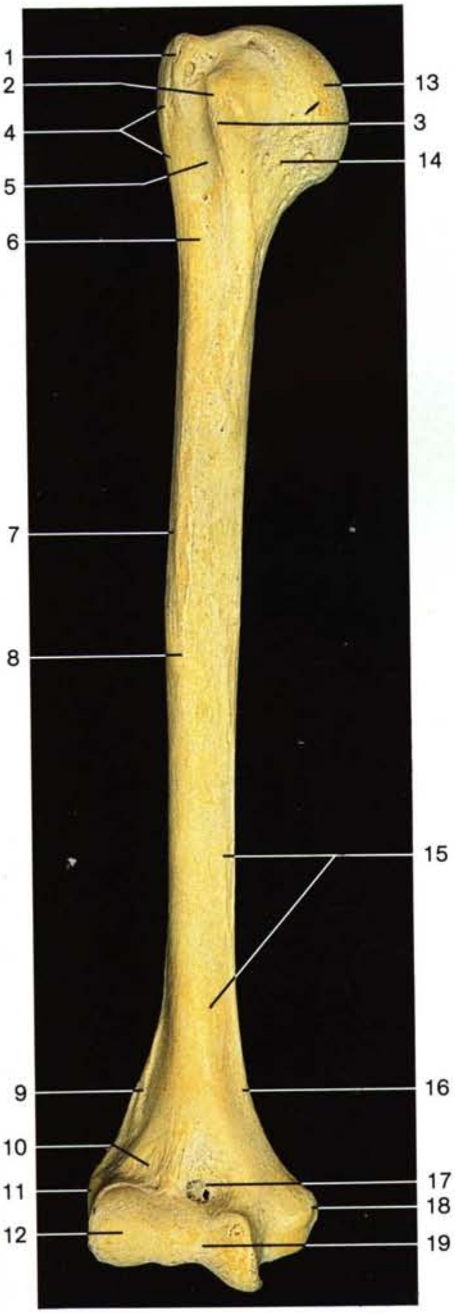
1
2
4
5
6
7
8
9
10
11
12
13
3
14
15
16
17
18
19

Right humerus (anterior aspect).

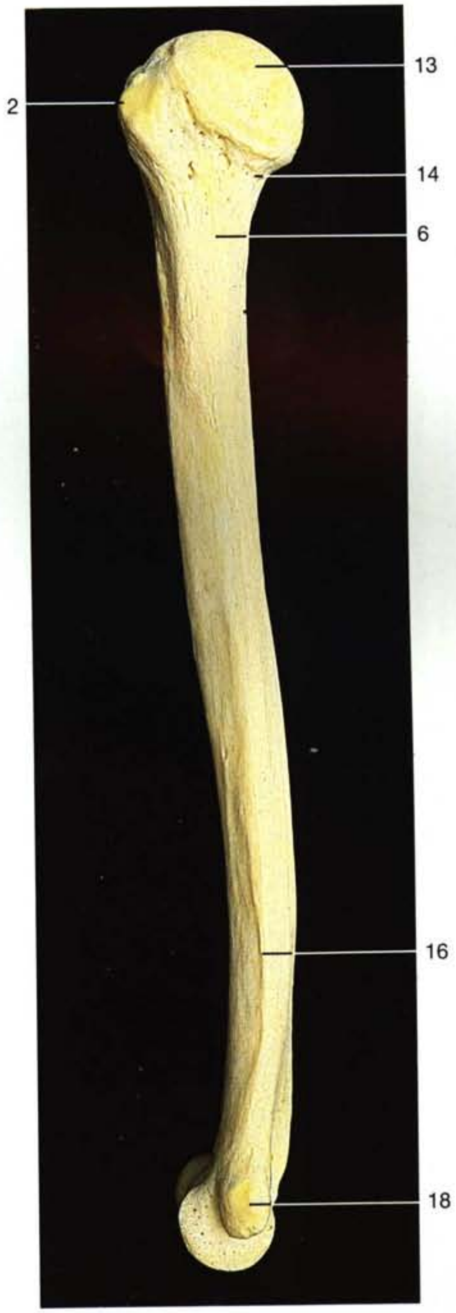
2
13
14
6
16
18

Right humerus (medial aspect).

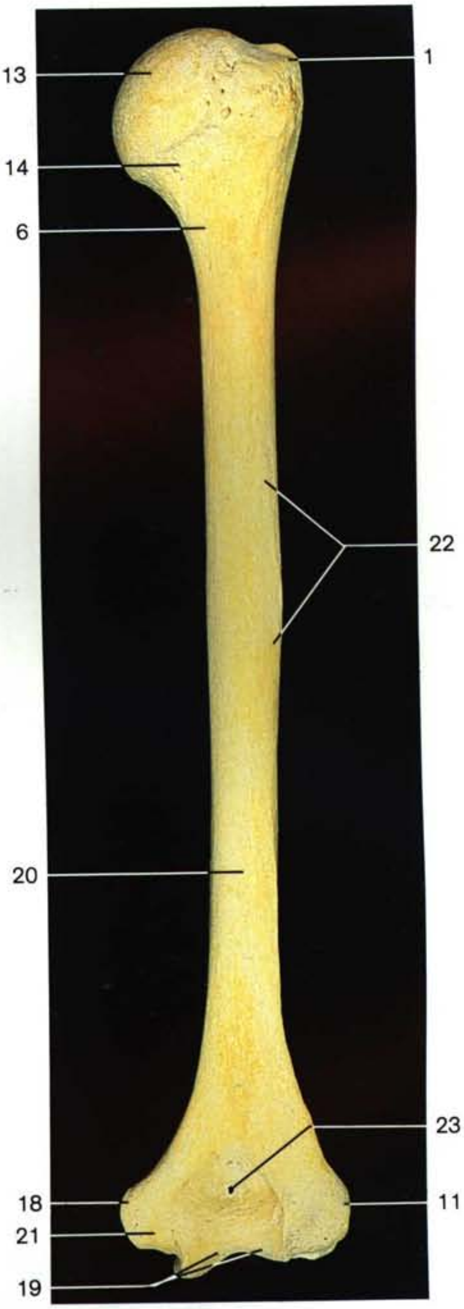
13
14
6
20
18
21
19
1
22
23
11

Right humerus (posterior aspect).

Humerus

1 Greater tubercle
2 Lesser tubercle
3 Crest of lesser tubercle
4 Crest of greater tubercle
5 Intertubercular sulcus
6 Surgical neck
7 Deltoid tuberosity
8 Anterolateral surface
9 Lateral supracondylar ridge
10 Radial fossa
11 Lateral epicondyle
12 Capitulum
13 Head
14 Anatomical neck
15 Anteromedial surface
16 Medial supracondylar ridge
17 Coronoid fossa
18 Medial epicondyle
19 Trochlea
20 Posterior surface
21 Groove for ulnar nerve
22 Groove for radial nerve
23 Olecranon fossa

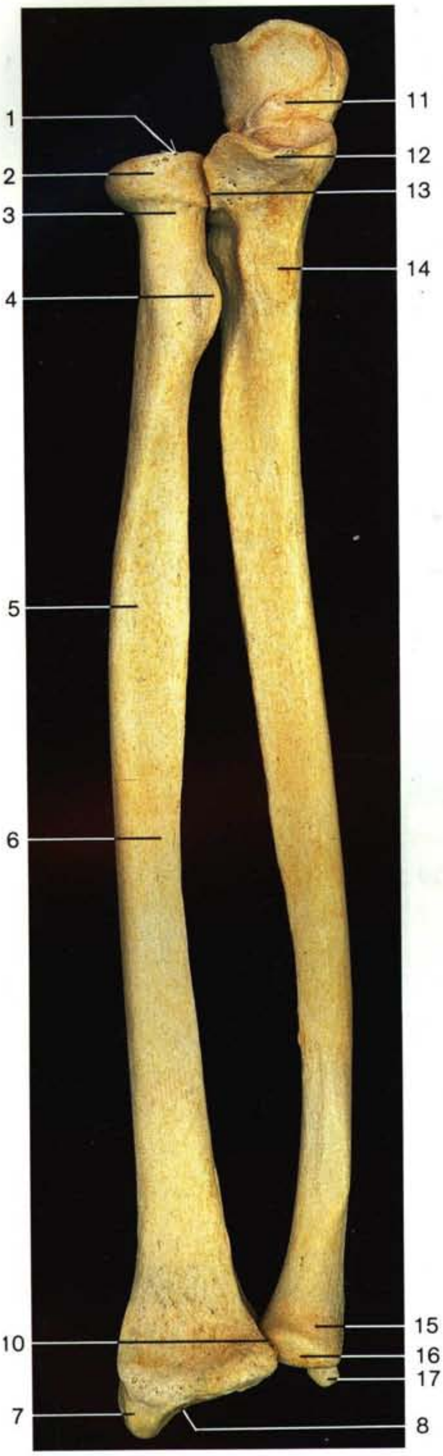

Bones of right forearm, radius and ulna (anterior aspect).

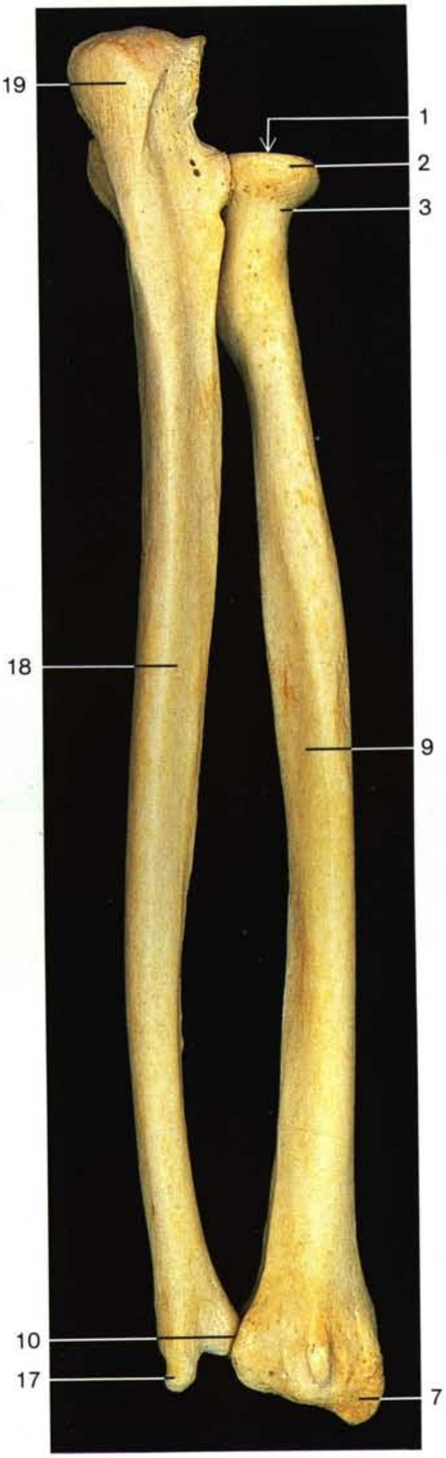

Bones of right forearm, radius and ulna (posterior aspect).

Radius

1 Head
2 Articular circumference
3 Neck
4 Radial tuberosity
5 Shaft
6 Anterior surface
7 Styloid process
8 Articular surface
9 Posterior surface
10 Ulnar notch

Ulna

11 Trochlear notch
12 Coronoid process
13 Radial notch
14 Ulnar tuberosity
15 Head
16 Articular circumference
17 Styloid process
18 Posterior surface
19 Olecranon

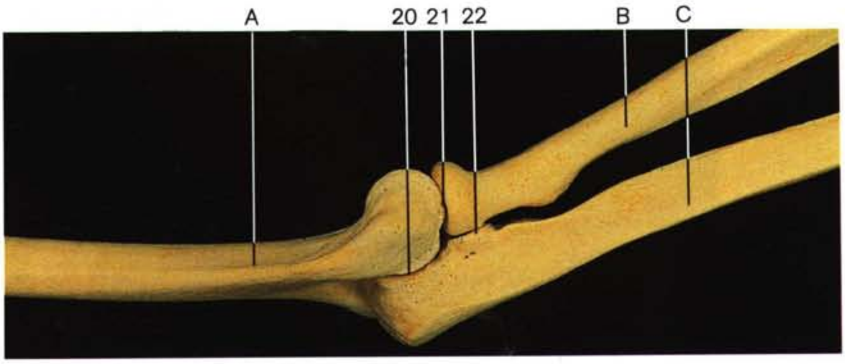

Bones of right elbow joint (lateral aspect).

Articulations at the right elbow

20 Site of humeroulnar joint
21 Site of humeroradial joint
22 Site of proximal radioulnar joint

A Humerus
B Radius
C Ulna

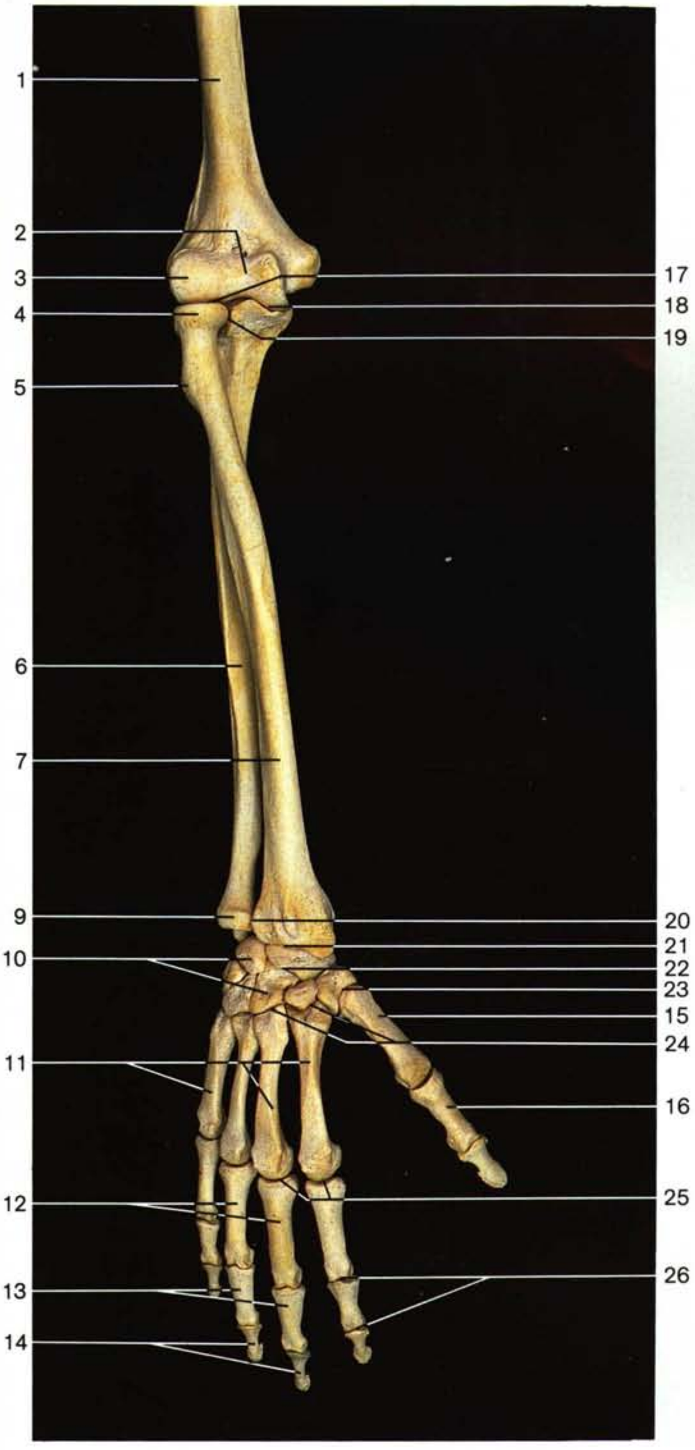
1
2
3
4
5
6
7
9
10
11
12
13
14
17
18
19
20
21
22
23
15
24
16
25
26

Skeleton of right forearm and hand in pronation.

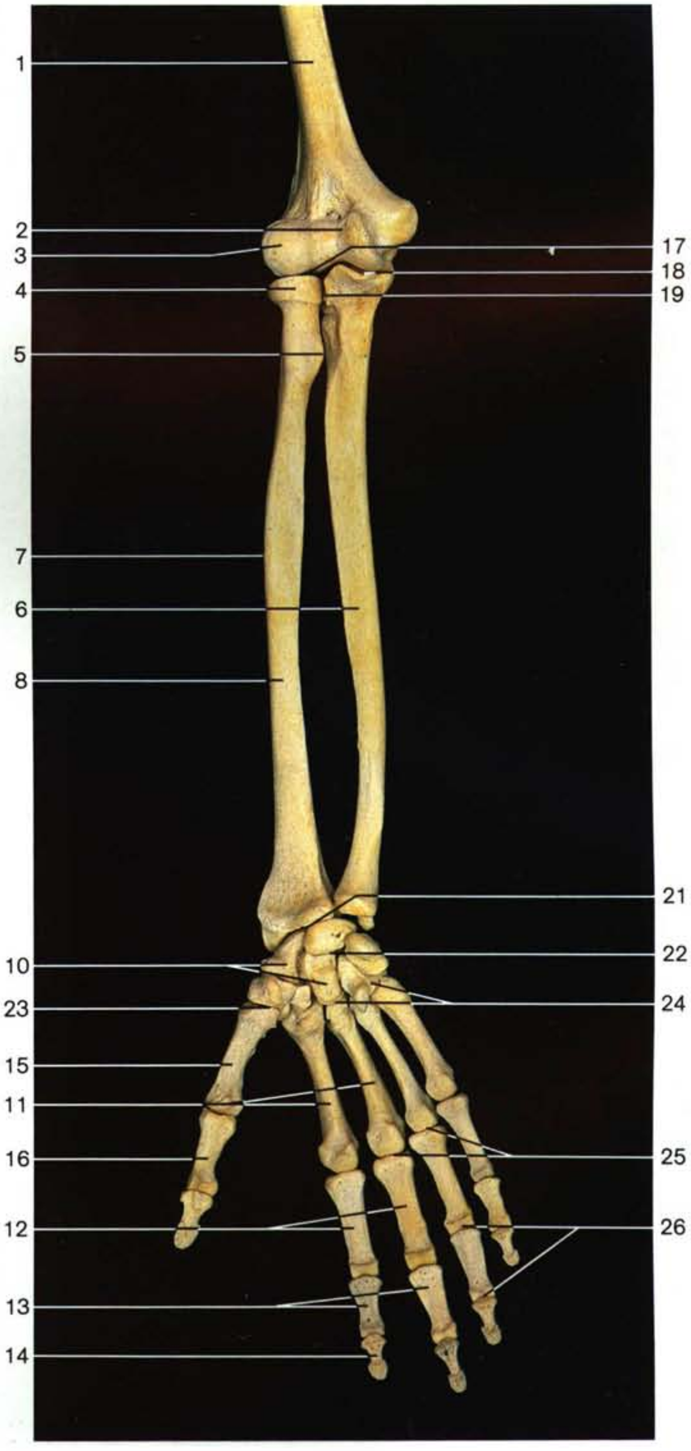
1
2
3
4
5
7
6
8
10
23
15
11
16
12
13
14
17
18
19
21
22
24
25
26

Skeleton of right forearm and hand in supination.

1 Humerus
2 Trochlea of humerus
3 Capitulum of humerus
4 Articular circumference of radius
5 Radial tuberosity
6 Anterior surface of **ulna**
7 Posterior surface of **radius**
8 Anterior surface of radius
9 Articular circumference of ulna
10 **Carpal bones**

11 **Metacarpal bones**
12 Proximal phalanges
13 Middle phalanges
14 Distal phalanges
15 Metacarpal bone of thumb
16 Proximal phalanx of thumb

Sites of joints
17 Humeroradial joint
18 Humeroulnar joint
19 Proximal radioulnar joint
20 Distal radioulnar joint
21 Radiocarpal joint
22 Midcarpal joint
23 Carpometacarpal joint of thumb
24 Carpometacarpal joints
25 Metacarpophalangeal joints
26 Interphalangeal joints of fingers

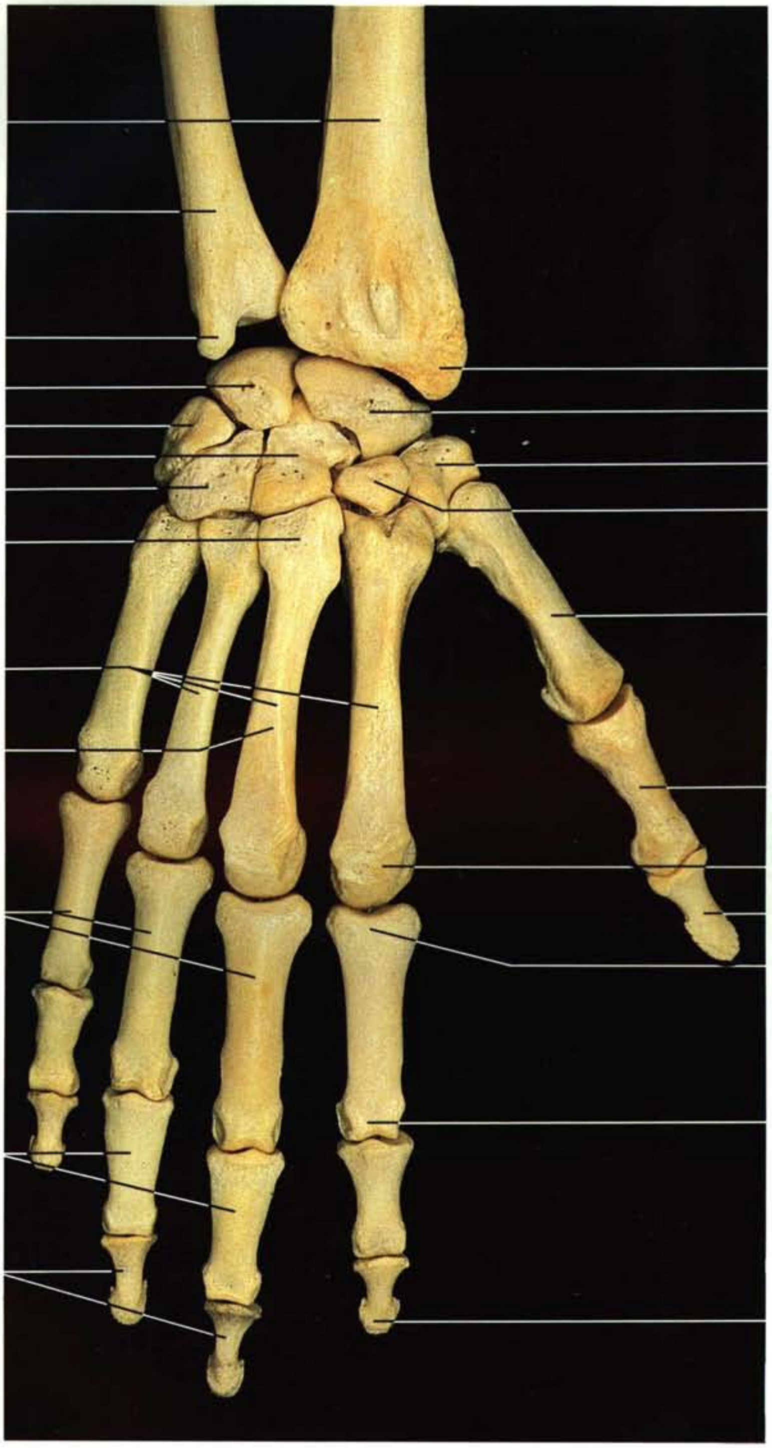
1
2
3
4
5
6
7
8
9
24
11
12
13
14
15
16
17
18
19
10
20
21
22
23

Skeleton of right wrist and hand (dorsal aspect).

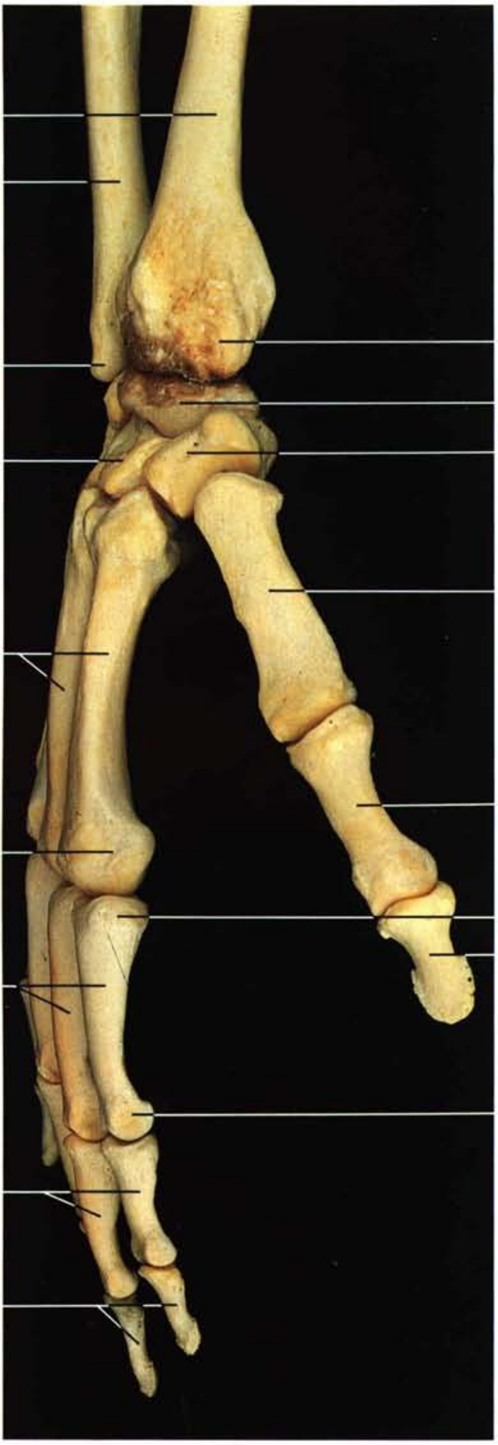
1
2
3
6
9
10
11
12
13
14
15
16
18
19
21
20
22

Skeleton of right wrist and hand (medial aspect).

1 **Radius**
2 **Ulna**
3 Styloid process of ulna
4 Lunate bone
5 Triquetral bone
6 Capitate bone
7 Hamate bone
(4–7) **Carpal bones**

8 Base of third metacarpal bone
9 **Metacarpal bones**
10 Head of metacarpal bone
11 Proximal phalanges of hand
12 Middle phalanges of hand
13 Distal phalanges of hand
14 Styloid process of radius

15 Scaphoid bone
16 The trapezium
17 Trapezoid bone
(15–17) **Carpal bones**
18 Metacarpal bone of thumb
19 Proximal phalanx of thumb
20 Distal phalanx of thumb
21 Base of second proximal phalanx

22 Head of second proximal phalanx
23 Tuberosity of distal phalanx
24 Body of third metacarpal bone

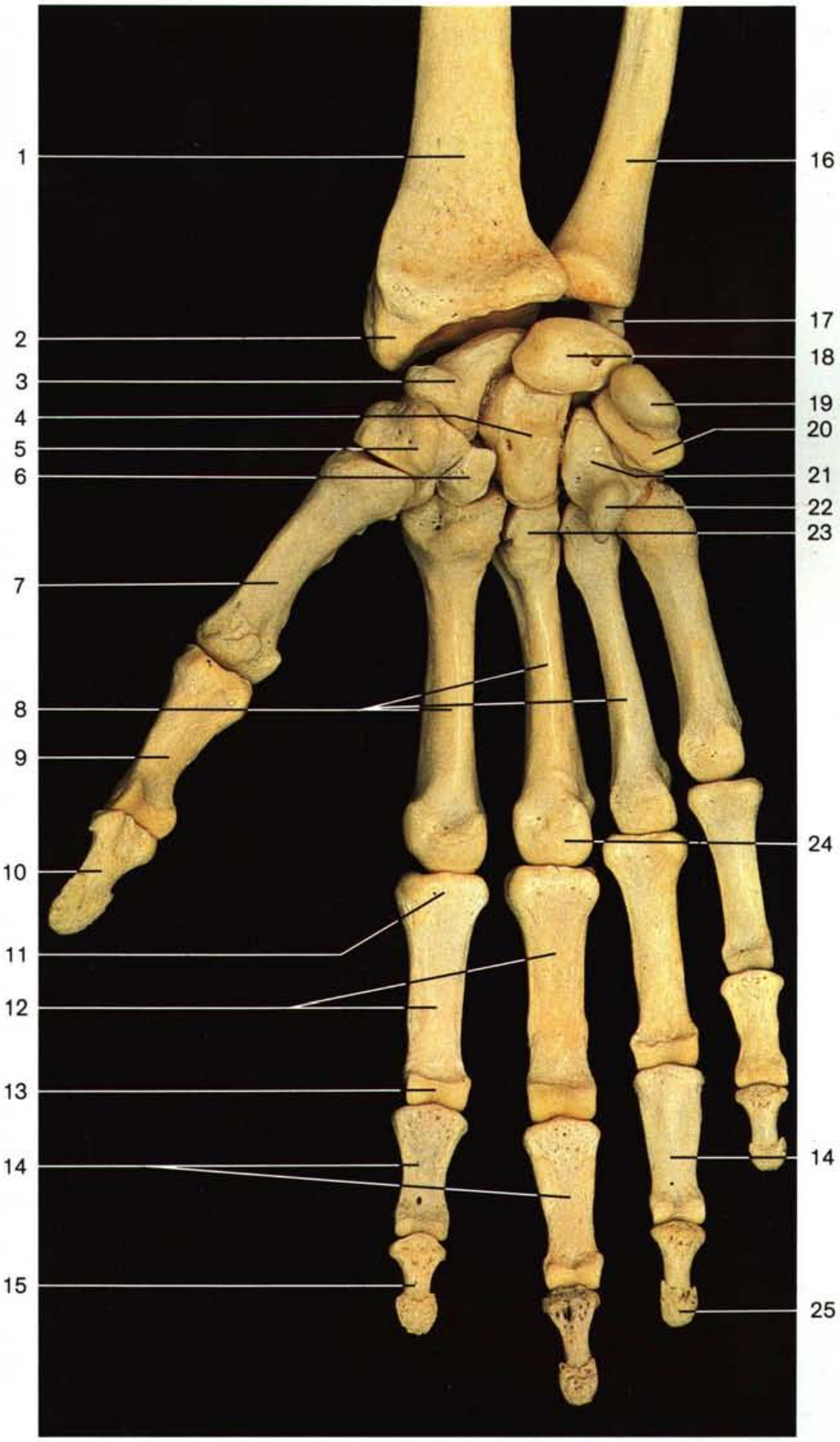

1 Radius
2 Styloid process of radius
3 Scaphoid bone
4 Capitate bone
5 Trapezium
6 Trapezoid bone
Carpal bones (3–6)
7 First metacarpal bone
8 Second to fourth metacarpal bones
9 Proximal phalanx of thumb
10 Distal phalanx of thumb
11 Base of second proximal phalanx
12 Proximal phalanges
13 Head of second proximal phalanx
14 Middle phalanges
15 Distal phalanx
16 Ulna
17 Styloid process of ulna
18 Lunate bone
19 Pisiform bone
20 Triquetral bone
21 Hamate bone
22 Hamulus or hook of hamate bone
Carpal bones (18–22)
23 Base of third metacarpal bone
24 Head of metacarpal bone
25 Tuberosity of distal phalanx

Skeleton of right wrist and hand (palmar aspect).

The human hand is one of the most admirable structures of the human body. The carpometacarpal joint of the thumb, a saddle joint, enjoys wide mobility, so that the thumb can get in contact with all other fingers, thus enabling the hand to become an instrument for grasping and psychologic expression. During evolution these newly developed functions appeared after the erect posture of the human body was achieved. An inevitable prerequisite for the development of human cultures is not only the differentiation of the brain but also the development of an organ capable of realizing its ideas: the human hand.

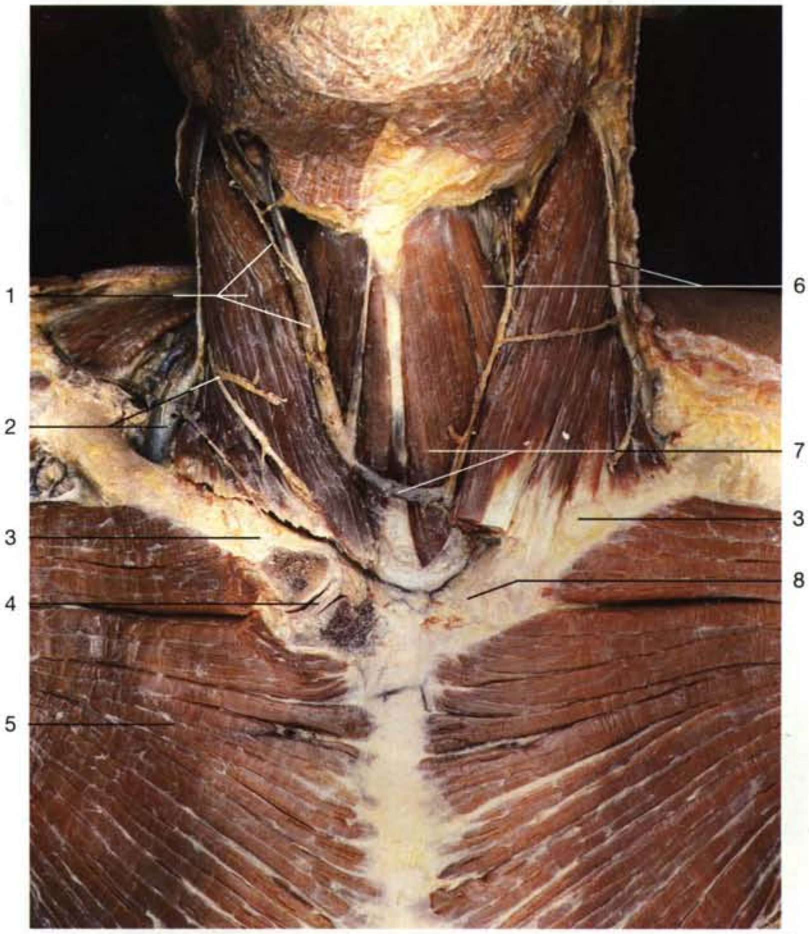

Right sternoclavicular joint (anterior aspect). On the right side the joint has been opened by a coronal section. Note the articular disk.

1 Sternocleidomastoid muscle, cervical branch of facial nerve and anterior jugular vein
2 External jugular vein and transverse cervical nerve (inferior branch)
3 Clavicle
4 **Sternoclavicular joint** (opened) with articular disk
5 Pectoralis major muscle
6 Omohyoid muscle and external jugular vein
7 Jugular venous arch and sternohyoid muscle
8 Sternoclavicular joint (not opened)
9 Acromial end of clavicle
10 **Acromioclavicular joint**
11 Acromion
12 Tendon of supraspinatus muscle (attached to the articular capsule)
13 Coraco-acromial ligament
14 Tendon of long head of biceps brachii muscle
15 Tendon of subscapularis muscle (attached to the articular capsule)
16 Intertubercular sulcus
17 Articular capsule of shoulder joint
18 Humerus
19 Trapezoid ligament
20 Coracoid process
21 Glenoid labrum
22 **Shoulder joint** (joint cavity)
23 Scapula
24 Supraspinatus muscle
25 Cartilage of glenoid cavity
26 Tendon of long head of triceps brachii muscle
27 Head of humerus (articular cartilage)
28 Epiphyseal line

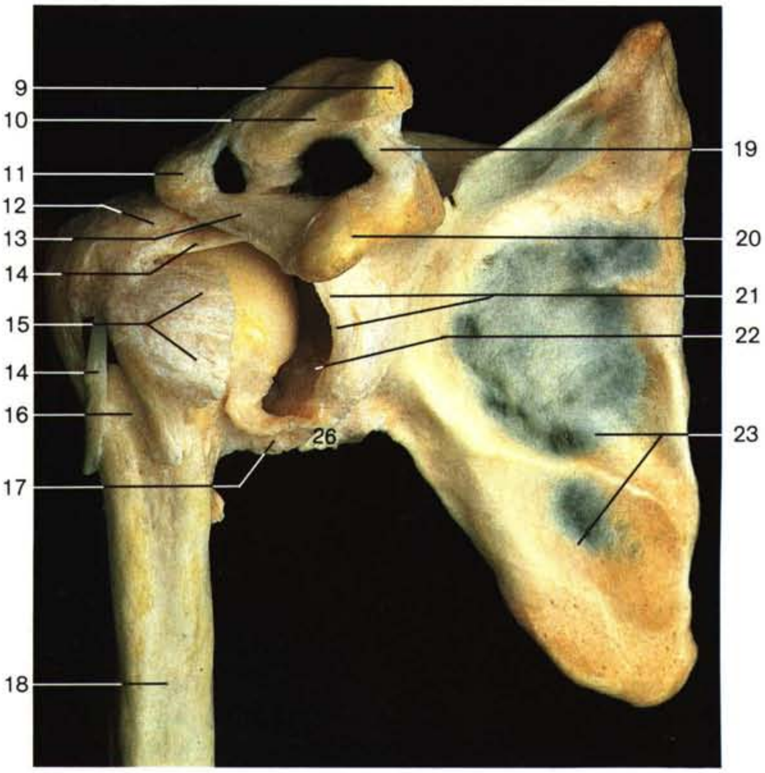

Right shoulder joint. The anterior part of the articular capsule has been removed and the head of the humerus has been slightly rotated outward to show the cavity of the joint.

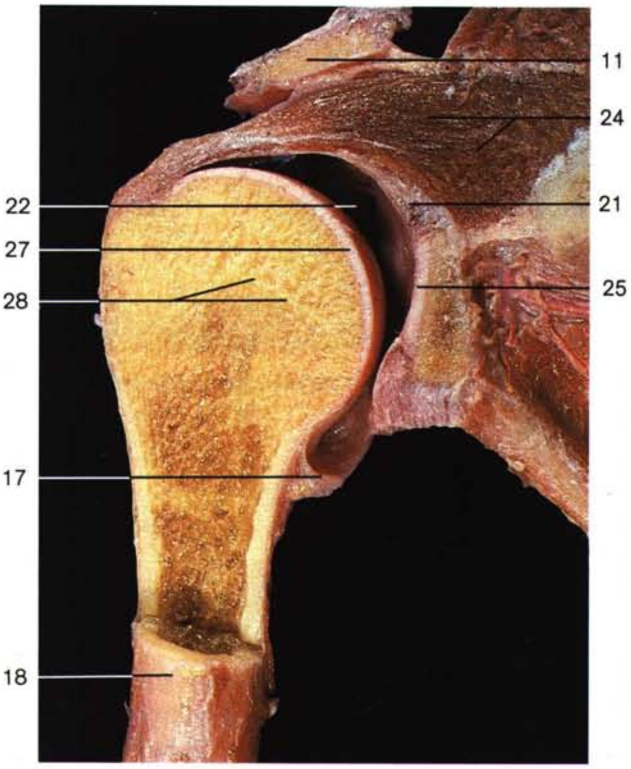

Coronal section of the right shoulder joint (anterior aspect).

1 Humerus
2 Lateral epicondyle of humerus
3 Articular capsule
4 Annular ligament of proximal radioulnar joint
5 Radius
6 Tendon of biceps brachii muscle
7 Medial epicondyle of humerus
8 Ulnar collateral ligament
9 Oblique cord
10 Ulna
11 Interosseous membrane
12 Radial fossa
13 Capitulum of humerus
14 Head of radius
15 Radial collateral ligament
16 Coronoid fossa
17 Trochlea of humerus
18 Coronoid process of ulna
19 Olecranon
20 Radial tuberosity

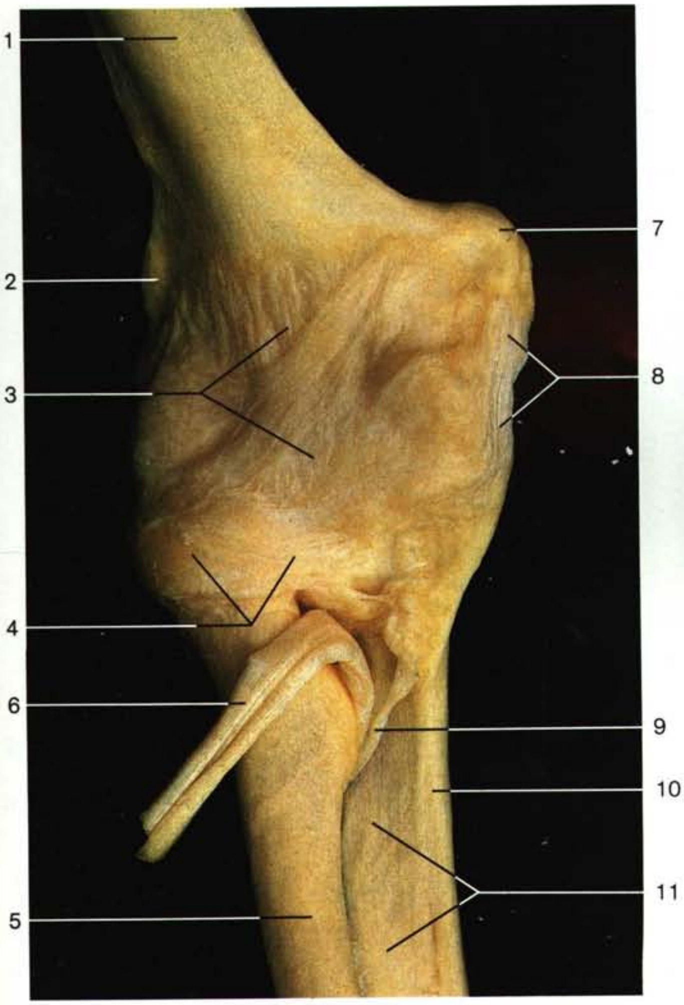

Ligaments of elbow joint (anterior aspect).

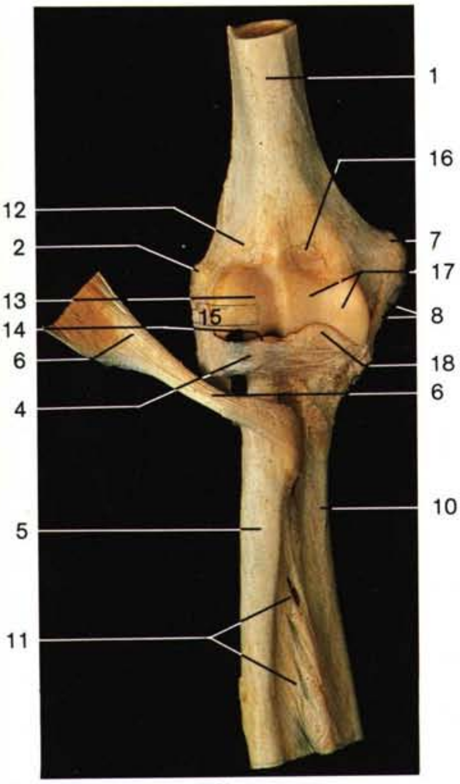

Elbow joint with ligaments (anterior aspect). Articular capsule has been removed to show the annular ligament.

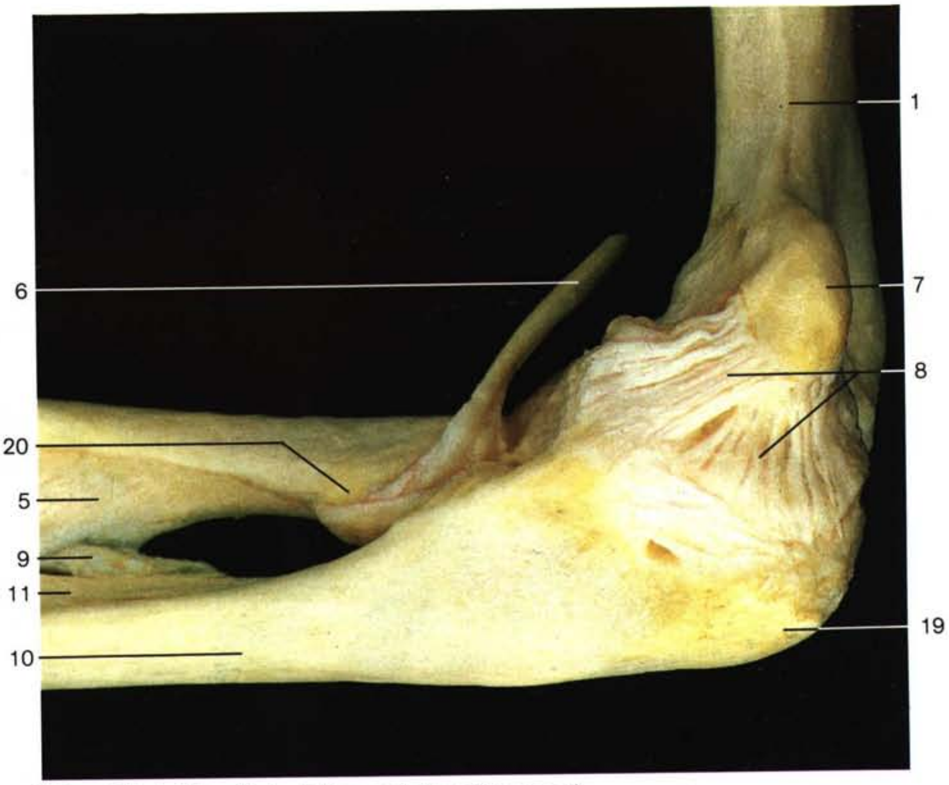

Elbow joint with collateral ligaments (medial aspect).

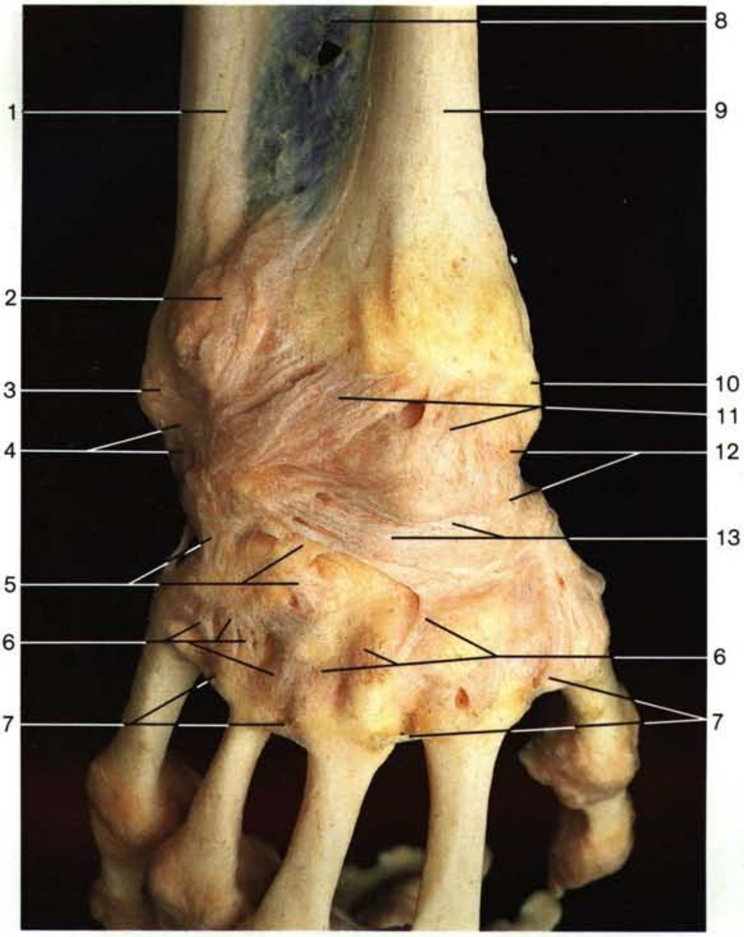

Ligaments of hand and wrist (dorsal aspect).

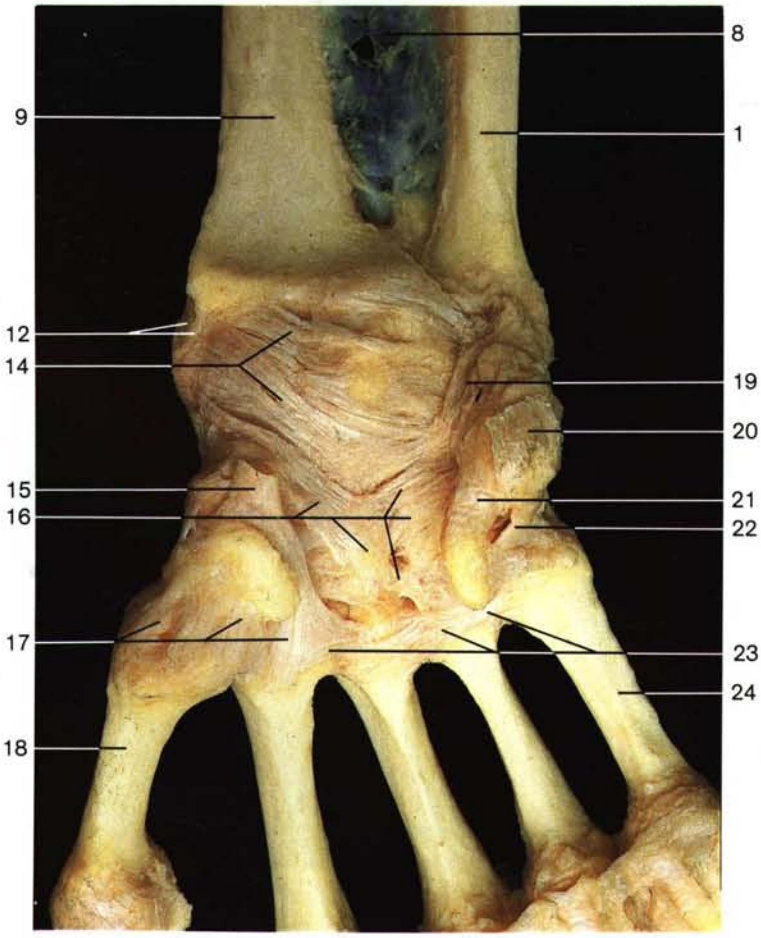

Ligaments of hand and wrist (palmar aspect).

1 Ulna
2 Exostosis (pathological)
3 Head of ulna
4 **Ulnar carpal collateral ligament**
5 Deep intercarpal ligaments
6 Dorsal carpometacarpal ligaments
7 Dorsal metacarpal ligaments
8 Interosseous membrane
9 Radius
10 Styloid process of radius
11 **Dorsal radiocarpal ligament**
12 Radial collateral ligament
13 Articular capsule and dorsal intercarpal ligaments
14 **Palmar radiocarpal ligament**
15 Tendon of flexor carpi radialis muscle (cut)
16 Radiating carpal ligament
17 Palmar carpometacarpal ligaments
18 1st metacarpal bone
19 Palmar ulnocarpal ligament
20 Tendon of flexor carpi ulnaris muscle (cut)
21 **Pisohamate ligament**
22 **Pisometacarpal ligament**
23 Palmar metacarpal ligaments
24 5th metacarpal bone

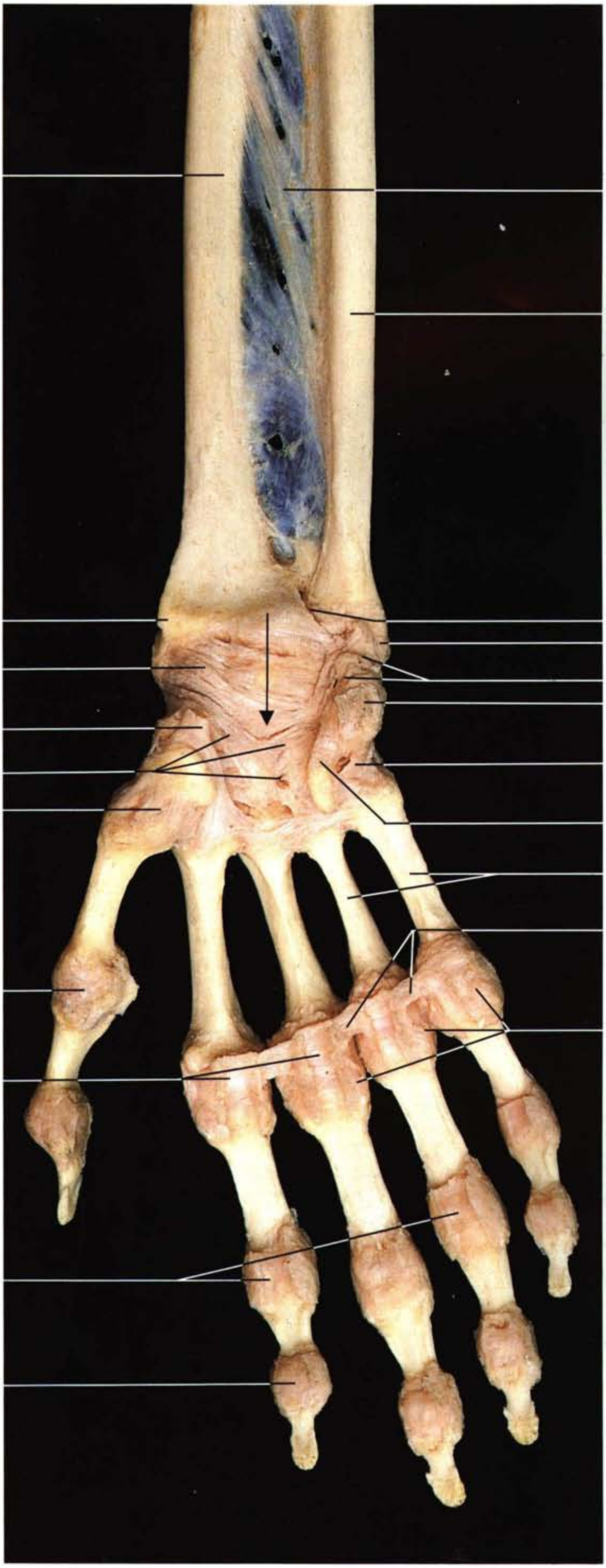

Ligaments of right forearm, hand and fingers (palmar aspect). The arrow indicates the location of the carpal tunnel.

1 Radius
2 Styloid process of radius
3 Palmar radiocarpal ligament
4 Tendon of flexor carpi radialis muscle (cut)
5 Radiating carpal ligament
6 Articular capsule of **carpometacarpal joint of thumb**
7 Articular capsule of **metacarpophalangeal joint of thumb**
8 Palmar ligaments and articular capsule of metacarpophalangeal joints
9 Palmar ligaments and articular capsule of interphalangeal joints
10 Articular capsule
11 Interosseous membrane
12 Ulna
13 **Distal radioulnar joint**
14 Styloid process of ulna
15 Palmar ulnocarpal ligament
16 **Pisiform bone with tendon of flexor carpi ulnaris muscle**
17 Pisometacarpal ligament
18 Pisohamate ligament
19 Metacarpal bone
20 Deep transverse metacarpal ligament
21 Tendons of extensor muscles and articular capsule
22 Collateral ligament of **interphalangeal joint**
23 Collateral ligaments of **metacarpophalangeal joints**
24 Second metacarpal bone

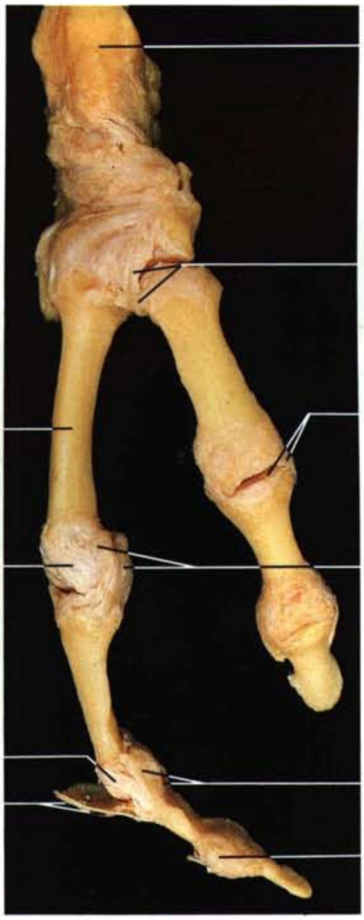

Ligaments of fingers (lateral aspect).

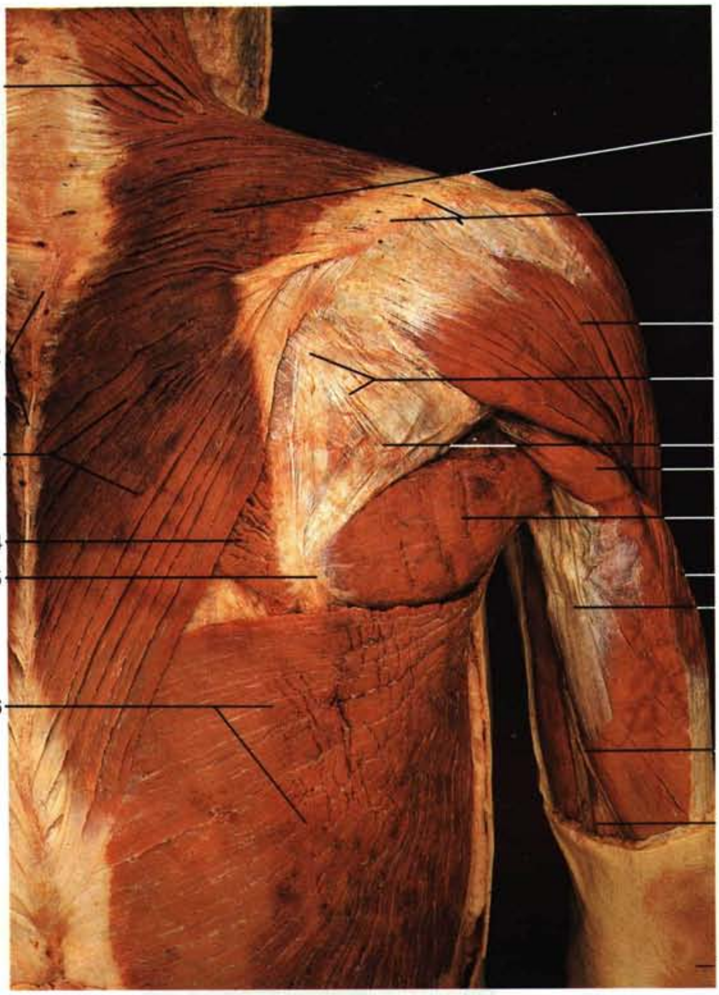

1 Descending fibers of **trapezius muscle**
2 Spinous processes of thoracic vertebrae
3 Ascending fibers of trapezius muscle
4 Rhomboid major muscle
5 Inferior angle of scapula
6 **Latissimus dorsi muscle**
7 Transverse fibers of trapezius muscle
8 Spine of scapula
9 Posterior fibers of **deltoid muscle**
10 Infraspinatus muscle and infraspinous fascia
11 Teres minor muscle and fascia
12 Long head of triceps brachii muscle
13 **Teres major muscle**
14 Lateral head of triceps brachii muscle
15 Medial head of triceps brachii muscle
16 Medial intermuscular septum
17 Ulnar nerve
18 Olecranon

Muscles of shoulder and arm, superficial layer (dorsal aspect).

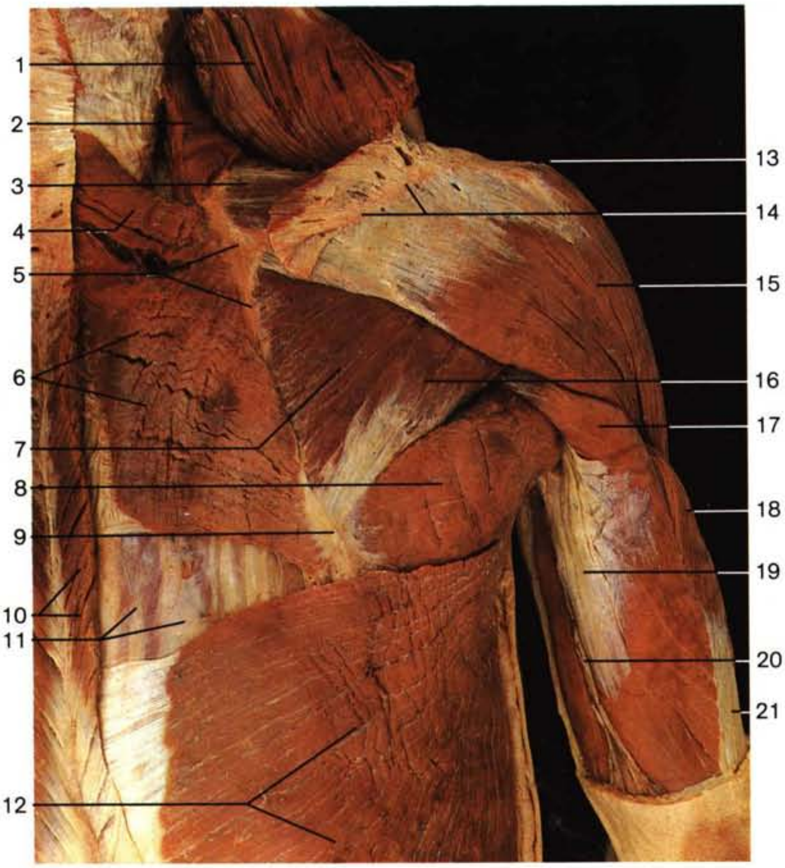

1 Trapezius muscle (reflected)
2 Levator scapulae muscle
3 Supraspinatus muscle
4 Rhomboid minor muscle
5 Medial border of scapula
6 **Rhomboid major muscle**
7 **Infraspinatus muscle**
8 **Teres major muscle**
9 Inferior angle of scapula
10 Cut edge of trapezius muscle
11 Intrinsic muscles of back with fascia
12 **Latissimus dorsi muscle**
13 Acromion
14 Spine of scapula
15 **Deltoid muscle**
16 Teres minor
17 Long head of triceps brachii muscle
18 Lateral head of triceps brachii muscle
19 Medial head of triceps brachii muscle
20 Medial intermuscular septum
21 Tendon of triceps brachii muscle

Muscles of shoulder and arm, deeper layer (right side, dorsal aspect). The trapezius has been cut near its origin at the vertebral column and reflected upward.

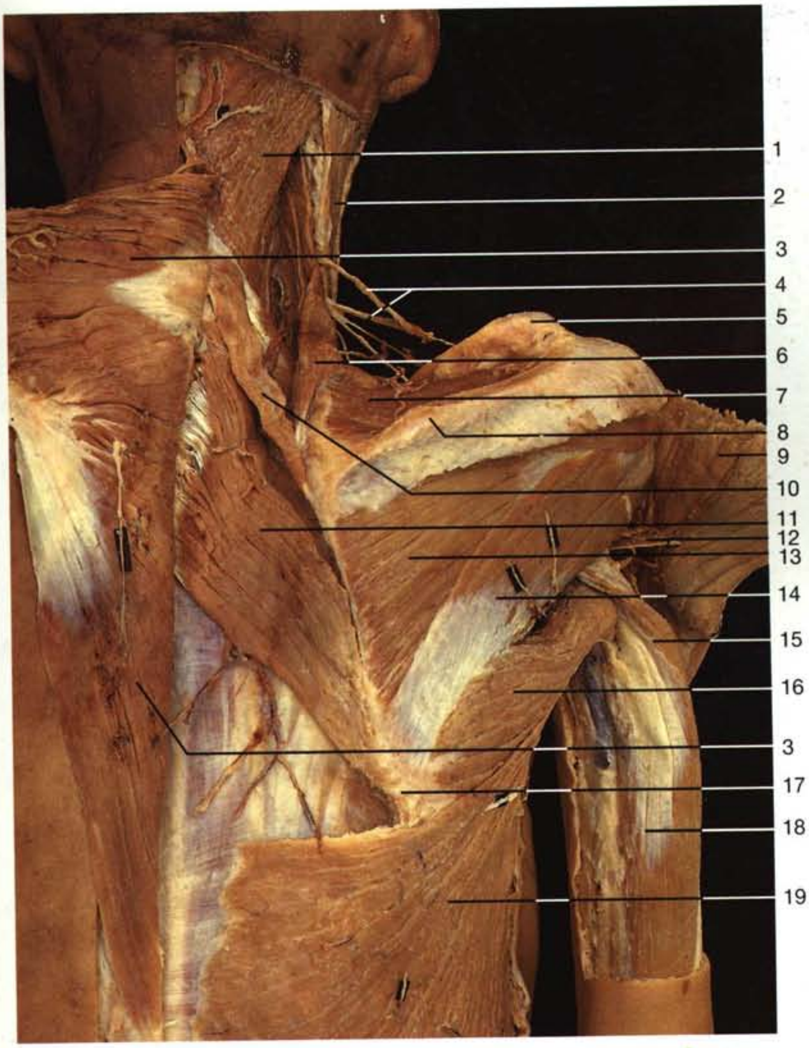

1 **Splenius capitis muscle**
2 Sternocleidomastoideus muscle
3 Trapezius muscle (reflected)
4 **Lateral supraclavicular nerves**
5 Clavicle
6 Levator scapulae muscle
7 Supraspinatus muscle
8 Spine of scapula
9 Deltoid muscle (reflected)
10 Rhomboid minor muscle
11 **Rhomboid major muscle**
12 **Axillary nerve and posterior circumflex humeral artery**
13 **Infraspinatus muscle**
14 **Teres minor muscle**
15 Long head of triceps brachii muscle
16 **Teres major muscle**
17 Inferior angle of scapula
18 Triceps brachii muscle
19 **Latissimus dorsi muscle**

Muscles of shoulder and arm, deeper layer, right side (dorsal aspect). The trapezius and deltoid muscles have been divided and reflected.

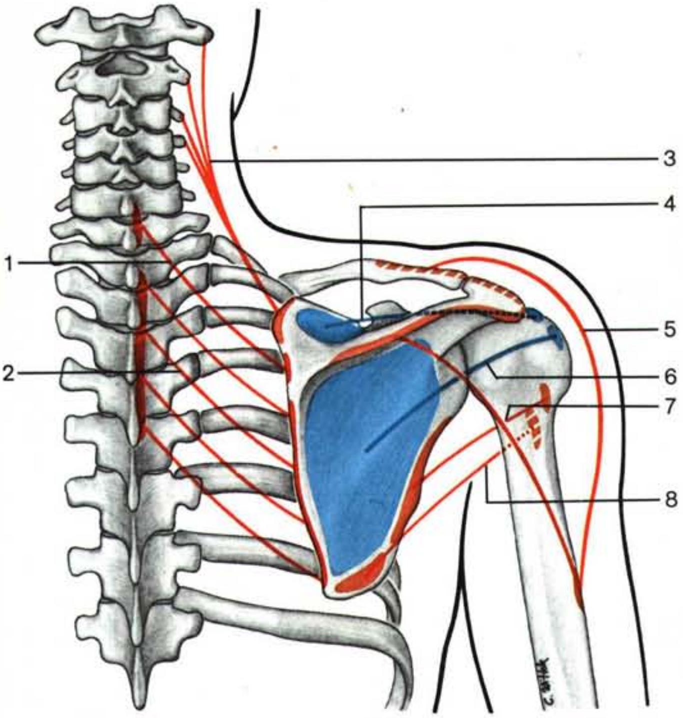

Shoulder muscles, schematic diagram illustrating the course of the main muscles of the dorsal aspect of the shoulder.

1 Rhomboid minor muscle (red)
2 Rhomboid major muscle (red)
3 Levator scapulae muscle (red)
4 Supraspinatus muscle (blue)
5 Deltoid muscle (red)
6 Infraspinatus muscle (blue)
7 Teres minor muscle (red)
8 Teres major muscle (red)

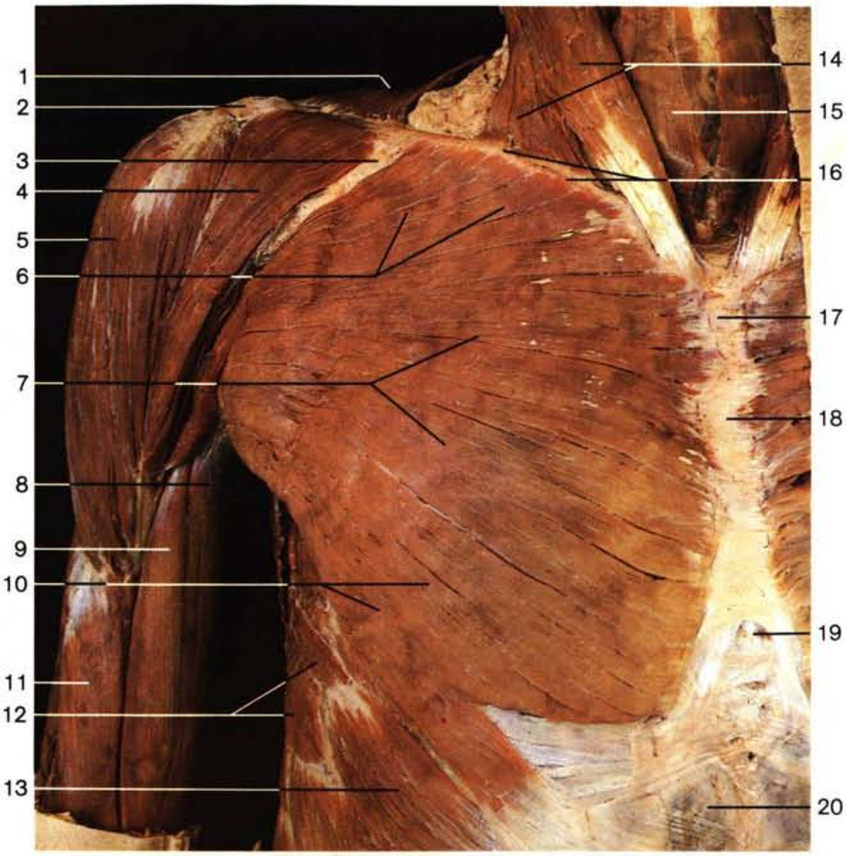

1 Trapezius muscle
2 Acromion
3 Deltopectoral triangle
4 Clavicular part of deltoid muscle (anterior fibers)
5 Acromial part of deltoid muscle (central fibers)
6 Clavicular part of pectoralis major muscle
7 Sternocostal part of pectoralis major muscle
8 Short head of biceps brachii muscle
9 Long head of biceps brachii muscle
10 Abdominal part of pectoralis major muscle
11 Brachialis muscle
12 Serratus anterior muscle
13 External abdominal oblique muscle
14 Sternocleidomastoid muscle
15 Infrahyoid muscles
16 Clavicle
17 Manubrium sterni
18 Body of sternum
19 Xiphoid process
20 Anterior layer of sheath of rectus abdominis muscle

Muscles of shoulder and arm, superficial layer (ventral aspect).

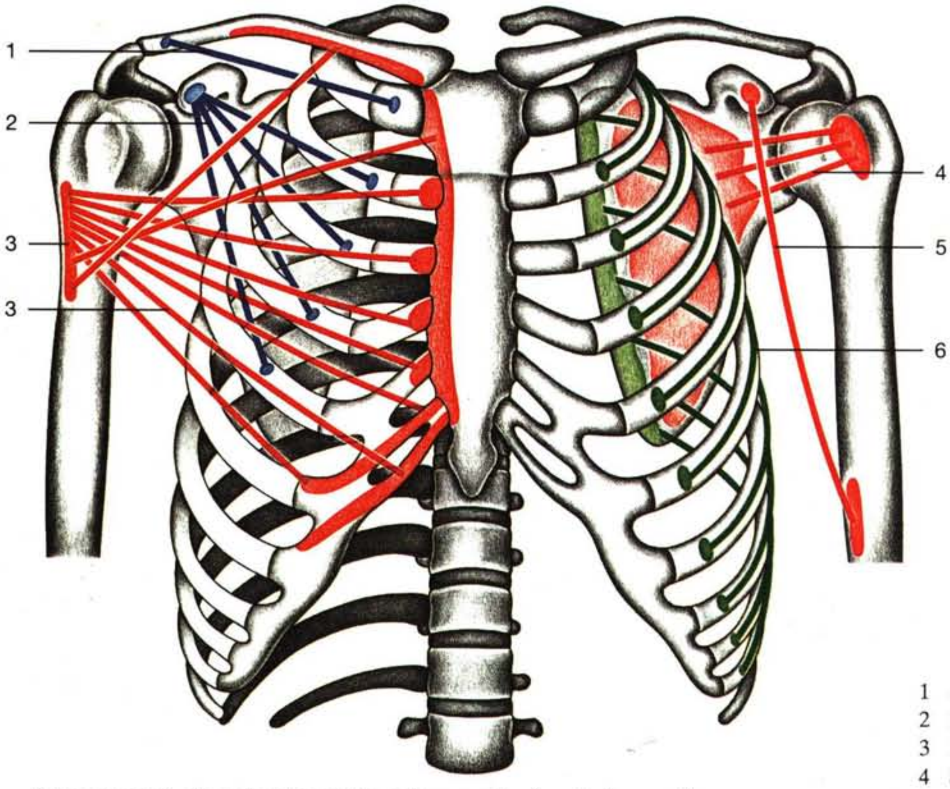

Arrangement of pectoral and shoulder muscles (ventral aspect).
(Schematic drawing.)

1 Subclavius muscle (blue)
2 Pectoralis minor muscle (blue)
3 Pectoralis major muscle (red)
4 Subscapularis muscle (red)
5 Coracobrachialis muscle (light-blue)
6 Serratus anterior muscle (green)

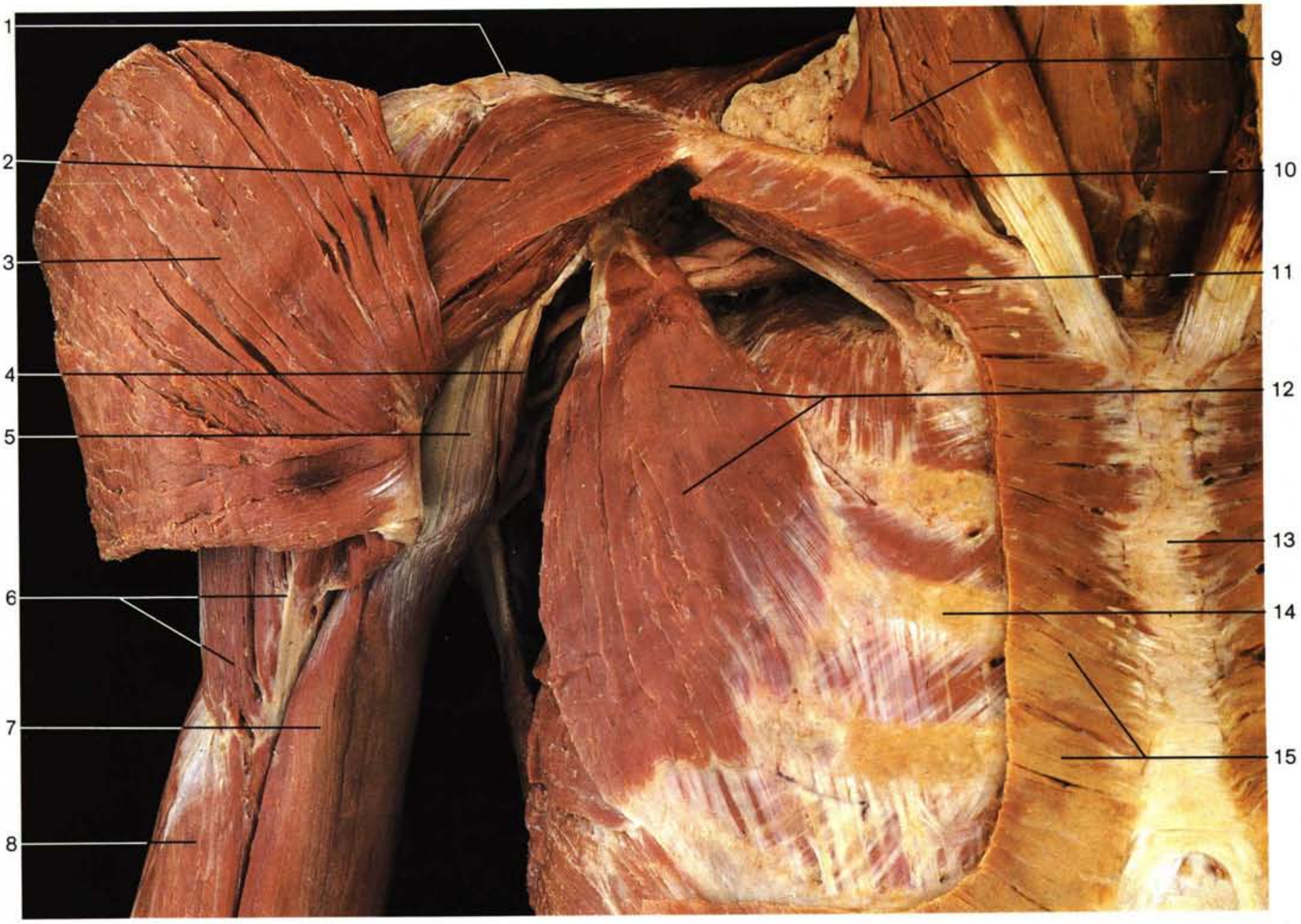

Muscles of shoulder and arm, deep layer (ventral aspect).

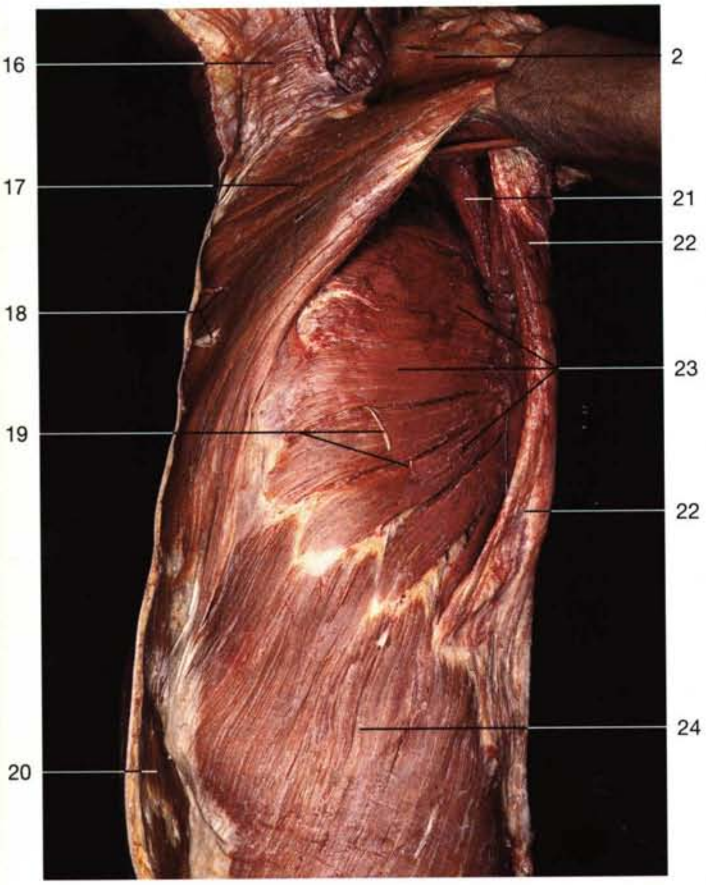

1 Acromion
2 Clavicular part of deltoid muscle
3 Pectoralis major muscle (reflected)
4 Coracobrachialis muscle
5 Short head of biceps brachii muscle
6 Deltoid muscle (insertion on humerus)
7 Long head of biceps brachii muscle
8 Brachialis muscle
9 Sternocleidomastoid muscle
10 Clavicle
11 Subclavius muscle
12 Pectoralis minor muscle
13 Sternum
14 3rd rib
15 Pectoralis major muscle
16 Platysma muscle
17 Pectoralis major muscle forming the anterior axillary fold
18 Anterior cutaneous branches of intercostal nerves
19 Lateral cutaneous branches of intercostal nerves
20 Rectus abdominis muscle
21 Subscapularis muscle
22 **Latissimus dorsi muscle** forming the posterior axillary fold
23 **Serratus anterior muscle** forming the medial wall of the axilla
24 External abdominal oblique muscle

Axillary fossa and serratus anterior muscle (left side, lateral aspect).

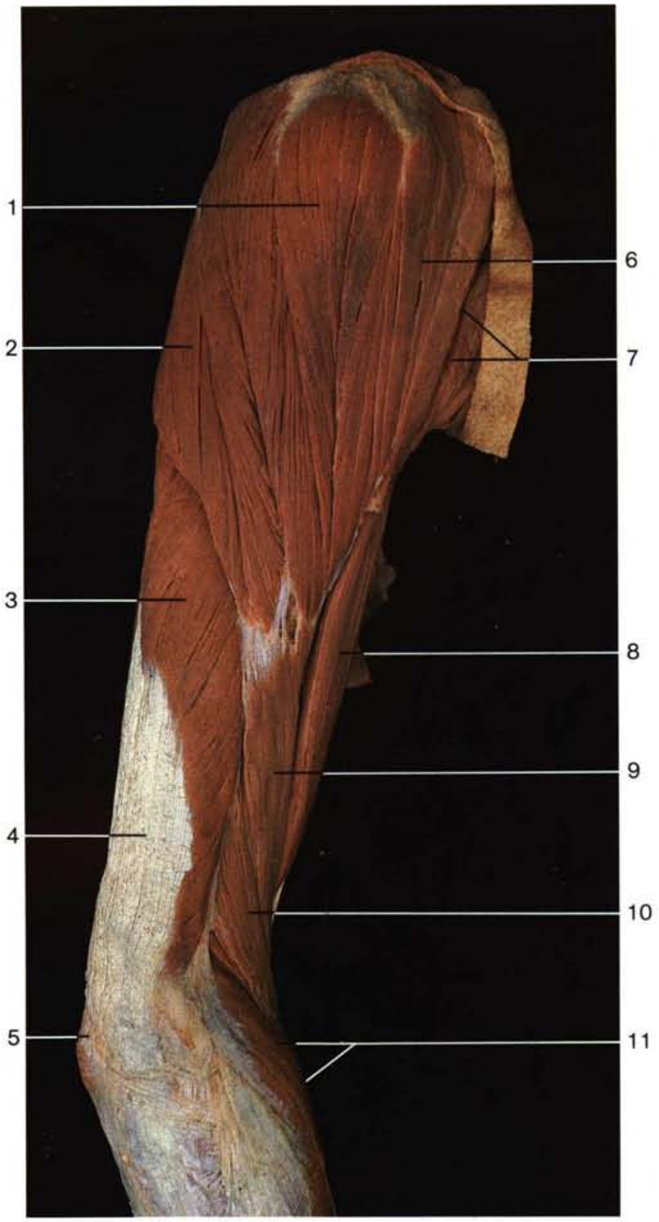
1
2
3
4
5
6
7
8
9
10
11

Muscles of right arm (lateral aspect).

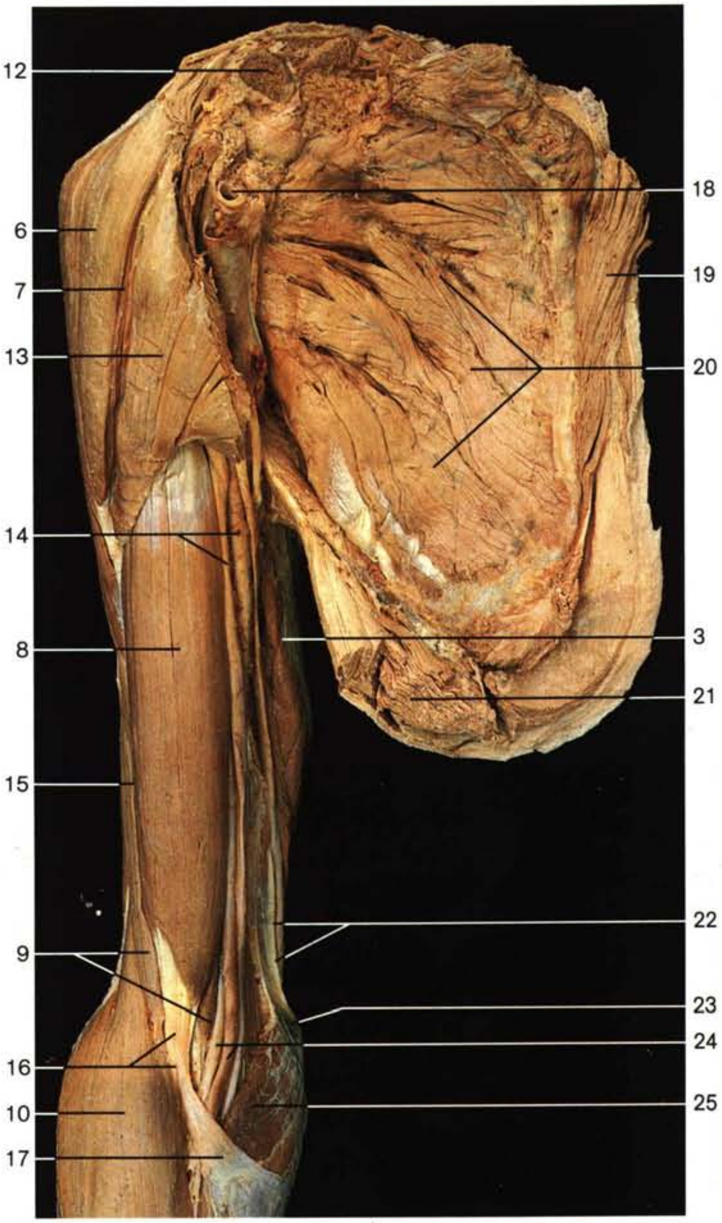
12
6
7
13
14
8
15
9
16
10
17
18
19
20
3
21
22
23
24
25

Muscles of right arm (ventral aspect). The arm with the scapula and attached muscles has been removed from the trunk.

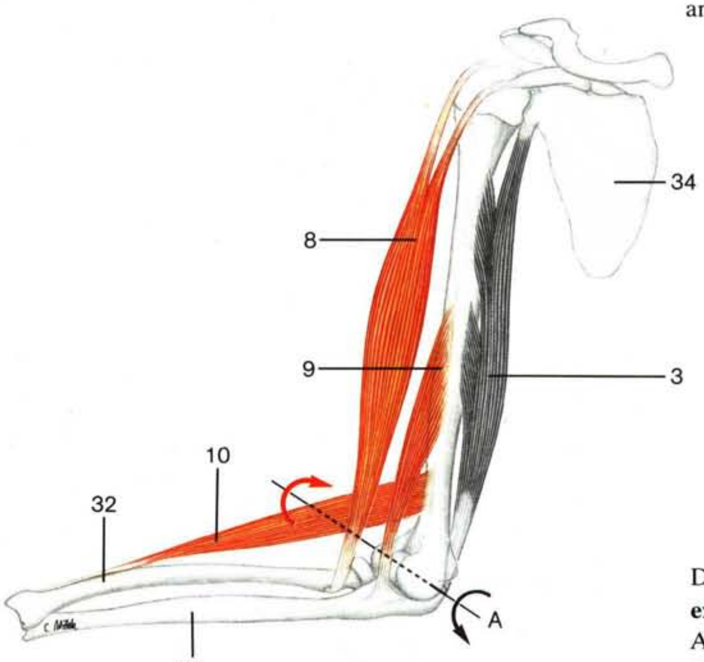
34
8
9
3
10
32
A
33

Diagram illustrating the position of the **flexor and extensor muscles of the arm** and their effect on the elbow joint. A = axis; arrows = direction of movements; red = flexion; black = extension.

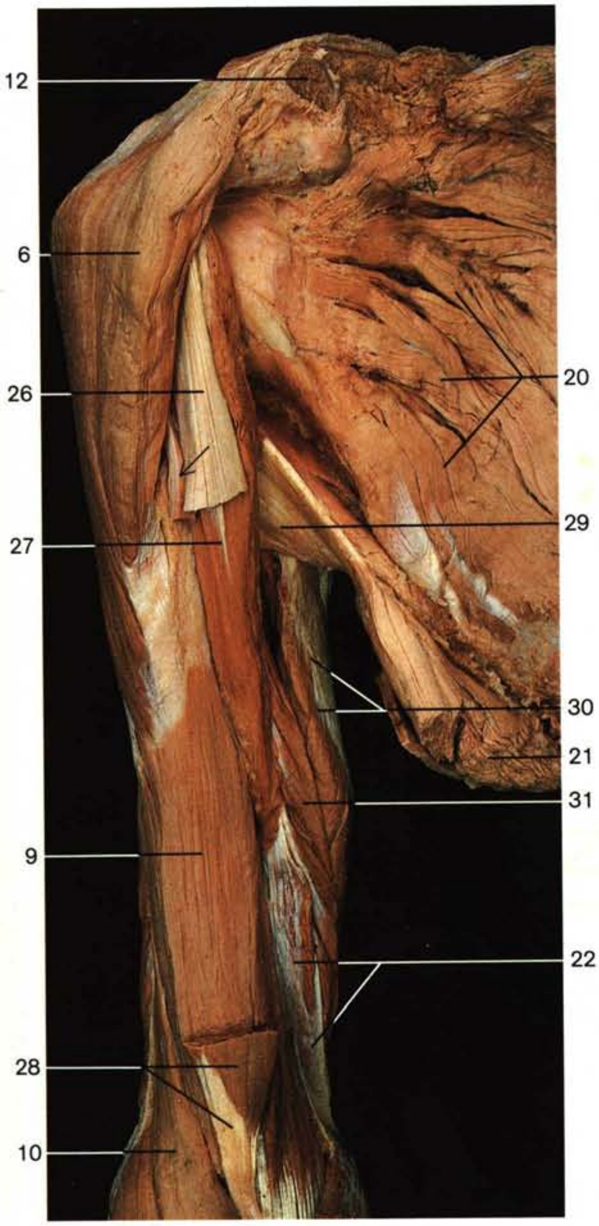

Flexor muscles of right arm (ventral aspect). Part of the biceps brachii muscle has been removed. Arrow: tendon of long head of biceps brachii muscle.

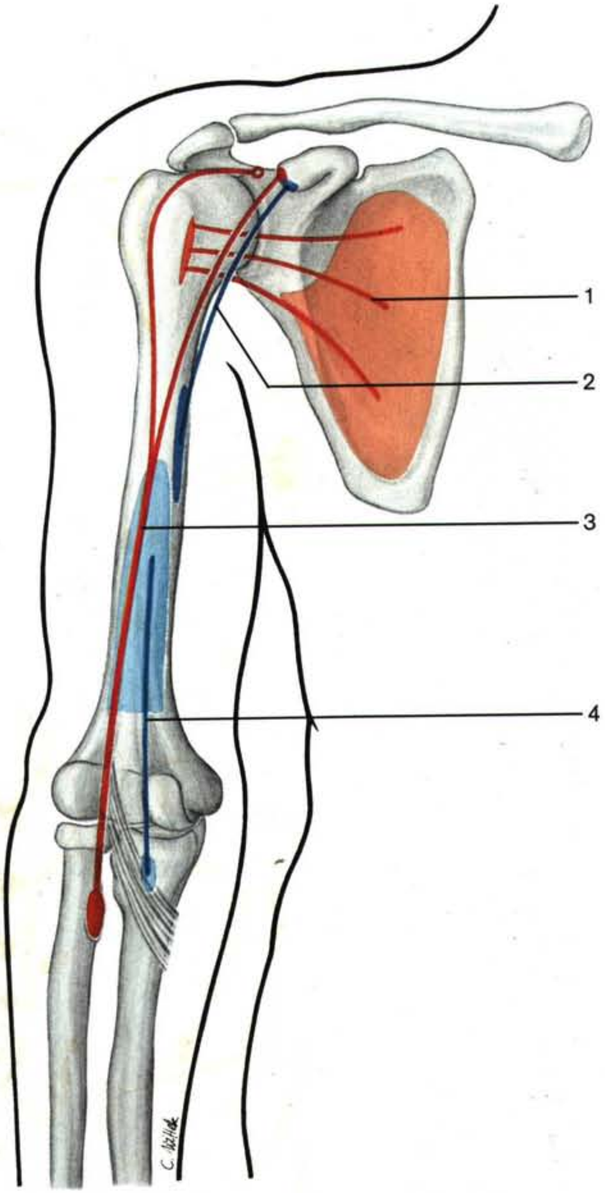

Position and course of flexors of arm. (Schematic drawing.)

1 Subscapularis muscle (red)
2 Coracobrachialis muscle (blue)
3 Biceps brachii muscle (red)
4 Brachialis muscle (blue)

1 Acromial part of deltoid muscle (central fibers)
2 Scapular part of deltoid muscle (posterior fibers)
3 Triceps brachii muscle
4 Tendon of triceps brachii muscle
5 Olecranon
6 Clavicular part of deltoid muscle (anterior fibers)
7 Deltopectoral groove
8 **Biceps brachii muscle**
9 **Brachialis muscle**
10 **Brachioradialis muscle**
11 Extensor carpi radialis longus muscle
12 Clavicle (divided)
13 Pectoralis major muscle
14 Medial intermuscular septum with vessels and nerves
15 Lateral intermuscular septum
16 Tendon of biceps brachii muscle
17 Bicipital aponeurosis
18 Axillary artery
19 Rhomboid major muscle
20 **Subscapularis muscle**
21 Latissimus dorsi muscle (divided)
22 Medial intermuscular septum
23 Medial epicondyle of humerus
24 Brachial artery and median nerve
25 Pronator teres muscle
26 Tendon of short head of biceps brachii muscle
27 Coracobrachialis muscle
28 Distal part of biceps brachii muscle
29 Teres major muscle
30 Long head of triceps brachii muscle
31 Medial head of triceps brachii muscle
32 Radius
33 Ulna
34 Scapula

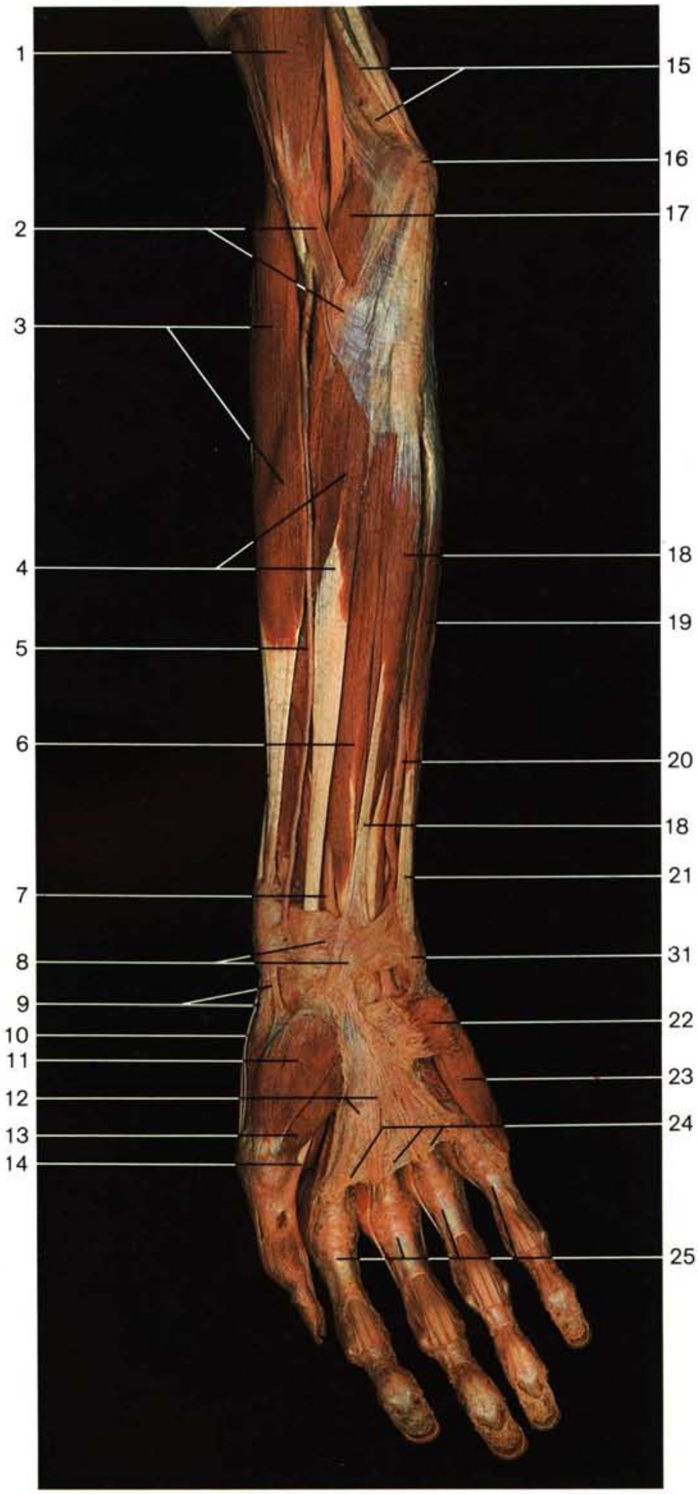
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
18
21
31
22
23
24
25

Forearm muscles, superficial layer (ventral aspect).

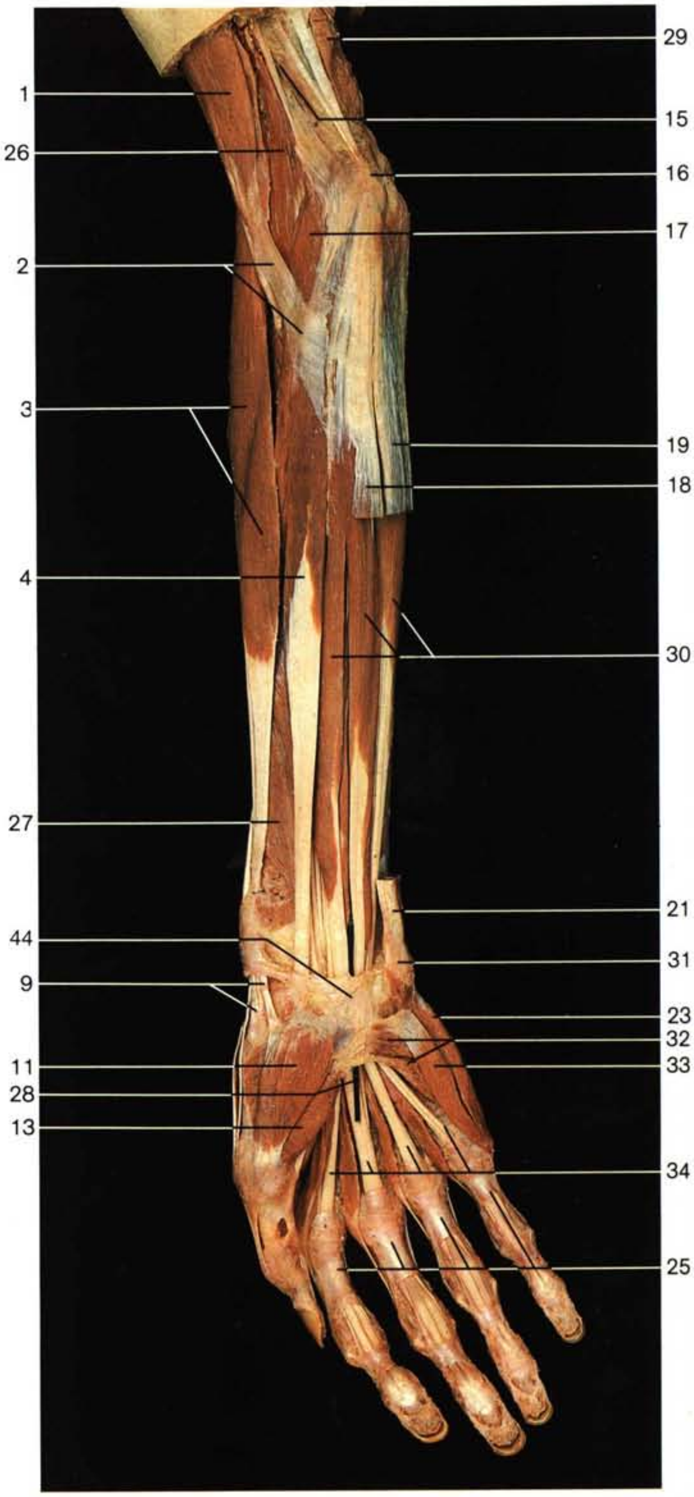
29
1
15
26
16
17
2
3
19
18
4
30
27
21
44
31
9
23
32
11
33
28
13
34
25

Forearm muscles, superficial layer (ventral aspect). The palmaris longus and flexor carpi ulnaris muscles have been removed.

1 Biceps brachii muscle
2 Bicipital aponeurosis
3 **Brachioradialis muscle**
4 **Flexor carpi radialis muscle**
5 Radial artery
6 **Flexor digitorum superficialis muscle**
7 Median nerve
8 Antebrachial fascia and tendon of palmaris longus muscle
9 Tendon of abductor pollicis longus muscle
10 Tendon of extensor pollicis brevis muscle
11 Abductor pollicis brevis muscle
12 **Palmar aponeurosis**
13 Superficial head of flexor pollicis brevis muscle
14 Tendon of flexor pollicis longus muscle
15 Medial intermuscular septum
16 Medial epicondyle of humerus
17 Humeral head of pronator teres muscle
18 **Palmaris longus muscle**
19 **Flexor carpi ulnaris muscle**
20 **Ulnar artery**
21 Tendon of flexor carpi ulnaris muscle
22 Palmaris brevis muscle

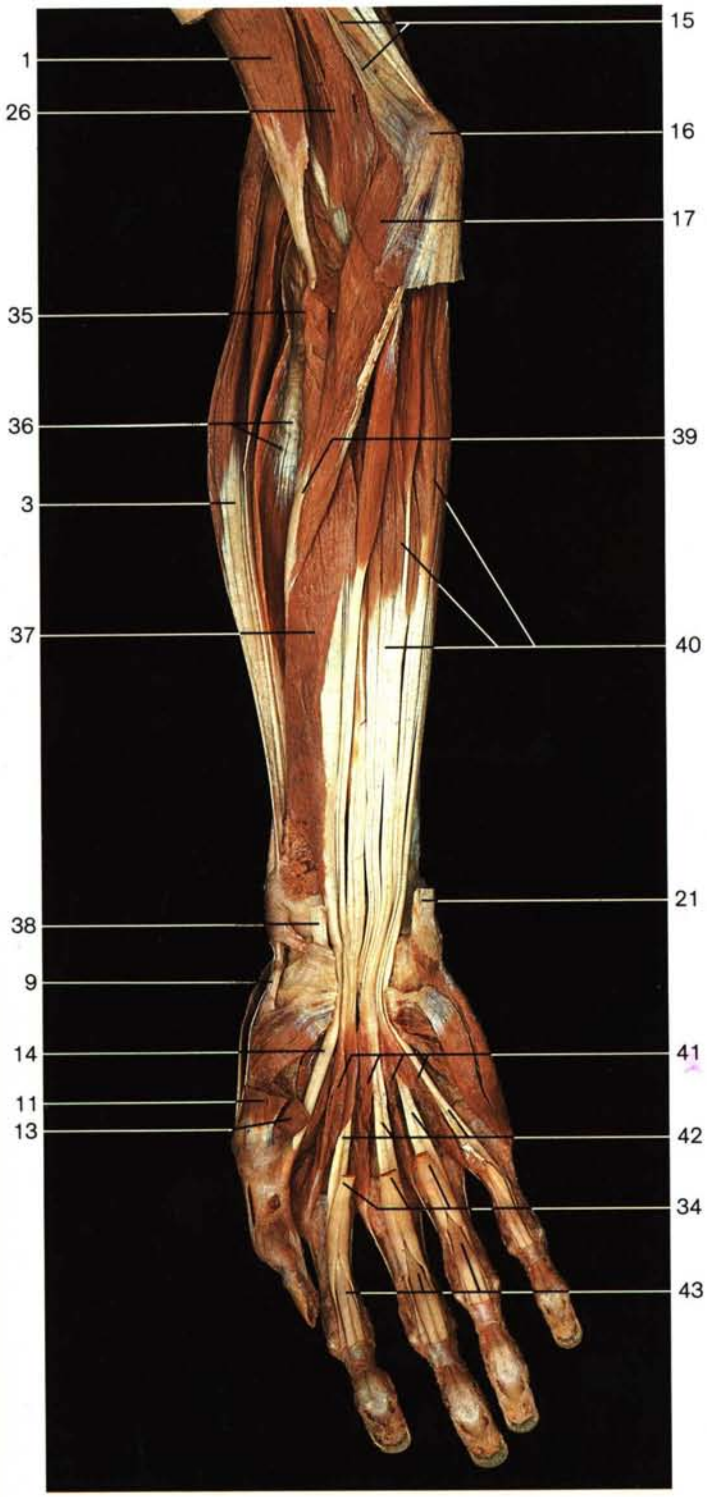

Forearm muscles, middle layer (ventral aspect). The palmaris longus, flexor carpi radialis and ulnaris muscles have been partly removed. The flexor retinaculum has been divided.

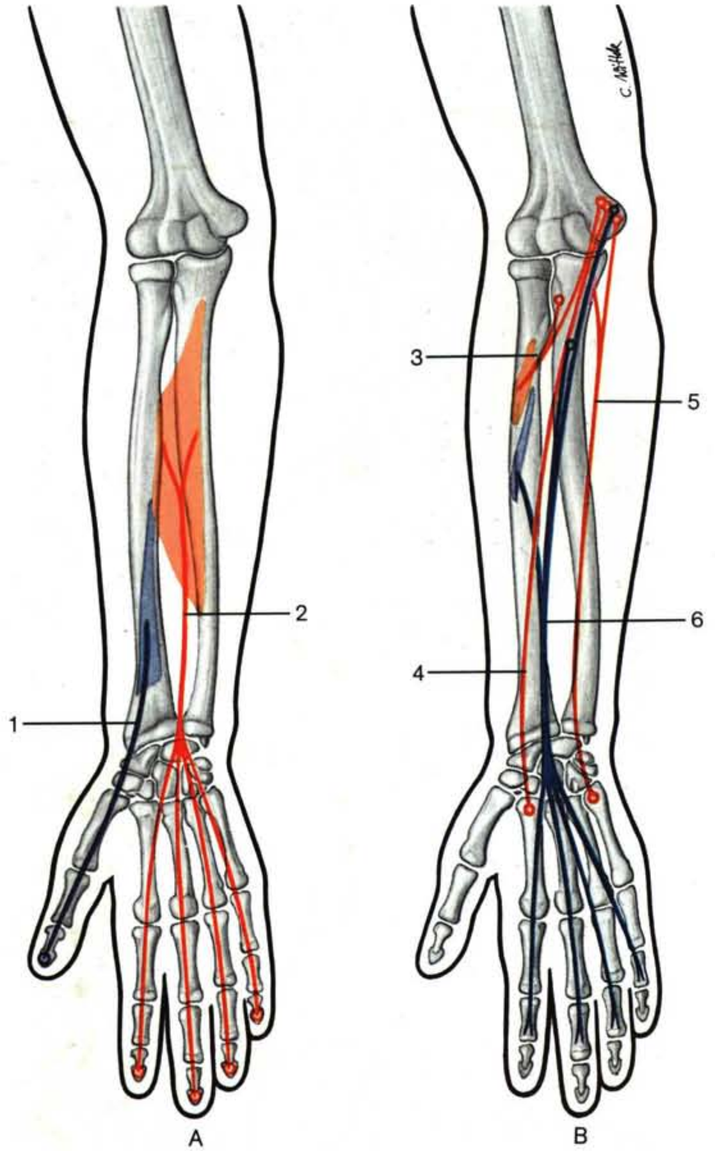

Position of flexors of fingers and hand. (Schematic drawing.)

A Deep layer
B Superficial layer
△
1 Flexor pollicis longus muscle (blue)
2 Flexor digitorum profundus muscle (red)
3 Pronator teres muscle (red)
4 Flexor carpi radialis muscle (red)
5 Flexor carpi ulnaris muscle (red)
6 Flexor digitorum superficialis muscle (blue)

23 Abductor digiti minimi muscle
24 Transverse fasciculi of palmar aponeurosis
25 Digital fibrous sheaths of tendons of flexor digitorum muscle
26 Brachialis muscle
27 Flexor pollicis longus muscle
28 Carpal tunnel (canalis carpi, probe)
29 Triceps brachii muscle
30 Flexor digitorum superficialis muscle
31 Pisiform bone
32 Opponens digiti minimi muscle
33 Flexor digiti minimi brevis muscle
34 Tendons of flexor digitorum superficialis muscle
35 Supinator muscle
36 Extensor carpi radialis brevis muscle
37 **Flexor pollicis longus muscle**
38 Tendon of flexor carpi radialis muscle
39 **Pronator teres muscle** (insertion of radius)
40 **Flexor digitorum profundus muscle**
41 **Lumbrical muscles**
42 Tendons of flexor digitorum profundus muscle
43 Tendons of flexor digitorum profundus muscle having passed through the divided tendons of the flexor digitorum superficialis muscle
44 Flexor retinaculum

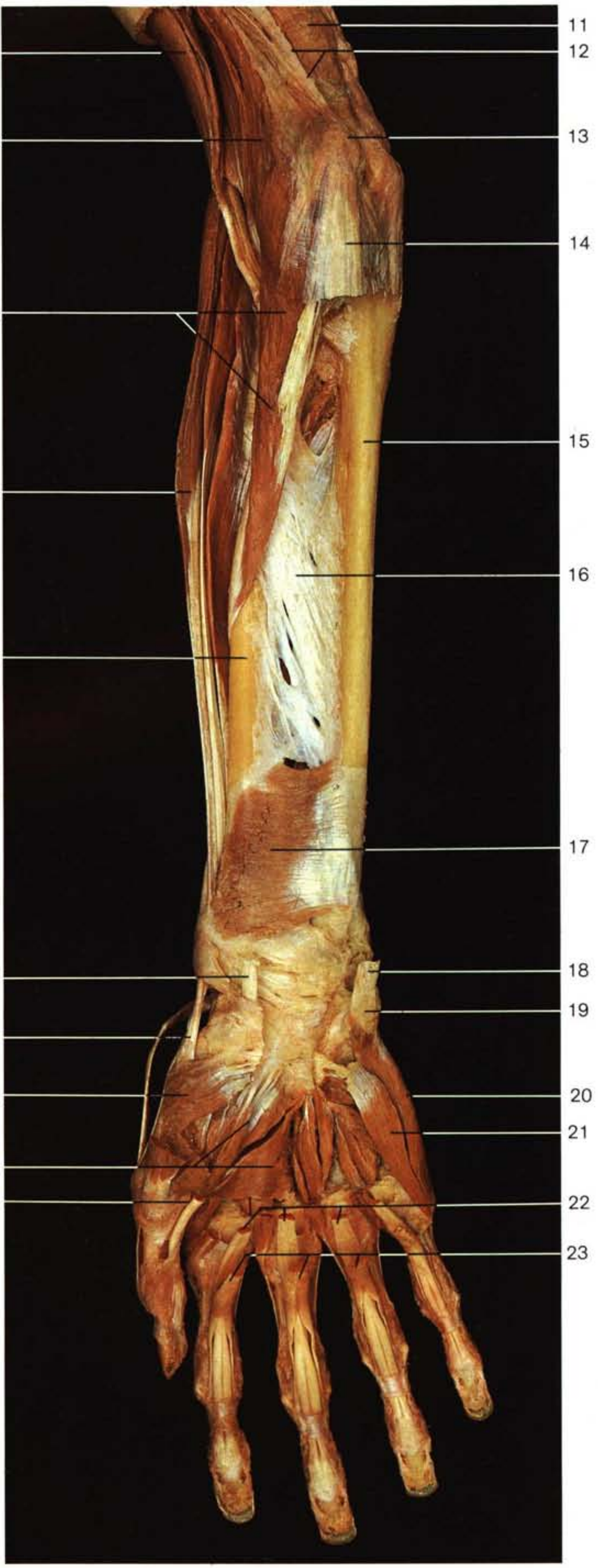

Muscles of the forearm, deep layer (ventral aspect). All flexors have been removed to display the pronator quadratus and pronator teres muscles together with the interosseous membrane. Forearm in supination.

1 Biceps brachii muscle
2 Brachialis muscle
3 **Pronator teres muscle**
4 Brachioradialis muscle
5 **Radius**
6 Tendon of flexor carpi radialis muscle
7 Tendon of abductor pollicis longus muscle
8 Opponens pollicis muscle
9 Adductor pollicis muscle
10 Tendon of flexor pollicis longus muscle
11 Triceps brachii muscle
12 Medial intermuscular septum
13 Medial epicondyle of humerus
14 Common flexor mass (divided)
15 **Ulna**
16 **Interosseous membrane**
17 **Pronator quadratus muscle**
18 Tendon of flexor carpi ulnaris muscle
19 Pisiform bone
20 Abductor digiti minimi muscle
21 Flexor digiti minimi brevis muscle
22 Tendons of flexor digitorum profundus muscle
23 Tendons of flexor digitorum superficialis muscle
24 Flexor retinaculum
25 Hypothenar muscles
26 Thenar muscles
27 Common synovial sheath of flexor tendons
28 Synovial sheath of tendon of flexor pollicis longus muscle
29 Digital synovial sheaths of flexor tendons

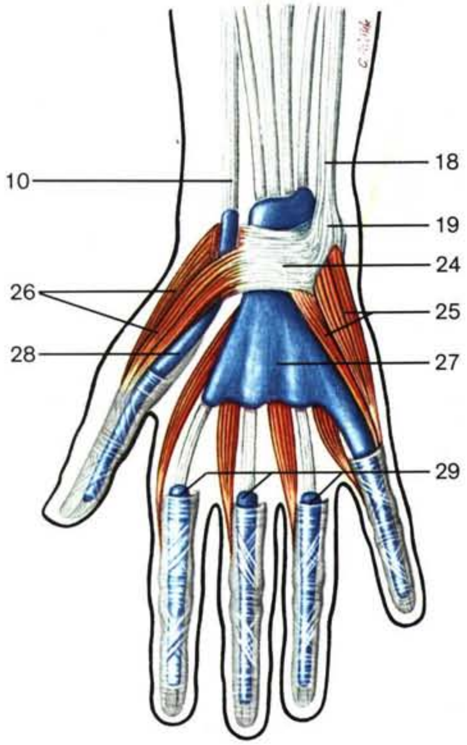

Synovial sheaths of flexor tendons (palmar aspect of right hand). (Semischematic drawing.)

1 Humerus
2 Lateral epicondyle of humerus
3 Articular capsule
4 Position of capitulum of humerus
5 Deep branch of radial nerve
6 **Supinator muscle**
7 Entrance of deep branch of radial nerve to extensor muscles
8 Radius and insertion of pronator teres muscle
9 Interosseous membrane
10 Median nerve
11 Triceps brachii muscle
12 Trochlea of humerus
13 Tendon of biceps brachii muscle
14 Brachial artery
15 **Pronator teres muscle**
16 Tendon of pronator teres muscle
17 Ulna
18 **Pronator quadratus muscle**
19 Tendon of flexor carpi radialis muscle
20 Thenar muscles
21 Synovial sheath of tendon of flexor pollicis longus muscle
22 Fibrous sheath of flexor tendons
23 **Digital synovial sheath of flexor tendons**
24 Flexor digitorum superficialis muscle
25 Tendon of flexor carpi ulnaris muscle
26 **Common synovial sheath of flexor tendons**
27 Position of pisiform bone
28 Flexor retinaculum
29 Hypothenar muscles

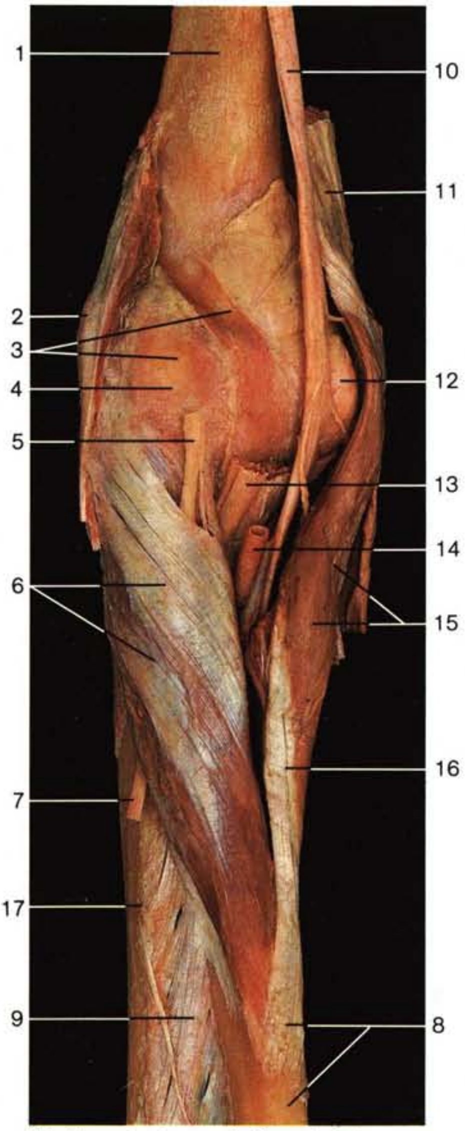

Right supinator and elbow joint (anterior aspect). Forearm in pronation.

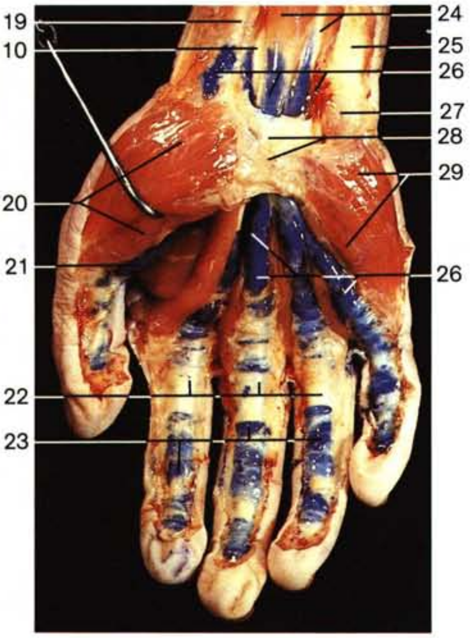

Synovial sheaths of flexor muscles (palmar aspect of right hand). Blue PVA-solution has been injected into the sheaths.

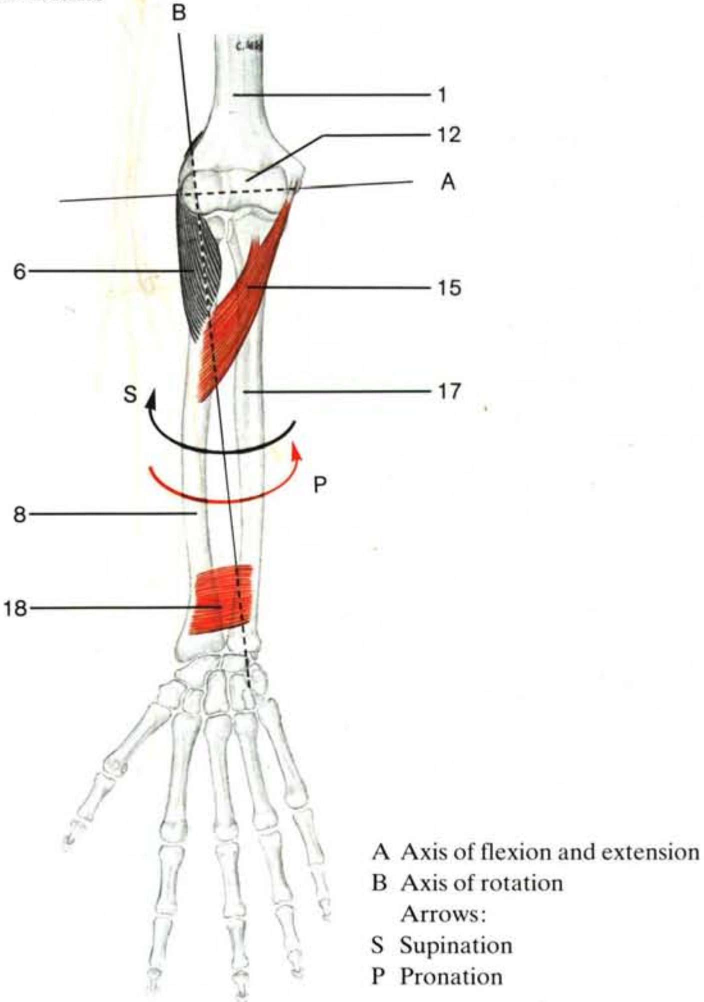

A Axis of flexion and extension
B Axis of rotation
Arrows:
S Supination
P Pronation

Diagram illustrating the two **axes of the elbow joint.**

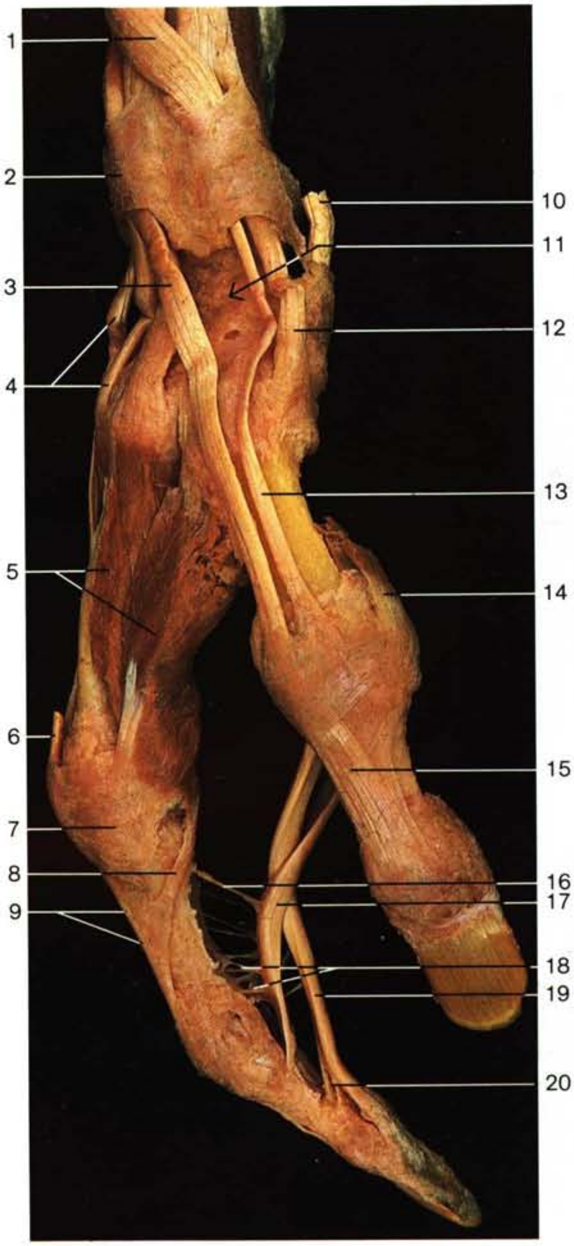

Muscles of thumb and index finger (medial aspect). The tendons of the extensor muscles of the thumb and the insertion of the flexor tendons of the index finger are displayed.

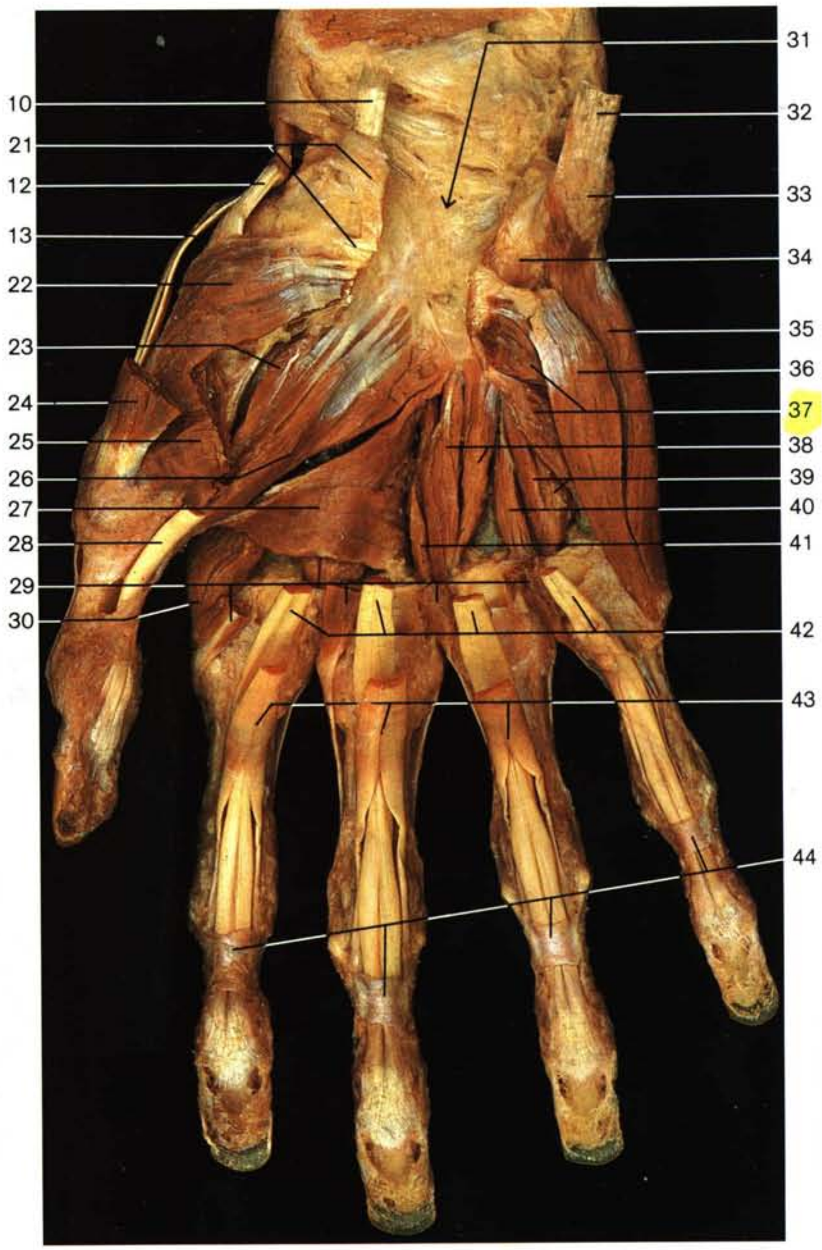

Muscles of hand (palmar aspect). The tendons of the flexor muscles and parts of the thumb muscles have been removed. The carpal tunnel has been opened.

1 Tendons of extensor pollicis brevis and abductor pollicis longus muscle
2 Extensor retinaculum
3 Tendon of extensor pollicis longus muscle
4 Tendons of extensor carpi radialis longus and brevis muscles
5 First dorsal interosseous muscle
6 Tendon of extensor digitorum muscle for index finger
7 Location of metacarpophalangeal joint
8 Tendon of lumbrical muscle
9 Extensor expansion of index finger
10 Tendon of flexor carpi radialis muscle (cut)
11 **Anatomical snuffbox**
12 Tendon of abductor pollicis longus muscle
13 Tendon of extensor pollicis brevis muscle
14 Tendon of abductor pollicis brevis muscle
15 Extensor expansion of extensor of thumb
16 Vinculum longum
17 Tendons of flexor digitorum superficialis muscle dividing to allow passage of deep tendons
18 Vincula of flexor tendons
19 Tendon of flexor digitorum profundus muscle
20 Vinculum breve
21 Radial carpal eminence (cut edge of flexor retinaculum)
22 Opponens pollicis muscle
23 Deep head of flexor pollicis brevis muscle
24 Abductor pollicis brevis muscle (cut)
25 Superficial head of flexor pollicis brevis muscle (cut)
26 Oblique head of adductor pollicis muscle
27 Transverse head of adductor pollicis muscle
28 Tendon of flexor pollicis longus muscle (cut)
29 Lumbrical muscles (cut)
30 First dorsal interosseous muscle
31 **Position of carpal tunnel**
32 Tendon of flexor carpi ulnaris muscle
33 Location of pisiform bone
34 Hook of hamate bone
35 Abductor digiti minimi muscle
36 Flexor digiti minimi brevis muscle
37 Opponens digiti minimi muscle
38 2nd palmar interosseous muscle
39 3rd palmar interosseous muscle
40 4th dorsal interosseous muscle
41 3rd dorsal interosseous muscle
42 Tendon of flexor digitorum profundus muscle (cut)
43 Tendons of flexor digitorum superficialis muscle (cut)
44 Fibrous flexor sheaths

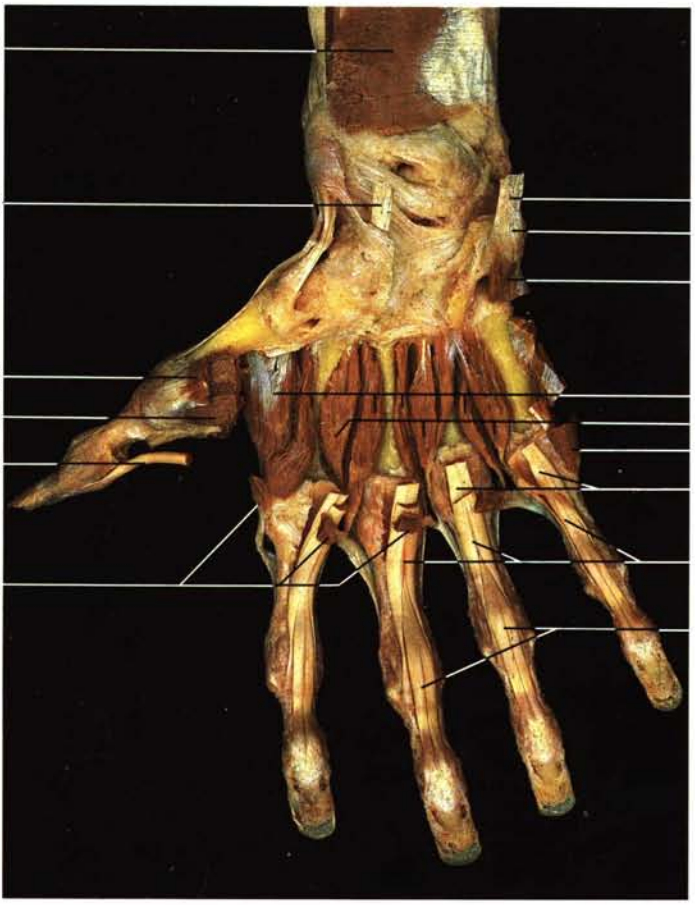

Muscles of right hand, deep layer (palmar aspect). The thenar and hypothenar muscles have been removed to display the interosseous muscles.

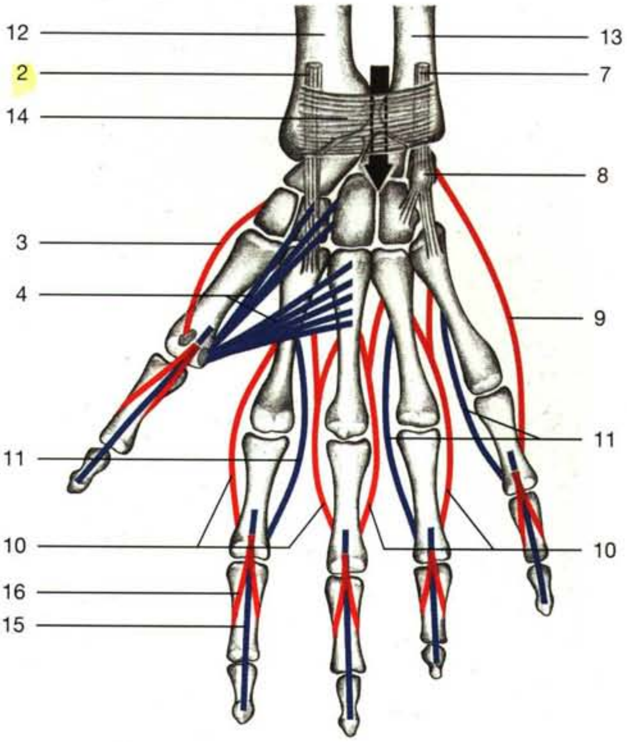

Actions of interosseous muscles in abduction and adduction of fingers (palmar aspect, schematic drawing). Arrow: carpal tunnel.
Red = **abduction**
(dorsal interosseous muscles, abductor digiti minimi and abductor pollicis brevis muscles).
Black = **adduction**
(palmar interosseous muscles, adductor pollicis muscle).

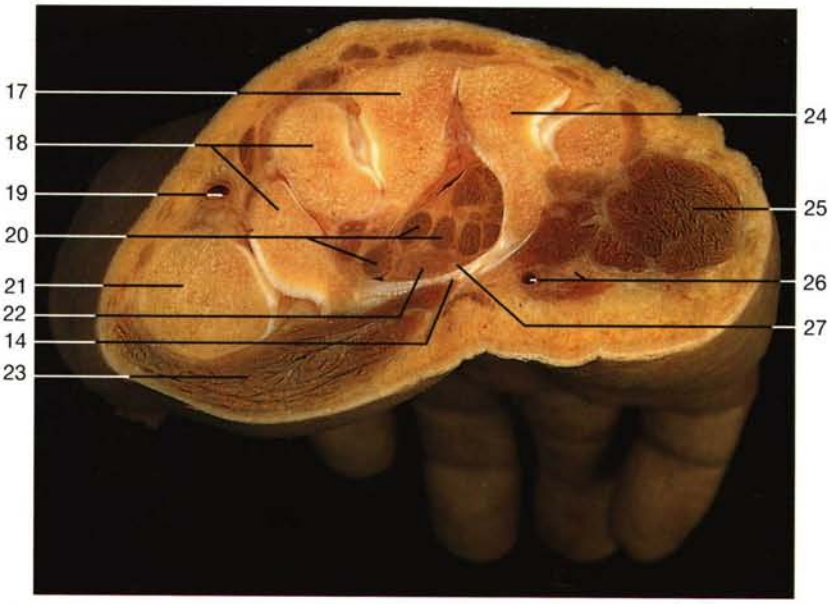

Transverse section through the right hand, showing the carpal tunnel (canalis carpi).

1 Pronator quadratus muscle
2 Tendon of flexor carpi radialis muscle
3 Abductor pollicis brevis muscle (divided)
4 Adductor pollicis muscle (divided)
5 Tendon of flexor pollicis longus muscle
6 Lumbrical muscles (cut)
7 Tendon of flexor carpi ulnaris muscle
8 Pisiform bone
9 Abductor digiti minimi muscle (divided)
10 **Dorsal interosseous muscles**
11 **Palmar interosseous muscles**
12 Radius
13 Ulna
14 **Flexor retinaculum**
15 Tendons of flexor digitorum profundus muscle
16 Tendons of flexor digitorum superficialis muscle
17 Capitate bone
18 Trapezium bone and trapezoid bone
19 Radial artery
20 Tendon of flexor muscles
21 First metacarpal bone
22 **Median nerve**
23 Thenar muscles
24 Hamate bone
25 Hypothenar muscles
26 Ulnar artery and nerve
27 **Carpal tunnel** (canalis carpi)

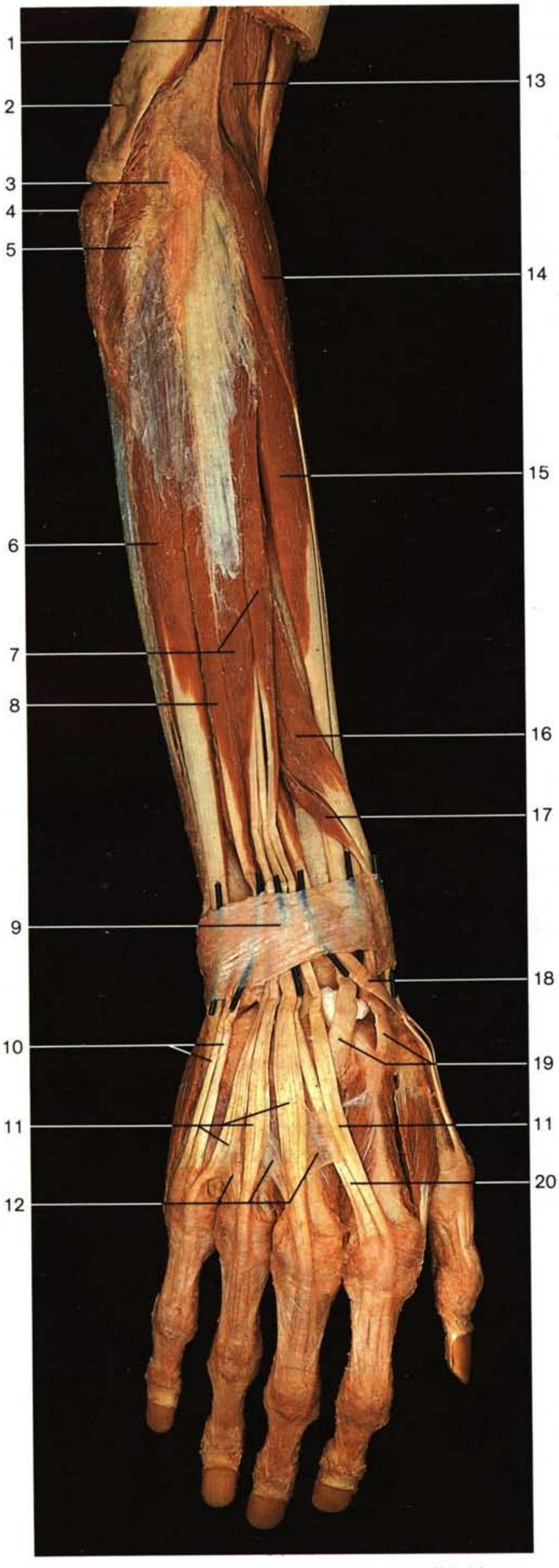

Extensor muscles of forearm and hand, superficial layer (dorsal aspect). Tunnels for extensor tendons indicated by probes.

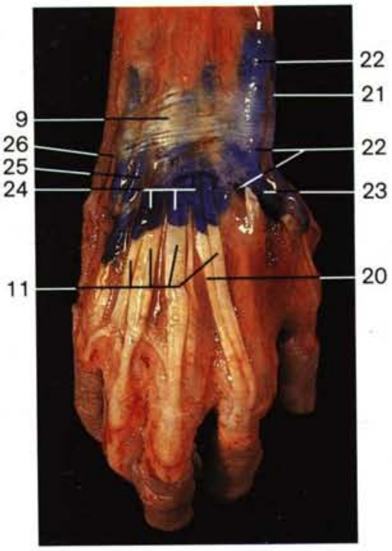

Synovial sheaths of extensor tendons. The sheaths have been injected with blue gelatin.

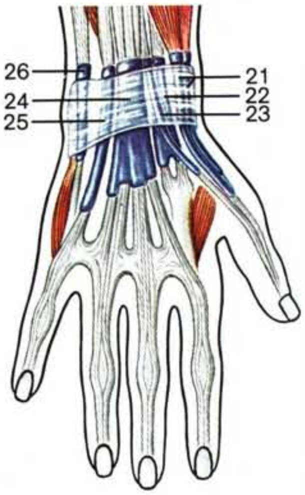

Synovial sheaths of extensor tendons on the back of right wrist (indicated in blue). Notice the six tunnels for the passage of the extensor tendons beneath the extensor retinaculum. (Semischematic drawing.)

1 Lateral intermuscular septum
2 Tendon of triceps brachii muscle
3 Lateral epicondyle of humerus
4 Olecranon
5 Anconeus
6 Extensor carpi ulnaris muscle
7 Extensor digitorum muscle
8 Extensor digiti minimi muscle
9 Extensor retinaculum
10 Tendons of extensor digiti minimi muscle
11 Tendons of extensor digitorum muscle
12 Intertendinous connections
13 Brachioradialis muscle
14 Extensor carpi radialis longus muscle
15 Extensor carpi radialis brevis muscle
16 Abductor pollicis longus muscle
17 Extensor pollicis brevis muscle
18 Tendon of extensor pollicis longus muscle
19 Tendons of both extensor carpi radialis longus and extensor carpi radialis brevis muscles
20 Tendon of extensor indicis muscle
21 1st tunnel: Abductor pollicis longus muscle
Extensor pollicis brevis muscle
22 2nd tunnel: Extensor carpi radialis longus and brevis muscles
23 3rd tunnel: Extensor pollicis longus muscle
24 4th tunnel: Extensor digitorum muscle
Extensor indicis muscle
25 5th tunnel: Extensor digiti minimi muscle
26 6th tunnel: Extensor carpi ulnaris muscle

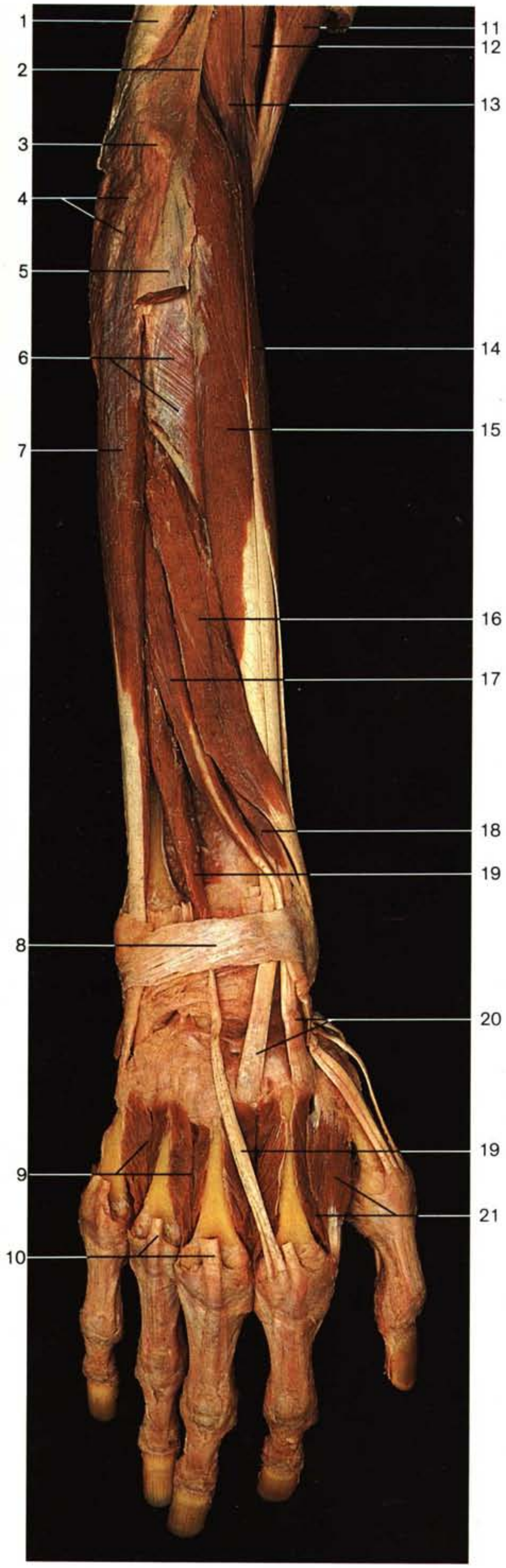

Extensor muscles of forearm and hand, deep layer (dorsal aspect).

1 Triceps brachii muscle
2 Lateral intermuscular septum
3 Lateral epicondyle of humerus
4 Anconeus muscle
5 Extensor digitorum and extensor digiti minimi muscles (cut)
6 Supinator muscle
7 Extensor carpi ulnaris muscle
8 Extensor retinaculum
9 3rd and 4th dorsal interosseous muscles
10 Tendons of extensor digitorum muscle (cut)
11 Biceps brachii muscle
12 Brachialis muscle
13 Brachioradialis muscle
14 Extensor carpi radialis longus muscle
15 Extensor carpi radialis brevis muscle
16 Abductor pollicis longus muscle
17 Extensor pollicis longus muscle
18 Extensor pollicis brevis muscle
19 Extensor indicis muscle
20 Tendons of the extensor carpi radialis longus and extensor carpi radialis brevis muscles
21 1st dorsal interosseous muscle

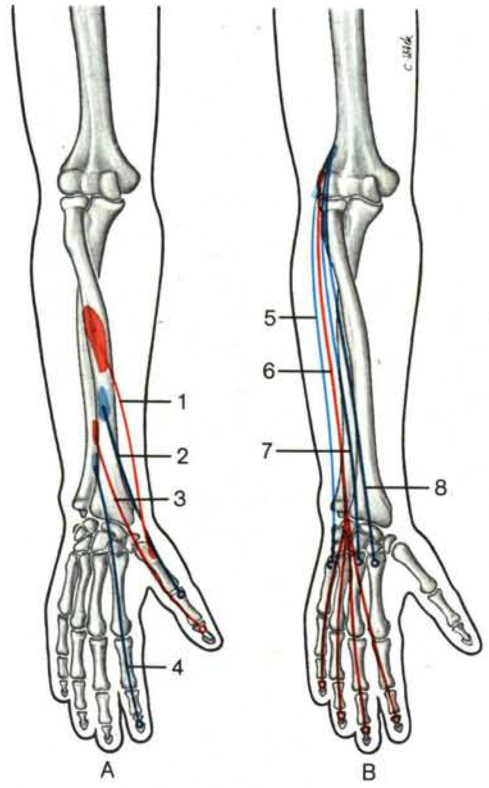

Position of extensor muscles of forearm and hand. (Semischematic drawing.)

A Extensors of thumb

1 Abductor pollicis longus muscle (red)
2 Extensor pollicis brevis muscle (blue)
3 Extensor pollicis longus muscle (red)
4 Extensor indicis muscle (blue)

B Extensors of fingers and hand

5 Extensor carpi ulnaris muscle (blue)
6 Extensor digitorum muscle (red)
7 Extensor carpi radialis brevis muscle (blue)
8 Extensor carpi radialis longus muscle (blue)

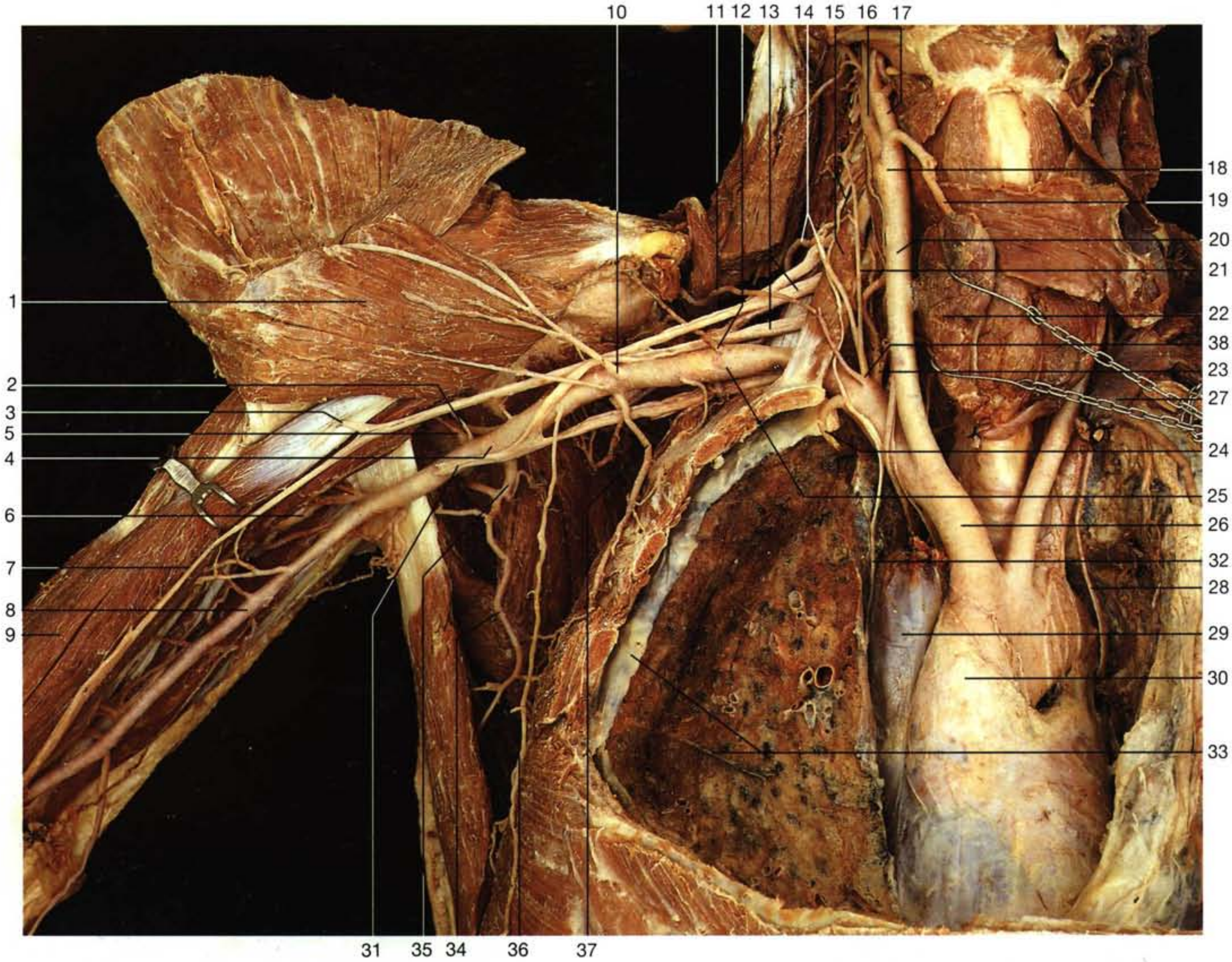

Main branches of right subclavian and axillary arteries (anterior aspect). Pectoralis muscles have been reflected, clavicle and anterior wall of thorax removed and right lung divided. Left lung with pleura and thyroid gland have been reflected laterally to display aortic arch and common carotid artery with their branches.

1 Pectoralis minor muscle (reflected)
2 Anterior circumflex humeral artery
3 Musculocutaneous nerve (divided)
4 **Axillary artery**
5 Posterior circumflex humeral artery
6 **Profunda brachii artery**
7 Median nerve (var.)
8 **Brachial artery**
9 Biceps brachii muscle
10 Thoracoacromial artery
11 Suprascapular artery
12 Descending scapular artery
13 Brachial plexus (middle trunk)
14 Transverse cervical artery
15 Scalenus anterior muscle and phrenic nerve
16 Right internal carotid artery
17 Right external carotid artery
18 Carotid sinus
19 Superior thyroid artery
20 Right common carotid artery
21 Ascending cervical artery
22 Thyroid gland
23 Inferior thyroid artery
24 Internal thoracic artery
25 Right subclavian artery
26 Brachiocephalic trunk
27 Left brachiocephalic vein (divided)
28 Left vagus nerve
29 Superior vena cava (divided)
30 **Ascending aorta**
31 Median nerve (divided)
32 Phrenic nerve
33 Right lung (divided) and pulmonary pleura
34 Thoracodorsal artery
35 Subscapular artery
36 Lateral mammary branches (variant)
37 Lateral thoracic artery
38 **Thyrocervical trunk**
39 Superior thoracic artery
40 Superior ulnar collateral artery
41 Inferior ulnar collateral artery
42 Middle collateral artery
43 Radial collateral artery
44 Radial recurrent artery
45 **Radial artery**
46 Anterior and posterior interosseous arteries
47 Princeps pollicis artery
48 **Deep palmar arch**
49 Common palmar digital arteries
50 Ulnar recurrent artery
51 Recurrent interosseous artery
52 Common interosseous artery
53 **Ulnar artery**
54 **Superficial palmar arch**
55 Median nerve and brachial artery
56 Biceps brachii muscle
57 Ulnar nerve
58 Flexor pollicis longus muscle
59 Palmar digital arteries
60 Anterior interosseous artery
61 Flexor carpi ulnaris muscle
62 Superficial palmar branch of radial artery

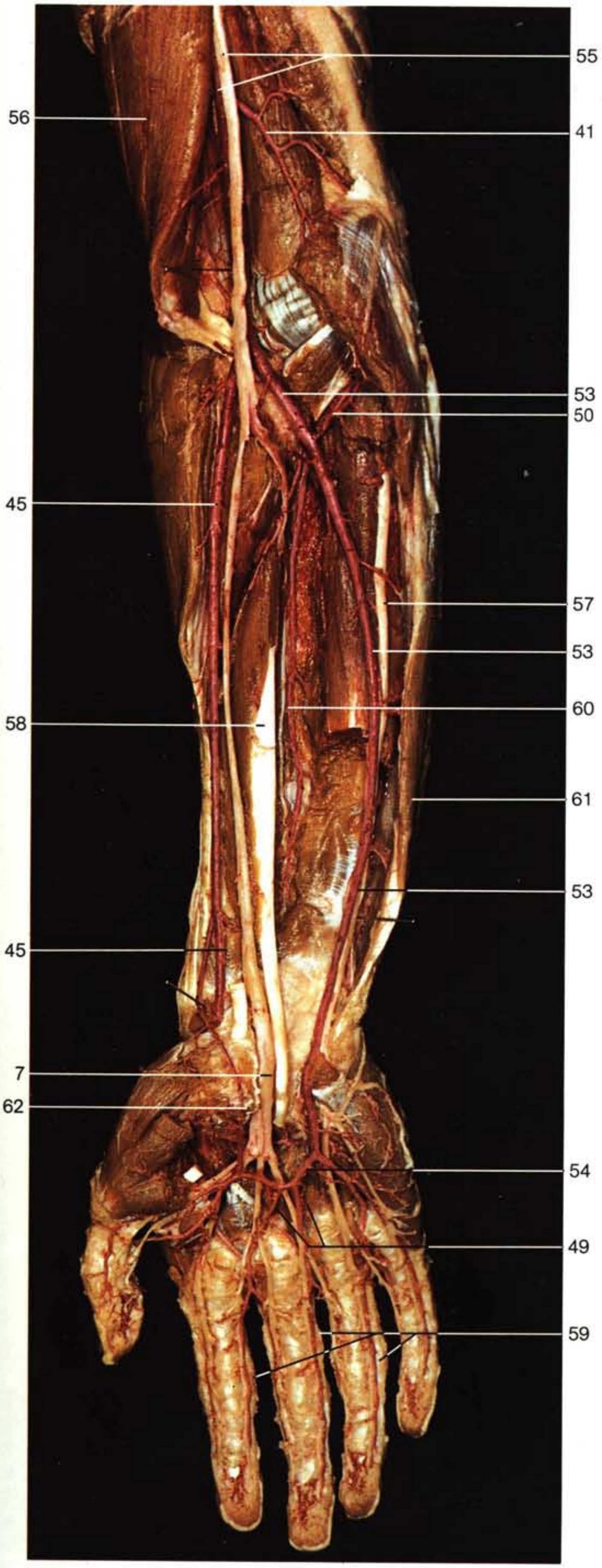
55
56
41
53
50
45
57
53
58
60
61
53
45
7
62
54
49
59

Dissection of the arteries of forearm and hand. The superficial flexor muscles have been removed, the carpal tunnel opened and the flexor retinaculum cut. The arteries have been filled with colored resin.

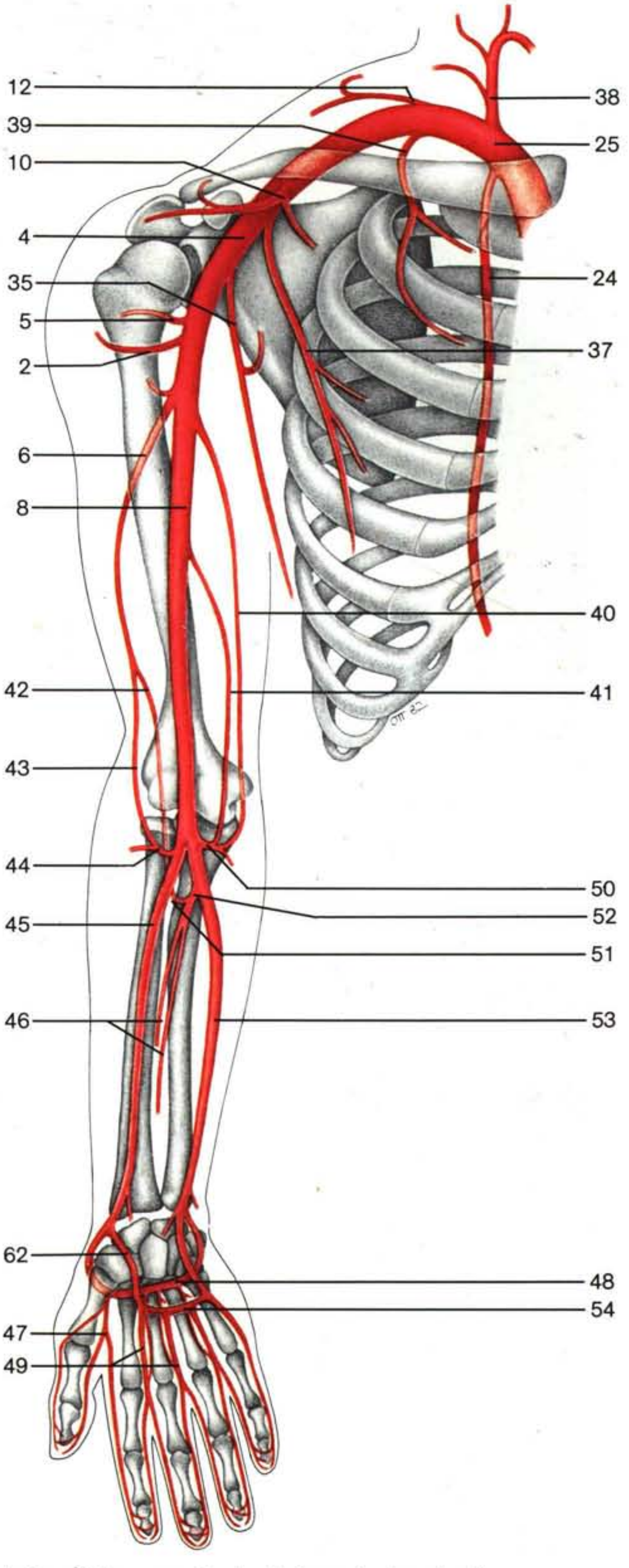
12
38
39
25
10
4
24
35
5
37
2
6
8
40
42
41
43
44
50
52
45
51
46
53
62
48
54
47
49

Arteries of the upper limb. (Schematic drawing.)

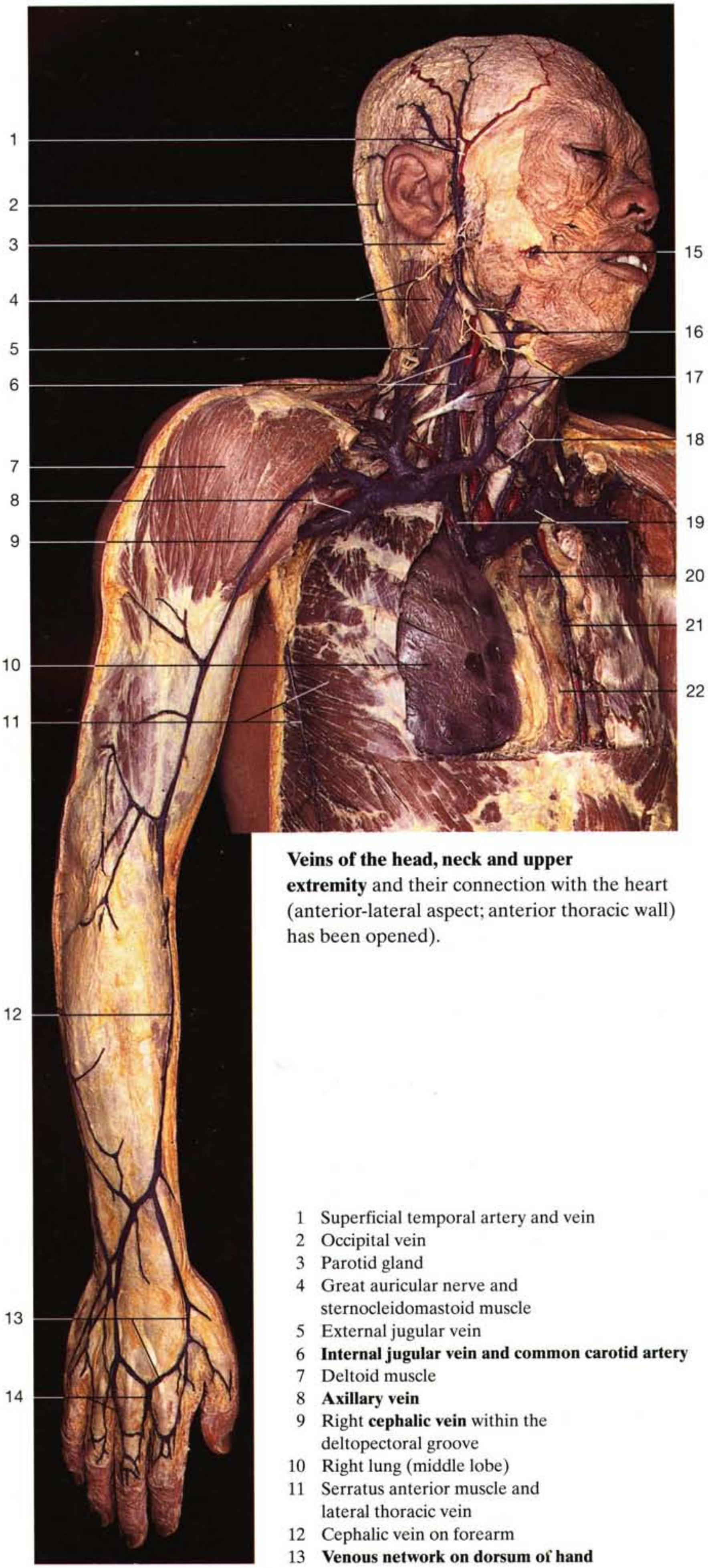

Veins of the head, neck and upper extremity and their connection with the heart (anterior-lateral aspect; anterior thoracic wall) has been opened).

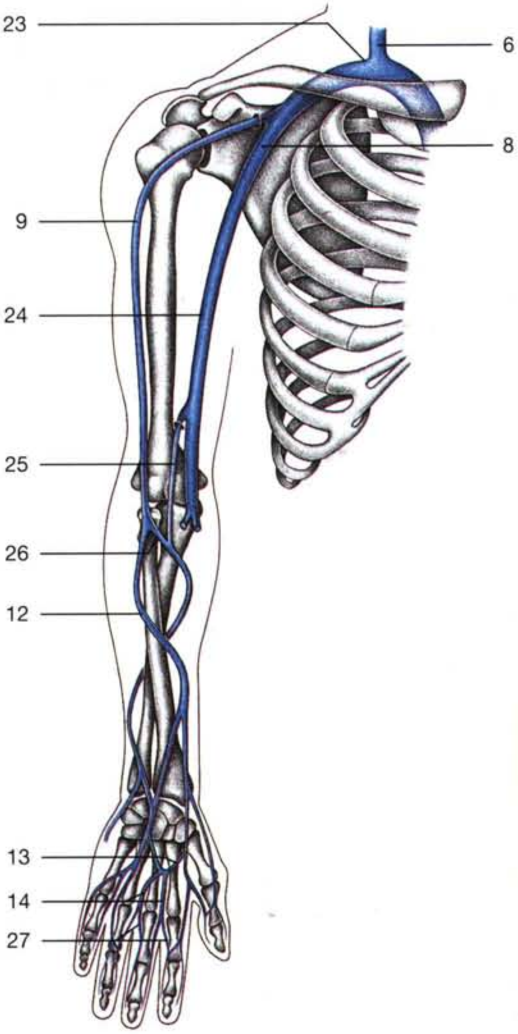

Superficial veins of upper limb. (Semischematic drawing.)

1 Superficial temporal artery and vein
2 Occipital vein
3 Parotid gland
4 Great auricular nerve and sternocleidomastoid muscle
5 External jugular vein
6 **Internal jugular vein and common carotid artery**
7 Deltoid muscle
8 **Axillary vein**
9 Right **cephalic vein** within the deltopectoral groove
10 Right lung (middle lobe)
11 Serratus anterior muscle and lateral thoracic vein
12 Cephalic vein on forearm
13 **Venous network on dorsum of hand**
14 Dorsal metacarpal veins
15 **Facial artery and vein**
16 Submandibular gland
17 Anterior jugular vein, hyoid bone and omohyoid muscle
18 Jugular venous arch and thyroid gland
19 Right and left brachiocephalic veins
20 Retrosternal body (remnant of thymus gland)
21 Internal thoracic artery and vein
22 Heart with pericardium
23 Right venous angle
24 **Brachial vein**
25 Basilic vein
26 Median cubital vein
27 Digital veins

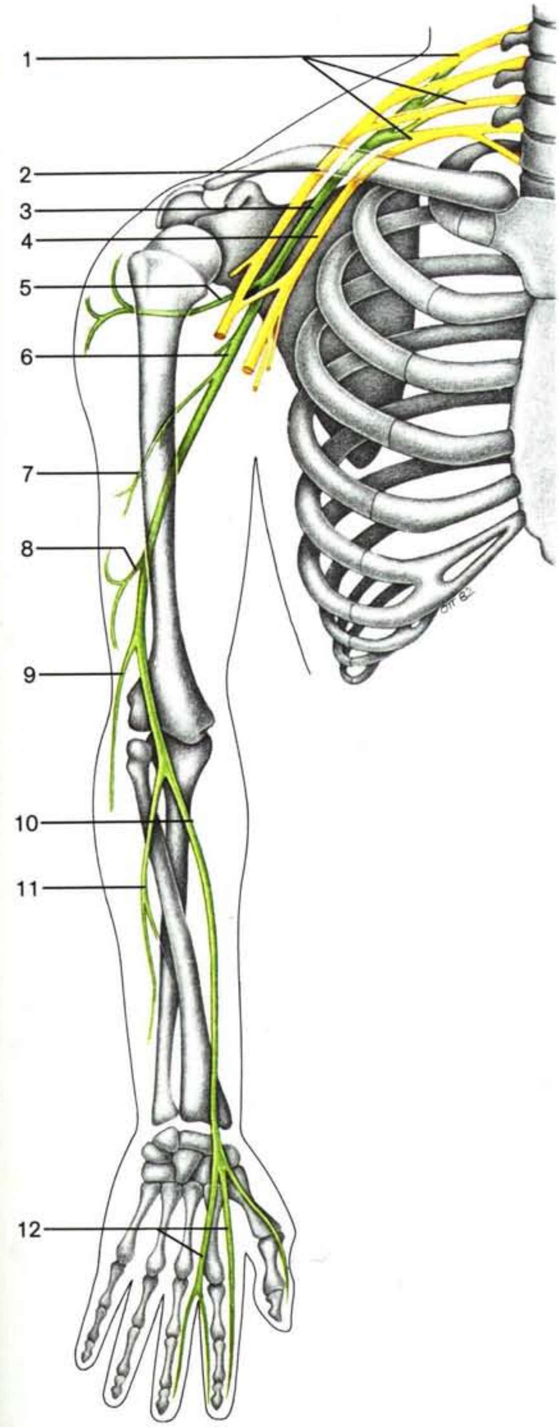

Main branches of radial nerve. (Schematic drawing.) Posterior divisions of trunks and posterior cord and its branches are indicated in green.

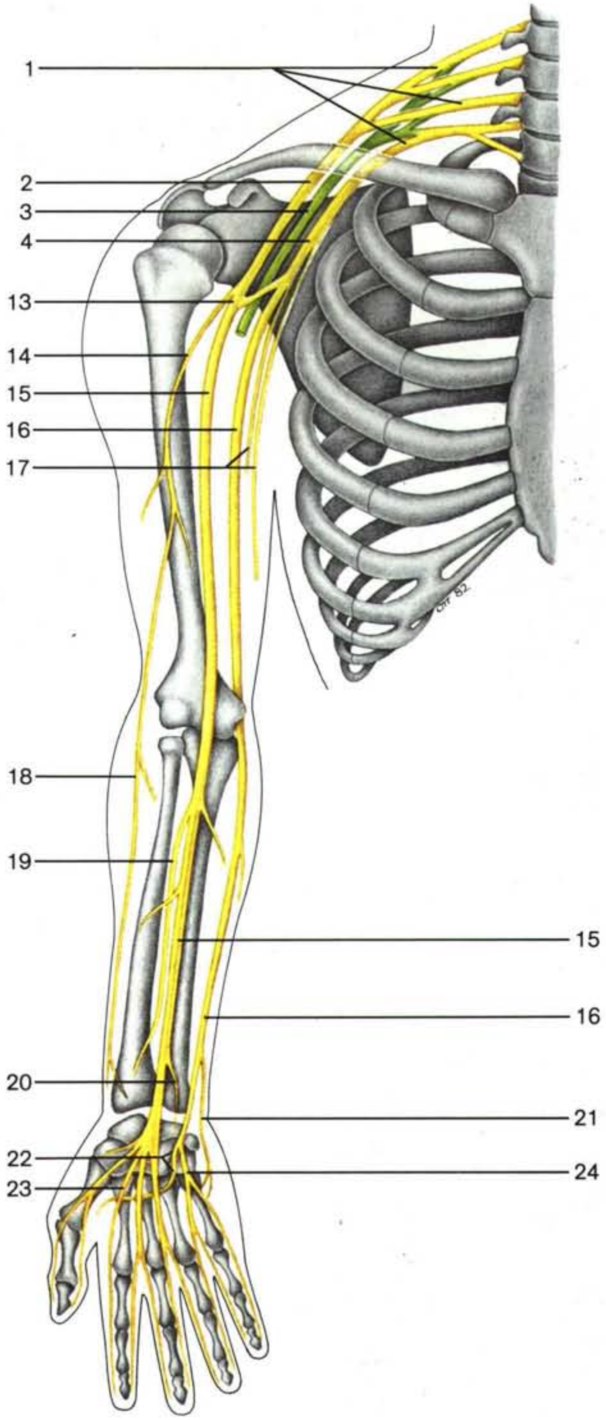

Main branches of musculocutaneous, median and ulnar nerves. (Schematic drawing.) Anterior divisions of the trunks and all the components arising from them are indicated in yellow.

1 Brachial plexus
2 Lateral cord of brachial plexus
3 Posterior cord of brachial plexus
4 Medial cord of brachial plexus
5 **Axillary nerve**
6 **Radial nerve**
7 Posterior cutaneous nerve of arm
8 Lower lateral cutaneous nerve of arm
9 Posterior cutaneous nerve of forearm
10 Superficial branch of radial nerve
11 Deep branch of radial nerve
12 Dorsal digital nerves

13 Roots of median nerve
14 Musculocutaneous nerve
15 **Median nerve**
16 **Ulnar nerve**
17 Medial cutaneous nerves of arm and forearm
18 Lateral cutaneous nerve of forearm
19 Anterior interosseous nerve
20 Palmar branch of median nerve
21 Dorsal branch of ulnar nerve
22 Deep branch of ulnar nerve
23 Common palmar digital nerves of median nerve
24 Superficial branch of ulnar nerve

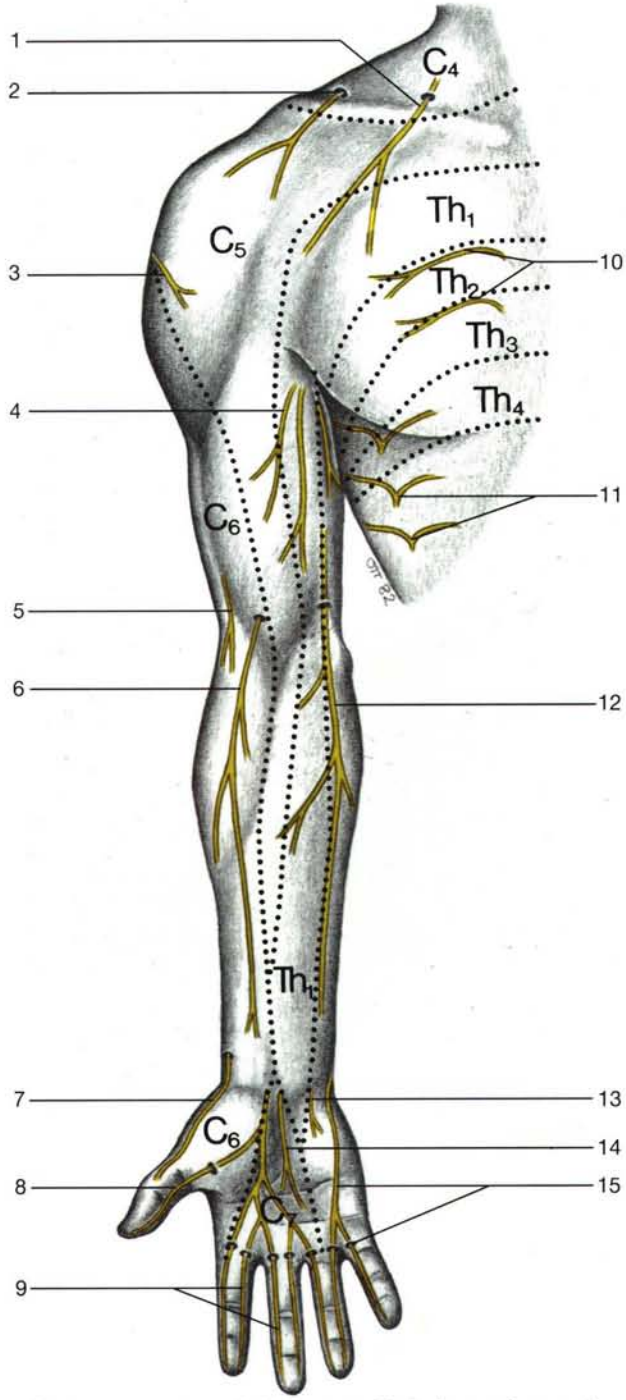

Cutaneous nerves of right upper limb (ventral aspect).
(Semischematic drawing.)

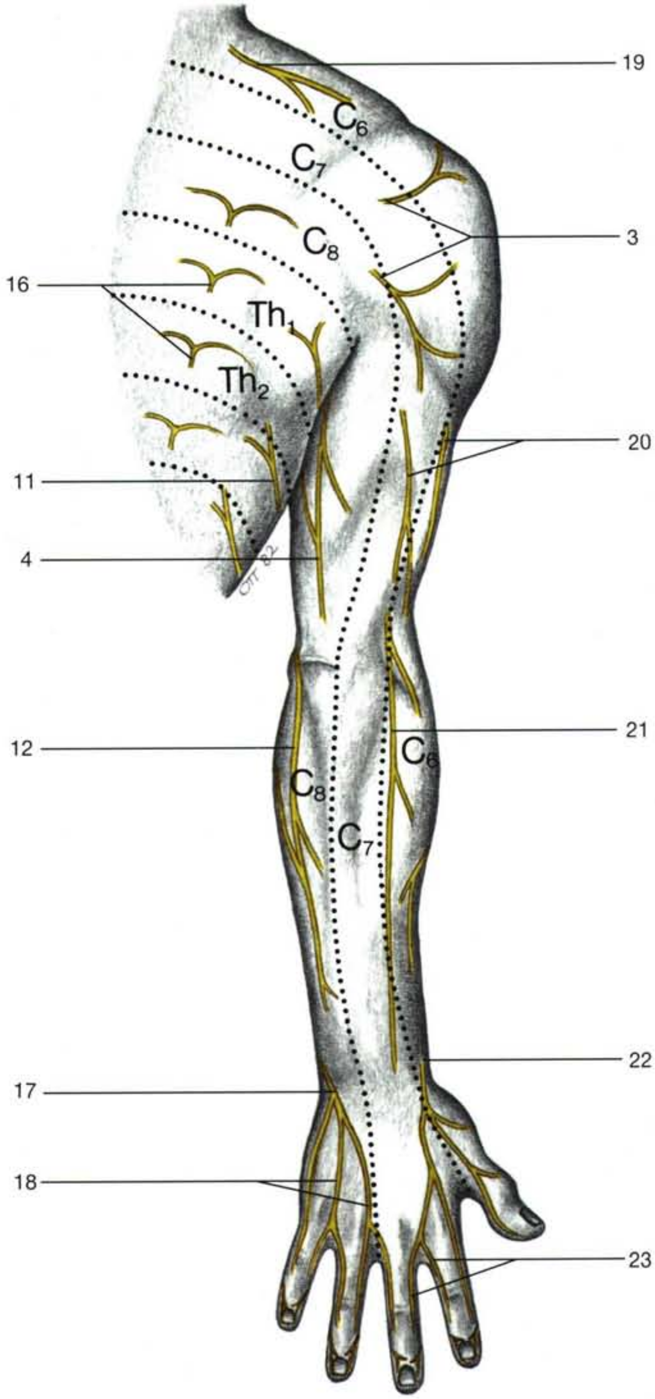

Cutaneous nerves of the upper limb (dorsal aspect).
(Semischematic drawing.)

1 Medial supraclavicular nerve
2 Intermediate supraclavicular nerve
3 Upper lateral cutaneous nerve of arm
4 Terminal branches of intercostobrachial nerves
5 Lower lateral cutaneous nerve of arm
6 Lateral cutaneous nerve of forearm
7 Terminal branch of superficial branch of radial nerve
8 Palmar digital nerve of thumb (branch of median nerve)
9 Palmar digital branches of median nerve
10 Anterior cutaneous branches of intercostal nerves
11 Lateral cutaneous branches of intercostal nerves
12 Medial cutaneous nerve of forearm

13 Palmar cutaneous branch of ulnar nerve
14 Palmar branch of median nerve
15 Palmar digital branches of ulnar nerve
16 Cutaneous branches of dorsal rami of spinal nerves
17 Dorsal branch of ulnar nerve
18 Dorsal digital nerves
19 Posterior supraclavicular nerve
20 Posterior cutaneous nerve of arm
21 Posterior cutaneous nerve of forearm
22 Superficial branch
23 Dorsal digital branches
} from radial nerve

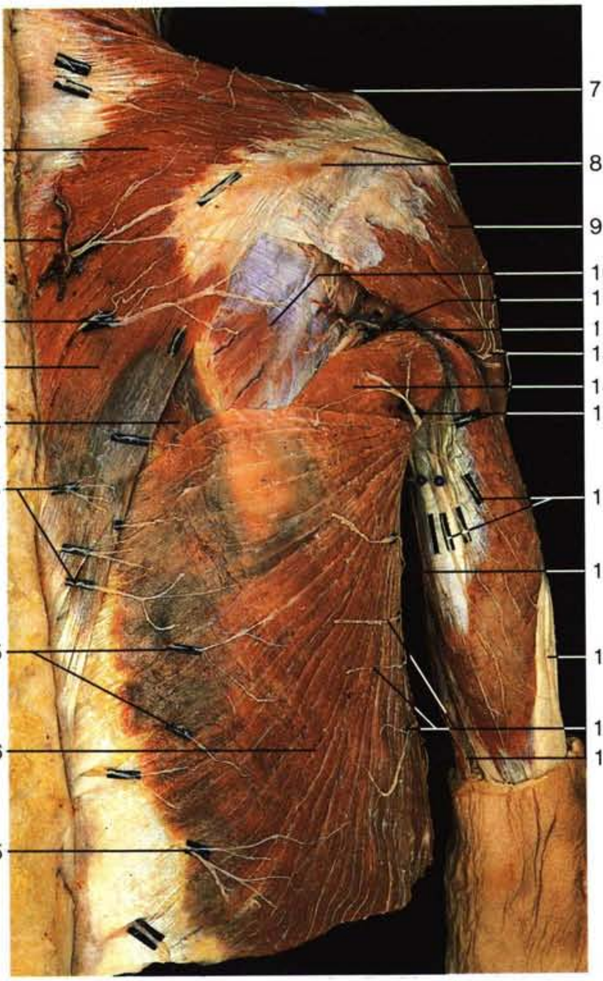

Dorsal region of shoulder, superficial layer. Note the segmental arrangement of the cutaneous nerves of the back.

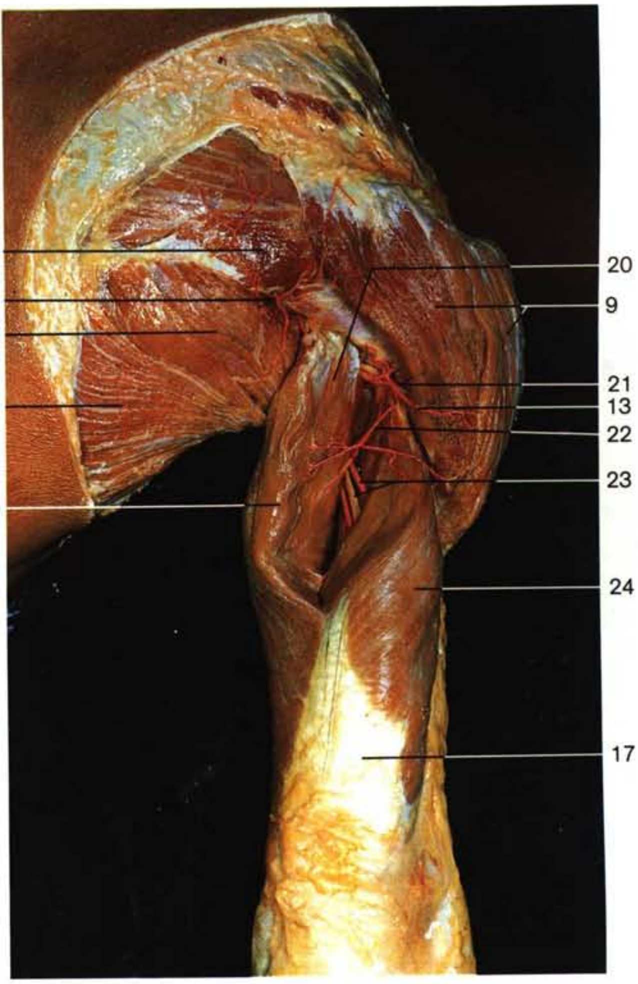

Shoulder and arm (dorsal aspect). Dissection of the quadrangular and triangular spaces of the axillary region.

1 Trapezius muscle
2 Dorsal branches of posterior intercostal artery and vein (medial cutaneous branches)
3 Medial branches of dorsal rami of spinal nerves
4 Rhomboid major muscle
5 Lateral branches of dorsal rami of spinal nerves
6 Latissimus dorsi muscle
7 Posterior supraclavicular nerves
8 Spine of scapula
9 Deltoid muscle
10 Infraspinatus muscle
11 Teres minor muscle
12 **Triangular space** with circumflex scapular artery and vein
13 Upper lateral cutaneous nerve of arm with artery
14 Teres major muscle
15 Terminal branches of intercostobrachial nerve
16 Medial cutaneous nerve of arm
17 Tendon of triceps brachii muscle
18 Lateral cutaneous branches of intercostal nerves
19 Medial cutaneous nerve of forearm
20 Long head of triceps brachii muscle
21 **Quadrangular space** with axillary nerve and posterior humeral circumflex artery
22 Anastomosis between profunda brachii artery and posterior humeral circumflex artery
23 Course of radial nerve and profunda brachii artery
24 Lateral head of triceps brachii muscle
25 Course of descending scapular artery and dorsal scapular nerve
26 Course of suprascapular nerve, artery and vein

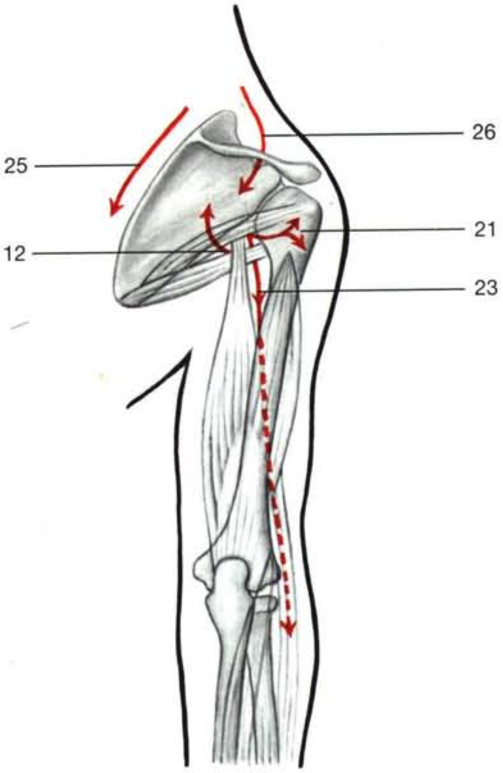

Course of vessels and nerves to shoulder and upper limb. (Schematic drawing.)

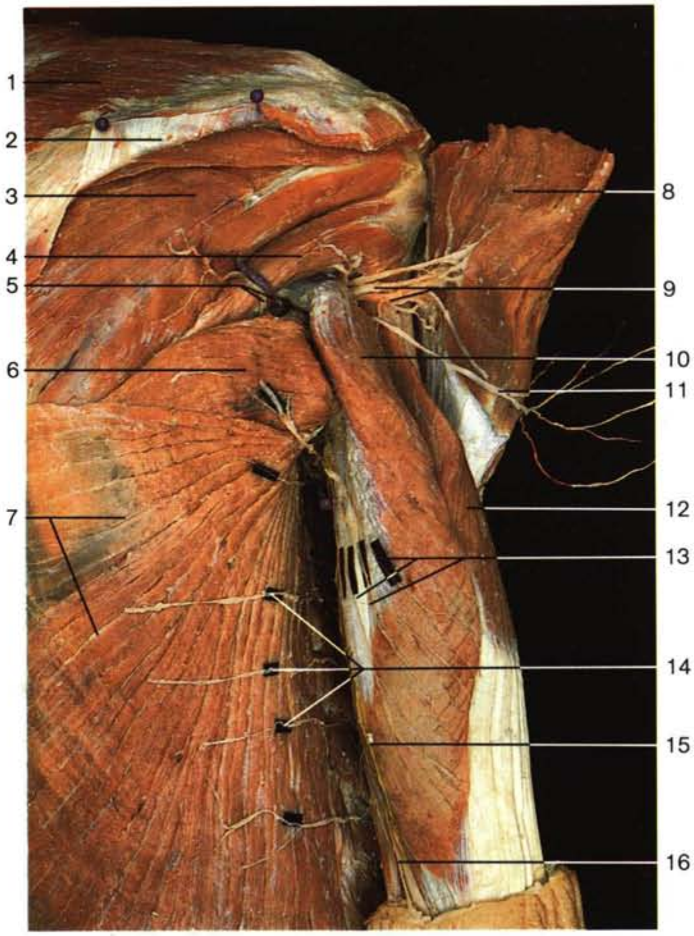

Scapular region, arm and shoulder, deep layer (dorsal aspect). Part of deltoid muscle has been cut and reflected to display the quadrangular and triangular spaces of the axillary region.

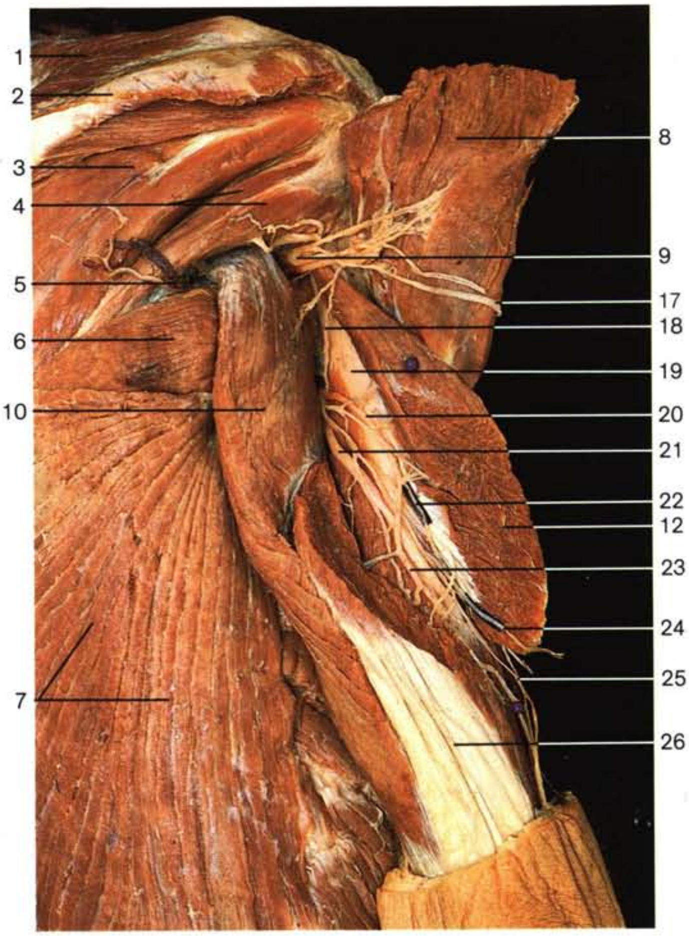

Scapular region and posterior brachial region; arm and shoulder, deep layer (dorsal aspect). The lateral head of the triceps brachii muscle has been cut to display the radial nerve and accompanying vessels.

1 Trapezius muscle
2 Spine of scapula
3 Infraspinatus muscle
4 Teres minor muscle
5 **Triangular space** containing circumflex scapular artery and vein
6 Teres major muscle
7 Latissimus dorsi muscle
8 Deltoid muscle (cut and reflected)
9 **Quadrangular space** containing axillary nerve and posterior circumflex humeral artery and vein
10 Long head of triceps brachii muscle
11 Cutaneous branch of axillary nerve
12 Lateral head of triceps brachii muscle
13 Terminal branches of intercostobrachial nerve
14 Lateral cutaneous branches of intercostal nerves
15 Medial cutaneous nerve of arm
16 Medial cutaneous nerve of forearm
17 Upper lateral cutaneous nerve of arm
18 Anastomosis between profunda brachii artery and posterior humeral circumflex artery
19 Humerus
20 **Profunda brachii artery**
21 **Radial nerve**
22 Radial collateral artery
23 Middle collateral artery
24 Lower lateral cutaneous nerve of arm
25 Posterior cutaneous nerve of forearm
26 Tendon of triceps brachii muscle

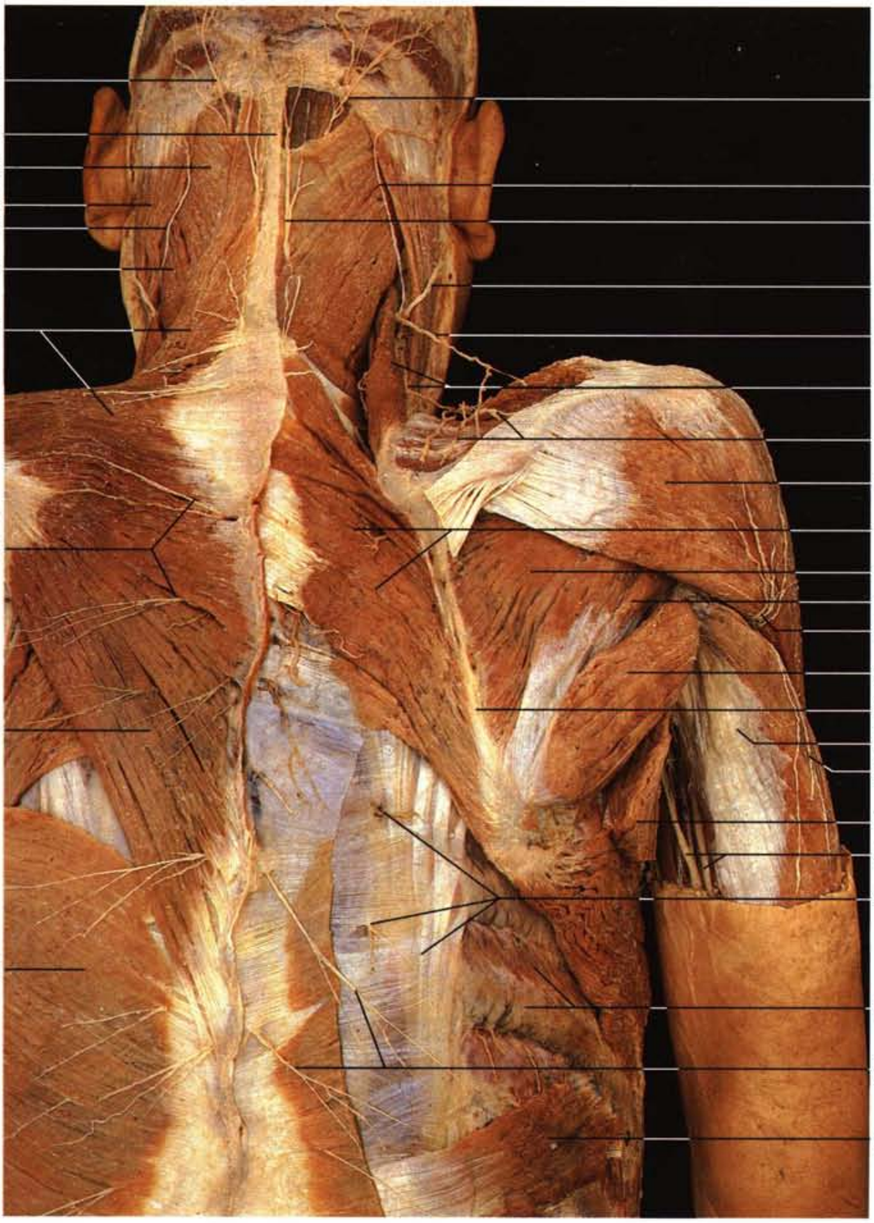

Dorsal regions of neck and shoulder (dorsal aspect). Left side: superficial layer. Right side: trapezius and latissimus dorsi muscles have been removed. Dissection of dorsal branches of spinal nerves.

1 **Greater occipital nerve**
2 Ligamentum nuchae
3 Splenius capitis muscle
4 Sternocleidomastoid muscle
5 Lesser occipital nerve
6 Splenius cervicis muscle
7 Descending and transverse fibers of trapezius muscle
8 **Medial cutaneous branches of dorsal rami of spinal nerves**
9 Ascending fibers of trapezius muscle
10 Latissimus dorsi muscle
11 Cutaneous branch of third occipital nerve
12 Great auricular nerve
13 Accessory nerve (n. XI)
14 Posterior supraclavicular nerve and levator scapulae muscle
15 Branches of suprascapular artery
16 Deltoid muscle
17 Rhomboid major muscle
18 Infraspinatus muscle
19 Teres minor muscle
20 Upper lateral cutaneous nerve of arm (branch of axillary nerve)
21 Teres major muscle
22 Medial margin of scapula
23 Long head of triceps muscle
24 Posterior cutaneous nerve of arm (branch of radial nerve)
25 Latissimus dorsi muscle (divided)
26 Ulnar nerve and brachial artery
27 **Lateral cutaneous branches of dorsal rami of spinal nerves** and iliocostalis thoracis muscle
28 External intercostal muscle and seventh rib
29 Serratus posterior inferior muscle

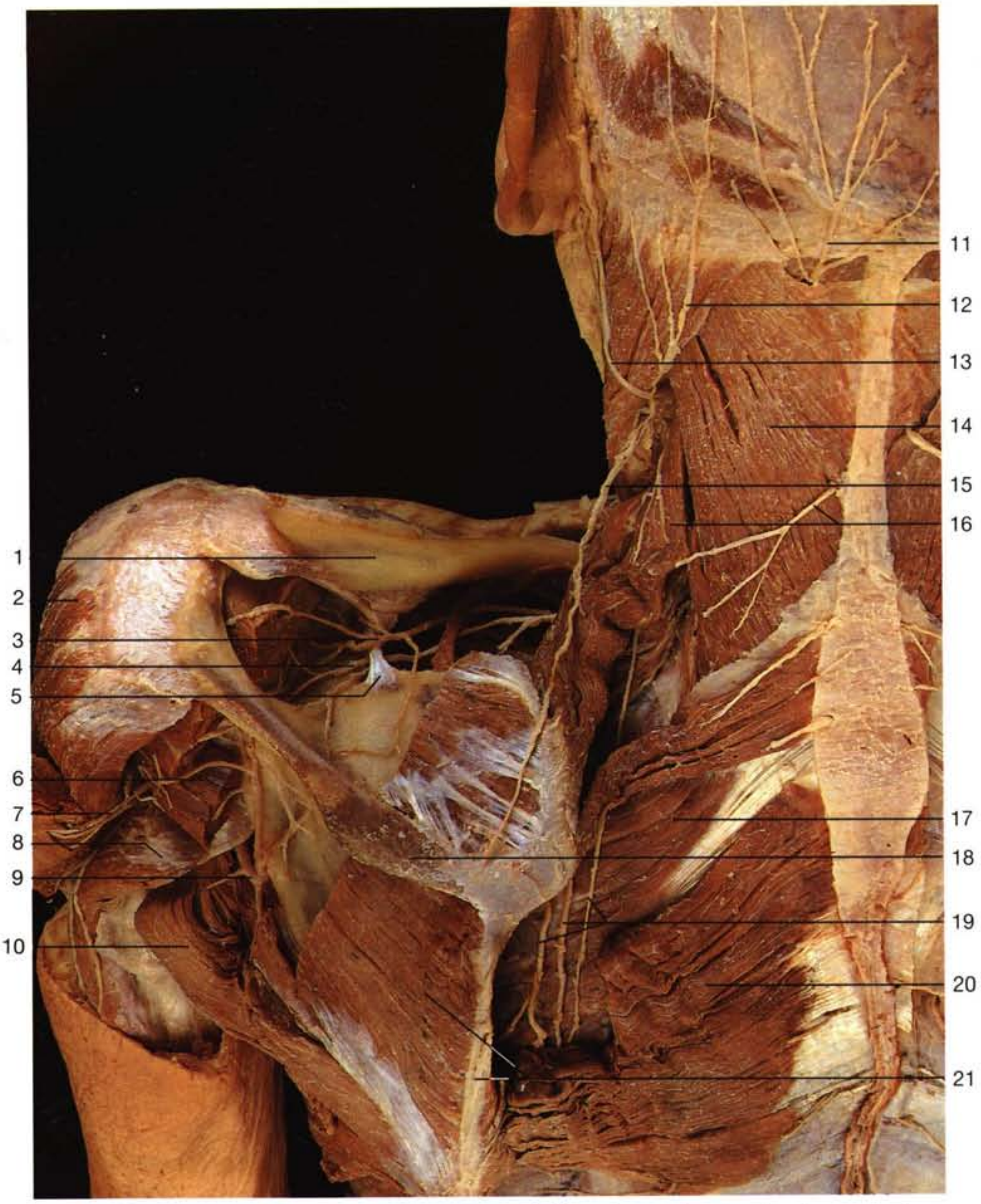

Dorsal region of shoulder; deepest layer. Rhomboid and scapular muscles fenestrated; posterior part of deltoid muscle reflected.

1 Clavicle
2 Deltoid muscle
3 **Suprascapular artery**
4 Suprascapular nerve
5 Superior transverse scapular ligament
6 Teres minor muscle
7 Axillary nerve and posterior circumflex humeral artery
8 Long head of triceps muscle
9 **Circumflex scapular artery**
10 Teres major muscle
11 Greater occipital nerve
12 Lesser occipital nerve
13 Great auricular nerve
14 Splenius capitis muscle
15 Accessory nerve (n. XI)
16 Third occipital nerve and levator scapulae muscle
17 Serratus posterior superior muscle
18 Spine of scapula
19 Descending scapular artery and dorsal scapular nerve
20 Rhomboid major muscle
21 Infraspinatus muscle and medial margin of scapula
22 Radial nerve and profunda brachii artery
23 Thoracodorsal artery
24 Thyrocervical trunk
25 Roots of brachial plexus

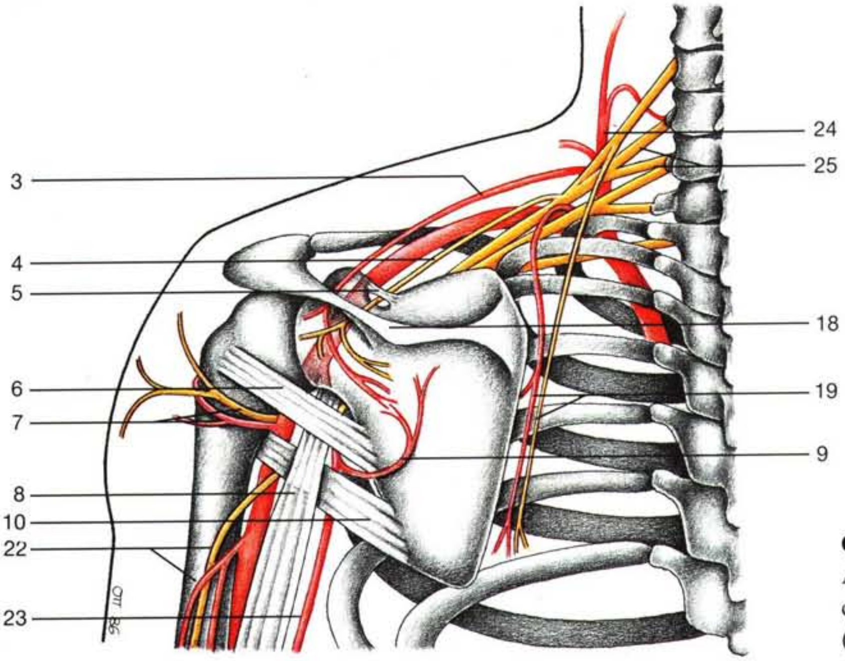

Collateral circulation of shoulder. Anastomosis of suprascapular and circumflex scapular arteries. (Semischematic drawing.)

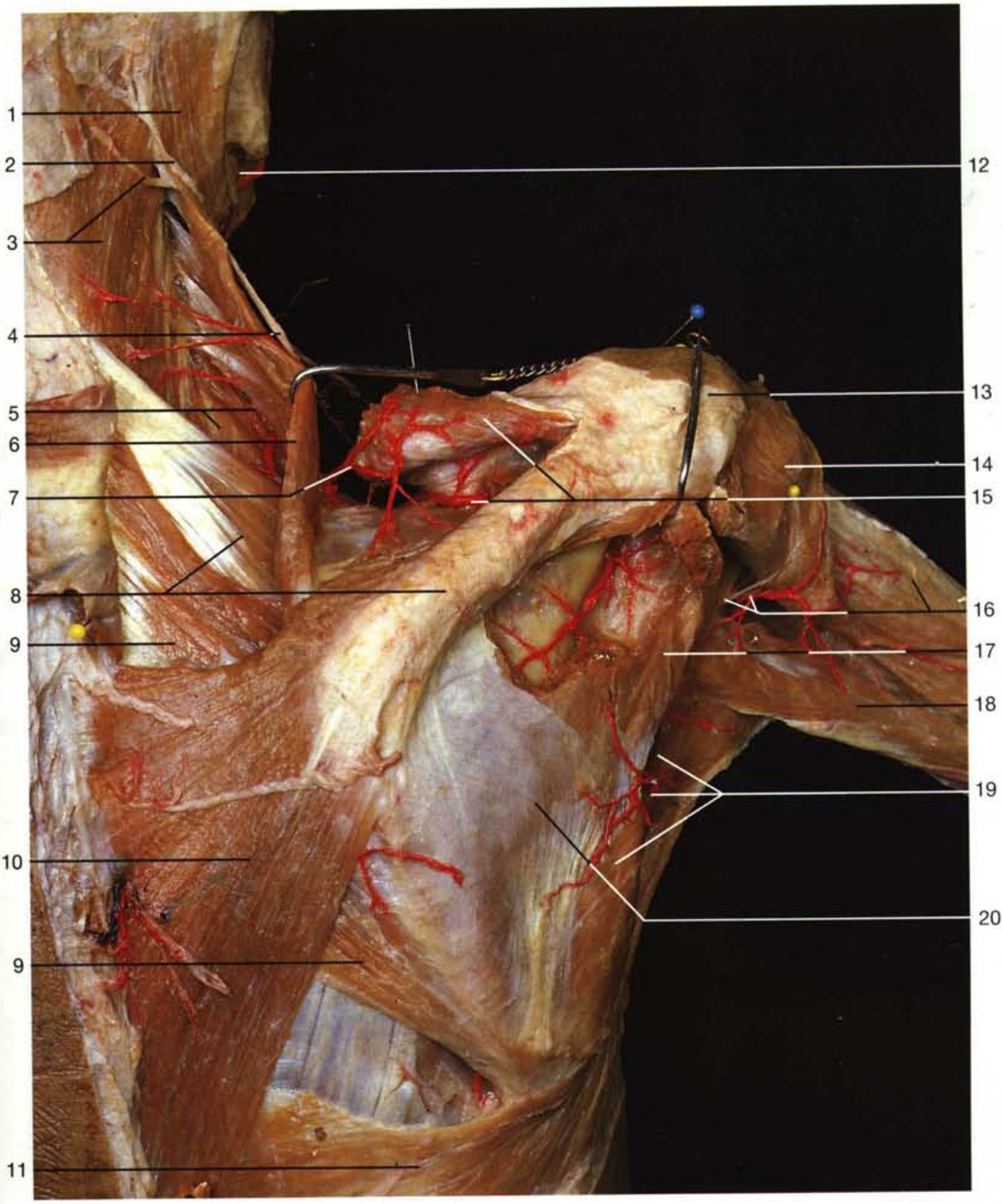

Scapular region, arm and shoulder (dorsal aspect). Arteries of scapular region are injected. Trapezius, deltoid and infraspinatus muscles are partially removed or reflected.

1	Sternocleidomastoid muscle	11	Latissimus dorsi muscle
2	Lesser occipital nerve	12	Facial artery
3	Splenius capitis muscle and third occipital nerve	13	Acromion
4	**Accessory nerve** (N. XI)	14	Deltoid muscle
5	Splenius cervicis muscle and transverse cervical artery (deep branch)	15	**Suprascapular artery** and supraspinatus muscle (reflected)
6	Levator of scapula muscle	16	**Axillary nerve,** posterior circumflex humeral artery and lateral head of triceps, brachii muscle
7	**Transverse cervical artery** (superficial branch)	17	Teres minor muscle
8	Spine of scapula and serratus posterior superior muscle	18	Long head of triceps brachii muscle
9	Rhomboid major muscle	19	**Circumflex scapular artery** and teres major
10	Trapezius muscle	20	Infraspinatus muscle

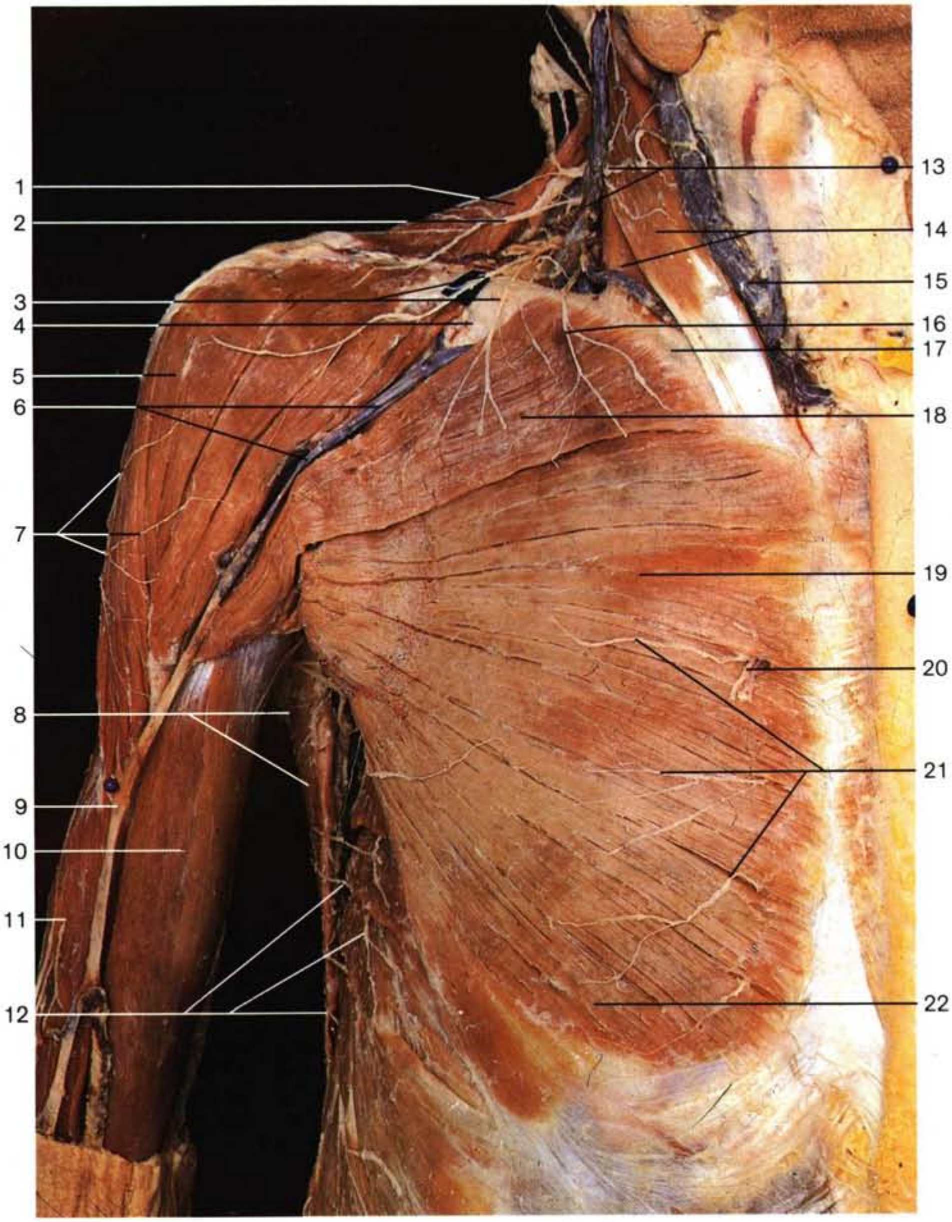

Right shoulder and thoracic wall, superficial layer (anterior aspect). Dissection of the cutaneous nerves and veins.

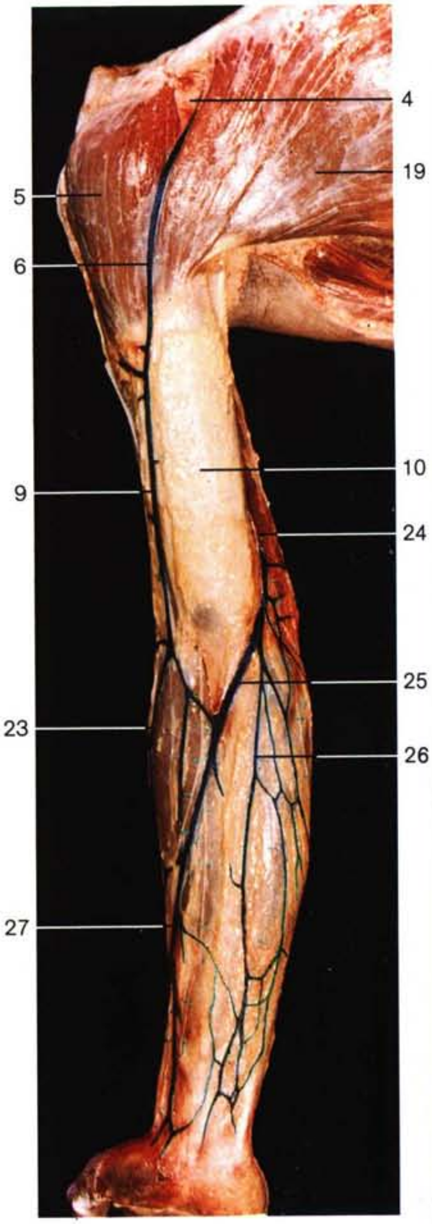

Superficial veins of right arm have been injected with blue gelatin.

1 Trapezius muscle
2 Posterior supraclavicular nerve
3 Middle supraclavicular nerve
4 Deltopectoral triangle
5 Deltoid muscle
6 Cephalic vein within the deltopectoral groove
7 Upper lateral cutaneous nerve of arm (branch of axillary nerve)
8 Latissimus dorsi muscle
9 **Cephalic vein**
10 Biceps brachii muscle
11 Triceps brachii muscle
12 Lateral cutaneous branches of intercostal nerves
13 Transverse cervical nerve and external jugular vein
14 Sternocleidomastoid muscle
15 **Anterior jugular vein**
16 Anterior supraclavicular nerve
17 Clavicle
18 Clavicular part of pectoralis major muscle
19 Sternocostal part of pectoralis major muscle
20 Perforating branch of internal thoracic artery
21 **Anterior cutaneous branches of intercostal nerves**
22 Abdominal part of pectoralis major muscle
23 Accessory cephalic vein
24 **Basilic vein**
25 **Median cubital vein**
26 Median vein of forearm
27 Cephalic vein in the forearm

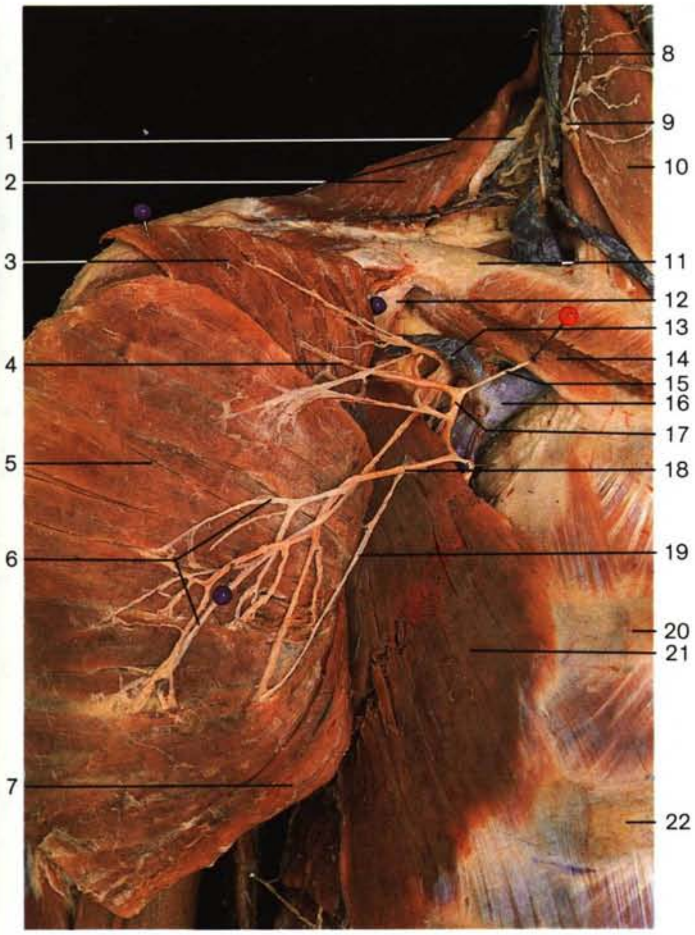

Right deltopectoral triangle, infraclavicular region (anterior aspect). The pectoralis major muscle has been cut and reflected.

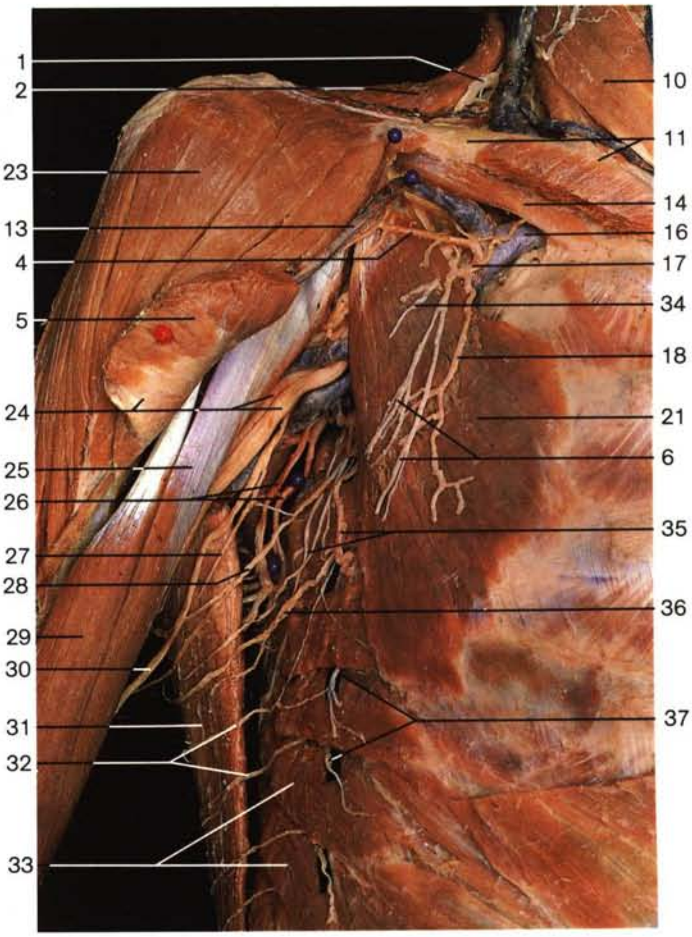

Thoracic wall and shoulder, deep layer. **Right axillary region** (anterior aspect). The pectoralis major muscle has been cut and partly removed.

1 Accessory nerve
2 Trapezius muscle
3 Pectoralis major muscle (clavicular part)
4 Acromial branch of thoracoacromial artery
5 Pectoralis major muscle
6 **Lateral pectoral nerves**
7 Abdominal part of pectoralis major muscle
8 External jugular vein
9 Cutaneous branches of cervical plexus
10 Sternocleidomastoid muscle
11 Clavicle
12 Clavipectoral fascia
13 **Cephalic vein**
14 Subclavius muscle
15 Clavicular branch of thoracoacromial artery
16 **Subclavian vein**
17 **Thoracoacromial artery**
18 Pectoral branch of thoracoacromial artery
19 Medial pectoral nerve
20 2nd rib
21 Pectoralis minor muscle
22 3rd rib
23 Deltoid muscle
24 Pectoralis major muscle (reflected), **brachial artery and median nerve**
25 Short head of biceps brachii muscle
26 Thoracodorsal artery and nerve
27 Medial cutaneous nerve of arm
28 Intercostobrachial nerve (T_2)
29 Long head of biceps brachii muscle
30 Medial cutaneous nerve of forearm
31 Latissimus dorsi muscle
32 Lateral cutaneous branches of intercostal nerves (posterior branches)
33 Serratus anterior muscle
34 Medial pectoral nerve
35 **Long thoracic nerve and lateral thoracic artery**
36 Intercostobrachial nerve (T_3)
37 **Lateral cutaneous branches of intercostal nerves** (anterior branches)

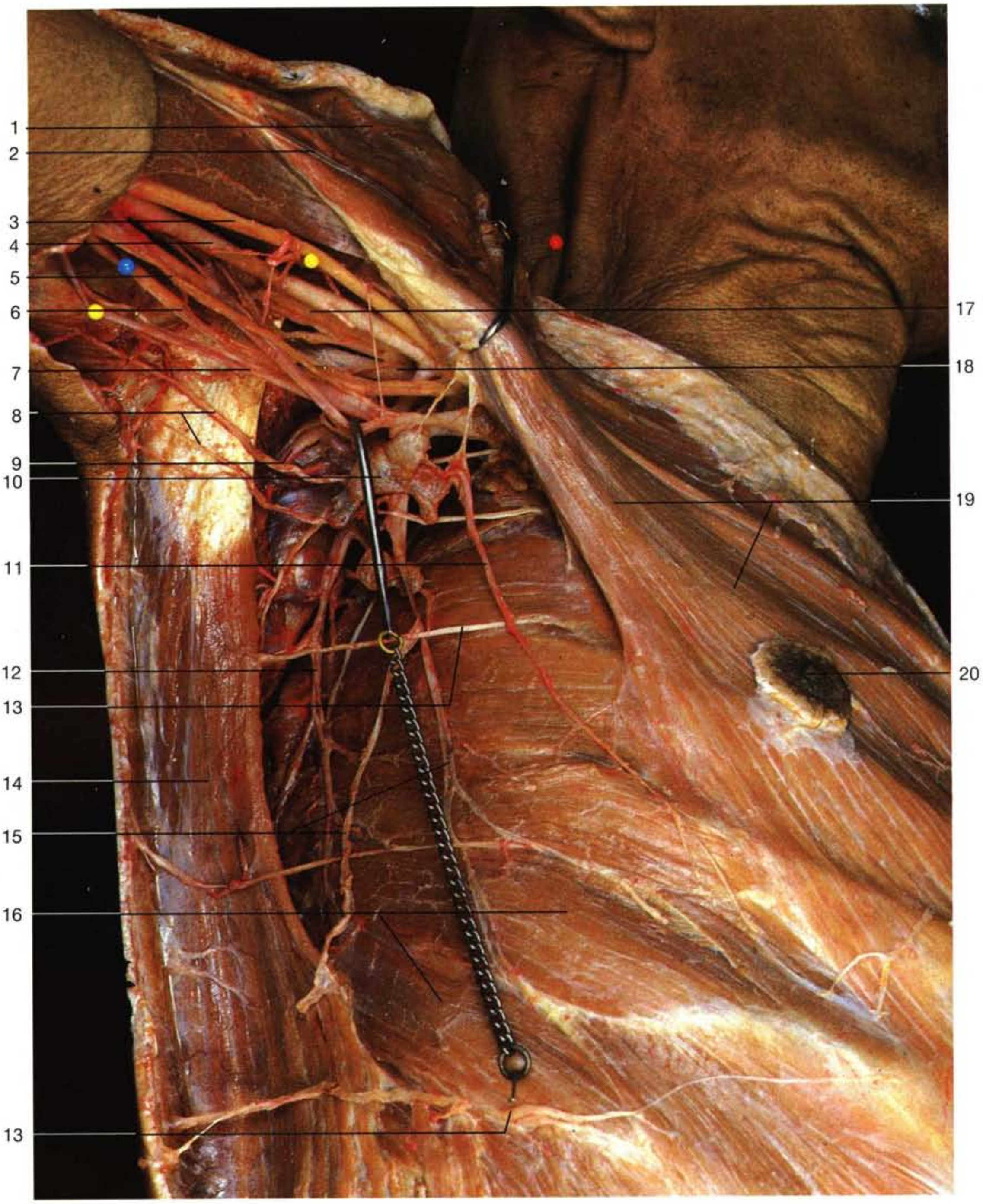

Right axillary region (inferior aspect). **Dissection of superficial axillary nodes and lymphatic vessels.** The pectoralis major muscle has been slightly elevated.

1	Deltoid muscle	12	Thoracodorsal artery
2	Cephalic vein	13	Lateral cutaneous branch of intercostal nerve
3	Median nerve	14	Latissimus dorsi muscle
4	Brachial artery	15	Thoracoepigastric vein
5	Medial cutaneous nerves of arm and forearm	16	Serratus anterior muscle
6	Ulnar nerve	17	Musculocutaneous nerve
7	Basilic Vein	18	Radial nerve
8	Intercostobrachial nerves	19	Pectoralis major muscle
9	Circumflex scapular artery	20	Nipple
10	**Superficial axillary nodes**		
11	Lateral thoracic artery		

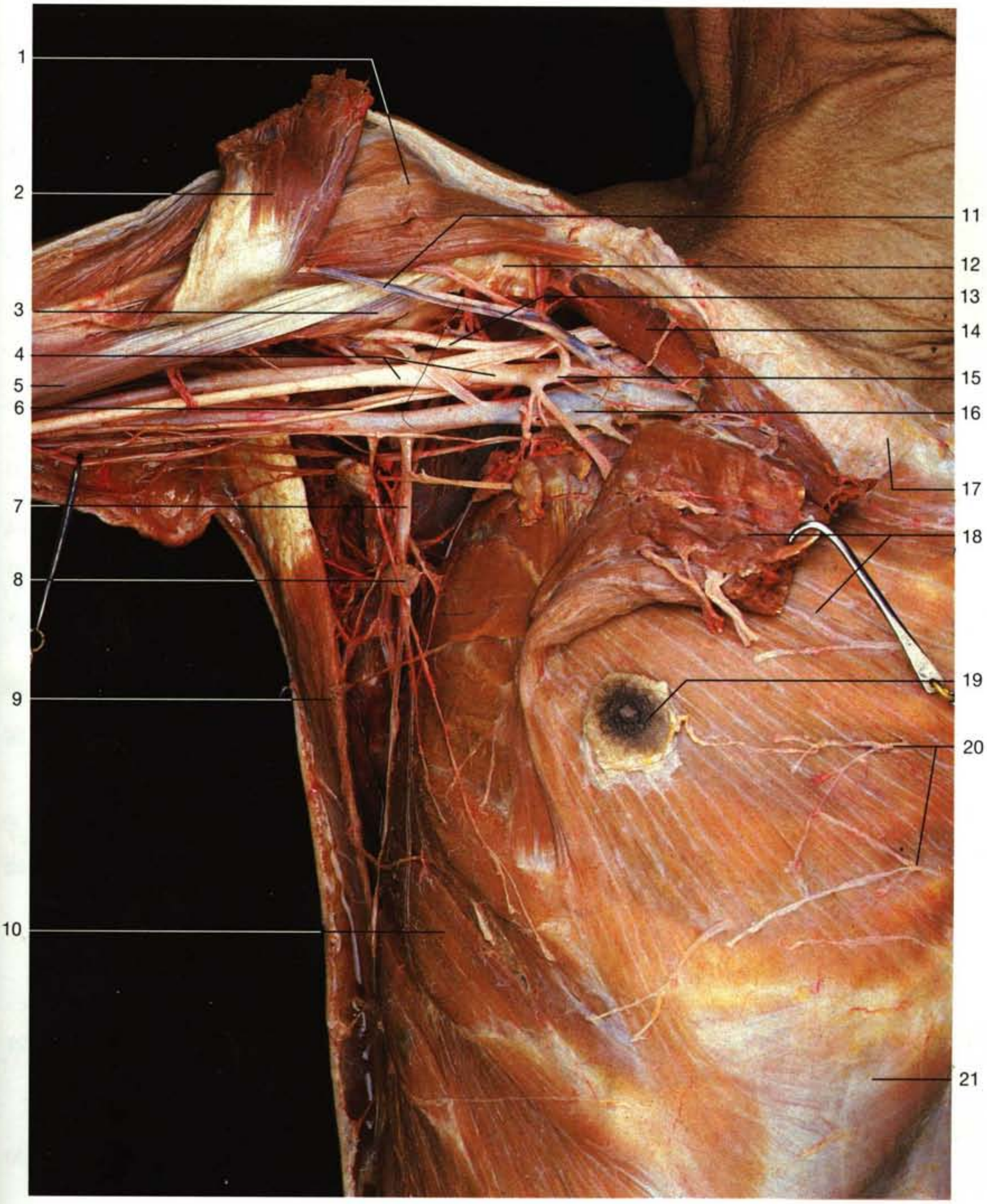

Axillary region (anterior aspect). Dissection of deep axillary nodes. Pectoralis major and minor divided and reflected. Shoulder girdle and arm elevated and reflected.

1 Deltoid muscle
2 Insertion of pectoralis major muscle
3 Coracobrachialis muscle
4 Roots of **median nerve, axillary artery**
5 Short head of biceps brachii muscle
6 **Ulnar nerve** and medial cutaneous nerve of forearm
7 Thoracoepigastric vein
8 Deep axillary node
9 Latissimus dorsi muscle
10 Serratus anterior muscle
11 Cephalic vein
12 Insertion of pectoralis minor muscle (coracoid process)
13 **Musculocutaneous nerve**
14 Subclavius muscle
15 Thoracoacromial artery
16 **Axillary vein**
17 Clavicle
18 Pectoralis major and minor muscles (reflected)
19 Nipple
20 Anterior cutaneous branches of intercostal nerves
21 Anterior layer of rectus sheaths

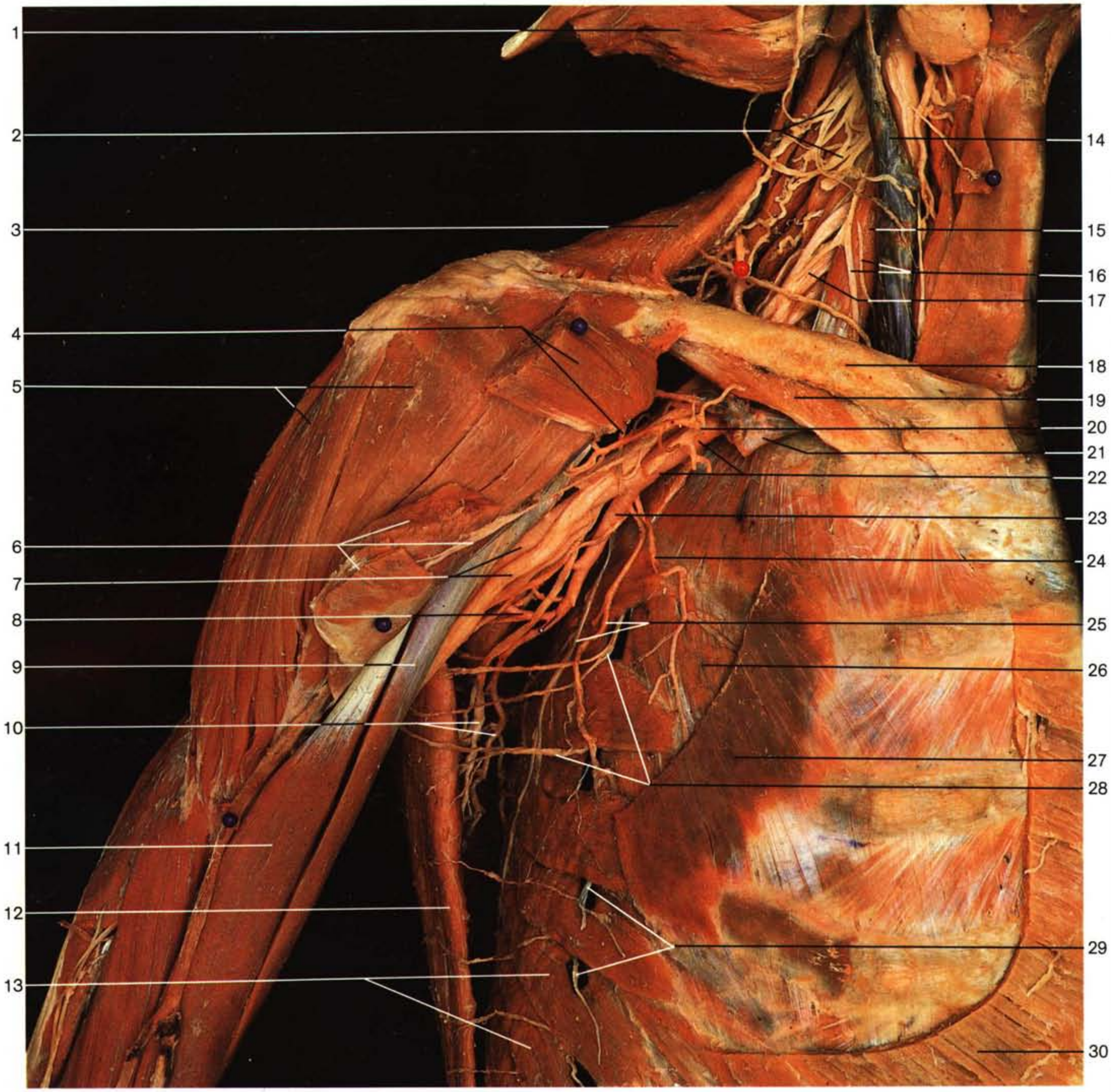

Right axillary region (anterior aspect). The pectoralis major and minor muscles have been cut and reflected to display the vessels and nerves of the axilla.

1 Sternocleidomastoid muscle (cut and reflected)
2 Cervical plexus
3 Trapezius muscle
4 Pectoralis minor muscle and medial pectoral nerve
5 Deltoid muscle
6 Pectoralis major muscle and lateral pectoral nerve
7 **Median nerve and brachial artery**
8 Circumflex scapular artery
9 Short head of biceps brachii muscle
10 **Thoracodorsal artery and nerve**
11 Long head of biceps brachii muscle
12 Latissimus dorsi muscle
13 Serratus anterior muscle
14 Internal jugular vein
15 **Scalenus anterior muscle**
16 Phrenic nerve and ascending cervical artery
17 Brachial plexus (at the levels of the trunks)
18 Clavicle
19 Subclavius muscle
20 **Thoracoacromial artery**
21 Subclavian vein (cut)
22 **Axillary artery**
23 Subscapular artery
24 Superior thoracic artery
25 **Lateral thoracic artery and long thoracic nerve**
26 External intercostal muscle
27 Insertion of pectoralis minor muscle
28 Intercostobrachial nerves
29 Lateral cutaneous branches of intercostal nerves
30 Insertion of pectoralis major muscle

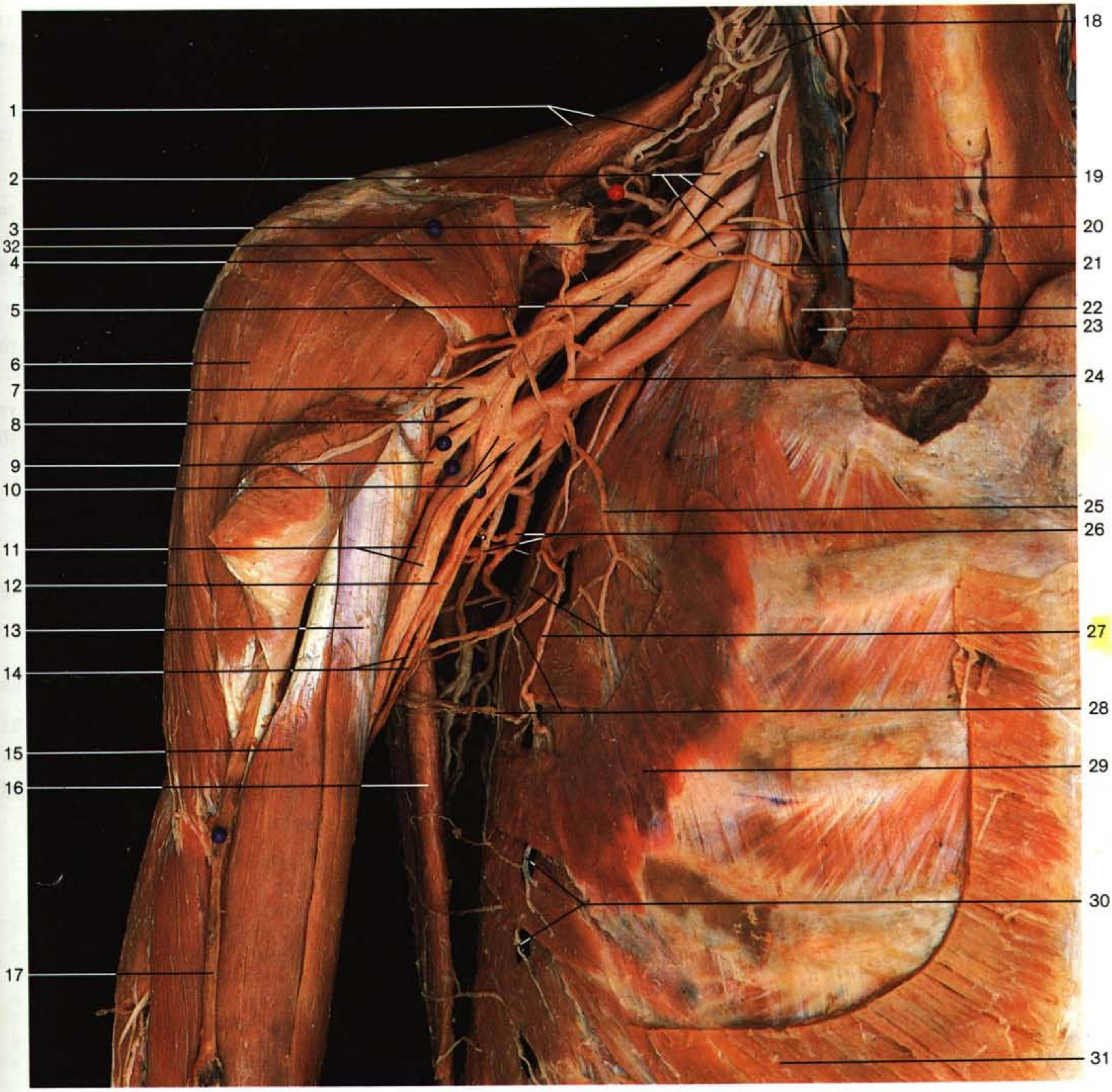

Shoulder and arm. Axillary region with brachial plexus (anterior aspect). The pectoralis major and minor muscles have been partly removed. The clavicle has been cut.

1	Trapezius muscle and accessory nerve	17	Cephalic vein
2	Brachial plexus	18	Cervical plexus
3	Clavicle (cut)	19	**Phrenic nerve and scalenus anterior muscle**
4	Pectoralis minor muscle	20	Descending scapular artery
5	**Axillary (subclavian) artery**	21	Suprascapular artery
6	Deltoid muscle	22	Internal thoracic artery
7	**Musculocutaneous nerve**	23	Subclavian vein
8	**Axillary nerve**	24	Thoracoacromial artery
9	**Radial nerve**	25	Superior thoracic artery
10	Medial and lateral roots of the median nerve	26	**Thoracodorsal artery and nerve**
11	**Median nerve** and **brachial artery**	27	**Lateral thoracic artery and long thoracic nerve**
12	**Ulnar nerve**	28	Intercostobrachial nerves
13	Short head of biceps brachii muscle	29	Insertion of the pectoralis minor muscle into the 2nd–5th ribs
14	Medial cutaneous nerves of arm and forearm	30	Lateral cutaneous branches of intercostal nerves
15	Long head of biceps brachii muscle	31	Abdominal part of pectoralis major muscle (cut)
16	Latissimus dorsi muscle	32	Suprascapular nerve and artery

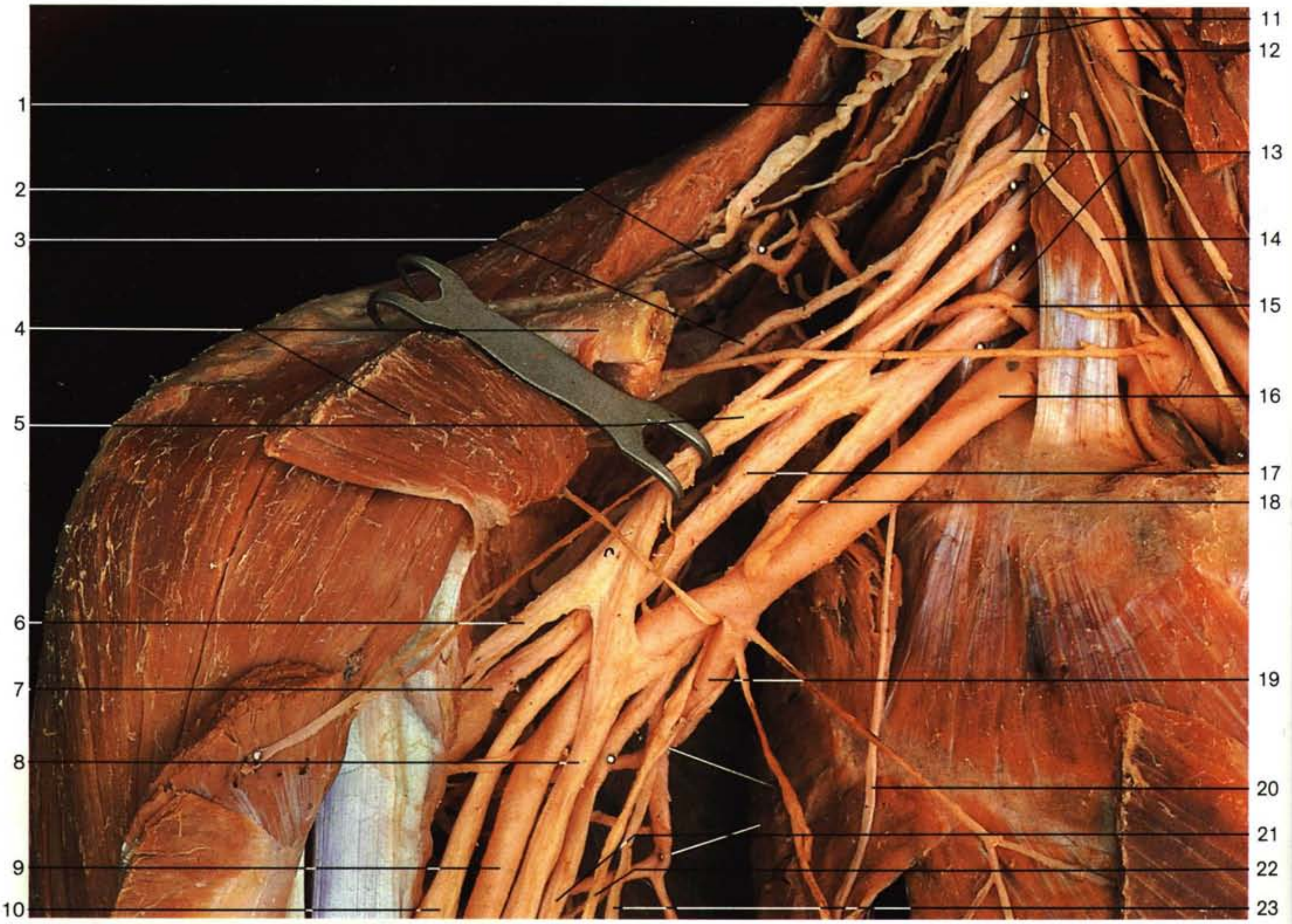

Brachial plexus (anterior aspect). Clavicle and the two pectoralis muscles have been partly removed.

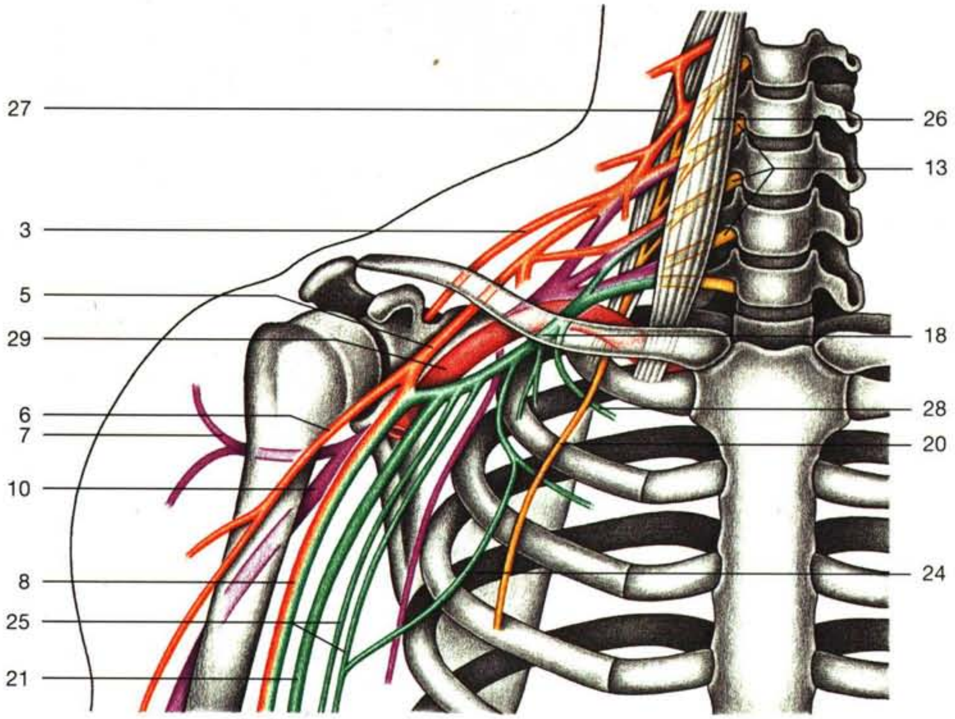

Main branches of brachial plexus. Posterior cord in red, lateral cord in yellow and medial cord in purple. (Schematic drawing.)

1 Accessory nerve
2 Dorsal scapular artery
3 Suprascapular nerve
4 Clavicle and pectoralis minor muscle
5 **Lateral cord** of brachial plexus
6 Musculocutaneous nerve
7 Axillary nerve
8 Median nerve
9 Brachial artery
10 Radial nerve
11 Cervical plexus
12 Common carotid artery
13 Roots of brachial plexus (C_5–T_1)
14 Phrenic nerve
15 Dorsal scapular artery
16 Subclavian artery
17 **Posterior cord** of brachial plexus
18 **Medial cord** of brachial plexus
19 Subscapular artery
20 Long thoracic nerve
21 Ulnar nerve
22 Medial cutaneous nerve of forearm
23 Thoracodorsal nerve
24 Intercostobrachial nerve
25 Medial cutaneous nerves of arm and forearm
26 Scalenus anterior muscle
27 Scalenus medius muscle
28 Intercostal nerve (T_1)
29 Axillary artery

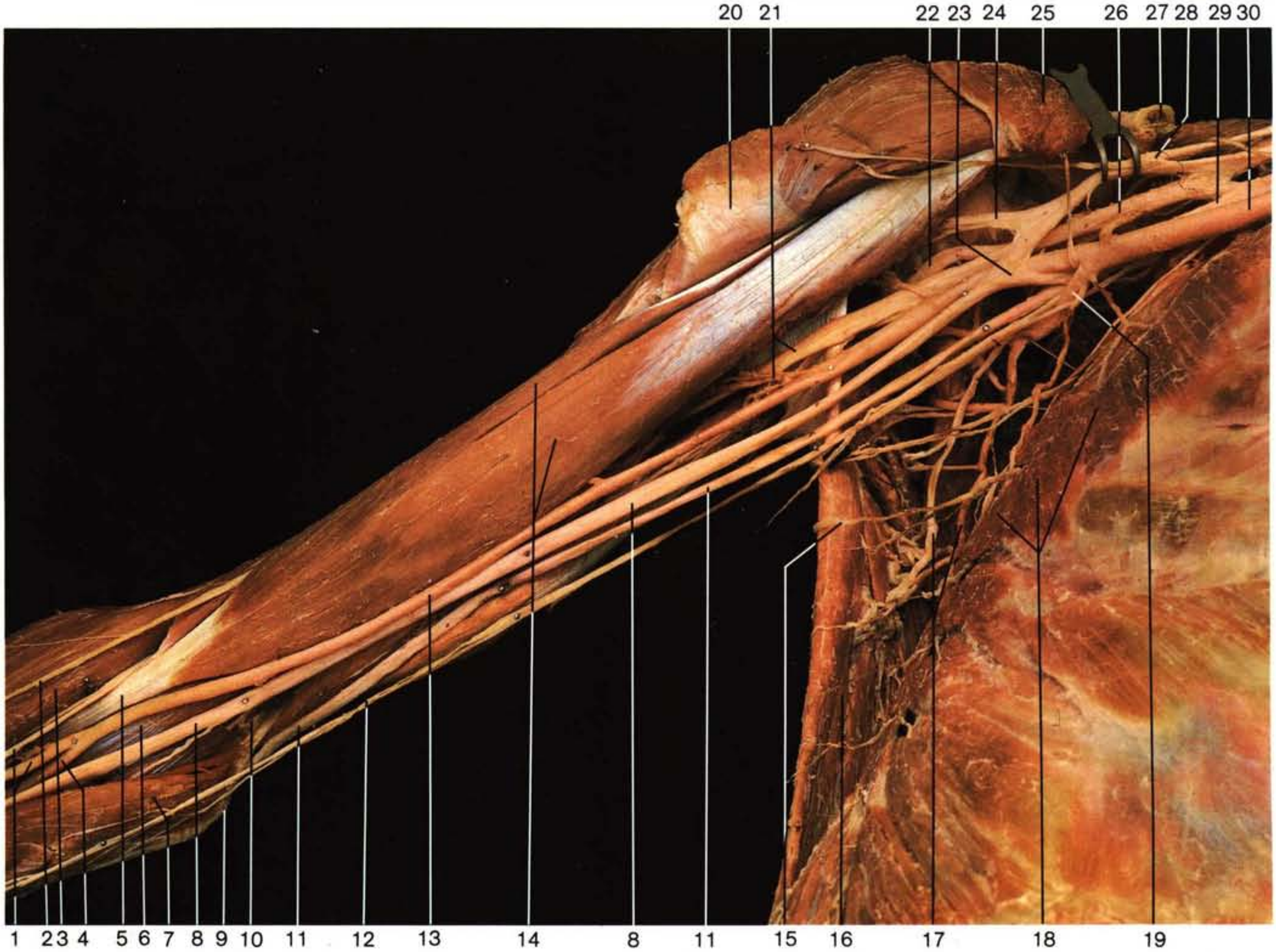

Right arm. Dissection of vessels and nerves (medial aspect). Shoulder girdle has been reflected slightly.

1 Radial artery and superficial branch of radial nerve
2 Lateral cutaneous nerve of forearm
3 Brachioradialis muscle
4 Ulnar artery
5 Tendon of biceps brachii muscle
6 Brachialis muscle
7 Pronator teres muscle
8 **Median nerve**
9 Medial epicondyle of humerus
10 Inferior ulnar collateral artery
11 **Ulnar nerve**
12 Medial cutaneous nerve of forearm
13 **Brachial artery**
14 Biceps brachii muscle
15 Intercostobrachial nerve (T_3)
16 Latissimus dorsi muscle
17 Thoracodorsal nerve and artery
18 Serratus anterior muscle
19 Subscapular artery
20 Pectoralis major muscle (reflected) and lateral pectoral nerve
21 Radial nerve and profunda brachii artery
22 Axillary nerve
23 Roots of the median nerve with axillary artery
24 **Musculocutaneous nerve**
25 Pectoralis minor muscle (reflected) and medial pectoral nerve
26 Posterior cord of brachial plexus
27 Clavicle (cut)
28 Lateral cord of brachial plexus
29 Medial cord of brachial plexus
30 Subclavian artery
31 Brachial vein

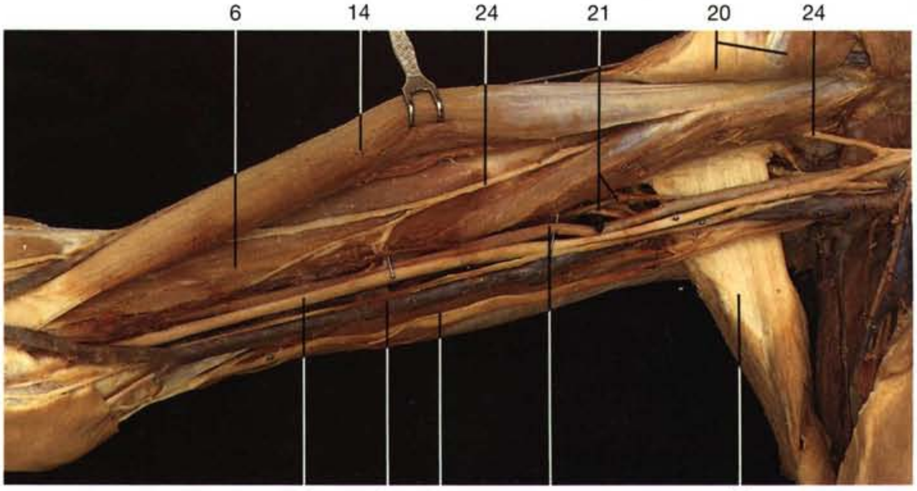

Right arm. Dissection of vessels and nerves, deeper layer. Biceps muscle has been reflected.

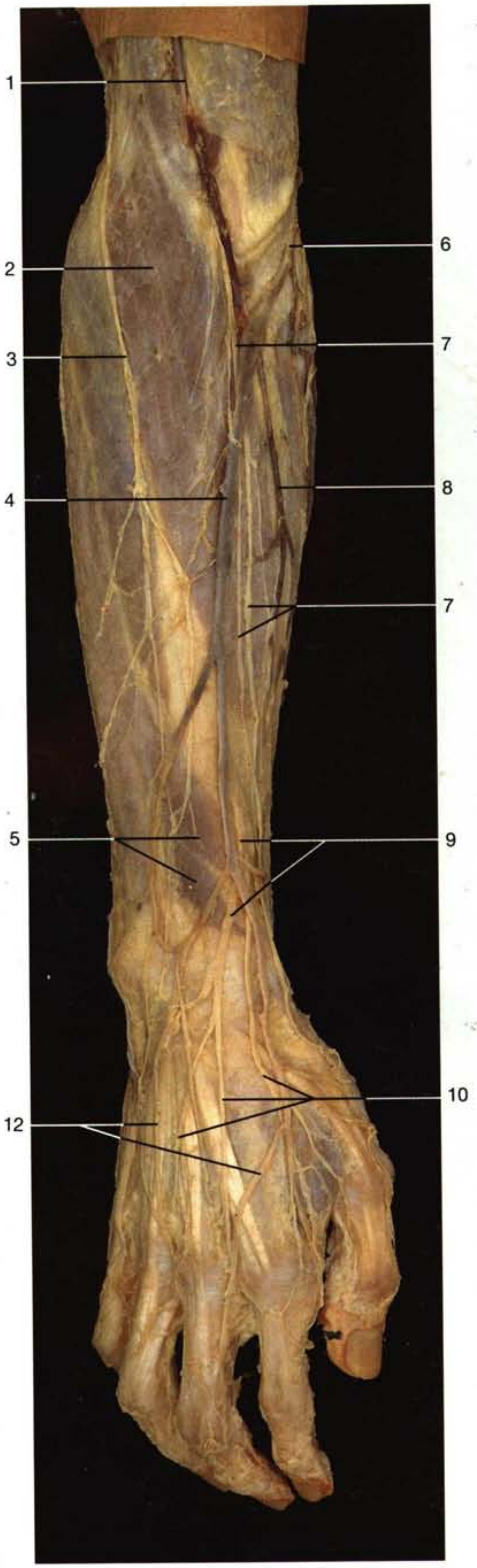

Superficial veins and cutaneous nerves of forearm and hand (anteromedial aspect).

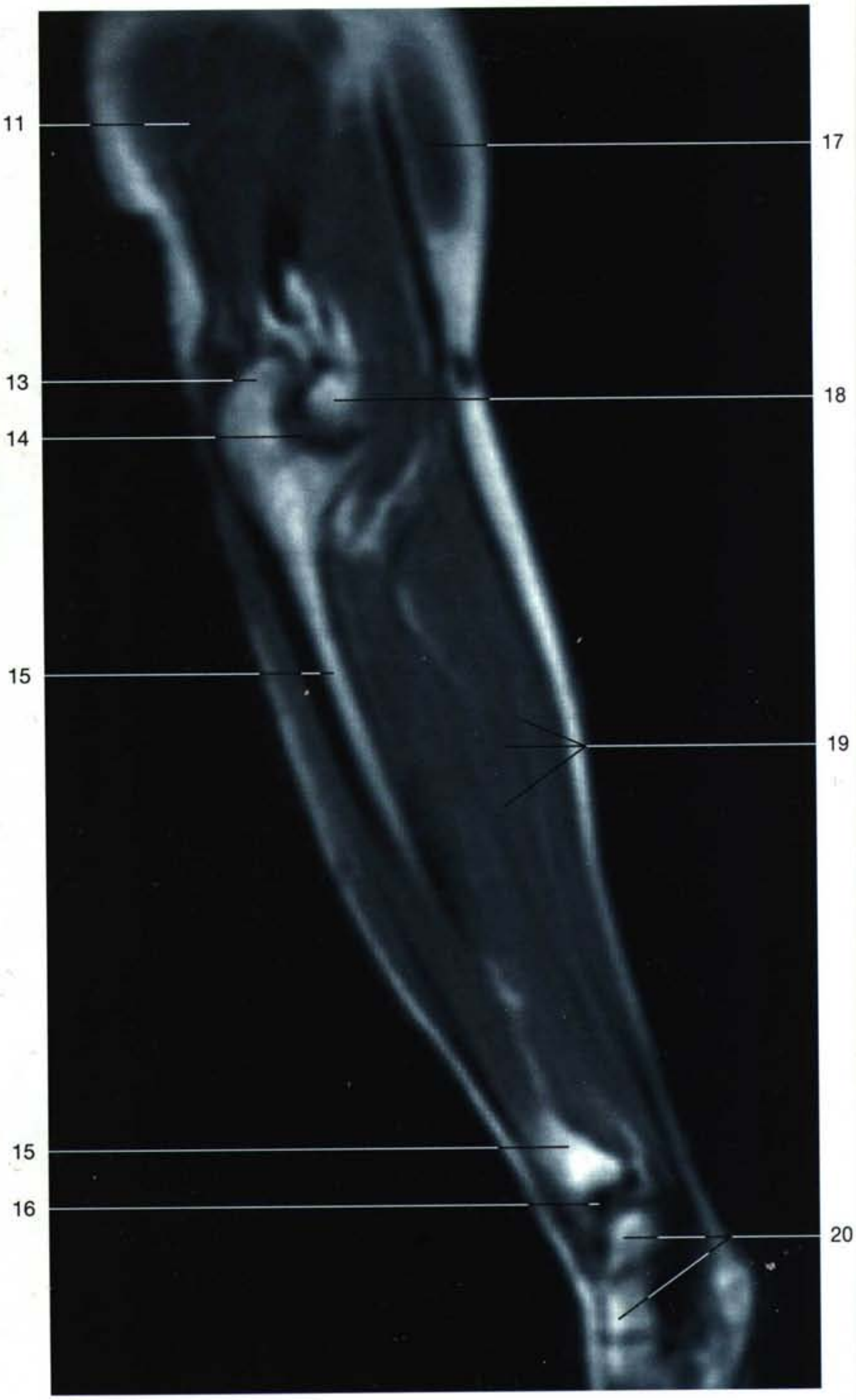

Sagittal section through the forearm at the level of the humeroulnar articulation (MR-Scan, after A. Heuck, G. Luttke, and J. W. Rohen, 1994).

1 **Cephalic vein**
2 Brachioradialis muscle covered by its fascia
3 **Posterior cutaneous nerve of forearm** (branch of radialis nerve)
4 Cephalic vein of forearm
5 Extensor pollicis longus and brevis muscles covered by their fascia
6 **Median cubital vein**
7 Lateral cutaneous nerves of forearm (branch of musculocutaneous nerve)
8 Intermedian vein of forearm
9 Superficial branch of radial nerve
10 **Dorsal digital branches of radial nerve**
11 Triceps brachii muscle
12 Venous network on the dorsum of the hand
13 Olecranon
14 **Humeroulnar articulation**
15 Ulna
16 Radiocarpal articulation
17 Biceps brachii muscle
18 Trochlea of humerus
19 Flexor muscles of forearm
20 Carpal bones

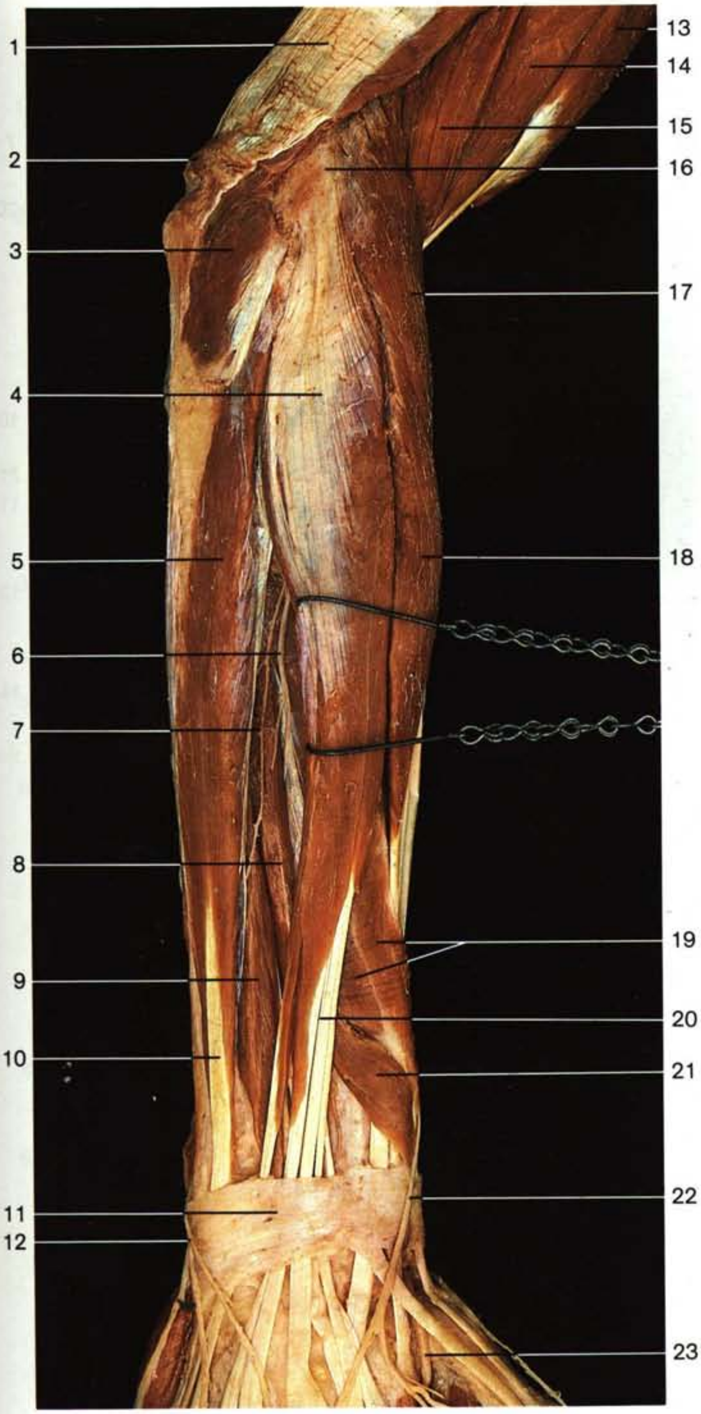

Vessels and nerves of right forearm, superficial layer (dorsal aspect).

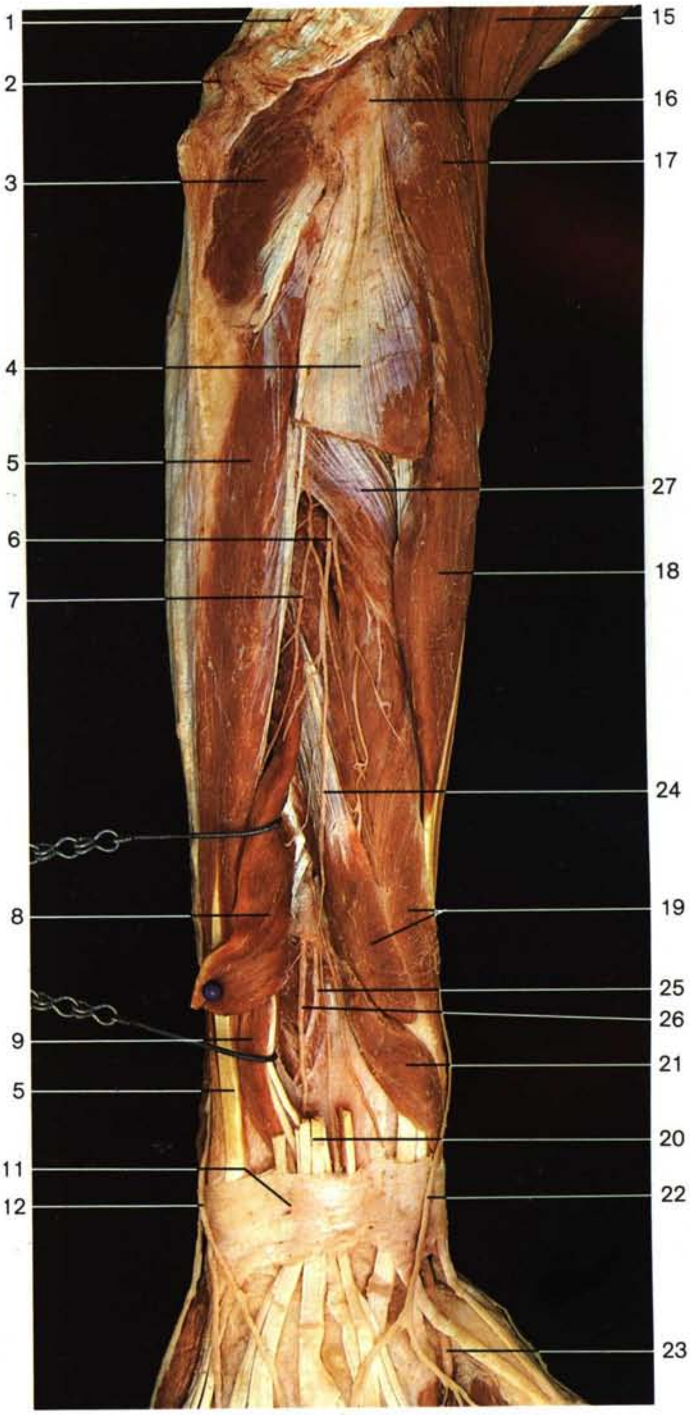

Vessels and nerves of right forearm, deep layer (dorsal aspect).

1 Tendon of triceps brachii muscle
2 Olecranon
3 Anconeus muscle
4 Extensor digitorum muscle
5 Extensor carpi ulnaris muscle
6 Deep branch of radial nerve
7 Posterior interosseous artery
8 Extensor pollicis longus muscle
9 Extensor indicis muscle
10 Tendon of extensor carpi ulnaris muscle
11 Extensor retinaculum
12 Dorsal branch of ulnar nerve
13 Biceps brachii muscle
14 Brachialis muscle
15 Brachioradialis muscle
16 Lateral epicondyle of humerus
17 Extensor carpi radialis longus muscle
18 Extensor carpi radialis brevis muscle
19 Abductor pollicis longus muscle
20 Tendons of extensor digitorum muscle
21 Extensor pollicis brevis muscle
22 Superficial branch of radial nerve
23 Radial artery
24 Posterior interosseous nerve
25 Posterior interosseous branch of radial nerve
26 Posterior branch of anterior interosseous artery
27 Supinator muscle

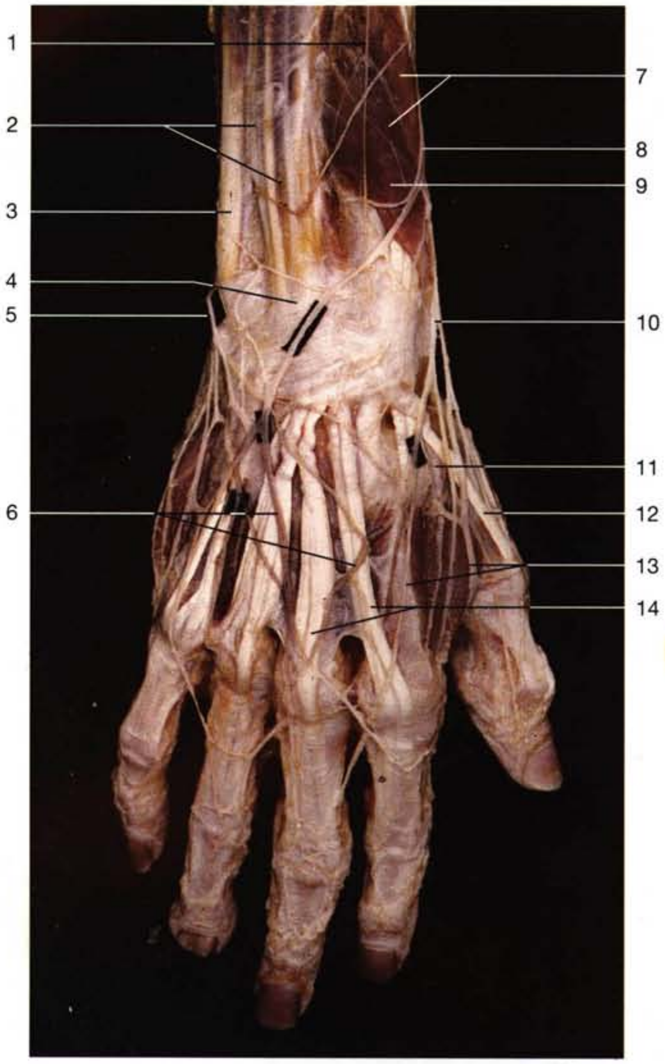

Cutaneous nerves and veins of forearm and hand (superficial layer, dorsal aspect).

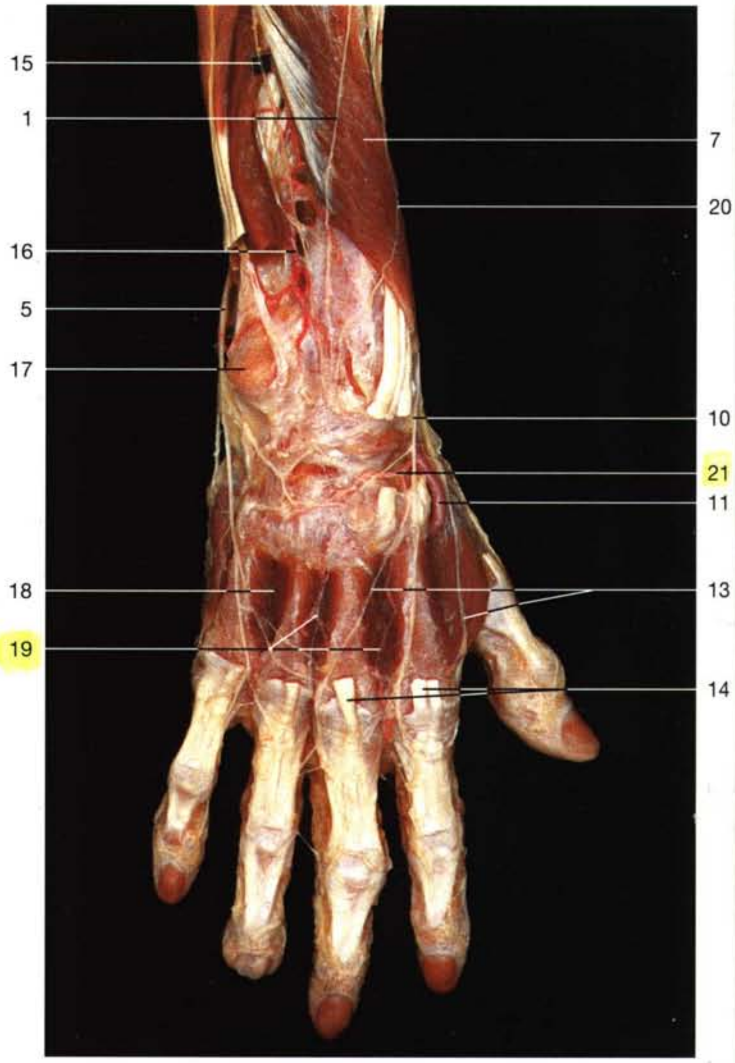

Dorsal region of forearm and hand (deeper layer). Extensor digitorum muscle has been partly removed.

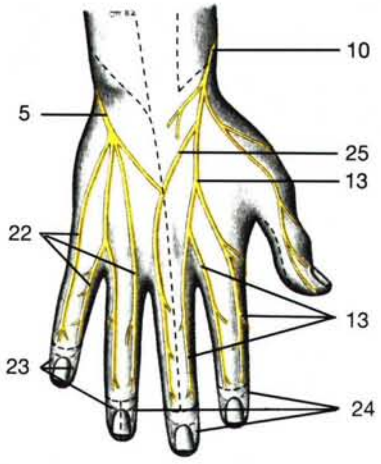

Innervation pattern of dorsal surfaces of hand
$2^{1}/_{2}$ digits by radial nerve, $2^{1}/_{2}$ digits by ulnar nerve. Note that the terminal branches to the dorsal surfaces of the distal phalanges are derived from the palmar digital nerves. The cutaneous distribution varies; often $3^{1}/_{2}$ digits are innervated by the radial and $1^{1}/_{2}$ digits by the ulnar nerve.

1 **Posterior cutaneous nerve of forearm** (branch of radial nerve)
2 Extensor digitorum muscle
3 Tendon of extensor carpi ulnaris muscle
4 Extensor retinaculum
5 **Ulnar nerve**
6 **Venous network of dorsum of hand**
7 Abductor pollicis longus muscle
8 Cephalic vein
9 Extensor pollicis brevis muscle
10 **Radial nerve,** superficial branch
11 **Radial artery**
12 Tendon of extensor pollicis longus muscle
13 Dorsal digital branches of radial nerve
14 Tendons of extensor digitorum muscle with intertendinous connections
15 Posterior interosseus nerve (branch of the deep radial nerve)
16 Posterior interosseous artery
17 Styloid process of ulna
18 Dorsal interosseus muscle IV
19 Dorsal carpal branch of radial artery
20 **Lateral cutaneous nerve of forearm** (branch of musculocutaneous nerve)
21 Dorsal metacarpal artery
22 Proper dorsal digital branches of ulnar nerve
23 Regions supplied by palmar digital nerves (ulnar nerve)
24 Regions supplied by palmar digital nerves (median nerve)
25 Communicating branch with ulnar nerve

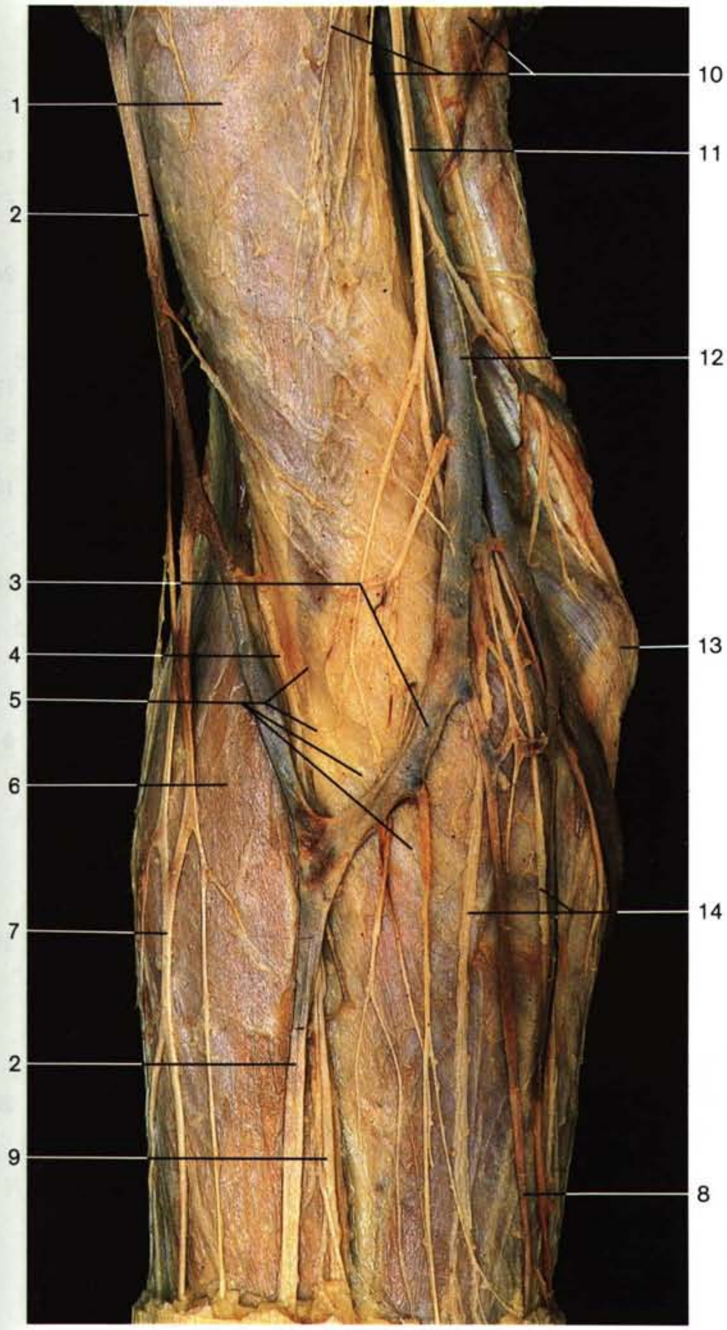

Cubital region (anterior aspect), dissection of cutaneous nerves and veins.

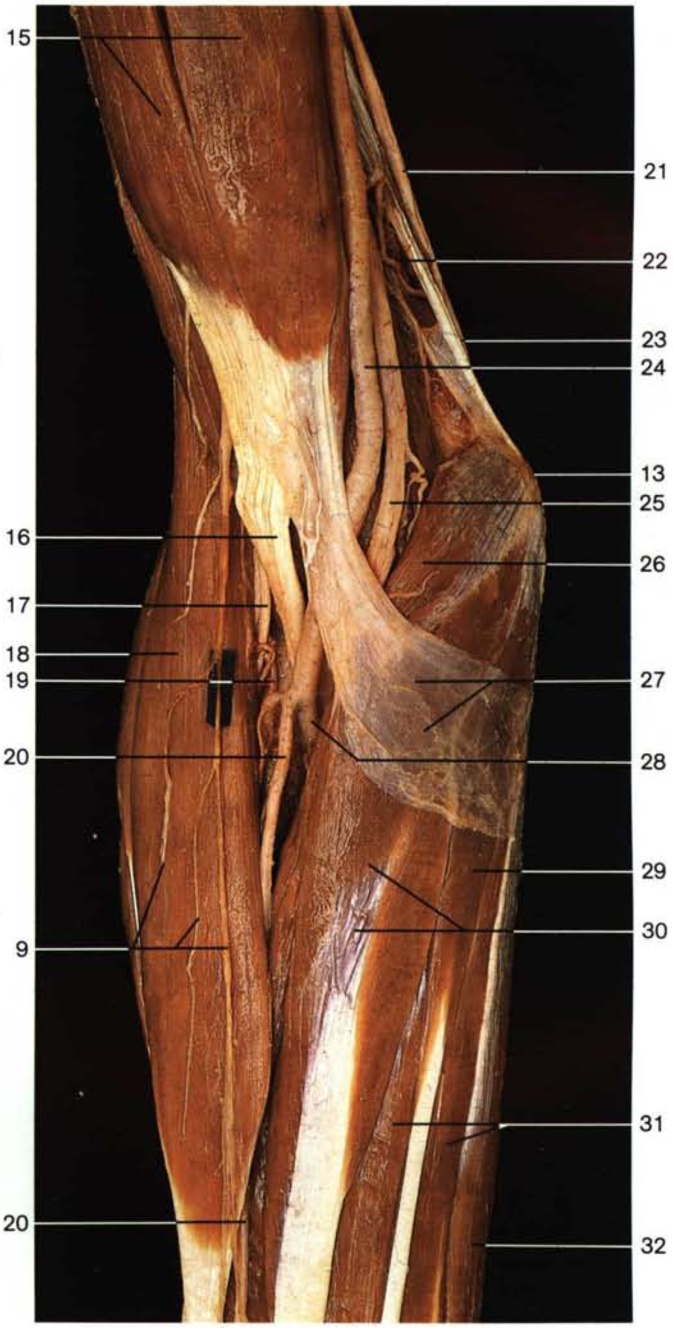

Cubital region, superficial layer (anterior aspect). The fasciae of the muscles have been removed.

1 Biceps brachii muscle with fascia
2 **Cephalic vein**
3 **Median cubital vein**
4 Lateral cutaneous nerve of forearm
5 Tendon and aponeurosis of biceps brachii muscle (covered by the antebrachial fascia)
6 Brachioradialis muscle with fascia
7 Accessory cephalic vein
8 Median vein of forearm
9 Branches of lateral cutaneous nerve of forearm
10 Terminal branches of medial cutaneous nerve of arm
11 **Medial cutaneous nerve of forearm**
12 **Basilic vein**
13 Medial epicondyle of humerus
14 Terminal branches of medial cutaneous nerve of forearm
15 Biceps brachii muscle

16 Tendon of biceps brachii muscle
17 Radial nerve
18 Brachioradiaiis muscle
19 Radial recurrent artery
20 **Radial artery**
21 **Ulnar nerve**
22 Superior ulnar collateral artery
23 Medial intermuscular septum
24 **Brachial artery**
25 **Median nerve**
26 Pronator teres muscle
27 Bicipital aponeurosis
28 Ulnar artery
29 Palmaris longus muscle
30 Flexor carpi radialis muscle
31 Flexor digitorum superficialis muscle
32 Flexor carpi ulnaris muscle

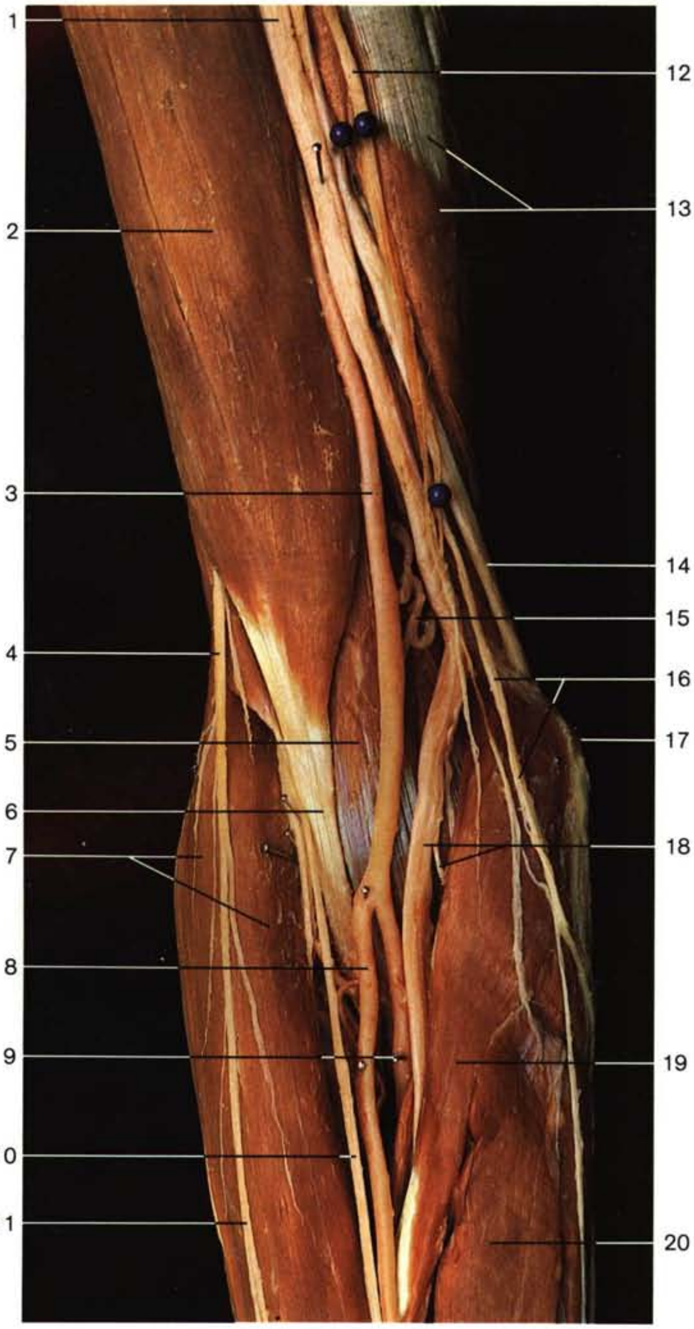

Cubital region, middle layer (anterior aspect). The bicipital aponeurosis has been removed.

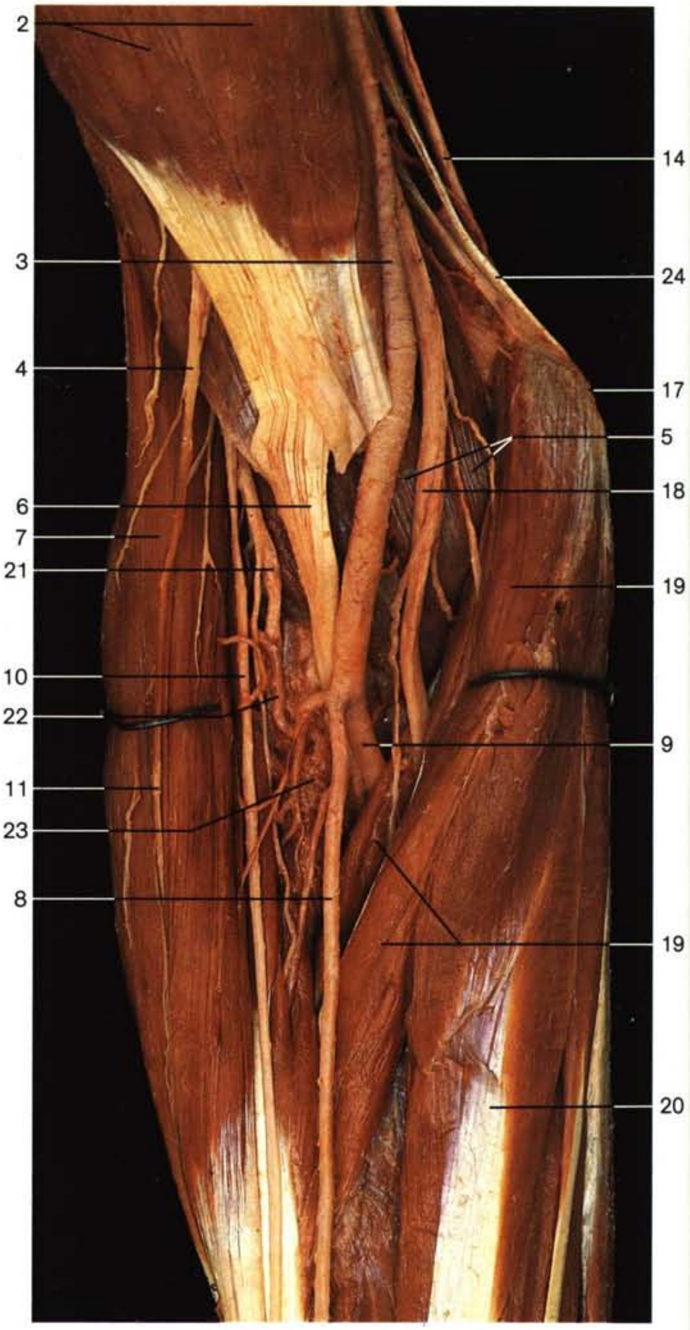

Cubital region, middle layer (anterior aspect). The pronator teres and brachioradialis muscles have been slightly reflected.

1 Median nerve
2 Biceps brachii muscle
3 Brachial artery
4 Lateral cutaneous nerve of forearm (terminal branch of **musculocutaneous nerve**)
5 Brachialis muscle
6 Tendon of biceps brachii muscle
7 Brachioradialis muscle
8 **Radial artery**
9 **Ulnar artery**
10 Superficial branch of radial nerve
11 Lateral cutaneous nerve of forearm
12 Medial cutaneous nerve of forearm
13 Triceps brachii muscle
14 **Ulnar nerve**
15 Inferior ulnar collateral artery
16 Anterior branch of medial cutaneous nerve of forearm
17 Medial epicondyle of humerus
18 **Median nerve** with branches to pronator teres muscle
19 Pronator teres muscle
20 Flexor carpi radialis muscle
21 Deep branch of radial nerve
22 Radial recurrent artery
23 Supinator muscle
24 Medial intermuscular septum of arm

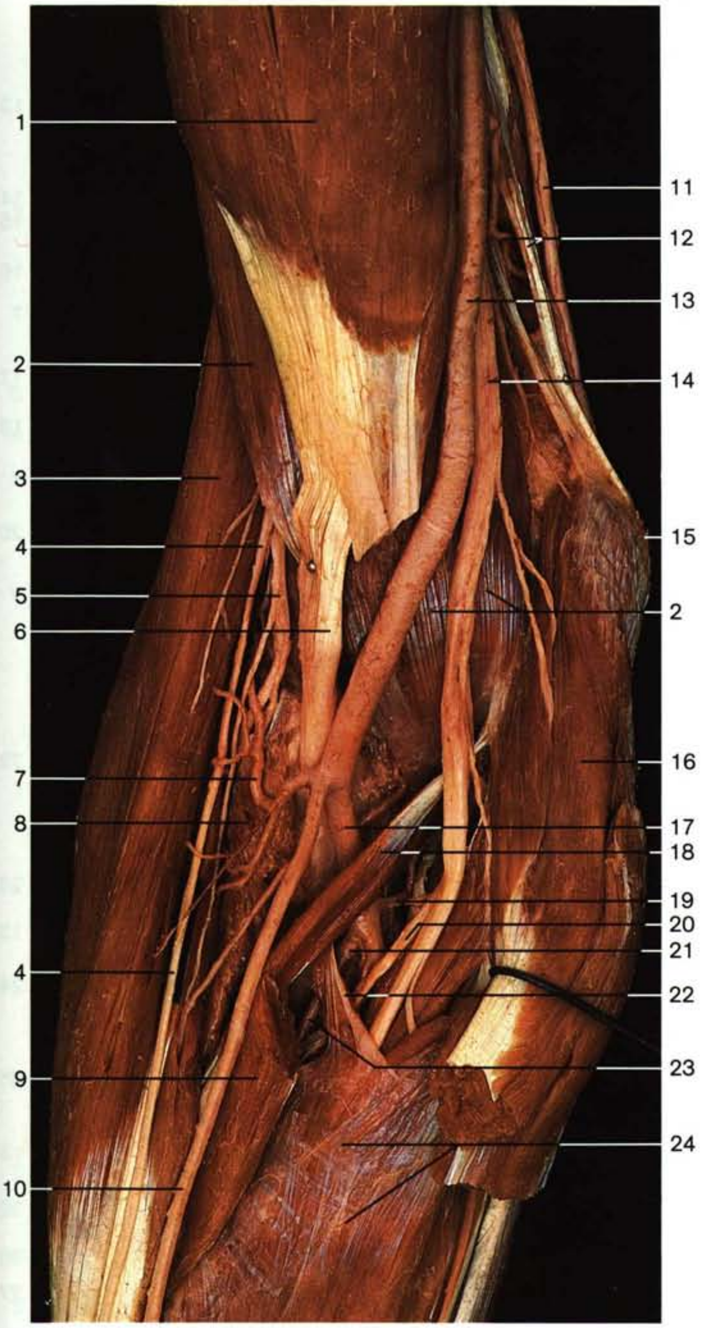

Cubital region, deep layer (anterior aspect). The pronator teres and flexor carpi ulnaris muscles have been cut and reflected.

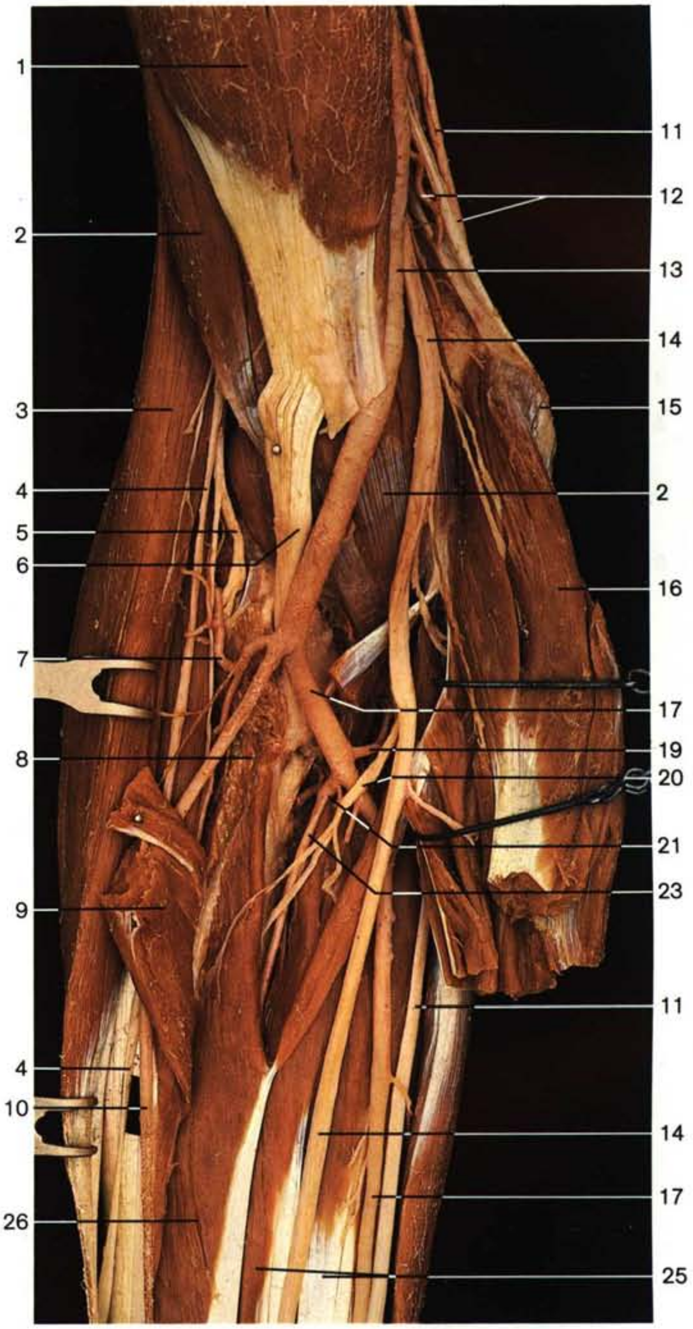

Cubital region, deepest layer (anterior aspect). The flexor digitorum superficialis and the ulnar head of the pronator teres have been cut and reflected.

1 Biceps brachii muscle
2 Brachialis muscle
3 Brachioradialis muscle
4 Superficial branch of radial nerve
5 Deep branch of radial nerve
6 Tendon of biceps brachii muscle
7 **Radial recurrent artery**
8 Supinator muscle
9 Insertion of pronator teres muscle
10 **Radial artery**
11 **Ulnar nerve**
12 Medial intermuscular septum of arm and superior ulnar collateral artery
13 **Brachial artery**
14 **Median nerve**
15 Medial epicondyle of humerus
16 Humeral head of pronator teres muscle
17 **Ulnar artery**
18 Ulnar head of pronator teres muscle
19 **Ulnar recurrent artery**
20 Anterior interosseous nerve
21 Common interosseous artery
22 Tendinous arch of flexor digitorum superficialis muscle
23 Anterior interosseous artery
24 Flexor digitorum superficialis muscle
25 Flexor digitorum profundus muscle
26 Flexor pollicis longus muscle

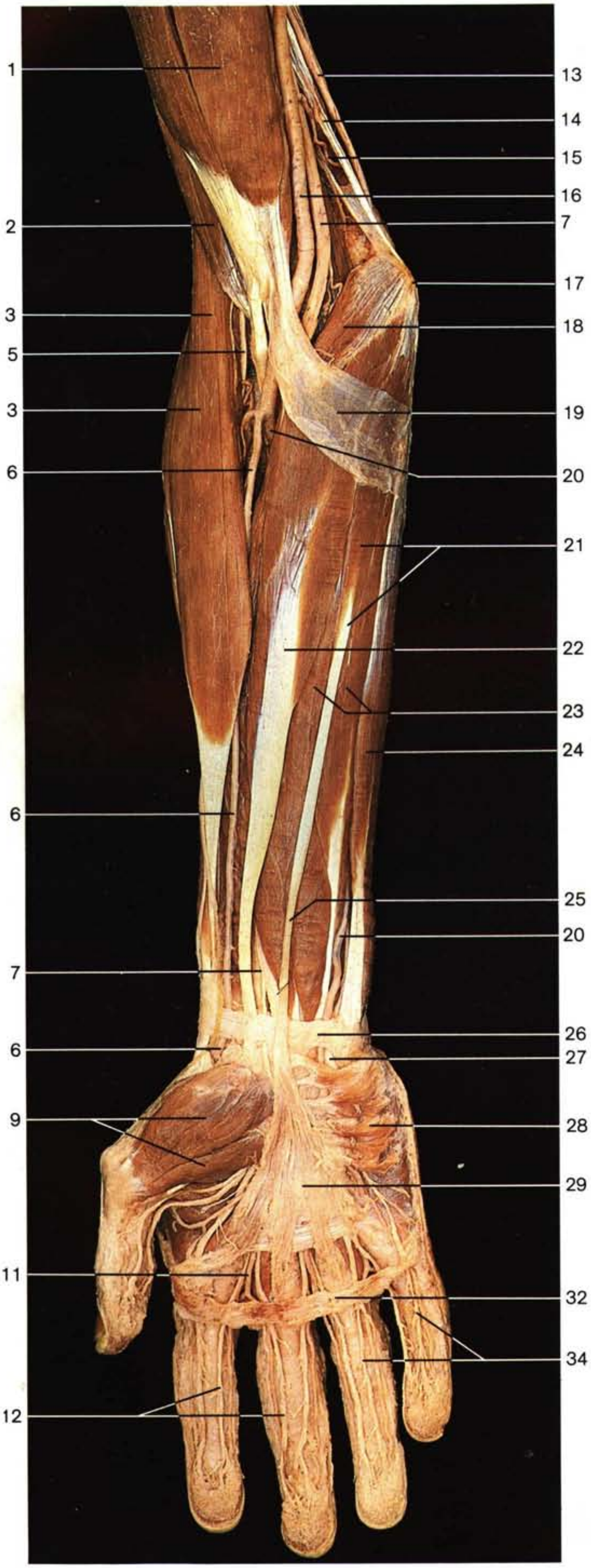
1
2
3
5
3
6
6
7
6
9
11
12
13
14
15
16
7
17
18
19
20
21
22
23
24
25
20
26
27
28
29
32
34

Vessels and nerves of right forearm and hand, superficial layer (palmar aspect).

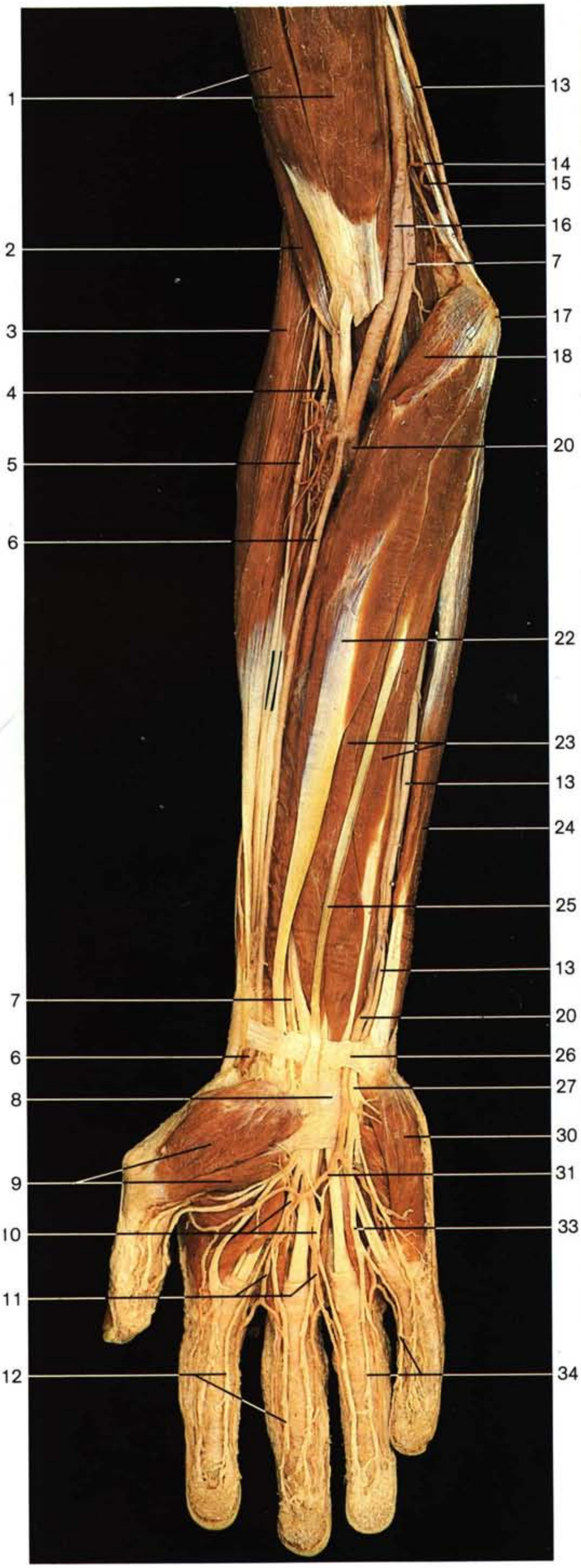
1
2
3
4
5
6
7
6
8
9
10
11
12
13
14
15
16
7
17
18
20
22
23
13
24
25
13
20
26
27
30
31
33
34

Vessels and nerves of right forearm and hand, superficial layer (palmar aspect). The palmar aponeurosis of the hand and the bicipital aponeurosis have been removed.

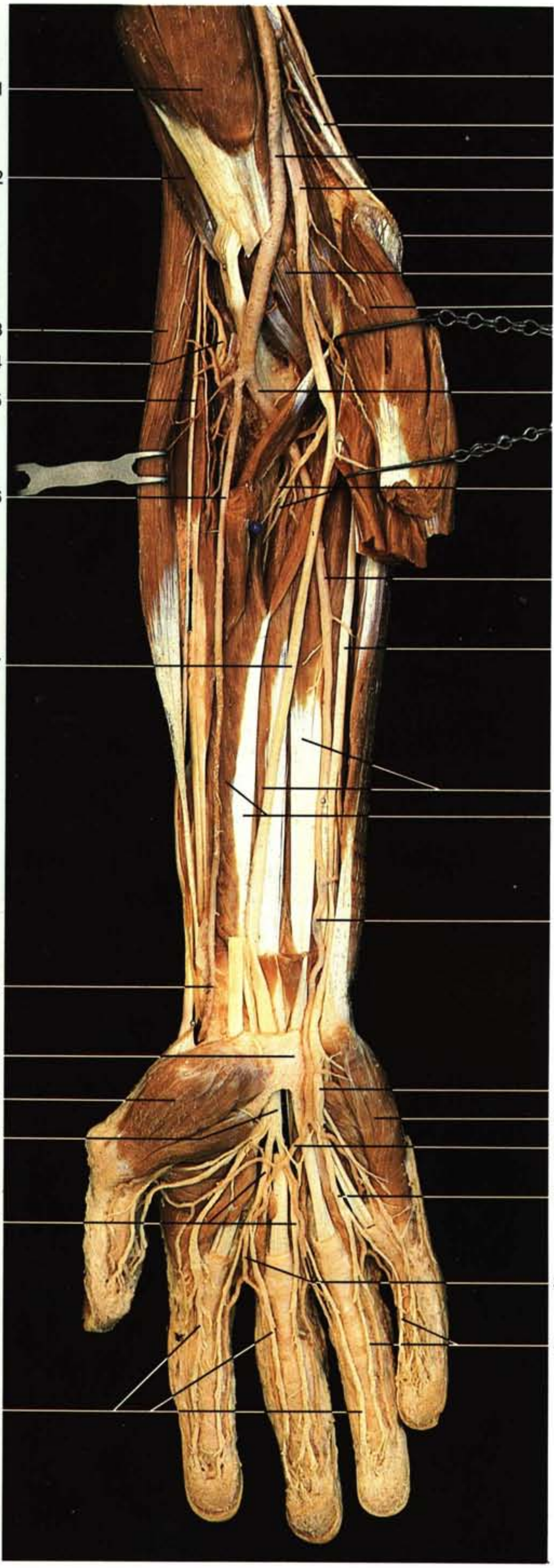

Vessels and nerves of forearm and hand, deep layer (anterior aspect). The superficial layer of the flexor muscles has been removed.

1 Biceps brachii muscle
2 Brachialis muscle
3 Brachioradialis muscle
4 **Deep branch of radial nerve**
5 **Superficial branch of radial nerve**
6 **Radial artery**
7 **Median nerve**
8 Flexor retinaculum
9 Thenar muscles
10 Common palmar digital branches of median nerve
11 Common palmar digital arteries
12 Proper palmar digital nerves (median nerve)
13 **Ulnar nerve**
14 Medial intermuscular septum of arm
15 Superior ulnar collateral artery
16 Brachial artery
17 Medial epicondyle of humerus
18 Pronator teres muscle
19 Bicipital aponeurosis
20 **Ulnar artery**
21 Palmaris longus muscle
22 Flexor carpi radialis muscle
23 Flexor digitorum superficialis muscle
24 Flexor carpi ulnaris muscle
25 Tendon of palmaris longus muscle
26 Remnant of antebrachial fascia
27 Superficial branch of ulnar nerve
28 Palmaris brevis muscle
29 Palmar aponeurosis
30 Hypothenar muscles
31 **Superficial palmar arch**
32 Superficial transverse metacarpal ligament
33 Common palmar digital branch of ulnar nerve
34 Proper palmar digital branches of ulnar nerve
35 Anterior interosseous artery and nerve
36 Flexor digitorum profundus muscle
37 Common palmar digital arteries
38 Palmar branch of median nerve
39 Flexor pollicis longus muscle
40 Palmar branch of ulnar nerve

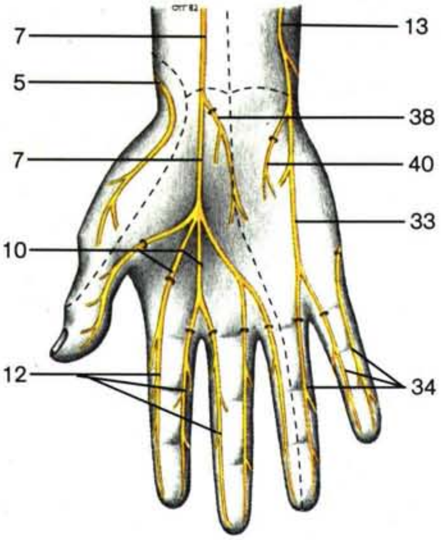

Cutaneous innervation of hand (palmar aspect). (Schematic drawing.)

Cutaneous innervation of palmar surface:
3½ digits by median nerve,
1½ digits by ulnar nerve.

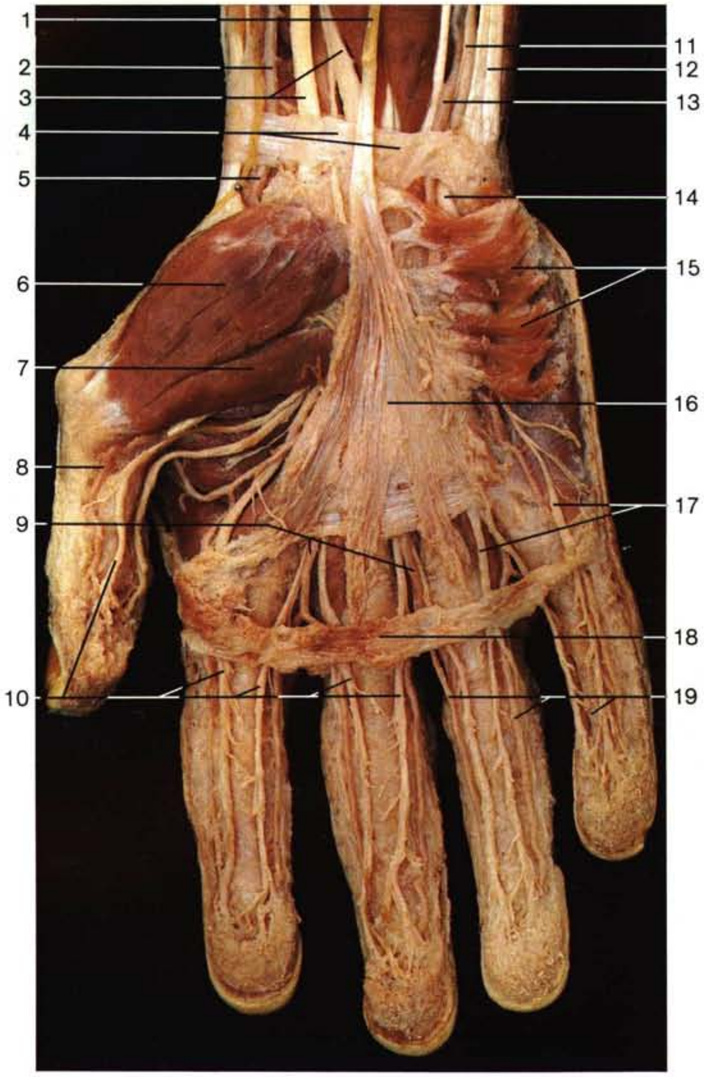

Right hand, superficial layer, dissection of vessels and nerves (palmar aspect).

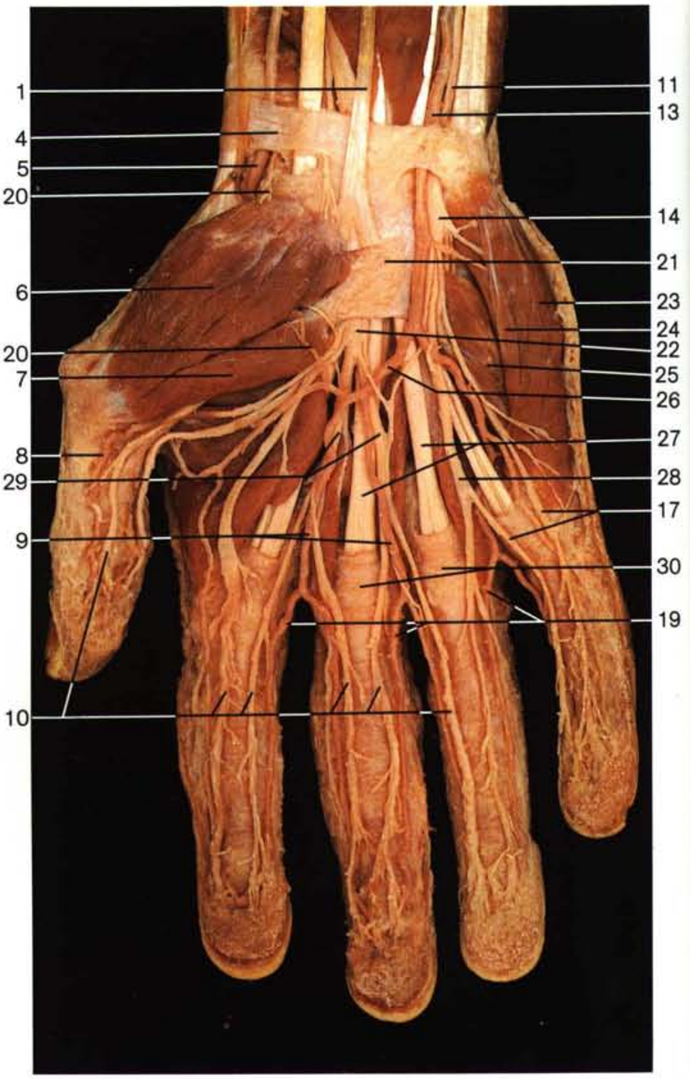

Right hand, superficial layer, dissection of vessels and nerves (palmar aspect). The palmar aponeurosis has been removed to display the superficial palmar arch.

1 Tendon of palmaris longus muscle
2 **Radial artery**
3 Tendon of flexor carpi radialis muscle and median nerve
4 Distal part of antebrachial fascia
5 Radial artery passing into the anatomical snuffbox
6 Abductor pollicis brevis muscle
7 Superficial head of flexor pollicis brevis muscle
8 Palmar digital artery of thumb
9 Common palmar digital arteries
10 Proper palmar digital nerves (median nerve)
11 **Ulmar nerve**
12 Tendon of flexor carpi ulnaris muscle
13 **Ulnar artery**
14 Superficial branch of ulnar nerve
15 Palmaris brevis muscle
16 Palmar aponeurosis
17 Palmar digital nerves (ulnar nerve)
18 Superficial transverse metacarpal ligament
19 Proper palmar digital arteries
20 Superficial palmar branch of radial artery (contributing to the superficial palmar arch)
21 Flexor retinaculum
22 **Median nerve**
23 Abductor digiti minimi muscle
24 Flexor digiti minimi brevis muscle
25 Opponens digiti minimi muscle
26 **Superficial palmar arch**
27 Tendons of flexor digitorum superficialis muscle
28 Common palmar digital branch of ulnar nerve
29 Common palmar digital branch of median nerve
30 Fibrous sheath of flexor tendons

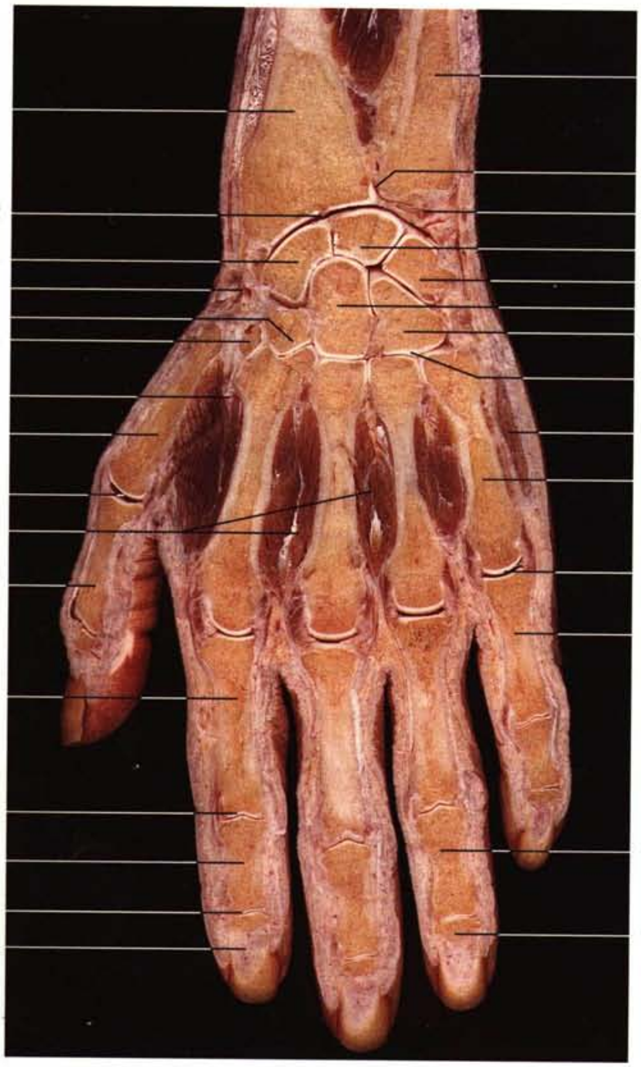
1
2
3
4
5
6
4
7
8
9
10
11
12
13
12
14
15
16
17
18
19
20
21
22
23
24
25
11
13
14

Coronal section through the right hand (palmar aspect).

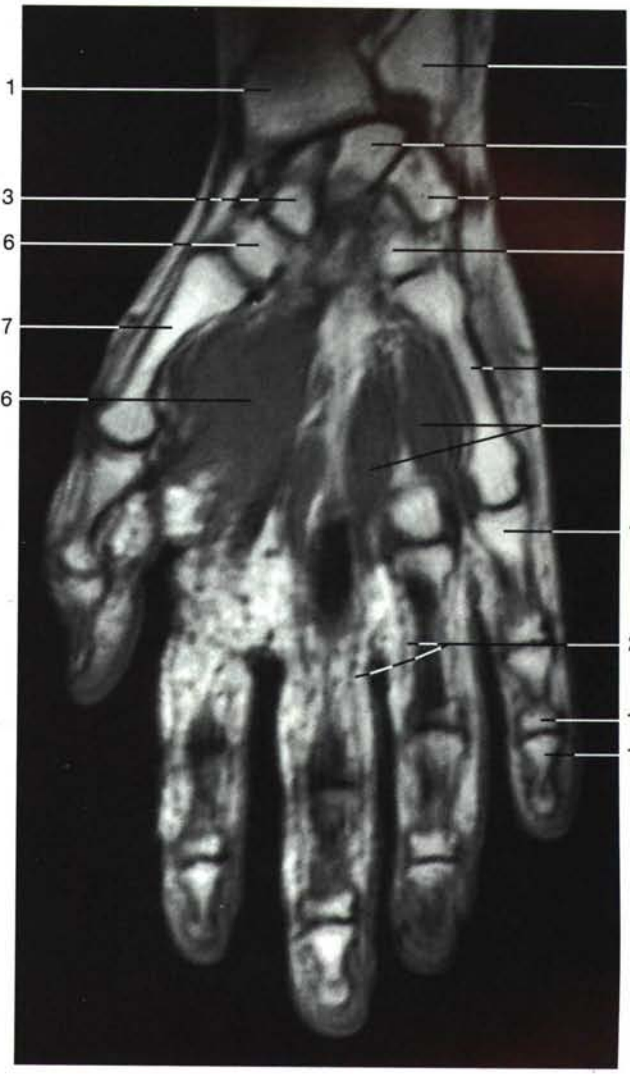
1
3
6
7
26
15
18
19
21
24
9
11
27
13
14

Coronal section through the right hand (palmar aspect) (MR-Scan, courtesy of Dr. A. Heuck, Munich).

1 Radius
2 **Radiocarpal articulation**
3 **Scaphoid** (navicular) **bone**
4 **Radial artery**
5 Trapezoid bone
6 **Trapezium bone**
7 First metacarpal bone
8 Metacarpophalangeal articulation of thumb
9 Interosseous muscles
10 Proximal phalanx of thumb
11 Proximal phalanx of fingers
12 Interphalangeal articulations
13 Middle phalanx
14 Distal phalanx
15 **Ulna**
16 Distal radioulnar articulation
17 Articular disk
18 **Lunate bone**
19 **Triangular bone**
20 **Capitate bone**
21 **Hamate bone**
22 Carpometacarpal articulations
23 Abductor digiti minimi muscle
24 Fifth metacarpal bone
25 Metacarpophalangeal articulation
26 Adductor pollicis muscle
27 Proper palmar digital arteries

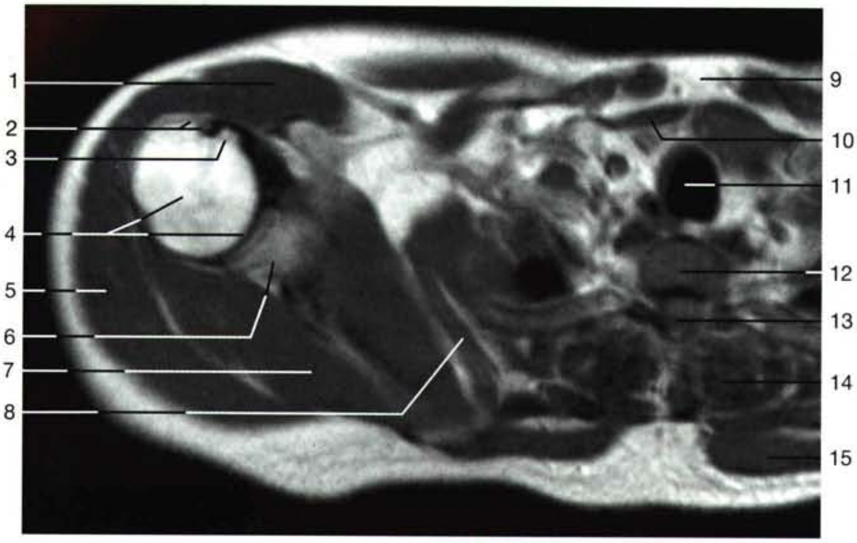
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Horizontal section through shoulder joint (section 1; MR-Image; inferior aspect).

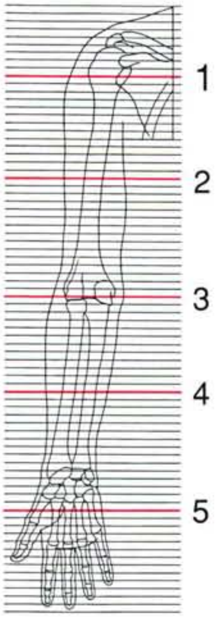

Upper extremity, location of sections 1–5 (MR-Scans, courtesy of Dr. A. Heuck, Munich).

1 Pectoralis major muscle
2 Greater tubercle and tendon of biceps muscle
3 Lesser tubercle
4 Head of humerus and articular cavity of shoulder joint
5 Deltoid muscle
6 Scapula
7 Infraspinatus muscle
8 Serratus anterior muscle
9 Sternum
10 Infrahyoid muscles
11 Trachea
12 Body of thoracic vertebra
13 Vertebral canal and spinal cord
14 Deep muscles of the back
15 Trapezius muscle
16 Brachialis muscle
17 Radial nerve and profunda brachii vessels

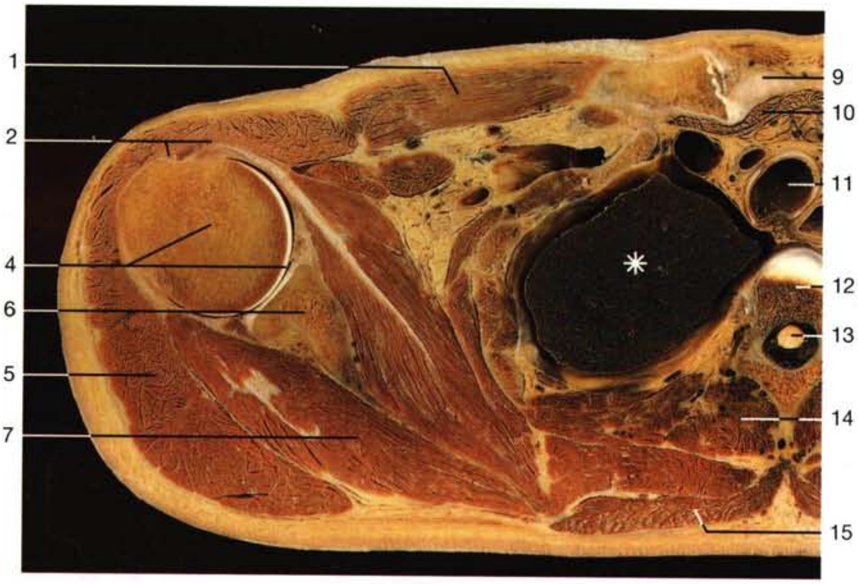
1
2
4
6
5
7
9
10
11
*
12
13
14
15

Horizontal section through the right shoulder at the level of T_1 (section 1; inferior aspect). * = Upper lobe of lung.

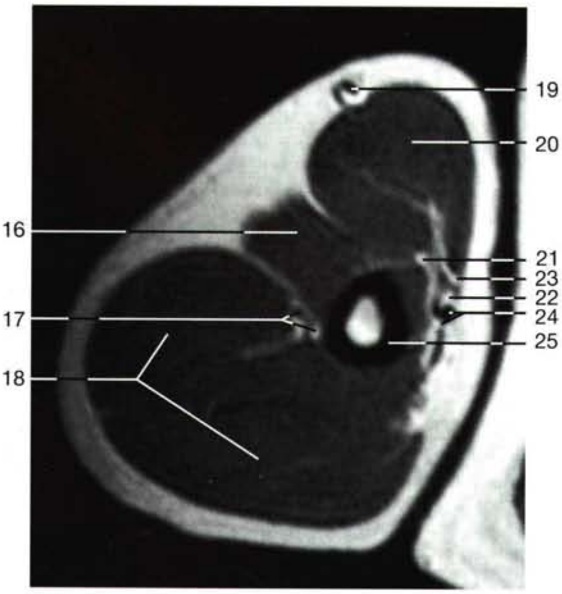
19
20
16
21
23
22
24
17
25
18

Axial section through the middle of the right arm (section 2; MR-Image, inferior aspect).

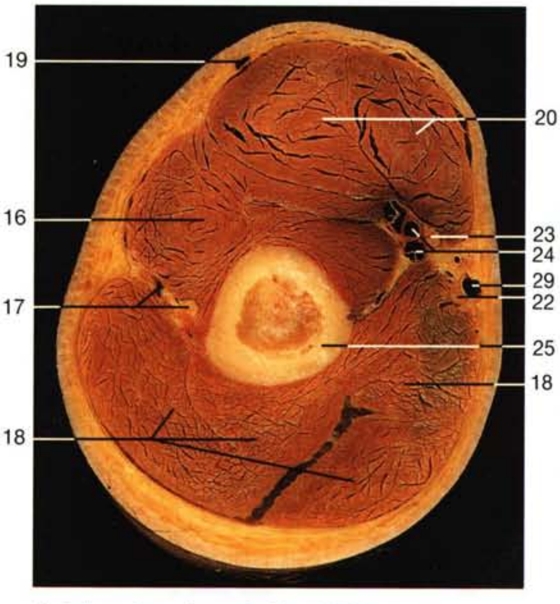
19
20
16
23
24
29
17
22
25
18
18

Axial section through the middle of the right arm (section 2; inferior aspect).

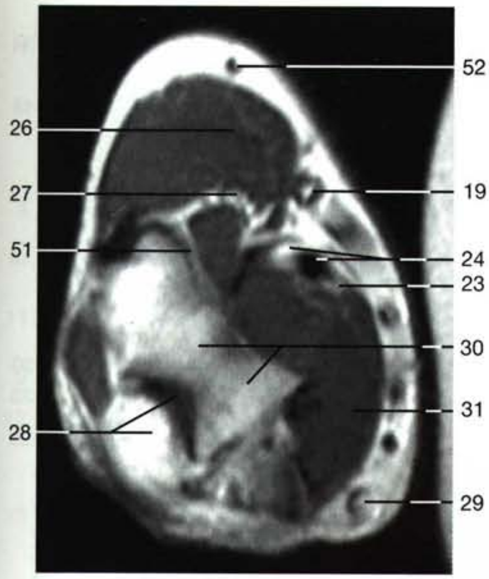

Axial section through the elbow joint (section 3; MR-Image; inferior aspect).

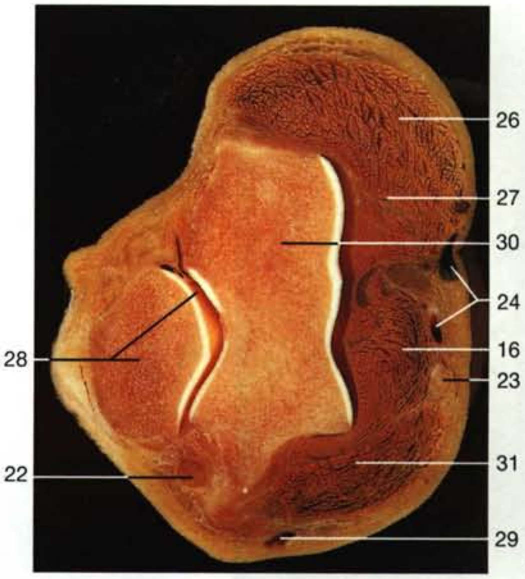

Axial section through the right elbow joint (section 3; inferior aspect).

18 Triceps brachii muscle
19 Cephalic vein
20 Biceps brachii muscle
21 Musculocutaneous nerve
22 **Ulnar nerve**
23 **Medianus nerve**
24 Brachial artery and vein
25 Shaft of humerus
26 Brachioradialis muscle
27 **Radial nerve**
28 Olecranon and articular cavity of elbow joint
29 Basilic vein
30 Humerus
31 Pronator teres muscle
32 Extensor muscles of forearm
33 Ramus profundus of radialis nerve
34 Anterior interosseus vessels and nerve
35 Interosseous membrane
36 Ulna
37 Radius
38 Radial artery and superficial branch of radial nerve
39 Flexor pollicis longus muscle
40 Flexor digitorum superficialis and profundus muscles
41 Ulnar nerve, ulnar artery and vein
42 Flexor carpi ulnaris muscle
43 Radial artery
44 Metacarpal bones III and IV
45 **Carpal canal** with tendons of flexor digitorum muscles
46 Hypothenar muscle
47 Median nerve
48 Interosseus muscles
49 First metacarpal bone
50 Thenar muscles
51 Articular cavity of humeroradial joint
52 Median cubital vein

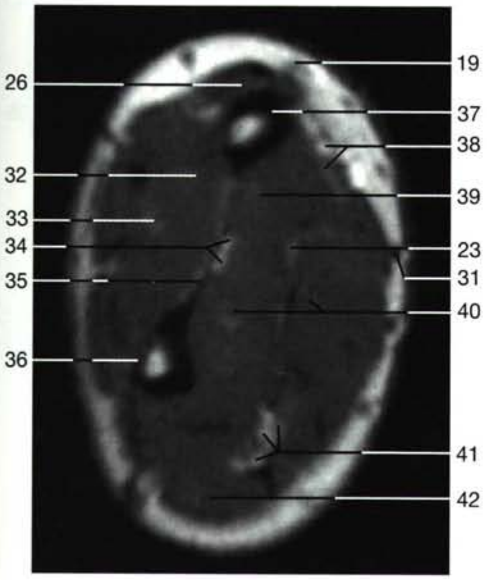

Axial section through the middle of the forearm (section 4; MR-Image; inferior aspect).

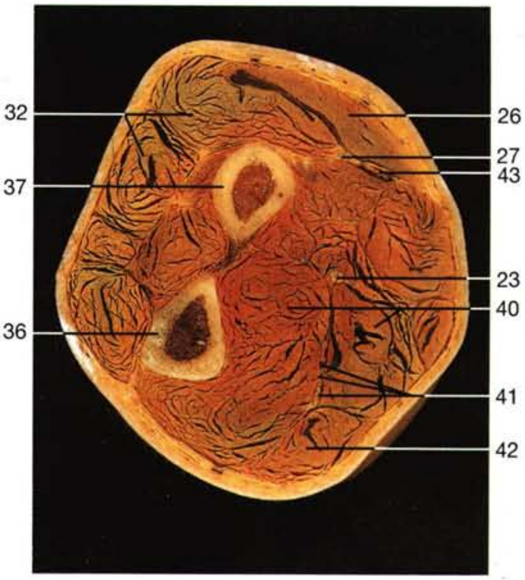

Axial section through the right forearm (section 4; inferior aspect).

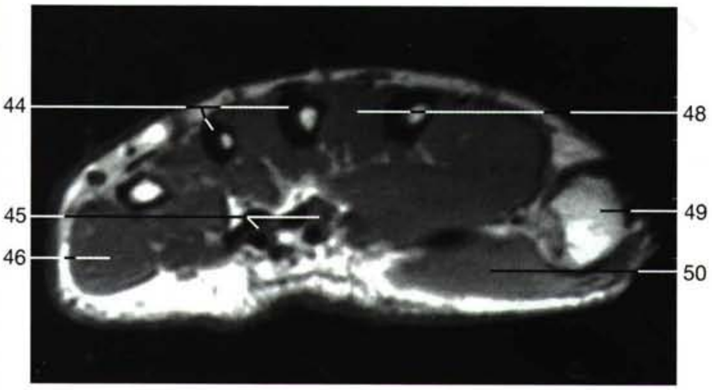

Axial section through the right hand (metacarpus; MR-Image; section 5; inferior aspect).

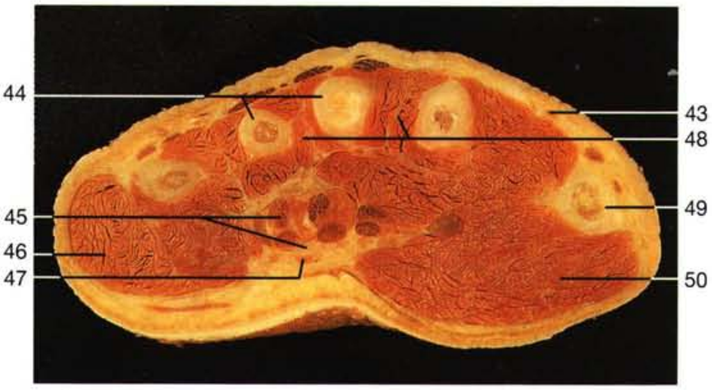

Axial section through the right hand at the level of the metacarpus (section 5; inferior aspect).

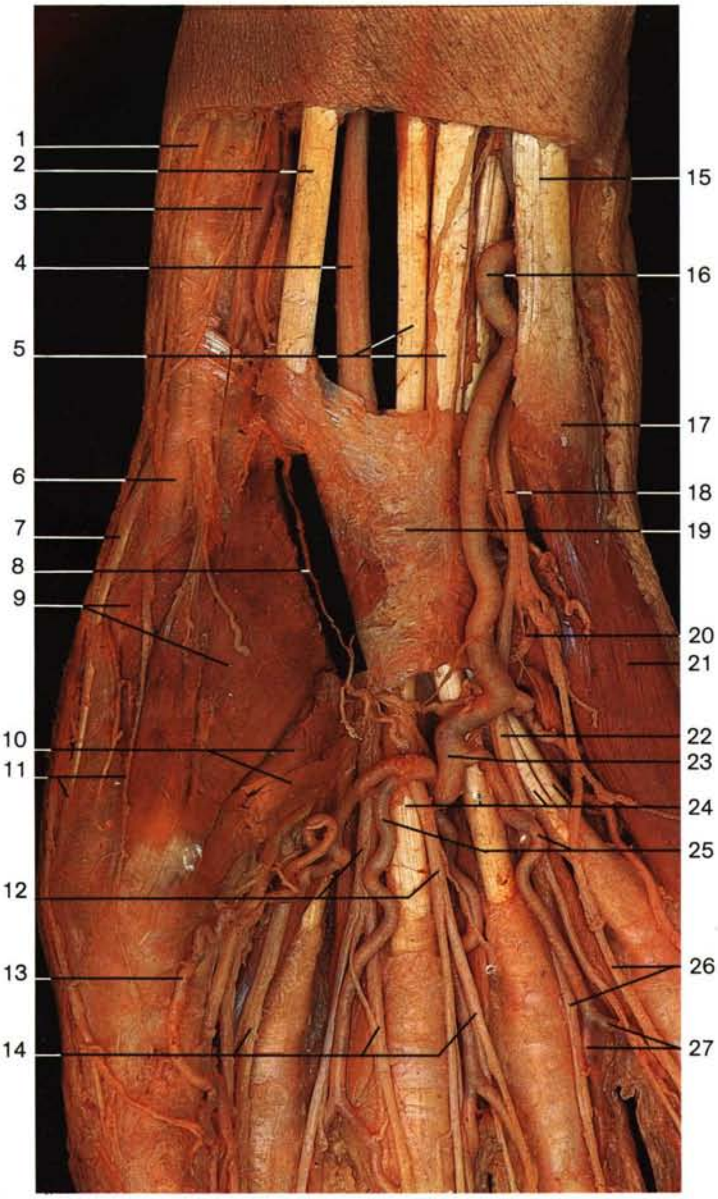
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

◁
Right hand, superficial layer (palmar aspect). Dissection of the superficial palmar arch.

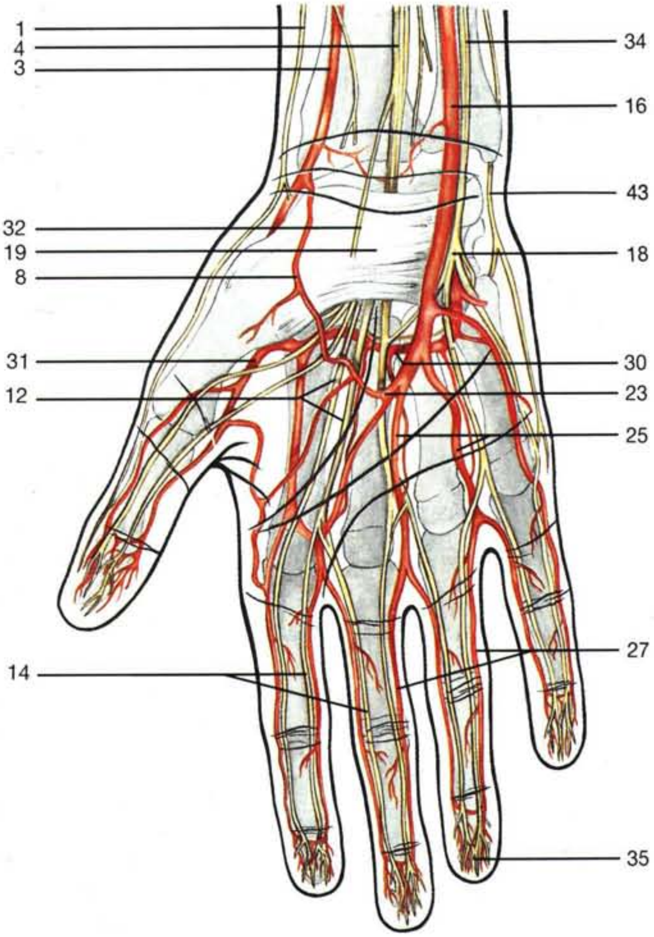
1
4
3
32
19
8
31
12
14
34
16
43
18
30
23
25
27
35

Arteries and nerves of the right hand (palmar aspect). (Schematic drawing.)

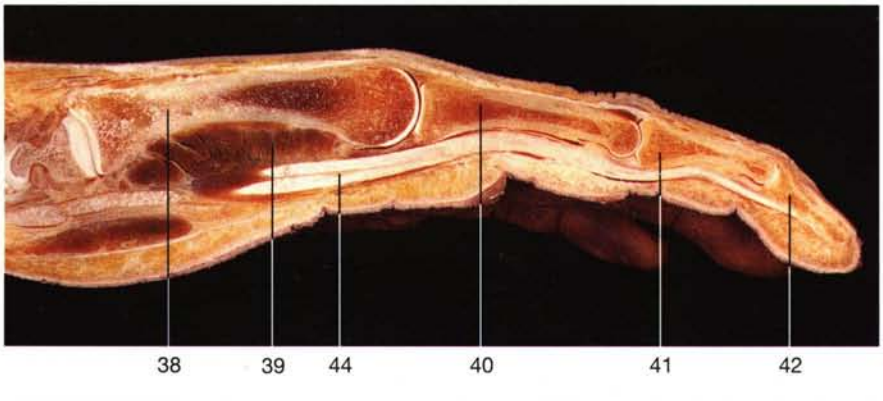
38
39
44
40
41
42

Longitudinal section through the hand at the level of the third finger

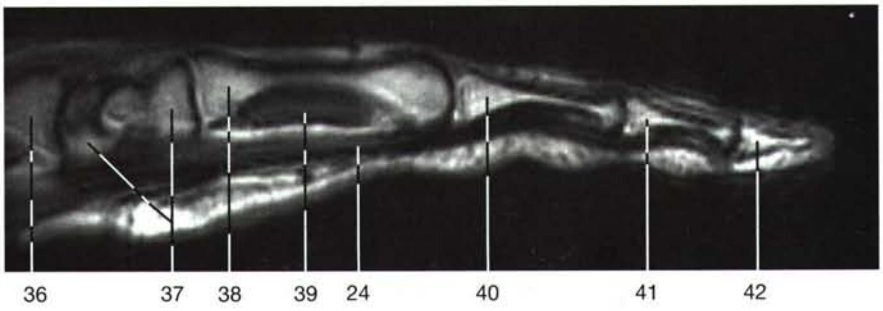
36
37
38
39
24
40
41
42

Longitudinal section through the hand at the level of the third finger (MR-Scan, courtesy of Dr. A. Heuck, Munich).

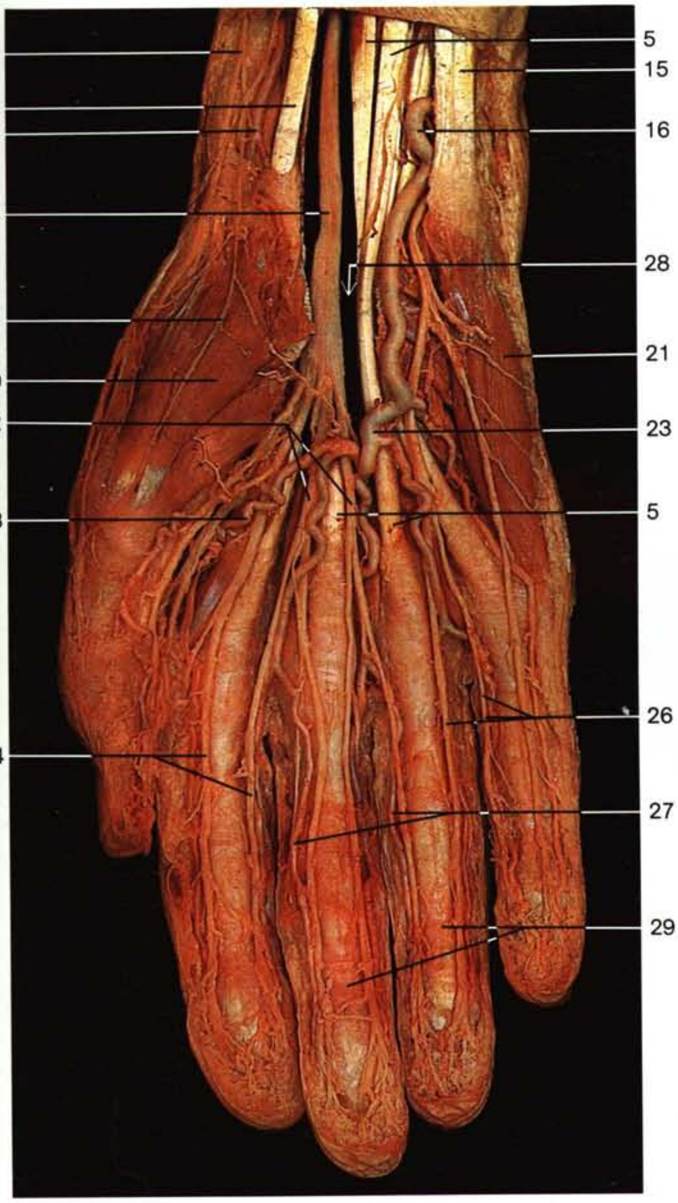

Right hand, middle layer (palmar aspect). The flexor retinaculum has been removed.

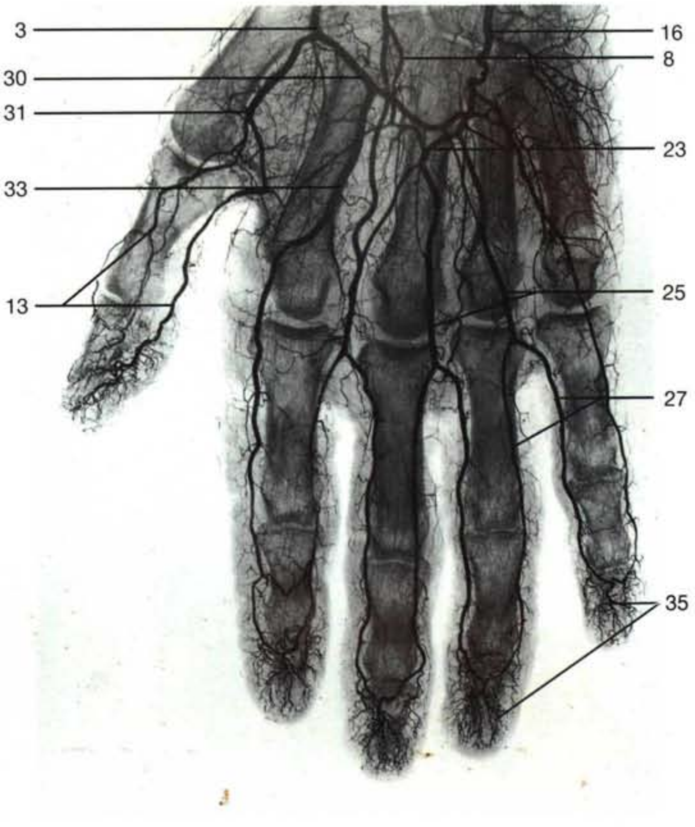

Arteriogram of the right hand (palmar aspect).

1 Superficial branch of radial nerve
2 Tendon of flexor carpi radialis muscle
3 **Radial artery**
4 **Median nerve**
5 Tendon of flexor digitorum superficialis muscle
6 Tendon of abductor pollicis longus muscle
7 Tendon of extensor pollicis brevis muscle
8 Superficial palmar branch of radial artery
9 Abductor pollicis brevis muscle
10 Superficial head of flexor pollicis brevis muscle
11 Terminal branches of superficial branch of radial nerve
12 Common palmar digital nerves (median nerve)
13 Proper palmar digital arteries of thumb
14 Proper palmar digital nerves (median nerve)
15 Tendon of flexor carpi ulnaris muscle
16 **Ulnar artery**
17 Position of pisiform bone
18 Superficial branch of ulnar nerve
19 Flexor retinaculum
20 Deep branch of ulnar nerve
21 Abductor digiti minimi muscle
22 Common palmar digital nerves (ulnar nerve)
23 **Superficial palmar arch**
24 Tendons of flexor digitorum muscles
25 Common palmar digital arteries
26 **Palmar digital nerves (ulnar nerve)**
27 **Proper palmar digital arteries**
28 **Carpal tunnel**
29 Fibrous sheaths for the tendons of flexor digitorum muscles
30 **Deep palmar arch**
31 Princeps pollicis artery
32 Palmar branch of median nerve
33 Common digital palmar artery
34 Ulnar nerve
35 Capillary network of finger
36 Radius
37 Carpal bones
38 Metacarpal bone
39 Interosseus muscles
40 Proximal phalanx
41 Middle phalanx
42 Distal phalanx
43 Dorsal branch of ulnar nerve
44 Tendons of flexor digitorum profundus (upper) and superficialis (lower) muscles

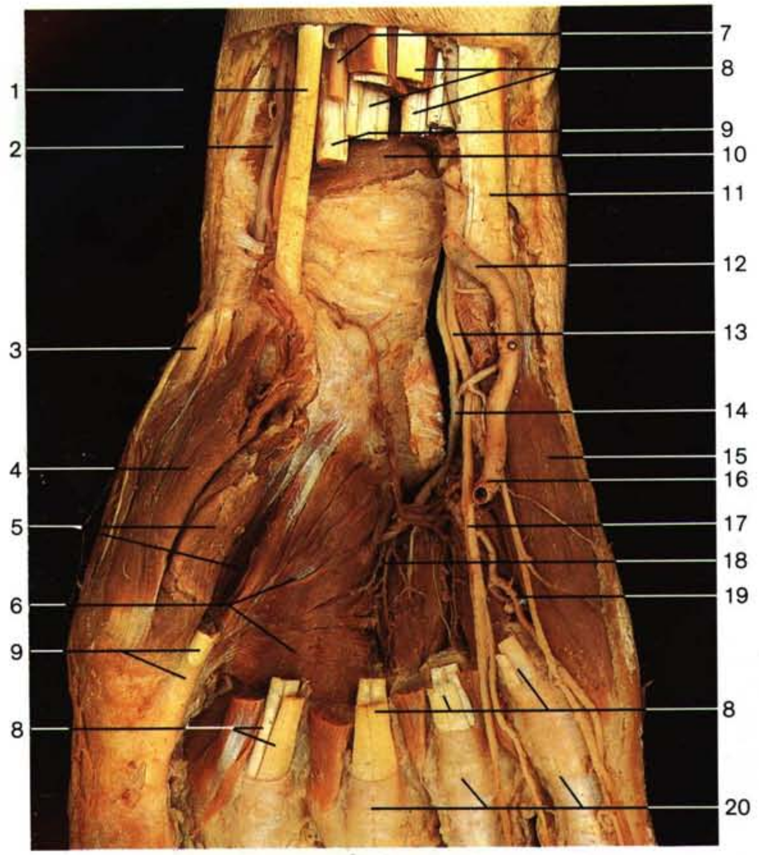

Right hand, deep layer (palmar aspect). The carpal tunnel has been opened, the tendons of the flexor muscles have been removed and the superficial palmar arch has been cut.

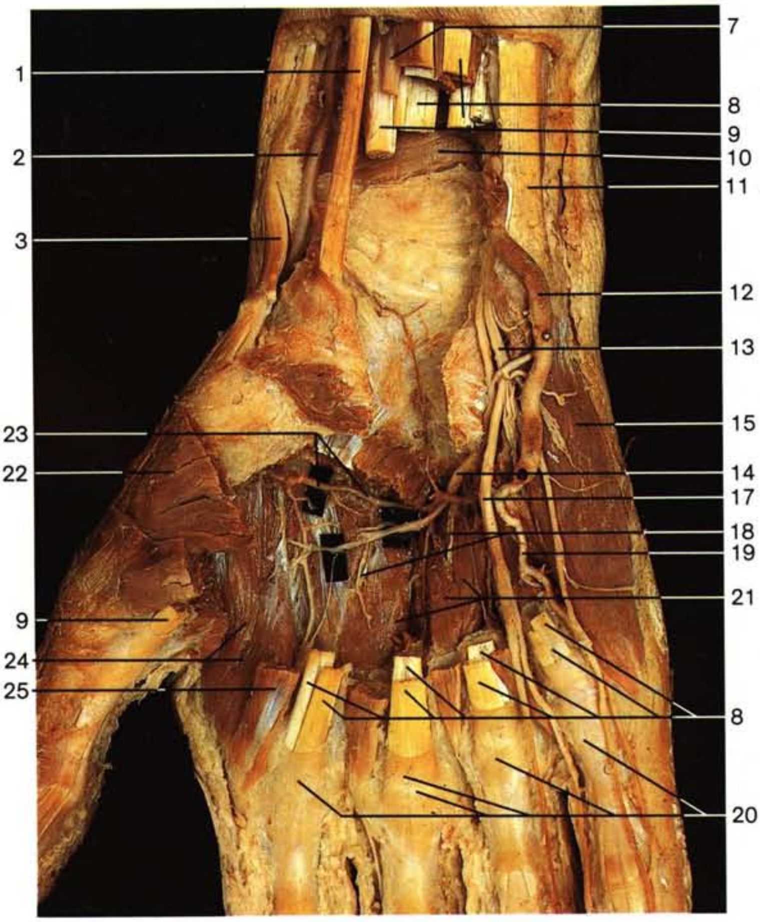

Right hand, deep layer (palmar aspect). Dissection of the deep palmar arch.

1 Tendon of flexor carpi radialis muscle
2 Radial artery
3 Tendon of abductor pollicis longus muscle
4 Abductor pollicis brevis muscle
5 Superficial and deep heads of flexor pollicis brevis muscle
6 Oblique and transverse heads of adductor pollicis muscle
7 Median nerve
8 Tendons of flexor digitorum superficialis and profundus muscles
9 Tendon of flexor pollicis longus muscle
10 Pronator quadratus muscle
11 Tendon of flexor carpi ulnaris muscle
12 Ulnar artery
13 Superficial branch of ulnar nerve
14 Deep branch of ulnar nerve
15 Abductor digiti minimi muscle
16 **Superficial palmar arch** (cut end)
17 Common palmar digital nerves (ulnar nerve)
18 Palmar metacarpal arteries of deep palmar arch
19 Palmar digital artery of the 5th finger
20 Fibrous sheaths of tendons of flexor muscles
21 Palmar interosseous muscles
22 Opponens pollicis muscle (cut)
23 **Deep palmar arch**
24 1st dorsal interosseous muscle
25 1st lumbrical muscle

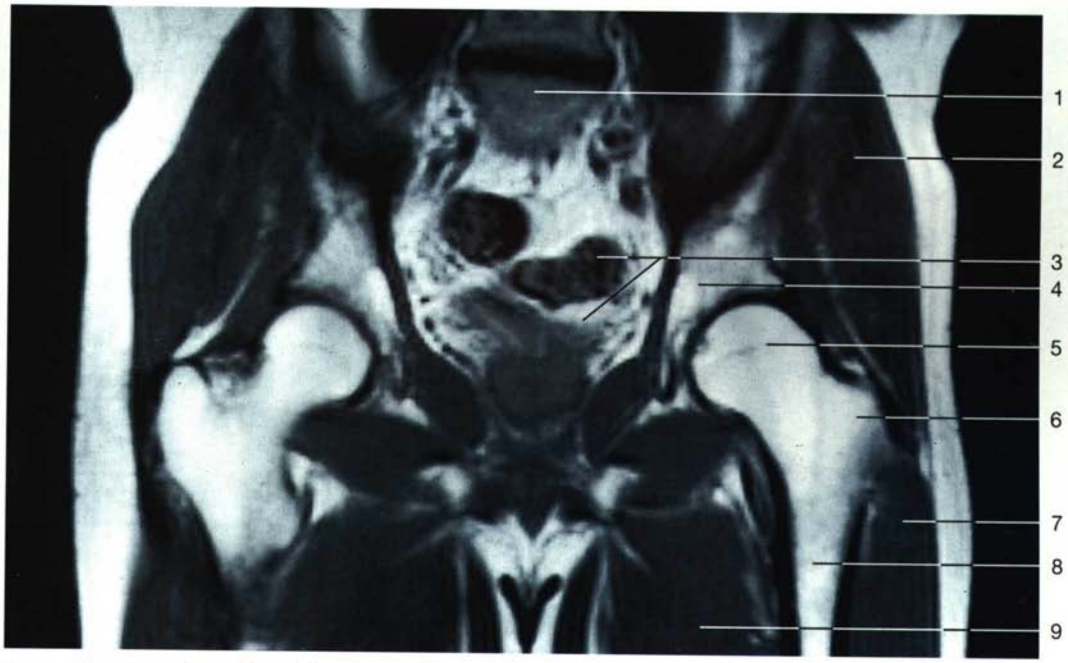

Coronal section through pelvis and thighs (MR-Scan of the male, courtesy of Dr. A. Heuck, Munich).

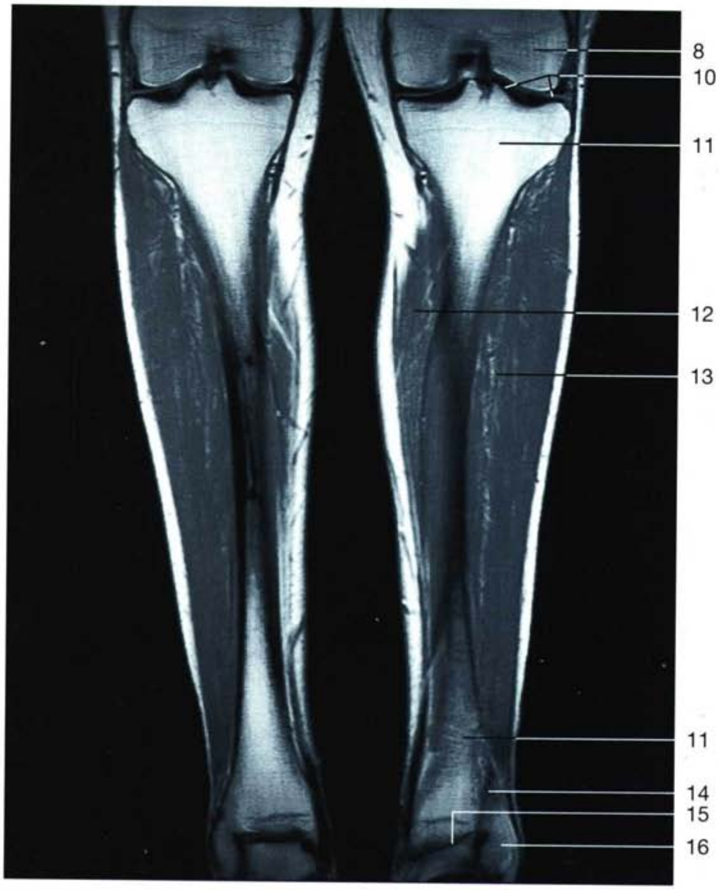

Coronal section through the legs (MR-Scan, courtesy of Dr. A. Heuck, Munich).

1 Sacral promontory
2 Gluteus medius muscle
3 Small intestine and urinary bladder
4 Acetabulum
5 **Head of femur**
6 Greater trochanter of femur
7 Vastus lateralis muscle
8 **Femur**
9 Adductor muscles
10 **Knee joint** with menisci
11 **Tibia**
12 Soleus muscle
13 Tibialis anterior muscle
14 Distal tibiofibular articulation
15 **Talocrural articulation**
16 Fibula (lateral malleolus)

The lower limb (extremity) is specialized for support of the upright posture, locomotion and maintaining balance. In contrast to the upper limb, the lower limb is more restricted in its movements, and the joints are more tight and fixed by strong ligaments. The hip joint is a ball and socket type of synovial joint between head of femur and acetabulum. The knee joint is a hinge type of synovial joint that permits only limited rotation. The talocrural joint is a hinge joint between talus, fibula, and tibia only allowing movements of flexion and extension.
The long axis of the foot is at right angle to that of the leg, thus, forming an effective arch for the upright stance of man.

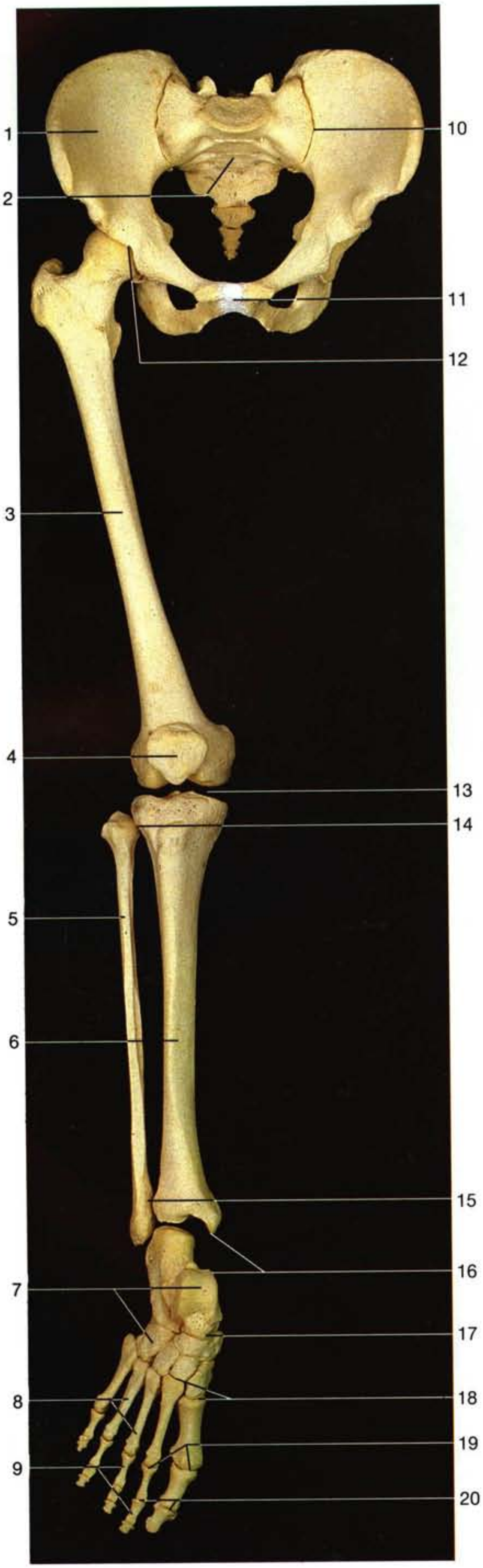

Skeleton of pelvic girdle and lower limb (anterior aspect). The talocrural joint has been dislocated.

A Pelvic girdle
B Thigh
C Leg
D Foot
1 Right hip bone
2 Sacrum
3 Femur
4 Patella
5 Fibula
6 Tibia
7 Tarsal bones
8 Metatarsal bones
9 Phalanges
10 Sacroiliac joint
11 Pubic symphysis
12 Hip joint
13 Knee joint
14 Proximal tibiofibular joint
15 Distal tibiofibular joint
16 Talocrural joint
17 Talocalcaneonavicular joint
18 Tarsometatarsal joints
19 Metatarsophalangeal joints
20 Interphalangeal joints

The pelvic girdle is firmly connected to the vertebral column at the sacroiliac joint. Therefore the body can be kept upright more easily even if only one limb is used for support (as in walking). The mobility of the lower limb is more limited than that of the upper limb.

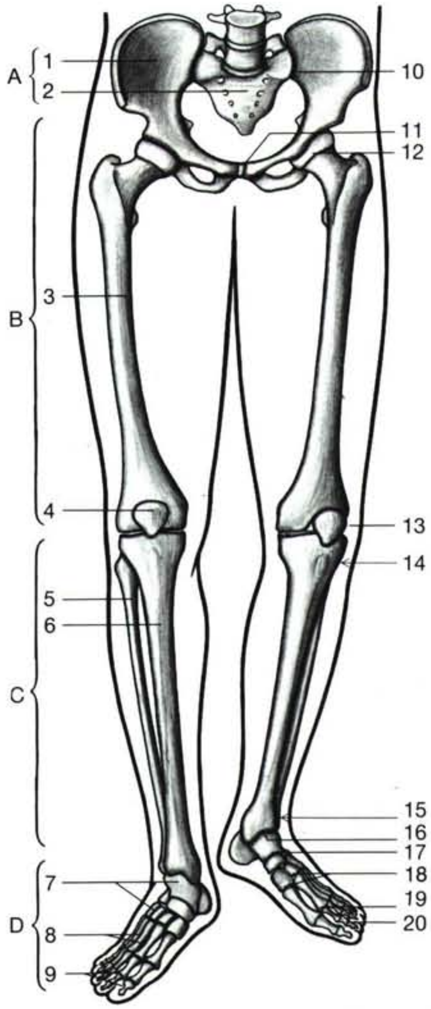

Organization of pelvic girdle and lower limb.

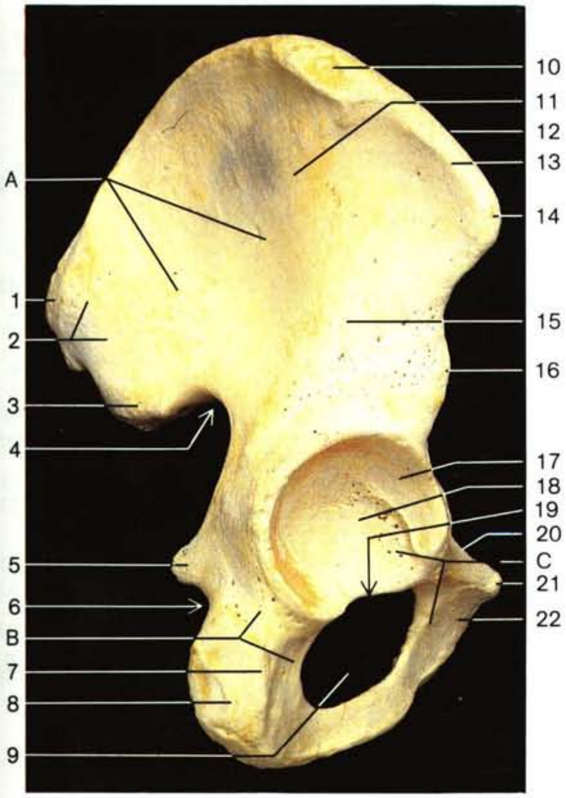

Right hip bone (lateral aspect).

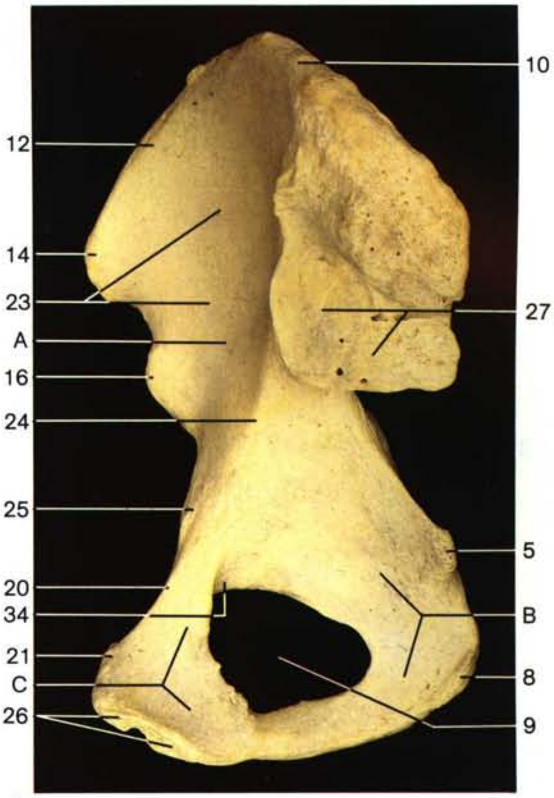

Right hip bone (medial aspect).

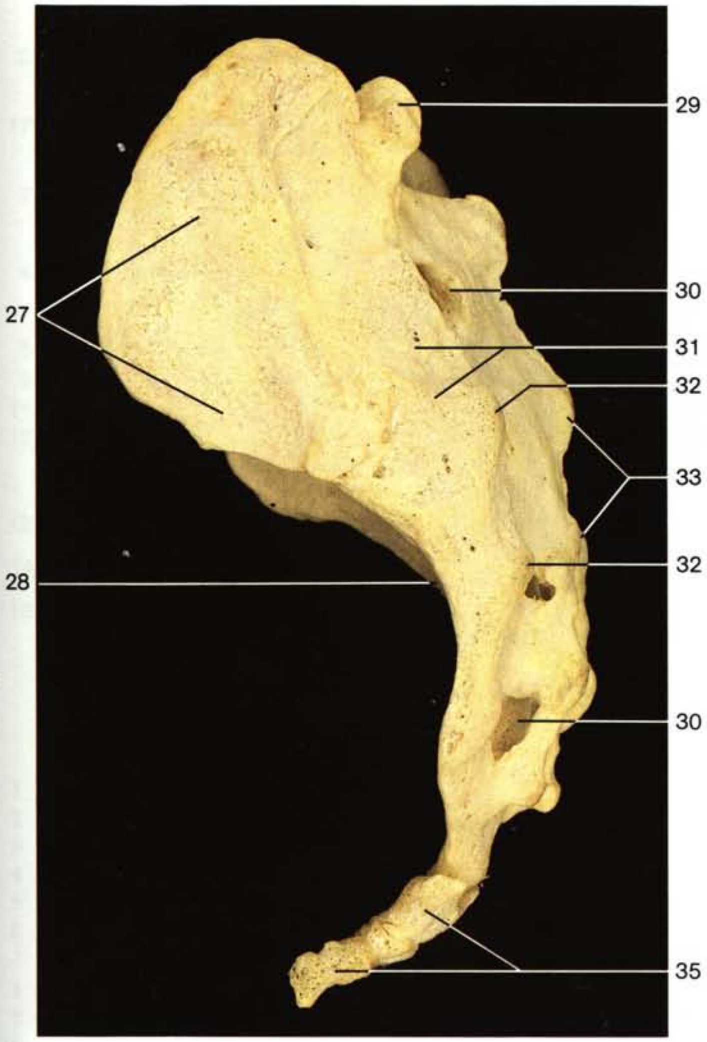

Sacrum and coccyx (lateral aspect).

A	Ilium
B	Ischium
C	Pubis
1	Posterior superior iliac spine
2	Posterior gluteal line
3	Posterior inferior iliac spine
4	Greater sciatic notch
5	Ischial spine
6	Lesser sciatic notch
7	Body of ischium
8	Ischial tuberosity
9	Obturator foramen
10	Iliac crest
11	Anterior gluteal line
12	Internal lip of iliac crest
13	External lip of iliac crest
14	Anterior superior iliac spine
15	Inferior gluteal line
16	Anterior inferior iliac spine
17	Lunate surface of acetabulum
18	Acetabular fossa
19	Acetabular notch
20	Pecten pubis
21	Pubic tubercle
22	Body of pubis
23	Iliac fossa
24	Arcuate line
25	Iliopubic eminence
26	Articular surface of pubis
27	Auricular surface of sacrum
28	Pelvic surface of sacrum
29	Superior articular process of sacrum
30	Dorsal sacral foramina
31	Sacral tuberosity
32	Lateral sacral crest
33	Median sacral crest
34	Obturator groove
35	Coccyx

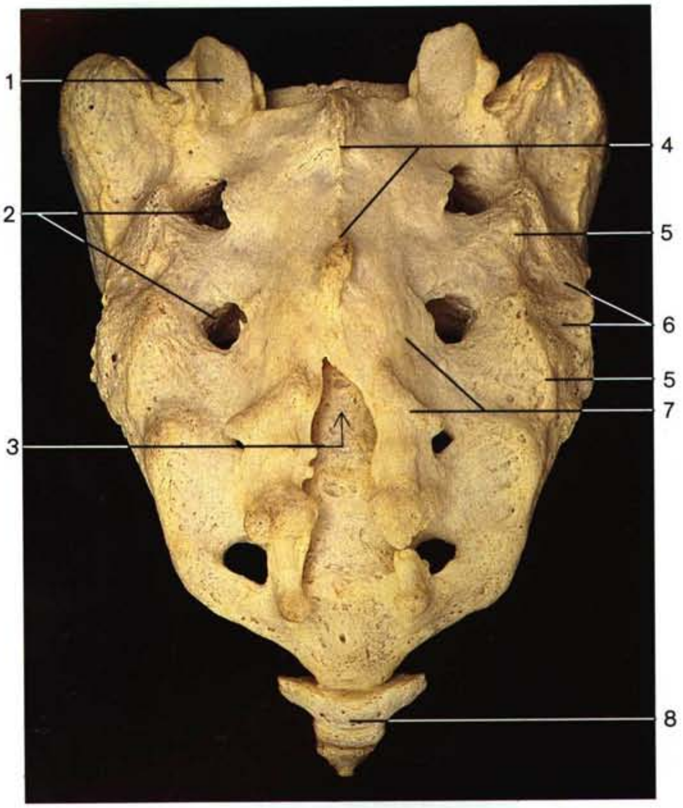

Sacrum (posterior aspect).

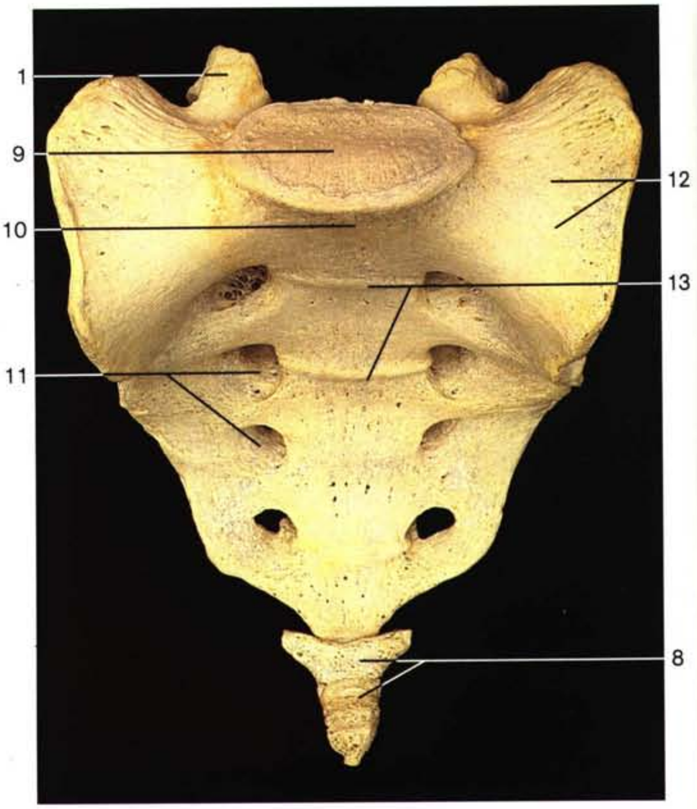

Sacrum (anterior aspect).

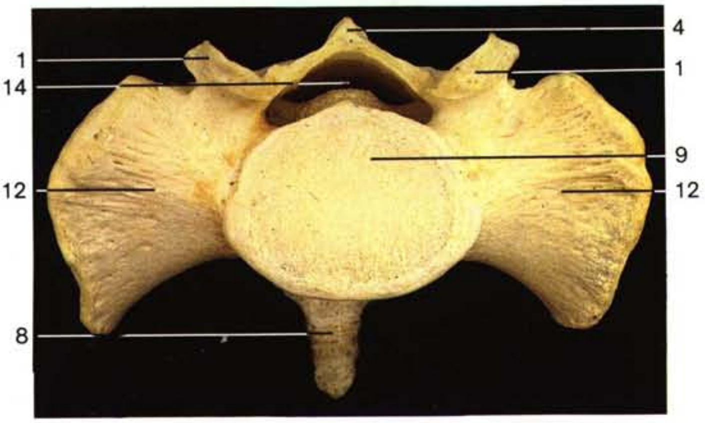

Sacrum (superior aspect).

1 Superior articular process of sacrum
2 Dorsal sacral foramina
3 Sacral hiatus
4 Median sacral crest
5 Lateral sacral crest
6 Sacral tuberosity
7 Intermediate sacral crest
8 Coccyx
9 Base of sacrum
10 Sacral promontory
11 Anterior sacral foramina
12 Lateral part of sacrum (ala)
13 Transverse line of sacrum
14 Sacral canal
15 Linea terminalis
16 True conjugate
17 Diagonal conjugate
18 Transverse diameter
19 Oblique diameter
20 Inferior pelvic aperture or outlet

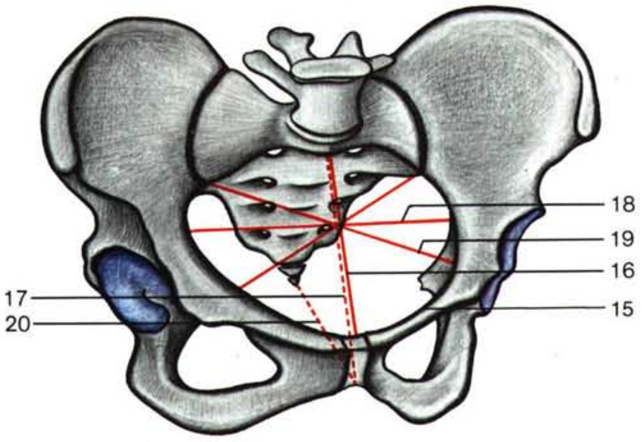

Diameters of pelvis (oblique superior aspect). (Schematic drawing.)

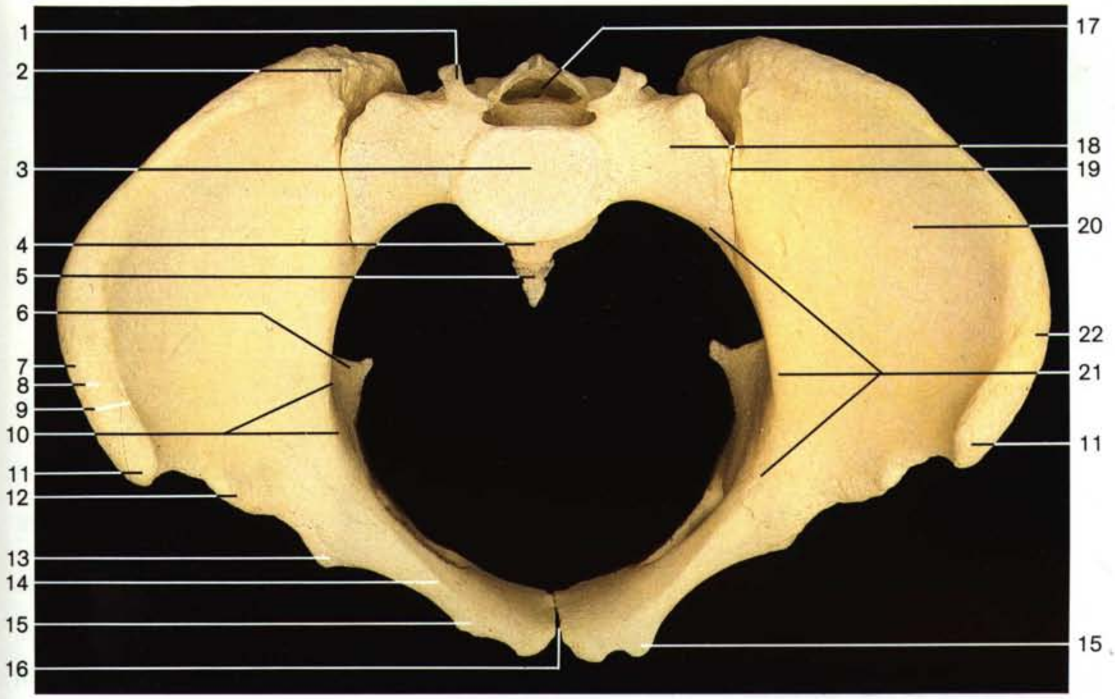
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
22
21
11
15

Female pelvis (superior aspect). Note the differences between the male and the female pelvis, predominantly in the form and dimensions of the sacrum, the superior and inferior apertures and the alae of the ilium.

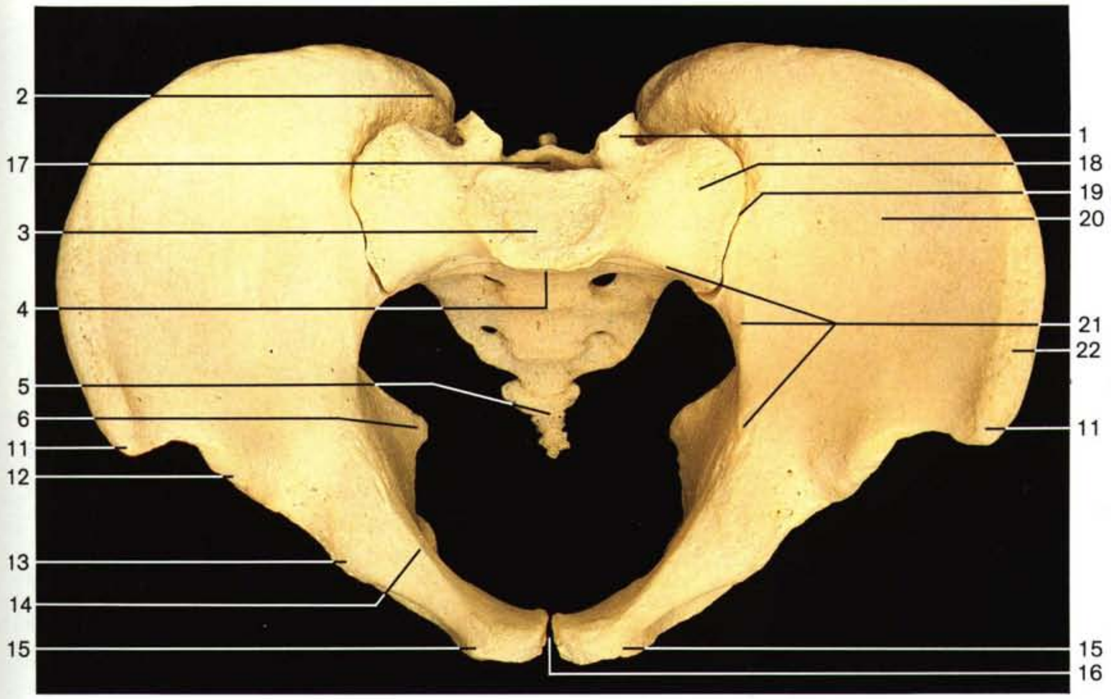
2
17
3
4
5
6
11
12
13
14
15
1
18
19
20
21
22
11
15
16

Male pelvis (superior aspect). Compare with the female pelvis (depicted above).

1 Superior articular process of sacrum
2 Posterior superior iliac spine
3 Base of sacrum
4 Sacral promontory
5 Coccyx
6 Ischial spine
7 External lip } of iliac crest
8 Intermediate line } of iliac crest
9 Internal lip } of iliac crest
10 Arcuate line
11 Anterior superior iliac spine
12 Anterior inferior iliac spine
13 Iliopubic eminence
14 Pecten pubis
15 Pubic tubercle
16 Pubic symphysis
17 Sacral canal
18 Ala of sacrum
19 Position of sacroiliac joint
20 Iliac fossa
21 Linea terminalis
22 Iliac crest

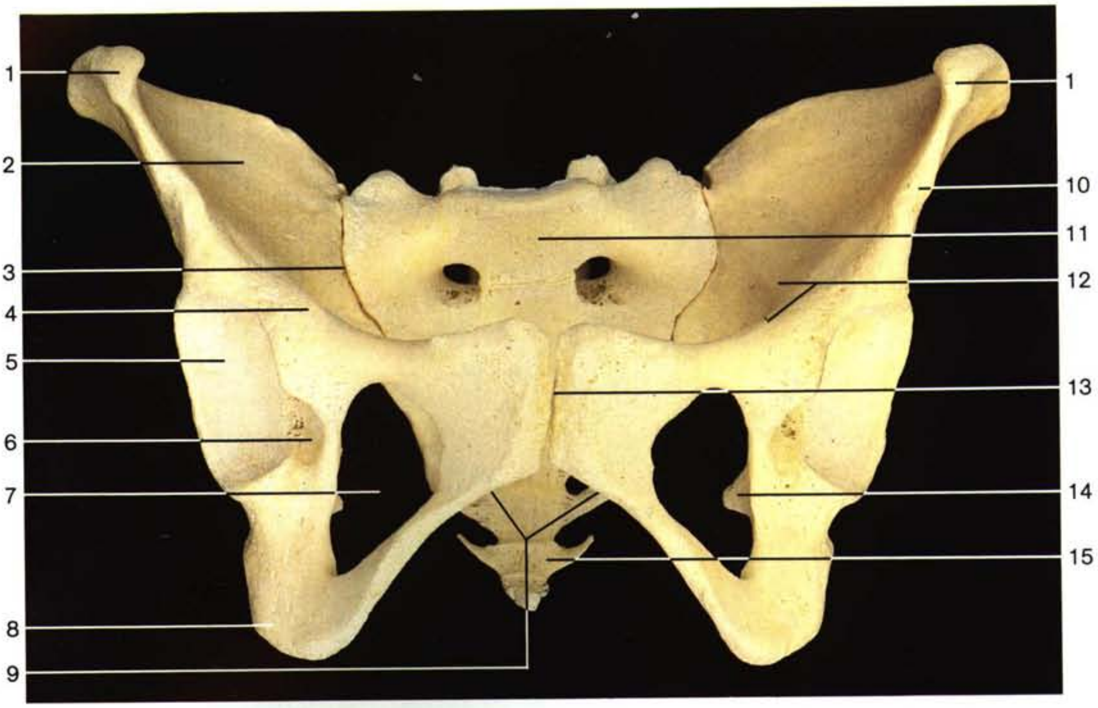

Female pelvis (anterior aspect). Note the differences between the form and dimensions of the male and the female pelvis. The female pubic arch is wider than the male. The obturator foramen in the female pelvis is triangular, while that in the male pelvis is ovoid.

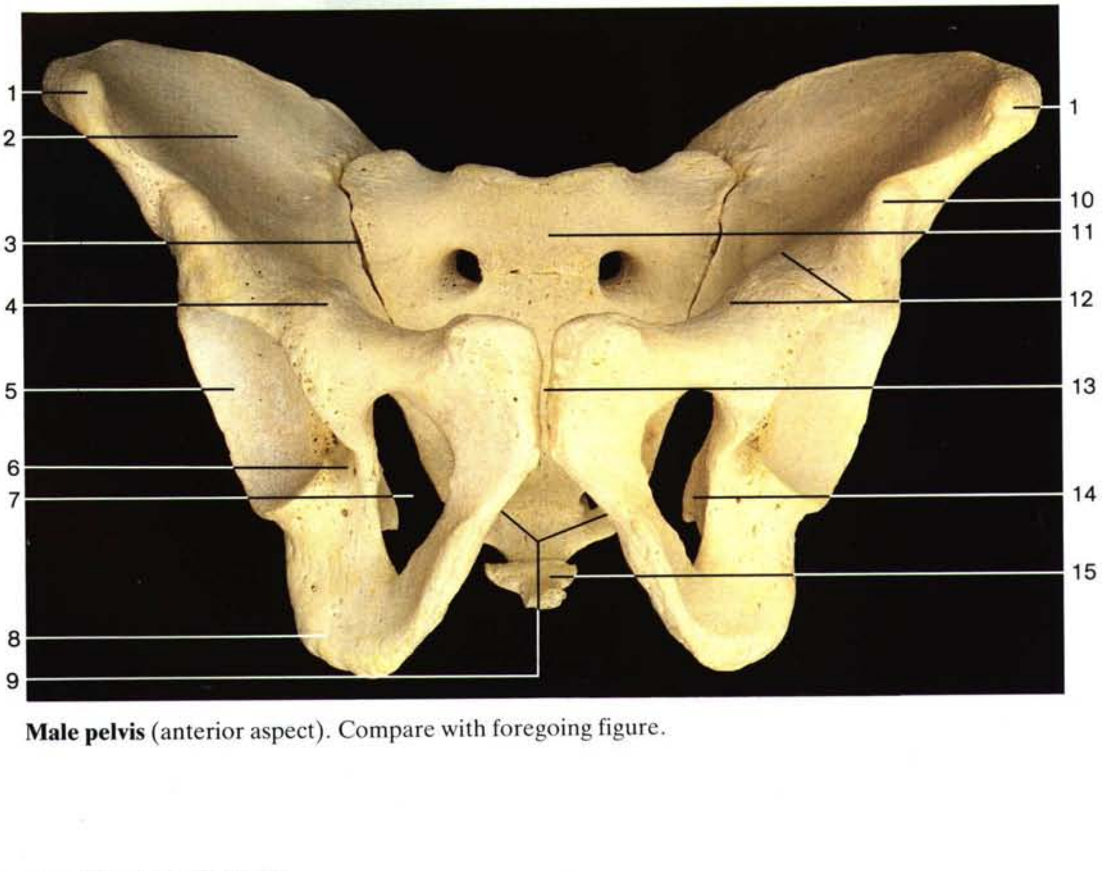

Male pelvis (anterior aspect). Compare with foregoing figure.

1 Anterior superior iliac spine
2 Iliac fossa
3 Position of sacroiliac joint
4 Iliopubic eminence
5 Lunate surface of acetabulum
6 Acetabular notch
7 Obturator foramen
8 Ischial tuberosity
9 Pubic arch
10 Anterior inferior iliac spine
11 Sacrum
12 Linea terminalis (at margin of superior aperture)
13 Pubic symphysis
14 Ischial spine
15 Coccyx

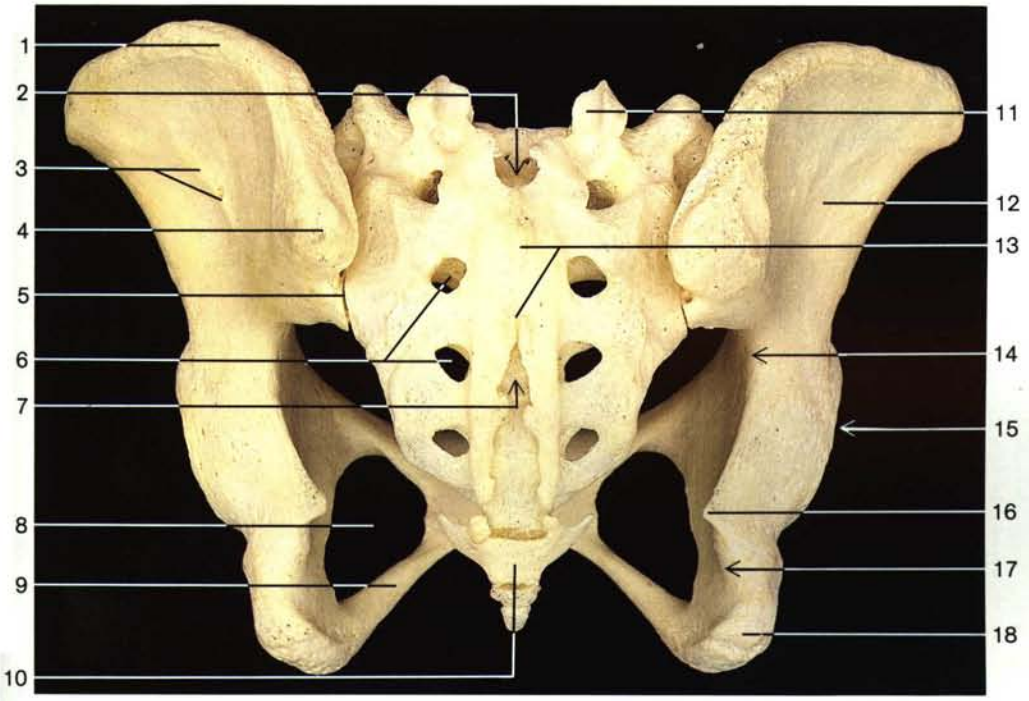

Female pelvis (posterior inferior aspect). Note the differences between the female and male pelvis, especially with respect to the inferior aperture, the shape of the sacrum, the two sciatic notches and the pubic arch.

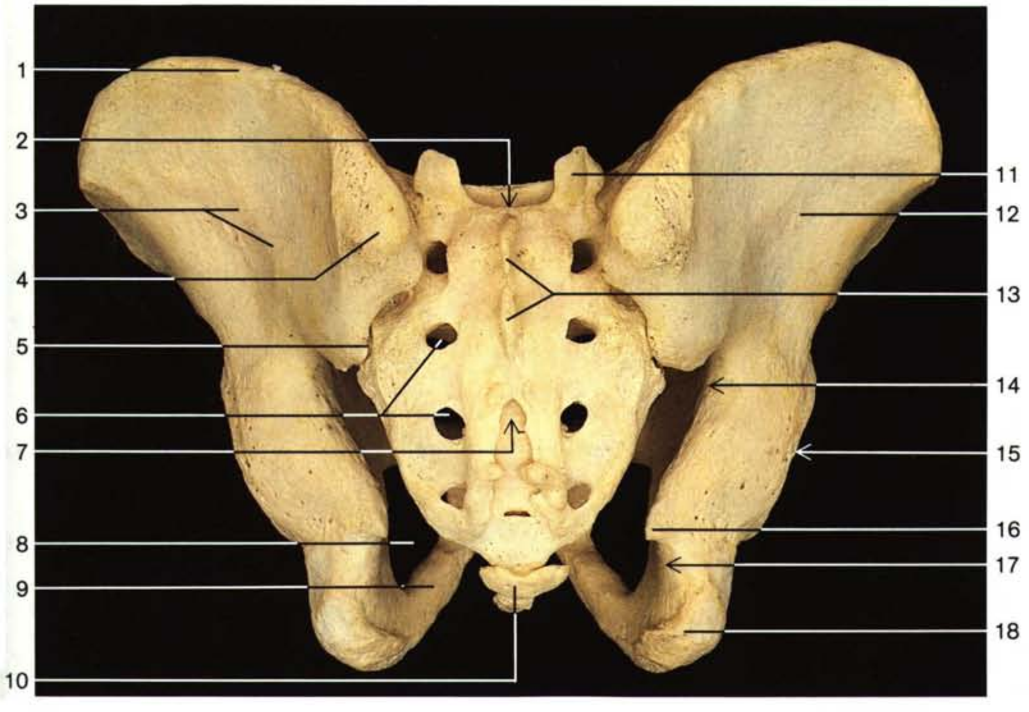

Male pelvis (posterior inferior aspect). Compare with the female pelvis (depicted above).

1	Iliac crest	10	Coccyx
2	Sacral canal	11	Superior articular process of sacrum
3	Posterior gluteal line	12	Gluteal surface of ilium
4	Posterior superior iliac spine	13	Median sacral crest
5	Position of sacroiliac joint	14	Greater sciatic notch
6	Dorsal sacral foramina	15	Position of acetabulum
7	Sacral hiatus	16	Ischial spine
8	Obturator foramen	17	Lesser sciatic notch
9	Ramus of ischium	18	Ischial tuberosity

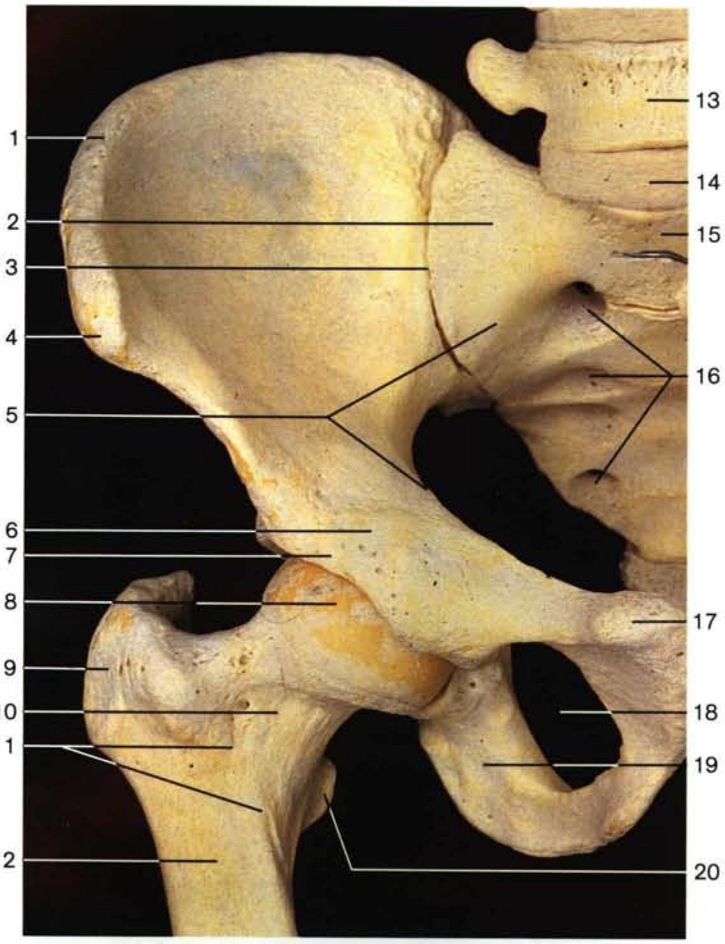

Bones of right hip joint (anterior aspect).

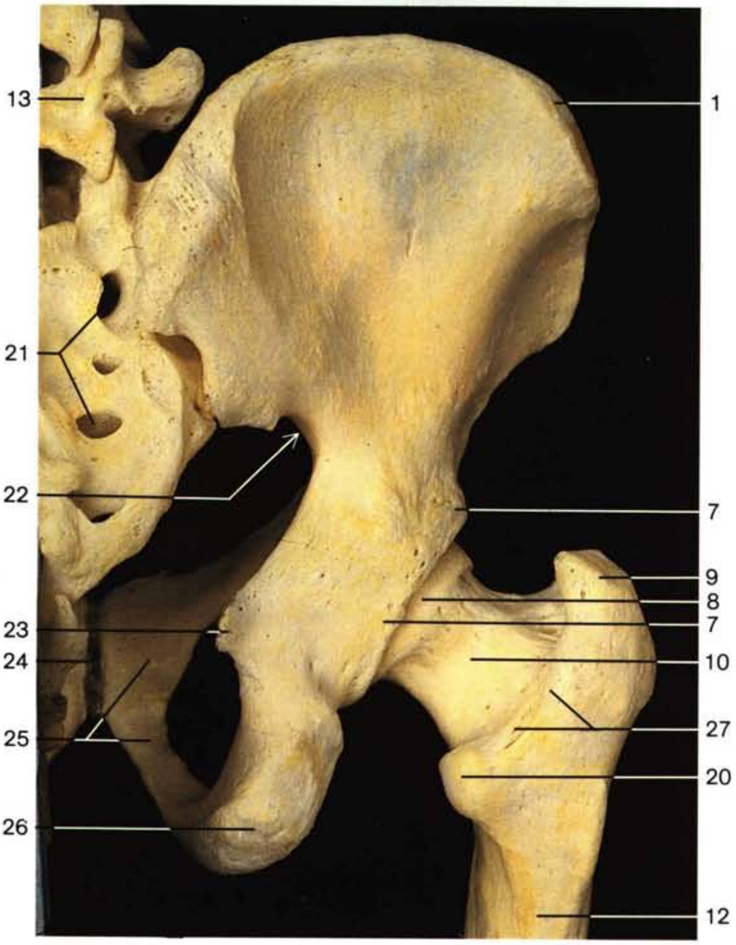

Bones of right hip joint (posterior aspect).

1 Iliac crest
2 Lateral part of sacrum (ala)
3 Position of sacroiliac joint
4 Anterior superior iliac spine
5 Linea terminalis
6 Iliopubic eminence
7 Bony margin of acetabulum
8 Head of femur
9 Greater trochanter
10 Neck of femur
11 Intertrochanteric line
12 Shaft of femur
13 5th lumbar vertebra
14 Imitation intervertebral disc between 5th lumbar vertebra and sacrum
15 Sacral promontory
16 Anterior sacral foramina
17 Pubic tubercle
18 Obturator foramen
19 Ramus of ischium
20 Lesser trochanter
21 Dorsal sacral foramina
22 Greater sciatic notch
23 Ischial spine
24 Pubic symphysis
25 Pubis
26 Ischial tuberosity
27 Intertrochanteric crest
28 Symphysial surface

Diameters of the pelvis

A True conjugate (11–11.5 cm) (Conjugata vera)
B Diagonal conjugate (12.5–13 cm)
C Largest diameter of pelvis
D Inferior pelvic aperture
E Pelvic inclination (60°)

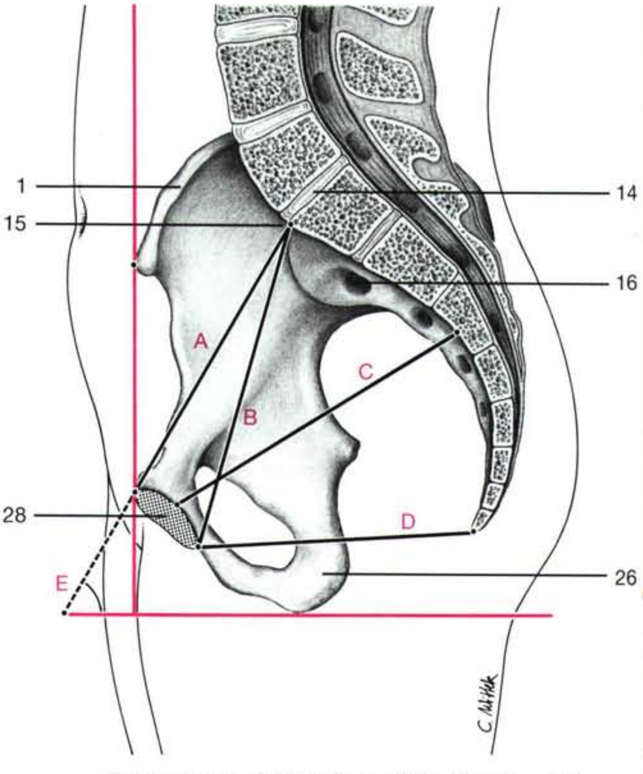

Inclination and diameters of the female pelvis, right half (medial aspect).

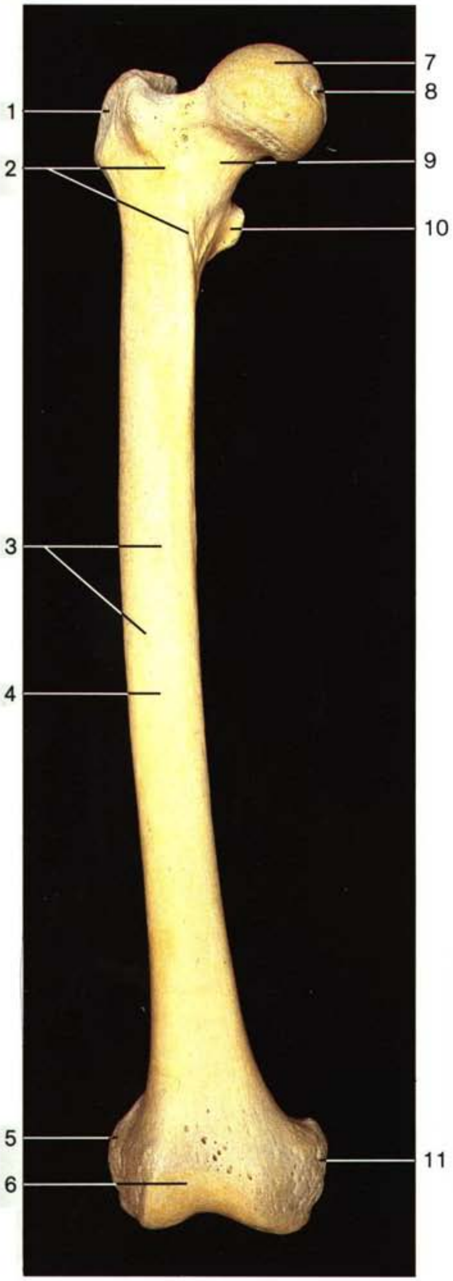
7
8
1
2
9
10
3
4
5
11
6

Right femur (anterior aspect).

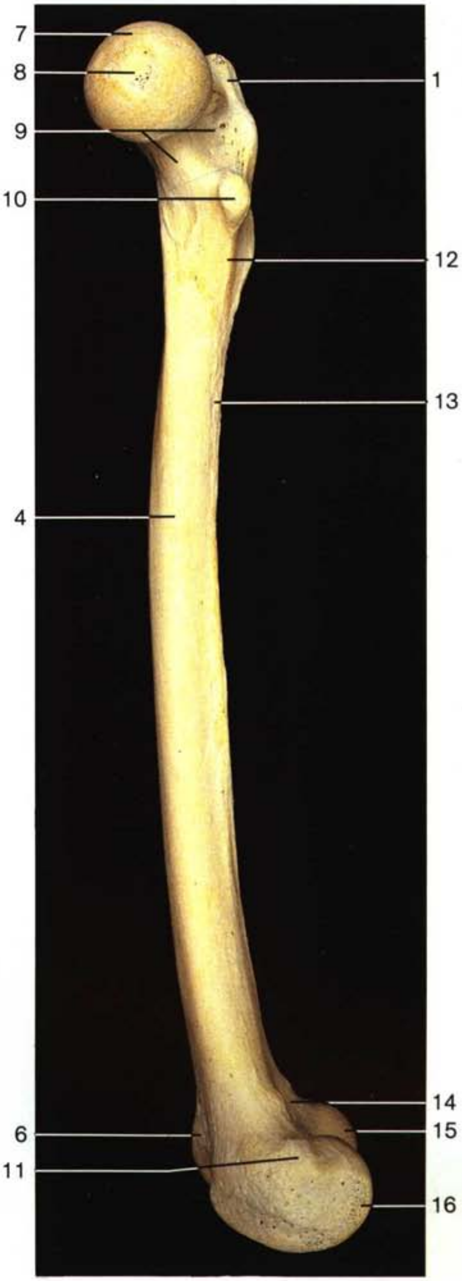
7
8
1
9
10
12
13
4
14
15
6
11
16

Right femur (medial aspect).

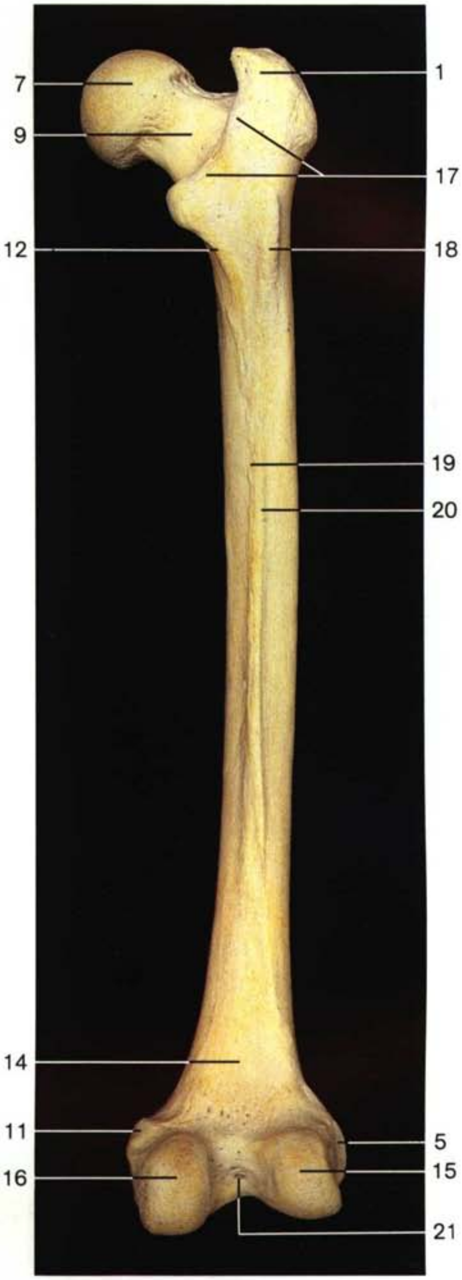
7
1
9
17
12
18
19
20
14
11
5
16
15
21

Right femur (posterior aspect).

1 Greater trochanter
2 Intertrochanteric line
3 Nutrient foramina
4 Shaft of femur (diaphysis)
5 Lateral epicondyle
6 Patellar surface
7 Head
8 Fovea of head
9 Neck
10 Lesser trochanter
11 Medial epicondyle
12 Pectineal line
13 Linea aspera
14 Popliteal surface
15 Lateral condyle
16 Medial condyle
17 Intertrochanteric crest
18 Third trochanter
19 Medial lip of linea aspera
20 Lateral lip of linea aspera
21 Intercondylar fossa

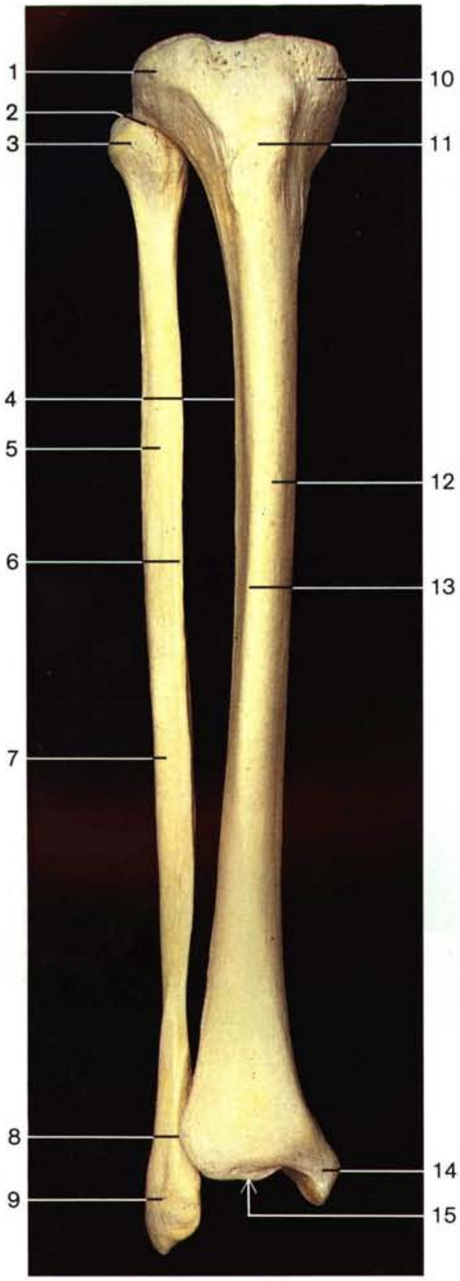

Bones of leg, right tibia and fibula (anterior aspect).

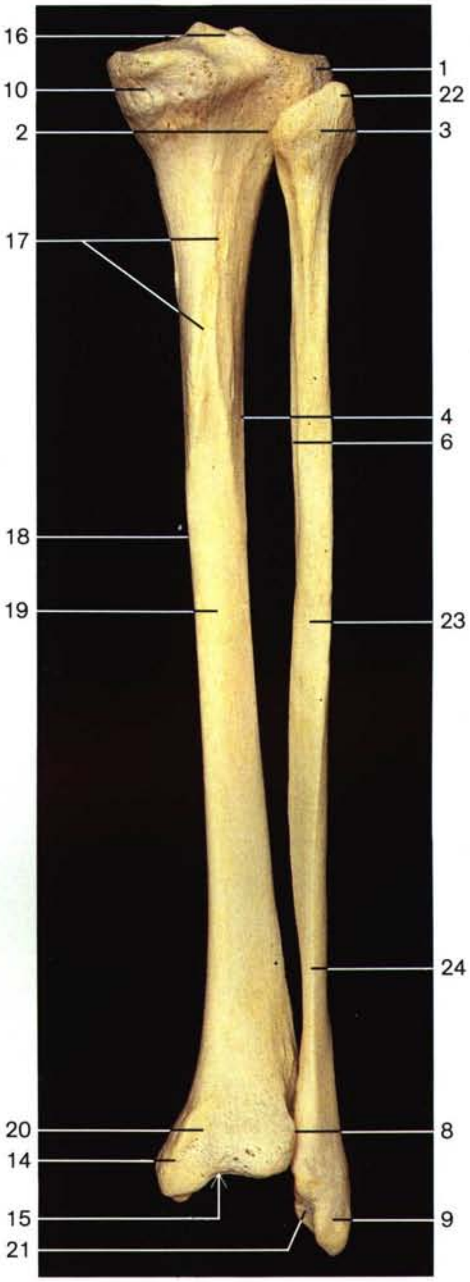

Bones of leg, right tibia and fibula (posterior aspect).

1 Lateral condyle of tibia
2 Position of tibiofibular joint
3 Head of fibula
4 Interosseous border of tibia
5 Shaft of fibula
6 Interosseous border of fibula
7 Lateral surface of fibula
8 Position of tibiofibular joint
9 Lateral malleolus
10 Medial condyle of tibia
11 Tuberosity of tibia
12 Shaft of tibia (diaphysis)
13 Anterior margin of tibia
14 Medial malleolus
15 Inferior articular surface of tibia
16 Intercondylar eminence
17 Soleal line
18 Medial border of tibia
19 Posterior surface of tibia
20 Malleolar sulcus of tibia
21 Malleolar articular surface of fibula
22 Apex of head of fibula
23 Posterior surface of fibula
24 Posterior border of fibula
25 Medial intercondylar tubercle
26 Posterior intercondylar area
27 Anterior intercondylar area
28 Lateral intercondylar tubercle

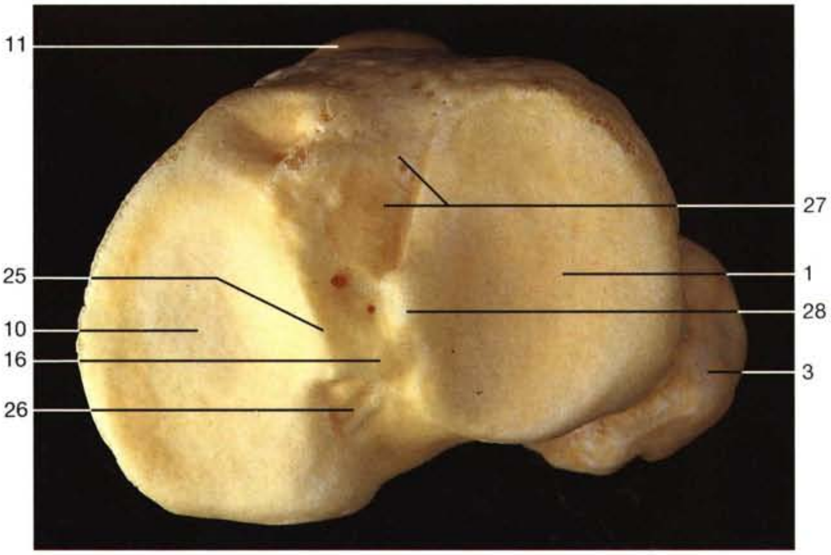

Upper end of right tibia with fibula (from above), anterior margin of tibia above. Superior articular surface of tibia.

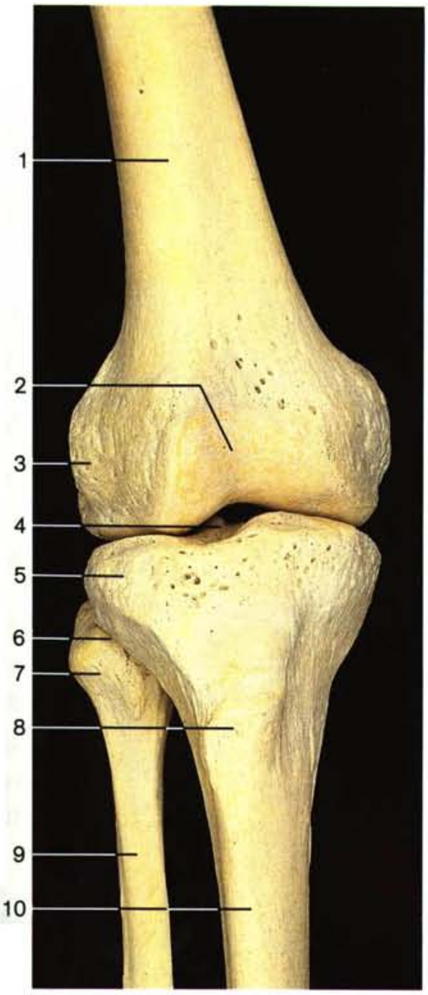

Bones of right knee joint
(anterior aspect).

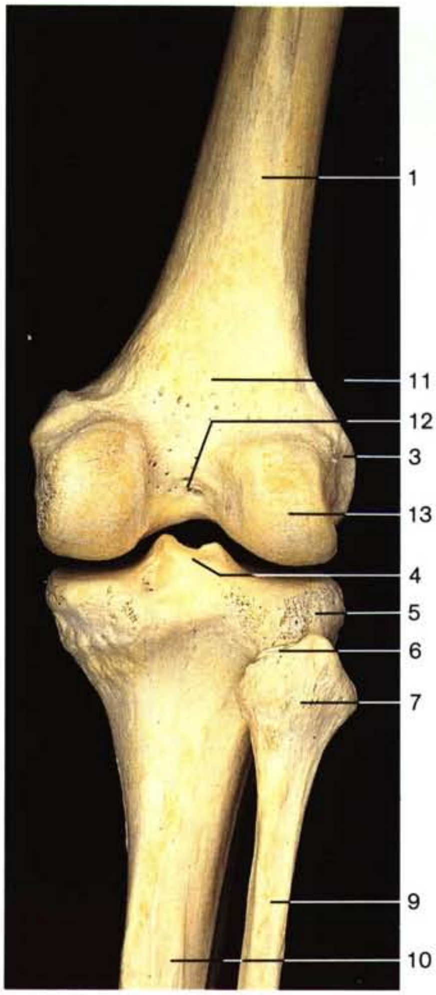

Bones of right knee joint
(posterior aspect).

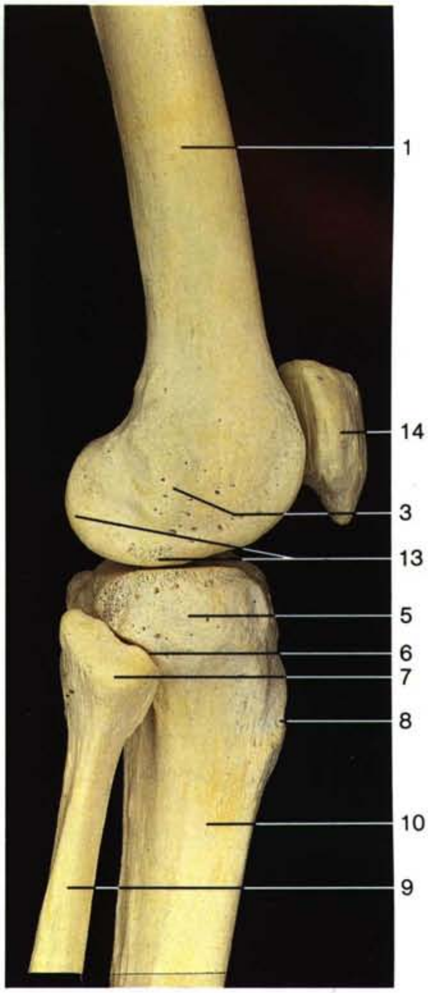

Bones of right knee joint
(lateral aspect).

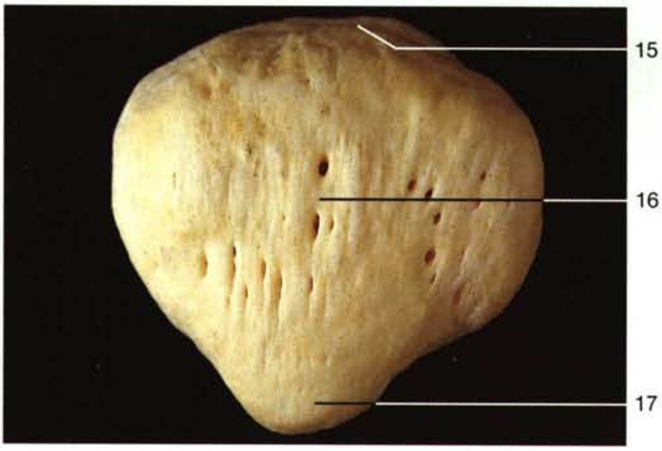

Right patella (anterior aspect).

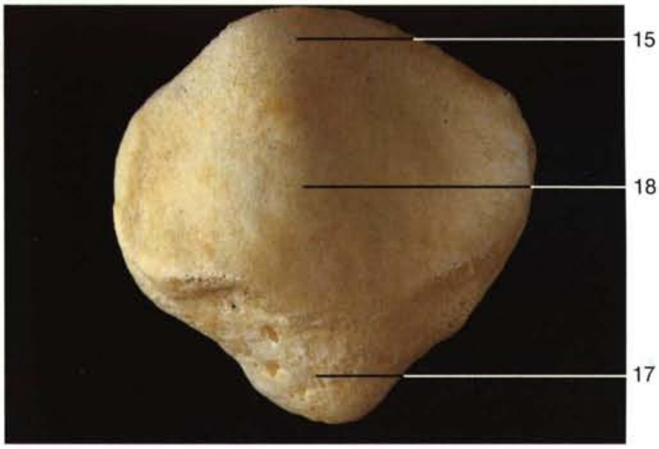

Right patella (posterior aspect).

1 **Femur**
2 Patellar surface of femur
3 Lateral epicondyle of femur
4 Intercondylar eminence of tibia
5 Lateral condyle of tibia
6 Position of tibiofibular joint
7 Head of fibula
8 Tuberosity of **tibia**
9 **Fibula**
10 Shaft of tibia
11 Popliteal surface of femur
12 Intercondylar fossa of femur
13 Lateral condyle of femur
14 **Patella**
15 Base of patella
16 Anterior surface of patella
17 Apex of patella
18 Articular surface of patella

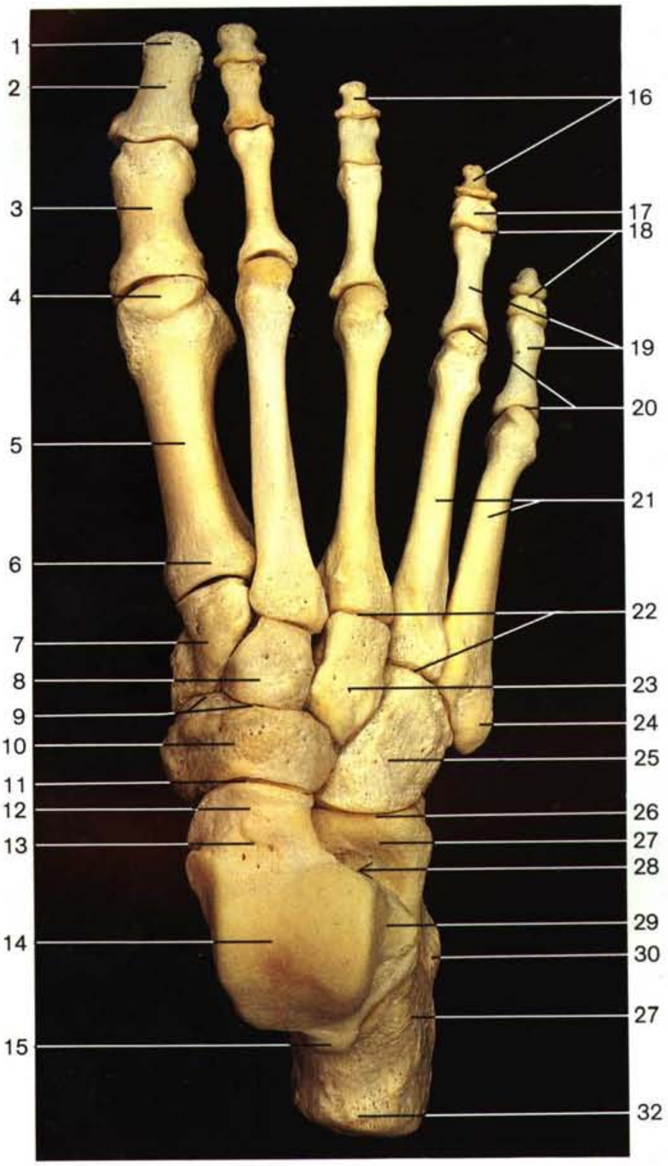
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
27
32

Bones of right foot (dorsal aspect).

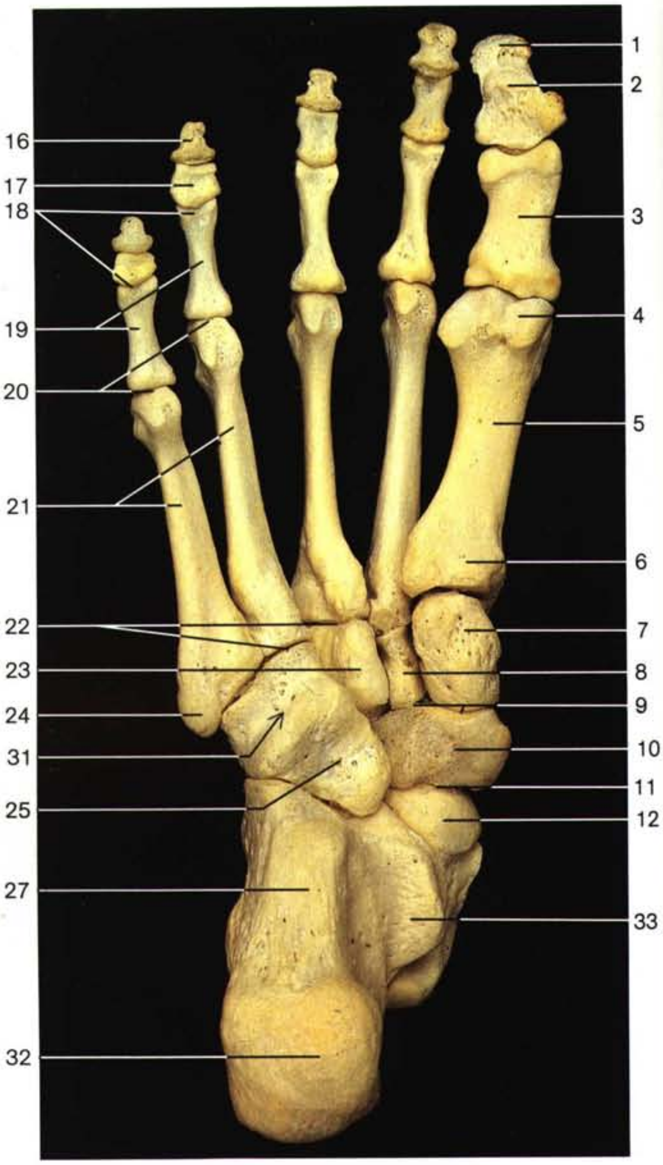
16
17
18
19
20
21
22
23
24
31
25
27
32
1
2
3
4
5
6
7
8
9
10
11
12
33

Bones of right foot (plantar aspect).

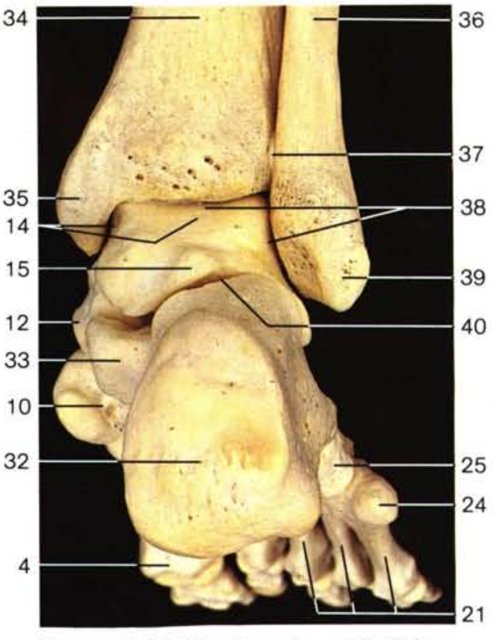
34
35
14
15
12
33
10
32
4
36
37
38
39
40
25
24
21

Bones of right foot together with tibia and fibula (posterior aspect).

1 Tuberosity of distal phalanx of great toe
2 Distal phalanx of great toe
3 Proximal phalanx of great toe
4 Head of first metatarsal bone
5 First metatarsal bone
6 Base of first metatarsal bone
7 Medial cuneiform bone
8 Intermediate cuneiform bone
9 Position of **cuneonavicular joint**
10 Navicular bone

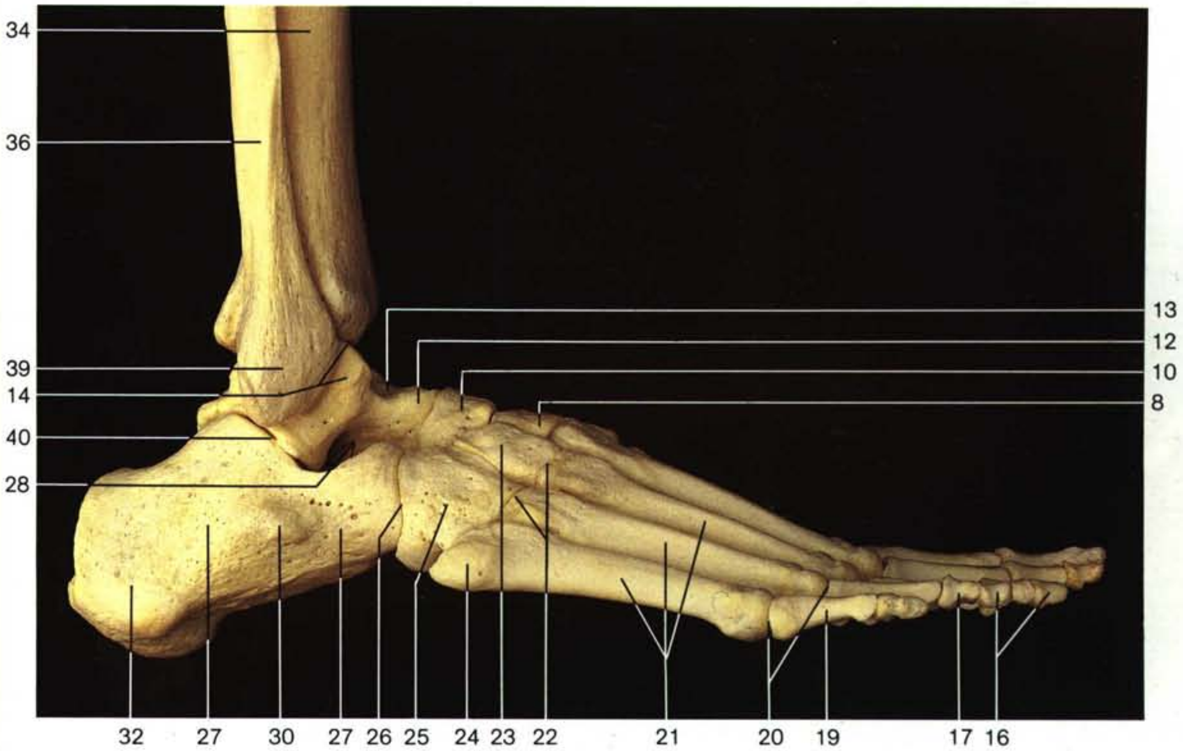

Bones of right foot, tibia and fibula (lateral aspect).

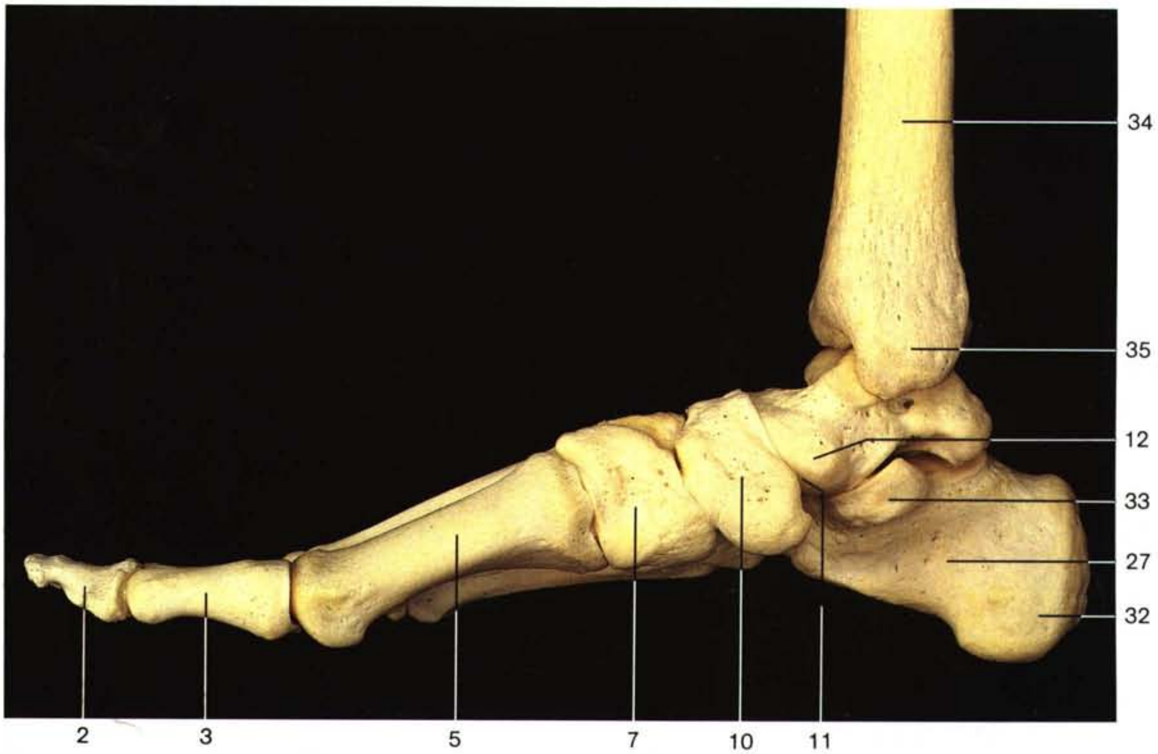

Bones of right foot, tibia and fibula (medial aspect).

11 Position of **talocalcaneonavicular joint**
12 Head of talus
13 Neck of talus
14 Trochlea of talus
15 Posterior talar process
16 Distal phalanges
17 Middle phalanges
18 Position of **interphalangeal joints**
19 Proximal phalanges
20 Position of **metatarsophalangeal joints**
21 Metatarsal bones
22 Position of **tarsometatarsal joints**
23 Lateral cuneiform bone
24 Tuberosity of 5th metatarsal bone
25 Cuboid bone
26 Position of **calcaneocuboid joint**
27 Calcaneus
28 Tarsal sinus
29 Lateral malleolar surface of talus
30 Peroneal trochlea of calcaneus
31 Groove for tendon of peroneus longus
32 Calcaneal tuberosity
33 Sustentaculum tali
34 Tibia
35 Medial malleolus
36 Fibula
37 Position of tibiofibular syndesmosis
38 Position of **talocrural joint**
39 Lateral malleolus
40 Position of **subtalar joint**

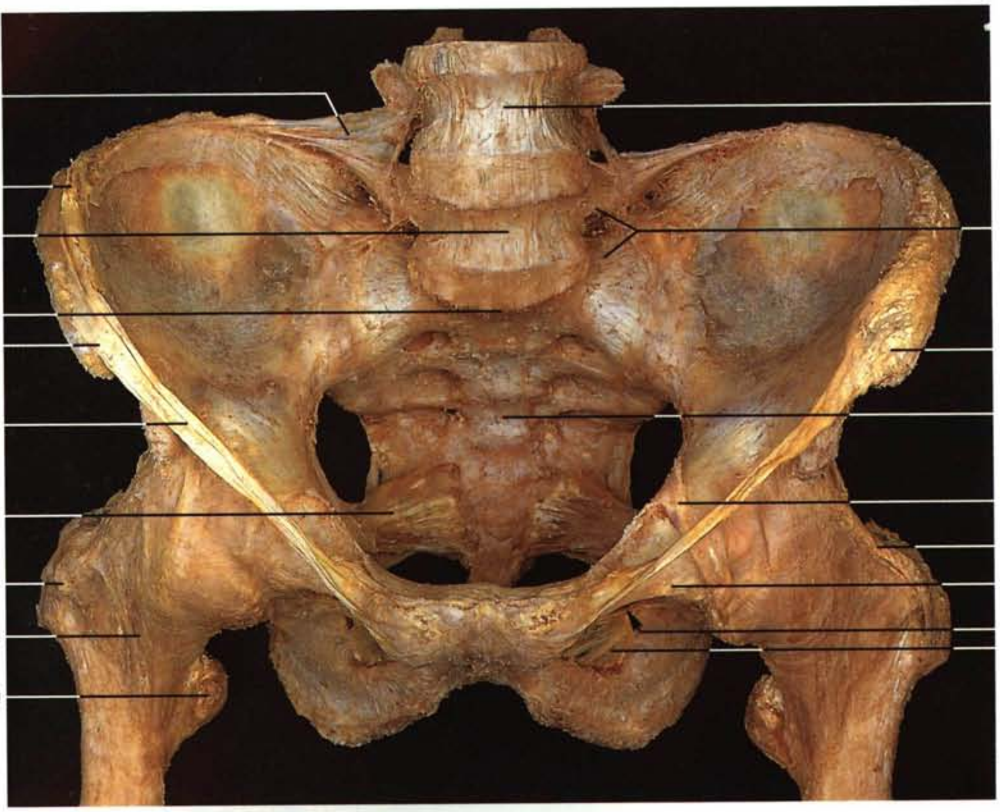

Ligaments of pelvis and hip joint (anterior aspect).

1 **Iliolumbar ligament**
2 Iliac crest
3 Fifth lumbar vertebra
4 Sacral promontory
5 Anterior superior iliac spine
6 Inguinal ligament
7 Sacrospinous ligament
8 Greater trochanter
9 **Iliofemoral ligament** (vertical band)
10 Lesser trochanter
11 Forth lumbar vertebra
12 Iliolumbar and ventral sacroiliac ligaments
13 Sacrum
14 Iliopectineal arch
15 **Iliofemoral ligament** (horizontal band)
16 Obturator canal
17 Obturator membrane
18 **Greater sciatic foramen**
19 Sacrospinous ligament
20 Sacrotuberous ligament
21 **Lesser sciatic foramen**
22 Ischial tuberosity
23 **Ischiofemoral ligament**
24 Intertrochanteric crest
25 Femur
26 Articular capsule of hip joint
27 Dorsal sacroiliac ligaments
28 Coccyx with superficial dorsal sacrococcygeal ligament
29 Head of femur
30 Articular cartilage of head of femur
31 Articular cavity of hip joint
32 Acetabular lip
33 Spongy bone
34 Ligament of head of femur
35 Pubofemoral ligament
36 Zona orbicularis

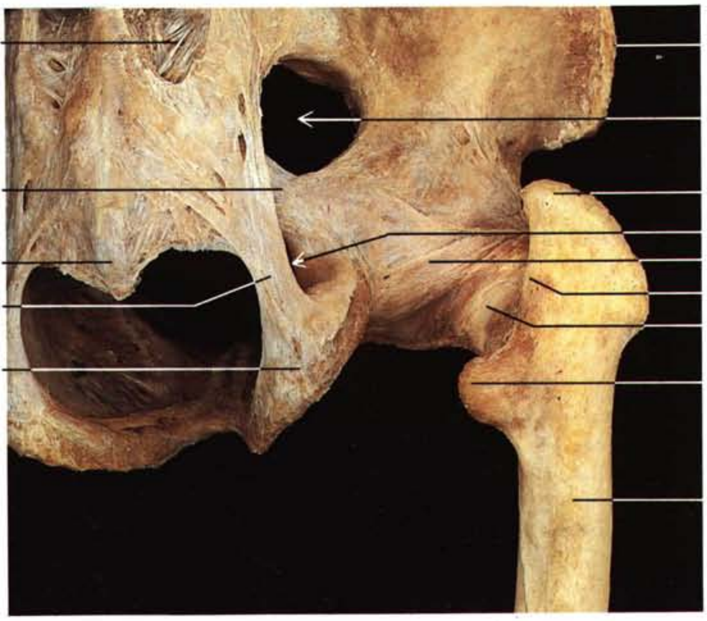

Ligaments of pelvis and hip joint (left anterior aspect).

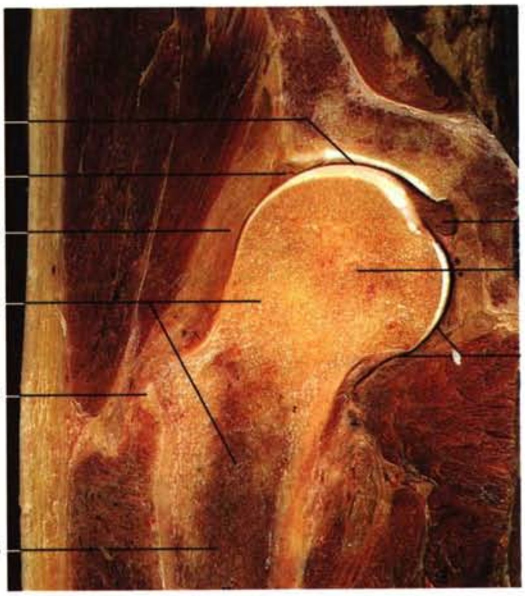

Coronal section of right hip joint (anterior view).

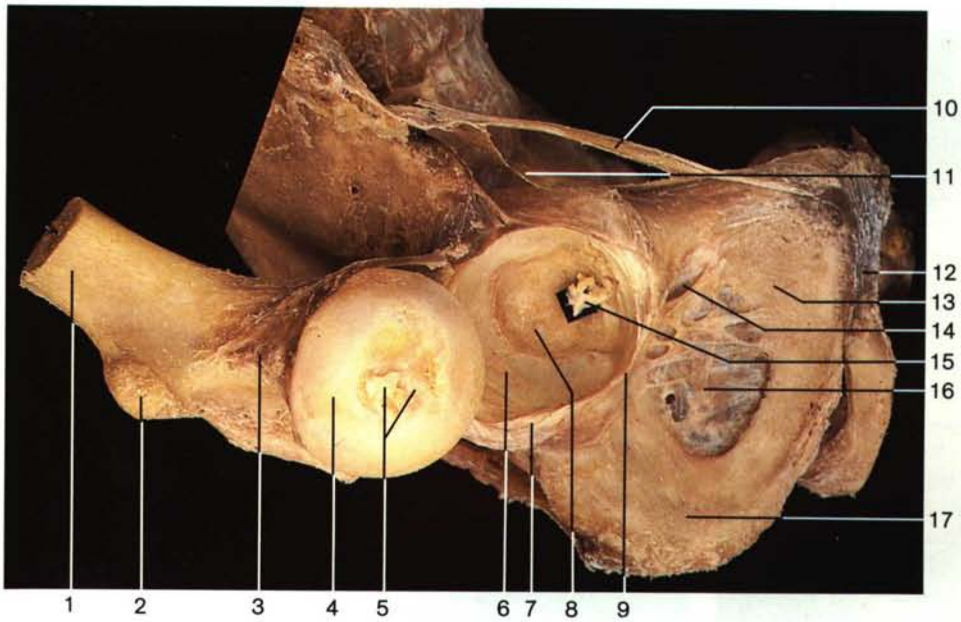

Right hip joint, opened (lateral anterior aspect). The ligament of the head of the femur has been divided and the femur has been posteriorly reflected.

1	Femur
2	Lesser trochanter
3	Neck of femur
4	Head of femur
5	Fovea of head with remnants of ligament of head
6	Lunate surface of acetabulum
7	Acetabular lip
8	Acetabular fossa
9	Transverse acetabular ligament
10	Inguinal ligament
11	Iliopectineal arch
12	Pubic symphysis
13	Pubic bone
14	Obturator canal
15	Ligament of head of femur
16	Obturator membrane
17	Ischium
18	Anterior longitudinal ligament (level of fifth lumbar vertebra)
19	Sacral promontory
20	Iliolumbar ligament
21	Iliac crest
22	Anterior superior iliac spine
23	Iliofemoral ligament (horizontal band)
24	Iliofemoral ligament (vertical band)
25	Greater trochanter
26	Pubofemoral ligament
27	Anterior inferior iliac spine
28	Ventral sacroiliac ligaments
29	**Sacrospinous ligament**
30	**Sacrotuberous ligament**
31	Intertrochanteric line
32	**Ischiofemoral ligament**
33	Zona orbicularis

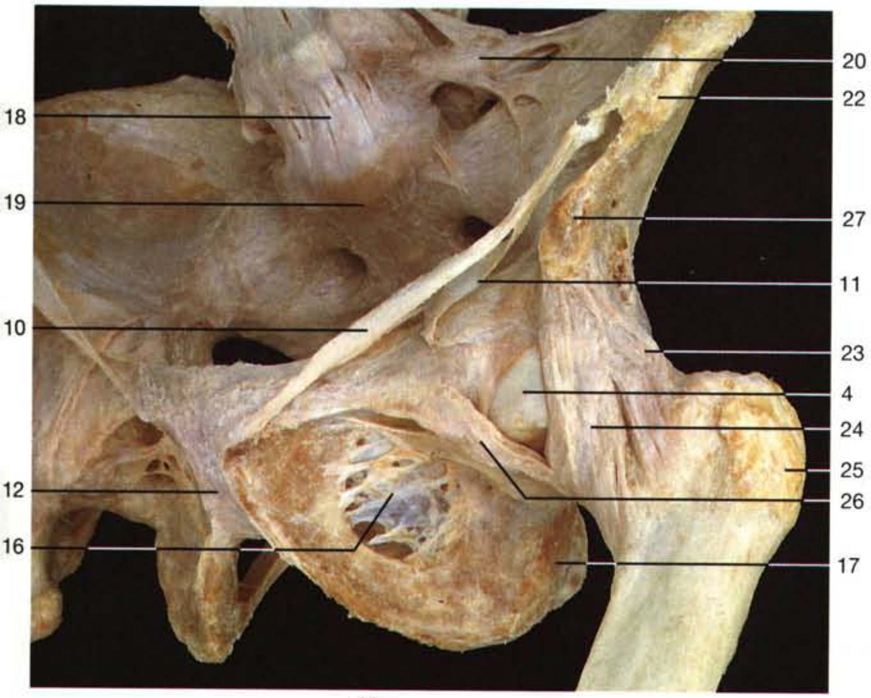

Ligaments of the pelvis and hip joint (antero-lateral aspect).

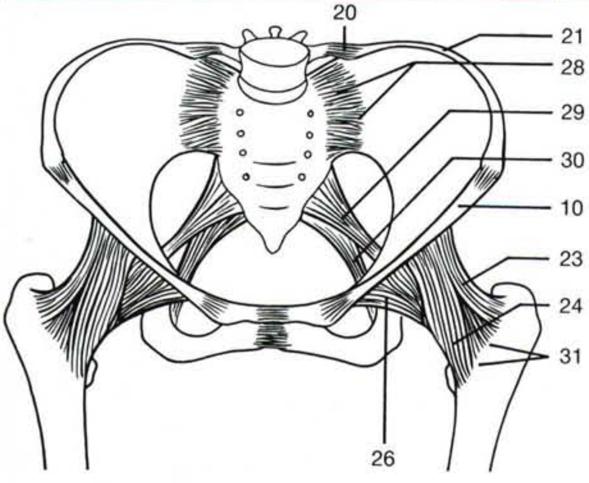

Ligaments of hip joint (anterior aspect). (Schematic drawing.)

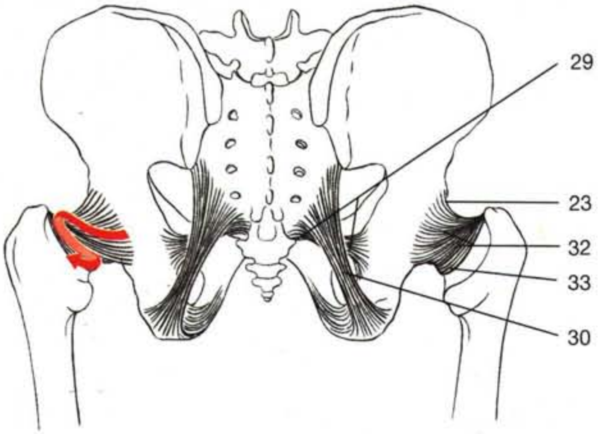

Ligaments of hip joints (posterior aspect). (Schematic drawing.)

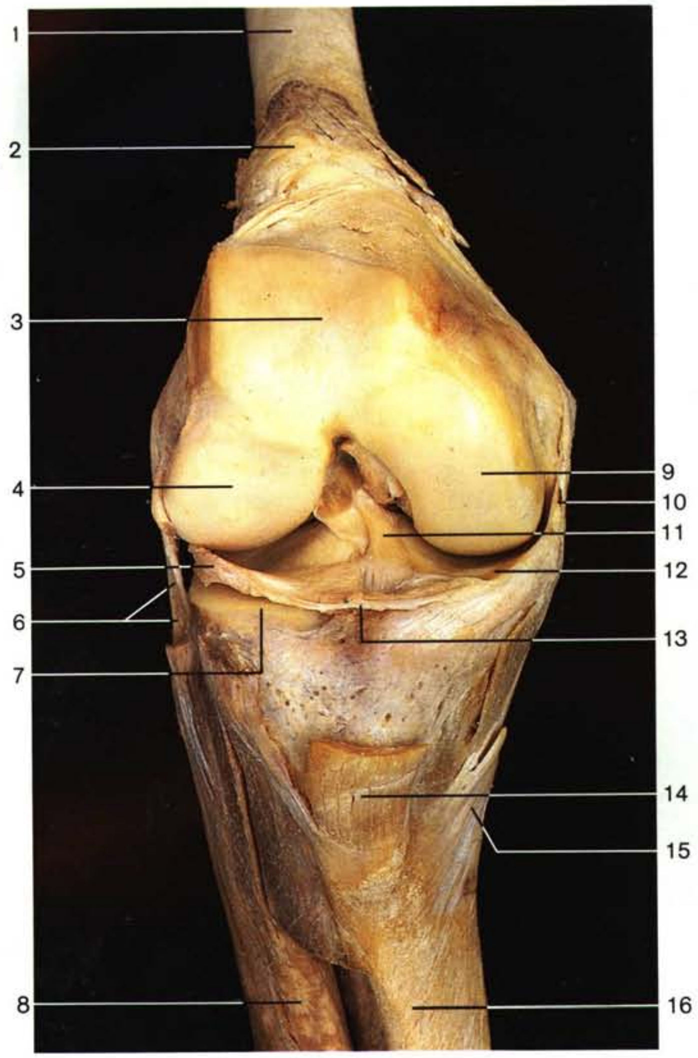

Right knee joint (opened) **with ligaments** (anterior aspect). The patella and articular capsule have been removed and the femur slightly flexed.

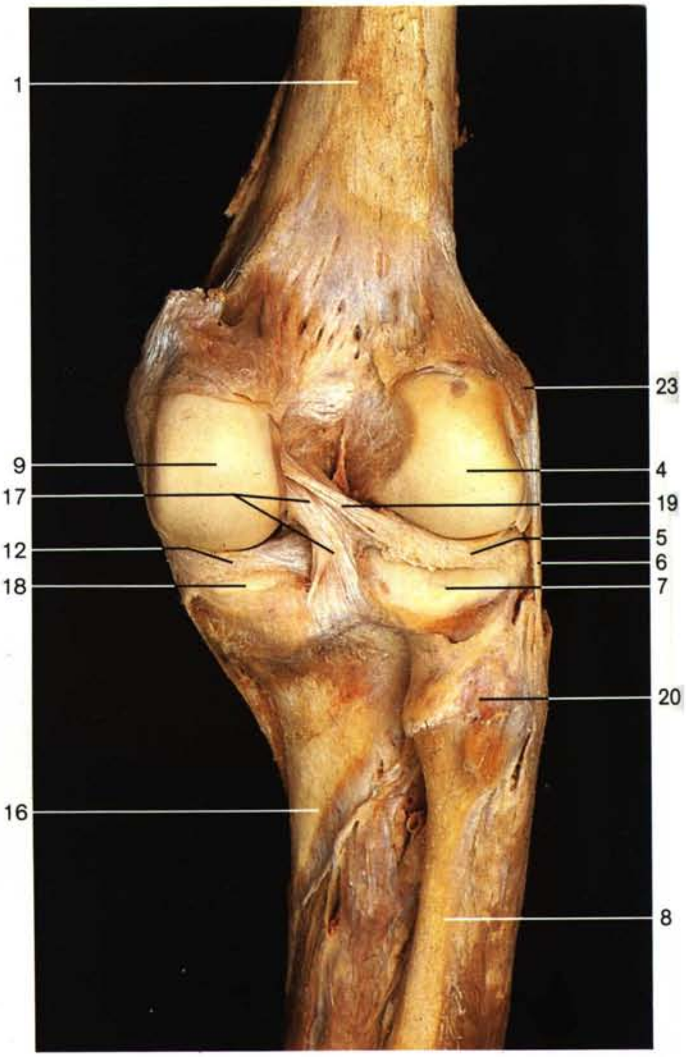

Right knee joint with ligaments (posterior aspect). The joint is extended and the articular capsule has been removed.

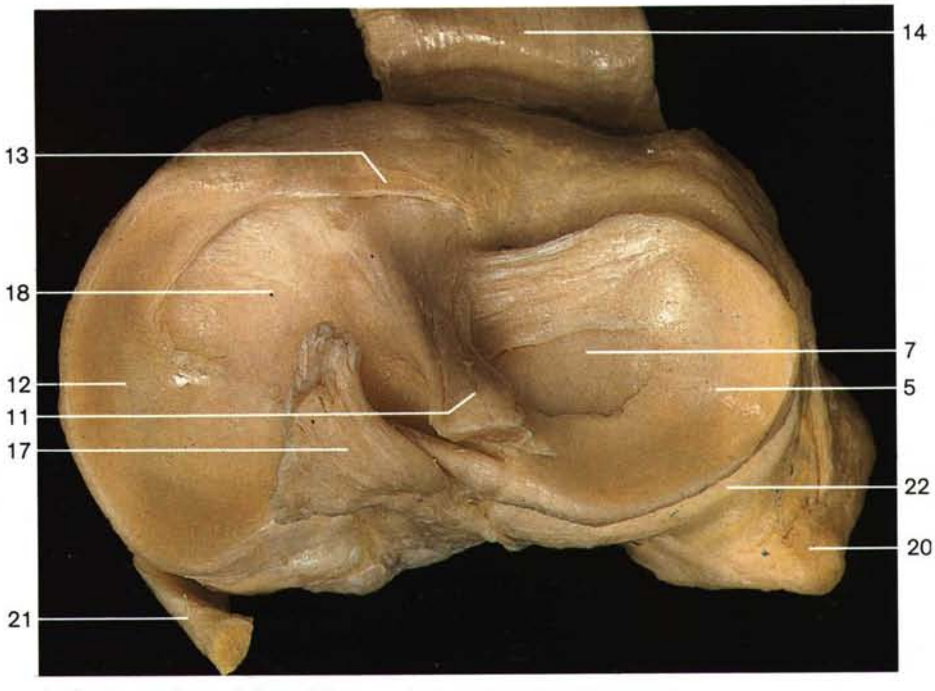

Articular surface of right tibia, menisci, and cruciate ligaments (superior aspect). Anterior margin of tibia above.

1 Femur
2 Articular capsule with **suprapatellar bursa**
3 Patellar surface
4 Lateral condyle of femur
5 **Lateral meniscus of knee joint**
6 Fibular collateral ligament
7 Lateral condyle of tibia (superior articular surface)
8 Fibula
9 Medial condyle of femur
10 Tibial collateral ligament
11 **Anterior cruciate ligament**
12 **Medial meniscus of knee joint**
13 Transverse ligament of knee
14 Patellar ligament
15 Common tendon of sartorius, semitendinosus and gracilis muscles
16 Tibia
17 **Posterior cruciate ligament**
18 Medial condyle of tibia (superior articular surface)
19 Posterior meniscofemoral ligament
20 Head of fibula
21 Tendon of semimembranosus muscle
22 Posterior attachment of articular capsule of knee joint
23 Lateral epicondyle of femur

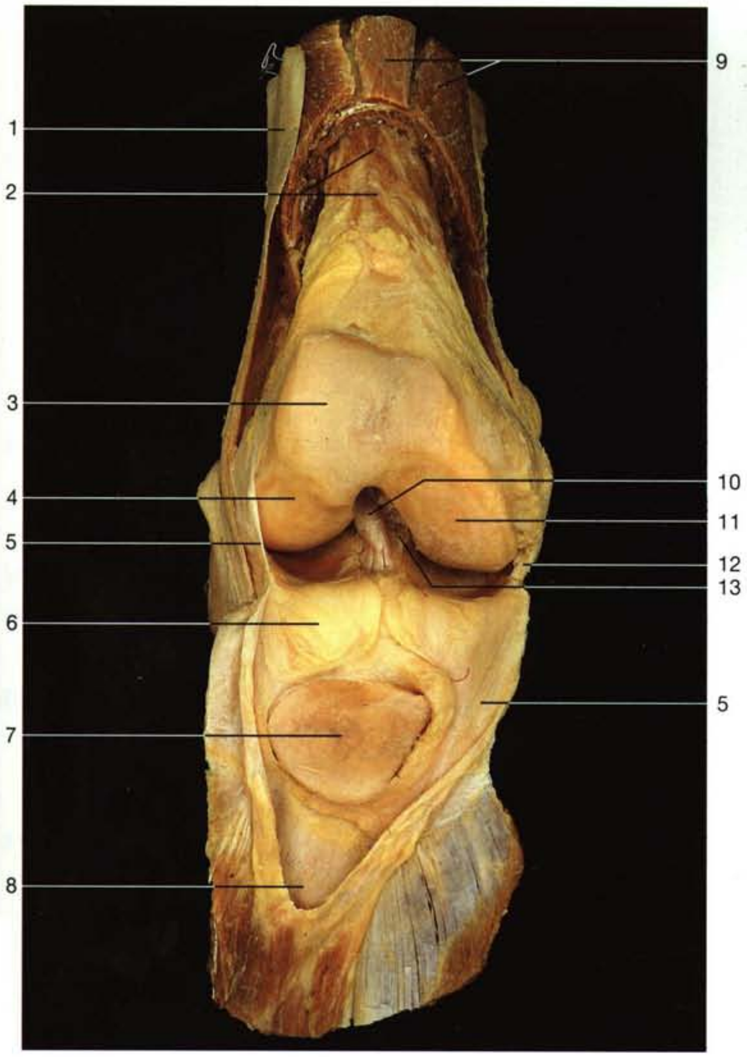

Right knee joint, opened (anterior aspect). Patellar ligament with patella reflected.

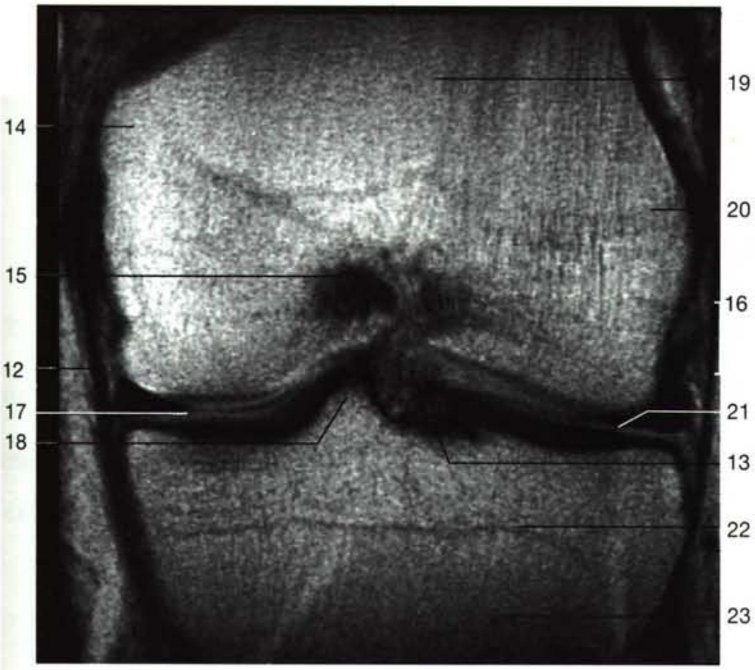

Right knee joint. Frontal section through the central part of the joint (posterior aspect, MR-Scan). (See also p. 10.)

1 Iliotibial tract
2 Articular muscle of knee
3 Patellar surface
4 Lateral condyle of femur
5 Articular capsule
6 Infrapatellar fat pad
7 **Patella** (articular surface)
8 **Suprapatellar bursa**
9 Quadriceps muscle of thigh (divided)
10 **Anterior cruciate ligament**
11 Medial condyle of femur
12 Fibular collateral ligament
13 **Posterior cruciate ligament**
14 Medial epicondyle of femur
15 Intercondylar fossa of femur
16 Tibial collateral ligament
17 **Medial meniscus** of knee joint
18 Medial intercondylar tubercle
19 Femur
20 Lateral epicondyle of femur
21 **Lateral meniscus** of knee joint
22 Epiphyseal line of tibia
23 Tibia

1 **Femur**
2 Quadriceps femoris muscle
3 Suprapatellar bursa and articular cavity
4 **Patella**
5 Patellar surface (articular cartilage)
6 Infrapatellar fat pad
7 Patellar ligament
8 **Tibia**
9 Tibial nerve
10 Adductor magnus muscle
11 Popliteal vein
12 Semitendinosus muscle
13 Semimembranosus muscle
14 **Popliteal artery**
15 Gastrocnemius muscle
16 Anterior cruciate ligament
17 Posterior cruciate ligament
18 Popliteus muscle
19 Soleus muscle
20 Deep flexor muscles of leg
21 Calcaneal tendon
22 Epiphyseal line of tibia
23 Calcaneus
24 **Talocrural joint**
25 Talus
26 Lateral meniscus
27 Epiphysial line

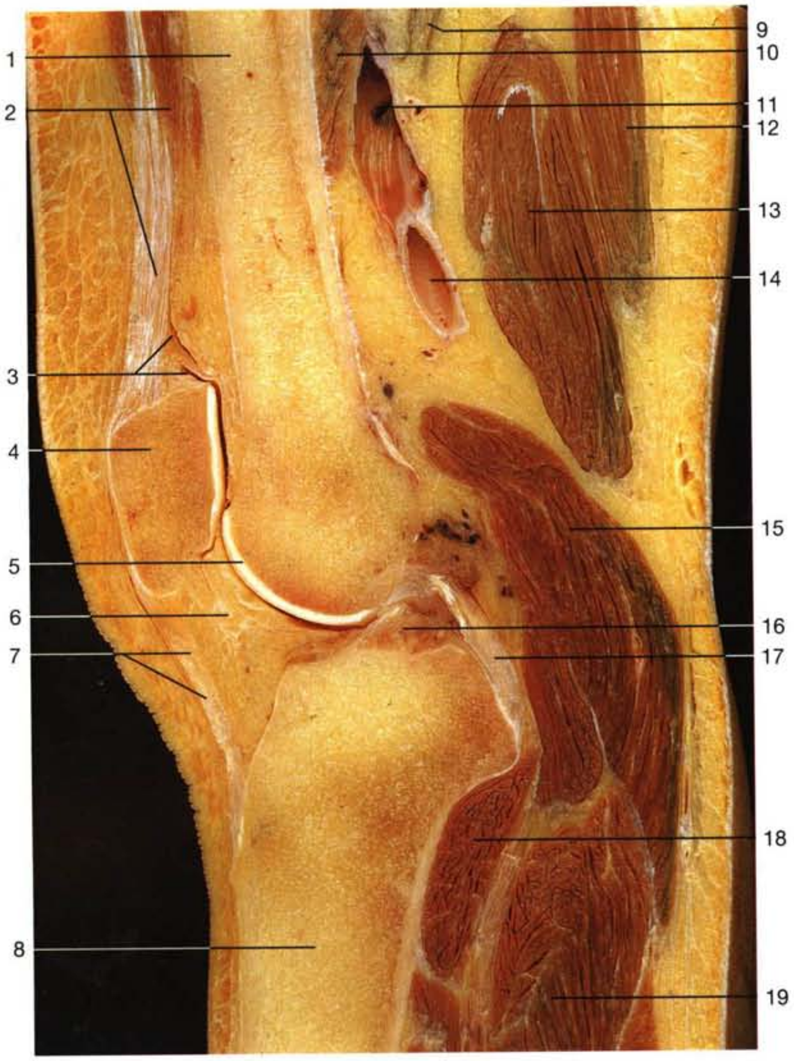

Sagittal section through the knee joint. Anterior surface to the left.

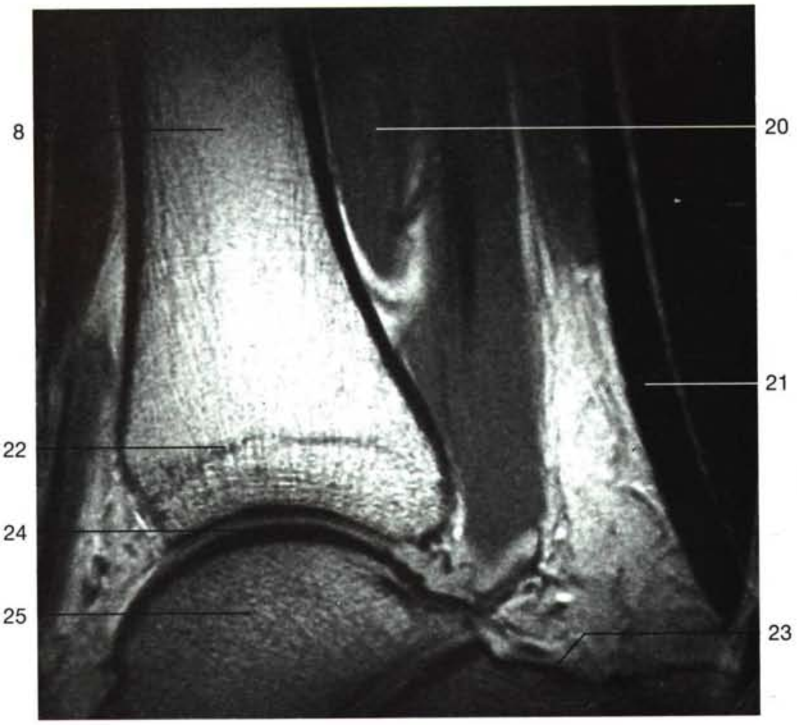

Talocrural joint. Sagittal section; anterior part to the left. (MR-Scan.)

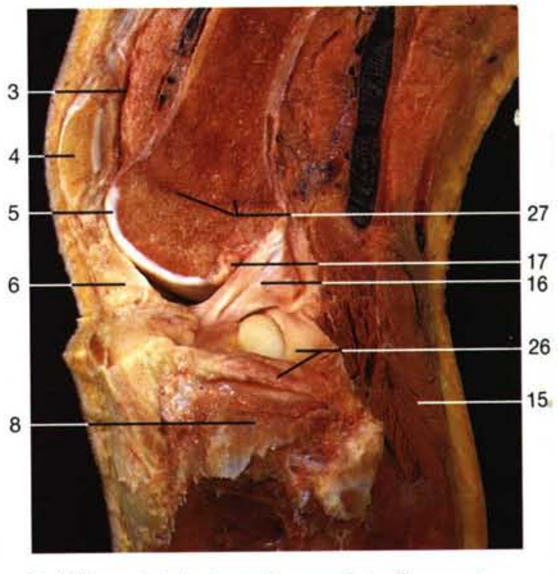

Left knee joint. Anterior cruciate ligament (lateral aspect).

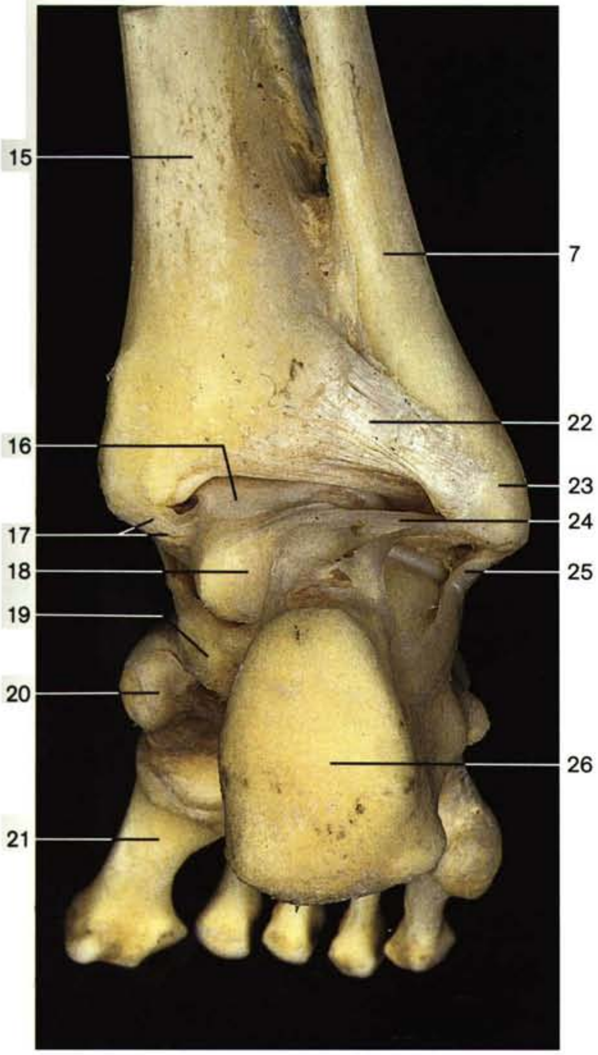

Ligaments of talocrural joint, right leg (posterior aspect).

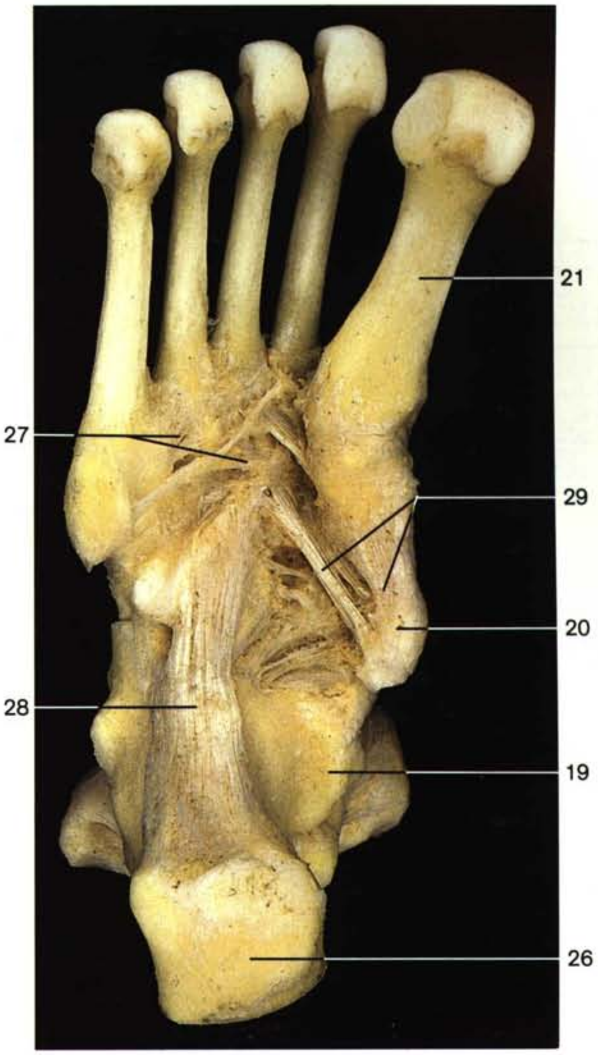

Deep ligaments of the foot, right foot (plantar aspect). The toes have been removed.

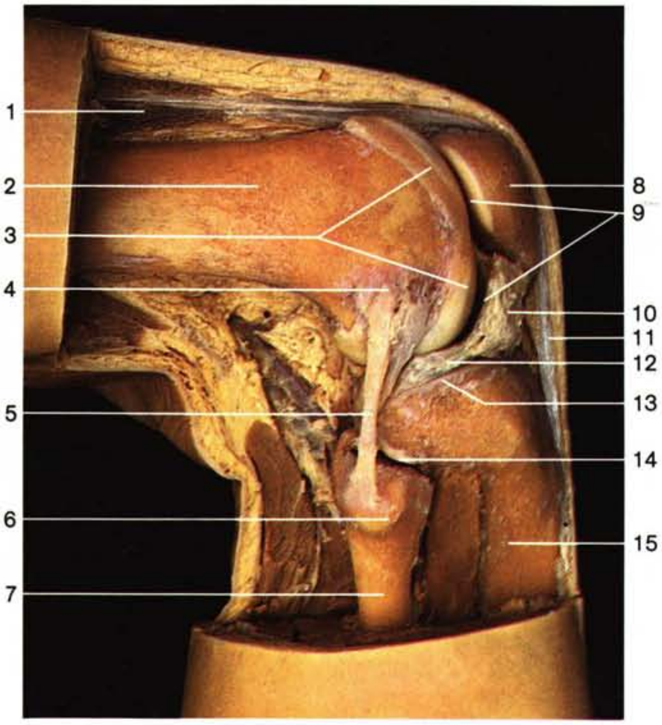

Right knee joint and tibiofibular joint with ligaments. Note the position of the lateral meniscus.

1 Quadriceps femoris muscle
2 Femur
3 Patellar surface
4 Lateral epicondyle of femur
5 **Fibular collateral ligament**
6 Head of fibula
7 **Fibula**
8 **Patella**
9 Articular cavity of knee joint
10 Infrapatellar fat pad
11 **Patellar ligament**
12 **Lateral meniscus of knee joint**
13 Lateral condyle of tibia (superior articular facet)
14 **Tibiofibular joint**
15 Tibia
16 Trochlea of talus (superior surface)
17 Deltoid ligament of ankle (posterior tibiotalar part)
18 Talus
19 **Sustentaculum tali**
20 Navicular bone
21 1st metatarsal bone
22 **Posterior tibiofibular ligament**
23 Lateral malleolus
24 Posterior talofibular ligament
25 Calcaneofibular ligament
26 Calcaneal tuberosity
27 Plantar tarsometatarsal ligaments
28 **Long plantar ligament**
29 Plantar cuneonavicular ligaments

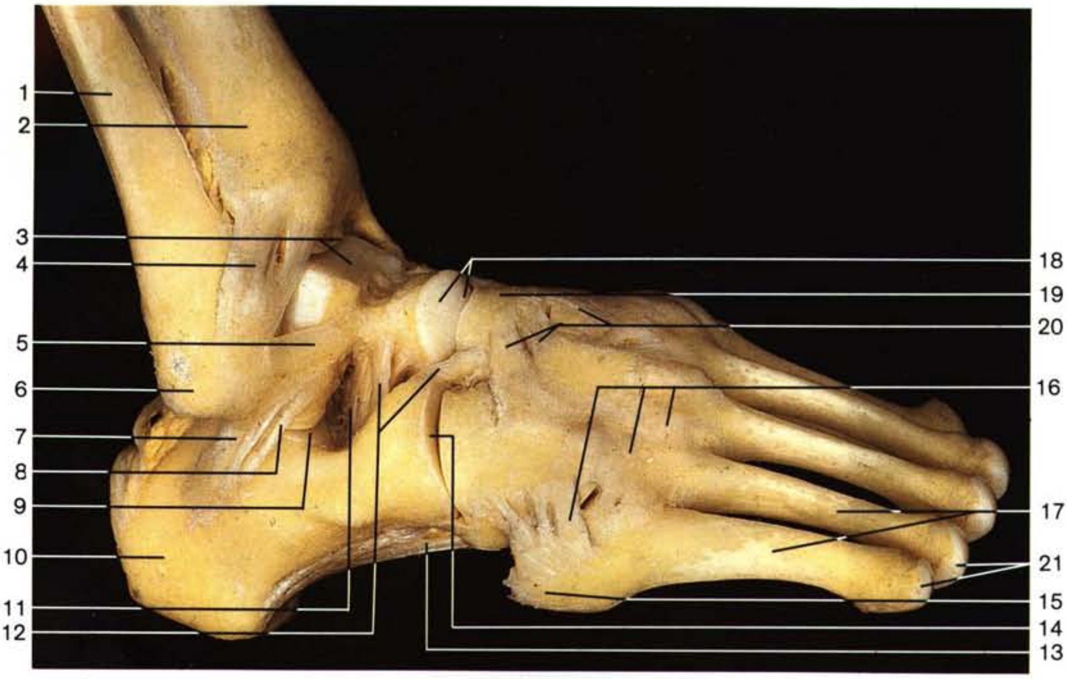

Ligaments of right foot (lateral aspect).

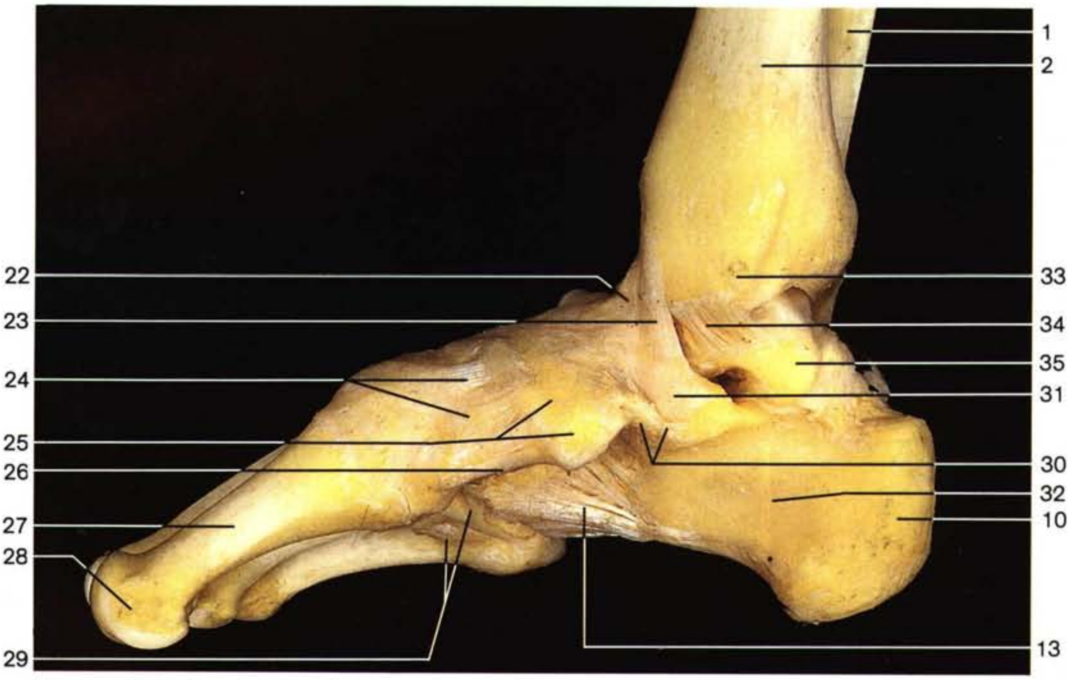

Ligaments of right foot (medial aspect).

1 Fibula
2 Tibia
3 Trochlea of talus and **talocrural joint**
4 Anterior tibiofibular ligament
5 Anterior talofibular ligament
6 Lateral malleolus
7 Calcaneofibular ligament
8 Lateral talocalcaneal ligament
9 **Subtalar joint**
10 Tuber calcanei
11 Interosseous talocalcaneal ligament
12 Bifurcate ligament
13 Long plantar ligament
14 **Calcaneocuboid joint**
15 Tuberosity of 5th metatarsal bone
16 Dorsal tarsometatarsal ligaments
17 Metatarsal bones
18 Head of talus and **talocalcaneonavicular joint**
19 Navicular bone
20 Dorsal cuneonavicular ligaments
21 Heads of metatarsal bones
22 Medial or deltoid ligament of ankle (tibionavicular part)
23 Medial or deltoid ligament of ankle (tibiocalcaneal part)
24 Dorsal cuneonavicular ligaments
25 Navicular bone
26 Plantar cuneonavicular ligament
27 1st metatarsal bone
28 Head of 1st metatarsal bone
29 Plantar tarsometatarsal ligaments
30 Plantar calcaneonavicular ligament
31 Sustentaculum tali
32 Calcaneus
33 Medial malleolus
34 Medial or deltoid ligament of ankle (posterior part)
35 Talus

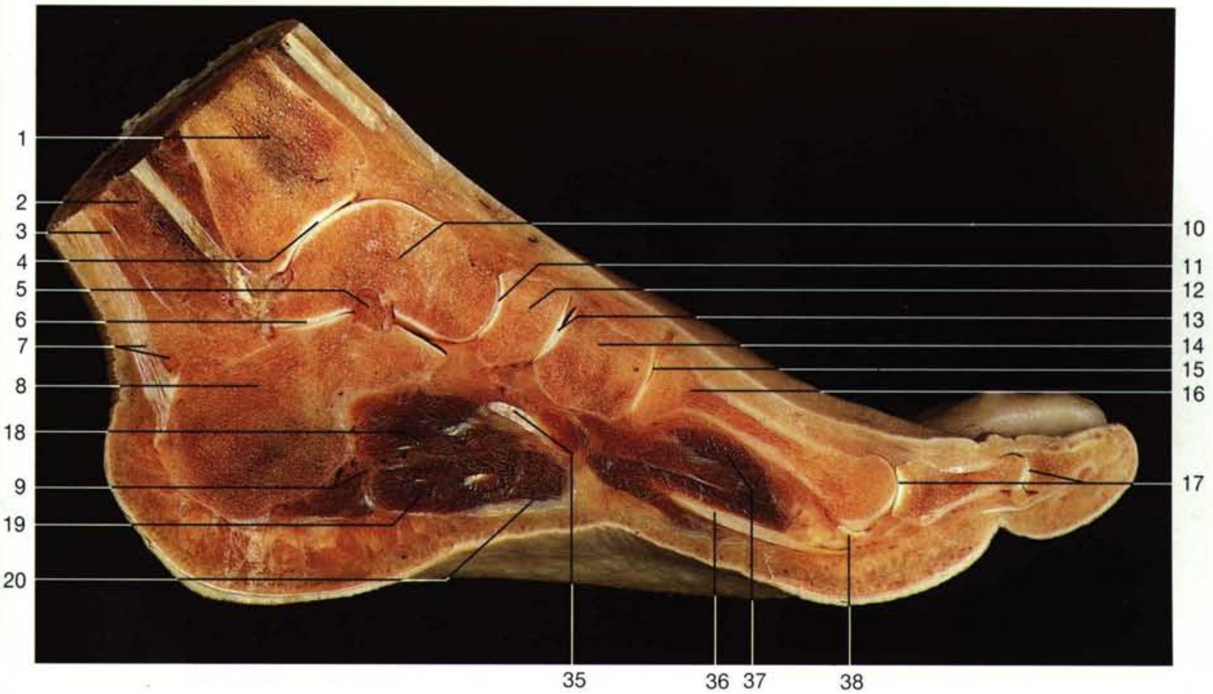

Longitudinal section through the foot at the level of first phalanx.

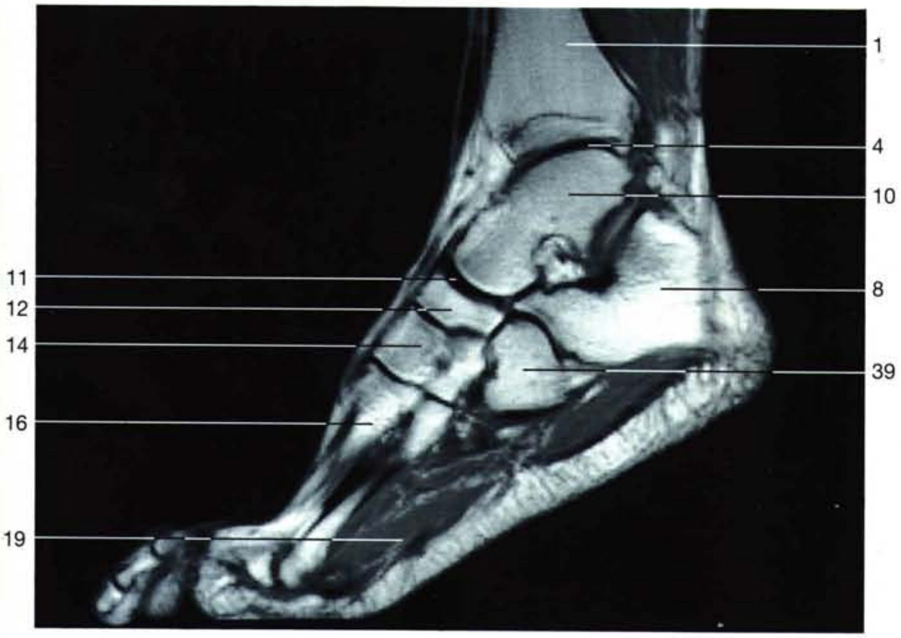

Sagittal section through the foot
(MR-Scan after A. Heuck, G. Luttke and J. W. Rohen, 1994).

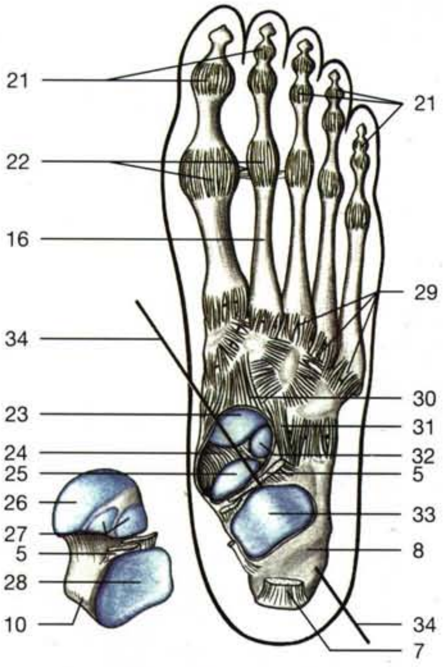

Talocalcaneonavicular joint.
The talus has been rotated to show the articular surfaces of the joint.

1 Tibia
2 Deep flexor muscles
3 Superficial flexor muscles
4 **Talocrural joint**
5 Interosseous talocalcaneal ligament
6 **Subtalar joint**
7 Calcaneal or Achilles tendon and bursa
8 Calcaneus
9 Vessels and nerves of foot
10 Talus
11 **Talocalcaneonavicular joint**
12 Navicular bone
13 **Cuneonavicular joint**
14 Intermediate cuneiform bone
15 **Tarsometatarsal joints**
16 Metatasal bones
17 **Metatarsophalangeal** and **interphalangeal joints**
18 Quadratus plantae muscle with flexor tendons
19 Flexor digitorum brevis muslce
20 Plantar aponeurosis
21 Articular capsules of interphalangeal joints
22 Articular capsules of metatarsophalangeal joints
23 Articular surface of navicular bone
24 Plantar calcaneonavicular ligament
25 Middle talar articular surface of calcaneus
26 Navicular articular surface of talus
27 Anterior and middle calcaneal surfaces of talus
28 Posterior calcaneal surface of talus
29 Dorsal tarsometatarsal ligaments
30 Talonavicular ligament
31 Bifurcate ligament
32 Anterior talar articular surface of calcaneus
33 Posterior talar articular surface of calcaneus
34 Axis for inversion and eversion
35 Tendon of tibialis posterior muscle
36 Tendon of flexor hallucis longus muscle
37 Flexor hallucis brevis muscle
38 Sesamoid bone
39 Cuboid bone

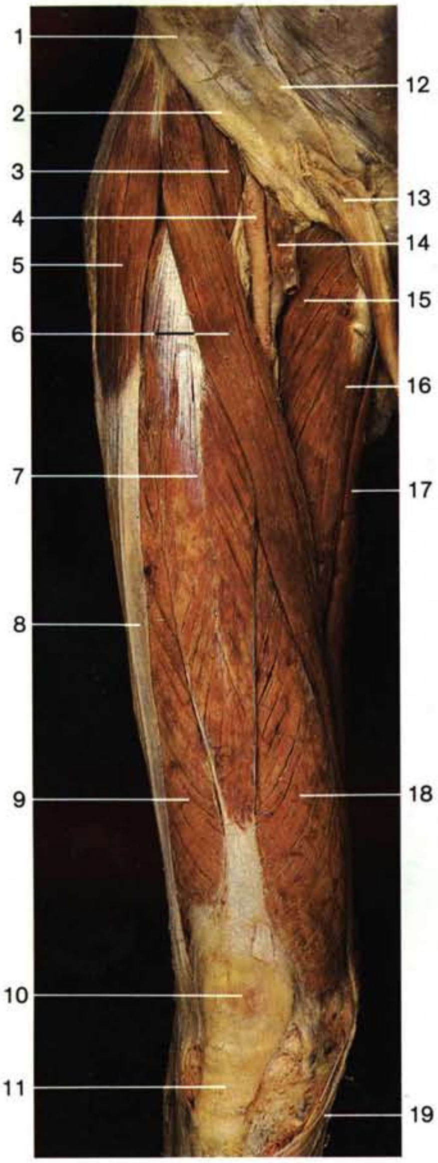

Extensor and adductor muscles of thigh, right thigh (anterior aspect).

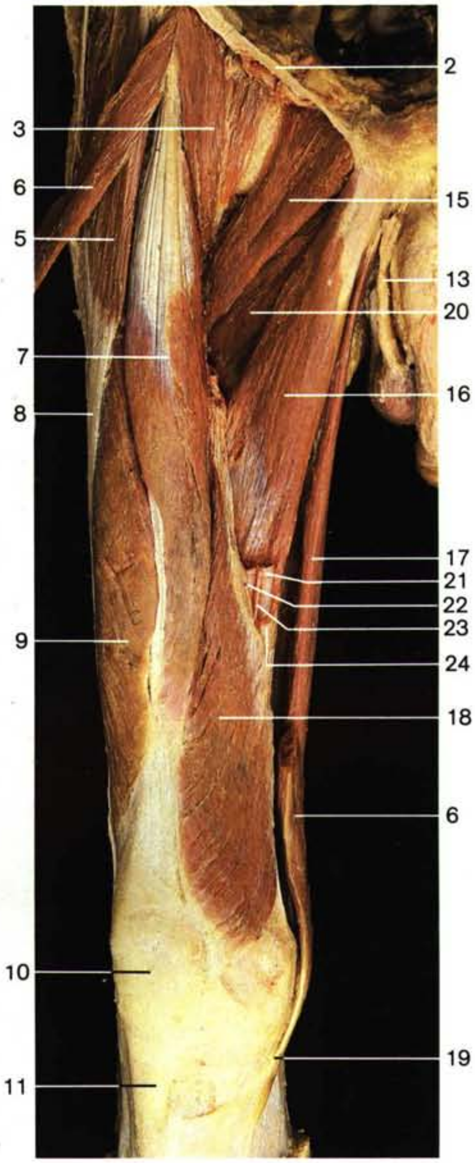

Quadriceps muscle and superficial layer of adductor muscles, right thigh (anterior aspect). The sartorius muscle has been divided.

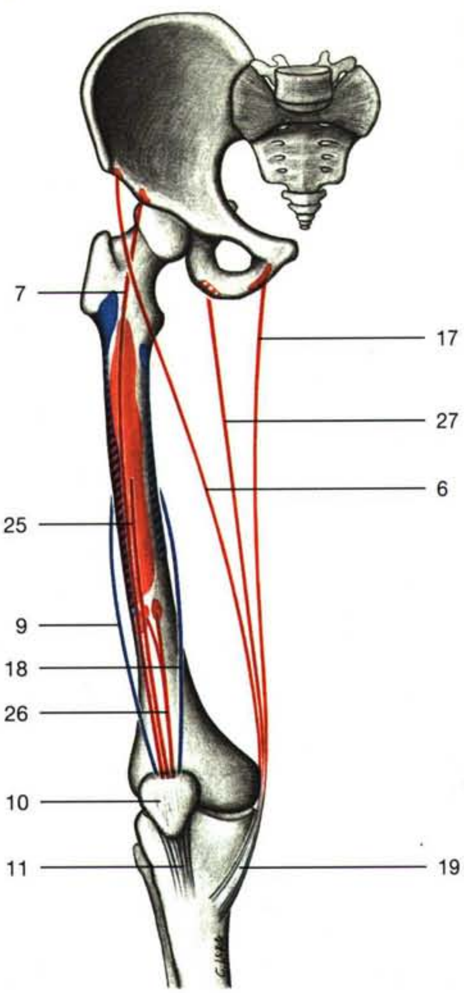

Course of extensor muscles of thigh and muscles inserting with common tendon on tibia.

1 Anterior superior iliac spine
2 Inguinal ligament
3 Iliopsoas muscle
4 Femoral artery
5 Tensor fasciae latae muscle
6 Sartorius muscle
7 **Rectus femoris muscle**
8 Iliotibial tract
9 Vastus lateralis muscle
10 Patella
11 Patellar ligament
12 Aponeurosis of external abdominal oblique muscle
13 Spermatic cord
14 Femoral vein
15 **Pectineus muscle**
16 **Adductor longus muscle**
17 Gracilis muscle
18 Vastus medialis muscle
19 Common tendon of sartorius, gracilis, and semitendinosus muscles (pes anserinus)
20 **Adductor brevis muscle**
21 Femoral artery } entering the adductor canal
22 Femoral vein
23 Saphenous nerve
24 Fascia of adductor canal
25 Vastus intermedius muscle
26 Articularis genus muscle
27 Semitendinosus muscle

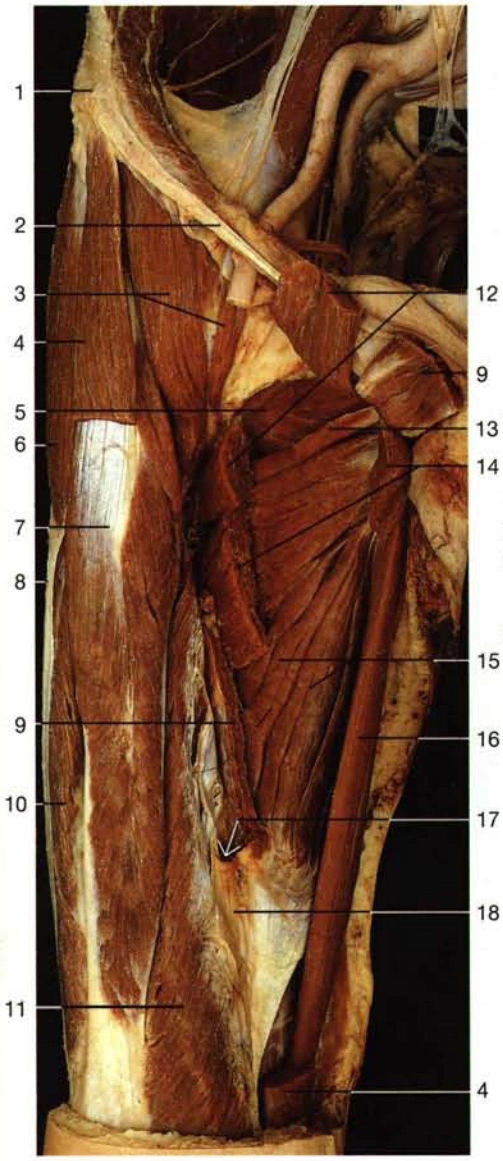

Deep layer of adductor muscles. Adductor magnus muscle (anterior aspect). Pectineus, adductor longus and brevis muscles have been divided.

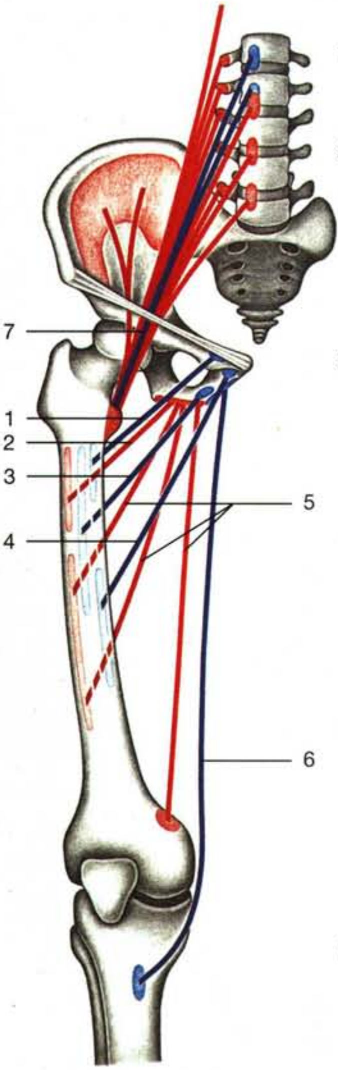

Course of adductor muscles. (Schematic drawing.)

1 Pectineus muscle (blue)
2 Adductor minimus muscle (red)
3 Adductor brevis muscle (blue)
4 Adductor longus muscle (blue)
5 Adductor magnus muscle (red)
6 Gracilis muscle (blue)
7 Iliopsoas muscle (red/blue)

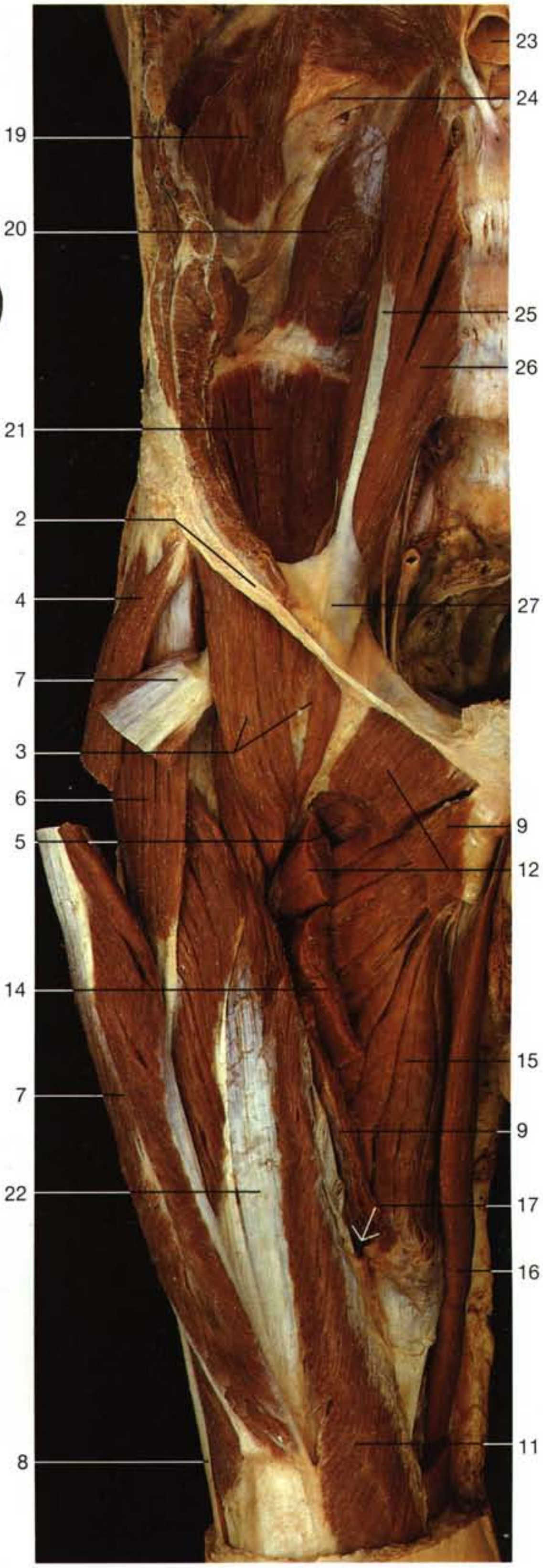

Iliopsoas and adductor muscles; deepest layer (anterior aspect). Pectineus, adductor longus and brevis, rectus femoris muscles have been divided.

1 Anterior superior iliac spine
2 Inguinal ligament
3 **Iliopsoas muscle**
4 Sartorius muscle
5 Obturator externus muscle
6 Tensor fasciae latae muscle
7 Rectus femoris muscle
8 Iliotibial tract
9 Adductor longus muscle (divided)
10 Vastus lateralis muscle
11 Vastus medialis muscle
12 Pectineus muscle (divided)
13 Adductor minimus muscle
14 Adductor brevis muscle (cut)
15 **Adductor magnus**
16 Gracilis muscle
17 Adductor hiatus
18 Vastoadductor membrane
19 Diaphragm
20 Quadratus lumborum muscle
21 **Iliacus muscle**
22 Vastus intermedius muscle
23 Aorta in aortic hiatus
24 Twelfth rib
25 **Psoas minor muscle**
26 **Psoas major muscle**
27 Iliopectineal arch

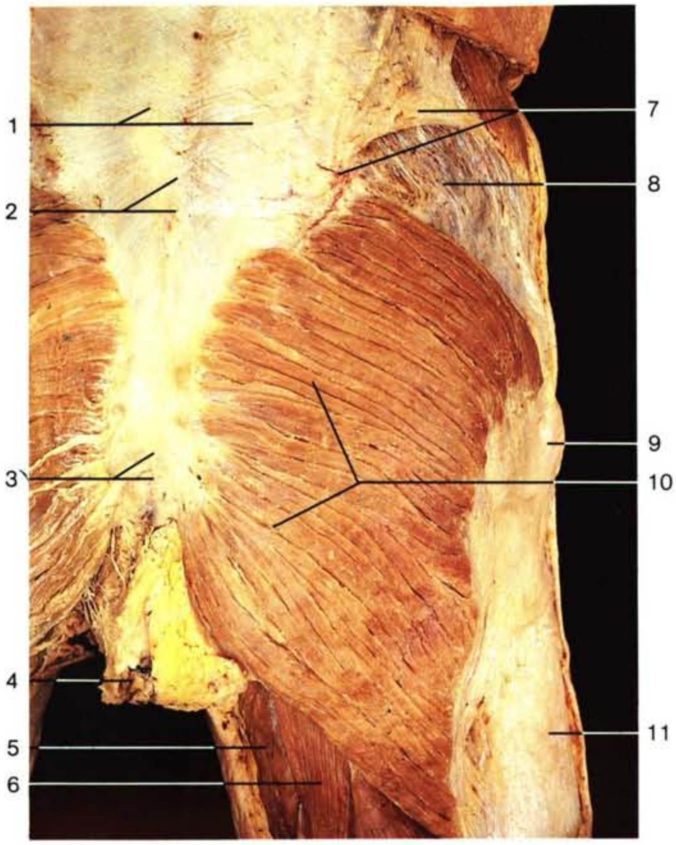
1
2
3
4
5
6
7
8
9
10
11

Gluteal muscles, superficial layer (posterior aspect).

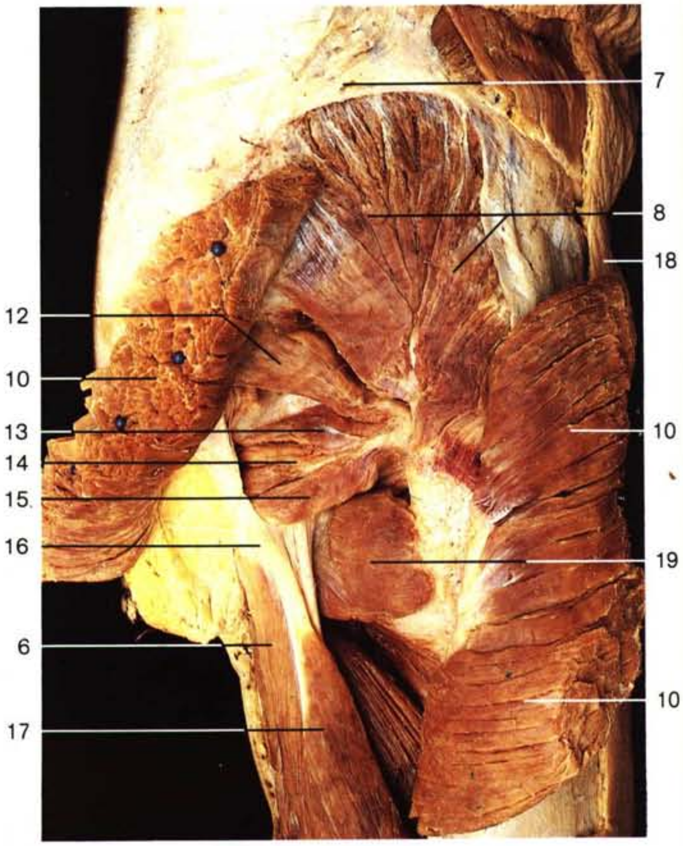
7
8
18
12
10
13
14
15
16
10
19
6
10
17

Gluteal muscles, deeper layer (posterior aspect).

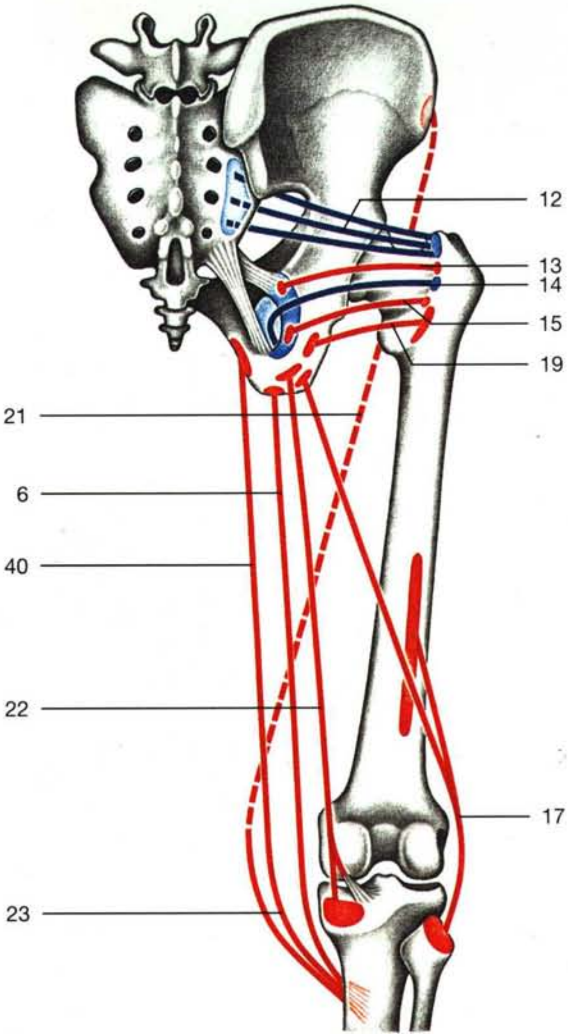
12
13
14
15
19
21
6
40
22
17
23

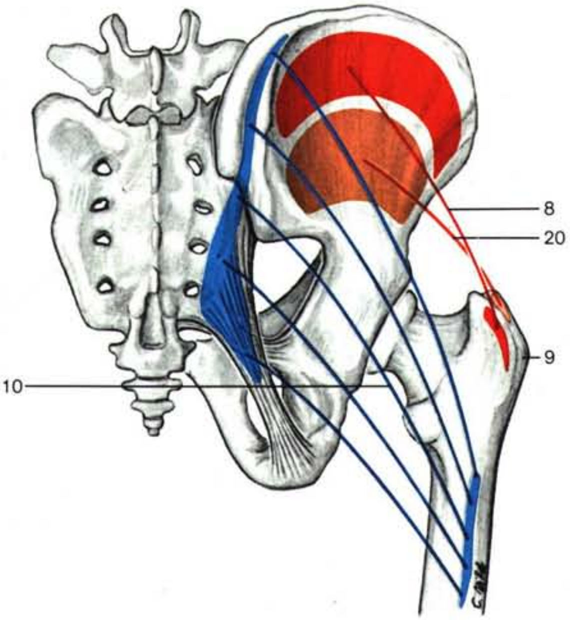

Course of gluteal muscles
(posterior aspect; schematic drawing).

◁

Course of gluteal muscles (deeper layer) **and of ischiocrural muscles** (posterior aspect).
Sartorius muscle is indicated by dotted line (schematic drawing).

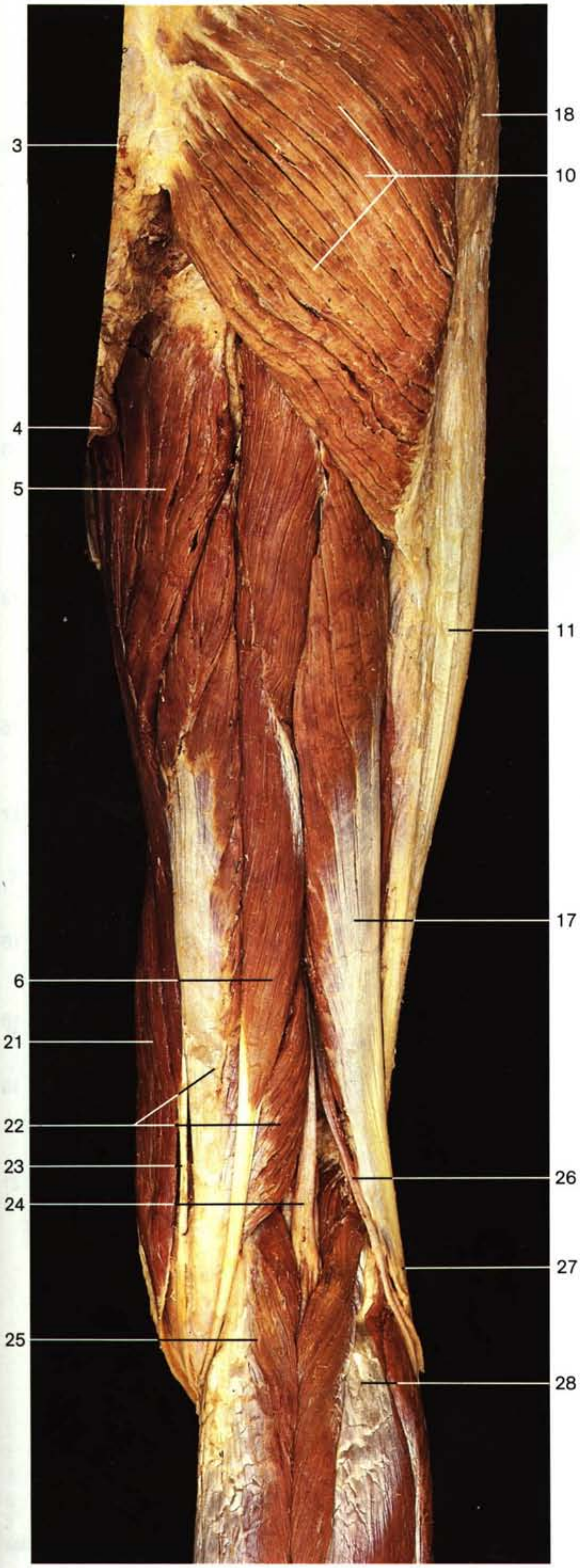

Flexors of the right thigh, superficial layer (dorsal aspect).

1	Thoracolumbar fascia
2	Spinous processes of lumbar vertebrae
3	Coccyx
4	Anus
5	Adductor magnus muscle
6	Semitendinosus muscle
7	Iliac crest
8	**Gluteus medius muscle**
9	Greater trochanter
10	**Gluteus maximus muscle**
11	Iliotibial tract
12	Piriformis muscle
13	Superior gemellus muscle
14	Obturator internus muscle
15	Inferior gemellus muscle
16	Ischial tuberosity
17	**Biceps femoris muscle**
18	Tensor fasciae latae muscle
19	Quadratus femoris muscle
20	**Gluteus minimus muscle**
21	Sartorius muscle
22	Semimembranosus muscle
23	Tendon of gracilis muscle
24	Tibial nerve
25	Medial head of gastrocnemius muscle
26	Common peroneal nerve
27	Tendon of biceps femoris muscle
28	Lateral head of gastrocnemius muscle
29	Rectus femoris muscle
30	Vastus medialis muscle
31	Vastus intermedius muscle
32	Vastus lateralis muscle
33	Sciatic nerve
34	Gluteus maximus muscle (insertion)
35	Great saphenous vein
36	Femoral artery
37	Femoral vein
38	Adductor longus muscle
39	Femur
40	Gracilis muscle
41	Septum between semitendinosus and semimembranosus muscles

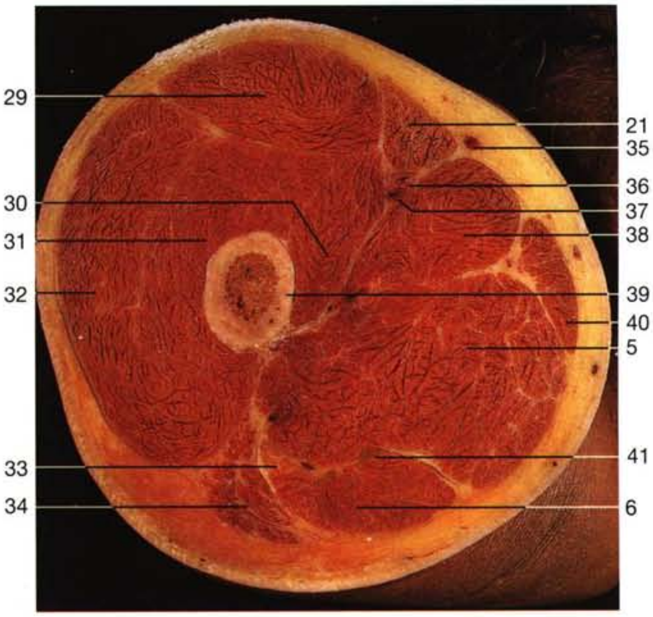

Cross section of right thigh (inferior aspect). Anterior side on top.

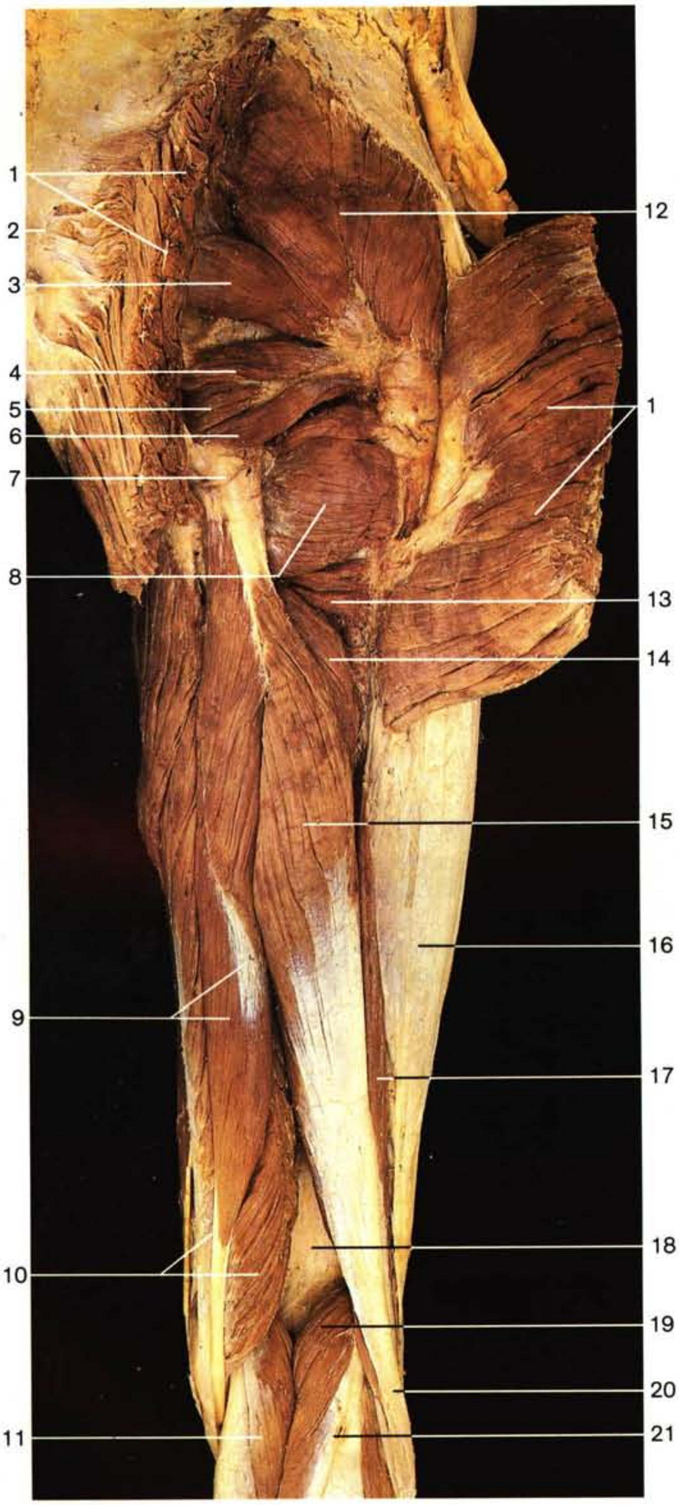

Dorsal muscles of right thigh (posterior aspect).
The gluteus maximus muscle has been cut and reflected.

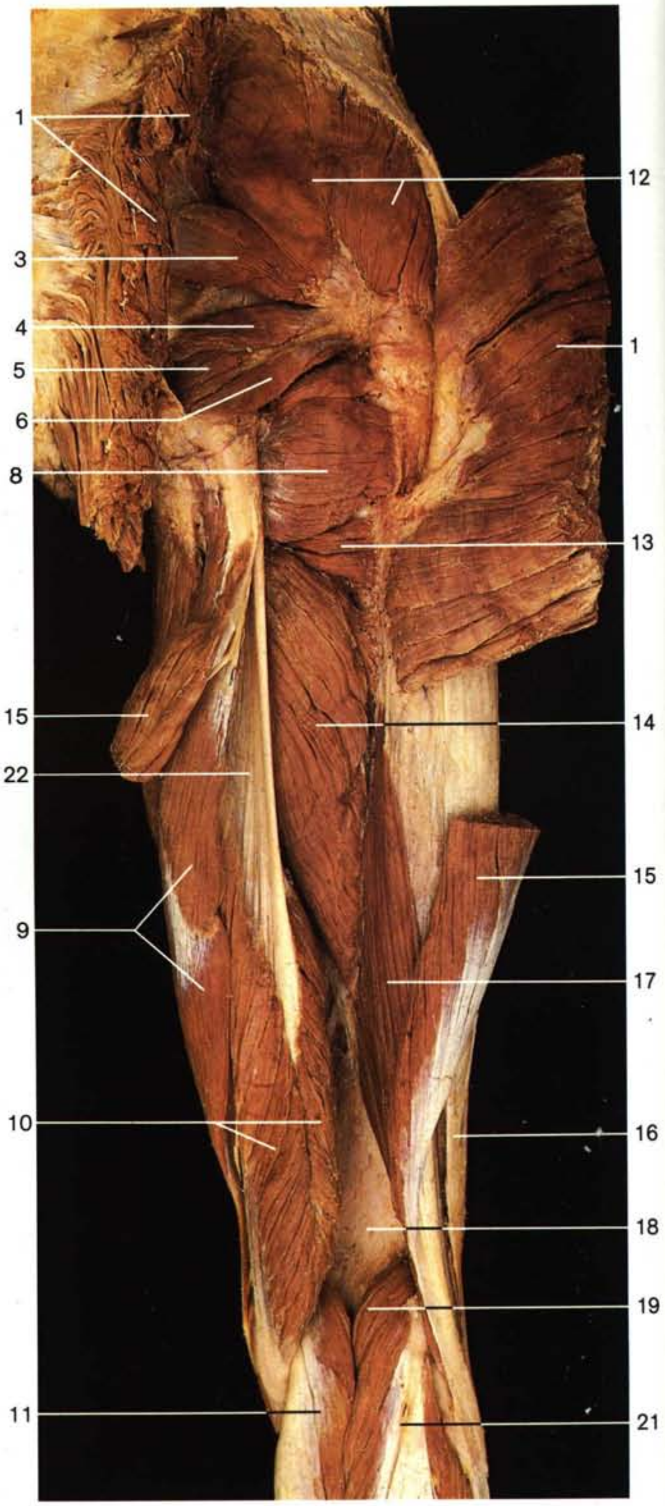

Dorsal muscles of right thigh (posterior aspect).
The gluteus maximus muscle and the long head of biceps femoris muscle have been divided and displaced.

1 Gluteus maximus muscle (divided)
2 Position of coccyx
3 Piriformis muscle
4 Superior gemellus muscle
5 Obturator internus muscle
6 Inferior gemellus muscle
7 Ischial tuberosity
8 Quadratus femoris muscle
9 Semitendinosus muscle with intermediate tendon
10 Semimembranosus muscle
11 Medial head of gastrocnemius muscle
12 Gluteus medius muscle
13 Adductor minimus muscle
14 Adductor magnus muscle
15 Long head of biceps femoris muscle
16 Iliotibial tract
17 Short head of biceps femoris muscle
18 Popliteal surface of femur
19 Plantaris muscle
20 Tendon of biceps femoris muscle
21 Lateral head of gastrocnemius muscle
22 Membranous part of semimembranosus muscle

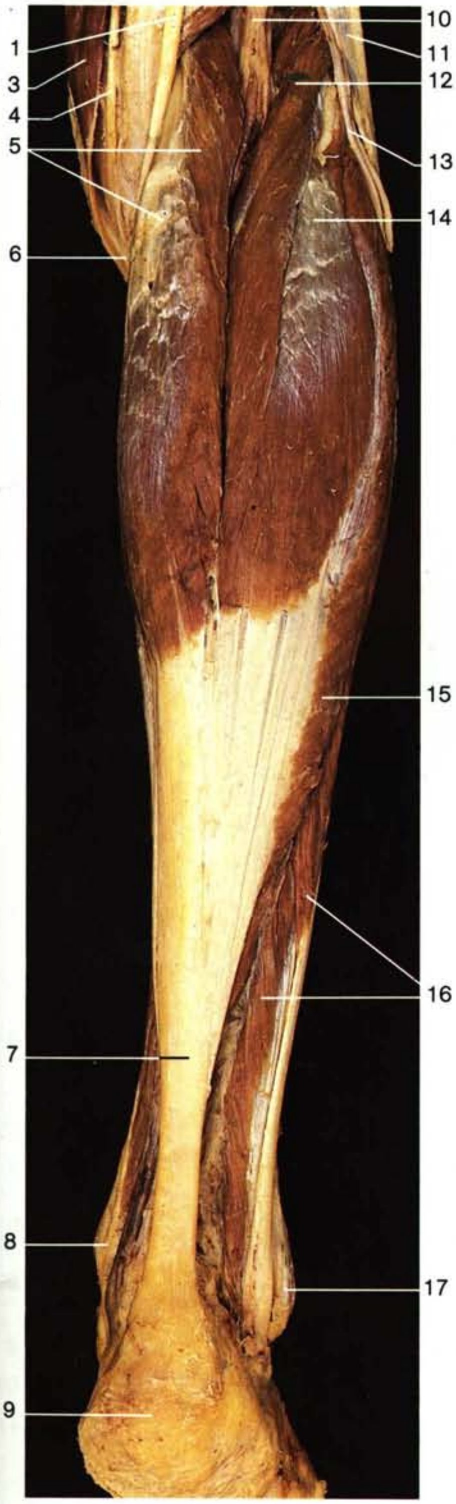

Flexor muscles of right leg
(posterior aspect).

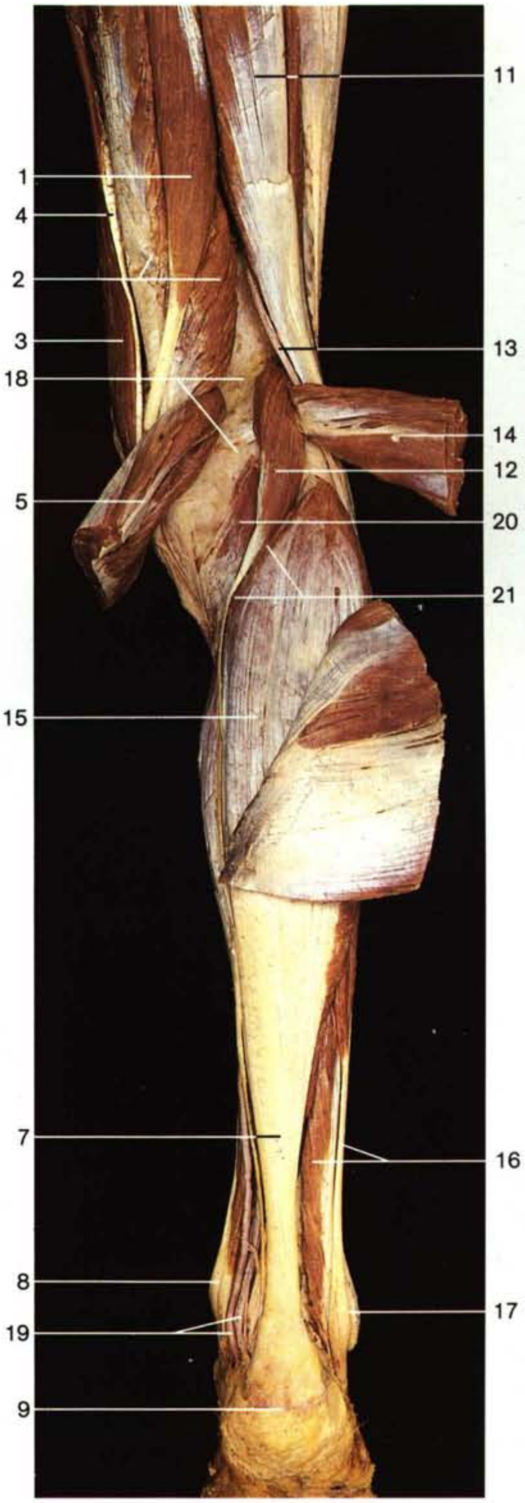

Flexor muscles of right leg (posterior aspect). Both heads of the gastrocnemius muscle have been cut and reflected.

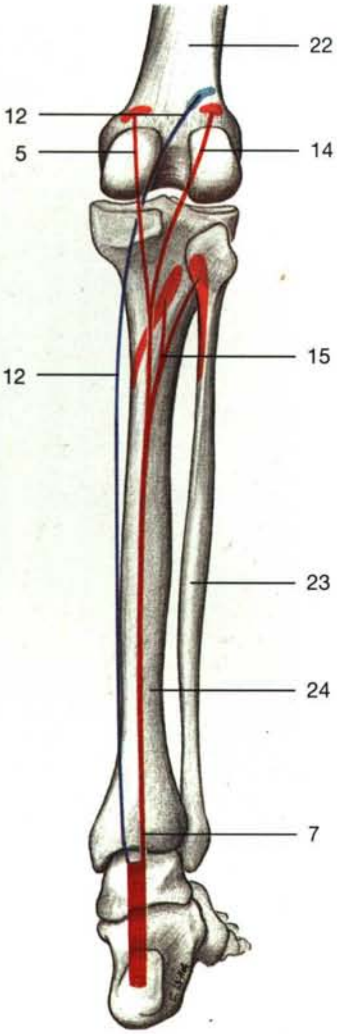

Flexor muscles of the leg
(right side).

1	Semitendinosus muscle
2	Semimembranosus muscle
3	Sartorius muscle
4	Tendon of gracilis muscle
5	**Medial head of gastrocnemius muscle**
6	Common tendon of gracilis, sartorius, and semitendinosus muscles
7	Calcaneal or Achilles tendon
8	Medial malleolus
9	Calcaneal tuberosity
10	Tibial nerve
11	Biceps femoris muscle
12	Plantaris muscle
13	Common peroneal nerve
14	**Lateral head of gastrocnemius muscle**
15	**Soleus muscle**
16	Peroneus longus and brevis muscles
17	Lateral malleolus
18	Popliteal fossa
19	Tibial nerve and posterior tibial artery
20	Popliteus muscle
21	Tendinous arch of soleus muscle
22	Femur
23	Fibula
24	Tibia

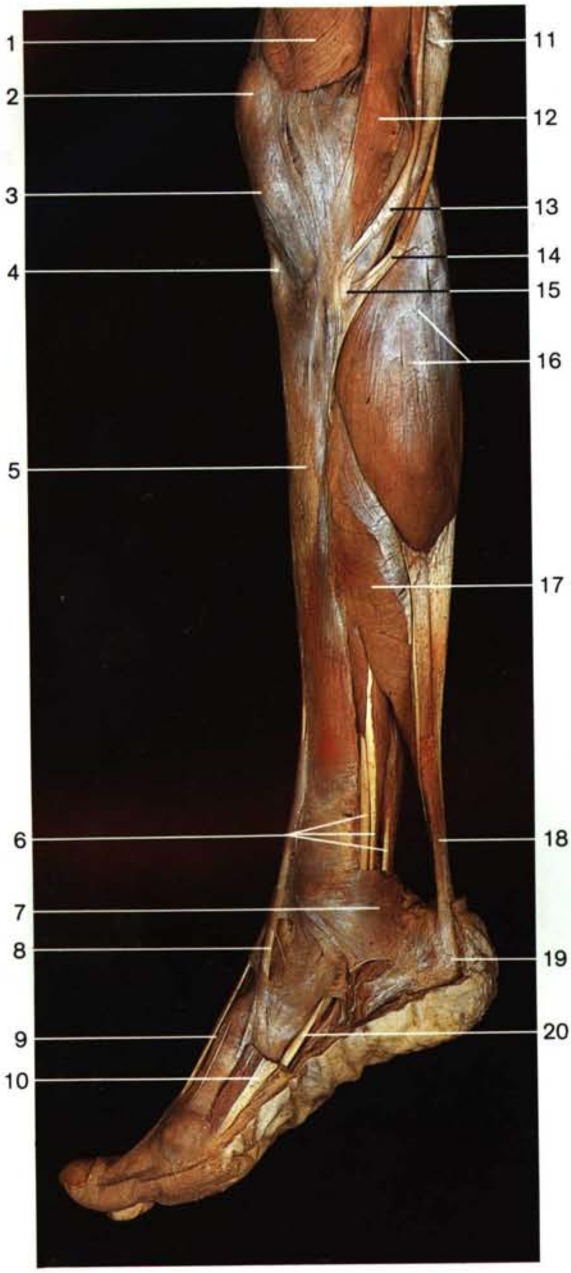

Muscles of right leg and foot (medial aspect).

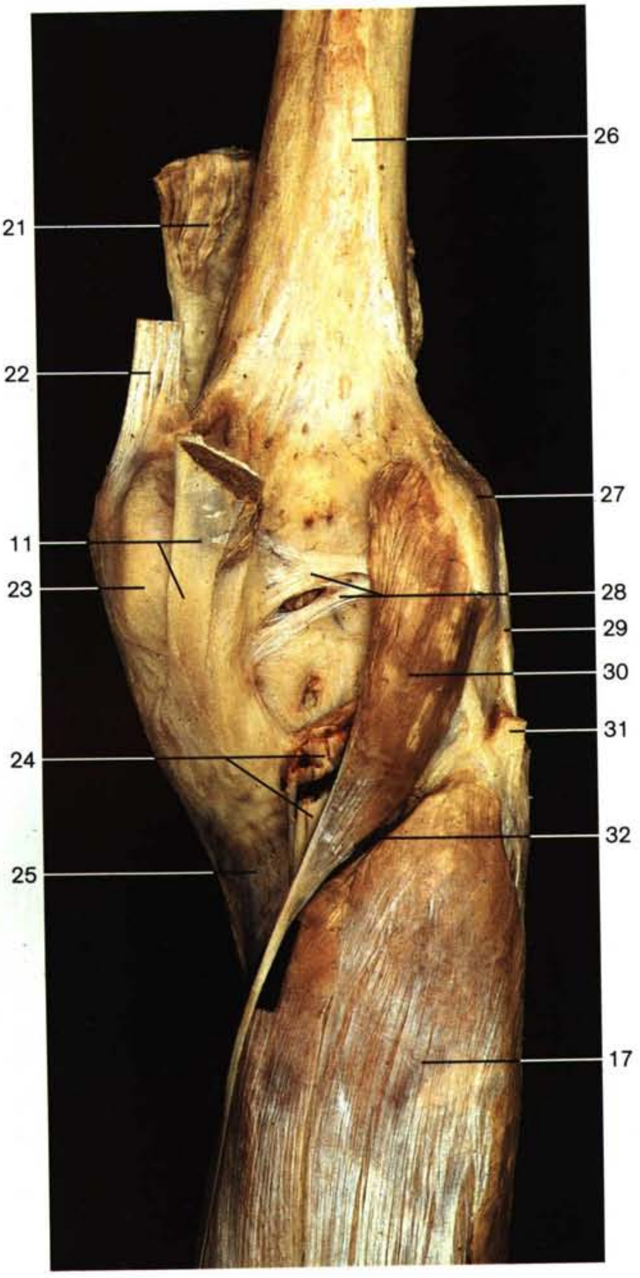

Popliteal region with plantaris and soleus, right side (dorsal aspect). Notice the insertion of the tendon of semimembranosus.

1 Vastus medialis muscle
2 Patella
3 Patellar ligament
4 Tibial tuberosity
5 Tibia
6 Tendons of deep flexor muscles (from anterior to posterior: 1. tibialis posterior; 2. flexor digitorum longus; 3. flexor hallucis longus muscles)
7 Flexor retinaculum
8 Tendon of tibialis anterior muscle
9 Tendon of extensor hallucis longus muscle
10 Abductor hallucis muscle
11 Semimembranosus muscle
12 Sartorius muscle
13 Tendon gracilis muscle
14 Tendon of semitendinosus muscle
15 Common tendon of gracilis, semitendinosus and sartorius muscles
16 **Medial head of gastrocnemius muscle**
17 **Soleus muscle**
18 Calcaneal or Achilles tendon
19 Calcaneus muscle
20 Tendon of flexor hallucis longus muscle
21 Quadriceps femoris muscle (divided)
22 Tendon of adductor magnus muscle (divided)
23 Medial condyle of femur
24 Popliteal artery and vein, tibial nerve
25 Tibia
26 Femur
27 Lateral epicondyle of femur
28 Oblique popliteal ligament
29 Lateral (fibular) collateral ligament
30 **Plantaris muscle**
31 Tendon of biceps femoris muscle (divided)
32 Tendinous arch of soleus muscle

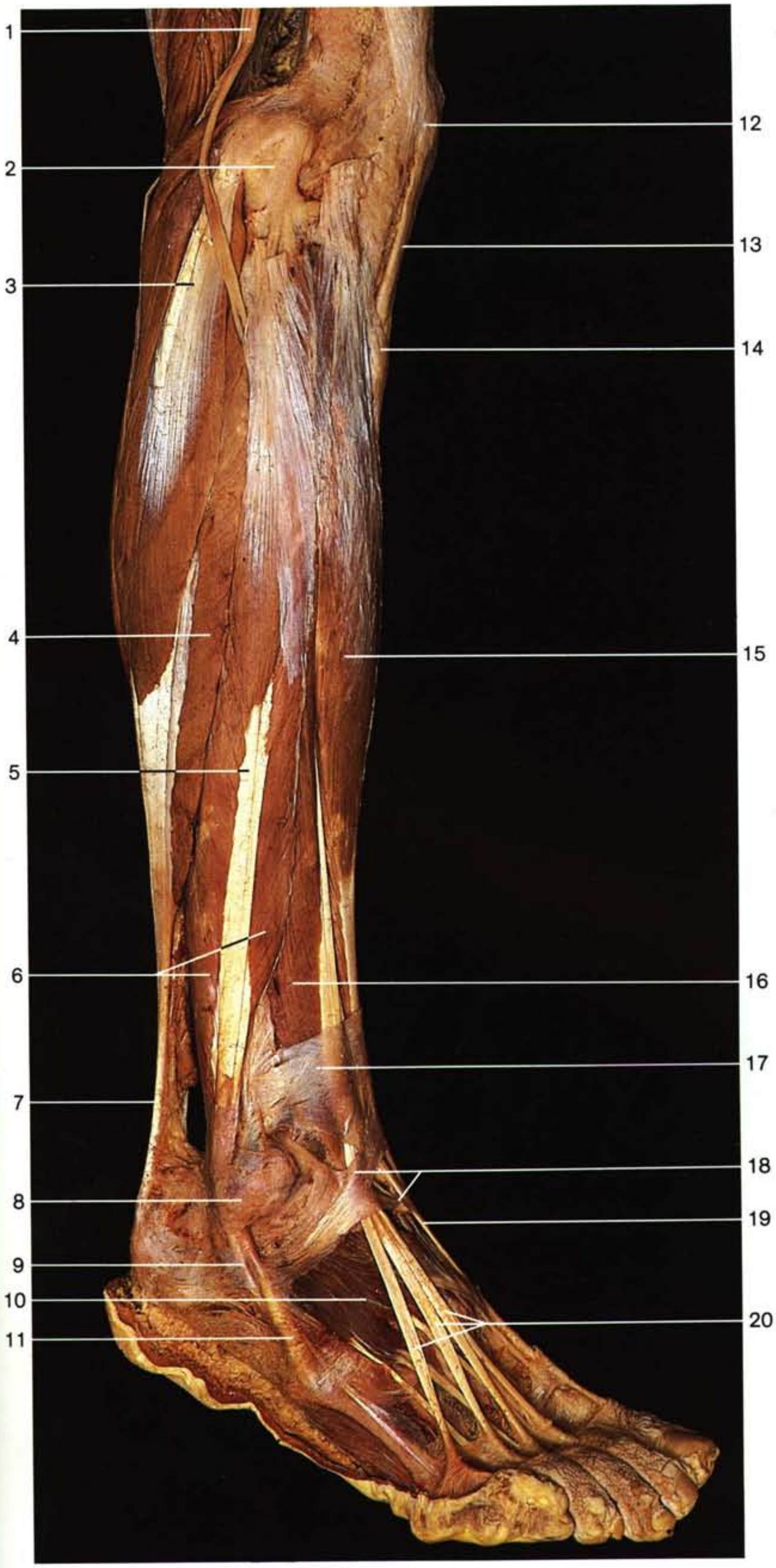

Muscles of right leg and foot (lateral aspect).

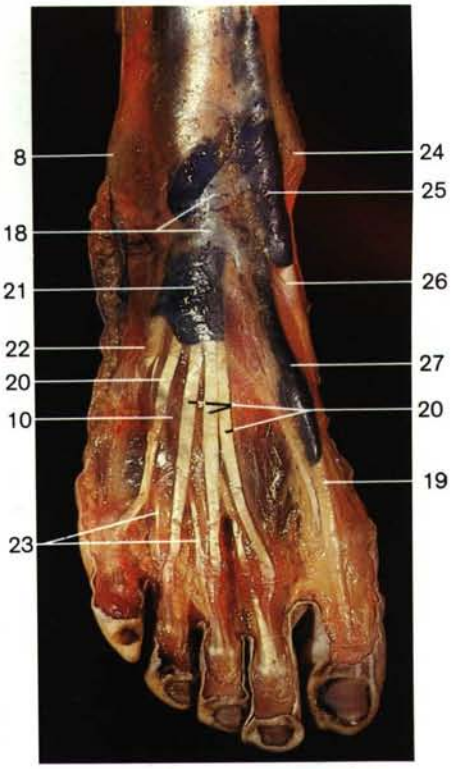

Right foot with synovial sheaths of extensor muscles (dorsal aspect). The synovial sheaths have been injected with blue solution.

1 Common peroneal nerve
2 Head of fibula
3 Lateral head of gastrocnemius muscle
4 Soleus muscle
5 Peroneus longus muscle
6 Peroneus brevis muscle
7 Calcaneal or Achilles tendon
8 Lateral malleolus muscle
9 Tendon of peroneus longus muscle
10 Extensor digitorum brevis muscle
11 Tendon of peroneus brevis muscle
12 Patella
13 Patellar ligament
14 Tuberosity of tibia
15 Tibialis anterior muscle
16 Extensor digitorum longus muscle
17 Superior extensor retinaculum
18 Inferior extensor retinaculum
19 Tendon of extensor hallucis longus muscle
20 Tendons of extensor digitorum longus muscle
21 Common synovial sheath of extensor digitorum longus muscle
22 Tendon of peroneus tertius muscle to the lateral margin of foot
23 Tendons of extensor digitorum brevis muscle
24 Medial malleolus
25 Synovial sheath of tendon of tibialis anterior muscle
26 Tendon of tibialis anterior muscle
27 Synovial sheath of tendon of extensor hallucis longus muscle

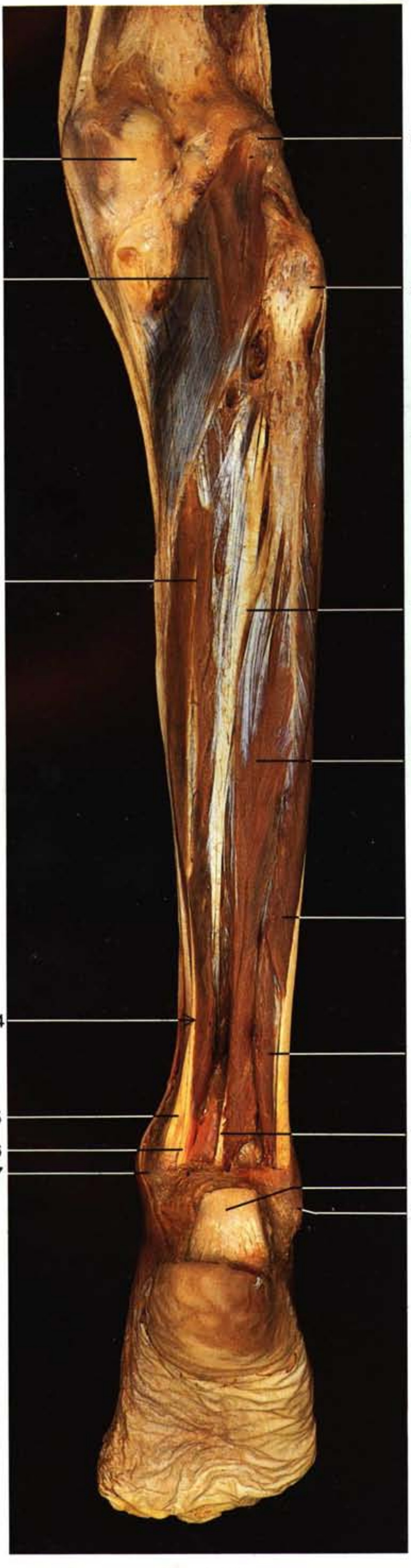

1 Medial condyle of femur
2 Popliteus muscle
3 Flexor digitorum longus muscle
4 Crosssing of tendons in leg
5 Tendon of tibialis posterior muscle
6 Tendon of flexor digitorum longus muscle
7 Medial malleolus
8 Lateral condyle of femur
9 Head of fibula
10 Tibialis posterior muscle
11 Flexor hallucis longus muscle
12 Peroneus longus muscle
13 Peroneus brevis muscle
14 Tendon of flexor hallucis longus muscle
15 Calcaneal tendon (divided)
16 Lateral malleolus

Deep flexor muscles of right leg (posterior aspect).

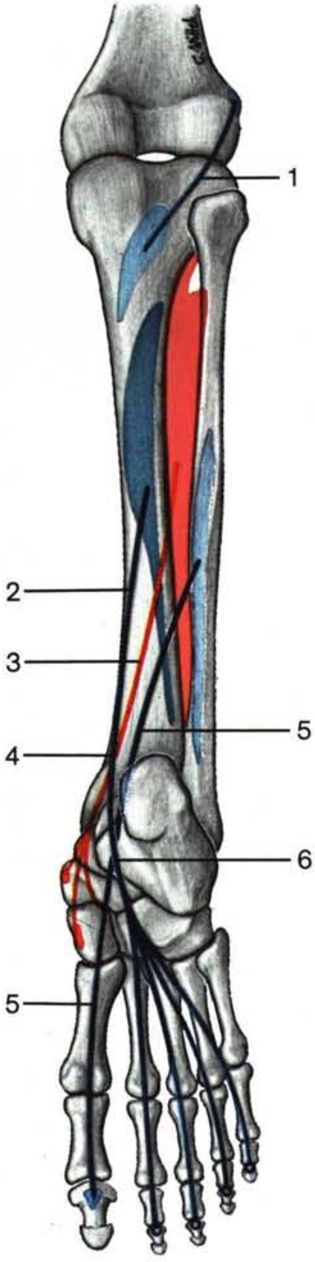

1 Popliteus muscle (blue)
2 Flexor digitorum longus muscle (blue)
3 Tibialis posterior muscle (red)
4 Crossing of tendons in leg
5 Flexor hallucis longus muscle (blue)
6 Crossing of tendons in sole

Course of deep flexor muscle of leg. (Schematic drawing.)

1 Medial condyle of femur
2 Tibia
3 Flexor digitorum longus muscle
4 Crossing of tendons in leg
5 Tendon of tibialis posterior muscle
6 Abductor hallucis muscle
7 Tendon of flexor hallucis longus muscle
8 Lateral condyle of femur
9 Head of fibula
10 Tibialis posterior muscle
11 Tendon of flexor digitorum longus muscle
12 Flexor retinaculum
13 Calcaneal tendon
14 Calcaneal tuberosity
15 Crossing of tendons in sole
16 Quadratus plantae muscle
17 Tendons of flexor digitorum longus muscle
18 Tendon of tibialis anterior muscle
19 Area of insertion of tibialis posterior muscle
20 Lumbrical muscles
21 Flexor hallucis longus muscle

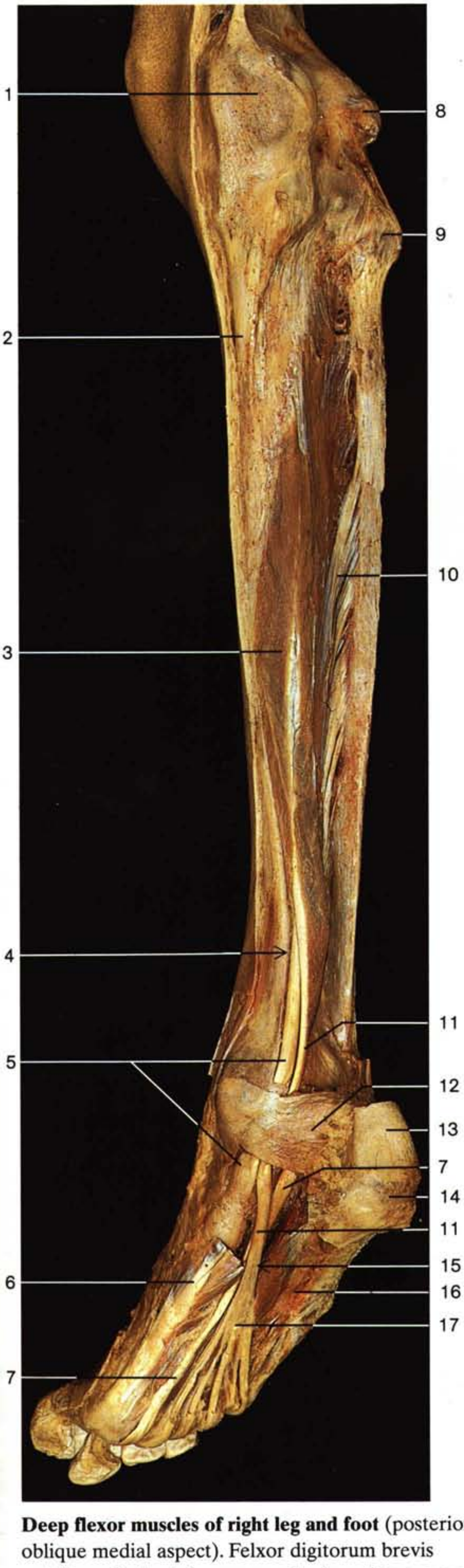

Deep flexor muscles of right leg and foot (posterior oblique medial aspect). Felxor digitorum brevis and flexor hallucis longus muscles have been removed.

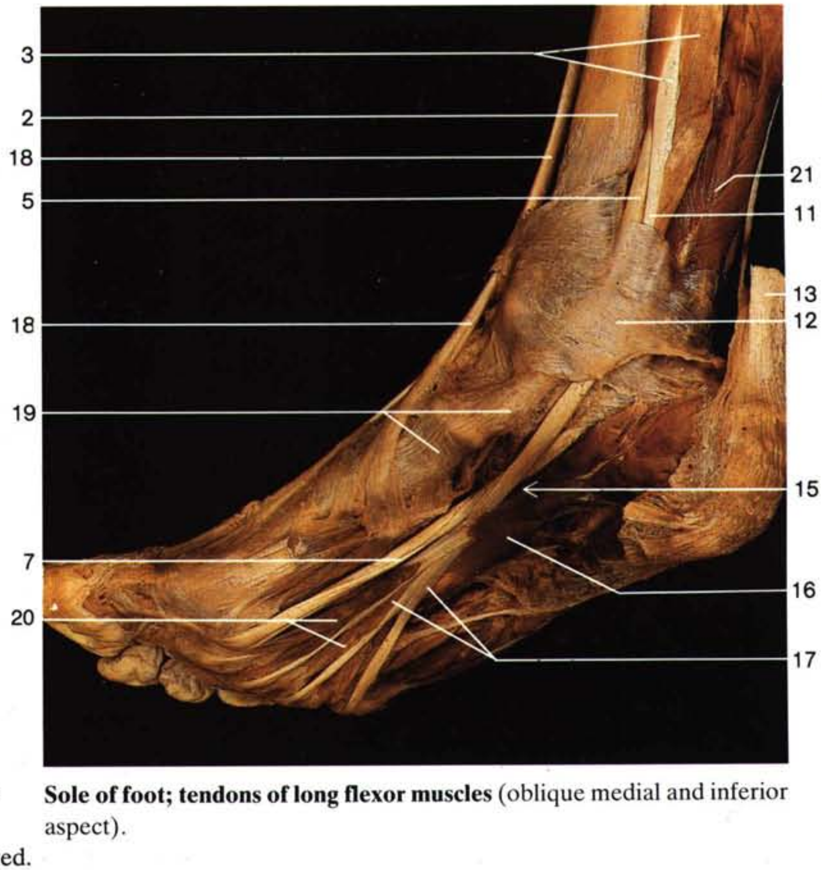

Sole of foot; tendons of long flexor muscles (oblique medial and inferior aspect).

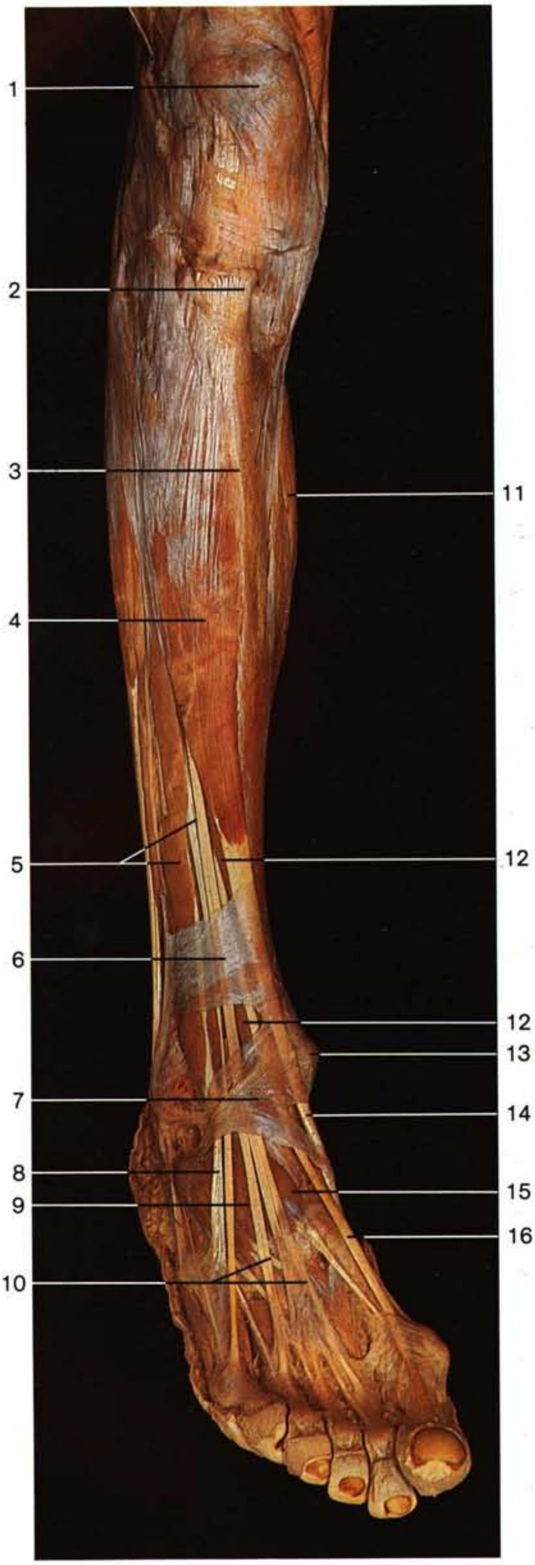

Extensor muscles of right leg and foot (oblique anterolateral aspect).

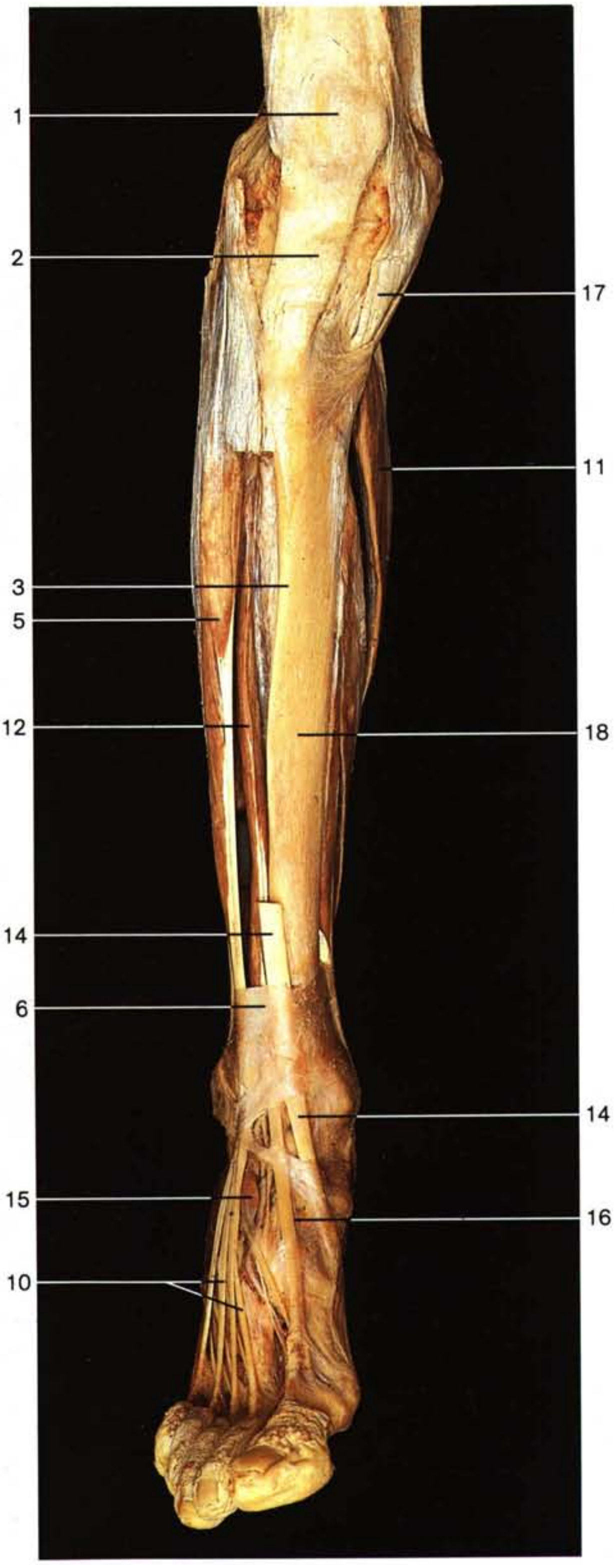

Extensors of the right leg and foot (anterior aspect). Part of the tibialis anterior muscle has been removed.

1 Patella
2 Patellar ligament
3 Anterior margin of tibia
4 Tibialis anterior muscle
5 Extensor digitorum longus muscle
6 Superior extensor retinaculum
7 Inferior extensor retinaculum
8 Tendon of peroneus tertius muscle
9 Extensor digitorum brevis muscle
10 Tendons of extensor digitorum longus muscle
11 Soleus
12 Extensor hallucis longus muscle
13 Medial malleolus
14 Tendon of tibialis anterior muscle
15 Extensor hallucis brevis muscle
16 Tendon of extensor hallucis longus muscle
17 Common tendon of gracilis, semitendinosus and sartoris muscles
18 Tibia

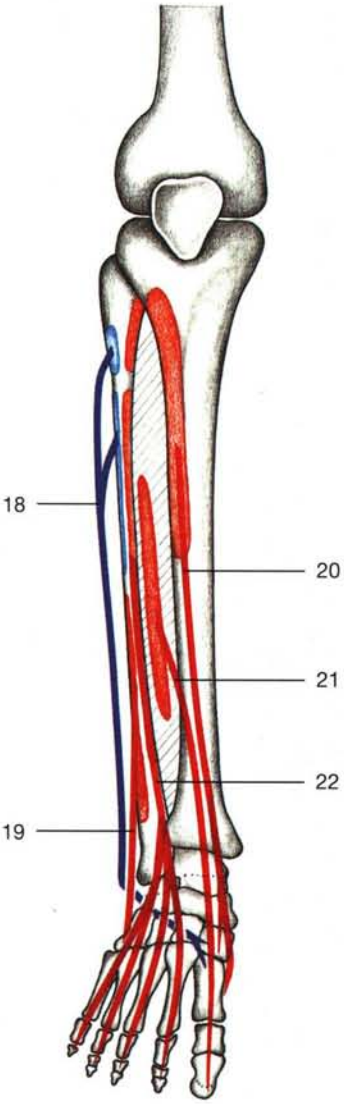

Extensor muscles of the leg (right side).

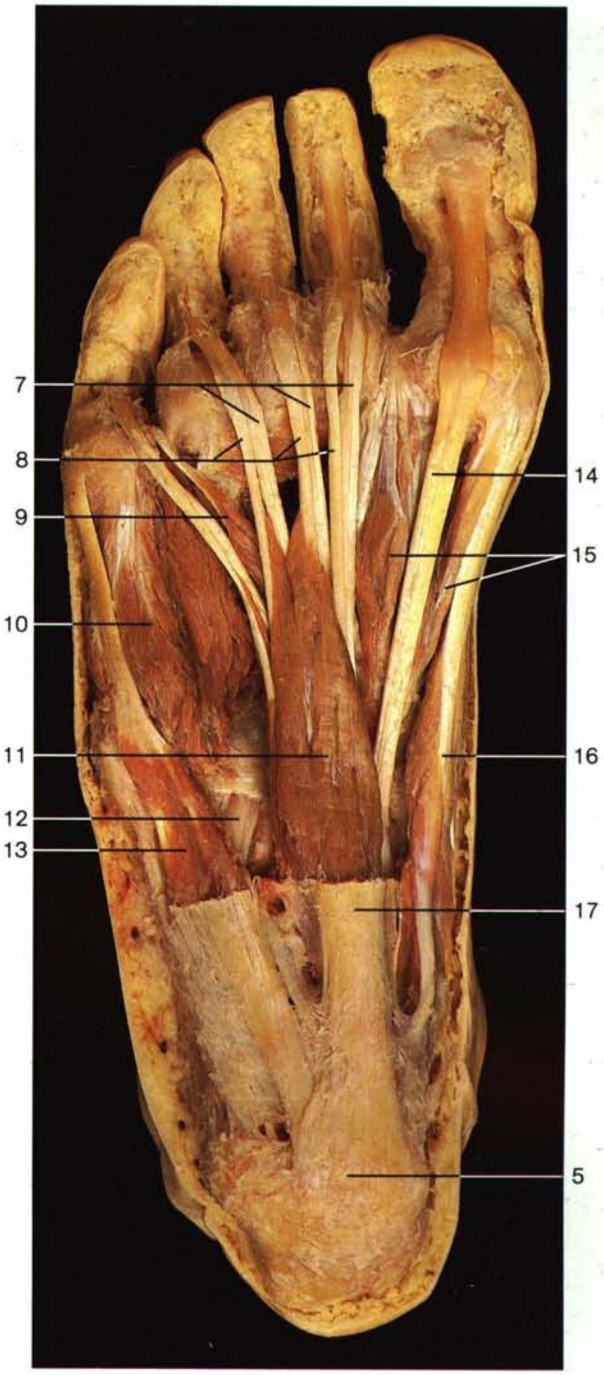

Sole of foot, first layer of muscles (from below). The plantar aponeurosis and the fasciae of the superficial muscles have been removed.

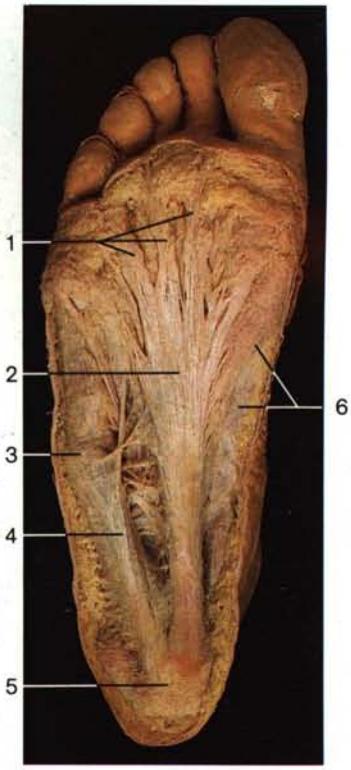

Sole of foot, plantar aponeurosis (from below).

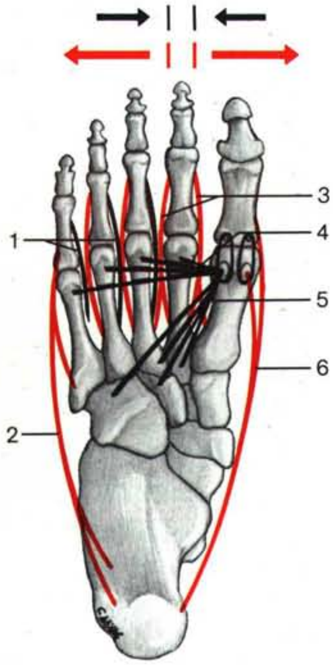

Course of abductor and adductor muscles of foot. (Schematic drawing.)
Red arrows = abduction.
Black arrows = adduction.

1 Longitudinal bands of plantar aponeurosis
2 Plantar aponeurosis
3 Position of tuberosity of 5th metatarsal bone
4 Muscles of 5th toe with fascia
5 Calcaneal tuberosity
6 Muscles of great toe with fascia
7 Tendons of flexor digitorum longus muscle
8 Tendons of flexor digitorum brevis muscle
9 Lumbrical muscle
10 Flexor digiti minimi brevis muscle
11 Flexor digitorum brevis muscle
12 Tendon of peroneus longus muscle
13 Abductor digiti minimi muscle
14 Tendon of flexor hallucis longus muscle
15 Flexor hallucis brevis muscle
16 Abductor hallucis muscle
17 Plantar aponeurosis (cut)
18 Peroneus longus muscle
19 Peroneus brevis muscle
20 Tibialis anterior muscle
21 Extensor hallucis longus muscle
22 Extensor digitorum longus muscle

1 Plantar interossei muscles (black)
2 Abductor digiti minimi muscles (red)
3 Dorsal interossei muscle (red)
4 Transverse head of adductor muscle (black)
5 Oblique head of adductor muscle (black)
6 Abductor hallucis muscle (red)

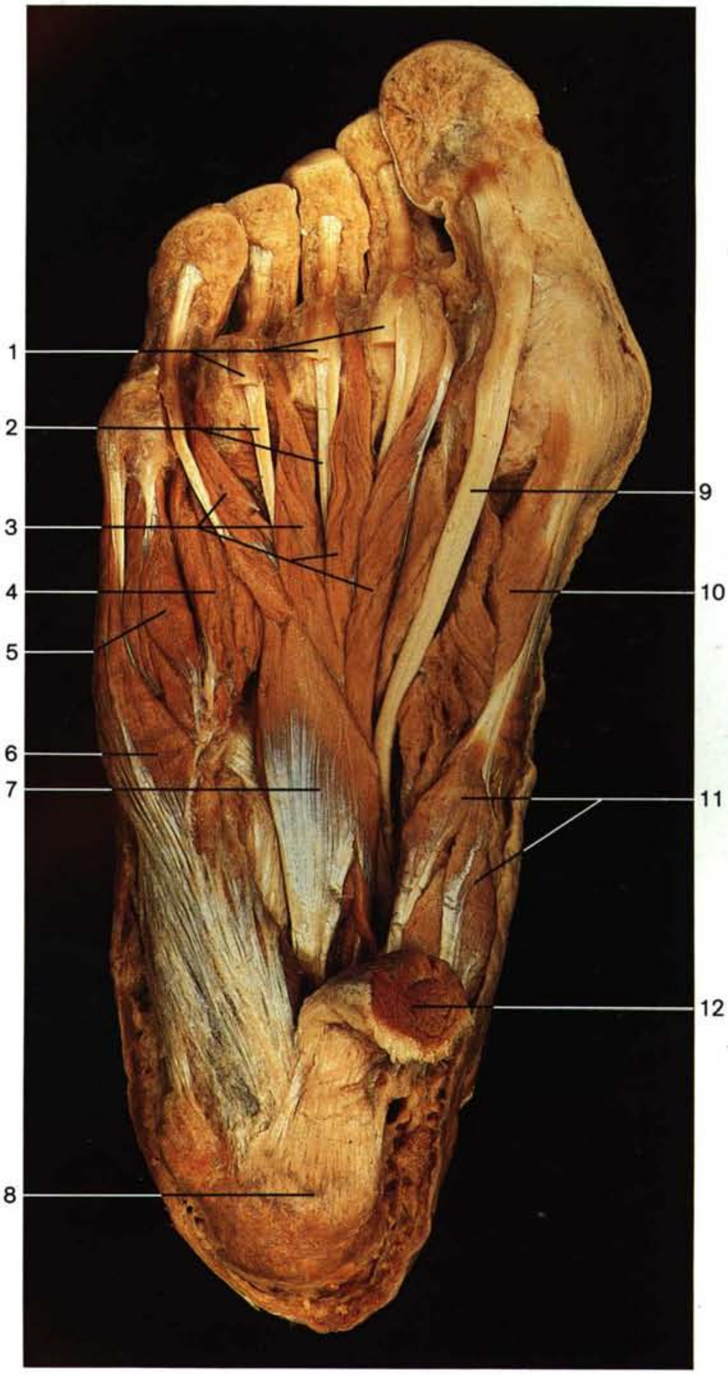

Muscles of sole of foot, second layer (from below). The flexor digitorum brevis muscle has been divided.

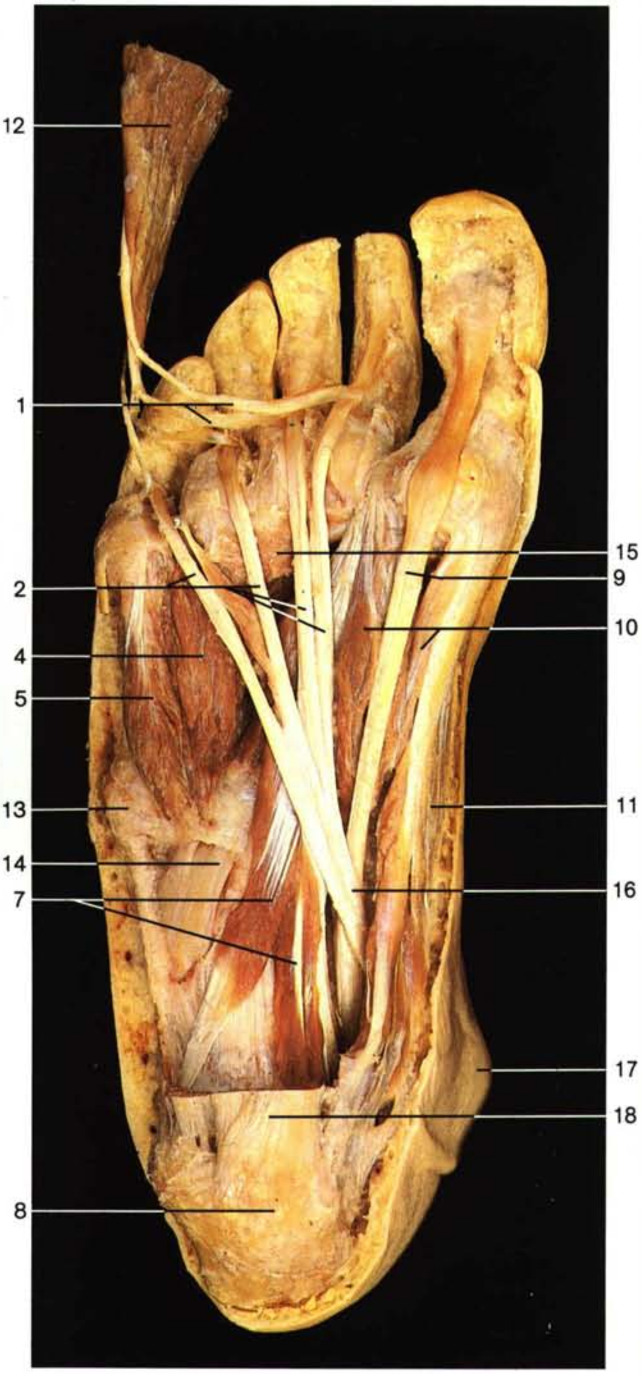

Muscles of sole of foot, second layer (from below). The tendons of the flexor muscles and the crossing of tendons are displayed. The flexor digitorum brevis muscle has been divided and reflected.

1 Tendons of flexor digitorum brevis muscle
2 Tendons of flexor digitorum longus muscle
3 Lumbrical muscles
4 Interossei muscles
5 Flexor digiti minimi brevis muscle
6 Abductor digiti minimi muscle
7 Quadratus plantae muscle
8 Calcaneal tuberosity
9 Tendon of flexor hallucis longus muscle
10 Flexor hallucis brevis muscle
11 Abductor hallucis muscle
12 Flexor digitorum brevis muscle (divided)
13 Tuberosity of 5th metatarsal bone
14 Tendon of peroneus longus muscle
15 Transverse head of adductor hallucis muscle
16 Crossing of tendons in sole of foot
17 Medial malleolus
18 Plantar aponeurosis (divided)

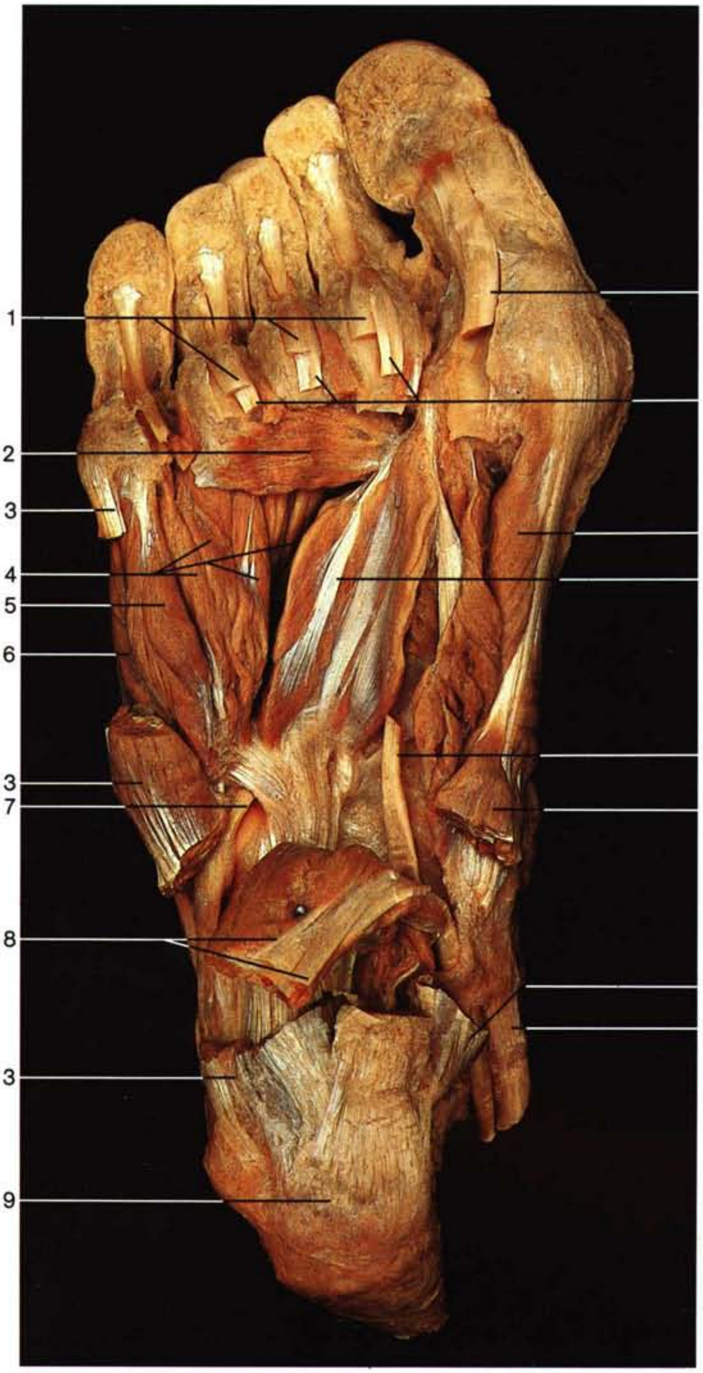

Muscles of sole of foot, third layer (from below). The flexor digitorum brevis muscle has been removed, and the quadratus plantae muscle and the abductor hallucis and digiti minimi muscles have been divided.

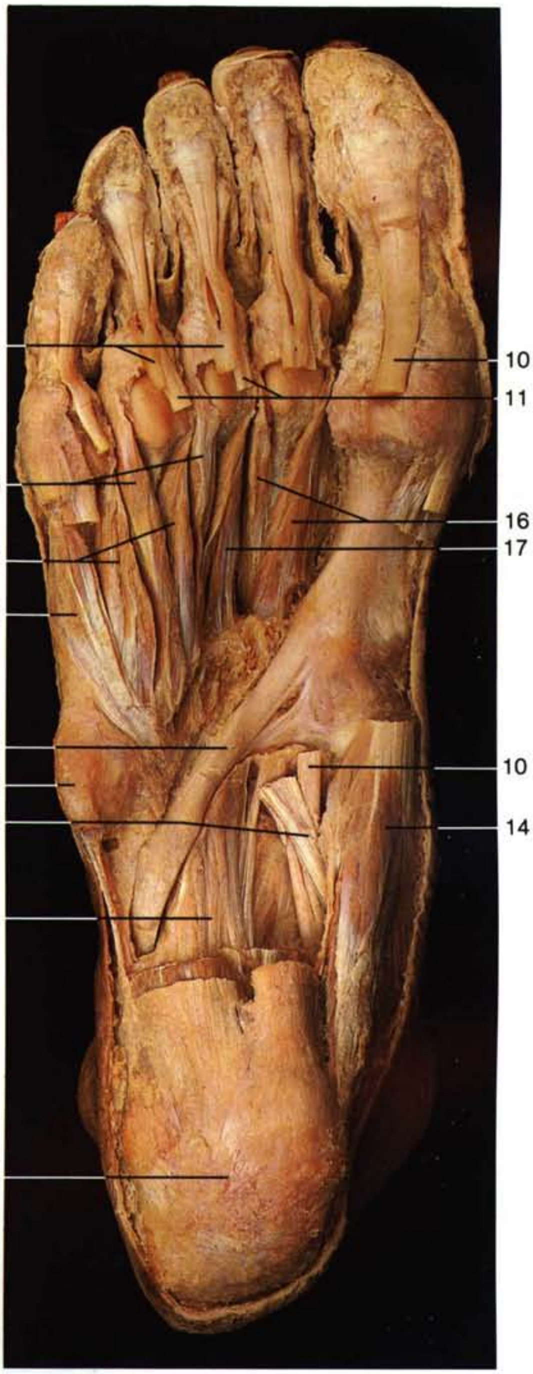

Muscles of sole of foot, fourth layer (from below). The interosseous muscles and the canal for the tendon of peroneus longus muscle are shown.

1 Tendons of flexor digitorum brevis muscle
2 Transverse head of adductor hallucis muscle
3 Abductor digiti minimi muscle
4 Interosseous muscles
5 Flexor digiti minimi brevis muscle
6 Opponens digiti minimi muscle
7 Tendon of peroneus longus muscle
8 Quadratus plantae muscle with tendon of flexor digitorum longus muscle
9 Calcaneal tuberosity
10 Tendons of flexor hallucis longus muscle (divided)
11 Tendon of flexor digitorum longus muscle
12 Flexor hallucis brevis muscle
13 Oblique head of adductor hallucis muscle
14 Abductor hallucis muscle (cut)
15 Tendon of tibialis posterior muscle
16 Dorsal interossei muscles
17 Plantar interossei muscles
18 Tuberosity of 5th metatarsal bone
19 Tendon of flexor digitorum longus muscle (crossing of plantar tendons)
20 Long plantar ligament

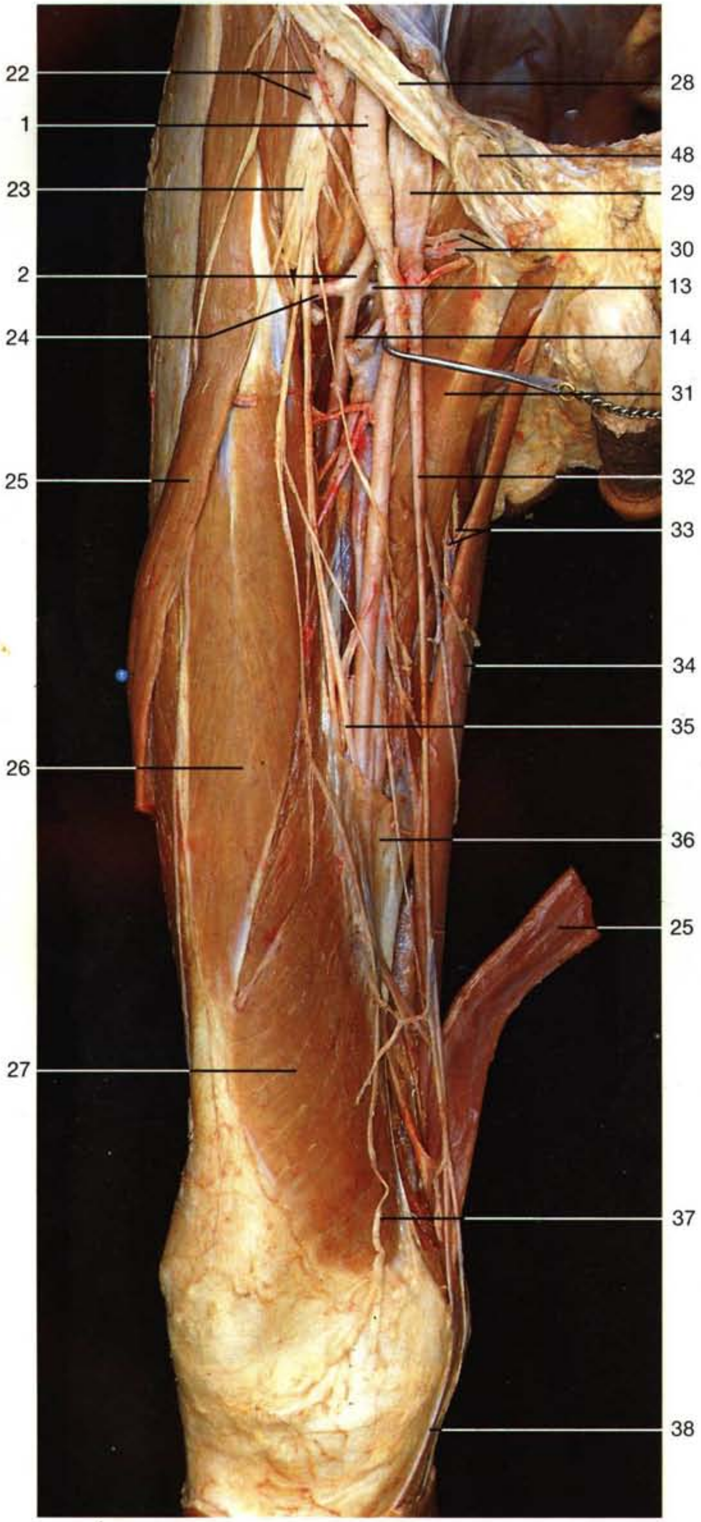
22
1
23
2
24
25
26
27
28
48
29
30
13
14
31
32
33
34
35
36
25
37
38

Main arteries and nerves of right thigh (anterior aspect). Sartorius muscle has been divided and reflected. The femoral vein has been partly removed to show the deep femoral artery. Notice: the vessels enter the adductor canal to reach the popliteal fossa.

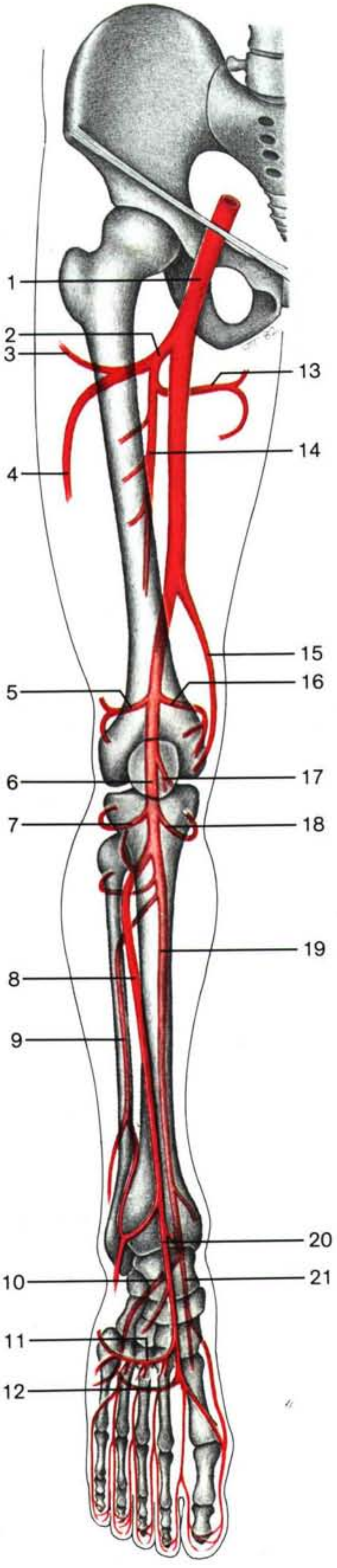
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Main arteries of lower extremity, right side (ventral aspect). (Schematic drawing.)

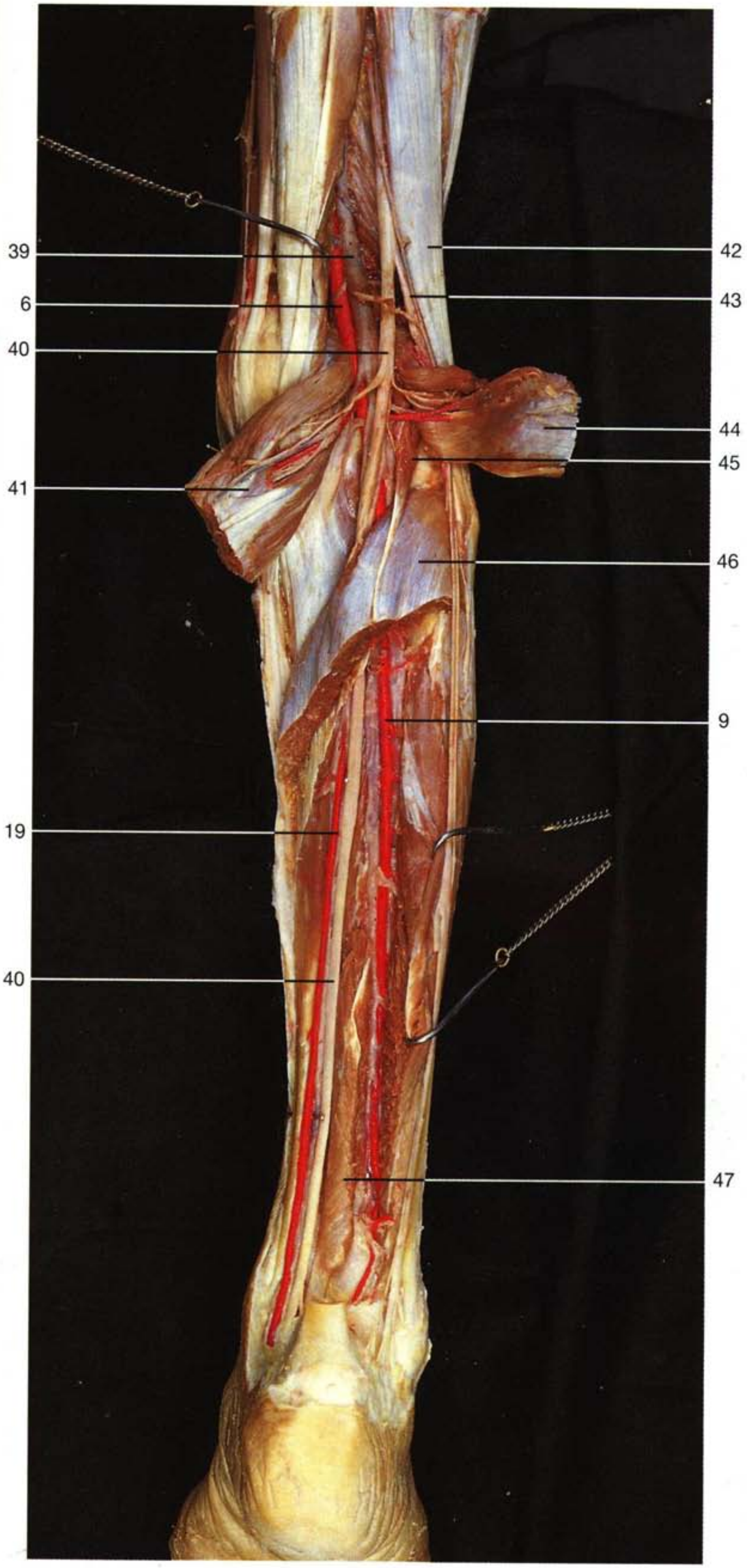

1 **Femoral artery**
2 **Profunda femoris artery**
3 Ascending branch of lateral circumflex femoral artery
4 Descending branch of lateral circumflex femoral artery
5 Lateral superior genicular artery
6 Popliteal artery
7 Lateral inferior genicular artery
8 **Anterior tibial artery**
9 **Peroneal artery**
10 **Lateral plantar artery**
11 Arcuate artery with dorsal metatarsal arteries
12 **Plantar arch** with plantar metatarsal arteries
13 Medial circumflex femoral artery
14 Profunda femoris artery with perforating arteries
15 Descending genicular artery
16 Medial superior genicular artery
17 Middle genicular artery
18 Medial inferior genicular artery
19 **Posterior tibial artery**
20 **Dorsalis pedis artery**
21 **Medial plantar artery**
22 Superficial and deep circumflex iliac arteries
23 **Femoral nerve**
24 Lateral circumflex femoral artery
25 Sartorius muscle (cut and reflected)
26 Rectus femoris muscle
27 Vastus medialis muscle
28 Inguinal ligament
29 **Femoral vein** (cut)
30 External pudendal artery and vein
31 Adductor longus muscle
32 Great saphenous vein
33 Obturator artery and nerve
34 Gracilis muscle
35 Saphenous nerve
36 Tendinous wall of adductor canal
37 Anterior cutaneous branch of femoral nerve
38 Infrapatellar branch of saphenous nerve
39 Popliteal vein
40 **Tibial nerve**
41 Medial head of gastrocnemius muscle
42 Biceps femoris muscle
43 Common peroneal nerve
44 Lateral head of gastrocnemius muscle
45 Plantaris muscle
43 Soleus muscle
47 Flexor hallucis longus muscle
48 Spermatic cord

Arteries of the right leg (posterior aspect).

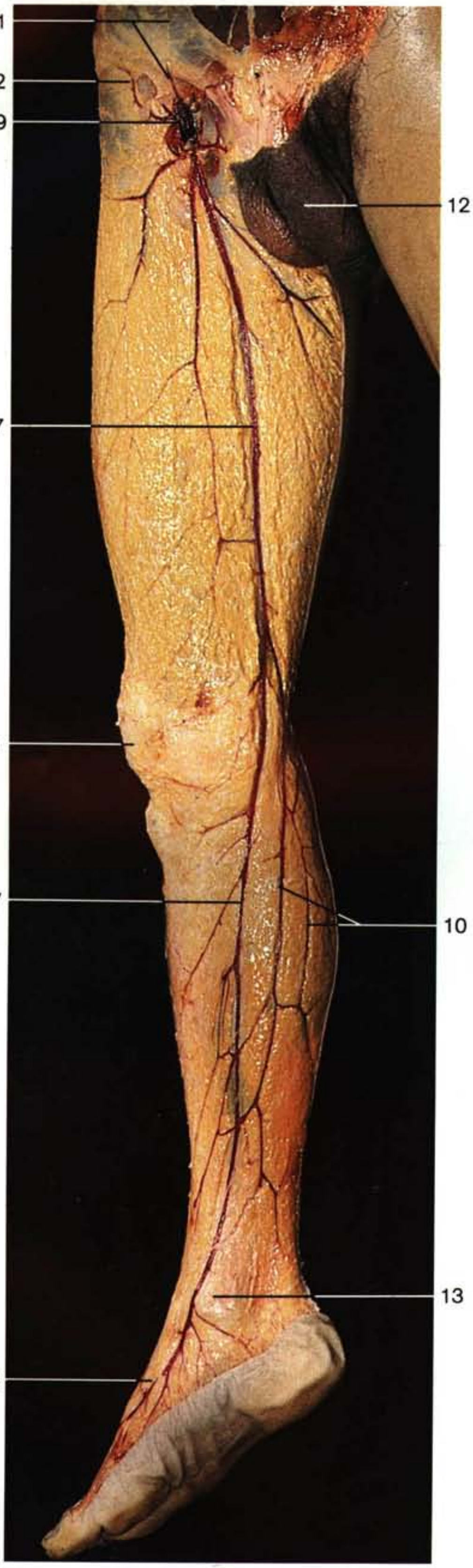

Superficial veins of lower limb, right side (medial anterior aspect). The veins have been injected with red solution.

Main veins of lower limb, right side (anterior aspect). (Schematic drawing.)

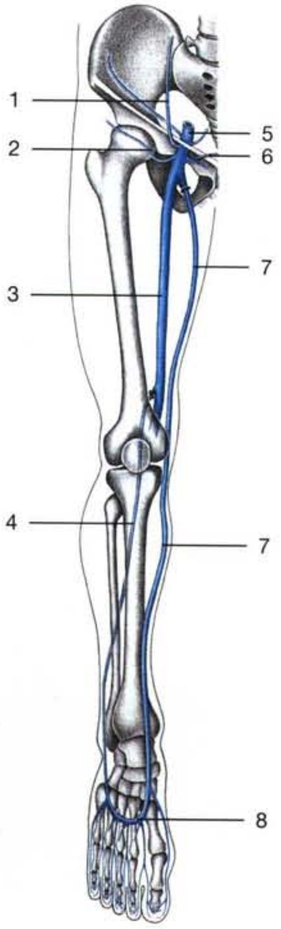

1	Superficial epigastric vein
2	Superficial circumflex iliac vein
3	Femoral vein
4	**Small saphenous vein**
5	External iliac vein
6	External pudendal vein
7	**Great saphenous vein**
8	Dorsal venous arch
9	Saphenous opening with femoral vein
10	Venous anastomoses of small saphenous vein with great saphenous vein
11	Patella
12	Penis
13	Medial malleolus
14	Popliteal fossa
15	Perforating veins
16	Lateral malleolus
17	Dorsal digital veins of foot
18	Dorsal venous arch of foot
19	Dorsal metatarsal veins of foot
20	Anterior tibial artery and veins
21	Tibia
22	**Posterior tibial artery and veins**
23	Fibula
24	Peroneal artery and vein
25	Deep layer of crural fascia
26	Superficial layer of crural fascia
27	Perforating veins I-III (of Cockett)
28	**Tibial nerve**
29	**Arcuate vein**
30	**Saphenous nerve**
31	Medial dorsal cutaneous nerve (branch of superficial peroneal nerve)
32	Posterior tibial vein

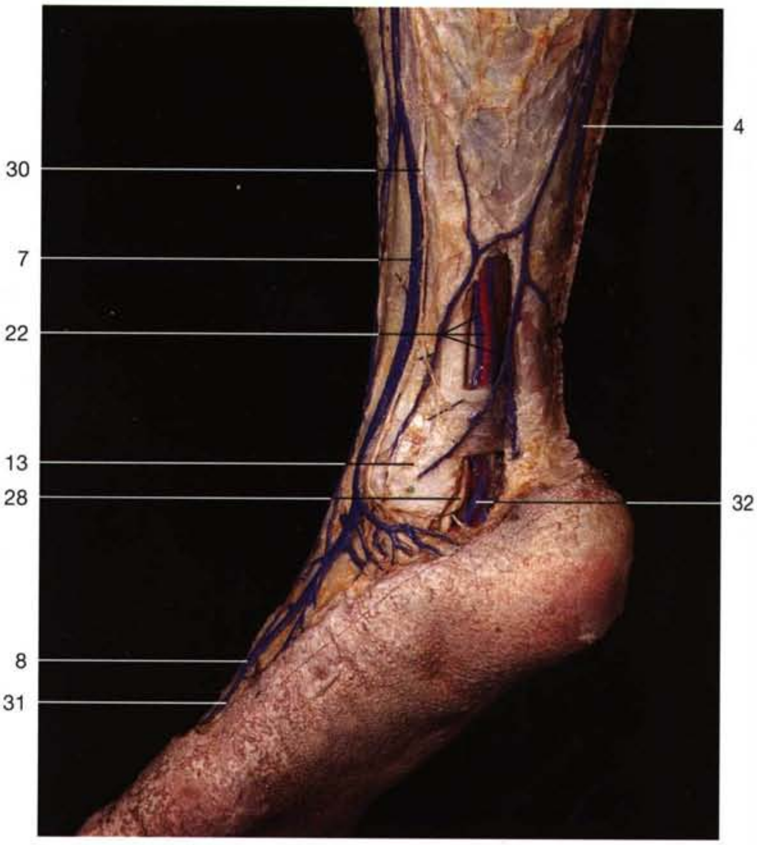

▷

Medial malleolar region. Dissection of tibial nerve, posterior tibial vessels, and great saphenous vein (veins injected with blue resin).

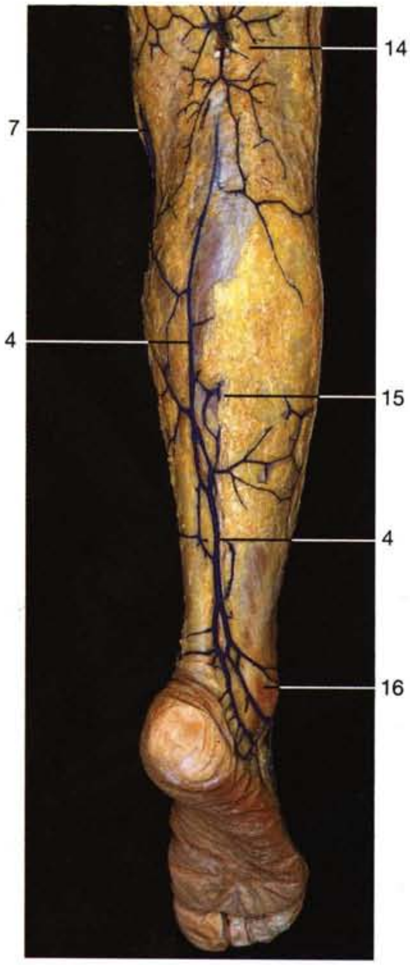

Superficial veins of leg (posterior aspect; injected with blue resin).

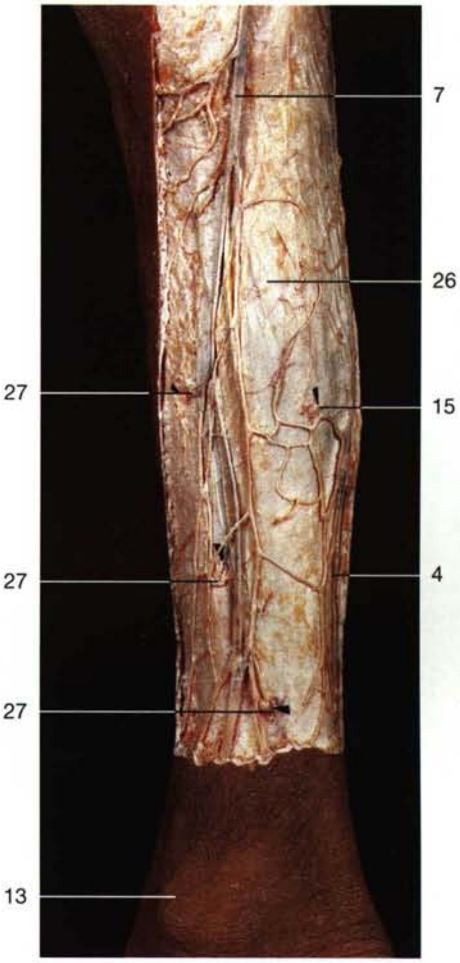
7
26
27
15
27
4
27
13

Superficial veins of leg. The perforating veins of Cockett have been dissected.

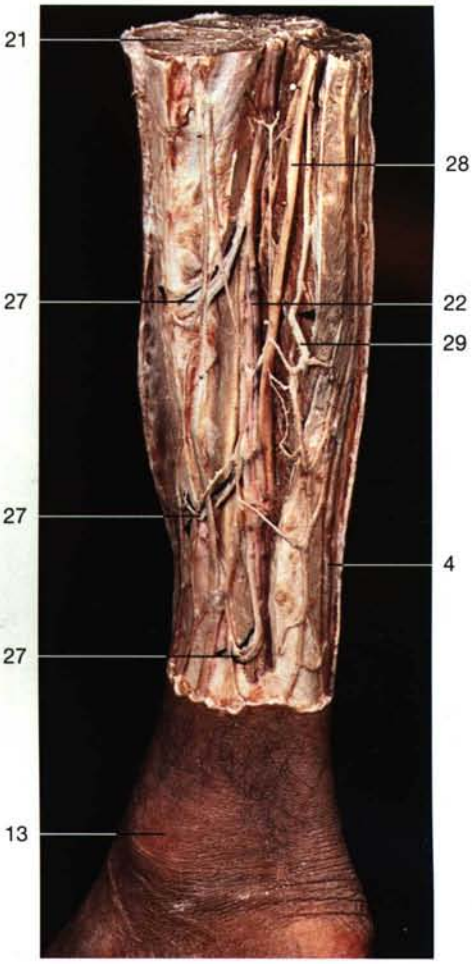
21
28
27
22
29
27
4
27
13

Veins of leg. The anastomoses between superficial and deeper veins are dissected.

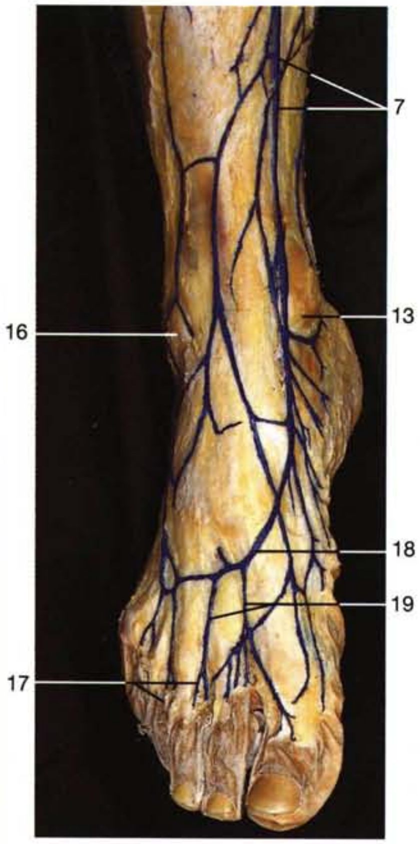
7
16
13
18
19
17

◁

Superficial veins on dorsum of foot (injected with blue resin).

Anastomoses between superficial and deep veins of the leg (after Aigner). (Schematic drawing.) Arrows: directions of blood flow. ▷

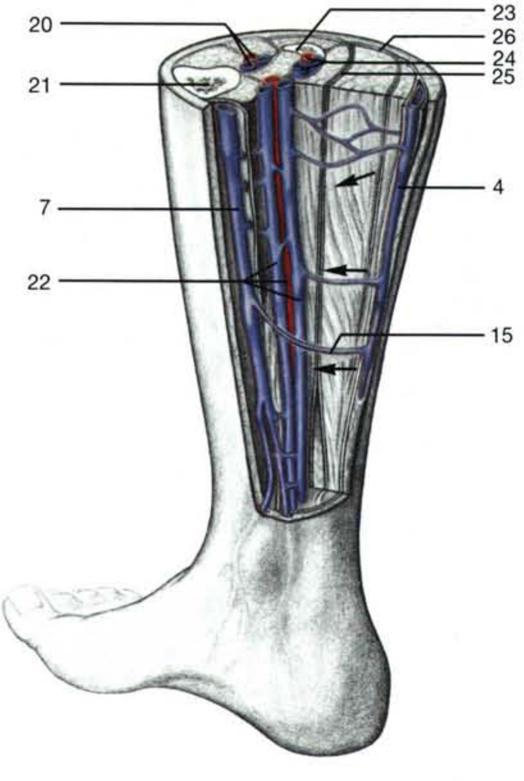
20
23
26
24
21
25
4
7
22
15

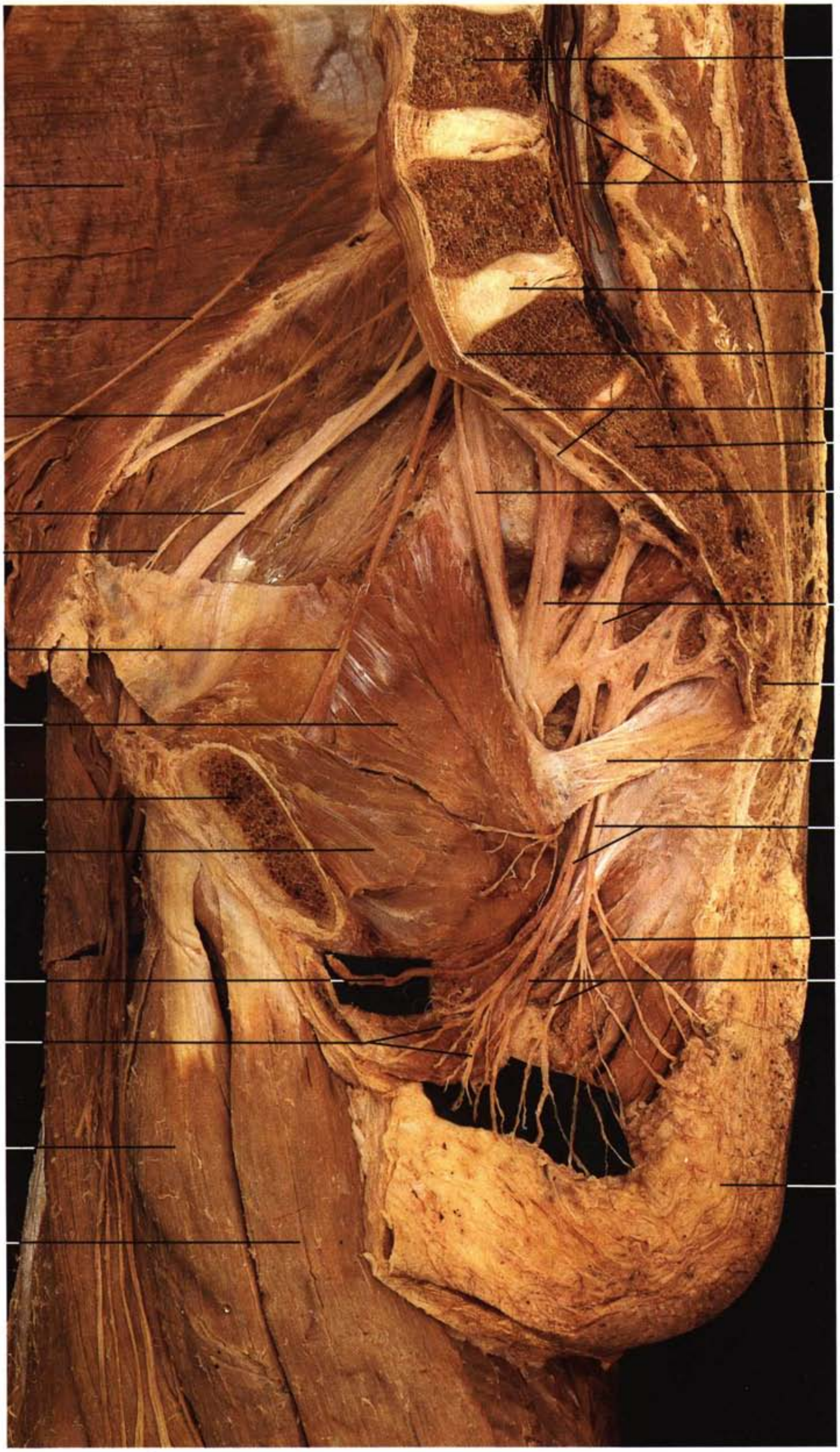

1 Transversus abdominis muscle
2 Iliohypogastric nerve
3 Ilioinguinal nerve
4 **Femoral nerve**
5 Lateral femoral cutaneous nerve
6 Obturator nerve
7 Obturator internus muscle
8 Pubic bone (cut edge)
9 Levator ani muscle (remnant)
10 Dorsal nerve of penis
11 Posterior scrotal nerves
12 Adductor longus muscle
13 Gracilis muscle
14 Body of 4th lumbar vertebra
15 Cauda equina
16 Intervertebral disc
17 Sacral promontory
18 Sympathetic trunk
19 Sacrum
20 **Lumbosacral trunk**
21 **Sciatic plexus**
22 Coccyx
23 Sacrospinous ligament
24 **Pudendal nerve**
25 Inferior rectal nerves
26 Perineal nerves
27 Subcutaneous fat tissue of gluteal region

Lumbosacral plexus in situ, right side (medial aspect).
Pelvic organs with peritoneum and part of the levator ani muscle have been removed.

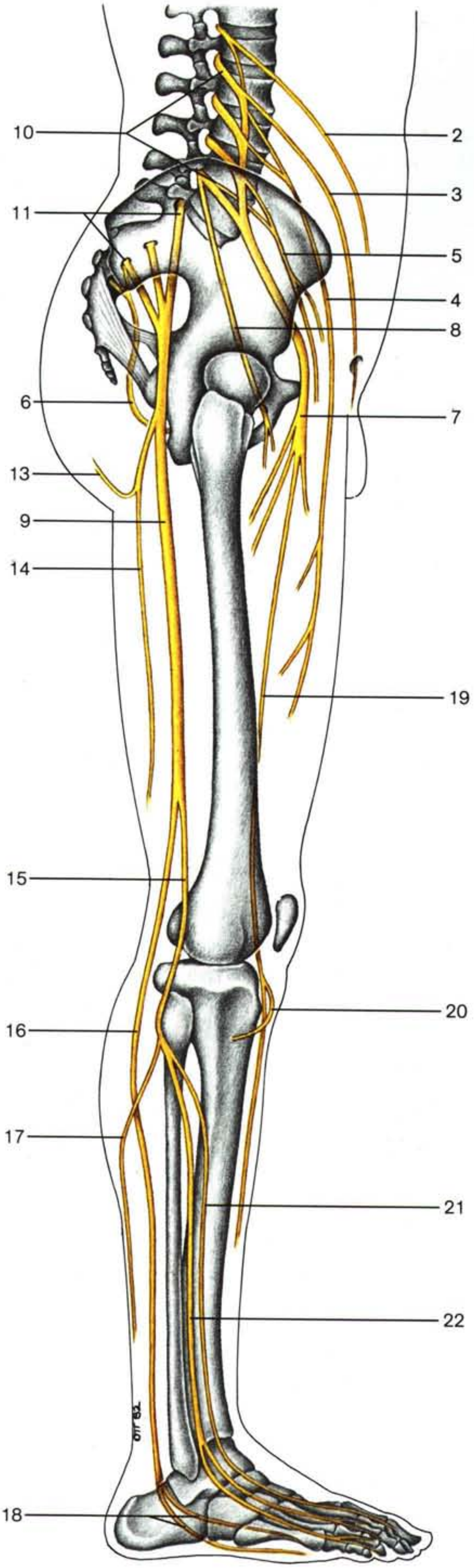

Nerves of lower limb, right side (lateral aspect). (Schematic drawing.)

1	Subcostal nerve	
2	Iliohypogastric nerve	
3	Ilioinguinal nerve	
4	Lateral femoral cutaneous nerve	
5	Genitofemoral nerve	
6	**Pudendal nerve**	
7	**Femoral nerve**	
8	Obturator nerve	
9	**Sciatic nerve**	
10	Lumbar plexus (L_1–L_4)	lumbosacral plexus
11	Sacral plexus (L_4–S_4)	
12	"Pudendal" plexus (S_2–S_4)	
13	Inferior cluneal nerve	
14	Posterior femoral cutaneous nerve	
15	Common peroneal nerve	
16	**Tibial nerve**	
17	Lateral sural cutaneous nerve	
18	Medial and lateral plantar nerves	
19	Saphenous nerve	
20	Infrapatellar branch of saphenous nerve	
21	**Deep peroneal nerve**	
22	**Superficial peroneal nerve**	

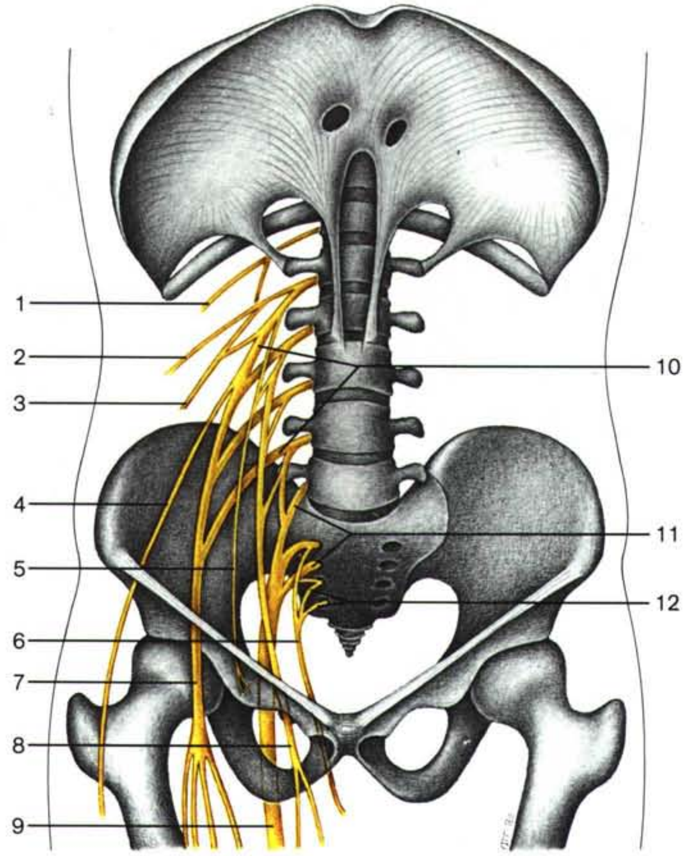

Main branches of lumbosacral plexus (ventral aspect). (Schematic drawing.)

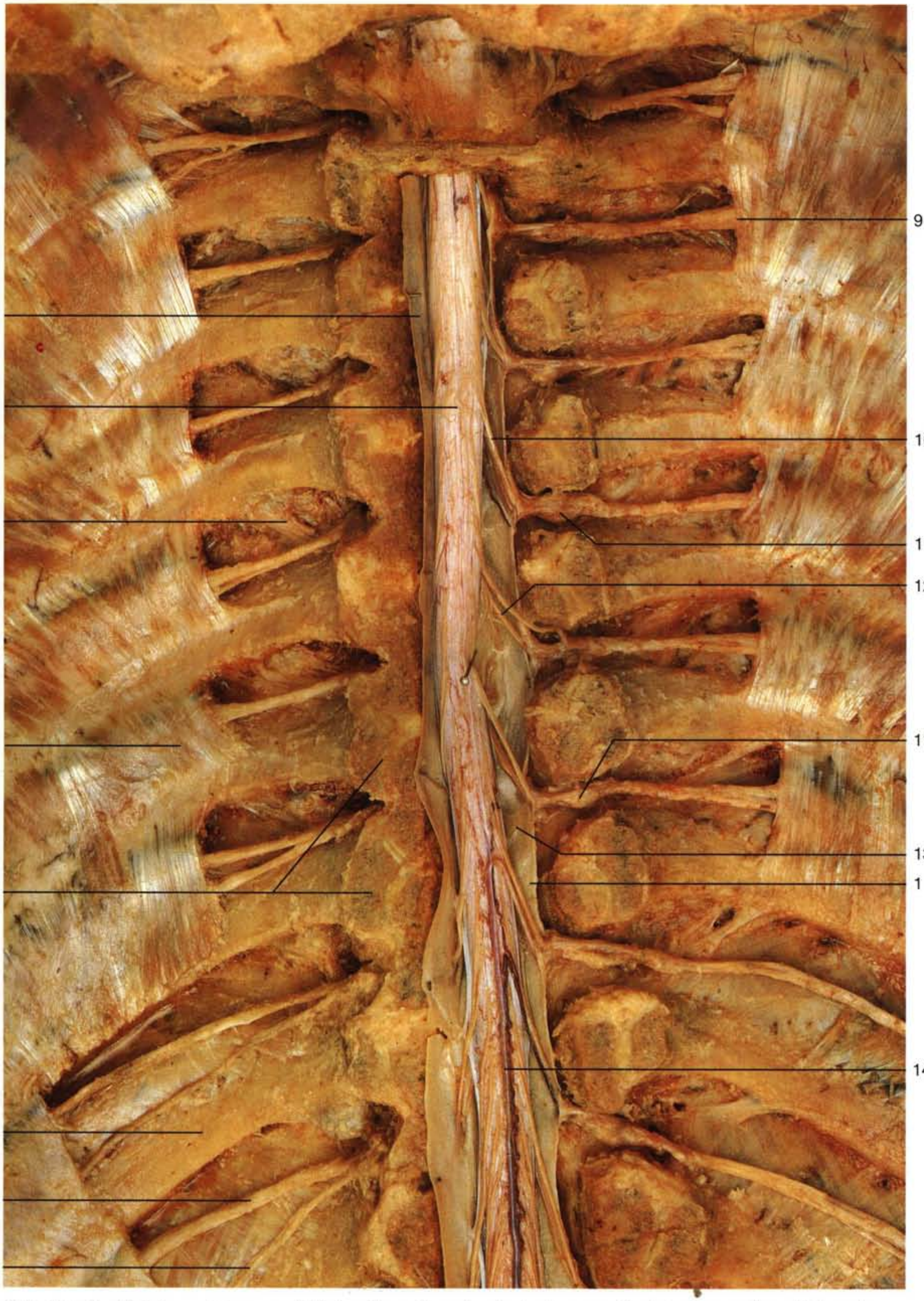

Spinal cord with intercostal nerves. Inferior thoracic region (anterior aspect). Anterior portion of thoracic vertebrae removed, dural sheath opened and spinal cord slightly reflected to the right to display the dorsal and ventral roots.

1 Dura mater
2 **Spinal cord**
3 Costotransverse ligament
4 Innermost intercostal muscle
5 Vertebral arches (cut surfaces)

6 Eleventh rib
7 **Intercostal nerve**
8 Collateral branch of intercostal nerve
9 Intercostal nerve (entering the intermuscular interval)

10 Anterior root filaments
11 Spinal (dorsal root) ganglion
12 Posterior root filaments
13 Arachnoid mater and denticulate ligament
14 Anterior spinal artery

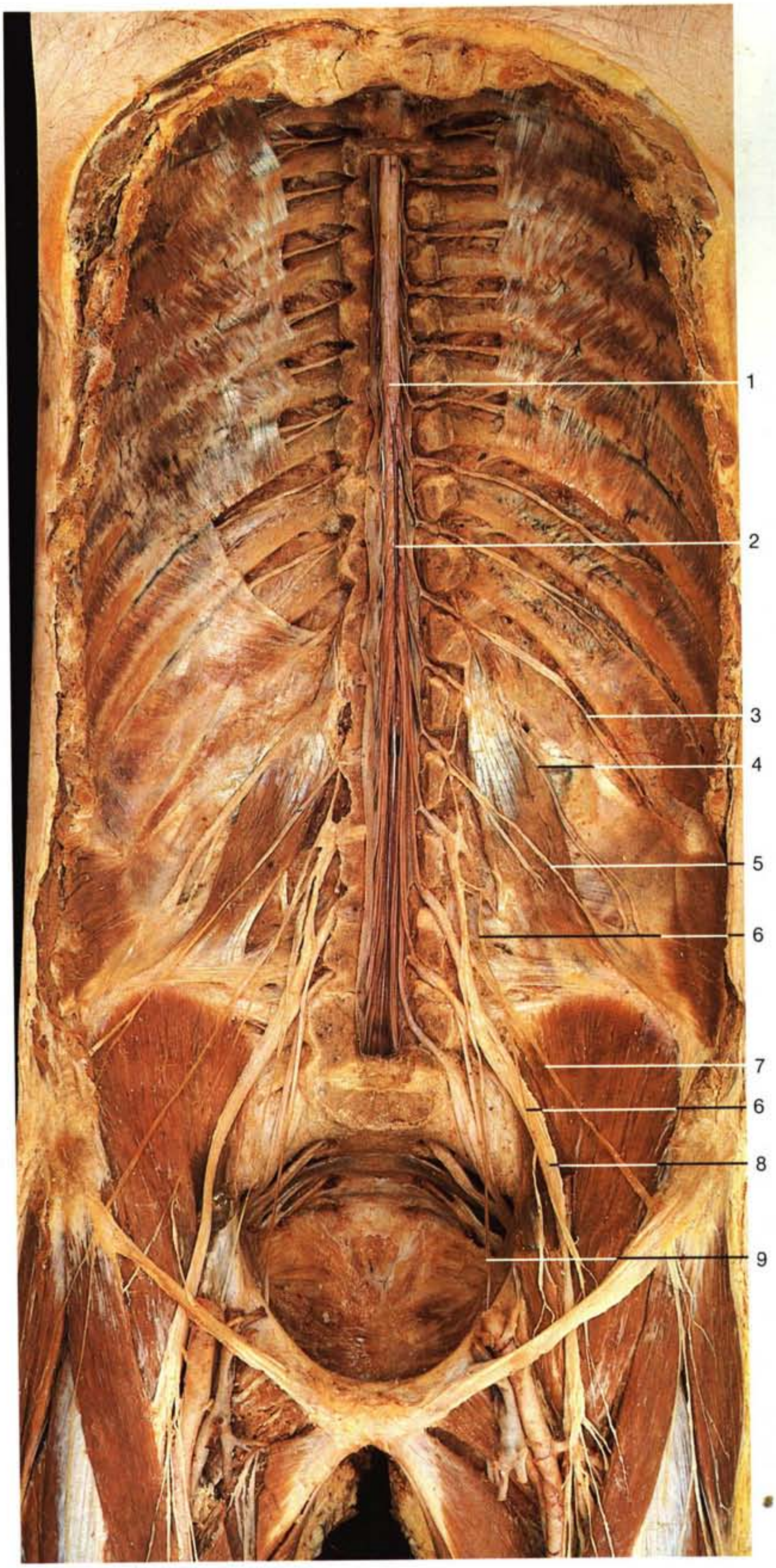
1
2
3
4
5
6
7
6
8
9

Spinal cord and lumbar plexus in situ (anterior aspect).

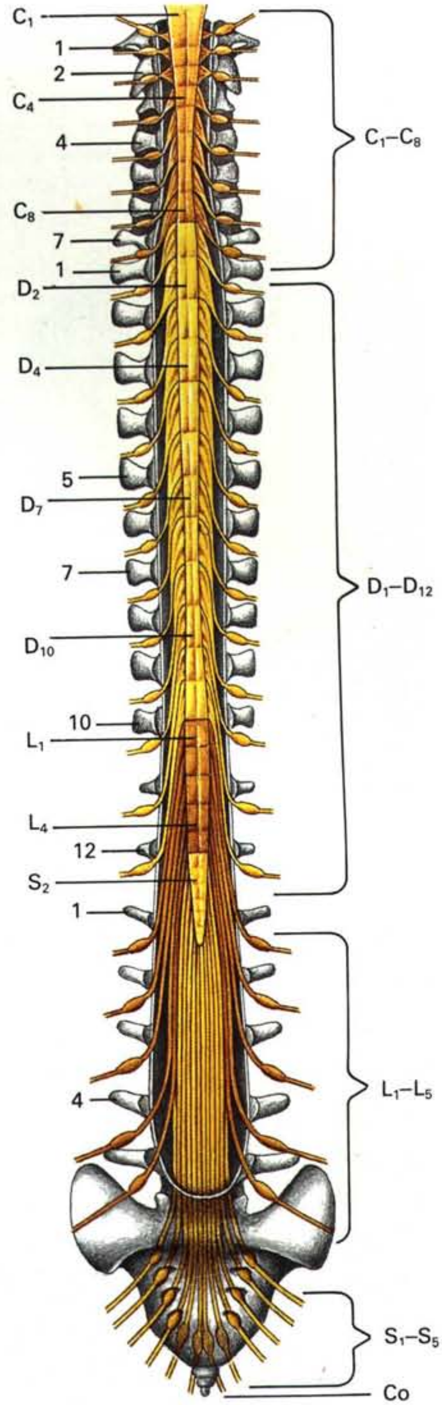
C_1
1
2
C_4
4
C_8
7
1
D_2
D_4
5
D_7
7
D_{10}
10
L_1
L_4
12
S_2
1
4
C_1–C_8
D_1–D_{12}
L_1–L_5
S_1–S_5
Co

Organization of spinal cord segments in relation to the vertebral column (anterior aspect). C = cervical; D = thoracic; L = lumbar; S = sacral segments; Co = coccygeal bone.

1 Conus medullaris
2 Filum terminale
3 Subcostal nerve
4 Iliohypogastric nerve
5 Ilioinguinal nerve
6 Genitofemoral nerve
7 Lateral femoral cutaneous nerve
8 Femoral nerve
9 Obturator nerve

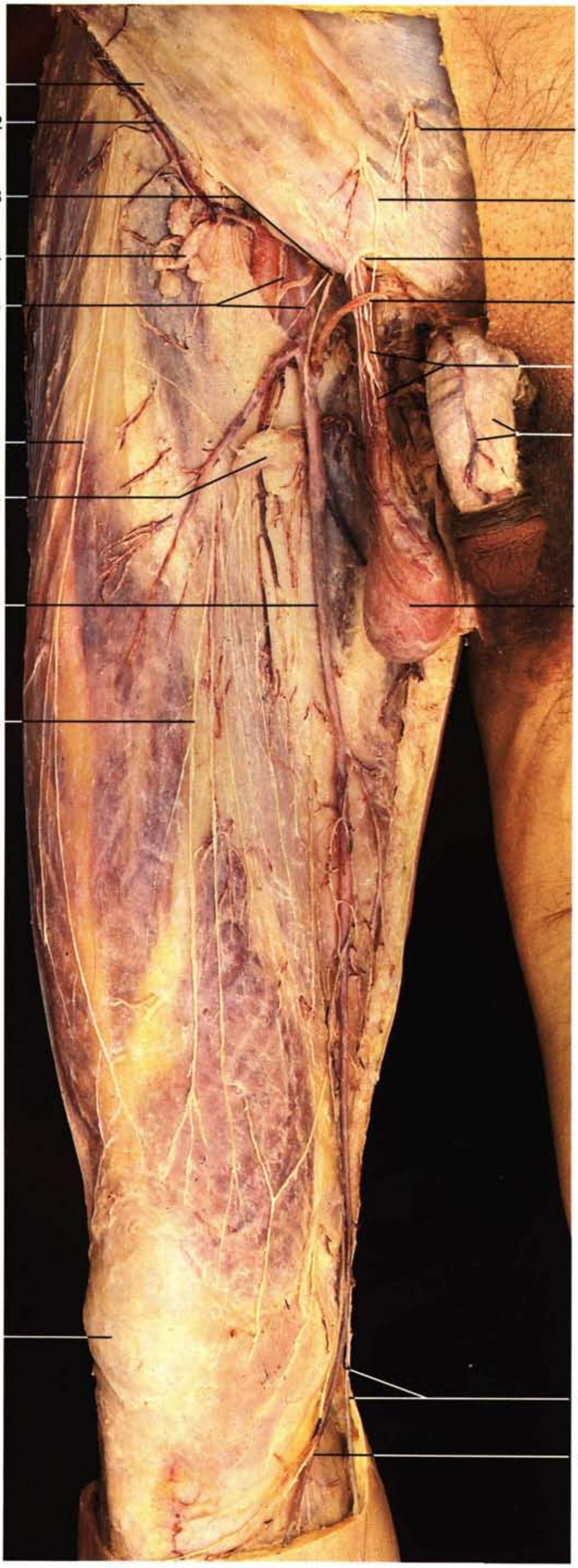
1
2
3
4
5
6
4
7
8
9
10
11
12
13
14
15
16
17
18

Cutaneous nerves and veins of thigh (anterior aspect).

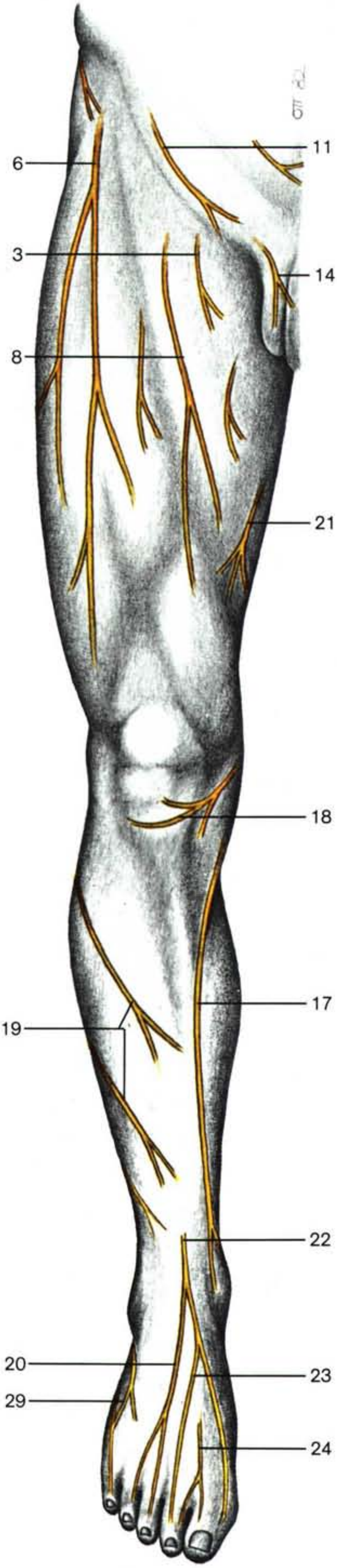
6
11
3
14
8
21
18
17
19
22
20
23
29
24

Cutaneous nerves of lower limb (anterior aspect). (Schematic drawing.)

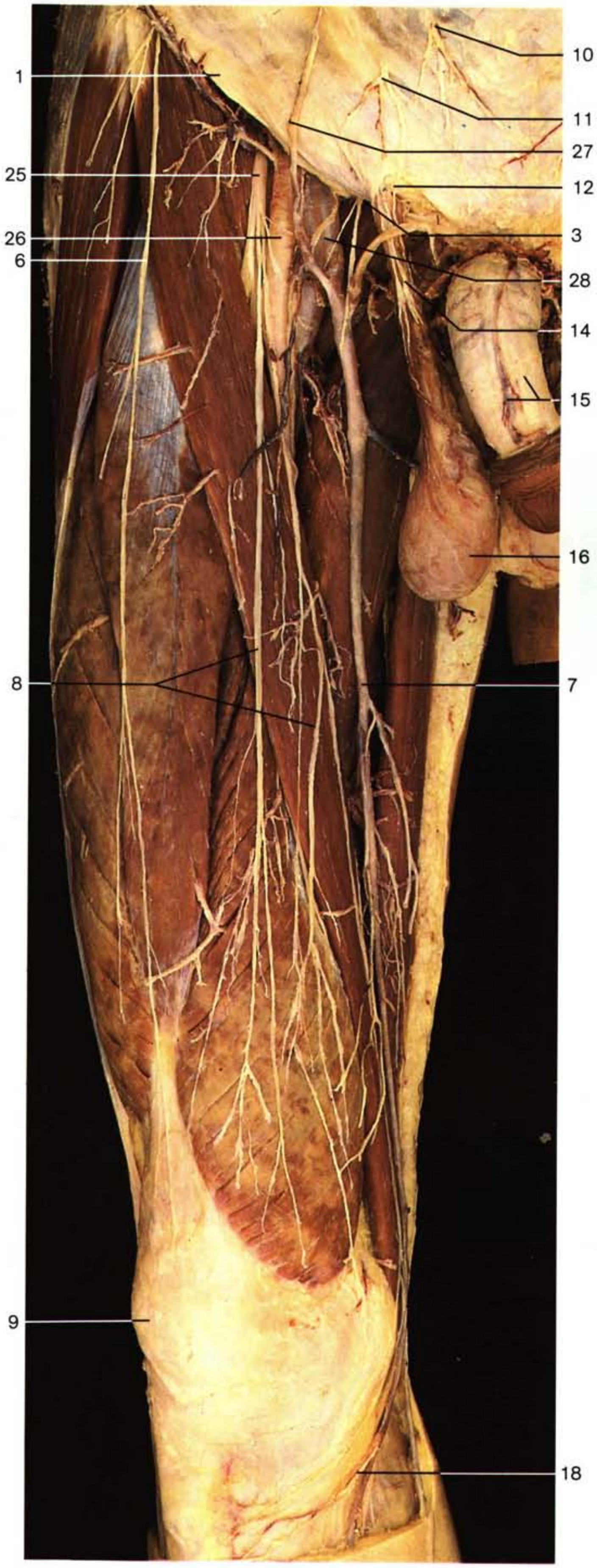

Cutaneous nerves and veins of thigh (anterior aspect). The fascia lata and fasciae of the thigh muscles have been removed.

1 Inguinal ligament
2 Superficial circumflex iliac vein
3 Femoral branch of genitofemoral nerve
4 Superficial inguinal lymph nodes
5 Saphenous opening with femoral artery and vein
6 Lateral femoral cutaneous nerve
7 Great saphenous vein
8 Anterior cutaneous branches of femoral nerve
9 Patella
10 Terminal branches of subcostal nerve
11 Terminal branches of iliohypogastric nerve
12 Superficial inguinal ring
13 External pudendal vein
14 Spermatic cord with genital branch of genitofemoral nerve
15 Penis with superficial dorsal vein of penis
16 Testis and its coverings
17 Saphenous nerve
18 Infrapatellar branch of saphenous nerve
19 Lateral sural cutaneous nerves
20 Intermediate dorsal cutaneous branch of superficial peroneal nerve
21 Cutaneous branch of obturator nerve
22 Superficial peroneal nerve
23 Medial dorsal cutaneous branch of superficial peroneal nerve
24 Deep peroneal nerve
25 **Femoral nerve**
26 **Femoral artery**
27 Superficial epigastric vein
28 **Femoral vein**
29 Lateral dorsal cutaneous branch of sural nerve
30 **Inguinal nodes** (enlarged)
31 Lymphatic vessels
32 Sartorius muscle

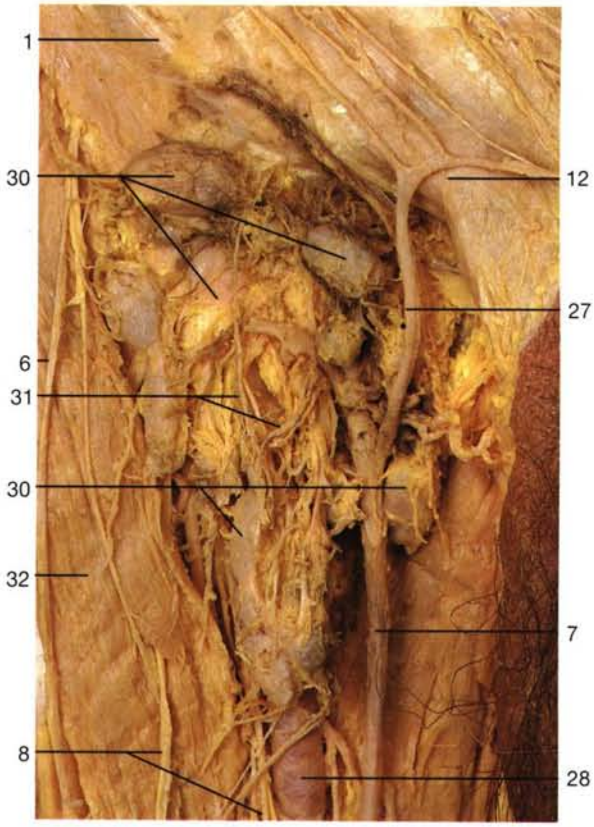

Inguinal nodes with lymphatic vessels (anterior aspect).

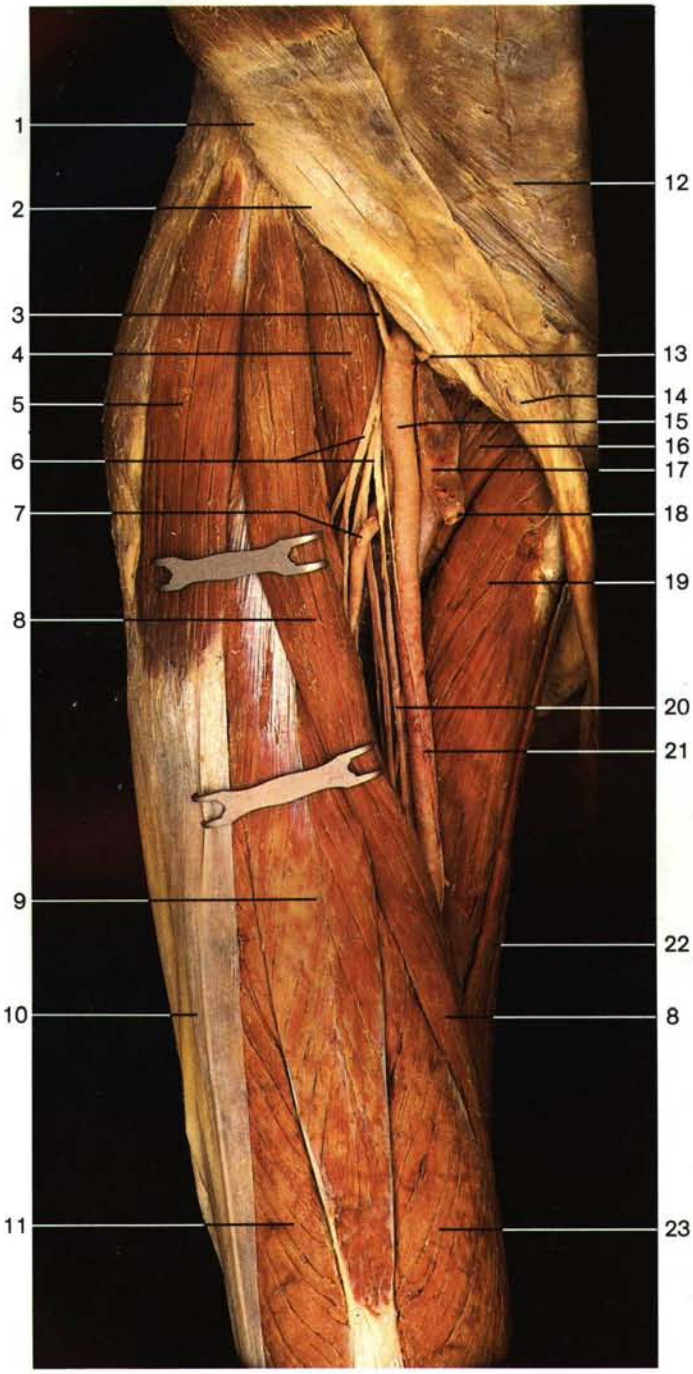

Anterior region of right thigh (anterior aspect). The fascia lata has been removed, and the sartorius muscle has been slightly reflected.

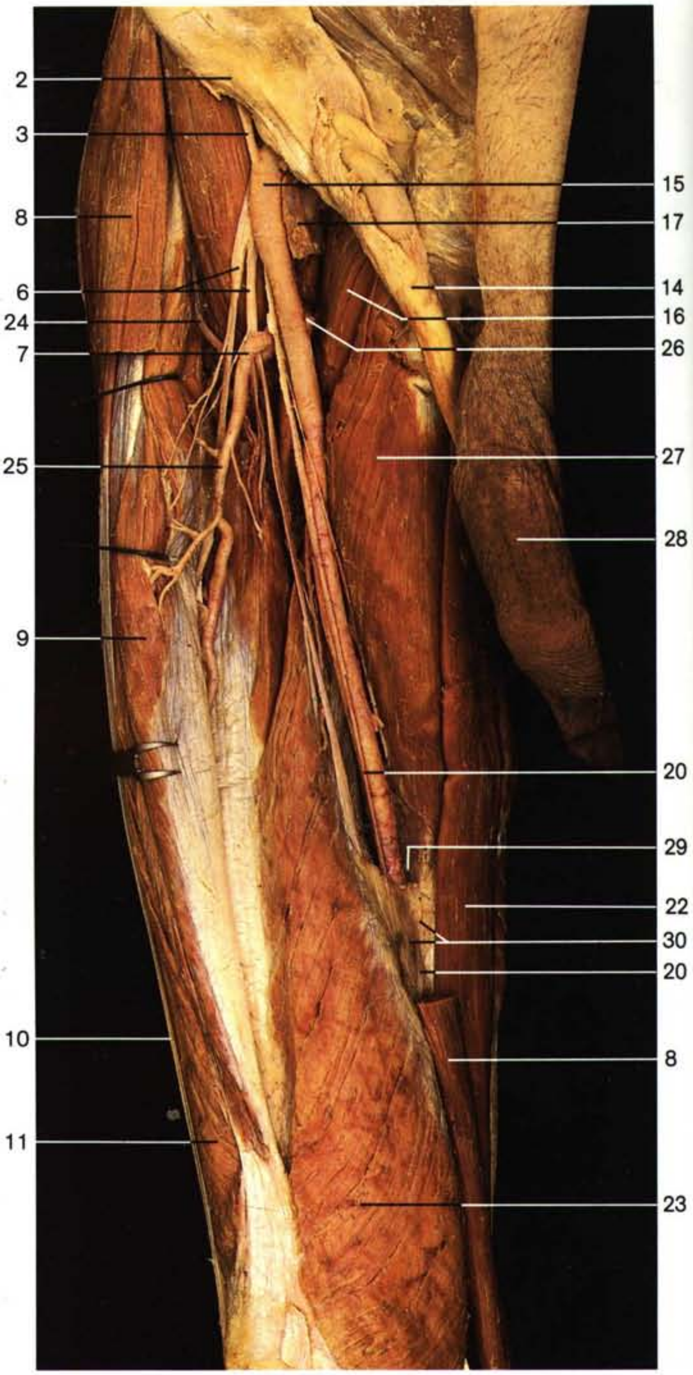

Anterior region of right thigh (anterior aspect). The fascia lata has been removed, and the sartorius muscle has been divided.

1	Anterior superior iliac spine
2	Inguinal ligament
3	**Deep circumflex iliac artery**
4	Iliopsoas muscle
5	Tensor fasciae latae muscle
6	**Femoral nerve**
7	**Lateral circumflex femoral artery**
8	Sartorius muscle
9	Rectus femoris muscle
10	Iliotibial tract
11	Vastus lateralis muscle
12	Anterior sheath of rectus abdominis muscle
13	Inferior epigastric artery
14	Spermatic cord
15	**Femoral artery**
16	Pectineus muscle
17	**Femoral vein**
18	Great saphenous vein (divided)
19	Adductor longus muscle
20	**Saphenous nerve**
21	Muscular branch of femoral nerve
22	Gracilis muscle
23	Vastus medialis muscle
24	Ascending branch of lateral circumflex femoral artery
25	Descending branch of lateral circumflex femoral artery
26	Medial circumflex femoral artery
27	Adductor longus muscle
28	Penis
29	Entrance to adductor canal
30	Vastoadductory lamina of fascia beneath sartorius muscle

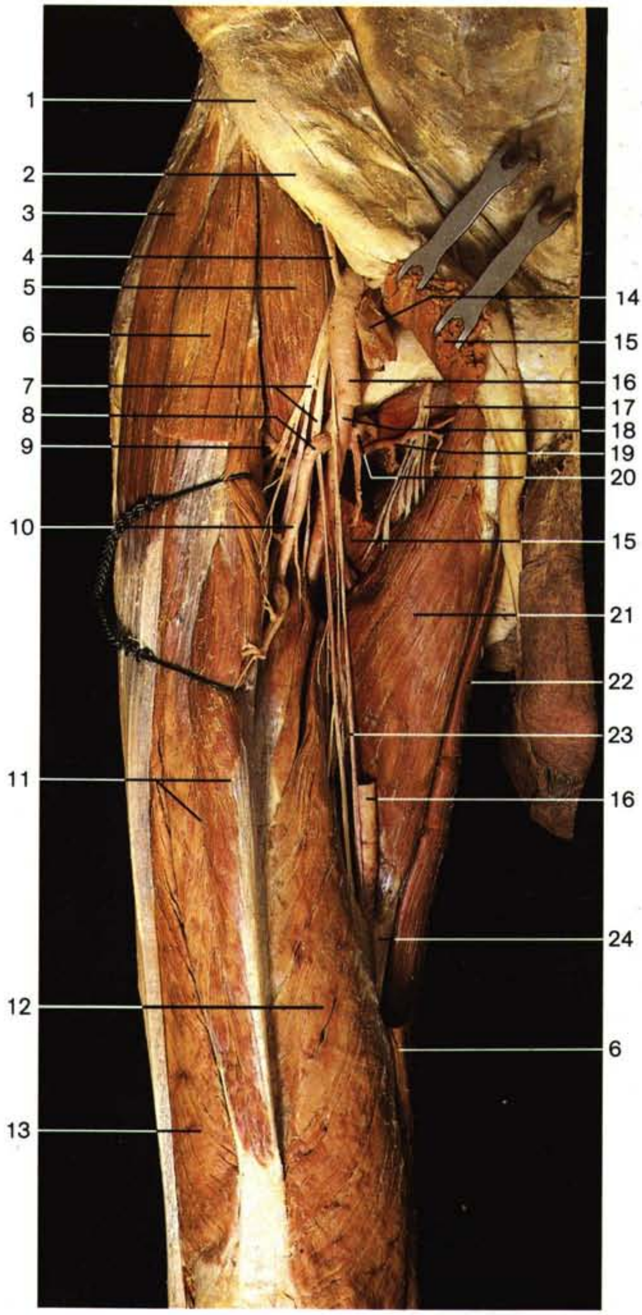

Anterior region of right thigh (anterior aspect). The fascia lata has been removed. Sartorius muscle, pectineus muscle and femoral artery have been cut to display the deep femoral artery with its branches. The rectus femoris muscle has been slightly reflected.

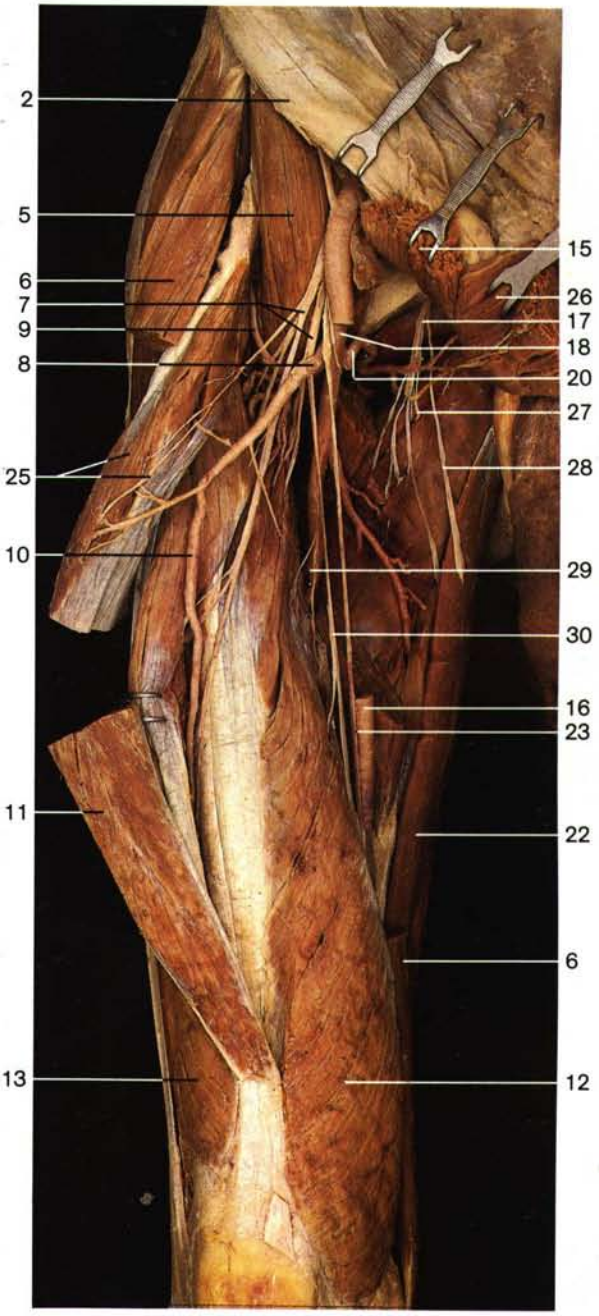

Anterior region of right thigh (anterior aspect). The sartorius, pectineus, adductor longus and rectus femoris muscles have been divided and reflected. The greater part of the femoral artery has been removed.

1 Anterior superior iliac spine
2 Inguinal ligament
3 Tensor fasciae latae muscle
4 Deep circumflex iliac artery
5 Iliopsoas muscle
6 Sartorius muscle (cut)
7 **Femoral nerve**
8 **Lateral circumflex femoral artery**
9 Ascending branch of lateral circumflex femoral artery
10 Descending branch of lateral circumflex femoral artery
11 Rectus femoris muscle
12 Vastus medialis muscle
13 Vastus lateralis muscle
14 Femoral vein
15 Pectineus muscle (cut)
16 Femoral artery (cut)
17 **Obturator nerve**
18 **Profunda femoris artery**
19 Ascending branch of medial circumflex femoral artery
20 **Medial circumflex femoral artery**
21 Adductor longus muscle
22 Gracilis muscle
23 Saphenous nerve
24 Distal part of vastoadductory lamina
25 Rectus femoris muscle with muscular branch of femoral nerve
26 Adductor longus muscle (divided)
27 Posterior branch of obturator nerve
28 Anterior branch of obturator nerve
29 Point at which perforating artery branches off from profunda femoris artery
30 Muscular branch to vastus medialis muscle

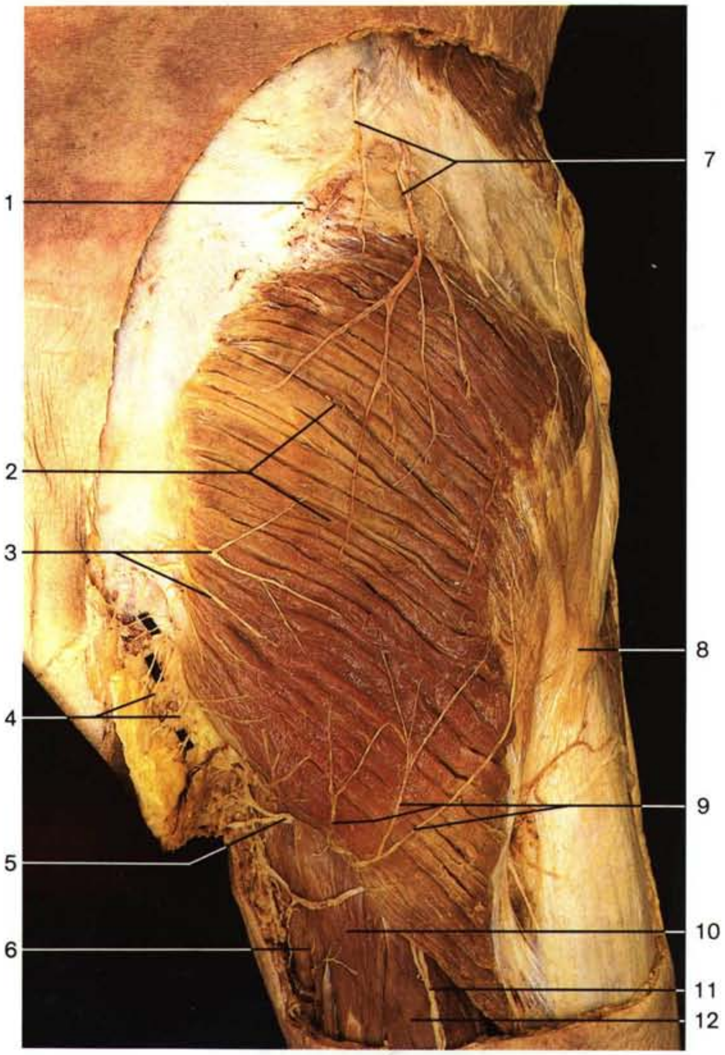

Gluteal region, right side (posterior aspect).

1 Iliac crest
2 Gluteus maximus muscle
3 **Middle cluneal nerves**
4 Anococcygeal nerves
5 Perineal branch of posterior femoral cutaneous nerve
6 Adductor magnus muscle
7 **Superior cluneal nerves**
8 Position of greater trochanter
9 **Inferior cluneal nerves**
10 Semitendinosus muscle
11 Posterior femoral cutaneous nerve
12 Long head of biceps femoris muscle

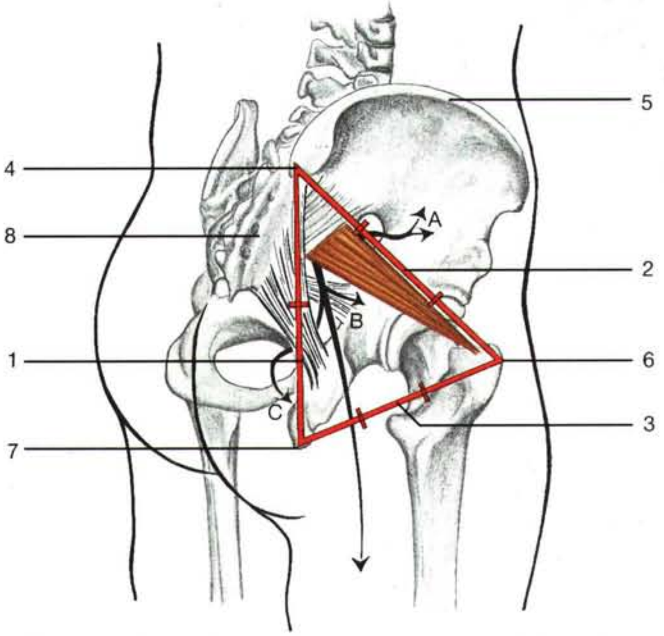

Gluteal region, right side (posterolateral aspect). Location of sciatic foramina in relation to the bones.
(Schematic drawing.)

A **Suprapiriform foramen**
of greater sciatic foramen
Superior gluteal artery, vein and nerve

B **Infrapiriform foramen**
of greater sciatic foramen
Sciatic nerve
Inferior gluteal artery, vein and nerve
Posterior femoral cutaneous nerve
Internal pudendal artery and vein
Pudendal nerve

C **Lesser sciatic foramen**
Pudendal nerve
Internal pudendal artery and vein

Red lines

1 **Spine-tuber line.**
In the middle of this line the infrapiriform foramen is situated
2 **Spine-trochanter line.**
In the upper third the suprapiriform foramen is located
3 **Tuber-trochanter line.**
Between the middle and posterior third, the ischiadic nerve can be found

Other structures

4 Posterior superior iliac spine
5 Iliac crest
6 Greater trochanter
7 Ischial tuberosity
8 Sacrum

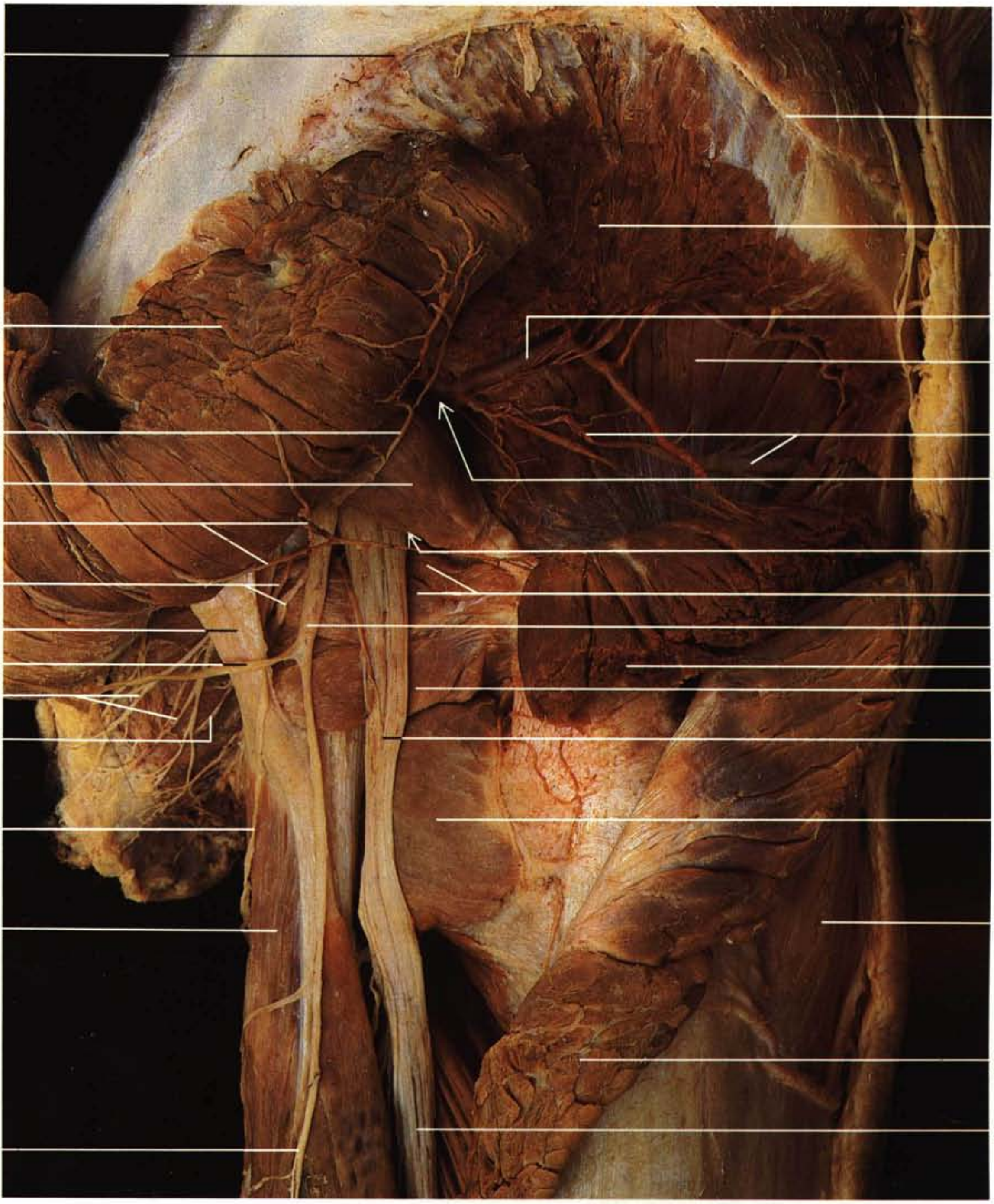

Gluteal region, right side (dorsal aspect). The gluteus maximus and gluteus medius muscles have been divided and reflected. Notice the position of the foramina above and below piriformis muscles and the lesser sciatic foramen.

1 Iliac crest
2 Gluteus maximus muscle (cut)
3 Inferior gluteal nerve
4 Piriformis muscle
5 Muscular branches of inferior gluteal artery
6 Pudendal nerve and internal pudendal artery within the **lesser sciatic foramen** (entrance to the pudendal canal)
7 Sacrotuberous ligament
8 Inferior cluneal nerve
9 Inferior rectal nerves
10 Inferior rectal arteries
11 Perforating cutaneous nerve
12 Long head of biceps femoris muscle

13 Gluteus medius muscle (cut)
14 Deep branch of superior gluteal artery
15 Gluteus minimus muscle
16 Superior gluteal nerve
17 **Suprapiriform foramen** } greater sciatic foramen
18 **Infrapiriform foramen** }
19 Tendon of obturator internus and superior gemellus muscles
20 Posterior femoral cutaneous nerve
21 Inferior gemellus muscle
22 **Sciatic nerve**
23 Quadratus femoris muscle
24 Tensor fasciae latae muscle

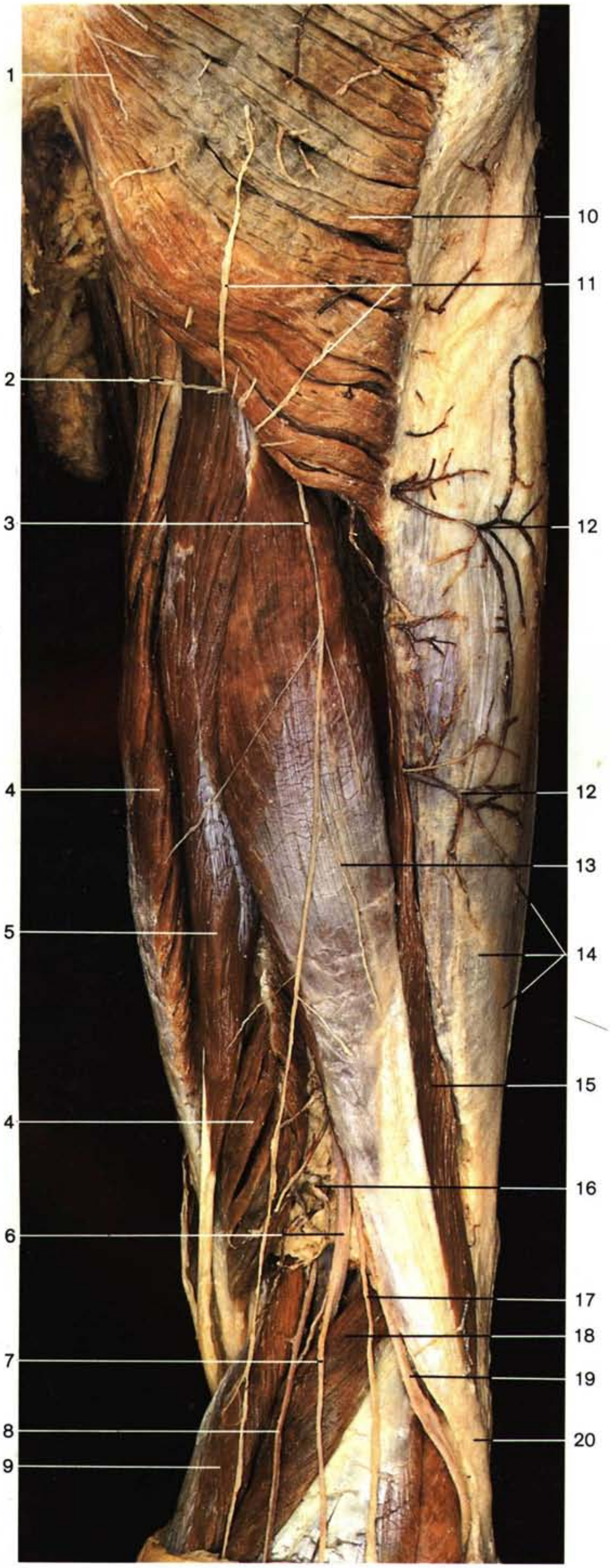

Cutaneous nerves of thigh (posterior aspect).
The fascia lata and the fasciae of muscles have been removed.

1 Middle cluneal nerves
2 Perineal branch of posterior femoral cutaneous nerve
3 **Posterior femoral cutaneous nerve**
4 Semimembranosus muscle
5 Semitendinosus muscle
6 **Tibial nerve**
7 Medial sural cutaneous nerve
8 **Small saphenous vein**
9 Medial head of gastrocnemius muscle
10 Gluteus maximus muscle
11 **Inferior cluneal nerve**
12 Cutaneous veins
13 Long head of biceps femoris muscle
14 Iliotibial tract
15 Short head of biceps femoris muscle
16 Popliteal fossa
17 Lateral sural cutaneous nerve
18 Lateral head of gastrocnemius muscle
19 **Common peroneal nerve**
20 Tendon of biceps femoris muscle
21 Inferior gluteal nerve
22 Sacrotuberous ligament
23 Inferior rectal branches of pudendal nerve
24 Anus
25 Gluteus medius muscle
26 Piriformis muscle
27 **Sciatic nerve**
28 Inferior gluteal artery
29 Gluteus maximus muscle (cut)
30 Quadratus femoris muscle
31 Sciatic nerve dividing into its two branches: the common peroneal nerve and the tibial nerve
32 Muscular branches of sciatic nerve to hamstring muscles
33 **Popliteal artery**
34 **Popliteal vein**
35 Small saphenous vein (cut)
36 Long head of biceps femoris muscle (cut)
37 **Superficial peroneal nerve**

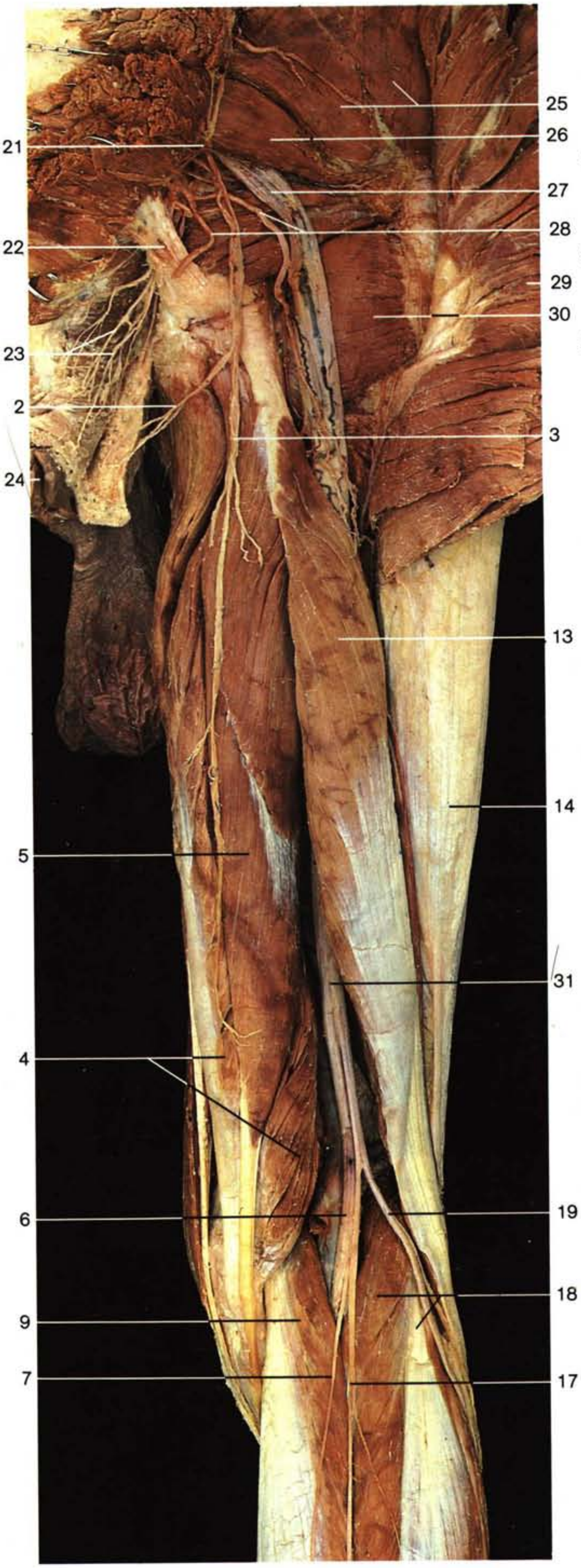
25
21
26
27
28
22
29
30
23
2
3
24
13
14
5
31
4
6
19
18
9
7
17

Posterior femoral region and gluteal region, right side (posterior aspect). The gluteus maximus muscle has been divided and reflected.

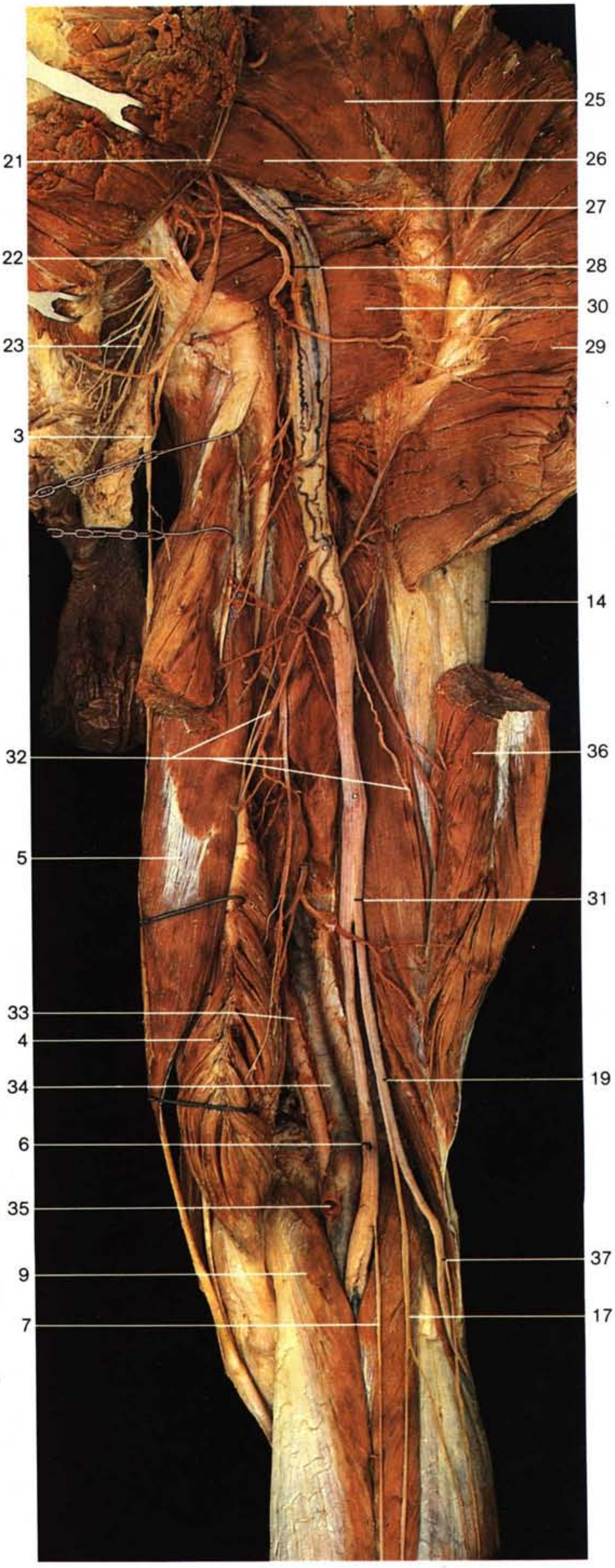
25
21
26
27
22
28
30
23
29
3
14
32
36
5
31
33
4
19
34
6
35
37
9
7
17

Posterior femoral region and gluteal region, right side (posterior aspect). The gluteus maximus muscle and the long head of biceps femoris muscle have been divided and reflected.

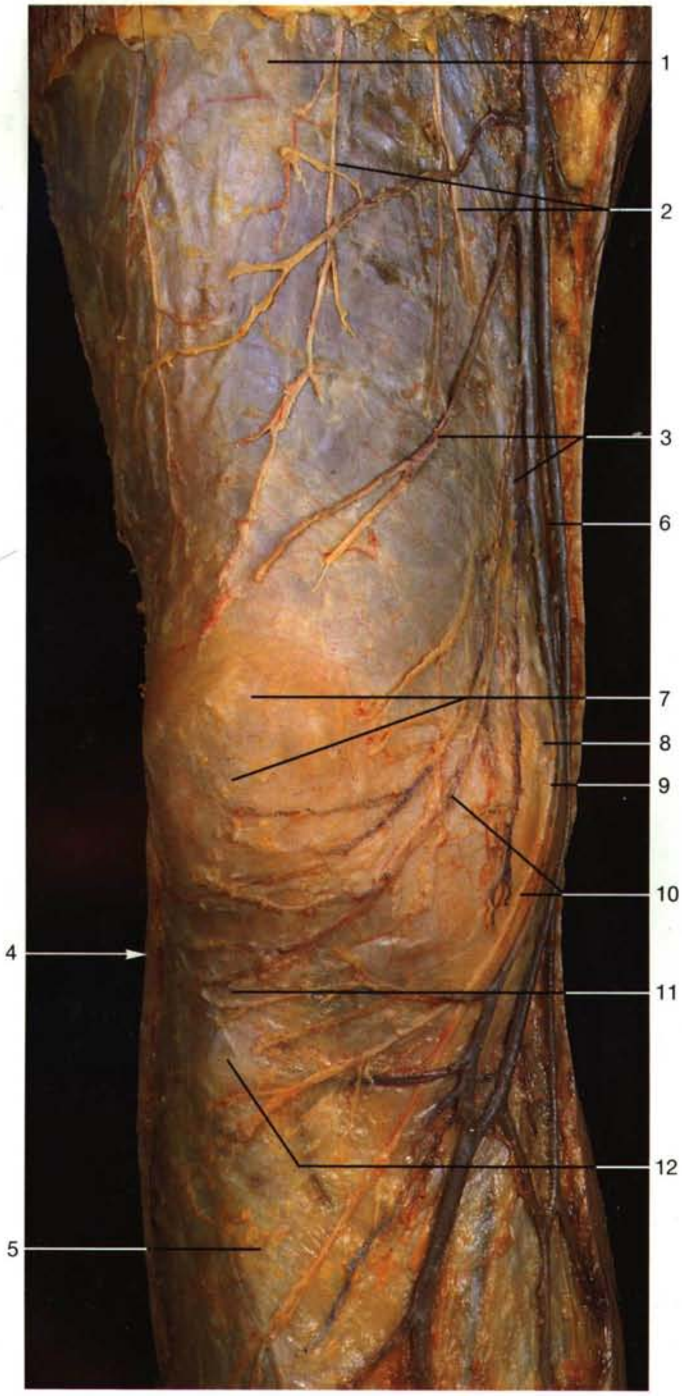

Anterior region of right knee; cutaneous nerves and veins (anterior aspect).

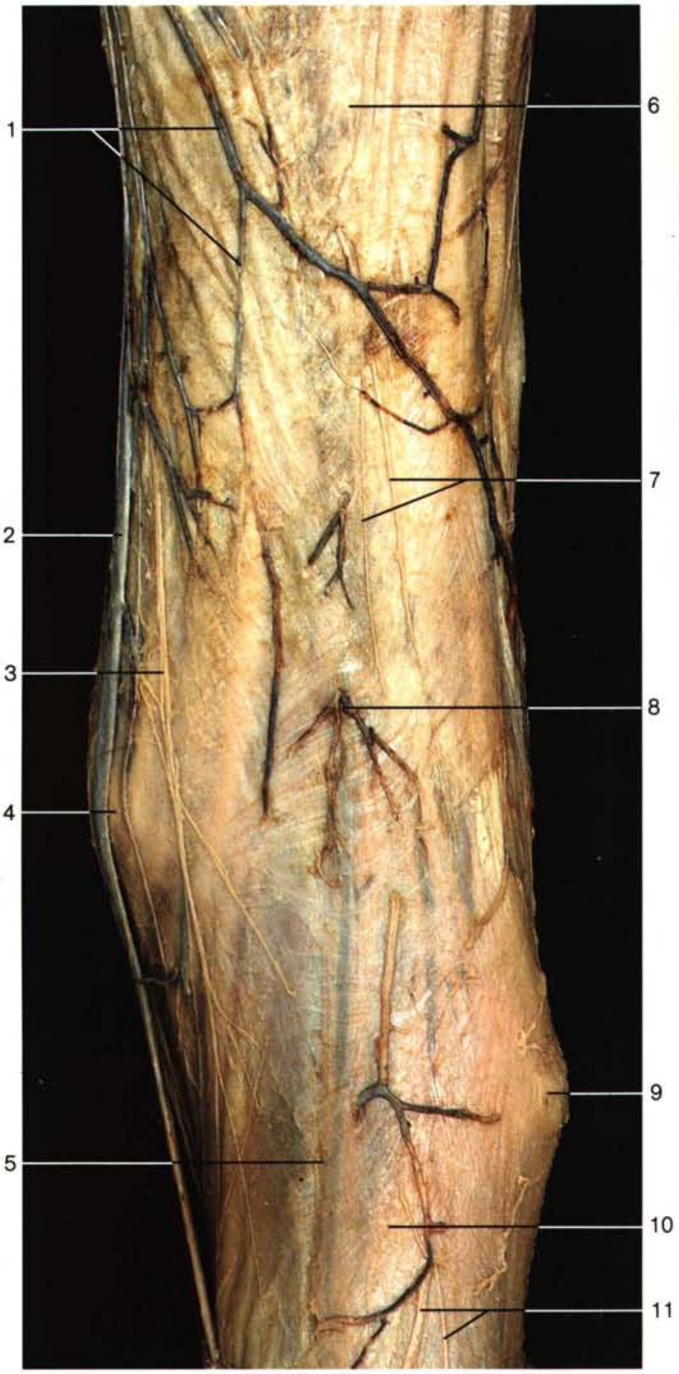

Posterior region of right kneee; cutaneous nerves and veins (posterior aspect).

1 Fascia lata
2 Terminal branches of anterior cutaneous branches of femoral nerve
3 Venous network around knee
4 Position of head of fibula
5 Superficial crural fascia
6 **Great saphenous vein**
7 Patella
8 Position of medial epicondyle of femur
9 Saphenous nerve
10 **Infrapatellar branches of saphenous nerve**
11 Patellar ligament
12 Position of tuberosity of tibia

1 Cutaneous veins (tributaries of great saphenous vein)
2 **Great saphenous vein**
3 Cutaneous branch of femoral nerve
4 Position of medial epicondyle of femur
5 Position of small saphenous vein
6 Fascia lata
7 Terminal branches of **posterior femoral cutaneous nerve**
8 Cutaneous veins of popliteal fossa
9 Position of head of fibula
10 Superficial layer of fascia cruris
11 Lateral sural cutaneous nerve

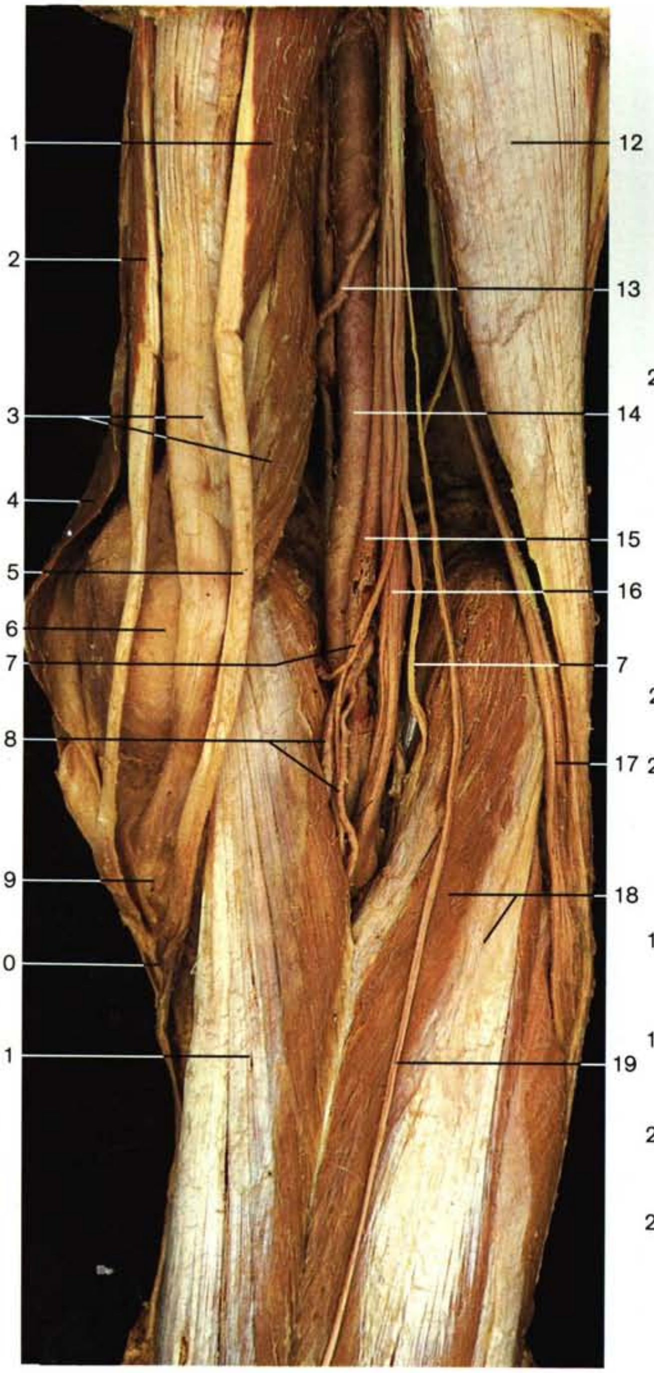

Right leg, posterior crural region (posterior aspect). The gastrocnemius muscle has been divided and reflected.

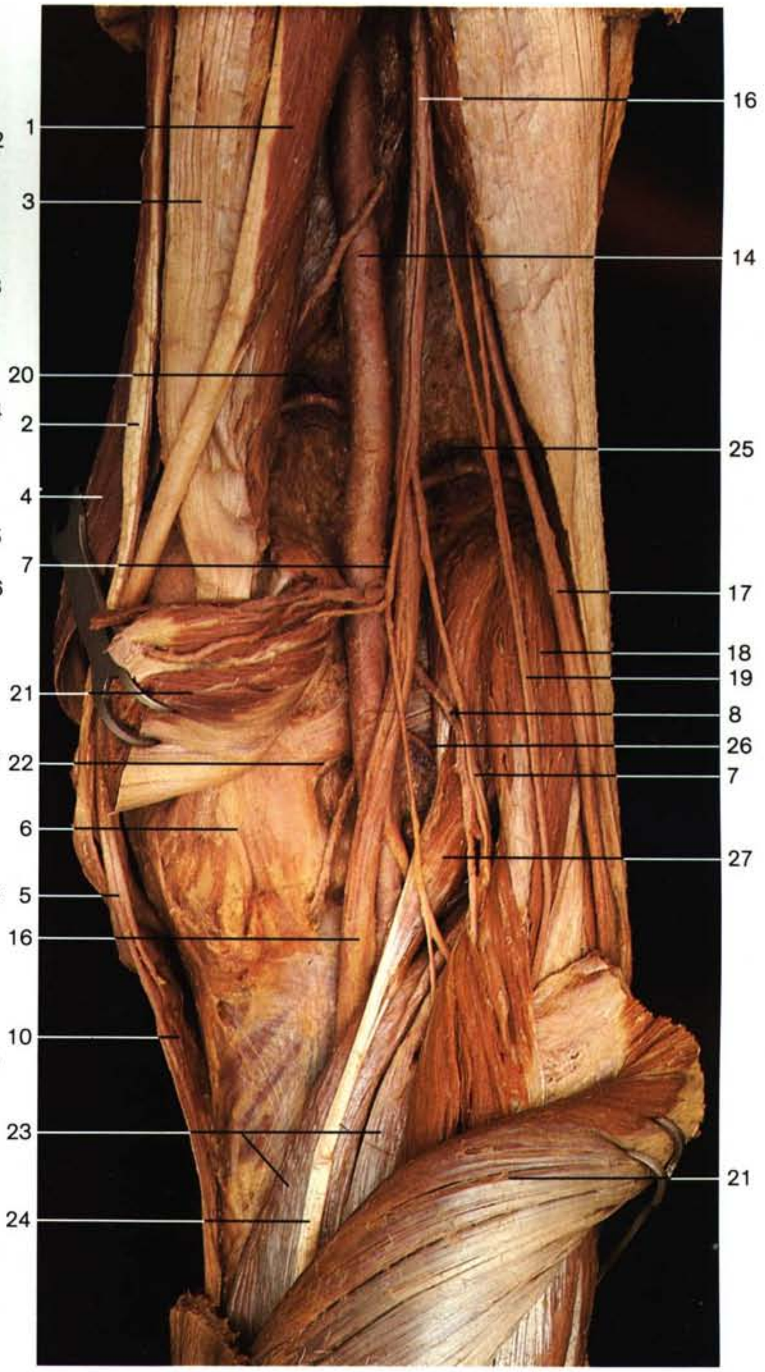

Right leg, posterior crural region, deep layer (posterior aspect). The gastrocnemius and the soleus muscles have been divided and reflected.

1 Semitendinosus muscle
2 Gracilis muscle
3 Semimembranosus muscle
4 Sartorius muscle
5 Tendon of semitendinosus muscle
6 Position of medial condyle of femur
7 Muscular branches of tibial nerve
8 Sural arteries and veins
9 Tendon of semimembranosus muscle
10 Common tendon of gracilis, semitendinosus and sartorius muscles
11 Medial head of gastrocnemius muscle
12 Biceps femoris muscle
13 Muscular branch of popliteal artery
14 **Popliteal artery**
15 **Popliteal vein**
16 **Tibial nerve**
17 **Common peroneal nerve**
18 Lateral head of gastrocnemius muscle
19 Medial sural cutaneous nerve
20 Medial superior genicular artery
21 Medial head of gastrocnemius muscle (cut and reflected)
22 Medial inferior genicular artery
23 Soleus muscle
24 Tendon of plantaris muscle
25 Lateral superior genicular artery
26 Lateral inferior genicular artery
27 Plantaris muscle

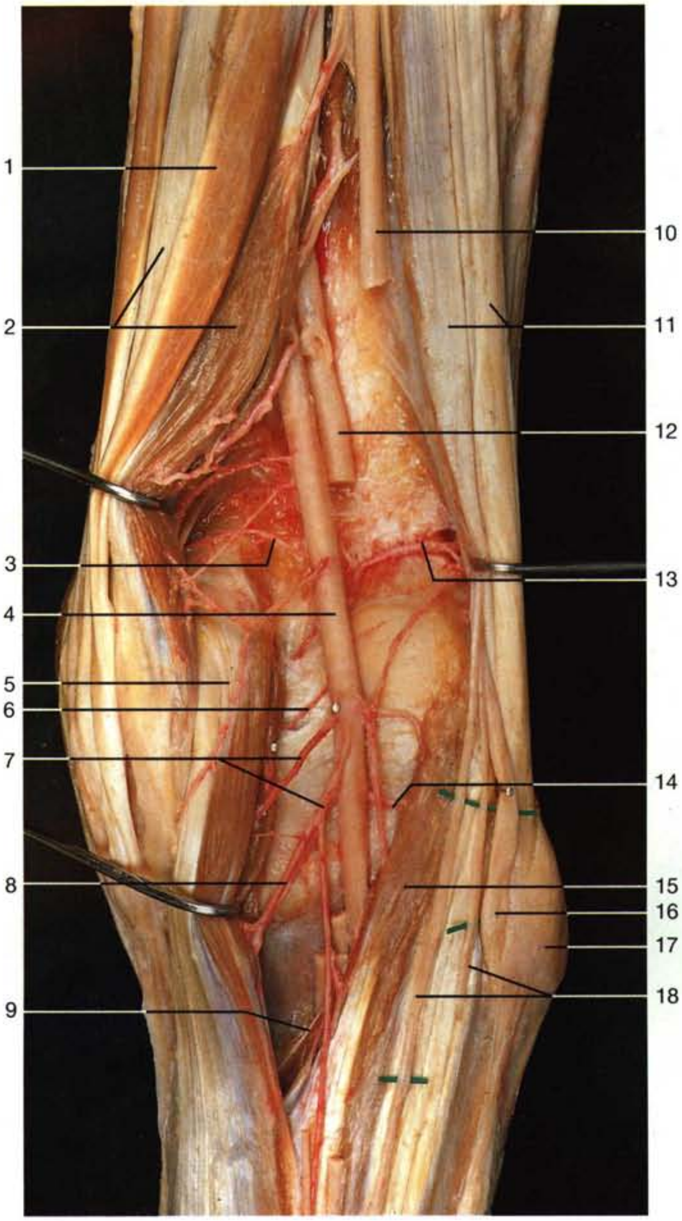

Right leg, popliteal fossa, deep layer (posterior aspect). The muscles have been reflected to display the genicular arteries.

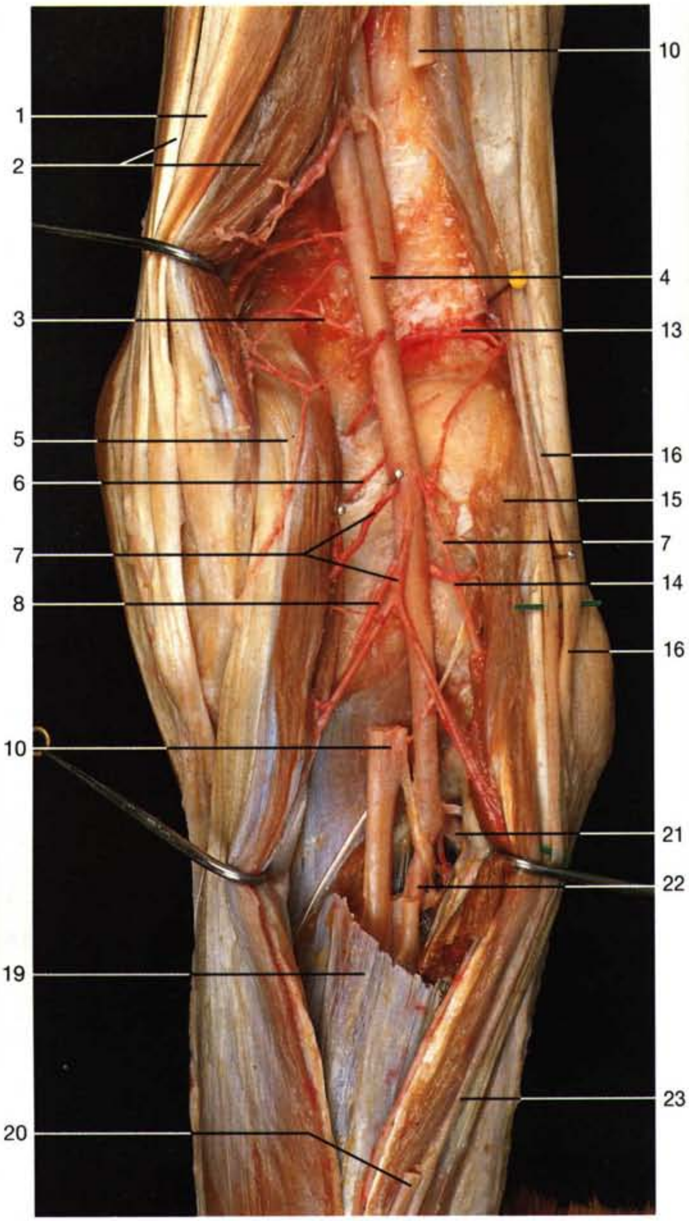

Right leg, popliteal fossa, deepest layer (posterior aspect). Tibial nerve and popliteal vein have been partly removed and a portion of the soleus muscle was cut away to display the anterior tibial artery.

1 Semitendinosus muscle
2 Semimembranosus muscle
3 Medial superior genicular artery
4 **Popliteal artery**
5 Medial head of gastrocnemius muscle
6 Middle genicular artery
7 Muscular branches
8 Medial inferior genicular artery
9 Tendon of plantaris muscle
10 **Tibial nerve** (cut)
11 Biceps femoris muscle

12 Popliteal vein (cut)
13 Lateral superior genicular artery
14 Lateral inferior genicular artery
15 Lateral head of gastrocnemius muscle
16 **Common peroneal nerve**
17 Head of fibula
18 Lateral sural cutaneous nerves
19 Soleus muscle
20 Medial sural cutaneous nerve
21 **Anterior tibial artery**
22 **Posterior tibial artery**
23 Lateral sural cutaneous nerve

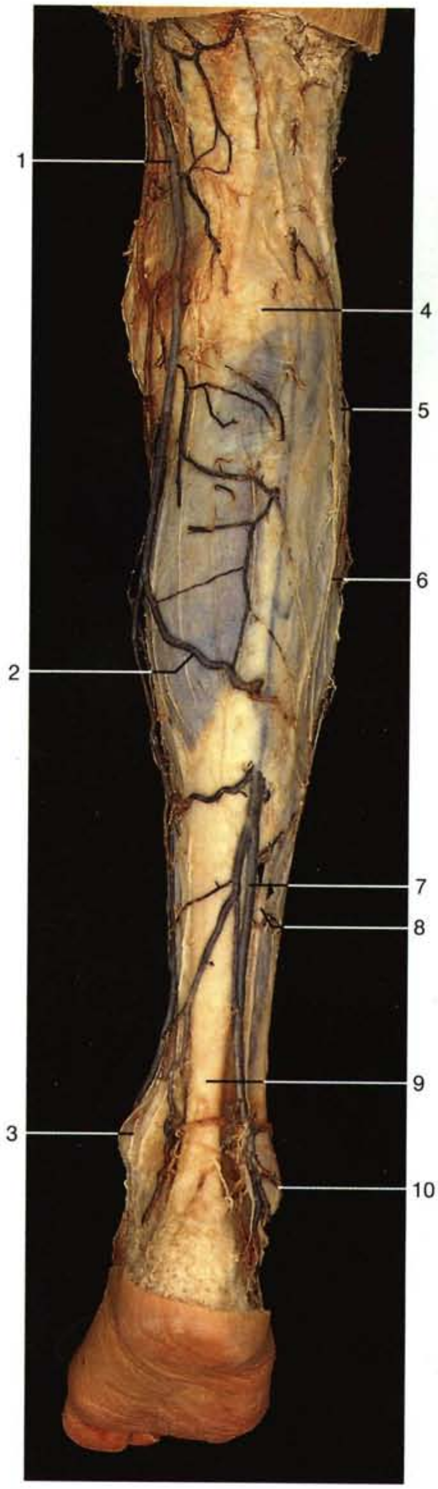

Right leg. Cutaneous veins and nerves (posterior aspect).

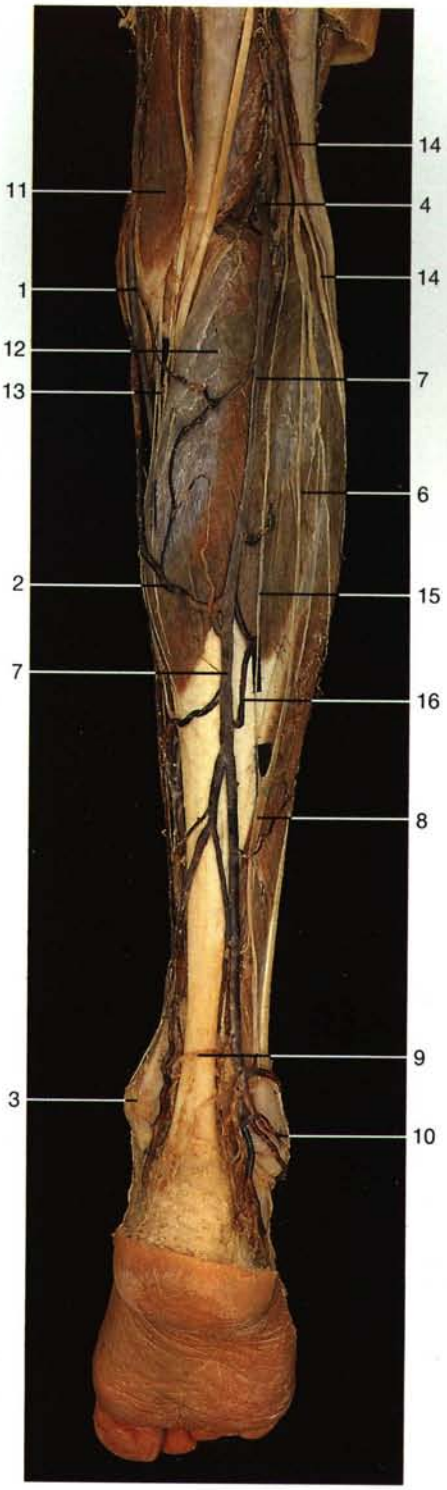

Right leg. Cutaneous nerves and veins (posterior aspect). The superficial layer of the crural fascia has been removed.

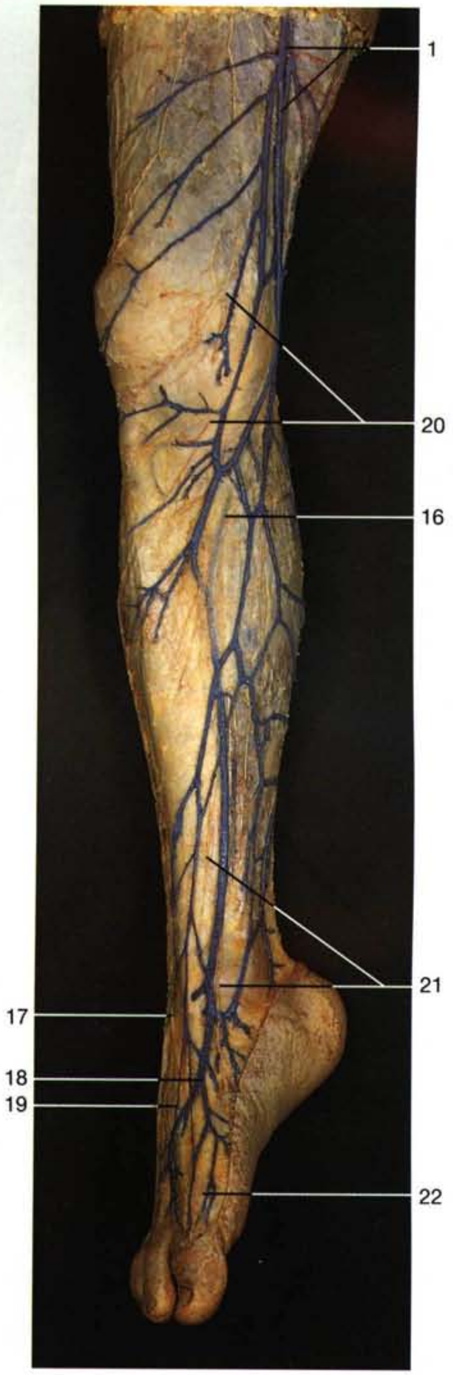

Cutaneous veins and nerves of the right leg (anterior-medial aspect; veins are colored).

1 **Great saphenous vein**
2 Venous anastomosis between small and great saphenous veins
3 Medial malleolus
4 Popliteal fossa
5 Position of head of fibula
6 Lateral sural cutaneous nerve
7 **Small saphenous vein**
8 **Sural nerve**
9 Calcaneal tendon
10 Lateral malleolus
11 Semitendinosus muscle
12 Medial head of gastrocnemius muscle
13 Saphenous nerve
14 **Common peroneal nerve**
15 Medial sural cutaneous nerve
16 Perforating veins
17 Superficial peroneal nerve
18 Dorsal venous arch
19 Intermediate dorsal cutaneous nerve
20 **Infrapatellar branches of saphenous nerve**
21 Terminal branches of saphenous nerve
22 Medial dorsal cutaneous nerve

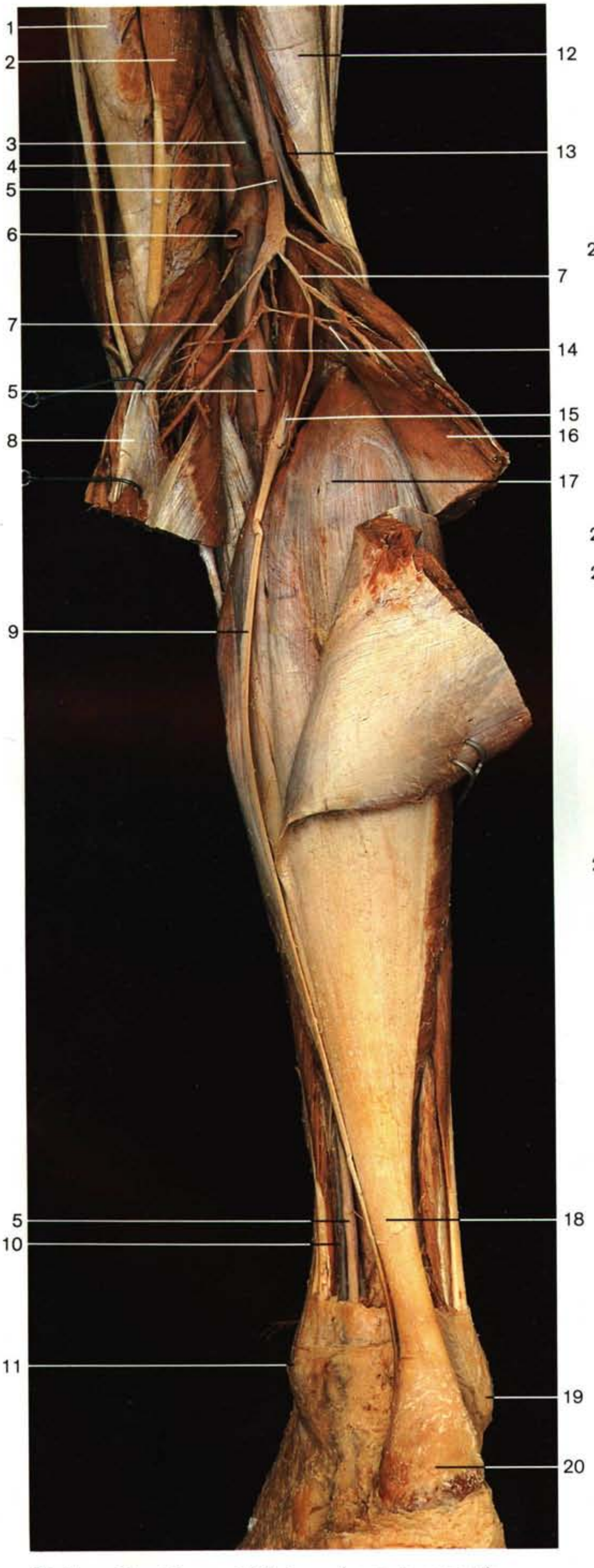
1
2
3
4
5
6
7
5
8
9
5
10
11
12
13
7
14
15
16
17
18
19
20

Right popliteal fossa, middle layer (posterior aspect). The cutaneous veins and nerves have been removed.

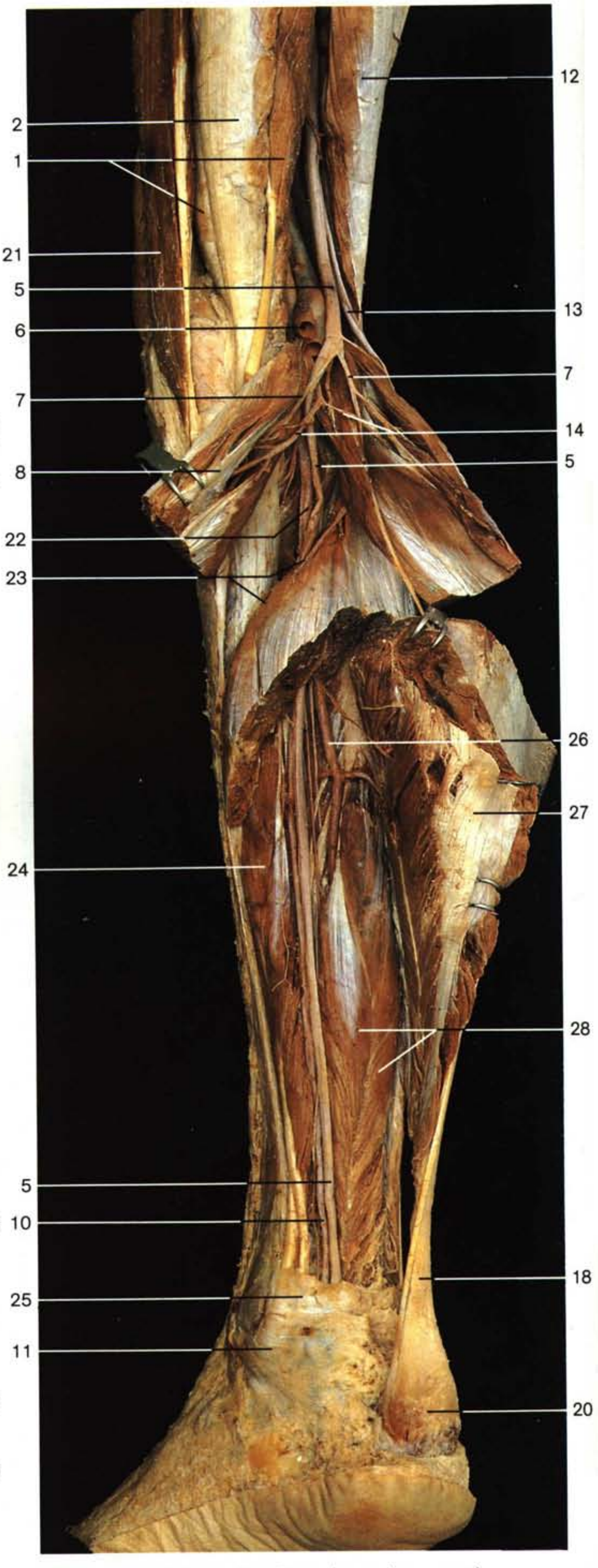
2
1
21
5
6
7
8
22
23
24
5
10
25
11
12
13
7
14
5
26
27
28
18
20

Right popliteal fossa, deep layer (posterior aspect). The medial head of gastrocnemius muscle has been divided and reflected.

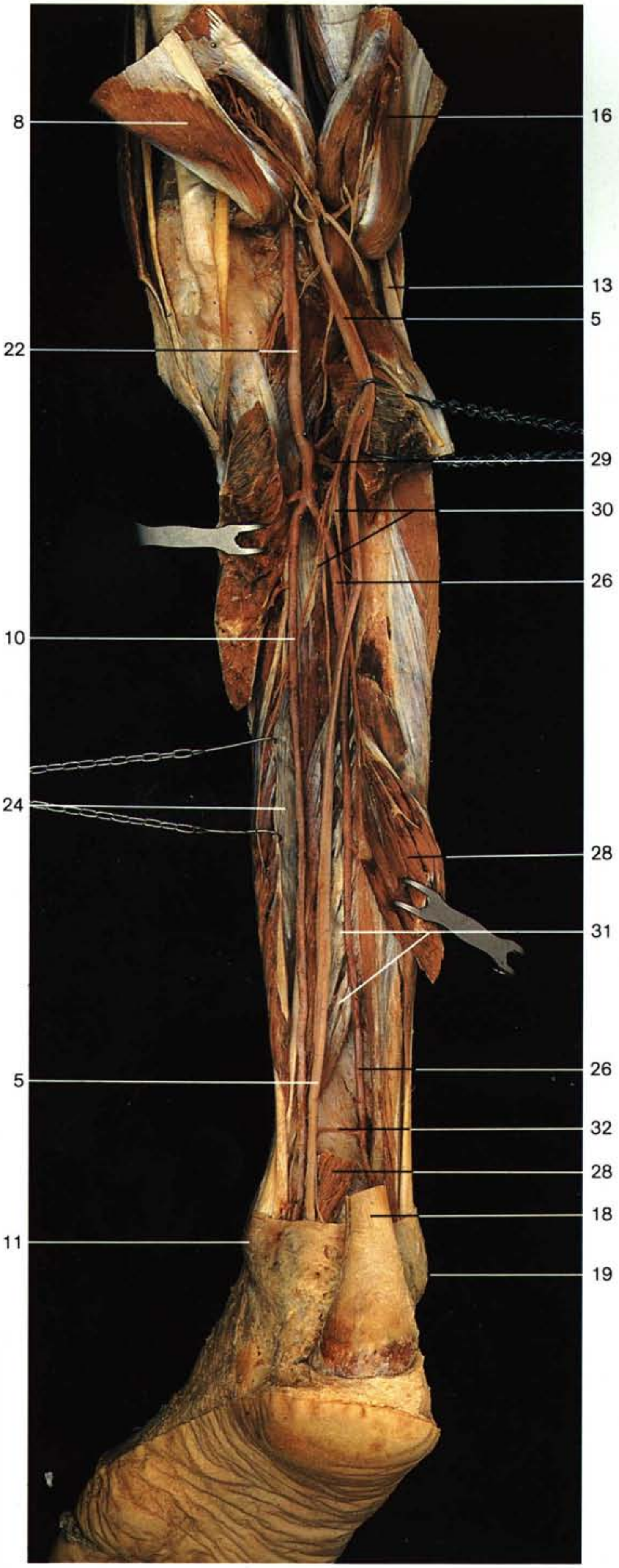

Right leg, posterior crural region, deepest layer (posterior aspect). Triceps surae (gastrocnemius and soleus) and flexor hallucis longus muscles have been cut and reflected.

1 Semimembranosus muscle
2 Semitendinosus muscle
3 **Popliteal vein**
4 Popliteal artery
5 **Tibial nerve**
6 Small saphenous vein (cut)
7 Muscular branch of tibial nerve
8 Medial head of gastrocnemius muscle
9 Tendon of plantaris muscle
10 **Posterior tibial artery**
11 Medial malleolus
12 Biceps femoris muscle
13 Common peroneal nerve
14 **Sural arteries**
15 Plantaris muscle
16 Lateral head of gastrocnemius muscle
17 Soleus muscle
18 Calcaneal tendon
19 Lateral malleolus
20 Calcaneal tuberosity
21 Sartorius muscle
22 **Popliteal artery**
23 Tendinous arch of soleus muscle
24 Flexor digitorum longus muscle
25 Flexor retinaculum
26 **Peroneal artery**
27 Soleus muscle
28 Flexor hallucis longus muscle
29 **Anterior tibial artery**
30 Muscular branches of tibial nerve
31 Tibialis posterior muscle
32 Communicating branch of peroneal artery
33 Tendon of tibialis anterior muscle
34 Tibia
35 Tendon of extensor hallucis longus muscle
36 Tendons of extensor digitorum longus muscle
37 Anterior tibialis artery
38 Fibula
39 Tendons of peroneus longus and brevis muscles

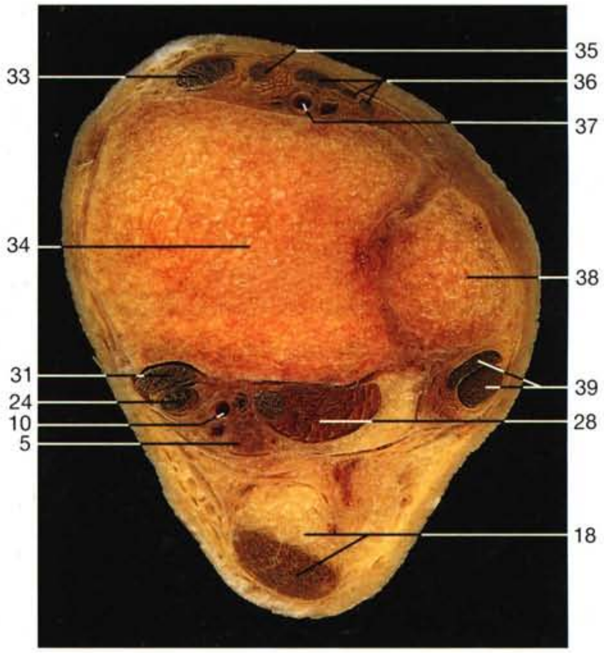

Cross-section of the leg, superior to the malleoli (from below).

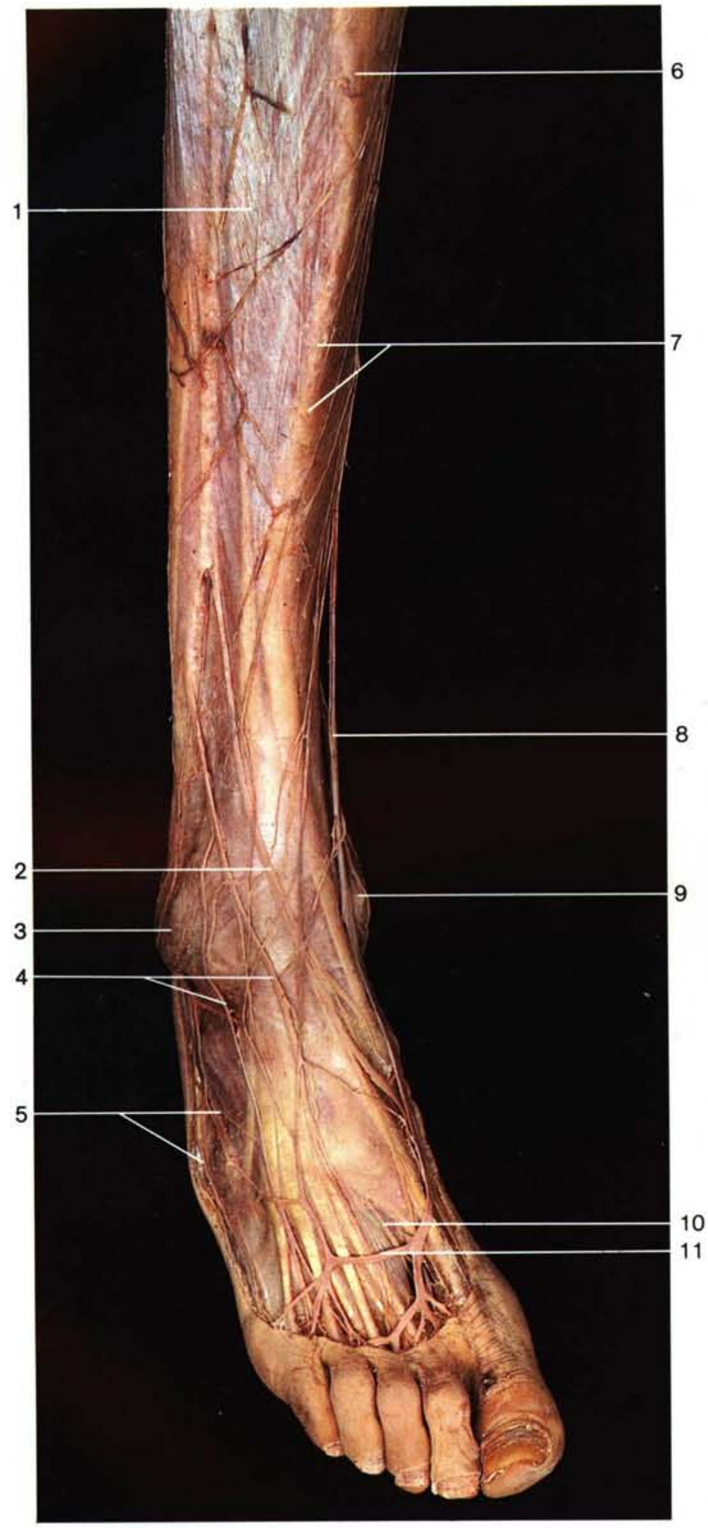

Right leg and foot, anterior crural region and dorsum of foot; cutaneous nerves and veins (anterior aspect).

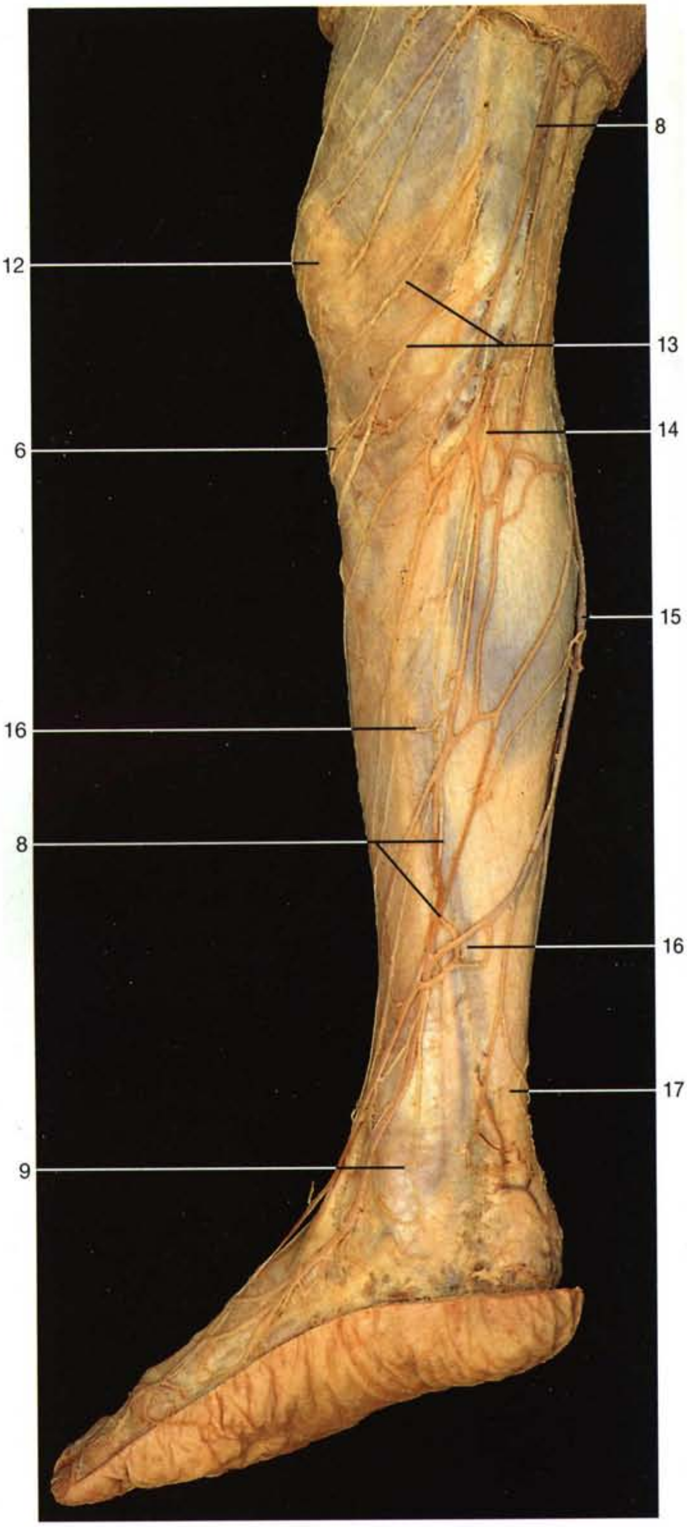

Right leg and foot; cutaneous nerves and veins (medial aspect).

1 Superficial crural fascia
2 Medial cutaneous branch of superficial peroneal nerve
3 Lateral malleolus
4 Lateral cutaneous branch of superficial peroneal nerve
5 Cutaneous branch of sural nerve
6 Position of tuberosity of tibia
7 Anterior margin of tibia
8 **Great saphenous vein**
9 Medial malleolus
10 Deep peroneal nerve
11 **Venous arch of dorsum of foot**
12 Position of patella
13 **Infrapatellar branches of saphenous nerve**
14 Saphenous nerve
15 Small saphenous vein
16 Perforating vein
17 Calcaneal tendon

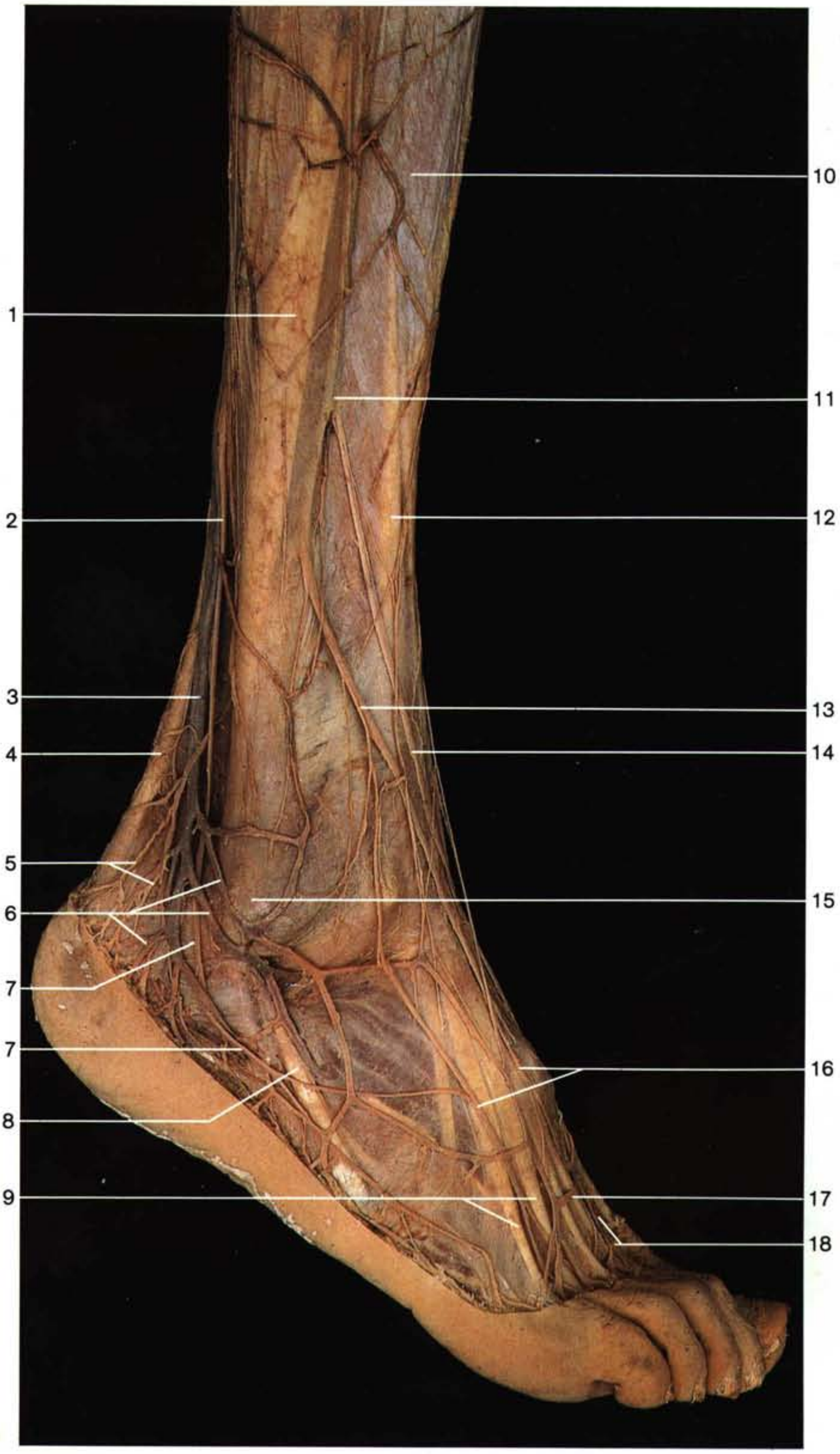

Right leg and dorsum of foot; cutaneous nerves and veins (lateral aspect).

1 Position of fibula
2 **Sural nerve**
3 **Small saphenous vein**
4 Calcaneal tendon
5 Lateral calcaneal branches of sural nerve
6 Venous network at lateral malleolus
7 Cutaneous branch of sural nerve
8 Tendon of peroneus brevis muscle
9 Tendons of extensor digitorum longus muscle

10 Fascia cruris
11 Superficial peroneal nerve
12 Position of tibia
13 Lateral cutaneous branch } of **superficial peroneal nerve**
14 Medial cutaneous branch }
15 Lateral malleolus
16 Dorsal digital nerves
17 Dorsal venous arch
18 Deep peroneal nerve

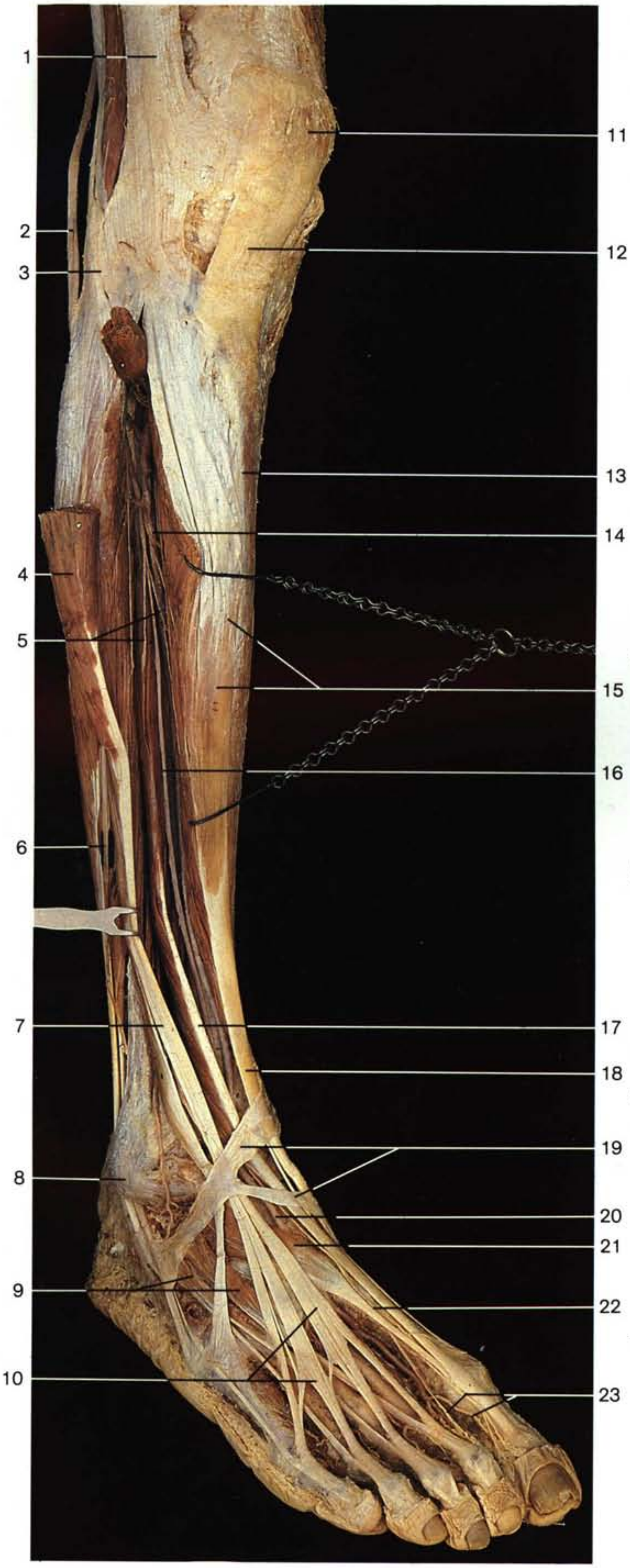
1
11
2
3
12
13
14
4
5
15
16
6
7
17
18
19
8
20
21
9
22
10
23

Right leg and dorsum of foot, middle layer (anterior lateral aspect). The extensor digitorum longus muscle has been divided and reflected laterally.

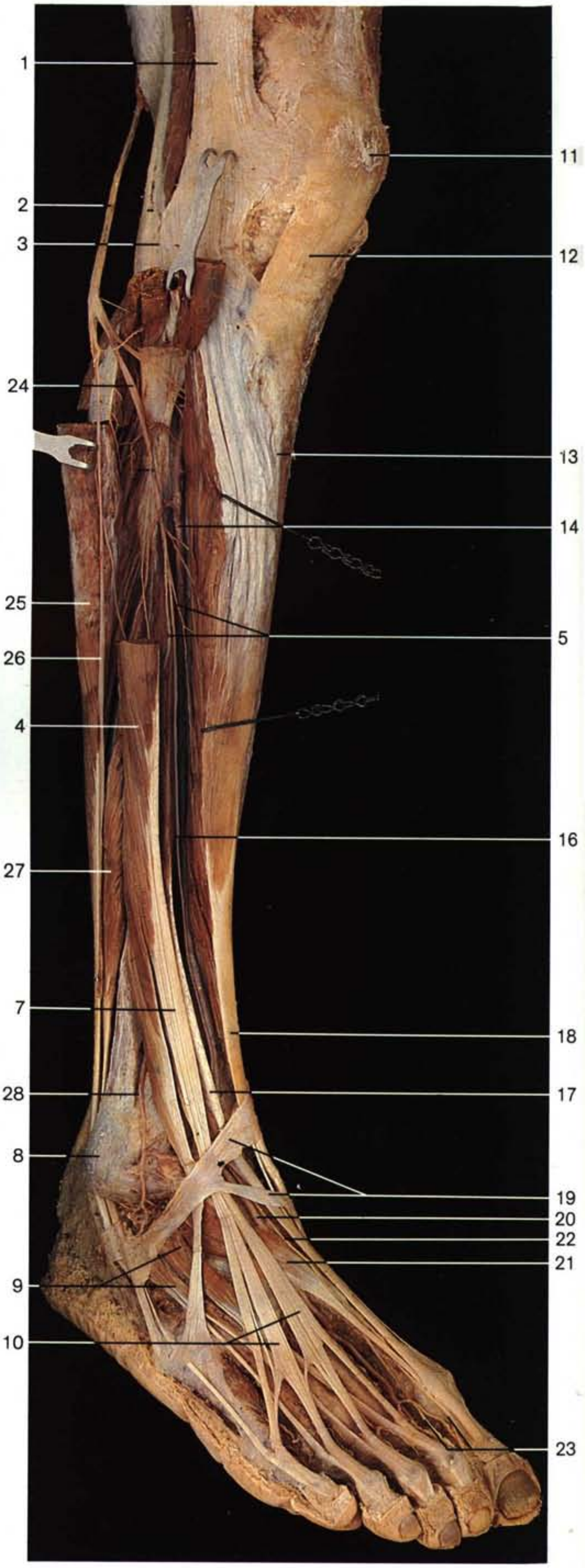
1
11
2
3
12
24
13
14
25
5
26
4
16
27
7
18
28
17
8
19
20
22
9
21
10
23

Right leg, deep layer (anterior lateral aspect). The extensor digitorum longus and the peroneus longus muscle have been divided or removed. The common peroneal nerve has been elevated to show its course around the head of fibula.

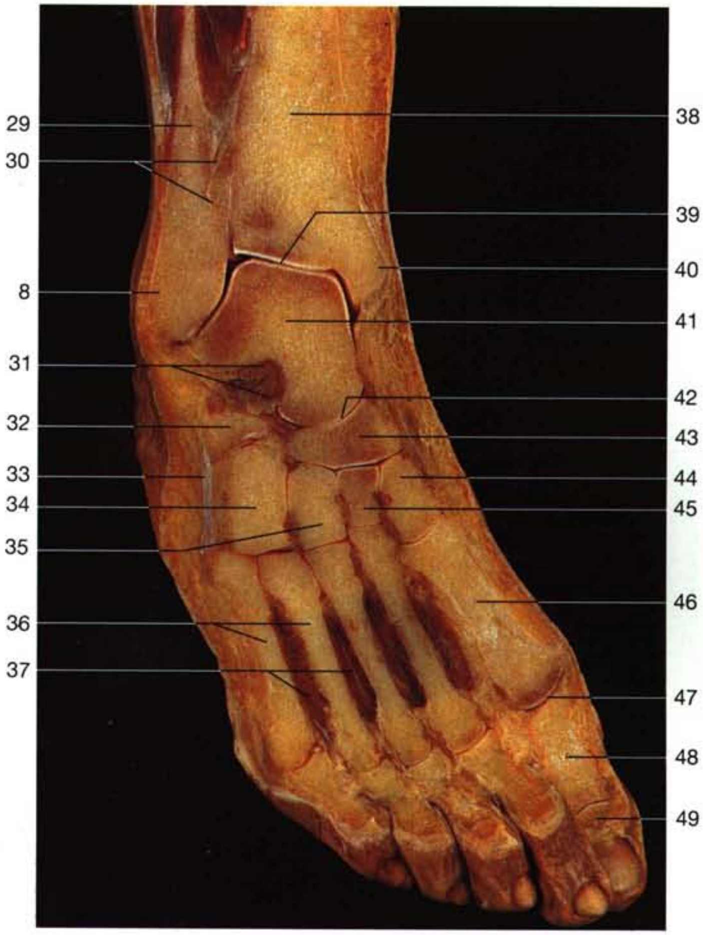

Coronal section through the foot and talocrural joint (anterior aspect).

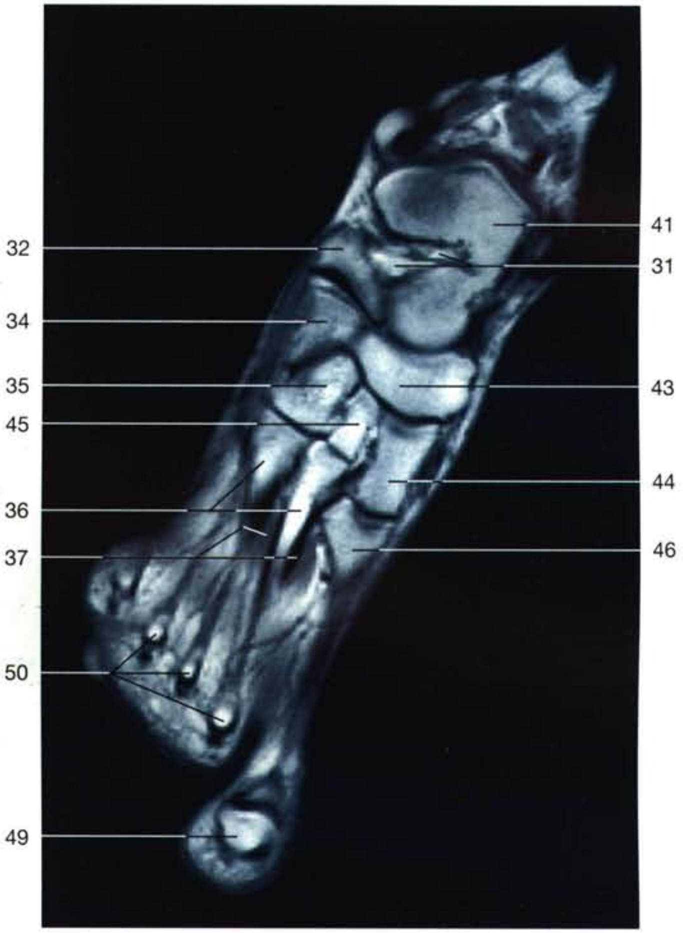

Coronal section through the foot (MR-Scan after A. Heuck, G. Luttke and J. W. Rohen, 1994).

1 Iliotibial tract
2 **Common peroneal nerve**
3 Position of head of fibula
4 Extensor digitorum longus muscle
5 Muscular branches of deep peroneal nerve
6 **Superficial peroneal nerve**
7 Tendon of extensor digitorum longus muscle
8 Lateral malleolus
9 Extensor digitorum brevis muscle
10 Tendons of extensor digitorum longus muscle
11 Patella
12 Patellar ligament
13 Anterior margin of tibia
14 **Anterior tibial artery**
15 Tibialis anterior muscle
16 Deep peroneal nerve
17 Extensor hallucis longus muscle
18 Tendon of tibialis anterior muscle
19 Extensor retinaculum
20 Dorsalis pedis artery
21 Extensor hallucis brevis muscle
22 Deep peroneal nerve (on dorsum of foot)
23 Dorsal digital nerves (terminal branches of deep peroneal nerve)
24 **Deep peroneal nerve**
25 Peroneus longus muscle (cut)
26 Superficial peroneal nerve (with peroneal muscles laterally reflected)
27 Peroneus brevis muscle
28 Lateral anterior malleolar artery
29 Fibula
30 Distal tibiofibular joint (syndesmosis)
31 Interosseous talocalcaneal ligament
32 Calcaneus
33 Tendon of peroneus brevis muscle
34 Cuboid bone
35 Lateral cuneiform bone
36 Metatarsal bones
37 Dorsal interosseous muscles
38 Tibia
39 **Talocrural joint**
40 Medial malleolus
41 Talus
42 **Talocalcaneonavicular joint**
43 Navicular bone
44 Medial cuneiform bone
45 Intermediate cuneiform bone
46 First metatarsal bone
47 Metatarsophalangeal joint of great toe
48 Proximal phalanx of great toe
49 Distal phalanx of great toe
50 Heads of metatarsal bones II-IV

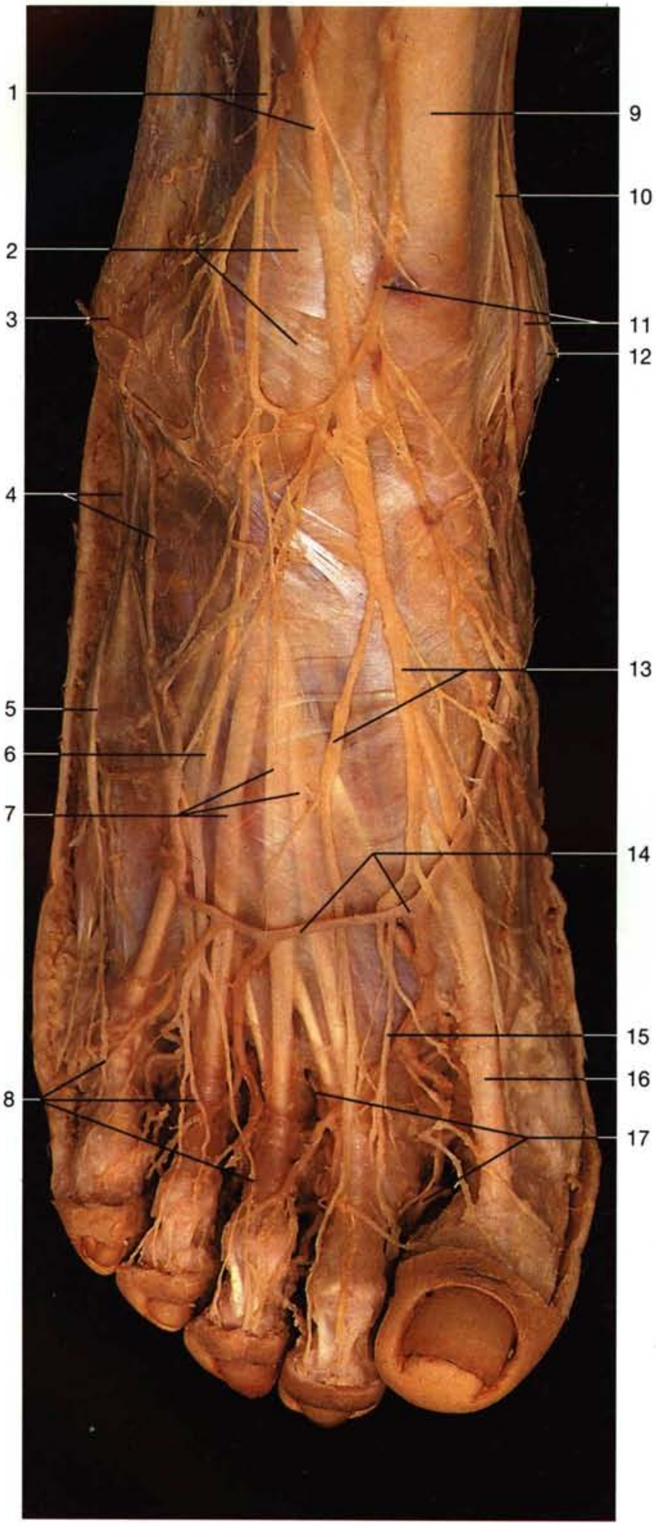

Dorsum of the right foot, superficial layer (anterior aspect).

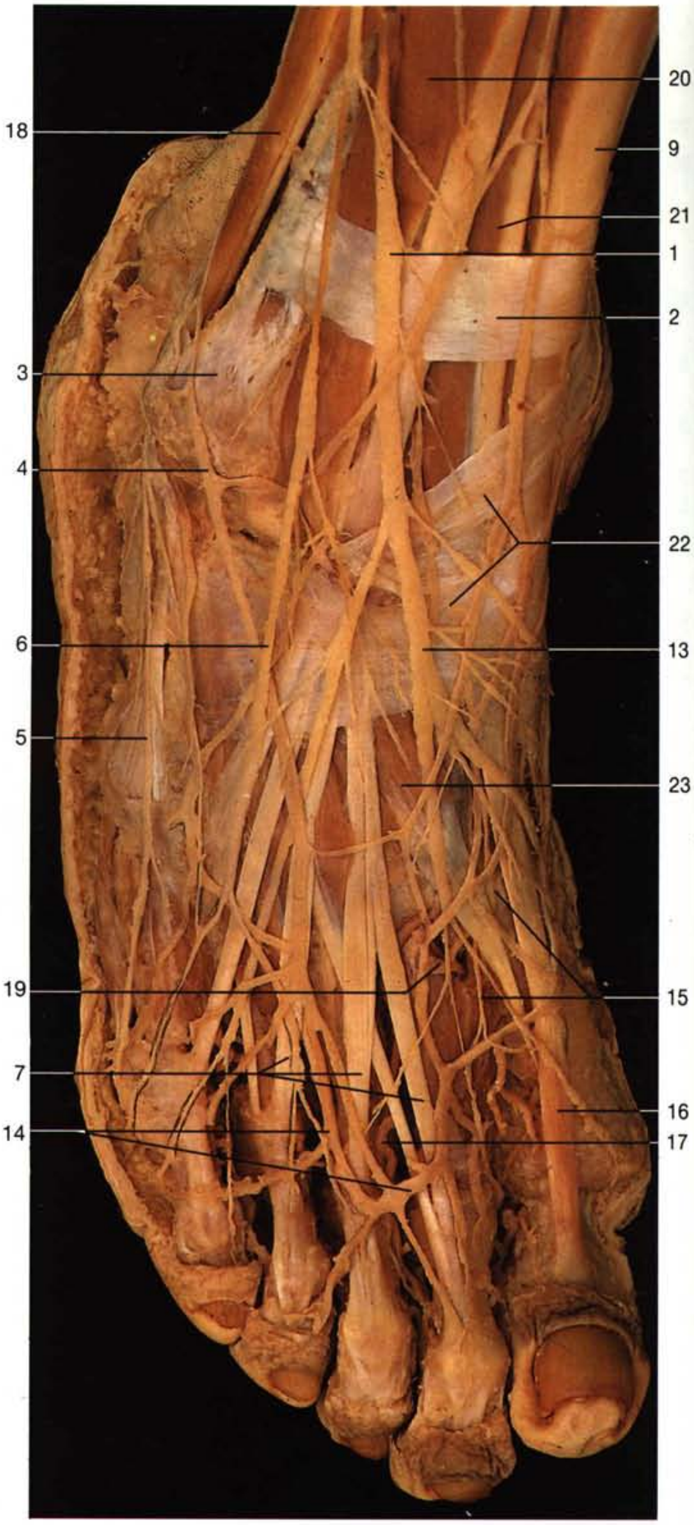

Dorsum of the right foot, superficial layer. The fascia of the dorsum has been removed.

1 **Superficial peroneal nerve**
2 Superior extensor retinaculum
3 Lateral malleolus
4 Venous network of lateral malleolus and tributaries of small saphenous vein
5 Lateral dorsal cutaneous nerve (branch of sural nerve)
6 Intermediate dorsal cutaneous nerve
7 Tendons of extensor digitorum longus muscle
8 Dorsal digital nerves
9 Tendon of tibialis anterior muscle
10 **Saphenous nerve**
11 Venous network of medial malleolus and tributaries of great saphenous vein
12 Medial malleolus
13 Medial dorsal cutaneous nerves
14 **Dorsal venous arch**
15 Dorsal digital nerve (of deep peroneal nerve)
16 Tendon of extensor hallucis longus muscle
17 Dorsal digital arteries
18 Peroneal muscles
19 Deep plantar branch of dorsalis pedis artery anastomosing with plantar arch
20 Extensor digitorum longus muscle
21 Extensor hallucis longus muscle
22 Inferior extensor retinaculum
23 Extensor hallucis brevis muscle

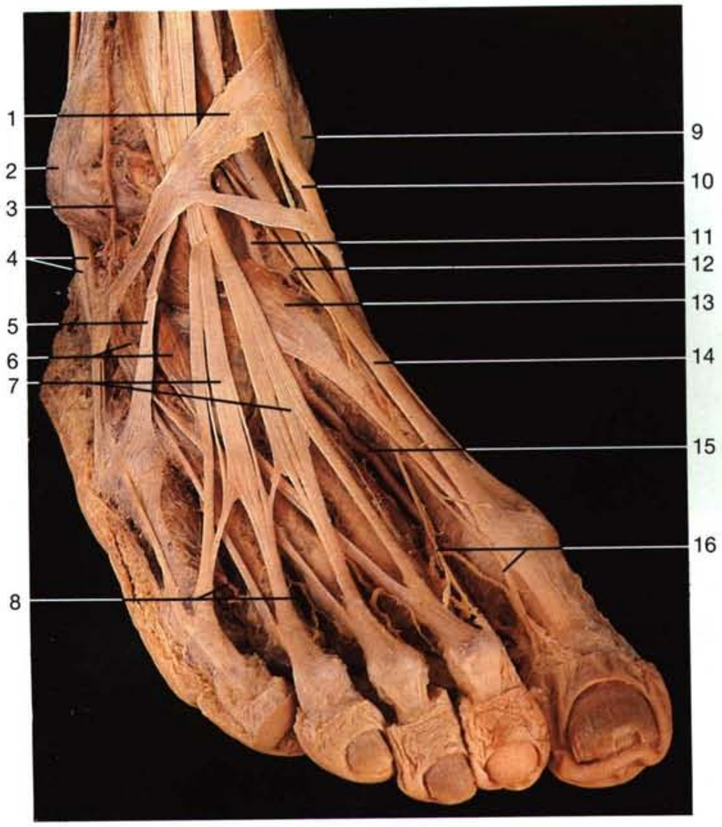

Dorsum of right foot, middle layer (anterior lateral aspect). The cutaneous nerves have been removed.

1 Extensor retinaculum
2 Lateral malleolus
3 Lateral anterior malleolar artery
4 Tendons of peroneal muscles
5 Tendon of peroneus tertius muscle
6 Extensor digitorum brevis muscle
7 Tendons of extensor digitorum longus muscle
8 Dorsal metatarsal arteries
9 Medial malleolus
10 Tendon of tibialis anterior muscle
11 **Dorsalis pedis artery**
12 Deep peroneal nerve (on dorsum of foot)
13 Extensor hallucis brevis muscle
14 Tendon of extensor hallucis longus muscle
15 Dorsalis pedis artery with deep plantar branch to the plantar arch
16 Dorsal digital nerves (terminal branches of deep peroneal nerve)
17 Lateral tarsal artery
18 Extensor digitorum brevis muscle (divided)
19 Arcuate artery
20 Dorsal interosseous muscles
21 **Deep peroneal nerve**
22 Medial cuneiform and 1st metatarsal bone
23 Tendon of peroneus longus muscle
24 Abductor hallucis and flexor hallucis brevis muscles
25 Medial plantar artery, vein and nerve
26 Fourth and fifth metatarsal bone
27 Adductor hallucis (oblique head)
28 Tendons of flexor digitorum longus muscle
29 Lateral plantar artery, vein and nerve
30 Flexor digitorum brevis muscle
31 Plantar aponeurosis

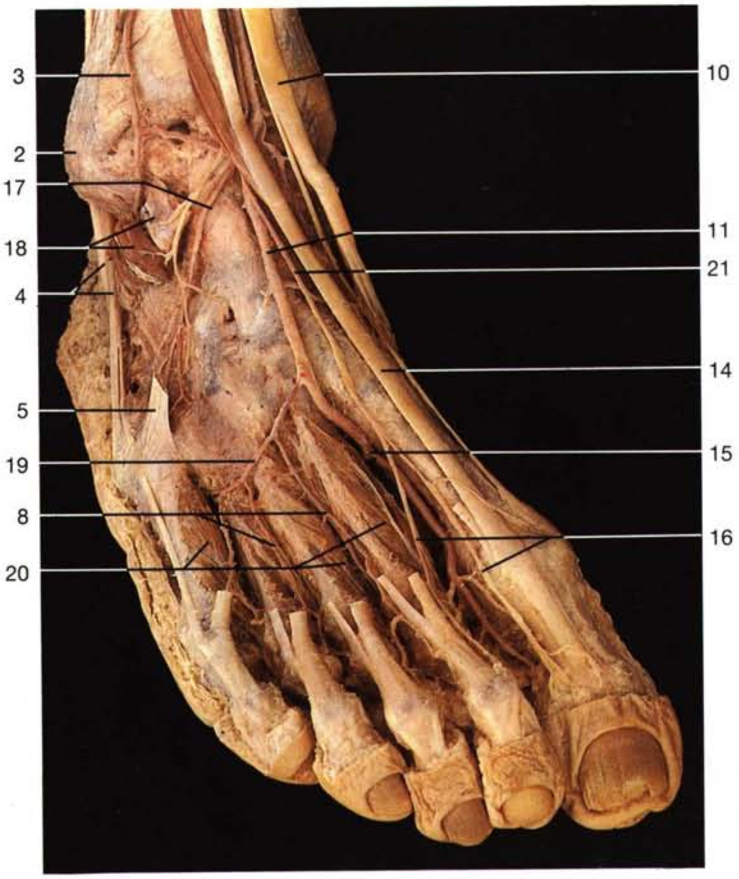

Dorsum of right foot, deep layer (anterior lateral aspect). The extensor digitorum and hallucis breves muscles have been removed.

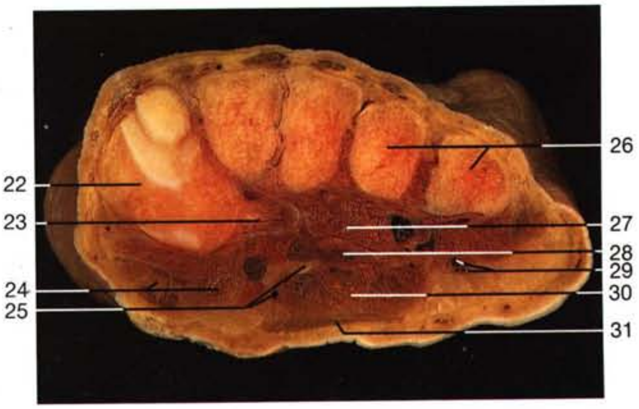

Cross-section of the rigt foot at the level of the metatarsal bones (posterior aspect).

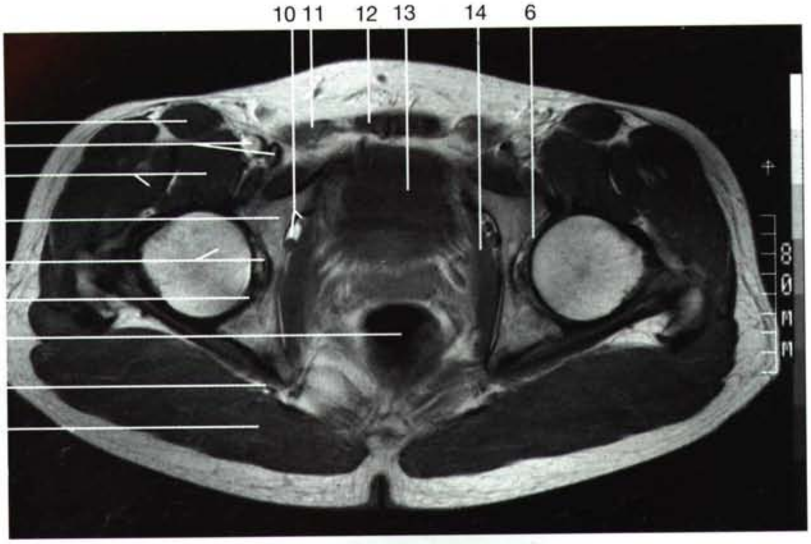
10
11
12
13
14
6
1
2
3
4
5
6
7
8
9

Axial section through the pelvis and the hip joints (section 1; MR-Image; inferior aspect).

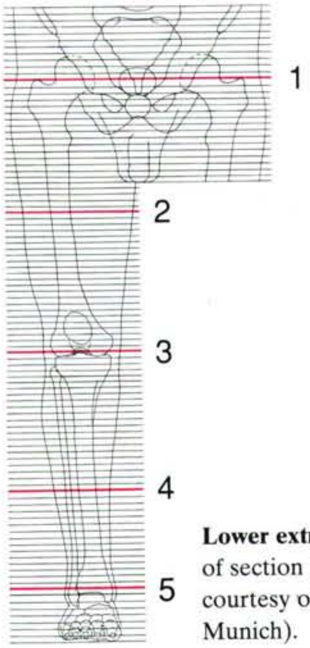

Lower extremity, location of section 1–5. (MR-Images, courtesy of Dr. A. Heuck, Munich).

1	Sartorius muscle
2	**Femoral artery and vein**
3	Iliopsoas muscle
4	Pelvic bone
5	**Femoral head** with ligament of femoral head
6	Articular cavity
7	Rectum
8	**Sciatic nerve** and accompanying artery
9	Gluteus maximus muscle
10	Obturator vessels and obturator nerve
11	Rectus abdominis muscle
12	Pyramidalis muscle
13	Urinary bladder
14	Obturator internus muscle
15	Rectus femoris muscle
16	Vastus intermedius and vastus lateralis of quadriceps femoris muscle

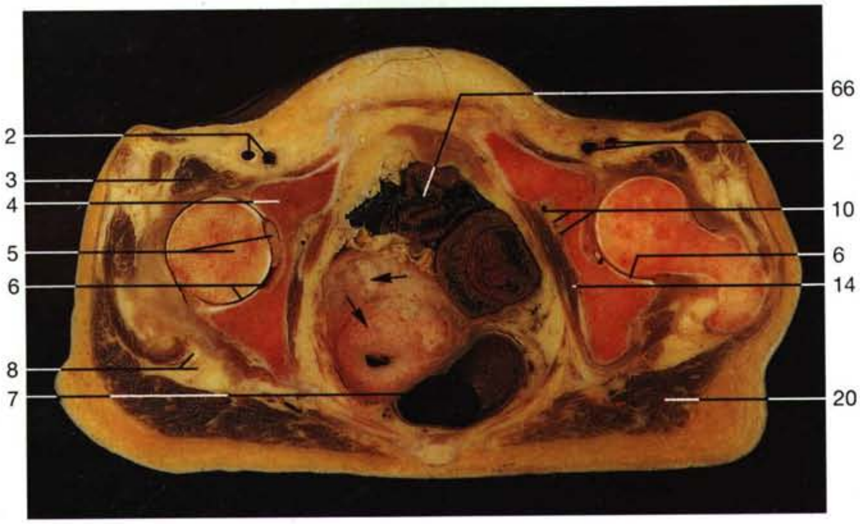
2
3
4
5
6
8
7
66
2
10
6
14
20

Axial section through the pelvis and hip joints in the female (section 1, inferior aspect). (Arrows: uterus, myometrium with myoma).

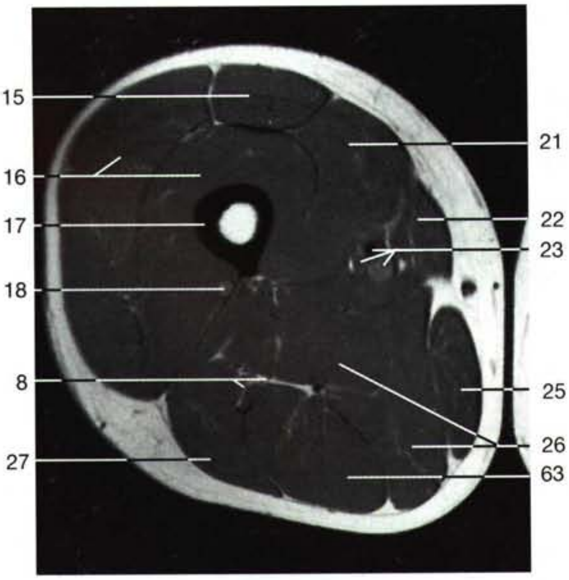
15
16
17
18
8
27
21
22
23
25
26
63

Axial section through the middle of the thigh (section 2; MR-Image, inferior aspect).

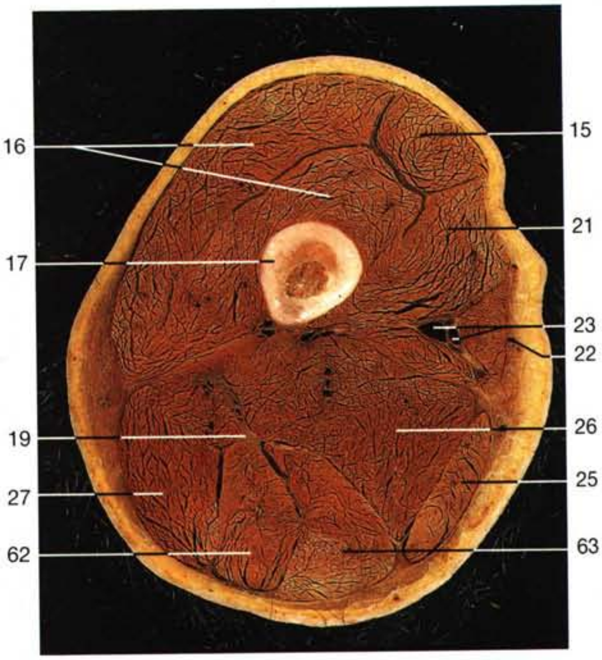
16
17
19
27
62
15
21
23
22
26
25
63

Axial section through the middle of the right thigh (section 2, inferior aspect).

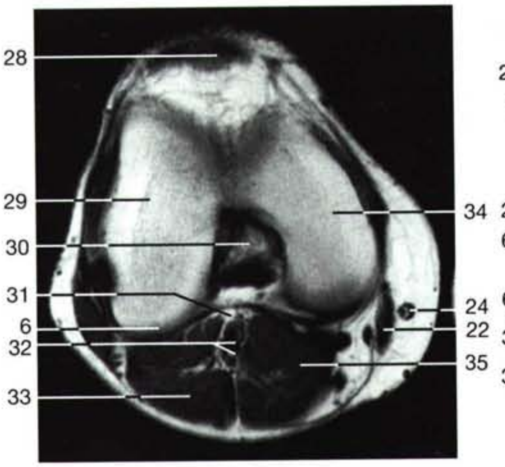

Axial section through the knee joint (section 3; MR-Image; inferior aspect).

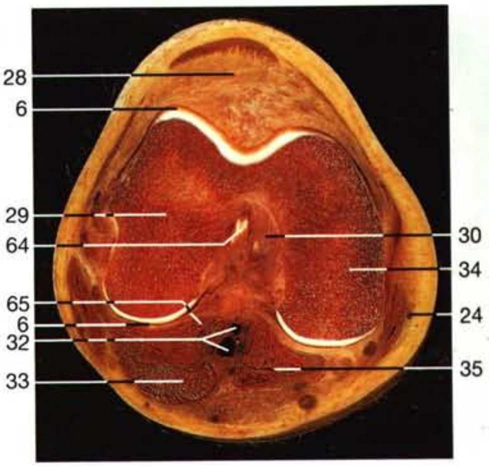

Axial section through the right knee joint (section 3, inferior aspect).

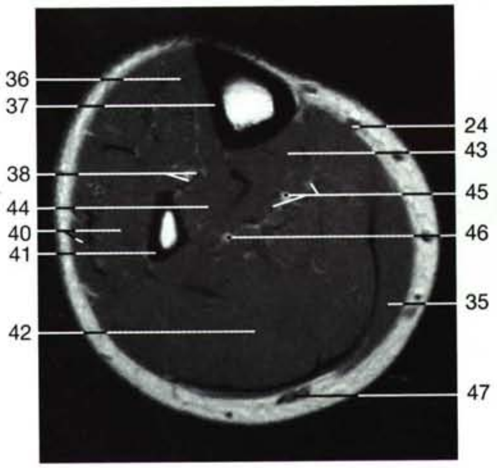

Axial section through the middle of the leg (section 4, MR-Image, inferior aspect).

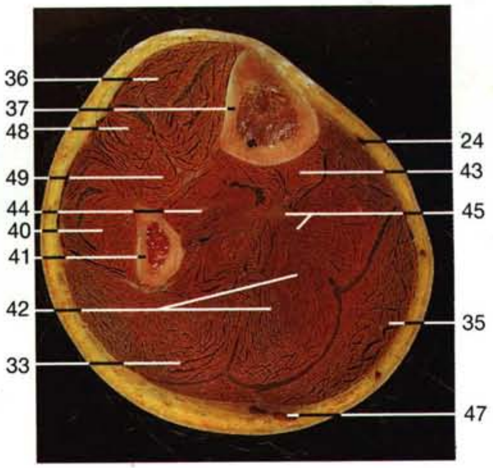

Axial section through the middle of the right leg (section 4, inferior aspect).

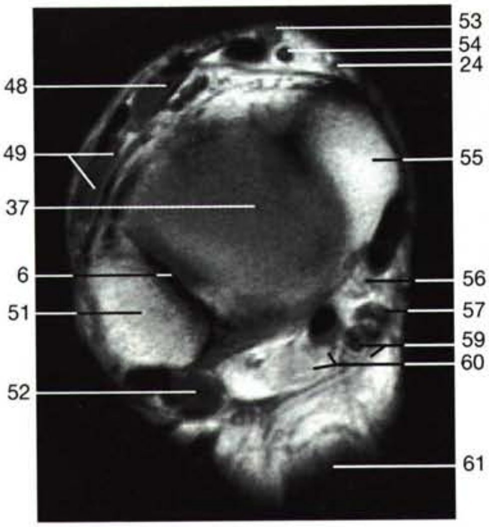

Axial section through the end of the right leg (section 5; MR-Image; inferior aspect).

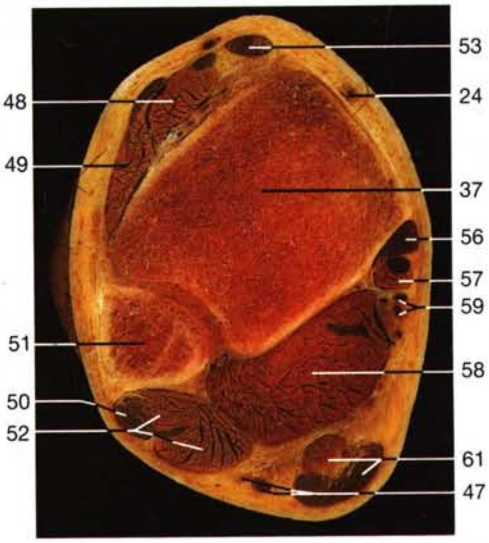

Axial section through the end of the right leg (section 5; inferior aspect).

17 **Femur**
18 Perforating artery
19 Sciatic nerve
20 Gluteus maximus muscle (insertion)
21 Vastus medialis muscle
22 Sartorius muscle
23 Femoral artery and vein
24 Great saphenous vein
25 Gracilis muscle
26 Adductor muscles
27 Biceps femoris muscle
28 Patellar ligament
29 Lateral condyle of femur
30 Posterior cruciate ligament
31 **Tibialis nerve**
32 Popliteal artery and vein
33 Lateral head of gastrocnemius muscle
34 Medial condyle of femur
35 Medial head of gastrocnemius muscle
36 Tibialis anterior muscle
37 **Tibia**
38 Deep peroneal nerve, anterior tibial artery and vein
40 Peroneus longus and brevis muscles
41 **Fibula**
42 Soleus muscle
43 Flexor digitorum longus muscle
44 Tibialis posterior muscle
45 Posterior tibial artery and vein and tibial nerve
46 Peroneal artery
47 Small saphenous vein and sural nerve
48 Extensor hallucis longus muscle
49 Extensor digitorum longus muscle
50 Tendon of peroneus longus muscle
51 Lateral malleolus (fibula)
52 Peroneus brevis muscle
53 Tibialis anterior muscle (tendon)
54 Dorsalis pedis artery
55 Medial malleolus (tibia)
56 Tibialis posterior muscle (tendon)
57 Flexor digitorum longus muscle (tendon with synovial sheath)
58 Flexor hallucis longus muscle
59 Posterior tibial artery and vein
60 Lateral and medial plantar nerves
61 Calcaneal tendon
62 Semitendinosus muscle
63 Semimembranosus muscle
64 Anterior cruciate ligament
65 Plantaris muscle
66 Small intestine

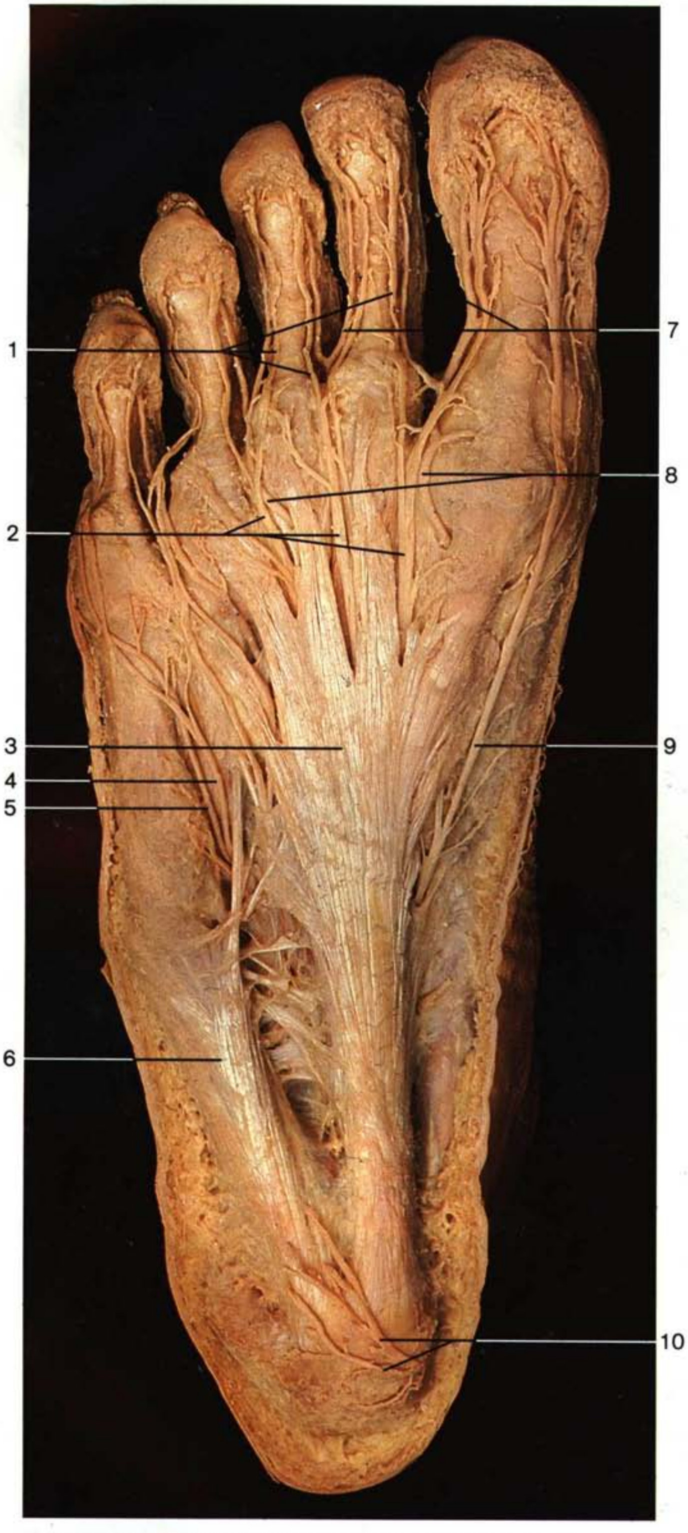
1
7
8
2
3
9
4
5
6
10

Sole of the right foot, superficial layer (from below): dissection of cutaneous nerves and vessels.

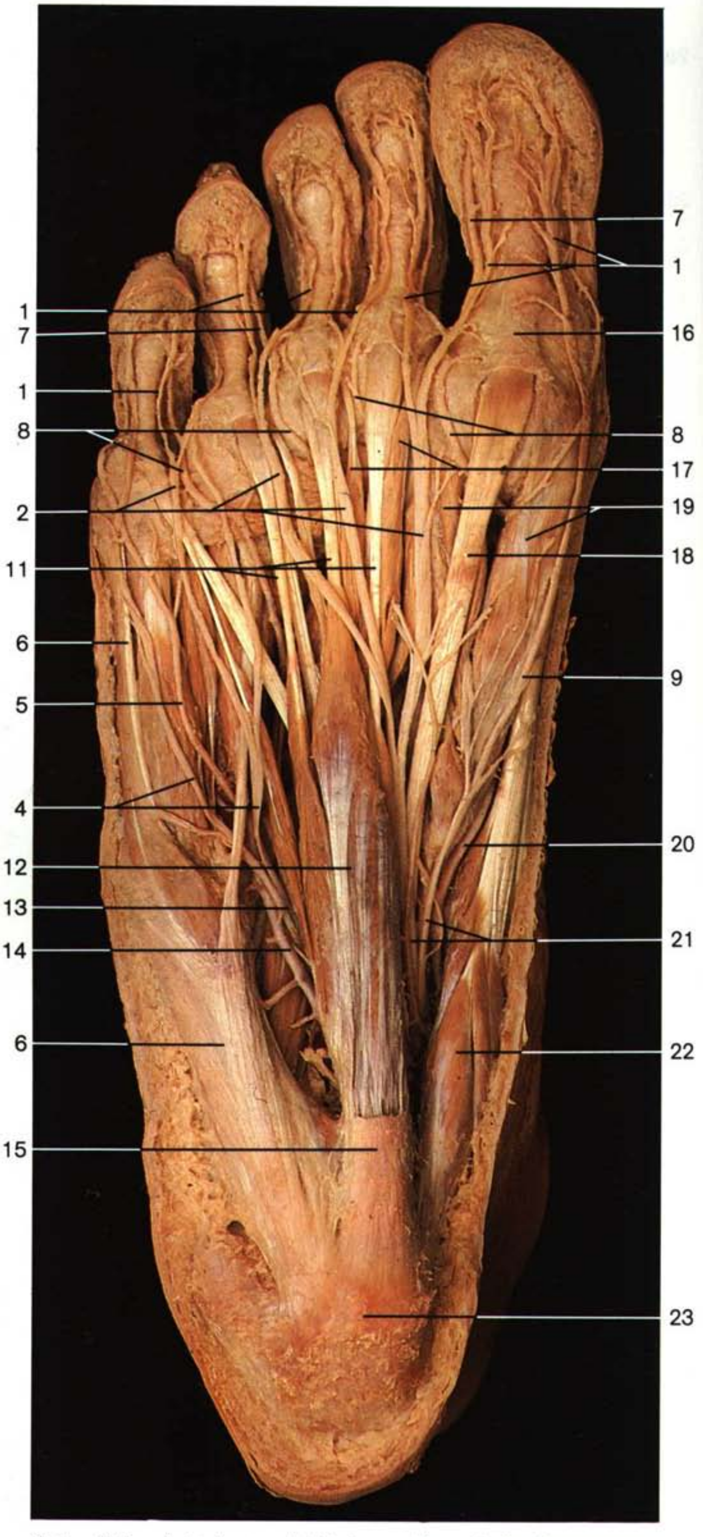
7
1
1
7
16
1
8
8
17
2
19
18
11
6
9
5
4
20
12
13
21
14
6
22
15
23

Sole of the right foot, middle layer (from below). The plantar aponeurosis has been removed.

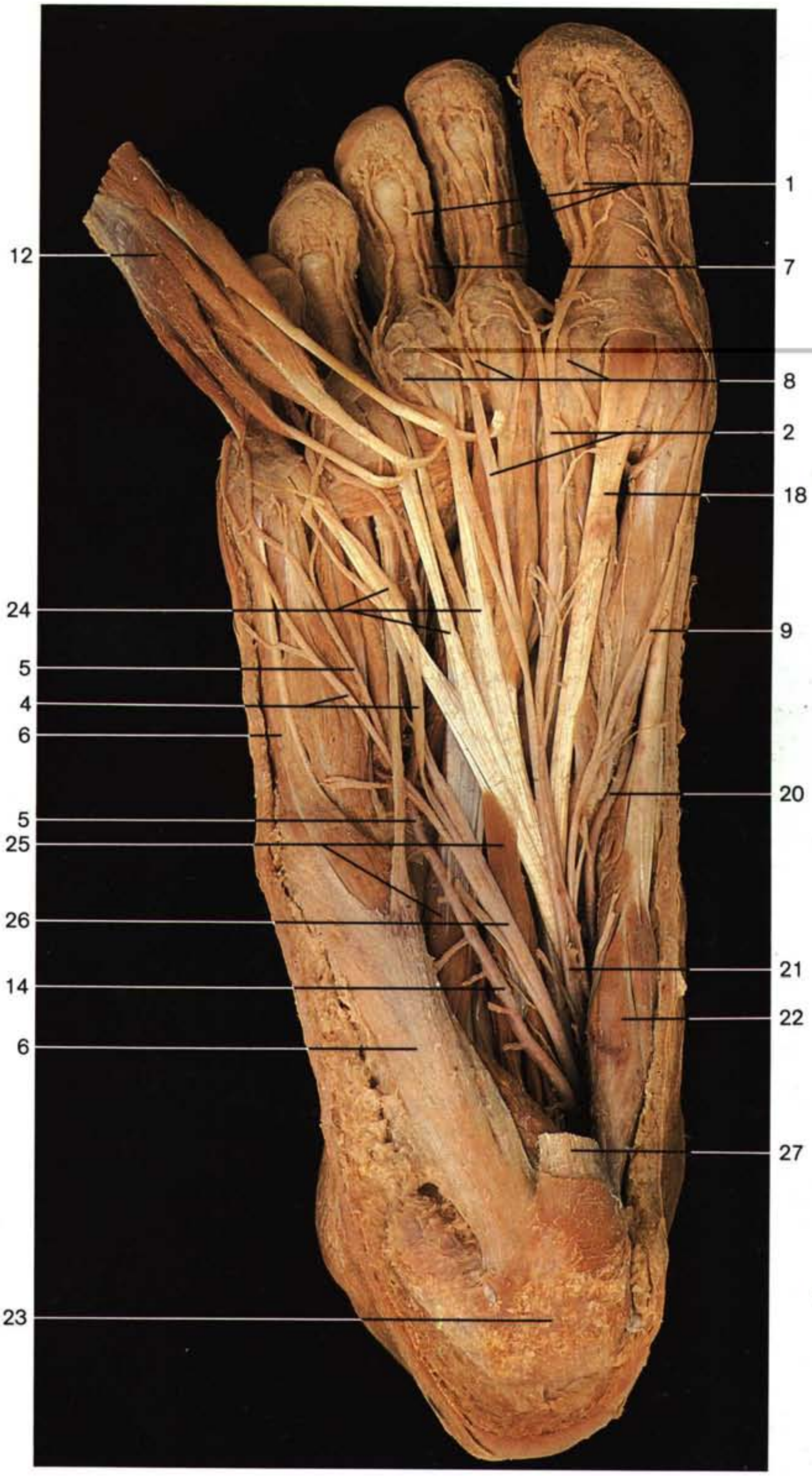

Sole of the right foot, middle layer (from below); dissection of vessels and nerves. The flexor digitorum brevis muscle has been divided and anteriorly reflected.

1 Proper plantar digital nerves
2 Common plantar digital nerves
3 Plantar aponeurosis
4 Superficial branch of lateral plantar nerve
5 Superficial branch of lateral plantar artery
6 Abductor digiti minimi
7 Proper plantar digital arteries
8 Common plantar digital arteries
9 Digital branch of medial plantar nerve to great toe
10 Medial calcaneal branches
11 Tendons of flexor digitorum brevis muscle
12 Flexor digitorum brevis muscle
13 **Superficial branch of lateral plantar nerve**
14 **Lateral plantar artery**
15 Plantar aponeurosis (remnant)
16 Digital synovial sheath
17 Lumbrical muscles
18 Tendon of flexor hallucis longus muscle
19 Flexor hallucis brevis muscle
20 **Medial plantar artery**
21 **Medial plantar nerve**
22 Abductor hallucis muscle
23 Calcaneal tuberosity
24 Tendons of flexor digitorum longus muscle
25 Quadratus plantae muscle
26 **Lateral plantar nerve**
27 Flexor digitorum brevis muscle (cut)
28 Synorial sheaths
29 Plantar arch

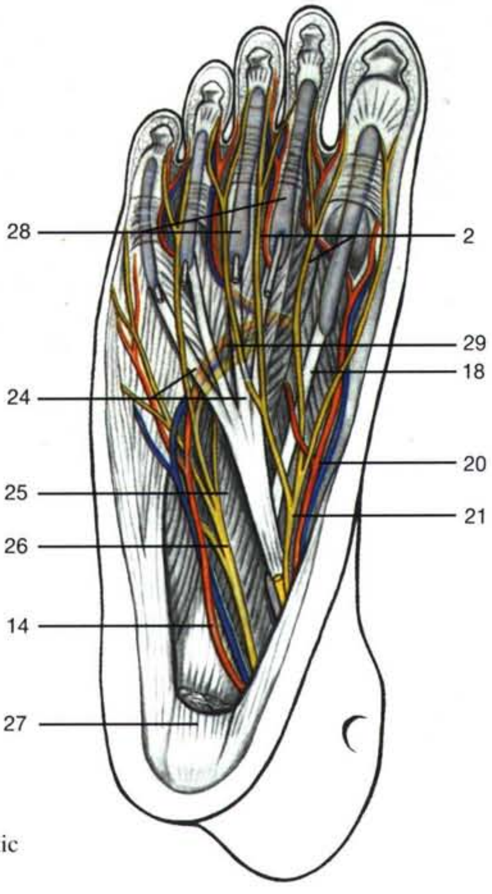

Sole of the right foot. Synovial sheaths of flexor tendons indicated in light blue. (Schematic drawing.)

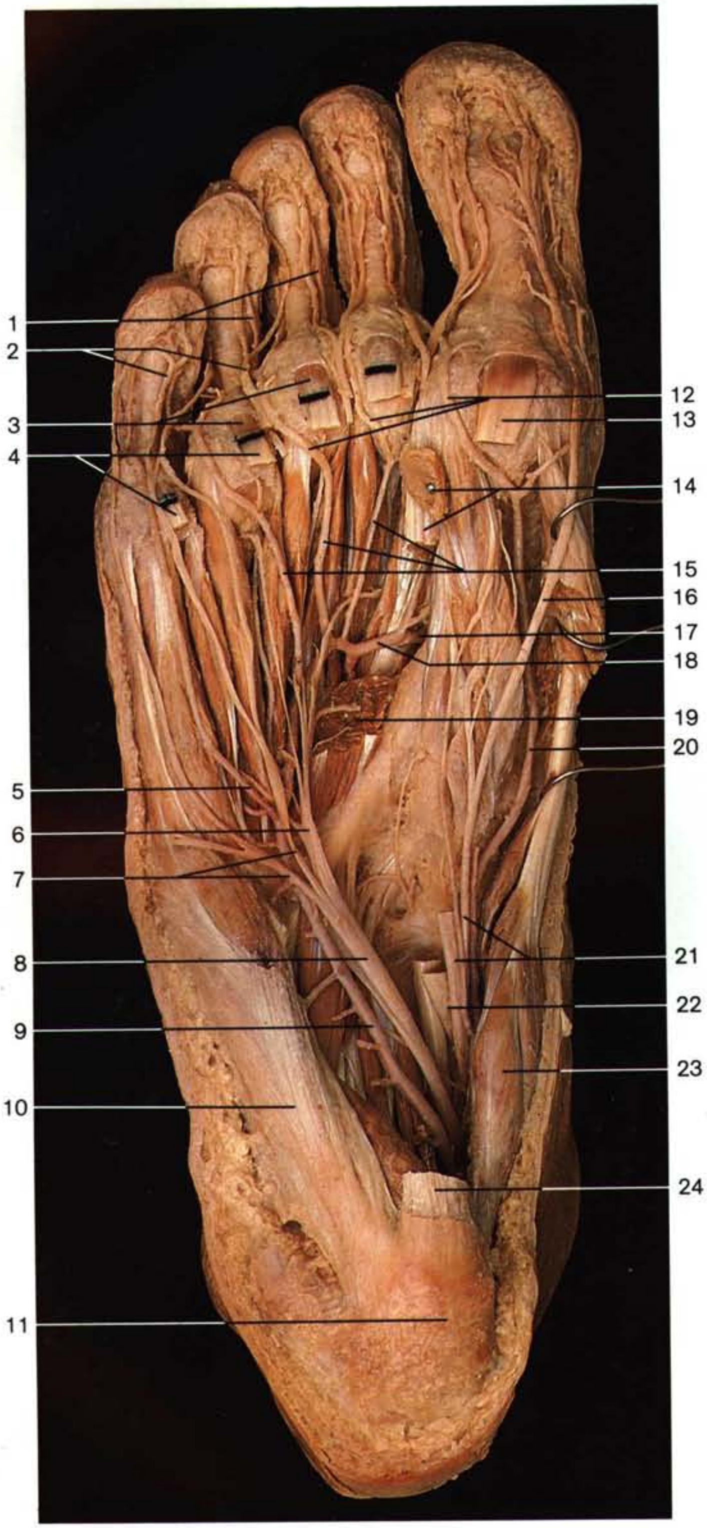

1 Proper plantar digital arteries
2 Proper plantar digital nerves
3 Tendons of flexor digitorum brevis muscle
4 Tendons of flexor digitorum longus muscle
5 Superficial branch of lateral plantar artery
6 Deep branch of lateral plantar nerve
7 Superficial branch of lateral plantar nerve
8 **Lateral plantar nerve**
9 **Lateral plantar artery**
10 Abductor digiti minimi muscle
11 Calcaneal tuberosity
12 Common plantar digital arteries
13 Tendon of flexor hallucis longus muscle
14 Insertion of both heads of adductor hallucis muscle
15 Plantar metatarsal arteries
16 Medial plantar nerve of great toe
17 Deep plantar branch of dorsalis pedis artery (perforating branch)
18 **Plantar arch**
19 Oblique head of adductor hallucis muscle (cut)
20 **Medial plantar artery**
21 **Medial plantar nerve**
22 Crossing of tendons in sole of foot (flexor hallucis longus and flexor digitorum longus muscles)
23 Abductor hallucis muscle
24 Origin of flexor digitorum brevis muscle

Sole of the right foot, deep layer (from below); dissection of vessels and nerves. The flexor digitorum brevis muscle, the quadratus plantae muscle with the tendons of the flexor digitorum longus muscle and some branches of the medial plantar nerve have been removed. The flexor hallucis brevis and adductor hallucis muscles have been cut and portions removed to show the somewhat atypical course of the medial plantar artery and deep muscles of the foot.

BLACK&DECKER

NEW FOR 2009

The Complete Photo Guide to
HOME IMPROVEMENT

More than 200 Value-Adding Remodeling Projects

BLACK&DECKER®

THE COMPLETE PHOTO GUIDE TO

HOME IMPROVEMENT

Creative Publishing
International

MINNEAPOLIS, MINNESOTA
www.creativepub.com

Creative Publishing
International

Copyright © 2009
Creative Publishing international, Inc.
400 First Avenue North, Suite 300
Minneapolis, Minnesota 55401
1-800-328-0590
www.creativepub.com
All rights reserved

Printed in China

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

The complete photo guide to home improvement.
p. cm.
"Black & Decker."
Includes index.
Summary: "Features more than 200 common do-it-yourself remodeling projects"--Provided by publisher.
ISBN-13: 978-1-58923-452-9 (hard cover)
ISBN-10: 1-58923-452-9 (hard cover)
1. Dwellings--Maintenance and repair--Amateurs' manuals.
2. Dwellings--Remodeling--Amateurs' manuals. I. Black & Decker Corporation (Towson, Md.) II. Title.

TH4817.3.C654 2008
643'.7--dc22

2008045755

President/CEO: Ken Fund
VP for Sales & Marketing: Kevin Hamric

Home Improvement Group

Publisher: Bryan Trandem
Managing Editor: Tracy Stanley
Senior Editor: Mark Johanson
Editor: Jennifer Gehlhar

Creative Director: Michele Lanci-Altomare
Senior Design Managers: Jon Simpson, Brad Springer
Design Managers: James Kegley

Lead Photographer: Joel Schrell
Photo Coordinator: Joanne Wawra
Shop Manager: Bryan McLain
Shop Assistant: Cesar Fernandez Rodriguez

Production Managers: Linda Halls, Laura Hokkanen

Page Layout Artist: Danielle Smith

The Complete Photo Guide to Home Improvement
Created by: The Editors of Creative Publishing international, Inc., in cooperation with Black & Decker.
Black & Decker® is a trademark of The Black & Decker Corporation and is used under license.

NOTICE TO READERS

For safety, use caution, care, and good judgment when following the procedures described in this book. The publisher and Black & Decker cannot assume responsibility for any damage to property or injury to persons as a result of misuse of the information provided.

The techniques shown in this book are general techniques for various applications. In some instances, additional techniques not shown in this book may be required. Always follow manufacturers' instructions included with products, since deviating from the directions may void warranties. The projects in this book vary widely as to skill levels required: some may not be appropriate for all do-it-yourselfers, and some may require professional help.

Consult your local building department for information on building permits, codes, and other laws as they apply to your project.

Contents

The Complete Photo Guide to Home Improvement

Introduction **7**

HOME IMPROVEMENT SKILLS

Anatomy of a House 10
Building Codes & Permits 16
Working with Drawings 20
Making Construction Plans 22
Working Safely 24

Techniques **27**

Demolition 28
Building Walls 40
Installing Wallboard 44
Understanding Plumbing 52
Working with Copper Pipe 56
Working with Plastic Pipe 64
Working with PEX Pipe 68
Planning New Plumbing 76
Understanding Wiring 82
Working with Wire & Cable 90
Working with Conduit 102
Working with Boxes 108
Working with Switches 116
Working with Receptacles 122
Planning New Wiring 128
Using Circuit Maps 138

HOME IMPROVEMENT BASICS

Flooring **149**

Installing Underlayment 150
Installing Radiant Floor Mats 156
Installing Hardwood Strip Flooring . . . 162
Installing Laminate Floors 166
Installing Parquet Flooring 170
Installing Sheet Vinyl 174
Evaluating an Existing Floor 176
Installing Resilient Tile 180
Installing Bamboo Flooring 186
Installing Ceramic Tile 191
Installing Mosaic Tile 198
Installing Carpet Squares 200
Refinishing Wood Floors 204

Walls & Ceilings **209**

Installing Paneling 210
Installing Beadboard Wainscoting . . . 214
Building Glass Block Walls 220
Applying Veneer Plaster 224
Paneling Ceilings 228
Installing Suspended Ceilings 232
Installing Acoustical Ceiling Tiles 236
Installing Metal Ceilings 239
Installing Base Molding 242
Installing Picture Rail 246
Installing Chair Rail 248
Installing Crown Molding 250

Windows & Doors **255**

Framing & Installing Windows 256
Installing Garden Windows 264
Installing Bay Windows 268
Installing Skylights 278
Installing Tubular Skylights 288
Installing Interior Doors 290

Installing Entry Doors 306
Installing Storm Doors 310
Installing Patio Doors 312
Installing Door & Window Casing. . . . 318
Lighting 323
Installing Ceiling Lights. 324
Installing Recessed Ceiling Lights . . . 327
Installing Track Lighting 330

MAJOR REMODELING

Kitchen Remodeling 337
Planning & Designing Kitchens 338
Installing Kitchen Cabinets. 350
Building Kitchen Islands 358
Building Tiled Islands 360
Installing Postform Countertops 364
Installing Laminate Countertops 368
Installing Tile on Countertops 376
Installing Tile Backsplashes 386
Installing Kitchen Sinks 390
Replacing Kitchen Faucets. 392
Connecting Drains 398
Hooking Up Dishwashers 402
Installing Food Disposers 407
Installing Undercabinet Lights 412
Bathroom Remodeling. . . . 421
Planning & Designing Bathrooms . . . 422
Installing Shower Kits. 430
Installing Alcove Bathtubs 438
Installing 3-piece Tub Surrounds 446
Installing Sliding Tub Doors 450
Tiling Bathroom Walls. 454
Installing Toilets 458
Installing Wall-hung Vanities 462
Installing Vessel Sinks 464
Installing Pedestal Sinks. 468
Installing Vanity Cabinets 470
Installing Bathroom Sink Faucets & Drains 472
Installing Tub Spouts 480
Installing Vanity Lights 484
Installing Grab Bars 486
Installing Glass Shelving. 488
Installing Towel Warmers 490
Basements & Attics 493
Evaluating Your Attic 494
Evaluating Your Basement 496
Planning the Framing 498
Installing Basement Floors. 500
Installing Attic Floors 506
Installing Basement Walls 510
Installing Attic Walls 514
Finishing Attic Ceilings. 516
Adding Basement Egress Windows . . . 518
Installing Baseboard Heaters 524
Exterior Improvements . . . 529
Estimating Roofing, Siding & Trim . . . 530
Preparing the Roof for Shingles 532
Shingling a Roof 538
Installing Vinyl Siding 546
Resources/Photo Credits. . . 554
Conversion Charts 555
Index 556

Introduction

Home improvement is a huge subject that demands a big book. *The Complete Photo Guide to Home Improvement* fits the bill. With over 200 projects and more than 2,000 color photos, all of the most popular home remodeling subjects and projects are included in this single resource. Here's one way to look at it: If a picture is indeed worth a thousand words, it would require an 8,000-page volume to convey in written form the information packed into the photos in this book. And unlike the tiny pictures found in other home improvement books, the photos within these pages are large and clear and very easy to follow.

Doing your own home remodeling work takes a combination of skills, planning, and specific information. For example, if you decide to remodel your kitchen (the most frequently remodeled room), you should start the process by defining the scale of the project and identifying the tasks within the larger job that you will perform yourself. This phase corresponds to the first section of this book, "Home Improvement Basics." If you love getting your hands dirty this won't be your favorite part, but it is quite important.

Once you have some background information under your belt and a rough plan in your head, it's time to tackle the skillbuilding most of us need. In "Techniques" you will find clear how-to information for just about every task, from demolition to stripping cable. Thorough information on home wiring and home plumbing fill out a large part of the section, followed by step-by-step instructions for installing flooring, finishing walls and ceilings, working with windows and doors, and—finally—upgrading ventilation and lighting.

In the final section of the book you will move from the general to the more specific. Individual chapters on kitchen remodeling and bathroom remodeling deal with projects, planning and information that are specific to each subject. Here you'll find a host of options for installing cabinets and countertops, for example, to supplement the information you've encountered in the Techniques portion of the book. Converting attics and basements also gets its own section of targeted information. And for a strong finish, we offer extended how-to sequences for the two most common exterior improvements—roofing and installing siding.

Whether you are simply freshening up the appearance of your house or embarking on a down-to-the studs remodel of your bathroom, you'll find the information you need. Even if you have special remodeling needs, such as creating accessible kitchens and baths, you'll see that the subjects are covered. More than a quarter of a million homeowners have relied on *The Complete Photo Guide to Home Improvement* as their principal guidebook through the remodeling process. And now, with this new 3rd edition, you'll be pleased that you chose to join them.

shatter resistant

HOME IMPROVEMENT SKILLS

Pionite
See Larger Sample
Storm Cirrus
AC281 Suede
AV741 Suede
AC361 Suede
Graphite Talc
10
11 MADE IN CHINA 12
FL02

Anatomy of a House

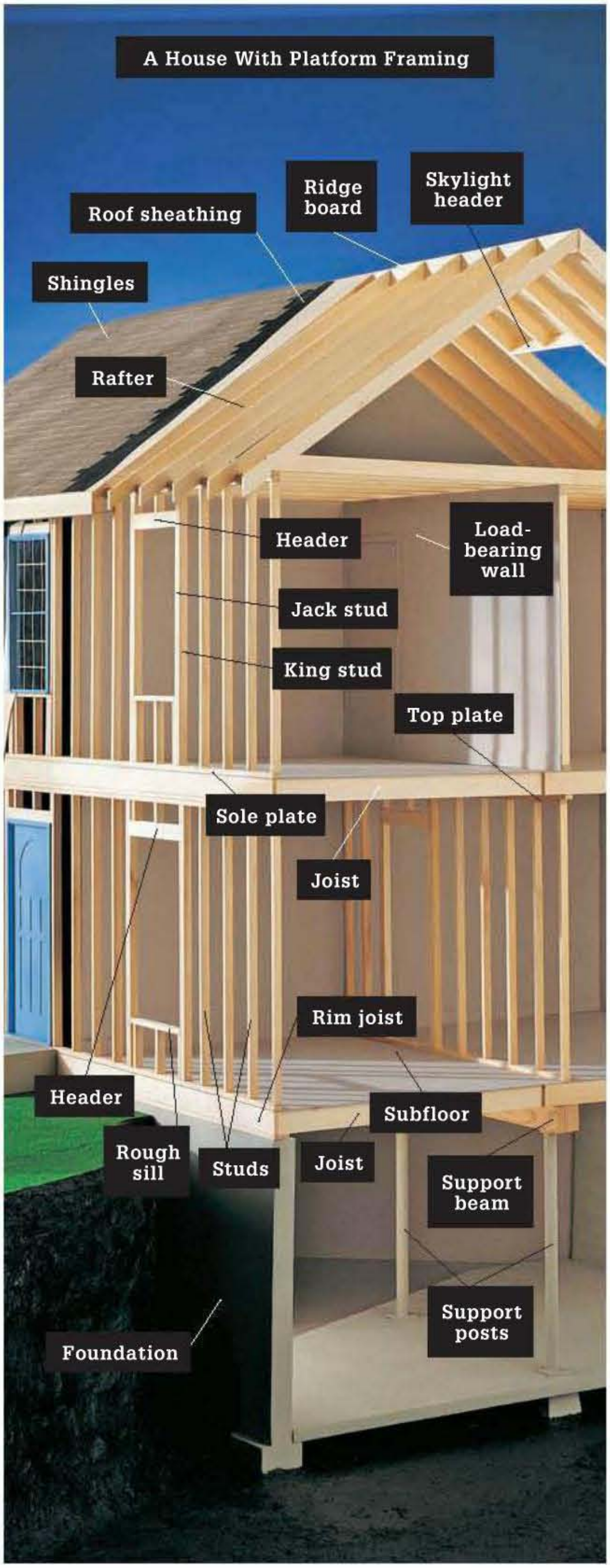
A House With Platform Framing
Roof sheathing
Ridge board
Skylight header
Shingles
Rafter
Header
Load-bearing wall
Jack stud
King stud
Top plate
Sole plate
Joist
Rim joist
Header
Subfloor
Rough sill
Studs
Joist
Support beam
Support posts
Foundation

Before you start a do-it-yourself home improvement project, you should familiarize yourself with a few basic elements of home construction and remodeling. Take some time to get comfortable with the terminology of the models shown on the next few pages. The understanding you will gain in this section will make it easier to plan your project, buy the right materials, and clear up any confusion you might have about the internal design of your home.

If your project includes modifying exterior or load-bearing walls, you must determine if your house was built using platform- or balloon-style framing. The framing style of your home determines what kind of temporary supports you will need to install while the work is in progress. If you have trouble determining what type of framing was used in your

ANATOMY OF A HOUSE WITH PLATFORM FRAMING

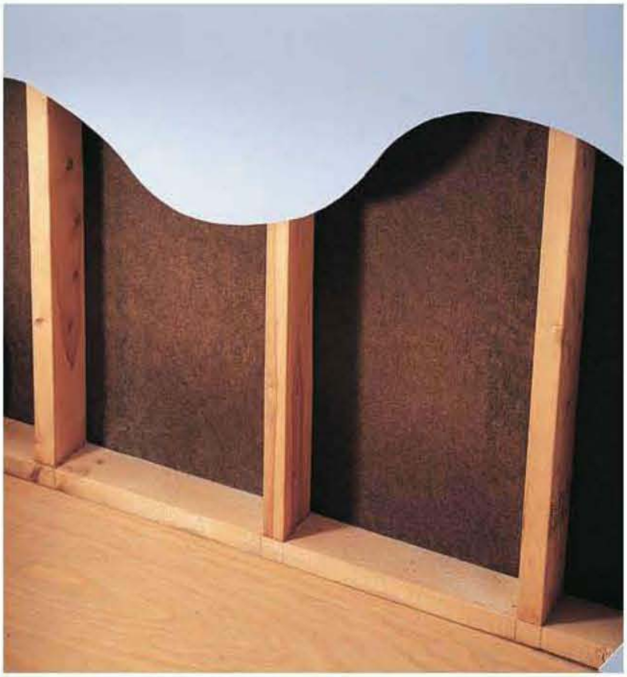

Platform framing (photos, left and above) is identified by the floor-level sole plates and ceiling-level top plates to which the wall studs are attached. Most houses built after 1930 use platform framing. If you do not have access to unfinished areas, you can remove the wall surface at the bottom of a wall to determine what kind of framing was used in your home.

home, refer to the original blueprints, if you have them, or consult a building contractor or licensed home inspector.

Framing in a new door or window on an exterior wall normally requires installing a header. Make sure that the header you install meets the requirements of your local building code, and always install cripple studs where necessary.

Floors and ceilings consist of sheet materials, joists, and support beams. All floors used as living areas must have joists with at least 2 × 8 construction. For modification of smaller joists see page 14.

There are two types of walls: load-bearing and partition. Load-bearing walls require temporary supports during wall removal or framing of a door or window. Partition walls carry no structural load and do not require temporary supports.

ANATOMY OF A HOUSE WITH BALLOON FRAMING

Balloon framing (photos, right and above) is identified by wall studs that run uninterrupted from the roof to a sill plate on the foundation, without the sole plates and top plates found in platform-framed walls (page opposite). Balloon framing was used in houses built before 1930, and it is still used in some new home styles, especially those with high vaulted ceilings.

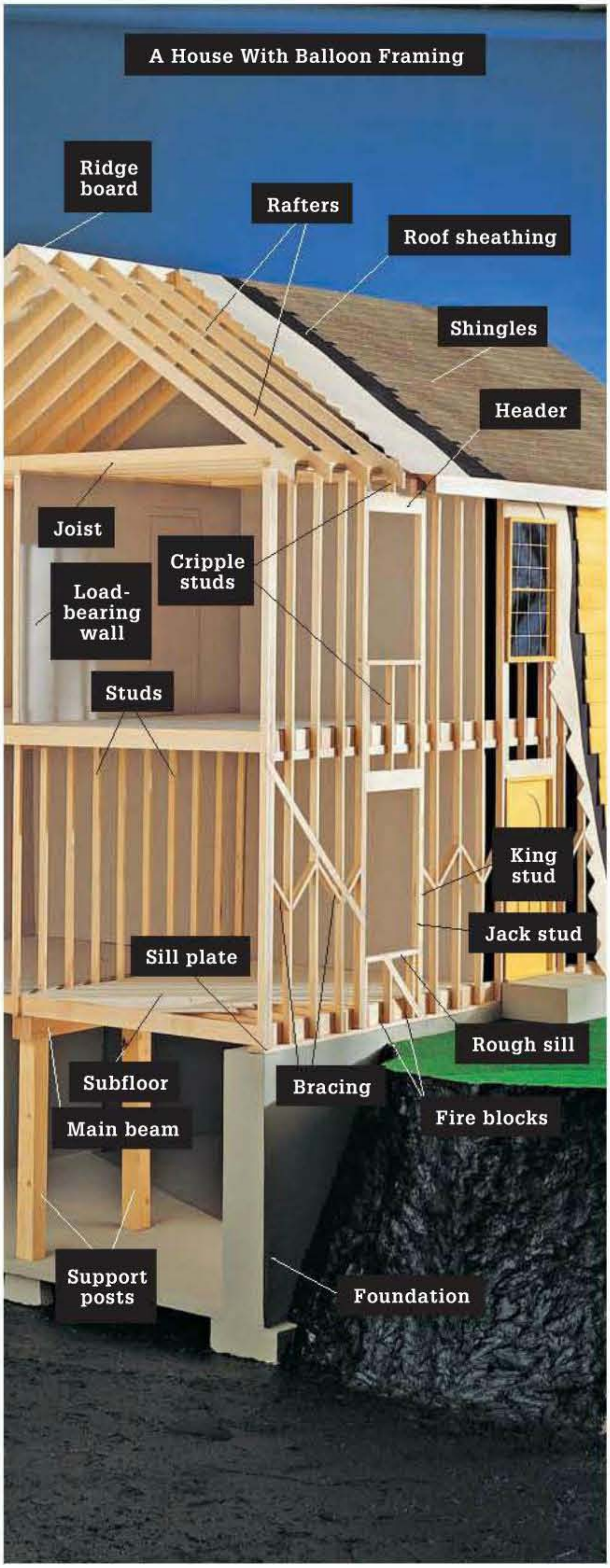
A House With Balloon Framing
Ridge board
Rafters
Roof sheathing
Shingles
Header
Joist
Cripple studs
Load-bearing wall
Studs
King stud
Jack stud
Sill plate
Rough sill
Subfloor
Bracing
Fire blocks
Main beam
Support posts
Foundation

Anatomy Details

Many remodeling projects, like adding new doors or windows, require that you remove one or more studs in a load-bearing wall to create an opening. When planning your project, remember that new openings require a permanent support beam called a header, above the removed studs, to carry the structural load directly.

The required size for the header is set by local building codes and varies according to the width of the rough opening. For a window or door opening, a header can be built from two pieces of 2" dimensional lumber sandwiched around ⅜" plywood (chart, right). When a large portion of a load-bearing wall (or an entire wall) is removed, a laminated beam product can be used to make the new header.

If you will be removing more than one wall stud, make temporary supports to carry the structural load until the header is installed.

Recommended Header Sizes

Rough Opening Width	Recommended Header Construction
Up to 3 ft.	⅜" plywood between two 2 × 4s
3 ft. to 5 ft.	⅜" plywood between two 2 × 6s
5 ft. to 7 ft.	⅜" plywood between two 2 × 8s
7 ft. to 8 ft.	⅜" plywood between two 2 × 10s

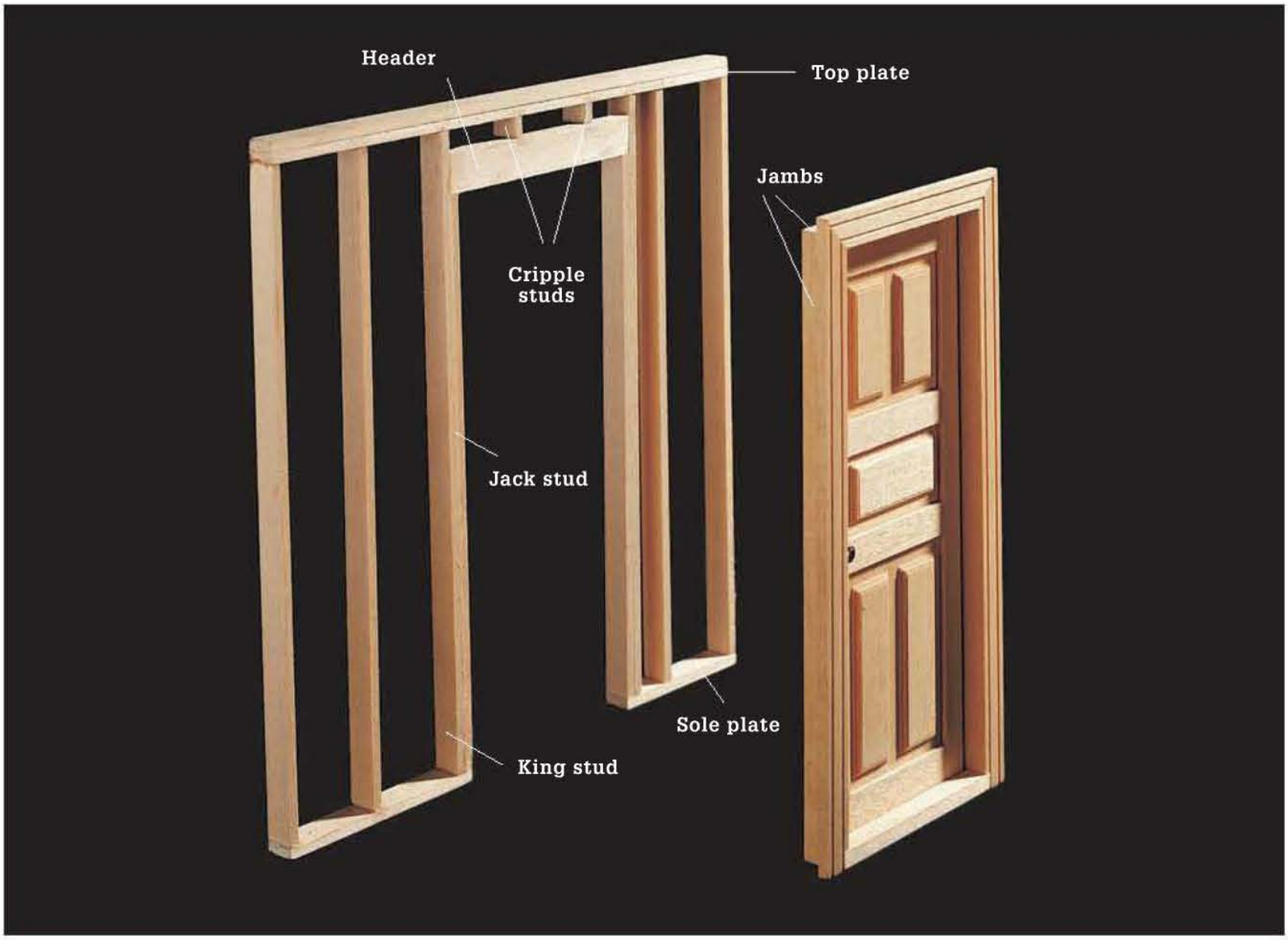

Door opening: The structural load above the door is carried by cripple studs that rest on a header. The ends of the header are supported by jack studs (also known as trimmer studs) and king studs that transfer the load to the sole plate and the foundation of the house. The rough opening for a door should be 1" wider and ½" taller than the dimensions of the door unit, including the jambs. This extra space lets you adjust the door unit during installation.

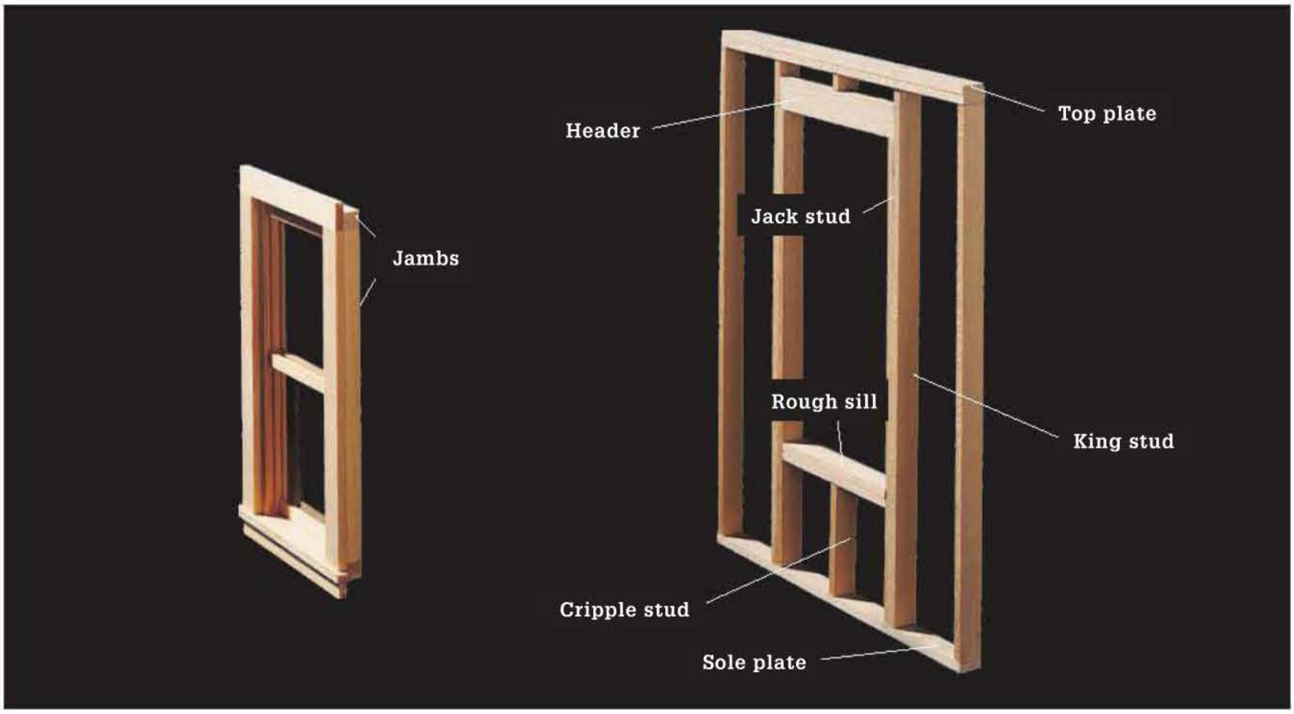

Window opening: The structural load above the window is carried by cripple studs resting on a header. The ends of the header are supported by jack studs and king studs, which transfer the load to the sole plate and the foundation of the house. The rough sill, which helps anchor the window unit but carries no structural weight, is supported by cripple studs. To provide room for adjustments during installation, the rough opening for a window should be 1" wider and ½" taller than the window unit, including the jambs.

Framing Options for Window & Door Openings (new lumber shown in yellow)

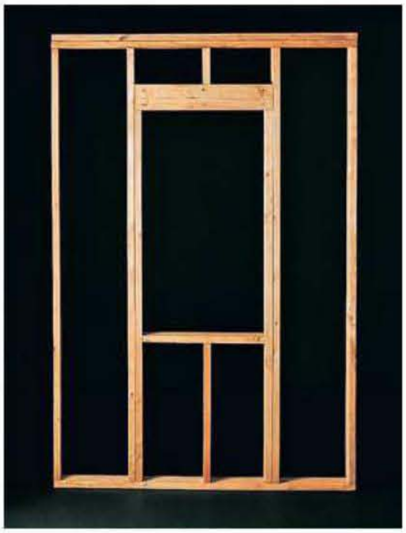

Using an existing opening avoids the need for new framing. This is a good option in homes with masonry exteriors, which are difficult to alter. Order a replacement unit that is 1" narrower and ½" shorter than the rough opening.

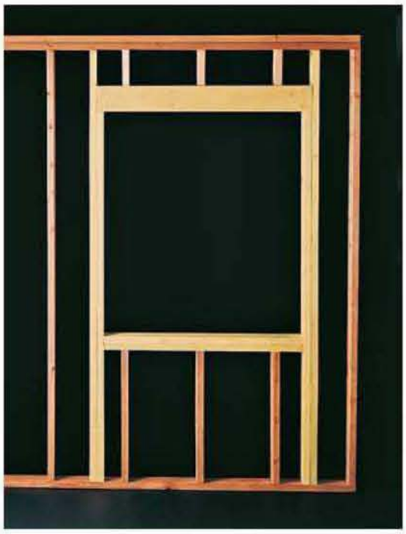

Framing a new opening is the only solution when you're installing a window or door where none existed or when you're replacing a unit with one that is much larger.

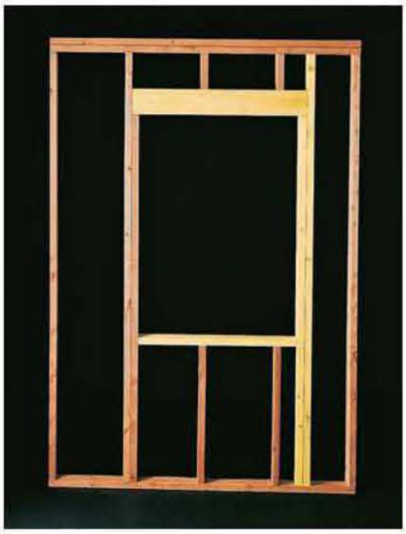

Enlarging an existing opening simplifies the framing. In many cases, you can use an existing king stud and jack stud to form one side of the new opening.

Floor & Ceiling Anatomy

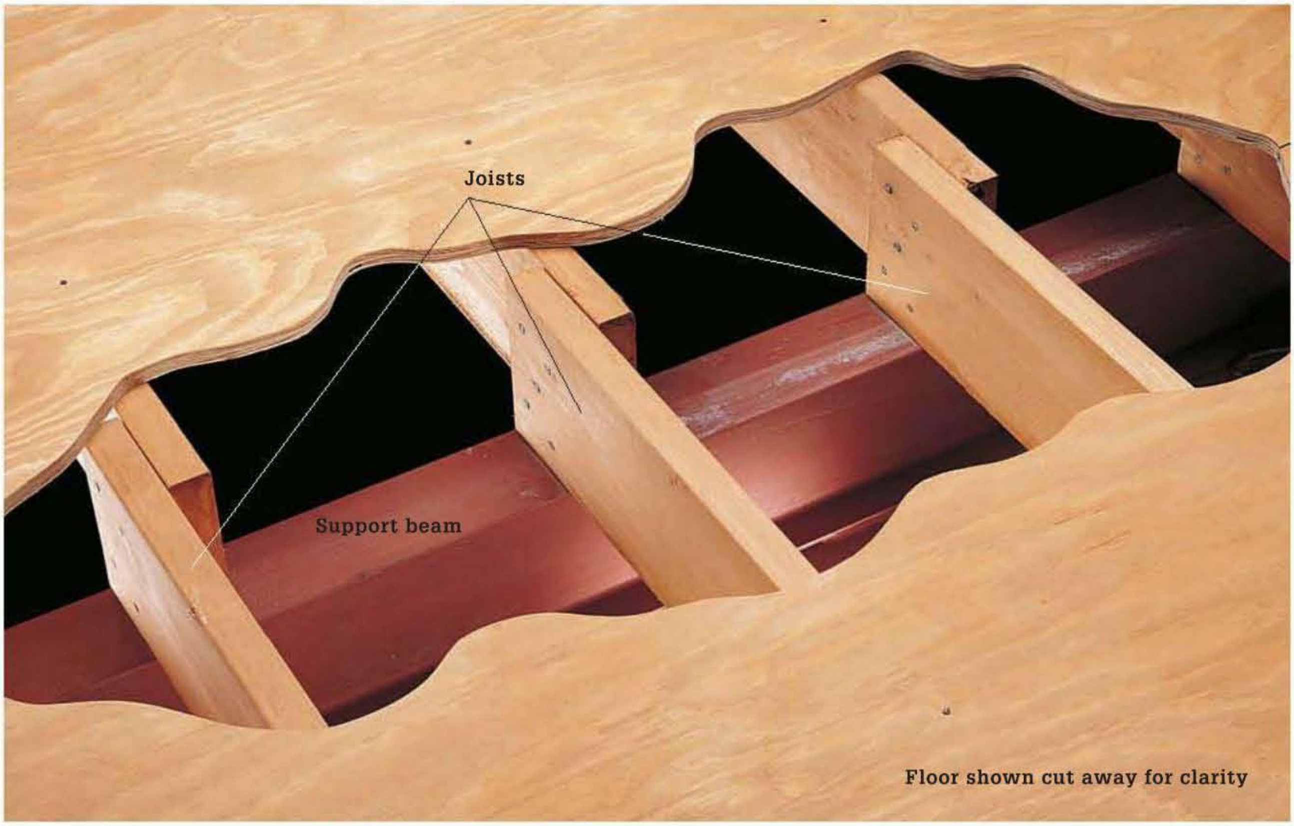
Joists
Support beam
Floor shown cut away for clarity

Joists carry the structural load of floors and ceilings. The ends of the joists rest on support beams, foundations, or load-bearing walls. Rooms used as living areas must be supported by floor joists that are at least 2 × 8" in size. Floors with smaller joists can be reinforced with sister joists (photos, below).

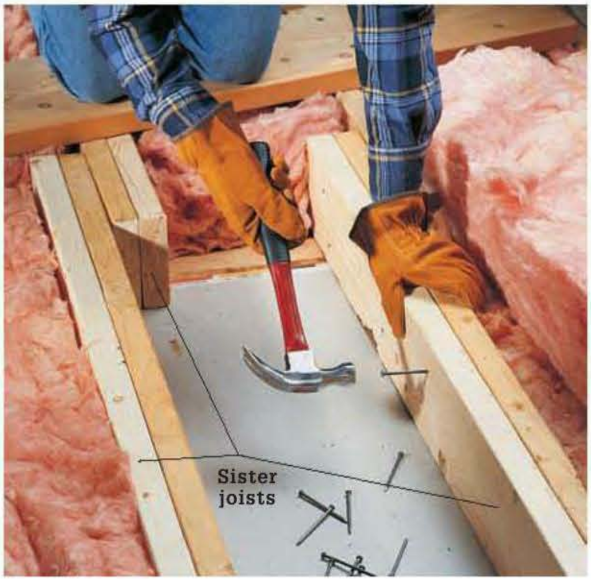
Sister joists

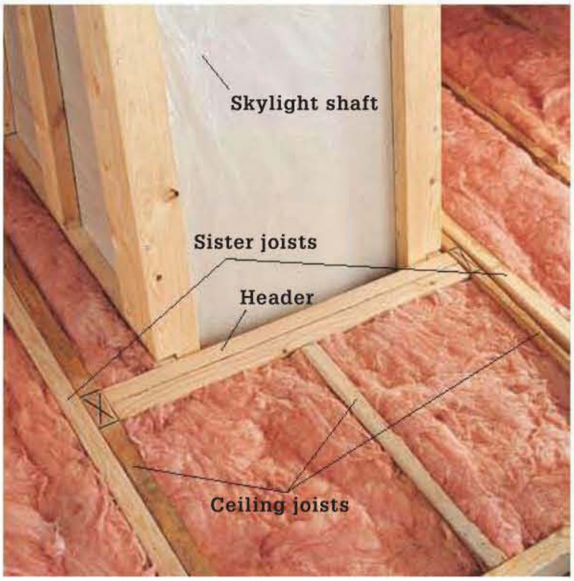
Skylight shaft
Sister joists
Header
Ceiling joists

Floors with 2 × 6 joists, like those sometimes found in attics, cannot support living areas unless a sister joist is attached alongside each original joist to strengthen it (above, left). This often is necessary when an attic is converted to a living area. Sister joists also are used to help support a header when ceiling joists must be cut, such as when framing a skylight shaft (above, right).

Roof Anatomy

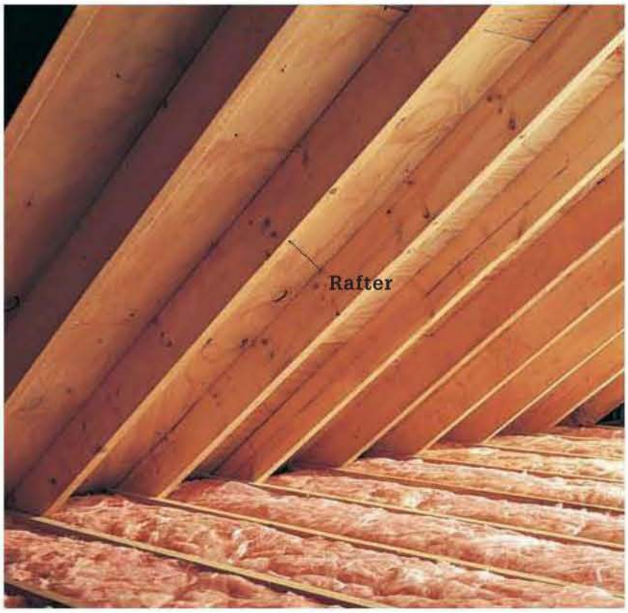

Rafters made from 2 × 4s or 2 × 6s spaced every 16" or 24" are used to support roofs in most houses built before 1950. If necessary, rafters can be cut to make room for a large skylight. Check in your attic to determine if your roof is framed with rafters or roof trusses (right).

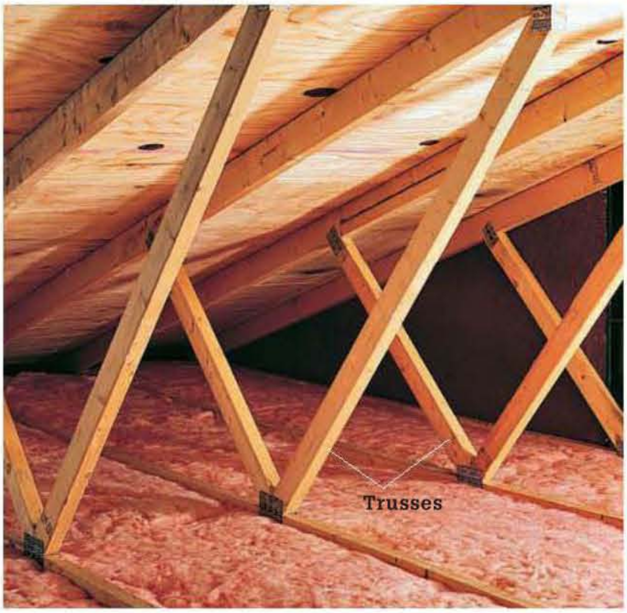

Trusses are prefabricated "webs" made from 2" dimensional lumber. They are found in many houses built after 1950. Never cut through or alter roof trusses. If you want to install a skylight in a house with roof trusses, buy a unit that fits in the space between the trusses.

Wall Anatomy

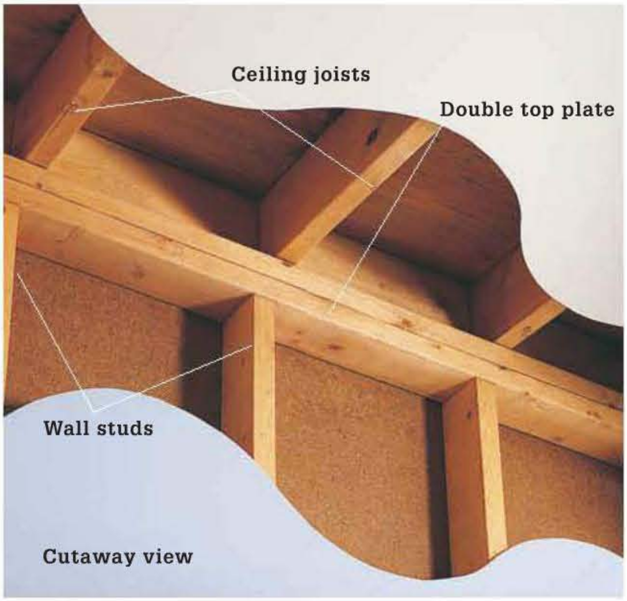
Ceiling joists
Double top plate
Wall studs
Cutaway view

Load-bearing walls carry the structural weight of your home. In platform-framed houses, load-bearing walls can be identified by double top plates made from two layers of framing lumber. Load-bearing walls include all exterior walls and any interior walls that are aligned above support beams.

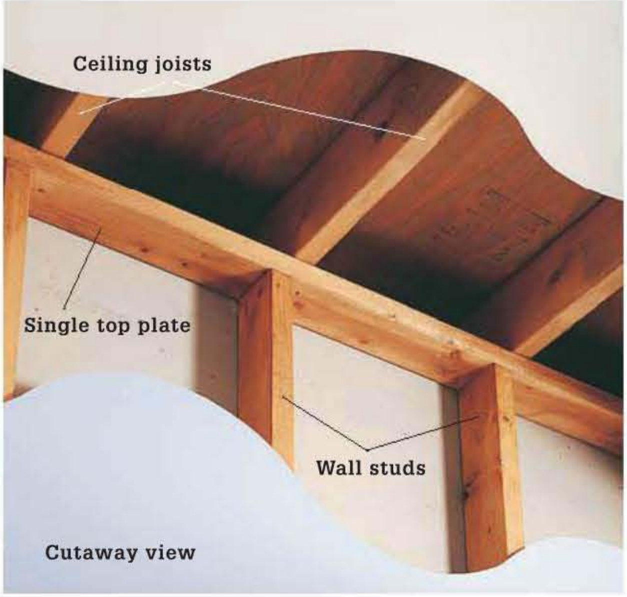
Ceiling joists
Single top plate
Wall studs
Cutaway view

Partition walls are interior walls that do not carry the structural weight of the house. They have a single top plate and can be perpendicular to the floor and ceiling joists but are not aligned above support beams. Any interior wall that is parallel to floor and ceiling joists is a partition wall.

Building Codes & Permits

Building permits are required for any remodeling project that involves a change or addition to your home's structure or mechanical systems. Building permits are issued to ensure your remodeling project meets local building codes, which establish material standards, structural requirements, and installation guidelines for your project. In short, they ensure that your (or your contractor's) work is done properly.

The areas outlined on pages 17 and 18—room dimensions, exits and openings, light and ventilation, and fire protection—are usually covered by general building permits. If your project involves major changes to your plumbing, electrical, or HVAC systems, you may be required to obtain separate permits from the respective administration departments.

Building permits are required by law, and getting caught without them can result in fines from the city and possible trouble with your insurance company. Also, work done without permits can cause problems if you try to sell your house.

Most local building codes follow the national codes, such as the National Electrical Code, but are adapted to meet the demands of local conditions and legislation. Keep in mind that local codes always supersede national codes. Always check with your local building department before finalizing your plans.

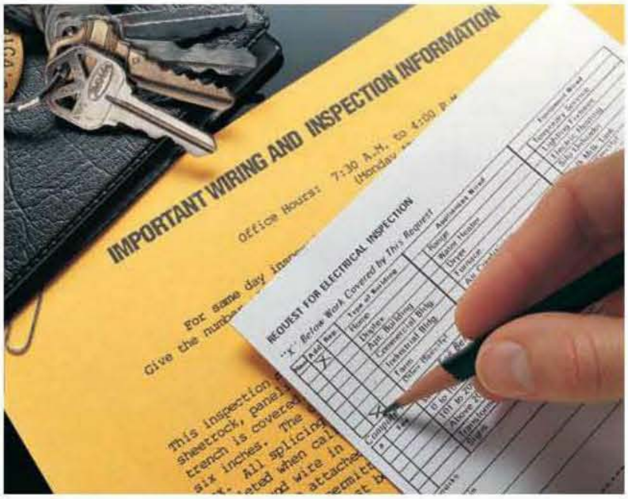

Before issuing permits, your local building department will require plans and cost estimates for your project. After your plans have been approved, you must pay permit fees, which are based on the cost of the project. You'll also learn what inspections are required and when you should call for inspections.

Once issued, a building permit typically is good for 180 days. You can apply for an extension by submitting a written request showing justifiable cause for the delay.

Tip ▸

Here are some tips to help you prepare for the permit process:

- To obtain a building permit, you must fill out a form from your local building department that includes a description of the project; your home's address, legal description, and occupancy; and an estimate of the project cost.
- The building department may require two to four sets of construction documents or drawings of your project—including floor and elevation plans—to be submitted for inspection and approval.
- A building inspector will examine all construction plans and stamp or send written notification of approval and acceptance.
- One set of approved documents is kept by the building official, one set is sent to the applicant, and one set is displayed at the site until the project is completed.
- Some permits are granted by phase of construction. After the work for one phase is completed and inspected, a permit for the next phase is issued. However, building officials will not guarantee issuance of subsequent permits.
- All work is inspected by a building official to ensure compliance with codes and permits.
- Your project is complete only after the local building inspector makes a final inspection and gives approval of your site.

Room Dimensions

- Habitable rooms must be at least 7 ft. wide and 7 ft. deep.
- Ceilings in all habitable rooms, hallways, corridors, bathrooms, toilet rooms, laundry rooms, and basements must be at least 7 ft., 6" high, measured from the finished floor to the lowest part of the ceiling.
- Beams, girders, and other obstructions that are spaced more than 4 ft. apart can extend 6" below the required ceiling height.
- In nonhabitable rooms, such as unfinished basements, ceilings may be 6 ft., 8" from the floor, and beams, girders, and ducts may be within 6 ft., 4" of the floor.
- Habitable rooms cannot have more than 50% of their floor area under sloped ceilings less than 7 ft., 6" high, and no portion of a floor area can be under a ceiling less than 5 ft. high.
- Finished floor is not considered measurable floor area when it is below sloped ceilings less than 5 ft. high or beneath furred ceilings less than 7 ft., 6" high.
- One habitable room in a home must have at least 120 square feet of gross floor area. Other habitable rooms can have gross floor space of 70 sq. ft. minimum.

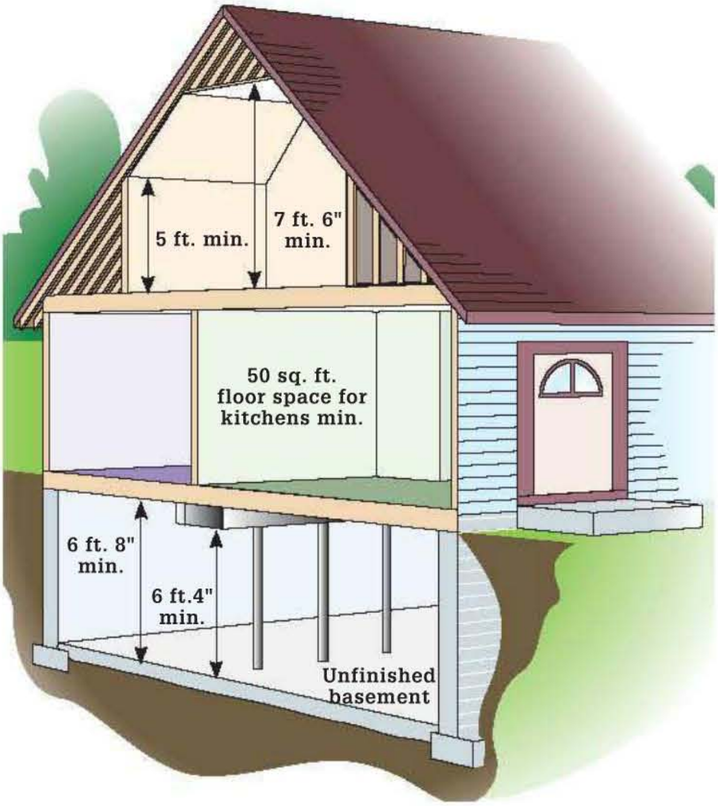

- Kitchens cannot have less than 50 sq. ft. of gross floor area.
- Hallways must be at least 3 ft. wide.

Exits & Openings

- Sleeping rooms and habitable basements must have at least one egress window or exterior door for emergency escape. Occupants must be able to open the exit from inside the home, without a key or tool.
- An egress window must have a net clear opening of at least 5.7 sq. ft., with a minimum height of 24" and a minimum width of 20".
- Window sills on egress windows cannot be more than 44" above the floor.
- Egress windows below ground level must have window wells. If the wells are deeper than 44", they must have permanent ladders or steps. The steps can project up to 6" into the well but must be usable when the window is fully opened. Steps must be at least 12" wide and project at least 3" from the wall. Ladder rungs must be less than 18" apart.
- Screens, bars, grills, and covers on emergency exits must open easily and be removable from inside the home, without tools or keys.
- Exit doors must be at least 3 ft. wide and 6 ft., 8" high. They must provide direct outside access and operate without special knowledge or tools.
- Bulkhead enclosures may serve as emergency exits in habitable basements if they provide direct access to the basement and meet the dimension requirements for emergency exits.

Natural Light & Ventilation

- Ventilation includes windows, doors, louvers, and other approved openings or mechanical systems.
- Windows must equal at least 8% of the floor area in habitable rooms. The minimum openable area of a window must equal at least 4% of the room's floor area.
- In bathrooms, windows must be at least 3 sq. ft., and at least half of the window must open.
- Windows must open and operate from inside the room, and they must exit to a street, alley, yard, court, or porch.
- Window light can be replaced by an artificial light if it produces 6.46 lux from 30" above the floor.
- Mechanical ventilation can replace operable windows. In bedrooms, ventilation must supply outside air at a rate of 15 cubic ft. per minute (cfm) for each occupant. In primary bedrooms, the rate is based on two occupants. In additional bedrooms, the rate is based on one occupant per room.
- In bathrooms, intermittent mechanical ventilation rates must be 50 cfm, and continuous rates must be 20 cfm. Bathroom ventilation must exhaust to the outside.

Fire Protection

- All concealed and interconnected spaces, such as soffits, drop and cove ceilings, stair stringers, and areas around vents, pipes, ducts, chimneys, and fireplaces must be fireblocked to prevent fire spread.
- Exterior walls must be constructed to resist fire for at least one hour, with exposure from both sides.
- Batts or blankets of fiberglass, mineral wool, or other approved material must be secured between wall studs and partitions, at the ceiling and floor level, and at 10-ft. intervals both vertically and horizontally.
- Foam insulation installed in interior walls covered with ½" wallboard or other approved material must have a flame-spread rating of 75 or less and a smoke-developing index of 450 or less.
- Other insulation, including facings, vapor barriers, and breather papers, must have a flame-spread index of 25 or less and a smoke-developing index of 450 or less.
- Loose-fill insulation mounted with screens or supports must have a flame-spread rating of 25 or less and a smoke-developing index of 450 or less.

- Wall and ceiling finishes must have a flame-classification rating of 200 or less and a smoke-developing index of 450 or less.
- Smoke alarms must be installed in bedrooms, in hallways near bedrooms, and on each full story of a home. Multiple alarms must be wired together so one activation triggers all alarms.

Universal Design ▸

Universal design is intended for all people. While standard home and product designs are based on the "average" person—that is, the average adult male—not everyone fits into that category. Some people are short, some tall; some have difficulty walking, while others walk ably but find bending difficult. And physical abilities change constantly, as do family situations. By incorporating universal design into your remodeling plans, you can create spaces that work better for everyone who lives in or visits your home, regardless of their size, age, or ability.

Universal design is simply good design that improves everyday situations. For example, wide doorways make passage easier for a person carrying a load of laundry as well as for someone in a wheelchair; a lowered countertop enables a child to help prepare dinner and allows a person who tires easily to sit while cooking. More a way of thinking than a set of rules, universal design can be applied to any area of your home—from room layouts to light fixtures to door hardware. In all cases, universal design encourages independence by creating a safe, comfortable environment.

Many people take on remodeling projects to accommodate changes in their households. Perhaps you are remodeling because your aging parents are coming to live with you or your grown children or grandchildren are coming for an extended visit. Or you may be preparing your home for your own retirement years. Considering both your current and future needs is an essential part of a fundamental universal design concept: creating a lifespan home—one that accommodates its residents throughout their lives. A lifespan home enables your aging parents to live comfortably with you now and will allow you to stay in your home as you grow older. And, while universal design makes your everyday life easier, it will also make your home more appealing to a wide range of potential buyers if you choose to sell.

Much of the universal design information in this book comes from universal design specialists, kitchen and bath designers, physical therapists, specialty builders and manufacturers, and organizations such as the National Kitchen and Bath Association (NKBA). Some suggestions are ADA (Americans with Disabilities Act) requirements; while these generally apply to public spaces, they often are used as guidelines for residential design. As always, be sure that all aspects of your project meet local code requirements.

For more help with planning with universal design, contact a qualified professional. Many kitchen and bath designers, home builders, and product manufacturers specialize in universal design.

Working with Drawings

Drawings are necessary for any remodeling project that involves construction, enlargement, alteration, repair, demolition, or change to any major system within your home. There are two basic types of construction drawings: floor plans and elevation drawings.

Floor plans show a room as seen from above. These are useful for showing overall room dimensions, layouts, and the relationships between neighboring rooms. Elevation drawings show a side view of a room, showing one wall per drawing. Elevations are made for both the interior and exterior of a house and generally show more architectural detail than floor plans.

Both floor plans and elevation drawings provide you with a method for planning and recording structural and mechanical systems for your project. They also help the local building department to ensure your project meets code requirements.

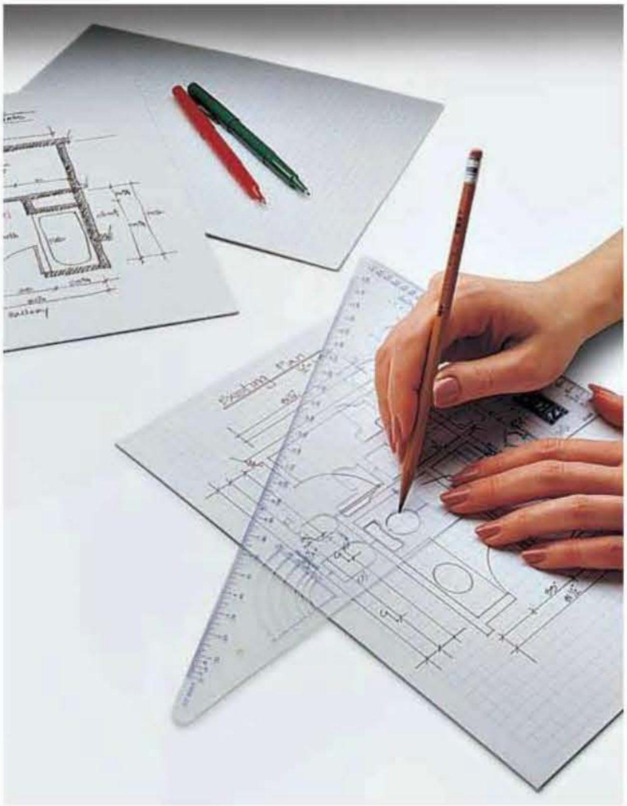

Use existing blueprints of your home, if available, to trace original floor plans and elevation drawings onto white paper. Copy the measurement scale of the original blueprints onto the traced drawings. Make photocopies of the traced drawings, then use the photocopies to experiment with remodeling ideas.

Before you draw up new plans, check with your home's architect, builder, or your local building department. These places often have copies of your home's floor plans on file. If your home is a historic building, your plans may be on file at a state or local historic office or university library.

If you are unable to obtain a copy of your home's floor plans, you can draw your own. This process provides you with a wealth of information about your home. Drawings let you see how changes will affect your home's overall layout and feel. They also help you plot your ideas, list materials, and solve design problems.

Follow the steps on page 21 to create floor plans and elevation drawings. Keep in mind that your plans may change as your ideas develop; until you have worked out all the elements of your design, consider your plans to be drafts. When you have arrived at a plan that meets your needs, draw up final floor and elevation plans to submit to your local building department.

To create floor plans, draw one story at a time. First, measure each room on the story from wall to wall. Transfer the rooms' dimensions to ¼" grid paper, using a scale of ¼" = 1 ft. Label each room for its use and note its overall dimensions. Include wall thicknesses, which you can determine by measuring the widths of door and window jambs—do not include the trim.

Next, add these elements to your drawings:

- Windows and doors; note which way the doors swing.
- Stairs and their direction as it relates to each story.
- Permanent features, such as plumbing fixtures, major appliances, countertops, built-in furniture, and fireplaces.
- Overhead features, such as exposed beams, or wall cabinets—use dashed lines.
- Plumbing, electrical, and HVAC elements. You may want a separate set of drawings for these mechanical elements and service lines.
- Overall dimensions measured from outside the house. Use these to check the accuracy of your interior dimensions.

To create elevation drawings, use the same ¼" = 1 ft. scale, and draw everything you see on one wall (each room has four elevations). Include:

- Ceiling heights and the heights of significant features such as soffits and exposed beams.
- Doors, including the heights (from the floor to the top of the opening) and widths.
- Windows, including the height of the sills and tops of the openings, and widths.
- Trim and other decorative elements.

When your initial floor plans and elevations are done, use them to sketch your remodeling layout options. Use overlays to show hidden elements or proposed changes to a plan. Photographs of your home's interior and exterior may also be helpful. Think creatively, and draw many different sketches; the more design options you consider, the better your final plans will be.

When you have completed your remodeling plans, draft your final drawings and create a materials list for the project.

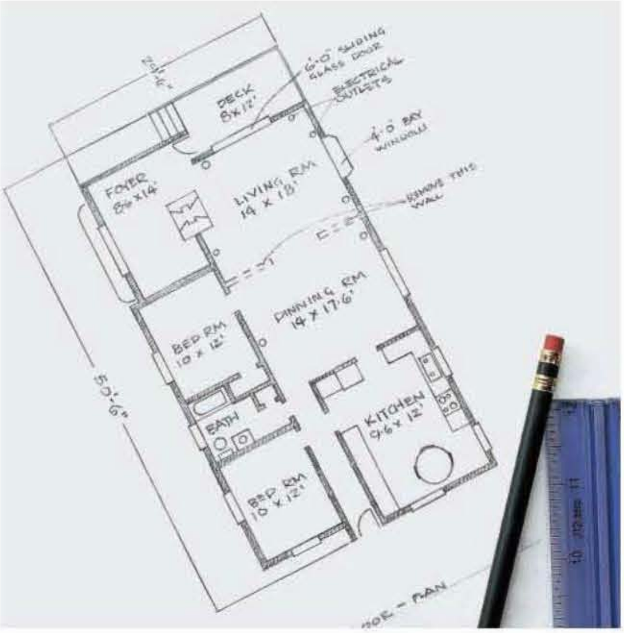
20'-6"
DECK 8x12'
6'-0" SLIDING GLASS DOOR
ELECTRICAL OUTLETS
4'-0" BAY WINDOW
FOYER 8½x14'
LIVING RM 14 x 18'
REMOVE THIS WALL
DINNING RM 14 x 17'6'
BED RM 10 x 12'
BATH
KITCHEN 9'6 x 12'
BED RM 10 x 12'
50'-6"
FLOOR - PLAN

Draft a detailed floor plan showing the layout of the area that will be remodeled, including accurate measurements. Show the location of new and existing doors and windows, wiring, and plumbing fixtures.

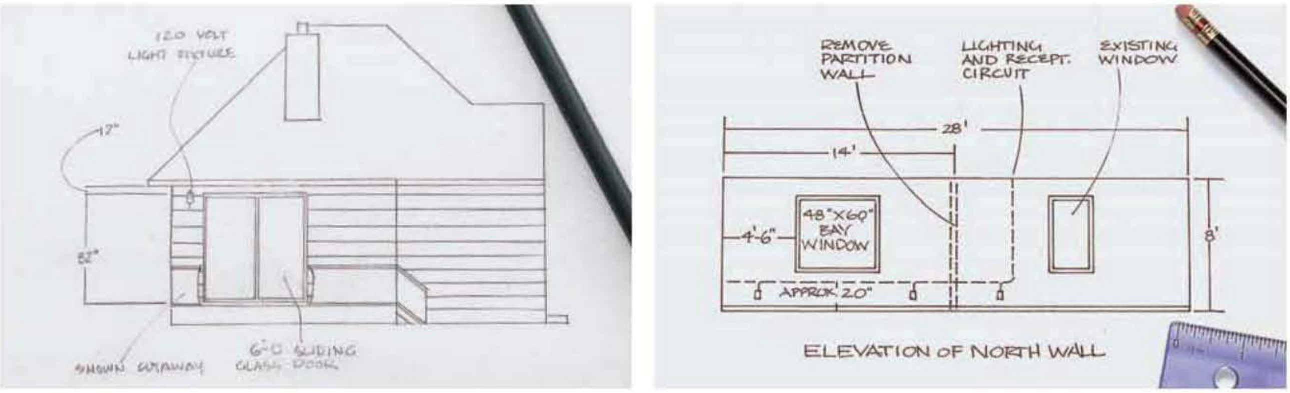
120 VOLT LIGHT FIXTURE
12"
82"
SHOWN CUTAWAY
6'-0 SLIDING GLASS DOOR
REMOVE PARTITION WALL
LIGHTING AND RECEPT. CIRCUIT
EXISTING WINDOW
28'
14'
4'-6"
48"X60" BAY WINDOW
8'
APPROX 20"
ELEVATION OF NORTH WALL

Create elevation drawings showing a side-view layout of windows and doors, as viewed from both inside and outside the home. Indicate the size of windows and doors, ceiling heights, and the location of wiring and plumbing fixtures.

Making Construction Plans

The best way to prepare for a remodel project is to create a construction plan. Having a complete construction plan enables you to view your entire project at a glance. It helps you identify potential problems, provides sense of the time involved, and establishes a logical order of steps. Without a construction plan, it's easy to make costly mistakes, such as closing up a wall with wallboard before the rough-ins have been inspected.

The general steps shown here follow a typical construction sequence. Your plan may differ at several points, but thinking through each of these steps will help you create a complete schedule.

1. CONTACT THE BUILDING DEPARTMENT

To avoid any unpleasant—and expensive—surprises, discuss you project with a building official. Find out about the building codes in your area and what you'll need to obtain the applicable permits. Explain how much of the work you plan to do yourself. In some states, plumbing, electrical, and HVAC work must be done by licensed professionals. Also determine what types of drawings you'll need to get permits and whether you'll need engineer's drawings and calculations.

2. CREATE YOUR DRAWINGS

Make your floor plans and elevation drawings (see pages 20 to 21). This step also involves most of the design work for your project; you may want to get help from a professional for this phase.

3. GET THE PERMITS

Have your final plans reviewed by the building inspector, and make any necessary adjustments required to obtain all of the permits for your project. This is also the time to schedule inspections. Find out what work must be inspected and when to call for inspections.

4. HIRE CONTRACTORS

If you're getting help with your project, it's best to find and hire the contractors early in the process, as their schedules will affect yours. It may be necessary for some contractors to obtain their own work permits from the building department. To avoid problems, make sure all of the contractors know exactly what work they are being hired to do and what work you will be doing yourself. Always check contractors' references and make sure any contractors are licensed and insured before hiring them. This is also the time to order materials and arrange for delivery.

5. COMPLETE THE FRAMING AND MAJOR MECHANICAL CHANGES

Begin the construction work with any major structural or mechanical changes. Move mechanical elements and reroute major service lines. Complete any rough-ins that must happen before the framing goes up, such as adding ducts, installing under-floor drains, or replacing old plumbing. Complete the new framing. Build the rough openings for windows and doors, and install the windows.

6. COMPLETE THE ROUGH-INS

Run drain, waste, and vent (DWV), water, and gas supply lines. Install electrical boxes, and run the wiring. Complete the HVAC rough-ins. Jot down measurements of pipes and locations of wiring, for future reference. Have the building inspector approve your work before you close up the walls. Install any fixtures that go in during the rough-in stage (others will come after the wall surfaces are installed).

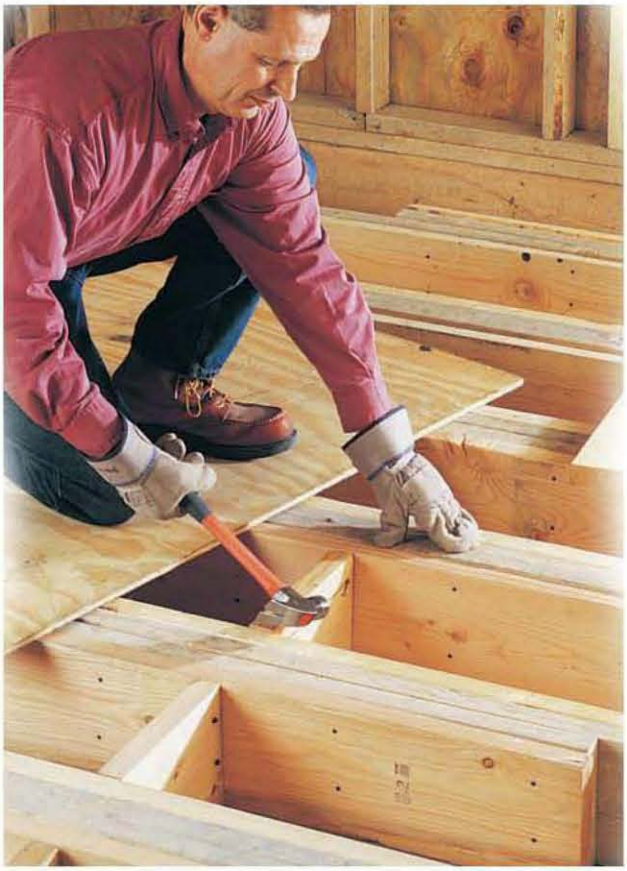

7. FINISH THE WALLS AND CEILINGS

After your work has passed inspection, insulate the walls, ceilings, and pipes. Install fiberglass insulation used as fireblocking. Make sure protector plates for pipes and wires running through framing are in place. Add vapor barriers as required by local code.

Make sure everything is in place before you cover up the framing, then finish the walls and ceilings. If you're installing wallboard, do the ceilings first, then the walls. Tape and finish the wallboard. Install other finish treatments. Texture, prime, and paint the wallboard when it's convenient. If you are installing a suspended ceiling, do so after you finish the walls.

8. ADD THE FINISHING TOUCHES

Install doors, moldings, woodwork, cabinets, and built-in shelving, and lay the floor coverings. The best order for these tasks will depend on the materials you're using and the desired decorative effects.

Install any new plumbing fixtures you have chosen for bathrooms, and complete the drain and supply hookups. Make electrical connections, and install all fixtures, devices, and appliances. Get a final inspection from the building inspector.

Joint Compound

Working Safely

Your personal safety when working on carpentry projects depends greatly on what safety measures you take. The power tools sold today offer many safety features, such as blade guards, locks to prevent accidental starts, and double insulation to reduce the risk of shock in the event of a short circuit. It's up to you to take advantage of these safety features. For example, never operate a saw with the blade guard removed. You risk injury from flying debris as well as from being cut by the blade.

Follow all precautions outlined in the owner's manuals for your tools and make sure you protect yourself with safety glasses, earplugs, and a dust mask or respirator to filter out dust and debris.

Keep your work environment clean. A cluttered work area is more likely to result in accidents. Clean your tools and put them away at the end of every work period, and sweep up dust and debris.

Some materials emit dangerous fumes or particles. Keep such materials stored away from heat sources and out of the reach of children; always use these products in a well-ventilated area.

Maintaining safety is an ongoing project. Take the time to update your first-aid kit and evaluate your workspace, tools, and safety equipment on a regular basis. To avoid accidents, repair and replace old and worn-out parts before they break.

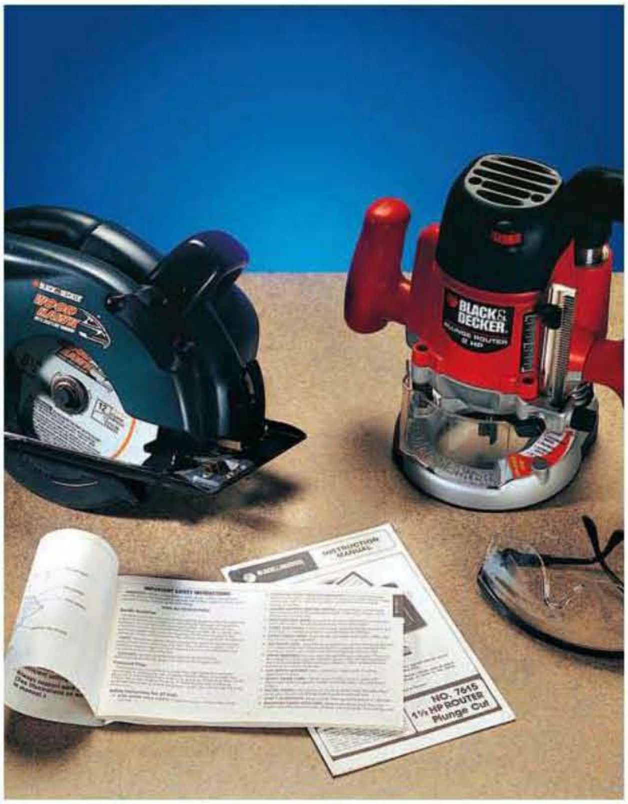

Read the owner's manual before operating any power tool. Your tools may differ in many ways from those described in this book, so it's best to familiarize yourself with the features and capabilities of the tools you own. Always wear eye and ear protection when operating a power tool. Wear a dust mask when the project will produce dust.

Some walls may contain asbestos. Many homes built or remodeled between 1930 and 1950 have older varieties of insulation that included asbestos. Consult a professional for removal of hazardous pollutants like asbestos, and if you find asbestos or materials that may contain asbestos, do not attempt to remove them on your own. Even if you determine that no asbestos is present, it is a good idea to wear a particle mask and other safety gear when doing demolition.

First-Aid Kits

Assemble a first-aid kit. Cuts from hand or power tools can be serious and require prompt and thoughtful attention. Be prepared for such situations with a well-equipped first-aid kit that is easy to find. Record any emergency telephone numbers on the first-aid kit or by the nearest phone so they are available in an emergency.

Equip your kit with a variety of items (photo right), including bandages, needles, tweezers, antiseptic ointment, cotton swabs, cotton balls, eye drops, a first-aid handbook, a chemical-filled cold pack, elastic bandages, first-aid tape, and sterile gauze.

For puncture wounds, cuts, burns, and other serious injuries, always seek medical attention as soon as first aid—such as washing and wrapping of cuts—has been provided.

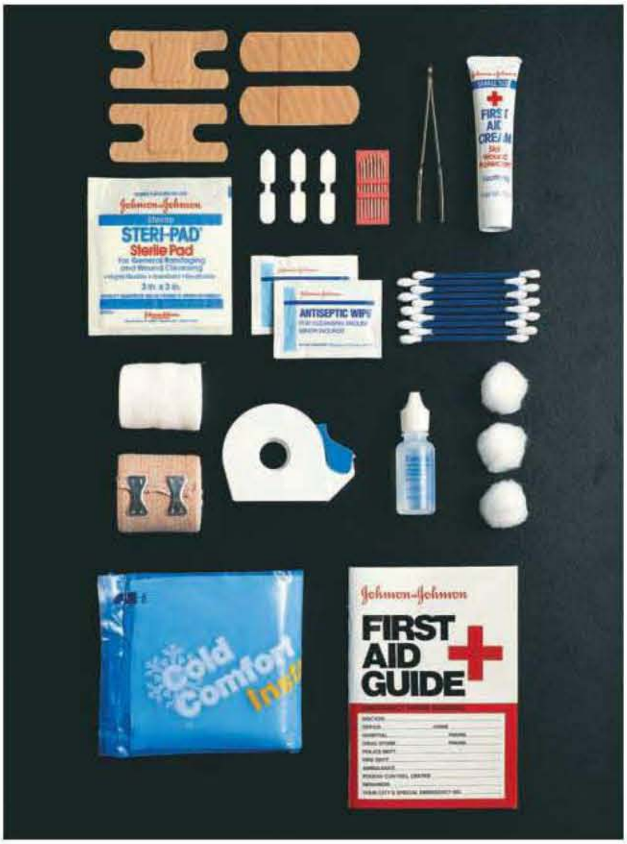

Safe Practices

Keep your tools sharp and clean. Accidents are more likely when blades are dull and tools are filled with sawdust and dirt.

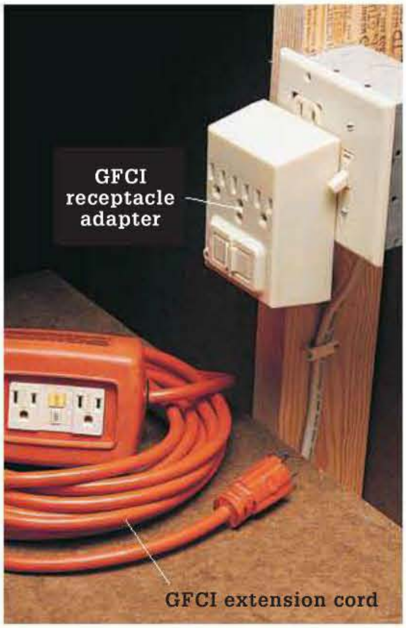

Use a GFCI receptacle, adapter, or extension cord to reduce the risk of shock while operating a power tool outdoors or in wet conditions.

Check with a neon circuit tester to make sure the power is off before removing cover plates, exposing wires, or drilling or cutting into walls that contain wiring.

Techniques

While every home improvement project is different from the others in some way, the basic skills you need come from a common set of DIY techniques. Building walls, running plumbing lines, and adding electrical circuits are some of the chief techniques you'll need to call upon. In this chapter you'll find a condensed version of the essential information you need to put these valuable techniques into practice and become an accomplished do-it-yourselfer.

Once the new walls are up (but before they're covered) you need to fill them with pipes and wires. In the plumbing section that follows you'll learn how to plan new lines and hookups as well as how to work with basic plumbing materials. In the wiring section you'll find useful instructions on working with electrical cables and devices. A handy series of circuit maps shows you exactly how your remodeled room should be wired.

In this chapter:

- Demolition
- Building Walls
- Installing Wallboard
- Understanding Plumbing
- Working with Copper Pipe
- Working with Plastic Pipe
- Working with PEX Pipe
- Planning New Plumbing
- Understanding Wiring
- Working with Wire & Cable
- Working with Conduit
- Working with Boxes
- Working with Switches
- Working with Receptacles
- Planning New Wiring
- Using Circuit Maps

Demolition

Many home remodeling projects actually begin with demolition. When you're remodeling, it's often necessary to cut or enlarge openings for new doors or windows or even to remove entire walls. The basic procedures for this type of demolition are the same whether you're working with doors and windows on exterior walls or altering interior walls.

Your first step will be to determine how your house was framed. House framing variations will dictate the proper procedures for creating openings in walls or removing walls altogether. Then, you'll need to inspect the walls for hidden mechanicals—wiring, plumbing, and HVAC lines.

After you've rerouted any utility lines, you're ready to remove the interior wall surfaces. If you're replacing old windows and doors, now is the time to remove them as well. Where necessary, you can now remove exterior wall surfaces, but don't remove any framing members yet.

The next step will depend on the nature of your project. If you are removing a load-bearing wall or creating a new or enlarged opening in one, you'll need to build temporary supports to brace the ceiling while the work is being done. This step won't be necessary if you are removing a non-loadbearing wall. Then you can remove any wall framing members, following the applicable procedures for load-bearing or non-loadbearing walls.

With the removal steps of the project completed, you'll be ready to install your new windows and doors.

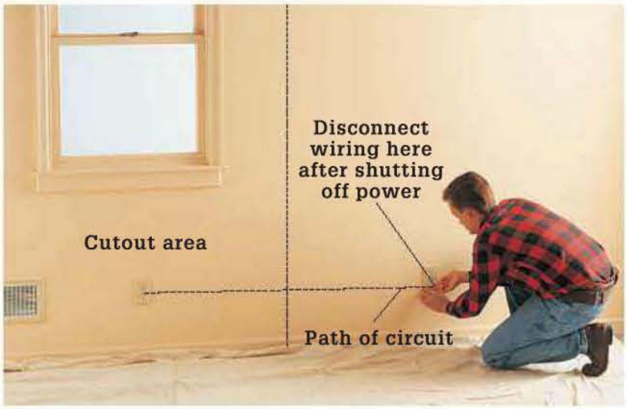
Disconnect wiring here after shutting off power
Cutout area
Path of circuit

Disconnect electrical wiring before you cut into walls. Trace the wiring back to a fixture outside the cutout area, then shut off the power and disconnect the wires leading into the cutout area. Turn the power back on and check for current with a circuit tester before cutting into the walls.

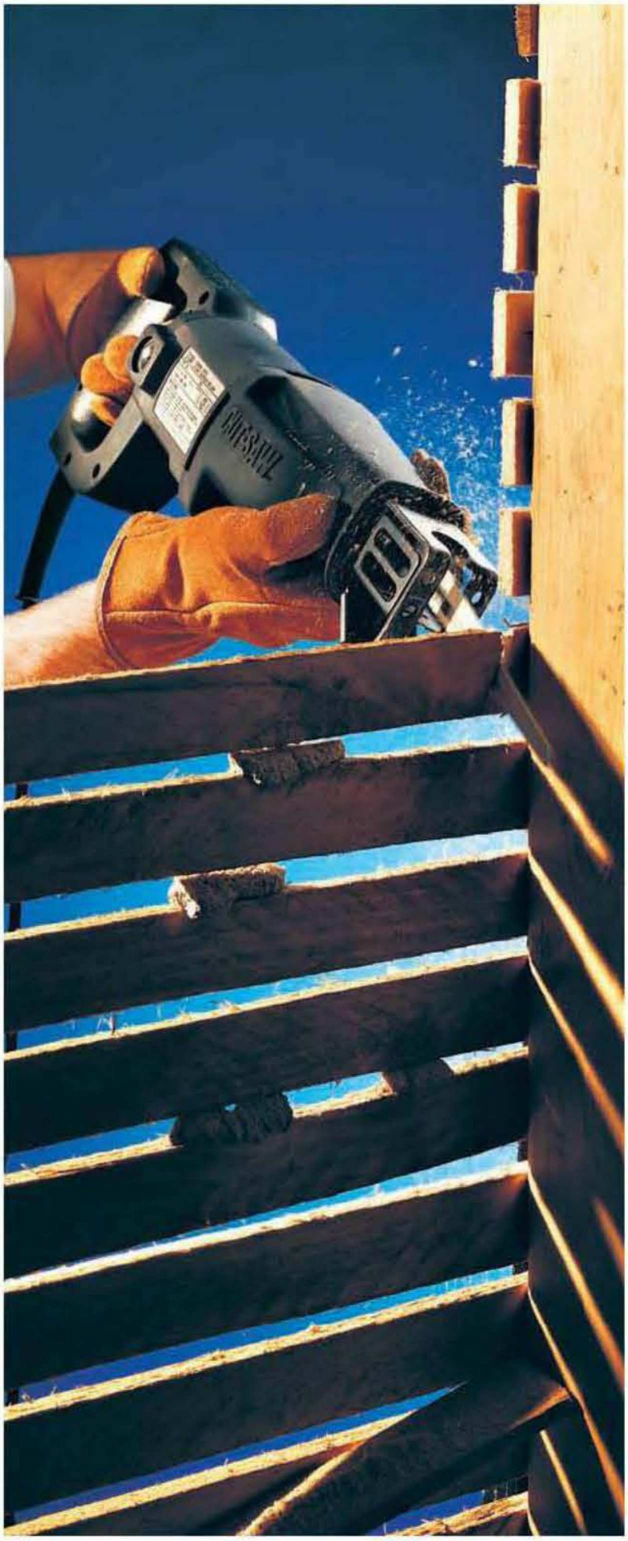

Demolition is the starting point of most projects where a window or door opening is enlarged.

How to Install Temporary Support (perpendicular joists)

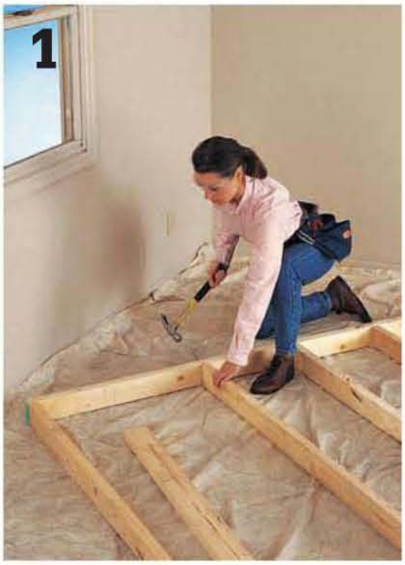

Build a 2 × 4 stud wall that is 4 ft. wider than the planned wall opening and 1¾" shorter than the distance from floor to ceiling.

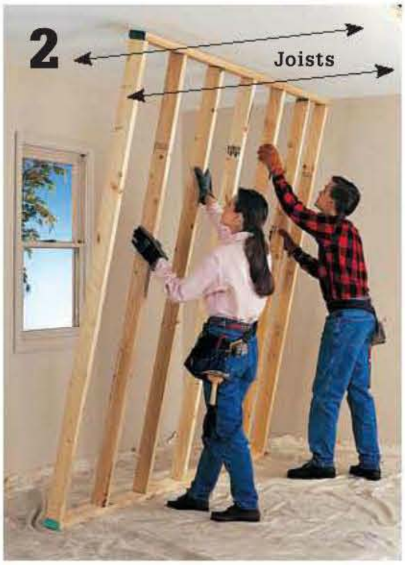

Raise the stud wall up and position it 3 ft. from the wall, centered on the planned rough opening.

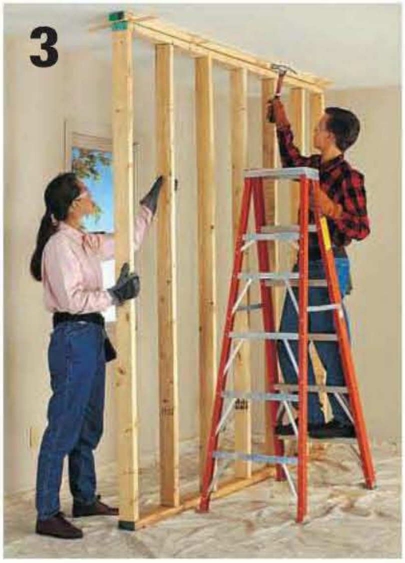

Slide a 2 × 4 top plate between the temporary wall and the ceiling. Check to make sure the wall is plumb, and drive shims under the top plate at 12" intervals until the wall is wedged tightly in place.

How to Install Temporary Support (parallel joists)

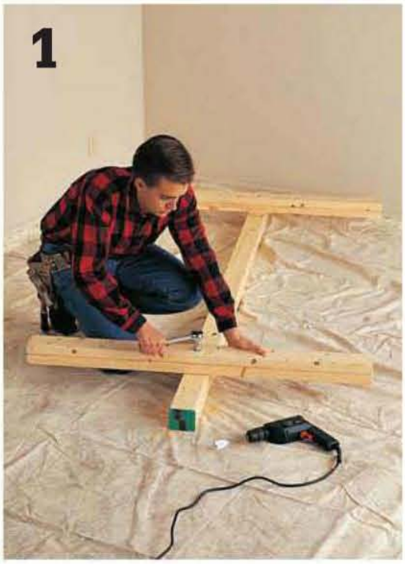

Build two 4-ft.-long cross braces, using pairs of 2 × 4s nailed together. Attach the cross braces to the double top plate, 1 ft. from the ends, using countersunk lag screws.

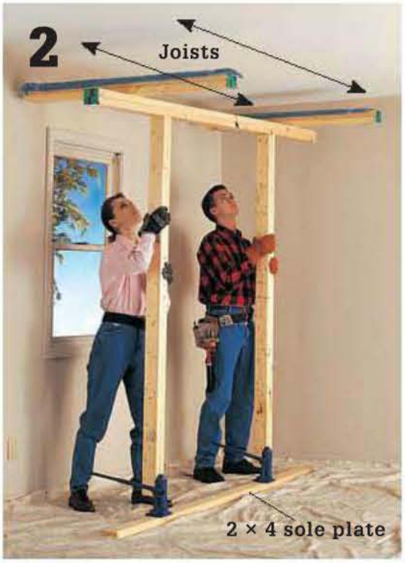

Place a 2 × 4 sole plate directly over a floor joist, then set hydraulic jacks on the sole plate. For each jack, build a post 8" shorter than the jack-to-ceiling distance. Nail the posts to the top plate, 2 ft. from the ends. Cover the braces with the cloth, and set the support structure on the jacks.

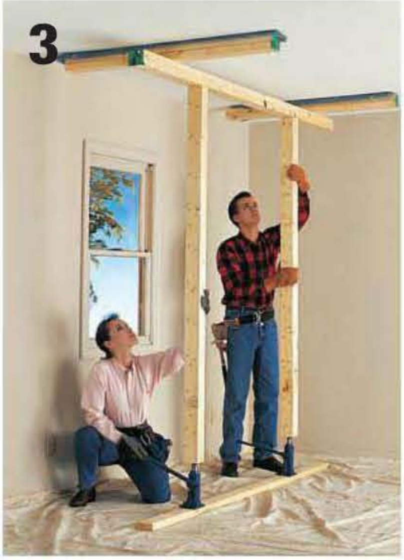

Adjust the support structure so the posts are exactly plumb, and pump the hydraulic jacks until the cross braces just begin to lift the ceiling. Do not lift too far or you may damage the ceiling or floor.

How to Remove Wallboard

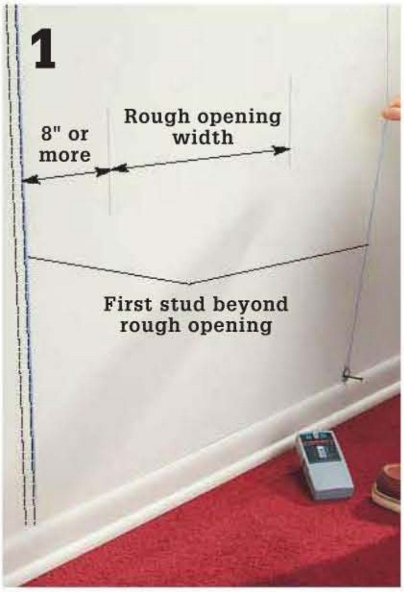

Mark the width of the rough opening on the wall and locate the first stud on either side of the planned rough opening. If the rough opening is more than 8" from the next stud, use a chalk line to mark a cutting line on the inside edge of the stud.

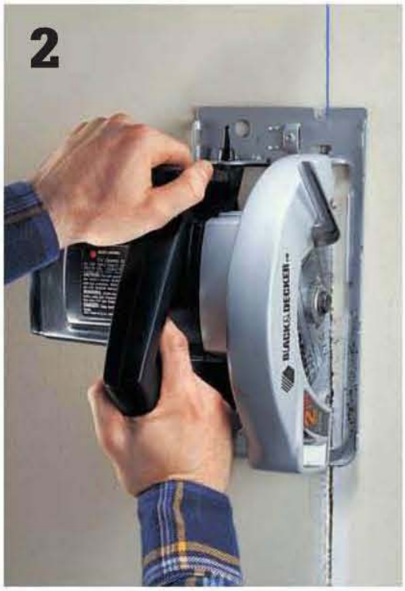

Remove the baseboards and other trim, and prepare the work area. Make a ¾"-deep cut from floor to ceiling along both cutting lines, using a circular saw. Use a utility knife to finish the cuts at the top and bottom and to cut through the taped horizontal seam where the wall meets the ceiling surface.

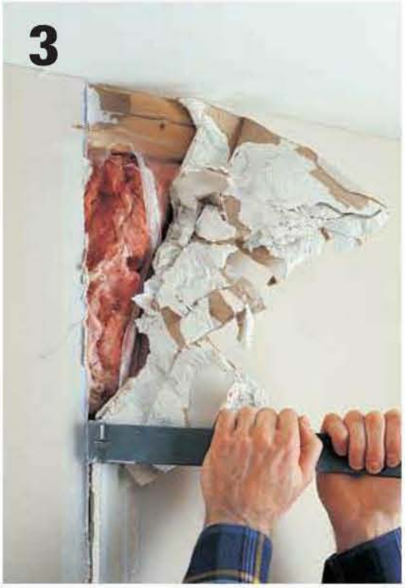

Insert the end of a pry bar into the cut near a corner of the opening. Pull the pry bar until the wallboard breaks, then tear away the broken pieces. Take care to avoid damaging the wallboard outside the project area.

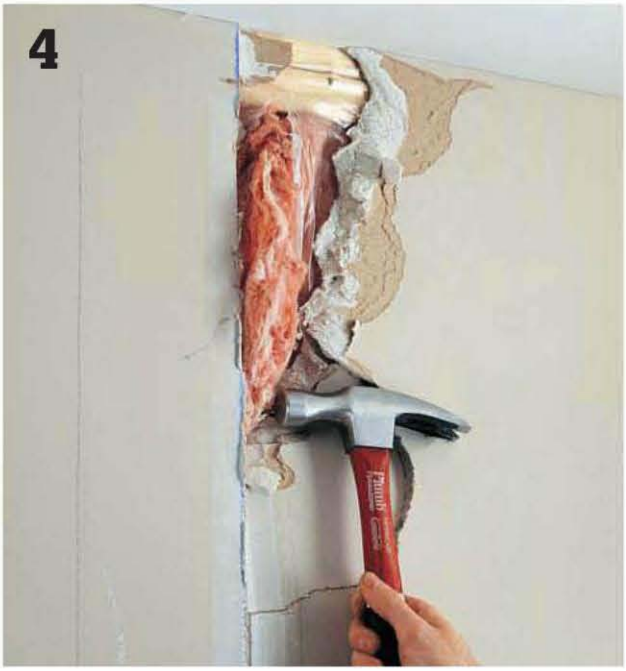

Continue removing the wallboard by striking the surface with the side of a hammer and pulling the wallboard away from the wall with the pry bar or your hands.

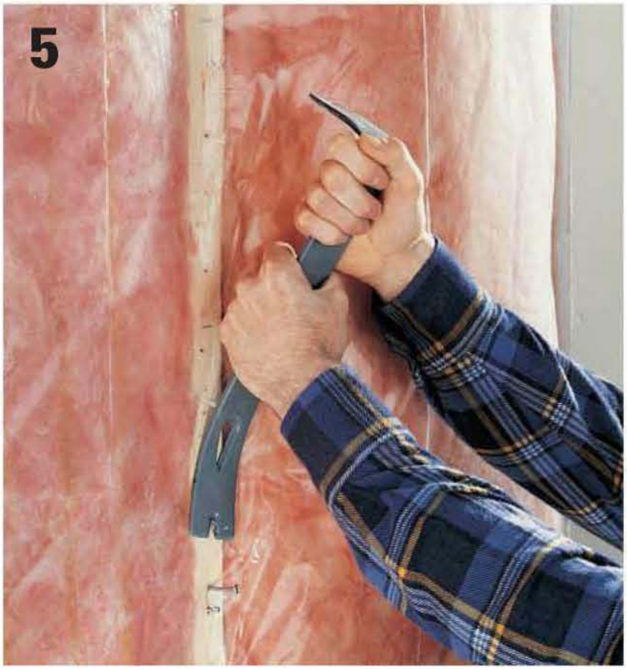

Remove nails, screws, and any remaining wallboard from the framing members, using a pry bar. Remove any vapor barrier and insulation.

How to Remove Plaster Walls

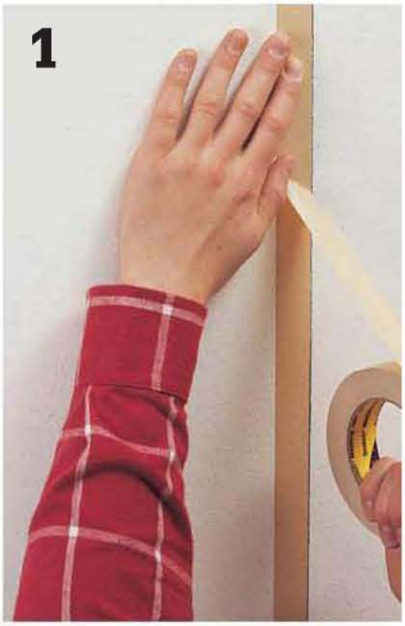

Shut off the power and inspect the wall for wiring and plumbing. Mark the wall area to be removed by following the directions on page 30. Apply a double layer of masking tape along the outside edge of each cutting line.

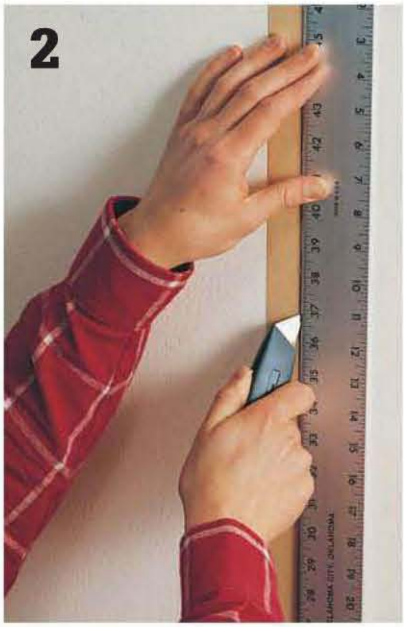

Score each line several times with a utility knife, using a straightedge as a guide. Scored lines should be at least ⅛" deep.

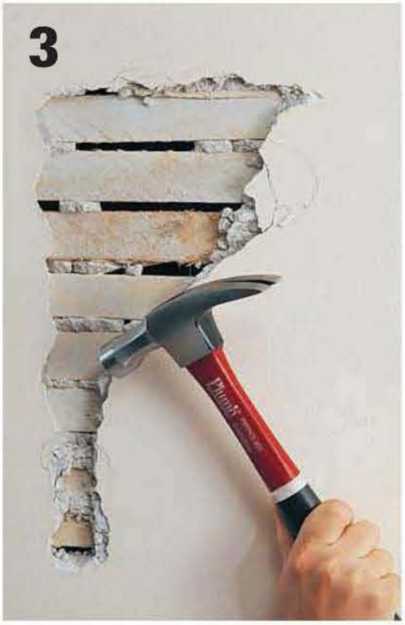

Beginning at the top of the wall in the center of the planned opening, break up the plaster by striking the wall lightly with the side of a hammer. Clear away all plaster from floor to ceiling to within 3" of the marked lines.

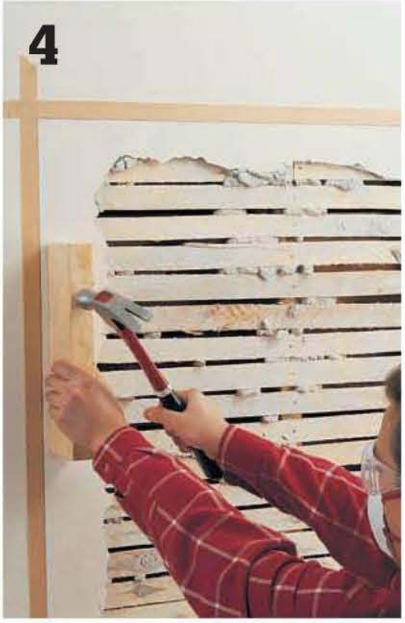

Break the plaster along the edges by holding a scrap piece of 2 × 4 on edge just inside the scored line and rapping it with a hammer. Use a pry bar to remove the remaining plaster.

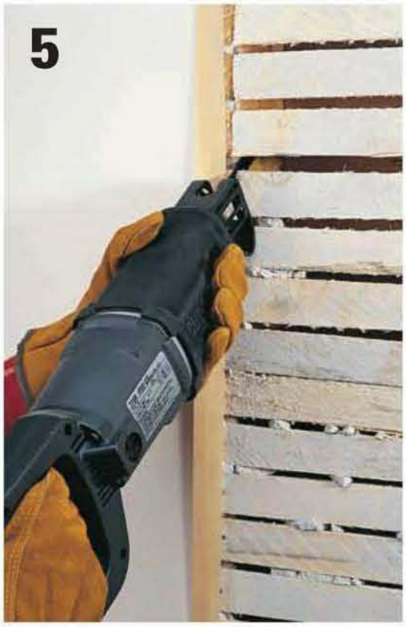

Cut through the lath along the edges of the plaster, using a reciprocating saw or jigsaw.

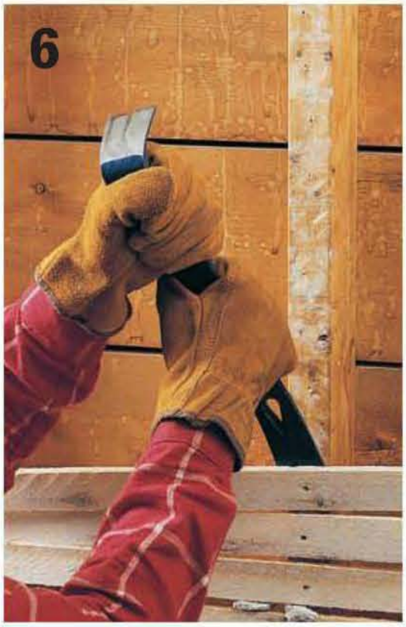

Remove the lath from the studs, using a pry bar. Pry away any remaining nails, and remove any vapor barrier and insulation.

How to Make an Opening in an Exterior Wall

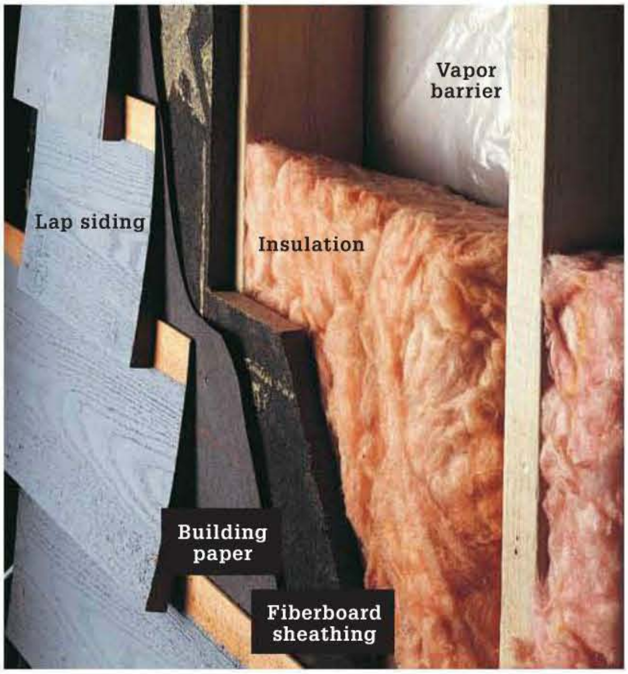

A cross-section view of a typical exterior wall with lap siding.

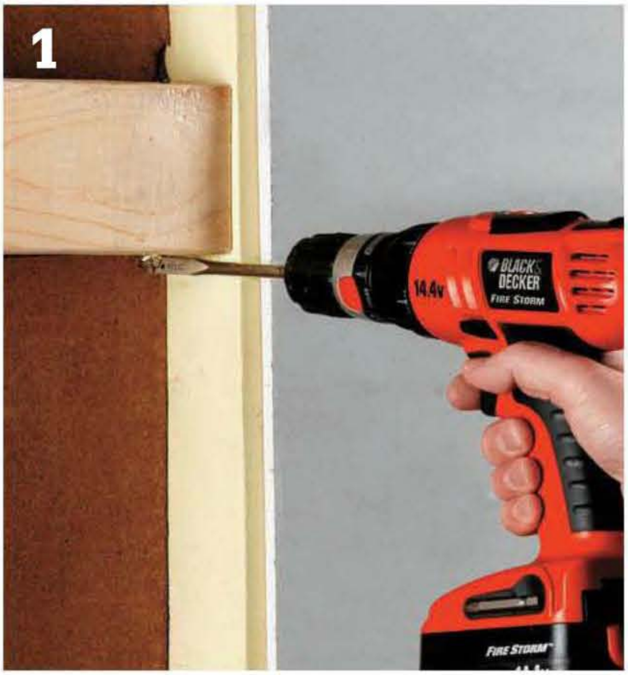

From inside the house, drill through the wall at the corners of the framed opening. Push casing nails through the holes to mark their location. For round-top windows, drill holes around the curved outline (see variation, page 195).

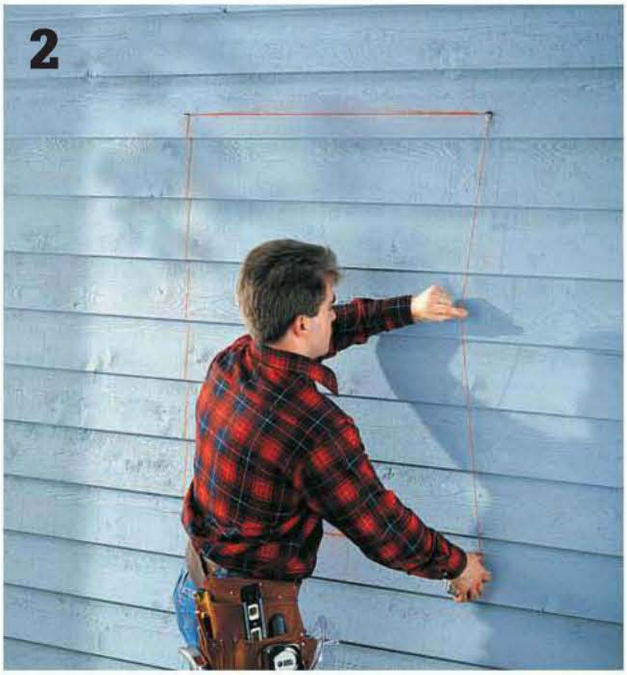

Measure the distance between the nails on the outside of the house to make sure the dimensions are accurate. Mark the cutting lines with a chalk line stretched between the nails. Push the nails back through the wall.

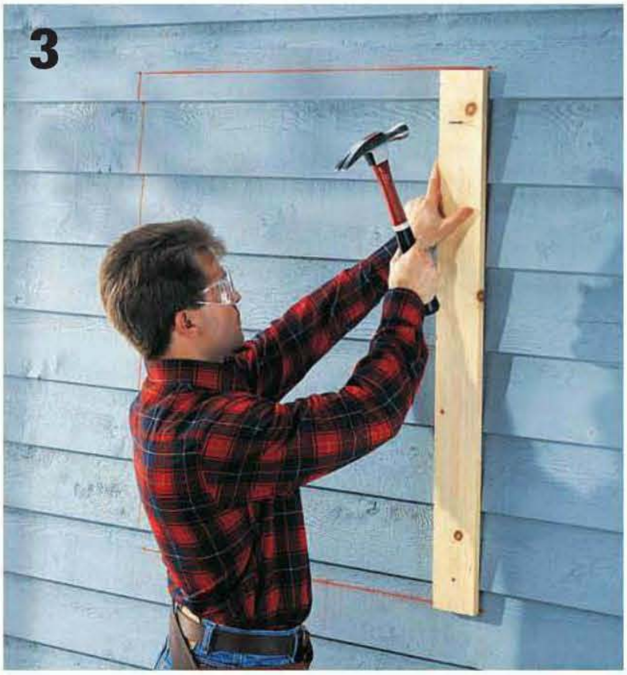

Nail a straight 1 × 4 flush with the inside edge of the right cutting line. Sink the nail heads with a nail set to prevent scratches to the foot of the saw. Set the circular saw to its maximum blade depth.

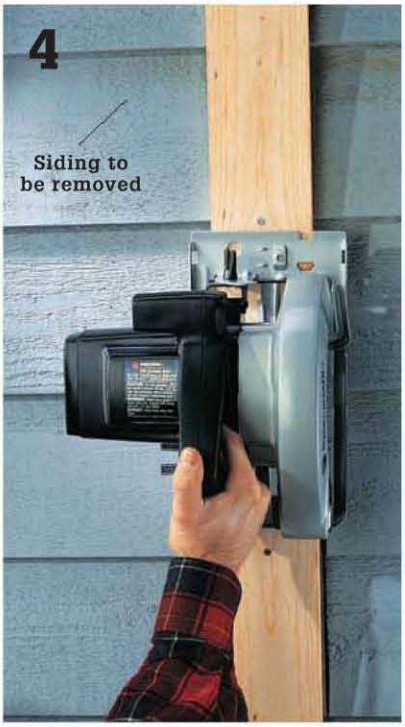

Rest the saw on the 1 × 4, and cut along the marked line, using the edge of the board as a guide. Stop the cuts about 1" short of the corners to keep from damaging the framing members.

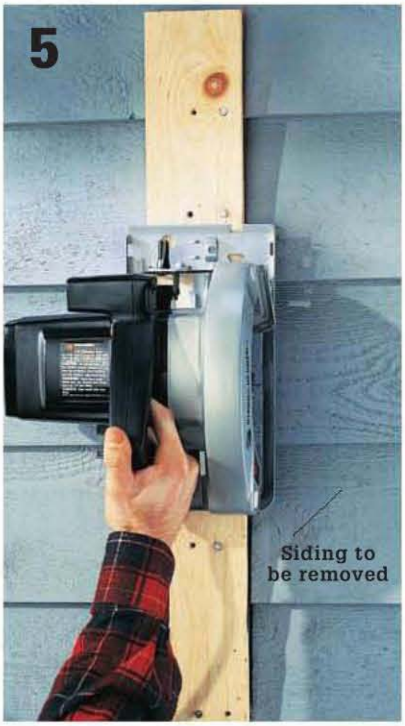

Reposition the 1 × 4, and make the remaining straight cuts. Drive nails within 1½" of the inside edge of the board, because the siding under this area will be removed to make room for door or window brick moldings.

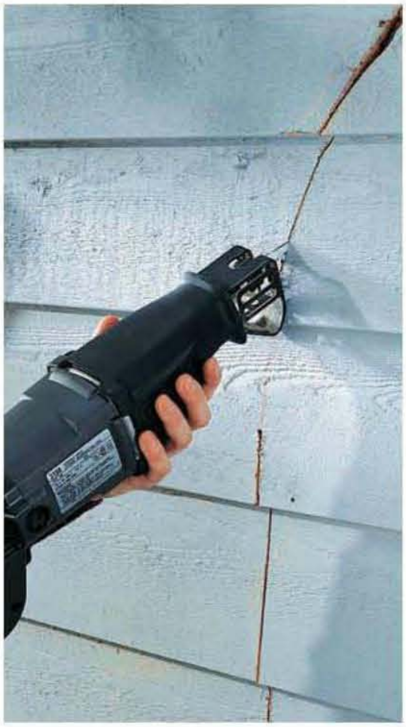

Variation: For round-top windows, make curved cuts using a reciprocating saw or jigsaw. Move the saw slowly to ensure smooth, straight cuts. To draw an outline for round-top windows, use a cardboard template.

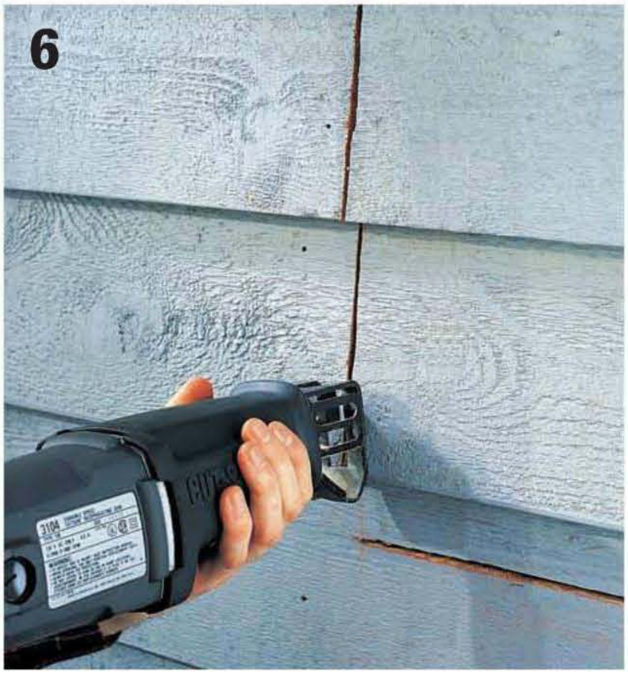

Complete the cuts at the corner with a reciprocating saw or jigsaw.

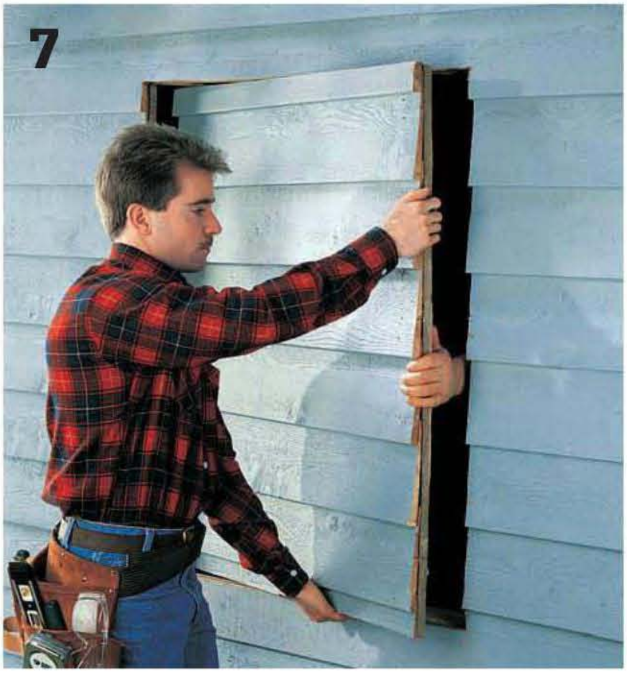

Remove the cut wall section. If you are working with metal siding, wear work gloves. If you wish, remove the siding pieces from the sheathing and save them for future use.

How to Make an Opening in a Stucco Wall

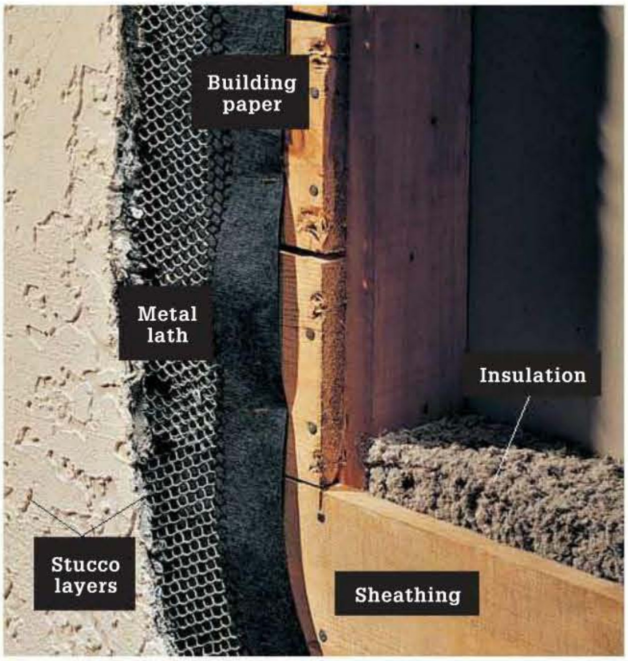

Stucco is a multiple-layer cement product applied to metal lath. Building paper is sandwiched between the metal lath and the sheathing to create a waterproof barrier. Stucco is extremely durable due to its cement base. But if you don't do the removal carefully, it's easy to crack the stucco past the outline for the new window or door.

From inside the house, drill through the wall at the corners of the framed opening. Use a twist bit to drill through the sheathing, then use a masonry bit to finish the holes. Push casing nails through the holes to mark their locations.

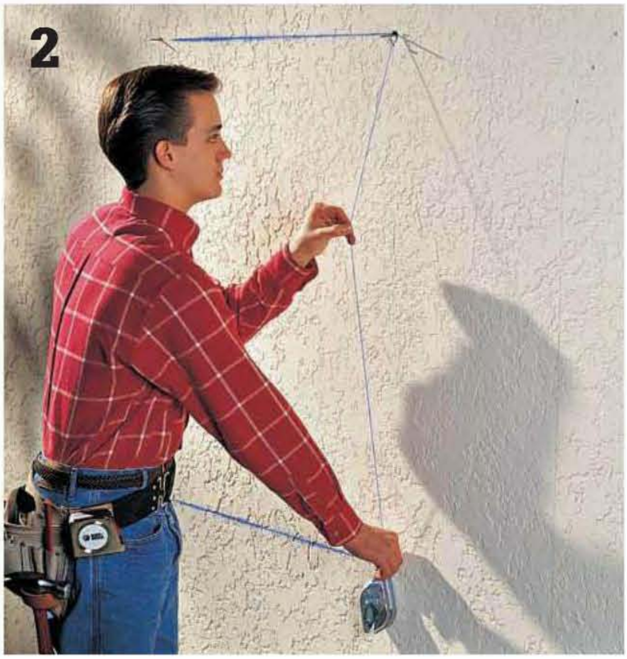

On the outside wall, measure the distance between the nails to make sure the rough opening dimensions are accurate. Mark cutting lines between the nails, using a chalk line.

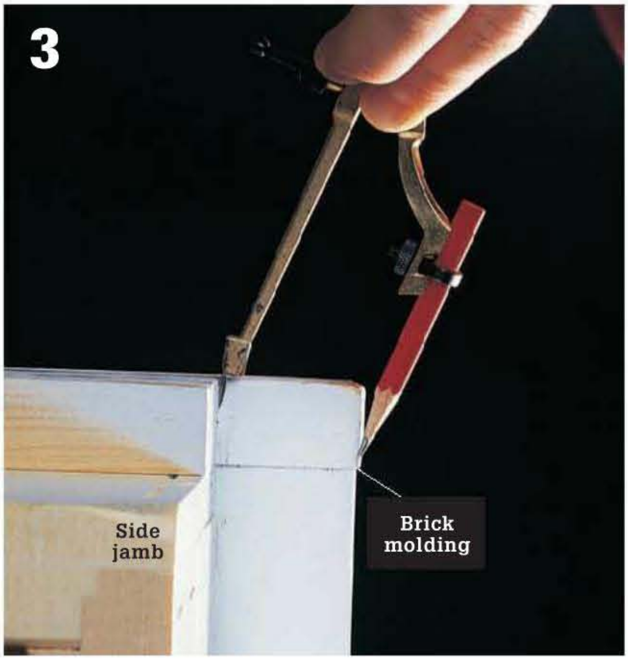

Match the distance between the side jambs and the edge of the brick molding on a window or door with the legs of a compass.

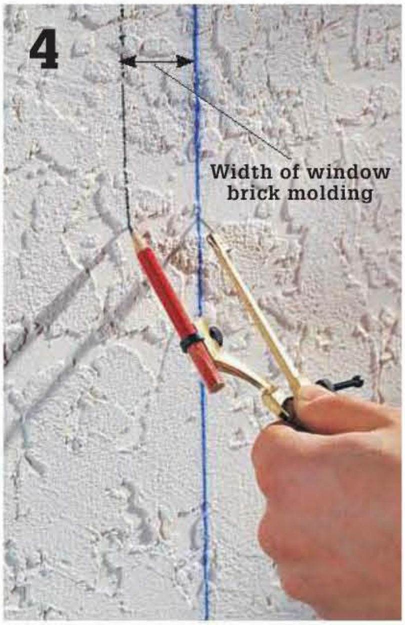

Scribe a cutting line on the stucco by moving the compass along the outline, with the compass point held on the marked line. This added margin will allow the brick molding to fit tight against the wall sheathing.

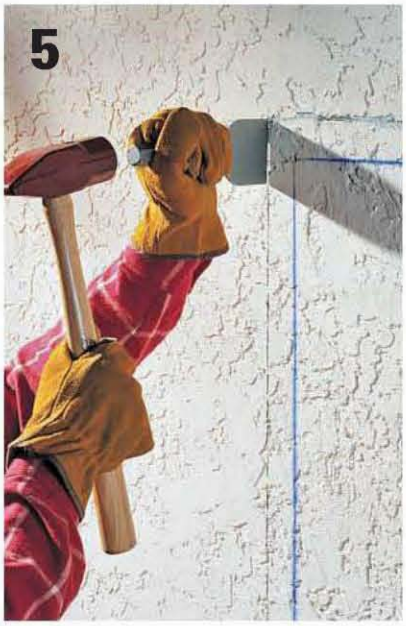

Score the stucco surface around the outside edge of the scribed line, using a masonry chisel and masonry hammer. The scored grooves should be at least ⅛" deep to serve as a guide for the circular saw blade.

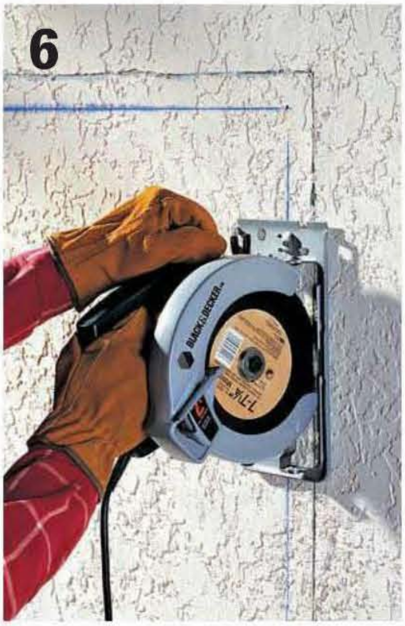

Make straight cuts using a circular saw and masonry-cutting blade. Make several passes with the saw, gradually deepening the cuts until the blade just cuts through the metal lath, causing sparks to fly. Stop cuts just ahead of the corners to avoid damaging the stucco past the cutting line; complete the cuts with a masonry chisel.

Variation: For round-top windows, mark the outline on the stucco, using a cardboard template, and drill a series of holes around the outline, using a masonry bit. Complete the cut with a masonry chisel.

Break up the stucco with a masonry hammer or sledgehammer, exposing the underlying metal lath. Use aviation snips to cut through the lath around the opening. Use a pry bar to pull away the lath and attached stucco.

Outline the rough opening on the sheathing, using a straightedge as a guide. Cut the rough opening along the inside edge of the framing members, using a circular saw or reciprocating saw. Remove the cut section of sheathing.

How to Remove a Partition Wall

Use a utility knife to score the intersections where the wall you're removing meets the ceiling to keep from damaging it during wall removal. Pry away baseboard trim and remove receptacle plates and switch covers to prepare for demolition.

Use the side of a hammer to punch a starter hole in the wallboard, then carefully remove the wallboard with a pry bar. Try to pull off large sections at a time to minimize dust. Remove any remaining wallboard nails or screws from the wall studs.

Reroute outlets, switches, plumbing, or ductwork. Have professionals do this for you if you are not experienced with these systems or confident in your skills. This work should be inspected after it is completed.

Locate the closest permanent studs on the adjacent wall or walls with a stud finder, and carefully remove the wallboard up to these studs. Score the wallboard first with a utility knife, then cut through it with a circular saw.

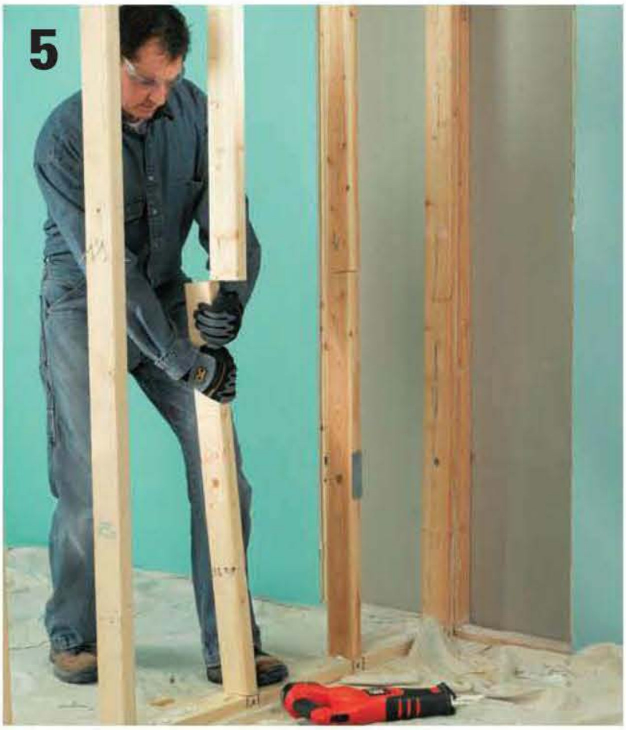

Remove the wall studs by cutting through them in the middle with a reciprocating saw and prying out the upper and lower sections. Remove the endmost studs where the wall meets an adjacent wall or walls.

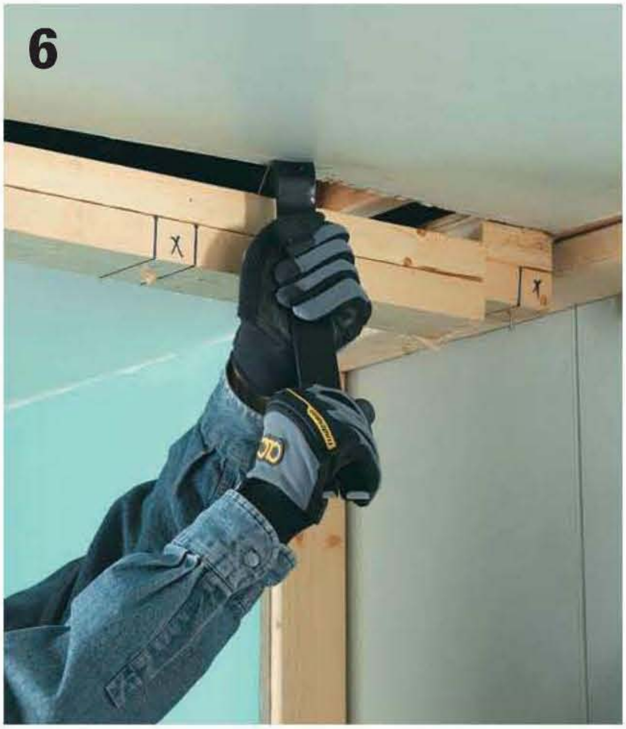

Cut through the wall's top plate with a circular saw or reciprocating saw. Pry out the top plate sections carefully to avoid damaging the ceiling.

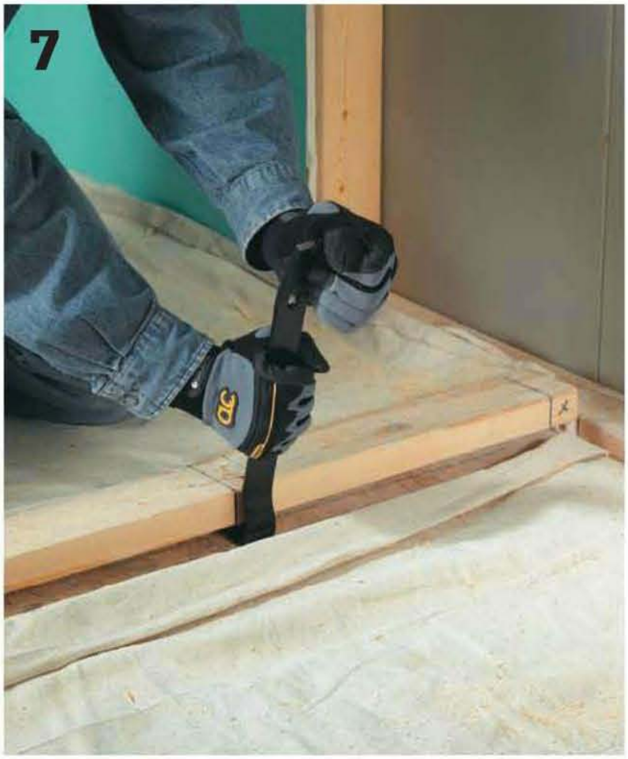

Remove the sole plate just as you did the top plate by cutting through it and prying up the long pieces.

Patch the walls and ceiling with strips of wallboard, and repair the floor as needed with new floor coverings.

Removing Trim

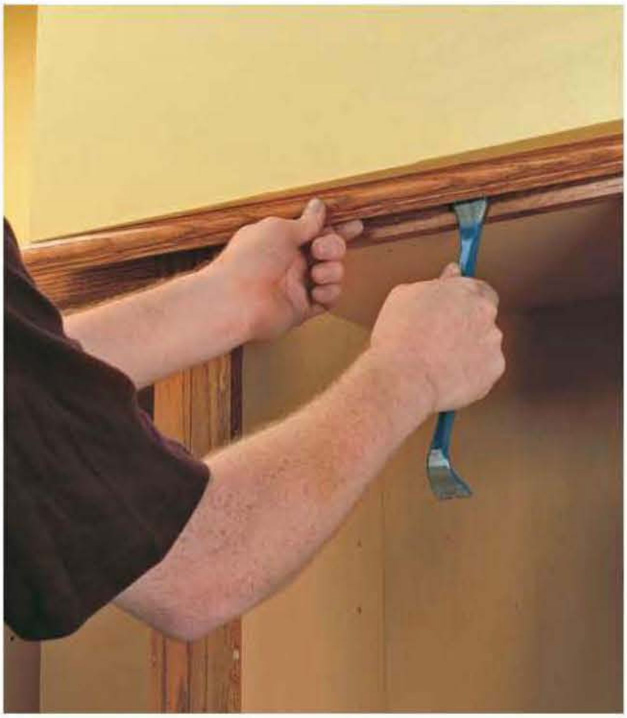

Remove trim moldings at the edges and tops of the cabinets with a flat pry bar or putty knife.

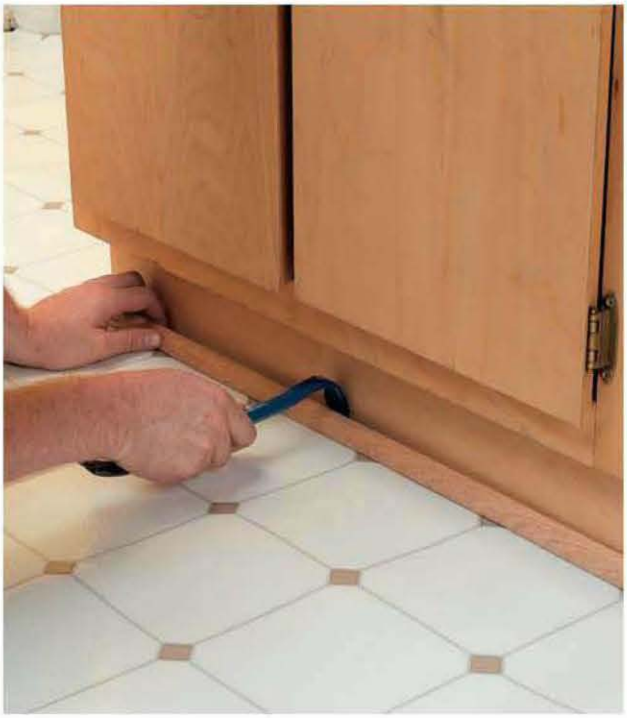

Remove base shoe from cabinet base if the molding is attached to the floor.

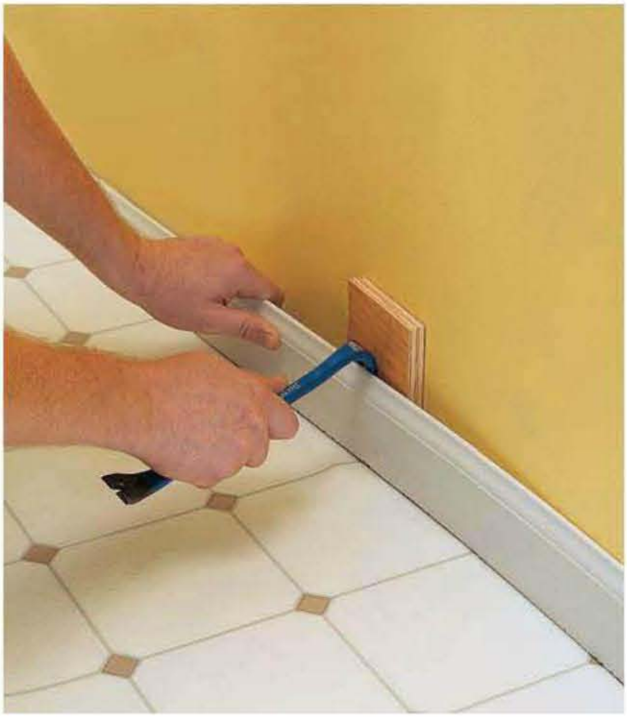

Remove baseboards and other trim moldings with a pry bar. Protect wall surfaces with scraps of wood.Label the trim boards on the back side so you can replace them correctly.

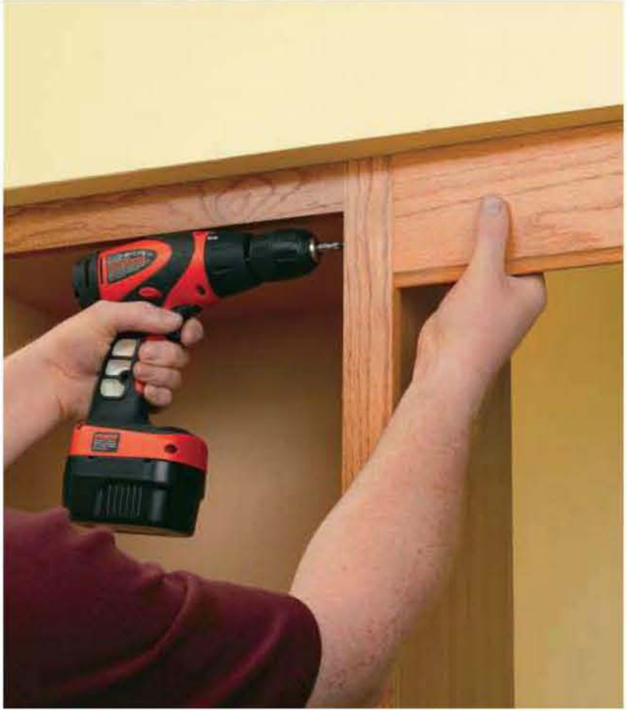

Remove valances above cabinets. Some valances are attached to the cabinets or soffits with screws. Others are nailed and must be pried loose.

How to Remove Cabinets

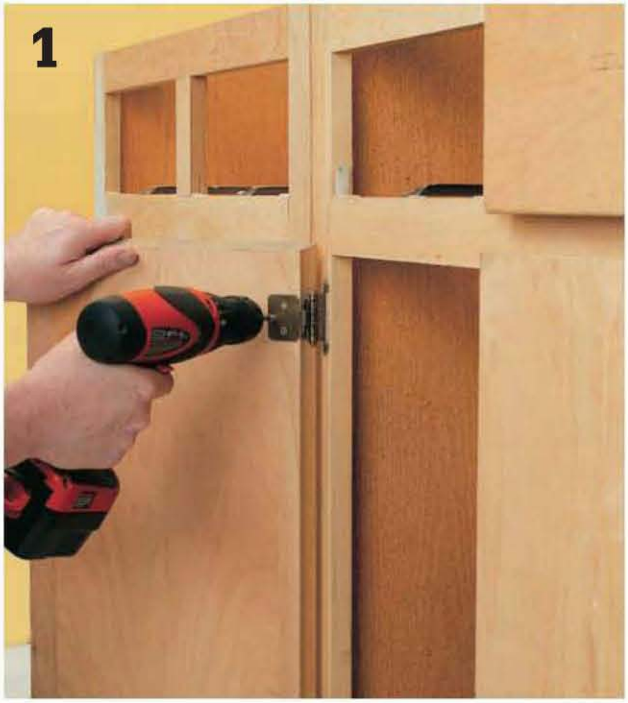

Remove doors and drawers to make it easier to get at interior spaces. You may need to scrape away old paint to expose hinge screws.

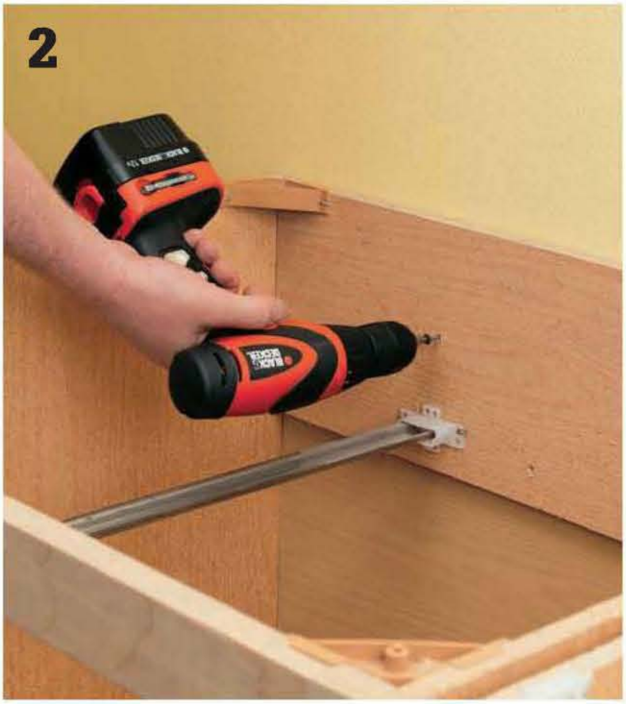

At the backs of cabinets, remove any screws holding the cabinet to the wall. Cabinets can be removed as a group, or can be disassembled.

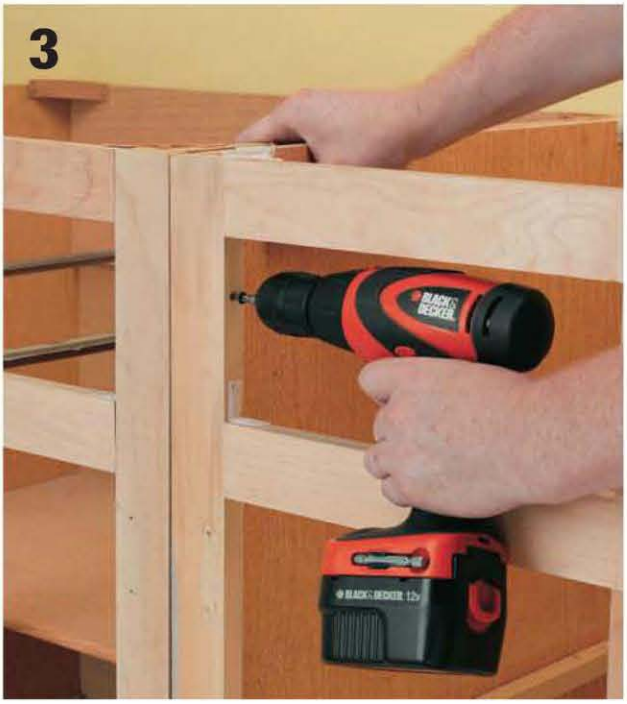

Detach individual cabinets by removing screws that hold face frames together.

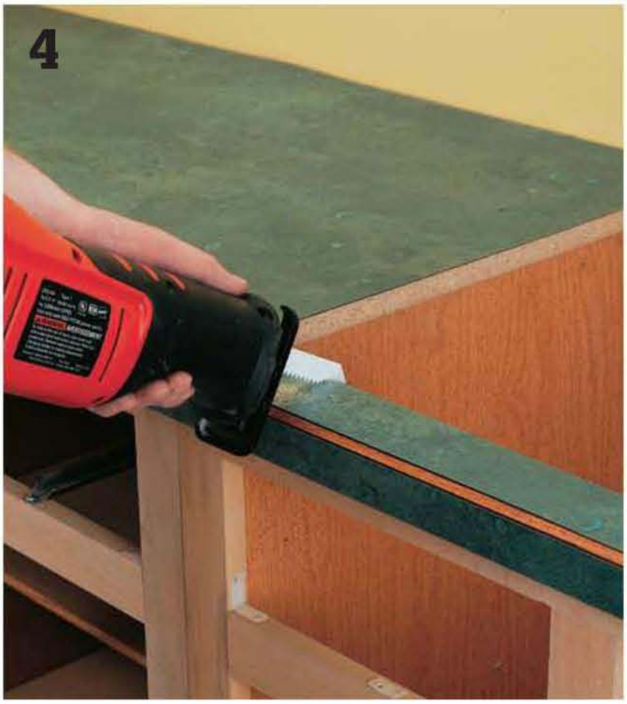

Countertops are usually not salvageable. Cut them into manageable pieces with a reciprocating saw, or take them apart piece by piece with a hammer and pry bar.

Building Walls

Partition walls are constructed between load-bearing walls to divide space. They should be strong and well made, but their main job is to house doors and to support wall coverings. The sole plate at the bottom of the wall should be fastened securely to the subfloor and joists with panel adhesive and 16d common nails. Use masonry fasteners on concrete basement floors.

Anchoring New Partition Walls

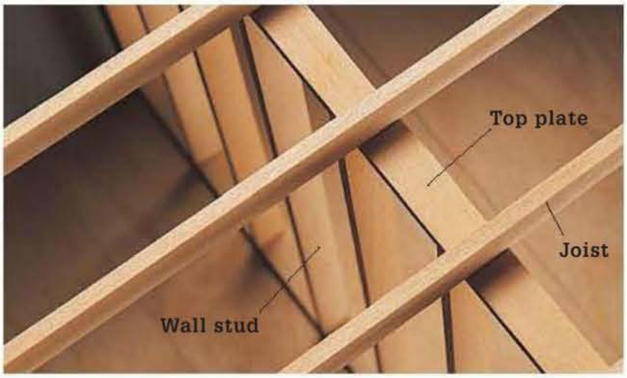

When a new wall is perpendicular to the ceiling or floor joists above, attach the top plate directly to the joists, using 16d nails.

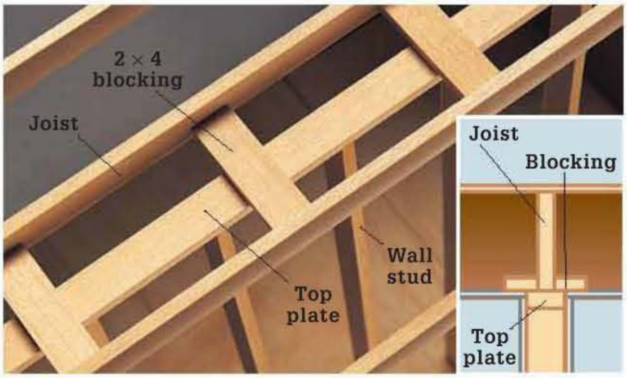

When a new wall falls between parallel joists, install 2 × 4 blocking between the joists every 24". If the new wall is aligned with a parallel joist, install blocks on both sides of the wall, and attach the top plate to the joist (inset).

How to Build a Partition Wall

Mark the location of the new wall on the ceiling, then snap two chalk lines or use a scrap piece of 2× lumber as a template to mark layout lines for the top plate. Use a stud finder to locate floor joists or roof framing above the ceiling, and mark these locations with tick marks or tape outside the layout lines.

Cut the top and sole plates to length and lay them side by side. Use a combination or framing square to draw pairs of lines across both plates to mark the stud locations. Space the studs at 16" intervals, on center.

Mark the location of any door framing on the top and sole plates. Refer to the door's rough opening specifications when marking the layout. Draw lines for both the king and jack studs.

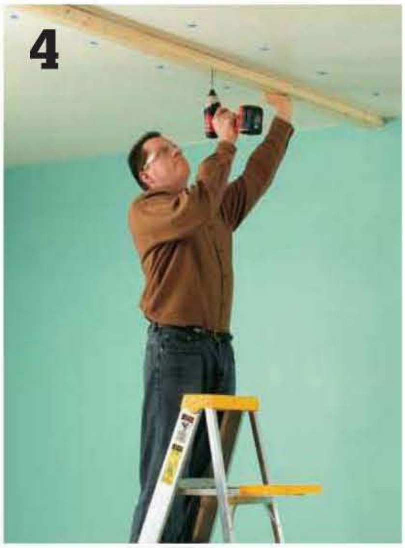

Fasten the top plate to the ceiling using 3" deck screws or 10d nails. Be sure to orient the plate so the stud layout faces down.

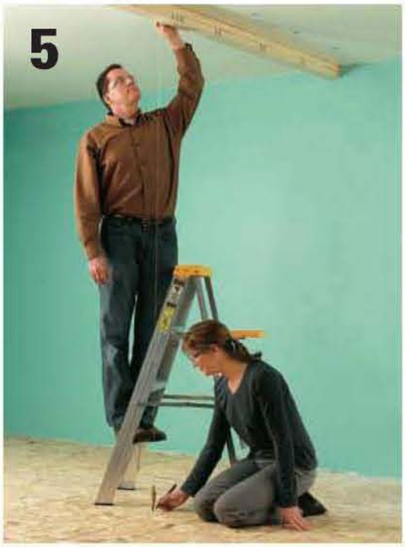

Hang a plumb bob from the edge of the top plate at several points along its length to find the sole plate location on the floor. The tip of the plumb bob should almost touch the floor. Wait until it stops moving before marking the sole plate reference point. Connect the points with a line to establish one edge of the sole plate. Use a piece of scrap 2× material as a template for marking the other edge.

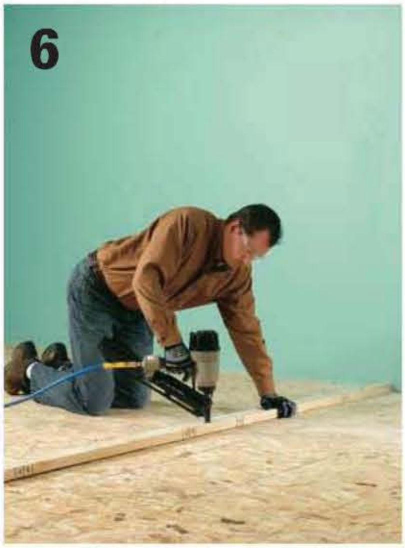

Cut away the portion of the sole plate where the new door will be, and nail or screw the two sections to the floor between the sole plate layout lines. Use the cutaway door section as a spacer for the door when fastening the plates. Drive the fasteners into the floor framing. For concrete floors, attach the sole plate with a powder-actuated nail gun or with hardened masonry screws.

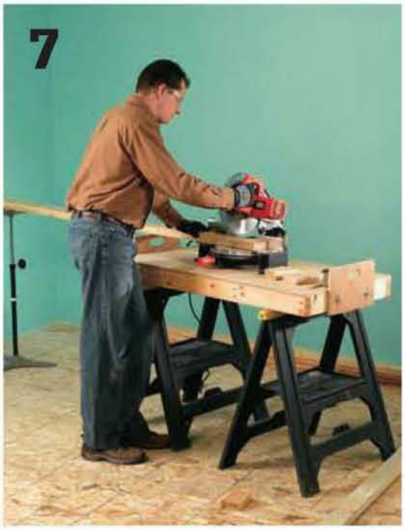

Measure the distance between the top and sole plates at several places along the wall to determine the stud lengths. The stud length distance may vary, depending on structural settling or an out-of-flat floor. Add $1/8$" to the stud length(s), and cut them to size. The extra length will ensure a snug fit between the wall plates.

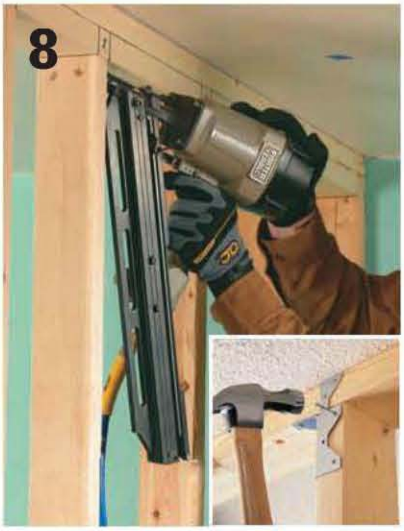

Install the framing members. Use a hammer to tap each stud into position, then toenail the studs to the top and sole plates. Fit, trim, and nail the studs, one at a time. *Tip: If the studs tend to shift during nailing, drive pilot holes for the nails first, or use 3" deck screws instead of nails. (Inset) An option for attaching wall studs to plates is to use metal connectors and 4d nails.*

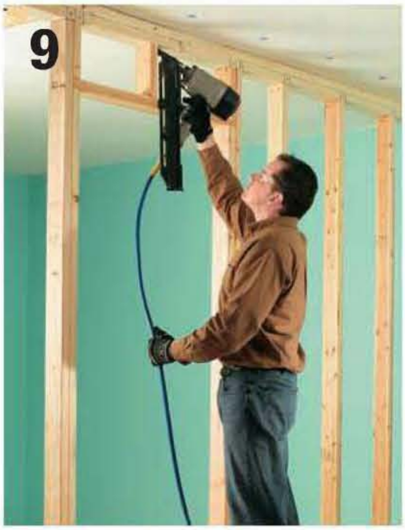

Nail the king studs, jack studs, a header and a cripple stud in place to complete the rough door framing. See page 12 for more information on framing a door opening.

(continued)

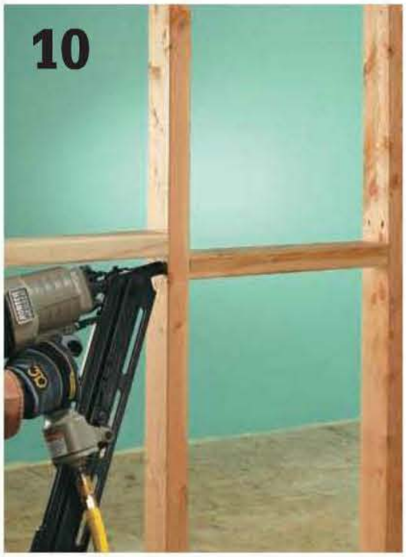

If building codes in your area require fire blocking, install 2× cutoff scraps between the studs, 4 ft. from the floor, to serve this purpose. Stagger the blocks so you can endnail each piece.

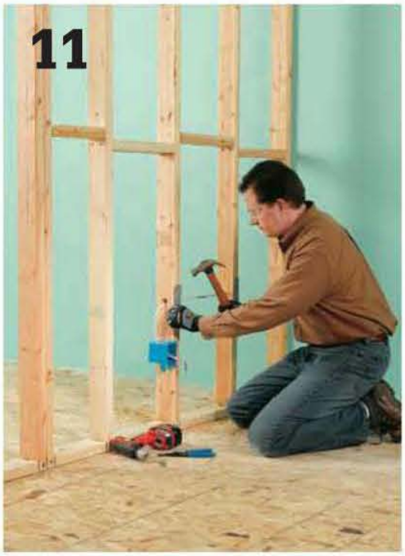

Drill holes through the studs to create raceways for wiring and plumbing. When this work is completed, fasten metal protector plates over these areas to prevent drilling or nailing through wiring and pipes later. Have your work inspected before proceeding with wallboard.

Cover the wall with wallboard. Plan the layout wisely to minimize waste and to avoid butted joints with untapered seams. If you are installing wallboard on a ceiling as well, do that first. For more on installing wallboard, see pages 44 to 51.

Tools & Materials for Framing with Steel

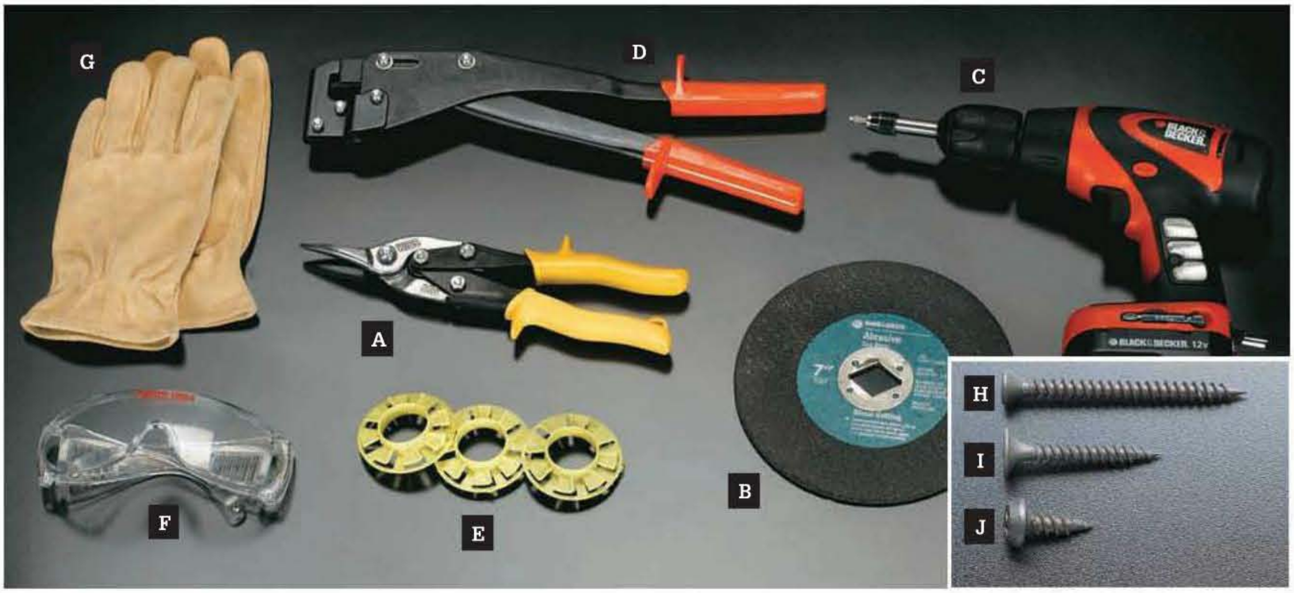

Steel framing requires a few specialty tools and materials. Aviation snips (A) are needed to cut tracks and studs, though a miter saw outfitted with a steel-cutting abrasive blade (B) can speed up the process. A drill or screw gun (C) is required for fastening framing. Handy for large projects, a stud crimper (D) creates mechanical joints between tracks and studs. Plastic grommets (E) are placed in knockouts to help protect utility lines. Protective eyewear and heavy work gloves (F, G) are necessities when working with the sharp edges of hand-cut steel framing. Use self-tapping screws (inset) to fasten steel components. To install wood trim, use type S trim-head screws (H); to fasten wallboard, type S wallboard screws (I); and to fasten studs and tracks together, $\frac{7}{16}$" type S panhead screws (J).

Joining Sections Using Steel Studs ▸

Steel studs and tracks have the same basic structure—a web that spans two flanged sides—however, studs also contain a ¼" lip to improve their rigidity.

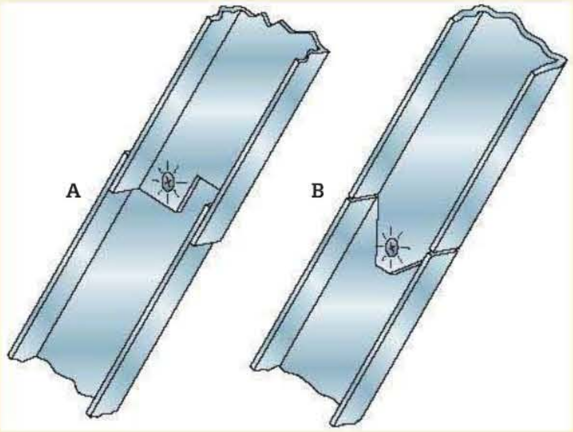

Join sections with a spliced joint (A) or notched joint (B). Make a spliced joint by cutting a 2" slit in the web of one track. Slip the other track into the slit and secure with a screw. For a notched joint, cut back the flanges of one track and taper the web so it fits into the other track; secure with a screw.

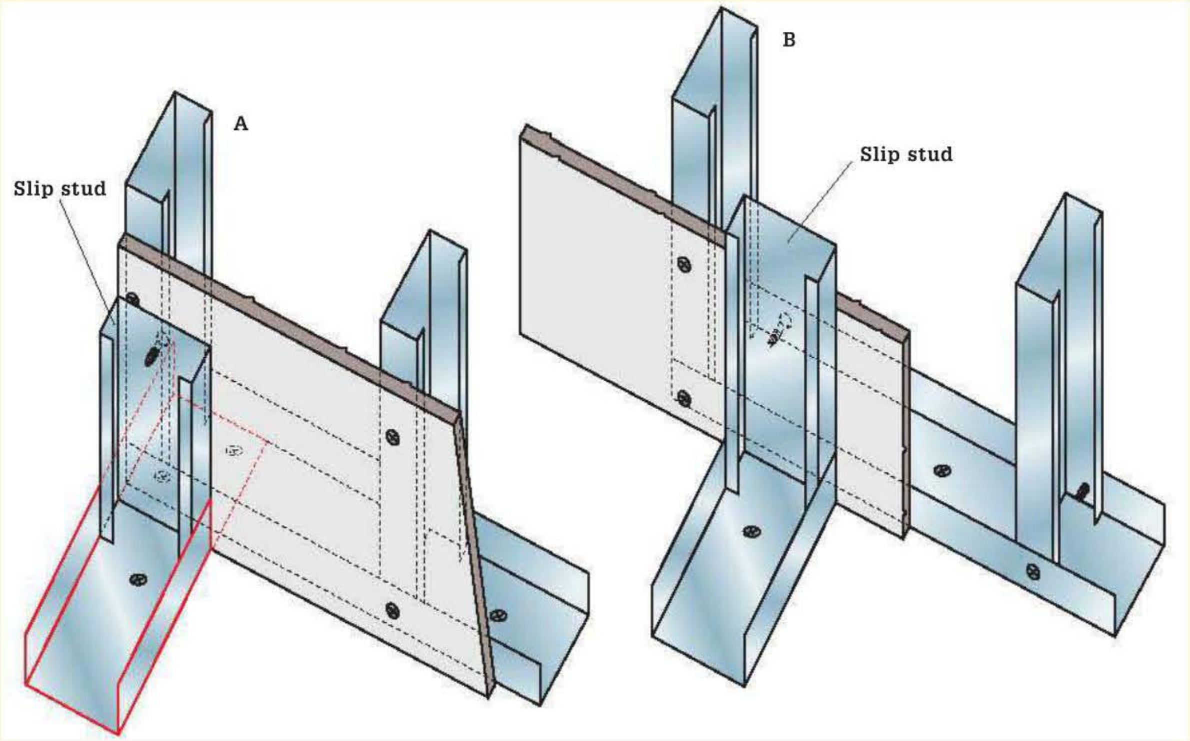

Build corners using a slip stud: A slip stud is not fastened until the adjacent drywall is in place. Form L-shaped corners (A) by overlapping the tracks. Cut off the flange on one side of one track, removing enough to allow room for the overlapping track and drywall. Form a T-shaped corner (B) by leaving a gap between the tracks for the drywall. Secure each slip stud by screwing through the stud into the tracks of the adjacent wall. Also screw through the back side of the drywall into the slip stud, if possible. Where there's no backing behind the slip stud, drive screws at a 45° angle through the back corners of the slip stud and into the drywall.

Installing Wallboard

Hanging wallboard is a project that can be completed quickly and easily with a little preplanning and a helping hand.

Planning the layout of panels will help you reduce waste and deal with problem areas. Where possible, install full panels perpendicular to the framing to add strength and rigidity to walls and ceilings. To save yourself time and trouble during the finishing process, avoid joints where two untapered panel ends are butted together. These are difficult to finish because there's no recess for the compound and tape. In small areas, install long sheets horizontally that run the full length of the walls. Or hang the panels vertically, which produces more seams that need taping but eliminates butted end joints. If butted joints are unavoidable, stagger the seams and locate them away from the center of the wall, or install back blocking to help mask unflattering effects.

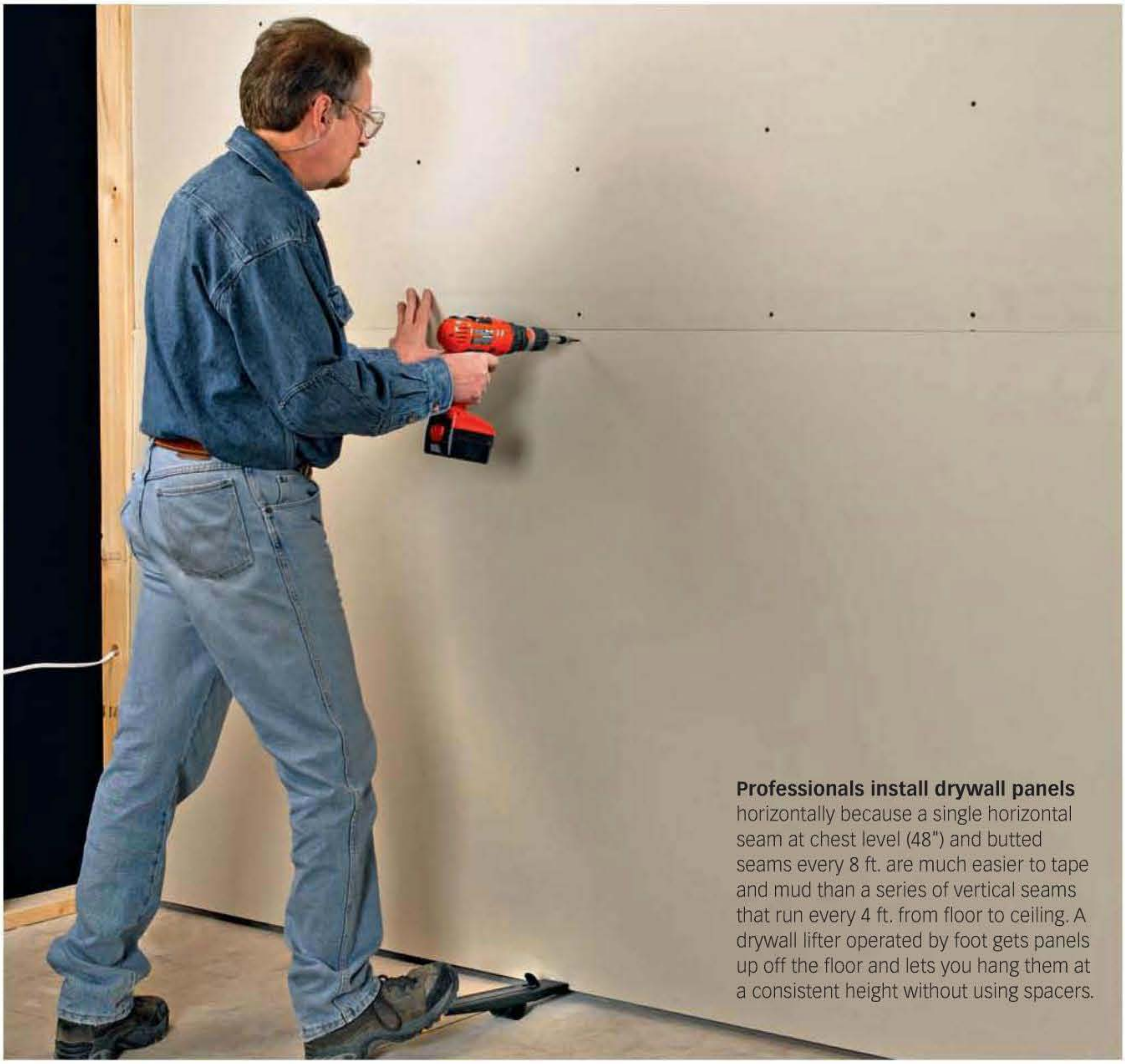

Professionals install drywall panels horizontally because a single horizontal seam at chest level (48") and butted seams every 8 ft. are much easier to tape and mud than a series of vertical seams that run every 4 ft. from floor to ceiling. A drywall lifter operated by foot gets panels up off the floor and lets you hang them at a consistent height without using spacers.

Tools & Prep Work

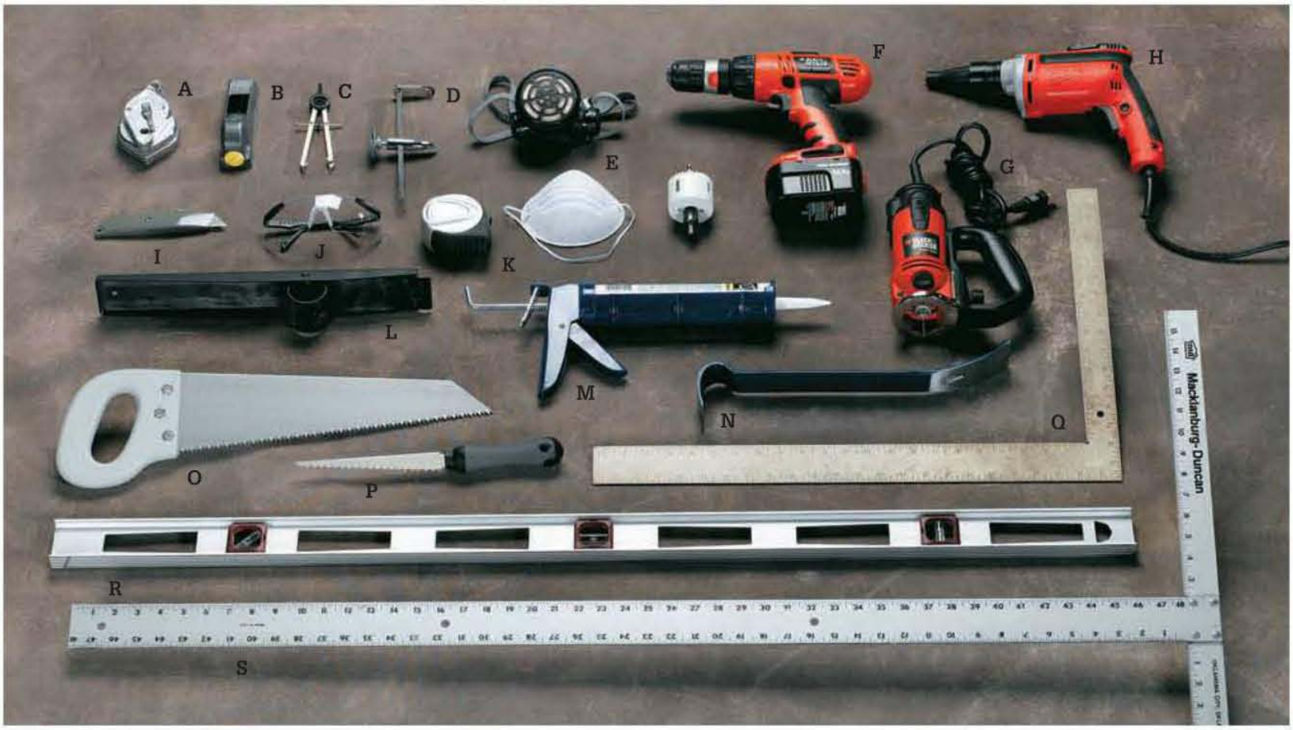

Tools for installing wallboard include: chalkline (A), surform plane (B), compass (C), wallboard compass (D), protective masks (E), drill with hole saw (F), trim router (G), drywall gun (H), utility knife (I), eye protection (J), tape measure (K), wallboard lifter (L), caulk gun (M), pry bar (N), wallboard saw (O), keyhole saw (P), framing square (Q), level (R), wallboard T-square (S).

Preparing for Wallboard Installation

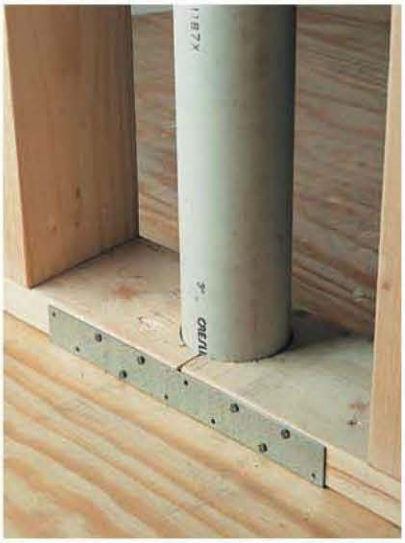

Use protector plates where wires or pipes pass through framing members less than 1¼" from the face. The plates prevent wallboard screws from puncturing wires or pipes.

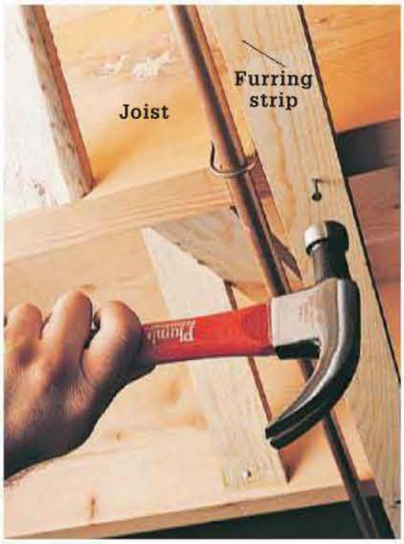

Nail furring strips to the framing to extend the wall surface beyond any obstructions such as water pipes or heating ducts.

Mark the locations of the studs on the floor with a carpenter's pencil or masking tape. After wallboard covers the studs, the marks indicate the stud locations.

How to Cut Wallboard

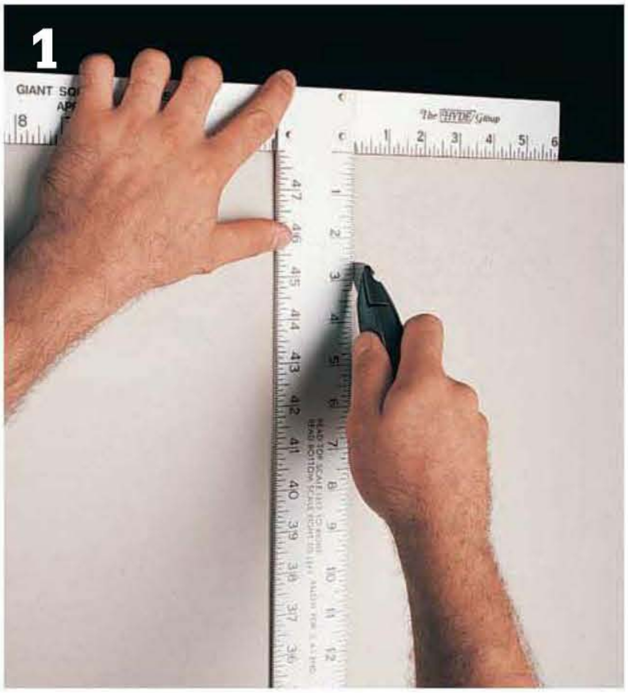

Position the wallboard T-square with the short arm flush against the edge. Use a utility knife to score the wallboard face paper along the arm of the square at the cutting point.

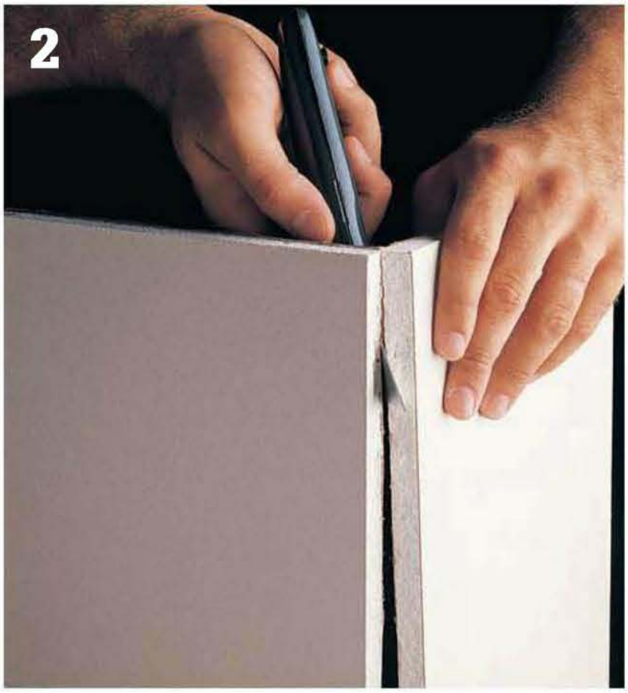

Bend the scored section with both hands to break the plaster core of the wallboard. Fold back the unwanted piece and cut through the back paper to separate the pieces.

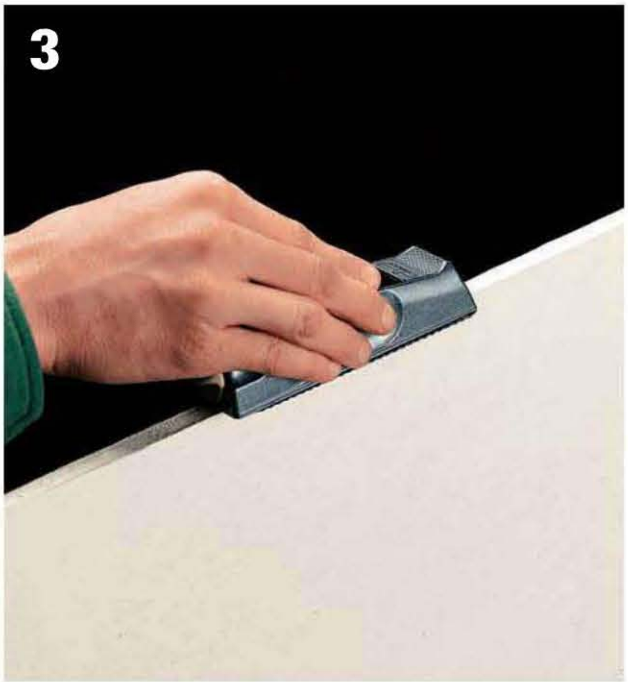

Smooth rough edges with a drywall rasp. One or two passes with the rasp should be sufficient. To help fit a piece into a tight space, bevel the edge slightly toward the back of the panel.

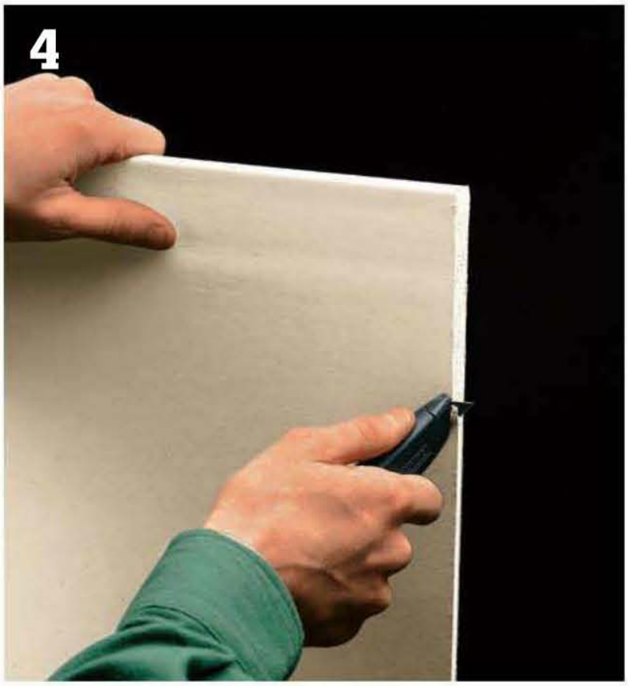

Where untapered panel ends will be butted together, bevel-cut the outside edges of each panel at 45°, removing about ⅛" of material. This helps prevent the paper from creating a ridge along the seam. Peel off any loose paper from the edge.

How to Install Wallboard on Flat Ceilings

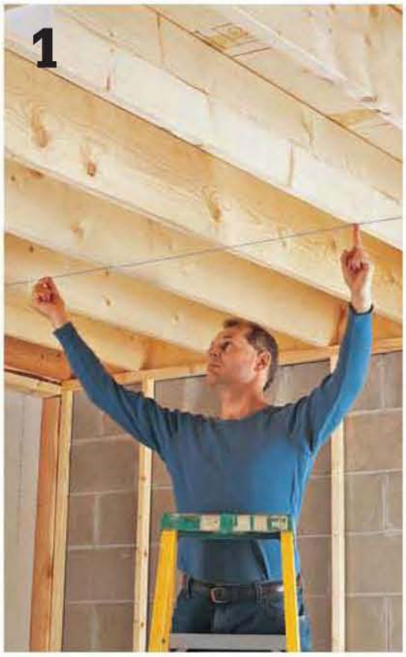

Snap a chalk line perpendicular to the joists, 48⅛" from the starting wall.

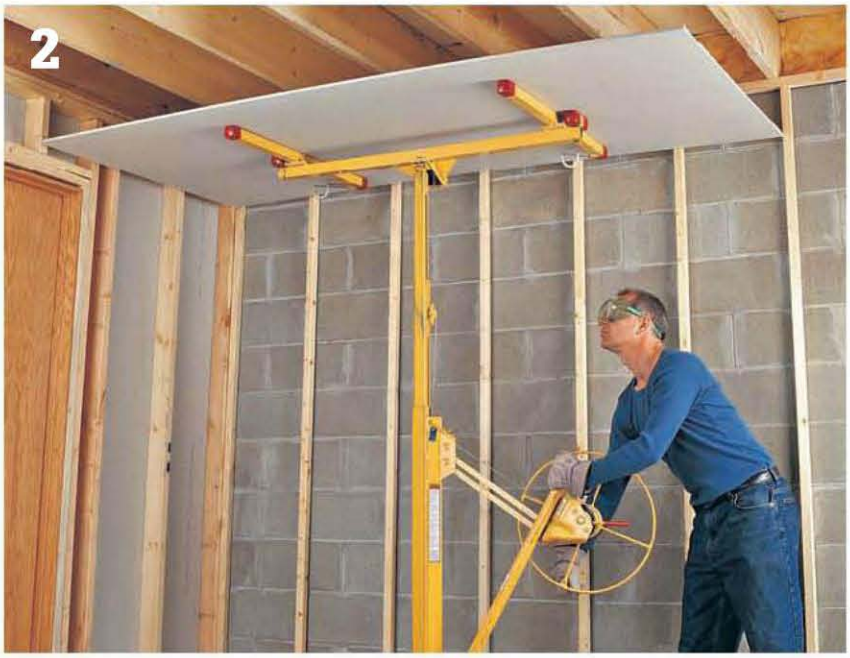

Measure to make sure the first panel will break on the center of a joist. If necessary, cut the panel on the end that abuts the side wall so the panel breaks on the next farthest joist. Load the panel onto a rented drywall lift, or use a helper, and lift the panel flat against the joists.

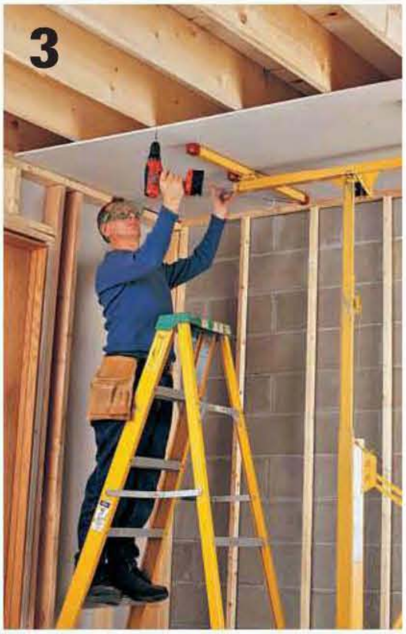

Position the panel with the leading edge on the chalk line and the end centered on a joist. Fasten the panel with appropriately sized screws driven every 12" along the joists.

After the first row of panels is installed, begin the next row with a half panel. This ensures that the butted end joints will be staggered between rows.

Tip ▸

Drywall stilts bring you within reach of ceilings, so you can fasten and finish the drywall without a ladder. Stilts are commonly available at rental centers and are surprisingly easy to use.

How to Install Wallboard on Walls

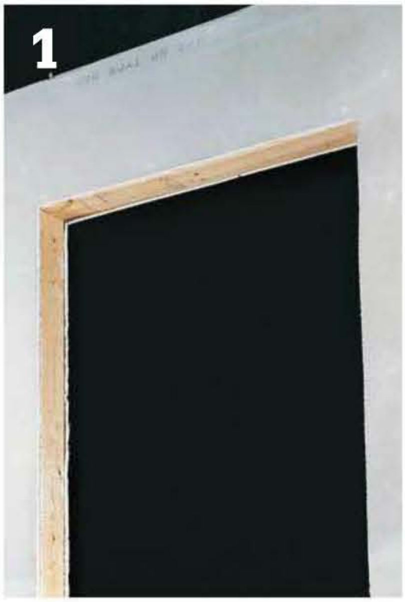

Plan wallboard placement so joints do not fall at the corners of doors or windows. Wallboard joints at corners often crack and cause bulges that interfere with miter joints in window or door trim.

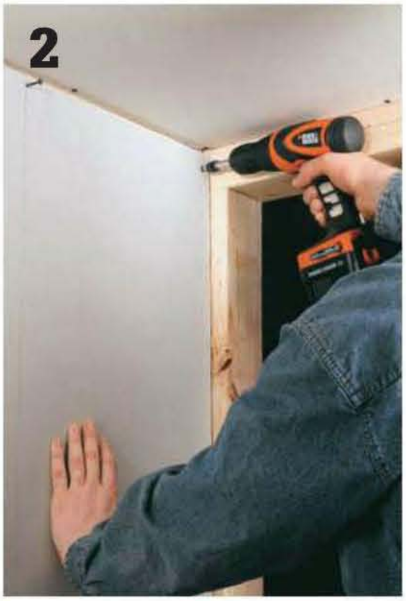

With a helper or a wallboard lift, hoist the first panel tight against the ceiling, making sure the side edge is centered on a stud. Push the panel flat against the framing and drive the starter screws to secure the panel. Make any cutouts, then fasten the field of the panel with drywall screws driven every 12".

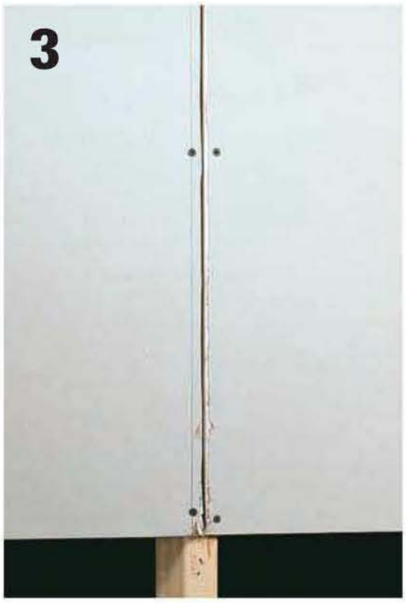

Measure, cut, and install the remaining panels along the upper wall. Bevel panel ends slightly, leaving a ⅛" gap between them at the joint. Butt joints can also be installed using back blocking to create a recess.

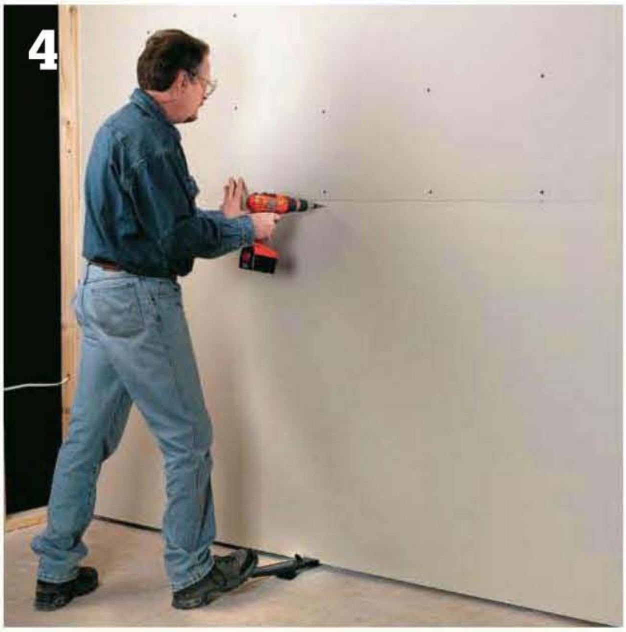

Measure, cut, and install the bottom row, butting the panels tight to the upper row and leaving a ½" gap at the floor. Secure to the framing along the top edge using the starter screws, then make all cutouts before fastening the rest of the panel.

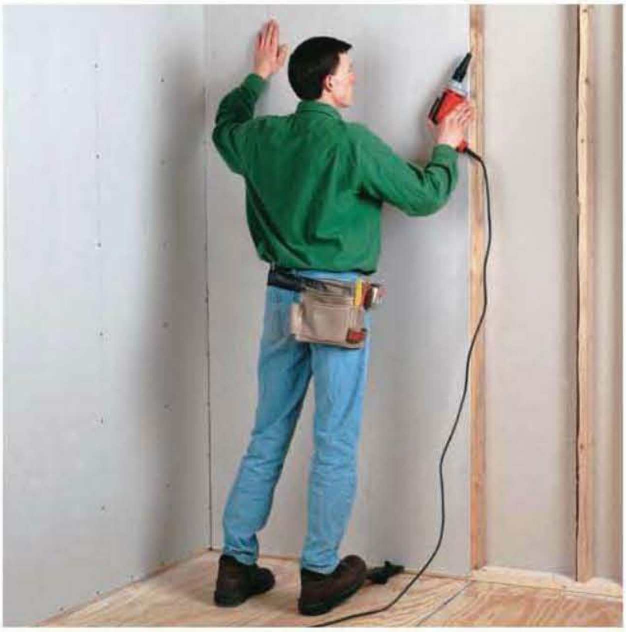

Variation: When installing wallboard vertically, cut each panel so it's ½" shorter than the ceiling height to allow for expansion. (The gap will be covered by base molding.) Avoid placing tapered edges at outside corners; they are difficult to finish.

How to Install Wallboard at Inside Corners

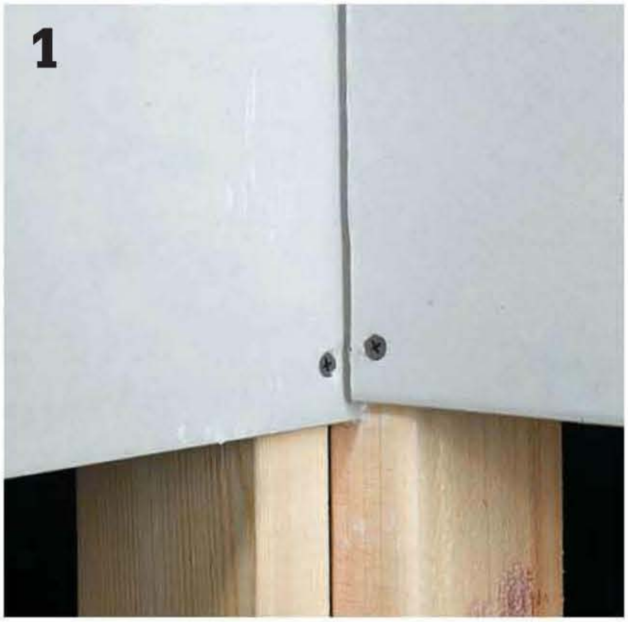

Standard 90° inside corners are installed with the first panel butted against the framing and the adjacent panel butted against the first. The screw spacing remains the same as on a flat wall (see page 48).

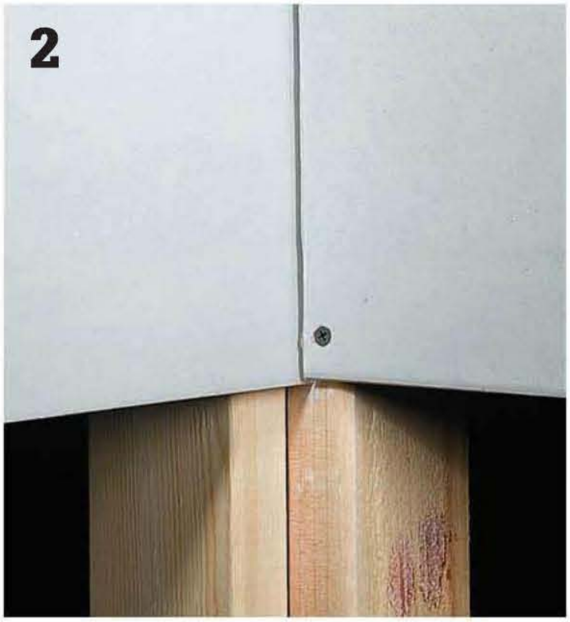

Use a "floating corner" to reduce the chances of popped fasteners and cracks. Install the first panel, fastening only to within one stud bay of the corner. Push the leading edge of the adjacent panel against the first to support the unfastened edge. Fasten the second panel normally, including the corner.

How to Install Wallboard at Outside Corners

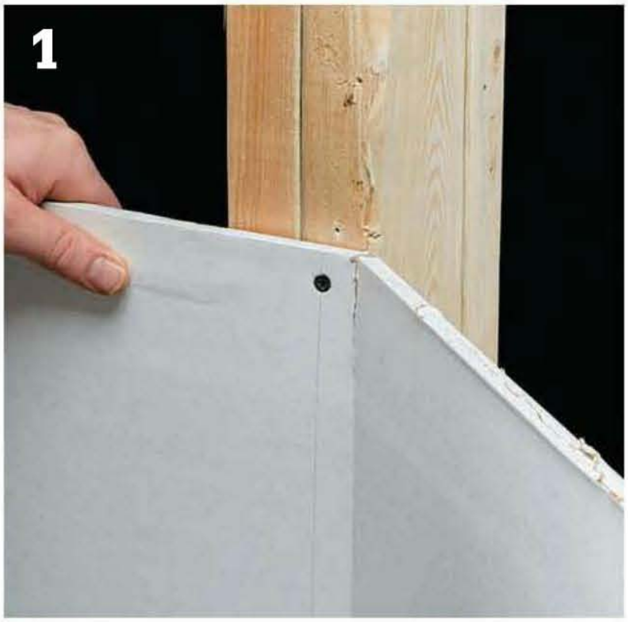

Cut the first panel so it is longer than the opening and extends past the corner framing. Fasten the panel in place, then score the backside and snap cut to remove the waste piece.

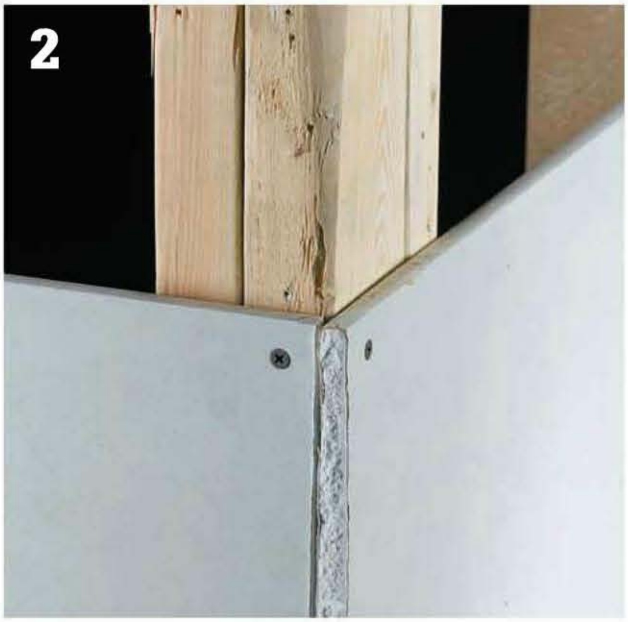

Install the second panel on the adjacent wall so it overlaps the end of the first panel slightly. Score and snap-cut so end is flush with first panel wall surface.

How to Tape Wallboard Joints

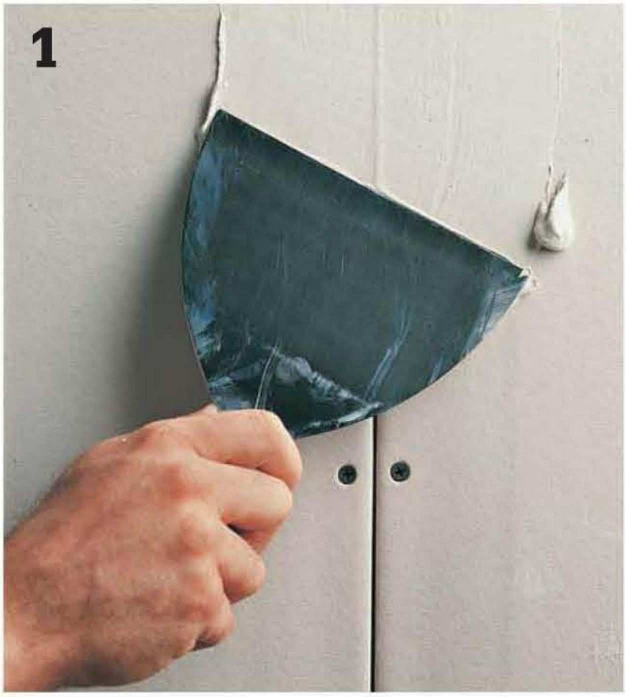
1

Apply a thin layer of wallboard compound over the joint with a 4" or 6" wallboard knife. To load the knife, dip it into a pan filled with wallboard compound.

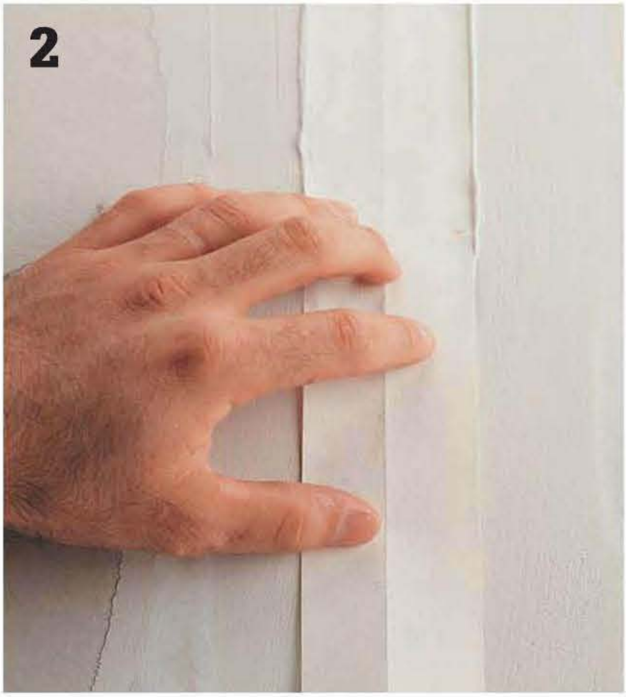
2

Press the wallboard tape into the compound immediately, centering the tape on the joint. Wipe away the excess compound and smooth the joint with a 6" knife. Let dry overnight.

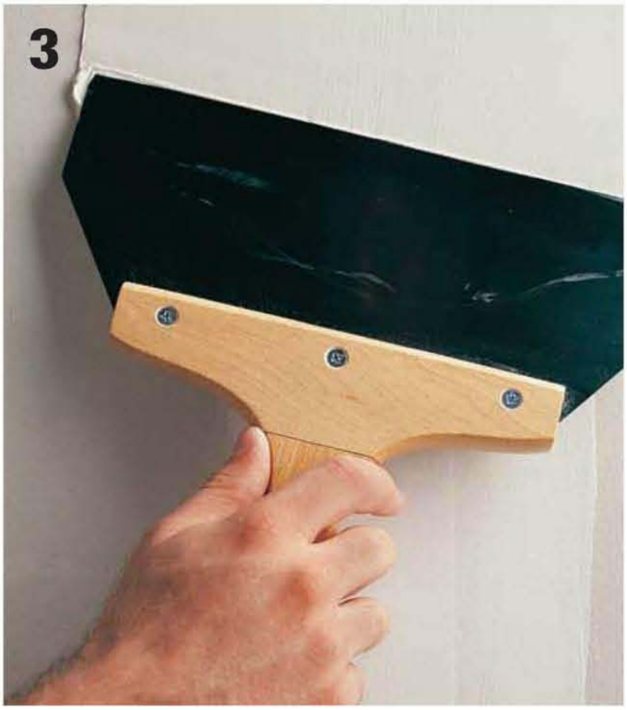
3

Apply a thin finish coat of compound with a 10" wallboard knife. Allow the second coat to dry and shrink overnight. Apply the last coat and let it harden slightly before wet-sanding.

Tip ▸

TRIKO PRODUCTS
TAPING • TOPPING • TEXTURE
REDI-MIX - ALL PURPOSE
JOINT COMPOUND

Use premixed wallboard compound for most taping and finishing jobs to eliminate mixing. Use paper wallboard tape when using premixed wallboard compound.

How to Finish Inside Corners

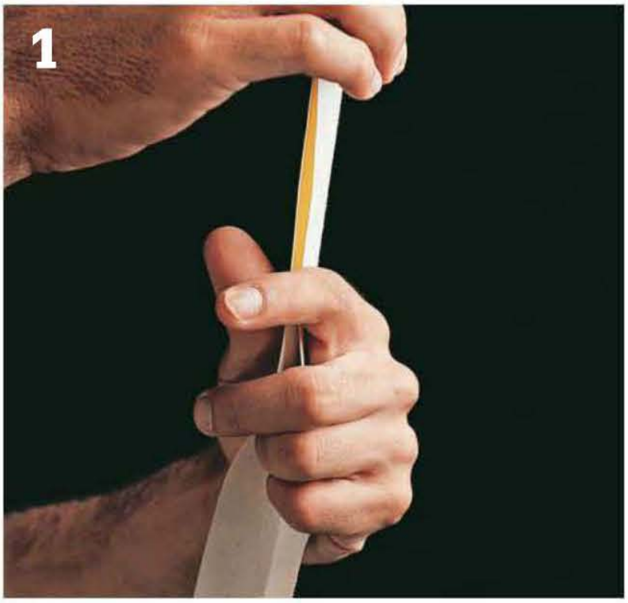

Fold a strip of paper wallboard tape in half by pinching the strip and pulling it between thumb and forefinger. Apply a thin layer of wallboard compound to both sides of the inside corner, using a 4" wallboard knife.

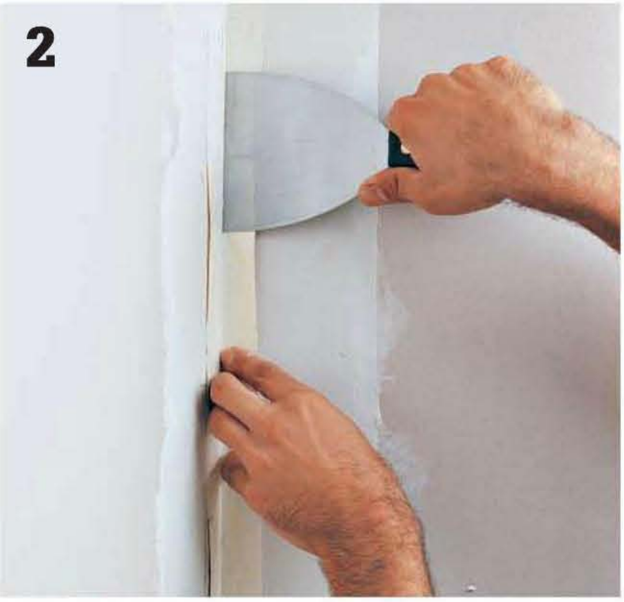

Position the end of the folded tape strip at the top of the joint and press the tape into the wet compound with the knife. Smooth both sides of the corner. Finish as described in step 3, preceding page.

How to Finish Outside Corners

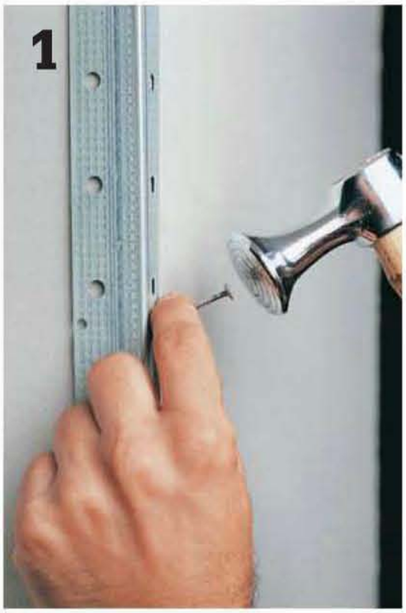

Position corner bead on outside corners. Using a level, adjust the bead so corner is plumb. Attach with 1¼" wallboard nails or screws spaced at 8" intervals. (Some corner beads are fastened with wallboard compound.)

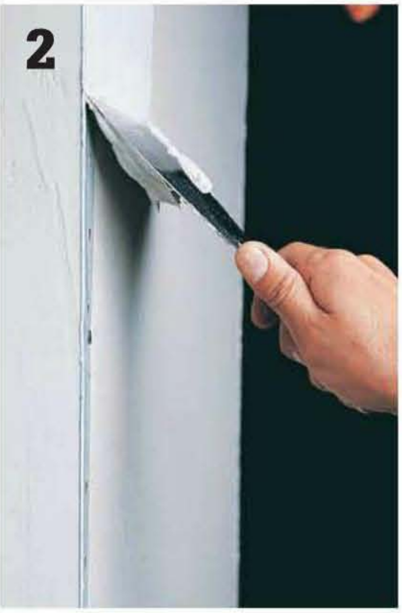

Cover the corner bead with three coats of wallboard compound, using a 6" or 10" wallboard knife. Let each coat dry and shrink overnight before applying the next coat. Smooth the final coat with a wet sander.

Tip ▸

Sand joints lightly after wallboard compound dries. Use a pole sander to reach high areas without a ladder. Wear a dust mask when dry-sanding.

Understanding Plumbing

Because most of a plumbing system is hidden inside walls and floors, it may seem to be a complex maze of pipes and fittings. In fact, home plumbing is simple and straightforward. Understanding how home plumbing works is an important first step toward doing routine maintenance and money-saving repairs.

A typical home plumbing system includes three basic parts: a water supply system, a fixture and appliance set, and a drain system. These three parts can be seen clearly in the photograph of the cut-away house on the opposite page.

Fresh water enters a home through a main supply line (1). This fresh water source is provided by either a municipal water company or a private underground well. If the source is a municipal supplier, the water passes through a meter (2) that registers the amount of water used. A family of four uses about 400 gallons of water each day.

Immediately after the main supply enters the house, a branch line splits off (3) and is joined to a water heater (4). From the water heater, a hot water line runs parallel to the cold water line to bring the water supply to fixtures and appliances throughout the house. Fixtures include sinks, bathtubs, showers, and laundry tubs. Appliances include water heaters, dishwashers, clothes washers, and water softeners. Toilets and exterior sillcocks are examples of fixtures that require only a cold water line.

The water supply to fixtures and appliances is controlled with faucets and valves. Faucets and valves have moving parts and seals that eventually may wear out or break, but they are easily repaired or replaced.

Waste water then enters the drain system. It first must flow past a drain trap (5), a U-shaped piece of pipe that holds standing water and prevents sewer gases from entering the home. Every fixture must have a drain trap.

The drain system works entirely by gravity, allowing waste water to flow downhill through a series of large-diameter pipes. These drain pipes are attached to a system of vent pipes. Vent pipes (6) bring fresh air to the drain system, preventing suction that would slow or stop drain water from flowing freely. Vent pipes usually exit the house at a roof vent (7).

All waste water eventually reaches a main waste and vent stack (8). The main stack curves to become a sewer line (9) that exits the house near the foundation. In a municipal system, this sewer line joins a main sewer line located near the street. Where sewer service is not available, waste water empties into a septic system.

TRIDENT
METER
GALLONS
10 GALLONS

Water meters and main shutoff valves are located where the main water supply pipe enters the house. The water meter is the property of your local municipal water company. If the water meter leaks, or if you suspect it is not functioning properly, call your water company for repairs.

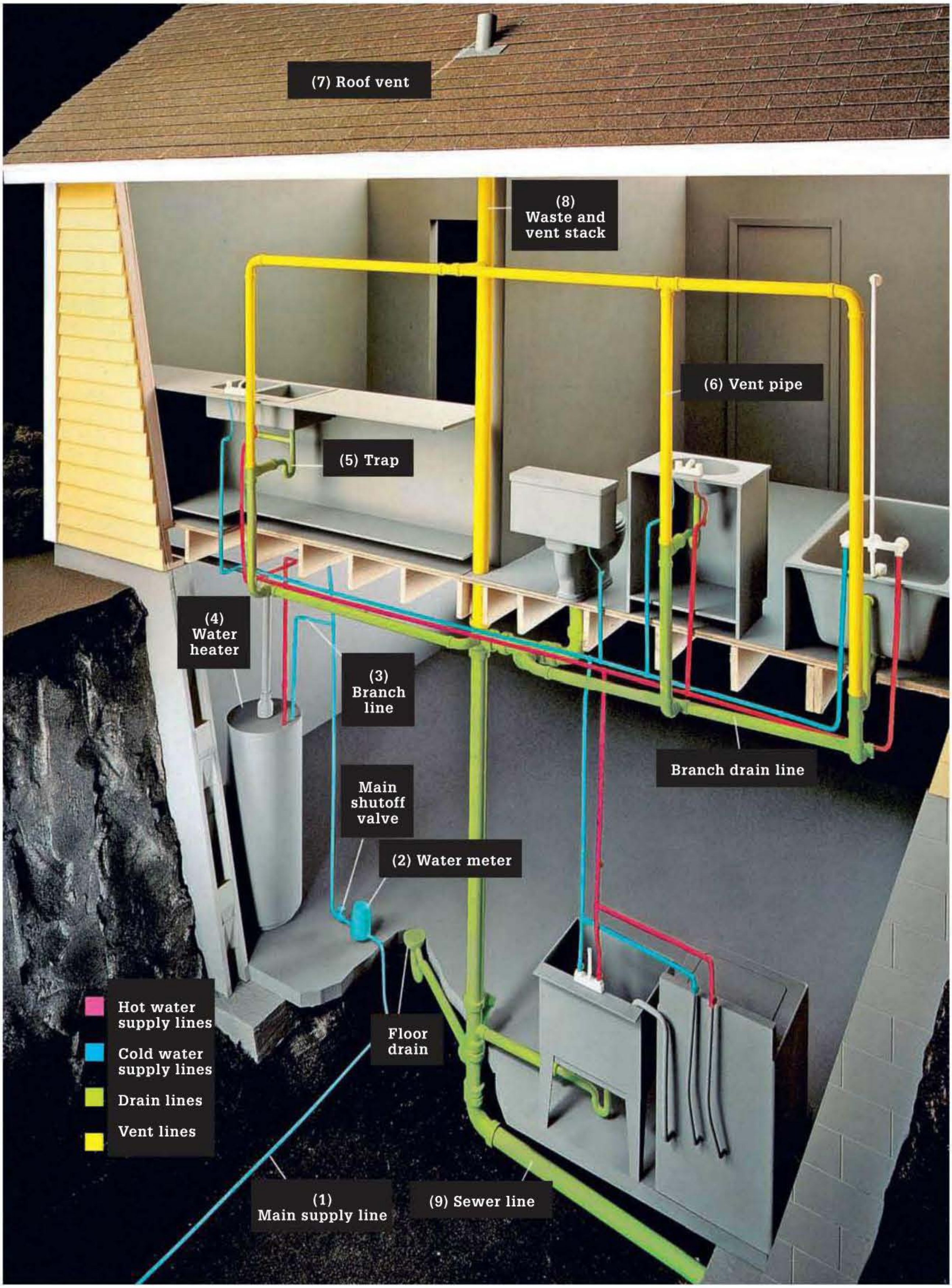
(7) Roof vent
(8) Waste and vent stack
(6) Vent pipe
(5) Trap
(4) Water heater
(3) Branch line
Branch drain line
Main shutoff valve
(2) Water meter
Floor drain
(1) Main supply line
(9) Sewer line
Hot water supply lines
Cold water supply lines
Drain lines
Vent lines

Water Supply System

Water supply pipes carry hot and cold water throughout a house. In homes built before 1960, the original supply pipes were usually made of galvanized iron. Newer homes have supply pipes made of copper. In most areas of the country, supply pipes made of rigid plastic or PEX are accepted by local plumbing codes.

Water supply pipes are made to withstand the high pressures of the water supply system. They have small diameters, usually ½" to 1", and are joined with strong, watertight fittings. The hot and cold lines run in tandem to all parts of the house. Usually, the supply pipes run inside wall cavities or are strapped to the undersides of floor joists.

Hot and cold water supply pipes are connected to fixtures or appliances. Fixtures include sinks, tubs, and showers. Some fixtures, such as toilets or hose bibs, are supplied only by cold water. Appliances include dishwashers and clothes washers. A refrigerator icemaker uses only cold water. Tradition says that hot water supply pipes and faucet handles are found on the left-hand side of a fixture, with cold water on the right.

Because it is pressurized, the water supply system is prone to leaks. This is especially true of galvanized iron pipe, which has limited resistance to corrosion.

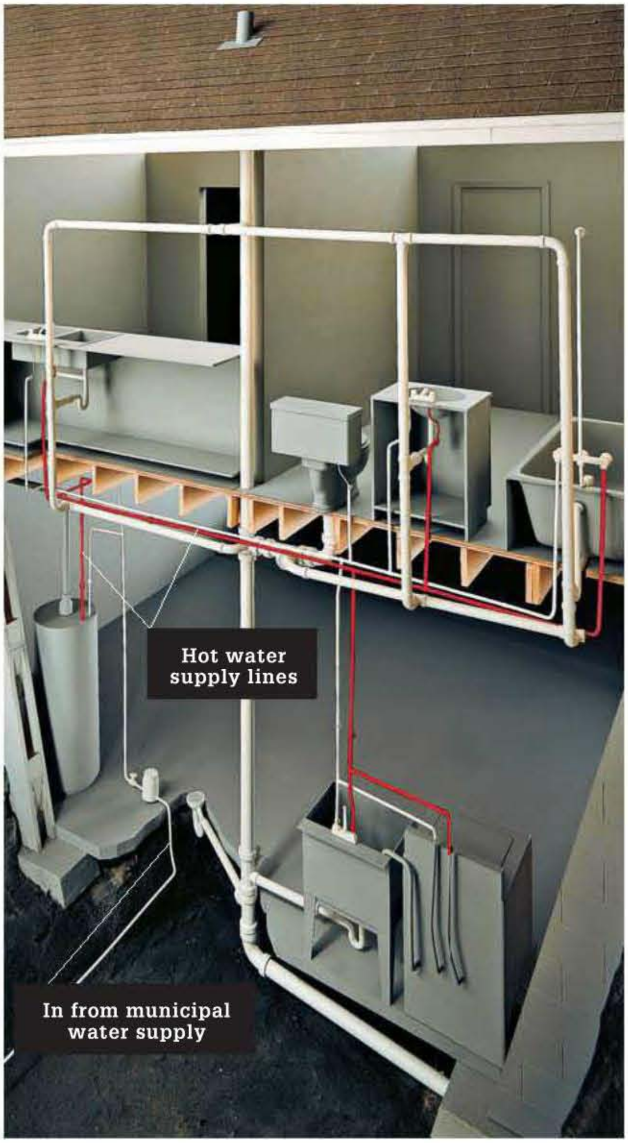
Hot water supply lines
In from municipal water supply

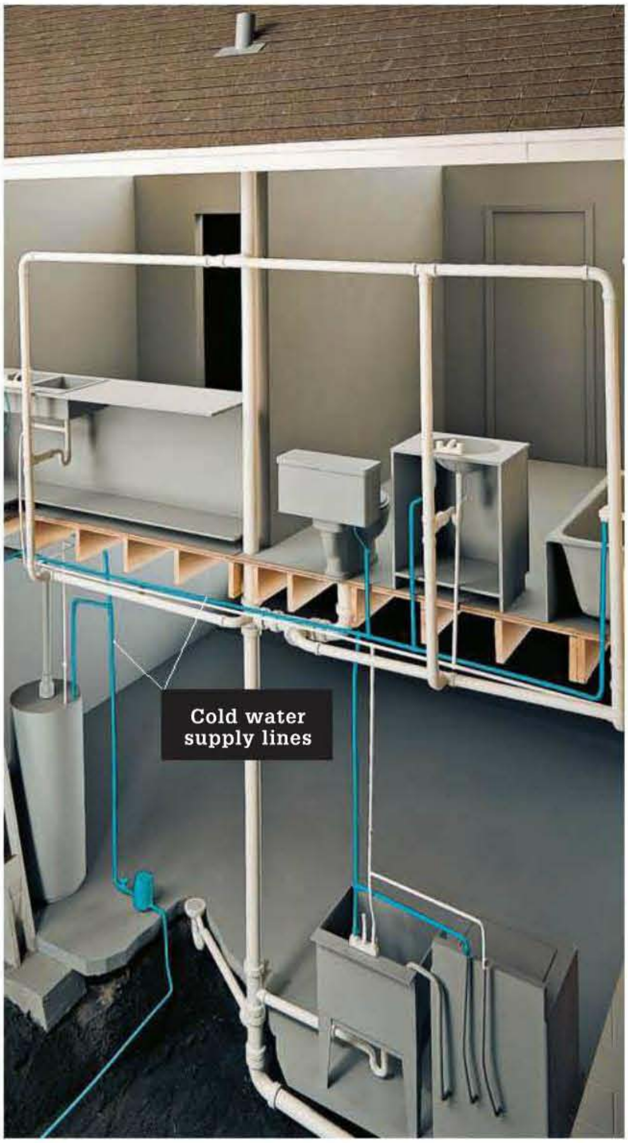

Drain-Waste-Vent System

Drain pipes use gravity to carry waste water away from fixtures, appliances, and other drains. This waste water is carried out of the house to a municipal sewer system or septic tank.

Drain pipes are usually plastic or cast iron. In some older homes, drain pipes may be made of copper or lead. Because they are not part of the supply system, lead drain pipes pose no health hazard. However, lead pipes are no longer manufactured for home plumbing systems.

Drain pipes have diameters ranging from 1¼" to 4". These large diameters allow waste to pass through easily.

Traps are an important part of the drain system. These curved sections of drain pipe hold standing water, and they are usually found near any drain opening. The standing water of a trap prevents sewer gases from backing up into the home. Each time a drain is used, the standing trap water is flushed away and is replaced by new water.

In order to work properly, the drain system requires air. Air allows waste water to flow freely down drain pipes.

To allow air into the drain system, drain pipes are connected to vent pipes. All drain systems must include vents, and the entire system is called the drain-waste-vent (DWV) system. One or more vent stacks, located on the roof, provide the air needed for the DWV system to work.

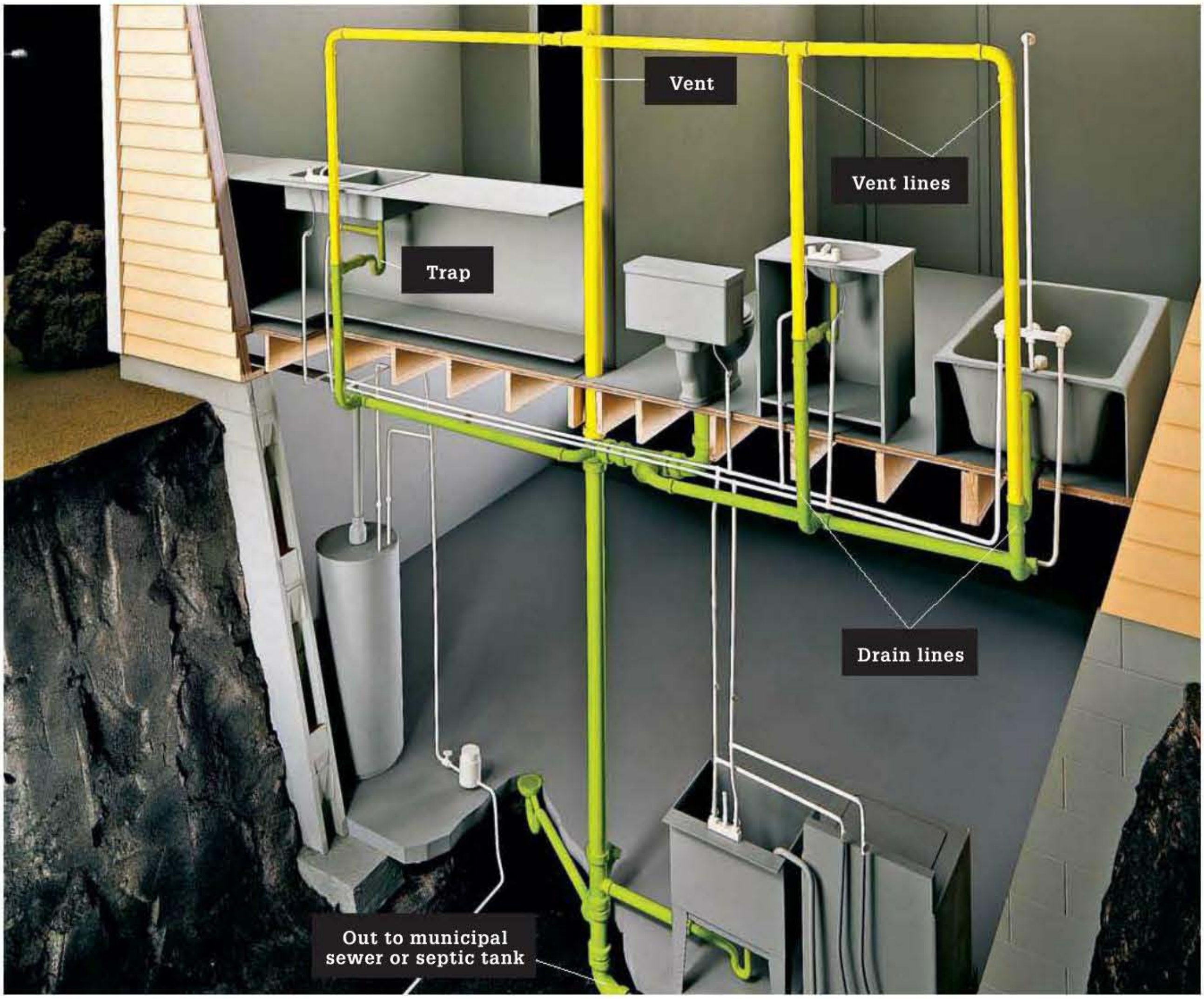

Working with Copper Pipe

Copper is the ideal material for water supply pipes. It resists corrosion and has smooth surfaces that provide good water flow. Copper pipes are available in several diameters (page 314), but most home water supply systems use ½" or ¾" pipe. Copper pipe is manufactured in rigid and flexible forms.

Rigid copper, sometimes called hard copper, is approved for home water supply systems by all local codes. It comes in three wall-thickness grades: Types M, L, and K. Type M is the thinnest, the least expensive, and a good choice for do-it-yourself home plumbing.

Rigid Type L usually is required by code for commercial plumbing systems. Because it is strong and solders easily, Type L may be preferred by some professional plumbers and do-it-yourselfers for home use. Type K has the heaviest wall thickness and is used most often for underground water service lines.

Flexible copper, also called soft copper, comes in two wall-thickness grades: Types L and K. Both are approved for most home water supply systems, although flexible Type L copper is used primarily for gas service lines. Because it is bendable and will resist a mild frost, Type L may be installed as part of a water supply system in unheated indoor areas, like crawl spaces. Type K is used for underground water service lines.

A third form of copper, called DWV, is used for drain systems. Because most codes now allow low-cost plastic pipes for drain systems, DWV copper is seldom used.

Copper pipes are connected with soldered, compression, or flare fittings (see chart below). Always follow your local code for the correct types of pipes and fittings allowed in your area.

Soldered fittings, also called sweat fittings, often are used to join copper pipes. Correctly soldered fittings (pages 278 to 280) are strong and trouble-free. Copper pipe can also be joined with compression fittings or flare fittings. See chart below.

Copper Pipe & Fitting Chart ▸

	Rigid Copper			Flexible Copper		
Fitting Method	**Type M**	**Type L**	**Type K**	**Type L**	**Type K**	**General Comments**
Soldered	yes	yes	yes	yes	yes	Inexpensive, strong, and trouble-free fitting method. Requires some skill.
Compression	yes	no	no	no	no	Makes repairs and replacement easy. More expensive than solder. Best used on flexible copper.
Flare	no	no	yes	yes	yes	Use only with flexible copper pipes. Usually used as a gas-line fitting. Requires some skill.

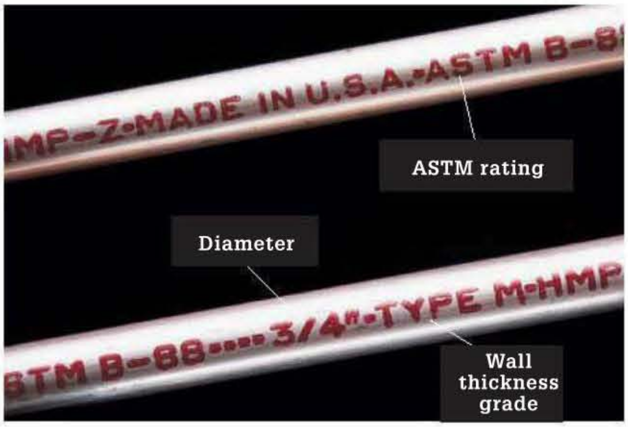
MP-Z-MADE IN U.S.A.-ASTM B-8
ASTM rating
Diameter
STM B-88----3/4"-TYPE M-HMP
Wall thickness grade

Grade stamp information includes the pipe diameter, the wall-thickness grade, and a stamp of approval from the ASTM (American Society for Testing and Materials). Type M pipe (shown) is identified by red lettering, Type L by blue lettering.

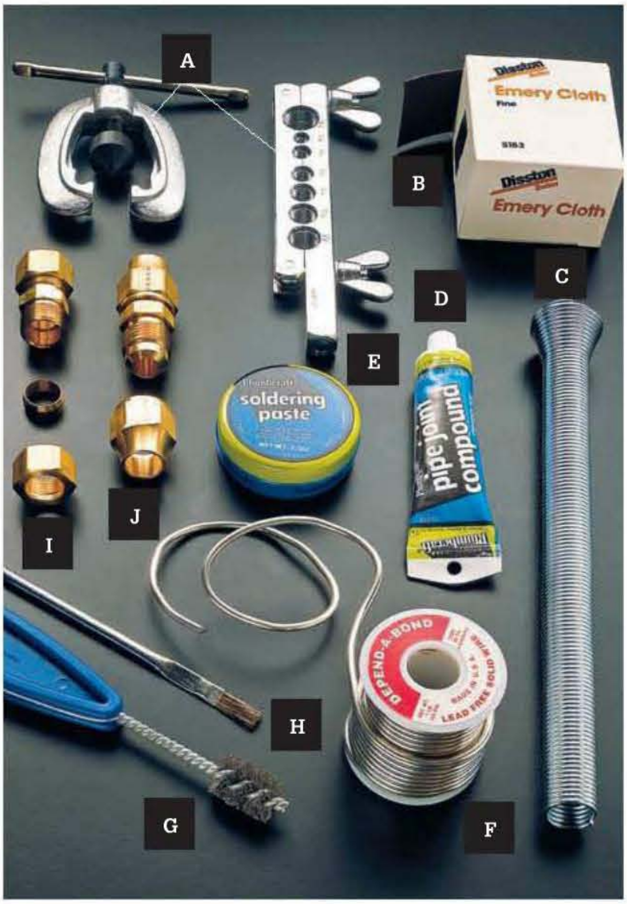
A
B
C
D
E
F
G
H
I
J
Emery Cloth
soldering paste
pipejoint compound

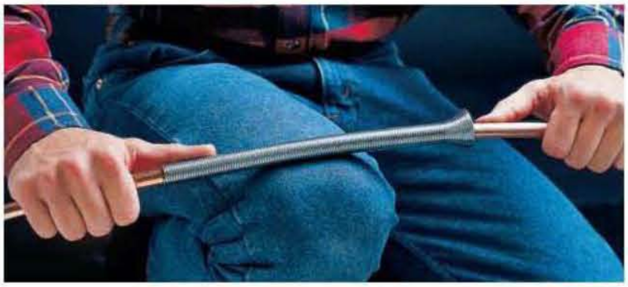

Bend flexible copper pipe with a coil-spring tubing bender to avoid kinks. Select a bender that matches the outside diameter of the pipe. Slip bender over pipe using a twisting motion. Bend pipe slowly until it reaches the correct angle, but not more than 90º.

Specialty tools and materials for working with copper include: flaring tools (A), emery cloth (B), coil-spring tubing bender (C), pipe joint compound (D), soldering paste (flux) (E), lead-free solder (F), wire brush (G), flux brush (H), compression fitting (I), flare fitting (J).

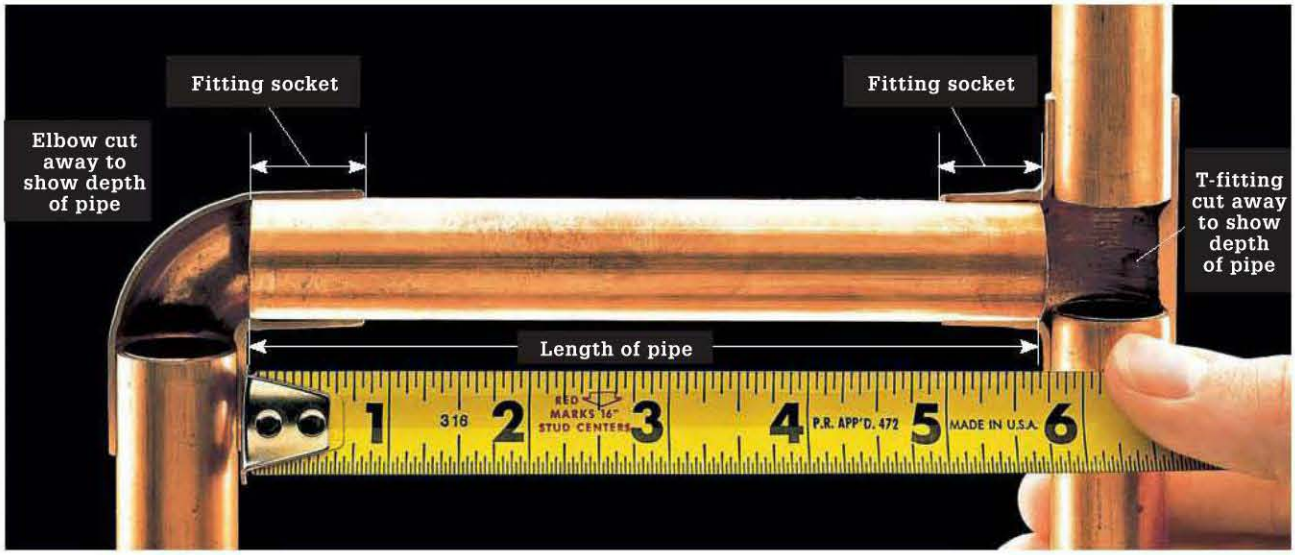
Fitting socket
Fitting socket
Elbow cut away to show depth of pipe
T-fitting cut away to show depth of pipe
Length of pipe

Find the length of copper pipe needed by measuring between the bottom of the copper fitting sockets (fittings shown in cutaway). Mark the length on the pipe with a felt-tipped pen.

Cutting & Soldering Copper

The best way to cut rigid and flexible copper pipe is with a tubing cutter. A tubing cutter makes a smooth, straight cut, an important first step toward making a watertight joint. Remove any metal burrs on the cut edges with a reaming tool or round file.

Copper can be cut with a hacksaw. A hacksaw is useful in tight areas where a tubing cutter will not fit. Take care to make a smooth, straight cut when cutting with a hacksaw.

A soldered pipe joint, also called a sweated joint, is made by heating a copper or brass fitting with a propane torch until the fitting is just hot enough to melt metal solder. The heat draws the solder into the gap between the fitting and pipe to form a watertight seal. A fitting that is overheated or unevenly heated will not draw in solder. Copper pipes and fittings must be clean and dry to form a watertight seal.

Tools & Materials ▸

- Tubing cutter with reaming tip (or hacksaw and round file)
- Wire brush
- Flux brush
- Propane torch
- Spark lighter (or matches)
- Round file
- Cloth
- Adjustable wrench
- Channel-type pliers
- Copper pipe
- Copper fittings
- Emery cloth
- Soldering paste (flux)
- Sheet metal
- Lead-free solder
- Rag

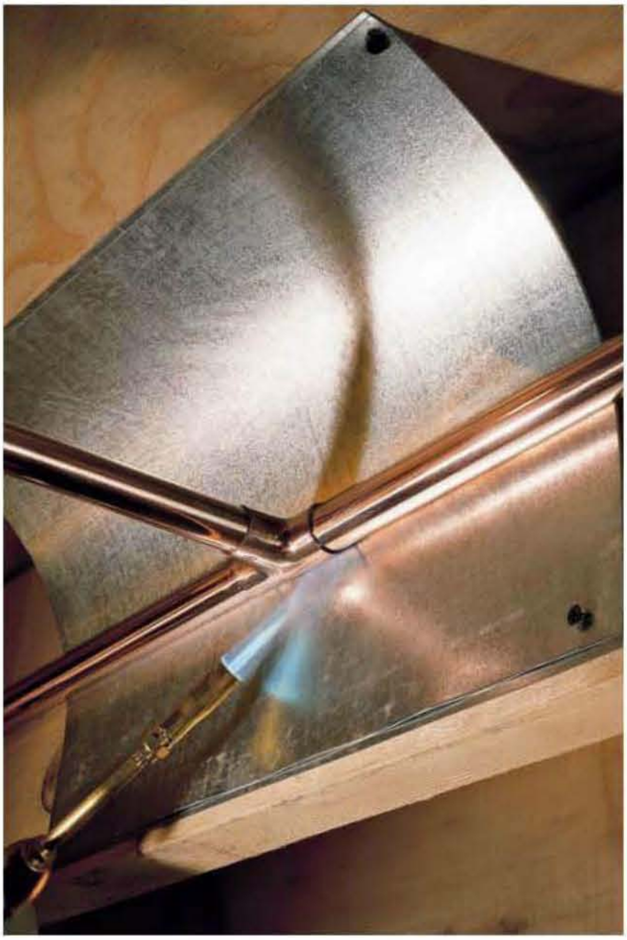

Protect wood from the heat of the torch flame while soldering, using a double layer (two 18" × 18" pieces) of 26-gauge sheet metal. Buy sheet metal at hardware stores or building supply centers and keep it to use with all soldering projects.

Soldering Tips ▸

Use caution when soldering copper. Pipes and fittings become very hot and must be allowed to cool before handling.

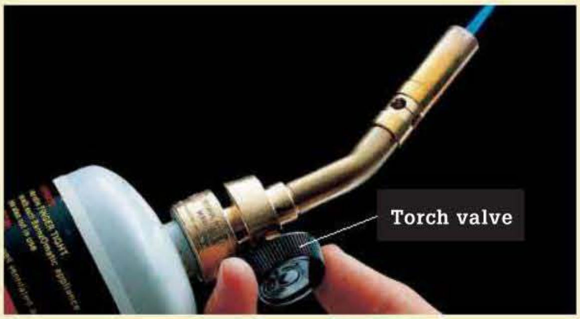

Prevent accidents by shutting off propane torch immediately after use. Make sure valve is closed completely.

How to Cut Rigid & Flexible Copper Pipe

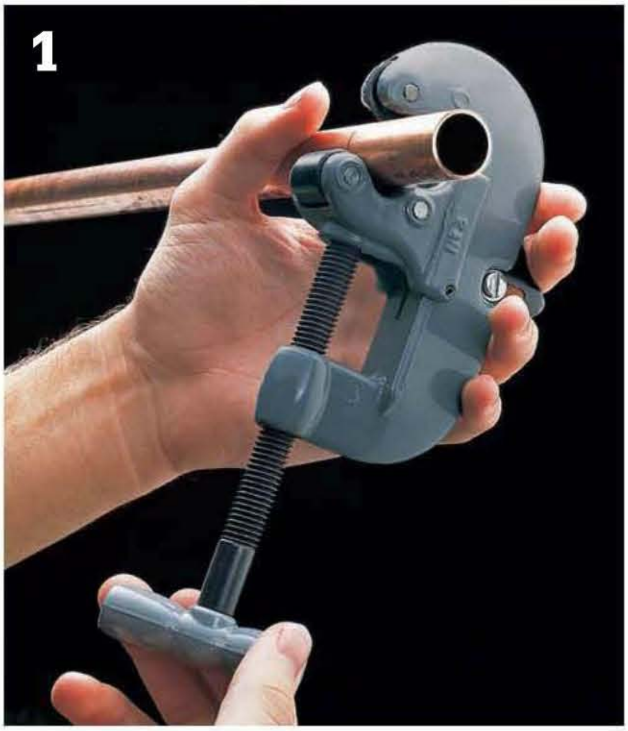

Place the tubing cutter over the pipe and tighten the handle so that the pipe rests on both rollers, and the cutting wheel is on the marked line.

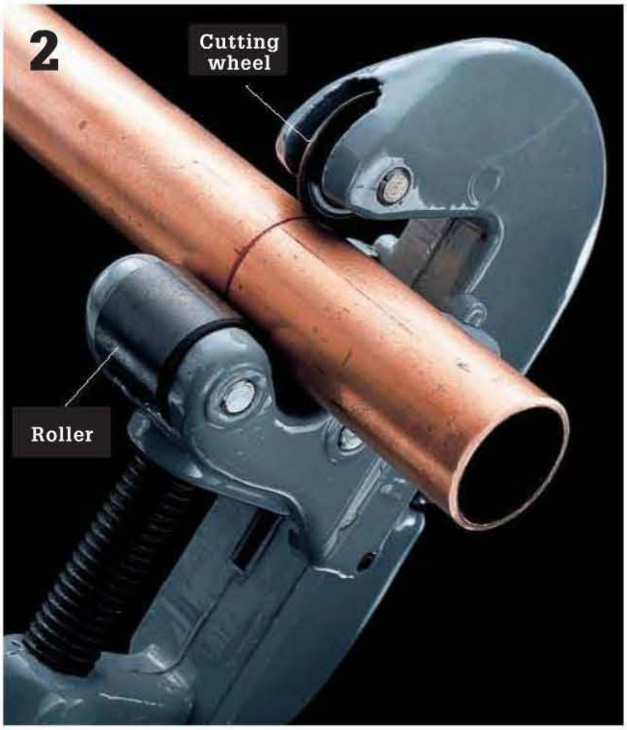

Turn the tubing cutter one rotation so that the cutting wheel scores a continuous straight line around the pipe.

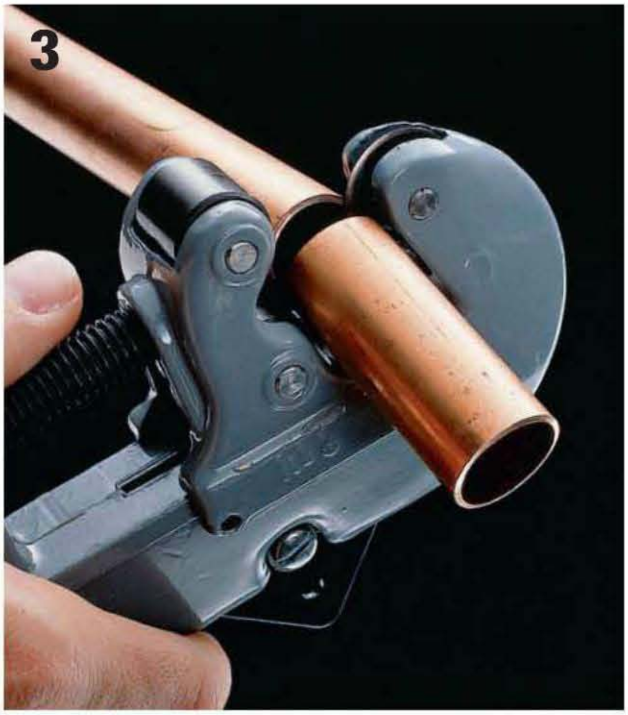

Rotate the cutter in the opposite direction, tightening the handle slightly after every two rotations, until the cut is complete.

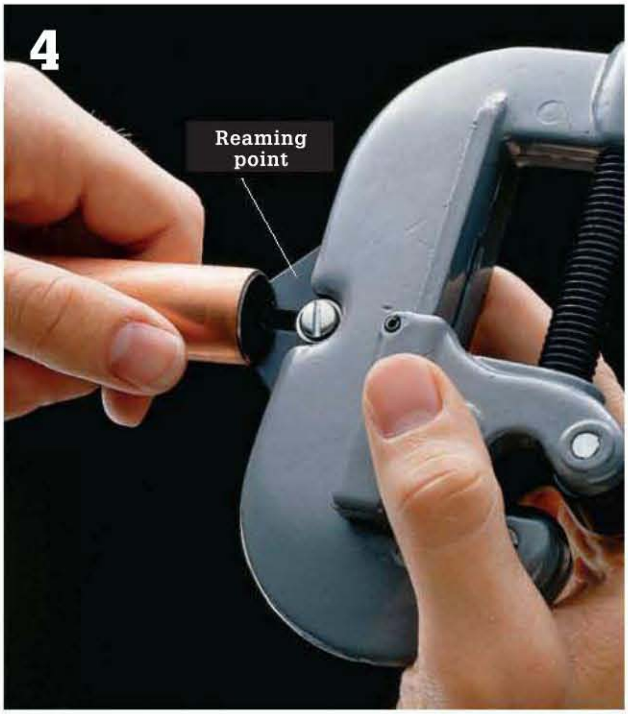

Remove sharp metal burrs from the inside edge of the cut pipe, using the reaming point on the tubing cutter, or a round file.

How to Solder Copper Pipes & Fittings

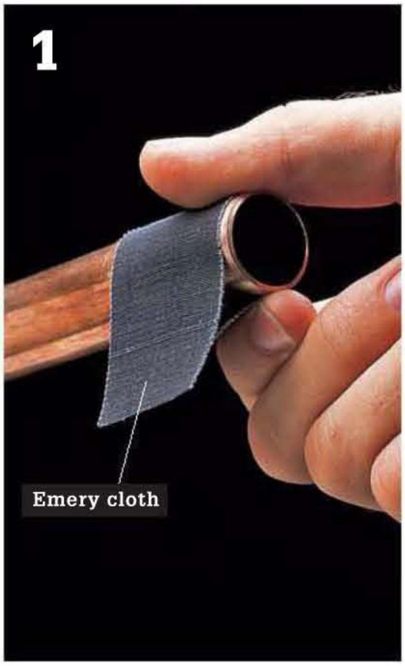
1
Emery cloth

Clean the end of each pipe by sanding with emery cloth. Ends must be free of dirt and grease to ensure that the solder forms a good seal.

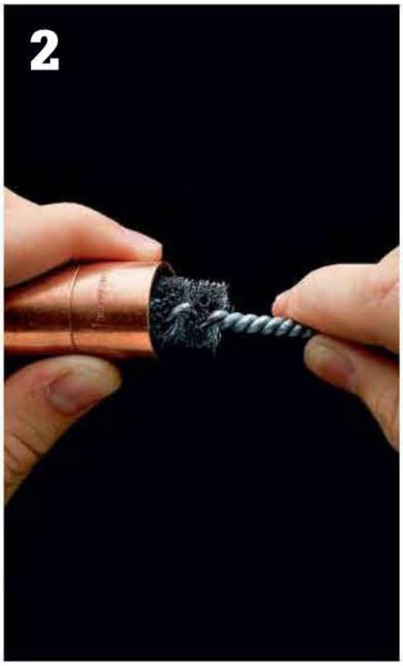

Clean the inside of each fitting by scouring with a wire brush or emery cloth.

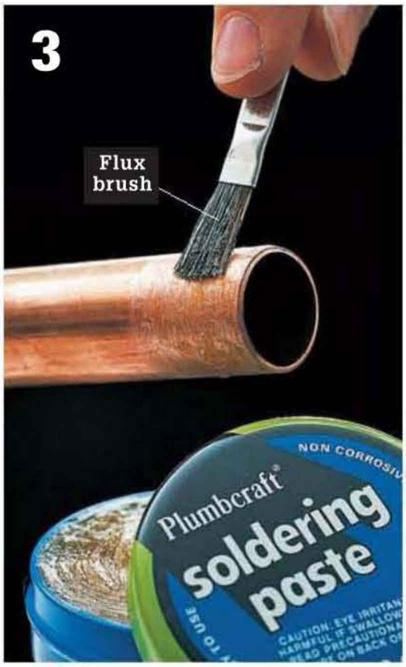
3
Flux brush
Plumbcraft
soldering paste
NON CORROSIVE

Apply a thin layer of soldering paste (flux) to end of each pipe, using a flux brush. Soldering paste should cover about 1" of pipe end.

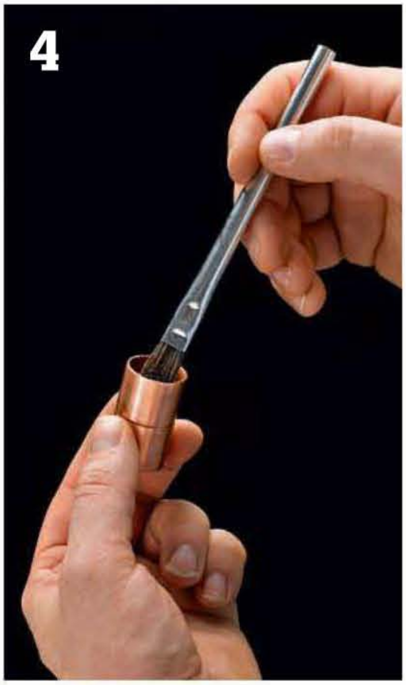

Apply a thin layer of flux to the inside of the fitting.

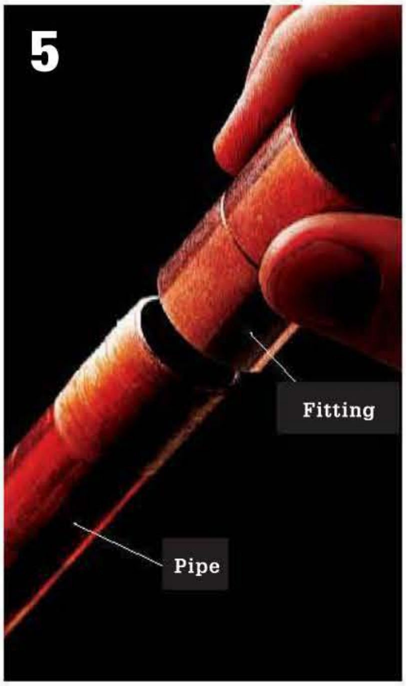
5
Fitting
Pipe

Assemble each joint by inserting the pipe into the fitting so it is tight against the bottom of the fitting sockets. Twist each fitting slightly to spread soldering paste.

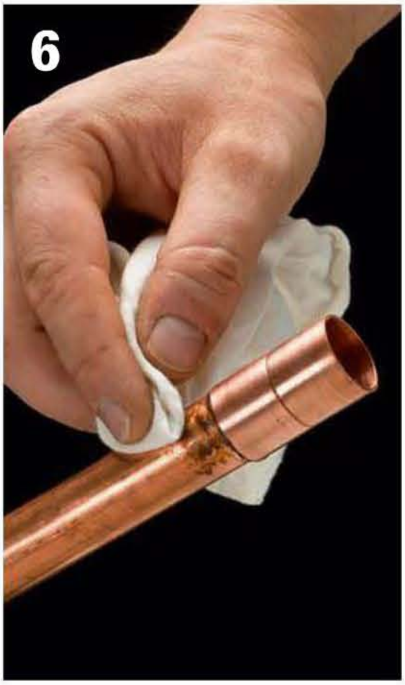

Use a clean dry cloth to remove excess flux before soldering the assembled fitting.

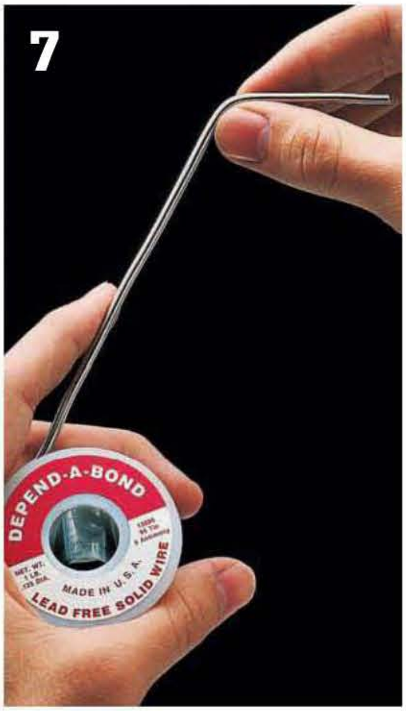
7
DEPEND-A-BOND
MADE IN U.S.A.
LEAD FREE SOLID WIRE

Prepare the wire solder by unwinding 8" to 10" of wire from spool. Bend the first 2" of the wire to a 90º angle.

Open the gas valve and trigger the spark lighter to ignite the torch.

9
Quickfire

Adjust the torch valve until the inner portion of the flame is 1" to 2" long.

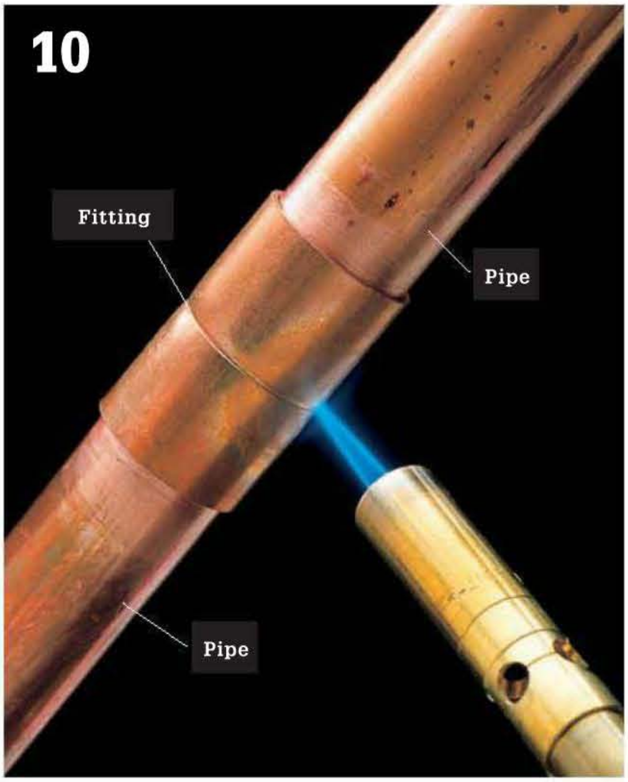
10
Fitting
Pipe
Pipe

Move the torch flame back and forth and around the pipe and the fitting to heat the area evenly.

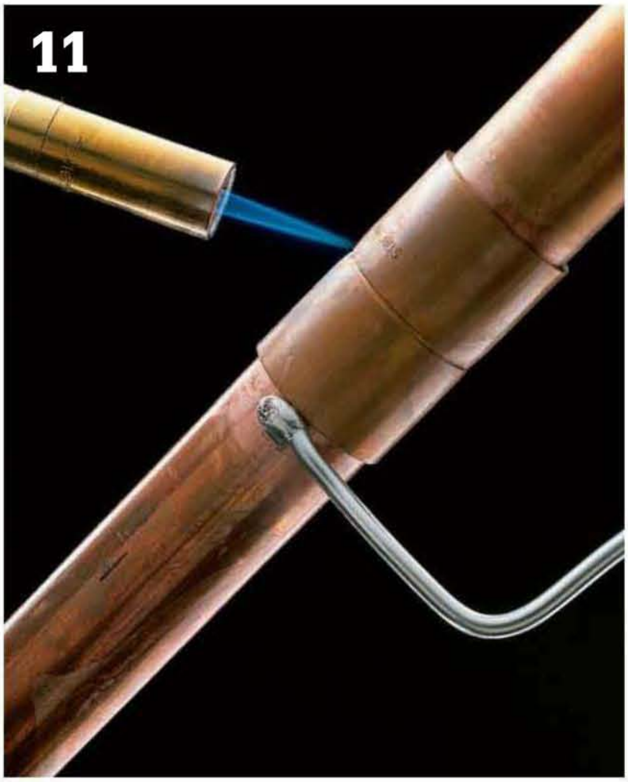

Heat the other side of the copper fitting to ensure that heat is distributed evenly. Touch solder to pipe. Solder will melt when the pipe is at the right temperature.

(continued)

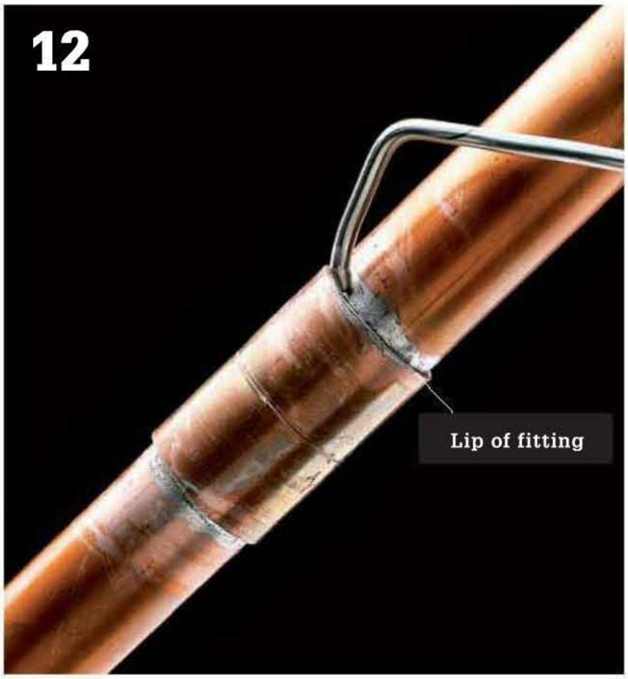

When solder melts, remove the torch and quickly push ½" to ¾" of solder into each joint. Capillary action fills the joint with liquid solder. A correctly soldered joint should show a thin bead of solder around the lips of the fitting.

Allow the joint to cool briefly, then wipe away excess solder with a dry rag. *Caution: Pipes will be hot.* If joints leak after water is turned on, disassemble and resolder.

How to Solder Brass Valves

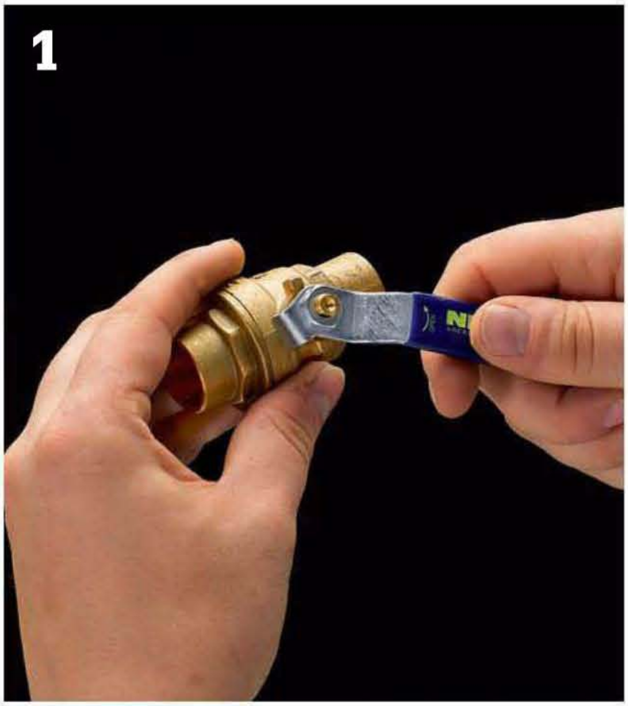

Valves should be fully open during all stages of the soldering process.

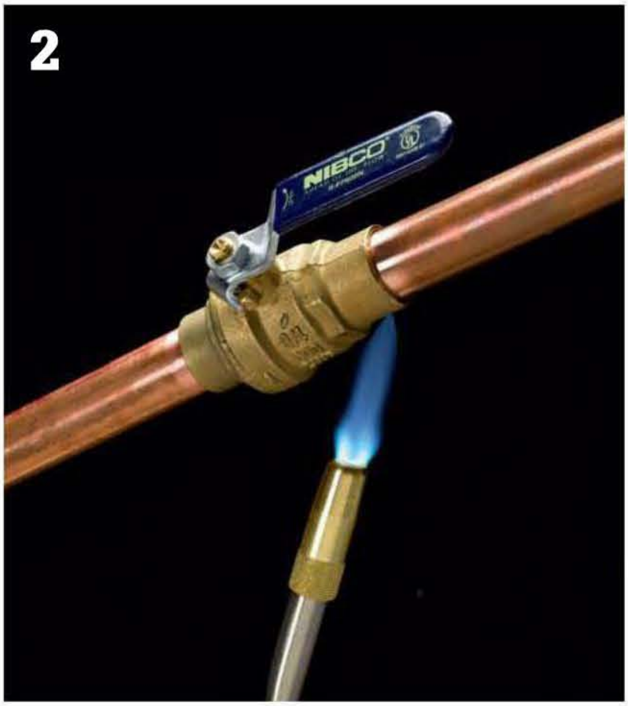

To prevent valve damage, quickly heat the pipe and the flanges of the valve, not the valve body. After soldering, cool the valve by spraying with water.

How to Take Apart Soldered Joints

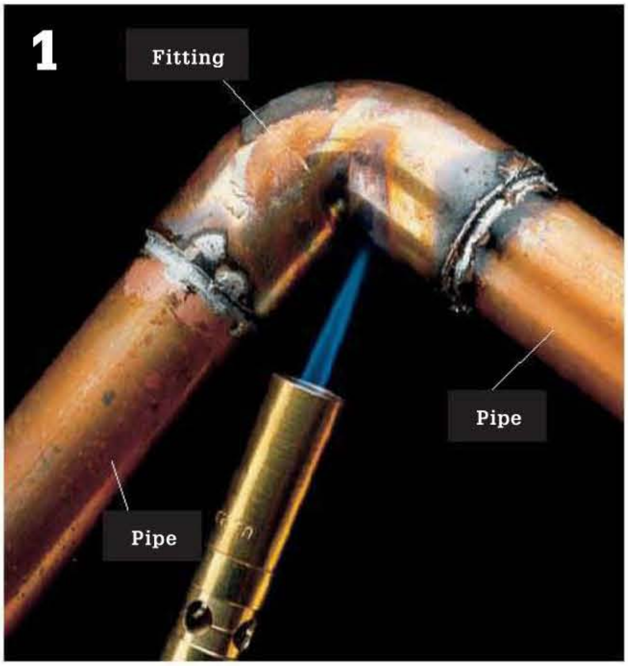

Turn off the water and drain the pipes by opening the highest and lowest faucets in the house. Light your torch. Hold the flame tip to the fitting until the solder becomes shiny and begins to melt.

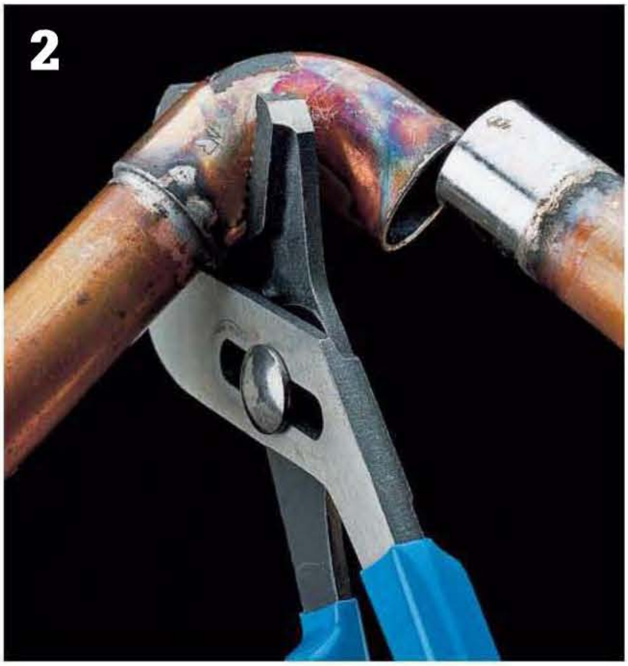

Use channel-type pliers to separate the pipes from the fitting.

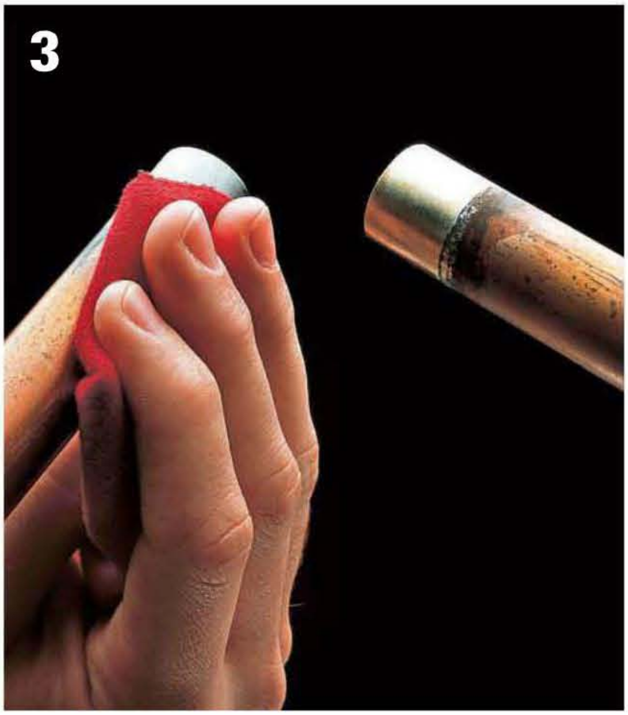

Remove old solder by heating the ends of the pipe with your torch. Use a dry rag to wipe away melted solder quickly. *Caution: Pipes will be hot.*

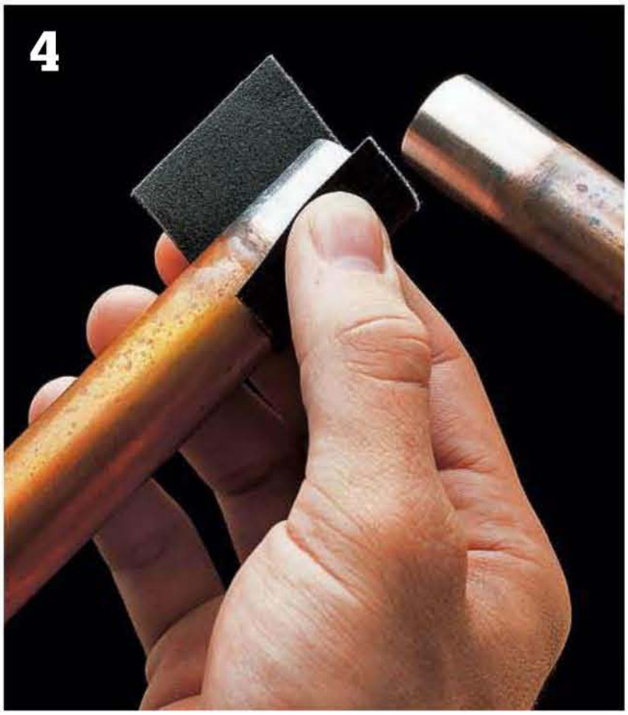

Use emery cloth to polish the ends of the pipe down to bare metal. Never reuse fittings.

Working with Plastic Pipe

Cut rigid ABS, PVC, or CPVC plastic pipes with a tubing cutter or with any saw. Cuts must be straight to ensure watertight joints.

Rigid plastics are joined with plastic fittings and solvent glue. Use a solvent glue that is made for the type of plastic pipe you are installing. For example, use ABS solvent on ABS pipe. Some solvent glues, called "all-purpose" or "universal" solvents, may be used on all types of plastic pipe, but create an inferior bond.

Solvent glue hardens in about 30 seconds, so test-fit all plastic pipes and fittings before gluing the first joint. For best results, the surfaces of plastic pipes and fittings should be dulled with emery cloth and liquid primer before they are joined.

Liquid solvent glues and primers are toxic and flammable. Provide adequate ventilation when fitting plastics, and store the products away from any source of heat.

Plastic grip fittings can be used to join rigid or flexible plastic pipes to copper plumbing pipes.

Tools & Materials ▸

- Tape measure
- Felt-tipped pen
- Tubing cutter
  - (or miter box or hacksaw)
- Utility knife
- Channel-type pliers
- Gloves
- Plastic pipe
- Fittings
- Emery cloth
- Plastic pipe primer
- Solvent glue
- Rag
- Petroleum jelly

Solvent welding is a chemical bonding process used to permanently join PVC pipes and fittings.

Oatey
MEDIUM • ORANGE
CPVC CEMENT
Oatey
BLACK
ABS CEMENT
MEDIUM • CLEAR
FLEXIBLE PVC CEMENT
Oatey
CLEANER
Oatey
PURPLE PRIMER
DANGER

Primer and solvent glue are specific to the plumbing material being used. Do not use all-purpose or multi-purpose products. Light to medium body glues are appropriate for DIYers as they allow the longest working time and are easiest to use. The products work best when fresh, so buy small containers and dispose of any unused product after a few months.

How to Cut Rigid Plastic Pipe

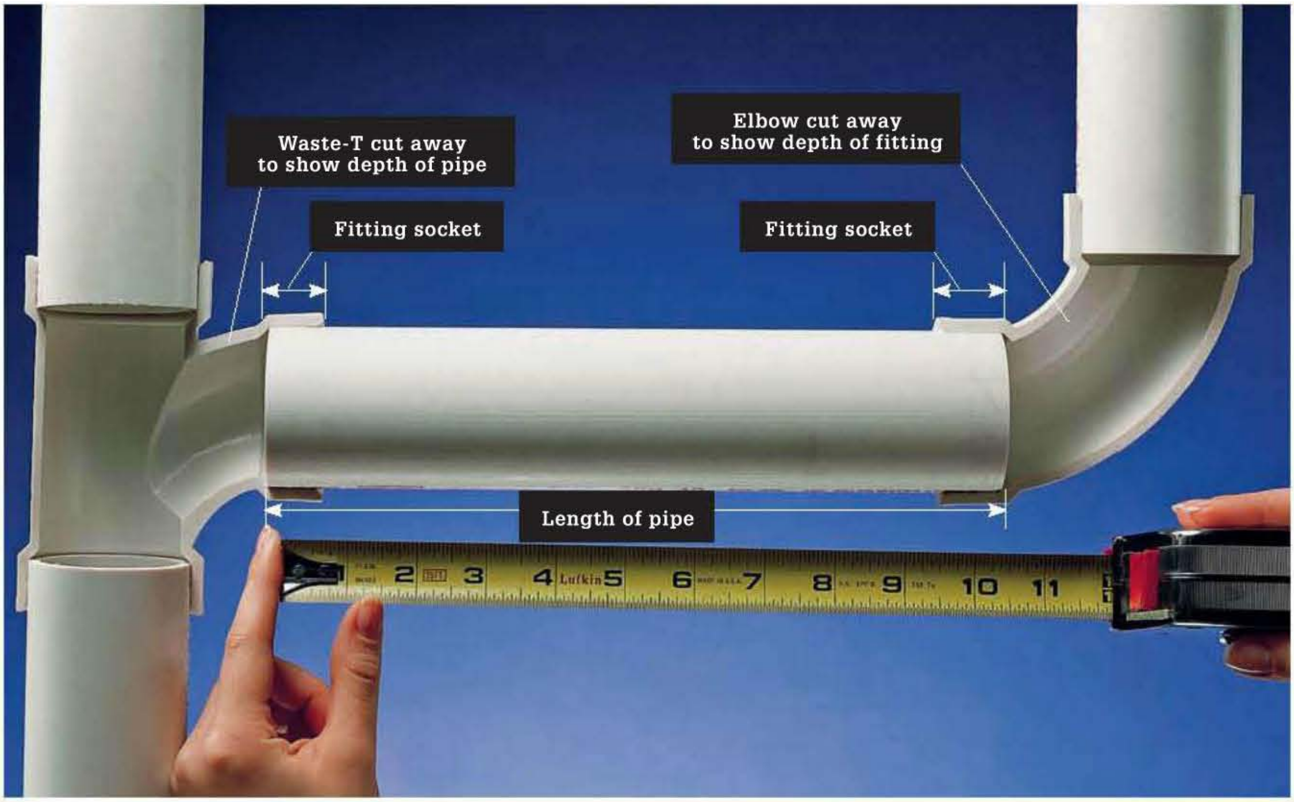
Waste-T cut away
to show depth of pipe
Elbow cut away
to show depth of fitting
Fitting socket
Fitting socket
Length of pipe

Find the length of plastic pipe needed by measuring between the bottoms of the fitting sockets (fittings shown in cutaway). Mark the length on the pipe with a felt-tipped pen.

The best cutting tool for plastic pipe is a power miter saw with a fine tooth woodworking blade or a plastic-specific blade.

A ratcheting plastic-pipe cutter can cut smaller diameter PVC and CPVC pipe in a real hurry. If you are plumbing a whole house you may want to consider investing in one. They also are sold only at plumbing supply stores.

How to Solvent-Glue Rigid Plastic Pipe

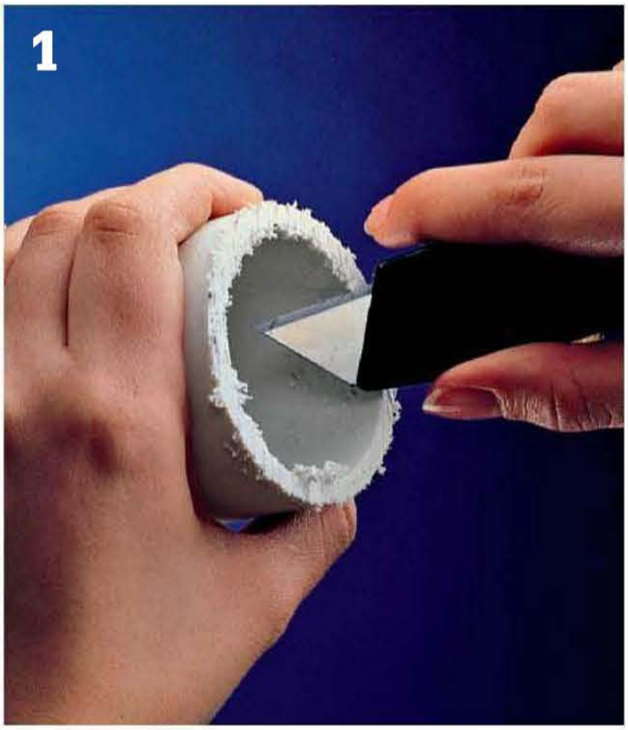

Remove rough burrs on cut ends of plastic pipe, using a utility knife.

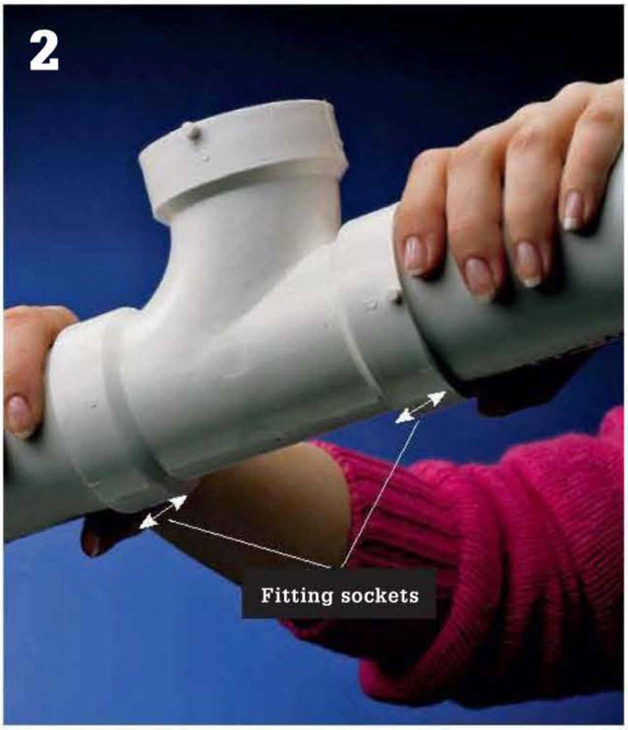

Test-fit all pipes and fittings. Pipes should fit tightly against the bottom of the fitting sockets.

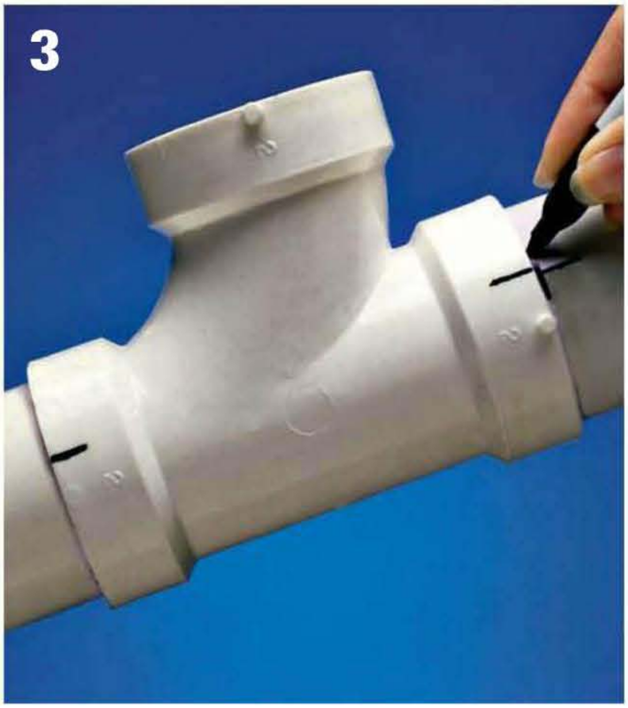

Mark the depth of the fitting sockets on the pipes. Take pipes apart. Clean the ends of the pipes and fitting sockets with emery cloth.

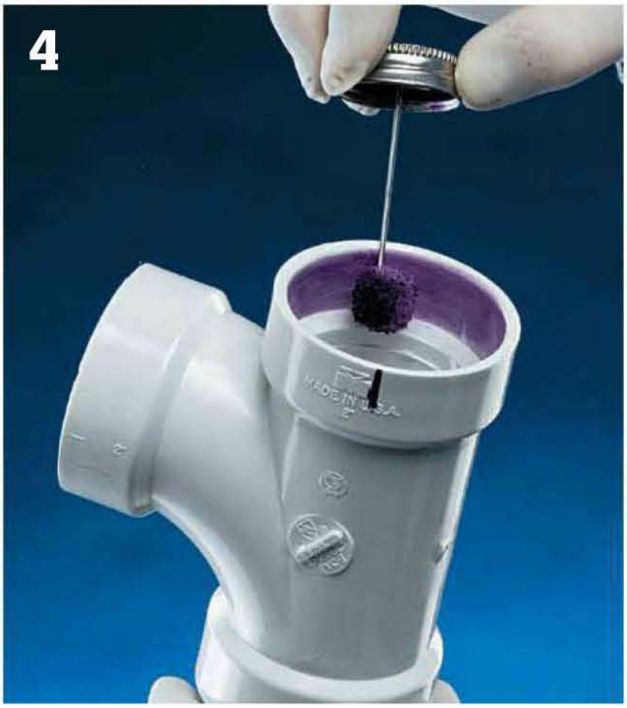

Apply a light coat of plastic pipe primer to the ends of the pipes and to the insides of the fitting sockets. Primer dulls glossy surfaces and ensures a good seal.

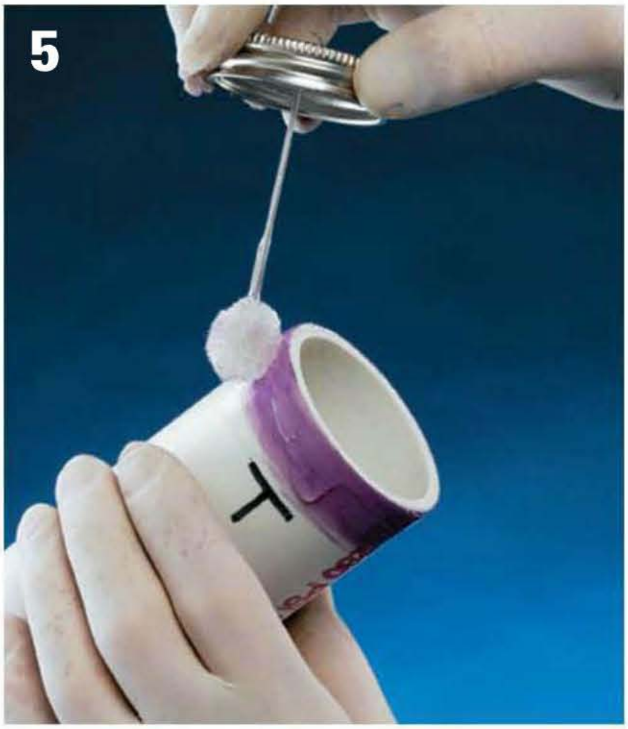

Solvent-glue each joint by applying a thick coat of solvent glue to the end of the pipe. Apply a thin coat of solvent glue to the inside surface of the fitting socket. Work quickly: solvent glue hardens in about 30 seconds.

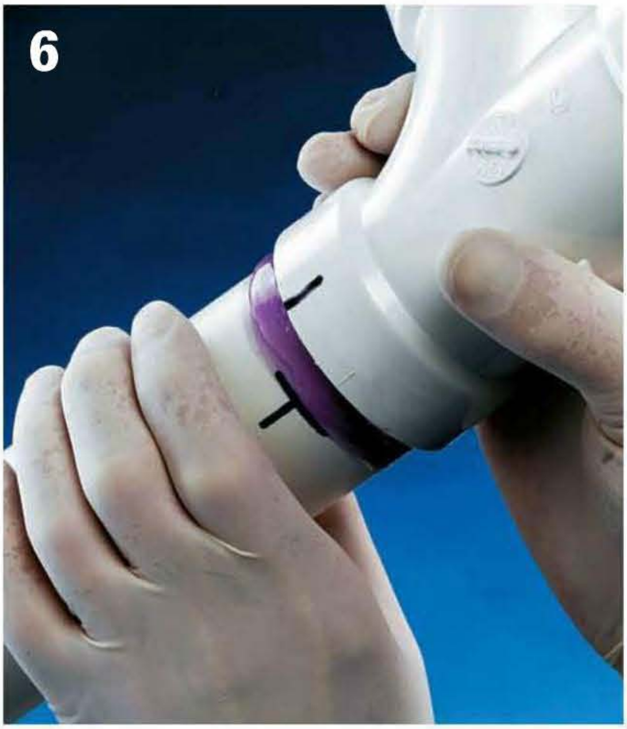

Quickly position the pipe and fitting so that the alignment marks are offset by about 2". Force the pipe into the fitting until the end fits flush against the bottom of the socket.

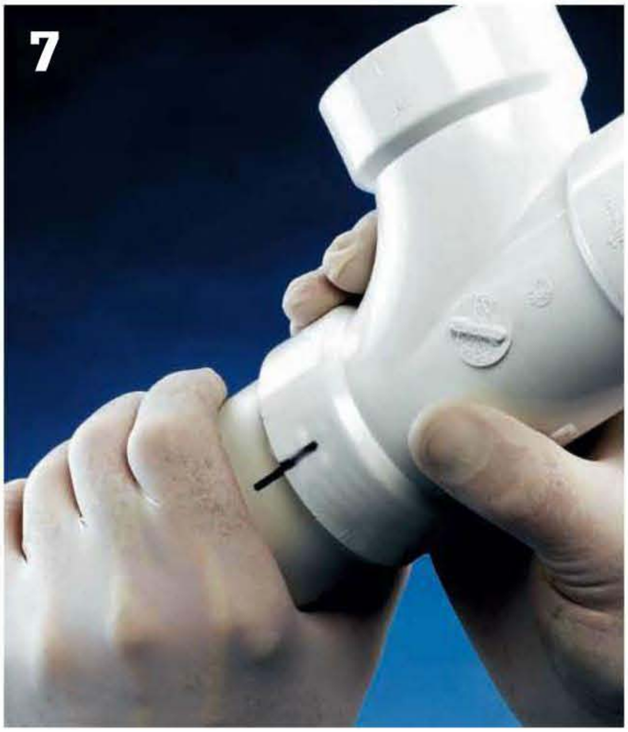

Spread solvent by twisting the pipe until the marks are aligned. Hold the pipe in place for about 20 seconds to prevent the joint from slipping.

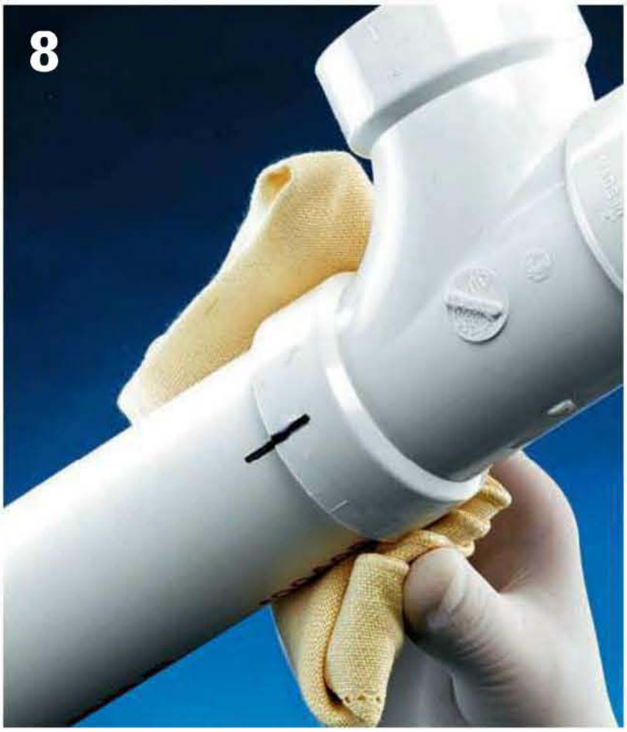

Wipe away excess solvent glue with a rag. Do not disturb the joint for 30 minutes after gluing.

Working with PEX Pipe

Cross-linked polyethylene (PEX) is growing quickly in acceptance as a supply pipe for residential plumbing. It's not hard to understand why. Developed in the 1960s but relatively new to the United States, this supply pipe combines the ease of use of flexible tubing with the durability of rigid pipe. It can withstand a wide temperature range (from subfreezing to 180° F); it is inexpensive; and it's quieter than rigid supply pipe.

PEX is flexible plastic (polyethylene, or PE) tubing that's reinforced by a chemical reaction that creates long fibers to increase the strength of the material. It has been allowed by code in Europe and the southern United States for many years, but has won approval for residential supply use in most major plumbing codes only recently. It's frequently used in manufactured housing and recreational vehicles and in radiant heating systems. Because it is so flexible, PEX can easily be bent to follow corners and make other changes in direction. From the water main and heater, it is connected into manifold fittings that redistribute the water in much the same manner as a lawn irrigation system.

For standard residential installations, PEX can be joined with very simple fittings and tools. Unions are generally made with a crimping tool and a crimping ring. You simply insert the ends of the pipe you're joining into the ring, then clamp down on the ring with the crimping tool. PEX pipe, tools, and fittings can be purchased from most wholesale plumbing suppliers and at many home centers. Coils of PEX are sold in several diameters from ¼" to 1". PEX tubing and fittings from different manufacturers are not interchangeable. Any warranty coverage will be voided if products are mixed.

Tools & Materials ▸

- Tape measure
- Felt-tipped pen
- Full-circle crimping tool
- Go/no-go gauge
- Tubing cutter
- PEX pipe
- Manifolds
- Protector plates
- PEX fittings
- Utility knife
- Plastic hangers
- Crimp ring

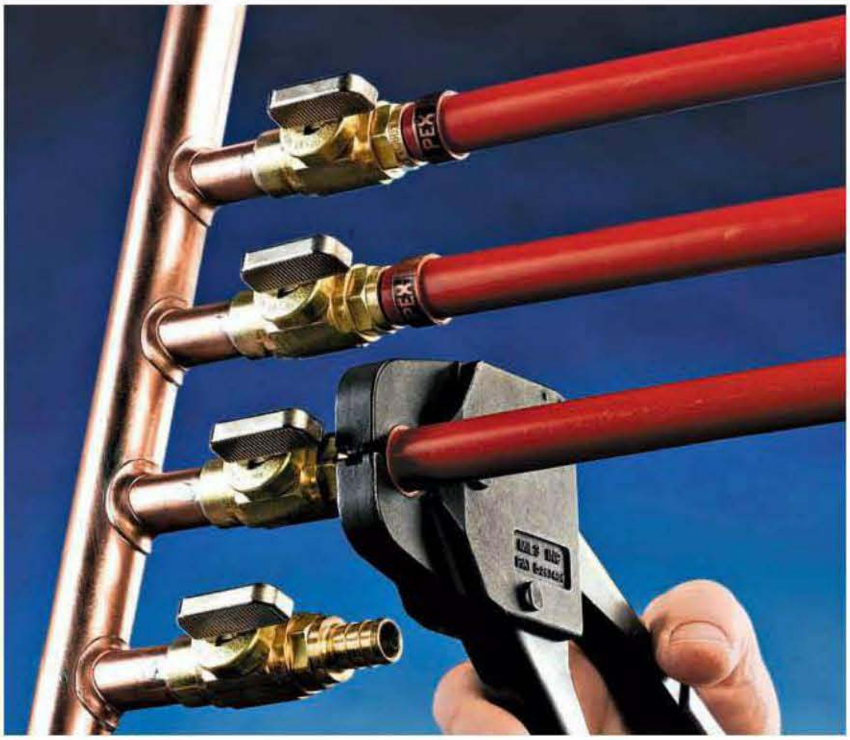

PEX pipe is a relatively new water supply material that's growing in popularity, in part because it can be installed with simple mechanical connections. In this example, we see a standard 4-outlet copper manifold, with branch water lines being attached with a crimping tool. To ensure leak-free joints, each joint is carefully measured with a go/no-go tool before the water is turned on.

PEX Tools & Materials

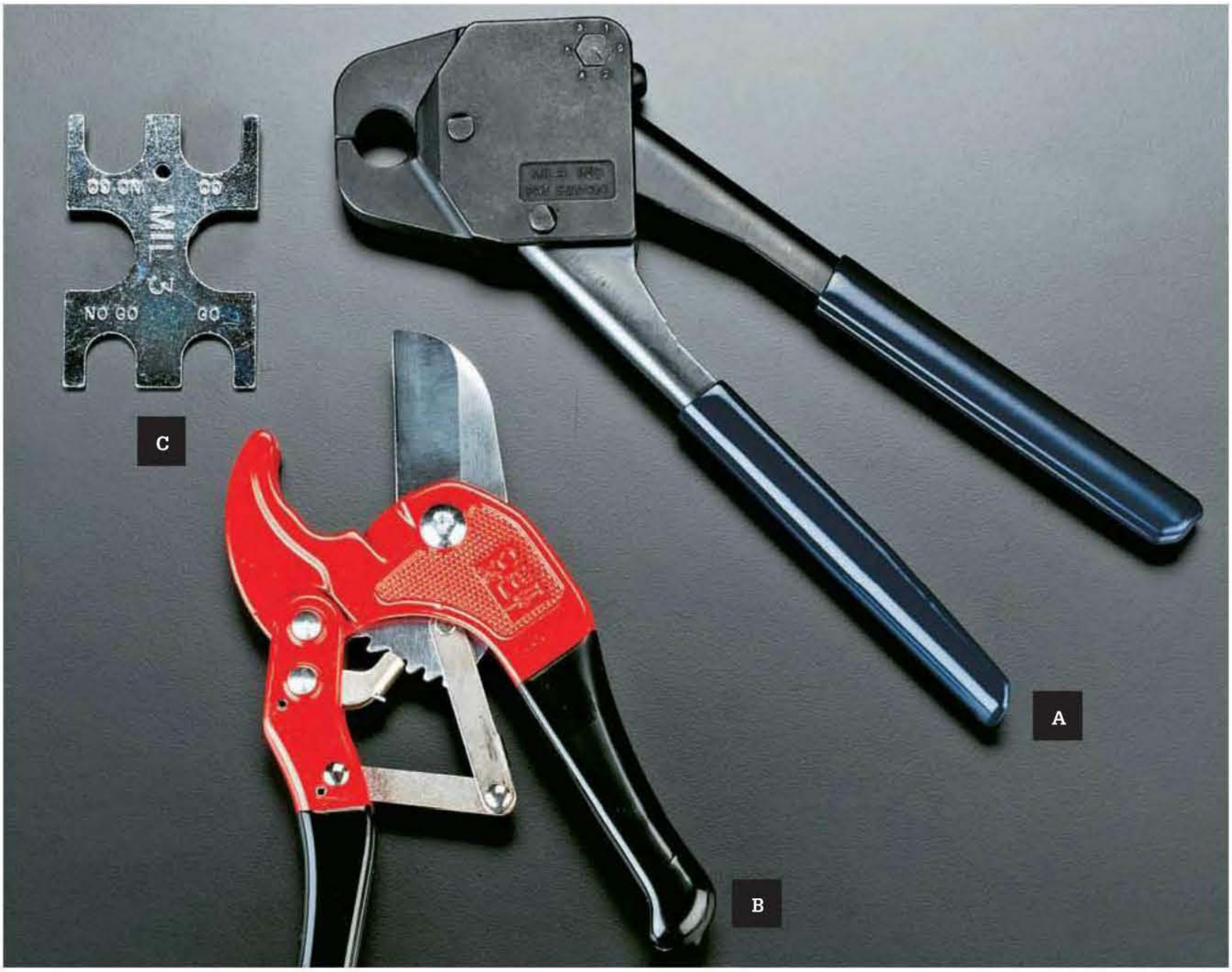

Specialty tools for installing PEX are available wherever PEX is sold. The basic set includes a full-circle crimping tool (A), a tubing cutter (B), and a go/no-go gauge (C) to test connections after they've been crimped.

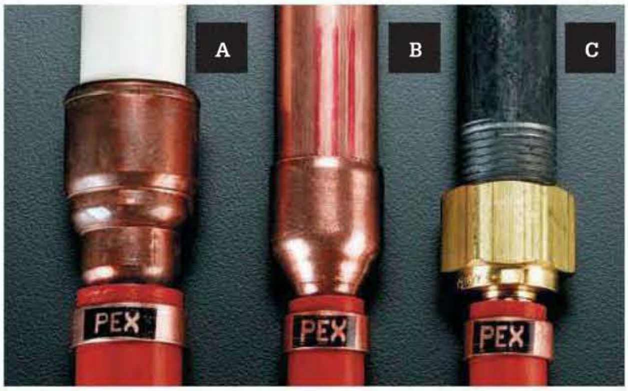

PEX is connected to other water supply materials with transition fittings, including CPVC-to-PEX (A), copper-to-PEX (B), and iron-to-PEX (C).

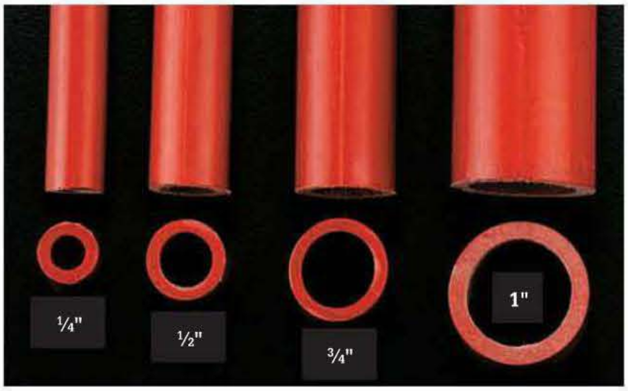

Generally, you should use the same diameter PEX as is specified for rigid supply tubing, but in some "home run" installations (see page 72) you can use ⅜" PEX where ½" rigid copper would normally be used.

PEX Installation

Check with your local plumbing inspector to verify that PEX is allowed in your municipality. PEX has been endorsed by all major plumbing codes in North America, but your municipality may still be using an older set of codes. Follow the guidelines below when installing PEX:

- Do not install PEX in above-ground exterior applications because it degrades quickly from UV exposure.
- Do not use PEX for gas lines.
- Do not use plastic solvents or petroleum-based products with PEX (they can dissolve the plastic).
- Keep PEX at least 12" away from recessed light fixtures and other potential sources of high heat.
- Do not attach PEX directly to a water heater. Make connections at the heater with metallic tubing (either flexible water-heater connector tubing or rigid copper) at least 18" long; then join it to PEX with a transition fitting.
- Do not install PEX in areas where there is a possibility of mechanical damage or puncture. Always fasten protective plates to wall studs that house PEX.
- Always leave some slack in installed PEX lines to allow for contraction and in case you need to cut off a bad crimp.
- Use the same minimum branch and distribution supply-pipe dimensions for PEX that you'd use for copper or CPVC, according to your local plumbing codes.

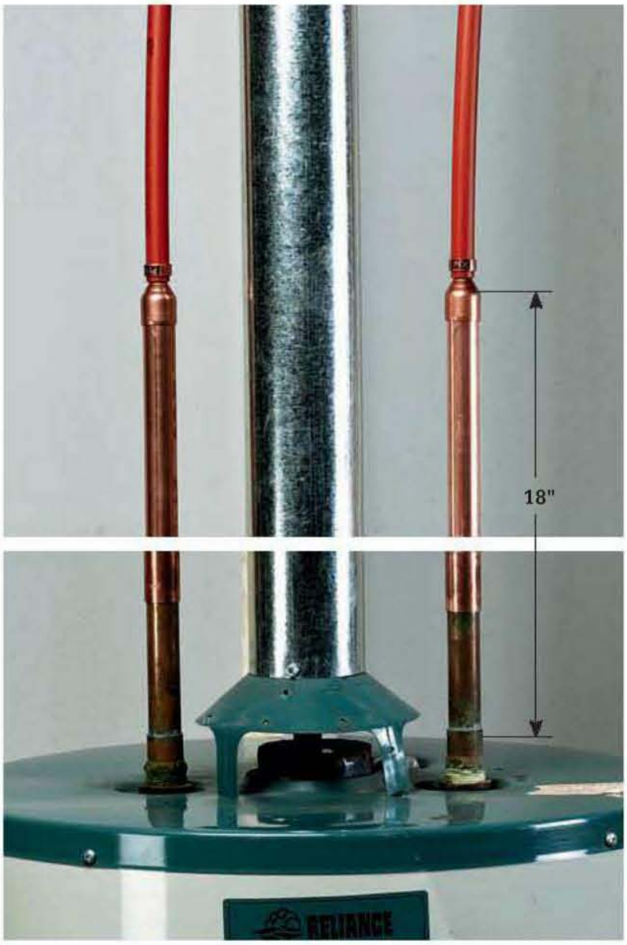

Do not connect PEX directly to a water heater. Use metal connector tubes. Solder the connector tubes to the water heater before attaching PEX. Never solder metal tubing that is already connected to PEX lines.

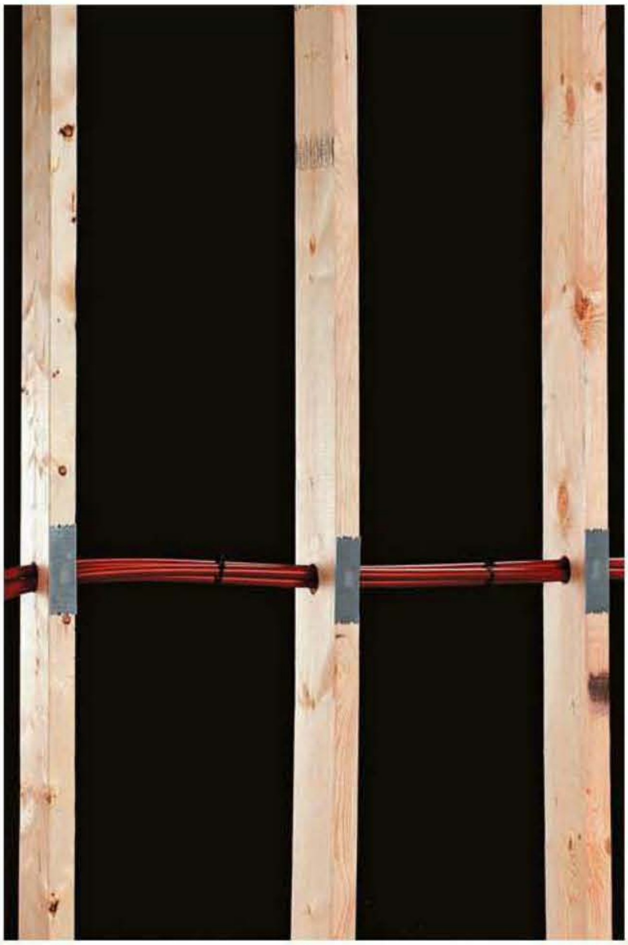

Bundle PEX together with plastic ties when running pipe through wall cavities. PEX can contract slightly, so leave some slack in the lines.

Buying PEX

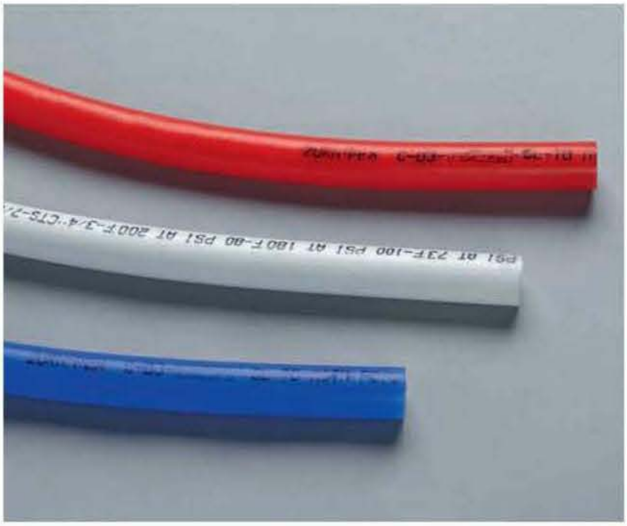

Color coding is a practice many PEX manufacturers have embraced to make identification easier. Because the material is identical except for the color, you can buy only one color (red is more common) and use it for both hot and cold supply lines.

PEX combines the flexibility of plastic tubing with the durability of rigid supply pipe. It is sold in coils of common supply-pipe diameters.

The PEX Advantage ▸

PEX supply tubing offers a number of advantages over traditional rigid supply tubing:

- Easy to install. PEX does not require coupling joints for long runs or elbows and sweeps for turns. The mechanical connections do not require solvents or soldering.
- Easy to transport. Large coils are lightweight and much easier to move around than 10-ft. lengths of pipe.
- Good insulation. The PEX material has better thermal properties than copper for lessened heat loss.
- Quiet. PEX will not rattle or clang from trapped air or kinetic energy.
- Good for retrofit jobs. PEX is easier to snake through walls than rigid supply tubing and is compatible with copper, PVC, or iron supply systems if the correct transition fittings are used. If your metal supply tubes are used to ground your electrical system, you'll need to provide a jumper if PEX is installed in midrun. Check with a plumber or electrician.
- Freeze resistance. PEX retains some flexibility in sub-freezing conditions and is less likely to be damaged than rigid pipe, but it is not frostproof.

General Codes for PEX ▸

PEX has been endorsed for residential use by all major building codes, although some municipal codes may be more restrictive. The specific design standards may also vary, but here are some general rules:

- For PEX, maximum horizontal support spacing is 32" and maximum vertical support spacing is 10 ft.
- Maximum length of individual distribution lines is 60 ft.
- PEX is designed to withstand 210° F water for up to 48 hours. For ongoing use, most PEX is rated for 180 degree water up to 100 pounds per square inch of pressure.
- Directional changes of more than 90 degrees require a guide fitting (see page 75).

System Designs

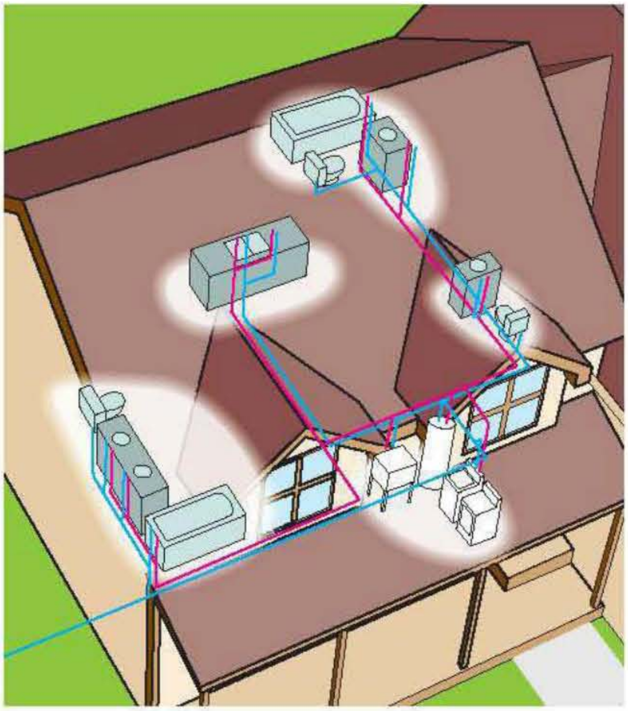

Trunk-and-branch systems are configured in much the same way as a traditional rigid copper or PVC supply systems. A main supply line (the trunk line) carries water to all of the outlets via smaller branch lines that tie into the trunk and serve a few outlets in a common location.

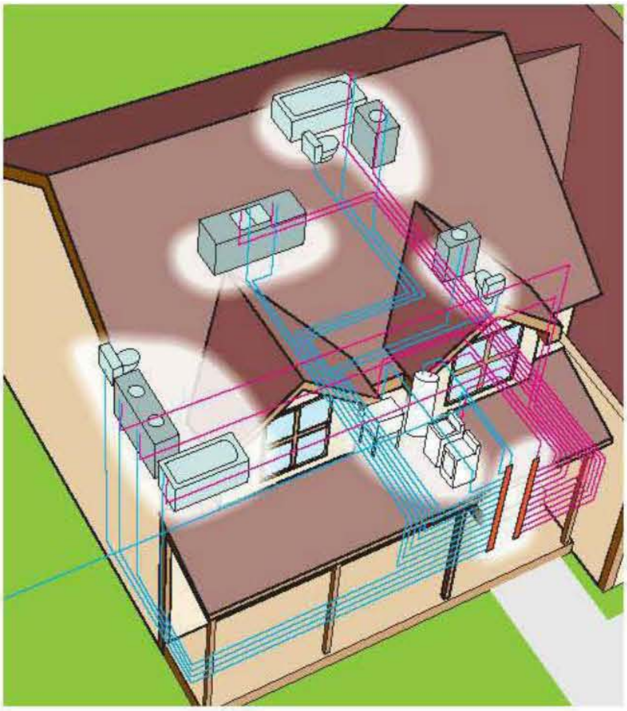

Home run systems rely on one or two central manifolds to distribute the hot and cold water very efficiently. Eliminating the branch fittings allows you to use thinner supply pipe in some situations.

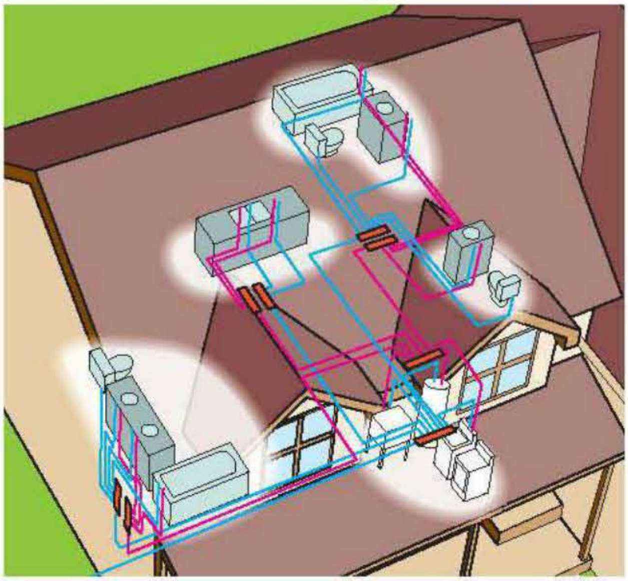

Remote manifold systems are a hybrid between traditional trunk-and-branch systems and home run systems. Instead of relying on just one or two manifolds, they employ several smaller manifolds downline from a larger manifold. Each smaller manifold services a group of fixtures, as in a bathroom or kitchen.

Choosing a PEX system ▸

- For maximum single-fixture water pressure: Trunk and branch
- For economy of materials: Trunk and branch or remote manifold
- For minimal wait times for hot water (single fixture): Home run
- For minimal wait times for hot water (multiple fixtures used at same approximate time): Trunk and branch or remote manifold
- For ease of shutoff control: Home run
- For lowest number of fittings and joints: Home run

How to Make PEX Connections

1

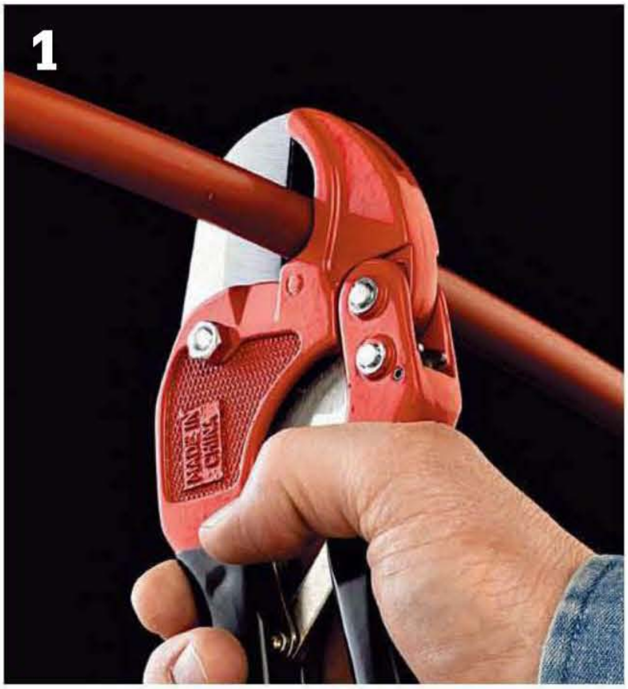

Cut the pipe to length, making sure to leave enough extra material so the line will have a small amount of slack once the connections are made. A straight, clean cut is very important. For best results, use a tubing cutter.

2

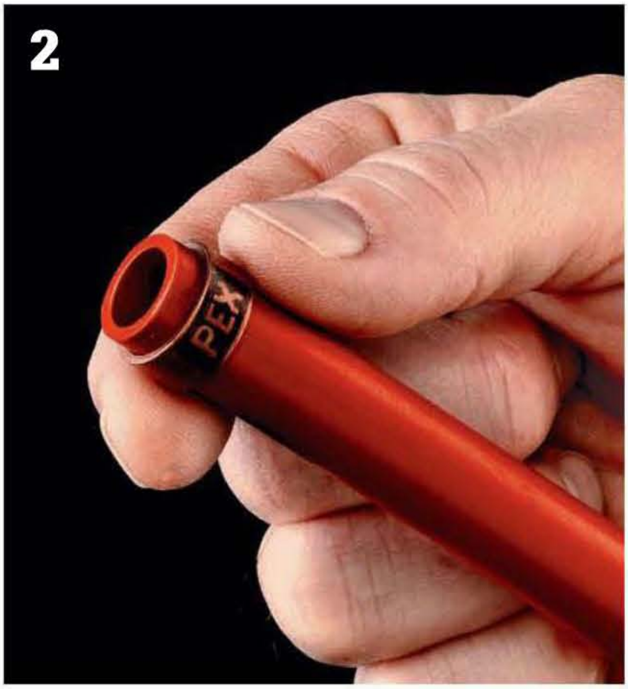

Inspect the cut end to make sure it is clean and smooth. If necessary, deburr the end of the pipe with a sharp utility knife. Slip a crimp ring over the end.

3

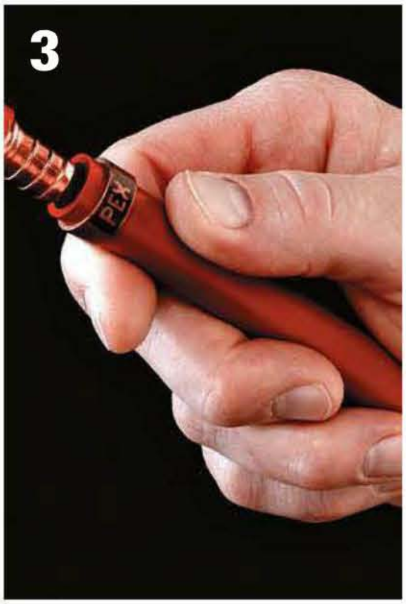

Insert the barbed end of the fitting into the pipe until it is snug against the cut edges.Position the crimp ring so it is ⅛" to ¼" from the end of the pipe, covering the barbed end of the fitting. Pinch the fitting to hold it in place.

4

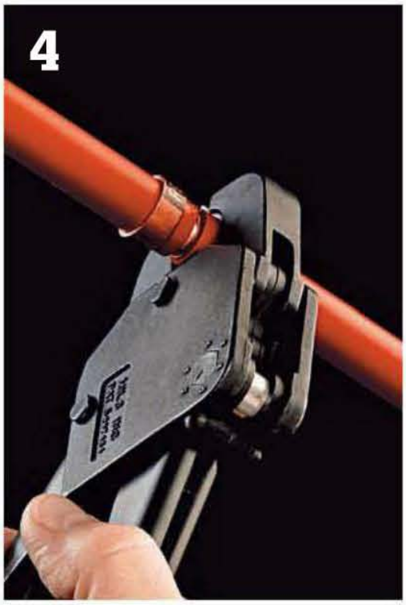

Align the jaws of a full-circle crimping tool over the crimp ring and squeeze the handles together to apply strong, even pressure to the ring.

5

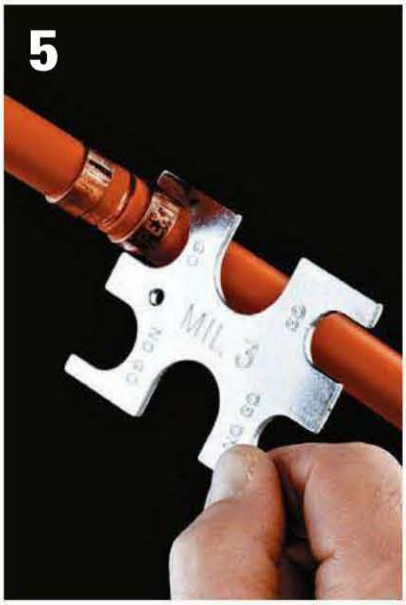

Test the connection to make sure it is mechanically acceptable, using a go/no-go gauge. If the ring does not fit into the gauge properly, cut the pipe near the connection and try again.

How to Plumb a PEX Water-Supply System

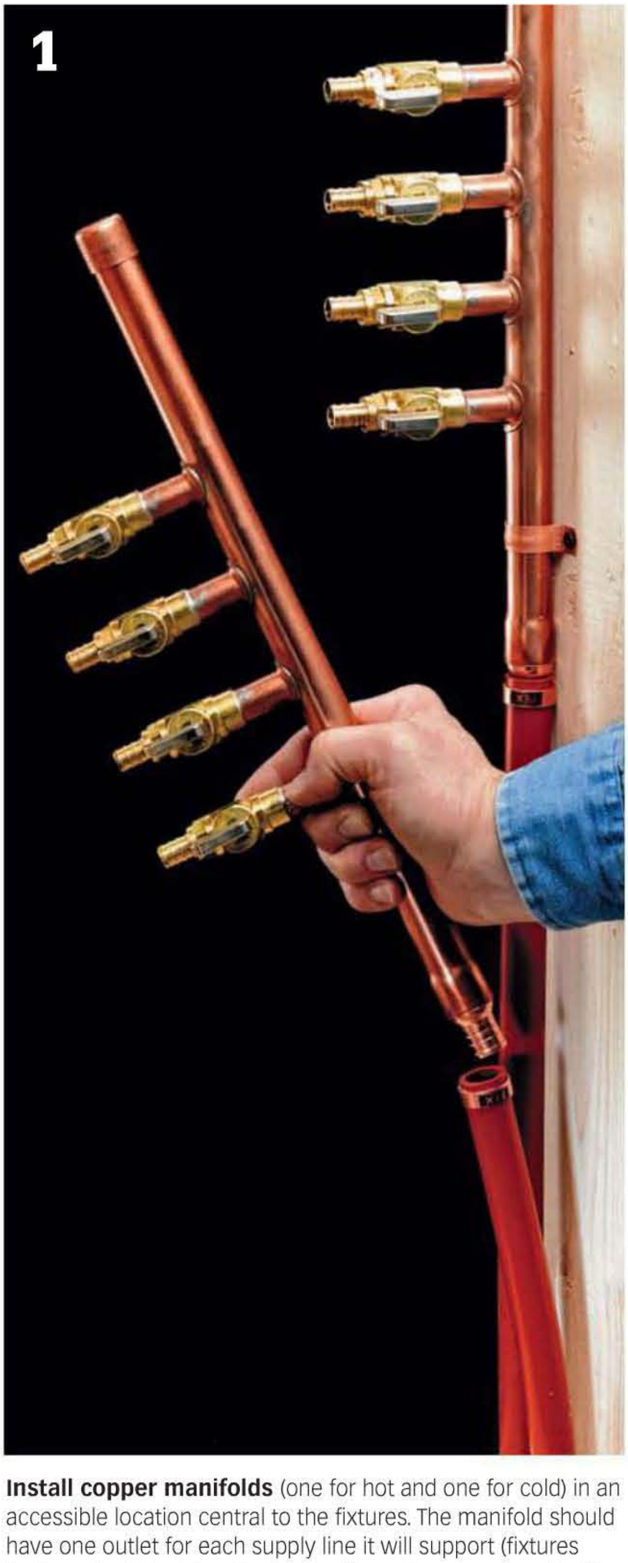

Install copper manifolds (one for hot and one for cold) in an accessible location central to the fixtures. The manifold should have one outlet for each supply line it will support (fixtures that require hot and cold supply will need a separate outlet for each). Run supply lines from the water heater and water main to the copper manifolds. Connect the supply pipes to the manifolds with crimp fittings.

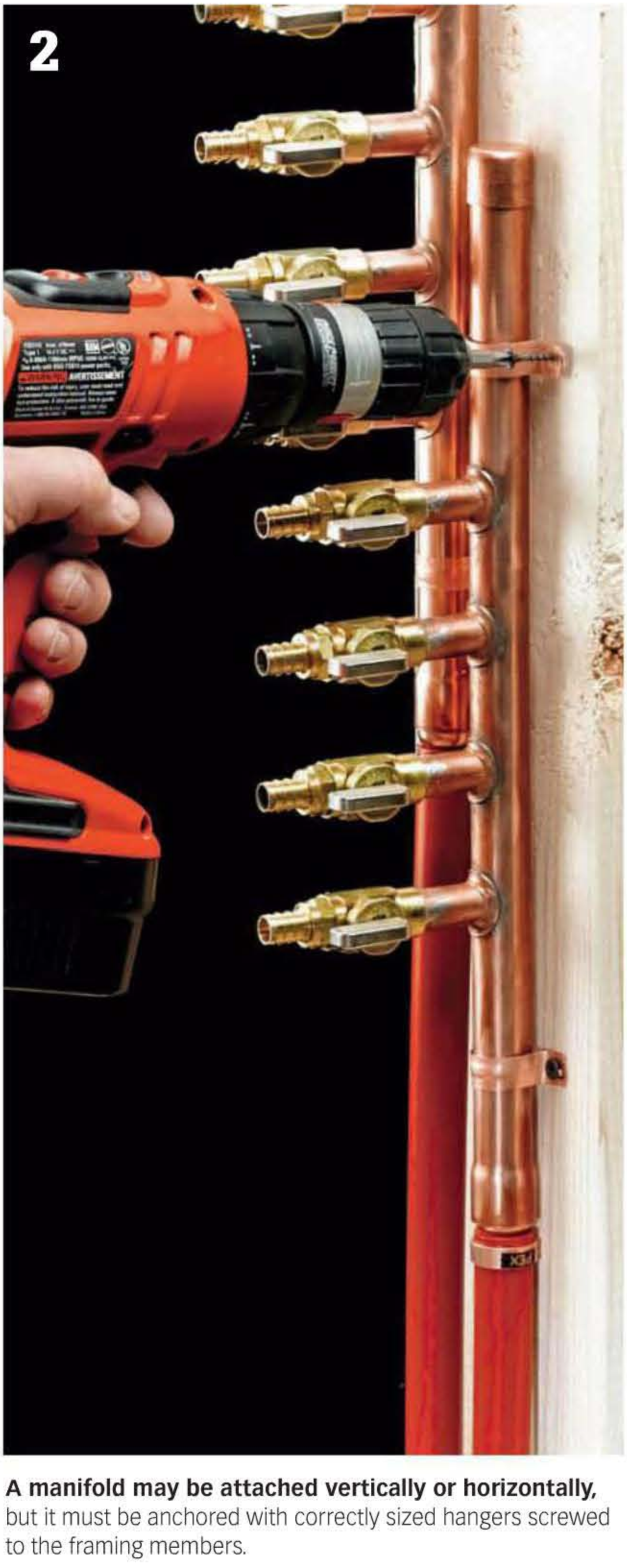

A manifold may be attached vertically or horizontally, but it must be anchored with correctly sized hangers screwed to the framing members.

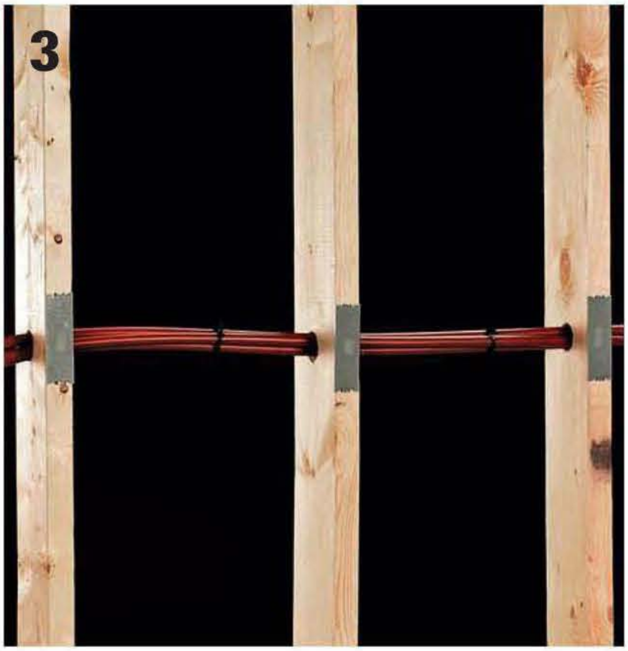

Starting at each fixture (and leaving at least 12" of extra pipe exposed), run appropriately sized PEX through holes in the framing to the manifolds. Pipes may be bundled together loosely with plastic ties. Protect the line with a nailing plate at each stud location. Be sure to leave some slack in the supply lines.

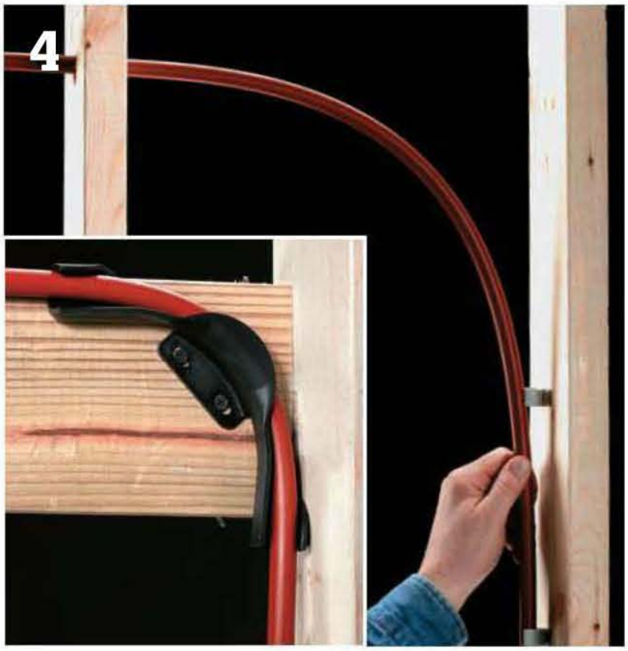

Support the pipe with a plastic hanger near every floor or ceiling and midway up vertical runs. Also use hangers to guide pipe near the beginnings and ends of curves and near fittings. Use a plastic guide for sharp curves (inset). Do not bend PEX so sharply that it kinks.

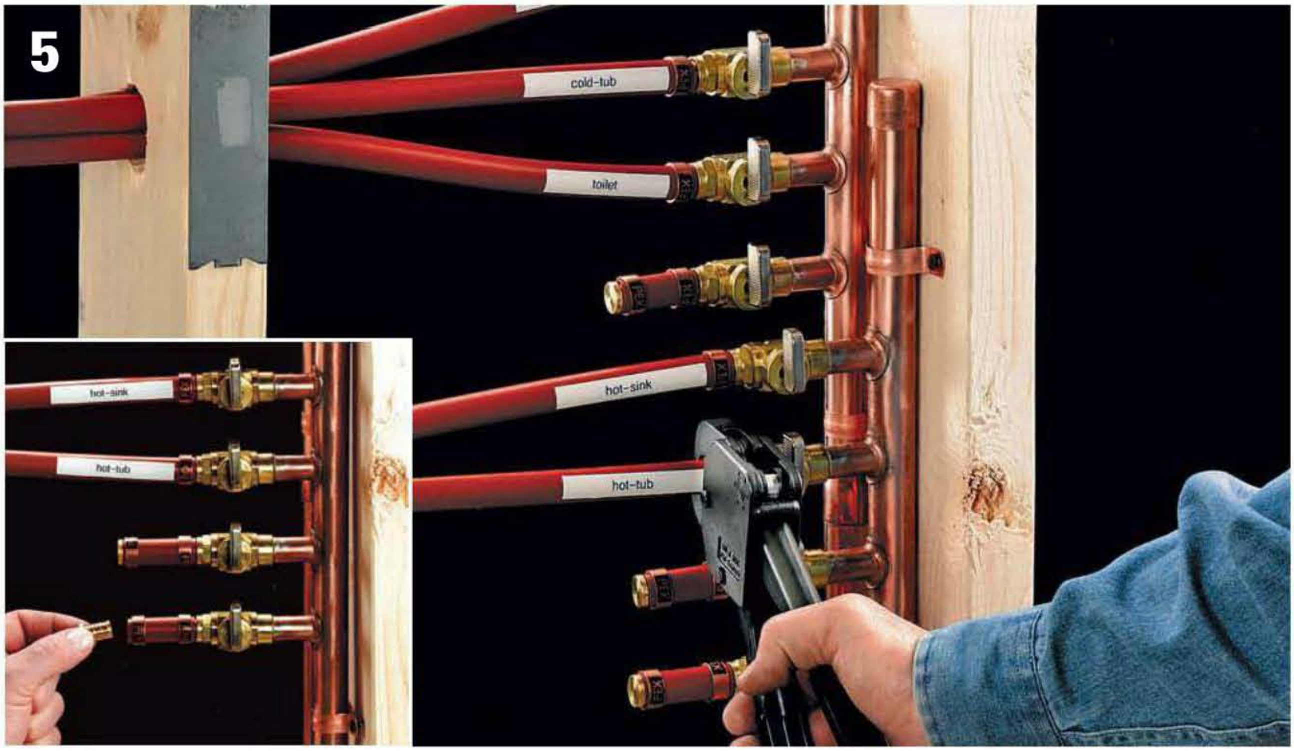

Cut each branch supply line to length (leave some extra in case you need to recrimp). Install shutoff valves for each outlet (most manifolds come with preattached valves). Connect the PEX branch supply lines to the shutoff valves. Label each pipe. Use a short length of PEX and a plug to seal any unused outlets (inset).

Planning New Plumbing

Start planning by drawing maps. Mapping your home's plumbing system is a good way to familiarize yourself with the plumbing layout and can help you when planning plumbing renovation projects. With a good map, you can envision the best spots for new fixtures and plan new pipe routes more efficiently. Maps also help in emergencies, when you need to locate burst or leaking pipes quickly.

Draw a plumbing map for each floor on tracing paper, so you can overlay floors and still read the information below. Make your drawings to scale and have all plumbing fixtures marked. Fixture templates and tracing paper are available at drafting supply stores.

Snoop around your basement for clues about the locations of supply, drain, vent, and gas pipes in your walls.

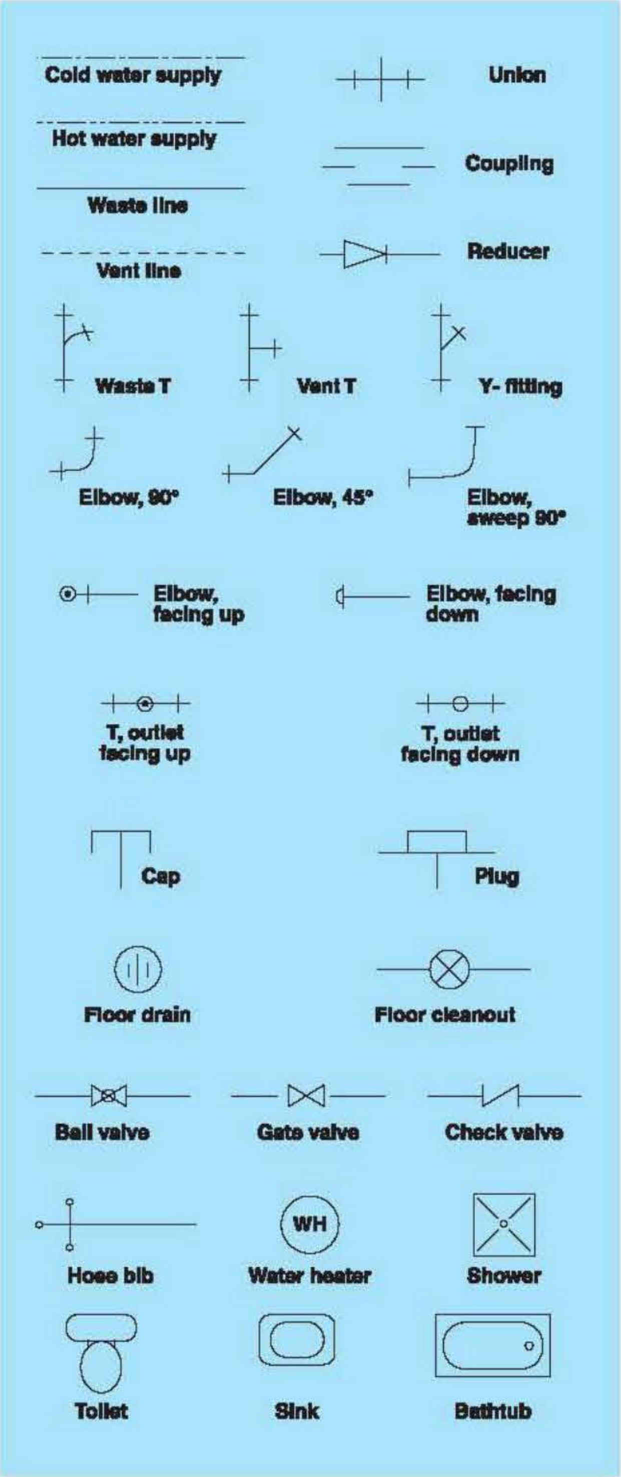
Cold water supply
Union
Hot water supply
Coupling
Waste line
Reducer
Vent line
Waste T
Vent T
Y- fitting
Elbow, 90°
Elbow, 45°
Elbow, sweep 90°
Elbow, facing up
Elbow, facing down
T, outlet facing up
T, outlet facing down
Cap
Plug
Floor drain
Floor cleanout
Ball valve
Gate valve
Check valve
Hose bib
WH
Water heater
Shower
Toilet
Sink
Bathtub

Use standard plumbing symbols on your map to identify the components of your plumbing system. These symbols will help you and your building inspector follow connections and transitions more easily.

Understanding Plumbing Codes

The plumbing code is the set of regulations that building officials and inspectors use to evaluate your project plans and the quality of your work. Codes vary from region to region, but most are based on the National Uniform Plumbing Code, the authority we used in the development of this book.

Code books are available for reference at bookstores and government offices. However, they are highly technical, difficult-to-read manuals. More user-friendly for do-it-yourselfers are the variety of code handbooks available at bookstores and libraries. These handbooks are based on the National Uniform Plumbing Code but are easier to read and include many helpful diagrams and photos.

Plumbing code handbooks sometimes discuss three different plumbing "zones" in an effort to accommodate variations in regulations from state to state. The states included in each zone are listed below.

Zone 1: Washington, Oregon, California, Nevada, Idaho, Montana, Wyoming, North Dakota, South Dakota, Minnesota, Iowa, Wisconsin, Nebraska, Kansas, Utah, Arizona, Colorado, New Mexico, Indiana, parts of Texas.

Zone 2: Alabama, Arkansas, Louisiana, Tennessee, North Carolina, Mississippi, Georgia, Florida, South Carolina, parts of Texas, parts of Maryland, parts of Delaware, parts of Oklahoma, parts of West Virginia.

Zone 3: Virginia, Kentucky, Missouri, Illinois, Michigan, Ohio, Pennsylvania, New York, Connecticut, Massachusetts, Vermont, New Hampshire, Rhode Island, New Jersey, parts of Delaware, parts of West Virginia, parts of Maine, parts of Maryland, parts of Oklahoma.

Remember that your local plumbing code always supersedes the national code. Local codes may be more restrictive than the national code. Your local building inspector is a valuable source of information and may provide you with a convenient summary sheet of the regulations that apply to your project.

INSPECTION RECORD
NOTICE
NATIONAL PLUMBING CODES HANDBOOK
UNIFORM BUILDING CODE

The plumbing inspector is the final authority when it comes to evaluating your work. By visually examining and testing your new plumbing, the inspector ensures that your work is safe and functional.

GETTING A PERMIT

To ensure public safety, your community requires that you obtain a permit for most plumbing projects, including most of the projects demonstrated in this book.

When you visit your city building inspection office to apply for a permit, the building official will want to review three drawings of your plumbing project: a site plan, a water supply diagram, and a drain-waste-vent diagram. These drawings are described on this page. If the official is satisfied that your project meets code requirements, he or she will issue you a plumbing permit, which is your legal permission to begin work. The building official also will specify an inspection schedule for your project. As your project nears completion, you will be asked to arrange for an inspector to visit your home while the pipes are exposed to review the installation and ensure its safety.

Although do-it-yourselfers often complete complex plumbing projects without obtaining a permit or having the work inspected, we strongly urge you to comply with the legal requirements in your area. A flawed plumbing system can be dangerous, and it can potentially threaten the value of your home.

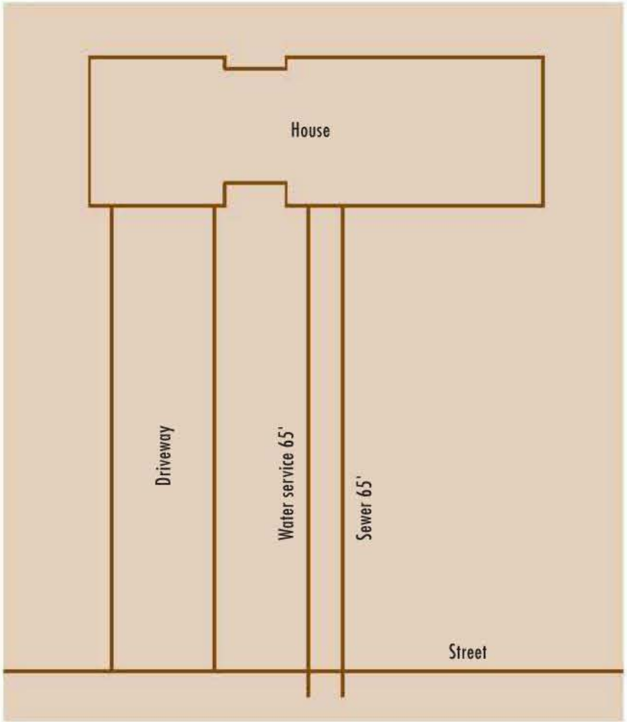
House
Driveway
Water service 65'
Sewer 65'
Street

The site plan shows the location of the water main and sewer main with respect to your yard and home. The distances from your foundation to the water main and from the foundation to the main sewer should be indicated on the site plan.

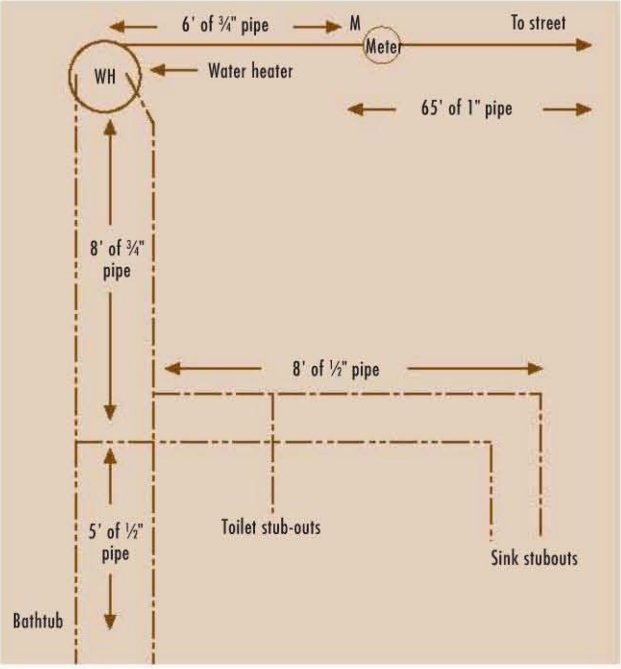
6' of ¾" pipe
M
Meter
To street
WH
Water heater
65' of 1" pipe
8' of ¾" pipe
8' of ½" pipe
5' of ½" pipe
Toilet stub-outs
Sink stubouts
Bathtub

The supply riser diagram shows the length of the hot and cold water pipes and the relation of the fixtures to one another. The inspector will use this diagram to determine the proper size for the new water supply pipes in your system.

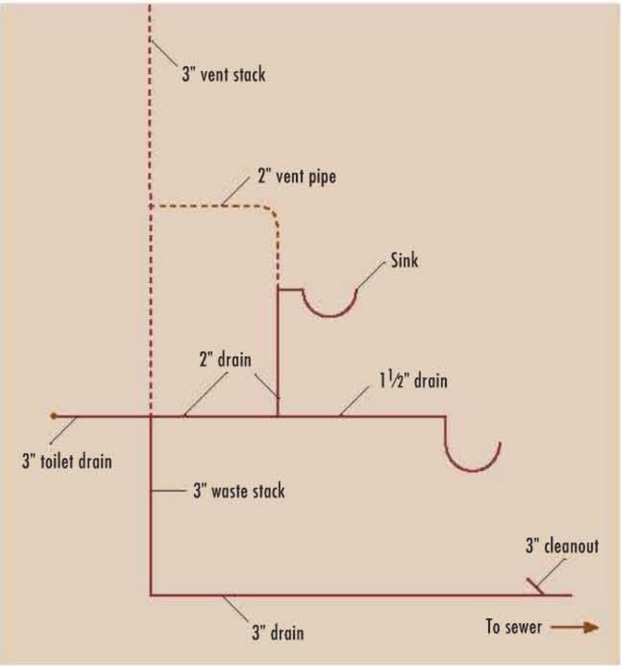
3" vent stack
2" vent pipe
Sink
2" drain
1½" drain
3" toilet drain
3" waste stack
3" cleanout
3" drain
To sewer

A DWV diagram shows the routing of drain and vent pipes in your system. Make sure to indicate the lengths of drain pipes and the distances between fixtures. The inspector will use this diagram to determine if you have properly sized the drain traps, drain pipes, and vent pipes in your project.

Sizing for Water Distribution Pipes

Fixture	Unit rating
Toilet	3
Vanity sink	1
Shower	2
Bathtub	2
Dishwasher	2
Kitchen sink	2
Clothes washer	2
Utility sink	2
Sillcock	3

Size of service pipe from street	Size of distribution pipe from water meter	Maximum length (ft.)—total fixture units					
		40	60	80	100	150	200
¾"	½"	9	8	7	6	5	4
¾"	¾"	27	23	19	17	14	11
¾"	1"	44	40	36	33	28	23
1"	1"	60	47	41	36	30	25
1"	1¼"	102	87	76	67	52	44

Water distribution pipes are the main pipes extending from the water meter throughout the house, supplying water to the branch pipes leading to individual fixtures. To determine the size of the distribution pipes, you must first calculate the total demand in "fixture units" (above, left) and the overall length of the water supply lines, from the street hookup through the water meter and to the most distant fixture in the house. Then, use the second table (above, middle) to find the minimum size for the water distribution pipes. Note that the fixture unit capacity depends partly on the size of the street-side pipe that delivers water to your meter.

Sizes for Branch Pipes & Supply Tubes

Fixture	Min. branch pipe size	Min. supply tube size
Toilet	½"	⅜"
Vanity sink	½"	⅜"
Shower	½"	½"
Bathtub	½"	½"
Dishwasher	½"	½"
Kitchen sink	½"	½"
Clothes washer	½"	½"
Utility sink	½"	½"
Sillcock	¾"	N.A.
Water heater	¾"	N.A.

Branch pipes are the water supply lines that run from the distribution pipes toward the individual fixtures. Supply tubes are the vinyl, chromed copper, or braided tubes that carry water from the branch pipes to the fixtures. Use the chart above as a guide when sizing branch pipes and supply tubes.

Valve Requirements

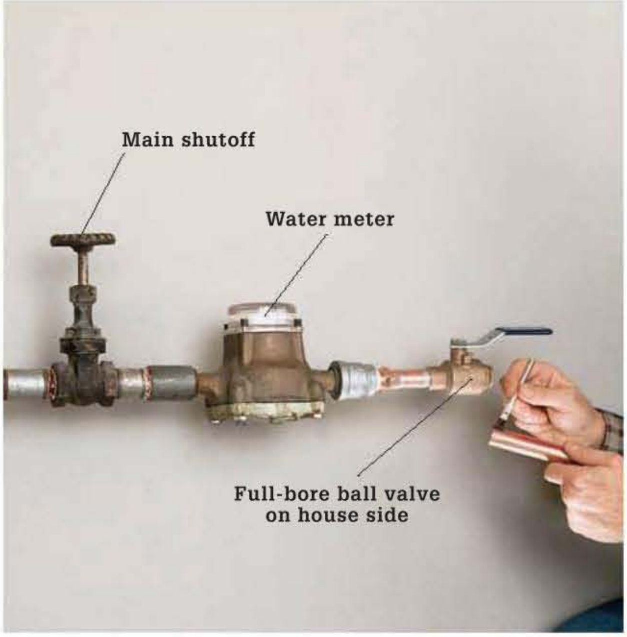

Full-bore gate valves or ball valves are required in the following locations: on both the street side and house side of the water meter; on the inlet pipes for water heaters and heating system boilers. Individual fixtures should have accessible shutoff valves, but these need not be full-bore valves. All sillcocks must have individual control valves located inside the house.

Drain cleanouts make your DWV system easier to service. In most areas, the plumbing code requires that you place cleanouts at the end of every horizontal drain run. Where horizontal runs are not accessible, removable drain traps will suffice as cleanouts.

Pipe Support Intervals

Type of pipe	Vertical-run support interval	Horizontal-run support interval
Copper	10 ft.	6 ft.
PEX	5 ft.	3 ft.
CPVC	10 ft.	3 ft.
PVC	10 ft.	4 ft.
Steel	12 ft.	10 ft.
Iron	15 ft.	5 ft.

Minimum intervals for supporting pipes are determined by the type of pipe and its orientation in the system. Remember that the measurements shown above are minimum requirements; local code may require supports at closer intervals.

Fixture Units & Minimum Trap Size

Fixture	Fixture units	Min. trap size
Shower	2	2"
Vanity sink	1	1¼"
Bathtub	2	1½"
Dishwasher	2	1½"
Kitchen sink	2	1½"
Kitchen sink*	3	1½"
Clothes washer	2	1½"
Utility sink	2	1½"
Floor drain	1	2"

*Kitchen sink with attached food disposer

Minimum trap size for fixtures is determined by the drain fixture unit rating, a unit of measure assigned by the plumbing code. *Note: Kitchen sinks rate 3 units if they include an attached food disposer, 2 units otherwise.*

Sizes for Horizontal & Vertical Drain Pipes

Pipe size	Maximum fixture units for horizontal branch drain	Maximum fixture units for vertical drain stacks
1¼"	1	2
1½"	3	4
2"	6	10
2½"	12	20
3"	20	30
4"	160	240

Drain pipe sizes are determined by the load on the pipes, as measured by the total fixture units. Horizontal drain pipes less than 3" in diameter should slope ¼" per foot toward the main drain. Pipes 3" or more in diameter should slope ⅛" per foot. *Note: Horizontal or vertical drain pipes for a toilet must be 3" or larger.*

Vent Pipe Sizes, Critical Distances

Size of fixture drain	Minimum vent pipe size	Maximum critical distance
1¼"	1¼"	2½ ft.
1½"	1¼"	3½ ft.
2"	1½"	5 ft.
3"	2"	6 ft.
4"	3"	10 ft.

Vent pipes are usually one pipe size smaller than the drain pipes they serve. Code requires that the distance between the drain trap and the vent pipe fall within a maximum "critical distance," a measurement that is determined by the size of the fixture drain. Use this chart to determine both the minimum size for the vent pipe and the maximum critical distance.

Vent Pipe Orientation to Drain Pipe

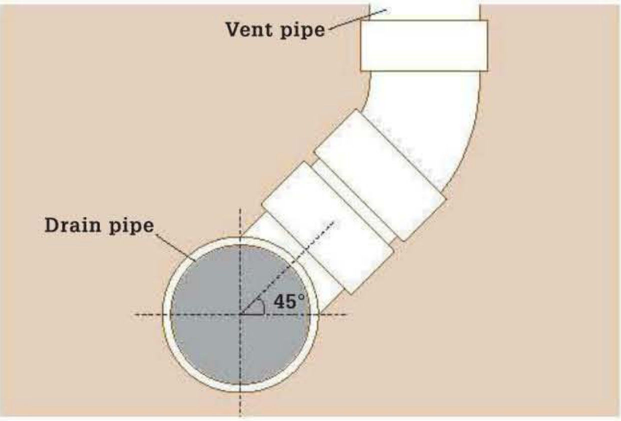

Vent pipes must extend in an upward direction from drains, no less than 45° from horizontal. This ensures that waste water cannot flow into the vent pipe and block it. At the opposite end, a new vent pipe should connect to an existing vent pipe or main waste-vent stack at a point at least 6" above the highest fixture draining into the system.

Wet Venting

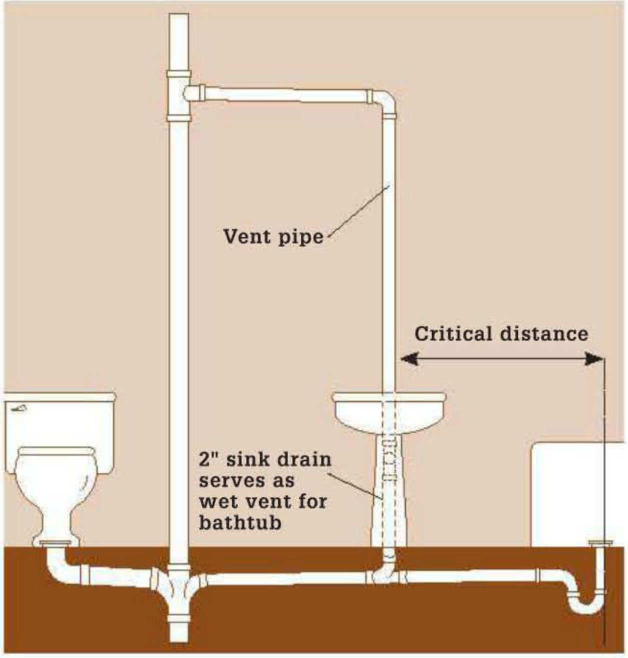

Wet vents are pipes that serve as a vent for one fixture and a drain for another. The sizing of a wet vent is based on the total fixture units it supports (opposite page): a 3" wet vent can serve up to 12 fixture units; a 2" wet vent is rated for 4 fixture units; a 1½" wet vent, for only 1 fixture unit. *Note: The distance between the wet-vented fixture and the wet vent itself must be no more than the maximum critical distance (chart above).*

Auxiliary Venting

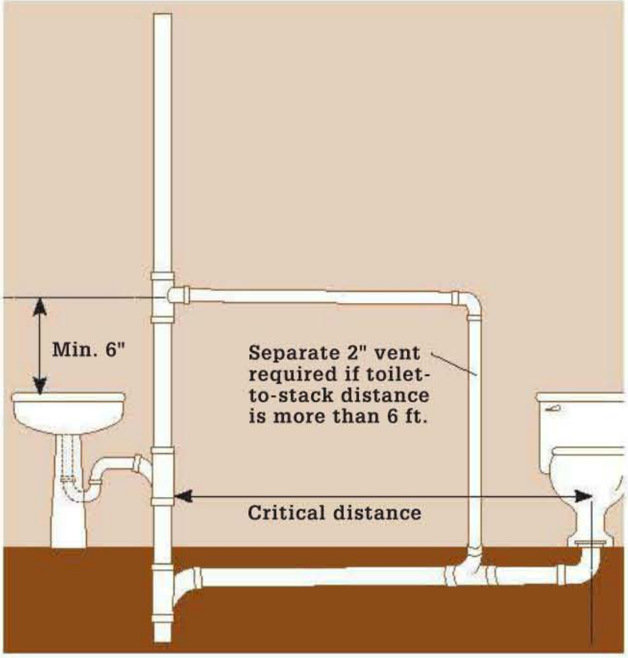
Min. 6"
Separate 2" vent required if toilet-to-stack distance is more than 6 ft.
Critical distance

Fixtures must have an auxiliary vent if the distance to the main waste-vent stack exceeds the critical distance (illustration above). A toilet, for example, should have a separate vent pipe if it is located more than 6 ft. from the main waste-vent stack. This secondary vent pipe should connect to the stack or an existing vent pipe at a point at least 6" above the highest fixture on the system.

Understanding Wiring

A household electrical system can be compared with a home's plumbing system. Electrical current flows in wires in much the same way that water flows inside pipes. Both electricity and water enter the home, are distributed throughout the house, do their "work," and exit.

In plumbing, water first flows through the pressurized water supply system. In electricity, current first flows along hot wires. Current flowing along hot wires also is pressurized. The pressure of electrical current is called *voltage.*

Large supply pipes can carry a greater volume of water than small pipes. Likewise, large electrical wires carry more current than small wires. This current-carrying capacity of wires is called *amperage.*

Water is made available for use through the faucets, spigots, and showerheads in a home. Electricity is made available through receptacles, switches, and fixtures.

Water finally leaves the home through a drain system, which is not pressurized. Similarly, electrical current flows back through neutral wires. The current in neutral wires is not pressurized and is said to be at zero voltage.

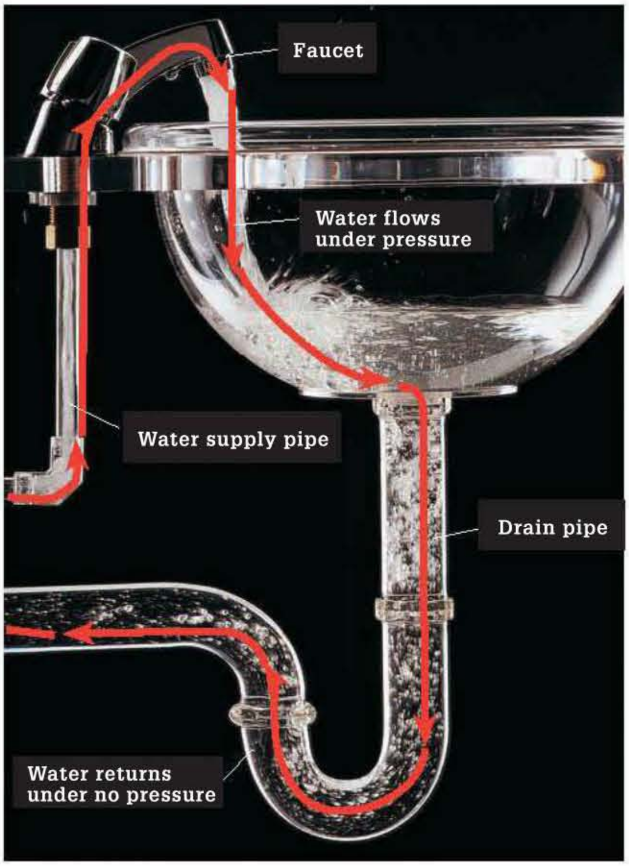
Faucet
Water flows under pressure
Water supply pipe
Drain pipe
Water returns under no pressure

Water and electricity both flow. The main difference is that you can see water (and touching water isn't likely to kill you). Like electricity, water enters a fixture under high pressure and exits under low pressure.

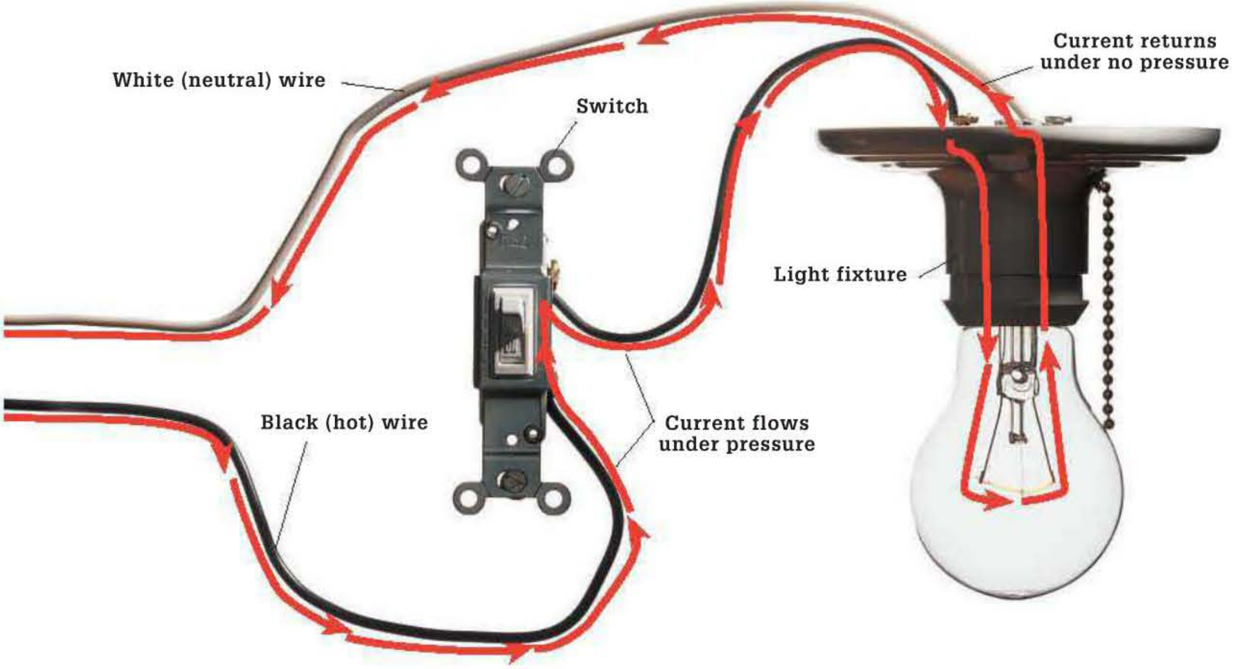
White (neutral) wire
Switch
Current returns under no pressure
Light fixture
Black (hot) wire
Current flows under pressure

The Delivery System

Electrical power that enters the home is produced by large power plants. Power plants are located in all parts of the country and generate electricity with turbines that are turned by water, wind, or steam. From these plants electricity enters large "step-up" transformers that increase voltage to half a million volts or more.

Electricity flows easily at these large voltages and travels through high-voltage transmission lines to communities that can be hundreds of miles from the power plants. "Step-down" transformers located at substations then reduce the voltage for distribution along street lines. On utility power poles, smaller transformers further reduce the voltage to ordinary 120-volt current for household use.

Lines carrying current to the house either run underground or strung overhead and attached to a post called a service mast. Most homes built after 1950 have three wires running to the service head: two power lines, each carrying 120 volts of current, and a grounded neutral wire. Power from the two 120-volt lines may be combined at the service panel to supply current to large 240-volt appliances like clothes dryers or electric water heaters.

Incoming power passes through an electric meter that measures power consumption. Power then enters the service panel, where it is distributed to circuits that run throughout the house. The service panel also contains fuses or circuit breakers that shut off power to the individual circuits in the event of a short circuit or an overload. Certain high-wattage appliances, like microwave ovens, are usually plugged into their own individual circuits to prevent overloads.

Voltage ratings determined by power companies and manufacturers have changed over the years. Current rated at 110 volts changed to 115 volts, then 120 volts. Current rated at 220 volts changed to 230 volts, then 240 volts. Similarly, ratings for receptacles, tools, light fixtures, and appliances have changed from 115 volts to 125 volts. These changes do not affect the performance of new devices connected to older wiring. For making electrical calculations, use a rating of 120 volts or 240 volts for your circuits.

Power plants supply electricity to thousands of homes and businesses. Step-up transformers increase the voltage produced at the plant, making the power flow more easily along high-voltage transmission lines.

Substations are located near the communities they serve. A typical substation takes current from high-voltage transmission lines and reduces it for distribution along street lines.

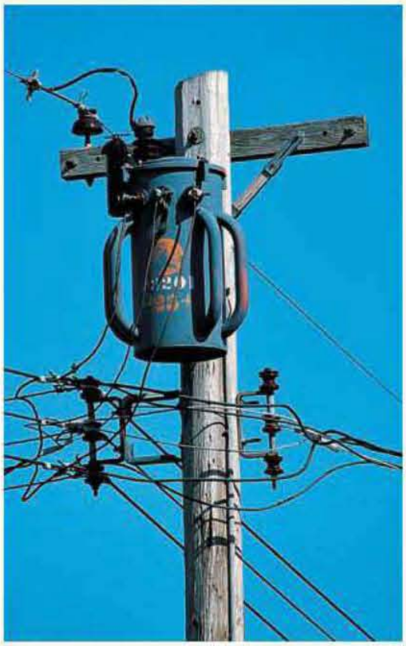
Utility pole transformers reduce the high-voltage current that flows through power lines along neighborhood streets. A utility pole transformer reduces voltage from 10,000 volts to the normal 120-volt current used in households.

Parts of the Electrical System

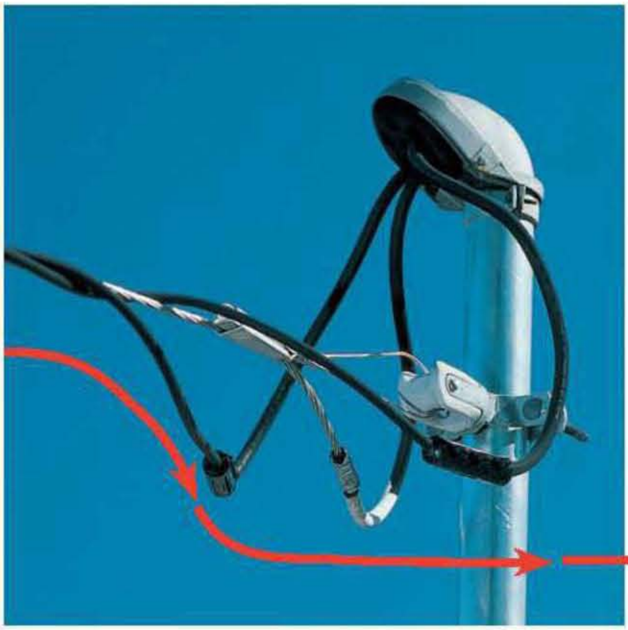

The service mast is the metal pole and weatherhead that create the entry point for electricity into your home. The mast is supplied with two wires carrying 120 volts and a third grounded neutral wire. The current runs to the service mast from the nearest transformer.

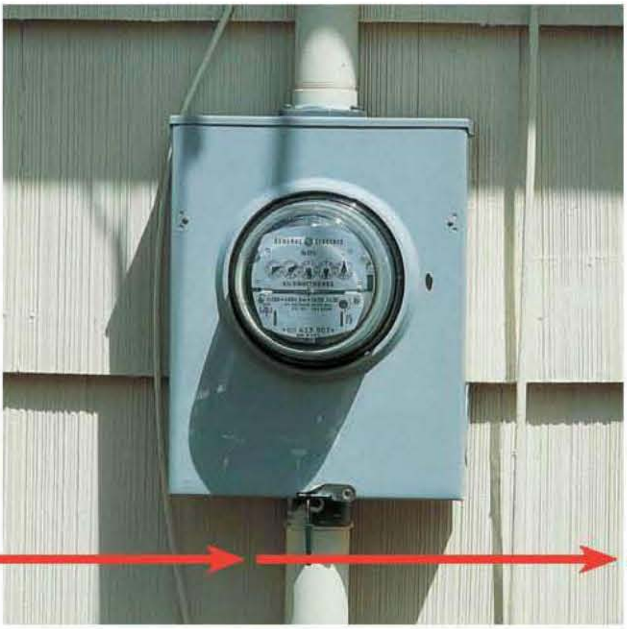

The electric meter measures the amount of electrical power consumed. It is usually attached to the side of the house, and connects to the service mast. A thin metal disc inside the meter rotates when power is used. The electric meter belongs to your local power utility company. If you suspect the meter is not functioning properly, contact the power company.

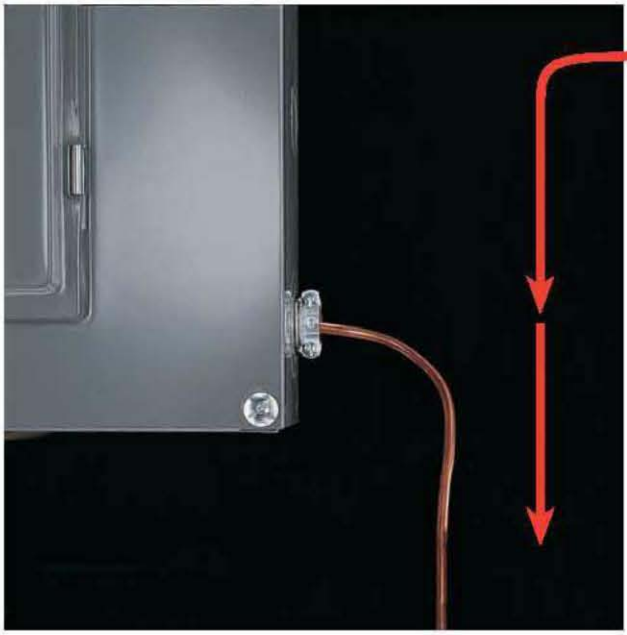

Grounding wire connects the electrical system to the earth through grounding rods or, in older systems, through a cold water pipe. In the event of an overload or short circuit, the grounding wire allows excess electrical power to find its way harmlessly to the earth.

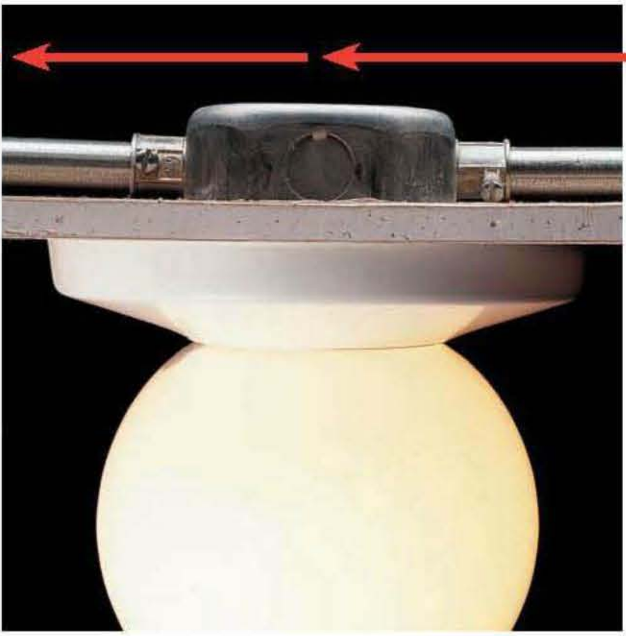

Light fixtures attach directly to a household electrical system. They are usually controlled with wall switches. The two most common types of light fixtures are incandescent and fluorescent.

The main service panel, in the form of a fuse box or breaker box, distributes power to individual circuits. Fuses or circuit breakers protect each circuit from short circuits and overloads. Fuses and circuit breakers also are used to shut off power to individual circuits while repairs are made.

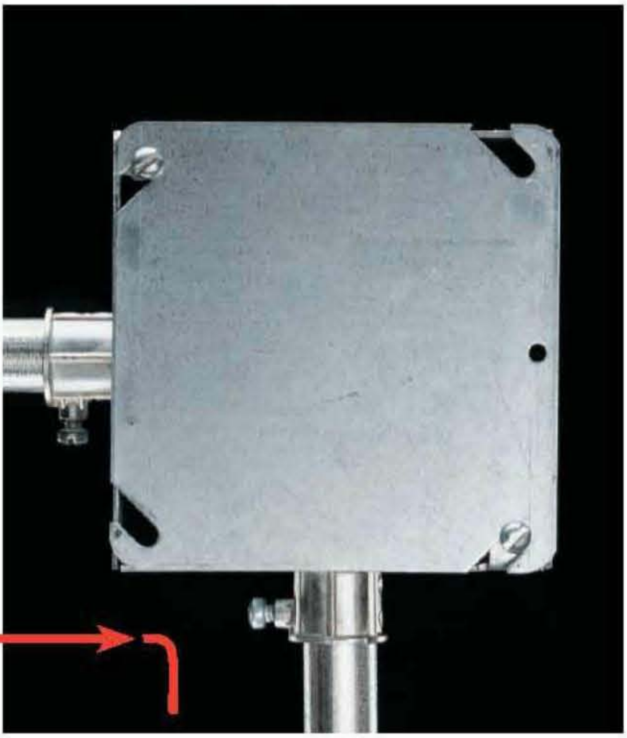

Electrical boxes enclose wire connections. According to the National Electrical Code, all wire splices or connections must be contained entirely in a covered plastic or metal electrical box.

Switches control electrical current passing through hot circuit wires. Switches can be wired to control light fixtures, ceiling fans, appliances, and receptacles.

Receptacles, sometimes called outlets, provide plug-in access to electrical power. A 120-volt, 15-amp receptacle with a grounding hole is the most typical receptacle in wiring systems installed after 1965. Most receptacles have two plug-in locations and are called *duplex receptacles.*

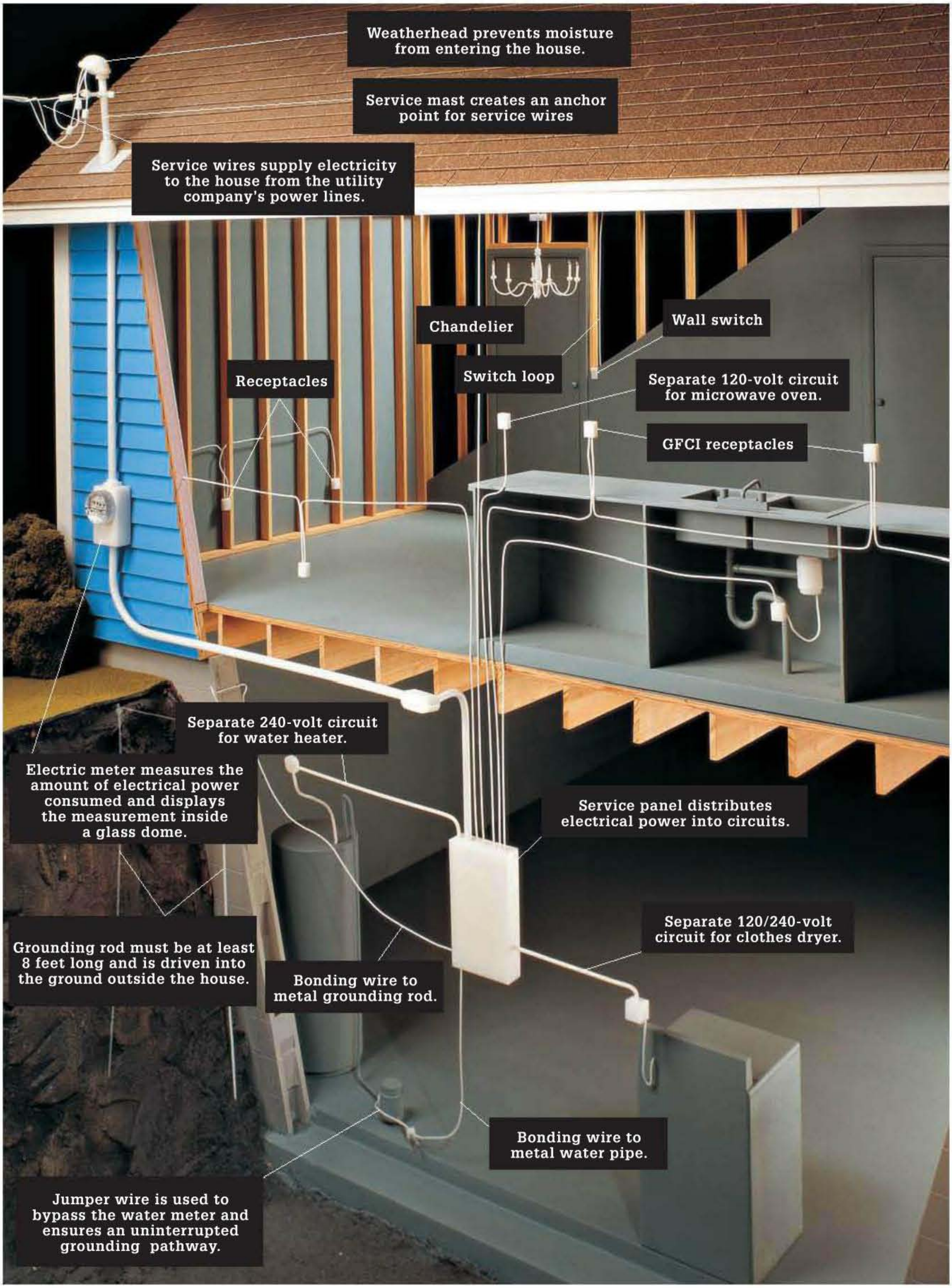
Weatherhead prevents moisture from entering the house.
Service mast creates an anchor point for service wires
Service wires supply electricity to the house from the utility company's power lines.
Chandelier
Wall switch
Receptacles
Switch loop
Separate 120-volt circuit for microwave oven.
GFCI receptacles
Separate 240-volt circuit for water heater.
Electric meter measures the amount of electrical power consumed and displays the measurement inside a glass dome.
Service panel distributes electrical power into circuits.
Separate 120/240-volt circuit for clothes dryer.
Grounding rod must be at least 8 feet long and is driven into the ground outside the house.
Bonding wire to metal grounding rod.
Bonding wire to metal water pipe.
Jumper wire is used to bypass the water meter and ensures an uninterrupted grounding pathway.

Glossary of Electrical Terms

Ampere (or amp): Refers to the rate at which electrical power flows to a light, tool, or appliance.

Armored cable: Two or more wires that are grouped together and protected by a flexible metal covering.

Box: A device used to contain wiring connections.

BX: See armored cable (Bx is the older term).

Cable: Two or more wires that are grouped together and protected by a covering or sheath.

Circuit: A continuous loop of electrical current flowing along wires or cables.

Circuit breaker: A safety device that interrupts an electrical circuit in the event of an overload or short circuit.

Conductor: Any material that allows electrical current to flow through it. Copper wire is an especially good conductor.

Conduit: A metal or plastic pipe used to protect wires.

Continuity: An uninterrupted electrical pathway through a circuit or electrical fixture.

Current: The movement of electrons along a conductor.

Duplex receptacle: A receptacle that provides connections for two plugs.

Feed wire: A conductor that carries 120-volt current uninterrupted from the service panel.

Fuse: A safety device, usually found in older homes, that interrupts electrical circuits during an overload or short circuit.

Greenfield: Materials used in flexible metal conduit. See armored cable.

Grounded wire: See neutral wire.

Grounding wire: A wire used in an electrical circuit to conduct current to the earth in the event of a short circuit. The grounding wire often is a bare copper wire.

Hot wire: Any wire that carries voltage. In an electrical circuit, the hot wire usually is covered with black or red insulation.

Insulator: Any material, such as plastic or rubber, that resists the flow of electrical current. Insulating materials protect wires and cables.

Junction box: See box.

Meter: A device used to measure the amount of electrical power being used.

Neutral wire: A wire that returns current at zero voltage to the source of electrical power. Usually covered with white or light gray insulation. Also called the grounded wire.

Outlet: See receptacle.

Overload: A demand for more current than the circuit wires or electrical device was designed to carry. Usually causes a fuse to blow or a circuit breaker to trip.

Pigtail: A short wire used to connect two or more circuit wires to a single screw terminal.

Polarized receptacle: A receptacle designed to keep hot current flowing along black or red wires, and neutral current flowing along white or gray wires.

Power: The result of hot current flowing for a period of time. Use of power makes heat, motion, or light.

Receptacle: A device that provides plug-in access to electrical power.

Romex: A brand name of plastic-sheathed electrical cable that is commonly used for indoor wiring. Commonly known as NM cable.

Screw terminal: A place where a wire connects to a receptacle, switch, or fixture.

Service panel: A metal box usually near the site where electrical power enters the house. In the service panel, electrical current is split into individual circuits. The service panel has circuit breakers or fuses to protect each circuit.

Short circuit: An accidental and improper contact between two current-carrying wires, or between a current-carrying wire and a grounding conductor.

Switch: A device that controls electrical current passing through hot circuit wires. Used to turn lights and appliances on and off.

UL: An abbreviation for Underwriters Laboratories, an organization that tests electrical devices and manufactured products for safety.

Voltage (or volts): A measurement of electricity in terms of pressure.

Wattage (or watt): A measurement of electrical power in terms of total energy consumed. Watts can be calculated by multiplying the voltage times the amps.

Wire connector: A device used to connect two or more wires together. Also called a *wire nut.*

Wiring Safety

Safety should be the primary concern of anyone working with electricity. Although most household electrical repairs are simple and straightforward, always use caution and good judgment when working with electrical wiring or devices. Common sense can prevent accidents.

The basic rule of electrical safety is: Always turn off power to the area or device you are working on. At the main service panel, remove the fuse or shut off the circuit breaker that controls the circuit you are servicing. Then check to make sure the power is off by testing for power with a current tester. Restore power only when the repair or replacement project is complete.

Follow the safety tips shown on these pages. Never attempt an electrical project beyond your skill or confidence level. Never attempt to repair or replace your main service panel or service entrance head. These are jobs for a qualified electrician and require that the power company shuts off power to your house.

Shut power OFF at the main service panel or the main fuse box before beginning any work.

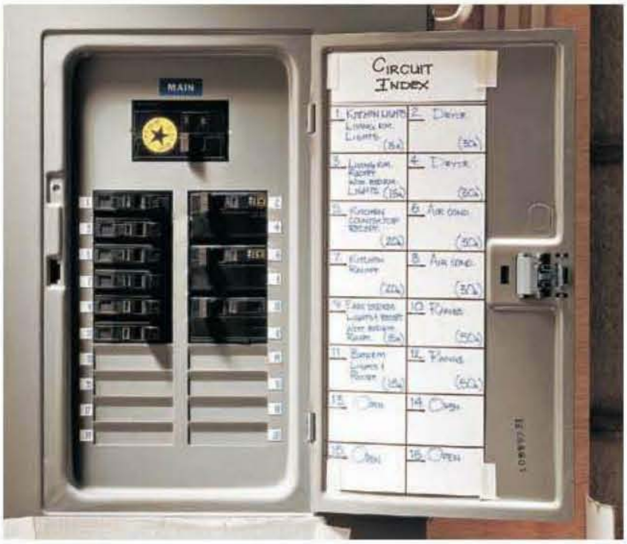

Create a circuit index and affix it to the inside of the door to your main service panel. Update it as needed.

Confirm power is OFF by testing at the outlet, switch, or fixture with a current tester.

Electracraft®
Quik-connect
double switch
QUIET Single Pole
UL

Use only UL-approved electrical parts or devices. These devices have been tested for safety by Underwriters Laboratories.

Wear rubber-soled shoes while working on electrical projects. On damp floors, stand on a rubber mat or dry wooden boards.

Use fiberglass or wood ladders when making routine household repairs near the service mast.

Extension cords are for temporary use only. Cords must be rated for the intended usage.

Electracraft®
plug
fuses
UL
TIME DELAY FUSE
20

Breakers and fuses must be compatible with the panel manufacturer and match the circuit capacity.

Never alter the prongs of a plug to fit a receptacle. If possible, install a new grounded receptacle.

Do not penetrate walls or ceilings without first shutting off electrical power to the circuits that may be hidden.

Working with Wire & Cable

Wires are made of copper, aluminum, or aluminum covered with a thin layer of copper. Solid copper wires are the best conductors of electricity and are the most widely used. Aluminum and copper-covered aluminum wires require special installation techniques and are rarely used in new installation, due to safety concerns.

A group of two or more wires enclosed in a metal, rubber, or plastic sheath is called a *cable* (photo, opposite page). The sheath protects the wires from damage. Metal conduit also protects wires, but it is not considered a cable.

Individual wires are covered with rubber or plastic vinyl insulation. An exception is a bare copper grounding wire, which does not need an insulation cover. The insulation is color coded (chart, below left) to identify the wire as a hot wire, a neutral wire, or a grounding wire.

In most wiring systems installed after 1965, the wires and cables are insulated with plastic vinyl. This type of insulation is very durable and can last as long as the house itself.

Before 1965, wires and cables were insulated with rubber. Rubber insulation has a life expectancy of about 25 years. Old insulation that is cracked or damaged can be reinforced temporarily by wrapping the wire with plastic electrical tape. However, old wiring with cracked or damaged insulation should be inspected by a qualified electrician to make sure it is safe.

Wires must be large enough for the amperage rating of the circuit (chart, below right). A wire that is too small can become dangerously hot. Wire sizes are categorized according to the American Wire Gauge (AWG) system. To check the size of a wire, use the wire stripper openings of a combination tool (page 30) as a guide.

Wire Color Chart ▸

Wire color		Function
	White	Neutral wire carrying current at zero voltage.
	Black	Hot wire carrying current at full voltage.
	Red	Hot wire carrying current at full voltage.
	White, black markings	Hot wire carrying current at full voltage.
	Green	Serves as a grounding pathway.
	Bare copper	Serves as a grounding pathway.

Individual wires are color-coded to identify their function. In some circuit installations, the white wire serves as a hot wire that carries voltage. If so, this white wire may be labeled with black tape or paint to identify it as a hot wire.

Wire Size Chart ▸

Wire gauge		Wire capacity & use
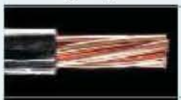	#6	60 amps, 240 volts; central air conditioner, electric furnace.
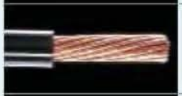	#8	40 amps, 240 volts; electric range, central air conditioner.
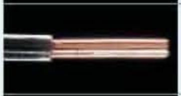	#10	30 amps, 240 volts; window air conditioner, clothes dryer.
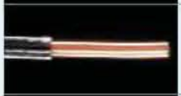	#12	20 amps, 120 volts; light fixtures, receptacles, microwave oven.
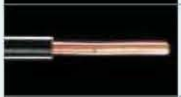	#14	15 amps, 120 volts; light fixtures, receptacles.
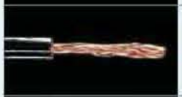	#16	Light-duty extension cords.
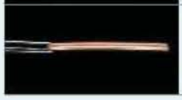	#18 to 22	Thermostats, doorbells, security systems.

Wire sizes (shown actual size) are categorized by the American Wire Gauge system. The larger the wire size, the smaller the AWG number.

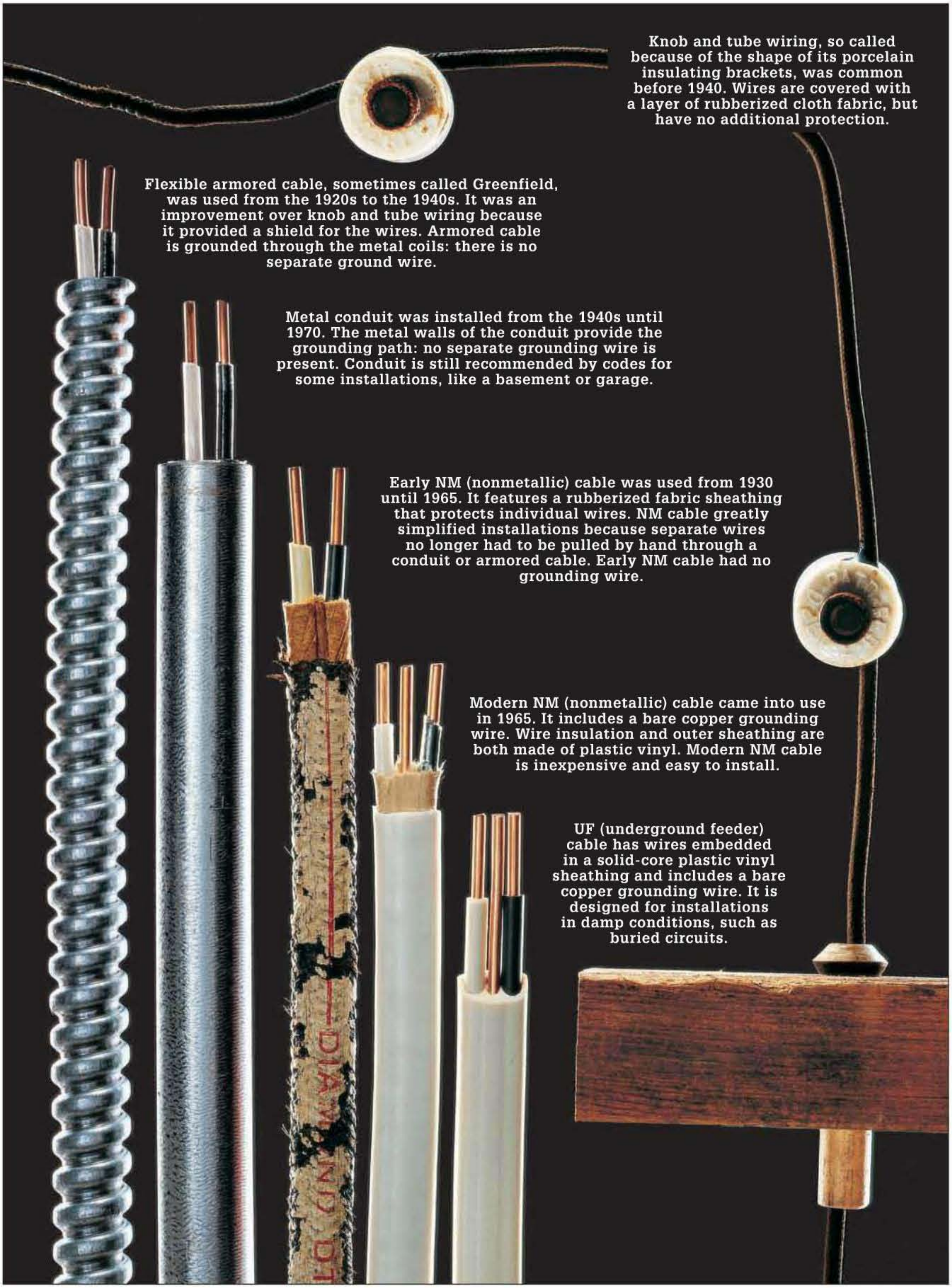

Knob and tube wiring, so called because of the shape of its porcelain insulating brackets, was common before 1940. Wires are covered with a layer of rubberized cloth fabric, but have no additional protection.

Flexible armored cable, sometimes called Greenfield, was used from the 1920s to the 1940s. It was an improvement over knob and tube wiring because it provided a shield for the wires. Armored cable is grounded through the metal coils: there is no separate ground wire.

Metal conduit was installed from the 1940s until 1970. The metal walls of the conduit provide the grounding path: no separate grounding wire is present. Conduit is still recommended by codes for some installations, like a basement or garage.

Early NM (nonmetallic) cable was used from 1930 until 1965. It features a rubberized fabric sheathing that protects individual wires. NM cable greatly simplified installations because separate wires no longer had to be pulled by hand through a conduit or armored cable. Early NM cable had no grounding wire.

Modern NM (nonmetallic) cable came into use in 1965. It includes a bare copper grounding wire. Wire insulation and outer sheathing are both made of plastic vinyl. Modern NM cable is inexpensive and easy to install.

UF (underground feeder) cable has wires embedded in a solid-core plastic vinyl sheathing and includes a bare copper grounding wire. It is designed for installations in damp conditions, such as buried circuits.

NM (nonmetallic) sheathed cable should be used for most indoor wiring projects in dry locations. NM cable is available in a wide range of wire sizes, and in either "2-wire with ground" or "3-wire with ground" types. NM cable is sold in boxed rolls that contain from 25 to 250 ft. of cable.

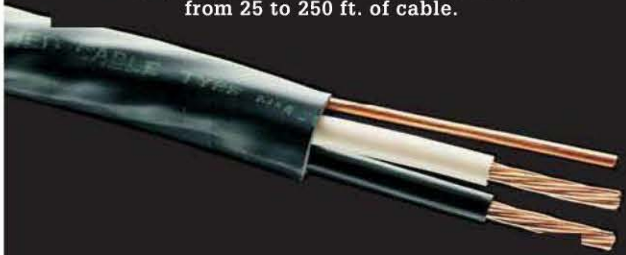

Large-appliance cable, also called SER cable, is used for kitchen ranges and other 50-amp or 60-amp appliances that require 8-gauge or larger wire. It is similar to NM cable, but each individual conducting wire is made from fine-stranded copper wires. Large-appliance cable is available in both 2-wire and 3-wire types.

UF (underground feeder) cable is used for wiring in damp locations, such as in an outdoor circuit. It has a white or gray solid-core vinyl sheathing that protects the wires inside. It also can be used indoors wherever NM cable is allowed.

Coaxial cable is used to connect cable television jacks. Coaxial cable is available in lengths up to 25 ft. with preattached F-connectors (A). Or you can buy bulk cable (B) in any length.

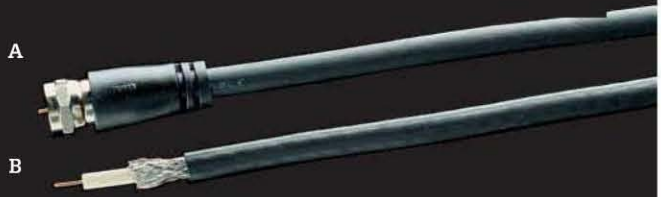

THHN/THWN wire can be used in all conduit applications. Each wire, purchased individually, is covered with a color-coded thermoplastic insulating jacket. Make sure the wire you buy has the THHN/THWN rating. Other wire types are less resistant to heat and moisture than THHN/THWN wire.

Telephone cable is used to connect telephone outlets. Your phone company may recommend four-wire cable (shown below) or eight-wire cable, sometimes called four-pair. Eight-wire cable has extra wires that are left unattached. These extra wires allow for future expansion of the system.

Tips for Working With Wire ▸

Wire gauge		Ampacity	Maximum wattage load
	14-gauge	15 amps	1440 watts (120 volts)
	12-gauge	20 amps	1920 watts (120 volts) 3840 watts (240 volts)
	10-gauge	30 amps	2880 watts (120 volts) 5760 watts (240 volts)
	8-gauge	40 amps	7680 watts (240 volts)
	6-gauge	50 amps	9600 watts (240 volts)

Wire "ampacity" is a measurement of how much current a wire can carry safely. Ampacity varies according to the size of the wires, as shown at left. When installing a new circuit, choose wire with an ampacity rating matching the circuit size. For dedicated appliance circuits, check the wattage rating of the appliance and make sure it does not exceed the maximum wattage load of the circuit.

Reading NM (Nonmetallic) Cable

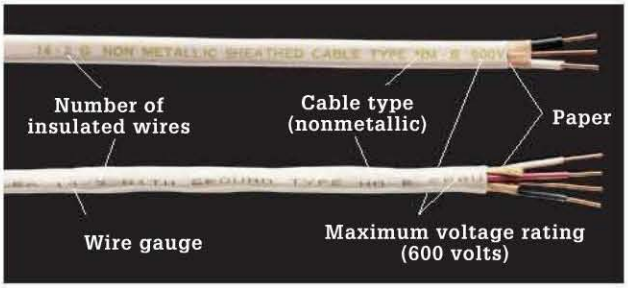
Number of insulated wires
Cable type (nonmetallic)
Paper
Wire gauge
Maximum voltage rating (600 volts)

NM (nonmetallic) cable is labeled with the number of insulated wires it contains. The bare grounding wire is not counted. For example, a cable marked 14/2 G (or 14/2 WITH GROUND) contains two insulated 14-gauge wires, plus a bare copper grounding wire. Cable marked 14/3 WITH GROUND has three 14-gauge wires plus a grounding wire. NM cable also is stamped with a maximum voltage rating, as determined by Underwriters Laboratories (UL).

Reading Unsheathed, Individual Wire

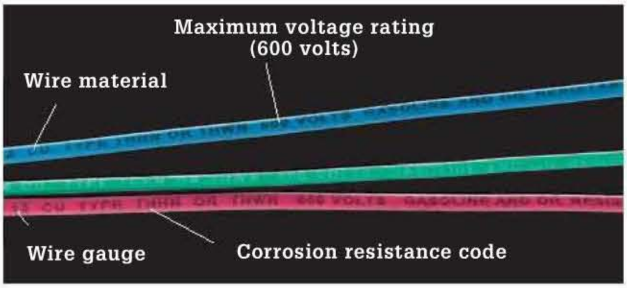
Maximum voltage rating (600 volts)
Wire material
Wire gauge
Corrosion resistance code

Unsheathed, individual wires are used for conduit and raceway installations. Wire insulation is coded with letters to indicate resistance to moisture, heat, and gas or oil. Code requires certain letter combinations for certain applications. T indicates thermoplastic insulation. H stands for heat resistance and two Hs indicate high resistance (up to 194° F). W denotes wire suitable for wet locations. Wire coded with an N is impervious to damage from oil or gas.

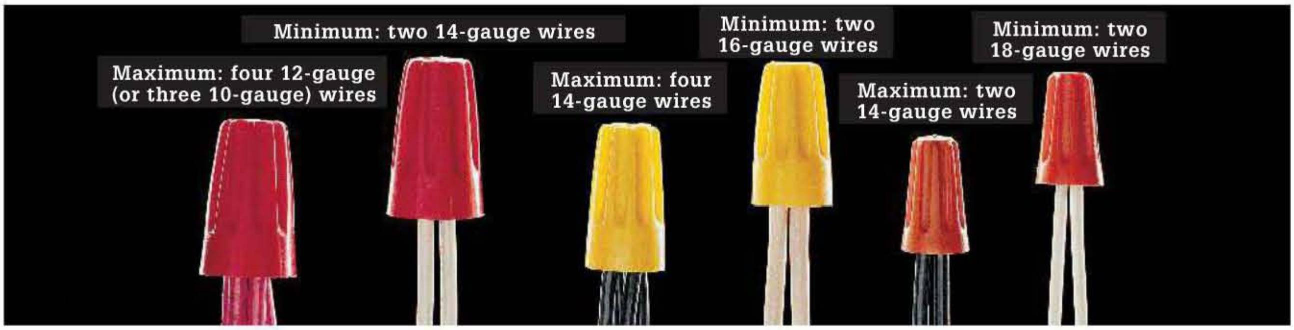
Maximum: four 12-gauge (or three 10-gauge) wires
Minimum: two 14-gauge wires
Maximum: four 14-gauge wires
Minimum: two 16-gauge wires
Maximum: two 14-gauge wires
Minimum: two 18-gauge wires

Use wire connectors rated for the wires you are connecting. Wire connectors are color-coded by size, but the coding scheme varies according to manufacturer. The wire connectors shown above come from one major manufacturer. To ensure safe connections, each connector is rated for both minimum and maximum wire capacity. These connectors can be used to connect both conducting wires and grounding wires. Green wire connectors are used only for grounding wires.

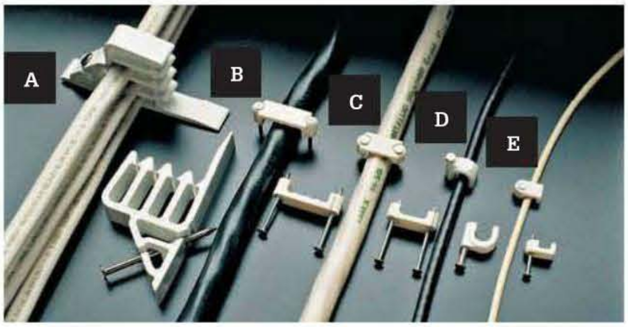

Use plastic cable staples to fasten cables. Choose staples sized to match the cables. Stack-It® staples (A) hold up to four 2-wire cables; ¾" staples (B) for 12/2, 12/3, and all 10-gauge cables; ½" staples (C) for 14/2, 14/3, or 12/2 cables; coaxial staples (D) for anchoring television cables; bell wire staples (E) for attaching telephone cables.

WAGO
100
773-164
4 x 12 AWG Max. Cu
50
773-166
6 x 12 AWG Max. Cu

Push-in connectors are a relatively new product for joining wires. Instead of twisting the bare wire ends together, you strip off about ¾" of insulation and insert them into a hole in the connector. The connectors come with two to four holes sized for various gauge wires. These connectors are perfect for inexperienced DIYers because they do not pull apart like a poorly twisted connection can.

How to Strip NM Sheathing & Insulation

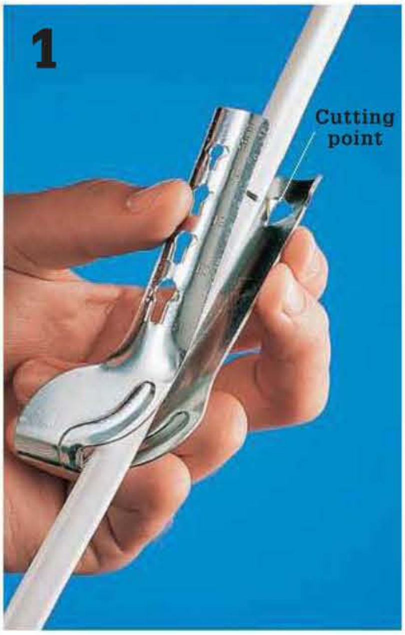

Measure and mark the cable 8 to 10" from end. Slide the cable ripper onto the cable, and squeeze tool firmly to force cutting point through plastic sheathing.

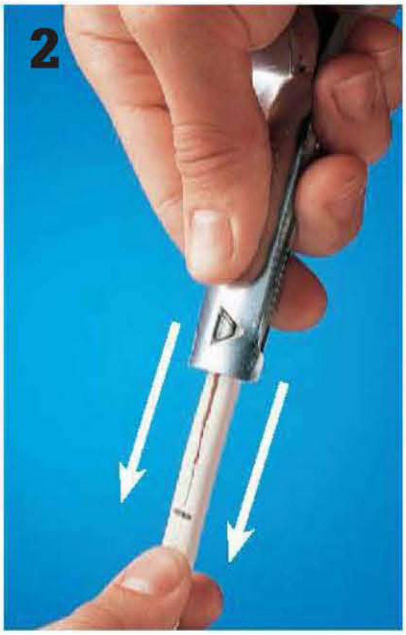

Grip the cable tightly with one hand, and pull the cable ripper toward the end of the cable to cut open the plastic sheathing.

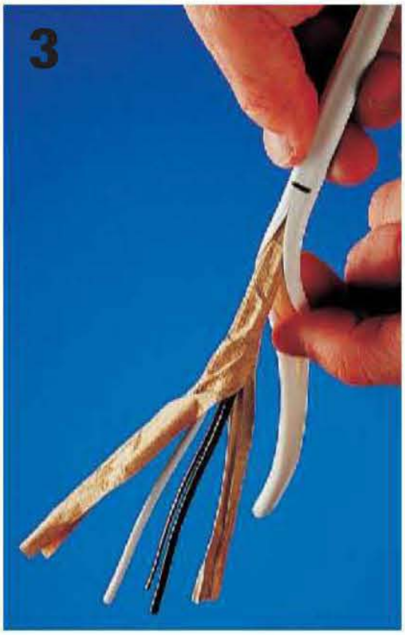

Peel back the plastic sheathing and the paper wrapping from the individual wires.

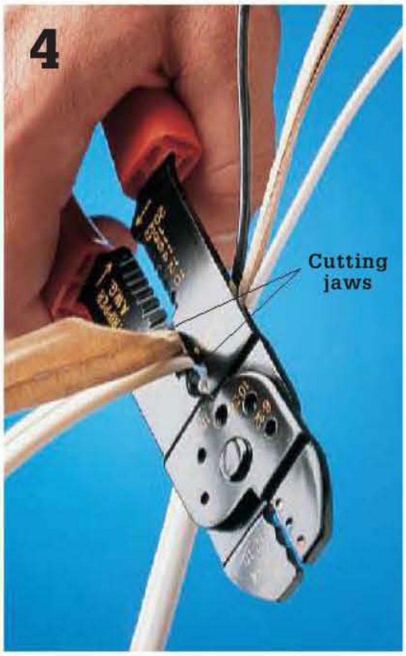

Cut away the excess plastic sheathing and paper wrapping, using the cutting jaws of a combination tool.

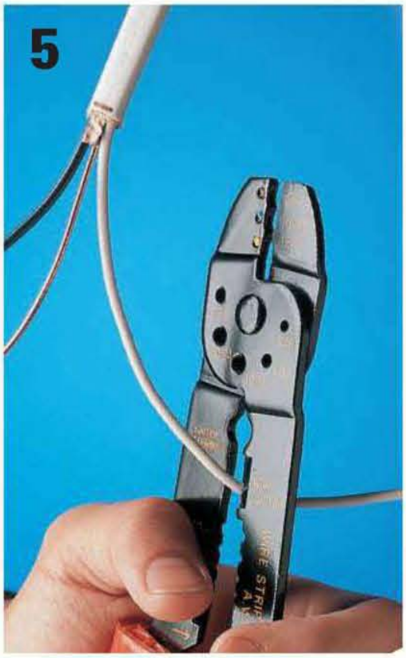

Cut individual wires as needed using the cutting jaws of the combination tool.

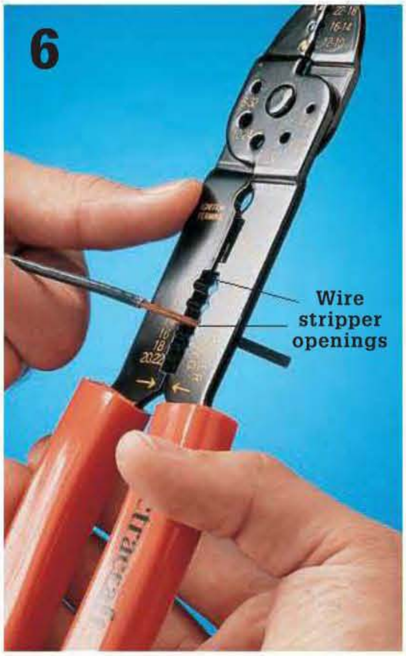

Strip insulation for each wire, using the stripper openings. Choose the opening that matches the gauge of the wire, and take care not to nick or scratch the ends of the wires.

How to Connect Wires to Screw Terminals

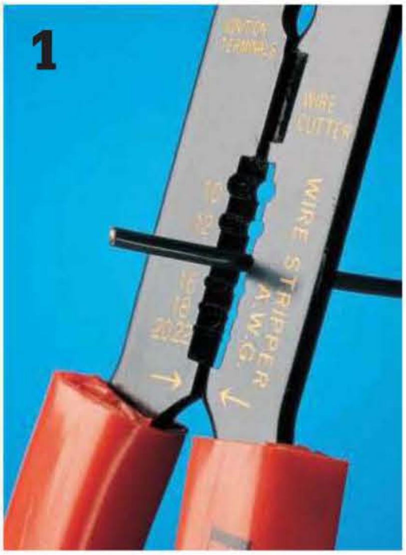
WIRE CUTTER
WIRE STRIPPER
A.W.G.
10
12
14
16
18
20
22

1

Strip about ¾" of insulation from each wire using a combination tool. Choose the stripper opening that matches the gauge of the wire, then clamp the wire in the tool. Pull the wire firmly to remove plastic insulation.

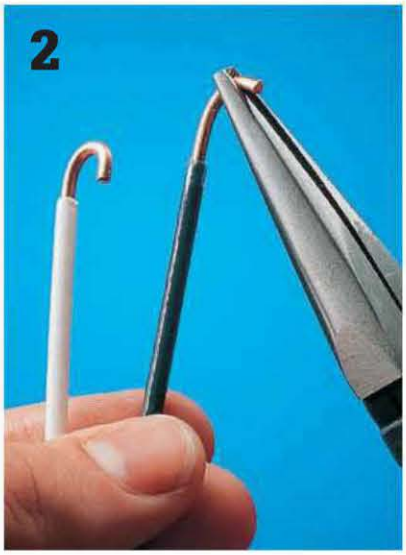

2

Form a C-shaped loop in the end of each wire using a needlenose pliers. The wire should have no scratches or nicks.

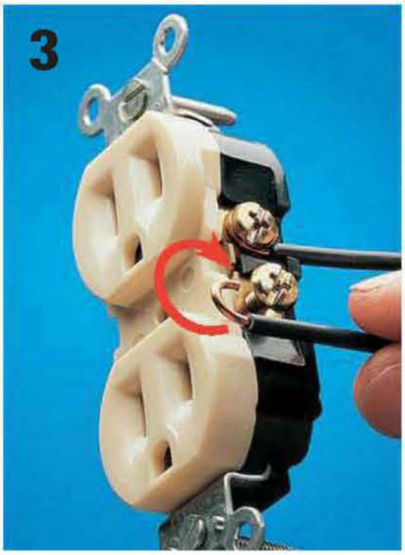

3

Hook each wire around the screw terminal so it forms a clockwise loop. Tighten screw firmly. Insulation should just touch head of screw. Never place the ends of two wires under a single screw terminal. Instead, use a pigtail wire (page 97).

How to Connect Wires with Push-ins

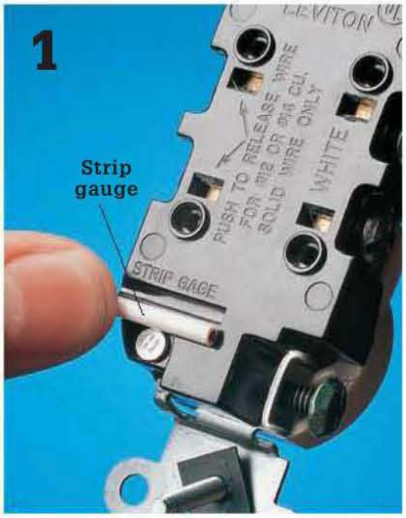
Strip gauge
LEVITON
PUSH TO RELEASE WIRE
FOR #12 OR #14 CU.
SOLID WIRE ONLY
WHITE
STRIP GAGE

1

Mark the amount of insulation to be stripped from each wire using the strip gauge on the back of the switch or receptacle. Strip the wires using a combination tool (step 1, above). Never use push-in fittings with aluminum wiring.

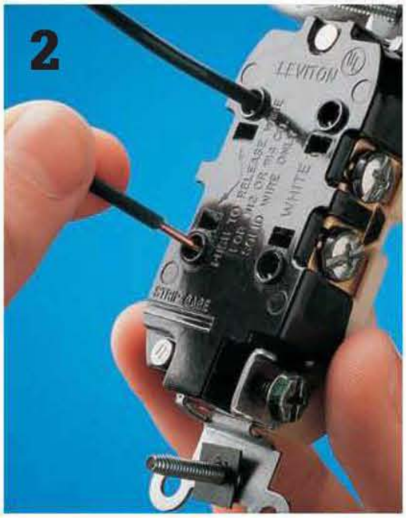
LEVITON
WHITE
STRIP GAGE

2

Insert the bare copper wires firmly into the push-in fittings on the back of the switch or receptacle. When inserted, wires should have no bare copper exposed. *Note: Although push-in fittings are convenient, most experts believe screw terminal connections (above) are more dependable.*

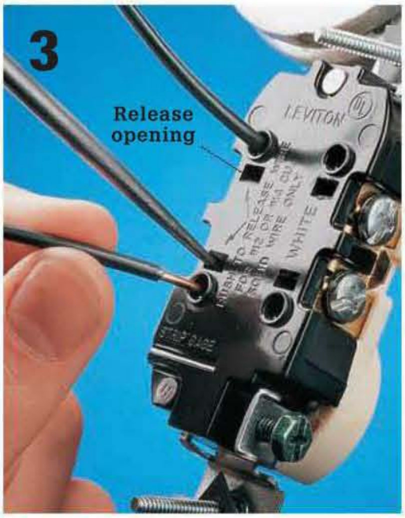
Release opening
LEVITON
WHITE
STRIP GAGE

3

Remove a wire from a push-in fitting by inserting a small nail or screwdriver in the release opening next to the wire. Wire will pull out easily.

How to Join Wires with a Wire Connector

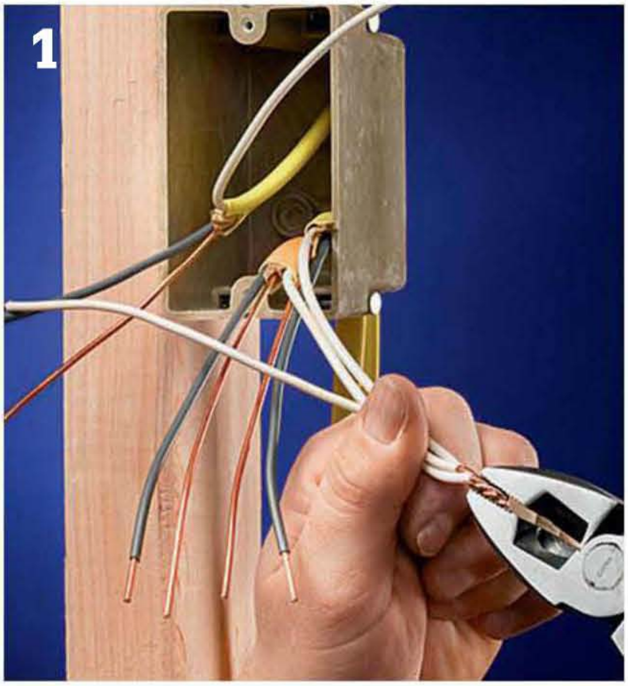

Ensure power is off and test for power. Grasp the wires to be joined in the jaws of a pair of linesman's pliers. The ends of the wires should be flush and they should be parallel and touching. Rotate the pliers clockwise two or three turns to twist the wire ends together.

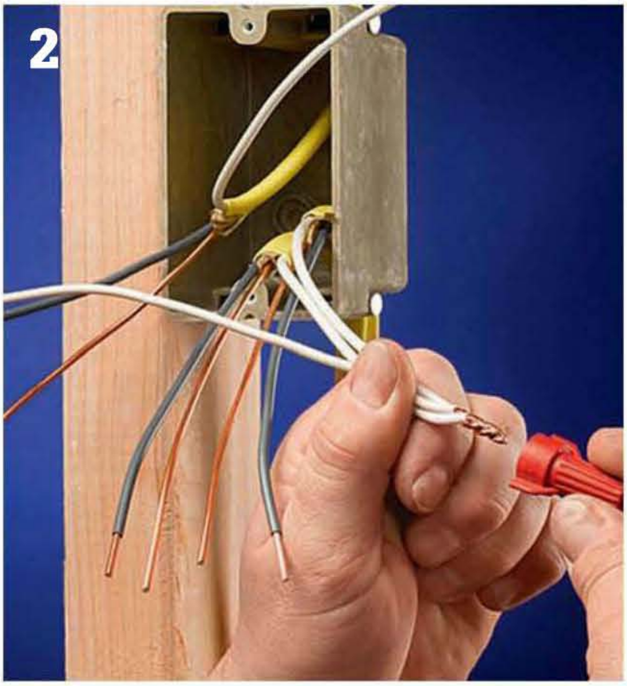

Twist a wire connector over the ends of the wires. Make sure the connector is the right size (see page 29). Hand-twist the connector as far onto the wires as you can. There should be no bare wire exposed beneath the collar of the connector. Do not overtighten the connector.

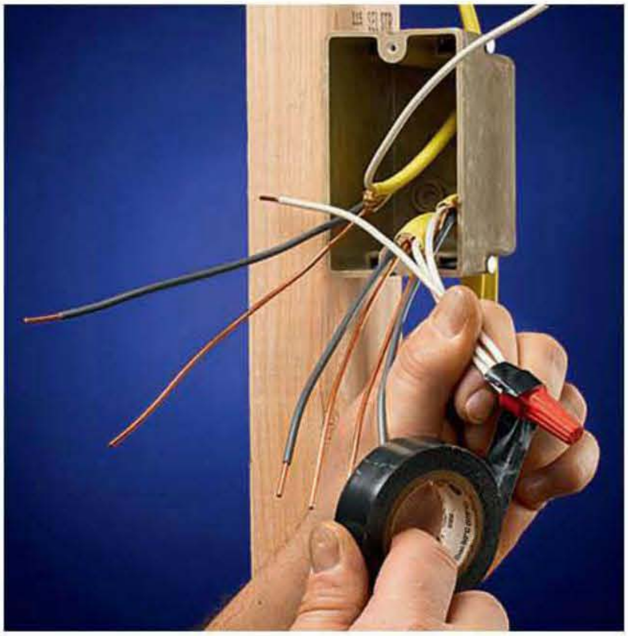

Option: Reinforce the joint by wrapping it with electrician's tape. By code, you cannot bind the wire joint with tape only, but it can be used as insurance. Few professional electricians use tape for purposes other than tagging wires for identification.

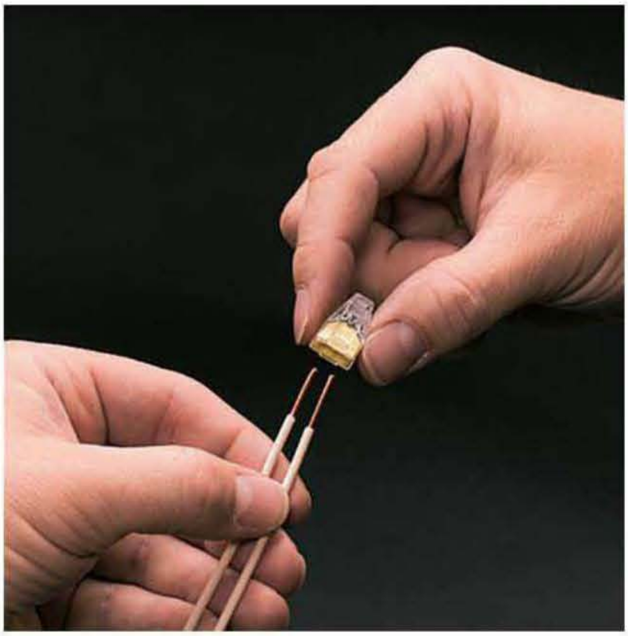

Option: Strip ¾" of insulation off the ends of the wires to be joined, and insert each wire into a push-in connector. Gently tug on each wire to make sure it is secure.

How to Pigtail Wires

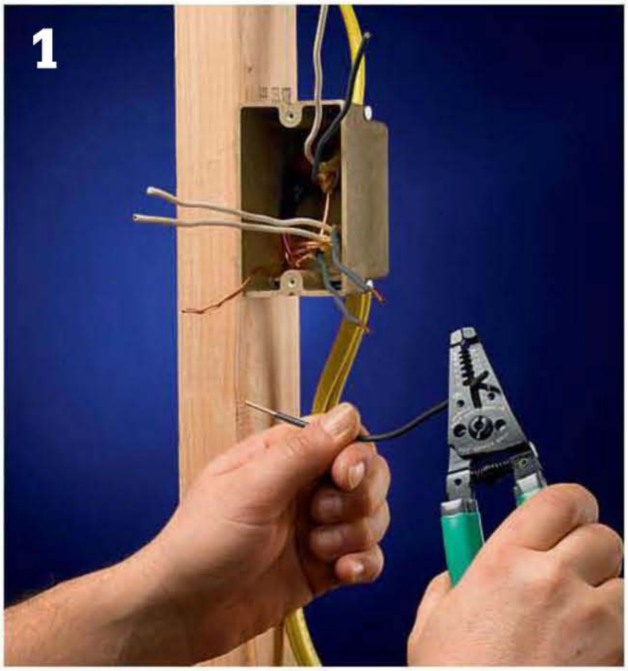

Cut a 6" length from a piece of insulated wire the same gauge and color as the wires it will be joining. Strip ¾" of insulation from each end of the insulated wire. *Note: Pigtailing is done mainly to avoid connecting multiple wires to one terminal, which is a code violation.*

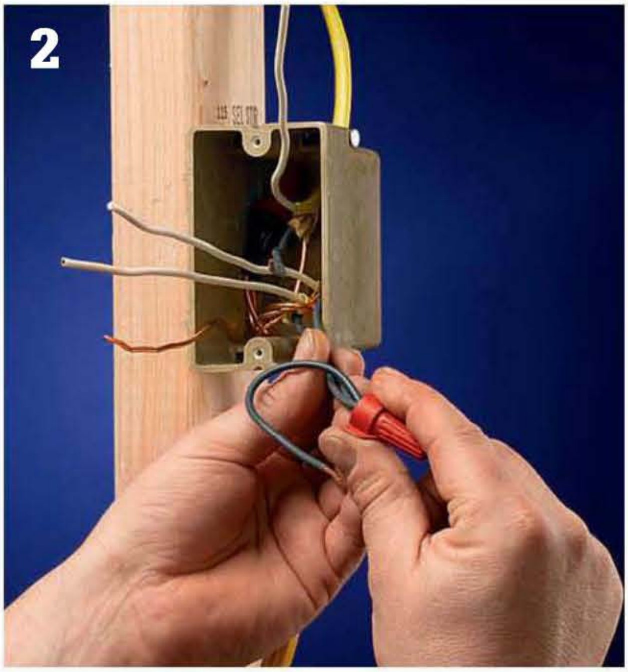

Join one end of the pigtail to the wires that will share the connection using a wire nut (see previous page).

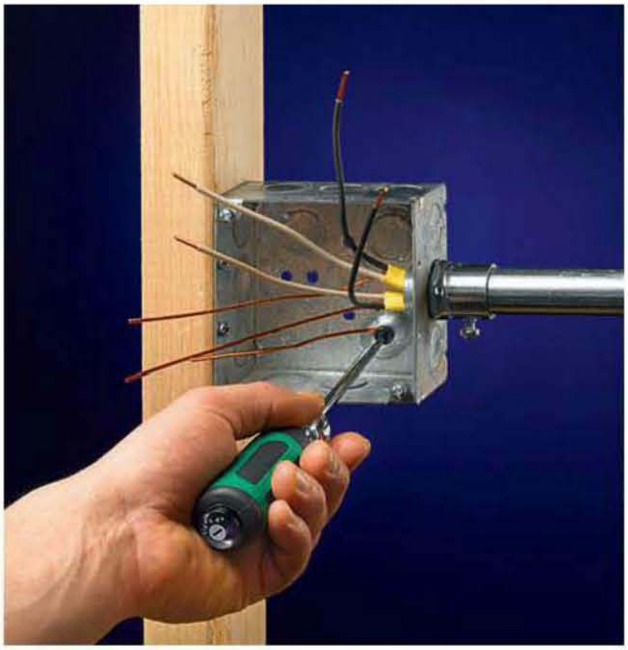

Alternative: If you are pigtailing to a grounding screw or grounding clip in a metal box, you may find it easier to attach one end of the wire to the grounding screw before you attach the other end to the other wires.

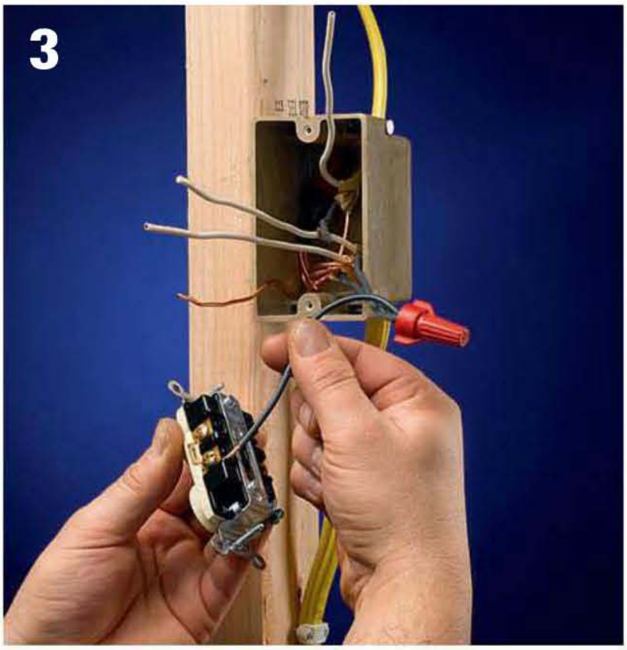

Connect the pigtail to the appropriate terminal on the receptacle or switch. Fold the wires neatly and press the fitting into the box.

How to Install NM Cable

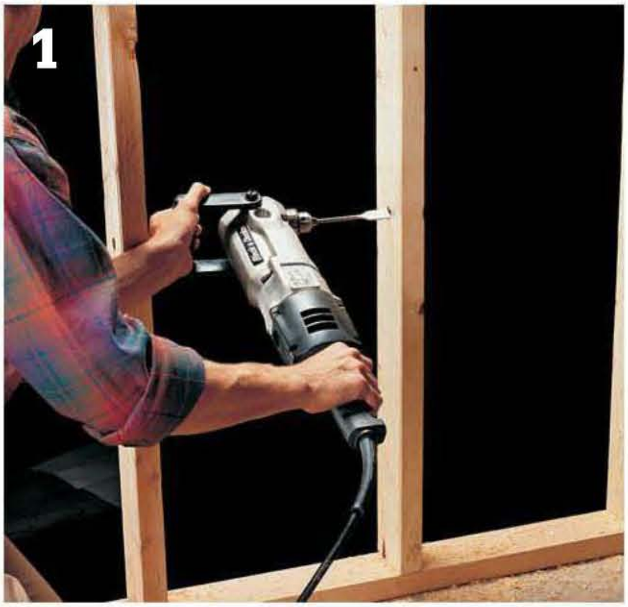

Drill ⅝" holes in framing members for the cable runs. This is done easily with a right-angle drill, available at rental centers. Holes should be set back at least 1¼" from the front face of the framing members.

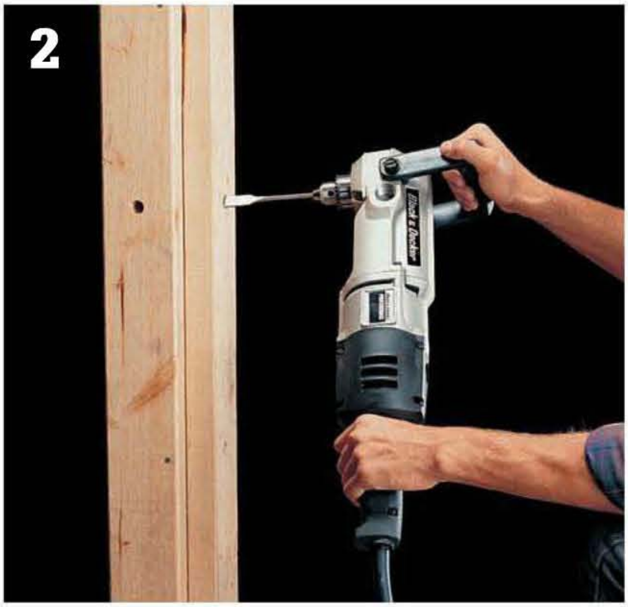

Where cables will turn corners (step 6, page 99), drill intersecting holes in adjoining faces of studs. Measure and cut all cables, allowing 2 ft. extra at ends entering the breaker panel and 1 foot for ends entering the electrical box.

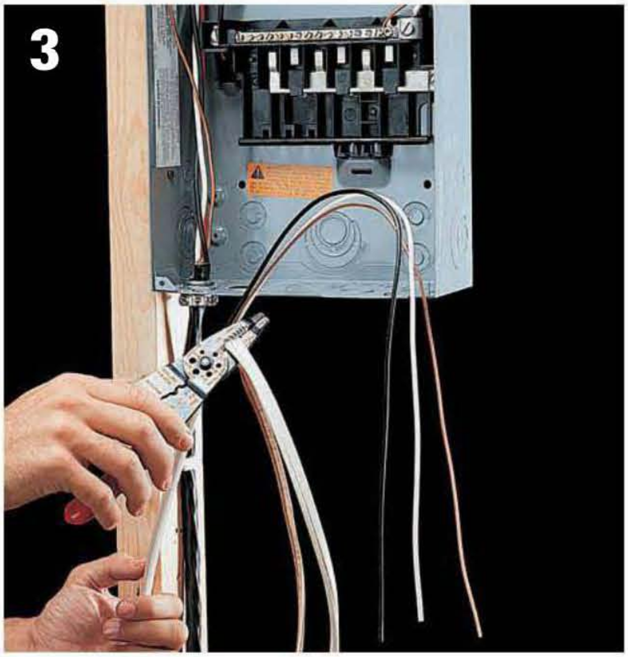

Shut off power to circuit breaker panel. Use a cable ripper to strip cable, leaving at least ¼" of sheathing to enter the circuit breaker panel. Clip away the excess sheathing.

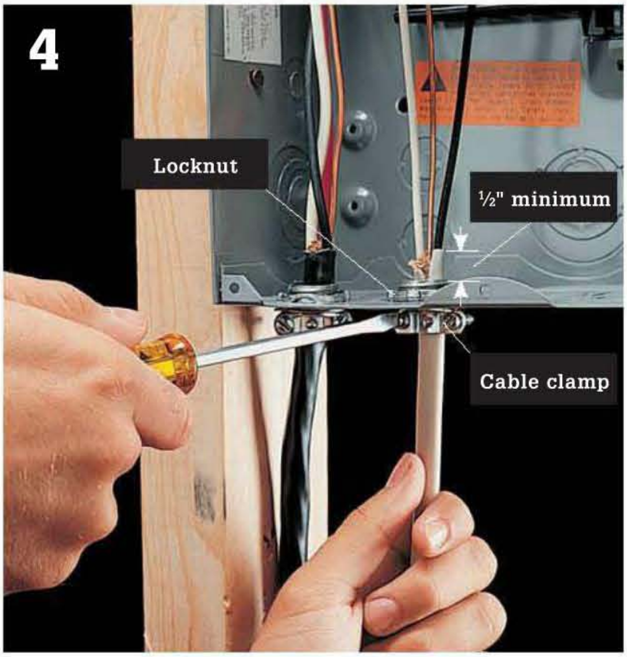

Open a knockout in the circuit breaker panel using a hammer and screwdriver. Insert a cable clamp into the knockout, and secure it with a locknut. Insert the cable through the clamp so that at least ½" of sheathing extends inside the circuit breaker panel. Tighten the mounting screws on the clamp so the cable is gripped securely but not so tightly that the sheathing is crushed.

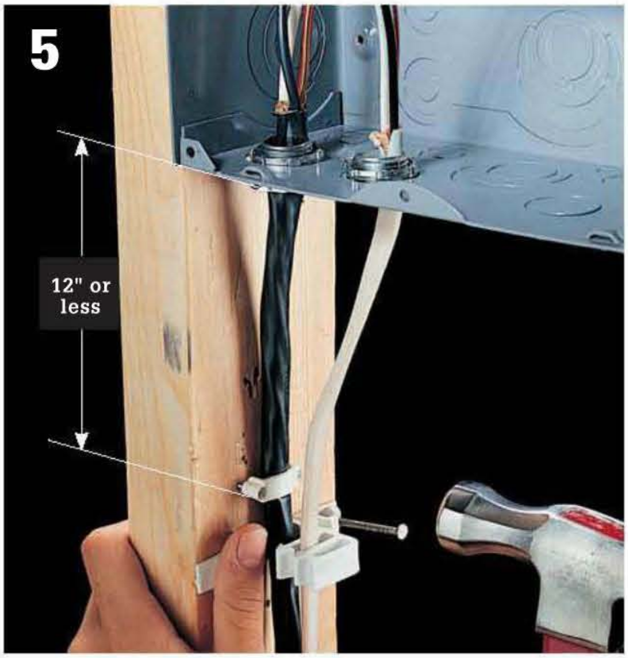

Anchor the cable to the center of a framing member within 12" of the circuit breaker panel using a cable staple. Stack-It® staples work well where two or more cables must be anchored to the same side of a stud. Run the cable to the first electrical box. Where the cable runs along the sides of framing members, anchor it with cable staples no more than 4 ft. apart.

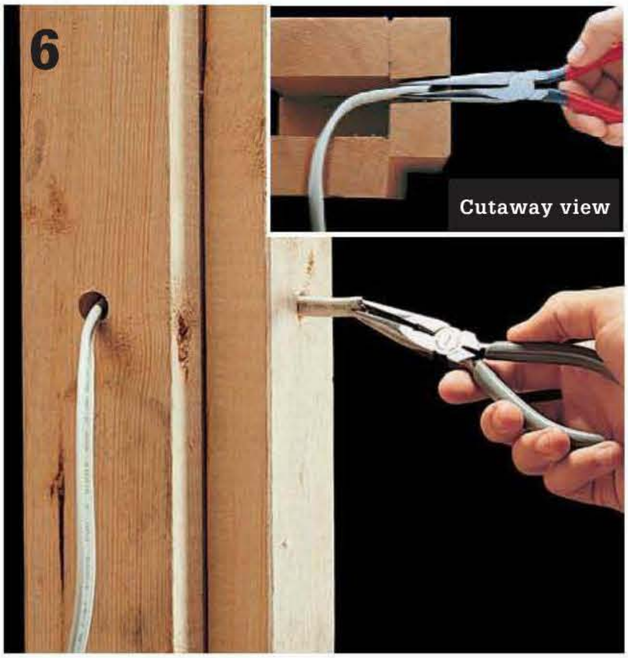

At corners, form a slight L-shaped bend in the end of the cable and insert it into one hole. Retrieve the cable through the other hole using needlenose pliers (inset).

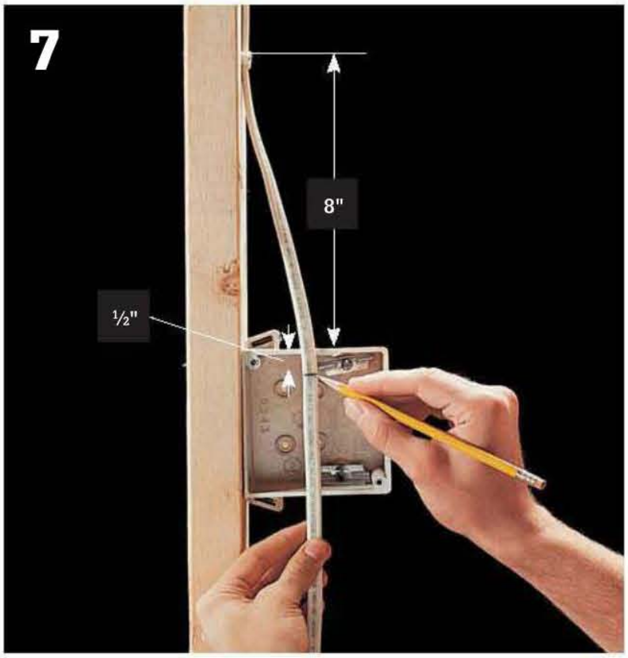

Staple the cable to a framing member 8" from the box. Hold the cable taut against the front of the box, and mark a point on the sheathing ½" past the box edge. Remove sheathing from the marked line to the end using a cable ripper, and clip away excess sheathing with a combination tool. Insert the cable through the knockout in the box.

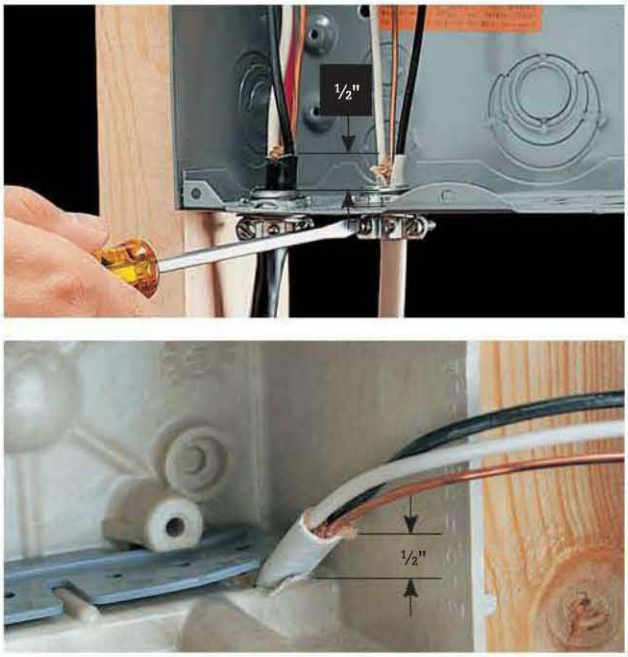

Variation: Different types of boxes have different clamping devices. Make sure cable sheathing extends ½" past the edge of the clamp to ensure that the cable is secure and that the wire won't be damaged by the edges of the clamp.

(continued)

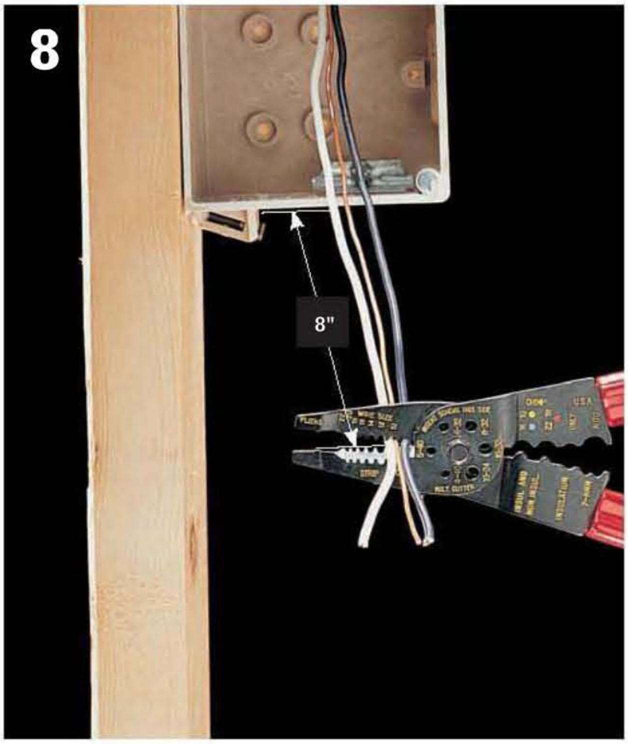
8

As each cable is installed in a box, clip back each wire so that 8" of workable wire extends past the front edge of the box.

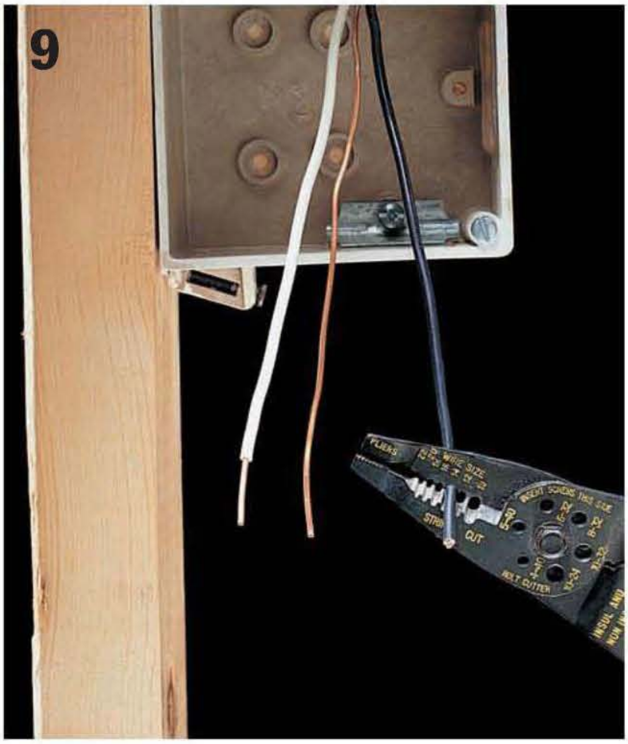
9

Strip ¾" of insulation from each circuit wire in the box using a combination tool. Take care not to nick the copper.

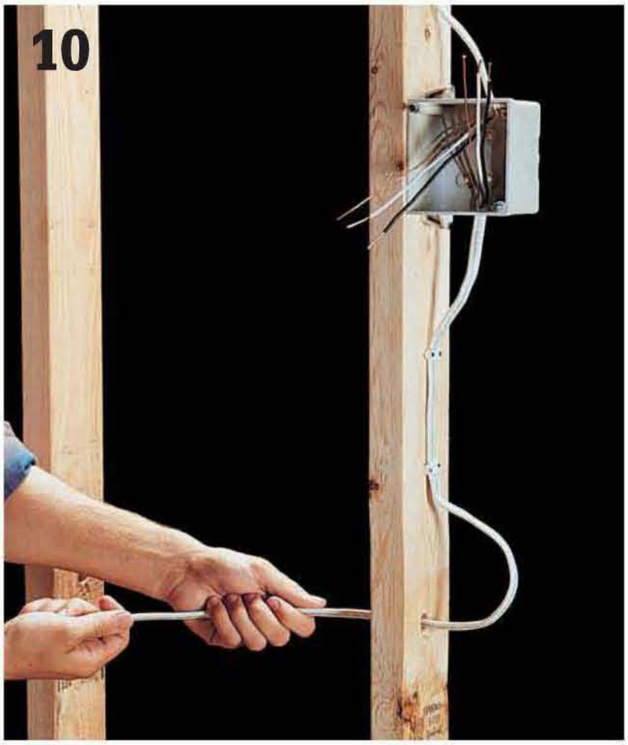
10

Continue the circuit by running cable between each pair of electrical boxes, leaving an extra 1 ft. of cable at each end.

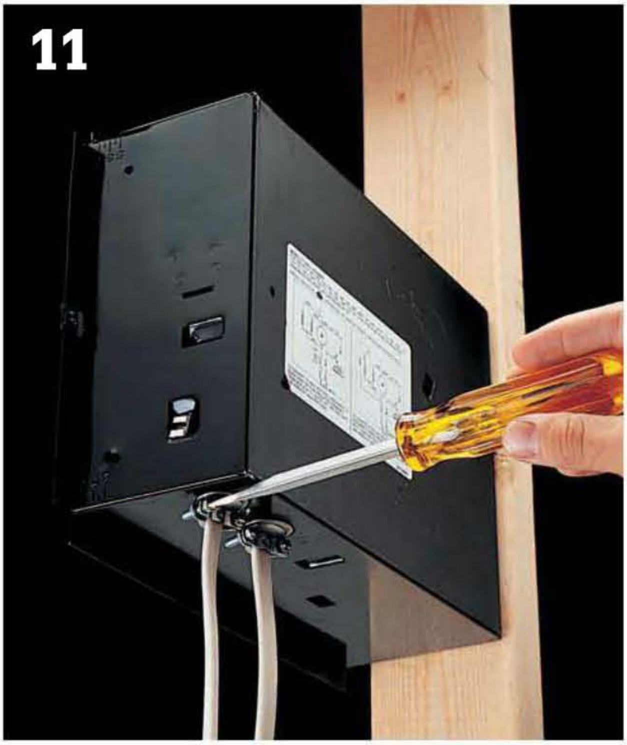
11

At metal boxes and recessed fixtures, open knockouts, and attach cables with cable clamps. From inside fixture, strip away all but ¼" of sheathing. Clip back wires so there is 8" of workable length, then strip ¾" of insulation from each wire.

12

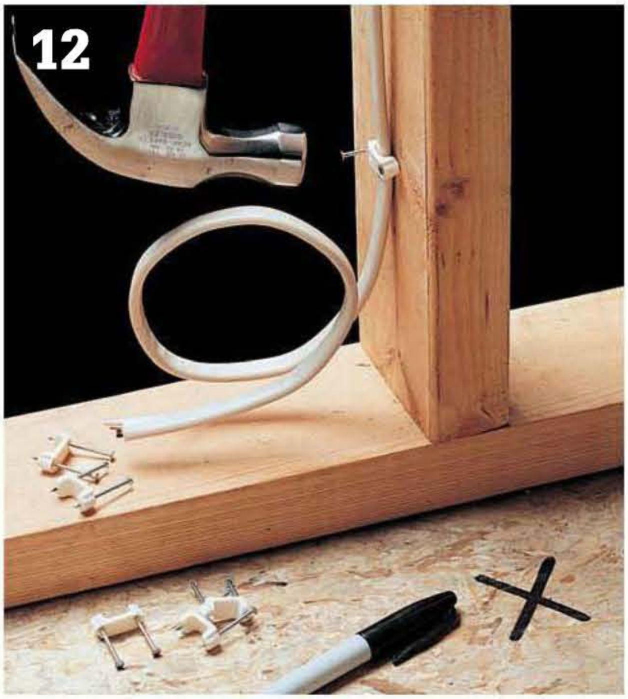

For a surface-mounted fixture like a baseboard heater or fluorescent light fixture, staple the cable to a stud near the fixture location, leaving plenty of excess cable. Mark the floor so the cable will be easy to find after the walls are finished.

13

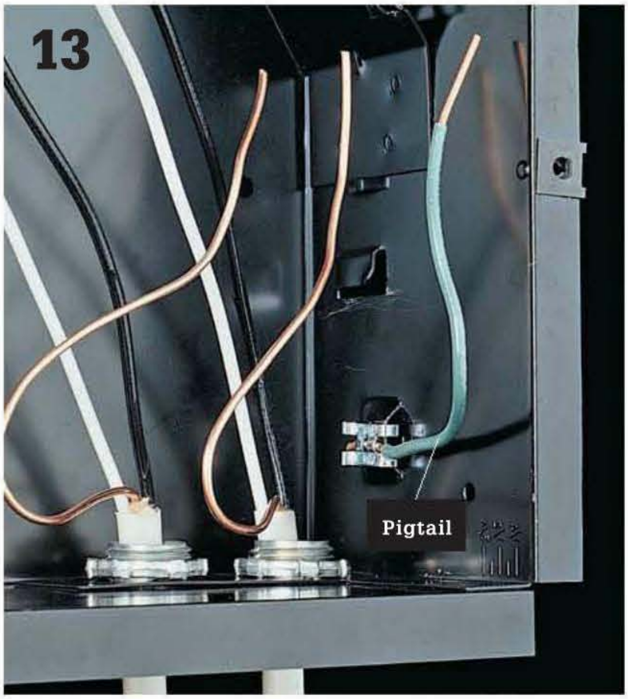

At each recessed fixture and metal electrical box, connect one end of a grounding pigtail to the metal frame using a grounding clip attached to the frame (shown above) or a green grounding screw.

14

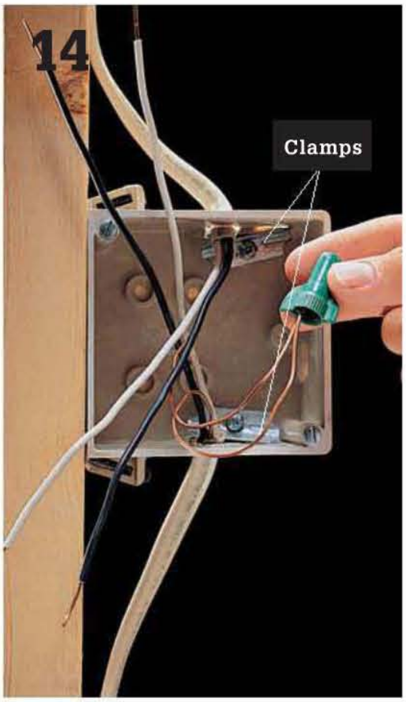

At each electrical box and recessed fixture, join grounding wires together with a wire connector. If the box has internal clamps, tighten the clamps over the cables.

15

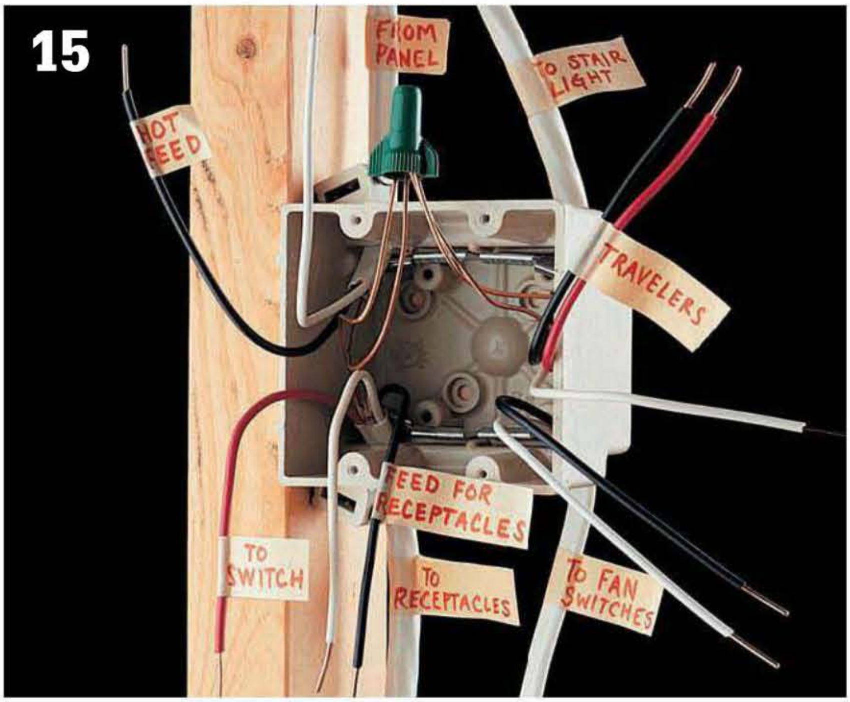
HOT FEED
FROM PANEL
TO STAIR LIGHT
TRAVELERS
FEED FOR RECEPTACLES
TO SWITCH
TO RECEPTACLES
TO FAN SWITCHES

Label the cables entering each box to indicate their destinations. In boxes with complex wiring configurations, also tag the individual wires to make final hookups easier. After all cables are installed, your rough-in work is ready to be reviewed by the electrical inspector.

Working with Conduit

Electrical wiring that runs in exposed locations must be protected by rigid tubing called conduit. For example, conduit is used for wiring that runs across masonry walls in a basement laundry and for exposed outdoor wiring. THHN/THWN wire (page 92) normally is installed inside conduit, although UF or NM cable can also be installed in conduit.

There are several types of conduit available, so check with your electrical inspector to find out which type meets code requirements in your area. Conduit installed outdoors must be rated for exterior use. Metal conduit should be used only with metal boxes, never with plastic boxes.

At one time, conduit could only be fitted by using elaborate bending techniques and special tools. Now, however, a variety of shaped fittings are available to let a homeowner join conduit easily.

Electrical Grounding in Metal Conduit

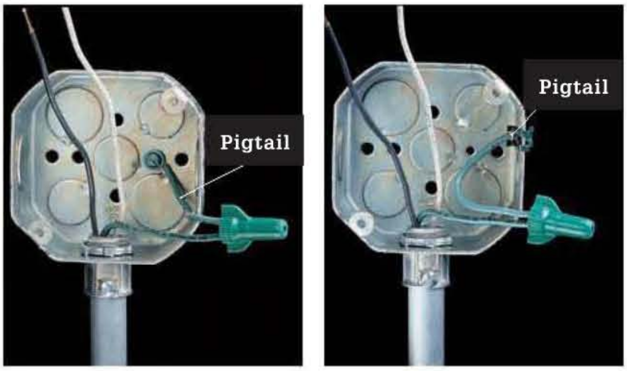

Install a green insulated grounding wire for any circuit that runs through metal conduit. Although code allows the metal conduit to serve as the grounding conductor, most electricians install a green insulated wire as a more dependable means of grounding the system. The grounding wires must be connected to metal boxes with a pigtail and grounding screw (left) or grounding clip (right).

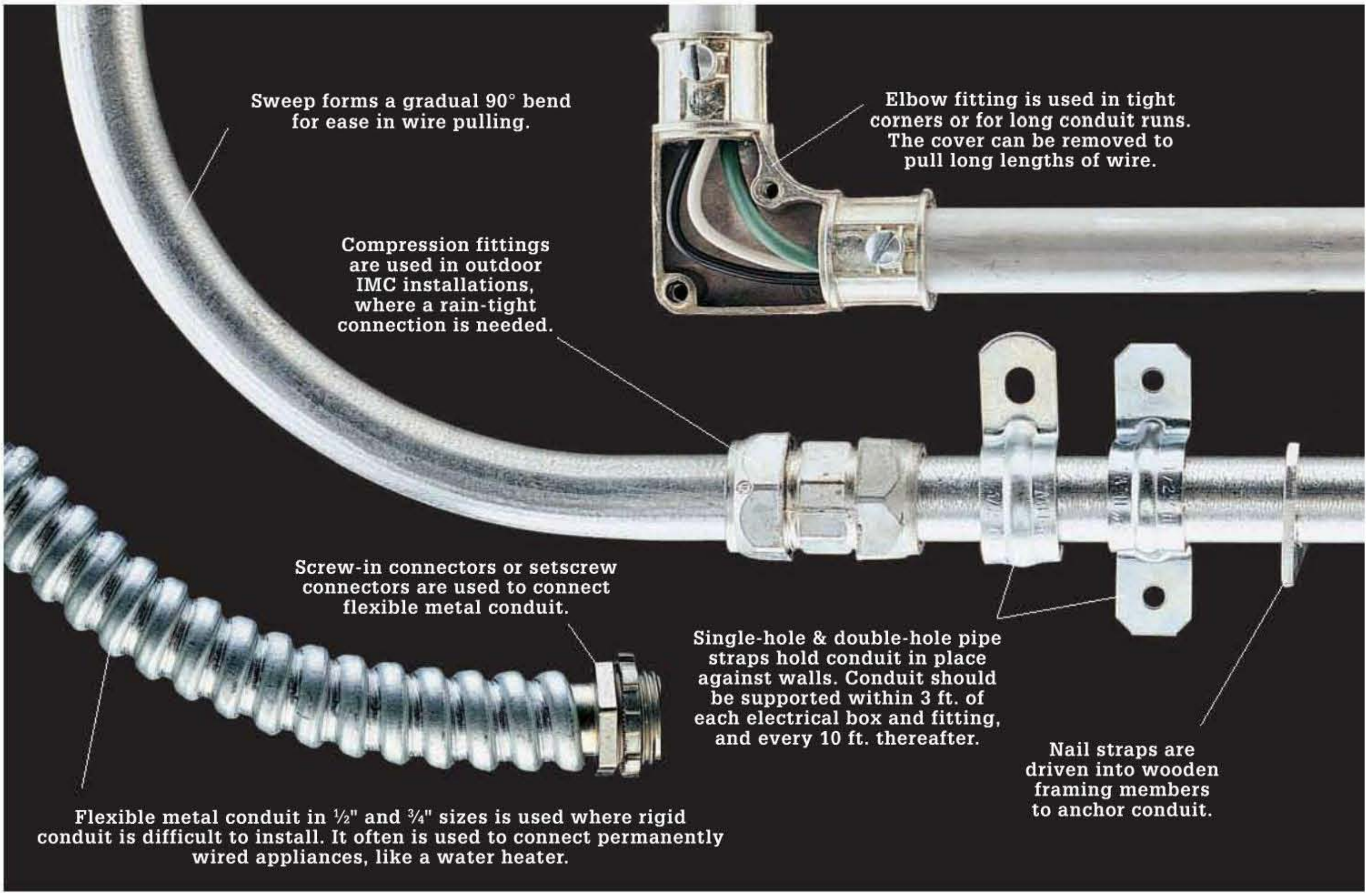

Sweep forms a gradual 90° bend for ease in wire pulling.

Elbow fitting is used in tight corners or for long conduit runs. The cover can be removed to pull long lengths of wire.

Compression fittings are used in outdoor IMC installations, where a rain-tight connection is needed.

Screw-in connectors or setscrew connectors are used to connect flexible metal conduit.

Single-hole & double-hole pipe straps hold conduit in place against walls. Conduit should be supported within 3 ft. of each electrical box and fitting, and every 10 ft. thereafter.

Nail straps are driven into wooden framing members to anchor conduit.

Flexible metal conduit in ½" and ¾" sizes is used where rigid conduit is difficult to install. It often is used to connect permanently wired appliances, like a water heater.

Fill Capacity

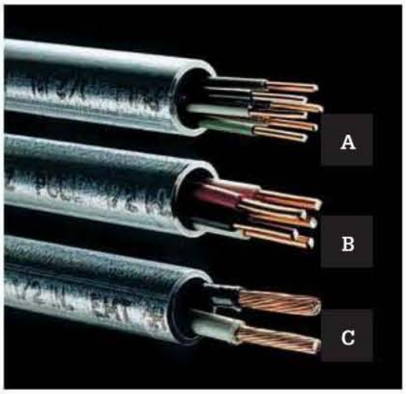

Conduit ½" in diameter can hold up to six 14-gauge or 12-gauge THHN/THWN wires (A), five 10-gauge wires (B), or two 8-gauge wires (C). Use ¾" conduit for greater capacity.

Metal Conduit

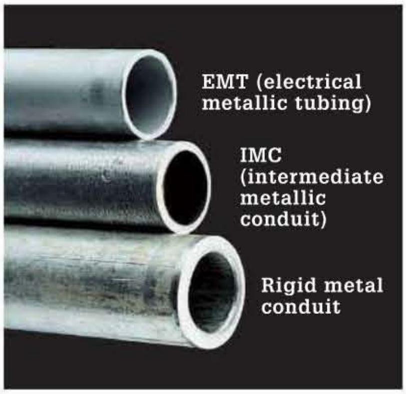
EMT (electrical metallic tubing)
IMC (intermediate metallic conduit)
Rigid metal conduit

EMT is lightweight and easy to install but should not be used where it can be damaged. IMC has thicker galvanized walls and is a good choice for exposed outdoor use. Rigid metal conduit provides the greatest protection for wires, but it is more expensive and requires threaded fittings.

Plastic Conduit

Plastic PVC conduit is allowed by many local codes. It is assembled with solvent glue and PVC fittings that resemble those for metal conduit. When wiring with PVC conduit, always run a green grounding wire.

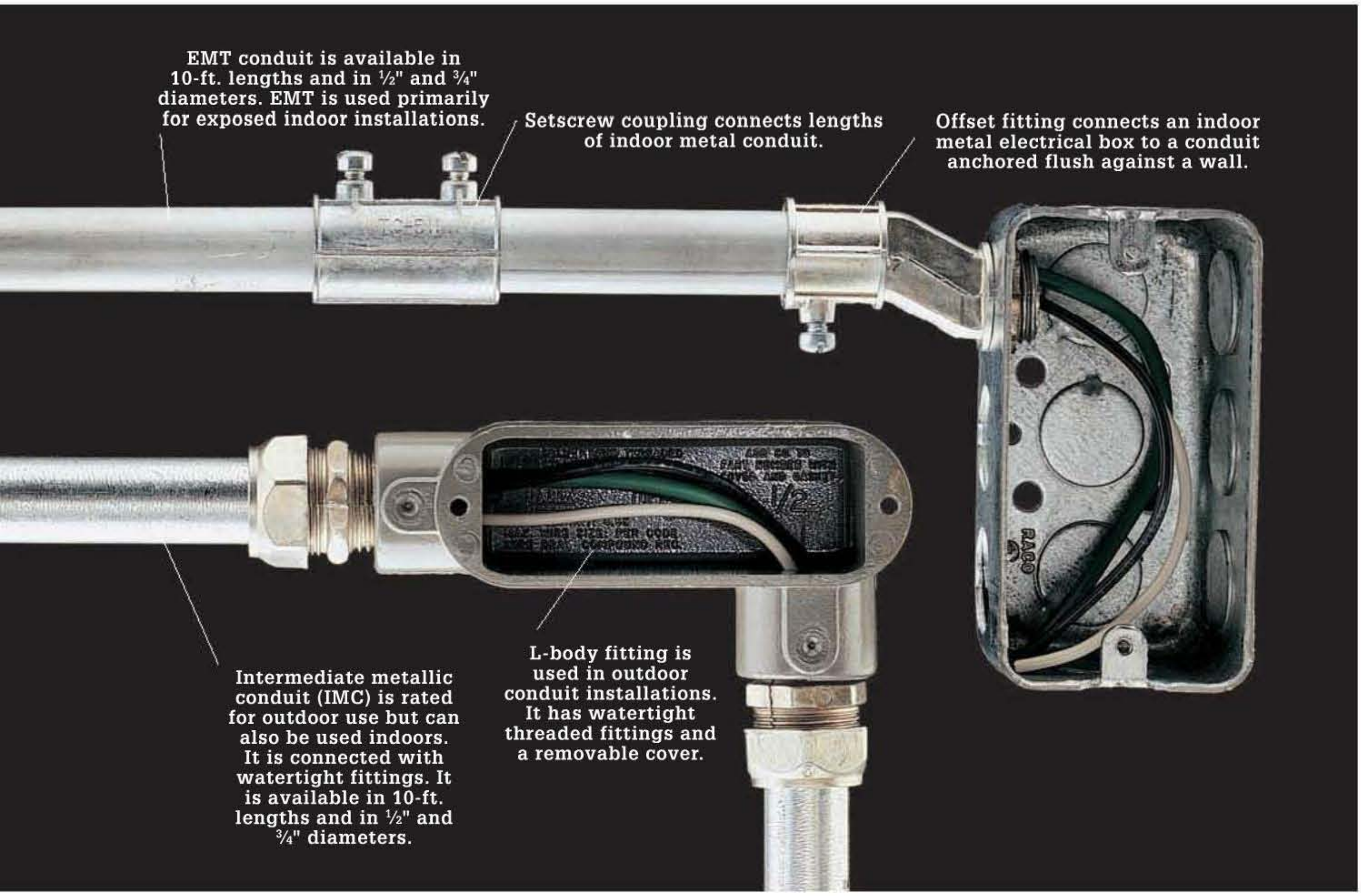

EMT conduit is available in 10-ft. lengths and in ½" and ¾" diameters. EMT is used primarily for exposed indoor installations.

Setscrew coupling connects lengths of indoor metal conduit.

Offset fitting connects an indoor metal electrical box to a conduit anchored flush against a wall.

Intermediate metallic conduit (IMC) is rated for outdoor use but can also be used indoors. It is connected with watertight fittings. It is available in 10-ft. lengths and in ½" and ¾" diameters.

L-body fitting is used in outdoor conduit installations. It has watertight threaded fittings and a removable cover.

Conduit Materials & Tools

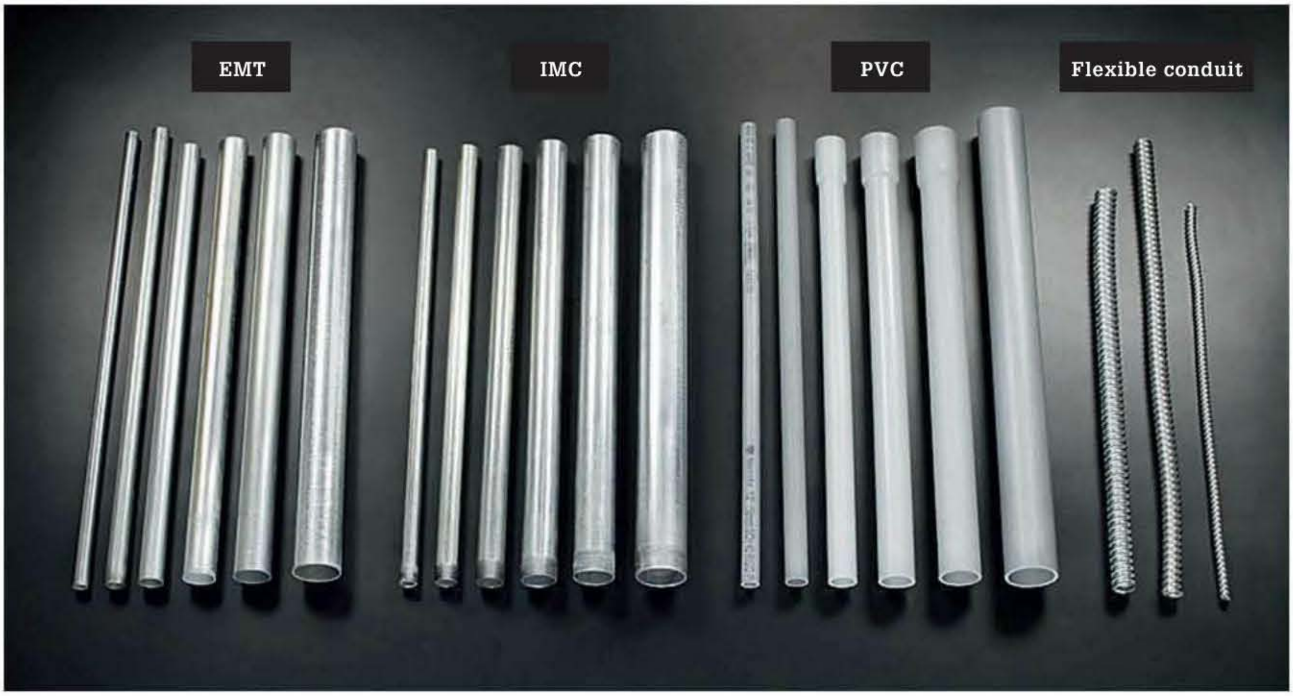
EMT
IMC
PVC
Flexible conduit

Conduit types used most in homes are EMT (electrical metallic tubing), IMC (intermediate metallic conduit), RNC (rigid nonmetallic conduit), and flexible metal conduit. The most common diameters by far are ½" and ¾", but larger sizes are stocked at most building centers.

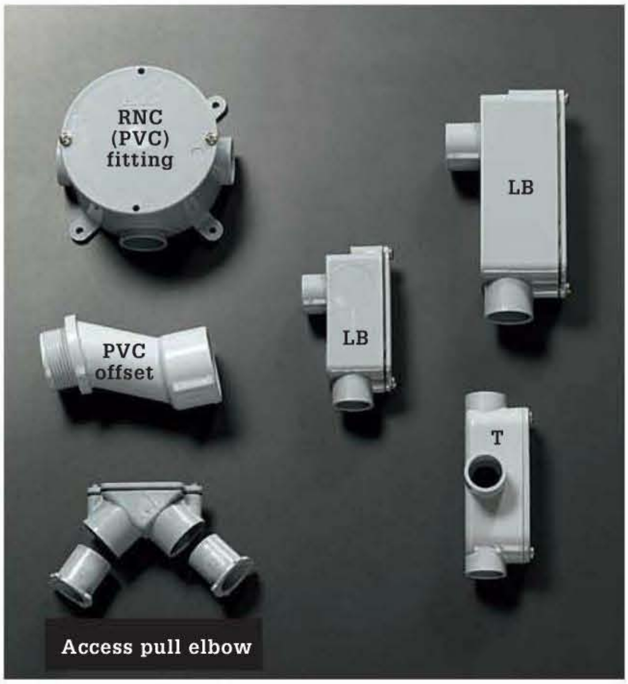
RNC (PVC) fitting
LB
LB
PVC offset
T
Access pull elbow

Nonmetallic conduit fittings typically are solvent welded to nonmetallic conduit, as opposed to metal conduit, which can be threaded and screwed into threaded fittings or attached with setscrews or compression fittings.

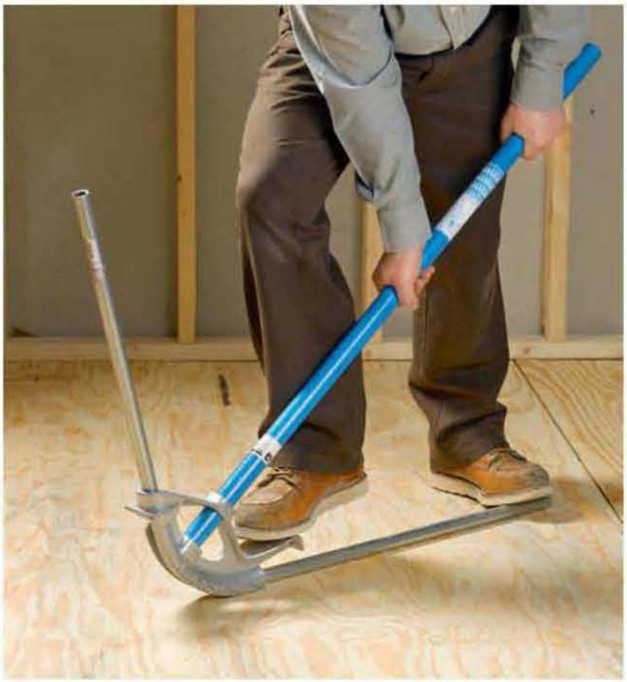

A thin-wall conduit bender is used to bend sweeps into EMT or IMC conduit.

How to Make Nonmetallic Conduit Connections

1

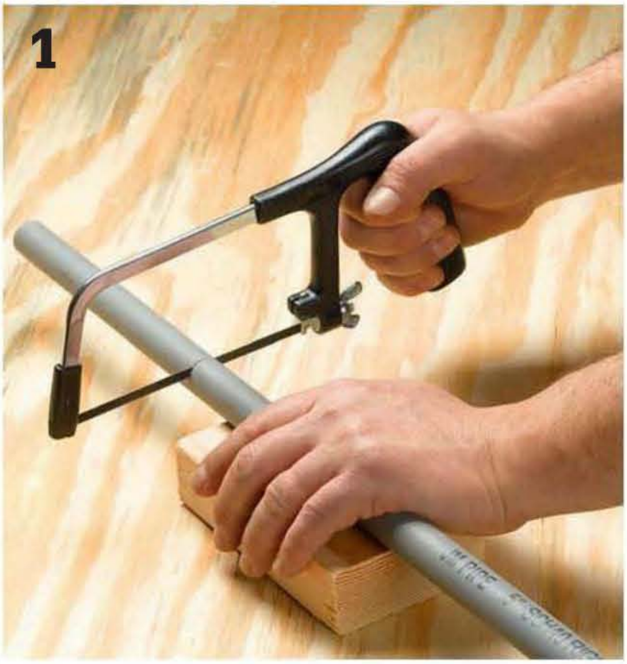

Cut the rigid nonmetallic conduit (RNC) to length with a fine-tooth saw, such as a hacksaw. For larger diameter (1½" and above), use a power miter box with a fine-tooth or plastic cutting blade.

2

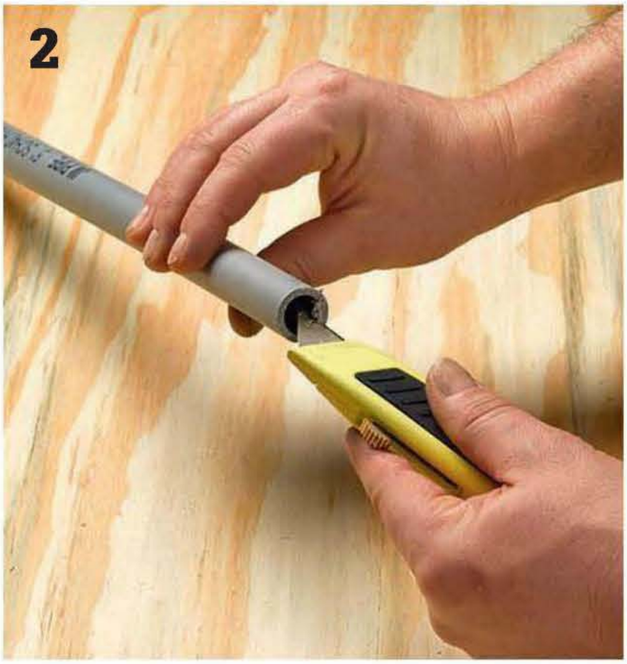

Deburr the cut edges with a utility knife or fine sandpaper such as emery paper. Wipe the cut ends with a dry rag. Also wipe the coupling or fitting to clean it.

3

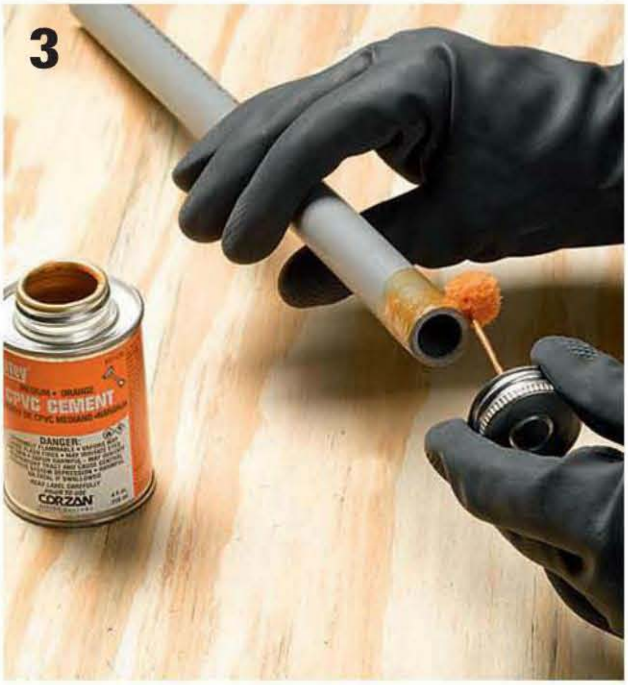

Apply a coat of PVC cement to the end of the conduit. Wear latex gloves to protect your hands. The cement should be applied past the point on the conduit where it enters the fitting or coupling.

4

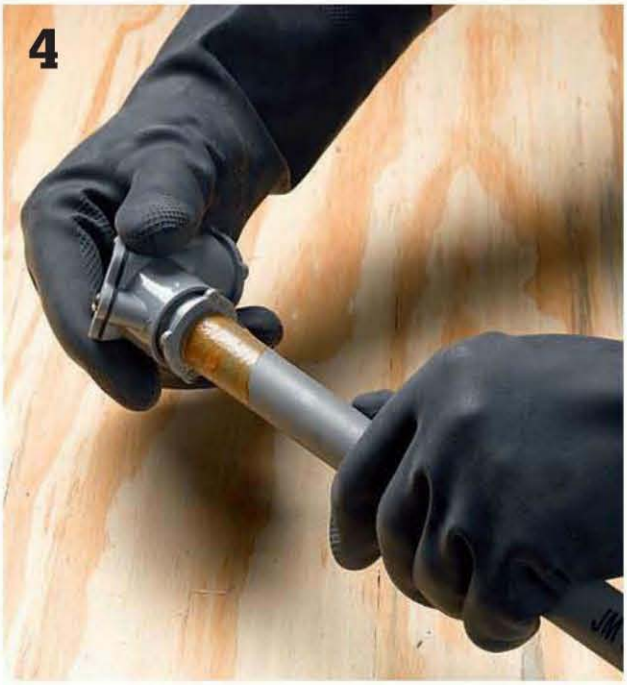

Insert the conduit into the fitting or coupling and spin it a quarter turn to help spread the cement. Allow the joint to set undisturbed for 10 minutes.

How to Install Conduit & Wires on a Concrete Wall

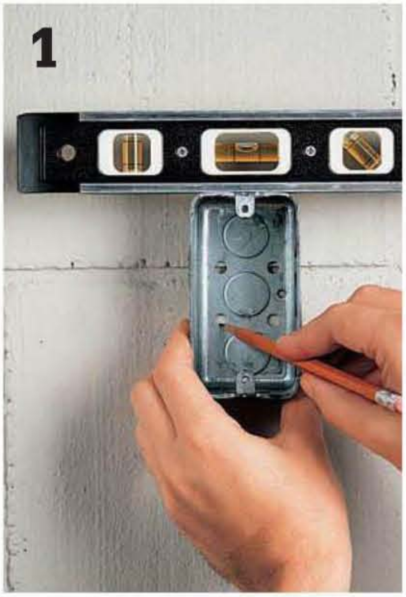

Measure from the floor to position electrical boxes on the wall, and mark location for mounting screws. Boxes for receptacles in an unfinished basement or other damp area are mounted at least 2 ft. from the floor. Laundry receptacles usually are mounted at 48".

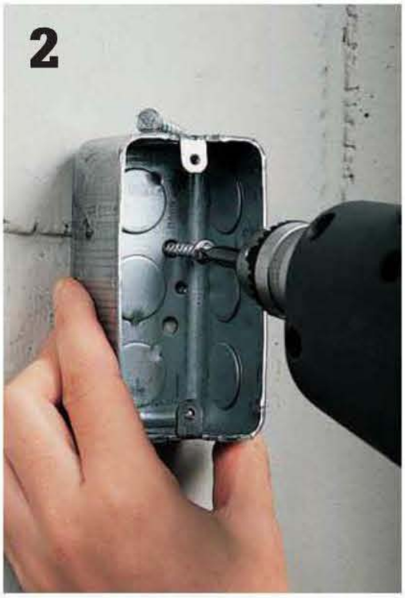

Drill pilot holes with a masonry bit, then mount the box against a masonry wall with masonry anchors. Or use masonry anchors and panhead screws.

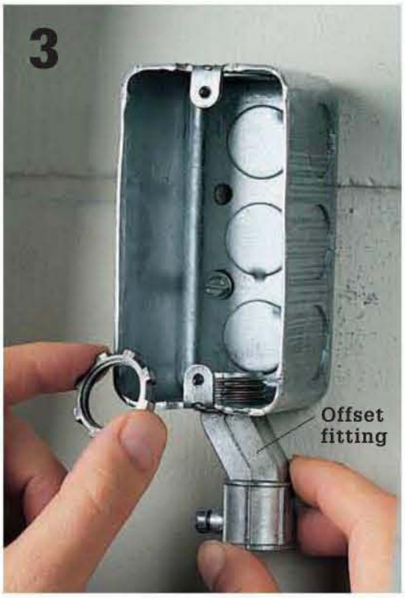

Open one knockout for each length of conduit that will be attached to the box. Attach an offset fitting to each knockout using a locknut.

Measure the first length of conduit and cut it with a hacksaw. Remove any rough inside edges with a pipe reamer or a round file. Attach the conduit to the offset fitting on the box, and tighten the setscrew.

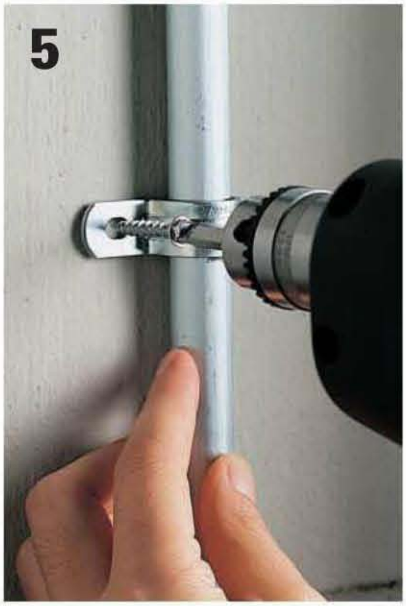

Anchor the conduit against the wall with pipe straps and masonry anchors. Conduit should be anchored within 3 ft. of each box and fitting, and every 10 ft. thereafter.

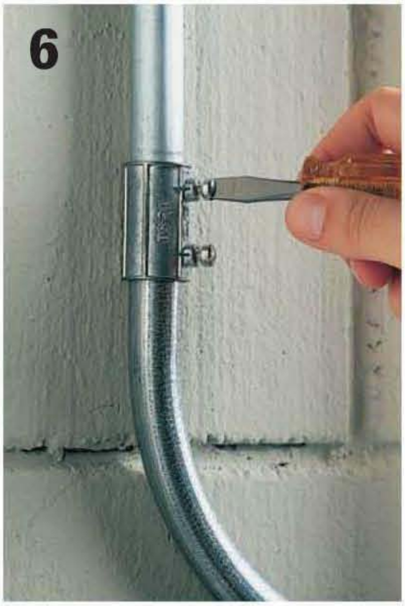

Make conduit bends by attaching a sweep fitting using a setscrew fitting or compression fitting. Continue conduit run by attaching additional lengths using setscrew or compression fittings.

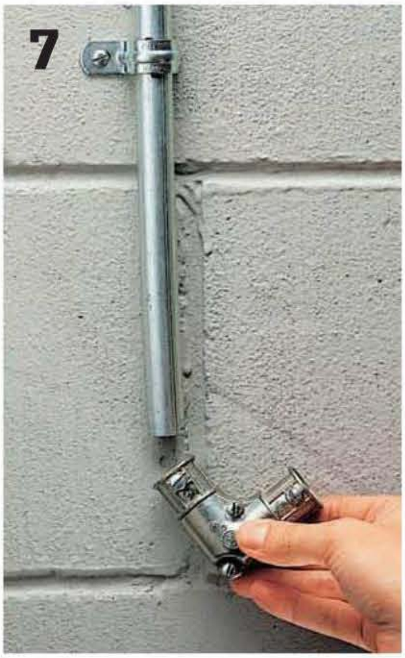

Use an elbow fitting in conduit runs that have many bends, or in runs that require very long wires. The cover on the elbow fitting can be removed to make it easier to extend a fish tape and pull wires.

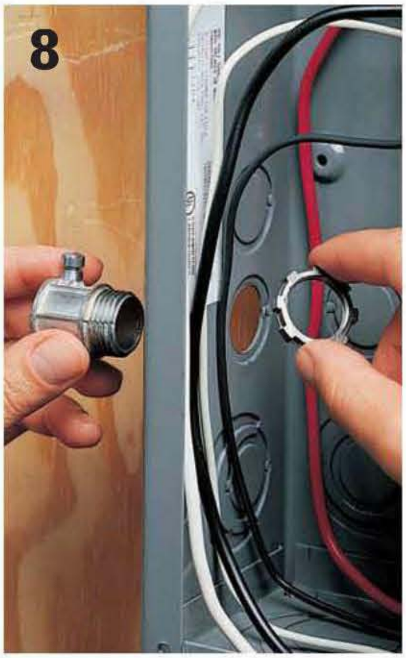

At the service breaker panel, turn the power off, then remove the cover and test for power. Open a knockout in the panel, then attach a setscrew fitting, and install the last length of conduit.

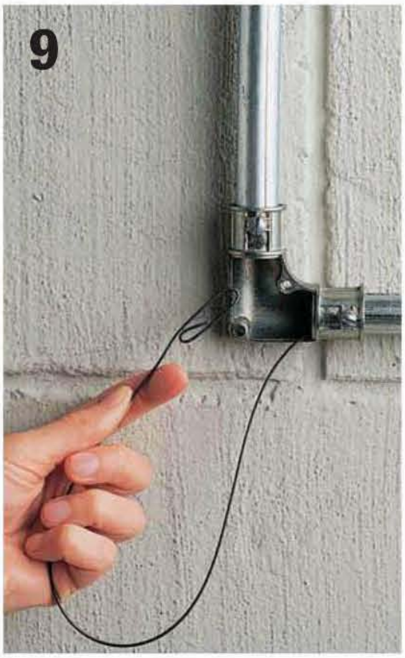

Unwind the fish tape and extend it through the conduit from the circuit breaker panel outward. Remove the cover on an elbow fitting when extending the fish tape around tight corners.

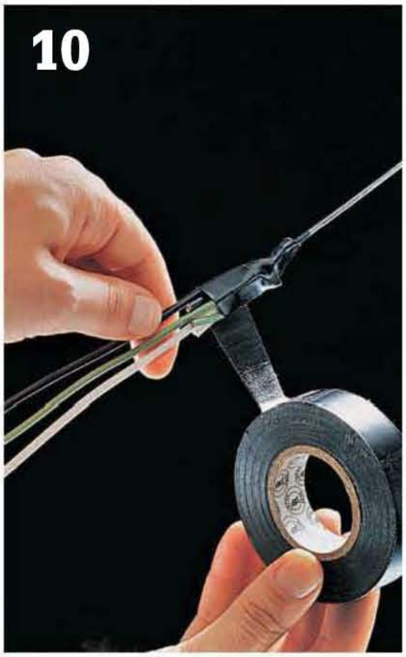

Trim back 2" of outer insulation from the end of the wires, then insert the wires through the loop at the tip of the fish tape.

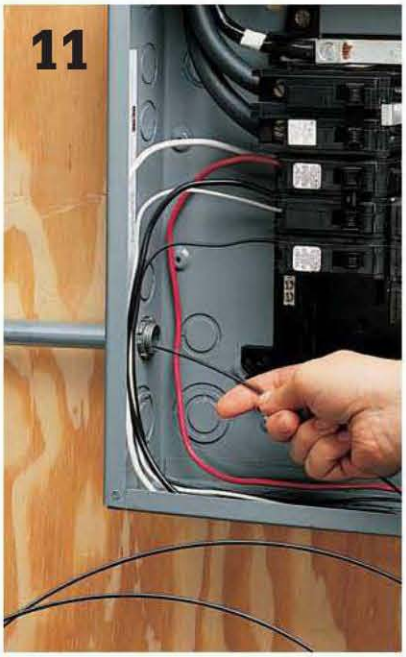

Retrieve the wires through the conduit by pulling on the fish tape with steady pressure. *Note: Use extreme care when using a metal fish tape inside a circuit breaker panel, even when the power is turned off.*

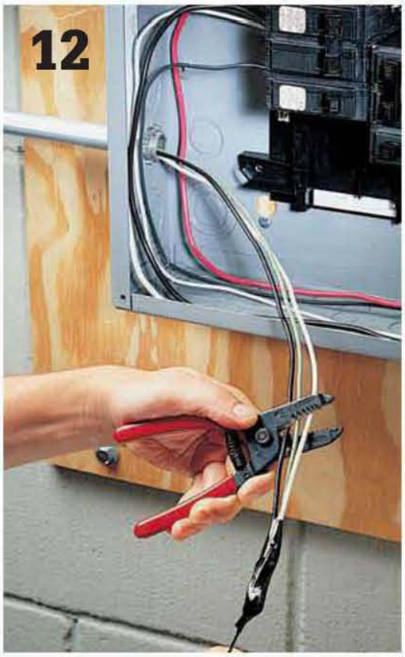

Clip off the taped ends of the wires. Leave at least 2 ft. of wire at the service panel and 8" at each electrical box.

Working with Boxes

The National Electrical Code requires that wire connections and cable splices be contained inside an approved metal or plastic box. This shields framing members and other flammable materials from electrical sparks.

Electrical boxes come in several shapes. Rectangular and square boxes are used for switches and receptacles. Rectangular (2 × 3") boxes are used for single switches or duplex receptacles. Square (4 × 4") boxes are used any time it is convenient for two switches or receptacles to be wired, or "ganged," in one box, an arrangement common in kitchens or entry hallways. Octagonal electrical boxes contain wire connections for ceiling fixtures.

All electrical boxes are available in different depths. A box must be deep enough so a switch or receptacle can be removed or installed easily without crimping and damaging the circuit wires. Replace an undersized box with a larger box using the Electrical Box Chart (right) as a guide. The NEC also says that all electrical boxes must remain accessible. Never cover an electrical box with drywall, paneling, or wallcoverings.

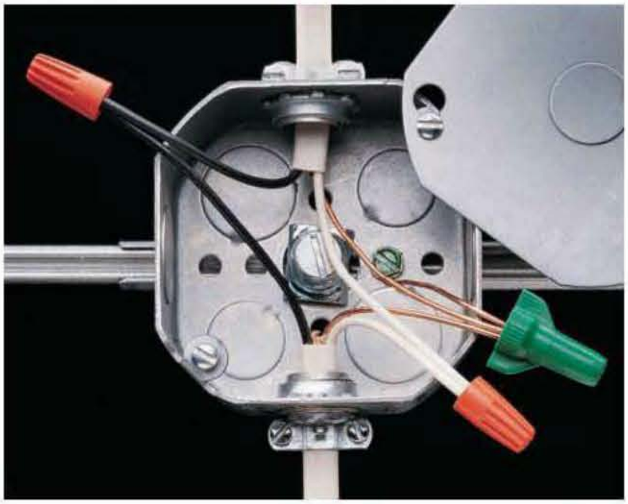

Octagonal boxes usually contain wire connections for ceiling fixtures. Cables are inserted into the box through knockout openings and are held with cable clamps. Because the ceiling fixture attaches directly to the box, the box should be anchored firmly to a framing member. Often, it is nailed directly to a ceiling joist. However, metal braces are available that allow a box to be mounted between joists or studs. A properly installed octagonal box can support a ceiling fixture weighing up to 35 pounds. Any box must be covered with a tightly fitting cover plate, and the box must not have open knockouts.

Electrical Box Fill Chart ▸

Box size and shape	Maximum number of conductors permitted (see Notes below)			
	18 AWG	16 AWG	14 AWG	12 AWG
Junction boxes				
4 × 1¼" R or O	8	7	6	5
4 × 1½" R or O	10	8	7	6
4 × 2⅛" R or O	14	12	10	9
4 × 1¼" S	12	10	9	8
4 × 1½" S	14	12	10	9
4 × 2⅛" S	20	17	15	13
4 11/16 × 1¼" S	17	14	12	11
4 11/16 × 1½" S	19	16	14	13
4 11/16 × 2⅛" S	28	24	21	18
Device boxes				
3 × 2 × 1½"	5	4	3	3
3 × 2 × 2"	6	5	5	4
3 × 2 × 2¼"	7	6	5	4
3 × 2 × 2½"	8	7	6	5
3 × 2 × 2¾"	9	8	7	6
3 × 2 × 3½"	12	10	9	8
4 × 2⅛ × 1½"	6	5	5	4
4 × 2⅛ × 1⅞"	8	7	6	5
4 × 2⅛ × 2⅛"	9	8	7	6

Notes:

- R = Round; O = Octagonal; S = Square or rectangular
- Each hot or neutral wire entering the box is counted as one conductor.
- Grounding wires are counted as one conductor in total—do not count each one individually.
- Raceway fittings and external cable clamps do not count. Internal cable connectors and straps count as either half or one conductor, depending on type.
- Devices (switches and receptacles mainly) each count as two conductors.
- Straps (yokes) from mounting devices each count as two conductors.
- When calculating total conductors, any nonwire components should be assigned the gauge of the largest wire in the box.
- For wire gauges not shown here, contact your local electrical inspections office.

Common Electrical Boxes

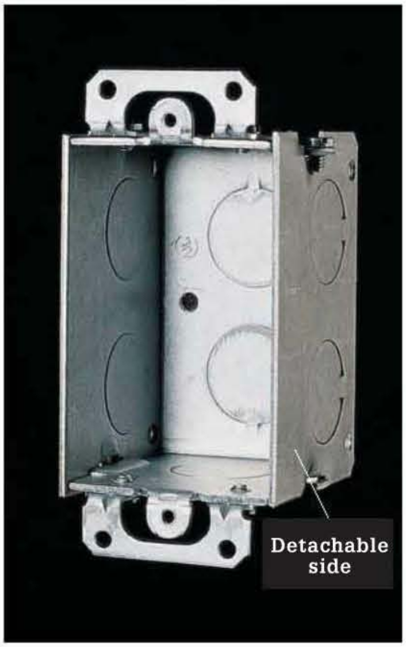

Rectangular boxes are used with wall switches and duplex receptacles. Single-size rectangular boxes (shown above) may have detachable sides that allow them to be ganged together to form double-size boxes.

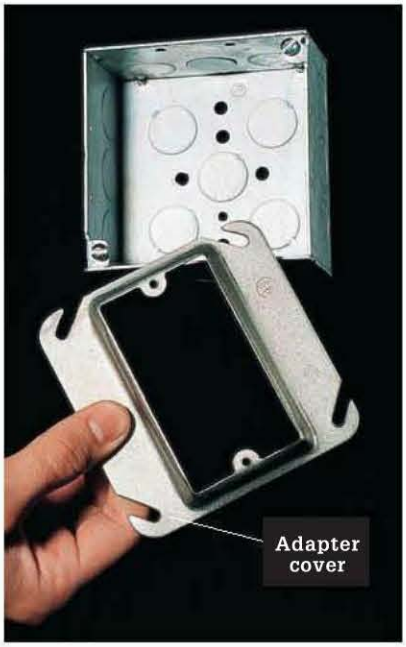

Square 4 × 4" boxes are large enough for most wiring applications. They are used for cable splices and ganged receptacles or switches. To install one switch or receptacle in a square box, use an adapter cover.

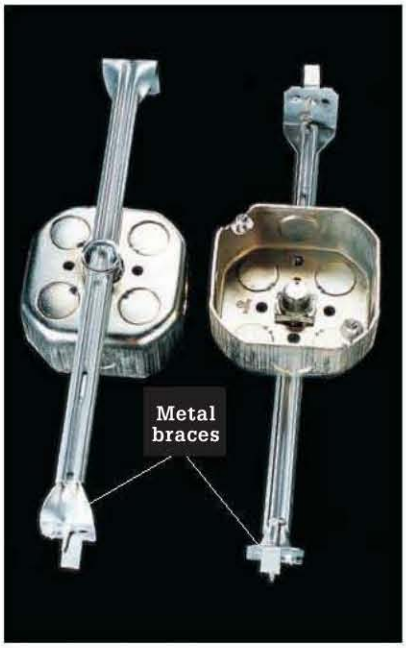

Braced octagonal boxes fit between ceiling joists. The metal braces extend to fit any joist spacing and are nailed or screwed to framing members.

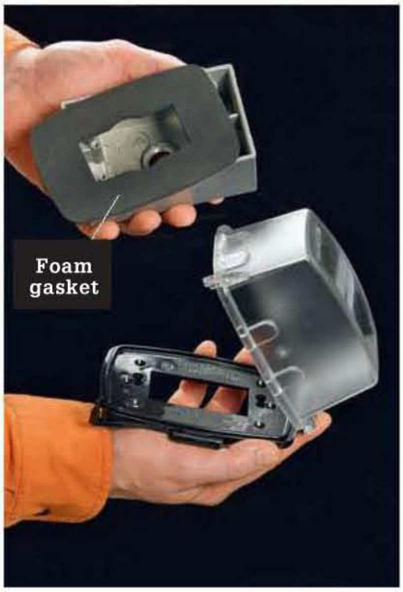

Outdoor boxes have sealed seams and foam gaskets to guard a switch or receptacle against moisture. Corrosion-resistant coatings protect all metal parts. Code-compliant models include a watertight hood.

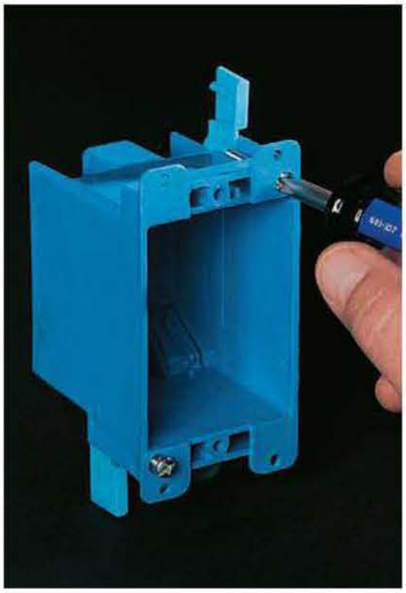

Retrofit boxes upgrade older boxes to larger sizes. One type (above) has built-in clamps that tighten against the inside of a wall and hold the box in place.

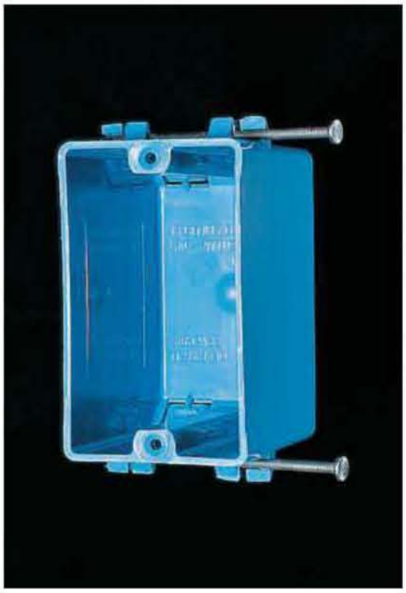

Plastic boxes are common in new construction. They can be used only with NM (nonmetallic) cable. The box may include preattached nails for anchoring it to framing members. Wall switches must have grounding screws if installed in plastic boxes.

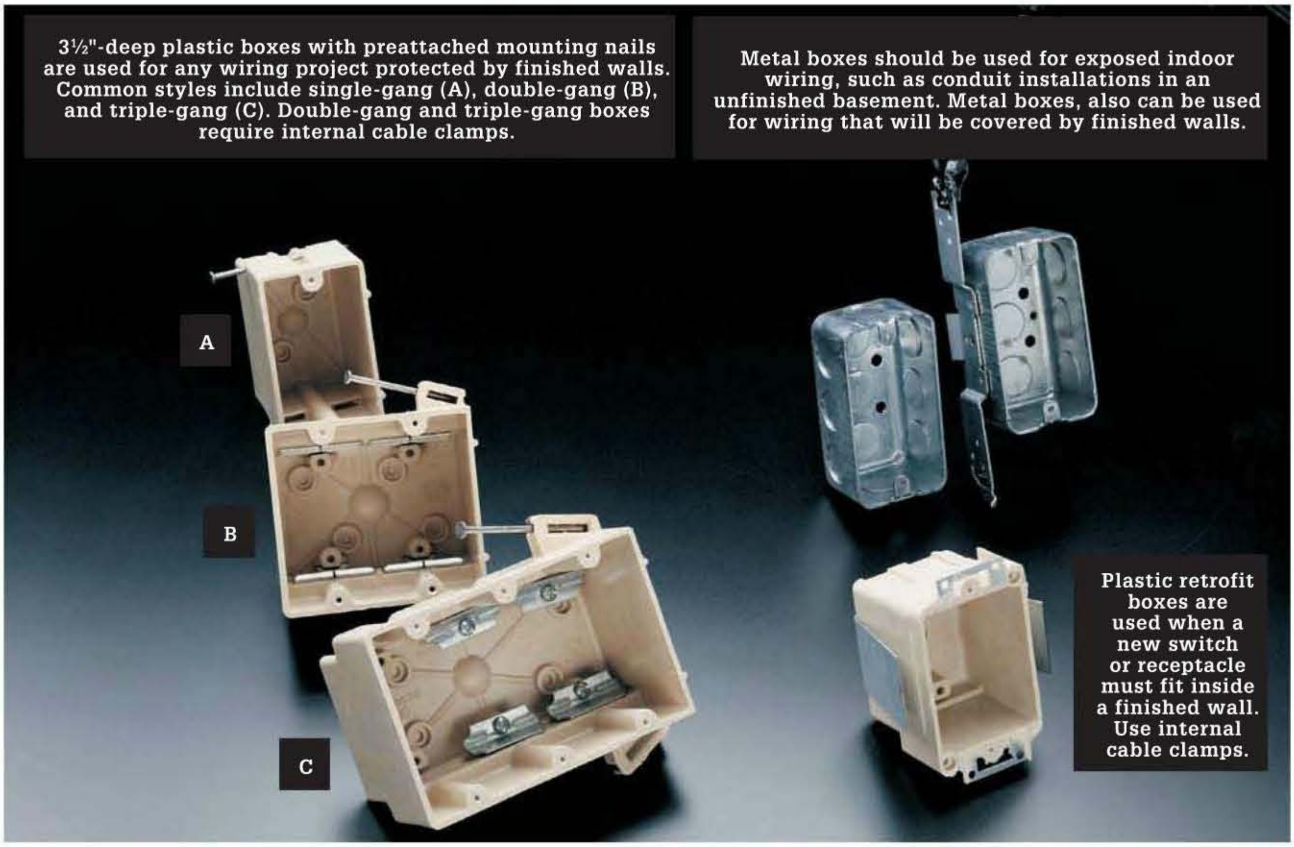

3½"-deep plastic boxes with preattached mounting nails are used for any wiring project protected by finished walls. Common styles include single-gang (A), double-gang (B), and triple-gang (C). Double-gang and triple-gang boxes require internal cable clamps.

Metal boxes should be used for exposed indoor wiring, such as conduit installations in an unfinished basement. Metal boxes, also can be used for wiring that will be covered by finished walls.

Plastic retrofit boxes are used when a new switch or receptacle must fit inside a finished wall. Use internal cable clamps.

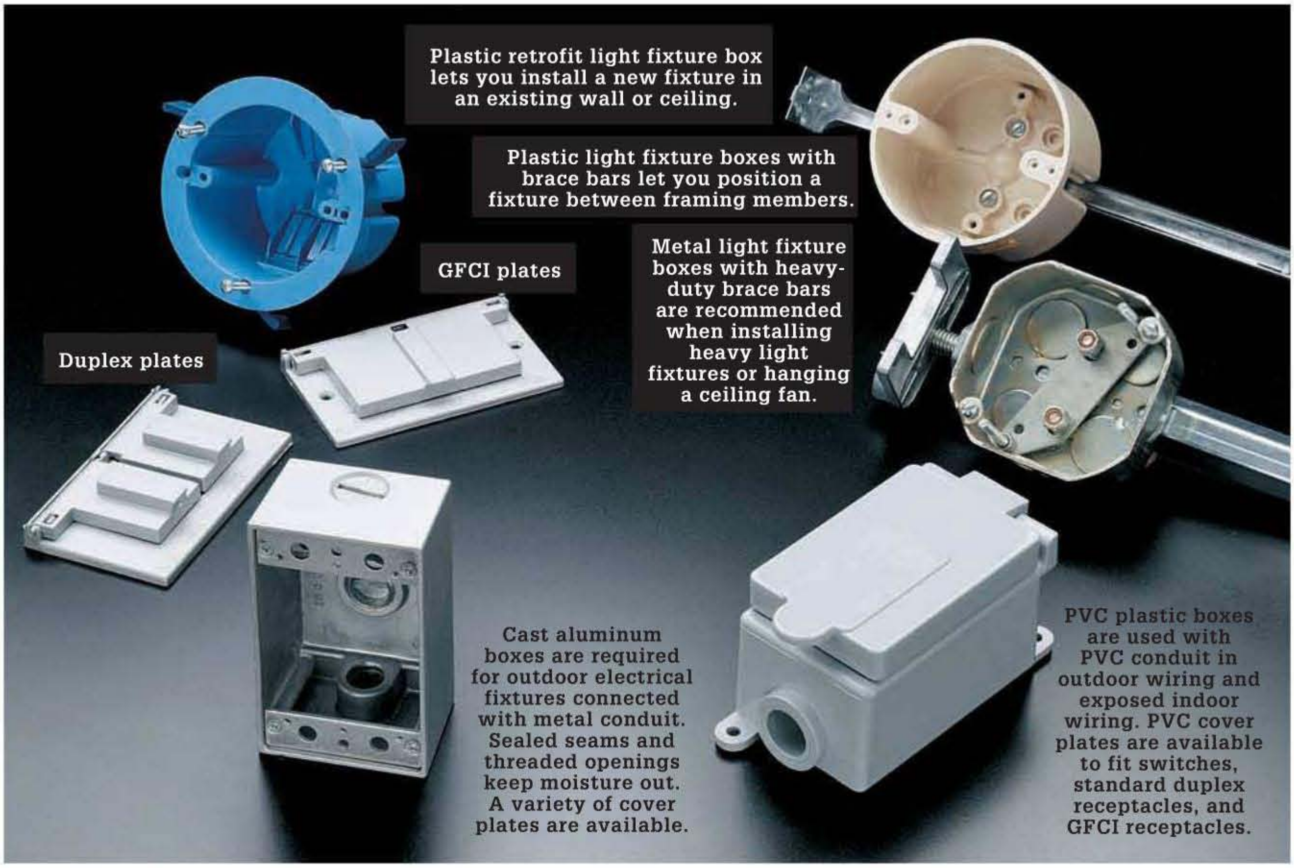

Plastic retrofit light fixture box lets you install a new fixture in an existing wall or ceiling.

Plastic light fixture boxes with brace bars let you position a fixture between framing members.

GFCI plates

Metal light fixture boxes with heavy-duty brace bars are recommended when installing heavy light fixtures or hanging a ceiling fan.

Duplex plates

Cast aluminum boxes are required for outdoor electrical fixtures connected with metal conduit. Sealed seams and threaded openings keep moisture out. A variety of cover plates are available.

PVC plastic boxes are used with PVC conduit in outdoor wiring and exposed indoor wiring. PVC cover plates are available to fit switches, standard duplex receptacles, and GFCI receptacles.

How to Locate Electrical Boxes

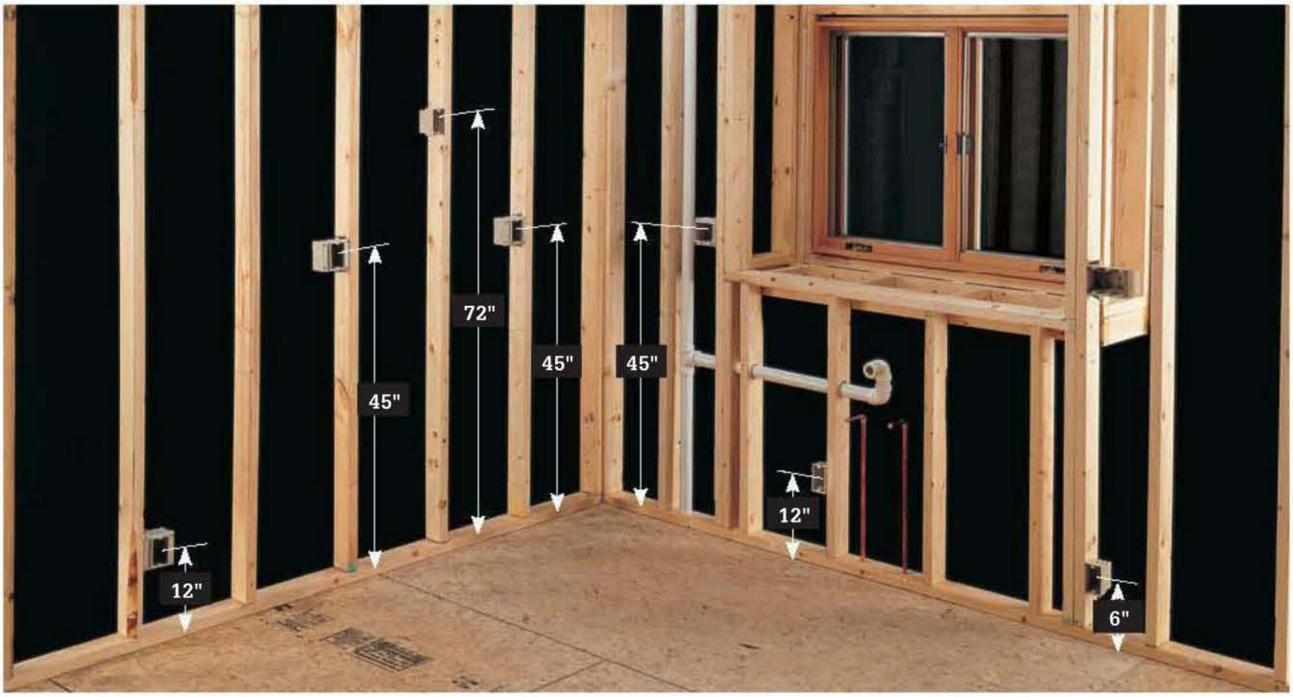

Heights of electrical boxes vary depending on use. In the kitchen shown here, boxes above the countertop are 45" above the floor, in the center of 18" backsplashes that extend from the countertop to the cabinets. All boxes for wall switches also are installed at this height. The center of the box for the microwave receptacle is 72" off the floor. The centers of the boxes for the range and food disposer receptacles are 12" off the floor, but the center of the box for the dishwasher receptacle is 6" off the floor.

Typical Wallcovering Thickness

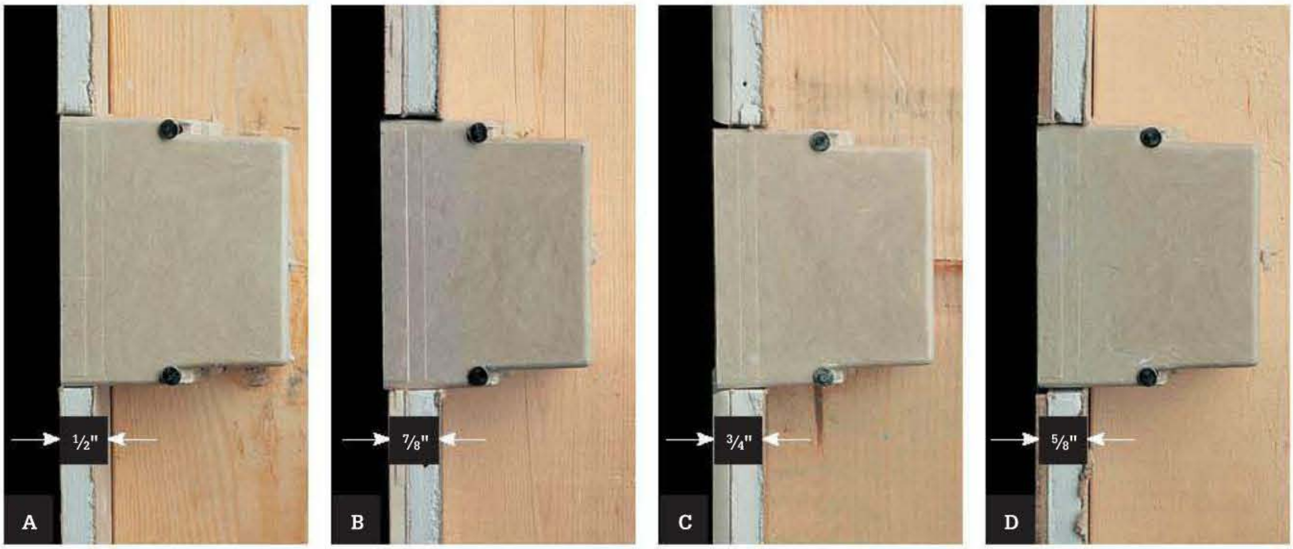

Consider the thickness of finished walls when mounting electrical boxes against framing members. Code requires that the front face of boxes be flush with the finished wall surface, so how you install boxes will vary depending on the type of wall finish that will be used. For example, if the walls will be finished with ½" wallboard (A), attach the boxes so the front faces extend ½" past the front of the framing members. With ceramic tile and wall board (B), extend the boxes ⅞" past the framing members. With ¼" Corian® over wallboard (C), boxes should extend ¾"; and with wallboard and laminate (D), boxes extend ⅝".

How to Install Electrical Boxes for Receptacles

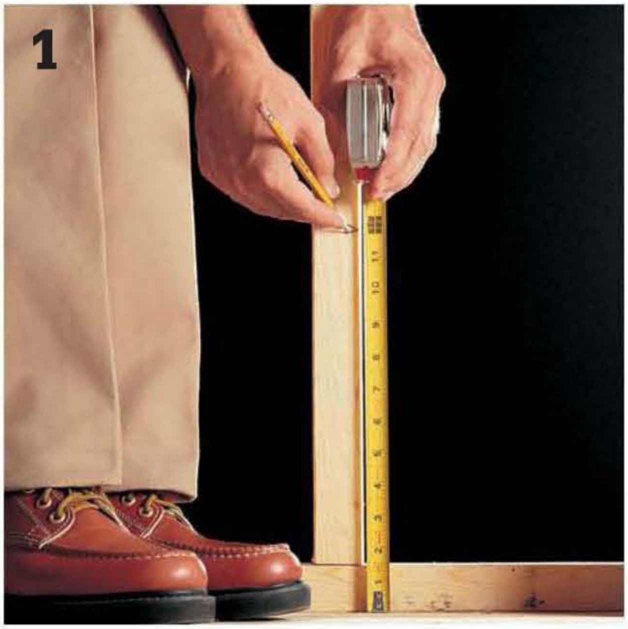

Mark the location of each box on studs. Standard receptacle boxes should be centered 12" above floor level. GFCI receptacle boxes in a bathroom should be mounted so they will be about 10" above the finished countertop.

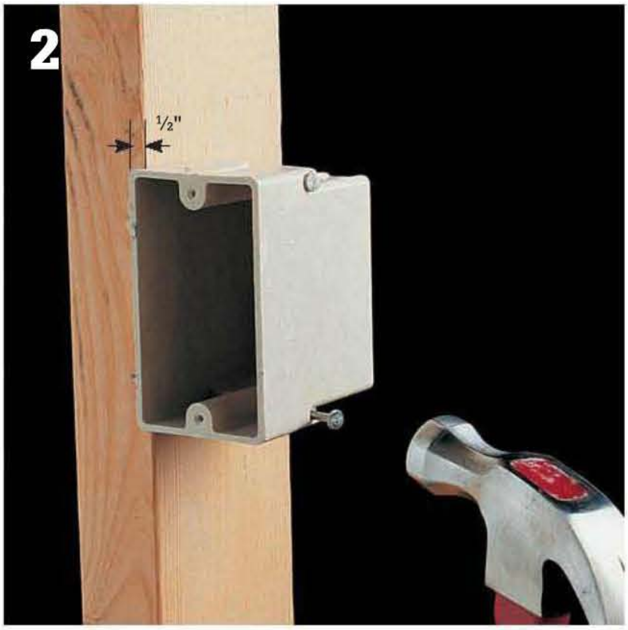

Position each box against a stud so the front face will be flush with the finished wall. For example, if you will be installing ½" wallboard, position the box so it extends ½" past the face of the stud. Anchor the box by driving the mounting nails into the stud.

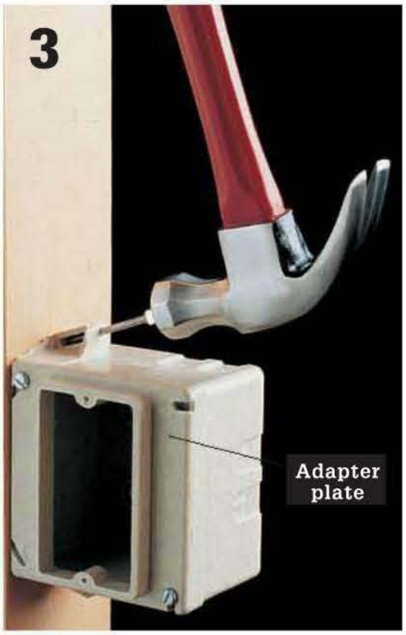

If installing square boxes, attach the adapter plates before positioning the boxes. Use adapter plates that match the thickness of the finished wall. Anchor the box by driving the mounting nails into the stud.

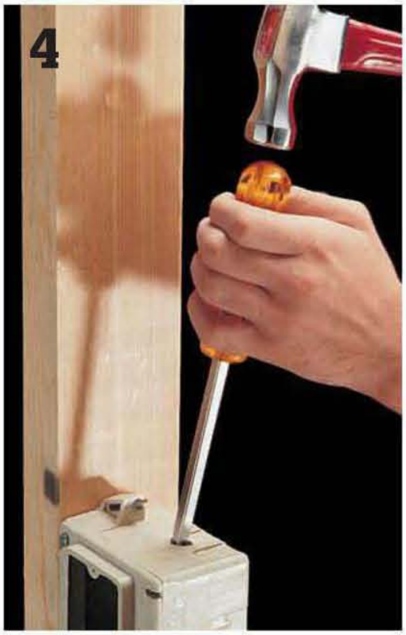

Open one knockout for each cable that will enter the box using a hammer and screwdriver.

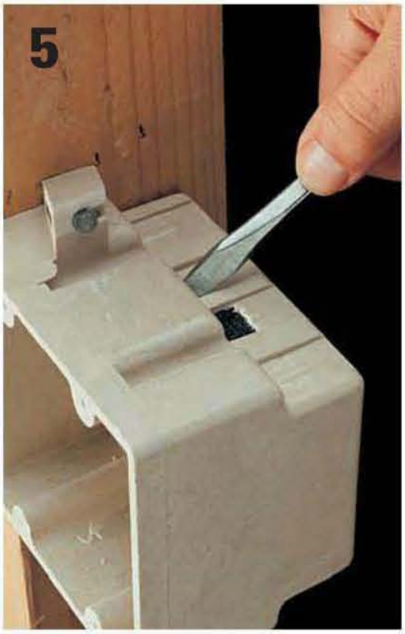

Break off any sharp edges that might damage vinyl cable sheathing by rotating a screwdriver in the knockout.

How to Install Electrical Boxes for Light Fixtures

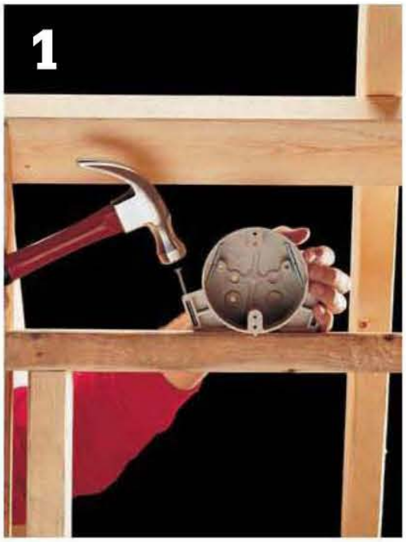

Position the light fixture box for a vanity light above the frame opening for a mirror or medicine cabinet. Place the box for a ceiling light fixture in the center of the room. Position each box against a framing member so the front face will be flush with the finished wall or ceiling, then anchor the box by driving the mounting nails into the framing.

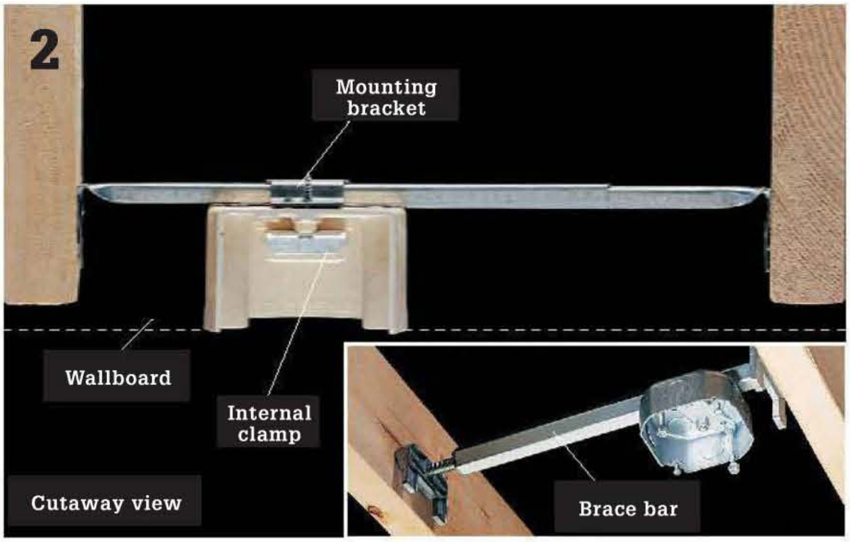

To position a light fixture between joists, attach an electrical box to an adjustable brace bar. Nail the ends of the brace bar to joists so the face of the box will be flush with the finished ceiling surface. Slide the box along the brace bar to the desired position, then tighten the mounting screws. Use internal cable clamps when using a box with a brace bar. *Note: For ceiling fans and heavy fixtures, use a metal box and a heavy-duty brace bar rated for heavy loads (inset photo).*

How to Install Electrical Boxes for Switches

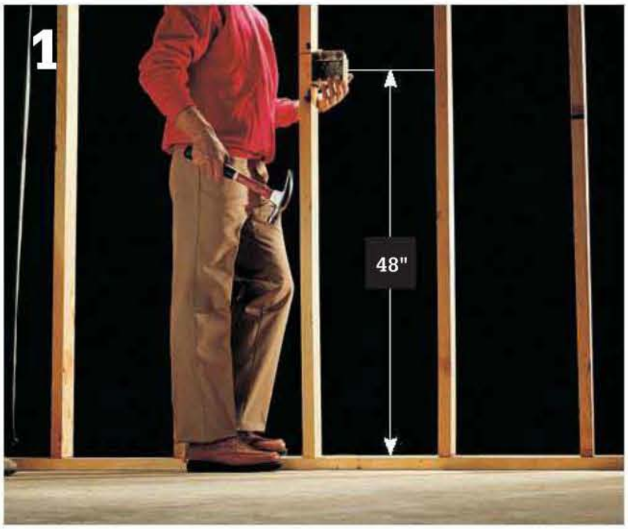

Install switch boxes at accessible locations, usually on the latch side of a door, with the center of the box 48" from the floor. The box for a thermostat is mounted at 48 to 60". Position each box against the side of a stud so the front face will be flush with the finished wall, and drive the mounting nails into the stud.

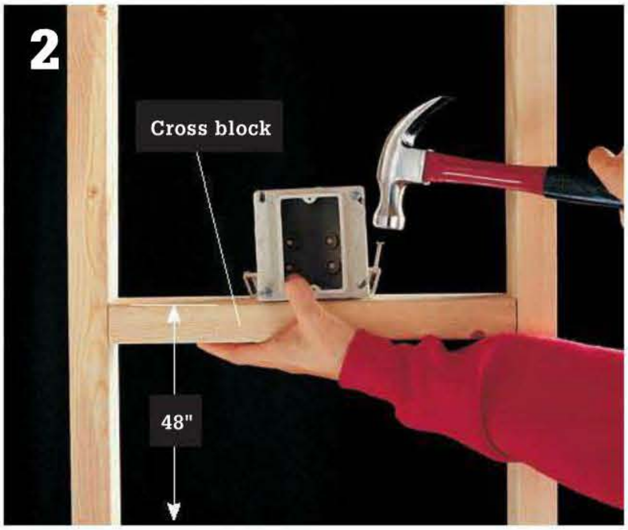

To install a switch box between studs, first install a cross block between studs, with the top edge 46" above the floor. Position the box on the cross block so the front face will be flush with the finished wall, and drive the mounting nails into the cross block.

Ceiling Boxes

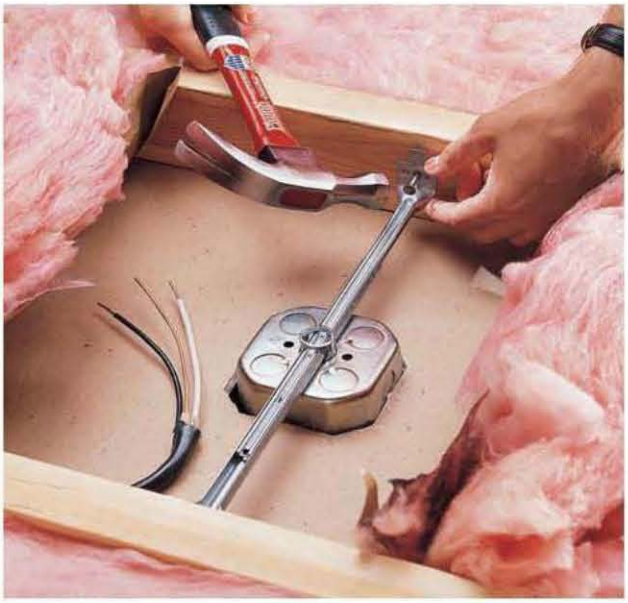

Ceiling boxes for lights are generally round or octagonal in shape to fit typical lamp mounting plates. The easiest way to install one is by nailing the brace to open ceiling joists from above. If the ceiling is insulated, pull the insulation away from the box if the fixture you're installing is not rated IC for insulation contact.

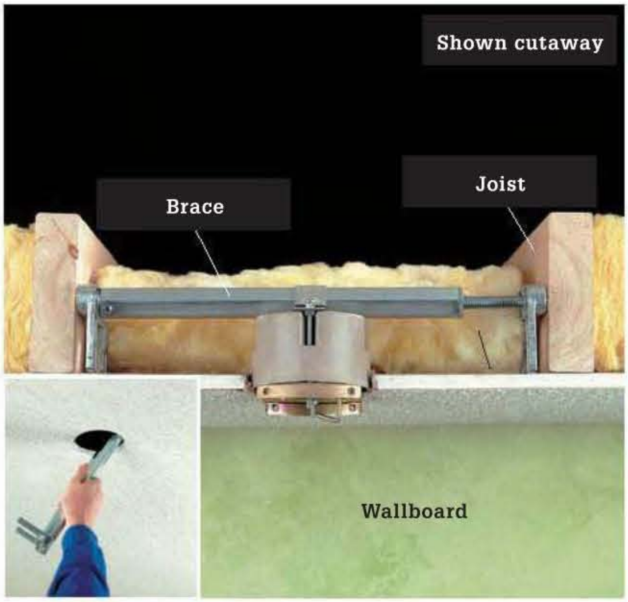

A heavy-duty brace is required for anchoring boxes that will support heavy chandeliers and ceiling fans. A remodeling brace like the one seen here is designed to install through a small cutout in the ceiling (inset photo).

How to Install a Junction Box

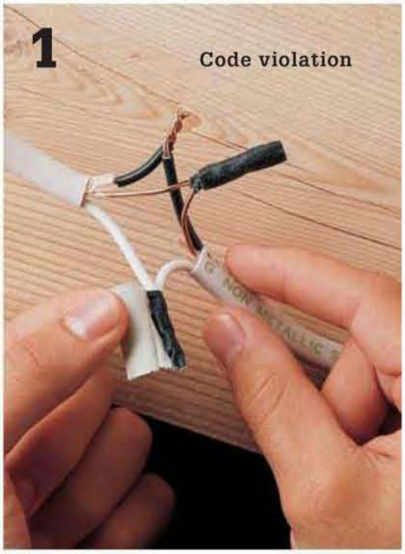

Turn off power to circuit wires at the main service panel. Carefully remove any tape or wire connectors from the exposed splice. Avoid contact with the bare wire ends until the wires have been tested for power.

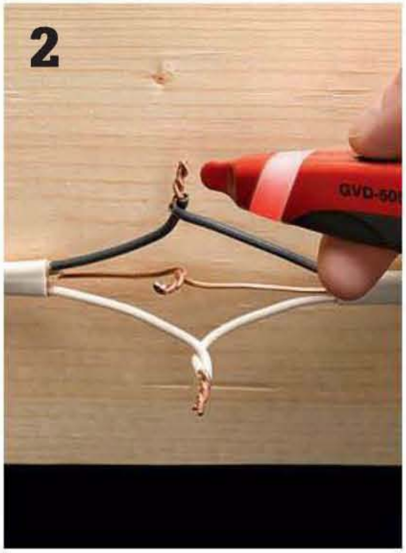

Test for power. The tester should not glow. If it does, the wires are still hot. Shut off power to correct circuit at the main service panel. Disconnect the illegally spliced wires.

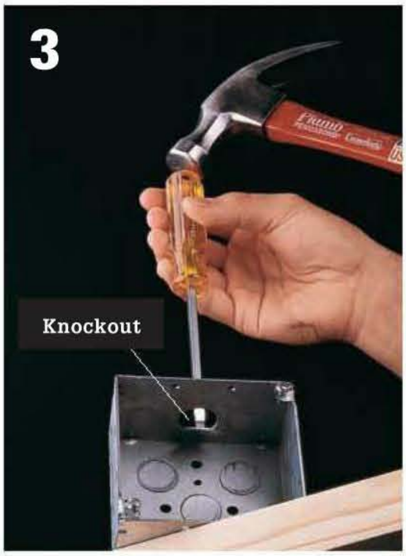

Open one knockout for each cable that will enter the box using a hammer and screwdriver. Any unopened knockouts should remain sealed.

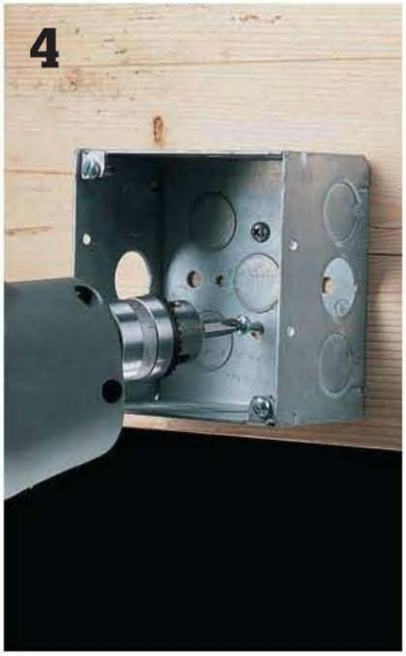

Anchor the electrical box to a wooden framing member using screws or nails.

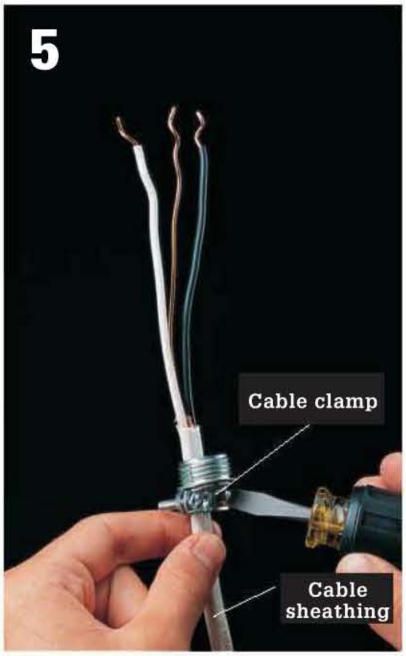
5
Cable clamp
Cable sheathing

Thread each cable end through a cable clamp. Tighten the clamp with a screwdriver. See if there is any slack in the cables so you can gain a little extra cable to work with.

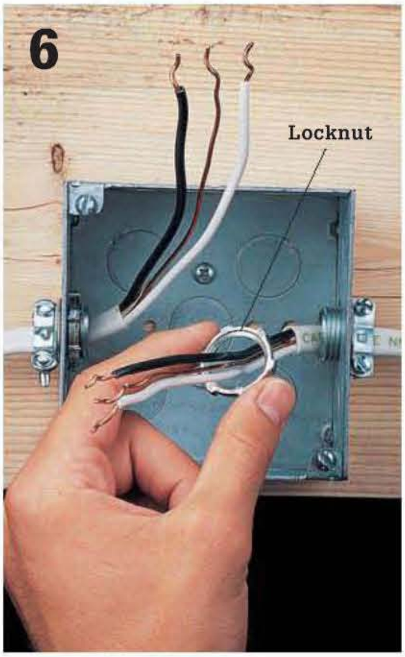

Insert the cables into the electrical box, and screw a locknut onto each cable clamp.

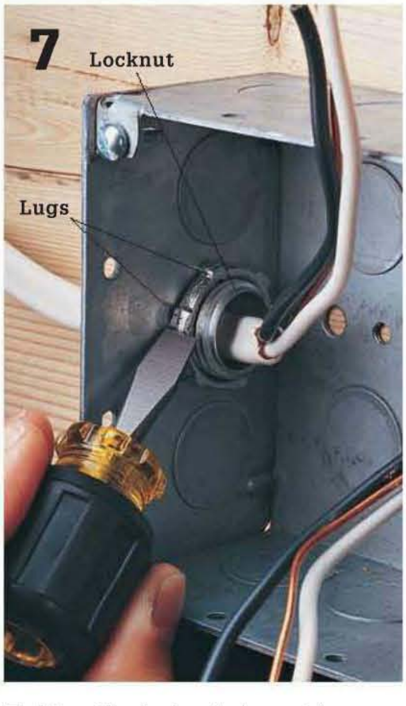
7
Locknut
Lugs

Tighten the locknuts by pushing against the lugs with the blade of a screwdriver.

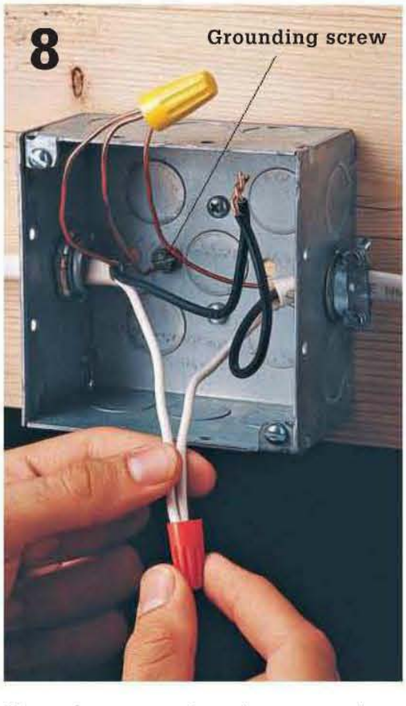
8
Grounding screw

Use wire connectors to reconnect the wires. Pigtail the copper grounding wires to the green grounding screw in the back of the box.

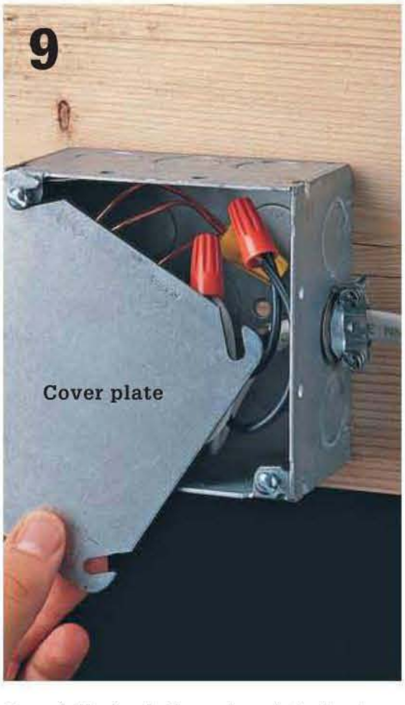
9
Cover plate

Carefully tuck the wires into the box, and attach the cover plate. Turn on the power to the circuit at the main service panel. Make sure the box remains accessible and is not concealed by finished walls or ceilings.

Working with Switches

Wall switches are available in three general types. To repair or replace a switch, it is important to identify its type.

Single-pole switches are used to control a set of lights from one location. Three-way switches are used to control a set of lights from two different locations and are always installed in pairs. Four-way switches are used in combination with a pair of three-way switches to control a set of lights from three or more locations.

Identify switch types by counting the screw terminals. Single-pole switches have two screw terminals, three-way switches have three screw terminals, and four-way switches have four. Most switches include a grounding screw terminal, which is identified by its green color.

When replacing a switch, choose a new switch that has the same number of screw terminals as the old one. The location of the screws on the switch body varies depending on the manufacturer, but these differences will not affect the switch operation.

Whenever possible, connect switches using the screw terminals rather than push-in fittings. Some specialty switches have wire leads instead of screw terminals. They are connected to circuit wires with wire connectors.

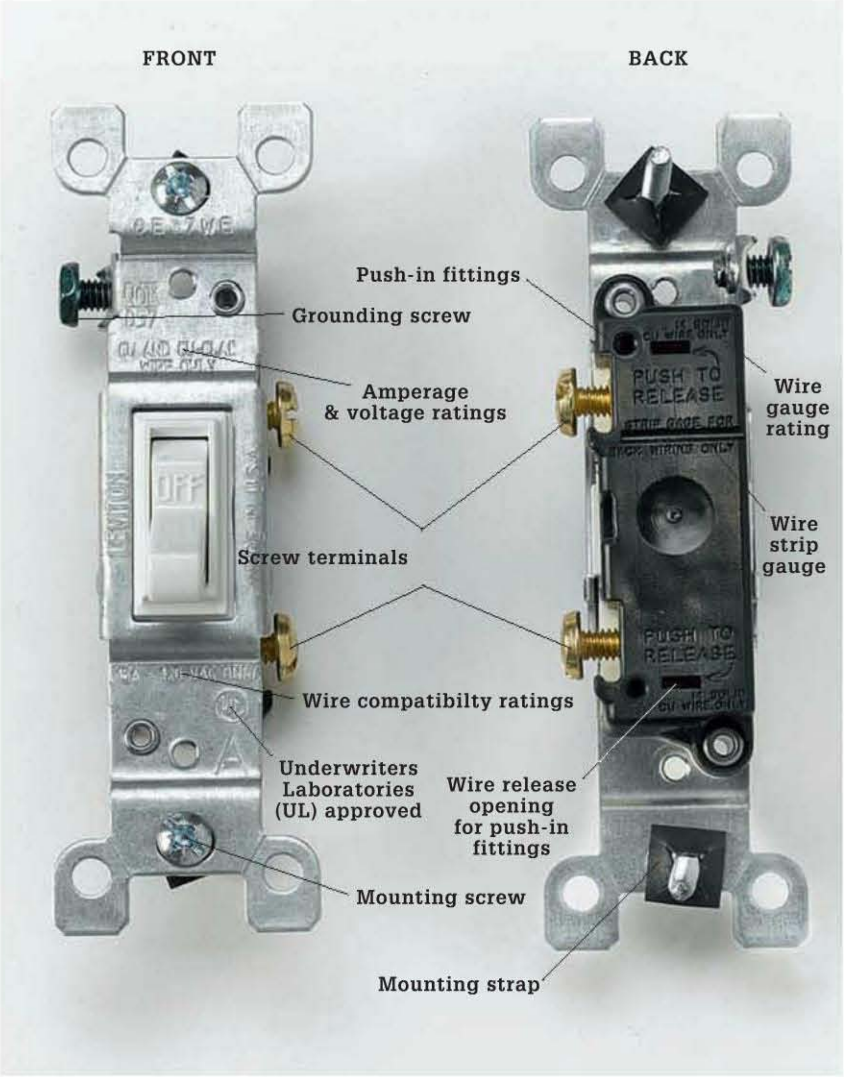
FRONT
BACK
Push-in fittings
Grounding screw
Amperage & voltage ratings
Wire gauge rating
Screw terminals
Wire strip gauge
Wire compatibilty ratings
Underwriters Laboratories (UL) approved
Wire release opening for push-in fittings
Mounting screw
Mounting strap

A wall switch is connected to circuit wires with screw terminals or with push-in fittings on the back of the switch. A switch may have a stamped strip gauge that indicates how much insulation must be stripped from the circuit wires to make the connections.

The switch body is attached to a metal mounting strap that allows it to be mounted in an electrical box. Several rating stamps are found on the strap and on the back of the switch. The abbreviation UL or UND. LAB. INC. LIST means that the switch meets the safety standards of the Underwriters Laboratories. Switches also are stamped with maximum voltage and amperage ratings. Standard wall switches are rated 15A or 125V. Voltage ratings of 110, 120, and 125 are considered to be identical for purposes of identification.

For standard wall switch installations, choose a switch that has a wire gauge rating of #12 or #14. For wire systems with solid-core copper wiring, use only switches marked COPPER or CU. For aluminum wiring, use only switches marked CO/ALR. Switches marked AL/CU can no longer be used with aluminum wiring, according to the National Electrical Code.

Single-pole Wall Switches

A single-pole switch is the most common type of wall switch. It has ON-OFF markings on the switch lever and is used to control a set of lights, an appliance, or a receptacle from a single location. A single-pole switch has two screw terminals and a grounding screw. When installing a single-pole switch, check to make sure the ON marking shows when the switch lever is in the up position.

In a correctly wired single-pole switch, a hot circuit wire is attached to each screw terminal. However, the color and number of wires inside the switch box will vary, depending on the location of the switch along the electrical circuit.

If two cables enter the box, then the switch lies in the middle of the circuit. In this installation, both of the hot wires attached to the switch are black.

If only one cable enters the box, then the switch lies at the end of the circuit. In this installation (sometimes called a switch loop), one of the hot wires is black, but the other hot wire usually is white. A white hot wire sometimes is coded with black tape or paint.

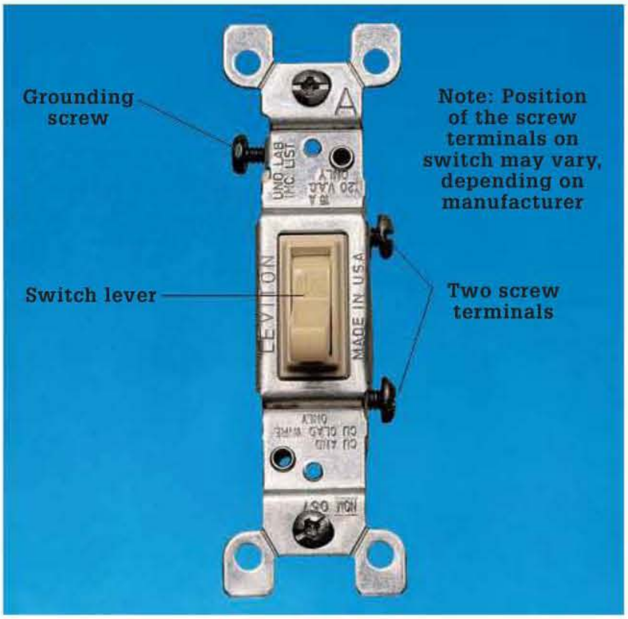
Grounding screw
Note: Position of the screw terminals on switch may vary, depending on manufacturer
Switch lever
Two screw terminals

A single-pole switch is essentially an interruption in the black power supply wire that is opened or closed with the toggle. Single-pole switches are the simplest of all home wiring switches.

Typical Single-pole Switch Installations

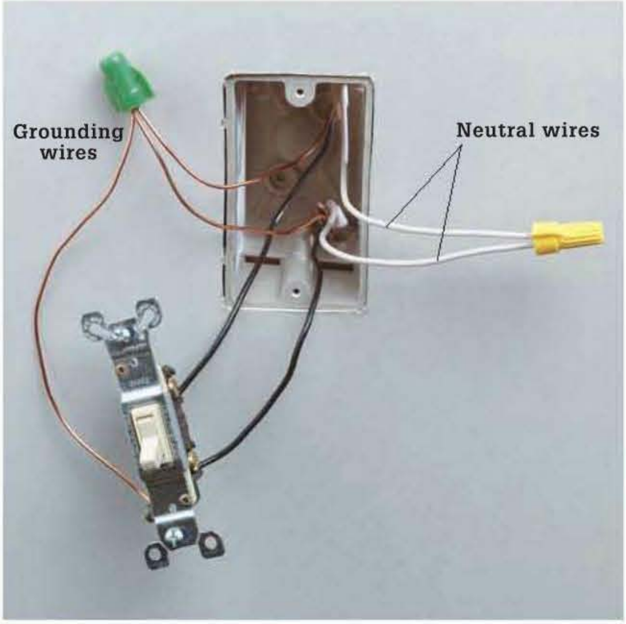

Two cables enter the box when a switch is located in the middle of a circuit. Each cable has a white and a black insulated wire, plus a bare copper grounding wire. The black wires are hot and are connected to the screw terminals on the switch. The white wires are neutral and are joined together with a wire connector. Grounding wires are pigtailed to the switch.

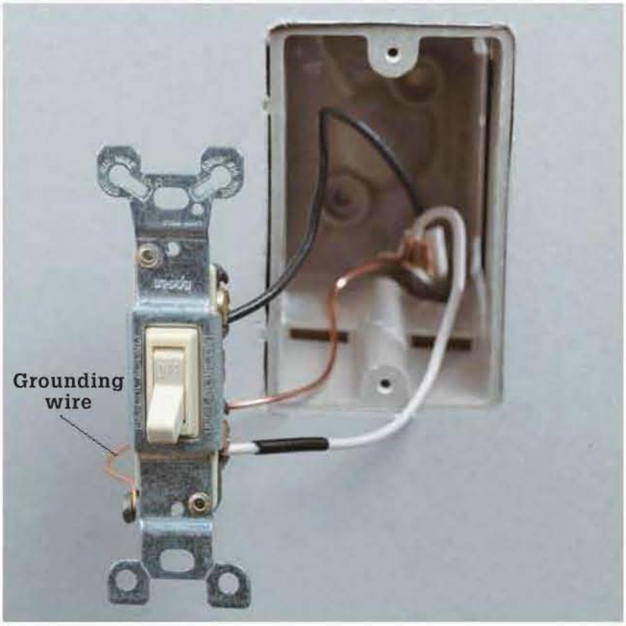

One cable enters the box when a switch is located at the end of a circuit. The cable has a white and a black insulated wire, plus a bare copper grounding wire. In this installation, both of the insulated wires are hot. The white wire may be labeled with black tape or paint to identify it as a hot wire. The grounding wire is connected to the switch grounding screw.

Three-way Wall Switches

Three-way switches have three screw terminals and do not have ON-OFF markings. Three-way switches are always installed in pairs and are used to control a set of lights from two locations.

One of the screw terminals on a three-way switch is darker than the others. This screw is the common screw terminal. The position of the common screw terminal on the switch body may vary, depending on the manufacturer. Before disconnecting a three-way switch, always label the wire that is connected to the common screw terminal. It must be reconnected to the common screw terminal on the new switch.

The two lighter-colored screw terminals on a three-way switch are called the traveler screw terminals. The traveler terminals are interchangeable, so there is no need to label the wires attached to them.

Because three-way switches are installed in pairs, it sometimes is difficult to determine which of the switches is causing a problem. The switch that receives greater use is more likely to fail, but you may need to inspect both switches to find the source of the problem.

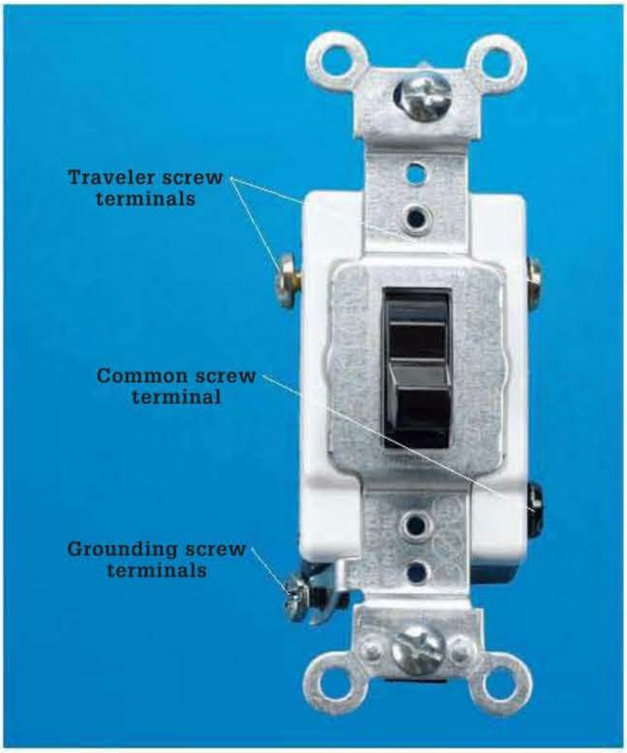
Traveler screw terminals
Common screw terminal
Grounding screw terminals

Typical Three-way Switch Installations

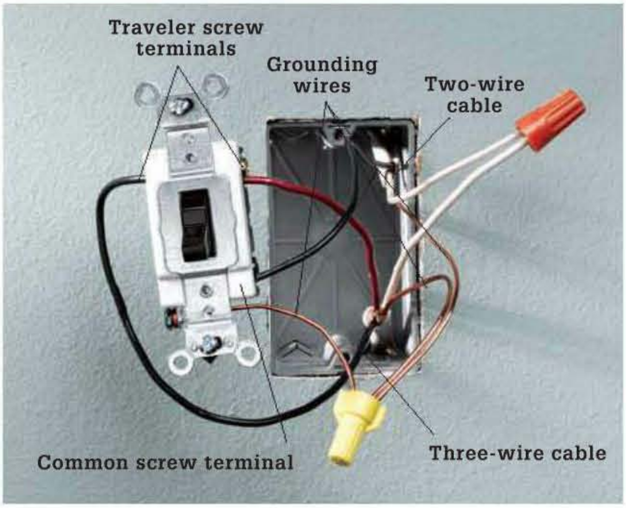
Traveler screw terminals
Grounding wires
Two-wire cable
Common screw terminal
Three-wire cable

Two cables enter the box if the switch lies in the middle of a circuit. One cable has two wires, plus a bare copper grounding wire; the other cable has three wires, plus a ground. The black wire from the two-wire cable is connected to the dark common screw terminal. The red and black wires from the three-wire cable are connected to the traveler screw terminals. The white neutral wires are joined together with a wire connector, and the grounding wires are pigtailed to the grounded metal box.

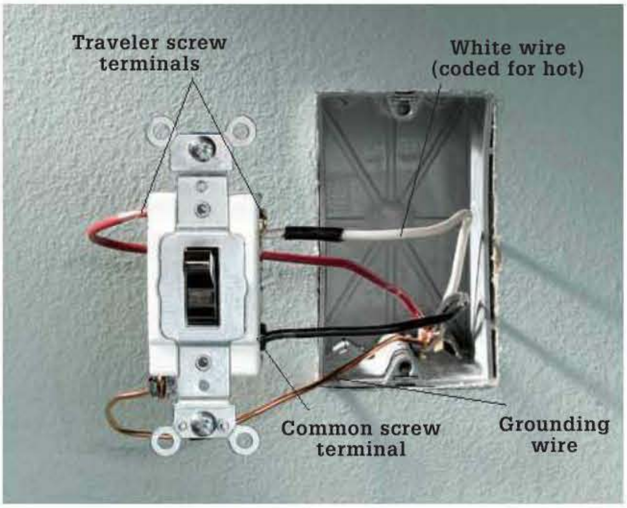
Traveler screw terminals
White wire (coded for hot)
Common screw terminal
Grounding wire

One cable enters the box if the switch lies at the end of the circuit. The cable has a black wire, red wire, and white wire, plus a bare copper grounding wire. The black wire must be connected to the common screw terminal, which is darker than the other two screw terminals. The white and red wires are connected to the two traveler screw terminals. The white wire is taped to indicate that it is hot. The bare copper grounding wire is connected to the grounded metal box.

Four-way Wall Switches

Four-way switches have four screw terminals and do not have ON-OFF markings. Four-way switches are always installed between a pair of three-way switches. This switch combination makes it possible to control a set of lights from three or more locations. Four-way switches are common in homes where large rooms contain multiple living areas, such as a kitchen opening into a dining room. Switch problems in a four-way installation can be caused by loose connections or worn parts in a four-way switch or in one of the three-way switches (facing page).

In a typical installation, there will be a pair of three-way cables that enter the box for the four-way switch. With most switches, the white and red wires from one cable should be attached to the bottom or top pair of screw terminals, and the white and red wires from the other cable should be attached to the remaining pair of screw terminals. However, not all switches are configured the same way, and wiring configurations in the box may vary, so always study the wiring diagram that comes with the switch.

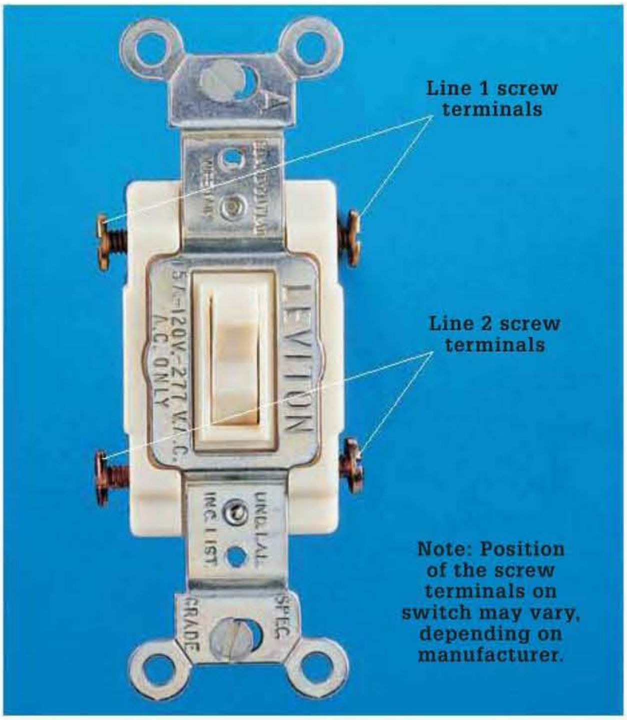

Note: Position of the screw terminals on switch may vary, depending on manufacturer.

Typical Four-way Switch Installation

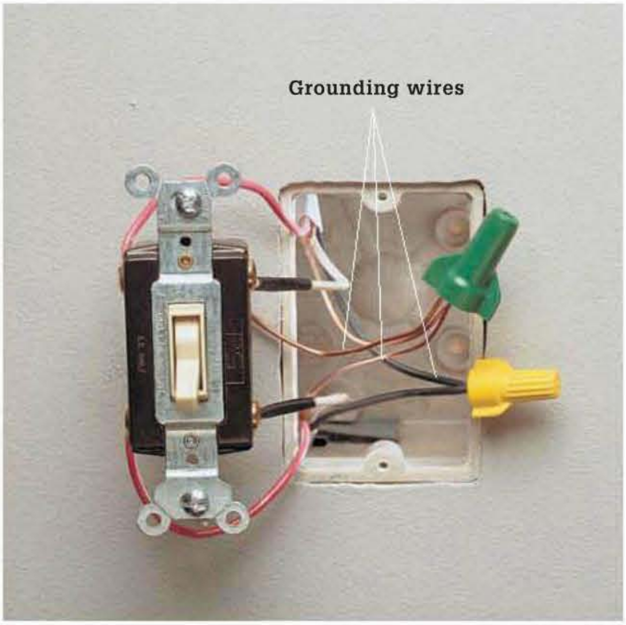

Four wires are connected to a four-way switch. The red and white wires from one cable are attached to the top pair of screw terminals, while the red and white wires from the other cable are attached to the bottom screw terminals.

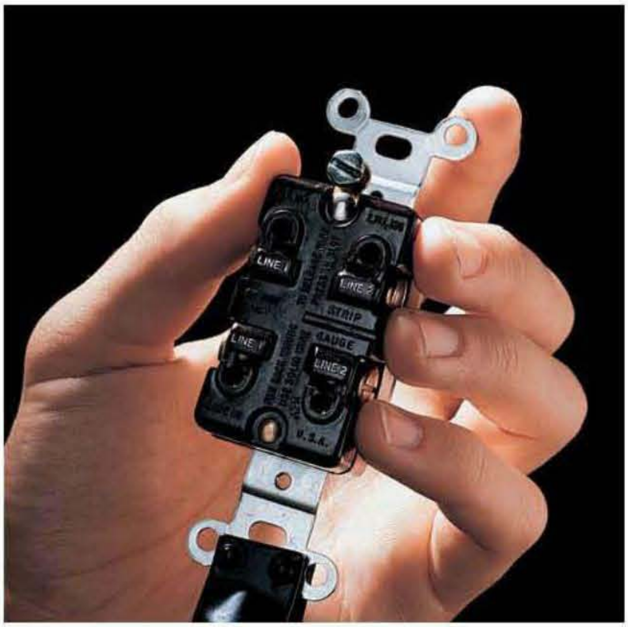

Switch variation: Some four-way switches have a wiring guide stamped on the back to help simplify installation. For the switch shown above, one pair of color-matched circuit wires will be connected to the screw terminals marked LINE 1, while the other pair of wires will be attached to the screw terminals marked LINE 2.

Testing Switches

A switch that does not work properly may have worn or broken internal parts. Test switches with a battery-operated continuity tester. The continuity tester detects any break in the metal pathway inside the switch. Replace the switch if the continuity tester shows the switch to be faulty.

Never use a continuity tester on wires that might carry live current. Always shut off the power and disconnect the switch before testing for continuity.

Some specialty switches, like dimmers, cannot be tested for continuity. Electronic switches can be tested for manual operation using a continuity tester, but the automatic operation of these switches cannot be tested.

How to Test a Single-pole Wall Switch

1

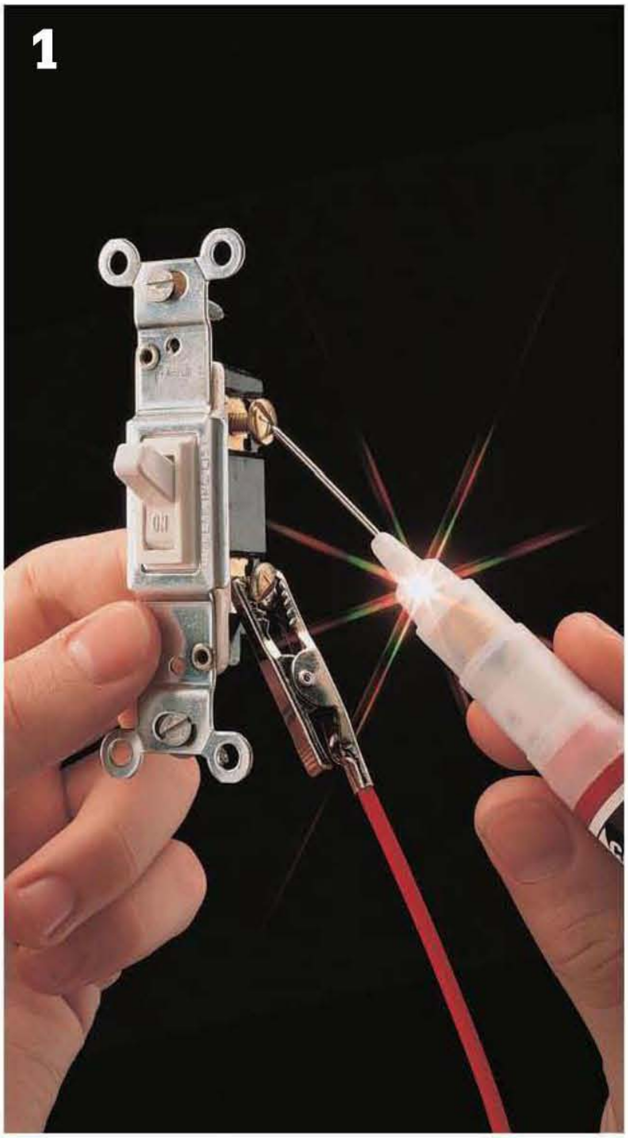

Attach clip of tester to one of the screw terminals. Touch the tester probe to the other screw terminal. Flip switch lever from ON to OFF. If switch is good, tester glows when lever is ON, but not when OFF.

2

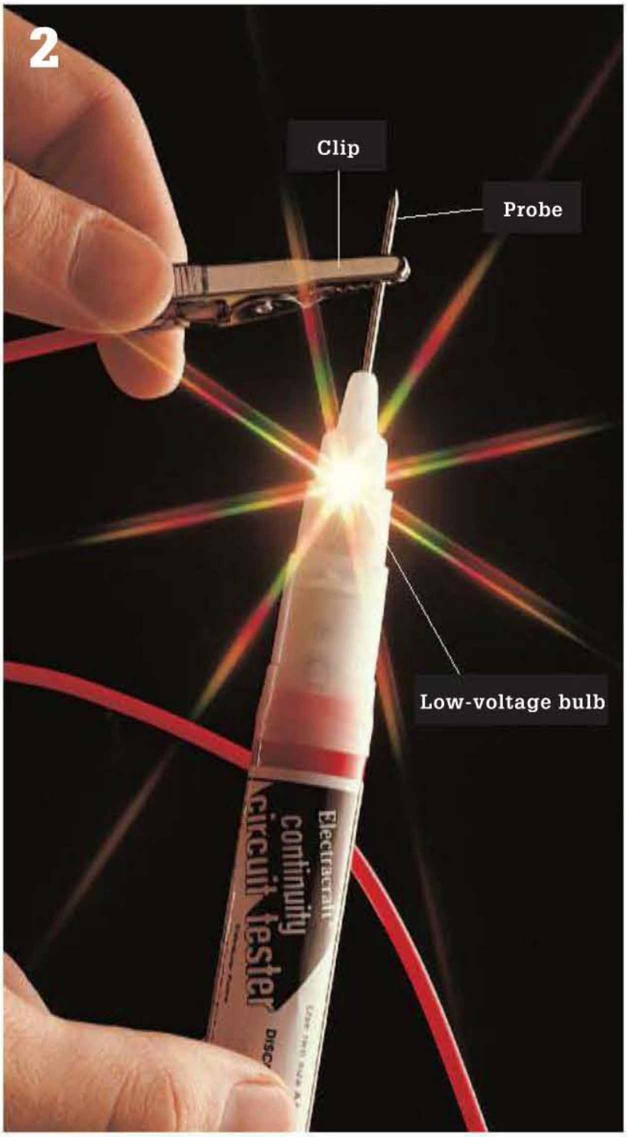

Continuity tester uses battery-generated current to test the metal pathways running through switches and other electrical fixtures. Always "test" the tester before use. Touch the tester clip to the metal probe. The tester should glow. If not, then the battery or lightbulb is dead and must be replaced.

How to Test a Three-way Wall Switch

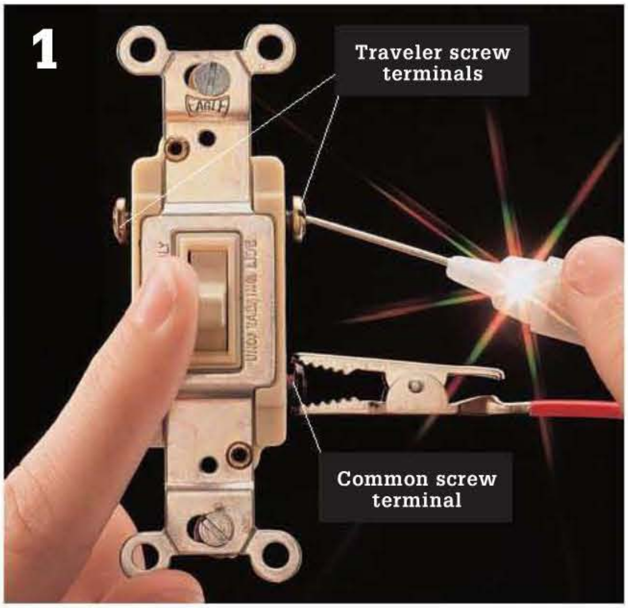

Attach tester clip to the dark common screw terminal. Touch the tester probe to one of the traveler screw terminals, and flip switch lever back and forth. If switch is good, the tester should glow when the lever is in one position, but not both.

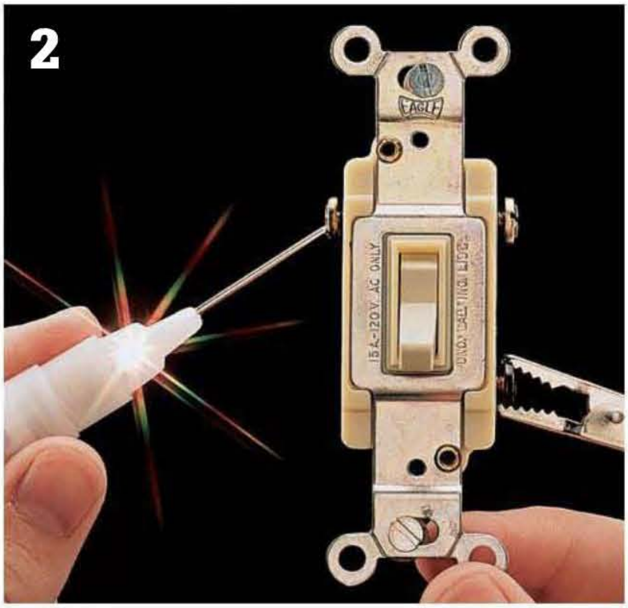

Touch probe to the other traveler screw terminal, and flip the switch lever back and forth. If switch is good, the tester will glow only when the switch lever is in the position opposite from the positive test in step 1.

How to Test a Four-way Wall Switch

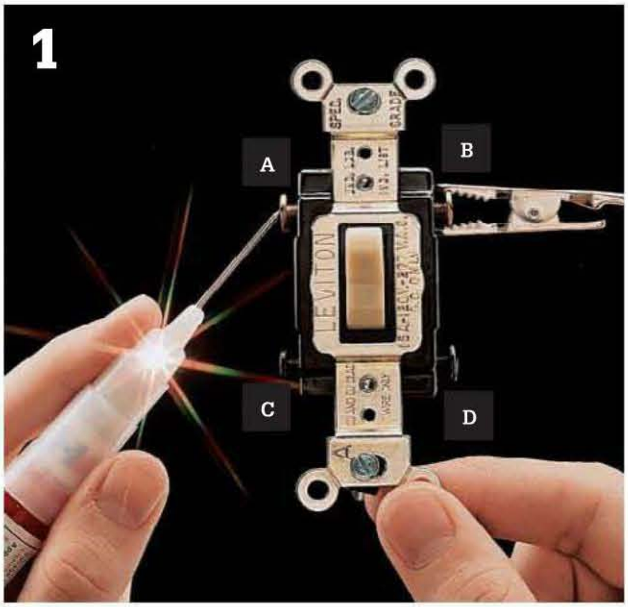

Test switch by touching probe and clip of continuity tester to each pair of screw terminals (A-B, C-D, A-D, B-C, A-C, B-D). The test should show continuous pathways between two different pairs of screw terminals. Flip lever to opposite position, and repeat test. Test should show continuous pathways between two different pairs of screw terminals.

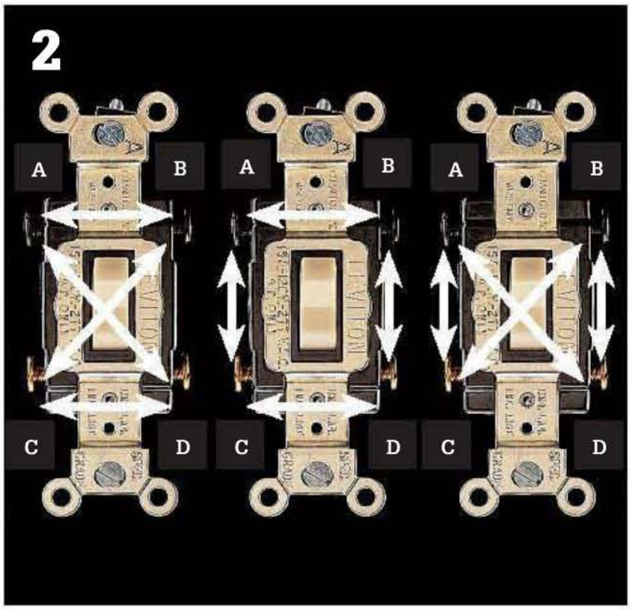

If switch is good, test will show a total of four continuous pathways between screw terminals—two pathways for each lever position. If not, then switch is faulty and must be replaced. (The arrangement of the pathways may differ, depending on the switch manufacturer. The photo above shows the three possible pathway arrangements.)

Working with Receptacles

A 120-volt duplex receptacle can be wired to the electrical system in a number of ways. The most common are shown on these pages.

Wiring configurations may vary slightly from these photographs, depending on the kind of receptacles used, the type of cable, or the technique of the electrician who installed the wiring. To make dependable repairs or replacements, use masking tape and label each wire according to its location on the terminals of the existing receptacle.

Receptacles are wired as either end-of-run or middle-of-run. These two basic configurations are easily identified by counting the number of cables entering the receptacle box. End-of-run wiring has only one cable, indicating that the circuit ends. Middle-of-run wiring has two cables, indicating that the circuit continues on to other receptacles, switches, or fixtures.

A split-circuit receptacle is shown on the next page. Each half of a split-circuit receptacle is wired to a separate circuit. This allows two appliances of high wattage to be plugged into the same receptacle without blowing a fuse or tripping a breaker. This wiring configuration is similar to a receptacle that is controlled by a wall switch. Code requires a switch-controlled receptacle in any room that does not have a built-in light fixture operated by a wall switch.

Split-circuit and switch-controlled receptacles are connected to two hot wires, so use caution during repairs or replacements. Make sure the connecting tab between the hot screw terminals is removed.

Two-slot receptacles are common in older homes. There is no grounding wire attached to the receptacle, but the box may be grounded with armored cable or conduit.

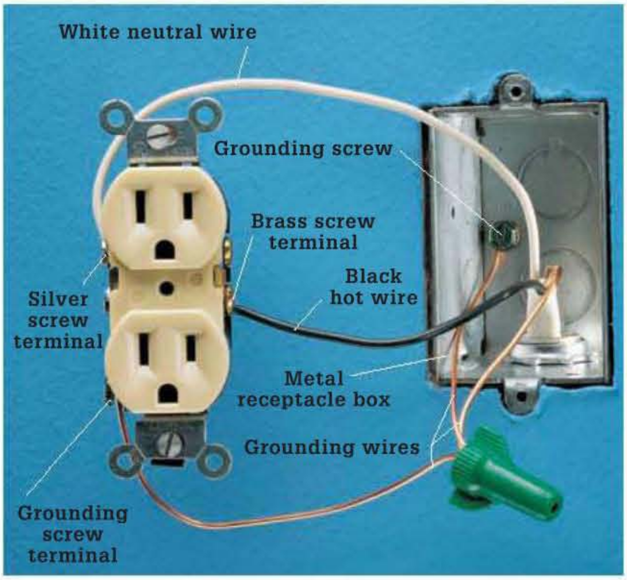
White neutral wire
Grounding screw
Brass screw terminal
Black hot wire
Silver screw terminal
Metal receptacle box
Grounding wires
Grounding screw terminal

Single cable entering the box indicates end-of-run wiring. The black hot wire is attached to a brass screw terminal, and the white neutral wire is connected to a silver screw terminal. If the box is metal, the grounding wire is pigtailed to the grounding screws of the receptacle and the box. In a plastic box, the grounding wire is attached directly to the grounding screw terminal of the receptacle.

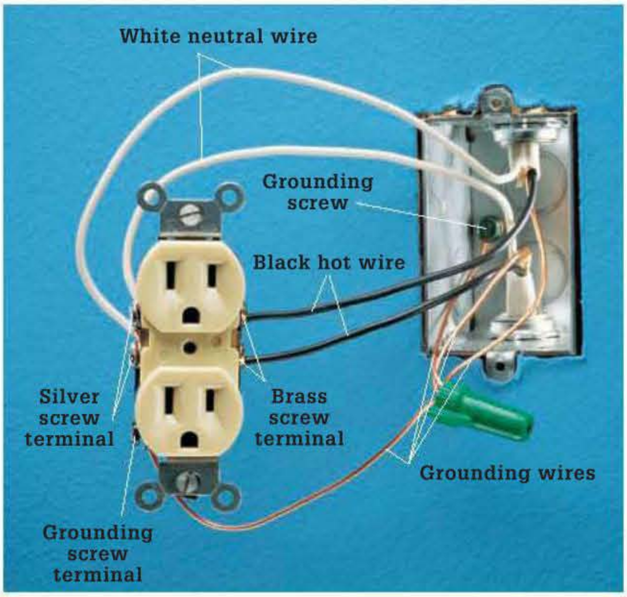
White neutral wire
Grounding screw
Black hot wire
Silver screw terminal
Brass screw terminal
Grounding wires
Grounding screw terminal

Two cables entering the box indicate middle-of-run wiring. Black hot wires are connected to brass screw terminals, and white neutral wires to silver screw terminals. The grounding wire is pigtailed to the grounding screws of the receptacle and the box.

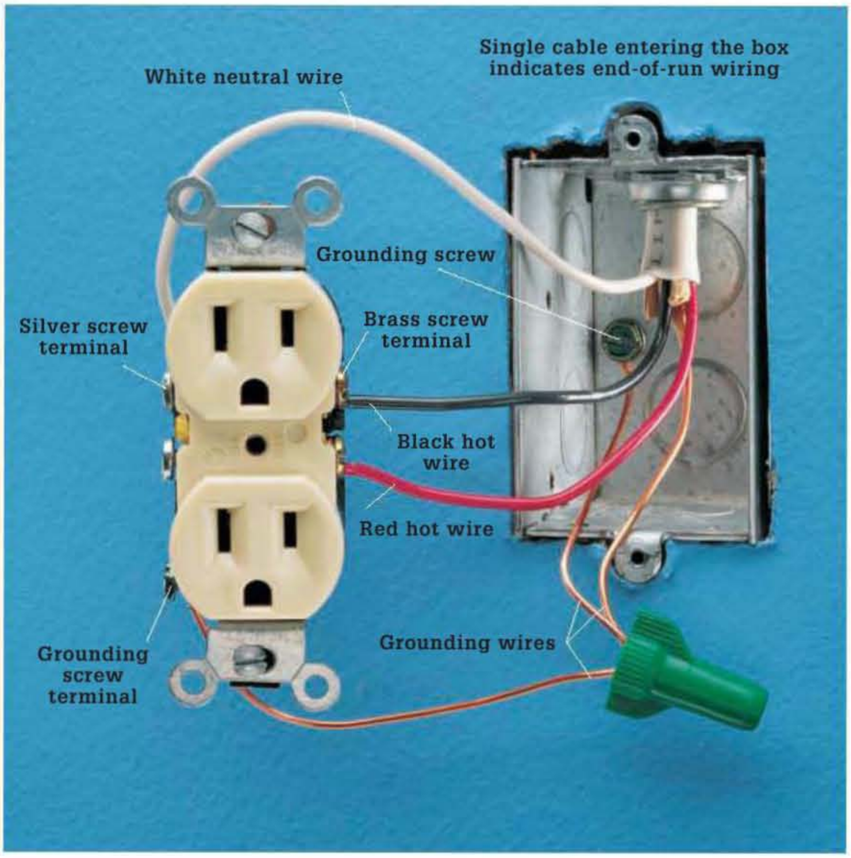
White neutral wire
Single cable entering the box indicates end-of-run wiring
Grounding screw
Silver screw terminal
Brass screw terminal
Black hot wire
Red hot wire
Grounding wires
Grounding screw terminal

A split-circuit receptacle is attached to a black hot wire, a red hot wire, a white neutral wire, and a bare grounding wire. The wiring is similar to a switch-controlled receptacle.
The hot wires are attached to the brass screw terminals, and the connecting tab or fin between the brass terminals is removed. The white wire is attached to a silver screw terminal, and the connecting tab on the neutral side remains intact. The grounding wire is pigtailed to the grounding screw terminal of the receptacle and to the grounding screw attached to the box.

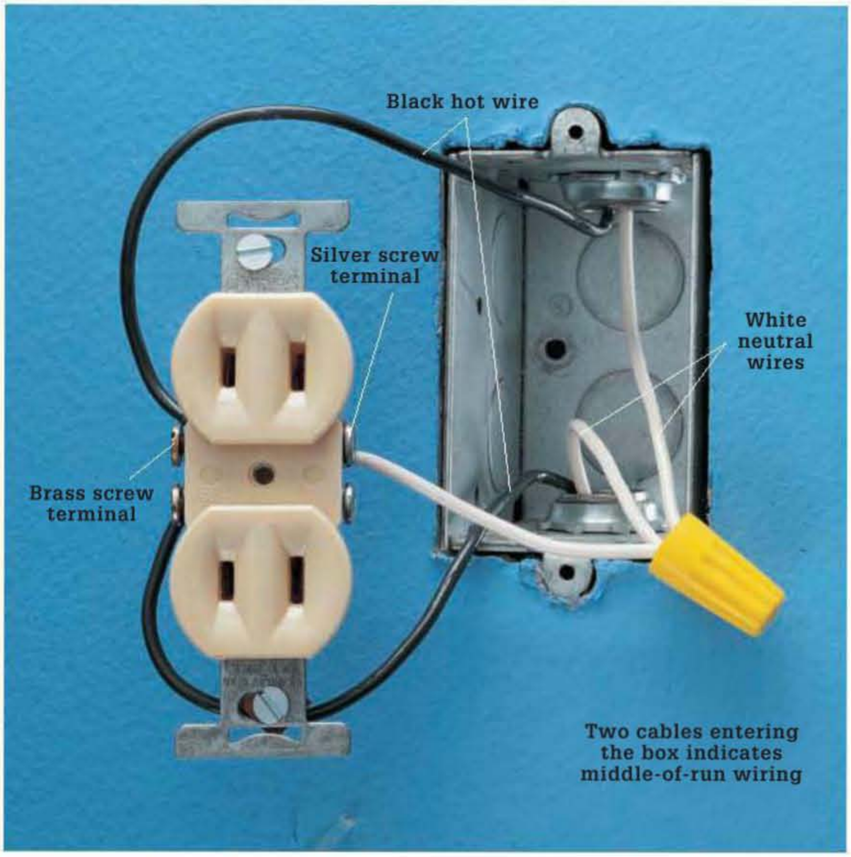
Black hot wire
Silver screw terminal
White neutral wires
Brass screw terminal
Two cables entering the box indicates middle-of-run wiring

A two-slot receptacle is often found in older homes. The black hot wires are connected to the brass screw terminals, and the white neutral wires are pigtailed to a silver screw terminal. Two-slot receptacles may be replaced with three-slot types, but only if a means of grounding exists at the receptacle box. In some municipalities, you may replace a two-slot receptacle with a GFCI receptacle as long as the receptacle has a sticker that reads "No equipment ground."

How to Install a New Receptacle

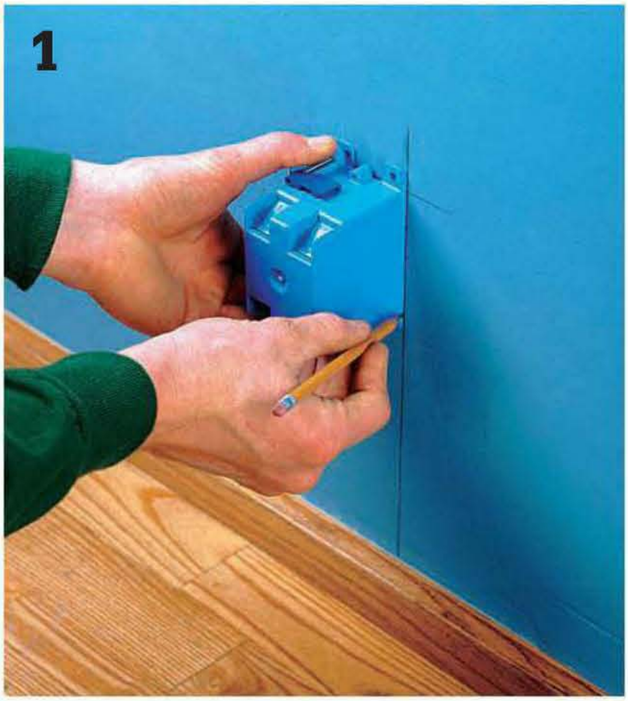

Position the new pop-in box on the wall and trace around it. Consider the location of hidden utilities within the wall before you cut.

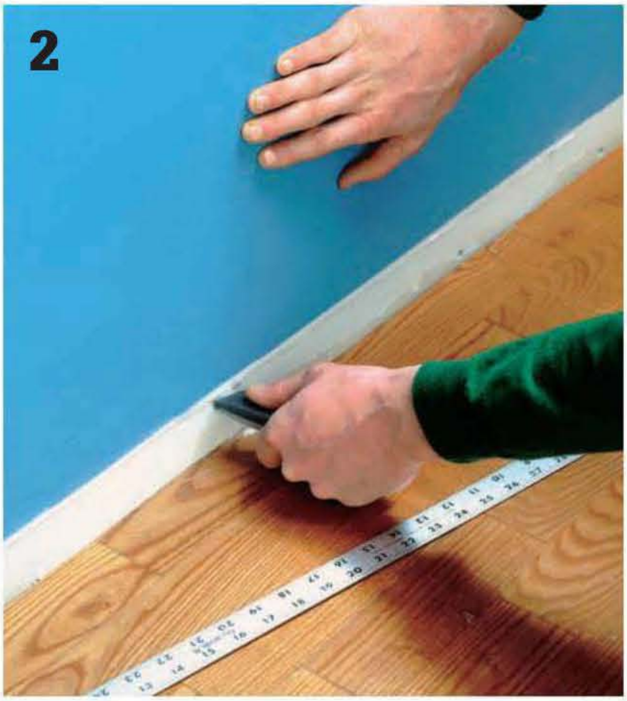

Remove baseboard between new and existing receptacle. Cut away the drywall about 1" below the baseboard with a jigsaw, wallboard saw, or utility knife.

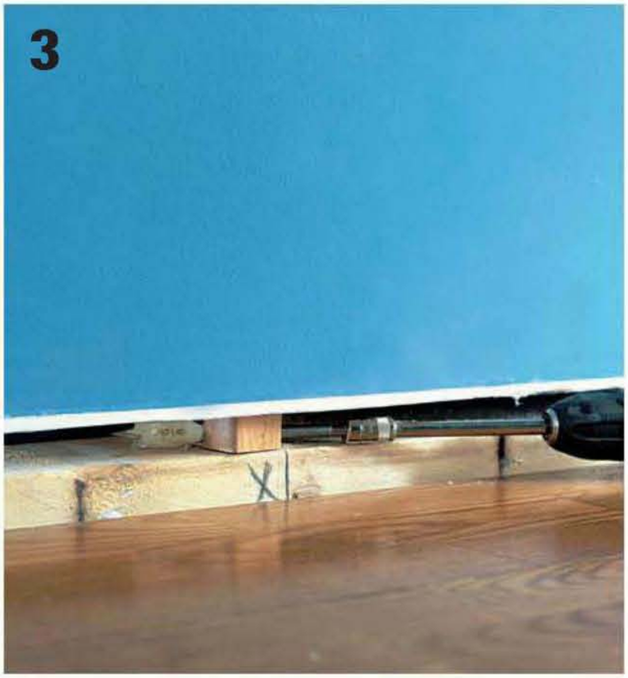

Drill a ⅝" hole in the center of each stud along the opening between the two receptacles. A drill bit extender or a flexible drill bit will allow you a better angle and make drilling the holes easier.

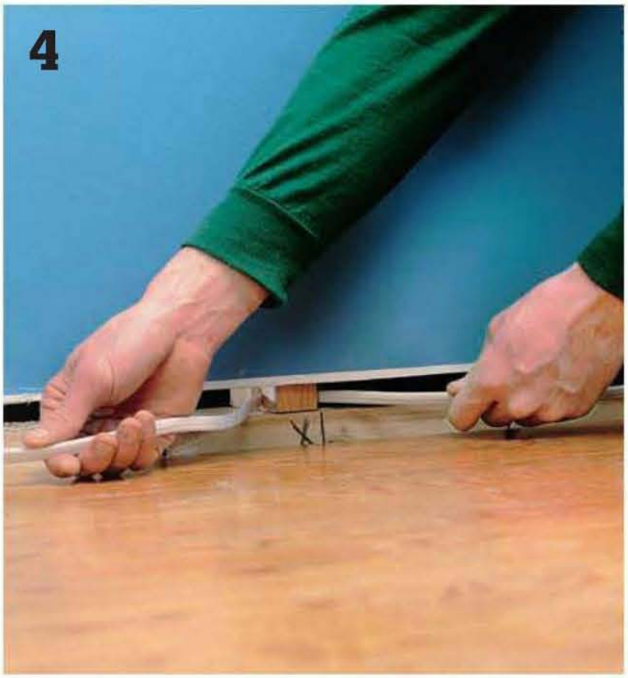

Run the branch cable through the holes from the new location to the existing receptacle. Staple the cable to the stud below the box. Install a metal nail plate on the front edge of each stud that the cable routes through.

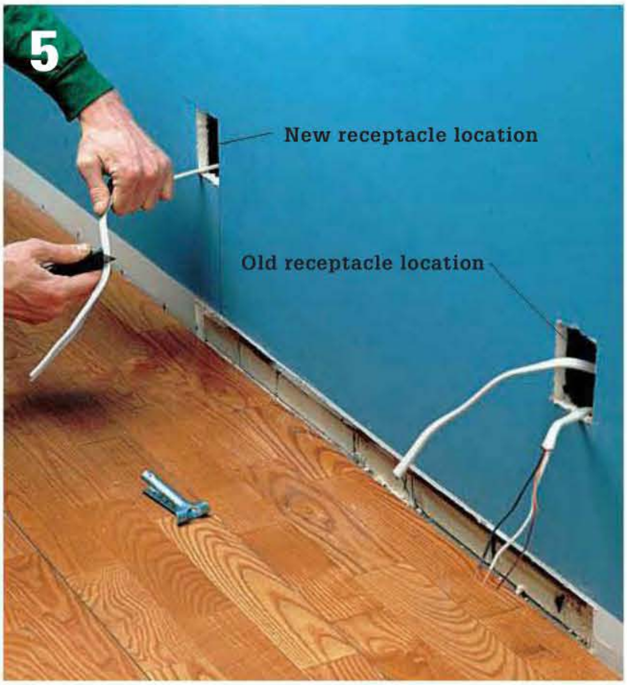

Turn off the power at the main panel and test for power. Remove the old receptacle and its box, and pull the new branch cable up through the hole. Remove sheathing and insulation from both ends of the new cable.

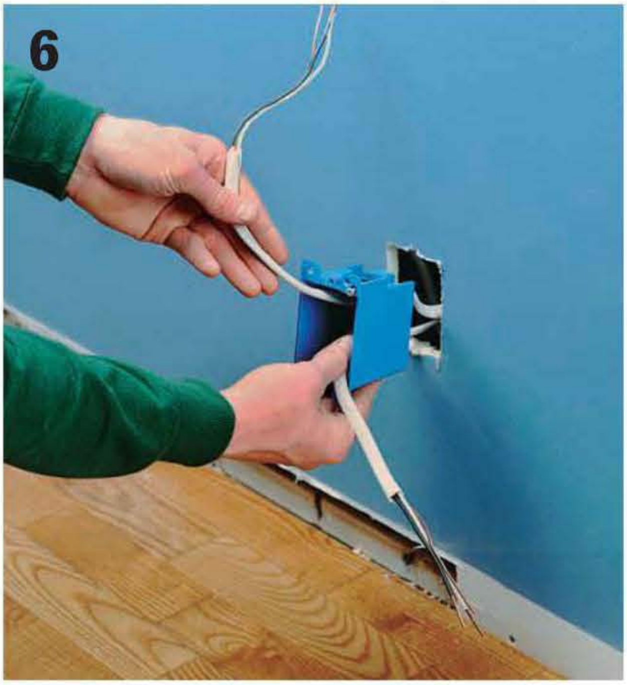

Thread the new and old cables into a pop-in box large enough to contain the added wires and clamp the cables. Fit the box into the old hole and attach it.

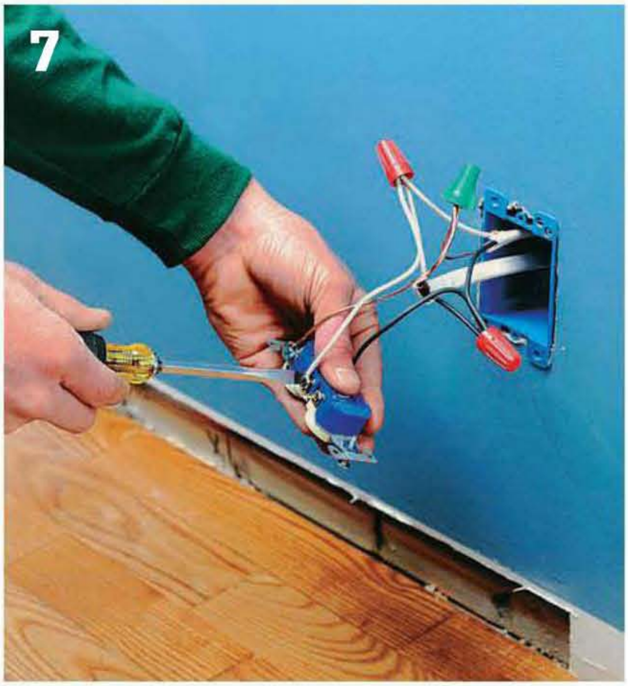

Reconnect the old receptacle by connecting its neutral, hot, and grounding screws to the new branch cable and the old cable from the panel with pigtails.

Pull the cable through another pop-in box for the new receptacle. Secure the cable and install the box. Connect the new receptacle to the new branch cable. Insert the receptacle into the box and attach the receptacle and cover plate with screws. Patch the opening with ½"-thick wood strips or drywall. Reattach the baseboard to the studs.

GFCI Receptacles

The ground-fault circuit-interrupter (GFCI) receptacle protects against electrical shock caused by a faulty appliance, or a worn cord or plug. It senses small changes in current flow and can shut off power in as little as 1⁄40 of a second.

GFCIs are now required in bathrooms, kitchens, garages, crawl spaces, unfinished basements, and outdoor receptacle locations. Consult your local codes for any requirements regarding the installation of GFCI receptacles. Most GFCIs use standard screw terminal connections, but some have wire leads and are attached with wire connectors. Because the body of a GFCI receptacle is larger than a standard receptacle, small crowded electrical boxes may need to be replaced with more spacious boxes.

The GFCI receptacle may be wired to protect only itself (single location), or it can be wired to protect all receptacles, switches, and light fixtures from the GFCI "forward" to the end of the circuit (multiple locations).

Because the GFCI is so sensitive, it is most effective when wired to protect a single location. The more receptacles any one GFCI protects, the more susceptible it is to "phantom tripping," shutting off power because of tiny, normal fluctuations in current flow.

Tools & Materials ▸

Circuit tester
Screwdriver
Wire connectors
Masking tape

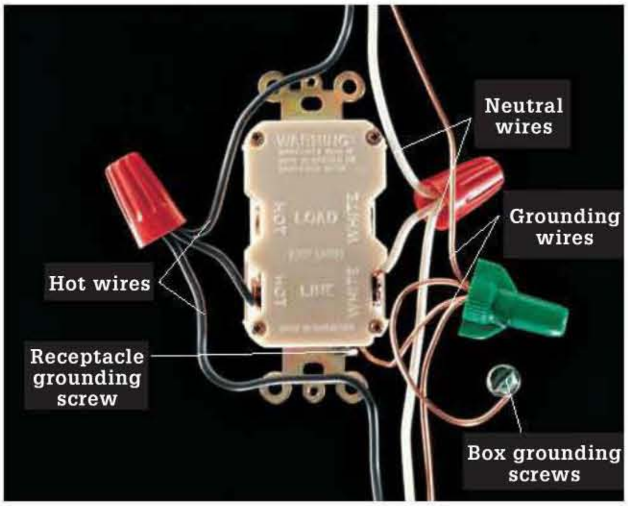
Neutral wires
Grounding wires
Hot wires
Receptacle grounding screw
Box grounding screws

A GFCI wired for single-location protection (shown from the back) has hot and neutral wires connected only to the screw terminals marked LINE. A GFCI connected for single-location protection may be wired as either an end-of-run or middle-of-run configuration (page 110).

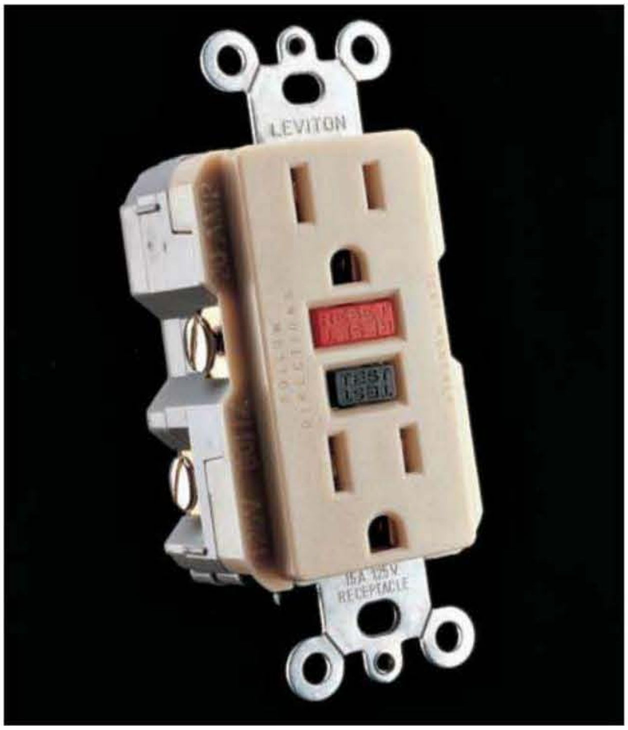
LEVITON
15A 125V RECEPTACLE
TEST

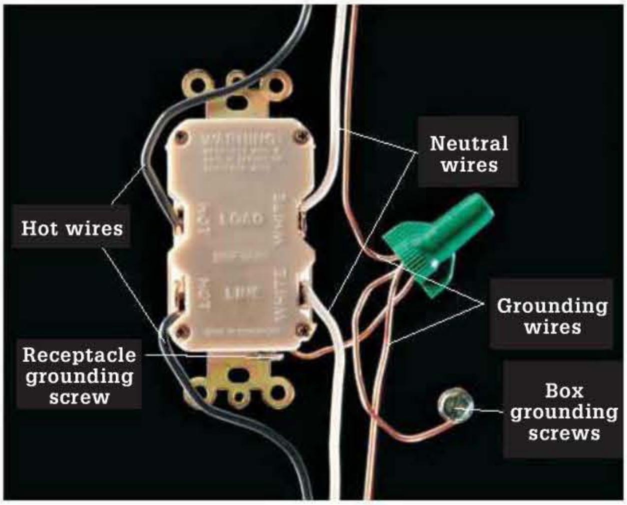
Neutral wires
Hot wires
Grounding wires
Receptacle grounding screw
Box grounding screws

A GFCI wired for multiple-location protection (shown from the back) has one set of hot and neutral wires connected to the LINE pair of screw terminals, and the other set connected to the LOAD pair of screw terminals. A GFCI receptacle connected for multiple-location protection may be wired only as a middle-of-run configuration.

How to Install a Single-location GFCI

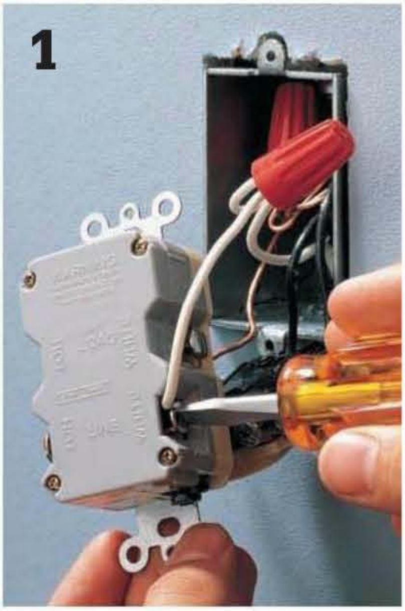

Pigtail all the white neutral wires together, and connect the pigtail to the terminal marked WHITE LINE on the GFCI (see photo on opposite page).

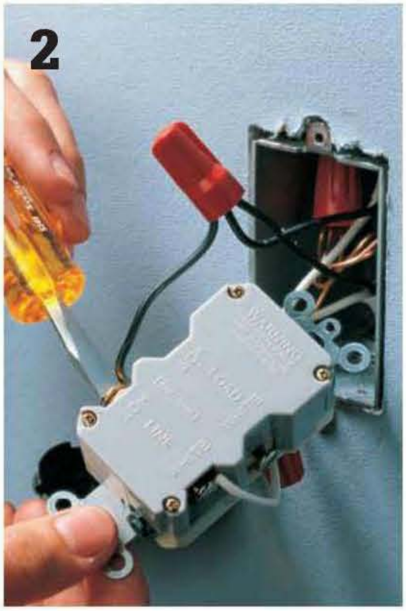

Disconnect all black hot wires from the brass screw terminals of the old receptacle. Pigtail these wires together, and connect them to the terminal marked HOT LINE on the GFCI.

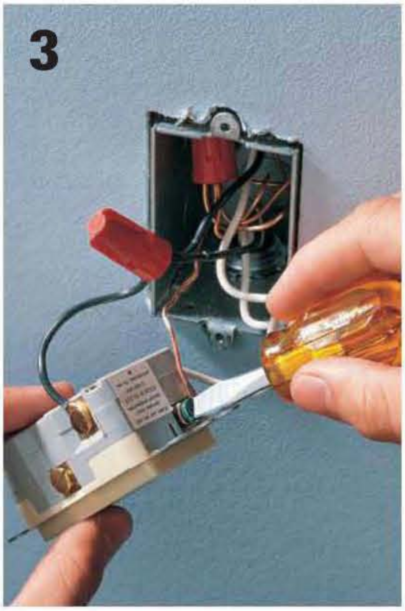

If a grounding wire is available, connect it to the green grounding screw terminal of the GFCI. Mount the GFCI in the receptacle box, and reattach the cover plate. Restore power, and test the GFCI according to the manufacturer's instructions.

How to Install Multiple-location GFCIs

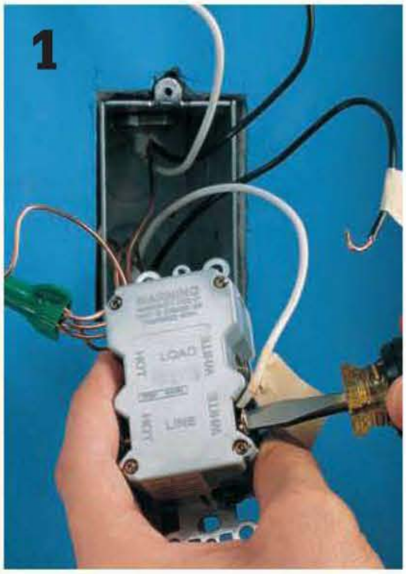

Connect the white feed wire to the terminal marked WHITE LINE on the GFCI. Connect the black feed wire to the terminal marked HOT LINE on the GFCI.

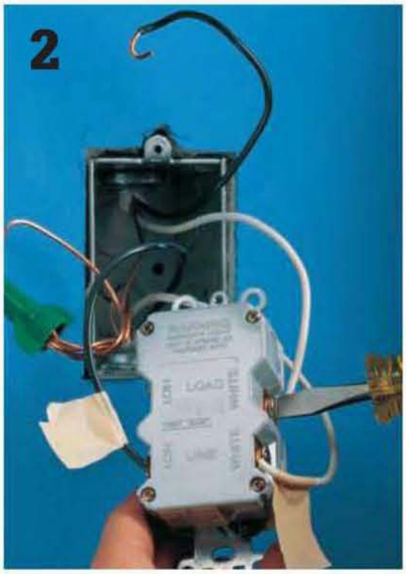

Connect the other white neutral wire to the terminal marked WHITE LOAD on the GFCI.

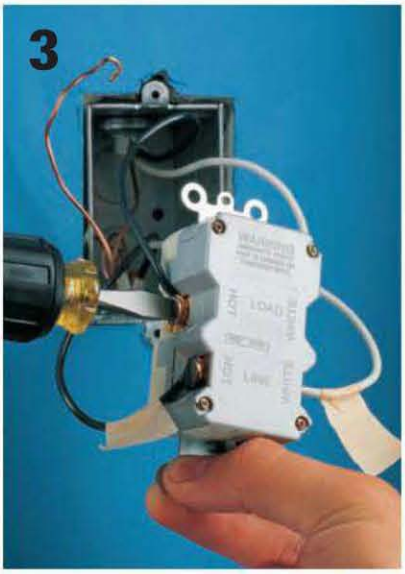

Connect the other black hot wire to the terminal marked HOT LOAD on the GFCI.

Planning New Wiring

Careful planning of a wiring project ensures you will have plenty of power for present and future needs. Whether you are adding circuits in a room addition, wiring a remodeled kitchen, or adding an outdoor circuit, consider all possible ways the space might be used, and plan for enough electrical service to meet peak needs.

For example, when wiring a room addition, remember that the way a room is used can change. In a room used as a spare bedroom, a single 15-amp circuit provides plenty of power, but if you ever choose to convert the same room to a family recreation space, you will need additional circuits.

When wiring a remodeled kitchen, it is a good idea to install circuits for an electric oven and countertop range, even if you do not have these electric appliances. Installing these circuits now makes it easy to convert from gas to electric appliances at a later date.

A large wiring project adds a considerable load to your main electrical service. In about 25 percent of all homes, some type of service upgrade is needed before new wiring can be installed. For example, many homeowners will need to replace an older 60-amp electrical service with a new service rated for 100 amps or more. This is a job for a licensed electrician but is well worth the investment. In other cases, the existing main service provides adequate power, but the main circuit breaker panel is too full to hold any new circuit breakers. In this case it is necessary to install a circuit breaker subpanel to provide room for hooking up added circuits.

This chapter gives an easy five-step method for determining your electrical needs and planning new circuits.

Five Steps for Planning a Wiring Project

1

MAIN
DANGER
TURN OFF POWER
WARNING

Examine your main service panel (page 130). The amp rating of the electrical service and the size of the circuit breaker panel will help you determine if a service upgrade is needed.

2

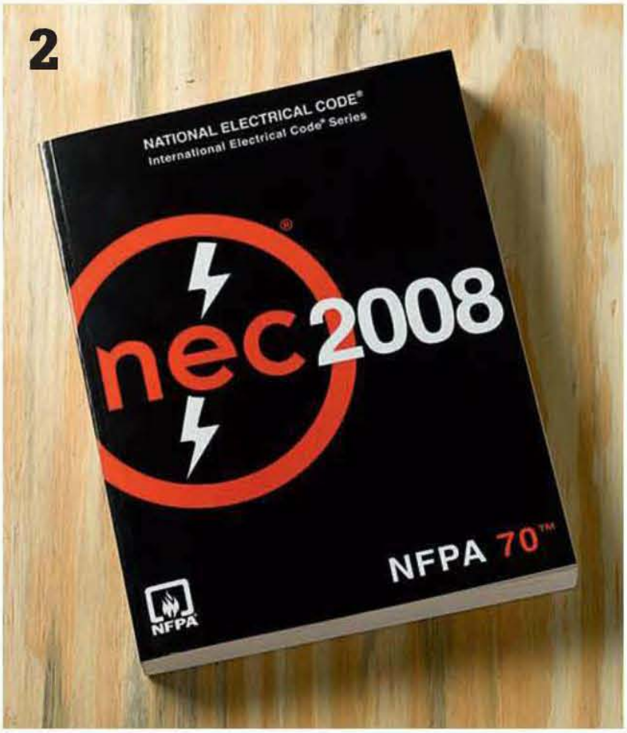
NATIONAL ELECTRICAL CODE®
International Electrical Code® Series
nec 2008
NFPA 70™
NFPA

Learn about codes (pages 131 to 135). The National Electrical Code (NEC), and local electrical codes and building codes, provide guidelines for determining how much power and how many circuits your home needs. Your local electrical inspector can tell you which regulations apply to your job.

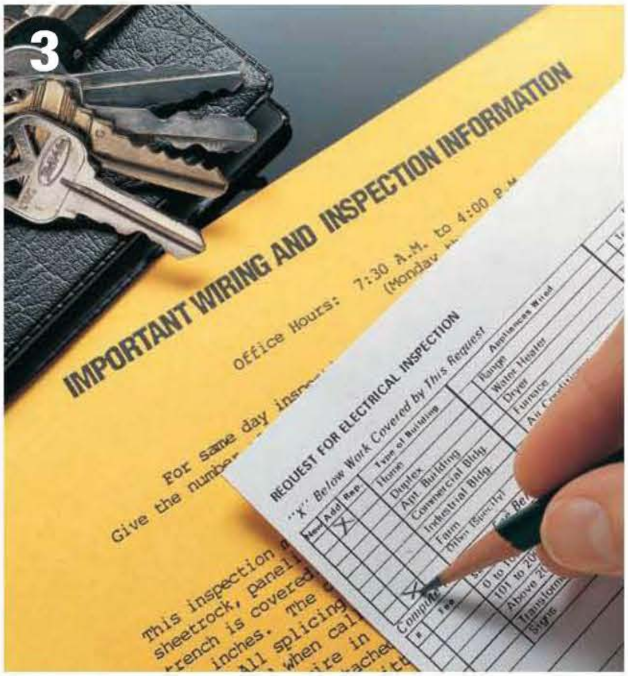
3
IMPORTANT WIRING AND INSPECTION INFORMATION
Office Hours: 7:30 A.M. to 4:00 P.M.
REQUEST FOR ELECTRICAL INSPECTION
"X" Below Work Covered by This Request
Home
Duplex
Apt Building
Commercial Bldg.
Industrial Bldg.
Farm
Range
Water Heater
Dryer
Furnace
Signs

Prepare for inspection. Remember that your work must be reviewed by your local electrical inspector. When planning your wiring project, always follow the inspector's guidelines for quality workmanship.

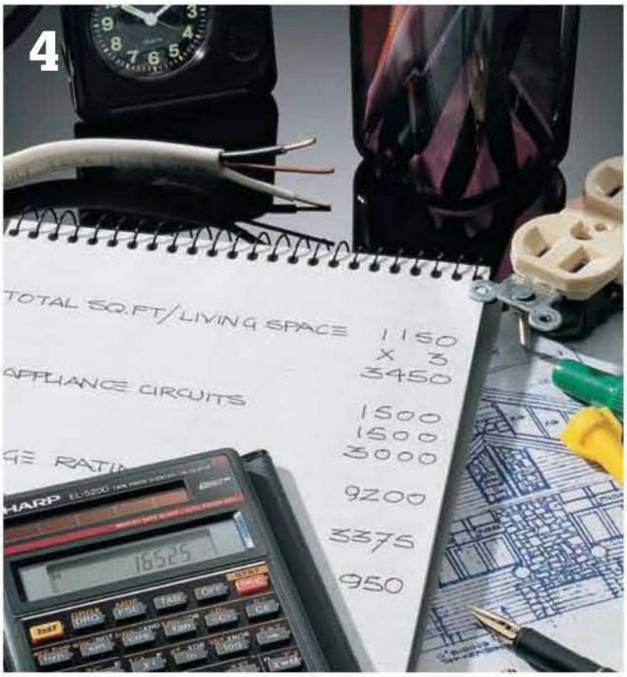
4
TOTAL SQ.FT/LIVING SPACE 1150
x 3
3450
APPLIANCE CIRCUITS
1500
1500
3000
9200
3375
950
SHARP
16525

Evaluate electrical loads. New circuits put an added load on your electrical service. Make sure that total load of the existing wiring and the planned new circuits does not exceed the main service capacity.

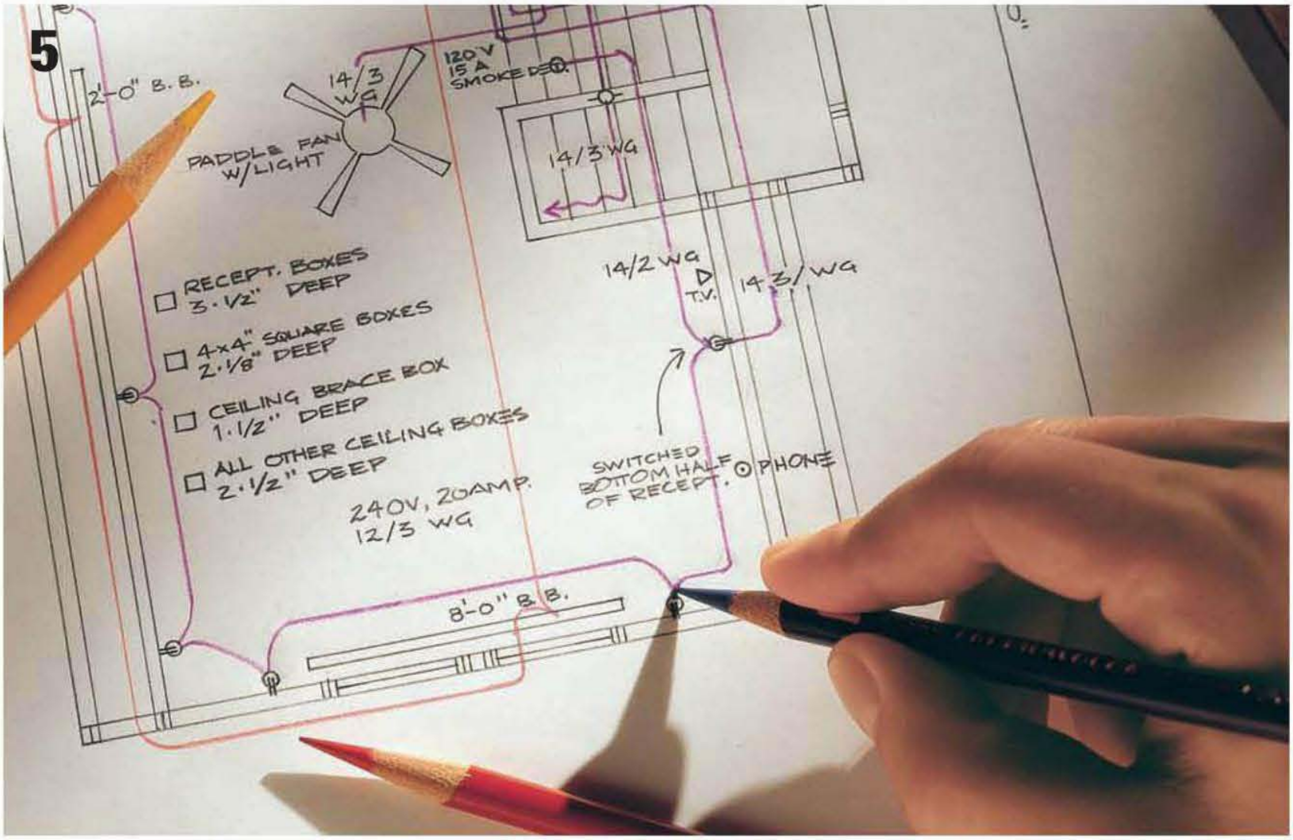
5
2'-0" B.B.
14/3 WG
120 V 15 A SMOKE DET.
PADDLE FAN W/LIGHT
14/3 WG
☐ RECEPT. BOXES 3·1/2" DEEP
☐ 4×4" SQUARE BOXES 2·1/8" DEEP
☐ CEILING BRACE BOX 1·1/2" DEEP
☐ ALL OTHER CEILING BOXES 2·1/2" DEEP
14/2 WG
T.V.
14/3/WG
SWITCHED BOTTOM HALF OF RECEPT.
PHONE
240V, 20AMP. 12/3 WG
8'-0" B.B.

Draw a wiring diagram and get a permit (pages 136 to 137). This wiring plan will help you organize your work.

Examine Your Main Service Panel

The first step in planning a new wiring project is to look in your main circuit breaker panel and find the size of the service by reading the amperage rating on the main circuit breaker. As you plan new circuits and evaluate electrical loads, knowing the size of the main service helps you determine if you need a service upgrade.

Also look for open circuit breaker slots in the panel. The number of open slots will determine if you need to add a circuit breaker subpanel.

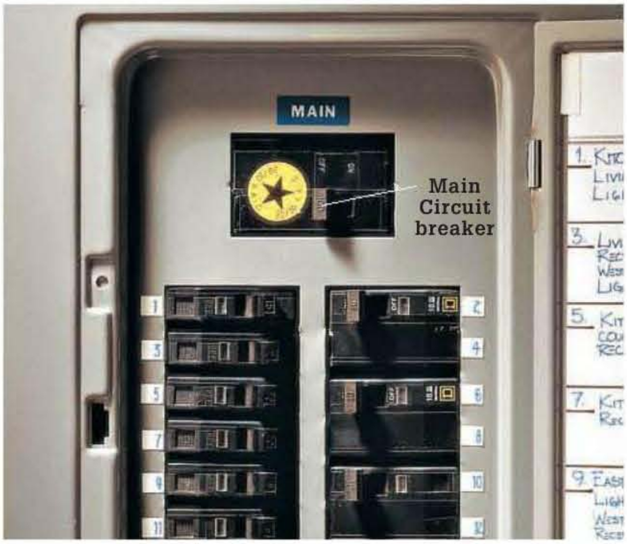

Find the service size by opening the main service panel and reading the amp rating printed on the main circuit breaker. In most cases, 100-amp service provides enough power to handle the added loads of projects like the ones shown in this book. A service rated for 60 amps or less may need to be upgraded.

Older service panels use fuses instead of circuit breakers. Have an electrician replace this type of panel with a circuit breaker panel that provides enough power and enough open breaker slots for the new circuits you are planning.

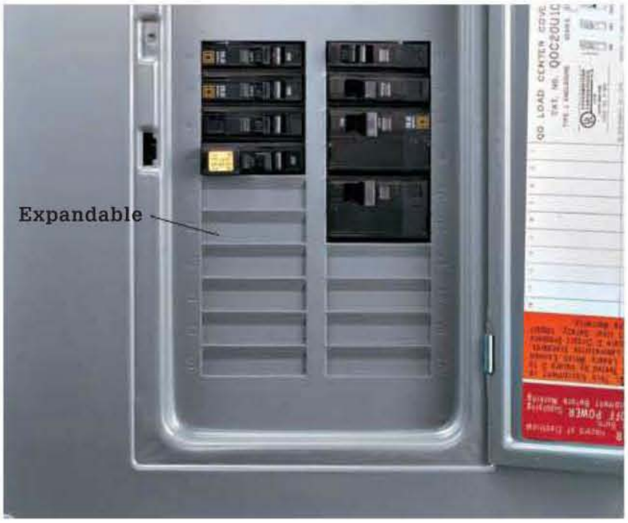

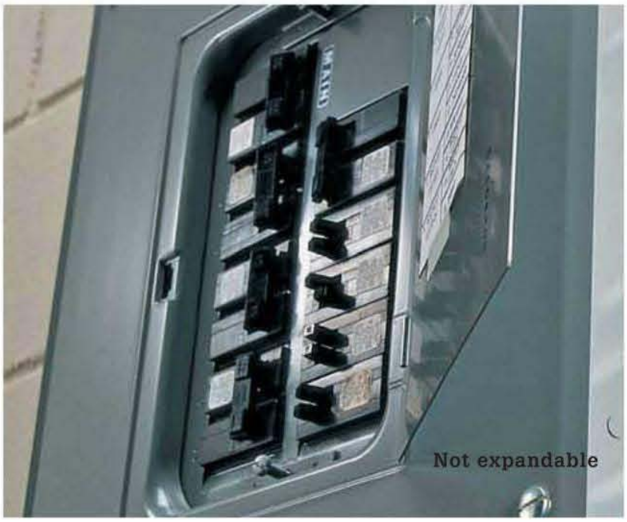

Look for open circuit breaker slots in the main circuit breaker panel or in a circuit breaker subpanel, if your home already has one. You will need one open slot for each 120-volt circuit you plan to install and two slots for each 240-volt circuit. If your main circuit breaker panel has no open breaker slots, install a subpanel to provide room for connecting new circuits.

Learn About Codes

To ensure public safety, your community requires that you get a permit to install new wiring and have the completed work reviewed by an appointed inspector. Electrical inspectors use the National Electrical Code (NEC) as the primary authority for evaluating wiring, but they also follow the local Building Code and Electrical Code standards.

As you begin planning new circuits, call or visit your local electrical inspector and discuss the project with him. The inspector can tell you which of the national and local code requirements apply to your job, and may give you a packet of information summarizing these regulations. Later, when you apply to the inspector for a work permit, he will expect you to understand the local guidelines as well as a few basic National Electrical Code requirements.

The National Electrical Code is a set of standards that provides minimum safety requirements for wiring installations. It is revised every three years. The national code requirements for the projects shown in this book are thoroughly explained on the following pages. For more information, you can find copies of the current NEC, as well as a number of excellent handbooks based on the NEC, at libraries and bookstores.

In addition to being the final authority of code requirements, inspectors are electrical professionals with years of experience. Although they have busy schedules, most inspectors are happy to answer questions and help you design well-planned circuits.

Basic Electrical Code Requirements

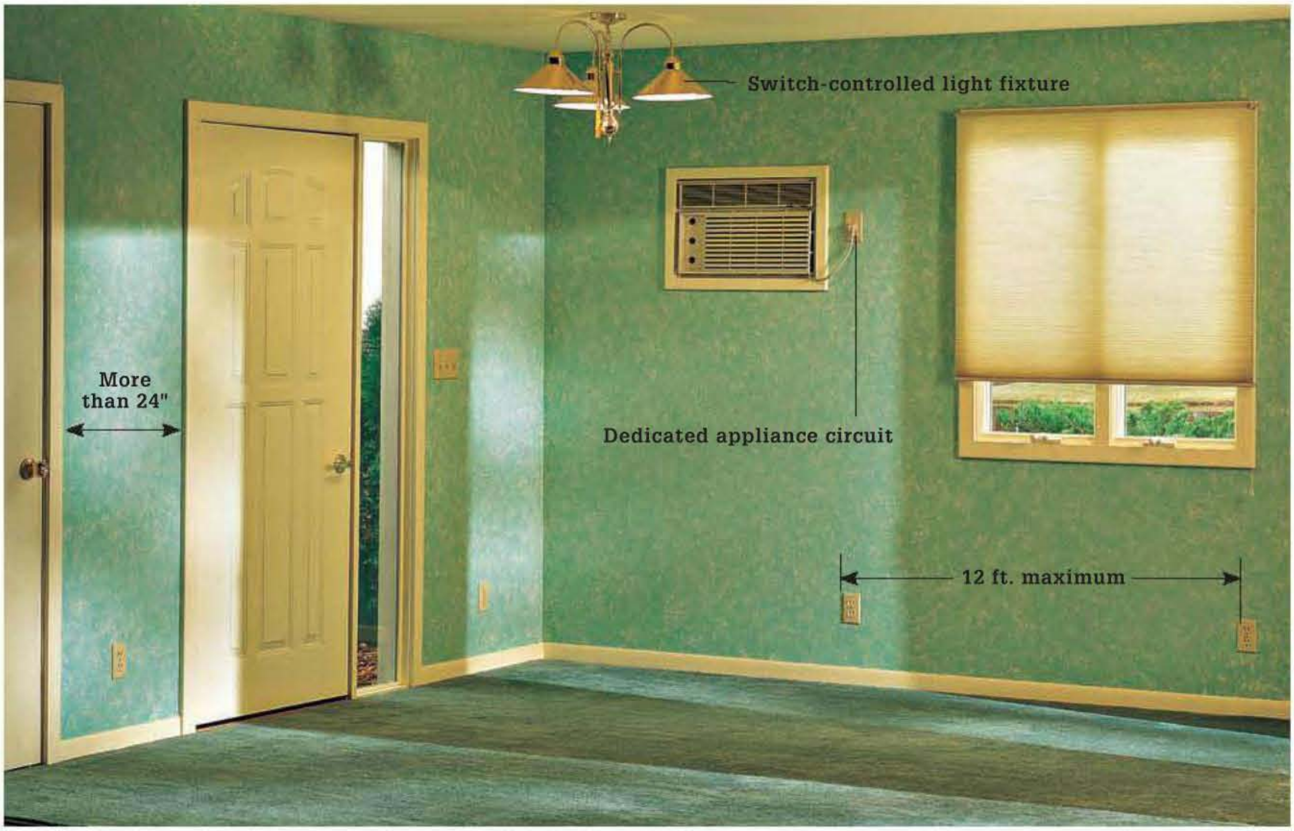

Electrical Code requirements for living areas: Living areas need at least one 15-amp or 20-amp basic lighting/receptacle circuit for each 600 sq. ft. of living space and should have a dedicated circuit for each type of permanent appliance, like an air conditioner, computer, or a group of baseboard heaters and within 6 ft. of any door opening. Receptacles on basic lighting/receptacle circuits should be spaced no more than 12 ft. apart. Many electricians and electrical inspectors recommend even closer spacing. Any wall more than 24" wide also needs a receptacle. Every room should have a wall switch at the point of entry to control either a ceiling light or plug-in lamp. Kitchens and bathrooms must have a ceiling-mounted light fixture.

Selected NEC Standards & Tips

Measure the living areas of your home, excluding closets and unfinished spaces. A sonic measuring tool gives room dimensions quickly and contains a built-in calculator for figuring floor area. You will need a minimum of one basic lighting/receptacle circuit for every 600 sq. ft. of living space. The total square footage also helps you determine heating and cooling needs for new room additions.

Three-way switches

Stairways with six steps or more must have lighting that illuminates each step. The light fixture must be controlled by three-way switches at the top and bottom landings.

Kitchen and bathroom receptacles must be protected by a ground-fault circuit-interrupter (GFCI). Also, all outdoor receptacles and general-use receptacles in an unfinished basement or crawl space and garages must be protected by a GFCI.

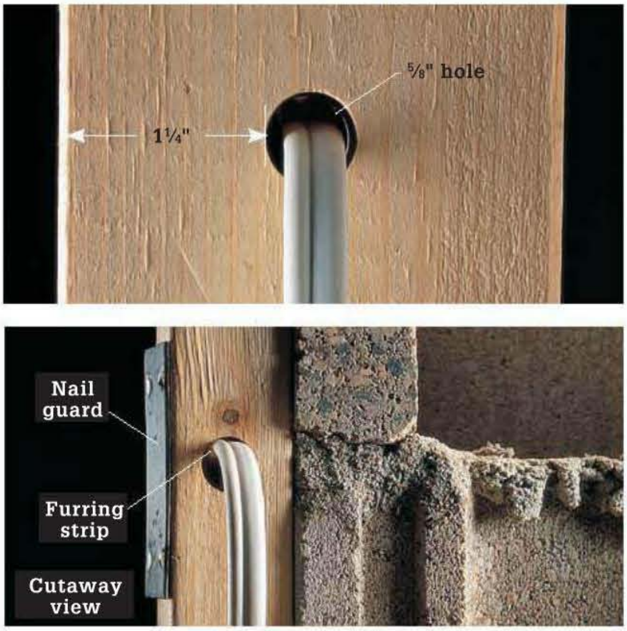

Cables must be protected against damage by nails and screws by at least 1¼" of wood (top). When cables pass through 2 × 2 furring strips (bottom), protect the cables with metal nail guards.

Closets and other storage spaces need at least one light fixture that is controlled by a wall switch near the entrance. Prevent fire hazards by positioning the light fixtures so the outer globes are at least 12" away from all shelf areas.

Hallways more than 10 ft. long need at least one receptacle. All hallways should have a switch-controlled light fixture.

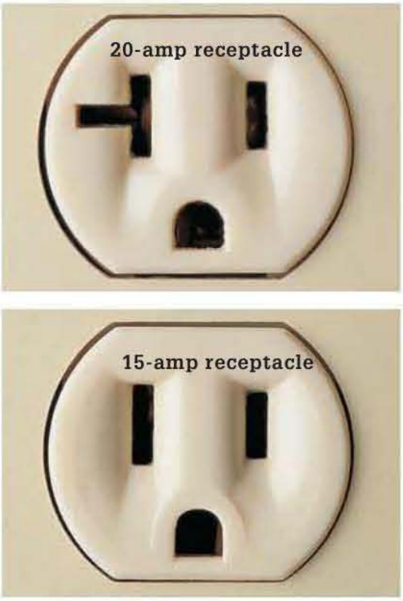
20-amp receptacle
15-amp receptacle

Amp ratings of receptacles must match the size of the circuit. A common mistake is to use 20-amp receptacles (top) on 15-amp circuits—a potential cause of dangerous circuit overloads because it allows you to plug in appliances that draw over 15 amps.

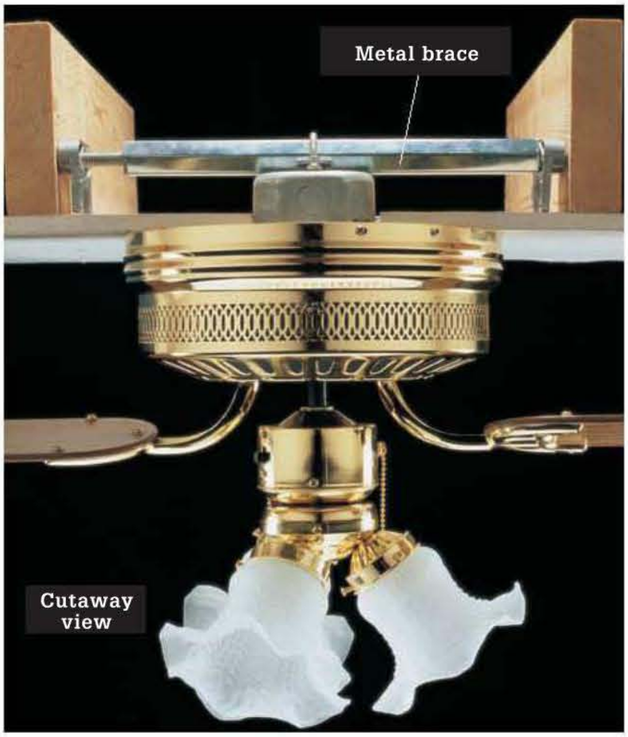
Metal brace
Cutaway view

A metal brace attached to framing members is required for ceiling fans and large light fixtures that are too heavy to be supported by an electrical box.

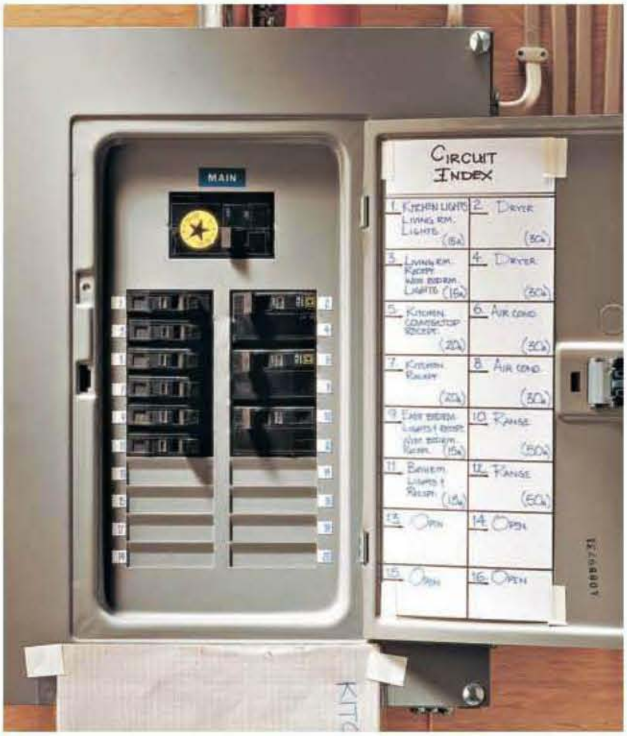
MAIN
CIRCUIT INDEX
1. KITCHEN LIGHTS LIVING RM. LIGHTS (15A)
2. DRYER (30A)
3. LIVING RM. RECEPT. W. BDRM. LIGHTS (15A)
4. DRYER (30A)
5. KITCHEN COUNTERTOP RECEPT. (20A)
6. AIR COND (30A)
7. KITCHEN RECEPT (20A)
8. AIR COND (30A)
9. EAST BDRM. LIGHTS & RECEPT. W. BDRM. RECEPT. (15A)
10. RANGE (50A)
11. BASEM. LIGHTS & RECEPT. (15A)
12. RANGE (50A)
13. OPEN
14. OPEN
15. OPEN
16. OPEN

Label new circuits on an index attached to the circuit breaker panel. List the rooms and appliances controlled by each circuit. Make sure the area around the panel is clean, well lit, and accessible.

Highlights of the National Electrical Code ▸

BY MATERIAL

Service Panel (page 130)

- Maintain a minimum 30 × 36" of clearance in front of the service panel.
- Ground all 120-volt and 240-volt circuits.
- Match the amperage rating of the circuit when replacing fuses.
- Locate service panels and subpanels a maximum of 79" above floor level.
- Use handle-tie breakers for 240-volt loads (line to line).
- Close all unused service panel openings.
- Label each fuse and breaker clearly on the panel.

Electrical Boxes (page 108)

- Use boxes that are large enough to accommodate the number of wires entering the box.
- Locate all receptacle boxes 12" above the finished floor (standard).
- Locate all switch boxes 48" above the finished floor (standard). For special circumstances, inspectors will allow switch and location measurements to be altered, such as a switch at 36" above the floor in a child's bedroom or receptacles at 24" above the floor to make them more accessible for someone in a wheelchair.
- Install all boxes so they remain accessible.
- Leave no gaps greater than ⅛" between wallboard and front of electrical boxes.
- Place receptacle boxes flush with combustible surfaces.
- Leave a minimum of 8" of usable cable or wire extending past the front of the electrical box.

Wires & Cables (page 90)

- Use wires that are large enough for the amperage rating of the circuit (see Wire Size Chart, page 90).
- Drill holes at least 2" back from the exposed edge of joists to run cables through. Do not attach cables to the bottom edge of joists.
- Do not run cables diagonally between framing members.
- Run cable between receptacles 20" above the floor.
- Use nail plates to protect cable that is run through holes drilled or cut into studs less than 1¼" from front edge of stud.
- Do not crimp cables sharply.
- Contain spliced wires or connections entirely in a plastic or metal electrical box.
- Use wire connectors to join wires.
- Use staples to fasten cables within 8" of an electrical box and every 48" along its run.
- Leave a minimum ¼" (maximum 1") of sheathing where cables enter an electrical box.
- Clamp cables and wires to electrical boxes with approved NM clamp. No clamp is necessary for one-gang plastic boxes if cables are stapled within 8".
- Label all cables and wires at each electrical box to show which circuits they serve for the rough-in inspection.
- Connect only a single wire to a single screw terminal. Use pigtails to join more than one wire to a screw terminal.

Switches (page 116)

- Use a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Use three-way switches at the top and bottom on stairways with six steps or more.
- Use switches with grounding screw with plastic electrical boxes.
- Locate all wall switches within easy reach of the room entrance.

Receptacles (page 122)

- Match the amp rating of a receptacle with the size of the circuit.
- Include receptacles on all walls 24" wide or greater.
- Include receptacles so a 6-ft. cord can be plugged in from any point along a wall or every 12-ft. along a wall.
- Include receptacles in any hallway that is 10-ft. long or more.
- Use three-prong, grounded receptacles for all 15- or 20-amp, 120-volt branch circuits.
- Include a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Install GFCI-protected receptacles in bathrooms, kitchens, garages, crawl spaces, unfinished basements, and outdoor receptacle locations.
- Install an isolated-ground circuit to protect sensitive equipment, like a computer, against tiny power fluctuations. Computers should also be protected by a standard surge protector.

Light Fixtures

- Use mounting straps that are anchored to the electrical boxes to mount ceiling fixtures.
- Keep non-IC-rated recessed light fixtures 3" from insulation and ½" from combustibles.
- Include at least one switch-operated lighting fixture in every room.

Grounding

- Ground all receptacles by connecting receptacle grounding screws to the circuit grounding wires.
- Use switches with grounding screws whenever possible. Always ground switches installed in plastic electrical boxes and all switches in kitchens, bathrooms, and basements.

BY ROOM

Kitchens/Dining Rooms

- Install a dedicated 40- or 50-amp, 120/240-volt circuit for a range (or two circuits for separate oven and countertop units).
- Install two 20-amp small appliance circuits.
- Install dedicated 15-amp, 120-volt circuits for dishwashers and food disposals (required by many local codes).
- Use GFCI receptacles for all accessible countertop receptacles; receptacles behind fixed appliances do not need to be GFCIs.
- Position receptacles for appliances that will be installed within cabinets, such as microwaves or food disposals, according to the manufacturer's instructions.
- Include receptacles on all counters wider than 12".
- Space receptacles a maximum of 48" apart above countertops and closer together in areas where many appliances will be used.
- Locate receptacles 18" above countertop. If backsplash is more than the standard 4" or the bottom of cabinet is less than 18" from countertop, center the box in space between countertop and bottom of wall cabinet.
- Mount one receptacle within 12" of the countertop on islands and peninsulas that are 12 × 24" or greater.
- Do not put lights on small appliance circuits.
- Install additional lighting in work areas at a sink or range for convenience and safety.

Bathrooms

- Install a separate 20-amp circuit.
- Ground switches in bathrooms.
- Use GFCI-protected receptacles.
- Install at least one ceiling-mounted light fixture.
- Place blower heaters in bathrooms well away from the sink and tub.

Utility/Laundry Rooms

- Install a separate 20-amp circuit for a washing machine.
- Install a minimum feed 30-amp #10 THHN wire for the dryer powered by a separate 120/240-volt major appliance circuit.
- Install metal conduit for cable runs in unfinished rooms.
- Use GFCI-protected receptacles, except for fixed appliances, such as freezers or dryers.

Living, Entertainment, Bedrooms

- Install a minimum of two 15-amp circuits in living rooms.
- Install a minimum of one 15- or 20-amp basic lighting/receptacle circuit for each 600 sq. ft. of living space.
- Install a dedicated circuit for each permanent appliance, like an air conditioner, computer, or group of electric baseboard heaters.
- Do not use standard electrical boxes to support ceiling fans.
- Include receptacles on walls 24" wide or more.
- Space receptacles on basic lighting/receptacle circuits a maximum of 12 ft. apart. For convenience you can space them as close as 6 ft.
- Position permanent light fixtures in the center of the room's ceiling.
- Install permanently wired smoke alarms in room additions that include sleeping areas.

Outdoors

- Check for underground utilities before digging.
- Use UF cable for outdoor wiring needs.
- Run cable in schedule 80 PVC plastic, as required by local code.
- Bury cables housed in conduit at least 18" deep; cable not in conduit must be buried at least 24" deep.
- Use weatherproof electrical boxes with watertight covers.
- Use GFCI-protected receptacles.
- Install receptacles a minimum of 12" above ground level.
- Anchor freestanding receptacles not attached to a structure by embedding the schedule 80 PVC plastic conduit in a concrete footing, so that it is at least 12" but no more than 18" above ground level.
- Plan on installing a 20-amp, 120-volt circuit if the circuit contains more than one light fixture rated for 300 watts, or more than four receptacles.

Stairs/Hallways

- Use three-way switches at the top and bottom on stairways with six steps or more.
- Include receptacles in any hallway that is 10 ft. long or more.
- Position stairway lights so each step is illuminated.

Draw a Diagram & Obtain a Permit

Drawing a wiring diagram is the last step in planning a circuit installation. A detailed wiring diagram helps you get a work permit, makes it easy to create a list of materials, and serves as a guide for laying out circuits and installing cables and fixtures. Use the circuit maps on pages 138 to 145 as a guide for planning wiring configurations and cable runs. Bring the diagram and materials list when you visit electrical inspectors to apply for a work permit.

Never install new wiring without following your community's permit and inspection procedure. A work permit is not expensive, and it ensures that your work will be reviewed by a qualified inspector to guarantee its safety. If you install new wiring without the proper permit, an accident or fire traced to faulty wiring could cause your insurance company to discontinue your policy and can hurt the resale value of your home.

When electrical inspectors look over your wiring diagram, they will ask questions to see if you have a basic understanding of the electrical code and fundamental wiring skills. Some inspectors ask these questions informally, while others give a short written test. Inspectors may allow you to do some, but not all, of the work. For example, they may ask that all final circuit connections at the circuit breaker panel be made by a licensed electrician, while allowing you to do all other work.

A few communities allow you to install wiring only when supervised by an electrician. This means you can still install your own wiring but must hire an electrician to apply for the work permit and to check your work before inspectors review it. The electrician is held responsible for the quality of the job.

Remember that it is the inspectors' responsibility to help you do a safe and professional job. Feel free to call them with questions about wiring techniques or materials.

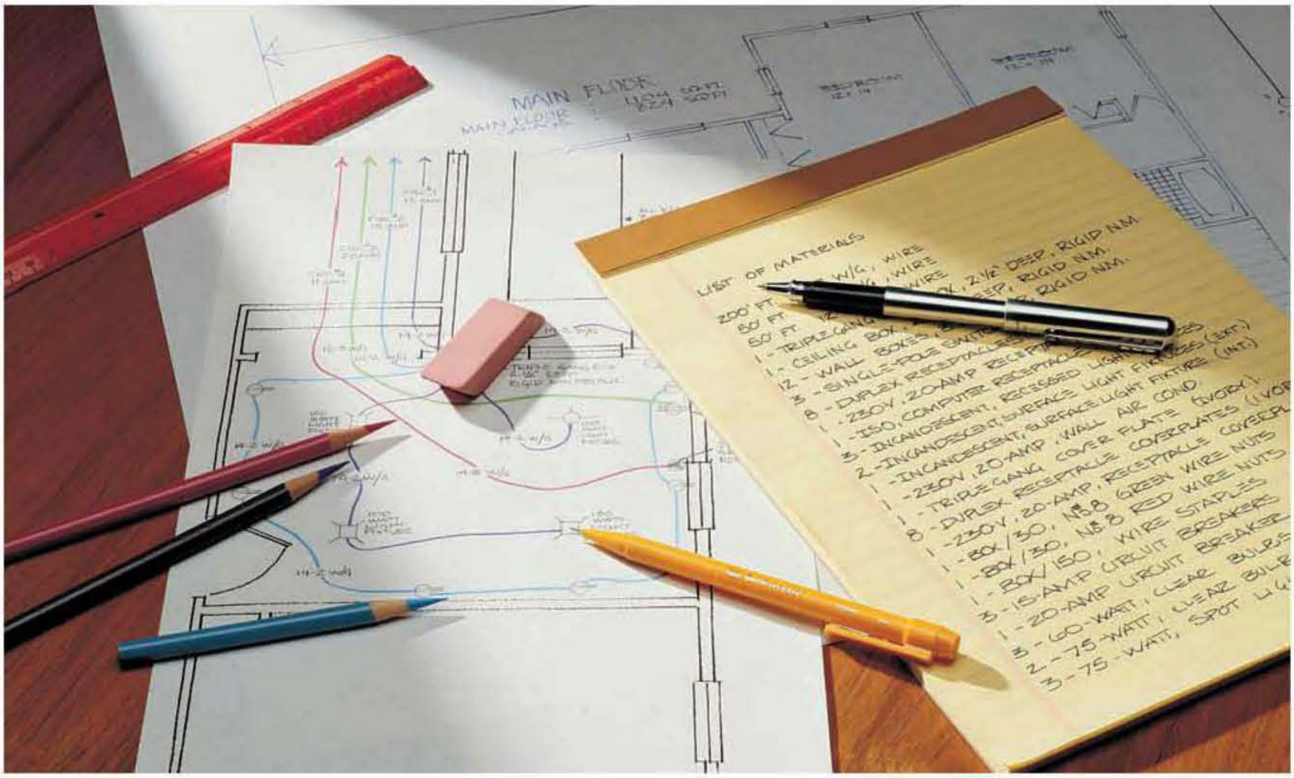

A detailed wiring diagram and a list of materials is required before electrical inspectors will issue a work permit. If blueprints exist for the space you are remodeling, start your electrical diagram by tracing the wall outlines from the blueprint. Use standard electrical symbols (next page) to clearly show all the receptacles, switches, light fixtures, and permanent appliances. Make a copy of the symbol key, and attach it to the wiring diagram for the inspectors' convenience. Show each cable run, and label its wire size and circuit amperage.

How to Draw a Wiring Plan

1

Draw a scaled diagram of the space you will be wiring, showing walls, doors, windows, plumbing pipes and fixtures, and heating and cooling ducts. Find the floor space by multiplying room length by width, and indicate this on the diagram. Do not include closets or storage areas when figuring space.

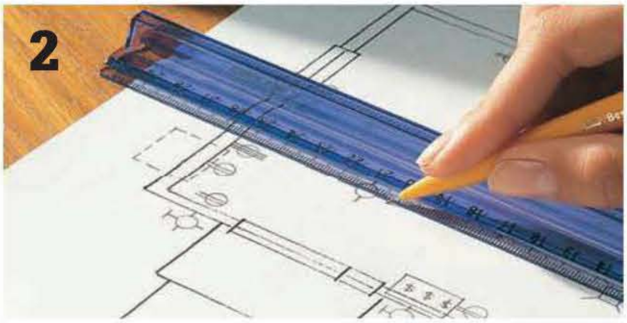

2

Mark the location of all switches, receptacles, light fixtures, and permanent appliances using the electrical symbols shown below. Where you locate these devices along the cable run determines how they are wired. Use the circuit maps on pages 138 to 145 as a guide for drawing wiring diagrams.

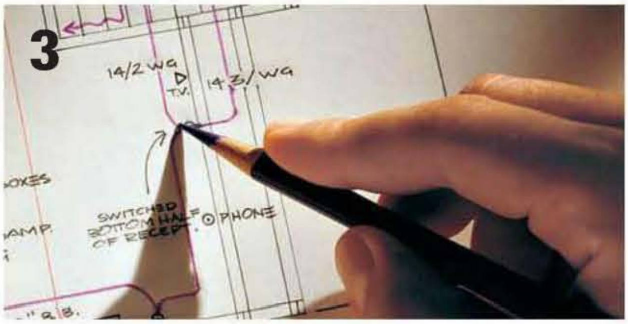

3

14/2 WG
T.V.
14-3/WG
SWITCHED BOTTOM HALF OF RECEPT.
PHONE

Draw in cable runs between devices. Indicate cable size and type, and the amperage of the circuits. Use a different-colored pencil for each circuit.

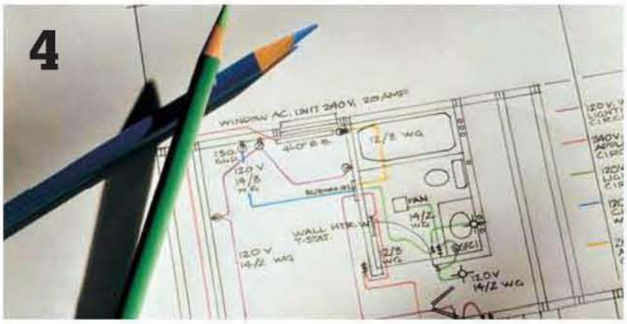

4

WINDOW AC. UNIT 240V, 20 AMP
12/2 WG
120 V 14/2 WG
WALL HTR. W/ T-STAT
FAN
14/2 WG
12/3 WG
GFCI
120V 14/2 WG

Identify the wattages for light fixtures and permanent appliances, and the type and size of each electrical box. On another sheet of paper, make a detailed list of all materials you will use.

Electrical Symbol Key ▸

(copy this key and attach it to your wiring plan)

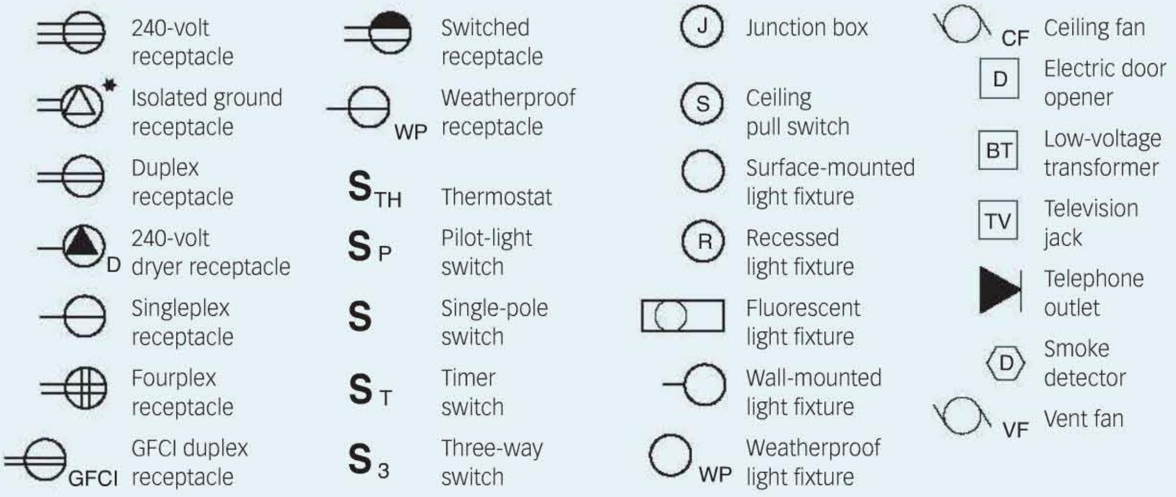

240-volt receptacle
Isolated ground receptacle
Duplex receptacle
D
240-volt dryer receptacle
Singleplex receptacle
Fourplex receptacle
GFCI
GFCI duplex receptacle
Switched receptacle
WP
Weatherproof receptacle
S TH
Thermostat
S P
Pilot-light switch
S
Single-pole switch
S T
Timer switch
S 3
Three-way switch
J
Junction box
S
Ceiling pull switch
Surface-mounted light fixture
R
Recessed light fixture
Fluorescent light fixture
Wall-mounted light fixture
WP
Weatherproof light fixture
CF
Ceiling fan
D
Electric door opener
BT
Low-voltage transformer
TV
Television jack
Telephone outlet
D
Smoke detector
VF
Vent fan

Using Circuit Maps

The arrangement of switches and appliances along an electrical circuit differs for every project. This means that the configuration of wires inside an electrical box can vary greatly, even when fixtures are identical.

The circuit maps on the following pages show the most common wiring variations for typical electrical devices. Most new wiring you install will match one or more of the maps shown. Find the maps that match your situation and use them to plan your circuit layouts.

The 120-volt circuits shown on the following pages are wired for 15 amps using 14-gauge wire and receptacles rated at 15 amps. If you are installing a 20-amp circuit, substitute 12-gauge cables and use receptacles rated for 20 amps.

In configurations where a white wire serves as a hot wire instead of a neutral, both ends of the wire are coded with black tape to identify it as hot. In addition, each of the circuit maps shows a box grounding screw. This grounding screw is required in all metal boxes, but plastic electrical boxes do not need to be grounded.

Note: For clarity, all grounding conductors in the circuit maps are colored green. In practice, the grounding wires inside sheathed cables usually are bare copper.

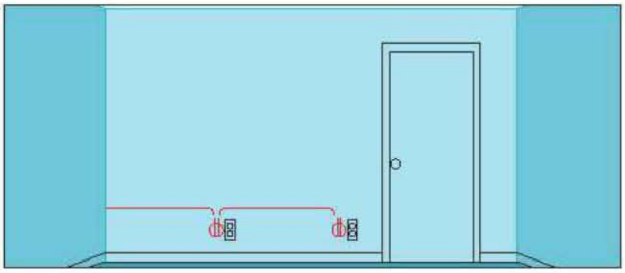

1. 120-VOLT DUPLEX RECEPTACLES WIRED IN SEQUENCE

Use this layout to link any number of duplex receptacles in a basic lighting/receptacle circuit. The last receptacle in the cable run is connected like the receptacle shown at the right side of the circuit map below. All other receptacles are wired like the receptacle shown on the left side. Requires two-wire cables.

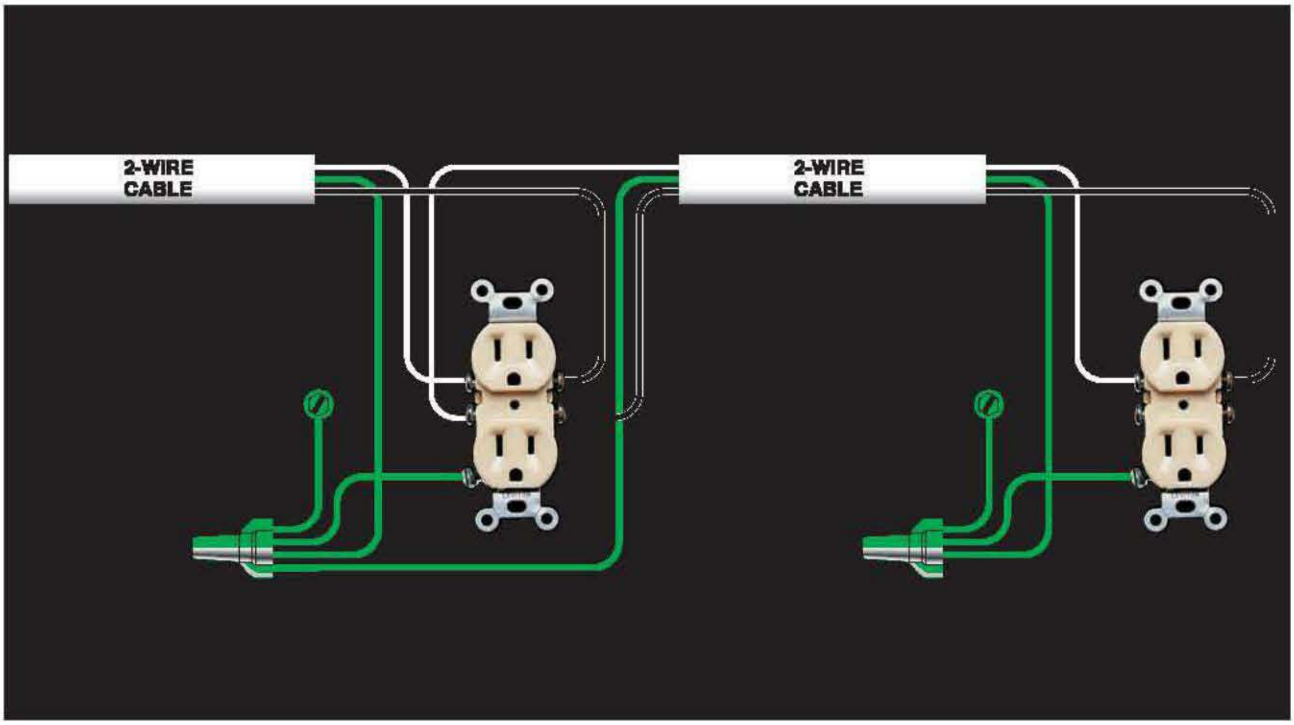

2. GFCI RECEPTACLES (Single-location Protection)

Use this layout when receptacles are within 6 ft. of a water source, like those in kitchens and bathrooms. To prevent nuisance tripping caused by normal power surges, GFCIs should be connected only at the line screw terminal so they protect a single location, not the fixtures on the load side of the circuit. Requires two-wire cables. Where a GFCI must protect other fixtures, use circuit map 3.

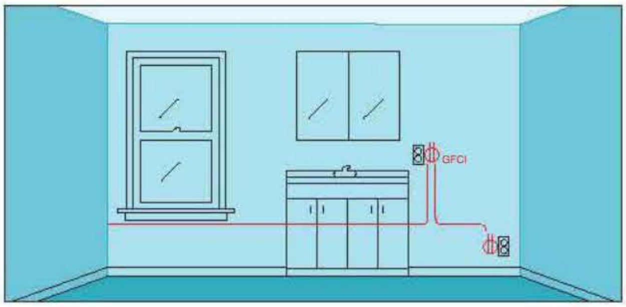

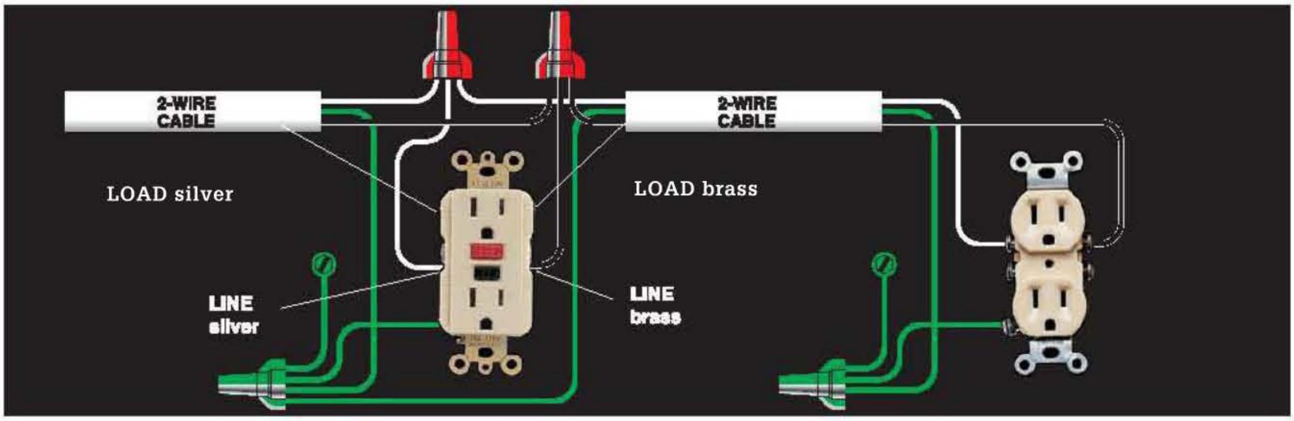
2-WIRE CABLE
2-WIRE CABLE
LOAD silver
LOAD brass
LINE silver
LINE brass

3. GFCI RECEPTACLE, SWITCH & LIGHT FIXTURE (Wired for Multiple-location Protection)

In some locations, such as an outdoor circuit, it is a good idea to connect a GFCI receptacle so it also provides shock protection to the wires and fixtures that continue to the end of the circuit. Wires from the power source are connected to the line screw terminals; outgoing wires are connected to load screws. Requires two-wire cables.

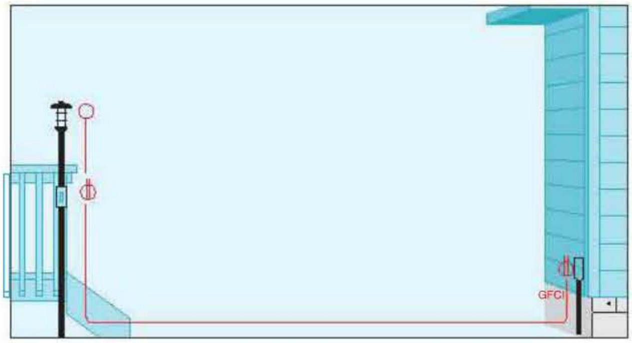

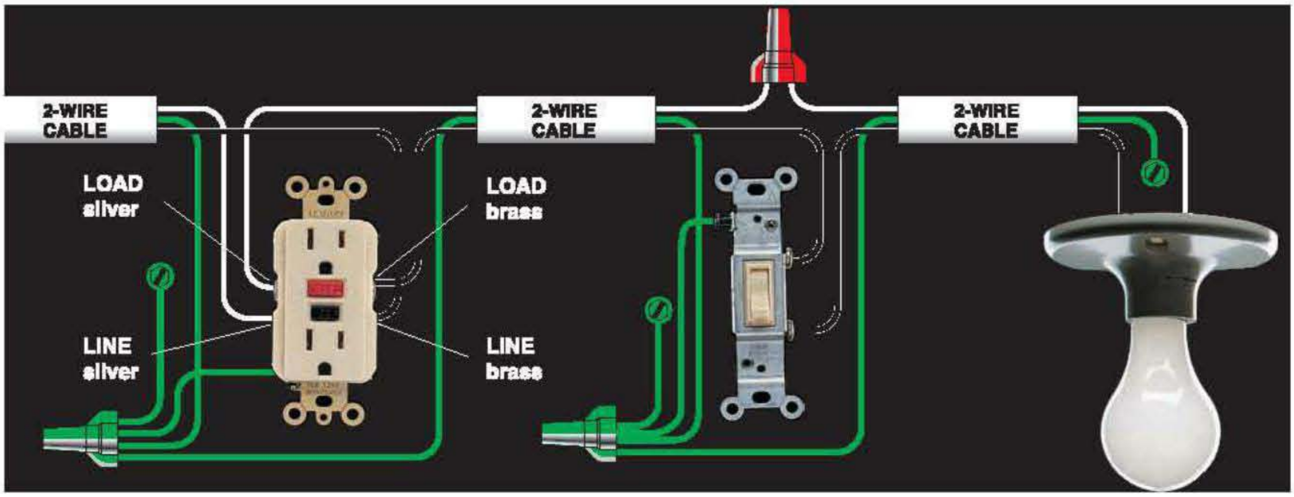
2-WIRE CABLE
2-WIRE CABLE
2-WIRE CABLE
LOAD silver
LOAD brass
LINE silver
LINE brass

4. SINGLE-POLE SWITCH & LIGHT FIXTURE (Light Fixture at End of Cable Run)

Use this layout for light fixtures in basic lighting/receptacle circuits throughout the home. It is often used as an extension to a series of receptacles (circuit map 1). Requires two-wire cables.

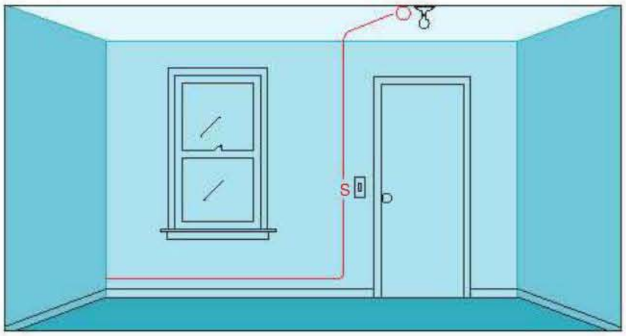

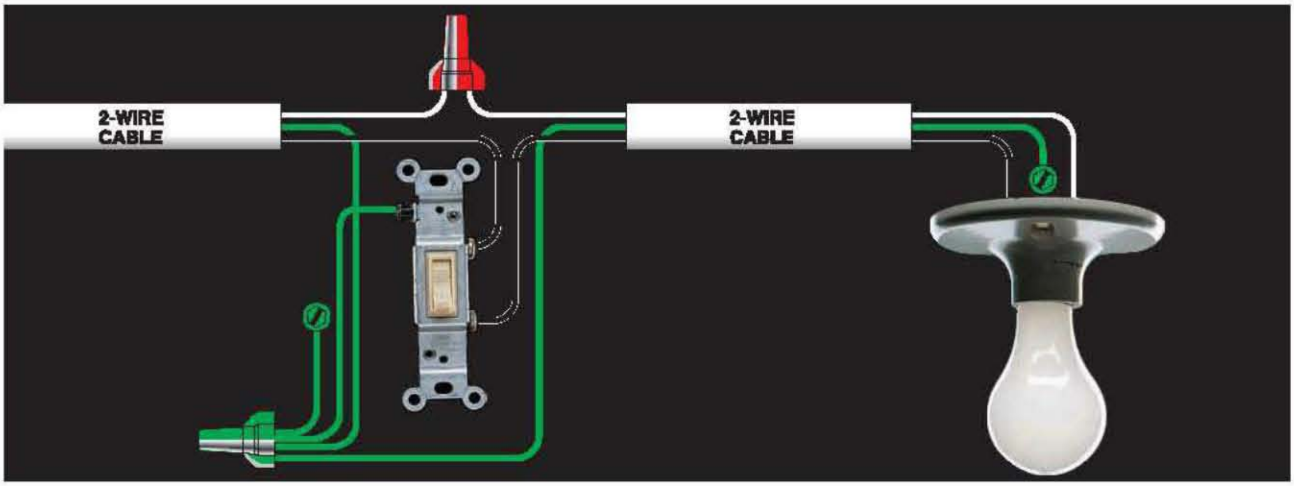
2-WIRE CABLE
2-WIRE CABLE

5. SINGLE-POLE SWITCH & LIGHT FIXTURE (Switch at End of Cable Run)

Use this layout, sometimes called a switch loop, where it is more practical to locate a switch at the end of the cable run. In the last length of cable, both insulated wires are hot; the white wire is tagged with black tape at both ends to indicate it is hot. Requires two-wire cables.

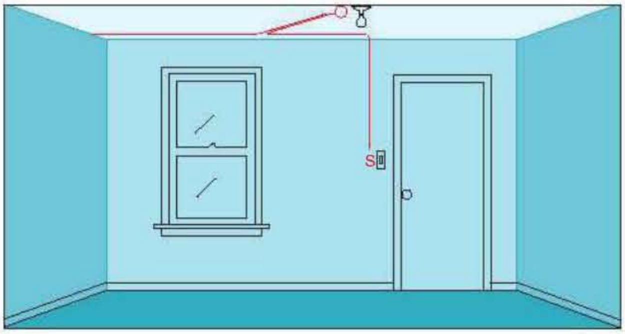

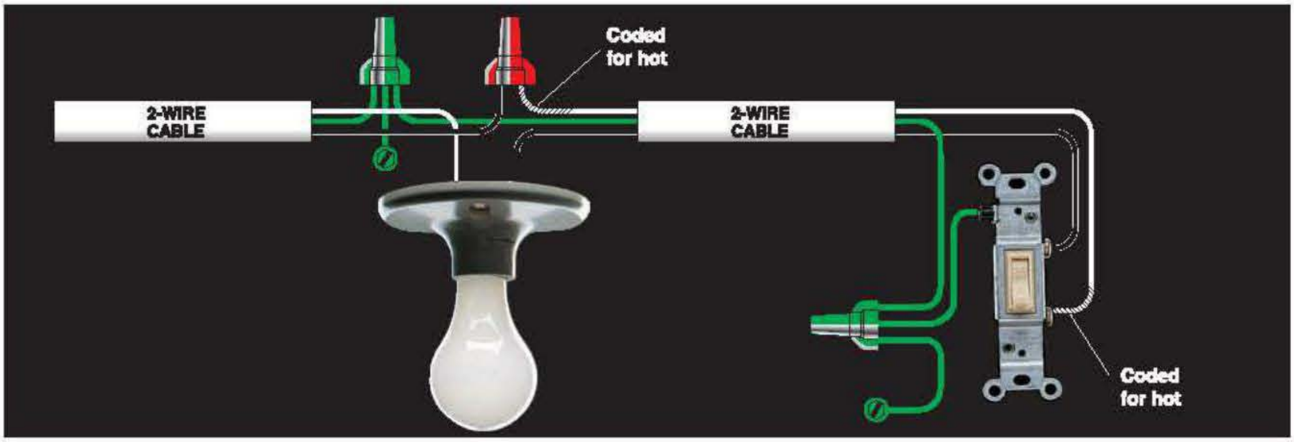
2-WIRE CABLE
Coded for hot
2-WIRE CABLE
Coded for hot

6. SINGLE-POLE SWITCH & TWO LIGHT FIXTURES (Switch Between Light Fixtures, Light at Start of Cable Run)

Use this layout when you need to control two fixtures from one single-pole switch and the switch is between the two lights in the cable run. Power feeds to one of the lights. Requires two-wire and three-wire cables.

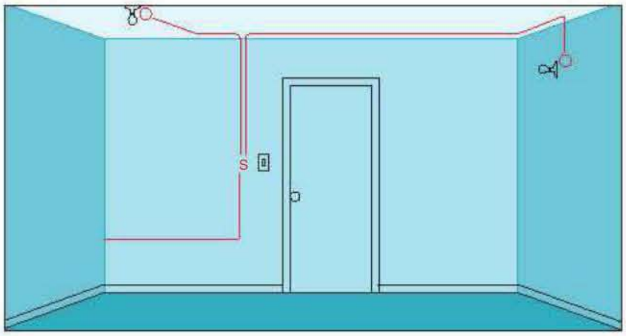

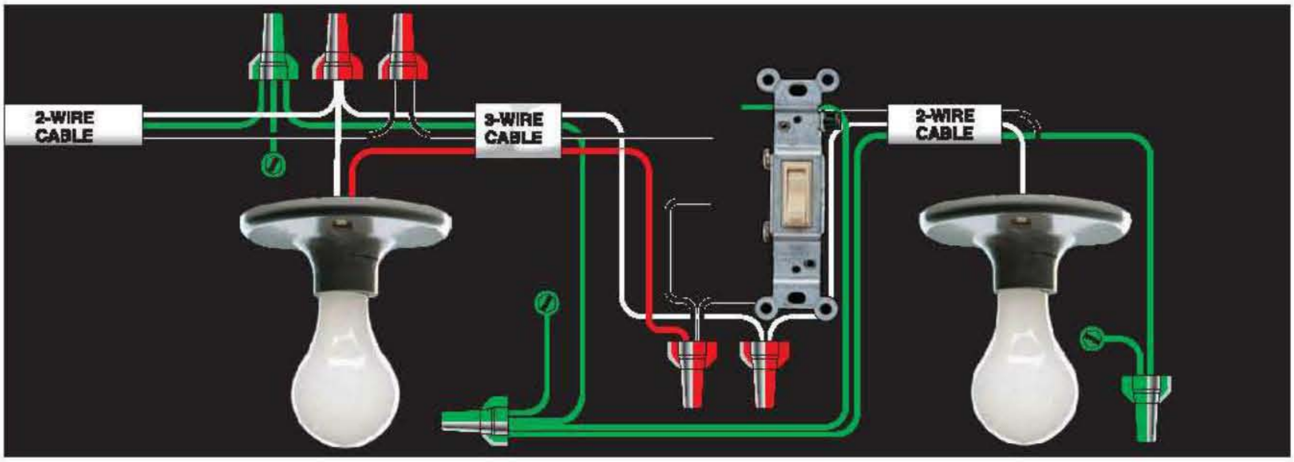
2-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE

7. SINGLE-POLE SWITCH & LIGHT FIXTURE, DUPLEX RECEPTACLE (Switch at Start of Cable Run)

Use this layout to continue a circuit past a switched light fixture to one or more duplex receptacles. To add multiple receptacles to the circuit, see circuit map 1. Requires two-wire and three-wire cables.

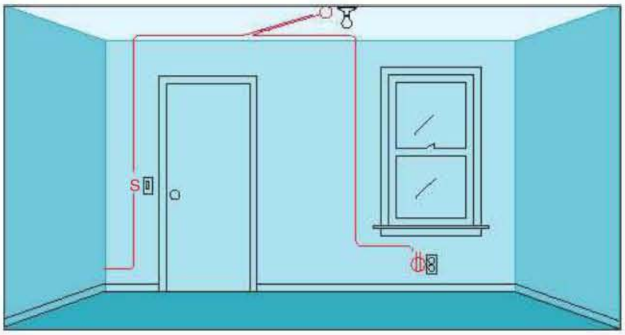

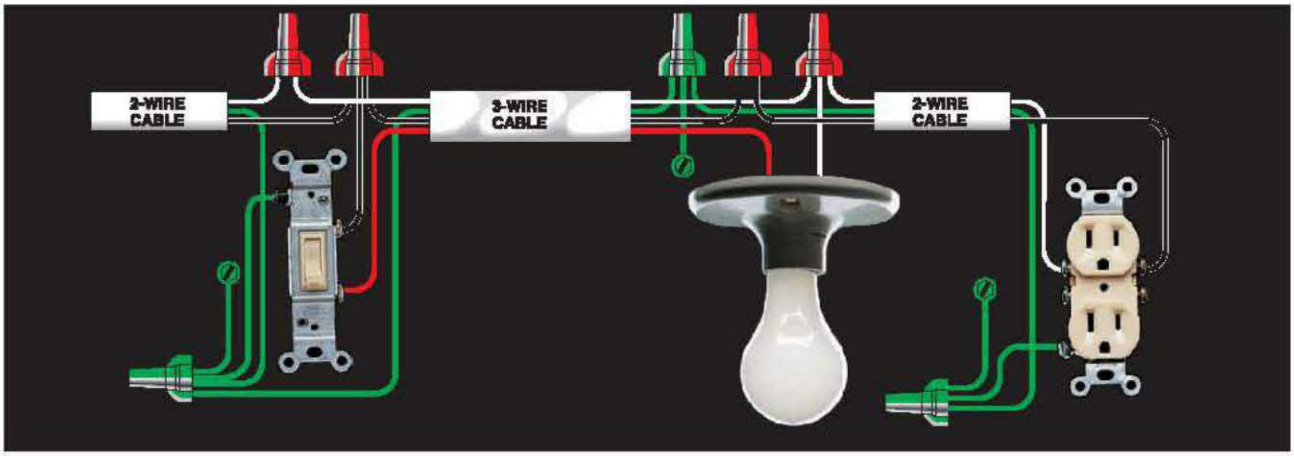
2-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE

8. DOUBLE RECEPTACLE SMALL-APPLIANCE CIRCUIT WITH GFCIs & SHARED NEUTRAL WIRE

Use this layout to wire a double receptacle circuit when code requires that some of the receptacles be GFCIs. The GFCIs should be wired for single-location protection (see circuit map 2). Requires three-wire and two-wire cables.

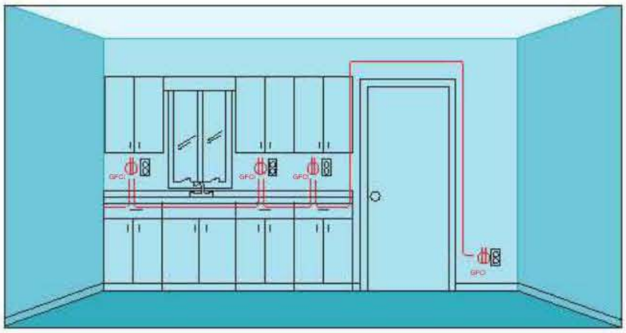
GFCI
GFCI
GFCI
GFCI

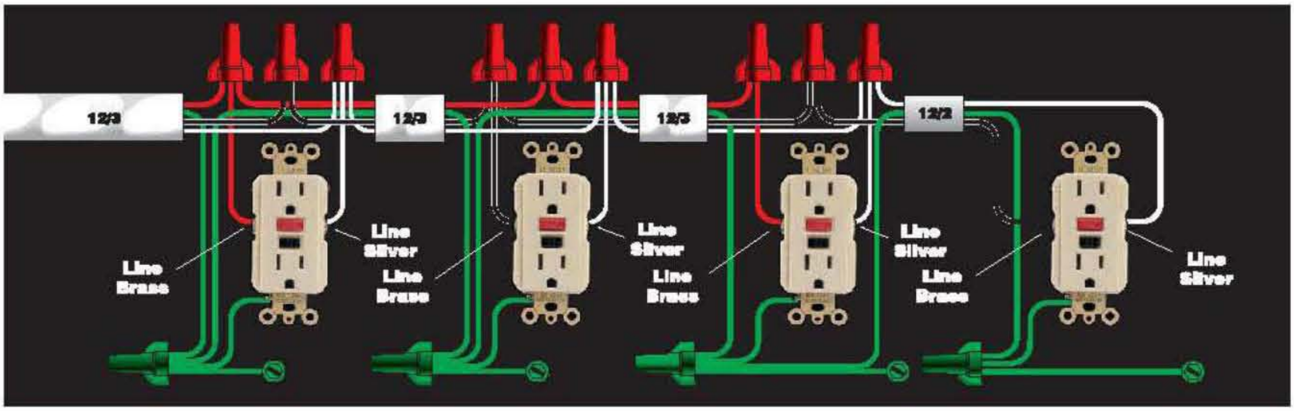
12/3
12/3
12/3
12/2
Line Brass
Line Silver
Line Brass
Line Silver
Line Brass
Line Silver
Line Brass
Line Silver

9. DOUBLE RECEPTACLE SMALL APPLIANCE CIRCUIT WITH GFCIs & SEPARATE NEUTRAL WIRES

If the room layout or local codes do not allow for a shared neutral wire, use this layout instead. The GFCIs should be wired for single-location protection (see circuit map 2). Requires two-wire cable.

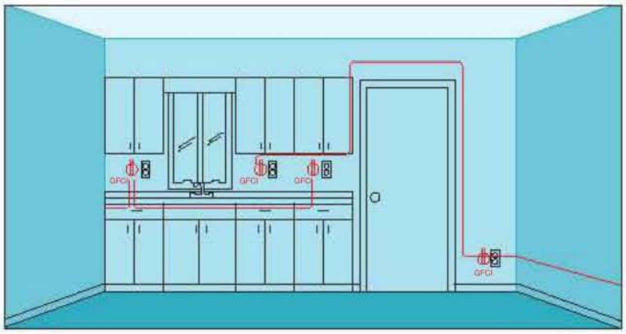
GFCI
GFCI
GFCI
GFCI

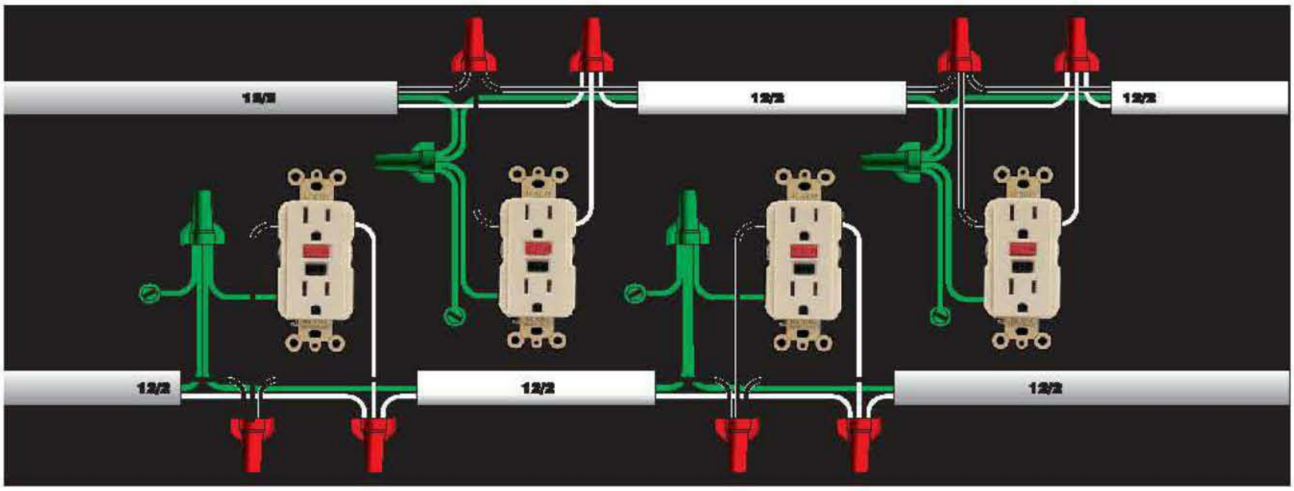
12/2
12/2
12/2
12/2
12/2
12/2

10. 120/240-VOLT RANGE RECEPTACLE

This layout is for a 50- or 60-amp, 120/240-volt dedicated appliance circuit wired with 6/3 cable, as required by code for a large kitchen range. The black and red circuit wires, connected to a double-pole circuit breaker in the circuit breaker panel, each bring 120 volts of power to the setscrew terminals on the receptacle. The white circuit wire attached to the neutral bus bar in the circuit breaker panel is connected to the neutral setscrew terminal on the receptacle.

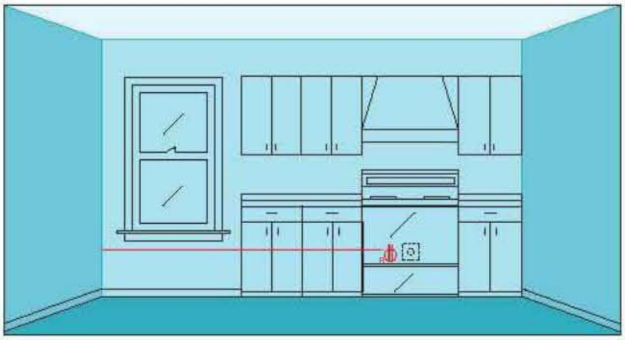

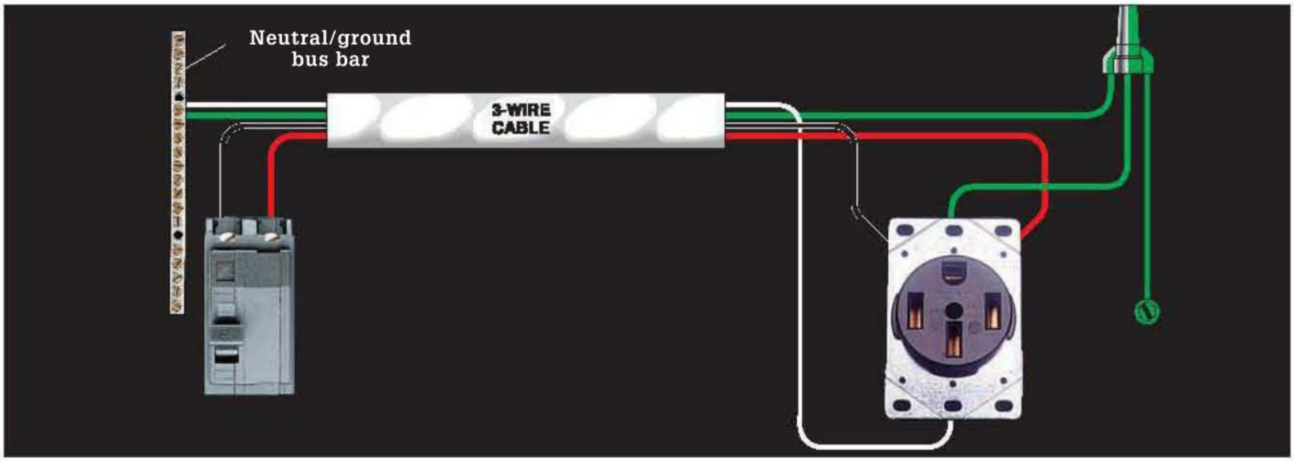
Neutral/ground bus bar
3-WIRE CABLE

11. 240-VOLT BASEBOARD HEATERS, THERMOSTAT

This layout is typical for a series of 240-volt baseboard heaters controlled by a wall thermostat. Except for the last heater in the circuit, all heaters are wired as shown below. The last heater is connected to only one cable. The size of the circuit and cables are determined by finding the total wattage of all heaters. Requires two-wire cable.

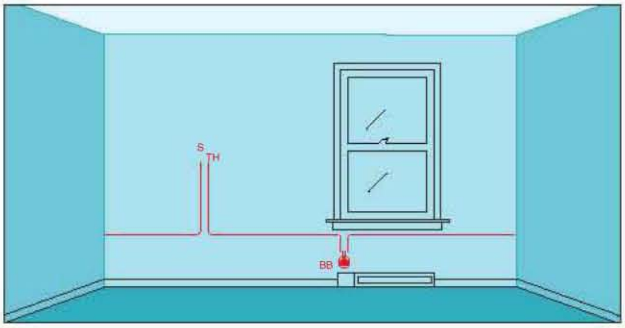

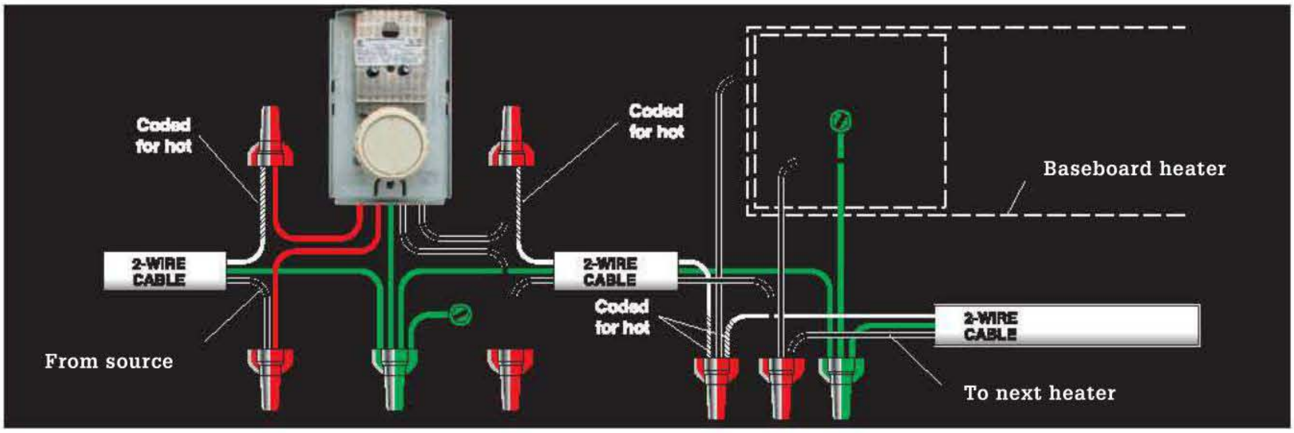
Coded for hot
Coded for hot
Baseboard heater
2-WIRE CABLE
2-WIRE CABLE
Coded for hot
2-WIRE CABLE
From source
To next heater

12. THREE-WAY SWITCHES & LIGHT FIXTURE (Fixture Between Switches)

This layout for three-way switches lets you control a light fixture from two locations. Each switch has one common screw terminal and two traveler screws. Circuit wires attached to the traveler screws run between the two switches, and hot wires attached to the common screws bring current from the power source and carry it to the light fixture. Requires two-wire and three-wire cables.

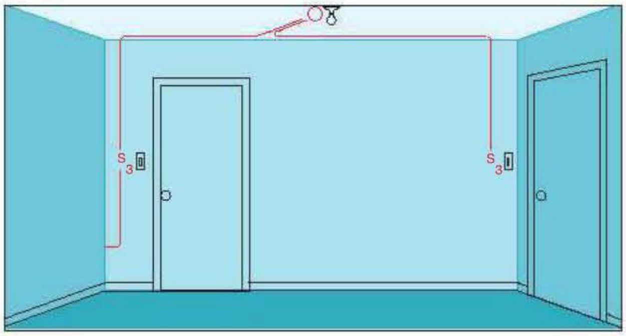

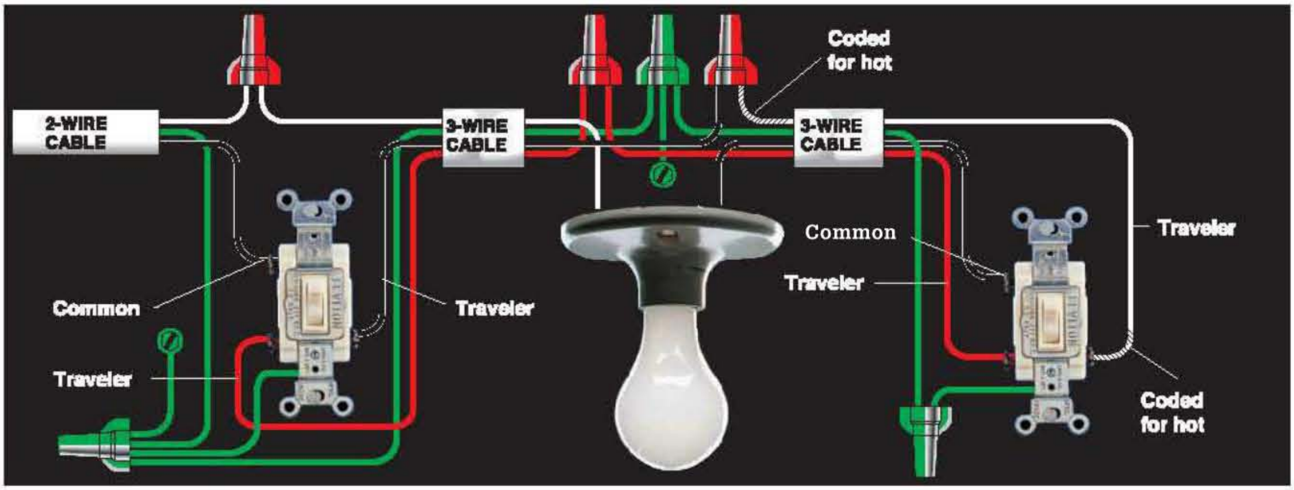
2-WIRE CABLE
3-WIRE CABLE
Coded for hot
3-WIRE CABLE
Common
Traveler
Common
Traveler
Traveler
Traveler
Coded for hot

13. THREE-WAY SWITCHES & LIGHT FIXTURE (Fixture at Start of Cable Run)

Use this layout when it is more convenient to locate the fixture ahead of the three-way switches in the cable run. Requires two-wire and three-wire cables.

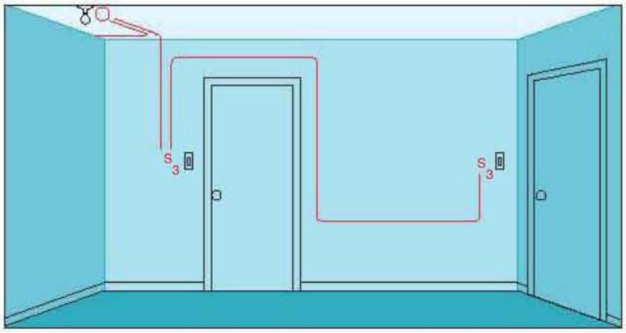

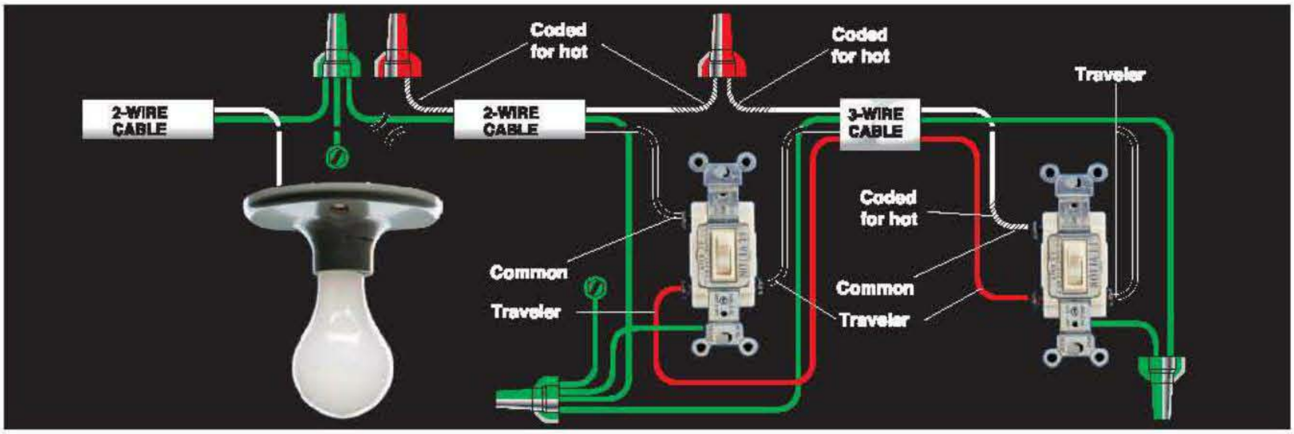
Coded for hot
Coded for hot
2-WIRE CABLE
2-WIRE CABLE
3-WIRE CABLE
Traveler
Coded for hot
Common
Traveler
Common
Traveler

14. FOUR-WAY SWITCH & LIGHT FIXTURE (Fixture at Start of Cable Run)

This layout lets you control a light fixture from three locations. The end switches are three-way and the middle is four-way. A pair of three-wire cables enter the box of the four-way switch. The white and red wires from one cable attach to the top pair of screw terminals (line 1) and the white and red wires from the other cable attaches to the bottom screw terminals (line 2). Requires two three-way switches and one four-way switch and two-wire and three-wire cables.

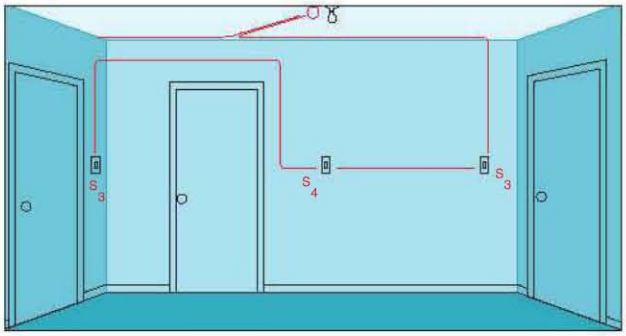

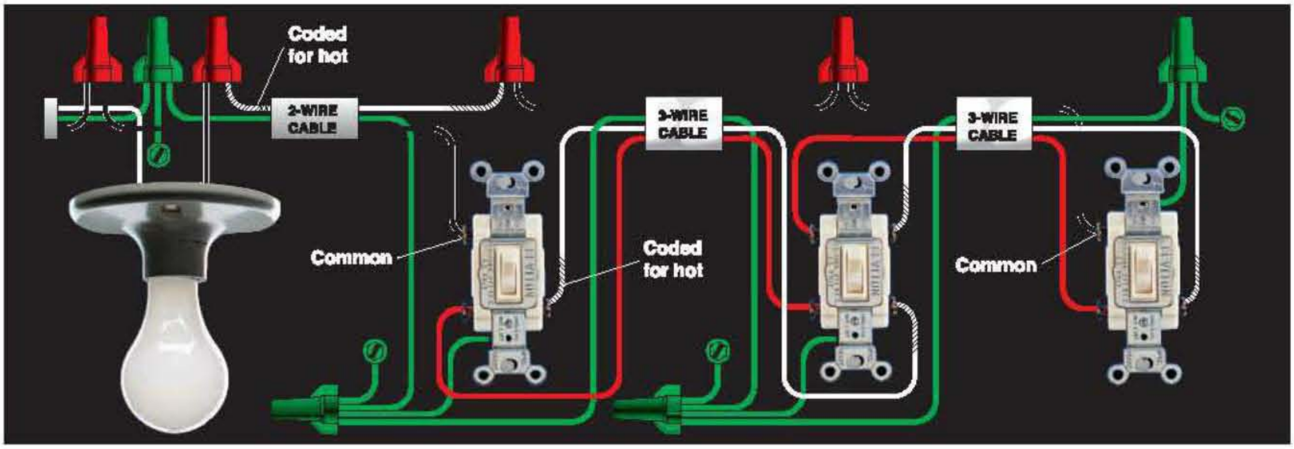

Coded for hot
2-WIRE CABLE
3-WIRE CABLE
3-WIRE CABLE
Common
Coded for hot
Common

15. FOUR-WAY SWITCH & LIGHT FIXTURE (Fixture at End of Cable Run)

Use this layout when it is more practical to locate the fixture at the end of the cable run. Requires two three-way switches and one four-way switch and two-wire and three-wire cables.

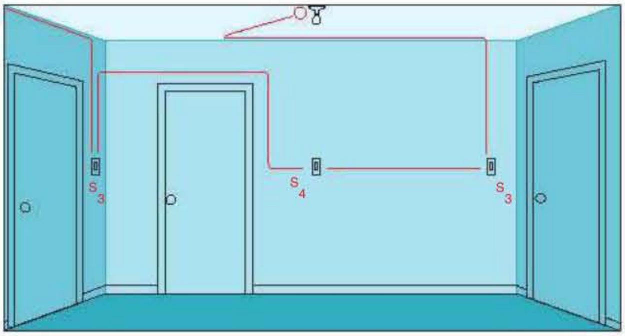

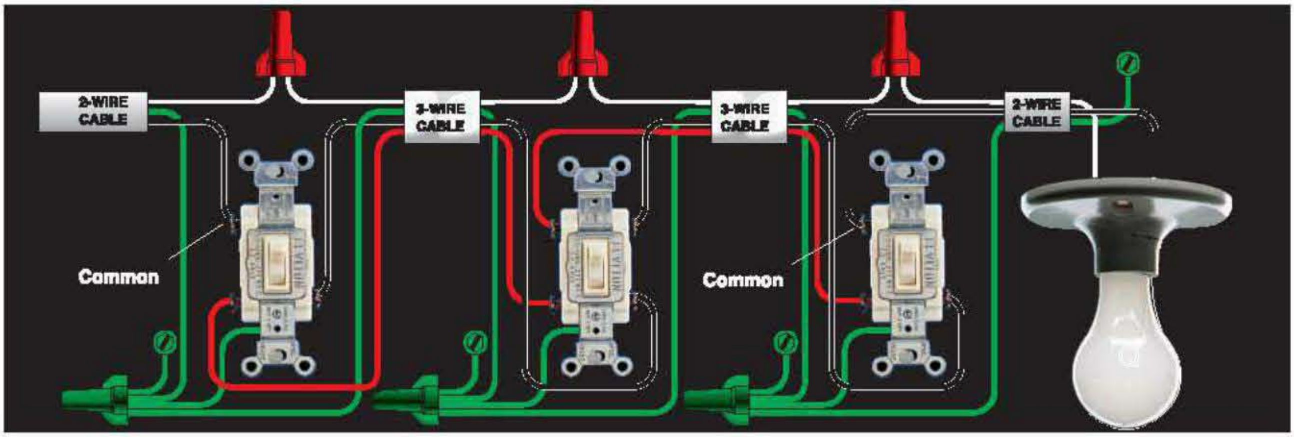

2-WIRE CABLE
3-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE
Common
Common

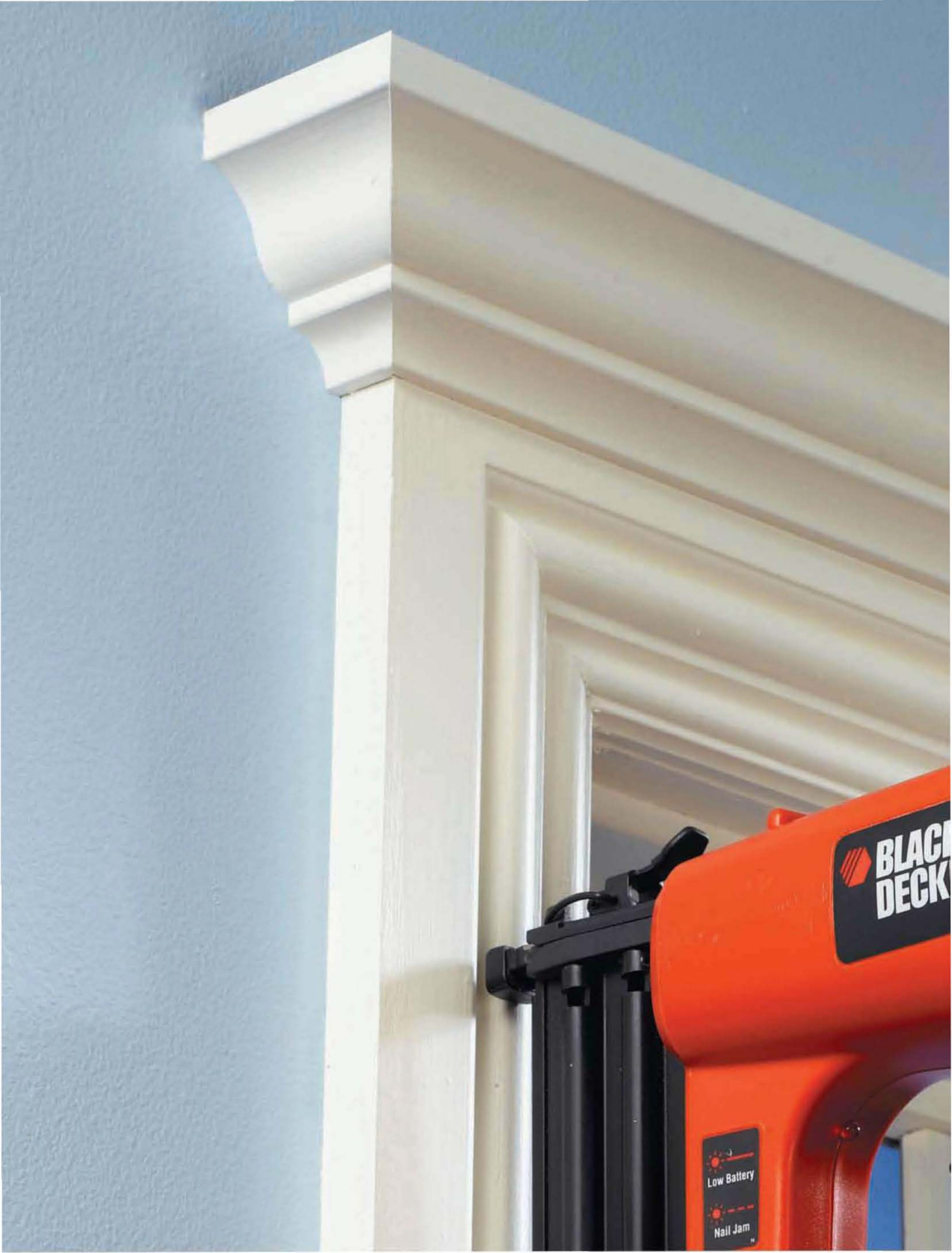
BLAC
DECK
Low Battery
Nail Jam

HOME IMPROVEMENT BASICS

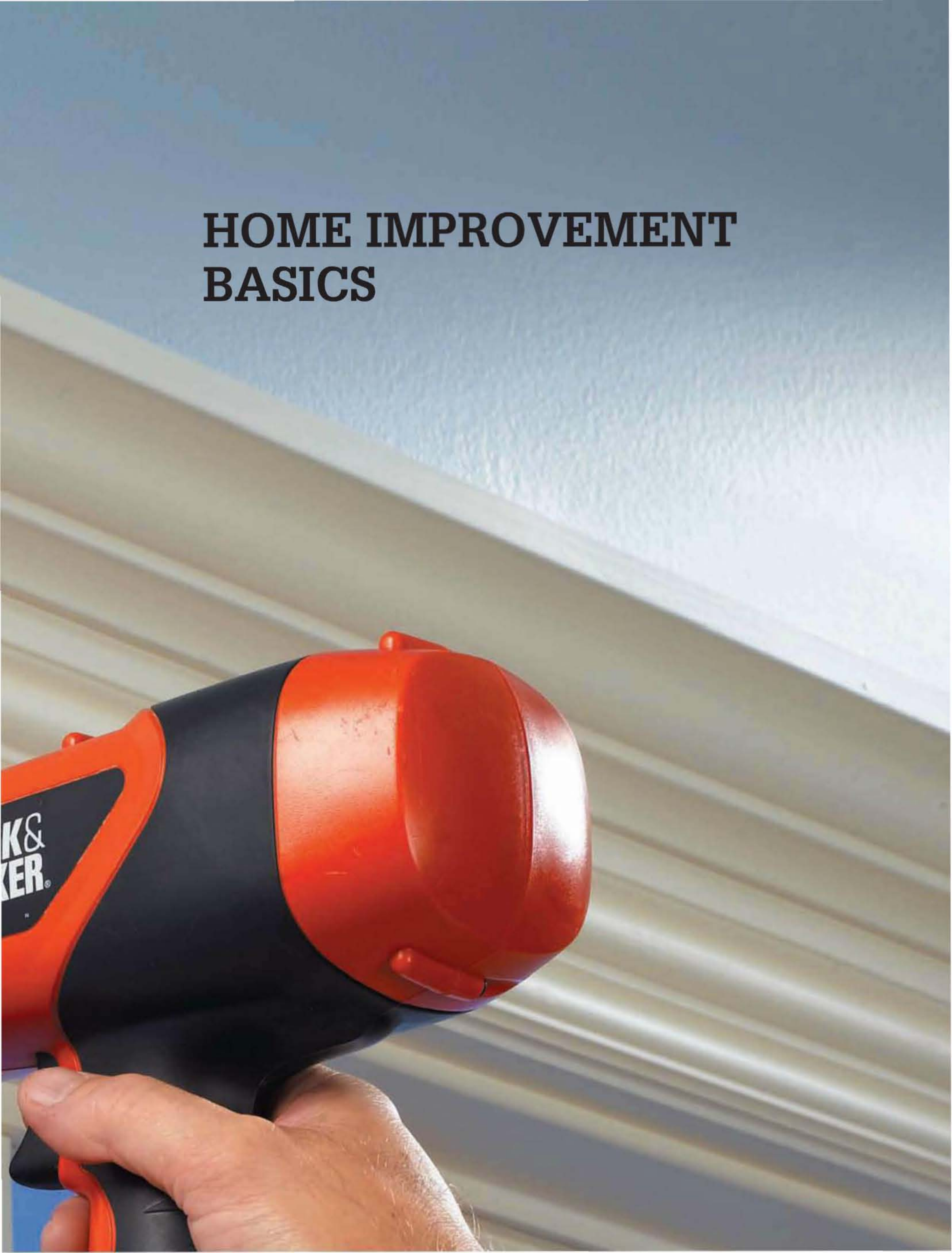

Flooring

Like any successful remodeling project, replacing your floor covering requires detailed planning and attention to design. Flooring is not separate from the rest of the room; it should fit into the overall design to create a desired effect. A floor can create excitement and become a focal point or it can serve as a background for the rest of the room.

Through careful planning, you can choose flooring that can be used successfully in multiple rooms, or select a pattern or design that is repeated throughout the room or in adjacent rooms.

Keep in mind that your flooring design will last a long time, especially if you install ceramic tile or wood. In most cases, the only way to change the design of your floor is to install a new floor covering.

In this chapter:

- Installing Underlayment
- Installing Radiant Floor Mats
- Installing Hardwood Strip Flooring
- Installing Laminate Floors
- Installing Parquet Flooring
- Installing Sheet Vinyl
- Evaluating an Existing Floor
- Installing Resilient Tile
- Installing Bamboo Flooring
- Installing Ceramic Tile
- Installing Mosaic Tile
- Installing Carpet Squares
- Refinishing Wood Floors

Installing Underlayment

Underlayment is a layer of sheeting screwed or nailed to the subfloor to provide a smooth, stable surface for the floor covering. The type of underlayment you choose depends in part on the type of floor covering you plan to install. Ceramic and natural stone tile floors usually require an underlayment that stands up to moisture, such as cementboard. For vinyl flooring, use a quality-grade plywood; most warranties are void if the flooring is installed over substandard underlayments. If you want to use your old flooring as underlayment, apply an embossing leveler to prepare it for the new installation. Most wood flooring and carpeting do not require underlayment and are often placed directly on a plywood subfloor.

When you install new underlayment, attach it securely to the subfloor in all areas, including under movable appliances. Notch the underlayment to fit the room's contours. Insert the underlayment beneath door casings and moldings. Once the underlayment is installed, use a latex patching compound to fill gaps, holes, and low spots. This compound is also used to cover screw heads, nail heads, and seams in underlayment. Some compounds include dry and wet ingredients that need to be mixed, while others are premixed. The compound is applied with a trowel or wallboard knife.

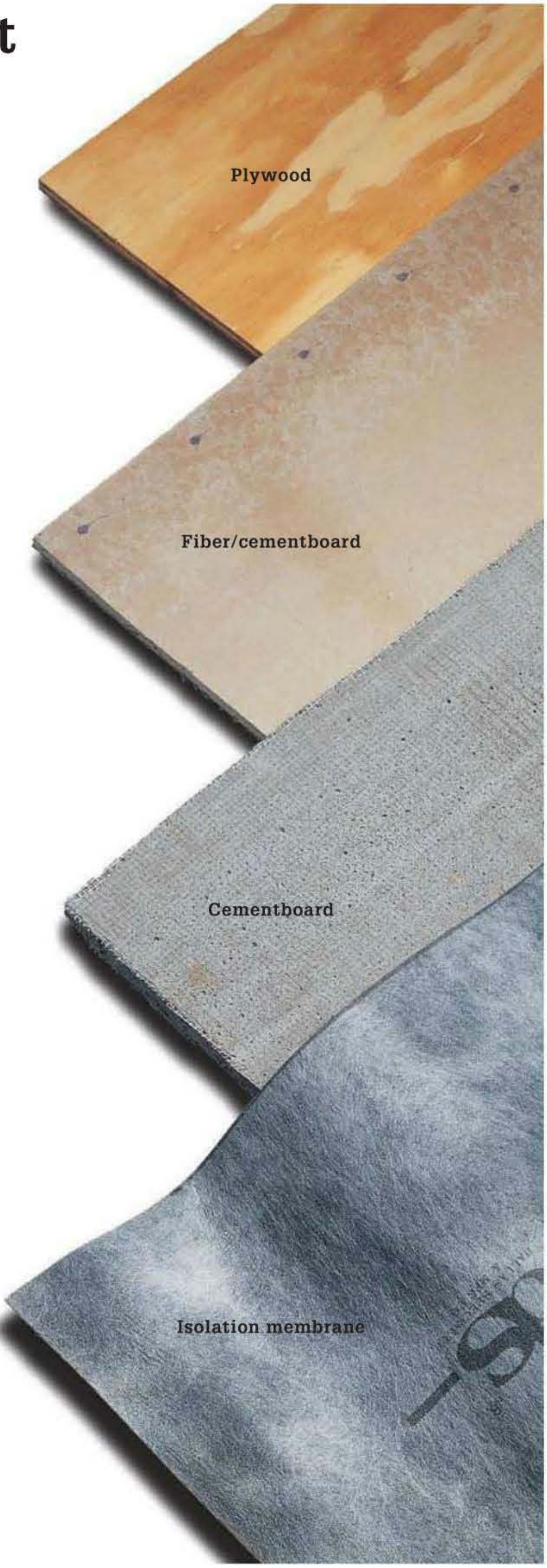

Tools & Materials ▸

- Drill
- Circular saw
- Wallboard knife
- Power sander
- ¼" notched trowel
- Straightedge
- Utility knife
- Jigsaw with carbide-tipped blade
- ⅛" notched trowel
- Flooring roller
- Underlayment
- 1" deck screws
- Floor-patching compound
- Latex additive
- Thin-set mortar
- 1½" galvanized deck screws
- Fiberglass-mesh wallboard tape

How to Install Plywood Underlayment

Plywood is the most common underlayment for vinyl flooring and some ceramic tile installations. For vinyl, use ¼" exterior-grade, AC plywood. This type has one smooth side for a quality surface. Wood-based floor coverings, like parquet, can be installed over lower-quality exterior-grade plywood. For ceramic tile, use ½" AC plywood. When installing plywood, leave ¼" expansion gaps at the walls and between sheets.

1

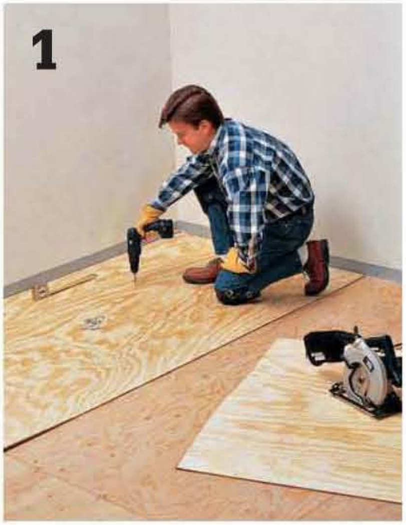

Install a full sheet of plywood along the longest wall, making sure the underlayment seams are not aligned with the subfloor seams. Fasten the plywood to the subfloor using 1" deck screws driven every 6" along the edges and at 8" intervals in the field of the sheet.

2

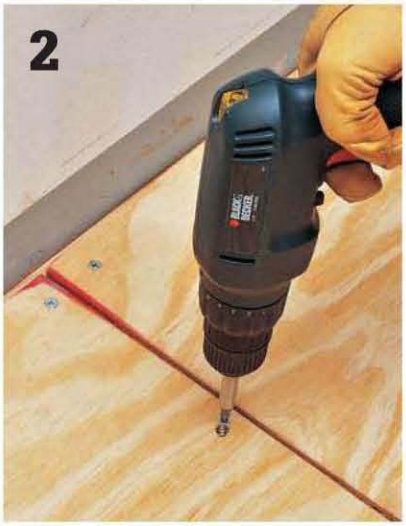

Continue fastening sheets of plywood to the subfloor, driving the screw heads slightly below the underlayment surface. Leave ¼" expansion gaps at the walls and between sheets. Offset seams in subsequent rows.

3

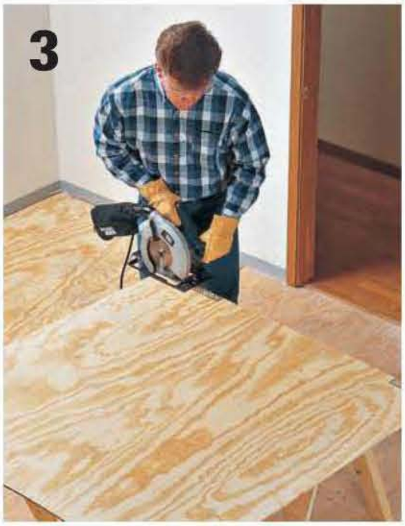

Using a circular saw or jigsaw, notch the plywood to meet the existing flooring in doorways. Fasten the notched sheets to the subfloor.

4

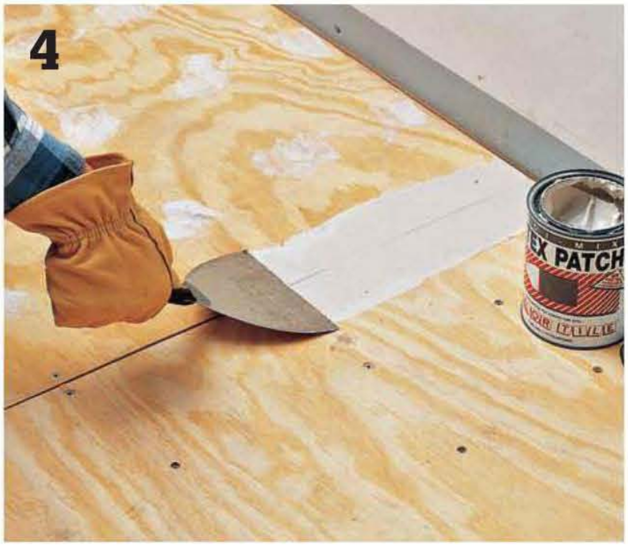

Mix floor-patching compound and latex or acrylic additive following the manufacturer's directions. Spread it over seams and screw heads, using a wallboard knife.

5

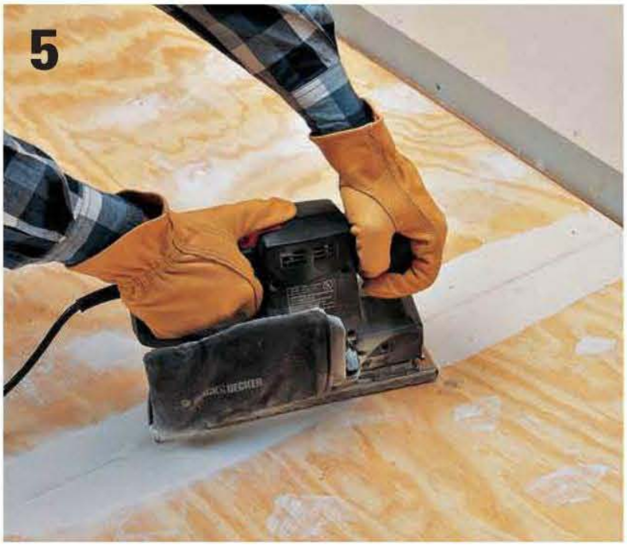

Let the patching compound dry, then sand the patched areas, using a power sander.

How to Install Cementboard

Ceramic and natural stone tile floors usually require an underlayment that stands up to moisture, such as cementboard. Fiber/cementboard is a thin, high-density underlayment used under ceramic tile and vinyl flooring in situations where floor height is a concern. Cementboard is used only for ceramic tile or stone tile installations. It remains stable even when wet, so it is the best underlayment to use in areas that are likely to get wet, such as bathrooms. Cementboard is more expensive than plywood, but a good investment for a large tile installation.

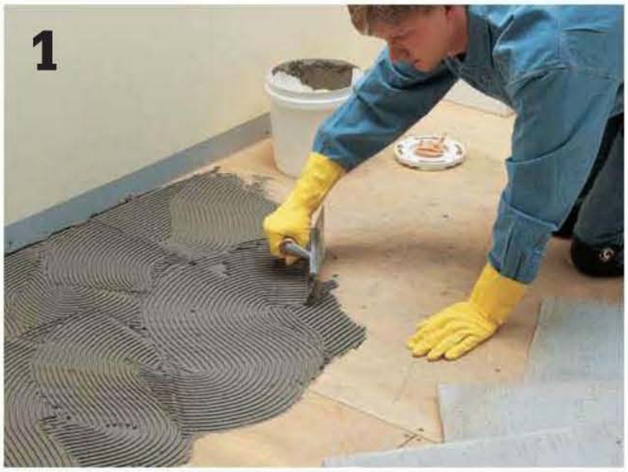

Mix thin-set mortar according to the manufacturer's directions. Starting at the longest wall, spread the mortar on the subfloor in a figure eight pattern using a ¼" notched trowel. Spread only enough mortar for one sheet at a time. Set the cementboard on the mortar with the rough side up, making sure the edges are offset from the subfloor seams.

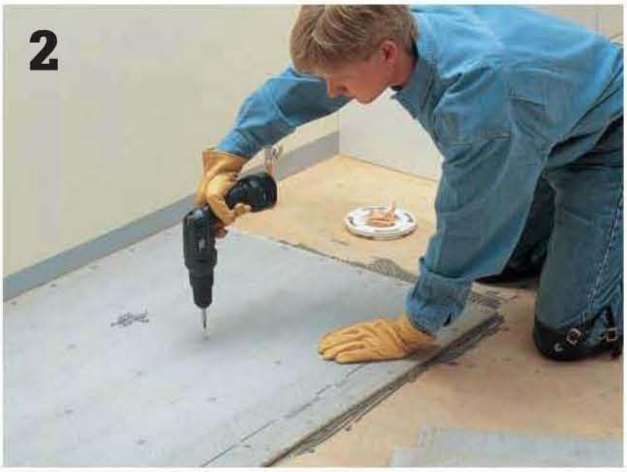

Fasten the cementboard to the subfloor using 1¼" cementboard screws driven every 6" along the edges and 8" throughout the sheet. Drive the screw heads flush with the surface. Continue spreading mortar and installing sheets along the wall. *Option: If installing fiber/cementboard underlayment, use a 3⁄16" notched trowel to spread the mortar, and drill pilot holes for all screws.*

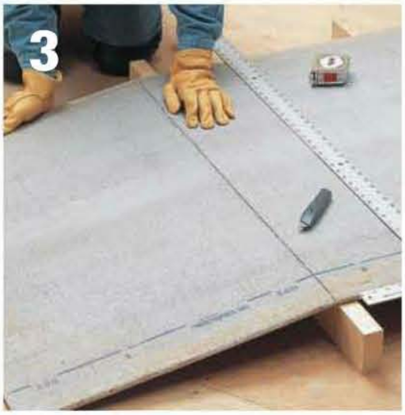

Cut cementboard pieces as necessary, leaving an ⅛" gap at all joints and a ¼" gap along the room perimeter. For straight cuts, use a utility knife to score a line through the fiber-mesh layer just beneath the surface, then snap the board along the scored line.

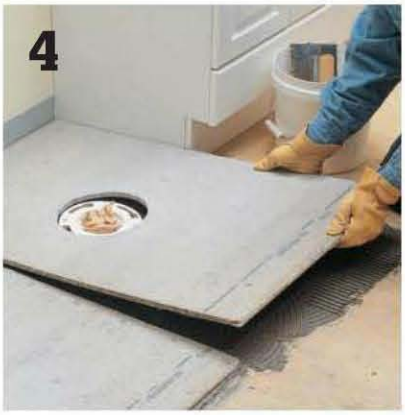

To cut holes, notches, or irregular shapes, use a jigsaw with a carbide-tipped blade. Continue installing cementboard sheets to cover the entire floor.

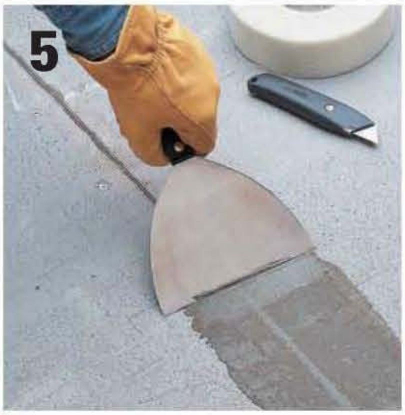

Place fiberglass-mesh wallboard tape over the seams. Use a wallboard knife to apply thin-set mortar to the seams, filling the gaps between sheets and spreading a thin layer of mortar over the tape. Allow the mortar to set for two days before starting the tile installation.

How to Install Isolation Membrane

Isolation membrane is used to protect ceramic tile installations from movement that may occur on cracked concrete floors. This product is used primarily for covering individual cracks, but it can be used over an entire floor. Isolation membrane is also available in a liquid form that can be poured over the project area.

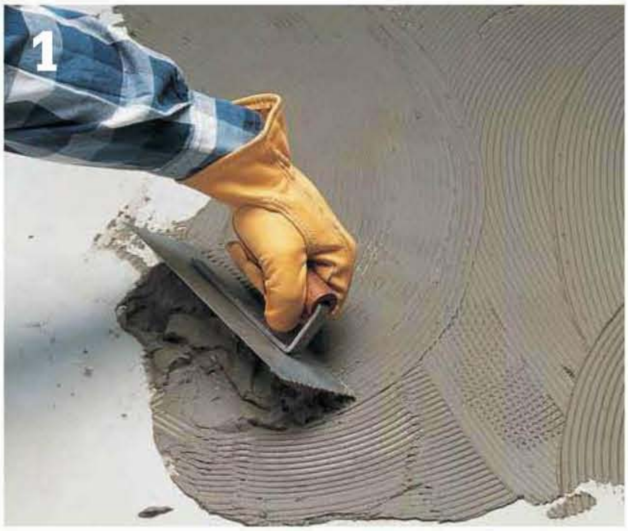

1

Thoroughly clean the subfloor, then apply thin-set mortar with a ⅛" notched trowel. Start spreading the mortar along a wall in a section as wide as the membrane and 8 to 10 ft. long. *Note: For some membranes, you must use a bonding material other than mortar. Read and follow manufacturer's directions.*

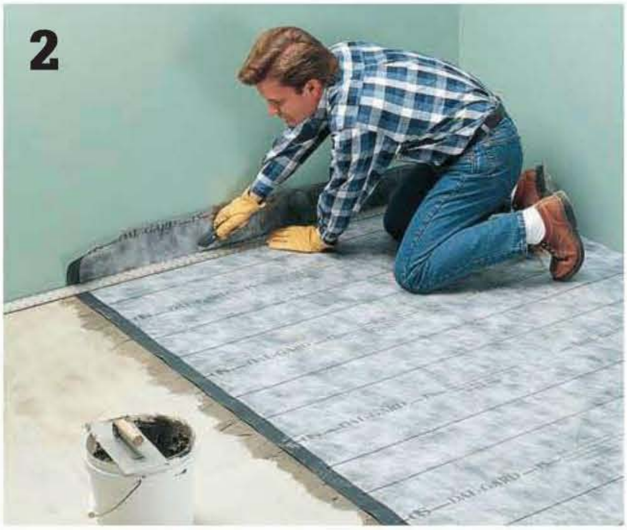

2

Roll out the membrane over the mortar. Cut the membrane to fit tightly against the walls, using a straightedge and utility knife.

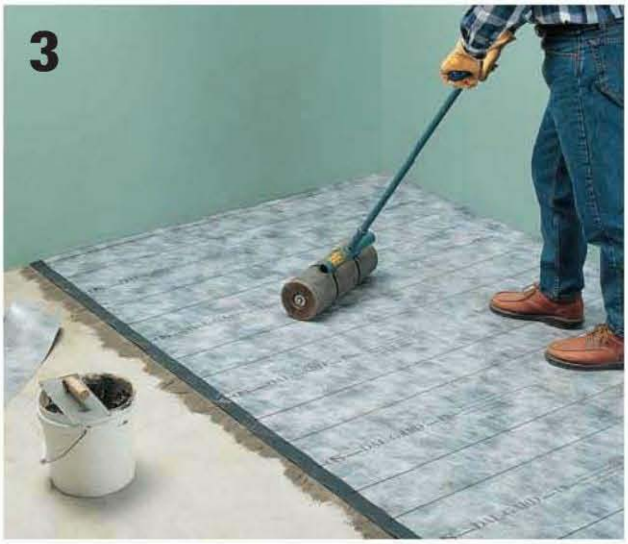

3

Starting in the center of the membrane, use a heavy floor roller to smooth out the surface toward the edges. This frees trapped air and presses out excess bonding material.

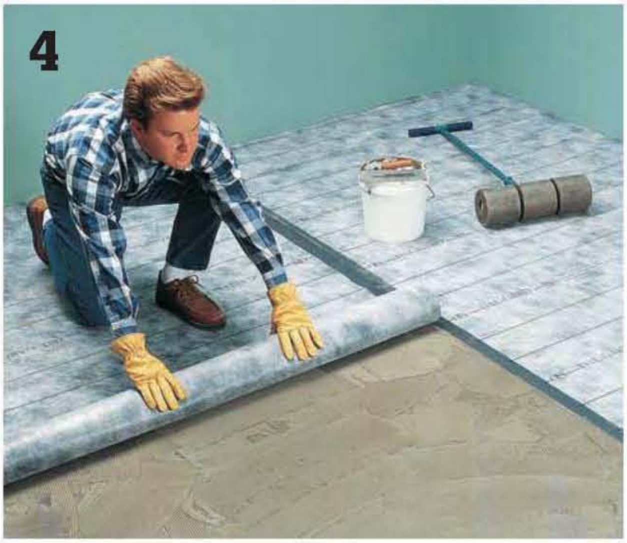

4

Repeat steps 1 through 3, cutting the membrane as necessary at the walls and obstacles, until the floor is completely covered with membrane. Do not overlap the seams, but make sure they're tight. Allow the mortar to set for two days before installing the tile.

Installing Raised Underlayment Panels

Concrete floors are practical and durable—and generally cold and uncomfortable. For a fast and easy makeover, you can now find raised underlayment panels that simply rest on the concrete and provide a surface for other flooring materials. The tongue-and-groove plywood panels have dimpled plastic on the bottom. This allows air to circulate underneath so that the concrete stays dry, and insulates the flooring above. The assembled panels can support laminates and resilient sheets or tiles. And you can install them in a weekend.

How to Install Raised Underlayment Panels

1

Start at one corner, and measure the length and width of the room from that starting point. Calculate the number of panels you will need to cover the space in both directions. If the starting corner is not square, trim the first row of panels to create a straight starting line.

2

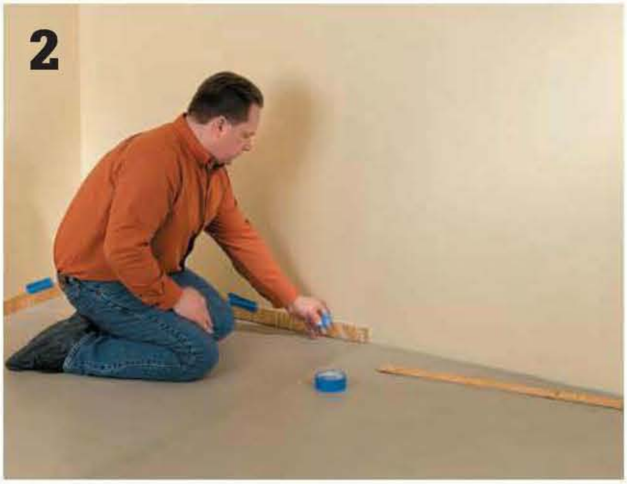

Create an expansion gap around the edges. Place ¼" spacers at all walls, doors, and other large obstacles. To make your own spacers, cut sheets of ¼" plywood to the thickness of the panels, and hold them in place temporarily with masking tape.

3

Dry-lay a row of panels across the room. If the last row will be less than 6" wide, balance it by trimming the first panel or the starting row, if necessary, to account for the row end pieces.

4

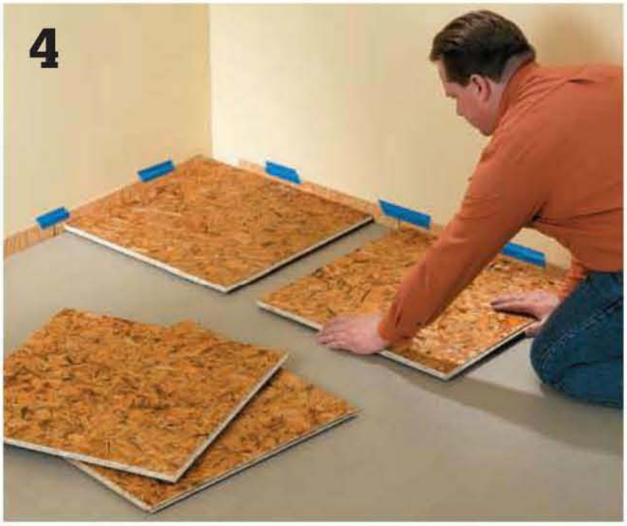

Starting in the corner, lay the first panel with the grooved side against the ¼" spacers. Slide the next panel into place and press-fit the groove of the second panel into the tongue of the first. Check the edges against the wall.

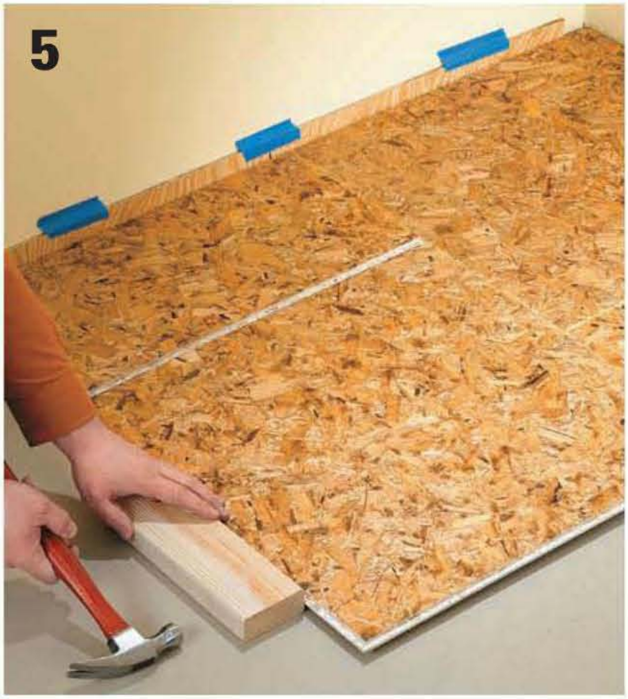

Repeat these steps to complete the first row. If necessary, tap the panels into place with a scrap piece of lumber and a rubber mallet or hammer—just be careful not to damage the tongue or groove edges. Starting with the second row, stagger the seams so that the panels interlock.

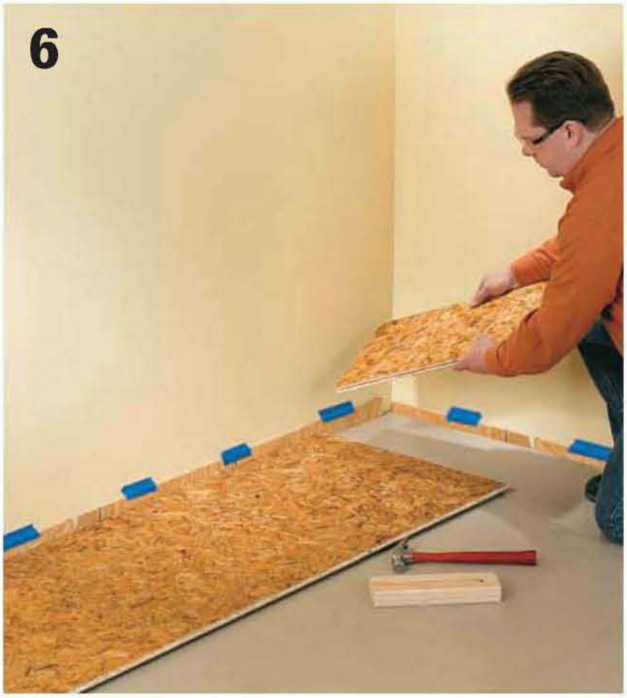

Cut the last panel to fit snugly between the next-to-last panel and the ¼" spacer on the far wall. Install the last panel at an angle and tap it down. Continue working from the starting point, checking after each row to be sure the panels are square and level.

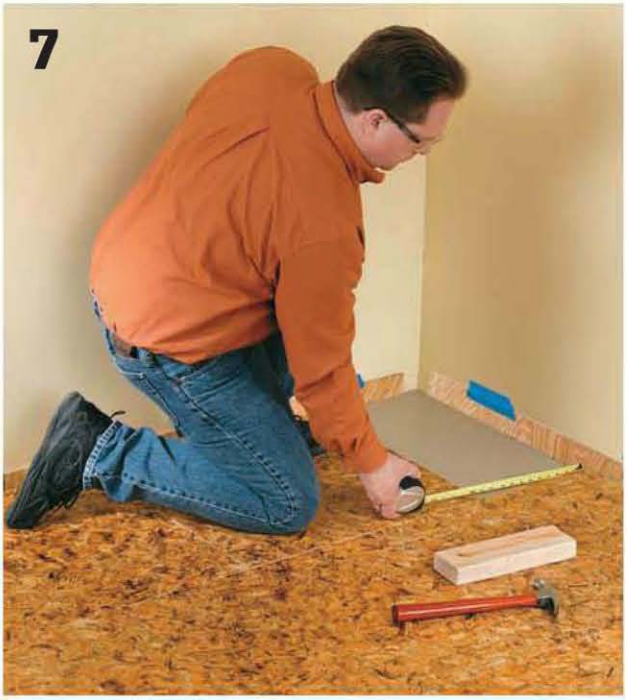

When you reach the last row and last panel to complete your installation, you may have to cut the panel to fit. Measure for fit, allowing for the ¼" expansion gap from the wall. Cut the panel and fit it into place.

When all the panels are in place and the finished floor is installed, remove the spacers from around the perimeter of the room.

Installing Radiant Floor Mats

Floor-warming systems require very little energy to run and are designed to heat floors only; they generally are not used as sole heat sources for rooms.

A typical floor-warming system consists of one or more thin mats containing electric resistance wires that heat up when energized like an electric blanket. The mats are installed beneath the tile and are hardwired to a 120-volt GFCI circuit. A thermostat controls the temperature, and a timer turns the system off automatically.

The system shown in this project includes two plastic mesh mats, each with its own power lead that is wired directly to the thermostat. Radiant mats may be installed over a plywood subfloor, but if you plan to install floor tile you should put down a base of cementboard first, and then install the mats on top of the cementboard.

A crucial part of installing this system is to use a multimeter to perform several resistance checks to make sure the heating wires have not been damaged during shipping or installation.

Electrical service required for a floor-warming system is based on size. A smaller system may connect to an existing GFCI circuit, but a larger one will need a dedicated circuit; follow the manufacturer's requirements.

To order a floor-warming system, contact the manufacturer or dealer (see Resources, page 554). In most cases, you can send them plans and they'll custom-fit a system for your project area.

Tools & Materials ▸

- Vacuum cleaner
- Multimeter
- Tape measure
- Scissors
- Router/rotary tool
- Marker
- Trowel or rubber float
- Notched trowel
- Staple gun
- Hot glue gun
- Electric wire fault indicator (optional)
- Radiant floor mats
- 12/2 NM cable
- Conduit
- Wire connectors
- Thinset mortar
- Thermostat with sensor
- Junction box(es)
- Tile or stone floorcovering

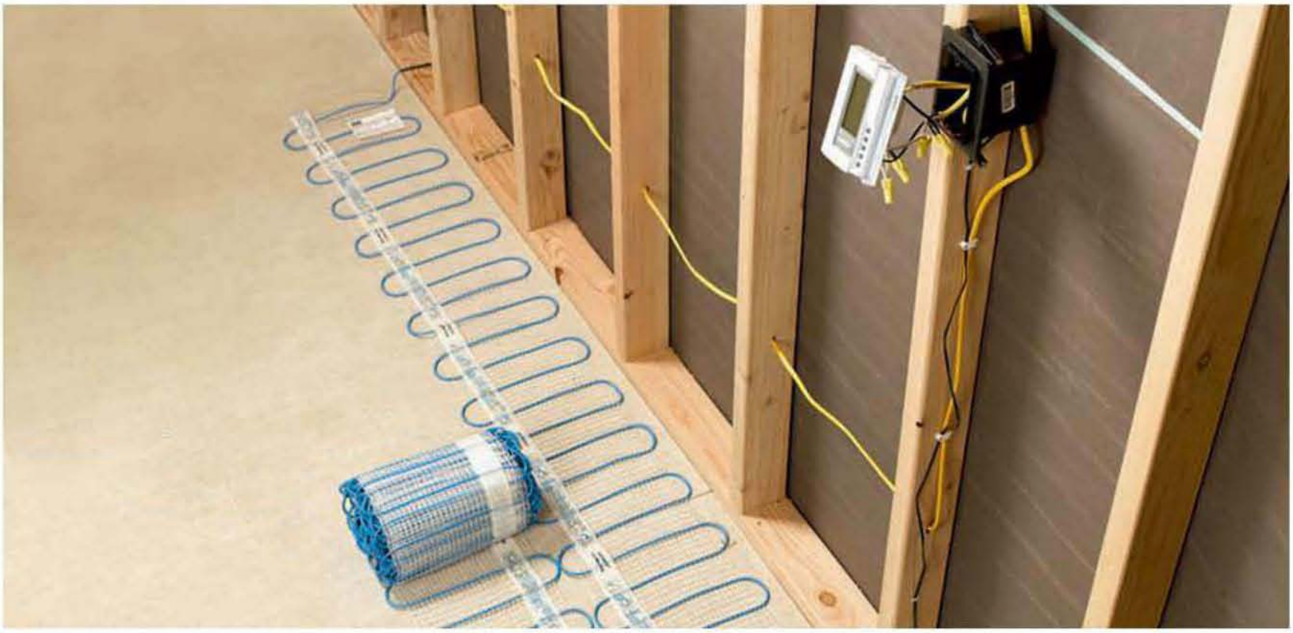

A radiant floor-warming system employs electric heating mats that are covered with floor tile to create a floor that's cozy under foot.

Installation Tips ▸

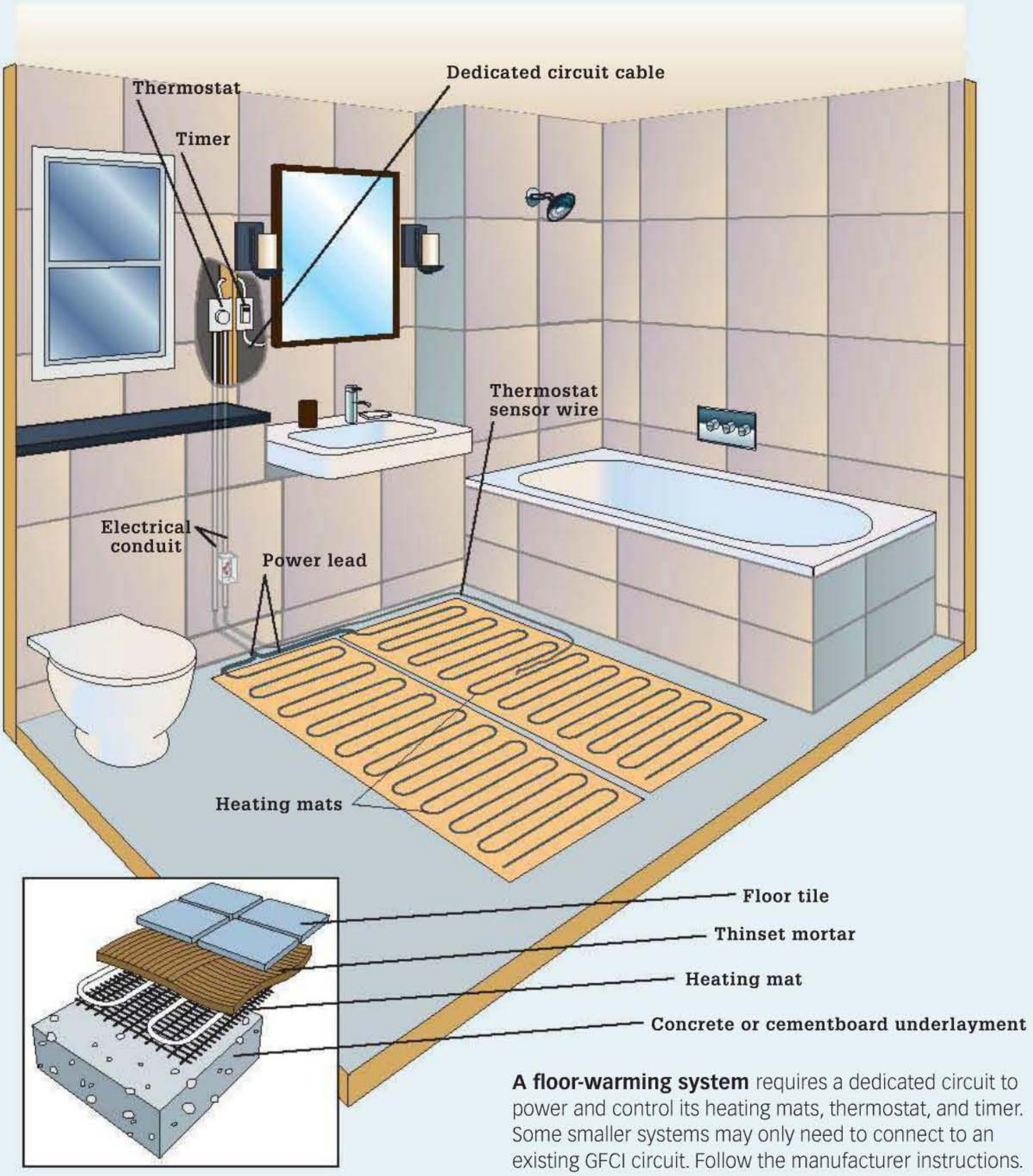
Thermostat
Timer
Dedicated circuit cable
Thermostat sensor wire
Electrical conduit
Power lead
Heating mats
Floor tile
Thinset mortar
Heating mat
Concrete or cementboard underlayment

A floor-warming system requires a dedicated circuit to power and control its heating mats, thermostat, and timer. Some smaller systems may only need to connect to an existing GFCI circuit. Follow the manufacturer instructions.

- Each radiant mat must have a direct connection to the power lead from the thermostat, with the connection made in a junction box in the wall cavity. Do not install mats in series.
- Do not install radiant floor mats under shower areas.
- Do not overlap maps or let them touch.
- Do not cut heating wire or damage heating wire insulation.
- The distance between wires in adjoining mats should equal the distance between wire loops measured center to center.

Installing a Radiant Floor-warming System

10,000 AIC
OFF
ON
PUSH TO TEST

Floor-warming systems must be installed on a circuit with adequate amperage and a GFCI breaker. Smaller systems may tie into an existing circuit, but larger ones need a dedicated circuit. Follow local building and electrical codes that apply to your project.

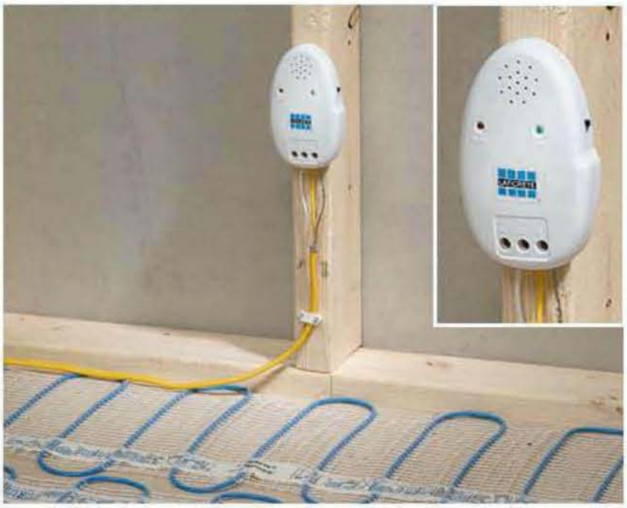

An electric wire fault indicator monitors each floor mat for continuity during the installation process. If there is a break in continuity (for example, if a wire is cut) an alarm sounds. If you choose not to use an installation tool to monitor the mat, test for continuity frequently using a multimeter.

How To Install a Radiant Floor-warming System

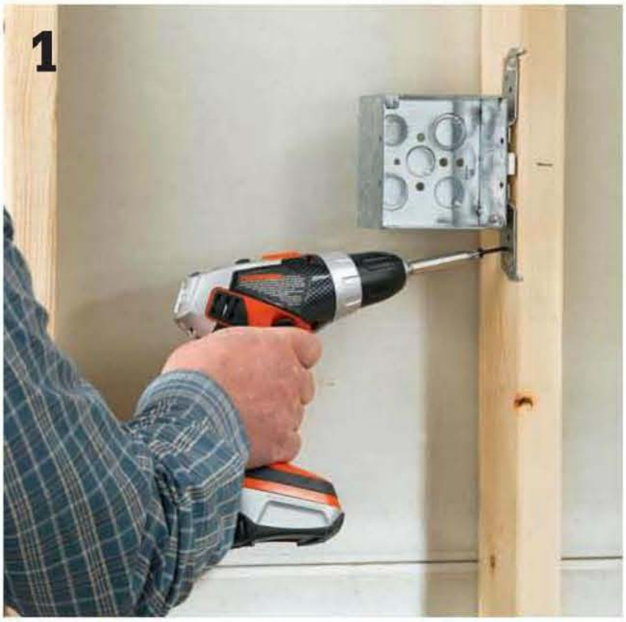

Install electrical boxes to house the thermostat and timer. In most cases, the box should be located 60" above floor level. Use a 4"-deep × 4"-wide double-gang box for the thermostat/timer control if your kit has an integral model. If your timer and thermostat are separate, install a separate single box for the timer.

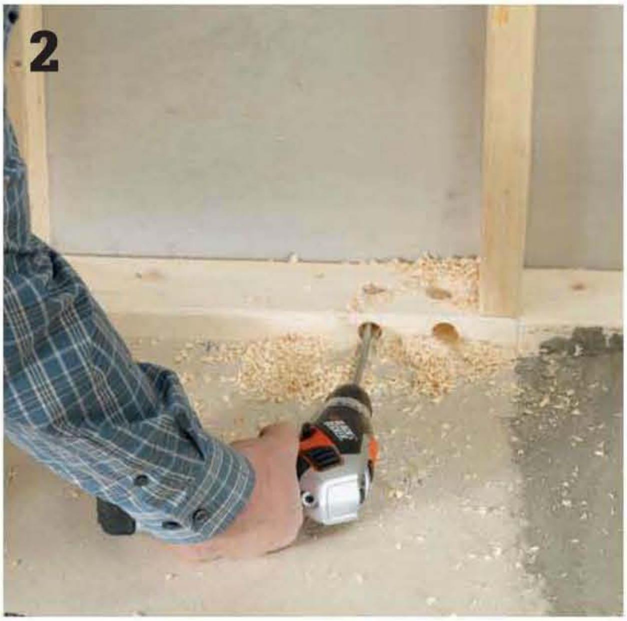

Drill access holes in the sole plate for the power leads that are preattached to the mats (they should be over 10 ft. long). The leads should be connected to a supply wire from the thermostat in a junction box located in a wall near the floor and below the thermostat box. The access hole for each mat should be located directly beneath the knockout for that cable in the thermostat box. Drill through the sill plate vertically and horizontally so the holes meet in an L-shape.

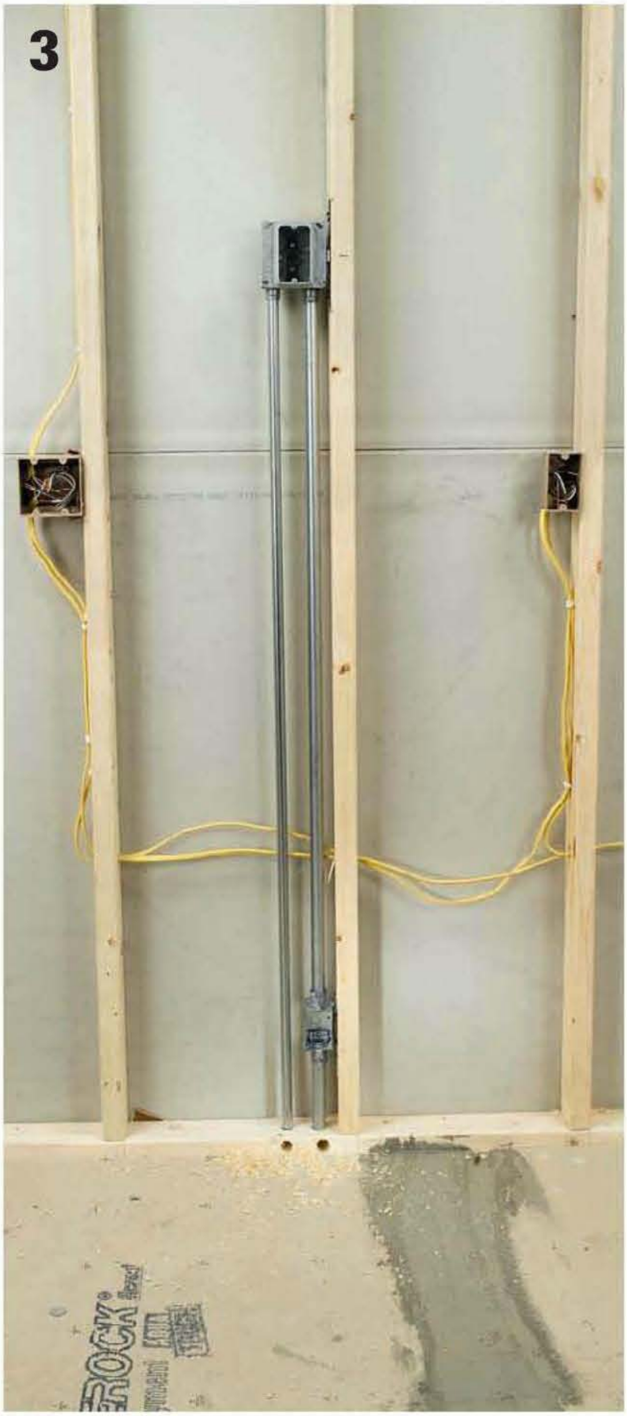

Run conduit from the electrical boxes to the sill plate. The line for the supply cable should be ¾" conduit. If you are installing multiple mats, the supply conduit should feed into a junction box about 6" above the sill plate and then continue into the ¾" hole you drilled for the supply leads. The sensor wire needs only ½" conduit that runs straight from the thermostat box via the thermostat. The mats should be powered by a dedicated 20-amp GFCI circuit of 12/2 NM cable run from your main service panel to the electrical box (this is for 120-volt mats—check your instruction manual for specific circuit recommendations).

Clean the floor surface thoroughly to get rid of any debris that could potentially damage the wire mats. A vacuum cleaner generally does a more effective job than a broom.

Test for resistance using a multimeter set to measure ohms. This is a test you should make frequently during the installation, along with checking for continuity. If the resistance is off by more than 10% from the theoretical resistance listing (see manufacturer's chart in installation instructions), contact a technical support operator for the kit manufacturer. For example, the theoretical resistance for the 1 × 50 ft. mat seen here is 19, so the ohms reading should be between 17 and 21.

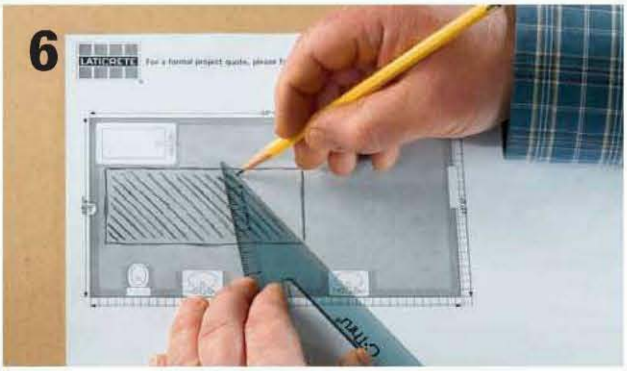

Finalize your mat layout plan. Most radiant floor warming mat manufacturers will provide a layout plan for you at the time of purchase, or they will give you access to an online design tool so you can come up with your own plan. This is an important step to the success of your project, and the assistance is free.

(continued)

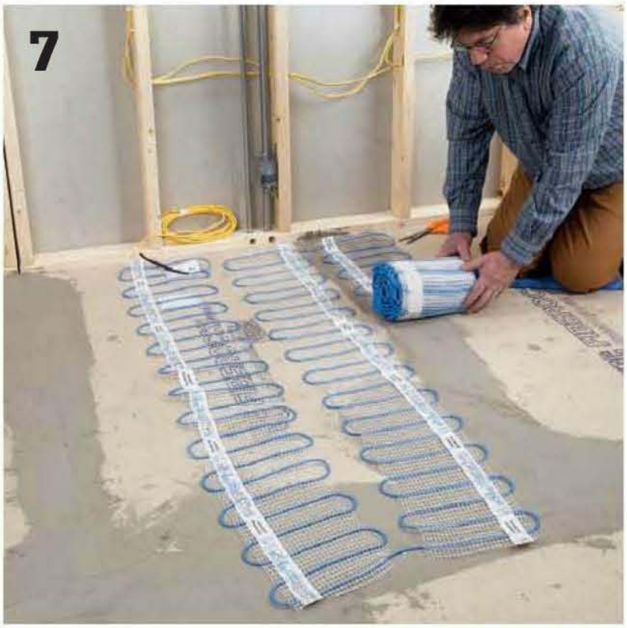

Unroll the radiant mat or mats and allow them to settle. Arrange the mat or mats according to the plan you created. It's okay to cut the plastic mesh so you can make curves or switchbacks, but do not cut the heating wire under any circumstances, not even to shorten it.

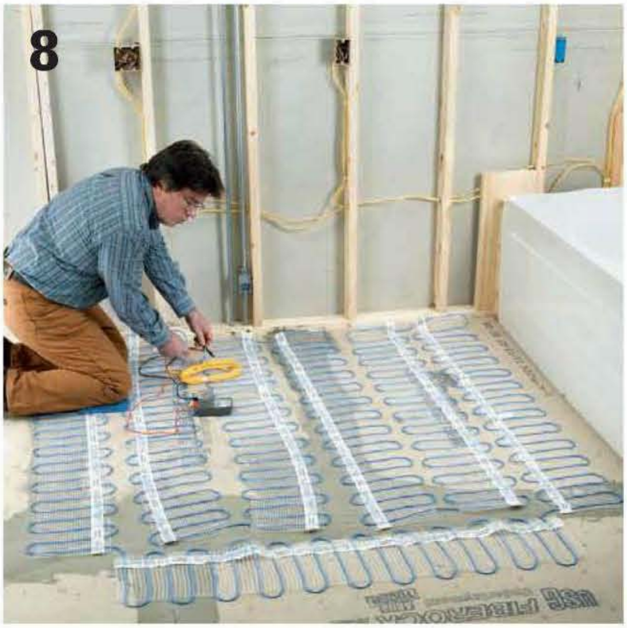

Finalize the mat layout and then test the resistance again using a multimeter. Also check for continuity in several different spots. If there is a problem with any of the mats, you should identify it and correct it before proceeding with the mortar installation.

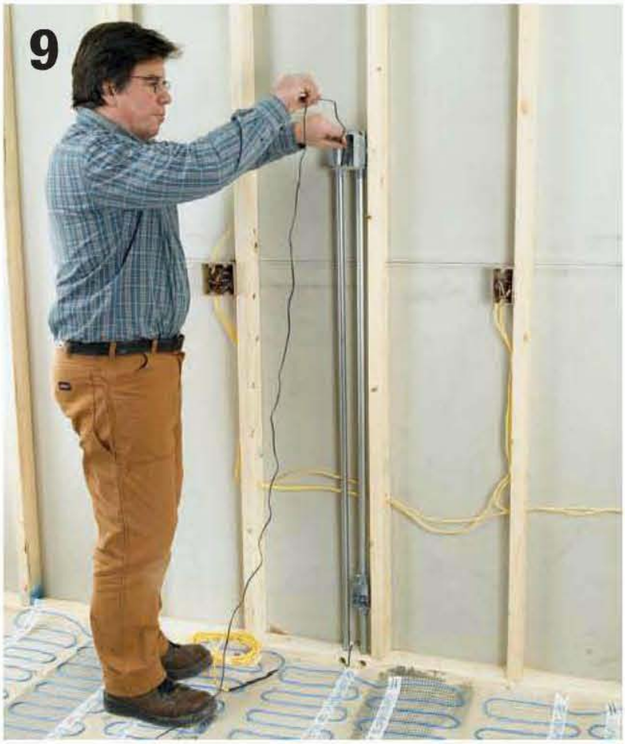

Run the thermostat sensor wire from the electrical box down the ½" conduit raceway and out the access hole in the sill plate. Select the best location for the thermostat sensor and mark the location onto the flooring. Also mark the locations of the wires that connect to and lead from the sensor.

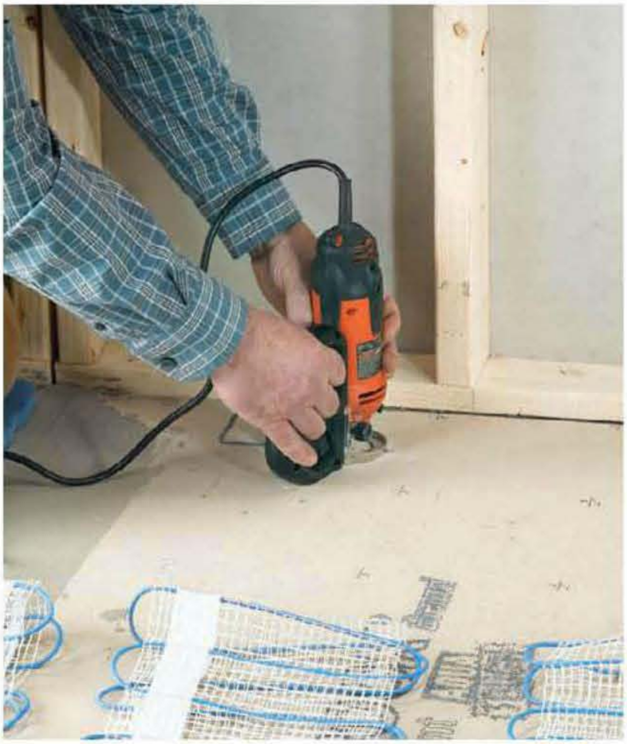

Variation: If your local codes require it, roll the mats out of the way and cut a channel for the sensor and the sensor wires into the floor or floor underlayment. For most floor materials, a spiral cutting tool does a quick and neat job of this task. Remove any debris.

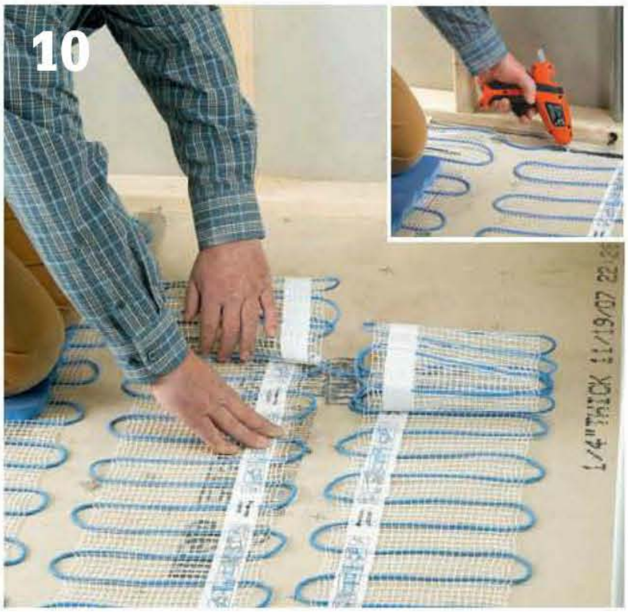

Bond the mats to the floor. If the mats in your system have adhesive strips, peel off the adhesive backing and roll out the mats in the correct position, pressing them against the floor to set the adhesive. If your mats have no adhesive, bind them with strips of double-sided carpet tape. The thermostat sensor and the power supply leads should be attached with hot glue (inset photo) and run up into their respective holes in the sill plate if you have not done this already. Test all mats for resistance and continuity.

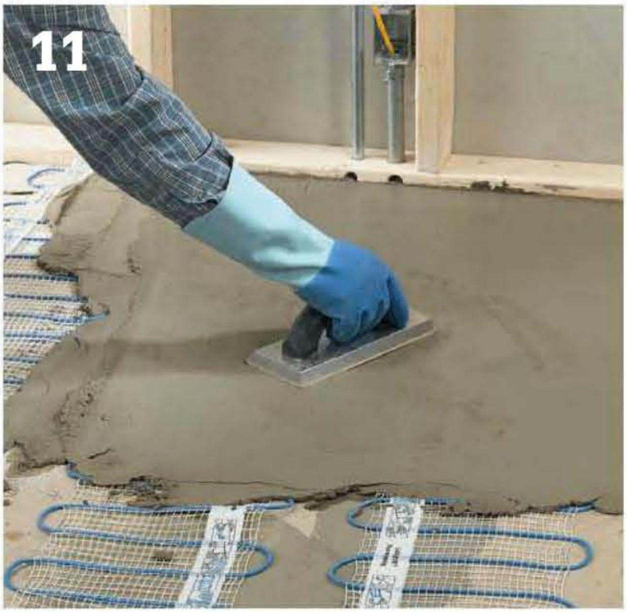

Cover the floor installation areas with a layer of thinset mortar that is thick enough to fully encapsulate all the wires and mats (usually around ¼" in thickness). Check the wires for continuity and resistance regularly and stop working immediately if there is a drop in resistance or a failure of continuity. Allow the mortar to dry overnight.

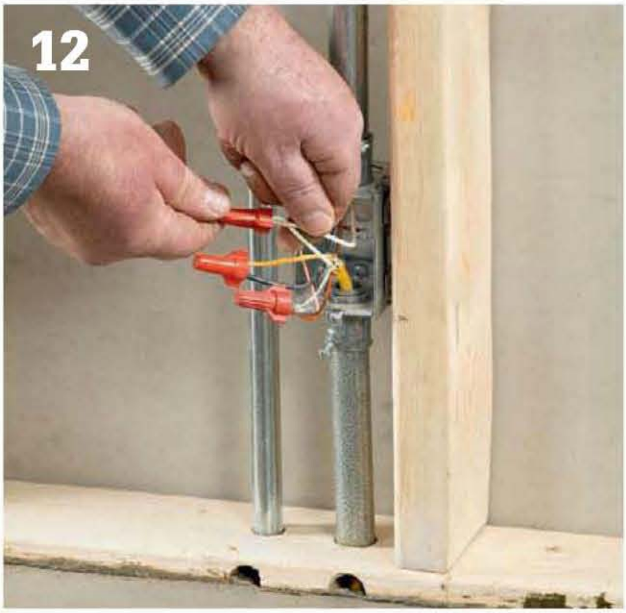

Connect the power supply leads from the mat or mats to the NM cable coming from the thermostat inside the junction box near the sill. Power must be turned off. The power leads should be cut so about 8" of wire feeds into the box. Be sure to use cable clamps to protect the wires.

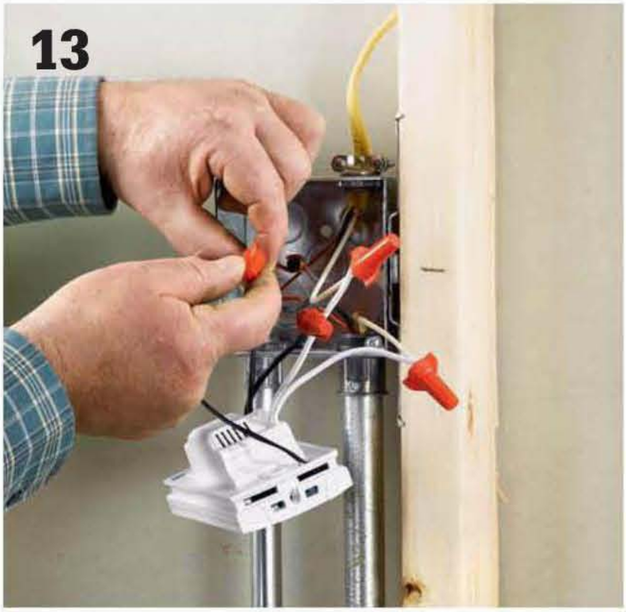

Connect the sensor wire and the power supply lead (from the junction box) to the thermostat/timer according to the manufacturer's directions. Attach the device to the electrical box, restore power, and test the system to make sure it works. Once you are convinced that it is operating properly, install flooring and repair the wall surfaces.

Installing Hardwood Strip Flooring

Tongue-and-groove hardwood flooring has always been popular with homeowners. It offers an attractive look, is one of the longest lasting floor coverings, and can be stripped and refinished to look like new.

Oak has been the most common type of strip flooring because of its durability and wood graining, and it's the species most people think of when hardwood is mentioned. Other woods, such as maple, cherry, and birch, are also becoming popular.

Exotic species of wood from around the world are now finding their way into American homes as people want a premium strip or plank floor that is unique and stylish, and expresses their personalities. The more than sixty exotic hardwoods include Brazilian cherry, Australian cypress, Honduran mahogany, tobaccowood, teak, zebrawood, and bamboo—which is not really wood but a type of grass.

This section describes how to install nailed-down tongue-and-groove flooring, how to install a decorative medallion, and how to install tongue-and-groove strip flooring over troweled-on adhesive. Customizing your floor with borders, accents, and medallions is easier than you may think. A number of manufacturers produce a variety of decorative options made to match the thickness of your floor.

Tools & Materials ▸

- Rosin paper
- Utility knife
- Chalk line
- 8d finish nails
- Drill
- Staple
- Nail set
- Hammer
- Pry bar
- Wood mallet
- Power nailer
- Pull bar

Real hardwood flooring has depth of beauty and warmth underfoot that even the highest quality laminate imitations struggle to match.

How to Install Tongue-and-Groove Hardwood Flooring

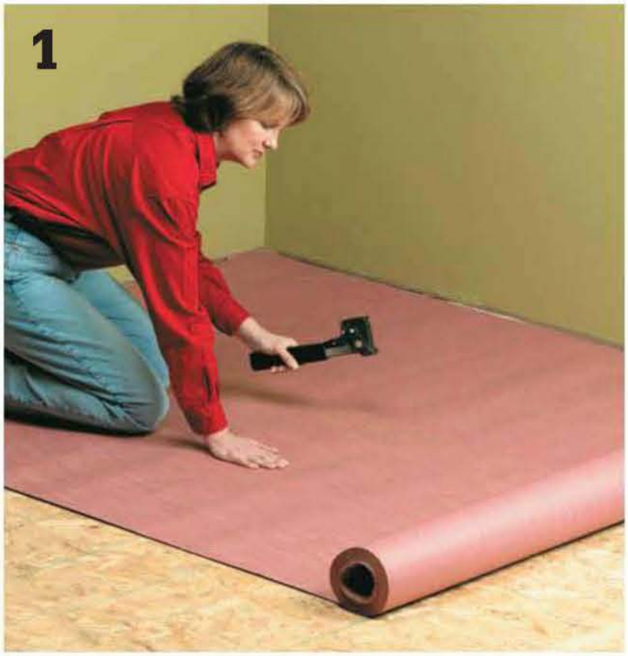
1

Cover the entire subfloor with rosin paper. Staple the paper to the subfloor, overlapping edges by 4". Cut the paper with a utility knife to butt against the walls.

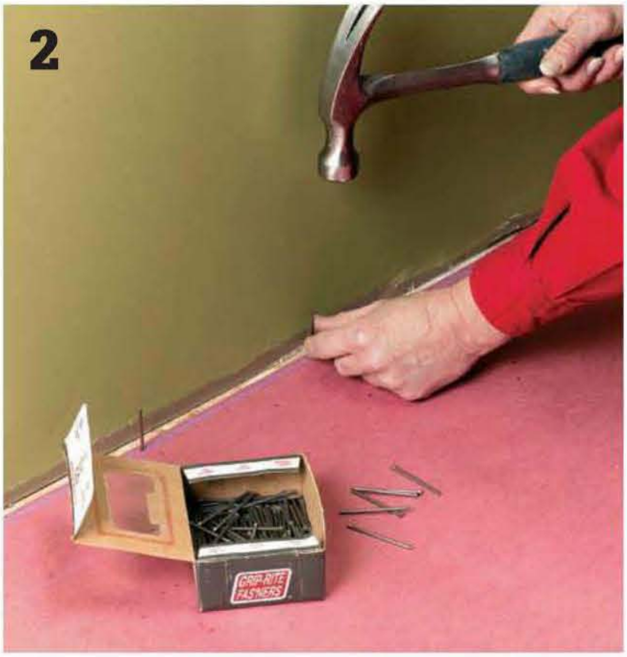
2

Make a mark on the floor ½" from the starter wall at both ends of the wall. Snap a chalk line between the marks. Nail 8d finish nails every 2" to 3" along the chalk line to mark the location for your first row.

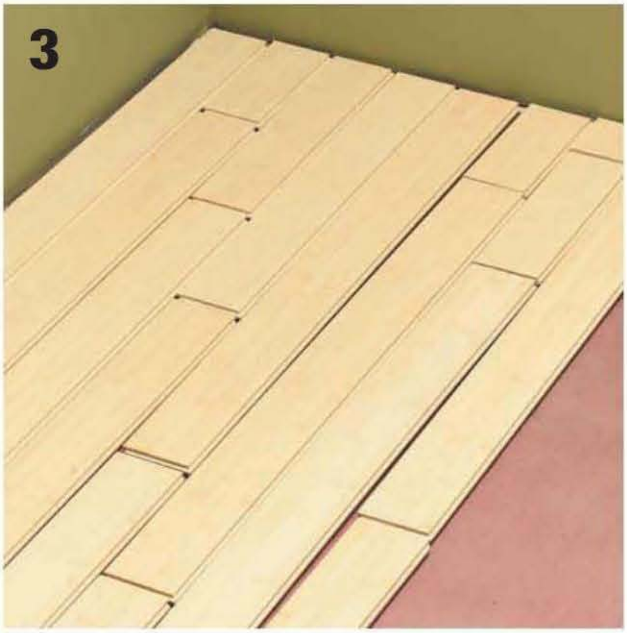
3

Lay out the first 8 rows of flooring in a dry run with the groove side facing the wall. Make sure the first row of boards is straight. Arrange the boards to get a good color and grain mix. Offset the ends by at least 6".

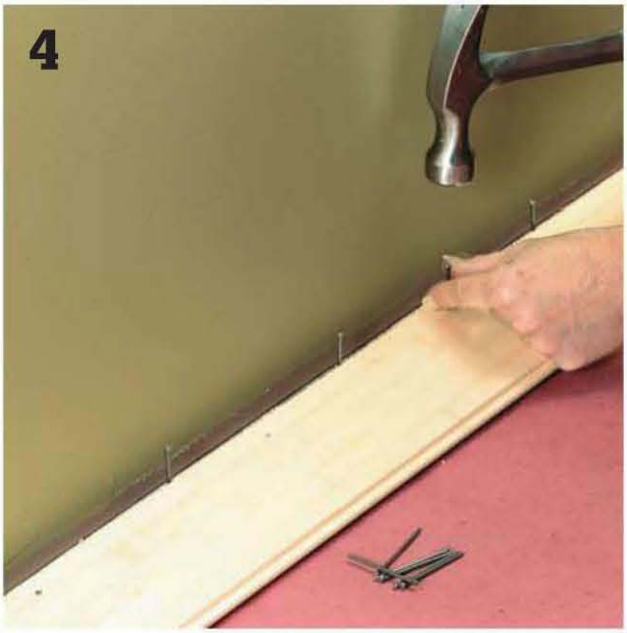
4

Place the starter row against the nails on the chalk line. Drill pilot holes in the flooring every 6" to 8", about ½" from the groove edge. Face nail the first row until the nail heads are just above the boards, then sink them using a nail set. (Be careful not to hit the boards with your hammer or you'll mar the surface.)

(continued)

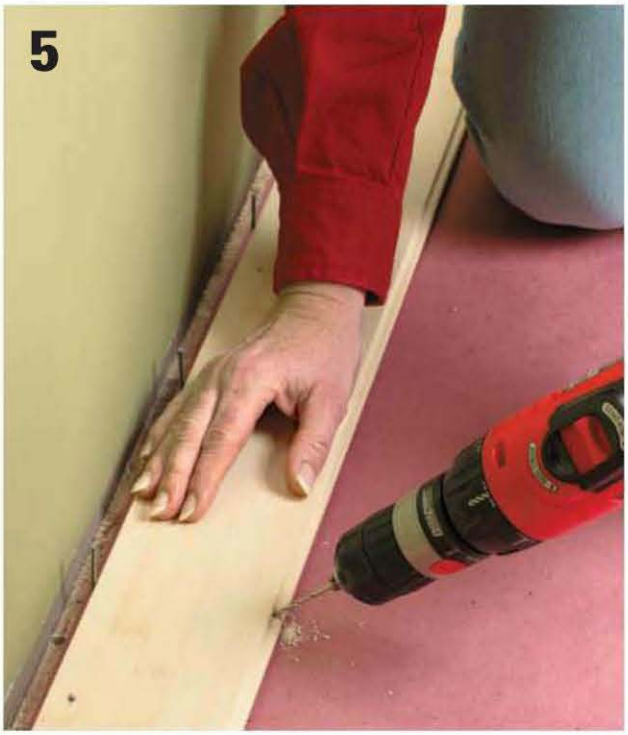

Drill pilot holes every 6" to 8" directly above the tongue, keeping the drill at a 45° angle.

Tip ▸

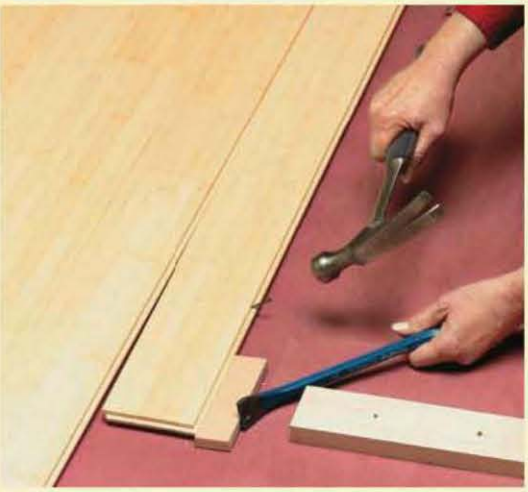

To install crooked boards, drill pilot holes above the tongue and insert nails. Fasten a scrap board to the subfloor using screws. Force the floor board straight using a pry bar and a scrap board placed in front of the flooring. With pressure on the floor board, blind nail it into place.

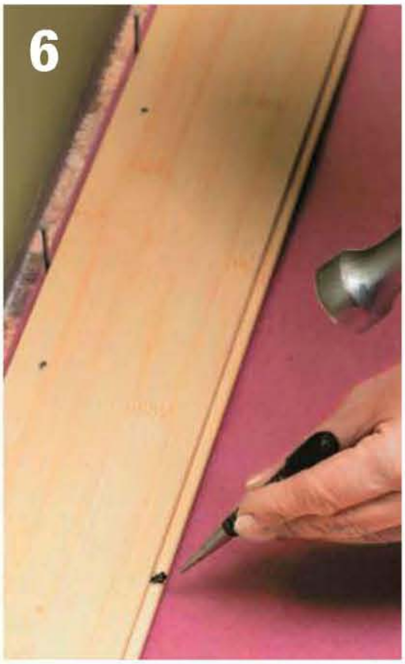

Blind nail a nail into each pilot hole. Keep the nail heads ½" out, then set them just below the surface, using a nail set.

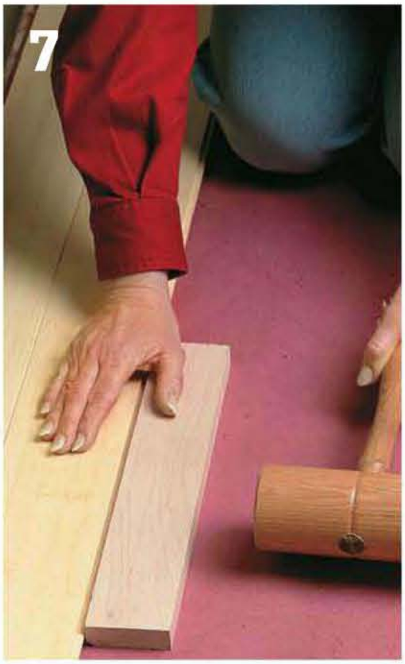

Set the second row of boards in place against the starter row, fitting together the tongue and groove connections. Use a scrap board and wood mallet to tap the floor boards together. Drill pilot holes and blind nail the boards. Do this for the next few rows.

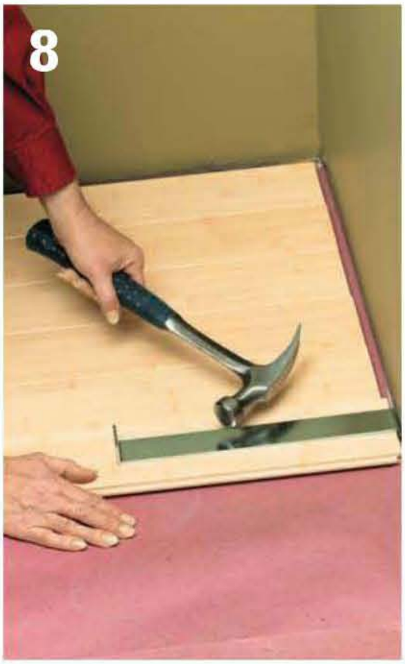

To install the last board in a row, place the tongue and groove joints together, then place a flooring pull bar over the end of the board. Hit the end of the pull bar with a hammer until the board slides into place. Stay ½" away from the walls.

9

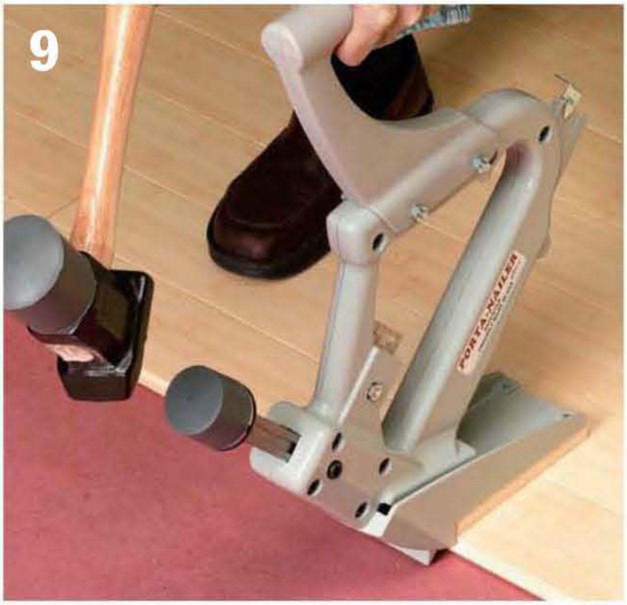

Once you have several rows installed and have enough room, use a power nailer. Place the nailer lip over the edge of the board and strike it with a mallet. Drive a nail 2" from the end of each board and about every 8" in the field. Keep a few rows of flooring laid out ahead of you as you work, and keep the joints staggered.

10

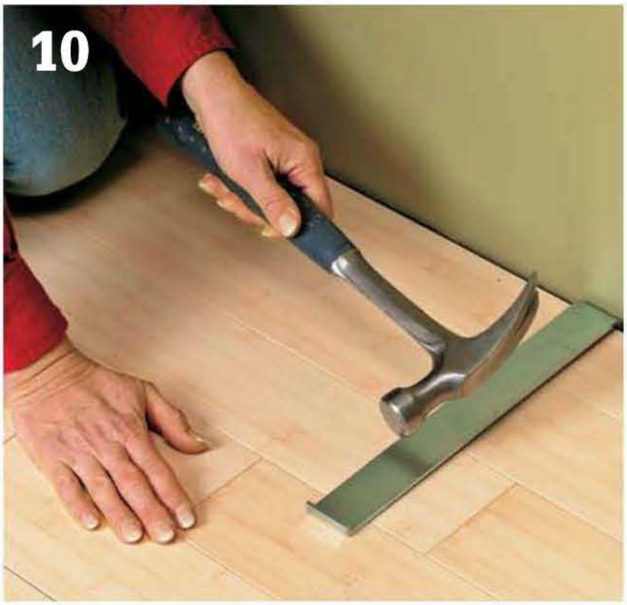

When you're out of room for the power nailer, drill pilot holes and blind nail the boards. For the last rows, drill pilot holes in the top of the boards, ½" from the tongue, and face nail them. The last row may need to be ripped to size. Pull the last row into place using the flooring pull bar, leaving a ½" gap along the wall. Drill pilot holes and face nail.

11

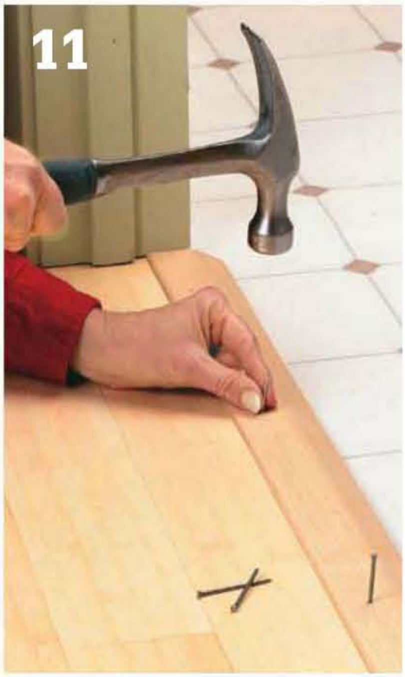

Install a reducer strip or transition strip between the wood floor and an adjoining room. Cut the strip to size. Fit the strip's groove over the floor board's tongue, then drill pilot holes and face nail. Set the nails with a nail set. Fill all visible nail holes with wood putty.

Tip ▸

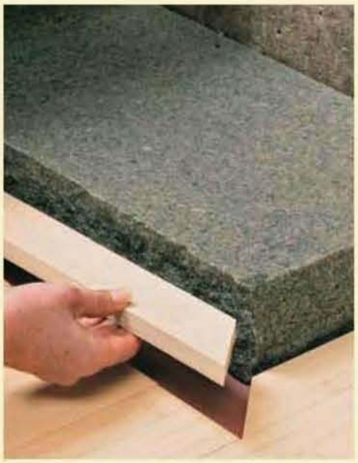

To install around an object, cut a notch in the board. For larger obstacles, cut 45° miters in boards so grooves face away from the object. Rip tongues off the boards. Set the boards against the object and flooring, fitting mitered ends together. Drill pilot holes and face nail in place. Apply silicone caulk between the floor and board edge.

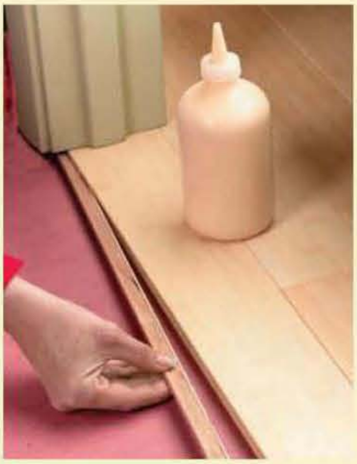

To reverse directions of the tongue and groove for doorways, glue a spline into the groove of the board. Fit the groove of the next board onto the spline, then nail the board in place.

Installing Laminate Floors

Laminate flooring comes in a floating system that is simple to install, even if you have no experience with other home-improvement projects. You may install a floating laminate floor right on top of plywood, concrete slab, sheet vinyl, or hardwood flooring. Just be sure to follow the manufacturer's instructions.

The pieces are available in planks or squares in a variety of different sizes, colors, and faux finishes—including wood and ceramic. The part you see is really a photographic print. Tongue-and-groove edges lock pieces together, and the entire floor floats on the underlayment. At the end of this project there are a few extra steps to take if your flooring manufacturer recommends using glue on the joints.

The rich wood tones of beautiful laminate planks may cause you to imagine hours of long, hard installation work, but this is a DIY project that you can do in a single weekend. Buy the manufactured planks at a home-improvement or flooring store and install laminate flooring with the step-by-step instructions offered in the following pages.

Tools & Materials ▸

Circular saw
Underlayment
½" spacers
Tapping block
Scrap foam
Speed square
Manufacturer glue
Painter's tape
Chisel
Rubber mallet
Drawbar
Finish nails
Nail set strap clamps
Threshold and screws

Laminate strip floors install quickly, wear well and are among the cheapest floor coverings you can find. Overall they have improved in appearance but the lower quality products continue to be fairly unconvincing imitations of natural materials.

How to Install a Floating Floor

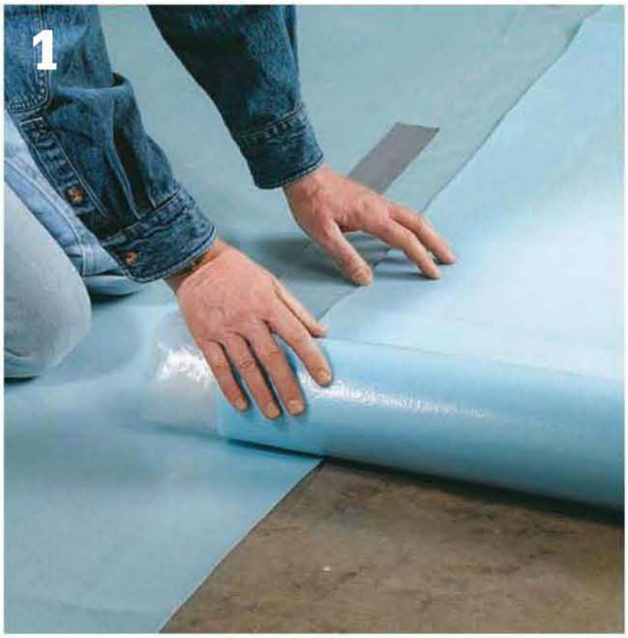

To install the underlayment, start in one corner and unroll the underlayment to the opposite wall. Cut the underlayment to fit, using a utility knife or scissors. Overlap the second underlayment sheet according to the manufacturer's recommendations, and secure the pieces in place with adhesive tape.

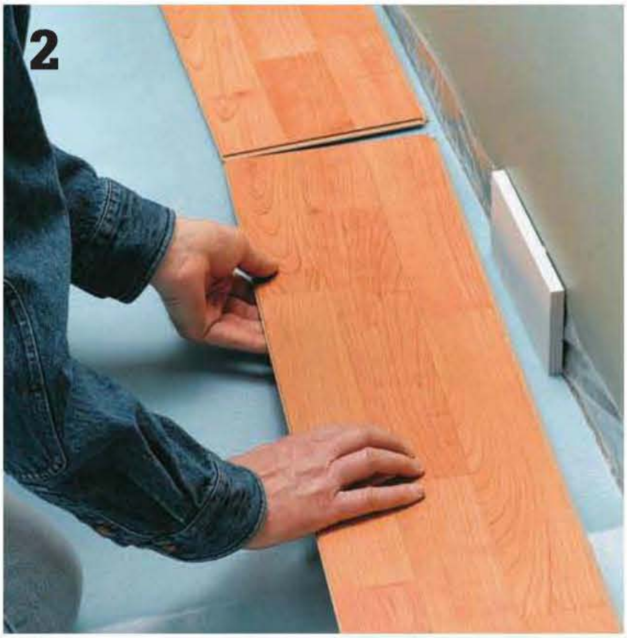

Working from the left corner of the room to right, set wall spacers and dry lay planks (tongue side facing the wall) against the wall. The spacers allow for expansion. If you are flooring a room more than 26 ft. long or wide, you need to buy appropriate-sized expansion joints. *Note: Some manufacturers suggest facing the groove side to the wall.*

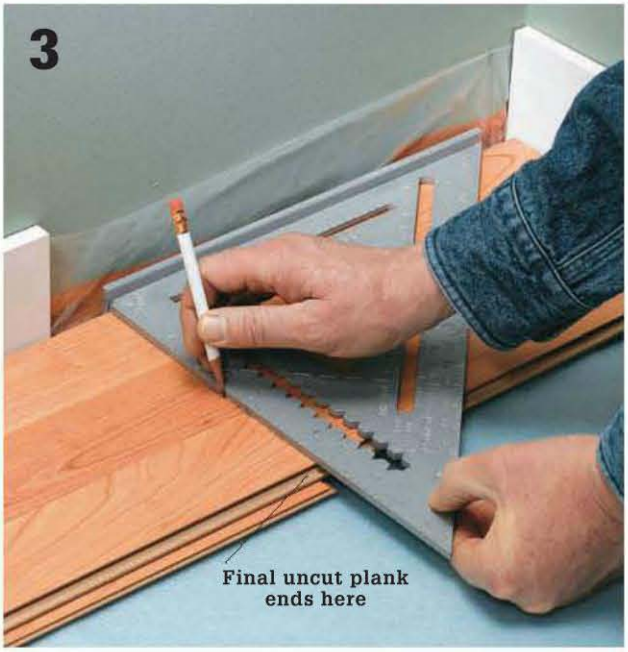
3
Final uncut plank ends here

Set a new plank right side up, on top of the previously laid plank, flush with the spacer against the wall at the end run. Line up a speed square with the bottom plank edge and trace a line. That's the cutline for the final plank in the row.

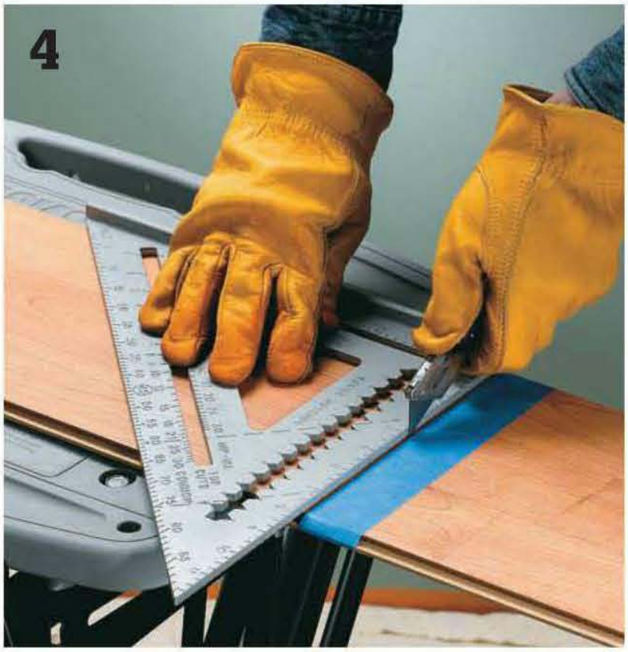

Press painter's tape along the cutline on the top of the plank to prevent chips when cutting. Score the line drawn in Step 3 with a utility knife. Turn the plank over and extend the pencil line to the backside.

(continued)

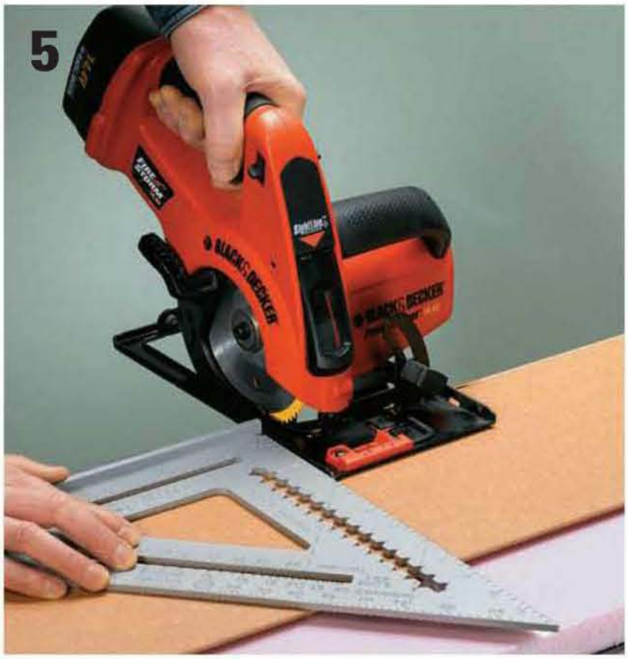

5

Clamp the board (face down) and rigid foam insulation or plywood to a work table. The foam reduces chipping. Clamp a speed square on top of the plank, as though you are going to draw another line parallel to the cutline—use this to make a straight cut. Place the circular saw's blade on the waste side of the actual cutline.

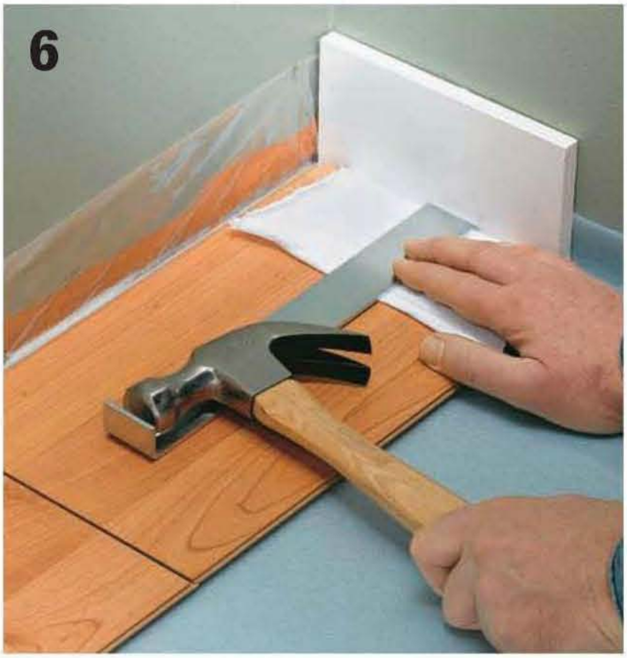

6

To create a tight fit for the last plank in the first row, place a spacer against the wall and wedge one end of a drawbar between it and the last plank. Tap the other end of the drawbar with a rubber mallet or hammer. Protect the laminate surface with a thin cloth.

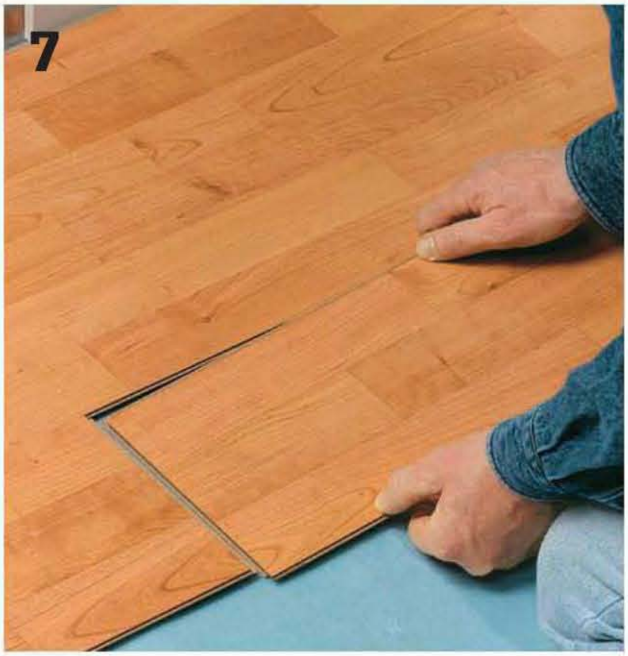

7

Continue to lay rows of flooring, making sure the joints are staggered. This prevents the entire floor from relying on just a few joints, and keeps the planks from lifting. Staggering also stengthens the floor, because the joints are shorter and more evenly distributed.

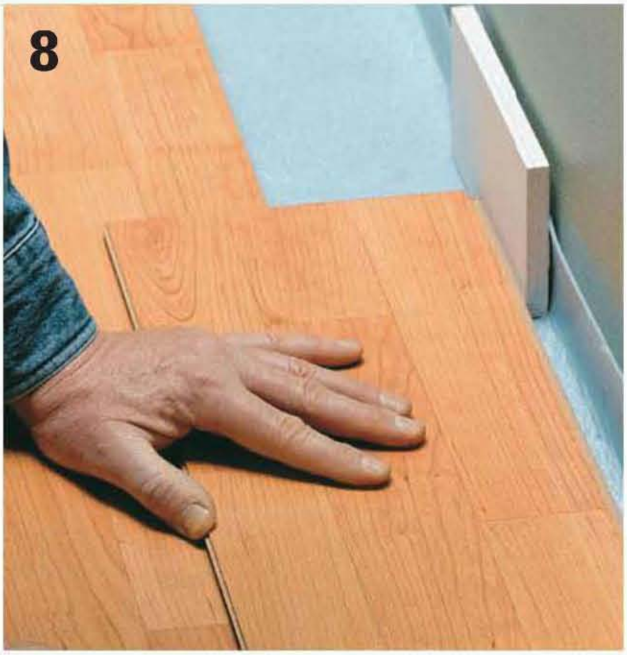

8

To fit the final row, place two planks on top of the last course; slide the top plank up against the wall spacer. Use the top plank to draw a cutline lengthwise on the middle plank. Cut the middle plank to size using the same method as in Step 3, just across the grain. The very last board must be cut lengthwise and widthwise to fit.

How to Work Around Obstacles

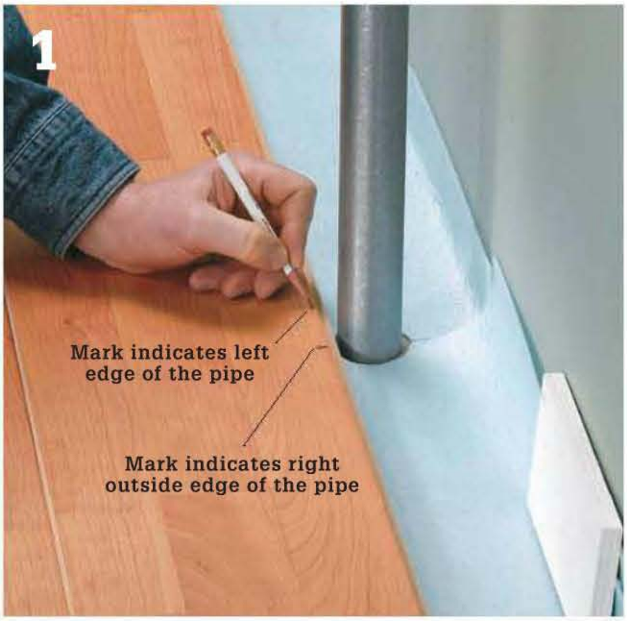
1
Mark indicates left edge of the pipe
Mark indicates right outside edge of the pipe

Position a plank and use a pencil to make two marks along the length of the plank, indicating the points where the obstacle begins and ends.

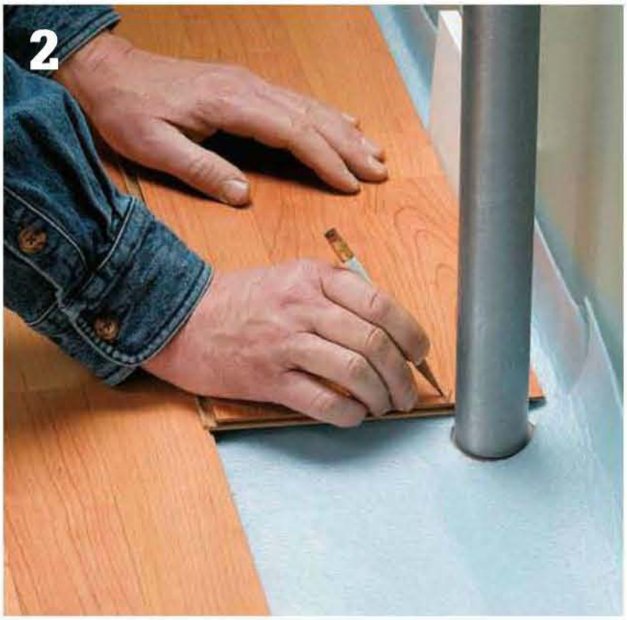

Position the plank end against the obstacle. Make two marks with a pencil, this time on the end of the plank to indicate where the obstacle falls along the width of the board.

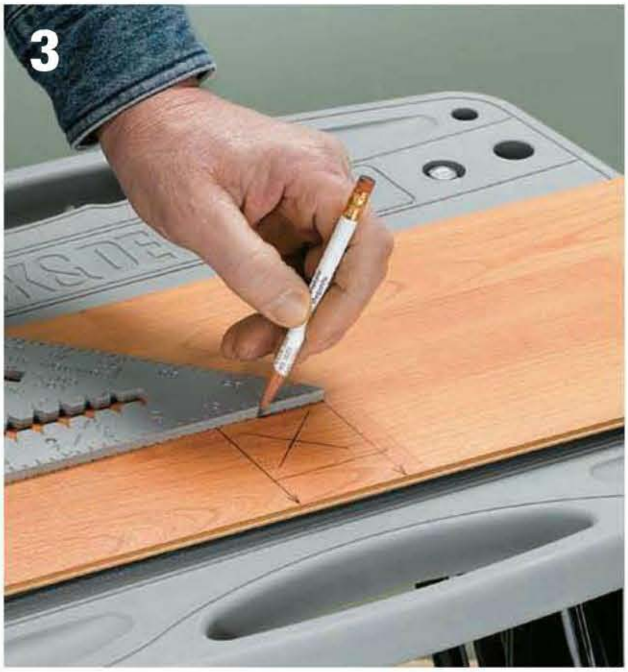

Use a speed square to extend the four lines. The space at which they intersect is the part of the plank that needs to be removed to make room for the obstacle to go through it. Use a drill with a Forstner bit, or a hole saw the same diameter as the space within the intersecting lines, and drill through the plank at the X. You'll be left with a hole; extend the cut to the edges with a jigsaw.

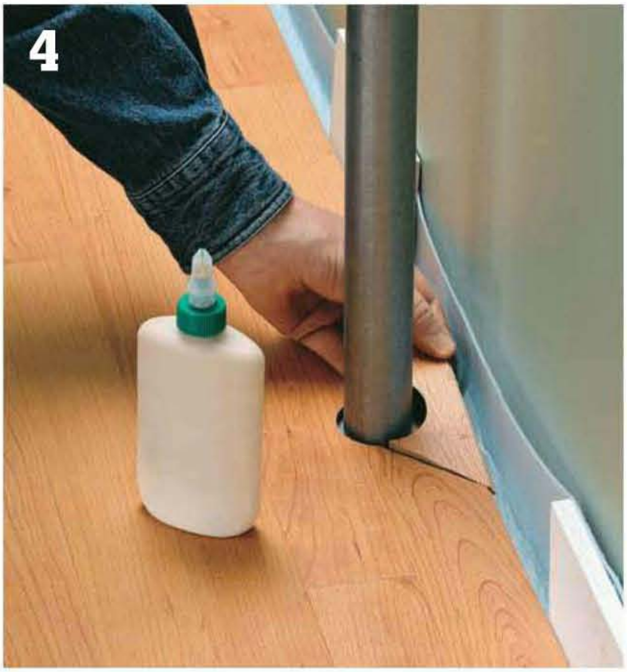

Install the plank by locking the tongue-and-groove joints with the preceding board. Fit the end piece in behind the pipe or obstacle. Apply manufacturer-recommended glue to the cut edges, and press the end piece tightly against the adjacent plank. Wipe away excess glue with a damp cloth.

Installing Parquet Flooring

For a hardwood floor with great design appeal, consider installing a parquet floor. It offers more visual interest than strip flooring without sacrificing the beauty and elegance of wood. Parquet comes in a variety of patterns and styles to create geometric designs. It can range from elaborate, custom-designed patterns on the high end, to the more common herringbone pattern, to the widely available and less expensive block design.

Parquet has experienced a radical transformation over the years. A few years ago, each individual piece of parquet was hand-cut and painstakingly assembled piece by piece. Today, parquet is prefabricated so the individual pieces making up the design are available as single tiles, which not only has reduced the cost, but has made the flooring easier to install.

Many types and designs of parquet floors are available, from custom-made originals to standard patterns, but they are all installed the same way—set in adhesive on a wood subfloor. The effort can be very rewarding: Parquet can be used to create shapes not possible with other wood flooring.

The finger block pattern is one of the most widely available parquet coverings and also one of the least expensive. The configuration of perpendicular strips of wood emphasizes the different grains and natural color variations.

Tools & Materials ▸

- Tape measure
- Chalk line
- Carpenter's square
- Parquet flooring
- Adhesive
- Notched trowel
- Putty knife
- Rubber mallet
- 100- to 150-pound floor roller
- Jigsaw
- Solvent

Parquet flooring tiles are made with real hardwood and can be a great vehicle for introducing exotic wood species at a reasonable cost. The custom parquet floor seen here is expensive, but a basic parquet floor with stock tiles is an affordable option and an easy DIY project.

How to Install Parquet Flooring

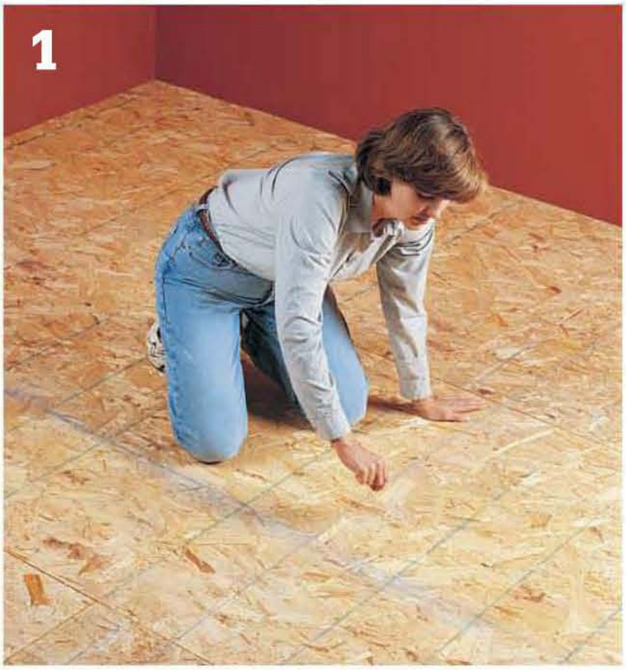

Mark the centerpoint of each wall. Snap chalk lines between the marks on opposite walls to establish your reference lines. Use the 3-4-5 triangle method to check the lines for squareness (see page 181).

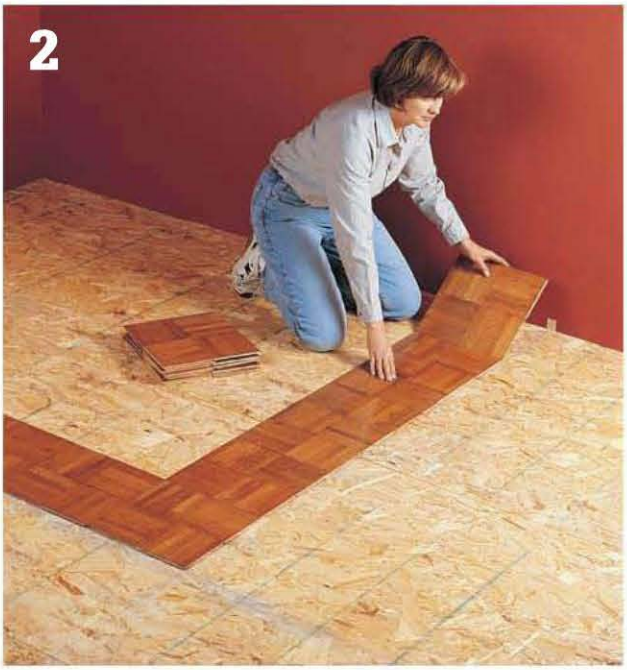

Lay out a dry run of panels from the center point along the reference lines to adjacent walls. Place ¼" spacers along the walls to allow for expansion. If more than half of the last panel needs to be cut off, adjust the lines by half the width of the panel. Snap new working lines, if necessary.

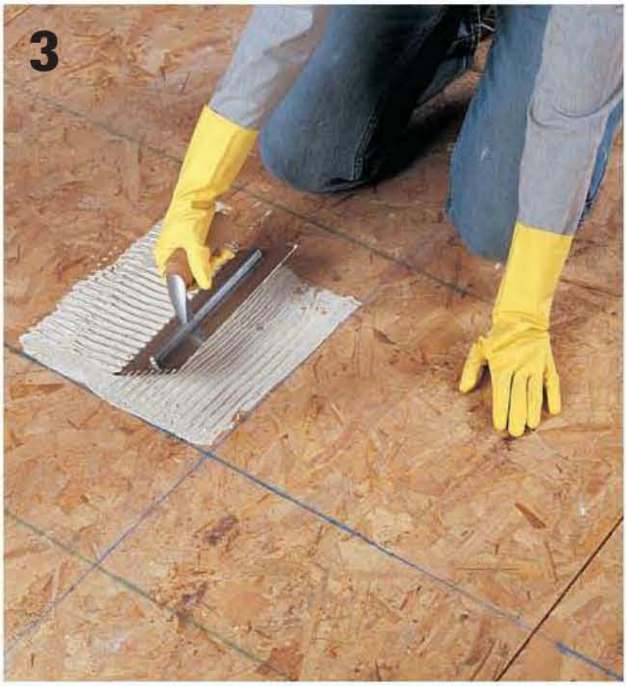

Put enough adhesive on the subfloor for your first panel, using a putty knife. Spread the adhesive into a thin layer with a notched trowel held at a 45° angle. Apply the adhesive right up to the working lines, but do not cover them.

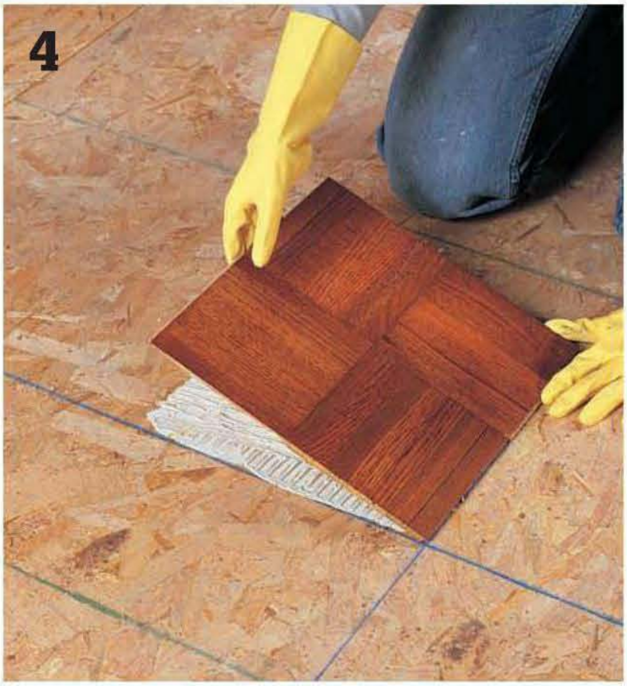

Place the first panel on the adhesive so two sides are flush with the working lines. Take care not to slide or twist the panel when setting it into place. This panel must be positioned correctly to keep the rest of your floor square.

(continued)

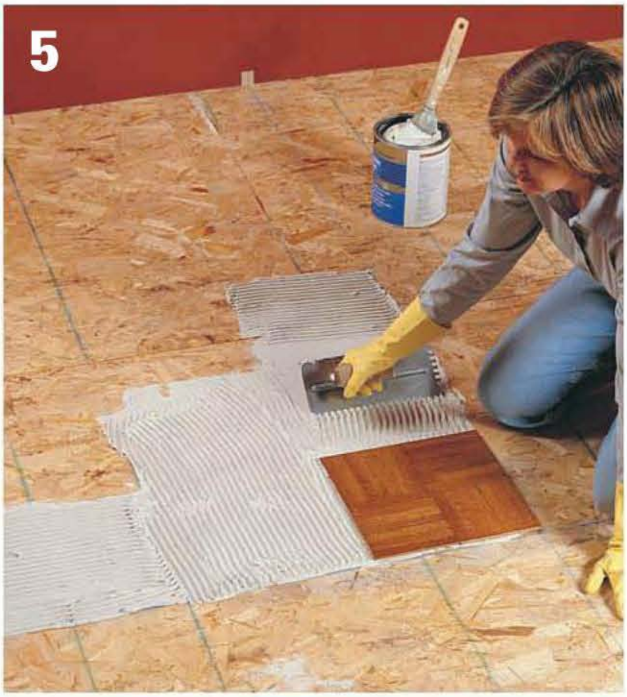

Apply enough adhesive for six to eight panels and spread it with a notched trowel.

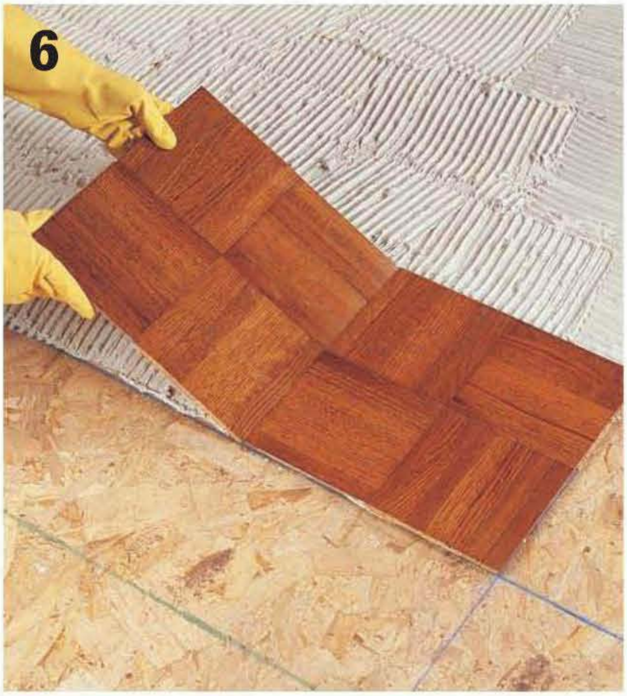

Set the next panel in place by holding it at a 45° angle and locking the tongue-and-groove joints with the first panel. Lower the panel onto the adhesive without sliding it. Install remaining panels the same way.

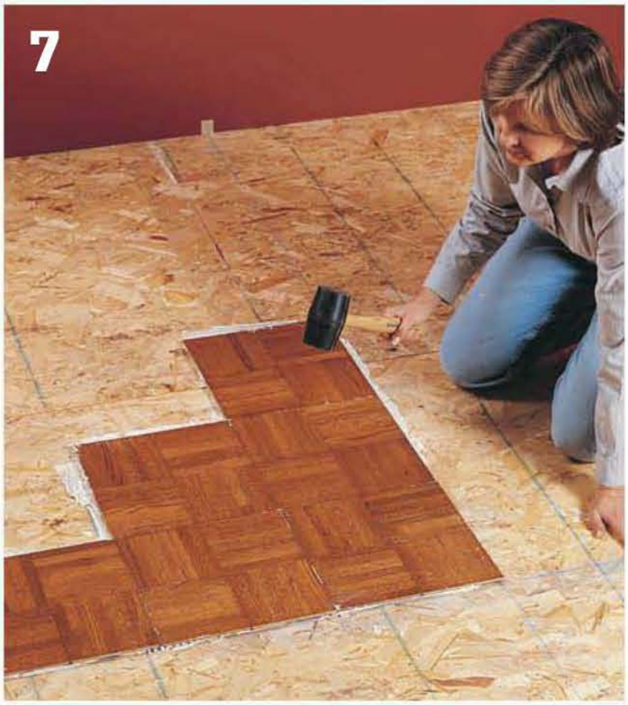

After every six to eight panels are installed, tap them into the adhesive with a rubber mallet.

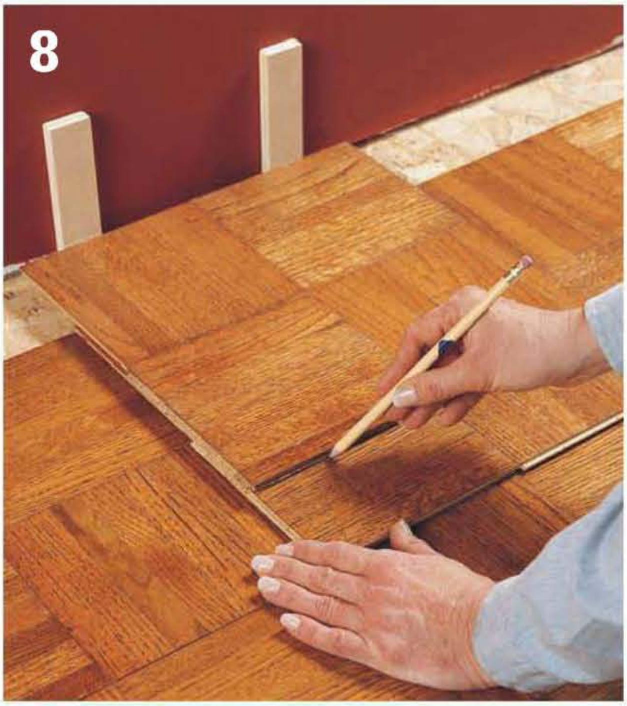

For the last row, align panels over the top of the last installed row. Place a third row over the top of these, with the sides butted against ½" spacers along the wall. Draw a line along the edge of the third panels onto the second row, cut the panels at the marks, and install.

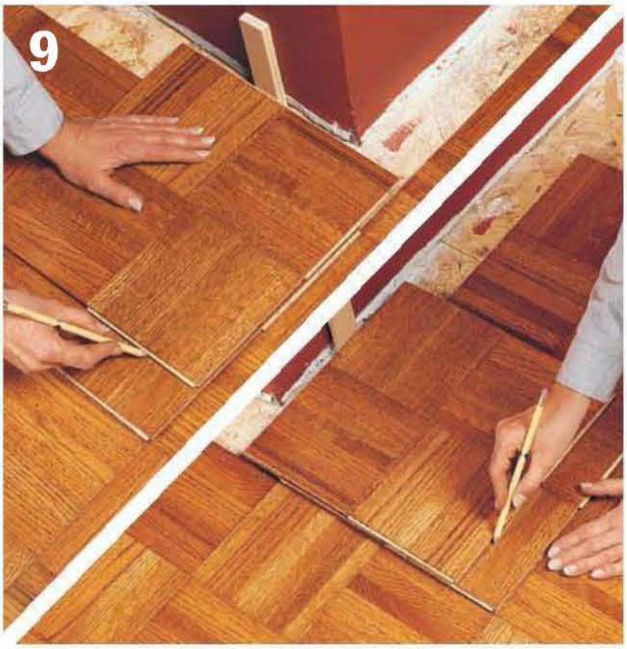

To work around corners or obstacles, align a panel over the last installed panel, then place another panel on top of it as in step 8. Keep the top panel ½" from the wall or obstacle and trace along the opposite edge onto the second panel (top). Move the top two panels to the adjoining side, making sure not to turn the top panel. Make a second mark on the panel the same way (bottom). Cut the tile with a jigsaw and install.

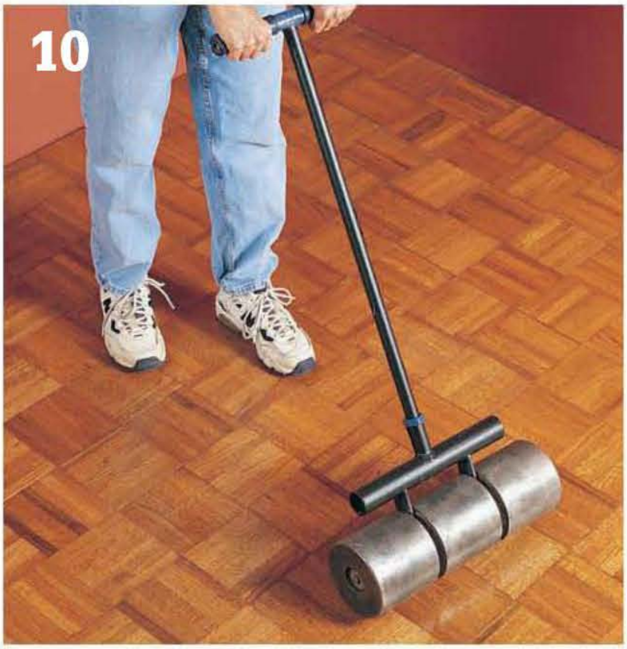

Within 4 hours of installing the floor, roll the floor with a 100- to 150-pound floor roller. Wait at least 24 hours before walking on the floor again.

How to Install Parquet in a Diagonal Layout

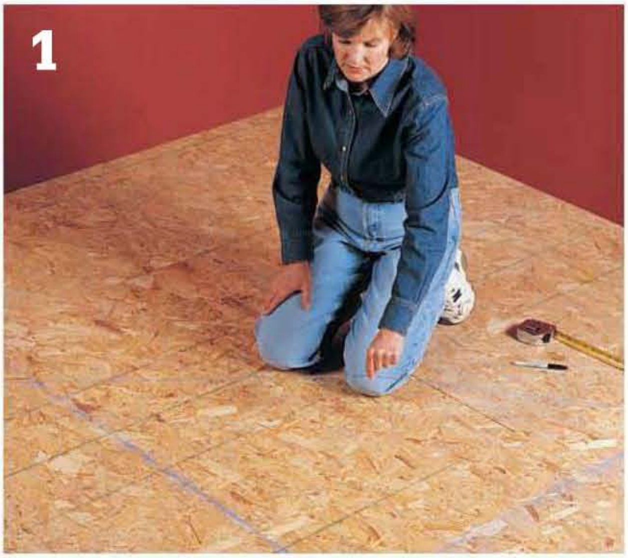

Establish perpendicular working lines following Step 1 on page 171. Measure 5 ft. from the centerpoint along each working line and make a mark. Snap chalk lines between the 5 ft. marks. Mark the centerpoint of these lines, then snap a chalk line through the marks to create a diagonal reference line.

Lay out a dry run of tiles along a diagonal line. Adjust your starting point as necessary. Lay the flooring along the diagonal line using adhesive, following the steps for installing parquet (pages 171 to 173). Make paper templates for tile along walls and in corners. Transfer the template measurements to tiles, and cut to fit.

Installing Sheet Vinyl

Preparing a perfect underlayment is the most important phase of resilient sheet vinyl installation. Cutting the material to fit the contours of the room is a close second. The best way to ensure accurate cuts is to make a cutting template. Some manufacturers offer template kits, or you can make your own. Be sure to use the recommended adhesive for the sheet vinyl you are installing. Many manufacturers require that you use their glue for installation. Use extreme care when handling the sheet vinyl, especially felt-backed products, to avoid creasing and tearing.

Tools & Materials ▸

- Linoleum knife
- Framing square
- Compass
- Scissors
- Non-permanent felt-tipped pen
- Utility knife
- Straightedge
- 1/4" V-notched trowel
- J-roller
- Stapler
- Flooring roller
- Chalk line
- Heat gun
- 1/16" V-notched trowel
- Straightedge
- Vinyl flooring
- Masking tape
- Heavy butcher or brown wrapping paper
- Duct tape
- Flooring adhesive
- 3/8" staples
- Metal threshold bars
- Nails

Sheet vinyl has no seams (or at least very few seams), which makes it a practical choice for wet areas such as bathrooms. Advancing technology has made a host of new patterns and styles available, including the mosaic tile look-alike pattern seen here.

Tools for Resilient Floors

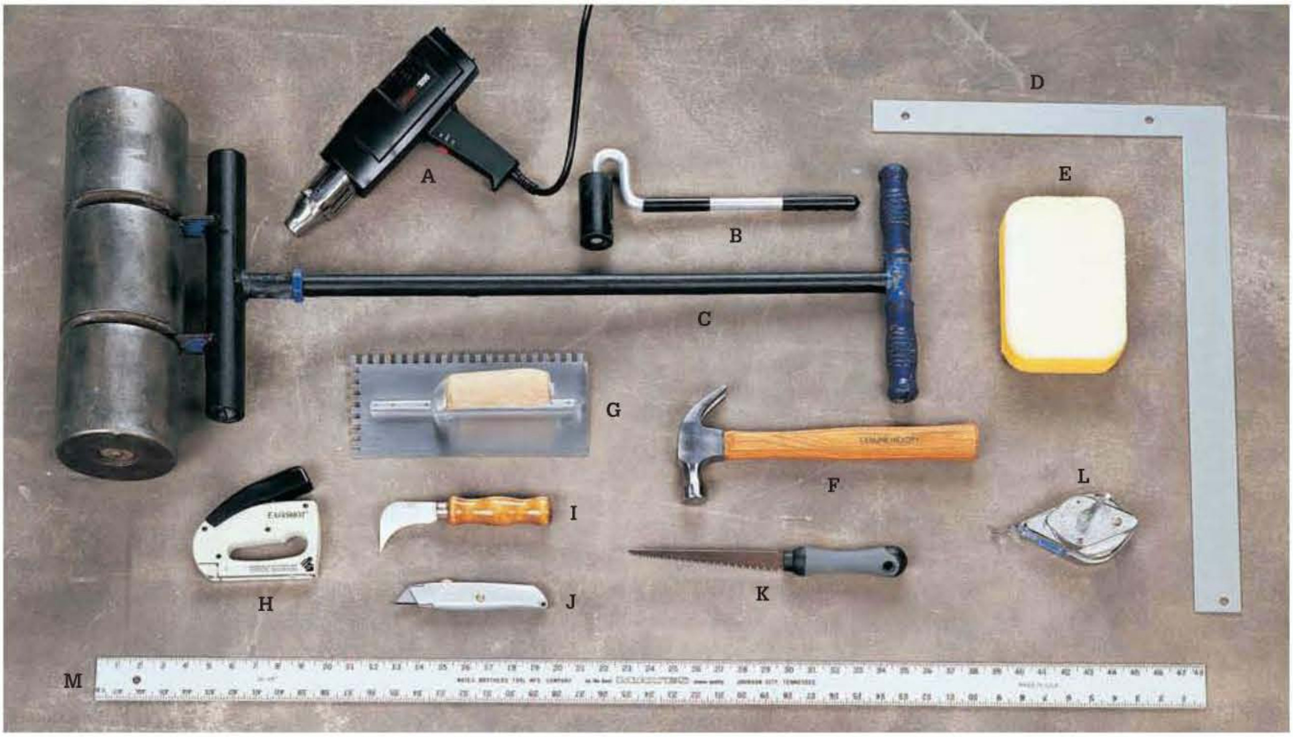

Tools for resilient flooring include: a heat gun (A), J-roller (B), floor roller (C), framing square (D), sponge (E), hammer (F), notched trowel (G), stapler (H), linoleum knife (I), utility knife (J), wallboard knife (K), chalk line (L), straightedge (M).

Buying & Estimating

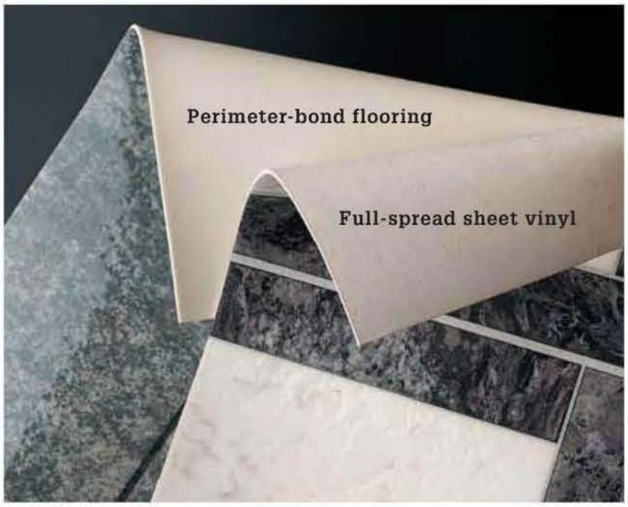

Resilient sheet vinyl comes in full-spread and perimeter-bond styles. Full-spread sheet vinyl has a felt-paper backing and is secured with adhesive that is spread over the floor before installation. Perimeter-bond flooring, identifiable by its smooth, white PVC backing, is laid directly on underlayment and is secured by a special adhesive spread along the edges and seams.

Dry-back tile
Self-adhesive tile

Resilient tile comes in self-adhesive and dry-back styles. Self-adhesive tile has a pre-applied adhesive protected by wax paper backing that is peeled off as the tiles are installed. Dry-back tile is secured with adhesive spread onto the underlayment before installation. Self-adhesive tile is easier to install than dry-back tile, but the bond is less reliable. Don't use additional adhesives with self-adhesive tile.

Evaluating an Existing Floor

The first step in preparing for a new floor covering is evaluating your old floor. A careful examination can help you decide whether to repair damaged areas, or replace the flooring altogether.

Evaluating your floor is a three-step process. Begin by identifying the existing floor material and the installation method used. Is your sheet vinyl attached using the full-spread method or the perimeter-bond method? Is your carpet glued down or stretched? Next, check the condition of the floor. Is it securely attached or is it loose in spots? Is it chipped or cracked? Finally, note the height of the existing floor in relation to adjoining floor surfaces. Is it significantly higher than surrounding floors?

A new floor covering or underlayment can often be installed on top of existing flooring. If the existing flooring is not sound or smooth, however, you will have to do some preparation work. Applying a floor leveler is one way to make your existing floor easier to use. More complex preparations may involve removing and replacing the underlayment or making spot repairs to the subfloor.

Warning ▸

Resilient flooring manufactured before 1986 may contain asbestos, which can cause severe lung problems if inhaled. The recommended method for dealing with asbestos-laden flooring is to cover it with an underlayment. If the flooring must be removed, do not do the work yourself. Instead, consult a certified asbestos-abatement contractor.

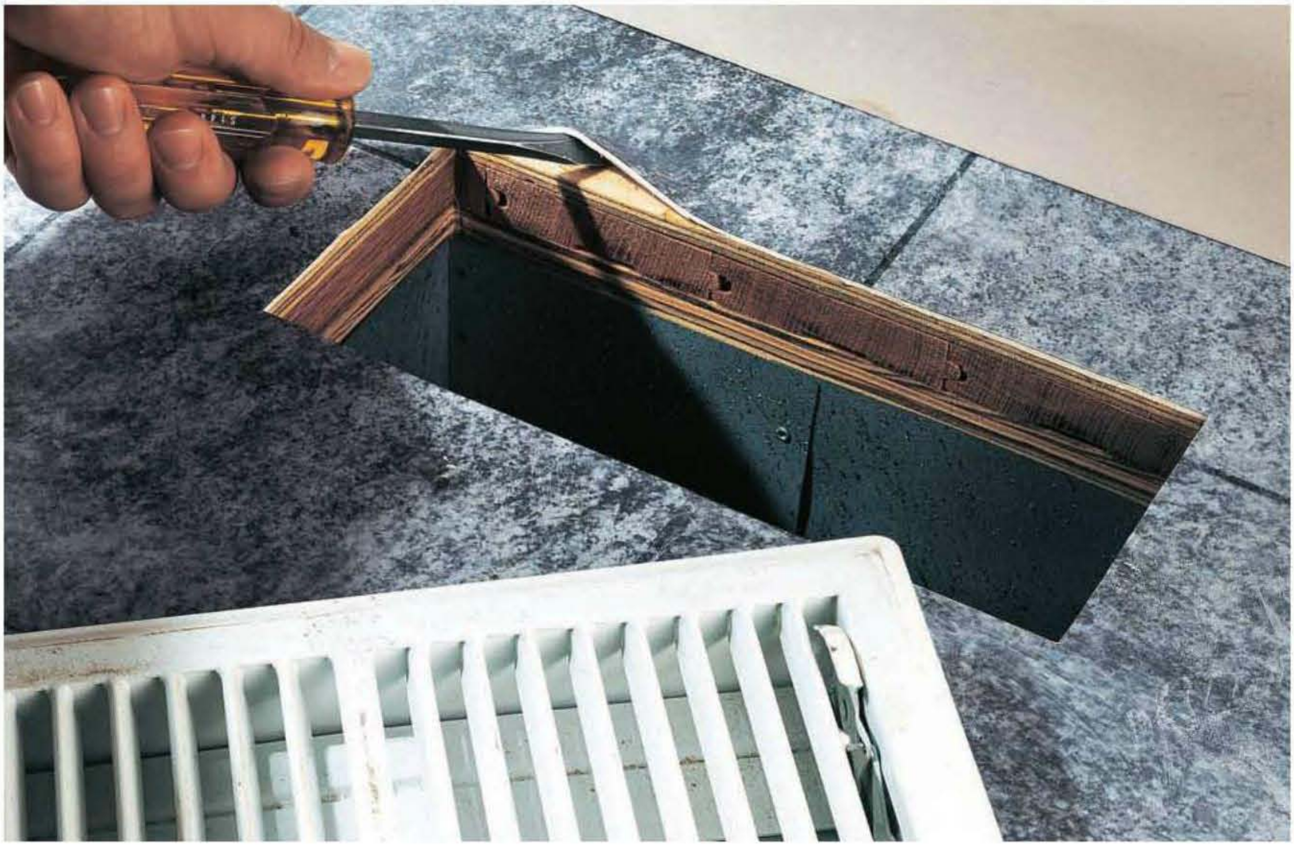

Determining the number and type of coverings already on your floor is an important early evaluation step. Too many layers of flooring and underlayment can stress floor joists and ultimately cause a new floor to fail. An easy way to check for old flooring is to remove floor vents.

How to Install Perimeter-bond Sheet Vinyl

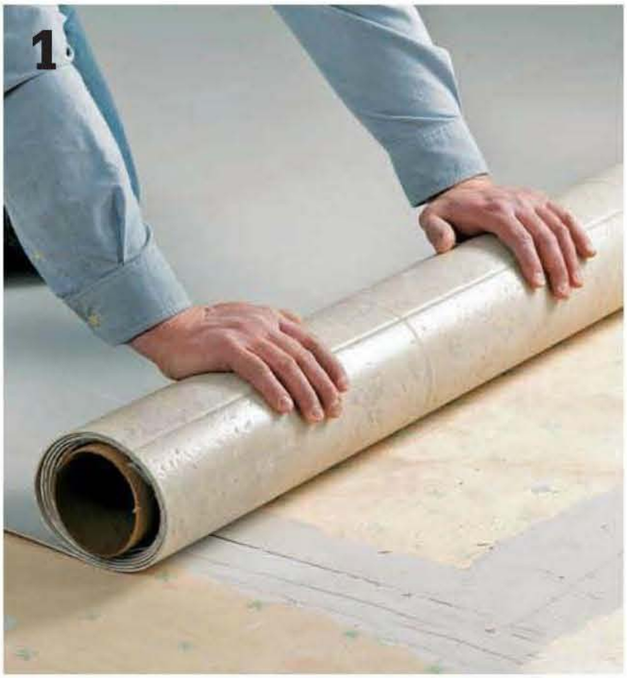

Unroll the flooring on any large, flat, clean surface. To prevent wrinkles, sheet vinyl comes from the manufacturer rolled with the pattern-side out. Unroll the sheet and turn it pattern-side up for marking.

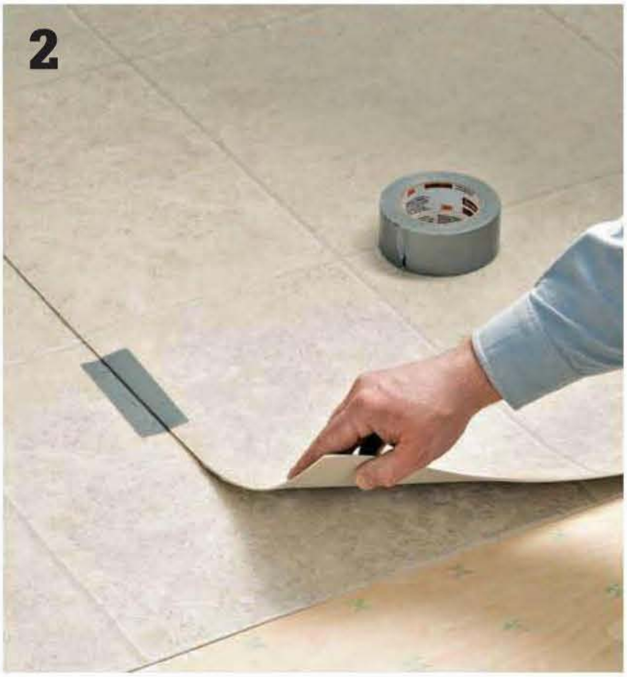

For two-piece installations, overlap the edges of the sheets by at least 2". Plan to have the seams fall along the pattern lines or simulated grout joints. Align the sheets so the pattern matches, then tape the sheets together with duct tape.

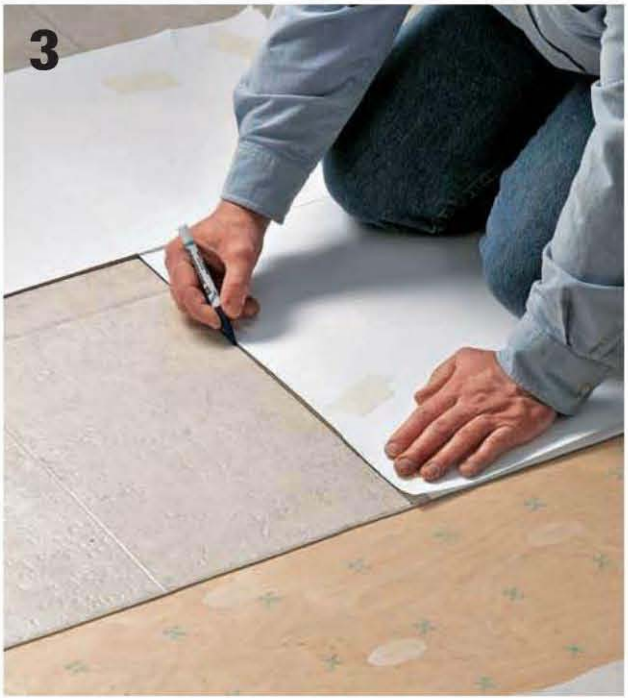

Position the paper template over the sheet vinyl and tape it in place. Trace the outline of the template onto the flooring using a non-permanent felt-tipped pen.

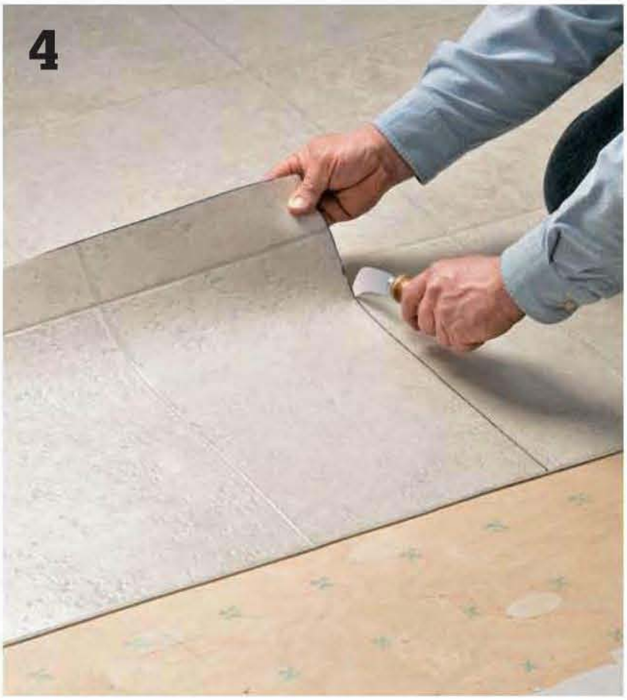

Remove the template. Cut the sheet vinyl with a sharp linoleum knife or a utility knife with a new blade. Use a straightedge as a guide for making longer cuts.

(continued)

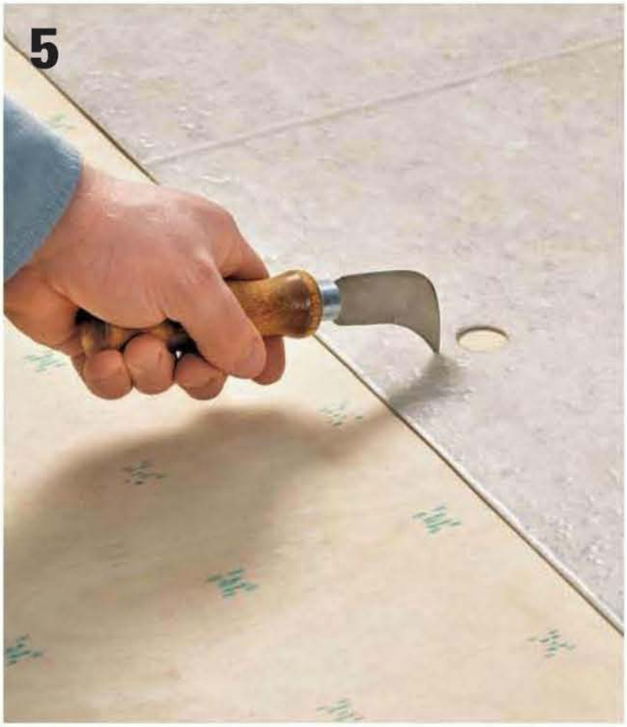

Cut holes for pipes and other permanent obstructions. Cut a slit from each hole to the nearest edge of the flooring. Whenever possible, make slits along pattern lines.

Roll up the flooring loosely and transfer it to the installation area. Do not fold the flooring. Unroll and position the sheet vinyl carefully. Slide the edges beneath door casings.

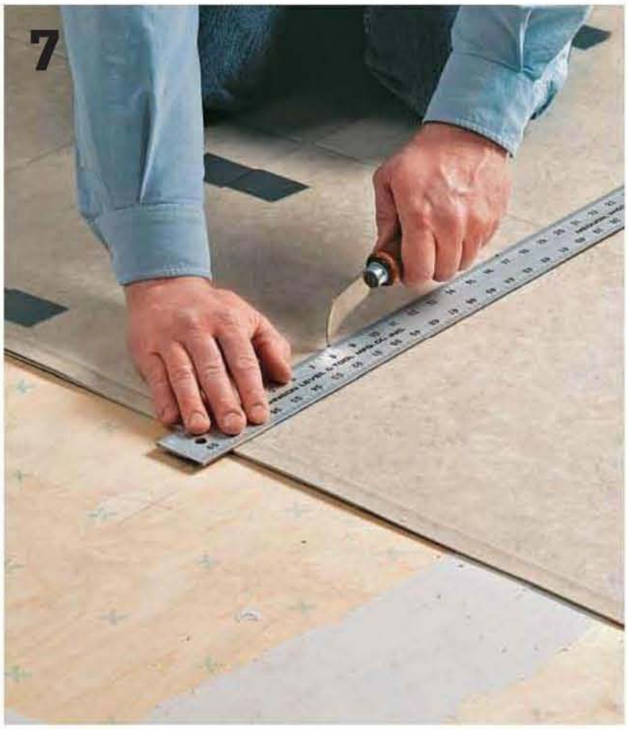

Cut the seams for two-piece installations using a straightedge as a guide. Hold the straightedge tightly against the flooring, and cut along the pattern lines through both pieces of vinyl flooring.

Remove both pieces of scrap flooring. The pattern should now run continuously across the adjoining sheets of flooring.

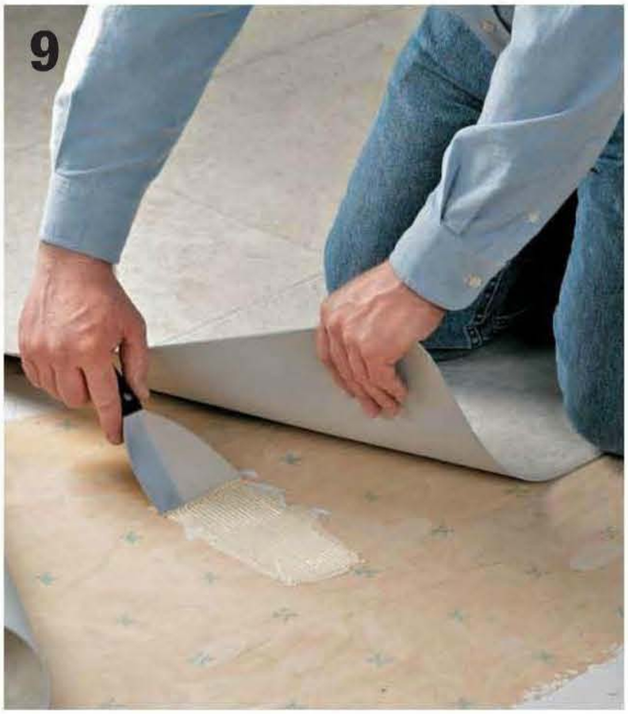

Fold back the edges of both sheets. Apply a 3" band of multipurpose flooring adhesive to the underlayment or old flooring, using a ¼" V-notched trowel or wallboard knife.

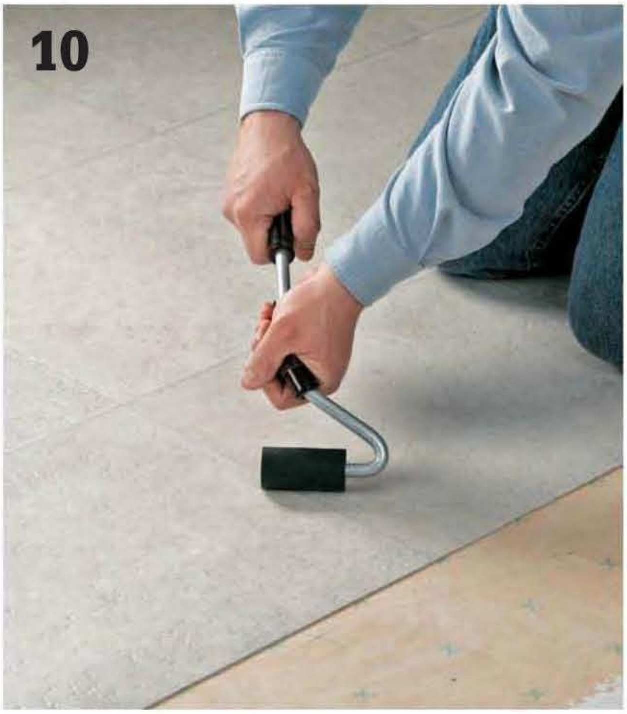

Lay the seam edges one at a time onto the adhesive. Make sure the seam is tight, pressing the gaps together with your fingers, if needed. Roll the seam edges with a J-roller or wallpaper seam roller.

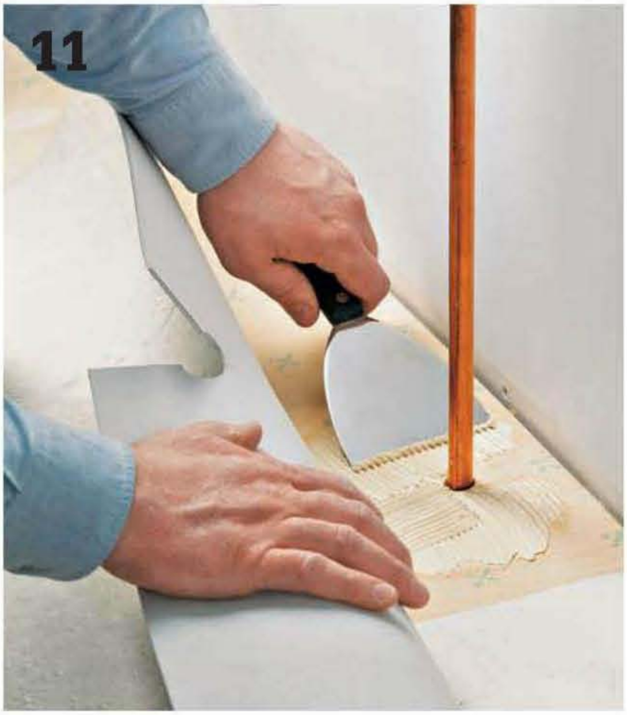

Apply flooring adhesive underneath flooring cuts at pipes or posts and around the entire perimeter of the room. Roll the flooring with the roller to ensure good contact with the adhesive.

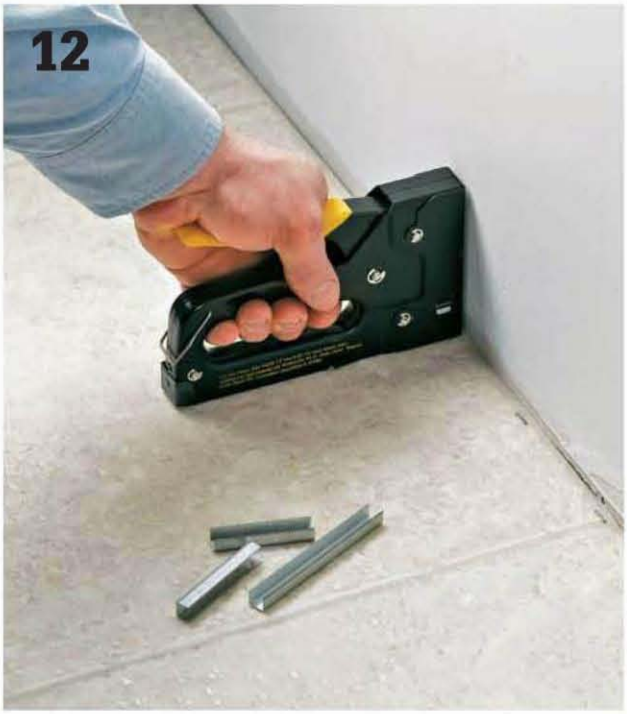

If you're applying flooring over a wood underlayment, fasten the outer edges of the sheet with ⅜" staples driven every 3". Make sure the staples will be covered by the base molding.

Installing Resilient Tile

As with any tile installation, resilient tile requires carefully positioned layout lines. Before committing to any layout and applying tile, conduct a dry run to identify potential problems.

Keep in mind the difference between reference lines (see opposite page) and layout lines. Reference lines mark the center of the room and divide it into quadrants. If the tiles don't lay out symmetrically along these lines, you'll need to adjust them slightly, creating layout lines. Once layout lines are established, installing the tile is a fairly quick process. Be sure to keep joints between the tiles tight and lay the tiles square.

Tiles with an obvious grain pattern can be laid so the grain of each tile is oriented identically throughout the installation. You can also use the quarter-turn method, in which each tile has its pattern grain running perpendicular to that of adjacent tiles. Whichever method you choose, be sure to be consistent throughout the project.

Tools & Materials ▸

- Tape measure
- Chalk line
- Framing square
- Utility knife
- 1⁄16" notched trowel
- Heat gun
- Resilient tile
- Flooring adhesive (for dry-back tile)

Resilient tiles have a pattern layer that is bonded to a vinyl base and coated with a transparent wear layer. Some come with adhesive pre-applied and covered by a paper backing, others have dry backs and are designed to be set into flooring adhesive.

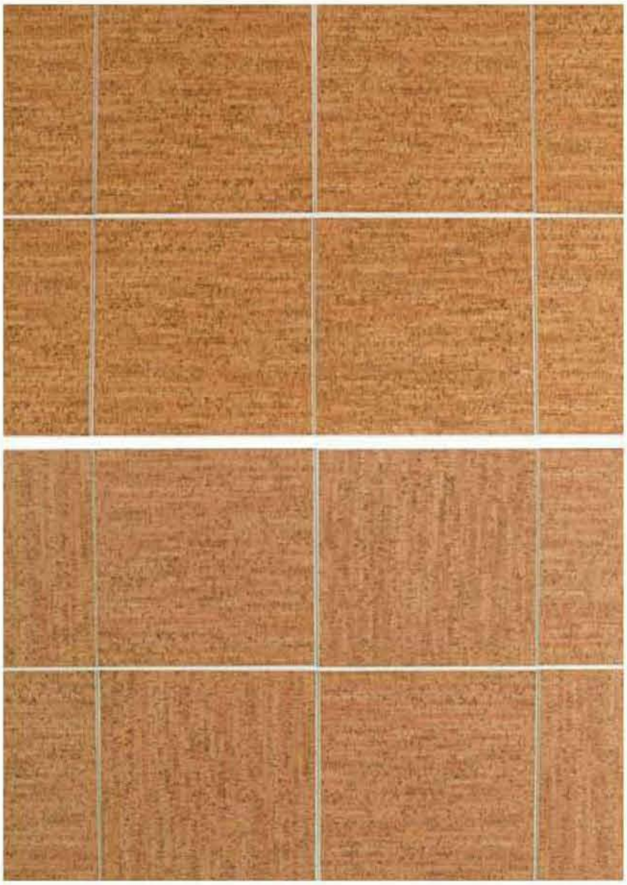

Check for noticeable directional features, like the grain of the vinyl particles. You can set the tiles in a running pattern so the directional feature runs in the same direction (top), or in a checkerboard pattern using the quarter-turn method (bottom).

How to Make Reference Lines for Tile Installation

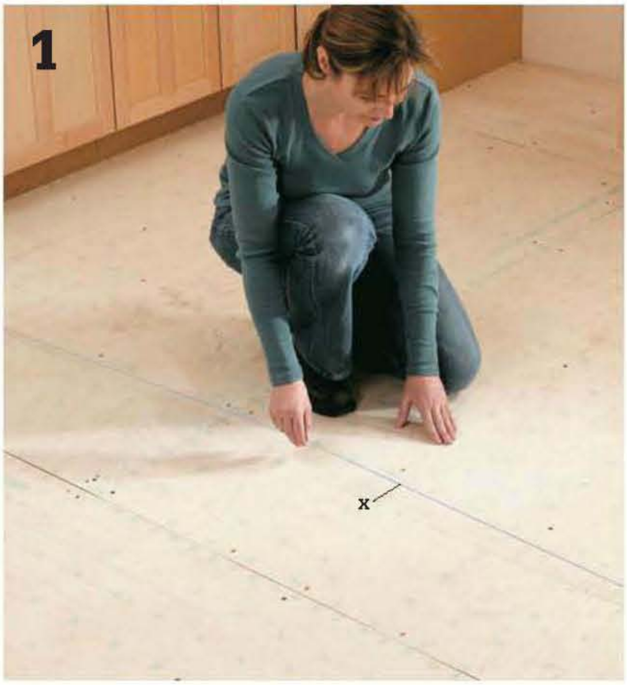

Position a reference line (X) by measuring along opposite sides of the room and marking the center of each side. Snap a chalk line between these marks.

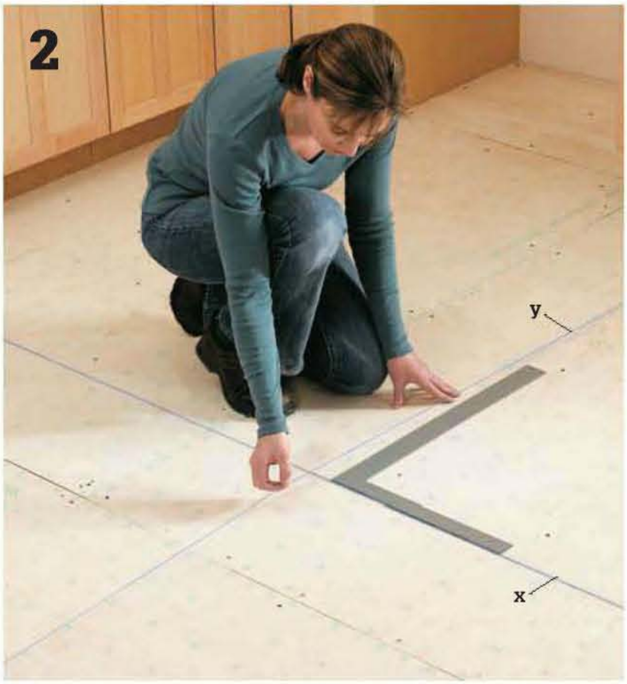

Measure and mark the centerpoint of the chalk line. From this point, use a framing square to establish a second reference line perpendicular to the first one. Snap the second line (Y) across the room.

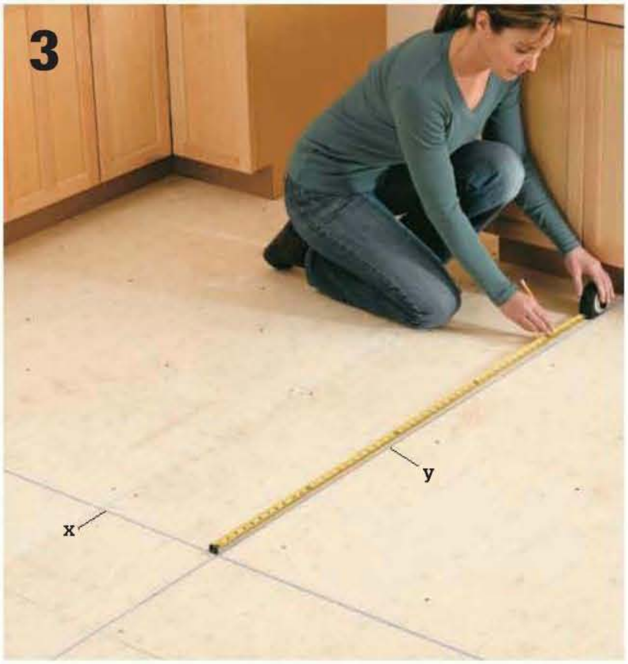

Check the reference lines for squareness using the 3-4-5 triangle method. Measure along reference line X and make a mark 3 ft. from the centerpoint. Measure from the centerpoint along reference line Y and make a mark at 4 ft.

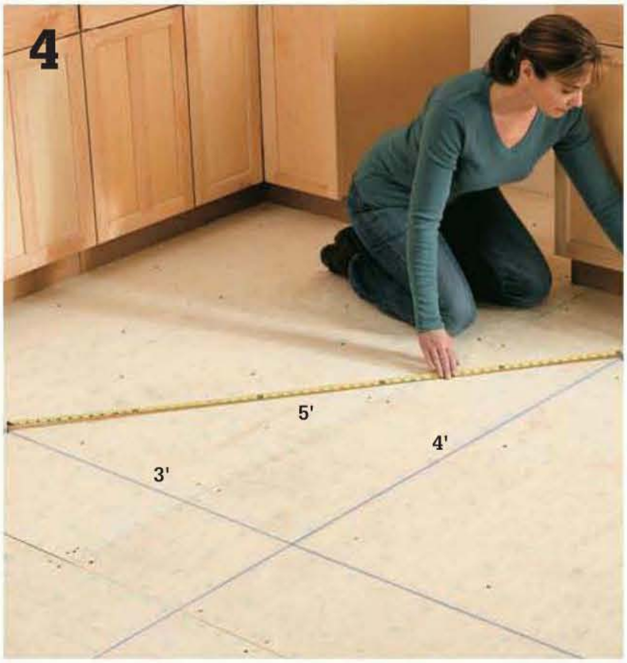

Measure the distance between the marks. If the reference lines are perpendicular, the distance will measure exactly 5 ft. If not, adjust the reference lines until they're exactly perpendicular to each other.

How to Install Dry-backed Resilient Tile

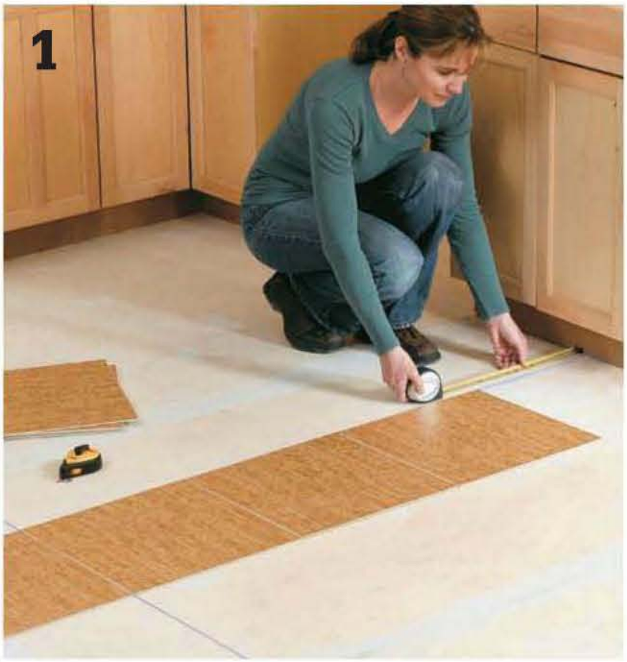

Snap perpendicular reference lines with a chalk line. Dry-fit tiles along layout line Y so a joint falls along reference line X. If necessary, shift the layout to make the layout symmetrical or to reduce the number of tiles that need to be cut.

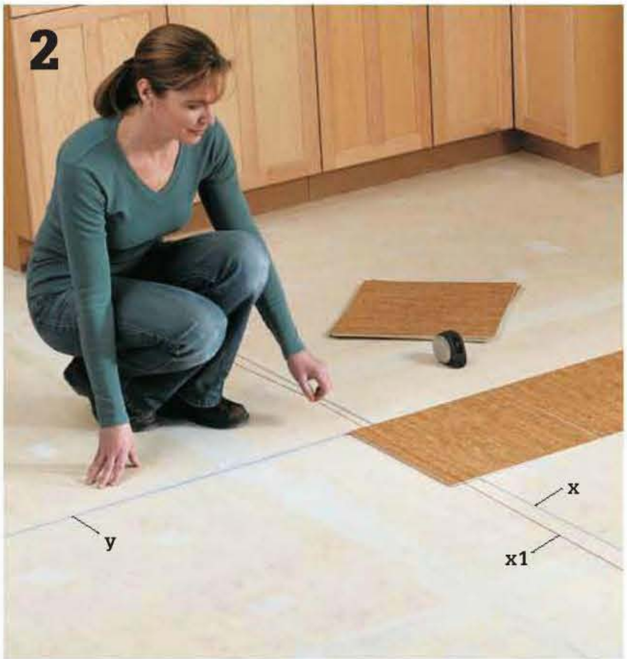

If you shift the tile layout, create a new line that is parallel to reference line X and runs through a tile joint near line X. The new line, X1, is the line you'll use when installing the tile. Use a different colored chalk to distinguish between lines.

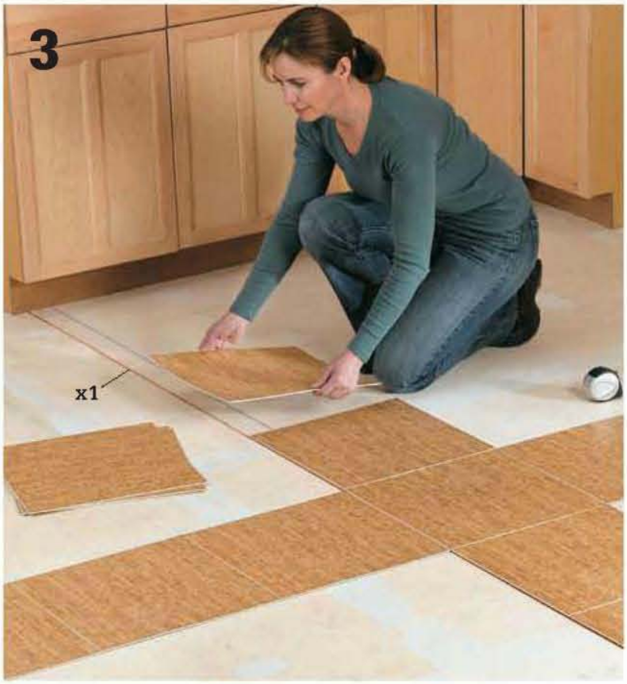

Dry-fit tiles along the new line, X1. If necessary, adjust the layout line as in steps 1 and 2.

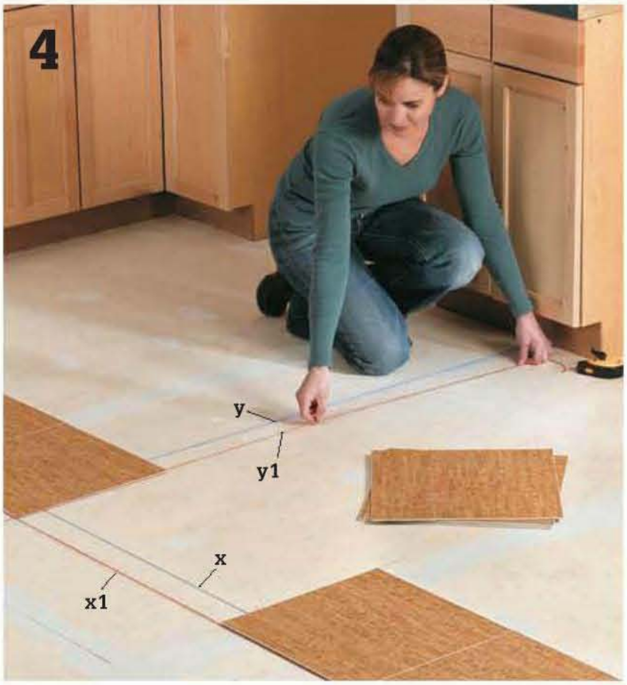

If you adjusted the layout along X1, measure and make a new layout line, Y1, that's parallel to reference line Y and runs through a tile joint. Y1 will form the second layout line you'll use during installation.

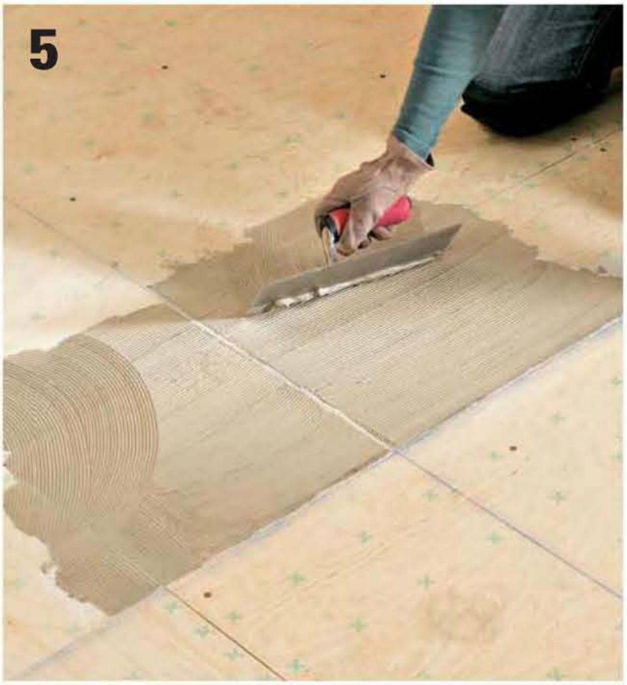

Apply adhesive around the intersection of the layout lines using a trowel with 1⁄16" V-shaped notches. Hold the trowel at a 45° angle and spread adhesive evenly over the surface.

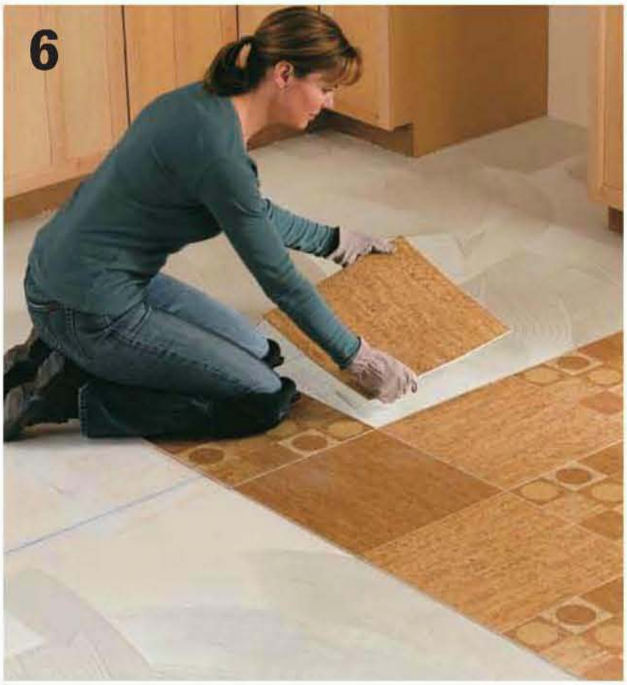

Spread adhesive over most of the installation area, covering three quadrants. Allow the adhesive to set according to the manufacturer's instructions, then begin to install the tile at the intersection of the layout lines. You can kneel on installed tiles to lay additional tiles.

When the first three quadrants are completely tiled, spread adhesive over the remaining quadrant, then finish setting the tile.

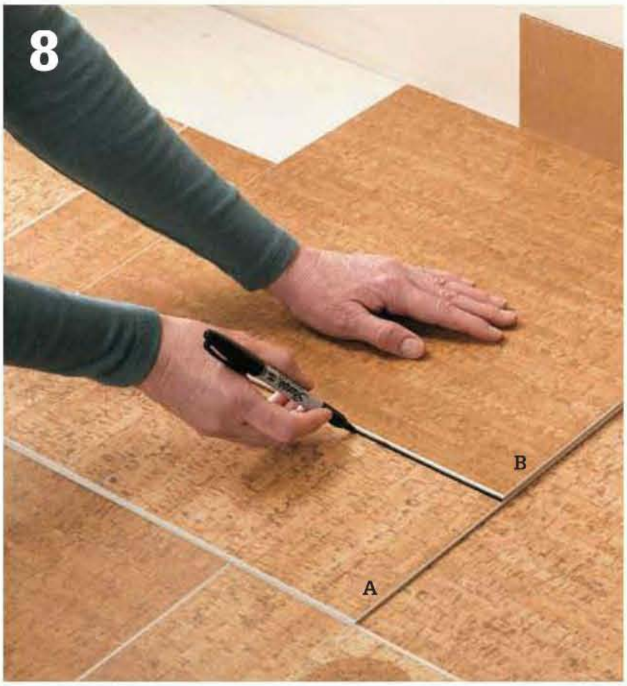

To cut tiles to fit along the walls, place the tile to be cut (A) face up on top of the last full tile you installed. Position a 1⁄8"-thick spacer against the wall, then set a marker tile (B) on top of the tile to be cut. Trace along the edge of the marker tile to draw a cutting line.

(continued)

Tip ▸

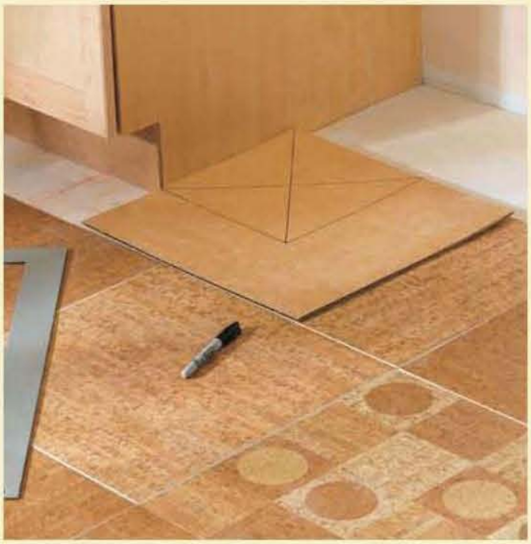

To mark tiles for cutting around outside corners, make a cardboard template to match the space, keeping a ⅛" gap along the walls. After cutting the template, check to make sure it fits. Place the template on a tile and trace its outline.

9

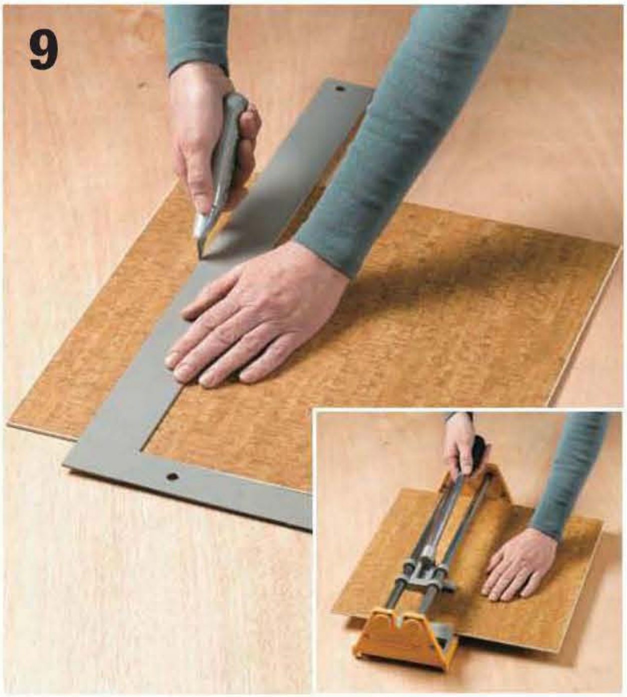

Cut tile to fit using a utility knife and straightedge. Hold the straightedge securely against the cutting line to ensure a straight cut. *Option: You can use a ceramic-tile cutter to make straight cuts in thick vinyl tiles (see inset).*

10

Install cut tiles next to the walls. If you're precutting all tiles before installing them, measure the distance between the wall and install tiles at various points in case the distance changes.

11

Continue installing tile in the remaining quadrants until the room is completely covered. Check the entire floor. If you find loose areas, press down on the tiles to bond them to the underlayment. Install metal threshold bars at room borders where the new floor joins another floor covering.

How to Install Self-adhesive Resilient Tile

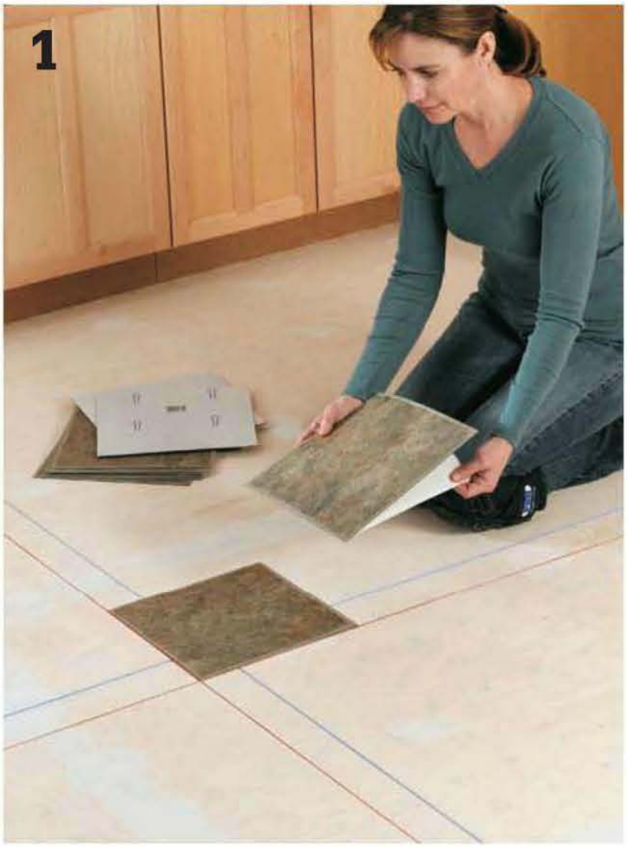

Once your reference lines are established, peel off the paper backing and install the first tile in one of the corners formed by the intersecting layout lines. Lay three or more tiles along each layout lines in the quadrant. Rub the entire surface of each tile to bond the adhesive to the floor underlayment.

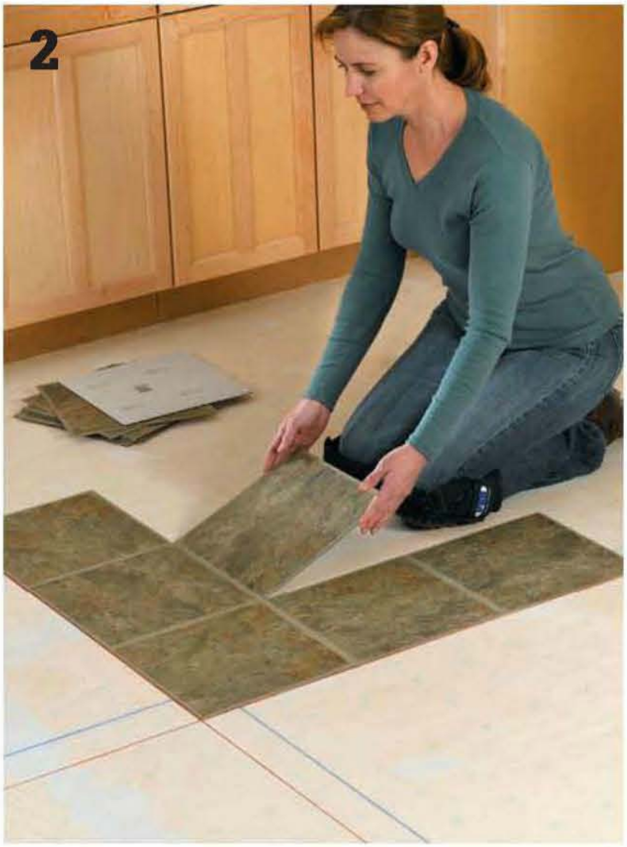

Begin installing tiles in the interior area of the quadrant. Keep the joints tight between tiles..

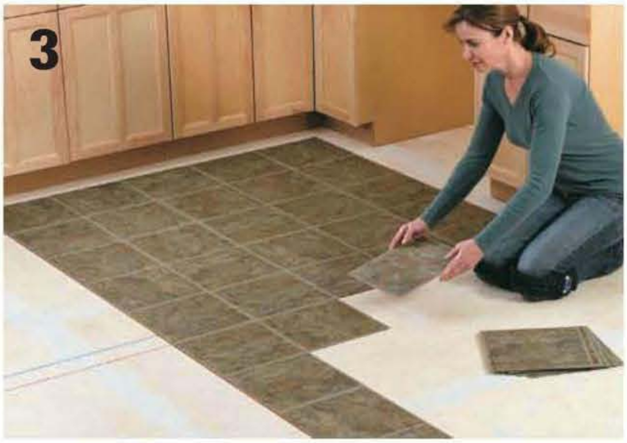

Finish setting full tiles in the first quadrant, then set the full tiles in an adjacent quadrant. Set the tiles along the layout lines first, then fill in the interior tiles.

Continue installing the tile in the remaining quadrants until the room is completely covered. Check the entire floor. If you find loose areas, press down on the tiles to bond them to the underlayment. Install metal threshold bars at room border where the new floor joins another floor covering.

Installing Bamboo Flooring

It looks like hardwood and is available in traditional tongue-and-groove form and in laminate planks. But bamboo is not wood. It's really a grass—and one of the most popular flooring materials today.

Bamboo flooring is made by shredding stalks of the raw material, then pressing them together with a resin that holds the shreds in their finished shape. Not only is bamboo a fast-growing and renewable crop, the companies that make bamboo flooring use binders with low emissions of volatile organic compounds (VOCs). The result is tough, economical, and ecologically friendly. In other words, it's just about perfect for flooring.

If you choose tongue-and-groove bamboo, the installation techniques are the same as for hardwoods. Bamboo is also available as a snap-fit laminate for use in floating floors. In this project we show Teragren Synergy Strand in Java (see Resources, page 554): thin, durable planks that are glued to the underlayment.

Tools & Materials ▸

- Adhesive
- Carpenter's level
- Carpenter's square
- Chalk line
- Cleaning supplies
- Flat-edged trowel
- Marking pen or pencil
- Measuring tape
- Moisture level meter
- Notched trowel
- Rubber mallet
- Scrap lumber
- Shims
- Straightedge
- Weighted roller

Tips for a Successful Installation ▸

60°
70°
RECOMMENDED TEMPERATURE RANGE
40%
60%
RECOMMENDED HUMIDITY RANGE

Bamboo plank flooring should be one of the last items installed on any new construction or remodeling project. All work involving water or moisture should be completed before floor installation. Room temperature and humidity of installation area should be consistent with normal, year-round living conditions for at least a week before installation. Room temperature of 60 to 70° F and humidity range of 40 to 60% is recommended.

About radiant heat: The subfloor should never exceed 85° F. Check the manufacturer's suggested guidelines for correct water temperature inside heating pipes. Switch on the heating unit three days before flooring installation. Room temperature should not vary more than 15° F year-round. For glue-down installations, leave the heating unit on for three days following installation.

How to Install Bamboo Planks

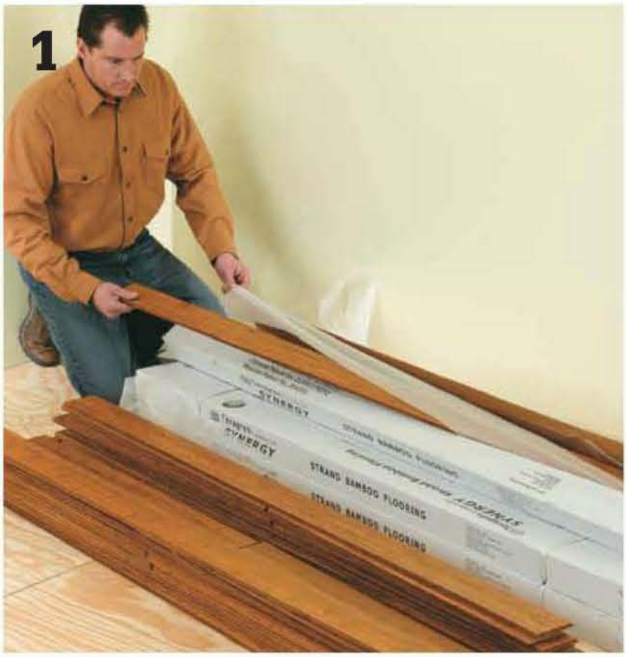

Give the bamboo time to adjust to installation conditions. Store it for at least 72 hours in or near the room where it will be installed. Open the packages for inspection, but do not store the planks on concrete or near outside walls.

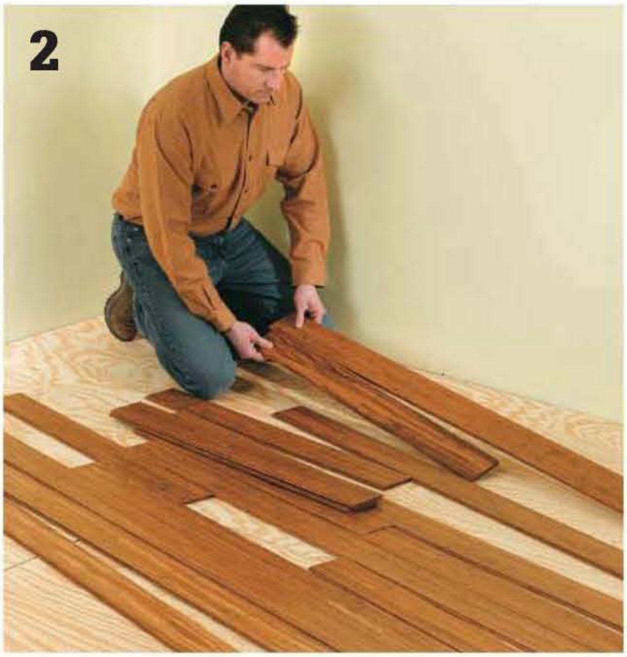

Even though thin-plank bamboo is an engineered material, it can vary in appearance. Buy all planks from the same lot and batch number. Then visually inspect the planks to make sure they match. Use the same lighting as you will have in the finished room.

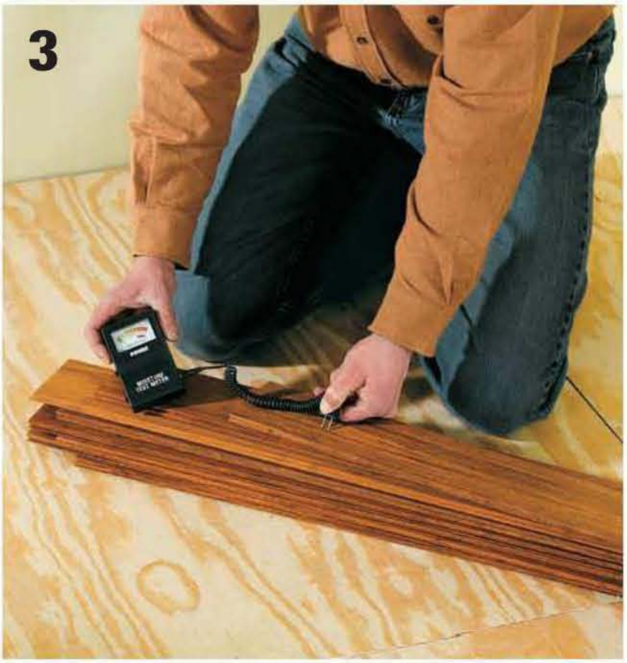

Inspect wood surfaces. The planks and underlayment should have no more than 12% moisture. Bamboo planks can be installed on plywood or oriented strandboard at least ¾" thick. The underlayment must be structurally sound.

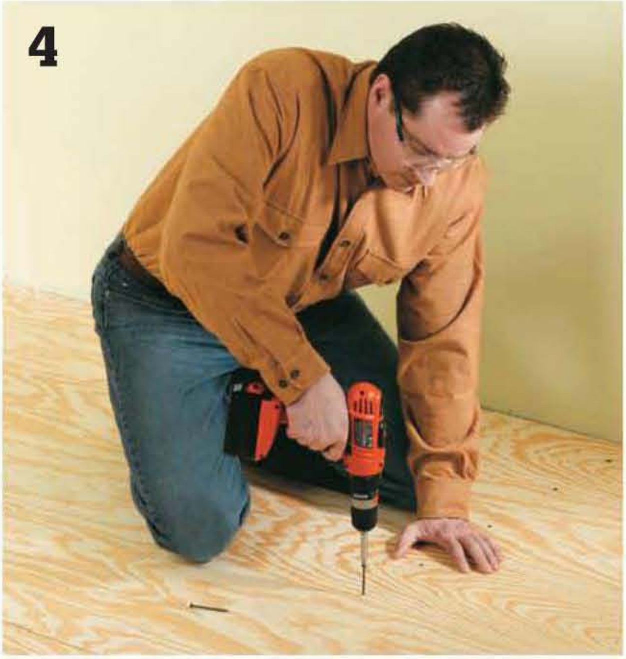

Make sure the underlayment is level. It should not change by more than ⅛" over 10 feet. If necessary, apply a floor leveler to fill any low places, and sand down any high spots. Prevent squeaks by driving screws every 6" into the subfloor below.

(continued)

Sweep and vacuum the floor surface, then measure all room dimensions.

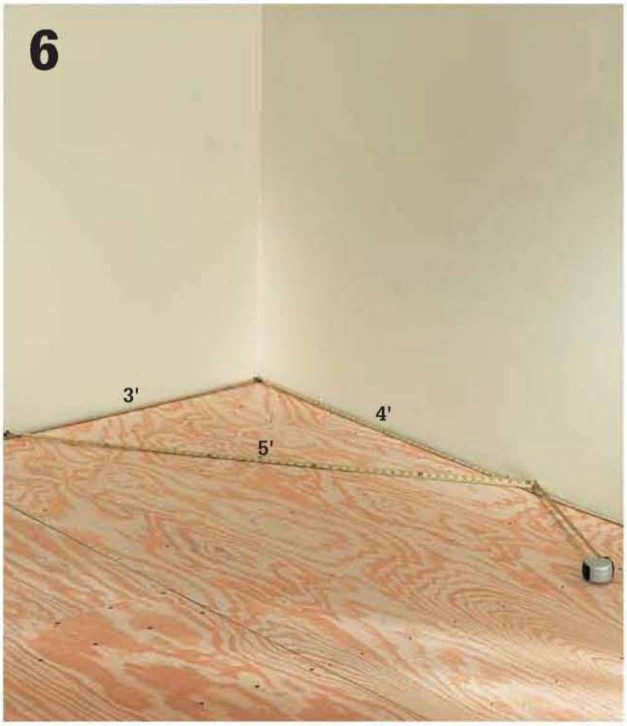

Check corners for squareness using the 3-4-5 triangle method.

The planks should be perpendicular to the floor joists below. Adjust your starting point if necessary. Snap a chalk line next to the longest wall. The distance from the wall should be the same at both ends, leaving ½" for expansion.

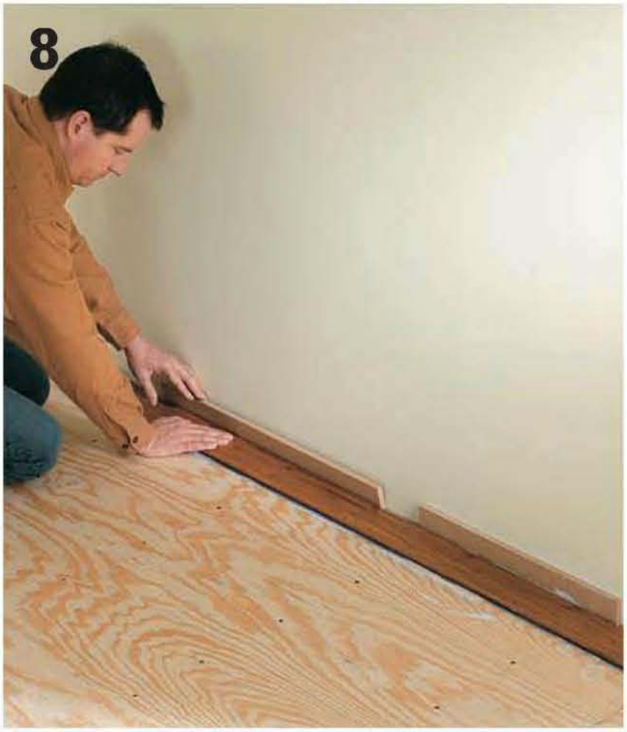

Lay the first course of planks with the tongue edge toward the wall. Align the planks with the chalk line. Hold the edge course in place with wedges or by nailing through the tongue edge. This row will anchor the others, so make sure it stays securely in place.

9

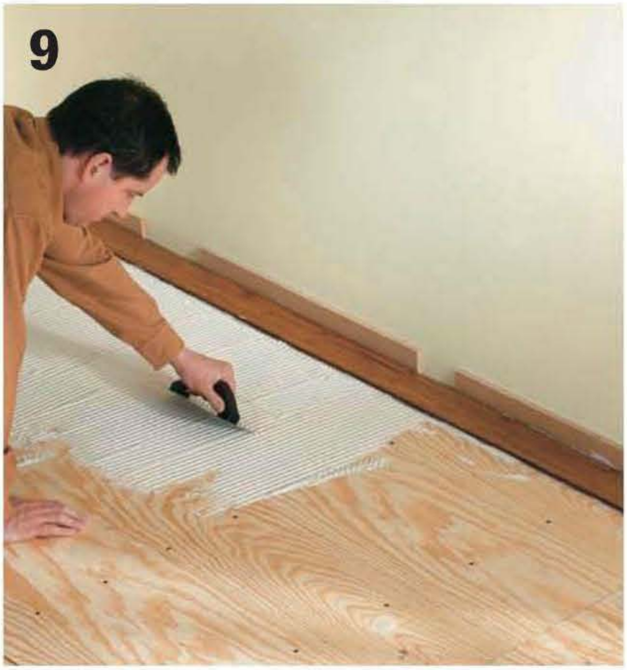

Once the starter row is in place, install the planks using a premium wood-flooring adhesive. Be sure to follow the manufacturer's instructions. Begin at the chalk line and apply enough adhesive to lay down one or two rows of planks. Spread the adhesive with a V-notched trowel at a 45° angle. Let the adhesive sit for the specified time.

10

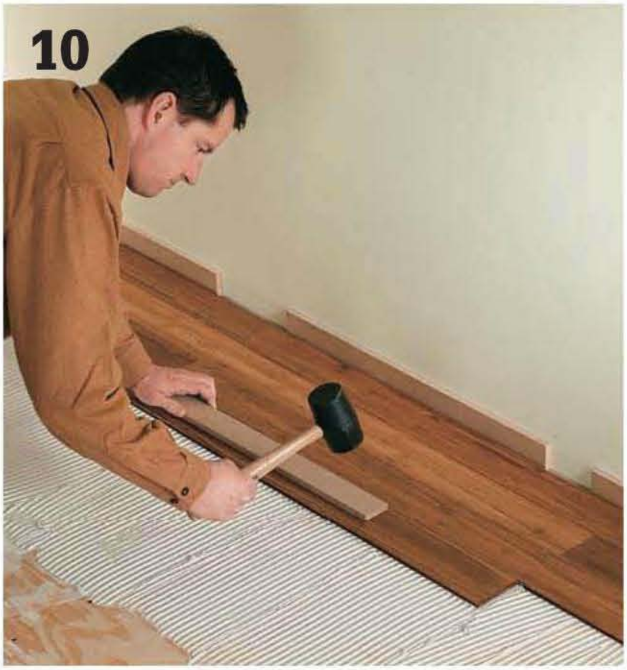

When the adhesive is tacky and ready to use, lay the first section of bamboo planks. Set each plank in the adhesive by placing a clean piece of scrap lumber on top and tapping it down with a rubber mallet. Check the edge of each section to make sure it keeps a straight line.

11

After you finish the first section, cover the next area with adhesive and give it time to become tacky. This slows down the project, but it prevents you from using more adhesive than you can use—and it allows the section you just finished to set up.

12

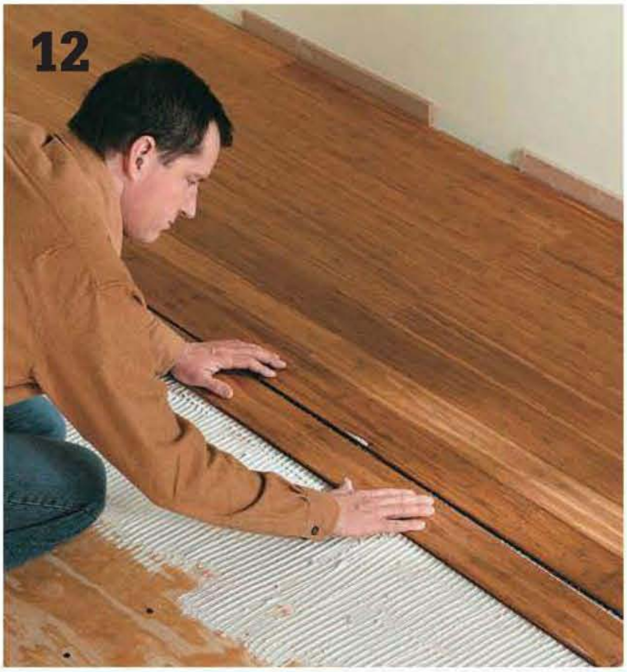

When the adhesive is ready, lay down the next section of planks. Fit the new planks tightly against the previous section, taking care not to knock the finished section out of alignment. If the planks have tongue-and-groove edges, fit them carefully into place.

(continued)

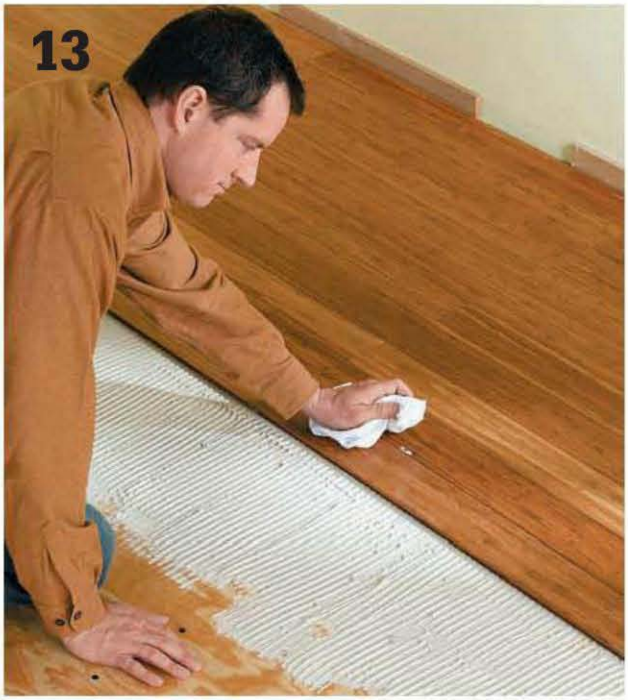

Continue applying adhesive and installing planks, one section at a time, to cover the entire floor. When adhesive gets on the flooring surface, wipe it off quickly.

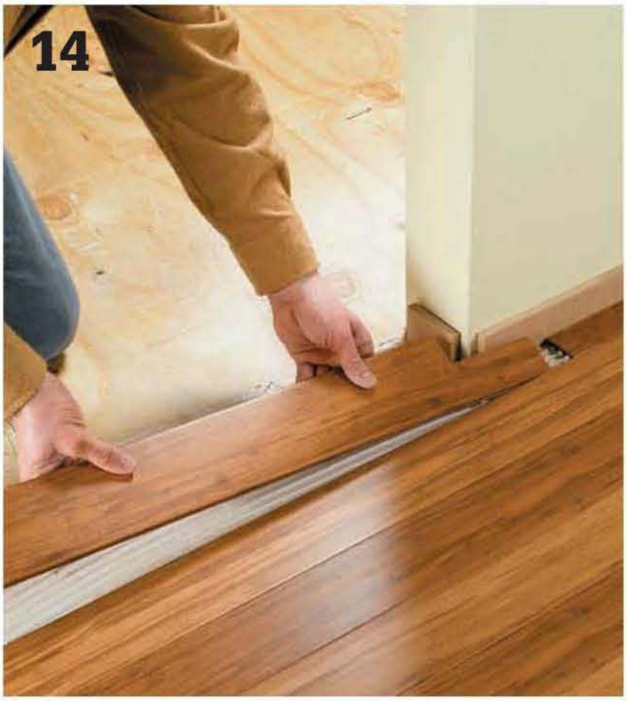

At the edges and around any fixed objects, such as doorways or plumbing pipes, leave a ½" gap for expansion. Use shims to maintain the gaps if needed. These spaces can be covered with baseboards, base shoes, and escutcheons.

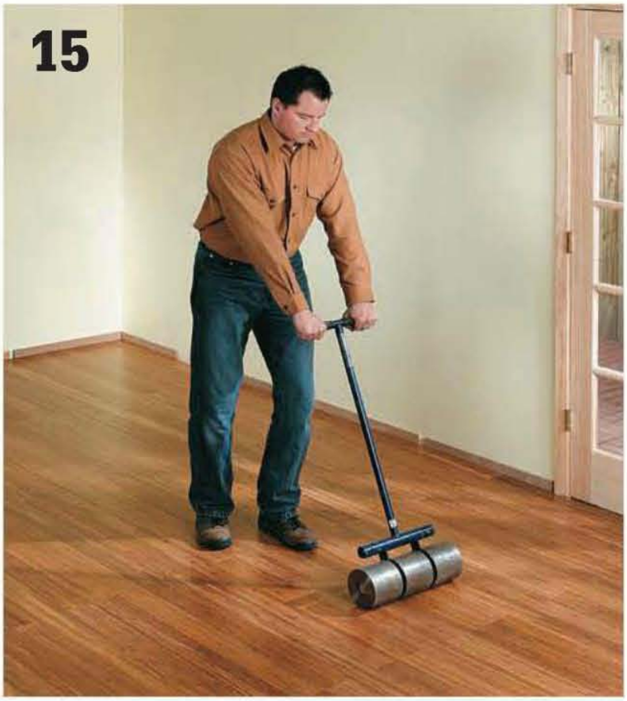

As you finish each section, walk across it a few times to maximize contact between the planks and the adhesive. When all the planks are in place, clean the surface and use a clean weighted roller. Push the roller in several directions, covering the entire surface many times.

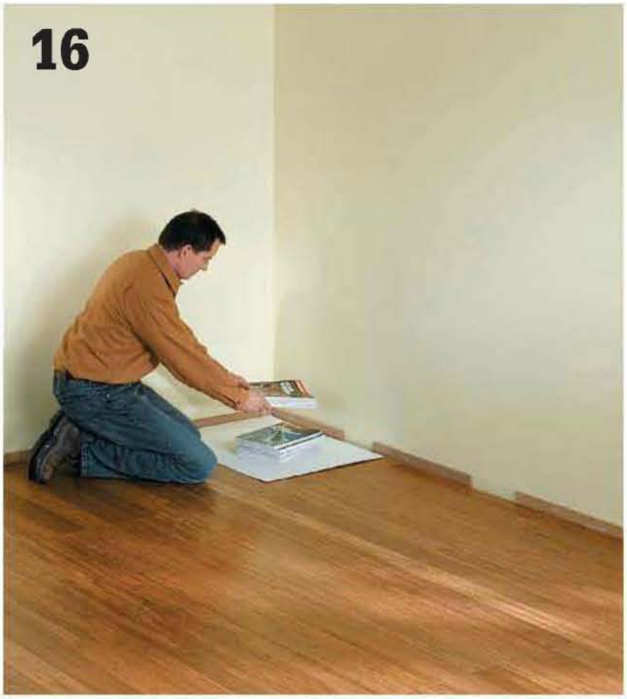

In places that are difficult to reach with a roller, lay down a sheet of protective material, such as butcher paper, and stack weights on the paper. Let the finished floor sit for at least 24 hours, then clean the surface and remove any spacers from the expansion gaps. Finally, install the finishing trim.

Installing Ceramic Tile

Ceramic tile installation starts with the same steps as installing resilient tile. You snap perpendicular reference lines and dry-fit tiles to ensure the best placement.

When setting tiles, work in small sections so the mortar doesn't dry before the tiles are set. Use spacers between tiles to ensure consistent spacing. Plan an installation sequence to avoid kneeling on set tiles. Be careful not to kneel or walk on tiles until the designated drying period is over.

Tools & Materials ▸

- ¼" square trowel
- Rubber mallet
- Tile cutter
- Tile nippers
- Hand-held tile cutter
- Needlenose pliers
- Grout float
- Grout sponge
- Soft cloth
- Small paint brush
- Thin-set mortar
- Tile
- Tile spacers
- Grout
- Latex grout additive
- Wall adhesive
- 2 × 4 lumber
- Grout sealer
- Tile caulk
- Sponge brush

Ceramic floor tile continues to be one of the most popular DIY floorcoverings. It is relatively easy to install, is very durable and, depending on the style you choose, can be quite affordable.

How to Install Ceramic Tile

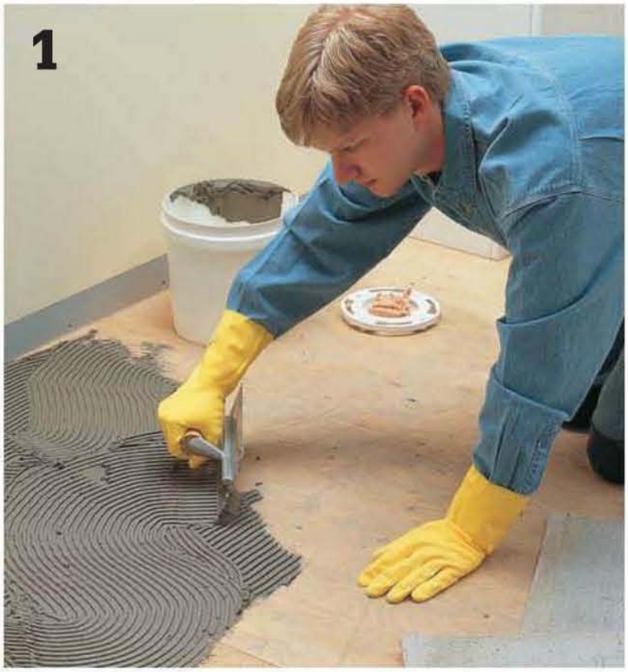

Make sure the subfloor is smooth, level, and stable. Spread thin-set mortar on the subfloor for one sheet of cementboard. Place the cementboard on the mortar, keeping a ¼" gap along the walls.

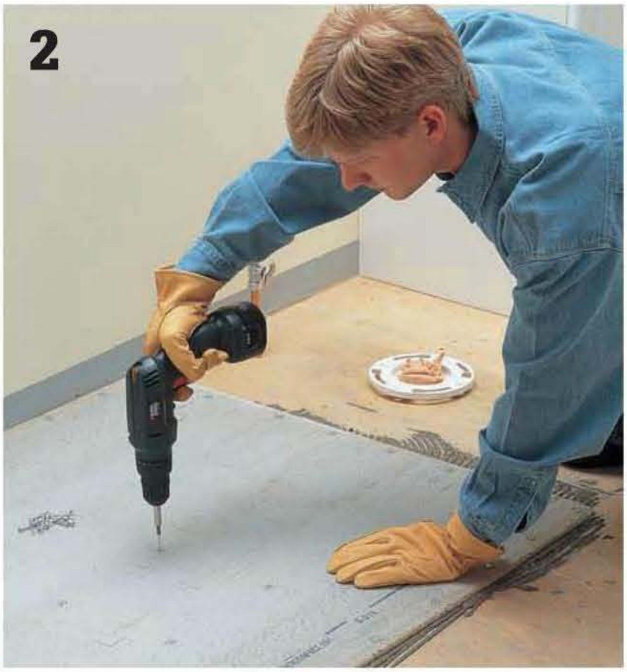

Fasten it in place with 1¼" cementboard screws. Place fiberglass-mesh wallboard tape over the seams. Cover the remainder of the floor, following the steps on page 152.

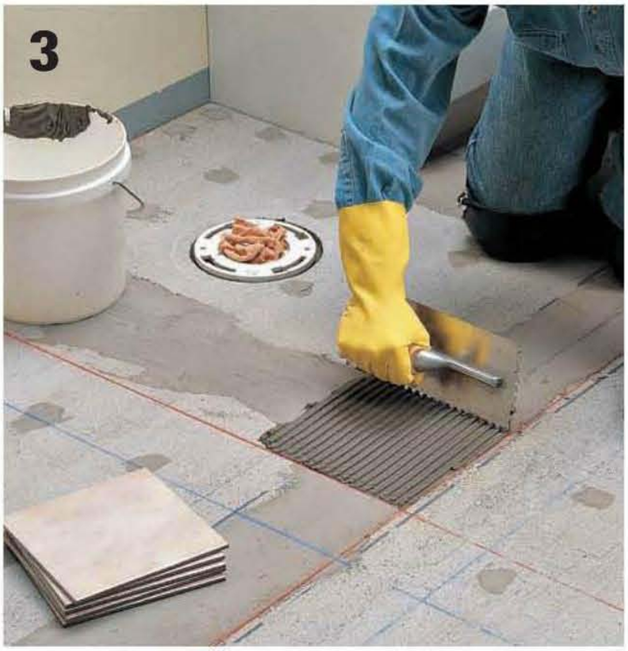

Draw reference lines and establish the tile layout (see page 181). Mix a batch of thin-set mortar, then spread the mortar evenly against both reference lines of one quadrant, using a ¼" square-notched trowel. Use the notched edge of the trowel to create furrows in the mortar bed.

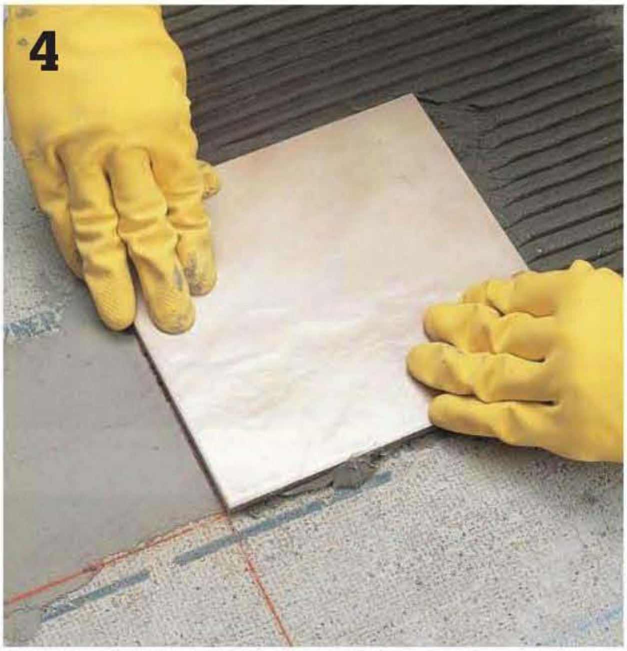

Set the first tile in the corner of the quadrant where the reference lines intersect. When setting tiles that are 8" square or larger, twist each tile slightly as you set it into position.

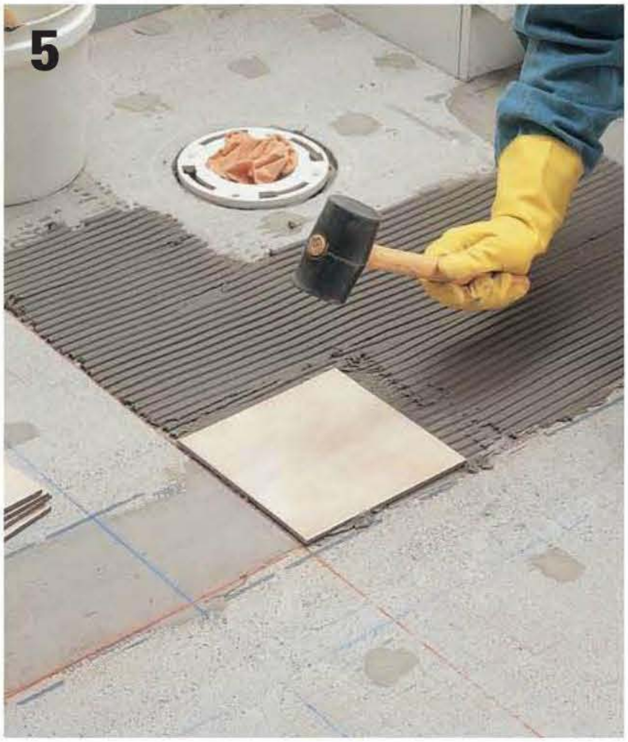

Using a soft rubber mallet, gently tap the central area of each tile a few times to set it evenly into the mortar.

Variation: For large tiles or uneven stone, use a larger trowel with notches that are at least ½" deep.

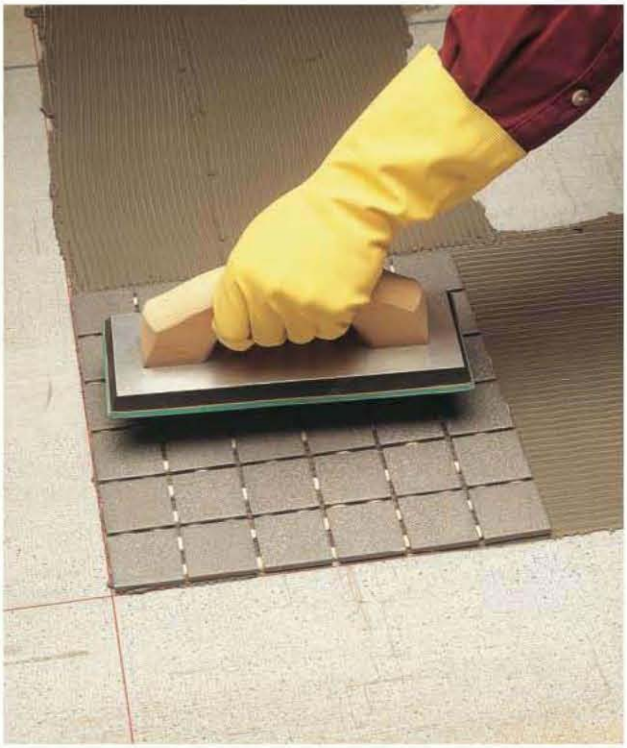

Variation: For mosaic sheets, use a $\frac{3}{16}$" V-notched trowel to spread the mortar and a grout float to press the sheets into the mortar. Apply pressure gently to avoid creating an uneven surface.

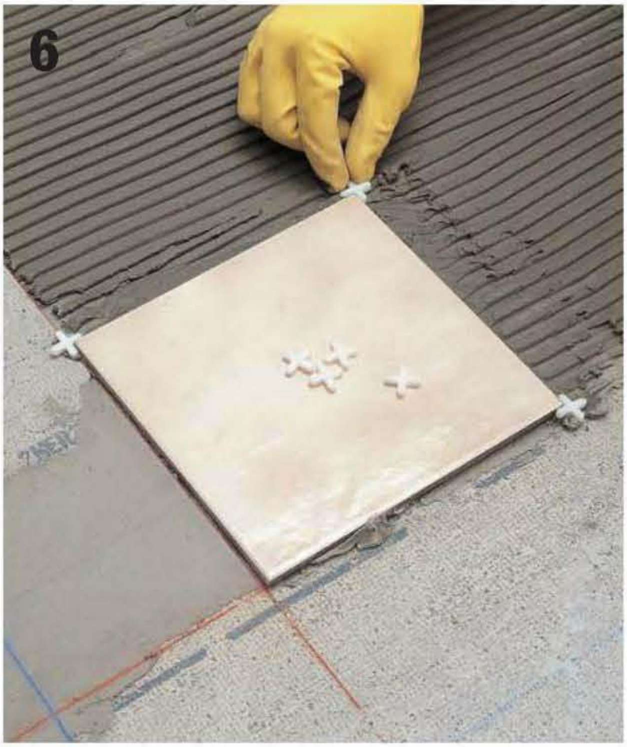

To ensure consistent spacing between tiles, place plastic tile spacers at the corners of the set tile. With mosaic sheets, use spacers equal to the gaps between tiles.

(continued)

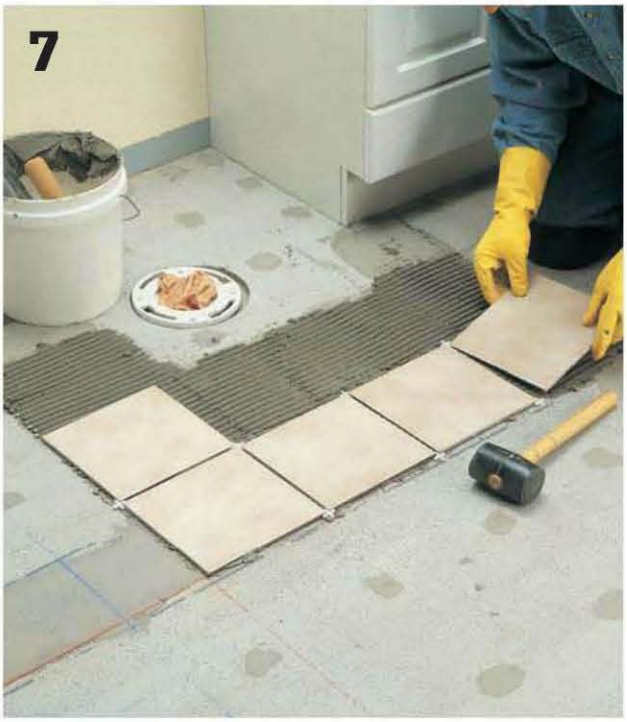

Position and set adjacent tiles into the mortar along the reference lines. Make sure the tiles fit neatly against the spacers.

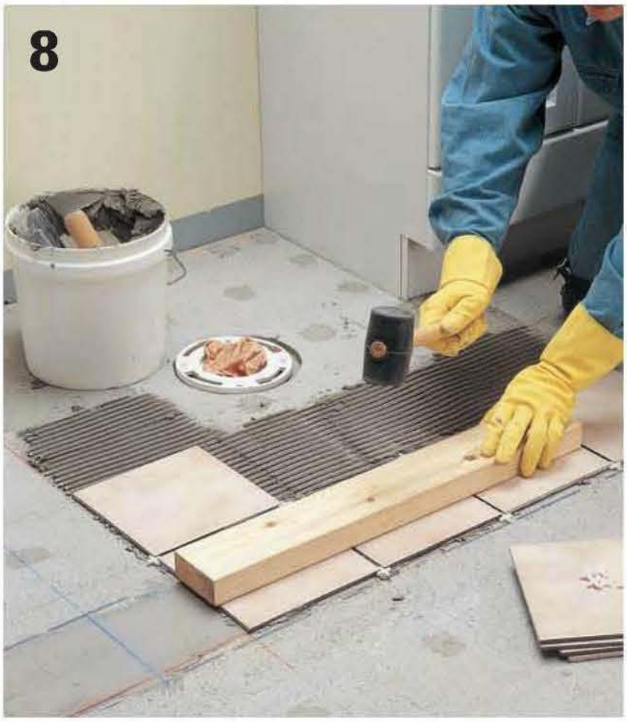

To make sure the tiles are level with one another, place a straight piece of 2 × 4 across several tiles, then tap the board with a mallet.

Lay tile in the remaining area covered with mortar. Repeat steps 2 to 7, continuing to work in small sections, until you reach walls or fixtures.

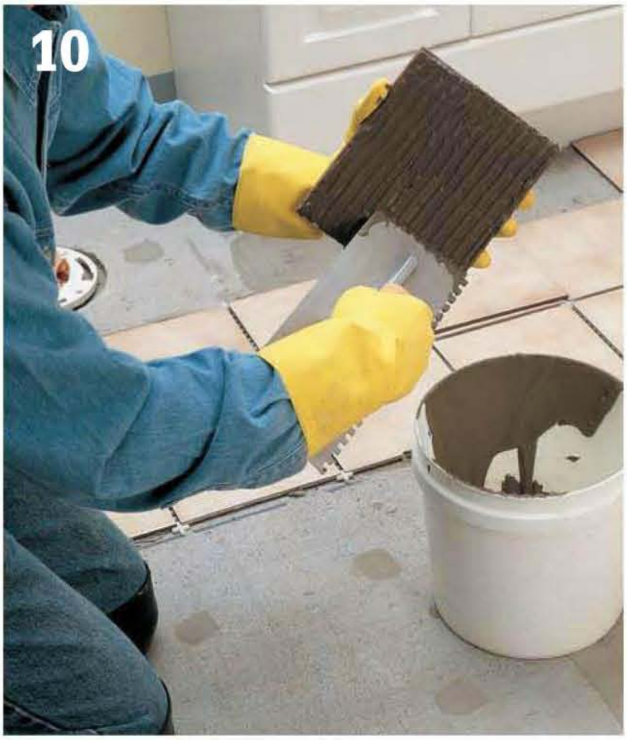

Measure and mark tiles to fit against walls and into corners (see pages 183 to 184). Cut the tiles to fit. Apply thin-set mortar directly to the back of the cut tiles, instead of the floor, using the notched edge of the trowel to furrow the mortar

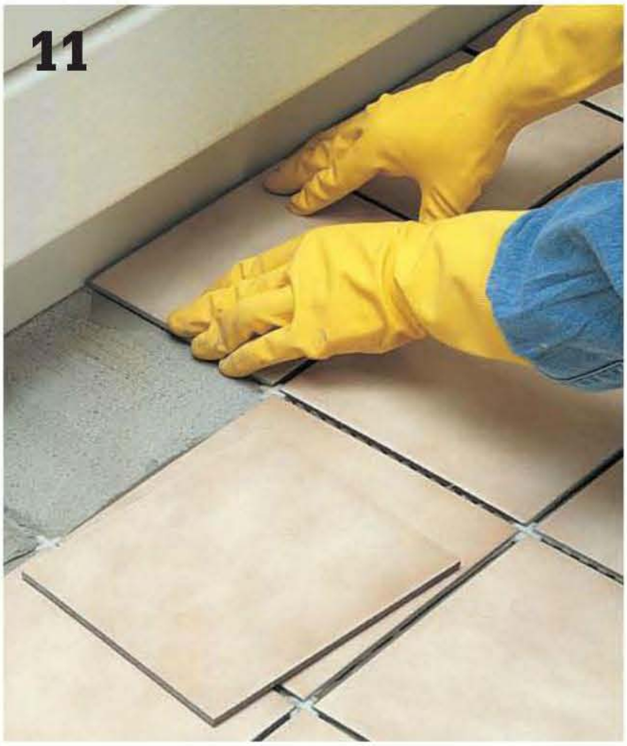

Set the cut pieces of tile into position. Press down on the tile until each piece is level with adjacent tiles.

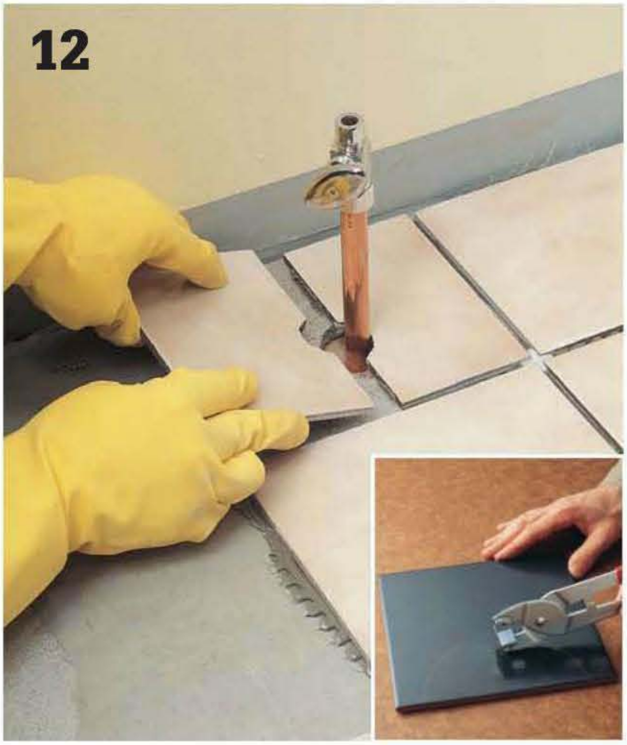

Measure, cut, and install tiles that require notches or curves to fit around obstacles (inset), such as exposed pipes or toilet drains.

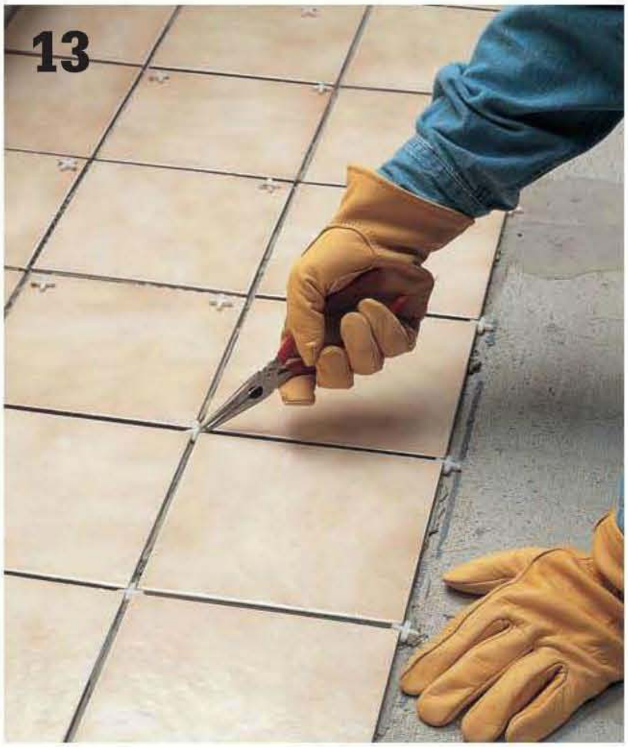

Carefully remove the spacers with needlenose pliers before the mortar hardens.

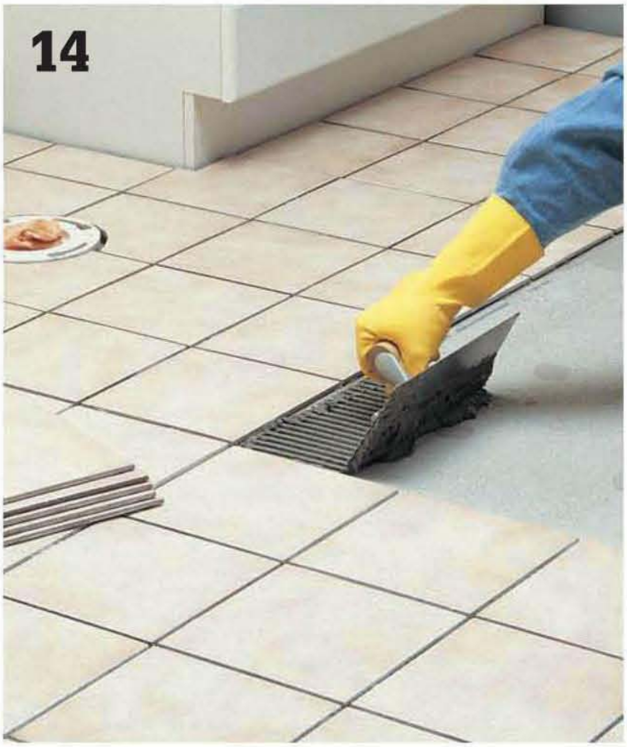

Apply mortar and set tiles in the remaining quadrants, completing one quadrant before starting the next. Inspect all of the tile joints and use a utility knife or grout knife to remove any high spots of mortar that could show through the grout.

(continued)

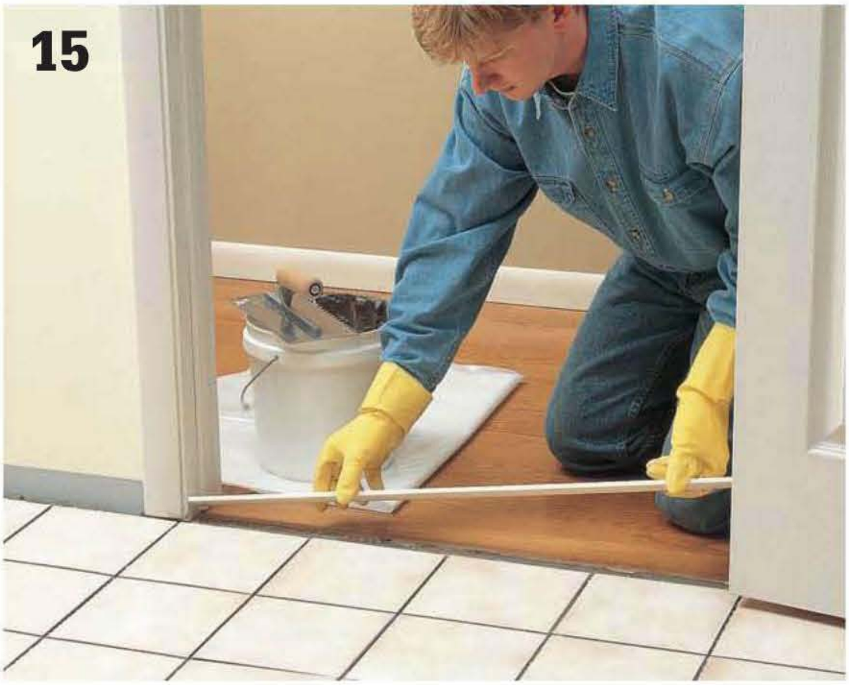

Install threshold material in doorways. If the threshold is too long for the doorway, cut it to fit with a jigsaw or circular saw and a tungsten-carbide blade. Set the threshold in thin-set mortar so the top is even with the tile. Keep the same space between the threshold as between tiles. Let the mortar set for at least 24 hours.

Prepare a small batch of floor grout to fill the tile joints. When mixing grout for porous tile, such as quarry or natural stone, use an additive with a release agent to prevent grout from bonding to the tile surfaces.

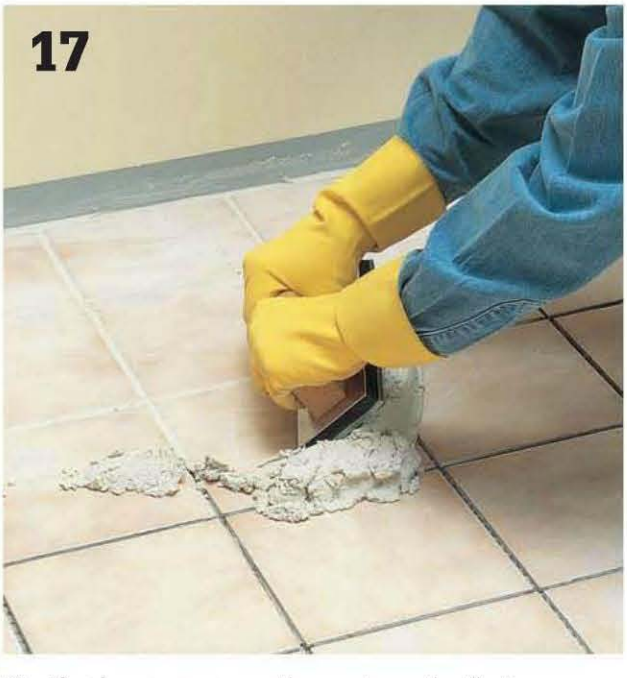

Starting in a corner, pour the grout over the tile. Use a rubber grout float to spread the grout outward from the corner, pressing firmly on the float to completely fill the joints. For best results, tilt the float at a 60° angle to the floor and use a figure eight motion.

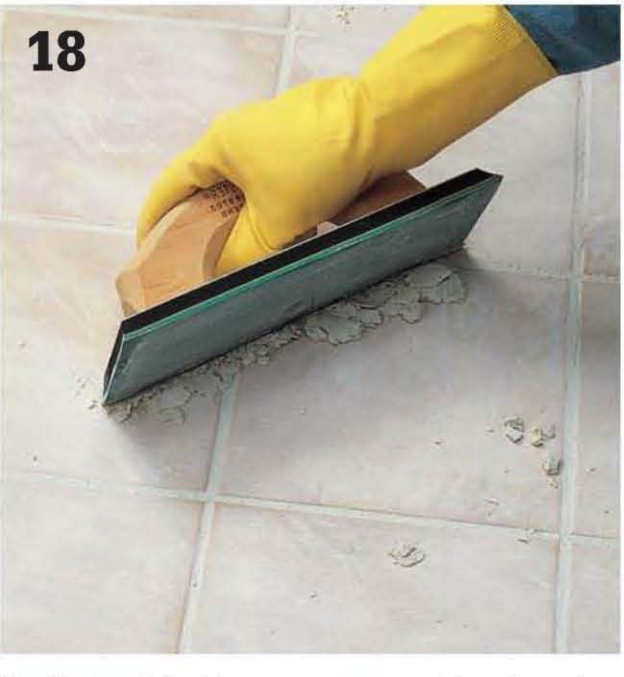

Use the grout float to remove excess grout from the surface of the tile. Wipe diagonally across the joints, holding the float in a near-vertical position. Continue applying grout and wiping off excess until about 25 square feet of the floor has been grouted.

Wipe a damp grout sponge diagonally over about 2 square feet of the floor at a time. Rinse the sponge in cool water between wipes. Wipe each area only once since repeated wiping can pull grout back out of joints. Repeat steps 17 to 18 to apply grout to remainder of floor.

Allow the grout to dry for about 4 hours, then use a soft cloth to buff the tile surface and remove any remaining grout film.

Apply grout sealer to the grout lines, using a small sponge brush or sash brush. Avoid brushing sealer on to the tile surfaces. Wipe up any excess sealer immediately.

Variation: Use a tile sealer to seal porous tile, such as quarry tile or unglazed tile. Following the manufacturer's instructions, roll a thin coat of sealer over the tile and grout joints, using a paint roller and extension handle.

Installing Mosaic Tile

Mosaic tile is an excellent choice for smaller areas. It requires the same preparation and handling as larger tiles, with a few differences. Sheets of mosaic tile are held together by a fabric mesh backing. This makes them more difficult to hold, place, and move. They may not be square with your guidelines when you first lay them down. And mosaic tiles require many more temporary spacers and much more grout.

A few cautions: Variations in color and texture are just as likely among mosaic tile as individual tiles, so buy all your tile from the same lot and batch. Mortar or mastic intended for ceramic tile may not work with glass mosaic tile. Finally, in projects where the finished project will be exposed to the elements, make sure you have adhesive and grout suitable for outdoor use.

Mosaic tiles come in sheets (usually 12 × 12") and can be made from ceramic, porcelain, glass or any number of designer materials. Normally installed for their appearance, mosaics are relatively high maintenance and prone to cracks because of all the grout lines.

Tools & Materials ▸

- Carpenter's square
- Chalk line
- Cleaning supplies
- Coarse sponge
- Craft/utility knife
- Grout sealer
- Marking pen or pencil
- Measuring tape
- Notched trowel
- Recommended adhesive
- Rubber mallet
- Sanded grout
- Scrap lumber
- Straightedge
- Tile nippers
- Tile spacers

How to Install Mosaic Tile

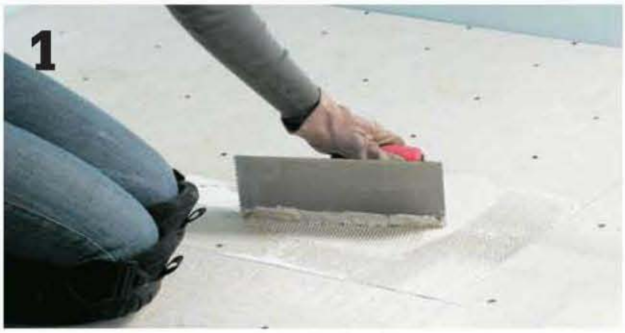

Clean and prepare the area and then draw reference lines (see page 181). Beginning at the center intersection, apply the recommended adhesive to one quadrant. Spread it outward evenly with a notched trowel. Lay down only as much adhesive as you can cover in 10 to 15 minutes.

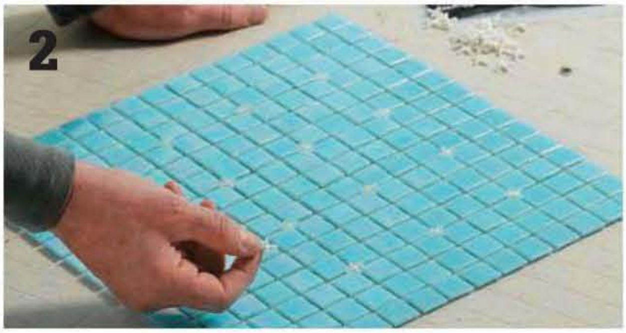

Select a sheet of mosaic tile. Place several plastic spacers within the grid so that the sheet remains square. Pick up the sheet of tiles by diagonally opposite corners. This will help you hold the edges up so that you don't trap empty space in the middle of the sheet.

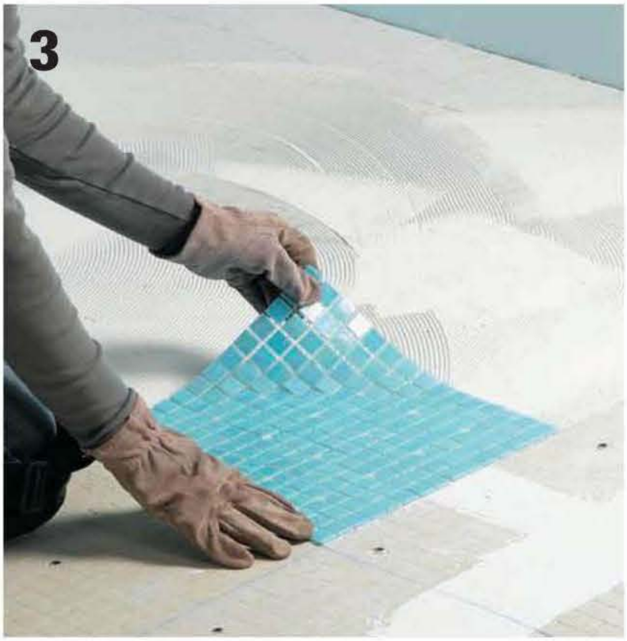

3

Gently press one corner into place on the adhesive. Slowly lower the opposite corner, making sure the sides remain square with your reference lines. Massage the sheet into the adhesive, being careful not to press too hard or twist the sheet out of position. Insert a few spacers in the outside edges of the sheet you have just placed. This will help keep the grout lines consistent.

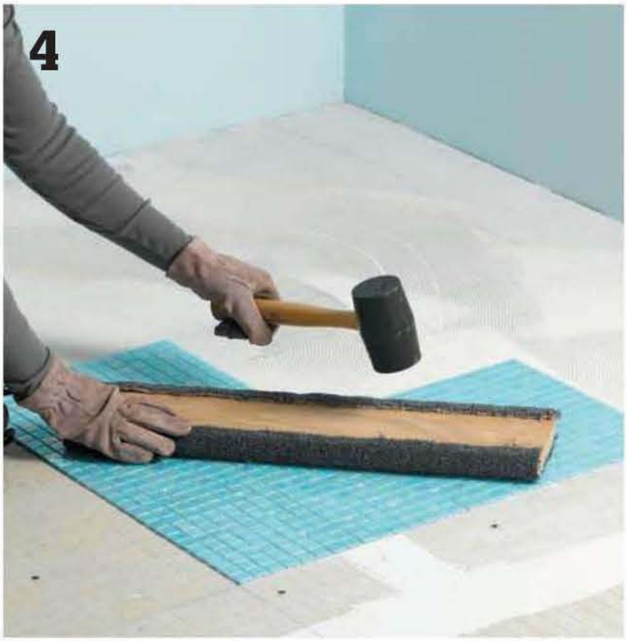

4

When you have placed two or three sheets, lay a scrap piece of flat lumber across the tops and tap the wood with a rubber mallet to set the fabric mesh in the adhesive, and to force out any trapped air.

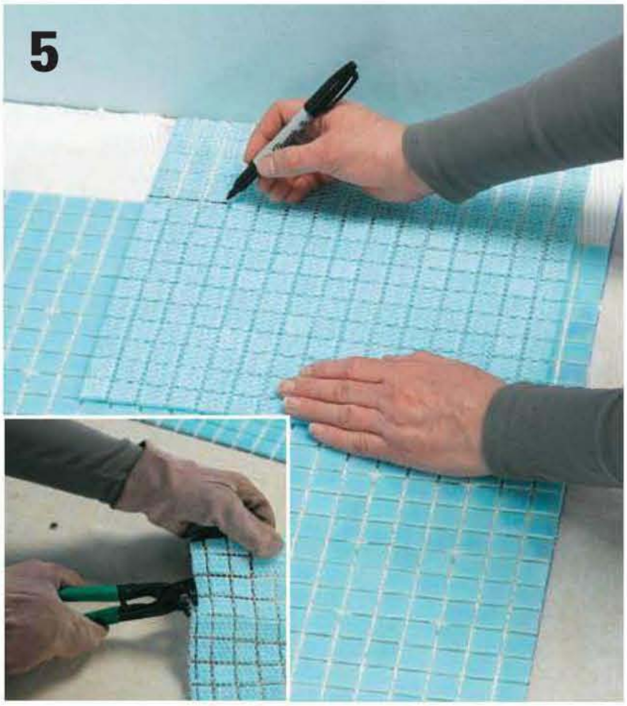

5

At the outer edges of your work area, you will probably need to trim one or more rows from the last sheet. If the space left at the edge is more than the width of a regular grout line, use tile nippers to trim the last row that will fit. Save these leftover tiles for repairs.

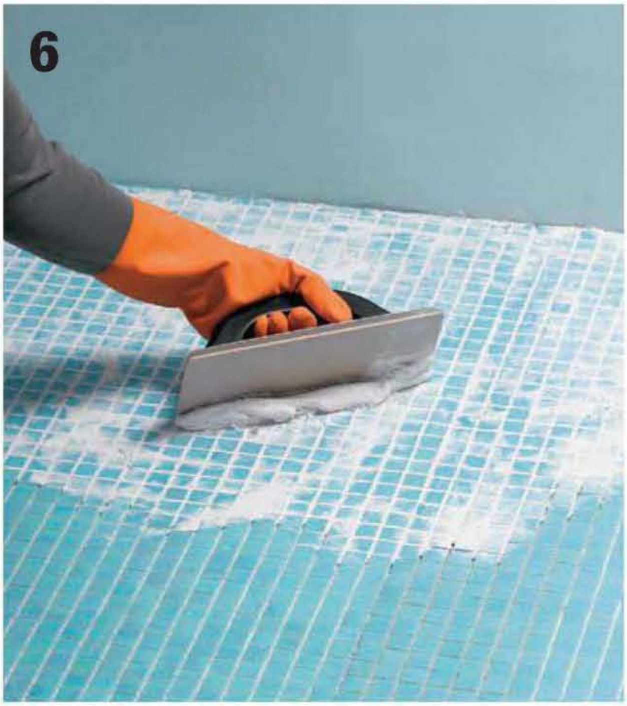

6

After the adhesive has cured, usually 24 to 48 hours, apply grout (see pages 196 to 197). With many more spaces, mosaic tiles will require more grout. Follow the manufacturer's instructions for spreading and floating the grout. Clean up using the instructions for individual tiles (see page 197).

Installing Carpet Squares

Most carpeting has a single design and is stretched from wall to wall. It covers more square feet of American homes than any other material. But if you want a soft floor covering that gives you more options, carpet squares are an excellent choice.

Manufacturers have found ways to create attractive new carpet using recycled fibers. This not only reuses material that would otherwise become landfill, it reduces waste in manufacturing as well. So, instead of adding to problems of resource consumption and pollution, carpet squares made from recycled materials help reduce them.

The squares are attached to each other and to the floor with adhesive dots. They can be installed on most clean, level, dry underlayment or existing floor. If the surface underneath is waxed or varnished, check with the manufacturer before you use any adhesives on it.

Tools & Materials ▸

- Adhesive
- Aviator's snips
- Carpenter's square
- Chalk line
- Cleaning supplies
- Craft/utility knife
- Flat-edged trowel
- Marking pen or pencil
- Measuring tape
- Notched trowel
- Straightedge

Carpet tiles combine the warmth and comfort of carpet with do-it-yourself installation, custom designs, and easy replacement. They can be laid wall-to-wall or in an area rug style, as shown above.

How to Install Carpet Squares

1

Take the squares out of the package. Be sure the room is well ventilated. Carpet squares should be at room temperature for at least 12 hours before you lay them down.

2

FLOR
FLOR
FLOR
FLOR

Check the requirements for the recommended adhesive. You can install carpet squares over many other flooring materials, including hardwood, laminates, and resilient sheets or tiles. The carpet squares shown here are fastened with adhesive dots, so almost any existing floor provides a usable surface.

3

Make sure the existing floor is clean, smooth, stable, and dry. Use floor leveler if necessary to eliminate any hills or valleys. If any part of the floor is loose, secure it to the subfloor or underlayment before you install the carpet squares. Vacuum the surface and wipe it with a damp cloth.

4

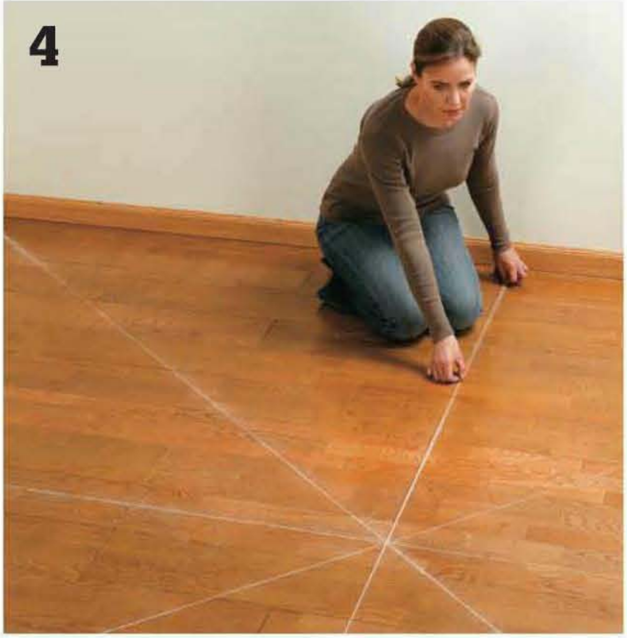

Snap chalk lines between diagonally opposite corners to find the center point for the room. In rooms with unusual shapes, determine the visual center and mark it. Next, snap chalk lines across the center and perpendicular to the walls. This set of guidelines will show you where to start.

(continued)

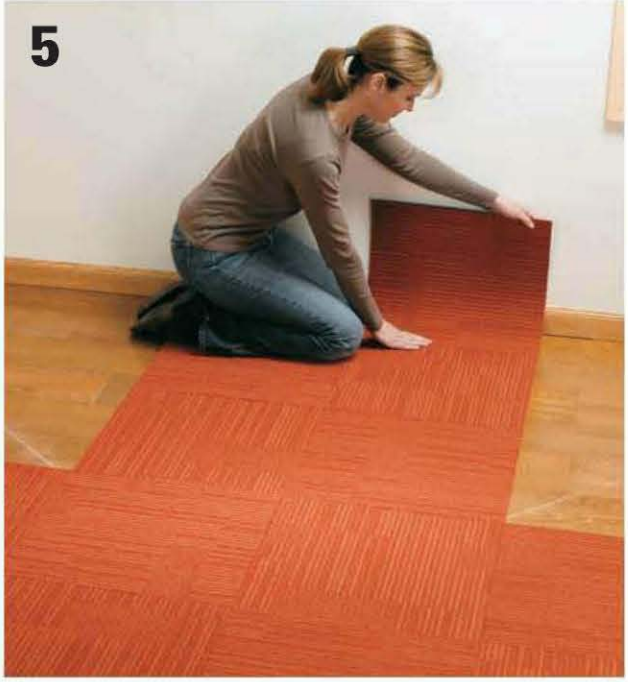

Lay a base row of carpet squares on each side of the two guidelines. When you reach the walls, make note of how much you will need to cut. You should have the same amount to cut on each side. If not, adjust the center point and realign the squares.

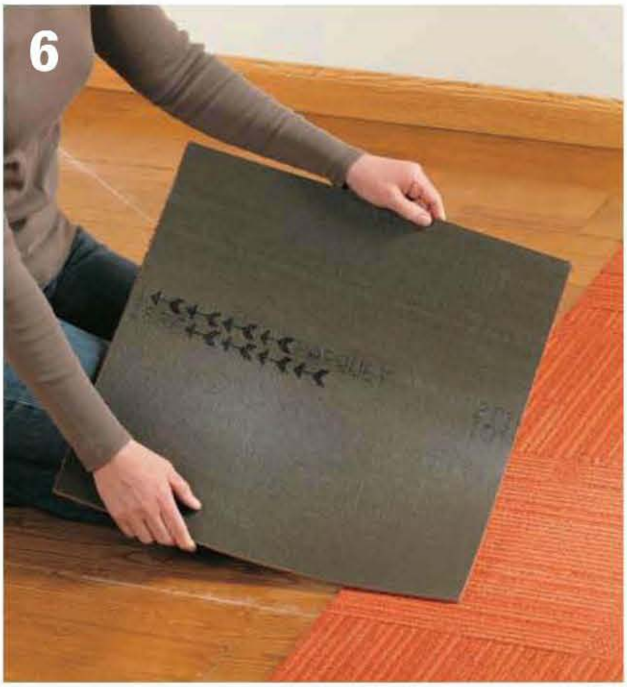

Check the backs of the squares before you apply any adhesive. They should indicate a direction, using arrows or other marks, so that the finished pile has a consistent appearance. If you plan to mix colors, this is the time to establish your pattern.

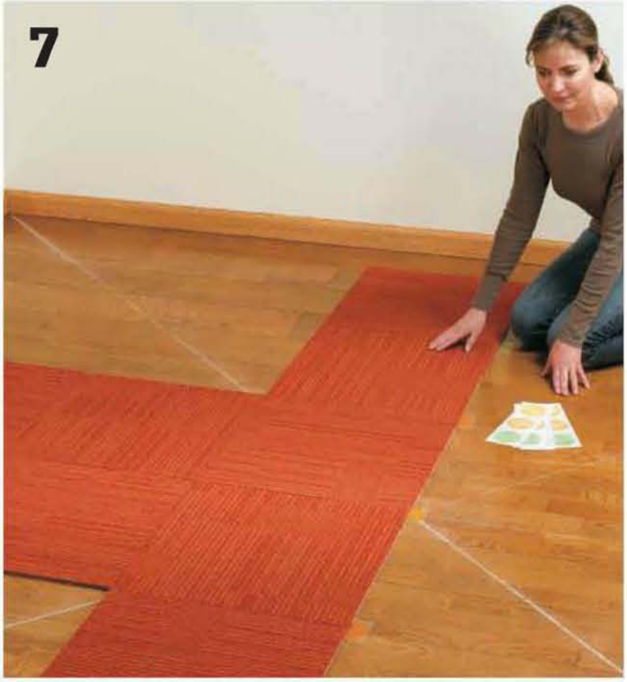

Fasten the base rows in place using the manufacturer's recommended adhesive. This installation calls for two adhesive dots per square. As you place each square, make sure it is aligned with the guidelines and fits tightly against the next square.

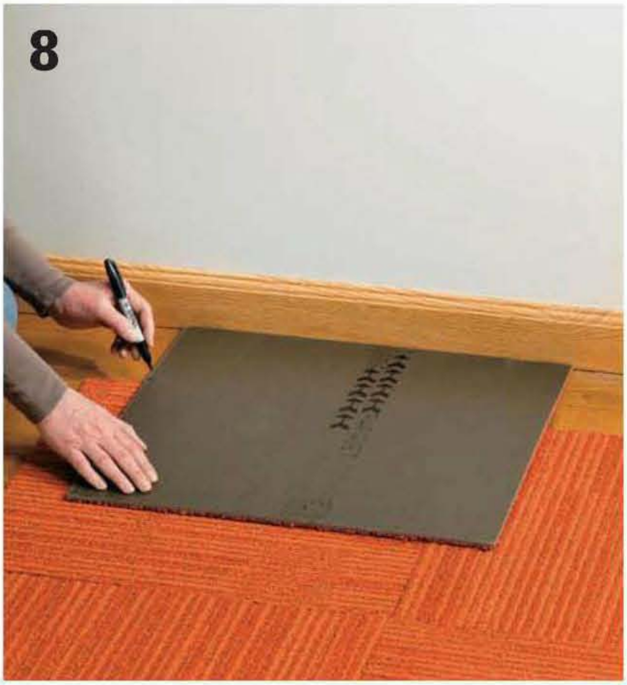

When you reach a wall, flip the last square over. Push it against the wall until it is snug. If you are planning a continuous pattern, align the arrows with the existing squares. If you are creating a parquet pattern, turn the new square 90 degrees before marking it.

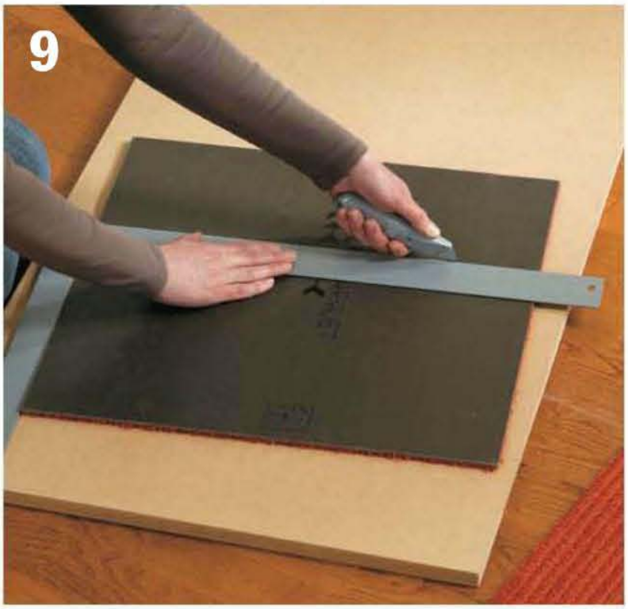

9

Mark notches or draw a line across the back where the new square overlaps the next-to-last one. Using a sharp carpet knife, a carpenter's square, and a tough work surface, cut along this line. The cut square should fit neatly in the remaining space.

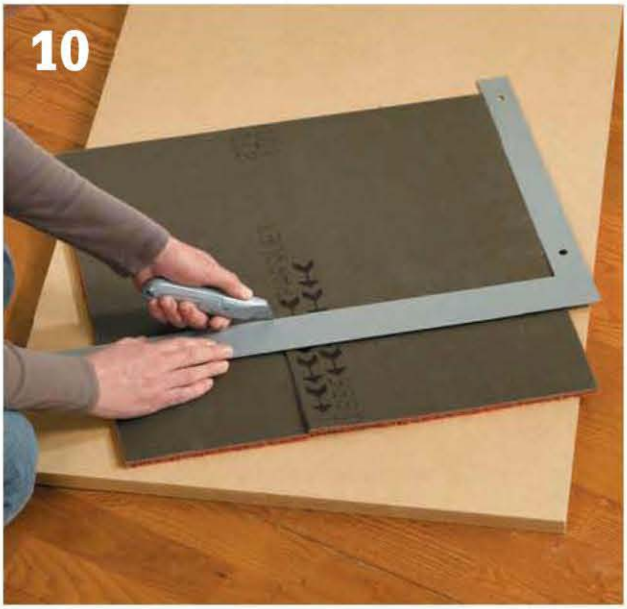

10

At a door jamb, place a square face up where it will go. Lean the square against the jamb and mark the point where they meet. Move the square to find the other cutline, and mark that as well. Flip the square over, mark the two lines using a carpenter's square, and cut out the corner.

11

Finish all four base rows before you fill in the rest of the room. As you work, check the alignment of each row. If you notice a row going out of line, find the point where the direction changed, then remove squares back to that point and start again.

12

Work outward from the center so that you have a known reference for keeping rows straight. Save the cut pieces from the ends. They may be useful for patching odd spaces around doorways, heat registers, radiator pipes, and when you reach the corners.

Refinishing Wood Floors

Refinishing hardwood floors is one of the first major home improvement projects homeowners attempt. The tools you need are readily available for rent, but perhaps the main reason so many DIYers attempt floor refinishing is that it is a very invasive project that requires a lot of planning and moving of furniture. Doing it yourself allows you to work gradually on your own schedule.

Tools & Materials ▸

- Staple gun
- Zip door
- Painter's tape
- Plastic
- Fan
- Pry bar
- 2 × 4 scrap
- Wood shims
- Nail set
- Drum sander
- Sandpaper
- Power edge sander
- Rotary buffer
- Paint scraper
- Sanding block
- Tack cloth

A drum sander (A) and an edge sander (B) are rental tools that can tackle just about any floorsanding project.

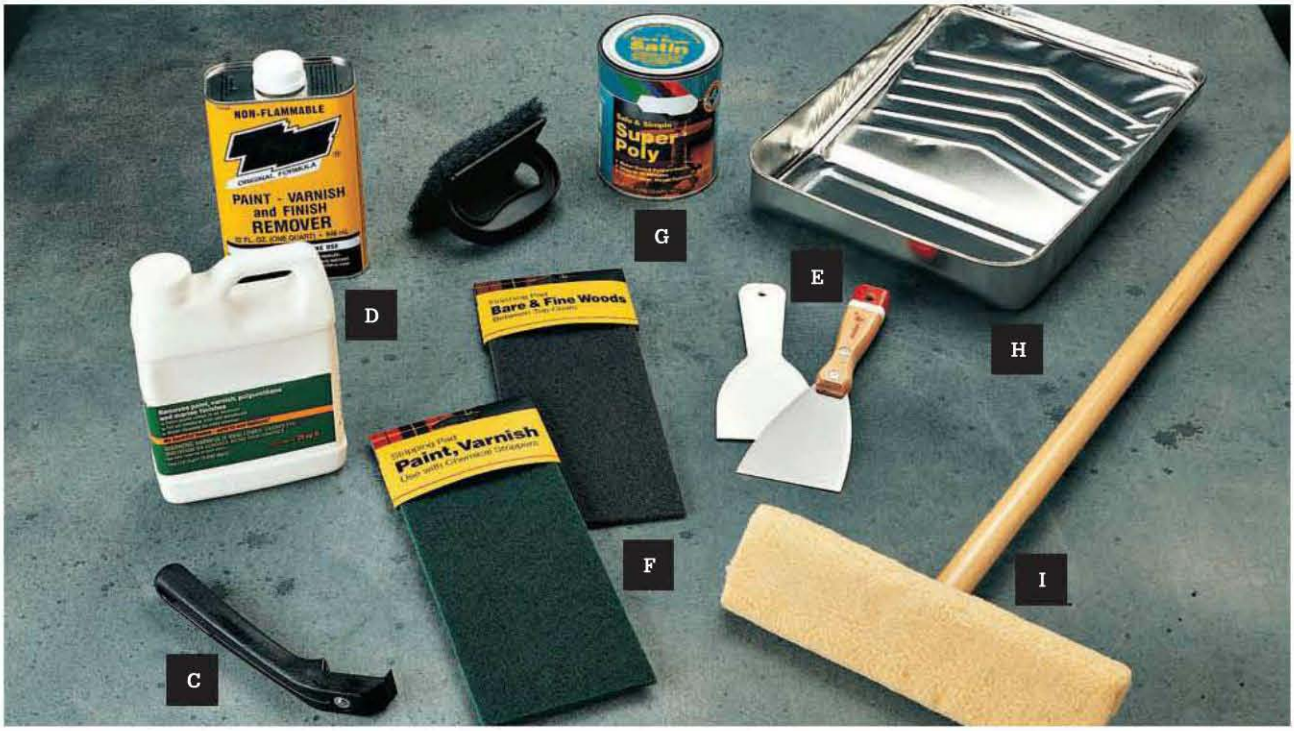

Other products and tools for resurfacing and refinishing floors: Paint scrapers (C) are helpful for removing old finish in corners and other areas that cannot be reached by sanders. When refinishing floors, chemical stripping products (D) are often a more efficient method that yields better results. This is especially true for floors that are uneven, or for parquet and veneered floors, which cannot be sanded. Stripper knives (E) and abrasive pads (F) are used with the stripping products. For the final finish, water-based polyurethane (G) is poured into a paint tray (H) and applied using a wide painting pad with a pole extension (I).

How to Refinish Wood Floors

1

Staple plastic on all doorways. Place a zip door over the entryway you plan to use for the duration of the project. Use painter's tape and plastic to cover heating and cooling registers, ceiling fans, and light fixtures. Finally, place a fan in a nearby window to blow the circulating dust outside.

2

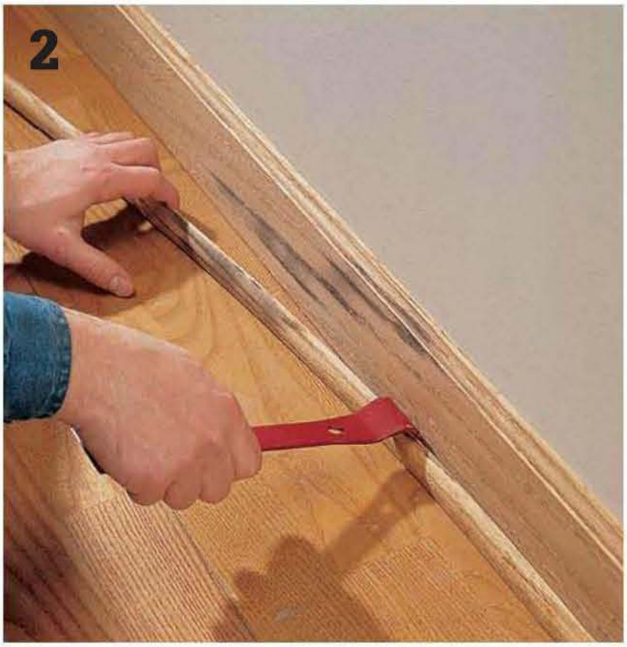

Wedge a prybar between the shoe molding and baseboards. Move along the wall as nails loosen. Once removed, place a scrap 2 × 4 board against the wall and, with a pry bar, pry out baseboard at nails. Maintain the gap with wood shims. Drive protruding nails in floor ⅛" below the surface with a nail set.

3

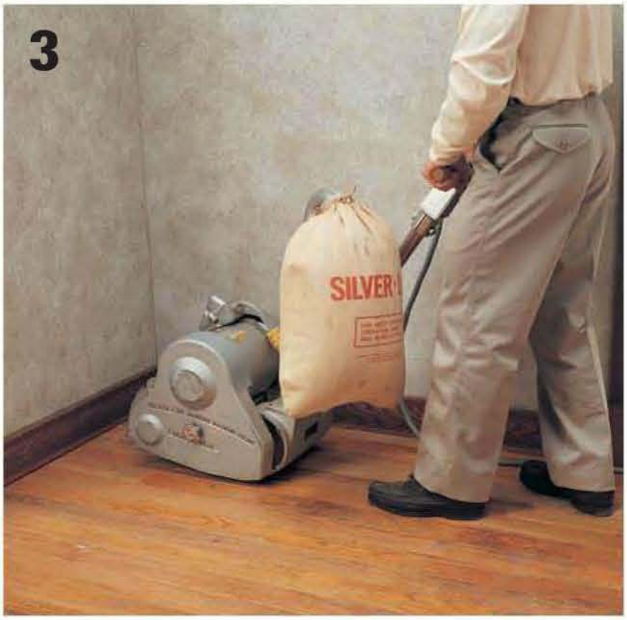

Practice with the drum sander turned off. Move forward and backward; tilt or raise it off the floor a couple of times. A drum sander is difficult to maneuver. Once it touches the floor it walks forward; if you stop it, it gouges the floor.

4

For the initial pass with the drum sander, sand with the grain, using 40- or 60-grit sandpaper. For large scratches use 20 or 30. Start two-thirds down the room length on the right side; work your way to the left. Raise drum. Start motor. Slowly lower drum to the floor. Lift the sander off the floor as you approach the wall. Move to the left 2 to 4" and then walk it backwards the same distance you just walked forward. Repeat.

(continued)

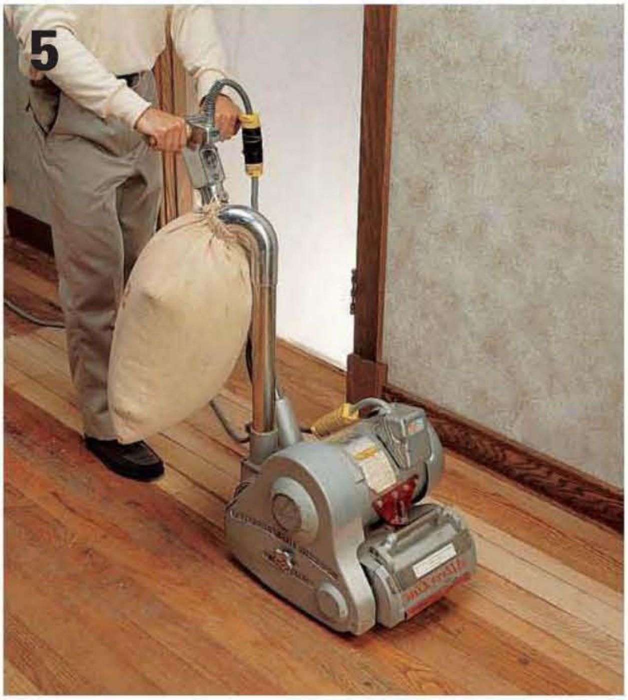

When you get to the far left side of the room, turn the machine around and repeat the process. Overlap the sanded two-thirds to feather out the ridgeline. Repeat this drum sanding process 3 or 4 times using 120-grit paper. For the final pass or two use finer paper (150-grit).

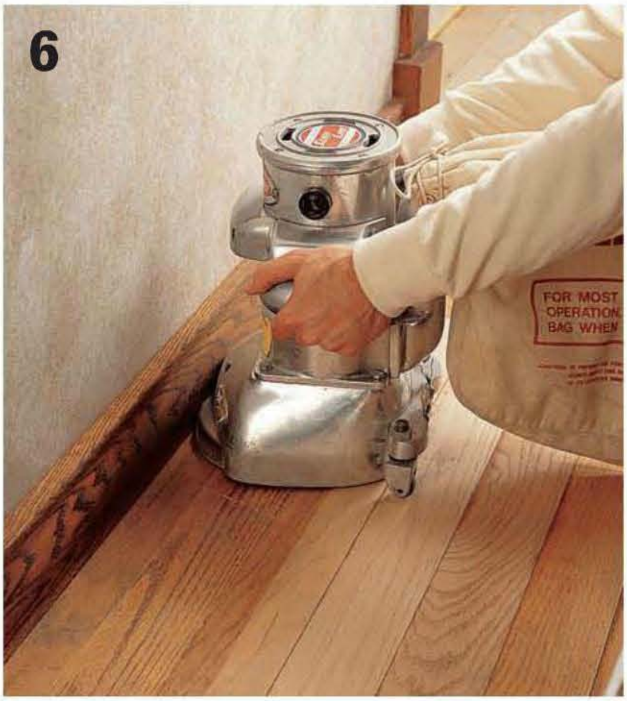

Use a power edge sander along the walls, using the same grit that you last used with the drum sander. Make a succession of overlapping half-circles as you move along the entire perimeter of the room. Next, run a rotary buffer over the floor twice: first with an 80-grit screen and then with a 100-grit screen. Finally, use a random orbital sander to smooth out the floor.

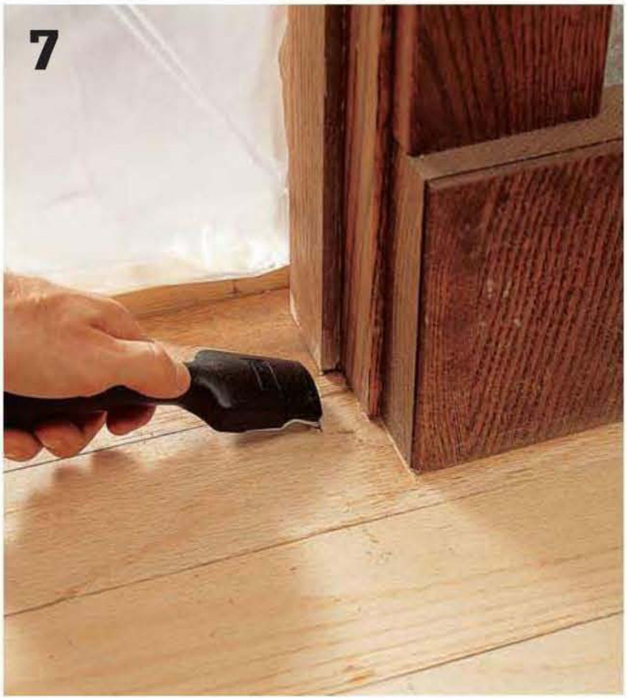

Use a paint scraper to get to corners and hard-to-reach nooks. Pull the scraper toward you with a steady downward pressure. Pull with the grain. Next, sand with a sanding block.

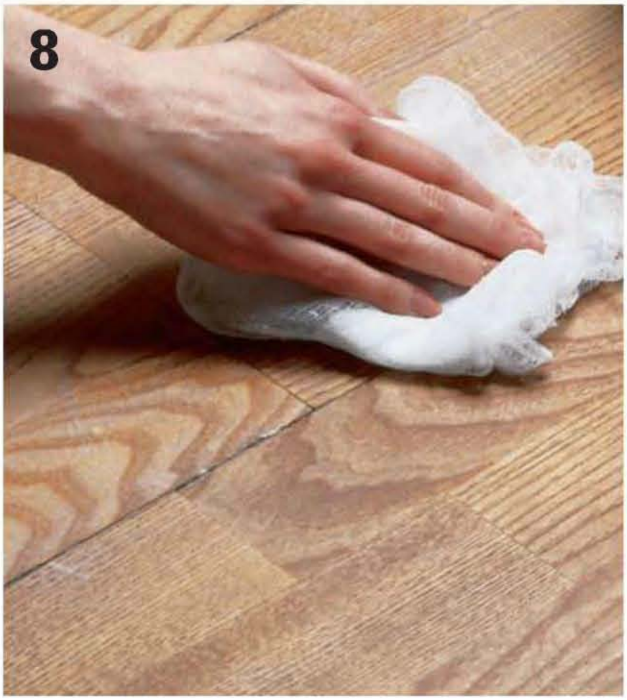

Remove all plastic on doors, windows, and fixtures and then sweep and vacuum to prepare the room for finish. Wipe up fine particles with a tack cloth.

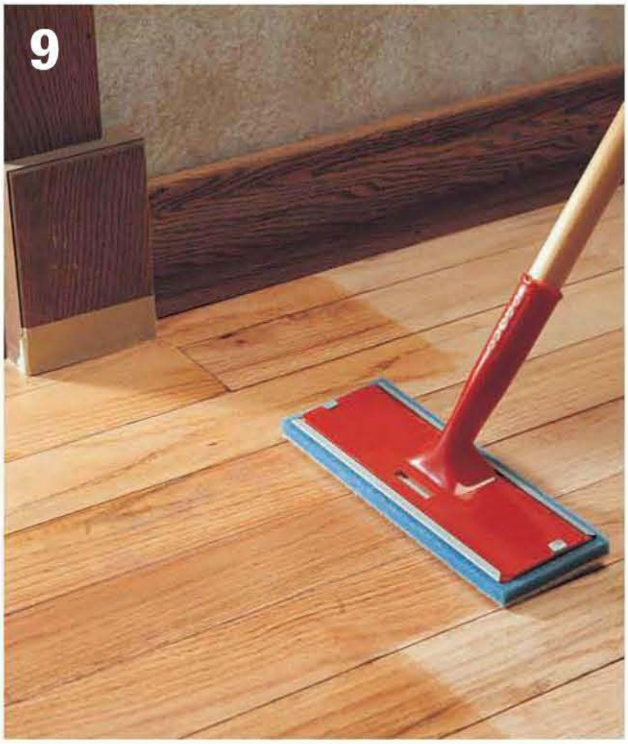

Seal the sanded wood with a 1:1 mixture of water-based polyurethane and water, using a painting pad and pole.

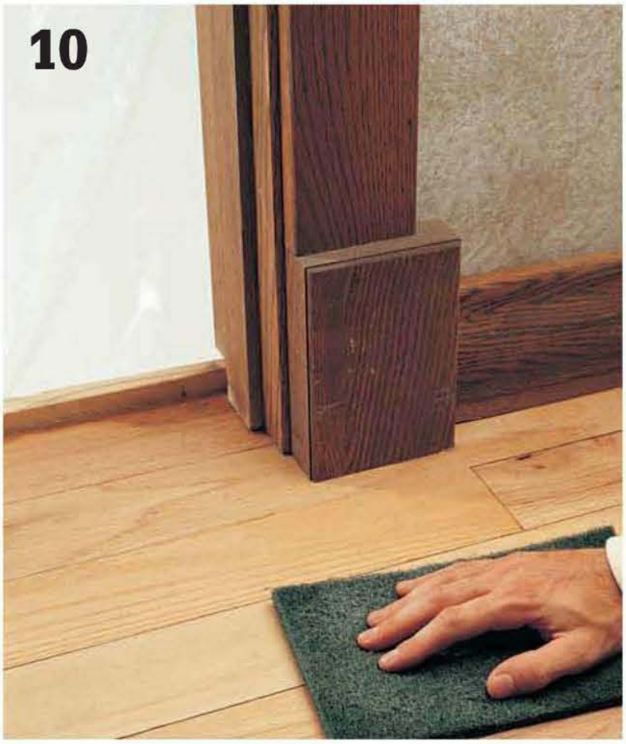

Allow the floor to dry. Buff the surface lightly to remove any raised wood grain, using a medium abrasive pad. Vacuum the surface, using a bristle attachment, then wipe with a tack cloth.

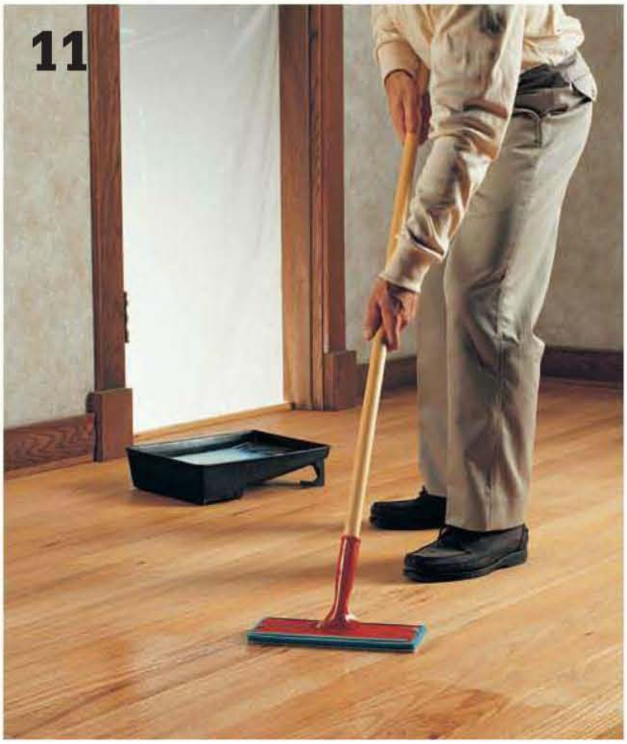

Apply at least two coats of undiluted polyurethane finish to get a hard, durable finish. Allow the finish to dry; repeat step 10 and then add a final coat. Do not overbrush.

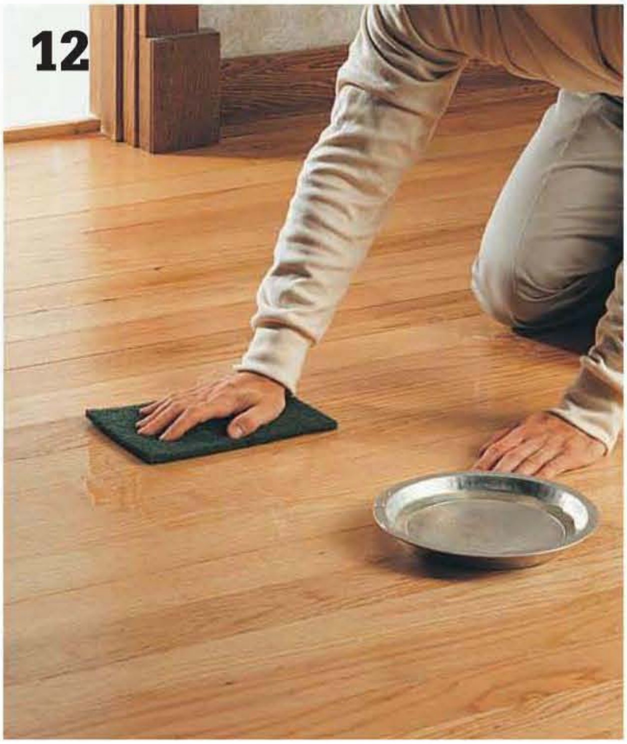

When the final coat is dry, buff the surface with water and fine abrasive pad. Wait at least 72 hours before replacing the shoe molding.

BDBN1202 Type 1 12V DC
Use only with B&D FSB12 or FS120BX battery pack.
5/8" - 2" (16-51mm) x 18 GA. FINISH NAILS
ACCEPTS ONLY STRAIGHT 18 GA FASTENERS
US Pat. No. 6705503, 6604666. Other patents pending.
Questions / Comments?: 1-800-54-HOW-TO
Black & Decker (U.S.) Inc.,
Towson, MD 21286 U.S.A.
Made in China
PRESS TO UNLOCK
SLIDE TO ADJUST
WARNING
AVERTISSEMENT

Walls & Ceilings

Walls and ceilings are much more than structural barriers dividing your house into separate areas. They are canvases awaiting an inspired treatment that can set mood, distinguish space, and reflect your personality within each individual room.

Through creative design and a thoughtful selection of colors, textures, and materials, nearly any wall and ceiling design you dream up can become a reality. From basic painted wallboard and traditional wood paneling to elegant glass block and distinct veneer plaster finishes, there is an ever-increasing variety of wall and ceiling finish materials available to you. Most home centers and lumberyards keep an impressive selection in stock and can easily accommodate special orders.

In this chapter:

- Installing Paneling
- Installing Beadboard Wainscoting
- Building Glass Block Walls
- Applying Veneer Plaster
- Paneling Ceilings
- Installing Suspended Ceilings
- Installing Acoustical Ceiling Tiles
- Installing Metal Ceilings
- Installing Base Molding
- Installing Picture Rail
- Installing Chair Rail
- Installing Crown Molding

Installing Paneling

Paneling is a versatile wall-surfacing material that comes in a wide range of styles, colors, and prices. Paneling sheets are made from a variety of materials for numerous applications:

Solid and veneer wood paneling is durable and easy to clean. Available in finished and unfinished sheets, wood paneling brings a warm, rich tone to any room. It is often used as wainscoting (pages 214 to 219) and also can be used as an inexpensive cover-up for damaged plaster.

Laminate panels are sheets of MDF, particleboard, or plywood faced with paper, print or vinyl. Laminates are available in hundreds of colors, styles, and patterns, providing a durable alternative to paint or wallcoverings.

FRP (fiberglass reinforced plastic), extruded plastic, and vinyl panels contain solid material throughout the panel, creating a low-maintenance, water-resistant wall surface for bathrooms, utility rooms, garages and workshops, as well as numerous commercial applications.

Tileboard is moisture-resistant hardboard coated with melamine, providing a durable, easy-to-clean plastic finish. It's designed to replicate the appearance of ceramic tile, for use in bathrooms, laundry rooms, and kitchens.

Bamboo paneling is gaining in popularity due to its unique look and green-friendly specs. Panels are constructed of strips of bamboo laminated to a fabric backing, which allows it to conform to any type of surface, flat or curved.

Most paneling is available in 4 × 8, 4 × 9, and 4 × 10 sheets. Some manufacturers also offer sheets in sixty-inch widths. Paneling that is ¼" or less in thickness requires a solid backer of at least ½" wallboard; paneling ⅜" thick or more is rigid enough to be fastened directly to framing with sixteen-inch O.C. spacing. Installation typically involves a panel adhesive, either applied in beads along the wall or framing, or troweled onto the back surface of the panel. Make sure to check the manufacturer's instructions for the product you purchase.

Tools & Materials ▸

- Pry bar
- Stud finder
- Tape measure
- Plumb bob
- Circular saw
- Straightedge
- Hammer
- Carpenter's level
- Compass
- Jigsaw
- Caulk gun
- Paneling sheets
- 4d finish nails
- Wood stain
- Panel adhesive
- Powdered chalk

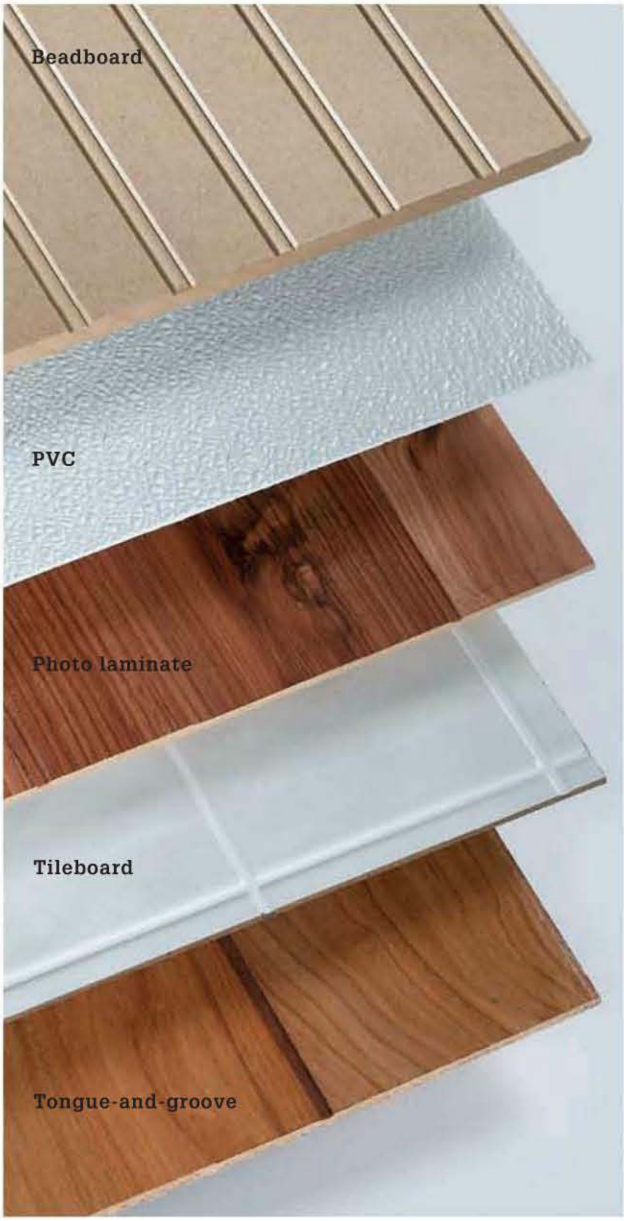

How to Install Wood Paneling

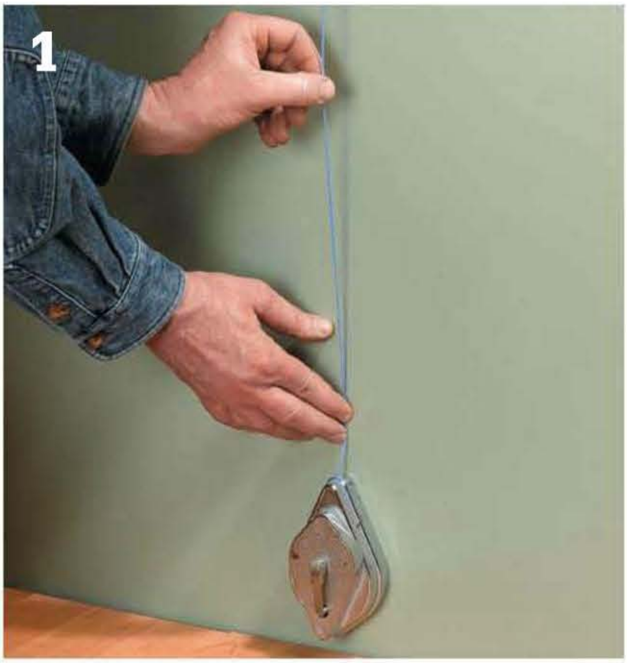

Starting in the corner farthest from the entry, use a stud finder to locate the center of the stud closest to, but less than 48" from, the corner. Find and mark stud centers every 48" from this first stud. Snap a plumb chalk line down the wall at each location. Paneling seams will fall along these lines.

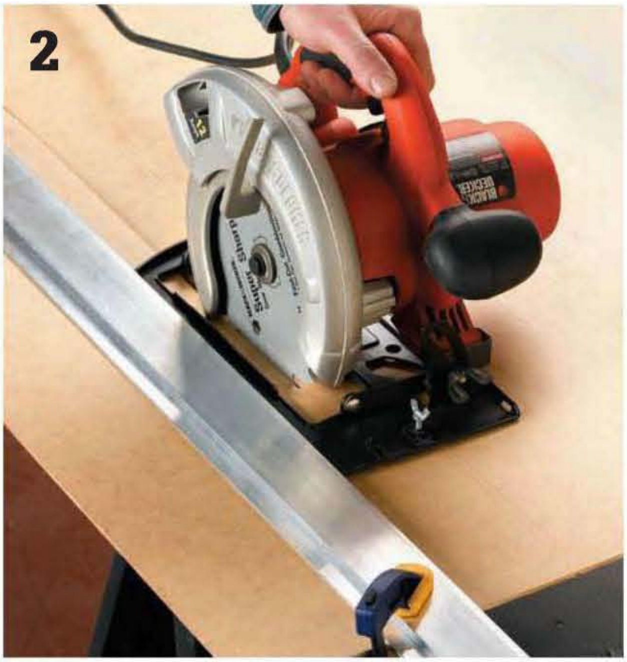

Lay the first paneling sheet face-side down. Measure the distance from corner to the first plumb mark and add 1" to allow for scribing. Use a circular saw and clamped straightedge to cut paneling to this measurement.

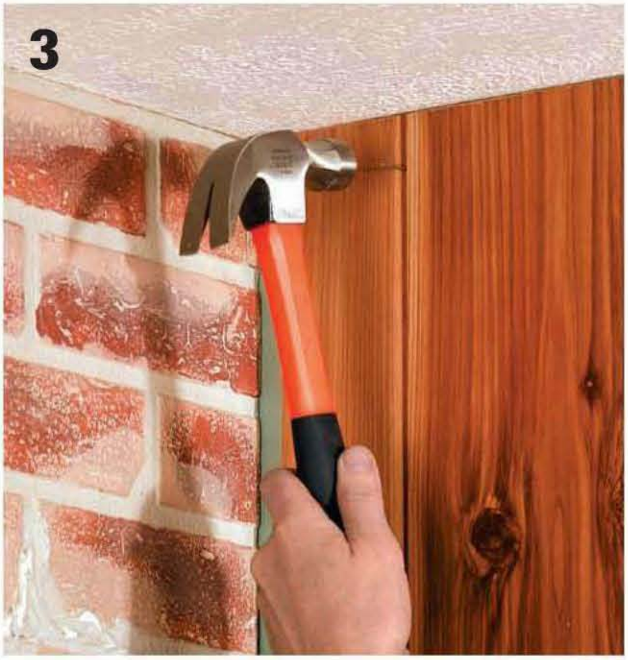

Position the first sheet of paneling against the wall so that the cut edge is 1" away from the corner, and the opposite, finished edge is plumb. Temporarily tack the top of the paneling to the wall.

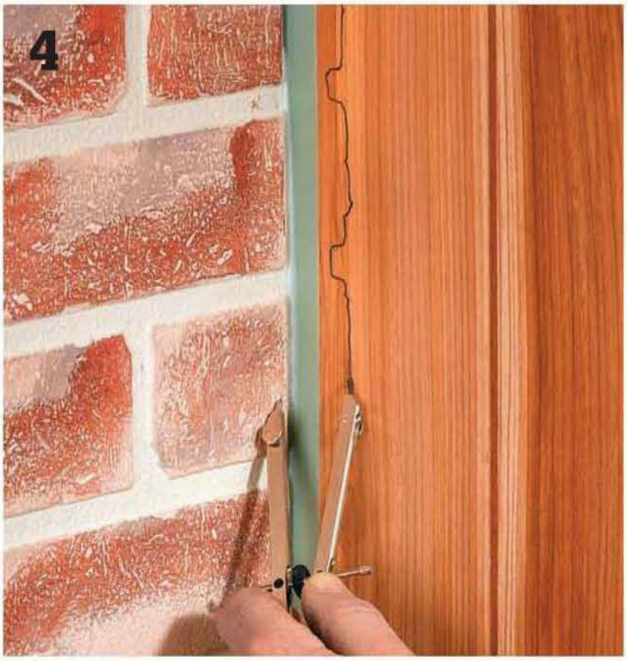

Spread the legs of a compass to 1¼", then run the compass down the full height of the wall to scribe the corner irregularities onto the face of the paneling. Remove paneling from wall.

(continued)

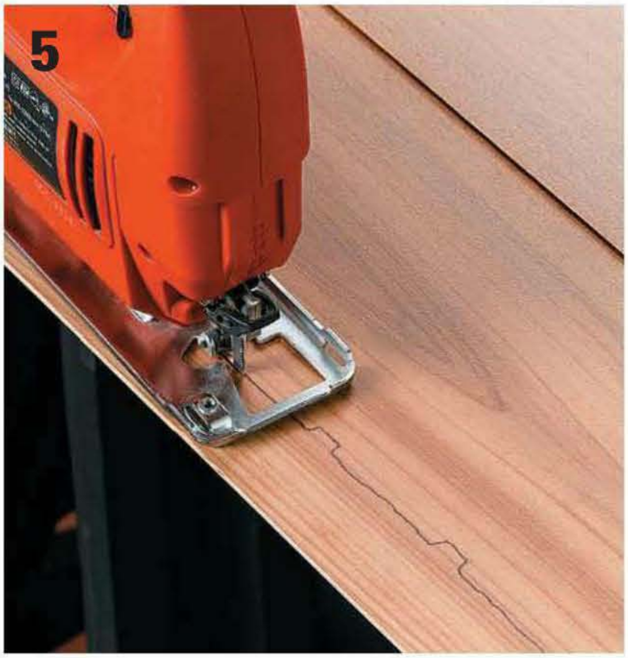

Lay the paneling face-side up and cut along the scribe line with a jigsaw. To prevent splintering, use a fine-tooth woodcutting blade. The scribed edge will fit perfectly against the wall corner.

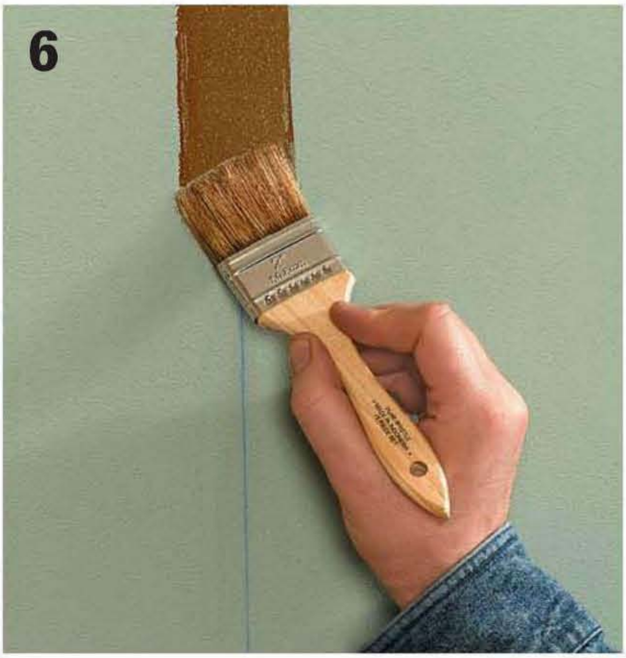

Apply stain or paint to the wall at the plumb lines so the backer will not show through the slight gaps at joints. Select a color that matches the color of the paneling edges, which may be darker than the paneling surface.

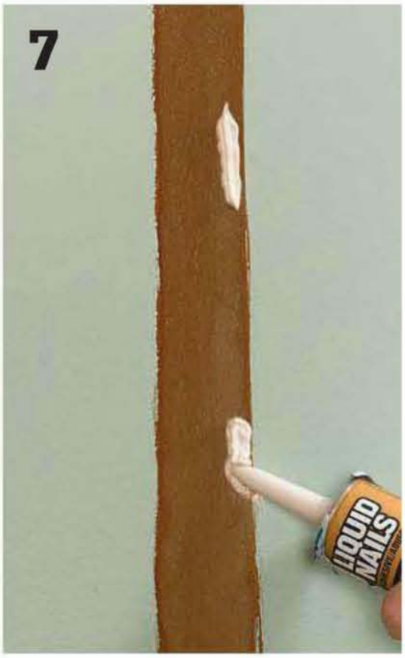

Use a caulk gun to apply 2"-long beads of panel adhesive to the wall at 6" intervals and in a continuous, wavy bead about 1" back from plumb lines (to prevent adhesive from seeping out through the joints). For new construction, apply adhesive directly to the studs.

Attach the paneling to the top of the wall, using 4d finishing nails driven every 16". Press the paneling against the adhesive, then pull it away from the wall. Press the paneling back against the wall when the adhesive is tacky, about 2 minutes.

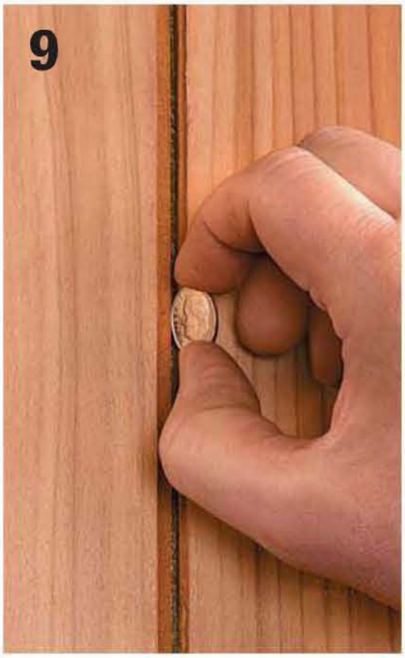

Hang the remaining paneling so that there is a slight space at the joints. This space allows paneling to expand in damp weather. Use a dime as a spacing gauge.

How to Cut Openings in Paneling

For window, door and other openings, measure the opening and mark the outline on the backside of the paneling. Cut to size using a circular saw and straightedge. Install as you would a full sheet of paneling (page 212).

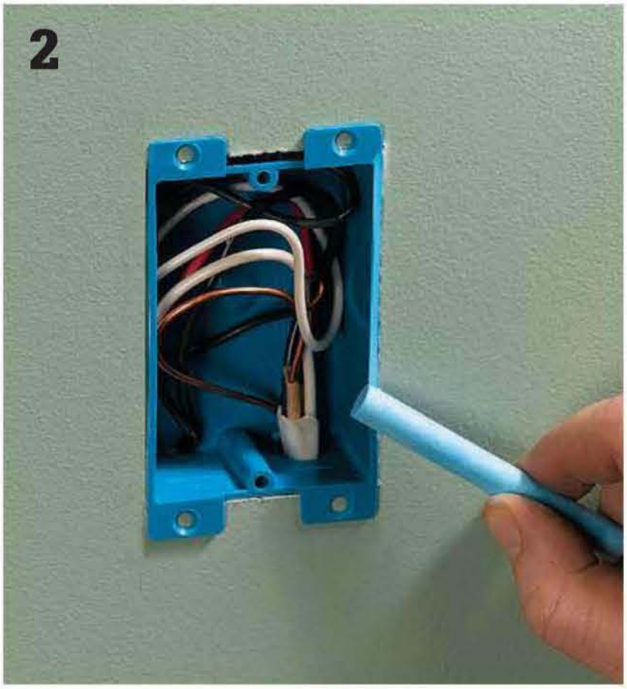

For receptacles, switches, fixtures, and heating vents, coat the edges of electrical boxes and ductwork with chalk.

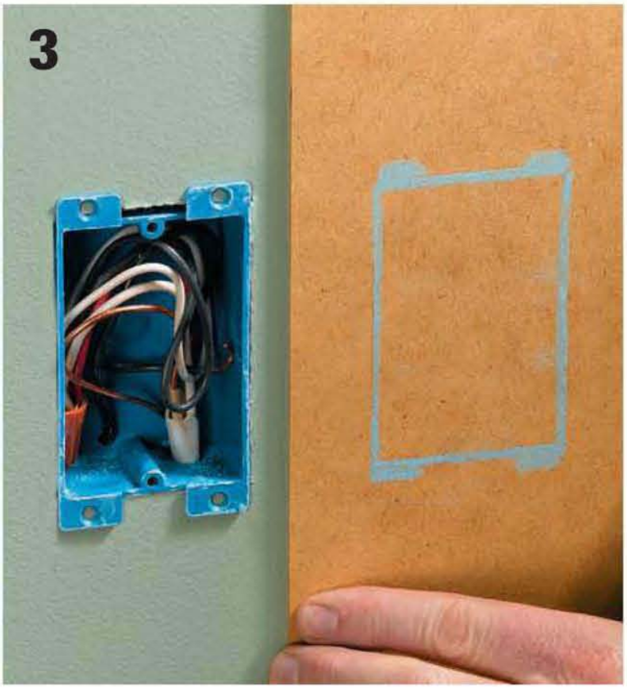

Press the paneling against the wall, so the backside presses against the outlet or vent—the chalk outline will transfer to the paneling.

4
STORAGE CASE

Lay the paneling face-side down. Drill a hole at one corner of each outline, then use a jigsaw with a fine-tooth woodcutting blade to make the cutouts.

Installing Beadboard Wainscoting

Wainscoting refers to virtually any specialized treatment of the lower three to four feet of interior walls. The form demonstrated here, using tongue-and-groove boards, first gained popularity in the early twentieth century. Recently, it has re-emerged as a stylish way to dress up a room.

Typical tongue-and-groove boards for wainscoting are made of pine, fir, or other softwoods and measure ¼" to ¾" thick. Each board has a tongue on one edge, a groove on the other, and usually a decorative bevel or bead on each edge. Boards are cut to length, then attached with nails, most of which are driven through the tongues of the boards. This technique, known as blindnailing, hides the nails from view.

Once installed, the wainscoting is capped at a height of 30" to 36" with a molding called a *cap rail.* The exact height of the wainscoting is a matter of personal preference. When installed to the height of the furniture in the room, wainscoting provides visual symmetry. It also allows the cap rail to double as a chair rail, protecting the lower portion of the walls from damage.

When installed over finished wallboard, wainscoting usually requires that nailers be fastened to the wall studs to provide a reliable backing for nailing. You can skip this step if you know there is consistent blocking between the studs to substitute for this backing. However, this is usually difficult to confirm unless the walls were framed with tongue-and-groove wainscoting in mind.

Wainscoting can be painted or stained. Oil-based stains can be applied before or after installation, since most of the stain will be absorbed into the wood and won't interfere with the tongue-and-groove joints. If you're painting, choose a latex-based paint; it will resist cracking as the joints expand and contract with changes in the weather.

Tools & Materials ▸

- Pencil
- Level
- Circular saw
- Miter saw
- Miter box
- Hammer
- Nail set
- Plane
- Circuit tester
- Pry bar
- Tape measure
- Paintbrush
- Tongue-and-groove boards
- Finish nails
- 1 × 3 furring strips
- 2" 10d finish nails
- Receptacle box extenders, as required
- Paint or stain

Tongue-and-groove wainscoting boards are milled with smooth faces or contoured to add additional texture to your walls. For staining, choose a wood species with a pronounced grain. For painting, poplar is a good choice, since it has few knots and a consistent, closed grain that accepts paint evenly.

How to Prepare for a Wainscoting Project

1
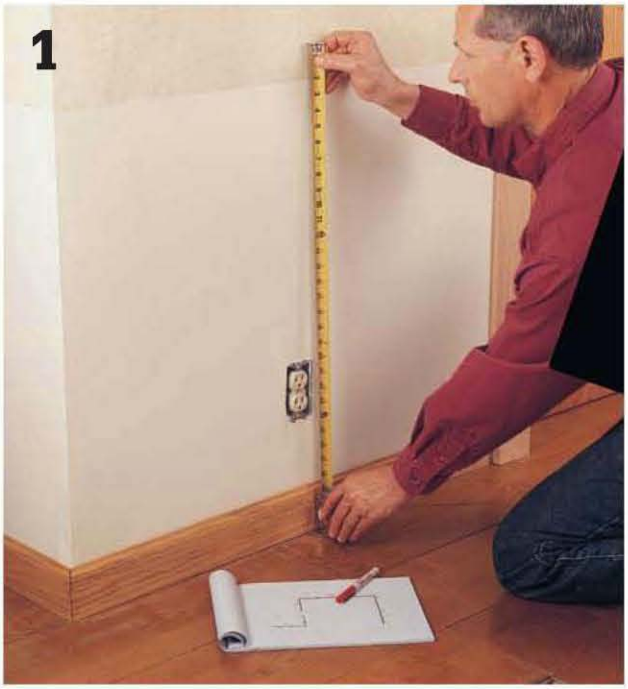

Measure to make a plan drawing of each wall in your project. Indicate the locations of fixtures, receptacles, and windows. Use a level to make sure the corners are plumb. If not, mark plumb lines on the walls to use as reference points.

2
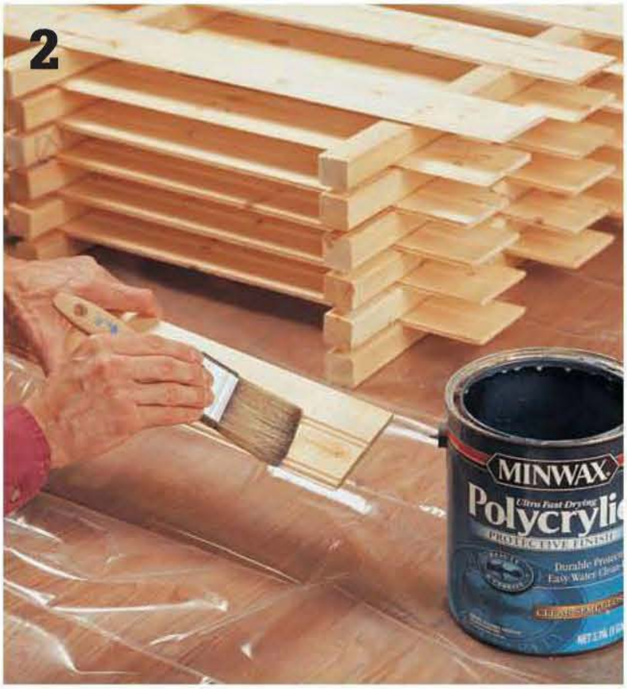

Condition the planking by stacking it in the room where it will be installed. Place spacers between the planks to let air circulate around each board, allowing the wood to adjust to the room's temperature and humidity. Wait 72 hours before staining or sealing the front, back, and edges of each plank.

3
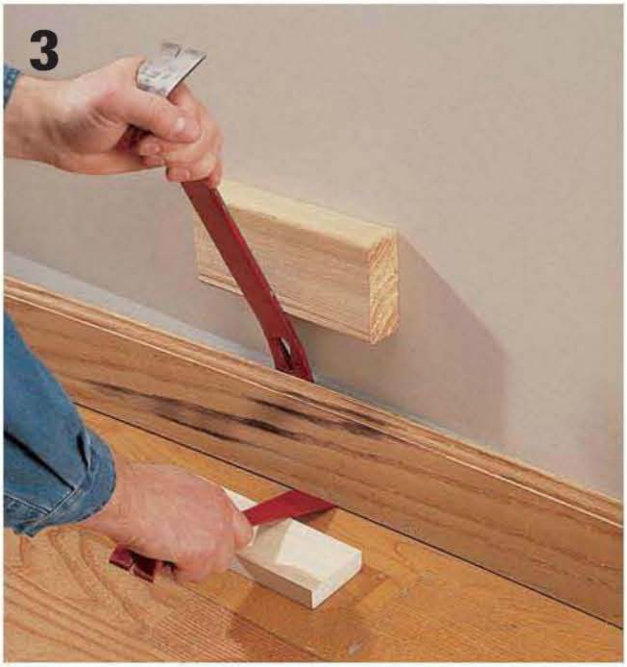

Remove the baseboard moldings, along with any receptacle cover plates, vent covers, or other wall fixtures within the area you plan to cover. Before you begin, turn off the electricity to the circuits in the area.

4
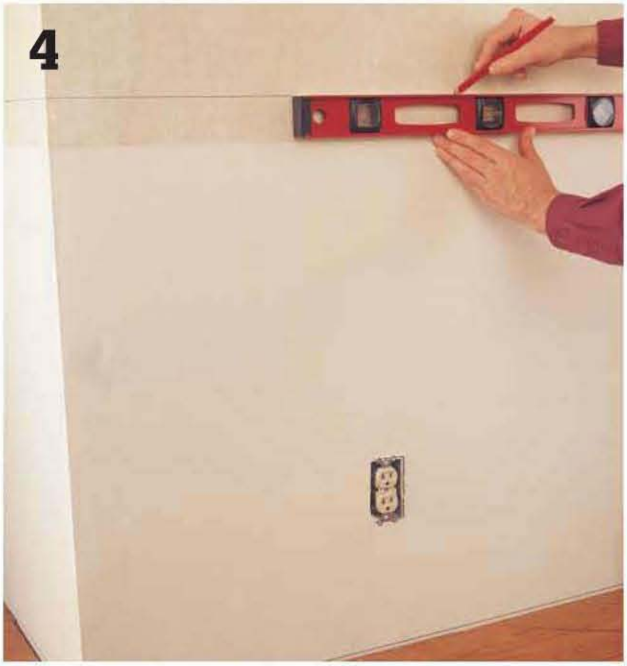

Mark the walls with level lines to indicate the top of the wainscoting. Mark a line ¼" from the floor to provide a small gap for expansion at the floor.

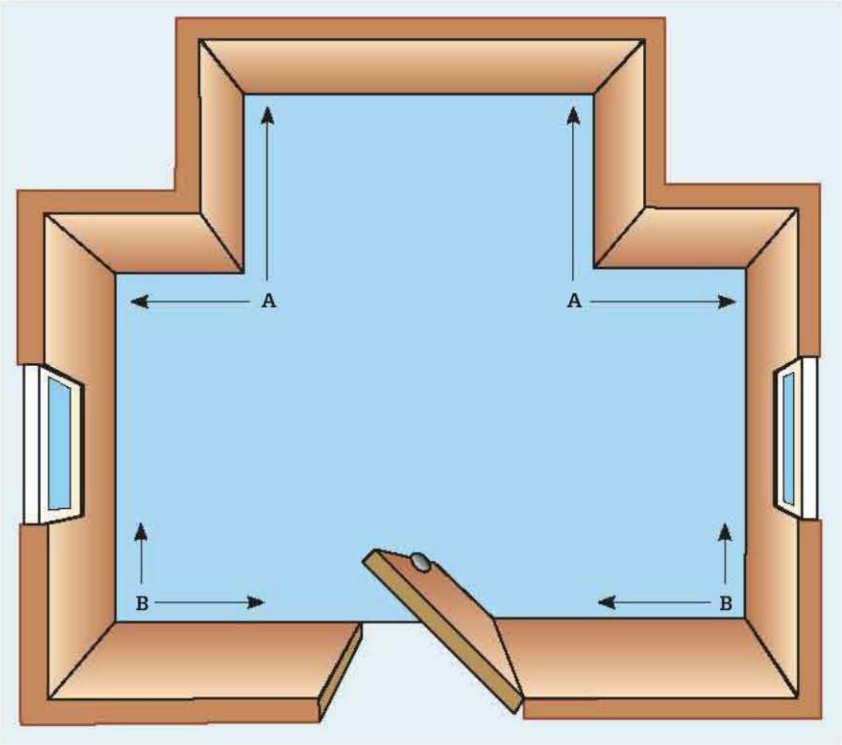

Begin installation at the corners. Install any outside corners (A) first, working your way toward the inside corners. In sections of a room that have no outside corners, start at the inside corners (B), and work your way toward the door and window casings. Calculate the number of boards required for each wall, using the measurements on the drawing you created earlier (length of wall divided by width of one plank). When making this calculation, remember that the tongues are removed from the corner boards. If the total number of boards for a wall includes a fraction of less than ½ of a board, plan to trim the first and last boards to avoid ending with a board cut to less than half its original width.

How to Install Wainscoting at Outside Corners

1

BLACK&DECKER
BLACK&DECKER
Workmate
Portable Project Center & Vise

Cut a pair of boards to the widths indicated in the calculations you developed during the planning process.

2

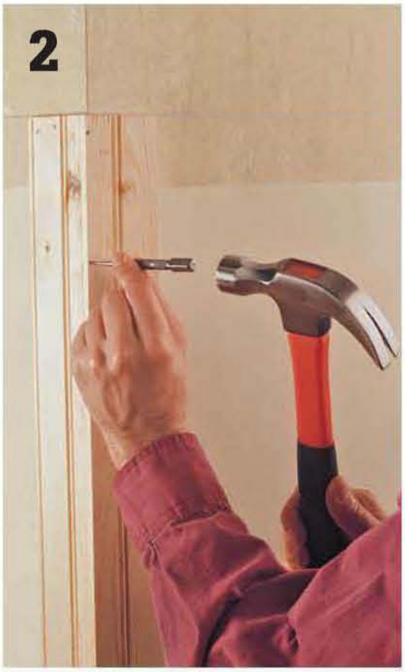

Position the boards at the corner, butting them to create a plumb corner. Facenail the boards in place, then nail the joint, using 6d finish nails. Drive the nails to within ⅛" of the face of the boards, then finish with a nail set.

3

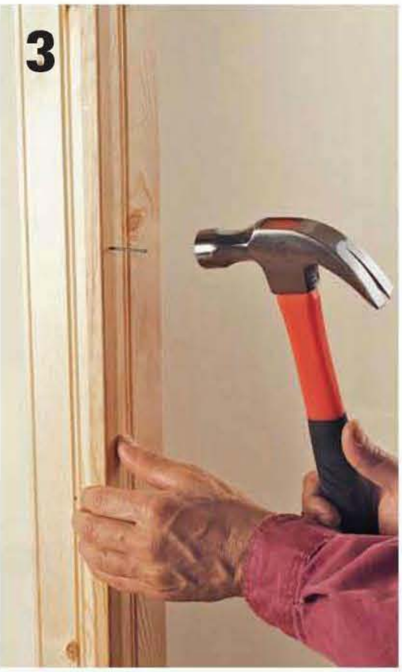

Position a piece of corner trim and nail it in place, using 6d finish nails. Install the remaining boards (opposite, steps 5 and 6).

How to Install Wainscoting at Inside Corners

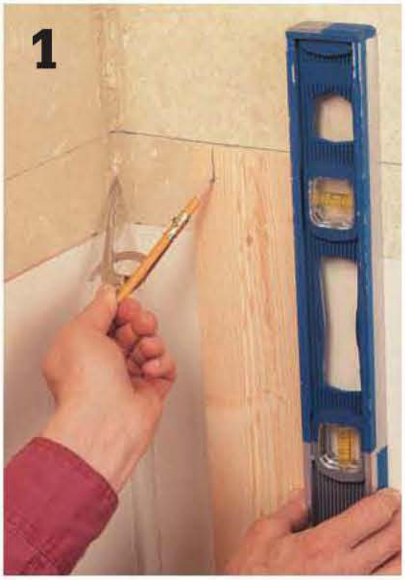

Hold a level against the first board and hold the board flush with the corner. If the wall is out of plumb, trim the board to compensate: Hold the board plumb, position a compass at the inside corner of the wall, and use it to scribe a line down the board.

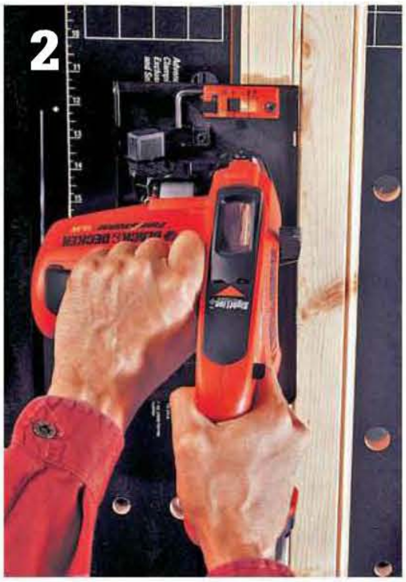

Cut along the scribed line with a circular saw. Subsequent boards may require minor tapering with a plane to adjust for plumb.

Hold the first board in the corner, leaving a ¼" gap for expansion, and facenail into the center of the board at each nailer location, using 6d finish nails. Drive the top nails roughly ½" from the edge so they'll be hidden from view once the cap rail is attached.

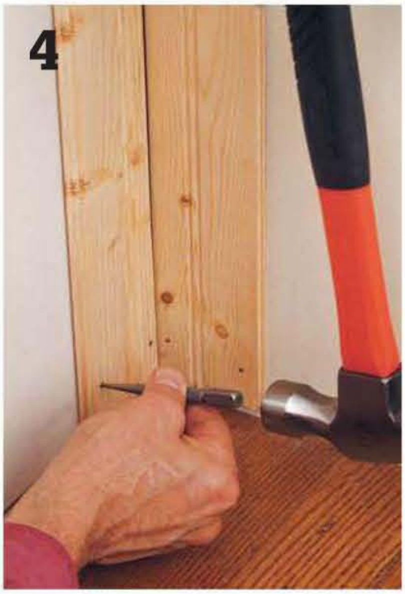

Install a second board at the corner by butting it against the first one, then facenailing in at least two locations. Nail to within ⅛" of the face of the board, then use a nail set to finish.

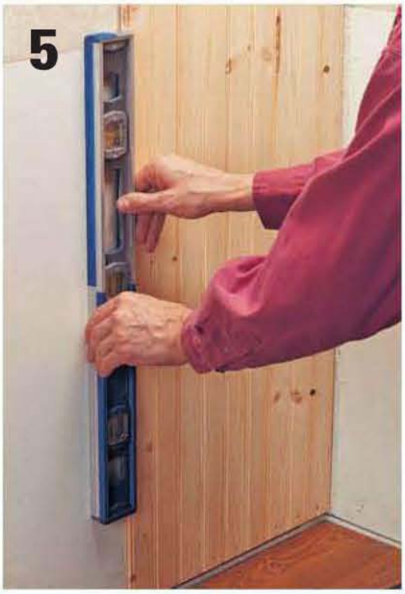

Position subsequent boards. Leave a 1⁄16" gap at each joint to allow for seasonal expansion. Use a level to check every third board for plumb. If the wainscoting is out of plumb, adjust the fourth board, as necessary, to compensate.

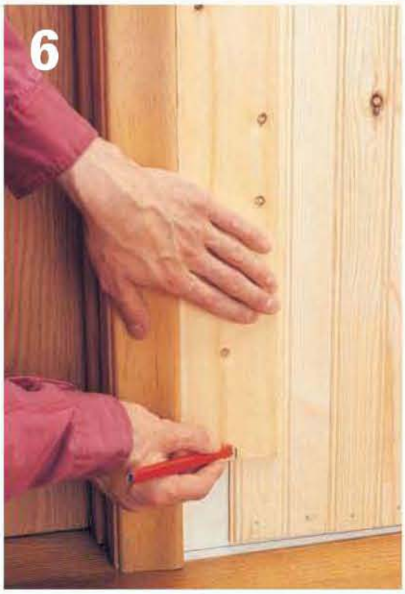

Mark and cut the final board to fit. If you're at a door casing, cut the board to fit flush with the casing (trim off at least the tongue). If you're at an inside corner, make sure it is plumb. If not, scribe and trim the board to fit.

How to Make a Cutout

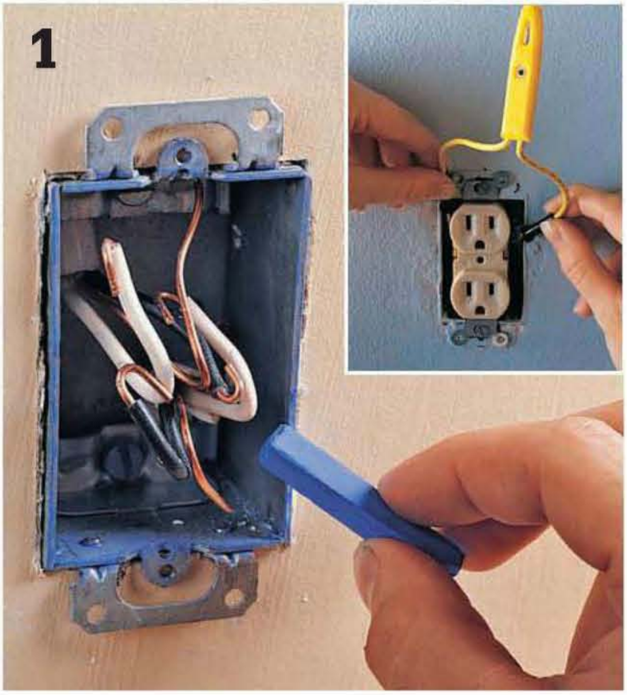

Test the receptacle (inset) to make sure the power is off. Then, unscrew and remove the receptacle from the box. Coat the edges of the electrical box with bright colored chalk.

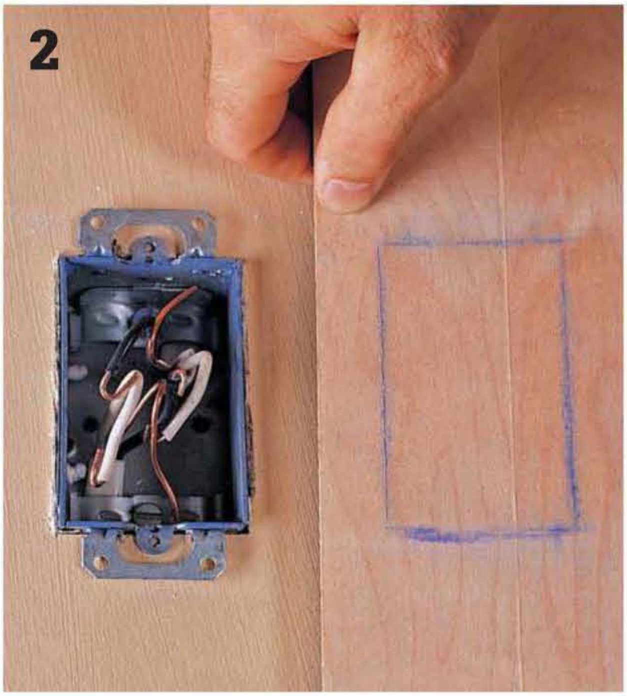

Press the back of the board that will be installed over the receptacle, directly against the electrical box, to create a cutting outline.

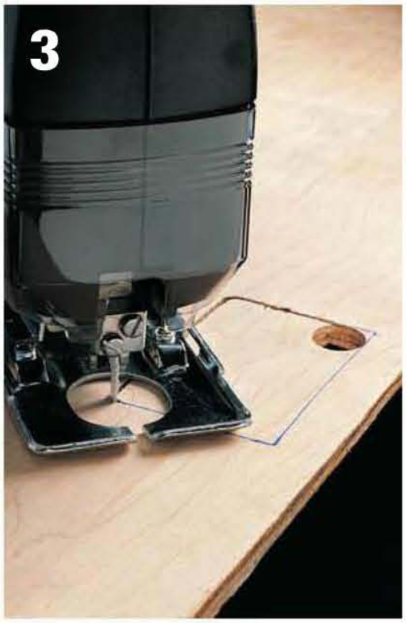

Lay the board face down and drill a large pilot hole near one corner of the outline. Use a jigsaw fitted with a fine-tooth woodcutting blade to make the cutout. Be careful not to cut outside the lines.

Facenail the wainscoting to the wall, then reattach the receptacle with the tabs overlapping the wainscoting so the receptacle is flush with the opening. You may need longer screws.

Tip ▸

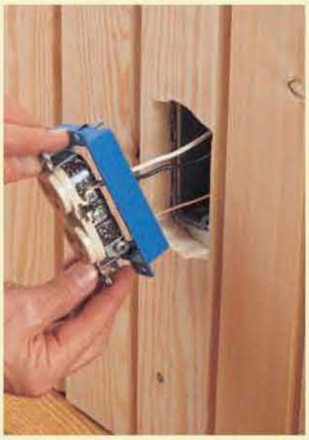

When paneling around a receptacle with thick stock, you will need to attach a receptacle box extender to the inside of the box, then reconnect the receptacle so it is flush with the opening in the paneling.

How to Install Wainscoting Around a Window

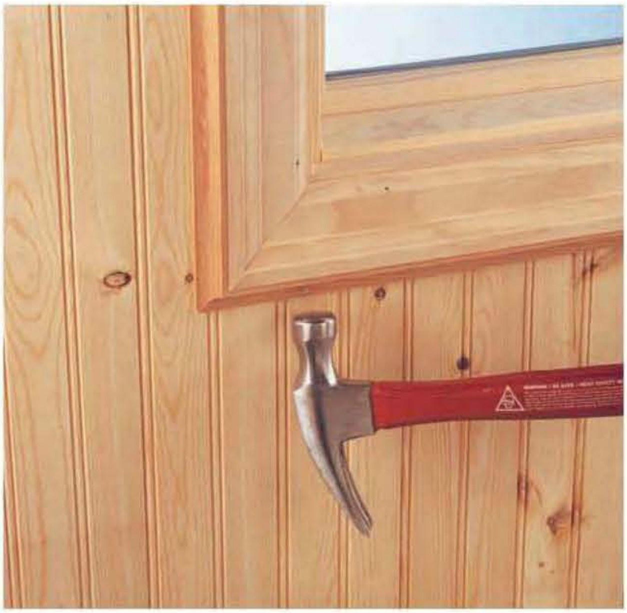

On casement windows, install wainscoting up to the casings on the sides and below the window. Install ½" cove molding, quarter round, or other trim to finish the edges.

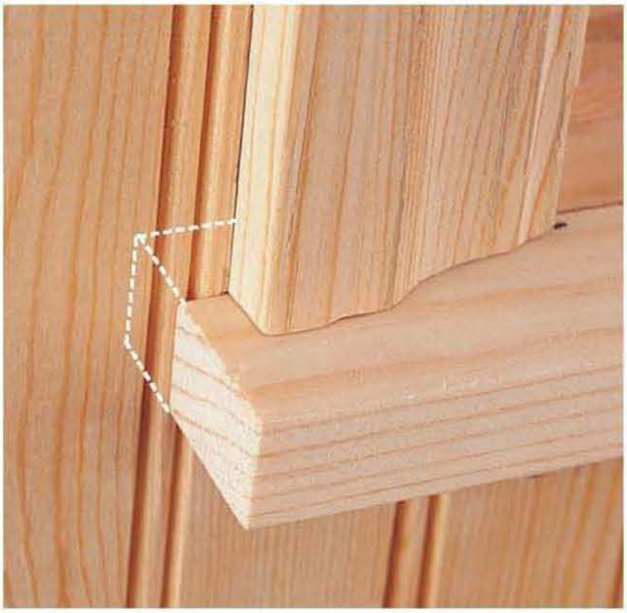

On double-hung windows, remove any window trim and install wainscoting up to the jambs on the sides and below the window. Cut the stool to fit over the wainscoting, then reinstall the apron.

How to Finish a Wainscoting Project

Cut baseboard moldings (pages 242 to 245) to fit over the wainscoting and attach them by nailing 6d finishing nails at the stud locations. If you plan to install a base shoe, leave a small gap at the floor.

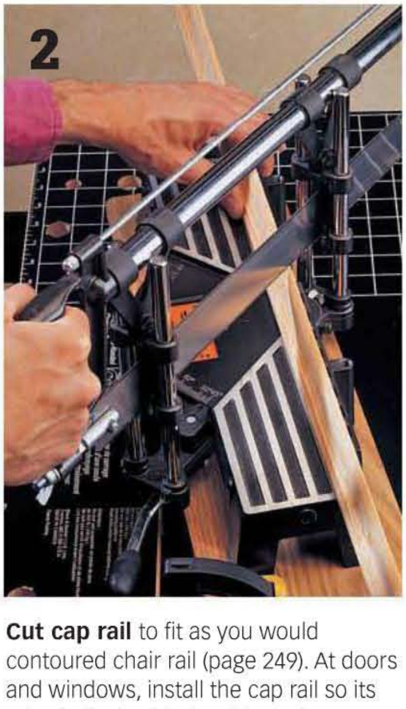

Cut cap rail to fit as you would contoured chair rail (page 249). At doors and windows, install the cap rail so its edge is flush with the side casings.

Attach the cap rail by nailing 4d finish nails through the flats of the moldings at the stud locations so that nails enter both the studs and the wainscoting. Set the nails with a nail set.

Building Glass Block Walls

With its ability to transmit light, a glass block partition wall defines separate living areas while maintaining a sense of openness. You can find glass block at specialty distributors and home centers in a variety of patterns, shapes, and sizes, along with all the products needed for the installation.

You can build your wall to any height. Top a low wall with a course of bullnose blocks to give it a finished rounded edge, or with flat block to create a shelf. To build a full-height wall, calculate the number of courses of block you'll have, then frame-in a header to fill the remaining space between the finished block and the ceiling.

Because of its weight, a glass block wall requires a sturdy foundation. A four-inch-thick concrete basement floor should be strong enough, but a wood floor may need to be reinforced. Contact the local building department for requirements in your area. Also bear in mind that glass block products and installation techniques vary by manufacturer—ask a glass block retailer or manufacturer for advice about the best products and methods for your project.

Tools & Materials ▸

Chalk line
Circular saw
Jigsaw
Paintbrush
Drill
Mixing box
Trowel
Level
Pliers
Jointing tool
Nylon- or natural-bristle brush
Sponge
2 × 6 lumber
16d common nails
Water-based asphalt emulsion
Panel anchors
2½" drywall screws
Foam expansion strips
Glass block mortar
8" glass blocks
¼" T-spacers
Board
Reinforcement wire
16-gauge wire
Caulk or wall trim
Baseboard

Glass block has many uses in modern up-scale homes and today's installation systems allow homeowners to easily work with it.

How to Build a Glass Block Wall

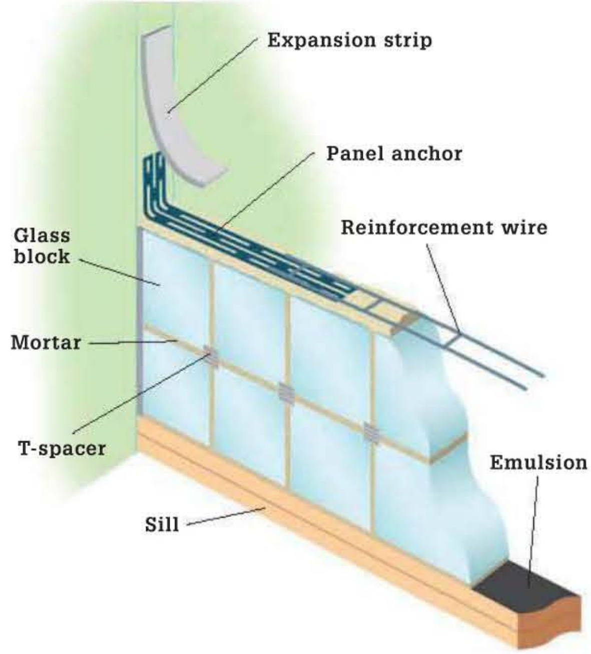

The block wall in this project has a sill made of two 2 × 6s cut to the exact width of the block. This provides a stable base to help resist floor movement and protect the lower courses. The block wall is secured to an anchor stud in an adjoining wall, by means of metal panel anchors. Expansion strips between the two walls allow for movement.

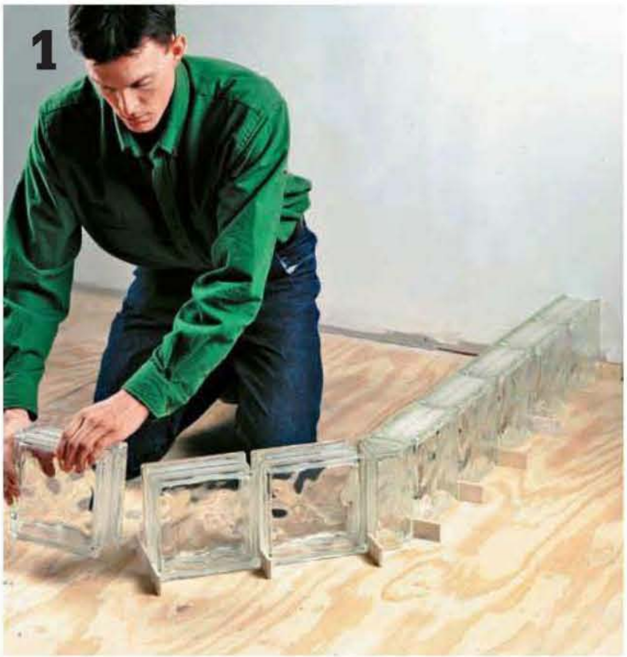

Dry-lay the first course of glass block, using a ⅜" wood spacer between the wall and the first block, and ¼" spacers between the remaining blocks, to set the gaps for the mortar joints. Mark the wall position onto the floor, then remove the blocks. Snap chalk lines along the marks to create the sill outline.

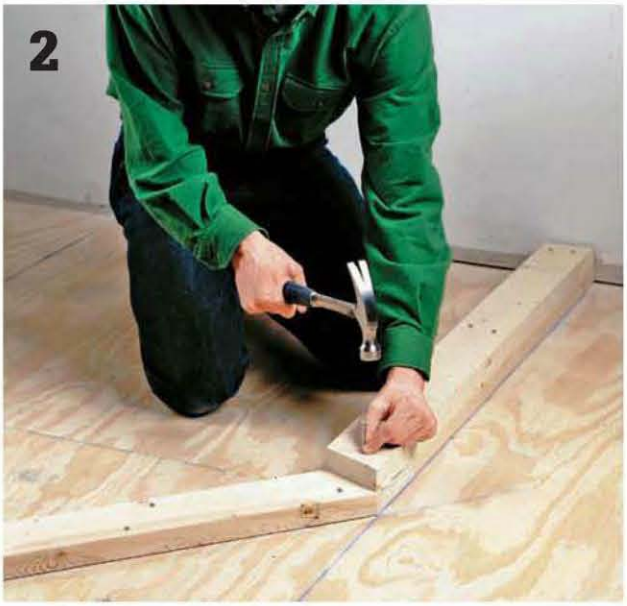

Determine the sill thickness based on the size of your baseboard and thickness of the floorcovering. Rip 2 × 6 lumber to the width of the block. If the end blocks are shaped, trim the sill pieces to match, using a jigsaw. Fasten the sill to the subfloor and framing below with 16d common nails. Apply asphalt emulsion to the sill, using a paintbrush.

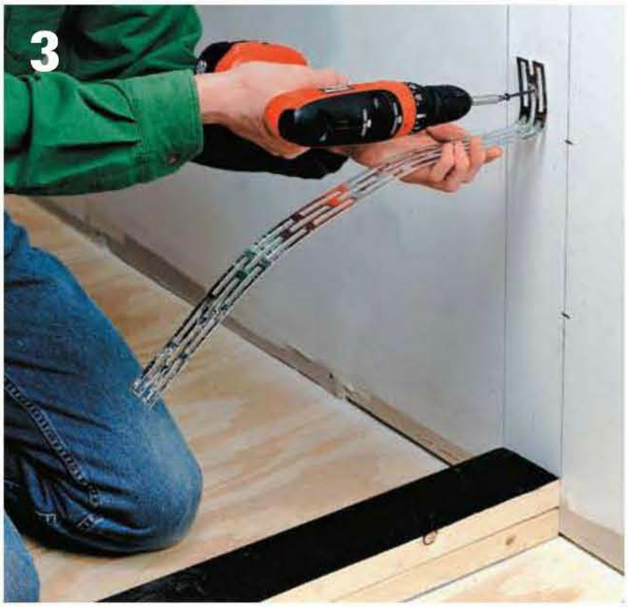

Mark plumb lines on the adjoining wall, straight up from sides of the sill. Mark the finished height of each course along the lines. Fasten a panel anchor to the anchor stud at the top of every second course, using 2½" drywall screws. Cut expansion strips to size and adhere them to the wall between the anchors.

(continued)

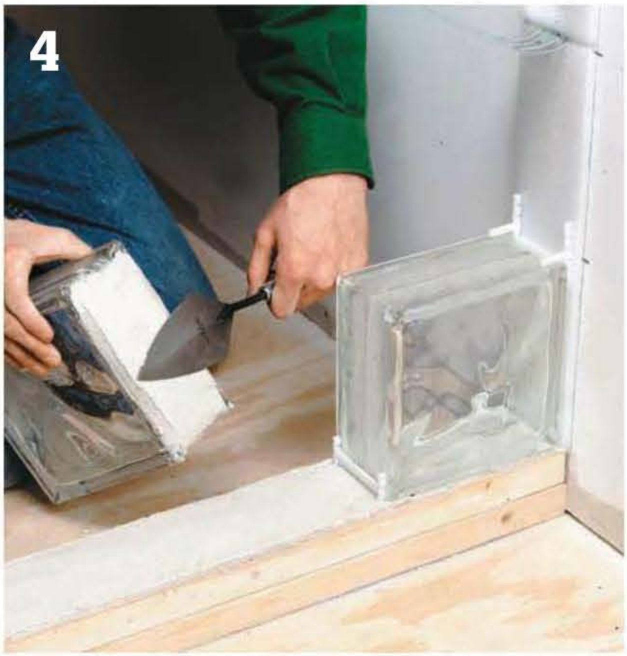

Mix only as much mortar as you can apply in about 30 minutes. Lay a ⅜"-thick mortar bed on the sill, enough for three or four blocks. Set the first block, using ¼" T-spacers at the mortar joint locations (follow the manufacturer's directions for modifying T-spacers at the bottom and sides of the wall). Do not place mortar between blocks and expansion strips. Butter the trailing edge of each subsequent block with enough mortar to fill the sides of both blocks.

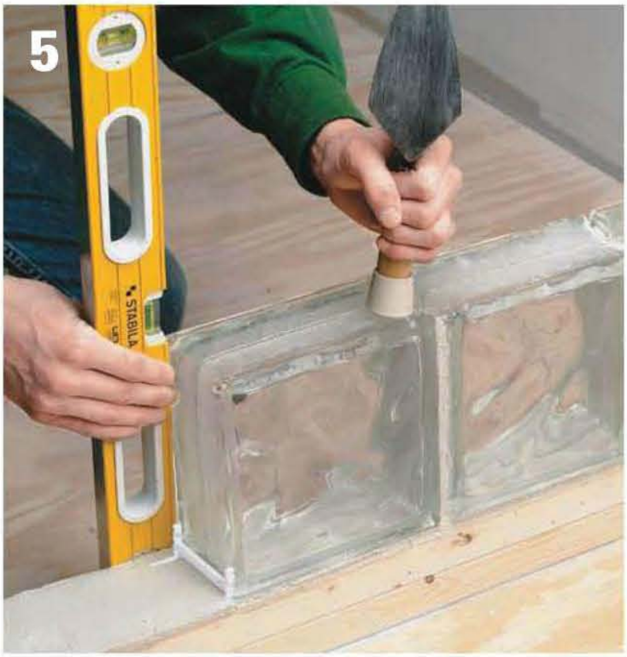

Lay the remainder of the course. If the wall has a corner, work from both ends toward the center, and install the corner piece last. Use ¼" T-spacers between blocks to maintain proper spacing. Plumb and level each block as you work, then check the entire course, using a flat board and a level. Tap blocks into place using a rubber mallet—do not strike them with a metal tool.

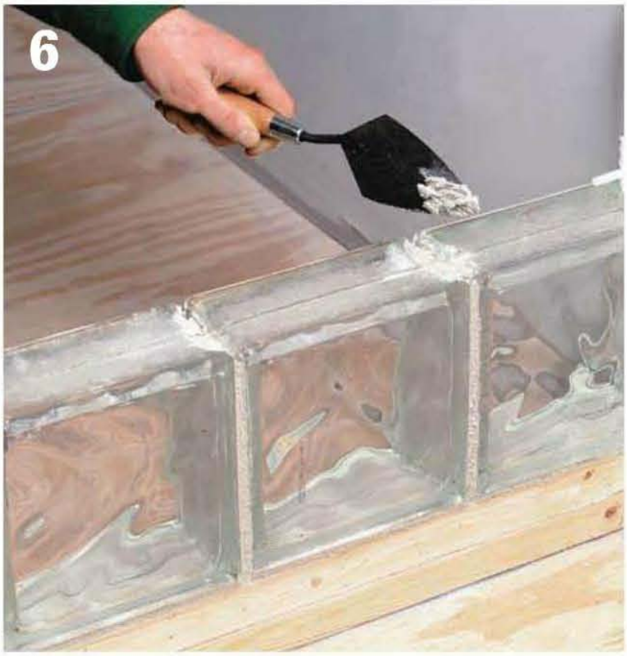

At the top of the course, fill the joints with mortar, and then lay a ¼" bed of mortar for the second course. Lay the block for the second course, checking each block for level and plumb as you work.

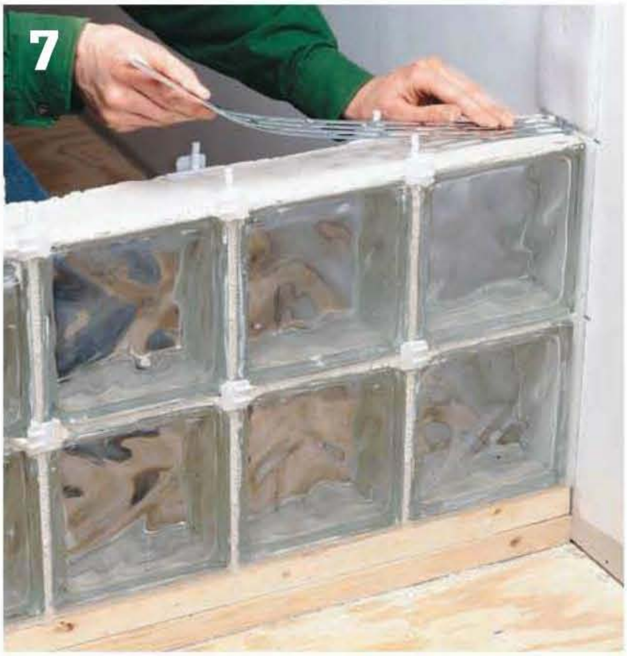

Apply a ⅛" bed of mortar over the second course, then press the panel anchor into the mortar. Repeat this process at each anchor location.

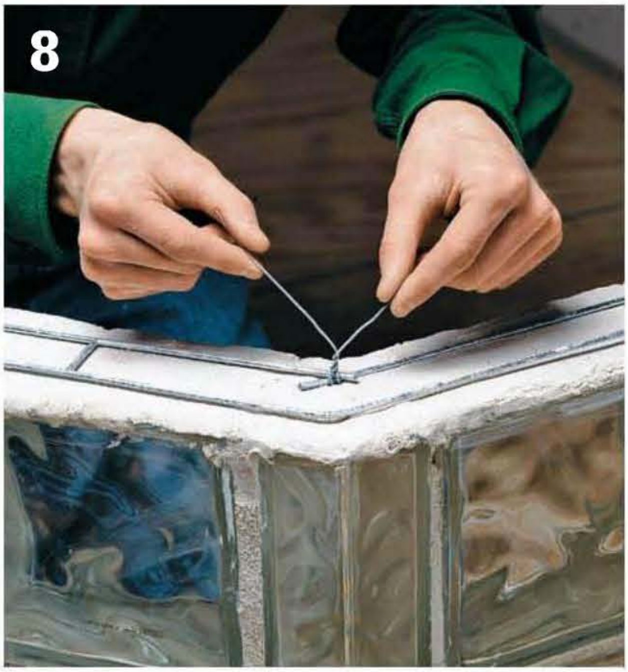

Add reinforcement wire in the same joints as the panel anchors, overlapping the anchors by 6". Also overlap the wire by 6" where multiple pieces are needed. At corners, cut the inner rail of the wire, bend the outer rail to follow the corner, then tie the inner rail ends together with 16-gauge wire. Add another ⅛" mortar bed, then lay the next course of block.

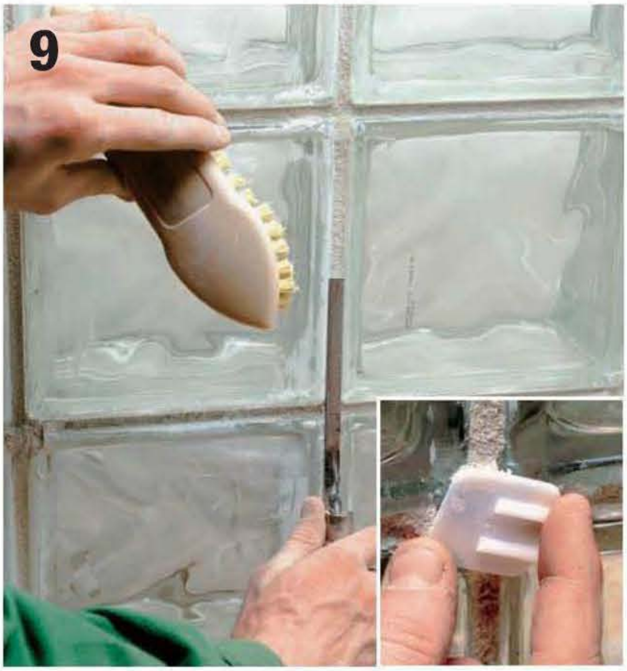

Build the wall in complete courses, checking the mortar after each course. When it is hard enough to resist light finger pressure (usually within 30 minutes), twist off the T-spacer tabs (inset) and pack mortar in the voids. Then, tool all of the joints with a jointing tool. Remove excess mortar from the glass, using a brush or damp sponge.

Clean the glass block thoroughly, using a wet sponge and rinsing it often. Allow the surface to dry, then remove cloudy residue with a clean, dry cloth. After the mortar has cured for two weeks, apply a sealant. Caulk the seam between the glass block and the adjoining wall, or cover the gap with trim.

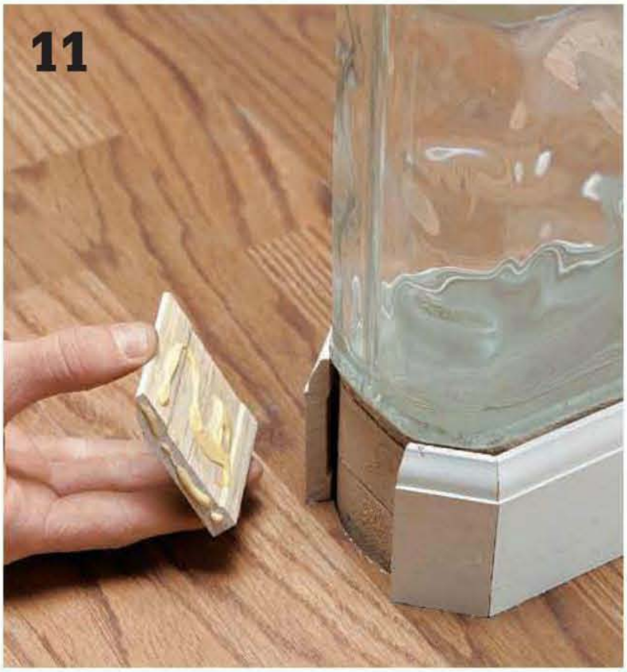

Reinstall the flooring, if necessary, then cut baseboard to fit around the sill (see pages 242 to 245). If the end of your wall has curved (bullnose) block, wrap the end with three pieces of trim.

Applying Veneer Plaster

While gypsum wallboard all but wiped out traditional plaster and lath in the 1940s, a new generation of plaster products now make plaster easier and cheaper to apply, leading to renewed popularity for this classic material.

Veneer plaster systems provide a solid, uniform wall surface that is highly resistant to nail pops, cracks and surface damage. A skim coat of plaster is troweled onto a gypsum wallboard base that has a distinctive blue color, commonly called blueboard. While blueboard is installed like standard wallboard, it has a highly absorptive face paper to which the wet-mix plaster bonds. Blueboard joints do not need to be taped as precisely as standard wallboard joints, and seams and fastener heads do not show through the finished plaster surface, a common problem with standard wallboard.

Veneer systems are available in one-coat and two-coat systems. One-coat systems have a single layer of finish plaster applied directly to the blueboard base; two-coat systems, a rough basecoat for the finish plaster to mechanically "key" or bond to, providing a more rigid surface. Finish plaster can be troweled smooth or tooled for a texture. Sand and other additives can be used to create coarser textures.

Applying veneer plaster effectively does take some time to master, but no more so than that of any masonry technique that requires troweling. The key is to apply the plaster in quick, short strokes, called "scratching in," and then to immediately trowel it over with a steady, even stroke to smooth the plaster to a consistent thickness, typically 1⁄16" to ⅛".

Tools & Materials ▸

Stapler
Hammer
Heavy-duty ½" drill with mixing paddle
16-gal. drum
Mortar hawk
12" trowel
Fine-wire rake or broom (for basecoat)
Spray water bottle
Metal corner bead with mesh flanges
1¼" wallboard screws
Non-adhesive fiberglass mesh tape
¼" staples
Clean potable water
Dry-mix veneer basecoat plaster (for two-coat application)
Dry-mix veneer finish plaster

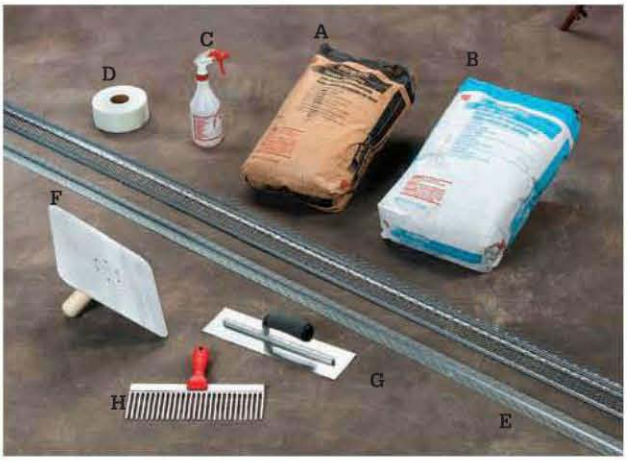

Tools and materials for installing veneer plaster include: dry-mix veneer plaster basecoats (A); finish plaster (B), available for smooth or textured applications; spray bottle for moistening surfaces (C); non-adhesive fiberglass mesh tape for covering blueboard panel seams and inside corners (D); outside corner beads with metal beads and mesh flanges (E); mortar hawk (F); 12" trowel (G); thin-wire rake for roughening the base coat (H).

ONE- AND TWO-COAT VENEER PLASTER SYSTEMS

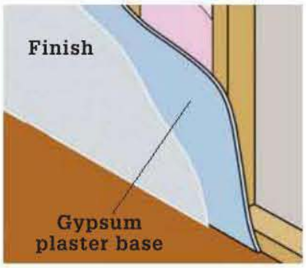

One-coat veneer plaster systems use a single, 1⁄16"- to 3⁄32"-thick coat of finish plaster applied directly to a blueboard base. The coat can be troweled smooth or textured, resulting in a hard, monolithic surface.

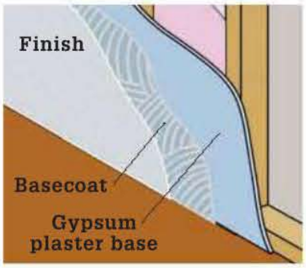

Two-coat veneer plaster systems are composed of a 1⁄16" to ⅛" basecoat plaster applied to blueboard, followed by a 1⁄16"- to 3⁄32"-thick coat of finish plaster. The finish coat bonds with the scratched basecoat surface, forming a more uniform and monolithic surface than that of a one-coat system.

How to Apply a One-Coat Veneer Plaster System

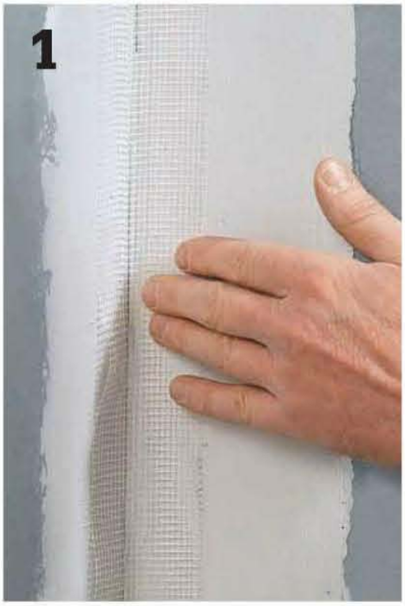

Cover all seams first. Apply a thin layer of plaster along all flat seams and corner bead, feathering out the edges by 6". For inside corners, apply a thin bed of plaster and embed the loose tape, then cover with another thin layer. Allow all taped seams to set.

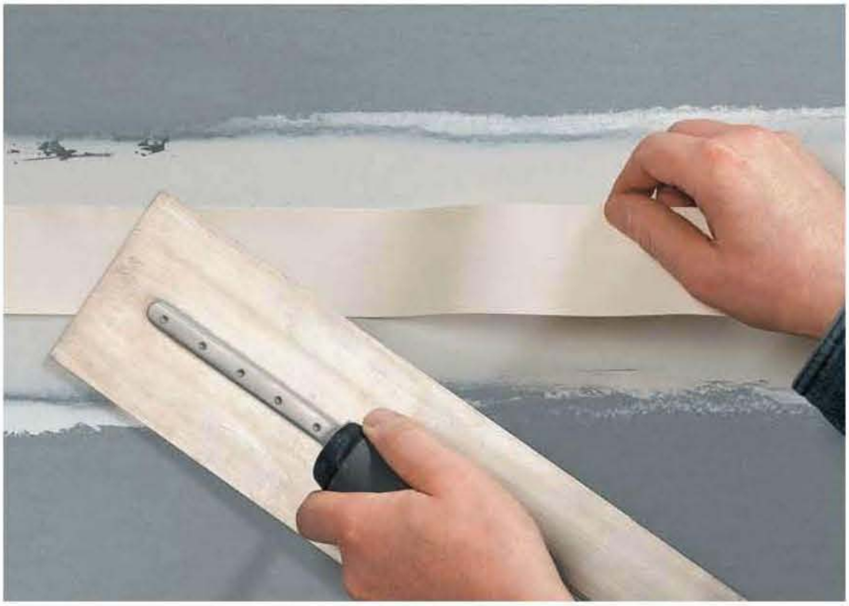

Variation: Blueboard joints can also be reinforced with paper tape. Embed the tape in a thin plaster bed, then and cover with another thin layer to conceal the tape fully. *Note: Some manufacturers recommend setting-type compound for embedding paper tape; always follow the manufacturer's directions for the products you use.*

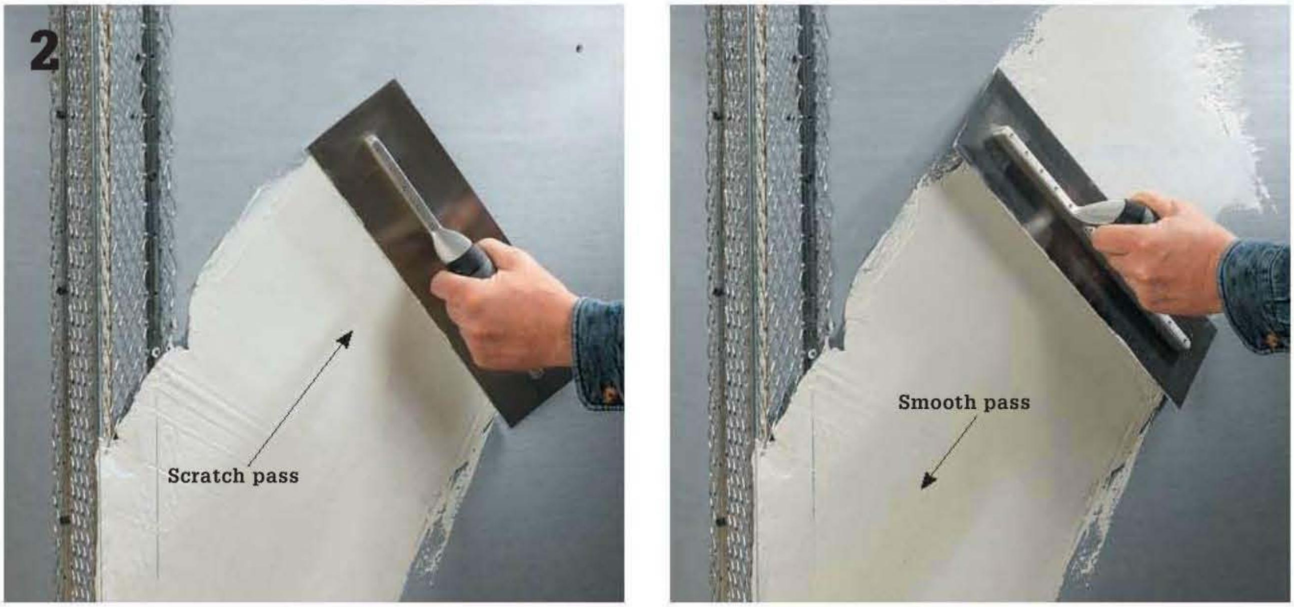

After the seams have set, begin plastering the surface, beginning at one corner and moving to the opposite. Start with ceilings and then do the walls, completing one entire surface before moving on to the next. To apply the plaster, tightly scratch in the material up the wall (photo left), then immediately double-back over it, smoothing over the material to a thickness of $\frac{1}{16}$ to $\frac{3}{32}$", as specified by the manufacturer. Use tight, quick strokes to apply the plaster during the "scratch pass" and long, even strokes to achieve consistency during the "smooth pass."

(continued)

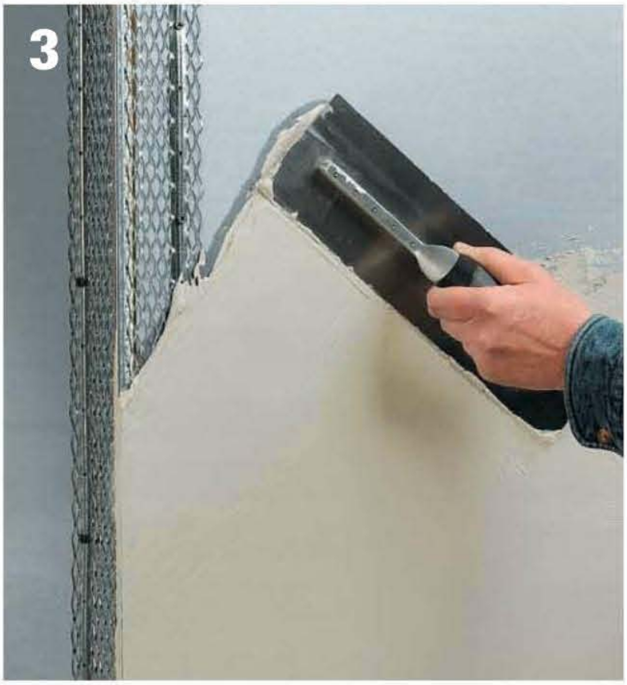
3

Continue to apply plaster by scratching in and smoothing over the surface. Don't worry about uniformity and trowel ridges at this point. Rather, make sure the entire surface is completely concealed with a relatively even plaster coat, $\frac{1}{16}$" to $\frac{3}{32}$"-thick.

4

Once the plaster begins to firm, trowel the surface to fill any voids and remove tooling marks and imperfections, integrating the surface into a uniform smoothness.

5

Prior to the plaster setting, make a final pass with the trowel to smooth the surface, using water sparingly. Do not over trowel; stop before the plaster begins to darken and sets.

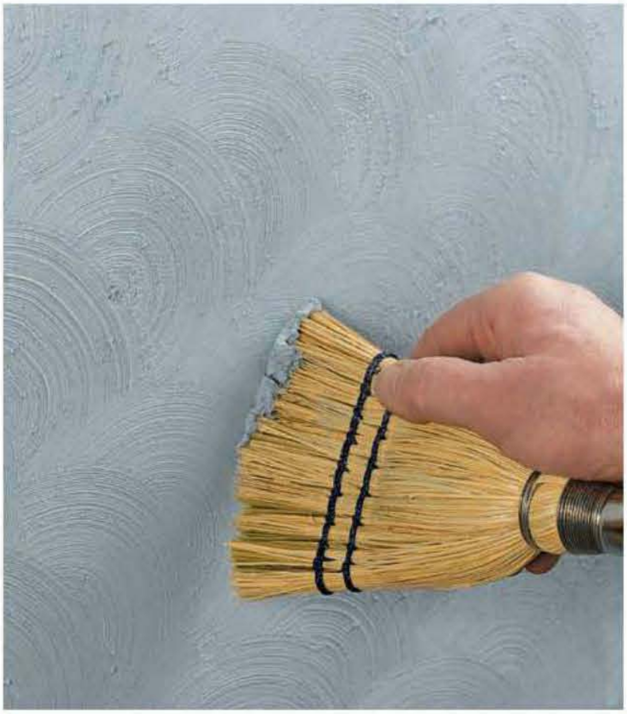

Variation: For textured surfaces, skip the final troweling and work the surface with a texturing tool to achieve the desired results. *Note: Sand or texture added to the plaster mixture does not require tooling.*

How to Apply Basecoat in a Two-Coat Veneer Plaster System

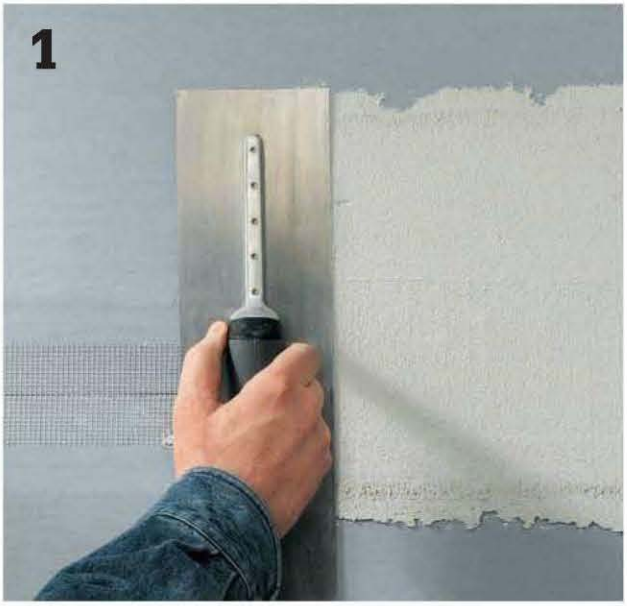

Apply a thin layer of basecoat along all flat seams and corner bead, feathering out the edges by 6". For inside corners, apply a thin bed of basecoat and embed the loose tape, then cover with another thin layer. Allow all taped seams to set.

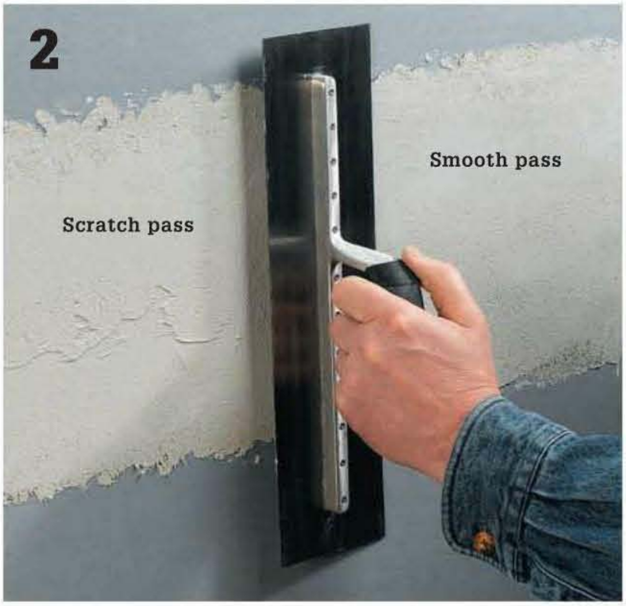

After the seams have set, tightly scratch in basecoat, then immediately double-back over it, smoothing over the material to a thickness of 1/16" to 1/8", as specified by the manufacturer. Use tight, quick strokes to apply basecoat for the "scratch pass" and long, even strokes to achieve consistency for the "smooth pass."

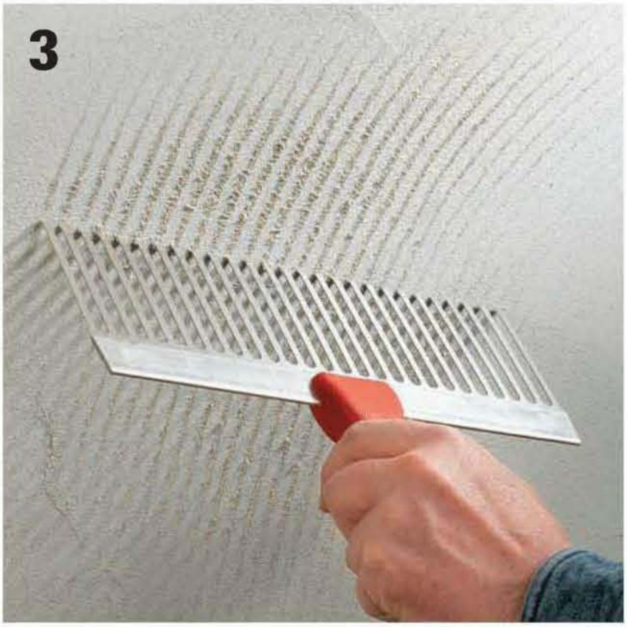

Once the plaster begins to firm or "take up," trowel the surface to fill any voids and remove tooling marks and imperfections, integrating the surface into a reasonably uniform surface—do not over-trowel to a smooth surface. Create keys for the final coat, using a thin-wire rake to roughen the basecoat.

Approximately two hours after the basecoat has set, the finish coat can be applied using the same techniques as for a one-coat veneer plaster system (see pages 225 to 226).

Paneling Ceilings

Tongue-and-groove paneling is a pleasing alternative to a wallboard ceiling, particularly in a knee-wall attic. Pine paneling is most common, but any tongue-and-groove material can be used. These materials are typically ⅜" to ¾" thick and are attached directly to ceiling joists and rafters (over faced insulation, when required). Most codes require you to install ⅜" wallboard as a fire stop under ceiling material thinner than ¼".

Allow for waste by purchasing 15% more material than the square footage of the ceiling; add more for waste if the ceiling requires many angled cuts. Since the tongue portion on most pieces slips into the groove on an adjacent piece, square footage is based on the exposed face (called the reveal) once the boards are installed. A compound miter saw is the best tool for ensuring clean cuts. This is especially important if the ceiling includes non-90° angles.

Tongue-and-groove boards are attached with flooring nails driven through the shoulder of the tongue into each rafter (called blindnailing because the nail heads are covered by the next board). Nailing through the board face is only necessary on the first and last course and on scarf joints.

Layout is very important to the success of a paneled surface, because the lines clearly reveal flaws such as pattern deviations, misaligned walls, and installation mistakes. Before beginning the installation, measure to see how many boards will be installed (using the reveal measurement). If the final board will be less than 2 inches wide, trim the first, or starter board, by trimming the long edge that abuts the wall.

If the angle of the ceiling peak is not parallel to the wall, you must compensate for the difference by ripping the starter piece at an angle so that the leading edge, and every piece thereafter, is parallel to the peak.

Tools & Materials ▸

- Chalk line
- Compound miter saw
- Circular saw
- Drill
- Nail set
- Tongue-and-groove paneling
- 1¾" spiral flooring nails
- Trim molding

Tongue-and-groove paneling can be installed directly over rafters or joists or over wallboard. In attic installations, it's important to insulate first, adding a separate vapor barrier if required by building codes. Local code may also require that paper-faced insulation behind a kneewall be covered with drywall or other material.

How to Panel a Ceiling

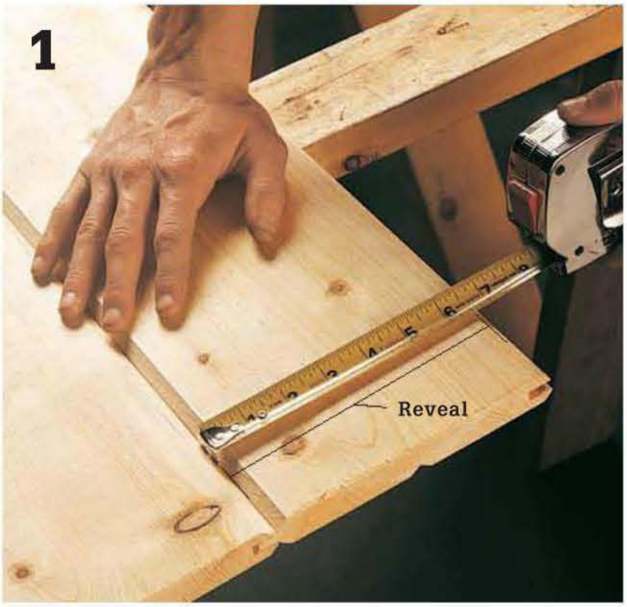

To plan your layout, first measure the reveal of the boards. Fit two pieces together and measure from the bottom edge of the upper board to the bottom edge of the lower board. Calculate the number of boards needed to cover one side of the ceiling by dividing the reveal dimension into the overall distance between the top of the wall and the peak.

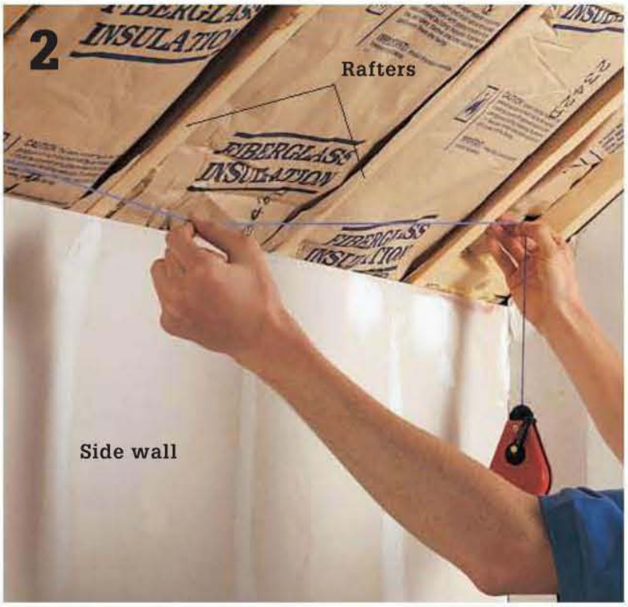

Use the calculation from step 1 to make a control line for the first row of panels—the starter boards. At both ends of the ceiling, measure down from the peak an equal distance, and make a mark to represent the top (tongue) edges of the starter boards. Snap a chalk line through the marks.

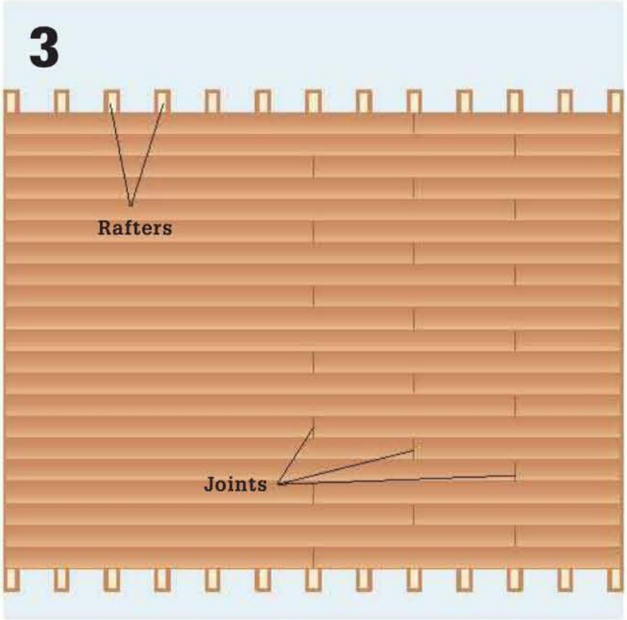

If the boards aren't long enough to span the entire ceiling, plan the locations of the joints. Staggering the joints in a three-step pattern will make them less conspicuous. Note that each joint must fall over the middle of a rafter. For best appearance, select boards of similar coloring and grain for each row.

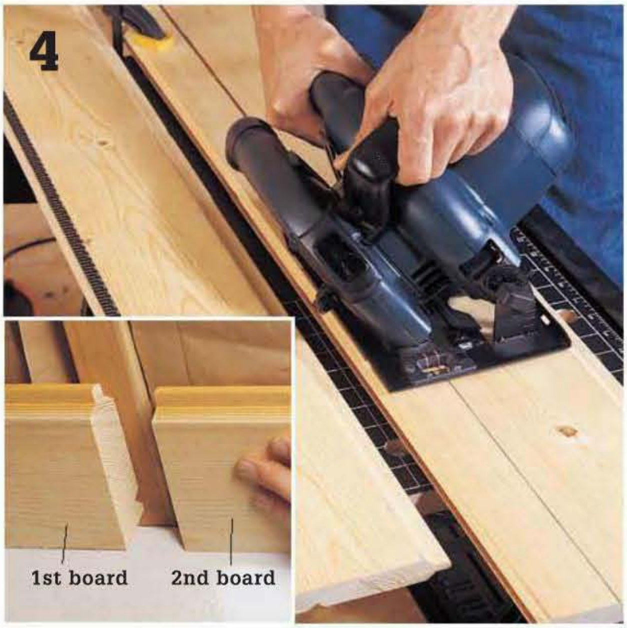

Rip the first starter board to width by bevel-cutting the bottom (grooved) edge. If the starter row will have joints, cut the board to length using a 30° bevel cut on the joint end only. Two beveled ends joined together form a scarf joint (inset), which is less noticeable than a butt joint. If the board spans the ceiling, square-cut both ends.

(continued)

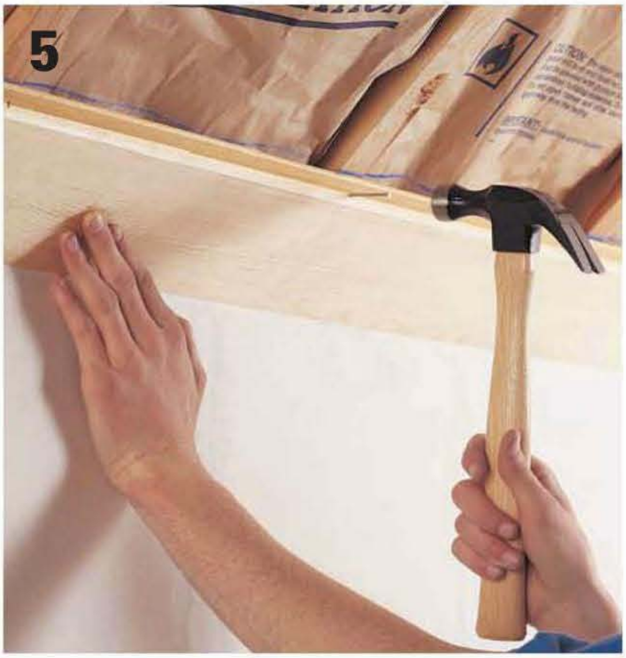

Position the first starter board so the tongue is on the control line. Leave a ⅛" gap between the square board end and the end wall. Fasten the board by nailing through its face about 1" from the grooved edge and into the rafters. Then, blind-nail through the base of the tongue into each rafter, angling the nail backward at 45°. Drive the nail heads beneath the wood surface, using a nail set.

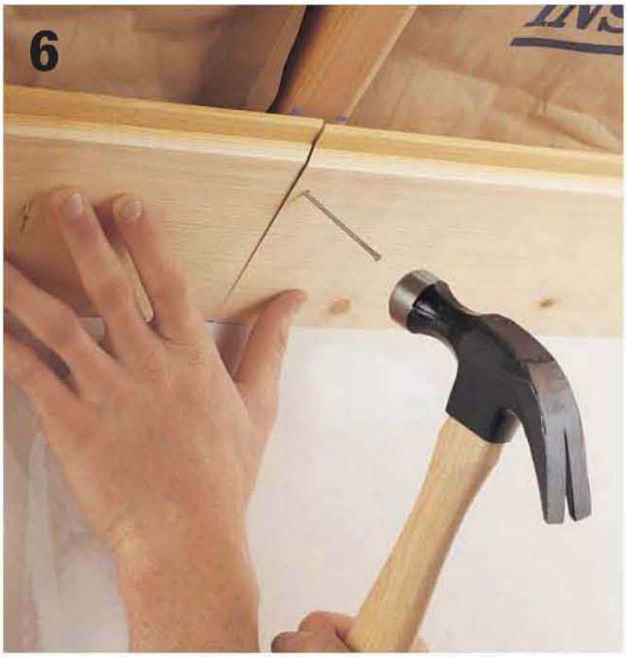

Cut and install any remaining boards in the starter row one at a time, making sure the scarf joints fit together tightly. At each scarf joint, drive two nails through the face of the top board, angling the nail to capture the end of the board behind it. If necessary, predrill the nail holes to prevent splitting.

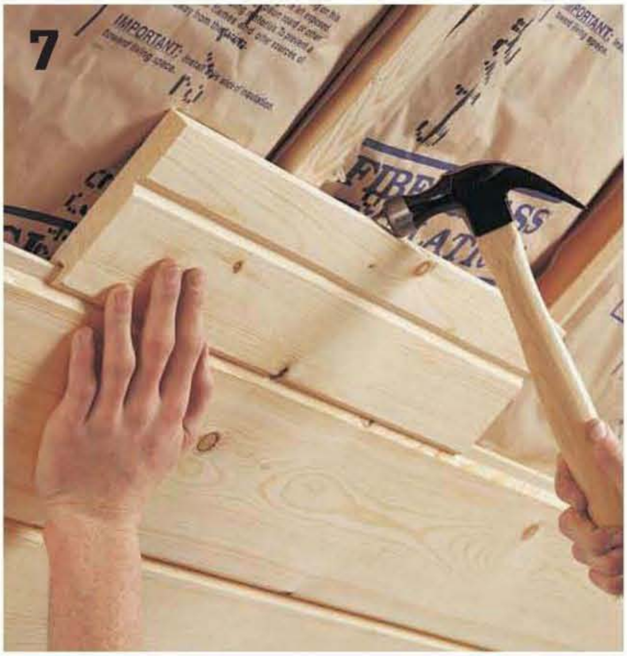

Cut the first board for the next row, then fit its grooved edge over the tongue of the board in the starter row. Use a hammer and a scrap piece of paneling to drive downward on the tongue edge, seating the grooved edge over the tongue of the starter board. Fasten the second row with blind-nails only.

8
FIBERGLASS INSULATION
FIBERGLASS INSULATION

As you install successive rows, measure down from the peak to make sure the rows remain parallel to the peak. Correct any misalignment by adjusting the tongue-and-groove joint slightly with each row. You can also snap additional control lines to help align the rows.

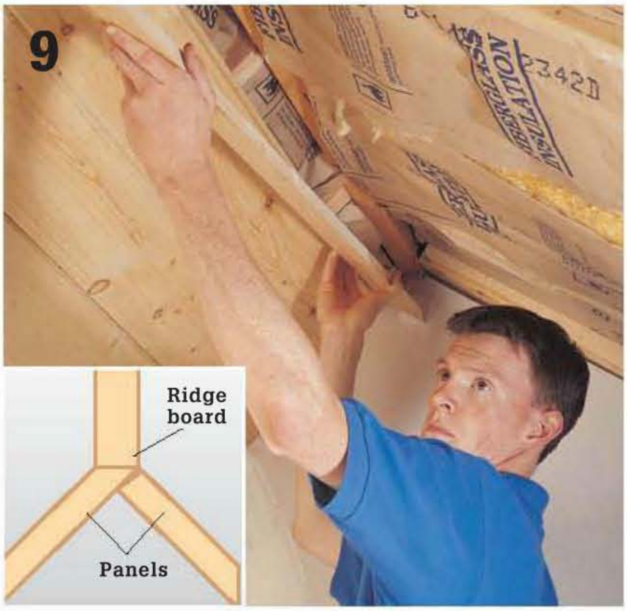

Rip the boards for the last row to width, beveling the top edges so they fit flush against the ridge board. Facenail the boards in place. Install paneling on the other side of the ceiling, then cut and install the final row of panels to form a closed joint under the ridge board (inset).

Install trim molding along walls, at joints around obstacles, and along inside and outside corners, if desired. (Select-grade 1 × 2 works well as trim along walls.) Where necessary, bevel the back edges of the trim or miter-cut the ends to accommodate the slope of the ceiling.

Tips for Paneling an Attic Ceiling ▸

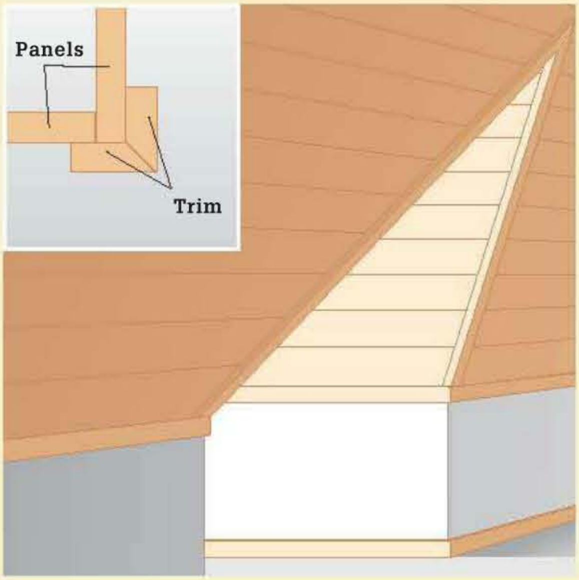

Use mitered trim to cover joints where panels meet at outside corners. Dormers and other roof elements create opposing ceiling angles that can be difficult to panel around. It may be easier to butt the panels together and hide the butt joints with custom-cut trim. The trim also makes a nice transition between angles.

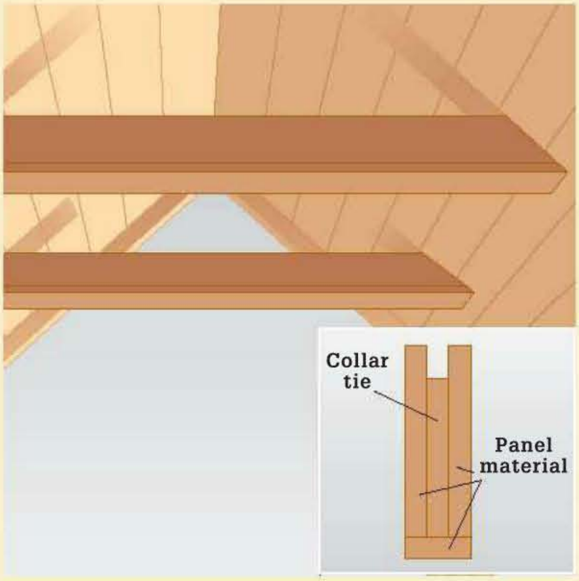

Wrap collar ties or exposed beams with custom-cut panels. Install the paneling on the ceiling first. Then, rip-cut panels to the desired width. You may want to include a tongue-and-groove joint as part of the trim detail. Angle-cut the ends of the trim so it fits tight to the ceiling panels.

Installing Suspended Ceilings

Suspended ceilings are traditionally popular ceiling finishes for basements and utility areas, particularly because they hang below pipes and other mechanicals while providing easy access to them. However, the commercial appearance and grainy texture of basic ceiling tiles make them an unlikely choice for formal areas such as living rooms. Basic tiles are not your only option.

Suspended ceiling tile manufacturers have a wide array of ceiling tiles to choose from that go above and beyond traditional institutional tiles. Popular styles mimic historical tin tiles and add depth to the ceiling while minimizing sound and vibration noise.

A suspended ceiling is a grid framework made of lightweight metal brackets hung on wires attached to ceiling or floor joists. The frame consists of T-shaped main beams (mains), cross-tees (tees), and L-shaped wall angles. The grid supports ceiling panels, which rest on the flanges of the framing pieces. Panels are available in 2 × 2-ft. or 2 × 4-ft. sizes, in a variety of styles. Special options include insulated panels, acoustical panels that absorb sound, and light-diffuser screens for use with fluorescent lights. Generally, metal-frame ceiling systems are more durable than ones made of plastic.

To begin your ceiling project, devise the panel layout based on the size of the room, placing equally sized trimmed panels on opposite sides to create a balanced look. Your ceiling must also be level. For small rooms, a 4-foot or 6-foot level will work, but a water level is more effective for larger jobs. You can make a water level with two water-level ends (available at hardware stores and home centers) attached to flexible plastic tubing.

Tools & Materials ▸

- Water level
- Chalk line
- Drill
- Aviation snips
- String
- Lock-type clamps
- Screw-eye driver
- Pliers
- Straightedge
- Utility knife
- Suspended ceiling kit (frame)
- Screw eyes
- Hanger wires
- Ceiling panels
- 1½" drywall screws or masonry nails

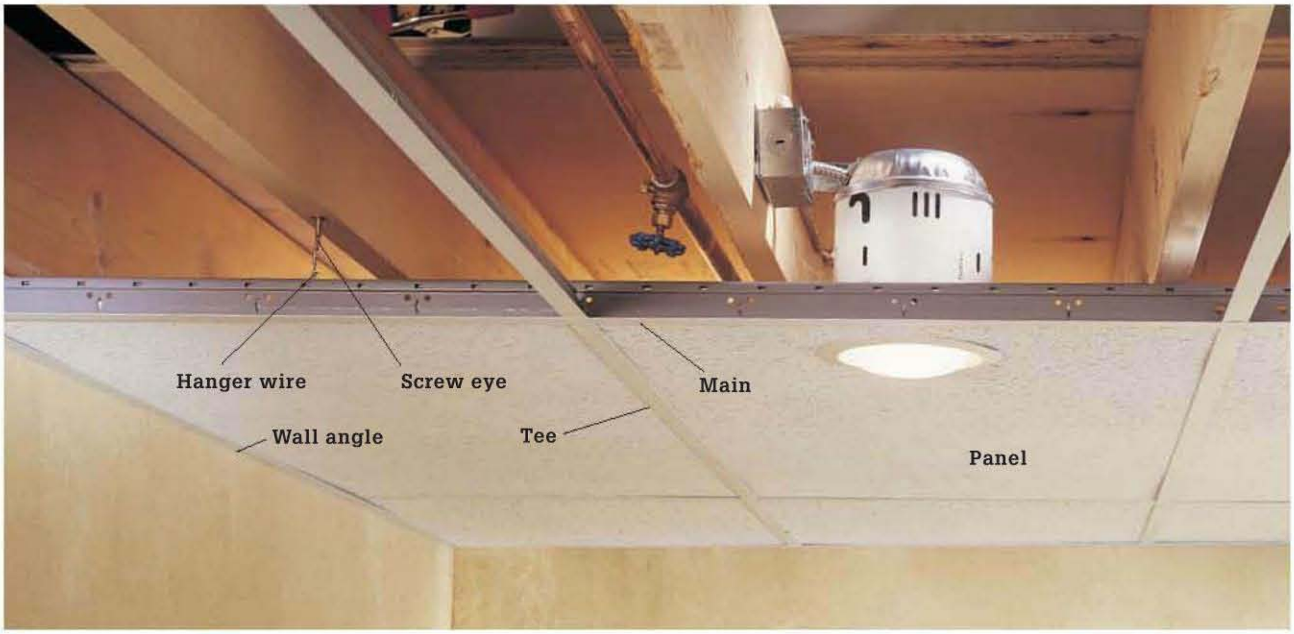

Suspended ceilings are the most common ceiling choice in areas, such as basements, where access to the joist cavities needs to be maintained. They're also installed as a way to conceal problem ceilings.

Installing a Suspended Ceiling

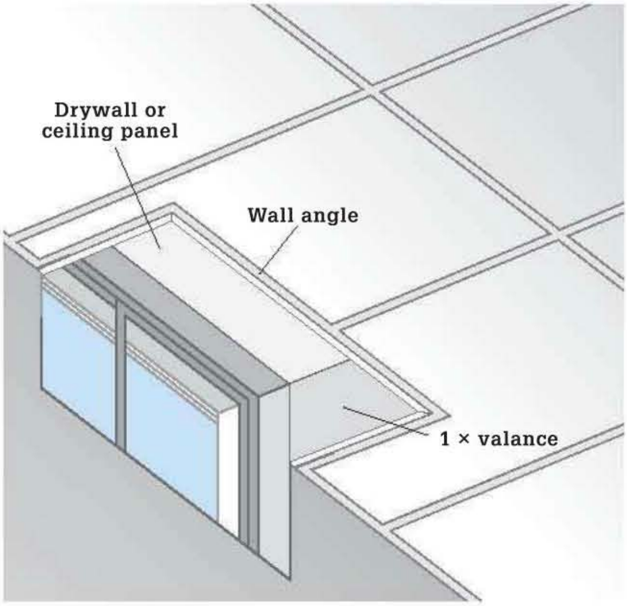
Drywall or ceiling panel
Wall angle
1 × valance

Build a valance around basement awning windows so they can be opened fully. Attach 1 × lumber of an appropriate width to joists or blocking. Install drywall (or a suspended-ceiling panel trimmed to fit) to the joists inside the valance.

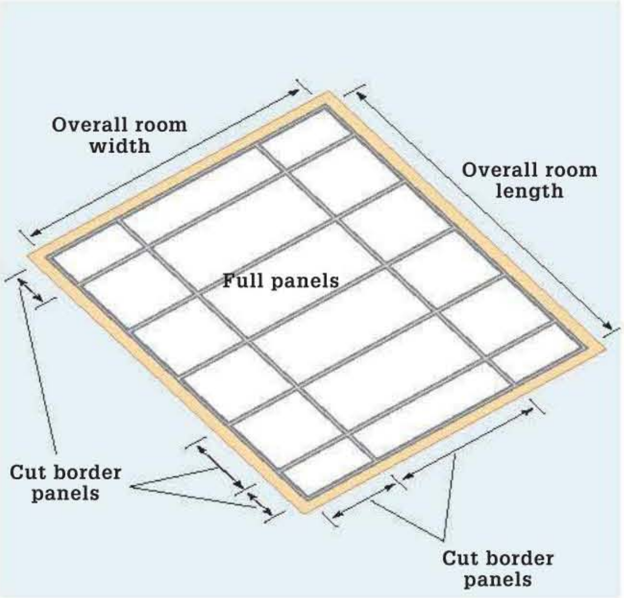
Overall room width
Overall room length
Full panels
Cut border panels
Cut border panels

Draw your ceiling layout on paper, based on the exact dimensions of the room. Plan so that trimmed border panels on opposite sides of the room are of equal width and length (avoid panels smaller than ½-size). If you include lighting fixtures in your plan, make sure they follow the grid layout.

How to Install a Suspended Ceiling

1

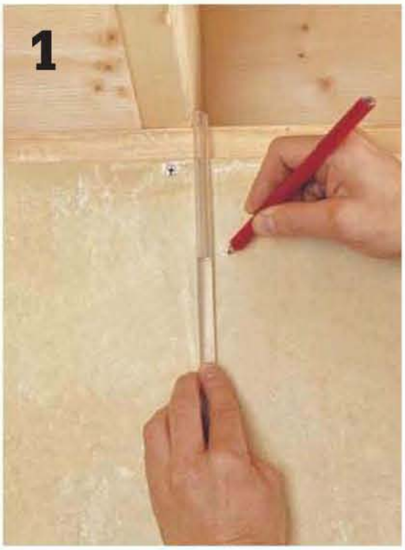

Make a mark on one wall that represents the ceiling height plus the height of the wall angle. Use a water level to transfer that height to both ends of each wall. Snap a chalk line to connect the marks. This line represents the top of the ceiling's wall angle.

2

Attach wall angle pieces to the studs on all walls, positioning the top of the wall angle flush with the chalk line. Use 1½" drywall screws (or short masonry nails driven into mortar joints on concrete block walls). Cut angle pieces using aviation snips.

Tip ▸

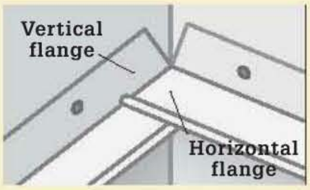
Vertical flange
Horizontal flange

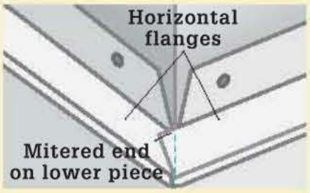
Horizontal flanges
Mitered end on lower piece

Trim wall angle pieces to fit around corners. At inside corners (top), back-cut the vertical flanges slightly, then overlap the horizontal flanges. At outside corners (bottom), miter-cut one horizontal flange, and overlap the flanges.

(continued)

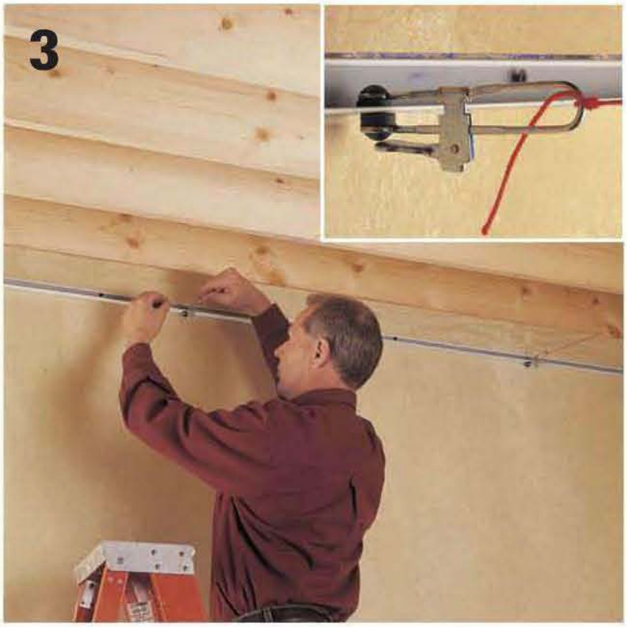

Mark the location of each main on the wall angles at the ends of the room. The mains must be parallel to each other and perpendicular to the ceiling joists. Set up a guide string for each main, using a thin string and lock-type clamps (inset). Clamp the strings to the opposing wall angles, stretching them very taut so there's no sagging.

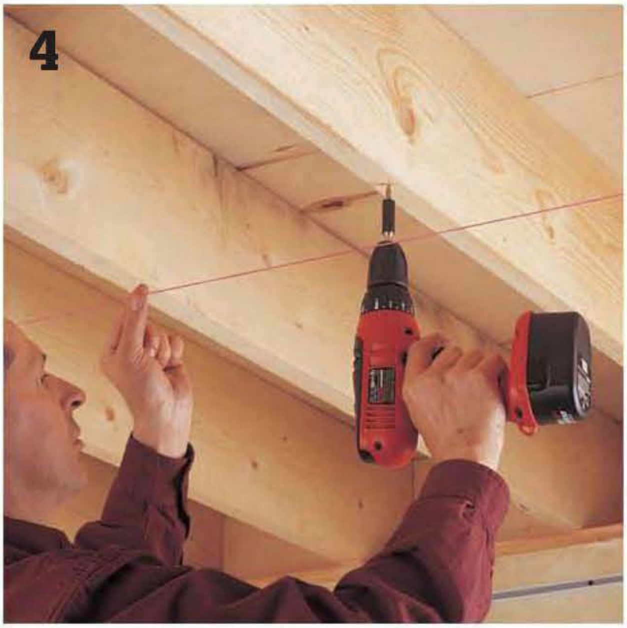

Install screw eyes for hanging the mains, using a drill and screw-eye driver. Drill pilot holes and drive the eyes into the joists every 4 ft., locating them directly above the guide strings. Attach hanger wire to the screw eyes by threading one end through the eye and twisting the wire on itself at least three times. Trim excess wire, leaving a few inches of wire hanging below the level of the guide string.

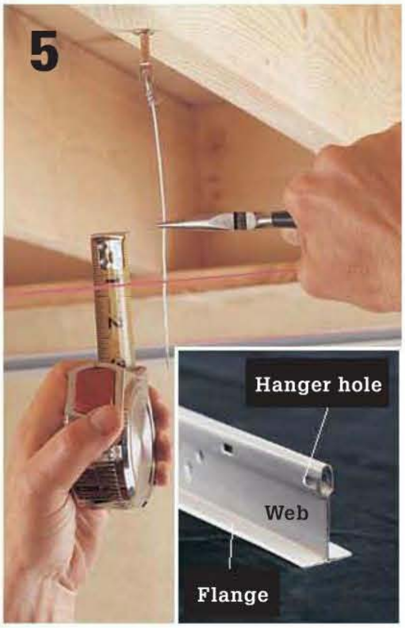

Measure the distance from the bottom of a main's flange to the hanger hole in the web (inset). Use this measurement to prebend each hanger wire. Measure up from the guide string and make a 90° bend in the wire, using pliers.

Following your ceiling plan, mark the placement of the first tee on opposite wall angles at one end of the room. Set up a guide string for the tee, using a string and clamps, as before. This string must be perpendicular to the guide strings for the mains.

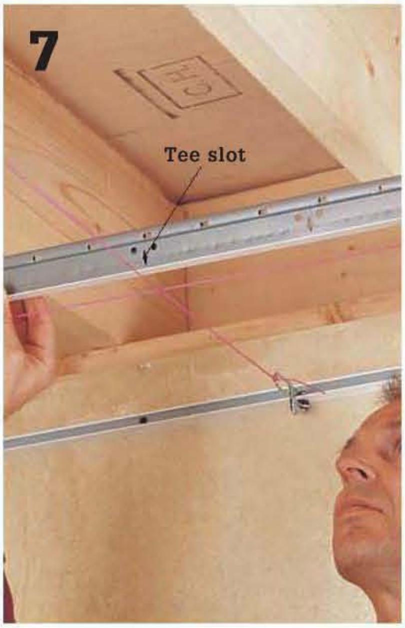

Trim one end of each main so that a tee slot in the main's web is aligned with the tee guide string, and the end of the main bears fully on a wall angle. Set the main in place to check the alignment of the tee slot with the string.

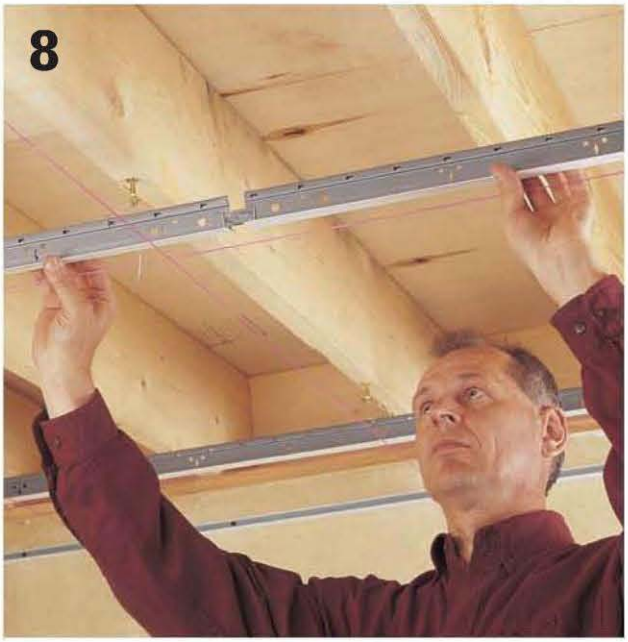

Cut the other end of each main to fit, so that it rests on the opposing wall angle. If a single main cannot span the room, splice two mains together, end-to-end (the ends should be fashioned with male-female connectors). Make sure the tee slots remain aligned when splicing.

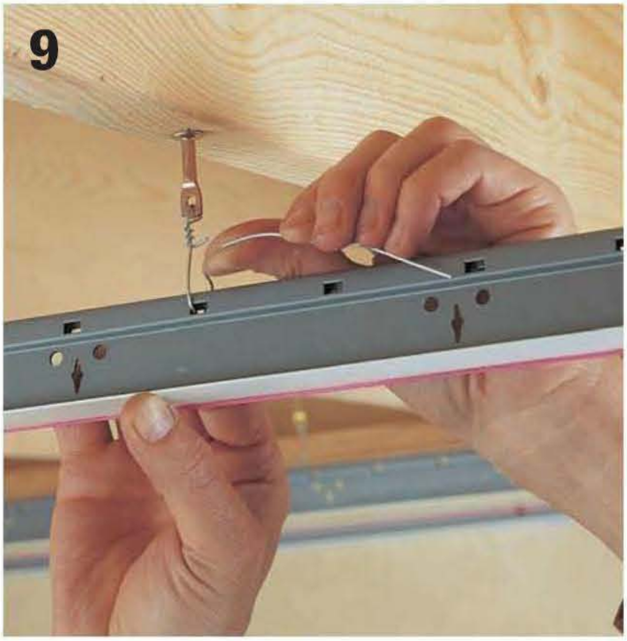

Install the mains by setting the ends on the wall angle and threading the hanger wires through the hanger holes in the webs. The wires should be as close to vertical as possible. Wrap each wire around itself three times, making sure the main's flange is level with the main guide string. Also install a hanger near each main splice.

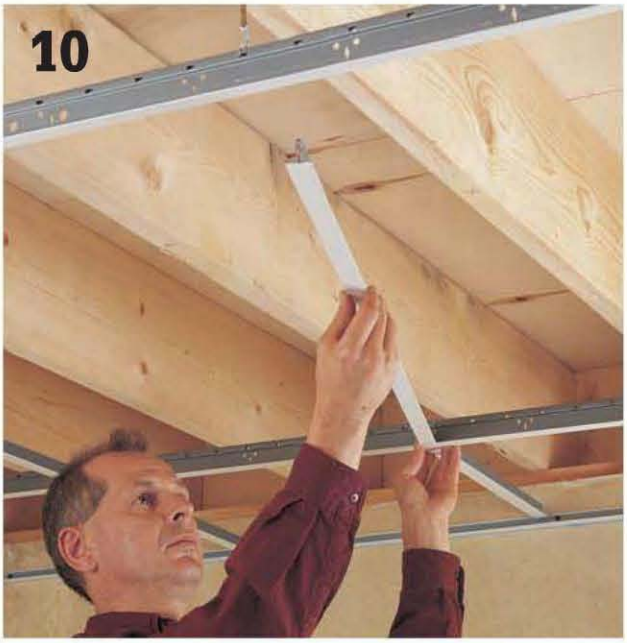

Attach tees to the mains, slipping the tabbed ends into the tee slots on the mains. Align the first row of tees with the tee guide string; install the remaining rows at 4-ft. intervals. If you're using 2 × 2-ft. panels, install 2-ft. cross-tees between the midpoints of the 4-ft. tees. Cut and install the border tees, setting the tee ends on the wall angles. Remove all guide strings and clamps.

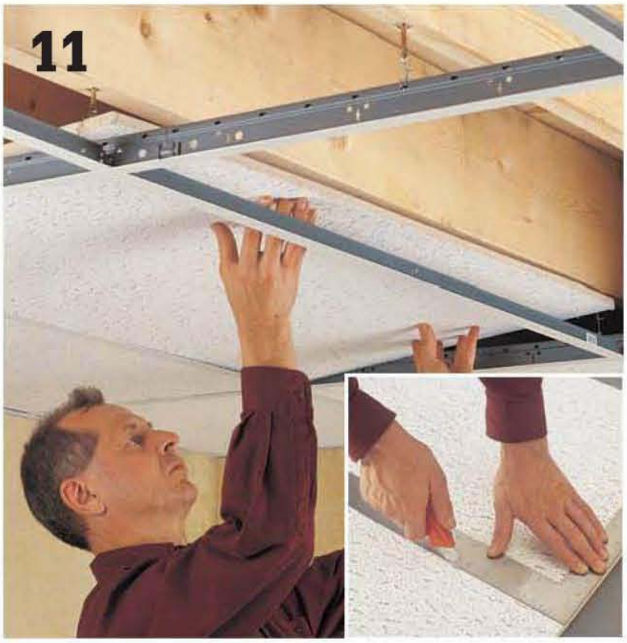

Place full ceiling panels into the grid first, then install the border panels. Lift the panels in at an angle, and position them so they rest on the frame's flanges. Reach through adjacent openings to adjust the panels, if necessary. To trim the border panels to size, cut them face-up, using a straightedge and utility knife (inset).

Installing Acoustical Ceiling Tiles

Easy-to-install ceiling tile can lend character to a plain ceiling or help turn an unfinished basement or attic into beautiful living space. Made of pressed mineral and fiberboard, ceiling tiles are available in a variety of styles. They also provide moderate noise reduction.

Ceiling tiles typically can be attached directly to a drywall or plaster ceiling with adhesive. If your ceiling is damaged or uneven, or if you have an unfinished joist ceiling, install 1 × 2 furring strips as a base for the tiles, as shown in this project. Some systems include metal tracks for clip-on installation.

Unless your ceiling measures in even feet, you won't be able to install the 12-inch tiles without some cutting. To prevent an unattractive installation with small, irregular tiles along two sides, include a course of border tiles along the perimeter of the installation. Plan so that tiles at opposite ends of the room are cut to the same width and are at least half the width of a full tile.

Most ceiling tile comes prefinished, but it can be painted to match any decor. For best results, apply two coats of paint using a roller with a ¼" nap. Wait 24 hours between coats.

Tools & Materials ▸

4-ft. level
Stepladder
Chalk line
Utility knife
Straightedge
Hammer or drill
Handsaw
Stapler
1 × 2 furring strips
8d nails or 2" screws
String
Ceiling tiles
Staples
Trim molding

Acoustic tile improves the sound quality inside a home theater and reduces the transmission of sound to the rooms surrounding it.

How to Install Ceiling Tile

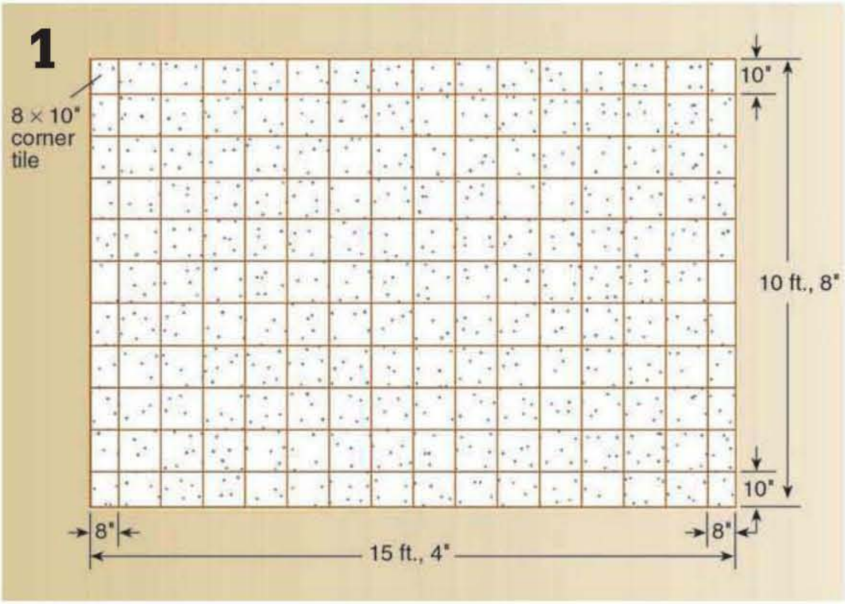

Measure the ceiling and devise a layout. If the length (or width) doesn't measure in even feet, use this formula to determine the width of the border tiles: add 12 to the number of inches remaining and divide by 2. The result is the width of the border tile. (For example, if the room length is 15 ft., 4", add 12 to the 4, then divide 16 by 2, which results in an 8" border tile.)

Install the first furring strip flush with the wall and perpendicular to the joists, fastening with two 8d nails or 2" screws at each joist. Measure out from the wall a distance equal to the border tile width minus ¾", and snap a chalk line. Install the second furring strip with its wall-side edge on the chalk line.

Install the remaining strips 12" on-center from the second strip. Measure from the second strip and mark the joist nearest the wall every 12". Repeat along the joist on the opposite side of the room, then snap chalk lines between the marks. Install the furring strips along the lines. Install the last furring strip flush against the opposite side wall. Stagger the butted end joints of strips between rows so they aren't all on the same joist.

Check the strips with a 4-ft. level. Insert wood shims between the strips and joists as necessary to bring the strips into a level plane.

(continued)

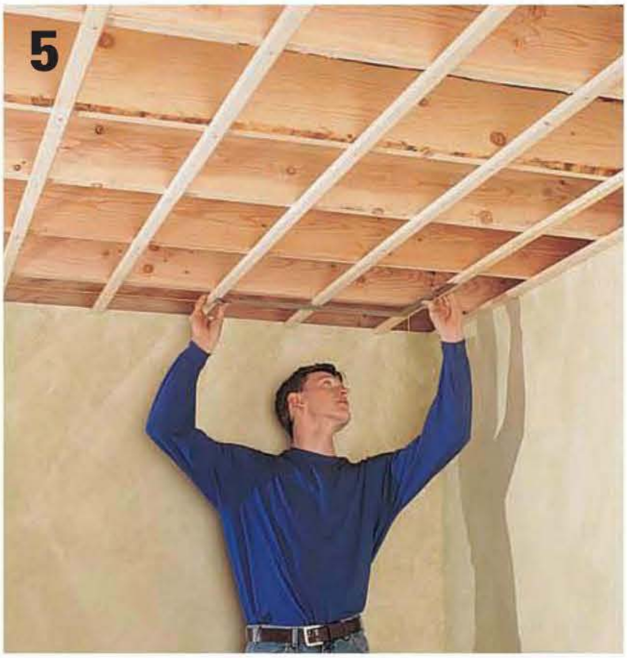

Set up taut, perpendicular string lines along two adjacent walls to help guide the tile installation. Inset the strings from the wall by a distance that equals that wall's border tile width plus ½". Use a framing square to make sure the strings are square.

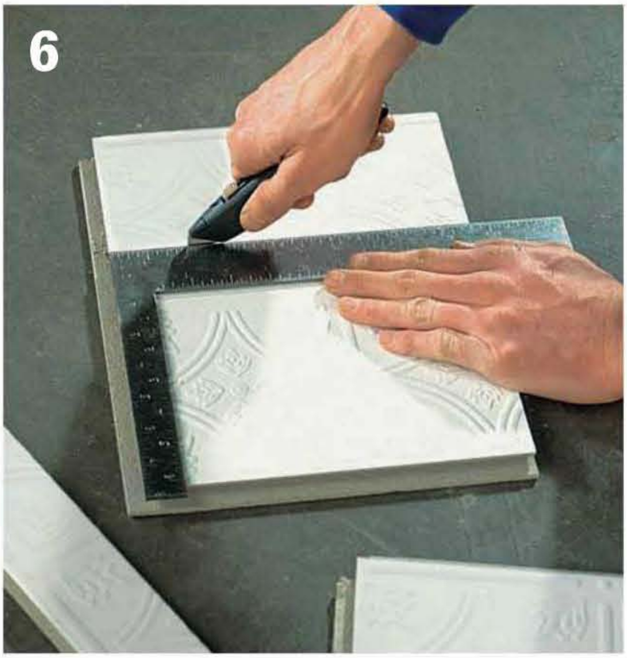

Cut the corner border tile to size with a utility knife and straightedge. Cutting the border tiles ¼" short will ease fitting them. The resulting gap between the tile and wall will be covered by trim. Cut only on the edges without the stapling flange.

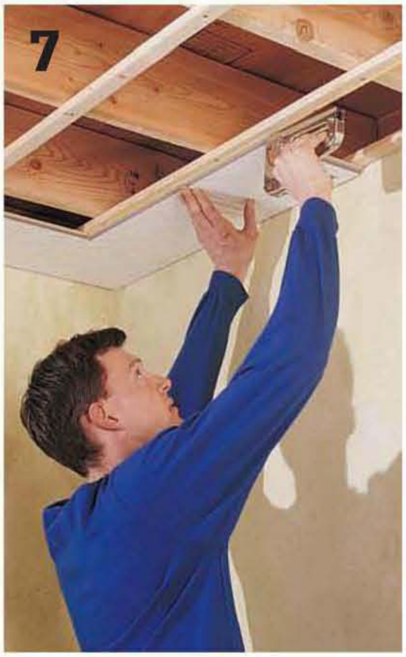

Position the corner tile with the flange edges aligned with the two string lines and fasten it to the furring strips with four ½" staples. Cut and install two border tiles along each wall, making sure the tiles fit snugly together.

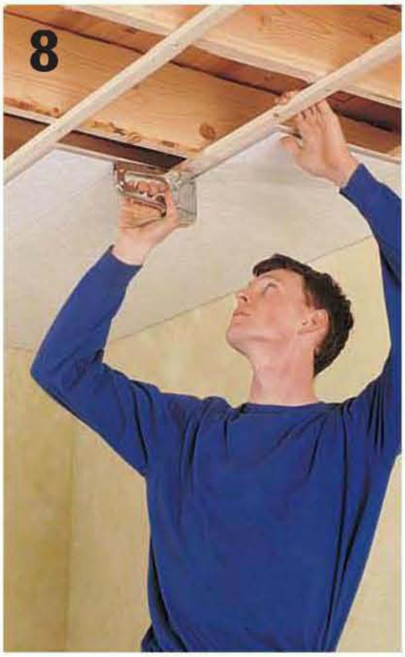

Fill in between the border tiles with full-size tiles. Continue working diagonally in this manner, toward the opposite corner. For the border tiles along the far wall, trim off the flange edges and staple through the faces of the tiles, close to the wall.

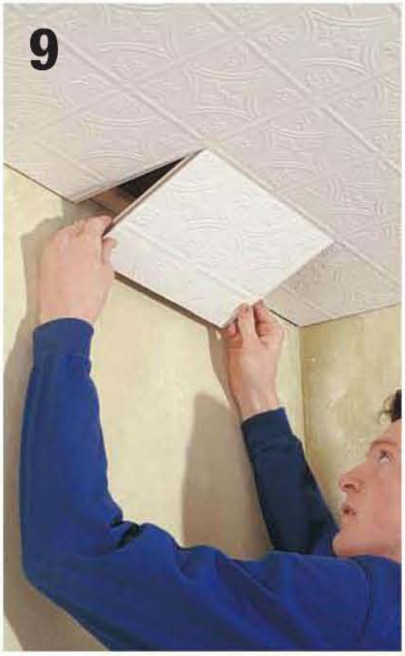

Install the final row of tiles, saving the far corner tile and its neighbor for last. Cut the last tile to size, then remove the tongue and nailing flange along the side edges. Finish the job by installing trim along the edges.

Installing Metal Ceilings

Today's metal ceilings offer the distinctive elegance of 19th-century tin tile in a durable, washable ceiling finish. Available at home centers and specialty distributors, metal ceiling systems include field panels (in 2 × 2-, 2 × 4-, and 2 × 8-ft. sizes), border panels that can be cut to fit your layout, and cornice molding for finishing the edges. The panels come in a variety of materials and finishes ready for installation, or they can be painted.

To simplify installation, the panels have round catches, called nailing buttons, that fit into one another to align the panels where they overlap. The buttons are also the nailing points for attaching the panels. Use 1" decorative conehead nails where nail heads will be exposed, and ½" wire nails where heads are hidden.

Install your metal ceiling over a smooth layer of ⅜" or ½" plywood, which can be fastened directly to the ceiling joists with drywall screws, or installed over an existing finish. The plywood provides a flat nailing surface for the panels. As an alternative, some manufacturers offer a track system for clip-on installation.

Begin your installation by carefully measuring the ceiling and snapping chalk lines to establish the panel layout. For most tile patterns, it looks best to cover the center of the space with full tiles only, then fill in along the perimeter with border panels, which are not patterned. Make sure your layout is square.

Tools & Materials ▸

Chalk line
Level
Tin snips
Drill with ⅛" metal bit
Compass
Metal file
⅜" or ½" plywood
2" drywall screws
Field panels
Border panels with molding edge
Cornice molding
Masking tape
½" wire nails
1" conehead nails
Wood block

Real metal ceilings have traditional embossed patterns with an unmistakable luxurious quality. But, they are expensive and a bit unwieldy to install. Acoustic panels with an embossed vinyl pattern layer are much cheaper and easier to install, but are also less authentic.

How to Install a Metal Tile Ceiling

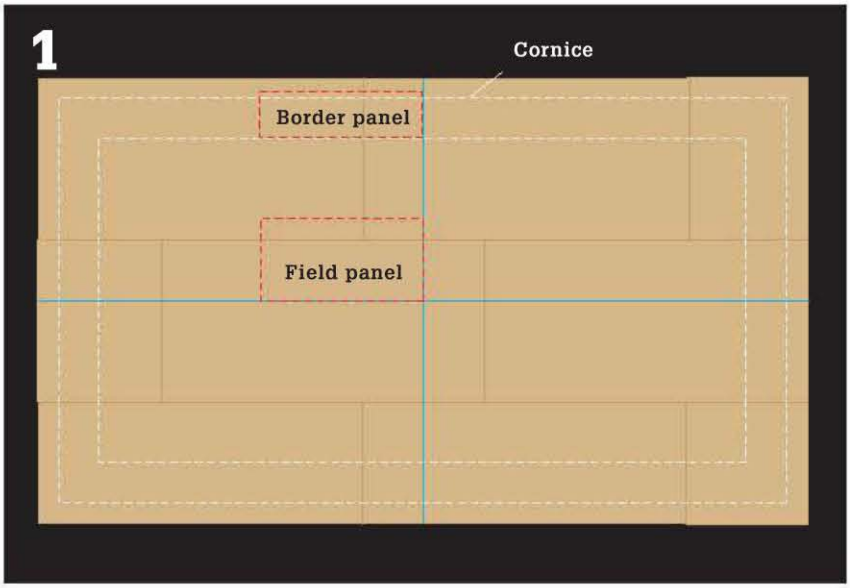

Measure to find the center of the ceiling, then snap perpendicular chalk lines intersecting the center. On the walls, mark a level reference line representing the bottom edges of the cornice molding. Where possible, plan to install the panels so they overlap toward the room's entrance, to help conceal the seams.

Align the first field panel with the chalk lines at the ceiling's center, and attach it with ½" wire nails along the edges where another panel will overlap it. Drive the nails beside the nailing buttons—saving the buttons for nailing the overlapping panel.

Continue to install field panels, working along the length of the area first, then overlapping the next row. Make sure the nailing buttons are aligned. Underlap panels by sliding the new panel into position beneath the installed panel, then fasten through both panels at the nailing buttons, using 1" conehead nails. Where field panels meet at corners, drill ⅛" pilot holes for the conehead nails.

Cut the border panels to width so they will underlap the cornice by at least 1". Use sharp tin snips, and cut from the edge without edge molding. Install the panels so the nailing buttons on the molding align with those on the field panels. Fasten through the buttons with conehead nails, and along the cut edge with wire nails. At corners, miter-cut the panels, and drive conehead nails every 6" along the seam.

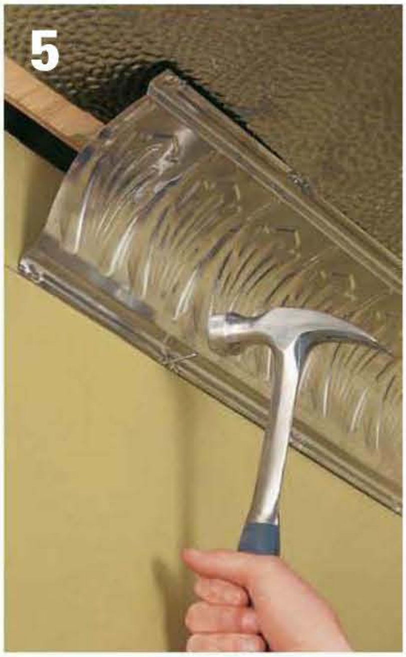

Install each cornice piece with its bottom edge on the level line. Drive 1" conehead nails through the nailing buttons and into the wall studs. Don't nail the ends until the succeeding piece is in place. Fasten the top edges to the ceiling.

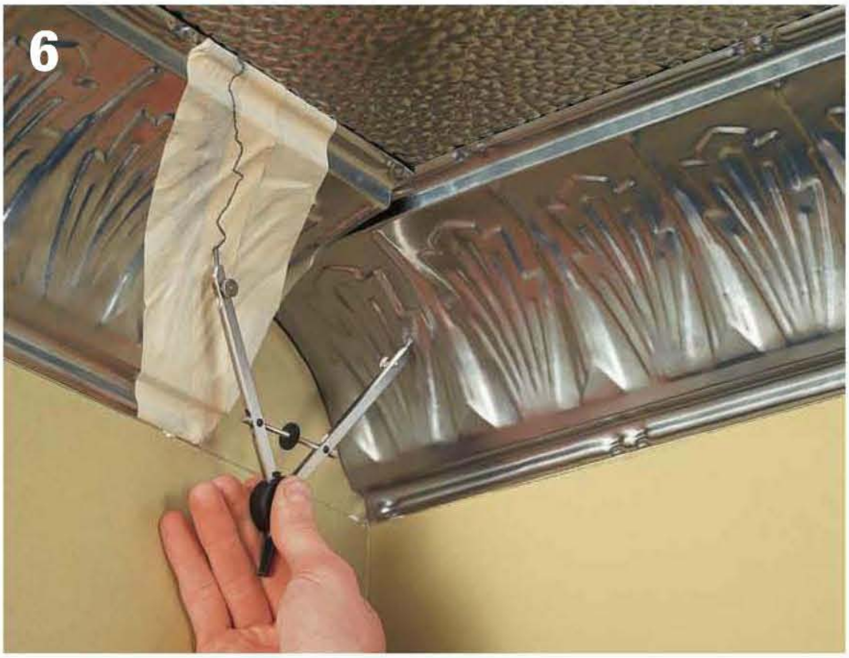

At inside corners, install one cornice piece tightly into the corner, then scribe the mating piece to fit, using masking tape and a compass. Cut along the scribed line with tin snips, and make minor adjustments with a metal file. You may have to cut the mating piece several times, so start with plenty of length. If you have several corners, use this technique to cut templates for the corner pieces.

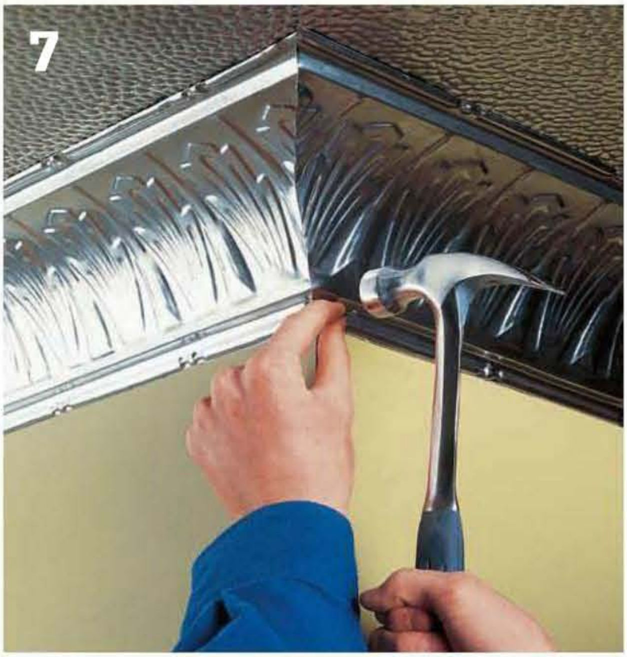

At outside corners, cut the ends of two scrap pieces at a 33° angle. Fit the pieces together at the corner, then trim and mark each piece in turn, making minor adjustments until they fit well. Use the scrap pieces as templates for marking the workpieces. Fasten near the corner only when both mating pieces are in place.

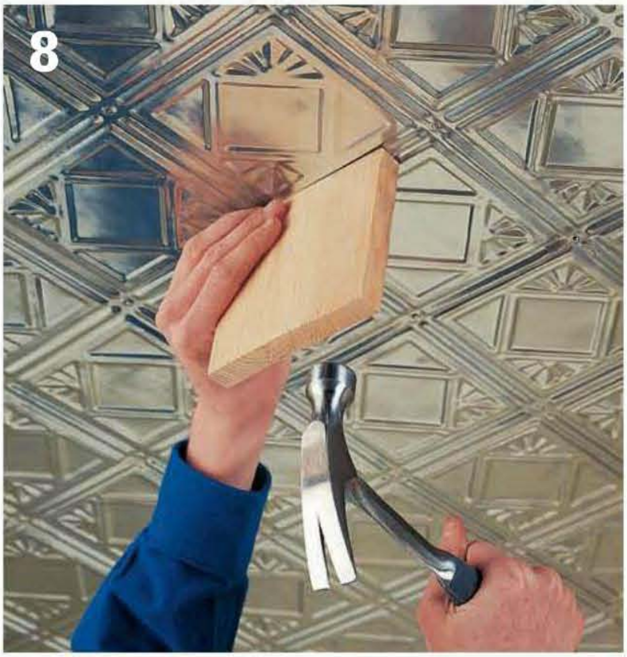

Using a hammer and a piece of wood, carefully tap any loose joints to tighten them. If the cornice will be left unpainted, file the joints for a perfect fit. If you're painting the ceiling, seal the seams with paintable silicone caulk, then apply two coats of paint using a roller with a ¼" nap. Allow the first coat to dry for 24 hours before applying the second coat.

Installing Base Molding

Baseboard trim is installed to conceal the joint between the finished floor and the wallcovering. It also serves to protect the wallboard at the floor. Installing plain, one-piece baseboard such as ranch-style base or cove base is a straightforward project. Outside corner joints are mitered, inside corners are coped, and long runs are joined with scarf cuts.

The biggest challenge to installing base is dealing with out-of-plumb and nonsquare corners. However, a T-bevel makes these obstacles easy to overcome.

Plan the order of your installation prior to cutting any pieces and lay out a specific piece for each length of wall. It may be helpful to mark the type of cut on the back of each piece so you don't have any confusion during the install.

Locate all studs and mark them with painter's tape, 6 inches higher than your molding height. If you need to make a scarf joint along a wall, make sure it falls on the center of a stud. Before you begin nailing trim in place, take the time to pre-finish the moldings. Doing so will minimize the cleanup afterward.

Tools & Materials ▸

- Pencil
- Tape measure
- Power miter saw
- T-bevel
- Coping saw
- Metal file set
- Pneumatic finish nail gun & compressor
- Moldings
- Pneumatic fasteners
- Carpenter's glue
- Finishing putty

How to Install One-piece Base Molding

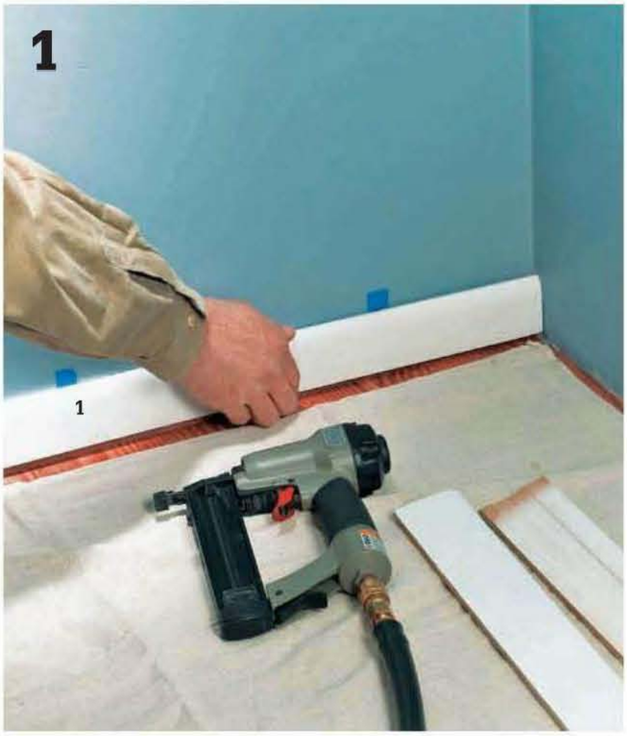

Measure, cut, and install the first piece of baseboard. Butt both ends into the corners tightly. For longer lengths, it is a good idea to cut the piece slightly oversized (up to ⅟₁₆" on strips over 10 ft. long) and "spring" it into place. Nail the molding in place with two nails at every stud location.

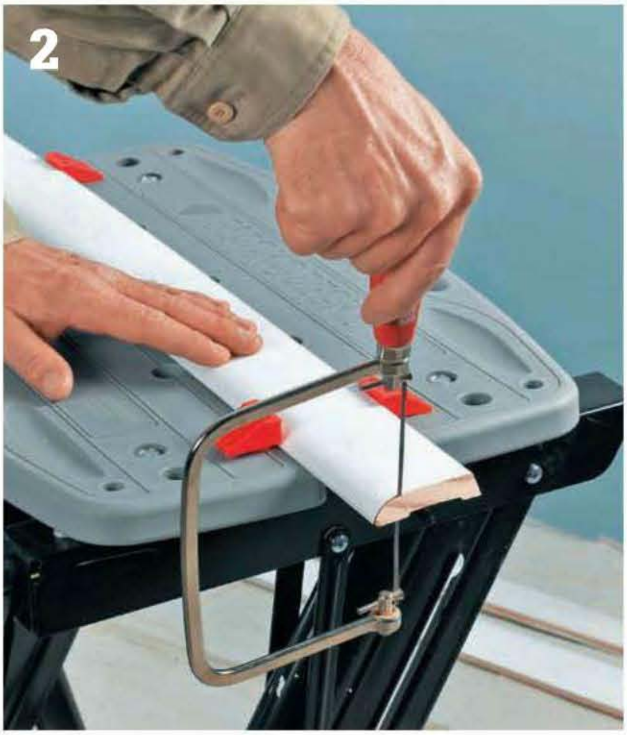

Cut the second piece of molding oversized by 6" to 10" and cope-cut the adjoining end to the first piece. Fine-tune the cope with a metal file and sandpaper. Dry-fit the joint, adjusting it as necessary to produce a tight-fitting joint.

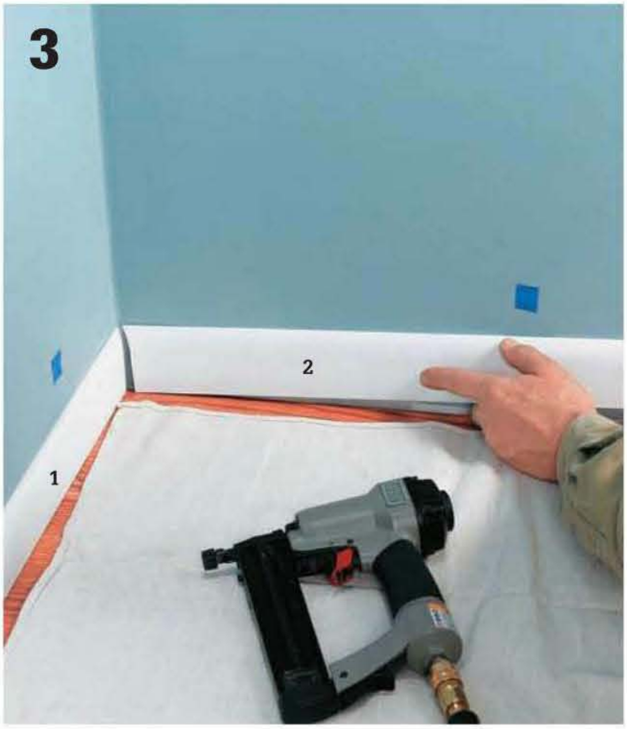

Check the corner for square with a framing square. If necessary, adjust the miter cut of your saw. Use a T-bevel to transfer the proper angle. Cut the second piece (coped) to length and install it with two nails at each stud location.

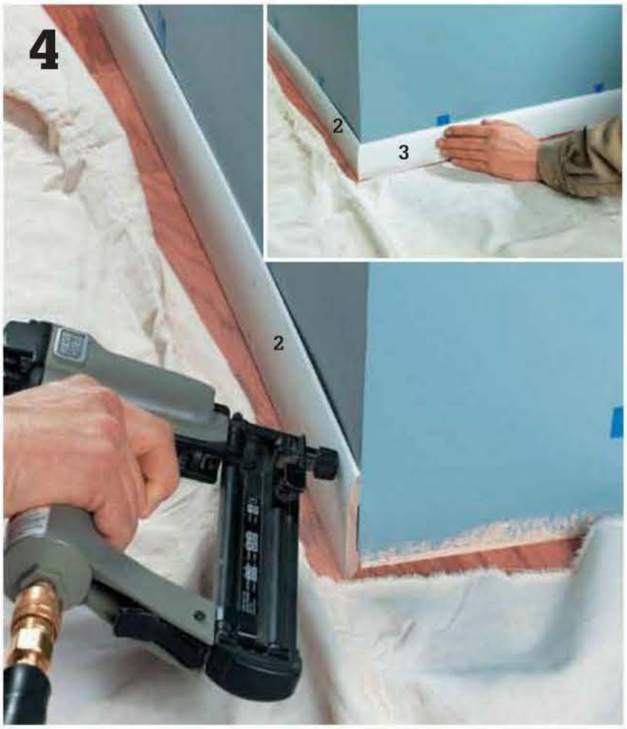

Adjust the miter angle of your saw to cut the adjoining outside corner piece (3). Test-fit the cut to ensure a tight joint (inset photo). Remove the mating piece of trim and fasten the first piece for the outside corner joint.

Lay out any scarf joints by placing the piece in position so that the previous joint is tight and then marking the center of a stud location nearest the opposite end. Set the angle of your saw to a 30° angle and cut the molding at the marked location.

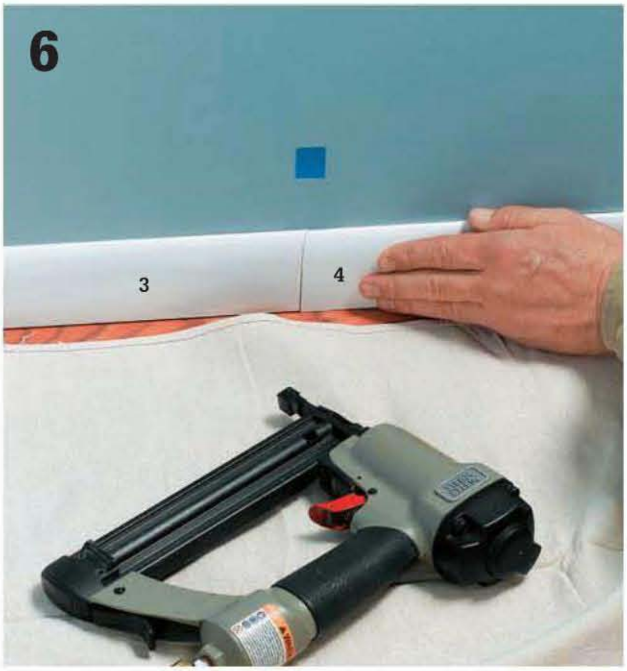

Nail the third piece in place, making sure the outside corner joint is tight. Cut the end of the fourth piece to match the scarf joint angle and nail it in place with two nails at each stud location. Add the remaining pieces of molding, fill the nail holes with putty, and apply a final coat of finish.

How to Install Built-up Base Molding

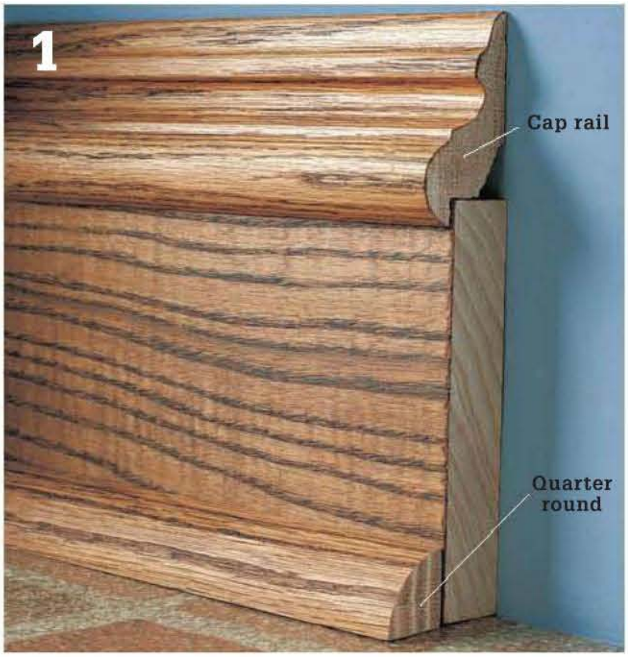
1
Cap rail
Quarter round

Dress up simple baseboard stock with cap moldings and base shoe or quarter round. The baseboard can be made of solid wood, as shown above, or from strips of veneered plywood, as shown at right.

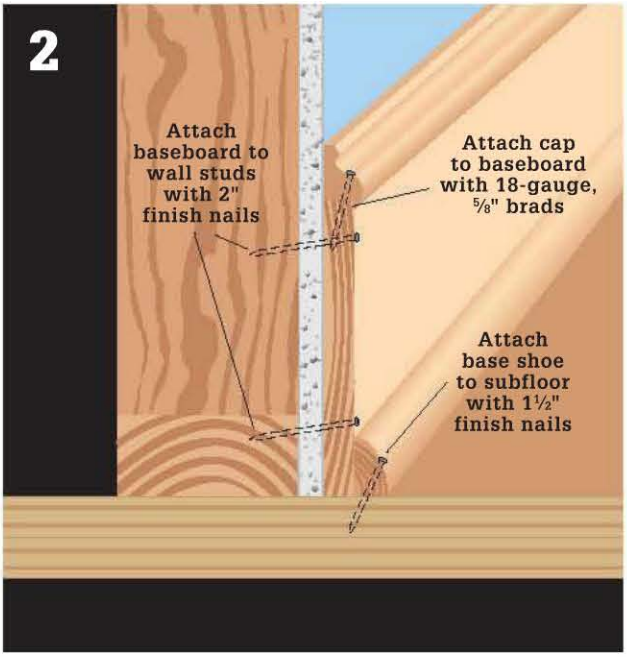
2
Attach baseboard to wall studs with 2" finish nails
Attach cap to baseboard with 18-gauge, ⅝" brads
Attach base shoe to subfloor with 1½" finish nails

Built-up baseboard requires more attention to the nailing schedule than simple one-piece baseboards. The most important consideration (other than making sure your nails are all driven into studs or other solid wood) is that the base shoe must be attached to the floor, while the baseboard is attached to the wall. This way, as the gap between the wall and floor changes, the parts of the built-up molding can change with them.

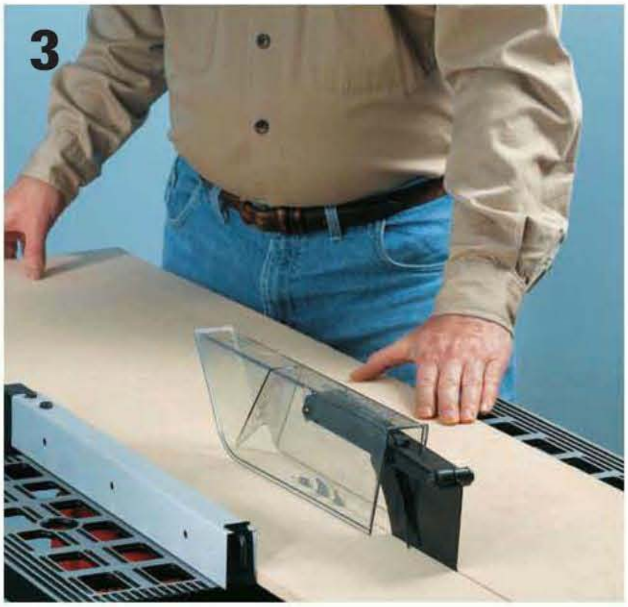

Cut the plywood panel into 6" strips with a table saw or a straightedge guide and a circular saw. Lightly sand the strips, removing any splinters left from the saw. Then, apply the finish of your choice to the moldings and the plywood strips.

Tip ▸

Baseboard can be built up on the back with spacer strips so it will project further out from the wall. This can allow you to match existing casings or to create the impression of a thicker molding. However, the cap rail needs to be thick enough to cover the plywood edge completely, or the core of the panel may be visible.

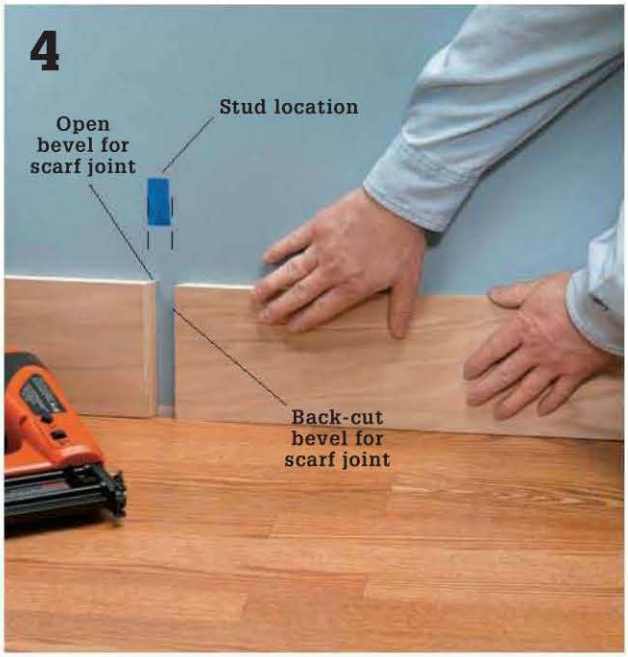
4
Open bevel for scarf joint
Stud location
Back-cut bevel for scarf joint

Install the plywood strips with 2" finish nails driven at stud locations. Use scarf joints on continuous runs, driving pairs of fasteners into the joints. Cut and install moldings so that all scarf joints fall at stud locations.

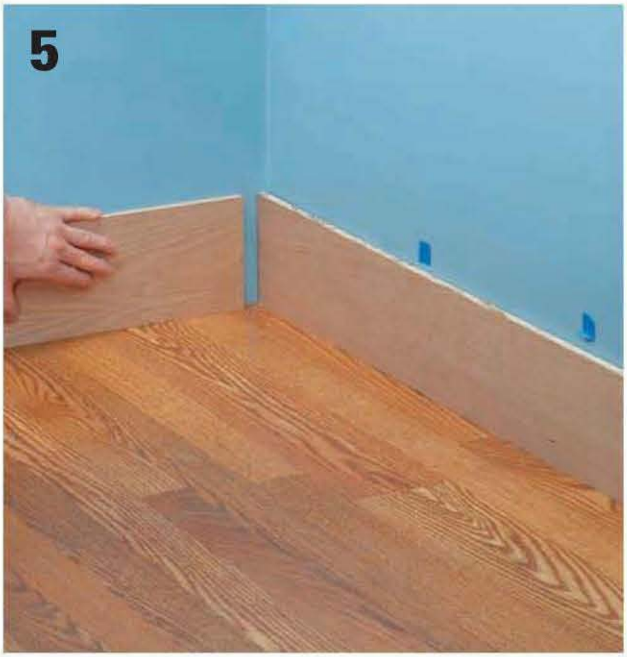

Test-fit inside corner butt joints before cutting a workpiece. If the walls are not square or straight, angle or bevel the end cut a few degrees to fit the profile of the adjoining piece. The cap molding will cover any gaps at the top of the joint.

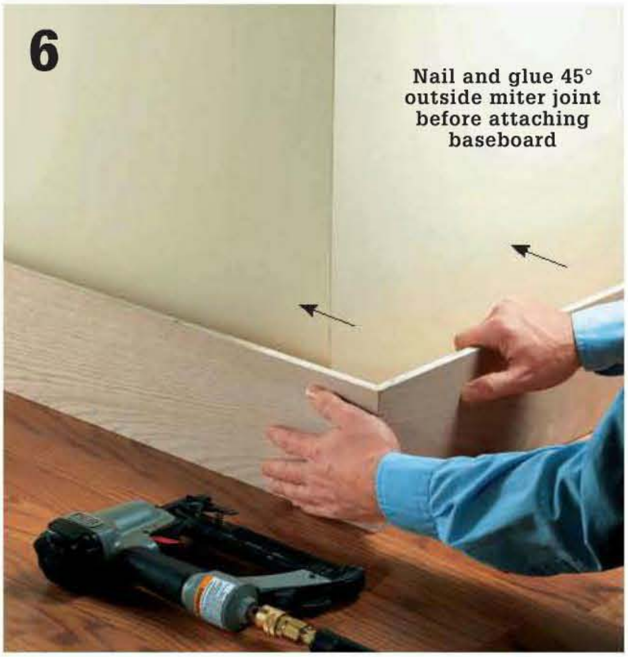
6
Nail and glue 45° outside miter joint before attaching baseboard

Miter outside corners squarely at 45°. Use wood glue and 1¼" brad nails to pull the mitered pieces tight, and then nail the base to the wall at stud locations with 2" finish nails. Small gaps at the bottom or top of the base molding will be covered with cap or base shoe.

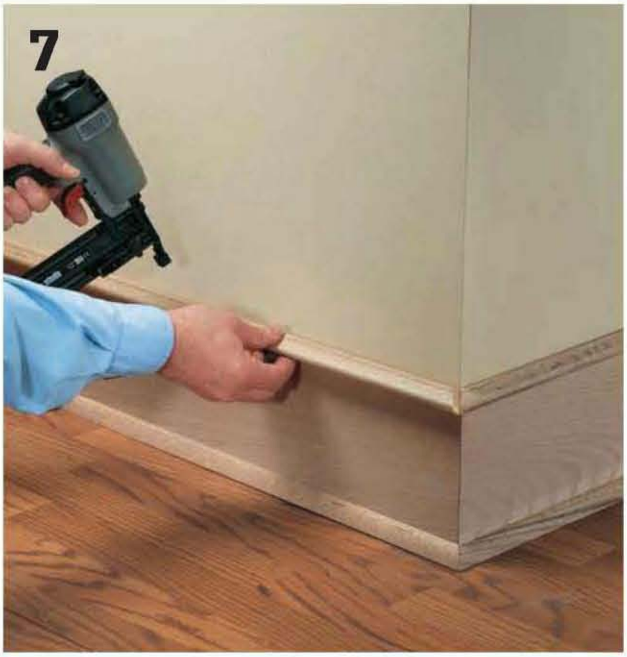

Use a brad nailer with 18-gauge, ⅝" brads to install the cap and base shoe moldings along the edges of the plywood base. Fit scarf joints on longer lengths, coped joints on inside corners, and miter joints on outside corners. Stagger the seams so that they do not line up with the base molding seams, following the suggested nailing pattern (right). Set any protruding nails with a nail set and fill all nail holes with putty.

Installing Picture Rail

Picture rail molding is a specialty molding that was installed in many older homes so the homeowners could avoid making nail holes in the finished walls. Picture rail molding is a simple but elegant way to add style to any room. Special picture hanging hooks slide over the molding and artwork may be hung with a cord over the hook. Picture rail molding also provides its own decorative touch, breaking up the vertical lines from floor to ceiling. For this reason, it is also installed as a decorative touch by itself.

Picture rail molding is easy to install but should be reinforced with screws, not brads or nails, especially if you are hanging large, heavy items. Depending upon the style of your home, picture rail can be hung anywhere from 1 ft. to a few inches down from the ceiling. In some homes, picture rail is added just below the cornice or crown molding to add an additional layer of depth. When applied this way, it is commonly referred to as a frieze board.

In the example shown, the picture rail is installed using a level line to maintain height. If your ceiling is uneven, you may choose to install picture rail a constant distance from the ceiling to avoid an uneven appearance.

Traditionally, picture rail (see top of photo) is used in conjunction with special hooks to hang artwork in formal rooms without penetrating the wall with fasteners.

Tools & Materials ▸

- Ladder
- Pencil
- Stud finder
- Tape measure
- Power miter saw
- T-bevel
- Pneumatic finish nail gun & compressor
- 4-ft. level or laser level
- Drill with bits
- Painter's tape
- Moldings
- Pneumatic fasteners
- $1\frac{5}{8}$" wallboard screws
- Hole filler

How to Install Picture Rail Molding

1

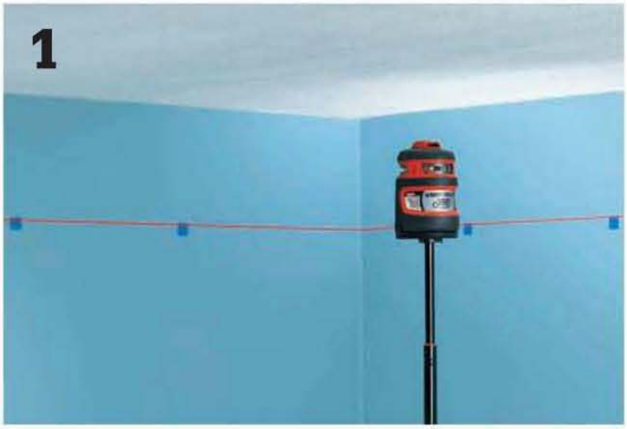

Measure down the desired distance from the ceiling and draw a level reference line around the room using a pencil and a 4-ft. level (or, take advantage of modern technology and use a laser level). While you are up there, use a stud finder to locate the framing members, and mark the locations on the walls with blue painter's tape.

2

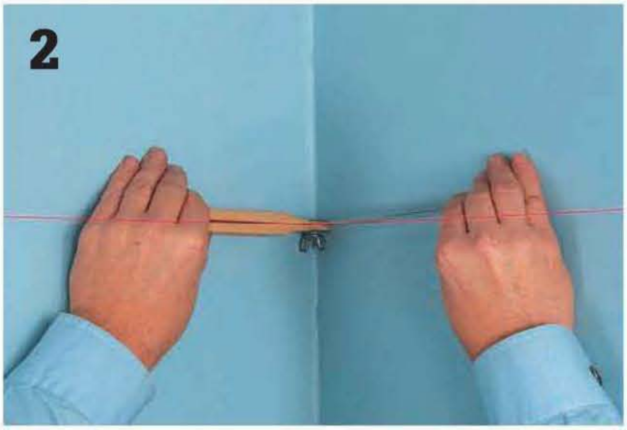

Most corners are close to 90°, but to cut a tight inside corner, the actual angle must be divided exactly in half. Use a T-bevel to measure the angle of the corner, tightening the lock nut with the blade and the handle on the reference line.

3

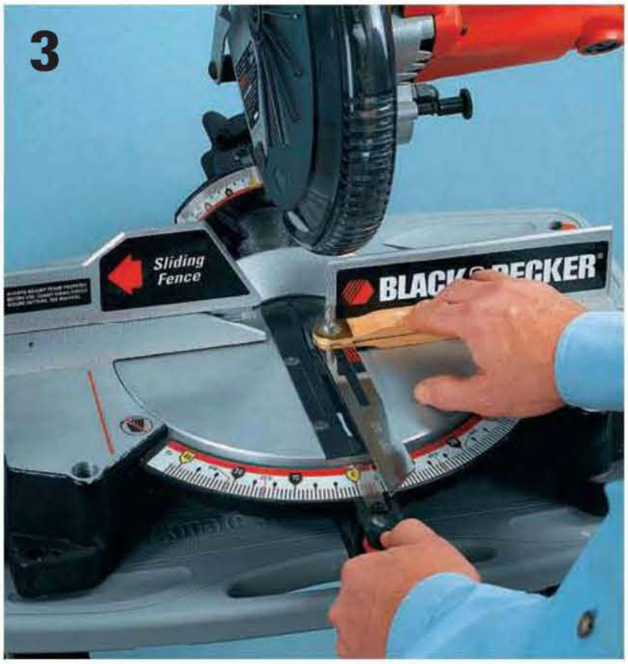

Place the T-bevel on the table of your power miter saw and adjust the miter blade so that it matches the angle. With the T-bevel tight to the fence, read the angle the saw is set to when it aligns with the T-bevel. If the blade is angled to the right of zero degrees the angle is larger than 90; to the left, smaller.

4

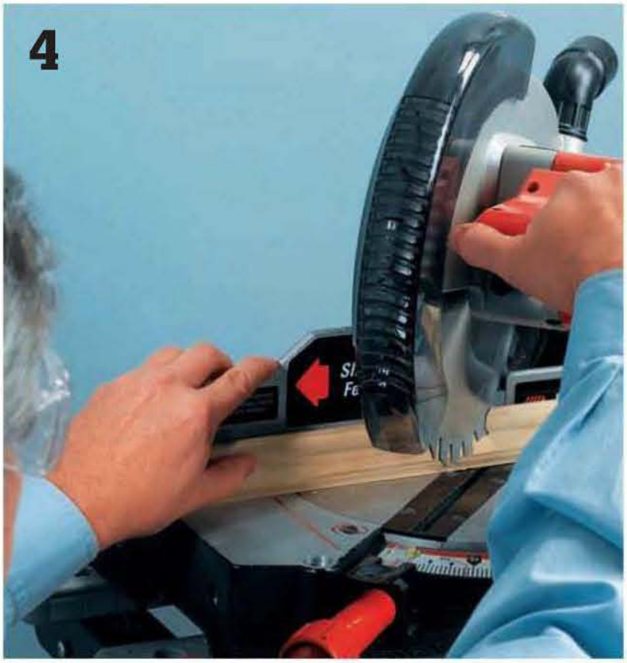

Read the angle from the miter saw table, divide the number by two, and add or subtract that number from 45 degrees to find the proper cutting angle for each corner. Cut each molding slightly longer than the measured length.

5

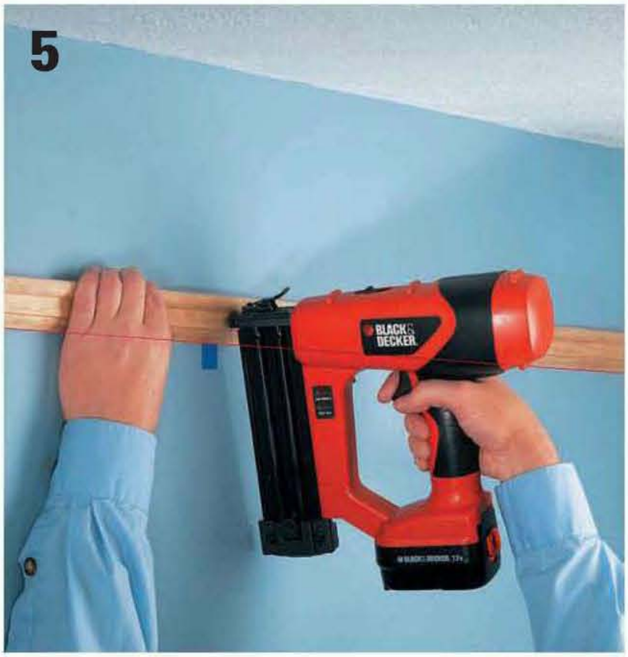

Nail the molding at the stud locations, covering the level line around the room (if you're using a laser level, you simply keep it in position and turned on to cast a reference line you can follow). After each molding is completely nailed in place, go back to each stud location and drive 1⅝" wallboard screws into the molding through counter-bored pilot holes.

6

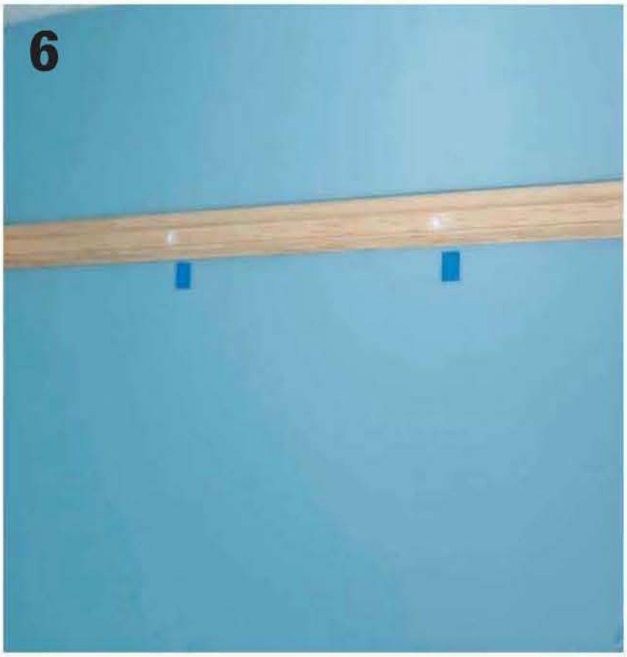

Fill nail holes with wood filler. Let the filler dry and sand it smooth. Then apply a final coat of paint over the molding face.

Installing Chair Rail

Chair rail molding typically runs horizontally along walls at a height of around 36" (the rule of thumb is to install it one-third of the way up the wall). Originally installed to protect walls from collisions with chair backs, today chair rail is commonly used to divide a wall visually. Chair rail may cap wainscoting, serve as a border for wallpaper, or divide two different colors on a wall. Interesting chair rail profiles can be effective alone on a one-color wall.

Stock chair rail moldings are available at most lumberyards and home centers. However, more intricate and elaborate chair rails can be created by combining multiple pieces of trim. Keep in mind the height of your existing furnishings when installing a chair rail. It would be disappointing to discover that the new molding has a bad visual effect with your couch or chair backs when the project is completed.

Chair rail once was installed to protect fragile walls from chair backs, but today it is mainly installed as a decorative accent that breaks up dull walls visually.

Tools & Materials ▸

Pencil
Stud finder
Tape measure
Power miter saw
4-ft. level
Air compressor
Finish nail gun
Metal file set
Moldings
Pneumatic fasteners
Painter's tape
Carpenter's glue
Finishing putty
Finishing materials

How to Install Chair Rail

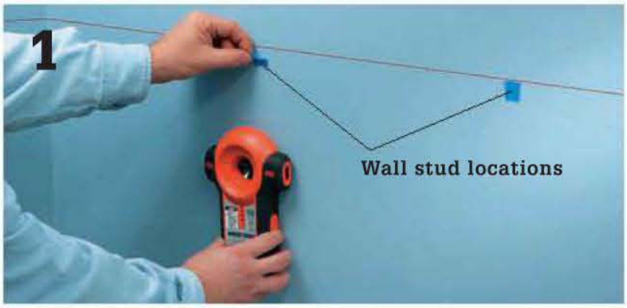

On the starting wall of your installation, measure up to the height at which you plan to install the chair rail, minus the width of the molding. Mark a level line at this height around the room. Locate all studs along the walls and mark their locations with painter's tape below the line.

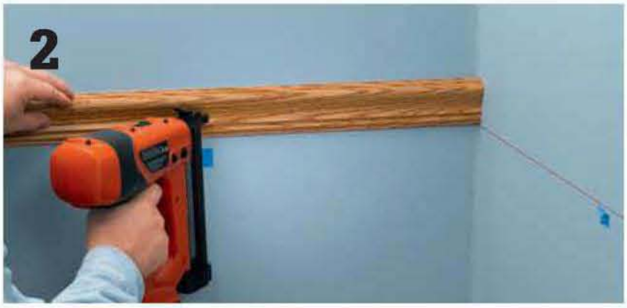

Measure, cut, and install the first piece of chair rail with the ends cut squarely, butting into both walls (in a wall run with two inside corners). Nail the molding in place with two 2" finish nails at each stud location.

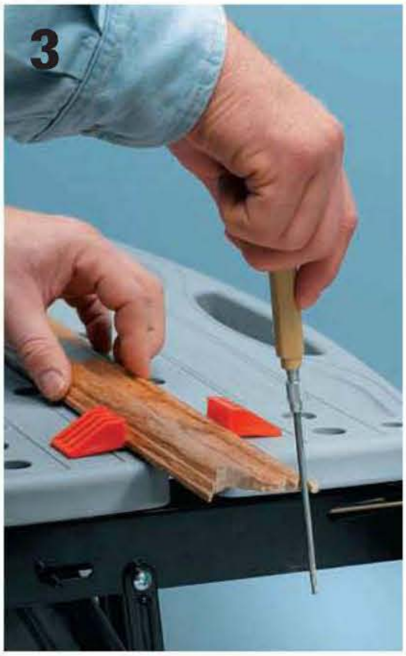

Miter-cut the second piece of molding with a power miter saw and then cope the end with a coping saw. Clean up the edge of the cope cut with a metal file to ensure a tight fit. Dry-fit the piece to check for any gaps in the joint.

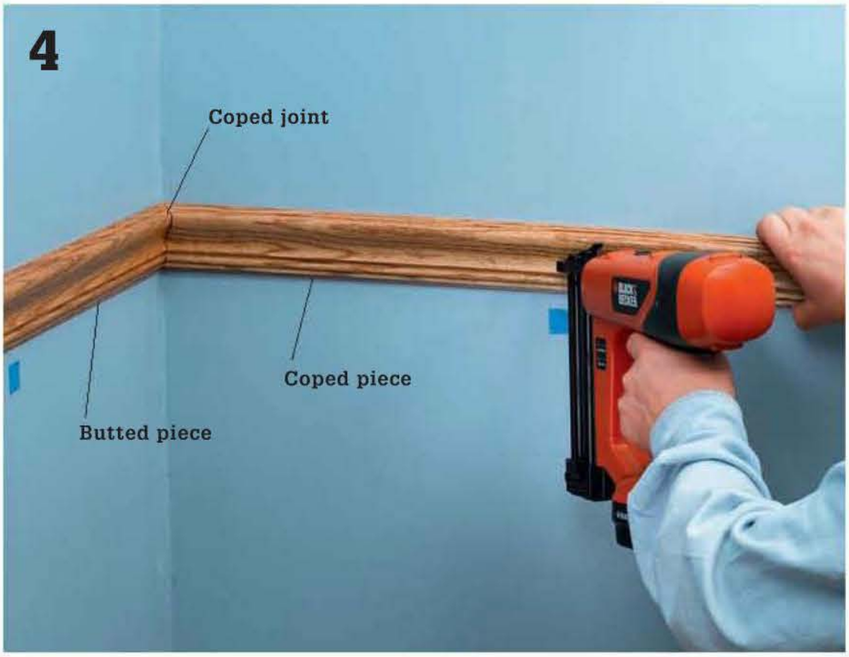

When the coped joint fits tightly, measure, mark, and cut the opposing end of the second piece of trim squarely with a miter saw. Nail the second piece in place with two nails at each stud location. Follow the level line with the bottom edge of the molding.

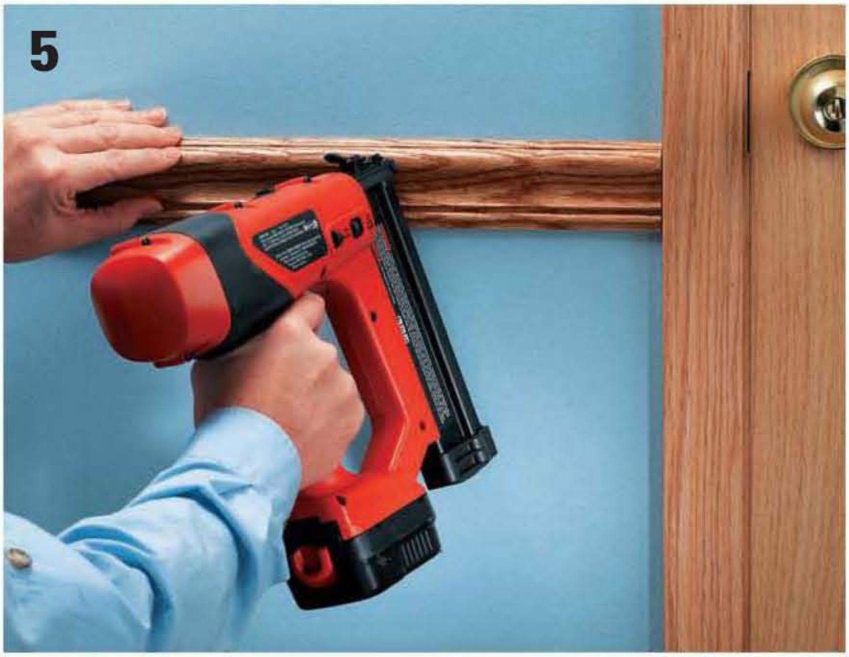

Install the third piece of chair rail with a cope cut at one end. Use a butt joint where the molding runs into door and window casings. Fill all nail holes with putty and apply a final coat of finish to the molding.

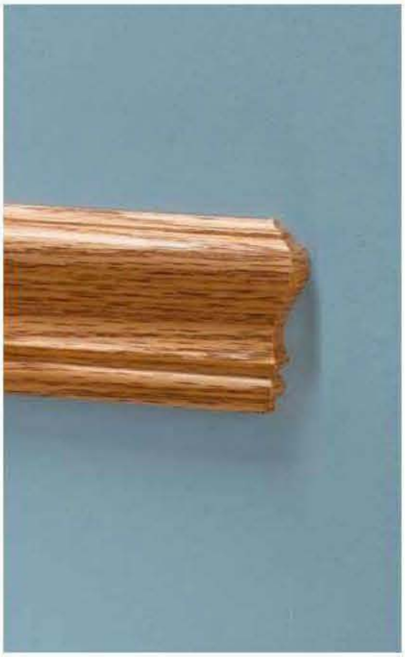

Option: Cut a mitered return for the chair rail in areas where it will end without joining into another molding. Cut the return with a miter saw and glue it in place, using painter's tape to hold it until the glue dries.

Installing Crown Molding

Simply put, crown molding is angled trim that bridges the joint between the ceiling and the wall. In order to cover this joint effectively, crown moldings are "sprung." This means that the top and bottom edges of the molding have been beveled, so when the molding is tilted away from the wall at an angle the tops and bottoms are flush on the wall and ceiling surfaces. Some crown moldings have a 45° angle at both the top and the bottom edges; another common style ("38° crown") has a 38° angle on one edge and a 52° angle on the other edge.

Installing crown molding can be a challenging and sometimes confusing process. Joints may be difficult for you to visualize before cutting, and wall and ceiling irregularities can be hard to overcome. If you have not worked on crown molding joints before, it is recommended that your first attempt be made with paint-grade materials. Stain-grade crown is commonly made of solid hardwood stock, which makes for expensive cutting errors, and difficulty concealing irregularities in joints.

Inside corner joints of crown molding should be cope cut, not mitered, except in the case of very intricate profile crown that is virtually impossible to cope (and must therefore be mitered). While mitering inside corners may appear to save time and produce adequate results, after a few changing seasons the joints will open up and be even more difficult to conceal.

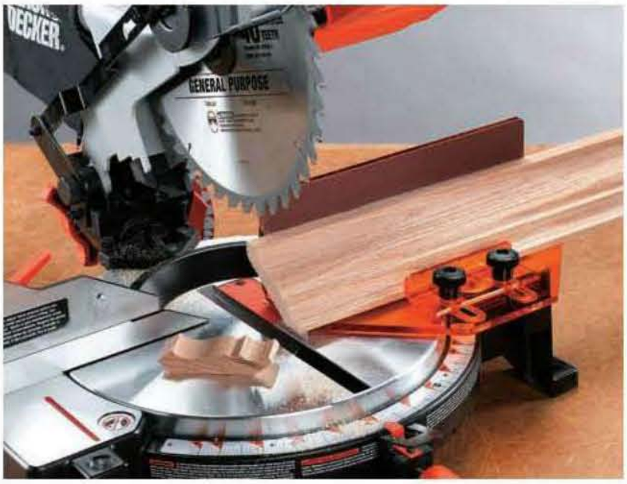

Cutting compound miters is tricky. Here, crown molding is cut with the workpiece held against a fence or fence extension. This hand-held approach is quick and effective, but takes some getting used to. A practically foolproof option is to use an adjustable jig, such as the compound miter jig shown here.

Tools & Materials ▸

- Pencil
- Tape measure
- Circular saw
- Straightedge guide
- Drill with bits
- Coping saw
- Power miter saw
- Pneumatic finish nail gun
- Framing square or combination square
- Nail set
- Hammer
- Metal files
- 2 × 4 material for backing
- 3" wallboard screws
- Carpenter's glue
- Crown molding
- 2", 1½" finish nails
- Fine-grit sandpaper
- Hole filler
- Paint and brushes

Basic crown molding softens the transitions between walls and ceilings. If it is made from quality hardwood crown molding can be quite beautiful when installed and finished with a clear topcoat. But historically, it is most often painted—either the same color as the ceiling (your eye tends to see it as a ceiling molding, not a wall molding) or with highly elaborate painted and carved details.

How to Install Basic Crown Molding

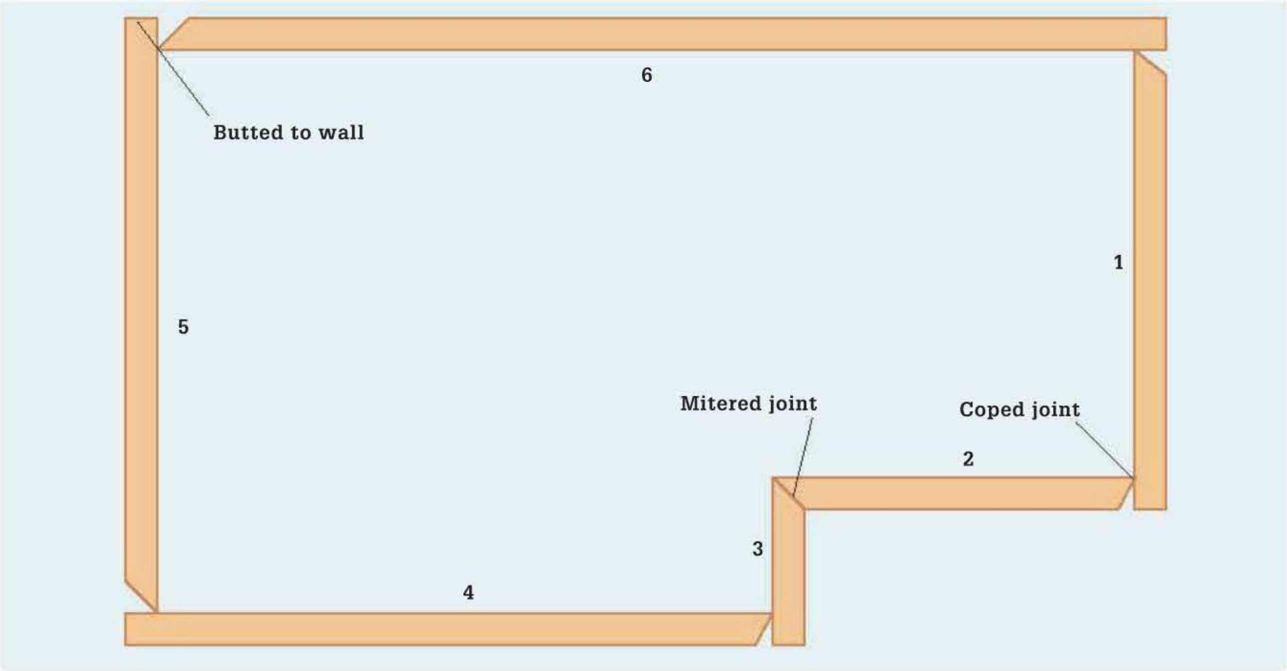

Plan the order of the installation to minimize the number of difficult joints on each piece and use the longest pieces for the most visible sections of wall. Notice that the left end of first piece is cope-cut rather than butted into the wall. Cope-cutting the first end eliminates the need to cope-cut both ends of the final piece, and places the cuts in the same direction. This simplifies your installation, making the method to cut each piece similar.

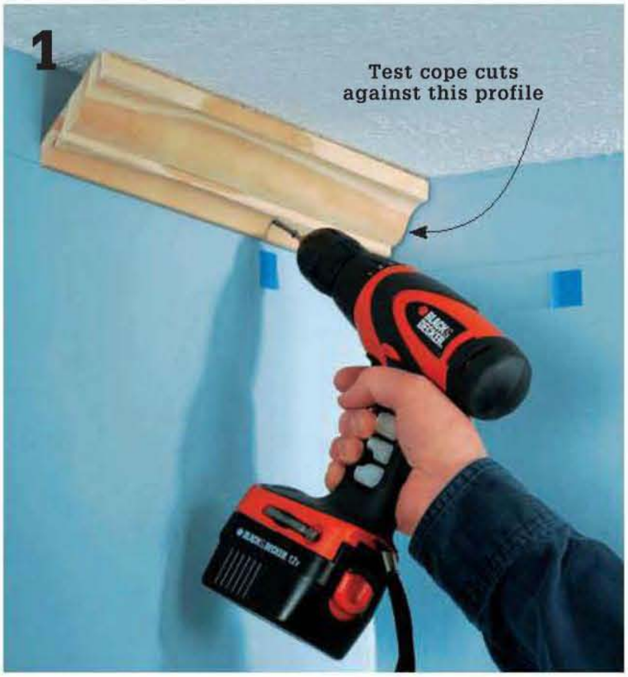

Cut a piece of crown molding about 1-ft. long with square ends. Temporarily install the piece in the corner of the last installation wall with two screws driven into the blocking. This piece serves as a template for the first cope cut on the first piece of molding.

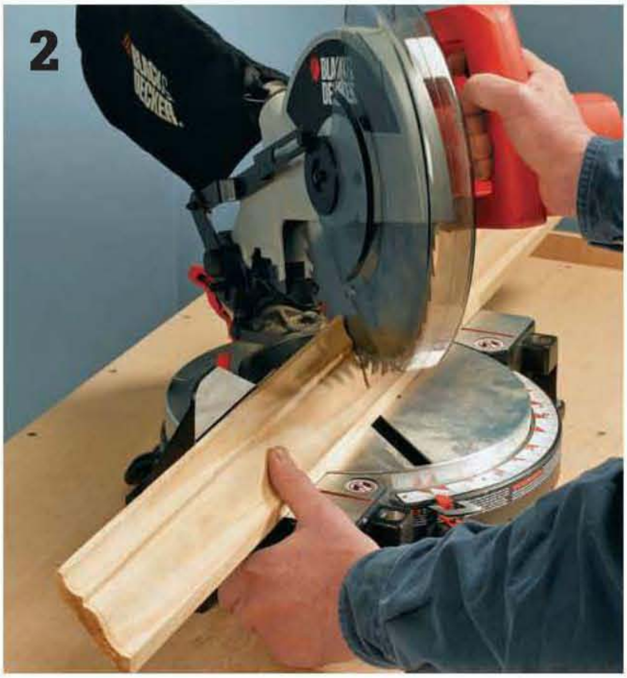

Place the first piece of molding upside down and sprung against the fence of the miter saw. Mark a reference line on the fence for placement of future moldings, and cut the first coped end with an inside miter cut to reveal the profile of the piece.

(continued)

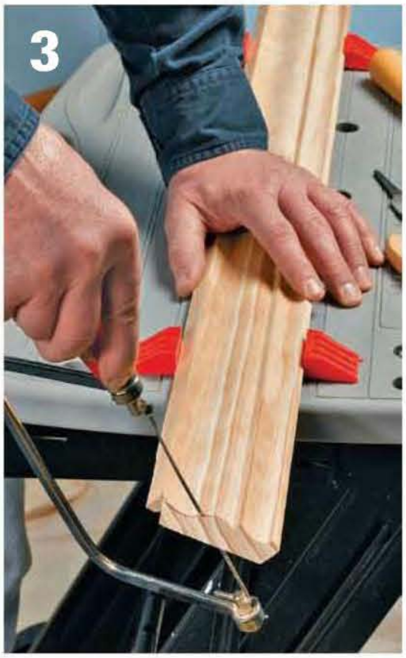

Cope-cut the end of the first piece with a coping saw. Carefully cut along the profile, angling the saw as you cut to back-bevel the cope. Test-fit the coped cut against the temporary scrap from Step 1. Fine-tune the cut with files and fine-grit sandpaper.

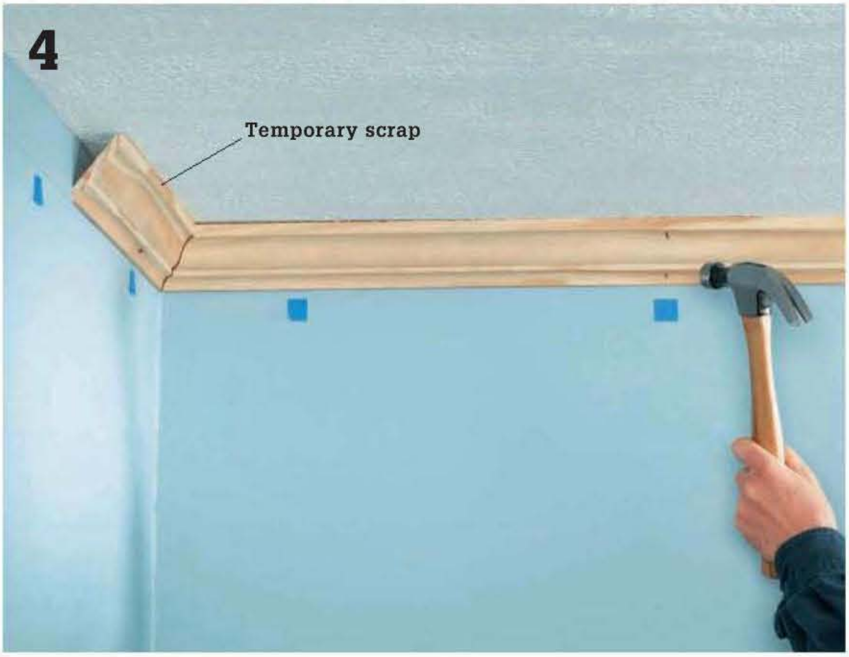

Measure, cut to length, and install the first piece of crown molding, leaving the end near the temporary scrap loose for final fitting of the last piece. Nail the molding at the top and bottom of each stud location.

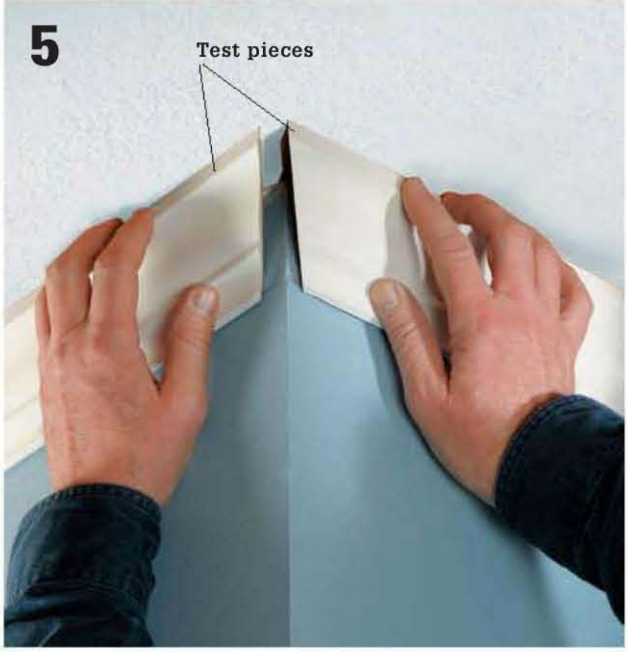

Cut two test pieces to check the fit of outside corners. Start with each molding cut at 45°, adjusting the angles larger or smaller until the joints are tight. Make sure the test moldings are properly aligned and are flush with the ceiling and walls. Make a note of your saw settings once the joint fits tightly.

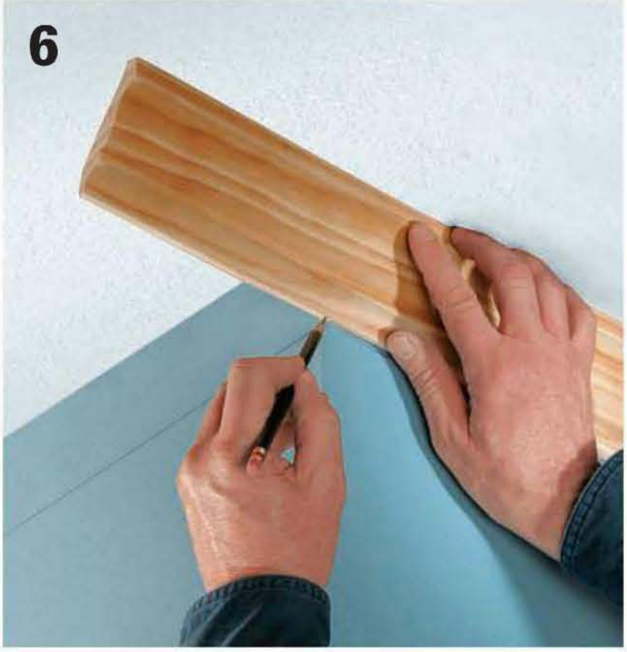

Position the actual stock so a cut end is flush against the wall at one end and, at the other end, mark the outside corner on the back edge of the molding. Miter-cut the piece at the mark, according to the angles you noted on the test pieces.

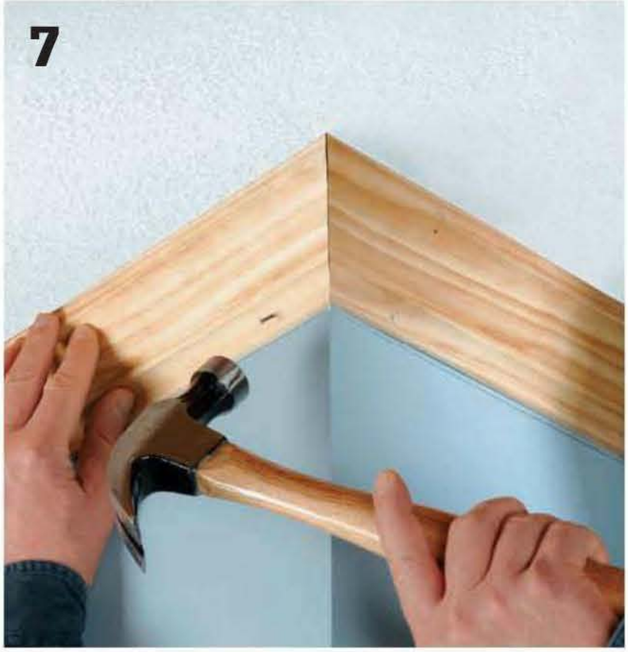

Measure and cut the third piece with an outside corner miter to match the angle of your test pieces. Cut the other end squarely, butting it into the corner. Install the piece with nails driven at stud locations. Install the subsequent pieces of crown molding, coping the front end and butting the other as you work around the room.

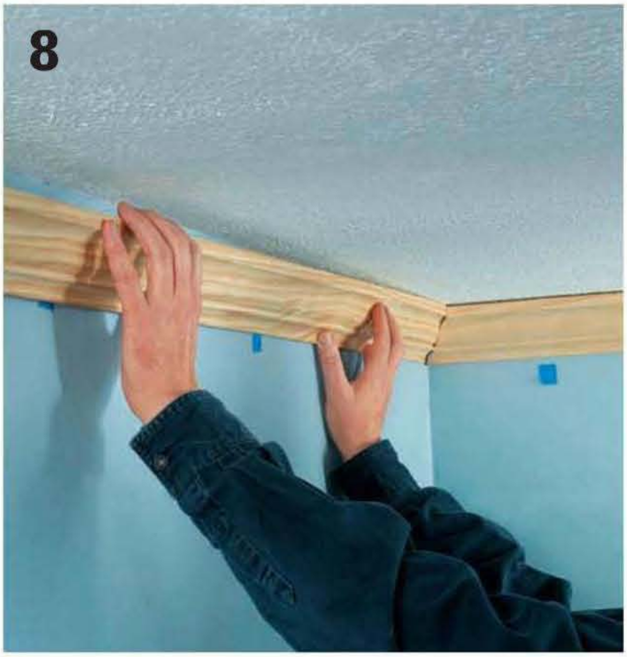

To fit the final piece, cope the end and cut it to length. Remove the temporary scrap piece from Step 3, and slide the last molding into position. Nail the last piece at the stud locations when the joints fit well, and finish nailing the first piece.

Fill all nail holes (use spackling compound if painting; wait until the finish is applied and fill with tinted putty for clear finishes). Use a putty knife to force spackling compound or tinted wood putty into loose joints and caulk gaps ⅛" or smaller between the molding and the wall or ceiling with flexible, paintable, latex caulk.

Lightly sand the filled nail holes and joint gaps using fine sandpaper. Sand the nail holes flush with the surface of the moldings and paint the entire project.

ELEVATION
1/2"=1'-0"
ELEVATION
1/2"=1'-0"
3-SEASON PORCH
DESIGN

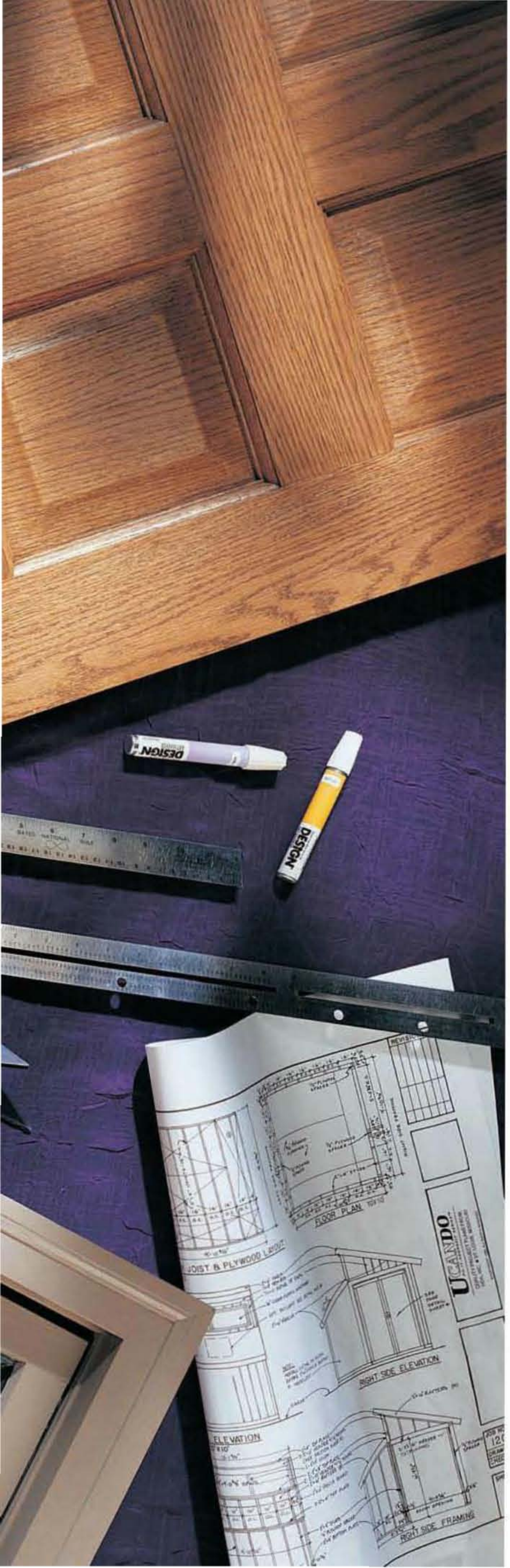

Windows & Doors

Once you've got a feel for the style of windows or doors you'd like to add to your home, the next step involves careful planning. You'll need to consider how a new door or window meets your needs for energy efficiency, durability, ease of maintenance, and security.

There's also the equally important matter of how a new window or door will impact your home's livability. Will it face a direction that takes advantage of sunlight and prevailing breezes or make the room stifling hot in mid July? Adding a window to a living space may also bring matters of egress into consideration. Building codes mandate that your project window must satisfy these requirements.

In this chapter:

- Framing & Installing Windows
- Installing Garden Windows
- Installing Bay Windows
- Installing Skylights
- Installing Tubular Skylights
- Installing Interior Doors
- Installing Entry Doors
- Installing Storm Doors
- Installing Patio Doors
- Installing Door & Window Casing

Framing & Installing Windows

Many windows must be custom-ordered several weeks in advance. To save time, you can complete the interior framing before the window unit arrives, but be sure you have the exact dimensions of the window unit before building the frame. Do not remove the outside wall surface until you have the window and accessories and are ready to install them.

Follow the manufacturer's specifications for rough opening size when framing for a window. The listed opening is usually 1" wider and ½" taller than the actual dimensions of the window unit. The following pages show techniques for wood-frame houses with platform framing.

If your house has balloon framing where wall studs pass continuously from one floor to the next, use the method shown on page 271 to install a header. Consult a professional to install a window on the second story of a balloon-framed house.

If your home's exterior has siding or is stucco, see pages 34 to 35 for tips on removing these surfaces and making the opening.

Correct framing techniques will ensure ease of installation and keep your windows operating smoothly.

Tools & Materials ▸

- Tape measure
- Pencil
- Combination square
- Hammer
- Level
- Circular saw
- Handsaw
- Pry bar
- Nippers
- Drill
- Reciprocating saw
- Stapler
- Nail set
- Caulk gun
- 10d common nails,
- 1" galvanized roofing nails
- Shims
- 2× lumber
- ⅛" plywood
- Building paper
- Drip edge
- 10d galvanized casing nails
- 8d casing nails
- Fiberglass insulation
- Paintable silicone caulk

How to Frame a Window Opening

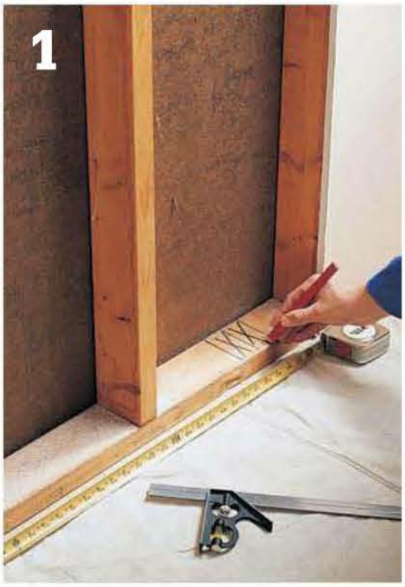

Prepare the project site and remove the interior wall surfaces (pages 30 to 31). Measure and mark the rough opening width on the sole plate. Mark the locations of the jack studs and king studs on the sole plate. Where practical, use the existing studs as king studs.

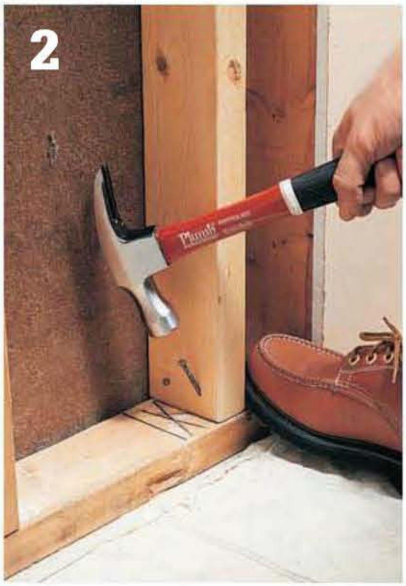

Measure and cut the king studs, as needed, to fit between the sole plate and the top plate. Position the king studs and toenail them to the sole plate with 10d nails.

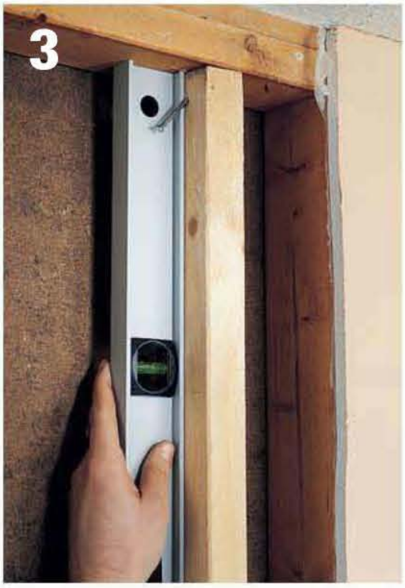

Check the king studs with a level to make sure they are plumb, then toenail them to the top plate with 10d nails.

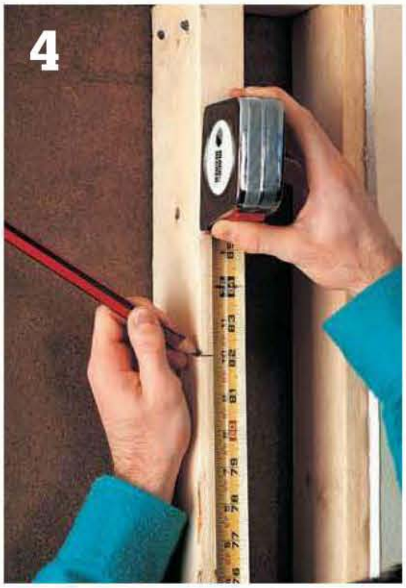

Measuring from the floor, mark the top of the rough opening on one of the king studs. This line represents the bottom of the window header. For most windows, the recommended rough opening is ½" taller than the height of the window frame.

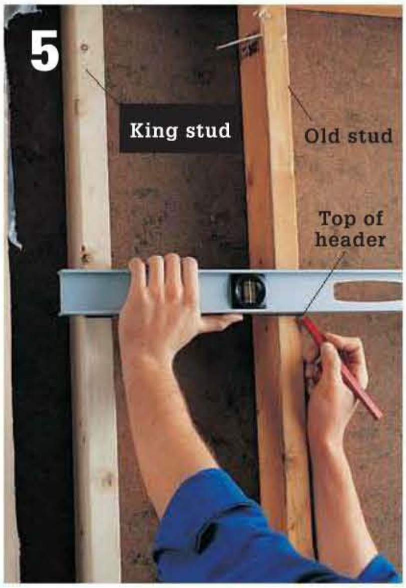

Measure and mark where the top of the window header will fit against the king studs. The header size depends on the distance between the king studs. Use a carpenter's level to extend the lines across the old studs to the opposite king stud.

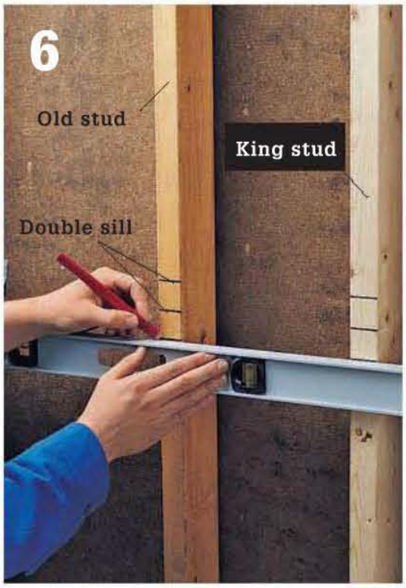

Measure down from the header line and mark the double rough sill on the king stud. Use a carpenter's level to extend the lines across the old studs to the opposite king stud. Make temporary supports (page 29) if removing more than one stud.

(continued)

7

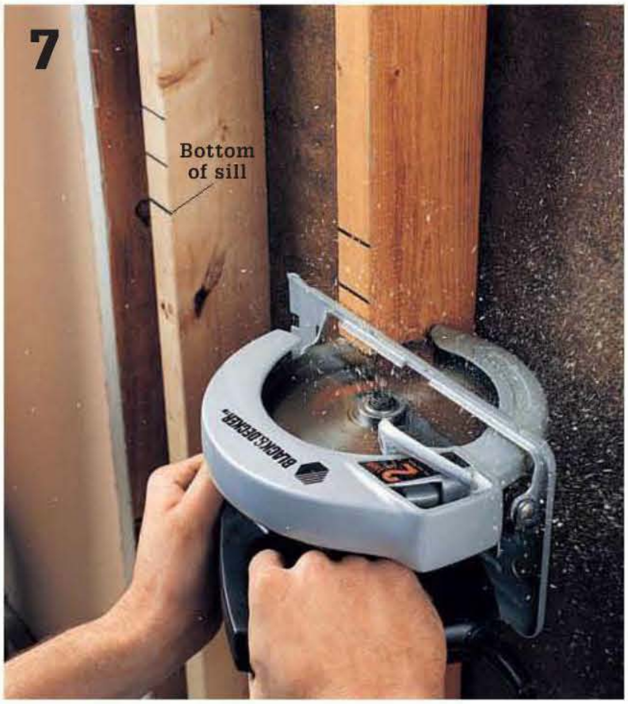
Bottom of sill

Set a circular saw to its maximum blade depth, then cut through the old studs along the lines marking the bottom of the rough sill and along the lines marking the top of the header. Do not cut the king studs. On each stud, make an additional cut about 3" above the sill cut. Finish the cuts with a handsaw.

8

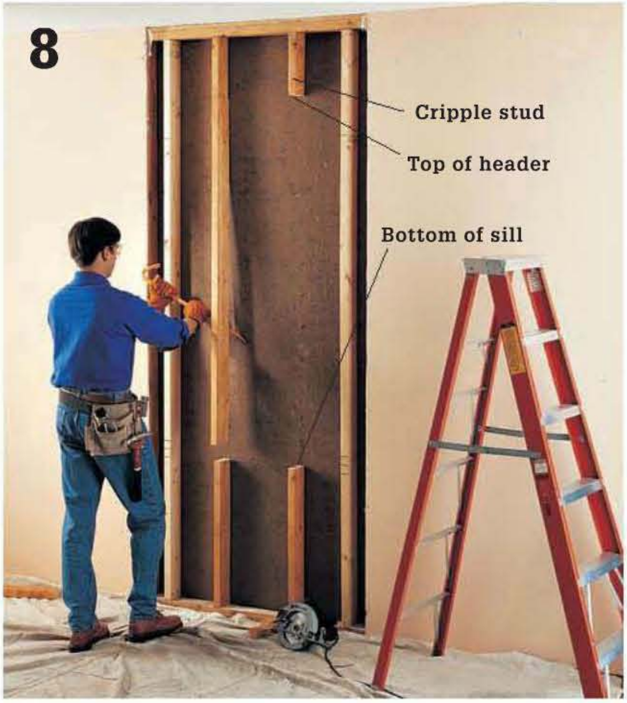
Cripple stud
Top of header
Bottom of sill

Knock out the 3" stud sections, then tear out the old studs inside the rough opening, using a pry bar. Clip away any exposed nails, using nippers. The remaining sections of the cut studs will serve as cripple studs for the window.

9

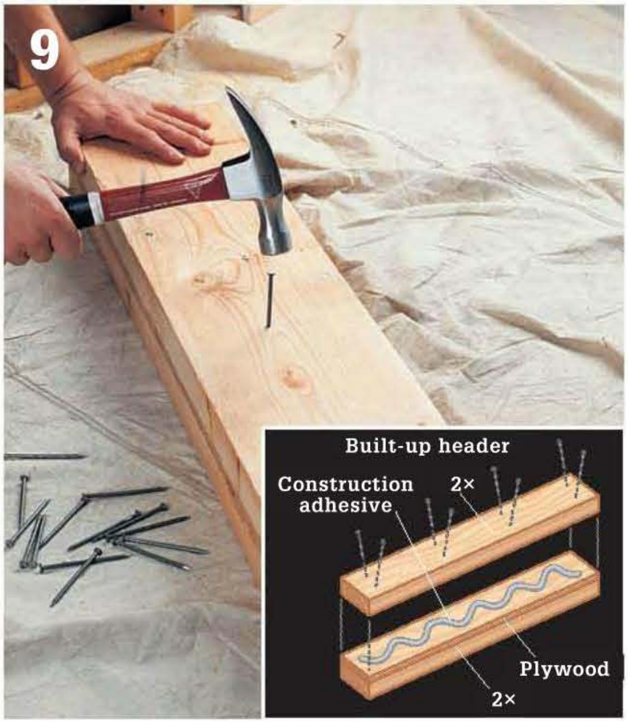
Built-up header
Construction adhesive
2×
Plywood
2×

Build a header to fit between the king studs on top of the jack studs, using two pieces of 2× lumber sandwiched around ½" plywood.

10

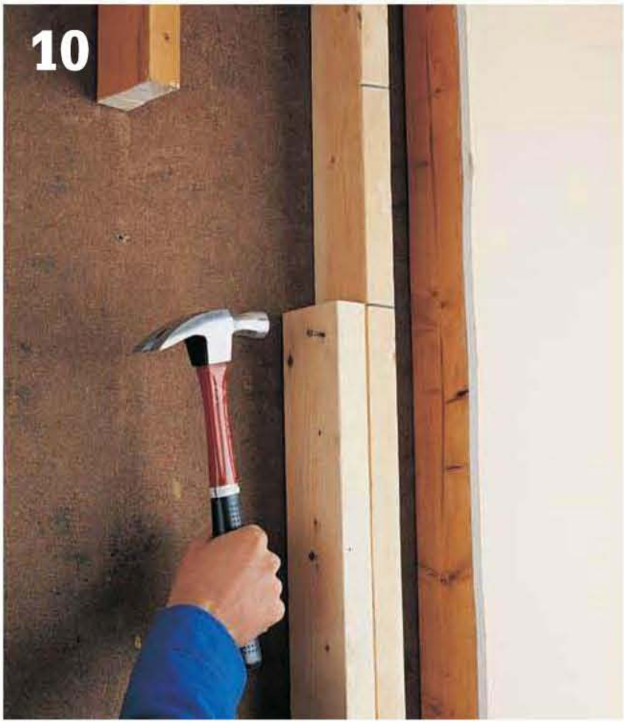

Cut two jack studs to reach from the top of the sole plate to the bottom header lines on the king studs. Nail the jack studs to the king studs with 10d nails driven every 12".

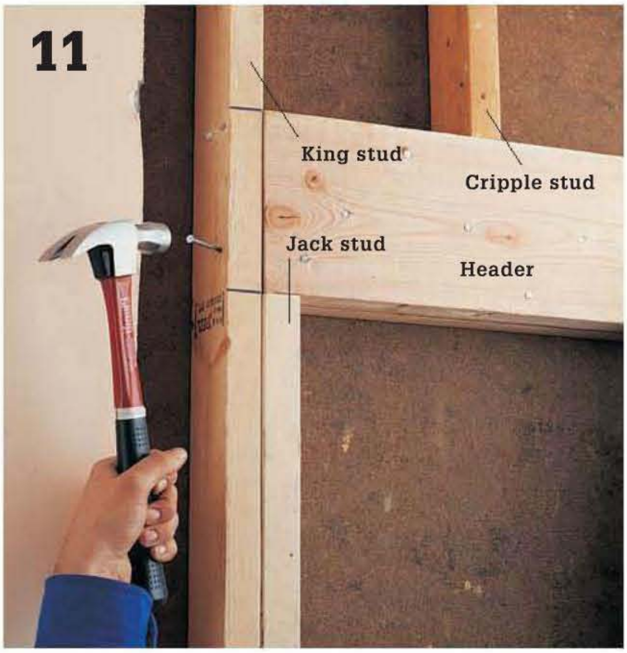

Position the header on the jack studs, using a hammer if necessary. Attach the header to the king studs, jack studs, and cripple studs, using 10d nails.

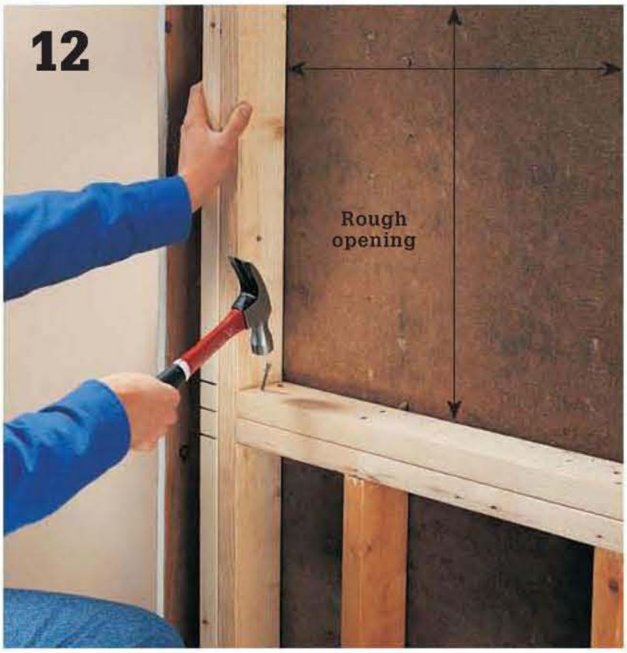

Build the rough sill to reach between the jack studs by nailing a pair of 2 × 4s together. Position the rough sill on the cripple studs, and nail it to the jack studs and cripple studs with 10d nails.

How to Install a Replacement Window with a Nailing Flange

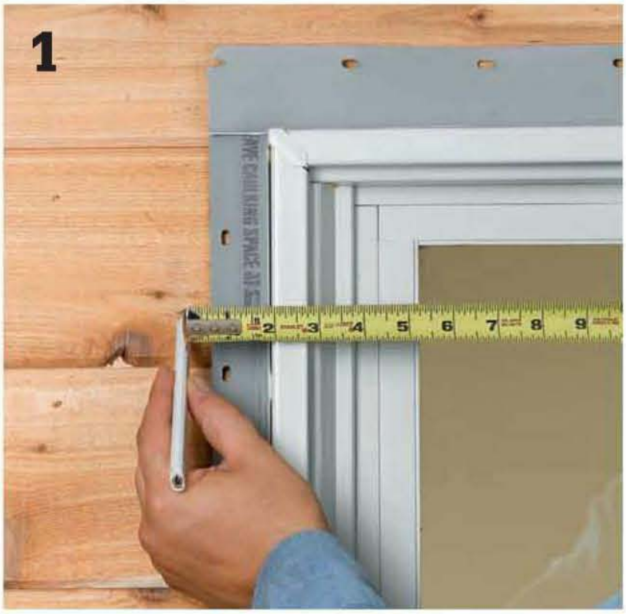

Remove the existing window and set the new window into the rough opening. Center it left to right, and shim beneath the sill to level it. On the exterior side, measure out from the window on all sides, and mark the siding for the width of the brick molding you'll install around the new window. Extend layout lines to mark where you'll cut the siding.

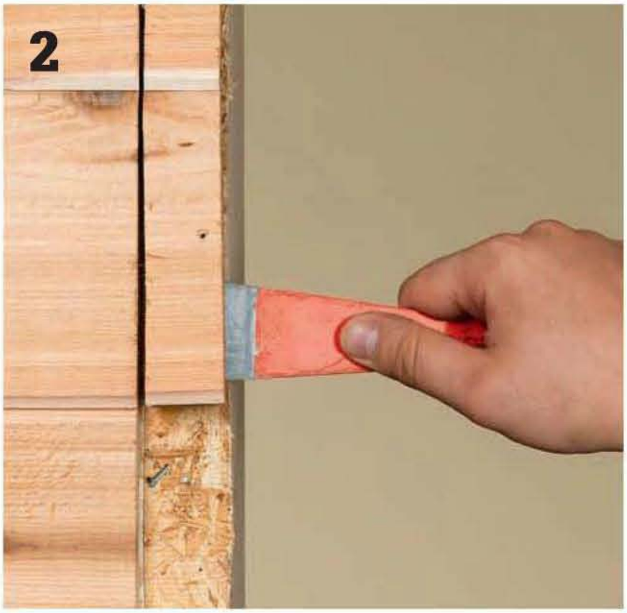

Remove exterior siding around the window area to expose the wall sheathing. Use a zip tool to separate vinyl siding for removal or use a pry bar and hammer to remove wood clapboard. For more on removing exterior surfaces, (see pages 32 to 33).

(continued)

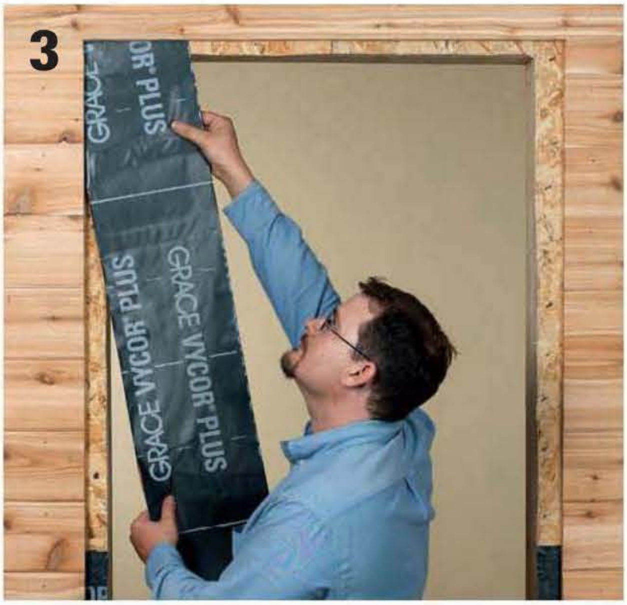

Cover the sill and rough opening framing members with self-adhesive, rolled flashing. Apply additional strips of flashing behind the siding and up the sill flashing. Finish flashing with a strip along the header. The flashing should cover the front edges and sides of the opening members.

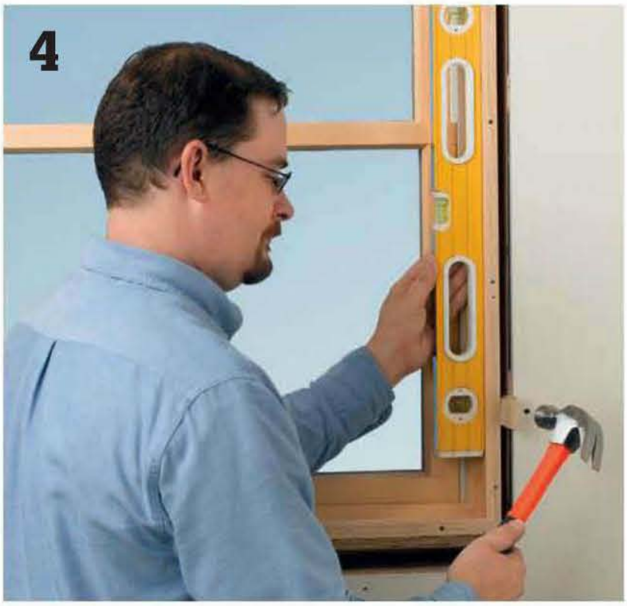

Apply a bead of silicone caulk around the back face of the window flange, then set it into the rough opening, centering it side-to-side in the opening. Tack the window in place by driving one roofing nail partway through the top flange. On the interior side, level and plumb the window, using shims to make any necessary adjustments.

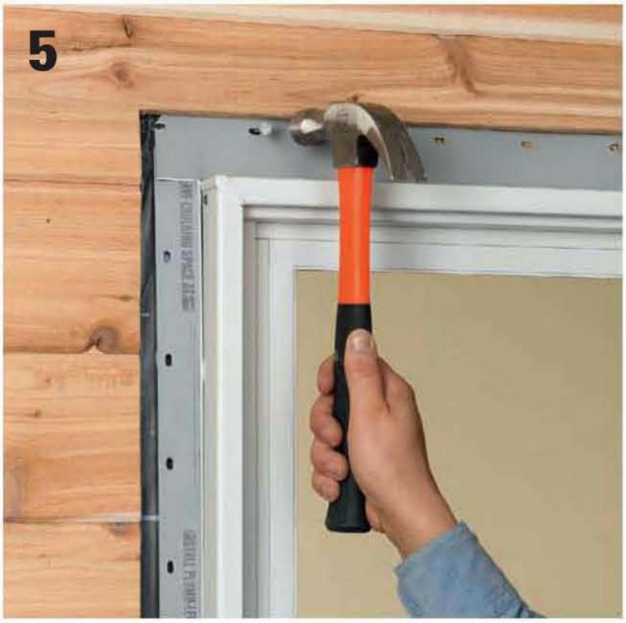

Tack the window to the header at one end of the nailing flange, using a 1" galvanized roofing nail. Drive a roofing nail through the other top corner of the flange to hold the window in place, then secure the flange all around the window with more roofing nails. Apply strips of rolled, self-adhesive flashing to cover the window flanges. Start with a strip that covers the bottom flange, then cover the side flanges, overlapping the bottom flashing and extending 8 to 10" above the window. Complete the flashing with a strip along the top, overlapping the side flashing.

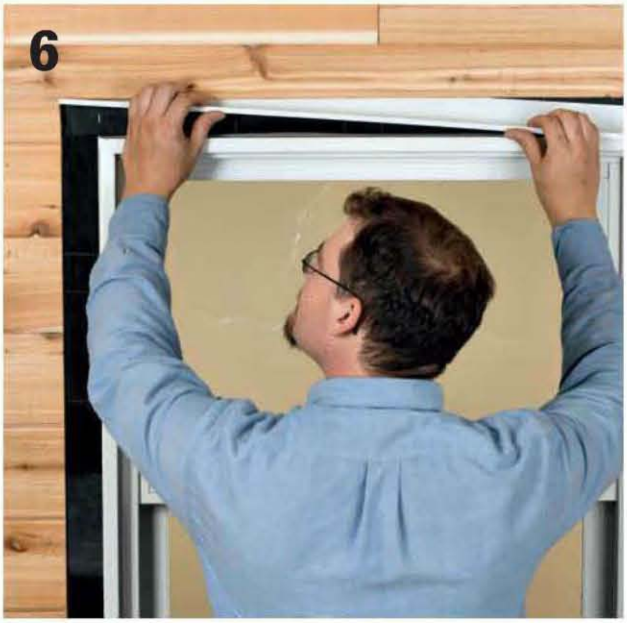

Install a piece of metal drip edge behind the siding and above the window. Secure it with silicone caulk only.

7

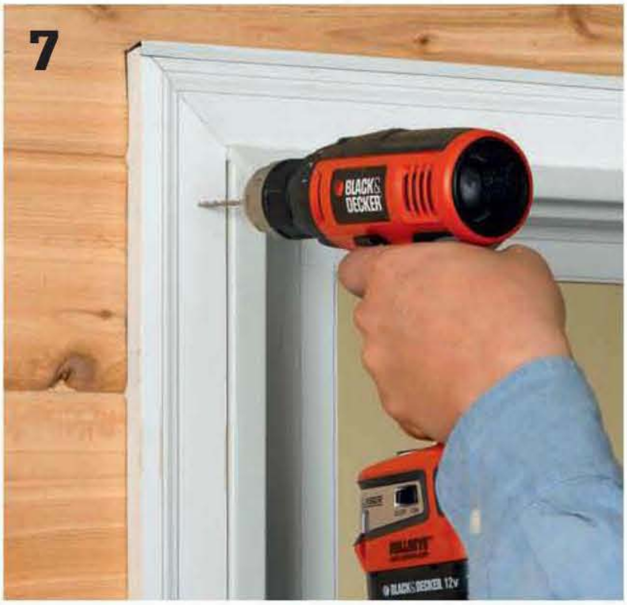

Cut and attach brick molding around the window, leaving a slight gap between the brick molding and the window frame. Use 8d galvanized casing nails driven into pilot holes to secure the brick molding to the rough framing. Miter the corner joints. Reinstall the siding in the window installation area, trimming as needed.

8

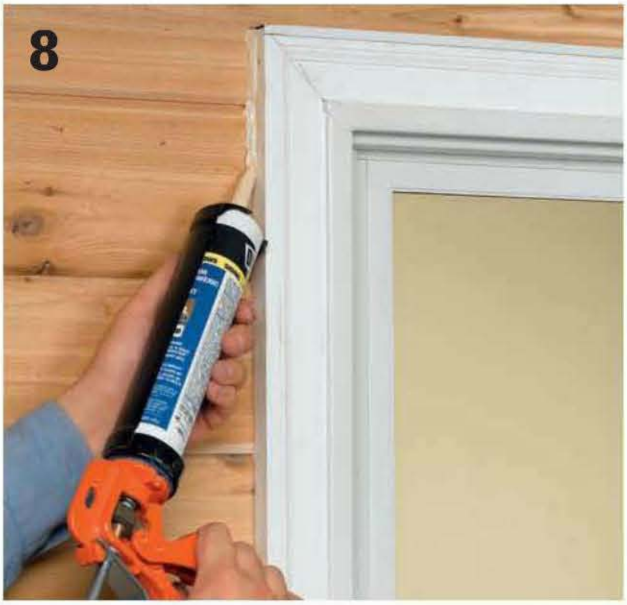

Use high-quality caulk to fill the gap between the brick molding and the siding. On the interior side, fill gaps between the window frame and surrounding framing with foam backer rod, low-expansion foam, or fiberglass insulation. Install the interior casing.

Tip Installation Variation: Masonry Clips ▸

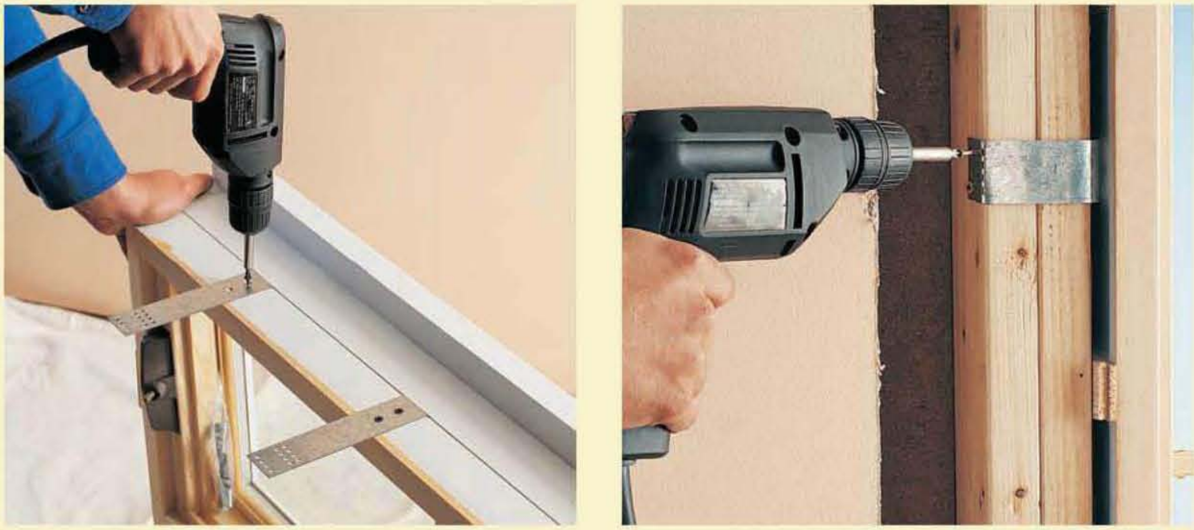

Use metal masonry clips when the brick molding on a window cannot be nailed because it rests against a masonry or brick surface. The masonry clips hook into precut grooves in the window jambs (above, left) and are attached to the jambs with screws. After the window unit is positioned in the rough opening, the masonry clips are bent around the framing members and anchored with screws (above, right). *Note: Masonry clips can also be used in ordinary lap siding installations if you want to avoid making nail holes in the smooth surface of the brick moldings. For example, windows that are precoated with polymer-based paint can be installed with masonry clips so that the brick moldings are not punctured with nails.*

How to Install a Round-Top Window

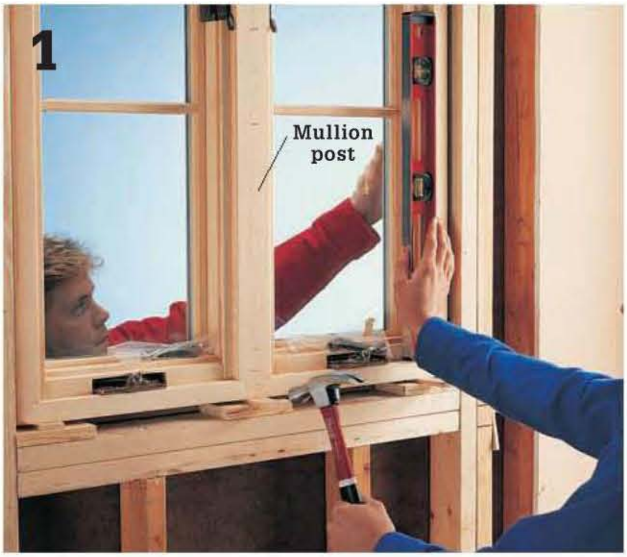

Remove the exterior wall surface as directed on pages 32 to 33, then test-fit the window, centering it in the rough opening. Support the window with wood blocks and shims placed under the side jambs and mullion post. Check to make sure the window is plumb and level, and adjust the shims, if necessary.

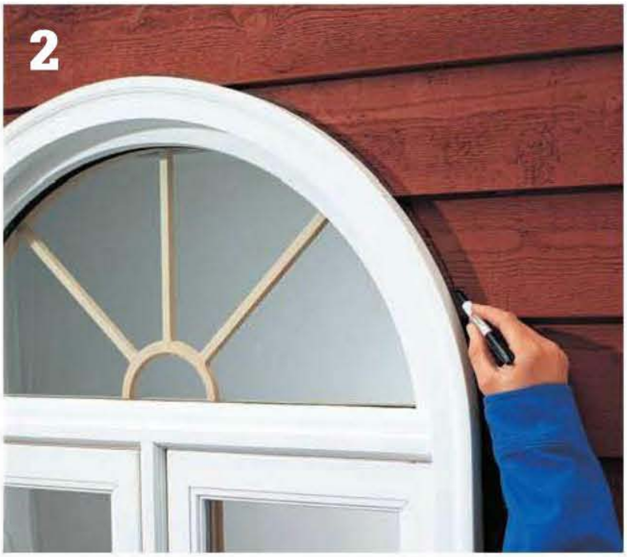

Trace the outline of the brick molding on the wood siding. Remove the window after finishing the outline. *Note: If you have vinyl or metal siding, you should have enlarged the outline to make room for the extra J-channel moldings required by these sidings.*

Tips for Framing a Round-top Window ▸

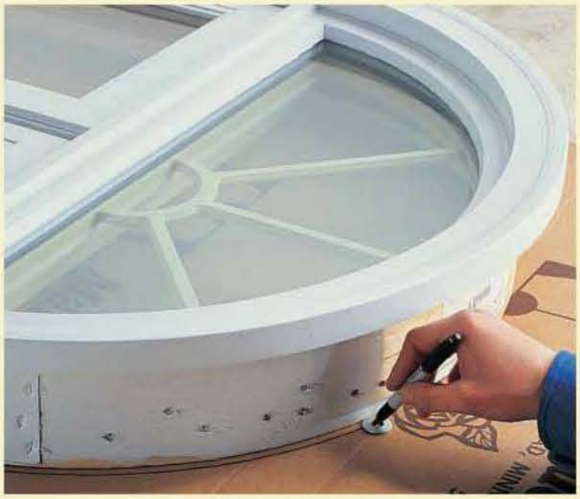

Create a template to help you mark the rough opening on the sheathing. Scribe the outline of the curved frame on cardboard, allowing an extra ½" for adjustments within the rough opening. A ¼ × 1¼" metal washer makes a good spacer for scribing the outline. Cut out the template along the scribed line.

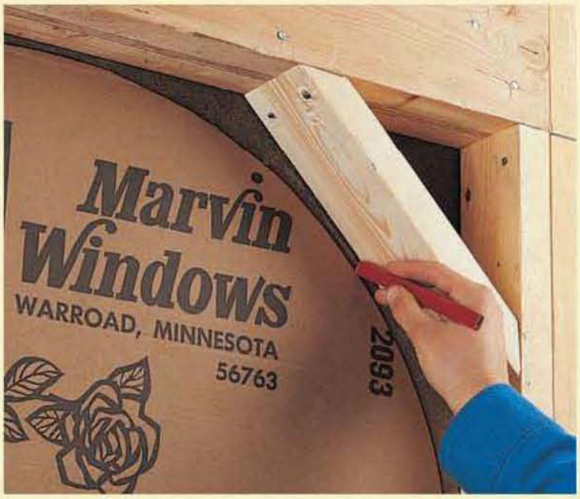

Tape the template to the sheathing, with the top flush against the header. Use the template as a guide for attaching diagonal framing members across the top corners of the framed opening. The diagonal members should just touch the template. Outline the template on the sheathing as a guide for cutting the exterior wall surface.

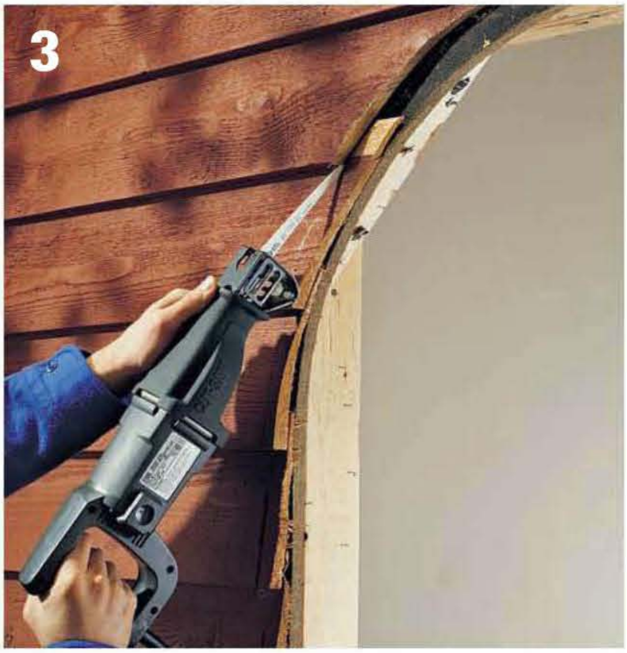

Cut the siding along the outline just down to the sheathing. For a round-top window, use a reciprocating saw held at a low angle. For straight cuts, use a circular saw adjusted so the blade cuts through only the siding. Use a sharp chisel to complete the cuts at the corners.

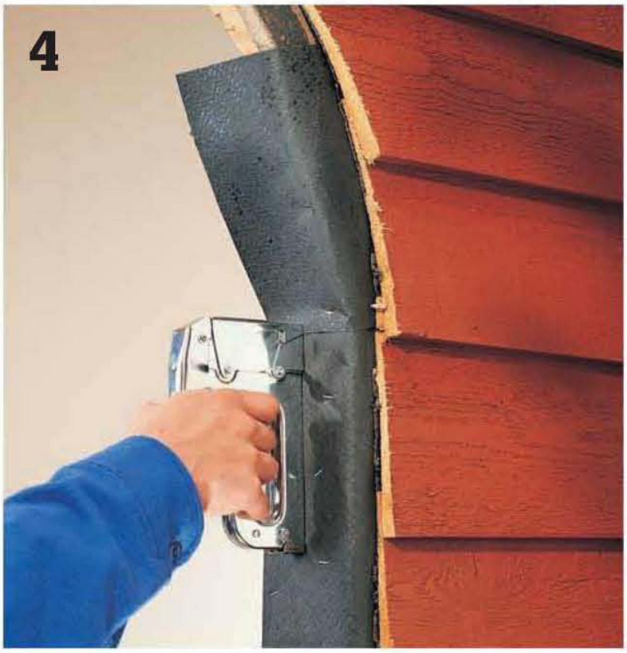

Cut 8"-wide strips of building paper and slide them between the siding and sheathing around the entire window opening. Bend the paper around the framing members and staple it in place. Work from the bottom up, so each piece overlaps the piece below. *Note: You can also use adhesive-backed, rolled flashing instead of building paper.*

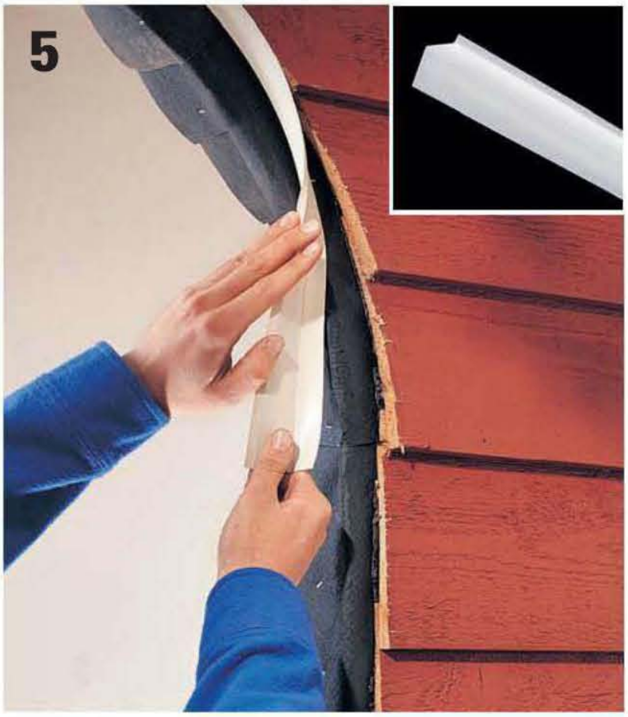

Cut a length of drip edge to fit over the top of the window, then slide it between the siding and building paper. For round-top windows, use flexible vinyl drip edge; for rectangular windows, use rigid metal drip edge (inset).

Insert the window in the opening, and push the brick molding tight against the sheathing. Nail through the brick molding, as usual, to secure the window in the opening.

Installing Garden Windows

Although often found in kitchens, a garden window is an attractive option for nearly any room in your home. Projecting out from the wall 16 to 24", garden windows add space to a room, making it feel larger. The glass roof and box-like design make them ideal growing environments for plants or display areas for collectibles. Garden windows also typically include front- or side-opening windows. These allow for ventilation and are usually available in either awning or casement style.

Home stores often stock garden windows in several common sizes. However, it may be difficult to locate a stock window that will fit in your existing window rough opening. In cases like this you must rebuild the rough opening to the proper size. It may be worth the added expense to custom-order your garden window to fit into the existing rough opening.

The large amount of glass in a garden window has a direct effect on the window's energy efficiency. When purchasing a garden window, as a minimum, look for double-pane glass with low-emissivity (low-E) coatings. More expensive super-efficient types of glass are available for severely cold climates.

Installation methods for garden windows vary by manufacturer. Some units include a nailing flange that attaches to the framing and holds the window against the house. Other models hang on a separate mounting frame that attaches to the outside of the house. In this project, the garden window has a built-in mounting sleeve that slides into the rough opening and is attached directly to the rough framing.

Tools & Materials ▸

Tape measure
Hammer
Level
Framing square
Circular saw
Wood chisel
Stapler
Drill and bits
Caulking gun
Utility knife
Garden window kit
Wood strips
2 × 4s
Shims
Exterior trim
Building paper
3" screws
Drip edge
Construction adhesive
4d siding nails
8d galvanized casing nails
Interior trim
Paintable silicone caulk

A garden window's glass roof makes it an ideal sun spot for houseplants. It can also help a room feel larger.

How to Install a Garden Window

Prepare the project site and remove the interior and exterior trim, then remove the existing window.

Check the rough opening measurements to verify the correct window sizing. The rough opening should be about ½" larger than the window height and width. If necessary, attach wood strips to the rough framing as spacers to bring the opening to the required size.

Use a level to check that the sill of the rough opening is level and the side jambs are plumb. Use a framing square to make sure each corner is square. The rough framing must be in good condition in order to support the weight of the garden window. If the framing is severely deteriorated or out of plumb or square, you may need to reframe the rough opening (pages 40 to 43).

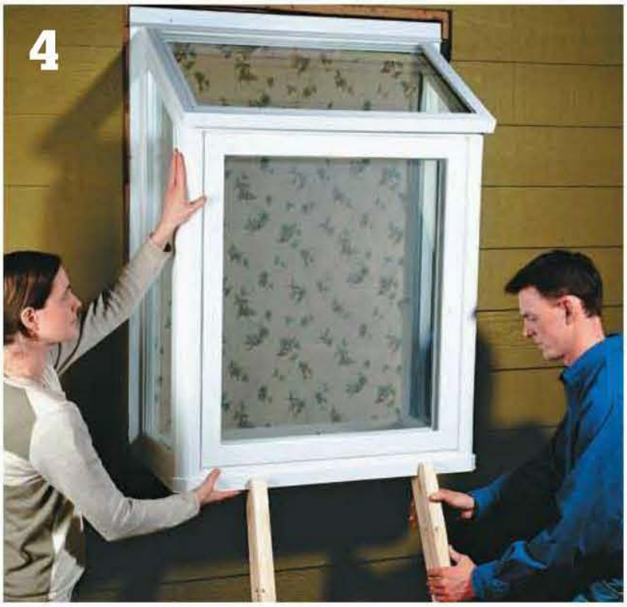

Insert the garden window into the opening, pressing it tight against the framing. Support the unit with notched 2 × 4s under the bottom edge of the window until it has been fastened securely to the framing.

(continued)

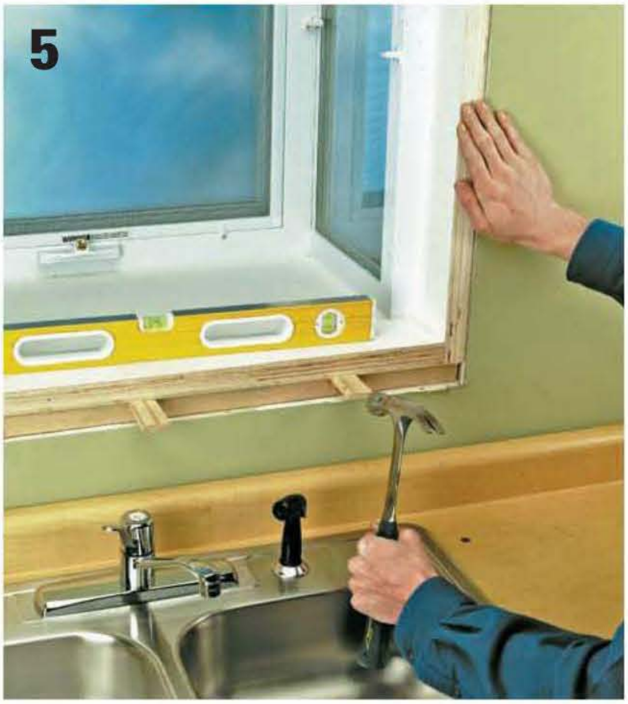

The inside edge of the window sleeve should be flush with the interior wall surface. Check the sill of the garden window for level. Shim beneath the lower side of the sill, if necessary, to make it level.

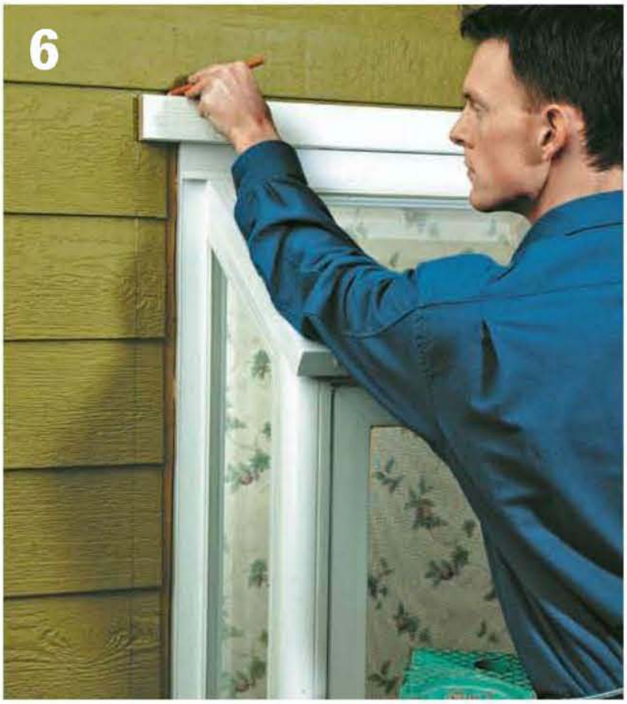

Once the garden window is in place and level, hold a piece of window trim in place along the exterior of the window and trace the outline onto the siding. Remove the window. Cut the siding down to the sheathing using a circular saw.

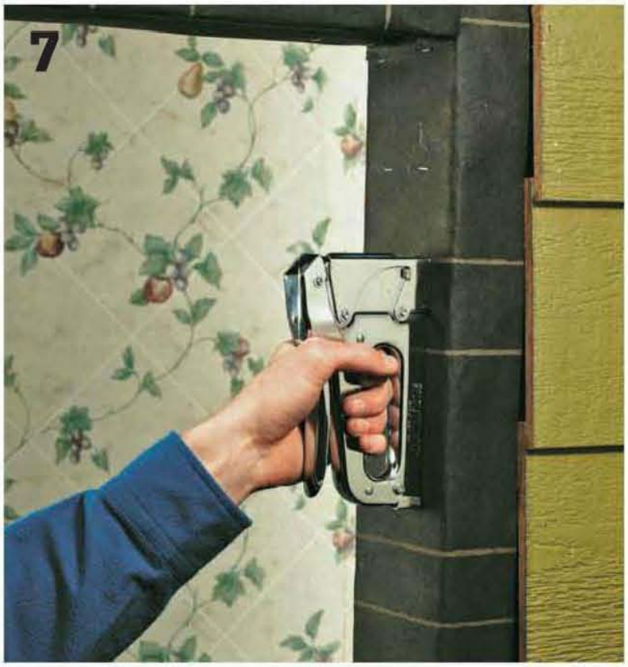

Install strips of building paper between siding and sheathing. Wrap them around the framing and staple them in place. On the sides, work from the bottom up so each piece overlaps the piece below. Reposition the window and reshim. Make sure the space between the window and the siding is equal to the width of the trim on all sides.

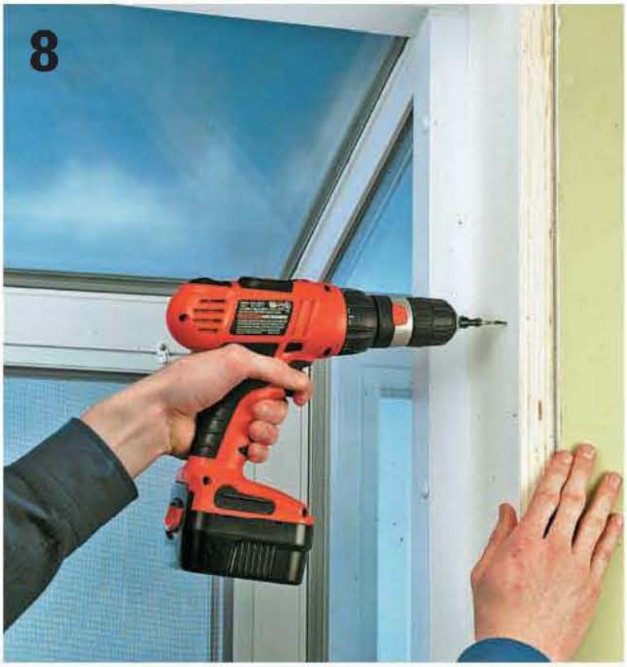

Drill countersunk pilot holes every 12" to 16" through the window sleeve into the rough header, jack studs, and sill.

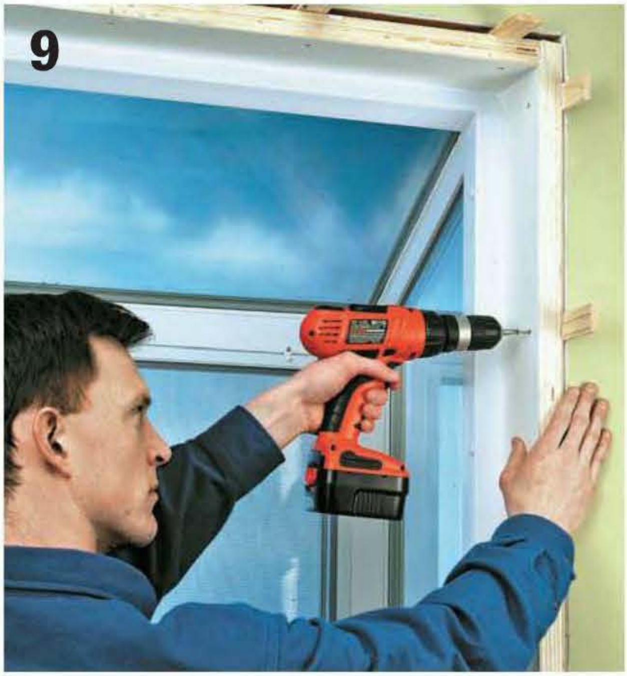

9

Insert shims between the window sleeve and rough frame at each hole location along the top and sides to prevent bowing of the window frame. Fasten the window to the framing using 3" screws. Continue checking for level, plumb, and square as the screws are tightened.

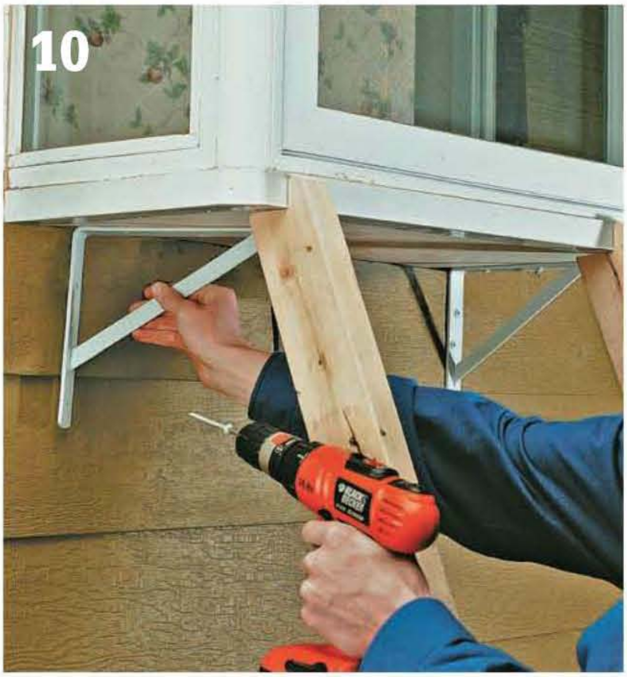

10

Locate and mark the studs nearest the edges of the window using a stud finder. Cut two pieces of siding to fit behind the brackets, and tack them in place over the marked studs with 4d siding nails. Position the support brackets with the shorter side against the siding and the longer side beneath the window. Fasten the brackets to the window and the studs using the included screws.

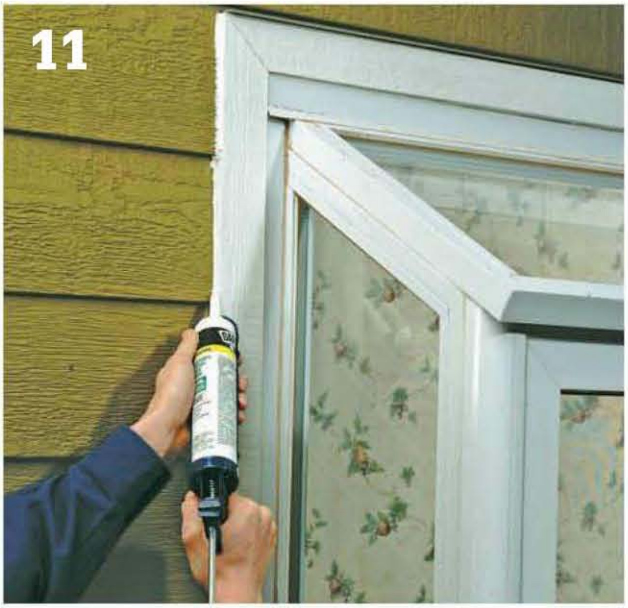

11

Cut a piece of drip edge to length, apply construction adhesive to its top flange, and slide it under the siding above the window. Cut each trim piece to size. Position the trim and attach it using 8d galvanized casing nails driven through pilot holes. Seal the edges of the trim with a bead of paintable silicone caulk, approximately ⅜" wide.

12

Cut all protruding shims flush with the framing using a utility knife or handsaw. Insulate or caulk gaps between the window sleeve and the wall. Finish the installation by reinstalling the existing interior trim or installing new trim.

Installing Bay Windows

Modern bay windows are preassembled for easy installation, but it will still take several days to complete an installation. Bay windows are large and heavy, and installing them requires special techniques. Have at least one helper to assist you, and try to schedule the work when there's little chance of rain. Using prebuilt bay window accessories will speed your work (see next page).

A large bay window can weigh several hundred pounds, so it must be anchored securely to framing members in the wall and supported by braces attached to framing members below the window. Some window manufacturers include cable-support hardware that can be used instead of metal support braces.

Before purchasing a bay window unit, check with the local building department regarding the code requirements. Many local codes require large windows and low bay windows with window seats to be glazed with tempered glass for safety.

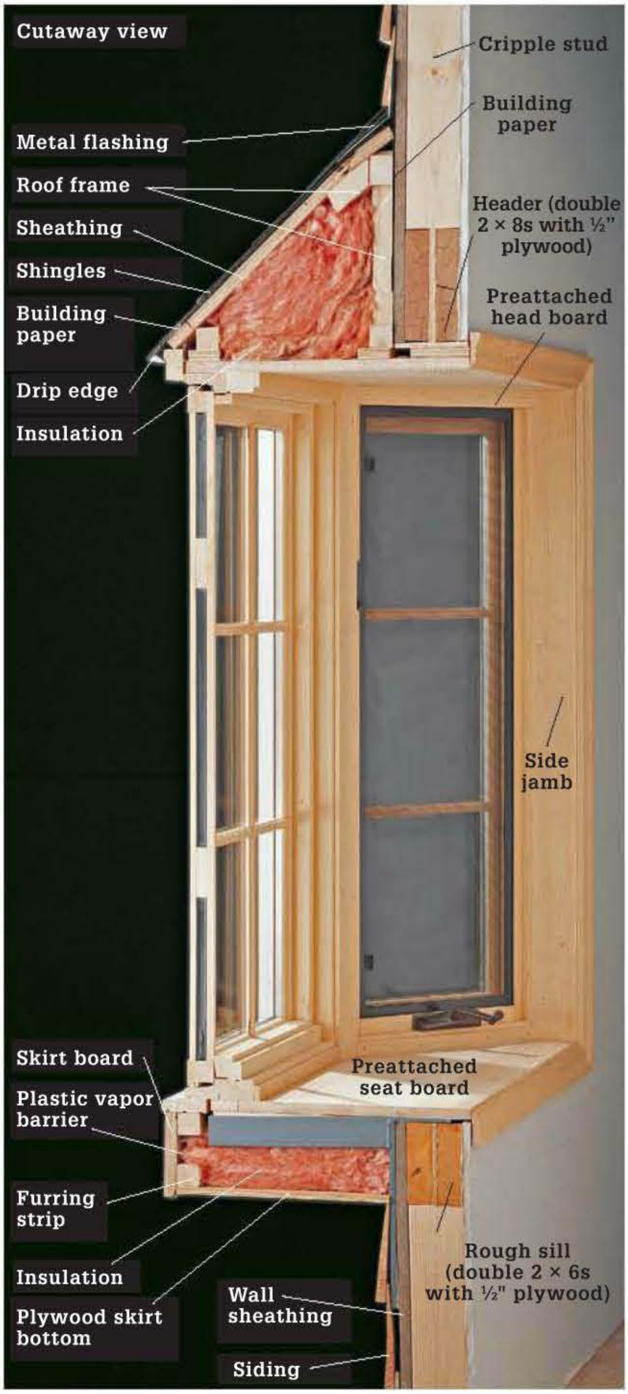
Cutaway view
Cripple stud
Metal flashing
Building paper
Roof frame
Sheathing
Header (double 2 × 8s with ½" plywood)
Shingles
Building paper
Preattached head board
Drip edge
Insulation
Side jamb
Skirt board
Preattached seat board
Plastic vapor barrier
Furring strip
Insulation
Rough sill (double 2 × 6s with ½" plywood)
Wall sheathing
Plywood skirt bottom
Siding

Tools & Materials ▸

- Straightedge
- Circular saw
- Wood chisel
- Pry bar
- Drill
- Level
- Nail set
- Stapler
- Aviation snips
- Roofing knife
- Caulk gun
- Utility knife
- T-bevel
- Bay window unit
- Prebuilt roof frame kit
- Metal support brackets
- 2× lumber
- 16d galvanized common nails
- 16d and 8d galvanized casing nails
- 3" and 2" galvanized utility screws
- 16d casing nails
- Tapered wood shims
- Building paper
- Fiberglass insulation
- 6-mil polyethylene sheeting
- Drip edge
- 1" roofing nails
- Step flashing
- Shingles
- Top flashing
- Roofing cement
- 2 × 2 lumber
- 5½" skirt boards
- ¾" exterior-grade plywood
- Paintable silicone caulk

Tips for Installing a Bay Window ▸

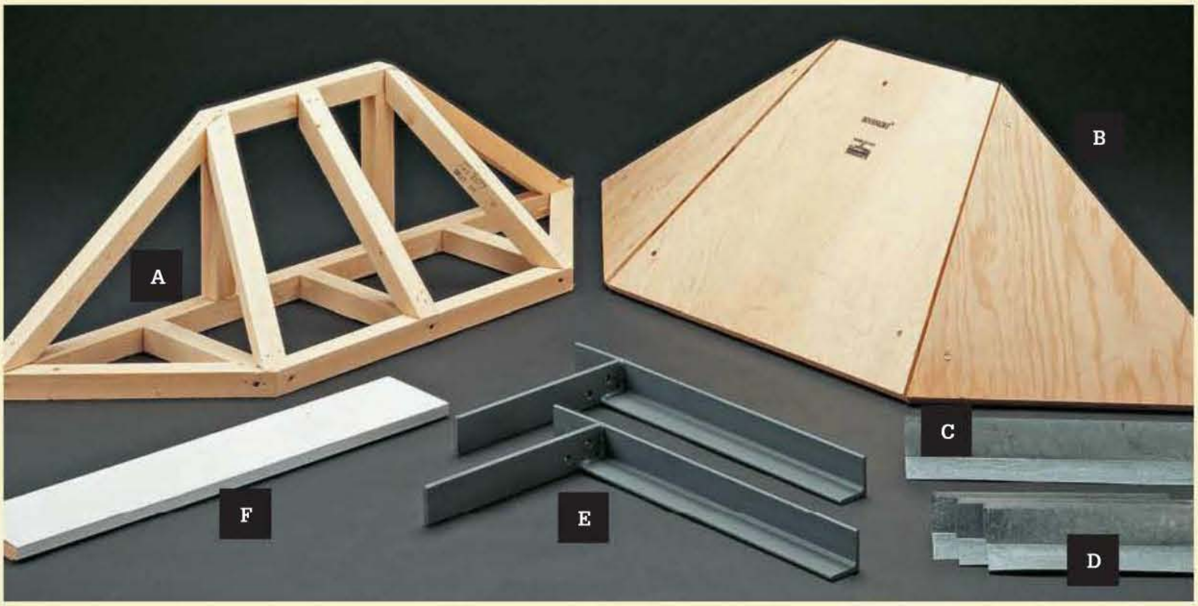

Use prebuilt accessories to ease installation of a bay window. Roof frames (A) come complete with sheathing (B), metal top flashing (C), and step flashing (D) and can be special-ordered at most home centers. You will have to specify the exact size of your window unit and the angle (pitch) you want for the roof. You can cover the roof inexpensively with building paper and shingles or order a copper or aluminum shell. Metal support braces (E) and skirt boards (F) can be ordered at your home center if not included with the window unit. Use two braces for bay windows up to 5 ft. wide and three braces for larger windows. Skirt boards are clad with aluminum or vinyl and can be cut to fit with a circular saw or miter saw.

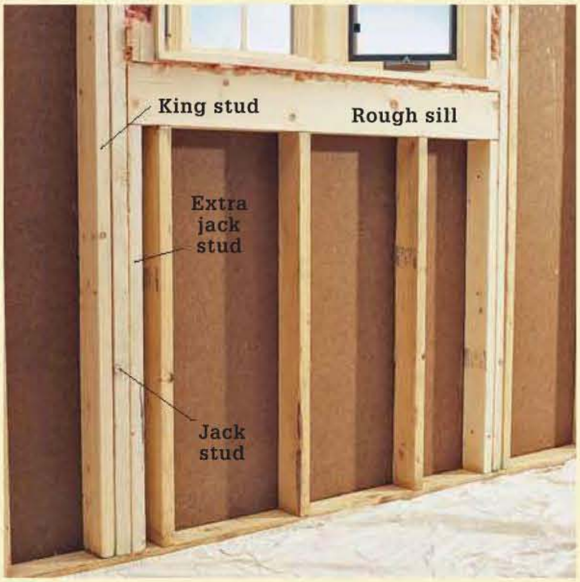

Construct a bay window frame similar to that for a standard window (see page 257) but use a built-up sill made from two 2 × 6s sandwiched around ½" plywood. Install extra jack studs under the sill ends to help carry the window's weight.

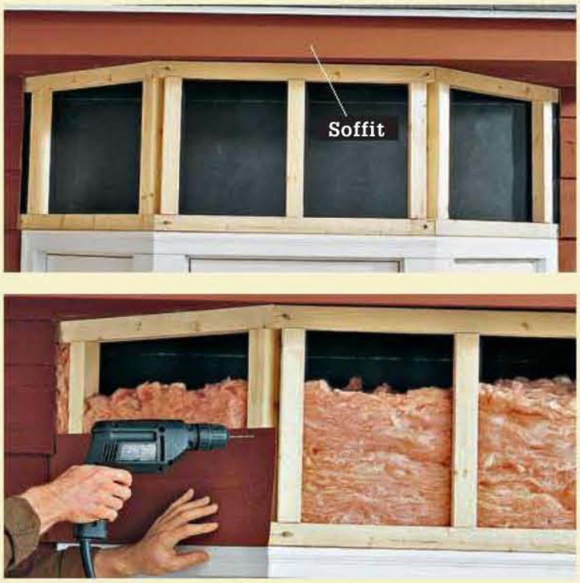

Build an enclosure above the bay window if the roof soffit overhangs the window. Build a 2 × 2 frame (top) to match the angles of the bay window, and attach the frame securely to the wall and overhanging soffit. Install a vapor barrier and insulation, then finish the enclosure so it matches the siding (bottom).

How to Install a Bay Window

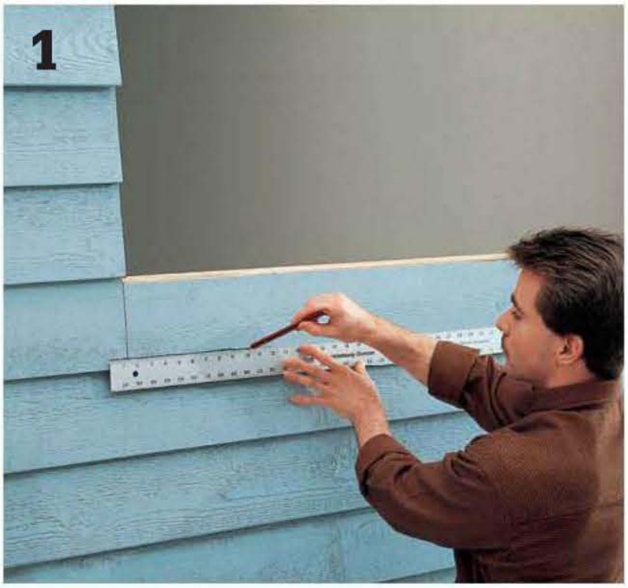

Prepare the project site and remove interior wall surfaces (pages 32 to 35), then frame the rough opening. Remove the exterior wall surfaces. Mark for removal a section of siding directly below the rough opening. The width of the marked area should equal that of the window unit and the height should equal that of the skirt board.

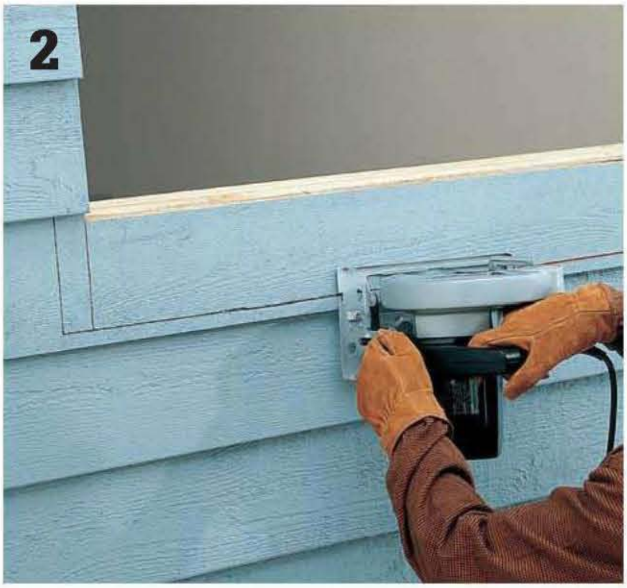

Set the blade on a circular saw just deep enough to cut through the siding, then cut along the outline. Stop just short of the corners to avoid damaging the siding outside the outline. Use a sharp chisel to complete the corner cuts. Remove the cut siding inside the outline.

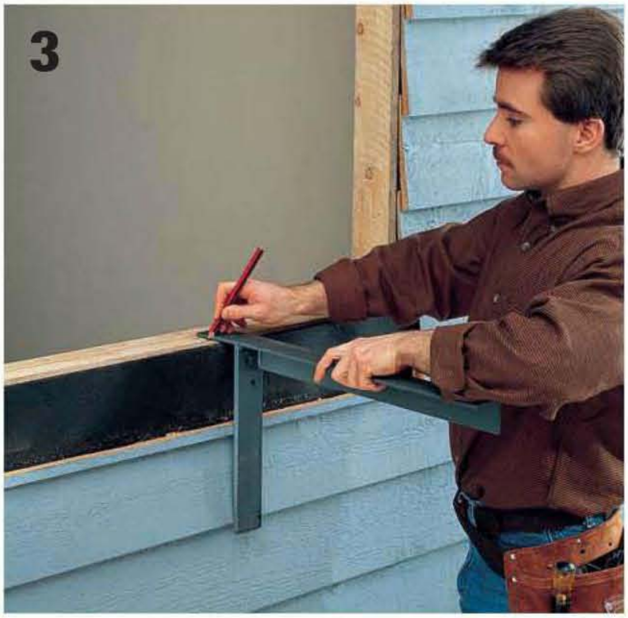

Position the support braces along the rough sill within the widest part of the bay window and above the cripple stud locations. Add cripple studs to match the support brace locations, if necessary. Draw outlines of the braces on the top of the sill. Use a chisel or circular saw to notch the sill to a depth equal to the thickness of the top arm of the support braces.

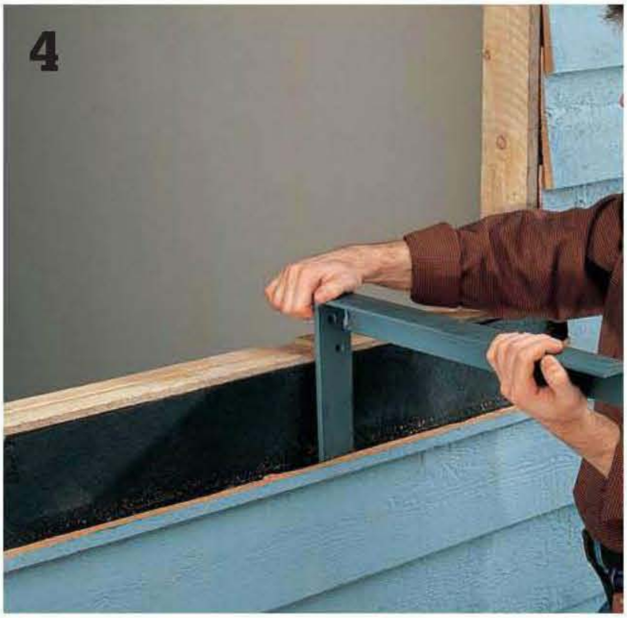

Slide the support braces down between the siding and the sheathing. Pry the siding material away from the sheathing slightly to make room for the braces, if necessary. *Note: On stucco, you will need to chisel notches in the masonry surface to fit the support braces.*

5

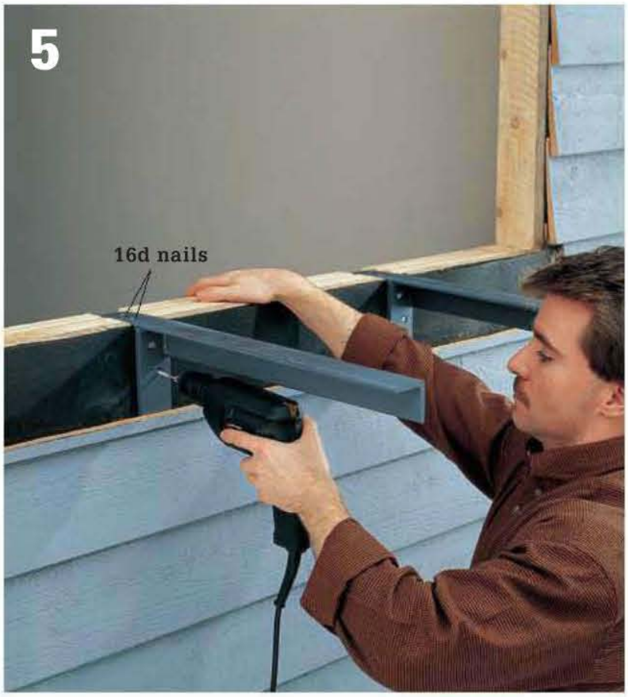

Attach the braces to the rough sill with galvanized 16d common nails. Drive 3" utility screws through the front of the braces and into the rough sill to prevent twisting.

6

Lift the bay window onto the support braces and slide it into the rough opening. Center the unit within the opening.

7

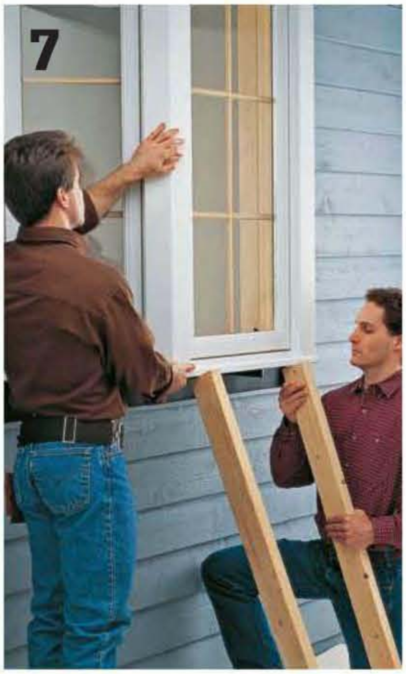

Check the window unit to make sure it is level. If necessary, drive shims under the low side to level the window. Temporarily brace the outside bottom edge of the unit with 2 × 4s to keep it from moving on the braces.

8

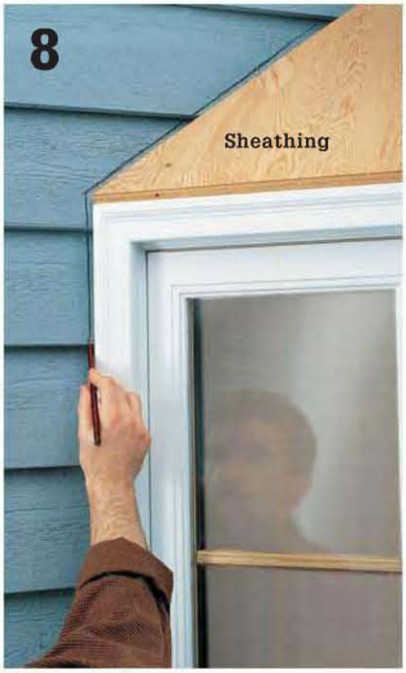

Set the roof frame on top of the window, with the sheathing loosely tacked in place. Trace the outline of the window and roof unit onto the siding. Leave a gap of about ½" around the roof unit to allow room for flashing and shingles.

9

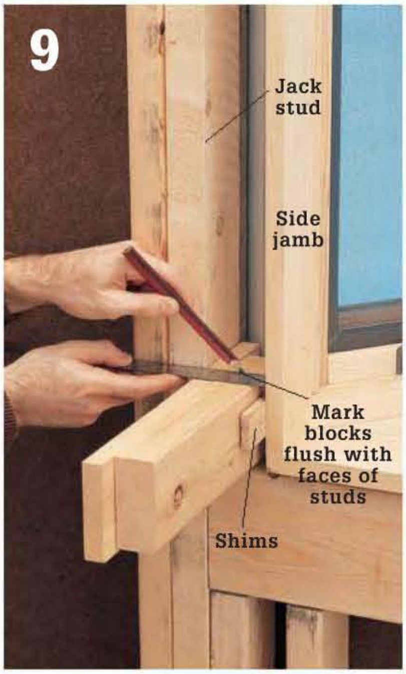

If the gap between the side jambs and jack studs is more than 1" wide, mark and cut wood blocks to bridge the gap (smaller gaps require no blocks). Leave a small space for inserting wood shims. Remove the window, then attach blocks every 12" along studs.

(continued)

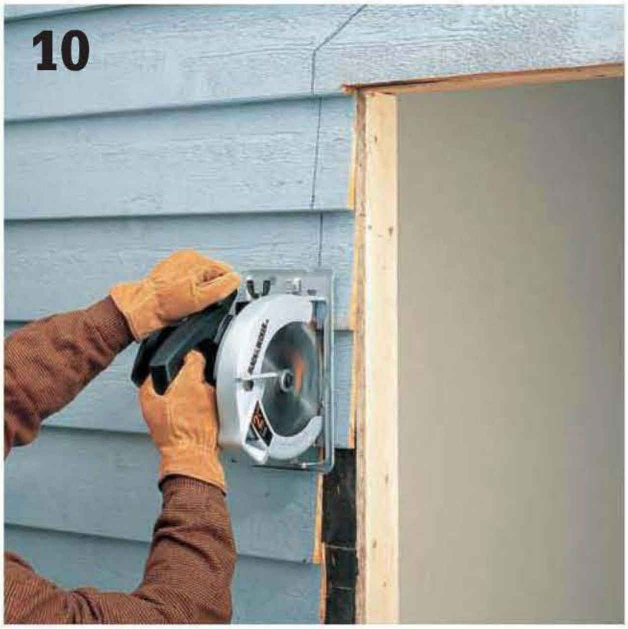

Cut the siding just down to the sheathing along the outline using a circular saw. Stop just short of the corners, then use a wood chisel to complete the corner cuts. Remove the cut siding. Pry the remaining siding slightly away from the sheathing around the roof outline to allow for easy installation of the metal flashing. Cover the exposed sheathing with 8"-wide strips of building paper (step 3, page 260).

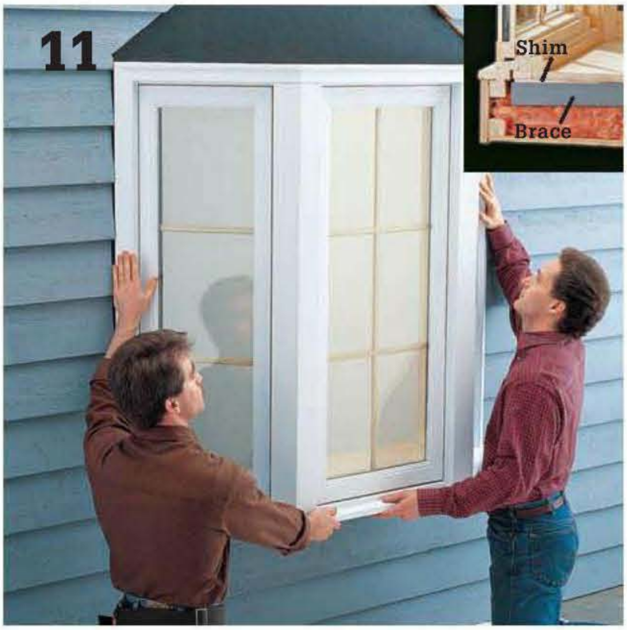

Set the bay window unit back on the braces, and slide it back into the rough opening until the brick moldings are tight against the sheathing. Insert wood shims between the outside end of the metal braces and the seat board (inset). Check the unit to make sure it is level, and adjust the shims, if necessary.

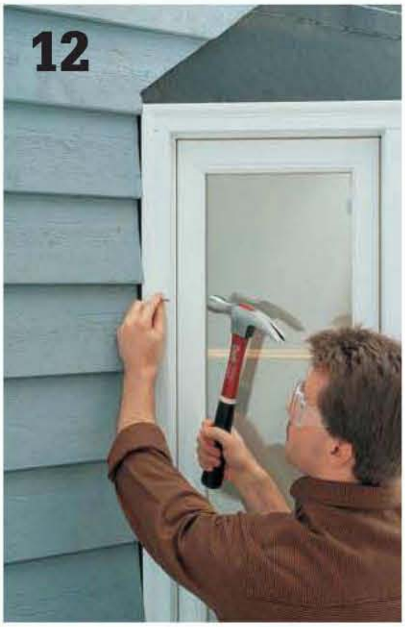

Anchor the window by drilling pilot holes and driving 16d casing nails through the brick molding and into the framing members. Space nails every 12", and use a nail set to drive the nail heads below the surface of the wood.

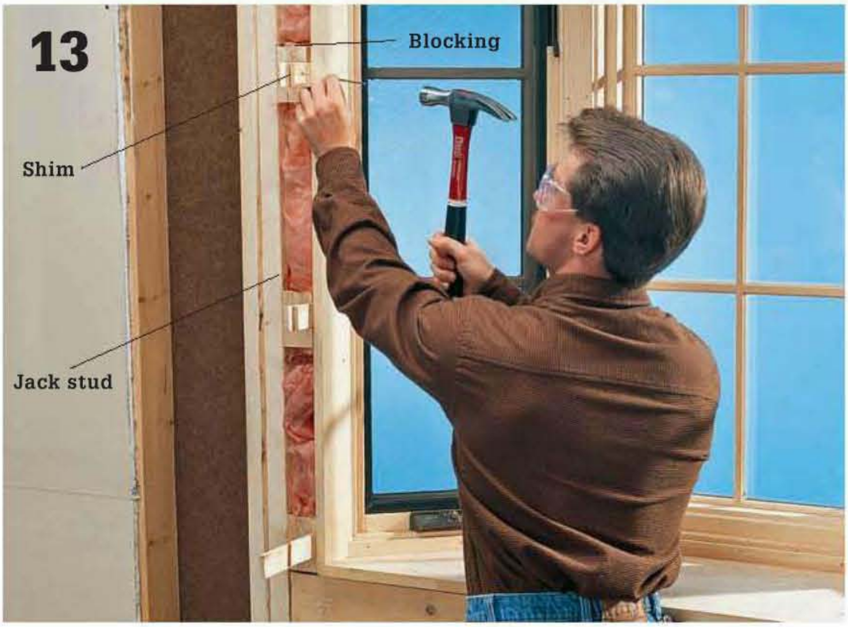

Drive wood shims into the spaces between the side jambs and the blocking or jack studs and between the headboard and header, spacing the shims every 12". Fill the spaces around the window with loosely packed fiberglass insulation. At each shim location, drive 16d casing nails through the jambs and shims and into the framing members. Cut off the shims flush with the framing members using a handsaw or utility knife. Use a nail set to drive the nail heads below the surface. If necessary, drill pilot holes to prevent splitting the wood.

14

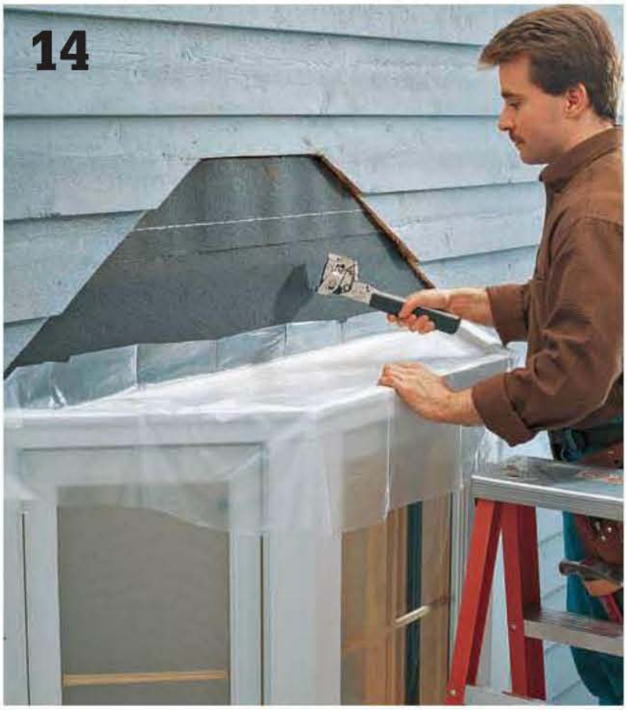

Staple sheet plastic over the top of the window unit to serve as a vapor barrier. Trim the edges of the plastic around the top of the window using a utility knife.

15

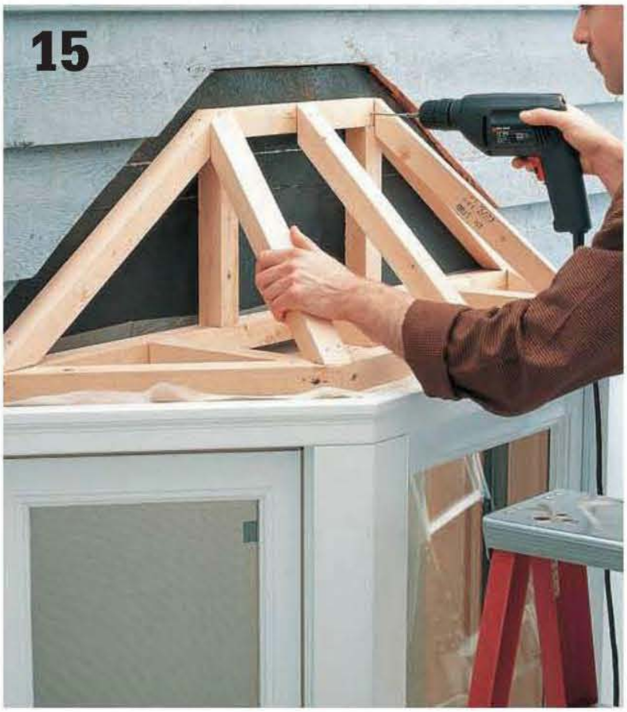

Remove the sheathing pieces from the roof frame, then position the frame on top of the window unit. Attach the roof frame to the window and to the wall at stud locations using 3" utility screws.

16

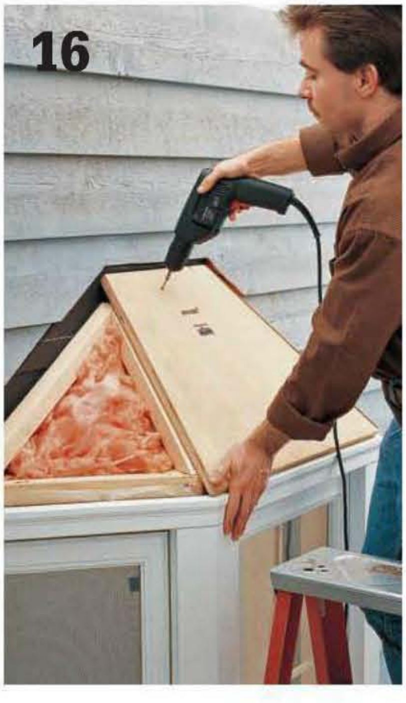

Fill the empty space inside the roof frame with loosely packed fiberglass insulation. Screw the sheathing back onto the roof frame using 2" utility screws.

17

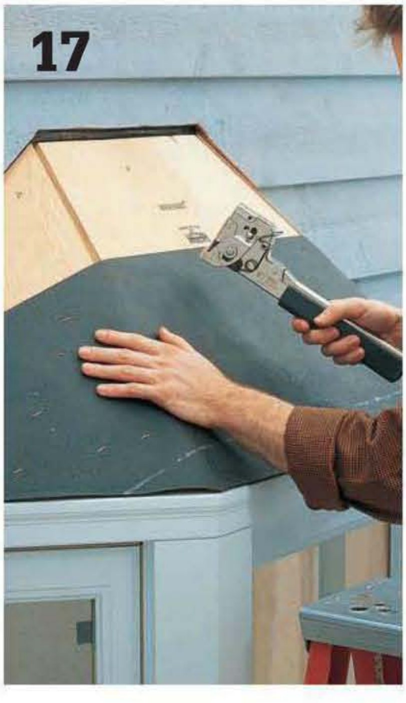

Staple asphalt building paper over the roof sheathing. Make sure each piece of building paper overlaps the one below by at least 5".

18

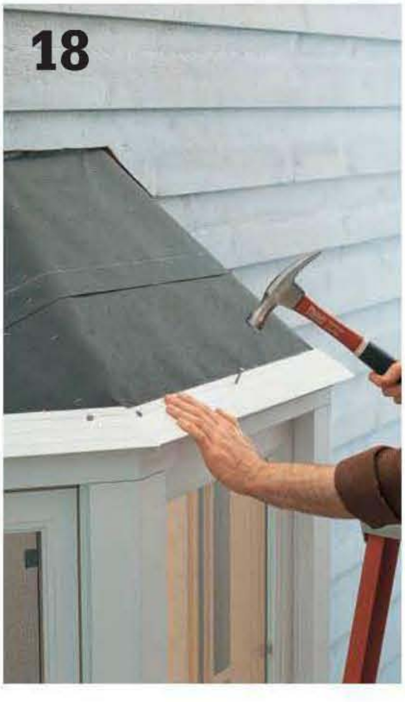

Cut drip edges with aviation snips, then attach them around the edge of the roof sheathing using roofing nails.

(continued)

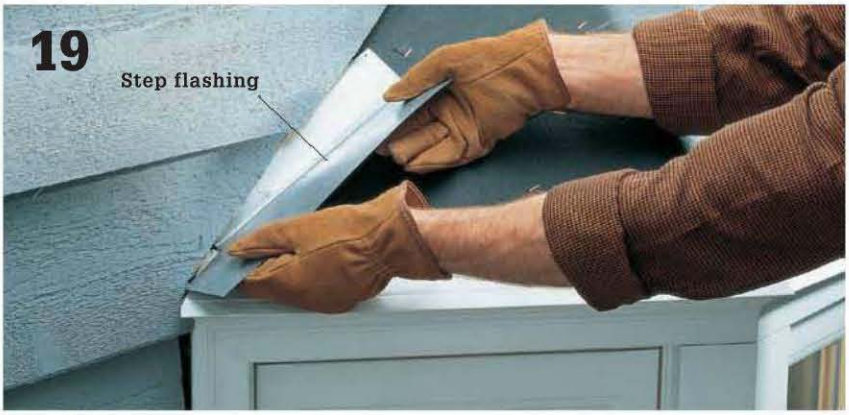

Cut and fit a piece of step flashing on each side of the roof frame. Adjust the flashing so it overhangs the drip edge by ¼". Flashings help guard against moisture damage.

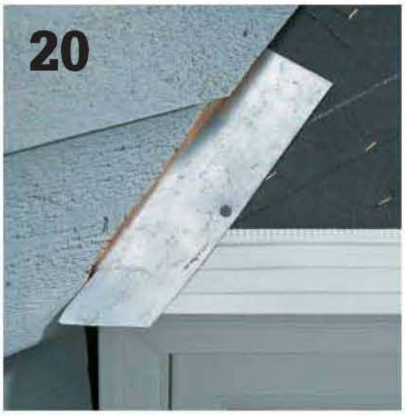

Trim the end of the flashing to the same angle as the drip edge. Nail the flashing to the sheathing with roofing nails.

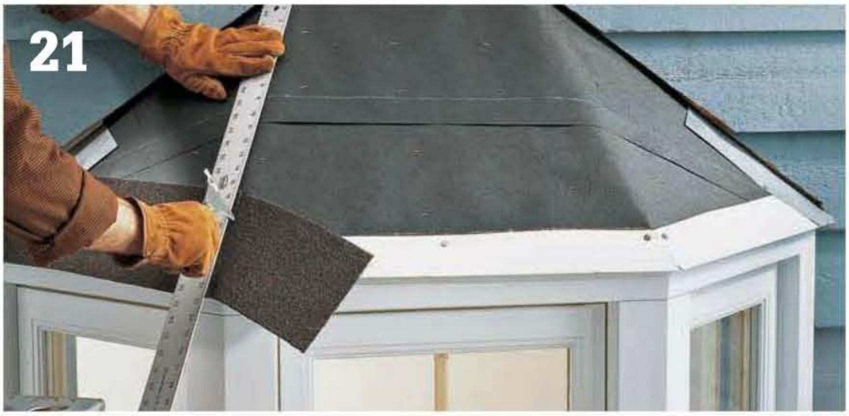

Cut 6"-wide strips of shingles for the starter row. Use roofing nails to attach the starter row shingles so they overhang the drip edge by about ½". Cut the shingles along the roof hips using a straightedge and a roofing knife.

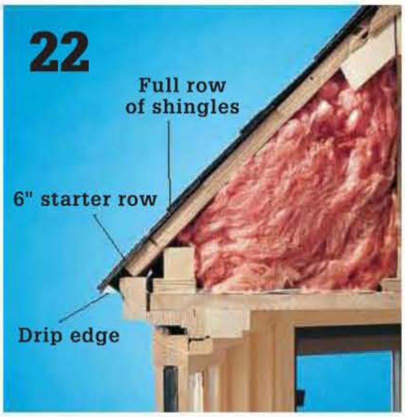

Nail a full row of shingles over the starter row, aligning the bottom edges with the bottom edge of the starter row. Make sure shingle notches are not aligned.

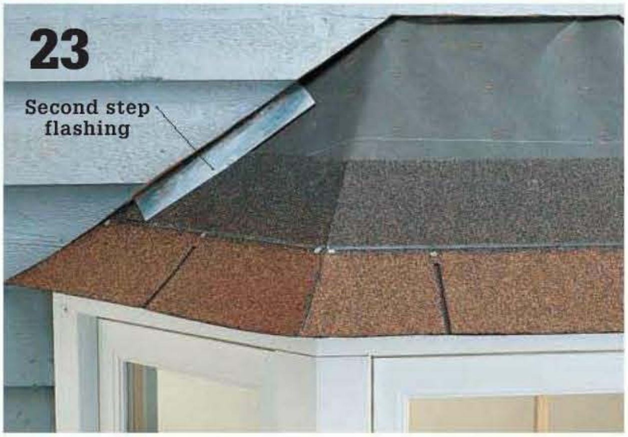

Install another piece of step flashing on each side of the roof, overlapping the first piece of flashing by about 5".

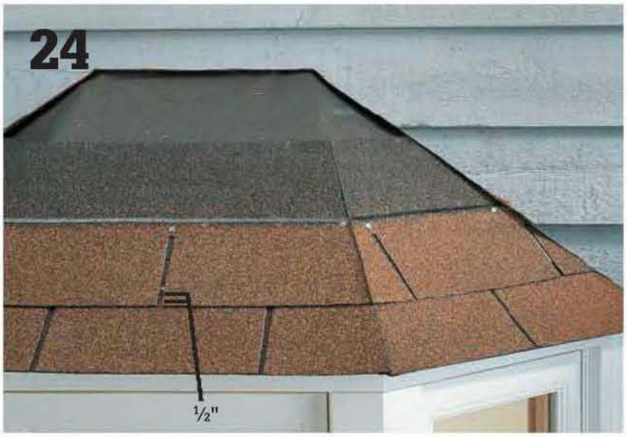

Cut and install another row of full shingles. The bottom edges should overlap the tops of the notches on the previous row by ½". Attach the shingles with roofing nails driven just above the notches.

25

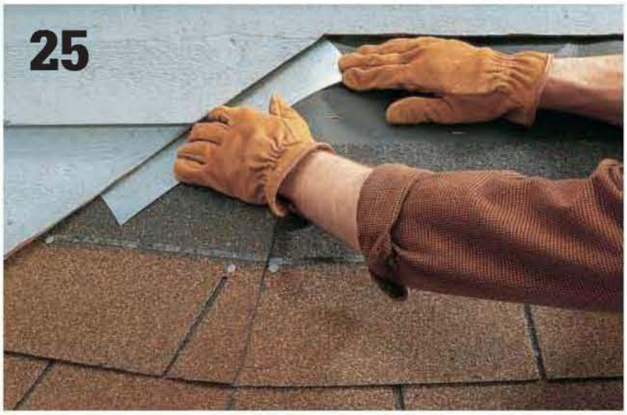

Continue installing alternate rows of step flashing and shingles to the top of the roof. Bend the last pieces of step flashing to fit over the roof hips.

26

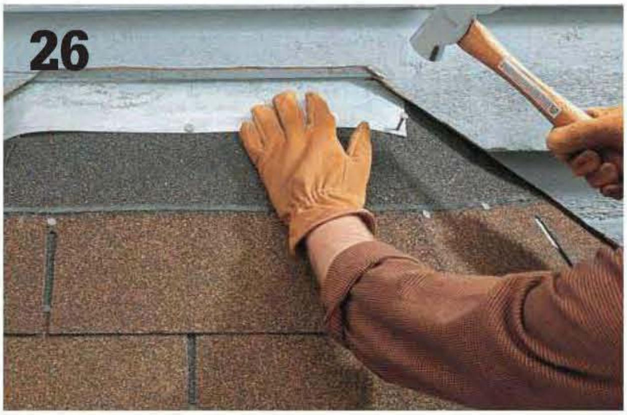

When the roof sheathing is covered with shingles, install the top flashing. Cut and bend the ends over the roof hips, and attach it with roofing nails. Attach the remaining rows of shingles over the top flashing.

27

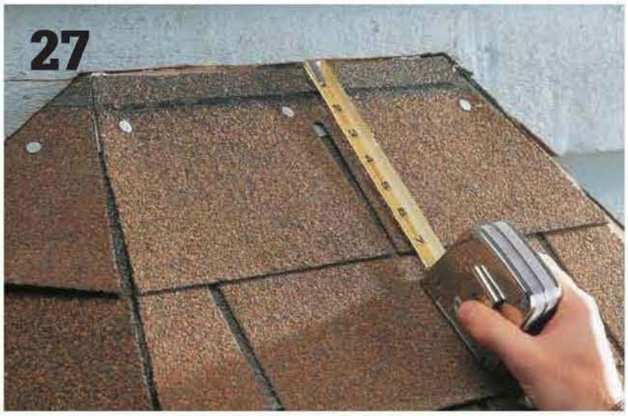

Find the height of the final row of shingles by measuring from the top of the roof to a point ½" below the top of the notches on the last installed shingle. Trim the shingles to fit.

28

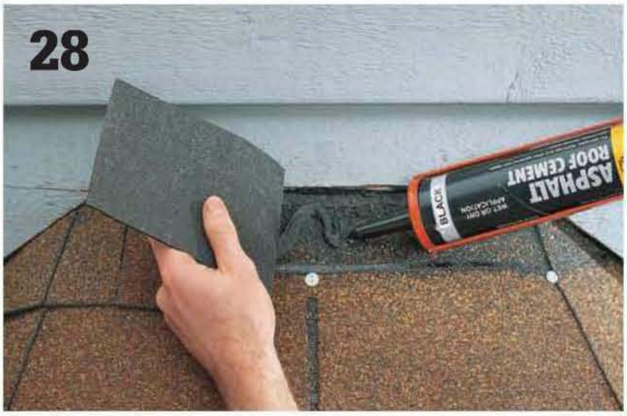

Attach the final row of shingles with a thick bead of roofing cement—not nails. Press firmly to ensure a good bond.

29

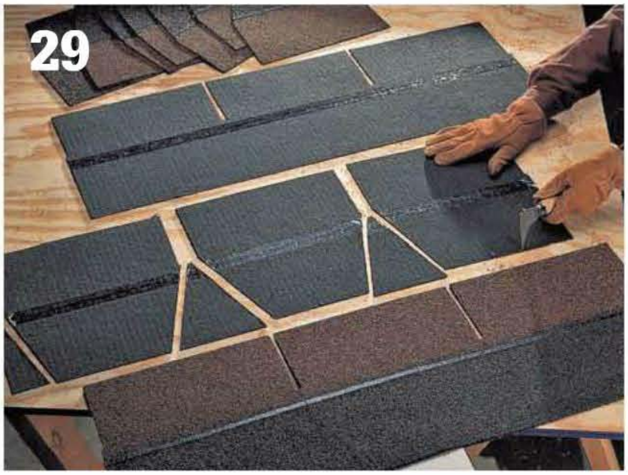

Make ridge caps by cutting shingles into 1-ft.-long sections. Use a roofing knife to trim off the top corners of each piece, so the ridge caps will be narrower at the top than at the bottom.

30

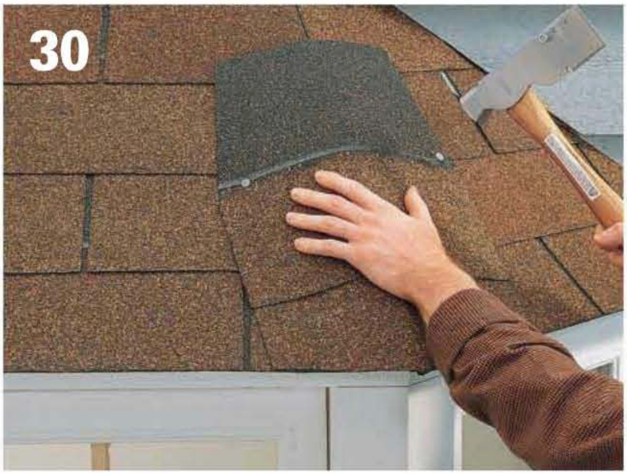

Install the ridge caps over the roof hips, beginning at the bottom of the roof. Trim the bottom ridge caps to match the edges of the roof. Keep the same amount of overlap with each layer.

(continued)

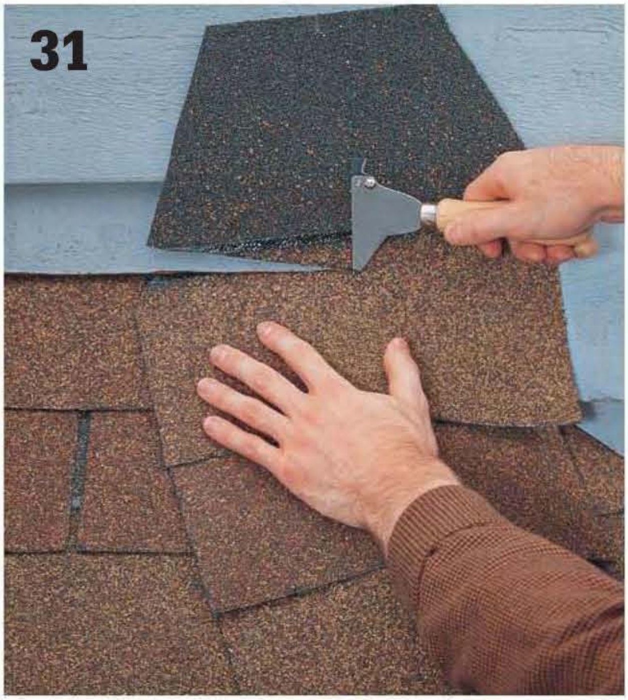

At the top of the roof hips, use a roofing knife to cut the shingles to fit flush with the wall. Attach the shingles with roofing cement—do not use any nails.

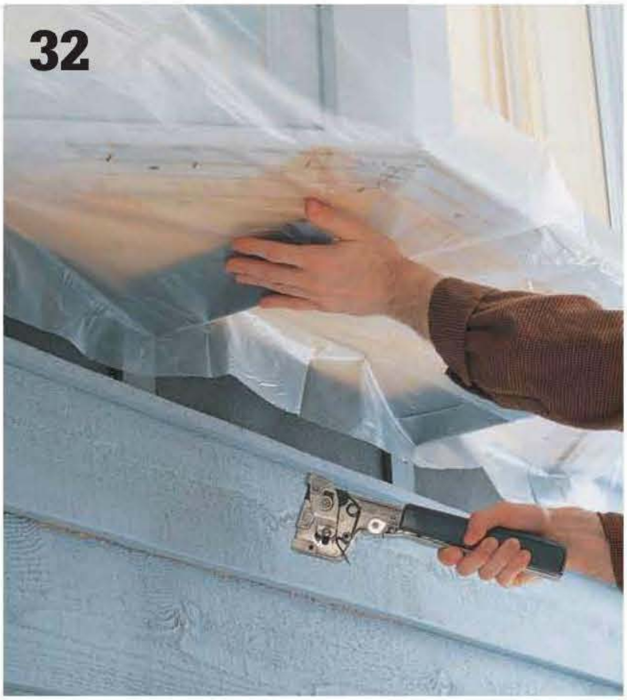

Staple sheet plastic over the bottom of the window unit to serve as a vapor barrier. Trim the plastic around the bottom of the window.

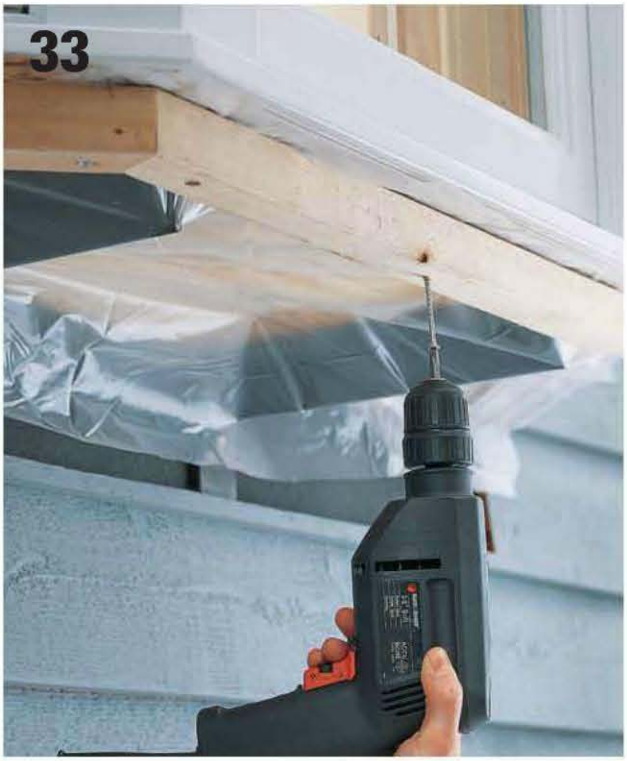

Cut and attach a 2 × 2 skirt frame around the bottom of the bay window using 3" galvanized utility screws. Set the skirt frame back about 1" from the edges of the window.

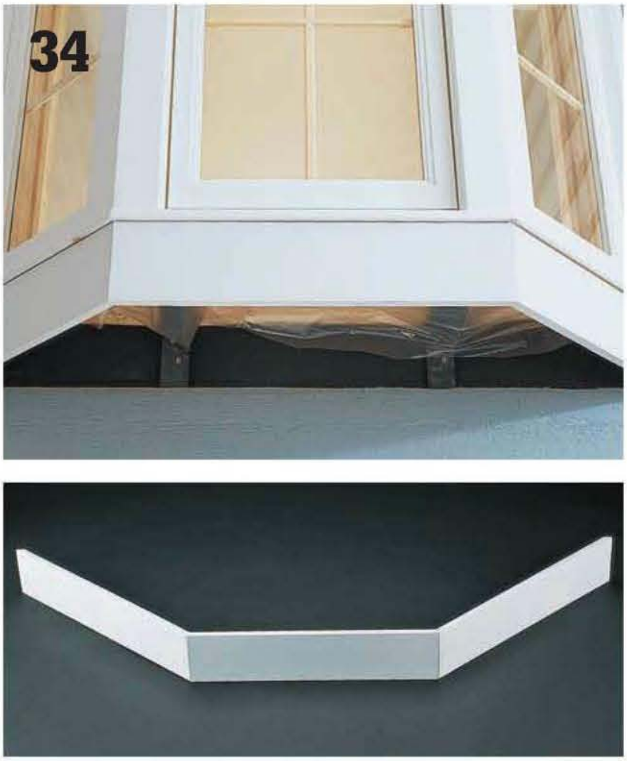

Cut skirt boards to match the shape of the bay window bottom, mitering the ends to ensure a tight fit. Test-fit the skirt board pieces to make sure they match the bay window bottom.

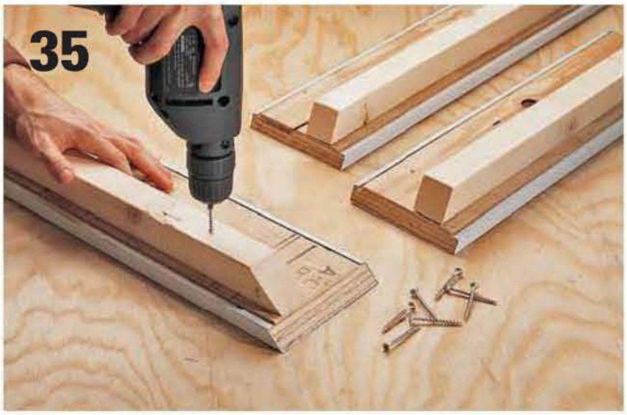

Cut a 2 × 2 furring strip for each skirt board. Miter the ends to the same angles as the skirt boards. Attach the furring strips to the back of the skirt boards, 1" from the bottom edges, using 2" galvanized utility screws.

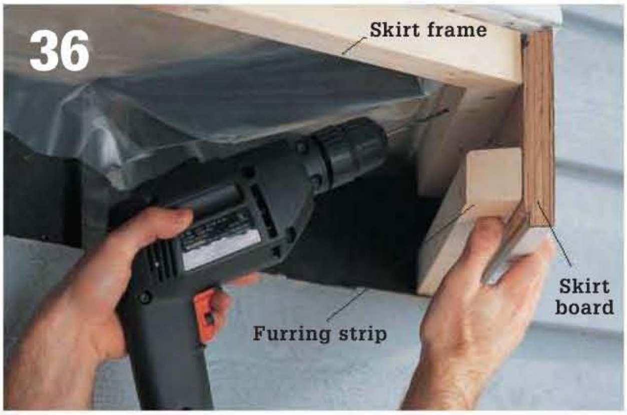

Attach the skirt board pieces to the skirt frame. Drill ⅛" pilot holes every 6" through the back of the skirt frame and into the skirt boards, then attach the skirt boards with 2" galvanized utility screws.

Measure the space inside the skirt boards using a T-bevel to duplicate the angles. Cut a skirt bottom from ¾" exterior-grade plywood to fit this space.

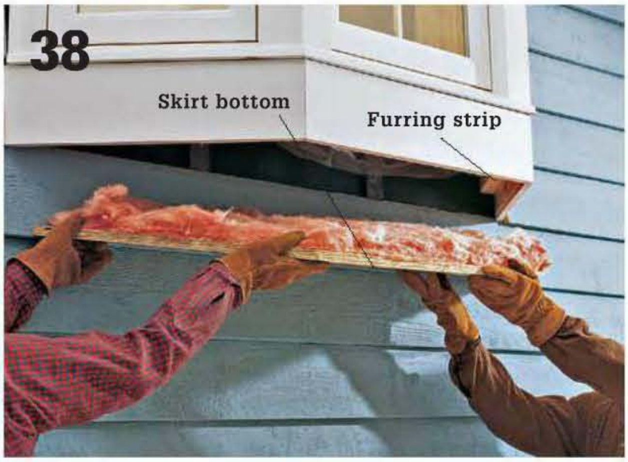

Lay fiberglass insulation on the skirt bottom. Position the skirt bottom against the furring strips and attach it by driving 2" galvanized utility screws every 6" through the bottom and into the furring strips.

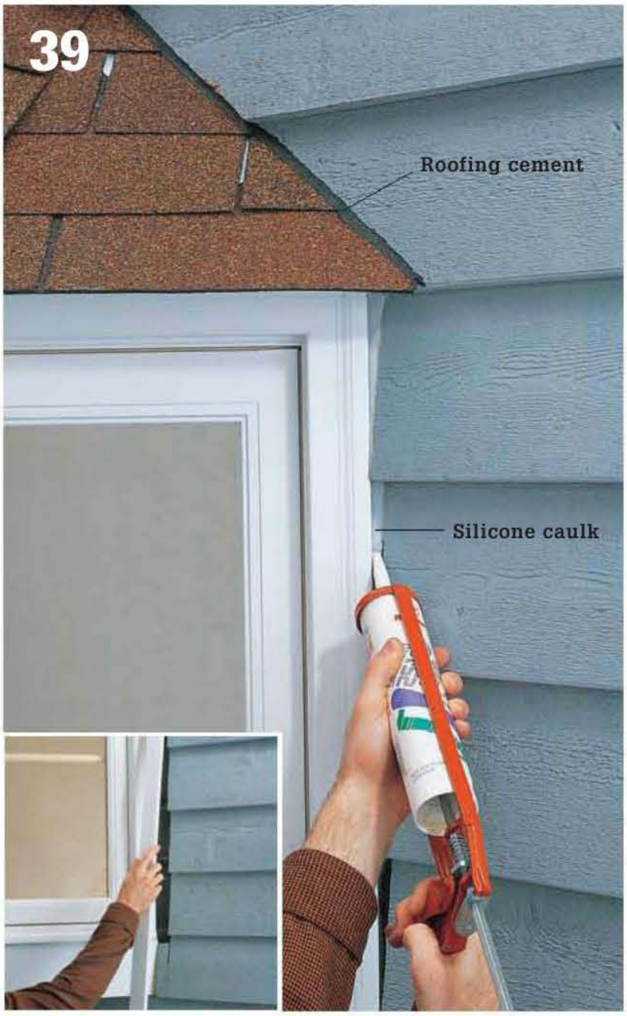

Install any additional trim pieces (inset) specified by your window manufacturer using 8d galvanized casing nails. Seal the roof edges with roofing cement, and seal around the rest of the window with paintable silicone caulk.

Installing Skylights

Since skylights let in so much light, their sizing and placement are important considerations. A skylight that's too big can quickly overheat a space, especially in an attic. The same is true of using too many skylights in any one room. For that reason it's often best to position a skylight away from the day's brightest sun. You may want an operable skylight that opens and closes to vent warm air.

When a skylight is installed above an unfinished attic space, a special skylight shaft must be constructed to channel light directly to the room below.

Skylights can offer warmth in the winter, cooling ventilation in the summer, and a view of the sky or the treetops around your house during any season. And, of course, skylights provide natural light.

Installing a skylight above finished space involves other considerations. First, the ceiling surface must be removed to expose the rafters. To remove wall and ceiling surfaces, see pages 28 to 31.

A skylight frame is similar to a standard window frame. It has a header and sill, like a window frame, but it has king rafters rather than king studs. Skylight frames also have trimmers that define the sides of the rough opening. Refer to the manufacturer's instructions to determine what size to make the opening for the skylight you select.

With standard rafter-frame roof construction, you can safely cut into one or two rafters as long as you permanently support the cut rafters, as shown in the following steps. If your skylight requires alteration of more than two rafters or if your roofing is made with unusually heavy material, such as clay tile or slate, consult an architect or engineer before starting the project.

Today's good-quality skylight units are unlikely to leak, but a skylight is only as leakproof as its installation. Follow the manufacturer's instructions, and install the flashing meticulously, as it will last a lot longer than any sealant.

Tools & Materials ▸

4-ft. level
Circular saw
Drill
Combination square
Reciprocating saw
Pry bar
Chalk line
Stapler
Caulk gun
Utility knife
Aviation snips
Plumb bob
Jigsaw
Wallboard tools
2× lumber
1 × 4
16d and 10d common nails
Building paper
Roofing cement
Skylight flashing
2", 1¼", and ¾" roofing nails
Finish nails
Fiberglass insulation
½" wallboard
Twine
Wallboard screws
6-mil polyethylene sheeting
Finishing materials

How to Install a Skylight

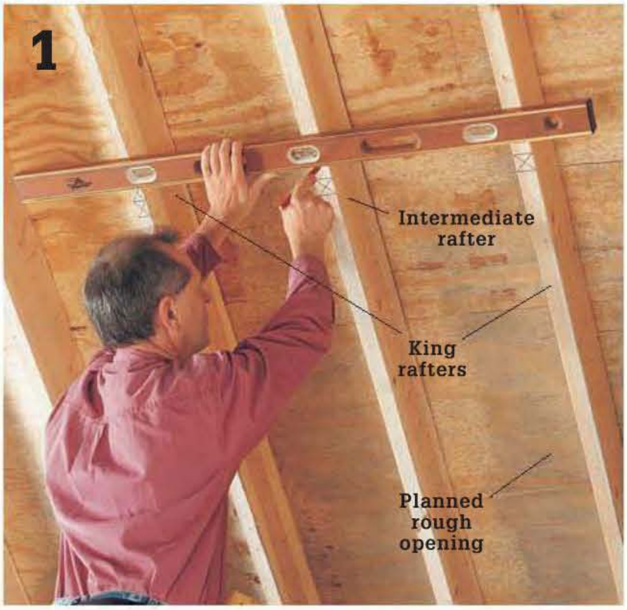

Use the first rafter on each side of the planned rough opening as a king rafter. Measure and mark where the double header and sill will fit against the king rafters. Then, use a level as a straightedge to extend the marks across the intermediate rafter.

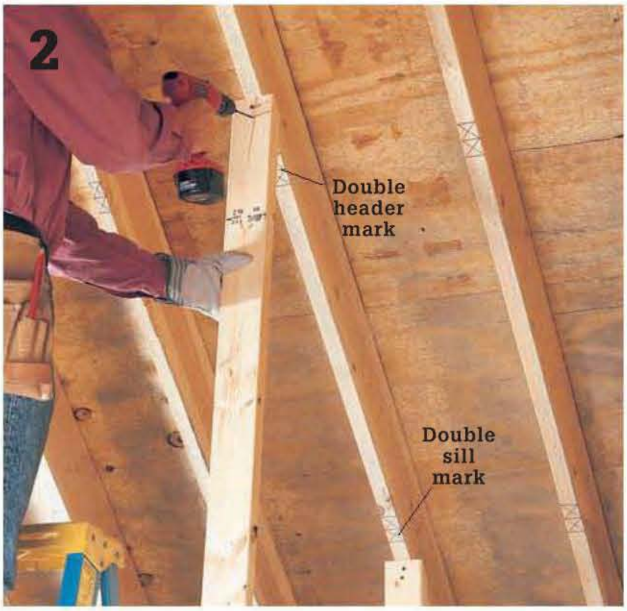

Brace the intermediate rafter by installing two 2 × 4s between the rafter and the attic floor. Position the braces just above the header marks and just below the sill marks. Secure them temporarily to the rafter and subfloor (or joists) with screws.

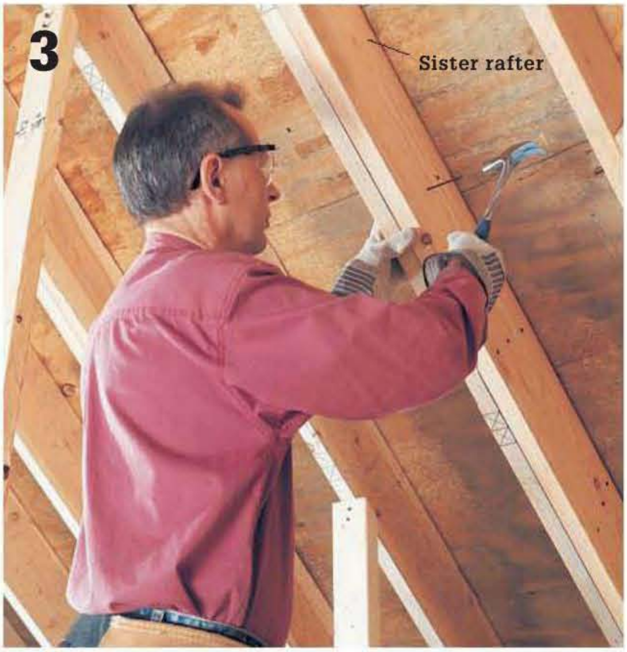

Reinforce each king rafter by attaching a full-length "sister" rafter against its outside face. Cut sister rafters from the same size of lumber as existing rafters, matching lengths and end cuts exactly. Work each one into position, flush against the outside face of the king rafters, then nail the sisters to the kings with pairs of 10d common nails spaced 12" apart.

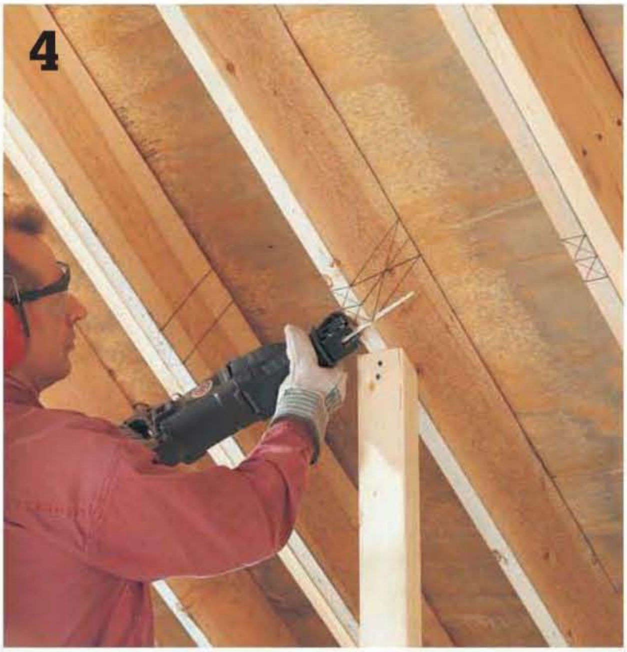

Use a combination square to transfer the sill and header marks across the face of the intermediate rafter, then cut along the outermost lines with a reciprocating saw. Do not cut into the roof sheathing. Carefully remove the cutout section with a pry bar. The remaining rafter portions will serve as cripple rafters.

(continued)

5

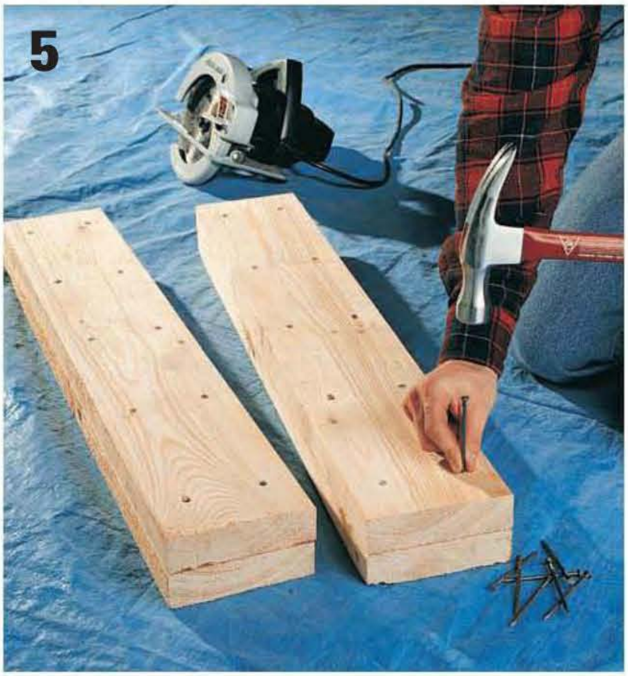

Build a double header and double sill to fit snugly between the king rafters, using 2× lumber that is the same size as the rafters. Nail the header pieces together using pairs of 10d nails spaced 6" apart.

6

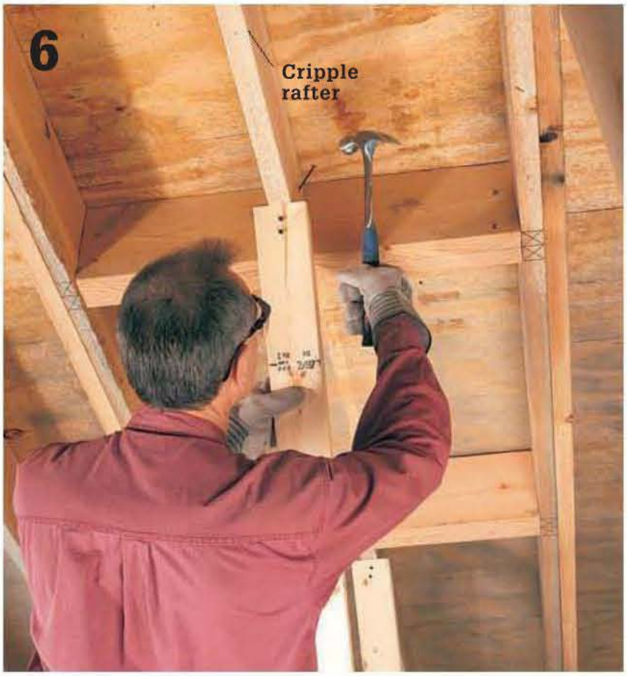

Install the header and sill, anchoring them to the king rafters and cripple rafters with 16d common nails. Make sure the ends of the header and sill are aligned with the appropriate marks on the king rafters.

7

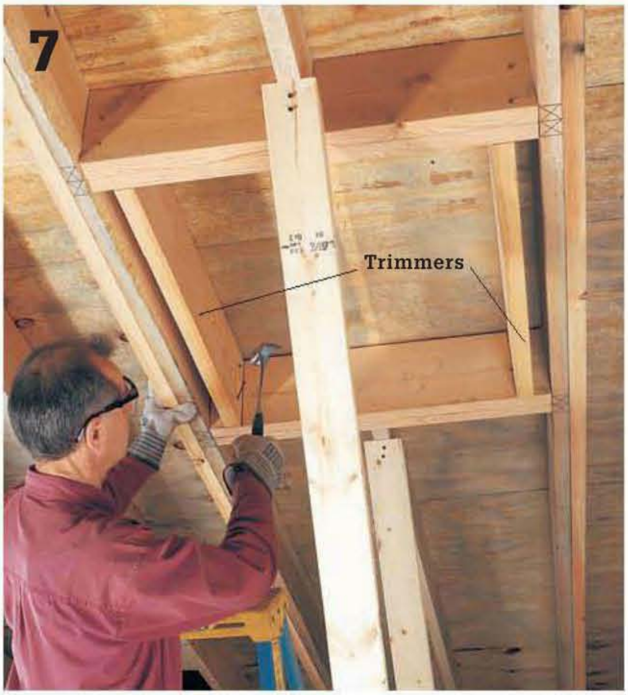

If your skylight unit is narrower than the opening between the king studs, measure and make marks for the trimmers: They should be centered in the opening and spaced according to the manufacturer's specifications. Cut the trimmers from the same 2× lumber used for the rest of the frame, and nail them in place with 10d common nails. Remove the 2 × 4 braces.

8

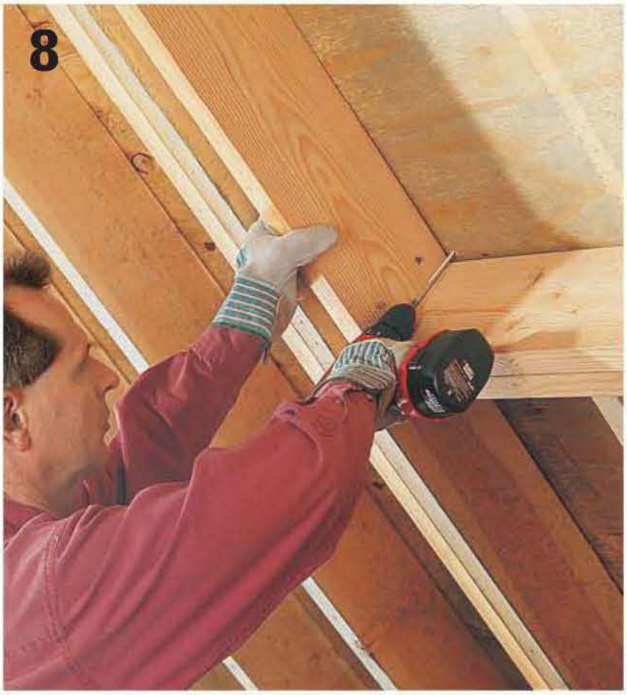

Mark the opening for the roof cutout by driving a screw through the sheathing at each corner of the frame. Then, tack a couple of scrap boards across the opening to prevent the roof cutout from falling and causing damage below.

9

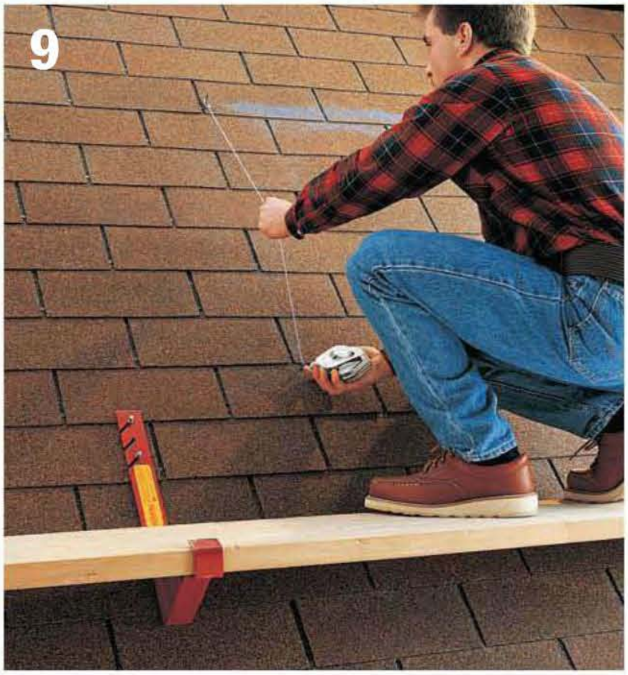

From the roof, measure between the screws to make sure the rough opening dimensions are accurate. Snap chalk lines between the screws to mark the rough opening, then remove the screws.

10

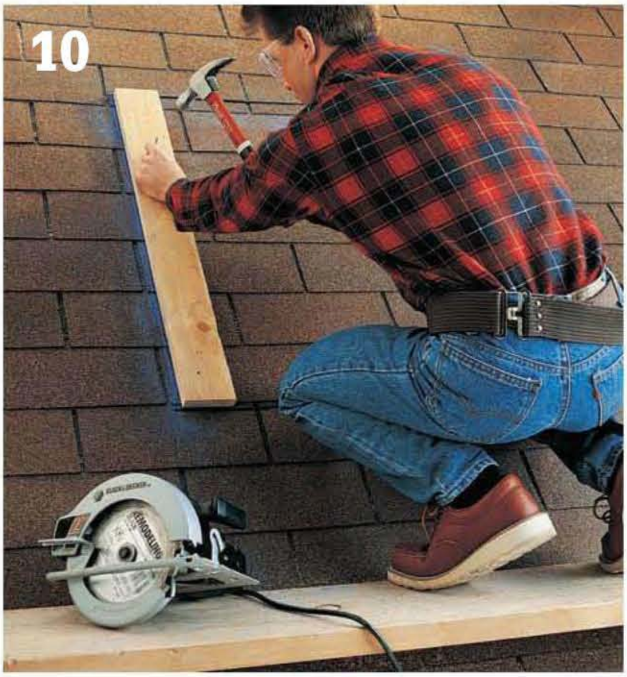

Tack a straight 1 × 4 to the roof, aligned with the inside edge of one chalk line. Make sure the nail heads are flush with the surface of the board.

11

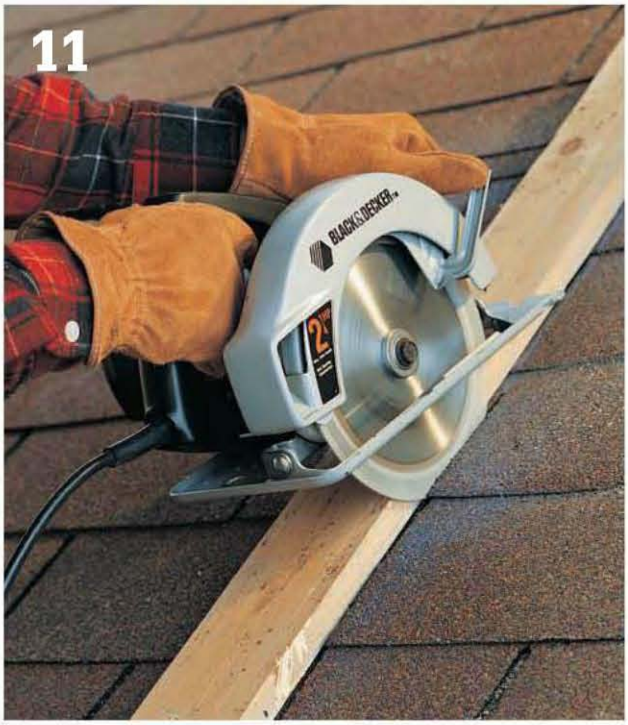

Cut through the shingles and sheathing along the chalk line using a circular saw and an old blade or a remodeling blade. Rest the saw foot on the 1 × 4, and use the edge of the board as a guide. Reposition the 1 × 4, and cut along the remaining lines. Remove the cutout roof section.

12

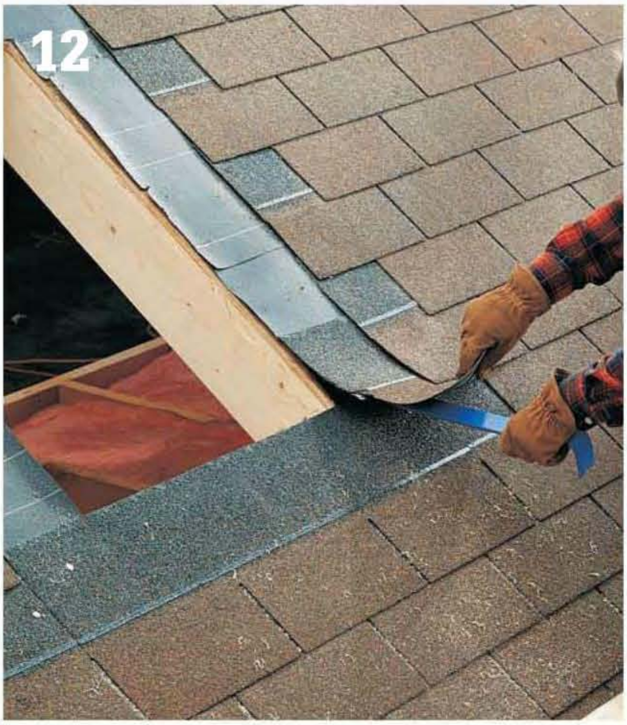

Remove the shingles around the rough opening with a flat pry bar, exposing at least 9" of building paper on all sides of the opening. Remove whole shingles rather than cutting them.

(continued)

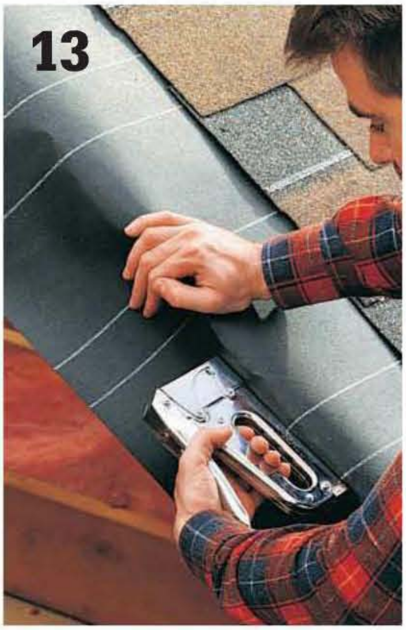

Cut strips of building paper and slide them between the shingles and existing building paper. Wrap the paper so that it covers the faces of the framing members, and staple it in place.

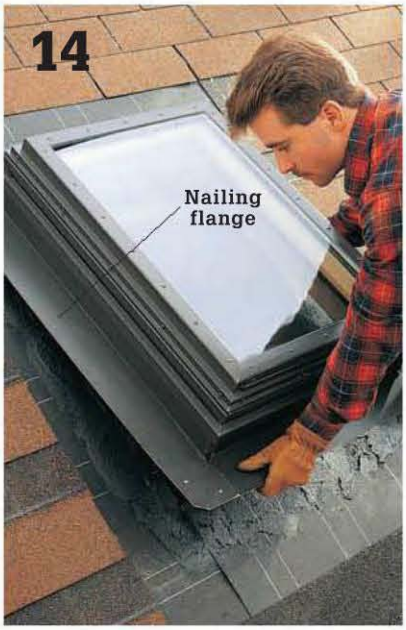

Spread a 5"-wide layer of roofing cement around the roof opening. Set the skylight into the opening so that the nailing flange rests on the roof. Finally, adjust the unit so it sits squarely in the opening.

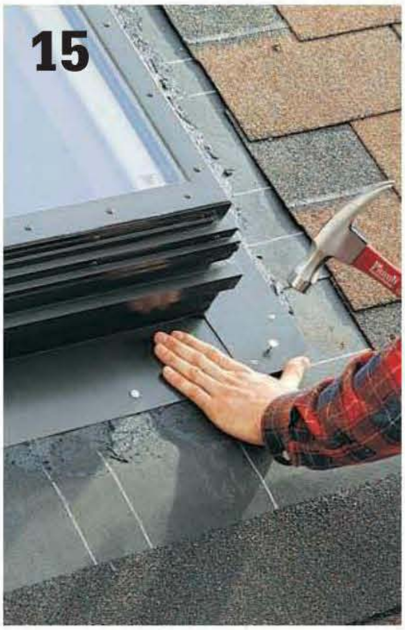

Nail through the flange and into the sheathing and framing members with 2" galvanized roofing nails spaced every 6". *Note: If skylight uses L-shaped brackets instead of a nailing flange, follow manufacturer's instructions.*

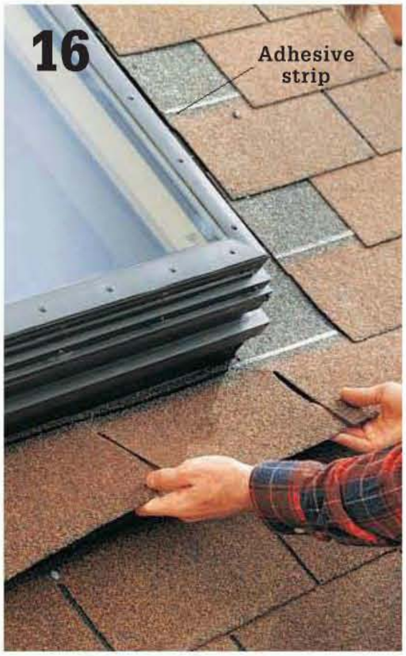

Patch in shingles up to the bottom edge of the skylight unit. Attach the shingles with 1¼" roofing nails driven just below the adhesive strip. If necessary, cut the shingles with a utility knife so they fit against the bottom of the skylight.

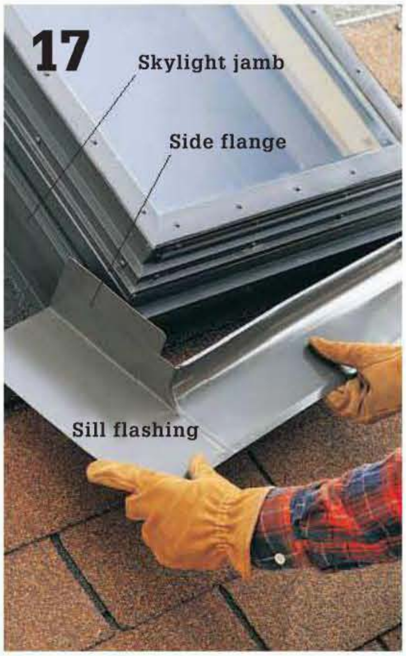

Spread roofing cement on the bottom edge of the sill flashing, then fit the flashing around the bottom of the unit. Attach flashing by driving ¾" galvanized roofing nails through the vertical side flange (near the top of the flashing) and into the skylight jambs.

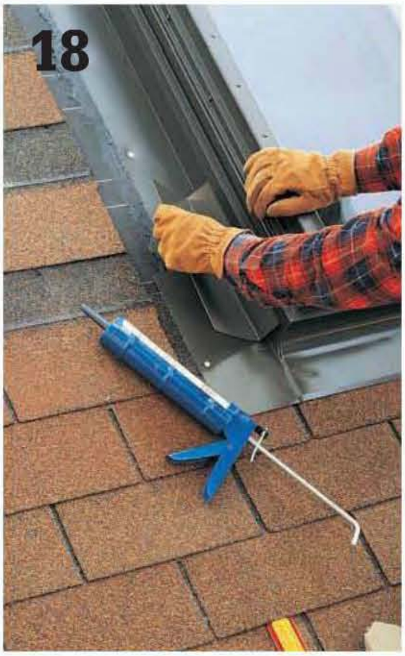

Spread roofing cement on the bottom of a piece of step flashing, then slide flashing under the drip edge on one side of the skylight. Step flashing should overlap sill flashing by 5". Press the step flashing down to bond it. Repeat on the opposite side of the skylight.

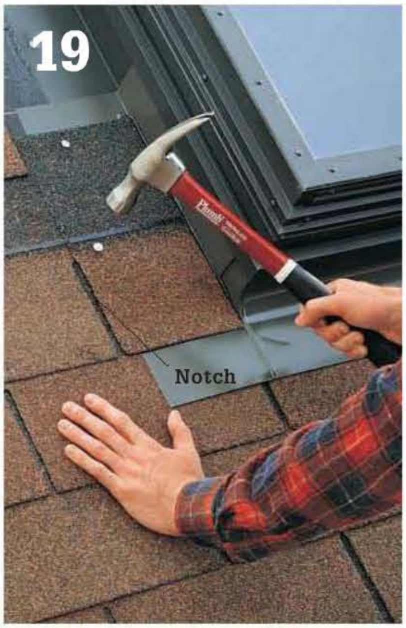

Patch in the next row of shingles on each side of the skylight, following the existing shingle pattern. Drive a 1¼" roofing nail through each shingle and the step flashing and into the sheathing. Drive additional nails just above the notches in the shingles.

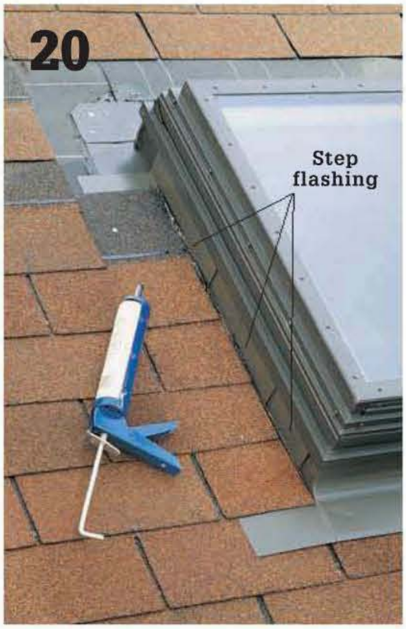

Continue applying alternate rows of step flashing and shingles using roofing cement and roofing nails. Each piece of flashing should overlap the preceding piece by 5".

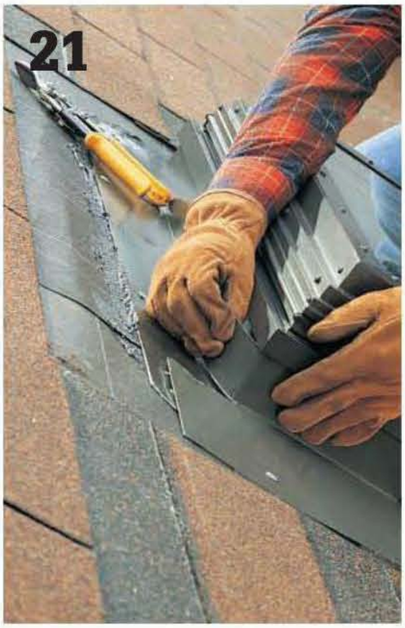

At the top of the skylight, cut and bend the last piece of step flashing on each side, so the vertical flange wraps around the corner of the skylight. Patch in the next row of shingles.

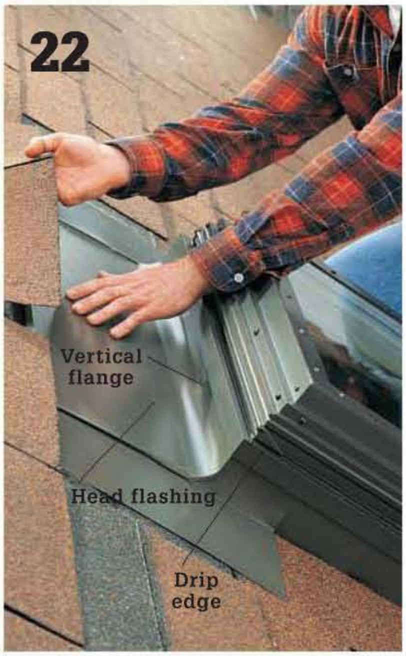

Spread roofing cement on the bottom of the head flashing to bond it to the roof. Place the flashing against the top of the skylight so the vertical flange fits under the drip edge and the horizontal flange fits under the shingles above the skylight.

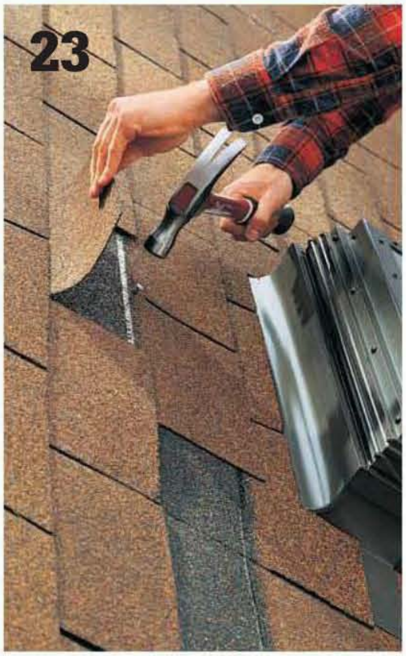

Fill in the remaining shingles, cutting them to fit, if necessary. Attach the shingles with roofing nails driven just above the notches.

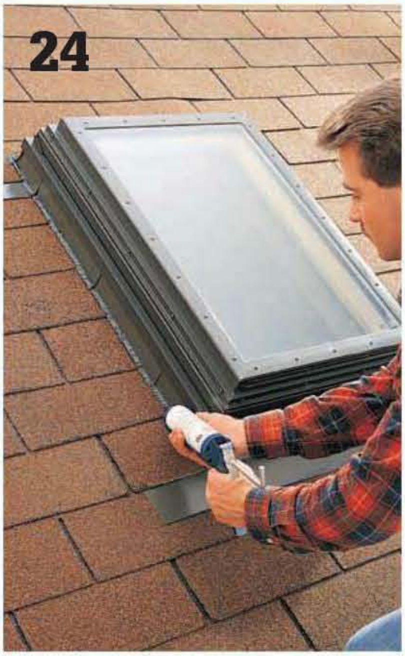

Apply a continuous bead of roofing cement along the joint between the shingles and the skylight. Finish the interior of the framed opening as desired.

How to Build a Skylight Shaft

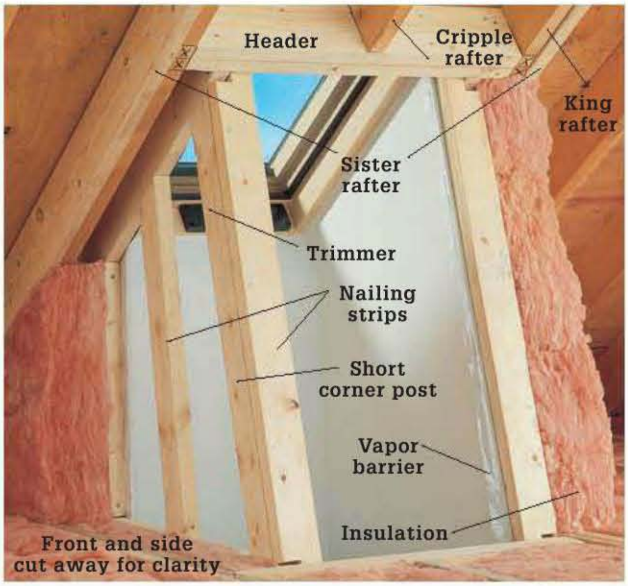
Header
Cripple rafter
King rafter
Sister rafter
Trimmer
Nailing strips
Short corner post
Vapor barrier
Insulation
Front and side cut away for clarity

A skylight shaft is made with 2 × 4 lumber and wallboard and includes a vapor barrier and fiberglass insulation. You can build a straight shaft with four vertical sides or an angled shaft that has a longer frame at ceiling level and one or more sides set at an angle. Since the ceiling opening is larger, an angled shaft lets in more direct light than a straight shaft.

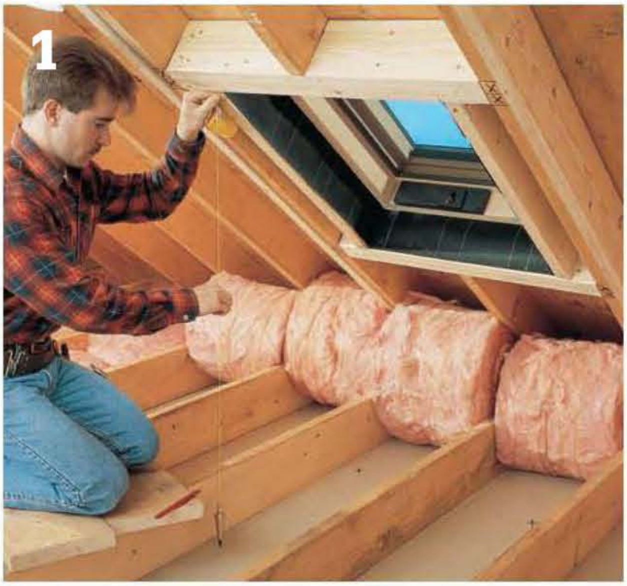

Remove any insulation in the area where the skylight will be located; turn off and reroute electrical circuits as necessary. Use a plumb bob as a guide to mark reference points on the ceiling surface, directly below the inside corners of the skylight frame.

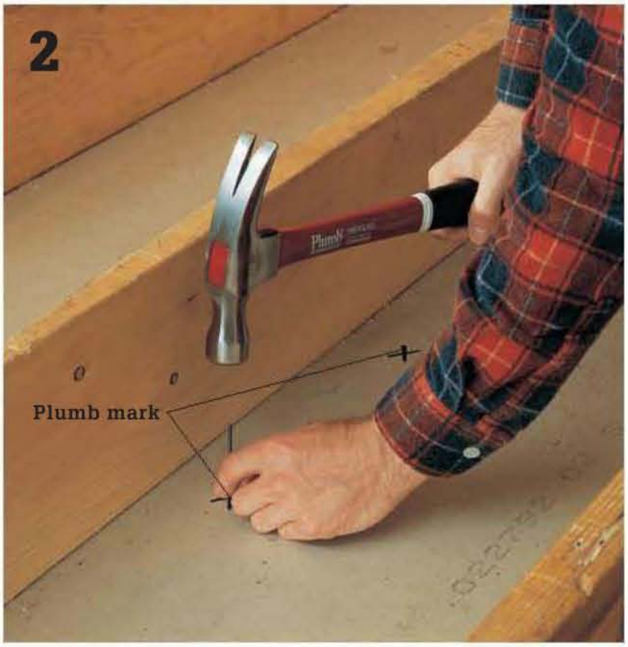
2
Plumb mark

If you are installing a straight shaft, use the plumb marks made in step 1 to define the corners of the ceiling opening; drive a finish nail through the ceiling surface at each mark. If you are installing an angled shaft, measure out from the plumb marks and make new marks that define the corners of the ceiling opening; drive finish nails at the new marks.

From the room below, mark cutting lines, then remove the ceiling surface.

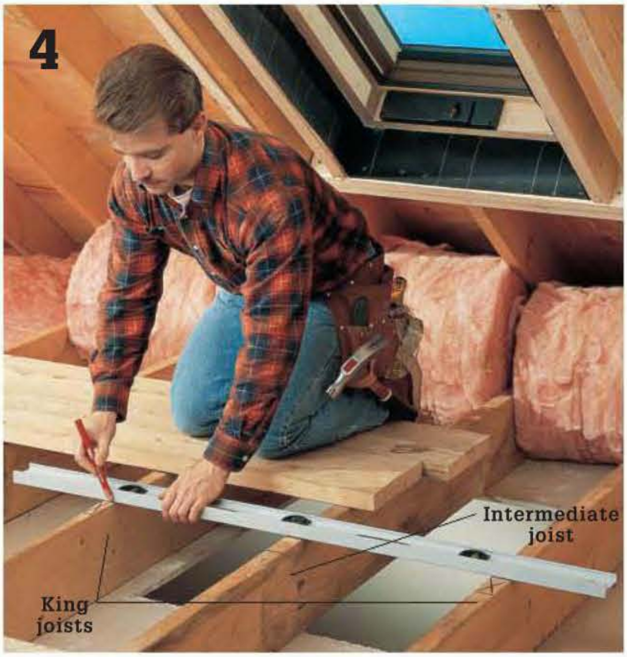

Use the nearest joists on either side of the ceiling opening to serve as king joists. Measure and mark where the double header and double sill will fit against the king joists and where the outside edge of the header and sill will cross any intermediate joists.

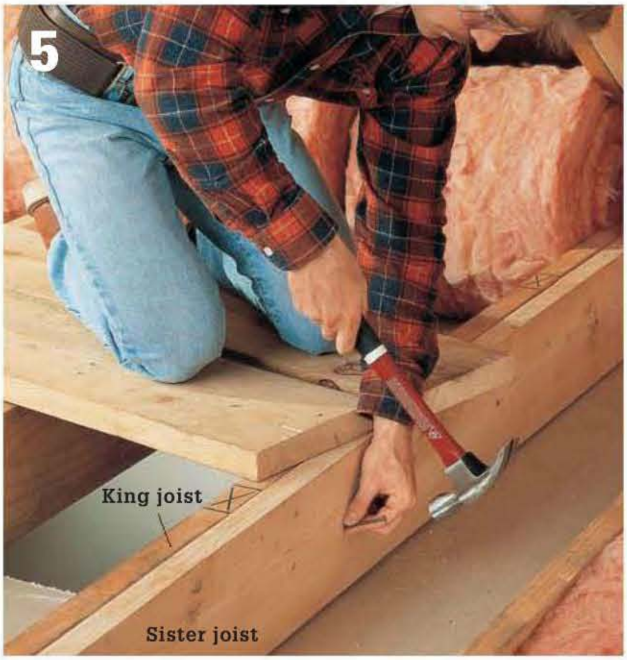

If you will be removing a section of an intermediate joist, reinforce the king joists by nailing full-length "sister" joists to the outside faces of the king joists using 10d nails.

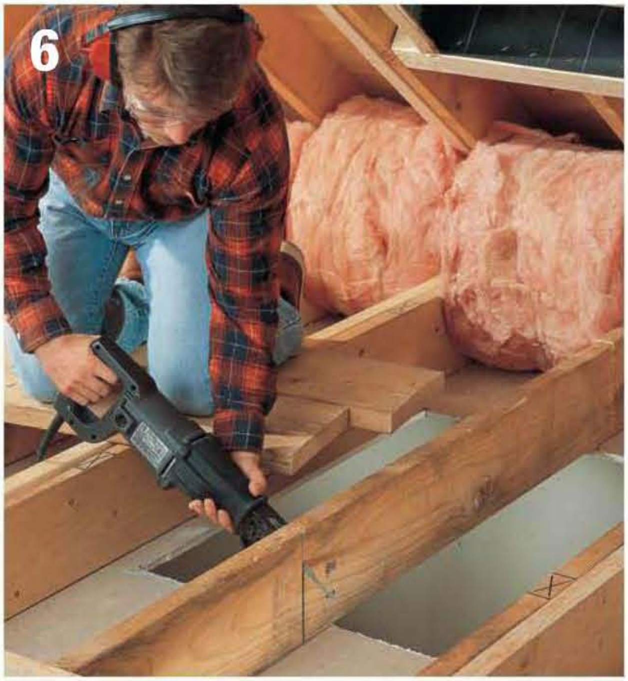

Install temporary supports below the project area to support the intermediate rafter on both sides of the opening. Use a combination square to extend cutting lines down the sides of the intermediate joist, then cut out the joist section with a reciprocating saw. Pry loose the cutout portion of the joist, being careful not to damage the ceiling surface.

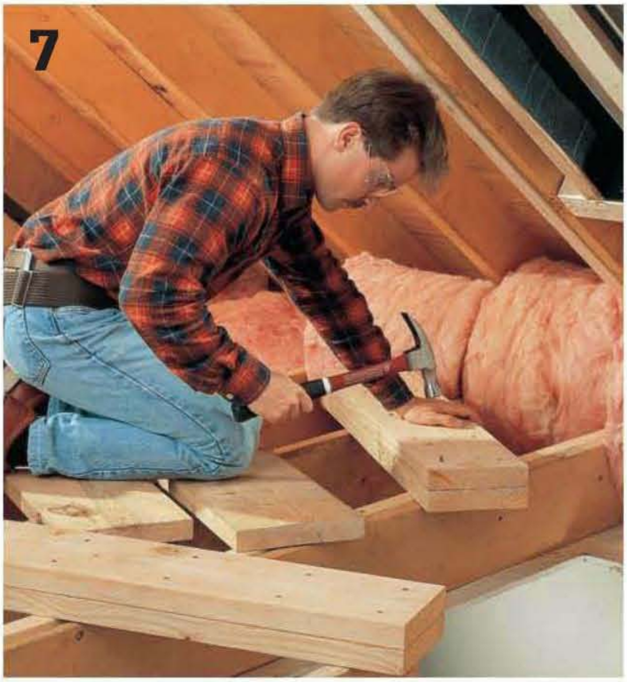

Build a double header and double sill to span the distance between the king joists using 2× dimensional lumber the same size as the joists.

(continued)

8

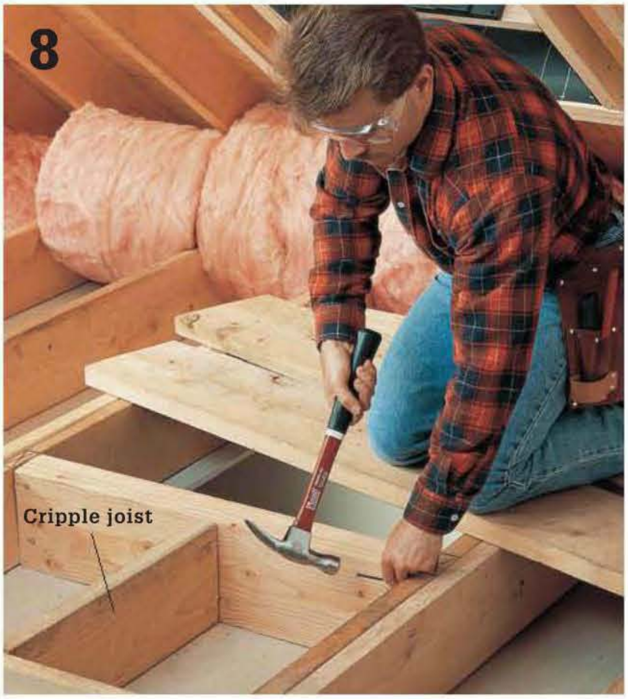
Cripple joist

Install the double header and double sill, anchoring them to the king joists and cripple joists with 10d nails. The inside edges of the header and sill should be aligned with the edge of the ceiling cutout.

9

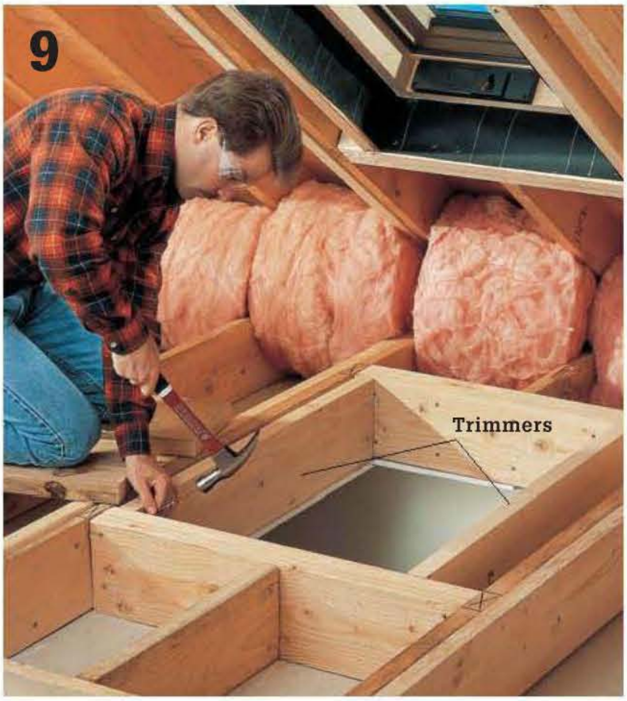

Complete the ceiling opening by cutting and attaching trimmers, if required, along the sides of the ceiling cutout between the header and sill. Toenail the trimmers to the header and sill with 10d nails.

10

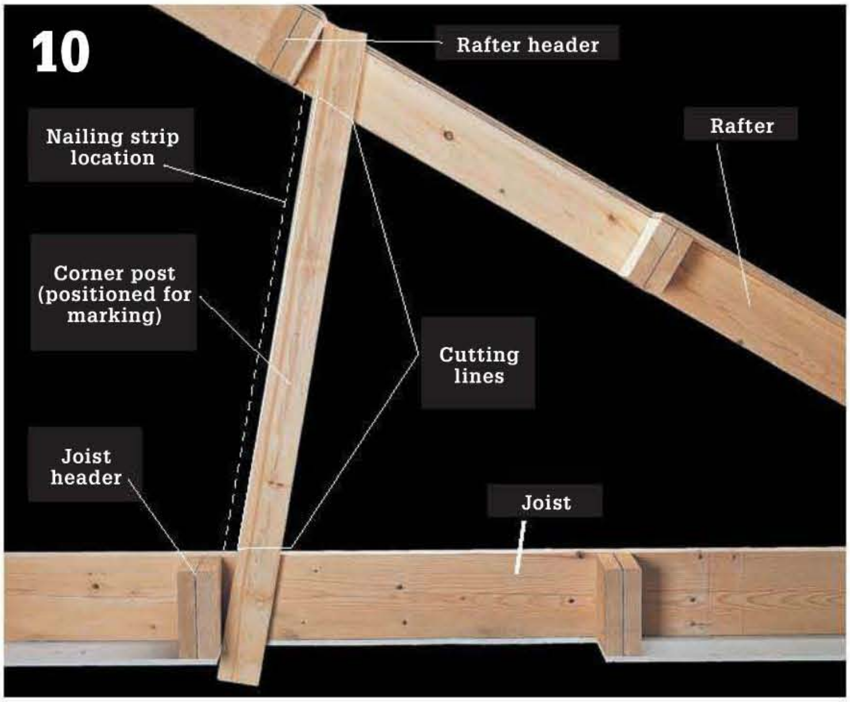
Rafter header
Nailing strip location
Rafter
Corner post (positioned for marking)
Cutting lines
Joist header
Joist

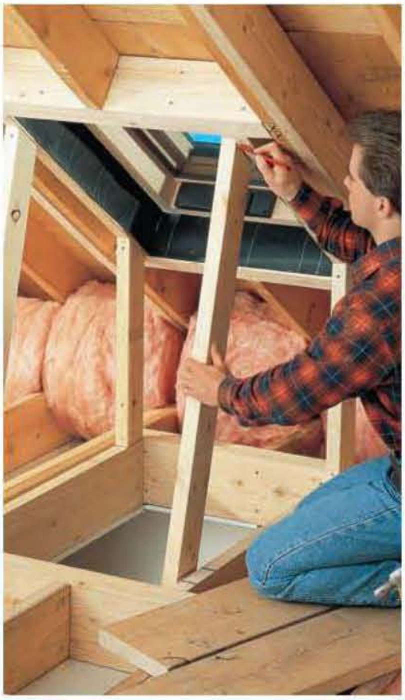

Install 2 × 4 corner posts for the skylight shaft. To measure for the posts, begin with a 2 × 4 that is long enough to reach from the top to the bottom of the shaft. Hold the 2 × 4 against the inside of the framed openings, so it is flush with the top of the rafter header and the bottom of the joist header (left photo). Mark cutting lines where the 2 × 4 meets the top of the joist or trimmer and the bottom of the rafter or trimmer (right photo). Cut along the lines, then toenail the posts to the top and bottom of the frame with 10d nails.

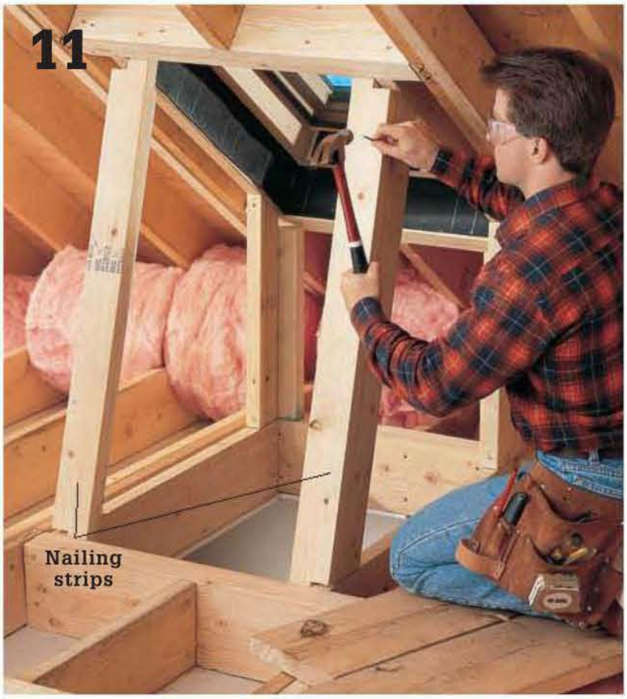
11
Nailing strips

Attach a 2 × 4 nailing strip to the outside edge of each corner post to provide a nailing surface for attaching the wallboard. Notch the ends of the nailing strips to fit around the trimmers; a perfect fit is not necessary.

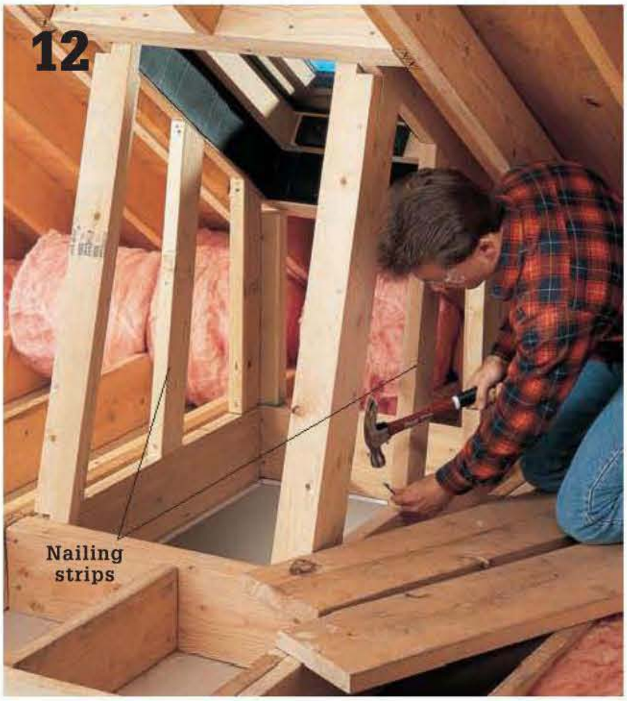
12
Nailing strips

Install additional 2 × 4 nailing strips between the corner posts if the distances between posts are more than 24". Miter the top ends of the nailing strips to fit against the rafter trimmers.

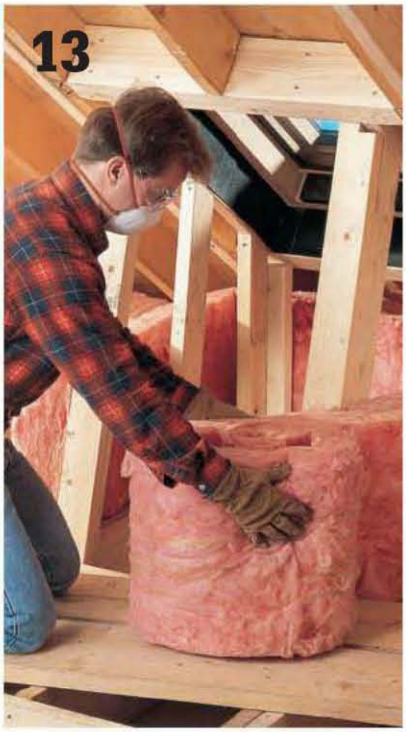

Wrap the skylight shaft with fiberglass insulation. Secure the insulation by wrapping twine around the shaft and insulation.

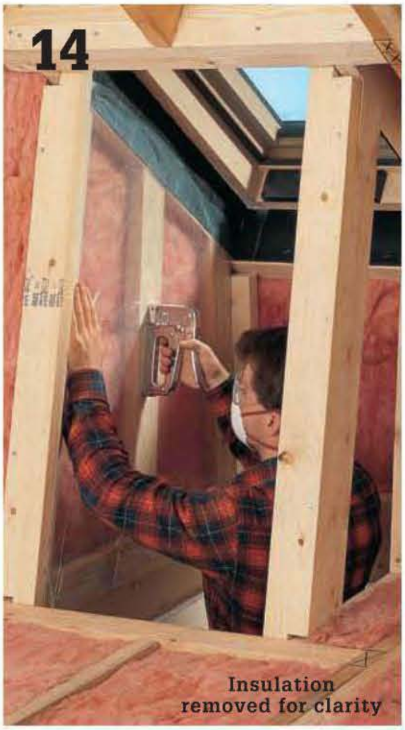
14
Insulation removed for clarity

From inside the shaft, staple a plastic vapor barrier of 6-mil polyethylene sheeting over the insulation.

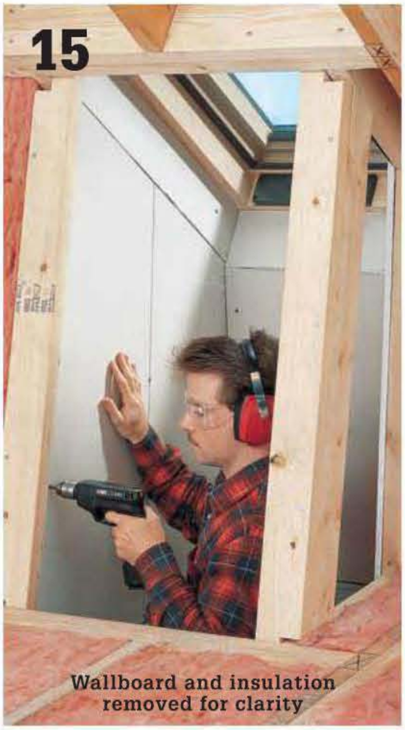
15
Wallboard and insulation removed for clarity

Finish the inside of the shaft with wallboard (pages 48 and 51). *Tip: To reflect light, paint the shaft interior with a light-colored, semigloss paint.*

Installing Tubular Skylights

Any interior room can be brightened with a tubular skylight. Tubular skylights are quite energy-efficient and are relatively easy to install, with no complicated framing involved.

The design of tubular skylights varies among manufacturers, with some using solid plastic reflecting tubes and others using flexible tubing. Various diameters are also available. Measure the distance between the framing members in your attic before purchasing your skylight to be sure it will fit.

This project shows the installation of a tubular skylight on a sloped, asphalt-shingled roof. Consult the dealer or manufacturer for installation procedures on other roof types.

A tubular skylight is an economical way to introduce more sunlight into a room without embarking on a major framing project.

Tools & Materials ▸

Pencil
Drill
Tape measure
Wallboard saw
Reciprocating saw
Pry bar
Screwdriver
Hammer
Wire cutters
Utility knife
Chalk
Tubular skylight kit
Stiff wire
2" roofing nails or flashing screws
Roofing cement

How to Install a Tubular Skylight

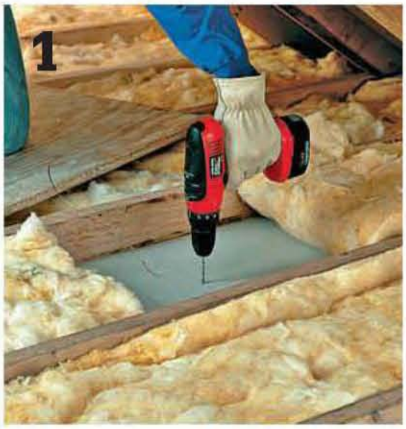

Drill a pilot hole through the ceiling at the approximate location for your skylight. Push a stiff wire up into the attic to help locate the hole. In the attic, make sure the space around the hole is clear of any insulation. Drill a second hole through the ceiling at the centerpoint between two joists.

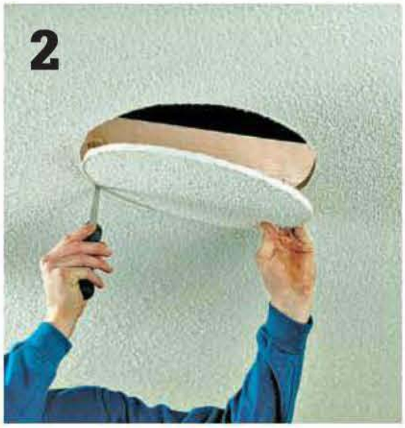

Center the ceiling ring frame over the hole and trace around it with a pencil. Carefully cut along the pencil line with a wallboard saw or reciprocating saw. Save the wallboard ceiling cutout to use as your roof-hole pattern. Attach the ceiling frame ring around the hole with the included screws.

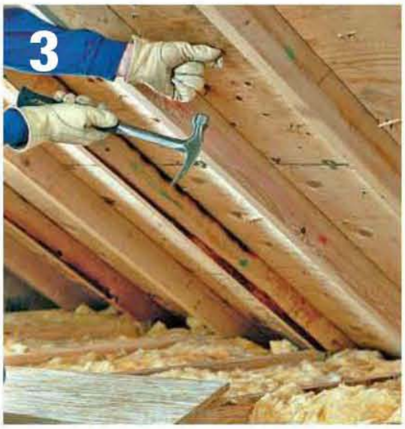

In the attic, choose the most direct route for the tubing to reach the roof. Find the center between the appropriate rafters and drive a nail up through the roof sheathing and shingles.

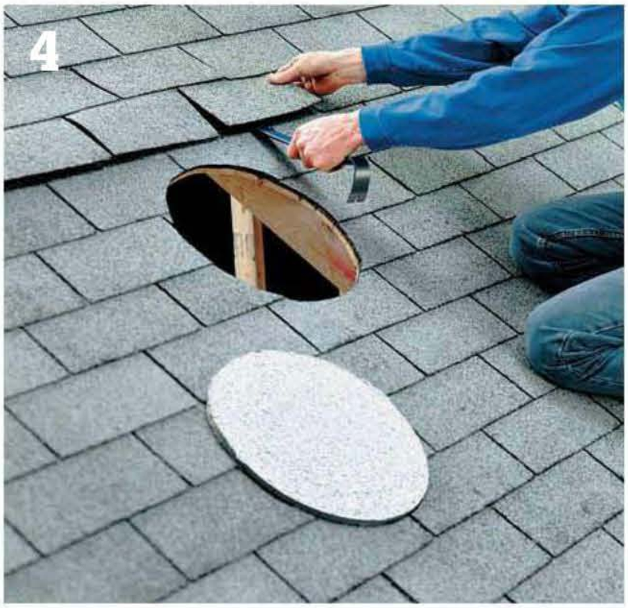

Use the wallboard ceiling cutout, centered over the nail hole, as a template for the roof opening. Trace the cutout onto the roof with chalk. Drill a starter hole to insert the reciprocating saw blade, then cut out the hole in the roof. Pry up the lower portion of the shingles above the hole. Remove any staples or nails around the hole edge.

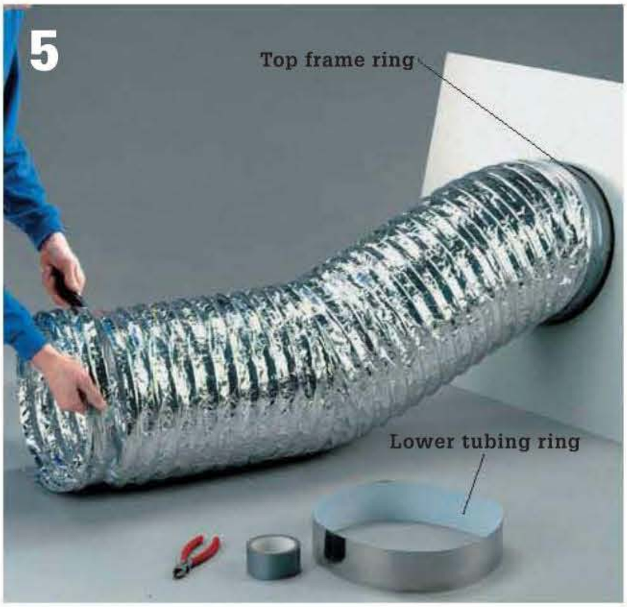

Pull the tubing over the top frame ring. Bend the frame tabs out through the tubing, keeping two or three rings of the tubing wire above the tabs. Wrap the junction three times around with included PVC tape. Then, in the attic, measure from the roof to the ceiling. Stretch out the tubing and cut it to length with a utility knife and wire cutters. Pull the loose end of tubing over the lower ring and wrap it three times with PVC tape.

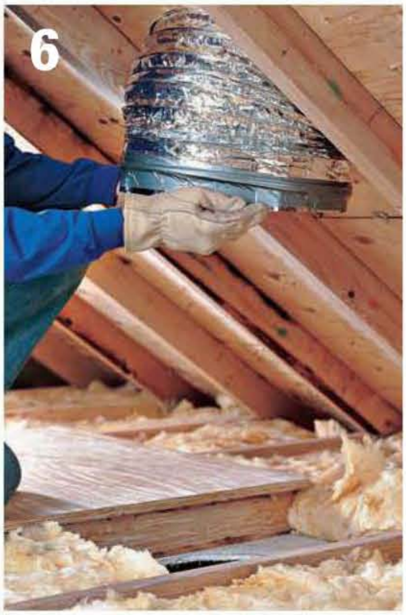

Lower the tubing through the roof hole and slide the flashing into place with the upper portion of the flashing underneath the existing shingles. This is easier with two people, one on the roof and one in the attic.

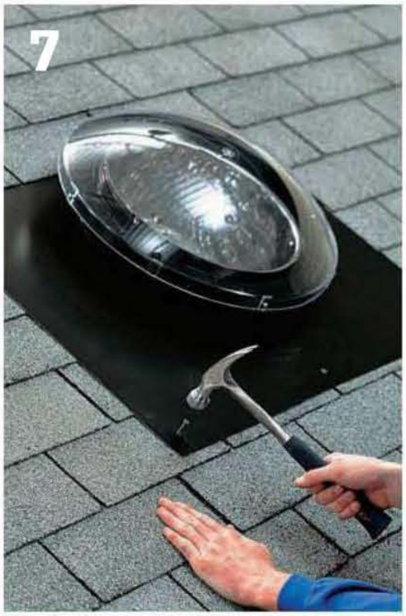

Secure the flashing to the roof with 2" roofing nails or flashing screws. Seal under the shingles and over all the nail heads with roofing cement. Attach the skylight dome and venting to the frame with the included screws.

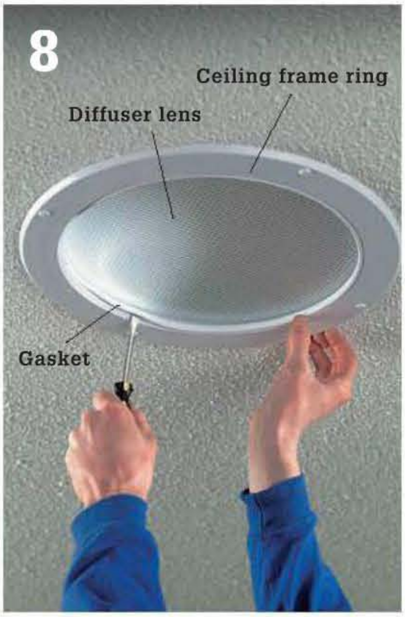

Pull the lower end of the tubing down through the ceiling hole. Attach the lower tubing ring to the ceiling frame ring and fasten it with screws. Attach the gasket to the diffuser lens and work the gasket around the perimeter of the ceiling frame. Repack any insulation around the tubing in the attic.

Installing Interior Doors

Creating an opening for a door in a wall involves building a framework about 1" wider and ½" taller than the door's jamb frame. This oversized opening, called a *rough opening,* will enable you to position the door easily and shim it plumb and level. Before framing a door, it's always a good idea to buy the door and refer to the manufacturer's recommendations for rough opening size.

Door frames consist of a pair of full-length king studs and two shorter jack studs that support the header above the door. A header provides an attachment point for wallboard and door casings. On load-bearing walls, it also helps to transfer the building's structural loads down into the wall framework and eventually the foundation.

Door framing requires flat, straight, and dry framing lumber, so choose your king, jack, and header pieces carefully. Sight down the edges and ends to look for warpage, and cut off the ends of pieces with splits.

Tools & Materials ▸

- Tape measure
- Framing square
- Hammer or nail gun
- Handsaw or reciprocating saw
- Framing lumber
- 10d or pneumatic framing nails
- $^3/_8$" plywood (for structural headers)
- Construction adhesive

Creating a square, properly sized opening for a door is the most important element of a successful door installation project.

How to Frame a Rough Opening for an Interior Prehung Door

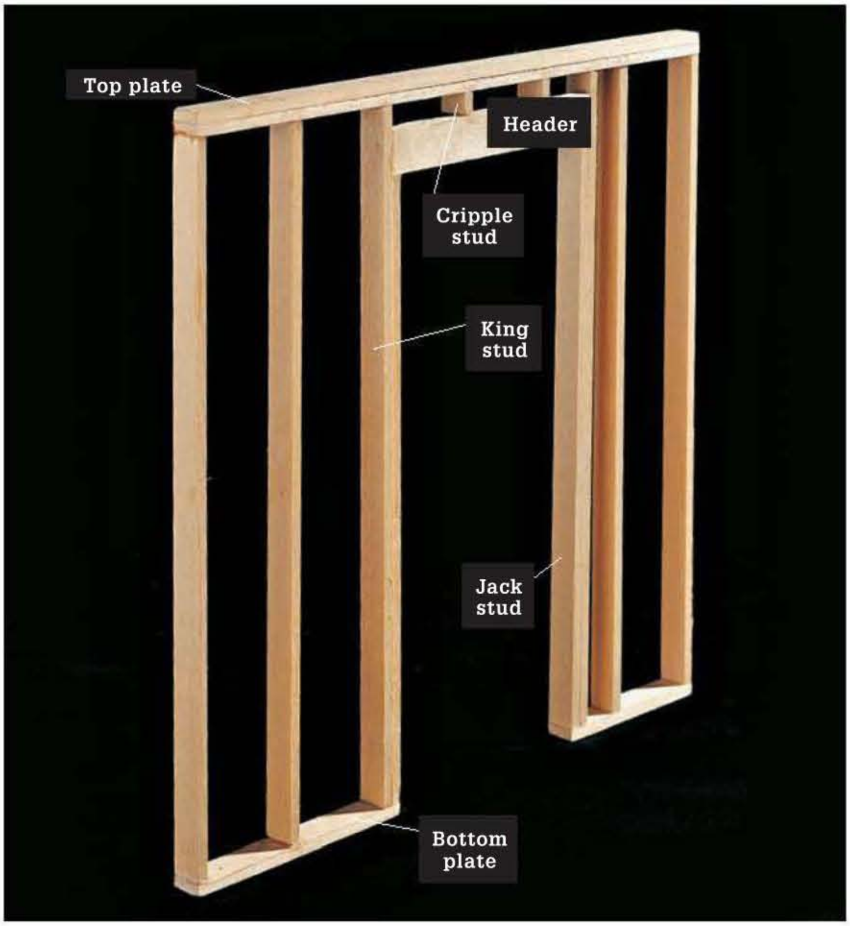
Top plate
Header
Cripple stud
King stud
Jack stud
Bottom plate

Door frames for prehung doors (left) start with king studs that attach to the top and bottom plates. Inside the king studs, jack studs support the header at the top of the opening. Cripple studs continue the wall-stud layout above the opening. In non-loadbearing walls, the header may be a 2 × 4 laid flat or a built-up header (below). The dimensions of the framed opening are referred to as the rough opening.

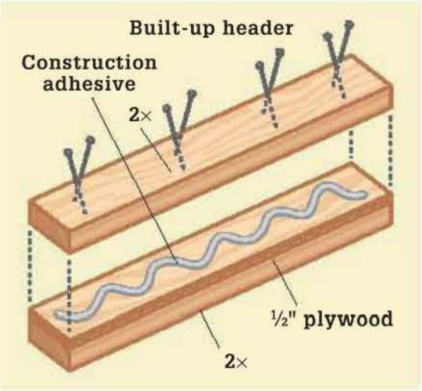
Built-up header
Construction adhesive
2×
½" plywood
2×

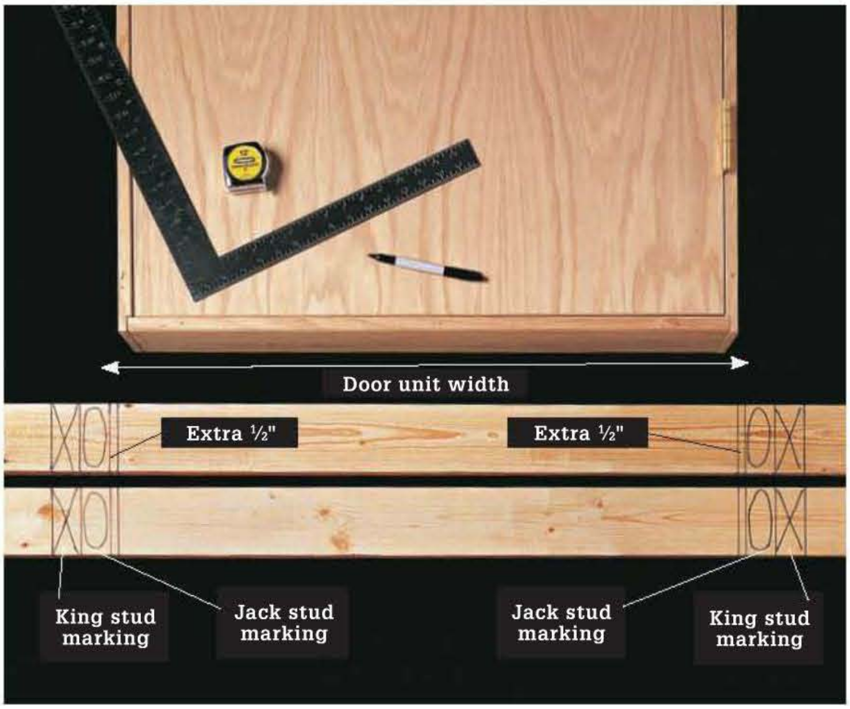
Door unit width
Extra ½"
Extra ½"
King stud marking
Jack stud marking
Jack stud marking
King stud marking

To mark the layout for the door frame, measure the width of the door unit along the bottom. Add 1" to this dimension to determine the width of the rough opening (the distance between the jack studs). This gives you a ½" gap on each side for adjusting the door frame during installation. Mark the top and bottom plates for the jack and king studs.

How to Frame a Prehung Door Opening

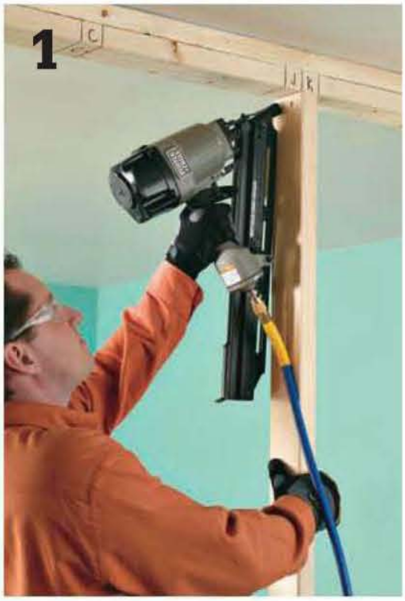

Mark layout lines for the king and jack studs on the wall's top and sole plates. Cut the king studs slightly longer than the distance between the wall plates, and toenail them in place with 10d nails or 3" pneumatic nails.

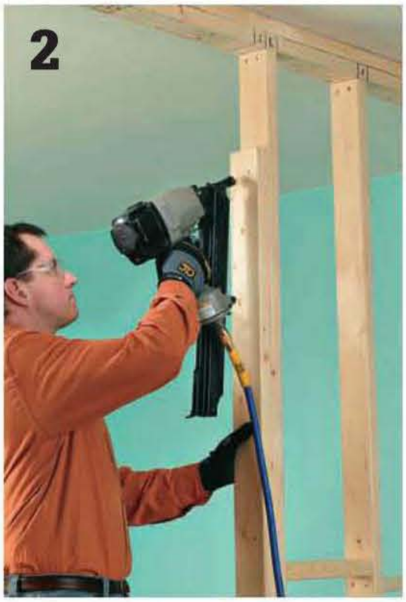

Cut the jack studs to length (they should rest on the sole plate). The height of a jack stud for a standard interior door is 83½", or ½" taller than the door. Nail the jack studs to the king studs.

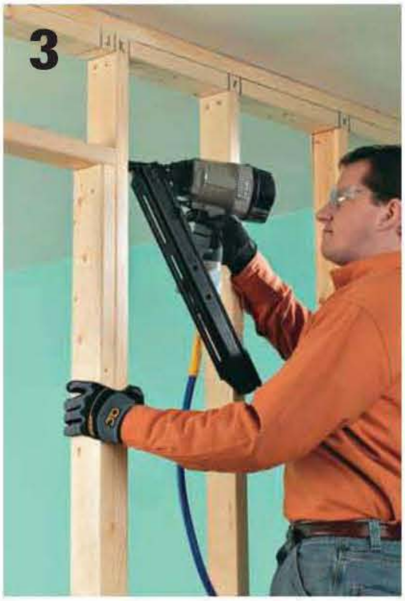

In a partition wall, the header can be a piece of 2× framing lumber that lays flat on top of the jack studs. Cut it to length, and install by endnailing through the king studs or down into the jack studs.

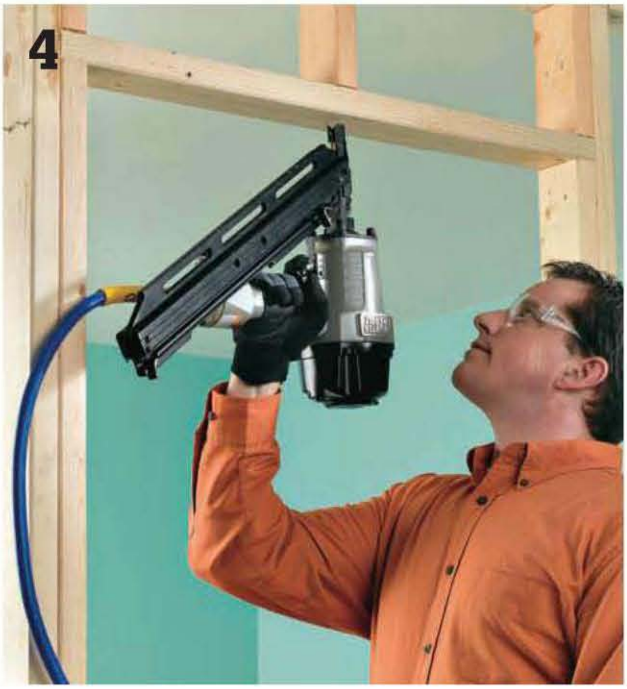

Fasten a cripple stud above the header, halfway between the king studs. It will prevent the header from warping. Toenail it into the top plate, and drive nails into it through the header.

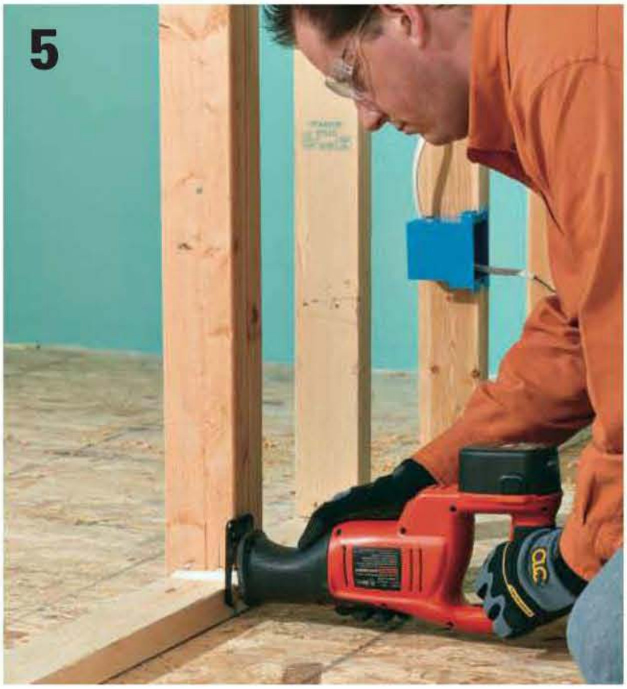

If you haven't cut a sole plate opening for the door yet, do that now with a reciprocating saw or handsaw. Trim the sole plate flush with the jack studs.

How to Frame an Opening for a Load-bearing Wall

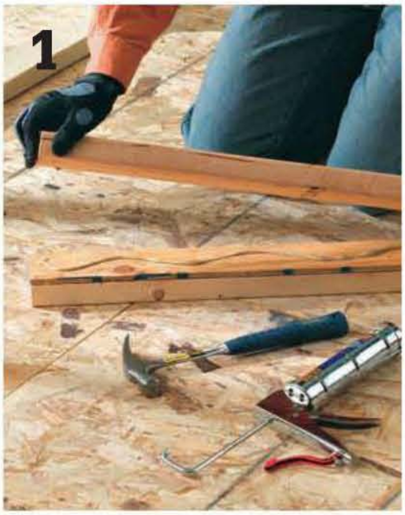

Door framing on load-bearing walls will require a structural header that transfers loads above the wall into the jack studs, sole plate, and down into the house foundation. Build it by sandwiching a piece of $^3/_8$" plywood between two 2 × 4s. Use construction adhesive and nails to fasten the header together.

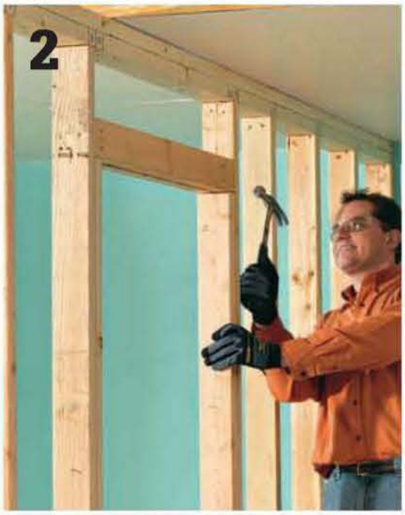

Install the built-up header by resting it on the jack studs and endnailing through the king studs. Use 10d nails or 3" pneumatic nails.

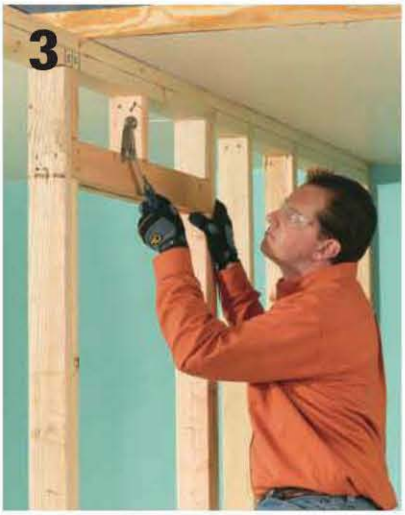

Toenail a cripple stud between the top plate and header, halfway between the king studs. It transfers structural loads into the header.

Option: Framing Openings for Sliding & Folding Doors

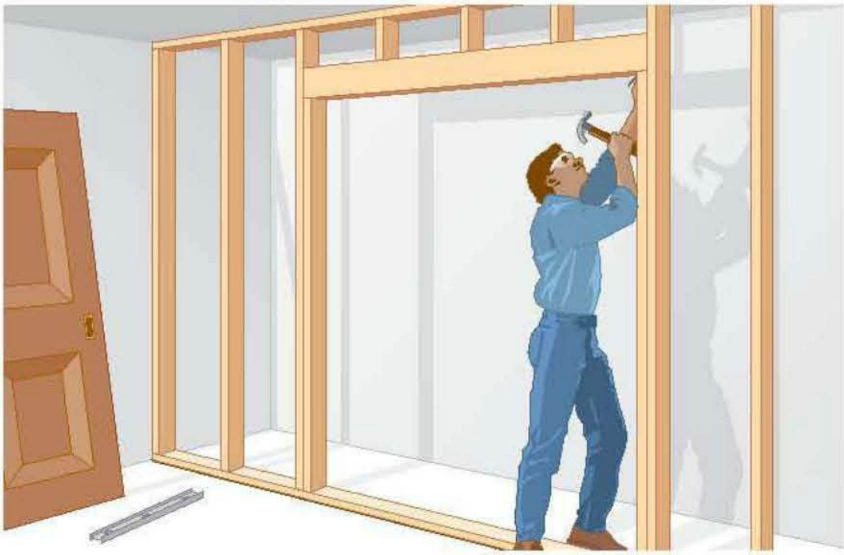

The same basic framing techniques are used, whether you're planning to install a sliding, bifold, pocket, or prehung interior door. The different door styles require different frame openings. You may need to frame an opening 2 to 3 times wider than the opening for a standard prehung door. Purchase the doors and hardware in advance, and consult the hardware manufacturer's instructions for the exact dimensions of the rough opening and header size for the type of door you select.

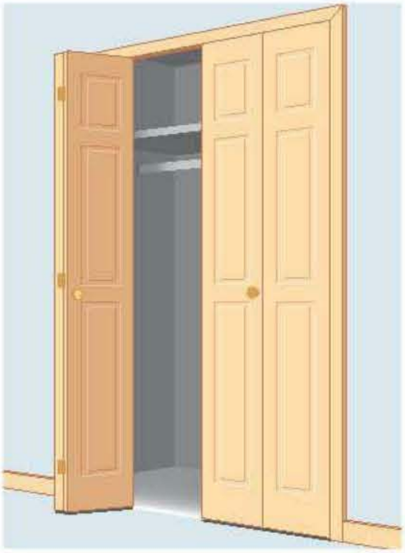

Most bifold doors are designed to fit in a 80"-high finished opening. Wood bifold doors have the advantage of allowing you to trim the doors, if necessary, to fit openings that are slightly shorter.

Installing a Prehung Interior Door

Install prehung interior doors after the framing work is complete and the wallboard has been installed. If the rough opening for the door has been framed accurately, installing the door takes about an hour.

Standard prehung doors have 4½"-wide jambs and are sized to fit walls with 2 × 4 construction and ½" wallboard. If you have 2 × 6 construction or thicker wall surface material, you can special-order a door to match, or you can add jamb extensions to a standard-sized door (photo, below).

Tools & Materials ▸

Level
Hammer
Handsaw

Prehung interior door
Wood shims
8d casing nails

Tip ▸

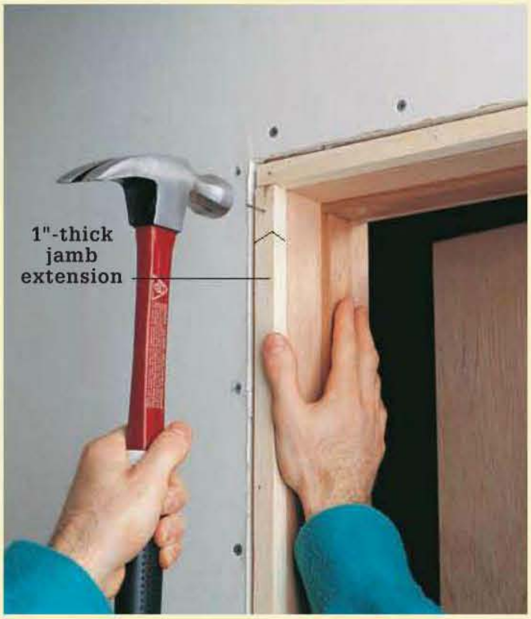

If your walls are built with 2 × 6 studs, you'll need to extend the jambs by attaching 1"-thick wood strips to the edges of the jamb after the door is installed. Use glue and 4d casing nails when attaching jamb extensions. Make the strips from the same wood as the jamb.

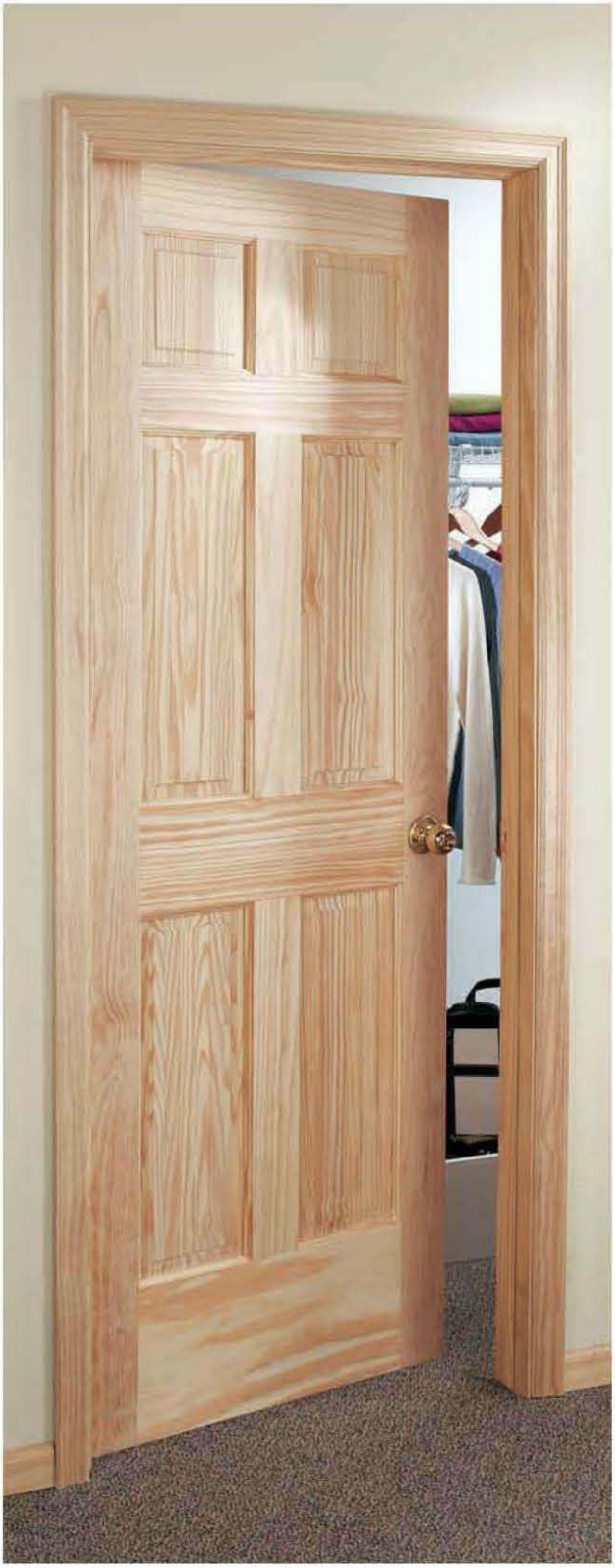

Prehung doors are shipped as single units with the door already hung on hinges attached to pre-installed jambs.

How to Install a Prehung Interior Door

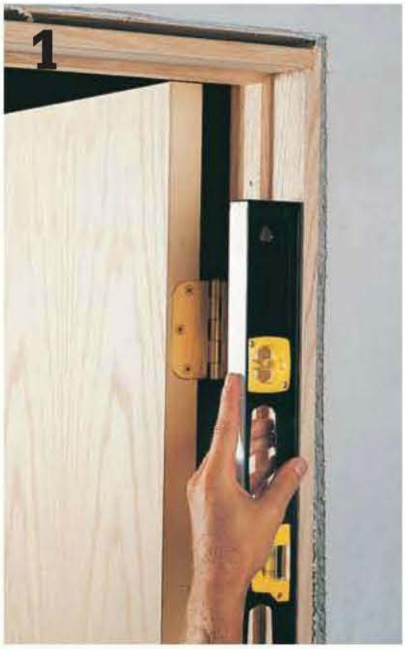

Slide the door unit into the framed opening so the edges of the jambs are flush with the wall surface and the hinge-side jamb is plumb.

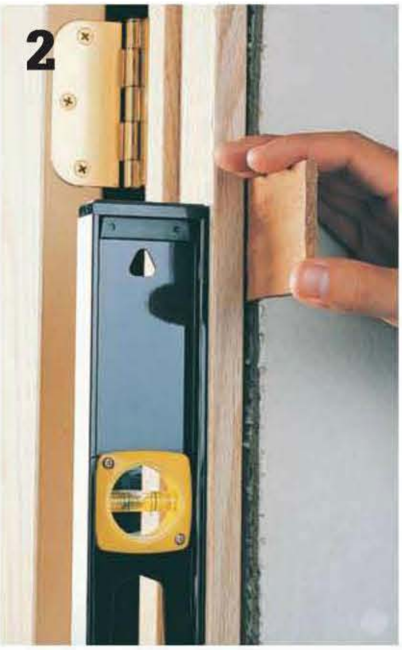

Insert pairs of wood shims driven from opposite directions into the gap between the framing members and the hinge-side jamb, spaced every 12". Check the hinge-side jamb to make sure it is still plumb and does not bow.

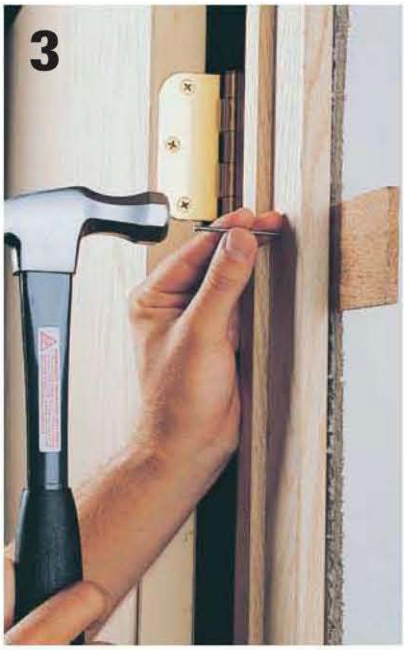

Anchor the hinge-side jamb with 8d casing nails driven through the jamb and shims and into the jack stud.

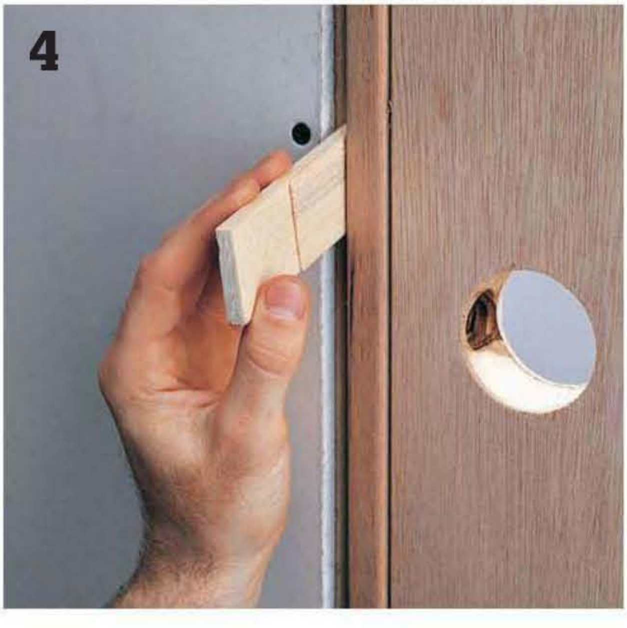

Insert pairs of shims in the gap between the framing members and the latch-side jamb and top jamb, spaced every 12". With the door closed, adjust the shims so the gap between door edge and jamb is ⅛" wide. Drive 8d casing nails through the jambs and shims, into the framing members.

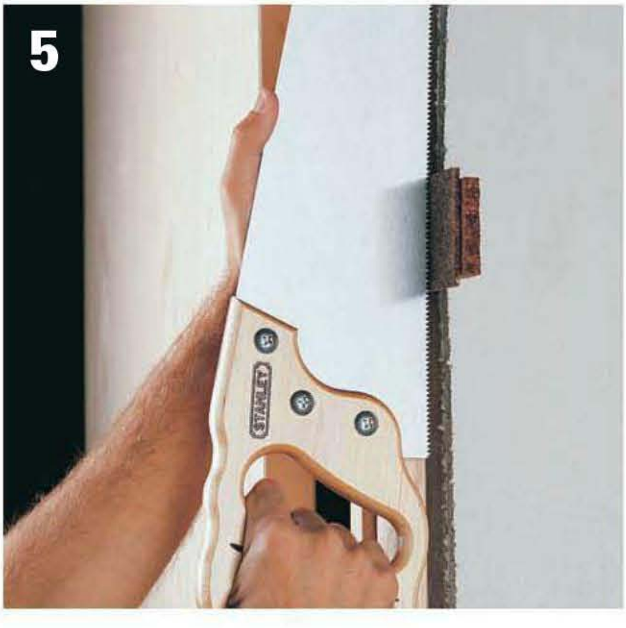

Cut the shims flush with the wall surface, using a handsaw. Hold the saw vertically to prevent damage to the door jamb or wall. Finish the door and install the lockset as directed by the manufacturer. See pages 318 to 321 to install trim around the door.

Pocket Doors

Pocket doors are a space-saving alternative to traditional hinged interior doors. Swinging doors can monopolize up to 16 square feet of floor space in a room, which is why pocket doors are a perfect choice for tight spaces, like small bathrooms. Installed in pairs, pocket doors can divide large rooms into more intimate spaces and can still be opened to use the entire area.

Pocket door hardware kits generally are universal and can be adapted for almost any interior door. In this project, the frame kit includes an adjustable track, steel-clad split studs, and all the required hanging hardware. The latch hardware, jambs, and the door itself are all sold separately. Pocket door frames can also be purchased as preassembled units that can be easily installed into a rough opening.

Framing and installing a pocket door is not difficult in new construction or a major remodel. But retrofitting a pocket door in place of a standard door or installing one in a wall without an existing door, is a major project that involves removing the wall material, framing the new opening, installing and hanging the door, and refinishing the wall. Hidden utilities, such as wiring, plumbing, and heating ducts, must be rerouted if encountered.

The rough opening for a pocket door is at least twice the width of a standard door opening. If you are installing the pocket door in a partition wall, see page 292 to learn how to frame the opening. If the wall is load bearing, you will need to install an appropriately sized header (page 293).

Because pocket doors are easy to open and close and require no threshold, they offer increased accessibility for wheelchair or walker users, provided the handles are easy to use (page 299). If you are installing a pocket door for this purpose, be aware that standard latch hardware may be difficult to use for some individuals.

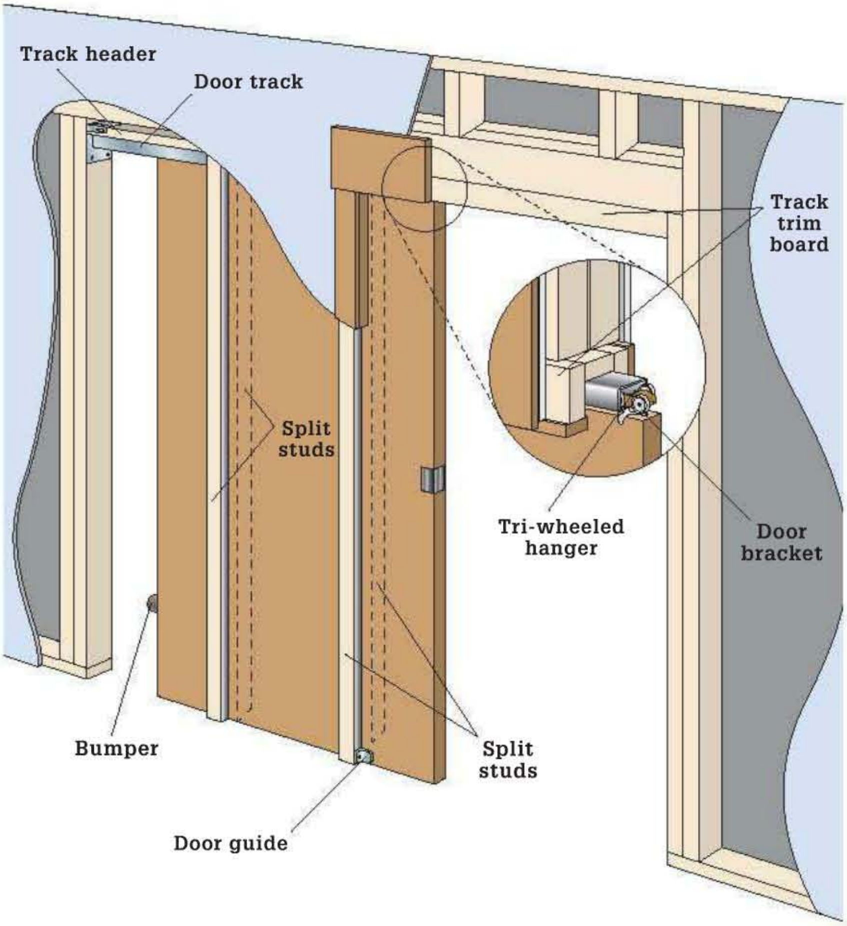

Tools & Materials ▸

- Tape measure
- Circular saw
- Hammer, nail set
- Screwdriver
- Level
- Drill
- Handsaw
- Hacksaw
- Wallboard tools
- 2 × 4 lumber
- 16d, 8d & 6d common nails
- Pocket door frame kit
- Door
- 1¼" wallboard screws
- Wallboard materials
- Manufactured pocket door jambs (or build jambs from 1× material)
- 8d & 6d finish nails
- 1½" wood screws
- Door casing

How to Install a Pocket Door

Prepare the project area and frame the rough opening to the manufacturer's recommended dimensions. Measuring from the floor, mark each jack stud at the height of the door plus ¾ to 1½" (depending on the door clearance above the floor) for the overhead door track. Drive a nail into each jack stud, centered on the mark. Leave about ⅛" of the nail protruding.

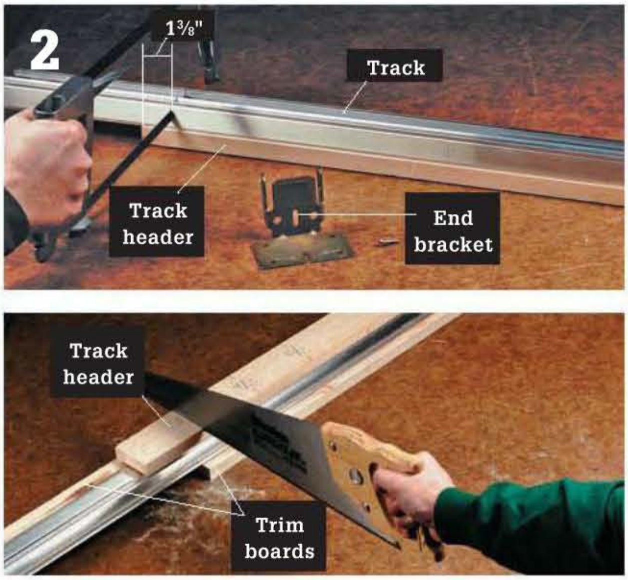

Remove the adjustable end bracket from the overhead door track. Cut the wooden track header at the mark that matches your door size. Turn the track over and cut the metal track 1⅜" shorter than the wooden track header using a hacksaw (top). Replace the end bracket. Cut the side trim boards along the marks corresponding to your door size, being careful not to cut the metal track (bottom).

Set end brackets of the track on the nails in the jack studs. Adjust the track to level and set the nails. Then drive 8d common nails through the remaining holes in the end brackets.

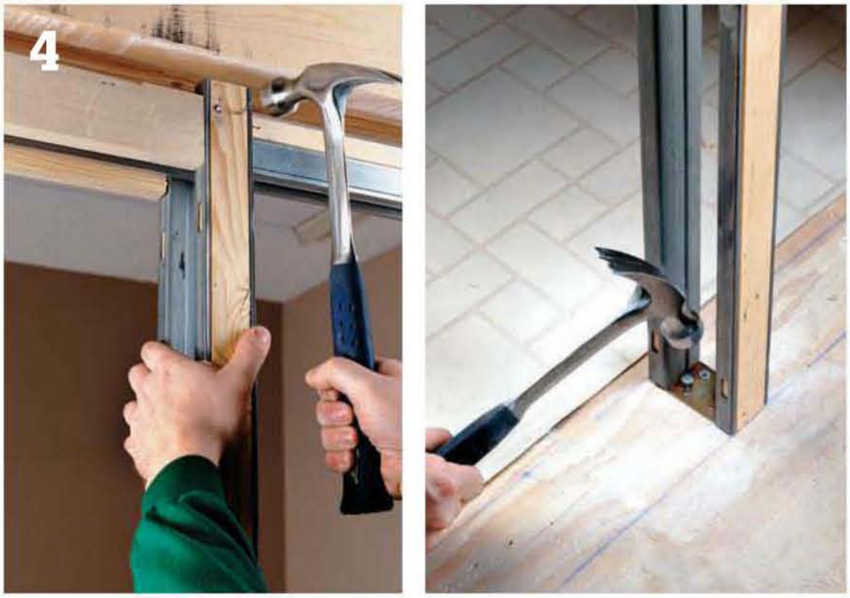

Snap chalk lines on the floor across the opening, even with the sides of the rough opening. Tap the floor plate spacers into the bottom ends of the pairs of steel-clad split studs. Butt one split stud pair against the door track trim board, check it for plumb, and fasten it to the track header using 6d common nails (left). Center the other split stud pair in the "pocket" and fasten it to the track header. Plumb the split studs again and attach them to the floor with 8d common nails or 2" screws driven through spacer plates (right).

(continued)

Cover the open framing with wallboard to the edge of the opening. You may want to leave the wallboard off one side of the wall to allow for door adjustment. Use 1¼" wallboard screws, which will not protrude into the pocket.

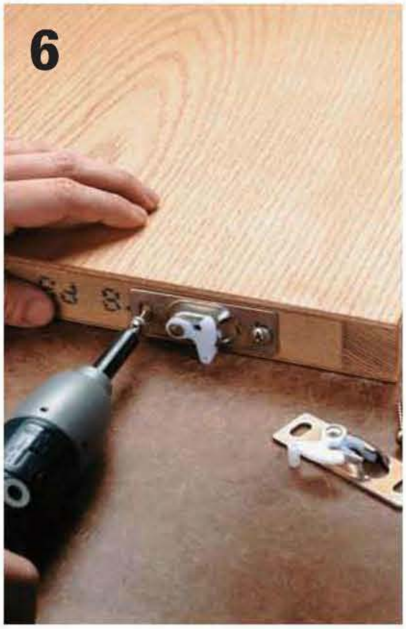

Paint or stain the door as desired. When the door has dried, attach two door brackets to the top of the door, using included screws driven through pilot holes. Install the rubber bumper to the rear edge of the door with its included screw.

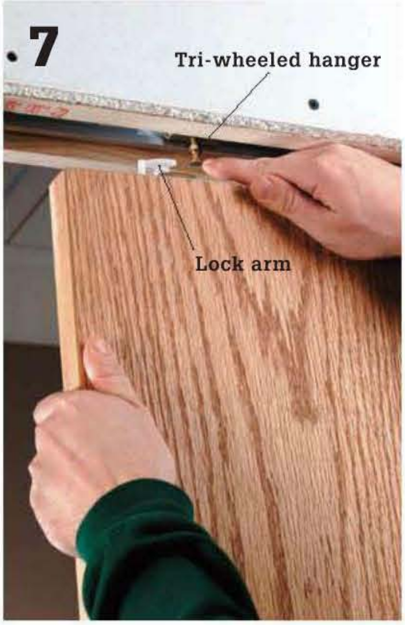

Slide two tri-wheeled hangers into the overhead door track. Set the door in the frame, aligning the hangers with the door brackets. Then raise the door and press each hanger into the door bracket until it snaps into place. Close the lock arm over the hanger.

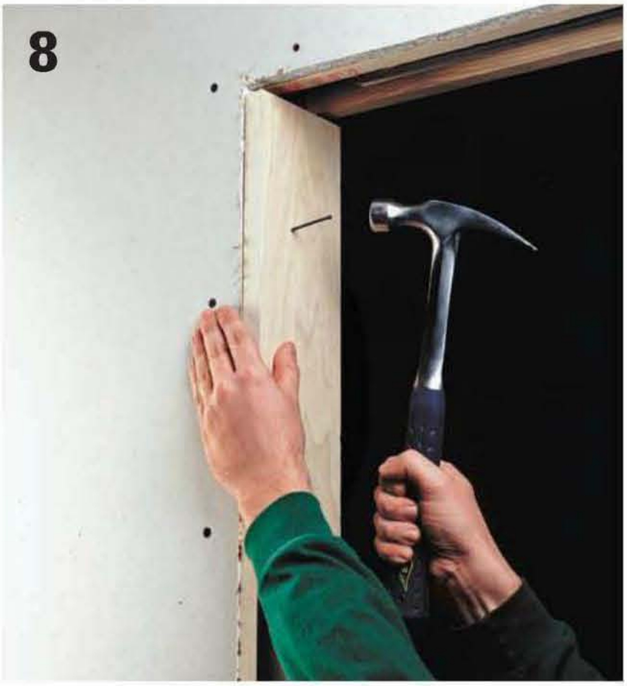

Cut the strike-side jamb to length and width. Fasten it to the jack stud using 8d finish nails, shimming the jamb to plumb as necessary. Close the door and adjust the hanger nuts to fine-tune the door height so the door is parallel with the jamb from top to bottom.

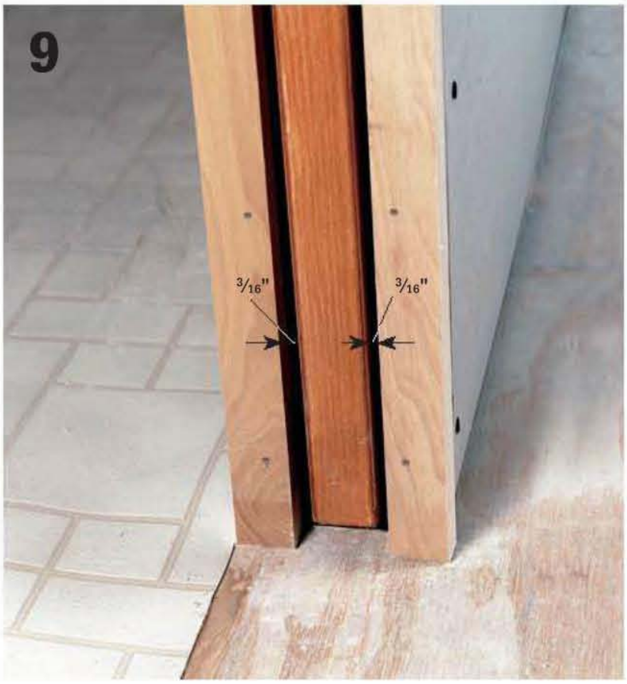

Measure and cut the split jambs to size. Fasten each split jamb to the front edge of the split stud using 8d finish nails. Maintain $\frac{3}{16}$" clearance on both sides of the door. If necessary, shim between the bumper and door until the door is flush with the jambs when open.

Measure and cut the split head jambs to size. Use 1½" wood screws driven through countersunk pilot holes to attach the head jamb on the side that has access to the lock arm of the hangers to allow for easy removal of the door. Attach the other head jamb using 6d finish nails. Maintain 3⁄16" clearance on each side of the door.

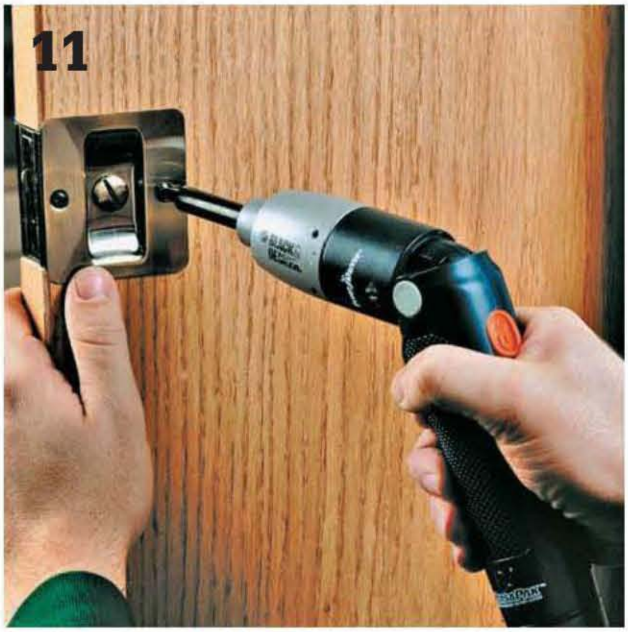

Install the included door guides on both sides of the door near the floor at the mouth of the pocket. Install the latch hardware according to the manufacturer's directions. Finish the wallboard and install casing around the door. Fill all nail holes, then paint or stain the jambs and casing as desired.

Improving Pocket Door Accessibility ▸

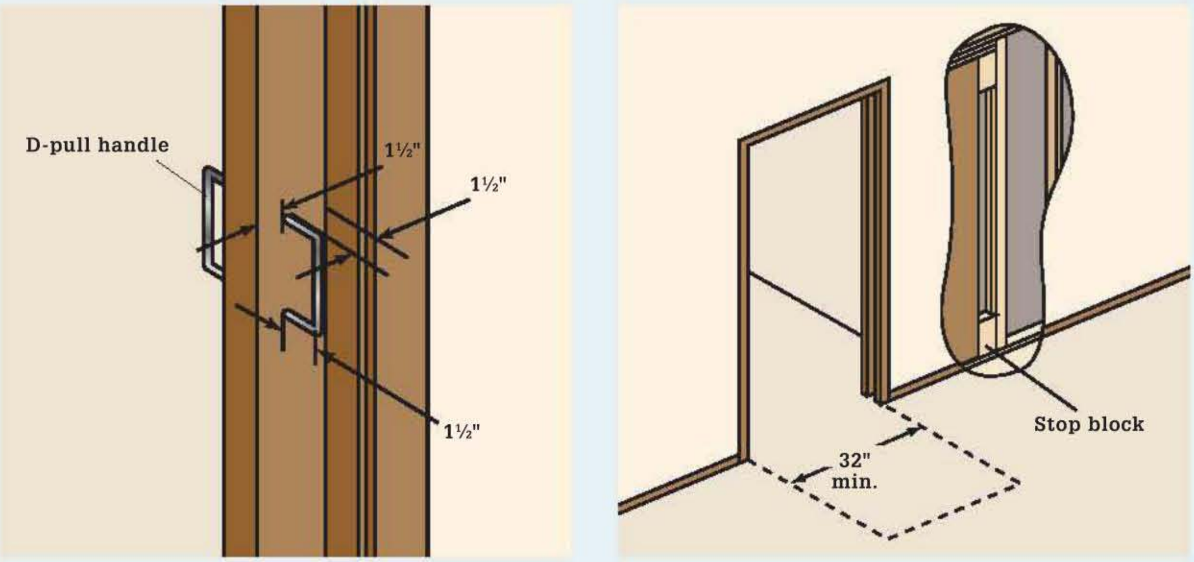

D-pull handles are easier to use than standard recessed hardware. Choose pulls that project at least 1½" from the door. Mount the pulls 1½" from the edge of the door to provide room for fingers when the door is closed (left). Install a stop block at the back of the frame (right), so the door stops 1½" short of the D-pull to provide room for fingers when the door is open. Because this design reduces the width of the door opening by 3", you must use a 36"-wide door to maintain the recommended doorway width of 32".

Bifold Doors

Bifold doors provide easy access to a closet without requiring much clearance for opening. Most home centers stock kits that include two pairs of prehinged doors, a head track, and all the necessary hardware and fasteners. Typically, the doors in these kits have predrilled holes for the pivot and guide posts. Hardware kits are also sold separately for custom projects. There are many types of bifold door styles, so be sure to read and follow the manufacturer's instructions for the product you use.

Tools & Materials ▸

Tape measure
Level
Circular saw
Straightedge (optional)
Drill
Plane
Screwdriver
Hacksaw
Prehinged bifold doors
Head track
Mounting hardware
Panhead screws
Flathead screws

A variety of designer bifold doors are available for installation between rooms and closets. They provide the same attractive appearance as French doors but require much less floor space.

How to Install Bifold Doors

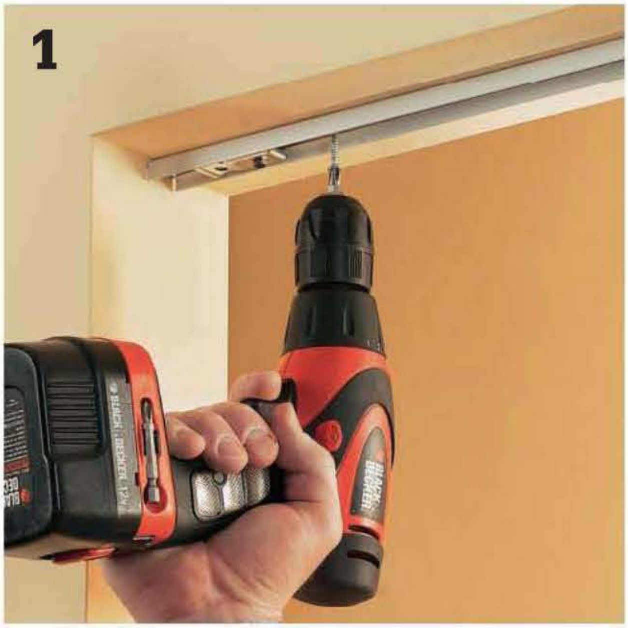

Cut the head track to the width of the opening using a hacksaw. Insert the roller mounts into the track, then position the track in the opening. Fasten it to the header using panhead screws.

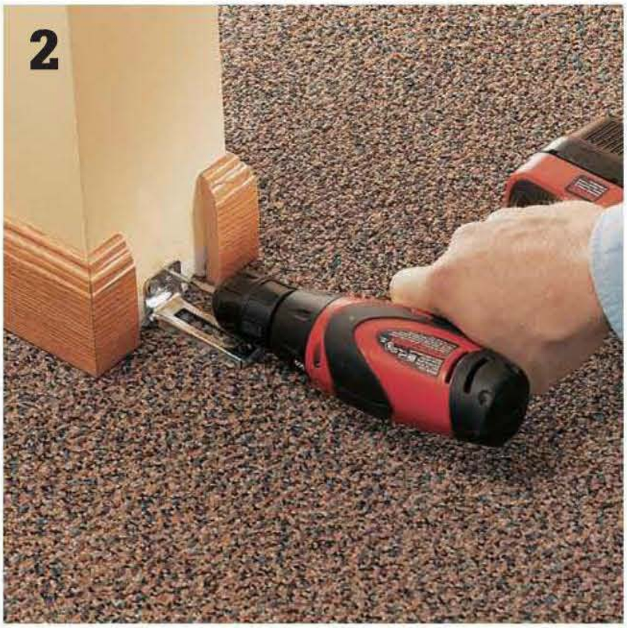

Measure and mark each side jamb at the floor for the anchor bracket so the center of the bracket aligns exactly with the center of the head track. Fasten the brackets in place with flathead screws.

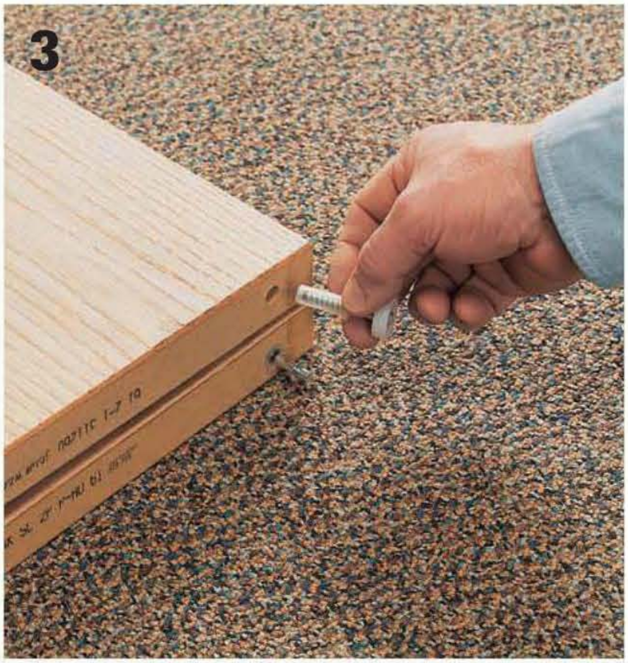

Check the height of the doors in the opening, and trim if necessary. Insert pivot posts into predrilled holes at the bottoms and tops of the doors. Insert guide posts at the tops of the leading doors. Make sure all posts fit snugly.

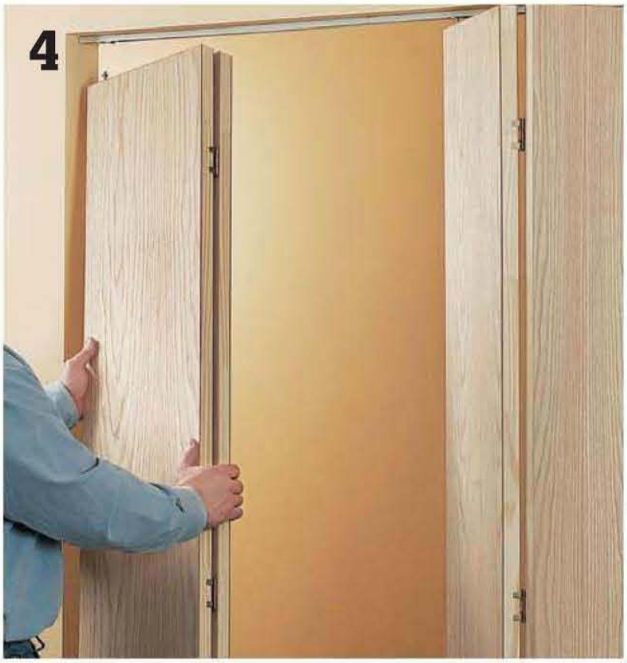

Fold one pair of doors closed and lift into position, inserting the pivot and guide posts into the head track. Slip the bottom pivot post into the anchor bracket. Repeat for the other pair of doors. Close the doors and check alignment along the side jambs and down the center. If necessary, adjust the top and bottom pivots following the manufacturer's instructions.

French Doors

French doors are made up of two separate doors, hinged on opposing jambs of a doorway. The doors swing out from the center of the doorway and into or out from a room. Like most doors, French doors are typically sold in prehung units, but are also available separately. They are generally available only in wood with a variety of designs and styles to choose from.

Before purchasing a prehung French door unit, determine the size of doors you will need. If you are planning to install the doors in an existing doorway, measure the dimensions of the rough opening from the unfinished framing members, then order the unit to size—the manufacturer or distributor will help you select the proper unit.

You can also pick the prehung unit first, then alter an existing opening to accommodate it (as shown in this project). In this case, build the rough opening a little larger than the actual dimensions of the doors to accommodate the jambs. Prehung units typically require adding 1" to the width and ½" to the height.

If the doorway will be in a load-bearing wall, you will need to make temporary supports (page 29) and install an appropriately sized header. Sizing the header (depth) is critical: it's based on the length of the header, the material it's made from, and the weight of the load it must support. For actual requirements, consult your local building department.

When installing French doors, it is important to have consistent reveals between the two doors and between the top of the doors and the head jamb. This allows the doors to close properly and prevents the hinges from binding.

Tools & Materials ▸

- Tape measure
- Circular saw
- 4-ft. level
- Hammer
- Handsaw
- Drill
- Utility knife
- Nail set
- Prehung French door unit
- 2 × 4 and 2 × 6 lumber
- ½" plywood
- 10d & 16d common nails
- Wood shims
- 8d finish nails

Traditionally, French doors open onto the patio or lush garden of a backyard. But you can create stylish entrances inside your home by bringing French doors to formal dining rooms, sitting rooms, dens, and master suites.

How to Install French Doors

1

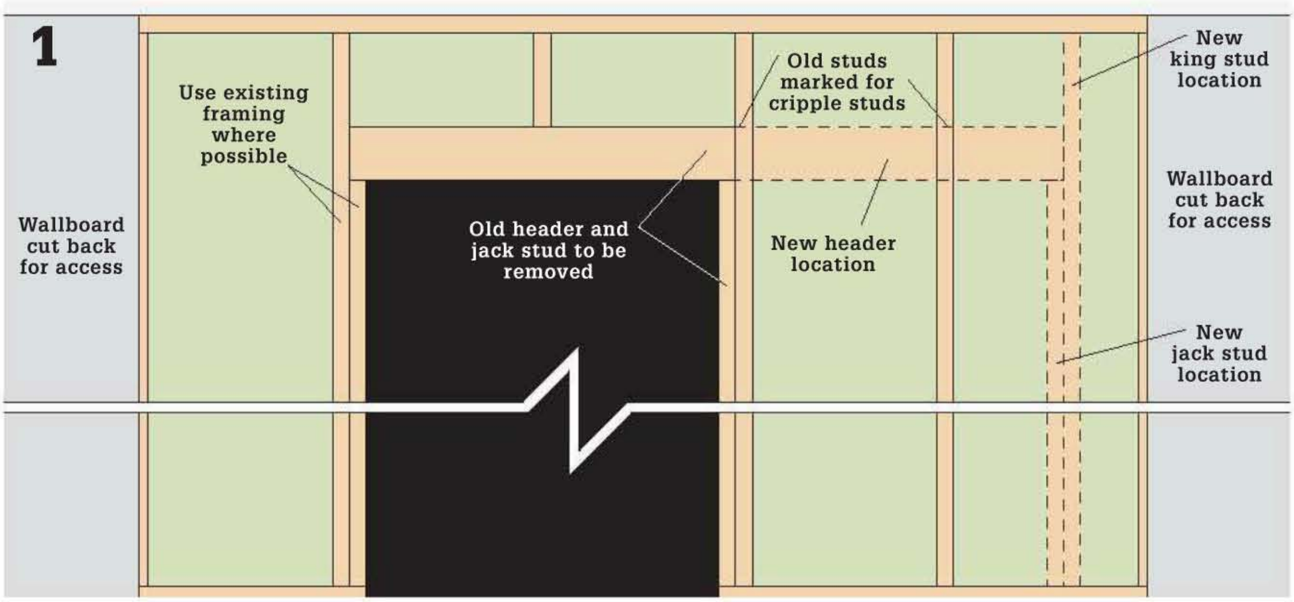
Use existing framing where possible
Wallboard cut back for access
Old header and jack stud to be removed
Old studs marked for cripple studs
New header location
New king stud location
Wallboard cut back for access
New jack stud location

Shut off power and water to the area. Remove the wall surfaces from both sides of the wall (pages 276 to 279), leaving one stud bay open on each side of the new rough opening. Also remove or reroute any wiring, plumbing, or ductwork. Lay out the new rough opening, marking the locations of all new jack and king studs on both the top and bottom plates. Where practical, use existing framing members. To install a new king stud, cut a stud to size and align with the layout marks; toenail to the bottom plate with 10d common nails, check for plumb, then toenail to the top plate to secure. Finally, mark both the bottom and top of the new header on one king stud, then use a level to extend the lines across the intermediate studs to the opposite king stud. If using existing framing, measure and mark from the existing jack stud.

2

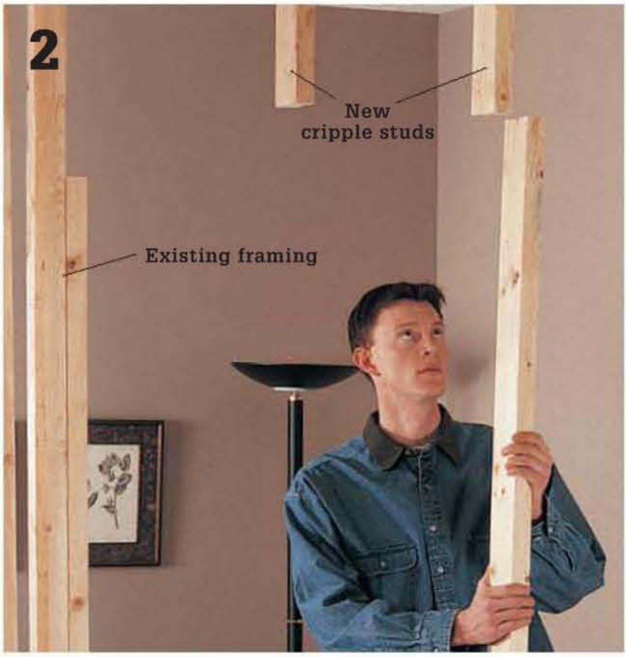
New cripple studs
Existing framing

Cut the intermediate studs at the reference marks for the top of the header using a reciprocating saw. Pry the studs away from the sole plates and remove—the remaining top pieces will be used as cripple studs.

3

To install a jack stud, cut the stud to fit between the sole plate and the bottom of the header, as marked on the king stud. Align it at the mark against the king stud, then fasten it in place with 10d common nails driven every 12".

(continued)

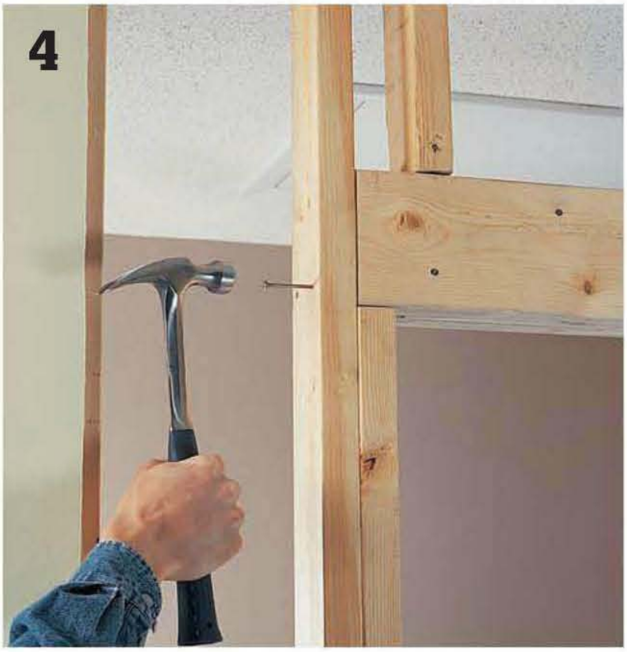

Build the header to size (pages 292 to 293) and install, fastening it to the jack studs, king studs, and cripple studs using 16d common nails. Use a handsaw to cut through the bottom plate so it's flush with the inside faces of the jack studs. Remove the cutout portion.

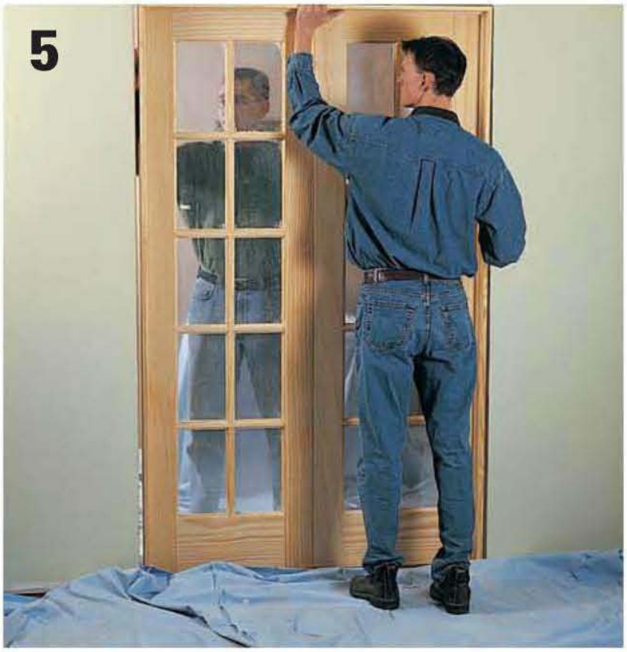

Finish the walls (for wallboard installation, see pages 44 to 51) before installing the doors, then set the prehung door unit into the framed opening so the jamb edges are flush with the finished wall surfaces and the unit is centered from side to side.

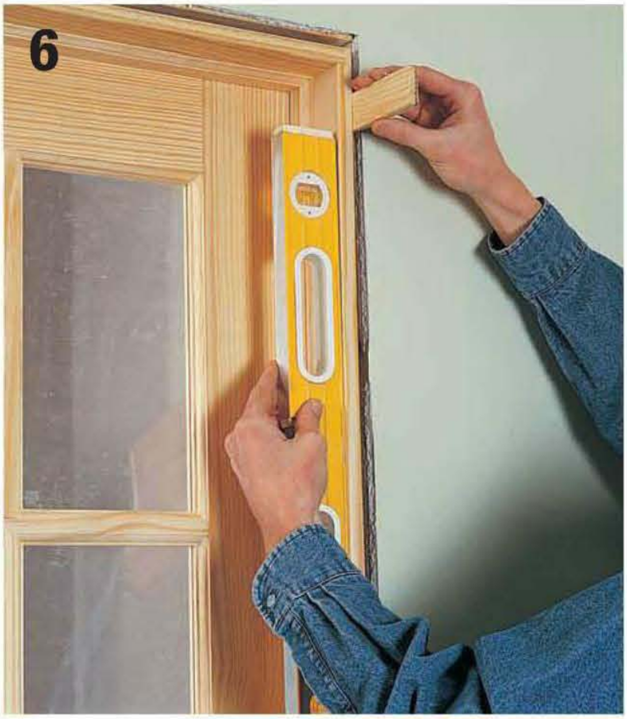

Using a level, adjust the unit to plumb one of the side jambs. Starting near the top of the door, insert pairs of shims driven from opposite directions into the gap between the framing and the jamb, sliding the shims until they are snug. Check the jamb to make sure it remains plumb and does not bow inward.

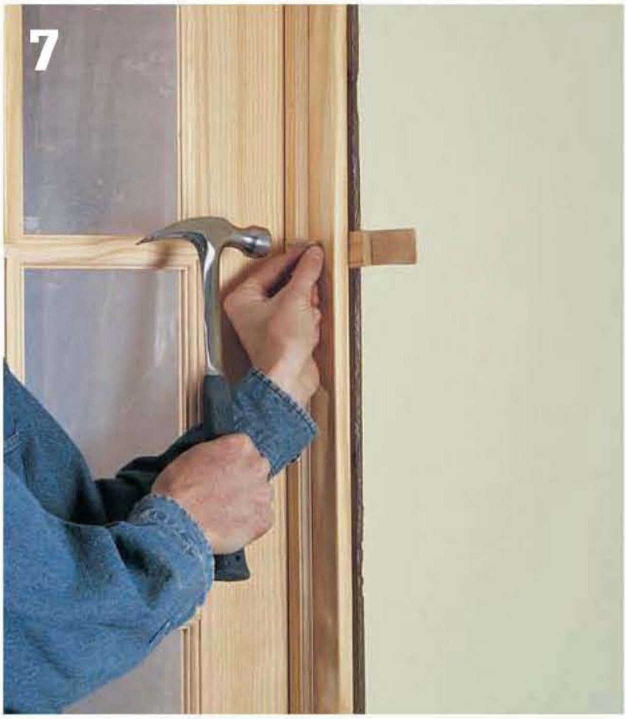

Working down along the jamb, install shims near each hinge and near the floor. Make sure the jamb is plumb, then anchor it with 8d finish nails driven through the jamb and shims and into the framing. Leave the nail heads partially protruding so the jamb can be readjusted later if necessary.

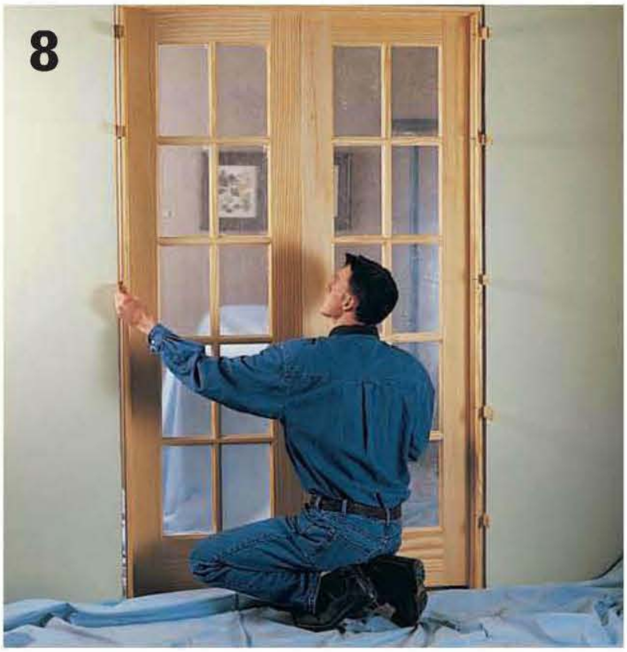

Install shims at the other side jamb, aligning them roughly with the shims of the first jamb. With the doors closed, adjust the shims so the reveal between the doors is even and the tops of the doors are aligned.

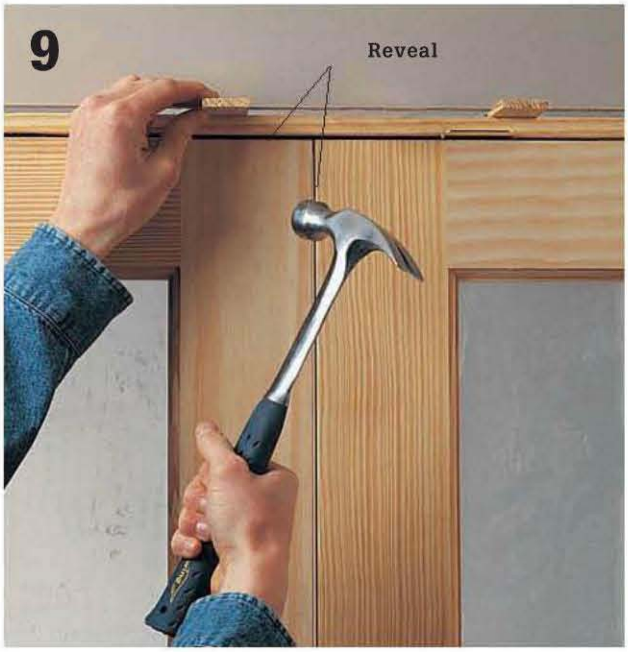

Shim the gap between the header and the head jamb to create a consistent reveal along the top when the doors are closed. Insert pairs of shims every 12". Drive 8d finish nails through the jambs and shims and into the framing members.

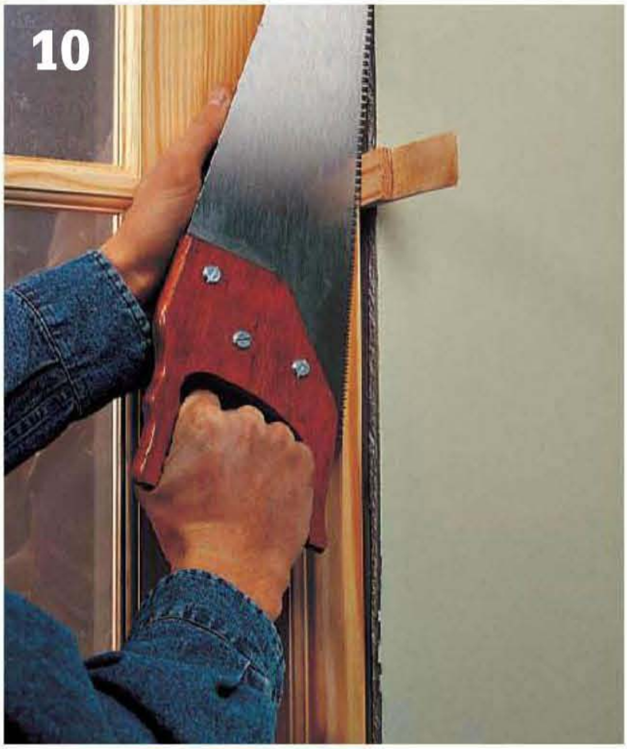

Drive all the nails fully, then set them below the surface of the wood with a nail set. Cut off the shims flush with the wall surface using a handsaw or utility knife. Hold the saw vertically to prevent damage to the door jamb or wall. Install the door casing.

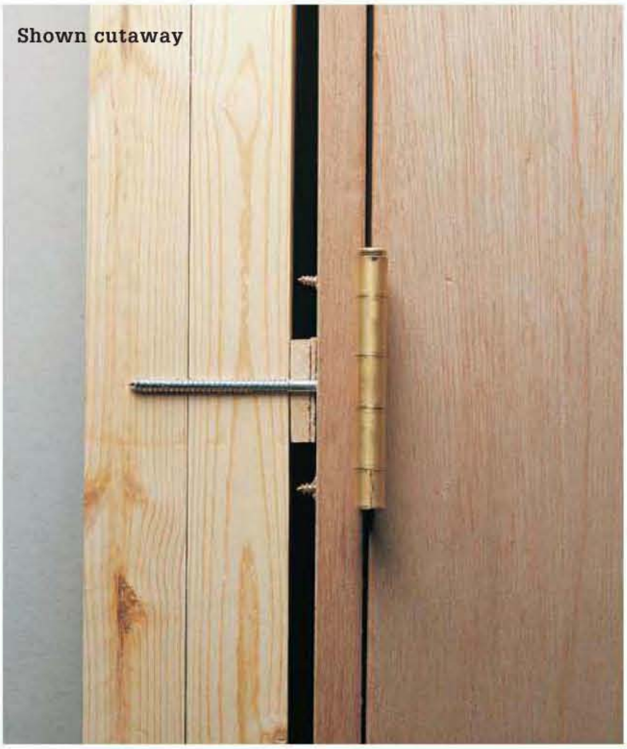

Option: Replace the center mounting screw on each hinge with a 3" wood screw to provide extra support for door hinges and jambs. These long screws extend through the side jambs and deep into the framing members. Be careful not to overtighten screws, which will cause the jambs to bow.

Installing Entry Doors

Few parts of a house have a more dramatic effect on the way your home is perceived than the main entry door. A lovely, well-maintained entryway that is tastefully matched architecturally to the house can utterly transform a home's appearance. In fact, industry studies have suggested that upgrading a plain entry door to a higher-end entry door system can pay back multiple times in the resale of your house. But perhaps more importantly, depending on your priorities, it makes a great improvement in how you feel about your home. Plus, it usually pays benefits in home security and energy efficiency as well.

If you are replacing a single entry door with a double door or a door with a sidelight or sidelights, you will need to enlarge the door opening (see pages 32 to 35). Be sure to file your plans with your local building department and obtain a permit. You'll need to provide temporary support from the time you remove the wall studs in the new opening until you've installed and secured a new door header that's approved for the new span distance.

The American Craftsman style door with sidelights (see Resources, page 554) installed in this project has the look and texture of a classic wood door, but it is actually created from fiber-glass. Today's fiberglass doors are quite convincing in their ability to replicate wood grain, while still offering the durability and low-maintenance of fiberglass.

Tools & Materials ▸

Tape measure
Level
Reciprocating saw
Caulk & caulk gun
Hammer
Shims
Framing nails
Finish nails
Nail set
Finishing materials

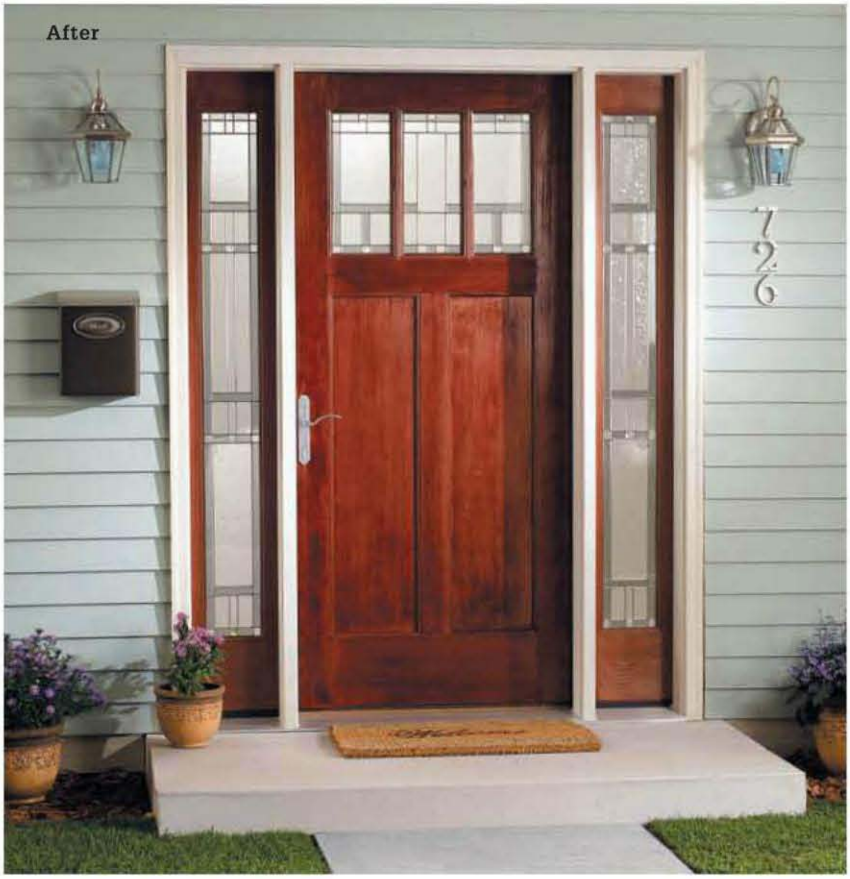

Replacing an ordinary entry door with a beautiful new upgrade has an exceptionally high payback in increased curb appeal and in perceived home value, according to industry studies.

How to Replace an Entry Door

Remove the old entry door by cutting through the fasteners driven into the jamb with a reciprocating saw (see pages 32 and 35). If the new door or door system is wider, mark the edges of the larger rough opening onto the wall surface. If possible, try to locate the new opening so one edge will be against an existing wall stud. Be sure to include the thickness of the new framing you'll need to add when removing the wall coverings.

Frame in the new rough opening for the replacement door (see pages 257 to 261). The instructions that come with the door will recommend a rough opening size, which is usually sized to create a ½" gap between the door and the studs and header. Patch the wall surfaces.

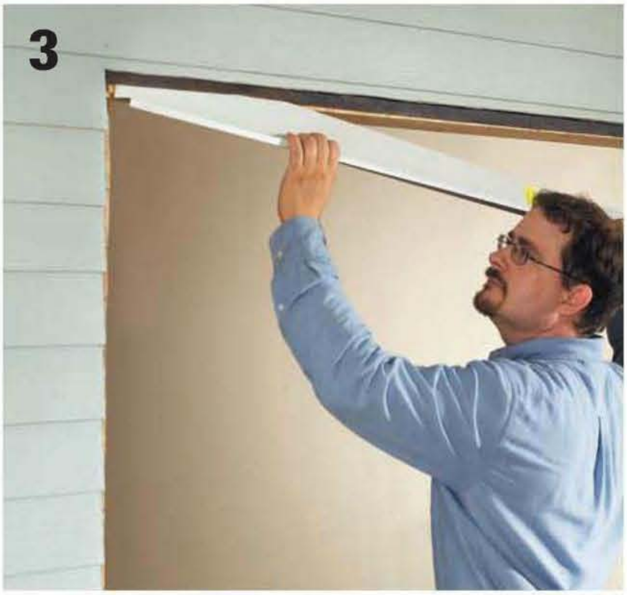

Cut metal door dripcap molding to fit the width of the opening and tuck the back edge up behind the wallcovering at the top of the door opening. Attach the dripcap with caulk only–do not use nails or screws.

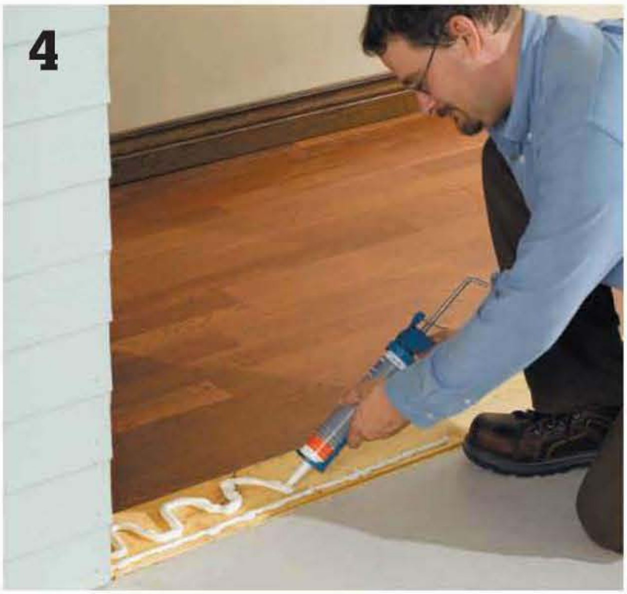

Unpack the door unit and set it in the rough opening to make sure it fits correctly. Remove it. Make sure the subfloor is clean and in good repair, and then apply heavy beads of caulk to the underside of the door sill and to the subfloor in the sill installation area. Use plenty of caulk.

(continued)

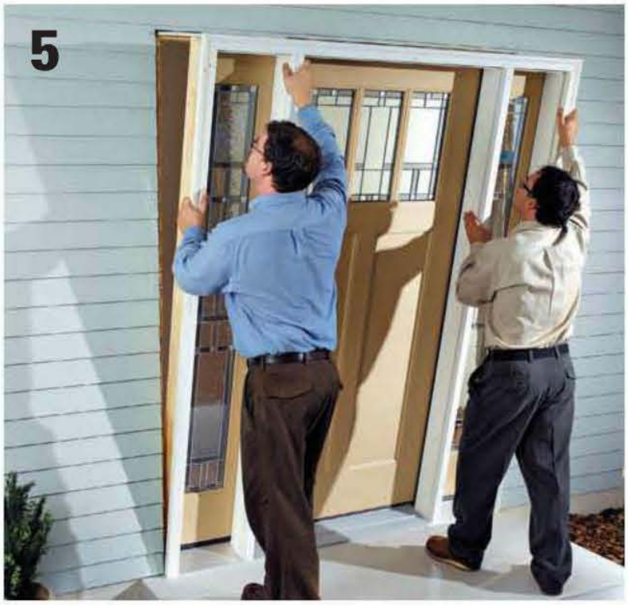

Set the door sill in the threshold and raise the unit up so it fits cleanly in the opening, with the exterior trim flush against the wall sheathing. Press down on the sill to seat it in the caulk and wipe up any squeeze-out with a damp rag

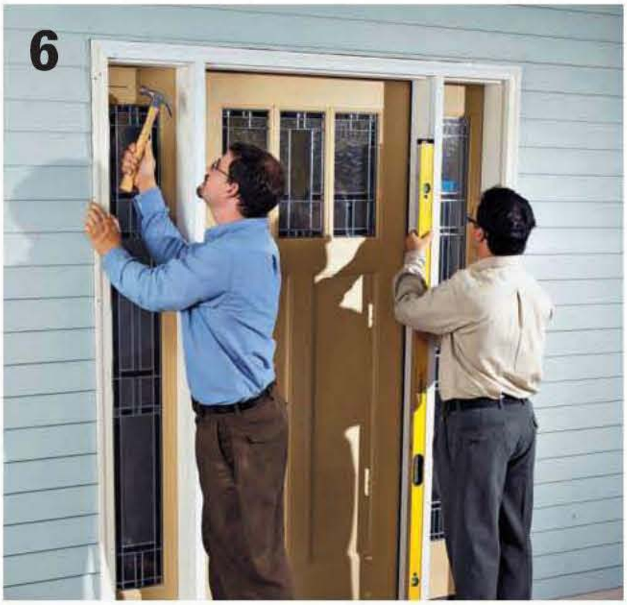

Use a 6-ft. level to make sure the unit is plumb and then tack it to the rough opening stud on the hinge side, using pairs of 10d nails driven partway through the casing on the weatherstripped side of the door (or the sidelight). On single, hinged doors, drive the nails just above the hinge locations. *Note: Many door installers prefer deck screws over nails when attaching the jambs. Screws offer more gripping strength and are easier to adjust, but covering the screw heads is more difficult than filling nail holes.*

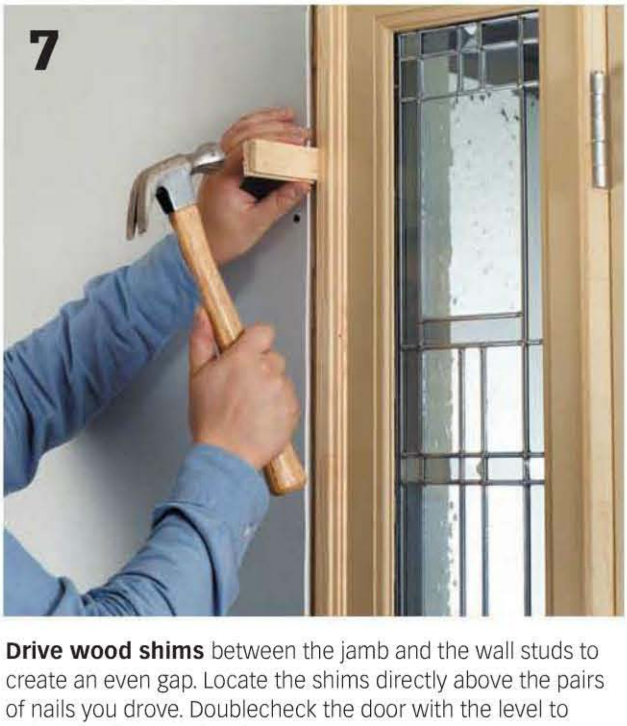

Drive wood shims between the jamb and the wall studs to create an even gap. Locate the shims directly above the pairs of nails you drove. Doublecheck the door with the level to make sure it is still plumb.

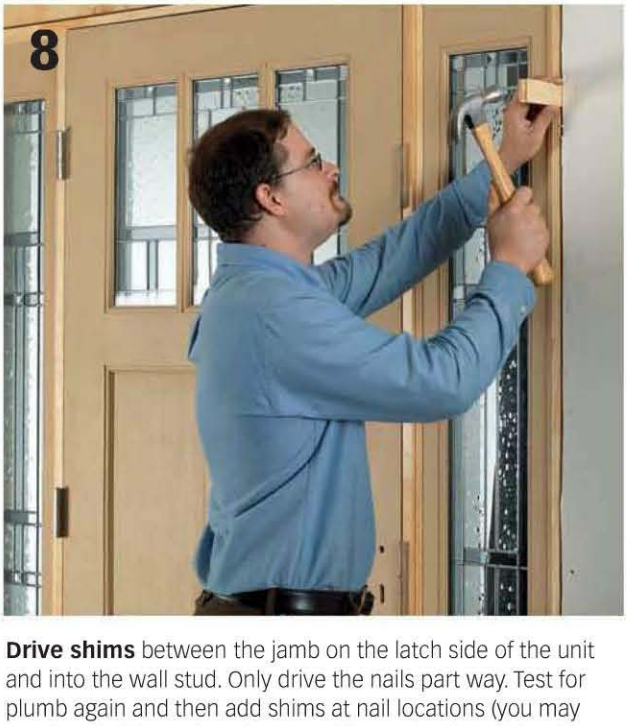

Drive shims between the jamb on the latch side of the unit and into the wall stud. Only drive the nails part way. Test for plumb again and then add shims at nail locations (you may need to double-up the shims, as this gap is often wider than the one on the hinge side). Check to make sure the door jamb is not bowed.

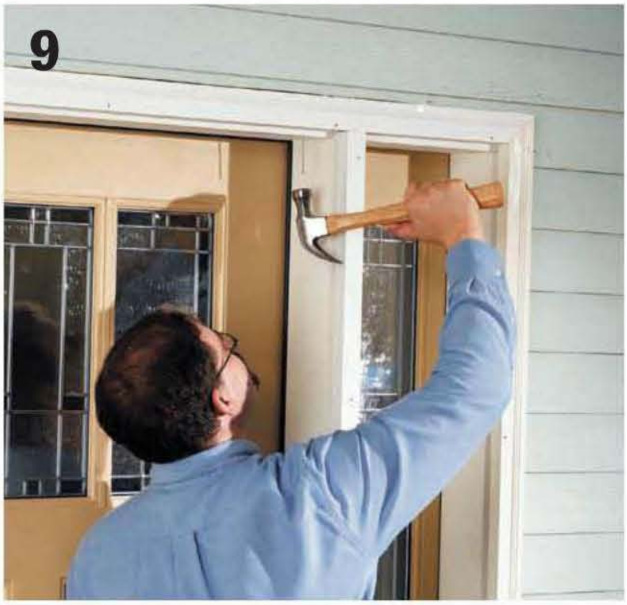

Drive finish nails at all remaining locations, following the nailing schedule in the manufacturer's installation instructions.

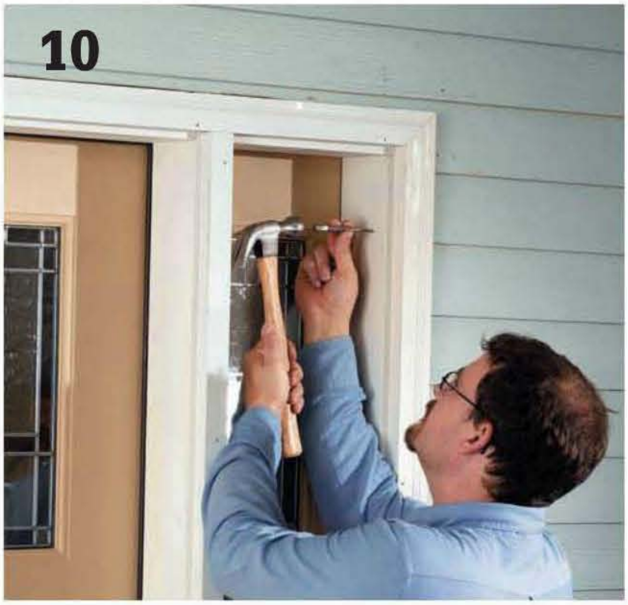

Use a nail set to drive the nail heads below the wood surface. Fill the nail holes with wood putty (you'll get the best match if you apply putty that's tinted to match stained wood after the finish is applied). The presence of the wood shims at the nail locations should prevent the jamb from bowing as you nail.

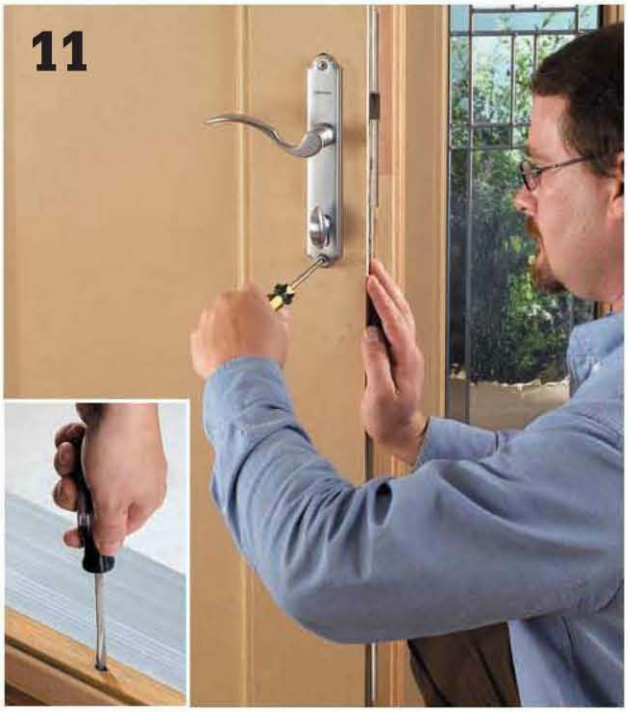

Install the lockset, strikeplates, deadbolts or multipoint locks, and any other door hardware. If the door finish has not been applied, you may want to do so first, but generally it makes more sense to install the hardware right away so the door can be operated and locked. Attach the door sill to the threshold and adjust it as needed, normally using the adjustment screws (inset).

Apply your door finish if it has not yet been applied. Read the manufacturer's suggestions for finishing very closely and follow the suggested sequences. Some manufacturers offer finish kits that are designed to be perfectly compatible with their doors. Install interior case molding and caulk all the exterior gaps after the finish dries.

Installing Storm Doors

Install a storm door to improve the appearance and weather resistance of an old entry door or to protect a newly installed door against weathering. In all climates, adding a storm door can extend the life of an entry door.

When buying a storm door, look for models that have a solid inner core and seamless outer shell construction. Carefully note the dimensions of your door opening, measuring from the inside edges of the entry door's brick molding. Choose a storm door that opens from the same side as your entry door.

Tools & Materials ▸

Tape measure
Pencil
Plumb bob
Hacksaw
Hammer
Drill and bits
Screwdrivers
Storm door unit
Wood spacer strips
4d casing nails

Adjustable sweeps help make storm doors weathertight. Before installing the door, attach the sweep to the bottom of the door. After the door is mounted, adjust the height of the sweep so it brushes the top of the sill lightly when the door is closed.

How to Cut a Storm Door Frame to Fit a Door Opening

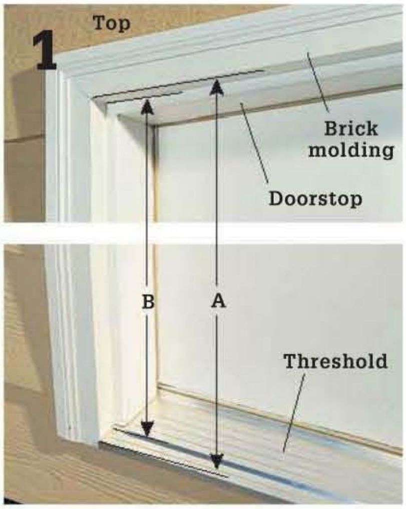

Because entry door thresholds are slanted, the bottom of the storm door frame needs to be cut to match the threshold angle. Measure from the threshold to the top of the door opening along the corner of the brick molding (A), then measure along the front edge of the entry doorstop (B).

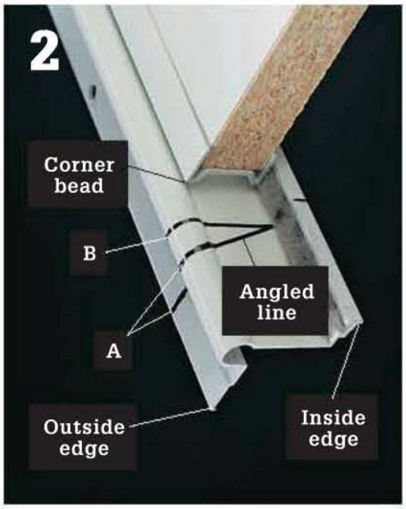

Subtract ⅛" from measurements A and B to allow for small adjustments when the door is installed. Measuring from the top of the storm door frame, mark the adjusted points A and B on the corner bead. Draw a line from point A to the outside edge of the frame and from point B to the inside edge. Draw an angled line from point A on the corner bead to point B on the inside edge.

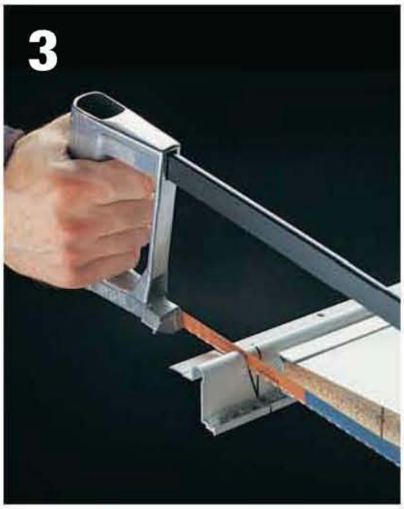

Use a hacksaw to cut down through the bottom of the storm door frame, following the angled line. Make sure to hold the hacksaw at the same slant as the angled line to ensure that the cut will be smooth and straight.

How to Fit & Install a Storm Door

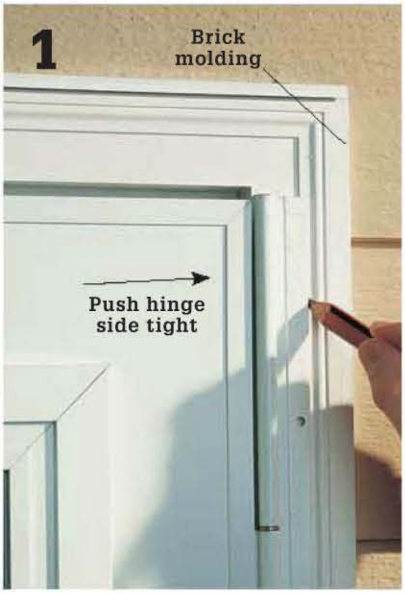

Position the storm door in the opening and push the frame tight against the brick molding on the hinge side of the storm door, then draw a reference line on the brick molding, following the edge of the storm door frame.

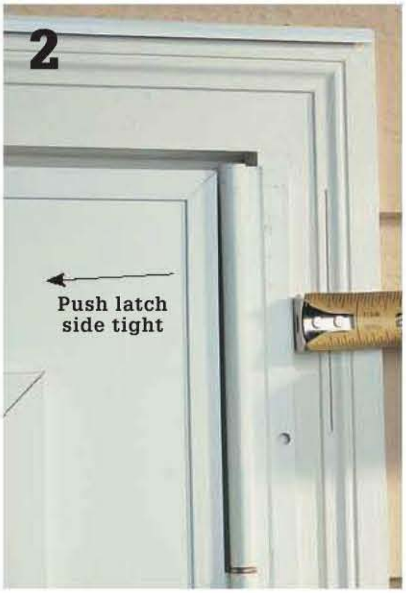

Push the storm door frame tight against the brick molding on the latch side, then measure the gap between the reference line and the hinge side of the door frame. If the distance is greater than ⅜", spacer strips must be installed to ensure the door will fit snugly.

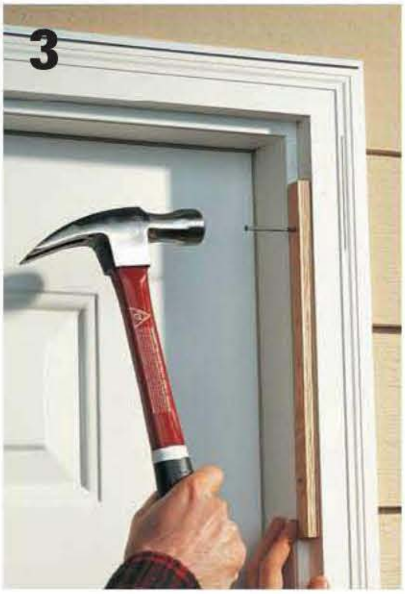

To install spacers, remove the door, then nail thin strips of wood to the inside of the brick molding at storm door hinge locations. The thickness of the wood strips should be ⅛" less than the gap measured in step 2.

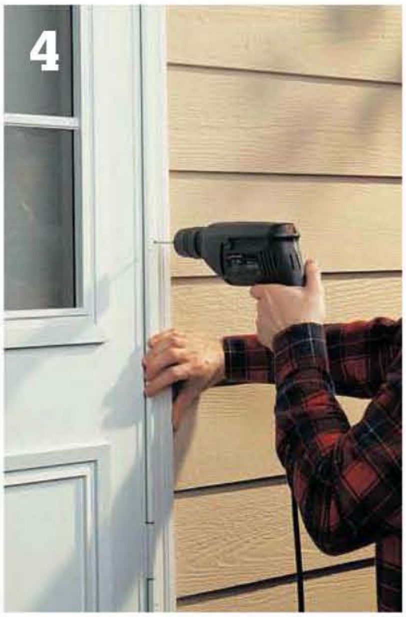

Replace the storm door and push it tight against the brick molding on the hinge side. Drill pilot holes through the hinge-side frame of the storm door and into the brick molding spaced every 12". Attach the frame with mounting screws.

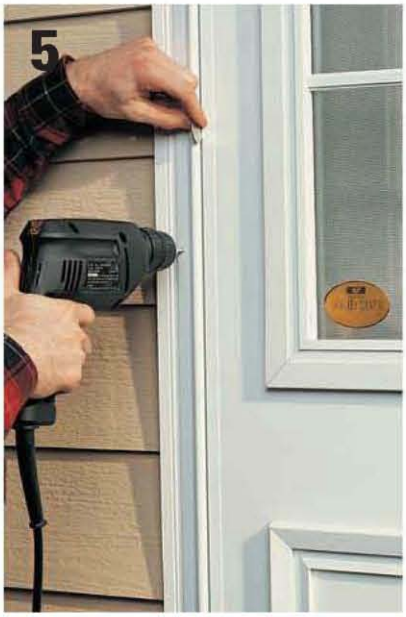

Remove any spacer clips holding the frame to the storm door. With the storm door closed, drill pilot holes and attach the latch-side frame to the brick molding. Use a coin to keep an even gap between the storm door and the storm door frame.

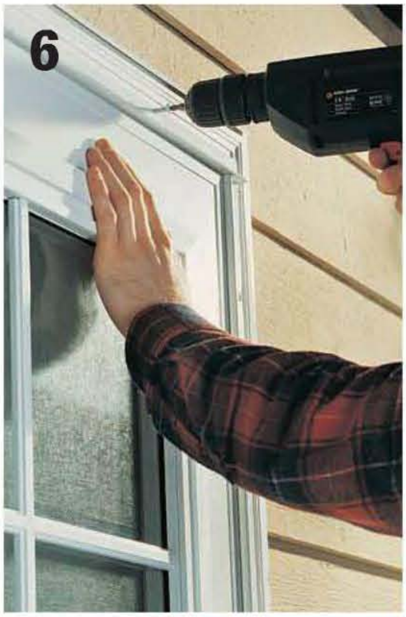

Center the top piece of the storm door frame on top of the frame sides. Drill pilot holes and screw the top piece to the brick molding. Adjust the bottom sweep, then attach the locks and latch hardware as directed by the manufacturer.

Installing Patio Doors

For easy installation, buy a patio door with the door panels already mounted in a preassembled frame. Try to avoid patio doors sold with frame kits that require complicated assembly.

Because patio doors have very long bottom sills and top jambs, they are susceptible to bowing and warping. To avoid these problems, be very careful to install the patio door so it is level and plumb and to anchor the unit securely to framing members. Yearly caulking and touch-up painting helps prevent moisture from warping the jambs.

Tools & Materials ▸

Pencil
Hammer
Circular saw
Handsaw
Wood chisel
Stapler
Caulk gun
Level
Pry bar
Cordless screwdriver
Drill and bits
Nail set
Shims
Drip edge
Building paper
Silicone and latex caulk
10d casing nails
3" wood screws
Sill nosing
Fiberglass insulation
Patio door kit

Patio doors offer the best qualities of both windows and doors—plenty of natural light, a great view, wide room access, and reasonable security.

If not included with the unit, screen doors can be ordered from most patio door manufacturers. Screen doors have spring-mounted rollers that fit into a narrow track on the outside of the patio door threshold.

Installing Sliding Patio Doors

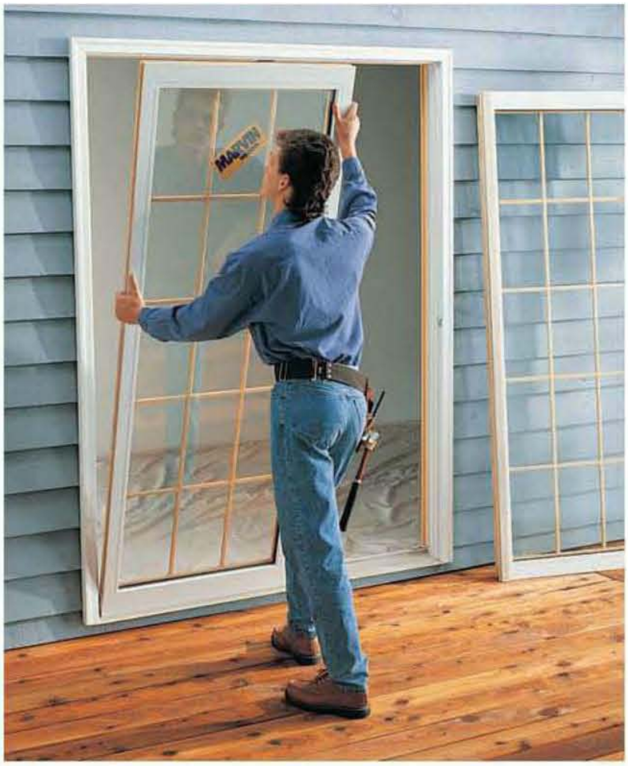

Remove heavy glass panels if you must install the door without help. Reinstall the panels after the frame has been placed in the rough opening and nailed at opposite corners. To remove and install the panels, remove the stop rail found on the top jamb of the door unit.

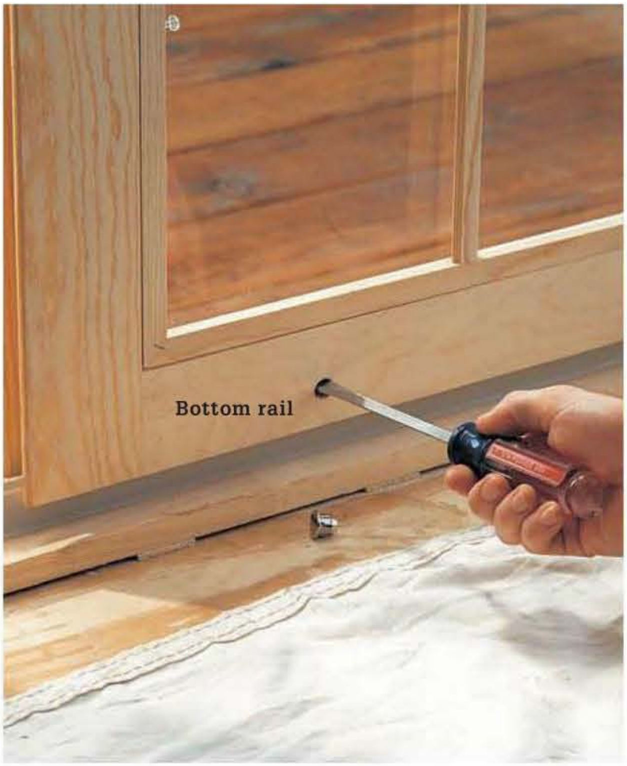

Adjust the bottom rollers after installation is complete. Remove the coverplate on the adjusting screw, found on the inside edge of the bottom rail. Turn the screw in small increments until the door rolls smoothly along the track without binding when it is opened and closed.

Tips for Installing Hinged Patio Doors ▸

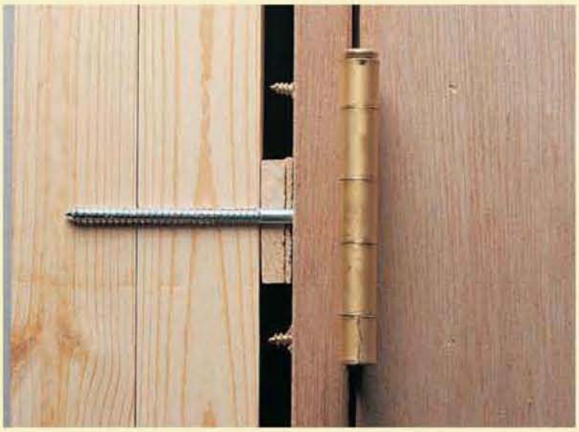

Provide extra support for door hinges by replacing the center mounting screw on each hinge with a 3" wood screw. These long screws extend through the side jambs and deep into the framing members.

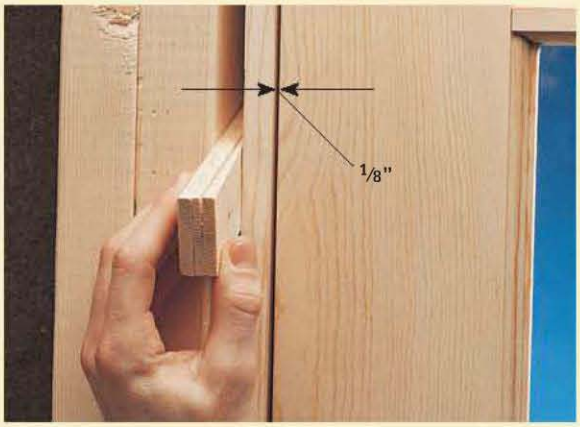

Keep a uniform ⅛" gap between the door, side jambs, and top jamb to ensure that the door will swing freely without binding. Check this gap frequently as you shim around the door unit.

How to Install a Patio Door

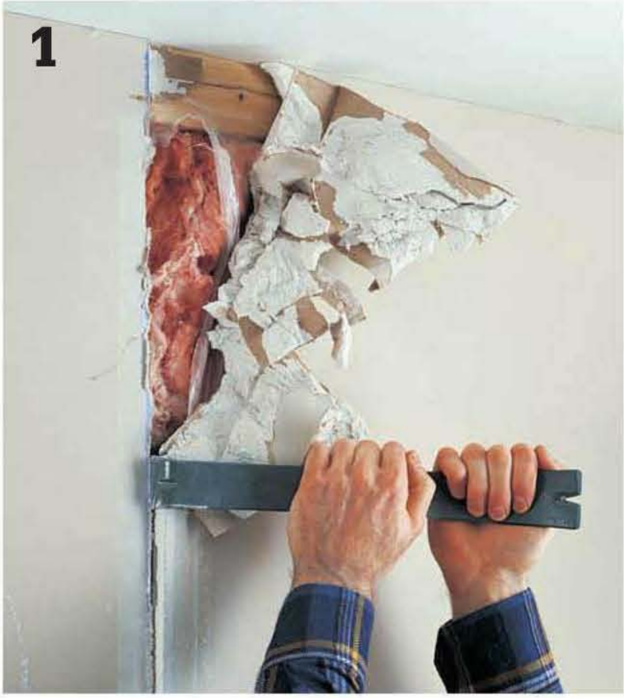

Prepare the work area and remove the interior wall surfaces, then frame the rough opening for the patio door. Remove the exterior surfaces inside the framed opening.

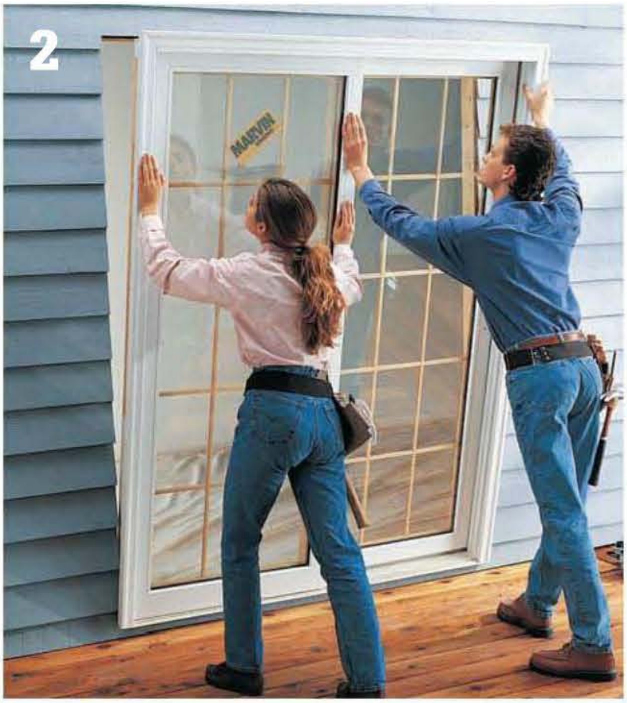

Test-fit the door unit, centering it in the rough opening. Check to make sure the door is plumb. If necessary, shim under the lower side jamb until the door is plumb and level. Have a helper hold the door in place while you adjust it.

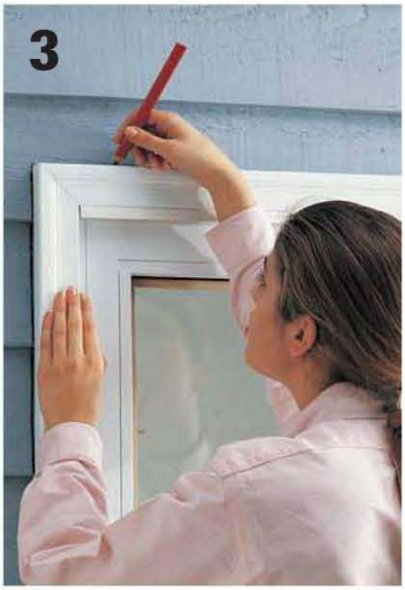

Trace the outline of the brick molding onto the siding, then remove the door unit. Note: *If you have vinyl or metal siding, see page 282 for advice on removing the siding.*

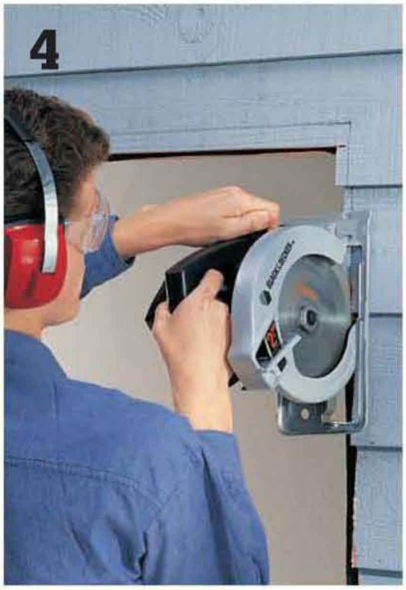

Cut the siding along the outline, just down to the sheathing using a circular saw. Stop just short of the corners to prevent damage to the remaining siding. Finish the cuts at the corners with a sharp wood chisel.

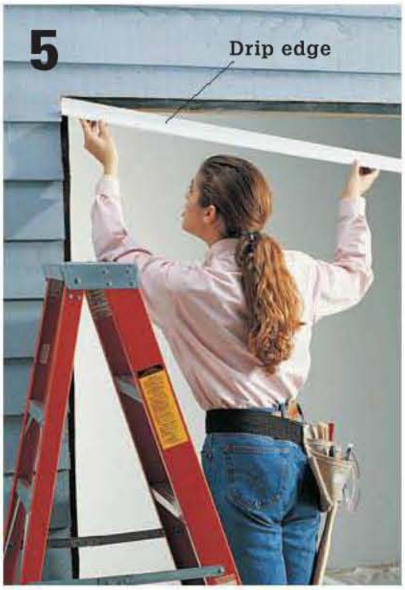

To provide an added moisture barrier, cut a piece of drip edge to fit the width of the rough opening, then slide it between the siding and the existing building paper at the top of the opening. Do not nail the drip edge.

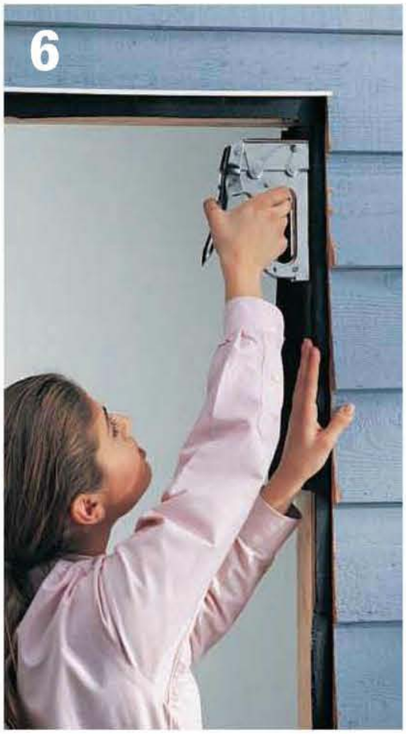

Cut 8"-wide strips of building paper and slide them between the siding and sheathing. Bend the paper around the framing members and staple it in place. Each piece overlaps the piece below it.

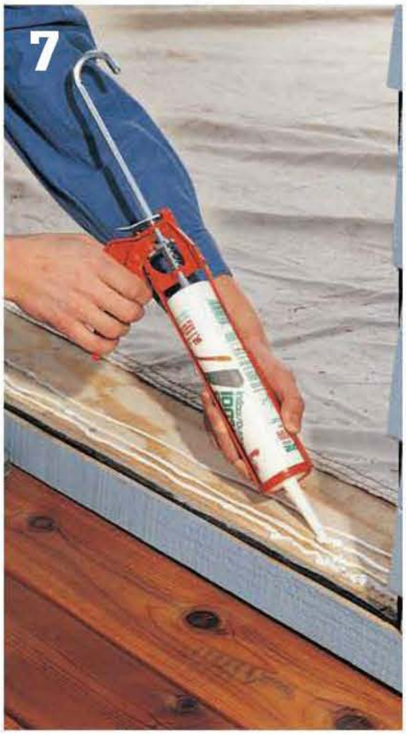

Apply several thick beads of silicone caulk to the subfloor at the bottom of the door opening.

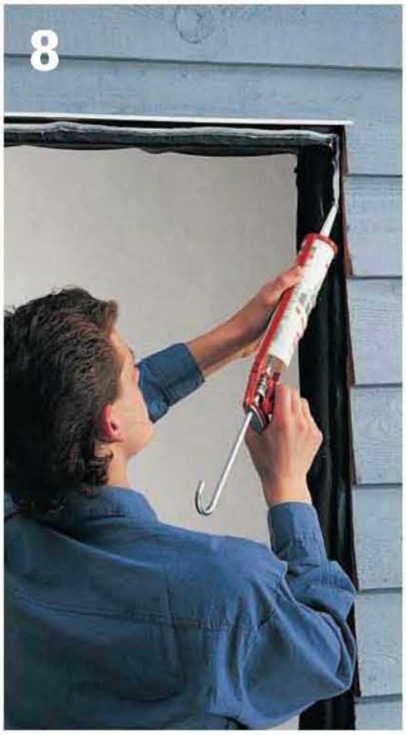

Apply silicone caulk around the front edge of the framing members where the siding meets the building paper.

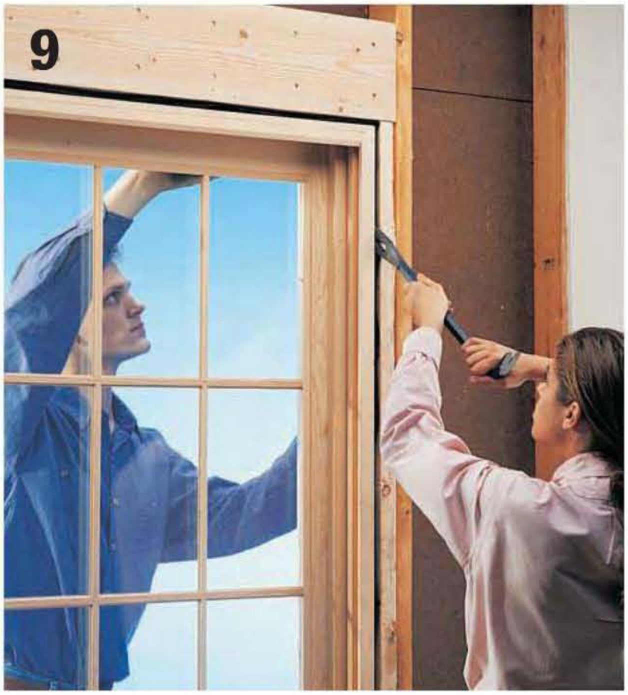

Use a pry bar to center the door in the rough opening so the brick molding is tight against the sheathing. Have a helper hold the door unit from outside.

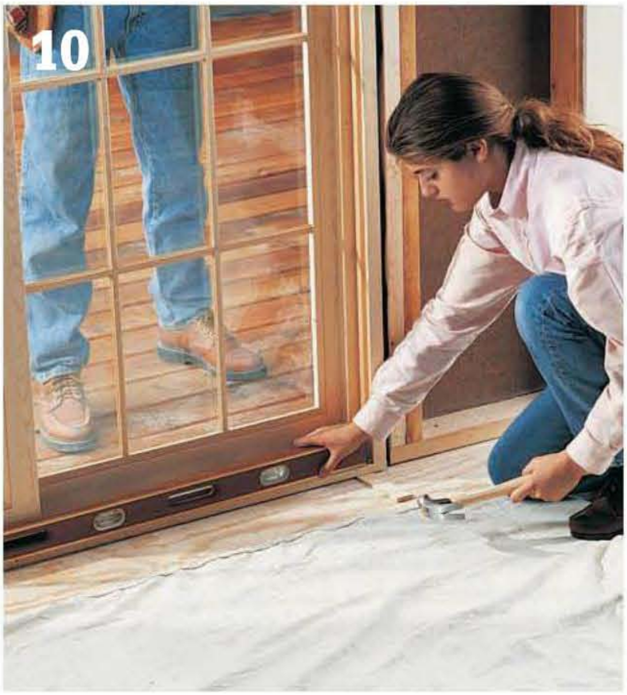

Check the door threshold to make sure it is level. If necessary, shim under the lower side jamb until the patio door unit is level.

(continued)

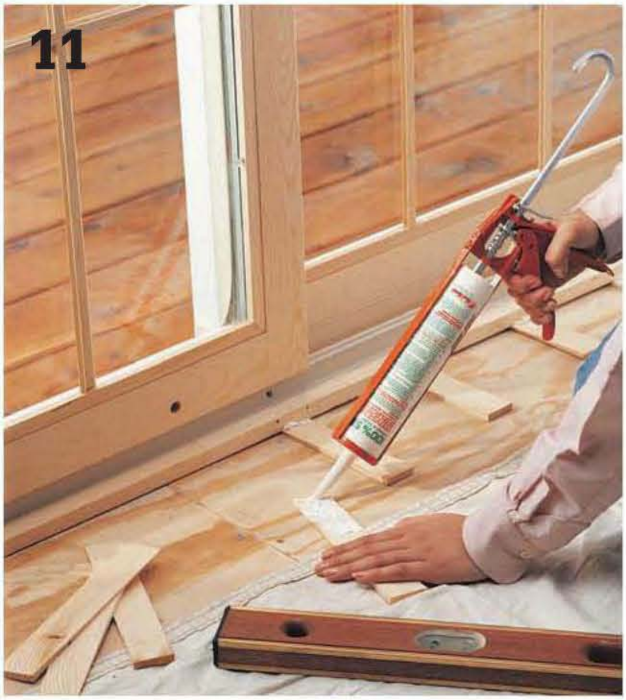

If there are gaps between the threshold and subfloor, insert shims coated with caulk into the gaps, spaced every 6". Shims should be snug, but not so tight that they cause the threshold to bow. Clear off excess caulk immediately.

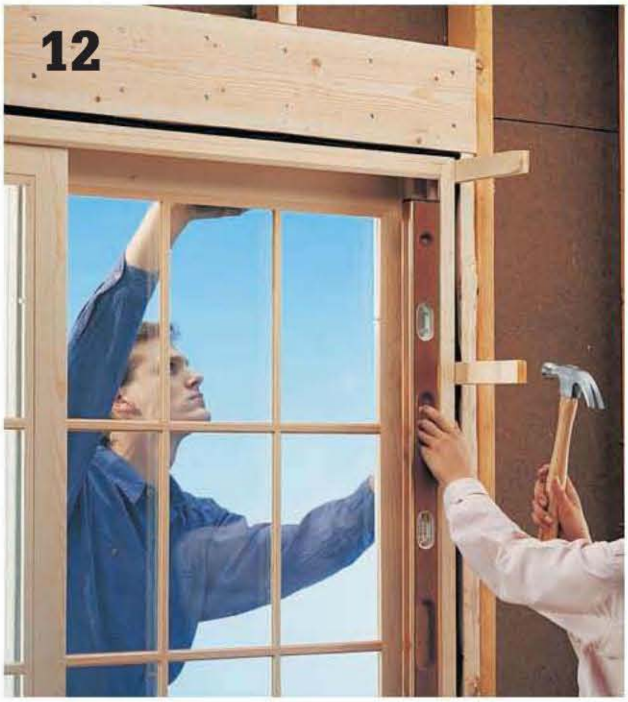

Place pairs of hardwood wedge shims together to form flat shims. Insert the shims every 12" into the gaps between the side jambs and the jack studs. For sliding doors, shim behind the strike plate for the door latch.

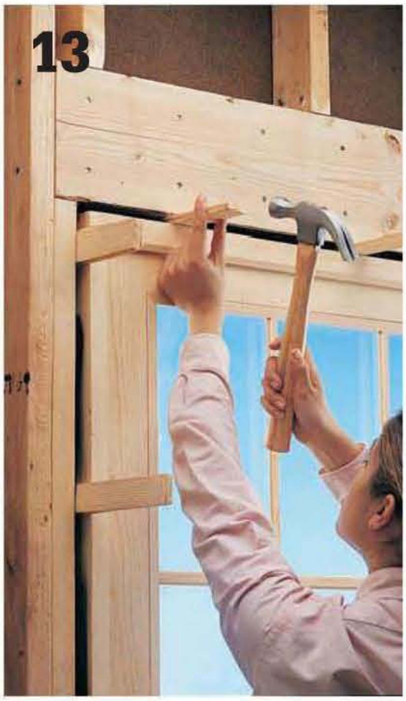

Insert shims every 12" into the gap between the top jamb and the header.

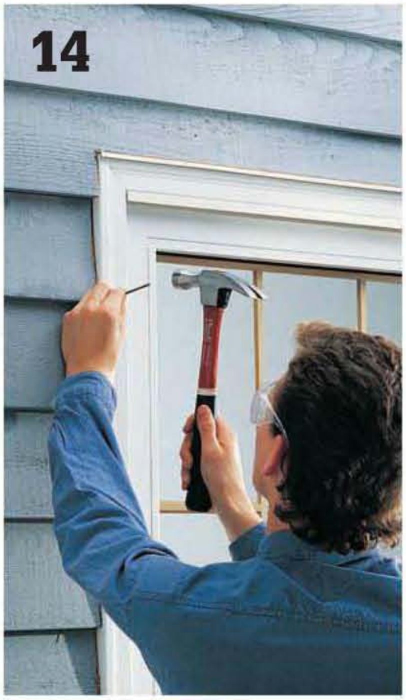

From outside, drive 10d casing nails, spaced every 12", through the brick molding and into the framing members. Use a nail set to drive the nail heads below the surface of the wood.

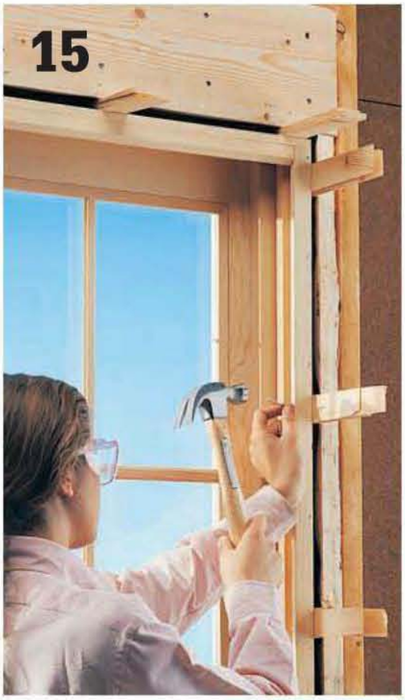

From inside, drive 10d casing nails through the door jambs and into the framing members at each shim location. Use a nail set to drive the nail heads below the surface of the wood.

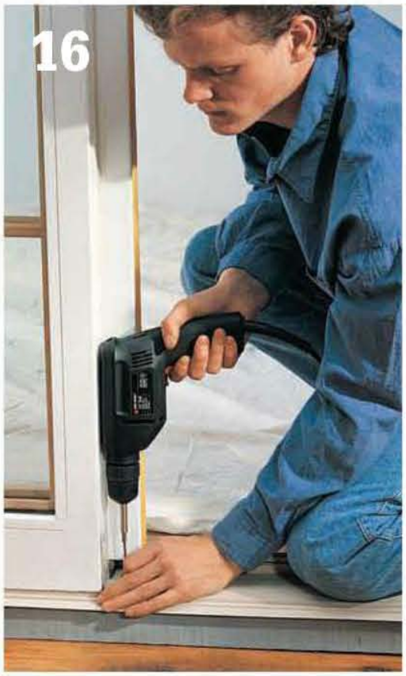

Remove one of the screws and cut the shims flush with the stop block found in the center of the threshold. Replace the screw with a 3" wood screw driven into the subfloor as an anchor.

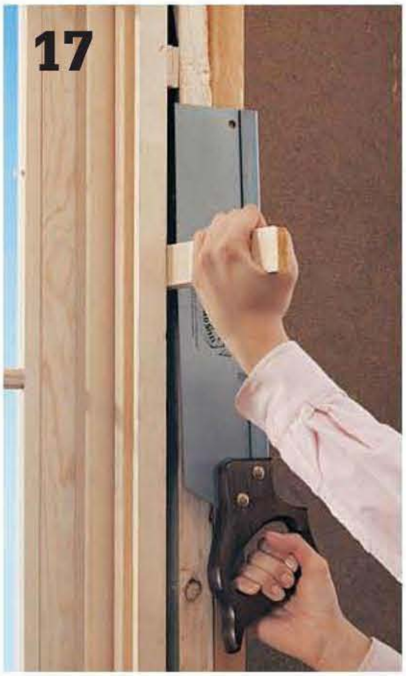

Cut off the shims flush with the face of the framing members using a handsaw. Fill gaps around the door jambs and beneath the threshold with loosely packed fiberglass insulation.

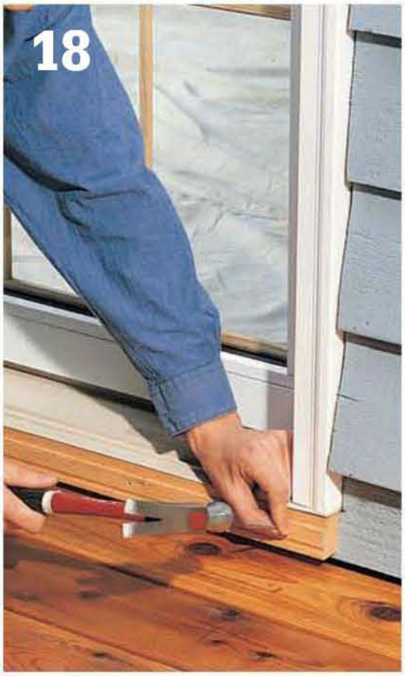

Reinforce and seal the edge of the threshold by installing sill nosing under the threshold and against the wall. Drill pilot holes and attach the sill nosing with 10d casing nails.

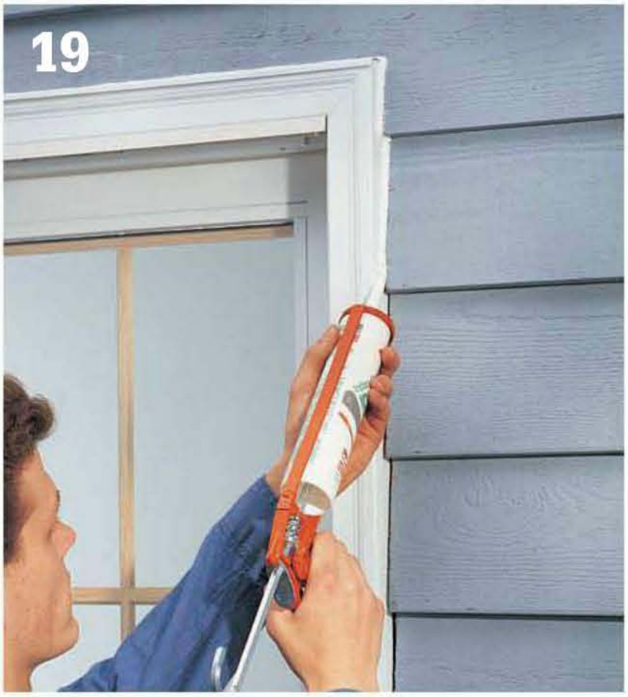

Make sure the drip edge is tight against the top brick molding, then apply paintable silicone caulk along the top of the drip edge and along the outside edge of the side brick moldings. Fill all exterior nail holes with caulk.

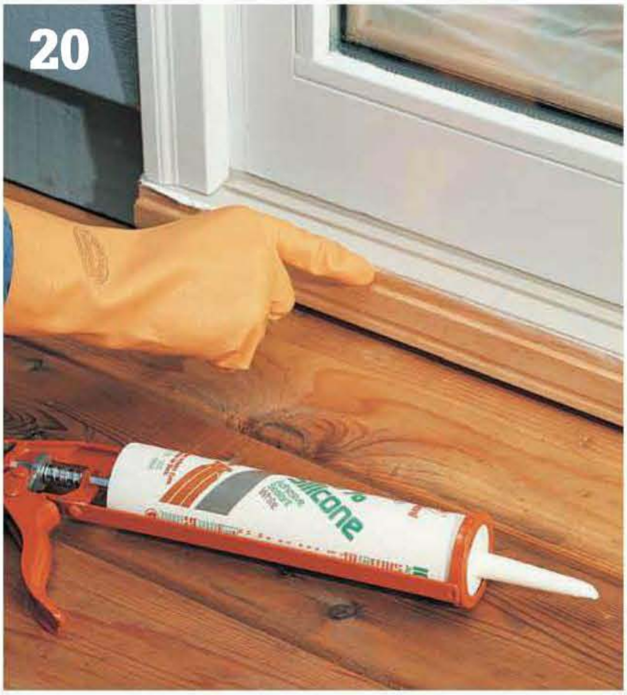

Caulk completely around the sill nosing using your finger to press the caulk into any cracks. As soon as the caulk is dry, paint the sill nosing. Finish the door and install the lockset as directed by the manufacturer.

Installing Door & Window Casing

Door and window casings provide an attractive border around doors and windows. They also cover the gaps between door or window jambs and the surfaces of surrounding walls.

Install door and window casings with a consistent reveal between the inside edges of the jambs and casings, making sure the casings are level and plumb.

In order to fit casings properly, the jambs and wallcoverings must lie in the same plane. If either one protrudes, the casings will not lie flush. To solve this problem, you'll need to remove some material from whichever surface is protruding.

Use a block plane to shave protruding jambs or a surface forming rasp to shave a protruding wallboard edge. Wallboard screws rely on the strength of untorn facing paper to support the wallboard. If the paper around the screws is damaged, drive additional screws nearby where the paper is still intact.

Tools & Materials ▸

- Tape measure
- Pencil
- Combination square
- Nail set
- Level
- Straightedge
- Power miter saw
- Hammer or pneumatic nailer
- Casing material
- Plinths and corner blocks (optional)
- 4d and 6d finish nails
- Wood putty

Case molding is installed around windows and doors to conceal the gaps between jambs and the wall. Venturing beyond the very common ranch-style casing offers some high design payback.

How to Install Mitered Casing on Doors & Windows

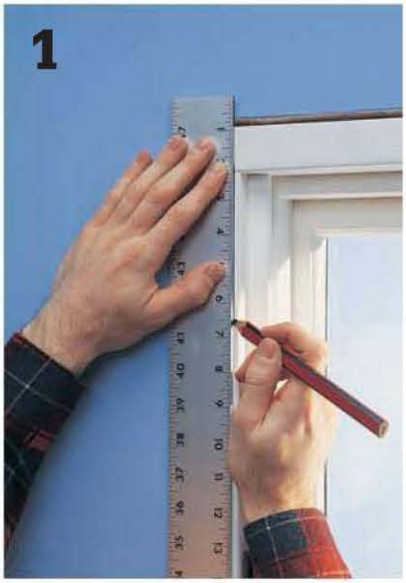

On each jamb, mark a reveal line ⅛" from the inside edge. The casings will be installed flush with these lines. *Note: On double-hung windows, the casings are usually installed flush with the edge of the jambs, so no reveal line is needed.*

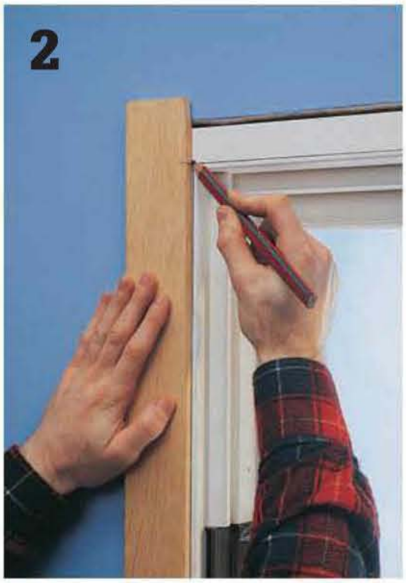

Place a length of casing along one side jamb, flush with the reveal line. At the top and bottom of the molding, mark the points where horizontal and vertical reveal lines meet. (When working with doors, mark the molding at the top only).

Make 45° miter cuts on the ends of the moldings. Measure and cut the other vertical molding piece, using the same method.

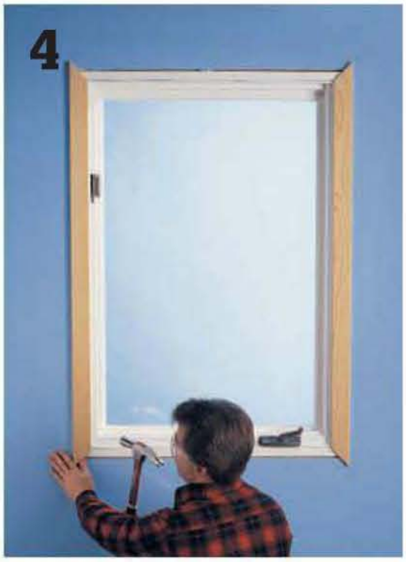

Drill pilot holes spaced every 12" to prevent splitting and attach the vertical casings with 4d finish nails driven through the casings and into the jambs. Drive 6d finish nails into framing members near the outside edge of the casings.

Measure the distance between the side casings, and cut top and bottom casings to fit, with ends mitered at 45°. If window or door unit is not perfectly square, make test cuts on scrap pieces to find the correct angle of the joints. Drill pilot holes and attach with 4d and 6d finish nails.

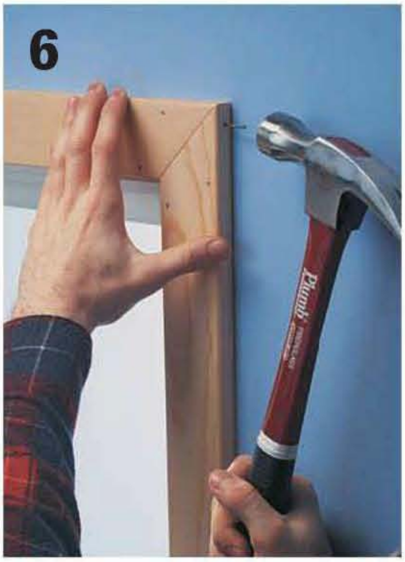

Locknail the corner joints by drilling pilot holes and driving 4d finish nails through each corner, as shown. Drive all nail heads below the wood surface, using a nail set, then fill the nail holes with wood putty.

How to Install Butted Door Casings

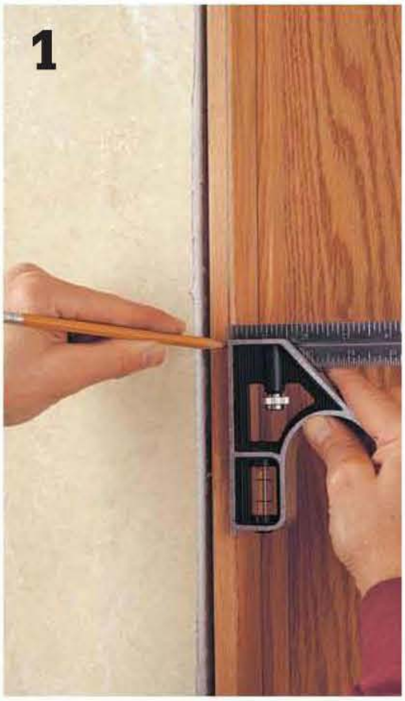

On each jamb, mark a reveal line ⅛" from the inside edge. The casings will be installed flush with these lines.

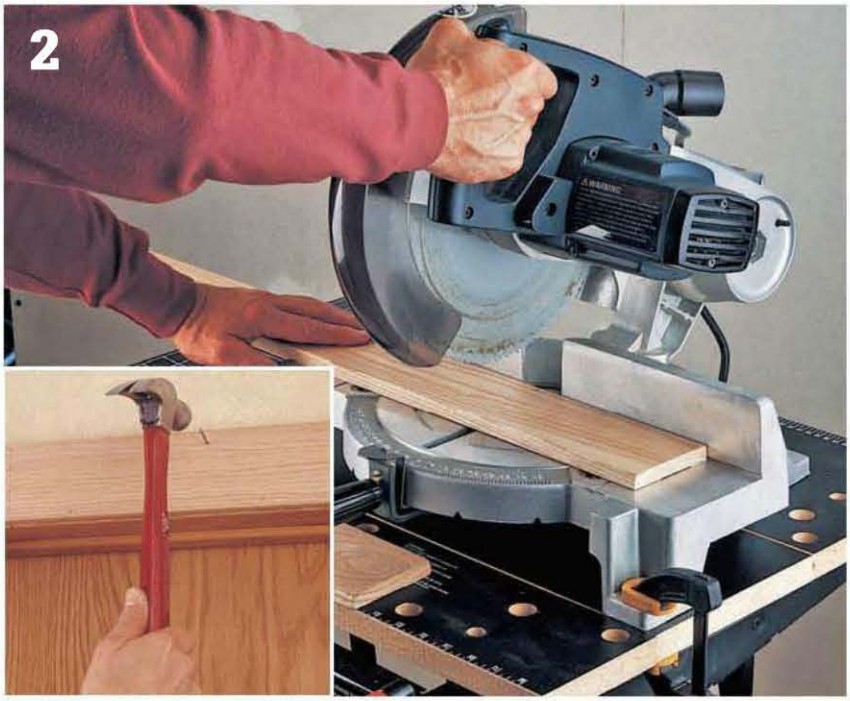

Cut the head casing to length. Mark the centerpoint of the head casing and the centerpoint of the head jamb. Align the casing with the head jamb reveal line, matching the centerpoints so that the head casing extends evenly beyond both side jamb casings. Nail the casing to the wall at stud locations and at the jamb (inset).

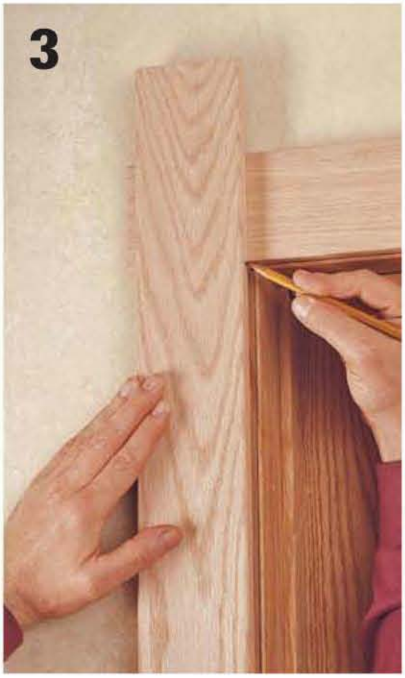

Hold the side casings against the head casing and mark them for cutting, then cut the side casings to fit.

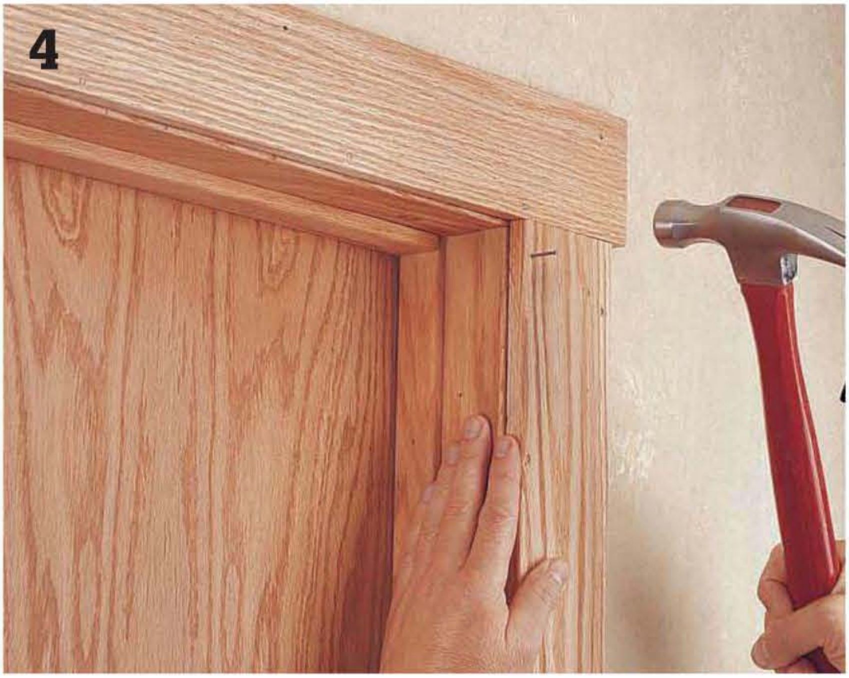

Align the side casings with the side jamb reveal lines, then nail the casings to the jambs and framing members. Set the nails, using a nail set. Fill the nail holes with wood putty.

Options for Installing Door & Window Casings

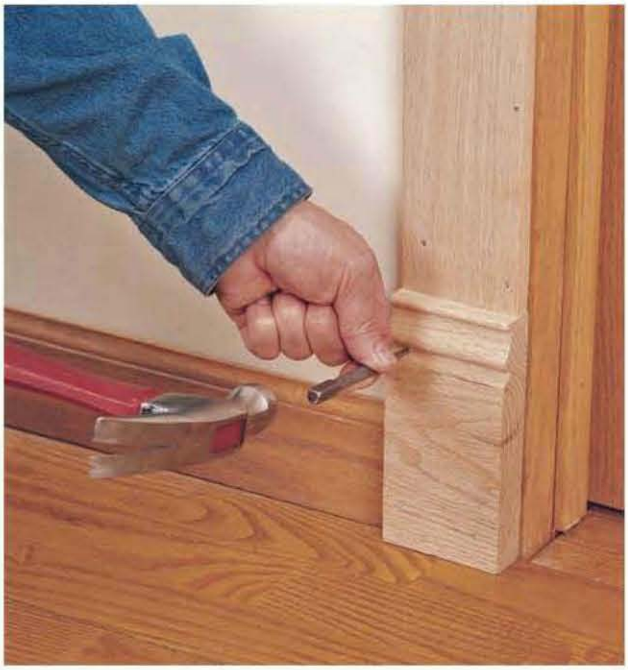

Dress up door casings by adding plinths. Cut the plinths from 1× stock and bevel one edge. Nail the plinths to the jambs with 2" 10d finish nails so the beveled edges are aligned with the reveal lines for the casings. Measure and cut the casings to fit.

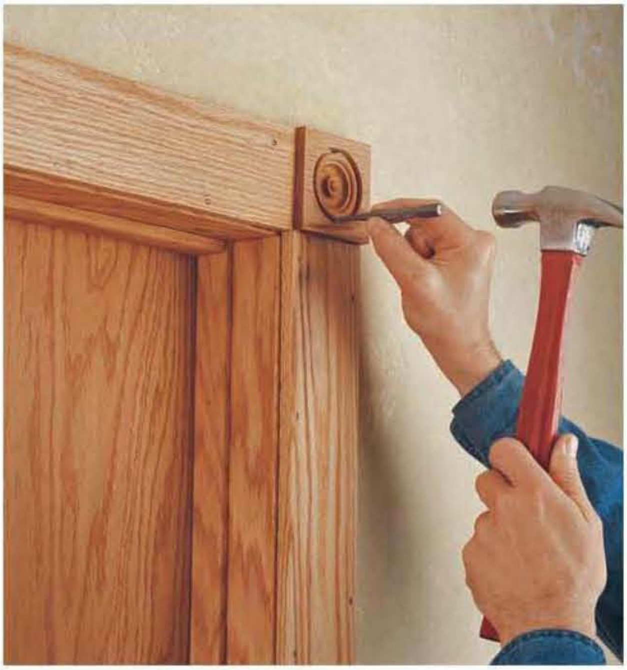

Add corner blocks, also known as rosettes, at the ends of the head casing. Attach the corner blocks once the side casings are in place, then cut the head casing to fit. Set the nails, using a nailset, after all pieces are installed.

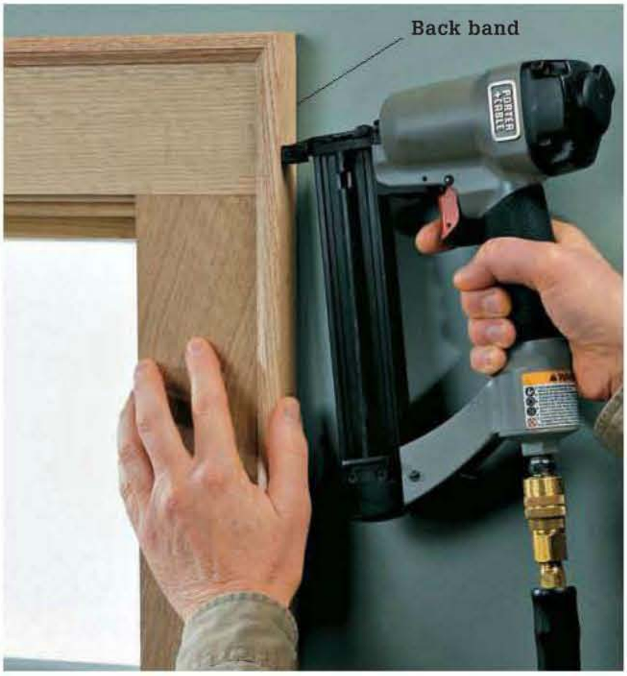

Backband molding can dress up butted window casings. Install the back band around the perimeter of the window, mitering the joints at the corners. Nail the back band in place with 4d finish nails or pneumatic brads.

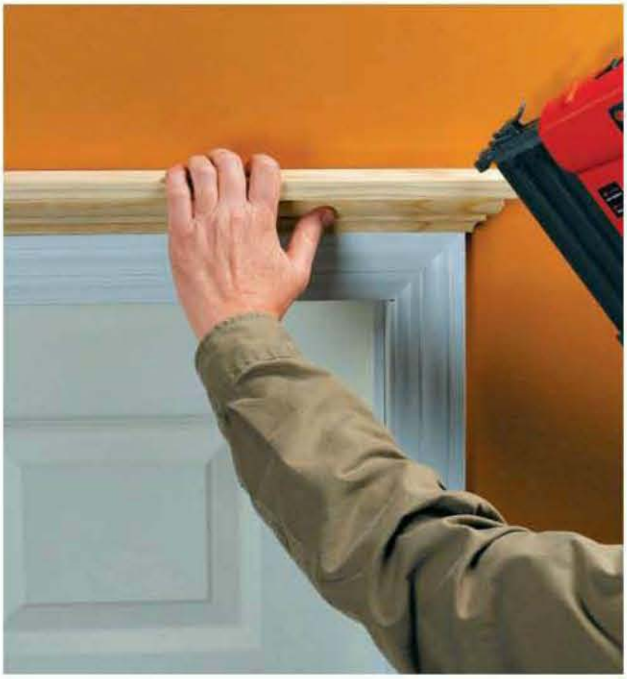

Create a decorative door header by nailing a combination of bed and lattice moldings over the top casing. Size the header to overhang the side casings.

Lighting

Artificial lighting is a necessary component of any room, but beyond simply providing illumination, lighting fixtures are an important part of a room's design. From the low-key but modern appearance of recessed canister lights to the elegance of a crystal chandelier, your choice of fixture will go a long way toward defining your remodeled room.

The easiest lighting improvement is to simply make a one-for-one replacement of an old fixture. The electrical connections are simple and, depending on the style you choose, you may not have to do any alterations to the room at all. Just make sure the light the new fixture provides is adequate for your room. In other words, don't replace an old fixture that holds three 60-watt bulbs with a single 25-watt pendant unless you are providing additional new light.

This chapter shows how to hook up the three types of light fixture you're most likely to include in your remodeling plan: an overhead ceiling light, recessed canister lights, and track lighting. If you need to add new wiring to supply the light, see the information on pages 82 through 147. Always shut off power at the main service panel before beginning even the simplest wiring project. Call a professional if you are unsure about your wiring job.

In this chapter:

- Installing Ceiling Lights
- Installing Recessed Ceiling Lights
- Installing Track Lighting

Installing Ceiling Lights

Ceiling fixtures don't have any moving parts and their wiring is very simple, so, other than changing bulbs, you're likely to get decades of trouble-free service from a fixture. This sounds like a good thing, but it also means that the fixture probably won't fail and give you an excuse to update a room's look with a new one. Fortunately, you don't need an excuse. Upgrading a fixture is easy and can make a dramatic impact on a room. You can substantially increase the light in a room by replacing a globe-style fixture by one with separate spot lights, or you can simply install a new fixture that matches the room's décor.

Tools & Materials ▸

Replacement light fixture
Wire stripper
Voltage sensor
Insulated screwdrivers
Wire connectors

Installing a new ceiling fixture can provide more light to a space, not to mention an aesthetic lift. It's one of the easiest upgrades you can do.

How to Replace a Ceiling Light

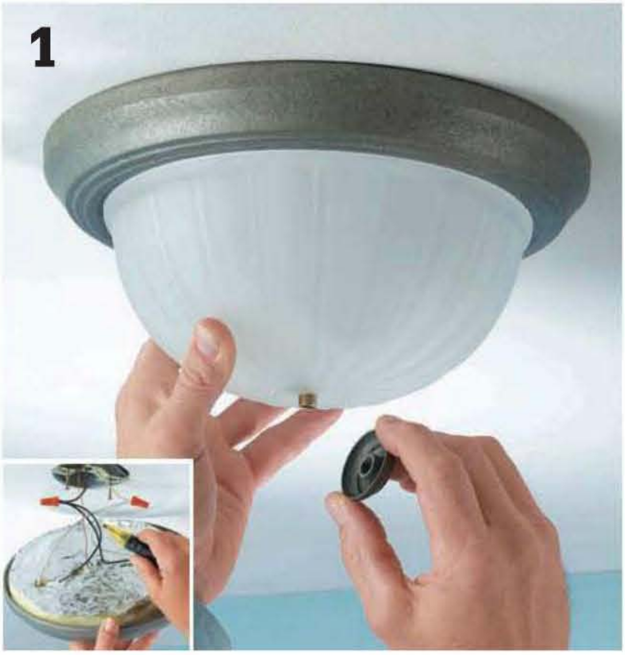

Shut off power to the ceiling light and remove the shade or diffuser. Loosen the mounting screws and carefully lower the fixture, supporting it as you work (do not let light fixtures hang by their electrical wires alone). Test with a voltage sensor to make sure no power is reaching the connections.

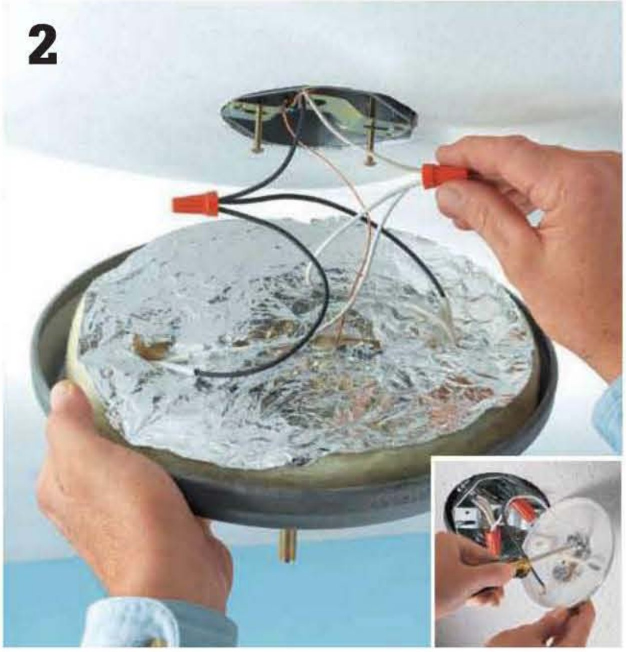

Remove the twist connectors from the fixture wires or unscrew the screw terminals and remove the white neutral wire and the black lead wire (inset).

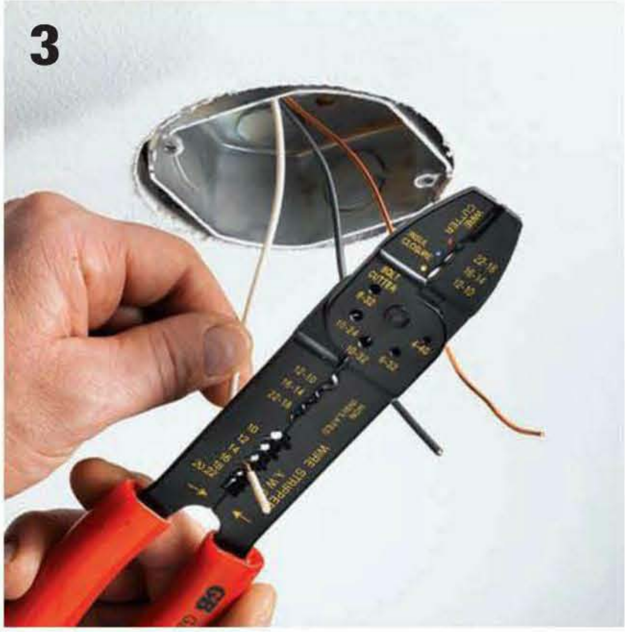

Before you install the new fixture, check the ends of the wires coming from the ceiling electrical box. They should be clean and free of nicks or scorch marks. If they're dirty or worn, clip off the stripped portion with your combination tool. Then strip away about ¾" of insulation from the end of each wire.

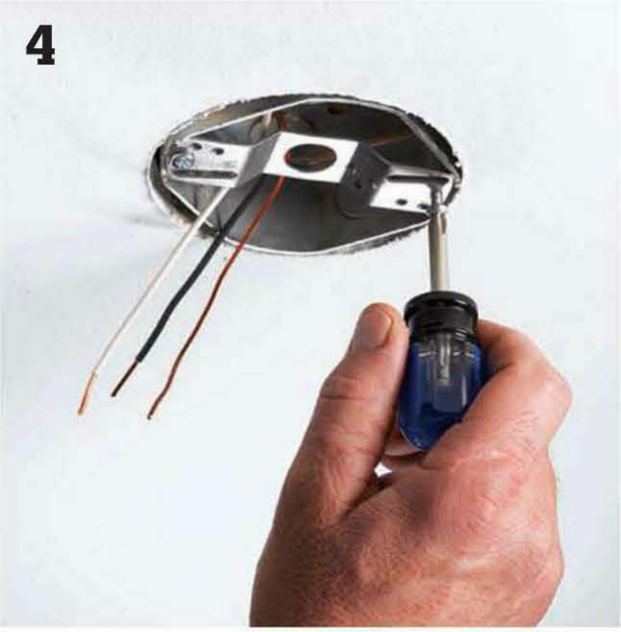

Attach a mounting strap to the ceiling fixture box if there is not one already present. Your new light may come equipped with a strap, otherwise you can find one for purchase at any hardware store.

(continued)

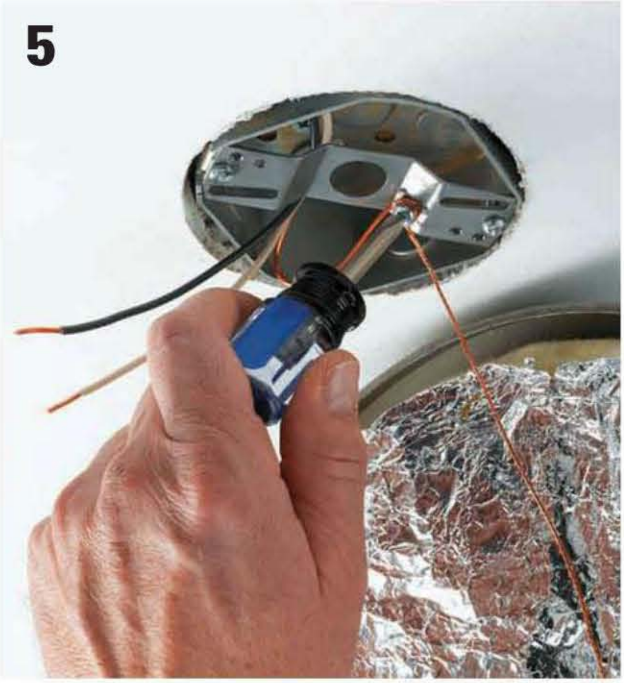

Lift the new fixture up to the ceiling (you may want a helper for this) and attach the bare copper ground wire from the power supply cable to the grounding screw or clip on the mounting strap. Also attach the ground wire from the fixture to the screw or clip.

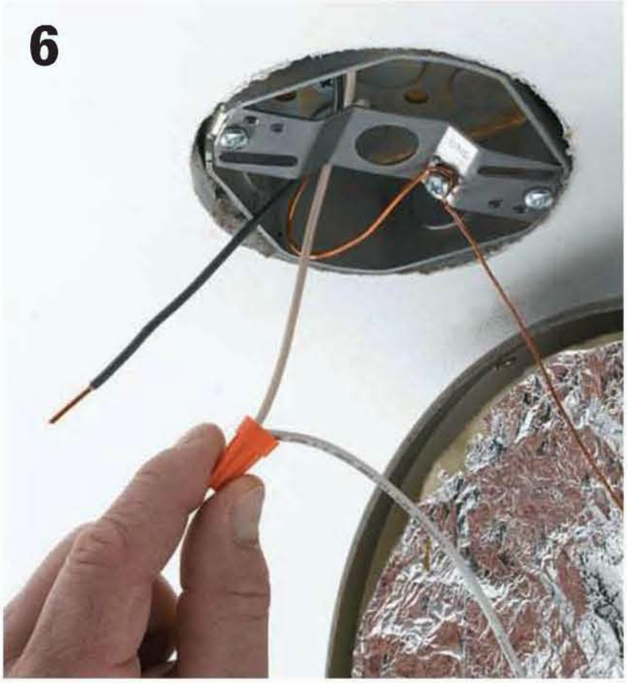

With the fixture supported by a ladder or a helper, join the white wire lead and the white fixture wire with a wire connector (often supplied with the fixture).

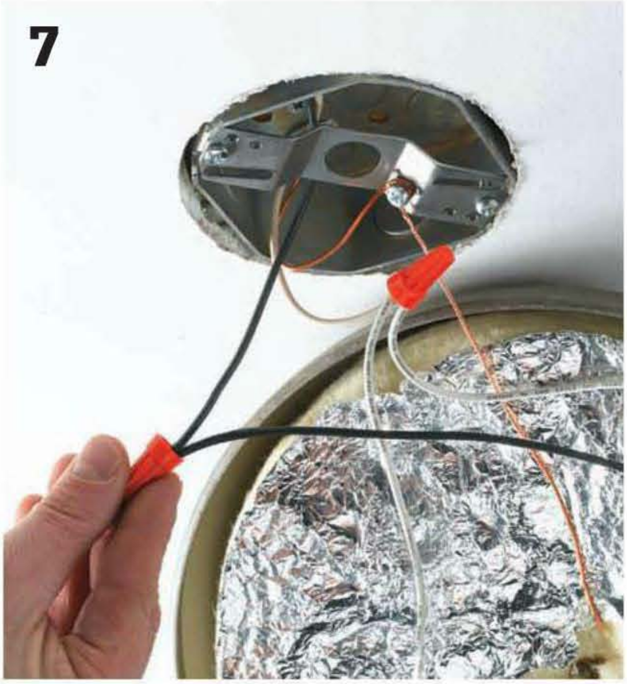

Connect the black power supply wire to the black fixture wire with a wire connector.

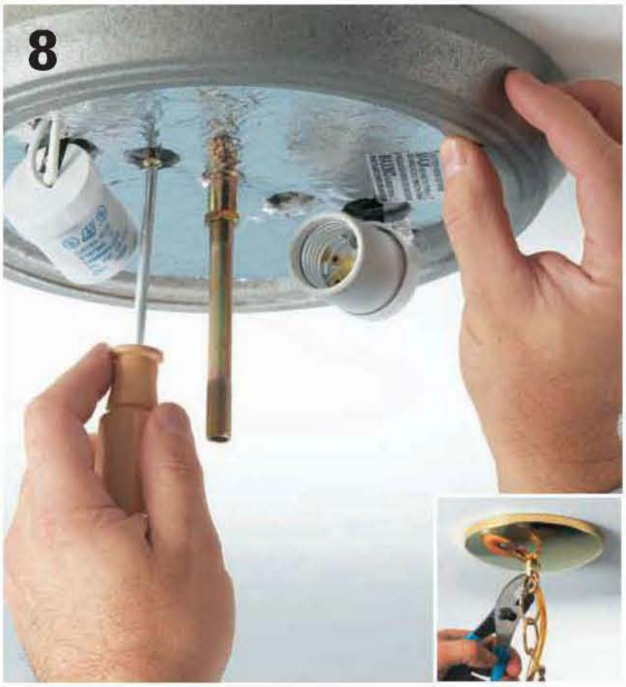

Position the new fixture mounting plate over the box so the mounting screw holes align. Drive the screws until the fixture is secure against the ceiling. *Note: Some fixtures are supported by a threaded rod or nipple in the center that screws into a female threaded opening in the mounting strap (inset).*

Installing Recessed Ceiling Lights

Recessed lights are versatile fixtures suited for a variety of situations. Fixtures rated for outdoor use can also be installed in roof soffits and overhangs for accent and security lighting. Recessed fixtures can also be installed over showers or tubs. Be sure to use fixture cans and trims rated for bathroom use.

There are recessed lighting cans in all shapes and sizes for almost every type of ceiling or cabinet. Cans are sold for unfinished ceilings (new construction) or for finished ceilings (retrofit installation). Cans are also rated as insulation compatible or for uninsulated ceilings. Be sure to use the correct one for your ceiling to prevent creating a fire hazard.

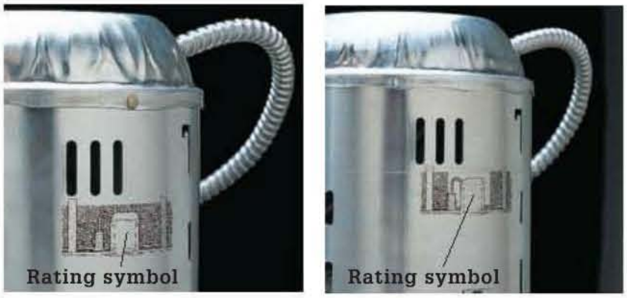

Choose the proper type of recessed light fixture for your project. There are two types of fixtures: those rated for installation within insulation (left), and those which must be kept at least 3" from insulation (right). Self-contained thermal switches shut off power if the unit gets too hot for its rating. A recessed light fixture must be installed at least ½" from combustible materials.

Tools & Materials ▸

- Recessed-lighting can for new construction or remodeling and trim
- Circuit tester
- Cable ripper
- Combination tool
- Pliers
- Fish tape
- Drywall saw
- NM cable

Recessed ceiling lights often are installed in series to provide exacting control over the amount and direction of light. Spacing the canisters in every other ceiling joist bay is a common practice.

How to Install Recessed Ceiling Lights

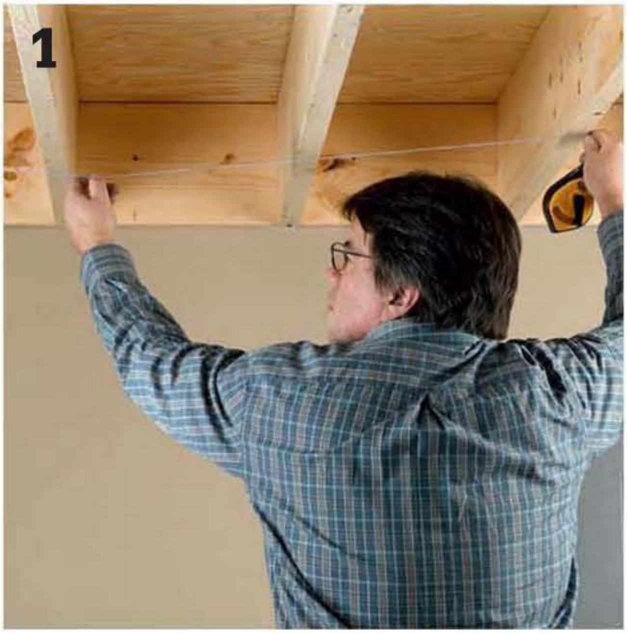

Mark the location for the light canister. If you are installing multiple lights, measure out from the wall at the start and end of the run, and connect them with a chalkline snapped parallel to the wall. If the ceiling is finished with a surface (wallboard), see next page.

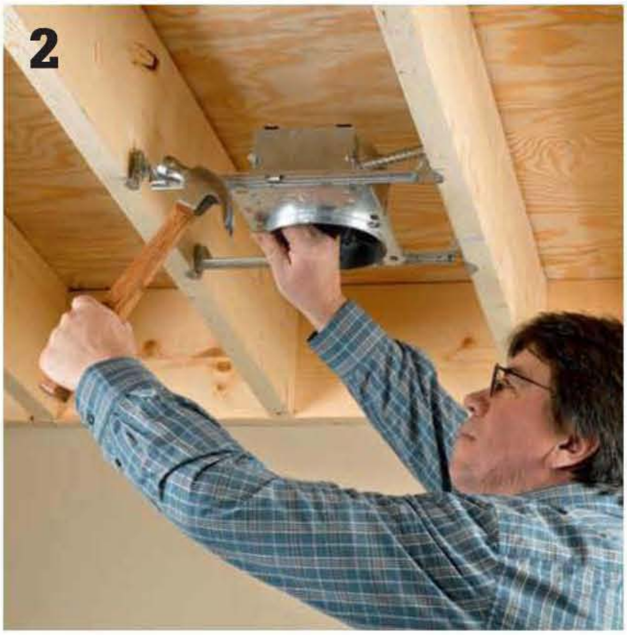

Install the housing for the recessed fixture. Housings for new construction (or remodeling installations where the installation area is fully accessible from either above or below) have integral hanger bars that you attach to the each joist in the joist bay.

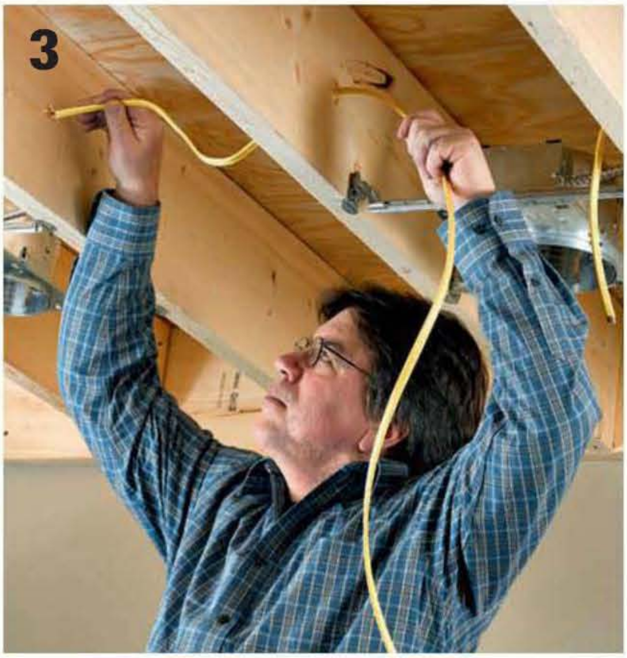

Run electric cable from the switch to each canister location. Multiple lights are generally installed in series so there is no need to make pigtail connections in the individual boxes. Make sure to leave enough extra cable at each location to feed the wire into the housing and make the connection.

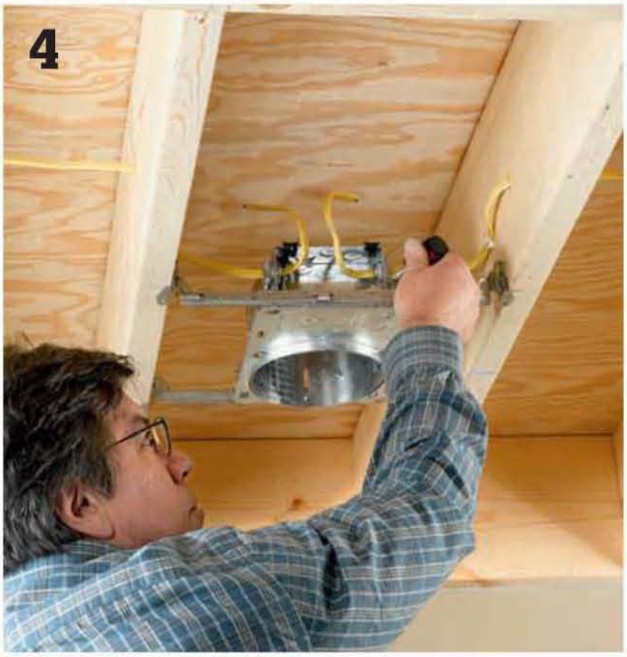

Run the feeder cables into the electrical boxes attached to the canister housings. You'll need to remove knockouts first and make sure to secure the cable with a wire staple within 8" of the entry point to the box.

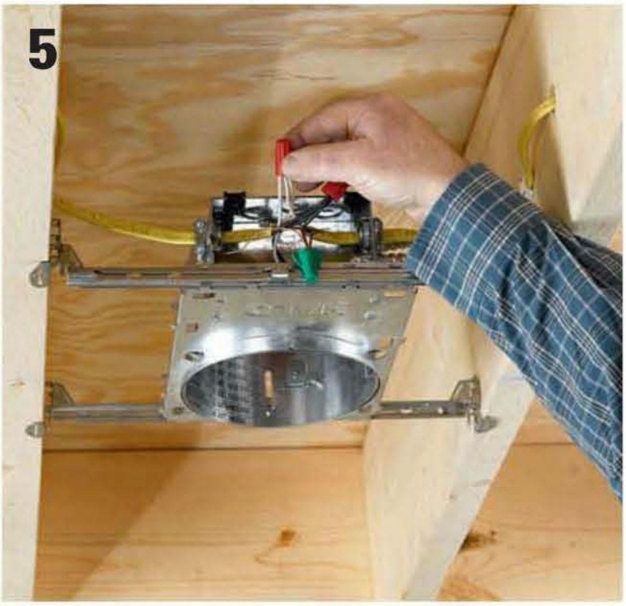

Connect the feeder wires to the fixture wires inside the junction box. Twist the hot lead together with the black fixture wire, as well as the black lead to other fixtures further downline. Also connect the neutral white wires. Join the ground wires and pigtail them to the grounding screw or clip in the box. Finish the ceiling, as desired.

Attach your trim kit of choice. Normally, these are hung with torsion spring clips from notches or hooks inside the canister. This should be done after the ceiling is installed and finished for new construction projects. With certain types of trim kits, such as eyeball trim, you'll need to install the light bulb before the trim kit.

How to Connect a Recessed Fixture Can in a Finished Ceiling

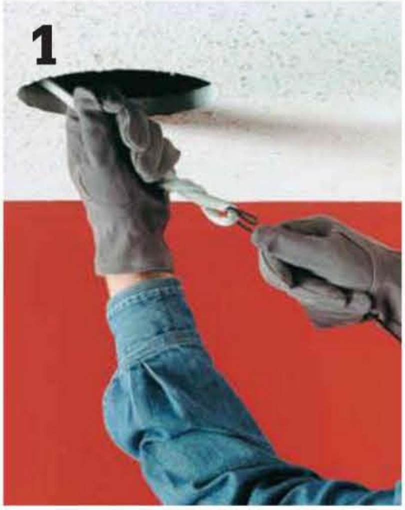

Make the hole for the can. Most fixtures will include a template for sizing the hole. Fish 14/2 cable from the switch location to the hole. Pull about 16" of cable out of the hole for making the connection.

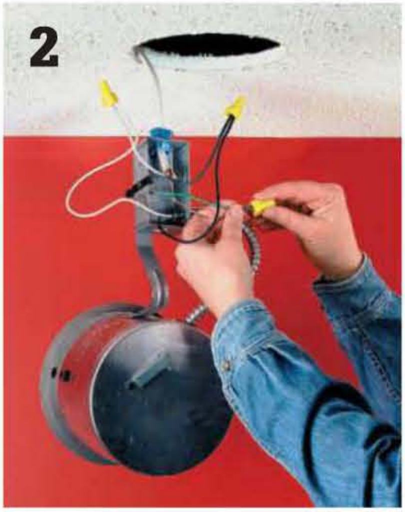

Remove a knockout from the electrical box attached to the can. Thread the cable into the box; secure it with a cable clamp. Remove sheathing insulation. Connect the black fixture wire to the black circuit wire, the white fixture wire to the white circuit wire, and then connect the ground wire to the grounding screw or grounding wire attached to the box.

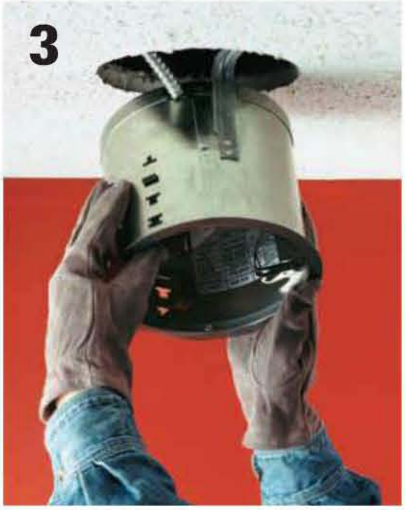

Retrofit cans secure themselves in the hole with spring-loaded clips. Install the can in the ceiling by depressing the mounting clips so the can will fit into the hole. Insert the can so that its edge is tight to the ceiling. Push the mounting clips back out so they grip the drywall and hold the fixture in place. Install the trim piece.

Installing Track Lighting

Track lighting offers a beautiful and functional way to increase the amount of light in a room or simply to update its look. A variety of fixture and lamp options lets you control the shape, color, and intensity of the light. Installing track lighting in place of an existing ceiling-mounted light fixture involves basic wiring and hand-tool skills, but the connections are even easier to make than with traditional light fixtures. Once installed, the system is very easy to upgrade or expand in the future.

Tools & Materials ▸

- Drill/driver and bits
- Wire stripper
- Screwdriver
- Voltage sensor
- Toggle bolts
- Prewired track and fittings
- Track light heads
- Wire connector
- Ceiling box

If you currently have a ceiling-mounted light fixture that is not meeting your lighting needs, it's simple to replace it with a track-lighting fixture. With track lighting you can easily change the type and number of lights, their position on the track, and the direction they aim. These fixtures come in many different styles, including short 3-ft. track systems with just one or two lights up to 12-ft. systems with five or more lights.

How to Install Track Lighting

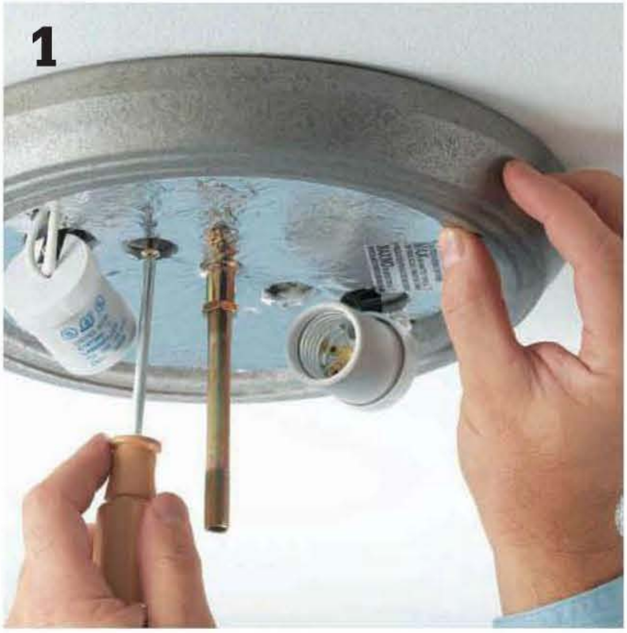

Disconnect the old ceiling light fixture (for remodeling projects) after shutting off power to the circuit at the main service panel. The globe or diffuser and the lamps should be removed before the fixture mounting mechanism is detached.

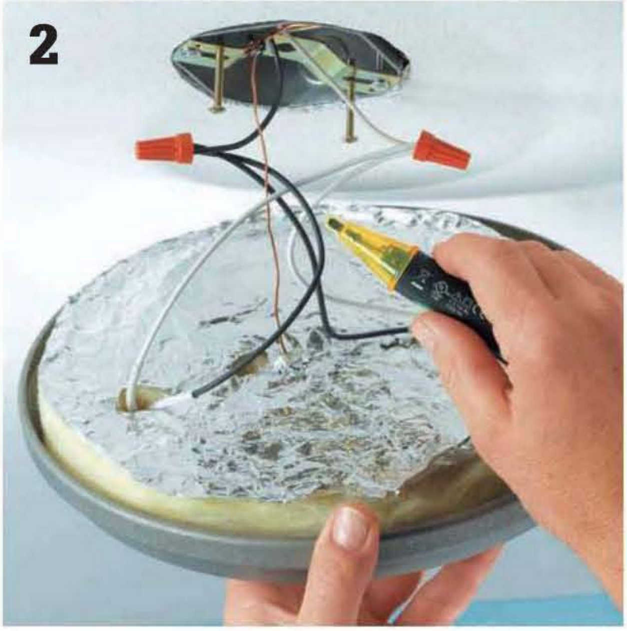

Test the fixture wires with a voltage sensor to make sure the circuit is dead. Support the fixture from below while you work—never allow a light fixture to hang by its electrical wires alone. Remove the wire connectors and pull the wires apart. Remove the old light fixture.

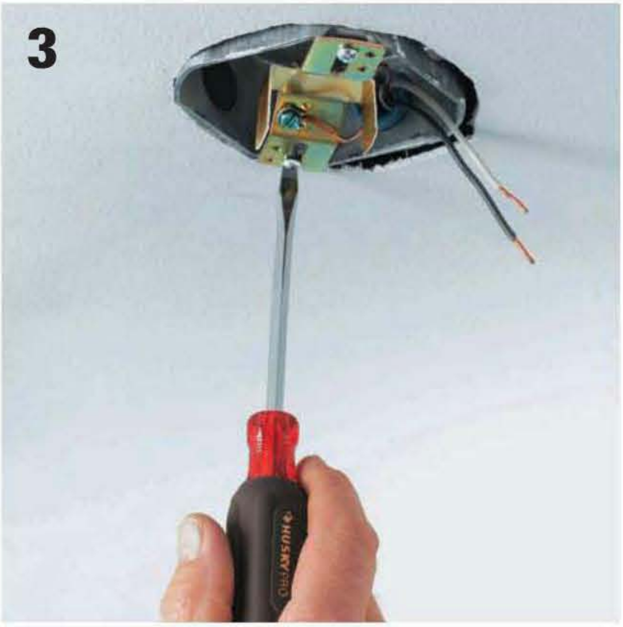

Attach the mounting strap for the new track light to the old ceiling box. If the mounting strap has a hole in the center, thread the circuit wires through the hole before screwing the strap to the box. The green or bare copper ground from the circuit should be attached to the grounding screw or clip on the strap or box.

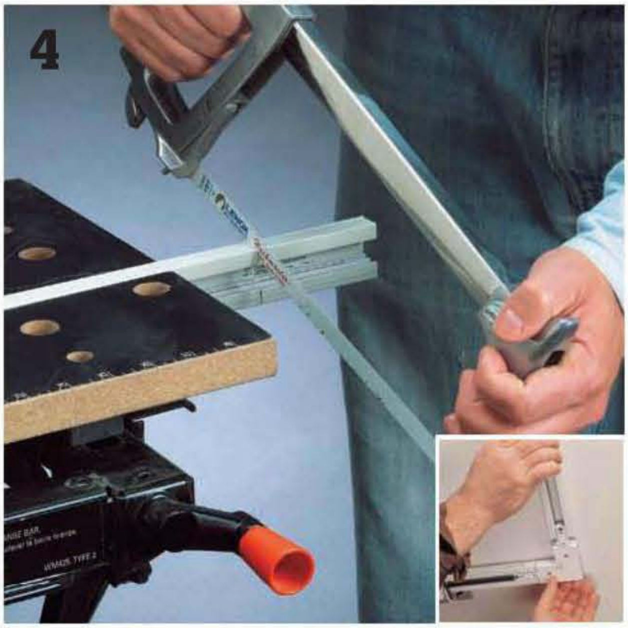

Cut the track section to length, if necessary, using a hack saw. Deburr the cut end with a metal file. If you are installing multiple sections of track, assemble the sections with the correct connector fittings (sold separately from your kit). You can also purchase T-fittings or L-fittings (inset photo) if you wish to install tracks in either of these configurations.

(continued)

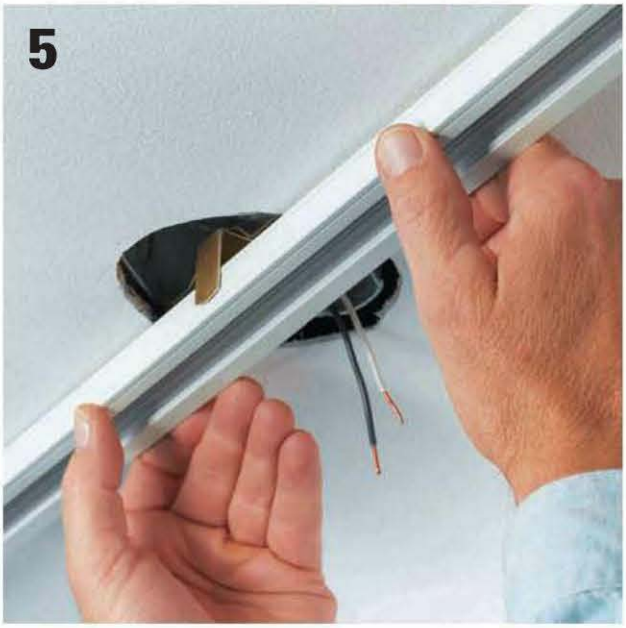

Position the track section in the mounting saddle on the mounting strap and hold it temporarily in place in the location where it will be installed. The track section will have predrilled mounting holes in the back. Draw a marking point on the ceiling at each of these locations. If your track does not have predrilled mounting holes, remove it and drill a 3/16" hole in the back every 16".

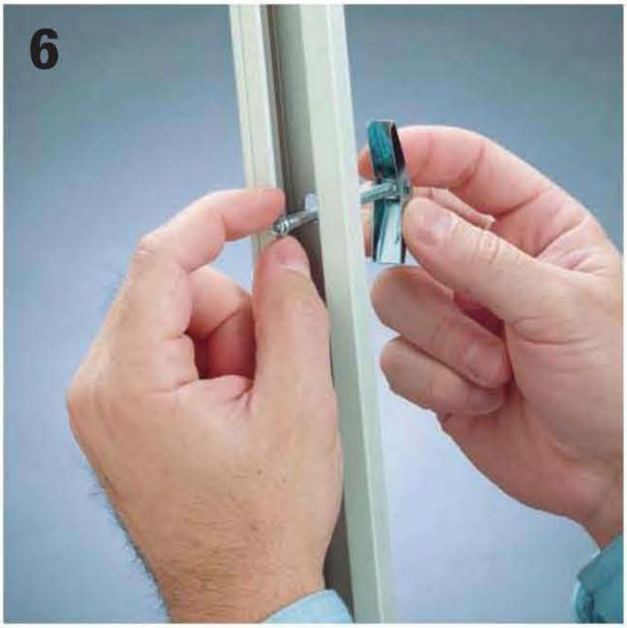

Insert the bolt from a toggle bolt or molly bolt into each predrilled screw location and twist the toggle or molly back onto the free end. These types of hardware have greater holding power than anchor sleeves. Drill a 5/8" dia. access hole in the ceiling at each of the mounting hole locations you marked on the ceiling in step 5.

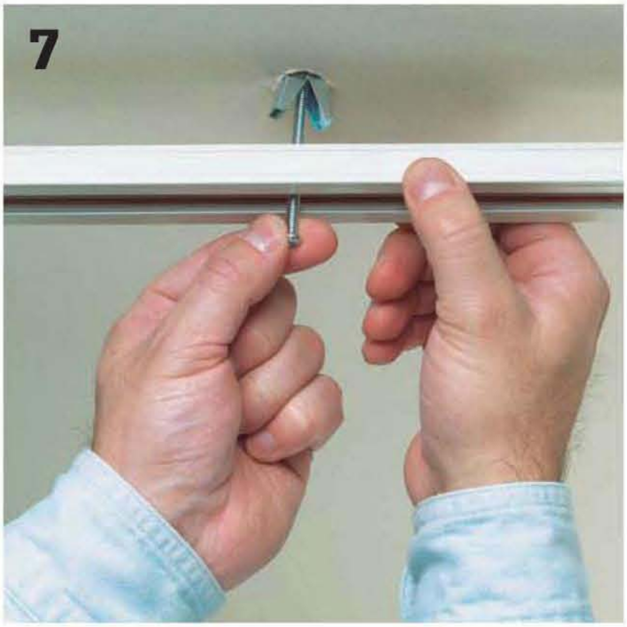

Insert the toggle or molly into the access hole far enough so it clears the top of the hole and the wings snap outward. Then tighten each bolt so the track is snug against the ceiling. If the mounting hole happens to fall over a ceiling joint, simply drive a wallboard screw at that hole location.

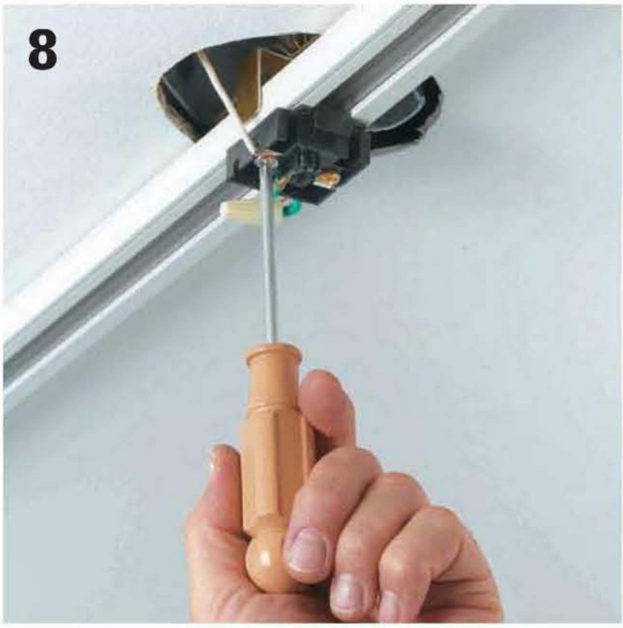

Hook up wires from the track's power supply fitting to the circuit wires. Connect black to black and white to white. The grounding wire from the power supply fitting can either be pigtailed to the circuit ground wire and connected to the grounding screw or clip, or it can be twisted together with the circuit grounding wire at the grounding terminal. Snap the fitting into the track if you have not already done so.

9

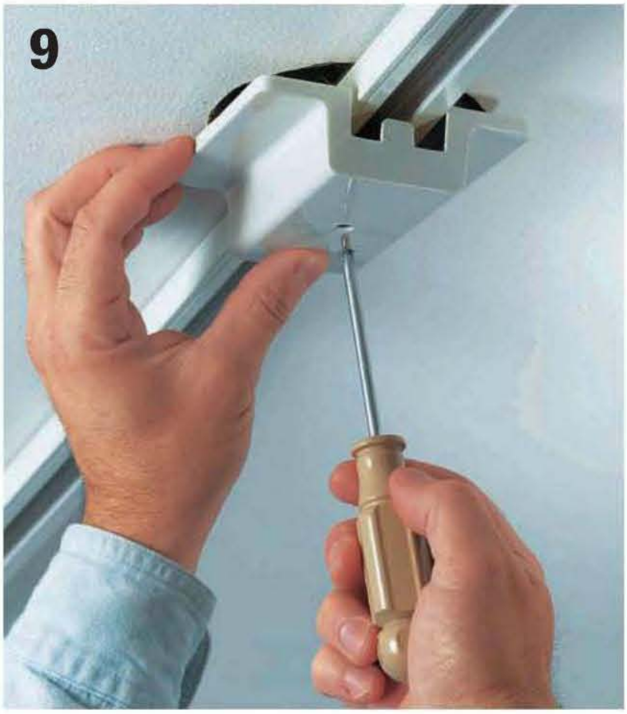

Attach the protective cover that came with your kit to conceal the ceiling box and the electrical connections. Some covers simply snap in place, others require a mounting screw.

10

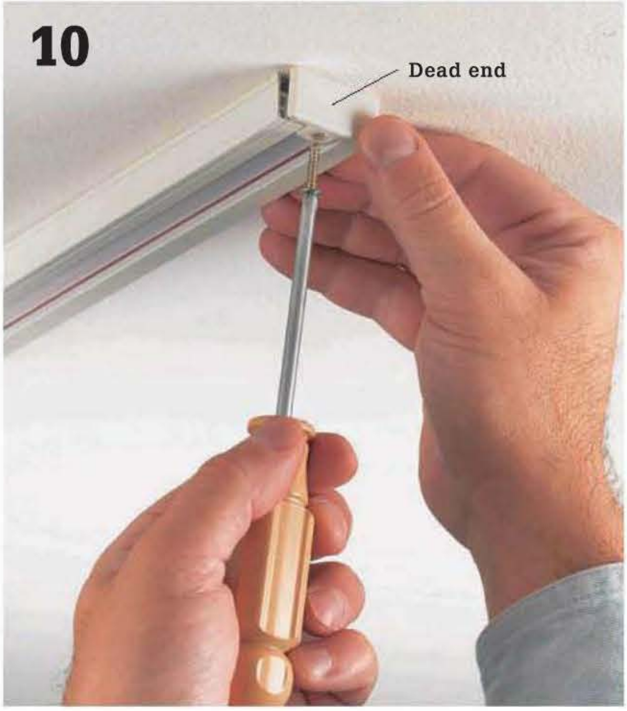

Cap the open ends of the track with a dead end cap fitting. These also may require a mounting screw. Leaving track ends open is a safety violation.

11

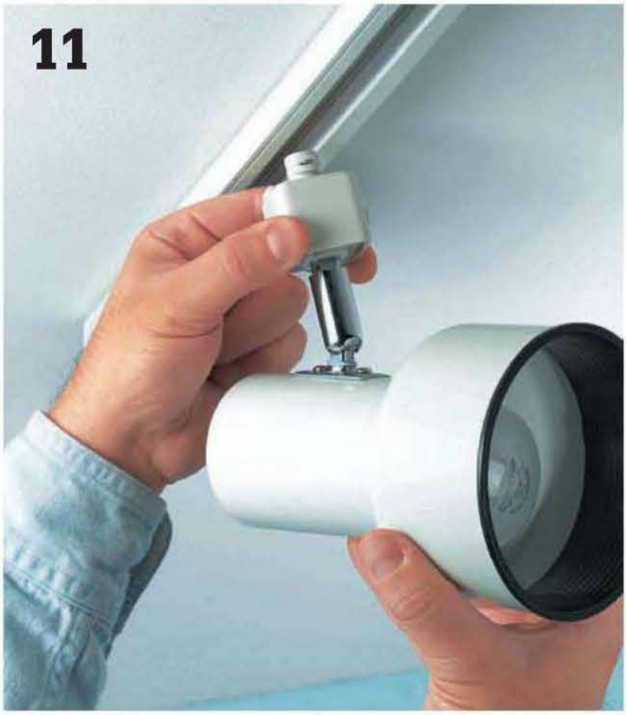

Insert the light heads into the track by slipping the stem into the track slot and then twisting it so the electrical contact points on the head press against the electrified inner rails of the track slot. Tug lightly on the head to make sure it is secure before releasing it.

12

Arrange the track light heads so their light falls in the manner you choose, and then depress the locking tab on each fixture to secure it in position. Restore power and test the lights.

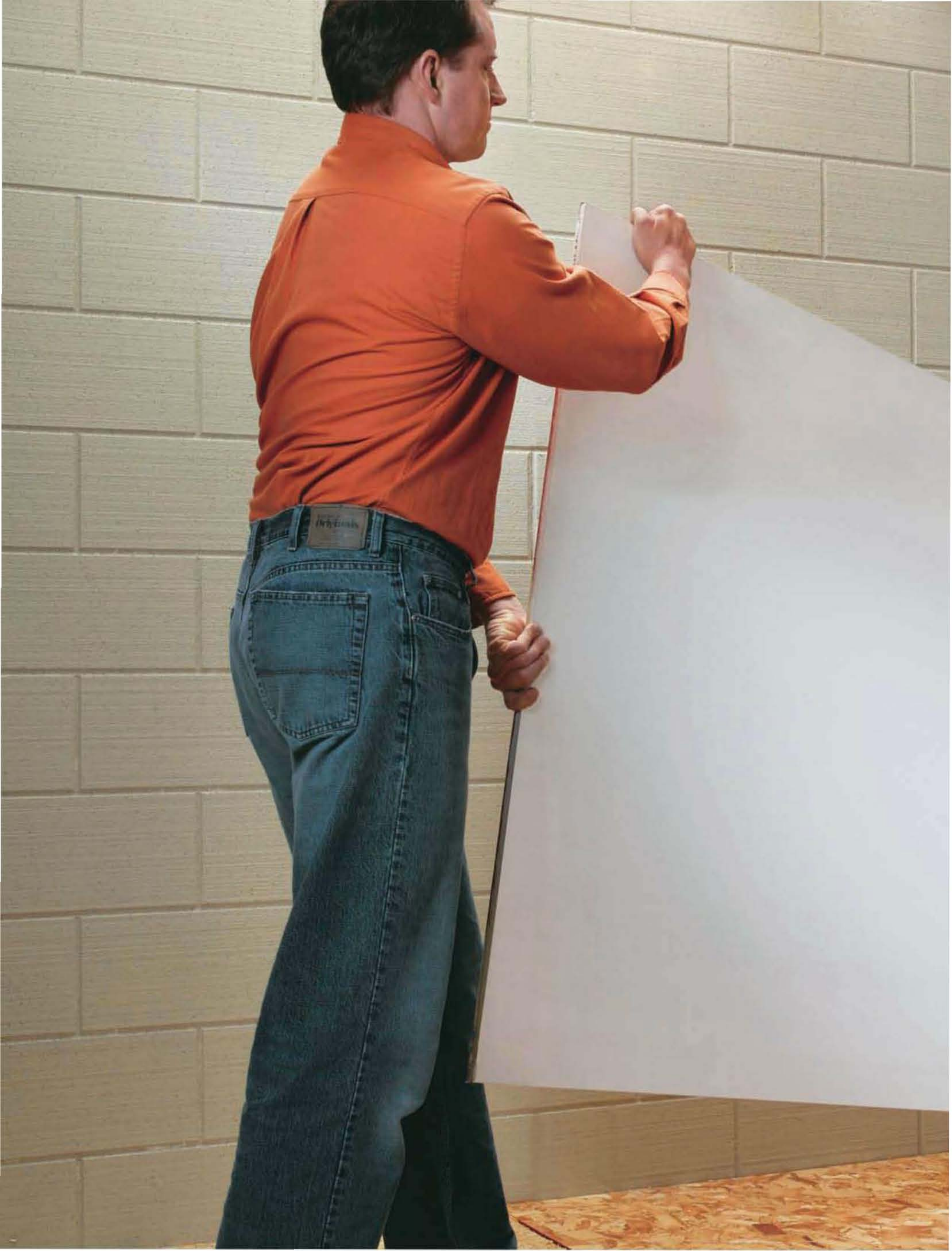

MAJOR REMODELING

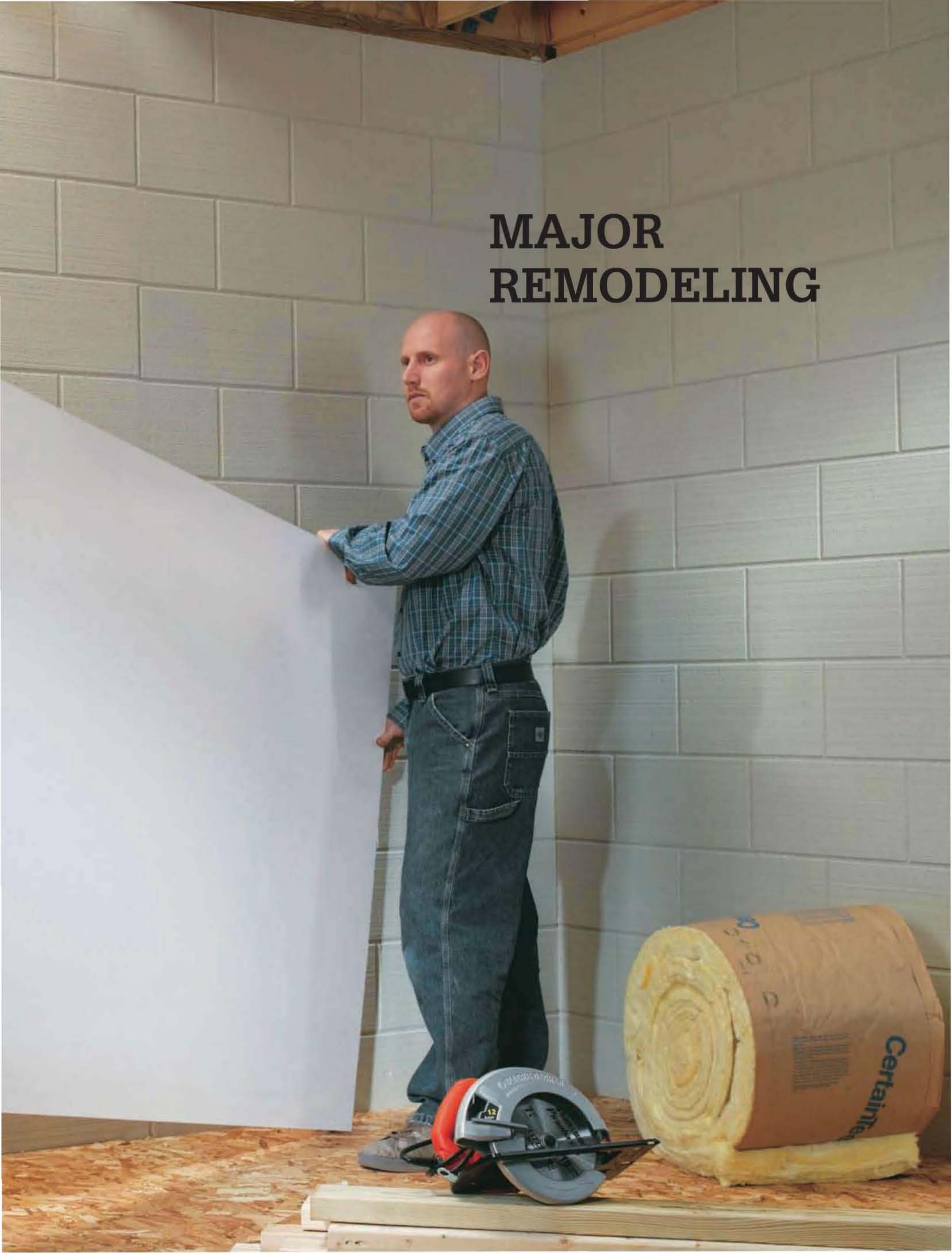

Kitchen Remodeling

Kitchens are remodeled with greater frequency than any other room in the house. This happens often for reasons of style: keeping up with design trends is important to many homeowners. But it also happens for more practical reasons. Kitchens receive a vast amount of wear and tear that results from high usage. Appliances stop working, flooring wears out, cabinets fail, even windows and doors suffer from constant wear. High humidity levels are also an issue in kitchens, where steam is a necessary byproduct of daily life.

The information in this chapter is unique to kitchen remodeling projects. But in practice, you will need to refer to the other chapters in this book to complete all of the design, planning and installation work that go into creating a new kitchen.

In this chapter:

- Planning & Designing Kitchens
- Installing Kitchen Cabinets
- Building Kitchen Islands
- Building Tiled Islands
- Installing Postform Countertops
- Installing Laminate Countertops
- Installing Tile on Countertops
- Installing Tile Backsplashes
- Installing Kitchen Sinks
- Replacing Kitchen Faucets
- Connecting Drains
- Hooking Up Dishwashers
- Installing Food Disposers
- Installing Undercabinet Lights

Planning & Designing Kitchens

Your motives for remodeling your kitchen probably fall into one of two categories: efficiency or appearance. In other words, either your kitchen is inconvenient for your family to use, or you just don't like the way it looks. Or maybe it's both dysfunctional and ugly. A logical place to start, then, is by documenting the elements that need improvement.

Begin by simply taking the time to observe how you now use your kitchen. Keep a notepad handy and jot down any major or minor problems and annoyances that prevent you from being as efficient or as comfortable as you'd like to be when cooking or eating. Also consider how your kitchen is used for special occasions and what issues arise at those times. Don't forget to make note of the positives. If you love having that window over your kitchen sink or the skinny cabinet next to the stove, jot down those observations as well.

Once you have documented the problems with your kitchen, give your imagination permission to roam. Don't worry about money yet. This is your dream kitchen, after all, and there will be plenty of time to bring your fantasies down to earth as you begin the planning stages. Now is the time to consider every possibility. Look at friends' kitchens, at magazine kitchens and at model kitchens. When looking at photographs like the ones included in this book, look at both the overall effect and the individual components. Don't let an odd color scheme scare you away from a faucet that you love. Use a file folder to collect pictures of kitchens or items you like. When it comes down to business, you'll have a refined idea of what you're looking for.

If your kitchen needs remodeling, you will know it. The trick in coming up with a workable design for your space is to get past the "ick" factor and really examine what you have to work with. When you assess your cabinets, you may notice immediately that they are very dark and outdated in appearance. But go past that and ask yourself if they are laid out efficiently. Imagine them as brand new cabinets in the same arrangement, and then try and think of ways you could improve the layout. Whether it's with cabinets or countertops or lighting or appliances, consider the components of your kitchen as generic elements.

Special Considerations

If you're going through the trouble of remodeling your kitchen—and the remodel is not solely for resale value—consider making room for some of your own interests and idiosyncrasies. A pull out recycling center with bins for separating different categories is always a smart idea. A built-in beer refrigerator with a custom tap is great for a home-brew enthusiast. A pull out shelf custom-sized to display a selection of tea is perfectly located next to the kettle. Another, perhaps more practical, investment is the careful measurement of counter heights and seat heights to ergonomically suit your personal needs. Finally, consider how your kitchen will be used over the next five to twenty years. The needs of a family with young children are much different than the needs of empty nesters.

It's okay to hang onto elements you love. Even if your favorite pullout shelf or work table or stool doesn't immediately appear to go with the new elements you hope to introduce, as the owner of the kitchen you're allowed to hang onto the things you're used to working with.

Design Standards

While you may imagine that the answer to every cramped kitchen is to knock down walls and add more space, that's often not the easiest or best solution. One alternative to tearing down the walls, is to add more windows for added light. An extra door, or even a pass-through window from the kitchen to an adjacent room, can also help make the space feel more airy. In short, the cramped feeling may only be one of perception.

Most kitchens fall into one of four categories: Galley, L-shaped, U-shaped and Open Plan. Whether they are small or large, old or new, these floor plans have proven to be popular models for efficient kitchens, though that doesn't mean they'll necessarily be the most efficient for your needs.

Galley: In small homes or city apartments, the galley kitchen is a space-saving choice that is ideal for one or two users. The components may all be lined up along one wall or divided between two parallel walls. For this floor plan to work best, the central galley space should be large enough to allow for all appliances to be open at the same time, with enough space remaining for someone to walk through the middle.

L-shape: This corner kitchen layout can feel roomier than it is in reality, because of the L-shape floor plan. However, the two "arms" might also create an awkward workstation, with little room to set items down mid-way through a meal. Consider taking advantage of the lost central space with a counter-height dining table that can double as a prep station or even a freestanding central island.

U-shape: This layout takes the benefits of a galley kitchen—space-saving solution plus accessibility for one user—and adds a third wall to create the ultimate triangular floor plan. In a compact space, placing the sink at the far end with the refrigerator and range on opposing walls creates a simple workstation. On a larger scale, a sizable island can anchor the center of the room and provide more storage.

Open Plan: Whether the kitchen is located in the middle of a larger great room or off to one side, an open plan layout works best with the addition of a central island or a dividing counter of some type. Perfect for family kitchens or entertaining spaces where the cooks can easily visit with guests, it's important for this floor plan to be well organized so that all necessities are close at hand.

Whatever layout plan you opt for, the most fundamental principle that you don't want to violate is to maintain plenty of free area between appliances, sinks and other elements of the work triangle. In a galley kitchen like the one seen here, the corridor between kitchen walls should be at least 4 ft. wide, and preferably wider.

Common Kitchen Layouts

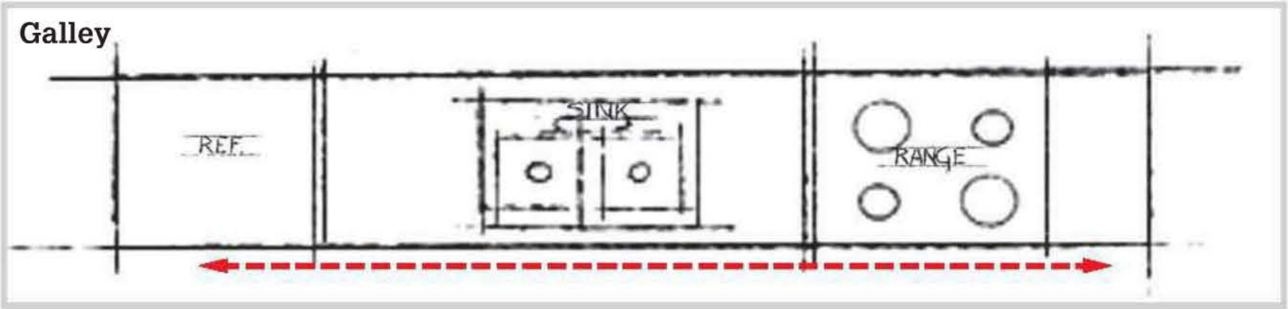
Galley
REF.
SINK
RANGE

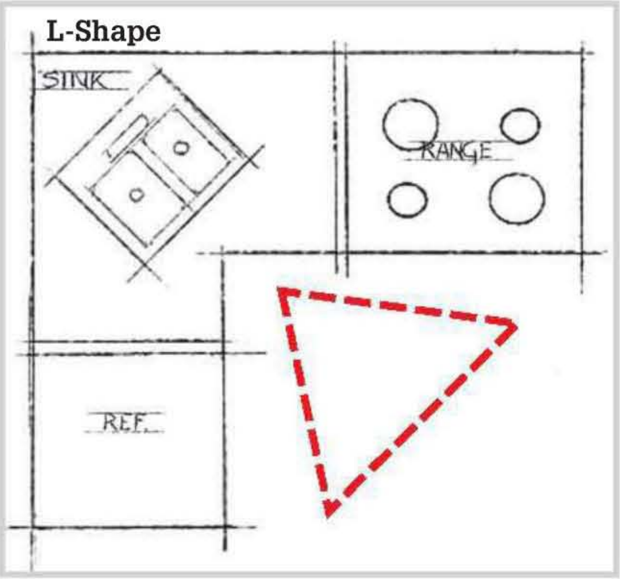
L-Shape
SINK
RANGE
REF.

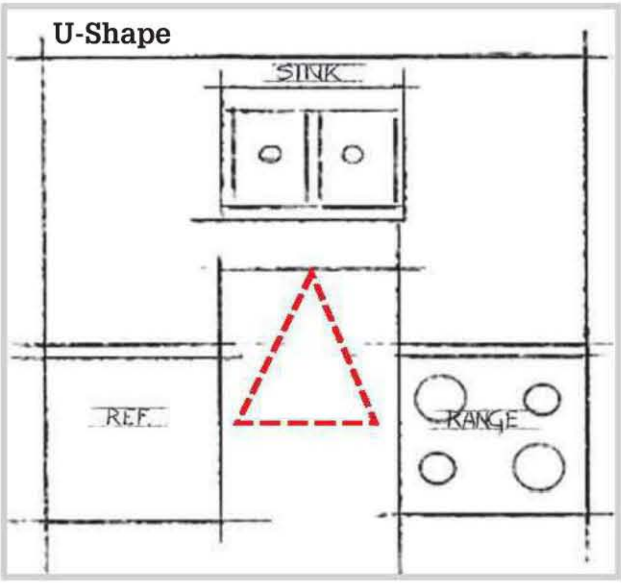
U-Shape
SINK
REF.
RANGE

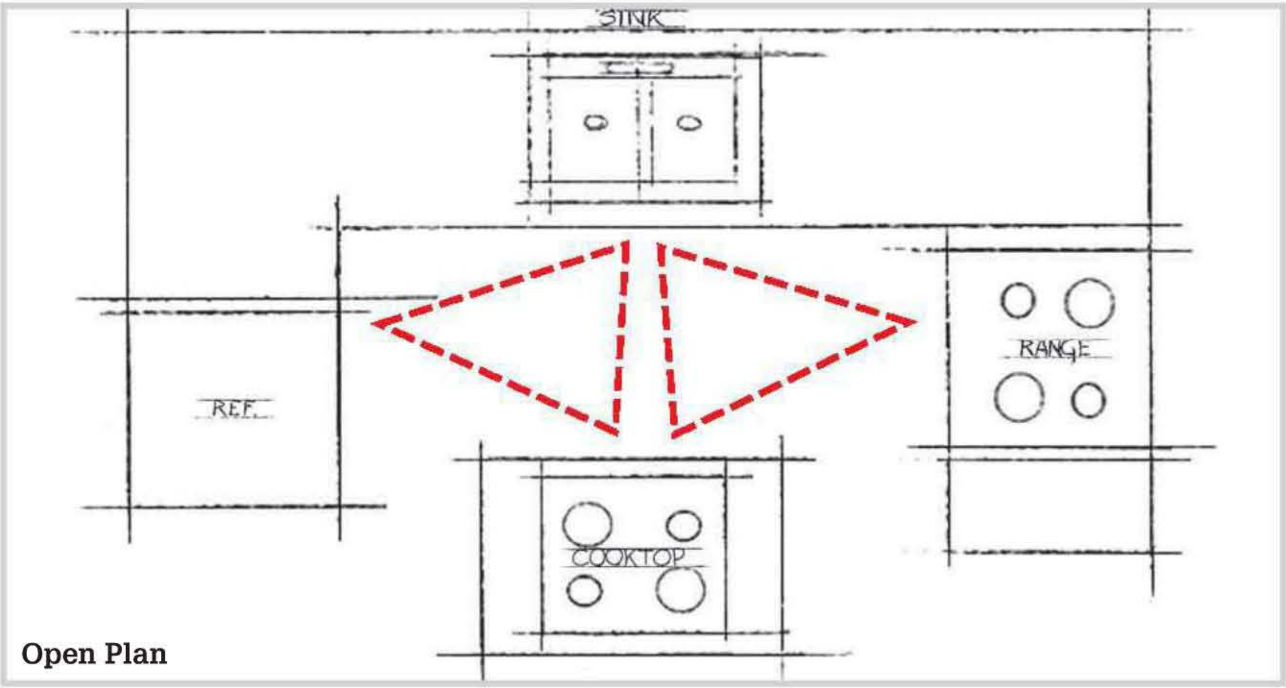
Open Plan
SINK
REF.
RANGE
COOKTOP

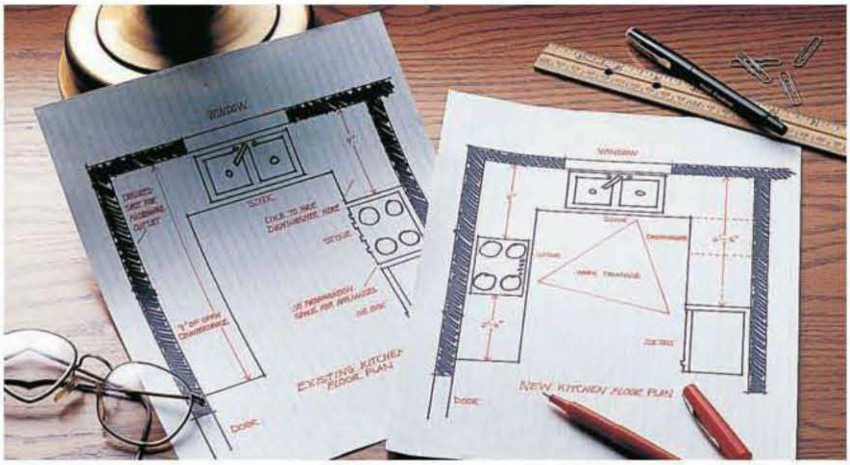

The work triangle is a layout concept that lets you develop a convenient arrangement of the range, sink and refrigerator in the kitchen.

Whether you are doing the work yourself or hiring others, once you have a good idea of the features you want in your new kitchen, it's time to create detailed plan drawings. Good planning drawings will help you in several phases of the planning process:

- Selecting cabinets and appliances to fit your kitchen layout.
- Soliciting accurate work bids when negotiating with plumbers, electricians and other subcontractors.
- Obtaining a building permit at your local Building Inspections office.
- Scheduling the stages of a remodeling project.
- Evaluating the work of contractors. If a carpenter or cabinetmaker fails to meet your expectations, your plan drawings serve as proof that the contractor did not complete the work as agreed.

CODES & STANDARDS

Creating plans for a kitchen can seem like an overwhelming challenge, but fortunately there are guidelines available to help you. Some of these guidelines are legal regulations specified by your local Building Code and must be followed exactly. Most codes have very specific rules for basic construction, as well as for plumbing and electrical installations.

Another set of guidelines, known as standards, are informal recommendations developed over time by kitchen designers, cabinetmakers and appliance manufacturers. These design standards suggest parameters for good kitchen layout, and following them helps ensure that your kitchen is comfortable and convenient to use.

GUIDELINES FOR LAYOUT

The goal of any kitchen layout is to make the cook's work easier and, where possible, to allow other people to enjoy the same space without getting in the way. Understanding the accepted design standards can help you determine whether your present layout is sufficient or if your kitchen needs a more radical layout change or expansion.

Work triangle & traffic patterns. A classic kitchen design concept, the work triangle theory proposes that the sink, range and refrigerator be arranged in a triangular layout according to the following guidelines:

- Position of the triangle should be such that traffic flow will not disrupt the main functions of the kitchen.
- Total distance between the corners of the triangle should be no more than 26 ft. and no less than 12 ft.
- Each side of the triangle should be between 4 and 9 ft. in length.

If two people frequently work in the kitchen simultaneously, the layout should include two work triangles. In a two-triangle kitchen, the triangles may share one side, but they should not cross one another.

Don't fret too much if you can't make the triangle layout work perfectly. Some kitchens, for example, may have four workstations instead of three, and others may not have enough space to accommodate the classic triangle.

For general traffic design, it is recommended to leave 4-ft. "corridors" between all stationary items for walking comfort. Some designers will allow this standard to be reduced to 3 feet in smaller kitchens.

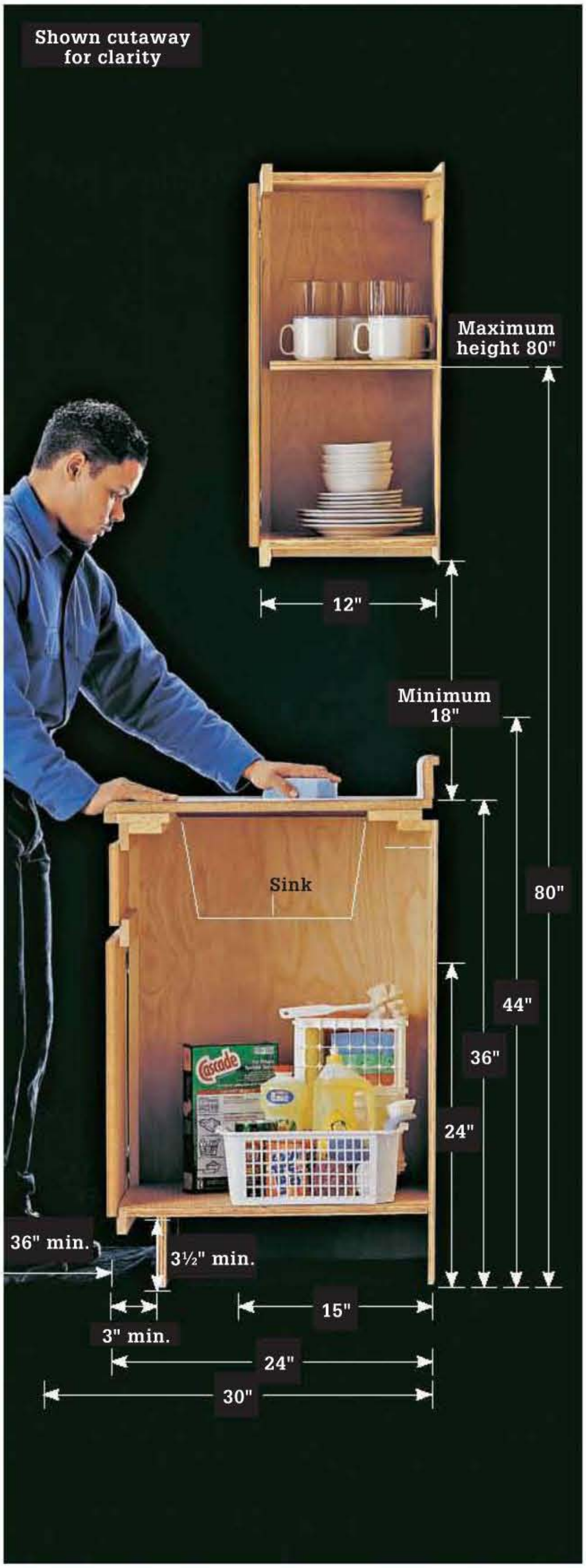

Dimensions and positions of cabinets follow accepted design standards, as shown here.

The sizes of base cabinets and wall cabinets are fairly uniform among manufacturers, and unless you have them custom-built in unusual sizes, they will conform to the following standards:

- Base cabinets: height—34½"; depth—23" to 24"; width—9" to 48", in 3" increments.
- Wall cabinets: height—12", 15", 18", 24", 30", 33", 42"; depth—12"; width—24", 30", 33", 36", 42", 48".
- Oven cabinets: height—84", 96"; depth—24"; width—27", 30", 33".
- Utility cabinets: height—84"; depth—12", 24"; width—18", 24", 36".

Not every manufacturer will offer all these sizes and styles, so it's a good idea to obtain product catalogs when planning the layout of cabinets. Some other tips:

- Use functional corner cabinets rather than "blind" cabinets that provide no access to the corner area.
- Include at least five storage/organizing units, such as swing-out pantry units, appliance garages and specialized drawers or shelves.

Eating areas. Kitchen tabletops and countertops used for dining are generally positioned 30", 36", or 42" above the floor, and the recommended space for each person varies according to the height of the surface.

Islands. A kitchen island should be positioned so there is at least 36" of clear space between the edges of its countertop and surrounding walls or cabinets

GUIDELINES FOR BASIC CONSTRUCTION

Plans for a major remodeling project that involves moving or adding walls, or building a new room addition must accurately show the locations and dimensions of the new walls and all doors and windows. This will allow the construction carpenter to give you an accurate bid on the work and will allow him to obtain the necessary building permits. If you will be moving walls or adding windows or doors, you must identify load-bearing walls and provide appropriate support during removal and rebuilding.

WINDOWS

Most Building Codes require that kitchens have at least one window, with at least 10 sq. ft. of glass area. Some local Building Codes, however, will allow windowless kitchens, so long as they have proper venting. Kitchen designers recommend that kitchens have windows, doors or skylights that together have a total glass surface area equal to at least 25% of the total floor area.

DOORS

Exterior entry doors should be at least 3 ft. wide and 6½ ft. high. Interior passage doors between rooms must be at least 2½ ft. wide. A kitchen must have at least two points of entry, arranged so traffic patterns don't intrude on work areas.

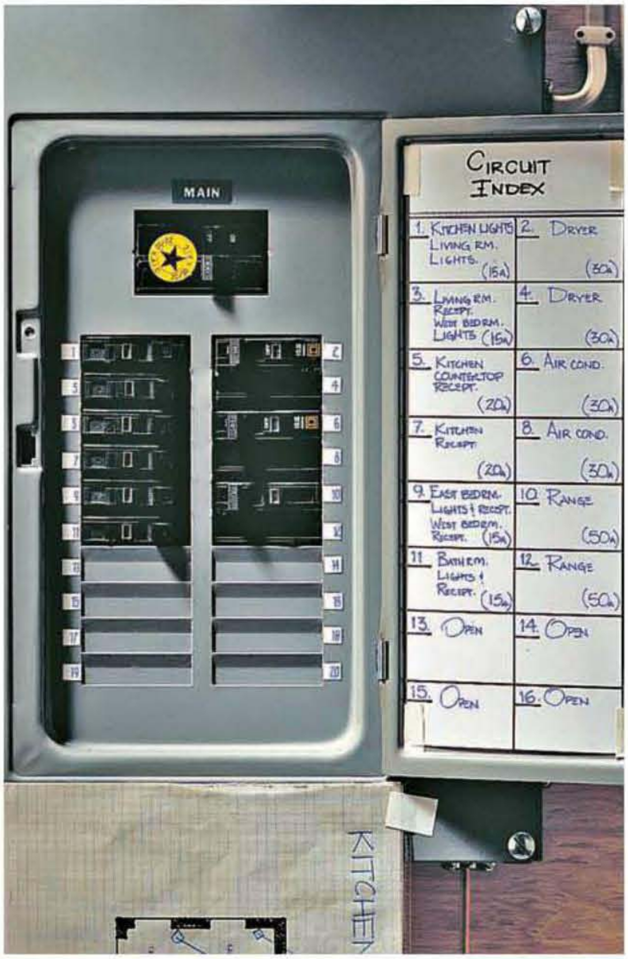
MAIN
CIRCUIT INDEX
1. KITCHEN LIGHTS LIVING RM. LIGHTS. (15A)
2. DRYER (30A)
3. LIVING RM. RECEPT. WEST BEDRM. LIGHTS (15A)
4. DRYER (30A)
5. KITCHEN COUNTERTOP RECEPT. (20A)
6. AIR COND. (30A)
7. KITCHEN RECEPT (20A)
8. AIR COND. (30A)
9. EAST BEDRM. LIGHTS & RECEPT. WEST BEDRM. RECEPT. (15A)
10. RANGE (50A)
11. BATHRM. LIGHTS & RECEPT. (15A)
12. RANGE (50A)
13. OPEN
14. OPEN
15. OPEN
16. OPEN
KITCHEN

Examine your circuit breaker panel. It may have an index that identifies circuits serving the kitchen. If your service panel has open slots, an electrician can add additional kitchen circuits relatively easily. If your service panel is full, he may have to install a new service panel at additional cost.

GUIDELINES FOR ELECTRICAL SERVICE & LIGHTING

Nearly any kitchen-remodeling project will require some upgrading of the electrical service. While your old kitchen may be served by a single 120-volt circuit, it's not uncommon for a large modern kitchen to require as many as seven individual circuits. And in a few cases, the extra demands of the new kitchen may require that the main electrical service for your entire house be upgraded by an electrician. By comparing the electrical service in your present kitchen with the requirements described below, you'll get an idea of how extensive your electrical service improvements will need to be. Your plan drawings should indicate the locations of all the outlets, lighting fixtures and electrical appliances in your new kitchen.

The National Electric Code requires the following for kitchens:

- Two small-appliance circuits (120-volt, 20-amp) to supply power for the plug-in countertop appliances.
- Wall outlets spaced no more than 12 ft. apart.
- Countertop outlets spaced no more than 4 ft. apart.
- GFCI (ground-fault circuit interrupter), protected receptacles installed in any general use outlet, whether above counter or at floor level.
- Dedicated circuits for each major appliance. Install a 20-amp, 120-volt circuit for a built-in microwave, a 15-amp circuit for the dishwasher and food disposer. An electrical range, cooktop or wall oven requires a dedicated 50-amp, 240-volt circuit.

The Electric Code only requires that a kitchen have some form of lighting controlled by a wall switch, but kitchen designers have additional recommendations:

- A general lighting circuit (120-volt, 15-amp) that operates independently from plug-in outlets.
- Plentiful task lighting, usually mounted under wall cabinets or soffits, to illuminate each work area.
- Decorative lighting fixtures to highlight attractive cabinets or other features of the kitchen.

GUIDELINES FOR PLUMBING

If your new kitchen layout changes the location of the sink, or if you are planning to add an additional sink or dishwasher, the water supply and drain pipes will need to be upgraded. Your plan drawings should indicate these intended changes.

Extending plumbing lines for a new kitchen is often fairly easy and surprisingly inexpensive, but there are some exceptions you should note:

Old pipes. If your present plumbing is more than 25 years old, there is a good chance the plumber will recommend replacing the pipes before installing the kitchen fixtures. Depending on circumstances, this can be an expensive proposition, but if you're faced with this decision, we strongly urge you to take a deep breath and do what the plumber suggests. Those corroded old pipes will need to be replaced someday, and this work is easier and cheaper if you're already in the process of remodeling the kitchen.

Outdated systems. Older plumbing systems may have drain trap and vent arrangements that violate modern Code requirements. If your plumber needs to run all-new vent pipes, this will increase the costs.

Island sinks. If your new kitchen will include an island sink, your plumber will need to run vent pipes beneath the floor. For this reason, plumbing an island sink is more expensive than plumbing a wall sink.

GUIDELINES FOR HEATING, VENTILATION & AIR-CONDITIONING

Your plan drawings should also show the locations of heating/air-conditioning registers or fixtures in your proposed kitchen. If you're planning a cosmetic make-over or a simple layout change, there is a pretty good chance you can get by with the same registers, radiators or heaters found in your present kitchen.

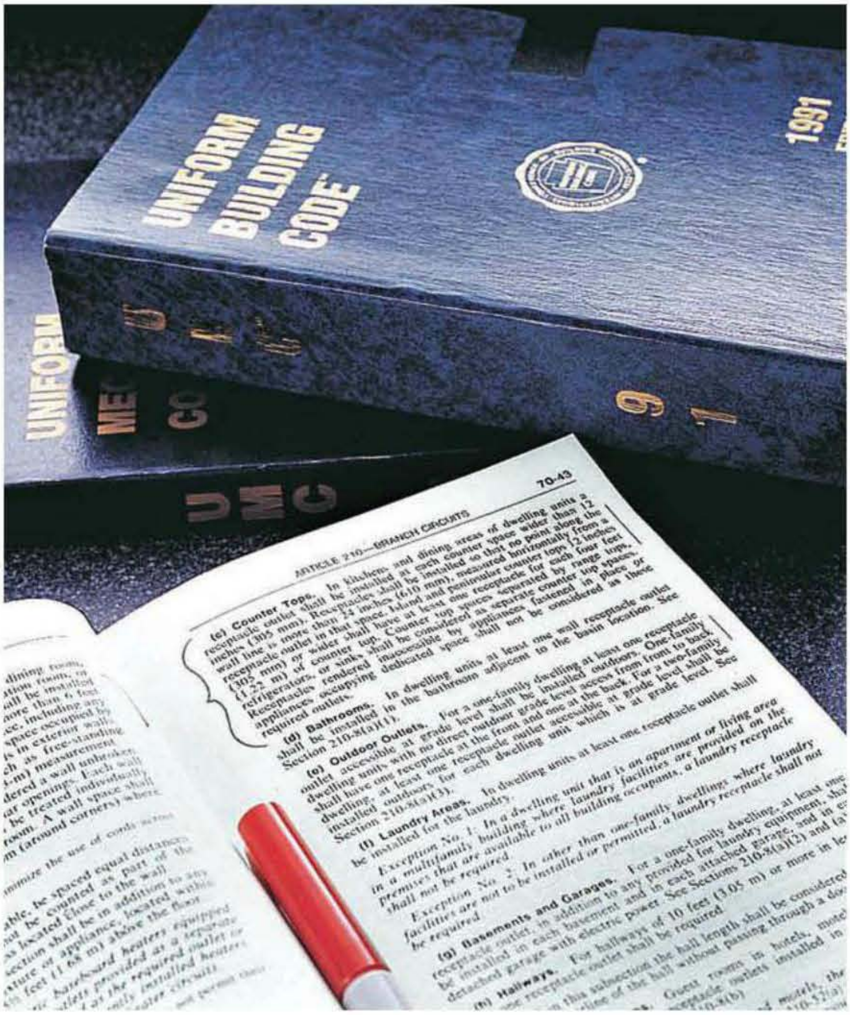

Code books can help you understand the structural, electrical and plumbing requirements for kitchens. In addition to the formal Code books, which are written for professional tradesmen, there are many Code handbooks available that are written for homeowners. Bookstores and libraries carry both the formal Code books and Code handbooks.

But if your new kitchen will be substantially larger than it is now, or if the ratio of wall space filled by glass windows and doors will be greater, it's possible that you'll need to expand its heating and cooling capacity.

Increasing your kitchen's heating and cooling can be as simple as extending ducts by a few feet, or as complicated as installing a new furnace. When installing a large room addition, for instance, you may learn that the present furnace is too small to adequately heat the increased floor space of your home.

How do you determine what your kitchen needs in the way of expanded heating and cooling? Unless you happen to be a mechanical engineer, you'll need to consult a professional to evaluate your heating/ventilation/air-conditioning (HVAC) system. The Code requirements for room heating are quite simple, but the methods used to calculate required energy needs of a room are fairly complex.

The Building Code requires simply that a room must be able to sustain a temperature of 70°F, measured at a point 3 ft. above the floor. HVAC contractors use a complicated formula to calculate the most efficient way to meet this Code requirement. You can make this estimation more accurate by providing the following information:

- The exact dimensions of your kitchen.
- The thickness and amount of insulation in the walls.
- The number of doors and windows, including their size and their energy ratings.
- The total square footage of your house.
- The heating and cooling capacity of your furnace and central air-conditioner, measured in BTUs.

This information, usually printed on the unit's access panel, will help the HVAC contractor determine if the system can adequately serve your new kitchen.

Finally, your cooktop should be equipped with an electric vent hood to exhaust cooking fumes and moisture from the kitchen. The volume of air moved by a vent fan is restricted by Code, so you should always check with a Building Inspector before selecting a vent hood.

Metal ductwork for the vent hood must be run through an exterior wall or through the ceiling. If your cooktop is located in an island cabinet, a special island vent fan is necessary.

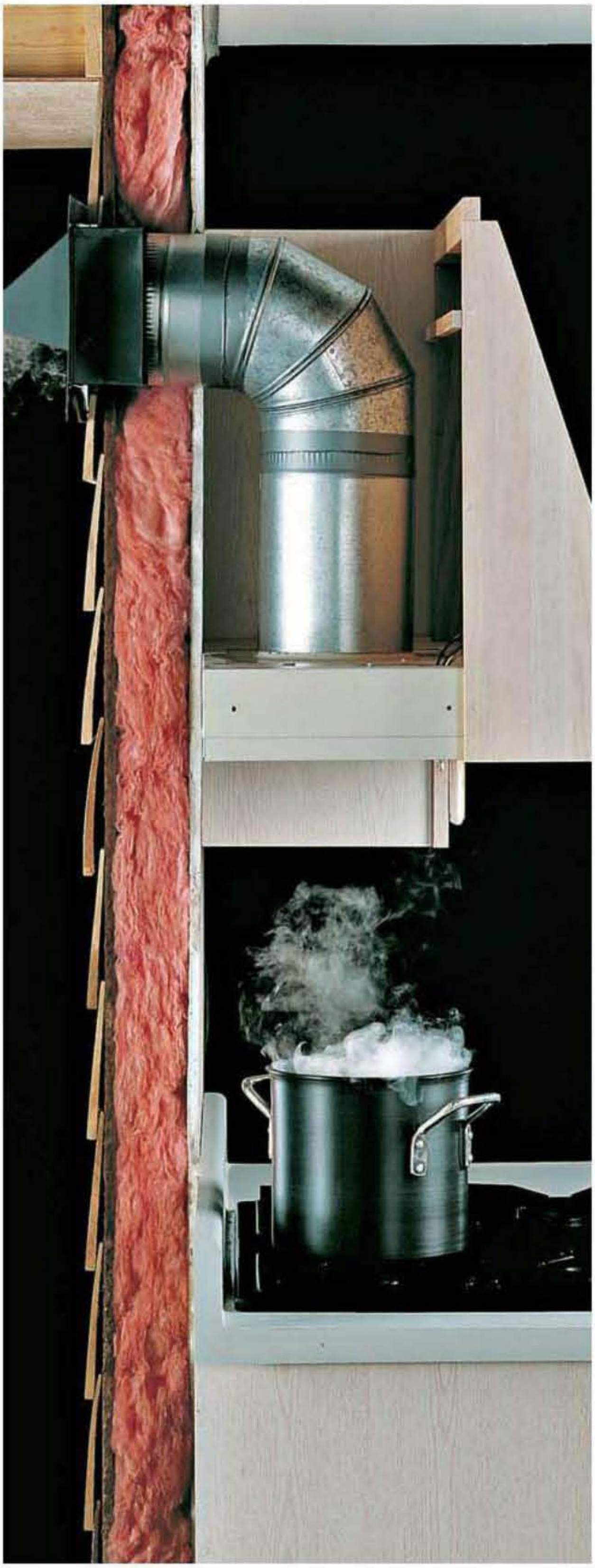

Vent hoods are required by some local codes on all ranges and cooktops. The vent fan exhausts cooking fumes and moisture to the outdoors.

Energy Efficiency

For the sake of both environmental resources and your own budget, selecting a kitchen design that incorporates energy- and water-saving appliances is a good idea. Most of the new appliances on the market are more energy efficient than older models, but the relative quality of their performance varies. Do some comparison-shopping when it comes to appliances and you are sure to outfit your kitchen with fixtures that save energy and money.

Lighting is another way to save energy. Consider the aesthetic benefits of adding a skylight or window to your kitchen; now think of the impact that natural lighting will have on your energy budget. Less artificial light will be needed, thus lowering electric bills.

Another way to cut down on energy requirements is to use fluorescent lighting or new innovations like LED lighting in the kitchen. These products not only reduce glare better than incandescent bulbs, they also use a fraction of the energy used by incandescent bulbs and need to be replaced less frequently. Newer versions of fluorescent bulbs produce a more true-to-life light color, rather than the old buzzing overhead tubes that gave everything a greenish hue. Some local building codes require a certain percentage of a kitchen's lighting to be fluorescent, so consider the benefits before lamenting this ordinance.

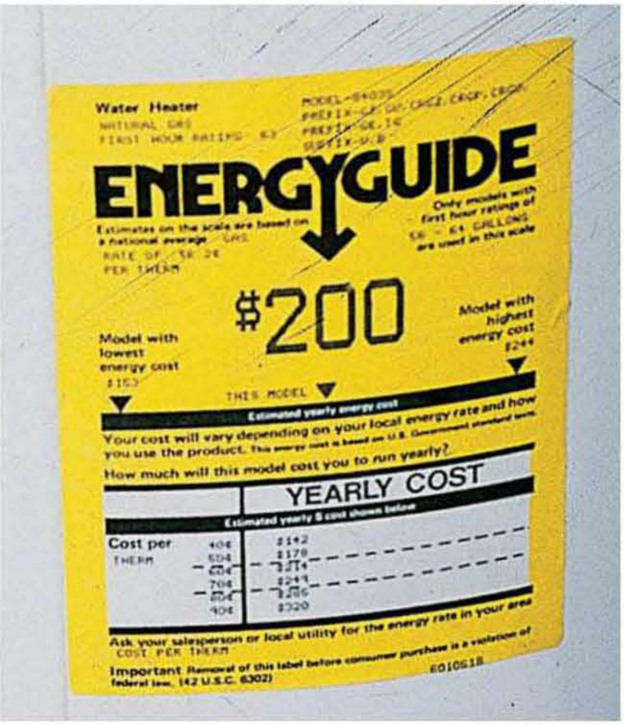

Energy usage labels can be found on most major appliances you'll encounter at the appliance store. More energy efficient models may qualify for a rebate from your local public utility company. Ask the sales staff for information on these programs.

As a general rule, newer appliances are more efficient when it comes to energy and water consumption. For example, this dishwasher has a feature that allows the user to wash only one of the racks, using less water and energy when smaller loads are being cleaned.

Accessible Kitchens ▸

Universal Design and Accessible Design are the terms used by architects, engineers, and product designers to describe making accommodations for the needs of every type of person, regardless of age and physical ability. For example, large handled tools, contrasting typeface, wide doorways, and smooth flooring are all results of Universal Design research. In the kitchen, the consideration of accessibility must be more extensive than almost anywhere else in the home. From safety lighting, to pull out shelves, to the height of counters, planning a kitchen that everyone can easily use requires careful forethought. There are tips on Universal Design throughout this book, offering advice on accessibility for many of the standard kitchen elements.

Universal design for kitchen cabinets puts the majority of items in the comfortable reach zone between 2 and 5 feet above the floor. Using pop-up and pull-down shelves can extend this area. Full-extension hardware on drawers and pull-out shelves eliminates reaching and fumbling for unseen items at the back of drawers and cabinets. Base cabinets installed at various heights serve users of different heights. Hanging some sections of upper cabinets at 12" to 15" above the countertop—rather than 18"—makes it easier to see and access items.

A full-length pantry also increases the amount of storage space in the comfortable reach zone. A pull-out pantry allows for easier viewing of the contents, but make sure the hardware is top quality and operates easily and smoothly. An appliance garage, a cabinet with a tambour door that sits on the countertop, is an excellent way to efficiently use corner space and store heavy or frequently used appliances out of sight. Make sure an electrical outlet or two are inside the garage.

UNIVERSAL TIPS

- Magnetic touch latches or c-shaped handles are the easiest to open.
- Avoid gloss cabinet finishes to reduce glare for the vision impaired.
- Countertops at a variety of heights should be ergonomically correct for a range of different users.

Countertop height and ease of fixture operation are two of the most important aspects of universal design. If your kitchen remodeling project needs to accommodate a person with special needs, most states have agencies that can offer you specific advice for your situation.

Universal Design

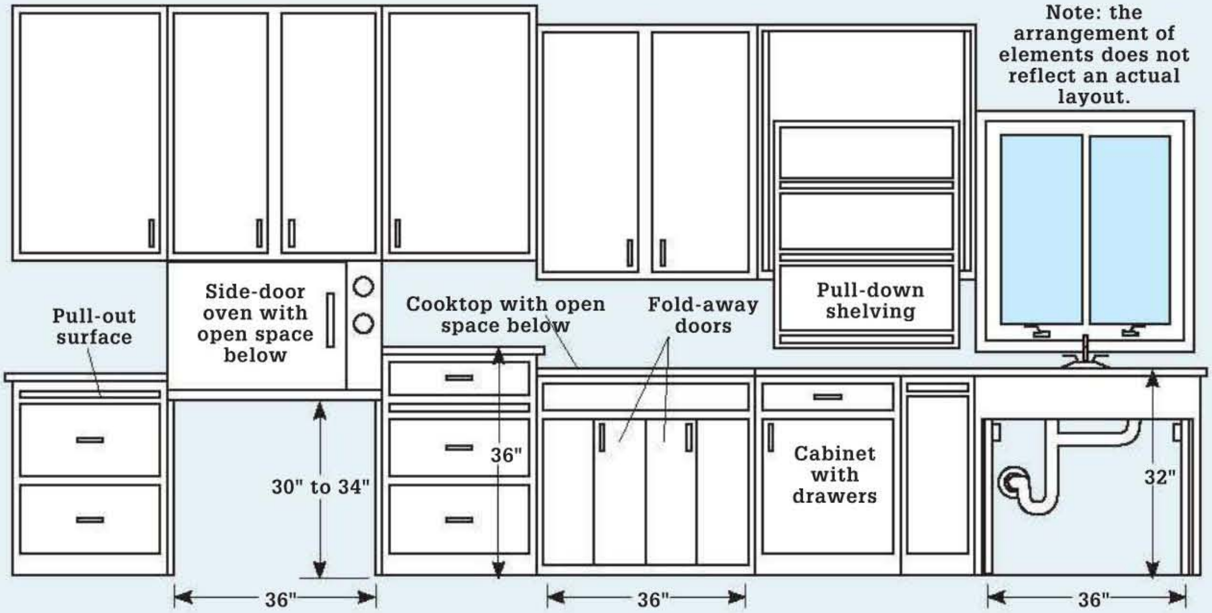
Note: the arrangement of elements does not reflect an actual layout.
Pull-out surface
Side-door oven with open space below
Cooktop with open space below
Fold-away doors
Pull-down shelving
30" to 34"
36"
Cabinet with drawers
32"
36"
36"
36"

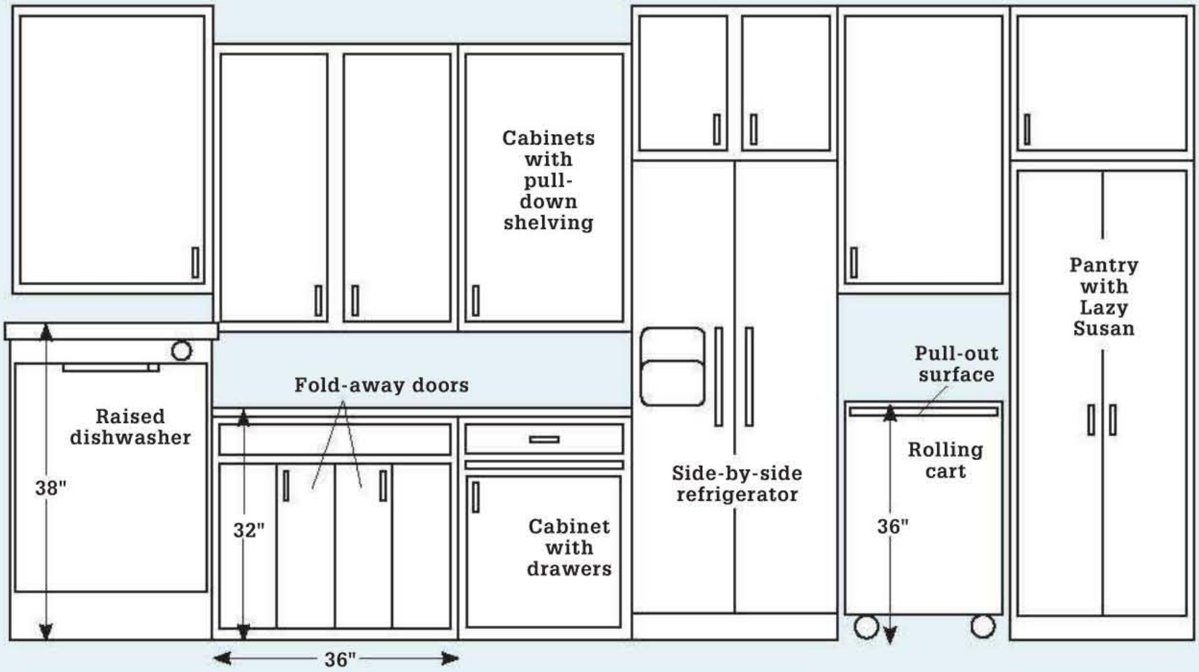
Cabinets with pull-down shelving
Pantry with Lazy Susan
Pull-out surface
Raised dishwasher
Fold-away doors
Rolling cart
Side-by-side refrigerator
38"
32"
Cabinet with drawers
36"
36"

Design your kitchen around a clear, circular space of at least 5 ft. in diameter to provide room for a wheelchair. If your kitchen doesn't have 60" of clear space, allow 48" for pathways. Plan for 30 to 48" of clear approach space in front of all appliances and workstations.

Installing Kitchen Cabinets

Cabinets must be firmly anchored to wall studs, and they must be plumb and level when installed. The best way to ensure this is by attaching a ledger board to the wall to assist in the installation. As a general rule, install the upper cabinets first so your access is not impeded by the base cabinets. (Although some pros prefer to install the base cabinets first so they can be used to support the uppers during installation.) It's also best to begin in a corner and work outward from there.

Stock cabinets are sold in boxes that are keyed to door and drawer packs (you need to buy these separately). It is important that you realize this when you are estimating your project costs at the building center (often a door pack will cost as much or more than the cabinet). Allow plenty of time for assembling the cabinets out of the box. It can take an hour or more to put some more complex cabinets together.

Tools & Materials ▸

- Handscrew clamps
- Level
- Hammer
- Utility knife
- Nail set
- Stepladder
- Drill
- Counterbore drill bit
- Cordless screwdriver
- Jigsaw
- Cabinets
- Trim molding
- Toe-kick molding
- Filler strips
- Valance
- 6d finish nails
- Finish washers
- #10 × 4" wood screws
- #8 × 2½" screws
- 3" drywall screws

How to Fit a Corner Cabinet ▸

Before installation, test-fit corner and adjoining cabinets to make sure doors and handles do not interfere with each other. If necessary, increase the clearance by pulling the corner cabinet away from the side wall by no more than 4". To maintain even spacing between the edges of the doors and the cabinet corner, cut a filler strip and attach it to the corner cabinet or the adjoining cabinet. Filler strips should be made from material that matches the cabinet doors and face frames.

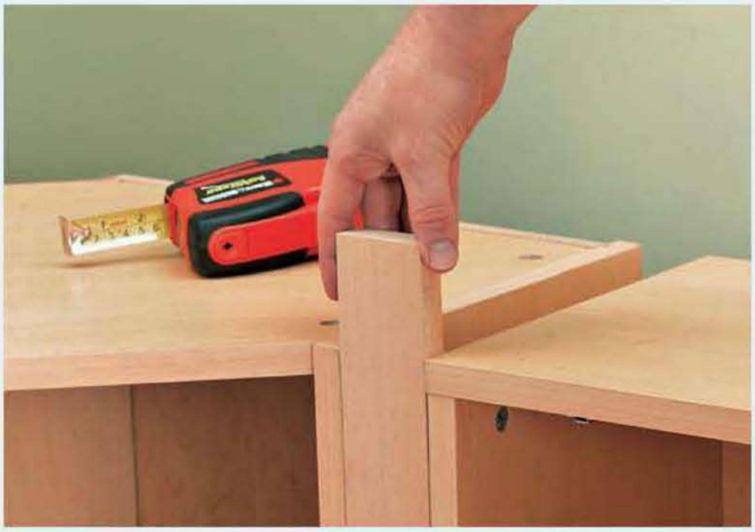

How to Prepare Walls

1

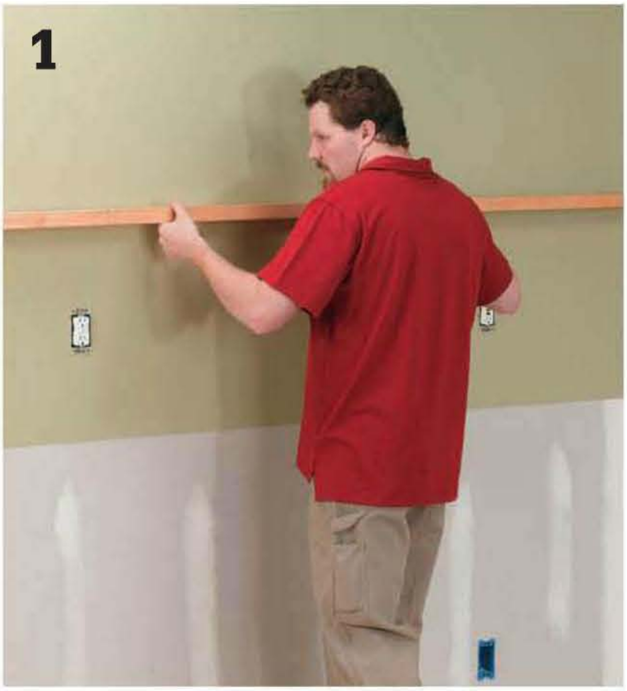

Find high and low spots on wall surfaces, using a long, straight 2 × 4. Sand down any high spots.

2

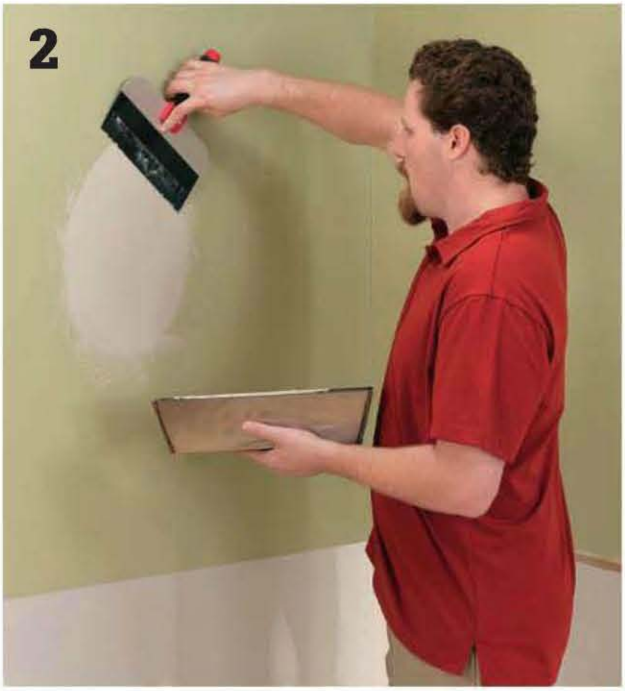

Fill in low spots of wall by applying wallboard compound using a taping knife. Let the compound dry, and then sand it lightly.

3

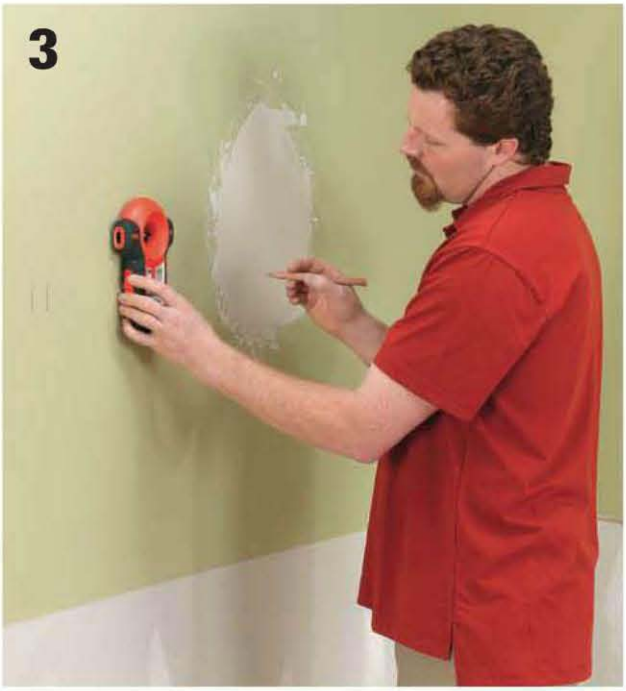

Locate and mark wall studs, using an electronic stud finder. Cabinets normally will be hung by driving screws into the studs through the back of the cabinets.

4

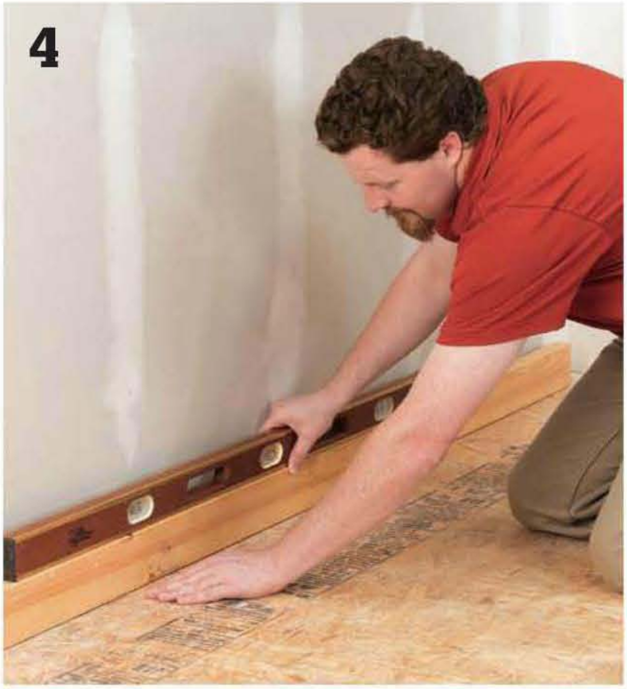

Find the highest point along the floor that will be covered by base cabinets. Place a level on a long, straight 2 × 4, and move the board across the floor to determine if the floor is uneven. Mark the wall at the high point.

(continued)

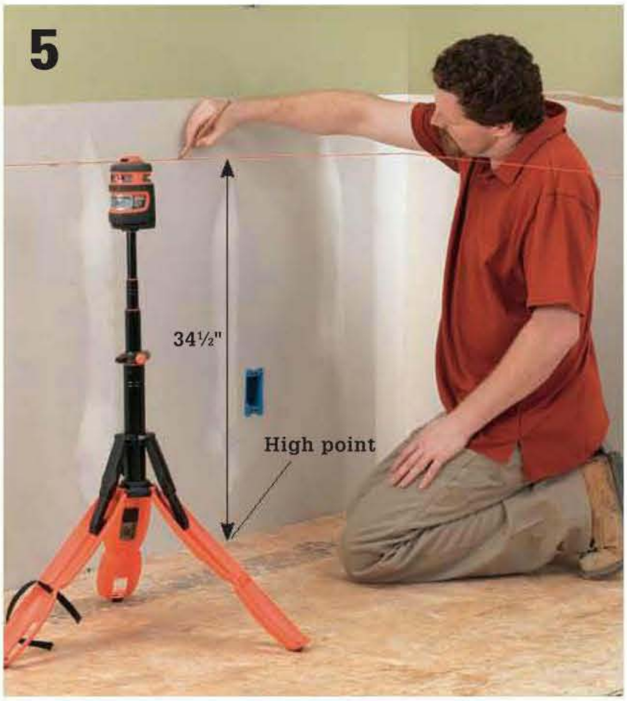

Measure up 34½" from the high-point mark (for standard cabinets). Use a level (a laser level is perfect) to mark a reference line on walls. Base cabinets will be installed with top edges flush against this line.

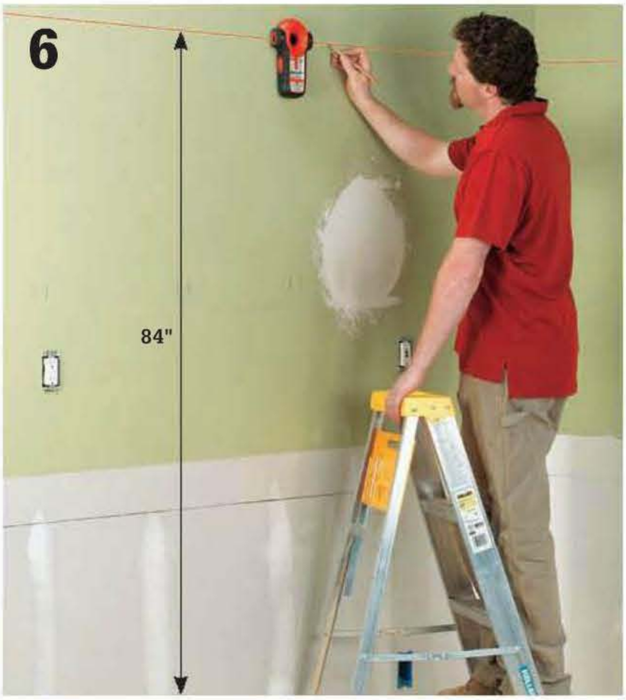

Measure up 84" from the high-point mark and draw a second reference line. Wall cabinets will be installed with their top edges flush against this line.

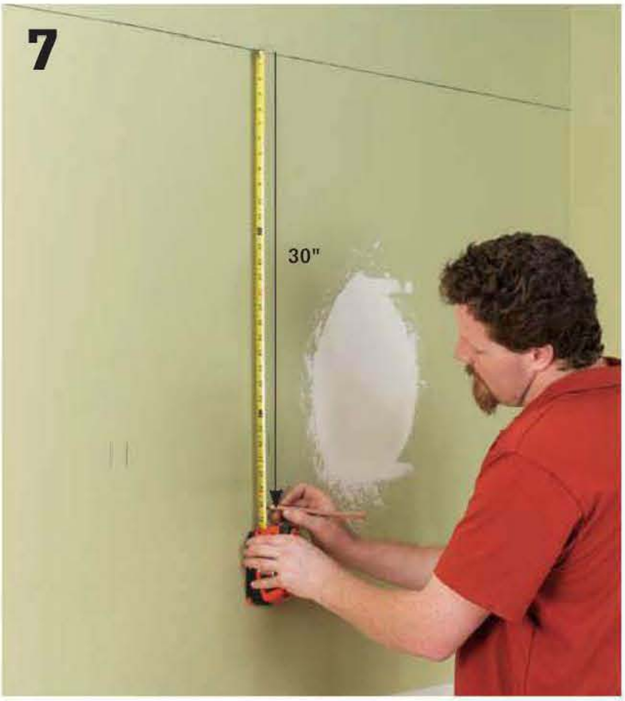

Measure down 30" from the wall-cabinet reference line and draw another level line where the bottoms of the cabinets will be. Temporary ledgers will be installed against this line.

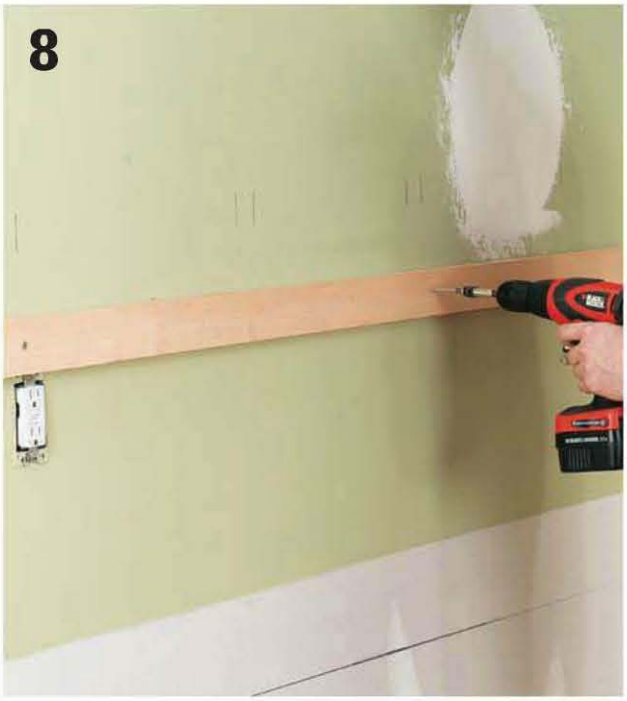

Install 1 × 3 temporary ledgers with top edges flush against the reference lines. Attach ledgers with 2½" wallboard screws driven into every other wall stud. Mark stud locations on ledgers. Cabinets will rest temporarily on ledgers during installation (the ledgers alone will not support them, however).

How to Hang Wall Cabinets

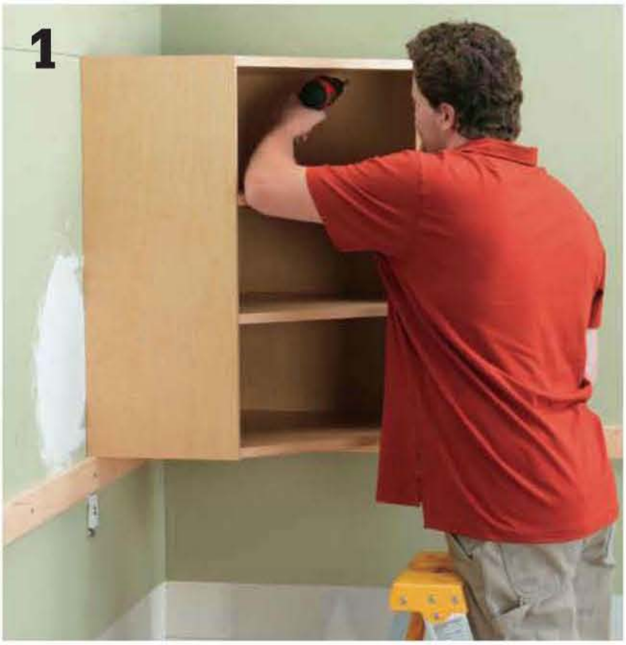

Position a corner upper cabinet on a ledger (see page 352) and hold it in place, making sure it is resting cleanly on the ledger. Drill 3⁄16" pilot holes into the wall studs through the hanging strips at the top, rear of cabinet. Attach the cabinet to the wall with 2½" screws. Do not tighten fully until all cabinets are hung.

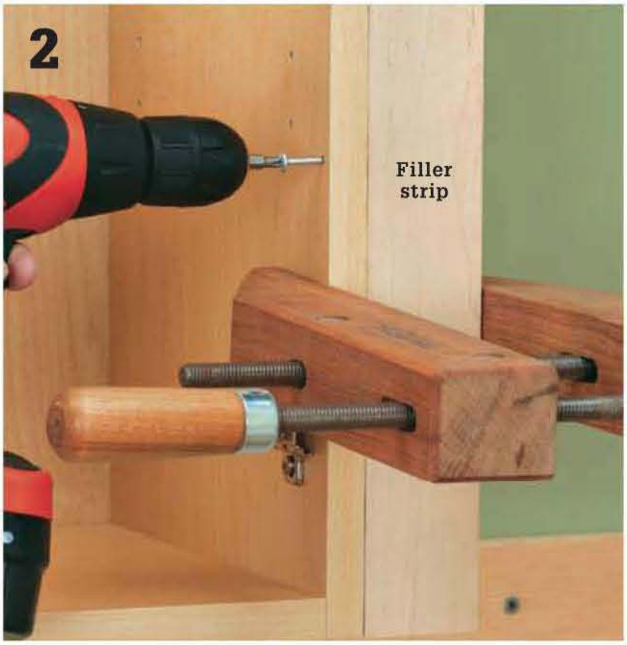

Attach a filler strip to the front edge of the cabinet, if needed (see page 350). Clamp the filler in place, and drill counterbored pilot holes through the cabinet face frame, near hinge locations. Attach filler to cabinet with 2½" cabinet screws or flathead wood screws.

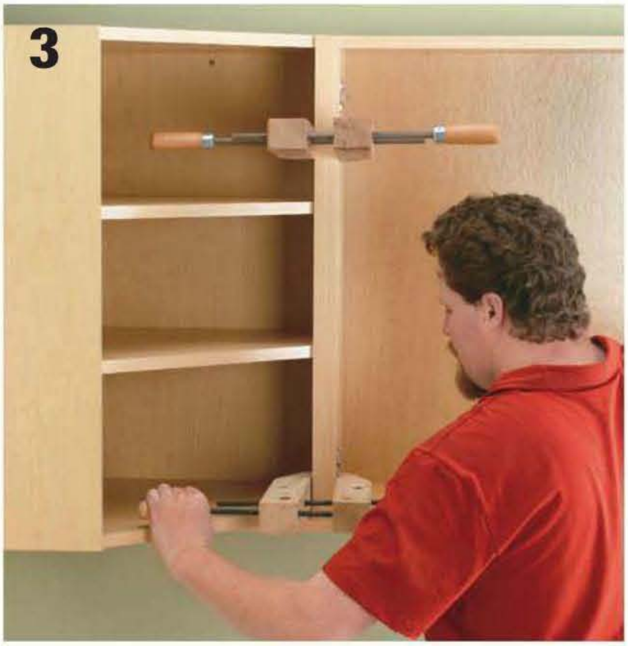

Position the adjoining cabinet on the ledger, tight against the corner cabinet or filler strip. Clamp the corner cabinet and the adjoining cabinet together at the top and bottom. Handscrew clamps will not damage wood face frames.

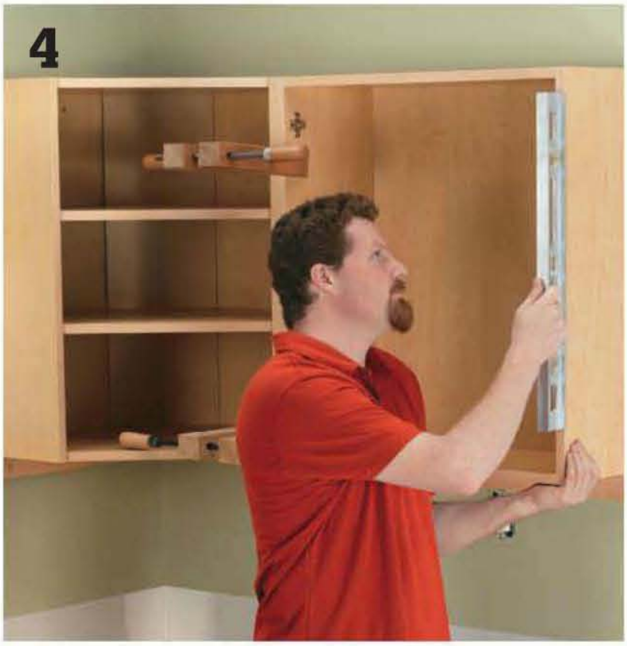

Check the front cabinet edges or face frames for plumb. Drill 3⁄16" pilot holes into wall studs through hanging strips in rear of cabinet. Attach cabinet with 2½" screws. Do not tighten wall screws fully until all cabinets are hung.

(continued)

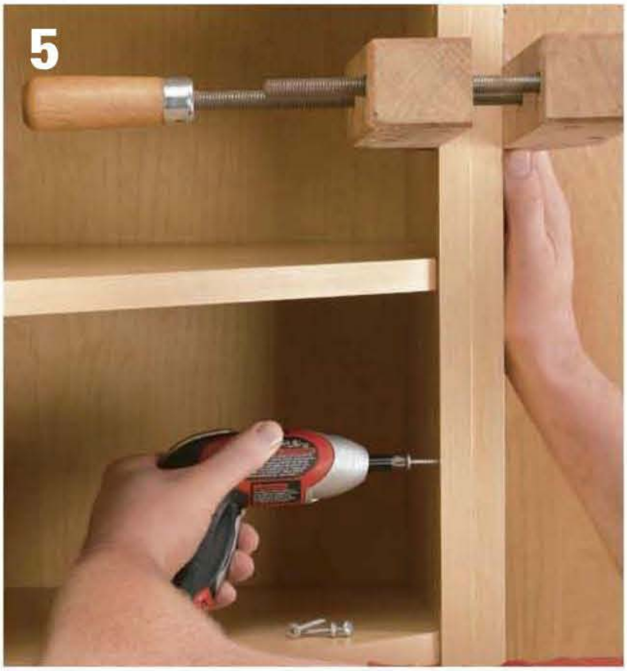

Attach the corner cabinet to the adjoining cabinet. From inside corner cabinet, drill pilot holes through face frame. Join cabinets with sheet-metal screws.

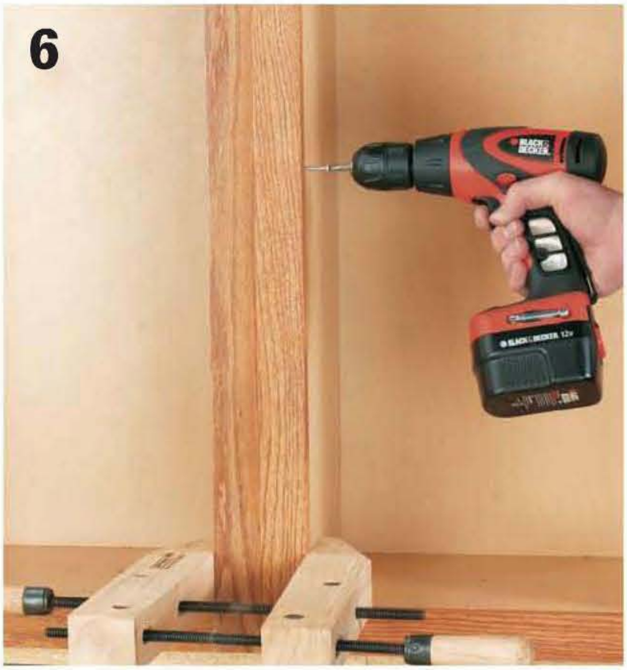

Position and attach each additional cabinet. Clamp frames together, and drill counterbored pilot holes through side of face frame. Join cabinets with wood screws. Drill 3/16" pilot holes in hanging strips, and attach cabinet to studs with wood screws.

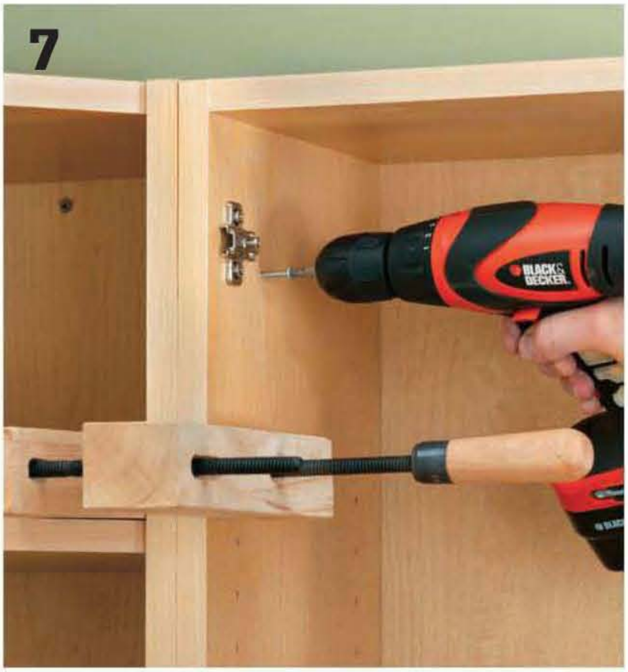

Join frameless cabinets with #8 × 1¼" panhead wood screws or wood screws with decorative washers. Each pair of cabinets should be joined by at least four screws.

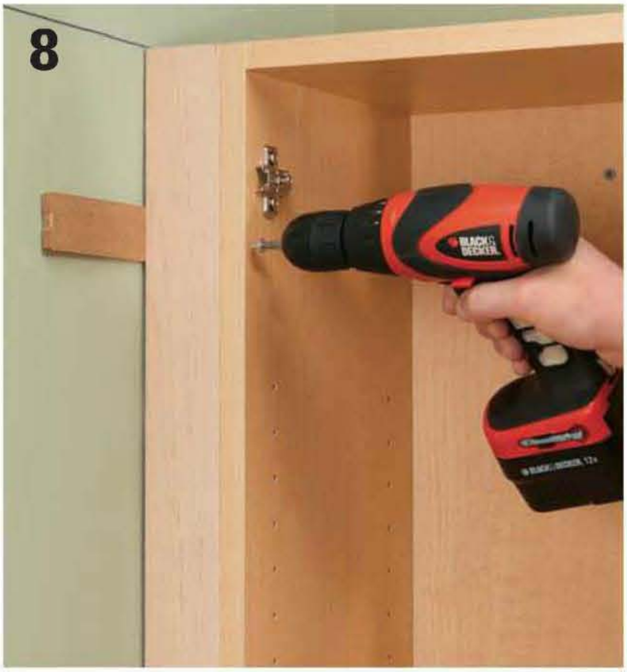

Fill gaps between the cabinet and wall or neighboring appliance with a filler strip. Cut the filler strip to fit the space, then wedge wood shims between the filler and the wall to create a friction fit that holds it in place temporarily. Drill counterbored pilot holes through the side of the cabinet (or the edge of the face frame) and attach filler with screws.

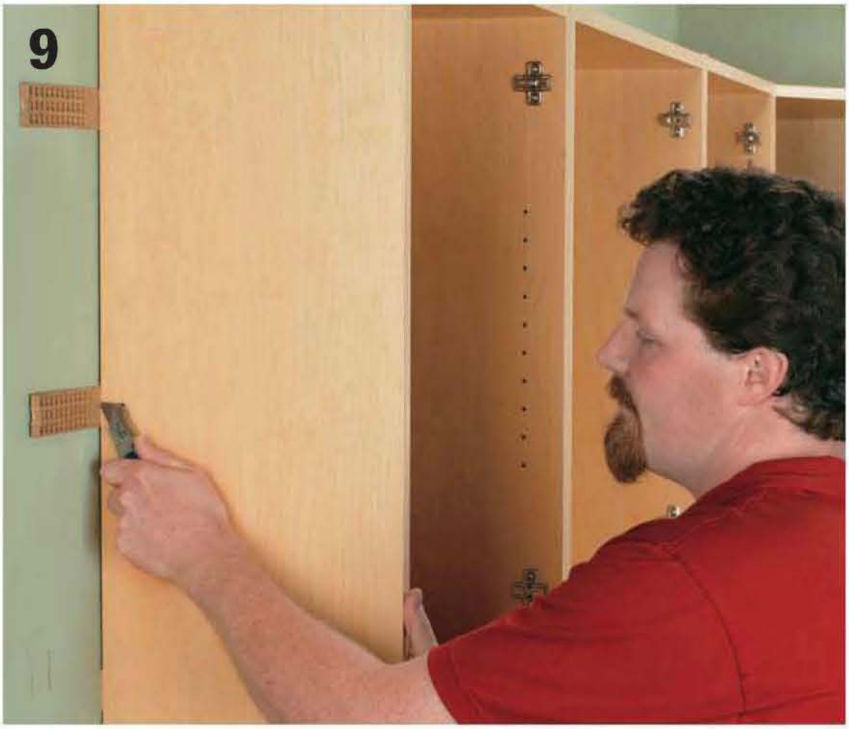

Remove the temporary ledger. Check the cabinet run for plumb, and adjust if necessary by placing wood shims behind cabinet, near stud locations. Tighten wall screws completely. Cut off shims with utility knife.

Use trim moldings to cover any gaps between cabinets and walls. Stain moldings to match cabinet finish.

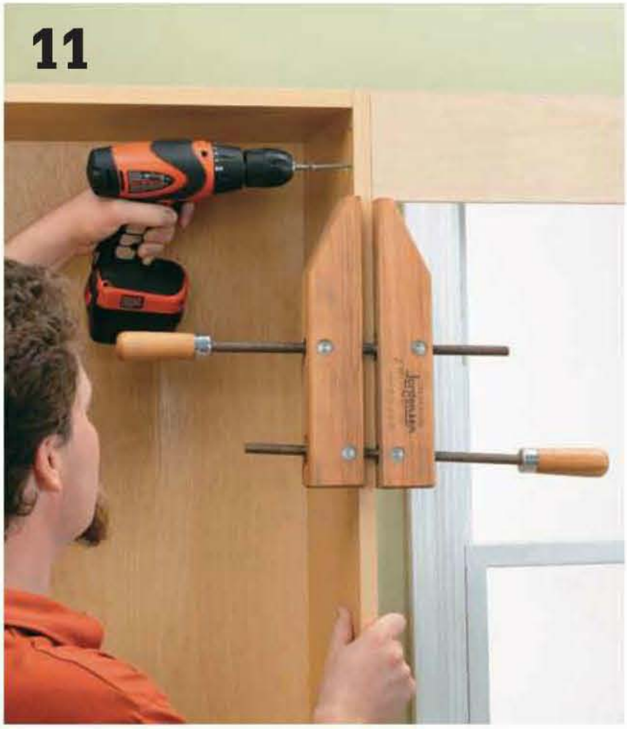

Attach decorative valance above sink. Clamp valance to edge of cabinet frames, and drill counterbored pilot holes through cabinet frames into end of valance. Attach with sheet-metal screws.

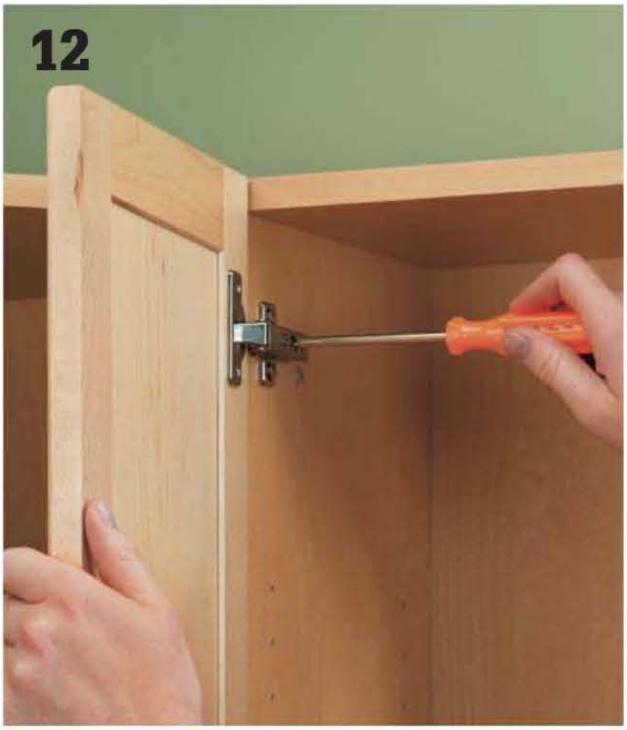

Install the cabinet doors. If necessary, adjust the hinges so that the doors are straight and plumb.

How to Install Base Cabinets

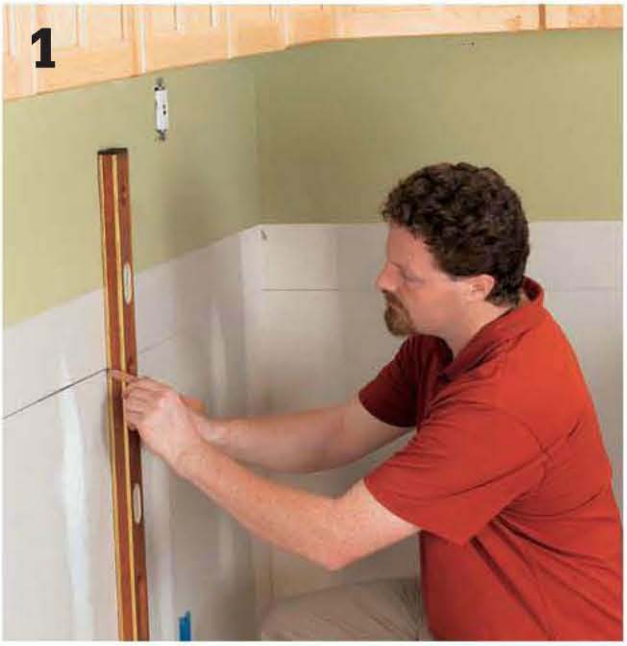

Begin the installation with a corner cabinet. Draw plumb lines that intersect the 34½" reference line (measured from the high point of the floor—see page 351) at the locations for the cabinet sides.

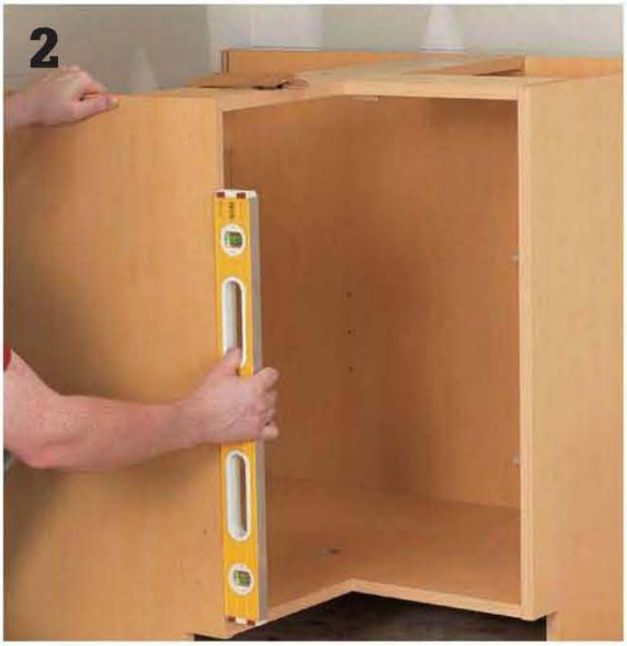

Place cabinet in corner. Make sure the cabinet is plumb and level. If necessary, adjust by driving wood shims under cabinet base. Be careful not to damage flooring. Drill 3⁄16" pilot holes through the hanging strip and into wall studs. Tack the cabinet to the wall with wood screws or wallboard screws.

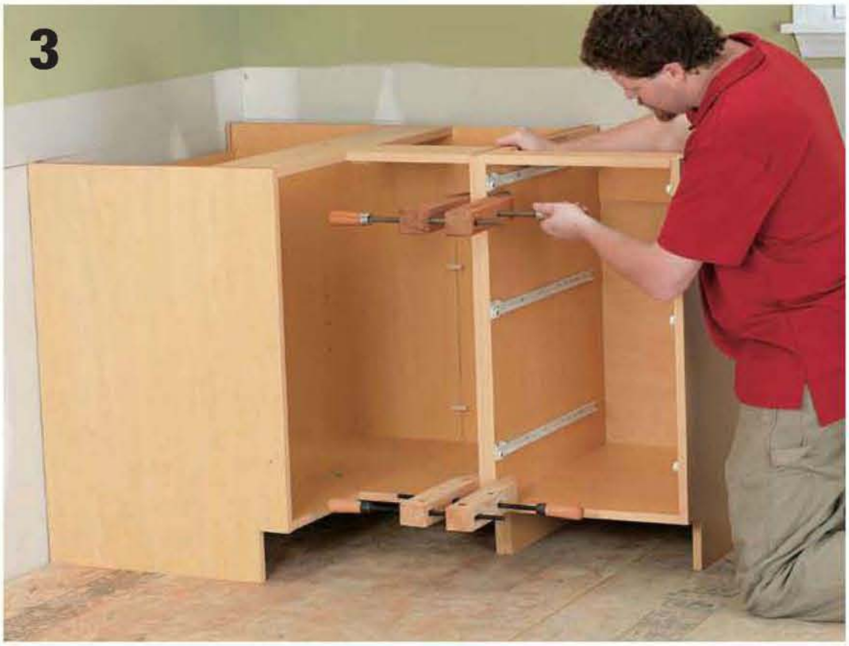

Clamp the adjoining cabinet to the corner cabinet. Make sure the new cabinet is plumb, then drill counterbored pilot holes through the cabinet sides or the face frame and filler strip. Screw the cabinets together. Drill 3⁄16" pilot holes through hanging strips and into wall studs. Tack the cabinets loosely to the wall studs with wood screws or wallboard screws.

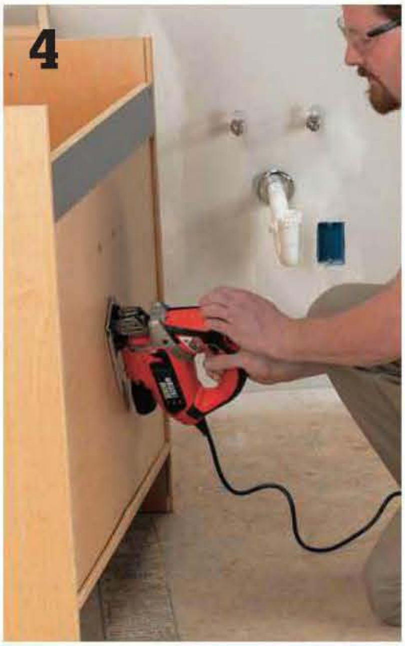

Use a jigsaw to cut any cabinet openings needed in the cabinet backs (for example, in the sink base seen here) for plumbing, wiring or heating ducts.

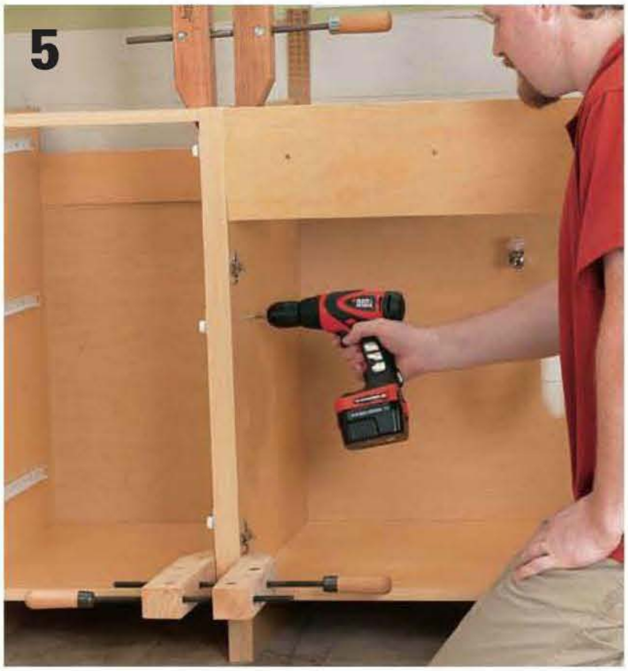
5

Position and attach additional cabinets, making sure the frames are aligned and the cabinet tops are level. Clamp cabinets together, then attach the face frames or cabinet sides with screws driven into pilot holes. Tack the cabinets to the wall studs, but don't drive screws too tight—you may need to make adjustments once the entire bank is installed.

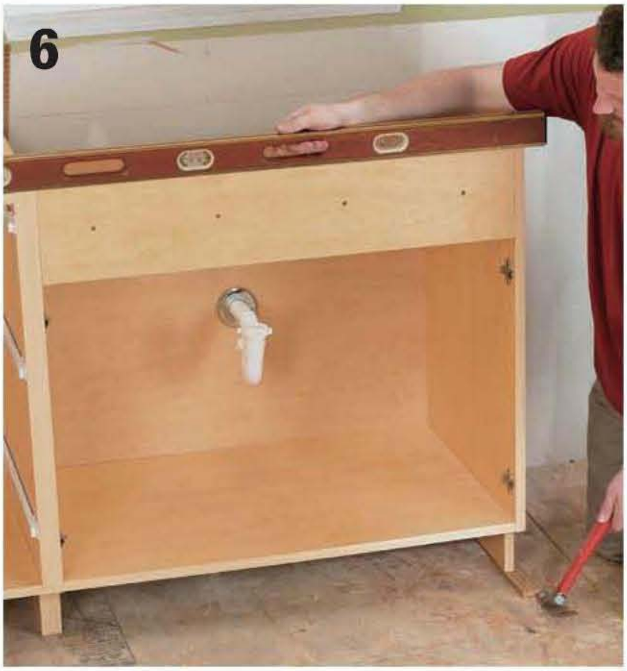
6

Make sure all cabinets are level. If necessary, adjust by driving shims underneath cabinets. Place shims behind the cabinets near stud locations to fill any gaps. Tighten wall screws. Cut off shims with utility knife.

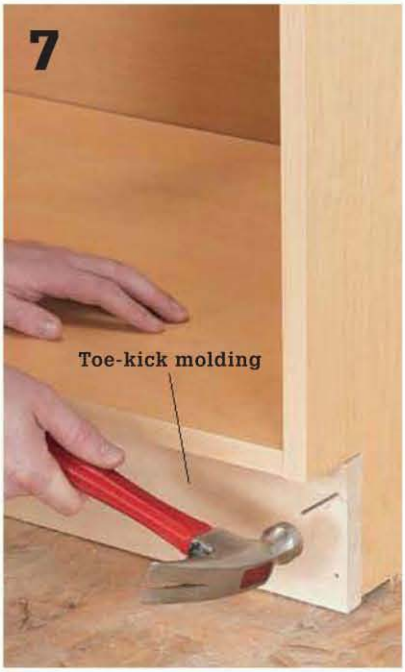
7

Use trim moldings to cover gaps between the cabinets and the wall or floor. The toe-kick area is often covered with a strip of wood finished to match the cabinets or painted black.

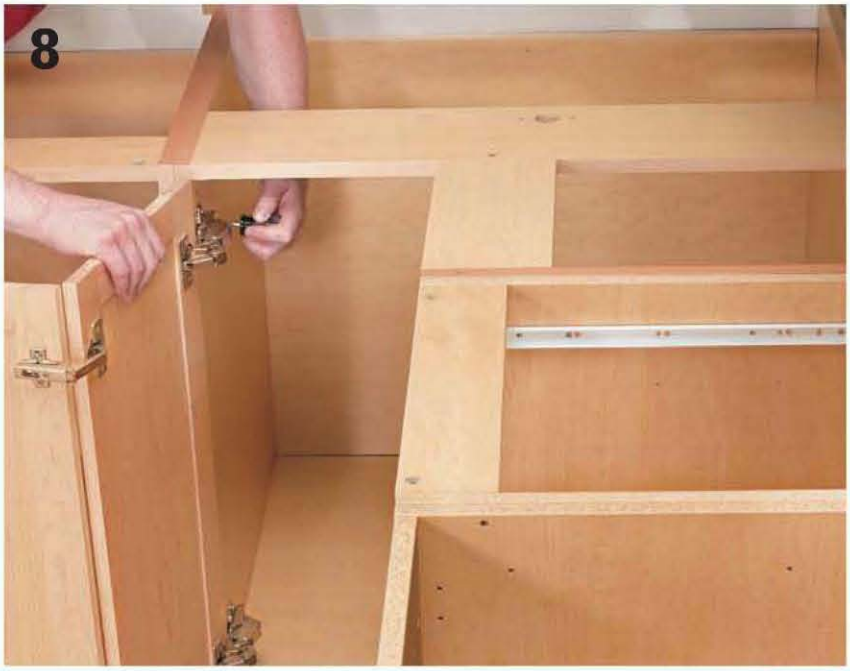
8

Hang cabinet doors and mount drawer fronts. Test to make sure the drawers close smoothly and the doors fit evenly and flush. Self-closing cabinet hinges (by far the most common type installed today) have adjustment screws that allow you to make minor changes to the hardware to correct any problems.

Building Kitchen Islands

Kitchen islands can be created using a whole range of methods, from repurposing an old table to fine, custom woodworking. But perhaps the easiest (and most failsafe) way to add the conveniences and conviviality of a kitchen island is to make one from stock base cabinets. The cabinets and countertops don't have to match your kitchen cabinetry, but that is certainly an option you should consider. When designing and positioning your new island, be sure to maintain a minimum distance of 3 ft. between the island and other cabinets (4 ft. or more is better).

Tools & Materials ▸

- Marker
- Drill/driver
- 2 × 4 cleats
- Pneumatic nailer and 2" finish nails or hammer and 6d finish nails
- 2 base cabinets (approx. 36" wide × 24" deep)
- Countertop
- Wallboard screws

Two base cabinets arranged back-to-back make a sturdy kitchen island base that's easy to install. When made with the same style cabinets and countertops as the rest of the kitchen, the island is a perfect match.

How to Create a Stock-cabinet Island

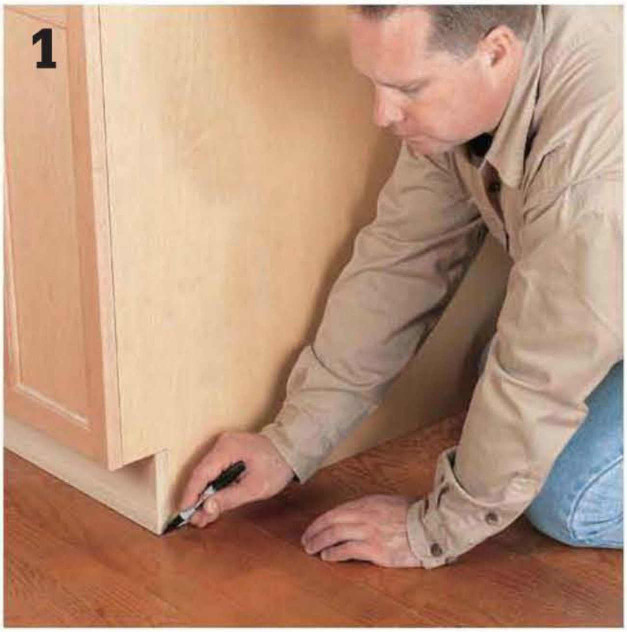

Set two base cabinets back-to-back in position on the floor and outline the cabinet corners onto the flooring. Remove the cabinets and draw a new outline inside the one you just created to allow for the thickness of the cabinet sides (usually ¾").

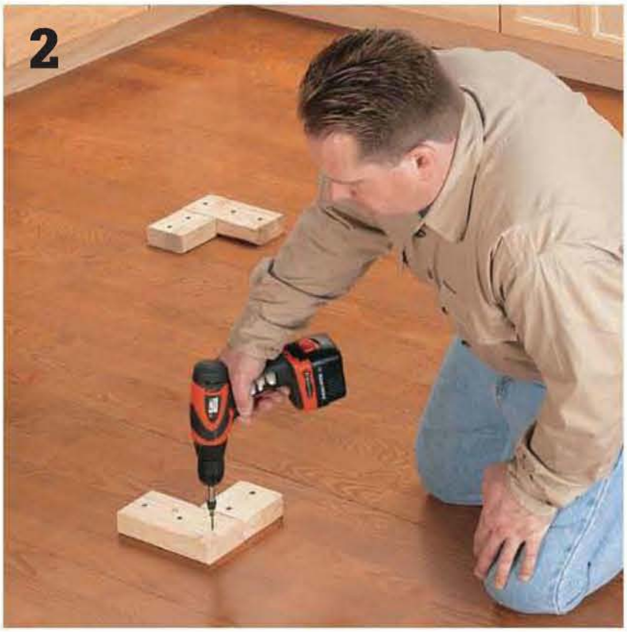

Cut 2 × 4 cleats to fit inside the inner outline to provide nailing surfaces for the cabinets. Attach the cleats to the floor with screws or nails. *Tip: Create an L-shape cleat for each inside corner.*

Join the two base cabinets by driving 1¼" wallboard screws through the nailing strips on the backs of the cabinets from each direction. Make sure the cabinet sides are flush and aligned. Lower the base cabinets over the cleats. Check the cabinets for level, and shim underneath the edges of the base if necessary.

Attach the cabinets to the floor cleats using 6d finish nails. Drill pilot holes for nails, and recess nail heads with a nail set. Install a countertop on top of the cabinets (see pages 362 to 385).

Building Tiled Islands

Islands are one of the most requested kitchen features. People love them for many reasons, including their value as bi-level counter space. In most cases, the lower level is used as work space and the upper as casual dining space. The upper level provides a little camouflage for the work space, something that's especially welcome in open-plan kitchens where meal preparation areas can be seen from other areas.

When planning casual dining space, remember that designers suggest at least 24" per person. For the work space, remember that standard design guidelines recommend at least 36" of uninterrupted work space to the side of a sink or cooktop.

On work surfaces, mosaic and other small tile is rarely the best choice. Larger tile requires fewer grout lines, always a good idea when it comes to cleaning and maintenance. But there is no rule that all three elements of a bi-level island have to use the same material. In fact, projects like this offer wonderful opportunities to mix materials or colors or textures. Choose floor tile or tile made especially for counters for the horizontal surfaces, and then branch out when it comes to the backsplash. Wall tile and mosaics work beautifully.

Tools & Materials ▸

- Tape measure
- Circular saw
- Drill
- Utility knife
- Straightedge
- Stapler
- Drywall knife
- Framing square
- Notched trowel
- Tile cutter
- Carpeted 2 × 4
- Mallet
- Rubber grout float
- Sponge
- Foam brush
- Caulk gun
- Birch plywood
- 1 × 2 hardwood
- 2 × 4 lumber
- Ceramic tile
- Tile spacers
- ¾" exterior-grade (CDX) plywood
- 4-mil polyethylene sheeting
- Packing tape
- ½" cementboard
- 1¼" deck screws
- 3" deck screws
- Fiberglass mesh tape
- Thinset mortar
- Grout with latex additive
- Silicone caulk
- Silicone grout sealer
- L-brackets
- 6d finish nails
- Drywall screws
- Glue

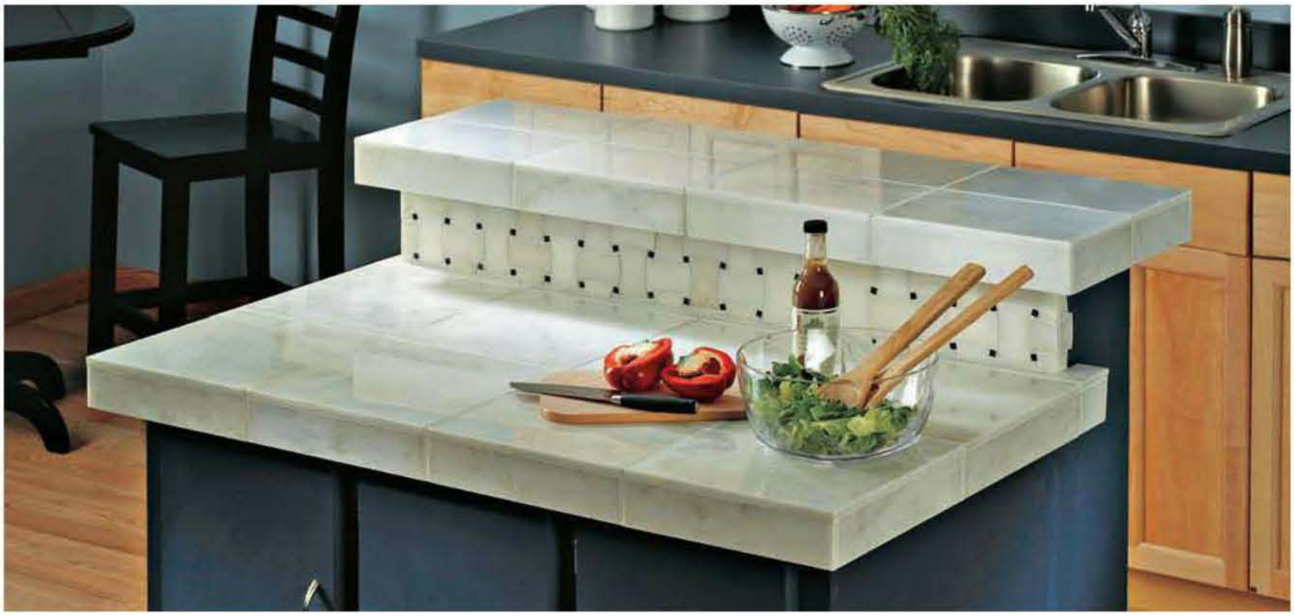

Attractive backsplashes complete tile contertops. Here, individual tiles seamlessly stack side by side to create one cohesive, interwoven pattern. Using the same material and tones as the surrounding countertop allows the pattern to add interest without cluttering the small island, resulting in a sophisticated finished design.

How to Build a Tiled Bi-level Island

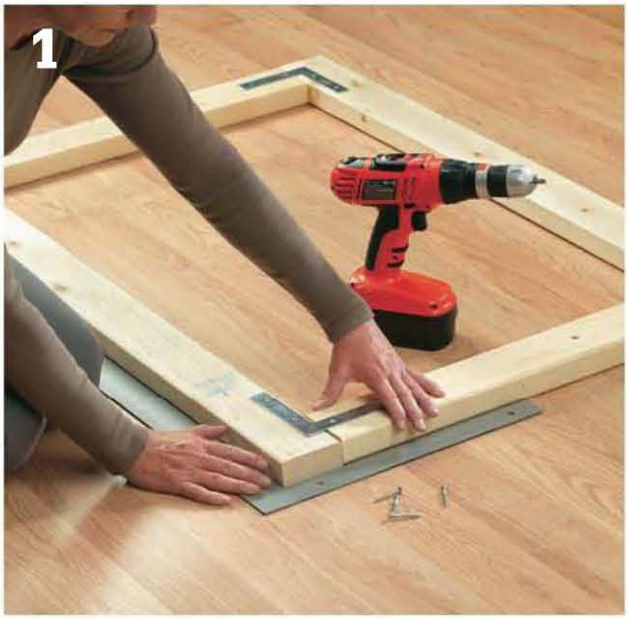
1

Build a 2 × 4 base for the island cabinet by cutting the 2 × 4s to length and joining them in a square frame that lays flat (wide sides down) on the floor. Use metal L-brackets to reinforce the joints. If you don't wish to move the island, fasten the frame to the floor in position with construction adhesive and/or deck screws.

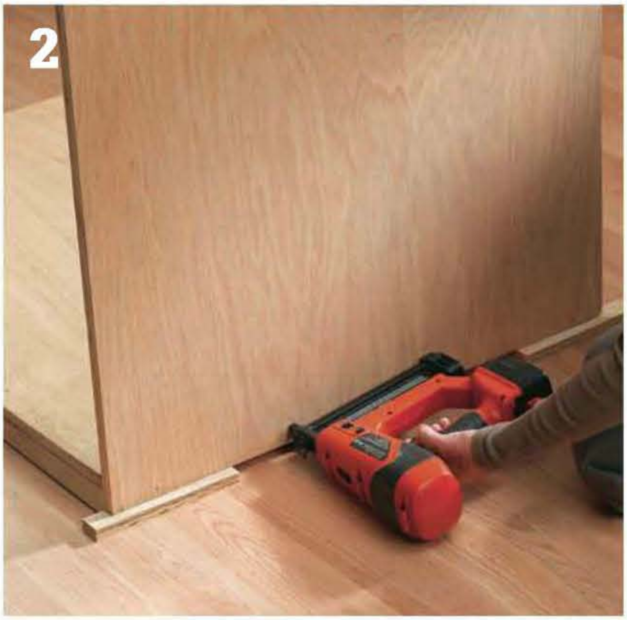
2

Cut bottom panels the same dimensions as the base frame from ¾" birch plywood. Attach it to the frame with finish nails. Then, cut the side panels to size and shape and fasten them to the edges of the curb with 6d finish nails and adhesive. Slip ¾" shims (scrap plywood works well) beneath the side panels before fastening them.

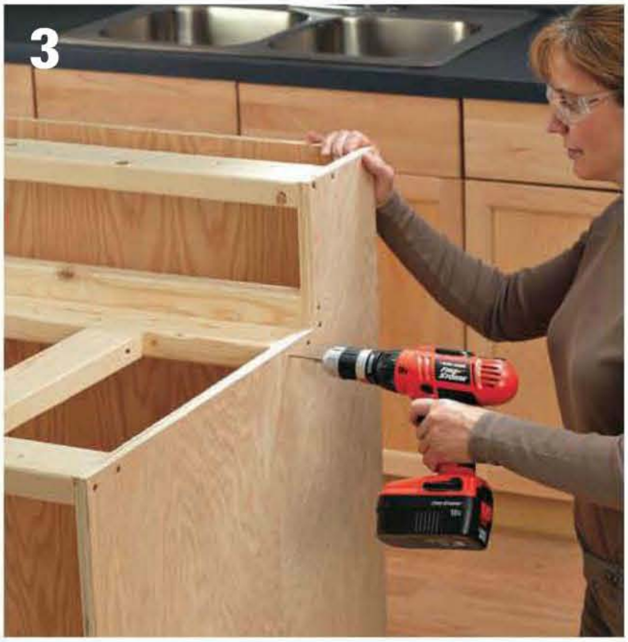
3

Cut the 2 × 4 cross supports to length and install them between the side panels at every corner, including the corners created by the L-shape cutout. Use 3" deck screws driven through the side panels and into the ends of the cross supports.

4

Lightly sand the cabinets and then clean off the dust. Prime and paint the cabinet interior and exterior.

(continued)

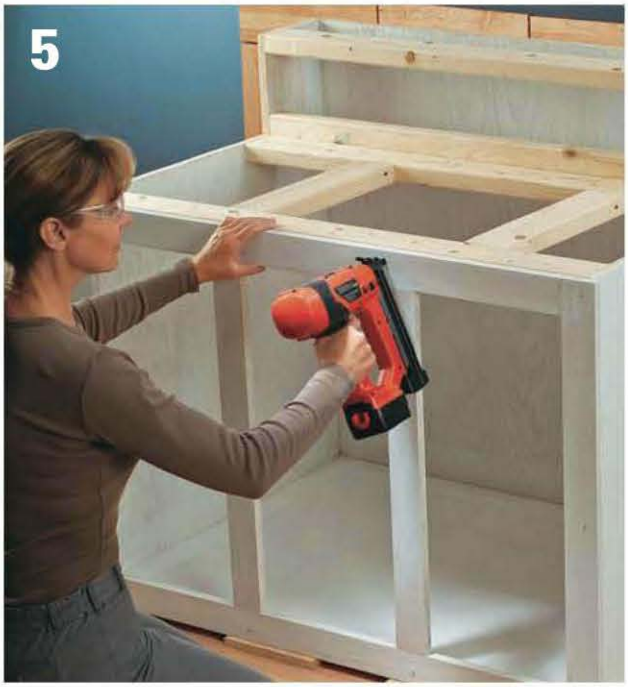

Build a face frame from 1 × 2 hardwood to fit the cabinet front. Attach it to the cabinet with 6d finish nails and hang cabinet doors (we installed three 13"-wide overlay doors).

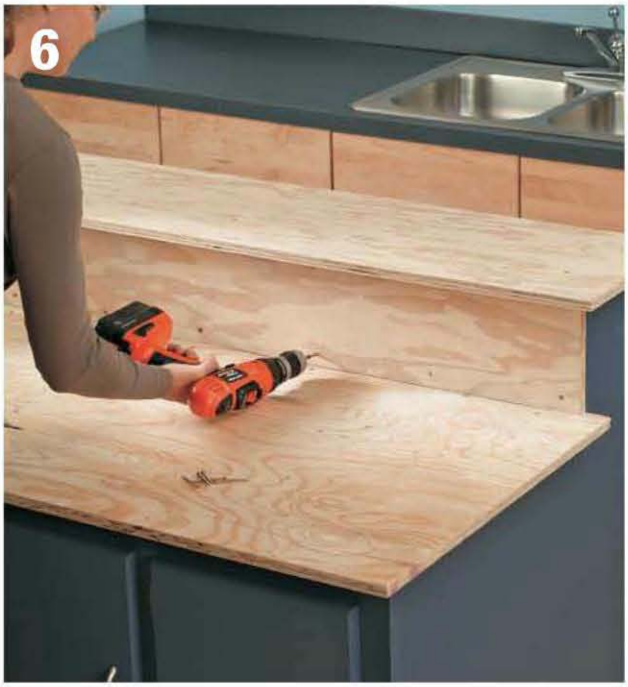

Cut strips of ¾" exterior plywood to make the subbases for the countertops and a backer for the backsplash. The lower counter subbase should over hang by 2" on the front and sides. The upper should overhang 2" on the sides and be centered on the cabinet front to back. Attach the backer and subbases with drywall screws driven down into the 2 × 4 cross supports.

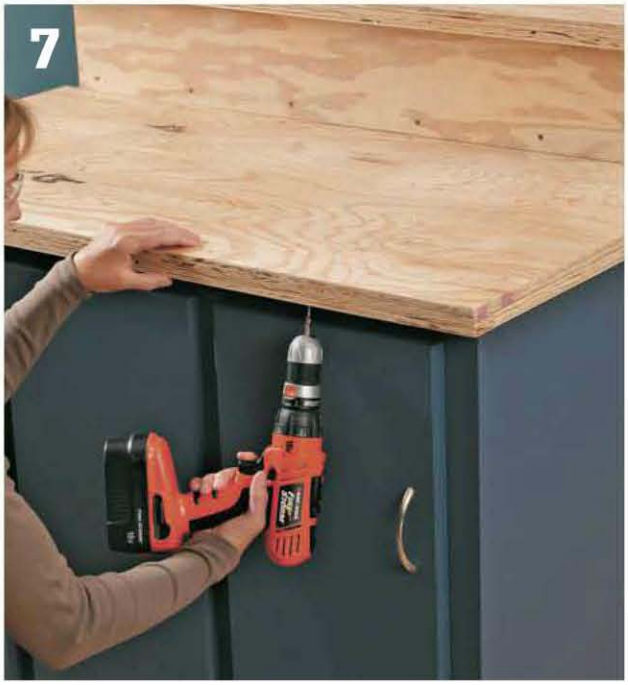

Cut 2" wide strips of plywood for buildup strips and attach to the undersides of the subbases with glue and screws.

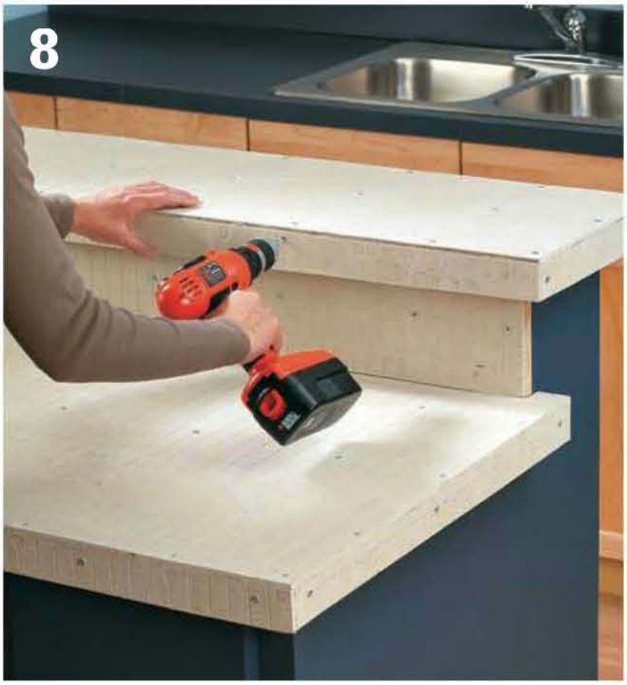

Attach tile backerboard to the counter subbases and the backsplash and tape seams and cover screws heads with compound (see page 380).

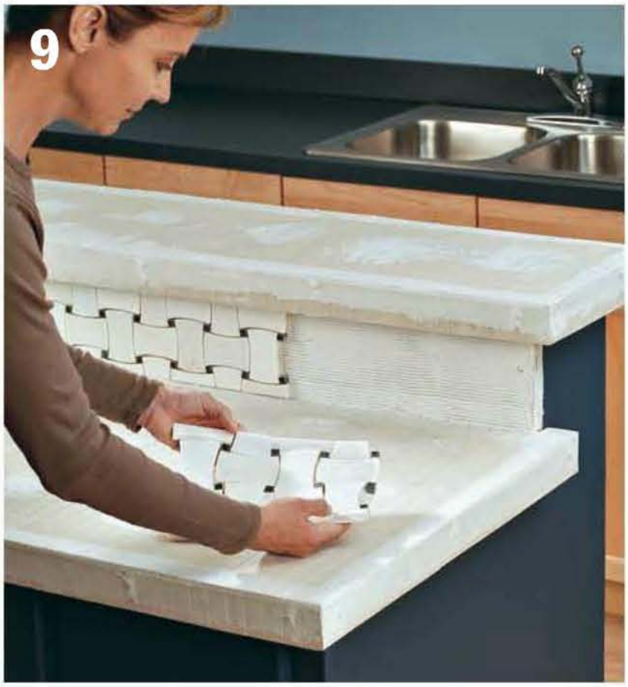

Cut mosaic sheets to fit the backsplash area and attach them with thinset adhesive (see Tiling a Backsplash, page 387).

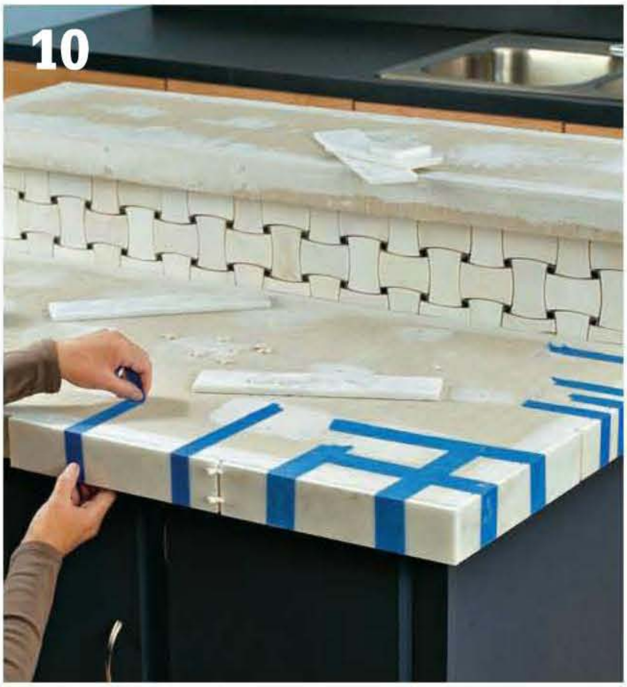

Cut edge tiles and fasten them around the perimeter of the subbase with thinset adhesive. Hold tiles in place with tape until adhesive sets up. The tiles should be flush or slightly below the bottoms of the buildup strips, and project past the top surfaces so they will be level with the field tiles.

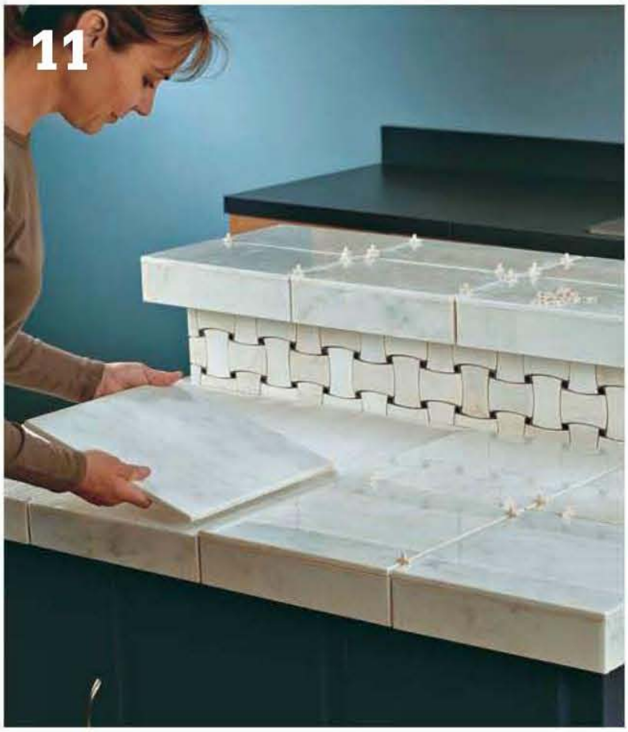

Install the field tiles for the countertops last (see pages 376 to 385).

Choose a suitable grout color and apply it to the tile with a grout float (page 385). Buff off excess once it has dried. Seal the grout with grout sealer.

Installing Postform Countertops

Postform laminate countertops are available in stock and custom colors. Pre-mitered sections are available for two- or three-piece countertops that continue around corners. If the countertop has an exposed end, you will need an endcap kit that contains a preshaped strip of matching laminate. Postform countertops have either a waterfall edge or a no-drip edge. Stock colors are typically available in 4-, 6-, 8-, 10- and 12-foot straight lengths and 6- and 8-foot mitered lengths.

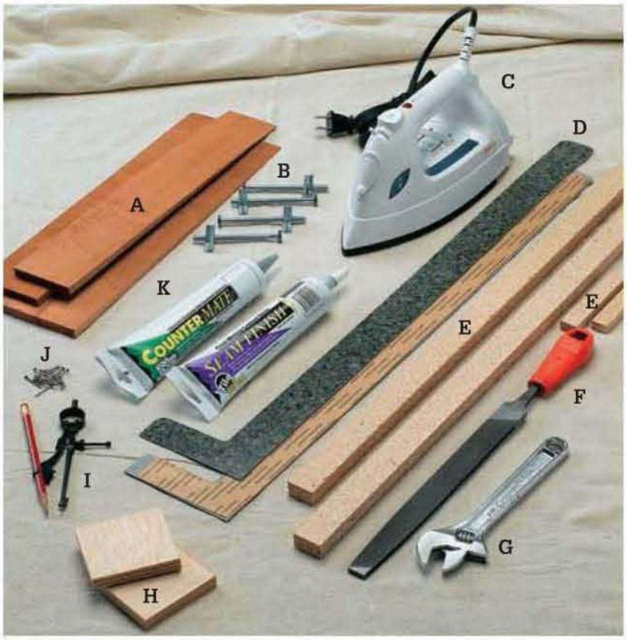

Materials and tools for installing a post-form countertop include: Wood for shimming (A), take-up bolts for drawing miters together (B), household iron (C), endcap laminate to match countertop (D), endcap battens (E), file (F), adjustable wrench (G), buildup blocks (H), compass (I), fasteners (J), silicone caulk and sealer (K).

Post-form countertops are among the easiest and cheapest to install. They are a good choice for beginning DIYers, but the design and color options are fairly limited.

How to Install a Postform Countertop

Tools & Materials ▸

- Tape measure
- Framing square
- Pencil
- Straightedge
- C-clamps
- Hammer
- Level
- Caulking gun
- Jigsaw
- Compass
- Adjustable wrench
- Belt sander
- Drill and spade bit
- Cordless screwdriver
- Postform countertop
- Wood shims
- Take-up bolts
- Drywall screws
- Wire brads
- Endcap laminate
- Silicone caulk
- Wood glue

BLACK & DECKER

Option: If the saw foot must rest on the good surface of the postform, use a jigsaw fitted with a downstroke blade to cut the postform. If you are unable to locate a downstroke blade, you can try applying tape over the cutting lines, but you are still likely to get tear-out from a normal upstroke jigsaw blade.

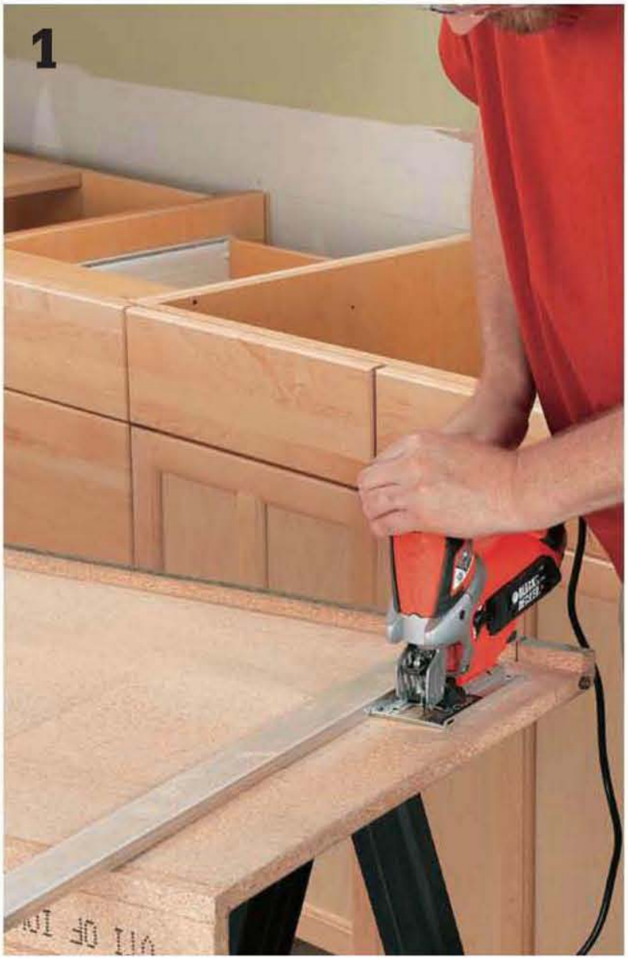
1

Use a framing square to mark a cutting line on the bottom surface of the countertop. Cut off the countertop with a jigsaw, using a clamped straight-edge as a guide.

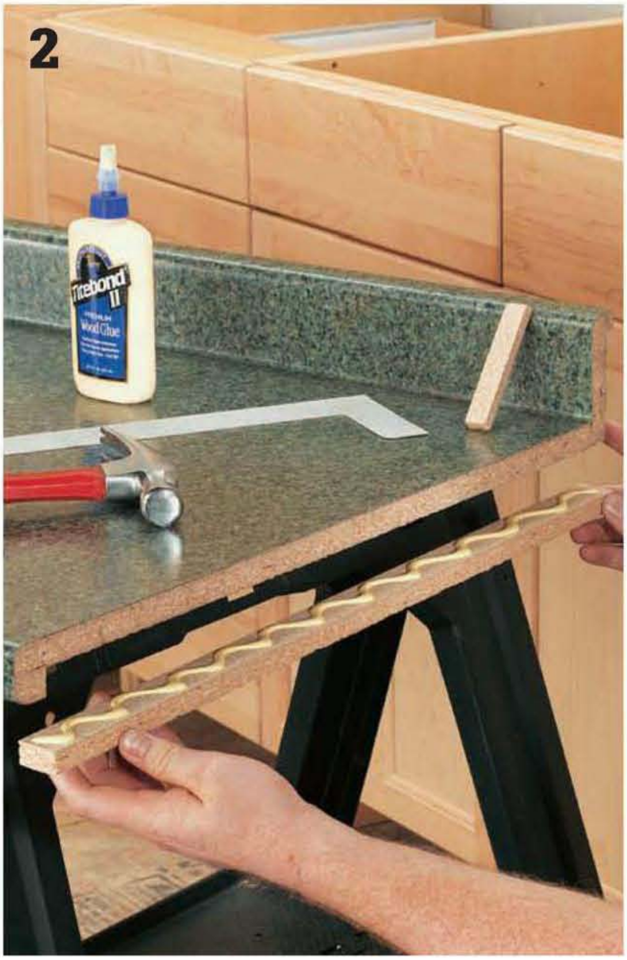
2

Attach the battens from the endcap kit to the edge of the countertop, using carpenter's glue and small brads. Sand out any unevenness with a belt sander.

(continued)

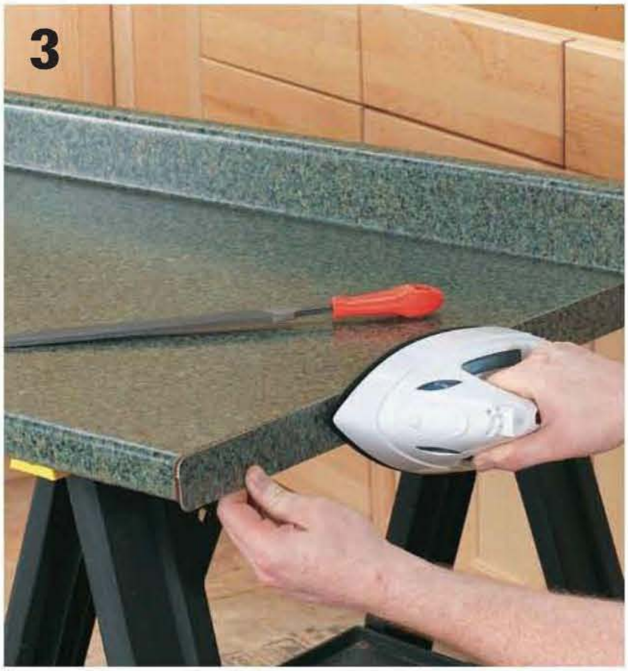

Hold the endcap laminate against the end, slightly overlapping the edges. Activate adhesive by pressing an iron set at medium heat against the endcap. Cool with a wet cloth, then file the endcap laminate flush with the edges of the countertop.

Position the countertop on base cabinets. Make sure the front edge of the countertop is parallel to the cabinet faces. Check the countertop for level. Make sure that drawers and doors open and close freely. If needed, adjust the countertop with shims.

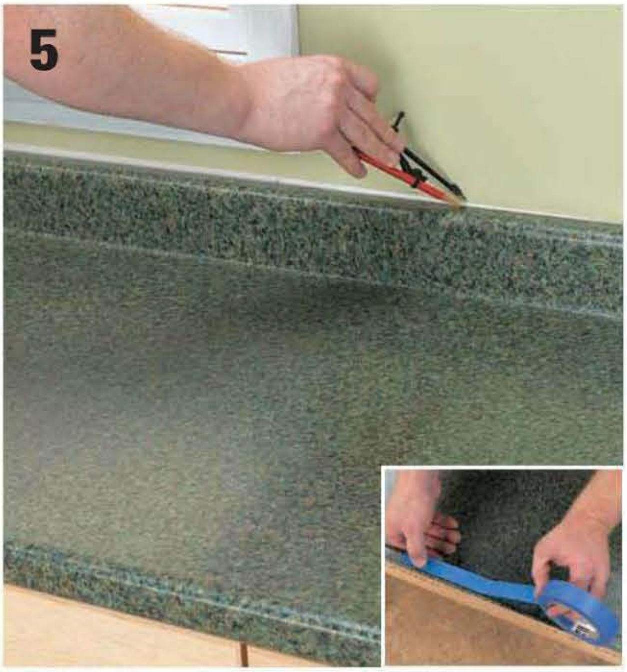

Because walls are usually uneven, use a compass to trace the wall outline onto the backsplash. Set the compass arms to match the widest gap, then move the compass along the length of the wall to transfer the outline to the top of the backsplash. Apply painter's tape to the top edge of the backsplash, following the scribe line (inset).

Remove the countertop. Use a belt sander to grind the backsplash to the scribe line.

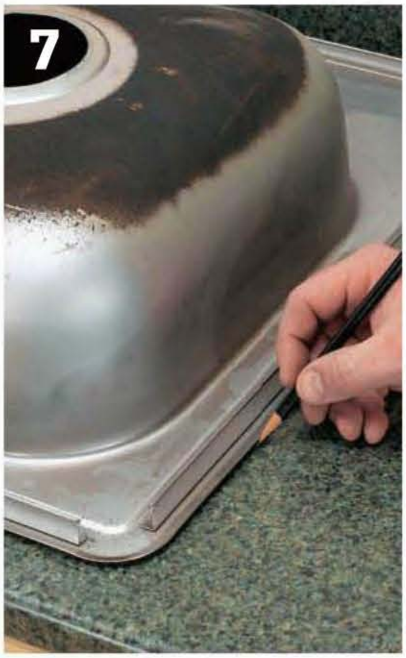

7

Mark cutout for self-rimming sink. Position the sink upside down on the countertop and trace its outline. Remove the sink and draw a cutting line ⅝" inside the sink outline.

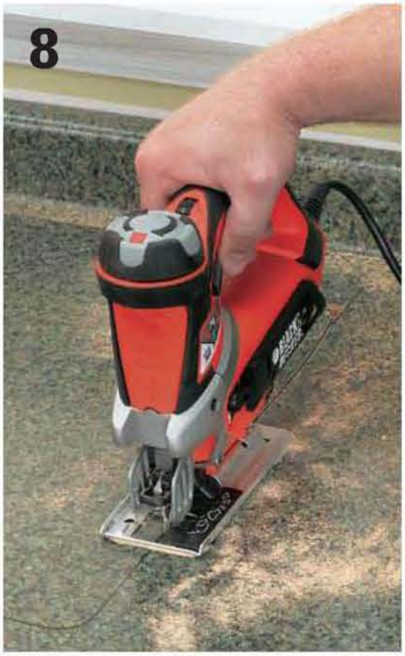

8

Drill a starter hole just inside the cutting line. Make sink cutouts with a jigsaw. Support the cutout area from below so that the falling cutout does not damage the cabinet.

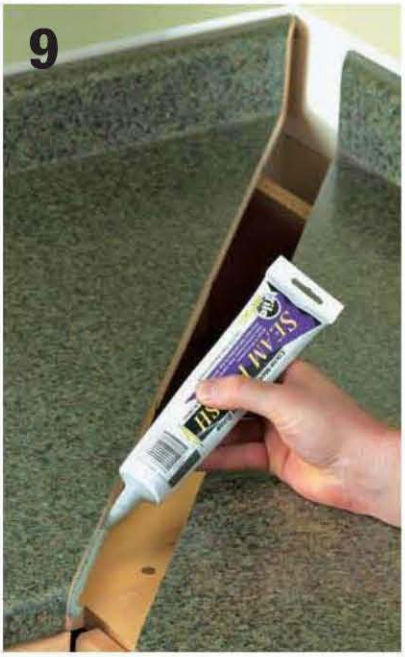

9

Apply a bead of silicone caulk to the edges of the mitered countertop sections. Force the countertop pieces tightly together.

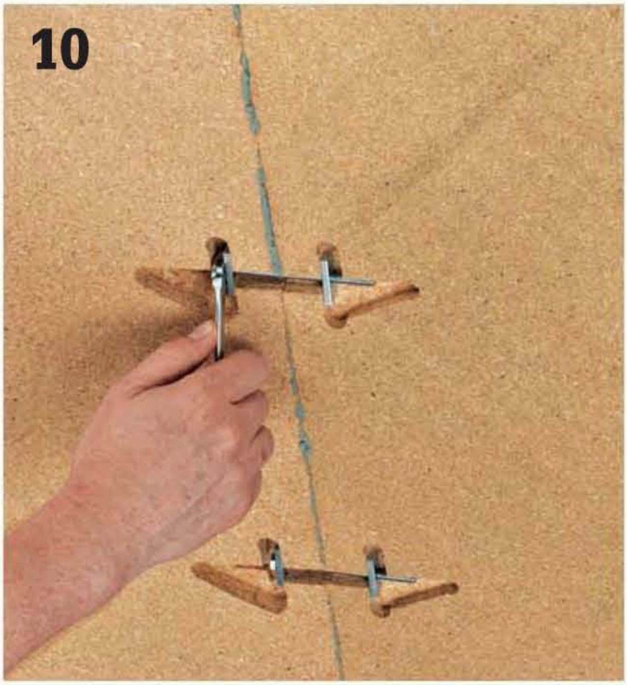

10

From underneath the countertop, install and tighten miter take-up bolts. Position countertop tightly against the wall and fasten it to the cabinets by driving wallboard screws up through corner brackets and into the countertop. Screws should be long enough to provide maximum holding power, but not long enough to puncture the laminate surface.

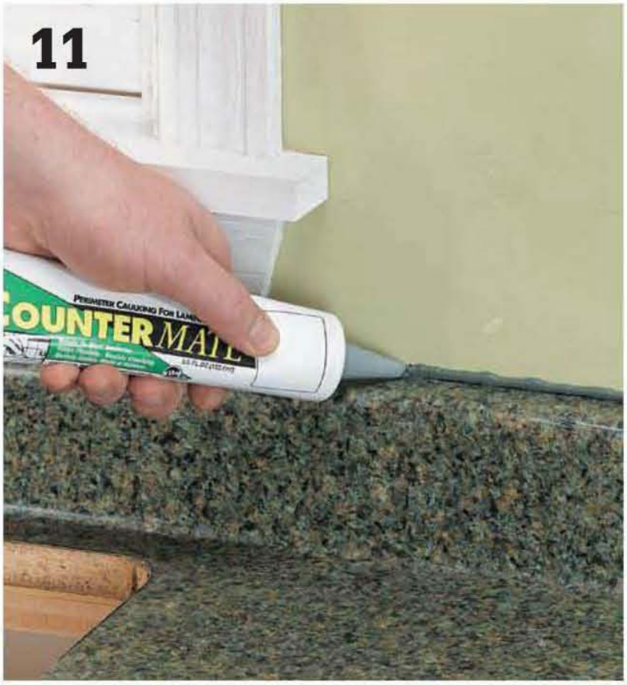

11

Seal the seam between the backsplash and the wall with silicone caulk. Smooth the bead with a wet fingertip. Wipe away excess caulk.

Installing Laminate Countertops

Tools & Materials ▸

- Tape measure
- Framing square
- Straightedge
- Scoring tool
- Paint roller
- 3-way clamps
- Caulk gun
- J-roller
- Miter saw
- Scribing compass
- Circular saw
- Screwdriver
- Belt sander
- File
- Router
- ¾" particleboard
- Sheet laminate
- Contact cement and thinner
- Wood glue
- Drywall screws

Building your own custom laminate countertop using sheets of plastic laminate and particleboard offers two advantages: the countertop you get will be less expensive than a custom-ordered countertop, and it will allow you more options in terms of colors and edge treatments. A countertop made with laminates also can be tailored to fit any space, unlike premade countertop material that is a standard width (usually 25").

Laminate commonly is sold in 8-ft. or 12-ft. lengths that are about 1/20" thick. In width, they range from 30" strips to 48" sheets. The 30" strips are sized specifically for countertops, allowing for a 25"-wide countertop, a 1½" wide front edge strip and a short backsplash.

The plastic laminate is bonded to the particle-board or MDF substrate with contact cement (although most professional installers use products that are available only to the trades). Water-base contact cement is nonflammable and nontoxic, but solvent-base contact cement (which requires a respirator and is highly flammable) creates a much stronger, more durable bond.

Fabricating your own custom countertop from particleboard and plastic laminate is not exactly an easy DIY project, but it gives you unlimited options and the results can be very satisfying.

Tips for Working with Laminate ▸

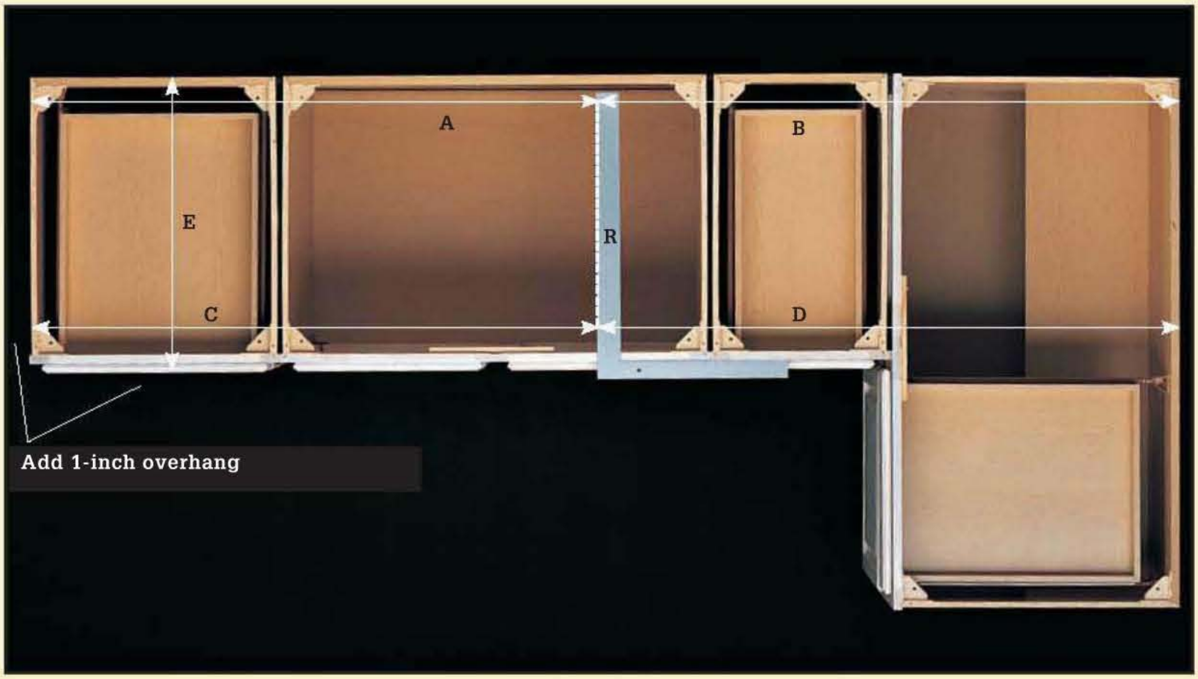

Measure along tops of base cabinets to determine the size of the countertop. If wall corners are not square, use a framing square to establish a reference line (R) near the middle of the base cabinets, perpendicular to the front of the cabinets. Take four measurements (A, B, C, D) from the reference line to the cabinet ends. Allow for overhangs by adding 1" to the length for each exposed end, and 1" to the width (E).

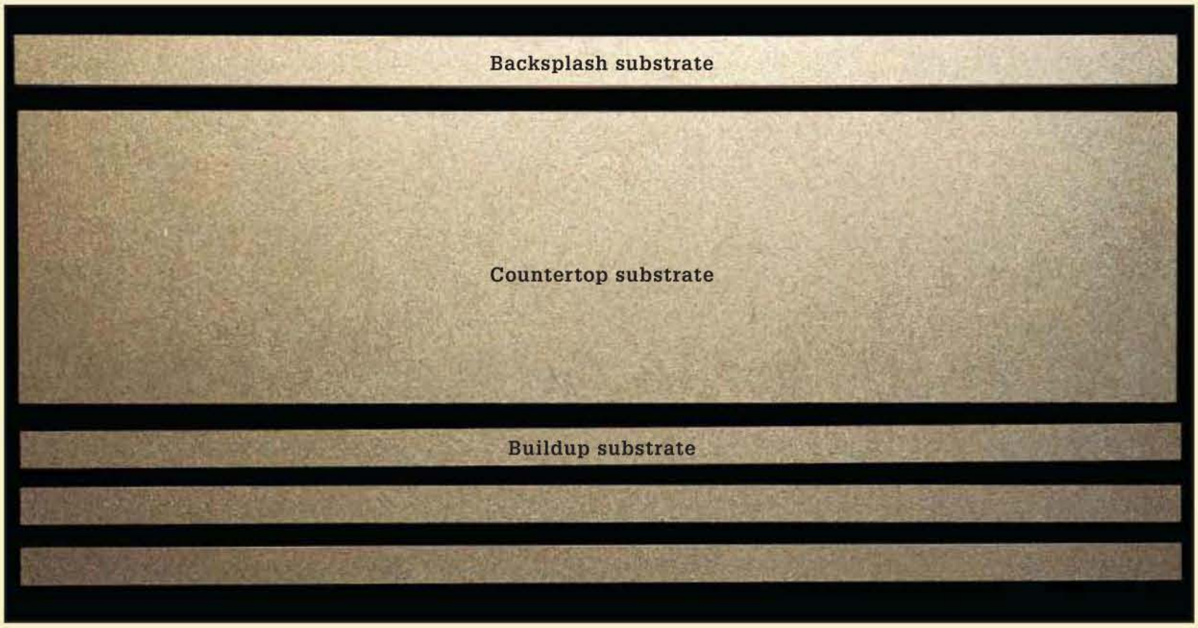

Lay out cutting lines on the particleboard so you can rip-cut the substrate and build-up strips to size, using a framing square to establish a reference line. Cut core to size using a circular saw with clamped straightedge as a guide. Cut 4" strips of particleboard for backsplash, and for joint support where sections of countertop core are butted together. Cut 3" strips for edge buildups.

How to Create a Laminate Countertop

1

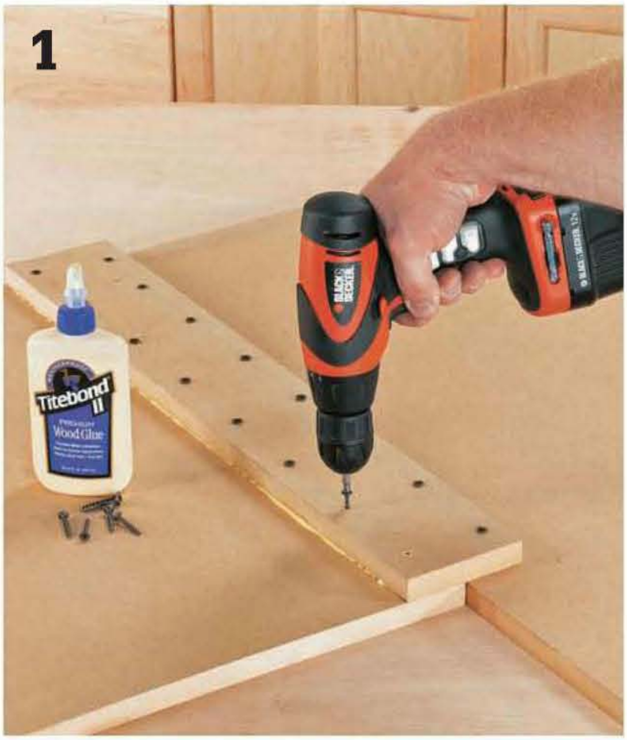

Join the countertop substrate pieces on the bottom side (see page 369 for sizing and cutting substrate). Attach a 4" particleboard joint support across the seam, using carpenter's glue and 1¼" wallboard screws.

2

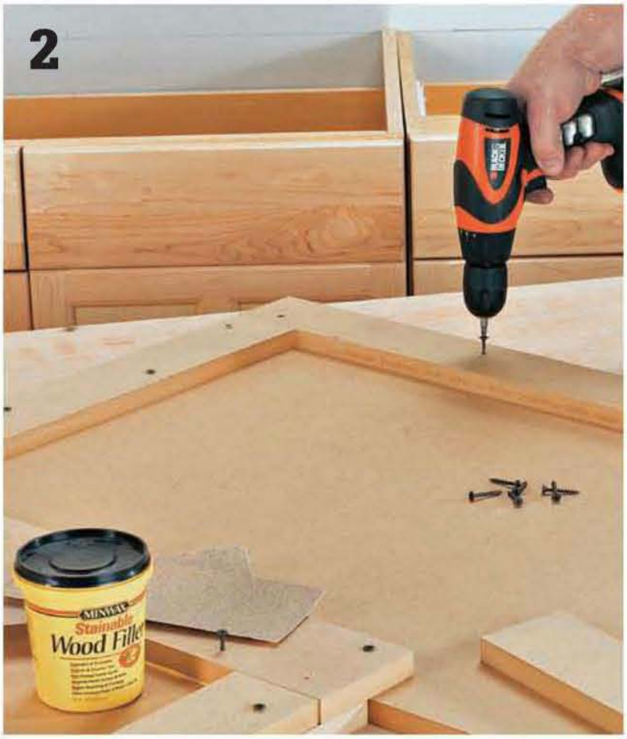

Attach 3"-wide edge buildup strips to the bottom of the countertop, using 1¼" wallboard screws. Fill any gaps on the outside edges with latex wood filler, and then sand the edges with a belt sander.

3

To determine the size of the laminate top, measure countertop substrate. Laminate seams should not overlap substrate. Add ½" trimming margin to both the length and width of each piece. Measure laminate needed for face and edges of backsplash, and for exposed edges of countertop substrate. Add ½" to each measurement.

4

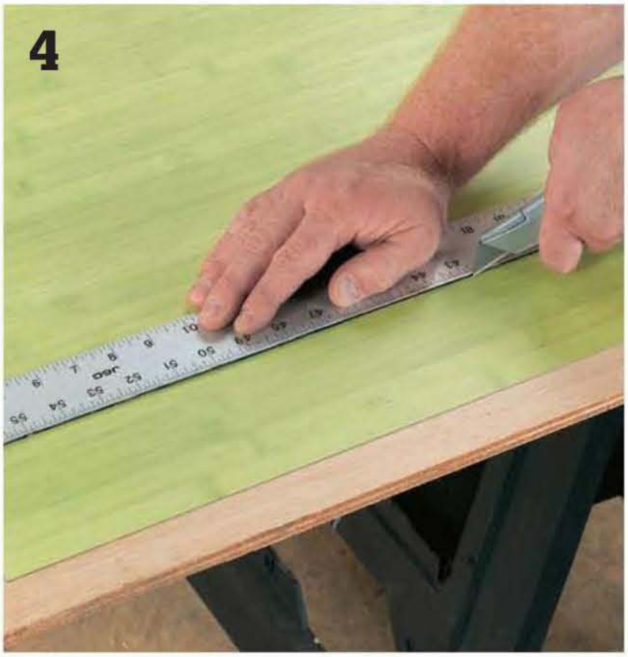

Cut laminate by scoring and breaking it. Draw a cutting line, then etch along the line with a utility knife or other sharp cutting tool. Use a straightedge as a guide. Two passes of scoring tool will help laminate break cleanly.

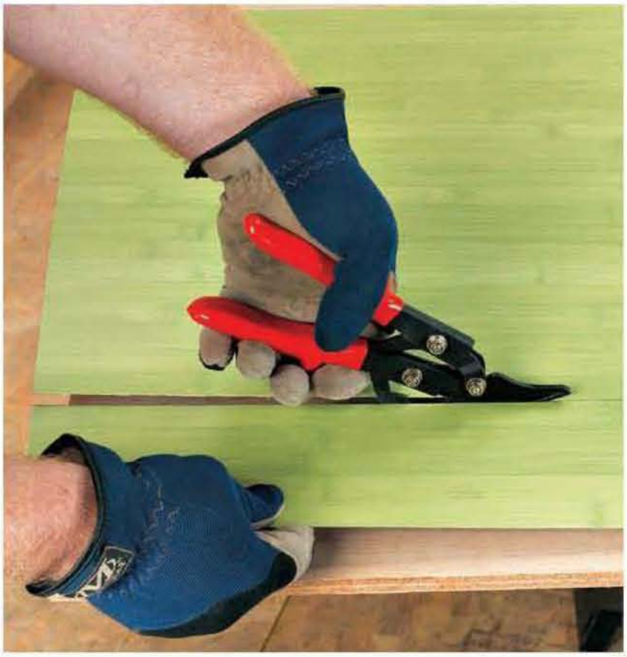

Option: Some laminate installers prefer to cut laminate with special snips that resemble avaiator snips. Available from laminate suppliers, the snips are faster than scoring and snapping, and less likely to cause cracks or tears in the material. You'll still need to square the cut edges with a trimmer or router.

5

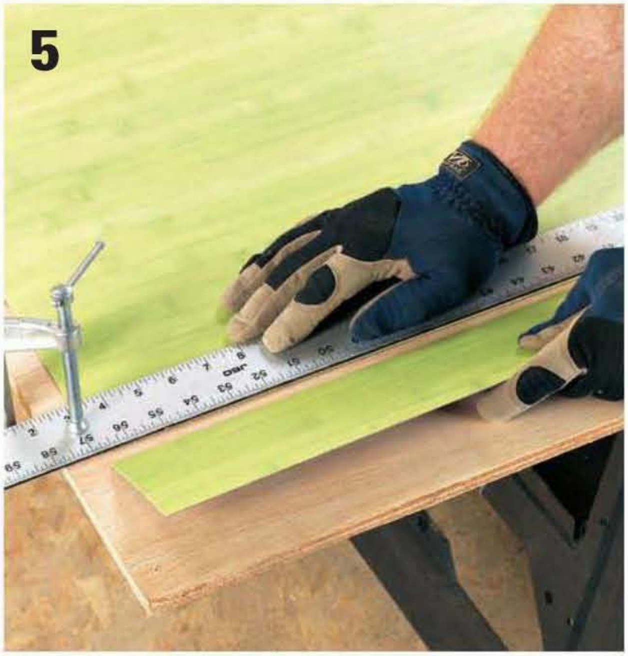

Bend laminate toward the scored line until the sheet breaks cleanly. For better control on narrow pieces, clamp a straightedge along scored line before bending laminate. Wear gloves to avoid being cut by sharp edges.

6

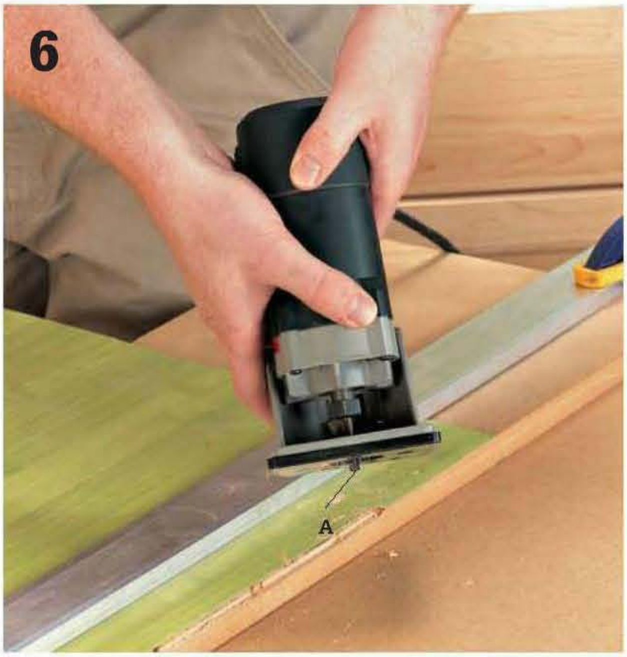

Create tight-piloted seams by using a router and a straight bit to trim edges that will butt together. Measure from cutting edge of the bit to edge of the router baseplate (A). Place laminate on scrap wood and align edges. To guide the router, clamp a straightedge on the laminate at distance A plus ¼", parallel to laminate edge. Trim laminate.

(continued)

7

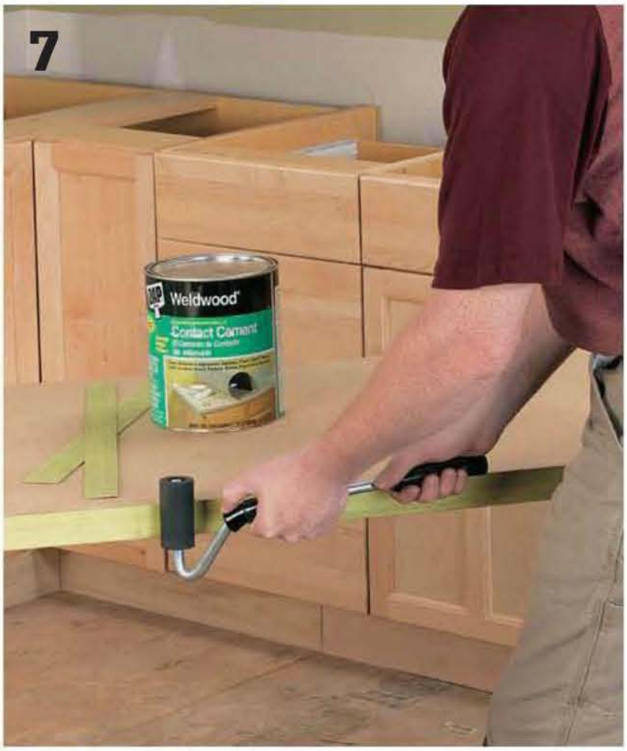

Apply laminate to sides of countertop first. Using a paint roller, apply two coats of contact cement to the edge of the countertop and one coat to back of laminate. Let cement dry according to manufacturer's directions. Position laminate carefully, then press against edge of countertop. Bond with J-roller.

8

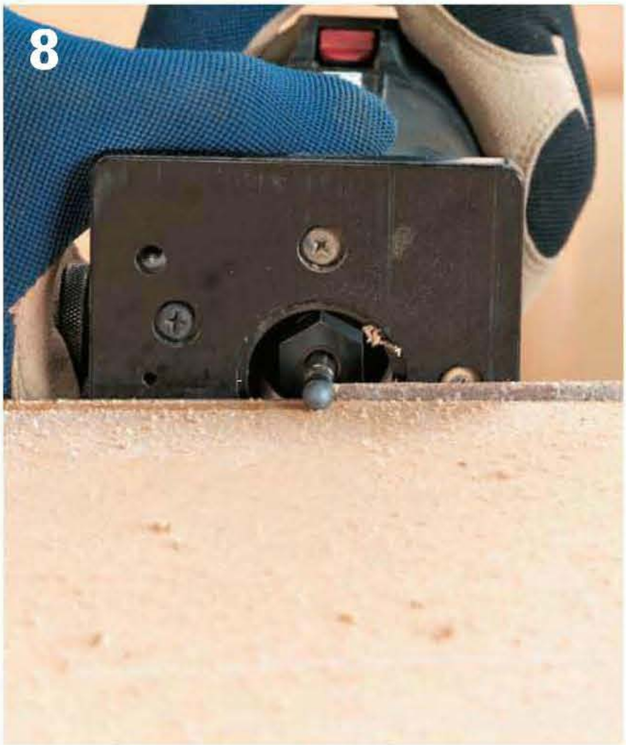

Use a router and flush-cutting bit to trim edge strip flush with top and bottom surfaces of countertop substrate. At edges where router cannot reach, trim excess laminate with a file. Apply laminate to remaining edges, and trim with router.

9

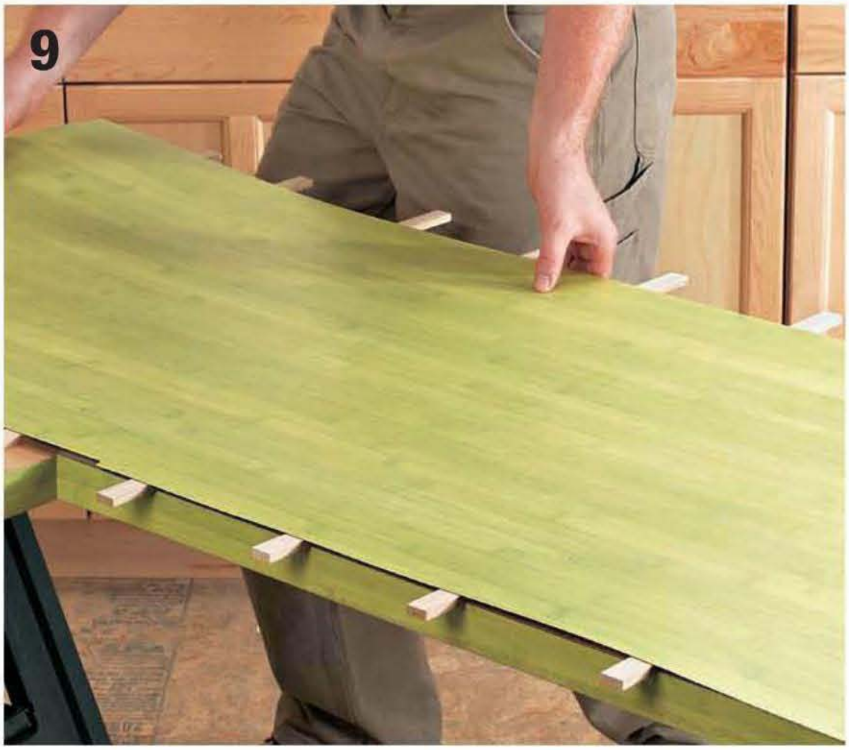

Test-fit laminate top on countertop substrate. Make sure laminate overhangs all edges. At seam locations, draw a reference line on core where laminate edges will butt together. Remove laminate. Make sure all surfaces are free of dust, then apply one coat of contact cement to back of laminate and two coats to substrate. Place spacers made of ¼"-thick scrap wood at 6" intervals across countertop core. Because contact cement bonds instantly, spacers allow laminate to be positioned accurately over core without bonding. Align laminate with seam reference line. Beginning at one end, remove spacers and press laminate to countertop core.

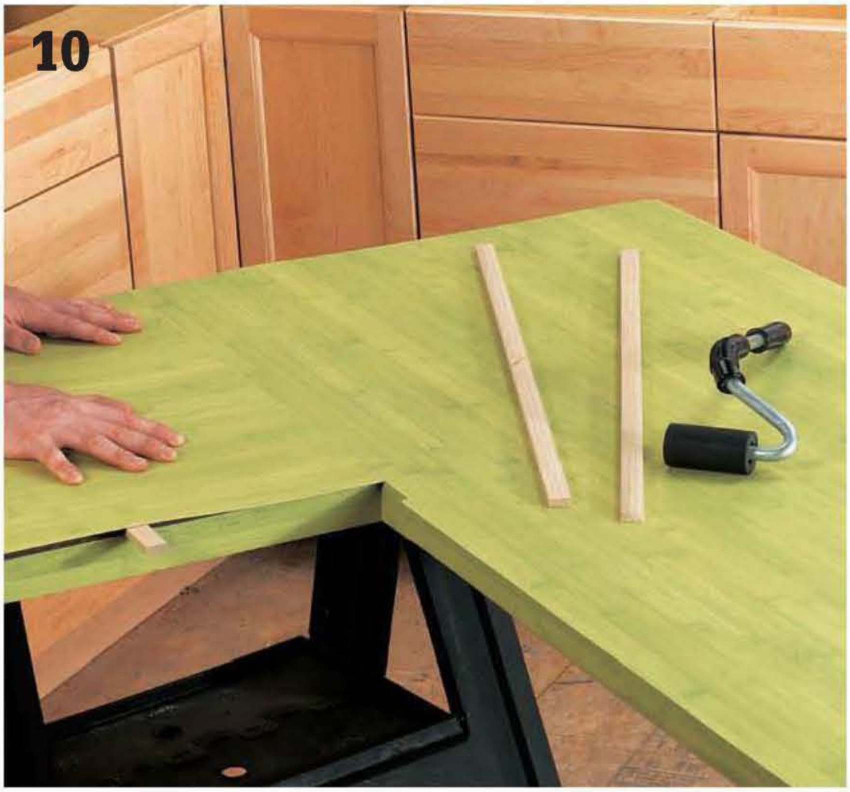

Apply contact cement to remaining substrate and next piece of laminate. Let cement dry, then position laminate on spacers, and carefully align butt seam. Beginning at seam edge, remove spacers and press laminate to the countertop substrate.

Roll the entire surface with a J-roller to bond the laminate to the substrate. Clean off any excess contact cement with a soft cloth and mineral spirits.

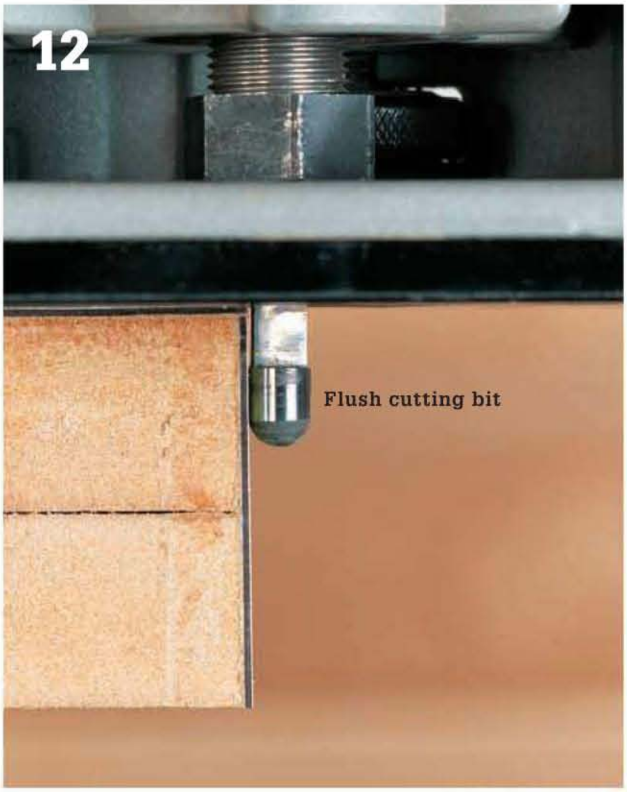

Remove excess laminate with a router and flush-cutting bit. At edges where router cannot reach, trim excess laminate with a file. Countertop is now ready for final trimming with bevel-cutting bit.

(continued)

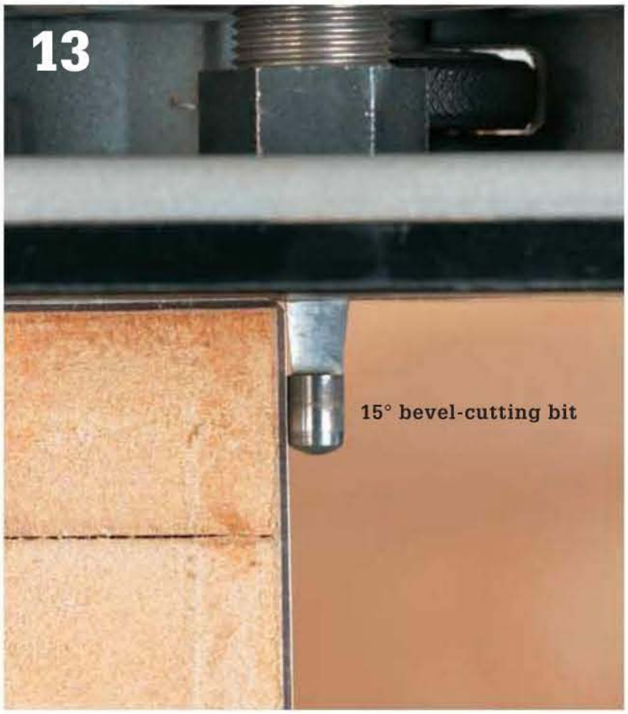

Finish-trim the edges with router and 15° bevel-cutting bit. Set bit depth so that the bevel edge is cut only on top laminate layer. Bit should not cut into vertical edge surface.

Tip ▸

File all edges smooth. Use downward file strokes to avoid chipping the laminate.

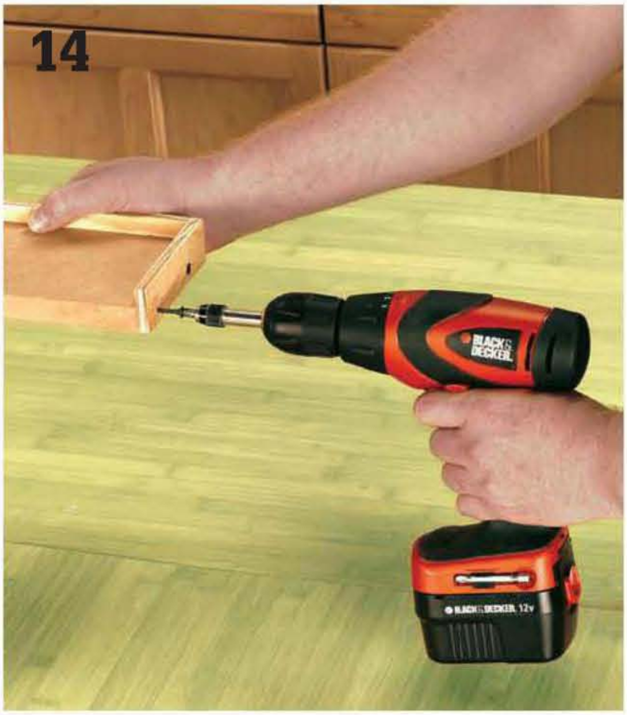

Cut 1¼"-wide strips of ¼" plywood to form an overhanging scribing strip for the backsplash. Attach to the top and sides of the backsplash substrate with glue and wallboard screws. Cut laminate pieces and apply to exposed sides, top and front of backsplash. Trim each piece as it is applied.

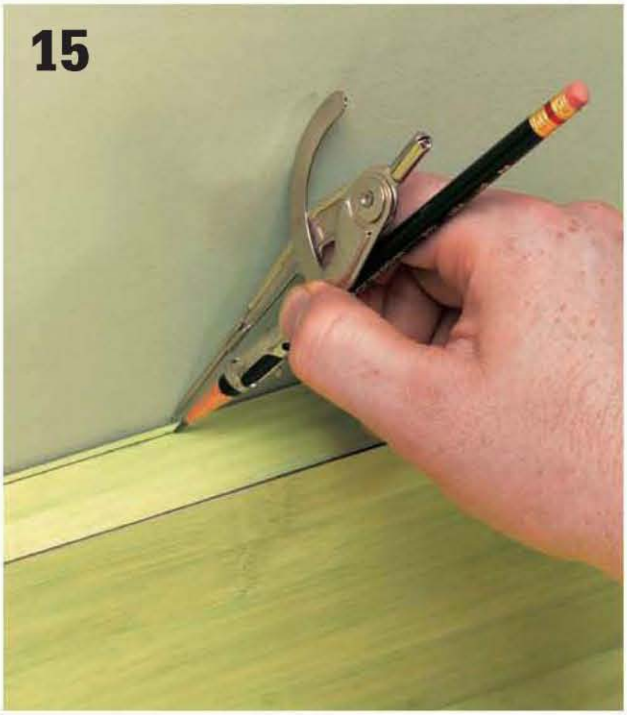

Test-fit the countertop and backsplash. Because your walls may be uneven, use a compass to trace the wall outline onto the backsplash scribing strip. Use a belt sander to grind backsplash to scribe line (see page 366).

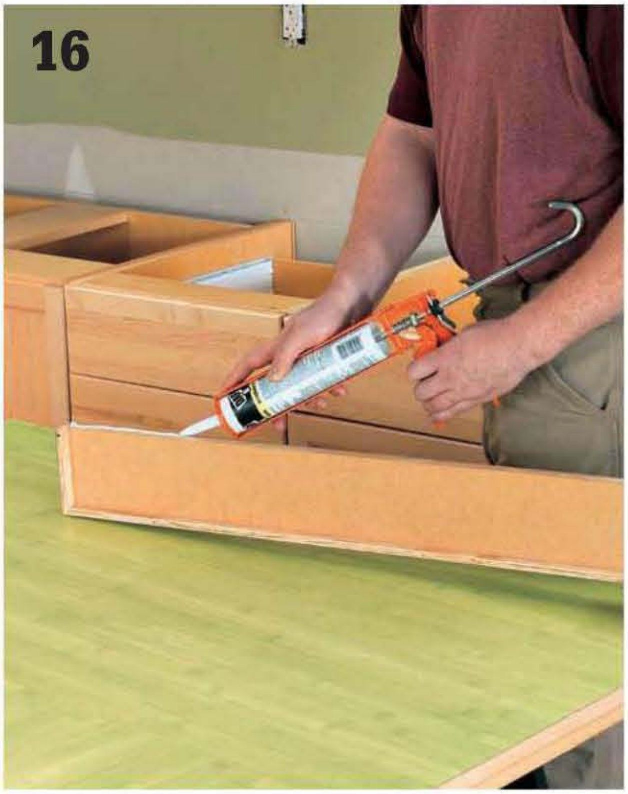
16

Apply a bead of silicone caulk to the bottom edge of the backsplash.

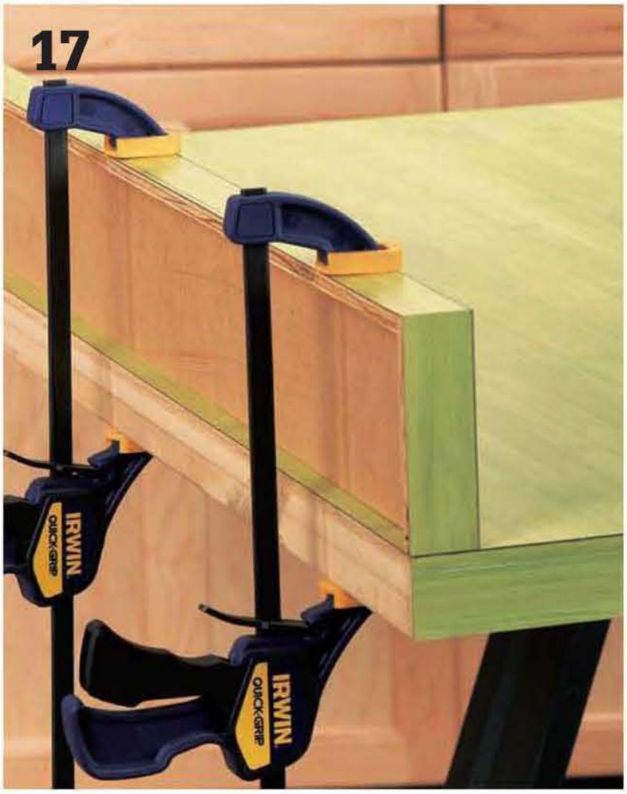

17

Position the backsplash on the countertop, and clamp it into place with bar clamps. Wipe away excess caulk, and let dry completely.

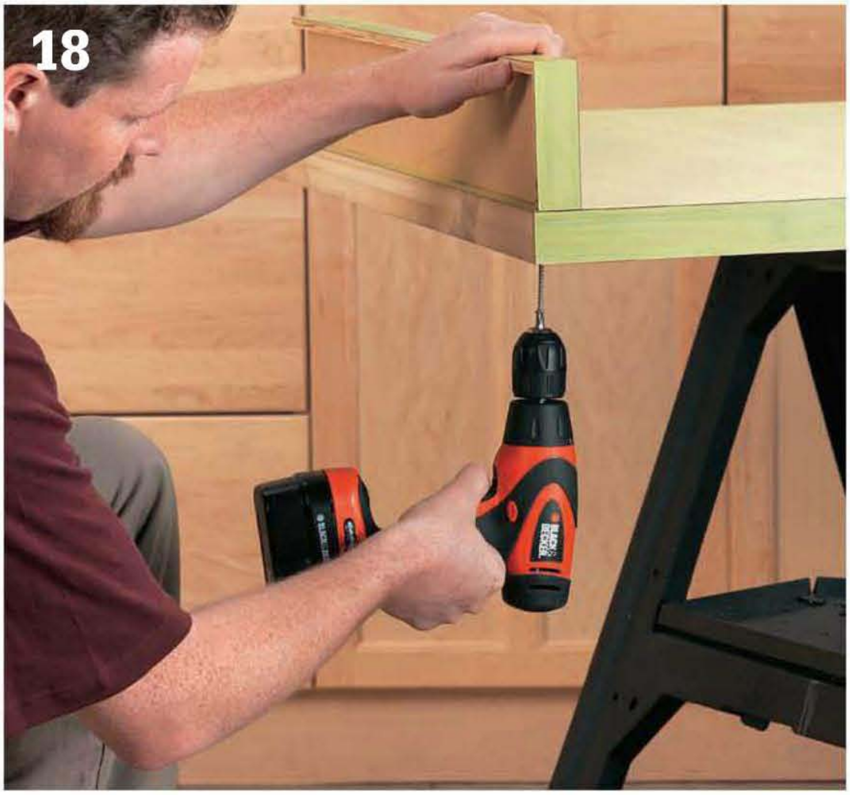
18

Drive 2" wallboard screws through the countertop and into the backsplash core. Make sure screw heads are countersunk completely for a tight fit against the base cabinet. Install countertops.

Installing Tile on Countertops

Ceramic and porcelain tile remain popular choices for countertops and backsplashes for a number of reasons: It's available in a vast range of sizes, styles and colors; it's durable and repairable; and some tile—not all—is reasonably priced. With careful planning, tile is also easy to install, making a custom countertop a good do-it-yourself project.

The best tile for most countertops is glazed ceramic or porcelain floor tile. Glazed tile is better than unglazed because of its stain resistance, and floor tile is better than wall tile because it's thicker and more durable.

While glazing protects tile from stains, the grout between tiles is still vulnerable because it's so porous. To minimize staining, use a grout that contains a latex additive, or mix the grout with a liquid latex additive. After the grout cures fully, apply a quality grout sealer, and reapply the sealer once a year thereafter. Choosing larger tiles reduces the number of grout lines. Although the selection is a bit limited, if you choose 13" × 13" floor tile, you can span from the front to the back edge of the countertop with a single seam.

The countertop in this project has a substrate of ¾" exterior-grade plywood that's cut to fit and fastened to the cabinets. The plywood is covered with a layer of plastic (for a moisture barrier) and a layer of ½"-thick cementboard. Cementboard is an effective backer for tile because it won't break down if water gets through the tile layer. The tile is adhered to the cementboard with thin-set adhesive. The overall thickness of the finished countertop is about 1½". If you want a thicker countertop, you can fasten an additional layer of plywood (of any thickness) beneath the substrate. Two layers of ¾" exterior-grade plywood without cementboard is also an acceptable substrate.

You can purchase tiles made specifically to serve as backsplashes and front edging. While the color and texture may match, these tiles usually come in only one length, making it difficult to get your grout lines to align with the field tiles. You can solve this problem by cutting your own edging and backsplash tiles from field tiles.

Tools & Materials ▸

- Tape measure
- Circular saw
- Drill
- Utility knife
- Straightedge
- Stapler
- Drywall knife
- Framing square
- Notched trowel
- Tile cutter
- Carpeted 2 × 4
- Mallet
- Rubber grout float
- Sponge
- Foam brush
- Caulk gun
- Ceramic tile
- Tile spacers
- ¾" exterior-grade (CDX) plywood
- 4-mil polyethylene sheeting
- Packing tape
- ½" cementboard
- 1¼" galvanized deck screws
- Fiberglass mesh tape
- Thin-set mortar
- Grout with latex additive
- Silicone caulk
- Silicone grout sealer

Ceramic or porcelain makes a durable countertop that is heat-resistant and relatively easy for a DIYer to create. By using larger tiles, you minimize the grout lines (and the cleaning that goes with them).

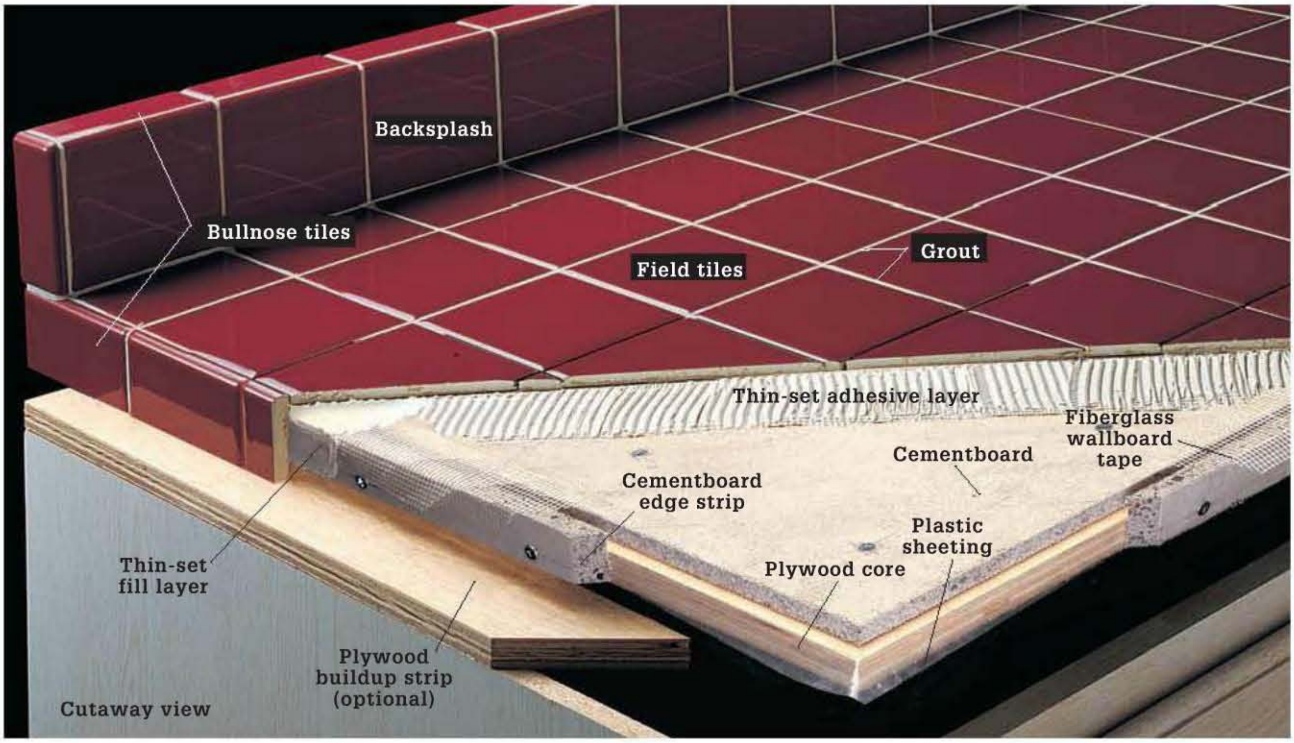
Backsplash
Bullnose tiles
Field tiles
Grout
Thin-set adhesive layer
Fiberglass wallboard tape
Cementboard
Cementboard edge strip
Plastic sheeting
Plywood core
Thin-set fill layer
Plywood buildup strip (optional)
Cutaway view

A ceramic tile countertop starts with a core of ¾" exterior-grade plywood that's covered with a moisture barrier of 4-mil polyethylene sheeting. Half-inch cementboard is screwed to the plywood, and the edges are capped with cementboard and finished with fiberglass mesh tape and thin-set mortar. Tiles for edging and backsplashes may be bullnose or trimmed from the factory edges of field tiles.

Options for Backsplashes & Countertop Edges

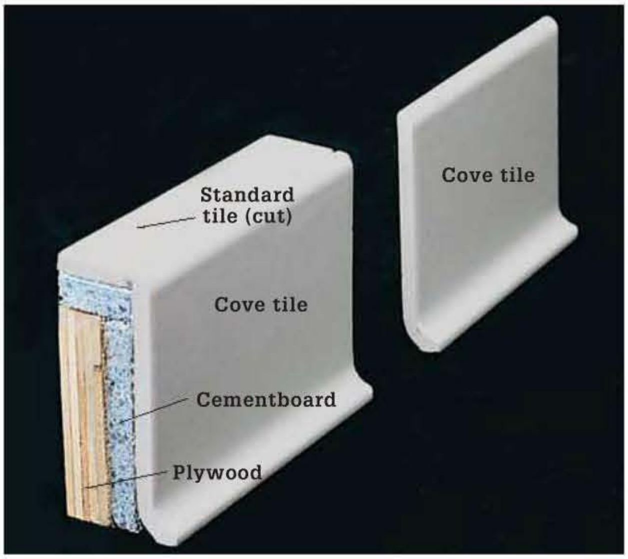
Standard tile (cut)
Cove tile
Cove tile
Cementboard
Plywood

Backsplashes can be made from cove tile attached to the wall at the back of the countertop. You can use the tile alone or build a shelf-type backsplash, using the same construction as for the countertop. Attach the plywood backsplash to the plywood core of the countertop. Wrap the front face and all edges of the plywood backsplash with cementboard before laying tile.

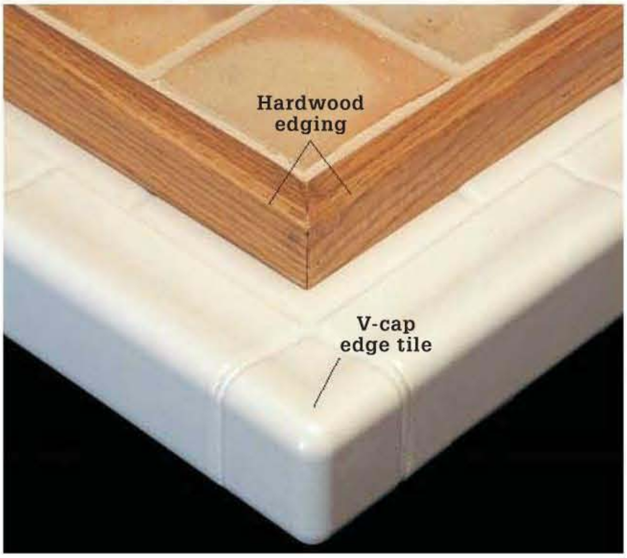
Hardwood edging
V-cap edge tile

Edge options include V-cap edge tile and hardwood strip edging. V-cap tiles have raised and rounded corners that create a ridge around the countertop perimeter—good for containing spills and water. V-cap tiles must be cut with a tile saw. Hardwood strips should be prefinished with at least three coats of polyurethane finish. Attach the strips to the plywood core so the top of the wood will be flush with the faces of the tiles.

Tips for Laying Out Tile ▸

- You can lay tile over a laminate countertop that's square, level and structurally sound. The laminate cannot have a no-drip edge. Use a belt sander with 60- or 80-grit sandpaper to rough up the surface before setting the tiles. If you're using a new substrate and need to remove your existing countertop, make sure the base cabinets are level front to back, side to side and with adjoining cabinets. If necessary, unscrew a cabinet from the wall and use shims on the floor or against the wall to level it.
- Installing battens along the front edge of the countertop helps ensure the first row of tile is perfectly straight. For V-cap tiles, fasten a 1 × 2 batten along the reference line, using screws. The first row of field tile is placed against this batten. For bullnose tiles, fasten a batten that's the same thickness as the edging tile, plus ⅛" for mortar thickness, to the face of the countertop so the top is flush with the top of the counter. The bullnose tiles are aligned with the outside edge of the batten. For wood edge trim, fasten a 1 × 2 batten to the face of the countertop so the top edge is above the top of the counter. The tiles are installed against the batten.
- Before installing any tile, lay out the tiles in a dry run, using spacers. If your counter is L-shaped, start at the corner and work outward. Otherwise, start the layout at a sink to ensure equal-sized cuts on both sides of the sink. If necessary, shift your starting point so you don't end up cutting very narrow tile segments.

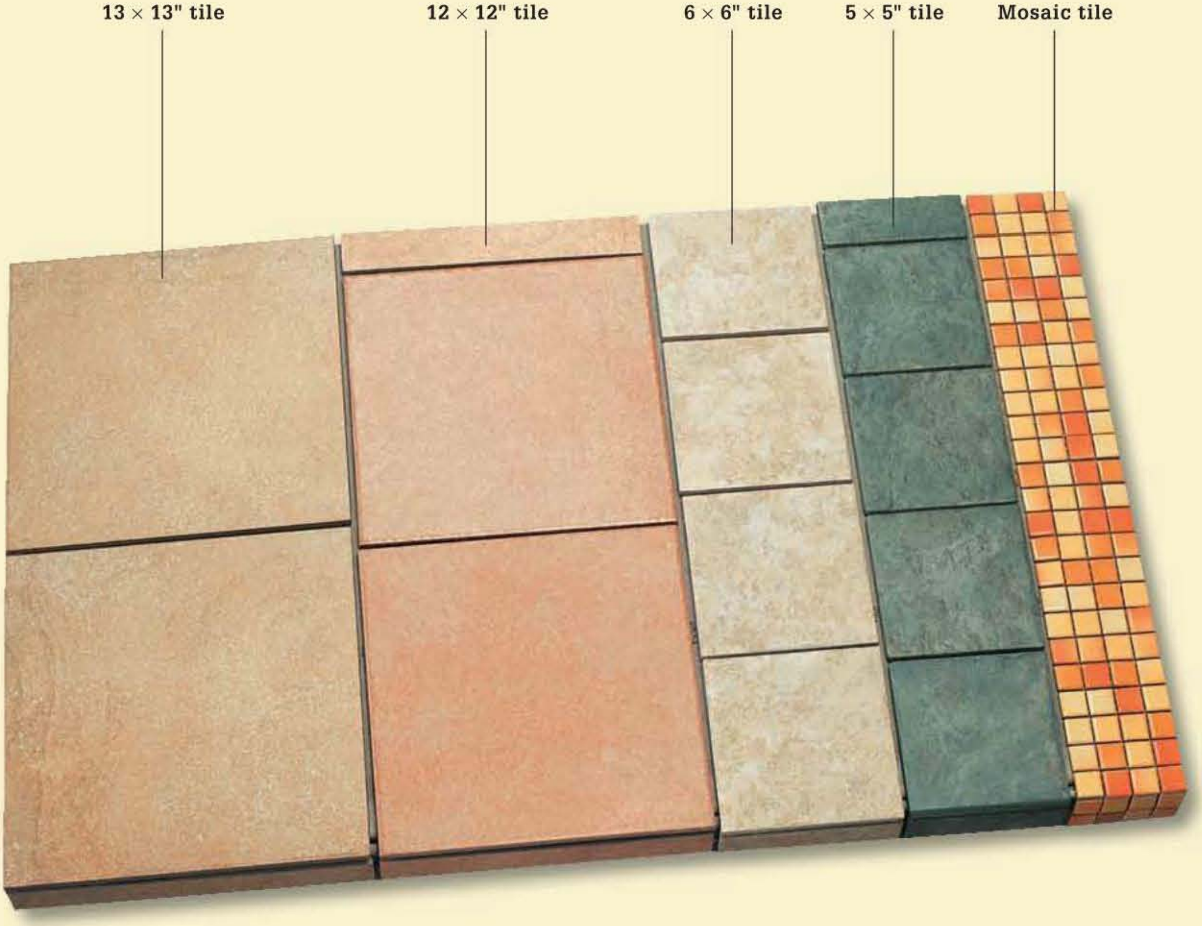

The bigger the tile, the fewer the grout lines. If you want a standard 25"-deep countertop, the only way to get there without cutting tiles is to use mosaic strips or 1" tile. With 13 × 13" tile, you need to trim 1" off the back tile but have only one grout line front to back. As the size of the tile decreases, the number of grout lines increases.

How to Build a Tile Countertop

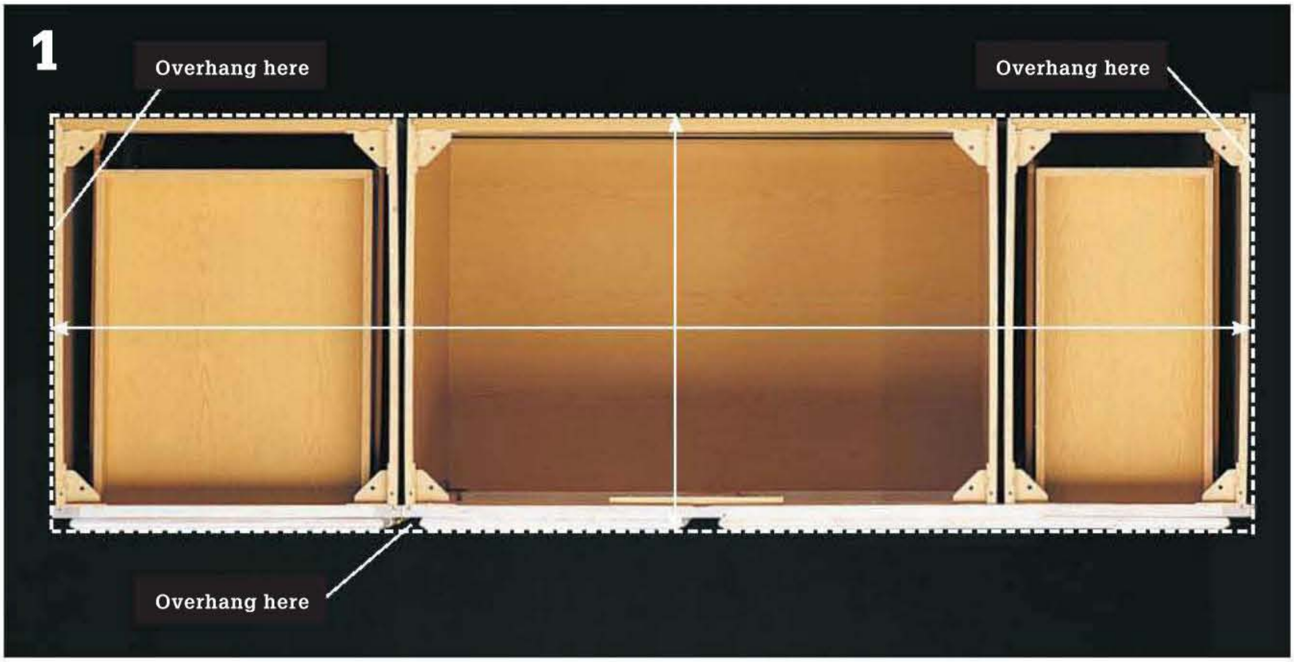

Determine the size of the plywood substrate by measuring across the top of the cabinets. The finished top should overhang the drawer fronts by at least ¼". Be sure to account for the thickness of the cementboard, adhesive and tile when deciding how large to make the overhang. Cut the substrate to size from ¾" plywood, using a circular saw. Make cutouts for sinks and other fixtures.

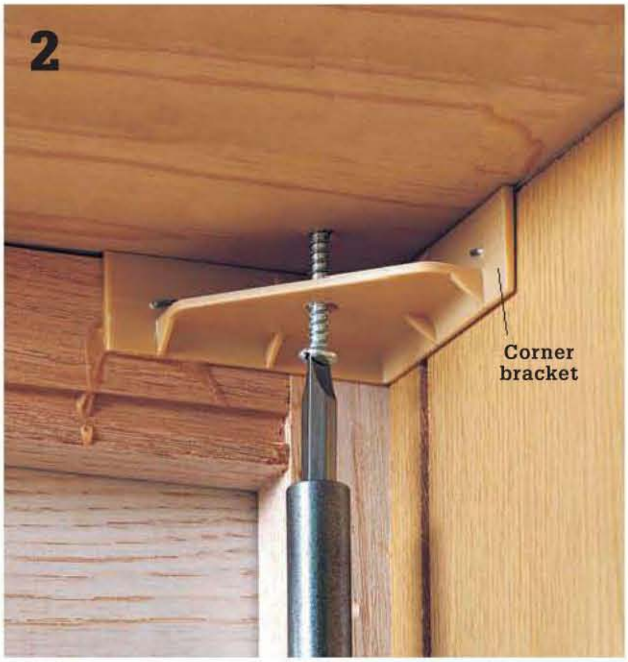

Set the plywood substrate on top of the cabinets, and attach it with screws driven through the cabinet corner brackets. The screws should not be long enough to go through the top of the substrate.

Cut cementboard using a straightedge and utility knife or a cementboard cutter with a carbide tip. Hold the straightedge along the cutting line, and score the board several times with the knife. Bend the piece backward to break it along the scored line. Back-cut to finish.

(continued)

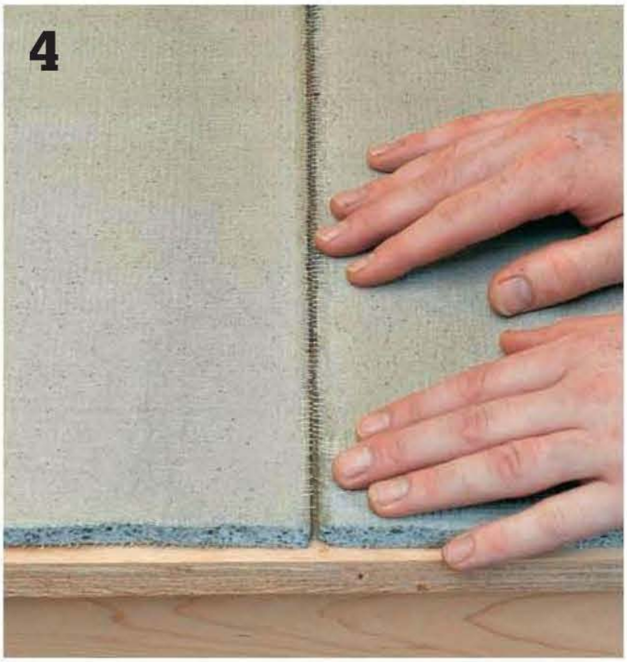

Dry-fit cementboard pieces on the plywood core with the rough sides of the panels facing up. Leave a ⅛" gap between the cementboard sheets and a ¼" gap along the perimeter.

Lay the 4-mil plastic moisture barrier over the plywood substrate, draping it over the edges. Tack it in place with a few staples. Overlap seams in the plastic by 6", and seal them with packing tape.

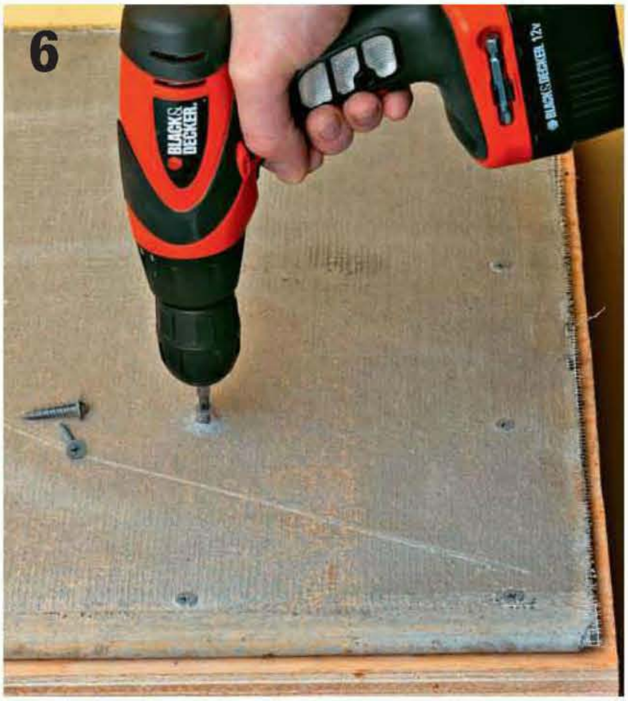

Lay the cementboard pieces rough-side up on the plywood and attach them with cementboard screws driven every 6". Drill pilot holes using a masonry bit, and make sure all screw heads are flush with the surface. Wrap the countertop edges with 1¼"-wide cementboard strips, and attach them to the core with cementboard screws.

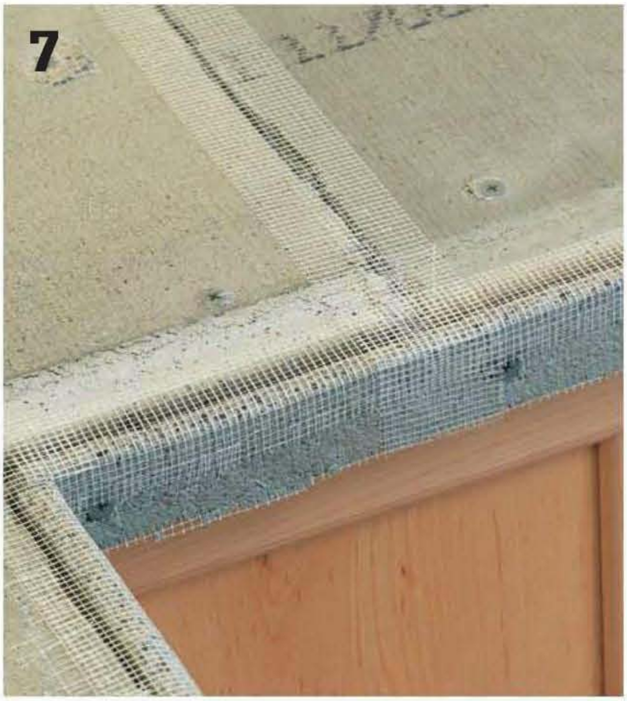

Tape all cementboard joints with fiberglass mesh tape. Apply three layers of tape along the front edge where the horizontal cementboard sheets meet the cementboard edging.

Fill all gaps and cover all of the tape with a layer of thin-set mortar. Feather out the mortar with a drywall knife to create a smooth, flat surface.

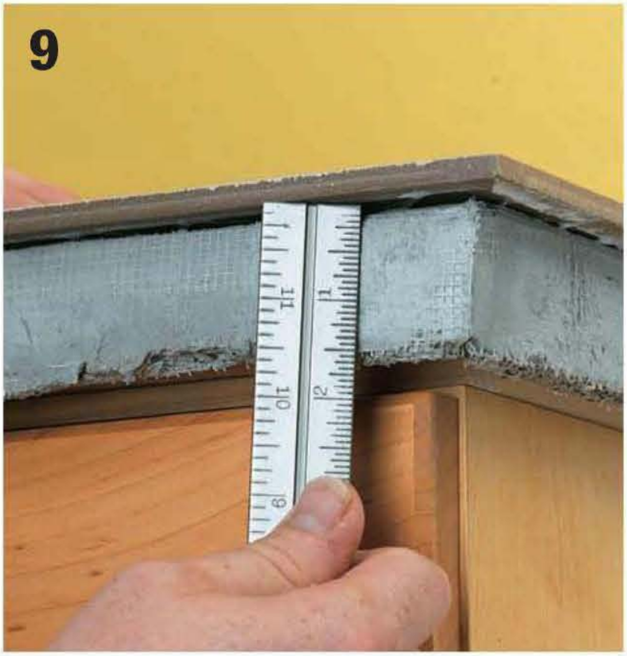

Determine the required width of your edge tiles. Lay a field tile onto the tile base so it overhangs the front edge by ½" or so. Then, hold a metal rule up to the underside of the tile and measure the distance from the tile to the bottom of the subbase. Your edge tiles should be cut to this width (the gap for the grout line will cause the edge tile to extend past the subbase, concealing it completely).

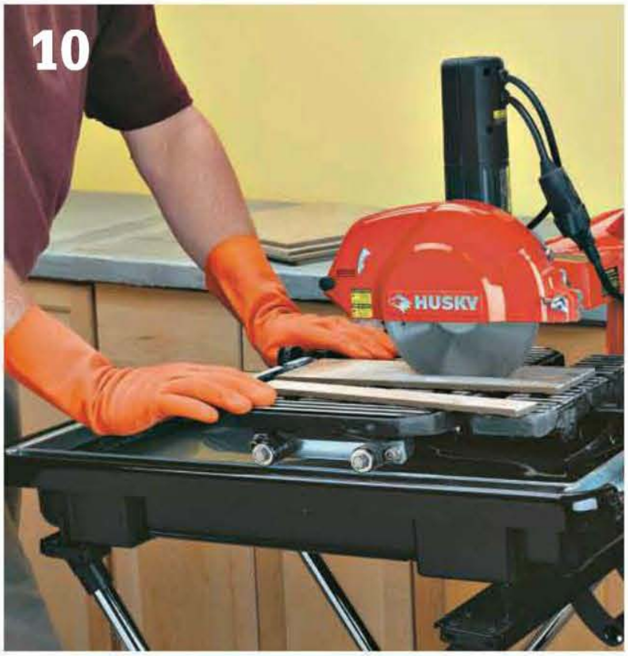

Cut your edge tiles to the determined width, using a tile saw. It's worth renting a quality wet saw for tile if you don't own one. Floor tile is thick and difficult to cut with a hand cutter (especially porcelain tiles).

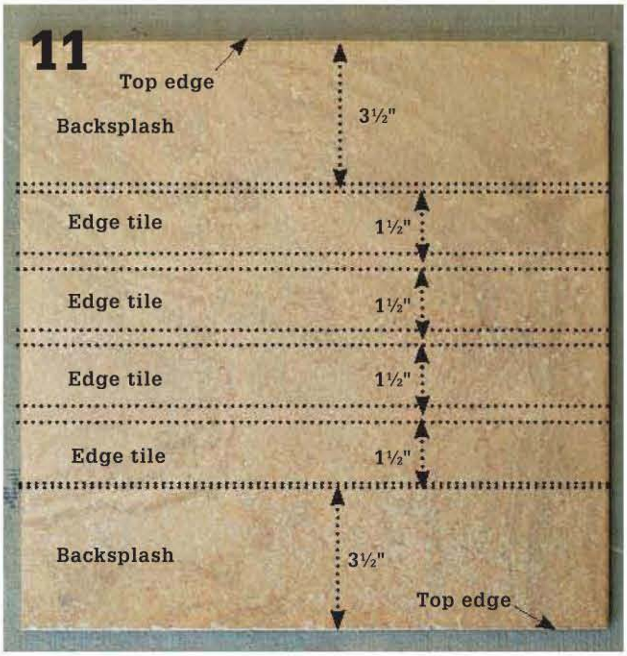

Cut tiles for the backsplash. The backsplash tiles (3½" wide in our project) should be cut with a factory edge on each tile that will be oriented upward when they're installed. You can make efficient use of your tiles by cutting edge tiles from the center area of the tiles you cut to make the backsplash.

(continued)

12

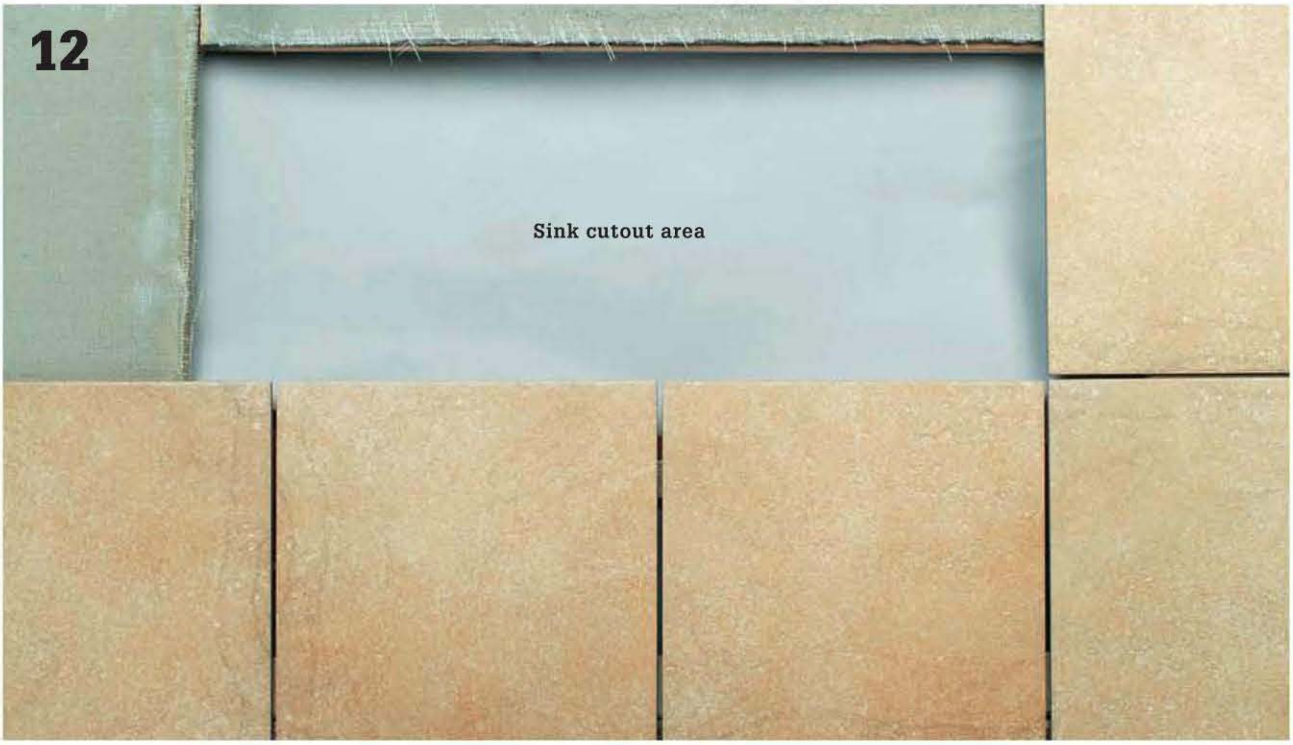

Dry-fit tiles on the countertop to find the layout that works best. Once the layout is established, make marks along the vertical and horizontal rows. Draw reference lines through the marks and use a framing square to make sure the lines are perpendicular.

Variation: Laying Out with Small Floor Tiles and Bullnose Edging ▸

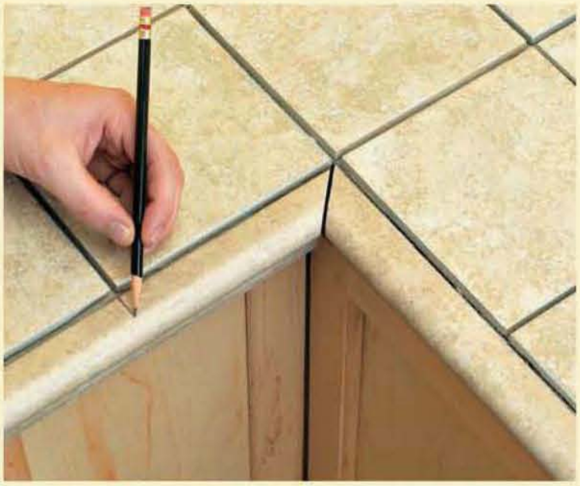

Lay out tiles and spacers in a dry run. Adjust starting lines, if necessary. If using battens, lay the field tile flush with the battens, then apply edge tile. Otherwise, install the edging first. If the countertop has an inside corner, start there by installing a ready-made inside corner or cutting a 45° miter in edge tile to make your own inside corner.

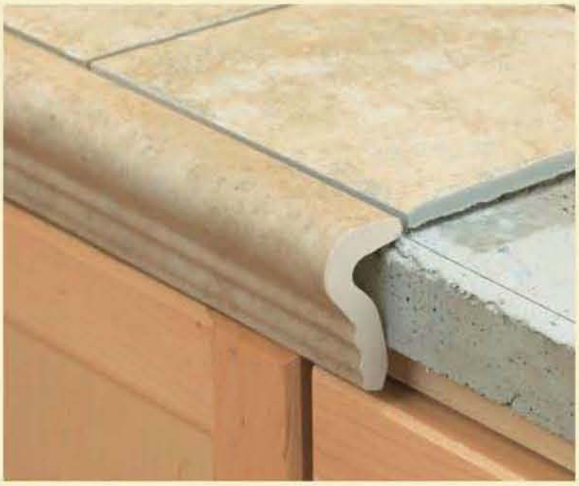

Place the first row of field tile against the edge tile, separating the tile with spacers. Lay out the remaining rows of tile. Adjust starting lines if necessary to create a layout using the least number of cut tiles.

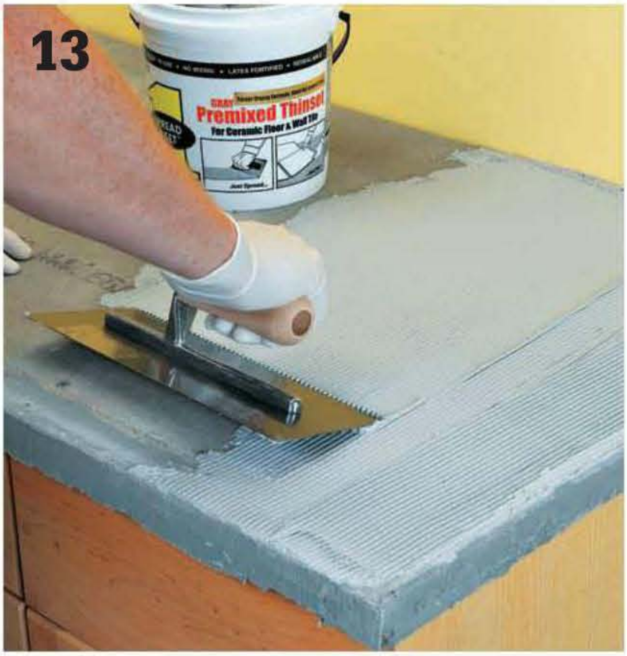

13

Use a ⅜" square notched trowel to apply a layer of thinset adhesive to the cementboard. Apply enough for two or three tiles, starting at one end. Hold the trowel at roughly a 30-degree angle and try not to overwork the adhesive or remove too much.

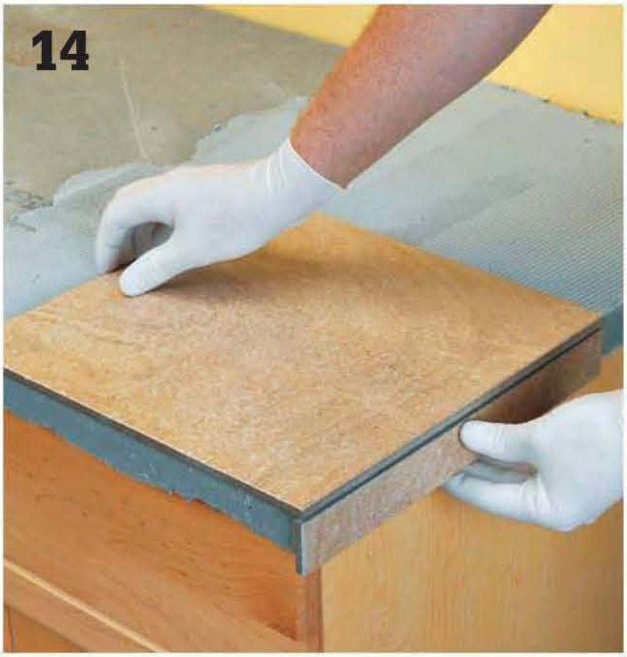

14

Set the first tile into the adhesive. Hold a piece of the edge in tile against the countertop edge as a guide to show you exactly how much the tile should overhang the edge.

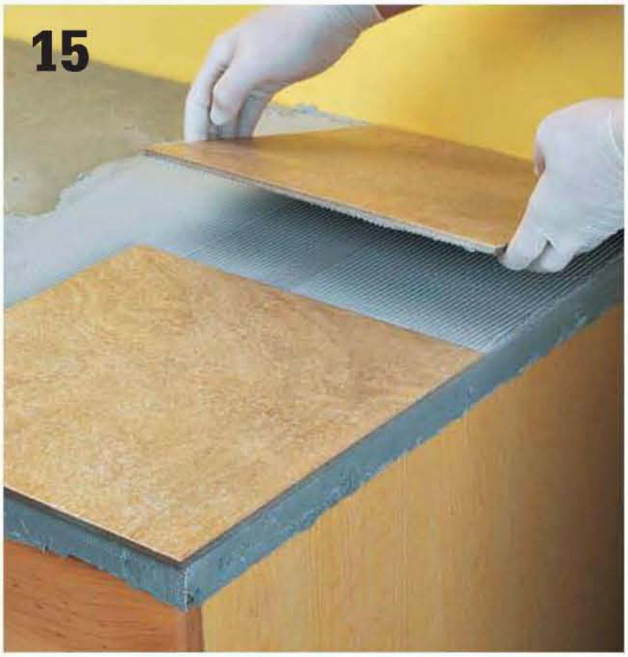

15

Cut all the back tiles for the layout to fit (you'll need to remove about 1" of a 13 × 13" tile) before you begin the actual installation. Set the back tiles into the thinset, maintaining the gap for groutlines created by the small spacer nubs cast into the tiles. If your tiles have no spacer nubs, see next step.

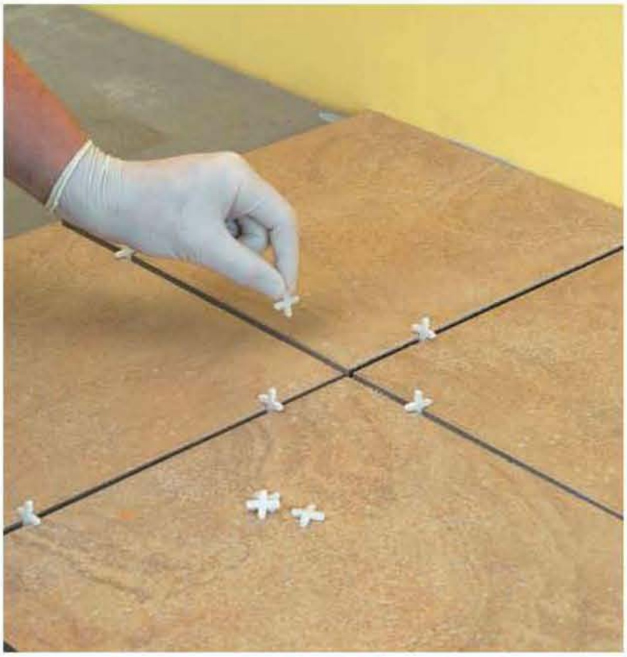

Option: To maintain even grout lines, some beginning tilers insert x-shaped plastic spacers at the joints. This is less likely to be useful with large tiles like those shown here, but it is effective with smaller tile. Many tiles today feature built-in spacing lugs, so the spacers are not necessary. If you use spacers, remove them before the thinset sets. If you leave them in place they will corrupt your grout lines.

(continued)

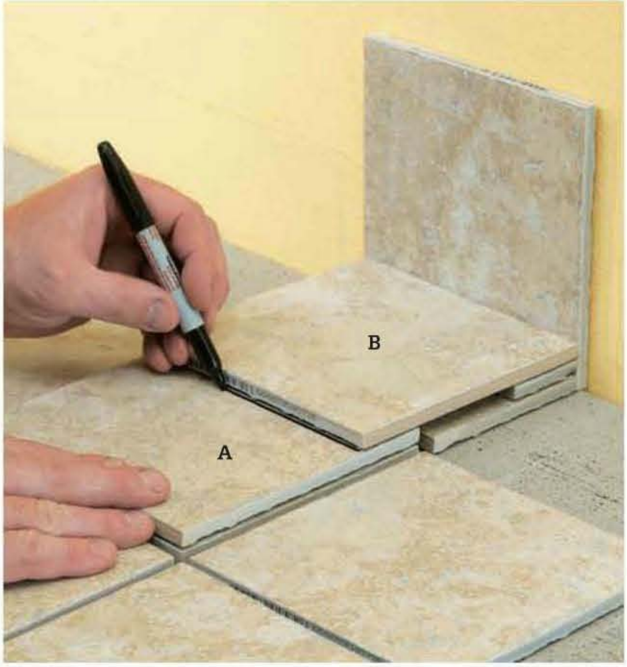

Variation: To mark border tiles for cutting, allow space for backsplash tiles, grout and mortar by placing a tile against the back wall. Set another tile (A) on top of the last full tile in the field, then place a third tile (B) over tile A and hold it against the upright tile. Mark and cut tile A and install it with the cut edge toward the wall. Finish filling in your field tiles.

16

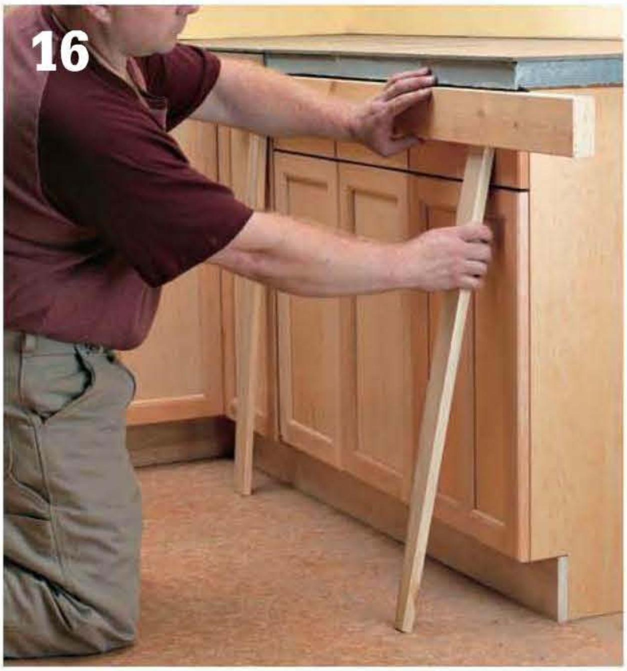

To create a support ledge for the edge tiles, prop pieces of 2 × 4 underneath the front edge of the substrate overhang, using wood scraps to prop the ledge tightly up against the substrate.

17

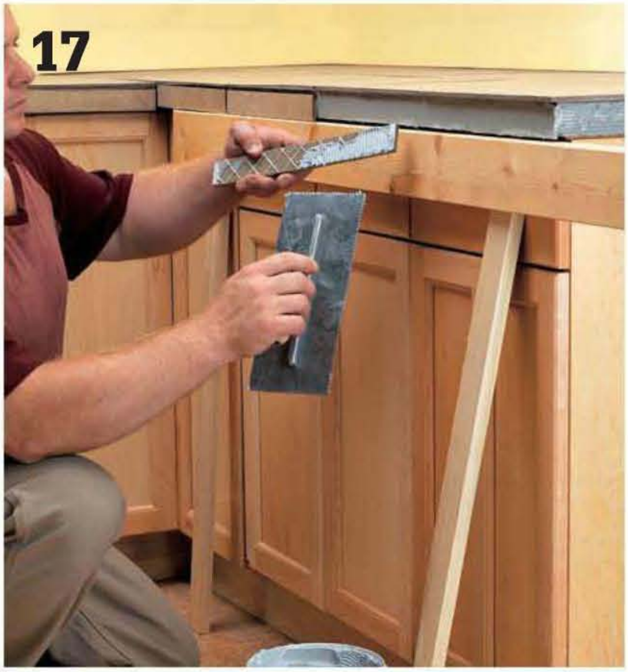

Apply a thick layer of thinset to the backside of the edge tile with your trowel. This is called "buttering" and it is easier and neater than attempting to trowel adhesive onto the countertop edge. Press the tiles into position so they are flush with the leading edges of the field tiles.

18

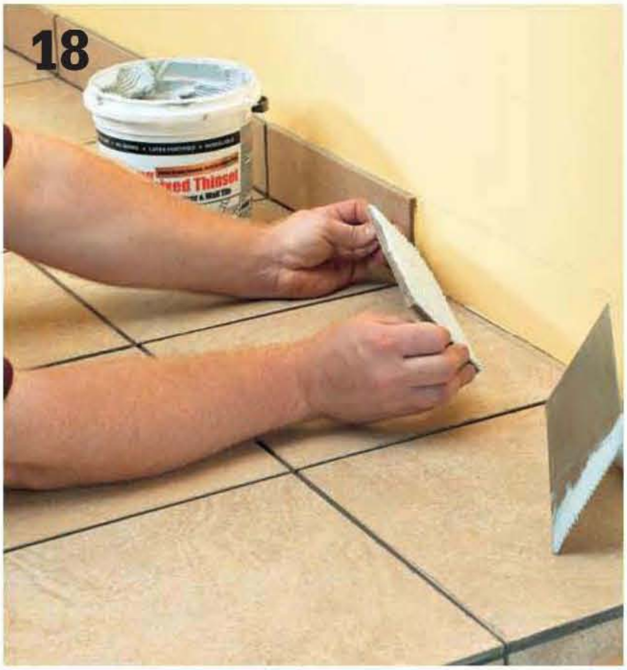

Butter each backsplash tile and press it into place, doing your best to keep all of the grout lines aligned.

19

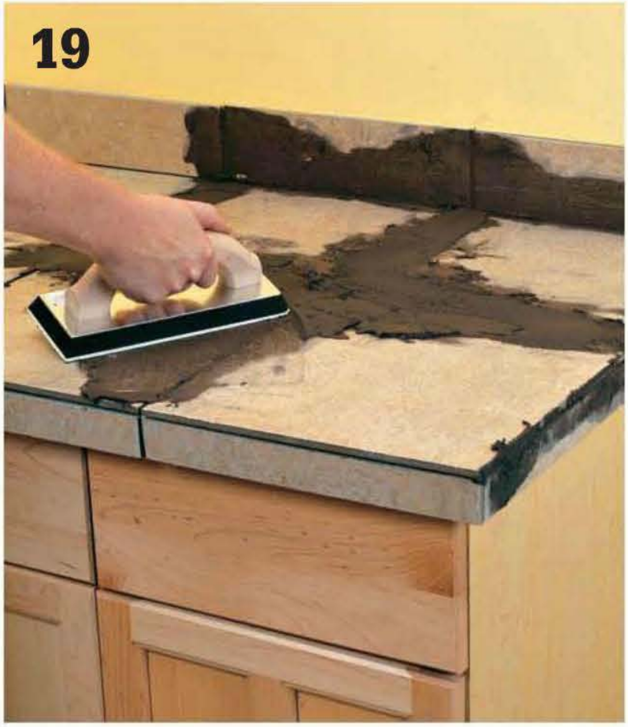

Mix a batch of grout to complement the tile (keeping in mind that darker grout won't look dirty as quickly as lighter grout). Apply the grout to the grout line areas with a grout float.

20

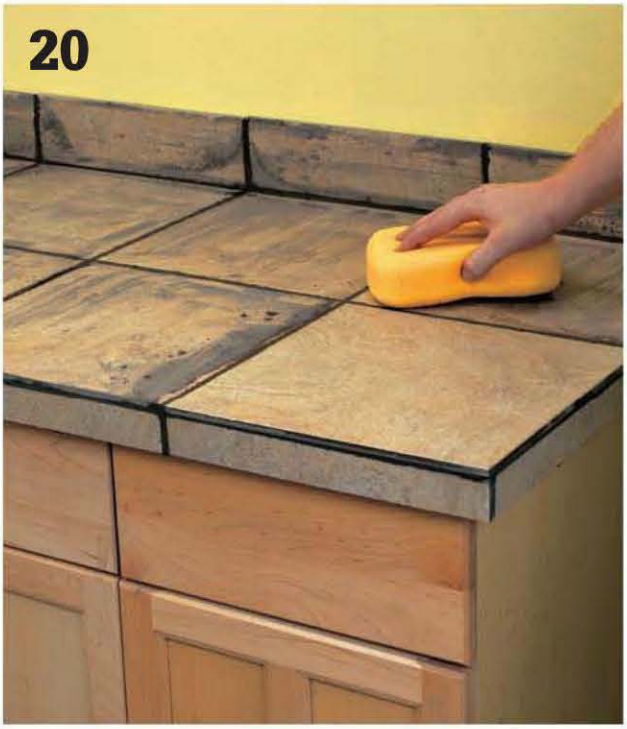

Let the grout dry until a light film is created on the countertop surface and then wipe off the excess grout using a sponge and warm, clean water.

21

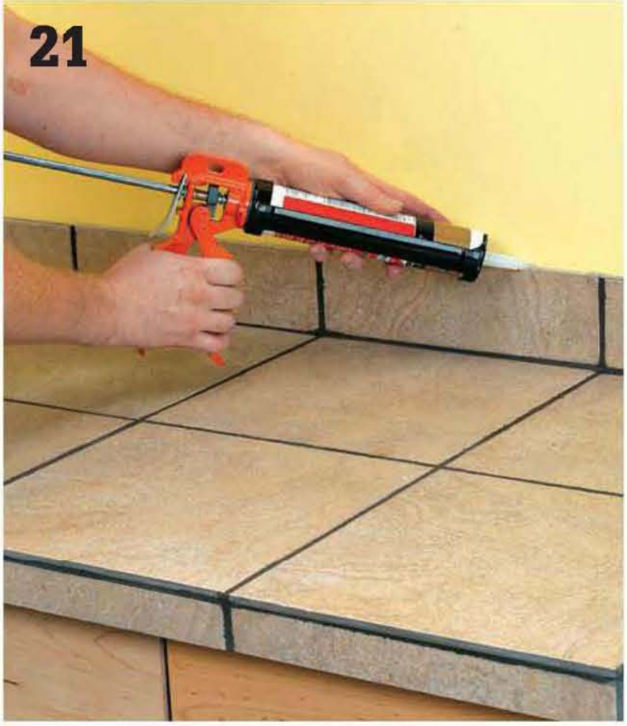

After the grout has dried run a bead of clear silicone caulk along the joint between the backsplash and the wall. Install your sink and faucet.

22

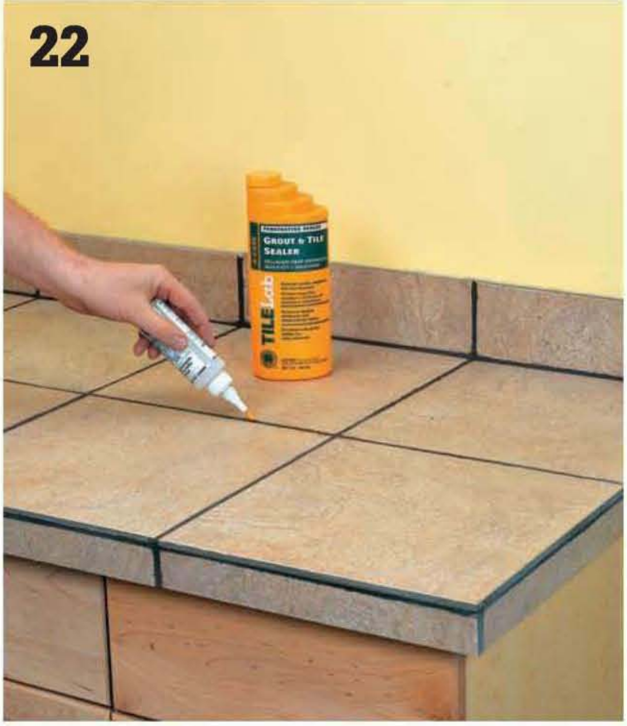

Wait at least one week and then seal the grout lines with a penetrating grout sealer. This is important to do. Sealing the tiles themselves is not a good idea unless you are using unglazed tiles (a poor choice for countertops, however).

Installing Tile Backsplashes

There are few spaces in your home with as much potential for creativity and visual impact as the space between your kitchen countertop and cupboards. A well-designed backsplash can transform the ordinary into the extraordinary.

Tiles for the backsplash can be attached directly to wallboard or plaster and do not require backerboard. When purchasing the tile, order 10 percent extra to cover breakage and cutting. Remove switch and receptacle coverplates and install box extenders to make up for the extra thickness of the tile. Protect the countertop from scratches by covering it with a drop cloth.

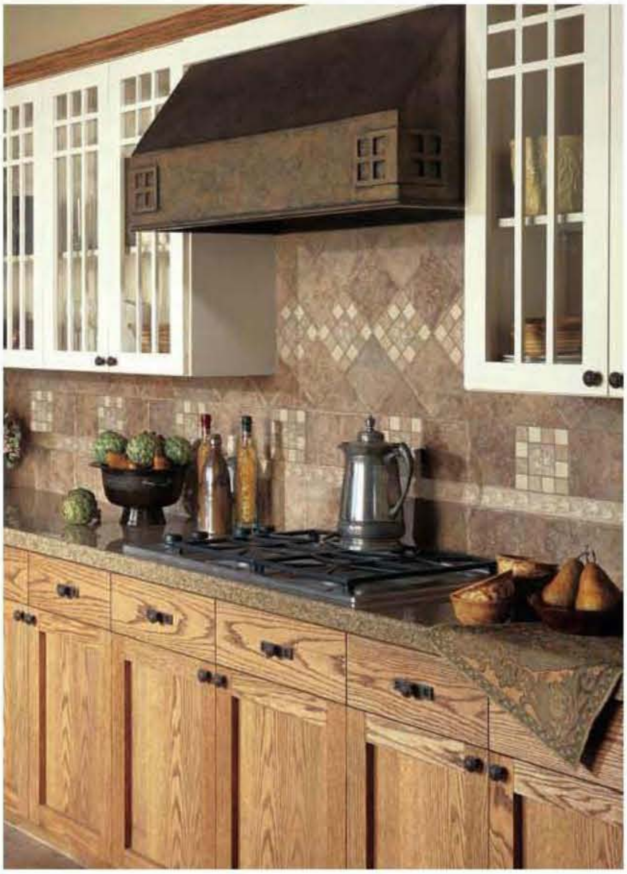

A tiled backsplash normally extends all the way from countertop to the bottoms of the wall cabinets. The tile pattern also can be extended to the wall underneath your range hood.

Tools & Materials ▸

- Level
- Tape measure
- Pencil
- Tile cutter
- Rod saw
- Notched trowel
- Rubber grout float
- Beating block
- Rubber mallet
- Sponge
- Bucket
- Straight 1 × 2
- Wall tile
- Tile spacers (if needed)
- Bullnose trim tile
- Mastic tile adhesive
- Masking tape
- Grout
- Caulk
- Drop cloth
- Grout sealer

Tips for Planning Tile Layouts ▸

Gather planning brochures and design catalogs to help you create decorative patterns and borders for the backsplash.

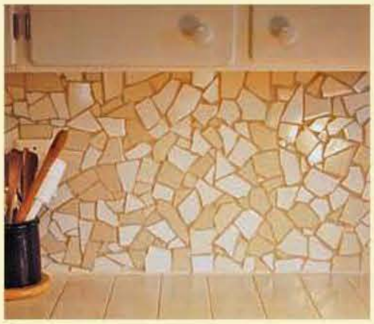

Break tiles into fragments and make a mosaic backsplash. Always use a sanded grout for joints wider than ⅛".

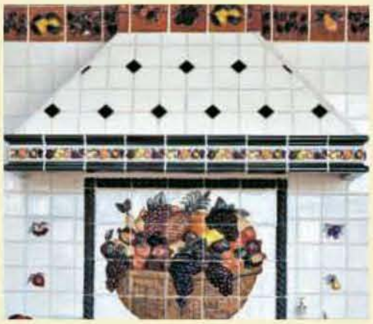

Add painted mural tiles to create a focal point. Mixing various tile styles adds an appealing contrast.

How to Tile a Backsplash

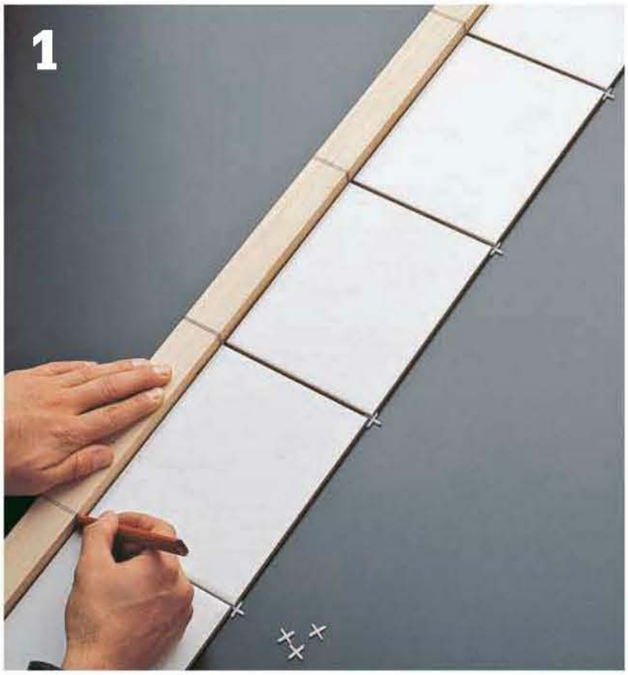

Make a story stick by marking a board at least half as long as the backsplash area to match the tile spacing.

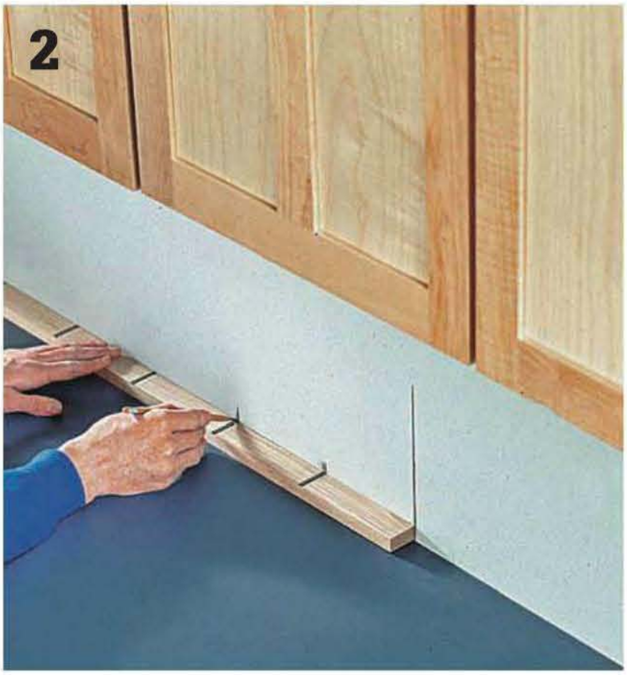

Starting at the midpoint of the installation area, use the story stick to make layout marks along the wall. If an end piece is too small (less than half a tile), adjust the midpoint to give you larger, more attractive end pieces. Use a level to mark this point with a vertical reference line.

While it may appear straight, your countertop may not be level and therefore is not a reliable reference line. Run a level along the counter to find the lowest point on the countertop. Mark a point two tiles up from the low point and extend a level line across the entire work area.

(continued)

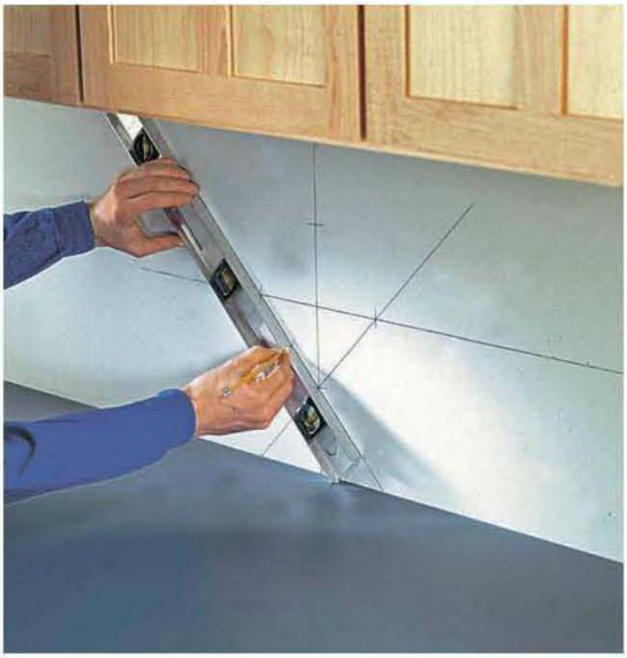

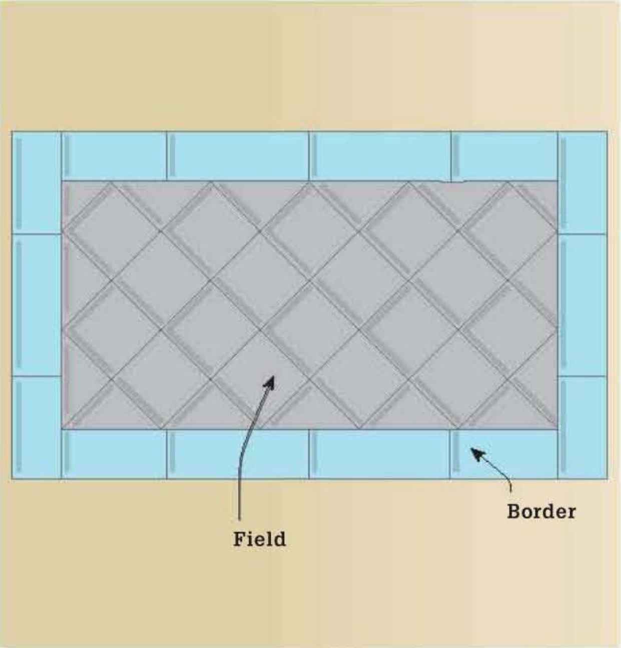

Variation: Diagonal Layout. Mark vertical and horizontal reference lines, making sure the angle is 90°. To establish diagonal layout lines, measure out equal distances from the crosspoint, then connect the points with a line. Additional layout lines can be extended from these as needed. To avoid the numerous, unattractive perimeter cuts common to diagonal layouts, try using a standard border pattern as shown. Diagonally set a field of full tiles, then cut enough half tiles to fill out the perimeter. Finally, border the diagonal field with tiles set square to the field.

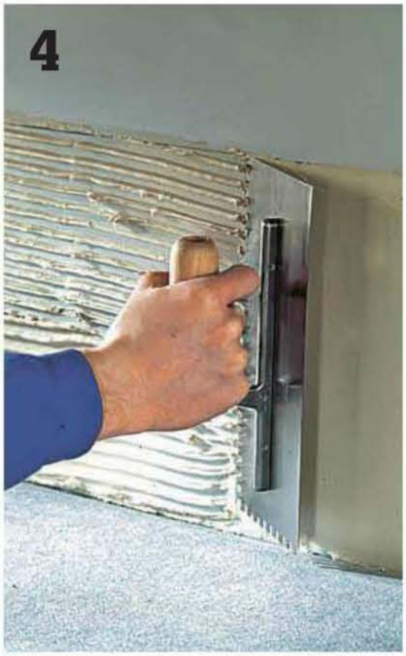

Apply mastic adhesive evenly to the area beneath the horizontal reference line, using a notched trowel. Comb the adhesive horizontally with the notched edge.

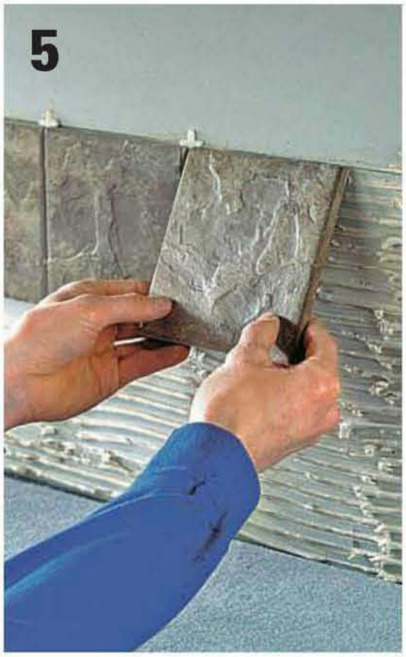

Starting at the vertical reference line, press tiles into the adhesive using a slight twisting motion. If the tiles are not self-spacing, use plastic spacers to maintain even grout lines. If the tiles do not hang in place, use masking tape to hold them in place until the adhesive sets.

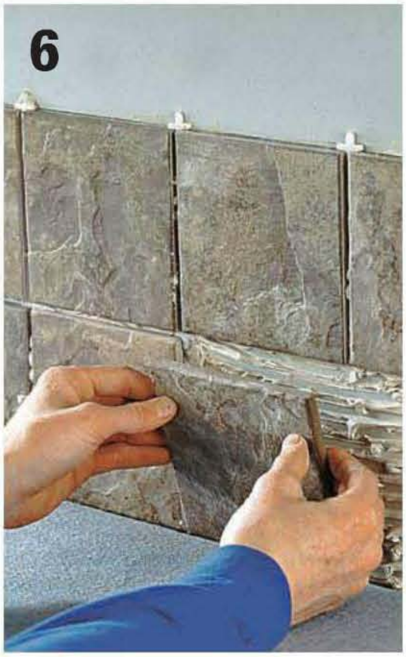

Install a whole row along the reference line, checking occasionally to make sure the tiles are level. Continue installing tiles below the first row, trimming tiles that butt against the countertop as needed.

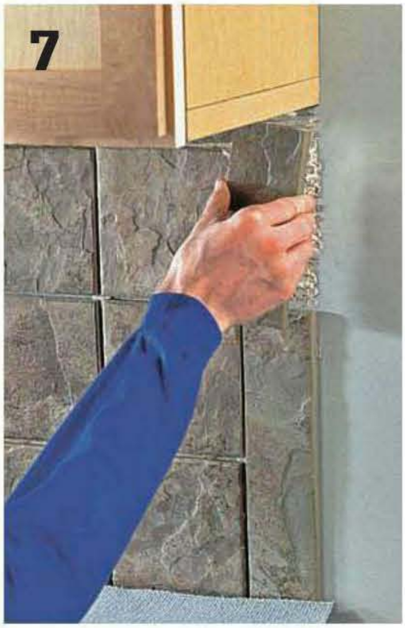

Apply adhesive to an area above the line and continue placing tiles, working from the center to the sides. Install trim tile, such as bullnose tile, to the edges of the rows.

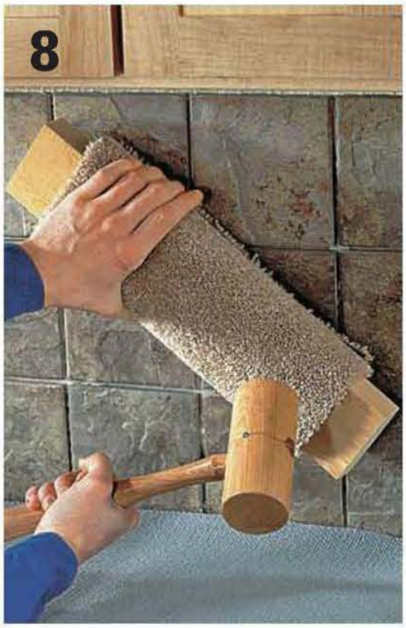

When the tiles are in place, make sure they are flat and firmly embedded by laying a beating block against the tile and rapping it lightly with a mallet. Remove the spacers. Allow the mastic to dry for at least 24 hours, or as directed by the manufacturer.

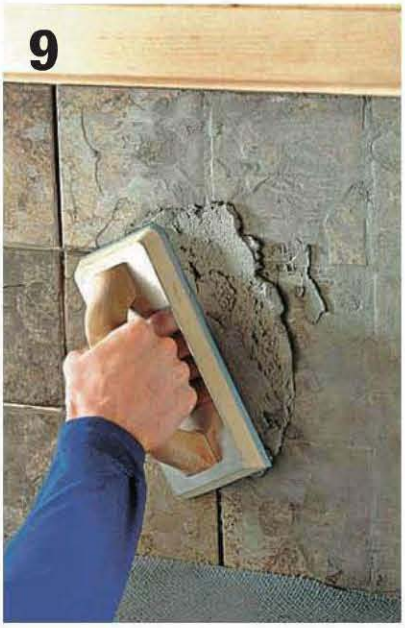

Mix the grout and apply it with a rubber grout float. Spread it over the tiles, keeping the float at a low 30° angle, pressing the grout deep into the joints. *Note: For grout joints ⅛" and smaller, be sure to use a non-sanded grout.*

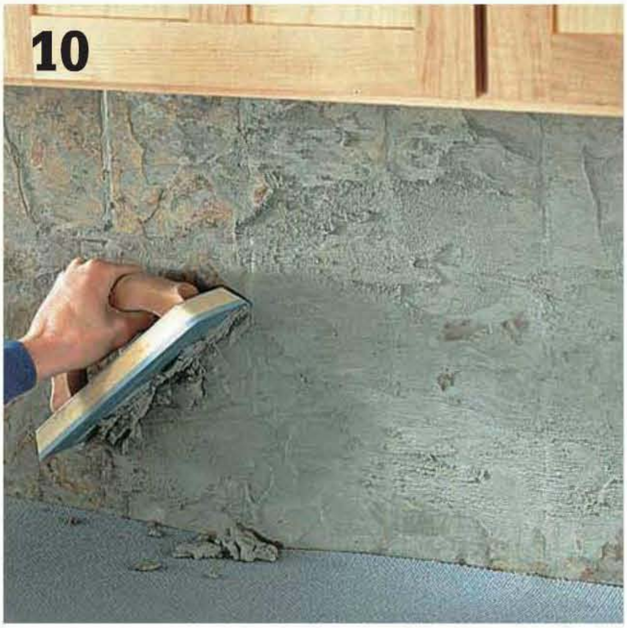

Wipe off excess grout, holding the float at a right angle to the tile, working diagonally so as not to remove grout from the joints. Clean any remaining grout from the tiles with a damp sponge, working in a circular motion. Rinse the sponge thoroughly and often.

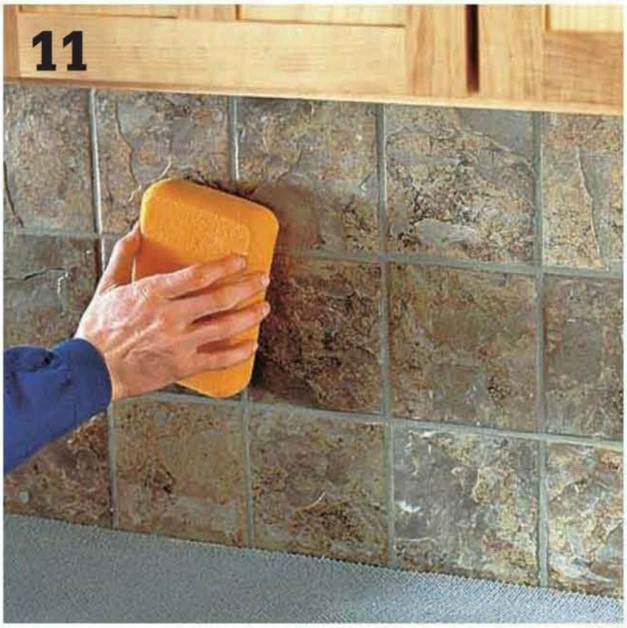

Shape the grout joints by making slow, short passes with the sponge, shaving down any high spots; rinse the sponge frequently. Fill any voids with a fingerful of grout. When the grout has dried to a haze, buff the tile clean with a soft cloth. Apply a bead of caulk between the countertop and tiles. Reinstall any electrical fixtures you removed. After the grout has completely cured, apply grout sealer.

Installing Kitchen Sinks

Most drop-in, self-rimming kitchen sinks are easily installed.

Drop-in sinks for do-it-yourself installation are made from cast iron coated with enamel, stainless steel, enameled steel, acrylic, fiberglass or resin composites. Because cast-iron sinks are heavy, their weight holds them in place and they require no mounting hardware. Except for the heavy lifting, they are easy to install. Stainless steel and enameled-steel sinks weigh less than cast-iron and most require mounting brackets on the underside of the countertop. Some acrylic and resin sinks rely on silicone caulk to hold them in place.

If you are replacing a sink but not the countertop, make sure the new sink is the same size or larger. All old silicone caulk residue must be removed with acetone or denatured alcohol, or the new caulk will not stick.

Shopping Tips ▸

- When purchasing a sink you also need to buy strainer bodies and baskets, sink clips and a drain trap kit.
- Look for basin dividers that are lower than the sink rim—this reduces splashing.
- Drain holes in the back or to the side leave more usable space under the sink.
- When choosing a sink, make sure the predrilled openings will fit your faucet.

Tools & Materials ▸

Caulk gun
Spud wrench
Screwdriver
Sink
Sink frame
Plumber's putty
or silicone caulk
Mounting clips
Jigsaw
Pen or pencil

Drop-in sinks, also known as self-rimming sinks, have a wide sink flange that extends beyond the edges of the sink cutout. They also have a wide back flange to which the faucet is mounted directly.

How to Install a Self-rimming Sink

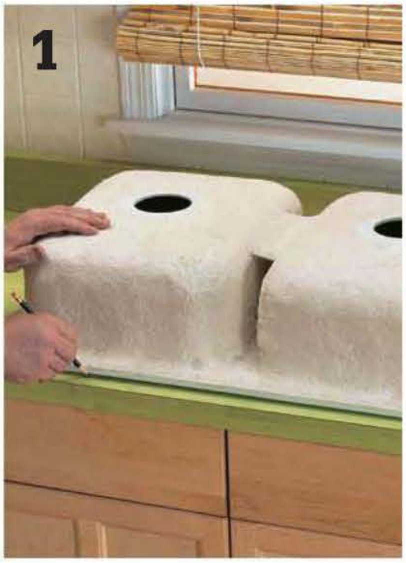

1

Invert the sink and trace around the edges as a reference for making the sink cutout cutting lines, which should be parallel to the outlines, but about 1" inside of them to create a 1" ledge. If your sink comes with a template for the cutout, use it.

2

Drill a starter hole and cut out the sink opening with a jigsaw. Cut right up to the line. Because the sink flange fits over the edges of the cutout, the opening doesn't need to be perfect, but as always you should try to do a nice, neat job.

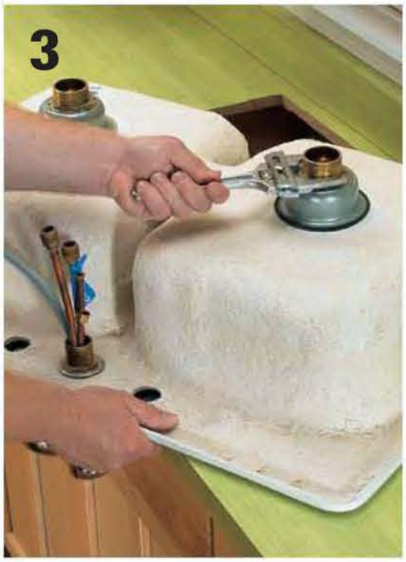

3

Attach as much of the plumbing as makes sense to install prior to setting the sink into the opening. Having access to the underside of the flange is a great help when it comes to attaching the faucet body, sprayer and strainer, in particular.

4

Apply a bead of silicone caulk around the edges of the sink opening. The sink flange most likely is not flat, so try and apply the caulk in the area that will make contact with the flange.

5

Place the sink in the opening. Try and get the sink centered right away so you don't need to move it around and disturb the caulk, which can break the seal. If you are installing a heavy cast-iron sink, it's best to leave the strainers off so you can hold the sink at the drain openings.

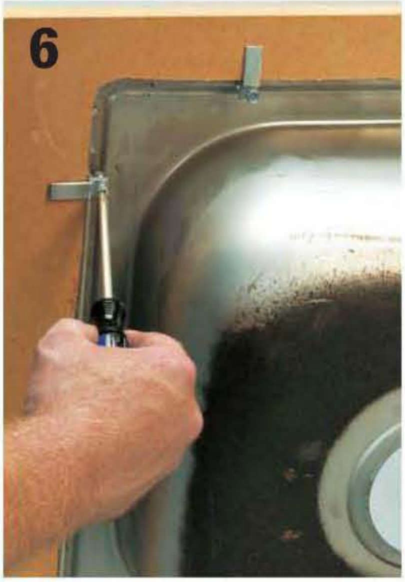

6

For sinks with mounting clips, tighten the clips from below using a screwdriver or wrench (depending on the type of clip your sink has). There should be at least three clips on every side. Don't overtighten the clips—this can cause the sink flange to flatten or become warped.

Replacing Kitchen Faucets

Most new kitchen faucets feature single-handle control levers and washerless designs that rarely require maintenance. Additional features include brushed metallic finishes, detachable spray nozzles, or even push-button controls.

Connect the faucet to hot and cold water lines with easy-to-install flexible supply tubes made from vinyl or braided steel. If your faucet has a separate sprayer, install the sprayer first. Pull the sprayer hose through the sink opening and attach to the faucet body before installing the faucet.

Tools & Materials ▸

- Adjustable wrench
- Basin wrench or channel-type pliers
- Hacksaw
- Faucet
- Putty knife
- Screwdriver
- Silicone caulk
- Scouring pad
- Scouring cleaner
- Plumber's putty
- Flexible vinyl or braided steel supply tubes
- Drain components
- Penetrating oil

Where local codes allow, use plastic tubes for drain hookups. A wide selection of extensions and angle fittings lets you easily plumb any sink configuration. Manufacturers offer kits that contain all the fittings needed for attaching a food disposer or dishwasher to the sink drain system.

Modern kitchen faucets tend to be single-handle models, often with useful features such as a pull-out head that functions as a sprayer. This Price Pfister™ model comes with an optional mounting plate that conceals sink holes when mounted on a predrilled sink flange.

Choosing a New Kitchen Faucet

When choosing a new kitchen faucet, you'll find many options. The best place to start the process is with your sink. In the past, most faucets were mounted directly to the sink deck, which had three or four predrilled holes to accommodate the faucets, spout, sprayer and perhaps a liquid soap dispenser or an air gap for your dishwasher. Modern kitchen faucets don't always conform to this set-up, with many of them designed to be installed in a single hole in the sink deck or in the countertop. If you plan to keep your old sink, look for a faucet that won't leave empty holes in the deck. You can buy snap-in plugs to fill unfilled stainless steel sink holes, or you can add a soap dispenser, but in general you'll have the best luck if you replace like for like.

The two most basic kitchen faucet categories are one-handle and two-handle. One handled models are much more popular now because you can adjust the water temperature easily with just one hand. Another difference is in the faucet body. Some faucets have the taps and the spout mounted onto a faucet body so the spacing between the tailpieces is preset. Others, called widespread faucets, have independent taps and spouts that can be configured however you please, as long as the tubes connecting the taps to the spout reach. This type is best if you are installing the faucet in the countertop (a common way to go about it with new countertops such as solid surface, quartz or granite).

In the past, kitchen faucets almost always had remote pull-out sprayers. The sprayer was attached to the faucet body with a hose directly below the mixing valve. While this type of sprayer is still fairly common, many faucets today have an integral pull-out spout that is very convenient and less prone to failure than the old-style sprayers.

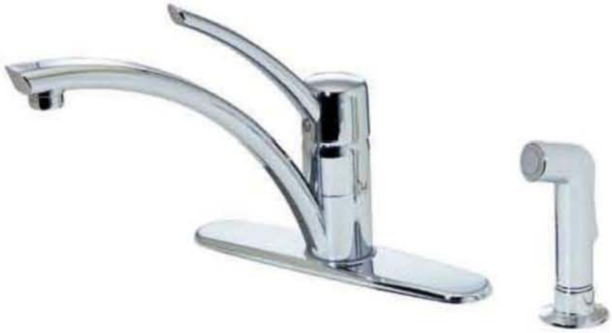

A single-handle, high arc faucet with traditional remote sprayer. The mounting plate is decorative and optional.

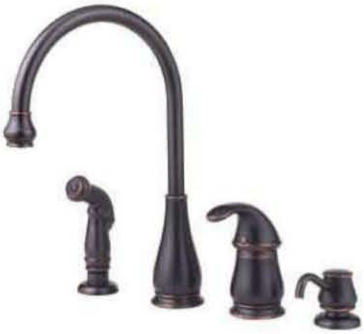

Single-handle faucets may require four holes, as does this model with its side sprayer and matching soap/lotion dispenser.

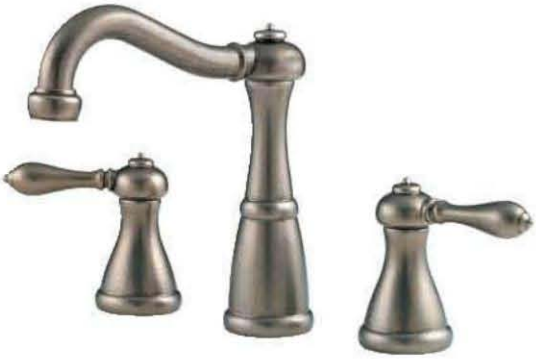

Two-handled faucets are less common, but remain popular choices for traditional kitchens. The gooseneck spout also has a certain elegance, but avoid this type if you have a shallow sink that's less than 8" deep.

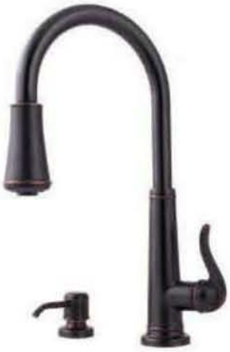

A single-handle faucet with pull-out spray head requires only one hole in your sink deck or countertop—a real benefit if your sink is not predrilled or if it is an undermount model.

How to Remove an Old Faucet

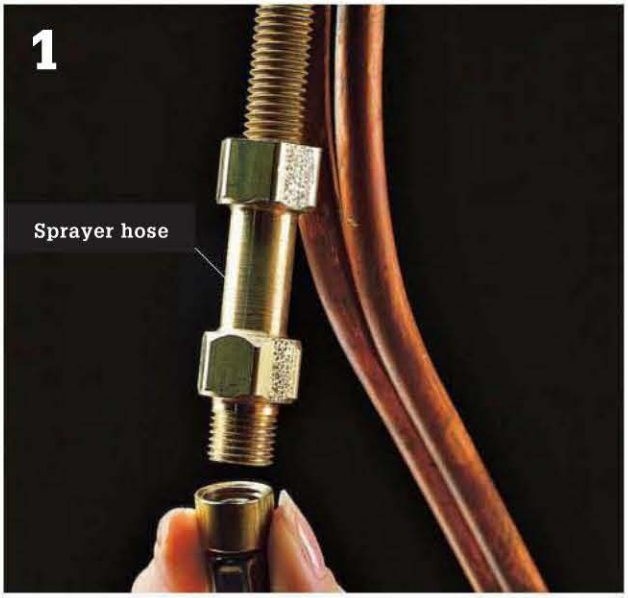

To remove the old faucet, start by clearing out the cabinet under the sink and laying down towels. Turn off the hot and cold stop valves and open the faucet to make sure the water is off. Detach the sprayer hose from the faucet sprayer nipple and unscrew the retaining nut that secures the sprayer base to the sink deck. Pull the sprayer hose out through the sink deck opening.

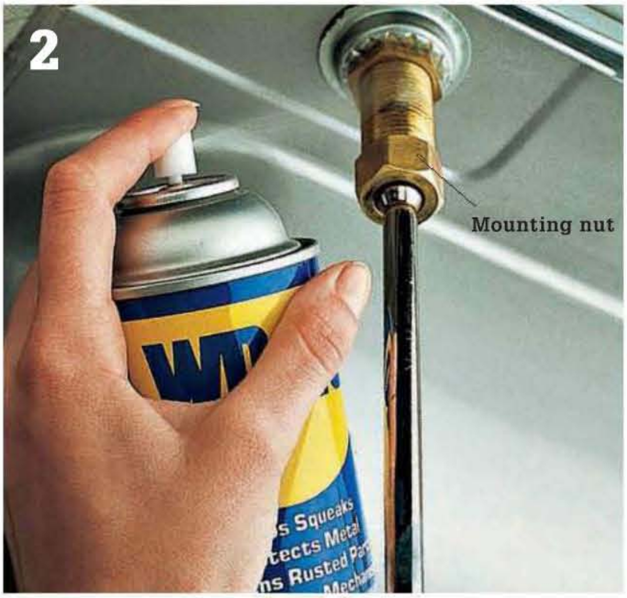

Spray the mounting nuts that hold the faucet or faucet handles (on the underside of the sink deck) with penetrating oil for easier removal. Let the oil soak in for a few minutes.

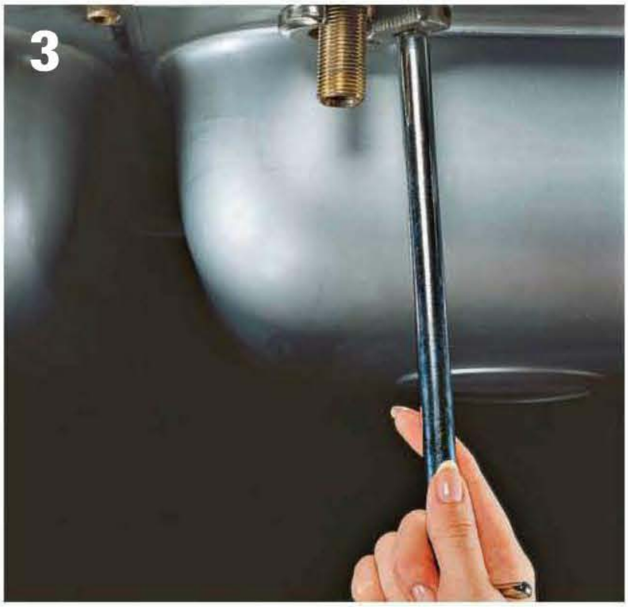

Unhook the supply tubes at the stop valves. Don't reuse old chrome supply tubes. If the stops are missing or unworkable, replace them. Then remove the coupling nuts and the mounting nuts on the tailpieces of the faucet with a basin wrench or channel-type pliers.

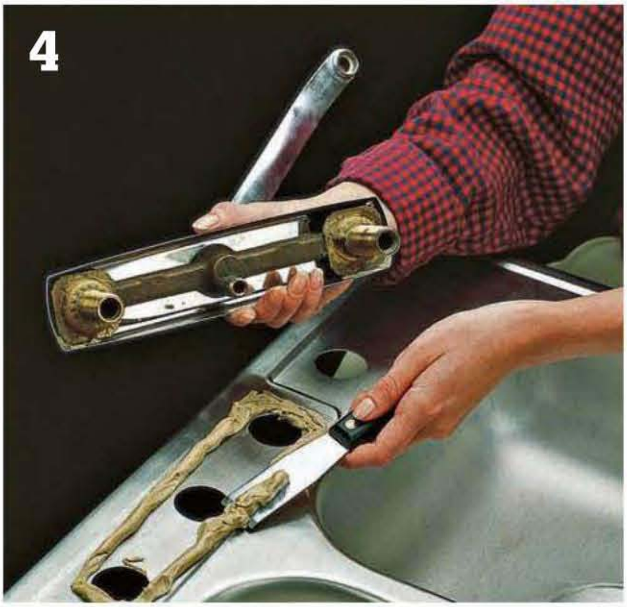

Pull the faucet body from the sink. Remove the sprayer base if you wish to replace it. Scrape off old putty or caulk with a putty knife and clean off the sink with a scouring pad and an acidic scouring cleaner like Bar Keeper's Friend®. *Tip: Scour stainless steel with a back and forth motion to avoid leaving unsightly circular markings.*

How to Install a Kitchen Sink Faucet

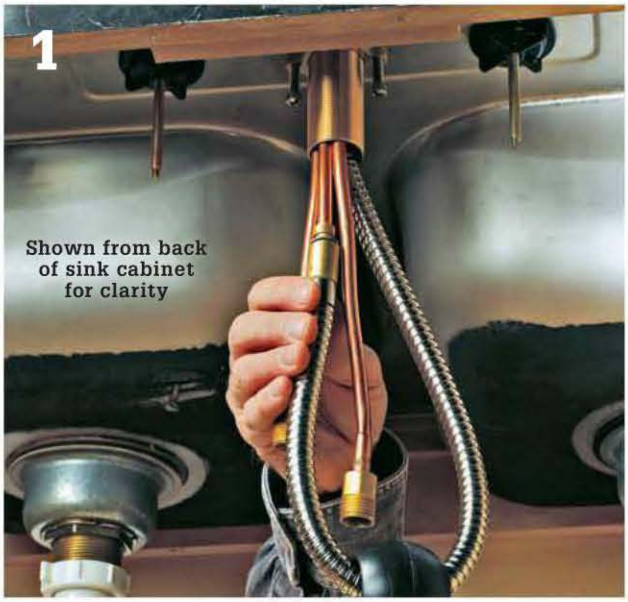

Shut off hot and cold water at the faucet stop valves. Assemble the parts of the deck plate that cover the outer mounting holes in your sink deck (unless you are installing a two-handle faucet, or mounting the faucet directly to the countertop, as in an undermount sink situation). Add a ring of plumber's putty in the groove on the underside of the base plate.

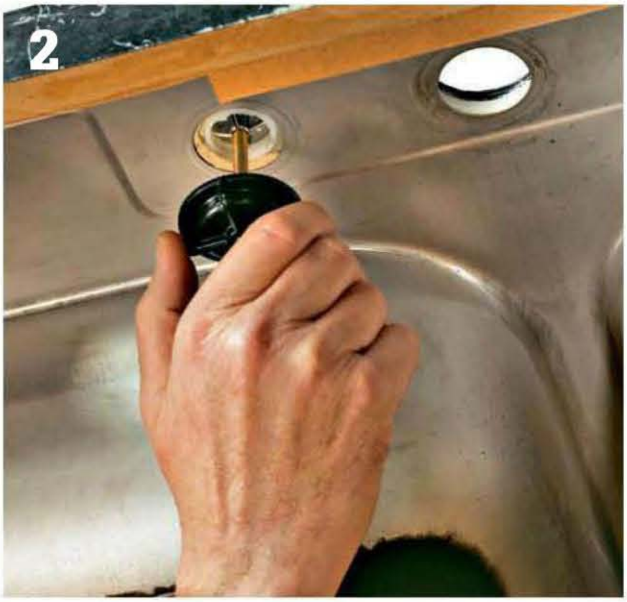

Set the base plate onto the sink flange so it is correctly aligned with the predrilled holes in the flange. From below, tighten the wing nuts that secure the deck plate to the sink deck.

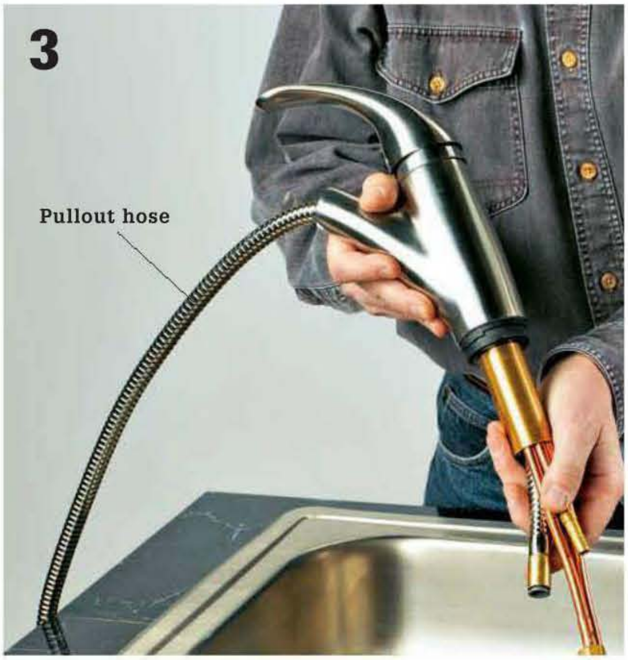

Retract the pullout hose by drawing it out through the faucet body until the fitting at the end of the hose is flush with the bottom of the threaded faucet shank. Insert the shank and the supply tubes down through the top of the deck plate.

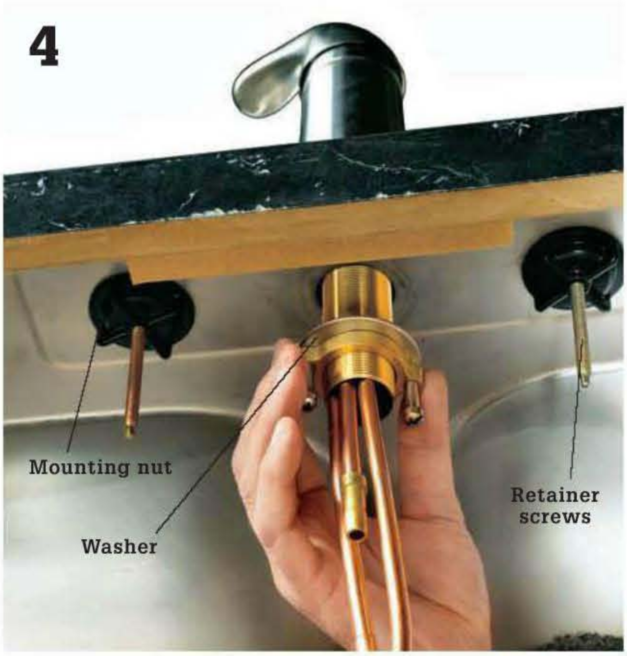

Slip the mounting nut and washer over the free ends of the supply tubes and pullout hose, then thread the nut onto the threaded faucet shank. Hand tighten. Tighten the retainer screws with a screwdriver to secure the faucet.

(continued)

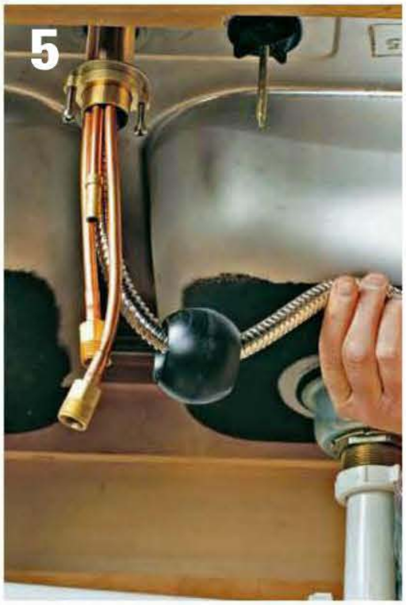

Slide the hose weight onto the pullout hose (the weight helps keep the hose from tangling and makes it easier to retract the hose).

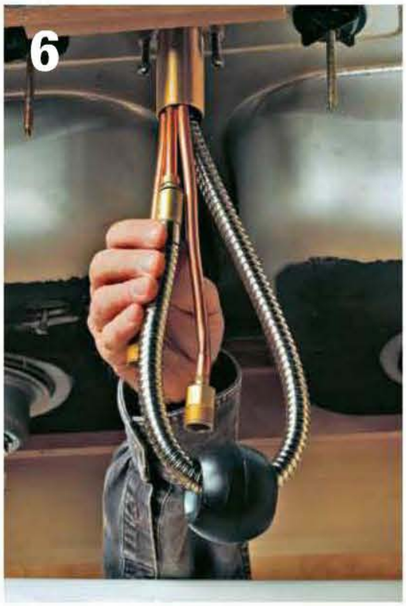

Connect the end of the pullout tube to the outlet port on the faucet body using a quick connector fitting.

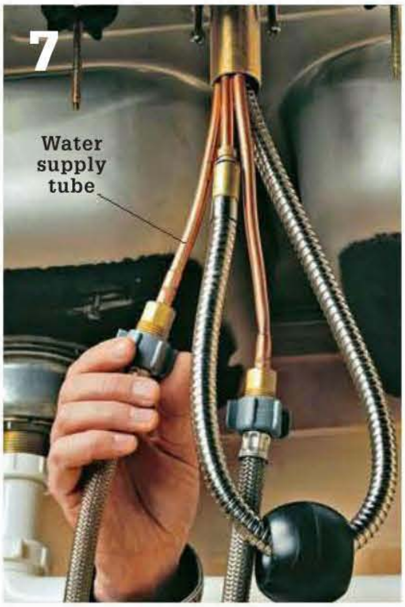

Hook up the water supply tubes to the faucet inlets. Make sure the lines are long enough to reach the supply risers without stretching or kinking.

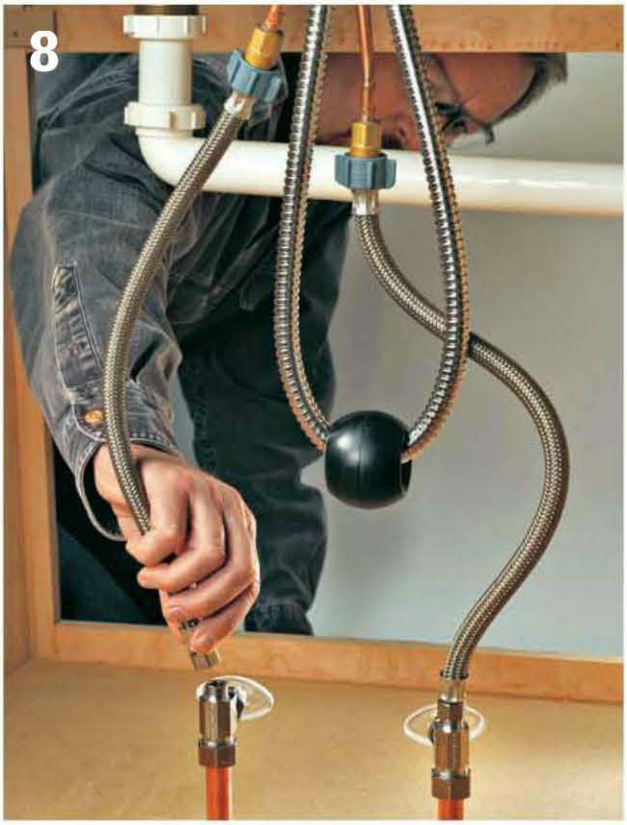

Connect the supply lines to the supply risers at the stop valves. Make sure to get the hot lines and cold lines attached correctly.

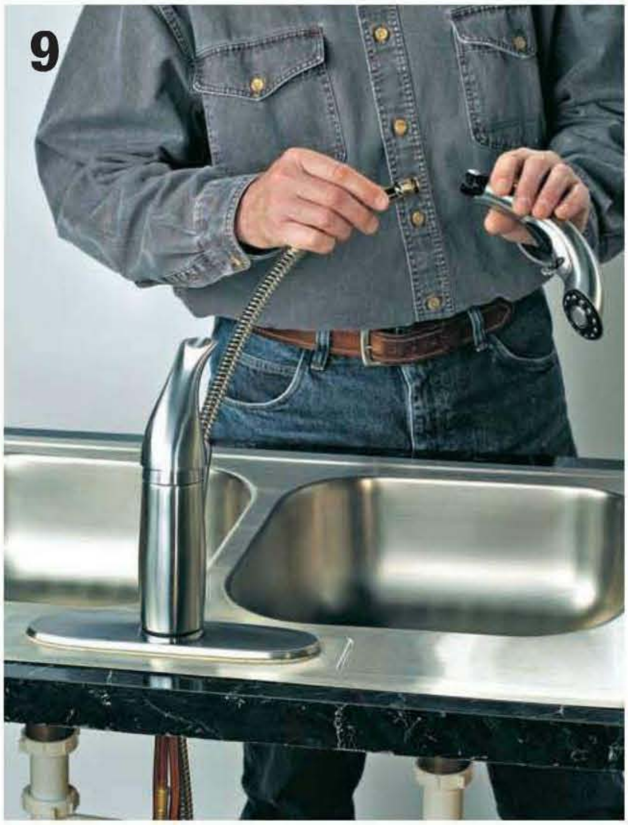

Attach the spray head to the end of the pullout hose and turn the fitting to secure the connection. Turn on water supply and test. *Tip: Remove the aerator in the tip of the spray head and run hot and cold water to flush out any debris.*

Variation: One-piece Faucet with Sprayer ▸

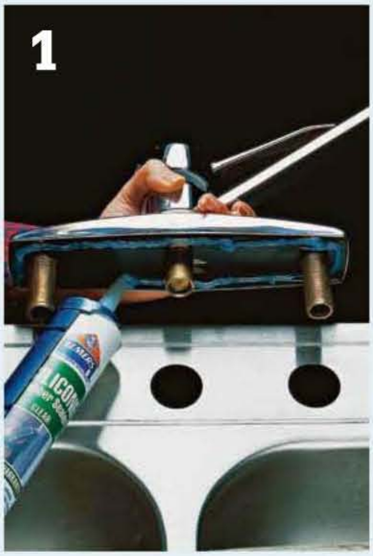

Apply a thick bead of silicone caulk to the underside of the faucet base, then insert the tailpieces of the faucet through the appropriate holes in the sink deck. Press down lightly on the faucet to set it in the caulk.

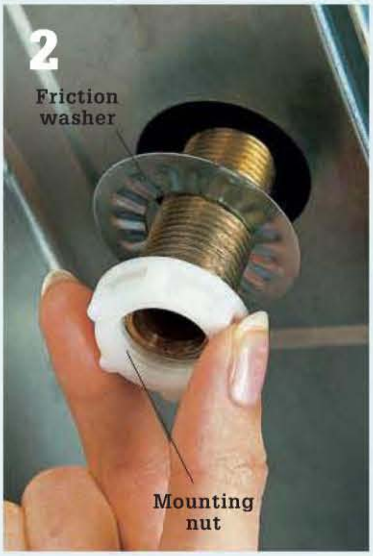

Slip a friction washer onto each tailpiece and then hand-tighten a mounting nut. Tighten the mounting nut with channel-type pliers or a basin wrench. Wipe up any silicone squeeze-out on the sink deck with a wet rag before it sets up.

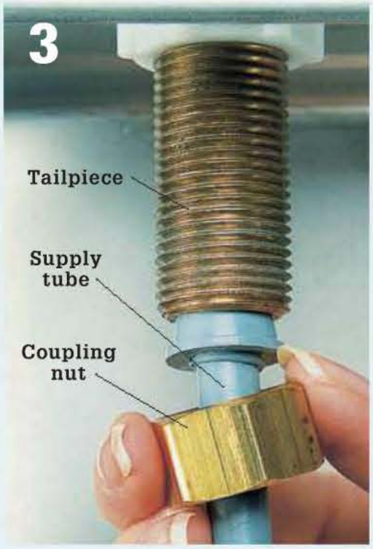

Connect supply tubes to the faucet tailpieces—make sure the tubes you buy are long enough to reach the stop valves and that the coupling nuts will fit the tubes and tailpieces.

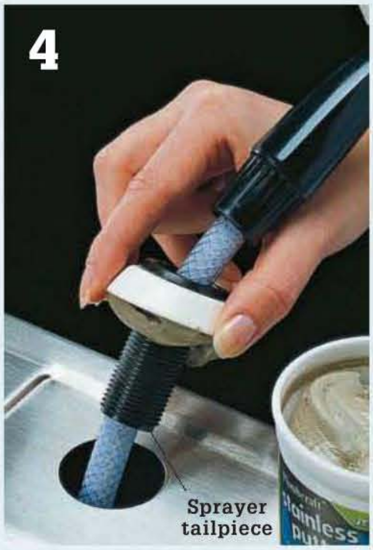

Apply a ¼" bead of plumber's putty or silicone caulk to the underside of the sprayer base. With the base threaded onto the sprayer hose, insert the tailpiece of the sprayer through the opening in the sink deck.

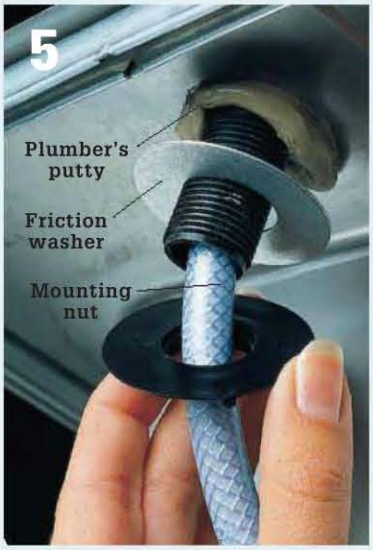

From beneath, slip the friction washer over the sprayer tailpiece and then screw the mounting nut onto the tailpiece. Tighten with channel-type pliers or a basin wrench. Wipe any excess putty or caulk on the sink deck from around the base.

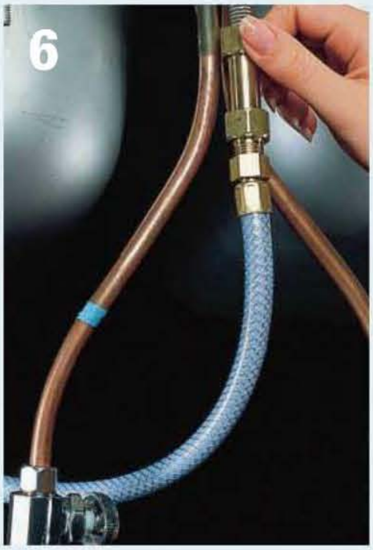

Screw the sprayer hose onto the hose nipple on the bottom of the faucet. Hand-tighten and then give the nut one quarter turn with pliers or a basin wrench. Turn on the water supply at the shutoff, remove the aerator and flush debris from the faucet.

Connecting Drains

Kitchen sink drains don't last forever, but they're very easy and inexpensive to replace. The most common models today are made of PVC plastic pipe and fittings held together with slip fittings. In addition to making the installation fairly forgiving, the slip fitting makes the drain easy to disassemble if it gets clogged. The project shown here is a bit unusual by today's standards, in that it does not include either a dishwasher drain or a garbage disposer. But you will see how to add each of these drain systems to your kitchen sink in the following two sections.

You can buy the parts for the kitchen drain individually (you can usually get better quality materials this way) or in a kit (see photo, next page). Because most kitchen sinks have two bowls, the kits include parts for plumbing both drains into a shared trap, often with a baffle in the T-fitting where the outlet line joins with the tailpiece from the other bowl. If you are installing a disposer, consider installing individual traps to eliminate the baffle, which reduces the flow capacity by half.

Tools & Materials ▸

Flat screwdriver
Spud wrench
Trap arm
Mineral spirits
Cloth
Strainer kit
Plumber's putty
Teflon tape
Washers
Waste-T fitting
S- or P-trap

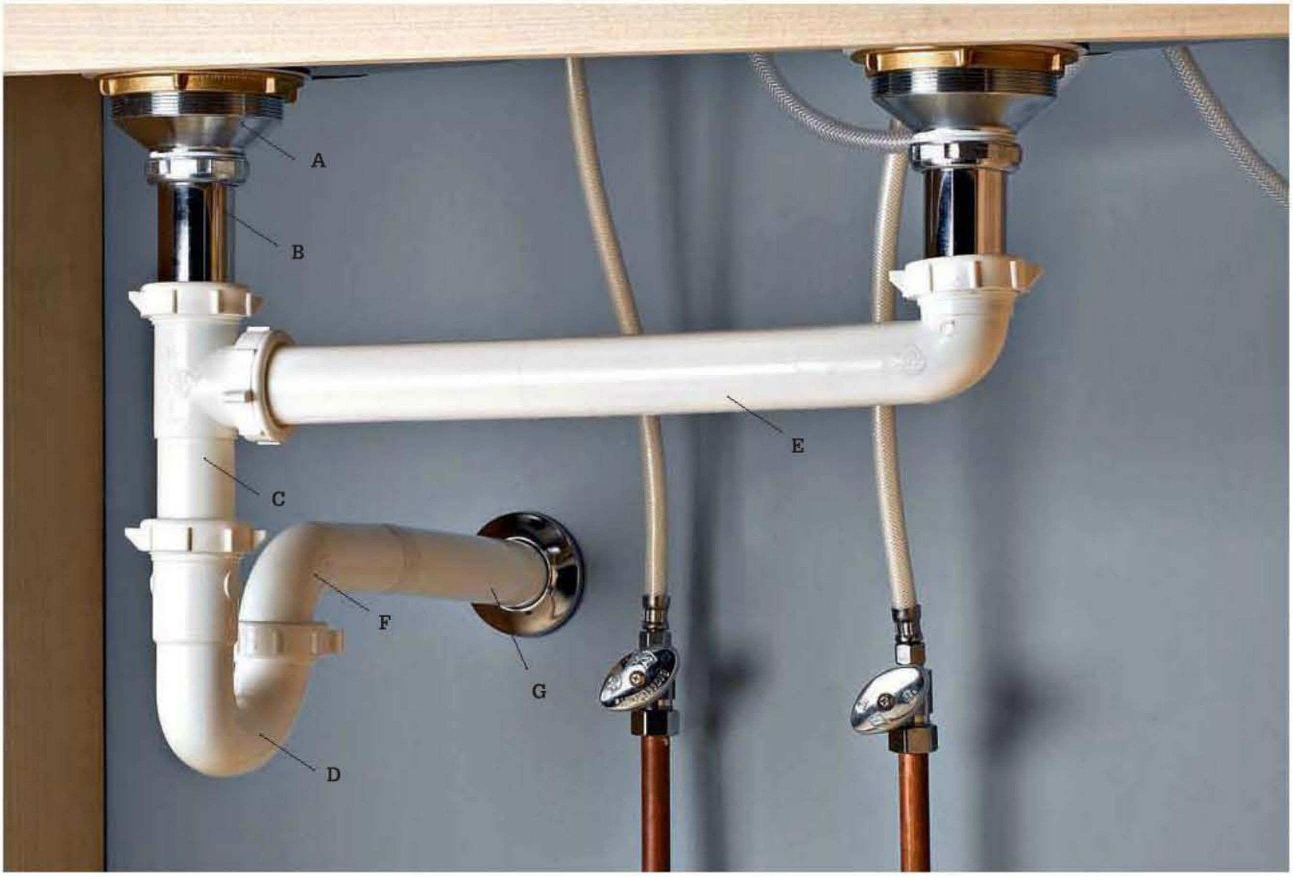

Kitchen sink drains include a strainer basket (A), tailpiece (B), continuous waste T (C), P- or S-trap (D), outlet drain lines (E), trap arm (F), and wall stubout (G).

Drain Kits ▸

Kits for installing a new sink drain include all the pipes, slip fittings, and washers you'll need to get from the sink tailpieces (most kits are equipped for a double bowl kitchen sink) to the trap arm that enters the wall or floor. For wall trap arms, you'll need a kit with a P-trap. For floor drains, you'll need an S-trap. Both drains normally are plumbed to share a trap. Chromed brass or PVC with slip fittings let you adjust the drain more easily and pull it apart and then reassemble if there is a clog. Kitchen sink drains and traps should be 1½" o.d. pipe—the 1¼" pipe is for lavatories and doesn't have enough capacity for a kitchen sink.

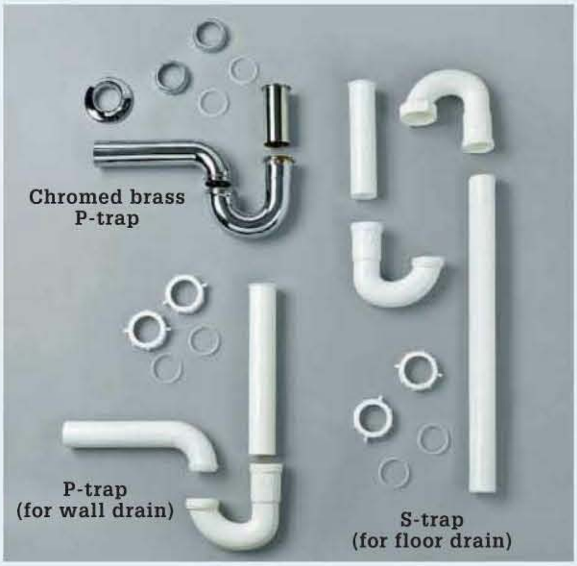

Tips for Choosing Drains ▸

Wall thickness varies in sink drain pipes. The thinner plastic material is cheaper and more difficult to obtain a good seal with than the thicker, more expensive tubing. The thin product is best reserved for lavatory drains, which are far less demanding.

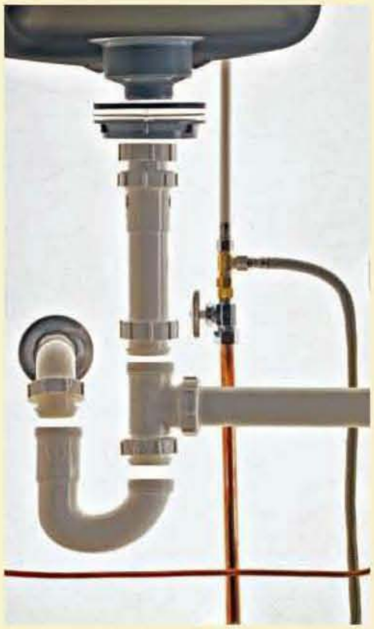

Slip joints are formed by tightening a male-threaded slip nut over a female-threaded fitting, trapping and compressing a beveled nylon washer to seal the joint.

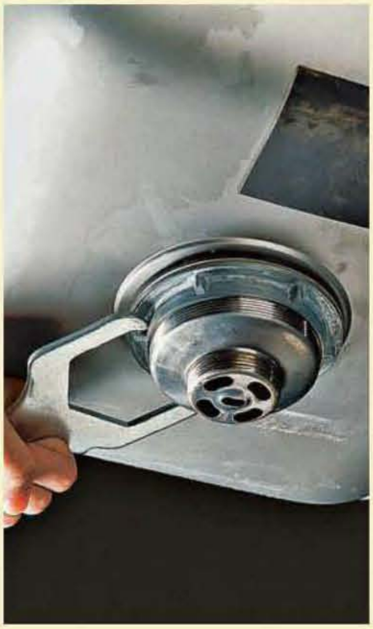

Use a spud wrench to tighten the strainer body against the underside of the sink bowl. Normally, the strainer flange has a layer of plumber's putty between it and the sink drain, and a pair of washers (one rubber, one fibrous) below.

How to Hook Up a Kitchen Sink Drain

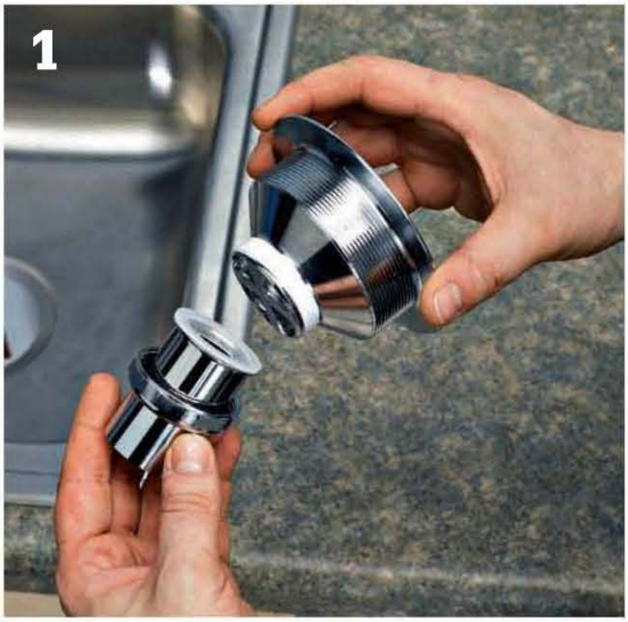

If you are replacing the sink strainer body, remove the old one and clean the top and bottom of the sink deck around the drain opening with mineral spirits. Attach the drain tailpiece to the threaded outlet of the strainer body, inserting a nonbeveled washer between the parts, if your strainer kits include one. Lubricate the threads or apply Teflon tape so you can get a good, snug fit.

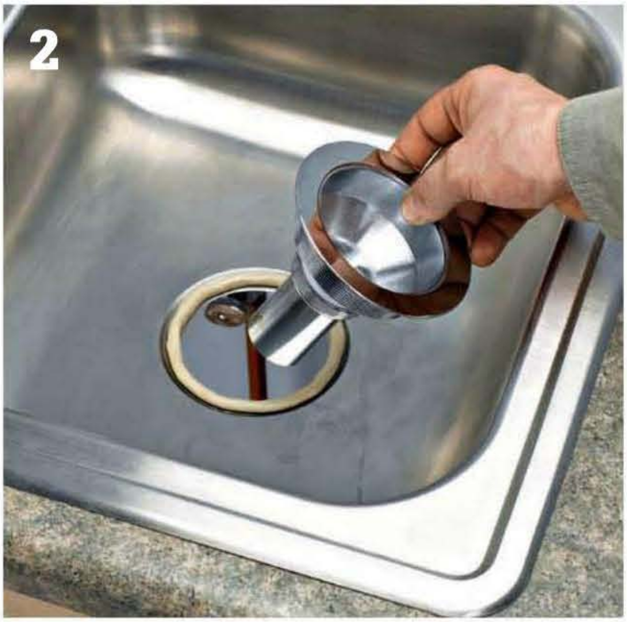

Apply plumber's putty around the perimeter of the drain opening and seat the strainer assembly into it. Add washers below as directed and tighten the strainer locknut using a spud wrench (see photo, previous page) or striking the mounting nubs at the top of the body with a flat screwdriver.

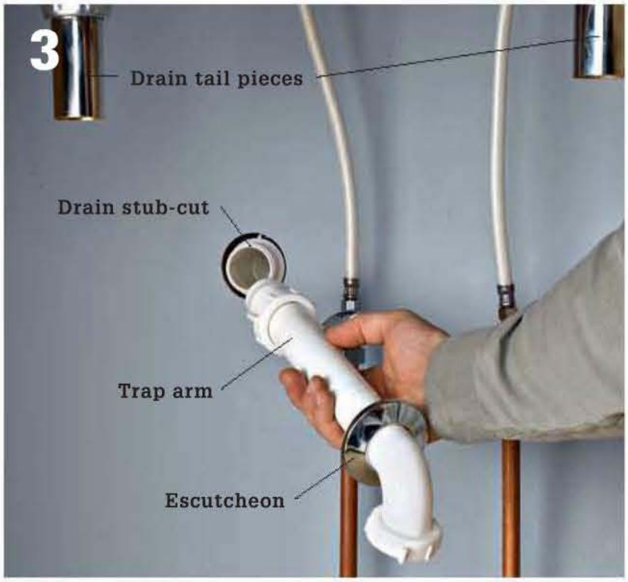

Attach the trap arm to the male-threaded drain stubout in the wall, using a slip nut and beveled compression washer. The outlet for the trap arm should point downward. *Note: The trap arm must be higher on the wall than any of the horizontal lines in the set-up, including lines to dishwasher, disposer, or the outlet line to the second sink bowl.*

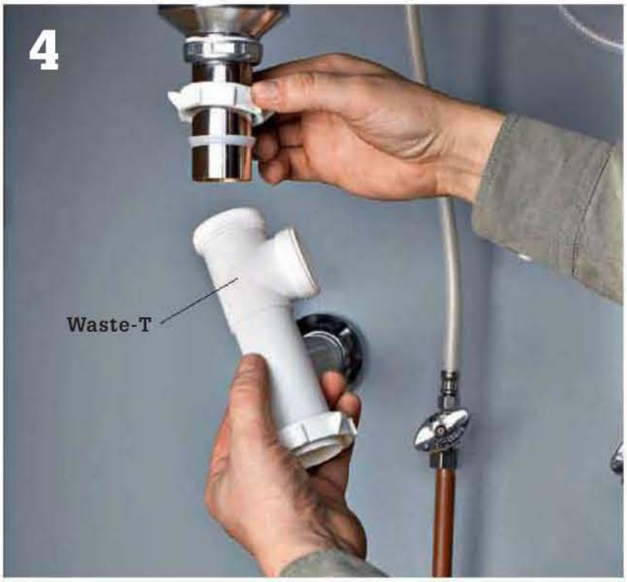

Attach a waste-T-fitting to the drain tailpiece, orienting the opening in the fitting side so it will accept the outlet drain line from the other sink bowl. If the waste-T is higher than the top of the trap arm, remove it and trim the drain tailpiece.

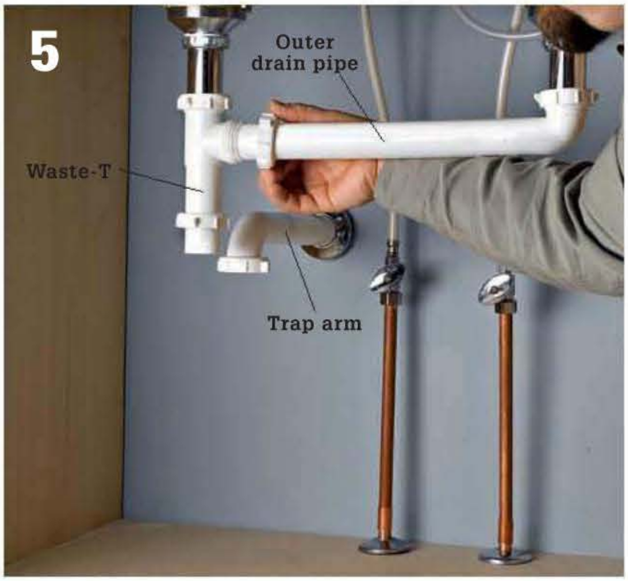
5
Outer drain pipe
Waste-T
Trap arm

Joint the short end of the outlet drain pipe to the tailpiece for the other sink bowl and then attach the end of the long run to the opening in the waste-T. The outlet tube should extend into the T ½" or so—make sure it does not extend far enough to block water flow from above.

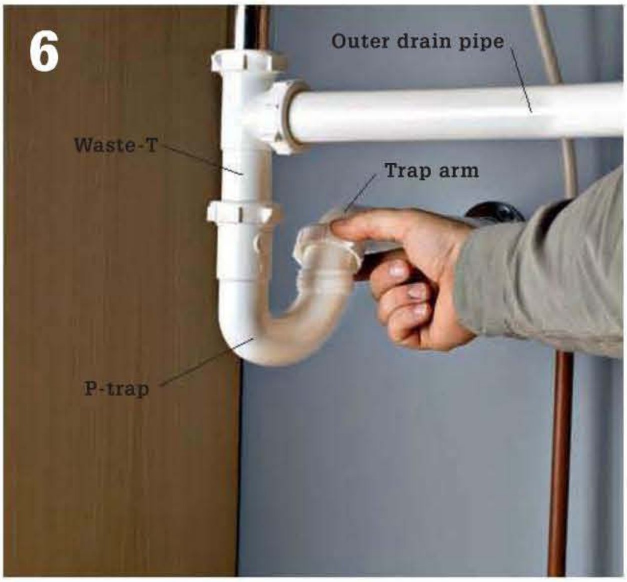
6
Outer drain pipe
Waste-T
Trap arm
P-trap

Attach the long leg of a P-trap to the waste-T and attach the shorter leg to the downward-facing opening of the trap arm. Adjust as necessary and test all joints to make sure they are still tight, and then test the system.

Variation: Drain in Floor ▸

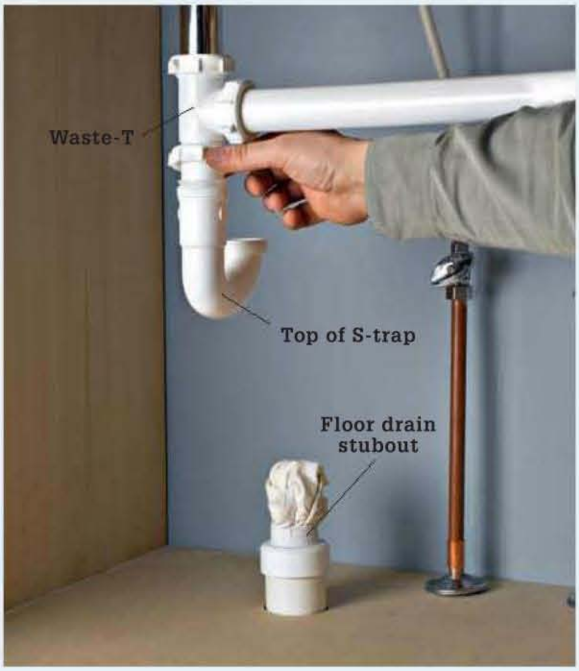
Waste-T
Top of S-trap
Floor drain stubout

If your drain stubout comes up out of the floor instead of the wall, you'll need an S-trap to tie into it instead of a P-trap. Attach one half of the S-trap to the threaded bottom of the waste-T.

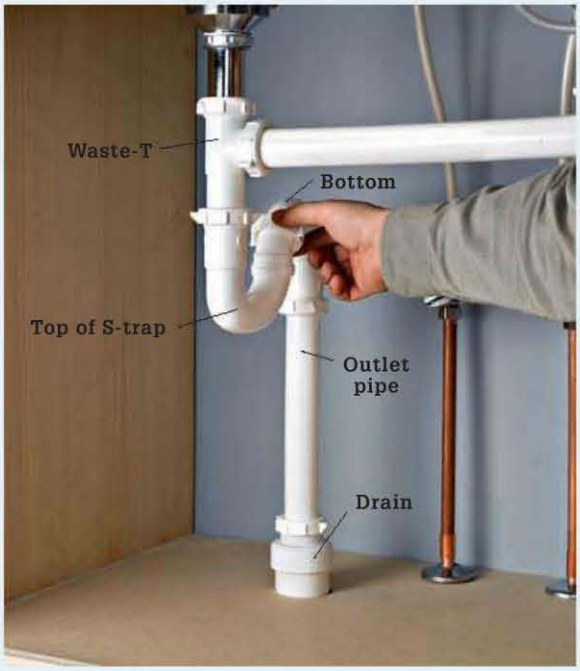
Waste-T
Bottom
Top of S-trap
Outlet pipe
Drain

Attach the other half of the S-trap to the stubout using a slip fitting. This should result in the new fitting facing downward. Join the halves of the S-trap with a slip nut, trimming the unthreaded end if necessary.

Hooking Up Dishwashers

A dishwasher that's past its prime may be inefficient in more ways than one. If it's an old model, it probably wasn't designed to be very efficient to begin with. But more significantly, if it no longer cleans effectively, you're probably spending a lot of time and hot water pre-rinsing the dishes. This alone can consume more energy and water than a complete wash cycle on a newer machine. Even if your old dishwasher still runs, replacing it with an efficient new model can be a good green upgrade.

In terms of sizing and utility hookups, dishwashers are generally quite standard. If your old machine is a built-in and your countertops and cabinets are standard sizes, most full-size dishwashers will fit right in. Of course, you should always measure the dimensions of the old unit before shopping for a new one to avoid an unpleasant surprise at installation time. Also, be sure to review the manufacturer's instructions before starting any work.

Tools & Materials ▸

Screwdrivers
Adjustable wrench
2-ft. level
⅝" automotive heater hose
Automotive heater hose
4"-length of ½" copper tubing
Cable connector
Teflon tape
Hose clamps
Wire connectors
Carpet scrap
Bowl

Replacing an old, inefficient dishwasher is a straightforward project that usually takes just a few hours. The energy savings begin with the first load of dishes and continue with every load thereafter.

How to Replace a Dishwasher

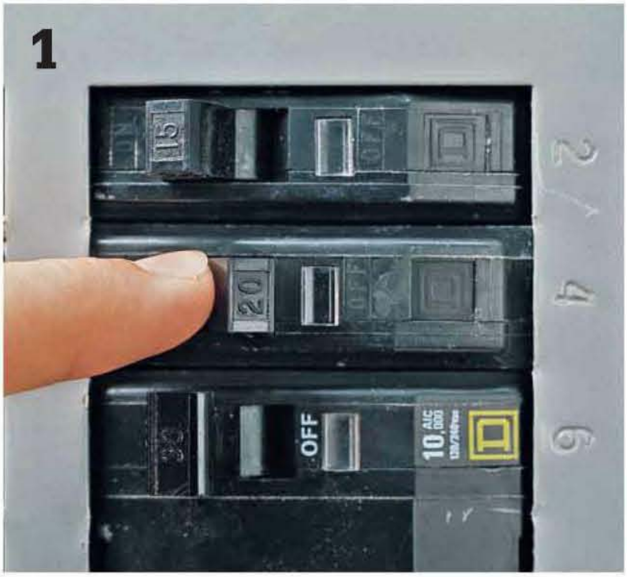

Start by shutting off the electrical power to the dishwasher circuit at the service panel. Also, turn off the water supply at the shutoff valve, usually located directly under the floor.

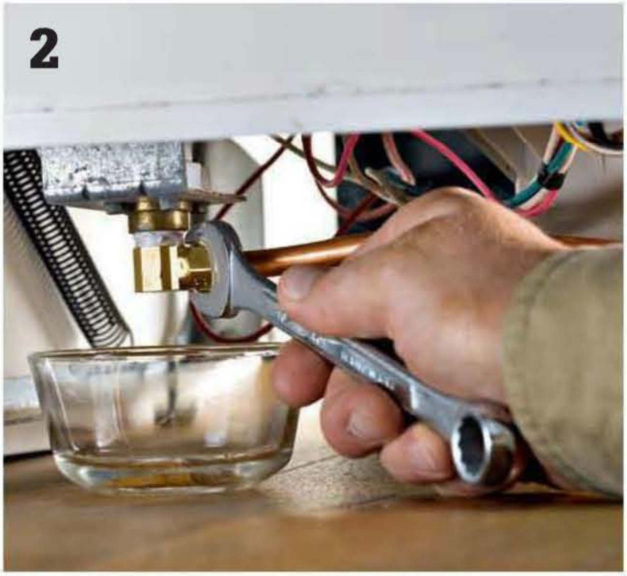

Disconnect old plumbing connections. First unscrew the front access panel. Once the access panel is removed, disconnect the water supply line from the L-fitting on the bottom of the unit. This is usually a brass compression fitting, so just turning the compression nut counterclockwise with an adjustable wrench should do the trick. Use a bowl to catch any water that might leak out when the nut is removed.

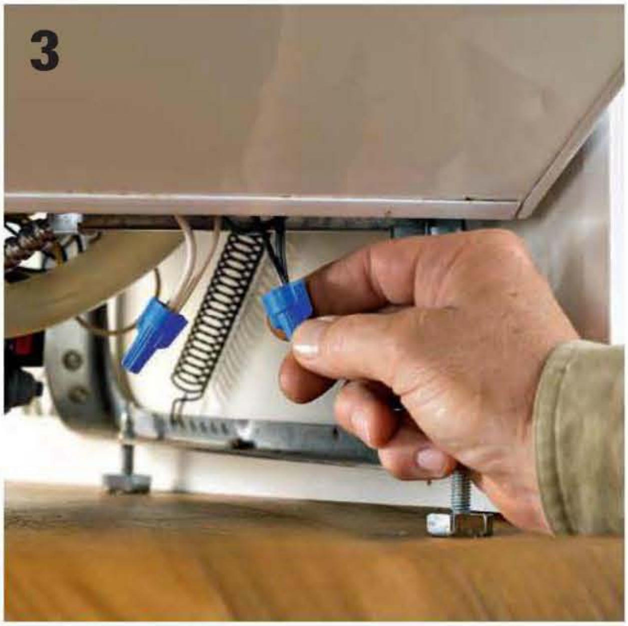

Disconnect old wiring connections. The dishwasher has an integral electrical box at the front of the unit where the power cable is attached to the dishwasher's fixture wires. Take off the box cover and remove the wire connectors that join the wires together.

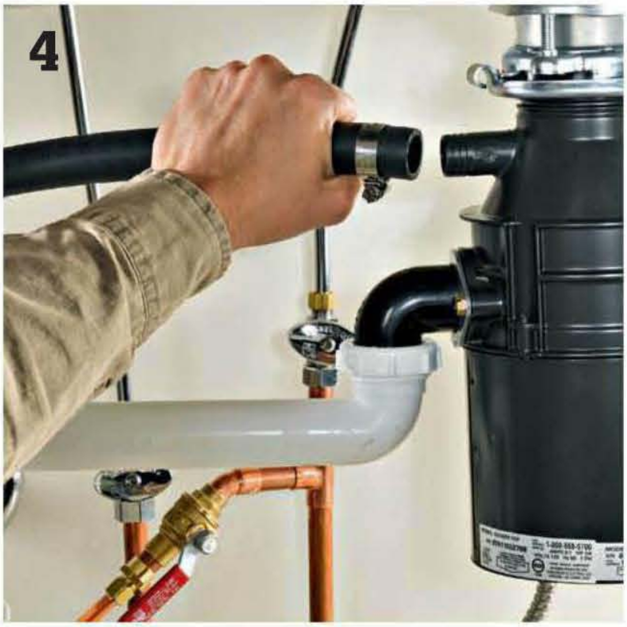

Disconnect the discharge hose, which is usually connected to the dishwasher port on the side of the garbage disposer. To remove it, just loosen the screw on the hose clamp and pull it off. You may need to push this hose back through a hole in the cabinet wall and into the dishwasher compartment so it won't get caught when you pull the dishwasher out.

(continued)

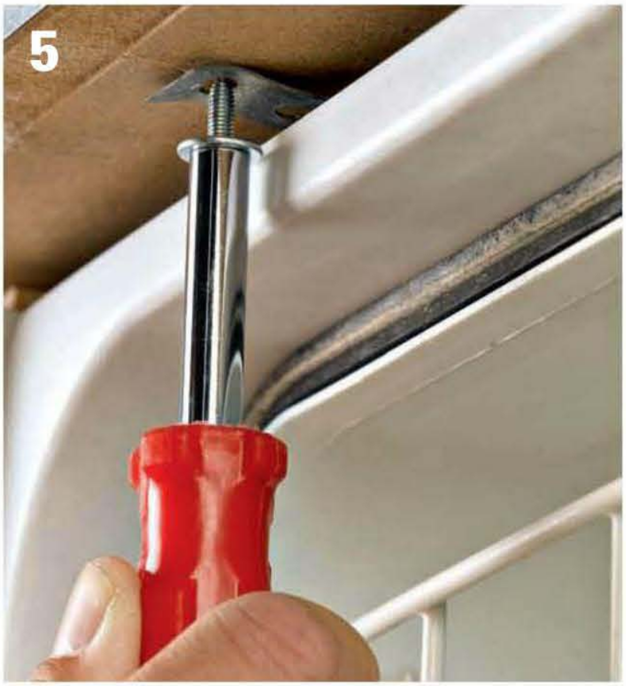

Detach the unit from surrounding cabinets before you pull it out. Remove the screws that hold the brackets to the underside of the countertop. Then put a piece of cardboard or old carpet under the front legs to protect the floor from getting scratched, and pull the dishwasher out.

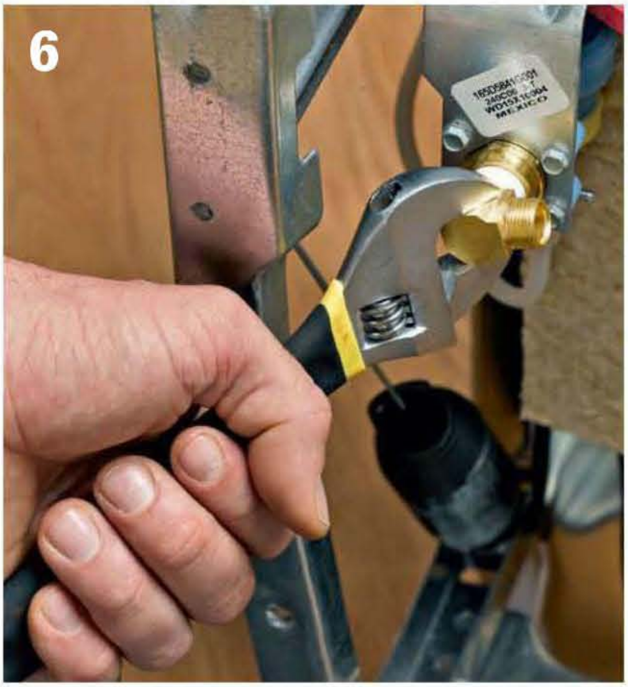

Prepare the new dishwasher. Tip it on its back and attach the new L-fitting into the threaded port on the solenoid. Apply some Teflon tape or pipe sealant to the fitting threads before tightening it, to prevent possible leaks.

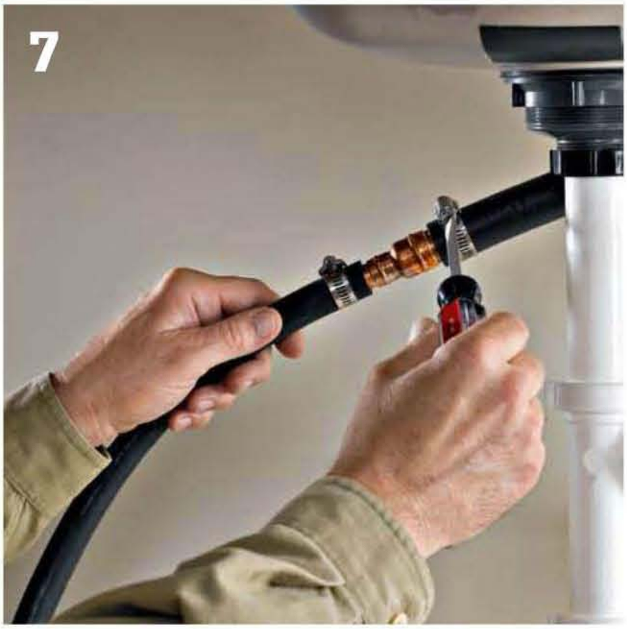

Attach a length of new automotive heater hose, usually ⅝" diameter, to the end of the dishwasher's discharge hose nipple with a hose clamp. The new hose you are adding should be long enough to reach from the discharge nipple to the port on the side of the kitchen sink garbage disposer.

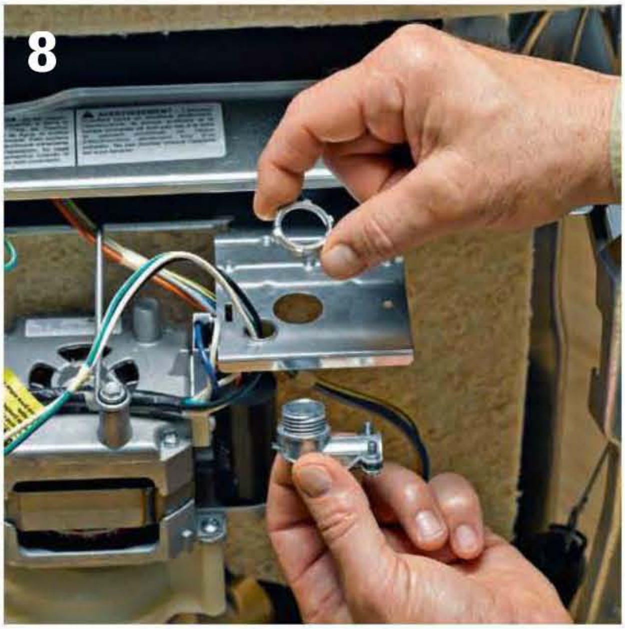

Prepare for the wiring connections. Like the old dishwasher, the new one will have an integral electrical box for making the wiring connections. To gain access to the box, just remove the box cover. Then install a cable connector on the back of the box and bring the power cable from the service panel through this connector. Power should be shut off at the main service panel at all times.

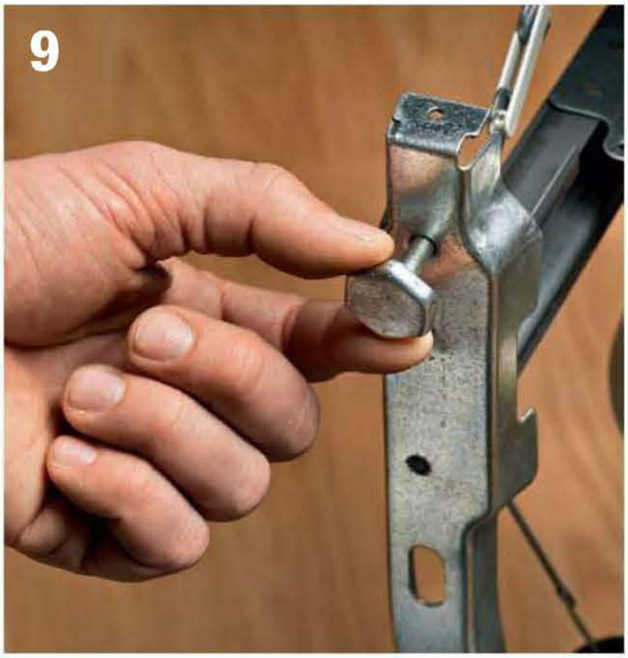

Install a leveling leg at each of the four corners while the new dishwasher is still on its back. Turn the legs into the threaded holes designed for them. Leave about ½" of each leg projecting from the bottom of the unit. (These will have to be adjusted later to level the appliance.) Tip the appliance up onto the feet and slide it into the opening. Check for level in both directions, and adjust the feet as required.

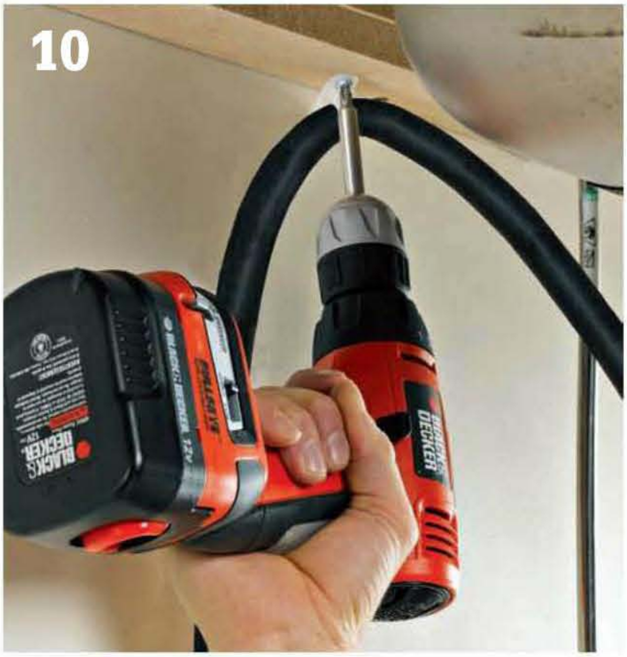

Once the dishwasher is level, attach the brackets to the underside of the countertop to keep it from moving. Then pull the discharge hose into the sink cabinet and install it so there's a loop that is attached with a bracket to the underside of the countertop. (This loop prevents waste water from flowing from the disposer back into the dishwasher.)

Lengthening a Discharge Hose ▸

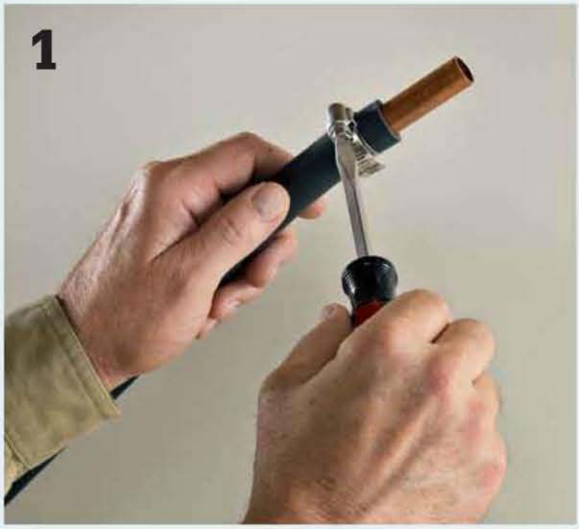

If the discharge hose has to be modified to fit onto the disposer port, insert a 4"-long piece of ½" copper tubing into the hose and hold it in place with a hose clamp. This provides a nipple for the rubber adapter that fits onto the disposer.

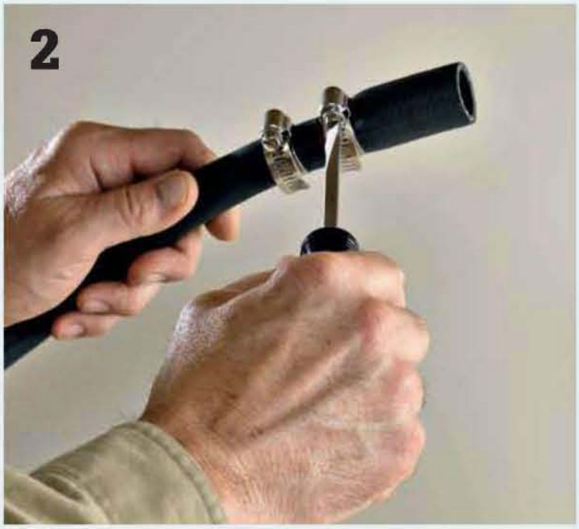

Clamp the rubber disposer adapter to the end of the copper tubing nipple. Then tighten the hose clamp securely.

(continued)

11

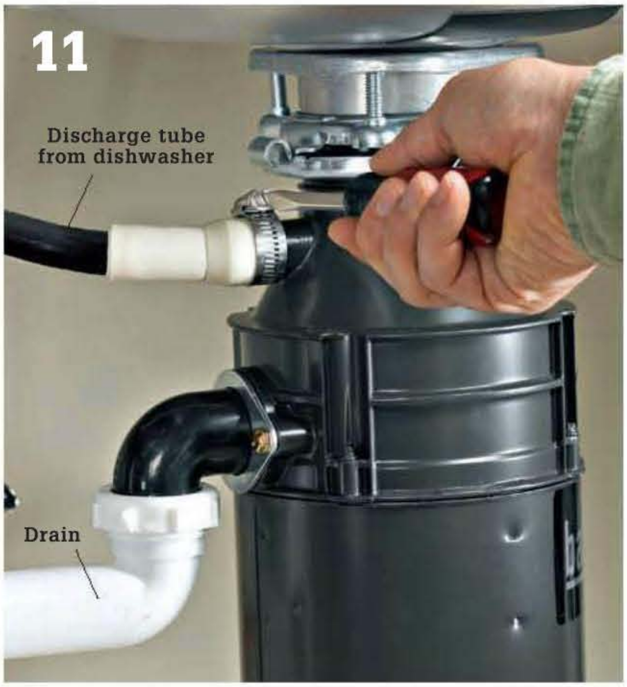

Push the adapter over the disposer's discharge nipple and tighten it in place with a hose clamp. If you don't have a disposer, this discharge hose can be clamped directly to a modified sink tailpiece that's installed below a standard sink strainer.

12

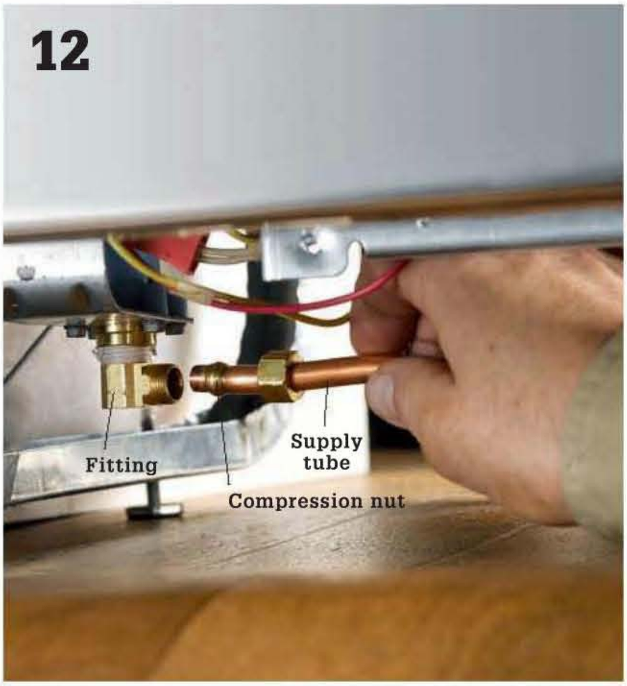

Adjust the L-fitting on the dishwasher's water inlet valve until it points directly toward the water supply tubing. Then lubricate the threads slightly with a drop of dishwashing liquid and tighten the tubing's compression nut onto the fitting. Use an adjustable wrench and turn the nut clockwise.

13

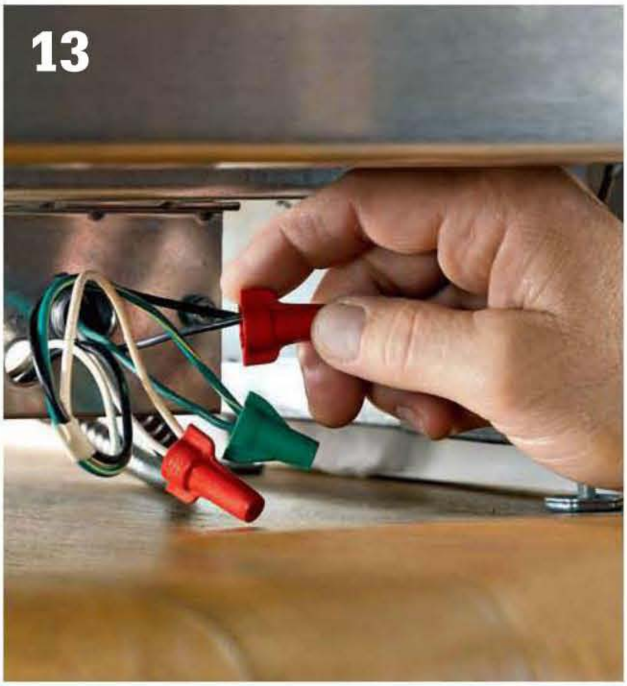

Complete the electrical connections by tightening the connector's clamp on the cable. Join the power wires to the fixture wires with wire connectors. Attach the ground wire (or wires) to the grounding screw on the box, and replace the cover.

14

Install the access panel, usually by hooking it on a couple of prongs just below the dishwasher's door. Install the screws (if any) that hold it in place, and turn on the water and power supplies. Replace the toe-kick panel at the bottom of the dishwasher.

Installing Food Disposers

Food disposers are standard equipment in the modern home, and most of us have come to depend on them to macerate our plate leavings and crumbs so they can exit the house along with waste water from the sink drain. If your existing disposer needs replacing, you'll find that the job is relatively simple, especially if you select a replacement appliance that is the same model as the old one. In that case, you can probably reuse the existing mounting assembly, drain sleeve, and drain plumbing.

Most food disposers are classified as "continuous feed" because they can only operate when an ON/OFF switch on the wall is being actively held down. Let go of the switch, and the disposer stops. Each appliance has a power rating between ⅓ and 1 HP (horsepower). More powerful models bog down less under load and the motors last longer because they don't have to work as hard. They are also costlier.

A disposer is hardwired to a switch mounted in an electrical box in the wall above the countertop. If your kitchen is not equipped for this, consult a wiring guide or hire an electrician. The actual electrical hookup of the appliance is quite simple (you only have to join two wires) but you can hire an electrician if you are not comfortable with the job.

Tools & Materials ▸

- Screwdriver
- Channel-type pliers
- Spud wrench (optional)
- Hammer
- Hacksaw or tubing cutter
- Kitchen drain supplies
- Drain auger
- Putty knife
- Mineral spirits
- Plumber's putty
- Wire caps
- Hose clamps
- Threaded Y-fitting
- Electrical tape

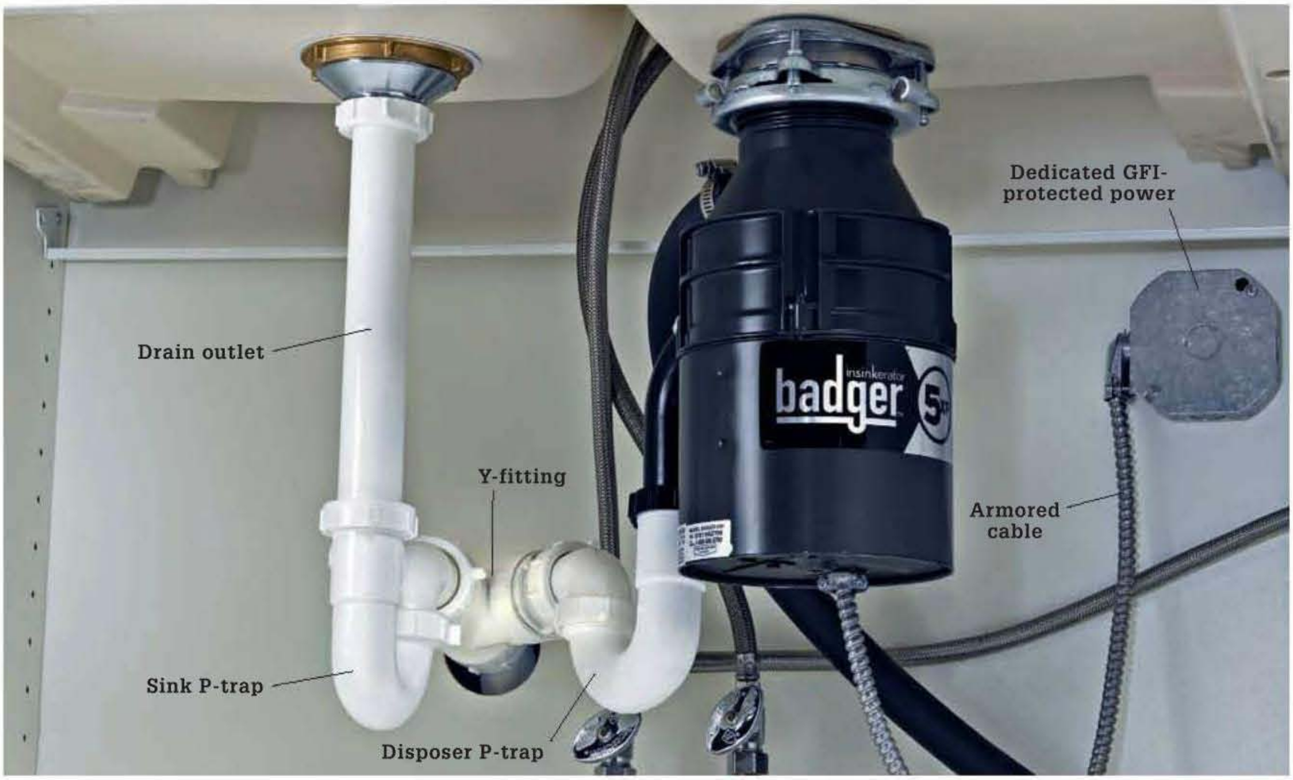

A properly functioning food disposer that's used correctly can actually help reduce clogs by ensuring that large bits of organic matter don't clog the drain system. Many plumbers suggest using separate P-traps for the disposer and the drain outlet tube as shown here.

How to Install a Food Disposer

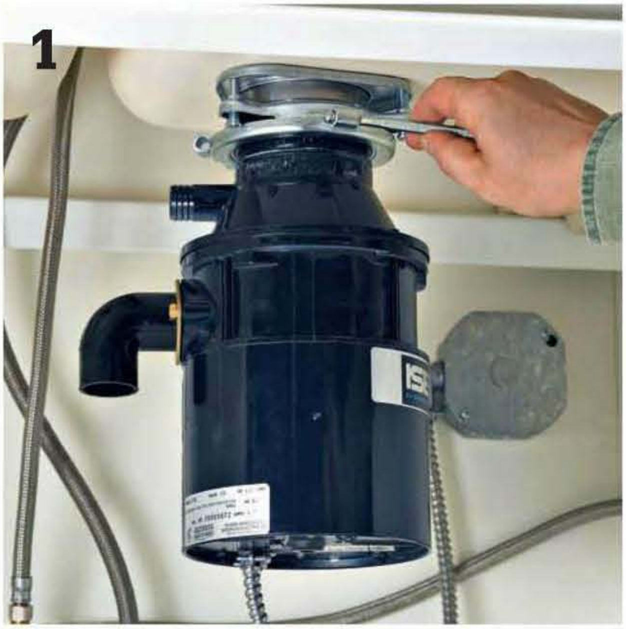

Remove the old disposer if you have one. You'll need to disconnect the drain pipes and traps first. If your old disposer has a special wrench for the mounting lugs, use it to loosen the lugs. Otherwise, use a screwdriver. If you do not have a helper, place a solid object directly beneath the disposer to support it before you begin removal. *Important: Shut off electrical power at the main service panel before you begin removal. Disconnect the wire leads, cap them, and stuff them into the electrical box.*

Tip ▸

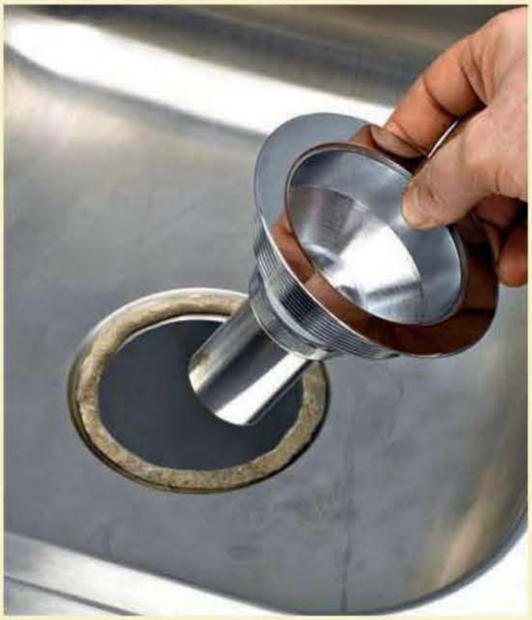

Alternate: If you are installing a disposer in a sink that did not previously have one, remove the old sink strainer and drain tailpiece. Scrape up any old plumbers putty and clean the sink thoroughly around the drain opening with mineral spirits.

Clear the drain lines all the way to the branch drain before you begin the new installation. Remove the trap and trap arm first.

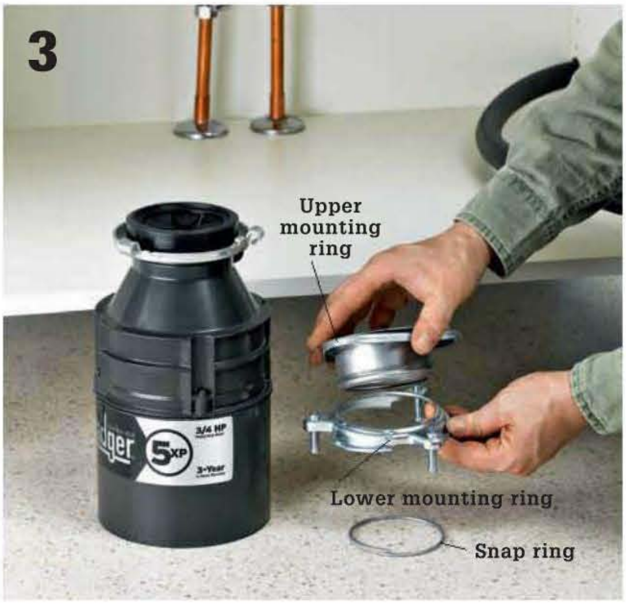

Disassemble the mounting assembly and then separate the upper and lower mounting rings and the backup ring. Also remove the snap ring from the sink sleeve.

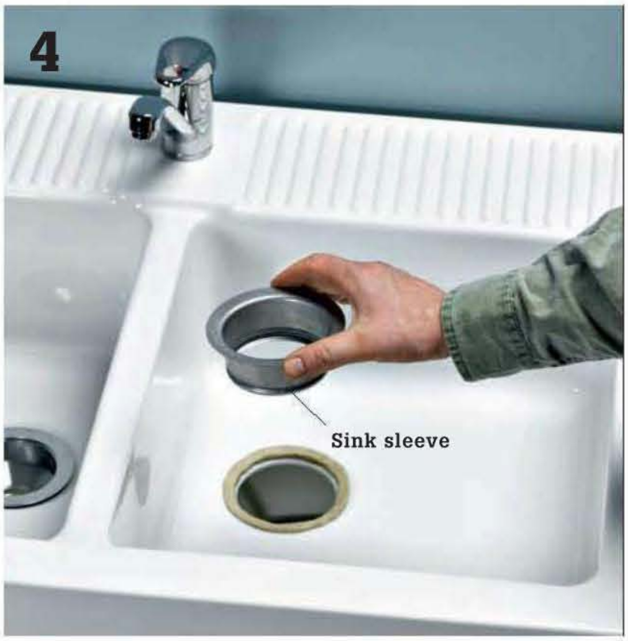

Press the flange of the sink sleeve for your new disposer into a thin coil of plumber's putty that you have laid around the perimeter of the drain opening. The sleeve should be well-seated in the putty.

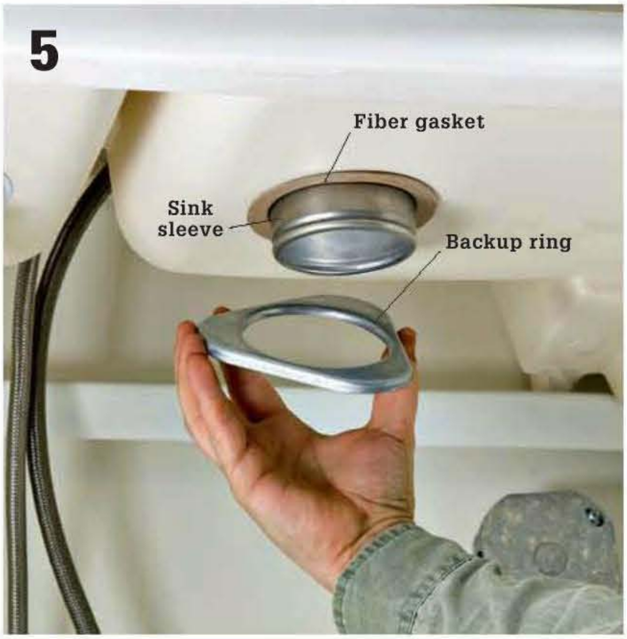

Slip the fiber gasket and then the backup ring onto the sink sleeve, working from inside the sink base cabinet. Make sure the backup ring is oriented the same way it was before you disassembled the mounting assembly.

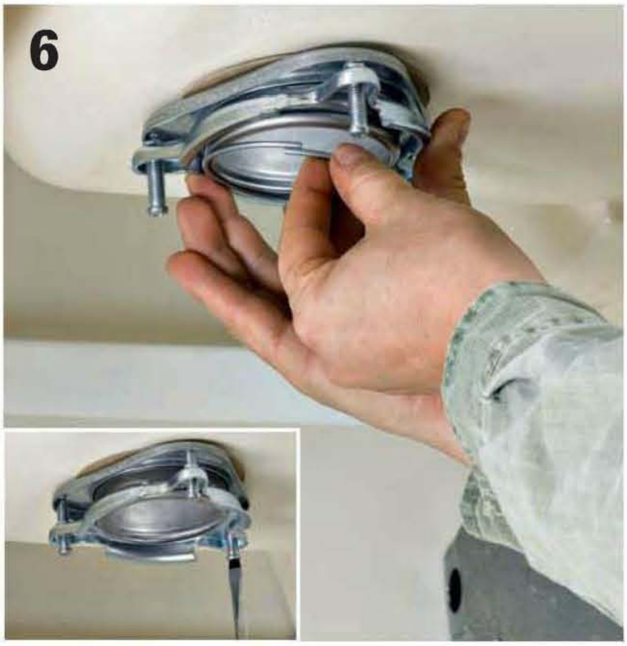

Insert the upper mounting ring onto the sleeve with the slotted ends of the screws facing away from the backup ring so you can access them. Then, holding all three parts at the top of the sleeve, slide the snap ring onto the sleeve until it snaps into the groove. Tighten the three mounting screws on the upper mounting ring until the tips press firmly against the backup ring (inset photo). It is the tension created by these screws that keeps the disposer steady and minimizes vibrating.

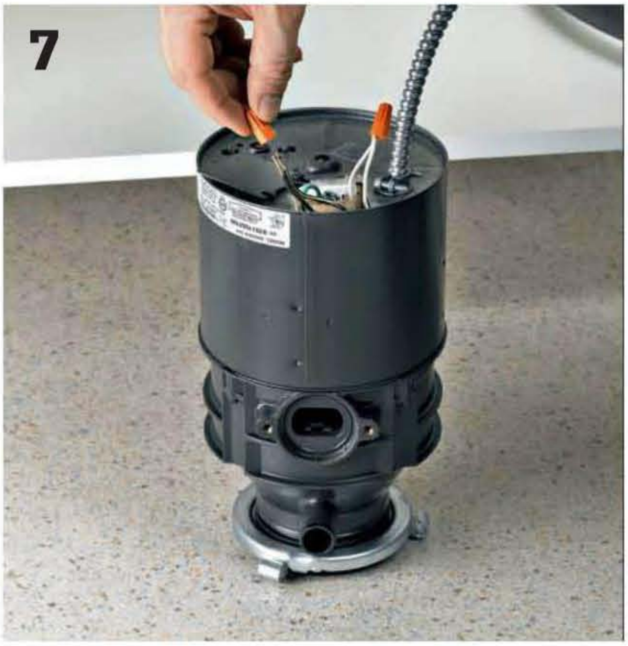

Make electrical connections before you mount the disposer unit on the mounting assembly. Make sure the power is still turned off at the service panel. Remove the access plate from the disposer. Attach the white and black feeder wires from the electrical box to the white and black wires (respectively) inside the disposer. Twist a small wire cap onto each connection and wrap it with electrical tape for good measure. Also attach the green ground wire from the box to the grounding terminal on your disposer.

(continued)

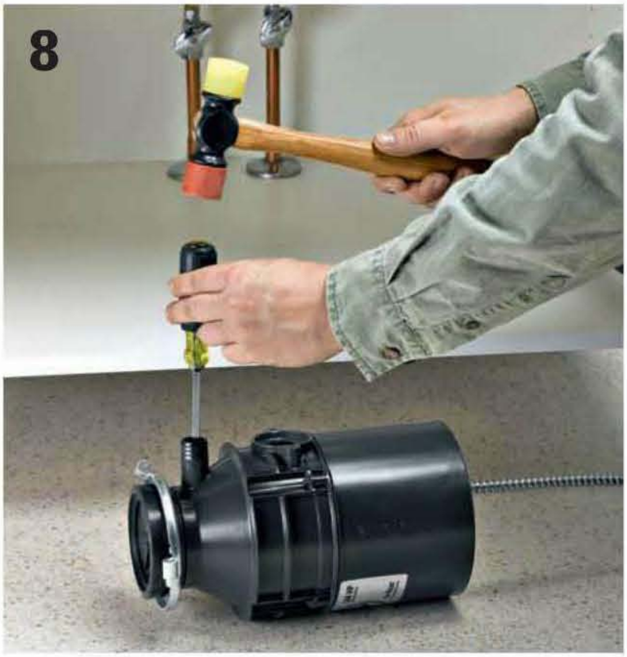
8

Knock out the plug in the disposer port if you will be connecting your dishwasher to the disposer. If you have no dishwasher, leave the plug in. Insert a large flathead screwdriver into the port opening and rap it with a mallet. Retrieve the knock plug from inside the disposer canister.

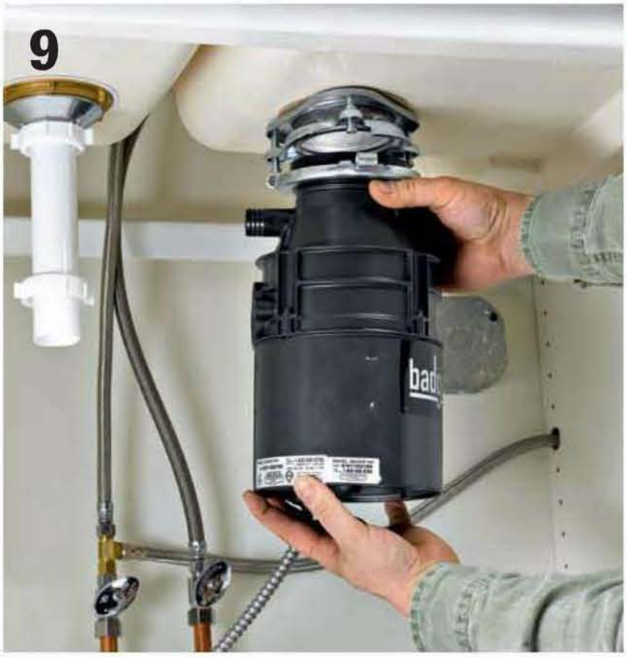
9

Hang the disposer from the mounting ring attached to the sink sleeve. To hang it, simply lift it up and position the unit so the three mounting ears are underneath the three mounting screws and then spin the unit so all three ears fit into the mounting assembly. Wait until after the plumbing hookups have been made to lock the unit in place.

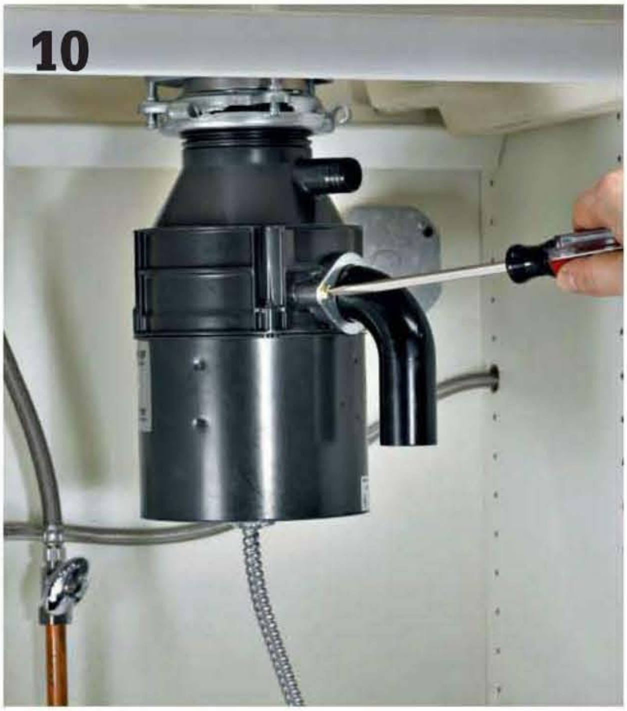
10

Attach the discharge tube to the disposer according to the manufacturer's instructions. It is important to get a very good seal here, or the disposer will leak. Go ahead and spin the disposer if it helps you access the discharge port.

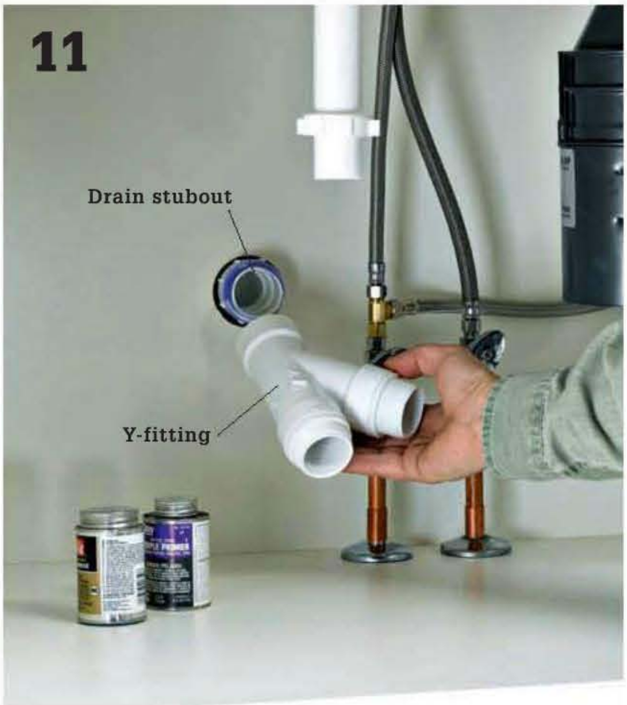
11

Attach a Y-fitting at the drain stubout. The Y-fitting should be sized to accept a drain line from the disposer and another from the sink. Adjust the sink drain plumbing as needed to get from the sink P-trap to one opening of the Y.

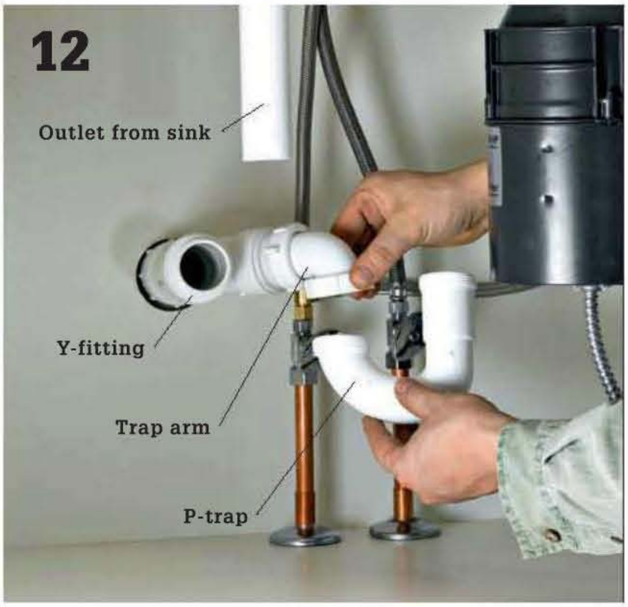

Install a trap arm for the disposer in the open port of the Y-fitting at the wall stubout. Then, attach a P-trap or a combination of a tube extension and a P-trap so the low end of the trap will align with the bottom of the disposer discharge tube.

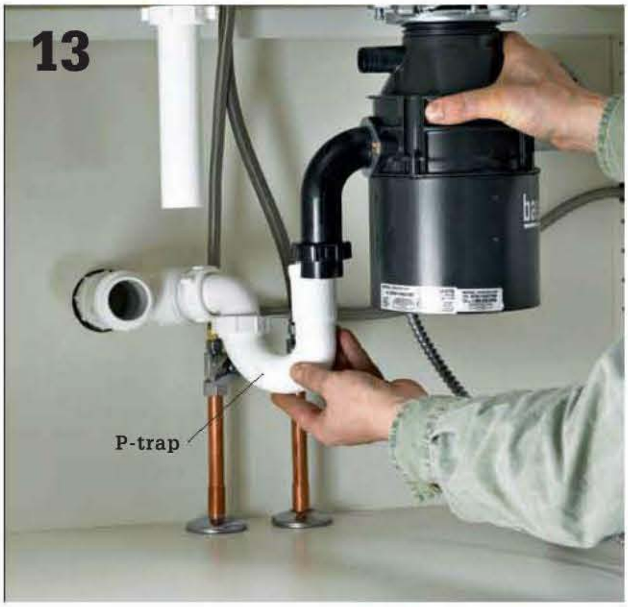

Spin the disposer so the end of the discharge tube is lined up over the open end of the P-trap and confirm that they will fit together correctly. If the discharge tube extends down too far, mark a line on it at the top of the P-trap and cut through the line with a hacksaw. If the tube is too short, attach an extension with a slip joint. You may need to further shorten the discharge tube first to create enough room for the slip joint on the extension. Slide a slip nut and beveled compression washer onto the discharge tube and attach the tube to the P-trap.

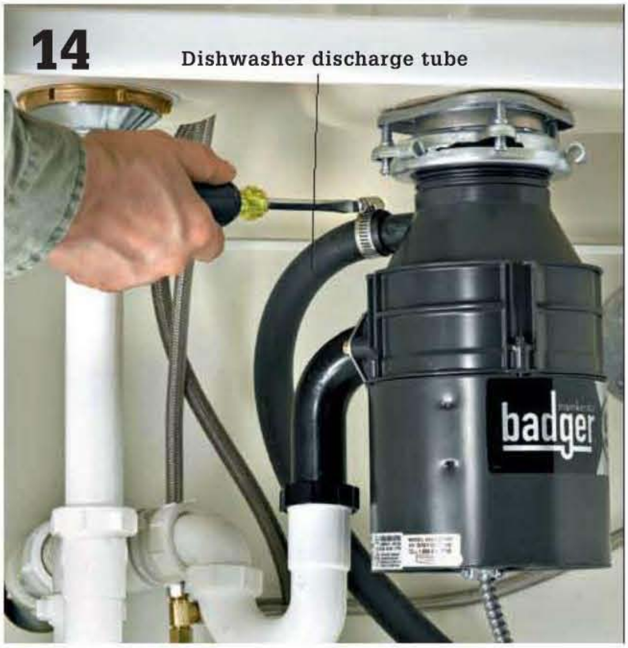

Connect the dishwasher discharge tube to the inlet port located at the top of the disposer unit. This may require a dishwasher hookup kit.

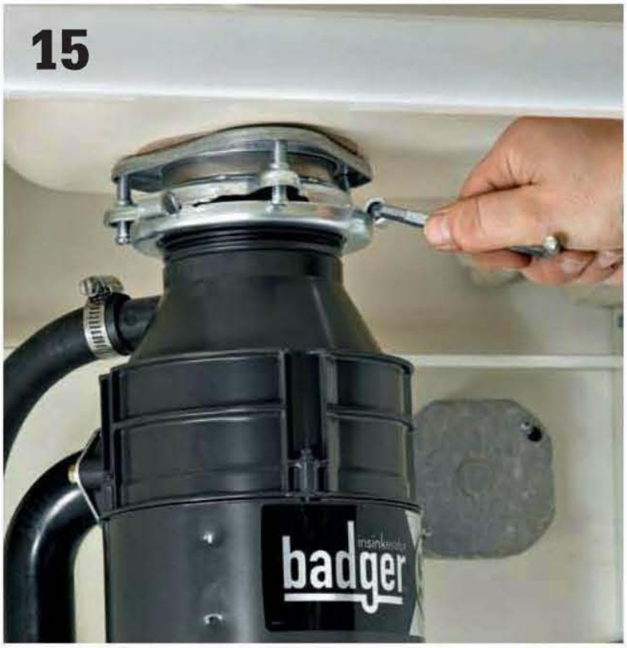

Lock the disposer into position on the mounting ring assembly once you have tested to make sure it is functioning correctly and without leaks. Lock it by turning one of the mounting lugs with a screwdriver until it makes contact with the locking notch.

Installing Undercabinet Lights

Hardwired undercabinet lights illuminate the kitchen countertop and sink areas that fall in the shadow of ceiling lights. Most of these light fixtures, which are often called *strip lights,* utilize fluorescent, halogen, or xenon bulbs that emit very low levels of heat and are therefore very efficient.

If you are doing a kitchen remodel with all-new cabinets, run the new light circuit wiring before the cabinets are installed. For a retrofit, you'll need to find an available power source to tie into. Options for this do not include the dedicated 20-amp small-appliance circuits that are required in kitchens. The best bet is to run new circuit wire from a convenient ceiling light switch box, but this will mean cutting into the walls to run cable. Another option is to locate a receptacle that's on the opposite side of a shared wall, preferably next to a location where a base cabinet is installed in the kitchen. By cutting an access hole in the cabinet back you can tie into the receptacle box and run cable through the wall behind the cabinets, up to the upper cabinet location, and out the wall to supply the fixture that's mounted to the underside of the upper cabinet.

You can purchase undercabinet lights that are controlled by a wall switch, but most products have an integral on/off button so you can control lights individually.

Tools & Materials ▸

- Circuit tester
- Utility knife
- Wallboard saw
- Hammer
- Screwdriver
- Drill and hole saw
- Jigsaw
- Wire stripper
- Undercabinet lighting kit
- 14/2 NM cable
- Wire connectors
- Switch box
- Switch

Undercabinet lights provide directed task lighting that bring sinks and countertop work surfaces out from the shadows. Hardwired lights may be controlled either by a wall switch or an onboard on/off switch located on the fixture. *Note: Do not supply power for lights from a small-appliance circuit.*

How to Install a Hardwired Undercabinet Light

1

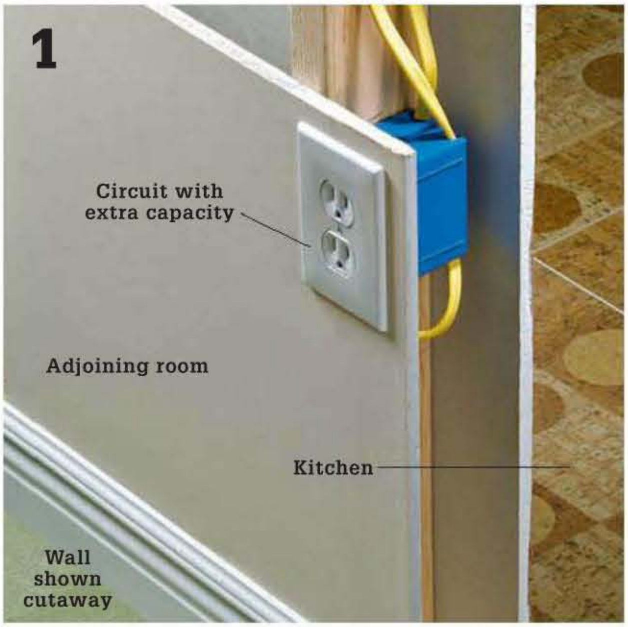

Look in an adjoining room for a usable power source in the form of a receptacle that has a box located in the wall behind your base cabinets. Unlike the small-appliance circuit with outlets in your backsplash area, these typically are not dedicated circuits (which can't be expanded). Make sure that the receptacle's circuit has enough capacity to support another load. Shut the power to the receptacle off at the main service panel and test for power.

2

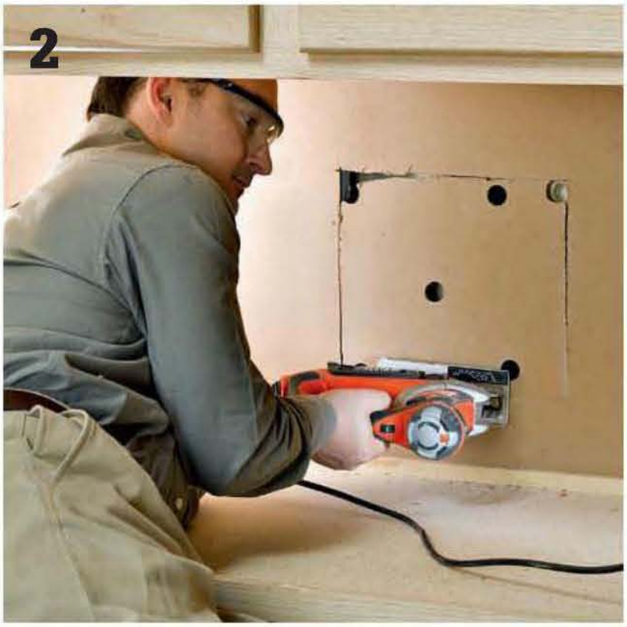

Cut a hole in the base cabinet back panel, roughly the area where you know the usable receptacle to be. Use a keyhole saw or drywall saw and make very shallow cuts until you have positively identified the locations of the electrical box and cables. Then finish the cuts with a jigsaw.

3

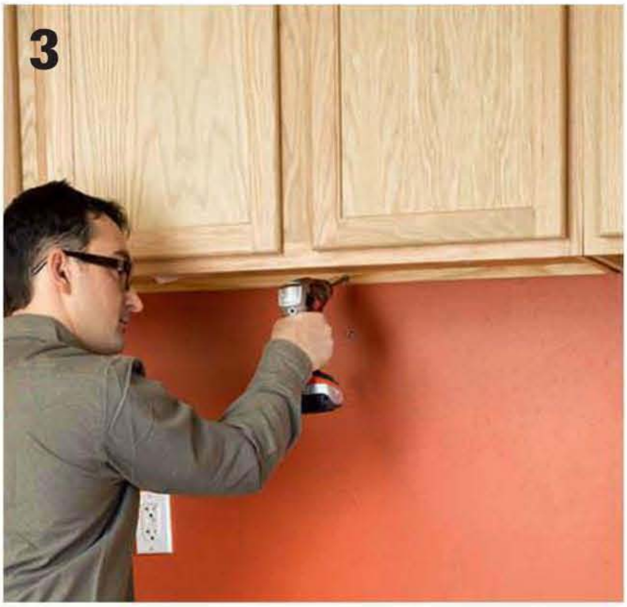

Drill an access hole through the cabinet frame and wall for the cable that will feed the undercabinet light. A ½" dia. hole should be about the right size if you are using 12-ga. or 14-ga. sheathed NM cable.

4

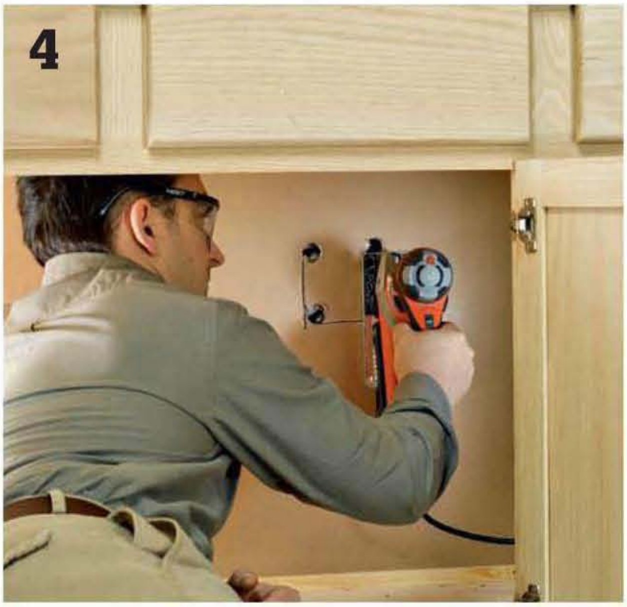

Cut a small access hole (4 × 4" or so) in the back panel of the base cabinet directly below the undercabinet light location.

(continued)

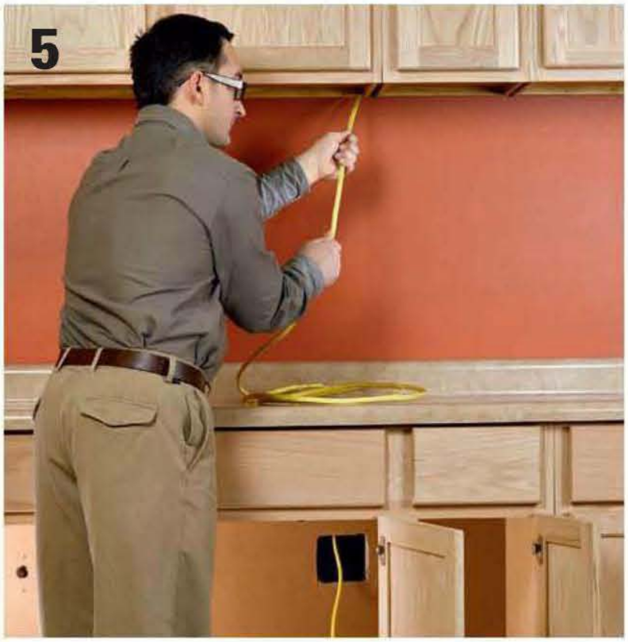

Feed cable into the access hole at the light location until the end reaches the access hole below. Don't cut the cable yet. Reach into the access hole and feel around for the free cable end and then pull it out through the access hole once you've found it. Cut the cable, making sure to leave plenty of extra on both ends.

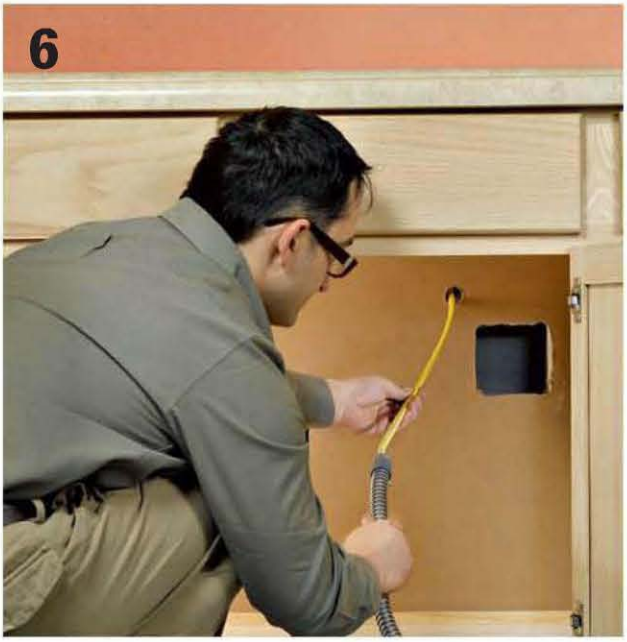

String the cable into a piece of flexible conduit that's long enough to reach between the two access holes in the base cabinets. To make patching the cabinet back easier, drill a new access hole for the cable near the square access hole. Attach a connector to each end of the conduit to protect the cable sheathing from the sharp edges of the cut metal.

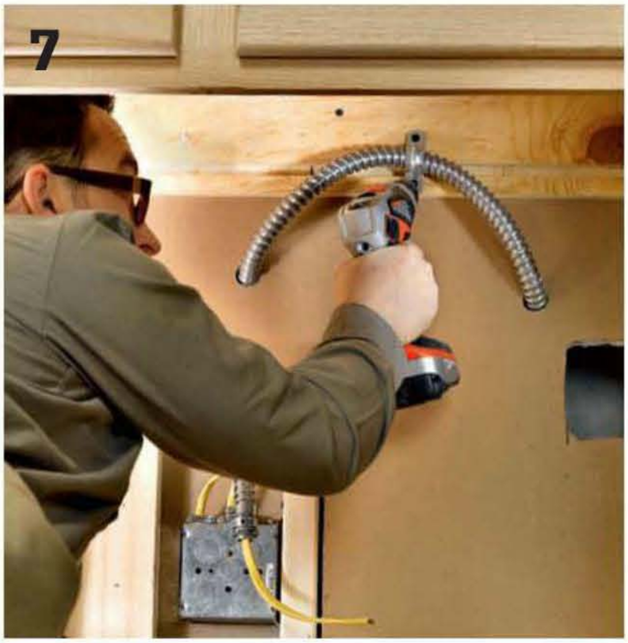

Hang the conduit with hanger straps attached to the base cabinet frame or back panel, drilling holes in the side walls of the cabinet where necessary to thread the conduit through. On back panels, use small screws to hang the straps instead of brads or nails. Support the conduit near both the entrance and the exit holes (the conduit should extend past the back panels by a couple of inches).

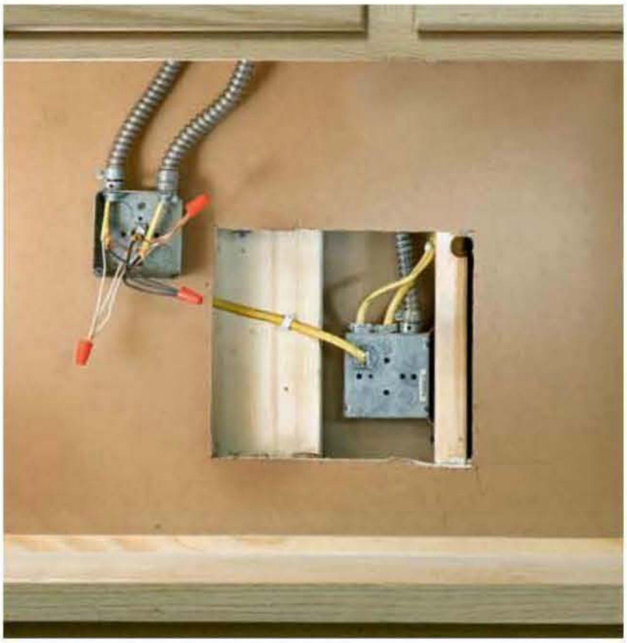

Variation: If you are installing more than one undercabinet light, run cable down from each installation point as you did for the first light. Mount an electrical junction box to the cabinet back near the receptacle providing the power. Run the power cables from each light through flexible conduit and make connections inside the junction box. Be sure to attach the junction box cover once the connections are made.

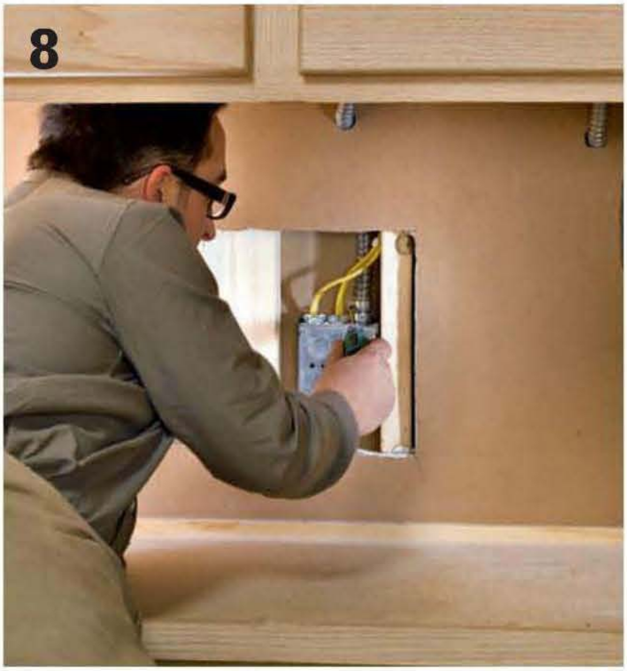

8

Remove the receptacle from the box you are tying into and insert the new circuit cable into one of the knockouts using a cable clamp. Check a wire capacity chart (see page 90) to make sure the box is big enough for the new conductors. Replace it with a larger box if necessary. Reinstall the receptacle once the connections are made.

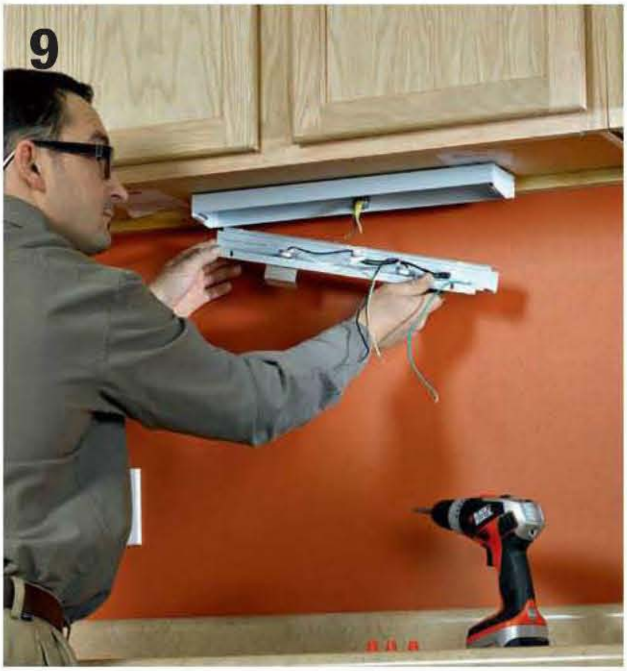

9

Install the undercabinet light. Some models have a removable diffuser that allows access to the fixture wires, and these should be screwed to the upper cabinet prior to making your wiring hookups. Other models need to be connected to the circuit wires before installation. Check your manufacturer's installations.

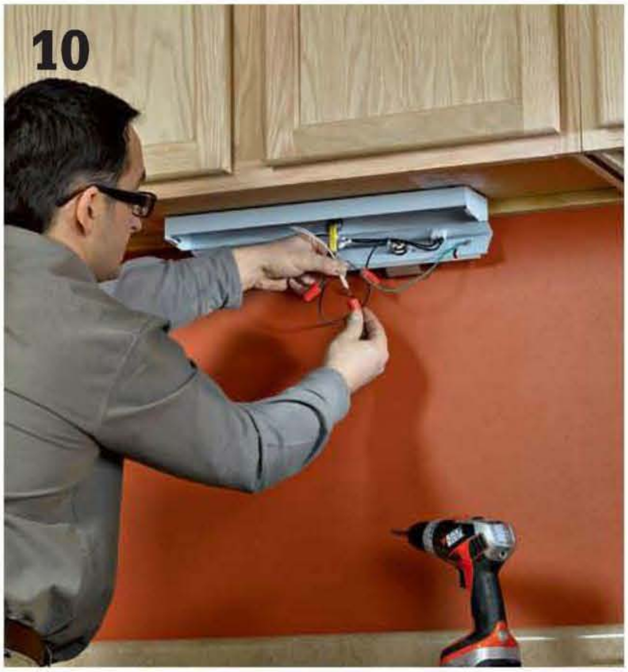

10

Connect wires inside the light fixture according to the light manufacturer's directions. Make sure the incoming cable is stapled just before it enters the light box and that a cable clamp is used at the knockout in the box to protect the cable. Restore power and test the light.

11

Cut patches of hardboard and fit them over the access holes, overlapping the edges of the cutouts. Adhere them to the cabinet backs with panel adhesive.

Green Kitchens ▸

In many ways, a green kitchen is just like any well-designed kitchen. There are windows for natural lighting and air circulation. Surfaces are durable, hygienic, and low-maintenance. The room is well lighted, with overhead fixtures and plenty of task lighting. And the overall plan is designed for maximum work efficiency and tailored to the way the homeowners use the kitchen on a daily basis. In terms of aesthetics, a green kitchen can be just as beautiful, if not more so, than a conventional kitchen.

So what's different about a green kitchen? For starters, it has highly efficient appliances, particularly the refrigerator. It contains sustainable and non-toxic materials, like formaldehyde-free cabinet cases and linoleum flooring. It's also ready to deal with the variety of waste products surrounding food preparation, including organic garbage, recyclable materials, and gray water.

Many green kitchens are further defined by what they don't have: trendy upgrades and extras that do little to improve anyone's daily experience in the kitchen. Depending on the type of cooking you do, you might get a lot of use out of special features, like a pot filler faucet, an extra prep sink, warming drawers, or a commercial-style range. But in reality, most cooks don't get full value from these pricey additions, which take up space and are often over-engineered and energy-inefficient (not to mention the environmental impact of producing them).

Whether you're planning a major kitchen remodel or simply replacing an item or two, here are some of the main elements you should consider for making the space more useful, healthful, and resource-efficient.

Efficiency and common sense are two of the hallmarks of any green room, but this is especially true in the kitchen. Keep the appliances as small as you can and look for natural materials, like this butcherblock countertop made from maple cutoffs.

PORTRAIT OF A GREEN KITCHEN

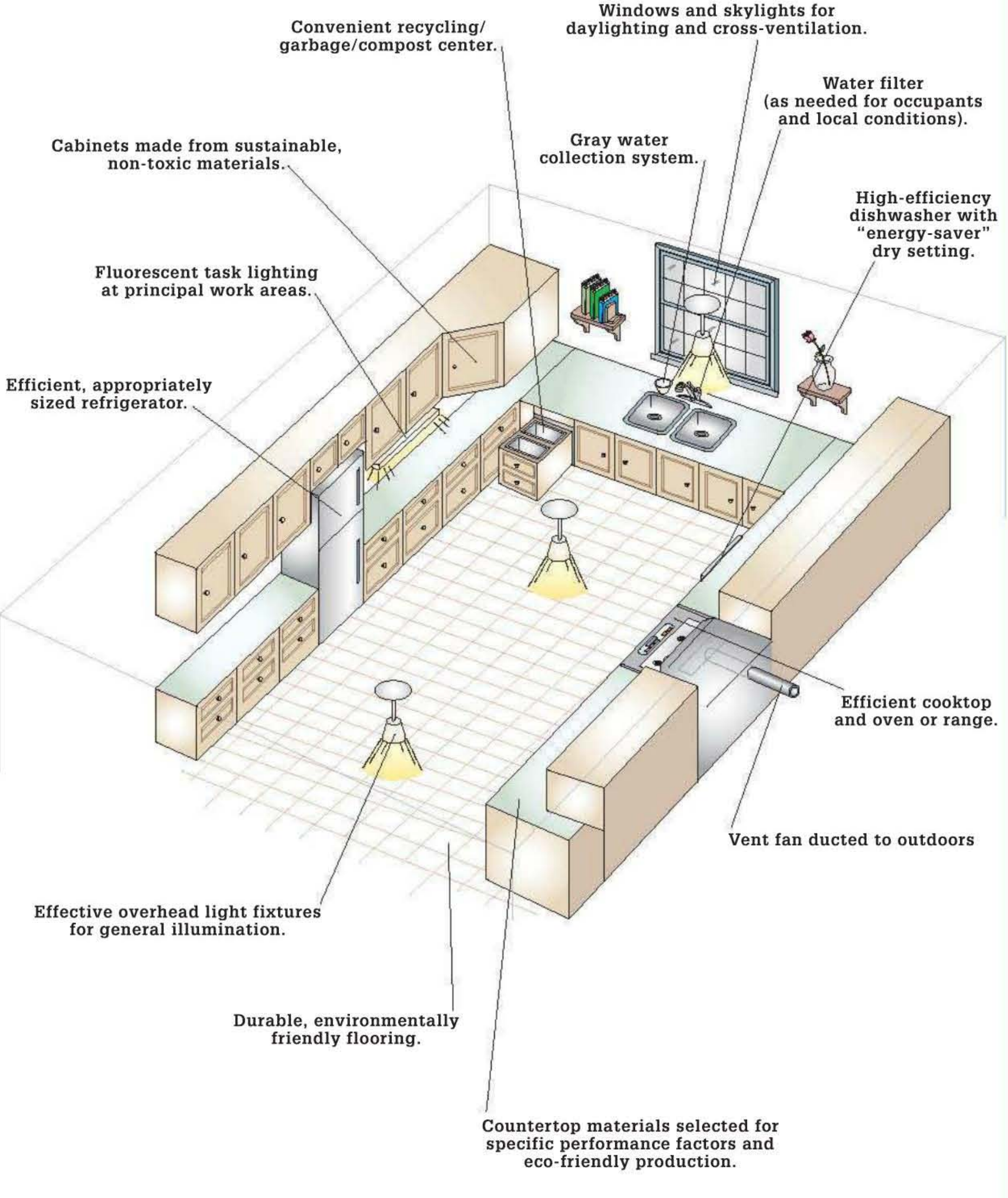
Convenient recycling/ garbage/compost center.
Windows and skylights for daylighting and cross-ventilation.
Water filter (as needed for occupants and local conditions).
Cabinets made from sustainable, non-toxic materials.
Gray water collection system.
High-efficiency dishwasher with "energy-saver" dry setting.
Fluorescent task lighting at principal work areas.
Efficient, appropriately sized refrigerator.
Efficient cooktop and oven or range.
Vent fan ducted to outdoors
Effective overhead light fixtures for general illumination.
Durable, environmentally friendly flooring.
Countertop materials selected for specific performance factors and eco-friendly production.

Green Kitchen Appliances ▸

Because it's full of hardworking appliances, the kitchen uses more energy than any other room in the house. Upgrading your old machines with more efficient models is one of the simplest ways to help make your kitchen green. Refrigerators are the big consumers, accounting for up to 15% of the average home's total electricity usage. Fortunately, you don't have to sacrifice quality or style to save energy. Highly efficient appliances, like many other machines, are simply better designed and better built than their hoggish counterparts. And you can't tell a miserly model by its looks.

Lots of do-it-yourselfers install their own appliances; see pages 402 to 407 for steps on installing a high-efficiency dishwasher. A refrigerator is the easiest to install: Just plug it in, slide it in place, and level it. If yours has an icemaker, you'll have to connect a small water line in the back of the unit (and check it for leaks in the beginning).

If you're replacing an existing appliance, find out where you can recycle the old unit. Appliances are a major source of recyclable steel and other materials, and besides, dumping an operable machine in the landfill just to make room for a new one isn't a very green move. Appliances that still work can find a great home through a local charity organization (this means an old, inefficient appliance is still in circulation, but it postpones the production of a new product, and it helps out those in need). To scrap an ailing or dead appliance, contact your local recycling authority or Earth 911, online at www.earth911.org.

Whatever you do, don't keep your old refrigerator and move it into the garage for backup storage. Keeping a second fridge plugged in will undo all energy-saving efforts from your kitchen upgrades many times over.

European sized refrigerators are becoming increasingly popular as a way to control energy usage. A typical model like the one above has a refrigerator compartment of 7-to-8 cubic feet in capacity, with a freezer that's 3 to 4 cu. ft.

Scale is important and many kitchen appliance manufacturers have begun marketing to this. Narrow dishwashers with smaller compartments offer ample space for some users, and they do their job with much less water and energy.

ENERGY STAR & ENERGYGUIDE LABELS

If you're in the market for a new appliance, it's time to start paying attention to those yellow-and-black ENERGYGUIDE labels you see pasted to the front of many new products. The labels, required by the Federal Trade Commission, help consumers quickly compare models for their energy efficiency. The box in the middle of each label provides the "energy use" of all models in a given class, with an arrow showing where that particular model falls within the range. The label also gives you an estimated annual operating cost for the appliance.

Energy Star is the Department of Energy's wide-reaching program that awards Energy Star status to the most efficient products in a given class. The program applies to major appliances, windows, doors, HVAC systems, and other household items. In general, Energy Star products exceed the federal government's energy-efficiency standards and perform within the top 25% of their category.

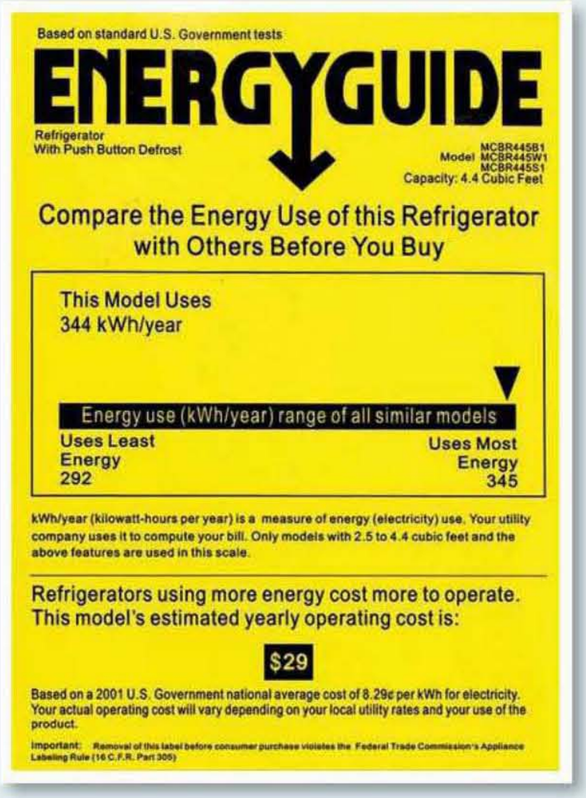

Based on standard U.S. Government tests

ENERGYGUIDE

Refrigerator
With Push Button Defrost

Model MCBR445B1
MCBR445W1
MCBR445S1
Capacity: 4.4 Cubic Feet

Compare the Energy Use of this Refrigerator with Others Before You Buy

This Model Uses
344 kWh/year

Energy use (kWh/year) range of all similar models

Uses Least Energy	Uses Most Energy
292	345

kWh/year (kilowatt-hours per year) is a measure of energy (electricity) use. Your utility company uses it to compute your bill. Only models with 2.5 to 4.4 cubic feet and the above features are used in this scale.

Refrigerators using more energy cost more to operate. This model's estimated yearly operating cost is:

$29

Based on a 2001 U.S. Government national average cost of 8.29¢ per kWh for electricity. Your actual operating cost will vary depending on your local utility rates and your use of the product.

Important: Removal of this label before consumer purchase violates the Federal Trade Commission's Appliance Labeling Rule (16 C.F.R. Part 305)

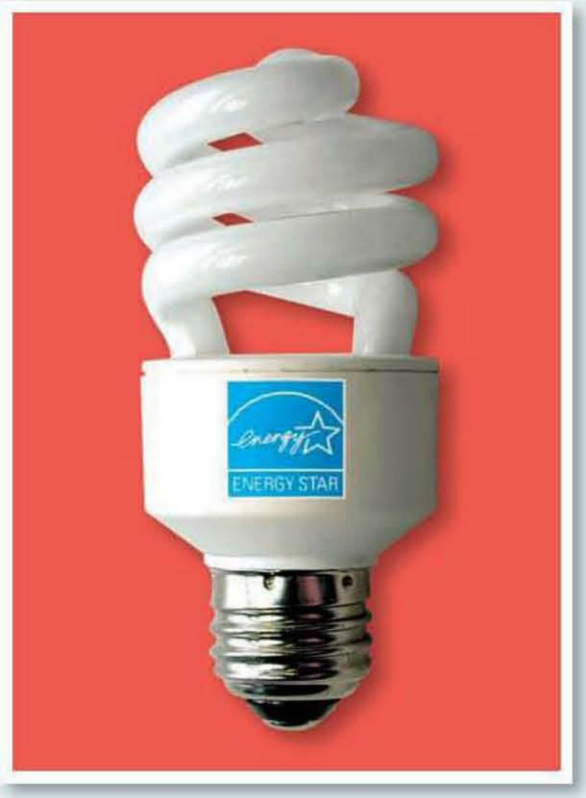

ADDITIONAL RESOURCES FOR CHOOSING & USING KITCHEN APPLIANCES

- U.S. Department of Energy (DOE): www.eere.energy.gov
- Energy Star: www.energystar.gov
- Consumer Reports, for unbiased analysis of appliance performance and other shopping considerations, online at www.consumerreports.org
- American Council for an Energy-Efficient Economy: www.aceee.org
- Your local utility provider
- Your state's energy authority; some states offer rebates on high-efficiency appliances
- Association of Home Appliance Manufacturers: www.aham.org
- The Green Guide: www.thegreenguide.com

Bathroom Remodeling

Bathroom fixtures and materials have changed dramatically in recent years. Tubs can rival healthclub spas, toilets and showerheads are more water efficient, and sinks run the gamut from stainless steel to wood basins.

Before you jump into a bathroom remodeling project, decide what's most important to you. Budget determines how luxurious you can get, but maybe it's worthwhile to splurge on that one item that would make you feel most pampered.

Home centers and kitchen and bath showrooms have a variety of bathroom setups you can browse through to see how the products look and feel. It's always a good idea to see a product in person before doling out the cash.

In this chapter:

- Planning & Designing Bathrooms
- Installing Shower Kits
- Installing Alcove Bathtubs
- Installing 3-piece Tub Surrounds
- Installing Sliding Tub Doors
- Tiling Bathroom Walls
- Installing Toilets
- Installing Wall-hung Vanities
- Installing Vessel Sinks
- Installing Pedestal Sinks
- Installing Vanity Cabinets
- Installing Bathroom Sink Faucets & Drains
- Installing Tub Spouts
- Installing Vanity Lights
- Installing Grab Bars
- Installing Glass Shelving
- Installing Towel Warmers

Planning & Designing Bathrooms

Once you've drawn up your plans and created a materials list, you'll need to have them reviewed by your local building department. Getting approval early in the process can save you time and expense later. To help ensure success, here are some design standards for you to follow:

The National Kitchen and Bath (NKBA) publishes a list of bathroom design standards to help people plan rooms that are safe and accessible to all users (see Resources, page 554).

Your bathroom probably won't conform to all of the recommended standards, but they can help guide your overall plan. What your plan must include is everything prescribed by the local building codes, including plumbing and wiring codes.

Bathroom Design Standards ▸

Codes and permits are necessary to ensure safety in any remodel. They're not the most fun to focus on—not like choosing just the right floor covering or deciding between granite or marble countertops—but they are important.

- Plan doorways with a clear floor space equal to the door's width on the push side and greater than the door's width on the pull side. *Note: Clear floor spaces within the bathroom can overlap.*
- Design toilet enclosures with at least 36" × 66" of space; include a pocket door or a door that swings out toward the rest of the bathroom.
- Install toiletpaper holders approximately 26" above the floor, toward the front of the toilet bowl.
- Place fixtures so faucets are accessible from outside the tub or shower. Add antiscald devices to tub and sink faucets (they are required for shower faucets).
- Avoid steps around showers and tubs, if possible.
- Fit showers and tubs with safety rails and grab bars.
- Install shower doors so they swing open into the bathroom, not the shower.
- Use tempered glass or another type of safety glass for all glass doors and partitions.
- Include storage for soap, towels, and other items near the shower, located within 15" to 48" above the floor. These should be accessible to a person in the shower or tub.
- Provide natural light equal to at least 10% of the floor area in the room.
- Illuminate all activity centers in the bathroom with task and ambient lighting.
- Provide a minimum clearance of 15" from the centerline of sinks to any sidewalls. Double-bowl sinks should have 30" clearance between bowls from centerline to centerline.
- Provide access panels for all electrical, plumbing, and HVAC systems connections.
- Include a ventilation fan that exchanges air at a rate of 8 air changes per hour.
- Choose countertops and other surfaces with edges that are smoothed, clipped, or radiused.

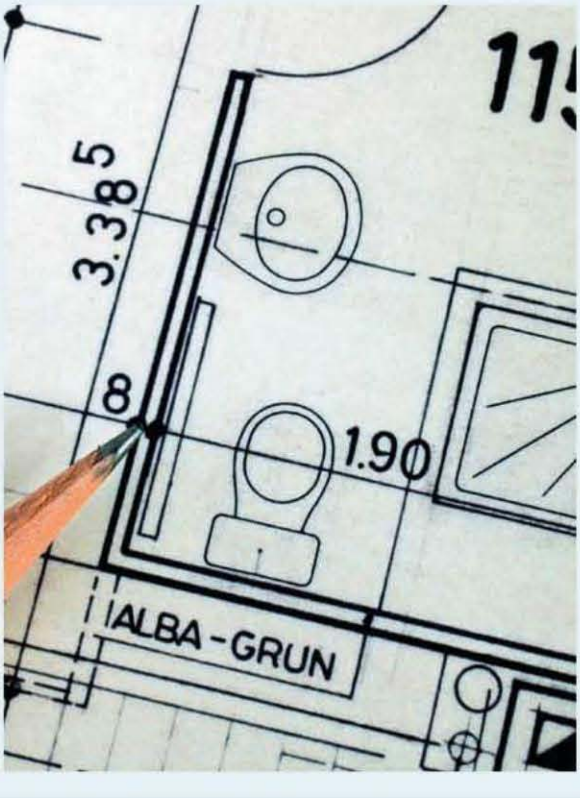

Building Codes for Bathrooms ▸

The following are some of the most common building codes for bathrooms. Contact your local building department for a list of all codes enforced in your area.

- The minimum ceiling height in bathrooms is 7 ft. Minimum floor area is determined by clearances around fixtures.
- Sinks must be at least 4" from side walls and have 21" of clearance in front.
- Sinks must be spaced 4" away from neighboring sinks and toilets, and 2" away from bathtubs.
- Toilets must be centered 15" from side walls and tubs, with 21" clearance in front.
- New and replacement toilets must be low-flow models (1.6 gal./flush).
- Shower stalls must be at least 30" × 30", with 24" of clearance in front of shower openings.
- Steps must be at least 10" deep and no higher than 7¼".
- Faucets for showers and combination tub/showers must be equipped with antiscald devices.
- Supply lines that are ½" in diameter can supply a single fixture, or one sink and one toilet.
- A ¾"-diameter supply line must be used to supply two or more fixtures.
- Waste and drain lines must slope ¼" per foot toward the main DWV stack to aid flow and prevent blockage.
- Each bathroom must be wired with at least one 20-amp circuit for GFCI-protected receptacles, and one 15-amp (minimum) circuit for light fixtures and vent fans without heating elements.
- All receptacles must be GFCI-protected.
- There must be at least one permanent light fixture controlled by a wall switch.
- Wall switches must be at least 60" away from bathtubs and showers.
- Toilet, shower, vanity, or other bathroom compartments must have adequate lighting.
- Light fixtures over bathtubs and showers must be vaporproof, with a UL rating for wet areas.
- Vanity light fixtures with built-in electrical receptacles are prohibited.
- Whirlpool motors must be powered by dedicated GFCI-protected circuits.
- Bathroom vent ducts must terminate no less than 10 ft. horizontally or 3 ft. vertically above skylights.

**Note: Codes for accessible bathrooms may differ (see page 429).*

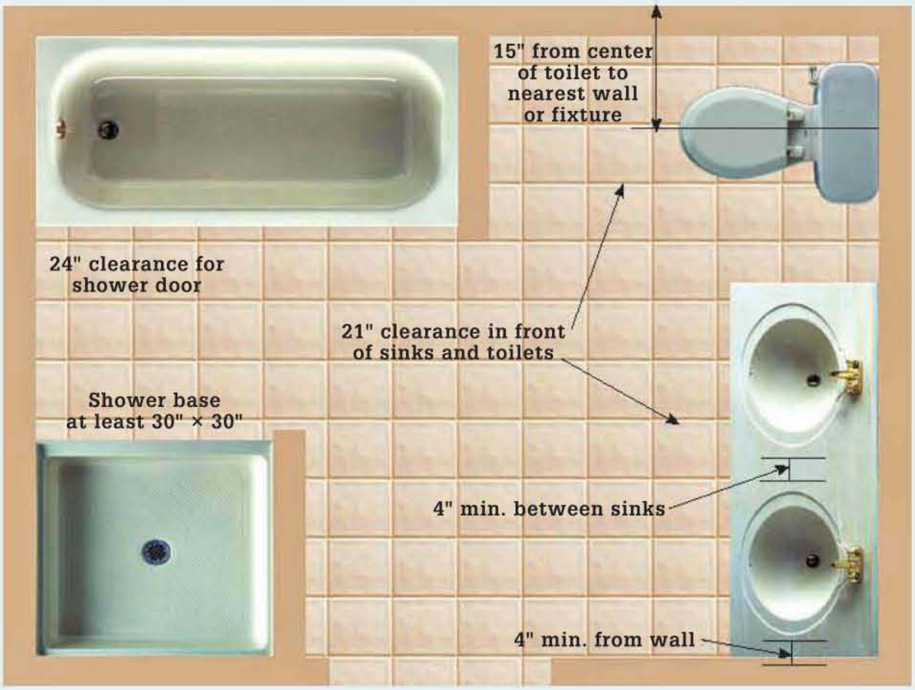
15" from center of toilet to nearest wall or fixture
24" clearance for shower door
21" clearance in front of sinks and toilets
Shower base at least 30" × 30"
4" min. between sinks
4" min. from wall

Follow minimum clearance and size guidelines when planning locations of bathroom fixtures. Easy access to fixtures is fundamental to creating a bathroom that is comfortable, safe, and easy to use.

Accessible Bathrooms (Universal Design)

The safety and accessibility of nearly all aspects of a bathroom can be improved by following universal design guidelines.

For safer floors, add a slip-resistant glaze to ceramic tile, and add nonslip adhesive strips or decals to shower floors. If you're replacing your floor, look for mosaic tiles, vinyl, and cork materials. Matte finishes tend to be less slippery than polished surfaces, and they reduce glare.

Toilets, faucets, sinks, cabinets, tubs, and showers can all be adapted or changed for increased usability by people who have experienced inflexibility or loss of strength.

Toilet height can be adjusted with the installation of a height adapter that raises a standard toilet seat 2 to 5". You can also consider an adjustable-height toilet or a model with a power-lift seat.

A bathroom designed for accessibility is an increasingly important element found in today's homes.

An integral personal hygiene system can be installed to help people with physical disabilities maintain independence.

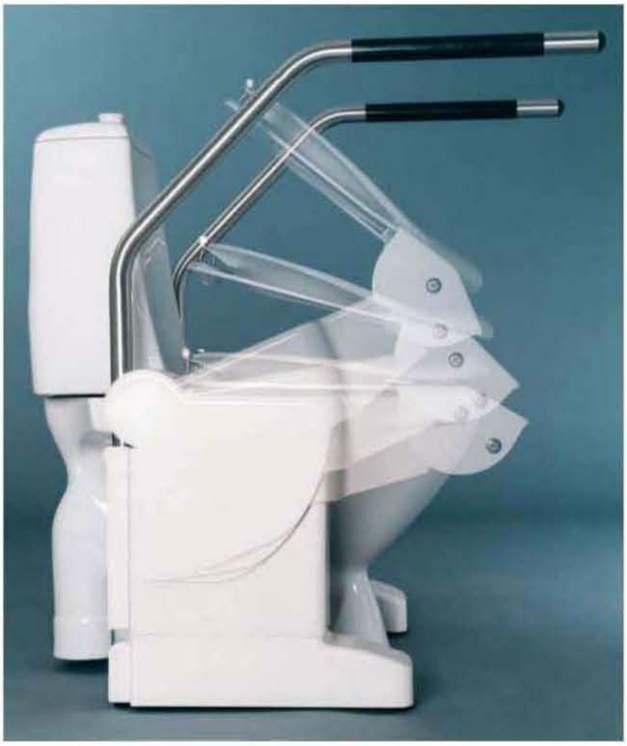

Power-lifts make toilet use easier for people with limited leg or joint strength.

Roll-out base cabinets provide the option of seating space at the countertop.

(continued)

Consider replacing your toilet with a wall-hung style that can be installed at any height, providing additional clear space for maneuvering a wheelchair or walker. To help people with physical disabilities maintain independent personal hygiene, install either a bidet or a toilet with an integral personal hygiene system. Grab bars are a must on walls around the toilet.

Install antiscald guard and volume-control devices on faucets and showerheads. Replace double-handle faucets with single-lever models, which are easier to use. A faucet with motion-sensor operation for hands-free use is a plus.

Change a standard showerhead to an adjustable showerhead mounted on a vertical slide-bar to accommodate people of all heights. Look for handheld models for best control.

Vanities and cabinets with C-shaped pulls or magnetic touch latches are easy to use. Add pull-down hardware to cabinets to bring items within reach for seated people or those with limited mobility.

Install your countertops at varied heights to serve both seated and standing users. Also consider mounting the faucet controls at the side of the sink, rather than the back, for easier access.

If you are replacing your sink, choose a style that is shallower at the front and deeper at the drain. Or install a wall-mounted sink positioned at 30 to 32". Providing a clear space under sinks and low sections of countertop allows seated users to comfortably reach the vanity. Use fold-away doors, remove face frames on base cabinets, or install roll-out base cabinets to gain clear space. Always finish the floor under fold-away or roll-out cabinets. Then insulate hot-water supply pipes or install a protective panel to prevent burns to seated users.

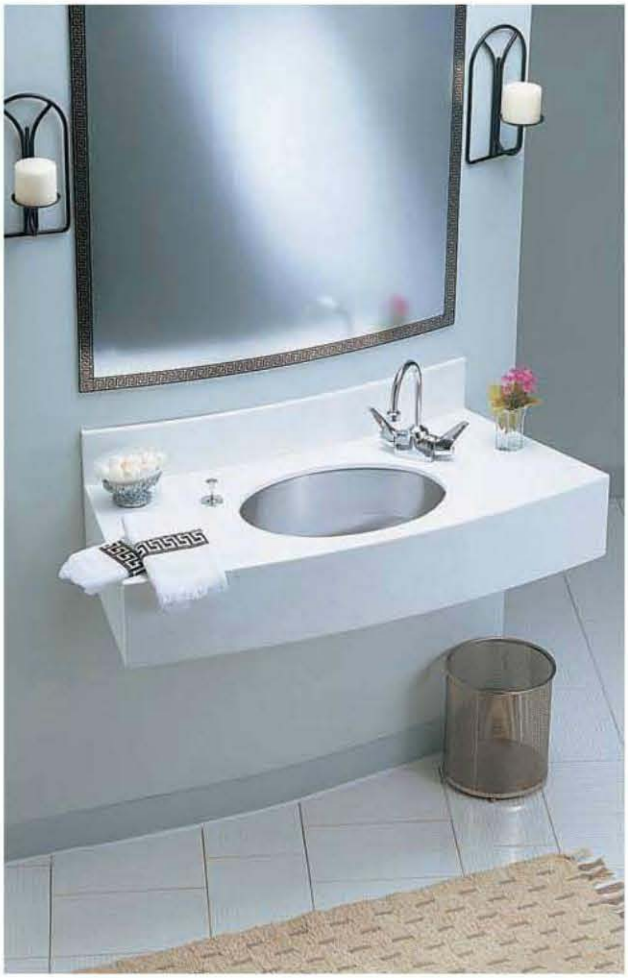

Side-mounted faucet and drain pulls in combination with a wall-mounted sink are easily accessible.

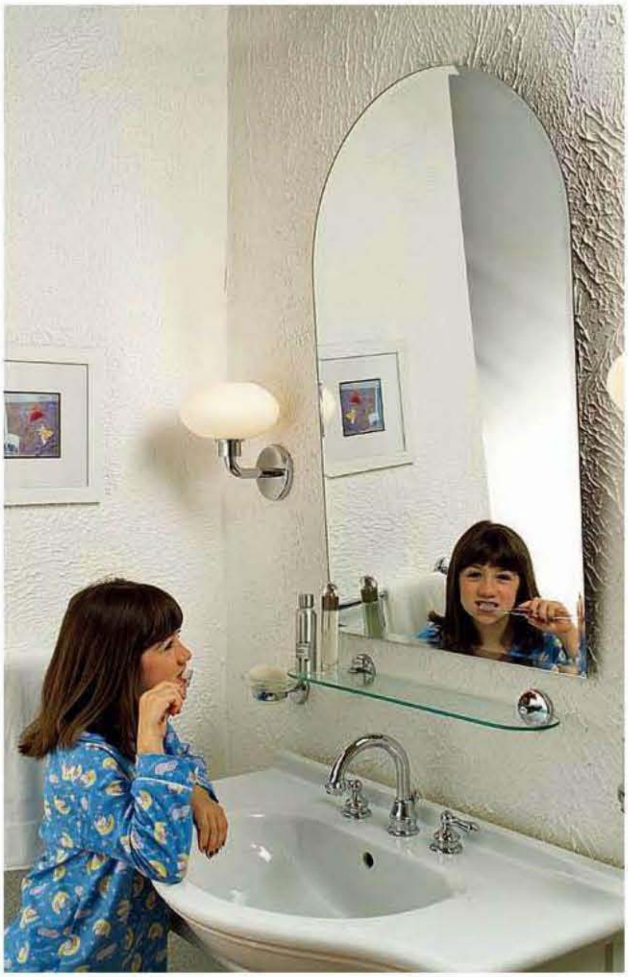

A tilted mirror allows seated or short users to easily see themselves.

Install grab bars in and around the shower and tub. Adding a shower seat or installing a pull-down or permanent seat in the shower allows elderly or disabled family members to sit while bathing. Install a permanently mounted shower seat at 18" high, and be sure the seat is at least 15" deep.

Water controls and faucets on tubs repositioned toward the outside edge at a height of 38 to 48" allows the water to be turned on and adjusted before getting in to bathe. If you're buying a new bathtub, consider one with a side-access door.

Consider replacing a combination tub and shower with a stand-alone shower with a floor that slopes gently toward the drain, rather than a curb to contain water. When fitted with a pull-down seat and an adjustable handheld showerhead, roll-in showers can accommodate people with a wide range of abilities.

Reverse door hinges so the doors open out. That way, the door swing won't be blocked if someone falls.

If needed, widen doorways to 32 to 36" so wheelchair and walker users can enter the bathroom easily. Or, replace a swing door with a pocket door to gain clear space.

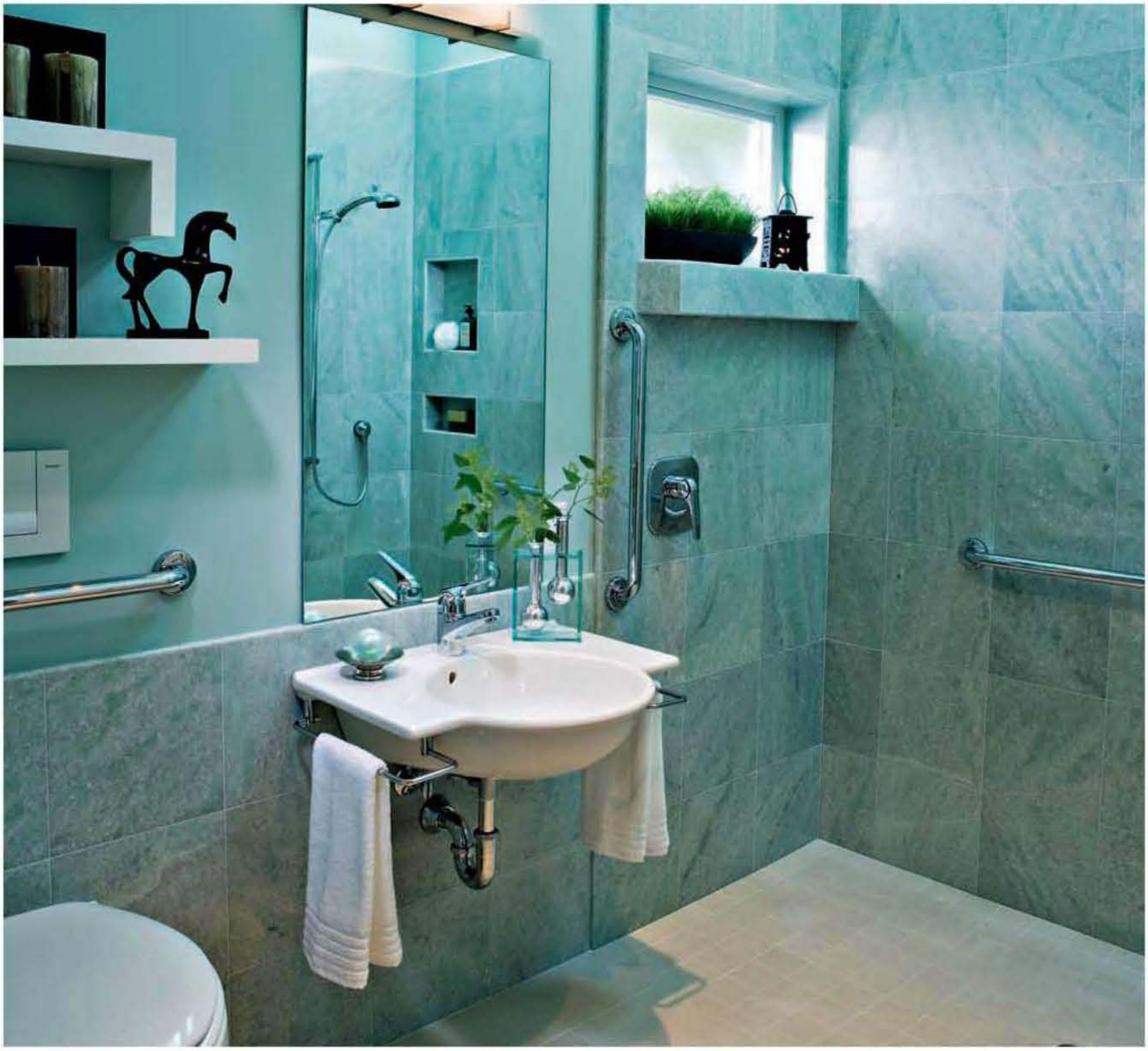

Roll-in shower designs, roll-under sinks, grab bars, adjustable slide-bar showers, and open floor space make bathing easier for people with disabilities.

(continued)

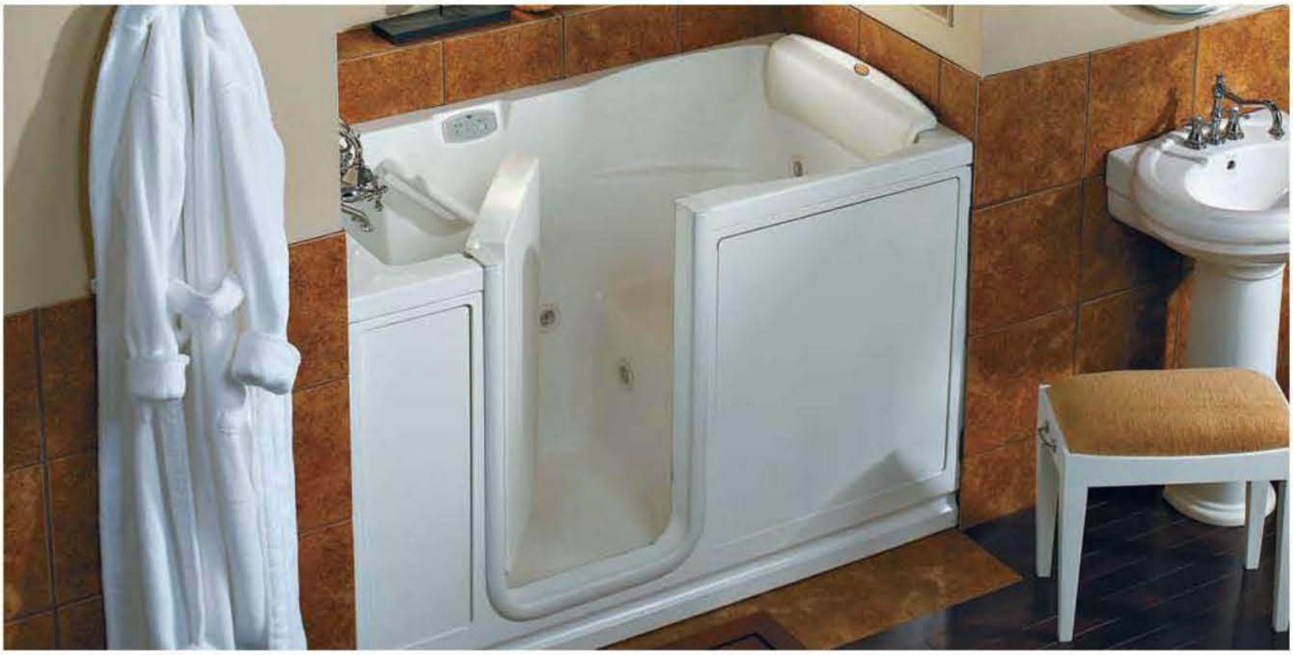

A swing-in door makes bathtubs accessible to practically anyone. The door creates a tight seal when closed and cannot be opened if the tub is full.

Install a pocket door in your bathroom to gain more clearance room for wheelchairs and walkers. Pocket doors do not require door stop molding, allowing for some additional clearance.

Recommended Clearances ▸

A bathroom should be planned with enough approach space and clearance room to allow a wheelchair or walker user to enter and turn around easily. The guidelines for approach spaces (patterned areas) and clearances shown here include some ADA guidelines and recommendations from universal design specialists.

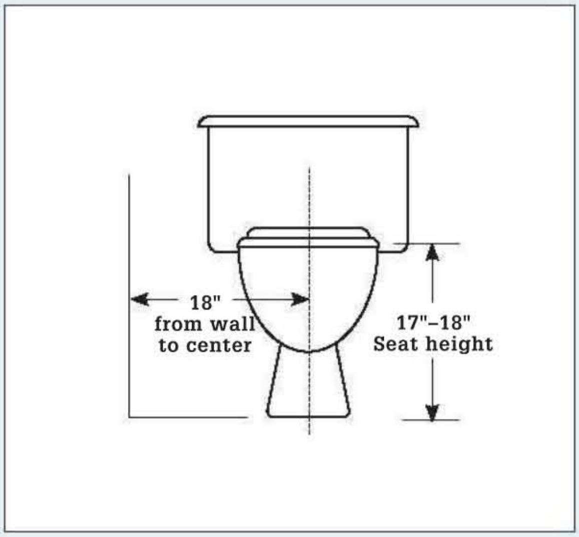
18"
from wall
to center
17"–18"
Seat height

Toilet

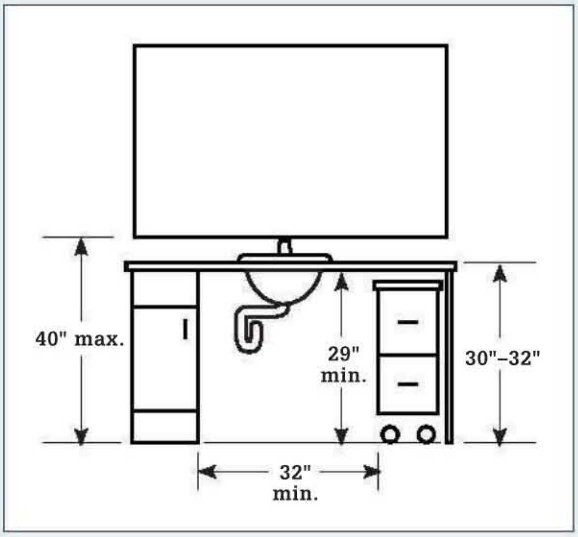
40" max.
29"
min.
30"–32"
32"
min.

Sink & Vanity

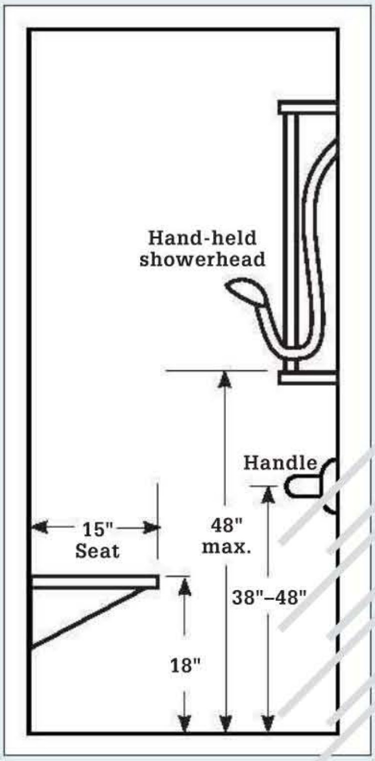
Hand-held
showerhead
Handle
15"
Seat
48"
max.
38"–48"
18"

Shower

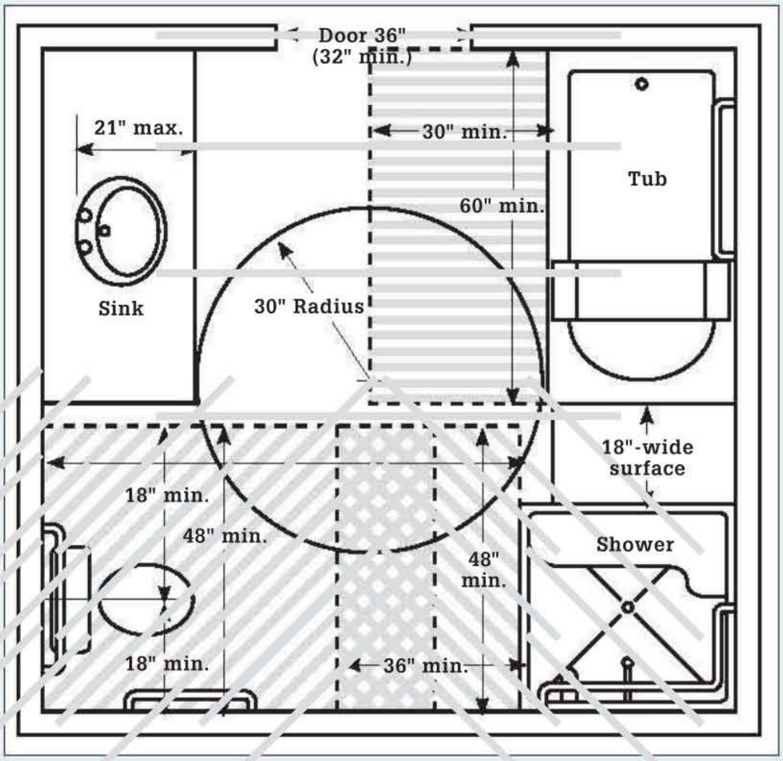
Door 36"
(32" min.)
21" max.
30" min.
Tub
60" min.
Sink
30" Radius
18"-wide
surface
18" min.
48" min.
48"
min.
Shower
18" min.
36" min.

Floor Plan

Installing Shower Kits

The fastest and easiest way to create a new shower in your bathroom is to frame in the stall area with lumber and wallboard and then install a shower enclosure kit. Typically consisting of three fiberglass or plastic walls, these enclosure kits snap together at the corners and nestle inside the flanges of the shower pan to create nearly foolproof mechanical seals. Often, the walls are formed with shelves, soap holders, and other conveniences.

If you are on a tight budget, you can find extremely inexpensive enclosure kits to keep costs down. You can even create your own custom enclosure using waterproof beadboard panels and snap-together connectors. Or, you can invest in a higher grade kit made from thicker material that will last much longer. Some kits are sold with the receptor (and perhaps even the door) included. The kit shown here is designed to be attached directly to wall studs, but others require a backer wall for support. The panels are attached to the backer with high-tack panel adhesive.

A paneled shower surround is inexpensive and easy to install. Designed for alcove installations, they often are sold with matching shower pans (called *receptors*).

Tools & Materials ▸

- Tape measure
- Pencil
- Hammer
- Carpenter's square
- Screwdrivers
- Pipe wrench
- Level
- Strap wrench
- Adjustable wrench
- Pliers
- Drill/driver
- Center punch
- File
- Utility knife
- Hacksaw
- Masking tape
- Silicone caulk and caulk gun
- Shower enclosure kit
- Shower door
- Showerhead
- Faucet
- Plumbing supplies
- Panel adhesive
- Spud wrench
- Large-head roofing nails
- Jigsaw
- Duct tape
- Miter box

How to Install a Shower Enclosure

1

Remove the wallcovering and floor covering in the installation area. Dispose of the materials immediately and thoroughly clean the area. *Note: Most kits can be installed over wallboard, but more professional-looking results can be achieved if it is removed.*

2

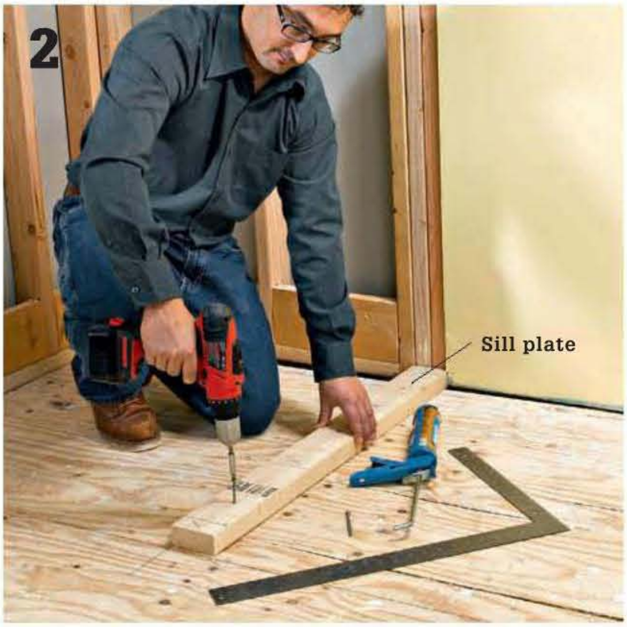

If you are adding a wall to create the alcove, lay out the locations for the studs and plumbing on the new wood sill plate. Also lay out the stud locations on the cap plate that will be attached to the ceiling. Refer to the enclosure kit instructions for exact locations and dimensions of studs. Attach the sill plate to the floor with deck screws and panel adhesive, making sure it is square to the back wall and the correct distance from the side wall.

3

Align a straight 2 × 4 right next to the sill plate and make a mark on the ceiling. Use a level to extend that line directly above the sill plate. Attach the cap plate at that point.

4

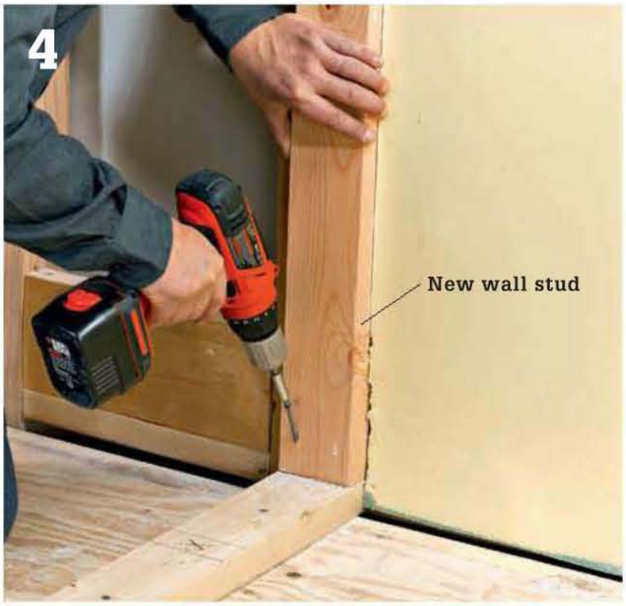

Install the 2 × 4 studs at the outlined locations. Check with a level to make sure each stud is plumb, and then attach them by driving deck screws toenail style into the sill plate and cap plate.

(continued)

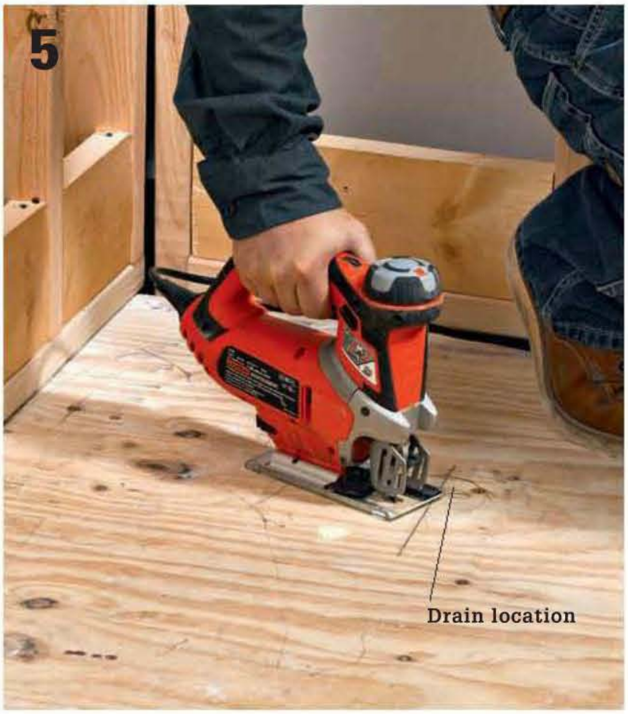

Cut an access hole in the floor for the drain, according to the installation manual instructions. Drill openings in the sill plate of the wet wall (the new wall in this project) for the supply pipes, also according to the instructions.

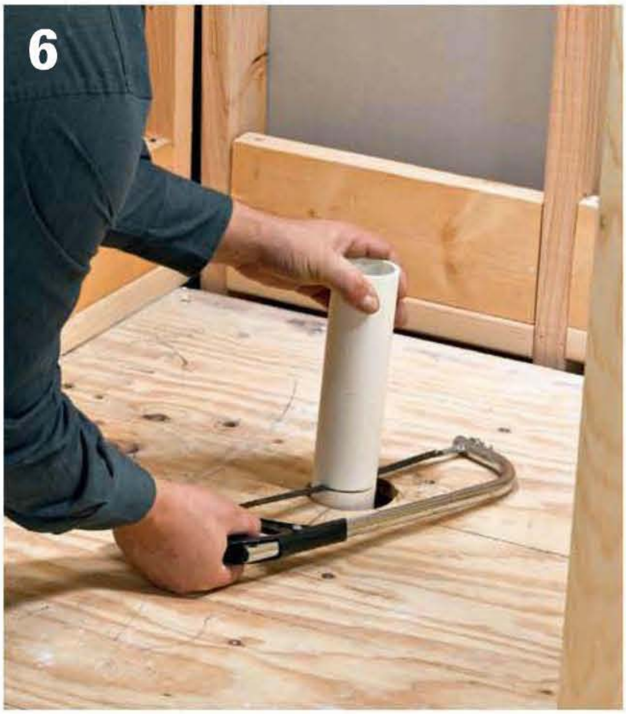

Install a drain pipe and branch line and then trim the drain pipe flush with the floor. If you are not experienced with plumbing, hire a plumber to install the new drain line.

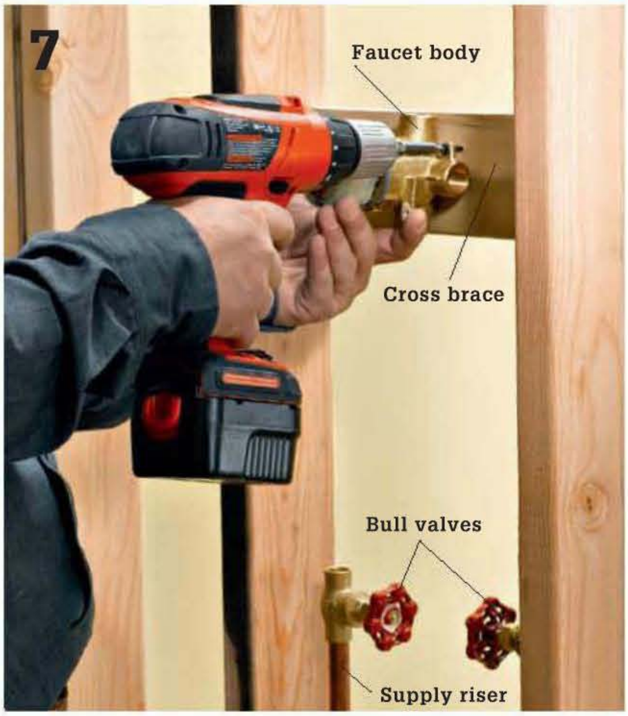

Install new supply risers as directed in the instruction manual (again, have a plumber do this if necessary). Also install cross braces between the studs in the wet wall for mounting the faucet body and shower arm.

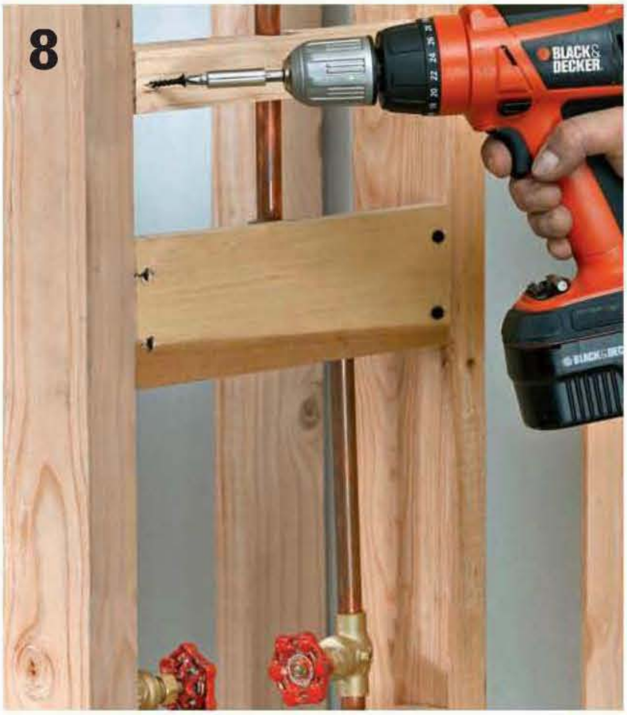

If the supply plumbing is located in a wall (old or new) that is accessible from the non-shower side, install framing for a removable access panel.

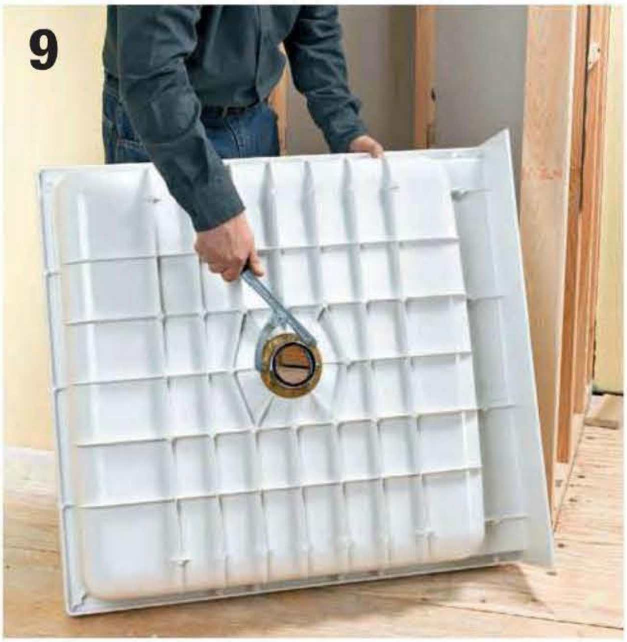

Attach the drain tailpiece that came with your receptor to the underside of the unit, following the manufacturer's instructions precisely. Here, an adjustable spud wrench is being used to tighten the tailpiece.

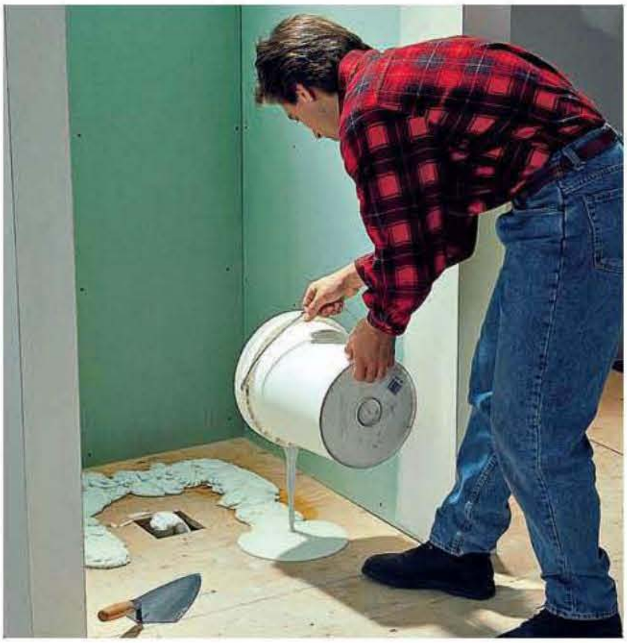

Option: To stabilize the receptor, especially if the floor is uneven, pour or trowel a layer of thinset mortar into the installation area, taking care to keep the mortar out of the drain access hole. Do not apply mortar in areas where the receptor has feet that are intended to make full contact with the floor.

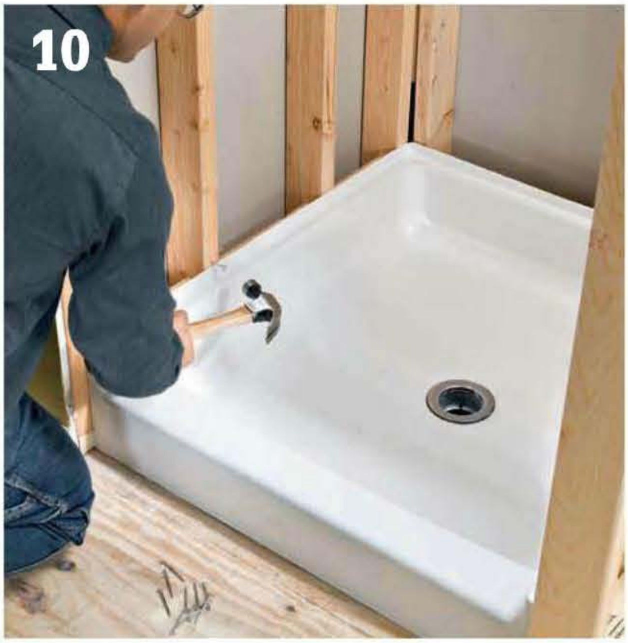

Set the receptor in place, check to make sure it is level, and shim it if necessary. Secure the receptor with large-head roofing nails driven into the wall stud so the nailheads pin the flange against the stud. Do not overdrive the nails.

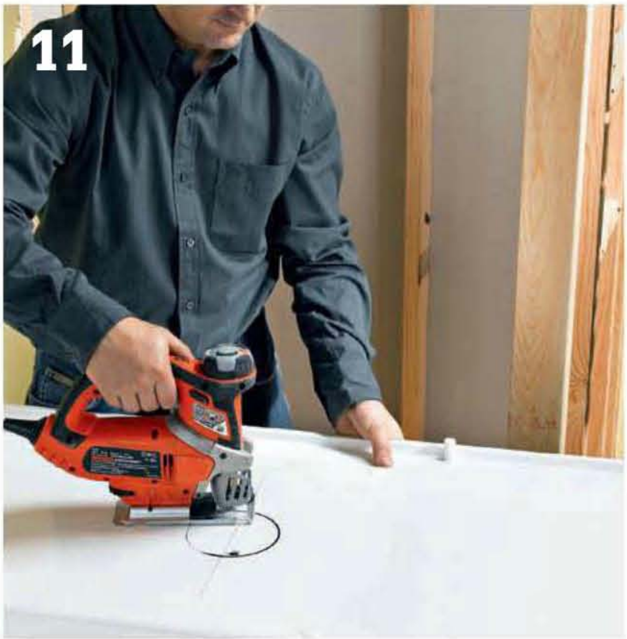

Lay out the locations for the valve hole or holes in the end wall panel that will be installed on the wet wall. Check your installation instructions. Some kits come with a template marked on the packaging carton. Cut the access hole with a hole saw and drill or with a jigsaw and fine-tooth blade. If using a jigsaw, orient the panel so the finish surface is facing down.

(continued)

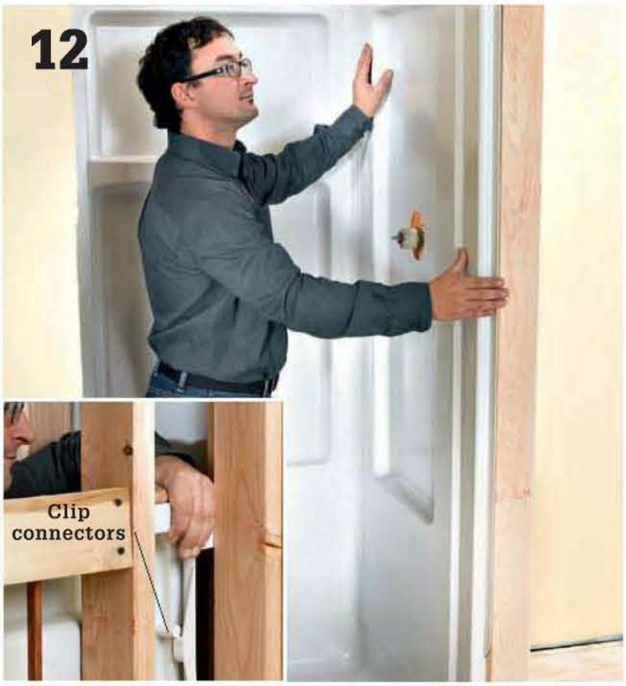

Position the back wall so there is a slight gap (about 1⁄32") between the bottom of the panel and the rim of the receptor—set a few small spacers on the rim if need be. Tack a pair of roofing nails above the top of the back panel to hold it in place (or, use duct tape). Position both end walls and test the fits. Make clip connections between panels (inset) if your kit uses them.

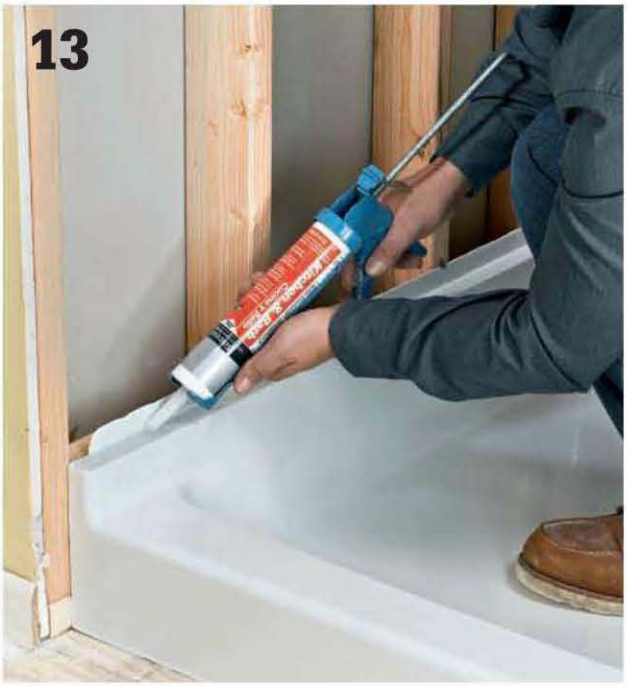

Remove the end walls so you can prepare the installation area for them. If your kit recommends panel adhesive, apply it to the wall or studs. In the kit shown here, only a small bead of silicone sealant on the receptor flange is required.

Reinstall the end panels, permanently clipping them to the back panel according to the kit manufacturer's instructions. Make sure the front edges of the end panels are flush with the front of the receptor.

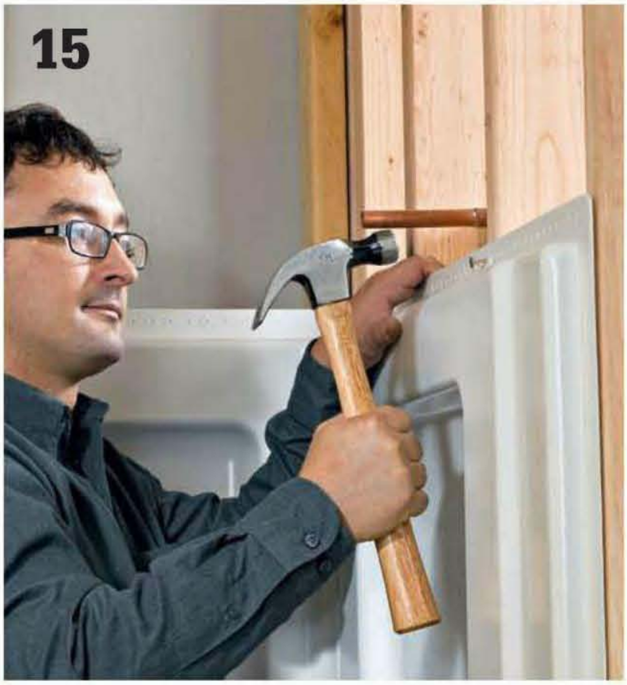

Once the panels are positioned correctly and snapped together, fasten them to the wall studs. If the panels have predrilled nail holes, drive roofing nails through them at each stud at the panel tops and every 4" to 6" along vertical surfaces.

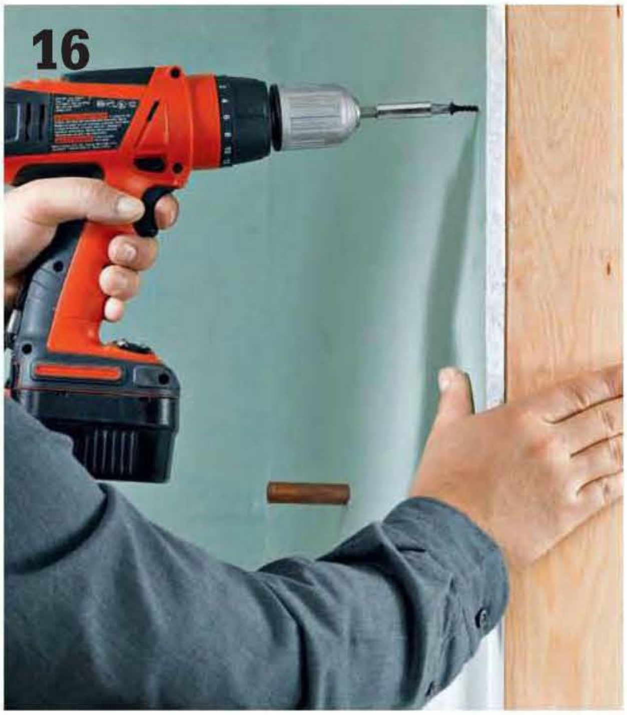

Install wallcovering material above the enclosure panels and anywhere else it is needed. Use moisture-resistant materials, and maintain a gap of ¼" between the shoulders of the top panel flanges and the wallcovering.

17
Kitchen & Bath
Cocina y Baño

Finish the walls and then caulk between the enclosure panels and the wallcoverings with silicone caulk.

Install the faucet handles and escutcheon and caulk around the escutcheon plate. Install the shower arm escutcheon and showerhead.

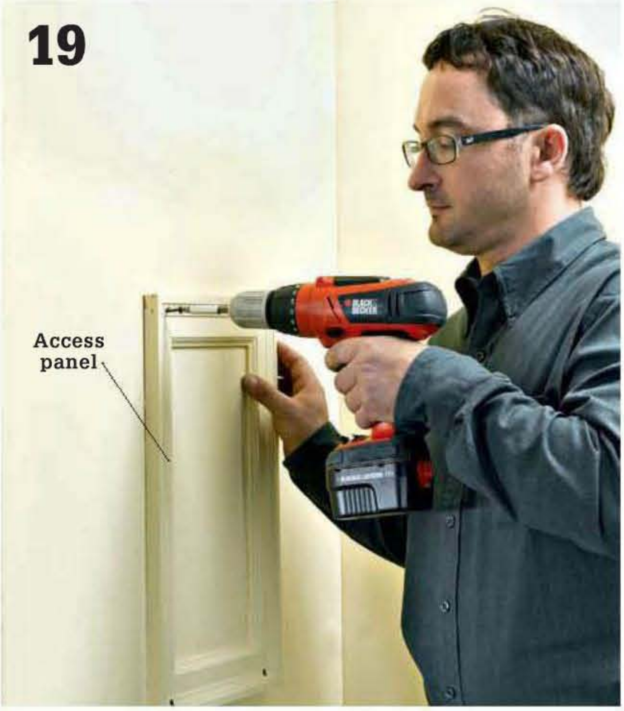

Make an access panel and attach it at the framed opening created in step 8. A piece of ¼" plywood framed with mitered case molding and painted to match the wall is one idea for access panel covers.

How to Install a Hinged Shower Door

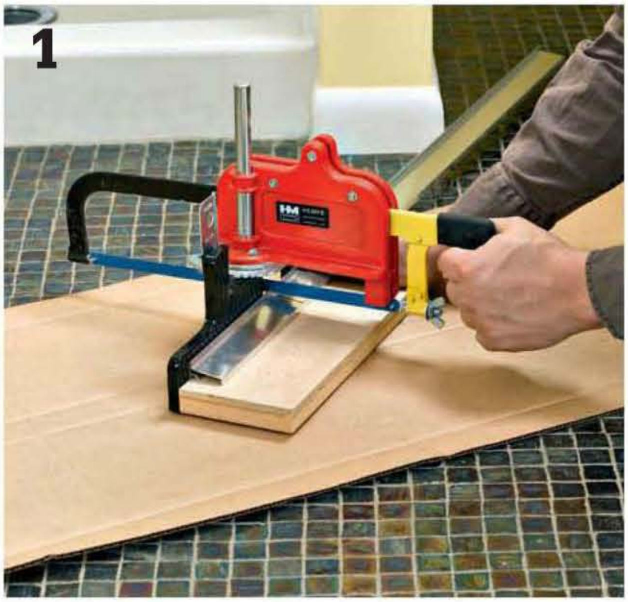

Measure the width of the shower opening. If the walls of the shower slope inward slightly before meeting the base, take your measurement from a higher point at the full width of the opening so you don't cut the door base too short. Cut the base piece to fit using a hacksaw and a miter box. File the cut ends if necessary to deburr them.

Identify which side jamb will be the hinge jamb and which will be the strike jamb according to the direction you want your hinged door to swing—an outward swing is preferred. Prepare the jambs for installation as directed in your instructions.

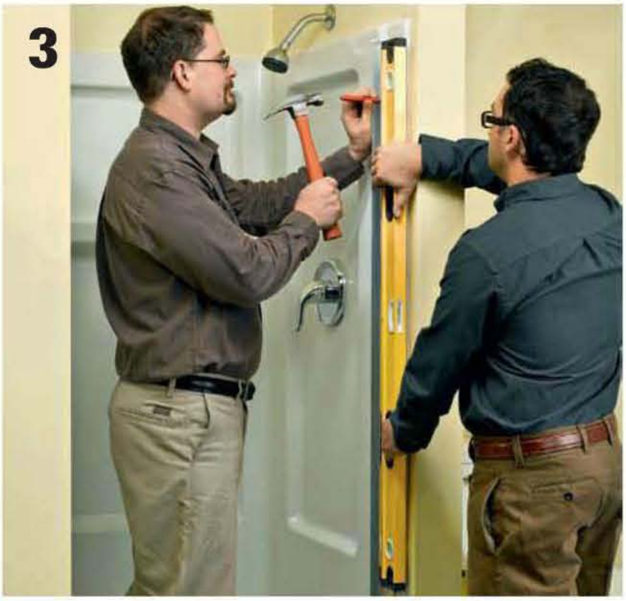

Place the base jamb on the curb of the shower base. If the joint where the wall meets the curb is sloped, you'll need to trim the corners of the base piece to follow the profile. Place a jamb carefully onto the base and plumb it with a level. Then, mark a drilling point by tapping a centerpunch in the middle of each nail hole in each jamb. Remove the jambs, drill pilot holes, and then attach the jambs with the provided screws.

Remove the bottom track and prepare the shower base curb for installation of the base track, following the manufacturer's directions. Permanently install the bottom track. Bottom tracks (not all doors have them) are usually attached to the side jambs or held in place with adhesive. Never use fasteners to secure them to curb.

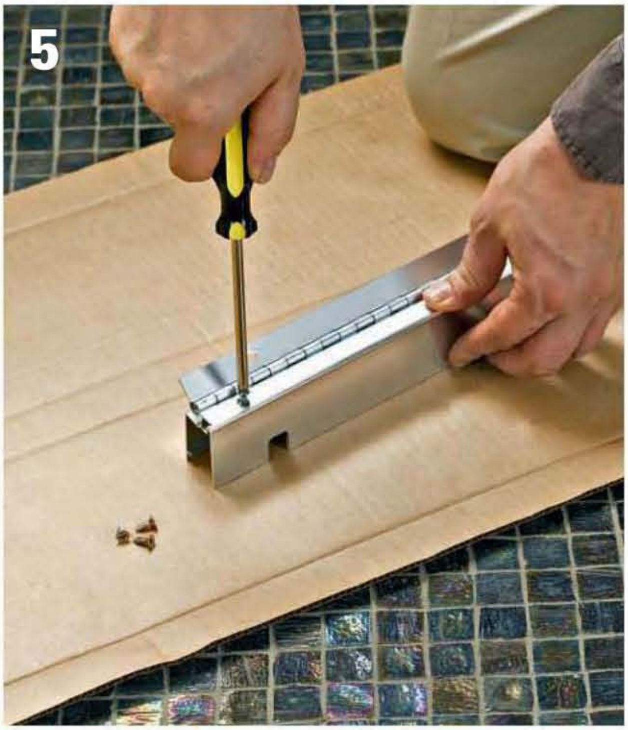

Working on the floor or another flat surface, attach the door hinge to the hinge jamb, if required. In most systems, the hinge is fitted over the hinge jamb after you attach it to the wall.

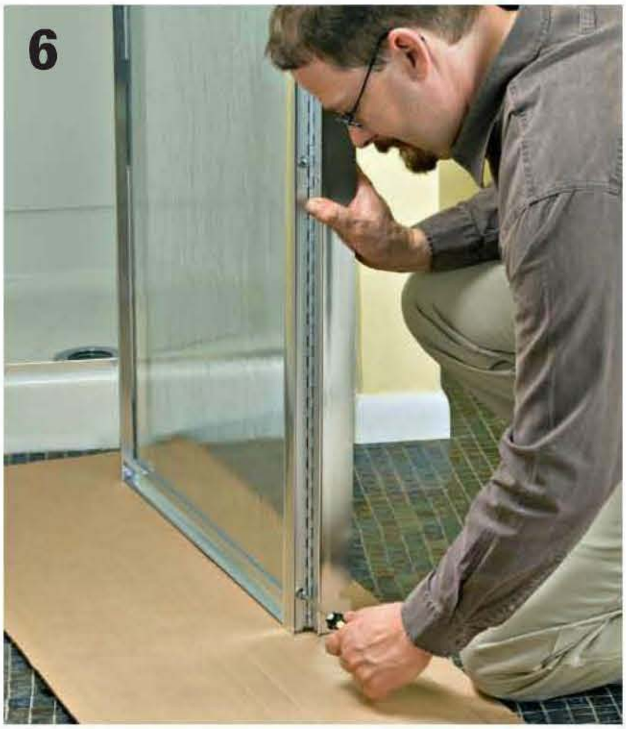

Attach the hinge to the door panel, according to the manufacturer's instructions. Attach any cap fitting that keeps water out of the jamb.

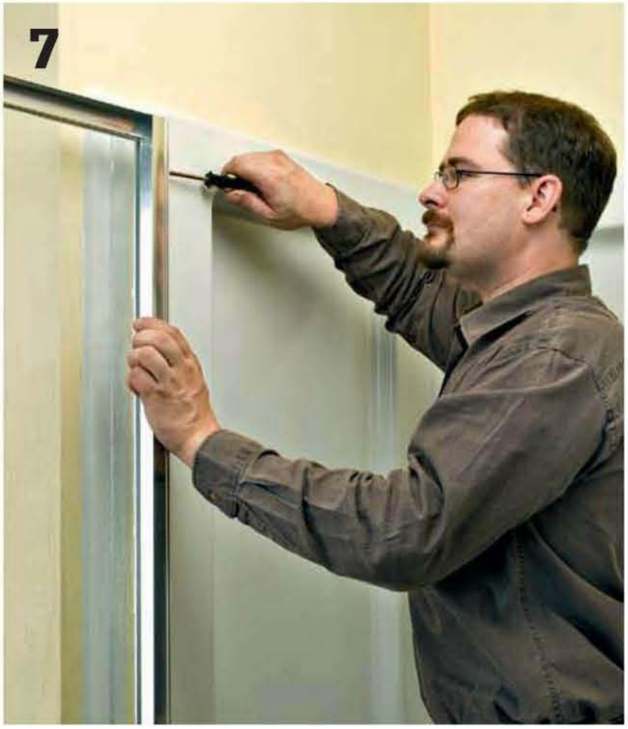

Fit the hinge jamb over the side jamb and adjust it as directed in your instruction manual. Once the clearances are correct, fasten the jambs to hang the door.

Install the magnetic strike plate and any remaining caps or accessories such as towel rods. Also attach the sweep that seals the passage, if provided.

Installing Alcove Bathtubs

Many of our homes are equipped with an alcove tub that includes a tub surround and shower feature. By combining the tub and the shower in one fixture, you conserve precious bathroom floorspace and simplify the initial installation. Plus, you only have one bathing fixture that needs cleaning.

But because tub/showers are so efficient, they do get a lot of use and tend to have fairly limited lifespans. The fact that the most inexpensive tubs on the market are designed for alcove use also reduces the average tub/shower lifespan. Pressed steel tubs have enamel finishes that crack and craze; plastic and fiberglass tubs get grimy and stained; even acrylic and composite tubs show wear eventually (and as with other fixtures, styles, and colors change too).

Plumbing an alcove tub is a relatively difficult job because getting access to the drain lines attached to the tub and into the floor is often very awkward. Although an access panel is required by most codes, the truth is that many tubs were installed without them or with panels that are too small or hard to reach to be of much use. If you are contemplating replacing your tub, the first step in the decision process should be to find the access panel and determine if it is sufficient. If it is not (or there is no panel at all), consider how you might enlarge it. Often, this means cutting a hole in the wall on the adjoining room and also in the ceiling below. This creates more work, of course, but compared to the damage caused by a leaky drain from a subpar installation, making an access opening is a small inconvenience.

Tools & Materials ▸

- Channel-type pliers
- Hacksaw
- Carpenter's level
- Pencil
- Tape measure
- Saw
- Screwdriver
- Drill
- Adjustable wrench
- Trowel
- Shims
- Galvanized deck screws
- Drain-waste-overflow kit
- 1 × 3, 1 × 4, 2 × 4 lumber
- Galvanized roofing nails
- Galvanized roof flashing
- Thinset mortar
- Tub & tile caulk
- Propane torch

By replacing a dingy old alcove tub with a fresh new one, you can make the tub and shower area as pleasant to use as it is efficient.

Tips for Installing Bathtubs ▸

Choose the correct tub for your plumbing setup. Alcove-installed tubs with only one-sided aprons are sold as either "left-hand" or "right-hand" models, depending on the location of the predrilled drain and overflow holes in the tub. To determine which type you need, face into the alcove and check whether the tub drain is on your right or your left.

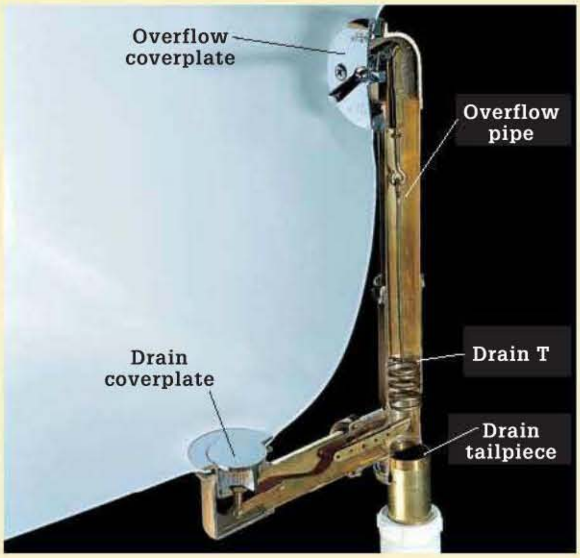
Overflow coverplate
Overflow pipe
Drain coverplate
Drain T
Drain tailpiece

A Drain-waste-overflow kit with stopper mechanism must be purchased separately and attached after the tub is set. Available in both brass and plastic types, most kits include an overflow coverplate, an overflow pipe that can be adjusted to different heights, a drain T-fitting, an adjustable drain tailpiece, and a drain coverplate that screws into the tailpiece.

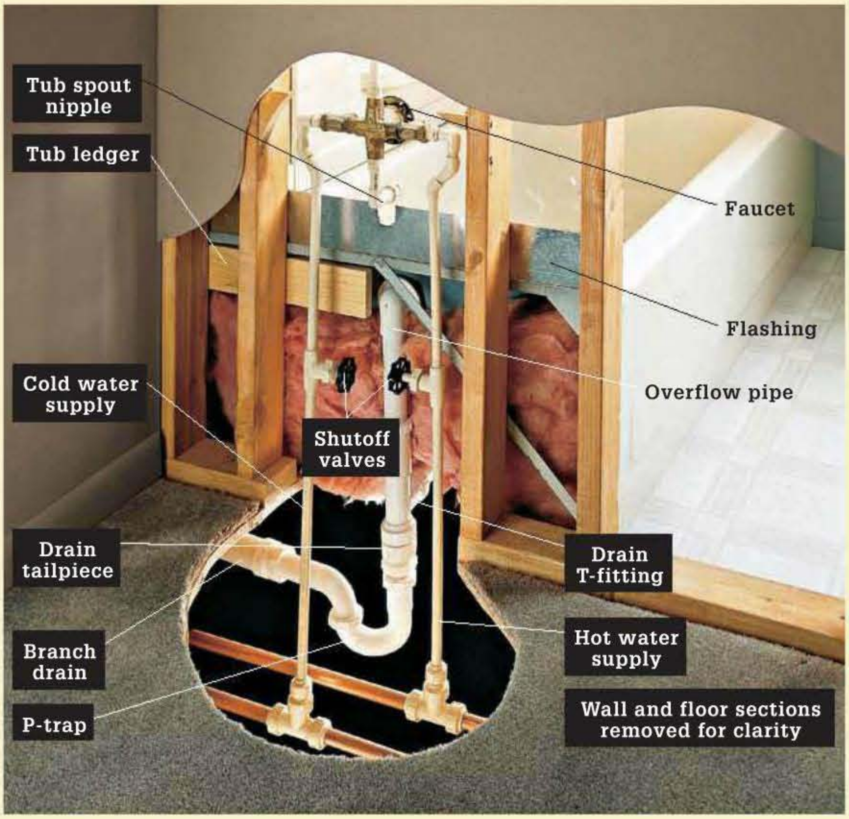
Tub spout nipple
Tub ledger
Faucet
Flashing
Cold water supply
Overflow pipe
Shutoff valves
Drain tailpiece
Drain T-fitting
Branch drain
Hot water supply
P-trap
Wall and floor sections removed for clarity

The supply system for a bathtub includes hot and cold supply pipes, shutoff valves, a faucet and handle(s), and a spout. Supply connections can be made before or after the tub is installed.

The drain-waste-overflow system for a bathtub includes the overflow pipe, drain T, P-trap, and branch drain. The overflow pipe assembly is attached to the tub before installation.

How To Remove an Alcove Bathtub

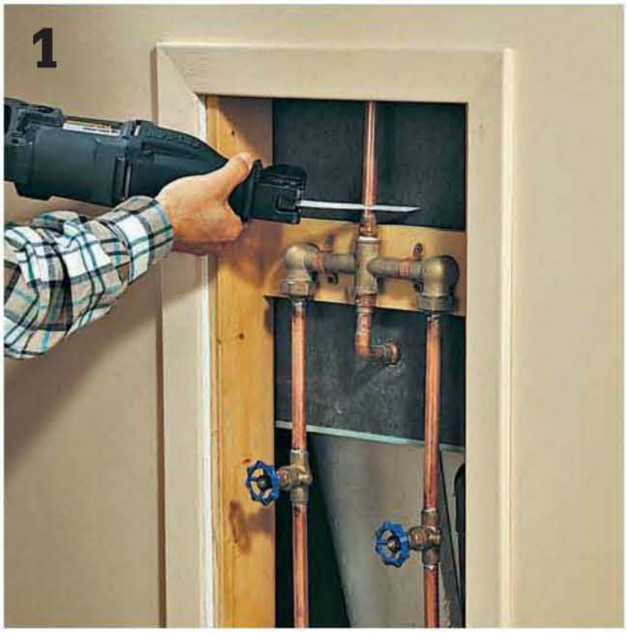

Cut the old supply tubes, if you have access to them, with a reciprocating saw and metal cutting blade or with a hacksaw. Be sure to shut off the water supply at the stop valves first. Cut the shower pipe just above the faucet body and cut the supply tubes just above the stop valves.

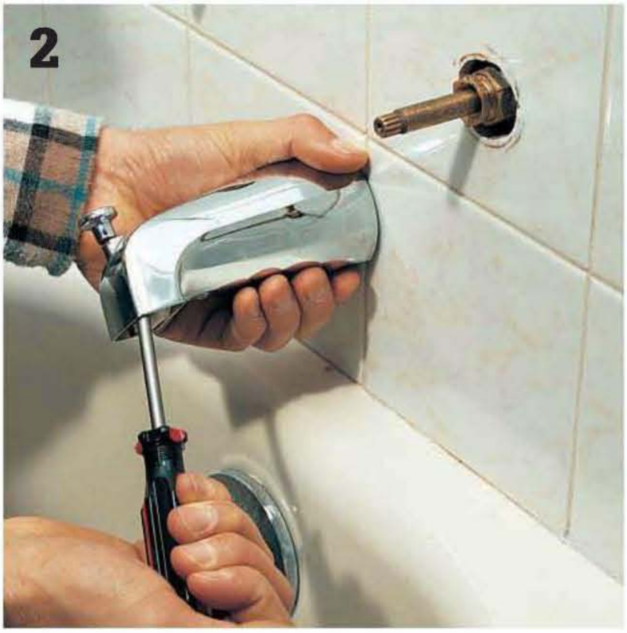

Remove the faucet handles, tub spout, shower head and escutcheon, and arm. For the spout, check the underside for a set screw and loosen it if you find one. Then, insert a long screwdriver into the spout and turn the spout counterclockwise.

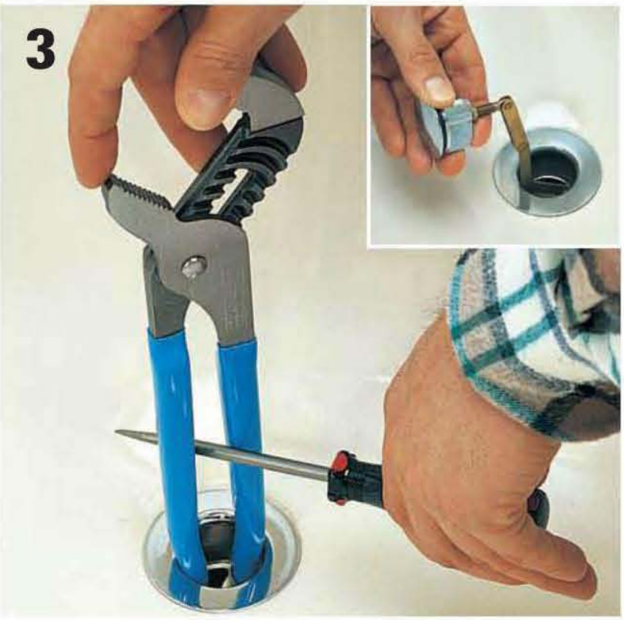

Remove the drain plug, working from the tub side. If the tub has a pop-up drain with linkage, twist the plug to disengage the linkage and remove the plug (inset). Then, insert the handles of a channel-type pliers into the drain opening and past the drain crosspiece. Twist the pliers counterclockwise to remove the plug.

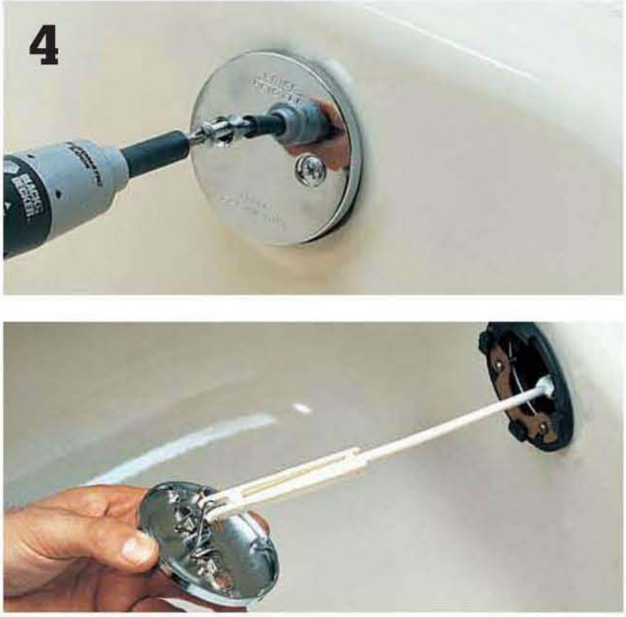

Remove the overflow coverplate (top photo) and then withdraw the pop-up drain linkage through the overflow opening (lower photo).

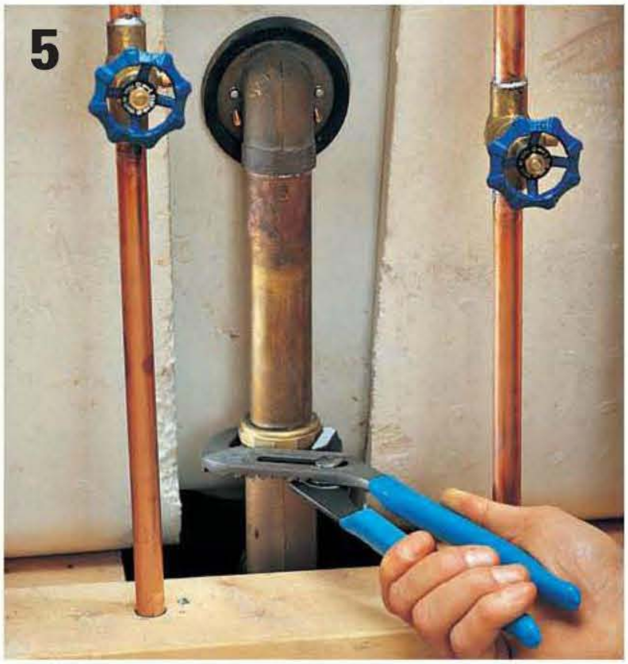

Disconnect the overflow pipe from the drain assembly and remove both parts (your access may not be as unrestricted as seen here). If you need to cut the pipes, go ahead and do it. In most cases, it is difficult to maneuver the tub out with the DWO assembly still attached.

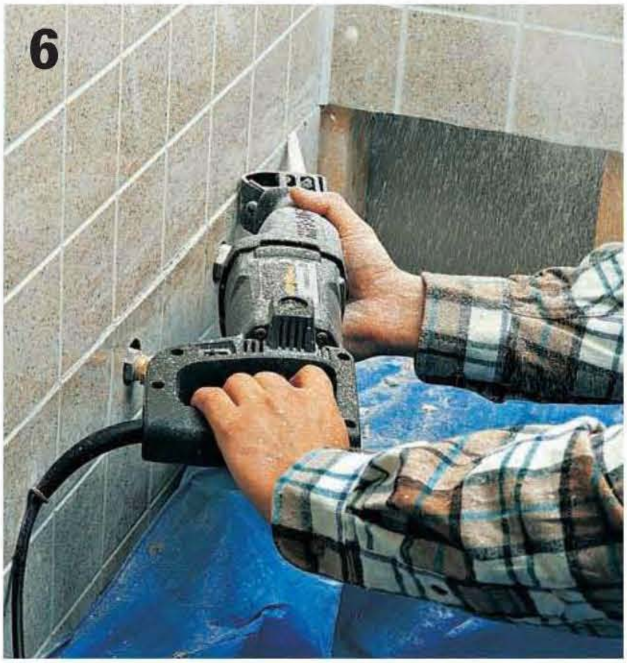

Cut the wall to a line about 6" above the tub rim. Alcove tubs are fastened to the wall studs with nails driven through or above a flange that extends up from the rim. You'll need to remove a bit of the wall covering so you can remove the fasteners.

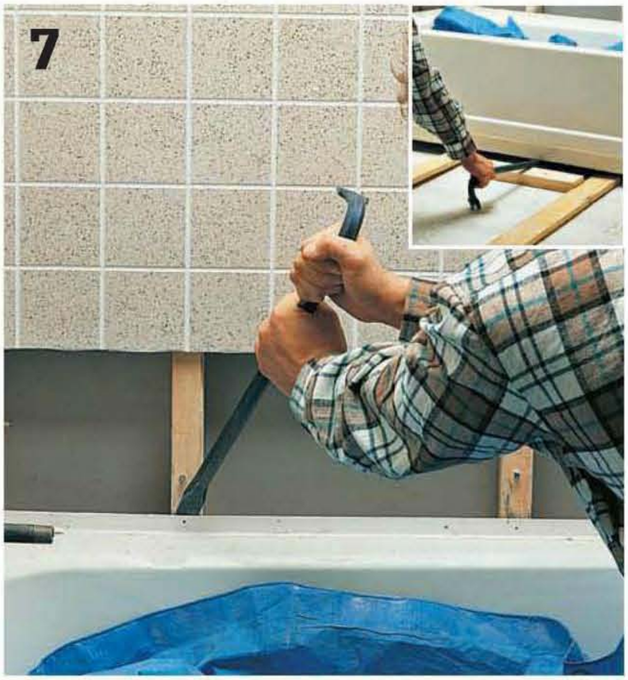

If you can, pry out fasteners and then pull the tub away from the walls by levering between the back rim of the tub and the back wall of the alcove. If it resists, check for adhesive caulk or even flooring blocking the bottom of the apron. If needed, raise the tub and slide a pair of 1 × 4 runners under the skirt edge (inset photo) to make it easier to slide out.

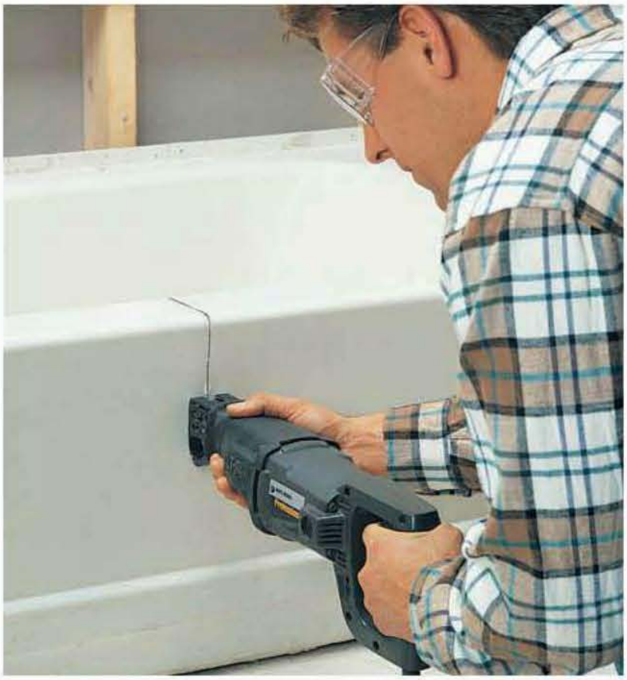

Option: Cut stubborn tubs in half to wrangle them out of the alcove. This has the added benefit of making the tubs easier to get out the door, down the stairs, and into the dumpster.

How to Install a New Alcove Tub

Prepare for the new tub. Inspect and remove old or deteriorated wall surfaces or framing members in the tub area. With today's mold-resistant wallboard products, it makes extra sense to go ahead and strip off the old alcove wallboard down to the studs so you can replace them. This also allows you to inspect for hidden damage in the wall and ceiling cavities.

Check the subfloor for level—if it is not level, use pour-on floor leveler compound to correct it (ask at your local flooring store). Make sure the supply and drain pipes and the shutoff valves are in good repair and correct any problems you encounter. If you have no bath fan in the alcove, now is the perfect time to add one.

Check the height of the crossbraces for the faucet body and the showerhead. If your family members needed to stoop to use the old shower, consider raising the brace for the showerhead. Read the instructions for your new faucet/diverter and check to see that the brace for the faucet body will conform to the requirements (this includes distance from the surround wall as well as height). Adjust the brace locations as needed.

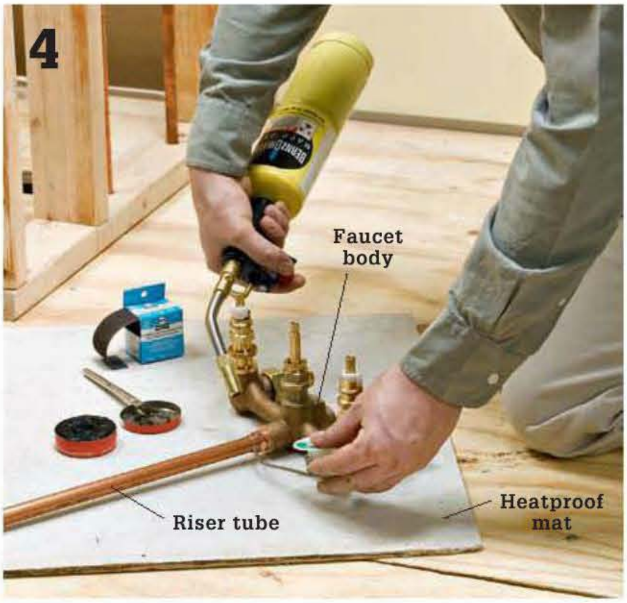

Begin by installing the new water supply plumbing. Measure to determine the required height of your shower riser tube and cut it to length. Solder the bottom of the riser to the faucet body and the top to the shower elbow.

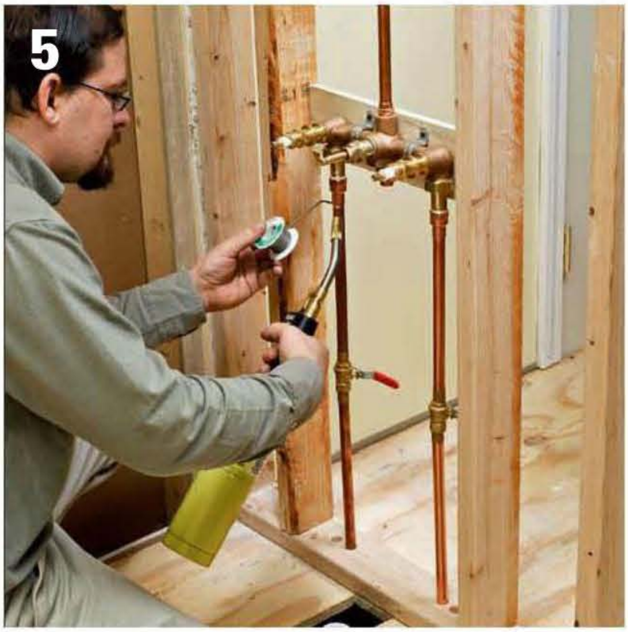
5

Attach the faucet body to the cross brace with pipe hanger straps. Then, attach supply tubing from the stop valves to the faucet body, making sure to attach the hot water to the left port and cold to the right port. Also secure the shower elbow to its cross brace with a pipe strap. Do not attach the shower arm yet.

6

Slide the bathtub into the alcove. Make sure tub is flat on the floor and pressed flush against the back wall. If your tub did not come with a tub protector, cut a piece of cardboard to line the tub bottom, and tape pieces of cardboard around the rim to protect the finish from shoes and dropped tools.

7

Mark locations for ledger boards. To do this, trace the height of the top of the tub's nailing flange onto the wall studs in the alcove. Then remove the tub and measure the height of the nailing flange. Measure down this same amount from your flange lines and mark new ledger board location.

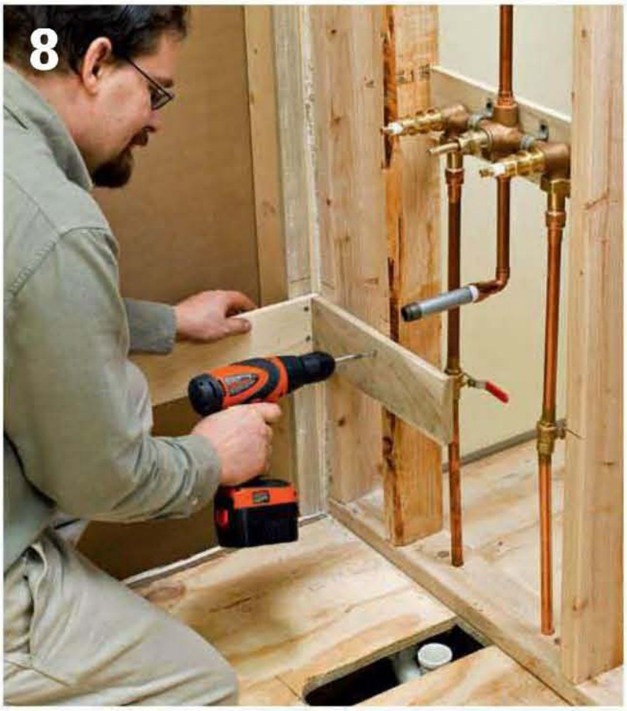
8

Install 1 × 4 ledger boards. Drive two or three 3"-galvanized deck screws through the ledger board at each stud. All three walls should receive a ledger. Leave an open space in the wet wall to allow clearance for the DWO kit.

(continued)

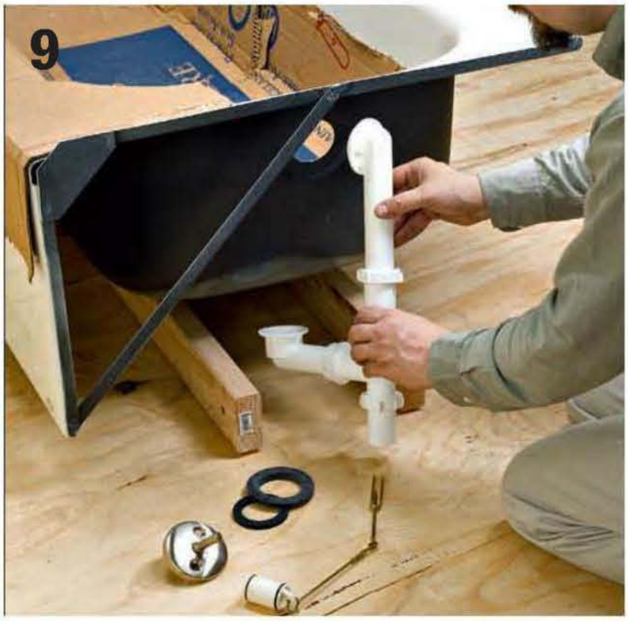

Install the drain-waste-overflow (DWO) pipes before you install the tub. Make sure to get a good seal on the slip nuts at the pipe joints. Follow the manufacturer's instructions to make sure the pop-up drain linkage is connected properly. Make sure rubber gaskets are positioned correctly at the openings on the outside of the tub.

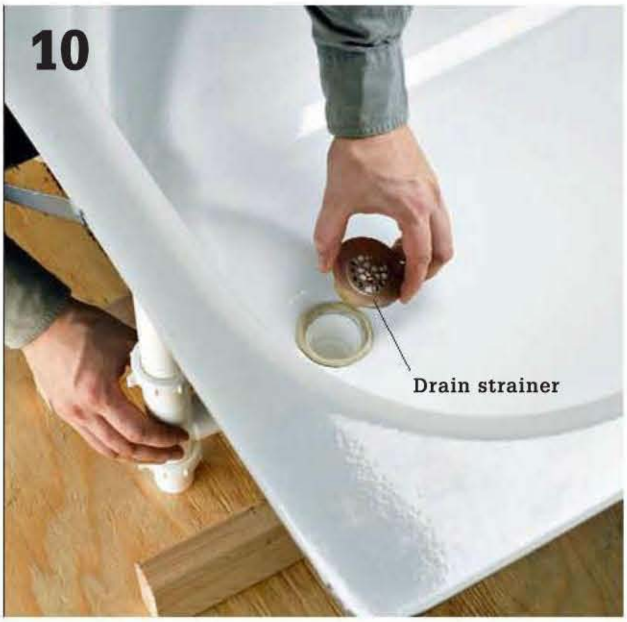

Thread the male-threaded drain strainer into the female-threaded drain waste elbow. Wrap a coil of plumber's putty around the drain outlet underneath the plug rim first. Hand tighten only.

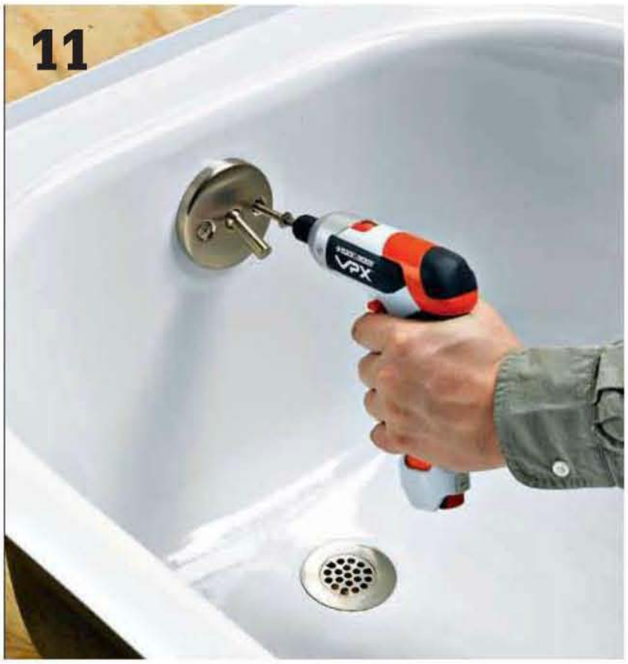

Attach the overflow coverplate, making sure the pop-up drain controls are in the correct position. Tighten the mounting screws that connect to the mounting plate to sandwich the rubber gasket snugly between the overflow pipe flange and the tub wall. Then, finish tightening the drain strainer against the waste elbow by inserting the handle of a pair of pliers into the strainer body and turning.

Place the tub back into the alcove, taking care not to bump the DWO assembly and disturb the connections. You definitely will want a helper for this job. If the drain outlet of the DWO assembly is not directly over the drain pipe when the tub is in position, you'll need to remove it and adjust the drain line location.

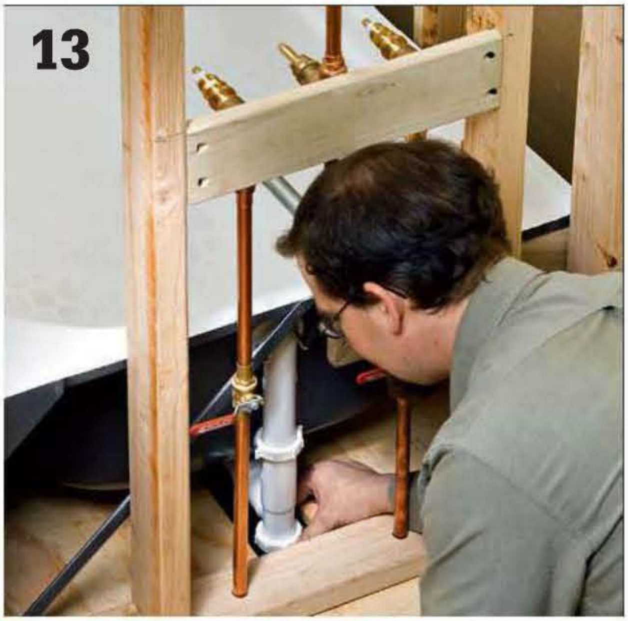

Attach the drain outlet from the DWO assembly to the drain P-trap. This is the part of the job where you will appreciate the time you spent creating a roomy access panel for the tub plumbing. Test the drain and overflow to make sure they don't leak. Also test the water supply plumbing, temporarily attaching the handles, spout, and shower arm so you can operate the faucet and the diverter.

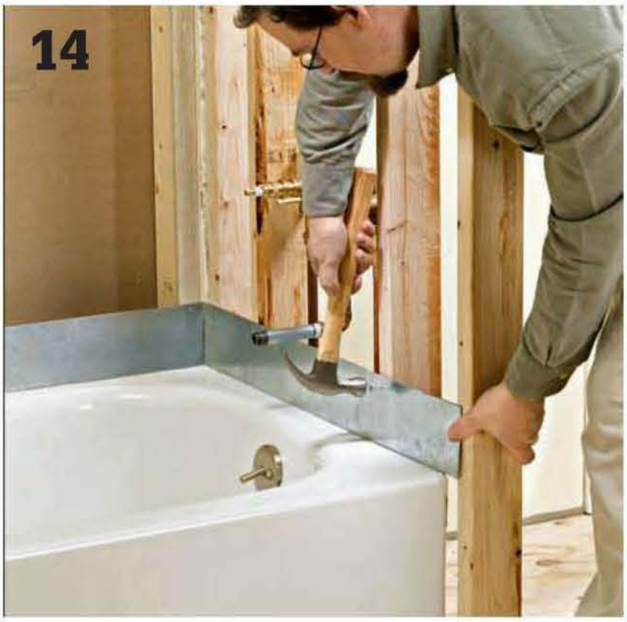

Drive a 1½" galvanized roofing nail at each stud location, just over the top of the tub's nailing flange. The nail head should pin the flange to the stud. Be careful here—an errant blow or overdriving can cause the enameled finish to crack or craze. *Option: You may choose to drill guide holes and nail through the flange instead.*

Install the wallcoverings and tub surround (see pages 446 to 449 for a 3-piece surround installation). You can also make a custom surround from tileboard or cementboard and tile.

Install fittings. First, thread the shower arm into the shower elbow and attach the spout nipple to the valve assembly. Also attach the shower head and escutcheon, the faucet handle/diverter with escutcheon, and the tub spout. Use thread lubricant on all parts.

Installing 3-piece Tub Surrounds

No one wants bathroom fixtures that are aging or yellowed from years of use. A shiny new tub surround can add sparkle and freshness to your dream bath.

Tub surrounds come in many different styles, materials, and price ranges. Choose the features you want and measure your existing bathtub surround for sizing. Surrounds typically come in three or five pieces. A three-panel surround is being installed here, but the process is similar for five-panel systems.

Surface preparation is important for good glue adhesion. Plastic tiles and wallpaper must be removed and textured plaster must be sanded smooth. Surrounds can be installed over ceramic tile that is well attached and in good condition, but it must be sanded and primed. All surfaces must be primed with a water-based primer.

Tools & Materials ▸

- Jigsaw
- Hole saw
- Drill
- Measuring tape
- Level
- Caulking gun
- Primer
- Adhesive
- Screwdriver
- Adjustable wrench
- Pry bar
- Hammer
- 3-piece tub surround

Three-piece tub surrounds are inexpensive and come in many colors and styles. The typical unit has two end panels and a back panel that overlap in the corners to form a watertight seal. They are formed from fiberglass, PVC, acrylic, or proprietary resin-based polymers. Five piece versions are also available and typically have more features such as integral soap shelves and even cabinets.

How to Install a 3-piece Tub Surround

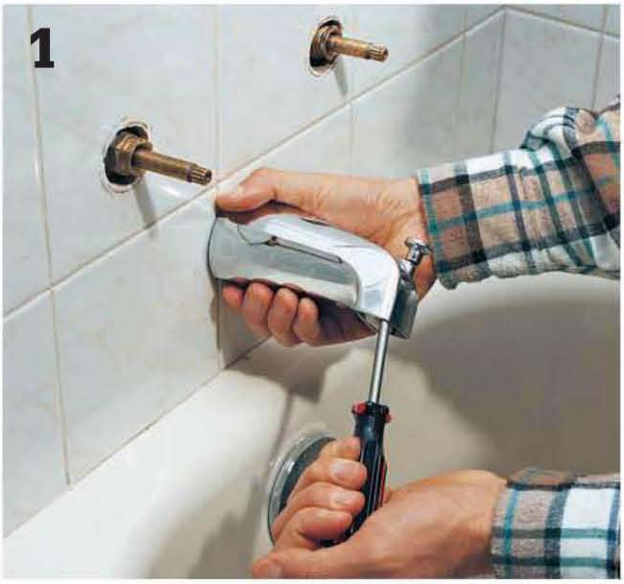

Remove the old plumbing fixtures and wallcoverings in the tub area. In some cases you can attach surround panels to old tileboard or even tile, but it is generally best to remove the wallcoverings down to the studs if you can, so you may inspect for leaks or damage.

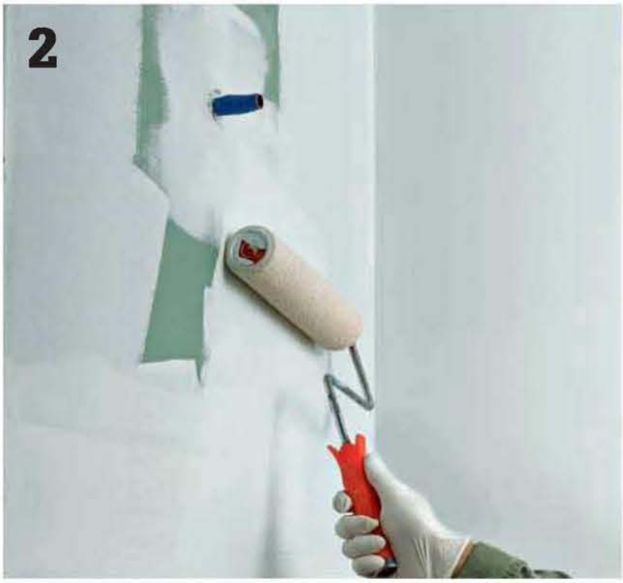

Replace the wallcoverings with appropriate materials, such as water and mold-resistant wallboard or cementboard (for ceramic tile installations). Make sure the new wall surfaces are smooth and flat. Some surround kit manufacturers recommend that you apply a coat of primer to sheet goods such as greenboard to create a better bonding surface for the panel adhesive.

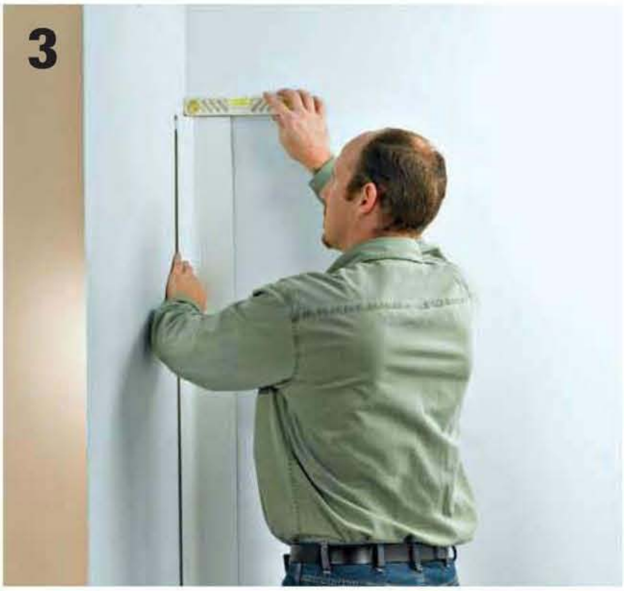

Test-fit the panels before you start; the tub may have settled unevenly or the walls may be out of plumb. Check the manufacturer's directions for distinguishing right and left panels. Place a panel in position on the tub ledge. Use a level across the top of the panel to determine if it is level. Create a vertical reference line to mark the edge of the panel on the plumbing end.

Test-fitting Tip ▸

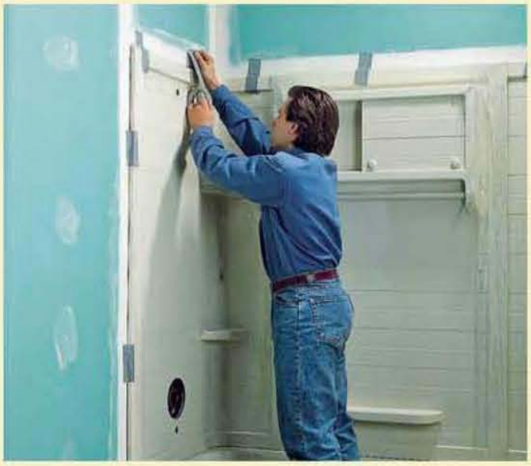

Ensure a perfect fit by taping the surround panels to the walls in the tub area. Make sure the tops are level when the overlap seams are aligned and that you have a consistent ⅛" gap between the panel bottoms and the tub flange. Mark the panels for cutting if necessary and, once the panels have been removed, make necessary adjustments to the walls.

(continued)

4

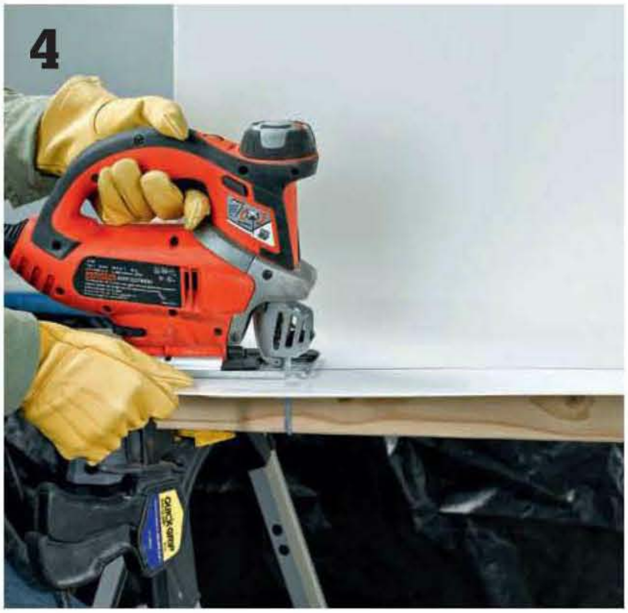

Some kits are created to fit a range of bathtub dimensions. After performing the test fit, check the fitting instructions to see if you need to trim any of the pieces. Follow the manufacturer's instructions for cutting. Here, we had to cut the corner panels because the instructions advise not to overlap the back or side panel over the corner panels by more than 3". Cut panels using a jigsaw and a fine-tooth blade that is appropriate for cutting fiberglass or acrylic tileboard. The cut panels should be overlapped by panels with factory edges.

5

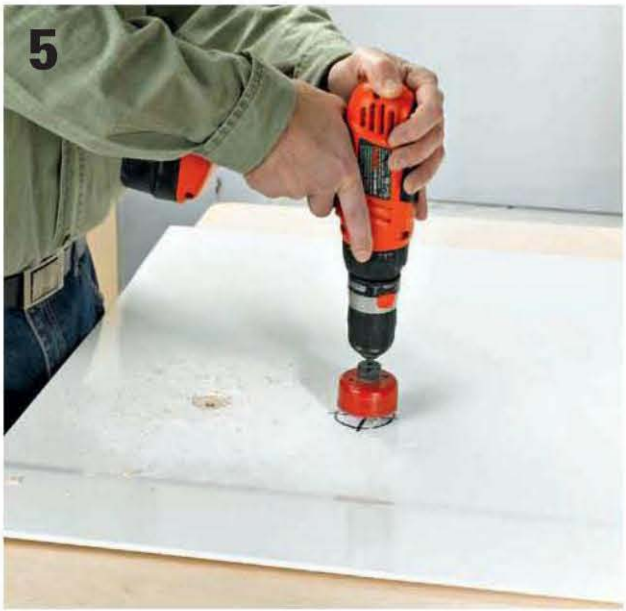

Measure and mark the location of the faucets, spout, and shower outlets. Measure in from the vertical reference line (made in step 3) and up from the top of the tub ledge. Re-measure for accuracy, as any cuts to the surround are final. Place the panel face-up on a sheet of plywood. Mark the location of the holes. Cut the holes ½" larger than the pipe diameter. If your faucet has a recessed trim plate (escutcheon), cut the hole to fit the recess. Using a hole saw or a jigsaw, cut out the plumbing outlets.

6

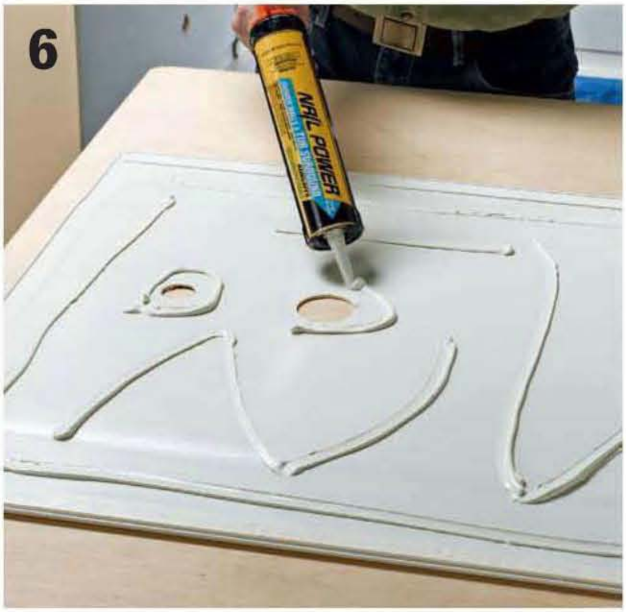

Install the plumbing end panel, test-fitting first. In this surround, the end panels are installed first. Apply adhesive to the back of the plumbing panel. Circle the plumbing outlet holes 1" from the edge. Do not apply adhesive closer than 1" to the double-sided tape (if your kit has tape) or to the bottom edge of the panel.

7

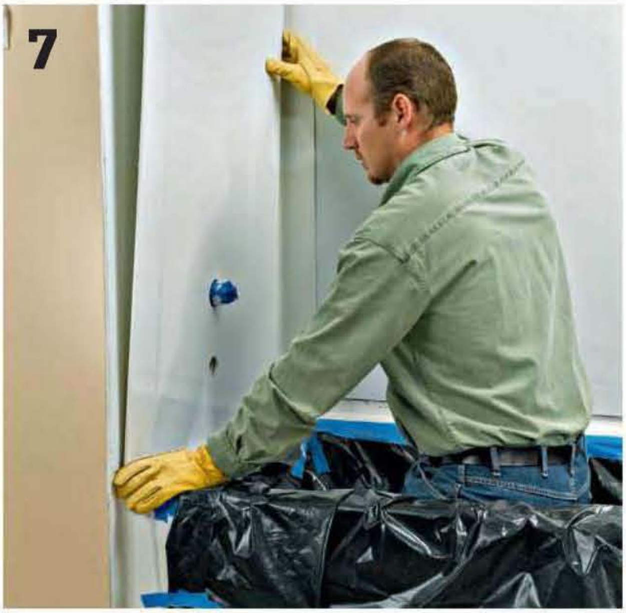

Remove the protective backing from the tape. Carefully lift the panel by the edges and place against the corner and top of the tub ledge. Press firmly from top to bottom in the corner, then throughout the panel.

8

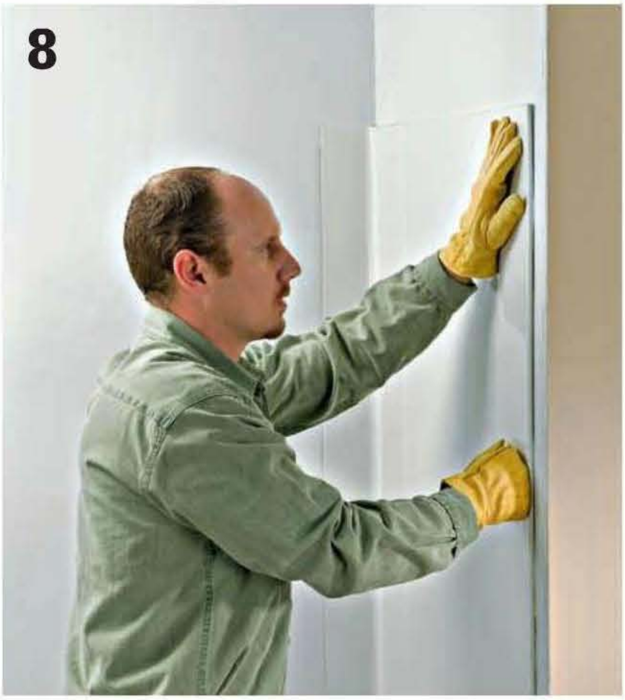

Test-fit the opposite end panel and make any necessary adjustments. Apply the adhesive, remove the protective backing from the tape, and put in place. Apply pressure to the corner first from top to bottom, and then apply pressure throughout.

9

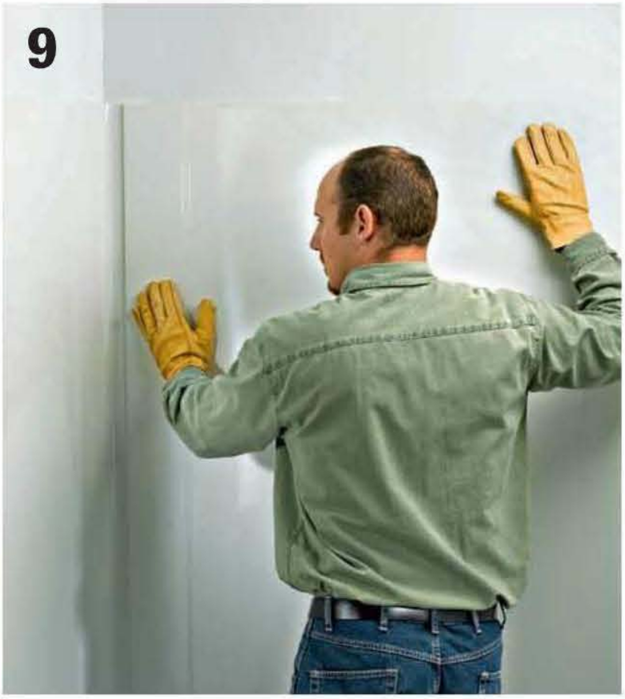

Apply adhesive to the back panel following the manufacturer's instructions. Remove protective backing from the tape. Lift the panel by the edges and carefully center between the two end panels. When positioned, firmly press in place from top to bottom.

10

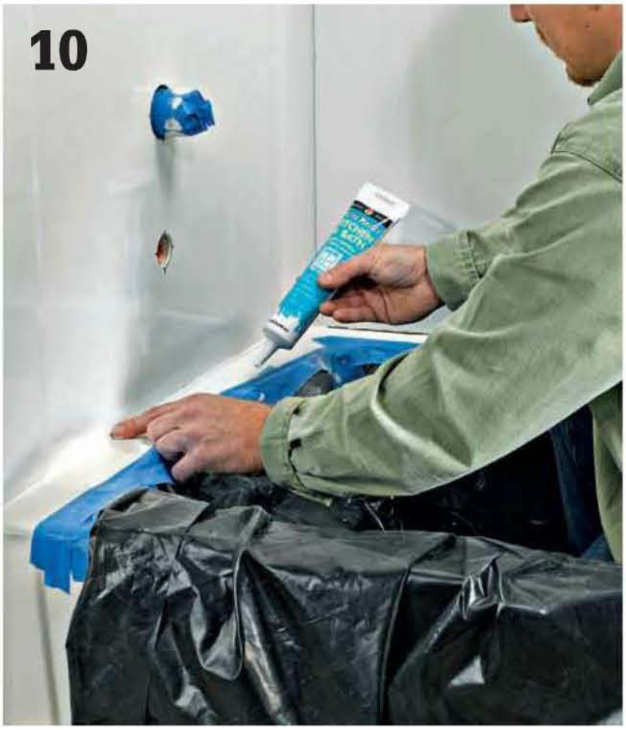

Apply caulk to the bottom and top edges of the panels and at panel joints. Dip your fingertip in water and use it to smooth the caulk to a uniform bead.

11

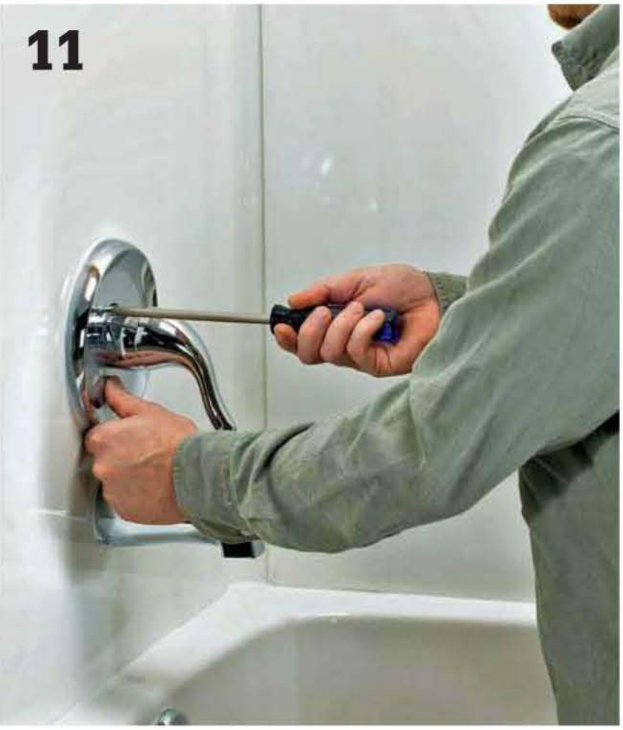

Apply silicone caulk to escutcheons or trim plates and reinstall them. Allow a minimum of 24 hours for caulk and adhesive to dry thoroughly before using the shower or tub.

Installing Sliding Tub Doors

Curtains on your bathtub shower are a hassle. If you forget to tuck them inside the tub, water flows freely onto your bathroom floor. If you forget to slide them closed, mildew sets up shop in the folds. And every time you brush against them, they stick to your skin. Shower curtains certainly don't add much elegance or charm to a dream bath. Neither does a deteriorated door. Clean up the look of your bathroom, and even give it an extra touch of elegance, with a new sliding tub door.

When shopping for a sliding tub door, you have a choice of framed or frameless. A framed door is edged in metal. The metal framing is typically aluminum but is available in many finishes, including those that resemble gold, brass, or chrome. Glass options are also plentiful. You can choose between frosted or pebbled glass, clear, mirrored, tinted, or patterned glass. Doors can be installed on ceramic tile walls or through a fiberglass tub surround.

Tools & Materials ▸

Measuring tape
Pencil
Hacksaw
Miter box
Level
Drill
Center punch
Razor blade
Marker
Masonry bit
  for tile wall
Phillips screwdriver
Caulk gun
Masking tape
Silicone sealant
  & remover
Tub door kit
Masking tape

A sliding tub door framed in aluminum gives the room a sleek, clean look and is just one of the available options.

How to Install Sliding Tub Doors

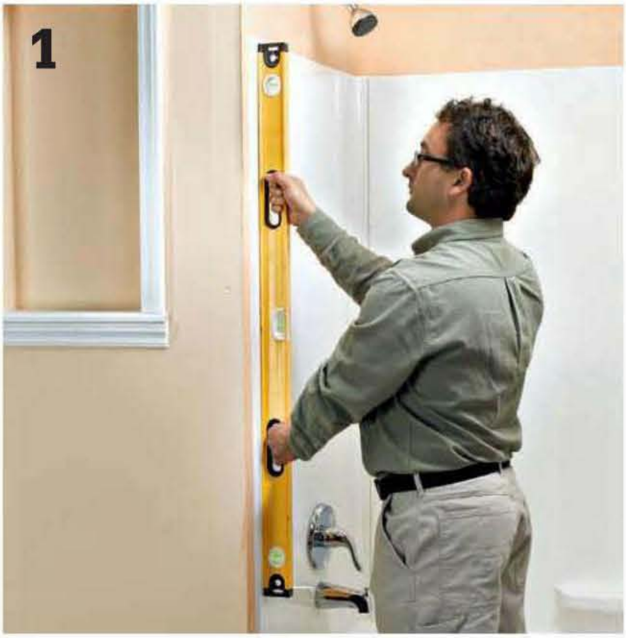

Remove the existing door and inspect the walls. Use a razor blade to cut sealant from tile and metal surfaces. Do not use a razor blade on fiberglass surfaces. Remove remaining sealant by scraping or pulling. Use a silicone sealant remover to remove all residue. Remove shower curtain rods, if present. Check the walls and tub ledge for plumb and level.

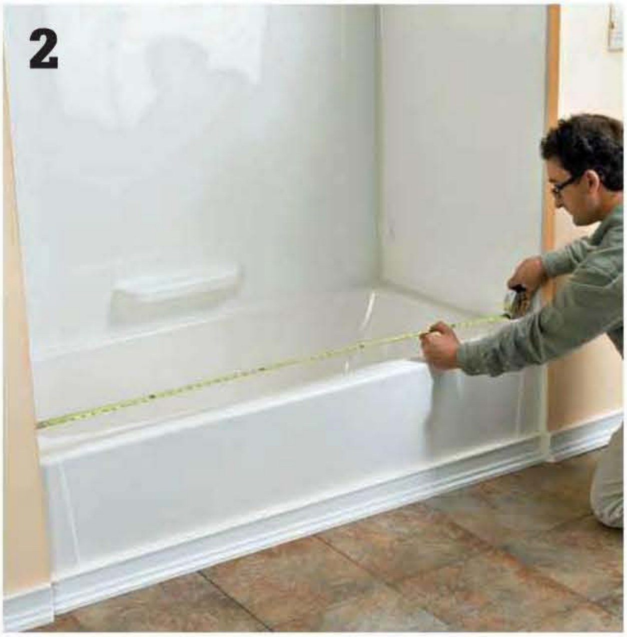

Measure the distance between the finished walls along the top of the tub ledge. Refer to the manufacturer's instructions for figuring the track dimensions. For the product seen here, $\frac{3}{16}$" is subtracted from the measurement to calculate the track dimensions.

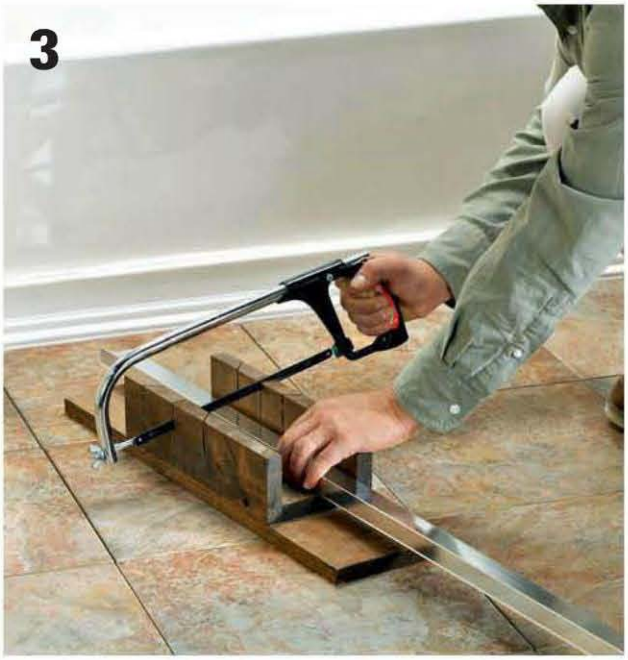

Using a hacksaw and a miter box, carefully cut the track to the proper dimension. Center the track on the bathtub ledge with the taller side out and so the gaps are even at each end. Tape into position with masking tape.

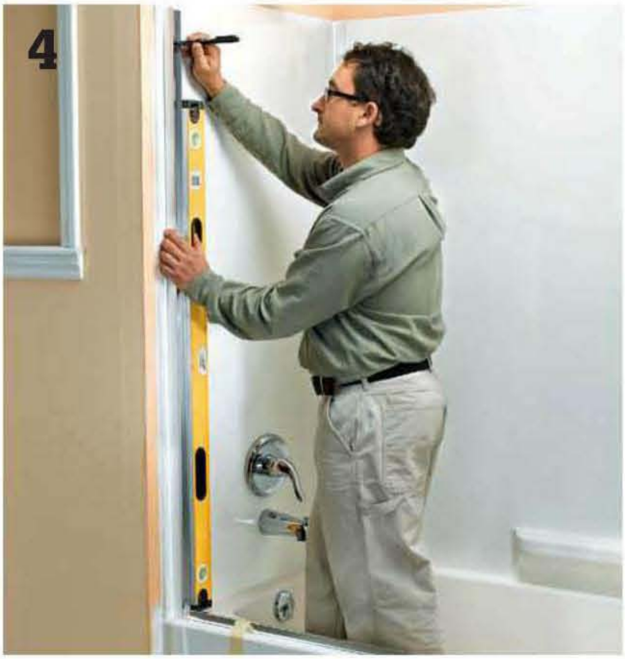

Place a wall channel against the wall with the longer side out and slide it into place over the track so they overlap. Use a level to check the channel for plumb, and then mark the locations of the mounting holes on the wall with a marker. Repeat for the other wall channel. Remove the track.

(continued)

5

Drill mounting holes for the wall channel at the marked locations. In ceramic tile, nick the surface of the tile with a center punch, use a ¼" masonry bit to drill the hole, and then insert the included wall anchors. For fiberglass surrounds, use a ⅛" drill bit; wall anchors are not necessary.

6

Apply a bead of silicone sealant along the joint between the tub and the wall at the ends of the track. Apply a minimum ¼" bead of sealant along the outside leg of the track underside.

7

Position the track on the tub ledge and against the wall. Attach the wall channels using the provided screws. Do not use caulk on the wall channels at this time.

8

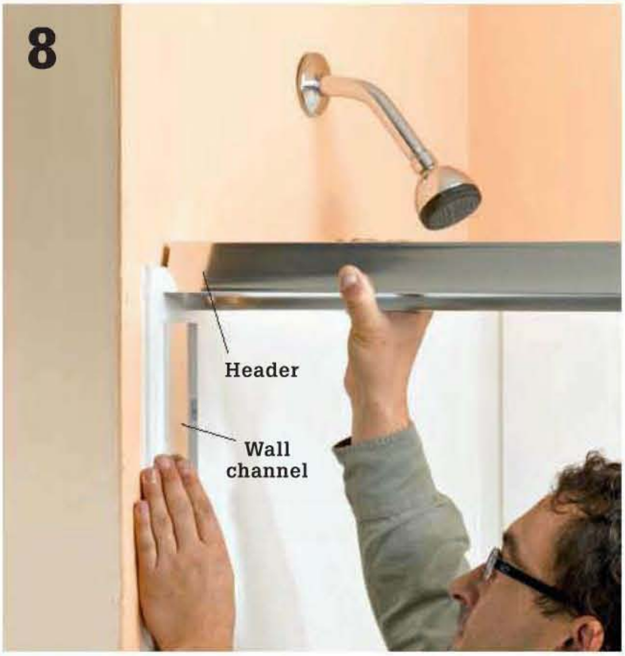

Cut and install the header. At a location above the tops of the wall channels, measure the distance between the walls. Refer to the manufacturer's instructions for calculating the header length. For the door seen here, the length is the distance between the walls minus $\frac{1}{16}$". Measure the header and carefully cut it to length using a hacksaw and a miter box. Slide the header down on top of the wall channels until seated.

9

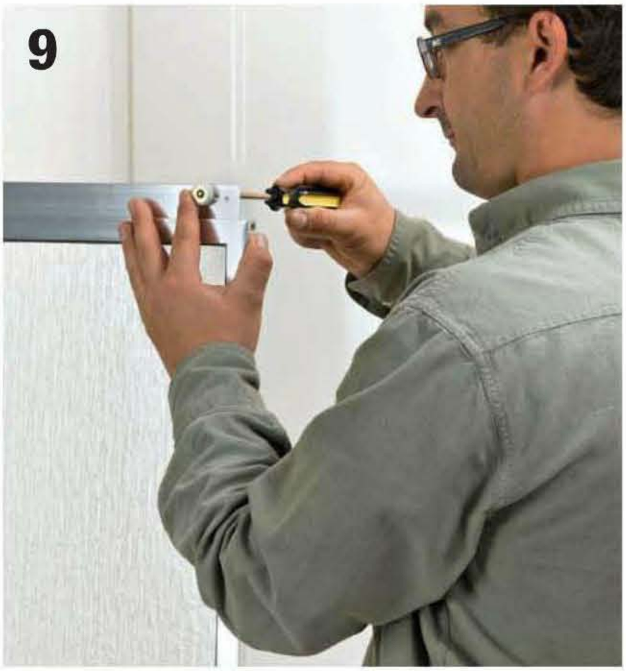

Mount the rollers in the roller mounting holes. To begin, use the second-from-the-top roller mounting holes. Follow the manufacturer's instructions for spacer or washer placement and orientation.

10

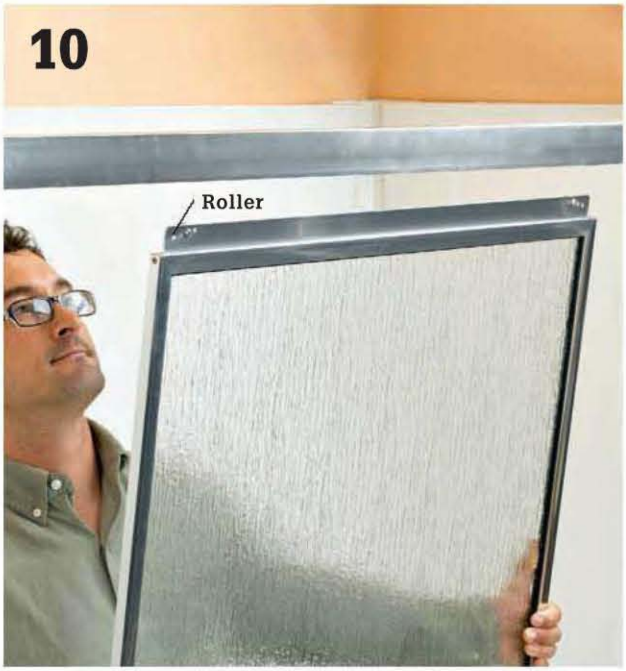

Carefully lift the inner panel by the sides and place the rollers on the inner roller track. Roll the door toward the shower end of the tub. The edge of the panel should touch both rubber bumpers. If it doesn't, remove the door and move the rollers to different holes. Drive the screws by hand to prevent overtightening.

11

Lift the outer panel by the sides with the towel bar facing out from the tub. Place the outer rollers over the outer roller track. Slide the door to the end opposite the shower end of the tub. If the door does not contact both bumpers, remove the door and move the rollers to different mounting holes.

12

Kitchen & Bath

Apply a bead of clear silicone sealant to the inside seam of the wall and wall channel at both ends and to the U-shaped joint of the track and wall channels. Smooth the sealant with a fingertip dipped in water.

Tiling Bathroom Walls

Tile is an ideal covering for walls in kitchens and bathrooms, but there's no reason to limit its use to those rooms. It's not as common in North American homes, but in Europe tile has been used in rooms throughout the house for generations. And why not? Beautiful, practical, easy to clean and maintain, tile walls are well suited to many spaces. On the preceding pages, you've seen some design ideas for tile walls.

When shopping for tile, keep in mind that tiles that are at least 6" × 6" are easier to install than small tiles, because they require less cutting and cover more surface area. Larger tiles also have fewer grout lines that must be cleaned and maintained. Check out the selection of trim and specialty tiles and ceramic accessories that are available to help you customize your project.

Most wall tile is designed to have narrow grout lines (less than ⅛" wide) filled with unsanded grout. Grout lines wider than ⅛" should be filled with sanded floor-tile grout. Either type will last longer if it contains, or is mixed with, a latex additive. To prevent staining, it's a good idea to seal your grout after it fully cures, then once a year thereafter.

You can use standard drywall or water-resistant drywall (called "greenboard") as a backer for walls in dry areas. In wet areas, install tile over cementboard. Made from cement and fiberglass, cementboard cannot be damaged by water, though moisture can pass through it. To protect the framing, install a waterproof membrane, such as roofing felt or polyethylene sheeting, between the framing members and the cementboard. Be sure to tape and finish the seams between cementboard panels before laying the tile.

See page 45 for information on planning and laying out tile walls.

Tools & Materials ▸

- Tile-cutting tools
- Marker
- Tape measure
- 4-ft. level
- Notched trowel
- Mallet
- Grout float
- Grout sponge
- Soft cloth
- Small paintbrush or foam brush
- Caulk gun
- Straight 1 × 2
- Dry-set tile mortar with latex additive
- Ceramic wall tile
- Ceramic trim tile (as needed)
- 2 × 4
- Carpet scrap
- Tile grout with latex additive
- Tub & tile caulk
- Alkaline grout sealer
- Cardboard
- Story stick/pole

Wall tile is very popular and practical in bathrooms, where it can be used on all surfaces or primarily as an accent. Installing custom-tiled tub and shower surrounds is a popular DIY project.

How to Install Wall Tile in a Bathtub Alcove

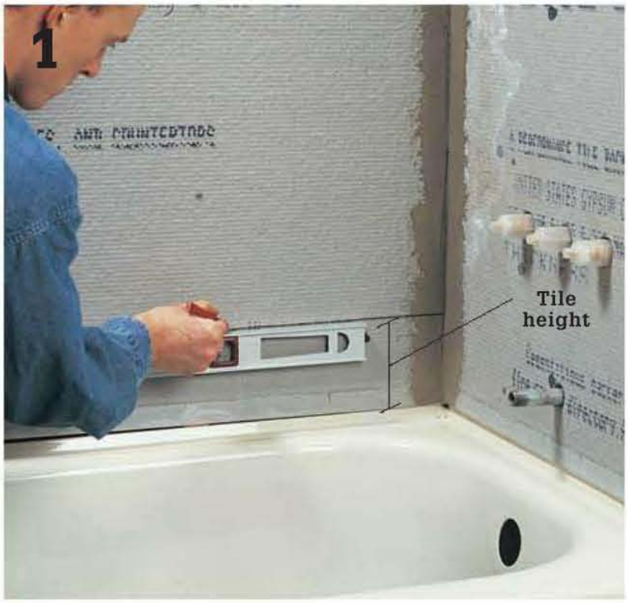

Beginning with the back wall, measure up and mark a point at a distance equal to the height of one ceramic tile (if the tub edge is not level, measure up from the lowest spot). Draw a level line through this point, along the entire back wall. This line represents a tile grout line and will be used as a reference line for making the entire tile layout.

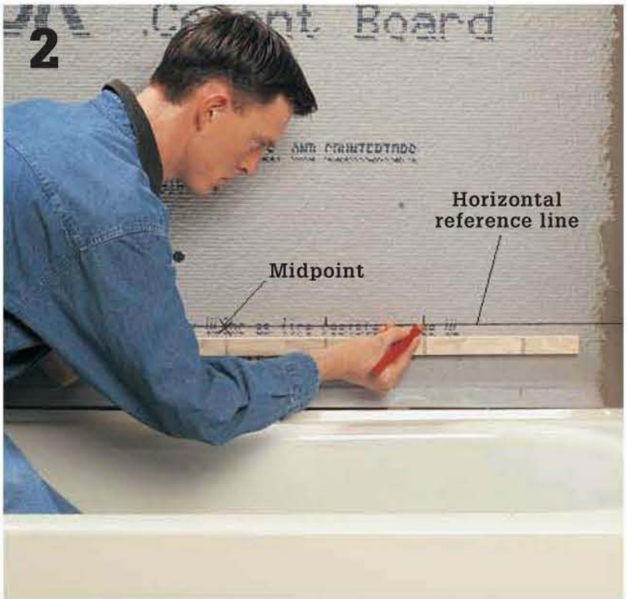

Measure and mark the midpoint on the horizontal reference line. Using a story stick, mark along the reference line where the vertical grout joints will be located. If the story stick shows that the corner tiles will be less than half of a full tile width, move the midpoint half the width of a tile in either direction and mark (shown in next step).

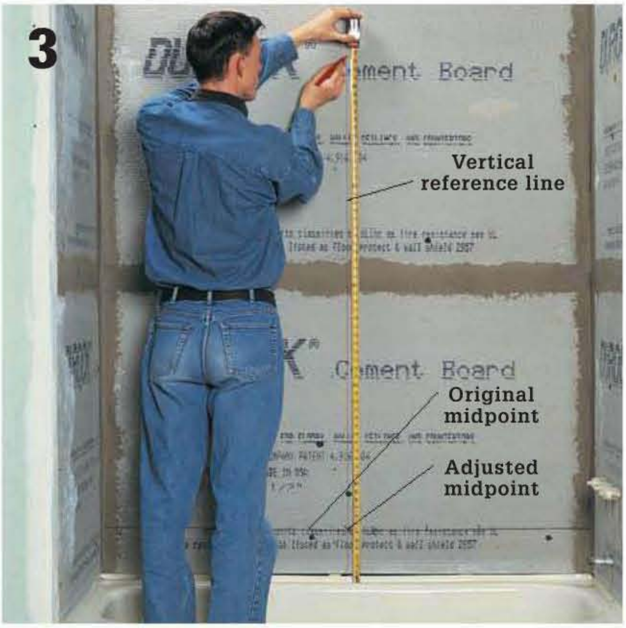

Use a level to draw a vertical reference line through the adjusted midpoint from the tub edge to the ceiling. Measure up from the tub edge along the vertical reference line and mark the rough height of the top row of tiles.

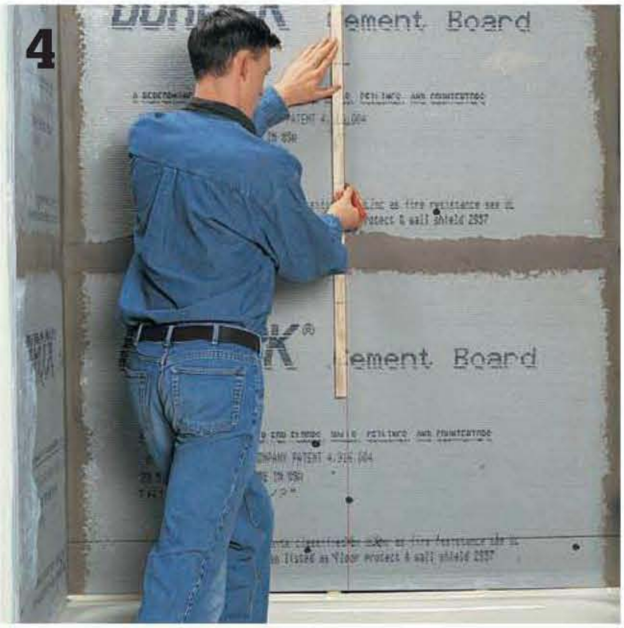

Use the story stick to mark the horizontal grout joints along the vertical reference line, beginning at the mark for the top row of tiles. If the cut tiles at the tub edge will be less than half the height of a full tile, move the top row up half the height of a tile. *Note: If tiling to a ceiling, evenly divide the tiles to be cut at the ceiling and tub edge, as for the corner tiles.*

(continued)

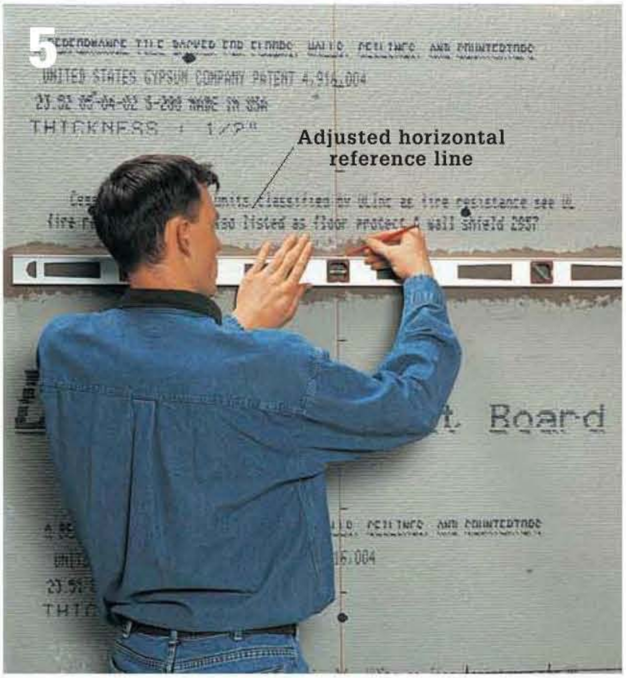

Use a level to draw an adjusted horizontal reference line through the vertical reference line at a grout joint mark close to the center of the layout. This splits the tile area into four workable quadrants.

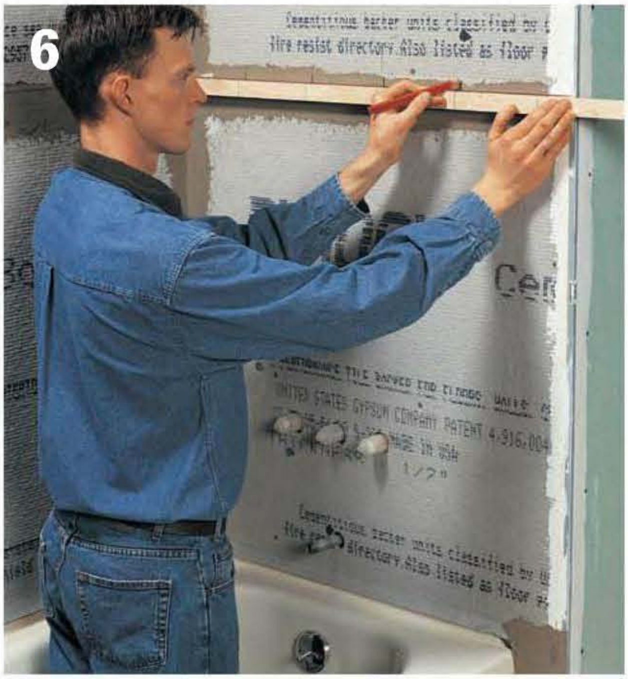

Use a level to transfer the adjusted horizontal reference line from the back wall to both side walls, then lay out both side walls. Adjust the layout as needed so the final column of tiles ends at the outside edge of the tub. Use only the adjusted horizontal and vertical reference lines for ceramic tile installation.

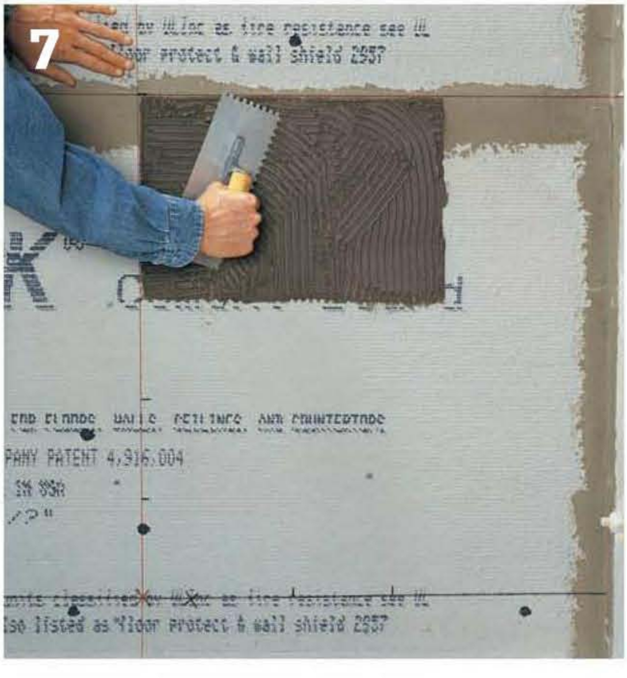

Mix a small batch of thinset mortar containing a latex additive. (Some mortar has additive mixed in by the manufacturer and some must have additive mixed separately.) Spread adhesive on a small section of the wall, along both legs of one quadrant, using a ¼" notched trowel.

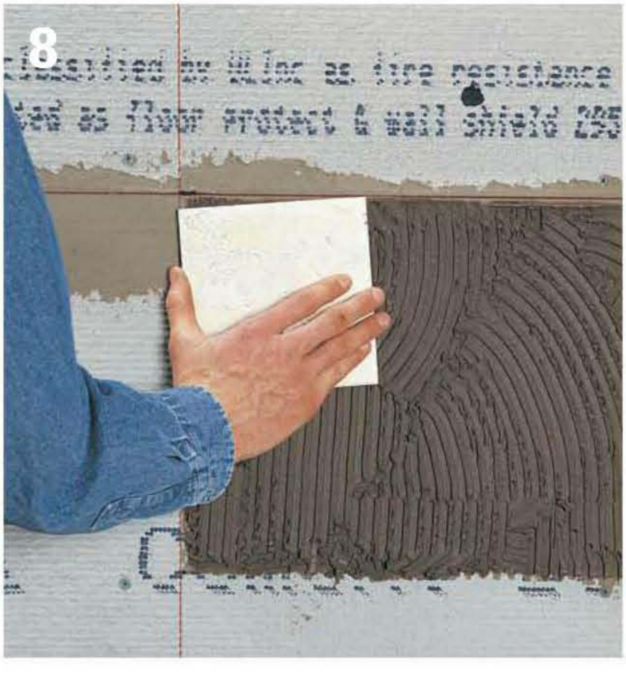

Use the edge of the trowel to create furrows in the mortar. Set the first tile in the corner of the quadrant where the lines intersect, using a slight twisting motion. Align the tile exactly with both reference lines. When placing cut tiles, position the cut edges where they will be least visible.

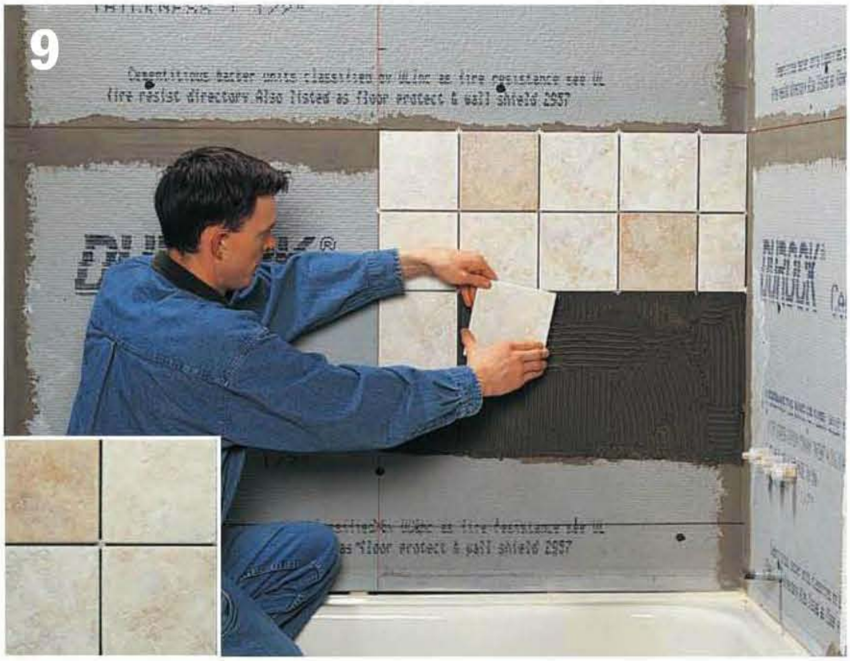

Continue installing tiles, working from the center out into the field of the quadrant. Keep the tiles aligned with the reference lines and tile in one quadrant at a time. If the tiles are not self-spacing, use plastic spacers inserted in the corner joints to maintain even grout lines (inset). The base row against the tub edge should be the last row of tiles installed.

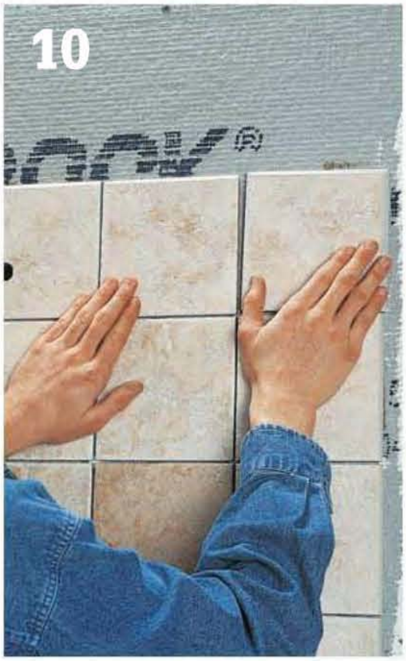

Install remaining tiles, including border areas. Wipe away excess mortar along the top edges of the edge tiles.

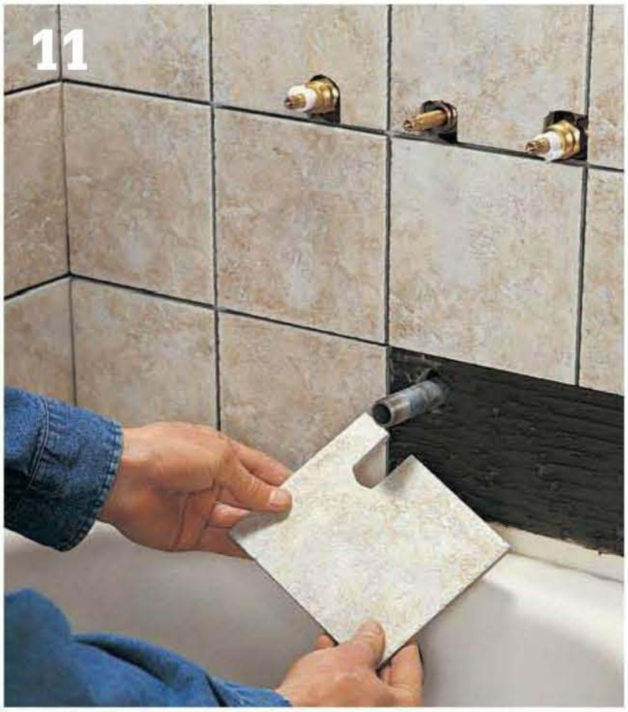

Mark and cut tiles to fit around all plumbing accessories or plumbing fixtures.

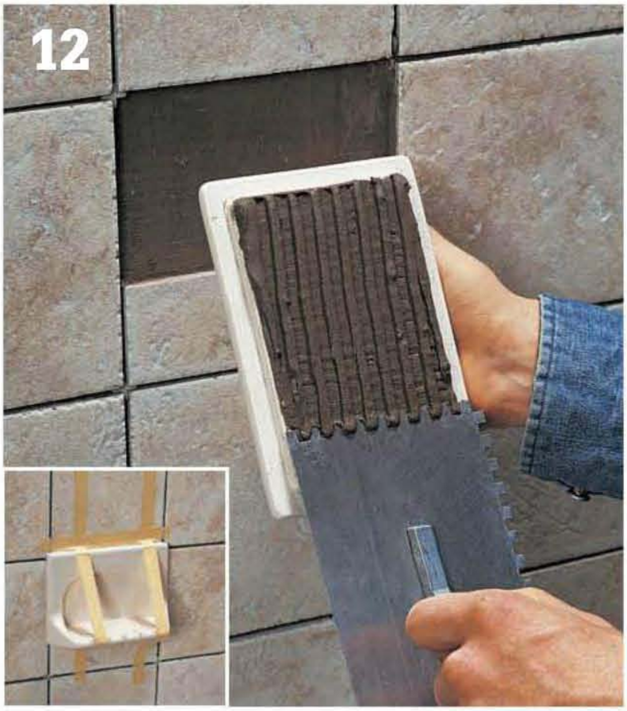

Install any ceramic accessories by applying thinset mortar to the back side, then pressing the accessory into place. Use masking tape to support the weight until the mortar dries (inset). Fill the tub with water, then seal expansion joints around the bathtub, floor, and corners with silicone caulk.

Installing Toilets

You can replace a poorly functioning or inefficient toilet with a high-efficiency, high-quality new toilet in just a single afternoon. All toilets made since 1996 have been required to use 1.6 gallons or less per flush, which has been a huge challenge for the industry. Today, the most evolved 1.6-gallon toilets have wide passages behind the bowl and wide (3") flush valve openings—features that facilitate short, powerful flushes. This means fewer second flushes and fewer clogged toilets. These problems were common complaints of the first generation of 1.6-gallon toilets and continue to beleaguer inferior models today. See what toilets are available at your local home center in your price range, then go online and see what other consumers' experiences with those models have been. New toilets often go through a "de-bugging" stage when problems with leaks and malfunctioning parts are more common. Your criteria should include ease of installation, good flush performance, and reliability. With a little research, you should be able to purchase and install a high-functioning economical gravity-flush toilet that will serve you well for years to come.

Tools & Materials ▸

- Adjustable wrench
- Bucket and sponge
- Channel-type pliers
- Hacksaw
- Penetrating oil
- Pliers
- Putty knife
- Rubber gloves
- Screwdriver
- Supply tube
- Teflon tape
- Toilet seat bolts
- Toilet seat
- Towels
- Utility knife
- Wax ring with flange

Replacing a toilet is simple, and the latest generation of 1.6-gallon water-saving toilets has overcome the performance problems of earlier models.

How to Install a Toilet

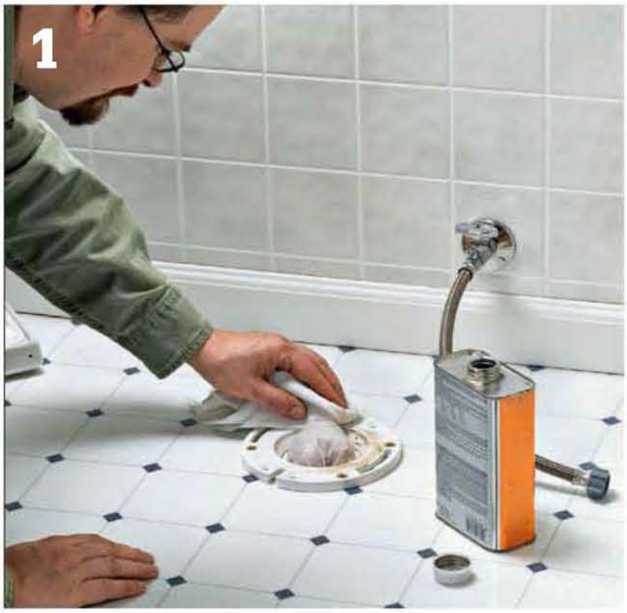

Clean and inspect the old closet flange. Look for breaks or wear. Also inspect the flooring around the flange. If either the flange or floor is worn or damaged, repair the damage. Use a rag and mineral spirits to completely remove residue from the old wax ring. Place a rag into a plastic bag and use it to block toxic sewer gas in the drain opening.

Tip ▸

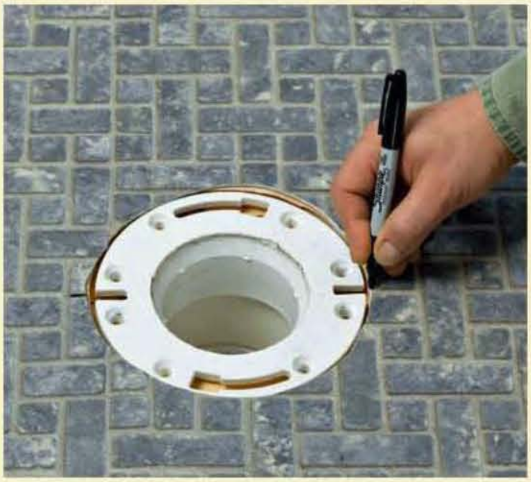

If you will be replacing your toilet flange or if your existing flange can be unscrewed and moved, orient the new flange so the slots are parallel to the wall. This allows you to insert bolts under the slotted areas, which are much stronger than the areas at the ends of the curved grooves.

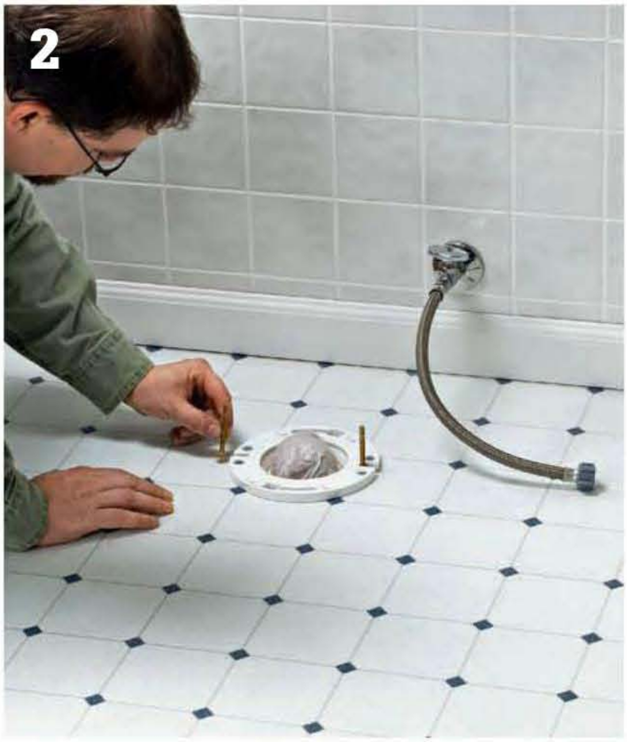

Insert new tank bolts (don't reuse old ones) into the openings in the closet flange. Make sure the heads of the bolts are oriented to catch the maximum amount of flange material.

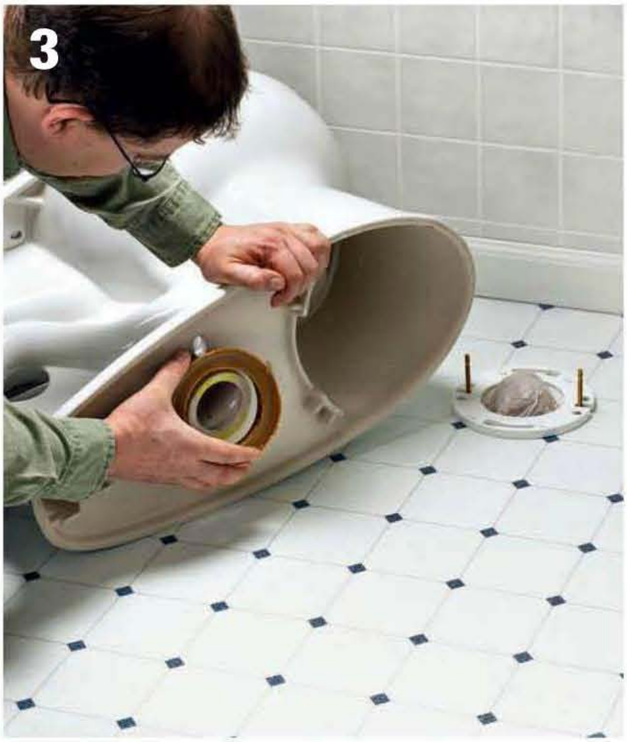

Remove the wax ring and apply it to the underside of the bowl, around the horn. Remove the protective covering. Do not touch the wax ring. It is very sticky. Remove the plastic bag.

(continued)

4

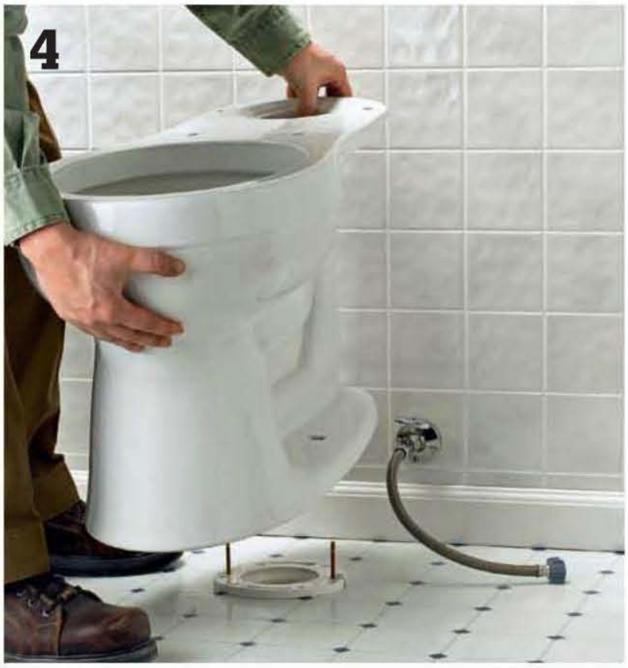

Lower the bowl onto the flange, taking care not to disturb the wax ring. The holes in the bowl base should align perfectly with the tank bolts. Add a washer and tighten a nut on each bolt. Hand tighten each nut and then use channel-type pliers to further tighten the nuts. Alternate back and forth between nuts until the bowl is secure. *Do not overtighten.*

5

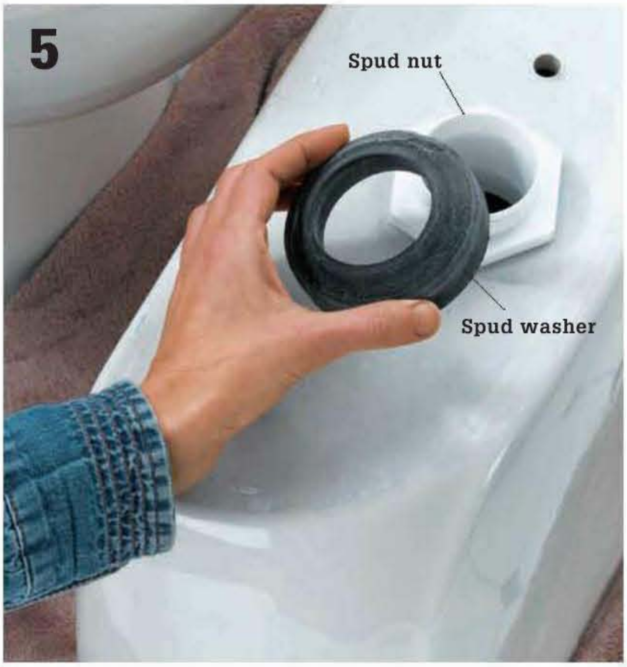

Attach the toilet tank. Some tanks come with a flush valve and a fill valve preinstalled. For models that do not have this, insert the flush valve through the tank opening and tighten a spud nut over the threaded end of the valve. Place a foam spud washer on top of the spud nut.

6

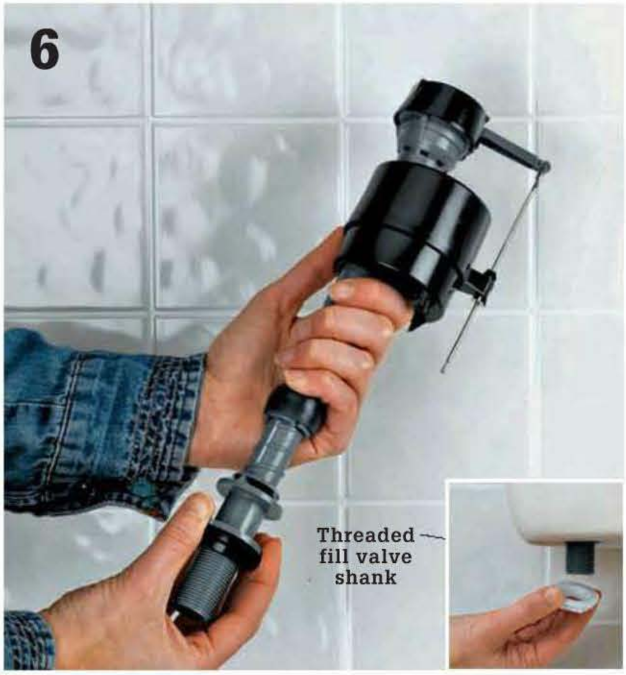

Adjust the fill valve as directed by the manufacturer to set the correct tank water level height and install the valve inside the tank. Hand tighten the nylon lock nut that secures the valve to the tank (inset photo) and then tighten it further with channel-type pliers.

7

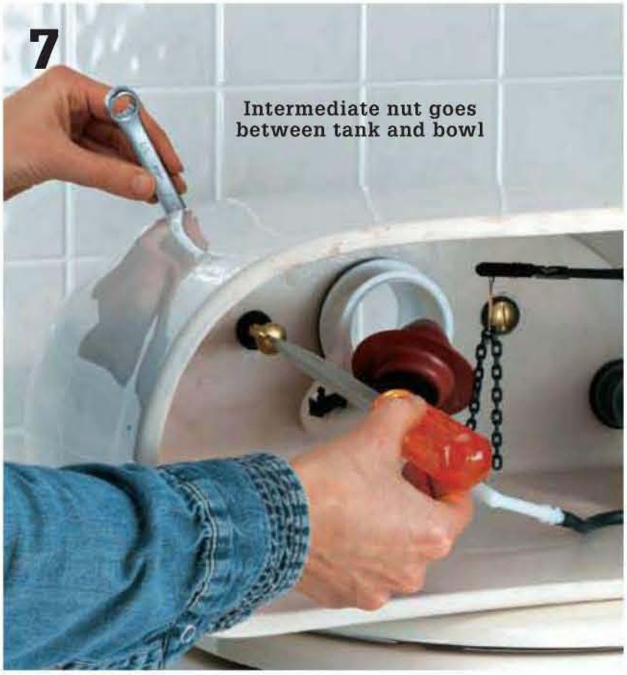

With the tank lying on its back, thread a rubber washer onto each tank bolt and insert it into the bolt holes from inside the tank. Then, thread a brass washer and hex nut onto the tank bolts from below and tighten them to a quarter turn past hand tight. Do not overtighten.

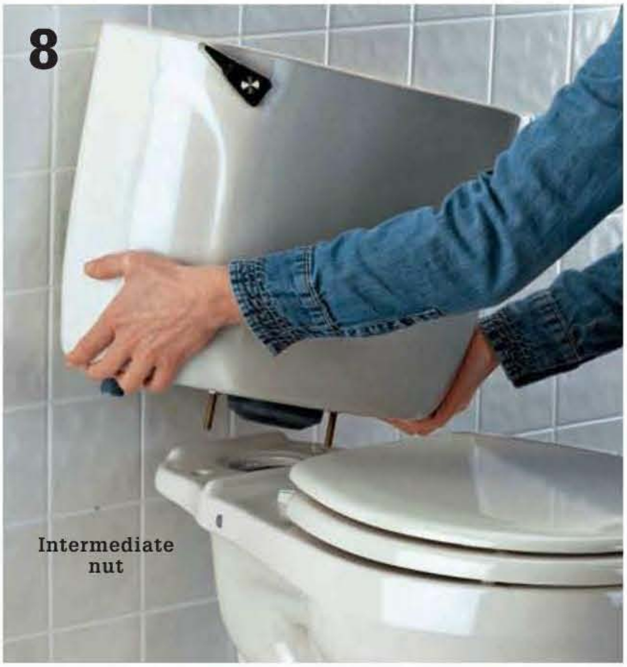

Position the tank on the bowl, spud washer on opening, bolts through bolt holes. Put a rubber washer, followed by a brass washer and a wing nut, on each bolt and tighten these up evenly.

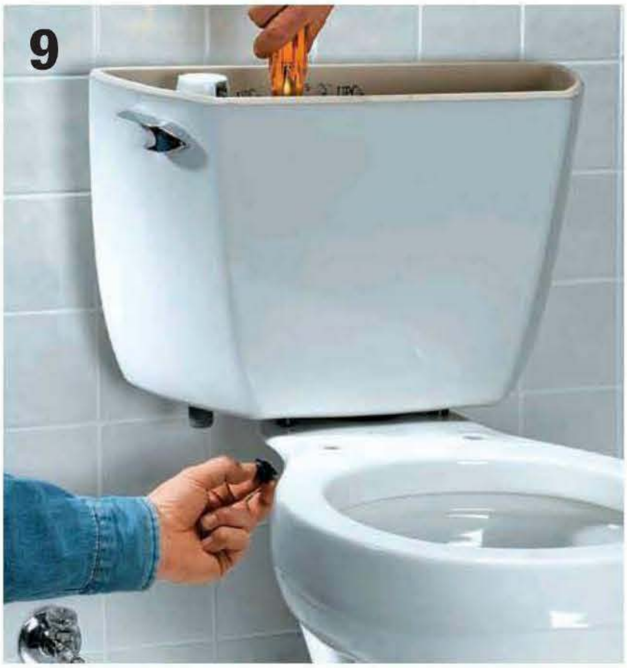

You may stabilize the bolts with a large slotted screwdriver from inside the tank, but tighten the nuts, not the bolts. You may press down a little on a side, the front, or the rear of the tank to level it as you tighten the nuts by hand. Do not overtighten and crack the tank. The tank should be level and stable when you're done.

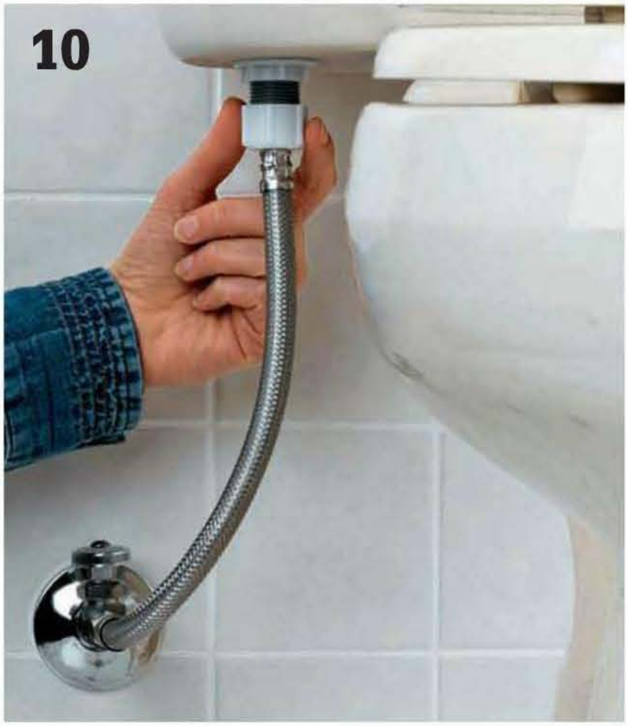

Hook up the water supply by connecting the supply tube to the threaded fill valve with the coupling nut provided. Turn on the water and test for leaks. Do not overtighten.

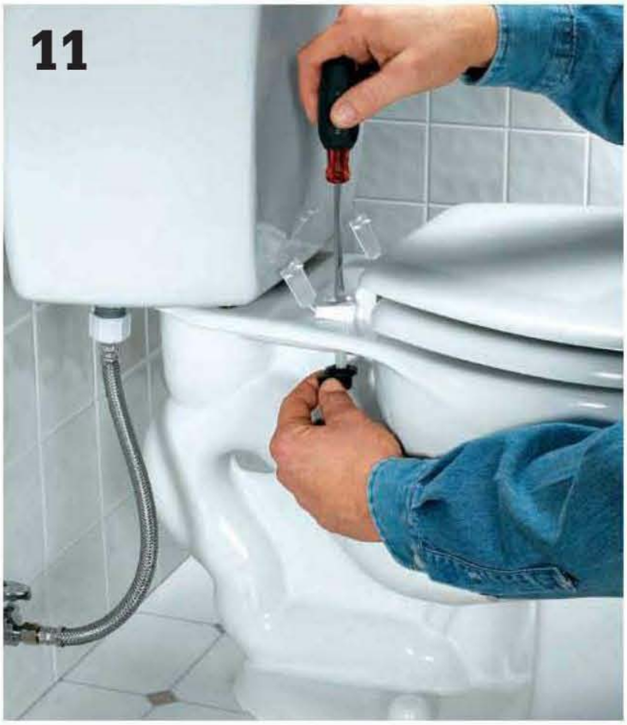

Attach the toilet seat by threading the plastic or brass bolts provided with the seat through the openings on the back of the rim and attaching the nuts.

Installing Wall-hung Vanities

Think of a wall-mounted sink or vanity cabinet and you're likely to conjure up images of public restrooms where these conveniences are installed to improve access for floor cleaning. However, wall-hung sinks and vanities made for home use are very different from the commercial installations.

Often boasting high design, beautiful modern vanities and sinks come in a variety of styles and materials, including wood, metal, and glass. Some attach with decorative wall brackets that are part of the presentation; others look like standard vanities without legs. Install wall-hung sinks and vanities by attaching them securely to studs or wood blocking.

Tools & Materials ▸

Studfinder	Level
Drill	Vanity

Today's wall-hung sinks are stylish and attractive. They must be mounted into studs or added blocking to keep them secure.

How to Install a Wall-hung Vanity Base

Remove the existing sink or fixture and inspect the wall framing. Also determine if plumbing supply and waste lines will need to be moved to accommodate the dimensions of the new fixture. Locate the studs in the sink location with a stud finder.

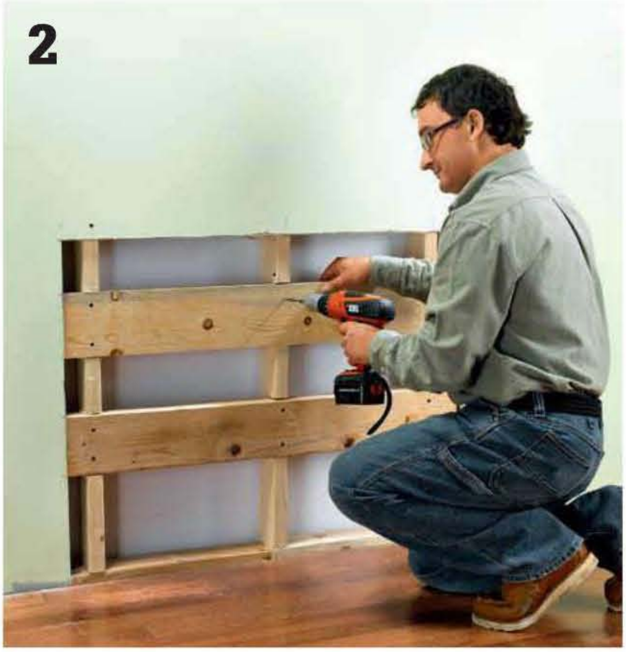

Hold the sink or cabinet in the installation area and check to see if the studs align with the sink or sink bracket mounting holes. If they do, skip to step 3. If the studs do not align, remove the wallboard behind the mounting area. Install 2 × 6 blocking between studs at the locations of the mounting screws. Replace and repair wallboard.

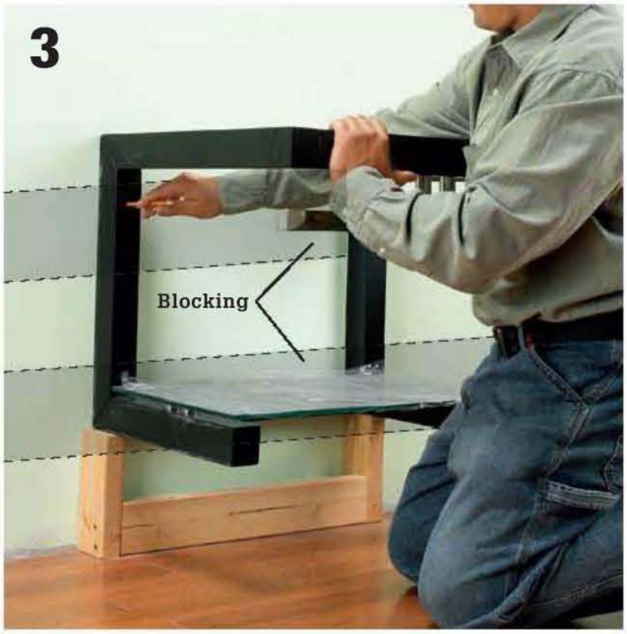

Mark the locations of the mounting holes on the wall using a template or by supporting the sink or vanity against the wall with a temporary brace (made here from scrap 2 × 4s) and marking through the mounting holes.

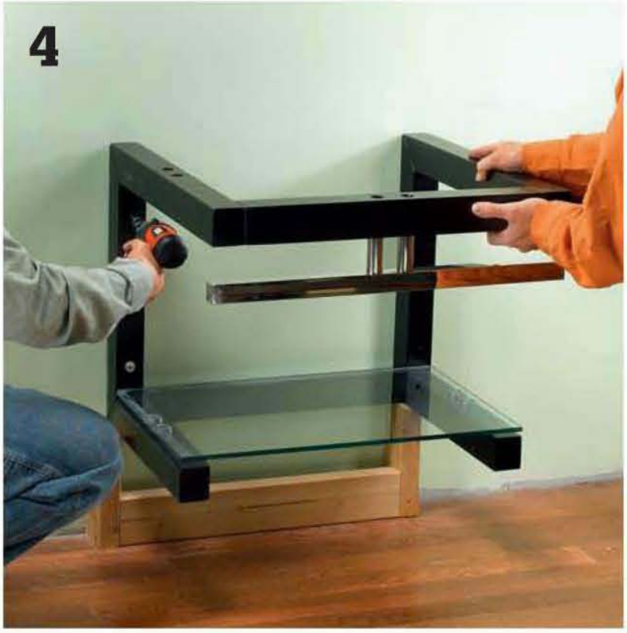

Drill pilot holes at the marks. Have a helper hold the vanity in place while you drive the mounting screws. Hook up the plumbing.

Installing Vessel Sinks

The vessel sink harkens back to the days of washstands and washbowls. Whether it's round, square, or oval, shallow or deep, the vessel sink offers great opportunity for creativity and proudly displays its style. Vessel sinks are a perfect choice for a powder room, where they will have high visibility.

Most vessel sinks can be installed on any flat surface—from a granite countertop to a wall-mounted vanity to an antique dresser. Some sinks are designed to contact the mounting surface only at the drain flange. Others are made to be partially embedded in the surface. Take care to follow the manufacturer's instructions for cutting holes for sinks and faucets.

A beautiful vessel sink demands an equally attractive faucet. Select a tall spout mounted on the countertop or vanity top or a wall-mounted spout to accommodate the height of the vessel. To minimize splashing, spouts should directly flow to the center of the vessel, not down the side. Make sure your faucet is compatible with your vessel choice. Look for a centerset or single-handle model if you'll be custom drilling the countertop—you only need to drill one faucet hole.

Tools & Materials ▸

- Jigsaw
- Trowel
- Pliers
- Wrench
- Caulk gun and caulk
- Sponge
- Drill
- Vanity or countertop
- Vessel sink
- Pop-up drain
- P-trap and drain kit
- Faucet
- Phillips screwdriver

Vessel sinks are available in countless styles and materials, shapes, and sizes. Their one commonality is that they all need to be installed on a flat surface.

Vessel Sink Options

This glass vessel sink embedded in a "floating" glass countertop is a stunning contrast to the strong and attractive frame anchoring it to the wall.

The natural stone vessel sink blends elegantly into the stone countertop and is enhanced by the sleek faucet and round mirror.

The stone vessel sink is complemented by the wall-hung faucet. The rich wood vanity on which it's perched adds warmth to the room.

Vitreous china with a glazed enamel finish is an economical and durable choice for a vessel sink (although it is less durable than stone). Because of the adaptability of both the material and the glaze, the design options are virtually unlimited with vitreous china.

How to Install a Vessel Sink

Secure the vanity cabinet or other countertop that you'll be using to mount the vessel sink (see pages 464 to 465).

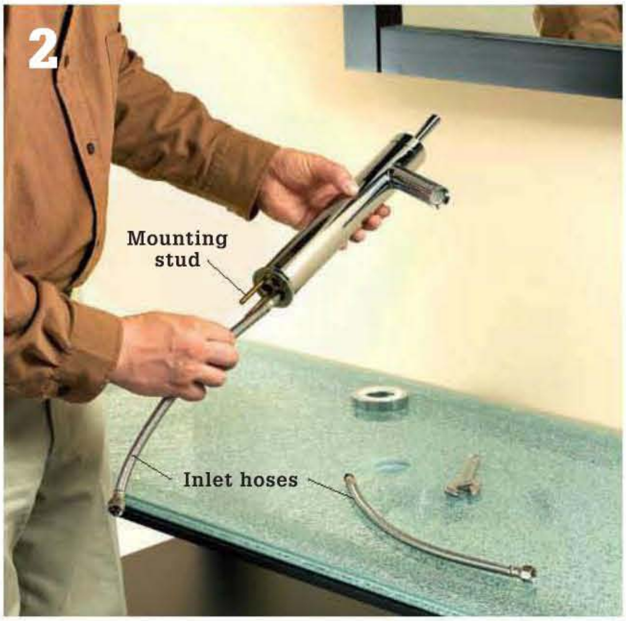

Begin hooking up the faucet. Insert the brass mounting stud into the threaded hole in the faucet base with the slotted end facing out. Hand tighten, and then use a screwdriver to tighten another half turn. Insert the inlet hoses into the faucet body and hand tighten. Use an adjustable wrench to tighten another half turn. Do not overtighten.

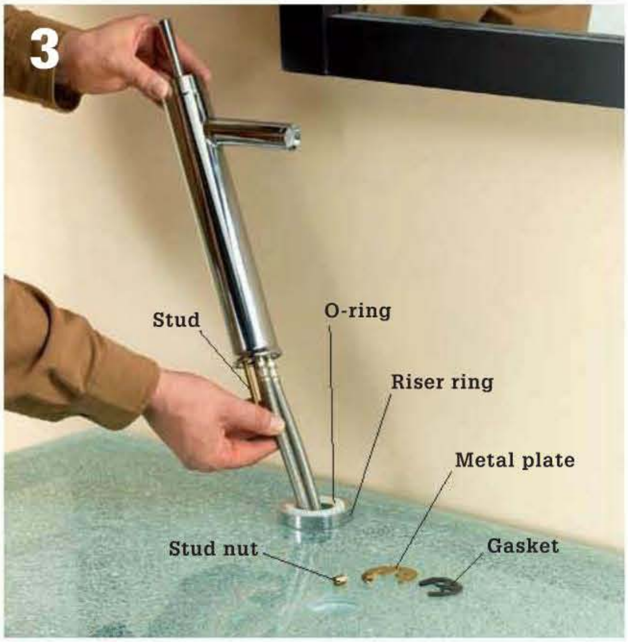

Place the O-ring on top of the riser-ring over the faucet cutout in the countertop. From underneath, slide the rubber gasket and the metal plate over the mounting stud. Thread the mounting stud nut onto the mounting stud and hand tighten. Use an adjustable wrench to tighten another half turn.

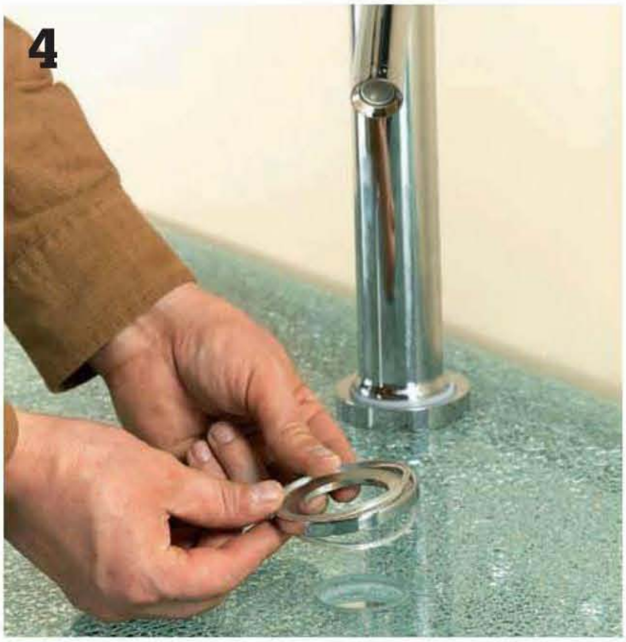

To install the sink and pop-up drain, first place the small metal ring between two O-rings and place over the drain cutout.

Place the vessel bowl on top of the O-rings. In this installation, the vessel is not bonded to the countertop.

Put the small rubber gasket over the drain hole in the vessel. From the top, push the pop-up assembly through the drain hole.

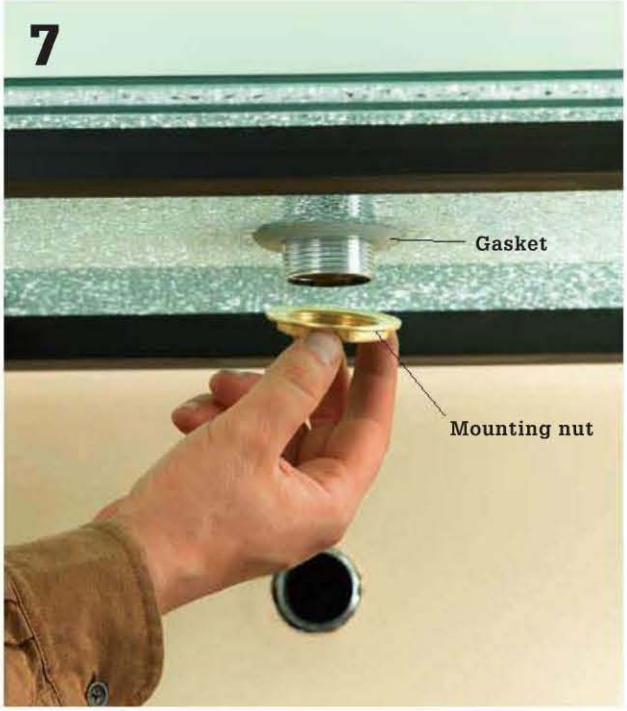

From underneath, push the large rubber gasket onto the threaded portion of the pop-up assembly. Thread the nut onto the pop-up assembly and tighten. Use an adjustable wrench or basin wrench to tighten an additional half turn. Thread the tailpiece onto the pop-up assembly.

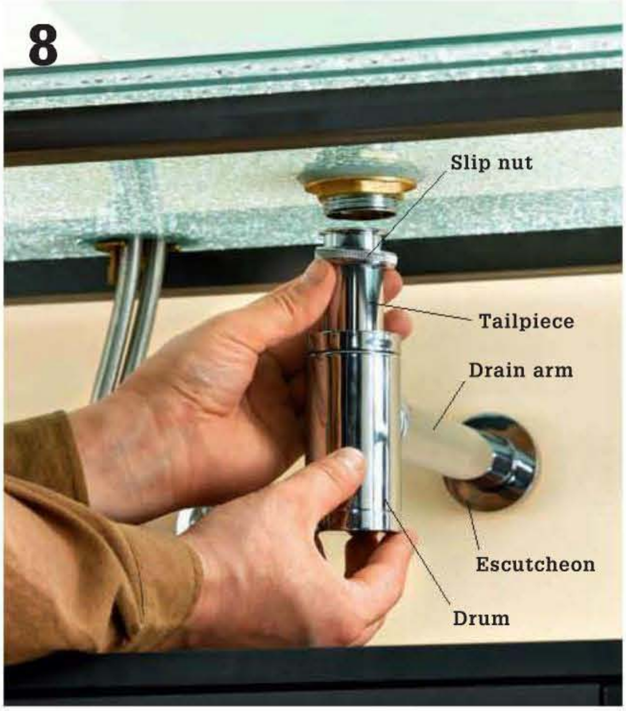

Install the drum trap. Loosen the rings on the top and outlet of the drum trap. Slide the drum trap top hole over the tailpiece. Slide the drain arm into the side outlet, with the flat side of the rubber gasket facing away from the trap. Insert the drain arm into the wall outlet. Hand tighten the rings.

Installing Pedestal Sinks

Pedestal sinks move in and out of popularity more frequently than other sink types, but even during times they aren't particularly trendy, they retain fairly stable demand. You'll find them most frequently in half baths, where their small footprint makes them an efficient choice. Designers are also discovering the appeal of tandem pedestal sinks of late, where the smaller profiles allow for his-and-hers sinks that don't dominate visually.

The primary drawback to pedestal sinks is that they don't offer any storage. Their chief practical benefit is that they conceal plumbing some homeowners would prefer not to see.

Pedestal sinks are mounted in two ways. Most of the more inexpensive ones you'll find at home stores are hung in the manner of wall-hung sinks. The pedestal is actually installed after the sink is hung and its purpose is only decorative. But other pedestal sinks (typically on the higher end of the design scale) have structurally important pedestals that bear most or all of the weight of the sink.

Pedestal sinks are available in a variety of styles and are a perfect fit for small half baths. They keep plumbing hidden, lending a neat, contained look to the bathroom.

How to Install a Pedestal Sink

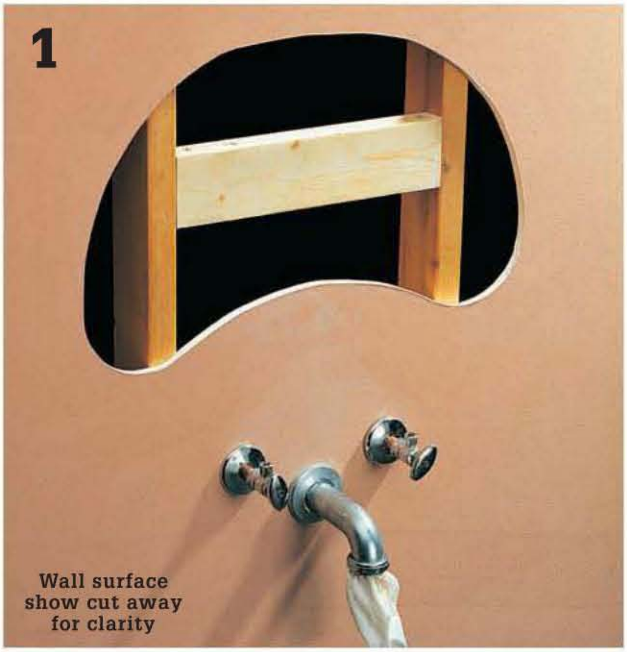

Install 2 × 4 blocking between the wall studs, behind the planned sink location. Cover the wall with water-resistant drywall. Waste and supply lines may need to be moved, depending on the sink.

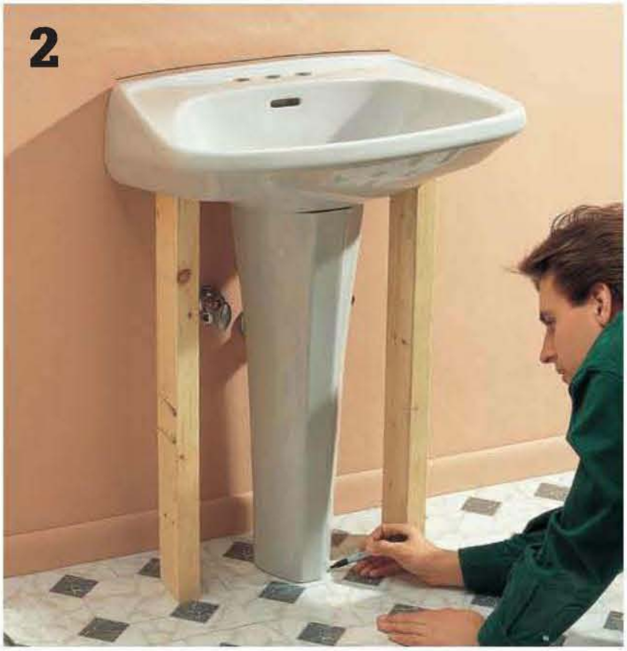

Set the basin and pedestal in position and brace it with 2 × 4s. Outline the top of the basin on the wall, and mark the base of the pedestal on the floor. Mark reference points on the wall and floor through the mounting holes found on the back of the sink and the bottom of the pedestal.

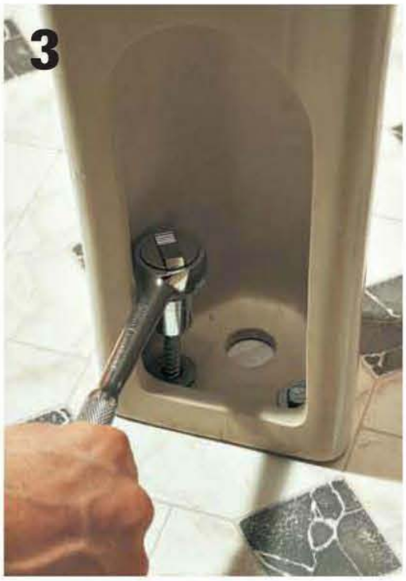

Set aside the basin and pedestal. Drill pilot holes in the wall and floor at the reference points, then reposition the pedestal. Anchor the pedestal to the floor with lag screws.

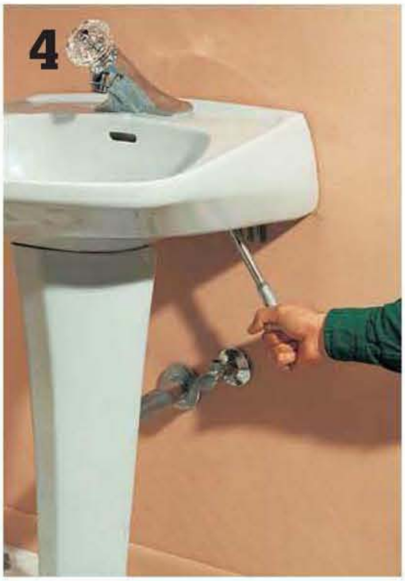

Attach the faucet, then set the sink on the pedestal. Align the holes in the back of the sink with the pilot holes drilled in the wall, then drive lag screws and washers into the wall brace using a ratchet wrench. Do not overtighten the screws.

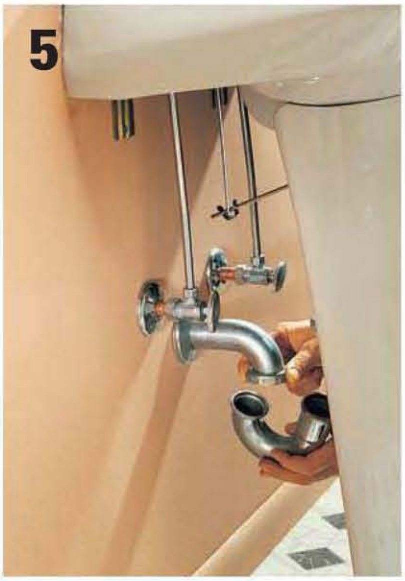

Hook up the drain and supply fittings. Caulk between the back of the sink and the wall when installation is finished.

Installing Vanity Cabinets

Most bathroom countertops installed today are integral (one-piece) sink-countertop units made from cultured marble or other solid materials, like solid surfacing. Integral sink-countertops are convenient, and many are inexpensive, but style and color options are limited.

Some remodelers and designers still prefer the distinctive look of a custom-built countertop with a self-rimming sink basin, which gives you a much greater selection of styles and colors. Installing a self-rimming sink is very simple.

Integral sink-countertops are made in standard sizes to fit common vanity widths. Because the sink and countertop are cast from the same material, integral sink-countertops do not leak, and do not require extensive caulking and sealing.

Tools & Materials ▸

- Pencil
- Scissors
- Carpenter's level
- Screwdriver
- Channel-type pliers
- Ratchet wrench
- Basin wrench
- Cardboard
- Masking tape
- Plumber's putty
- Lag screws
- Tub and tile caulk
- Plumber's putty

How to Install a Vanity Cabinet

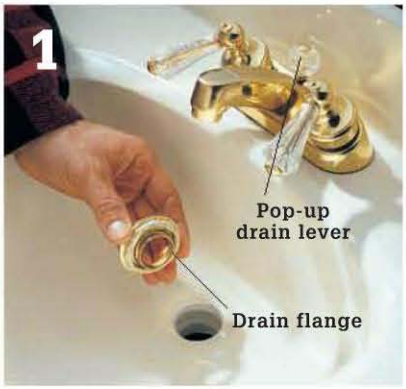

Set the sink-countertop unit onto sawhorses. Attach the faucet and slip the drain lever through the faucet body. Place a ring of plumber's putty around the drain flange, then insert the flange in the drain opening.

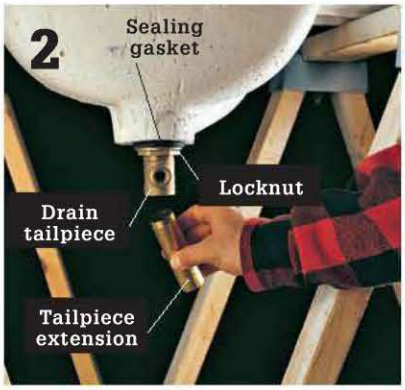

Thread the locknut and sealing gasket onto the drain tailpiece, then insert the tailpiece into the drain opening and screw it onto the drain flange. Tighten the locknut securely. Attach the tailpiece extension. Insert the pop-up stopper linkage.

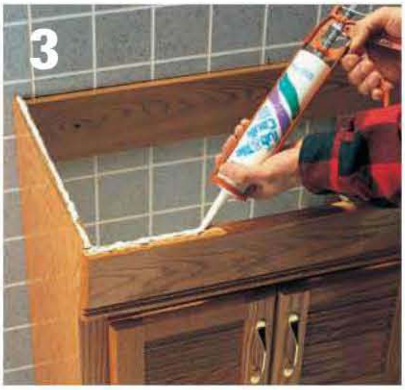

Apply a layer of tub and tile caulk (or adhesive, if specified by the countertop manufacturer) to the top edges of the cabinet vanity, and to any corner braces.

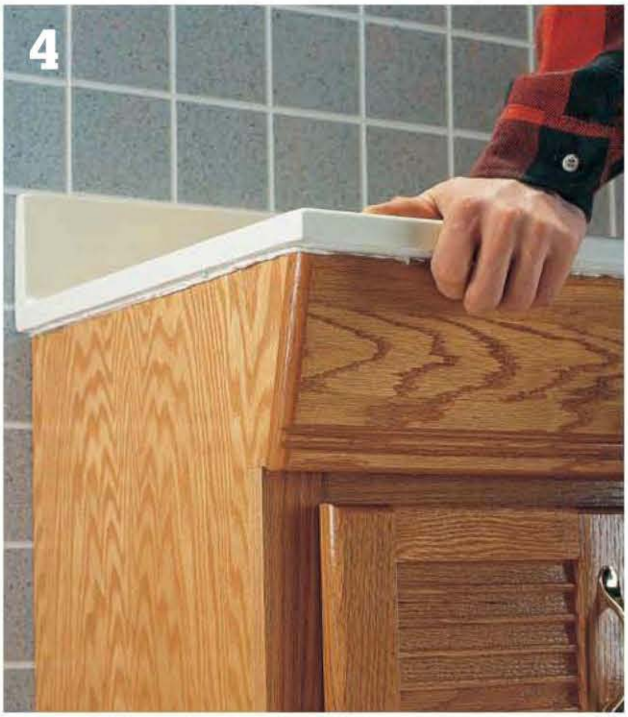

Center the sink-countertop unit over the vanity so the overhang is equal on both sides and the backsplash of the countertop is flush with the wall. Press the countertop evenly into the caulk.

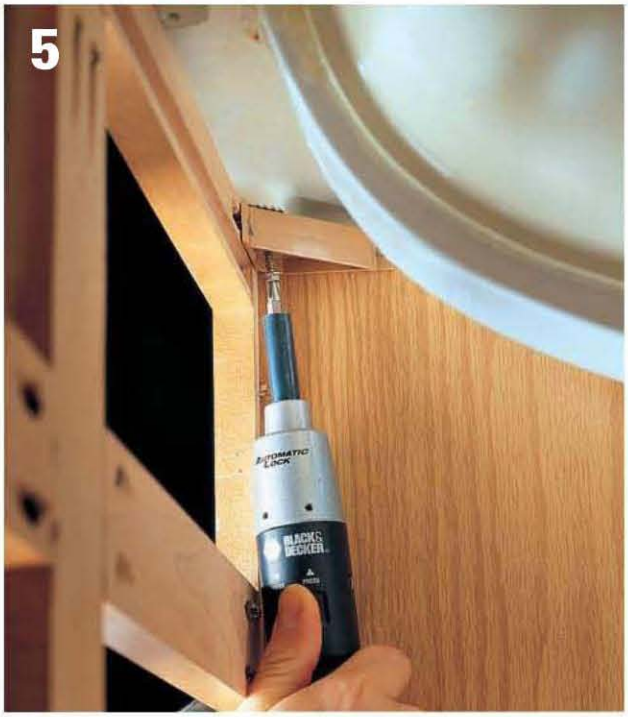
5
Automatic Lock
Black & Decker

Cabinets with corner braces: Secure the countertop to the cabinet by driving a mounting screw through each corner brace and up into the countertop. *Note: Cultured marble and other hard countertops require predrilling and a plastic screw sleeve.*

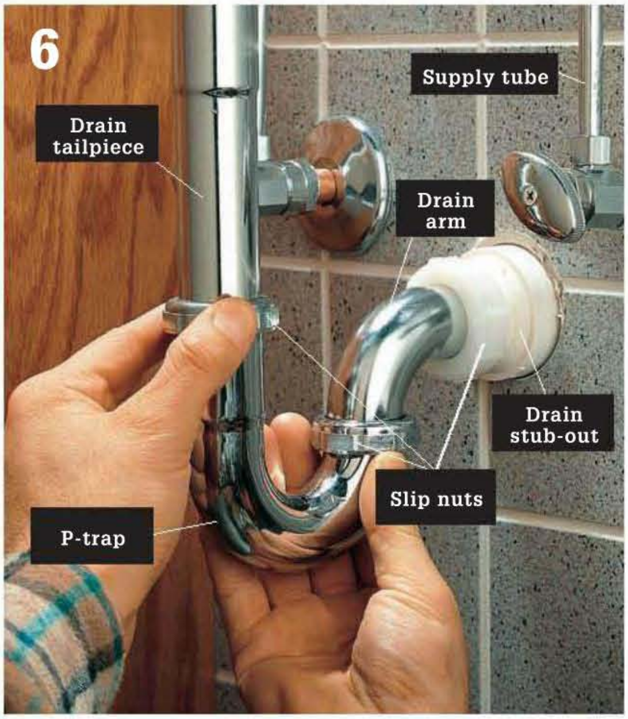
6
Supply tube
Drain tailpiece
Drain arm
Drain stub-out
Slip nuts
P-trap

Attach the drain arm to the drain stub-out in the wall, using a slip nut. Attach one end of the P-trap to the drain arm, and the other to the tailpiece of the sink drain, using slip nuts. Connect supply tubes to the faucet tailpieces.

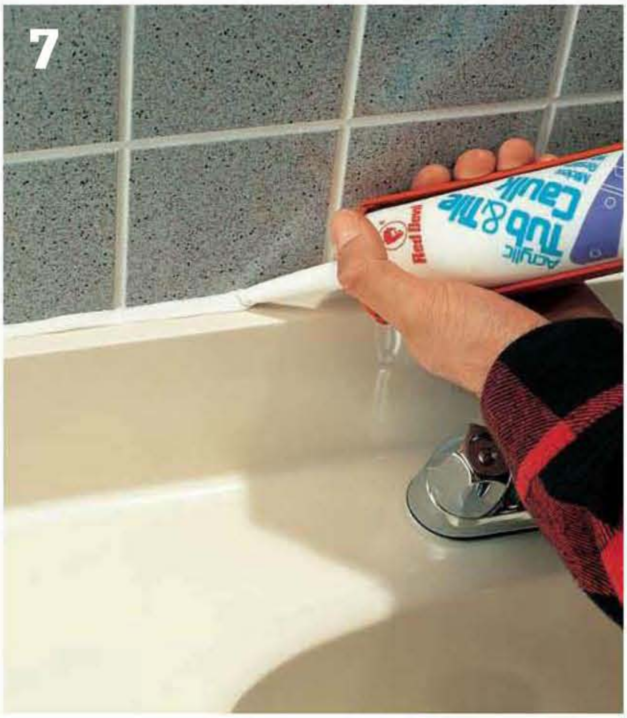
7
Red Devil
Acrylic Tub & Tile Caulk

Seal the gap between the backsplash and the wall using tub and tile caulk.

Installing Bathroom Sink Faucets & Drains

One-piece faucets, with either one or two handles, are the most popular fixtures for bathroom installations.

"Widespread" faucets with separate spout and handles are being installed with increasing frequency, however. Because the handles are connected to the spout with flex tubes that can be 18" or longer, widespread faucets can be arranged in many ways.

Tools & Materials ▸

- Hacksaw or tin snips
- Channel-type pliers
- Pliers
- Basin wrench
- Adjustable wrench
- Screwdriver
- Plumber's putty
- Teflon tape
- Faucet kit
- Pipe joint compound
- Flexible supply tubes
- Heat-proof grease
- Loctite (adhesive)

Bathroom sink faucets come in two basic styles: the widespread with independent handles and spout (top); and the single-body, deck-mounted version (bottom).

Bathroom Faucet & Drain Hookups

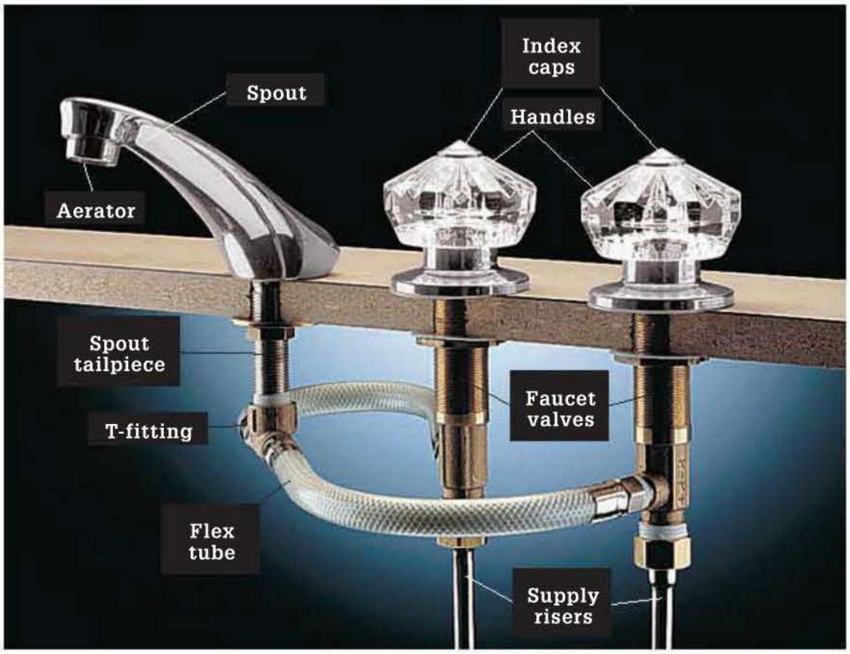
Index caps
Spout
Handles
Aerator
Spout tailpiece
Faucet valves
T-fitting
Flex tube
Supply risers

Widespread lavatory faucets have valves that are independent from the spout so they can be configured however you choose, provided that your flex tube connectors are long enough to span the distance.

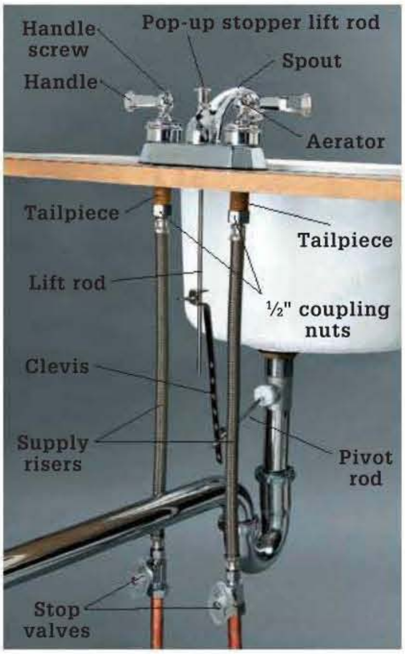
Handle screw
Pop-up stopper lift rod
Spout
Handle
Aerator
Tailpiece
Tailpiece
Lift rod
½" coupling nuts
Clevis
Supply risers
Pivot rod
Stop valves

Single-body lavatory faucets have both valves and the spout permanently affixed to the faucet body. They do not offer flexibility in configurations, but they are very simple to install.

The pop-up stopper fits into the drain opening so the stopper will close tightly against the drain flange when the pop-up handle is lifted up.

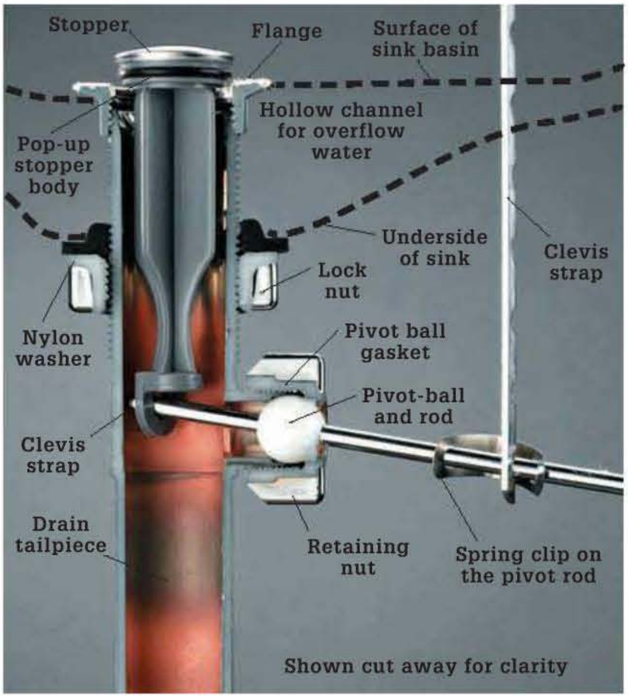
Stopper
Flange
Surface of sink basin
Hollow channel for overflow water
Pop-up stopper body
Underside of sink
Clevis strap
Lock nut
Nylon washer
Pivot ball gasket
Pivot-ball and rod
Clevis strap
Drain tailpiece
Retaining nut
Spring clip on the pivot rod
Shown cut away for clarity

The linkage that connects the pop-up stopper to the pop-up handle fits into a male-threaded port in the drain tailpiece. Occasionally the linkage will require adjustment or replacement.

How to Install a Widespread Faucet

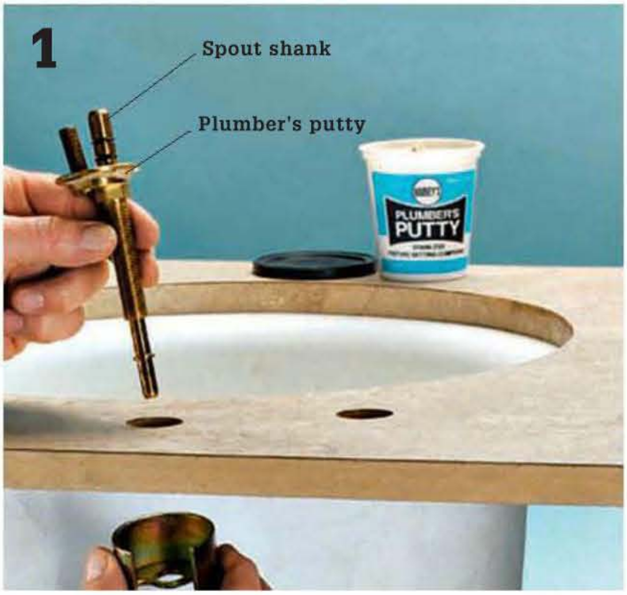
1
Spout shank
Plumber's putty
PLUMBER'S PUTTY

Insert the shank of the faucet spout through one of the holes in the sink deck (usually the center hole but you can offset it in one of the end holes if you prefer). If the faucet is not equipped with seals or O-rings for the spout and handles, pack plumber's putty on the undersides before inserting the valves into the deck. *Note: If you are installing the widespread faucet in a new sink deck, drill three holes of the size suggested by the faucet manufacturer.*

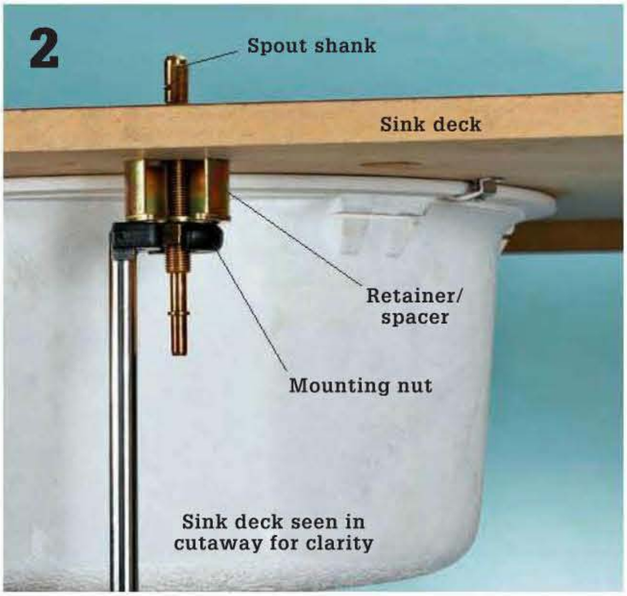
2
Spout shank
Sink deck
Retainer/ spacer
Mounting nut
Sink deck seen in cutaway for clarity

In addition to mounting nuts, many spout valves for widespread faucets have an open-retainer fitting that goes between the underside of the deck and the mounting nut. Others have only a mounting nut. In either case, tighten the mounting nut with pliers or a basin wrench to secure the spout valve. You may need a helper to keep the spout centered and facing forward.

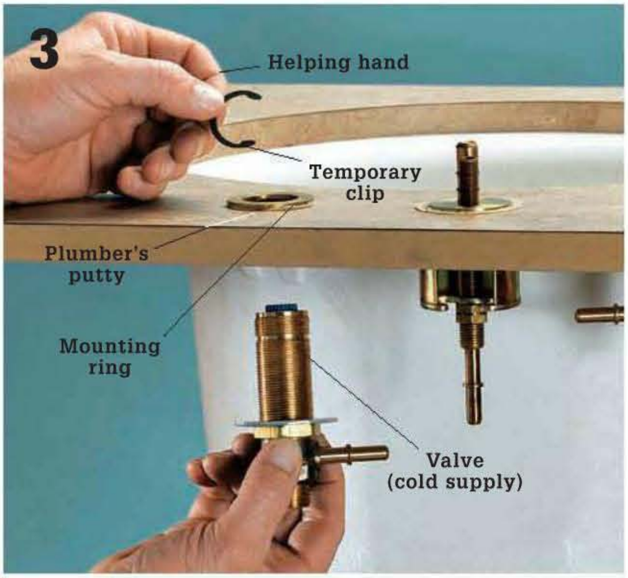
3
Helping hand
Temporary clip
Plumber's putty
Mounting ring
Valve (cold supply)

Mount the valves to the deck using whichever method the manufacturer specifies (it varies quite a bit). In the model seen here, a mounting ring is positioned over the deck hole (with plumber's putty seal) and the valve is inserted from below. A clip snaps onto the valve from above to hold it in place temporarily (you'll want a helper for this).

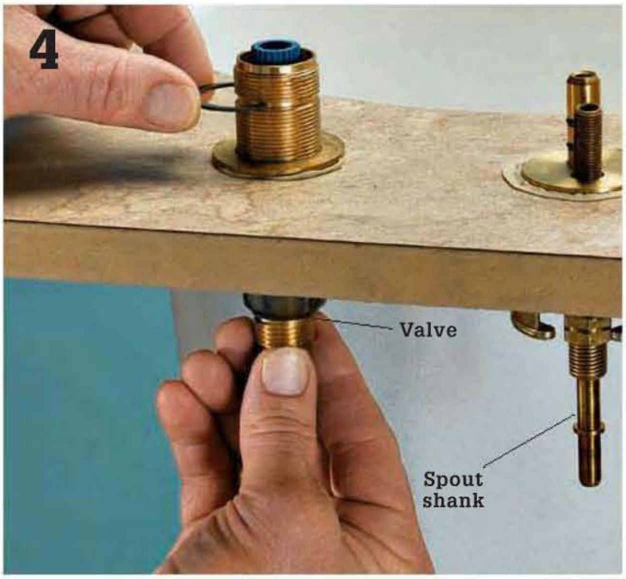
4
Valve
Spout shank

From below, thread the mounting nuts that secure the valves to the sink deck. Make sure the cold water valve (usually has a blue cartridge inside) is in the right-side hole (from the front) and the hot water valve (red cartridge) is in the left hole (from the front). Install both valves.

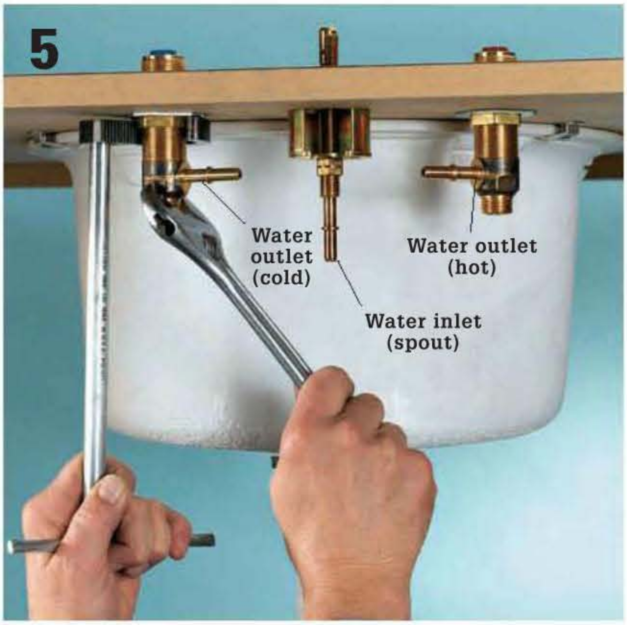
5
Water outlet (cold)
Water outlet (hot)
Water inlet (spout)

Once you've started the nut on the threaded valve shank, secure the valve with a basin wrench, squeezing the lugs where the valve fits against the deck. Use an adjustable wrench to finish tightening the lock nut onto the valve. The valves should be oriented so the water outlets are aimed at the inlet on the spout shank.

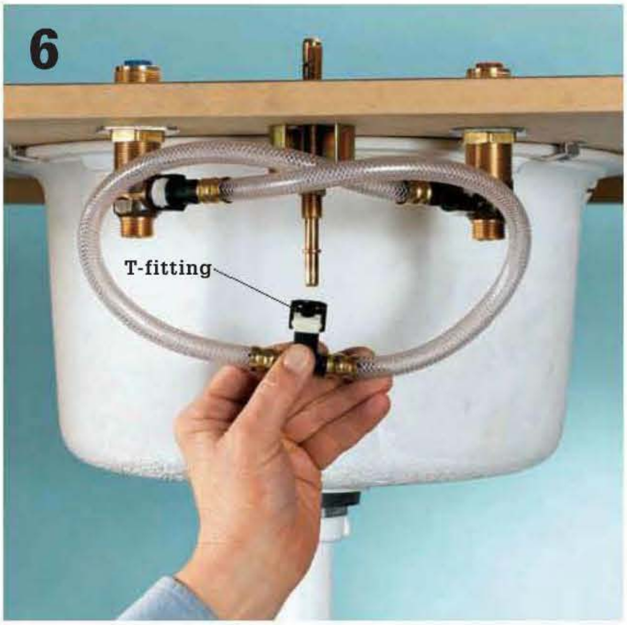

Attach the flexible supply tubes (supplied with the faucet) to the water outlets on the valves. Some twist onto the outlets, but others (like the ones above) click into place. The supply hoses meet in a T-fitting that is attached to the water inlet on the spout.

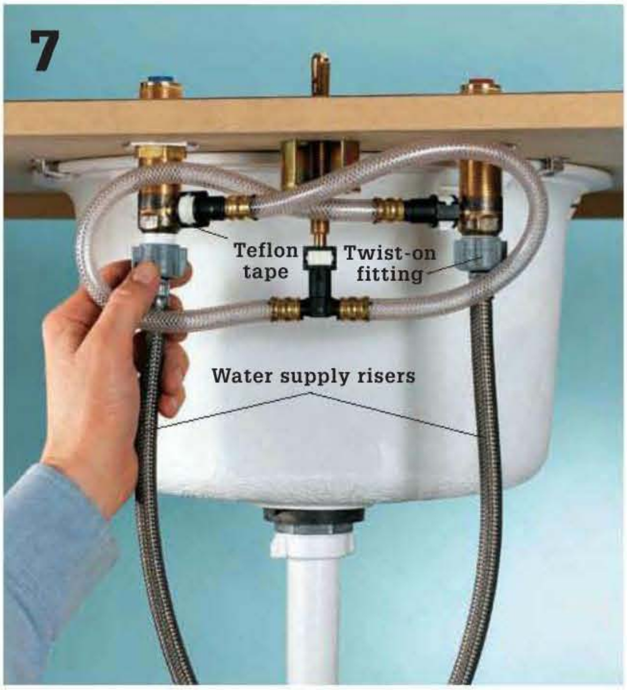
7
Teflon tape
Twist-on fitting
Water supply risers

Attach flexible braided-metal supply risers to the water stop valves and then attach the tubes to the inlet port on each valve (usually with Teflon tape and a twist-on fitting at the valve end of the supply riser).

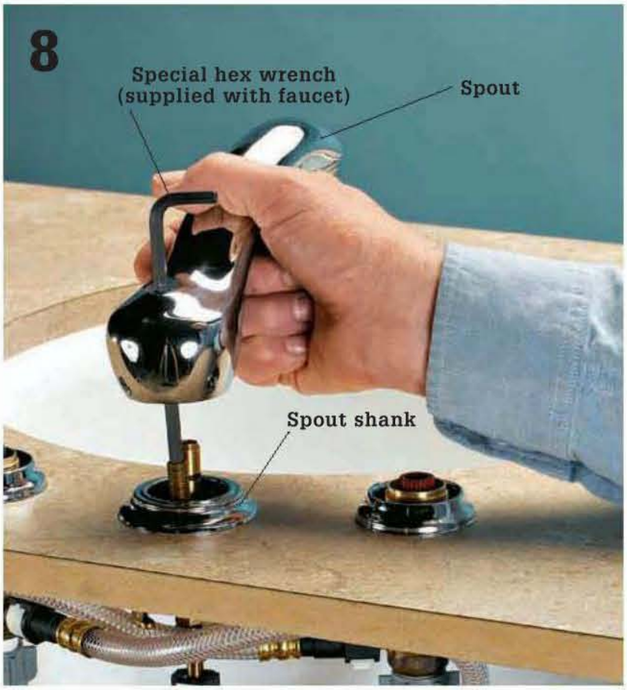
8
Special hex wrench (supplied with faucet)
Spout
Spout shank

Attach the spout. The model shown here comes with a special hex wrench that is threaded through the hole in the spout where the lift rod for the pop-up drain will be located. Once the spout is seated cleanly on the spout shank, tighten the hex wrench to secure the spout. Different faucets will use other methods to secure the spout to the shank.

(continued)

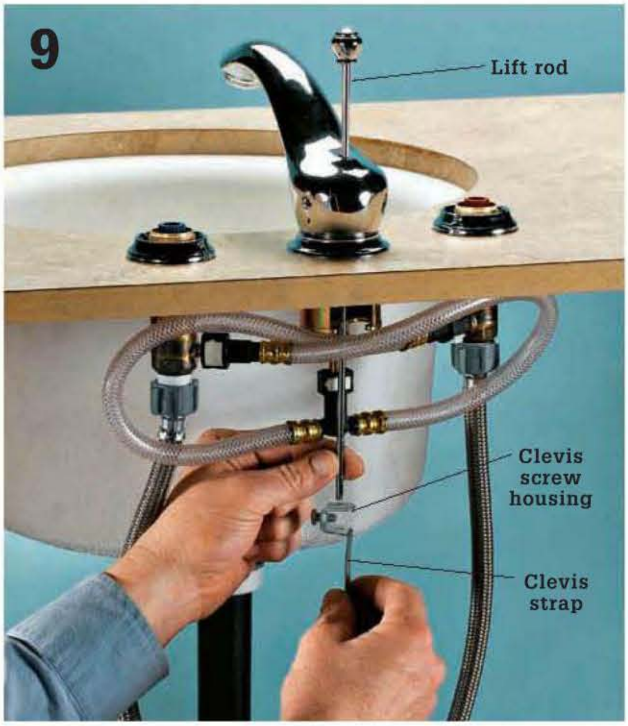

If your sink did not have a pop-up stopper, you'll need to replace the sink drain tailpiece with a pop-up stopper body (often supplied with the faucet). See page 473. Insert the lift rod through the hole in the back of the spout and, from below, thread the pivot rod through the housing for the clevis screw.

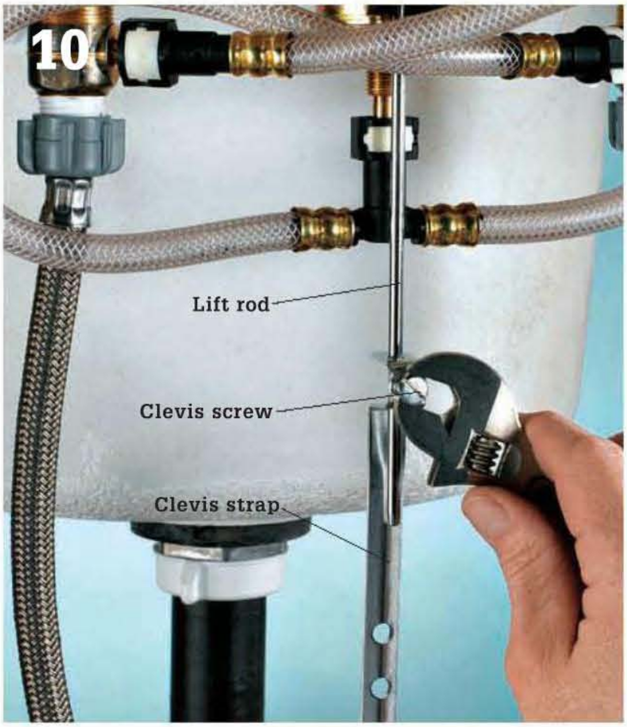

Attach the clevis strap to the pivot rod that enters the pop-up drain body, and adjust the position of the strap so it raises and lowers properly when the lift rod is pulled up. Tighten the clevis screw at this point. It's hard to fit a screwdriver in here, so you may need to use a wrench or pliers.

Attach the faucet handles to the valves using whichever method is required by the faucet manufacturer. Most faucets are designed with registration methods to ensure that the handles are symmetrical and oriented in an ergonomic way once you secure them to the valves.

Turn on the water supply and test the faucet. Remove the faucet aerator so any debris in the lines can clear the spout.

Variation: How to Install a Single-body Faucet ▸

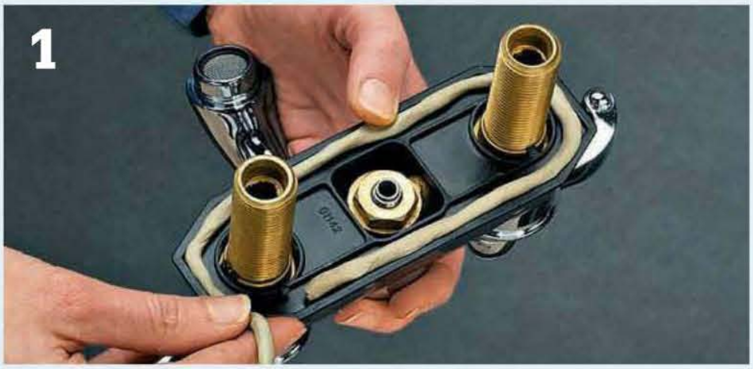

Most faucets come with a plastic or foam gasket to seal the bottom of the faucet to the sink deck. These gaskets will not always form a watertight seal. If you want to ensure no water gets below the sink, discard the seal and press a ring of plumber's putty into the sealant groove built into the underside of the faucet body.

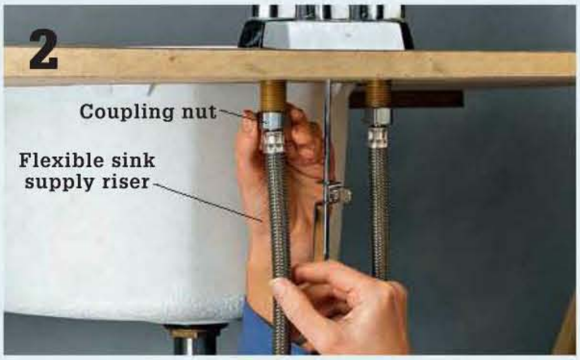

Insert the faucet tailpieces through the holes in the sink. From below, thread washers and mounting nuts over the tailpieces, then tighten the mounting nuts with a basin wrench until snug. Put a dab of pipe joint compound on the threads of the stop valves and thread the metal nuts of the flexible supply risers to these. Wrench tighten about a half-turn past hand tight. Overtightening these nuts will strip the threads. Now tighten the coupling nuts to the faucet tailpieces with a basin wrench.

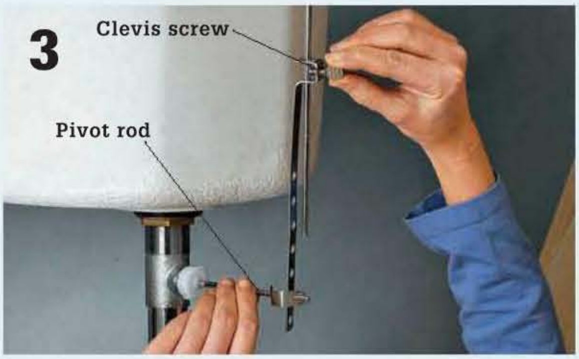

Slide the lift rod of the new faucet into its hole behind the spout. Thread it into the clevis past the clevis screw. Push the pivot rod all the way down so the stopper is open. With the lift rod also all the way down, tighten the clevis to the lift rod.

Grease the fluted valve stems with heatproof grease, then put the handles in place. Put a drop of Loctite on each handle screw before tightening it on. (This will keep your handles from coming loose.) Cover each handle screw with the appropriate index cap—Hot or Cold.

Unscrew the aerator from the end of the spout. Turn the hot and cold water taps on full. Turn the water back on at the stop valves and flush out the faucet for a couple of minutes before turning off the water at the faucet. Check the riser connections for drips. Tighten a compression nut only until the drip stops.

How to Install a Pop-up Drain

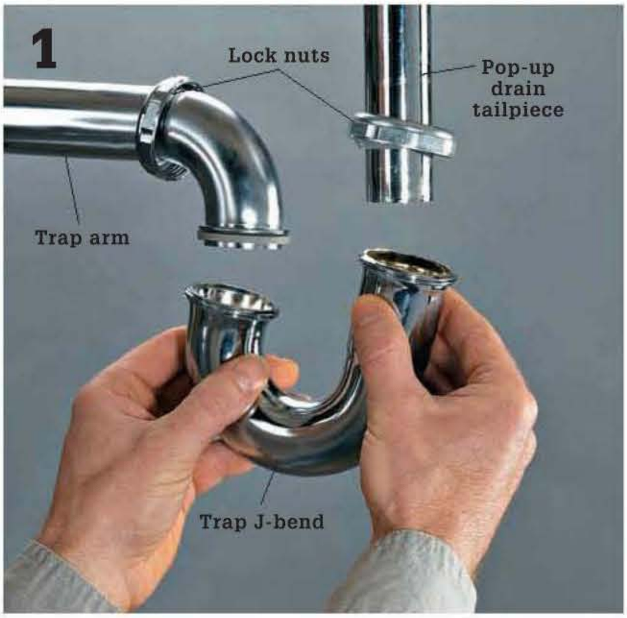
1
Lock nuts
Pop-up drain tailpiece
Trap arm
Trap J-bend

Put a basin under the trap to catch water. Loosen the nuts at the outlet and inlet to the trap J-bend by hand or with channel-type pliers and remove the bend. The trap will slide off the pop-up body tailpiece when the nuts are loose. Keep track of washers and nuts and their up/down orientation by leaving them on the tubes.

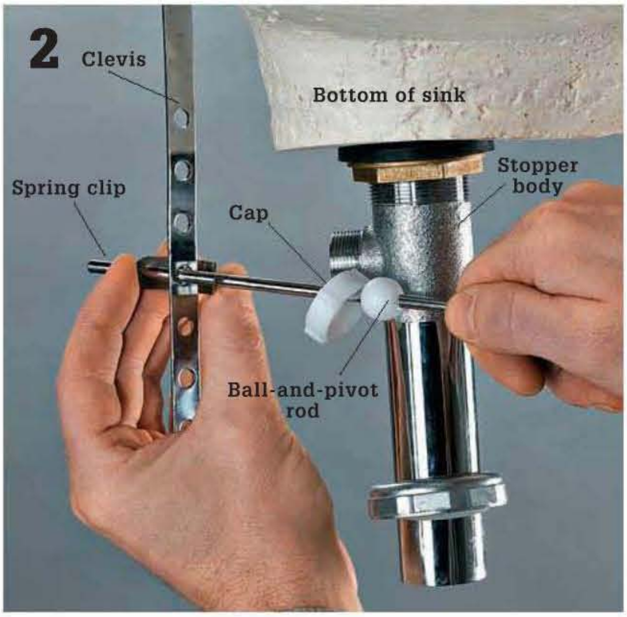
2
Clevis
Bottom of sink
Stopper body
Spring clip
Cap
Ball-and-pivot rod

Unscrew the cap holding the ball-and-pivot rod in the pop-up body and withdraw the ball. Compress the spring clip on the clevis and withdraw the pivot rod from the clevis.

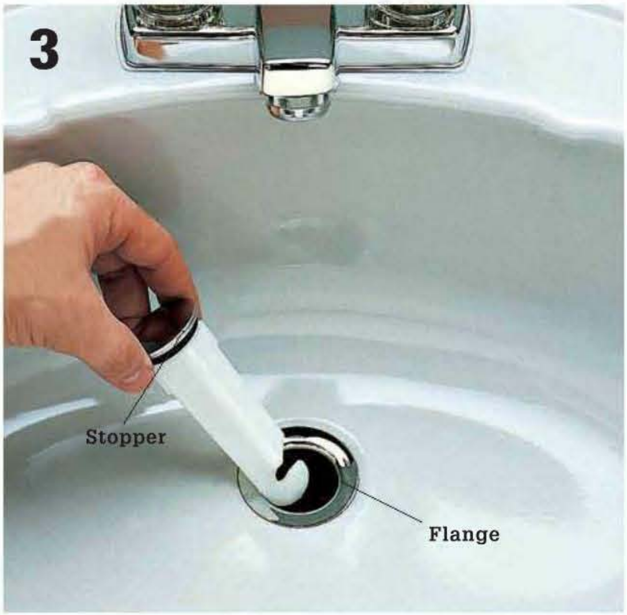
3
Stopper
Flange

Remove the pop-up stopper. Then, from below, remove the lock nut on the stopper body. If needed, keep the flange from turning by inserting a large screwdriver in the drain from the top. Thrust the stopper body up through the hole to free the flange from the basin, and then remove the flange and the stopper body.

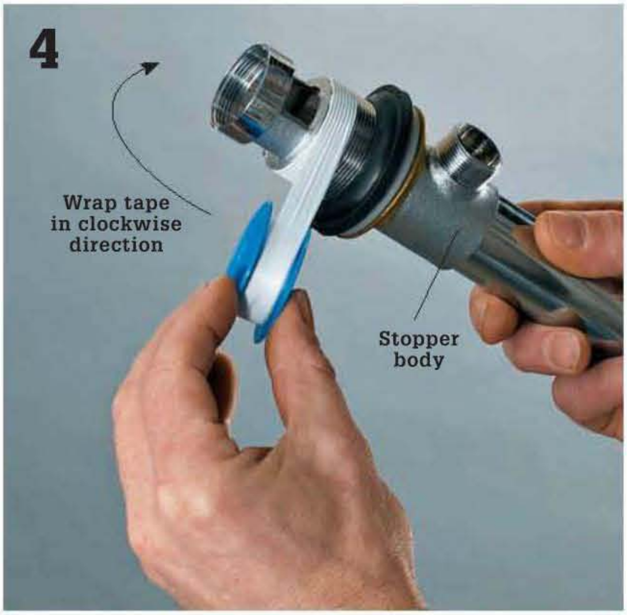
4
Wrap tape in clockwise direction
Stopper body

Clean the drain opening above and below, and then thread the locknut all the way down the new pop-up body, followed by the flat washer and the rubber gasket (beveled side up). Wrap three layers of Teflon tape clockwise onto the top of the threaded body. Make a ½"-dia. snake from plumber's putty, form it into a ring, and stick the ring underneath the drain flange.

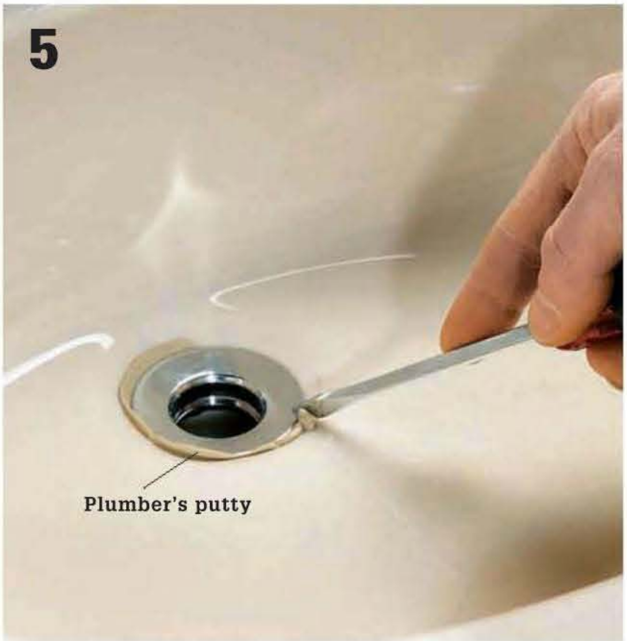

From below, face the pivot rod opening directly back toward the middle of the faucet and pull the body straight down to seat the flange. Thread the locknut/washer assembly up under the sink, then fully tighten the locknut with channel-type pliers. Do not twist the flange in the process, as this can break the putty seal. Clean off the excess plumber's putty from around the flange.

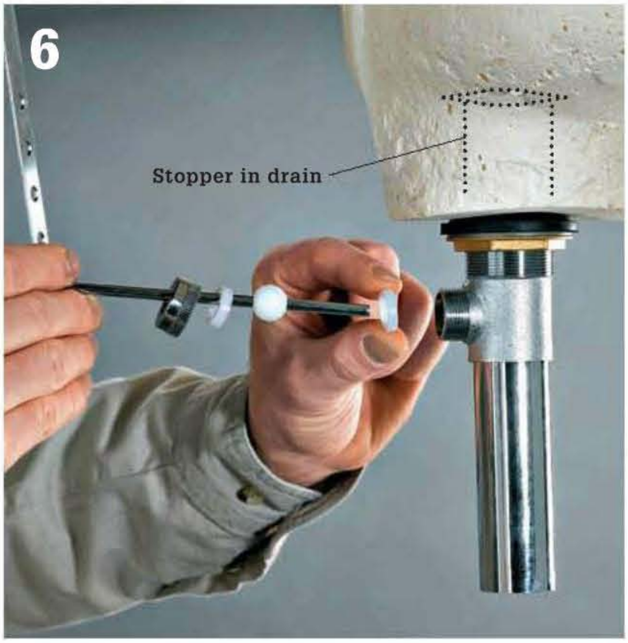

Drop the pop-up stopper into the drain hole so the hole at the bottom of its post is closest to the back of the sink. Put the beveled nylon washer into the opening in the back of the pop-up body with the bevel facing back.

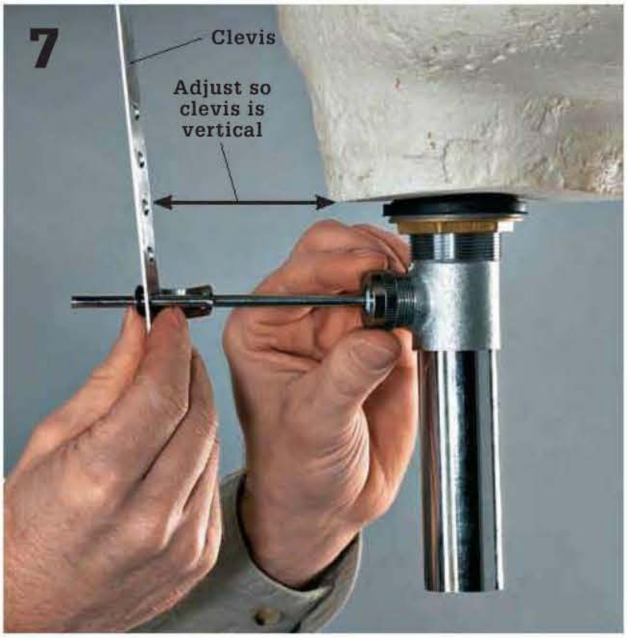

Put the cap behind the ball on the pivot rod as shown. Sandwich a hole in the clevis with the spring clip and thread the long end of the pivot rod through the clip and clevis. Put the ball end of the pivot rod into the pop-up body opening and into the hole in the the stopper stem. Screw the cap on to the pop-up body over the ball.

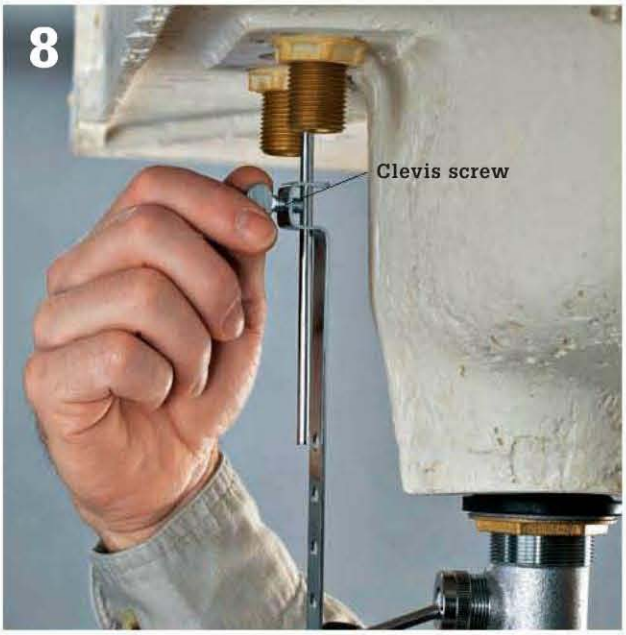

Loosen the clevis screw holding the clevis to the lift rod. Push the pivot rod all the way down (which fully opens the pop-up stopper). With the lift rod also all the way down, tighten the clevis screw to the rod. If the clevis runs into the top of the trap, cut it short with your hacksaw or tin snips. Reassemble the J-bend trap.

Installing Tub Spouts

In many situations, replacing a bathtub spout can be almost as easy as hooking up a garden hose to an outdoor spigot. There are some situations where it is a bit more difficult, but still pretty simple. The only time it's a real problem is when the spout is attached to a plain copper supply nipple, rather than a threaded nipple. You'll know this is the case if the spout has a setscrew on the underside where it meets the wall. Many bathtub spouts are sold in kits with a matching showerhead and handle or handles. But for a simple one-for-one replacement, spouts are sold separately. You just need to make sure the new spout is compatible with the existing nipple.

Tub spouts can be relatively complicated plumbing fittings, often performing three or four important functions. The spout itself is simple enough, since its only function is to deliver bathwater to the tub. But the diverter network and pop-up drain contain multiple moving parts that require precise adjustment and occasional repair or replacement (see photo, next page). The diverter is basically a stop valve that's activated by a lever or knob to block flow of water from the spout, forcing it up to a showerhead or out through a handheld showerhead, as seen here.

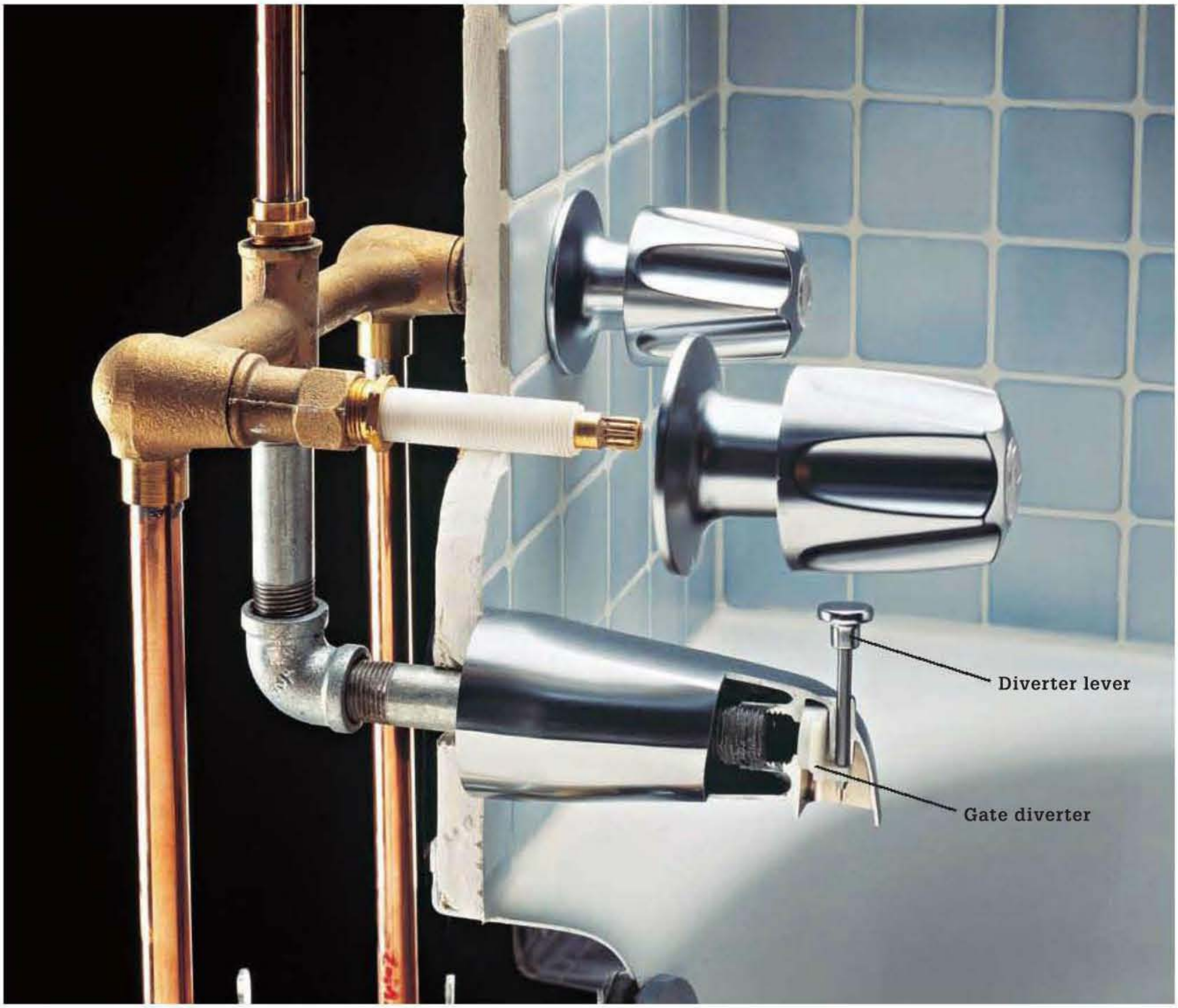

In many bathtub/shower plumbing systems, the spout has the important job of housing the diverter—a gate inside the spout that is operated by a lever with a knob for pulling. An open gate allows water to come out of the spout when the faucet is turned on. When the diverter is pulled shut, the water is redirected up a riser pipe and to the showerhead. Failure of the diverter is one of the most common reasons for replacing a spout.

Tools & Materials ▸

- Pliers
- Adjustable pliers
- Channel-type pliers
- Rags
- Masking tape
- Screwdrivers
- Allen wrenches
- Replacement spout
- Measuring tape
- Teflon tape
- Utility knife

Installation Tip ▸

If you are installing a brand new tub/shower faucet, it likely contains an antiscald device. Most of these devices work by sensing a decrease in cold water flow and compensating by stemming the hot water flow too. As you install the new plumbing, be aware that faucets with antiscald protection will not function at all if both water supply tubes are not turned on at the stop valve. In other words: you can't test the hot and cold supply independently.

How to Install a Slip-fit Spout

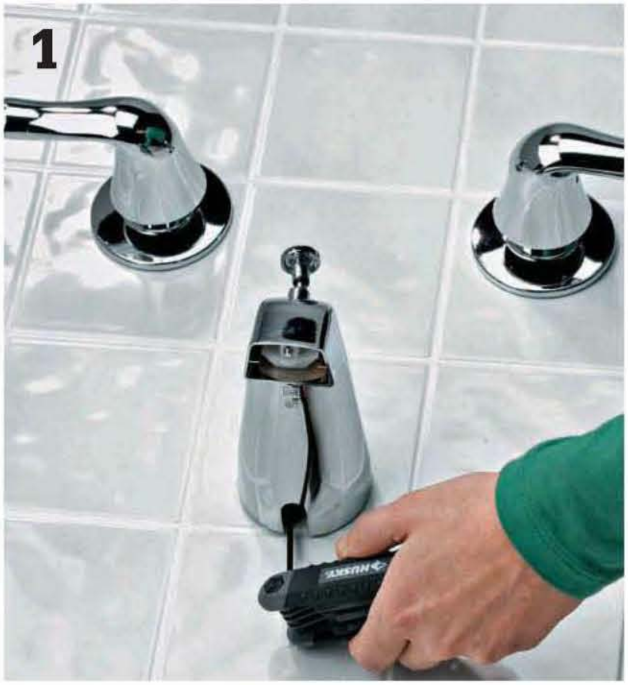

Slip fitting: Check underneath the tub spout to look for an access slot or cutout, which indicates the spout is a slip-fit style that is held in place with a setscrew and mounted on a copper supply nipple. Loosen the screw with a hex (Allen) wrench. Pull off the spout.

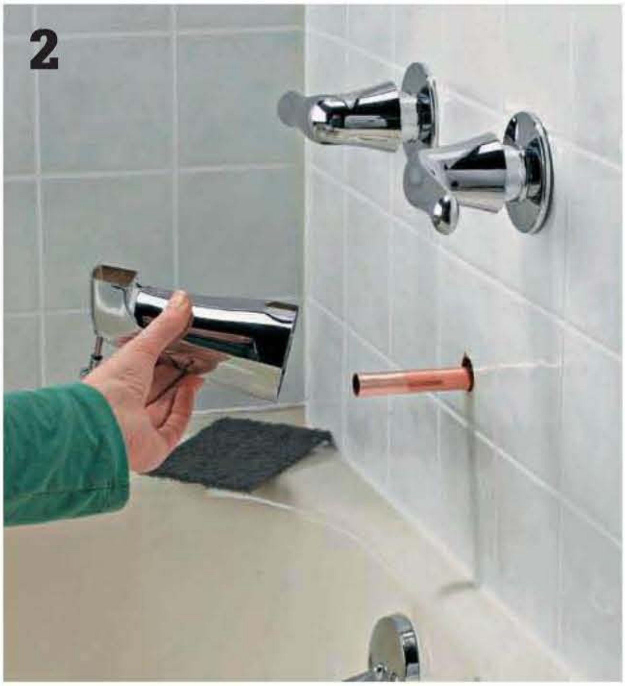

Clean the copper nipple with steel wool. If you find any sharp edges where the nipple was cut, smooth them out with emery paper. Then, insert the O-ring that comes with the spout onto the nipple (see the manufacturer's instructions) and slide the spout body over the nipple in an upside-down position.

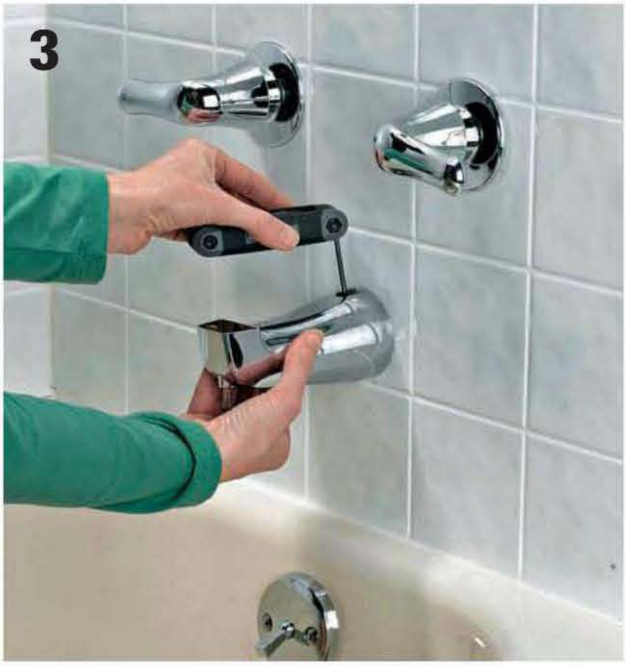

With the spout upside down for ease of access, tighten the setscrews on the clamp, working through the access slot or cutout, until you feel resistance.

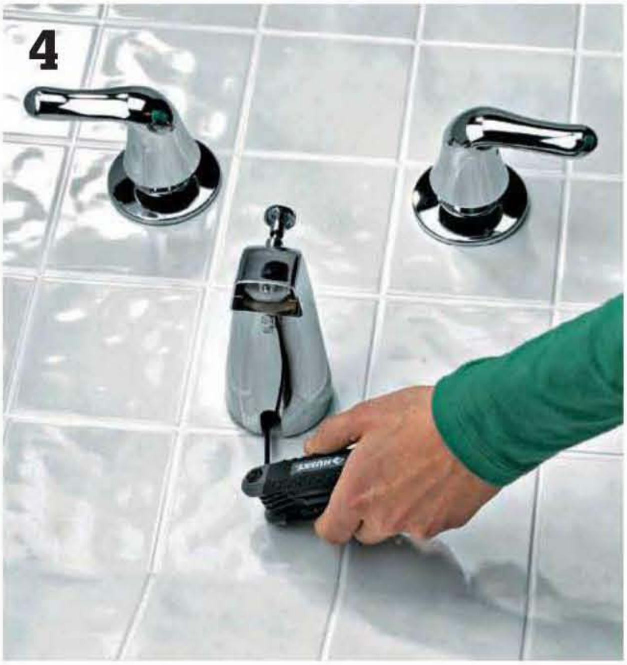

Spin the spout so it's right-side up and then tighten the setscrew from below, making sure the wall end of the spout is flush against the wall. Do not overtighten the setscrew.

How to Install a Threaded Spout

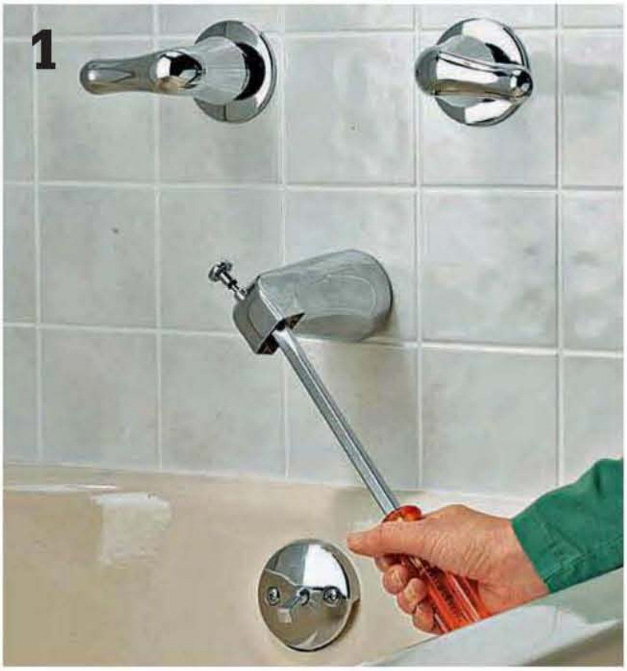

If you see no setscrew or slot on the underside of the spout, it is attached to a threaded nipple. Unscrew the tub spout by inserting a heavy-duty flat screwdriver into the spout opening and spinning it counterclockwise.

Tool Tip ▸

Alternatively, grip the spout with a padded pipe wrench or channel-type pliers. Buy a compatible replacement spout at a home center or hardware store.

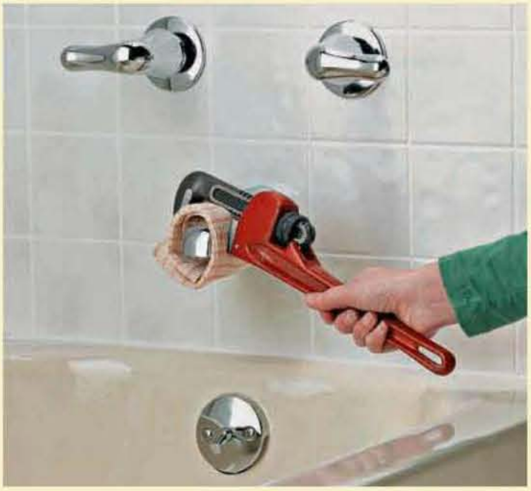

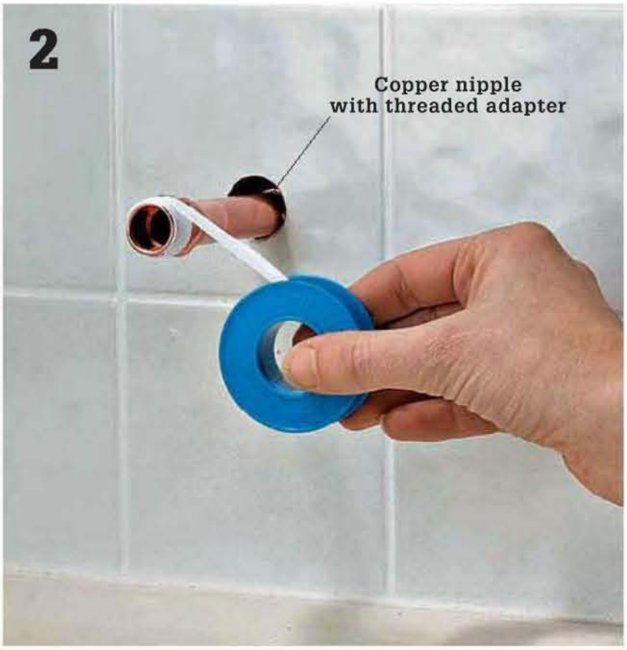

Wrap several courses of Teflon tape clockwise onto the pipe threads of the nipple. Using extra Teflon tape on the threads creates resistance if the spout tip points past six o'clock when tight.

Twist the new spout onto the nipple until it is flush against the wall and the spout is oriented properly. If the spout falls short of six o'clock, you may protect the finish of the spout with tape and twist it a little beyond hand tight with your channel-type pliers—but don't overdo it; the fitting can crack.

Installing Vanity Lights

Many bathrooms have a single fixture positioned above the vanity, but a light source in this position casts shadows on the face and makes grooming more difficult. Placing light fixtures on either side of the mirror is a better arrangement.

For a remodel, mark the mirror location, run cable, and position boxes before drywall installation. You can also retrofit by installing new boxes and drawing power from the existing fixture.

The light sources should be at eye level; 66" is typical. The size of your mirror and its location on the wall may affect how far apart you can place the sconces, but 36 to 40" apart is a good guideline.

Tools & Materials ▸

Drywall saw
Drill
Combination tool
Circuit tester
Screwdrivers
Hammer
Electrical boxes and braces
Vanity light fixtures
NM cable
Wire connectors

Vanity lights on the sides of the mirror provide even lighting.

How to Replace Vanity Lights in a Finished Bathroom

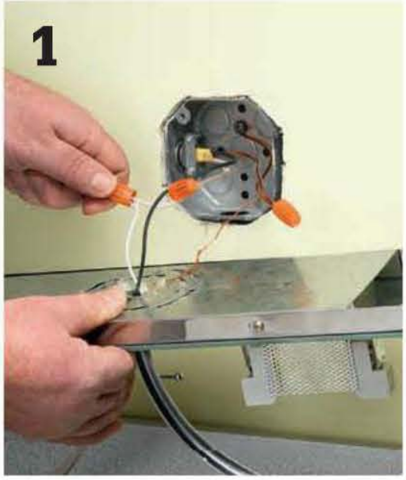

Turn off the power at the service panel. Remove the old fixture from the wall and test to make sure that the power is off. Then remove a strip of drywall from around the old fixture to the first studs beyond the approximate location of the new fixtures. Make the opening large enough that you have room to route cable from the existing fixture to the boxes.

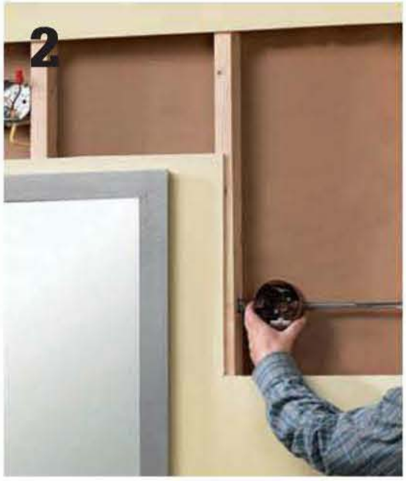

Mark the location for the fixtures and install new boxes. Install the boxes about 66" above the floor and 18 to 20" from the centerline of the mirror (the mounting base of some fixtures is above or below the bulb, so adjust the height of the bracing accordingly). If the correct location is on or next to a stud, you can attach the box directly to the stud, otherwise you'll need to install blocking or use boxes with adjustable braces (shown).

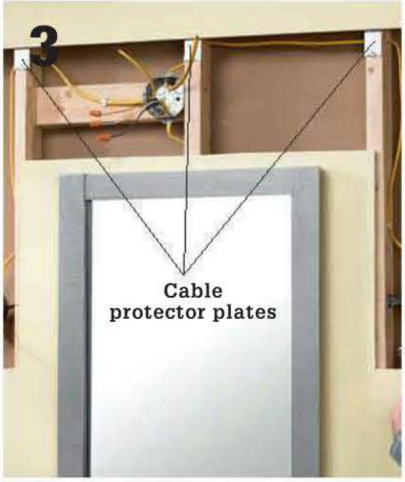

Open the side knockouts on the electrical box above the vanity. Then drill ⅝" holes in the centers of any studs between the old fixture and the new ones. Run two NM cables from the new boxes for the fixtures to the box above the vanity. Protect the cable with metal protector plates. Secure the cables with cable clamps, leaving 11" of extra cable for making the connection to the new fixtures. Remove sheathing and strip insulation from the ends of the wires.

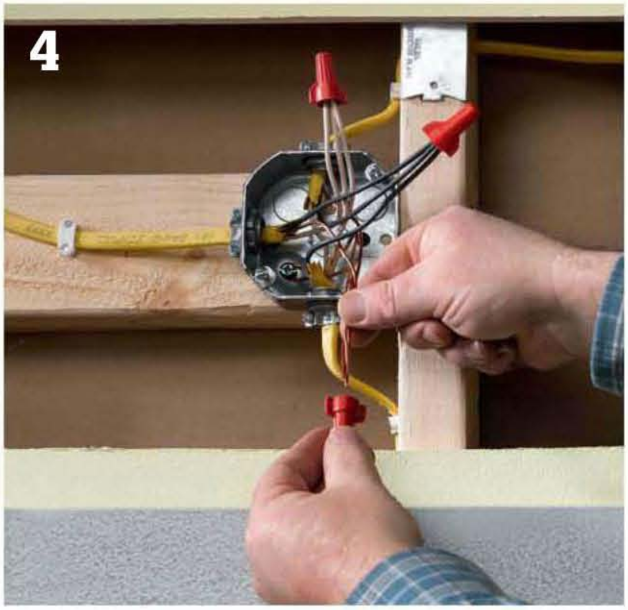

Connect the white wires from the new cables to the white wire from the old cable, and connect the black wires from the new cables to the black wire from the old cable. Connect the ground wires. Cover all open boxes and then replace the drywall, leaving openings for the fixture and the old box. (Cover the old box with a solid junction box cover plate.)

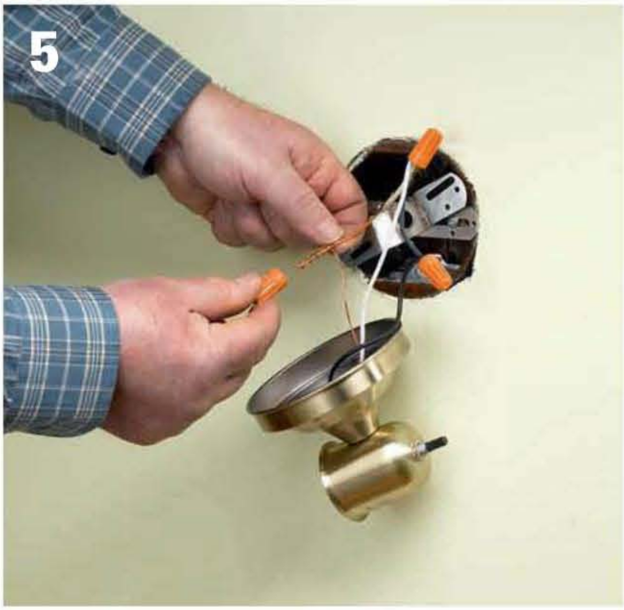

Install the fixture mounting braces on the boxes. Attach the fixtures by connecting the black circuit wire to the black fixture wire, and connecting the white circuit wire to the white fixture wire. Connect the ground wires. Position each fixture over each box, and attach with the mounting screws. Restore power and test the circuit.

Installing Grab Bars

Bathrooms are beautiful with their shiny ceramic tubs, showers, and floors, but add water and moisture to the mix and you've created the perfect conditions for a fall. The good news is many falls in the bathroom can be avoided by installing grab bars at key locations.

Grab bars help family members steady themselves on slippery shower, tub, and other floor surfaces. Plus, they provide support for people transferring from a wheelchair or walker to the shower, tub, or toilet.

Grab bars come in a variety of colors, shapes, sizes, and textures. Choose a style with a 1¼ to 1½" diameter that fits comfortably between your thumb and fingers. Then properly install it 1½" from the wall with anchors that can support at least 250 pounds.

The easiest way to install grab bars is to screw them into wall studs or into blocking or backing attached to studs. Blocking is a good option if you are framing a new bathroom or have the wall surface removed during a major remodel (see Illustration A). Use 2 × 6 or 2 × 8 lumber to provide room for adjustments, and fasten the blocks to the framing with 16d nails. Note the locations of your blocking for future reference.

As an alternative, cover the entire wall with ¾" plywood backing secured with screws to the wall framing, so you can install grab bars virtually anywhere on the wall (see Illustration B).

Grab bars can be installed in areas without studs. For these installations, use specialized heavy-duty hollow-wall anchors designed to support at least 250 pounds.

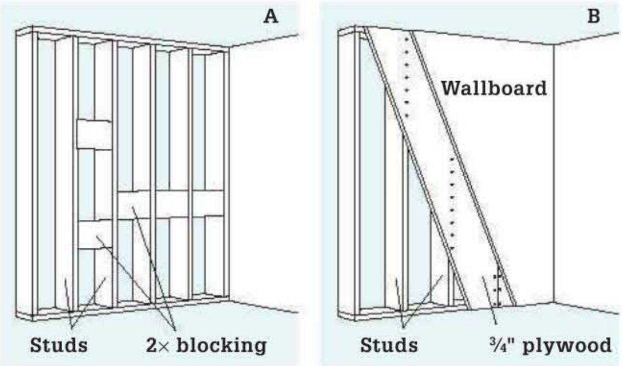

Blocking or backing is required for secure grab bars. If you know where the grab bars will be located, add 2× blocking between studs (Illustration A). You also can cover the entire wall with ¾" plywood backing, which allows you to install grab bars virtually anywhere on the wall.

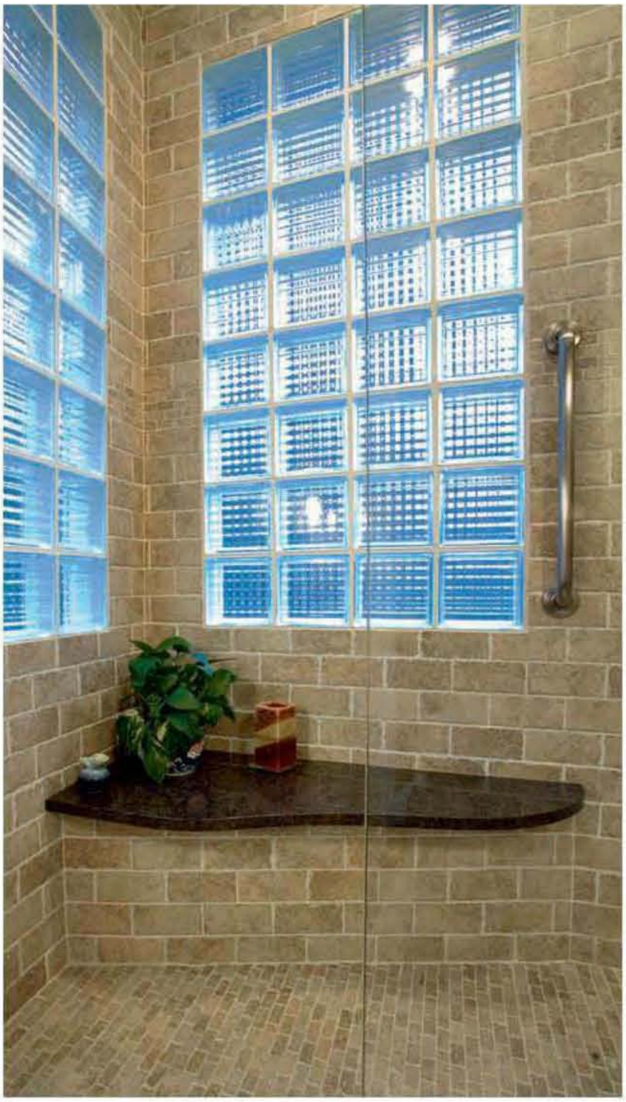

Grab bars promote independence in the bathroom, where privacy is especially important. Grab bars not only help prevent slips and falls, they also help people steady themselves in showers and lower themselves into tubs.

Tools & Materials ▸

Measuring tape
Pencil
Stud finder
Level
Drill
Masonry bit
Grab bar
Hollow-wall anchors
#12 stainless steel screws
Silicone caulk

How to Install Grab Bars

Locate the wall studs in the installation area, using a stud finder. If the area is tiled, the stud finder may not detect studs, so try to locate the studs above the tile, if possible, then use a level to transfer the marks lower on the wall. Otherwise, you can drill small, exploratory holes through grout joints in the tile, then fill the holes with silicone caulk to seal them. Be careful not to drill into pipes.

Mark the grab bar height at one stud location, then use a level to transfer the height mark to the stud that will receive the other end of the bar. Position the grab bar on the height marks so at least two of the three mounting holes are aligned with the stud centers. Mark the mounting hole locations onto the wall.

Drill pilot holes for the mounting screws. If you are drilling through tile, start with a small bit (about ⅛"), then redrill the hole with the larger bit. For screws that won't hit studs, drill holes for wall anchors, following the manufacturer's directions for sizing. Install anchors, if necessary.

Apply a continuous bead of silicone caulk to the back side of each bar end (inset). Secure the bar to the studs using #12 stainless steel screws (the screws should penetrate the stud by at least 1"). Install a stainless steel screw or bolt into the wall anchors. Test the bar to make sure it's secure.

Installing Glass Shelving

Glass shelving is unobtrusive, so it can fit many styles of bathrooms—from sleek modern to Victorian. You can find a wide variety of shelving available in home stores and online.

Most glass shelves are held in place with metal mounts. How the shelves are secured to the mounts and how the mounts are attached to the wall differs. Most shelves have a hidden bracket that is secured to the wall. The mount then slips over the bracket and is secured with a setscrew. The most basic models may have mounts that are screwed directly into the wall with exposed screws. The directions here are for shelving that uses hidden brackets.

If you are installing shelves on a tiled wall, mount the brackets in grout lines if at all possible to minimize the possibility of cracking the tiles. Many glass shelves have some variability in the distance between the mounts.

Tools & Materials ▸

Level
Drill
Pencil
Glass shelves

Glass shelves fit any style and size of bathroom. They are held in place with metal mounts attached to the walls.

How to Install Glass Shelves

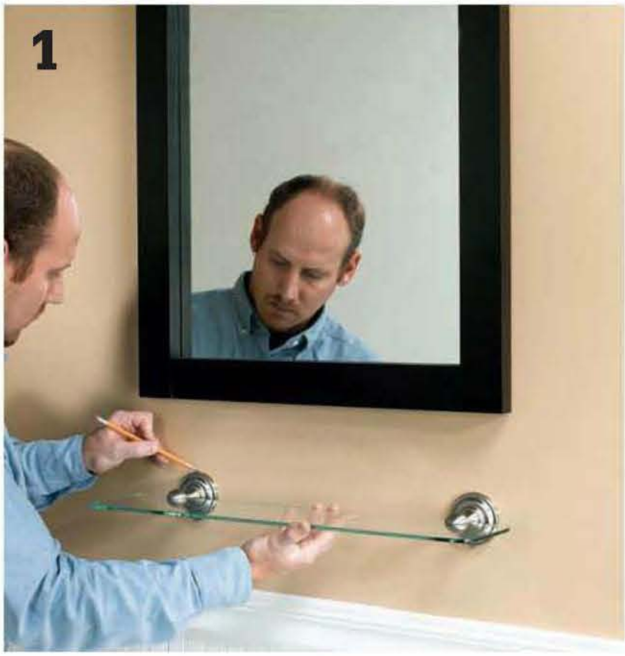

Assemble the shelf and shelf holders (not the brackets). Hold the shelf against the wall in the desired location. On the wall, mark the center point of each holder, where the setscrew is. Remove the shelf from the holders and set aside.

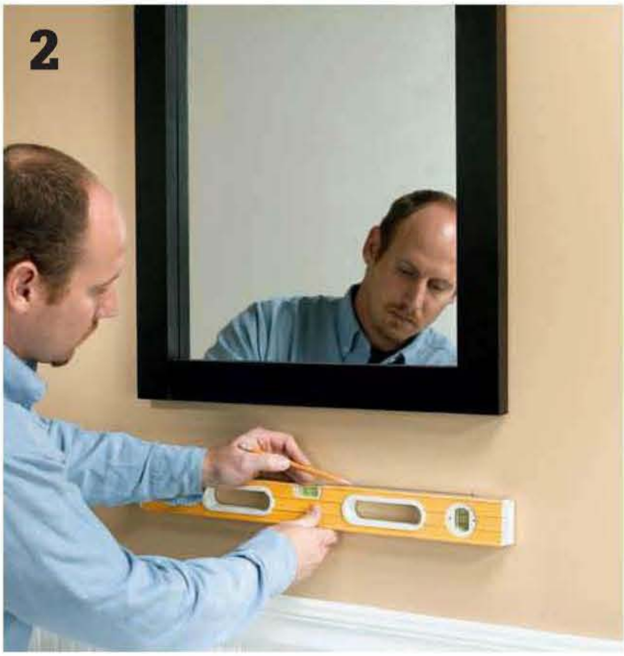

Remove the shelves and use the level to extend the mark into a 3" vertical line. Use the level to mark a horizontal line across the centers of these lines.

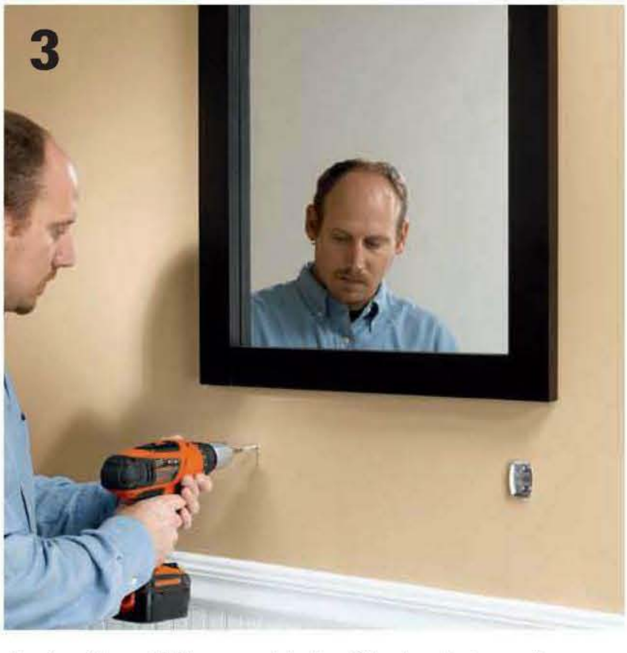

Center the middle round hole of the bracket over the intersection of the vertical and horizontal lines. Mark the center of each of the oblong holes. Put the bracket aside and drill a ¼" hole at each mark. Insert the included wall anchors in the holes. Replace the bracket and insert the screws into the wall anchors and drive the screws. Repeat for the second bracket.

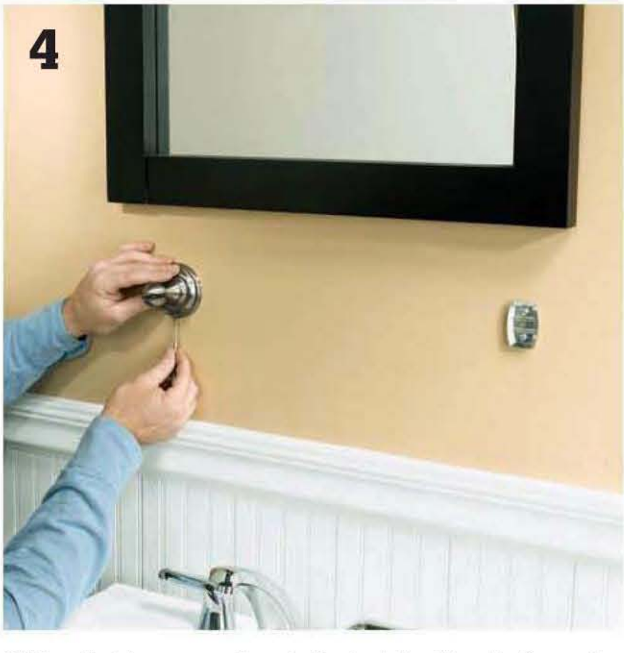

Slide a holder over a bracket, check that the shelf mount is level, and tighten the setscrew. Repeat with the other holder. Insert the shelf and fix in place. Check the shelf for level. If it's not level, remove one holder and loosen the bracket screws. Slide the bracket up or down to make the unit level. Replace the holder and shelf.

Installing Towel Warmers

Here's a little bit of luxury that need not be limited to high-end hotel stays. You can have toasty towels in your own bathroom with an easy-to-install towel warmer. In a relatively cold room, this can make stepping out of the shower a much more pleasant experience.

Heated towel racks are available in a wide range of styles and sizes. Freestanding floor models as well as door- and wall-mounted versions can be plugged in for use when desired. Hardwired wall-mounted versions can be switched on when you enter the bathroom so your towels are warm when you step out of the shower. Although installing them requires some electrical skills, the hardwired models do not need to be located near wall receptacles and they do not have exposed cords or extension cords hanging on the wall. However, if you locate the warmer directly above an existing receptacle, you can save a lot of time and mess by running cable up from the receptacle to the new electrical box for the warmer.

Before installing hardwired models, check your local electrical codes for applicable regulations. If you are not experienced with home wiring, have an electrician do this job for you or opt for a plug-in model.

Tools & Materials ▸

- Drill
- Level
- Keyhole saw
- Wiring tools
- Phillips screwdriver
- Stud finder
- Wire connectors
- Retrofit electrical outlet box
- NM cable
- Towel warmer
- Pencil
- Masking tape

A hard-wired towel warmer offers the luxury of heated towels without the safety concerns of a plug-in device.

How to Install a Hardwired Towel Warmer

1

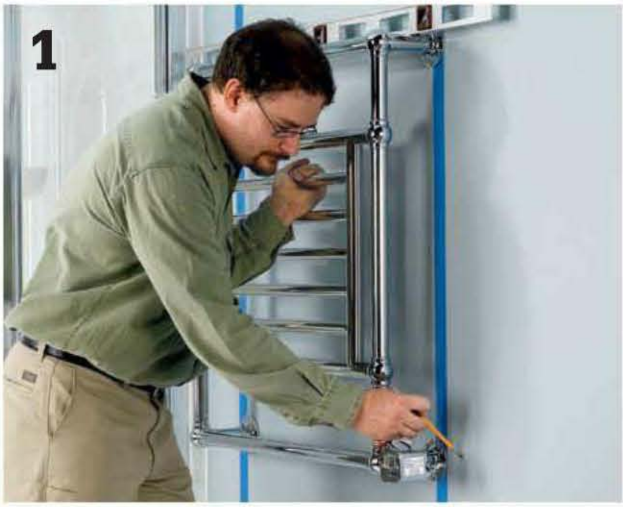

Use a stud finder to locate the studs in the area you wish to place the towel warmer. Mark the stud locations with masking tape or pencil lines. Attach the wall brackets to the towel warmer and hold the unit against the wall at least 7" from the floor and 3" from the ceiling or any overhang. Mark the locations of the wall bracket outlet plate (where the electrical connection will be made) and the mounting brackets.

2

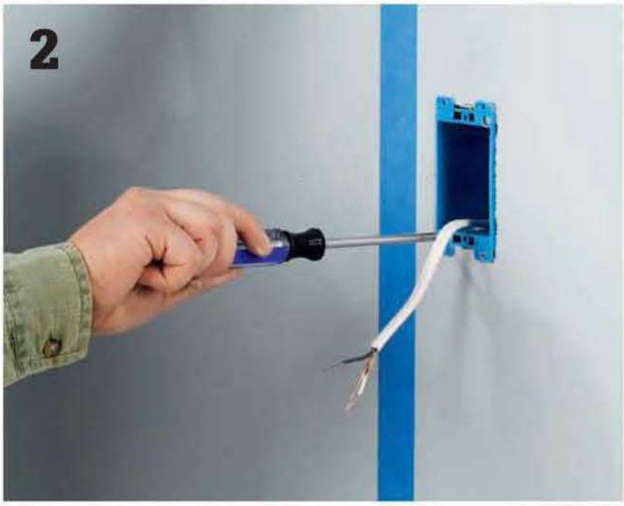

Shut off electrical power at the main service panel. At the mark for the wall bracket outlet, cut a hole in the wallcovering for a retrofit electrical box. Run NM cable from the opening to a GFCI-protected circuit (here, we ran cable down to a receptacle directly beneath it), or install a separate GFCI-protected circuit (you'll need to consult a wiring book or an electrician). Pull the cable through the hole in the retrofit box, and then tighten the cable clamp. Place the box in the hole flush with the wall surface and tighten the mounting screw in the rear of the box. Cut the wires so about 5" extends into the box and strip the insulation off at ⅜" from the end of each wire.

3

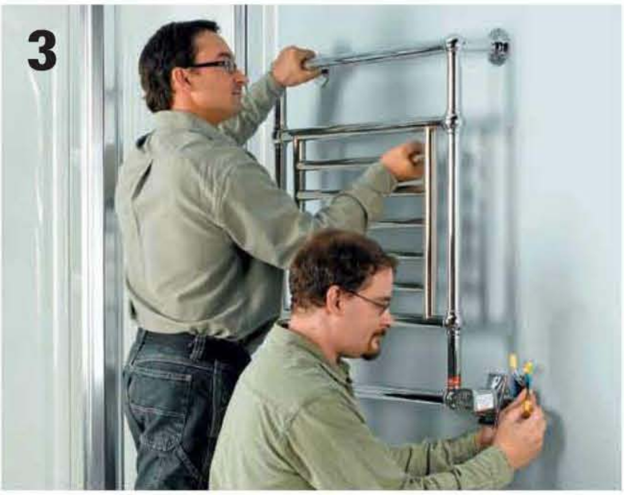

Position the towel warmer over the outlet box and mark the locations of the screw holes for the wall brackets. Make sure the appliance is level. Remove the warmer and drill ¼" pilot holes at the marked locations. If the marks are located over studs, drill ⅛" pilot holes. If not, push wall anchors into the holes. Thread the mounting screws through the brackets. Have a helper hold the towel rack in place and use wire connectors to connect the wires, including the ground wire, according to the instructions.

4

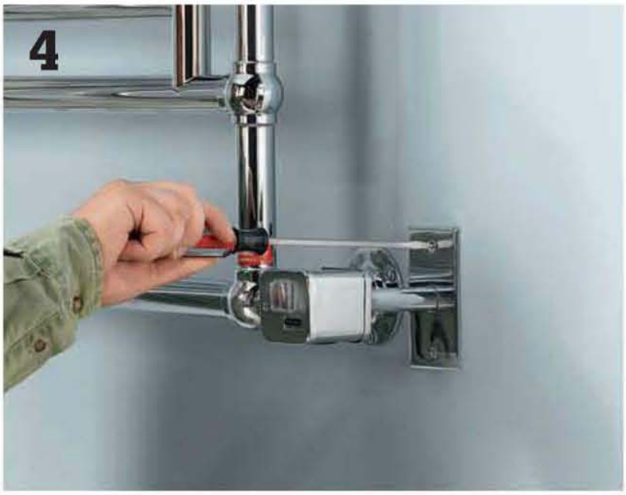

After the electrical connections are made, fasten the towel rack brackets to the wall. Turn on power and test the towel warmer. Finally, attach the electrical cover plate with integral on/off switch.

Basements & Attics

Is your house lacking just a little more space but you love it too much to think about moving? Has your life changed since you bought the house and now you find you need a home office for your freelance work or telecommuting, a place for the children in your life to play, or an extra bedroom for a relative? Or maybe now that your obligations have changed, you find yourself fantasizing about a space dedicated to your hobbies?

This chapter can help you determine how best to use your tantalizingly empty attic or basement, inspire you with great design ideas, and give you the practical how-to to tackle the projects on your own.

Over half of houses in North America are more than 30 years old. The attics and basements in these older homes are often significantly underutilized spaces. Even in new construction, attics and basements often are left unfinished so that the new homeowners can determine for themselves how best to incorporate the space.

In this chapter:

- Evaluating Your Attic
- Evaluating Your Basement
- Planning the Framing
- Installing Basement Floors
- Installing Attic Floors
- Installing Basement Walls
- Installing Attic Walls
- Finishing Attic Ceilings
- Adding Basement Egress Windows
- Installing Baseboard Heaters

Evaluating Your Attic

Start your attic evaluation with a quick framing inspection. If the roof is framed with rafters, you can continue to the next test. If it's built with trusses, however, consider remodeling your basement instead. The problem is that the internal supports in trusses leave too little space to work with, and trusses cannot be altered.

The next step is to check for headroom and overall floor space. Most building codes call for 7½ ft. of headroom over 50% of the "usable" floor space, which is defined as any space with a ceiling height of at least 5 ft. Remember that these minimums apply to the finished space—after the flooring and ceiling surfaces are installed. Other things can affect headroom, as well, such as reinforcing the floor frame, and increasing rafter depth for strength or insulation.

You may also find various supports in your attic that are there to strengthen your roof but may limit your space. Collar ties are horizontal boards that join two rafters together in the upper third of the rafter span. They prevent rafter uplift in high winds. Often collar ties can be moved up a few inches but cannot be removed. Rafter ties join rafters in the lower third of their span to prevent spreading. In most attics, the ceiling or floor joists serve as rafter ties. Purlins are horizontal boards that run at right angles to the rafters and are supported by struts. These systems shorten the rafter span, allowing the use of smaller lumber for the rafters. You may be allowed to substitute kneewalls for purlins and struts. If you'll need to have any support system altered or moved, consult an architect or engineer.

The rafters themselves also need careful examination. Inspect them for signs of stress or damage, such as cracks, sagging, and insect infestation. Look for dark areas indicating roof leaks. If you find leaks or you know your roofing is past its useful life, have it repaired or replaced before you start the finishing process. And even if the rafters appear healthy, they may be too small to support the added weight of finish materials. Small rafters can also be a problem if they don't provide enough room for adequate insulation.

At this point, it's a good idea to have a professional check the structural parts of your attic, including the

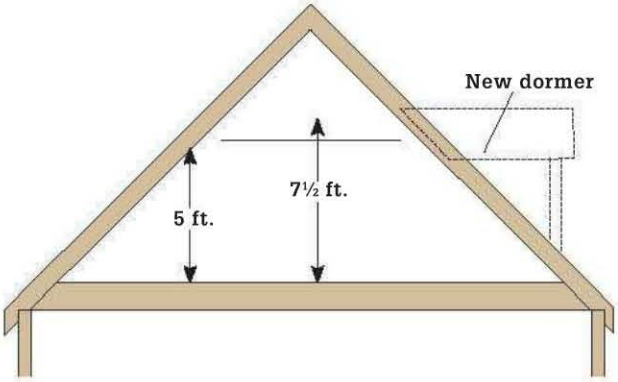

Habitable rooms must be at least 70 sq. ft. total and measure at least 7 ft. in any one direction. To meet headroom requirements, 50% of the usable floor space must have a ceiling height of 7½ ft. You can add to your floor space and headroom by adding protruding windows called dormers. In addition to space, dormers add light and ventilation to your attic.

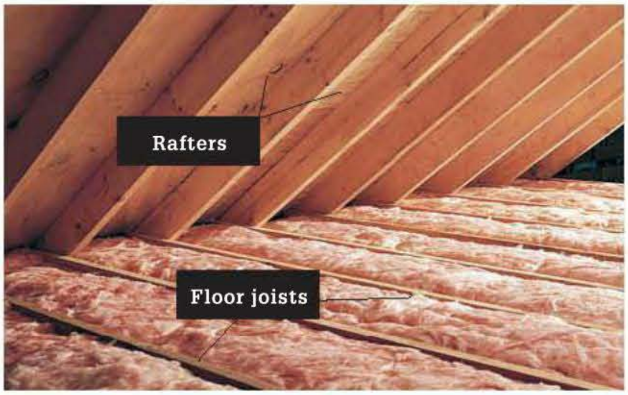

Rafter framing creates open space in an attic because the rafters carry most of the roof's weight.

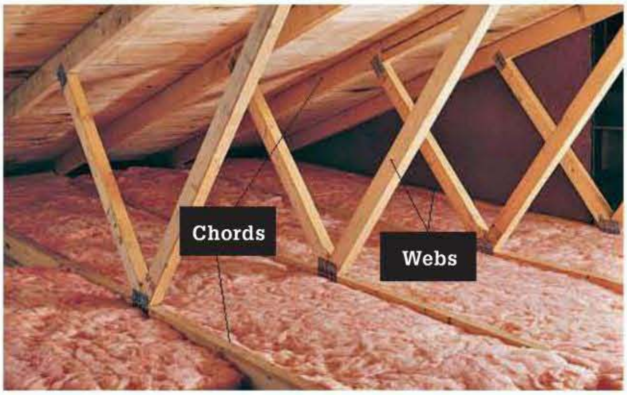

Trusses are made of interconnected cords and webs, which close off most of the attic space.

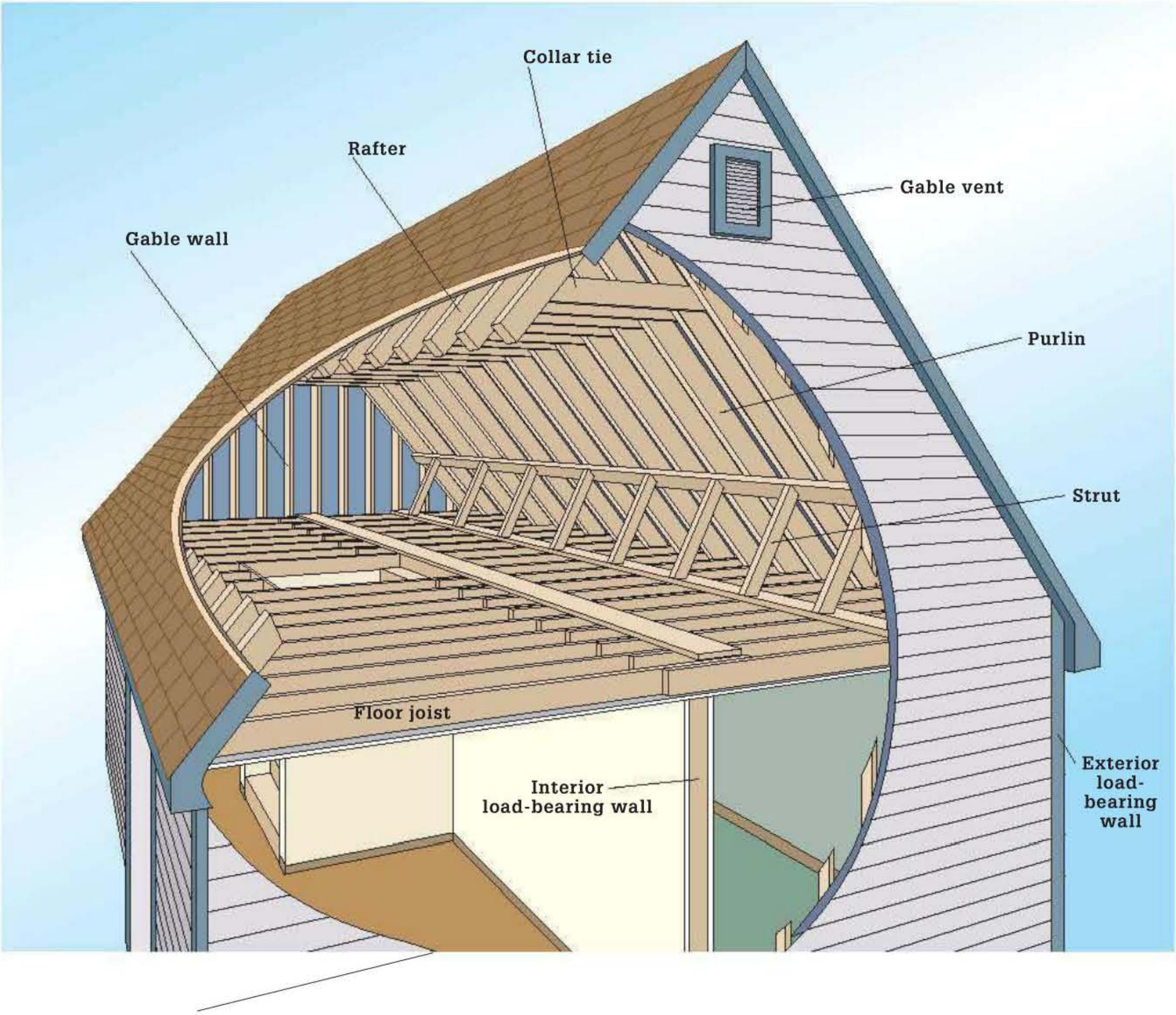

rafters and everything from the floor down. In some cases, finishing an attic is like adding a story to your home, which means that the structure must have adequate support for the new space. Attic floors are often built as ceiling frames for the level below and are not intended to support living space. Floors can be strengthened with additional joists, known as sister joists or with new joists installed between the existing ones.

Support for the attic floor is provided by the load-bearing walls below and, ultimately, by the foundation. If these elements can't support the finished attic, they'll need to be reinforced. This may be as simple as strengthening the walls with plywood panels or as complicated as adding support posts and beams or reinforcing the foundation.

In addition to these structural matters, there are a few general code requirements you should keep in mind as you inspect your attic. If you plan to add a bedroom, it will need at least one exit to the outside. This can be a door leading to an outside stairwell or an egress window. Most codes also have minimum requirements for ventilation and natural light, which means you may have to add windows or skylights.

One of the largest expenses of finishing an attic is in providing access: You'll need a permanent stairway at least 36" wide, with room for a 36" landing at the top and bottom. This is an important planning issue because adding a stairway affects the layout and traffic patterns of the attic as well as the floor below.

Finally, take an inventory of existing mechanicals in your attic. While plumbing and wiring runs can be moved relatively easily, other features, such as chimneys, must be incorporated into your plans. This is a good time to have your chimney inspected by a fire official and to obtain the building code specifications for framing around chimneys.

Evaluating Your Basement

The two things that put an end to most basement finishing plans are inadequate headroom and moisture. Begin your evaluation by measuring from the basement floor to the bottom of the floor joists above. Most building codes require habitable rooms to have a finished ceiling height of 7½ ft., measured from the finished floor to the lowest part of the finished ceiling. However, obstructions, such as beams, soffits, and pipes, (spaced at least 4 ft. on center) usually can hang down 6" below that height. Hallways and bathrooms typically need at least 7-ft. ceilings.

While it's impractical to add headroom in a basement, there are some ways of working around the requirements. Ducts and pipes often can be moved, and beams and other obstructions can be incorporated into walls or hidden in closets or other uninhabitable spaces. Also, some codes permit lower ceiling heights in rooms with specific purposes, such as recreation rooms. If headroom is a problem, talk to the local building department before you dash your dreams.

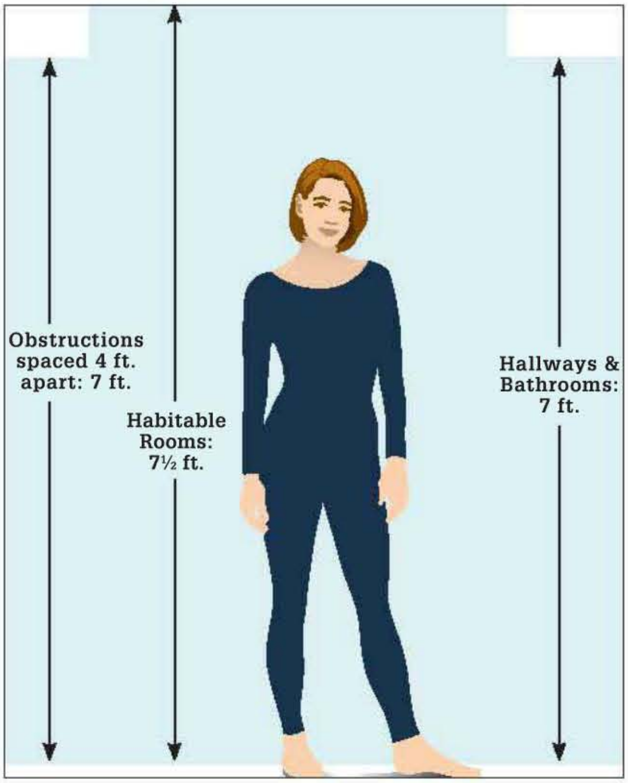
Obstructions spaced 4 ft. apart: 7 ft.
Habitable Rooms: 7½ ft.
Hallways & Bathrooms: 7 ft.

Basement headroom is often limited by beams, ducts, pipes, and other elements. Typical minimums for ceiling height are shown here: 7½ ft. for habitable rooms; 7 ft. for bathrooms and hallways; 7 ft. for obstructions spaced no less than 4 ft. apart.

A well-built basement is structurally sound and provides plenty of support for finished space, but before you cover up the walls, floor, and ceiling, check for potential problems. Inspect the masonry carefully. Large cracks may indicate shifting of the soil around the foundation; severely bowed or out-of-plumb walls may be structurally unsound. Small cracks usually cause moisture problems rather than structural woes, but they should be sealed to prevent further cracking. Contact an engineer or foundation contractor for help with foundation problems. If you have an older home, you may find sagging floor joists overhead or rotted wood posts or beams; any defective wood framing will have to be reinforced or replaced.

Your basement's mechanicals are another important consideration. The locations of water heaters, pipes, wiring, circuit boxes, furnaces, and ductwork can have a significant impact on the cost and difficulty of your project. Can you plan around components, or will they have to be moved? Is there enough headroom to install a suspended ceiling so mechanicals can remain accessible? Or, will you have to reroute pipes and ducts to increase headroom? Electricians and HVAC contractors can assess your systems and suggest modifications.

Aside from being dark and scary places, unfinished basements often harbor toxic elements. One of the most common is radon, a naturally occurring radioactive gas that is odorless and colorless. It's believed that prolonged exposure to high levels of radon can cause lung cancer. The Environmental Protection Agency has free publications to help you test for radon and take steps to reduce the levels in your house. For starters, you can perform a "short-term" test using a kit from a hardware store or home center. Look for the phrase "Meets EPA Requirements" to ensure the test kit is accurate. Keep in mind that short-term tests are not as conclusive as professional, long-term tests. If your test reveals high levels of radon, contact a radon specialist.

Another basement hazard is insulation containing asbestos, which was commonly used in older homes for insulating ductwork and heating pipes. In most cases, this insulation can be left alone provided it's in good condition and is protected from damage. If you fear the insulation in your basement poses a hazard, contact an asbestos abatement contractor to have it evaluated or safely removed.

Also check the local codes for exits from finished basements—most codes require two. The stairway commonly serves as one exit, while the other can be a door to the outside, an egress window, or a code-compliant bulkhead (an exterior stairway with cellar doors). Each bedroom will also need an egress window or door for escape.

Stairways must also meet the local code specifications. If yours doesn't, you'll probably have to hire someone to rebuild it.

As a final note, if you're planning to finish the basement in a new house, ask the builder how long you should wait before starting the project. Poured concrete walls and floors need time to dry out before they can be covered. Depending on where you live, you may be advised to wait up to two years, just to be safe.

Rerouting service lines and mechanicals adds quickly to the expense of a project, so consider your options carefully.

Weakened or undersized joists and other framing members must be reinforced or replaced.

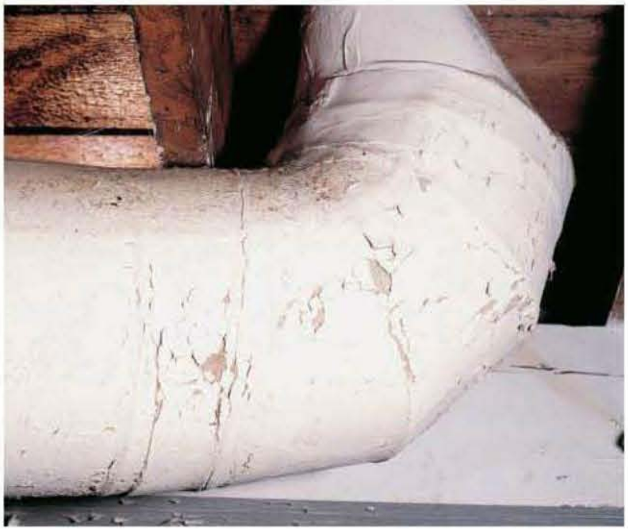

Old insulation containing asbestos poses a serious health risk if it is deteriorating or is disturbed.

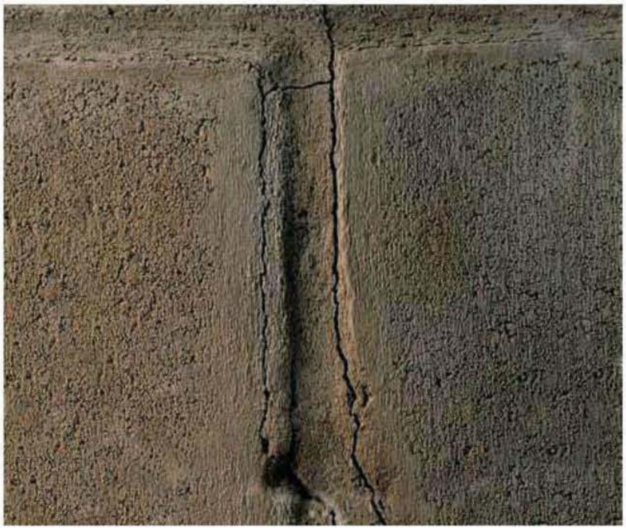

Minor cracks such as these in masonry walls and floors usually can be sealed and forgotten, while severe cracking may indicate serious structural problems.

Planning the Framing

Use walls to define your new spaces. Walls can create quiet private retreats or comfortable bathrooms or serve as barriers between formal living areas and dusty, unfinished storage spaces. To determine where your walls should go, start with a thorough investigation of the unfinished space. All obstacles, such as mechanical systems, service lines, floor drains, support columns, chimneys, and roof framing, must be considered. As you work with different layouts, think about which of these elements can be enclosed by walls, which can be hidden within a wall or concealed by a soffit or chase, and which, if any, can be moved.

One technique to help you get started is to draw full-scale "walls" onto your basement floor, using children's sidewalk chalk (on wood attic floors, use wide masking tape instead of chalk). This can help you visualize the planned spaces and give you a better sense of room sizes. Complete the proposed layout in chalk, then walk through the rooms to test the traffic patterns. As you plan your rooms, keep in mind that most building codes require habitable rooms to have at least 70 sq. ft. of floor space and measure a minimum of 7 ft. in any direction.

The next step is to draw floor plans. This doesn't require drafting skills—just a tape measure, a ruler, graph paper, and some pencils. Simply measure your basement or attic floor space, then scale down the dimensions and transfer them to the graph paper. Add all obstacles, windows, doors, and other permanent fixtures. When everything is in place, start experimenting with different layouts. If you have your home's original blueprints, trace the floor plans onto tracing paper and work on new layouts from there.

Creating a successful layout takes time and often requires some creative problemsolving. To help generate ideas for your remodel, study the before-and-after drawings on page 499. While these floor plans may not look like your basement, they include many of the common elements and obstacles involved in a finishing project. They also show how carefully placed walls can transform an unfinished space into several livable areas that still leave room for storage and mechanical elements.

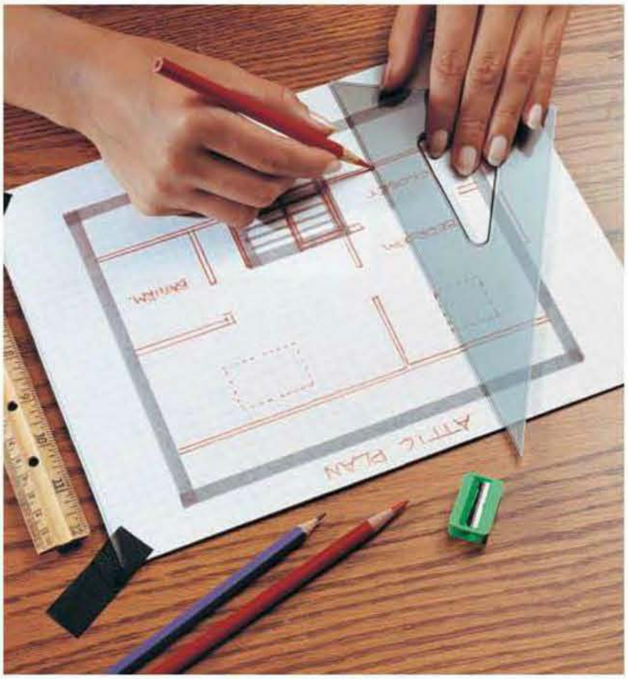

Use graph paper to sketch your wall layouts. Scale your floor plan (aerial view) drawings at ¼" equals 1 ft. For elevation drawings (wall details as viewed from the side), use a scale of ½" equals 1 ft.

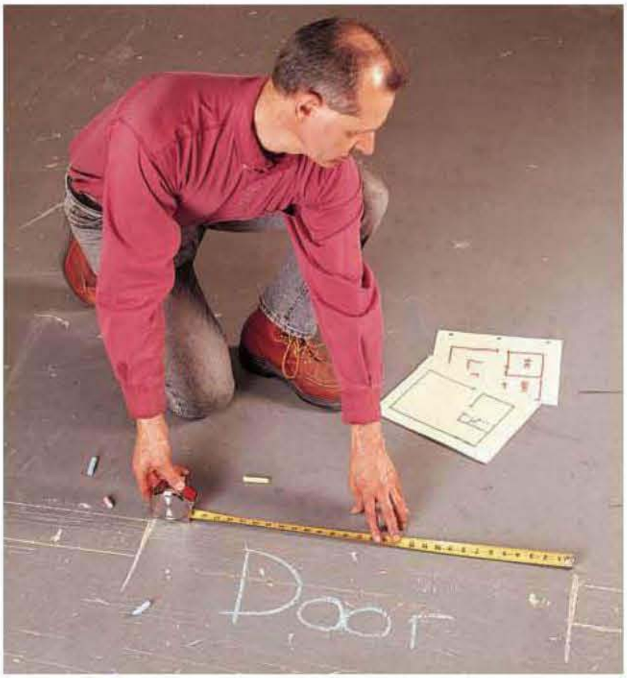

Draw layouts onto your basement floor with sidewalk chalk. Use different colors to represent elements other than walls, such as doors, windows, and ceiling soffits. Remove the chalk with a damp rag.

Basement Layouts: Before and After

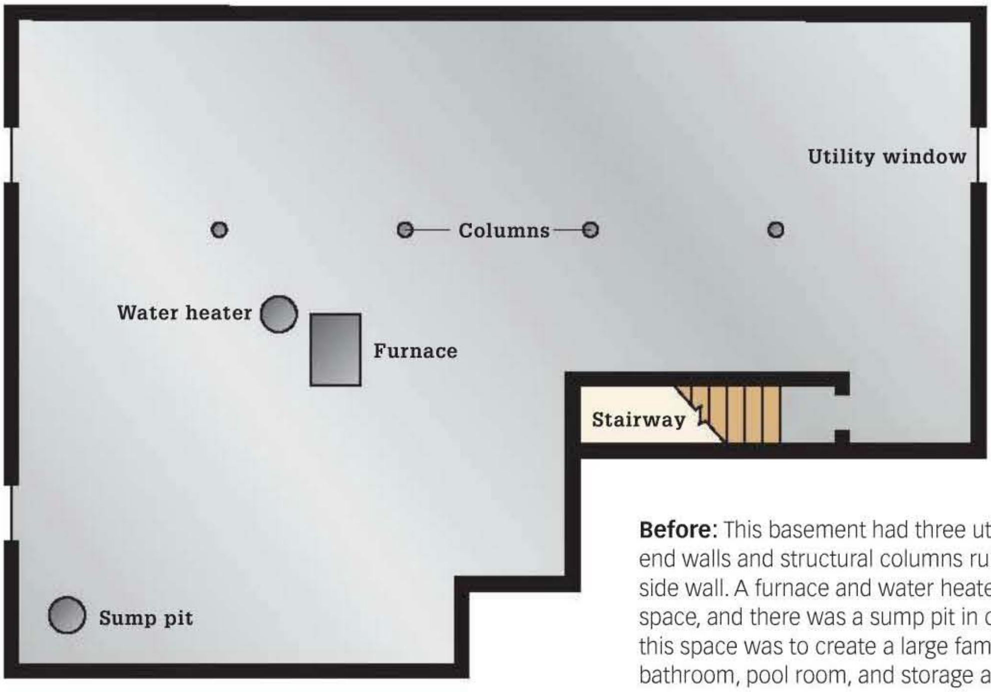
Utility window
Columns
Water heater
Furnace
Stairway
Sump pit

Before: This basement had three utility windows along the end walls and structural columns running parallel to the long side wall. A furnace and water heater broke up the central space, and there was a sump pit in one corner. The goal for this space was to create a large family room, home office, bathroom, pool room, and storage area.

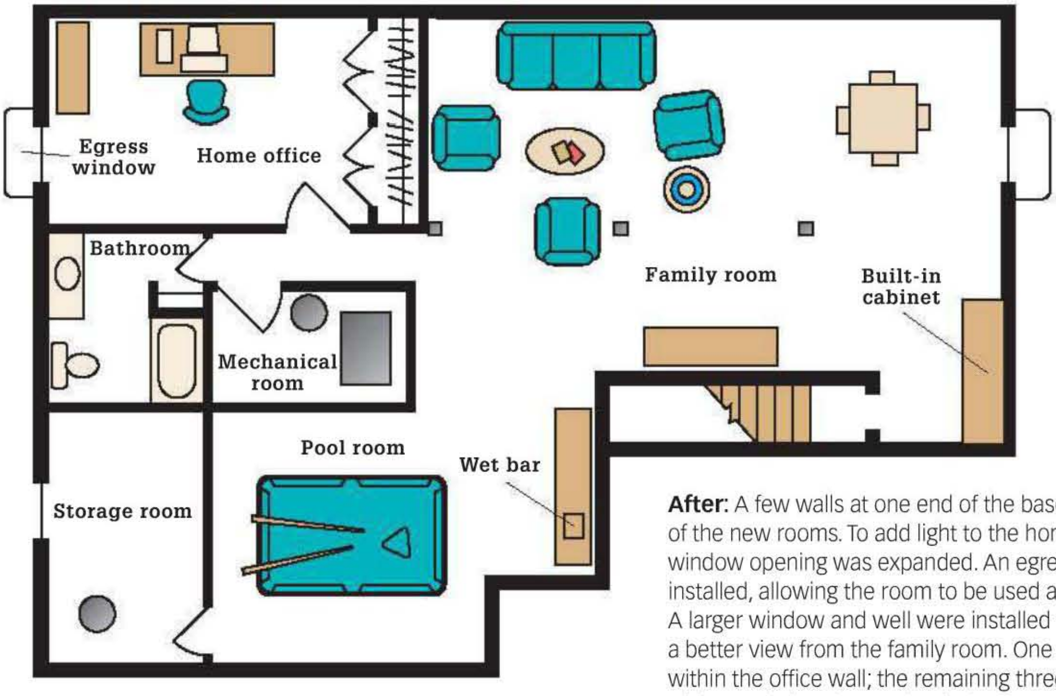
Egress window
Home office
Bathroom
Mechanical room
Family room
Built-in cabinet
Pool room
Wet bar
Storage room

After: A few walls at one end of the basement define several of the new rooms. To add light to the home office, the existing window opening was expanded. An egress window was installed, allowing the room to be used as a bedroom as well. A larger window and well were installed to provide light and a better view from the family room. One column was hidden within the office wall; the remaining three were wrapped with wood trim. The mechanical room contains the furnace and water heater, with plenty of space for servicing the units. Next to the office is a full bathroom, designed with a square layout that leaves a comfortable amount of space between the fixtures. The pool room occupies a well-defined space, where games won't disrupt activity in the family room. A wet bar can easily be accessed from both the pool room and family room. The stairway needed only a new handrail to become code-compliant. At the bottom of the stairs, a built-in cabinet provides storage and adds a decorative touch to the basement entrance.

Installing Basement Floors

Preparing a concrete floor—for carpet, laminate, vinyl, or wood flooring—has changed for the better in dramatic fashion, thanks to new subfloor products that have built-in vapor barriers and cleats that create a slight air gap between the subfloor and the concrete slab. This system allows air to circulate, protecting the finished flooring from any slab moisture. The older method of laying ¾" plywood over a frame of 2 × 4 "sleepers" was difficult, time-consuming, and raised the floor level by 2" or more—a significant drawback in basements where ceiling levels may already be too low. The new dry-floor subfloor systems are less than 1" thick and are very easy to install. There are several types of these dry-floor systems available, but the one most readily available and easiest to use is a product sold in 2 × 2-ft. tongue-and-groove squares.

Although subfloor panels can be adjusted for slight irregularities in the concrete slab, they can't overcome problems with a floor that is badly cracked and heaved. Nor is the built-in air gap beneath the system a solution to a basement that has serious water problems. A badly heaved slab will need to be leveled with a cement-based leveling compound, and serious water problems will need to be rectified before you consider creating finished living space in a basement.

Allow the subfloor panel squares to acclimate in the basement for at least 24 hours with the plastic surfaces facing down before installing them. In humid summer months, the squares—as well as the finished wood flooring product, if that's what you'll be installing—should be allowed to acclimate for a full two weeks before installation.

To estimate the number of 2-ft.-square subfloor panels you'll need, calculate the size of the room in square feet (multiply the width times the length of the room), then divide by 3.3 to determine the number of panels required.

Tools & Materials ▸

- Long straightedge
- Straightedge trowel
- Circular saw or jigsaw
- Hammer
- Carpenter's square
- Flat pry bar
- Dust mask
- Eye protection
- Portland-cement–based leveling compound
- Dry-floor subfloor squares
- Leveling shims
- Flooring spacers

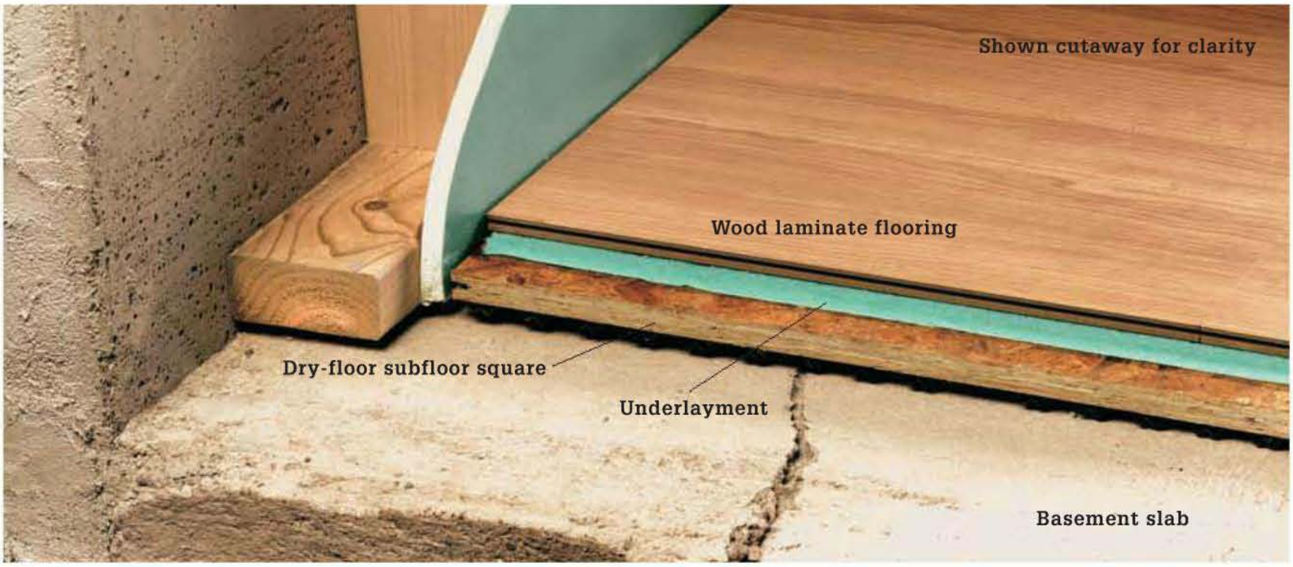

Most basement floors need some preparation before flooring can be laid. Patching compound and floor leveler can smooth rough concrete, while a subfloor system creates a new surface that is safe from moisture and feels like a framed wood floor.

Preparing Concrete Floors

Vacuum the floor thoroughly, then use a long straightedge to look for areas of the floor with serious dips or heaves. *Note: Any old floor coverings that may trap moisture should be removed before installing subfloor panels.*

Mix a batch of cement-based mortar and apply the compound to low areas with a straightedge trowel. After the patch dries, scrape the edges to feather the patch into the surrounding floor.

How to Install Subfloor Panels

1

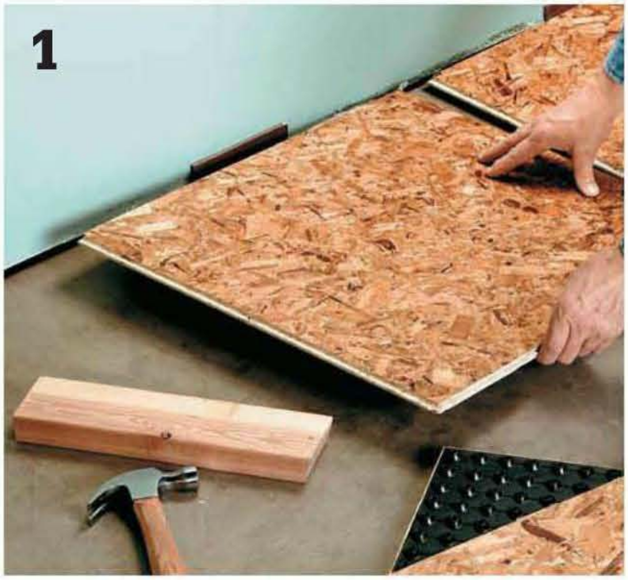

Beginning with the longest straight wall in the room, check one corner for square. If necessary, cut the first panel to match the angle of the corner. Position the panel in the corner, using ¼" spacers to create a gap between the panel and the walls, with the grooves of the panel against the spacers. Slide in the next panel, and use a piece of wood and a hammer to tap it firmly against the first panel. Repeat this placement along the entire wall. Cut the last panel to size.

2

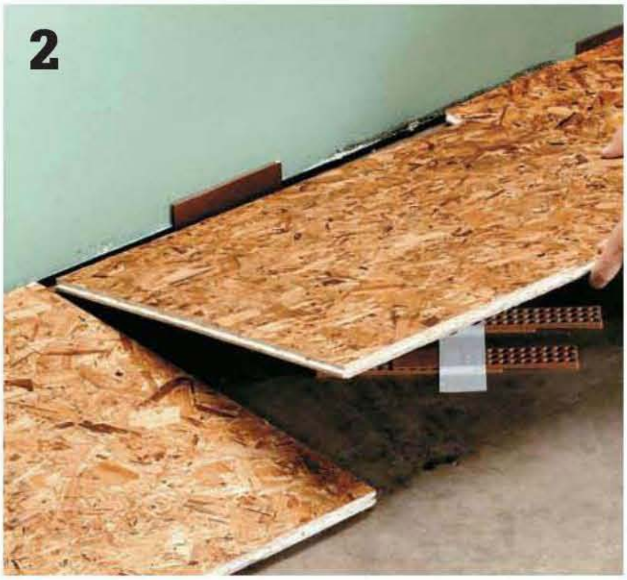

By hand, press along the entire row of panels, looking for low spots. At points where the subfloor flexes, lift the panels and place leveling shims on the floor. In some areas, you may find it necessary to stack two, three, or more shims to create the correct thickness to fill the low spot. Tape the shims to the floor to keep them in place as you reposition the panels.

(continued)

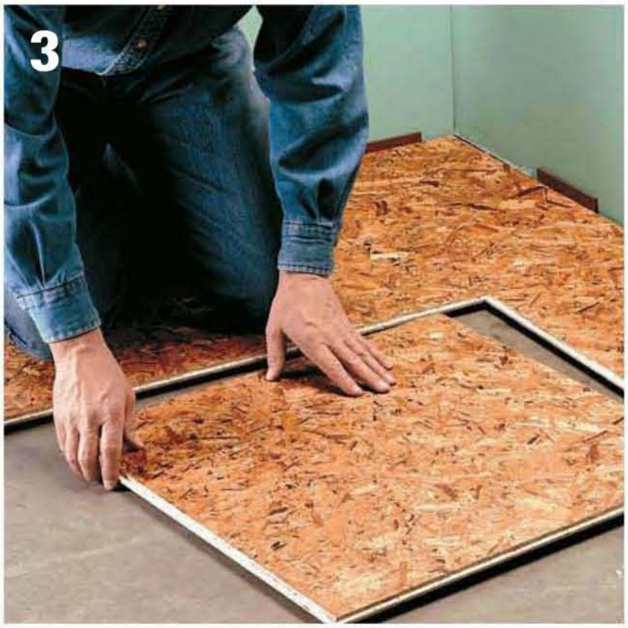

Begin the next row of panels by cutting a half-wide panel, creating offset joints between the rows. Finish the second row, test it for flatness, and shim if necessary. Move on to the third row, this time beginning with a full panel. Work across the entire room in this fashion, testing each row for flatness, shimming where necessary, and making sure joints are offset between rows.

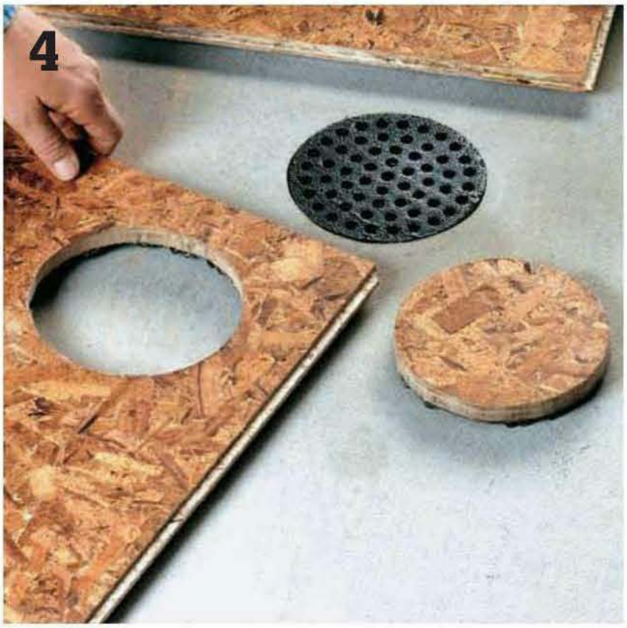

If the room has a floor drain, building code requires that you cut a round patch in the panel that falls directly over the floor drain. This patch can be removed to allow access to the floor drain should it ever be necessary.

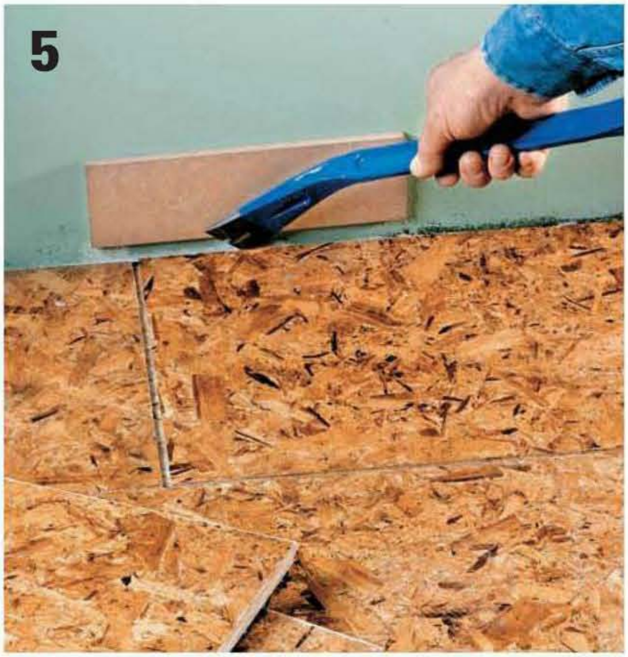

For the last row of panels, measure and cut panels to fit, maintaining a ¼" gap between the panels and the wall. Work each panel into place with a flat pry bar. When all panels are in place, you can remove all spacers and continue with the installation of the finished flooring. Laminate flooring can be installed directly over the subfloor panels.

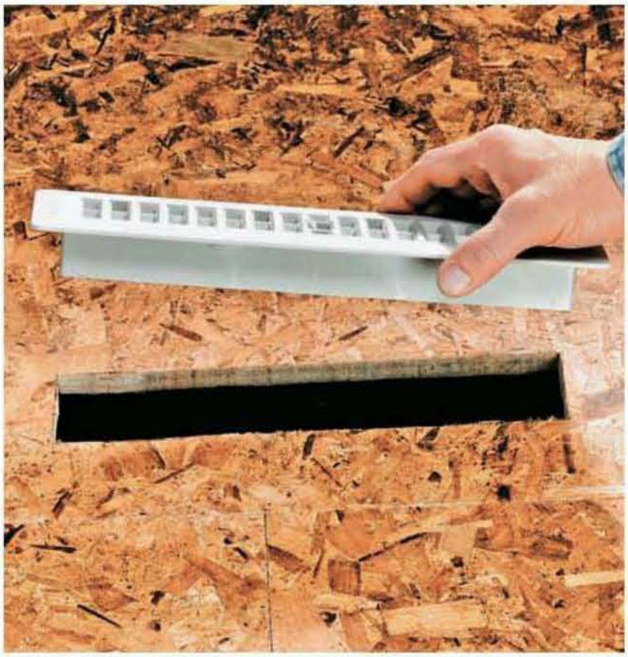

Variation: Where slab moisture is known to be a problem, you can cut vent openings spaced every 8 to 12", sized to match standard metal floor vent covers. Trace the vent opening onto the panel, and cut the opening with a jigsaw.

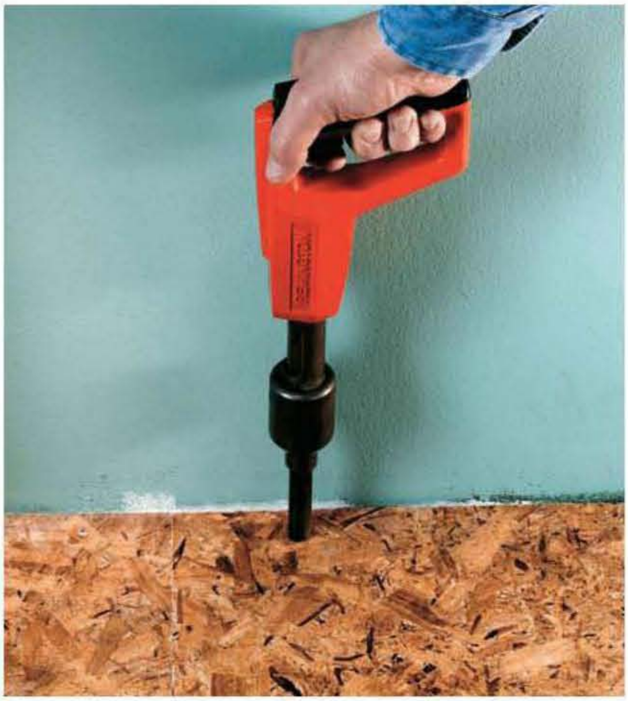

Variation: If carpeting will be stretched over the dry-floor panels, anchor the panels to the floor around the perimeter of the room and at the center of the room using concrete masonry anchors. Use a powder-actuated nailer to drive 2" nails through the panels and into the concrete slab.

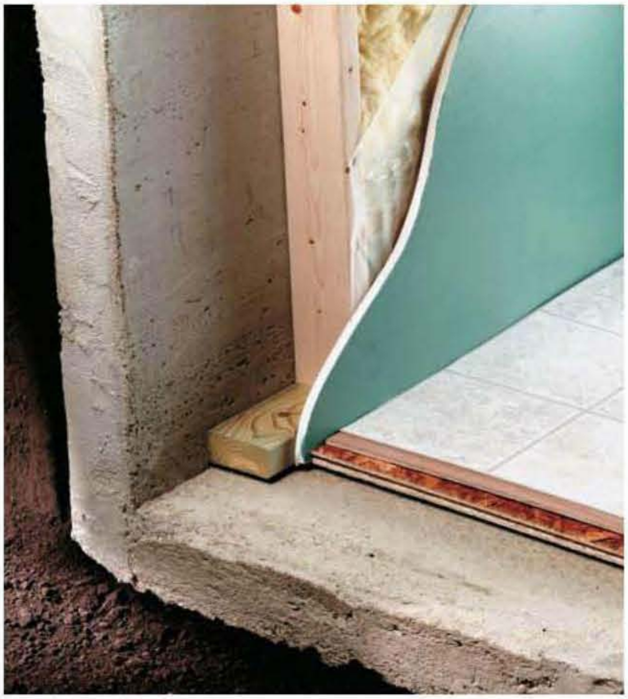

Variation: For vinyl flooring or for engineered wood flooring, install ¼" plywood underlayment over the dry-floor panels, using ½" screws or nails, which won't penetrate the moisture barrier on the underside of the panels. Don't glue the plywood to the subfloor panels, and never glue vinyl flooring directly to the subfloor panels.

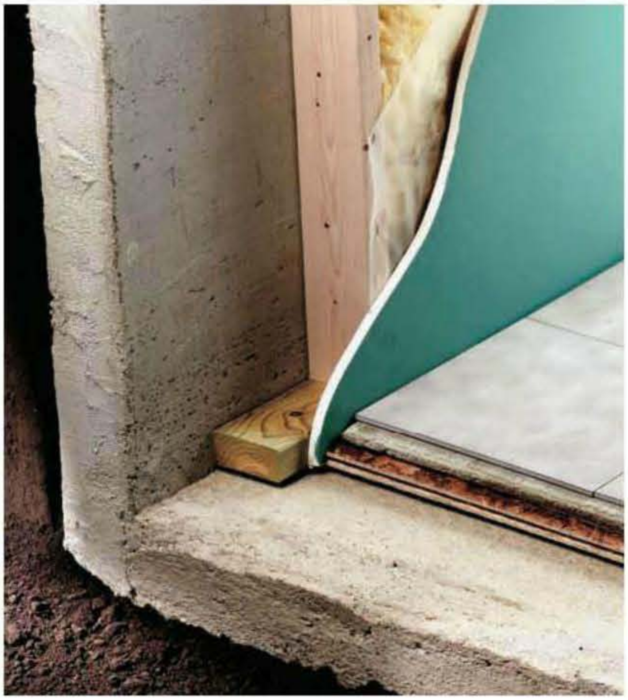

Variation: For ceramic tile, install ½" cementboard over the subfloor panels, and attach it using ¾" screws.

6

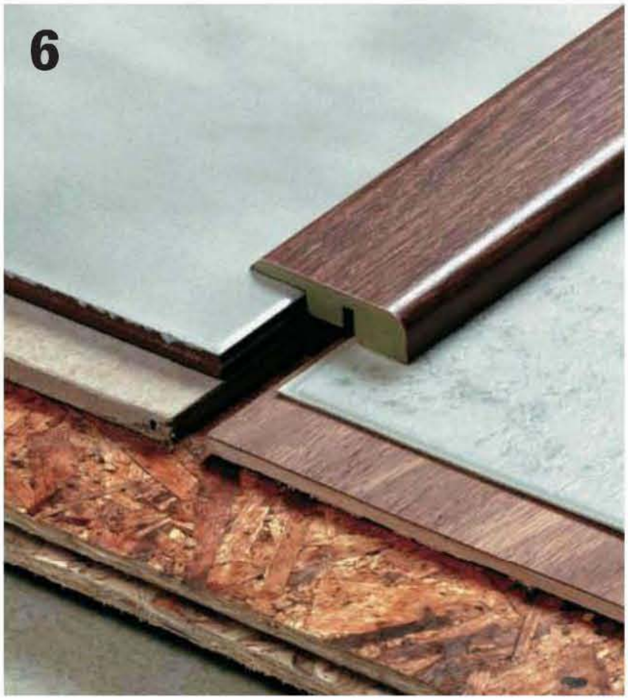

Where floor heights change, install transition strips or reducers, available in flooring departments of home improvement centers, to bridge the changes in floor heights.

How to Install a Sleeper Floor

Chip away loose or protruding concrete with a masonry chisel and hammer, then vacuum the floor. Roll out strips of 6-mil polyethylene sheeting, extending them 31" up each wall, Overlap strips by 6", then seal the seams with packing tape. Temporarily tape the edges along the walls. Be careful not to damage the sheeting.

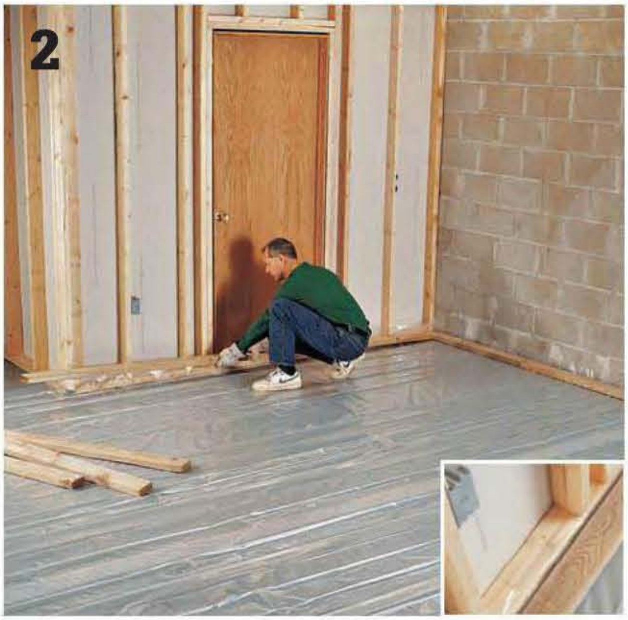

Lay out pressure-treated 2 x 4s along the perimeter of the room. Position the boards ½" in from all walls (inset). *Note: Before laying out the sleepers, determine where the partition walls will go. If a wall will fall between parallel sleepers, add an extra sleeper to support the planned wall.*

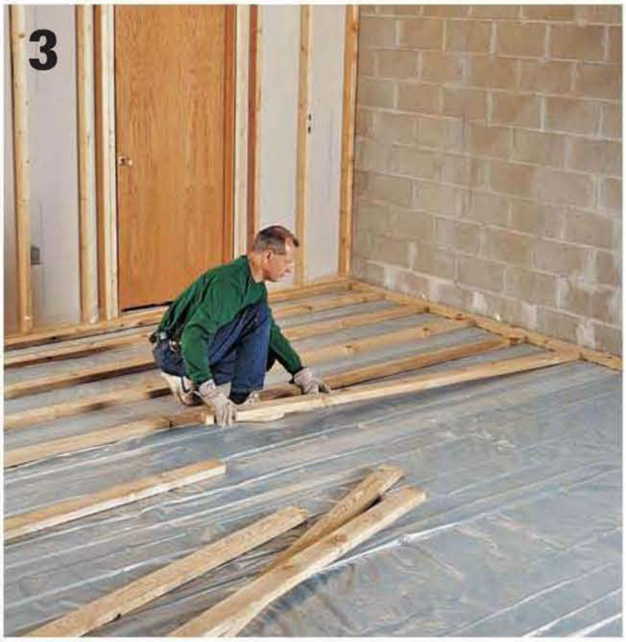

Using a circular saw, cut the sleepers to fit between the perimeter boards, leaving a ¼" gap at each end. Position the first sleeper so its center is 16" from the outside edge of the perimeter board. Lay out the remaining sleepers, using 16"-on-center spacing.

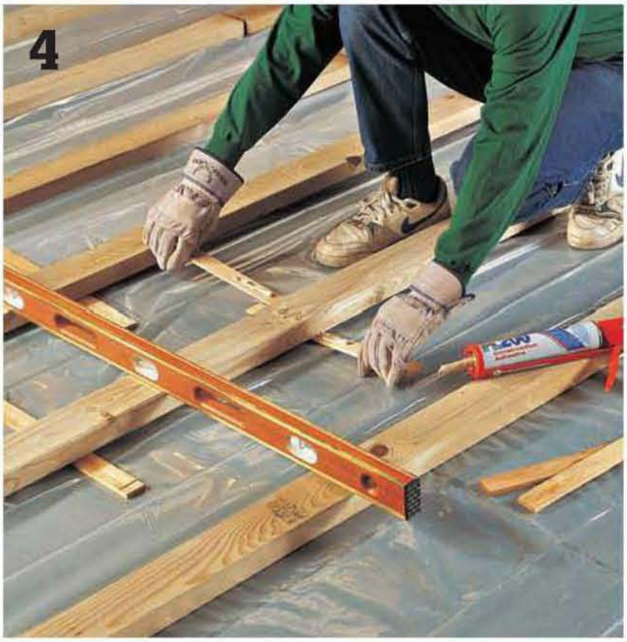

Where necessary, use tapered cedar shims to compensate for dips and variations in the floor. Place a 4-ft. level across neighboring sleepers. Apply construction adhesive to two wood shims. Slide the shims under the board from opposite sides until the board is level with adjacent sleepers.

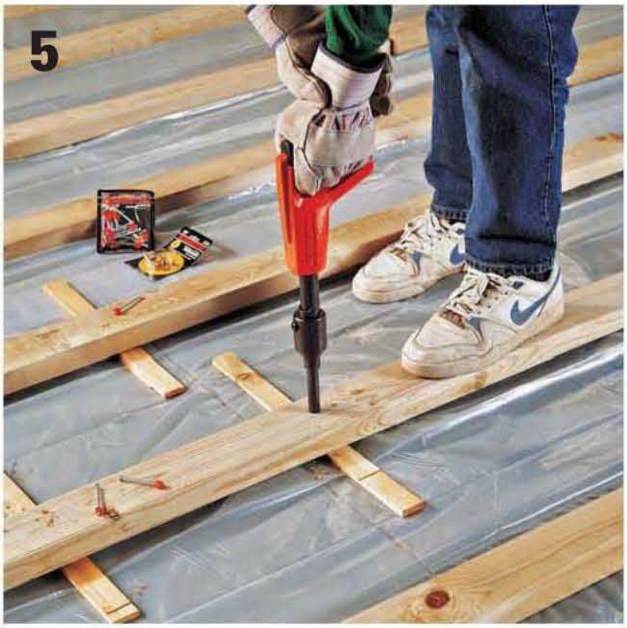

Fasten the perimeter boards and sleepers to the floor using a powder-actuated nailer or masonry screws. Drive a fastener through the center of each board at 16" intervals. Fastener heads should not protrude above the board's surface. Place a fastener at each shim location, making sure the fastener penetrates both shims.

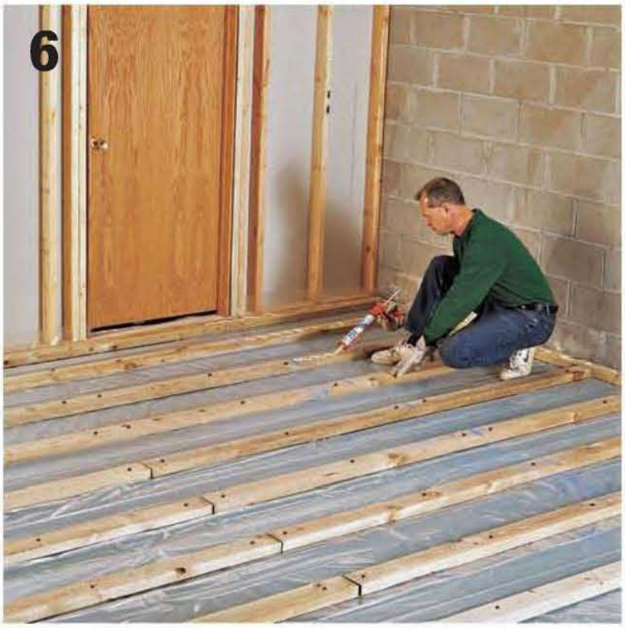

Establish a control line for the first row of plywood sheets by measuring 49" from the wall and marking the outside sleeper at each end of the room. Snap a chalk line across the sleepers at the marks. Run a ¼"-wide bead of adhesive along the first six sleepers, stopping just short of the control line.

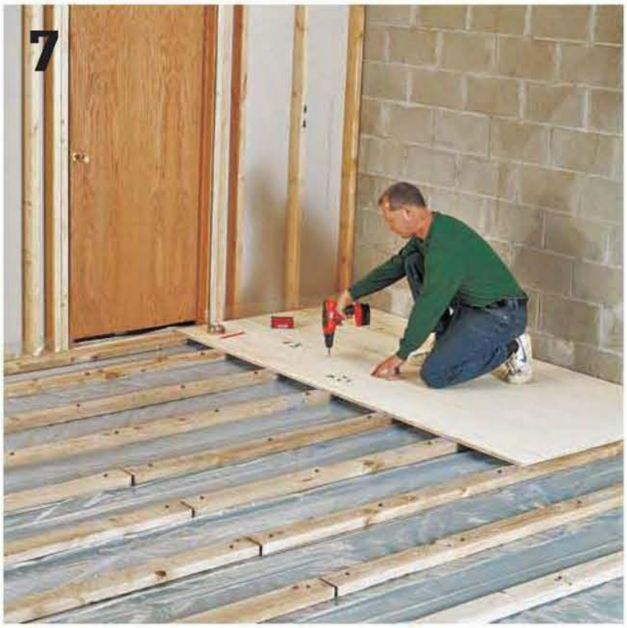

Position the first sheet of plywood so the end is ½" away from the wall and the grooved edge is flush with the control line. Fasten the sheet to the sleepers using 2" wallboard screws. Drive a screw every 6" along the edges and every 8" in the field. Don't drive screws along the grooved edge until the next row of sheeting is in place.

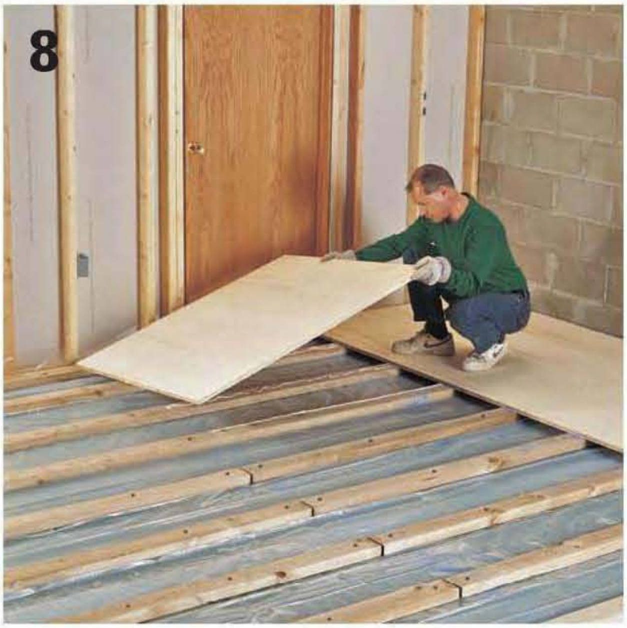

Install the remaining sheets in the first row, maintaining an ⅛" gap between ends. Begin the second row with a half-sheet (4 ft. long) so the end joints between rows are staggered. Fit the tongue of the half sheet into the groove of the adjoining sheet. If necessary, use a sledgehammer and wood block to help close the joint. After completing the second row, begin the third row with a full sheet. Alternate this pattern until the subfloor is complete.

Installing Attic Floors

Before you build the walls that will define the rooms in your attic, you'll need a sturdy floor beneath it all. Existing floors in most unfinished attics are merely ceiling joists for the floor below and are too small to support a living space.

There are several options for strengthening your attic's floor structure. The simplest method is to install an additional, identically sized joist next to each existing joist, connecting the two with nails. This process is known as *sistering*.

Sistering doesn't work when joists are smaller than 2 × 6s, are spaced too far apart, or where there are obstructions, such as plaster keys, from the ceiling below. An alternative is to build a new floor by placing larger joists between the existing ones. By resting the joists on 2 × 4 spacers, you avoid obstructions and minimize damage to the ceiling surface below. However, be aware that the spacers will reduce your headroom by 1½", plus the added joist depth.

To determine the best option for your attic, consult an architect, engineer, or building contractor, as well as a local building inspector. Ask what size of joists you'll need and which options are allowed in your area. Joist sizing is based on the span (the distance between support points), the joist spacing (typically 16" or 24" on-center), and the type of lumber used. In most cases, an attic floor must be able to support 40 pounds per sq. ft. of live load (occupants, furniture) and 10 psf dead load (wallboard, floor covering).

The floor joist cavities offer space for concealing the plumbing, wiring, and ductwork servicing your attic, so consider these systems as you plan. You'll also need to locate partition walls to determine if any additional blocking between joists is necessary.

When the framing is done, the mechanical elements and insulation are in place, and everything has been inspected and approved, complete the floor by installing ¾" tongue-and-groove plywood. If your remodel will include kneewalls, you can omit the subflooring behind the kneewalls, but there are good reasons not to: A complete subfloor will add strength to the floor, and will provide a sturdy surface for storage.

Tools & Materials ▸

Circular saw
Rafter square
Drill
Caulk gun
2 × joist lumber
2 × 4 lumber
16d, 10d, and 8d common nails
¾" T&G plywood
Construction adhesive
2¼" wallboard screws

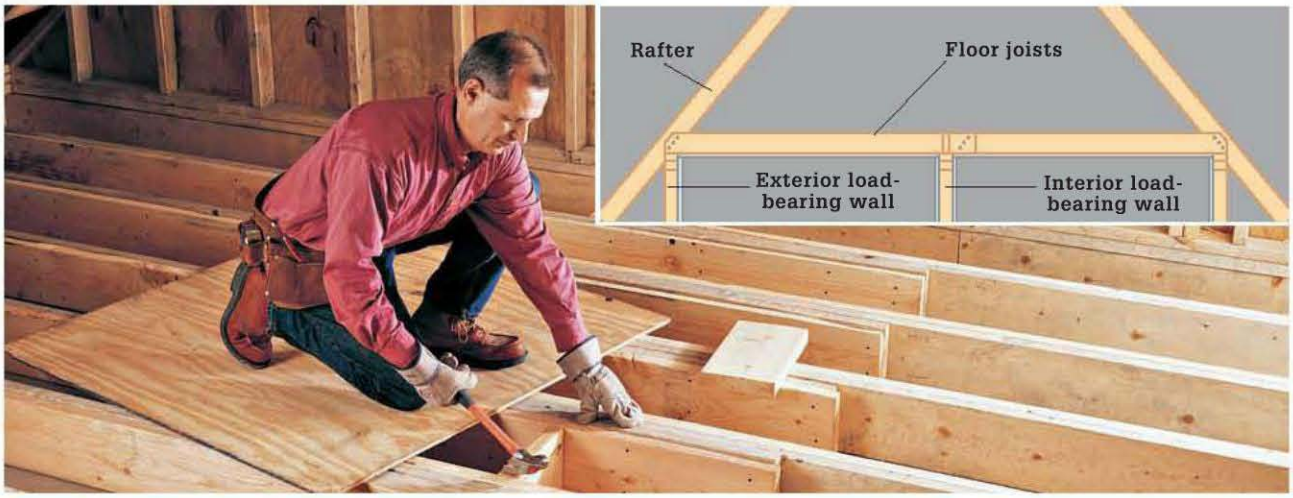

Attic joists typically rest on top of exterior walls and on an interior load-bearing wall, where they overlap from side to side and are nailed together. Always use a sheet of plywood as a platform while working over open joists.

How to Add Sister Joists

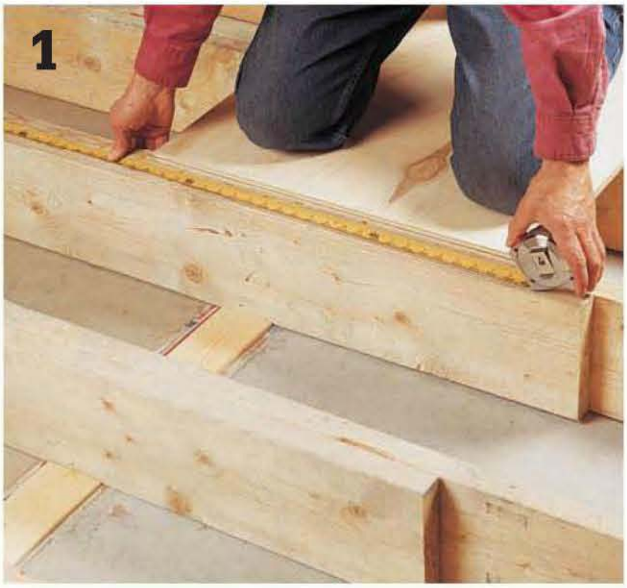

Remove all insulation from the joist cavities and carefully remove any blocking or bridging between the joists. Determine the lengths for the sister joists by measuring the existing joists. Also measure at the outside end of each joist to determine how much of the top corner was cut away to fit the joist beneath the roof sheathing. *Note: Joists that rest on a bearing wall should overlap each other by at least 3".*

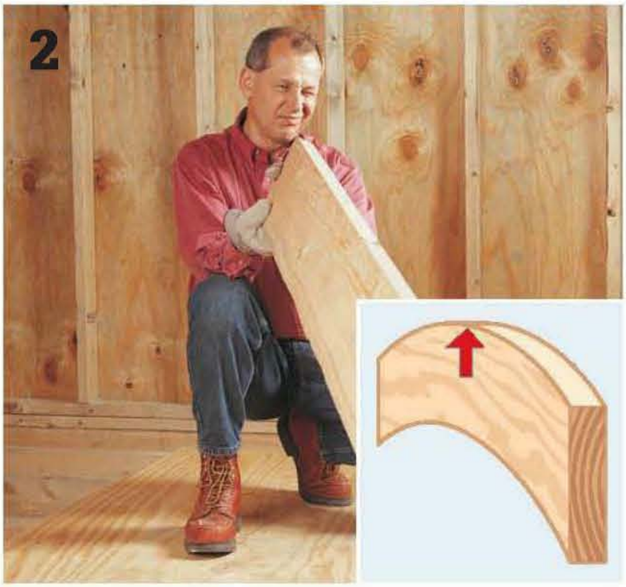

Before cutting, sight down both narrow edges of each board to check for crowning—upward arching along the length of the board. Draw an arrow that points in the direction of the arch. Joists must be installed "crown-up;" this arrow designates the top edge. Cut the board to length, then clip the top, outside corner to match the existing joists.

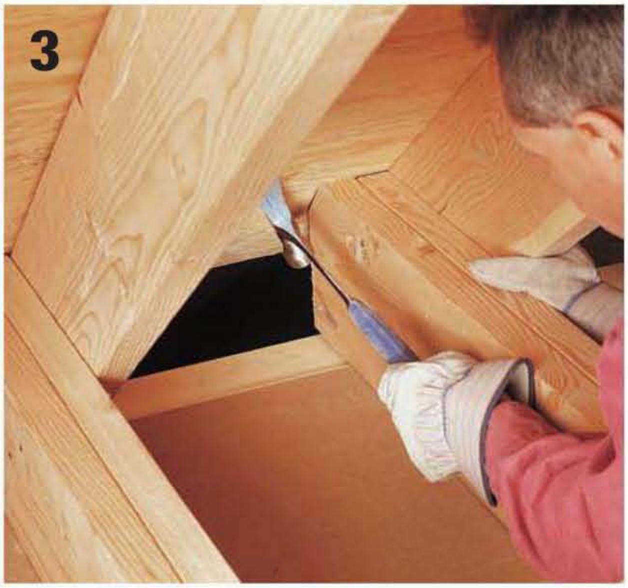

Set the sister joists in place, flush against the existing joists and with their ends aligned. Toenail each sister joist to the top plates of both supporting walls, using two 16d common nails.

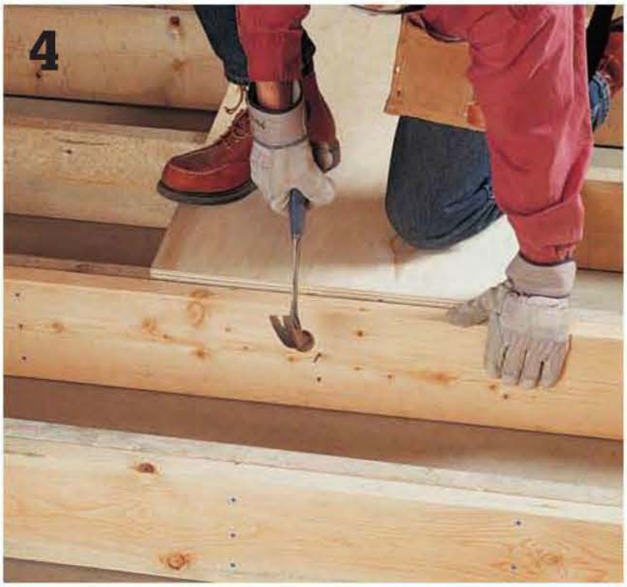

Nail the joists together using 10d common nails. Drive three nails in a row, spacing the rows 12 to 16" apart. To minimize damage to the ceiling surface below caused by the hammering (such as cracking and nail popping), you can use an air-powered nail gun (available at rental stores), or 3" lag screws instead of nails. Install new blocking between the sistered joists, as required by the local building code.

How to Build a New Attic Floor

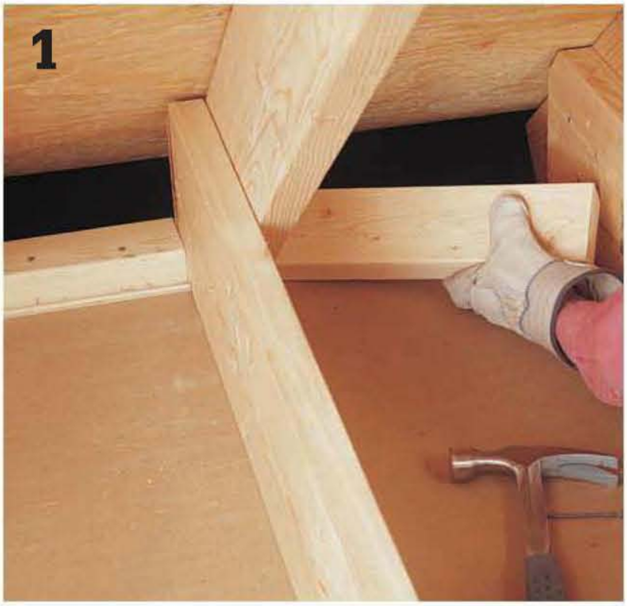

Remove any blocking or bridging from between the existing joists, being careful not to disturb the ceiling below. Cut 2 × 4 spacers to fit snugly between each pair of joists. Lay the spacers flat against the top plate of all supporting walls, and nail them in place with 16d common nails.

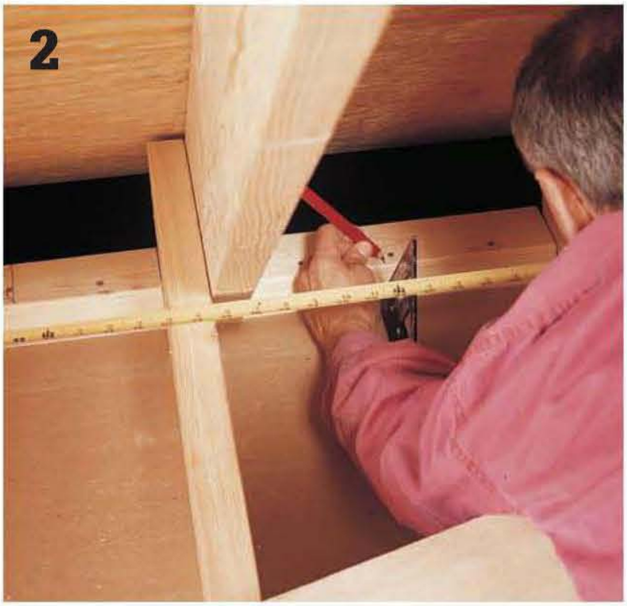

Create a layout for the new joists by measuring across the tops of the existing joists and using a rafter square to transfer the measurements down to the spacers. Following 16"-on center spacing, mark the layout along one exterior wall, then mark an identical layout onto the interior bearing wall. Note that the layout on the opposing exterior wall will be offset 1½", to account for the joist overlap at the interior wall.

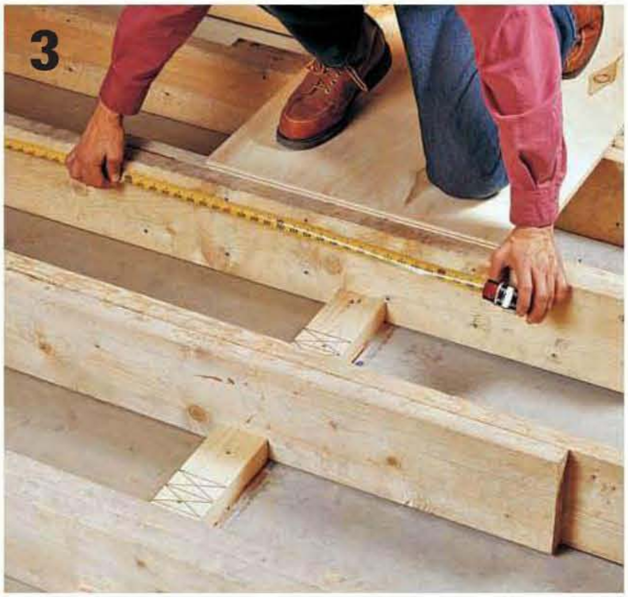

To determine joist length, measure from the outer edge of the exterior wall to the far edge of the interior bearing wall. The joists must overlap each other above the interior wall by 3". Before cutting, mark the top edge of each joist. Cut the joists to length, then clip the top outside corners so the ends can fit under the roof sheathing.

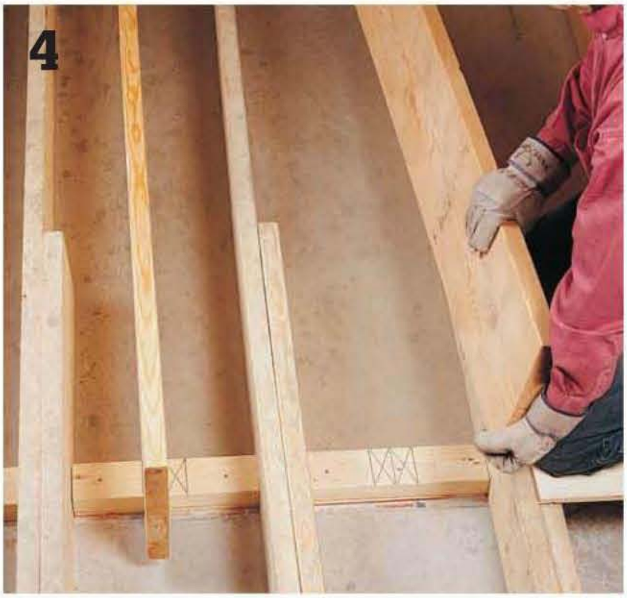

Set the joists in place on their layout marks. Toenail the outside end of each joist to the spacer on the exterior wall, using three 8d common nails.

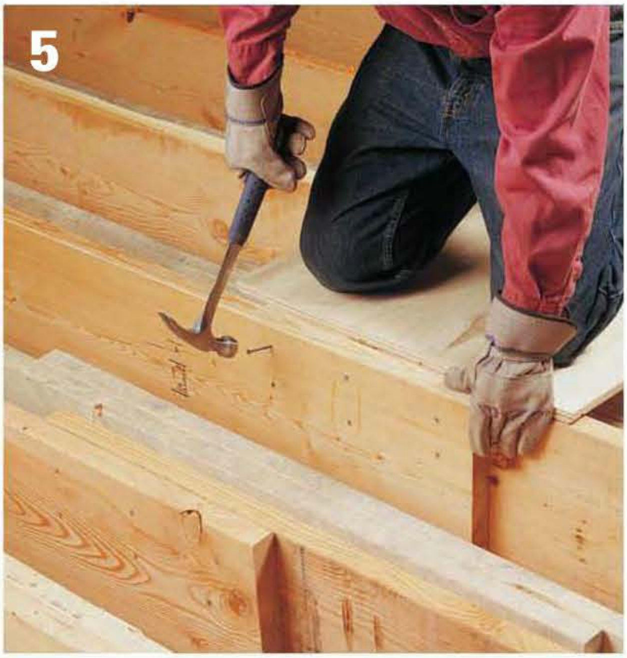

Nail the joists together where they overlap atop the interior bearing wall, using three 10d nails for each. Toenail the joists to the spacers on the interior bearing wall, using 8d nails.

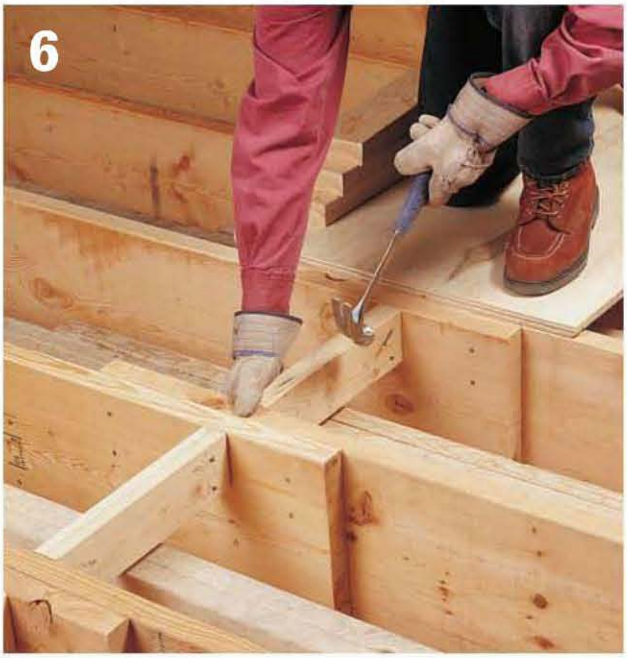

Install blocking or bridging between the joists, as required by the local building code. As a suggested minimum, the new joists should be blocked as close as possible to the outside ends and where they overlap at the interior wall.

How to Install Subflooring

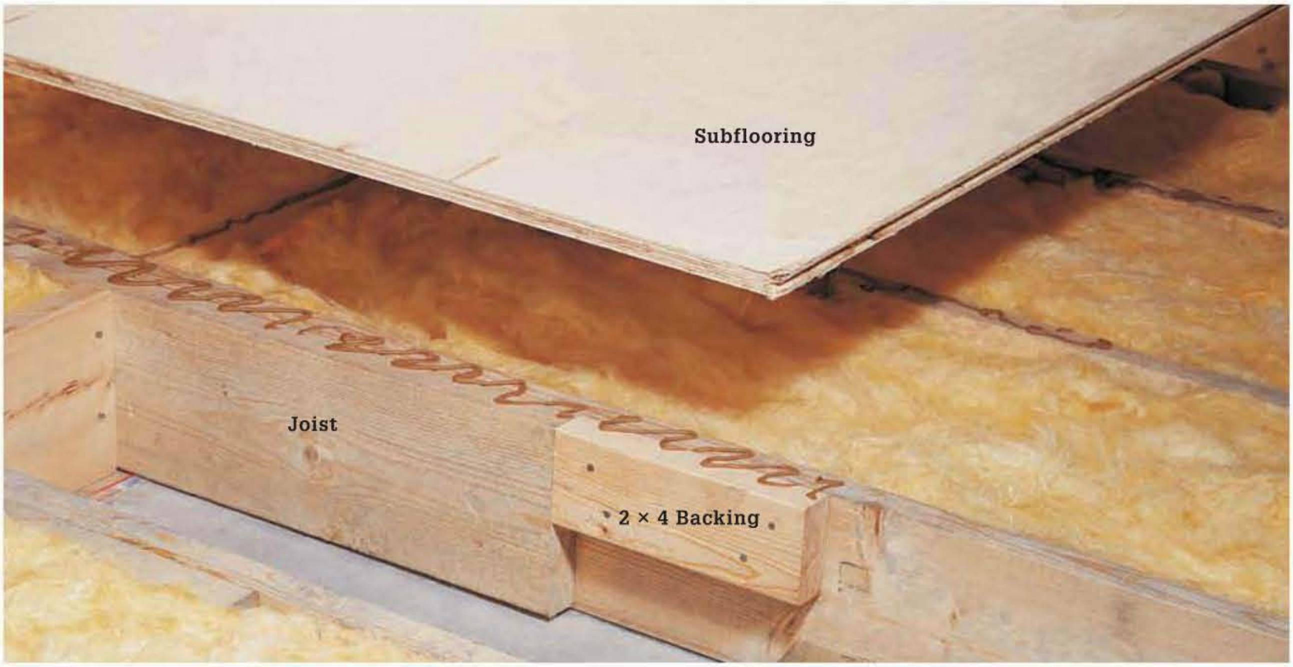

Install the subflooring only after all framing, plumbing, wiring, and ductwork is completed and has received the required building inspections. Also install any insulation and complete any caulking necessary for soundproofing. Fasten the sheets with construction adhesive and 2¼" wallboard or deck screws, making sure the sheets are perpendicular to the joists and the end joints are staggered between rows. Where joists overlap at an interior bearing wall, add backing as needed to compensate for the offset in the layout. Nail a 2 × 4 or wider board to the face of each joist to support the edges of the intervening sheets.

Installing Basement Walls

There are two common methods for covering foundation walls. Because it saves space, the more popular method is to attach 2 × 2 furring strips directly to the masonry wall. These strips provide a 1½"-deep cavity between strips for insulation and service lines, as well as a framework for attaching wallboard. The other method is to build a complete 2 × 4 stud wall just in front of the foundation wall. This method offers a full 3½" for insulation and lines, and it provides a flat, plumb wall surface, regardless of the foundation wall's condition.

To determine the best method for your project, examine the foundation walls. If they're fairly plumb and flat, you can consider furring them. If the walls are wavy or out of plumb, however, it may be easier to build stud walls. Also check with the local building department before you decide on a framing method. There may be codes regarding insulation minimums and methods of running service lines along foundation walls.

A local building official can also tell you what's recommended—or required—in your area for sealing foundation walls against moisture. Common types of moisture barriers include masonry waterproofers that are applied like paint and plastic sheeting installed between masonry walls and wood framing. The local building code will also specify whether you need a vapor barrier between the framing and the wallboard.

Before you shop for materials, decide how you'll fasten the wood framing to your foundation walls and floor. If you're covering a large wall area, it will be worth it to buy or rent a powder-actuated nailer for the job.

Tools & Materials ▸

- Caulk gun
- Trowel
- Paint roller
- Circular saw
- Drill
- Powder-actuated nailer
- Plumb bob
- Silicone caulk
- Paper-faced insulation
- Hydraulic cement
- Masonry waterproofer
- 2 × 2 and 2 × 4 lumber
- 2½" wallboard screws
- Construction adhesive
- Concrete fasteners
- Insulation

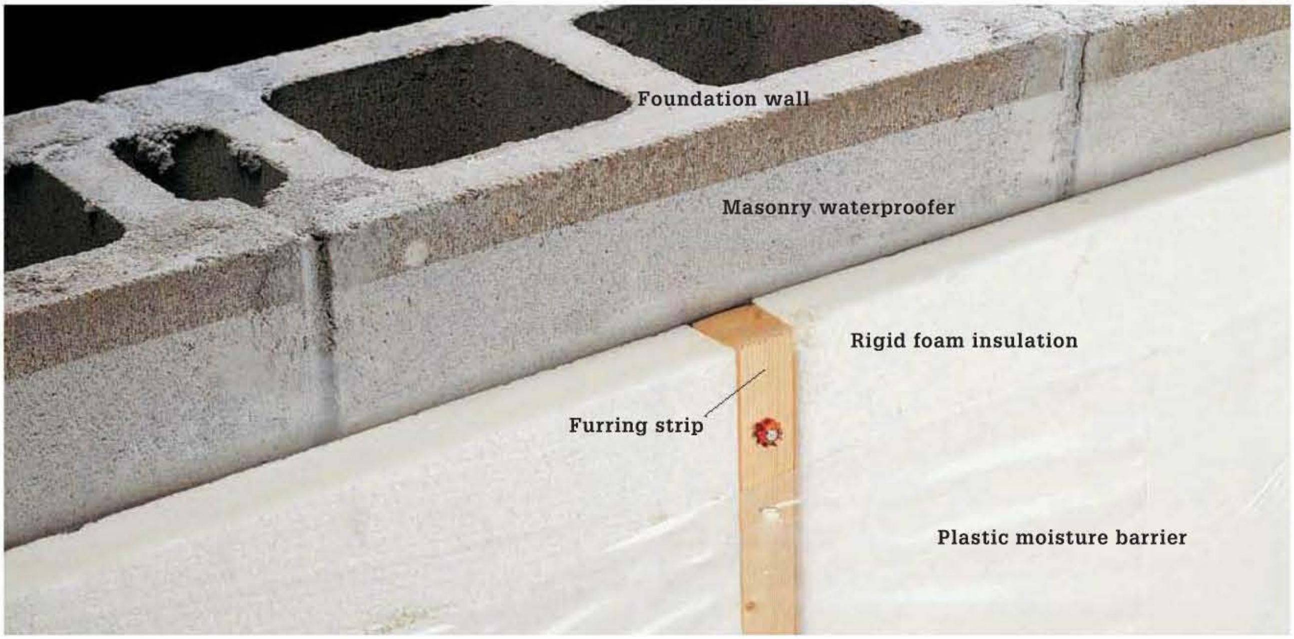

Basement walls can be made more livable by installing rigid foam insulation board between wood furring strips and then adding mold-resistant wallcovering. Although codes and practices vary dramatically, the current recommendation is to avoid the use of vapor barriers on the interior side of the wall.

How to Seal and Prepare Masonry Walls

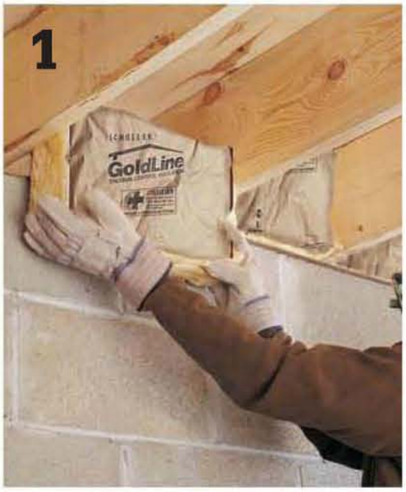

Insulate the rim-joist cavities (above the foundation walls) with solid pieces of paper-faced fiberglass insulation. Make sure the paper, which serves as a vapor barrier, faces the room.

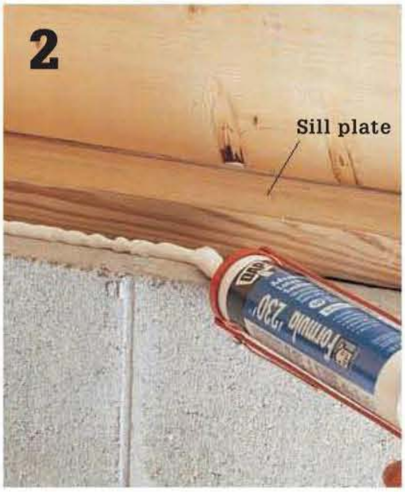

Apply silicone caulk to the joints between the sill plates and the foundation walls.

Fill small cracks with hydraulic cement or masonry caulk, and smooth the excess with a trowel. Ask the building department whether masonry waterproofer or a plastic moisture barrier is required in your area. Apply waterproofer as directed by the manufacturer, or install plastic sheeting following code specifications.

Options for Attaching Wood to Masonry

Masonry nails are the cheapest way to attach wood to concrete block walls. Drive the nails into the mortar joints for maximum holding power and to avoid cracking the blocks. Drill pilot holes through the strips if the nails cause splitting. Masonry nails are difficult to drive into poured concrete.

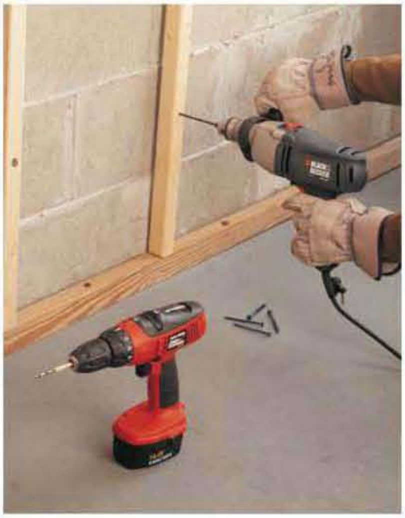

Self-tapping masonry screws hold well in block or poured concrete, but they must be driven into predrilled holes. Use a hammer drill to drill holes of the same size in both the wood and the concrete after the wood is positioned. Drive the screws into the web portion of the blocks.

(continued)

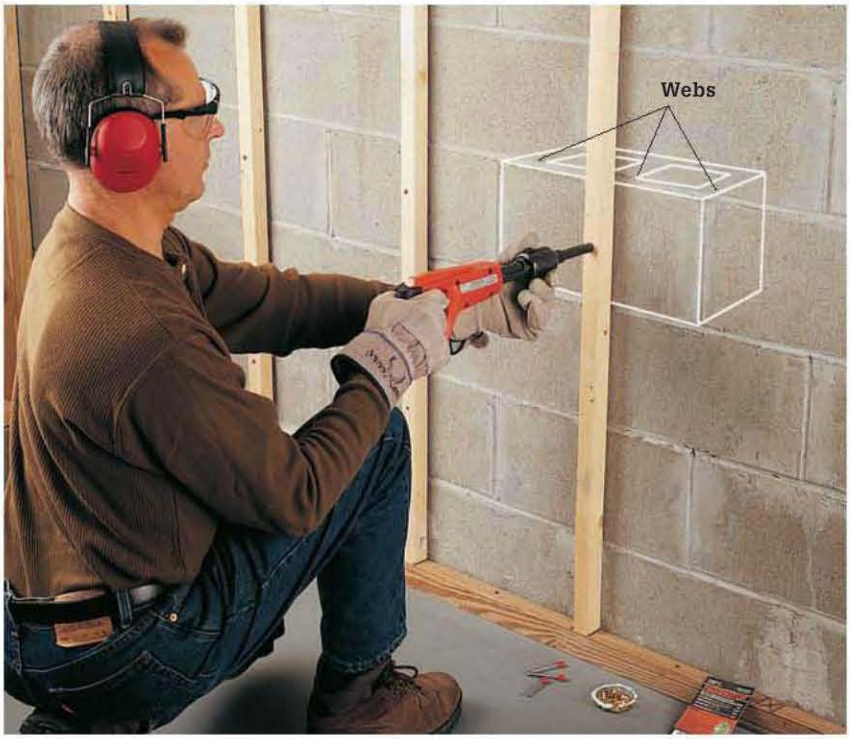

Powder-actuated nailers offer the quickest and easiest method for fastening framing to block, poured concrete, and steel. They use individual caps of gunpowder—called loads—to propel a piston that drives a hardened-steel nail (pin) through the wood and into the masonry. The loads are color-coded for the charge they produce, and the pins come in various lengths. *Note: Always drive pins into the solid web portions of concrete blocks, not into the voids. Trigger-type nailers, like the one shown here, are easiest to use, but hammer-activated types are also available. You can buy nailers at home centers and hardware stores, or rent them from rental centers. (Ask for a demonstration at the rental center.) Always wear hearing and eye protection when using these extremely loud tools.*

How to Install Furring Strips on Masonry Walls

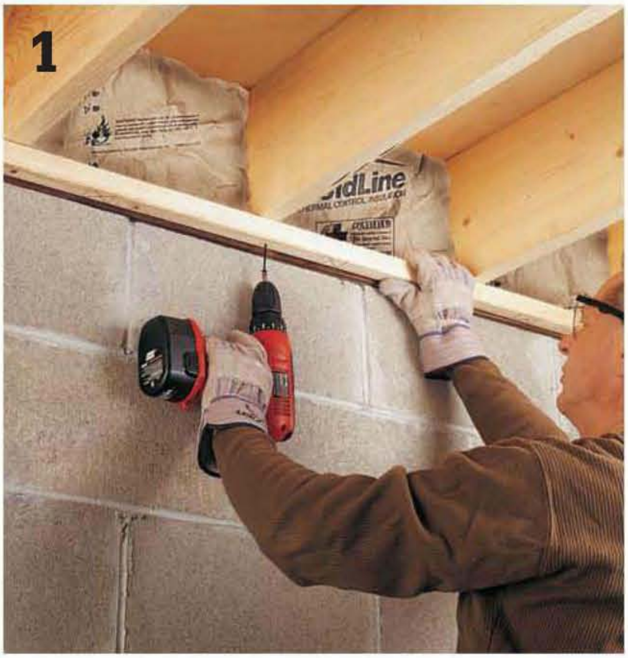

Cut a 2 × 2 top plate to span the length of the wall. Mark the furring-strip layout onto the bottom edge of the plate, using 16"-on-center spacing. Attach the plate to the bottom of the joists with 2½" wallboard screws. The back edge of the plate should line up with the front of the blocks.

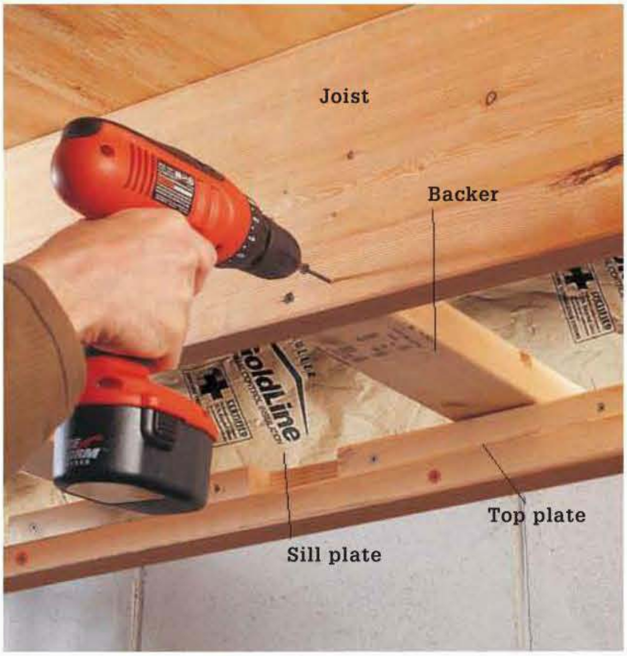

Note: If the joists run parallel to the wall, you'll need to install backers between the outer joist and the sill plate to provide support for ceiling wallboard. Make T-shaped backers from short 2 × 4s and 2 × 2s. Install each so the bottom face of the 2 × 4 is flush with the bottom edge of the joists. Attach the top plate to the foundation wall with its top edge flush with the top of the blocks.

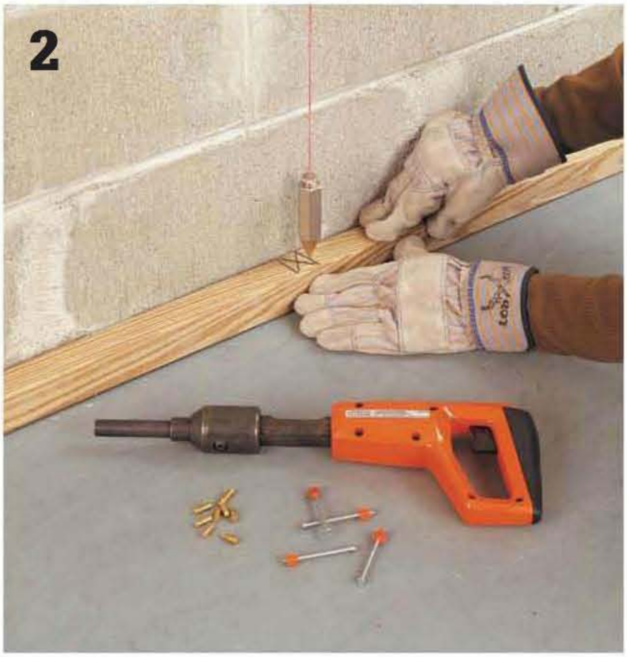

Install a bottom plate cut from pressure-treated 2 × 2 lumber so the plate spans the length of the wall. Apply construction adhesive to the back and bottom of the plate, then attach it to the floor with a nailer. Use a plumb bob to transfer the furring-strip layout marks from the top plate to the bottom plate.

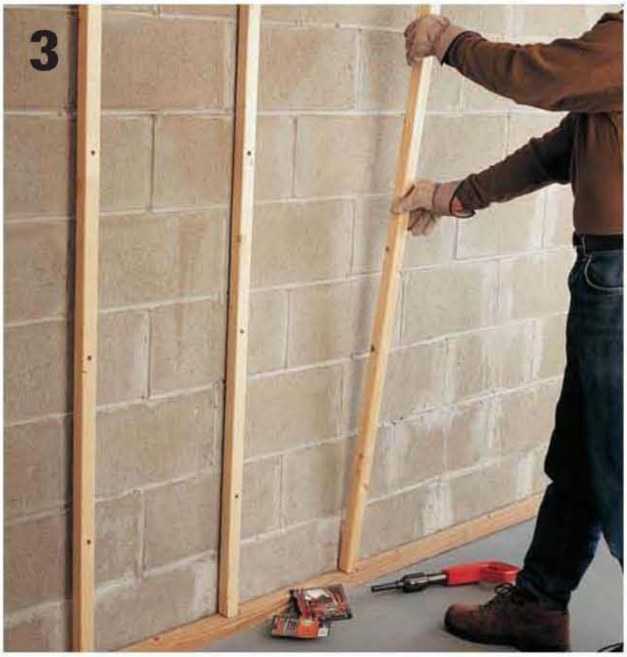

Cut 2 × 2 furring strips to fit between the top and bottom plates. Apply construction adhesive to the back of each furring strip, and position it on the layout marks on the plates. Nail along the length of each strip at 16" intervals.

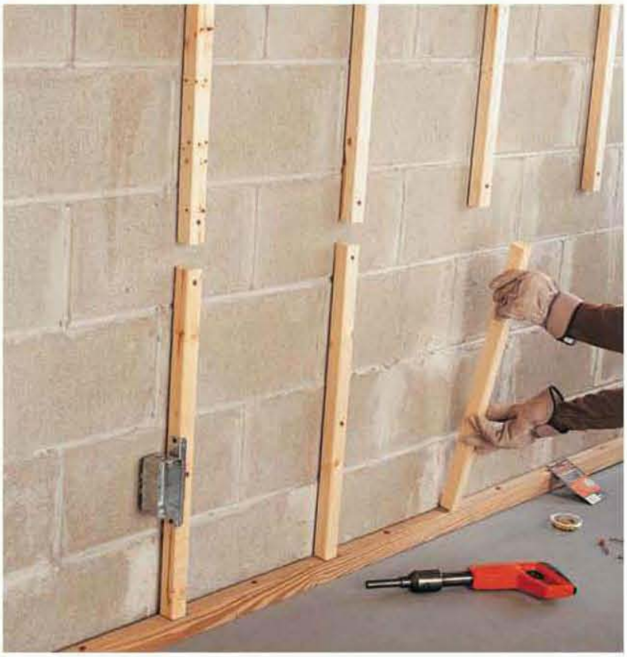

Option: Leave a channel for the installation of wires or supply pipes by installing pairs of vertically aligned furring strips with a 2" gap between each pair. *Note: Consult local codes to ensure proper installation of electrical or plumbing materials.*

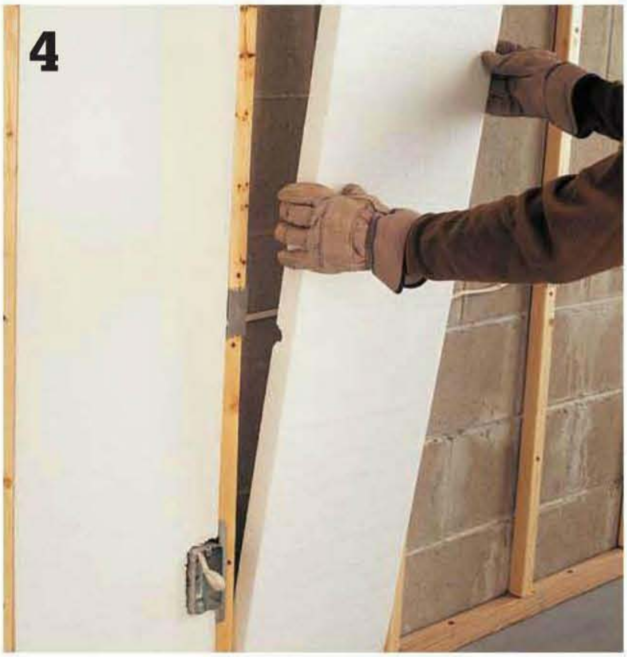

Fill the cavities between furring strips with rigid insulation board. Cut the pieces so they fit snugly within the framing. If necessary, make cutouts in the insulation to fit around mechanical elements, and cover any channels with metal protective plates before attaching the wall surface. Add a vapor barrier if required by local building code.

Installing Attic Walls

Attic kneewalls are short walls that extend from the attic floor to the rafters. They provide a vertical dimension to attic rooms, and without them, attics tend to feel cramped. Kneewalls are typically 5 ft. tall, for a couple of reasons: That's the minimum ceiling height for usable floor space according to most building codes, and it defines a comfortable room without wasting too much floor space. The unfinished space behind kneewalls doesn't have to go to waste: It's great for storage and for concealing service lines. To provide access to this space, create a framed opening in the wall during the framing process.

Kneewalls are similar to partition walls, except they have beveled top plates and angle-cut studs that follow the slope of the rafters. The added stud depth created by the angled cut requires a 2 × 6 top plate. Before starting on your kneewall project, it may help to review the techniques for building a partition wall.

Tools & Materials ▸

- Circular saw
- Level
- Chalk line
- T-bevel
- 2 × 4 and 2 × 6 lumber
- 16d
- 8d common nails

Attic kneewalls are just the right height to be backdrops for furniture, and they make a perfect foundation for built-in storage units.

How to Build a Kneewall

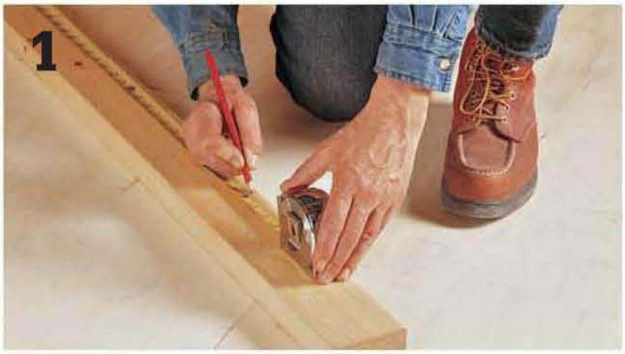

Create a storyboard using a straight 2 × 4. Cut the board a few inches longer than the planned height of the wall. Measure from one end and draw a line across the front edge of the board at the exact wall height.

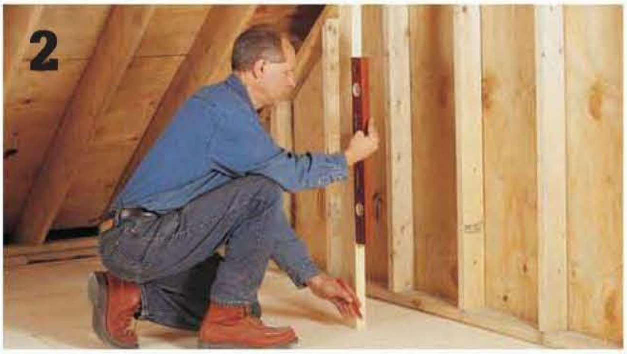

At one end of the room, set the storyboard flat against the outer rafter. Plumb the storyboard with a level while aligning the height mark with the bottom edge of the rafter. Transfer the height mark onto the rafter edge, then make a mark along the front edge of the storyboard onto the subfloor. These marks represent the top and bottom wall plates.

3

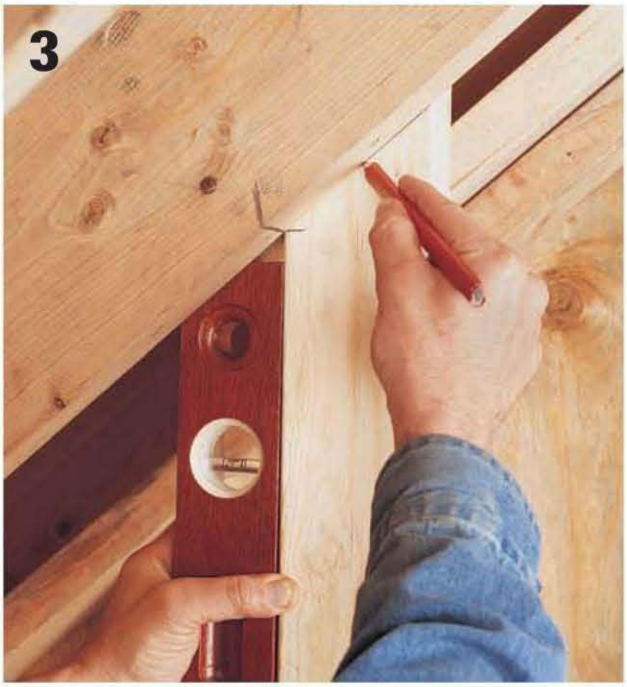

Holding the storyboard perfectly plumb, trace along the bottom edge of the rafter to transfer the rafter slope onto the face of the storyboard.

4

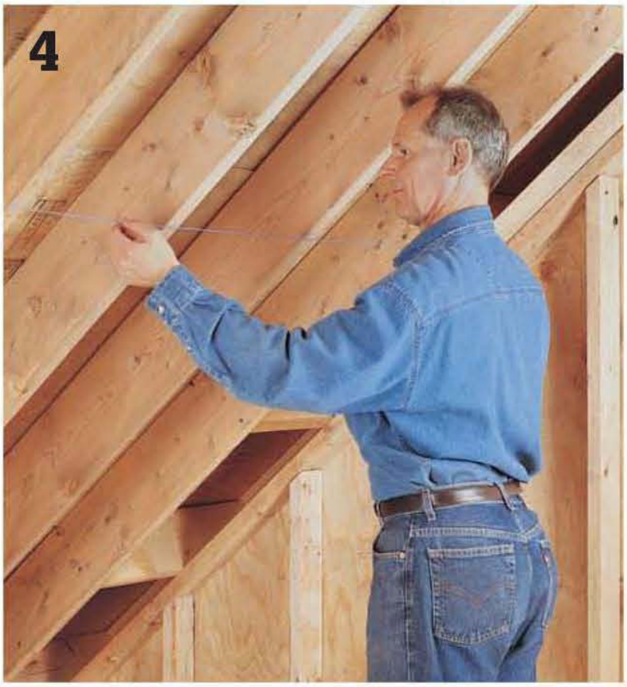

Repeat the wall-plate marking process on the other end of the room. Snap a chalk line through the marks—across the rafters and along the subfloor. If necessary, add backing for fastening the top plate to the gable wall.

5

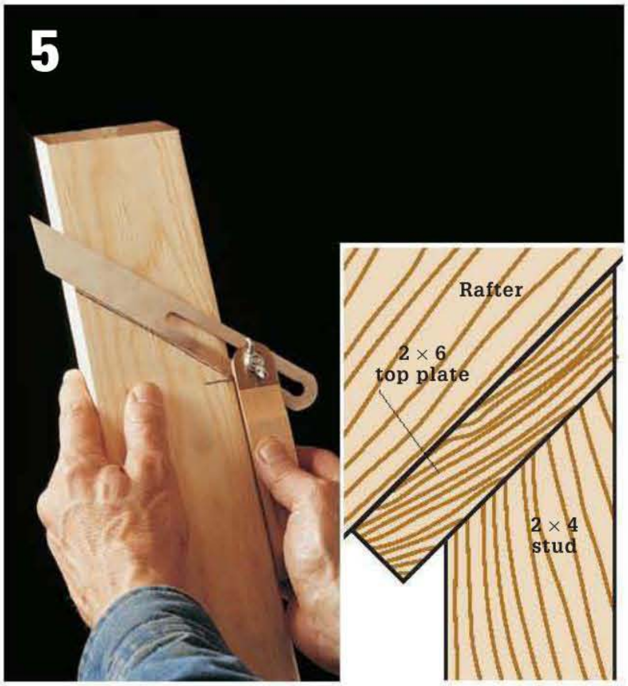

To cut a beveled edge on the top wall plate, set a T-bevel to match the rafter-slope line on the storyboard. Use the T-bevel to adjust the blade of a circular saw or table saw to the proper angle. Then, bevel-cut one edge of the 2 × 6 top plate. *Note: When the top plate is laid flat across the rafters, the front edge should be perpendicular to the floor.*

6

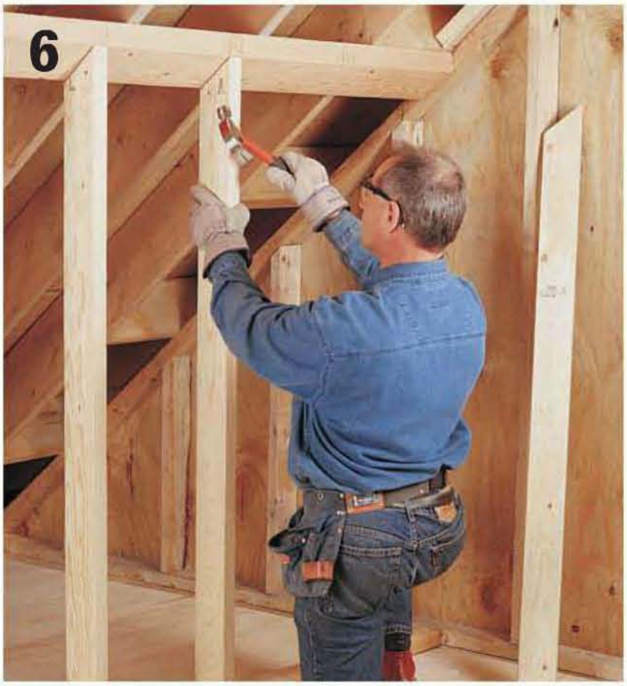

Mark the stud locations on the wall plates. Install the plates along the chalk lines, fastening them to the rafters and floor joists, respectively, using 16d nails. Measure and cut each stud to fit, angle-cutting the top end so that it meets flush with the top plate. Toenail each stud in place with three 8d nails.

Finishing Attic Ceilings

By virtue of sloping roofs, most attics naturally have "cathedral" ceilings. It's up to you whether to leave the peaks intact—and apply a finish surface all the way up to the ridge—or to frame-in a horizontal ceiling, creating a flat surface that's more like a standard ceiling. Before deciding, consider the advantages and disadvantages of each treatment.

If your attic has collar ties—horizontal braces installed between opposing rafters—your planning should start with those. Are the ties high enough to meet the code requirements for attic headroom? If not, consult an architect or engineer to see if you can move them up a few inches (do not move or remove them without professional guidance). If the ties are high enough, you can incorporate them into a new ceiling or leave them exposed and wrap them with a finish material, such as wallboard or finish-grade lumber. Do not use collar ties as part of your ceiling frame.

A peaked ceiling is primarily an aesthetic option. Its height expands the visual space of the room, and its rising angles provide a dramatic look that's unique in most homes. Because a peaked ceiling encloses the rafter bays all the way up to the ridge, this treatment may require additional roof vents to maintain proper ventilation.

By contrast, a flat ceiling typically offers a cleaner, more finished appearance closer to that of a conventional room, and flat ceilings offer some practical advantages over peaked styles. First, they provide a concealed space above the ceiling, great for running service lines. If there are vents high on the gable walls, this open space can help ventilate the roof (make sure to insulate above the ceiling). The ceiling itself can hold recessed lighting fixtures or support a ceiling fan. And if your plans call for full-height partition walls, you may want a ceiling frame to enclose the top of the wall.

When determining the height of flat-ceiling framing, be sure to account for the floor and ceiling finishes. And remember that most building codes require a finished ceiling height of at least 90".

Tools & Materials ▸

- 4-ft. level
- Chalk line
- Circular saw
- 2 × 4 and 2 × 6 lumber
- 10d common nails

Flat attic ceilings provide space for recessed light fixtures, vents, and speakers.

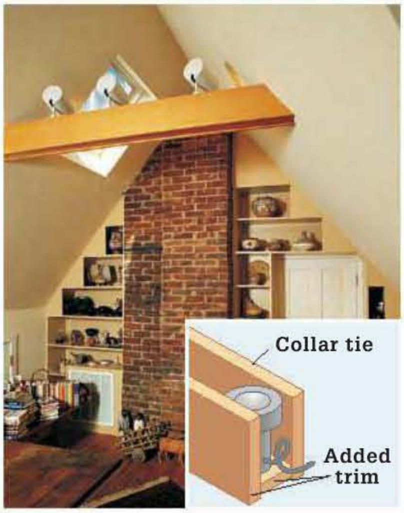

Exposed collar ties can add an interesting architectural element to a peaked ceiling. By adding to the existing ties, you can create a channel for holding small light fixtures.

How to Frame a Flat Ceiling

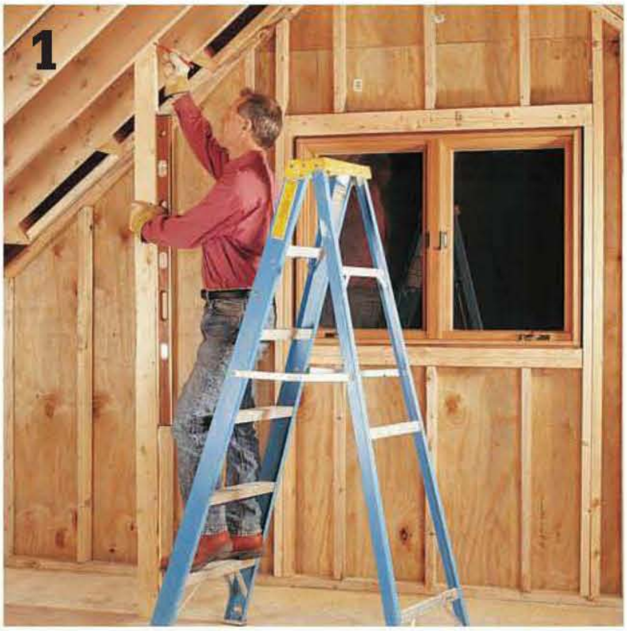

Make a storyboard for the planned height of the ceiling frame. At one end of the attic, hold the storyboard plumb and align the height mark with the bottom edge of a rafter. Transfer the mark to the rafter. Repeat at the other end of the attic, then snap a chalk line through the marks. This line represents the bottom edge of the ceiling frame.

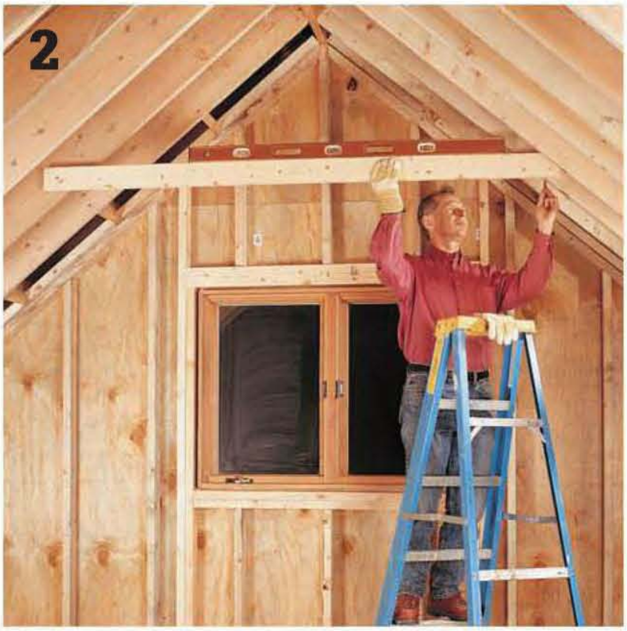

Using a level and the storyboard, level over from the chalk line and mark two outside rafters on the other side of the attic. Snap a chalk line through the marks. *Note: The storyboard is used merely as a straightedge for this step.*

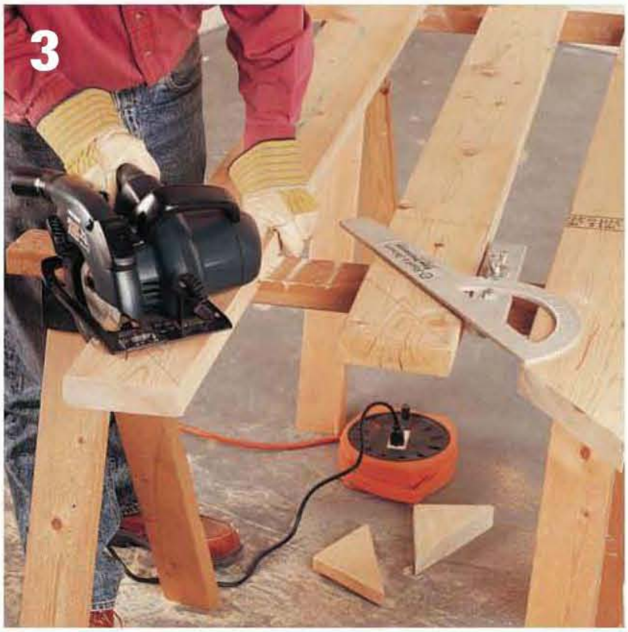

Cut 2 × 6 joists to span across the rafters, angle-cutting the ends to follow the roof pitch. Check each joist for crowning to make sure you're cutting it so it will be installed with the crowned edge up. Make the overall length about ½" short so the ends of the joists won't touch the roof sheathing.

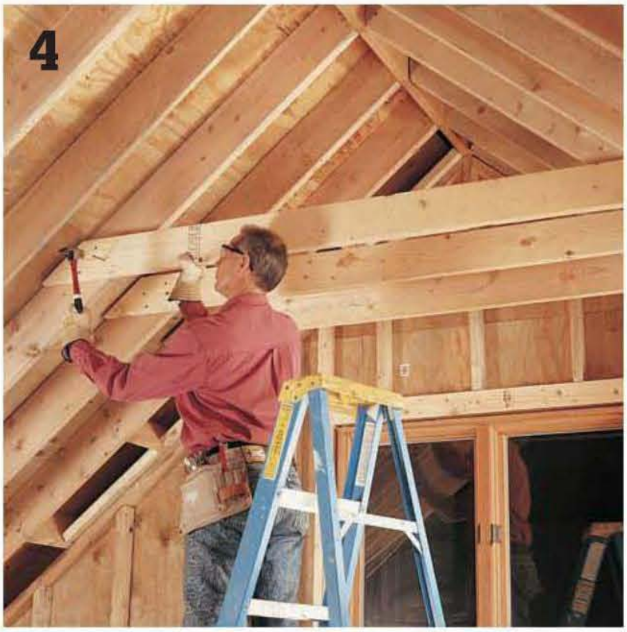

Nail each joist to the rafters with three 10d common nails at each end. Be sure to maintain 16"- or 24"- on-center spacing between joists to provide support for attaching wallboard or other finish material.

Adding Basement Egress Windows

An egress window brings a pleasant source of natural light and ventilation to a dark, dank basement. More importantly, it can provide a life-saving means of escape in the event of a fire. Before you proceed with this project, read more about building code issues regarding basement egress on page 17. Contact your local building department to apply for the proper permits and to learn more about the code requirements for your area.

As long as the window opens wide enough to meet minimum standards for egress, the particular window style is really up to you. Casement windows are ideal, because they crank open quickly and provide unobstructed escape. A tall, double-hung window or wide sliding window can also work. Select a window with insulated glass and clad with vinyl or aluminum for durability; it will be subject to humidity and temperature fluctuations just like any other above-grade window in your home.

The second fundamental component of a basement egress window project is the subterranean escape well you install outside the foundation. There are several options to choose from: prefabricated well kits made of lightweight plastic that bolt together and are easy to install; corrugated metal wells are a lower-cost option; or, you can build a well from scratch using concrete, stone or landscape timber.

Installing an egress window involves four major steps: digging the well, cutting a new or larger window opening in the foundation, installing the window, and finally, installing the well. You'll save time and effort if you hire a backhoe operator to excavate the well. In most cases, you'll also need a large concrete saw (available at most rental stores) to cut the foundation wall.

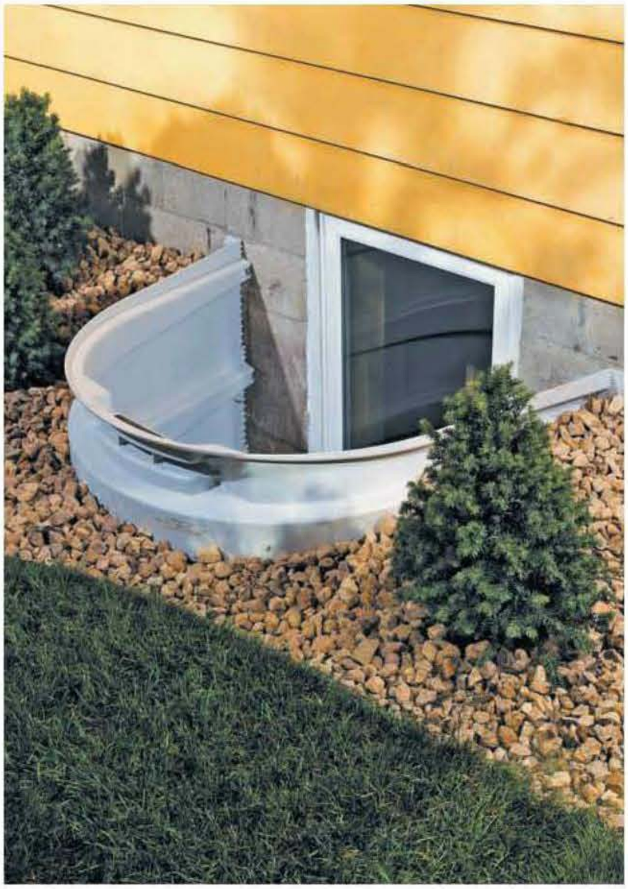

Replacing a small basement window with an egress window is a big job, but it is required if you want to convert part of a basement into livable space, especially a bedroom.

Tools & Materials ▸

- Tape measure
- 4-ft. level
- Stakes and string line
- Shovel
- Colored masking tape
- Hammer drill with ½". dia. × 12- to 16"-long masonry bit
- Concrete saw
- Hand maul
- Wide masonry chisel
- Trowel, miter saw
- Drill/driver, hammer
- Caulk and caulk gun
- Gloves
- Window well and window
- Pea gravel
- Plastic sheeting
- Self-tapping masonry screws
- 2× pressure-treated lumber
- Shims
- Insulation materials
- Concrete sleeve anchors

How to Install an Egress Window & Window Well

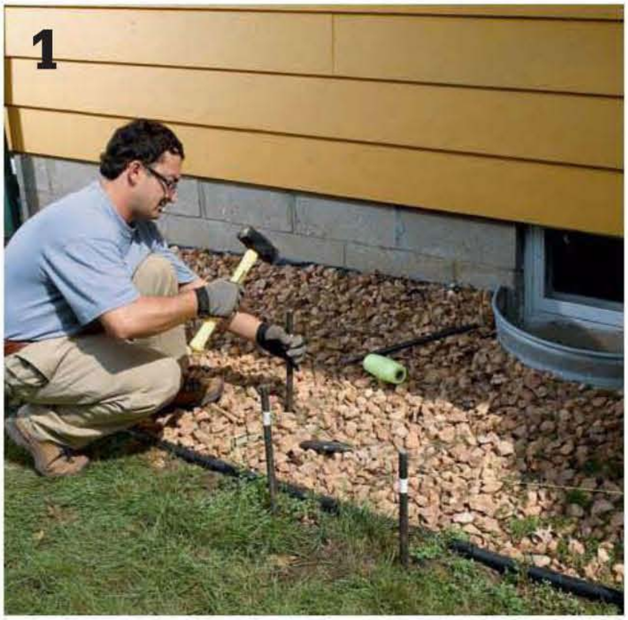

Lay out the border of the window well area with stakes and string. Plan the length and width of the excavation to be several feet larger than the window well's overall size to provide extra room for installation and adjustment.

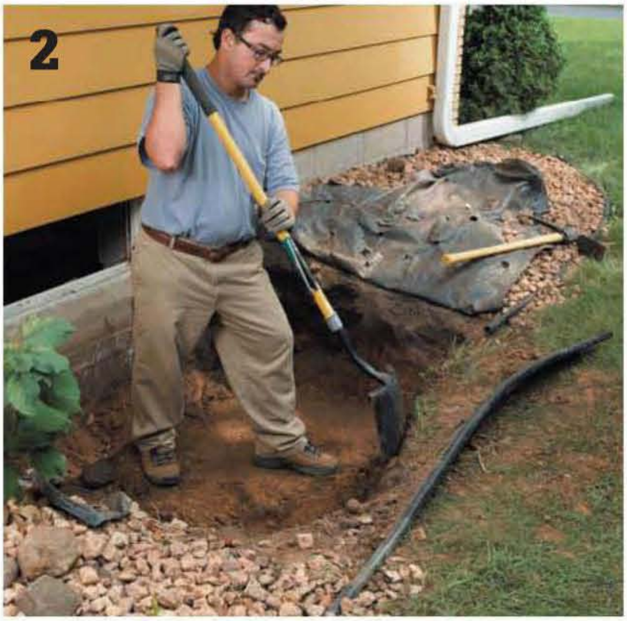

Excavate the well to a depth 6 to 12" deeper than the well's overall height to allow room for drainage gravel. Make sure to have your local public utilities company inspect the well excavation area and okay it for digging before you start.

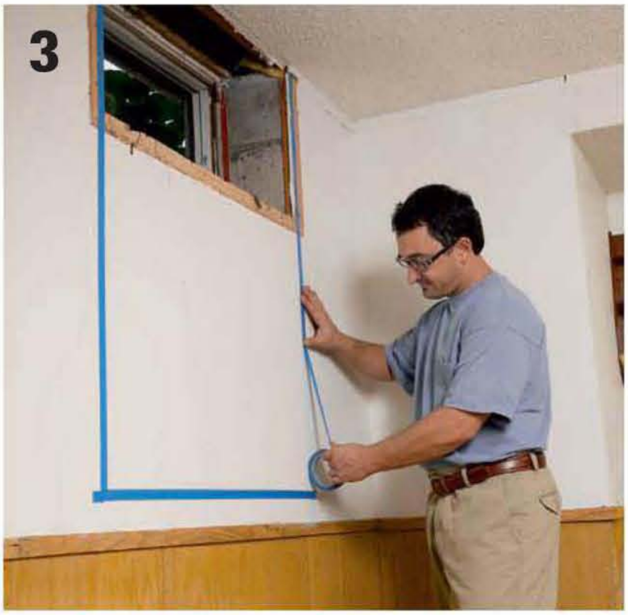

Measure and mark the foundation wall with brightly colored masking tape to establish the overall size of the window's rough opening (here, we're replacing an existing window). Be sure to take into account the window's rough opening dimensions, the thickness of the rough framing (usually 2x stock) and the width of the structural header you may need to build. Remember also that sill height must be within 44" of the floor. Remove existing wall coverings inside the layout area.

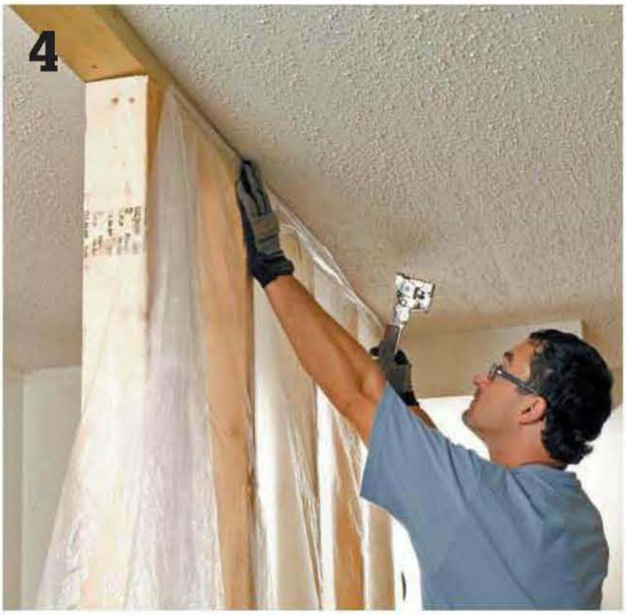

If the floor joists run perpendicular to your project wall, build a temporary support wall parallel to the foundation wall and 6 to 8 ft. from it (for more on building support walls, see page 29). Staple sheet plastic to the wall and floor joists to form a work tent that will help control concrete dust.

(continued)

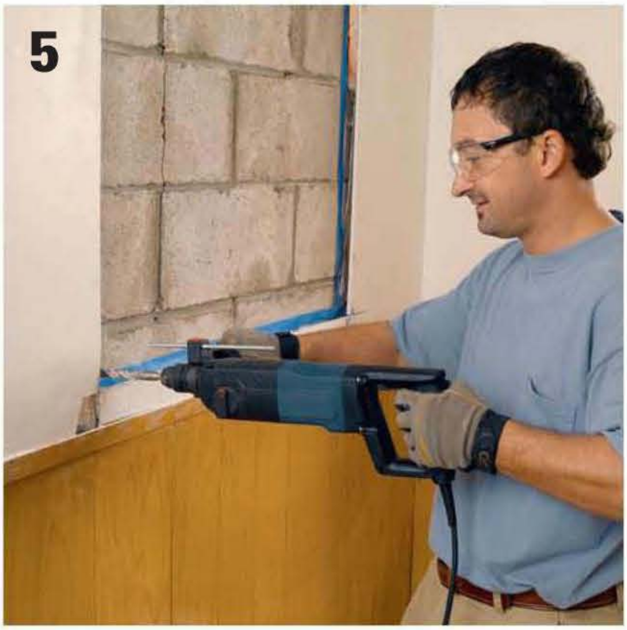

You'll need to cut through the foundation from both the inside and outside so the blocks will break cleanly. Drill a hole through the wall at each bottom corner with a hammer drill and long masonry bit to give you reference points for marking the outside cuts.

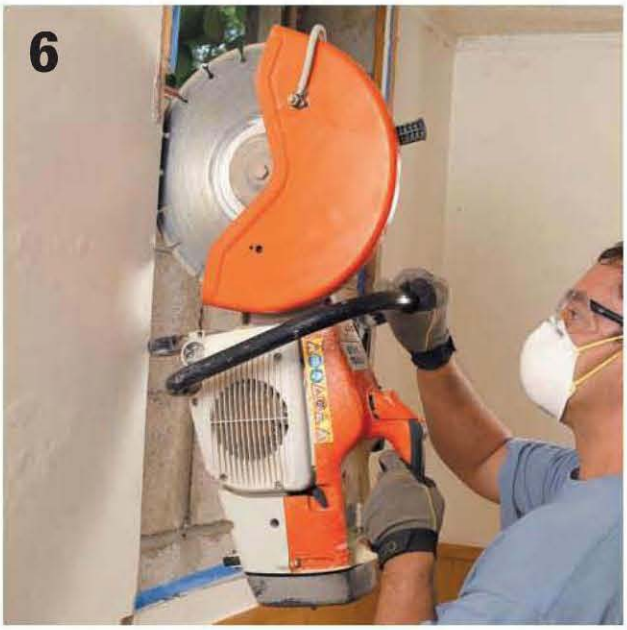

Equip a masonry cutting saw (or large angle grinder) with a diamond blade and set it for a ½" cut to score the blocks first, then reset the saw to full depth and make the final bottom and side cuts through the blocks. Wear a tight-fitting particle mask, ear and eye protection, and gloves for all of this cutting work; the saw will generate a tremendous amount of thick dust. Feed the saw slowly and steadily. Stop and rest periodically so the dust can settle.

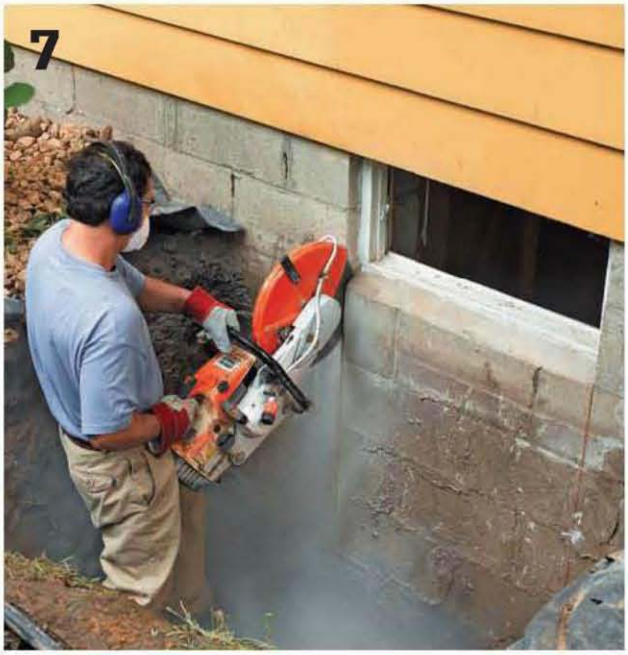

Connect the holes you drilled previously with a level and plumb lines to mark the outside foundation wall for cutting. Measure the opening between the holes to make sure it is the correct size and that it is square. Score the cuts first, then make the full-depth cuts.

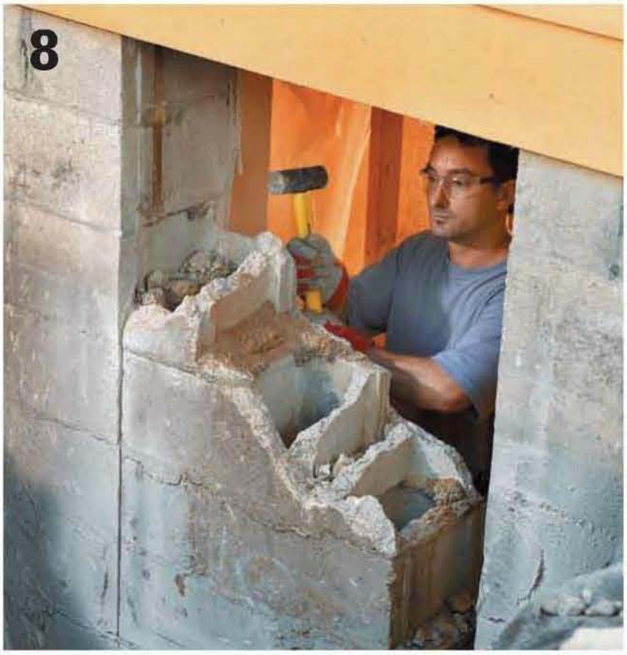

Strike the blocks with a hand maul to break or loosen the block sections. When all the blocks are removed, carefully chip away remaining debris with a cold chisel to create flat surfaces.

9

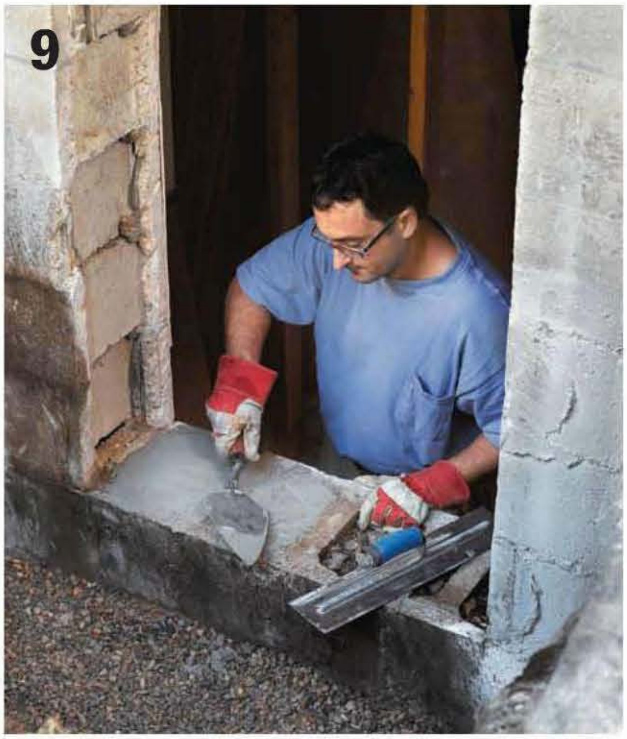

In concrete block walls, fill the hollow voids with broken pieces of block, then level and smooth the voids by trowelling on a fresh layer of quick-curing concrete. Flatten the surfaces, and allow the concrete to dry overnight.

10

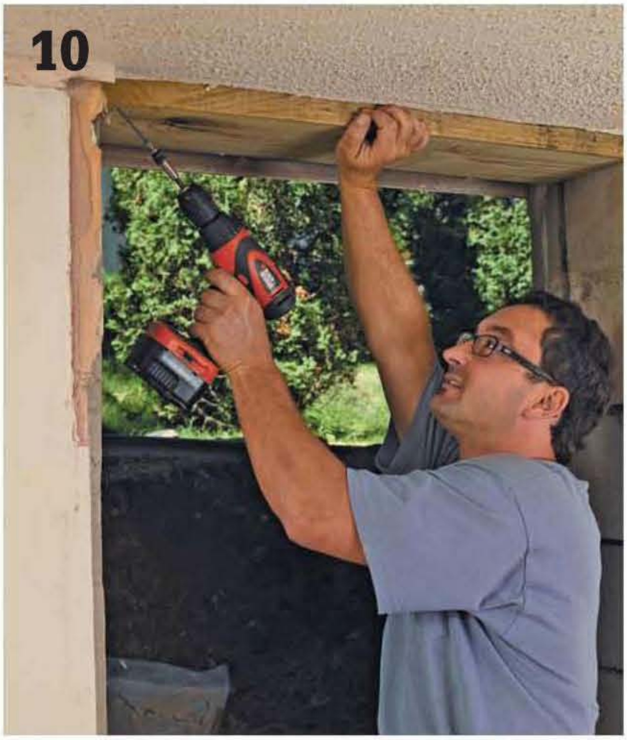

If your project requires a new header above the new window, build it from pieces of 2× lumber sandwiching some ½" plywood and fastened together with construction adhesive and 10d nails. Slip it into place and tack it temporarily to the mudsill with 3½" deck screws driven toenail style.

11

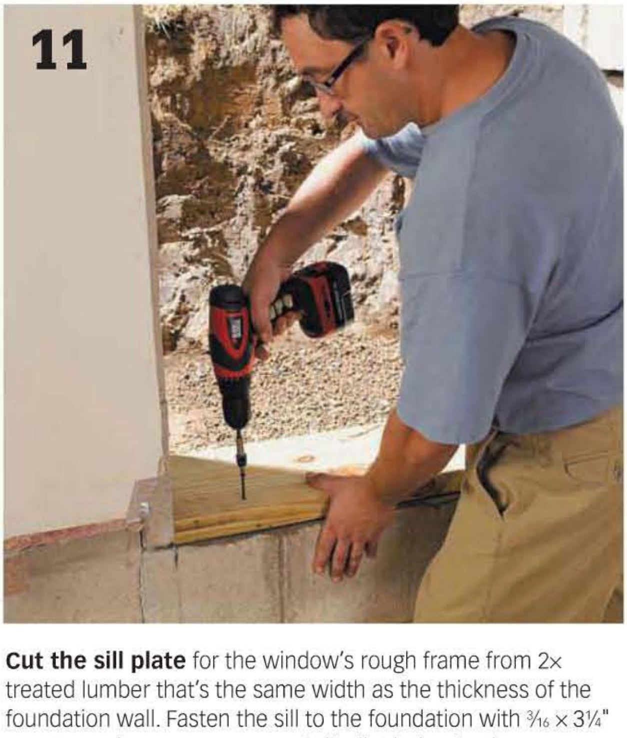

Cut the sill plate for the window's rough frame from 2× treated lumber that's the same width as the thickness of the foundation wall. Fasten the sill to the foundation with 3⁄16 × 3¼" countersunk masonry screws. Drill pilot holes for the screws first with a hammer drill.

12

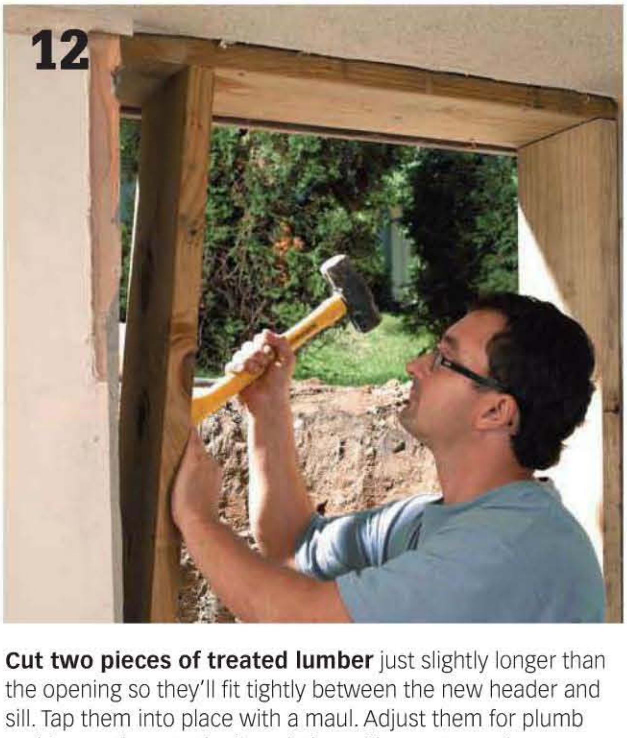

Cut two pieces of treated lumber just slightly longer than the opening so they'll fit tightly between the new header and sill. Tap them into place with a maul. Adjust them for plumb and fasten them to the foundation with countersunk masonry screws or powder-actuated fasteners.

(continued)

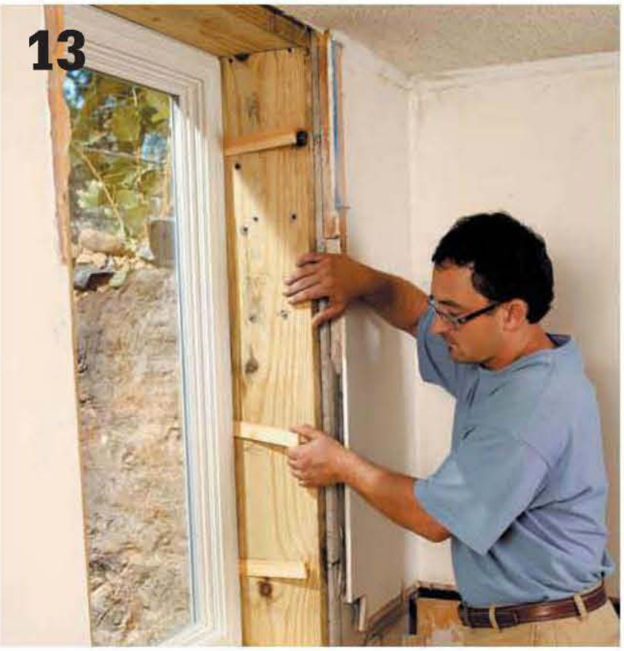

13

Apply a thick bead of silicone caulk around the outside edges of the rough frame and set the window in its opening, seating the nailing flanges into the caulk. Shim the window so the frame is level and plumb. Test the action of the window to make sure the shims aren't bowing the frame.

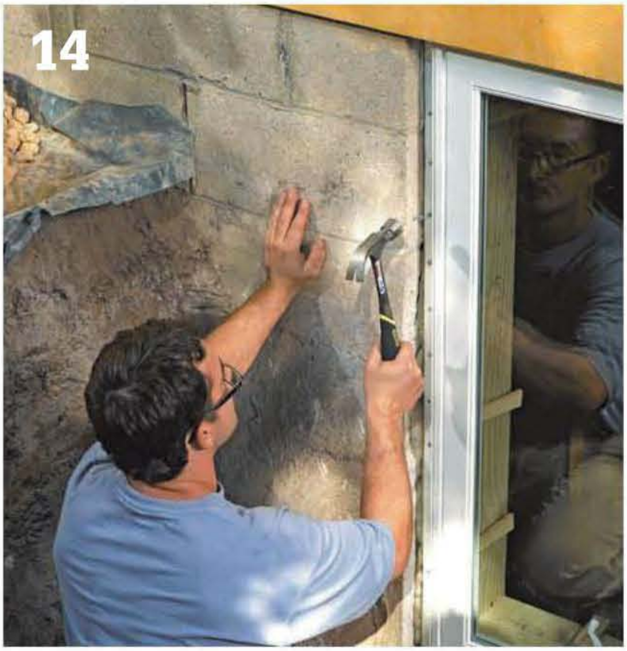

14

Attach the window's nailing flanges to the rough frame with screws or nails, as specified by the manufacturer. Check the window action periodically as you fasten it to ensure that it still operates smoothly.

15

Seal gaps between the rough frame and the foundation with a bead of exterior silicone or polyurethane caulk. If the gaps are wider than ¼", insert a piece of backer rod first, then cover it with caulk. On the interior, fill gaps around the window shims with strips of foam backer rod, fiberglass insulation, or a bead of minimally expanding spray foam. Do not distort the window frame.

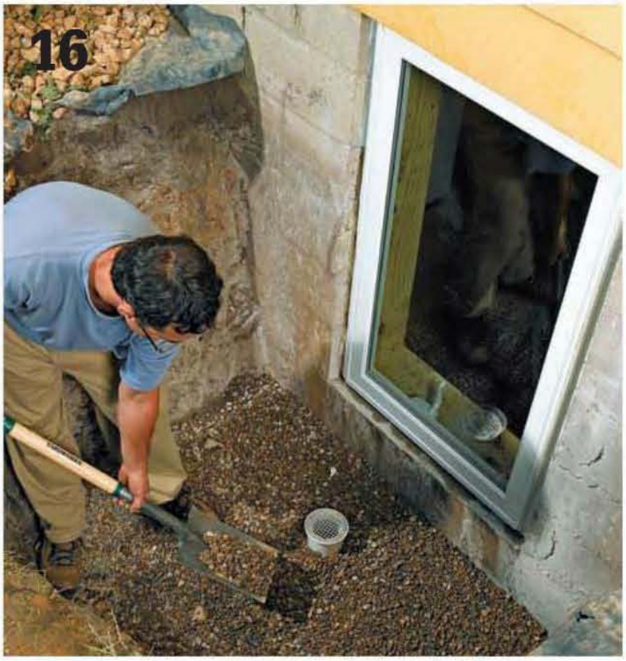

16

Fill the well excavation with 6 to 12" of pea gravel. This will serve as the window's drain system. Follow the egress well kit instructions to determine the exact depth required; you may need to add more gravel so the top of the well will be above the new window. *Note: We added a drain down to the foundation's perimeter tile for improved drainage as well.*

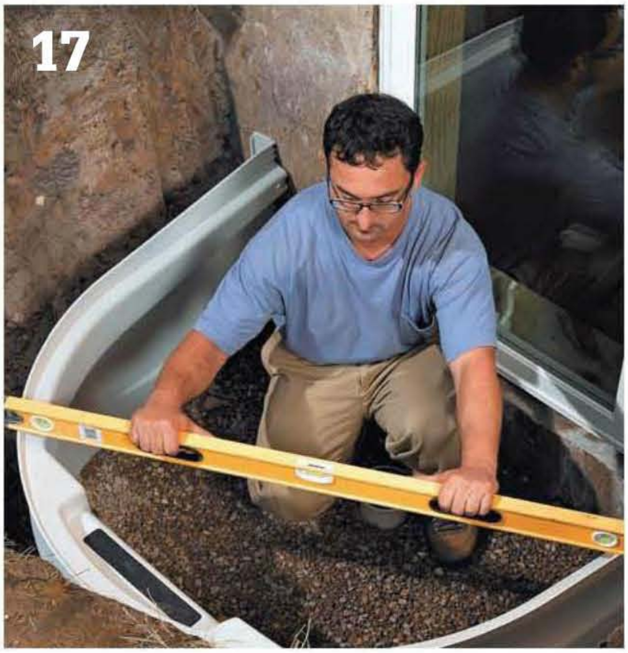

Set the bottom section of the well into the hole, and position it evenly from left to right relative to the window. Adjust the gravel surface to level the well section carefully.

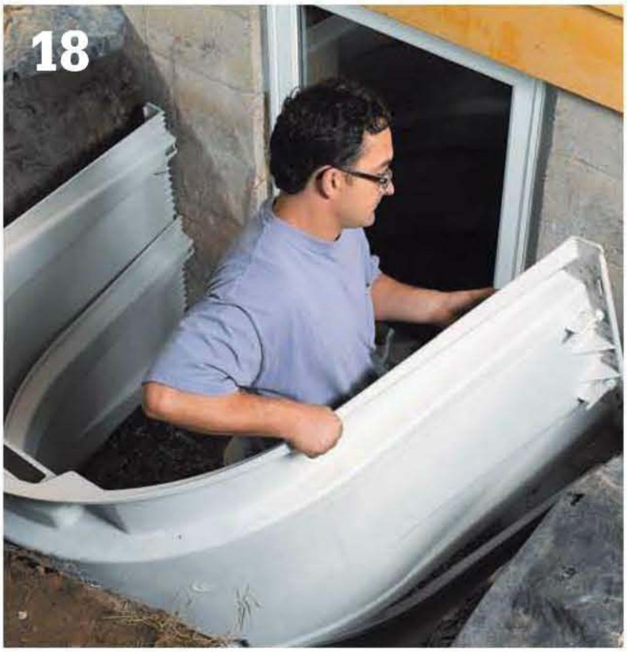

Stack the second well section on top of the first, and connect the two with the appropriate fasteners.

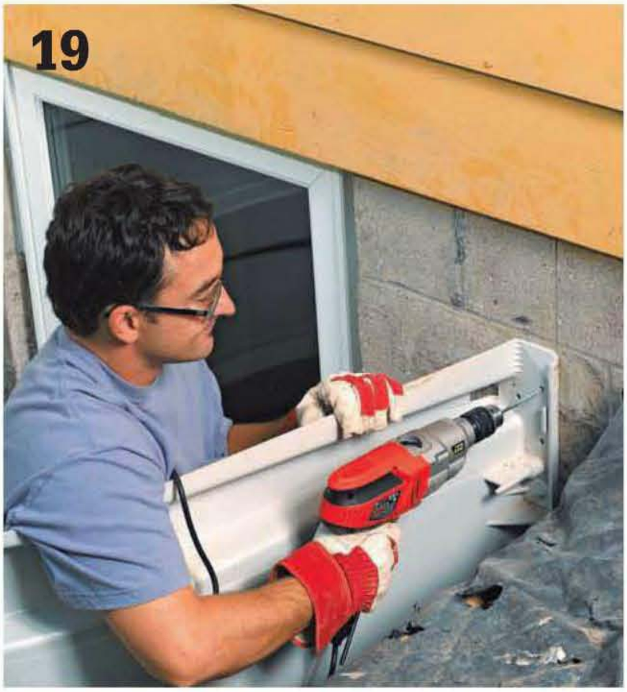

Fasten the window well sections to the foundation wall with concrete sleeve anchors driven into pre-bored pilot holes. You could also use masonry nails driven with a powder actuated nailer.

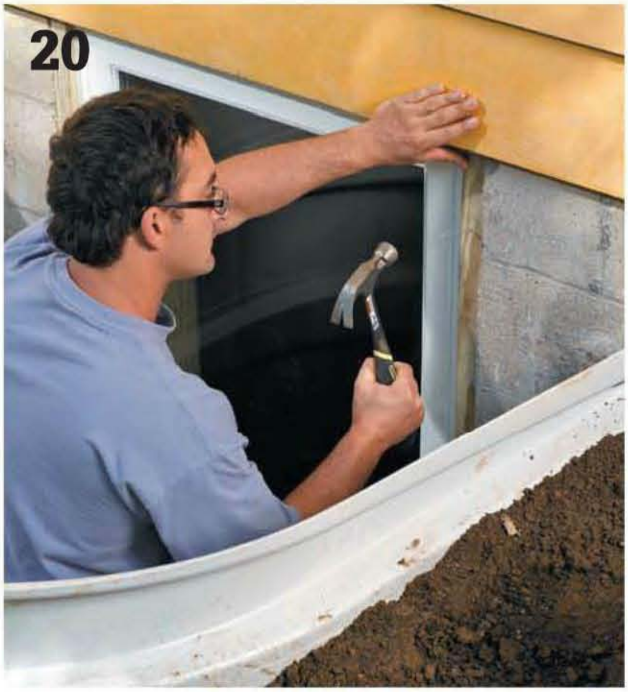

When all the well sections are assembled and secured, nail pieces of treated lumber trim around the window frame to hide the nailing flange. Complete the well installation by using excavated dirt to backfill around the outside of the well. Pack the soil with a tamper, creating a slope for good drainage. If you are installing a window well cover, set it in place and fasten it according to the manufacturer's instructions. (The cover must be removable.)

Installing Baseboard Heaters

Baseboard heaters are a popular way to provide additional heating for an existing room or primary heat to a converted attic or basement.

Heaters are generally wired on a dedicated 240-volt circuit controlled by a thermostat. Several heaters can be wired in parallel and controlled by a single thermostat (see circuit map 11, page 143).

Baseboard heaters are generally surface-mounted without boxes, so in a remodeling situation, you only need to run cables before installing wallboard. Be sure to mark cable locations on the floor before installing drywall. Retrofit installations are also not difficult. You can remove existing baseboard and run new cable in the space behind.

Tools & Materials ▸

Drill/driver
Wire stripper
Cable ripper
Wallboard saw
Baseboard heater or heaters
240-thermostat (in-heater or in-wall)
12/2 NM cable
Electrical tape
Basic wiring supplies

Baseboard heaters can provide primary or supplemental heat for existing rooms or additions. Install heaters with clear space between the heater and the floor.

Baseboard Thermostats

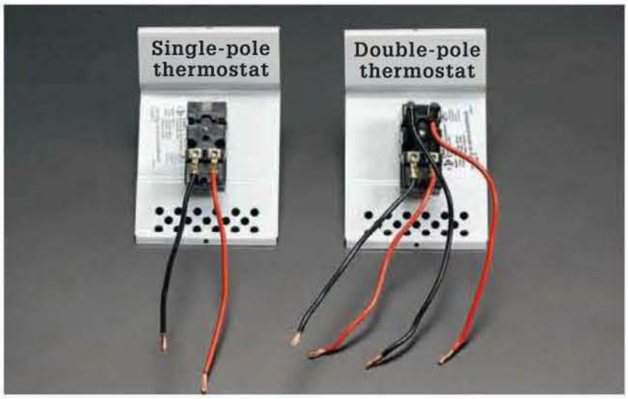

Single-pole and double-pole thermostats work in a similar manner, but double-pole models are safer. The single-pole model will open the circuit (causing shutoff) in only one leg of the power service. Double-pole models have two sets of wires to open both legs, lessening the chance that a person servicing the heater will contact a live wire.

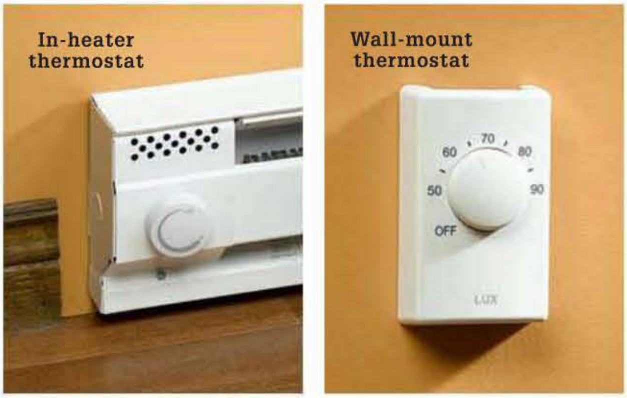

In-heater and wall-mount are the two types of baseboard thermostats you can choose from. If you are installing multiple heaters, a single wall-mount thermostat is more convenient. Individual in-heater thermostats give you more zone control, which can result in energy savings.

How Much Heater Do You Need? ▸

If you don't mind doing a little math, determining how many lineal feet of baseboard heater a room requires is not hard.

1. Measure the area of the room in square feet (length × width): ________
2. Divide the area by 10 to get the baseline minimum wattage: ________
3. Add 5% for each newer window or 10% for each older window: ________
4. Add 10% for each exterior wall in the room: ________
5. Add 10% for each exterior door: ________
6. Add 10% if the space below is not insulated: ________
7. Add 20% if the space above is not well insulated: ________
8. Add 10% if ceiling is more than 8 ft. high: ________
9. Total of the baseline wattage plus all additions: ________
10. Divide this number by 250 (the wattage produced per foot of standard baseboard heater): ________
11. Round up to a whole number. This is the minimum number of feet of heater you need. ________

Note: It is much better to have more feet of heater than is required than fewer. Having more footage of heater does not consume more energy; it does allow the heaters to work more efficiently.

Planning Tips for Baseboard Heaters ▸

- 240-volt heaters are much more energy efficient than 120-volt heaters.
- Baseboard heaters require a dedicated circuit. A 20-amp, 240-volt circuit of 12-gauge copper wire will power up to 16 ft. of heater.
- Do not install a heater beneath a wall receptacle. Cords hanging down from the receptacle are a fire hazard.
- Do not mount heaters directly on the floor. You should maintain at least 1" of clear space between the baseboard heater and the floor covering.
- Installing heaters directly beneath windows is a good practice.
- Locate wall thermostats on interior walls only, and do not install directly above a heat source.

How to Install a 240-volt Baseboard Heater

At the heater locations, cut a small hole in the drywall 3 to 4" above the floor. Pull 12/2 NM cables through the first hole: one from the thermostat, the other to the next heater. Pull all the cables for subsequent heaters. Middle-of-run heaters will have two cables, while end-of-run heaters have only one cable.

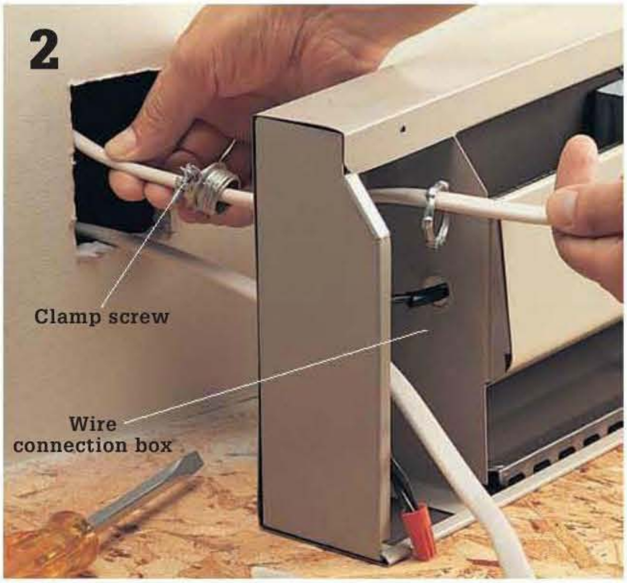

Remove the cover on the wire connection box. Open a knockout for each cable that will enter the box, then feed the cables through the cable clamps and into the wire connection box. Attach the clamps to the wire connection box, and tighten the clamp screws until the cables are gripped firmly.

Anchor heater against wall about 1" off floor by driving flathead screws through back of housing and into studs. Strip away cable sheathing so at least ½" of sheathing extends into the heater. Strip ¾" of insulation from each wire using a combination tool.

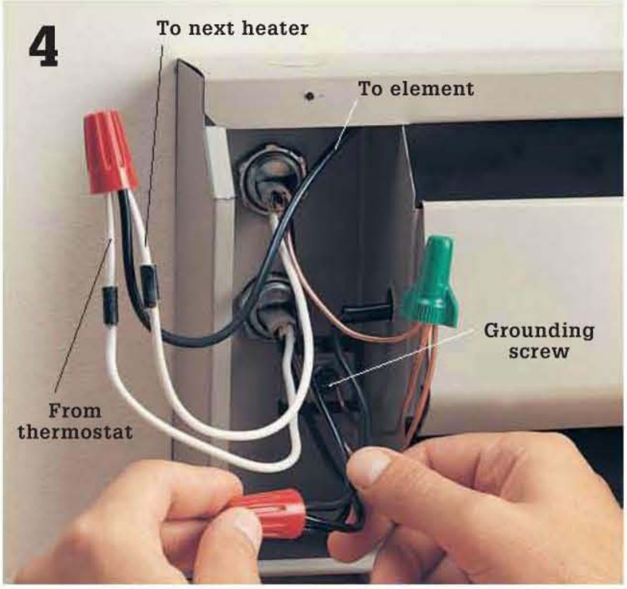

Make connections to the heating element if the power wires are coming from a thermostat or another heater controlled by a thermostat. See next page for other wiring schemes. Connect the white circuit wires to one of the wire leads on the heater. Tag white wires with black tape to indicate they are hot. Connect the black circuit wires to the other wire lead. Connect a grounding pigtail to the green grounding screw in the box, then join all grounding wires with a wire connector. Reattach cover.

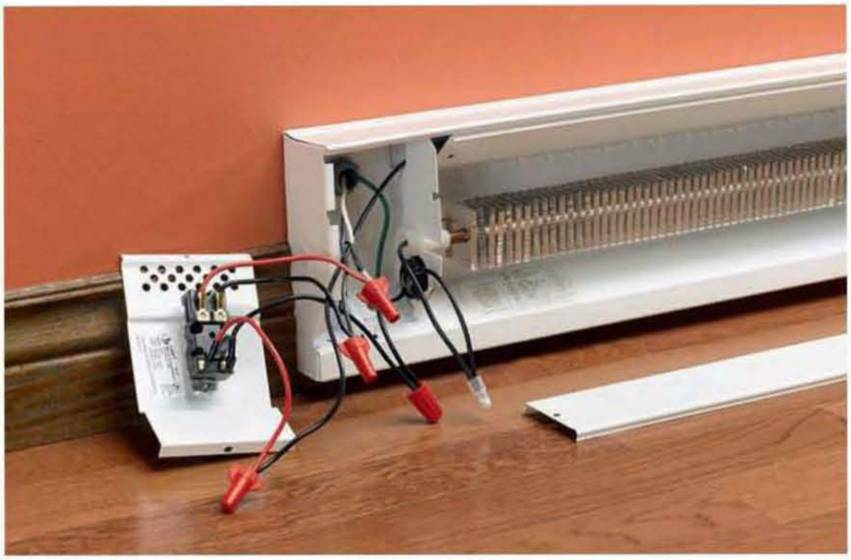

One heater with end-cap thermostat. Run both power leads (black plus tagged neutral) into the connection box at either end of the heater. If installing a single-pole thermostat, connect one power lead to one thermostat wire and connect the other thermostat wire to one of the heater leads. Connect the other hot LINE wire to the other heater lead. If you are installing a double-pole thermostat, make connections with both legs of the power supply.

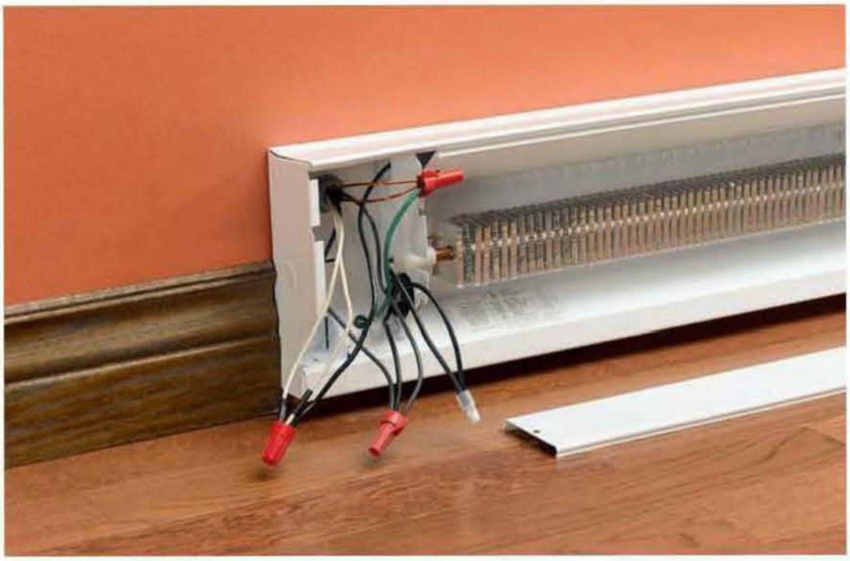

Multiple heaters. At the first heater, join both hot wires from the thermostat to the wires leading to the second heater in line. Be sure to tag all white neutrals hot. Twist copper ground wires together and pigtail them to the grounding screw in the baseboard heater junction box. This parallel wiring configuration ensures that power flow will not be interrupted to the downstream heaters if an upstream heater fails.

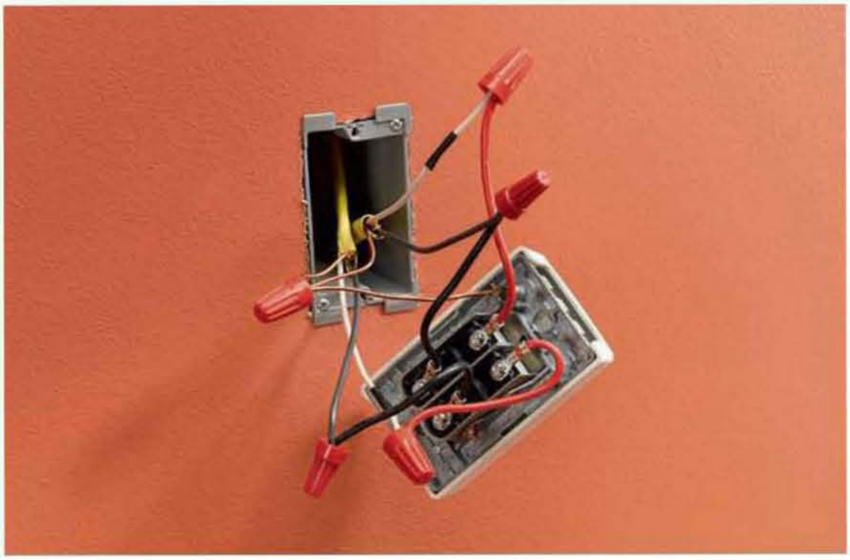

Wall-mounted thermostat. If installing a wall-mounted thermostat, the power leads should enter the thermostat first and then be wired to the individual heaters singly or in series. Hookups at the heater are made as shown in step 4. Be sure to tag the white neutral as hot in the thermostat box as well as in the heater box.

Exterior Improvements

Roofing, siding, and exterior trim details create the outer envelope of your home, safeguarding it from the elements. Aside from their protective purposes, these three systems should also blend together aesthetically to improve your home's appearance. Roofing and siding replacement, in particular, are among the more costly projects you'll face as a homeowner, and generally they're investments you'll make only every few decades or so. For these reasons, it's wise to plan your projects carefully, so you can keep all the relevant factors in mind before you begin.

This chapter will help you get started on the right foot. You'll begin the planning process by evaluating your needs and estimating both price and material quantities required for the job. Then, you'll take a more in-depth look at the wide variety of product options available. New forms and styles of roofing and siding continue to come to market, and there may be some you simply aren't aware of yet. A roofing, siding, or trim project will usually require at least some work at heights. You'll learn important tips for working safely, how to set up scaffolds, and how to prepare your job site to minimize damage, manage debris, and work efficiently.

In this chapter:

- Estimating Roofing, Siding & Trim
- Preparing the Roof for Shingles
- Shingling a Roof
- Installing Vinyl Siding

Estimating Roofing, Siding & Trim

Ordering Roofing Materials

Roofing materials are ordered in squares, with one square equaling 100 square feet. To determine how many squares are needed, first figure out the square footage of your roof. The easiest way to make this calculation is to multiply the length by the width of each section of roof, and then add the numbers together.

For steep roofs and those with complex designs, do your measuring from the ground and multiply by a number based on the slope of your roof. Measure the length and width of your house, include the overhangs, then multiply the numbers together to determine the overall square footage. Using the chart at the lower right, multiply the square footage by a number based on the roof's slope. Add 10 percent for waste, then divide the total square footage by 100 to determine the number of squares you need. Don't spend time calculating and subtracting the areas that won't be covered, such as skylights and chimneys. They're usually small enough that they don't impact the number of squares you need. Besides, it's good to have extra materials for waste, mistakes, and later repairs.

To determine how much flashing you'll need, measure the length of the valley to figure valley flashing, the lengths of the eaves and rakes to figure drip edge, and the number and size of vent pipes to figure vent flashing.

TIMBERLINE
TIMBERLINE

Asphalt shingles come in packaged bundles weighing around 65 pounds each. For typical three-tab shingles, three bundles will cover one square (100 sq. ft.) of roof.

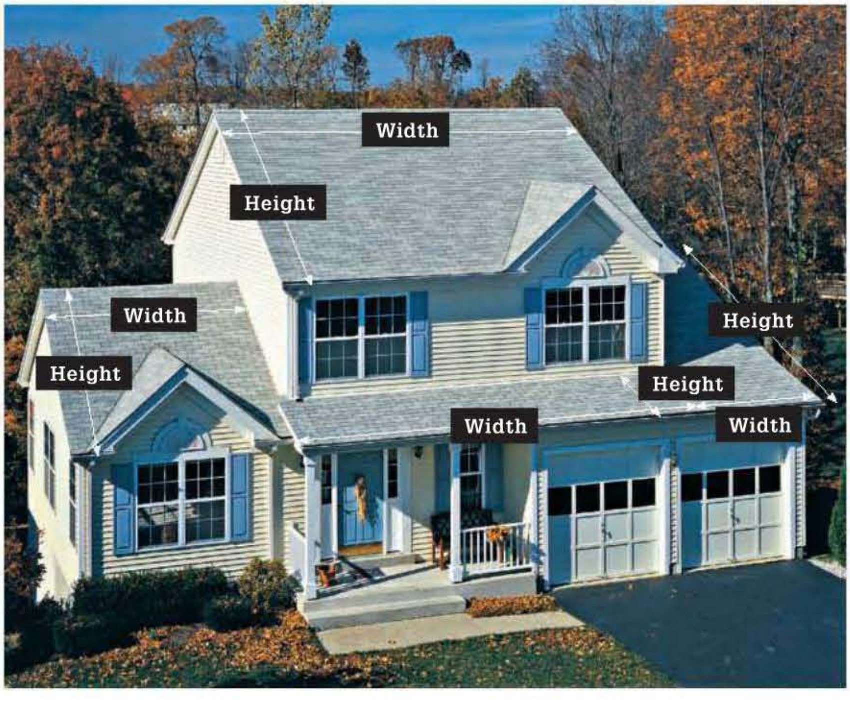

Calculate the roof's surface by multiplying the height of the roof by the width. Do this for each section, then add the totals together. Divide that number by 100, add 10 percent for waste, and that's the number of squares of roofing materials you need.

Conversion Chart

Slope	Multiply by	Slope	Multiply by
2 in 12	1.02	8 in 12	1.20
3 in 12	1.03	9 in 12	1.25
4 in 12	1.06	10 in 12	1.30
5 in 12	1.08	11 in 12	1.36
6 in 12	1.12	12 in 12	1.41
7 in 12	1.16		

Ordering Siding Materials

Siding is sold by the linear foot, square foot, board foot, or square, depending on the type of material. To determine the amount of siding materials you'll need, calculate the square footage of each wall, then add them together to get your total surface area.

To figure the square footage for a wall, multiply the length of the wall by the height. Calculate the area on the gable end of a wall by using the formula for triangles, which is one-half the length of the base multiplied by the height of the triangle. You won't need to subtract for areas covered by windows and doors, and you won't need to add an extra 10 percent for waste. The window and door areas are roughly equal to the amount of waste you'll need.

Use your height and length measurements from the walls to determine how many feet of starter strip, channels, corner posts, and trim you'll need. Keep in mind you'll probably also need to apply trim or channels around all of your doors and windows.

Depending on the siding material you're ordering, you may need extra material to allow for overlap. For example, if you want to install 10" lap siding with an 8" exposure, you'll need to account for a 2" overlap for each board. Likewise, the exposure rate for wood shakes or shingles will determine how many squares you'll need. Depending on the exposure rate, one square of shingles could cover 80 sq. ft, 100 sq. ft, or even 120 sq. ft. Most manufacturers have charts that show how much material is needed to cover a specified number of square feet at various rates of exposure. If you have trouble estimating how much material you need to purchase, ask your supplier for help.

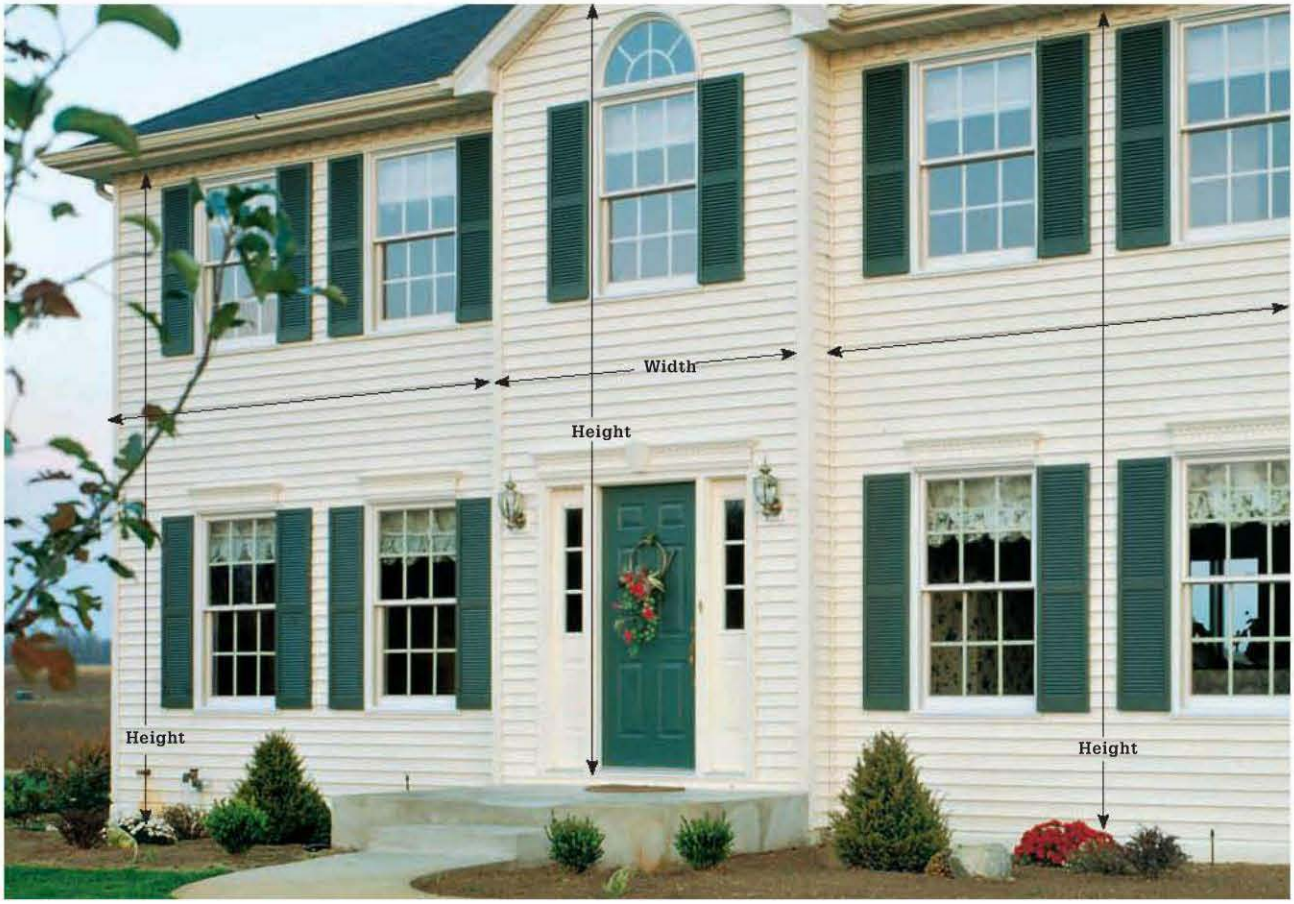

Estimate the amount of siding you'll need by calculating the square footage of each wall, then adding the numbers together. To determine the square footage of a wall, multiply the wall length by its height. Subtract the square footage of all windows and doors, and then add 10 percent for waste to get square footage. Don't forget about corner trim, J-channel and other trim pieces.

Preparing the Roof for Shingles

Working conditions on a roof can be arduous, so make the job as easy as possible by gathering the right tools and equipment before you begin.

Some of these tools, such as a pneumatic nailer and roofer's hatchet, are specific to roofing projects. If you don't have them and don't want to buy them, you can rent them from a rental center.

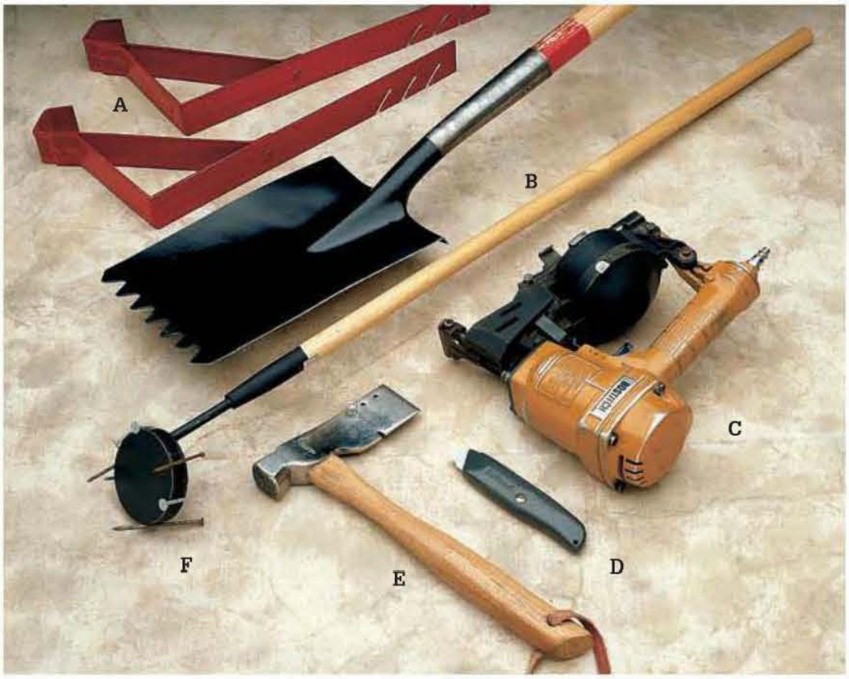

Specialty roofing tools include roof jacks (A) for use on steep roofs, roofing shovel (B) for tearoff work, pneumatic nailer (C), utility knife with hooked blade (D) for trimming shingles, roofing hammer with alignment guides and hatchet blade (E) for shingle installation, and a release magnet for site cleanup (F).

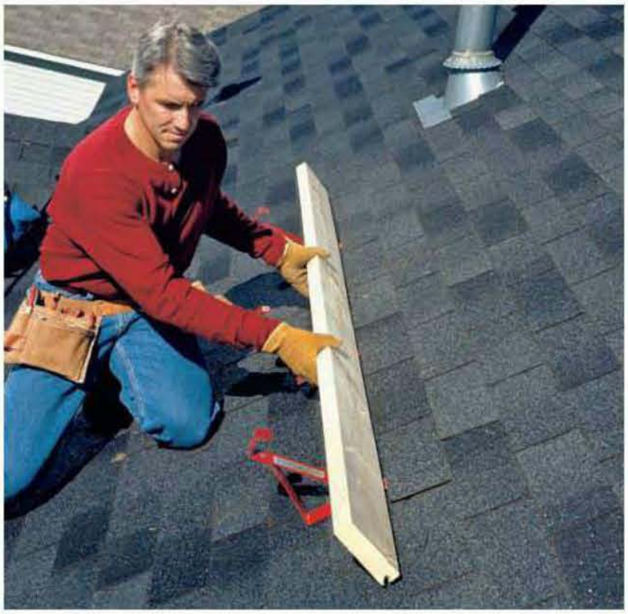

Use roof jacks on steep roofs. Nail the supports at the fourth or fifth course of shingles, and add the widest board the supports will hold.

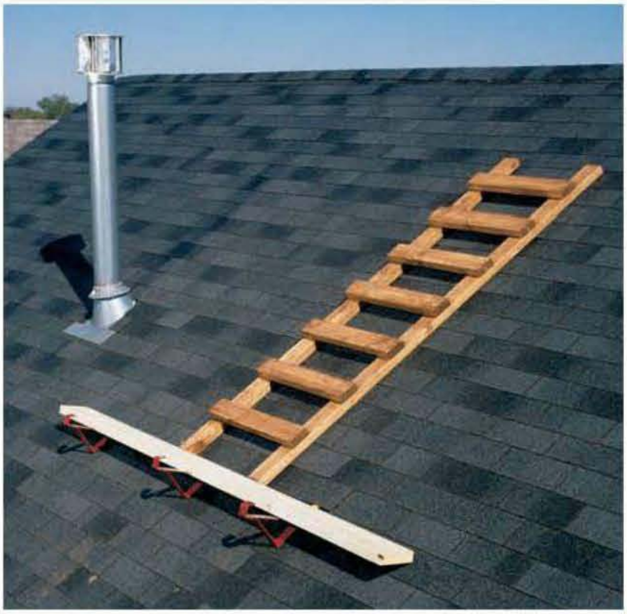

For more secure footing, fashion a roofing ladder by nailing wood strips across a pair of 2 × 4s. Secure the ladder to the roof jacks, and use it to maintain your footing.

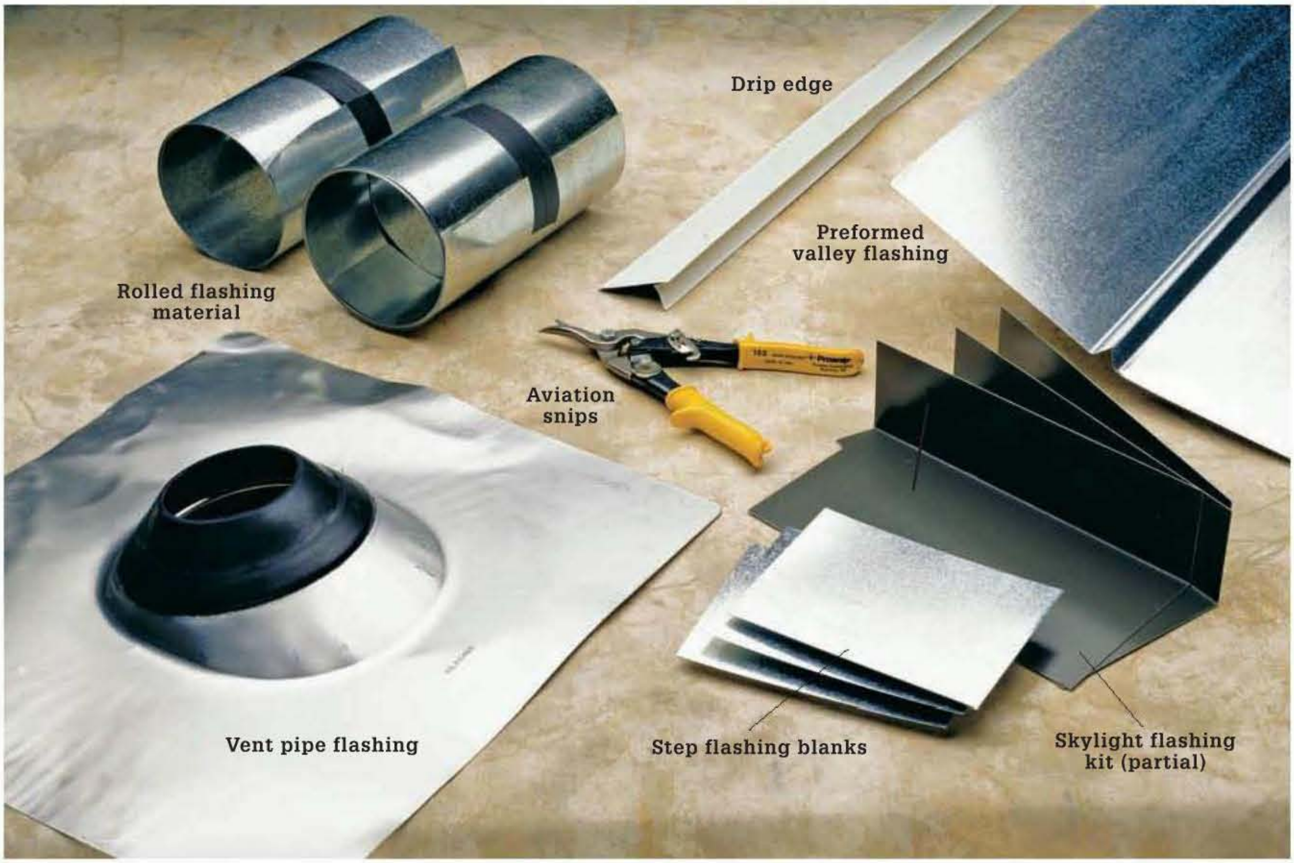
Drip edge
Preformed valley flashing
Rolled flashing material
Aviation snips
Vent pipe flashing
Step flashing blanks
Skylight flashing kit (partial)

Roof flashing can be hand cut or purchased in preformed shapes and sizes. Long pieces of valley flashing, base flashing, top saddles, and other nonstandard pieces can be cut from rolled flashing material using aviation snips. Step flashing blanks can be bought in standard sizes and bent to fit. Drip edge and vent pipe flashing are available preformed. Skylight flashing usually comes as a kit with the window. Complicated flashings, such as chimney crickets, can be custom fabricated by a metalworker.

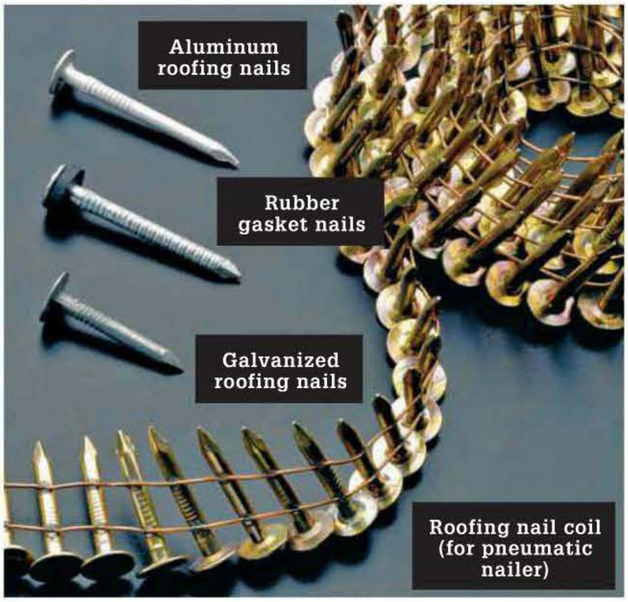
Aluminum roofing nails
Rubber gasket nails
Galvanized roofing nails
Roofing nail coil (for pneumatic nailer)

Different fasteners are specially developed for different jobs. Use galvanized roofing nails to hand nail shingles; use aluminum nails for aluminum flashing; use rubber gasket nails for galvanized metal flashing; and use nail coils for pneumatic nailers.

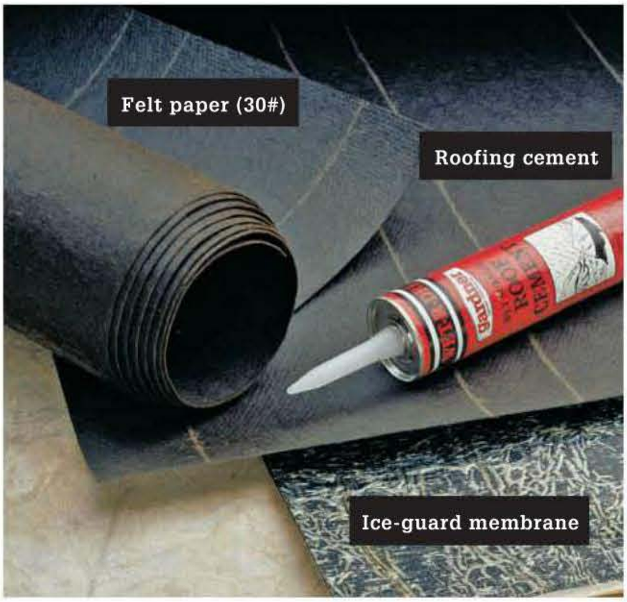
Felt paper (30#)
Roofing cement
Ice-guard membrane

Common roofing materials include 30# felt paper for use as underlayment; ice-guard membrane for use as underlayment in cold climates; and tubes of roofing cement for sealing small holes, cracks, and joints.

How to Tear Off Old Shingles

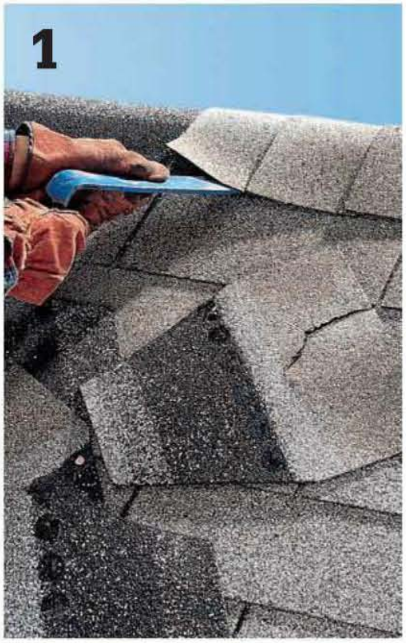

Remove the ridge cap using a flat pry bar. Pry up the cap shingles at the nail locations.

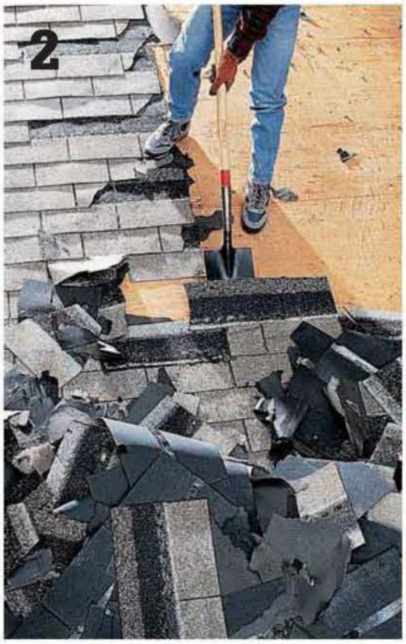

Working downward from the peak, tear off the felt paper and old shingles with a roofing shovel or pitchfork.

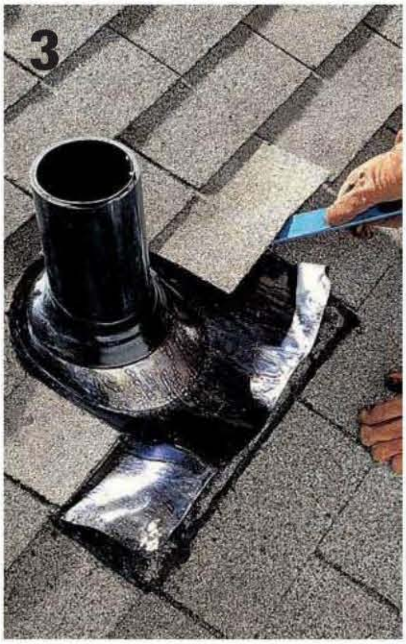

Unless flashing is in exceptional condition, remove it by slicing through the roofing cement that attaches it to the shingles. You may be able to salvage flashing pieces, such as chimney saddles and crickets, and reuse them.

After removing the shingles, felt paper, and flashing from the entire tear-off section, pry out any remaining nails and sweep the roof with a broom.

If an unexpected delay keeps you from finishing a section before nightfall, cover any unshingled sections, using tarps weighted down with shingle bundles.

How to Replace Damaged Sheathing

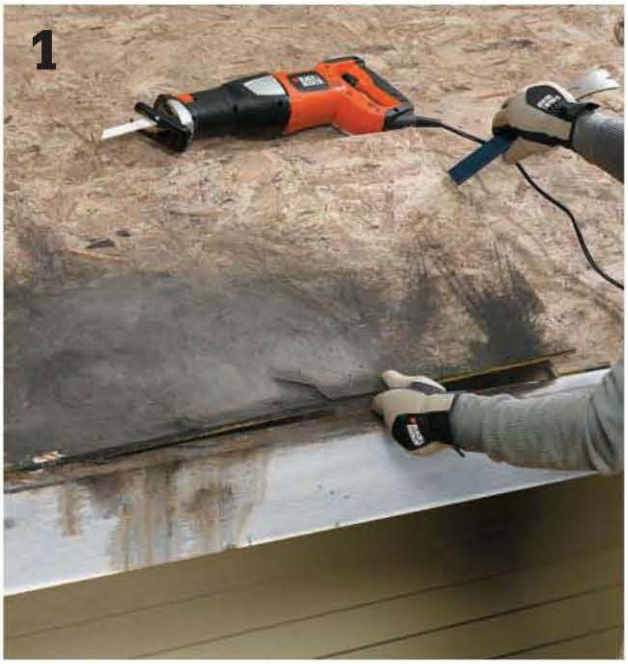

Use a reciprocating saw to cut next to the rafters in an area that extends well beyond the damaged area. Pry out the damaged sections using a pry bar.

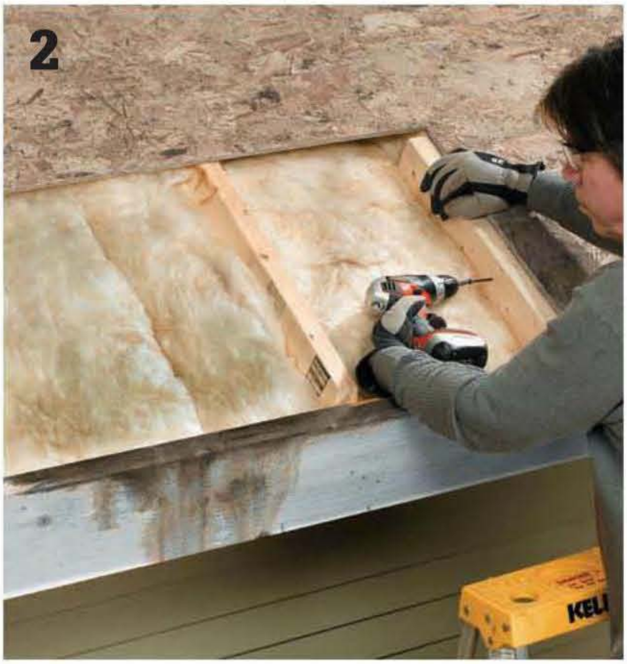

Attach 2 × 4 nailing strips to the inside edges of the rafters using 3" deck screws

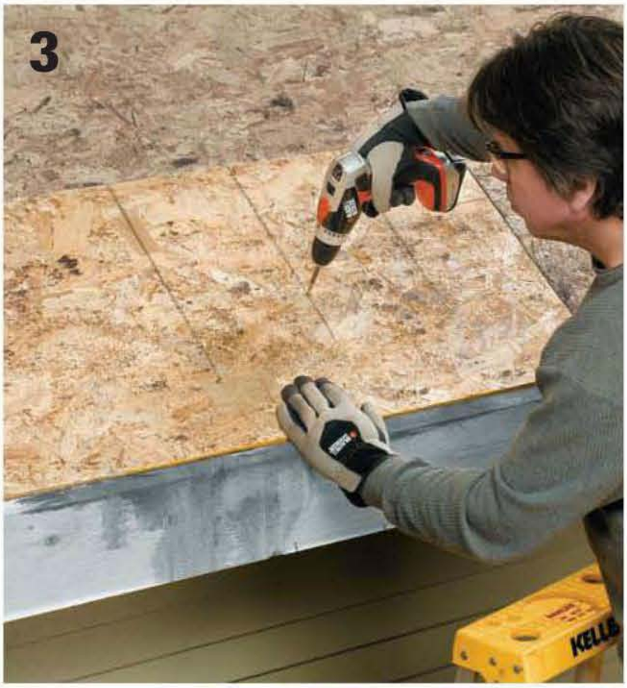

Use exterior grade plywood to make a patch. Measure the cutout area, allow for a ⅛" gap on all sides for expansion, and cut the patch to size. Attach the patch to the rafters and nailing strips using 2¼" deck screws or 8d ring-shank siding nails.

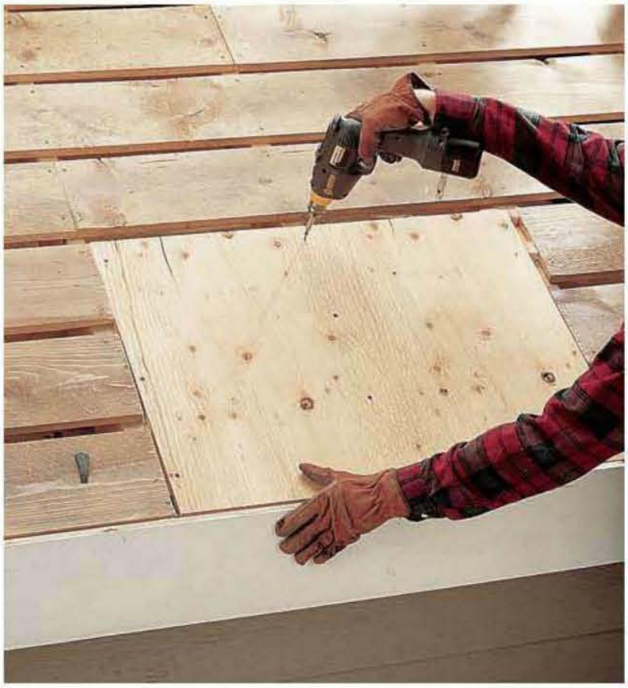

Option: If your existing roof deck is made of boards (1 × 6 was common before plywood took over the market), it is perfectly acceptable to use plywood when replacing a section of the deck. The plywood should be the same thickness as the boards, generally, ¾".

How to Install Underlayment

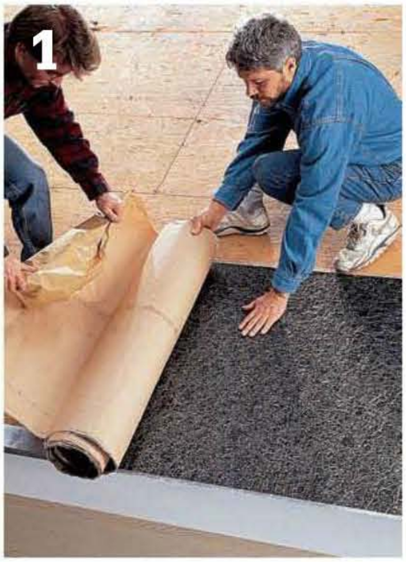

Snap a chalk line 35⅝" up from the eaves, so the first course of the 36"-wide membrane will overhang the eaves by ⅜". Install a course of ice and water shield, using the chalk line as a reference, and peeling back the protective backing as you unroll it.

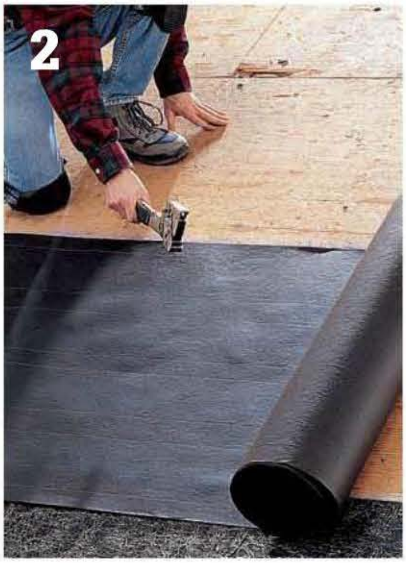

Measuring up from the eaves, make a mark 32" above the top of the last row of underlayment, and snap another chalk line. Roll out the next course of felt paper (or ice guard, if required) along the chalk line, overlapping the first course by 4". *Tip: Drive staples every 6 to 12" along the edges of felt paper, and one staple per sq. ft. in the field area.*

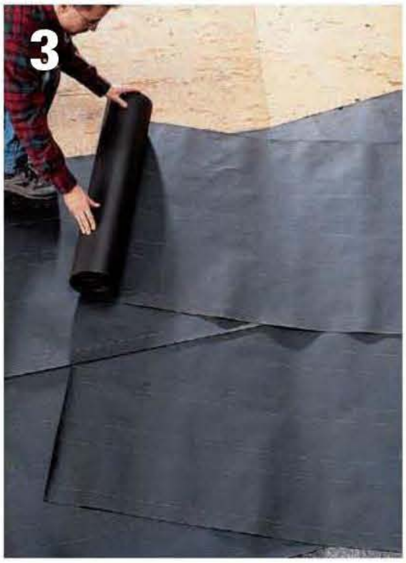

At valleys, roll felt paper across from both sides, overlapping the ends by 36". Install felt paper up to the ridge—ruled side up—snapping horizontal lines every two or three rows to check alignment. Overlap horizontal seams by 4", vertical seams by 12", and hips and ridges by 6". Trim the courses flush with the rake edges.

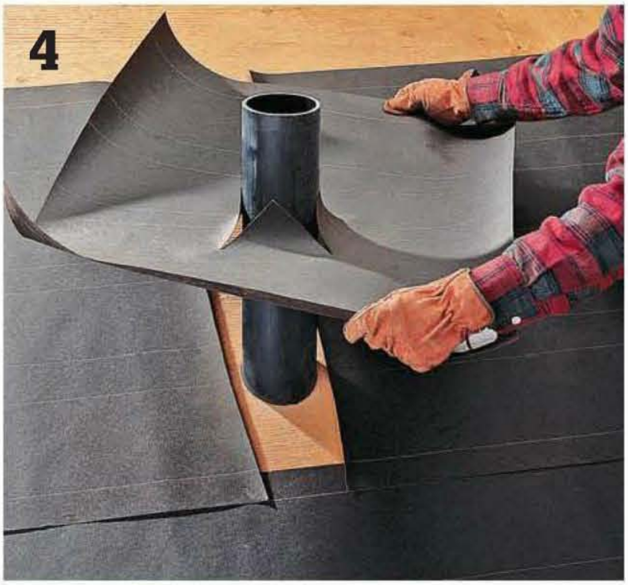

Apply felt paper up to an obstruction, then resume laying the course on the opposite side (make sure to maintain the line). Cut a patch that overlaps the felt paper by 12" on all sides. Make a crosshatch cutout for the obstruction. Position the patch over the obstruction, staple it in place, then caulk the seams with roofing cement.

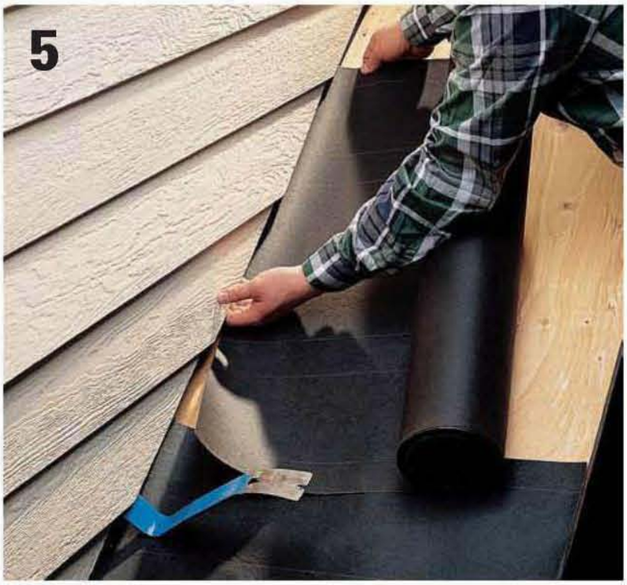

At the bottom of dormers and sidewalls, tuck the felt paper under the siding where it intersects with the roof. Carefully pry up the siding and tuck at least 2" of paper under it. Also tuck the paper under counter flashing or siding on chimneys and skylights. Leave the siding or counter flashing or siding unfastened until after you install the step flashing.

How to Install Drip Edge

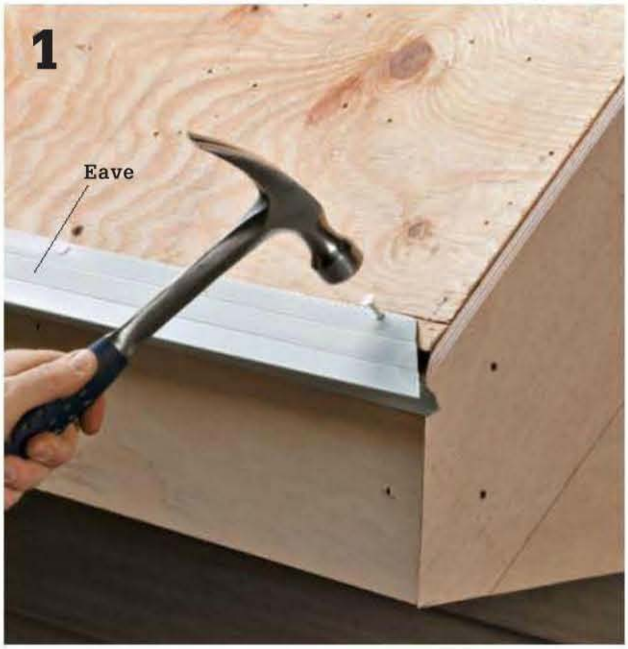

Cut a 45° miter at one end of the drip edge using aviation snips. Place the drip edge along the eaves end of the roof, aligning the mitered end with the rake edge. Nail the drip edge in place every 12".

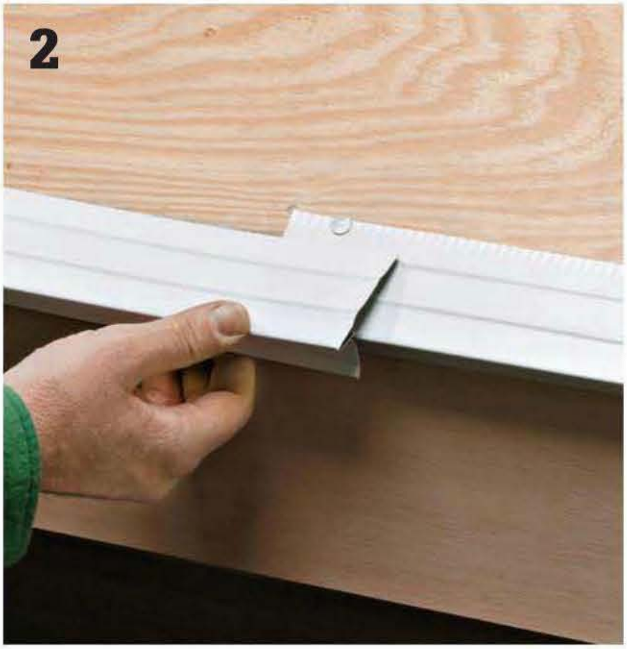

Overlap pieces of drip edge by 2". Install drip edge across the entire eaves, ending with a mitered cut on the opposite corner.

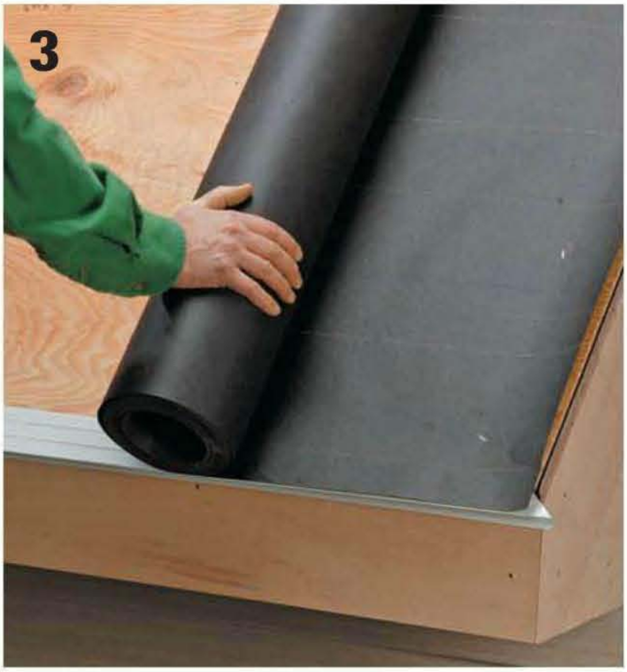

Apply felt paper, and ice guard if needed, to the roof, overhanging the eaves by ⅜".

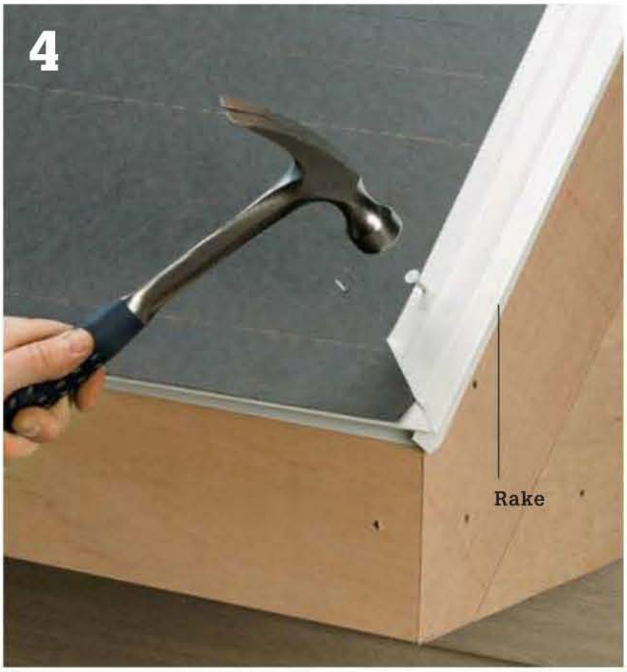

Cut a 45° miter in a piece of drip edge and install it along the rake edge, forming a miter joint with the drip edge along the eaves. Overlap pieces by 2", making sure the higher piece is on top at the overlap. Apply drip edge all the way to the peak. Install drip edge along the other rake edges the same way.

Shingling a Roof

If you want to install asphalt shingles on your roof, then you're in good company. Asphalt shingles, also known as composition shingles, are the roofing of choice for nearly four out of five homeowners in America. They perform well in all types of climate, are available in a multitude of colors, shapes, and textures to complement every housing design, and are less expensive than most other roofing products.

Asphalt shingles are available as either fiberglass shingles or organic shingles. Both types are made with asphalt, the difference being that one uses a fiberglass reinforcing mat, while the other uses a cellulose-fiber mat. Fiberglass shingles are lighter, thinner, and have a better fire rating. Organic shingles have a higher tear strength, are more flexible in cold climates, and are used more often in northern regions.

Although the roofing market has exploded with innovative new asphalt shingle designs, such as the architectural or laminated shingle that offers a three-dimensional look, the standard three-tab asphalt shingle is still the most common, which is the project we're featuring here. The tabs provide an easy reference for aligning shingles for installation.

To help the job get done faster, rent an air compressor and pneumatic roofing gun. This will greatly reduce the time you spend nailing.

Tools & Materials ▸

Aviation snips
Carpenter's square
Chalk line
Flat bar
Roofer's hatchet or pneumatic nailer
Utility knife
Straightedge
Tape measure
Chalk gun
Flashing
Shingles
Nailing cartridges
Roofing cement
Roofing nails (⅞", 1¼")
Rubber gasket nails

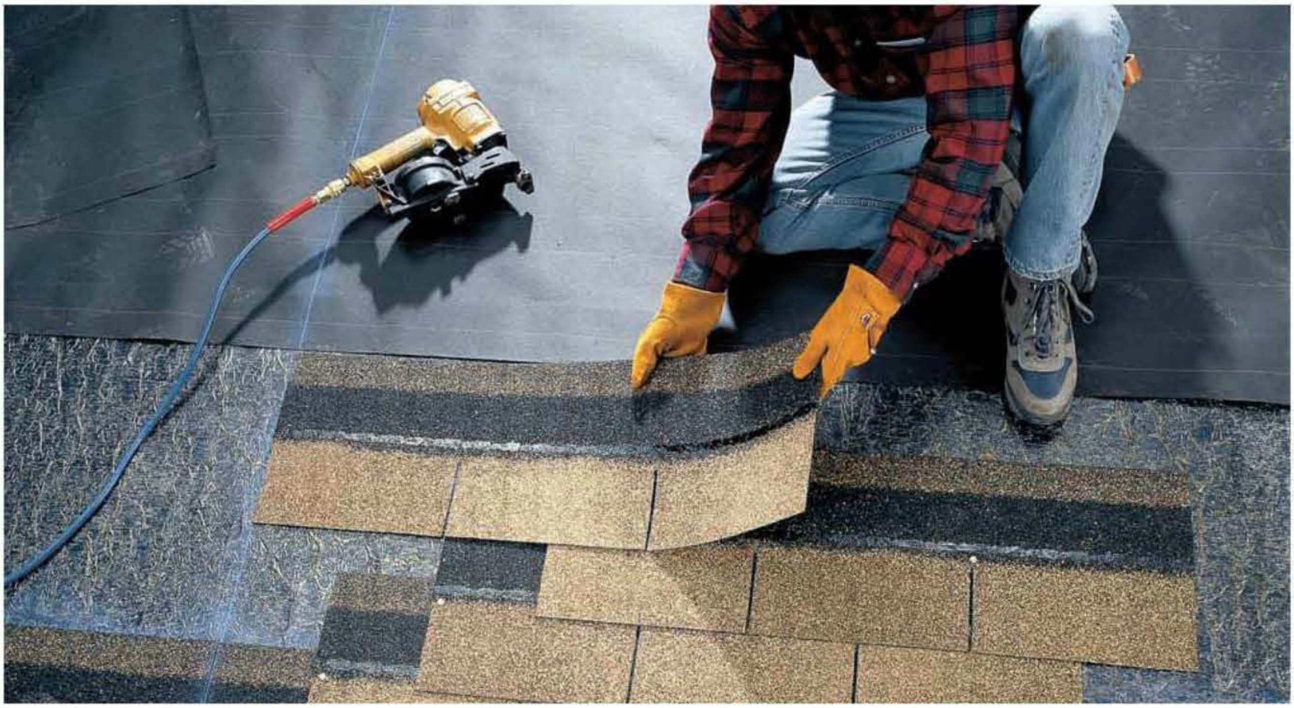

Stagger shingles for effective protection against leaks. If the tab slots are aligned in successive rows, water forms channels, increasing erosion of the mineral surface of the shingles. Creating a 6" offset between rows of shingles—with the three-tab shingles shown above—ensures that the tab slots do not align.

How to Install Three-tab Shingles

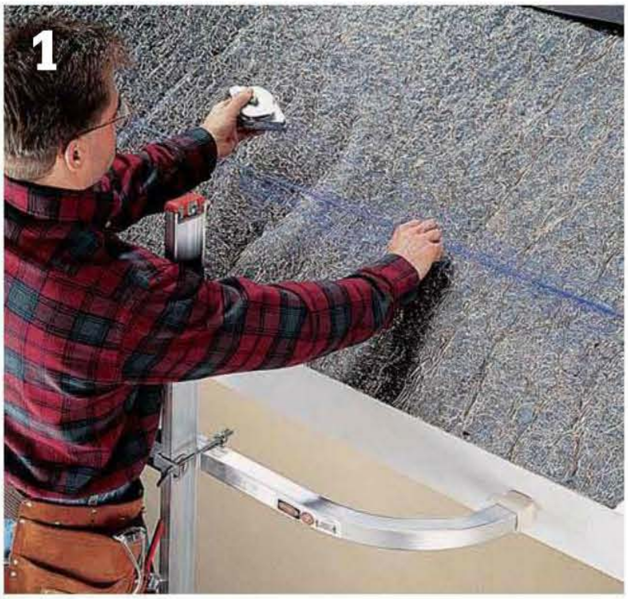

Cover the roof with felt paper (page 536) and install drip edge (page 537). Snap a chalk line onto the felt paper or ice guard 11½" up from the eaves edge, to mark the alignment of the starter course. This will result in a ½" shingle overhang for standard 12" shingles. *Tip: Use blue chalk rather than red. Red chalk will stain roofing materials.*

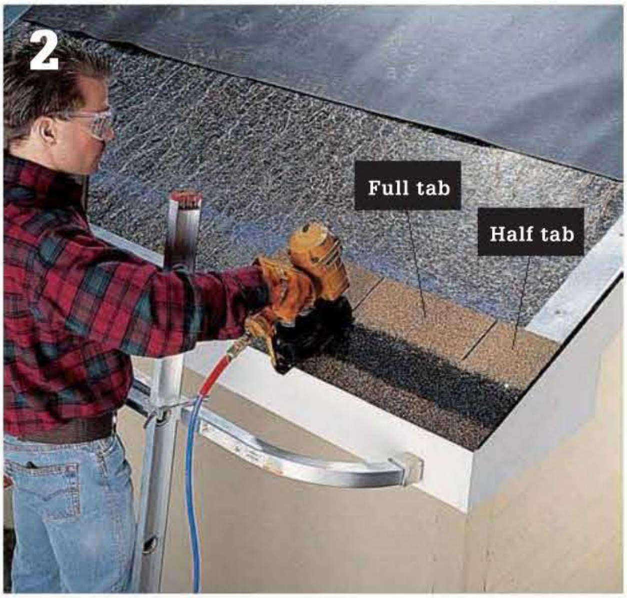

Trim off one-half (6") of an end tab on a shingle. Position the shingle upside down, so the tabs are aligned with the chalk line and the half-tab is flush against the rake edge. Drive ⅞" roofing nails near each end, 1" down from each slot between tabs. Butt a full upside-down shingle next to the trimmed shingle, and nail it. Fill out the row, trimming the last shingle flush with the opposite rake edge.

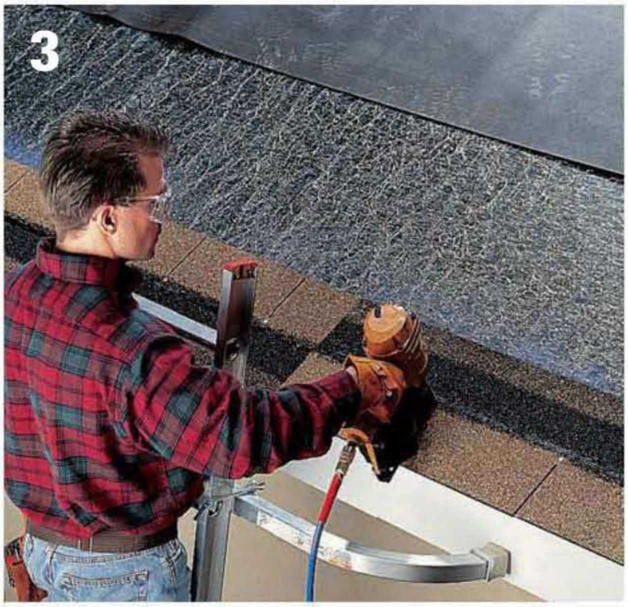

Apply the first full course of shingles over the starter course with the tabs pointing down. Begin at the rake edge where you began the starter row. Place the first shingle so it overhangs the rake edge by ⅜" and the eaves edge by ½". Make sure the top of each shingle is flush with the top of the starter course, following the chalk line.

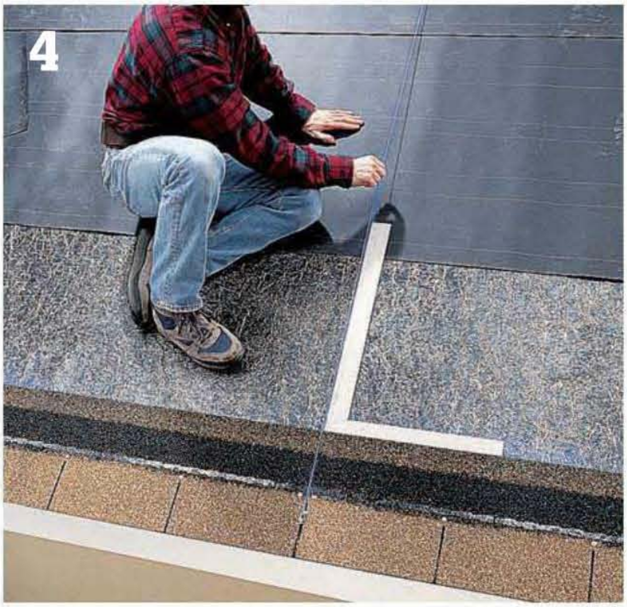

Snap a chalk line from the eaves edge to the ridge to create a vertical line to align the shingles. Choose an area with no obstructions, as close as possible to the center of the roof. The chalk line should pass through a slot or a shingle edge on the first full shingle course. Use a carpenter's square to establish a line perpendicular to the eaves edge.

(continued)

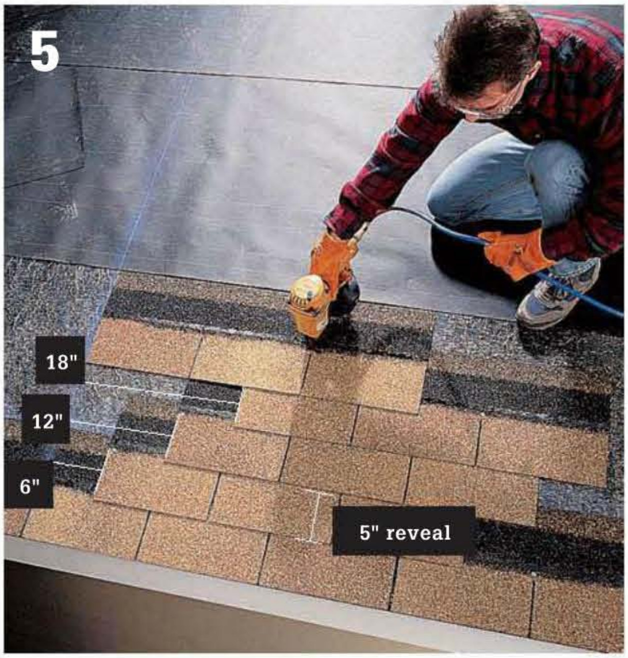

Use the vertical reference line to establish a shingle pattern with slots that are offset by 6" in succeeding courses. Tack down a shingle 6" to one side of the vertical line, 5" above the bottom edge of the first-course shingles to start the second row. Tack down shingles for the third and fourth courses, 12" and 18" from the vertical line. Butt the fifth course against the line.

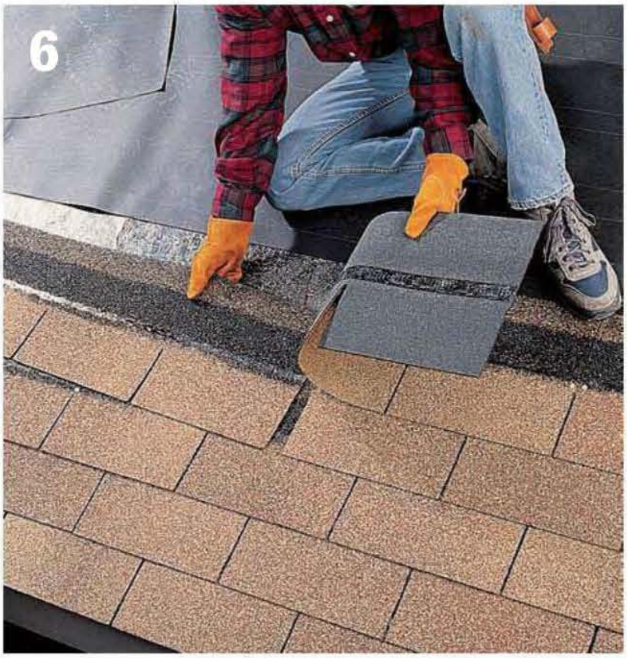

Fill in shingles in the second through fifth courses, working upward from the second course and maintaining a consistent 5" reveal. Slide lower-course shingles under any upper-course shingles left partially nailed, and then nail them down. *Tip: Install roof jacks, if needed, after filling out the fifth course.*

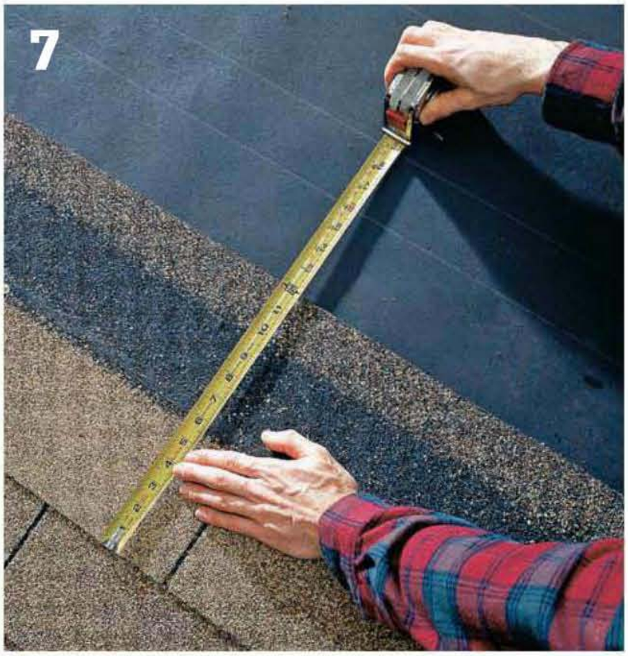

Check the alignment of the shingles after each four-course cycle. In several spots on the last installed course, measure from the bottom edge of a shingle to the nearest felt paper line. If you discover any misalignment, make minor adjustments over the next few rows until it's corrected.

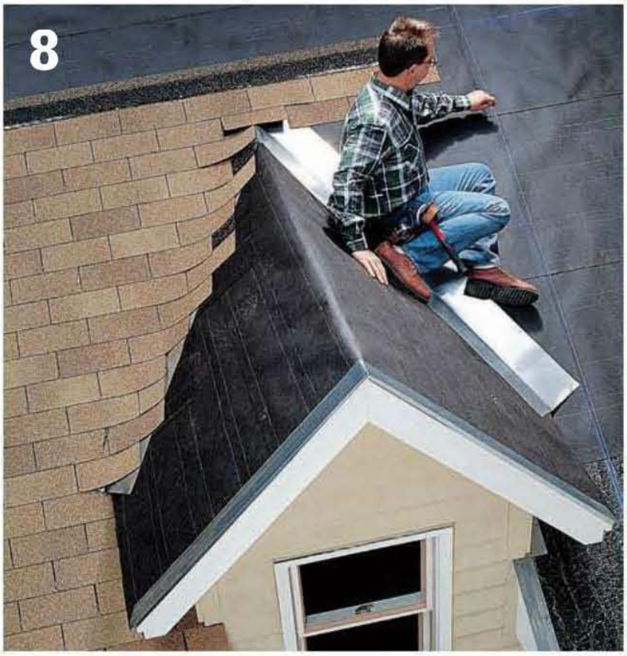

When you reach obstructions, such as dormers, install a full course of shingles above them so you can retain your shingle offset pattern. On the unshingled side of the obstruction, snap another vertical reference line using the shingles above the obstruction as a guide.

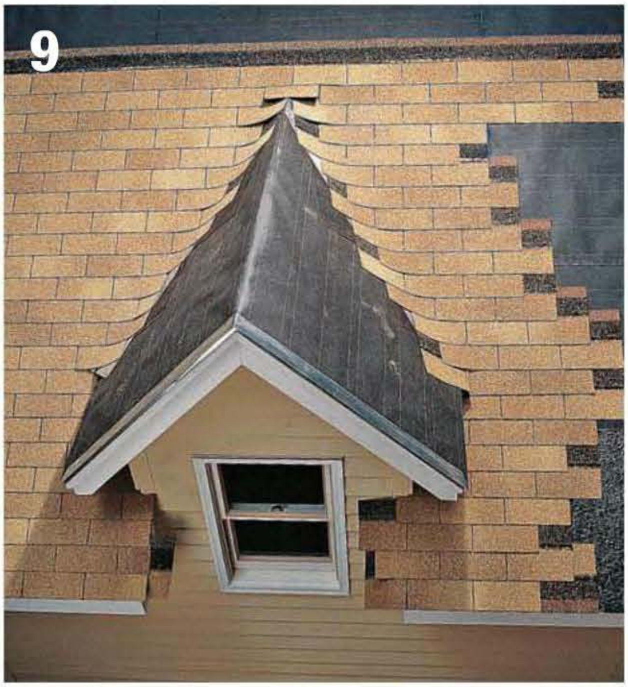

9

Shingle upward from the eaves on the unshingled side of the obstruction using the vertical line as a reference for re-establishing your shingle slot offset pattern. Fill out the shingle courses past the rake edges of the roof, then trim off the excess.

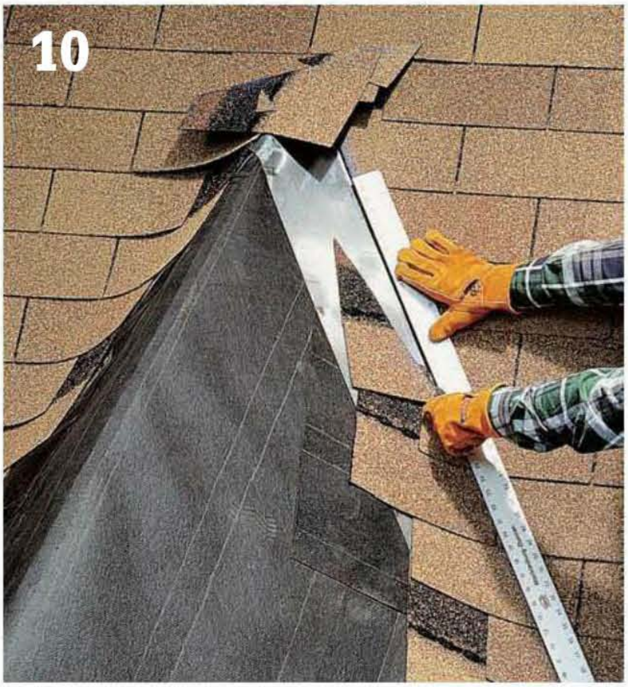

10

Trim off excess shingle material at the V in the valley flashing using a utility knife and straightedge. Do not cut into the flashing. The edges will be trimmed back farther at a slight taper after both roof decks are completely shingled.

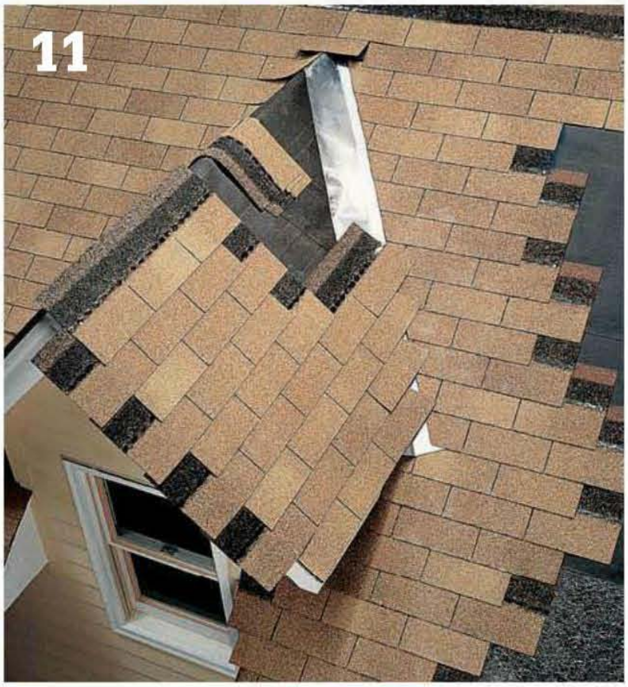

11

Install shingles on adjoining roof decks, starting at the bottom edge using the same offset alignment pattern shown in steps 1 to 6. Install shingles until courses overlap the center of the valley flashing. Trim shingles at both sides of the valley when finished.

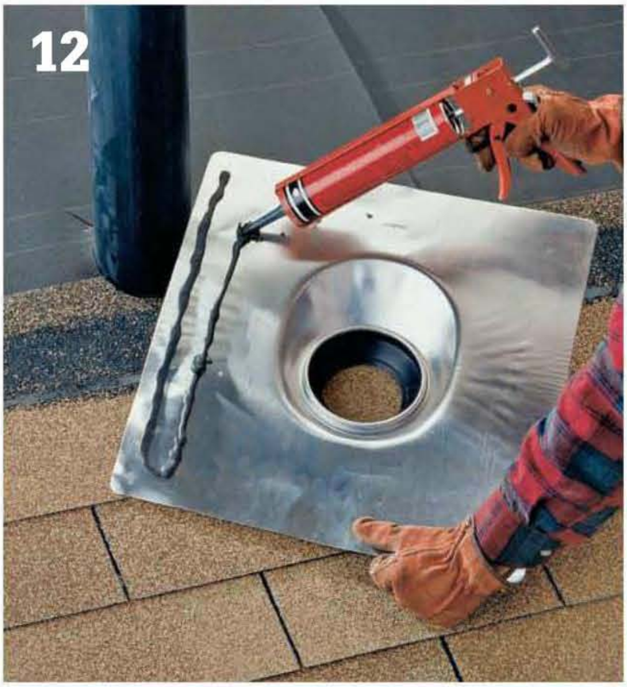

12

Install shingles up to the vent pipe so the flashing rests on at least one row of shingles. Apply a heavy double bead of roofing cement along the bottom edge of the flange.

(continued)

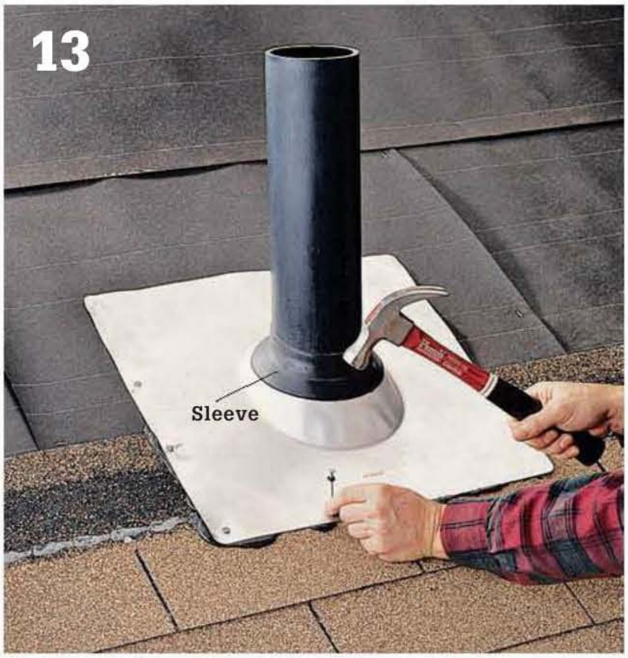

Place the flashing over the vent pipe. Position the flashing collar so the longer portion of the tapered neck slopes down the roof and the flange lies over the shingles. Nail the perimeter of the flange using rubber gasket nails.

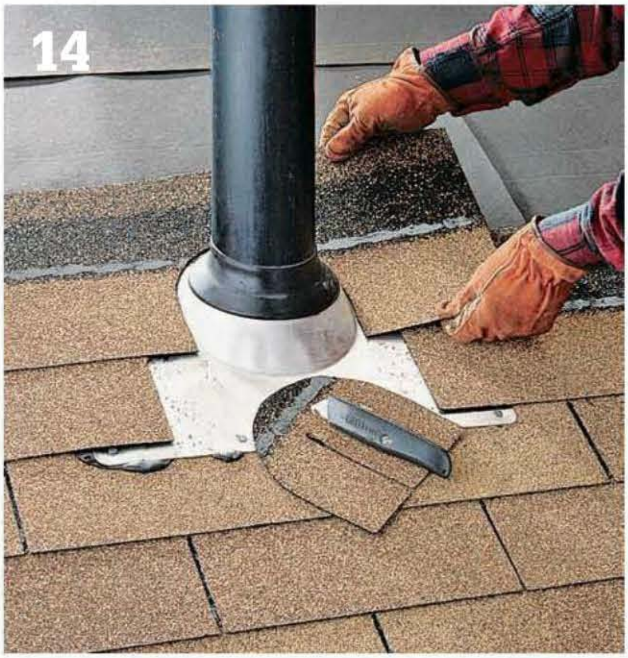

Cut shingles to fit around the neck of the flashing so they lie flat against the flange. Do not drive roofing nails through the flashing. Instead, apply roofing cement to the back of shingles where they lie over the flashing.

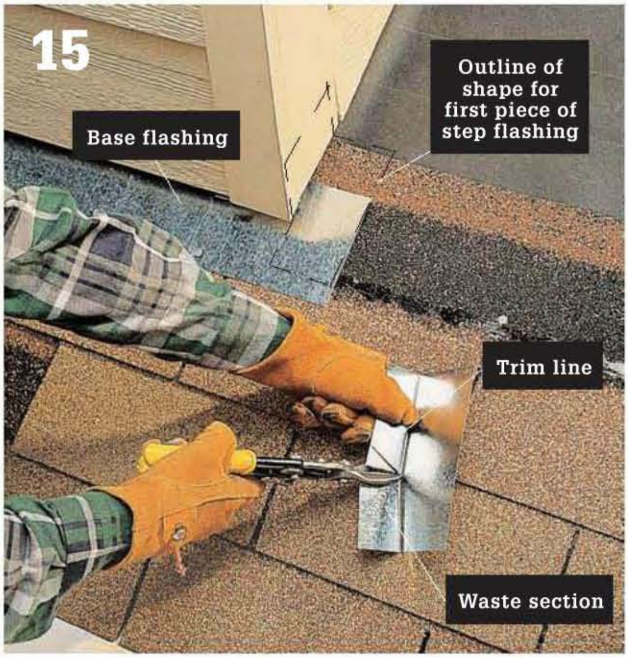

Shingle up to an element that requires flashing so the top of the reveal areas are within 5" of the element. Install base flashing using the old base flashing as a template. Bend a piece of step flashing in half and set it next to the lowest corner of the element. Mark a trim line on the flashing, following the vertical edge of the element. Cut the flashing to fit.

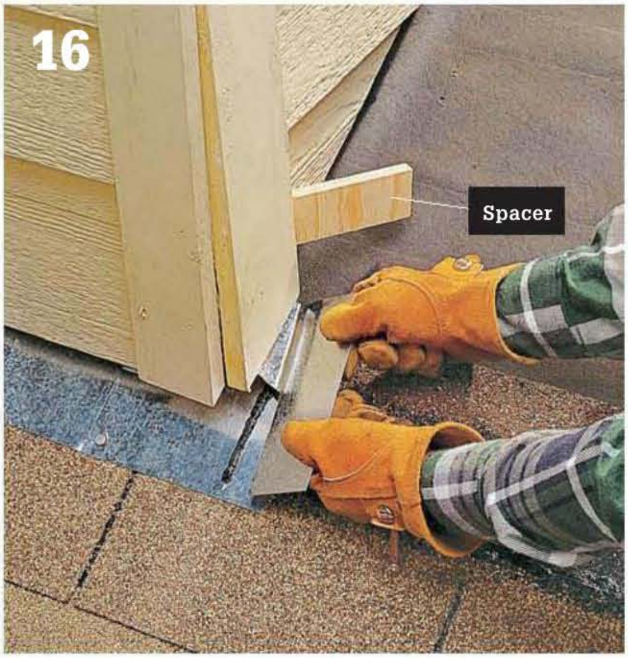

Pry out the lowest courses of siding and any trim at the base of the element. Insert spacers to prop the trim or siding away from the work area. Apply roofing cement to the base flashing in the area where the overlap with the step flashing will be formed. Tuck the trimmed piece of step flashing under the propped area, and secure the flashing. Fasten the flashing with one rubber gasket nail driven near the top and into the roof deck.

17

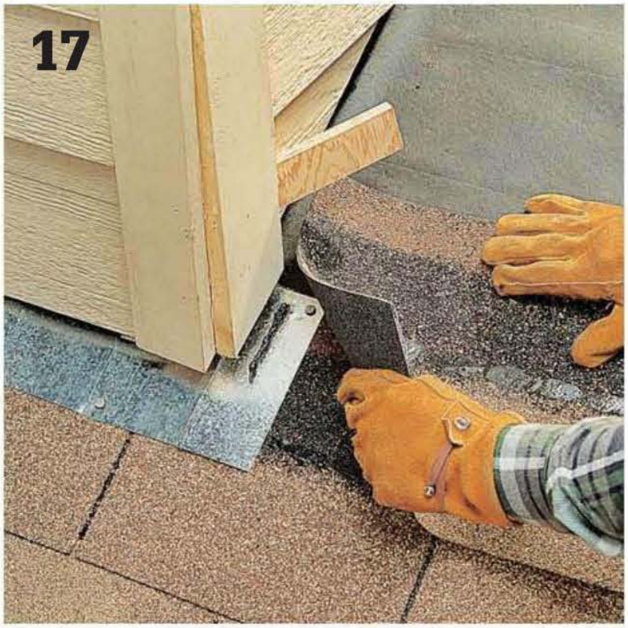

Apply roofing cement to the top side of the first piece of step flashing where it will be covered by the next shingle course. Install the shingle by pressing it firmly into the roofing cement. Do not nail through the flashing underneath.

18

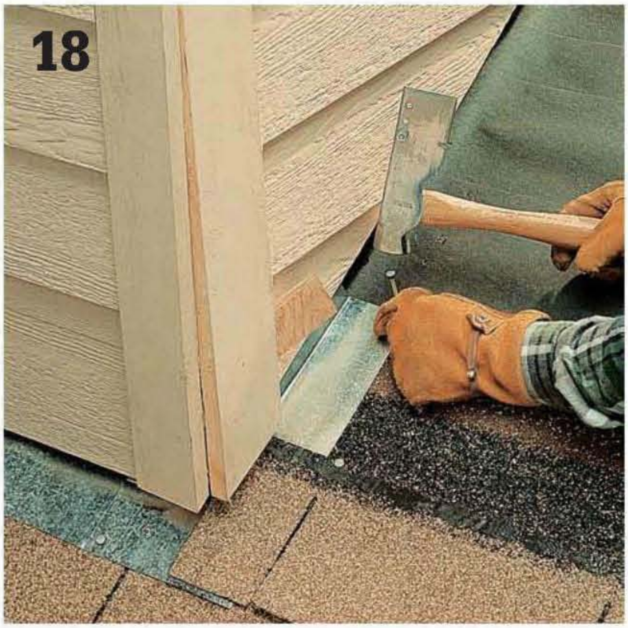

Tuck another piece of flashing under the trim or siding, overlapping the first piece of flashing at least 2". Set the flashing into roofing cement applied on the top of the shingle. Nail the shingle in place without driving nails through the flashing. Install flashing up to the top of the element the same way. Trim the last piece of flashing to fit the top corner of the element. Reattach the siding and trim.

19

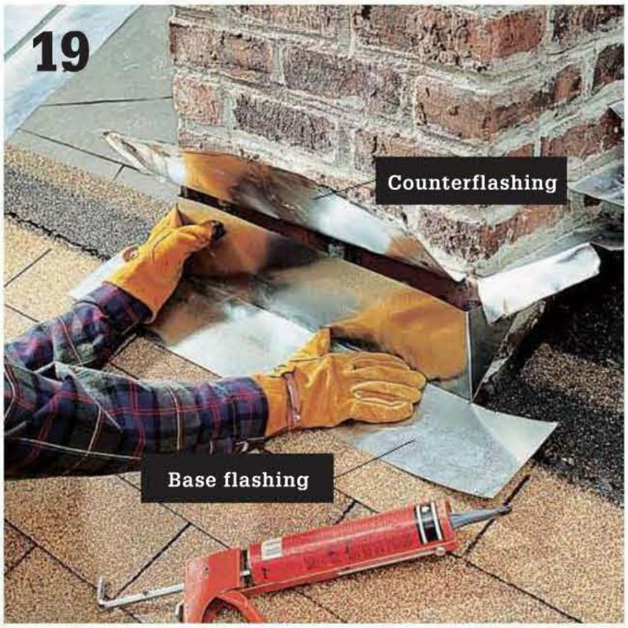

Shingle up to the chimney base. Use the old base flashing as a template to cut new flashing. Bend up the counter flashing. Apply roofing cement to the base of the chimney and the shingles just below the base. Press the base flashing into the roofing cement and bend the flashing around the edges of the chimney. Drive rubber gasket nails through the flashing flange into the roof deck.

20

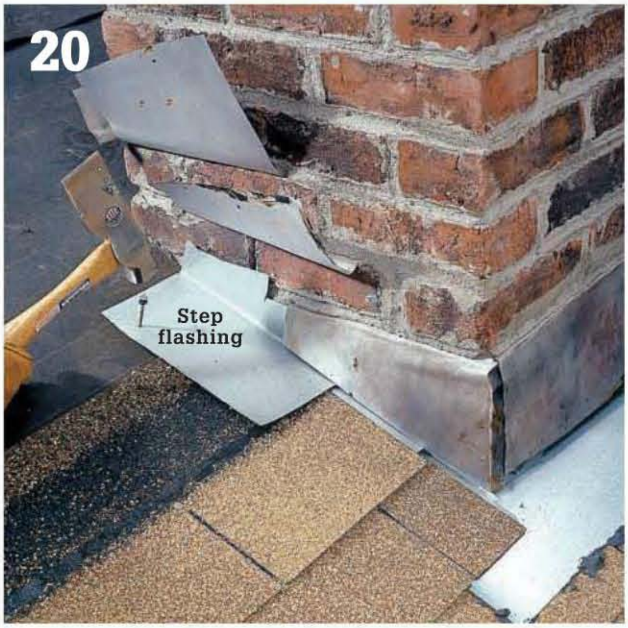

Install step flashing and shingles, working up to the high side of the chimney. Fasten flashing to the chimney with roofing cement. Fold down the counter flashing as you go.

(continued)

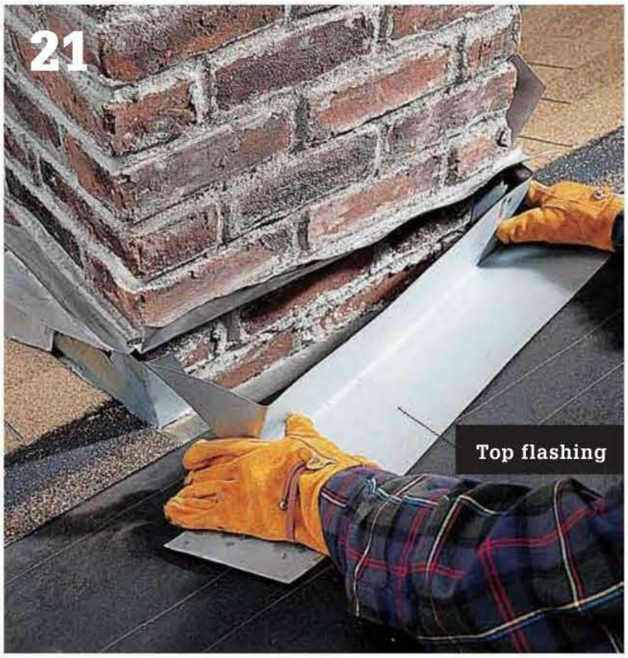

Cut and install top flashing (also called a saddle) around the high side of the chimney. Overlap the final piece of flashing along each side. Attach the flashing with roofing cement applied to the deck and chimney and with rubber gasket nails driven through the flashing base into the roof deck. Shingle past the chimney using roofing cement (not nails) to attach shingles over the flashing.

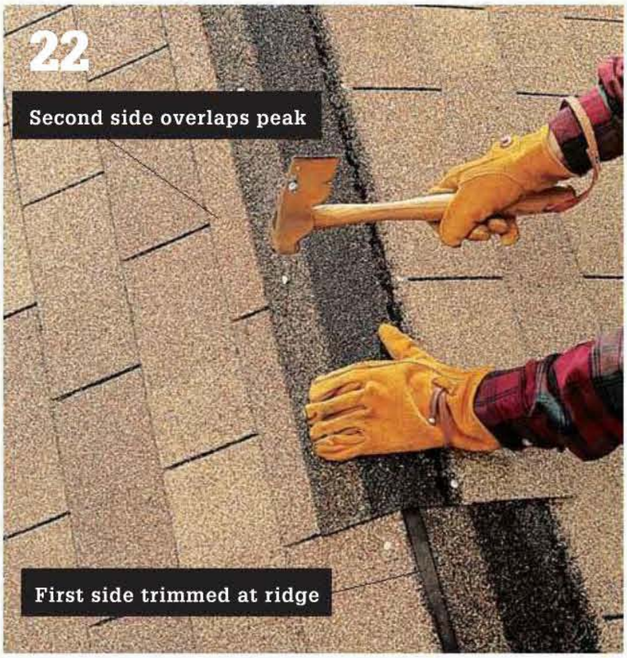

When you reach a hip or ridge, shingle up the first side until the top of the uppermost reveal area is within 5" of the hip or ridge. Trim the shingles along the peak. Install shingles on the opposite side of the hip or ridge. Overlap the peak no more than 5".

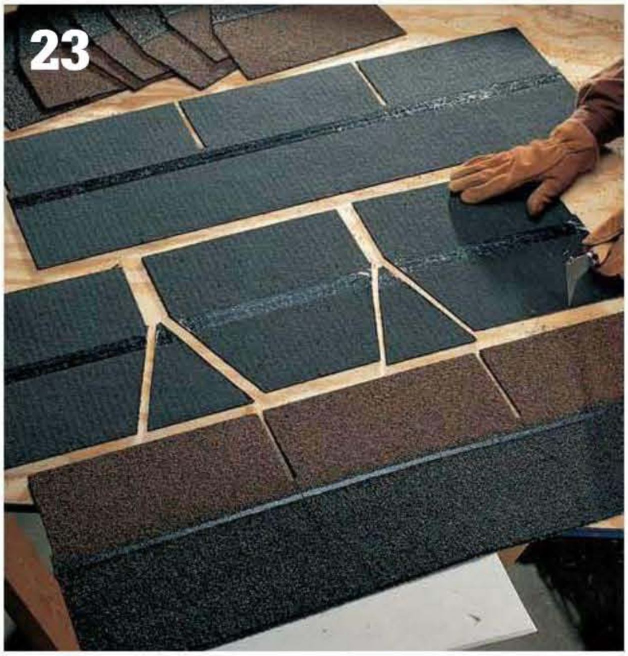

Cut three 12"-sq. cap shingles from each three-tab shingle. With the back surface facing up, cut the shingles at the tab lines. Trim the top corners of each square with an angled cut, starting just below the seal strip to avoid overlaps in the reveal area.

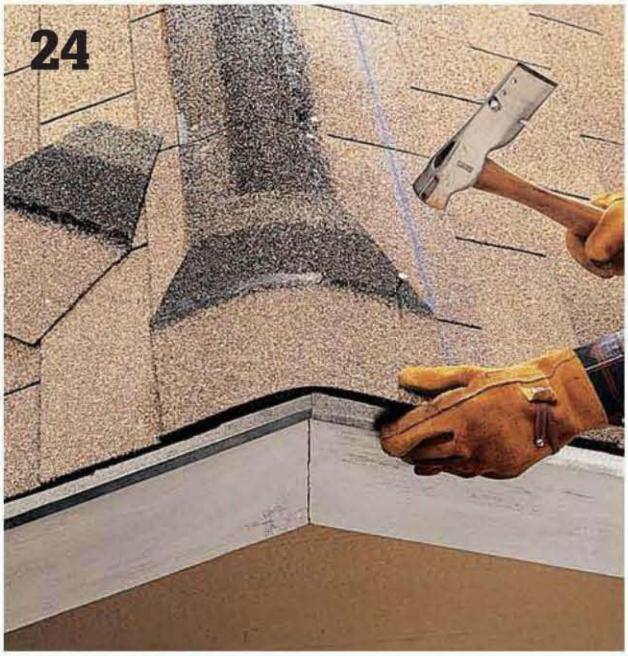

Snap a chalk line 6" down from the ridge, parallel to the peak. Attach cap shingles, starting at one end of the ridge, aligned with the chalk line. Drive two 1¼" roofing nails per cap about 1" from each edge, just below the seal strip.

25

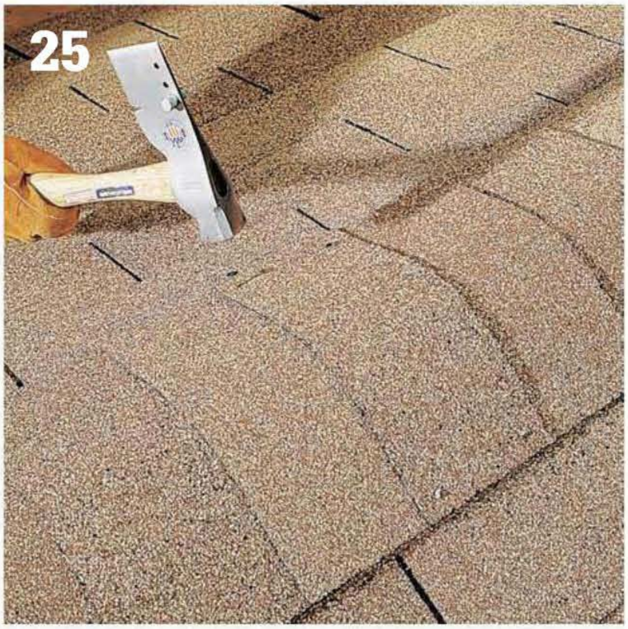

Following the chalk line, install cap shingles halfway along the ridge, creating a 5" reveal for each cap. Then, starting at the opposite end, install caps over the other half of the ridge to meet the first run in the center. Cut a 5"-wide section from the reveal area of a shingle tab, and use it as a "closure cap" to cover the joint where the caps meet.

26

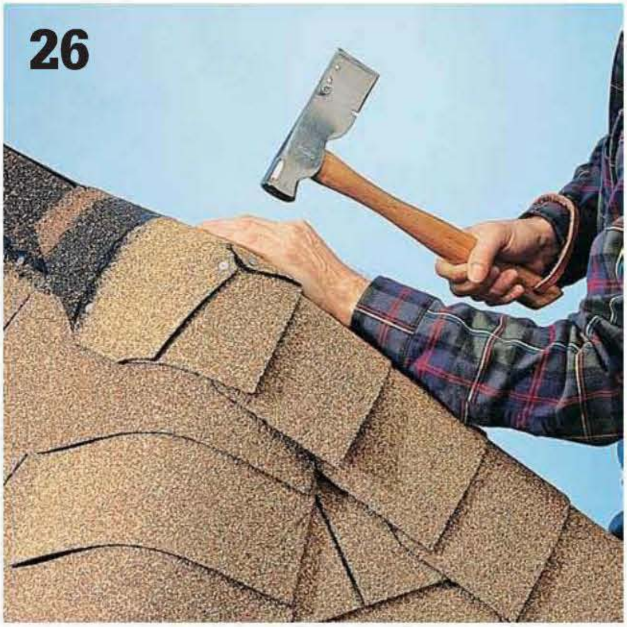

Shingle the hips in the same manner using a chalk reference line and cap shingles. Start at the bottom of each hip and work to the peak. Where hips join with roof ridges, install a custom shingle cut from the center of a cap shingle. Set the cap at the end of the ridge and bend the corners so they fit over the hips. Secure each corner with a roofing nail, and cover the nail heads with roofing cement.

27

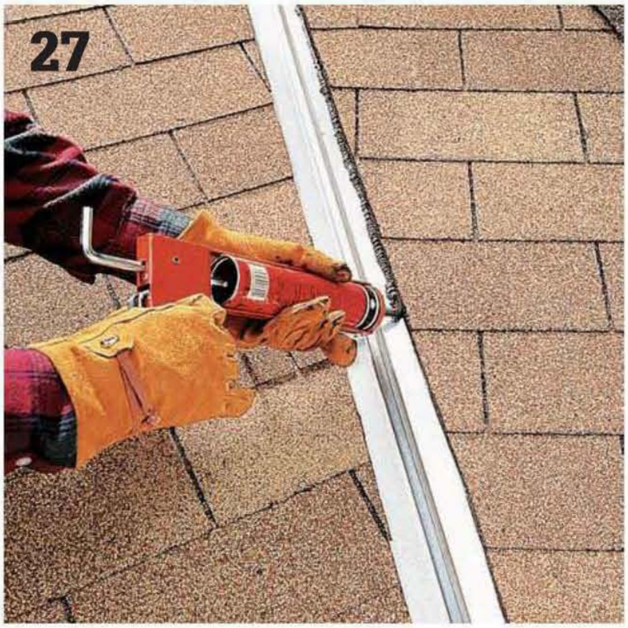

After all shingles are installed, trim them at the valleys to create a gap that's 3" wide at the top and widens at a rate of ⅛" per foot as it moves downward. Use a utility knife and straightedge to cut the shingles, making sure not to cut through the valley flashing. At the valleys, seal the undersides and edges of shingles with roofing cement. Also cover exposed nail heads with roofing cement.

28

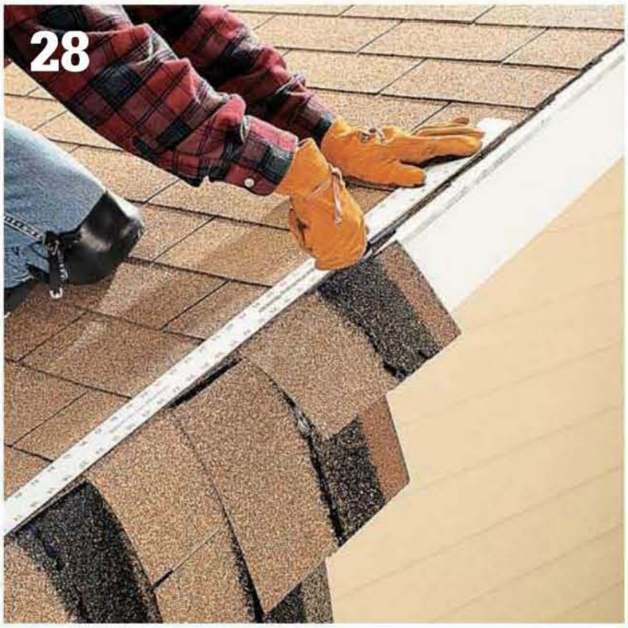

Mark and trim the shingles at the rake edges of the roof. Snap a chalk line ⅜" from the edge to make an overhang, then trim the shingles.

Installing Vinyl Siding

Vinyl has become one of the most popular sidings due to its low cost, uniform appearance, and maintenance-free durability. Installation is fairly simple, with each row locking onto the lip of the underlying course, then nailed along the top.

There are a couple of key factors that will make or break your siding project. First, the sheathing must be straight and solid before the siding is applied. The siding will only look as straight and smooth as the wall it's on. Second, determine how the siding should overlap to hide the seams from the main traffic patterns. This usually means starting in the back and working toward the front of the house.

Do not nail the siding tight to the house. The panels need to slide back and forth as they expand and contract with changes in the temperature. If the siding can't move, it will bow and need to be reinstalled. Keep a 1⁄32" gap between the head of the nail and the siding.

Vinyl siding is available in a wide variety of colors and styles, and with a lot of accessories such as trim, fluted lineals, vertical columns, crown molding, and band boards. The most common vinyl siding is horizontal lap siding, which is shown starting on page 548. This project shows a foam underlayment, which reduces outside noise, protects the siding from dents, and adds an insulation value. Vertical vinyl siding is also available. It's specifically made for vertical applications.

Tools & Materials ▸

- Hammer
- Circular saw
  or radial-arm saw
- Clamps
- Tape measure
- String
- Straightedge
- Utility knife
- Aviation snips
- Level
- Chalk line
- Framing square
- Nail slot punch
- Zip-lock tool
- Snap lock punch
- Caulk gun
- Vinyl siding
- J-channel
- Corner posts
- Undersill
- Starter strip
- Nails
- Cutting table
- Safety glasses
- Silicone caulk

Vinyl siding can look very similar to wood lap siding, but it doesn't require regular upkeep. Vinyl can be installed on any type and style of house.

Cutting Vinyl Siding

Use a fine-tooth blade installed backward in the saw to cut vinyl siding. Use a radial-arm saw, power miter saw, or a circular saw, and move the blade slowly through the siding. Always wear safety glasses when cutting siding.

Support the siding on a cutting table when cutting. Vinyl siding is too flimsy to be placed across sawhorses without support. You can build a cutting table by fastening a long piece of scrap plywood between two sawhorses.

Using Specialty Tools

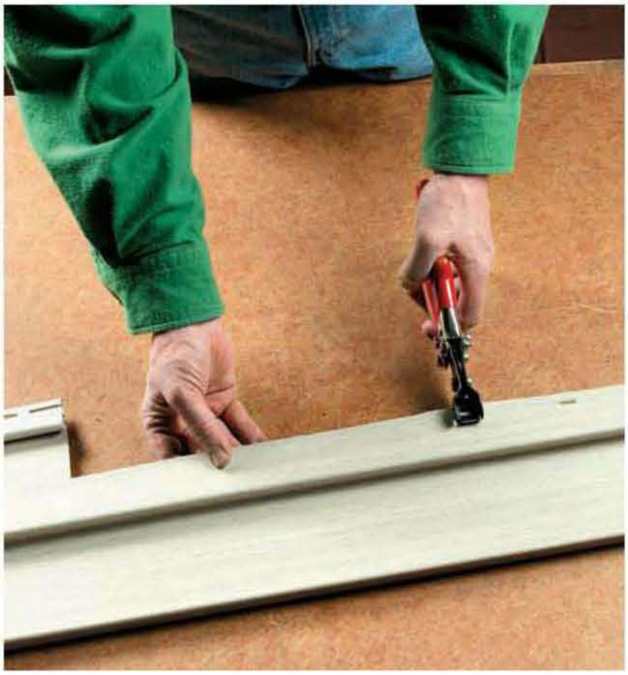

A snap lock punch is used to make raised tabs, or dimples, in a cut edge of siding where the nailing hem has been removed. This eliminates the need to facenail the panel.

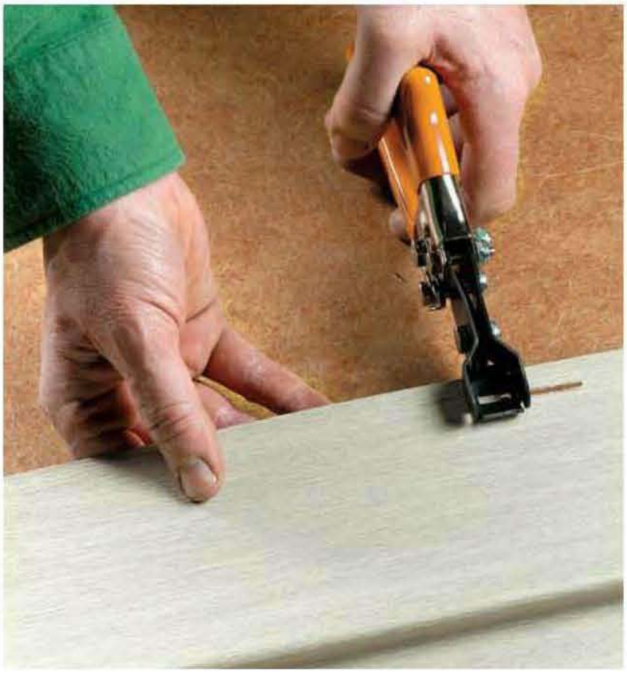

A nail slot punch is used to make horizontal nail slots in the face of panels. It can also be used to add or elongate the opening of an existing nail slot to match irregular stud spacing.

How to Install Vinyl Siding

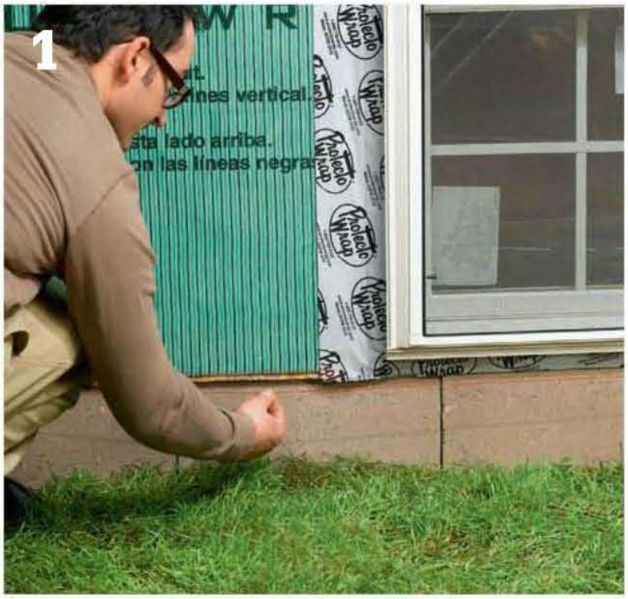

Install housewrap following manufacturer instructions. Identify the lowest corner of the house that has sheathing, and partially drive a nail 1½" above the bottom edge of the sheathing. Run a level string to the opposite corner of the wall and partially drive a nail. Do this around the entire house. Snap chalk lines between the nails.

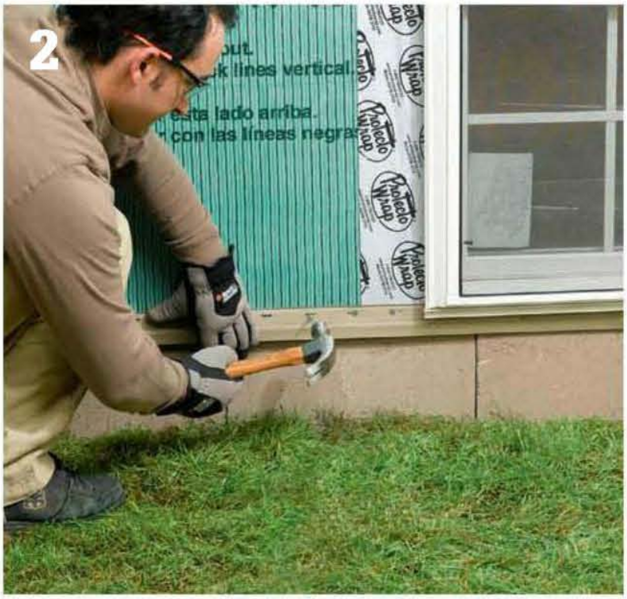

Place the top edge of the starter strip along the chalk line and nail every 10". Nail in the center of the slots and don't nail tight to the house. Keep a ¼" gap between strips, and leave space at the corners for a ½" gap between starter strips and corner posts.

Option: Install foam vinyl siding underlayment on the house using cap nails. Align the bottom of the underlayment with the starting strip. To cut panels to size, score them with a utility knife, then break them over your cutting table. Some panels need to be taped at the seams. Follow manufacturer's recommendations.

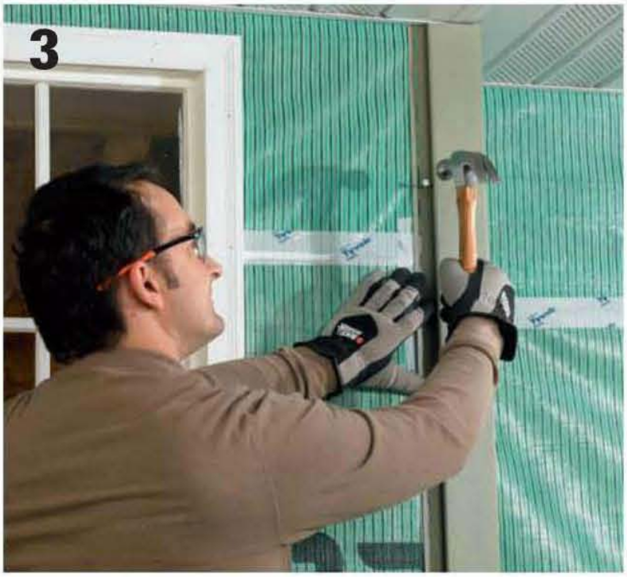

Install a corner post, keeping a ¼" gap between the top of the post and the soffit. Extend the bottom of the post ¼" below the bottom of the starter strip. Drive a nail at the top end of the uppermost slot on each side of the post (the post hangs from these nails). Make sure the post is plumb on both sides using a level. Secure the post by driving nails every 8 to 12" in the center of the slots. Do not nail the post tight. Install the other posts the same way.

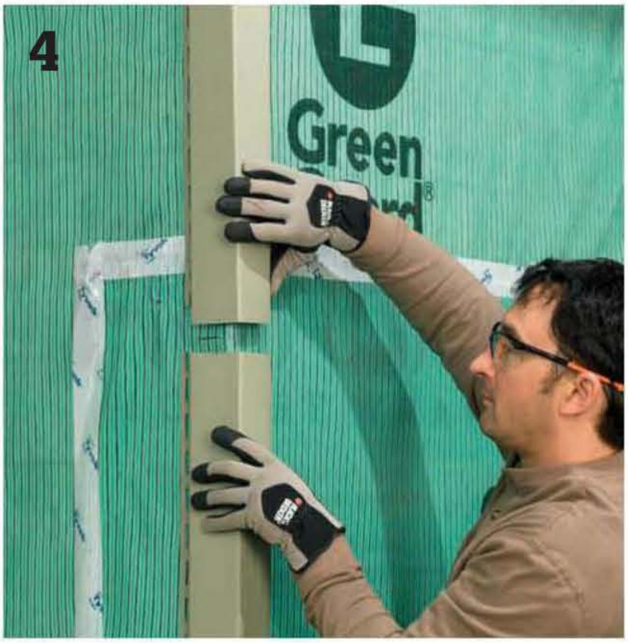

4

If more than one corner post is needed to span the length of a corner, the upper post overlaps the lower post. For an outside corner post, cut off 1" from the nailing flanges on the bottom edge of the top post. For an inside corner post, cut off 1" from the nailing flange on the upper edge of the bottom post. Overlap the posts by ¾", leaving ¼" for expansion.

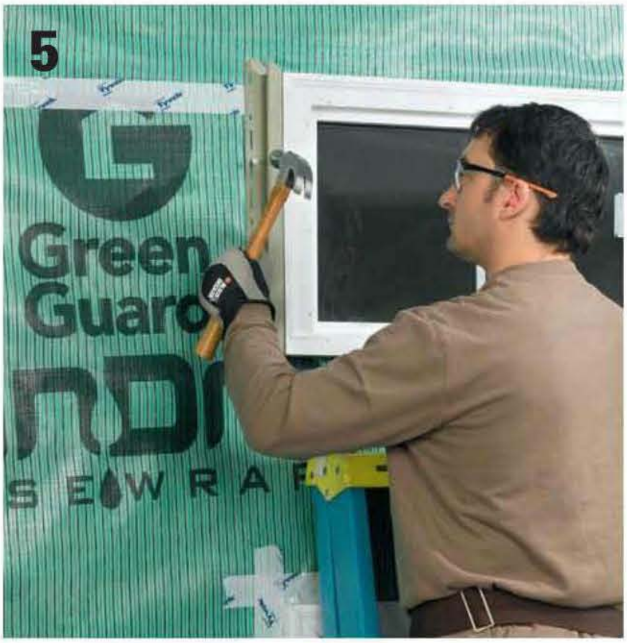

5

Measure and cut two J-channels that are the length of a window plus the width of the J-channel. Place one of the J-channels against the side of the window, aligning the bottom edge with the bottom edge of the window. Nail the channel in place. Nail the second J-channel against the opposite side of the window the same way.

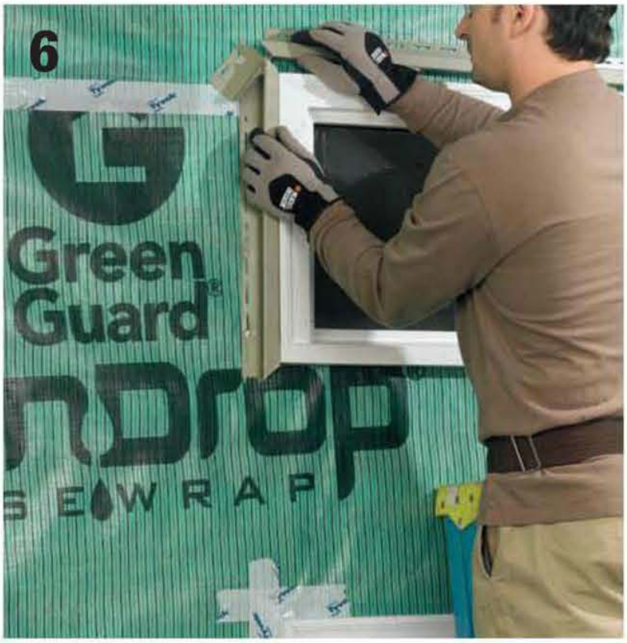

6

At the top of the window, measure between the outside edges of the side J-channels and cut a piece of J-channel to fit. Cut a ¾" tab at each end. Bend the tabs down to form a drip edge. Miter cut the face at each end at 45°. Center the J-channel over the window and nail it in place. The top J-channel overlaps the side pieces, and the drip edges fit inside the side pieces. Do this for each window and door.

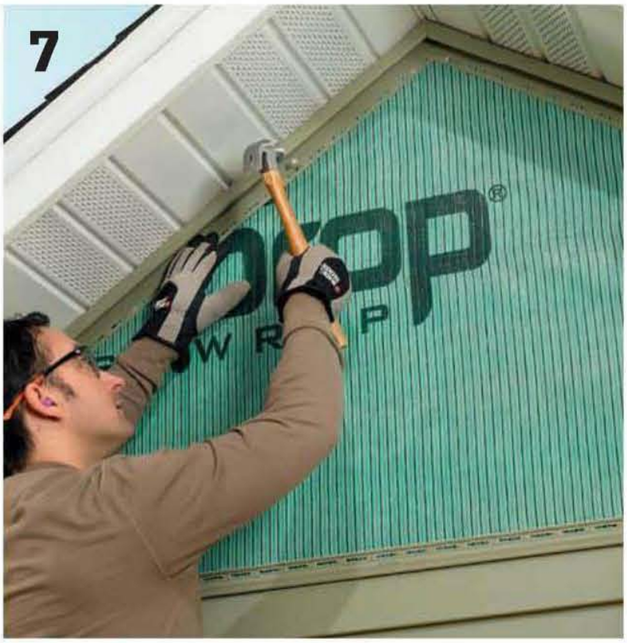

7

Measure, cut, and install J-channel along the gable ends. Nail the channels every 8 to 12". To overlap J-channels, cut 1" from the nailing hem. Overlap the channels ¾", leaving ¼" for expansion. At the gable peak, cut one channel at an angle to butt against the peak. Miter the channel on the opposite side to overlap the first channel.

(continued)

8

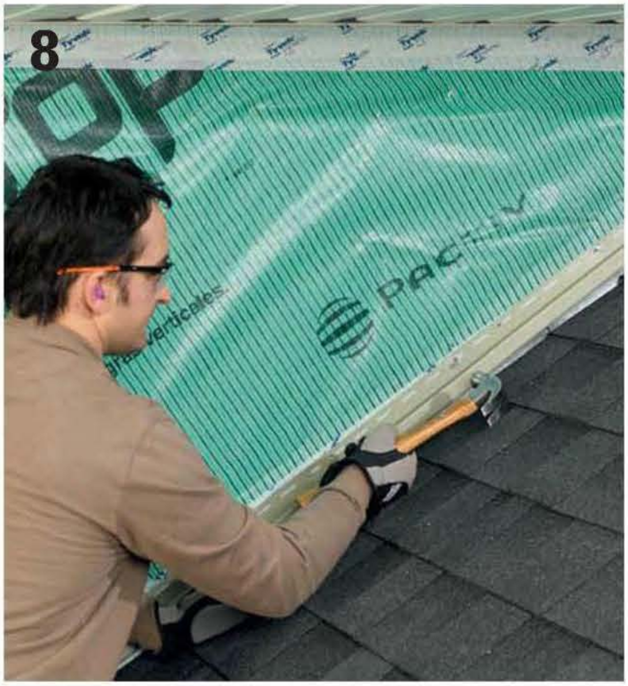

To install J-channel over a roof line, snap a chalk line along the roof flashing ½" above the roof. Align the bottom edge of the J-channel along the chalk line, and nail the channel in place. Make sure the channel does not make direct contact with the shingles.

9

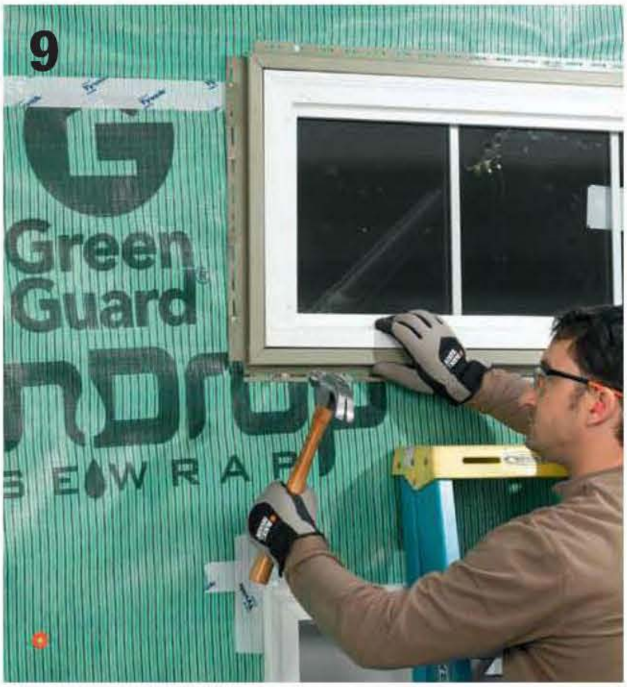

Measure, cut, and install undersill beneath each window. The undersill should be flush with the outside lip of the side channels.

10

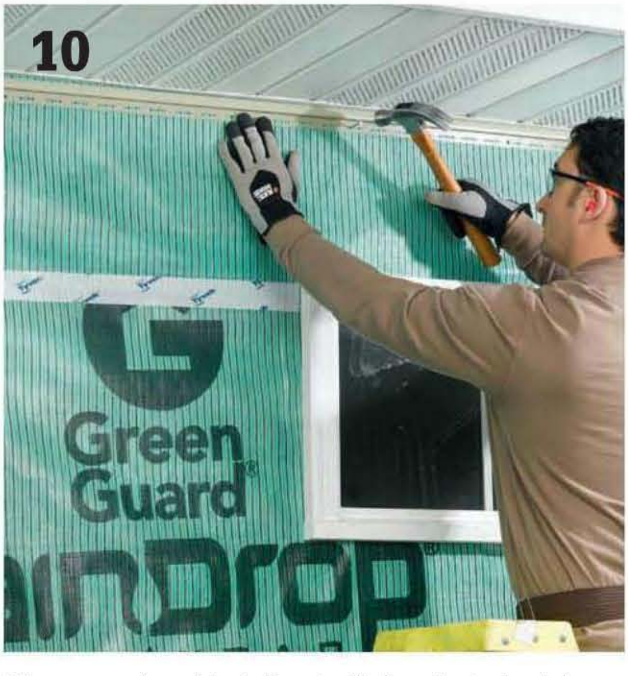

Measure, cut, and install undersill along the horizontal eaves on the house. If more than one undersill is needed, cut the nailing hem 1¼" from the end of one undersill. Overlap the undersills by 1".

11

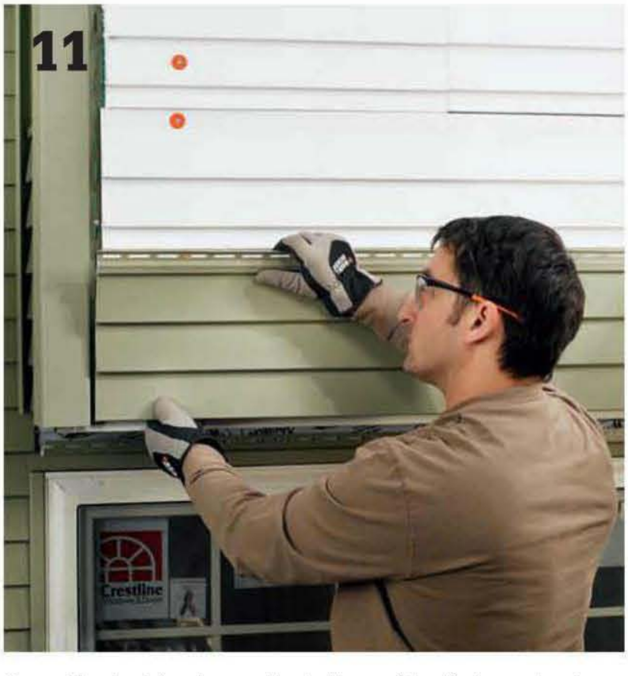

Snap the locking leg on the bottom of the first panel onto the starter strip, making sure it's securely locked in place. Keep a ¼" gap between the end of the panel and the corner post. Nail the panel a minimum of every 16" on center. Don't drive the nails tight. *Note: This installation shows a vinyl siding underlayment in place.*

12

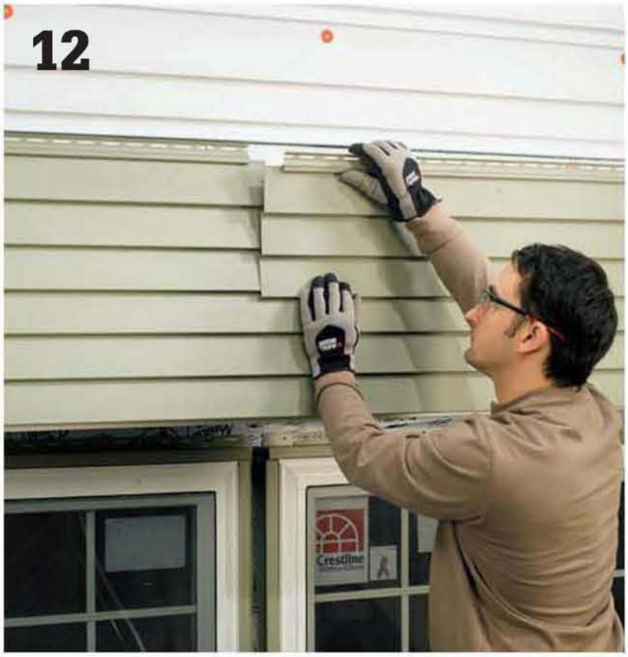

Overlap panels by 1". Cut panels so the factory cut edge is the one that's visible. Keep nails at least 6" from the end of panels to allow for smooth overlap. Do not overlap panels directly under a window.

13

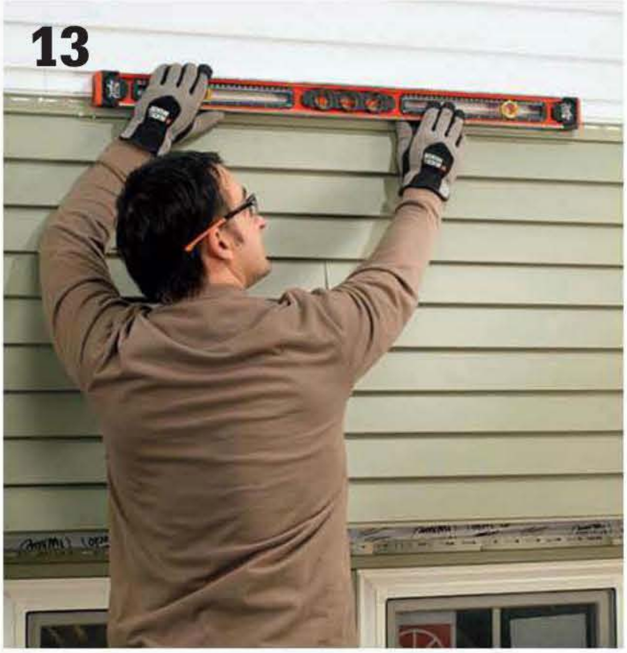

Place the second row over the first, snapping the locking leg into the lock of the underlying panels. Leave ¼" gap at corners and J-channels. Install subsequent rows, staggering seams at least 24" unless separated by more than three rows. Check every several rows for level. Make adjustments in slight increments, if necessary.

14

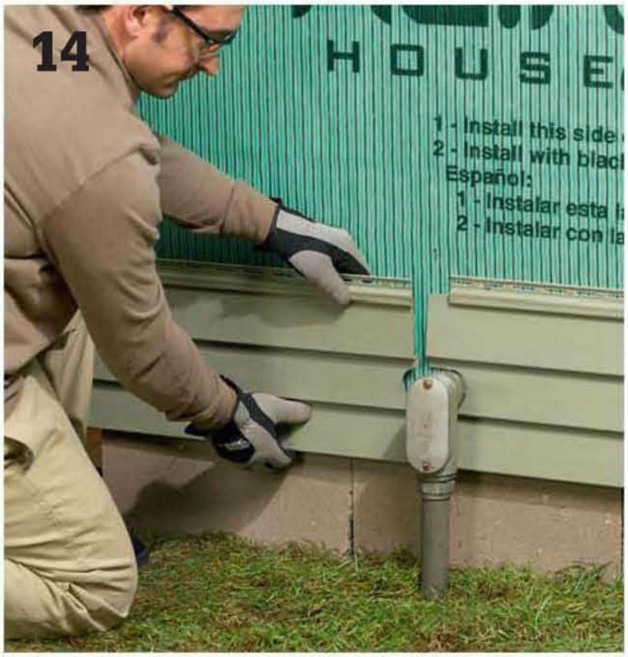

For hose spigots, pipes, and other protrusions, create a seam at the obstacle. Begin with a new panel to avoid extra seams. Cut an opening ¼" larger than the obstacle, planning for a 1" overlap of siding. Match the shape and contour as closely as possible. Fit the panels together around the obstruction and nail in place.

15

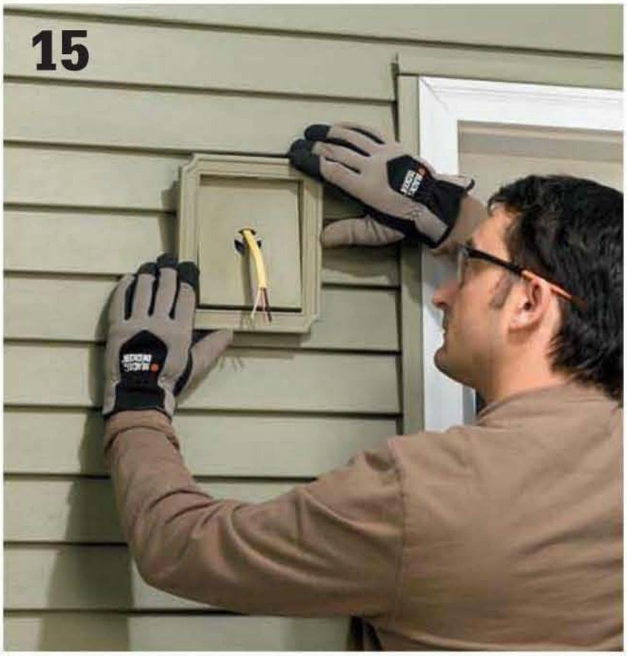

Place mounting blocks around outlets, lights, and doorbells. Assemble the base around the fixture, making sure it's level, and nail in place. Install siding panels, cutting them to fit around the mounting block with a ¼" gap on each side. Fasten the cover by snapping it over the block.

(continued)

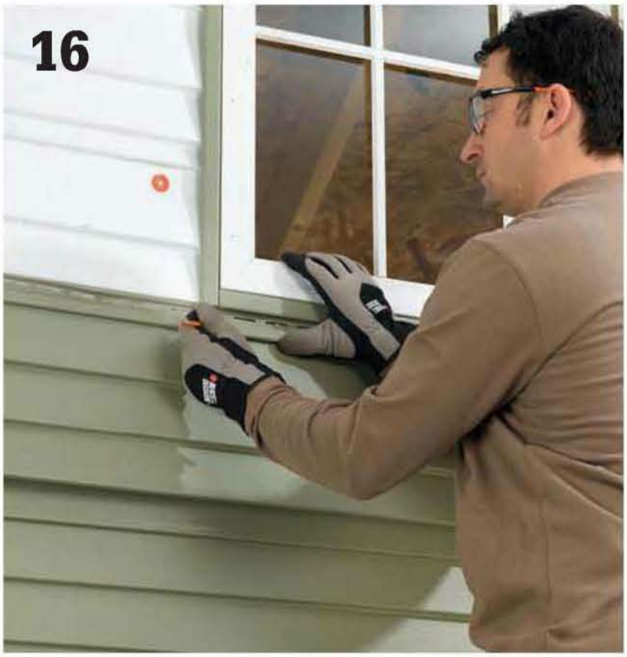

Where panels must be notched to fit below a window, position the panel below the window and mark the edges of the window, allowing for a ¼" gap. Place a scrap piece of siding alongside the window and mark the depth of the notch, keeping a ¼" gap. Transfer the measurement to the panel, mark the notch, and cut it out. Create tabs on the outside face every 6" using a snap lock punch. Install the panel, locking the tabs into the undersill.

Install cut panels between windows and between windows and corners as you would regular panels. Avoid overlapping panels and creating seams in small spaces. The panels need to align with panels on the opposite side of the window.

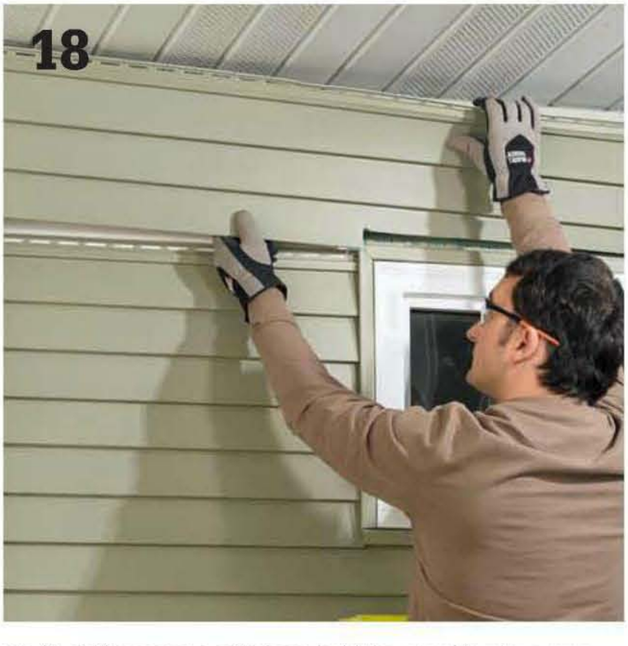

To fit siding over a window, hold the panel in place over the window and mark it. Use a scrap piece of siding to mark the depth of the cut. Transfer the measurement to the full panel and cut the opening. Fit the cut edge into the J-channel above the window, lock the panel in place, and nail it.

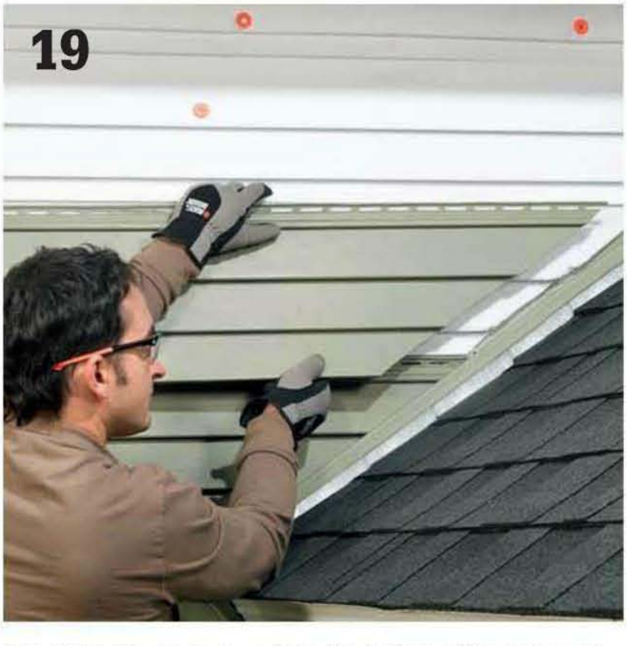

For dormers, measure up from the bottom of the J-channel the height of a panel and make a mark. Measure across to the opposite J-channel. Use this measurement to mark and cut the panel to size. Cut and install panels for the rest of the dormer the same way.

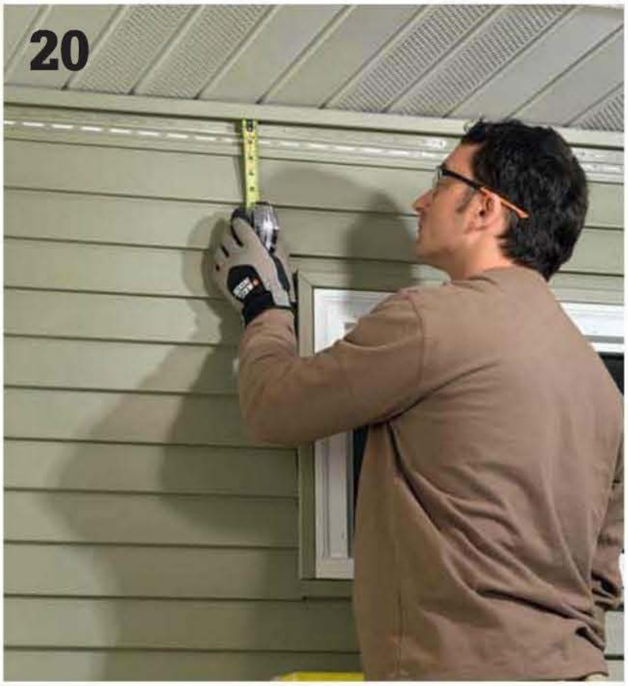

Measure the distance between the lock on the last fully installed panel and the top of the undersill under the horizontal eaves. Subtract ¼", then mark and rip a panel to fit. Use a snap lock punch to punch tabs on the outside face every 6". Install the panel, locking the tabs into the undersill.

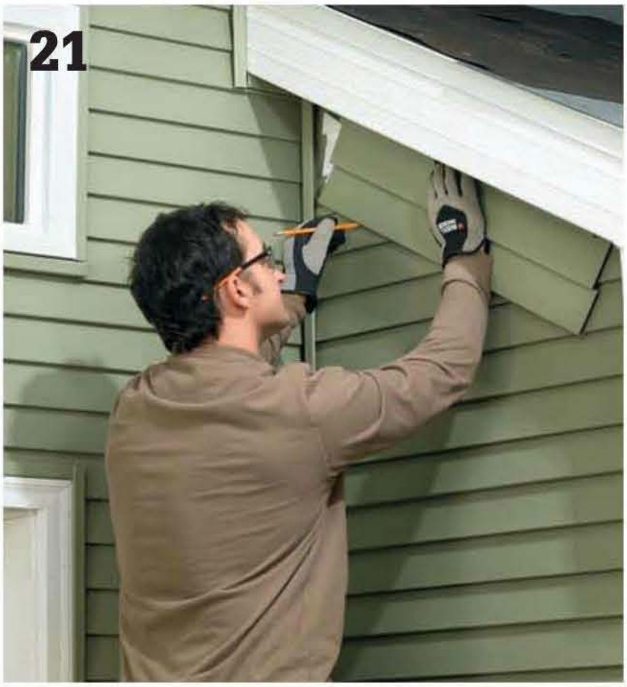

Place a scrap panel in the J-channel along the gable end of the house. Place another scrap over the last row of panels before the gable starts, slide it under the first scrap, and mark the angle where they intersect. Transfer this angle to full panels. Make a similar template for the other side. Cut the panels and set the cut edge into the J-channel, leaving a ¼" gap.

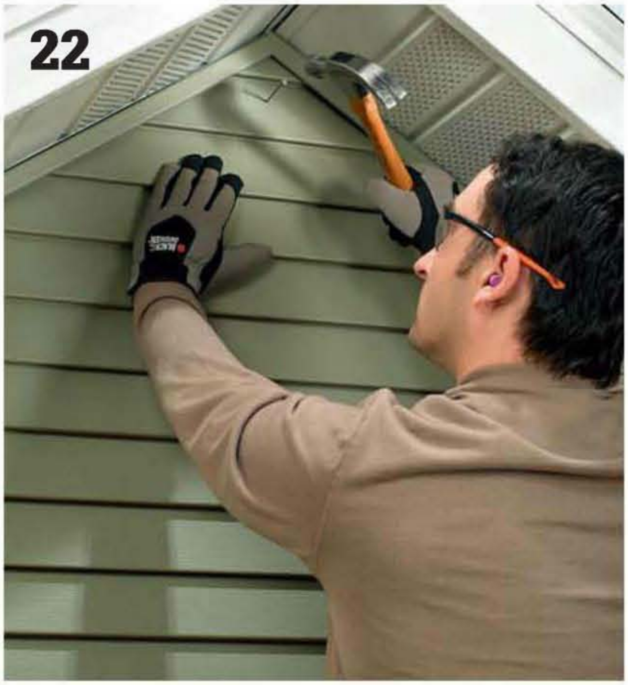

Cut the last piece of siding to fit the gable peak. Drive a single aluminum or stainless steel finish nail through the top of the panel to hold it in place. This is the only place where you will facenail the siding.

Apply caulk between all windows and J-channel, and between doors and J-channel.

Resources

American Institute of Architects
800-364-9364
www.aiaonline.com

American Lighting Association
800-724-4484
www.americanlightingassoc.com

American Society of Interior Designers
202-546-3480
www.asid.org

Association of Home Appliance Manufacturers
202-872-5955
www.aham.org

Certified Forest Products Council
503-224-2205
www.certifiedwood.org

Construction Materials Recycling Association
630-548-4510
www.cdrecycling.org

Energy & Environmental Building Association
952-881-1098
www.eeba.org

International Residential Code (book)
International Conference of Building Officials
800-284-4406
www.icbo.com

Kohler Co.
800-456-4537
www.kohler.com

Laticrete International
(radiant floor mats)
800-243-4788
www.laticrete.com

Light Gauge Steel Engineers Association
615-279-9251
www.lgsea.com

National Association of the Remodeling Industry (NARI)
847-298-9200
www.nari.org

National Fire Protection Agency
617-770-3000
www.nfpa.org

National Kitchen & Bath Association (NKBA)
800-843-6522
www.nkba.org

National Wood Flooring Association
800-422-4556
www.woodfloors.org

North American Insulation Manufacturers Association
703-684-0084
www.naima.org

Tile Council of America
864-646-8453
www.tileusa.com

U.S. Environmental Protection Agency—Indoor Air Quality
www.epa.gov/iaq/

Photo Credits

Alcoa Home Exteriors, Inc.
p. 531, 546

American Standard
p. 438

Armstrong World Industries
p. 174, 180 (left), 236

Brian Greer's Tin-Ceilings
519 570 1447
p. 239

Broan Nutone
p. 327

Ceramic Tiles of Italy
p. 454

CertainTeed, Corp.
p. 530 (lower)

Cherry Tree Design
p. 428 (lower)

Daltile
p. 386

Economic Mobility, Inc.
p. 425 (lower left)

Elkay
p. 348, 426 (left)

The Energy Conservatory
www.energyconservatory.com
p. 218 (right) Energy Star Program

Fotolia / www.fotolia.com
p. 480

General Electric
p. 418 (right)

Ginger
p. 426 (right)

Harrell Remodeling, Inc.
p. 427, 486

IKEA Home Furnishings
p. 416

ISTOCK Photo / www.istock.com,
p. 324 George Peters, 420, 422, 468 Nicola Gavin, 528

Jacuzzi
p. 428 (top)

Kentucky Wood Floors
p. 170

Kohler
p. 19, 392, 424, 425 (lower right), 446, 484

Kraftmaid
p. 339

Marvin Windows
p. 318

MIRAGE Prefinished Hardwood Floors
p. 162

Oasis Montana
p. 418 (left)

Robert Perron, photographer / www.bobperron.com
p. 516 (left & right)

Price Pfister
p. 393 (all), 472 (top & lower)

Seattle Glass Block
p. 220

SieMatic
p. 336

Swanstone
p. 470 (top)

Toto
p. 425 (top)

Urban Homes, NY
p. 340

Brian Vanden Brink, photographer / www.brianvandenbrink.com
p. 492

VELUX-America
p. 278

Weather Shield
p. 18 (top)

Western Red Cedar Lumbar Association
p. 228

Conversion Charts

Metric Equivalent

Inches (in.)	1/64	1/32	1/25	1/16	1/8	1/4	3/8	2/5	1/2	5/8	3/4	7/8	1	2	3	4	5	6	7	8	9	10	11	12	36	39.4
Feet (ft.)																								1	3	3 1/12
Yards (yd.)																									1	1 1/12
Millimeters (mm)	0.40	0.79	1	1.59	3.18	6.35	9.53	10	12.7	15.9	19.1	22.2	25.4	50.8	76.2	101.6	127	152	178	203	229	254	279	305	914	1,000
Centimeters (cm)							0.95	1	1.27	1.59	1.91	2.22	2.54	5.08	7.62	10.16	12.7	15.2	17.8	20.3	22.9	25.4	27.9	30.5	91.4	100
Meters (m)																								.30	.91	1.00

Converting Measurements

To Convert:	To:	Multiply by:
Inches	Millimeters	25.4
Inches	Centimeters	2.54
Feet	Meters	0.305
Yards	Meters	0.914
Miles	Kilometers	1.609
Square inches	Square centimeters	6.45
Square feet	Square meters	0.093
Square yards	Square meters	0.836
Cubic inches	Cubic centimeters	16.4
Cubic feet	Cubic meters	0.0283
Cubic yards	Cubic meters	0.765
Pints (U.S.)	Liters	0.473 (Imp. 0.568)
Quarts (U.S.)	Liters	0.946 (Imp. 1.136)
Gallons (U.S.)	Liters	3.785 (Imp. 4.546)
Ounces	Grams	28.4
Pounds	Kilograms	0.454
Tons	Metric tons	0.907

To Convert:	To:	Multiply by:
Millimeters	Inches	0.039
Centimeters	Inches	0.394
Meters	Feet	3.28
Meters	Yards	1.09
Kilometers	Miles	0.621
Square centimeters	Square inches	0.155
Square meters	Square feet	10.8
Square meters	Square yards	1.2
Cubic centimeters	Cubic inches	0.061
Cubic meters	Cubic feet	35.3
Cubic meters	Cubic yards	1.31
Liters	Pints (U.S.)	2.114 (Imp. 1.76)
Liters	Quarts (U.S.)	1.057 (Imp. 0.88)
Liters	Gallons (U.S.)	0.264 (Imp. 0.22)
Grams	Ounces	0.035
Kilograms	Pounds	2.2
Metric tons	Tons	1.1

Converting Temperatures

Convert degrees Fahrenheit (F) to degrees Celsius (C) by following this simple formula: Subtract 32 from the Fahrenheit temperature reading. Then mulitply that number by $5/9$. For example, 77°F - 32 = 45. 45 × $5/9$ = 25°C.

To convert degrees Celsius to degrees Fahrenheit, multiply the Celsius temperature reading by $9/5$, then add 32. For example, 25°C × $9/5$ = 45. 45 + 32 = 77°F.

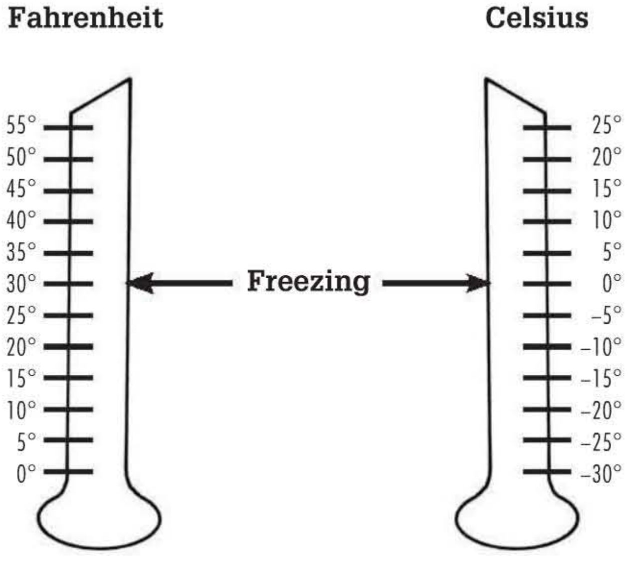

Contributing Editors, Art Directors, Set Builders, and Photographers

Phil Aarrestad
Dan Anderson
Glenn Austin
Randy Austin
Kim Bailey
Mark Biscan
Stewart Block
Charlie Boldt
Scott Boyd
Rose Brandt
Gary Branson
Greg Breining
Timothy Bro
Dave Brus
Keith Bruzelius
Ron Bygness
Peter Caley
Jennifer Caliandro
Rudy Calin
Tom Carpenter
Julie Caruso
Dan Cary
Janice Cauley
Marcia Chambers
Scott Christiansen
Tom Cooper
Paul Currie
Cy DeCosse
Jim Destiche
Doug Deutsche
Diane Dreon
Arthur Durkee
Barbara Falk
Jerri Farris
Lyle Ferguson
Mary Firestone
Steve Galvin
Kim Gerber
Abby Gnagey
Patricia Goar
Paul Gorton
John Haglof
Lynne Hanauer
Mark Hardy
David Hartley
Carol Harvatin
Rebecca Hawthorne
Tom Heck
Jon Hegge
Mike Hehner
Tami Helmer
Paul Herda
John Hermansen
Tim Himsel
Jonathan Hinz
Sara Holle
Lori Holmberg
Jim Huntley
Rex Irmen
Troy Johnson
Kari Johnston
Rob Johnstone
William B. Jones
Phil Juntti
Andrew Karre
Patrick Kartes
Geoffrey Kinsey
Tony Kubat
Karl Larson
John Lauenstein
Tom Lemmer
Bill Lindner
Earl Lindquist
Daniel London
Barbara Lund
Curtis Lund
Sarah Lynch
Mark Macemon
Bernice Maehren
Dave Mahoney
Paul Markert
Chris Marshall
Brett Martin
Jamey Mauk
Jeanette Moss McCurdy
John Nadeau
Paul Najlis
Bill Nelson
Charles Nields
Mette Nielsen
Kristen Olson
Carol Osterhus
Brad Parker
Mike Parker
Christian Paschke
Matthew Paymar
Mike Peterson
Greg Pluth
Anne Price-Gordon
John Riha
Joe Robillard
Mary Rohl
Tom Rosch
Susan Roth
Andrea Rugg
Gary Sandin
Joel Schmarje
Philip Schmidt
Mark Scholtes
Ned Scubic
Gina Seeling
Cathleen Shannon
Mike Shaw
Hugh Sherwood
Dane Smith
Danielle Smith
Steve Smith
Angela Spann
Dick Sternberg
Ruth Strother
Lori Swanson
Andrew Sweet
Dianne Talmage
Glenn Terry
Keith Thompson
Adrianne Truthe
Gregory Wallace
Kevin Walton
Robert Weaver
John Webb
Brad Webster
Wayne Wendland
John Whitman
Dan Widerski
Steve Willson
Christopher Wilson
Mike Woodside

Index

A
ABS pipes. *See* Plastic pipes (ABS, PVC, CPVC)
Acoustical ceiling tiles, installing, 236–238
Air-conditioning, guidelines for kitchen, 345–346
Ampacity (amperage) of wires, 90, 91
Appliances
  circuit maps for, 142–143
  dishwashers, hooking up, 402–406
  energy efficient, 347, 418–419
  food disposers, installing, 407–411
  minimum trap sizes for, 80
Asbestos, 24, 176
Attics
  finishing ceilings, 516–517
  installing
    baseboard heaters, 524–527
    floors, 506–509
    walls, 514–515
  planning remodeling, 494–495, 498

B
Balloon framing, elements of, 11
Bamboo flooring, installing, 186–190
Bamboo panels, 210
Baseboard heaters
  circuit maps for, 143
  installing, 524–527
Basements
  adding egress windows, 518–523
  installing
    baseboard heaters, 524–527
    floors, 500–505
    walls, 510–513
  planning remodeling, 496–499
Base molding, installing, 242–245
Bathrooms
  accessibility of, 424–429, 486–487
  codes & standards, 135, 422–423
  installing
    alcove bathtubs, 438–445
    drains, 472–473, 478–479
    faucets, 472–477
    flooring
      ceramic tile, 191–197
      mosaic tile, 198–199
    floor-warming systems, 156–161
    glass shelving, 488–489
    grab bars, 486–487
    showers
      enclosure kits, 430–435
      hinged doors for, 436–437
    sinks
      pedestal, 468–469
      vessel, 464–467
    sliding tub doors, 450–453
    3-piece tub surrounds, 446–449
    toilets, 458–461
    towel warmers, 490–491
    tub spouts, 480–483
    vanity cabinets, 470–471
    vanity lights, 484–485
    wall-hung vanities, 462–463
  tiling walls, 454–457
Bay windows, installing, 268–277
Beadboard wainscoting, installing, 214–219
Bedrooms, highlights of National Electrical Code about, 135
Bifold doors, installing interior, 300–301
Branch pipes, sizing for, 79
Brass valves, soldering, 62
Building codes & permits, 16–19
  bathrooms, 135, 422–423
  electrical
    basic requirements, 131–131
    highlights, 134–135
    learning about, 128, 131
  kitchens, 135, 342–343
  for PEX, 71
  for plumbing, 77–78
Built-up base molding, installing, 244–245
Butted casings on doors, installing, 319

C
Cabinets
  installing bathroom vanity, 470–471
  installing kitchen
    base cabinets, 356–357
    hanging cabinets, 353–355
    overview of, 350
    preparing walls, 351–352
  removing, 39
Cables
  entering receptacles, 122
  highlights of National Electrical Code about, 134
  installing NM (Nonmetallic), 98–101
  overview of, 90–92
Carpet squares, installing, 200–203
Ceilings
  finishing attic, 516–517
  installing
    acoustical tiles, 236–238
    crown molding, 250–253
    electrical boxes on, 114
    light fixtures on, 324–326
    metal, 239–241
    panels on, 228–231
    recessed light fixtures on, 327–329
    skylights, 278–283
      building shafts for, 284–287
      tubular, 288–289
    suspended, 232–235
    track lighting on, 330–333
    wallboard on flat, 47
Cementboard underlayment, installing, 152
Ceramic tile flooring, installing, 191–197
Chair rails, installing, 248–249
Circuit maps, using, 138–145
Codes. *See* Building codes & permits
Color chart of wires, 90
Concrete walls, installing wires on, 106–107
Conductors, maximum number permitted per electrical box, 108
Conduit
  connecting nonmetallic, 105
  installing on concrete walls, 106–107
  overview of, 102–104
Construction plans, 22–23
Copper pipes
  cutting, 59
  soldering, 56, 58, 60–62
  working with, 56–57
Countertops, installing
  laminate, 368–375
  postform, 364–367
  tiled, 376–385
CPVC pipes. *See* Plastic pipes (ABS, PVC, CPVC)
Crown molding, installing, 250–253

D
Demolition
  making opening for exterior doors, 32–35
  overview of, 28
  removing
    cabinets, 39
    partition walls, 36–37
    plaster walls, 31
    trim, 38
    wallboard, 30
  support, installing temporary, 29
Dining rooms, highlights of National Electrical Code about, 135
Dishwashers, hooking up, 402–406

Doors
  codes for exit, 17
  exterior
    making opening for in lap siding, 32–33
    making opening for in stucco, 34–35
  framing, 12–13
  framing interior, 290–291
    for load-bearing walls, 293
    prehung, 292
    sliding & folding doors, 293
  guidelines for kitchen, 344
  installing casings, 318–321
  installing entry, 306–309
  installing interior
    bifold doors, 300–301
    French doors, 302–305
    overview of, 290
    pocket doors, 296–299
    prehung, 294–295
  installing paneling around, 213
  installing patio, 312–317
  installing storm, 310–311
Drains
  hooking up kitchen sink, 398–401
  installing bathroom sink, 472–473, 478–479
Drain-waste-vent system
  cleanouts for, 80
  drain pipe sizes, 80
  overview of, 55
  vent pipe sizes & orientation, 80
Drawings, 20–21
Dry-backed resilient flooring tiles, installing, 175, 176, 180–184

E

Electrical boxes
  common types of, 108, 109, 110
  highlights of National Electrical Code about, 134
  installing
    on ceilings, 114
    junction boxes, 114–115
    for light fixtures, 113
    for receptacles, 112
    for switches, 113
  locating, 111
  maximum number of conductors permitted per box, 108
  overview of, 85, 108
  wall covering thickness and, 111
Electrical symbols, 137
Electrical system
  delivery from power plants, 83
  glossary of terms, 87
  guidelines for kitchen service, 344
  meters described, 84
  overview of home, 86
  parts of, 84–85
  safety, 25, 28
Entertainment rooms, highlights of National Electrical Code about, 135
Entry doors, installing, 306–309
Exits & openings, codes for, 17

F

Faucets
  installing bathroom, 472–477
  installing kitchen
    overview of, 392–393
    putting in new faucet, 395–397
    removing old faucet, 394
Fire protection, codes for, 18
First-aid kits, 25
Fittings
  methods used with copper pipes, 56
  soldering copper pipes &, 56, 58, 60–62
Fixtures
  installing panels & wainscoting around, 213, 218
  minimum trap sizes for plumbing, 80
  See also Light fixtures
Flexible copper pipes
  cutting, 59
  overview of, 56, 57
Flooring
  elements of, 14
  installing
    bamboo planks, 186–190
    carpet squares, 200–203
    ceramic tile, 191–197
    hardwood strip, 162–165
    laminate, 166–169
    mosaic tile, 198–199
    parquet, 170–173
    radiant floor-warming systems, 156–161
    resilient tiles, 175, 176, 180–185
    sheet vinyl, 174–179
  installing attic, 506–509
  installing basement, 500–505
  refinishing wood, 204–207
  underlayment for
    installing
      cementboard, 152
      isolation membrane, 153
      panels, 154
      plywood, 151
      raised panels, 154–155
    overview of, 150
Food disposers, installing, 407–411
Four-way switches
  circuit maps for, 145
  installing of, 119
  overview of, 119
  testing, 121
Framing
  doors, 12–13, 290–293
  elements of, 10–11
  with steel, 42–43
  windows, 12–13, 256–259
French doors, installing interior, 302–305
FRP panels, 210

G

Garden windows, installing, 264–267
GFCI (ground-fault circuit-interrupter) receptacles
  circuit maps for, 139, 142
  installing, 127
  overview of, 126
Glassblock walls, building, 220–223
Glass shelving, installing, 488–489
Grab bars, installing, 486–487
Green kitchens, 347, 416–419
Grounding
  highlights of National Electrical Code about, 135
  wires
    described, 84
    installing through metal conduits, 102

H

Hallways, highlights of National Electrical Code about, 135
Handicapped persons. *See* Universal design
Hard copper pipes. *See* Rigid copper pipes
Hardwood strip flooring, installing, 162–165
Headers, overview of, 12–13
Heating
  guidelines for kitchen, 345–346
  installing baseboard heaters, 524–527
  installing panels & wainscoting around vents, 213, 218
  installing towel warmers, 490–491

I

Inspections, 16
Isolation membrane underlayment, installing, 153

J
Joints, taking apart soldered, 63
Joists, installing temporary, 29
Junction boxes, installing, 114–115

K
Kitchens
  accessibility of, 348–349
  appliances
    circuit maps for, 142–143
    dishwashers, hooking up, 402–406
    energy efficient, 347, 418–419
    food disposers, installing, 407–411
    minimum trap sizes for, 80
  building islands
    stock-cabinet, 358–359
    tiled, 360–363
  codes & standards
    electrical, 135, 344
    heating, ventilation & air-conditioning, 345–346
    plumbing, 345
  designing
    general guidelines, 342–344
    layouts, 340–343
    overview of, 338–339
  dishwashers, hooking up, 402–406
  drains, hooking up sink, 398–401
  green, 347, 416–419
  installing backsplashes, 386–389
  installing cabinets
    base, 356–357
    hanging, 353–355
    overview of, 350
    preparing walls, 351–352
  installing countertops
    laminate, 368–375
    postform, 364–367
    tiled, 376–385
  installing faucets
    overview of, 392–393
    putting in new faucet, 395–397
    removing old faucet, 394
  installing food disposers, 407–411
  installing sinks, 390–391
  installing undercabinet lights, 412–415

L
Laminate flooring, installing, 166–169
Laminate panels, 210
Lap siding, making opening for exterior doors in, 32–33
Laundry rooms, highlights of National Electrical Code about, 135
Light fixtures
  circuit maps for, 139–141, 144–145
  described, 84
  highlights of National Electrical Code about, 134
  installing
    bathroom vanity, 484–485
    on ceilings, 324–329
    electrical boxes for, 113, 114
    track on ceilings, 330–333
    undercabinet in kitchens, 412–415
Lighting
  codes for natural, 18
  guidelines for kitchen, 344
Living rooms, highlights of National Electrical Code about, 135

M
Materials
  for applying veneer plaster, 224
  for framing with steel, 42
  for wood floors, refinishing, 204
  for working with conduit, 104
  for working with PEX pipes, 68
Mitered casings on doors & windows, installing, 319
Molding
  installing base, 242–243
  installing built-up base, 244–245
  installing crown, 250–253
Mosaic tile flooring, installing, 198–199

N
National Electrical Code highlights, 134–135
NM (Nonmetallic) cable
  about, 91, 92
  installing, 98–101
  reading, 93
  stripping sheathing & insulation, 94

O
One-coat veneer plaster system, applying, 224–226
Outdoors
  highlights of National Electrical Code about, 135
  installing ceramic tiles, 191–197
Outlets. *See* Receptacles

P
Paneling
  installing
    on ceilings, 228–231
    metal on ceilings, 239–241
    on walls, 211–213
  types, 210
Parquet flooring, installing, 170–173
Partition walls
  building, 40–42
  removing, 36–37
Patio doors, installing, 312–317
Perimeter-bond sheet vinyl, institution, 174–175, 177–179
Permits. *See* Building codes & permits
PEX pipes
  buying, 71
  connecting, 73
  installing, 70
  overview of, 68
  plumbing water-supply system with, 74–75
  system designs for, 72
  tools & materials for, 68
Picture rails, installing, 246–247
Pipes
  copper
    cutting, 59
    soldering, 56, 58, 60–62
    working with, 56–57
  PEX
    buying, 71
    connecting, 73
    installing, 70
    overview of, 68
    plumbing water-supply system with, 74–75
    system designs for, 72
    tools & materials for, 68
  plastic
    cutting, 65
    joining with solvent-glue, 66–67
    overview of, 64
  sizing for, 79
  supports for, 80
Plaster, applying veneer, 224–227
Plaster walls, removing, 31
Plastic pipes (ABS, PVC, CPVC)
  cutting, 65
  joining with solvent-glue, 66–67
  overview of, 64
Platform framing, elements of, 10
Plumbing
  codes & permits, 77–78, 423
  guidelines for kitchen, 345
  mapping system, 76
  overview of, 52–55
  symbols, 76
  water supply system with PEX pipes, 74–75
Plywood underlayment, installing, 151
Pocket doors, installing interior, 296–299
Prehung doors, installing interior, 294–295
PVC pipes. *See* Plastic pipes (ABS, PVC, CPVC)

R
Radiant floor mats
  bamboo flooring and, 186
  installing, 156–161
Raised underlayment panels, installing, 154–155
Ranges, circuit maps for, 143

Receptacles
  circuit maps for, 138–139, 142
  highlights of National Electrical Code about, 134
  installing
    electrical boxes for, 112
    GFCI, 127
    new, 124–125
    panels & wainscoting around, 213, 218
  overview of, 85, 122–123
Recessed ceiling lights, installing, 327–329
Remodeling steps, 20–23
Resilient flooring tiles, installing, 175, 176, 180–185
Resilient sheet vinyl flooring, installing, 174–179
Rigid copper pipes
  cutting, 59
  overview of, 56
Rigid plastic pipes. *See* Plastic pipes (ABS, PVC, CPVC)
Roofs
  elements of, 15
  estimating material needed, 531
  shingling
    installing drip edge, 537
    installing three-tab shingles, 538–545
    installing underlayment, 536
    preparation, 532–533
    removing old shingles, 534–535
Room dimensions, codes for, 17
Round-top windows, installing, 262–263

S
Safety, 24–25
  asbestos and, 24, 176
  wiring, 88–89
Self-adhesive resilient flooring tiles, installing, 175, 176, 180, 185
Self-rimming sinks, installing, 390–391
Service masts, described, 84
Service panels
  described, 85
  highlights of National Electrical Code about, 134
  new wiring and main, 128, 130
Sheet vinyl flooring, installing, 174–179
Showers, installing
  enclosure kits, 430–435
  hinged doors for, 436–437
Siding
  estimating material needed, 532
  installing vinyl, 548–553
    overview of, 546
    tools for, 547
    making opening for exterior doors in lap, 32–33
Single-pole wall switches
  circuit maps for, 140–141
  installing, 117
  overview of, 117
  testing, 120
Sinks
  hooking up drains, 398–401
  installing
    bathroom pedestal, 468–469
    bathroom vessel, 464–467
    self-rimming, 390–391
Skylights
  building shafts, 284–287
  installing, 278–283
    tubular, 288–289
Sliding patio doors, installing, 312–317
Small appliances, circuit maps for, 142
Soft copper pipes. *See* Flexible copper pipes
Soldered joints, taking apart, 63
Soldering copper pipes, 56, 58, 60–62
Solvent-glue, joining plastic pipes with, 66–67
Split-circuit receptacles, 122, 123
Stairs, highlights of National Electrical Code about, 135
Steel, framing with, 42–43
Storm doors, installing, 310–311
Stucco, making opening for exterior doors in, 34–35
Studs, joining sections using steel, 43
Supply tubes, sizing for, 79
Suspended ceilings, installing, 232–235
Switches
  four-way
    circuit maps for, 145
    installing of, 119
    overview of, 119
    testing, 121
  highlights of National Electrical Code about, 134
  installing electrical boxes for, 113
  installing panels & wainscoting around, 213, 218
  overview of, 85, 116
  single-pole wall
    circuit maps for, 140–141
    installing, 117
    overview of, 117
    testing, 120
  testing, 120–121
  three-way
    circuit maps for, 144
    installing, 118
    overview of, 118
    testing, 121
Symbols
  electrical, 137
  plumbing, 76

T
Thermostats, circuit maps for, 143
Three-way switches
  circuit maps for, 144
  installing, 118
  overview of, 118
  testing, 121
Tileboard panels, 210
Tiles
  building tiled kitchen islands, 360–363
  installing
    acoustical ceiling, 236–238
    backsplashes, 386–389
    on bathroom walls, 454–457
    flooring
      ceramic, 191–197
      mosaic, 198–199
    metal ceiling, 239–241
    tiled countertops, 376–385
Toilets, installing, 458–461
Tongue-and groove hardwood flooring, installing, 162–165
Tools
  for applying veneer plaster, 224
  for framing with steel, 42
  installing vinyl siding, 547
  for installing wallboard, 45
  for resilient floors, 175
  safety and, 25
  stilts, 47
  for wood floors, refinishing, 204
  for working with conduit, 104
  for working with PEX pipes, 68
Towel warmers, installing, 490–491
Track lighting, installing on ceilings, 330–333
Trim, removing, 38
Tubs, installing
  alcove, 438–445
  sliding doors, 450–453
  spouts, 480–483
  3-piece surrounds, 446–449
Two-coat veneer plaster system, applying basecoat in, 224, 226
Two-slot receptacles, 122, 123

U
Undercabinet kitchen lights, installing, 412–415
Underlayment for floors
    installing
    cementboard, 152
    isolation membrane, 153
    panels, 154

- plywood, 151
- raised panels, 154–155
- overview of, 150

Underlayment for roofs, installing, 536
Universal design
- in bathrooms, 424–429, 486–487
- in kitchens, 348–349
- overview of, 19

Utility rooms, highlights of National Electrical Code about, 135

V

Valve requirements, 79
Vanities
- installing lights for, 484–485
- installing wall-hung, 462–463

Veneer plaster, applying, 224–227
Ventilation
- codes for, 18
- guidelines for kitchen, 345–346

Vent system. *See* Drain-waste-vent system
Vinyl siding, installing, 546–553
Voltage ratings, 83

W

Wainscoting, installing, 214–219
Wallboard
- cutting, 46
- installing
  - at corners, 49, 51
  - finishing & taping, 50–51
  - on flat ceilings, 47
  - overview of, 44
  - tools & prep work for, 45
  - on walls, 48
- removing, 30

Wall-hung vanities, installing, 462–463
Walls
- applying veneer plaster, 224–227
- building
- glassblock, 220–223
  - partition, 40–42
- cabinet installation preparation of, 351–352
- elements of, 15
- installing
  - attic, 514–515
  - basement, 510–513
  - base molding, 242–245
  - beadboard wainscoting, 214–219
  - chair rails, 248–249
  - crown molding, 250–253
  - picture rails, 246–247
  - wallboard on, 48
  - wires on concrete, 106–107
  - wood paneling, 210, 211–213
- removing partition, 36–37
- tiling bathroom, 454–457

Waste system. *See* Drain-waste-vent system
Water supply system
- copper pipes, 56–57
- overview of, 54
- plumbing with PEX pipes, 74–75
- sizing for distribution pipes, 79

Windows
- adding basement egress, 17, 518–523
- framing, 12–13, 256–259
- guidelines for kitchen, 344
- installing
  - bay, 268–277
  - casings, 318–321
  - garden, 264–267
  - paneling around, 213
  - replacements with masonry clips, 261
  - replacements with nailing flange, 259–261
  - round-top, 262–263
  - wainscoting around, 219

Wires/wiring
- connecting
  - with push-ins, 95
  - to screw terminals, 95
  - using wire connector, 96
- connector types, 93
- grounding
  - described, 84
  - installing through metal conduits, 102
- installing on concrete walls, 106–107
- new
  - drawing plans, 136–137
  - examining main service panel, 130
  - National Electrical Code
    - basic requirements, 131–133
    - highlights, 134–135
  - planning steps, 128–129
- overview of, 82, 90–92
- pigtailing, 97
- reading unsheathed individual, 93
- safety, 88–89

Wood floors, refinishing, 204–207

Also From **CREATIVE PUBLISHING international**

- Complete Guide to A Green Home
- Complete Guide to Attics & Basements
- Complete Guide to Basic Woodworking
- Complete Guide Build Your Kids a Treehouse
- Complete Guide to Contemporary Sheds
- Complete Guide to Creative Landscapes
- Complete Guide to Custom Shelves & Built-Ins
- Complete Guide to Decorating with Ceramic Tile
- Complete Guide to Decks
- Complete Guide to DIY Projects for Luxurious Living
- Complete Guide to Dream Bathrooms
- Complete Guide to Dream Kitchens
- Complete Guide to Finishing Walls & Ceilings
- Complete Guide to Floor Décor
- Complete Guide to Gazebos & Arbors
- Complete Guide to Home Carpentry
- Complete Guide to Landscape Construction
- Complete Guide Maintain Your Pool & Spa
- Complete Guide to Masonry & Stonework
- Complete Guide to Outdoor Wood Projects
- Complete Guide to Painting & Decorating
- Complete Guide to Patios
- Complete Guide to Plumbing
- Complete Guide to Roofing & Siding
- Complete Guide to Trim & Finish Carpentry
- Complete Guide to Windows & Entryways
- Complete Guide to Wiring
- Complete Guide to Wood Storage Projects
- Complete Guide to Yard & Garden Features
- Complete Outdoor Builder
- Complete Photo Guide to Home Repair
- Complete Photo Guide to Home Improvement
- Complete Photo Guide to Homeowner Basics

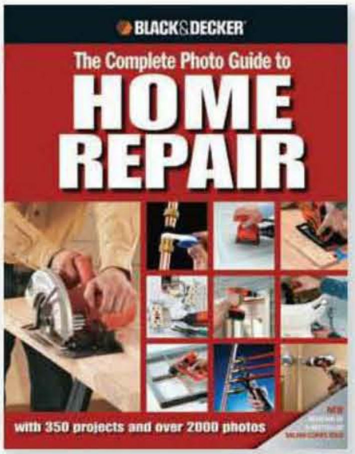
BLACK&DECKER
The Complete Photo Guide to
HOME REPAIR
with 350 projects and over 2000 photos

ISBN 1-58923-417-0

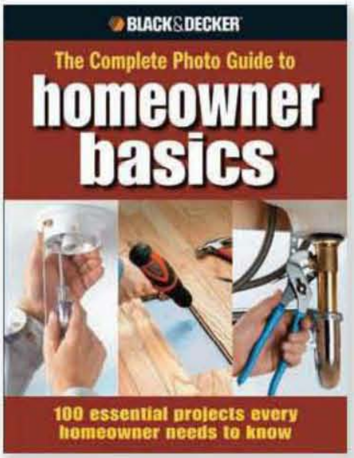
BLACK&DECKER
The Complete Photo Guide to
homeowner basics
100 essential projects every homeowner needs to know

ISBN 1-58923-376-X

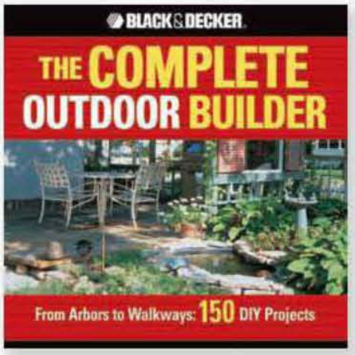
BLACK&DECKER
THE COMPLETE OUTDOOR BUILDER
From Arbors to Walkways: 150 DIY Projects

ISBN 1-58923-264-X

Creative Publishing international

400 First Avenue North • Suite 300 • Minneapolis, MN 55401 • www.creativepub.com

Taunton's

REVISED EDITION

WIRING COMPLETE

EXPERT ADVICE FROM START TO FINISH

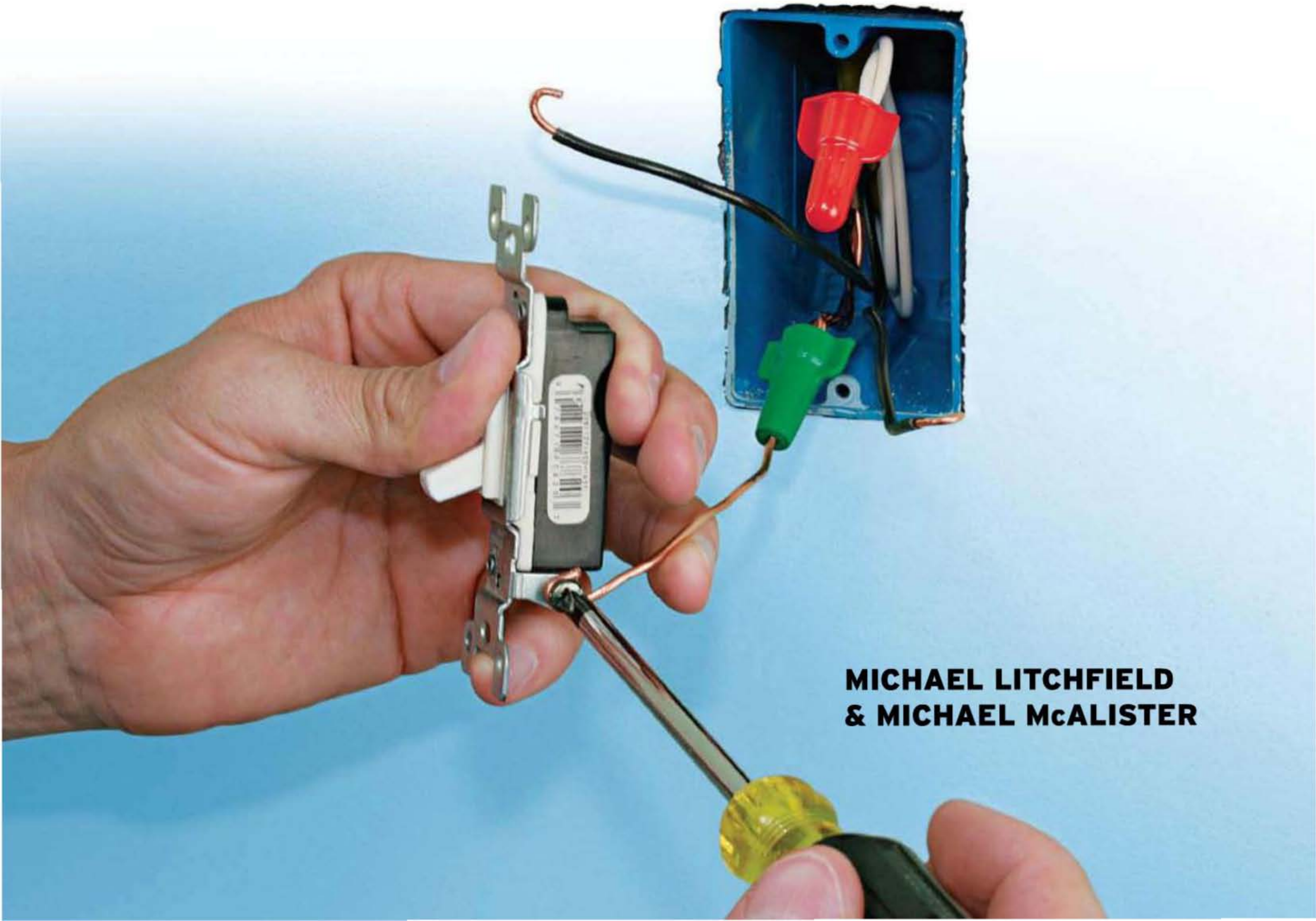

MICHAEL LITCHFIELD
& MICHAEL McALISTER

Taunton's

REVISED EDITION

WIRING COMPLETE

EXPERT ADVICE FROM START TO FINISH

Text by Michael W. Litchfield and Michael McAlister,
© 2008 by The Taunton Press, Inc.
Photographs © 2008 by The Taunton Press, Inc., except as noted.
Illustrations © 2008 by The Taunton Press, Inc.
All rights reserved.

The Taunton Press
Inspiration for hands-on living®

The Taunton Press, Inc.,
63 South Main Street, PO Box 5506,
Newtown, CT 06470-5506
e-mail: tp@taunton.com

Editors: Matthew Teague, Helen Albert
Copy editor: Candace B. Levy
Indexer: Jay Kreider
Jacket/Cover design: Kimberly Adis
Interior design: Kimberly Adis
Layout: Cathy Cassidy
Illustrator: Trevor Johnston
Photographer: Michael Litchfield, except where noted.

Library of Congress Cataloging-in-Publication Data
Litchfield, Michael.
Wiring complete / Michael Litchfield and Michael McAlister. -- Rev. ed.
p. cm.
Includes index.
ISBN 978-1-60085-256-5
1. Electric wiring, Interior. 2. Dwellings--Electric equipment. I. McAlister, Michael.
II. Title.
TK3285.L54 2008b
621.319'24--dc22
2008054144

Printed in the United States of America
10 9 8 7 6 5 4 3 2

Dozens of people made this book possible. We are especially indebted to the electricians and builders who got us onto job sites and allowed us to photograph them at work. Thanks to old friend Rafael Maldonado of Blue Electric in Berkeley, California, for sharing his vast knowledge and his crew, particularly Isaac Castro, Jorge Dominguez, and Carlos Guerrero. Huge thanks to the unflappable Daniel Kealey, Nathan Parker, and Din Abdullah of MRM Electrical, Berkeley. Likewise, Kevin McCarthy, Simon Jordan, Ki Soo An, and Jimmy Stuart. A million thanks to Mike Zelinka and colleagues at Laner Electric in Richmond, California.

We are grateful to Jana Olson, Karen Cornell, and John Nicoles of Omega Too Lighting in Berkeley for the excellent sequences of lamp and chandelier repair. Thanks also to Nowell's Antique Lighting in Sausalito, California.

Hats off to Jamie Carlen and Chris D'Andrea of Jamie Carlen & Co., Berkeley; and Chip Harley, Jesus Beltran, and Gelber Guzman of Holland-Harley. Thanks to supremely talented architect Gary Earl Parsons of Berkeley and his associate Cecil Lee.

Thanks to Roger Robinson of the Star Inspection Group, Oakland, California, for letting us cherry-pick his photo archive of code violations; Casablanca Fans for product shots; Muffy Kibbey for photos of the handsome kitchen in the planning chapter; and Ken Gutmaker for photos that first appeared in *Renovation, 3rd Edition* (Taunton, 2005).

Thanks to homeowners Catherine Moncrieff and Stuart Brotman, Martha and Dean Rutherford, Judith and Stanley Lubman, Laurie Case, and Andy Baker.

Finally, high-fives to the Taunton Press family and friends, including Executive Editor Helen Albert, Wendi Mijal, Matthew Teague, Jennifer Peters, Nicole Palmer, Julie Hamilton, Sandor Nagyszalanczy, Rex Cauldwell, Clifford Popejoy, Joseph Truini, Mike Guertin, and Fernando Pagés Ruiz.

—Michael Litchfield and Michael McAlister

The following manufacturers/names used in *Wiring Complete* are trademarks: Adapt-Touch™, American Aldes™, BernzOmatic™, Broan-NuTone℠, Capsylite®, Casablanca® Fan Co., Channellock®, Clikstrip®, *Code Check*®, Condulet®, COPALUM®, Cutler-Hammer®, Decora®, DeWALT®, Fantech®, Federal Pacific®, Fluke®, Greenlee® Nail Eater™, Grip-lok™, Halogen Capsylite®, Hubbell®, Hunter®, Ideal®, InSinkErator®, International Code Council®, Juno®, Kelvar®, Klein Tools™, Knipex®, Kobe®, Legrand®, Lemnis®, Leviton®, Lightolier®, Lutron®, Milwaukee®, National Electrical Code®, Nouvelle® fan, Panasonic®, Panduit®, Phillips®, PowerFast®, Romex®, Roto-Spilt®, Rotozip®, Scandia® fan, Seatek®, Snap-Tite®, Speed® Square, Square D®, Strap Gun®, Tech Lighting®, Tech Lighting℠, Vol-Con® Lite, Watt Stopper®, WD-40®, Wiggy®, Wire-Nut®, Zinsco®

Wiring is inherently dangerous. Using hand or power tools improperly or ignoring safety practices can lead to permanent injury or even death. Don't try to perform operations you learn about here (or elsewhere) unless you're certain they are safe for you. If something about an operation doesn't feel right, don't do it. Look for another way. We want you to enjoy working on your home, so please keep safety foremost in your mind.

contents

INTRODUCTION 3

WORKING WITH ELECTRICITY 4

BASICS

6 Understanding Electricity
7 Working Safely
8 Service Panels
10 Home Electrical Systems
12 Grounding Basics
14 Evaluating the Electrical System

CUTTING POWER

16 Cutting Power at the Panel

TESTING

17 Testing with a Neon Tester
18 Testing with an Inductance Tester

TOOLS & MATERIALS 20

TOOLS

22 Hand Tools
23 Power Tools
24 Electrical Testers

ELECTRICAL BOXES

25 Choosing Electrical Boxes
27 Installing New Work Boxes
28 Removing Knockouts
29 Mounting Devices

CABLES & WIRES

30 Cable & Conduit
32 Cable Clamps & Clips
33 Securing Cable to Framing
34 Stripping Cable with a Utility Knife
35 Stripping Cable with a Cable Ripper
36 Choosing Wire
37 Stripping & Splicing Wires

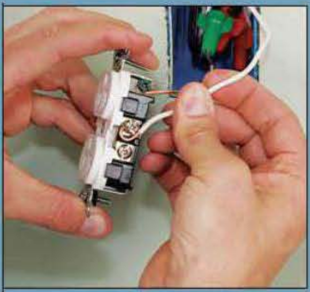

RECEPTACLES & SWITCHES 38

TOOLS & MATERIALS

40 Choosing Receptacles & Switches
42 Back-Wired Devices
43 Testing for Power

WIRING RECEPTACLES

44 Two Ways to Wire a Receptacle
45 Wiring a Duplex Receptacle
47 Making Receptacle Boxes Flush to Drywall
48 Wiring a GFCI Receptacle
49 Wiring a Split-Tab Receptacle

TESTING DEVICES

50 Testing a Single-Pole Switch
51 Testing a Three-Way Switch
51 Testing a Four-Way Switch

WIRING SWITCHES

52 Wiring a Single-Pole Switch
53 Single-Pole Switch with Back-Fed Wiring
54 Replacing a Single-Pole Switch
56 Wiring a Linear Slide Dimmer
58 Roughing In a Three-Way Switch
59 Wiring a Three-Way Switch
60 Wiring a Back-Fed Three-Way Switch
61 Wiring a Four-Way Switch

WIRING SPECIALTY SWITCHES

62 Wiring a Switch/Receptacle Combo
63 Wiring a Double Switch
63 Wiring a Motion Detector
64 Wiring an Electronic Timer Switch
65 Wiring a Manual Timer

>> >> >> >>

LIGHTING 66

BASICS

68 Fixture Wiring
69 Bulb Types

REPAIRS

70 Testing Lamps & Removing Old Wires
72 Rewiring a Lamp
74 Replacing a Plug
75 Refurbishing a Chandelier

BASIC FIXTURES

78 Mounting Light Fixtures
80 Installing a Simple Porcelain Fixture

WALL SCONCES

82 Removing an Existing Wall Sconce
83 Connecting a New Sconce
84 Installing Halogen Bulbs & Accessories

CHANDELIERS

85 Prepping & Mounting a Ceiling Box
86 Disconnecting a Chandelier
87 Installing a Chandelier

RECESSED LIGHTING

88 Cutting a Ceiling Hole for Recessed Lighting
90 Wiring a Recessed Fixture
91 Securing the Can

UNDERCABINET FLUORESCENTS

92 Undercabinet Fluorescent Fixtures

LOW-VOLTAGE FIXTURES

94 Low-Voltage Systems
95 Installing Low-Voltage Cable Lighting
95 Anchoring Low-Voltage Standoffs
96 Running Cable
97 Tensioning Cables
98 Locating the Canopy
99 Mounting the Box & Wiring the Transformer
101 Installing the Canopy
102 Attaching the Feed Rods & Fixtures

MONORAIL TRACK LIGHTING

103 Installing Low-Voltage Monorail Track Lighting
104 Locating the Track & Standoffs
105 Mounting & Wiring the Transformer
106 Hanging the Track
107 Installing the Light Fixtures

REPAIRING LOW-VOLTAGE CONTROLS 108

DOORBELLS

110 Troubleshooting a Doorbell
113 Replacing a Doorbell Switch
114 Installing a Chime Unit
115 Installing a Doorbell Transformer

THERMOSTATS

116 Testing an Old Thermostat
117 Installing a Programmable Thermostat

FURNACES

118 Replacing a Furnace Transformer

MULTIMEDIA 120

TOOLS & MATERIALS

122 Multimedia Cables & Connectors
123 Stripping Cable

ADDING PHONE JACKS

124 Extending a Phone Line
126 Mounting a New Jack

DISTRIBUTION PANELS

127 Mounting a Media Distribution Panel
130 Running Cables to a Media Distribution Panel
131 Connecting Cables at the Panel

INSTALLING A MULTIMEDIA OUTLET

132 Locating the Multimedia Outlet
133 Installing Cable Connectors
134 Connecting Panduit Jacks
135 Installing Leviton Jacks
136 Attaching Jacks to an Outlet Plate

FANS
138

WIRING TIPS

140 Before You Begin

BATHROOM VENT FANS

141 Choosing a Bathroom Fan
142 Locating a Bathroom Fan
143 Installing the Roof Vent
145 Mounting the Fan Box
146 Retrofitting a Bathroom Fan
147 Wiring a Bathroom Fan

RANGE HOODS

149 Mounting & Ducting a Range Hood
150 Wiring the Range Hood

CEILING FANS

152 Installing the Ceiling Fan Box
154 Mounting the Ceiling Fan

PLANNING NEW WORK
156

BEFORE YOU BEGIN

158 Inspecting the Fuse Box or Breaker Panel
160 Assessing Wiring Condition
161 Is the System Adequately Sized?

ACCORDING TO CODE

163 Common Code Requirements
164 General-Use Circuit Requirements

WORKING DRAWINGS

165 Developing a Floor Plan
168 Electrical Notation
170 Receptacles
172 Switches & Lights
174 Receptacles, Switches & Lights

ROUGH-IN WIRING
176

GETTING STARTED

178 Rough-In Basics
180 Tools for Rough-In
182 Materials for Rough-In
184 Laying Out the Job

OUTLET BOXES

186 Installing Wall Boxes
187 Installing Ceiling Boxes

RUNNING CABLE

188 Drilling for Cable
189 Pulling Cable
190 Feeding Cable through Corners
191 Fishing Cable behind Finish Walls
192 Fishing Cable to a Ceiling Fixture

RETROFIT WORK

194 Cutting a Wall Box into Plaster
194 Cutting a Wall Box into Drywall
195 Retrofitting a Ceiling Box
196 Creating a Wiring Trench
197 Retrofitting Boxes & Pulling Cable

MAKING UP BOXES

198 Making Up an Outlet Box
199 Making Up a Light Can
200 Making Up a Single Switch Box
201 Making Up a Multigang Switch Box

METAL CABLE & CONDUIT

202 Flexible Metal Cable
203 Installing AC Cable
204 Installing MC Cable
205 Working with EMT Steel Conduit
206 Making Turns with Metal Conduit
207 Fishing Cable & Conduit
208 Prepping Receptacles for Surface Metal Boxes
209 Wiring Outlets

>> >> >> >>

APPLIANCES 210

BASICS

212 A Quick Look at Kitchen Appliances
213 Preparing an Unfinished Appliance Cord

GARBAGE DISPOSERS

214 Installing a Garbage Disposer

OVENS

216 Roughing In an Oven Outlet
217 Wiring a Drop-In Oven

HOUSEHOLD APPLIANCES

218 Wiring Laundry Setups
220 Installing Baseboard Heaters
221 Electric Water Heaters

OUTDOOR WIRING 222

BEFORE YOU BEGIN

224 Planning Your Outdoor System
225 Choosing Outdoor Lights
226 Tools & Materials

OUTDOOR OUTLETS

228 Tapping into an Existing Outlet
230 Mounting & Wiring an Exterior Outlet

OUTBUILDINGS

231 Running Power to an Outbuilding

OUTDOOR PATH LIGHTS

233 Installing Low-Voltage Path Lights

MOTION-SENSOR LIGHTS

235 Installing a Motion-Sensor Light

PANELS & SUBPANELS 236

BEFORE YOU BEGIN

238 Understanding Service Panels & Subpanels
240 Sizing Panels, Subpanels & Conductors

CIRCUIT BREAKERS

242 Adding a Circuit Breaker

SUBPANELS

244 Installing a Subpanel

GLOSSARY 247

INDEX 248

INTRODUCTION

Electrical wiring requires attention to detail, patience, and a little dexterity, but it's nothing the average homeowner can't tackle. Before you buckle on that tool belt, however, take a few moments to read the first section, which provides an overview of electrical systems and a handful of essential safety tips. Section Two walks you through the tools you'll need and various techniques you'll use again and again. The remainder of the book takes you through every step of the wiring process—from replacing an old light fixture to wiring an entire house.

Before you buy tools and materials, however, check with local building code authorities. Although most local building codes do not forbid an owner's doing his or her own electrical work, most require a rough inspection—that is, before wires are connected to switches, receptacles and so on—and a final inspection when everything is wired, trimmed, and tested. Besides, building inspectors are usually knowledgeable: They can tell you if local codes conform to the National Electrical Code® or, if not, how they vary. Finally, check with your insurance agent to make sure that doing your own electrical work won't jeopardize your homeowners insurance coverage.

WORKING WITH ELECTRICITY

BEFORE WORKING WITH electricity, you should have a basic understanding of how it works. This chapter is designed to give you a quick overview of the electrical system in your home, including the major components. Since grounding is essential to keeping you safe, we cover that in detail. We'll also show how to do a basic inspection of your home for wiring problems. Respecting the power of electricity is essential to working safely. Always follow the instructions carefully, use appropriate safety equipment, and when in doubt consult a licensed electrician.

Before beginning work, check with local building authorities to make sure local regulations allow you to do your own work and that you are conforming to local code requirements.

BASICS

Understanding electricity, p. 6

Working safely, p. 7

Service panels, p. 8

Home electrical systems, p. 10

Grounding basics, p. 12

Evaluating the electrical system, p. 14

CUTTING POWER

Cutting power at the panel, p. 16

TESTING

Testing with a neon tester, p. 17

Testing with an inductance tester, p. 18

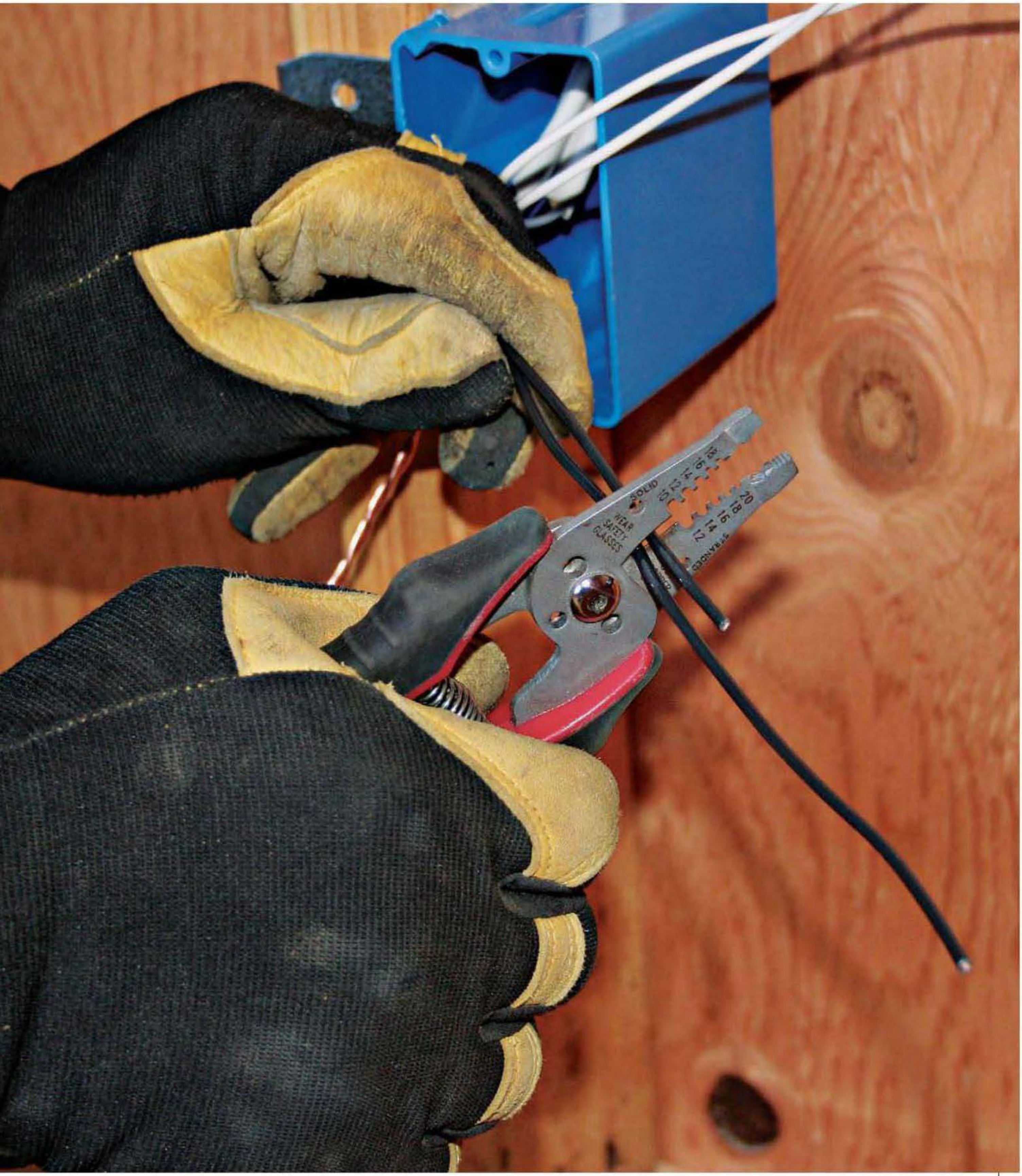
SOLID
18
16
14
12
10
WEAR
SAFETY
GLASSES
20
18
16
14
12
STRANDED

UNDERSTANDING ELECTRICITY

The easiest way to understand how electricity flows is to visualize water flowing through a pipe. Electricity (flowing electrons called *current*) moves through a circuit like water in a pipe. Just as the flow of water is measured in gallons per minute, the electrical flow of electrons is measured in "amperes" or *amps*. Water pressure is measured in pounds per square inch, and the force of electrons in a circuit is measured in *volts*. The larger the pipe, the more water that can flow through it; likewise, the larger wires allow a greater flow of electrical current. Too much water can rupture a pipe. Wiring that is too small will resist the flow of current. If that resistance (measured in *ohms*) is too great, the wires can melt and cause a fire.

AC (alternating current) electrical systems can be thought of as a loop that runs from the generation point (or power source) through utilization equipment (a light bulb for instance) and back to the generation point. Along the way it may cross the country through great transmission lines, through the power lines on your street, and through the cables in the walls of your house. In your home, the main loop, that is the service to your home, is broken into smaller loops called circuits. Typically, a *hot* wire (usually black or red) carries current to the utilization equipment from the service panel, and a *neutral* wire (typically white or light gray) carries current back to the service panel.

CIRCUIT BASICS

Electricity always flows in a circuit. The hot wire carries the current from the power source to the fixture and the neutral wire returns it to the power source.

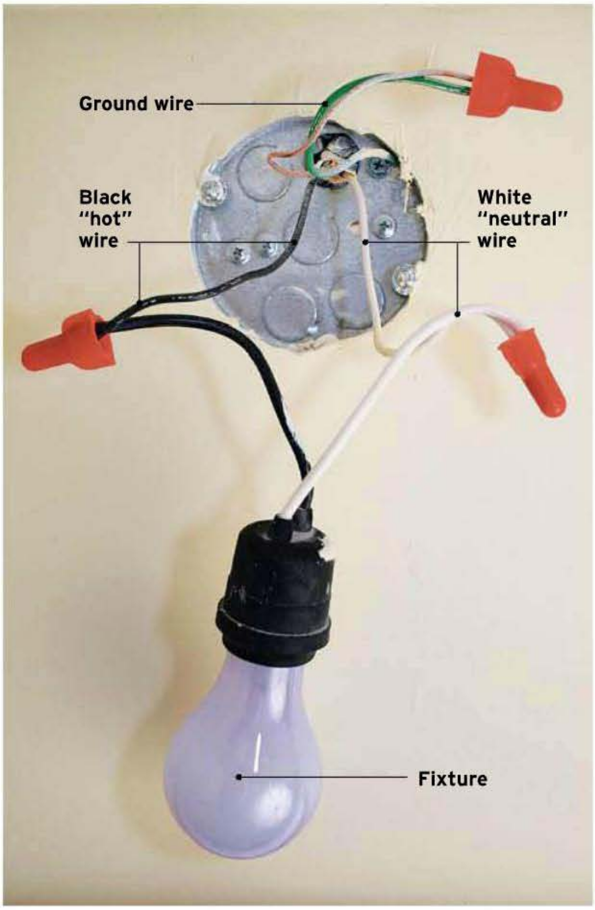

Key Terms

Power (Volt-Amps or VA)
The potential in the system to create motion (motors), heat (heaters), light (fixtures or lamps), etc. Volt-amps = available volts x available amps (VA).

Watts
Are a measure of power consumed. Watts are very similar to volt-amps, the main difference being that some energy is lost through heat and/or inherent inefficiencies in equipment.

Voltage
Voltage is the *pressure* of the electrons in a system. Voltage is measured in *volts*.

Amperes (Amps)
The measure of the volume of electrons flowing through a system (current).

Current
The flow of electrons in a system. Current is measured in amperes *(amps)*. There are two types of current: DC (direct current) and AC (alternating current). Typically AC is found in homes and buildings.

Ohms
Ohms is the measure of resistance to the flow of electrons (current) in a material (like a cable). The higher the resistance, the lower the flow of electrons.

WORKING SAFELY

To work safely with electricity you must respect its power. If you understand its nature and heed the safety warnings in this book—especially shutting off the power and testing with a voltage tester to make sure power is off—you can work with it safely. The cardinal rule of home-improvement projects, which goes double for electrical work, is: Know your limitations. Unless you have previous experience doing electrical work and feel confident about your skills, you should leave certain projects to a pro. Working inside a service panel or removing its cover can be especially dangerous. There is an area around the main breaker that remains hot even after the breaker is set to the off position. Never attempt to remove or repair the main service panel or the service entrance head. When in doubt, call a licensed electrician rather than risk harm.

For most projects, careful attention to following the instructions in this book will keep you safe and ensure good results. Where there is a particular risk in a project, we will call attention to it with a warning, such as the following:

WARNING

Please read this information carefully as it could save you from serious injury.

Always wear appropriate safety gear, including rubber-soled shoes, gloves (if working with cables, wires, or metal boxes); and safety glasses and a dust mask (when sawing or working overhead). And remember that current flows most easily along a path of least resistance, but it will follow any path that's available . . . including you!

Turn off the power to the circuit at the main service panel before removing receptacle, switch, or fixture covers.

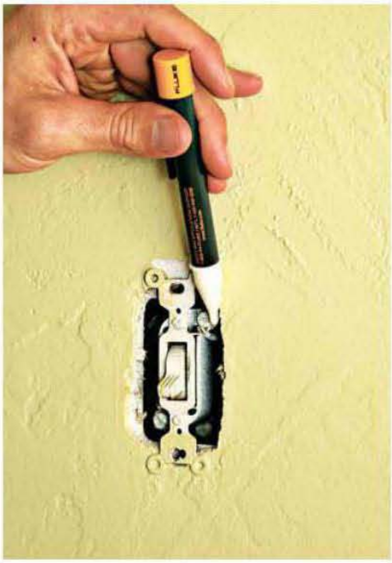

After removing the plate or cover, test again to make sure power is off.

Wear gloves to protect your hands from the sharp edges of wires, cables, and metal boxes.

Wear safety glasses when sawing or drilling, especially when working overhead.

SERVICE PANELS

At the main service panel, the two hot cables from the meter base attach to lug terminals atop the main breaker. The incoming neutral cable attaches to the main lug of the neutral/ground bus. In main service panels, neutral/ground buses must be bonded, usually by a main bonding jumper. In subpanels and all other locations downstream from the main service panel, ground and neutral buses must be electrically isolated from each other.

In a main fuse box, the hot cables attach to the main power lugs, and the neutral cable to the main neutral lug. Whether the panel has breakers or fuses, metal buses run from the bottom of the main breaker/main fuse. Running down the middle of the panel, buses distribute power to the various branch circuits. Similarly, neutral/ground buses are long aluminum strips with many screws, to which ground and neutral wires attach.

Each fuse or breaker is rated at a specific number of amps, such as 15 or 20. When a circuit becomes overloaded, its current flow becomes excessively high. This causes its breaker to trip or its fuse strip to melt, thereby cutting voltage to the hot wires. All current produces heat; but as current doubles, the heat generated quadruples. If there were no breakers or fuses, current would continue flowing until the wires overheated and a fire started. Thus the amperage ratings of breakers and fuses are matched to the size (cross-sectional area) of the circuit wires.

The main breaker

All electricity entering a house goes through the main breaker, which is usually located at the top of a main panel. In an emergency, throw the main breaker switch to cut all power to the house. The main breaker is also

INSIDE THE SERVICE PANEL

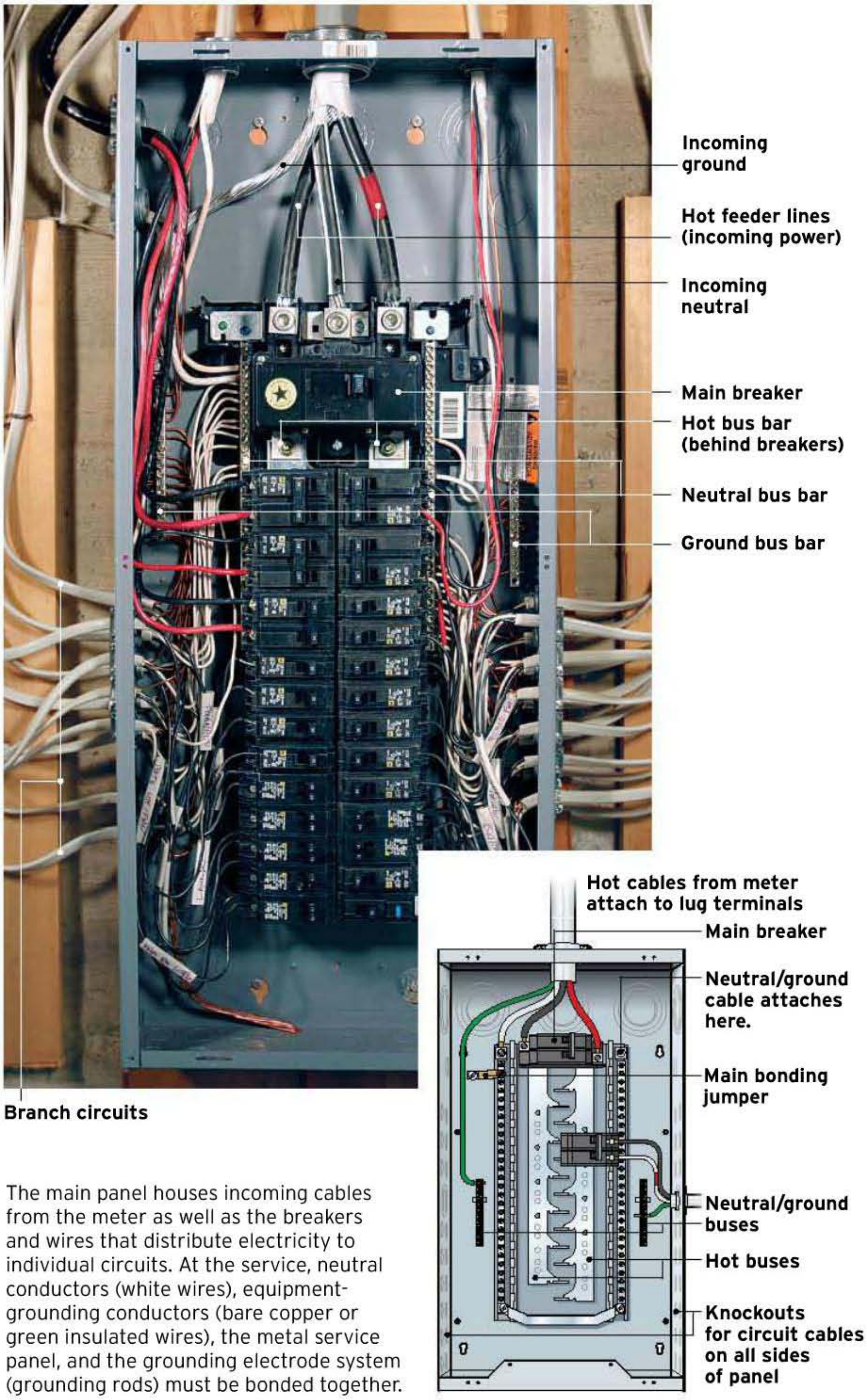
Incoming ground
Hot feeder lines (incoming power)
Incoming neutral
Main breaker
Hot bus bar (behind breakers)
Neutral bus bar
Ground bus bar
Branch circuits
Hot cables from meter attach to lug terminals
Main breaker
Neutral/ground cable attaches here.
Main bonding jumper
Neutral/ground buses
Hot buses
Knockouts for circuit cables on all sides of panel

The main panel houses incoming cables from the meter as well as the breakers and wires that distribute electricity to individual circuits. At the service, neutral conductors (white wires), equipment-grounding conductors (bare copper or green insulated wires), the metal service panel, and the grounding electrode system (grounding rods) must be bonded together.

the primary overcurrent protection for the electrical system and is rated accordingly. (The rating is stamped on the breaker handle.) Thus if the main breaker for a 200-amp panel senses incoming current that exceeds its overload rating, the breaker will automatically trip and shut off all power.

Meter-main combos

Increasingly common are meter-main combos, which house a meter base and a main service panel in a single box. Meter-main combos allow a homeowner to put the main breaker outside the house, where it can be accessed in an emergency—say, if firefighters want to cut the power to a blazing house before they enter it. When the service panel is located outside, electricians typically locate the subpanel close by, inside the house, to minimize runs of large SER (service entrance) cable to the subpanel.

Fuse boxes

Breaker service panels are the most common type you'll find in today's homes, but many older homes still have fuse boxes. Fuses are among the earliest overcurrent devices, and they come as either Edison-type (screw-in) fuses or cartridge (slide-in) fuses. The Edison-style fuses that screw in like a light bulb are more common, and they have little windows that you can peer into to see if the filament is separated. Separation means that the circuit was overloaded and the fuse has blown. A blackened (from heat) interior could mean a short circuit—a potentially dangerous situation calling for the intervention of a licensed electrician. The less common cartridge fuses are used to control 240-volt (240v) circuits and are usually part of the main disconnect switch.

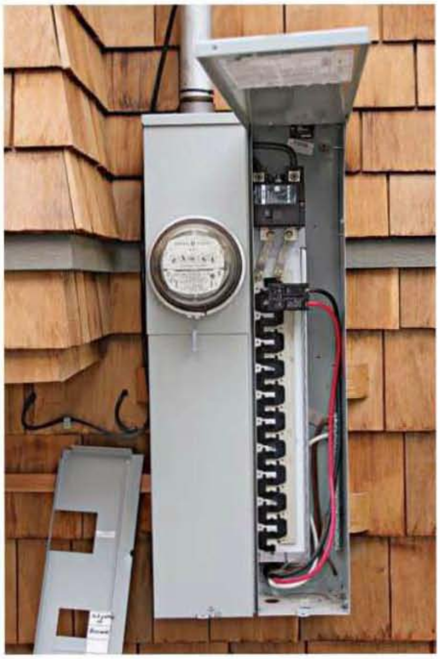

A meter-main combo, placed outside the house, provides easy access for service or emergencies.

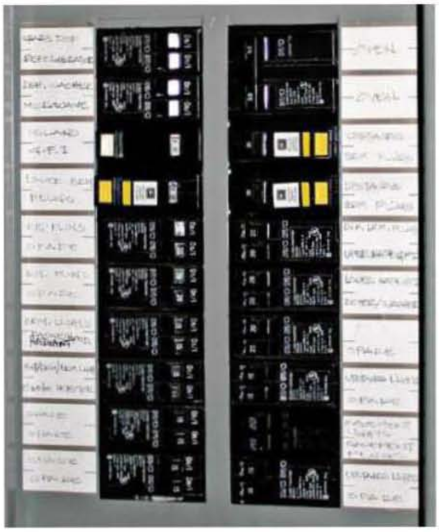

Labeling breakers speeds identification of the switch that turns off power to a device.

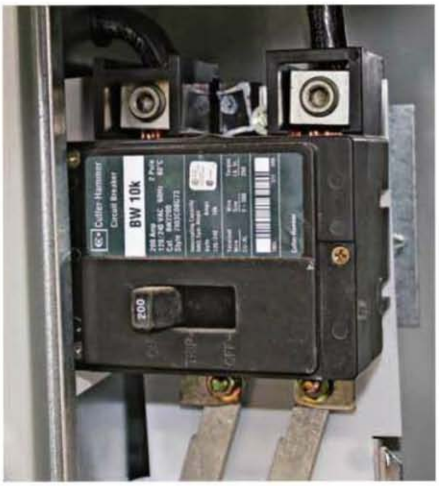

Stay away from the area around the main breaker switch. It remains hot and is extremely dangerous.

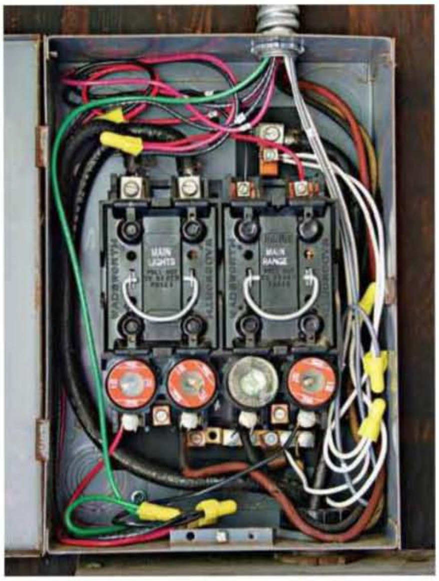

A fuse box in a subpanel.

HOME ELECTRICAL SYSTEMS

Power from the utility service is commonly delivered through three large cables, or conductors, which may enter the house overhead or underground. Overhead service cables are called a *service drop*. These cables run to a weatherhead atop a length of rigid conduit. When fed underground, service conductors are installed in buried conduit or run as underground service-entrance (USE) cable. Whether it arrives overhead or underground, three-wire service delivers 240v.

Service conductors attach to a meter base and then to the service panel. Straddling the two sets of terminals on its base, the meter measures the wattage of electricity as it is consumed. The service panel also routes power to various circuits throughout the house.

The utility company will install cables to the building and will install the meter. The homeowner is responsible for everything beyond that, including the meter base and entrance panel, which a licensed electrician should install.

WARNING

Whether the connection is an SER cable clamped to the lugs of a main breaker or between wires spliced together in an outlet box, the connection must be tight to be safe. Otherwise, electricity can leap a gap—it's called arcing—and that could lead to a house fire.

GFCI receptacles detect minuscule current leaks and shut off power almost instantaneously. They're important in kitchens and bathrooms, where water and dampness increase the potential for shocks.

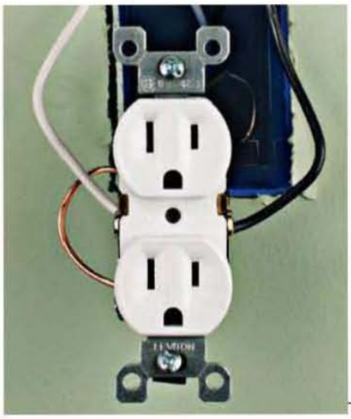

Modern three-prong outlets provide a low impedance path to ground in case a fault occurs.

Some appliances require 30-amp service. This is a 125/250v dryer receptacle. The breaker for this circuit must also be rated for 30 amps.

WIRING INSIDE OUT

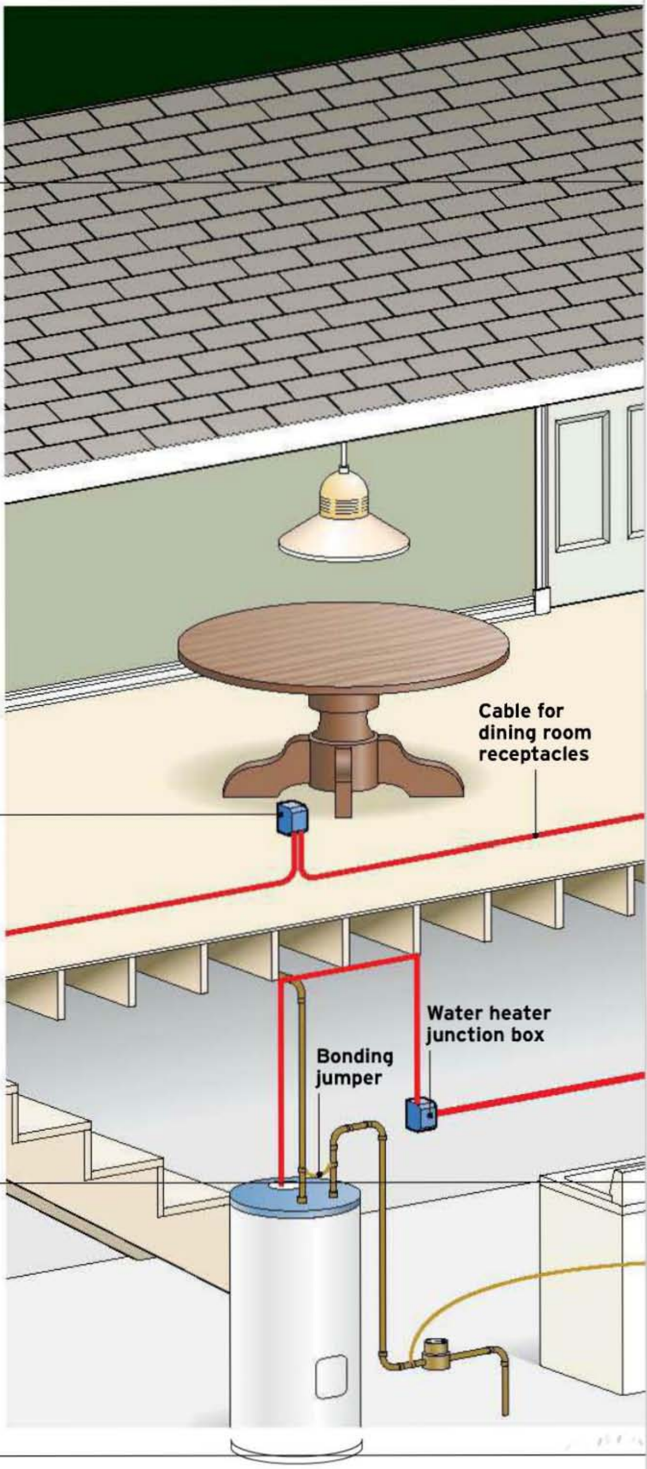

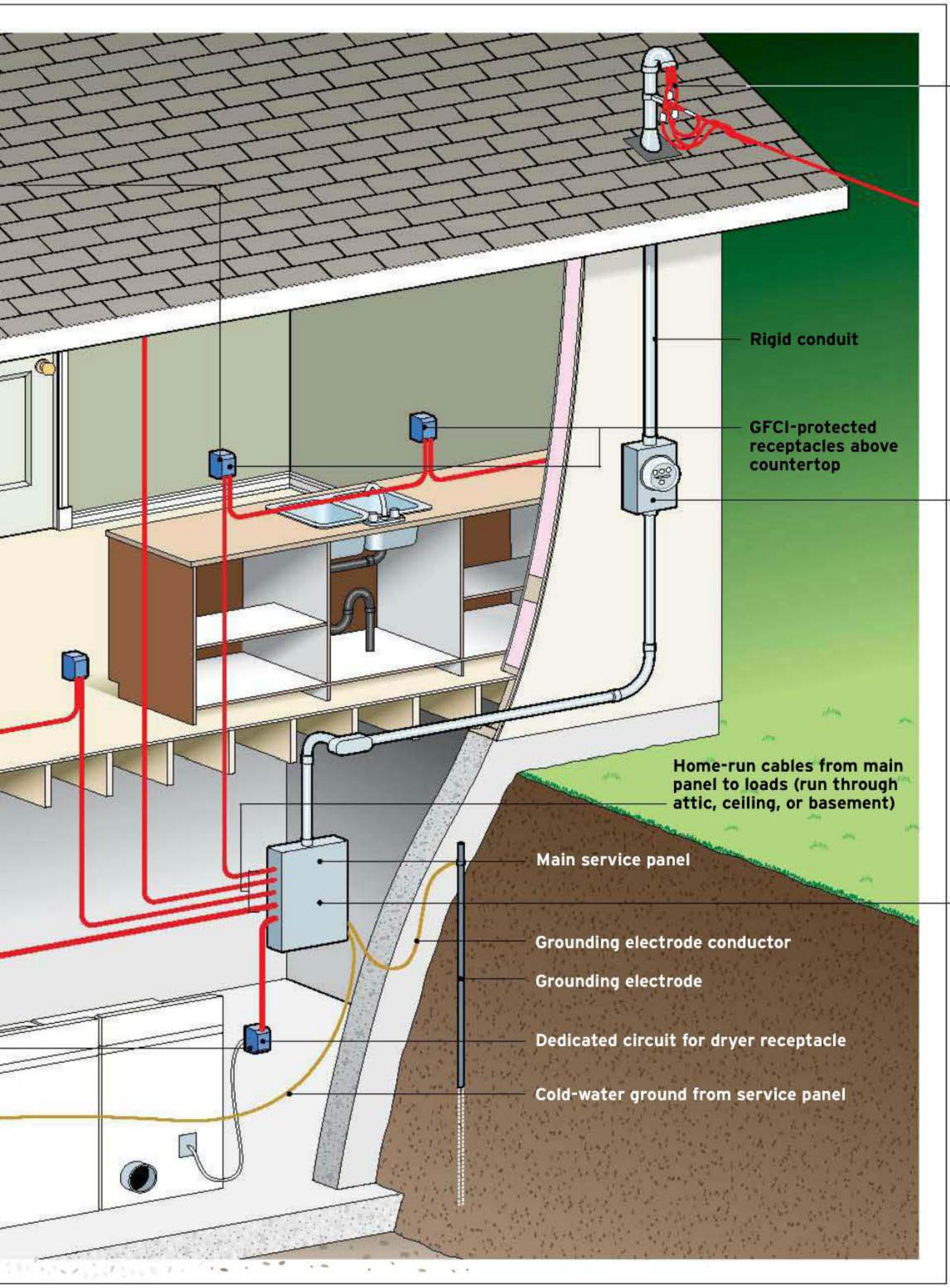
Rigid conduit
GFCI-protected receptacles above countertop
Home-run cables from main panel to loads (run through attic, ceiling, or basement)
Main service panel
Grounding electrode conductor
Grounding electrode
Dedicated circuit for dryer receptacle
Cold-water ground from service panel

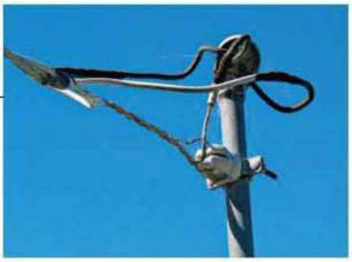

A typical three-wire service assembly has two insulated hot conductors wrapped around a bare messenger cable, which also serves as the neutral.

The meter provides a measure of the electrical power consumed. Positioned outside the house, it allows the power company to monitor consumption.

The service panel distributes power to circuits throughout the house. Breakers interrupt power if the circuits become overloaded.

GROUNDING BASICS

Because electricity moves in a circuit, it will return to its source unless the path is interrupted. The return path is through the white neutral wires that bring current back to the main panel. In the event that the current seeks to return to the panel through a path other than the neutral wire, ground wires provide the current with an alternative low-resistance path.

Why is this important? Before equipment-grounding conductors (popularly called *ground wires*) were widespread, people could be electrocuted when they came in contact with fault currents that unintentionally energized the metal casing of a tool or an electrical appliance. Ground wires bond all electrical devices and potentially current-carrying metal surfaces. This bonding creates a path with such low impedance (resistance) that fault currents zip along it, quickly tripping breakers or fuses and interrupting power. Contrary to popular misconceptions, the human body has a relatively high impedance (compared to copper wire), so if electricity is offered a path with less resistance (a copper ground wire), it will take it.

➔ **See "Avoiding Electrical Shocks," below.**

Individual ground wires connect to every part of the electrical system that could become a potential conductor—metal boxes, receptacles, fixtures—and, through three-pronged plugs, the metallic covers and frames of tools and appliances. The conductors, usually bare copper or green insulated wire, create an effective path back to the main service panel in case the equipment becomes energized.

The neutral/ground bus

The ground wires attach to a neutral/ground bus bar, which is itself bonded to the metal panel via a *main bonding jumper*. If there's a ground fault in the house, the main bonding

AVOIDING ELECTRICAL SHOCKS

GROUND FAULTS CAN KILL

Current flowing through an unintended conductor, such as a person, is called a *ground fault*. Because only a little current flowing through your heart can kill you, ground faults can be fatal. Likewise, you can get badly shocked by touching an energized wire or device with one hand and touching the other hand to a neutral or ground wire. In this case, you become part of the circuit and current flows through you.

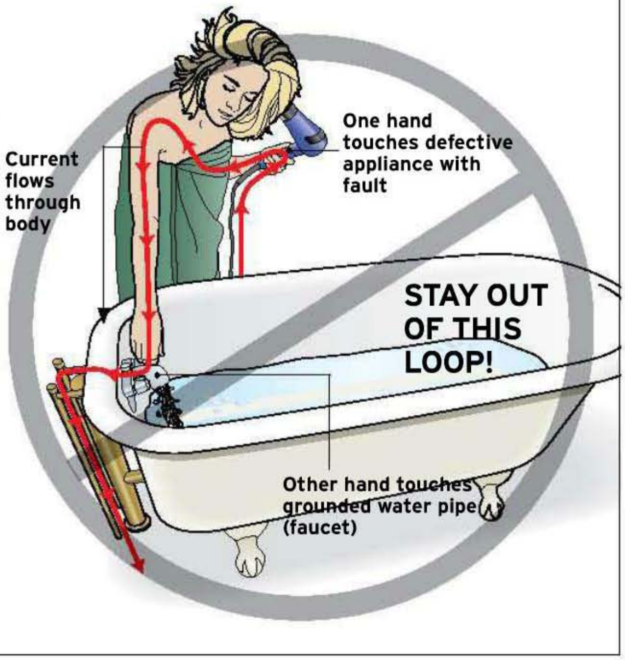
Current flows through body
One hand touches defective appliance with fault
STAY OUT OF THIS LOOP!
Other hand touches grounded water pipe (faucet)

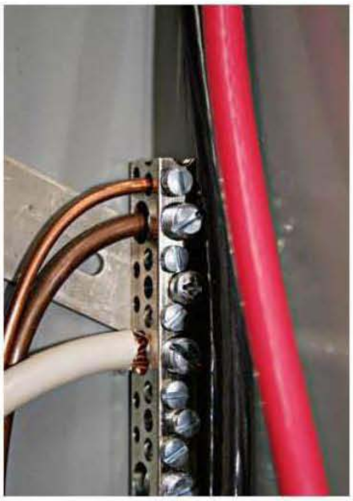

The smaller copper wire at the top runs to a ground rod; the thicker copper wire below and the neutral wire feed a subpanel within the house.

The main ground wire from the service panel clamps to an 8-ft. rod-grounding electrode (or ground rod) driven into the earth. It diverts outside voltage, such as lightning strikes.

Connections to cold-water pipes prevent shocks should the pipes become inadvertently energized.

jumper will ensure that the current can be safely directed to the ground—away from the house and the people inside. It is probably the single most important connection in the entire electrical system.

Also attached to the neutral/ground bus in the service panel is a large, bare copper ground wire—the grounding electrode conductor (GEC)—that clamps to a grounding electrode (also called a *ground rod*)—which is driven into the earth or attached to steel rebar in the footing of a foundation. The electrode's primary function is to divert lightning and other outside high voltages before they can damage the building's electrical system. Although the grounding electrode system (GES) is connected to the equipment grounding system at the service panel, the GES has virtually nothing to do with reducing shock hazards.

The National Electrical Code (NEC) sizes grounding electrode conductors based on the sizes and types of conductors in the service. Typically, residential GECs are size 6 American wire gauge (6AWG) copper. Ground rods are typically 5⁄8-in. to 3⁄4-in. copper-clad steel rods 8 ft. to 10 ft. long; the longer the rod, the more likely it will reach moist soil, whose resistance is less than that of dry soil. Be sure to install multiple rod systems in lightning-prone areas.

GFCIs

Ground-fault circuit interrupters (GFCIs) are sensitive devices that can detect even small current leaks and shut off power almost instantaneously. The NEC now requires GFCI protection on all bathroom receptacles; kitchen receptacles within 4 ft. of a sink; all receptacles serving kitchen counters; all outdoor receptacles; all accessible basement or garage receptacles; and all receptacles near pools, hot tubs, and the like.

→ **See "Wiring a GFCI Receptacle," on p. 48**

MAJOR GROUNDING ELEMENTS

The equipment-grounding system acts as an expressway for stray current. By bonding conductors or potential conductors, the system provides a low-resistance path for fault currents. The abnormally high amperage (current flow) that results trips a breaker or blows a fuse, disconnecting power to the circuit.

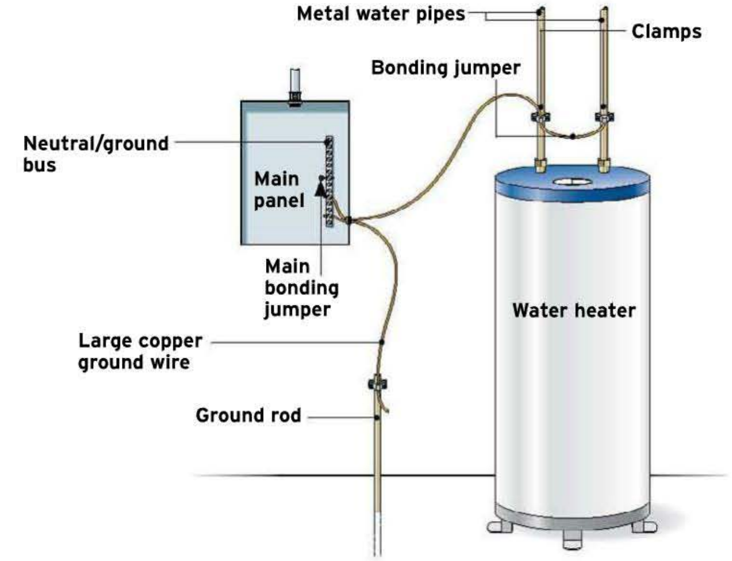
Metal water pipes
Clamps
Bonding jumper
Neutral/ground bus
Main panel
Main bonding jumper
Water heater
Large copper ground wire
Ground rod

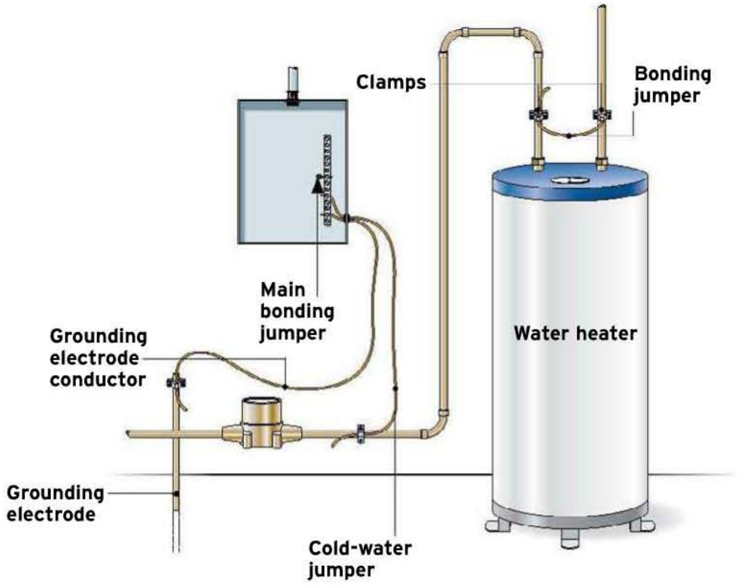
Clamps
Bonding jumper
Main bonding jumper
Water heater
Grounding electrode conductor
Grounding electrode
Cold-water jumper

EVALUATING THE ELECTRICAL SYSTEM

Before you start working on your electrical system, you should have a general knowledge of its condition. Certain situations could have a big impact on your safety as you work. Others, such as the electrical requirements of your home, could affect whether you can add new circuits or certain kinds of devices. For your peace of mind and safety, you may want to hire a licensed electrician or qualified home inspector to advise you about what needs to be done. Start by considering the following aspects of your existing service.

The service

Is there two-wire or three-wire service to the house? In older homes, electrical service is often undersize for the demands of a modern household. If your house has only two cables running from the utility pole—one hot and one grounded neutral conductor—it has only 120v service and should be upgraded to three-wire, 240v service. A 100-amp, circuit-breaker service panel is considered minimal today.

General condition of panels

If you see scorch marks, rust stains, wires with frayed or cracked insulation, or condensation on the service panel or damp conditions around it, the service is unsafe. Dampness is particularly unsafe, and many electricians will refuse to work on a panel

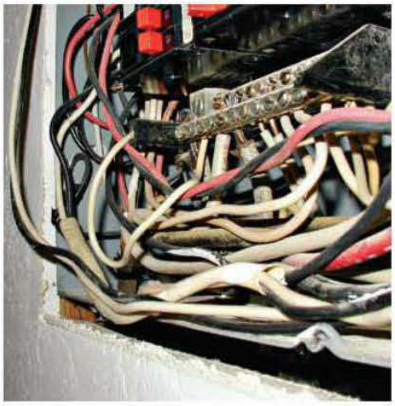

This panel is so overloaded that it will be tough to replace the cover.

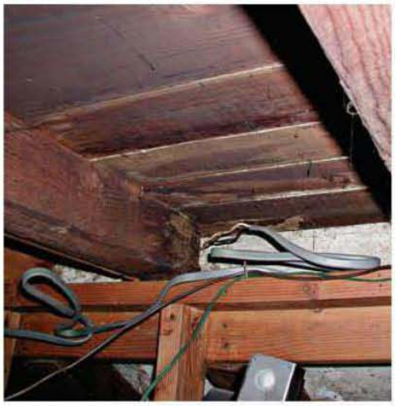

Sloppy, unstapled cable can work loose or be damaged.

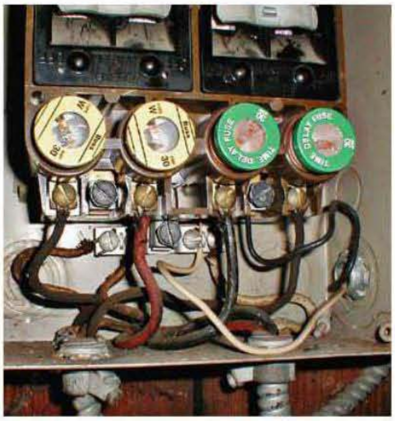

These fuses are too large (30 amps) for the load rating of the circuit wires. The wires could melt and start a fire.

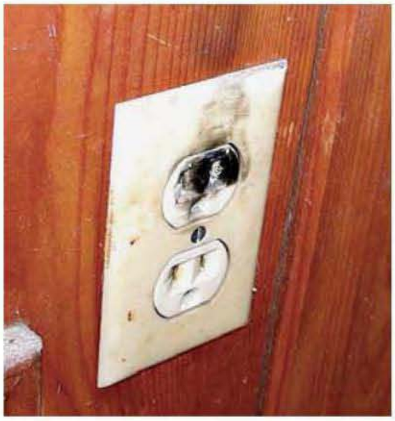

These scorch marks indicate electrical arcing (electrical leaps between connections that have worked loose).

Aluminum wiring Widely used in house circuits in the 1960s and 1970s, aluminum wiring expands and contracts excessively, which leads to loose connections, arcing, overheating, and—in many cases—house fires. The most common symptoms will be receptacle or switch cover plates that are warm to the touch, flickering lights, and an odd smell around electrical outlets. Once arcing begins, wire insulation deteriorates quickly. An electrician who checks the wiring may recommend adding COPALUM® connectors, CO/ALR-rated outlets and switches, or replacing the whole system. Aluminum service cable, however, is not a problem when terminated correctly and is still used today.

until the surrounding dampness is remedied. Look for melted plastic around wires, which could indicate overheating. Make sure the wire gauges match up with the size of the breaker (#14 wire on a 15-amp circuit, #12 on a 20-amp, and so forth). If the panel seems so stuffed with wires that you can barely close it, you should call a pro to do an inspection.

Current usage

Installing fuses too big for a circuit to prevent blown fuses is a fool's bargain; such circuits could overheat and start a fire. Make sure that heavy-duty circuits for ranges, dryers, and air-conditioners have the appropriately sized breakers installed. Likewise note overloaded receptacles, extension cords under rugs, and the like inside the house. These are invitations for a fire and a definite sign that more receptacles are needed in the area.

Grounding

Is the main service panel grounded? There should be a large grounding wire running from the panel and clamped to a cold-water pipe and/or a grounding rod.

Is there fixture and outlet grounding? Grounding the panel is not enough. For the entire electrical system to be grounded, there must be continuous ground wires running to every device in the house.

If the house has only two-slot receptacles, the system may be grounded by armor-clad cables. Test the receptacles by inserting the probe of a voltage tester into the short slot and touch the other to the mounting screw. (Make sure it's not covered with paint.) If the tester lights, you should install a grounded receptacle. Otherwise, consider upgrading the circuit. If there are three-slot receptacles, use a receptacle analyzer (p. 24) to check whether the plugs are grounded and polarized.

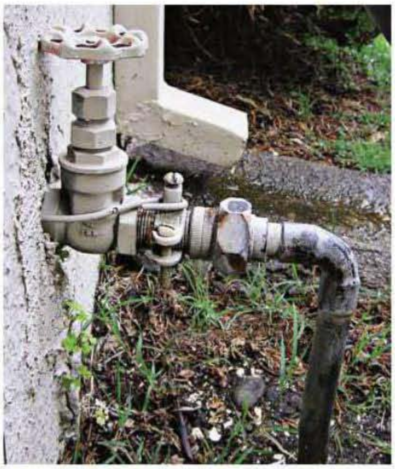

The ground is clamped to the wrong side of a dielectric union (coupling) on this copper cold-water pipe, so there's no ground.

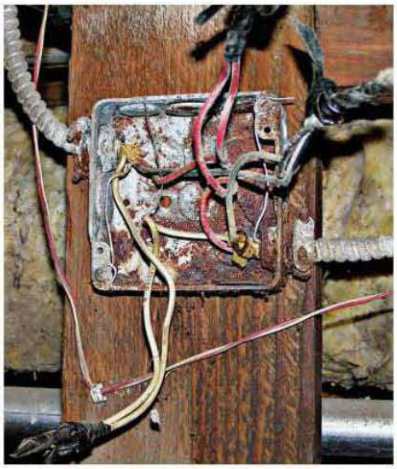

Unlike this installation, all wire splices must be protected by covered boxes that are installed in dry locations.

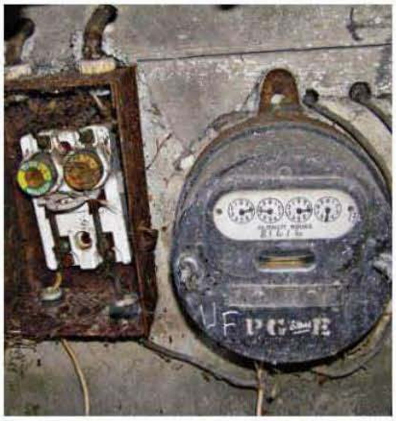

Moisture wicking through this masonry surface has rusted the fuse box and compromised electrical safety.

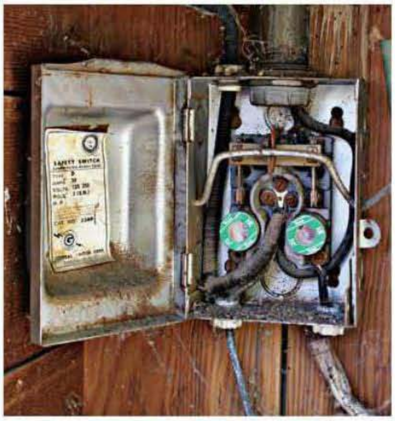

A 30-amp main is inadequate for modern uses; 200-amp service would be more appropriate.

Arcing

When nails puncture wires or electrical connections are loose or corroded, electricity can arc (jump) between points and cause fires. The NEC now requires arc-fault circuit interrupter (AFCI) protection for all 15-amp and 20-amp bedroom circuits. AFCIs detect minute fluctuations in current associated with arcing and de-energize the circuits before a fire can start. Installing AFCI breakers is essentially the same as installing GFCI circuit breakers.

CUTTING POWER AT THE PANEL

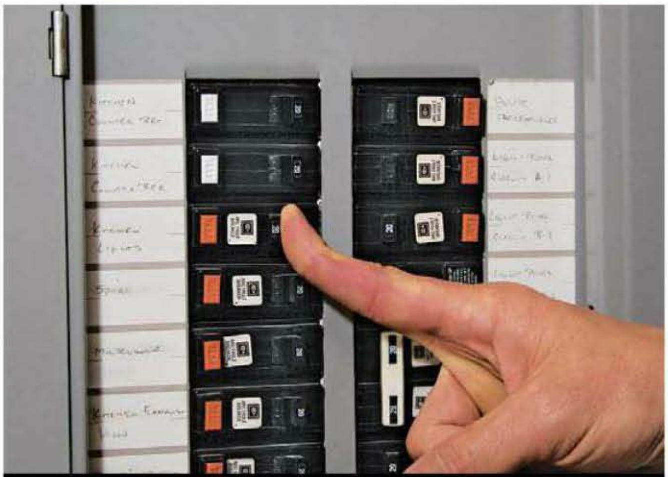

A In a normal breaker box, simply flip the relevant breaker to the off position.

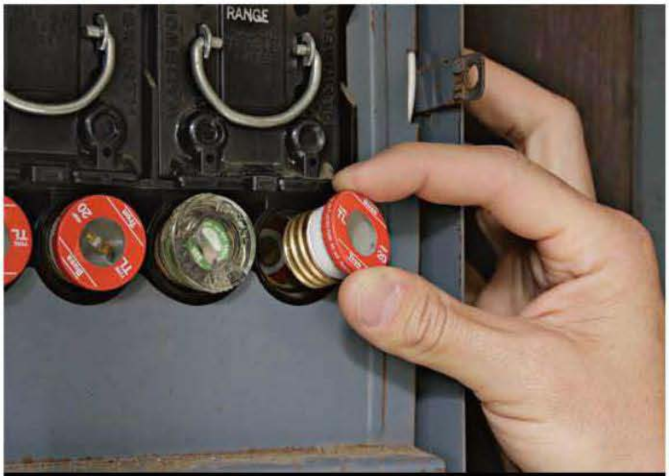

B If you're working on a fuse panel, unscrew the fuse that protects the circuit you're working on.

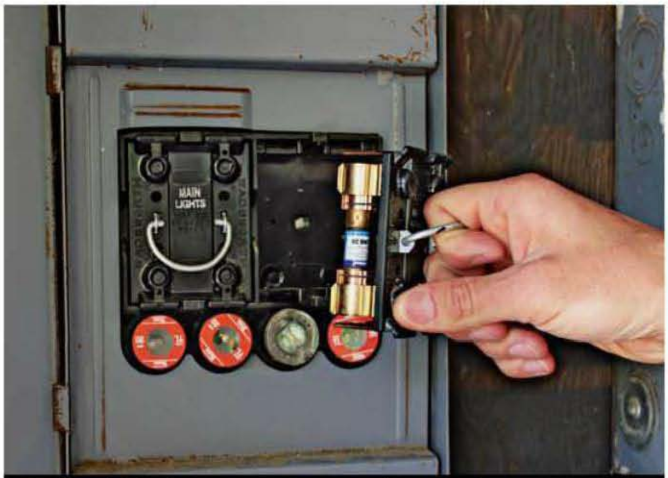

C If cartridge fuses are used, remove the cartridge or cartridge block.

D Once the power is off, post a warning sign to prevent others from turning the power back on while you're still working on it.

Always shut off the power to an outlet before working on it—and then test with an electrical tester to be sure there's no current present. Because individual devices such as receptacles, switches, and fixtures can give false readings if they are defective or incorrectly wired, the only safe way to shut off the electricity is by flipping a breaker in the service panel or subpanel.

Turning off the power at a breaker panel is usually straightforward. After identifying the breaker controlling the circuit, push the breaker's switch into the *off* position Ⓐ. The breaker switch should click loudly into place: If it doesn't, flip it again until you hear a loud click. (A breaker that won't snap into place may be worn out or defective and should be replaced by an electrician.)

If your home has a fuse panel instead, remove the fuse that controls the circuit Ⓑ, and tape a warning sign to the panel cover. Partially unscrewing a fuse is not a solution, because the fuse body is still in contact with the socket and possibly could be jiggled or rescrewed so that current resumes flowing through it. Remove the fuse. Likewise, if circuits are controlled by cartridge fuses, pull the cartridge or cartridge block out of the panel Ⓒ.

In any case, once you've cut off the power, shut the panel cover and tape a sign to it, telling others to stay out Ⓓ. Otherwise, someone not aware of the situation could flip the switch on, energizing the circuit you're working on.

TESTING WITH A NEON TESTER

When using a neon tester, hold the insulated part of the tester, and insert its probes into the receptacle slots. To test a three-slot receptacle, first insert the probes into the hot (short) and neutral (long) slots ❶. If the tester light does not glow, insert one probe into the hot slot and the other probe into the round ground slot. If there's current present, the tester will light up ❷. Next, insert one probe into the neutral slot and the other into the ground slot. Here, the tester will glow only if there is current and the hot and neutral wires have been reversed ❸.

If the tester does not light up for any of these three combinations, there's probably no voltage present. However, if the receptacle is faulty or a wire is loose, the previous tests may not detect voltage actually present in the outlet box. To be certain, unscrew the receptacle cover plate and the two screws holding the receptacle to the outlet box. Pull the receptacle out from the outlet box, being careful not to touch bare wires, receptacle screw terminals, or metal outlet boxes.

Then apply tester probes to the bare wire ends on both sides of the receptacle and to the gold and silver screw terminals ❹. Finally, if the outlet box is metal, touch one probe to the hot wire/screw and the other probe to the metal box ❺. If the tester doesn't glow during these tests, the power is off.

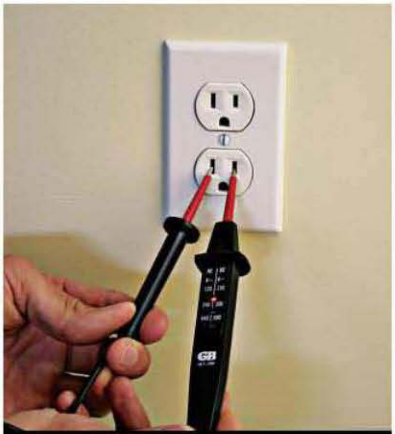
1 Insert probes into both the hot and the neutral slots on the receptacle.

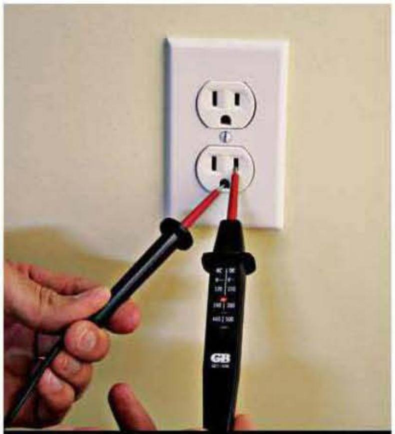
2 Test the hot (shorter) and the ground slots.

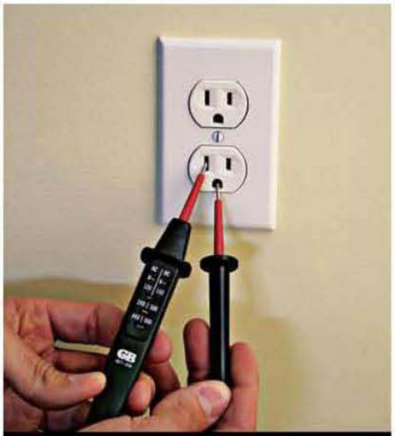
3 Test the neutral (longer) and the ground slots.

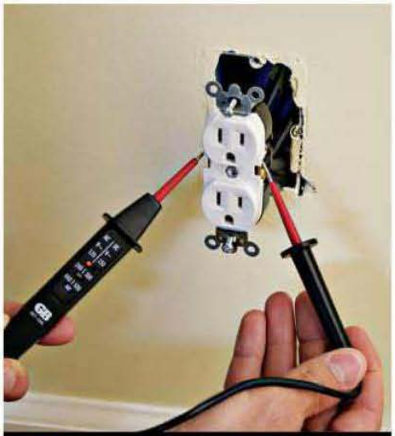
4 Test the gold and the silver screws.

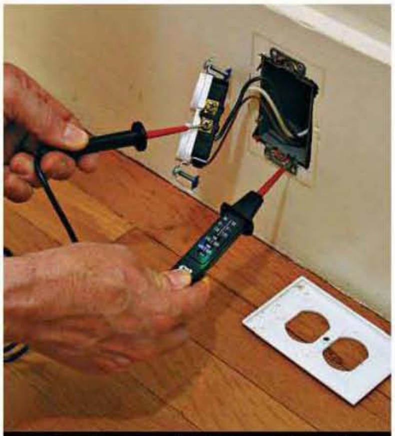
5 Touch one probe to the hot wire/screw and the other to the metal box.

WARNING

To avoid shock, never hold the tester probes by their bare tips. Always use the insulated shafts.

WARNING

Test the tester first. No matter what kind of tester you're using, test it on a circuit that you know is hot to make sure the tester is working properly.

TESTING WITH AN INDUCTANCE TESTER

An inductance tester is a popular battery-operated voltage tester that is reliable, inexpensive, and fits in a shirt pocket. Its tip glows when touched to a hot (energized) screw terminal or receptacle slot. Each time you use an inductance tester—or any voltage tester—first test its accuracy on a receptacle that you know is hot ❶.

After shutting off the power at the panel, first insert the tester tip into the short (hot) slot of a receptacle ❷. If the tester tip does not glow, there is probably no voltage present. To be sure, next insert the tester tip into the receptacle's long (neutral) slot ❸. This second insertion should protect you in case the receptacle was incorrectly wired.

If you need to remove the receptacle—say, to replace it—remove its cover. Then, being careful not to touch the sides of the receptacle, unscrew the two mounting screws holding the receptacle to the outlet box. (If the box is metal, avoid touching it, too.) Grasp the mounting straps and gently pull the receptacle out of the box. First touch the tip of the tester to the hot (black) wire or the gold screw terminal ❹, then touch the tip to the white (neutral) wire or silver screw terminal ❺. If the tester tip does not glow, it's safe to handle the receptacle and the wires feeding it.

An inductance tester will often glow when its tip is merely near a hot wire—that is, it can "read" current through a wire's insulation. Thus you can sometimes detect electrical current at a switch or fixture without removing the outlet cover ❻.

➔ **For more on testing switches and light fixtures, see Chapters 3 and 4.**

TRADE SECRET

Using an inductance tester is far safer than using a two-prong neon tester because it's possible to get shocked by touching the bare probes of neon testers.

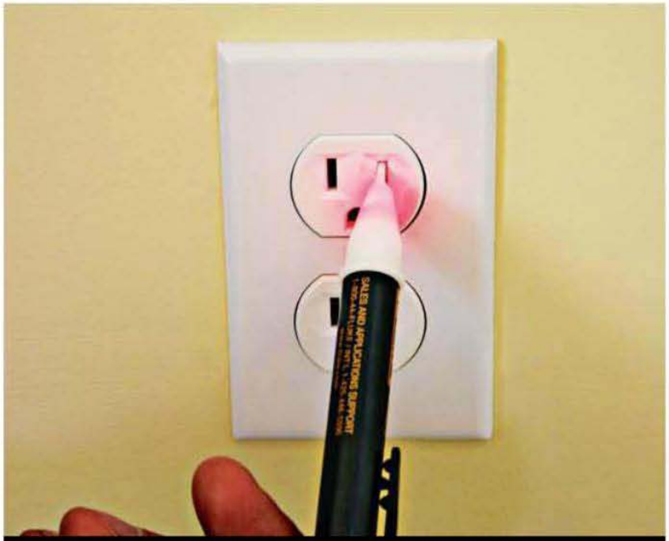

1 Test the tester on an outlet you know to be hot.

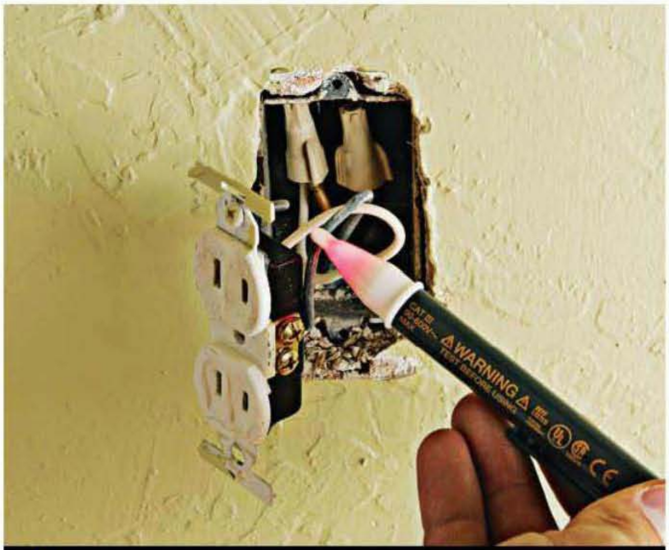

4 Test the hot connector. You can often detect current by touching the insulation, but touch the wire hold-down screw to be sure.

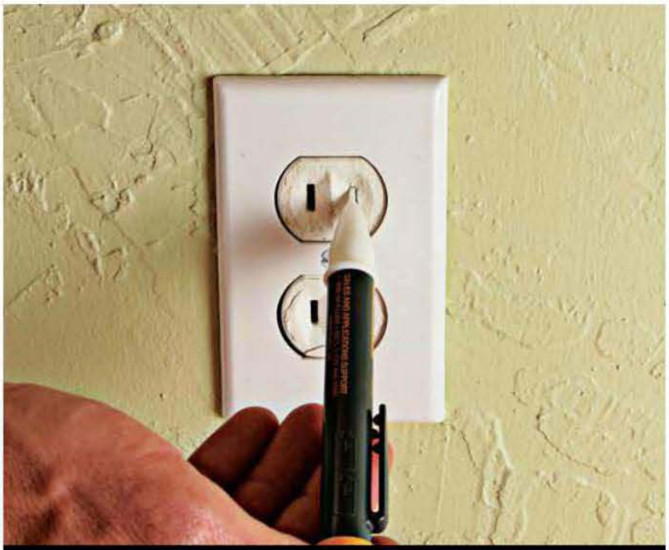

2 When checking to make sure power is off, test the hot slot first.

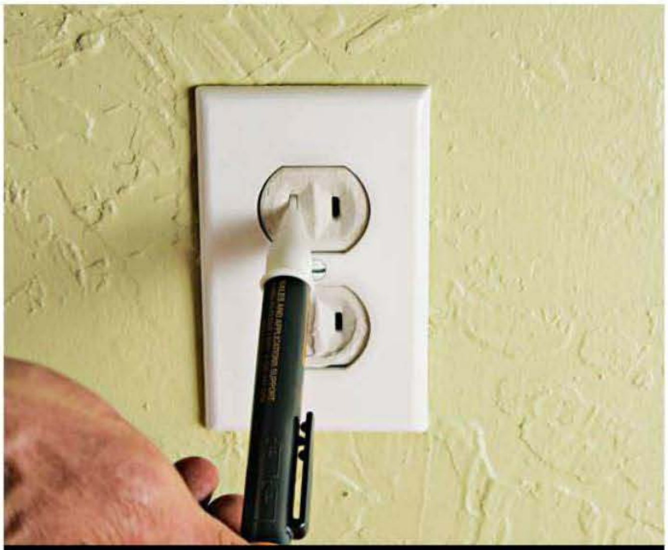

3 Then try the neutral slot.

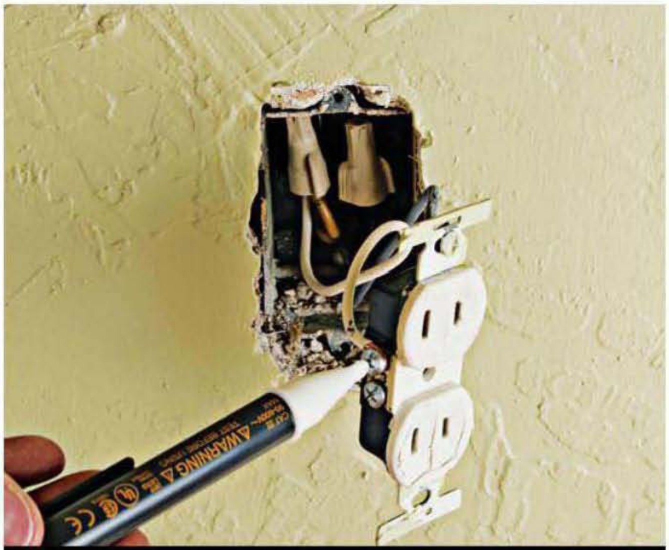

5 Then test the neutral connector.

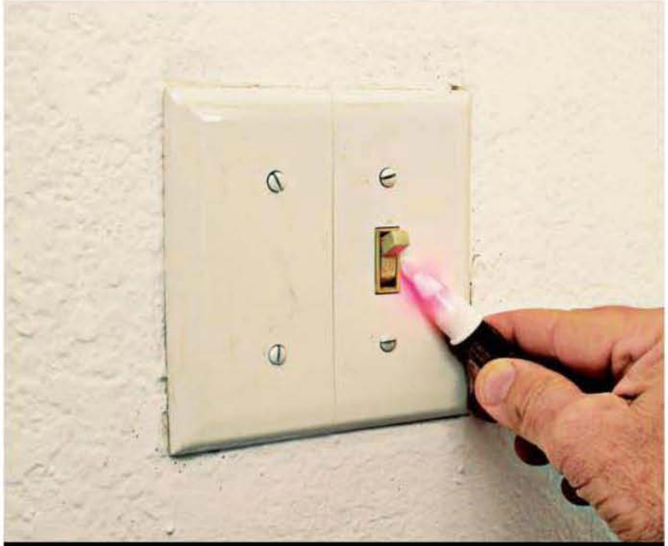

6 You can often detect current through an outlet cover.

TOOLS & MATERIALS

YOU DON'T NEED A LOT OF expensive tools to wire a house successfully. And there's little uniformity among the tools electricians prefer. Some pros carry a dozen different pliers and wire strippers in their tool belts, whereas others streamline their movements and save time by using the fewest tools possible. This chapter introduces the basic tools and materials you'll need and a few of the basic techniques you'll perform repeatedly. All materials should bear the Underwriters' Laboratory (UL) stamp, which indicates that a component meets the safety standards of the electrical industry.

The first test of any tool is to fit your hand comfortably; the second, that it feels solid and well made. Better tools tend to be a bit heftier and cost more.

TOOLS

Hand tools, p. 22
Power tools, p. 23
Electrical testers, p. 24

ELECTRICAL BOXES

Choosing electrical boxes, p. 25
Installing new work boxes, p. 27
Removing knockouts, p. 28
Mounting devices, p. 29

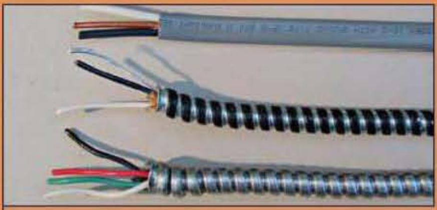

CABLES & WIRES

Cable & conduit, p. 30
Cable clamps & clips, p. 32
Securing cable to framing, p. 33
Stripping cable with a utility knife, p. 34
Stripping cable with a cable ripper, p. 35
Choosing wire, p. 36
Stripping & splicing wires, p. 37

ELECTRICAL

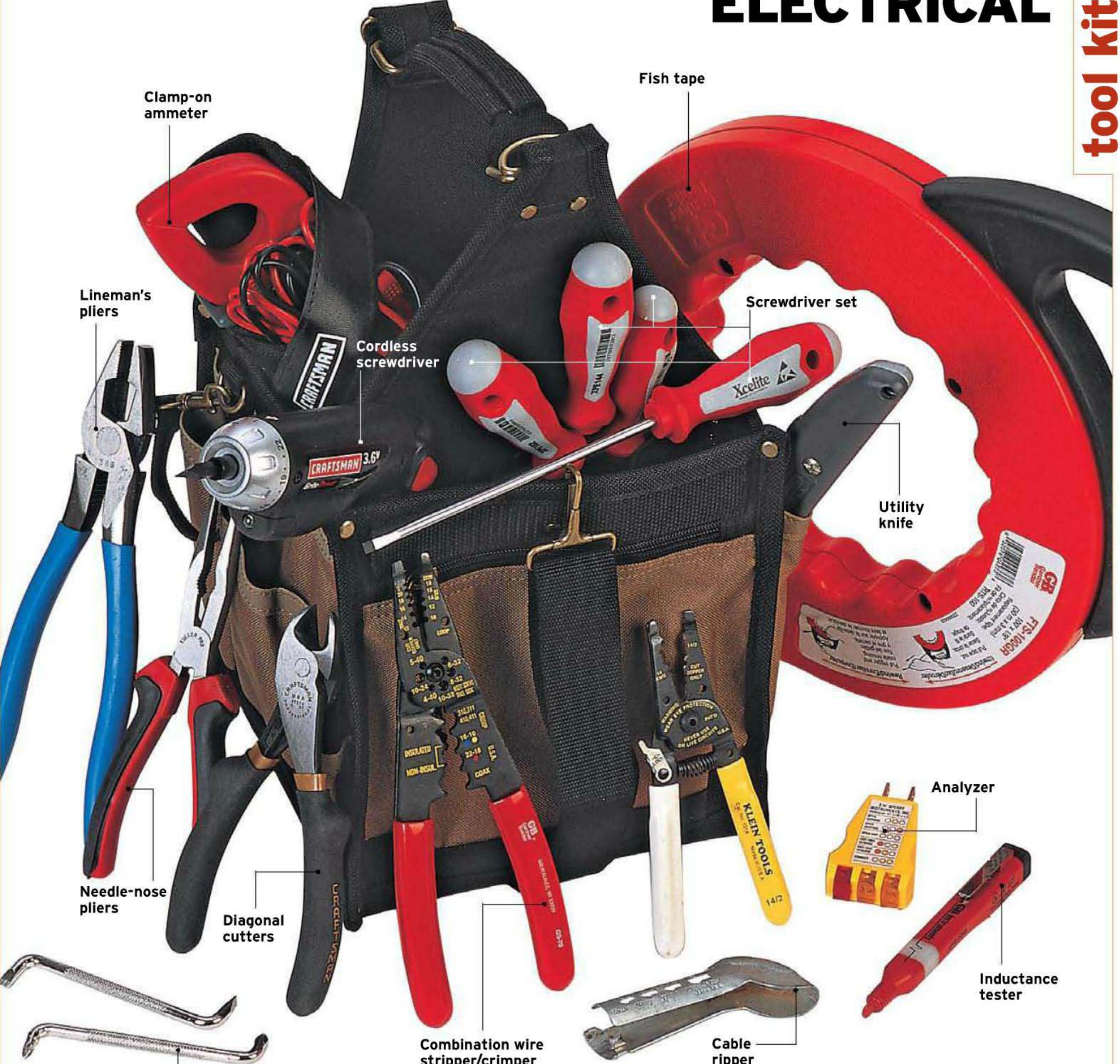
Clamp-on ammeter
Fish tape
Lineman's pliers
Screwdriver set
Cordless screwdriver
Utility knife
Analyzer
Needle-nose pliers
Diagonal cutters
Inductance tester
Combination wire stripper/crimper
Cable ripper
Offset screwdrivers

HAND TOOLS

A tool belt is especially helpful for wiring work. Assign each tool a specific place in the belt and you can easily retrieve it without searching.

Phillips-head screwdrivers

Torpedo level

Retractable tape measure

Drywall saw

7-in-1 tool (useful to free device plates encased in paint)

Hammer

All hand tools should have insulated handles and fit your hand comfortably. Manufacturers now make tools in various sizes; to fit everyone from the largest male to the most petite female. Here again, don't scrimp on quality.

Pliers and strippers

Lineman's pliers are the workhorse of an electrician's toolbox. They can cut wire, hold wires fast as you splice them, and twist out box knockouts. Needle-nose (long-nose) pliers can grasp and pull wire in tight spaces. These pliers can loop wire to fit around receptacle and switch screws. A large pair can also loosen and remove knockouts in metal outlet boxes. Diagonal-cutting and end-cutting pliers can cut wires close in tight spaces; end cutters (sometimes called nippers) also pull out staples easily. A multipurpose or combination stripping tool is used to strip individual wires of insulation, cut wire, crimp connections, and quickly loop wire around screw terminals.

A combination stripping tool not only cuts and strips wire but also crimps connectors, cuts small machine screws, and more.

A cable ripper (see p. 21) strips the plastic sheathing from Romex® cable without harming the insulation on the individual wires inside. Many pros use a utility knife to strip sheathing, but that takes practice and a light touch to avoid nicking the insulation of individual wires. To strip armored cable, use a Roto-Split® *cable stripper* (p. 202); it's vastly superior to the old method of using a hacksaw and diagonal cutters.

Other useful tools

No two electrician's tool belts will look the same, but most contain a tape measure, flashlight, small level, hammer, Speed® Square, and a large felt-tipped marker. In the course of a wiring job, you may need several sizes of slot-head and Phillips®-head screwdrivers, plus an offset screwdriver and a nut driver.

Fish tape is used to run cable behind finish surfaces and pull wire through conduit. A fish tape is invoked in almost every old wiring how-to book on the market. Modern-day pros, however, swear by a pulling grip, also called a swivel kellum (p. 132). For remodel work, you may need a plaster chisel, flat bar, and a drywall saw.

POWER TOOLS

Buy power tools that are appropriate to your strength and to the task at hand. More powerful tools tend to be heavier and harder to manage; and for wiring, they're often overkill. If possible, test-drive a friend's power tool before buying your own.

Drills

A 1/2-in. right-angle drill allows you to fit the drill head between studs or joists and drill perpendicular to the face of the lumber.

→ **To see a right-angle drill in action, see p. 177.**

Buy a drill with a clutch. Unless there's a clutch to disengage the motor, a drill bit that suddenly jams or lodges against a nail shank, could cause the body of the tool to torque powerfully and thus could injure you.

Drill bits

Spade bits cut quickly, but tend to snap in hard wood. For this reason most electricians prefer auger bits. Self-feeding chipper bits drill doggedly through hard, old wood but won't last long if they hit nails. A 7/8-in. Greenlee® Nail Eater™ bit is your best bet if old lumber is nail infested.

Reciprocating saw

A reciprocating saw is indispensable for most remodeling jobs, whether to cut box openings in plaster or to notch framing. Choose a blade that's appropriate for the material you're cutting: coarser teeth for cutting wood, finer for plaster and metal. Special bimetal remodel blades can cut through occasional nails without destroying the blade.

Rotary cutter

Using slightly different bits, a rotary cutter can cut through plaster or drywall to create box openings. Typically, drywallers install panels over installed outlet boxes, then use a rotary cutter to trim around the outside of a box; this method is much quicker than hand-cutting openings beforehand. Set the tool's depth to avoid nicking wires.

Cordless power tools

Cordless drills and saws enable you to keep working when the powers is off or not yet connected. They don't need an extension cord and won't electrocute you if you inadvertently drill or cut into a live wire. Cordless reciprocating saws can cut anything from plaster lath to studs.

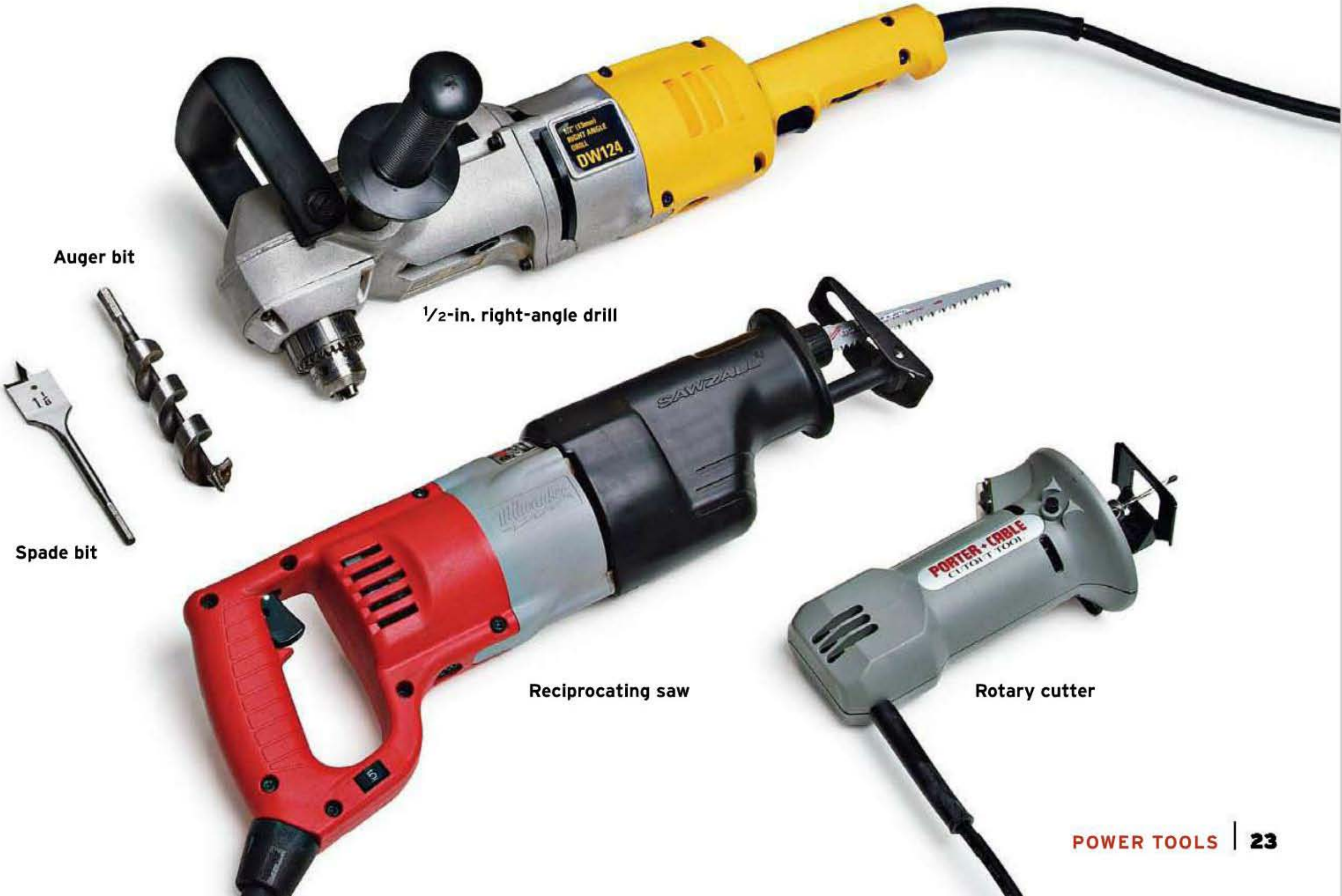

ELECTRICAL TESTERS

Testing to see if a circuit or device is energized is crucial to safety and correct wiring. There are several to choose from, and some perform multiple functions.

Neon voltage testers

Neon voltage testers are inexpensive and widely available, but there's a danger of touching the bare metal probes and getting shocked. Better neon testers have insulated handles. To use this tool, insert the probes into the receptacle slots or touch them to the screw terminals or to a metal outlet box to see if the unit is hot (energized).

Plug-in circuit analyzers

Plug-in circuit analyzers can be used only with three-hole receptacles, but they quickly tell you if a circuit is correctly grounded and, if not, what the problem is. Different light combinations on the tester indicate various wiring problems, such as no ground and hot and neutral reversed. They're quite handy for quick home inspections.

Solenoid voltage testers

Solenoid voltage testers (often called *wiggies*) test polarity as well as AC voltage. They also test DC voltage from 100v to 600v. Most models vibrate and light a bulb when current is present. Solenoid testers don't use batteries, so readings can't be compromised by low battery power. However, because of their low impedance, solenoid testers will trip ground-fault circuit interrupters (GFCIs).

Inductance testers

Inductance detectors provide a reading without directly touching a conductor. They often allow you to detect electrical currents without having to remove cover plates and expose receptacles or switches. Touch the tool's tip to an outlet, a fixture screw, or an electrical cord. If the tip glows red, it means there's voltage present. Inductance testers rely on battery power.

Multimeters

A multimeter, as the name suggests, offers precise readings in multiple scales, which you select beforehand. Some models are even *autoranging*, meaning that they select the correct scale for you. Extremely sensitive, multimeters can detect minuscule amounts of voltage. Better models test AC and DC voltage, resistance, continuity, capacitance, and frequency.

Continuity tester

In addition to voltage testers, get a continuity tester to test wire runs and connectors for short circuits or other wiring flaws. And be sure to do all your testing before a circuit is connected to power.

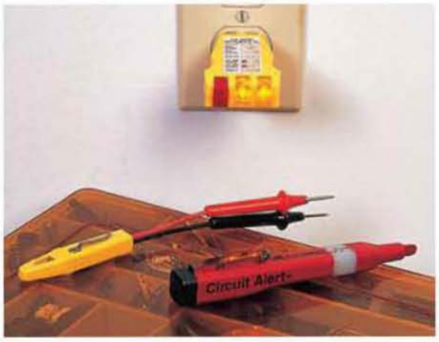

Some common electrical testers. A circuit analyzer (top) shows problems in the wiring of a 3-pronged plug. Test whether an outlet is energized with a neon voltage tester (left) or check for power with an inductance tester (right).

Digital multimeters provide precise readings in multiple scales.

Solenoid voltage testers don't require batteries to give a reading.

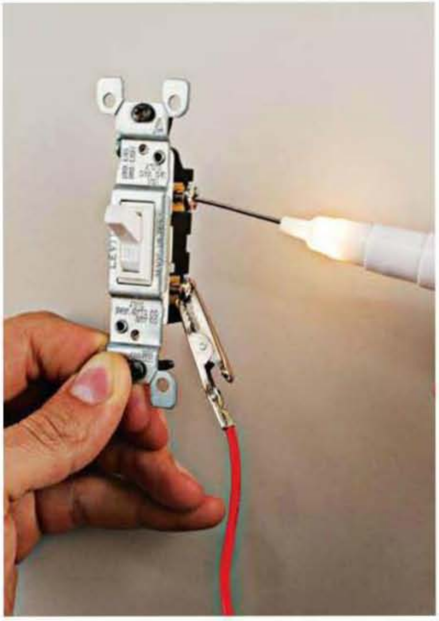

A continuity tester enables you to determine whether an electrical device or length of wire is defective.

CHOOSING ELECTRICAL BOXES

There is a huge selection of boxes, varying by size, shape, mounting device, and composition. But of all the variables to consider when choosing boxes, size (capacity) usually trumps the others. Correctly sized boxes are required by Code and faster to wire because you don't have to struggle to fit wires and devices.

Box capacity

The most common shape is a single-gang box. A single-gang box 3½ in. deep has a capacity of roughly 22½ cu. in.; enough space for a single device (receptacle or switch), three 12–2 w/grd cables, and two wire connectors. Double-gang boxes hold two devices; triple-gang boxes hold three devices. *Remember:* Everything that takes up space in a box must fit without cramping—devices, cable wires, wire connectors, and cable clamps—so follow NEC recommendations for the maximum number of conductors per box.

You can get the capacity you need in a number of ways. Some pros install shallow four-squares (4 in. by 4 in. by 1½ in. deep) throughout a system because such boxes are versatile and roomy. If a location requires a single device, pros simply add a mud-ring cover.

Because of their shallow depth, these boxes can also be installed back to back within a standard 2x4 wall. This allows you to keep even back-to-back switch boxes at the same height from one room to the next. Shallow pancake boxes (4 in. in diameter by ½ in. deep) are commonly used to flush-mount light fixtures.

Metal vs. plastic boxes

Metal boxes are sturdy and are available in more sizes than are plastic boxes. Some metal boxes can be interlocked for larger capacity. Also, metal boxes are usually favored for mounting ceiling fixtures because steel is stronger than plastic. If Code requires steel conduit, armored cable (BX), or MC cable, you *must* use steel boxes. All metal boxes must be grounded.

» » »

Single-gang plastic

Double-gang plastic

Triple-gang plastic

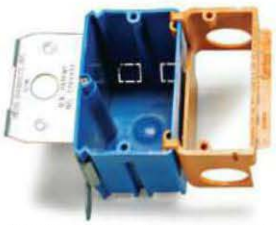

Single-gang adjustable with (orange) snap-on data ring

Double-gang adjustable

Single-gang metal

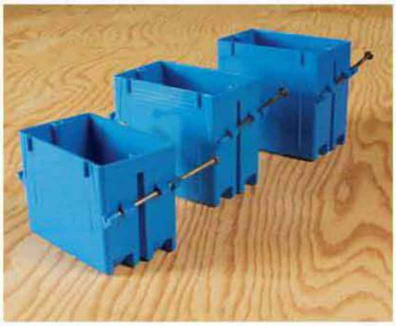

Single gang boxes come in three sizes, 18 cu. in., 20.4 cu. in. and 22.5 cu. in.

Throw a single or double gang mud (plaster) ring on a 4-in. box and it's hard to overfill.

BOX FILL WORKSHEET[*]

Item	Size (cu. in.)	Number	Total
#14 conductors exiting box	2.00		
#12 conductors exiting box	2.25		
#10 conductors exiting box	2.50		
#8 conductors exiting box	3.00		
#6 conductors exiting box	5.00		
Largest grounding device; count only one		1	
Devices; two times connected conductor size			
Internal clamps; one based on largest wire present		1	

[*] Table based on NEC 370-16(b) and adapted with permission from Redwood Kardon, Douglas Hansen, and Mike Casey, *Code Check: Electrical* (2005, The Taunton Press).

CHOOSING ELECTRICAL BOXES (CONTINUED)

For most other installations, plastic is king. (Plastic boxes may be PVC, fiberglass, or thermoset.) Electricians use far more plastic boxes because they are less expensive. Also, because they are nonconductive, they're quicker to install because they don't need to be grounded. However, even if a box doesn't need to be grounded, all electrical devices inside must be grounded by a continuous ground. Another reason to buy plastic: Box volumes are stamped on the outside.

Cut-in boxes

The renovator's mainstay are *cut-in boxes* because they mount directly to finish surfaces. These boxes are indispensable when you want to add a device but don't want to destroy a large section of a ceiling or wall to attach it to the framing. Most cut-in boxes have plastic ears that keep them from falling into the wall cavity; what vary are the tabs or mechanisms that hold them snug to the back side of the wall: screw-adjustable ears, metal-spring ears, swivel ears, or bendable metal tabs (Grip-Lok™ is one brand).

➔ For information on installing remodel boxes, see p. 194.

WARNING

All cut-in boxes, whether plastic or metal, must contain cable clamps inside that fasten cables securely. That is, it's impossible to staple cable to studs and joists when they are covered by finish surfaces, so you need clamps to keep the cables from getting tugged or chafed.

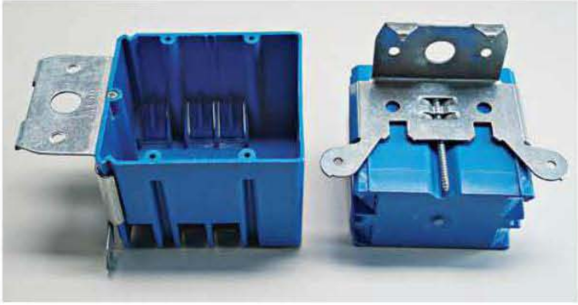

The screw on the side of an adjustable box enables you to raise or lower the face of the box to make it flush to the finish wall.

A remodel box doesn't mount to a stud. Instead, a pair of ears flips up at the turn of a screw and clamps the box to the wall.

CUT-IN REMODEL BOXES

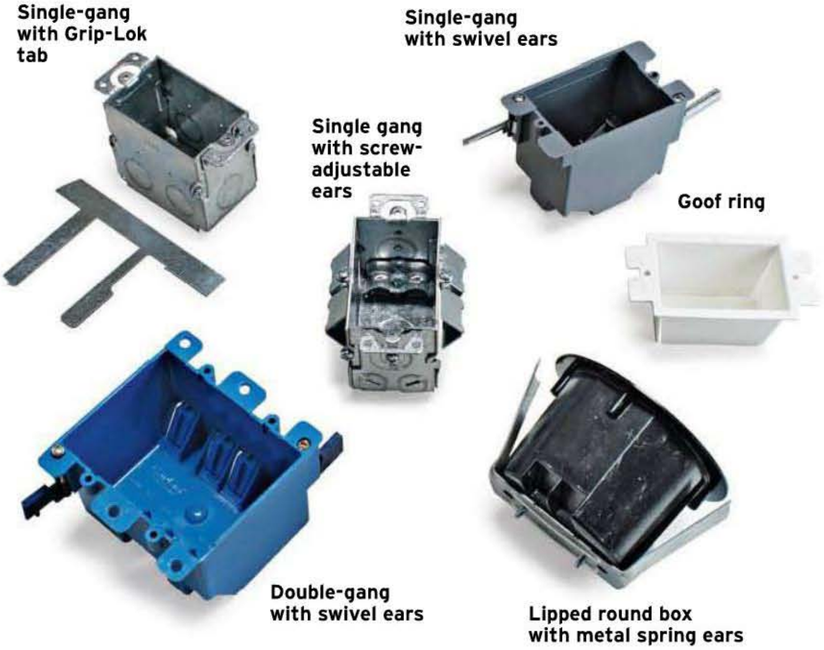
Single-gang with Grip-Lok tab
Single-gang with swivel ears
Single gang with screw-adjustable ears
Goof ring
Double-gang with swivel ears
Lipped round box with metal spring ears

INSTALLING NEW WORK BOXES

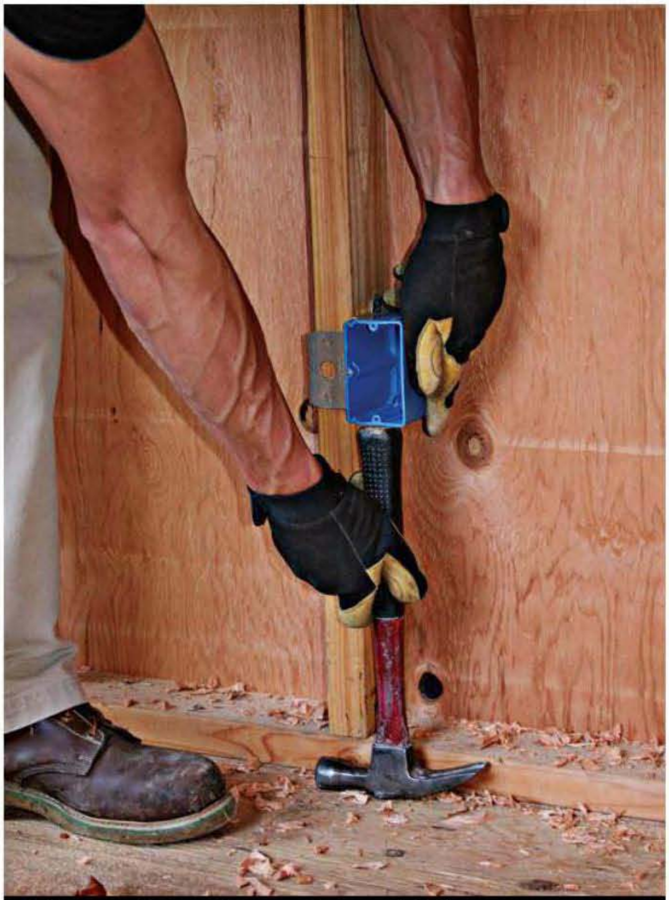

1 For convenience, set boxes at hammer height.

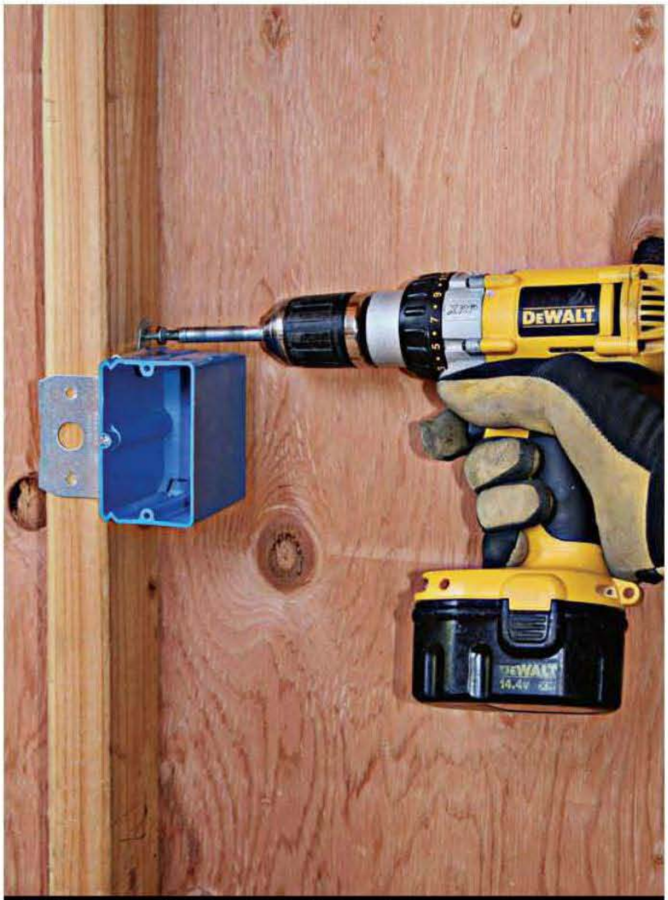

2 Screen an adjustable box to the framing.

There are few set rules about locating boxes. In general-use rooms, set the bottom of outlet boxes 12 in. above the subfloor—which you can approximate by resting a box atop a hammer held on end ❶. In housing for disabled occupants, outlet box bottoms should be a minimum of 18 in. above the subfloor.

For outlets over kitchen and bath counters, set box bottoms 42 in. from the finished floor, so they'll end up 8 in. above counters and 4 in. above a standard 4-in.-high backsplash. Locate wall switches on the lock side of a door (opposite the hinges).

→ **For more about locating boxes, see p. 184.**

Mount boxes so that they'll be flush with finish surfaces. Most boxes have tabs or gauge marks stamped on the side to indicate different surface thicknesses. If that's not the case, hold a scrap of the finish material—for example, 5/8-in. drywall—next to the front edge of the box as a depth gauge. Unless you're installing nail-in boxes, use screws to mount the boxes so you can make adjustments if you need to ❷. As noted elsewhere, adjustable boxes can be tweaked after the drywall is up.

ACCORDING TO CODE

Before positioning outlet boxes, check to see if local building codes require them to be set at a certain height.

REMOVING KNOCKOUTS

Once you've mounted boxes, you'll need to remove the appropriate number of box knockouts and install cable connectors (clamps). Single-gang, new construction plastic boxes don't need clamps: Simply strike a screwdriver handle with the heel of your hand to drive out the knockout. To remove a metal-box knockout, jab it with the nose of needle-nose pliers to loosen it Ⓐ, then use the plier's jaws to twist it free Ⓑ.

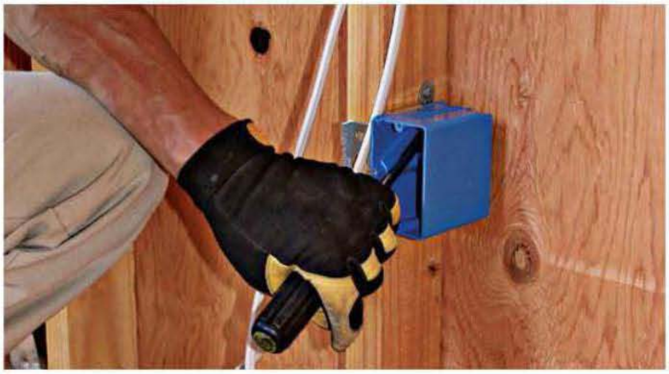

Use a screwdriver to remove a plastic-box knockout.

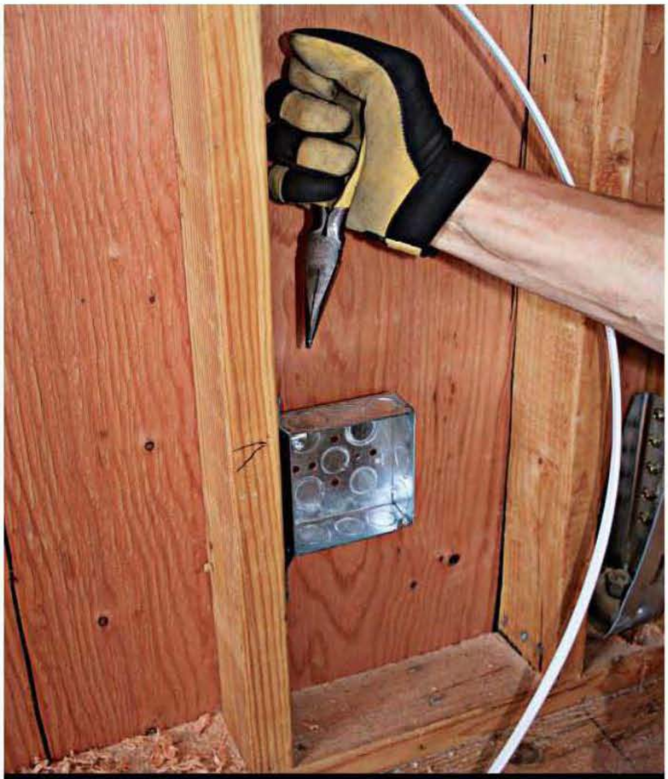

A Strike a metal knockout to loosen it.

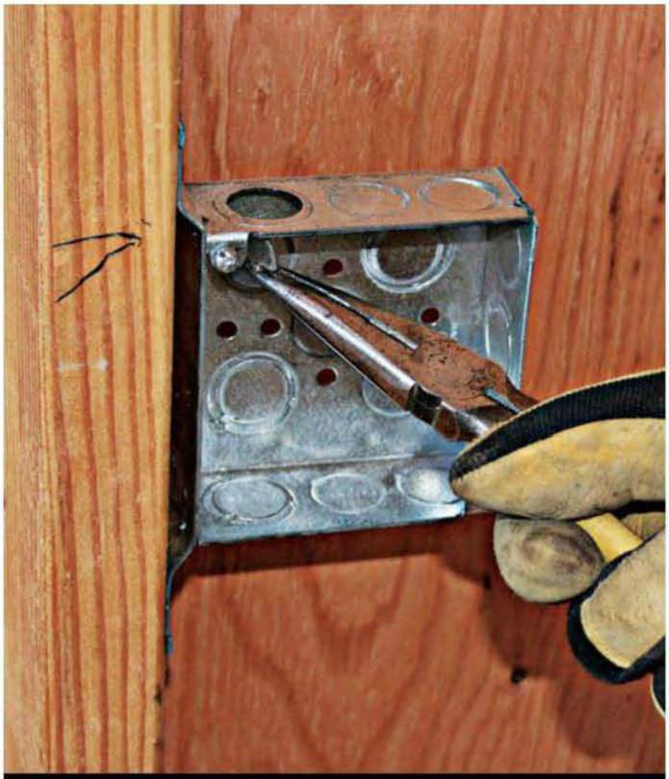

B Once the knockout is loose, remove it using pliers.

MOUNTING DEVICES

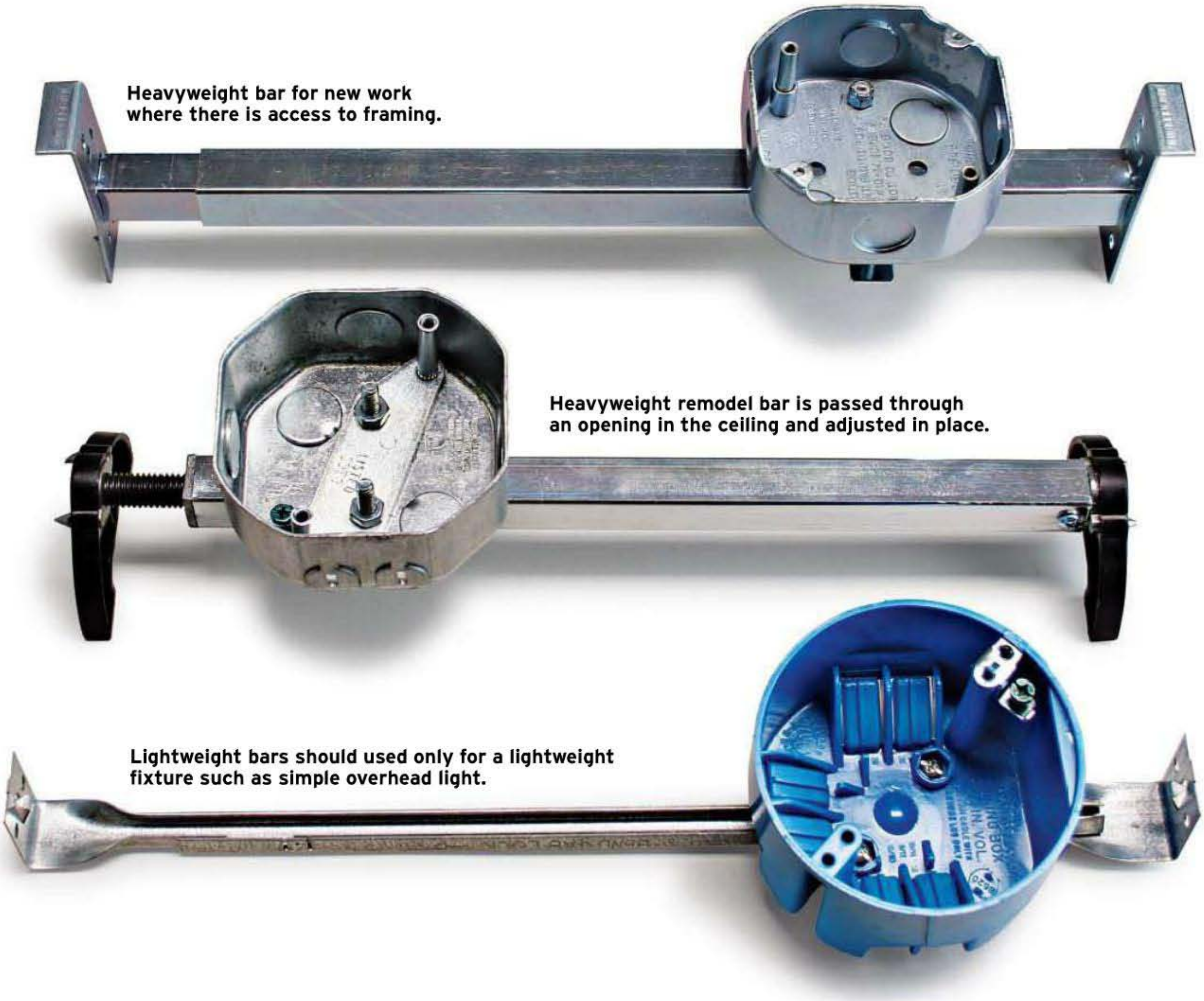

Heavyweight bar for new work where there is access to framing.

Heavyweight remodel bar is passed through an opening in the ceiling and adjusted in place.

Lightweight bars should used only for a lightweight fixture such as simple overhead light.

The type of mounting bracket, bar, or tab you use depends on whether you're mounting a box to finish surfaces or structural members. When you're attaching a box to an exposed stud or joist, you're engaged in new construction, or *new work* (as distinguished from *old work*, or remodel work), even if the house is old. New-work boxes are usually side-nailed or face-nailed through a bracket; nail-on boxes have integral nail holders. The mounting bracket for adjustable boxes is particularly ingenious. Once attached to framing, the box depth can be screw-adjusted until it's flush to the finish surface.

Adjustable bar hangers enable you to mount boxes between joists and studs; typically, hangers adjust from 14 in. to 22 in. Boxes mount to hangers via threaded posts or, more simply, by being screwed to the hangers. Bar hangers vary in thickness and strength, with heavier strap types favored to support ceiling fans and heavier fixtures.

A mounting bar is screwed into the ceiling joists.

CABLE & CONDUIT

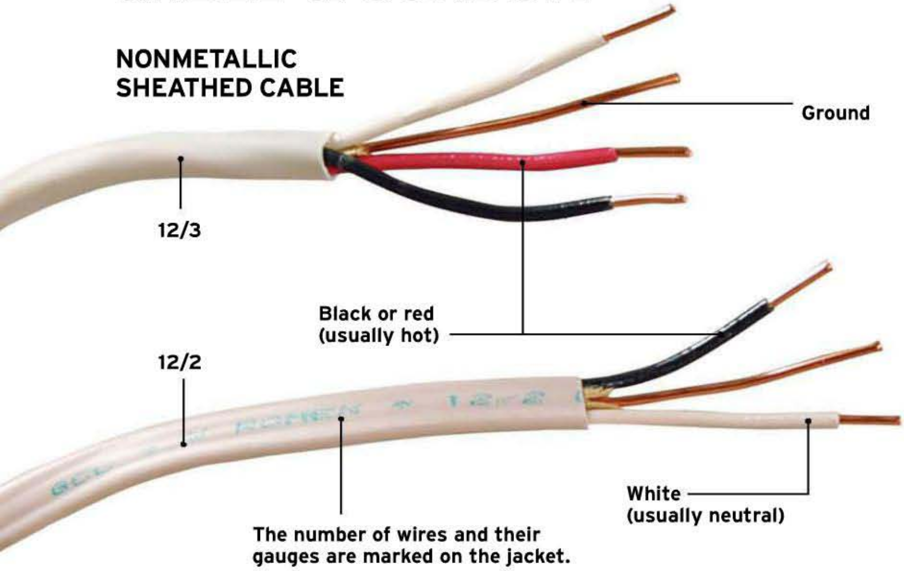

Some old houses still have knob-and-tube wiring, which doesn't necessarily need to be replaced. Have it tested to make sure it's still in good shape.

Most modern house wiring is flexible cable, but you may find any—or all—of the wiring types described here in older houses. Inside cables or conduits are individual wires, or conductors, that vary in thickness according to the load (amperage) they carry. Here's a quick overview.

Cable

Nonmetallic sheathed cable (NM or Romex) is by far the most common flexible cable. Covered with a flexible thermoplastic sheathing, Romex is easy to route, cut, and attach. Cable designations printed on the sheathing and the sheathing color indicate the gauge and the number of conducting wires inside.

➔ **See "Reading a Cable," on the facing page.**

Most of the NM cable used for a standard 15-amp lights-and-outlets circuit is 14/2 w/grd. For a 20-amp circuit, use 12-gauge cable. Three-way switches are wired with 14/3 or 12/3. The third conductor is red. Either the red or the black wire can be hot, depending on the switch position.

➔ **For more on three- and four-way switches, see p. 51.**

Remember that you can wire 15-amp circuits with 12-gauge wire, but you can't use 14-gauge wire for 20-amp circuits. Metal-clad (MC) cable is often specified where wiring is exposed. Some codes still allow armored cable (AC), but that's increasingly rare.

Knob-and-tube

Knob-and-tube wiring is no longer installed, but there's still plenty of it in older houses. If its sheathing is intact and not cracked, it may still be serviceable. You may even be able to extend it, but have an electrician do the work. Knob and tube is eccentric, requiring experience and a skilled hand.

Circuit wiring

Copper is the preferred conductor for residential circuit wiring. Aluminum cable is frequently used at service entrances, but it is not recommended for branch circuits. Individual wires within a cable or conduit are color coded. White or light gray wires are neutral conductors. Black or red wires denote hot, or load-carrying, conductors. Green or bare (uninsulated) wires are ground wires, which must be connected continuously throughout an electrical system.

Because most of the wiring in a residence is 120v service, most cables will have three wires: two insulated wires (one black and one white) plus a ground wire, usually uninsulated. Other colors are employed when a hookup calls for more than two wires—for example, 240v circuits and three- or four-way switches.

Conduit

Conduit may be specified to protect exposed wiring; it is commonly thin-wall steel (electrical metallic tubing; EMT), aluminum, or polyvinyl chloride (PVC) plastic. Metal conduit serves as its own ground. Apart from service entrances, conduit is seldom used in home wiring. When connected with weather-tight fittings and boxes, conduit can be installed outdoors.

➔ **For more on 240v circuits, see "Appliances," pp. 210–221.**

CABLE AND CONDUIT FOR SERVICE PANELS

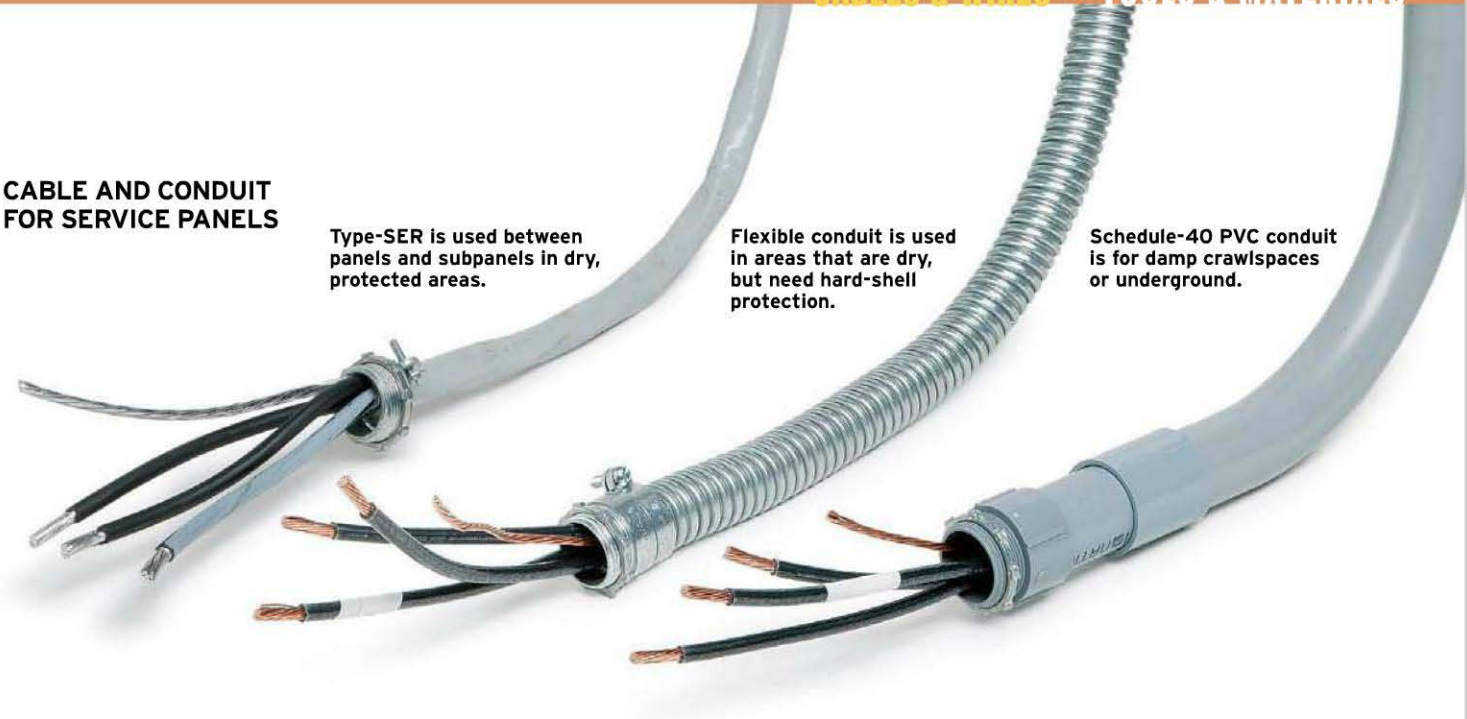

Type-SER is used between panels and subpanels in dry, protected areas.

Flexible conduit is used in areas that are dry, but need hard-shell protection.

Schedule-40 PVC conduit is for damp crawlspaces or underground.

READING A CABLE

Cables provide a lot of information in the abbreviations stamped into their sheathing—for example, *NM* indicates nonmetallic sheathing, and *UF* (for underground feeder) can be buried. The size and number of individual conductors inside a cable are also noted: *12/2 w/grd* or *12-2 W/G*, for example, indicates two insulated 12AWG wires plus a ground wire. Cable stamped *14/3 W/G* has three 14AWG wires plus a ground wire. (The higher the number, the smaller the wire diameter.) The maximum voltage, as in 600v, may also be indicated.

Individual wires within cable have codes, too. *T* (thermoplastic) wire is intended for dry, indoor use, and *W* means wet; thus *TW* wire can be used in dry and wet locations. *H* stands for heat resistant. *N*, for nylon jacketed, indicates a tough wire than can be drawn through conduit without being damaged.

Finally, make sure the Romex cable is marked *NM-B*. Cable without the final *B* has an old-style insulation that is not as heat resistant as NM-B cable.

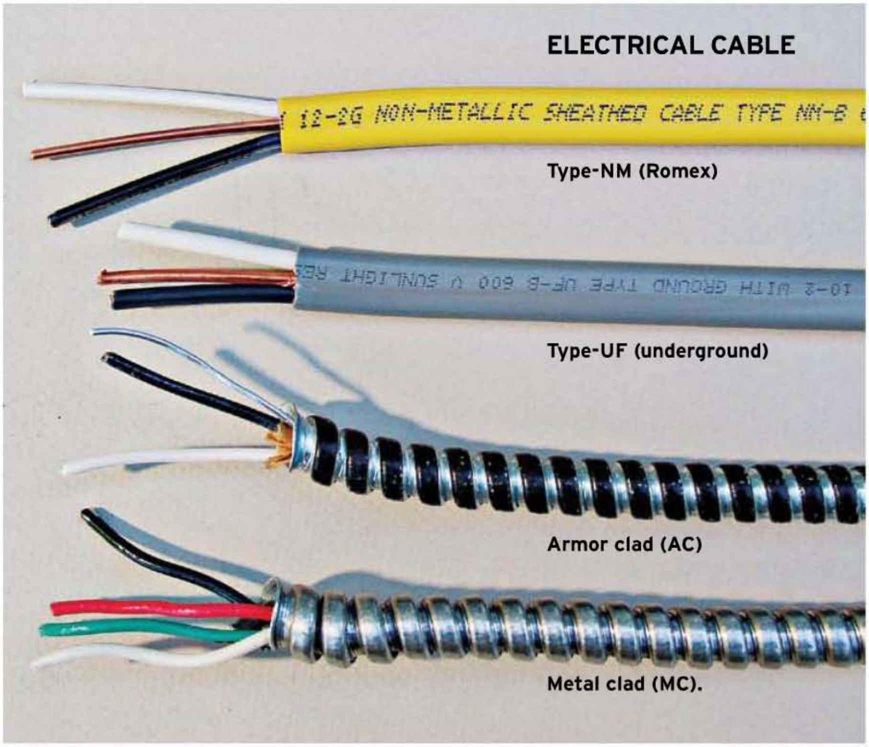

The silver wire in the AC cable is a bonding wire, not a ground. In the MC cable, the green wire is ground, the white is neutral, and the red and black are hot.

CABLE CLAMPS & CLIPS

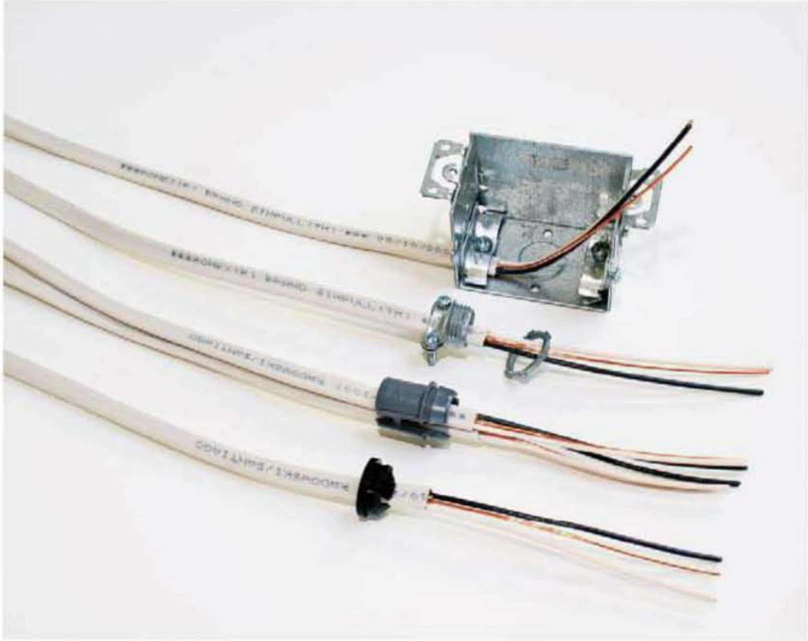

Romex cable connectors. **From bottom to top: plastic push-in connector, two-cable hit-lock connector, $^3/_8$-in. NM clamp with locknut, metal box with internal clamps. Cable connectors are set in box knockouts to prevent wires from wearing against sharp edges.**

A A metal locknut connector consists of two separate pieces.

B Plastic snap-in connectors snap easily into place.

It's important to secure cable as you run it and to protect it from puncture or being accidently severed. It's also essential to make sure cable is tightly secured into boxes.

Clamps

Clamps secure cable to boxes to protect connections inside the box so wire splices or connections to devices cannot get yanked apart or otherwise compromised. Every wiring system—whether nonmetallic (Romex), MC, or conduit—has clamps (connectors) specific to that system. Cable clamps in metal boxes also keep wires from being nicked by burrs created when metal box knockouts are removed. (Use a screwdriver to start knockouts and lineman's pliers to twist them free—or a pair of heavy-duty needlenose pliers for both tasks.)

The exception to this rule is single-gang plastic boxes. If framing is exposed (new construction) and cable can be stapled within 12 in. of the box, Code doesn't require cable clamps in a single-gang plastic box. However, two-gang plastic boxes must have cable clamps—typically, a plastic tension clip that keeps cables from being pulled out. And, as noted in the preceding chapter, all cut-in boxes must contain cable clamps.

Two-piece locknut connectors Ⓐ are still the most common cable clamp, but professional electricians who are racing the clock swear by plastic snap-in cable connectors Ⓑ, which seat instantly and grip NM cable tight.

➔ **For more on positioning boxes, see p. 27.**

WARNING

Cordless drills and screwdrivers reduce the tedium of screwing wires to terminals, attaching devices to boxes, putting on cover plates, and connecting myriad other items. But always tighten cable clamps by hand to avoid overtightening them and damaging the incoming wires.

SECURING CABLE TO FRAMING

The quickest way to secure cable is to staple it. The trick is to staple it correctly—flat and not so tight that the sheathing is squeezed against the framing. Staples should be snug but not too tight. Use enough staples to secure cable, remembering that code requires staples at least every 54 in. Use particular care when stapling cable overhead ①. Avoid making a sharp bend immediately after a staple, and leave yourself enough slack around boxes.

When a large number of cables run into a single box, it can be difficult to staple them within the 12 in. required by Code. Here's where cable clips come in handy. Simply nail or screw the clip to the stud where the box is attached ②. Thread the cables into the clip for a neat, organized box ③.

There are several styles of cable clips with different capacities. Avoid exceeding the capacity of the clip; it should hold the cable snugly but be loose enough to dissipate heat.

ACCORDING TO CODE

Cables should be fastened to framing at least every 54 in. Cables must also be fastened within 12 in. of a box.

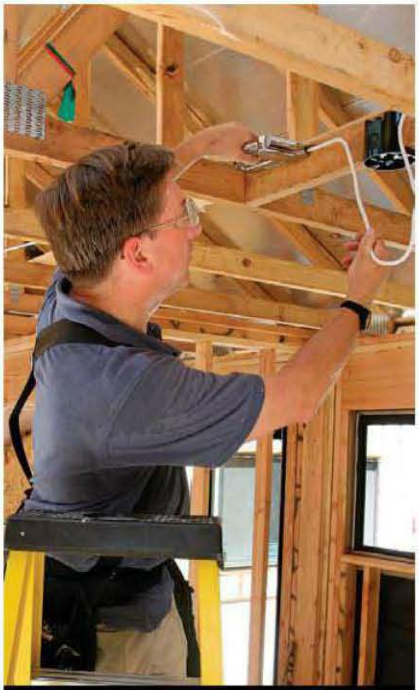

1 Staple snugly but not tightly enough to squeeze the sheathing.

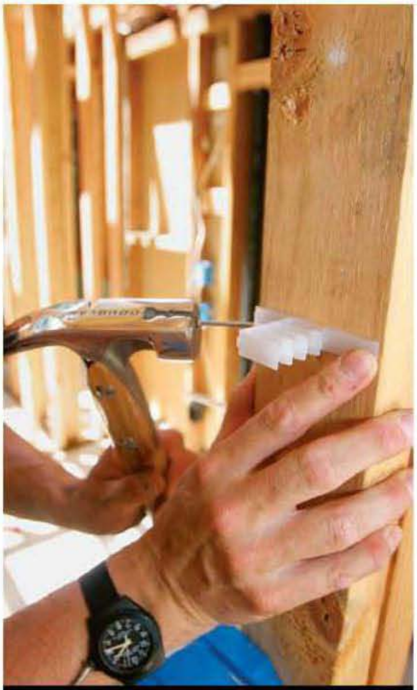

2 Nail cable clips within 12 in. of the box the cables will enter.

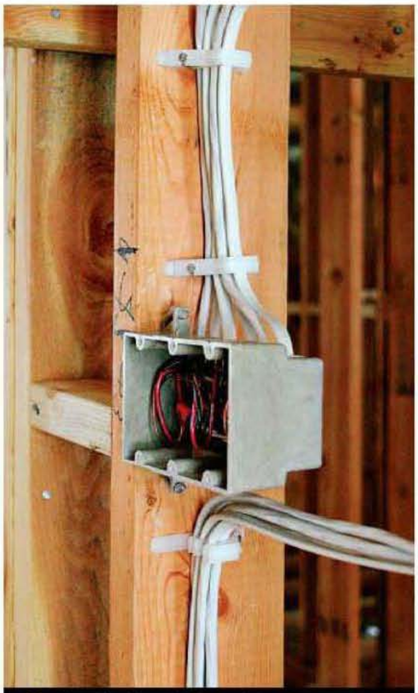

3 Neatly thread the cable into the clip.

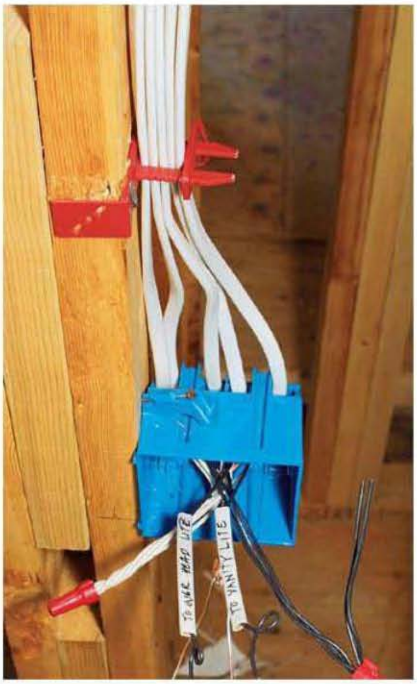

Another style of cable clip. It holds the cable near enough to the box but out of harm's way.

STRIPPING CABLE WITH A UTILITY KNIFE

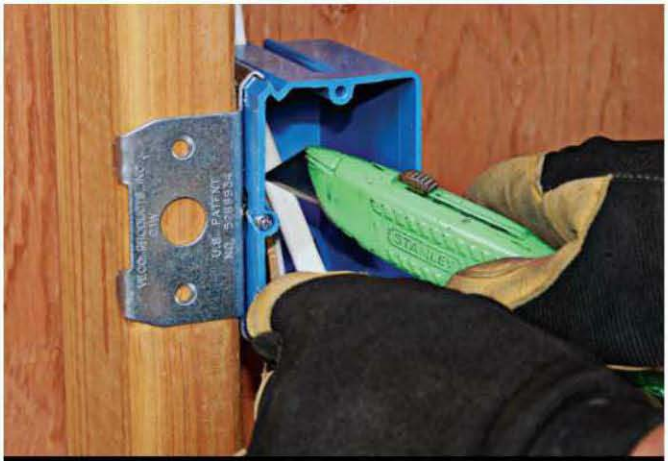

1 Score the front and back of the sheathing.

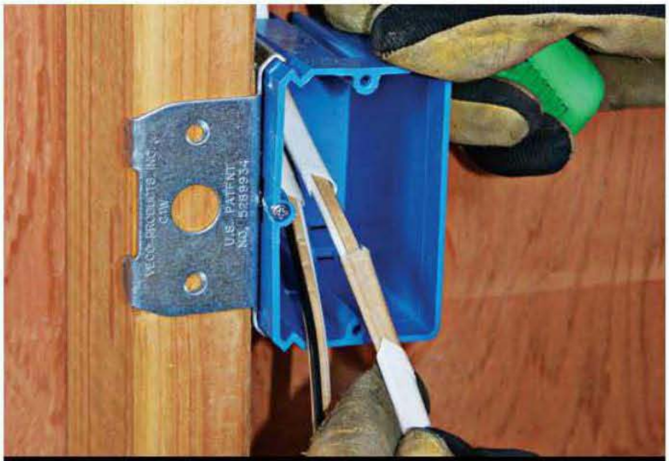

2 Slide the sheathing off of the cable.

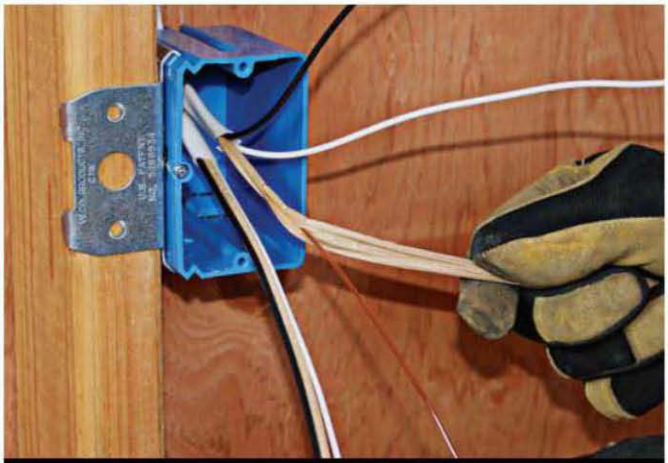

3 Remove the kraft paper covering the ground wires, and you're ready to make connections.

Many electricians use a utility knife to slit and remove NM cable sheathing, but it requires a light touch that takes a lot of practice. Typically, pros hold the blade at a low angle to the cable and lightly run the blade tip down the middle of the cable and over the bare ground wire inside. Alternatively, one can score the front and back face of the sheathing with diagonal slits ❶ and then yank the sheathing and slide it off ❷.

Once the sheathing is off, tear off the kraft paper covering the bare ground wires ❸. Then tuck the cables back into the box until you're ready to wire the receptacle or switch.

Cable strippers remove the wire's jacket without damaging the conductor insulation.

STRIPPING CABLE WITH A CABLE RIPPER

To remove sheathing with less risk of nicking wire insulation, use a cable ripper to slit the sheathing along its length ❶. Because the ripper's tooth is intentionally dull (so it won't nick wire insulation), it usually takes several pulls to slit the sheathing completely. Once that's done, pull back the sheathing and the kraft paper. You can snip both off using diagonal cutters ❷.

Because cable clamps grip sheathing—not individual wires—there should be at least 1/2 in. of sheathing still peeking out from under cable clamps when you're done. If you didn't tighten cable clamps earlier, do so now.

If there is only one cable entering a box, simply cut individual wires to length (typically, 8 in.) and tuck them into the box. If the box is metal, first bond the cable's ground wire to the box, using a grounding clip or a green grounding screw ❸. Once the wires are stripped and the box is grounded, fold the rest of the wires back into the box until you're ready to wire switches and outlets ❹.

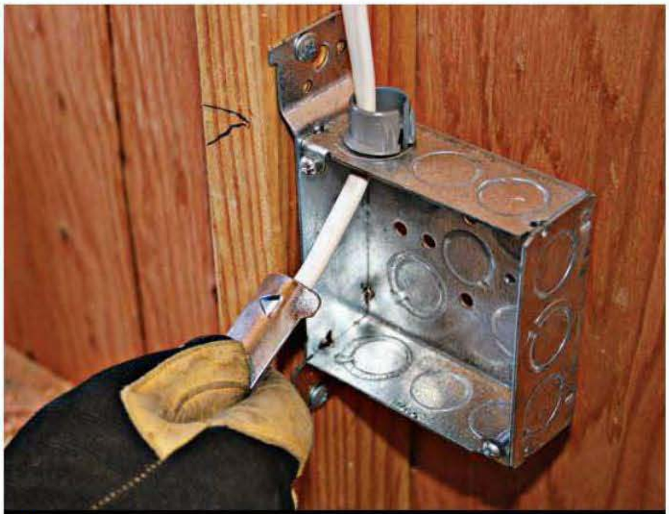

1 Pull a cable ripper along the length of the last few inches of cable.

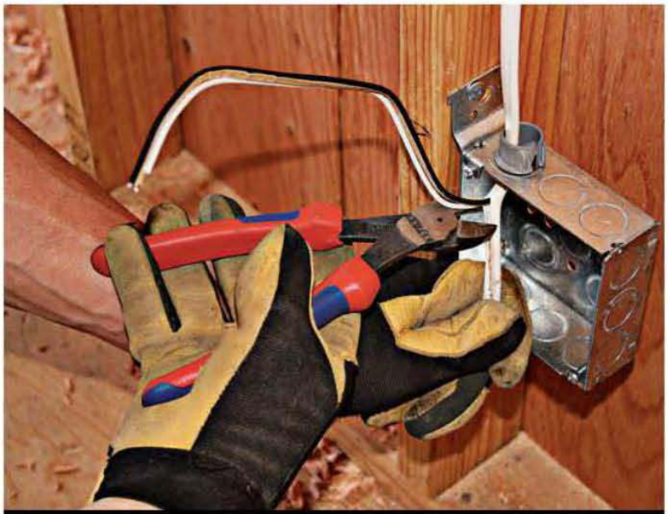

2 Snip both the sheathing and the paper.

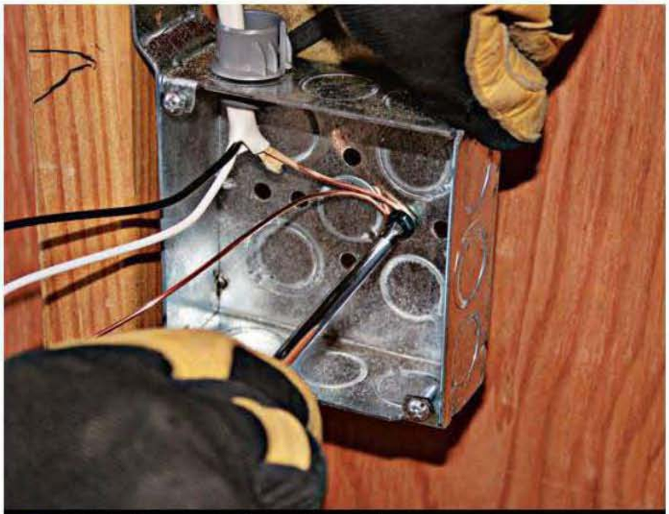

3 Attach the ground wire to the metal box.

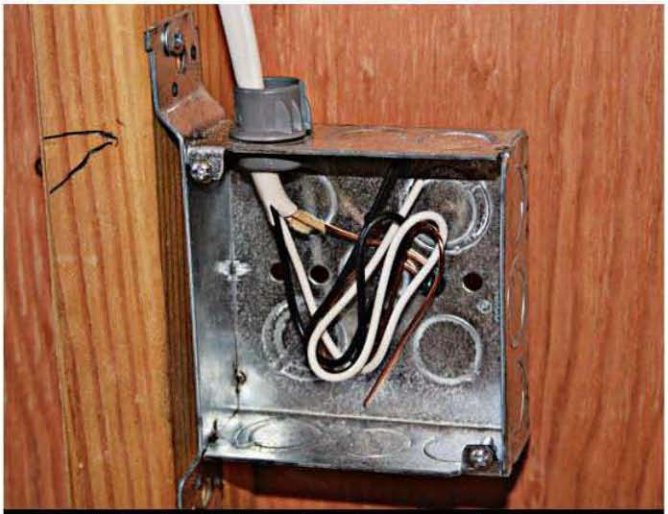

4 Tuck wires into the box until you're ready to wire the outlet.

CHOOSING WIRE

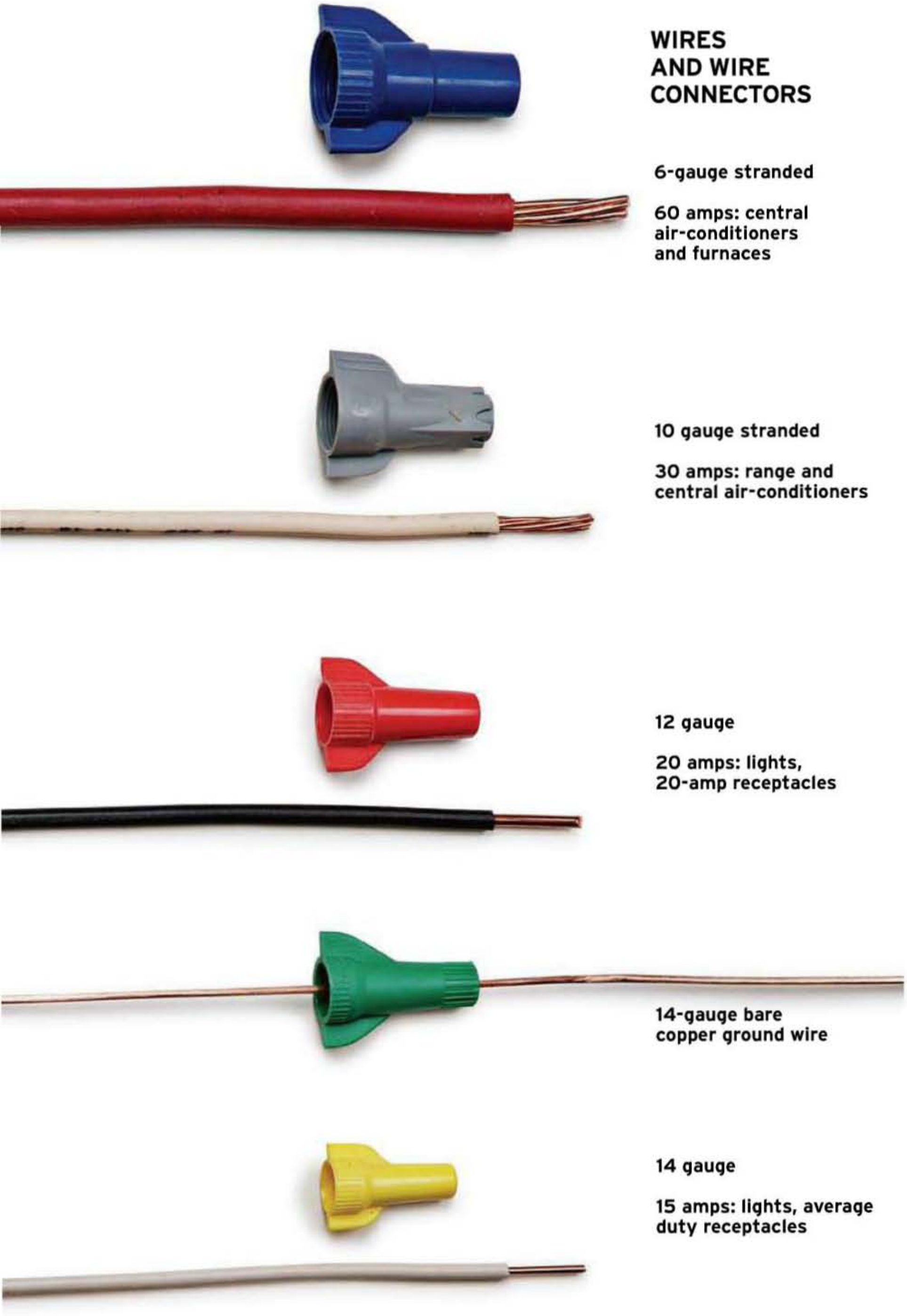

WIRES AND WIRE CONNECTORS

6-gauge stranded

60 amps: central air-conditioners and furnaces

10 gauge stranded

30 amps: range and central air-conditioners

12 gauge

20 amps: lights, 20-amp receptacles

14-gauge bare copper ground wire

14 gauge

15 amps: lights, average duty receptacles

Twist-on wire connectors are color coded to fit wires of different sizes. Green connectors, used to splice ground wires, have a hole in the cap that facilitates running a bare ground wire to a device or a metal box.

Wire comes in several gauges meant for different amp circuits. The higher the gauge, the smaller the wire. Larger wires can carry greater amperage, just as a larger pipe can carry greater water volume. If you use too small a wire, the resistance (measured in ohms) is too great and the wire can melt, causing a house fire. That's why it's important to use the right gauge wire for the load.

Wire connectors

Wire connectors, sometimes called by the popular brand Wire-Nut®, twist onto a group of like-colored wires to splice them together and ensure a solid mechanical connection. The importance of solid connections between spliced wires (or between wires and devices) can't be overstated. If wires work loose, electricity can leap the gaps between them and cause a house fire. Wire connectors are sized according to the number of wires and/or wire gauge they can accommodate; each size is color coded.

A divided pouch transforms a 5-gal. bucket into a portable hardware store of wire connectors, cable clamps, screws, staples, and other small items.

STRIPPING & SPLICING WIRES

Typically, electricians first splice the ground wires, which are usually bare copper. (If they're green insulated wires, first strip approximately ¾ in. of insulation off their ends.) If you use standard wire connectors, trim the ground wires and butt their ends together, along with a 6-in. pigtail, which you'll connect later to the green ground screw of a receptacle. However, many pros prefer to twist the ground wires together, leave one ground long, and thread it through the hole in the end of a special wire connector ❶. Splicing hot and neutral wire groups is essentially the same. Trim hot wires to the same length ❷. Strip ¾ in. of insulation off the cable wires and the pigtail ❸, and use lineman's pliers to twist the wires ❹. Then screw on a wire connector ❺.

When all the wire groups are spliced, gently fold the wires—rather like an accordion fold—and push them into the back of the box, where they'll be safe from drywall saws and nails ❻.

TRADE SECRET

Unless the manufacturer's instructions say otherwise, whenever you splice solid wires with a wire connector, twist the wires together before you twist the connector into place. This guarantees a solid connection between the wires should the wire connector come loose.

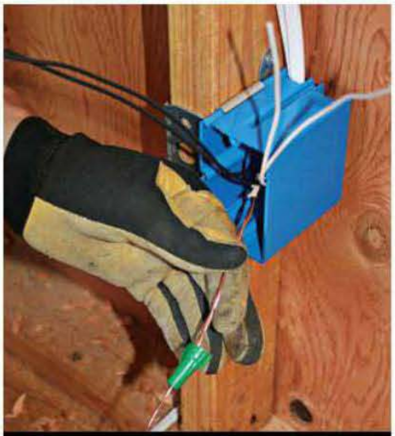

1 Twist ground wires and splice with a connector.

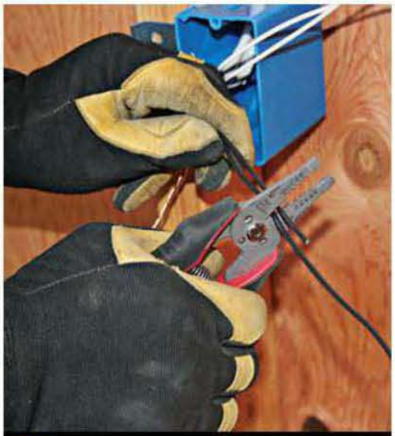

2 Trim hot wires, leaving enough length to work around the device.

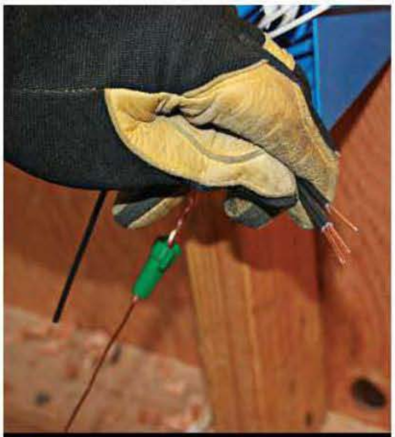

3 Strip wire ends using wire stripper (approximately ¾ in.).

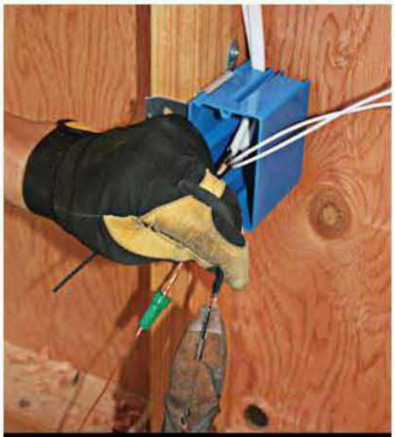

4 Twist the wires together with lineman's pliers.

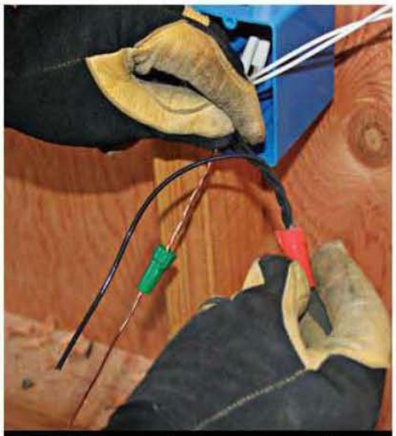

5 Twist on a wire connector until it's tight.

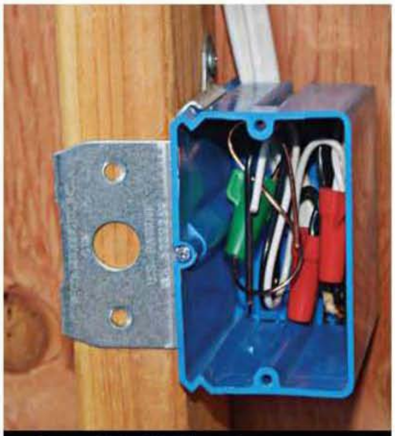

6 Carefully fold the wires into the box.

RECEPTACLES & SWITCHES

RECEPTACLES AND SWITCHES are the most-used electrical devices in a house. They're generally reliable and offer a long life, but they are often replaced when they become cracked or outdated or, eventually, cease to work. Fortunately, replacing them is straightforward and safe if you first shut off the power to the circuits that feed them—and test with a voltage tester to be sure that power is, in fact, off. This chapter shows you how to safely remove receptacles and switches, how to use a continuity tester to see if devices are defective, and how to install new receptacles or switches.

Wiring an electrical device is considered part of finish wiring—also called the trim-out stage—when finish walls are in place and painted. At the trim-out stage, everything should be ready so that the electrician needs only a pair of strippers and a screwdriver or screw gun.

TOOLS & MATERIALS

Choosing receptacles & switches, p. 40

Back-wired devices, p. 42

Testing for power, p. 43

WIRING RECEPTACLES

Two ways to wire a receptacle, p. 44

Wiring a duplex receptacle, p. 45

Making receptacle boxes flush to drywall, p. 47

Wiring a GFCI receptacle, p. 48

Wiring a split-tab receptacle, p. 49

TESTING DEVICES

Testing a single-pole switch, p. 50

Testing a three-way switch, p. 51

Testing a four-way switch, p. 51

WIRING SWITCHES

Wiring a single-pole switch, p. 52

Single-pole switch with back-fed wiring, p. 53

Replacing a single-pole switch, p. 54

Wiring a linear slide dimmer, p. 56

Roughing in a three-way switch, p. 58

Wiring a three-way switch, p. 59

Wiring a back-fed three-way switch, p. 60

Wiring a four-way switch, p. 61

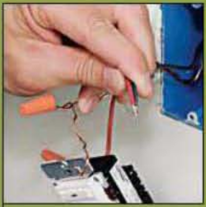

WIRING SPECIALTY SWITCHES

Wiring a switch/receptacle combo, p. 62

Wiring a double switch, p. 63

Wiring a motion detector, p. 63

Wiring an electronic timer switch, p. 64

Wiring a manual timer, p. 65

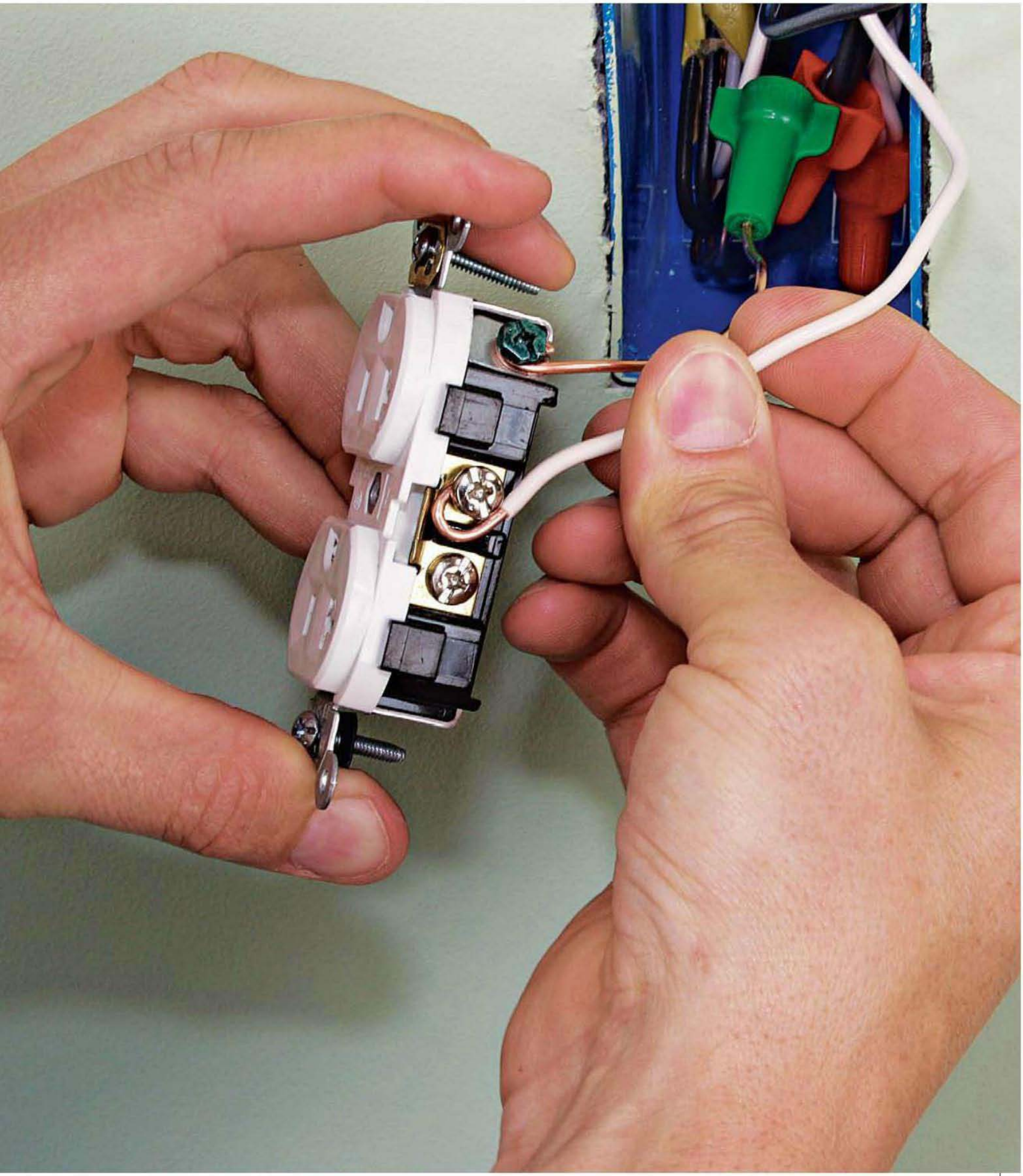

CHOOSING RECEPTACLES & SWITCHES

The difference in quality from one receptacle or switch to another can vary greatly. Over the life of the device, the difference in price is trivial, but the difference in performance can be substantial. For this reason, buy quality. As you can see in the top photo at right, cheap receptacles are pretty much all plastic, their thin metal mounting tabs will distort easily, and they tend to crack if subjected to heavy use.

On the other hand, quality receptacles and switches tend to have heavier nylon faces and may be reinforced with metal support yokes that reinforce the back of the devices.

Another telling detail is how wires are attached—whether they're screwed to terminals on the side of the device, inserted into the back of the device and held by internal clamps (back-wired), or some combination of the two. Again, better quality devices have better mechanisms for gripping wire.

➔ **For more on back-wired devices, see p. 42.**

Most household receptacles are rated for 15-amp circuits and wired with 14AWG or 12AWG wire. The National Electric Code (NEC) specifies 20-amp protection for kitchen appliance, garage, and workshop circuits.

The NEC specifies ground-fault circuit interrupter (GFCI) protection for many locations, including bathroom, outdoor, and kitchen-counter receptacles, so there are also 15-amp and 20-amp GFCI receptacles. In addition, you can obtain various kinds of specialty 15-amp receptacles, including childproof models that require an adult's grip to uncover them, weatherproof receptacles that combine cover and receptacle in an integral unit, recessed receptacles in which clock wires can be hidden, and covered floor receptacles. In addition, there are receptacles specifically matched to the

ACCORDING TO CODE

All bathrooms and kitchens should have GFCI receptacles. All outdoor outlets and some garage outlets must also be GFCIs. Your local building code will have the final say on GFCIs.

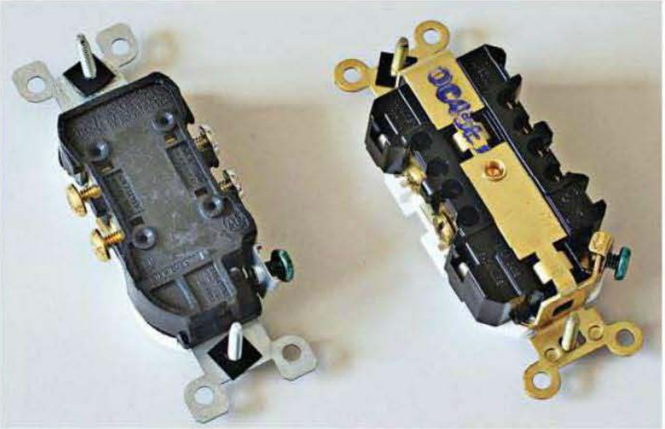

Better quality receptacles and switches are usually heftier and more reliable. The quality receptacle on the right has a nylon face and its back is reinforced with a brass yoke.

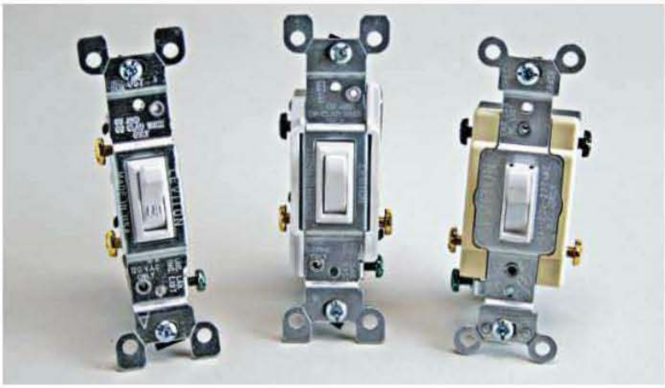

Switches (from left): single-pole, three-way, four-way.

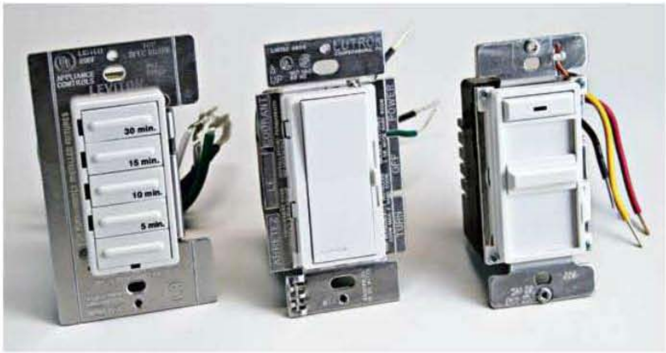

Specialty switches (from left): A timer switch, paddle-switch dimmer with small slide dimmer, linear slide dimmer.

plugs of 30-amp, 40-amp, and 50-amp appliances. Your electrical supplier can help you find the right receptacle for your needs.

Matching load ratings

Circuit components must be matched according to their load ratings. That is, a 20-amp receptacle must be fed by 12AWG cable, which is also rated at 20-amps, and protected by a 20-amp breaker or fuse. A 15-amp receptacle or switch must be fed by 14AWG cable, which is rated for 15-amps, and protected by a 15-amp breaker or fuse.

NM Cable sheathing is color coded to help you correctly match wire size and devices: White sheathing denotes 14-gauge wire; yellow sheathing, 12-gauge; orange sheathing, 10-gauge; and so on.

Polarized receptacles

Receptacles, plugs, and fixtures are *polarized* so they can fit together only one way. A receptacle's gold screw terminal connects to hot wires and, internally, to the hot (narrow) prong of a polarized plug. The receptacle's silver screw terminal connects to neutral wires and, internally, to the neutral (wide) prong of a polarized plug. Finally, the green ground screw connects to the ground wire and the U-shaped grounding prong of the plug.

POLARIZED RECEPTACLES

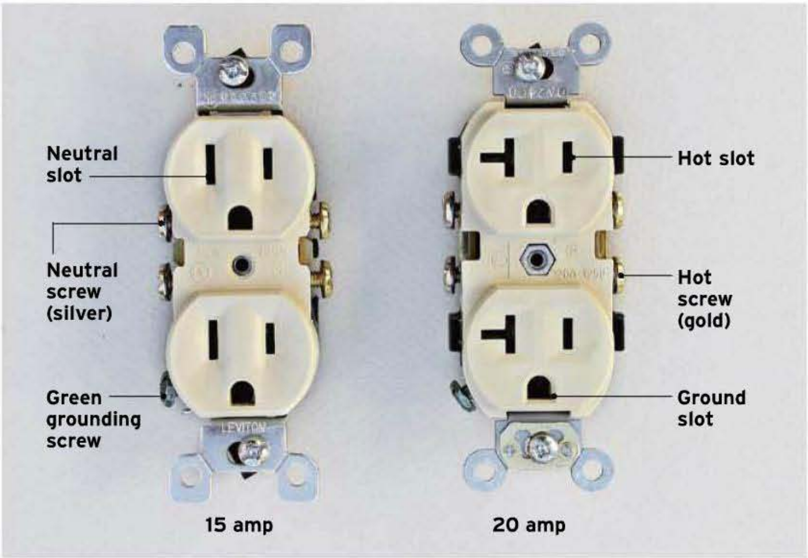
Neutral slot
Neutral screw (silver)
Green grounding screw
Hot slot
Hot screw (gold)
Ground slot
15 amp
20 amp

The 20-amp receptacle (at right) has a T-shaped neutral slot so it can receive a special 20-amp plug in addition to standard 15-amp plugs. But 15-amp receptacles cannot receive 20-amp plugs. Both receptacles are also polarized, so that only the large blade of a plug can fit into the large slot of the receptacle.

RECEPTACLES FOR DIFFERENT LOADS

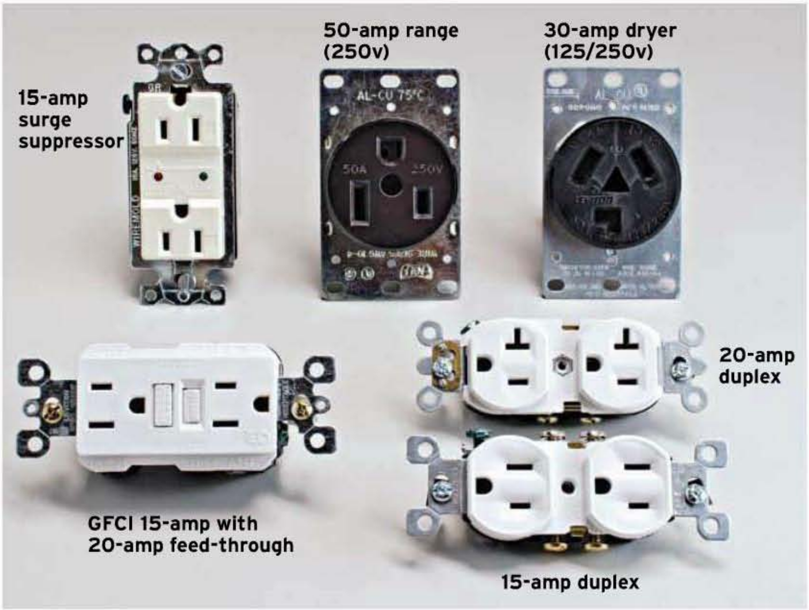
15-amp surge suppressor
50-amp range (250v)
30-amp dryer (125/250v)
20-amp duplex
GFCI 15-amp with 20-amp feed-through
15-amp duplex

BACK-WIRED DEVICES

Back-wiring receptacles or switches is a faster alternative to wrapping wires around screw terminals. Back-wired devices have holes in the back, into which you insert stripped wire ends. But although back-wiring is quicker, many electricians—and some local electrical codes—consider it unsafe, especially for receptacles. Their primary objection is that receptacles' internal tension clamps are made of thin metal strips, which can fatigue, leading to loose wires, flickering lights, and arcing. Moreover, each time users insert or remove plugs, the receptacles move slightly, which increases the likelihood of clamp failure. Manufacturers have stopped making back-wired 20-amp receptacles, so think twice about installing back-wired 15-amp receptacles.

Of course, there's an exception to every rule. A high-quality device such as the GFCI receptacle shown in photo ❶ on p. 48 allows you to loop wires around its screw terminals or insert wires into holes on the back of the device. Here, back-wiring is acceptable because you must screw down screws on either side to tighten internal clamps that grip the wire and ensure a solid connection.

Finally, back-wired switches are acceptable to Code. They rarely fail because switches aren't subject to the stresses of inserting and removing plugs; thus wire connections stay solid. Nonetheless, many pros don't like back-wired switches because their tension clamps can also fatigue and loosen. Use a screw terminal, they argue, and you're guaranteed a solid wire connection.

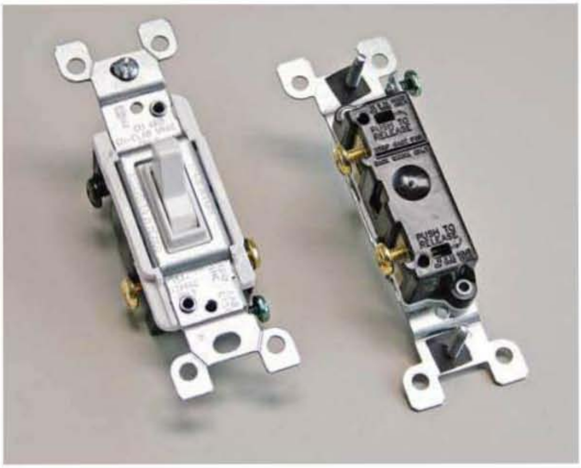

Many switches give installers the option of connecting wires to screw terminals on the side or of back-feeding wires into holes in the back of the switch body.

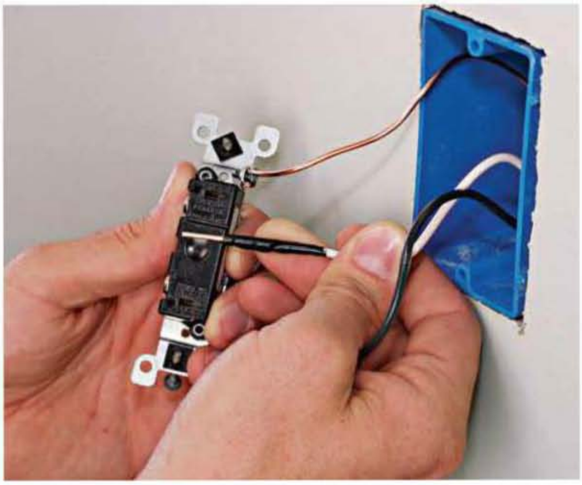

Back-wired switches are acceptable to most electrical codes. Use the stripping gauge on the back of the switch to determine how much insulation to strip from the wire.

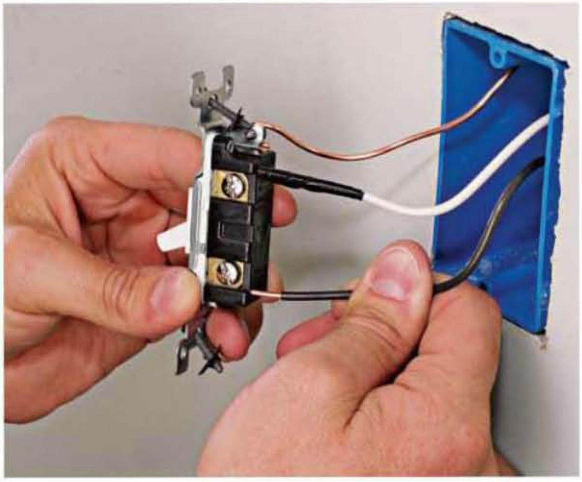

After stripping the wire end, insert it into a back port/hole until the wire bottoms, then pull gently to make sure that the device's internal clamp has gripped the wire securely.

TESTING FOR POWER

To identify the circuit that serves a particular receptacle, insert tester prongs into the receptacle and have a helper at the service panel flip breakers until the tester light goes out. To identify the circuit that serves a particular switch, turn on the fixture it controls and flip breakers until the light goes out. If that test is inconclusive or you aren't sure the receptacle or switch is operable, remove the cover plate and the two screws holding the device to the box. Being careful not to touch screw terminals or wires with your fingers, pull the receptacle out of the box. Touch tester prongs to each screw and to spliced wire groups. Here, an inductance tester is superior to a two-prong voltage tester because the inductance tester can usually read current through wire insulation or a wire connector.

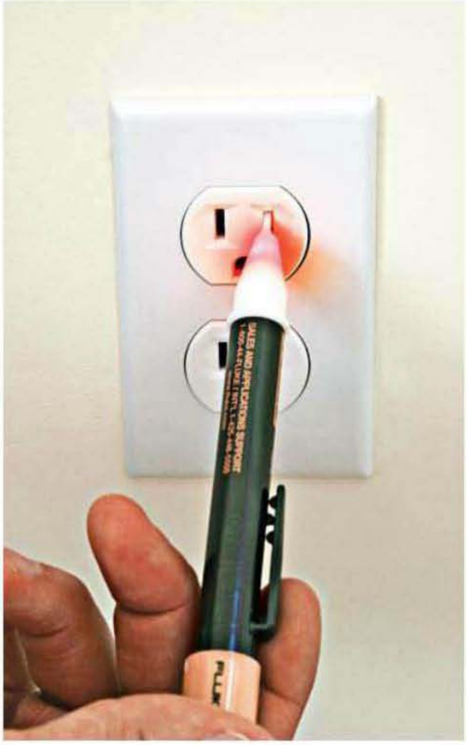

Before touching a wired receptacle, switch, or fixture, use an inductance tester to see if power is present.

WARNING

It's worth noting that in some old houses, the neutral wires—rather than the hot wires—may be attached (incorrectly) to receptacles or switches, in violation of Code. When testing existing receptacles, switches or fixtures, test *all* wires for voltage.

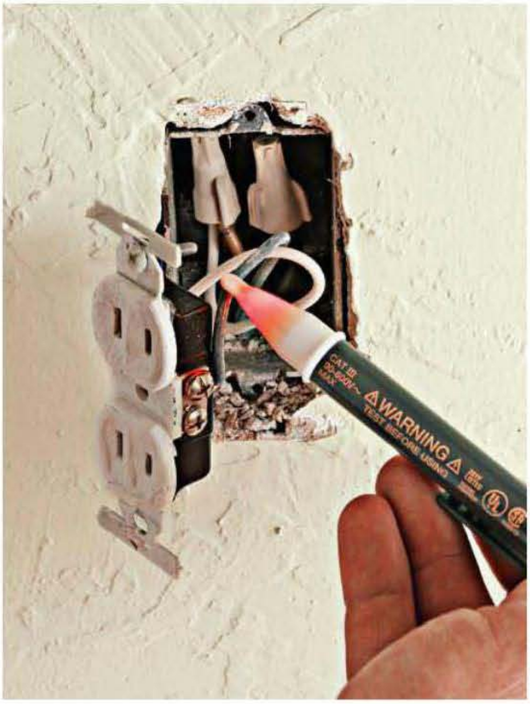

TWO WAYS TO WIRE A RECEPTACLE

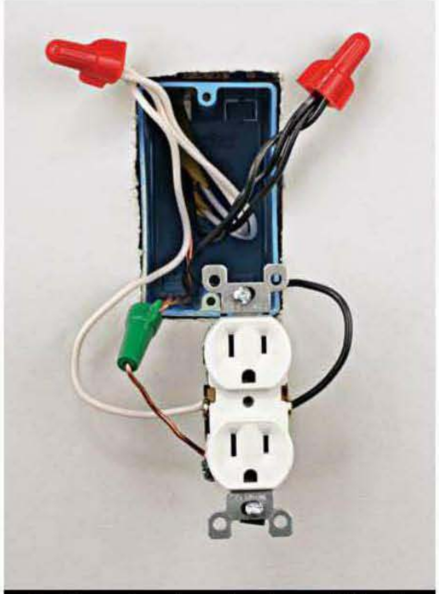

A The preferred way to wire a midcircuit receptacle.

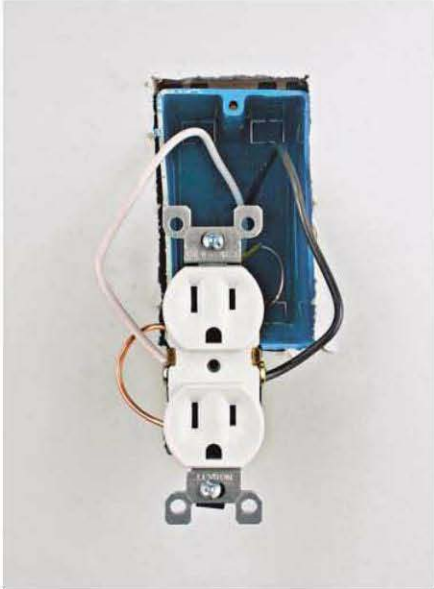

B At the end of a circuit, wires from the cable attach directly to the receptacle.

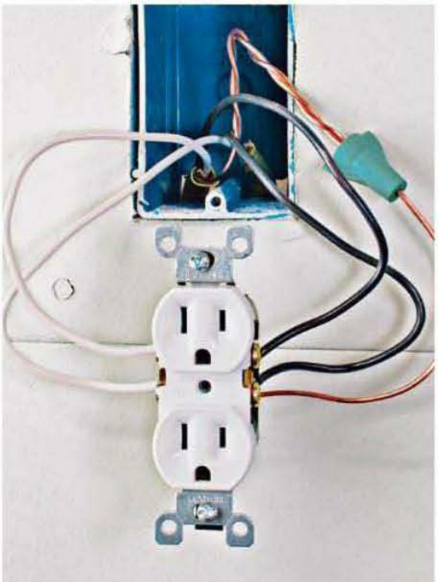

C An alternate to wiring a midcut receptacle, without using pigtails. (Not recommended).

The duplex receptacle is the workhorse of house wiring, because it enables you to plug in a variety of energy users at locations around the house. Receptacles are so indispensable to modern life that Code dictates that no space along a wall in a habitable room should be more than 6 ft. from a receptacle and any wall at least 2 ft. wide must have a receptacle.

➔ For more receptacle requirements, see p. 40.

The NEC requires that all conductors (wires) be solidly connected to devices such as receptacles, but several different wire configurations are acceptable. In brief, you can use pigtails—short lengths of wire running from wire splices to a device—or attach conductors directly to devices. Using pigtails to connect conductors ensures continuous power downstream, but sometimes, at the end of a run, for example, it makes sense to connect conductors directly to the device.

Most electricians prefer using pigtails to connect conductors to midcircuit receptacles because it ensures continuous power and, if it's necessary to replace the receptacle at some future date, there are only three pigtails to disconnect. Wiring with pigtails is shown in the photo Ⓐ.

At the end of a circuit, where only one cable feeds an outlet, there's no need for pigtails. Just attach incoming wires directly to the receptacle as shown in photo Ⓑ. As with pigtail wiring, connect the ground wire first, then the neutral, then the hot wire.

Route the circuit through the receptacle

Feeding circuits through receptacles is a widespread wiring method because, on the whole, it is quicker, requires fewer conductors, and results in boxes that are less crowded than those wired with spliced wires and pigtails Ⓒ.

But this method has detractors, who argue that in a circuit so wired, if a receptacle upstream fails or a wire comes loose, receptacles downstream will lose power. For this reason, feeding a circuit through a receptacle is particularly risky if you also use back-wired receptacles whose internal clamping mechanisms can weaken and result in loose connections. Moreover, there is a voltage drop of about 1 percent per receptacle wired in this manner because receptacles offer more resistance to current flow than do wires. If you have 6 or 10 receptacles daisy-chained in this manner, those overheated connections can waste energy and increase your electricity bill.

➔ See also "Back-Wired Devices," on p. 42.

WIRING A DUPLEX RECEPTACLE

When a duplex receptacle is in the middle of a circuit, there will be two 12/2 or 14/2 cables entering the box—one from the power source and the other running downstream to the next outlet. In the sequence shown here, there are two 12/2 cables because the box houses a 20-amp receptacle.

To ensure continuity downstream, all wire groups will have been spliced with wire connectors during the rough-in stage. A pigtail from each splice will need to be connected to a screw terminal on the receptacle. Unless the small tab between screw pairs has been removed, you need attach only one conductor to each side of the receptacle.

Loop and install the ground wire to the receptacle's green grounding screw first. Place the loop clockwise on the screw shaft so that when the screw is tightened down, the screw head will grip—rather than dislodge—the wire ❶.

Next, loop and attach a neutral conductor to a silver screw terminal. Tighten down the screw that you don't use to avoid electromagnetic interference on radio receivers and the like. Then flip the receptacle over to access the brass screw terminals on the other side. If a looped wire end is too wide, use needle-nose pliers to close it ❷.

>> >> >>

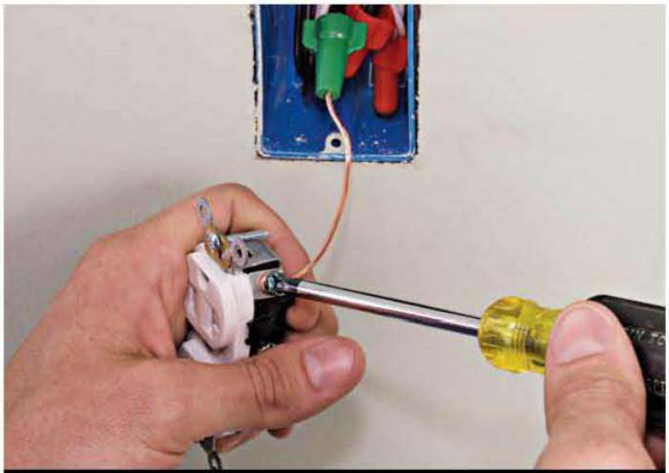

1 Connect the ground wire first.

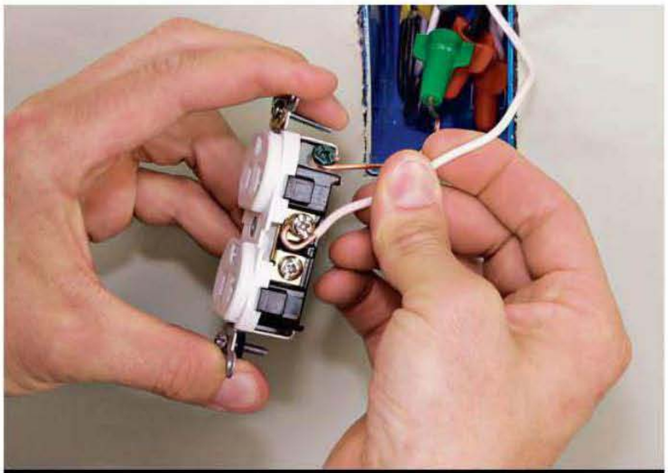

2 Connect the neutral wire next.

Wiring in an orderly way Any habit that increases your safety is worth adopting. When connecting wires to devices, most electricians connect the ground wire first, then the neutral wire, and then the hot wire. When disconnecting wires, they reverse the order: Disconnect the hot first, then the neutral, then the ground wire. Because the ground wire offers the lowest impedance path to ground, it makes sense to leave it connected as long as possible. Even if you're working on circuits that are disconnected, as veteran electricians say, "Treat every conductor as if it were live and you'll stay alive."

WIRING A DUPLEX RECEPTACLE (CONTINUED)

Screw down the brass screw so that it grips the hot wire. Pros frequently use screw guns for this operation, but weekend electricians should tighten the screw by hand to ensure a solid connection ❸.

Push the wired receptacle into the box by hand, keeping the receptacle face parallel to the wall ❹. Then hand-screw the device to the box. Avoid the temptation to use a screw gun because it can strip the screw holes in a plastic box ❺. Finally, install a cover plate to protect the electrical connections in the box and to prevent someone from inadvertently touching a bare wire end or the end of a screw terminal.

TRADE SECRET

It doesn't matter whether you install three-slot (grounded) receptacles with the ground slot up or down—just be consistent throughout the house.

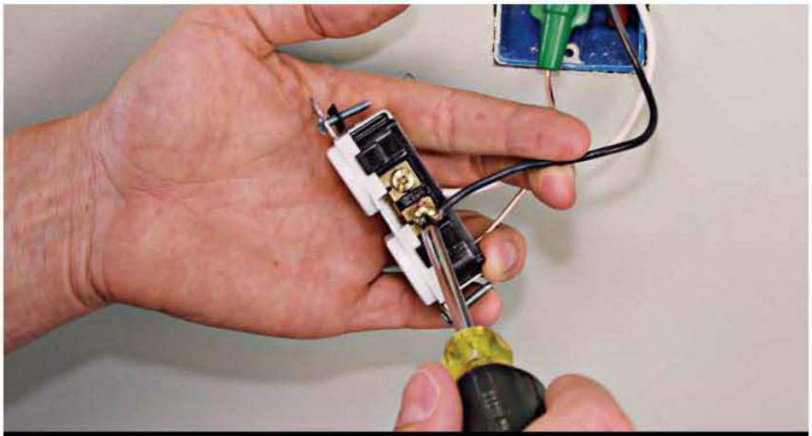

3 Tighten the hot wire.

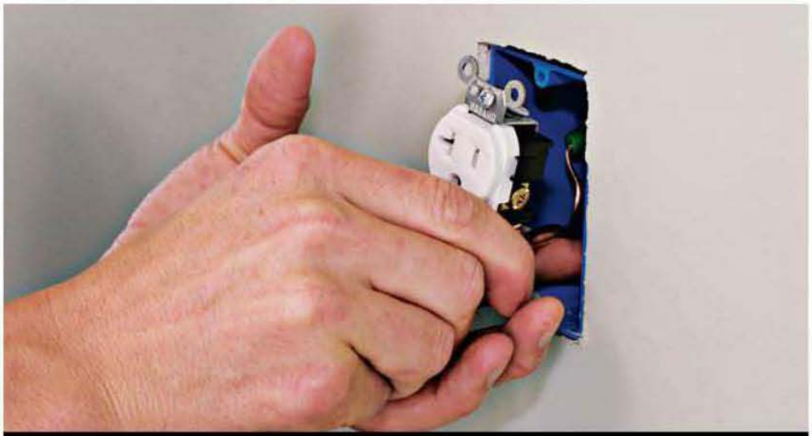

4 Push the device into place by hand.

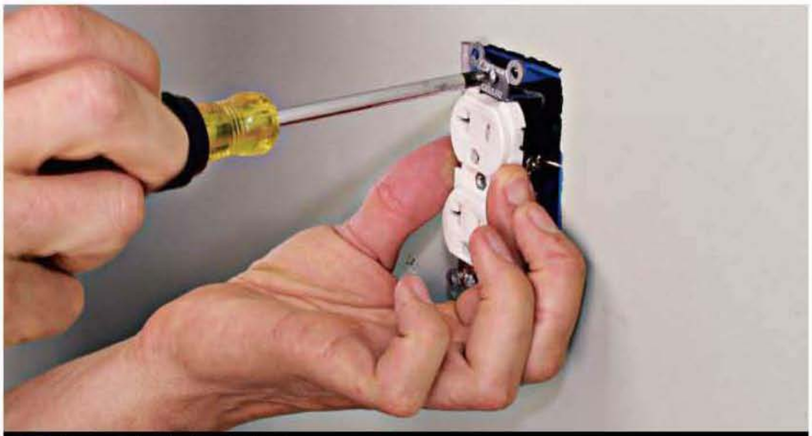

5 Secure the device to the box.

MAKING RECEPTACLE BOXES FLUSH TO DRYWALL

Outlet boxes are frequently installed below the surface of the drywall. If you're using adjustable boxes in new work, that's not a problem because you can turn a screw to raise the box until it's flush with the drywall. It's imperative to bring the device flush to the drywall and to mount it securely. After a plug is inserted into the receptacle a few times, the receptacle moves and the cover plate cracks, which is both unsightly and unsafe.

If you're using a typical nail-on box, you can use plastic spacers (often called *caterpillars*) to build up the level of the receptacle so its mounting tab is flush to the drywall. These spacers take up the space between the mounting plate on the box and the device, so the mounting plate can be flush. Break off the pieces from the strip. This style folds. Insert the spacer behind the screw tabs.

TRADE SECRET

If a box is recessed more than 1/4 in. from the surface, you must use a "goof ring" (p. 26).

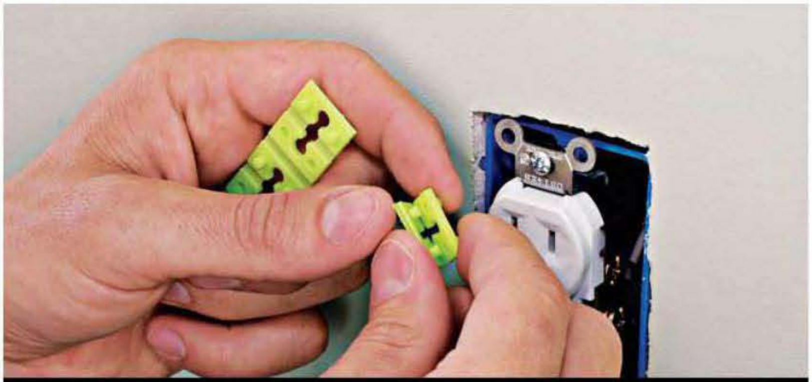

1 **If nail-on boxes are set too low, use plastic spacers to build up the level of the receptacle or switch so its mounting tab is flush to the drywall.**

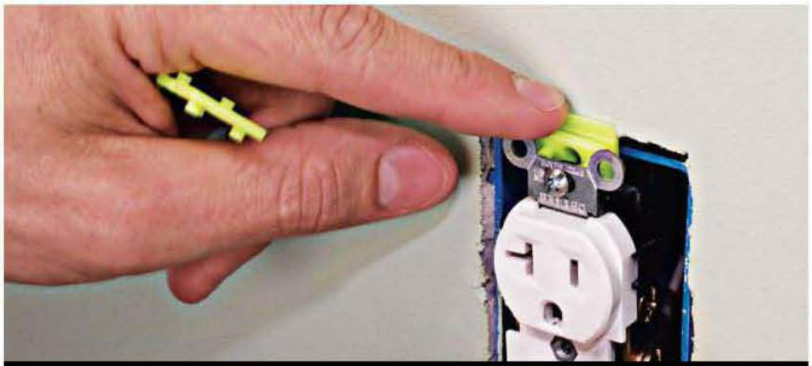

2 **After inserting the spacers between the device's mounting tab and the edge of the box, screw the device solidly to the box.**

TWO-SLOT RECEPTACLES

Receptacles with two slots (instead of three) are ungrounded.

Receptacles with only two slots are ungrounded. Because they are fed by two-wire cable without a ground wire, they are inherently less safe than three-slot receptacles fed with a grounded cable. If existing cables and receptacles are correctly wired and in good condition, most codes allow you to keep using them. Should you add circuits, however, Code requires that they be wired with grounded cable (12/2 w/grd or 14/2 w/grd) and three-slot receptacles.

Replacing a two-slot receptacle with a GFCI receptacle can be a cost-effective way to add protection to a two-wire circuit. There will still not be a ground wire on the circuit, but the GFCI will trip and cut the power if it detects a ground fault. You can wire the GFCI receptacle to protect just that outlet or that outlet and all outlets downstream (away from the power source).

→ **For more on wiring GFCI receptacles, see p. 48.**

Note: If one slot of a two-slot receptacle is longer, the receptacle will be polarized. That is, a receptacle's gold screw terminal will connect to a hot wire and, internally, to the hot (narrow) prong of a polarized two-prong plug. The receptacle's silver screw terminal connects to neutral wires and, internally, to the neutral (wide) prong of a polarized plug.

→ **For more about polarity, see p. 41.**

WIRING A GFCI RECEPTACLE

When wiring a GFCI receptacle, it's important to connect incoming wires (from the power source) to the terminals marked "line" on the back of the receptacle. Attach outgoing wires (to outlets downstream) to terminals marked "load." To distinguish line and load wires during rough-in, write each term on small pieces of the cable sheathing and slip them over the appropriate wires before folding them into the box.

The GFCI shown here is something of a hybrid because it allows you to loop stripped wire ends around the screw terminals or to leave the stripped wire ends straight and insert them into holes in the back of the device—also known as back-wiring. In this case, back-wiring is acceptable because you must tighten screws on either side to engage internal clamps that grip the wire, thus ensuring a solid connection.

→ **For more on back-wired receptacles, see p. 42.**

If the GFCI is going to protect users at a single outlet, attach wires to only one set of screw terminals ❶. The yellow tape across one set of screws indicates that they are load terminals: If you are hooking up the device to protect only a single point of use, leave the tape in place and connect wires only to the screw terminals marked "line." After attaching the ground pigtail, screw down the silver screw to secure the neutral pigtail.

Connect the hot pigtail to the brass screw last, then push the device into the box carefully, hand-screw it to the box, and install a cover plate ❷.

Feeding the circuit through a GFCI receptacle

If you want a GFCI receptacle to protect the outlet and all outlets downstream, feed the circuit through the receptacle. That is, connect incoming and outgoing cable wires directly to the device, rather than using pigtails. Again, it's important to connect incoming wires to the terminals marked "line," and outgoing wires to terminals marked "load."

→ **For more on wiring through the receptacle, see p. 44.**

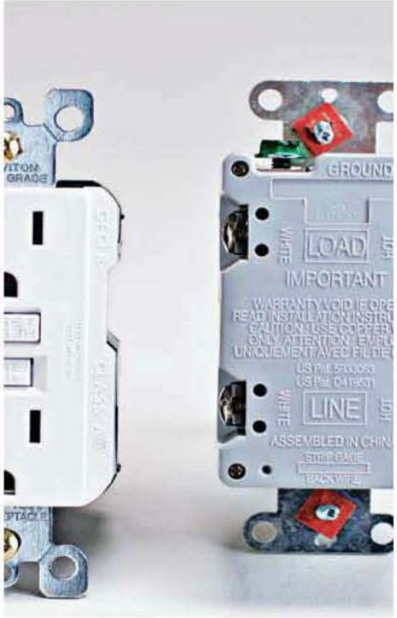

Quality GFCIs can be back-wired or side-wired.

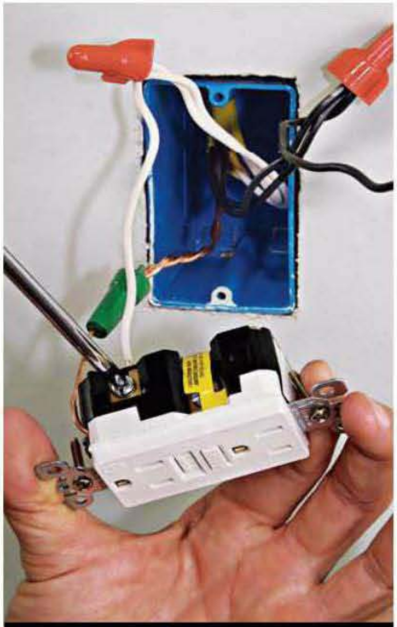

1 A GFCI receptacle can be wired to protect a single outlet.

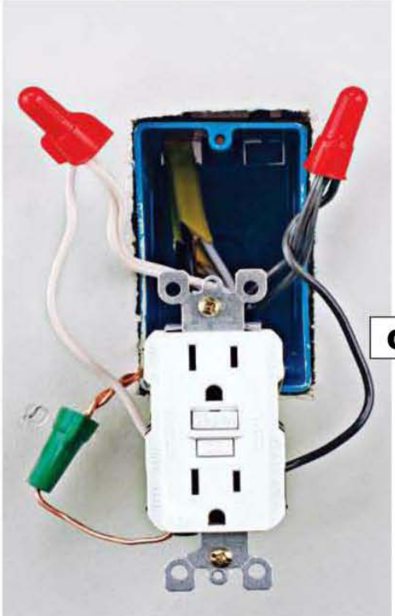

2 Connect the hot pigtail last.

or

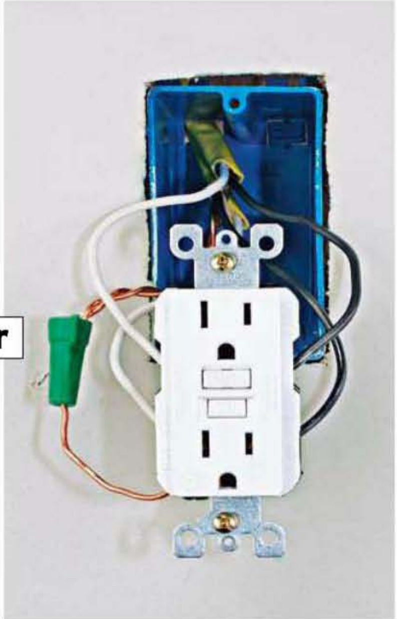

Circuits can be fed through a GFCI receptacle.

WIRING A SPLIT-TAB RECEPTACLE

Standard duplex receptacles have a small metal tab between the brass screw terminals, which conducts power to both terminals, even if you connect only a hot wire to just one terminal. However, if you break off and remove that tab, you isolate the two terminals and create, in effect, two single receptacles—each of which requires a hot lead wire to supply power.

This technique, known as split-tab wiring, is often used to provide separate circuits from a single outlet, a configuration commonly used when connecting a disposal and a dishwasher. The disposal receptacle is almost always controlled by a switch, which allows you to turn off the disposal at another location. To supply two hot leads to a split-tab receptacle, electricians usually run a 12/3 or 14/3 cable.

➔ **For more on split-tab circuits, see the drawings on pp. 174 and 175.**

To create a split-tab receptacle, use needle-nose pliers to twist off the small metal tab between the brass screws ❶. Next connect the bare ground wire to the green grounding screw on the device and connect the white neutral wire to a silver screw. If you keep a slight tension on the wires as you tighten each screw, they'll be less likely to slip off ❷.

Flip the receptacle over to expose the brass screws on the other side, and connect a hot lead to each brass screw. If you're running 12/3 or 14/3 cable, one hot wire will typically be red and the other black ❸. Finally, push the device into the box by hand until it's flush and install the cover plate ❹.

Though this 15-amp split-tab receptacle is fed with 12/3 cable (rated for 20 amps), there's no danger of the load exceeding the rating of the receptacle. Because of the configuration of its slots, the receptacle can receive only a 15-amp plug.

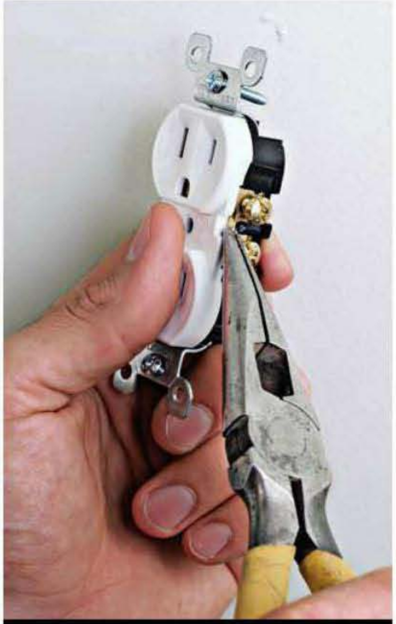

1 Twist off the tab.

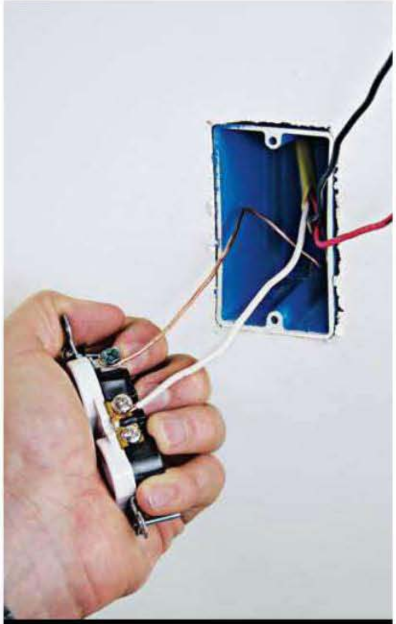

2 Connect the ground, then the neutral.

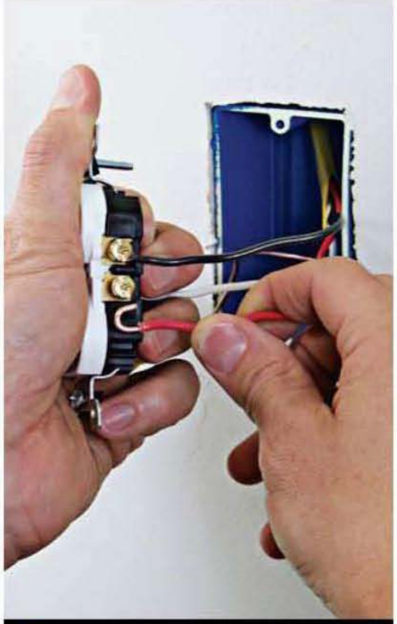

3 Connect the hot wires last.

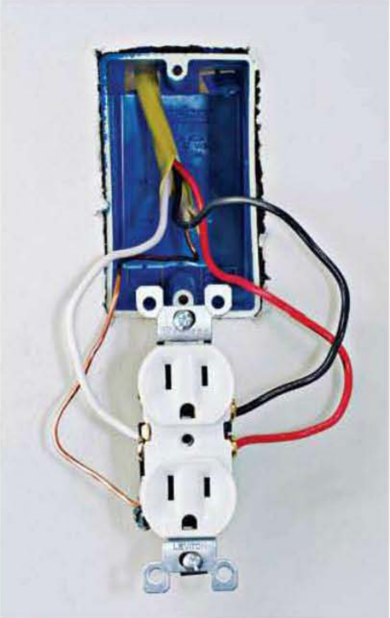

4 Correctly wired split-tab receptacle.

TESTING A SINGLE-POLE SWITCH

To test a disconnected switch, attach the tester clip to one screw or lead wire of the switch and touch the tester probe (point) to the other ❶. Turn the switch off and on ❷. Repeat the process several times. If the tester doesn't light—or doesn't go off in either position—the switch is a dud. Replace it.

Because they're always racing the clock, many pros will simply replace a switch or receptacle that's suspect. But testing a device makes sense for homeowners. Continuity testers are inexpensive and easy to use. Testing a switch can save amateur electricians—who are more likely to miswire a device—a trip to the hardware store to replace a switch that isn't defective.

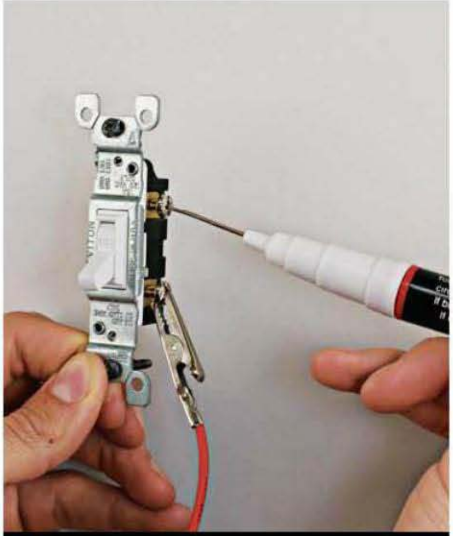

1 Switch off, light off.

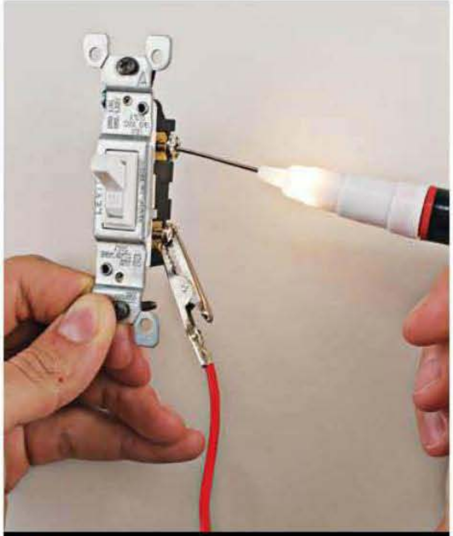

2 Switch on, light on.

WARNING

Use continuity testers only on devices that are not connected to wiring.

Test the tester

Before you start, make sure the tester is working properly. Touch the point and the clip of the continuity tester to each other to be sure the tool is working. This completes the circuit and lights the tester bulb. If the bulb doesn't light, replace the bulb and/or the battery or get a new tester—they're inexpensive.

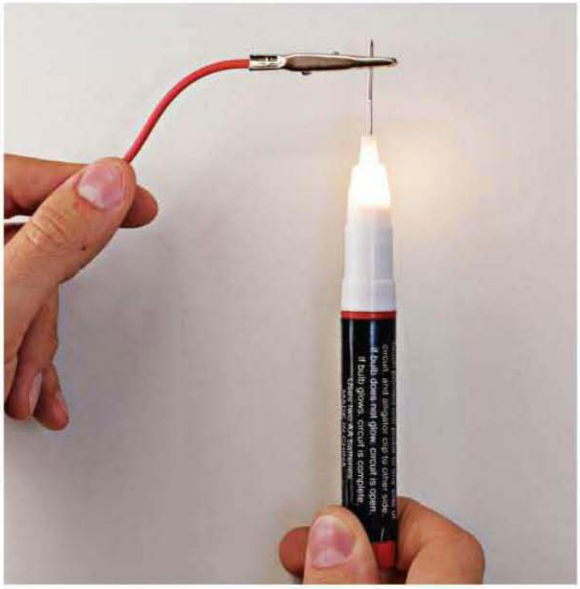

TESTING A THREE-WAY SWITCH

Three-way switches allow you to operate a fixture from two locations—from the top and bottom of stairs, for example, or from opposite ends of a room. When testing a three-way switch, attach the tester clip to the common terminal and touch the probe to one of the traveler screw terminals. The tester should light ❶. If the switch is functioning properly, flip the switch and the light should turn off ❷. Leave the clip in place, touch the probe to the other traveler screw, and flip the switch again. If all that works, the switch is good.

➔ See "Closeup: Three-Way Switch," on p. 172.

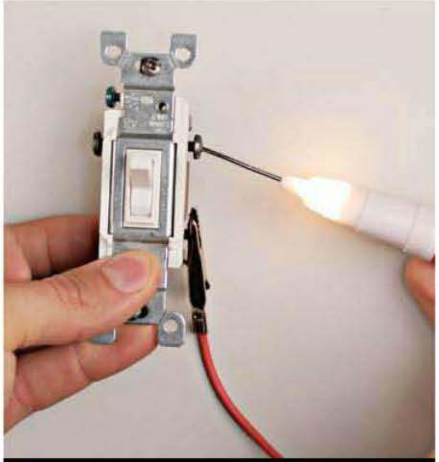

1 Test the common terminal and one traveler screw.

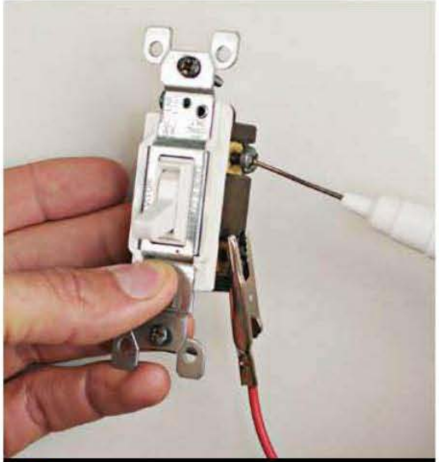

2 Toggle the switch off and the bulb should not light.

TESTING A FOUR-WAY SWITCH

In addition to its green grounding screw, a four-way switch has four screws: two common (dark) screws at the top, and two traveler (brass) screws at the bottom Ⓐ. To successfully test a four-way switch, move the test clip and probe to test all possible combinations.

Start by testing both screws on each side of the switch Ⓑ. Flipping the switch toggle should turn the switch off or on. After testing screw pairs on the same side of the switch, move the clip and probe diagonally. Again, flip the switch to turn the light off (if it was on) or on (if it was off).

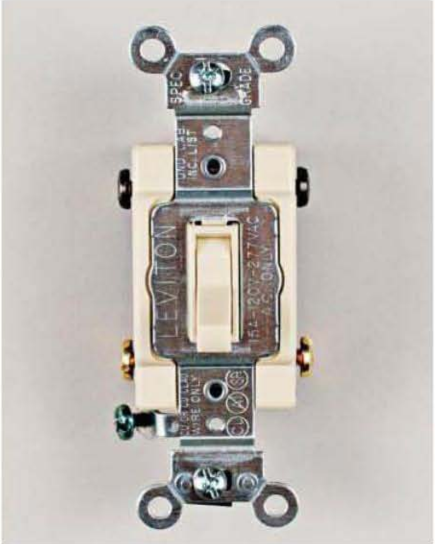

A Four-way switches have four terminals.

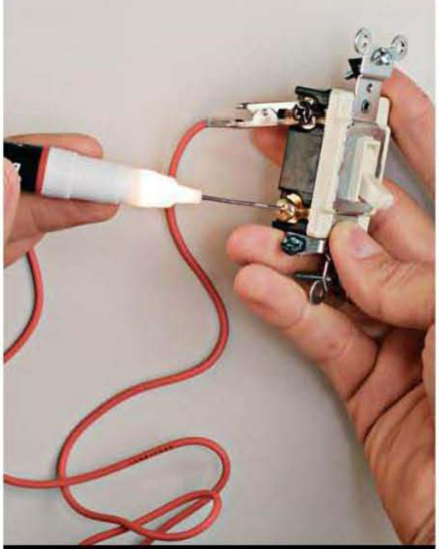

B Test both same-side screws at the same time, then repeat the test on the others.

WIRING A SINGLE-POLE SWITCH

The most commonly installed switch, a single-pole, is straightforward to wire. Spliced together during the rough-in stage, the neutral wires stay tucked in the outlet box. Pull ground and hot-wire groups out of the roughed-in box. Use the hole in the handle of your wire strippers or use needle-nose pliers to loop the conductor ends so they can be wrapped around the screw terminals.

First attach the ground wire to the green grounding screw on the switch ❶. Orient the wire loop in a clockwise direction—the same direction the screw tightens. A loop facing the wrong way may be dislodged by the pressure of the screw head as it tightens down.

Next, connect the hot wires to the switch terminals, again orienting wire loops clockwise ❷. One black wire is hot (power coming in), and the other is the switch leg (power going out to the fixture). With a single-pole switch, however, it doesn't matter which wire you attach to which screw. Generally, pros attach the hot wire last, much as they attach the hot wire on a receptacle last.

Once the ground and hot wires are connected to the device, they're ready to be tucked into the box ❸. Always push the device into the box by hand until it's flush to the wall. Don't use screws to draw a device to a box because the device may not lie flat, and it's easy to strip the screw holes in a plastic box. Likewise, although using a screw gun is faster than a screwdriver, hand screw switches until you get the hang of it ❹.

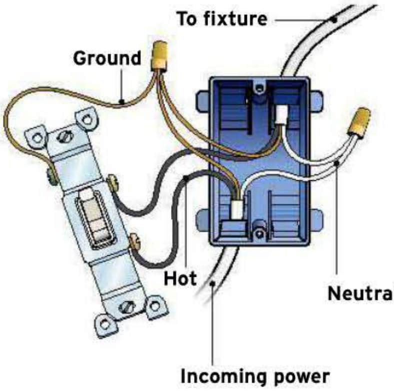

This switch controls a fixture at the end of a cable run. (See p. 172 for complete diagram).

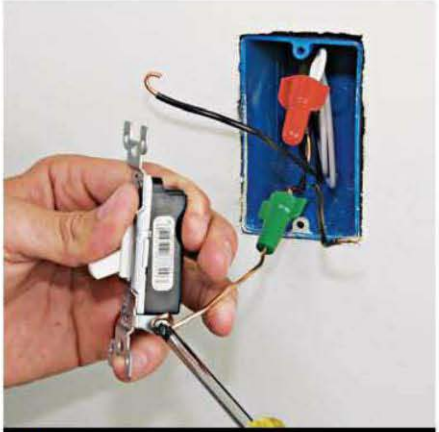

1 Always attach the ground first.

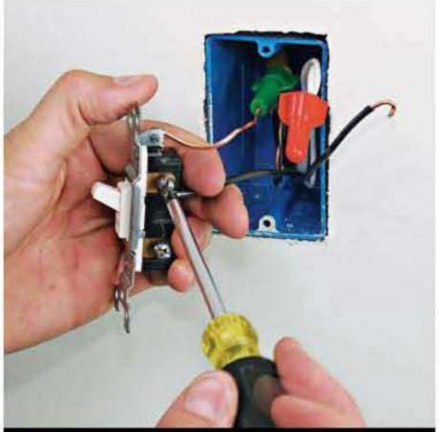

2 Place the wire loops clockwise on screws.

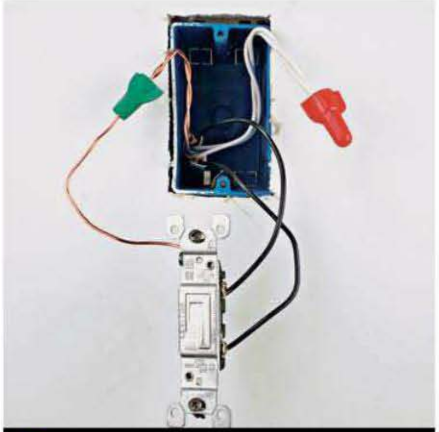

3 Wires ready to be tucked into the box.

4 Screw switches to the box by hand.

Wiring switches **Before connecting or disconnecting wires to a switch, use an inductance tester or a voltage tester to make sure that the power to the switch outlet is off. Test with the switch both on and off to be sure.**

Because switches interrupt only hot wires, you'd think they'd all be easy to wire. As you'll see, however, switch wiring can also be quite complex, especially three-way and four-way switches and switches with electronic components. Follow the manufacturer's instructions carefully.

SINGLE-POLE SWITCH WITH BACK-FED WIRING

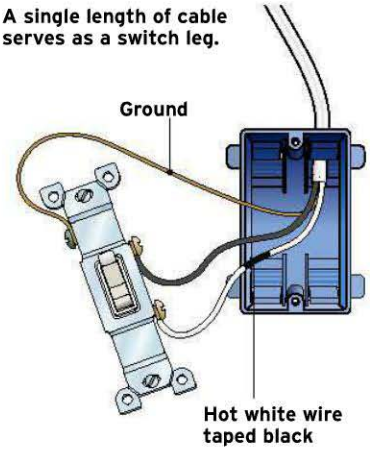
A single length of cable serves as a switch leg.
Ground
Hot white wire taped black

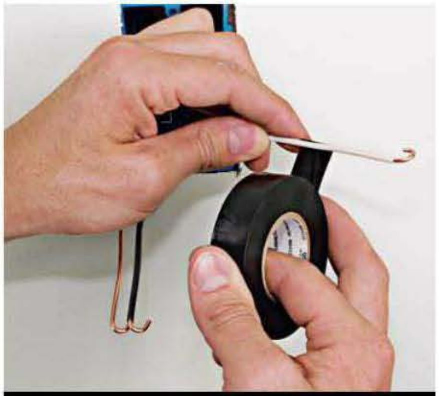

1 Tape the white wire black.

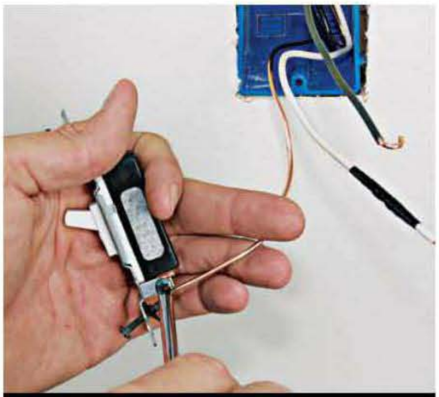

2 Connect the ground wire first.

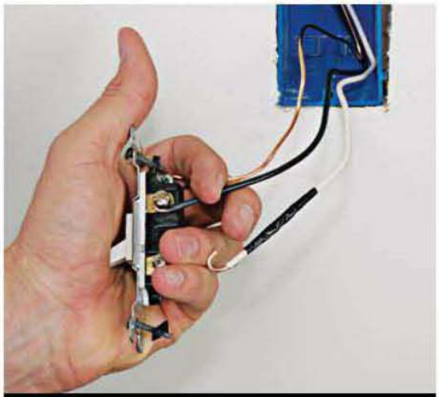

3 Connect the switch leg.

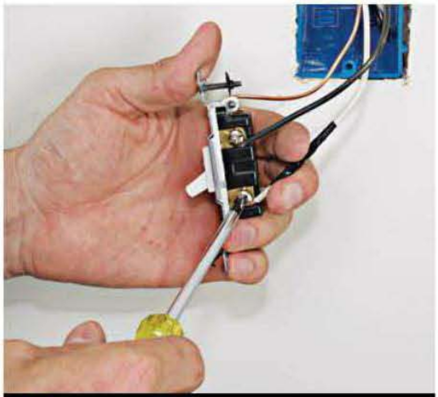

4 Connect the hot wire last.

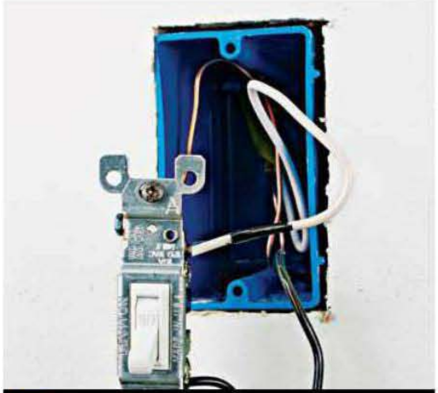

5 The wires are ready to be tucked into the box.

When an outlet or fixture box is closer to the power source than to the switch box, it's common to run a single length of 12/2 or 14/2 cable as a switch loop. This means bringing the power down from the fixture to and through the switch and then back up to the fixture. As such, the black wire in the 12/2 or 14/2 switch loop functions as the incoming hot wire, and the normally neutral white wire acts to return the power to the fixture. The white wire is actually a hot wire and is marked to identify it as such.

At the outlet or fixture box, splice all the grounds together. Attach the source neutral wire to the fixture neutral wire. Attach the source hot wire to the white wire (taped black) of the switch loop. Last, connect the switch loop black wire to the black fixture wire.

➔ **See also "Back-Fed Switch," on p. 172.**

Note: Here, for convenience, we bend the rule of using a white wire only as a neutral wire and instead wind black tape on each end of the white wire to show that, in this case, the white wire is being used as a hot wire.

At the switch, start by stripping and looping the wire ends in the switch loop. Next, tape the white wire with black electrician's tape to indicate that it is serving as a hot wire to the back-fed switch. Convention dictates that the white wire in back-fed wiring is *always* the hot lead (power coming in) ❶. The black wire, on the other hand, is the switch leg that runs back to the fixture.

First, connect the ground wire to the green ground screw on the back-fed switch ❷. Next, connect the switch-leg wire (black) ❸, then the hot wire (white taped black) to the switch terminals. To keep looped wire ends snug against the screw shaft as you tighten down the screw, pull gently on wires, as shown ❹. Not fumbling with wire ends saves time.

Finally, tuck the wires into the box, screw the switch to the box, and install the cover plate ❺.

REPLACING A SINGLE-POLE SWITCH

You can replace a single-pole toggle switch with a convertible dimmer that's wired as a single-pole dimmer. You can use the existing wires, but first turn off the power to the circuit. Use an inductance tester to see if voltage is present at the switch box ❶ (most inductance testers can detect the presence of power even before you remove the cover plate). If the tester glows, there's power present: Turn it off at the fuse box or breaker panel. Test again. If the power's off, unscrew the old switch and pull it out from the box ❷.

Disconnect the switch wires and note their condition ❸. If the cable's fiber sheathing is frayed but individual wire insulation is intact, the wires are probably safe to attach to the replacement switch. If there's debris present in the box, sweep or vacuum it out.

Connect the wires to the new switch ❹. There may not be a ground wire to attach to the new switch's ground screw but Code doesn't require grounding a switch if there's

Cover all connections

All electrical connections not ending at a switch, fixture, or receptacle must be housed inside a covered junction box so they can't be disturbed. Often, electricians will use an existing light box as a junction box in which to splice a cable feeding a new fixture. When there's not enough room in an existing box, use a separate junction box to house the splices.

Code requires all electrical connections to be housed in a junction box.

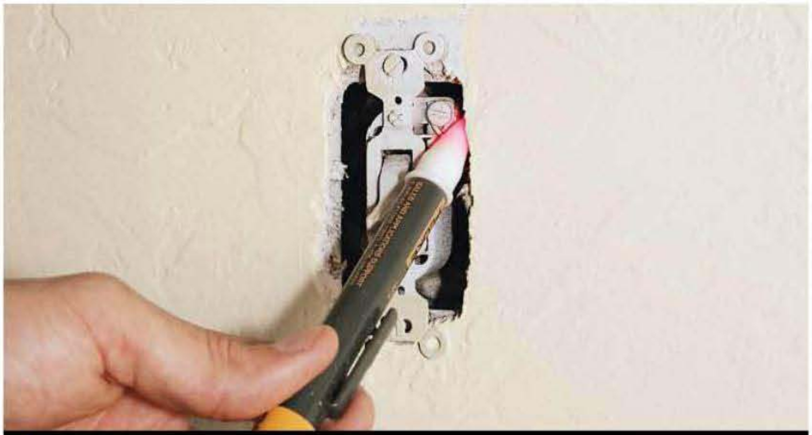

1 Test to make sure the power is off. Here the tester shows energized wires.

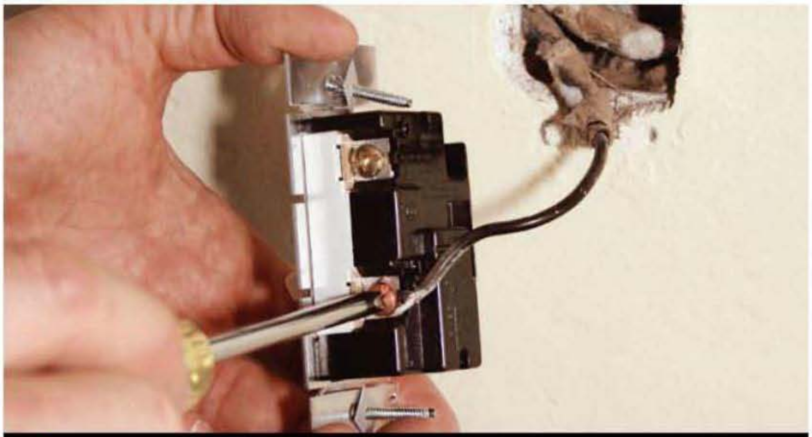

4 Wire the new switch to the existing wiring.

7 Install the cover plate.

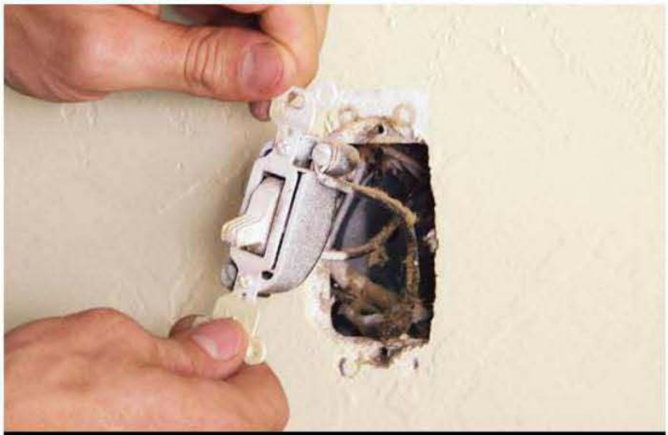

2 When the power is off, pull the switch from the box.

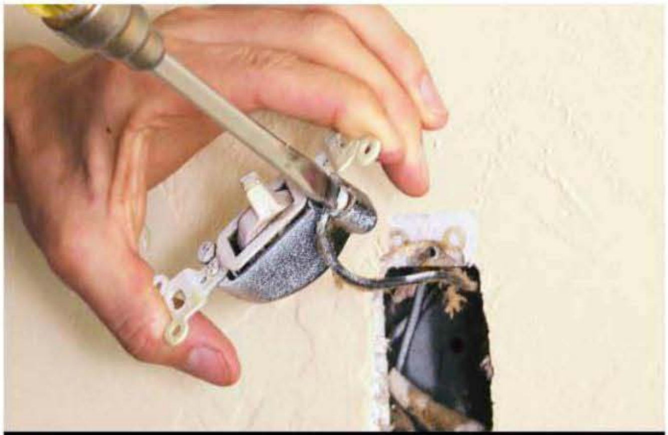

3 Disconnect the old switch.

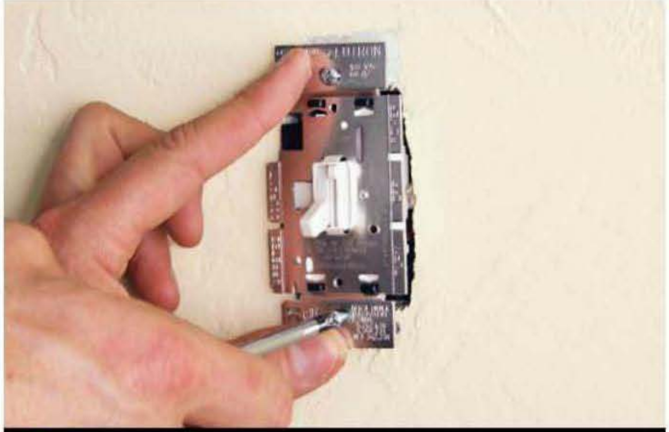

5 Screw the dimmer to the box.

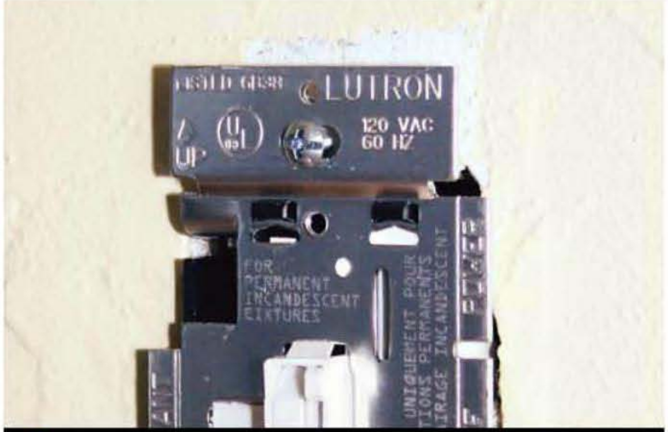
LUTRON
120 VAC
60 HZ
UP
FOR
PERMANENT
INCANDESCENT
FIXTURES
UNIQUEMENT POUR
TIONS PERMANENTS
IRAGE INCANDESCENT

6 The dimmer's rating is stamped on its face.

no ground wire feeding the box. Once the dimmer's connected, set it flush to the wall, and screw it to the box ❺.

Note: A dimmer must match the type of fixture it controls, whether incandescent or halogen or low-voltage. Typically, the dimmer rating is stamped on its face ❻. In this case, the rating specifies, "For permanent incandescent fixtures."

Finally, install the cover plate to protect the connections in the box and to prevent switch users from inadvertently touching the wire ends or dimmer terminals ❼.

TRADE SECRET

A convertible switch can be wired as a single-pole or a three-way switch. This type costs a bit more than a fixed switch but gives you more options on the job site. This switch may even save you a trip to the hardware store.

WIRING A LINEAR SLIDE DIMMER

Newer, more sophisticated dimmers often require different connectors—wire leads rather than screw terminals—but they're still basically switches and so interrupt only hot wires. For standard single-pole switches, it doesn't matter which screw terminals you connect a switch leg or hot wire to, but it does matter which wire you attach to dimmer leads.

Today's dimmers are sophisticated and expensive, so always read the directions that come with them. The slide dimmer shown in photo ① can be wired as a single-pole or three-way switch, depending on which wires you connect. Thus it has a bare-wire ground, red and black hot wires, and a yellow wire that is used to wire the dimmer as a three-way device ②.

If the convertible device will be used as a single-pole dimmer, you won't need the yellow wire. So cap it with a wire connector ③.

Splice the ground pigtail to the device's bare ground lead. Then splice the switch leg from the box to the red lead on the device. On devices with wire leads, typically a red lead attaches to the switch leg ④.

Finally, attach the incoming hot wire to the other hot lead (black) on the device ⑤. Carefully fold the wires into the box and push the wired dimmer into the box. Screw the device to the box, and install the cover plate.

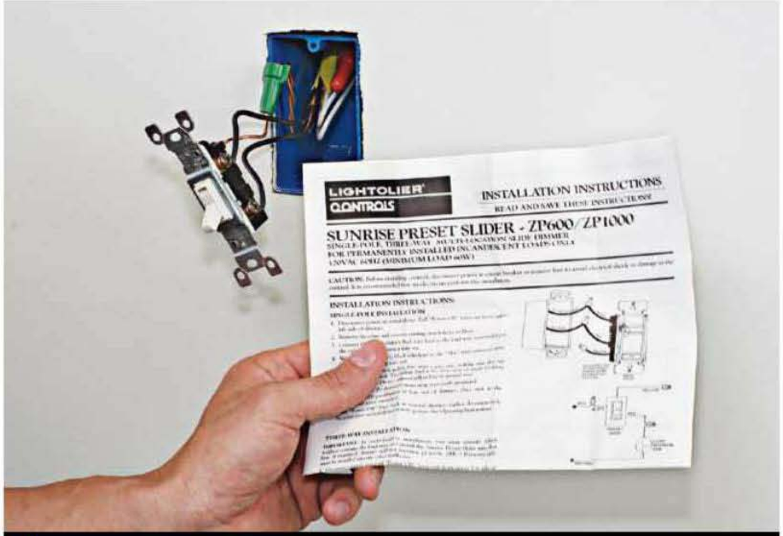
LIGHTOLIER CONTROLS
INSTALLATION INSTRUCTIONS
READ AND SAVE THESE INSTRUCTIONS
SUNRISE PRESET SLIDER - ZP600/ZP1000

1 Dimmer specifics can differ from one manufacturer to the next. Read the directions before installation.

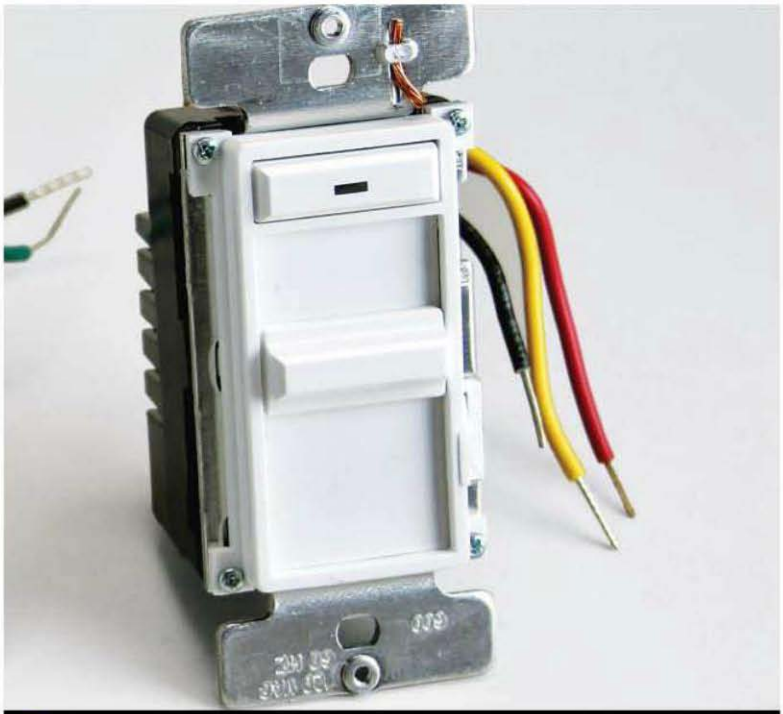

2 Prepare the leads before you begin attaching wires.

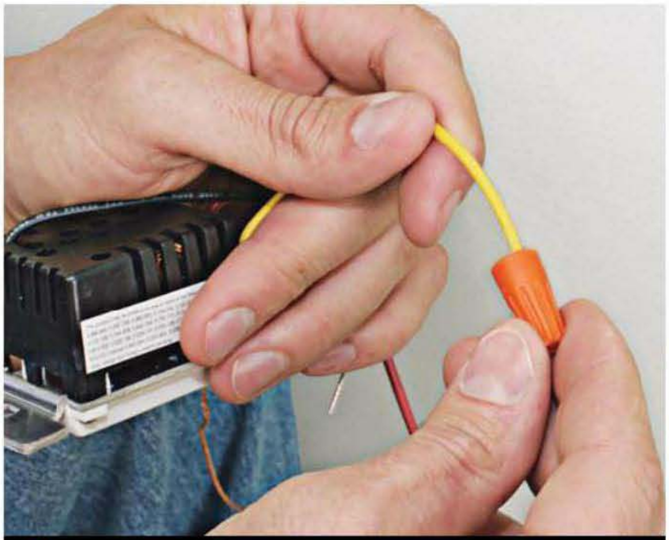

3 Cap any lead you won't connect.

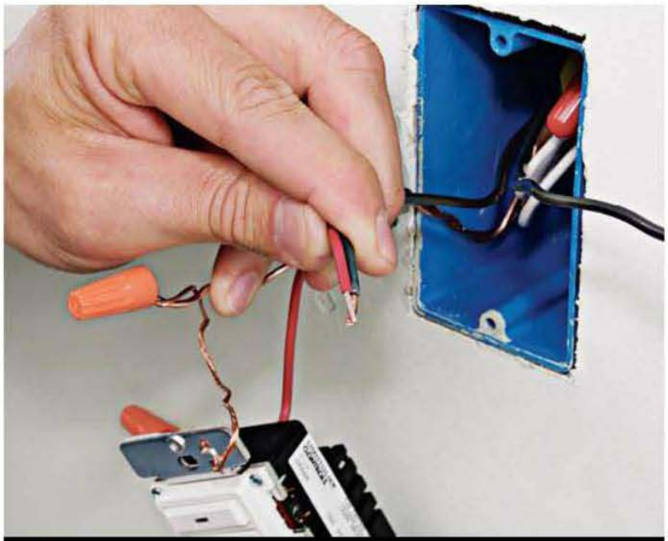

4 Attach the ground wire, then connect the red lead to the switch.

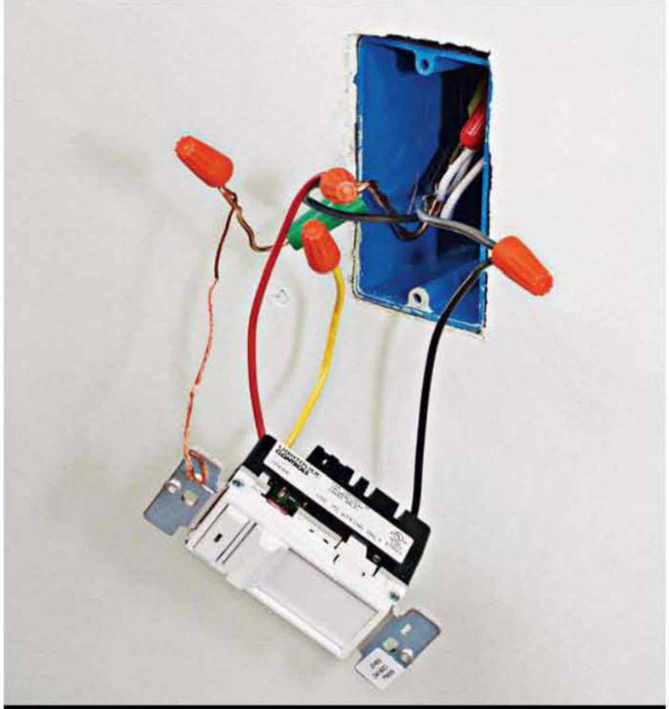

5 Connect the black lead to hot wire.

Dimmer leads It's critical which circuit wires you attach to dimmer leads. So during the rough-in stage, distinguish which wire is the switch leg and which is the incoming hot wire for each single-pole dimmer that you install. To distinguish these wires, many electricians wrap the wires together and bend back the switch leg wire. Typically, the red dimmer lead attaches to the switch leg wire, and the black dimmer lead attaches to the incoming hot wire. Or slip a short sleeve of cable sheathing back onto the stripped wires and use a felt-tipped marker to label what each wire does.

ROUGHING IN A THREE-WAY SWITCH

Three-way switches allow you to operate a light from two locations. They're often used at the top and bottom of a set of stairs or at two entrances to a room. If you get confused about which wire goes where, refer to the wiring schematics on pp. 172–174 or make a drawing of your own.

In new wiring, wires are roughed in when the framing is still exposed. Here, 12/2 and 12/3 cables were fished in to feed a three-way switch that was added after the drywall was up.

For more on fishing cables, see p. 191.

After stripping sheathing from the 12/2 cable, strip the 12/3 cable. Removing 12/3 sheathing is a little different: Start by lightly *scoring* the 12/3 cable along its length, up into the box ①. Then, when you reach the end of the cable, *cut* through the sheathing. Because you'll soon be stripping the ends of individual wires, cutting through the sheathing end won't compromise wire insulation.

For more on stripping cable, see p. 34–35.

Starting at the cut-through sheathing at the end of cable, pull the sheathing free of the wires within ②. The sheathing will separate easily along the scored line. At the cable's upper end in the box, carefully cut free the sheathing.

Twist together the ground wires before splicing them with a wire connector ③. To use a green wire connector, cut one of the leads shorter than the other so that it sticks out of the hole in the end of the wire connector. Then connect that ground lead to the switch's green ground screw. Next, strip wire insulation from the neutrals, splice them, cap them with a wire connector, and push them into the box ④. (Neutrals don't connect to standard switches.)

After professional electricians strip cable sheathing, many wrap individual wires in a distinct fashion so any other electrician will know which wires are travelers and which are switch-leg wires. Group and twist traveler wires clockwise, then wrap the switch-leg wire counterclockwise about the travelers. This way there's no need to put tape or labels on the wires to identify them or to pull out all the cables and figure out which wire is what ⑤.

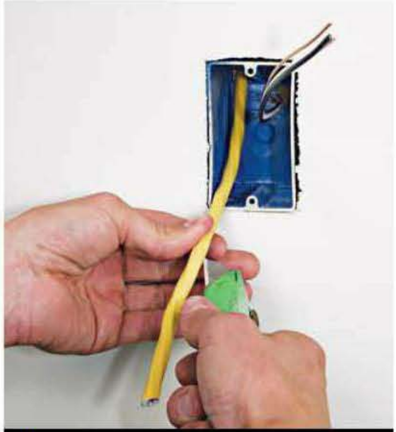

1 **Score the 12/3 cable along its length.**

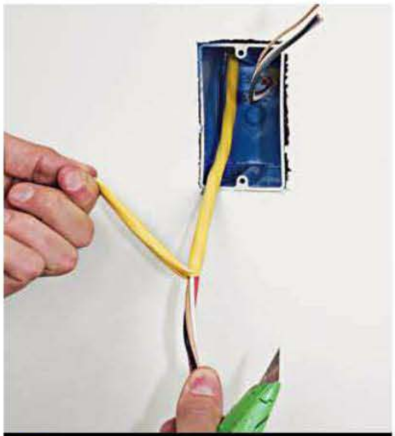

2 **Pull the sheathing free.**

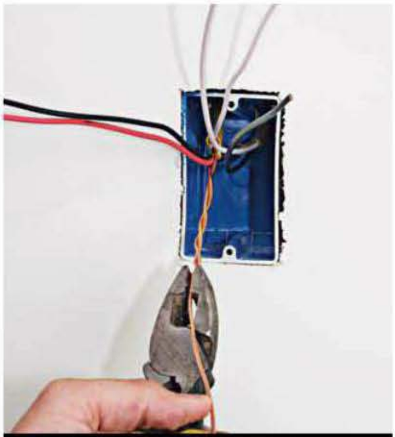

3 **Twist the ground wires before splicing them.**

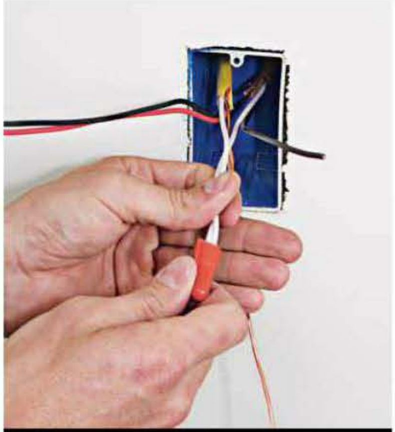

4 **Splice and cap the neutrals.**

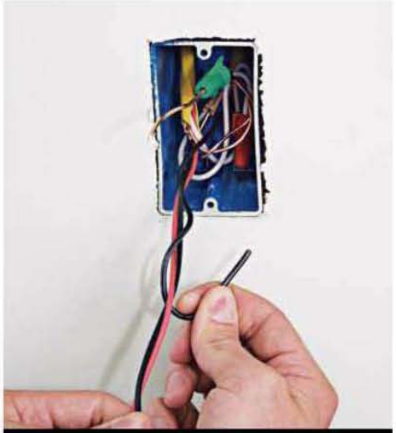

5 **Wrap the switch-leg wires around the travelers.**

WIRING A THREE-WAY SWITCH

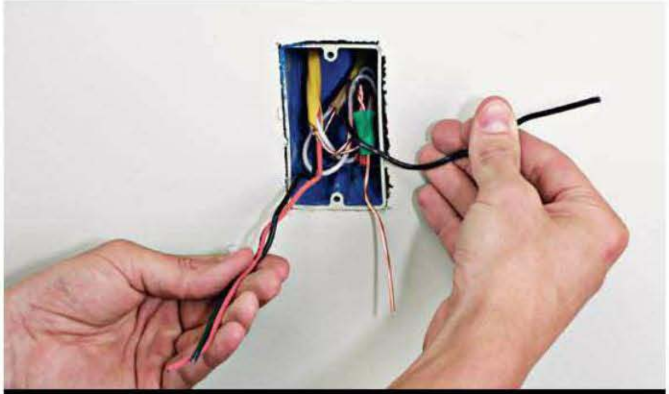

1 Separate the traveler and switch-leg wires.

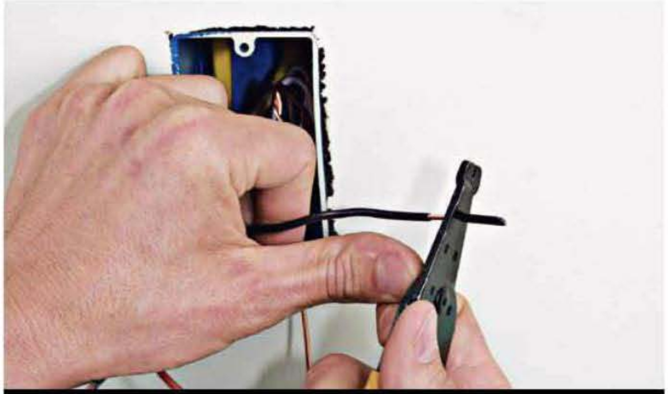

2 Strip the ends of the wires and loop them.

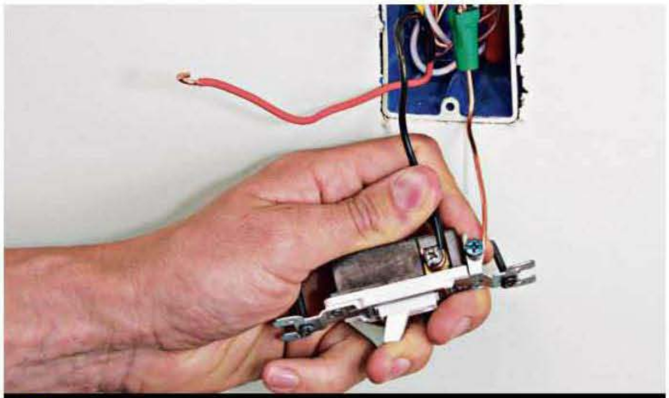

3 Attach the ground, then a traveler.

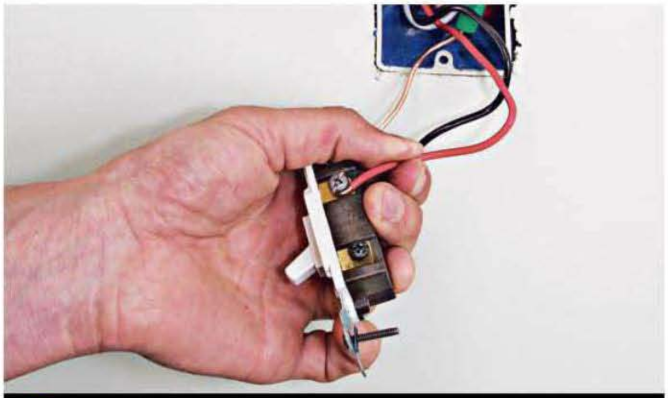

4 Connect the second traveler.

To wire a three-way switch, pull the conductors out of the outlet box. Unwrap the switch leg wire from around the travelers and separate the wires so the travelers are on one side and the switch leg on the other ❶. If individual wire ends weren't stripped during the rough-in phrase, do so now. Give the strippers a quick twist, use your thumb as a fulcrum to push the wire insulation off, and loop the wire ends ❷.

After attaching the bare ground wire to the green grounding screw on the device, attach the first traveler wire ❸. Loop wires clockwise around the screw shafts, and they'll be less likely to slip off when the screws are tightened. Flip the switch over and connect the second (red) traveler ❹. (*Note:* It doesn't matter which traveler wire goes where—you'll still be able to turn lights off and on. The only critical connection is the common terminal.)

If you connect traveler wires in the same position on two three-way switches—say, you attach the red traveler to the first terminal, as just described—the lights will be off when switch toggles are both up or both down. This is a fine point and most people needn't agonize about it: The three-way switches will still work as long as you attach traveler wires to traveler terminals.

After attaching the second traveler, connect the hot conductor (switch leg or hot wire) to the common screw terminal, which is color coded black ❺. Push the device into the box by hand, screw the device to the box, and install a cover plate.

Photo ❻ is a frontal view of the switch we just wired; it's typical for a three-way that's located between the power source and the light fixture it controls.

→ **Also see "Wiring Three-Way Switches," p. 172–174.**

>> >> >>

WIRING A THREE-WAY SWITCH (CONTINUED)

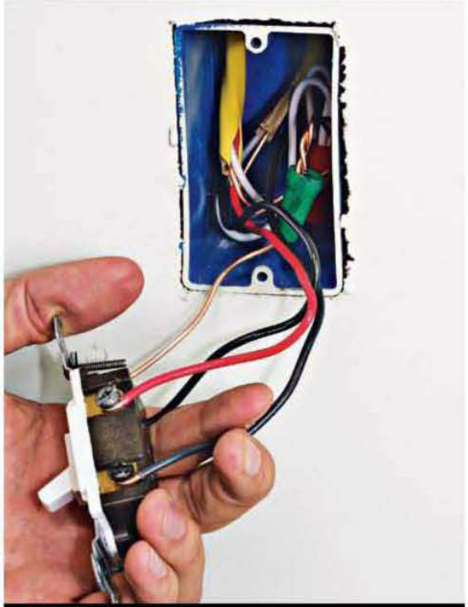

5 Connect the hot conductor last.

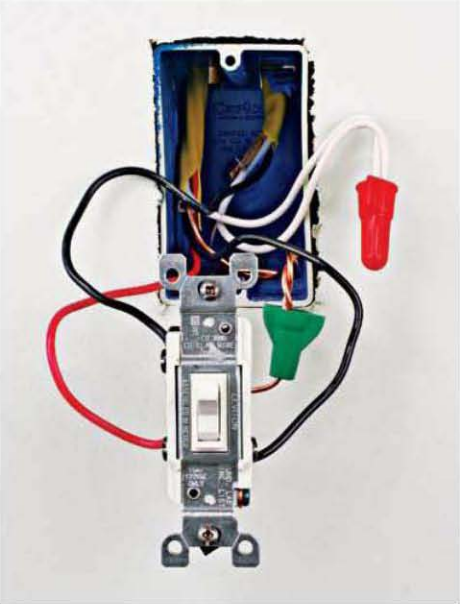

6 Typical connections for a three-way switch.

WIRING A BACK-FED THREE-WAY SWITCH

For a back-fed three-way switch, a single three-wire cable feeds the switch located beyond the light fixture. Hot wires run from a splice in the fixture box. In this case, the red and black wires are travelers. The white wire here is not a neutral; thus it is taped black to indicate that it is the hot conductor connected to the common terminal.

➔ For more on wiring back-fed switches, see the drawing on p. 172.

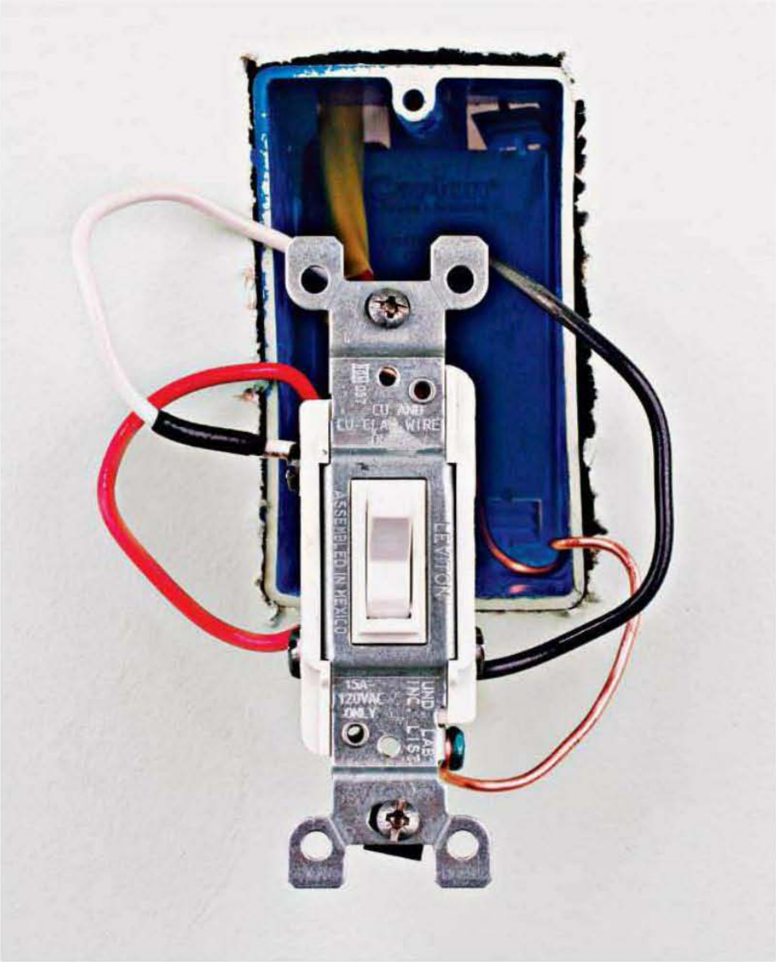

This is a typical three-way switch with back-fed wiring.

WIRING A FOUR-WAY SWITCH

Four-way switches have two travelers incoming (from the power source) and two travelers outgoing (to a second four-way switch). Thus there will usually be two three-wire cables entering the box. (In the sequence shown here, you know we're working with 14/3 cable because the cable sheathing, visible in the box, is white.)

Twist and splice the neutral wires with wire connectors and push them out of the way, into the back of the box. Then strip and loop the ends of the ground wire and the hot conductors and twist them so they turn clockwise around the screw shaft. Next, connect the ground to the four-way switch, as you do for all devices ❶.

Connect a set of travelers on one side of the switch. On each side of a four-way switch, there is a dark terminal and a brass terminal: Connect similar wires to similar terminals. You can connect red wires to dark terminals or black wires to dark terminals—it doesn't matter—just be consistent on both sides of the box. Note how the electrician exerts a slight tension on the wires to keep them from slipping off as he tightens the screw terminals ❷.

To summarize how the wires connect to the four-way switch in this photo sequence: Red travelers are attached to brass screws; black travelers, to dark screws ❸. When you've got this many wires in a box, it's helpful if you partially accordion fold the wires before you push them into the box. Screw the switch to the outlet box, then install the cover plate.

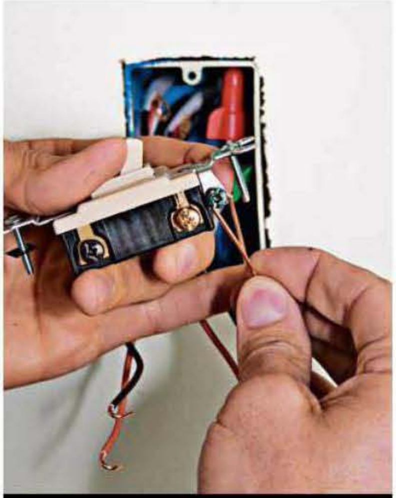

1 Connect the ground wire first.

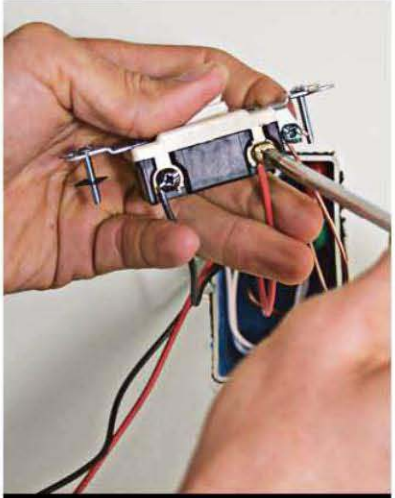

2 Connect one set of travelers.

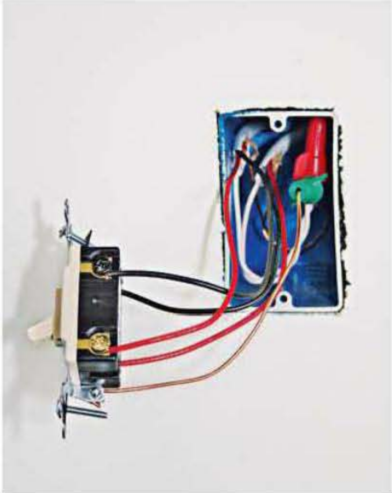

3 Similar wires connect to similar terminals.

TRADE SECRET

When replacing a defective or outdated three- or four-way switch, use a felt-tipped marker to note which wires connect to which switch terminal *before* disconnecting the wires. Of course, turn off power to the switch before you begin.

WIRING A SWITCH/RECEPTACLE COMBO

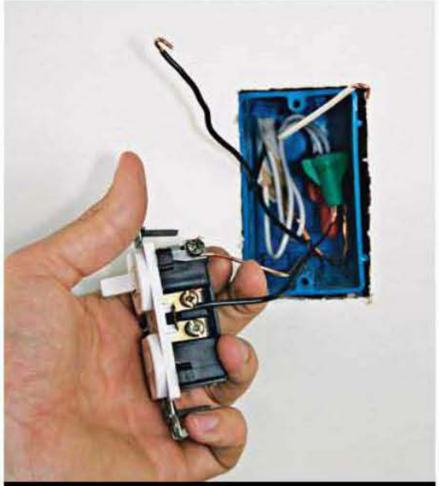

1 Connect the ground wire and then the hot wires to their respective terminals.

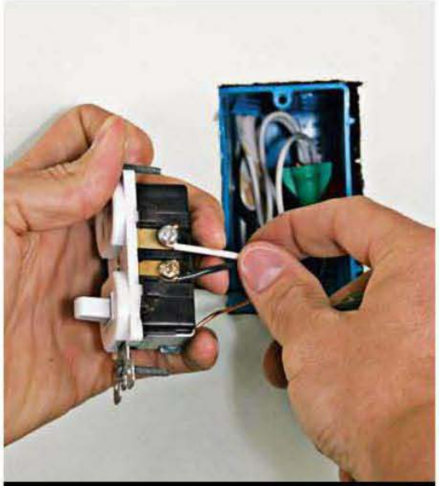

2 Attach the neutral and then the switch leg.

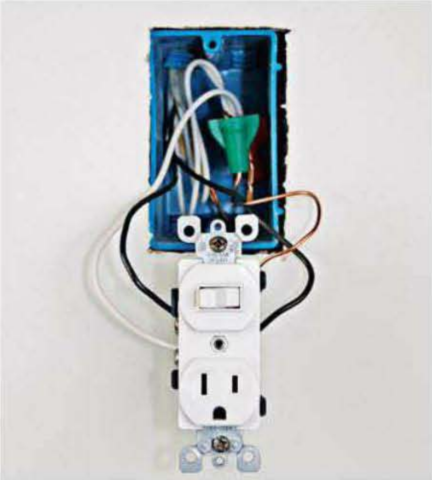

3 Front view of a switch/receptacle combo.

Switch/receptacle combinations are often installed along kitchen counters, where space is at a premium, because it enables you to fit a switch and a receptacle into a single outlet box. In effect, you'll be wiring both a single-pole switch—which interrupts hot wires—and a receptacle that will always be hot. The switch requires a hot wire (from the power source) coming in and a switch leg going out, and the receptacle requires a hot and a neutral wire. A ground wire connects to the device, too.

➔ Specialty switches are shown in the bottom photo on p. 40.

In the sequence shown here, two 14/2 or 12/2 cables feed the setup. During the rough-in stage, splice ground and neutral wires and run a pigtail from each group, which will be connected to the device later.

Connect the ground pigtail to the green ground screw on the device; then connect the hot leg ❶. *Note:* The tab between the two brass screws has not been broken out, so the hot wire will feed both the switch and the receptacle.

Turn over the switch. Notice that, on this side of the device, there is a brass screw and a silver screw. (The two screws are physically isolated from each other inside the device.) Attach the neutral (white wire) to the silver screw, which serves the receptacle. Then connect the black switch leg to the brass screw nearest the switch ❷.

The wired device as seen from the front is shown in photo ❸. On this device, the two screw terminals of the single-pole switch are on opposite sides.

WIRING A PILOT LIGHT

Pilot light switches are used when the light fixture is on the other side of a door that is usually closed, such as a basement door or the door to a walk-in cooler. This specialty switch requires a neutral wire so its tiny pilot light will glow. Wiring this device is similar to wiring a switch/receptacle combo: A single hot wire feeds both the switch and the tiny light bulb, a neutral serves the light, and a black switch leg runs from the switch to the light fixture in the next room.

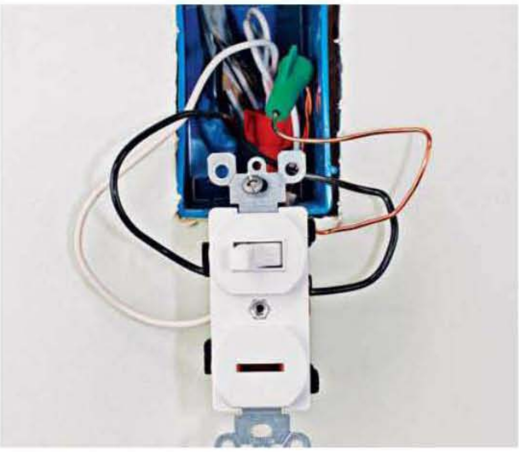

The pilot lights let you know whether the light on the other side of the door is on.

WIRING A DOUBLE SWITCH

This economy switch has only push-in (back-wired) terminals and no ground screw. On one side is a single hot lead coming in, and on the other side are switch legs running to two different energy users. This switch might be used in a bathroom with a light/fan combination; one switch would control the fan and the other, the light. This setup would require one 12/2 cable coming in to supply power and a 12/3 cable going out.

Note: Putting a bath fan on a timer is preferable to wiring it with an on/off switch. A timer allows the fan to continue running a while after the user has left the bathroom and turned off the light.

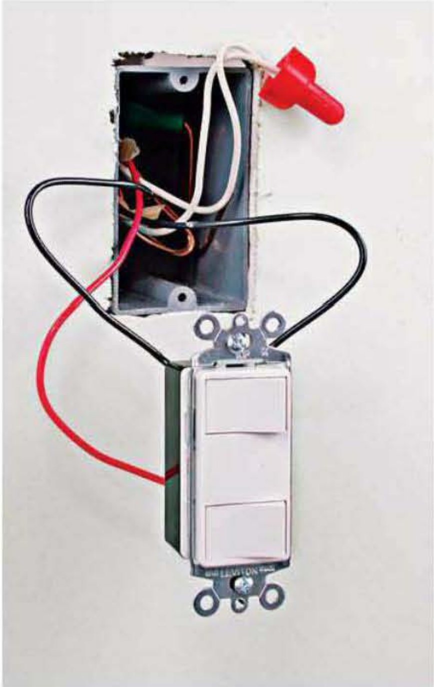

A stacked or double switch controls power to two energy users.

WIRING A MOTION DETECTOR

Motion detectors are specialized switches that sense motion by infrared or ultrasonic sensors and then turn on a light. The unit will remain on as long as there is sound or motion present and for a fixed interval thereafter. Then it will shut itself off automatically.

Because manufacturing details vary widely, it's difficult to generalize about wiring motion detectors; the switch's location in the layout—midcircuit or back fed—will also dictate the type of cables feeding it and how they're connected to switch leads. Some motion detectors require a neutral connection; others don't. The motion detector shown here has no neutral connection. Its green wire is a ground lead, the back lead connects to the incoming hot wire, and the blue wire connects to the switch leg ❶.

This motion detector's sensor is the large "eye" at the top of the unit. Finally, secure the cover plate ❷.

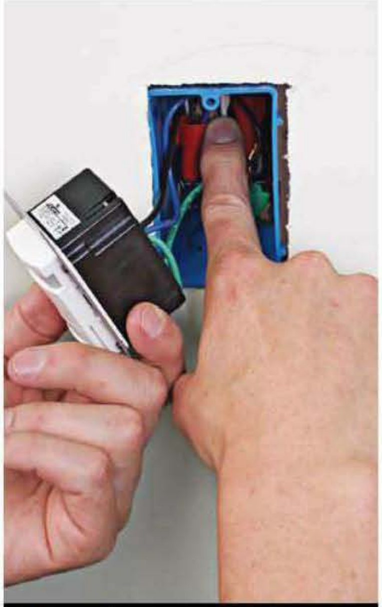

1 Connect the blue wire to the switch leg.

2 Secure the cover plate to protect the components.

WARNING

Specialty switches often combine several functions and so may require more complex hookups. Reading the directions will ensure a correct installation and prevent you from damaging the switch. Switches with solid-state electronic components are particularly sensitive.

WIRING AN ELECTRONIC TIMER SWITCH

An electronic timer switch is a sophisticated piece of equipment. In addition to a control screen and several programming buttons, the device is surrounded by a metal fin called a *heat sink,* which dissipates the heat generated by resistance within the switch, thus prolonging the life of the device. Better quality dimmers and programmable switches tend to have heat sinks.

This timer switch has four leads, including a neutral, so two 12/2 cables feed this box. Start by splicing the stranded green ground lead directly to the incoming ground wire group—no need to run a separate ground pigtail to the switch ❶.

Unlike most mechanical switches, this timer switch has electronic components that require a neutral wire to operate—hence the timer's neutral (white) lead is spliced to a neutral wire group from the cable feeds. Splice the hot (black) wire to the black lead of the switch and the switch leg to the blue wire of the switch ❷. Again, read the directions to be sure you're installing it right.

After carefully folding the spliced wires into the box, hold the face of the switch flush to the wall and screw it into the box ❸. Then install the cover plate ❹. (*Note:* The timer switch's control screen, large override button, and three smaller programming buttons are at the bottom.)

TRADE SECRET

When splicing a stranded wire to solid conductors, as shown in photo 1, place the tip of the stranded wire slightly beyond the solid conductors so that when you twist on the wire connector, its threads will fully engage the stranded wire and ensure a solid splice.

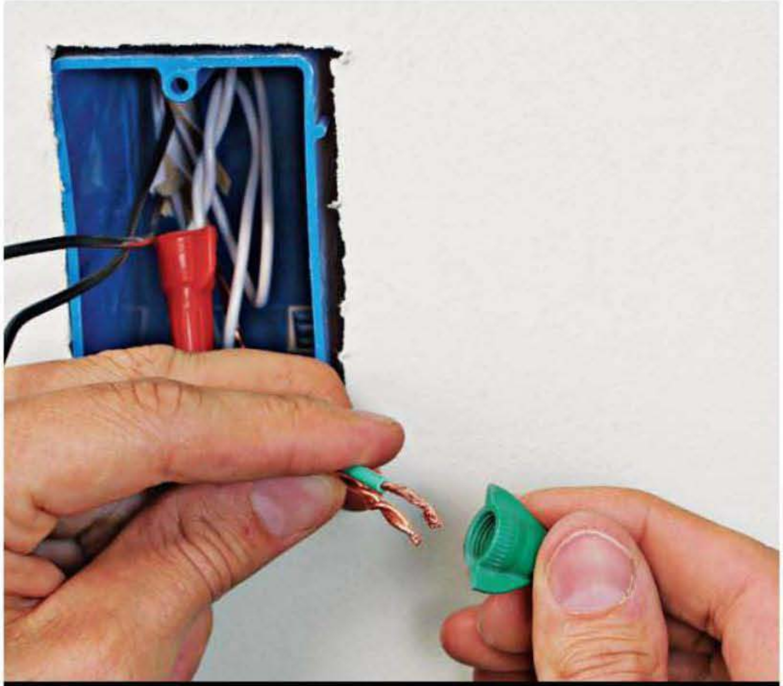

1 Connect the stranded wire ground first.

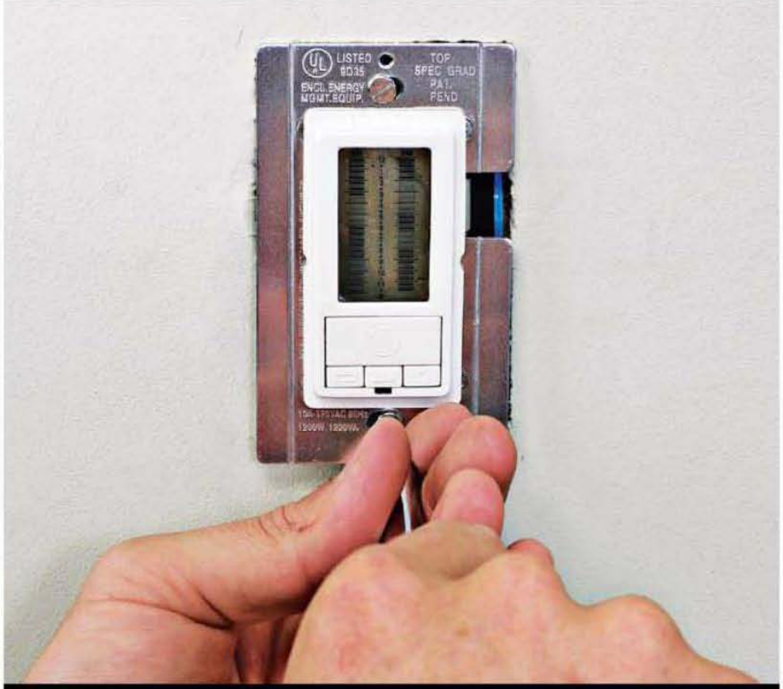

3 Hold the timer face flush.

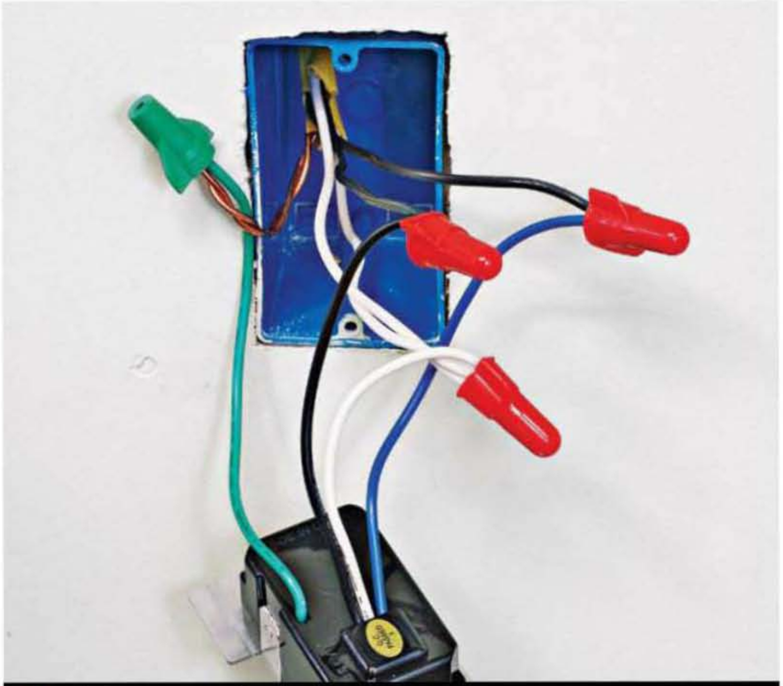

2 This timer requires four leads.

4 Install the cover plate.

WIRING A MANUAL TIMER

Manual timers tend to be inexpensive and simple to wire: They're basically a single-pole switch. The model shown here has no ground lead, and its lead wires attach to a hot wire and a switch leg. Because both leads are black, it probably doesn't matter which lead you connect to the hot wire or switch leg. After installing the cover plate, snap the plastic dial onto the metal post in the middle of the unit.

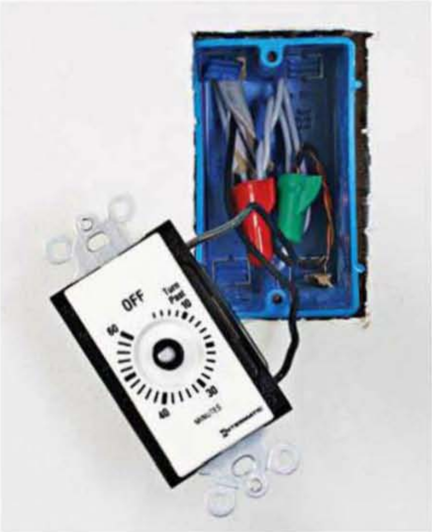

A manual timer turns off a light after a set period of time.

LIGHTING

BASICS

Fixture wiring, p. 68
Bulb types, p. 69

REPAIRS

Testing lamps & removing old wires, p. 70
Rewiring a lamp, p. 72
Replacing a plug, p. 74
Refurbishing a chandelier, p. 75

BASIC FIXTURES

Mounting light fixtures, p. 78
Installing a simple porcelain fixture, p. 80

WALL SCONCES

Removing an existing wall sconce, p. 82
Connecting a new sconce, p. 83
Installing halogen bulbs & accessories, p. 84

CHANDELIERS

Prepping & mounting a ceiling box, p. 85
Disconnecting a chandelier, p. 86
Installing a chandelier, p. 87

RECESSED LIGHTING

Cutting a ceiling hole for recessed lighting, p. 88
Wiring a recessed fixture, p. 90
Securing the can, p. 91

UNDER-CABINET FLUORESCENTS

Undercabinet fluorescent fixtures, p. 92

LOW-VOLTAGE FIXTURES

Low-voltage systems, p. 94
Installing low-voltage cable lighting, p. 95
Anchoring low-voltage standoffs, p. 95
Running cable, p. 96
Tensioning cables, p. 97
Locating the canopy, p. 98
Mounting the box & wiring the transformer, p. 99
Installing the canopy, p. 101
Attaching the feed rods & fixtures, p. 102

MONORAIL TRACK LIGHTING

Installing low-voltage monorail track lighting, p. 103
Locating the track & standoffs, p. 104
Mounting & wiring the transformer, p. 105
Hanging the track, p. 106
Installing the light fixtures, p. 107

FIXTURE WIRING

The temporary lighting socket in the photo below shows the basics of connecting a light fixture to *supply or source wires.* The supply ground wire connects to a green box pigtail, its neutral (white) supply wire connects to the neutral pigtail, and its hot (black) supply wire connects to the hot pigtail.

Connect wires in this order: ground, neutral, and then hot. And disconnect them in the reverse order: hot, neutral, and then ground. In this manner, the grounding system stays connected as long as possible to protect you. Remember to ground metal boxes, mounting plates, and fixtures.

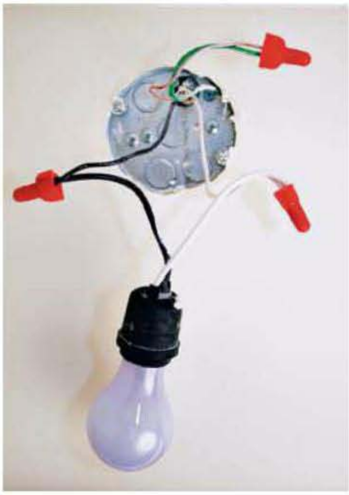

The basics of light fixture wiring. **(The bare copper and green wires are partially covered with white paint.)**

Wiring, Lighting, and Low-Voltage Terms

Hickey
A threaded coupling that joins two lengths of threaded tubing.

Lamp
For most of us, a light fixture that sits on a table or floor. To an electrician, *lamp* is the preferred term for a light bulb: A standard incandescent bulb is thus a "type A lamp." To avoid confusion in this book, a lamp is a fixture, not a bulb.

Leads
Wires preattached to a fixture. Leads are spliced to supply wires to energize the fixture. Typically, a fixture has ground, neutral, and hot leads. Fixture leads are often stranded wire.

Light Box
An outlet box that serves a light fixture.

Line Voltage
The standard current in most house circuits: 120v.

Lumens
A measure of light on the surface of a bulb.

Nipple
Short section of hollow threaded rod.

Lo Vo or LV
Low voltage. These systems are typically 12v but sometimes 24v.

Omnidirectional Flare
Light from a standard incandescent bulb that radiates in all directions; as opposed to a controlled or directional beam.

Primary Wires
In a lo-vo system, wires running from a 120v power source to the transformer.

Running Thread
Hollow threaded rod or pipe in the center of a light fixture, which provides a conduit for wires and a way to connect various fixture parts.

Secondary Wires
Wires running from a transformer to lo-vo tracks, cables, or fixtures.

Source or Supply Wires
Wires from the power source, typically 120v.

Transformer
An electrical device that reduces line voltage to low voltage. All lo-vo lighting systems require a transformer.

BULB TYPES

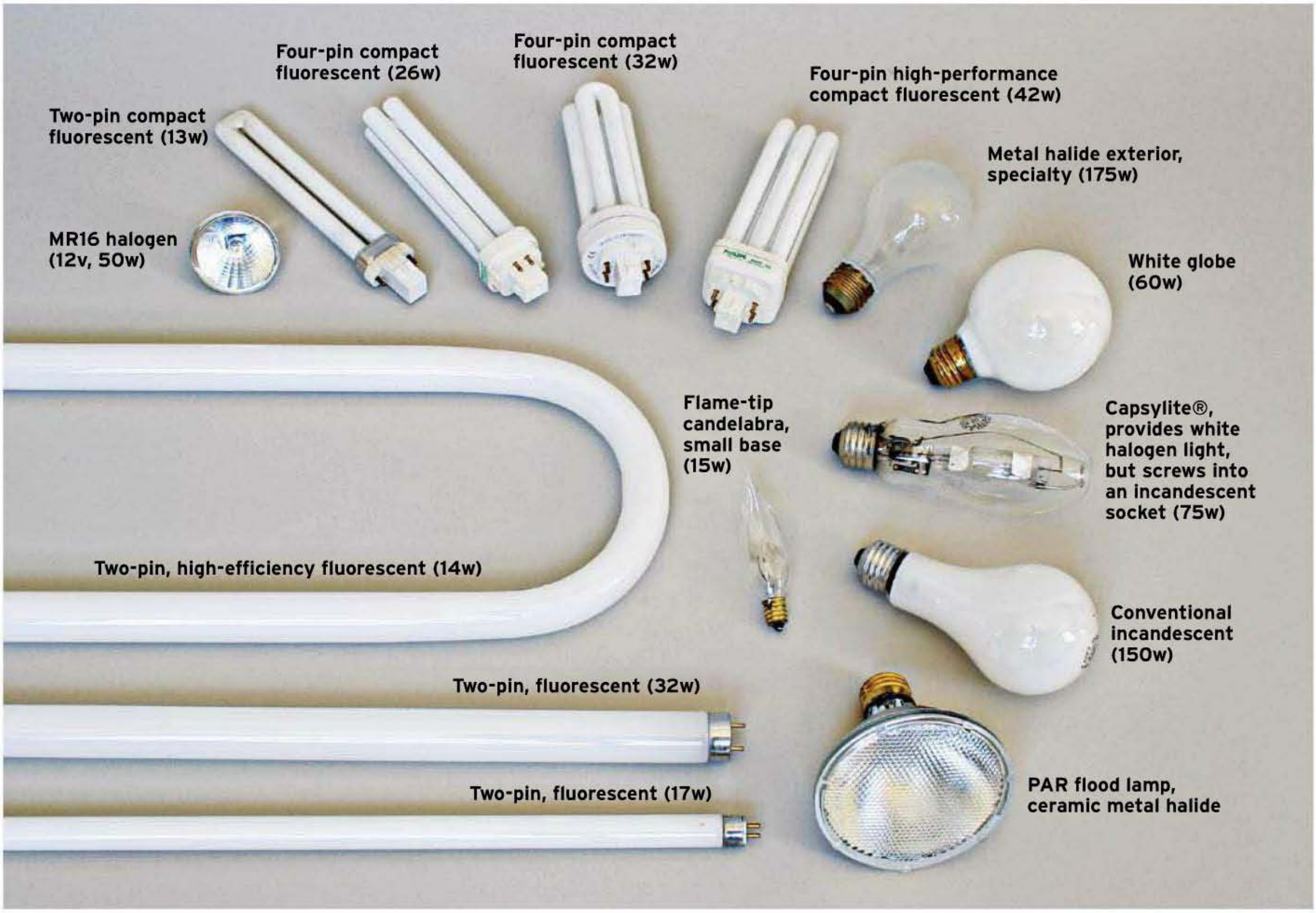
Four-pin compact fluorescent (26w)
Four-pin compact fluorescent (32w)
Four-pin high-performance compact fluorescent (42w)
Two-pin compact fluorescent (13w)
Metal halide exterior, specialty (175w)
MR16 halogen (12v, 50w)
White globe (60w)
Flame-tip candelabra, small base (15w)
Capsylite®, provides white halogen light, but screws into an incandescent socket (75w)
Two-pin, high-efficiency fluorescent (14w)
Conventional incandescent (150w)
Two-pin, fluorescent (32w)
Two-pin, fluorescent (17w)
PAR flood lamp, ceramic metal halide

There are more bulb types than there's room to describe them, so we'll stick to the three most common types here: incandescent, fluorescent, and halogen. When changing burned-out bulbs, make sure that the replacement bulb fits the fixture socket because, even within each bulb type, there are variations in diameter, pin size (for fluorescents and halogens), and so on. You should *never* have to force a bulb to make it fit a socket.

Incandescent bulbs

Incandescent bulbs screw into threaded sockets and complete a circuit by making contact with the aluminum or brass shell (neutral) and the contact tab (hot) in the bottom of the socket. Typically, the socket is energized with house current (120v). Incandescent bulb wattage is stamped on the top of the bulb.

Fluorescent bulbs

Fluorescent bulbs have contact pins at either end that slide into fixture sockets and then rotate into final position; two-pin bulbs plug into magnetic ballasts, four-pin bulbs plug into electronic ballasts. A fluorescent tube increases wattage as it increases in length, it can also coil to increase wattage. Fluorescent bulbs are more energy efficient than incandescent ones, so many building codes now mandate that a certain percentage of house lighting be fluorescent.

Halogen bulbs

Halogen bulbs can be installed in both standard and low-voltage (lo-vo) systems. But even on reduced current, they burn bright—and hot. The filament in a halogen bulb reaches in excess of 1200°F, so be sure to install these bulbs where they can't be accidentally touched.

TESTING LAMPS & REMOVING OLD WIRES

If a lamp flickers or doesn't light at all, you can perform a series of tests to identify the problem. A chandelier is more complicated because it will have several sockets, but investigating its problems is essentially the same. Perform these tests with the lamp unplugged or the chandelier disconnected.

Scrub the inside of an old socket with steel wool to remove corrosion and improve electrical contact. Remove any steel wool particles as they can cause shorts. Use needle-nose pliers or a screwdriver point to lift the contact tab in the middle of the socket. Plug the lamp in and screw in a light bulb to see if these repairs help. If not, test the socket.

If necessary, press the sides of the socket shell to pop it out of the base. On most brass shells, "press here" is typically stamped on the socket shell. Slide off both the shell and the cardboard liner. Pull the socket loose ❶. You can conduct the next test with the wires disconnected from the socket, but you don't need to remove them yet. Attach the continuity tester clip to the socket's brass screw terminal (hot) and the tester probe to the tab in the bottom of the socket. Turn the switch off and on ❷. If the tester doesn't light, the socket is defective and should be replaced.

If the socket is okay, test the cord. Turn the socket on and connect the tester clip to the narrow plug prong and the tester point alternately to either the bare wire ends held down by the screws or to the contact tab and the shell. Touch the tester point to both wires (or to the tab and shell), move the clip to the wide prong, and repeat the test ❸. If there's no continuity in either (or both) of the wires, there's a break in the wire or faulty connections to the plug. Replace the cord and plug.

If the lamp base is metal, test for shorts by attaching the tester clip to the narrow (hot) prong of the plug, and touching the tester point to the lamp metal base. Switch the lamp off and on. If there's a short, the tester will light ❹. If the plug is not polarized. repeat the test with the other prong.

To replace the switch, unscrew the wires from their terminals and install a new switch ❺. However, if you're replacing the cord and plug as well, simply snip one end of the lamp cord and pull the socket and cord out of the lamp.

To replace entire socket assembly, unscrew the set screw that holds the socket base to the threaded rod that runs through the center of the lamp. Then turn the socket base counterclockwise to remove it ❻.

Use steel wool to clean and improve electrical contacts.

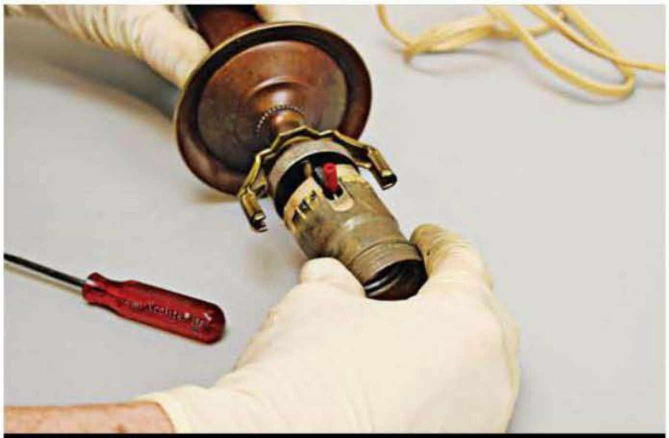

1 **Slide off the socket shell and cardboard liner.**

TRADE SECRET

Take photos of lamps and fixtures before you disassemble them, and the fixtures will be easier to put back together. To avoid losing hard-to-find and tiny parts, put similar pieces in plastic bags and label them.

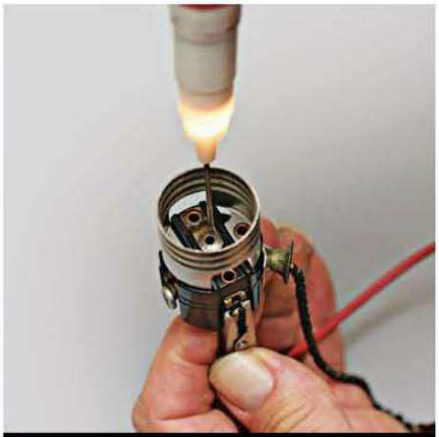

2 Test the brass screw and socket tab.

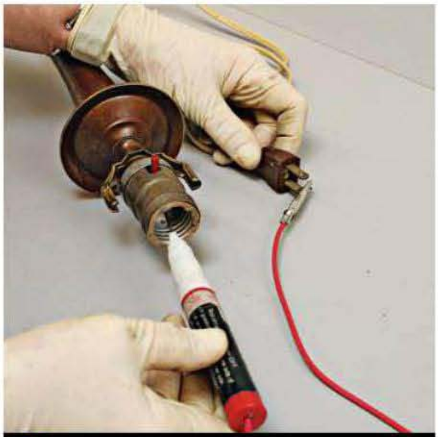

3 Test the cord for continuity.

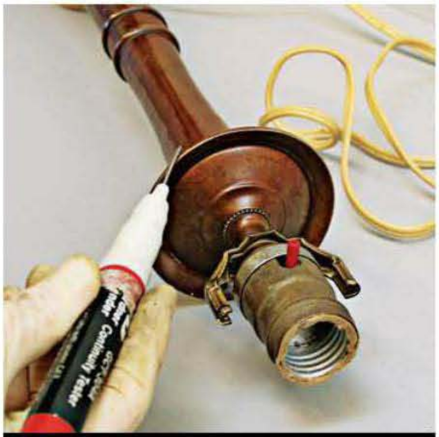

4 Test metal lamps for shorts.

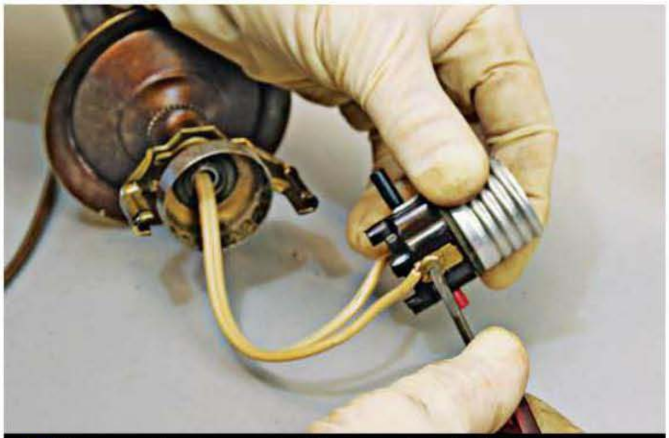

5 Unscrew wires from each of the screw terminals.

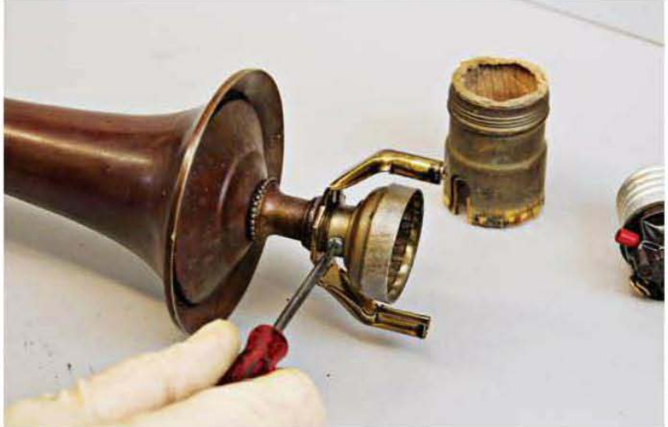

6 Unscrew the socket base from the threaded rod.

Work safely Unplug lamps or shut off power to chandeliers before working on them. If a chandelier is presently connected to a circuit, use a voltage tester to make sure the power is off before disconnecting the fixture. If a fixture is heavy, by all means get help supporting it and lowering it after it has been safely disconnected.

Wear work gloves to protect your hands; wires, sockets, and other metal parts are often sharp. Many repair pros favor disposable rubber or latex gloves, which are flexible enough for fine work, but impervious to cleaning solvents and toxic metal dusts. If you use power tools to drill, grind, or polish metal or if you solder connections, wear a respirator mask and eye protection.

REWIRING A LAMP

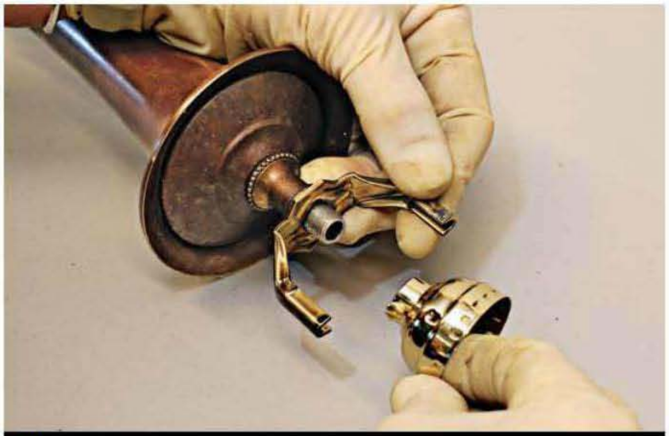

1 Screw on the new socket base.

2 Pull the clipped end of the replacement cord through the lamp.

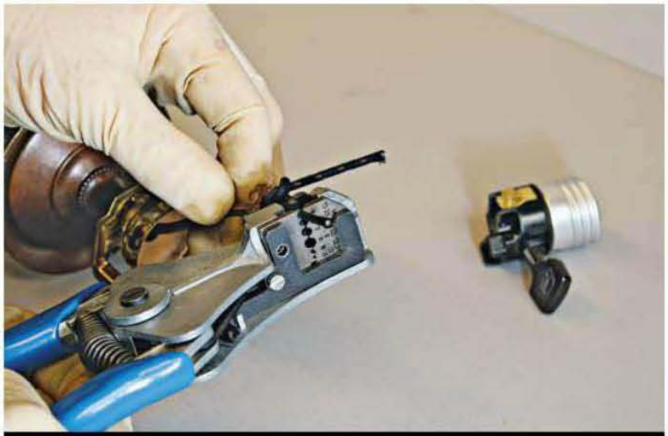

5 Strip the wires. Here, an automatic stripper is used.

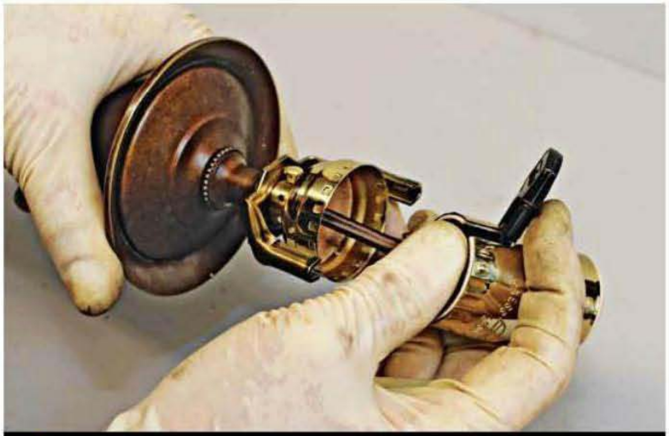

6 Fasten the wires to the screws, press them into the socket, and assemble the liner and shell.

After the old wires have been removed, a fixture is ready for rewiring. Screw the new socket base onto the threaded rod until the base is snug against the curved harp saddle (the harp, when fitted into its saddle, supports the lamp shade) ❶. Cut the replacement cord to length. Typically, that's about 9 ft.: 6 ft. of cord from the lamp to the plug, 2 ft. for the cord hidden inside the lamp, and roughly 1 ft. above the socket, which you'll strip, tie, and trim ❷.

If the cord has a fabric sheathing, use a utility knife to slit it along its length. Then pull apart the two wires inside the cord ❸. Next, tie an underwriter's knot, which prevents the cord from being pulled back into the lamp and stressing connections to switch terminals ❹. Trim the wire so there's about 1½ in. above the knot. Then strip the wire sheathing to expose bare wire. Although you can use a standard wire stripper, the automatic wire stripper shown here severs and pops off the sheathing with one squeeze of the tool ❺.

Twist and loop the bare wire ends. Then place the loops on the screw terminals of the switch—make sure they loop clockwise. Tighten the screws down, using your thumb to press the wires into the bottom of the socket. *Note:* Attach the marked or ridged wire to the silver (neutral) screw on the switch. Slide the cardboard liner over the wired switch. Then slide the socket shell over the liner ❻. Rock the shell into its base until you hear two clicks ❼. After snapping a socket shell into its base, gently wiggle the shell to make sure it's seated solidly.

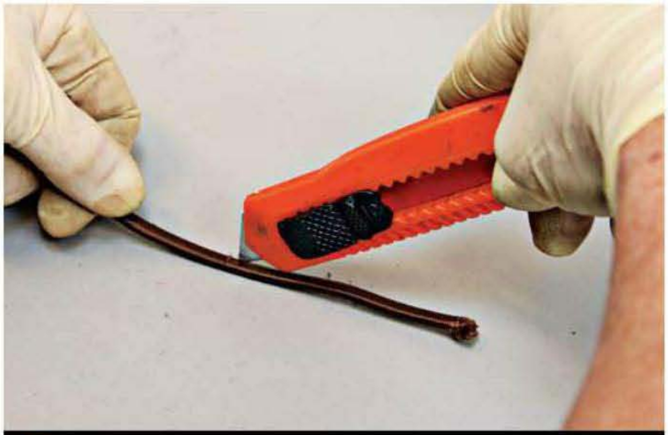

3 Slit the lamp cord with a utility knife. Then separate the wires.

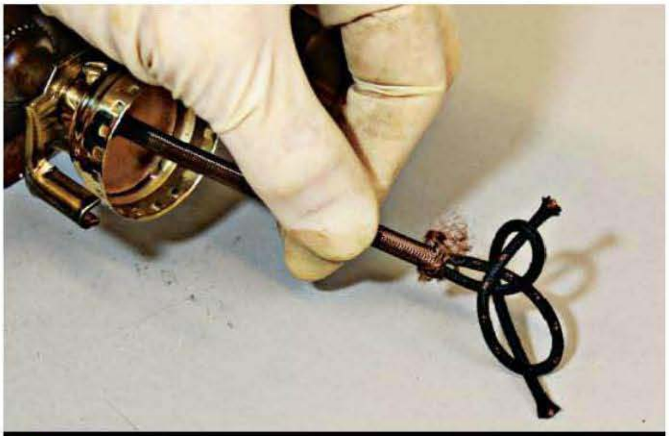

4 Tie an underwriter's knot to secure the cord in the lamp.

7 With completely new wiring, this old lamp is as good as new.

Tinning Tinning, often referred to as soldering, stranded wires is not absolutely necessary, but pros do it because it fills in spaces between wire strands. In effect, soldering makes stranded wire solid, so it's unlikely to crush and spread out when a screw head tightens on it. Tinned strands are also easier to twist wire connectors onto. *Note:* Use only rosin-core solder for electrical connections; acid-core solder will corrode them. And make sure you solder only on a heat-resistant surface.

Soldering makes stranded wire solid and easier to work.

REPLACING A PLUG

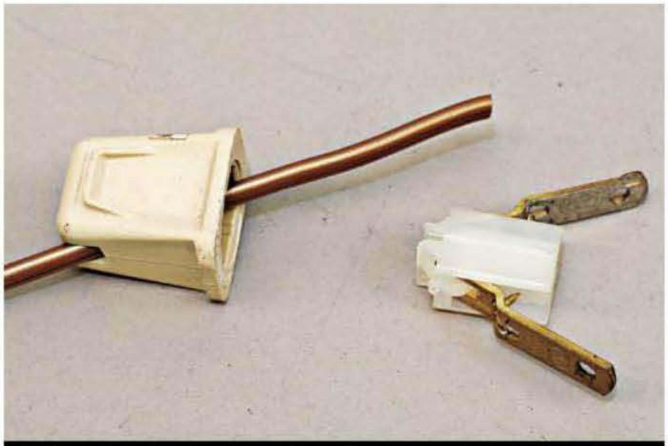

1 The points of a quick plug pierce the lamp cord.

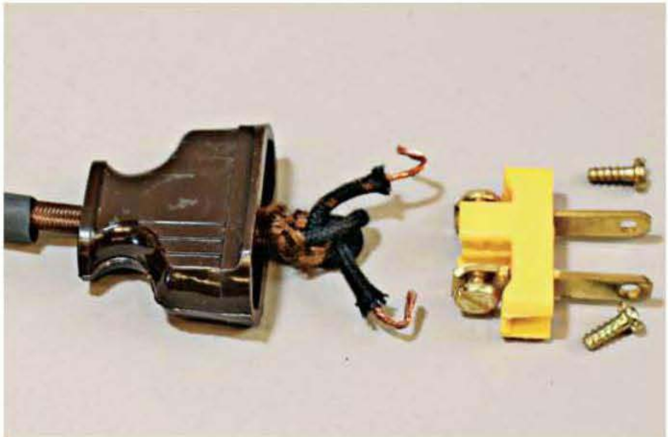

2 This plug is more traditional and comes with screw terminals.

You should replace any plug that is cracked, difficult to remove, or whose cord is damaged near the plug. All lamp cord is 18 gauge and will fit any type of lamp plug. If the plug's prongs are different sizes—one prong is wider—then the plug is polarized.

Individual wires in lamp cord are also differentiated so you can attach them correctly to switch and plug terminals.

If the lamp cord is plastic sheathed, the wire to be connected to neutral terminals will be ridged; the wire to be connected to hot terminals is smooth. If the cord is sheathed with fabric, remove the sheathing and you'll see that one wire is striped or marked in some way—that's the neutral.

There are two main types of lamp cord plugs. Connect plastic-sheathed lamp cords to a "quick plug" ❶. Cut the end of the cord square, thread it through the plug shell, and insert it into the body of the plug. As you squeeze the plug prongs together, two sharp points on the prongs pierce the sheathing to create an electrical connection. Slide the plug shell over the body to lock the cord and prongs in place.

If you have fabric-sheathed cord, however, use a plug with screw connections. Strip about 1 in. of fabric sheathing, then strip about 1/2 in. of insulation off each wire end to expose bare wire. Thread the wire through the plug body, tie an underwriter's knot, loop each wire clockwise around a screw terminal, then tighten the screws ❷.

Polarized fixtures

Light fixtures parts are polarized, so they fit together only one way. On a correctly wired light fixture, the neutral wire of the lamp cord connects to the silver screw on the socket, whereas the hot wire connects to the brass screw on the socket.

If the hidden parts of the socket were visible, you'd see that the silver screw makes contact with the sides of the socket shell (the part that the bulb screws into), and the brass screw makes contact with the tab in the bottom of the socket. Should you inadvertently touch the side of the shell while changing a bulb, there's no harm done if the socket is correctly wired. However, if you reverse the order in which wires are attached—known as reversing the polarity—the outside of the screw shell becomes hot (energized) and could shock someone changing a bulb. Make sure you have the wires properly connected, paying attention to the striped, ribbed, or otherwise marked neutral wire and connecting it to the neutral (silver) screw.

Finally, the neutral wire of lamp cord connects to the wider prong of a plug; the hot wire connects to the narrow prong of a plug.

TRADE SECRET

Near the plug, slide a piece of heat-shrink tubing over the end of a fabric-covered lamp cord to keep it from unraveling.

REFURBISHING A CHANDELIER

Chandeliers vary widely, especially those designed in the early decades of the 20th century. Fortunately, you can get replacement parts for many chandeliers and substitute modern equivalents for many other parts. Still, it's wise to save all original parts of a fixture. Photograph the fixture before you take it apart, and you'll know what goes where when it's time to reassemble it.

Survey the chandelier for missing or damaged parts. As you disassemble the unit, put like parts in plastic bags and label them. In many cases, old switch and socket assemblies will still work. Use a continuity tester to test old switches and lightly rub steel wool inside sockets to remove corrosion and improve electrical contacts.

Old wiring is rarely worth saving, however. At best, it will be brittle and, at worst, broken or unsafe. Snip the wires running to each socket ❶ (see p. 76) so you can remove the sockets from the shade holders. Removing sockets takes finesse and patience. Socket shells snap into socket bases, so rock the shells gently from side to side until they unsnap. Then angle the shells out of the holder ❷ (see p. 76). Chances are, the cardboard liners (sleeves) inside the shells will have deteriorated and should be discarded.

In many older chandeliers, source and switch wires are spliced together and housed in the fixture body. To disconnect the body, unscrew the finial at the bottom. Separate the cover (top) and pan (bottom) of the body, disconnect the wire splices and remove the old wires. Be sure to save the nipples (short threaded rods) and hickey (U-shaped coupling), because you'll need them to reassemble the chandelier body.

Rewiring the fixture

Before rewiring a chandelier, refinish it or simply clean up its parts. These days, leaving the surface patina intact is in vogue; the fixture seen here was washed with a simple household cleaner, rubbed with steel wool to

>> >> >>

WIRING A VICTORIAN LIGHT FIXTURE

Connecting incoming wires to fixture leads is standard: hot to hot, neutral to neutral. Grounds are spliced and connected to metal boxes and fixture ground screws, if any. *Note:* Fixture bodies and mounting devices vary considerably.

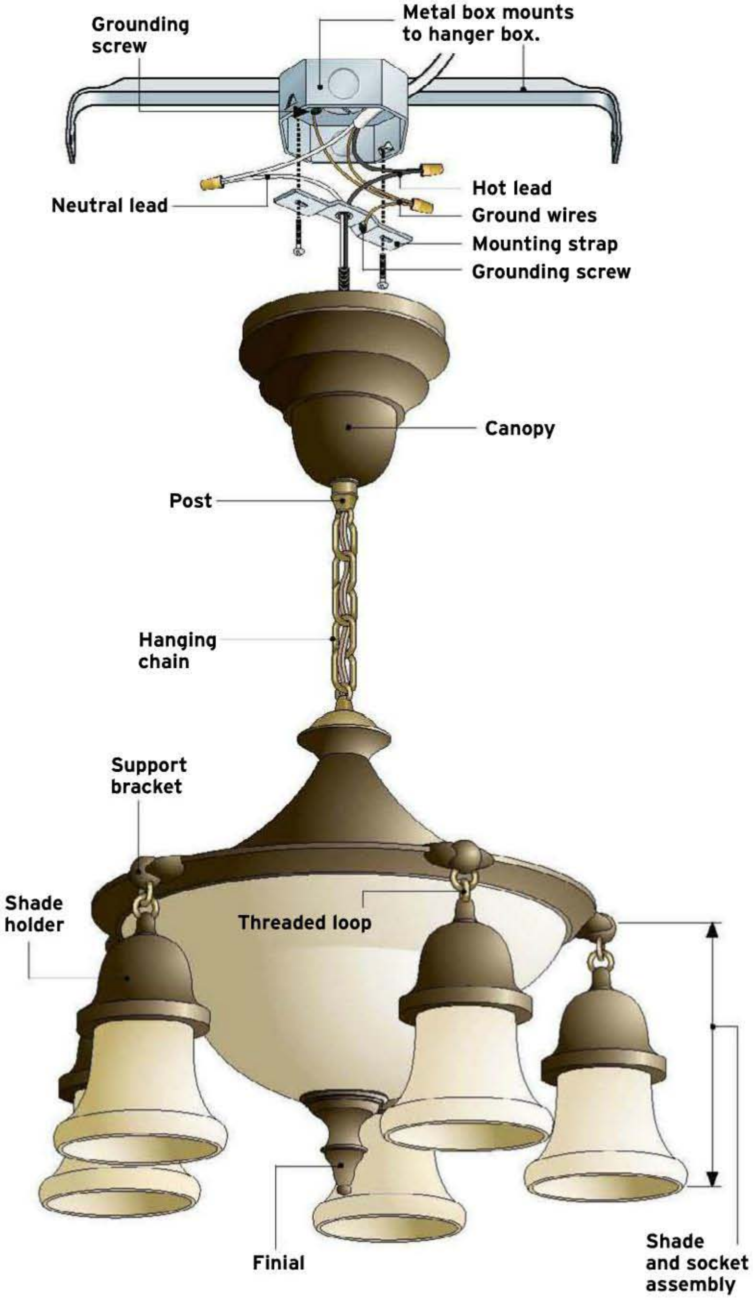

REFURBISHING A CHANDELIER (CONTINUED)

remove rusted spots and other obvious flaws, and then lacquered. Because the fixture's sockets still worked after 80 years, the restorer decided to reinstall them and wire them with rayon lamp cord. Use a utility knife to remove about 1½ in. of the fabric sheathing, and then use wire strippers to strip ½ in. of insulation off the individual wire ends. Loop the wire ends and place them clockwise onto the socket screws—in the same direction that the screws tighten ❸. For most lamp cord, individual wires are different colors; in this example, the installer made the blue wire neutral and the brown wire hot. It doesn't matter what color you designate hot or neutral, as long as your designations are consistent.

After wiring each socket, slide on a new cardboard liner to insulate the screw terminals. The socket base should also have a new liner ❹. Snap each socket shell to a base. Then use a small screwdriver to tighten the threaded part of the socket base to a male threaded loop linked to the fixture body. (A screwdriver won't mar the fixture surface as pliers would.) ❺

Feed the wires from each socket into the fixture body ❻. Group the neutral and hot wires from all the sockets. Then splice each group to the main lead wires, which you'll eventually connect to the source wires (neutral to neutral, hot to hot). Although it's not essential, professionals prefer to tin (solder) the ends of stranded wires, using rosin-core solder, so the wires will splice better. Use wire connectors to splice each group, tug gently to be sure wires are solidly spliced, then wrap each splice generously with electrician's tape ❼.

Reassemble the fixture

Once you've spliced socket wires to main lead wires, feed those leads through the opening in the hickey and into the threaded rod that runs through the top of the fixture body ❽. Then stack the chandelier pan, cover, and related parts onto the hickey-and-threaded-rod assembly ❾.

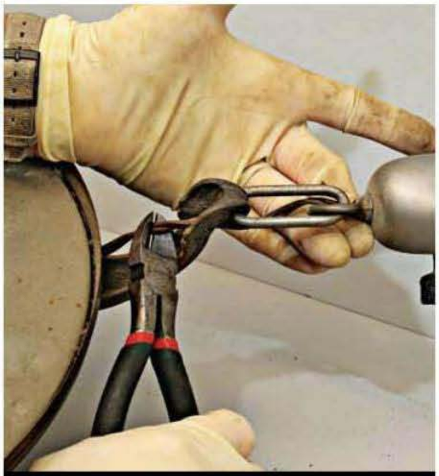

1 Snip old wires.

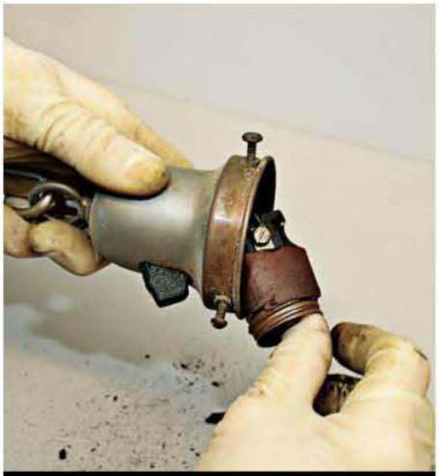

2 Carefully rock the sockets out of their holders.

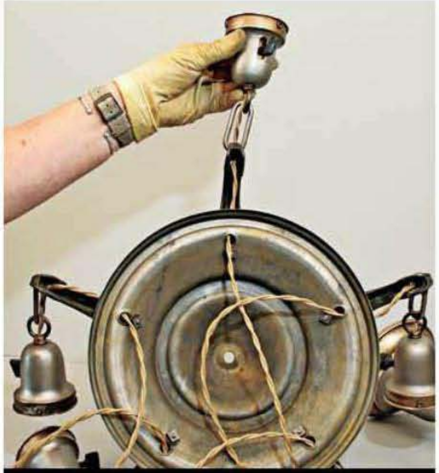

6 Feed the socket wires into the body of the fixture.

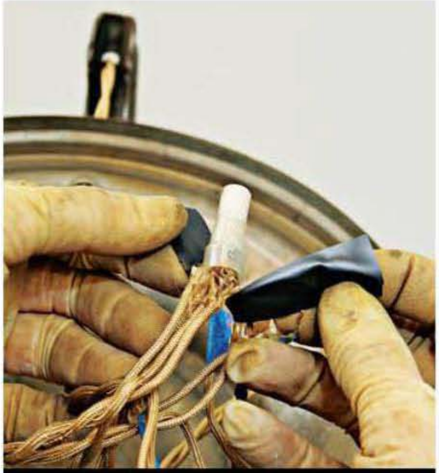

7 Wrap each of the splices with electrical tape.

Chandelier assemblies vary, so be guided by the photograph you took of your fixture before disassembling it. Typically, a finial screws to the bottom nipple, and a female threaded loop to the top nipple. After the main leads emerge from the top nipple and loop, thread the leads through the hanging chain that supports the chandelier body. Then feed the leads into the hollow post that the canopy screws to. Leave roughly 1 ft. of free cord above the post, to strip and attach to source wires. (The power will be off, of course.)

Chandeliers' mounting details also vary, so, again, be guided by your fixture. If your old fixture has a crow's foot, remove it, because it was intended to screw directly to a ceiling joist and doesn't allow much flexibility. Replace it with a standard mounting bar, which screws to a properly rated

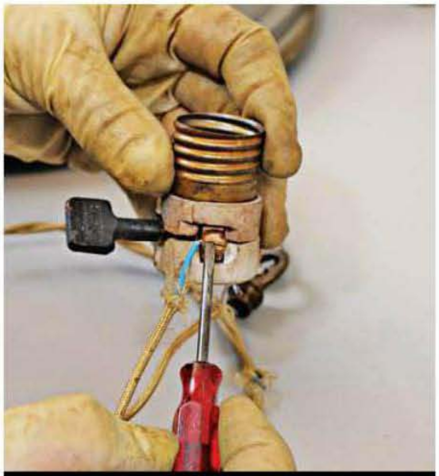

3 Connect the new wires to the old socket.

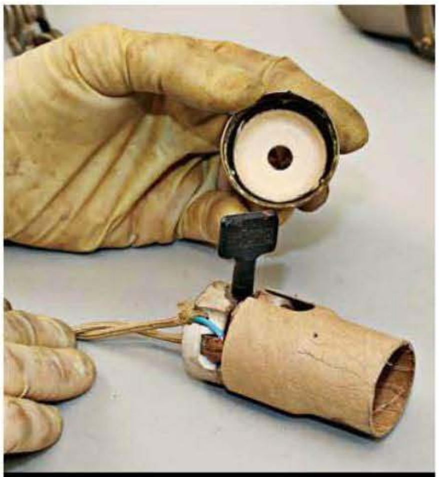

4 Install the liners on the socket and base.

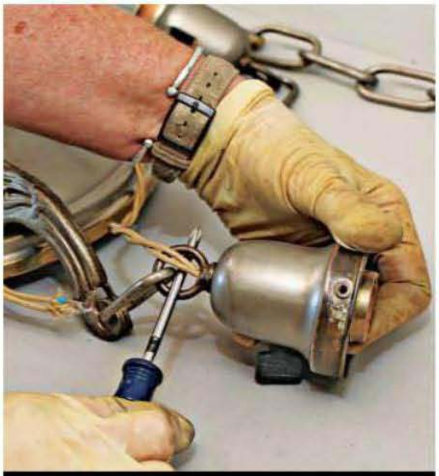

5 A screwdriver tightens a socket base without marring it.

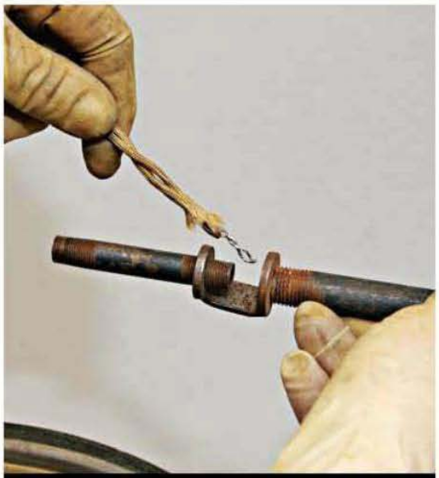

8 Feed main leads into the hickey.

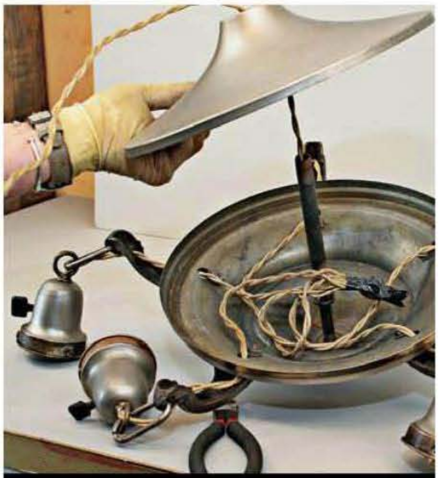

9 Stack the various parts of the fixture onto the threaded rods.

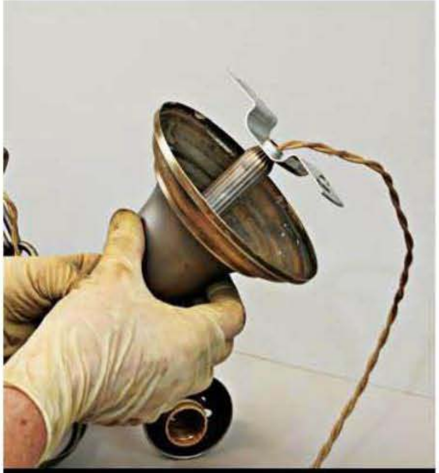

10 Remove the old crow's foot mount, replace with new mounting bar.

ceiling box. The threaded post under the canopy screws into a threaded hole in the center of the mounting bar ⑩.

Shades are usually held in place with three small thumbscrews; tighten them just snug and then back off the screws a fraction so the shades will have room to expand without cracking when they heat up ⑪.

11 The refurbished chandelier is as good as new but has an antique charm.

MOUNTING LIGHT FIXTURES

CEILING FIXTURE ELEMENTS

In this basic setup, the ceiling box mounts to an adjustable bar, which is screwed to ceiling joists. The fixture, in turn, screws to a mounting bracket, which screws to the ceiling box. All metal boxes and brackets must be grounded to be safe. Many electricians use grounding screws in both the box and the bracket, but one ground is sufficient: The metal mounting screws provide grounding continuity to box and bracket.

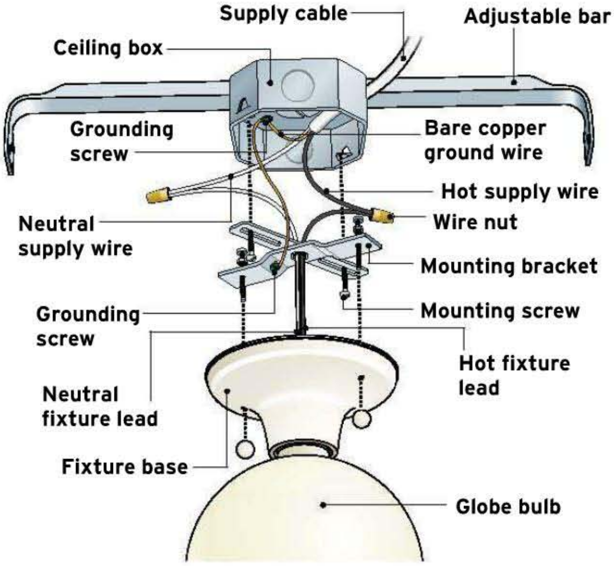
Supply cable
Adjustable bar
Ceiling box
Grounding screw
Bare copper ground wire
Hot supply wire
Neutral supply wire
Wire nut
Mounting bracket
Grounding screw
Mounting screw
Hot fixture lead
Neutral fixture lead
Fixture base
Globe bulb

Fixture boxes must be mounted to framing to adequately support the weight. (Boxes that are not solidly mounted are a hazard because in time they'll move, and that movement could compromise electrical connections inside.) There are many mounting options for boxes: The main choice is whether you nail or screw the box directly to a stud or ceiling joist or use an extendable mounting bar to which the box attaches, as shown here.

Either method works fine, but because the box slides along a mounting bar, you can position the box—and hence the light fixture—where you want. The other choice for a recessed can, which is also adjustable, is a housed unit such as that shown in the bottom photo on the facing page.

Mounting fixtures to boxes

If mounting screws on all light fixtures were exactly the same diameter and spacing as the screw holes on all boxes, life would be simple and you'd screw the fixture directly to the box. But there are many different box sizes and configurations, and light fixtures vary considerably. Consequently, there are many mounting brackets to reconcile these differences. Always examine existing outlet boxes before buying new fixtures and make sure that fixture hardware can mount to existing boxes. Otherwise, a routine installation could turn into a long, drawn-out affair with a lot of trips to the hardware store.

MOUNTING BRACKETS

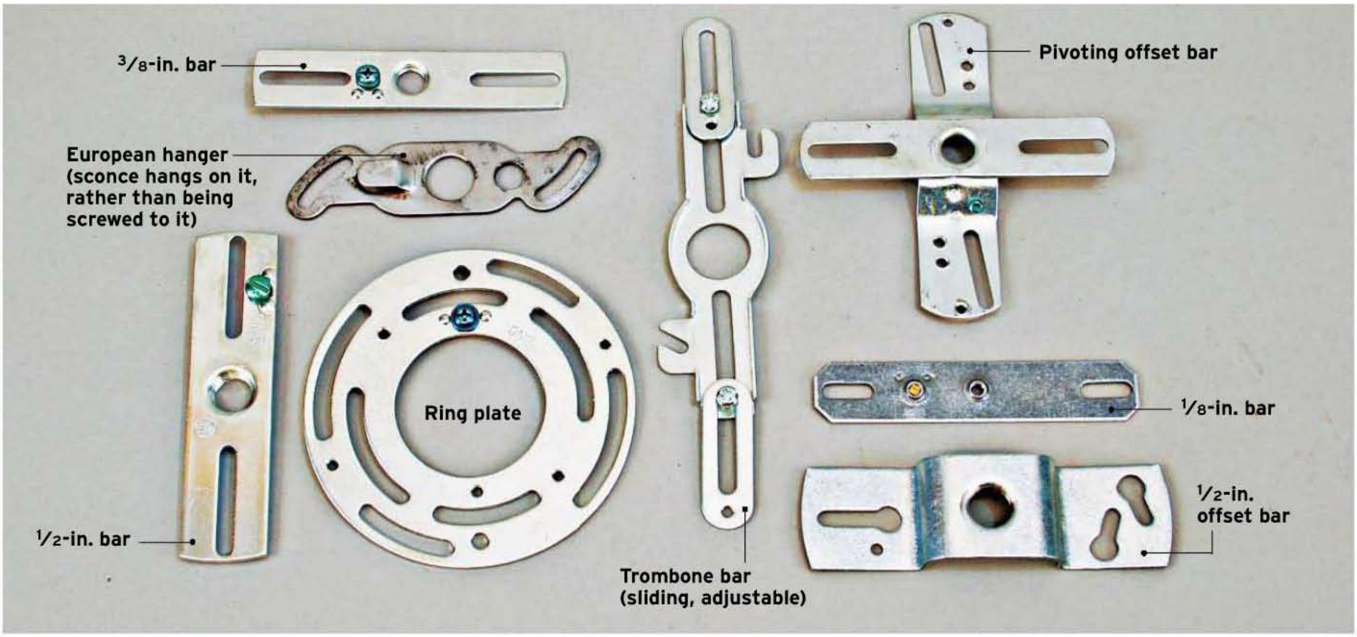
3/8-in. bar
European hanger (sconce hangs on it, rather than being screwed to it)
Ring plate
1/2-in. bar
Trombone bar (sliding, adjustable)
Pivoting offset bar
1/8-in. bar
1/2-in. offset bar

What follows is an overview of how various fixtures mount to outlet boxes. Later in the chapter (see p. 85), we cover how to attach a standard octagonal outlet box for a ceiling fixture.

All metal brackets, boxes, and lamp fixtures must be grounded to be safe. The green grounding screws have a specified thread count to ensure a positive connection to metal boxes or plates.

Flat-mounting brackets

Typically, a mounting bracket screws to an outlet box, and the fixture attaches to the bracket, either by machine screws or, as is more common for chandeliers, by a threaded post that screws into a threaded hole in the center of the mounting bracket. Brackets can be as simple as a flat bar with screw slots; but some adjust by sliding, whereas others are offset slightly to provide a little more room for electrical connections—and fingers. Ring brackets can be rotated so the slots line up perfectly with outlet box and fixture screw holes.

Even simple brackets give you several mounting options. The flat bar shown in the top photo at right, for example, is slotted to receive fixture machine screws. It also receives a threaded nipple to which a chandelier will mount.

➔ **For more on mounting ceiling boxes, see p. 85.**

Matched brackets

Some mounting brackets are specifically matched to a fixture, as with the clever martini wall sconce shown on p. 83. Because the fixture designer didn't want screws on the face of the fixture base, he specified a flanged ring bracket, which receives screws on the side, where they'll be less visible.

No brackets

Some fixtures, such as the recessed lighting fixture shown on p. 88, don't require a mounting bracket. The fixture's can (housing) is its own junction box; and, once inserted into a hole cut in the ceiling, the fixture is supported by the ceiling it sits on. The fixture can is further secured by integral clips and trim pieces that pull it tight to the plaster or drywall ceiling. Recessed cans are IC rated (they may be covered with insulation) or non-IC-rated (cannot be covered with insulation). (*IC* stands for "insulated contact.")

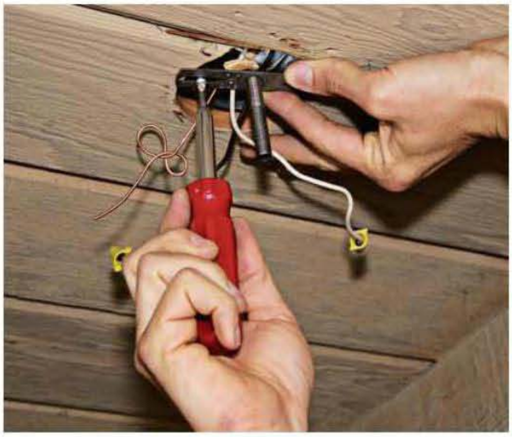

Screwing a mounting bracket with a threaded nipple to a ceiling box.

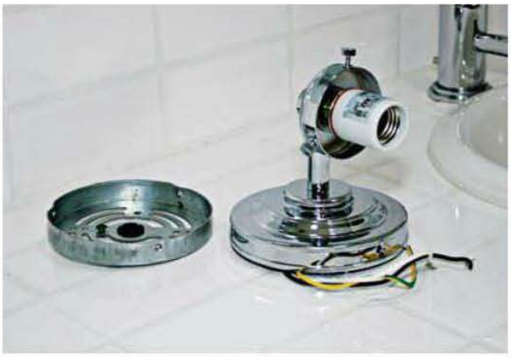

This wall sconce requires a special flanged ring bracket to mount it.

An IC-rated incandescent fixture can be covered with insulation.

WARNING

Most bar and box assemblies are rated for 50 lb; if your light fixture weighs more than that, install a fan box instead.

INSTALLING A SIMPLE PORCELAIN FIXTURE

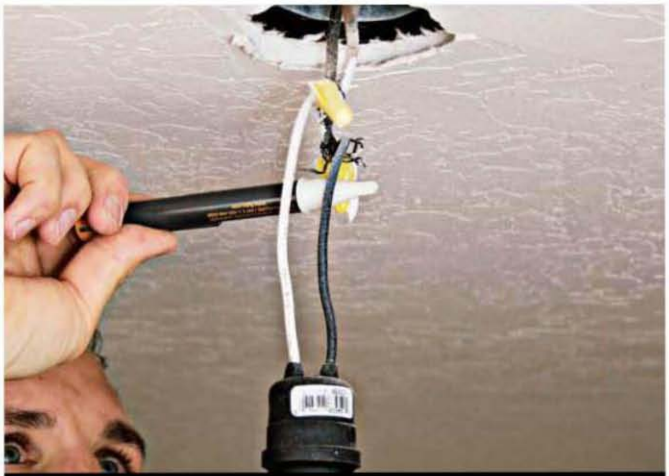

1 Test the hot wire for power.

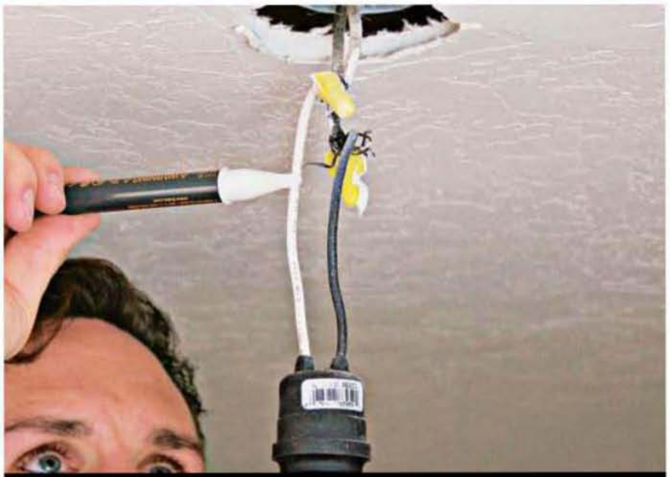

2 Then test the neutral wire.

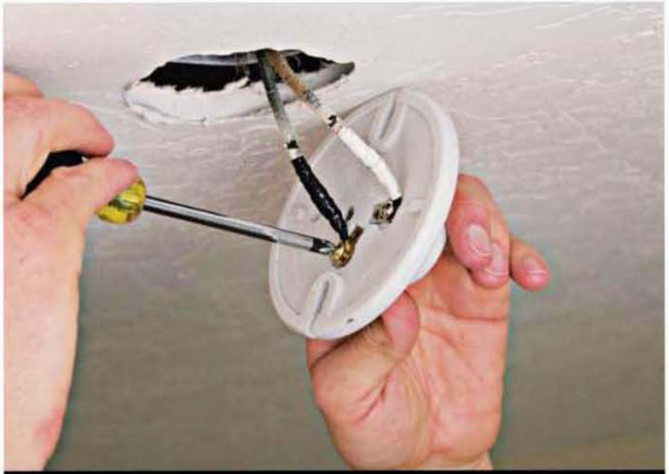

5 Connect the supply wires to the screw terminals.

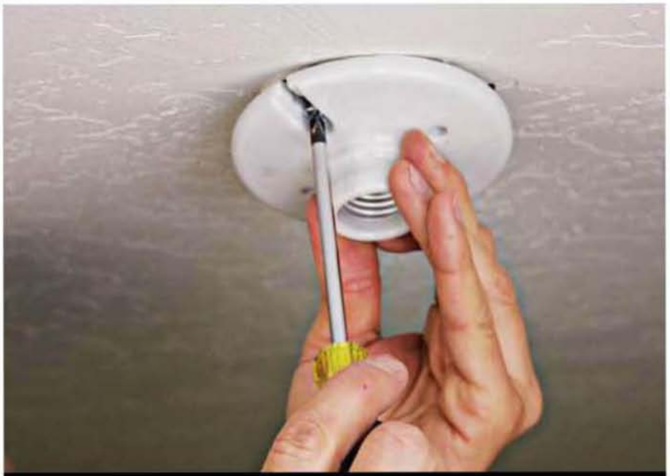

6 Screw the fixture to the box.

Before replacing or installing any fixture, shut off the power to the outlet and test to be sure it's off. Use an inductance tester to test both hot ① and neutral ② wires for power because there's no guarantee that the fixture was wired correctly.

If the existing cable has fraying fabric sheathing but the thermoplastic insulation around individual wires is intact, it's safe to reattach the wires to a new fixture. To keep the fabric from fraying further, wrap the sheathing with electrical tape. Use black tape on the hot (black) wire ③ and white tape on the neutral (white) so there will be no future confusion about which wire is which ④. Of course, if the sheathing is intact, it's not necessary to wrap wires with tape.

Connect wires to the appropriate screw terminals on the fixture: neutral wire to the silver screw, hot wire to the brass screw ⑤. If a fixture has two or more mounting screws, start each screw before tightening any one screw all the way down. It will be easier to line up the screws to the holes in the outlet box or mounting bracket ⑥.

With porcelain fixtures, don't overtighten the screws that hold it to the wall or ceiling. Likewise, never overtighten a bulb. A snug fit is fine ⑦.

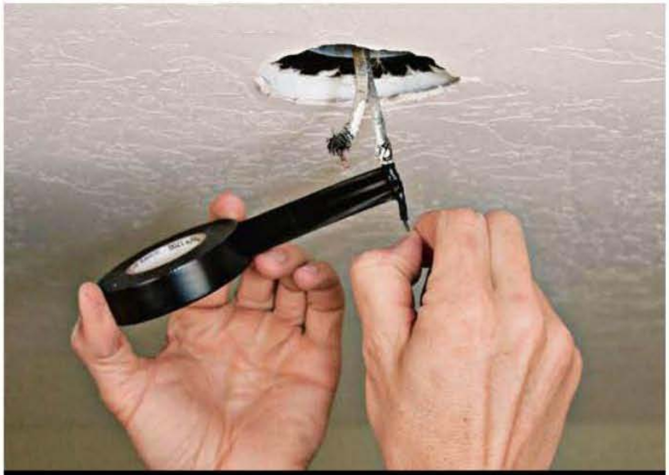

3 Wrap black tape on hot wire.

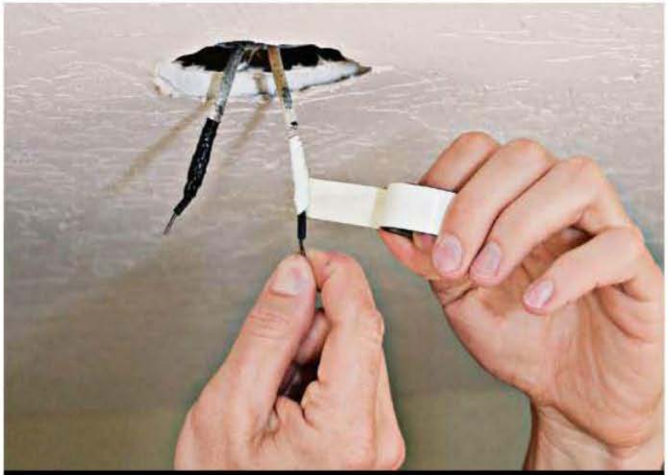

4 White tape marks the neutral wire.

7 Don't overtighten the bulb.

PLAY IT SAFE!

Before working on any electrical fixture, shut off power to the circuit by flipping the breaker or removing the fuse controlling that circuit. Then use an inductance tester to be sure the power is off.

Electrical codes require that all fixtures and devices—everything that gets installed—must be "listed" and must be recognized by the authority having jurisdiction (AHJ), usually the local inspector checking your installation. (Typically, light fixtures will have an Underwriters Laboratories [UL] listing.) If an inspector doesn't see a UL sticker, he or she could ask you to remove the fixture.

Pay attention to a fixture's wattage rating, usually specified on a sticker on the fixture's base. Substituting a bulb with a higher wattage can overheat and damage the fixture and, in some cases, ignite nearby combustible surfaces.

WARNING

Porcelain fixtures have endured for decades because they're inexpensive and reliable. But they are relatively fragile. Avoid overtightening the machine screws that hold a porcelain fixture base to a box, because if you fracture the base you'll have to replace it. When the screw head is just snug, stop turning.

REMOVING AN EXISTING WALL SCONCE

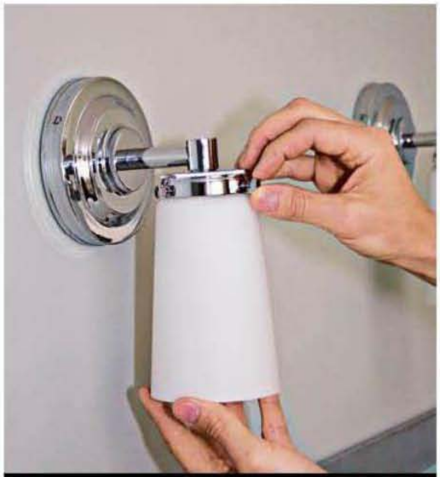

1 Support the shades as you unscrew their holding screws.

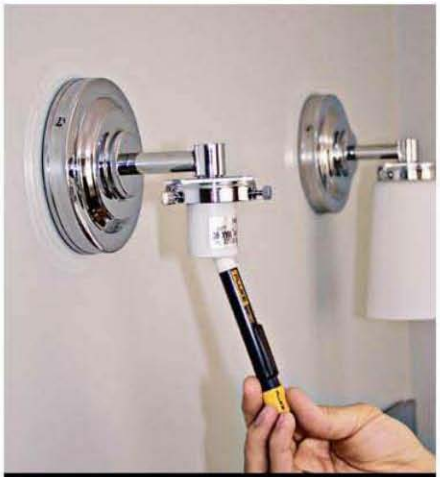

2 Test to make sure the power is off.

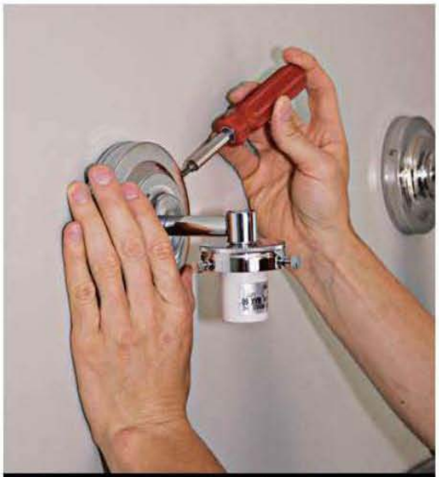

3 Unscrew the existing fixture.

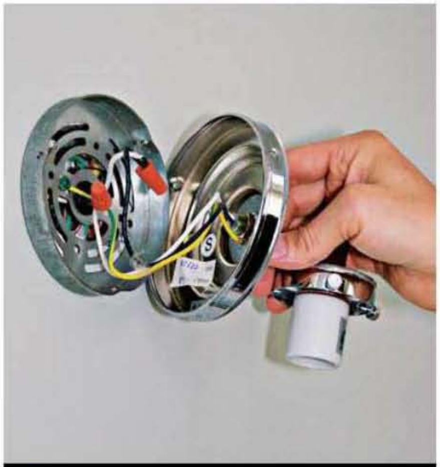

4 Remove the fixture to expose the wires.

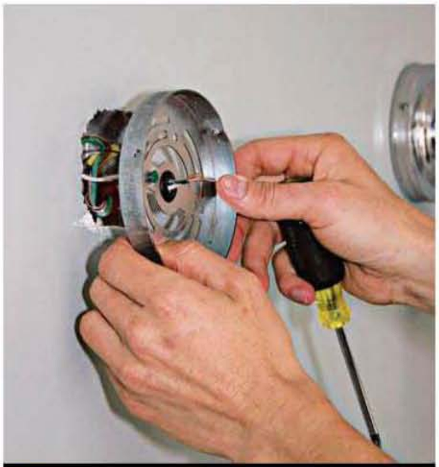

5 Disconnect the wires and remove the old bracket.

When removing an existing sconce, save all the old screws and incidental hardware—you never know what you might need when you attach the new one. Choose a sconce that suits your taste. In this case, we used something a bit more playful—a sconce whose shade is a martini glass and whose halogen bulb shines through colored glass ice cubes. Because the new sconce is a low-voltage unit, it also requires a transformer, which fits under the fixture base.

First remove the glass shade from the existing fixture; most are held on by small setscrews. Always support the shade—especially if it's inverted—to prevent its falling out and breaking ❶. Turn off power to the outlet, then use an inductance tester to make sure the power is off ❷. If the tester doesn't light, it's safe to disassemble the fixture and handle its wires.

Remove the fixture's mounting screws; in this case, they're on the side of the fixture base ❸. Pull the fixture away from the wall to reveal its mounting brackets and the wire connectors that splice the supply and fixture wires ❹. (The green-and-yellow striped wire is the fixture's ground lead.)

After disconnecting the splices to the fixture leads, remove the special mounting bracket because it won't be needed to mount the new fixture ❺. (Its holes won't line up with the new fixture's mounting screws.) Save this bracket in case you want to reinstall the old sconce elsewhere.

TRADE SECRET

It's always best to turn off the circuit breaker feeding the fixture rather than relying on the light switch to disconnect power.

CONNECTING A NEW SCONCE

Once the old sconce has been removed, you're ready to connect the new one. Install the new sconce's mounting bracket ❶. In this case, the new bracket has tabs on each side rather than a continuous flange, like the old one.

If you hold the fixture one handed as shown, it frees your other hand to make connections ❷. Here, the fixture takes a 12v, bi-pin halogen bulb, so it requires a transformer (the black box) to reduce the house's 120v power. (The transformer here is an electronic, solid-state device—essentially, a circuit board—with no moving parts.)

Splice the ground wires, then the neutral wires, and then the hot wires ❸. After you finish splicing the wire groups, push the excess wire and the wire connectors into the outlet box behind the mounting bracket so there will be room for the transformer under the fixture base ❹.

Line up the holes in the fixture base to the holes in the mounting bracket, and screw the base to the bracket ❺. Don't tighten one screw all the way down until you've at least started the screw on the other side.

WHAT CAN GO WRONG

On occasion, a mounting bracket is smaller than the hole or the outlet box will have been set below the level of the drywall. To prevent the bracket's being drawn into the wall, put washers behind the screw holes of the bracket.

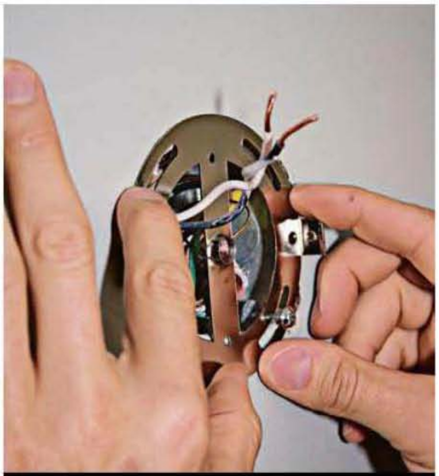

1 Install the new mounting bracket.

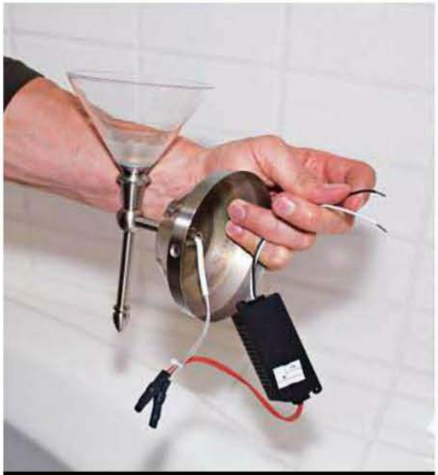

2 Prepare the wires before you place the fixture on the wall.

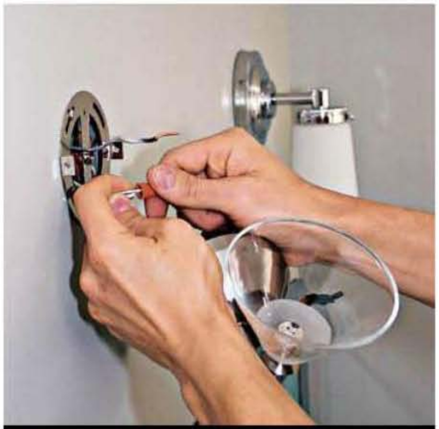

3 Use wire connectors to splice wires.

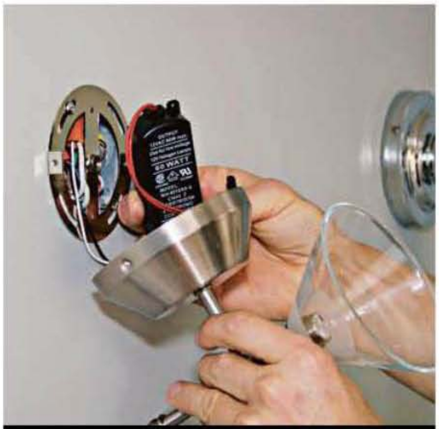

4 The transformer fits under the base.

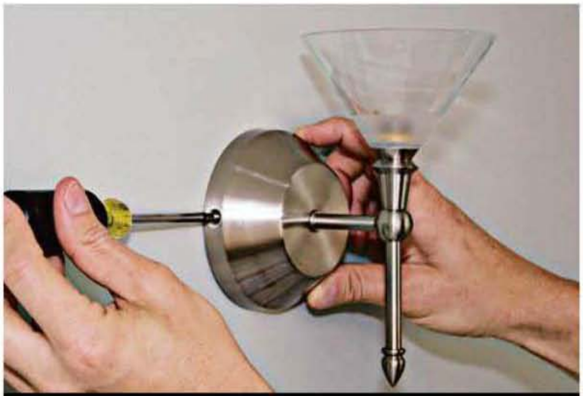

5 Screw the fixture to the bracket.

INSTALLING HALOGEN BULBS & ACCESSORIES

Before inserting any bulb into a fixture socket, check its rating. As noted earlier in this chapter, if the halogen bulb pins don't fit the socket, don't force them—you may have the wrong bulb. Grip the protective plastic wrapping—not the bulb—as you insert the halogen bulb into the socket ❶. As you remove the bulb from its packaging, gently pinch the end of the bulb—a little like squeezing a fast-food packet of ketchup—until its pins stick out through the plastic. Once you've pressed the bulb into the socket, you can easily slide the plastic off ❷.

Lo-vo bulb pins are so tiny that if they become oxidized, carbon can build up in the socket, causing the bulb to flicker or not shine at all. To avoid replacing a socket (which means rewiring the lamp), electricians routinely apply an antioxidant paste to the lo-vo pins before inserting them into a socket.

If your fixture has novelty items such as these glass ice cubes, take care when installing them so you don't damage the halogen bulb ❸. The cubes are tempered glass, so they can withstand heat. Always allow the lamp and cubes to cool before handling them or you'll burn yourself badly ❹.

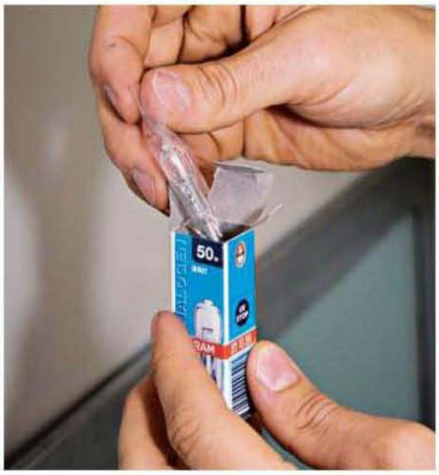

Touching a halogen bulb with bare fingers shortens the bulb's life. Instead, slit the plastic bag the bulb is shipped in and grip the bag as you insert the bulb.

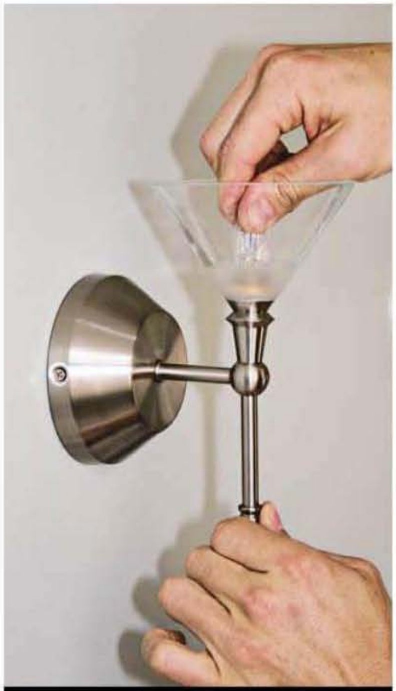

1 Grip the plastic wrapping when inserting the bulb.

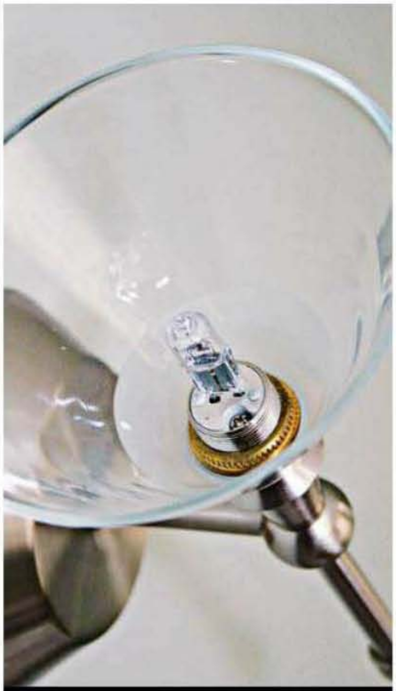

2 A bi-pin bulb must be properly seated in the socket.

3 Carefully place the glass cubes.

4 Party time!

PREPPING & MOUNTING A CEILING BOX

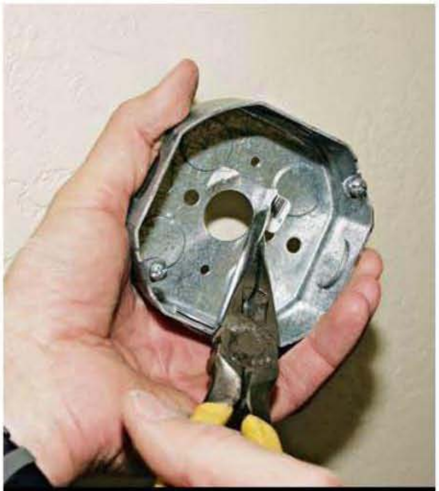

1 Remove the center knockout.

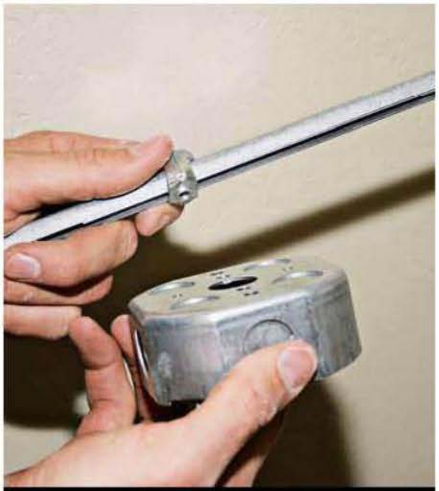

2 Insert the bar fitting into the knockout.

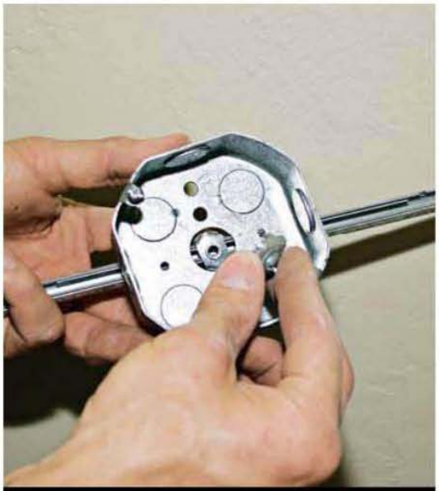

3 Screw the box bracket onto the bar fitting.

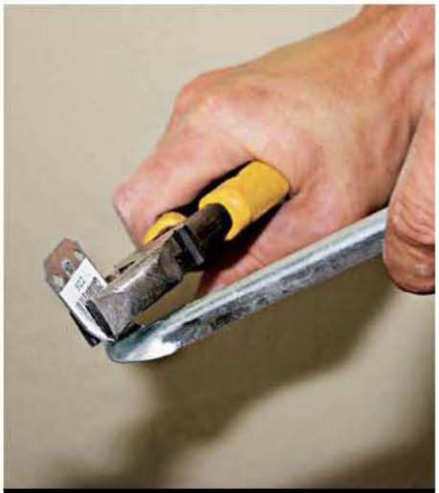

4 Bend bar tabs upward for retrofit applications.

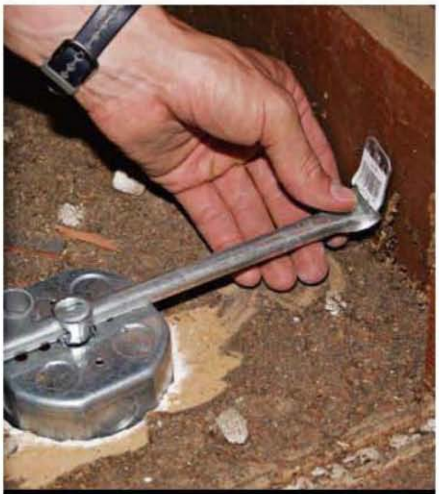

5 Extend the bars and screw them to joists.

Standard ceiling box and bar assemblies are rated for 50 lb. If your light fixture weighs more than that, install a fan box instead.

➔ **For more on mounting fan boxes, see p. 145.**

Start by using a sturdy pair of needlenose pliers or lineman's pliers to knock loose and twist out the center knockout in the box ①. To attach the box to the bar, line up the threaded fitting on the bar to the knockout in the center of the box ②.

Most bar or bar assemblies come with a center screw and washer designed to attach the box to the bar. Fasten the screw and washer to the threaded fitting on the bar ③. Mounted together, the box and bar are rated to support a 50-lb. light fixture.

In new construction, this bar and box assembly would be installed before the ceilings are covered, from below. Thus the bar's tabs face down. In a retrofit, however, you'll be screwing or nailing the tabs to joists from above (if there's access). So when retrofitting a ceiling box, bend up the tabs and they'll be easier to secure ④.

Extend the bar until both ends are snug against the joists. Measure the thickness of the ceiling, adjust the height of the mounting bar so the box will be flush to the ceiling below, and screw the bar tabs to the joists ⑤.

DISCONNECTING A CHANDELIER

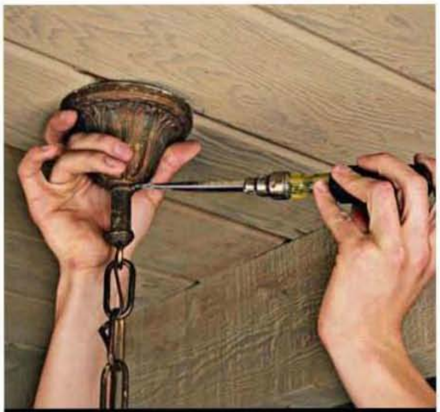

1 Unscrew the canopy and remove it.

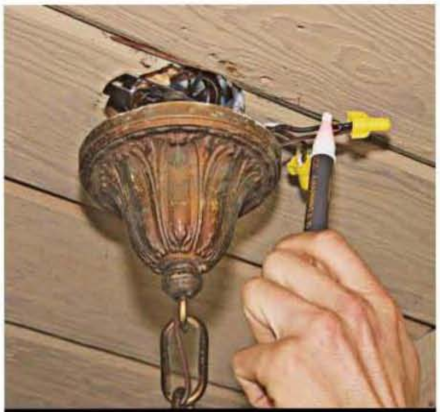

2 Test to make sure the power is off.

3 Remove the wire connectors.

4 Support the chandelier while removing the fixture.

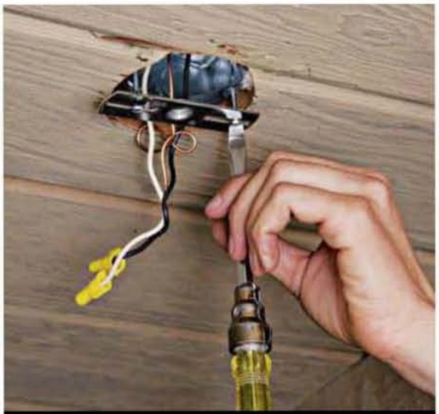

5 If necessary, remove the existing bracket.

TRADE SECRET

If a chandelier canopy is on the small side, make sure the ceiling box is large enough to house the wires (and wire connectors) to be connected. Conversely, if the box is on the small side, a large canopy can hide the connections.

In the sequence shown here, the homeowners swapped a 1920s gothic chandelier for a large, handcrafted Mexican one. Both fixtures were quirky, and the canopy of the Mexican fixture was small for a fixture of its size. Fortunately, the ceiling box was deep enough to house the supply cable and the wires feeding the four chandelier arms. Had there been a shallow "pancake box" in the ceiling, the box would have been too small to house all the wire connections and would have had to be replaced by a deeper one.

Because a helper didn't show up, the electrician had to support the very heavy fixture with one hand while he attached wire connectors with the other. Sometimes you've got to make do with the resources at hand. But if you install a chandelier this big or unwieldy, by all means get help. Before disconnecting a chandelier, turn off the breaker controlling the fixture. Unscrew the setscrew holding the canopy to the fixture post ❶ and slide the canopy down the post to expose the wire connections behind it.

After lowering the canopy, pull out the wire splices—being careful not to touch bare wires—and use an inductance tester to see if there's power ❷. If the tester glows, there's power: Flip the circuit breaker or remove the fuse controlling the circuit and test again. Once you're sure the power is off, twist off the wire connectors and separate the wires ❸. To avoid misplacing them, temporarily twist the wire connectors onto source wires.

Remove the chandelier. In most cases, that means unscrewing two mounting screws; but here the fixture post was threaded, so it was necessary to spin the whole fixture to unscrew the post from a threaded bracket in the box ❹.

See if the replacement chandelier can be mounted to the existing bracket; if not, remove the bracket ❺. Also, if the outlet box is not flush to the ceiling, now is the time to rectify that condition.

INSTALLING A CHANDELIER

Before installing a chandelier, preassemble the mounting bracket for the new fixture. This bracket is typical: A threaded nipple screws into the mounting bracket; a chandelier fitting will screw onto that nipple. Note, too, the green grounding screw, which will secure a pigtail from the ground-wire splice ❶. Then screw the mounting bracket to the ceiling box ❷.

Get help if the fixture's heavy, especially if it screws directly to the box rather than to a nipple. But if you must install it single-handedly, assemble all the tools and parts beforehand so you can focus on lifting the fixture and securing it quickly ❸. Make sure the wires won't obstruct the nipple in the box, and have the cap screw handy. As you raise the fixture, support its canopy ❹ rather than holding individual lamp supports, which may be assembled in sections and come apart.

If the nipple supporting the fixture is long enough, there will be adequate room to reach in and splice wires. Splice the ground wires first, then the neutrals, and then the hot wires ❺. When the splices are complete, tuck the conductors into the ceiling box or behind the canopy, hold the canopy flush to the ceiling, and tighten the mounting screws all the way down. Here, a cap nut covers the end of the threaded nipple ❻.

WARNING

When dealing with any fixture, disconnect the power at the service panel or fuse box before mounting the fixture. If the wires are jammed behind the canopy, there's a chance of nicking a wire and possibly creating a ground fault.

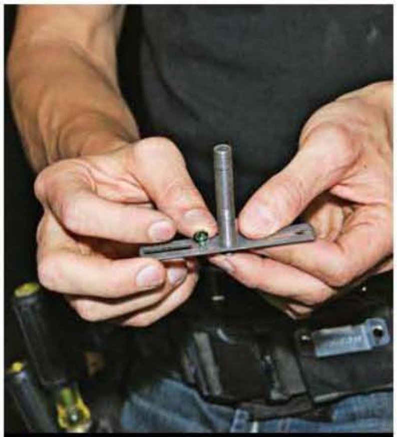

1 The mounting bracket is equipped with a grounding screw.

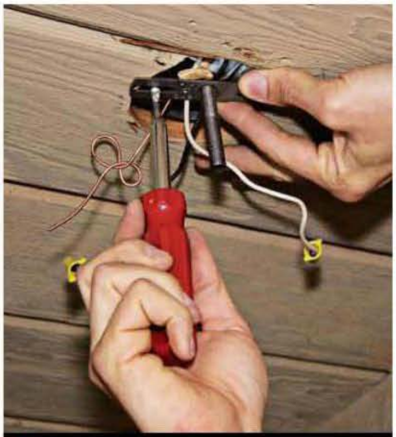

2 Attach the bracket for the new chandelier.

3 Look—and think—before you lift. What must you do first?

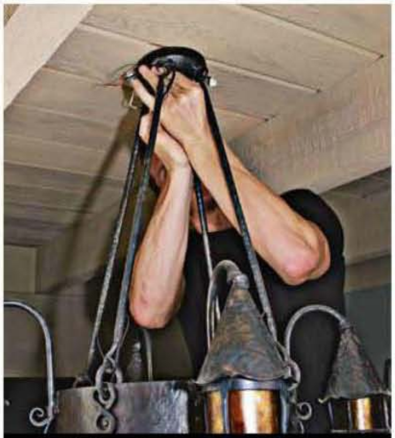

4 Tighten the canopy nut just a few turns, leaving the wire accessible.

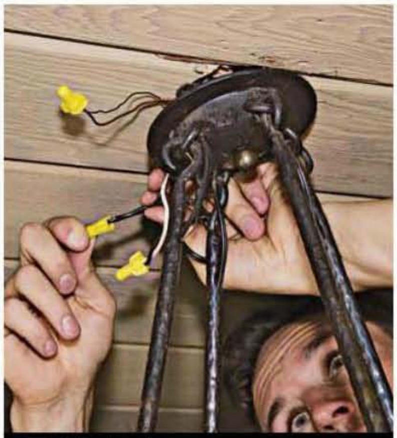

5 With the fixture supported, splice the wires.

6 Tuck in the wires. Tighten the canopy flush to the ceiling.

CUTTING A CEILING HOLE FOR RECESSED LIGHTING

RECESSED LIGHT FIXTURE

Recessed light fixtures vary. The low-voltage model in the photo sequences has a transformer at the end of its assembly to reduce house voltage. The drawing shows a model that runs on house voltage (120v), so it has no transformer. If the unit is watertight, it will have additional trim or lens elements. Closely follow the installation instructions provided with your fixture.

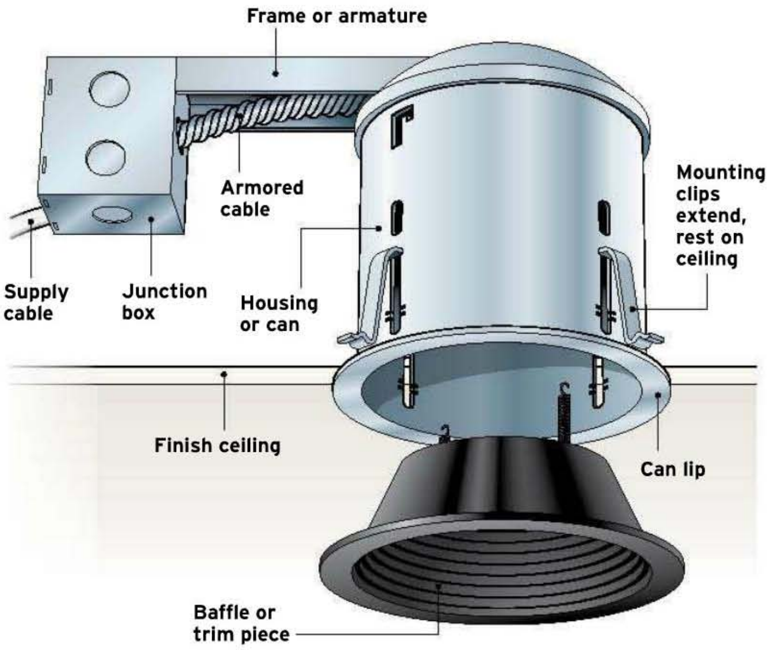

As the name implies, recessed lighting fixtures fit up into space above the ceiling and so don't disturb the flat plane below. Recessed fixtures distribute light evenly and thus are frequently used to illuminate work areas or tight spaces. (The watertight unit shown here is installed in a shower alcove.) Because a retrofit recessed light can (housing) doesn't weigh much, it rests lightly atop ceiling drywall or plaster. Spring-loaded clips or trim pieces draw the can lip tight to the ceiling surface. Of course, all connections are done with the power off.

There's no absolute on where to place a recessed light; but in a small space, such as a shower alcove, a fixture centered in one direction or another will look best ❶. In addition, you may want to use a stud finder to avoid hitting ceiling joists above.

Drill a pilot hole to see what's above ❷, and to make sure there's room for the can. Make the hole small because if there's an obstruction above it, you'll need to patch it. After drilling the hole, you can insert a 4-in. piece of bent wire and rotate it to see if it hits a ceiling joist. Also, drill a small pilot hole to keep the point of a hole-saw blade from drifting.

Keep the drill vertical, and the circle of the sawblade parallel to the ceiling ❸. There are special carbide hole saws for drilling through plaster. A bimetal hole saw will also cut through drywall or plaster, but it'll destroy the saw in the process. Wear goggles.

If the hole saw is the right size for the can, you won't need to enlarge it. But for the light shown here, the saw was a shade too small, so the installer used a jab saw to enlarge the hole slightly ❹. In a pinch, you can also use just a jab saw.

Test-fit the unit ❺. Although you want the can to fit snugly, the unit's junction box and transformer also need to fit through. The black box about to enter the hole is the transformer.

Retrofitting recessed lighting **In retrofit installations, the supply cable to the recessed lighting unit typically comes from an existing ceiling box or nearby switch box. The supply cable feeds to an integral junction box on the fixture. Finding the nearest power source and fishing the wires to the fixture are always an adventure if there's not accessible space above. If the recessed fixture is a low-voltage unit, such as the one shown here, it will come with a transformer, which reduces the 120v current of the supply cable.**

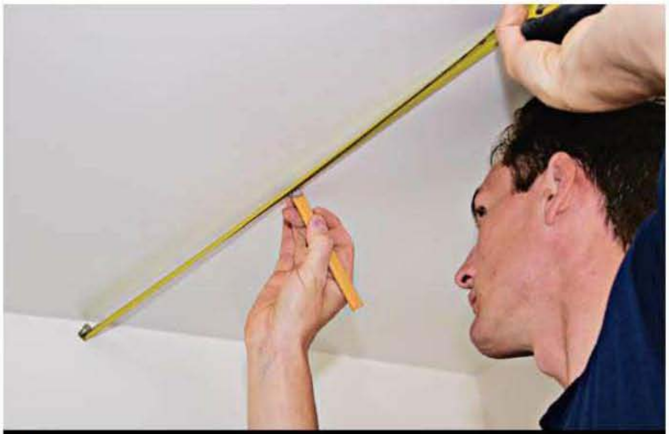

1 Measure to center the light fixture.

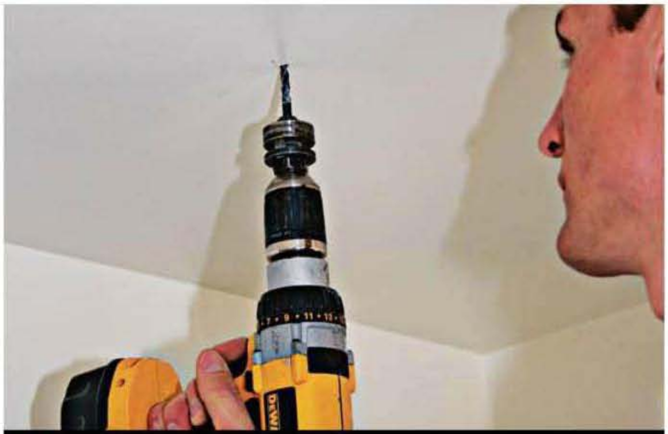

2 Drill a pilot hole.

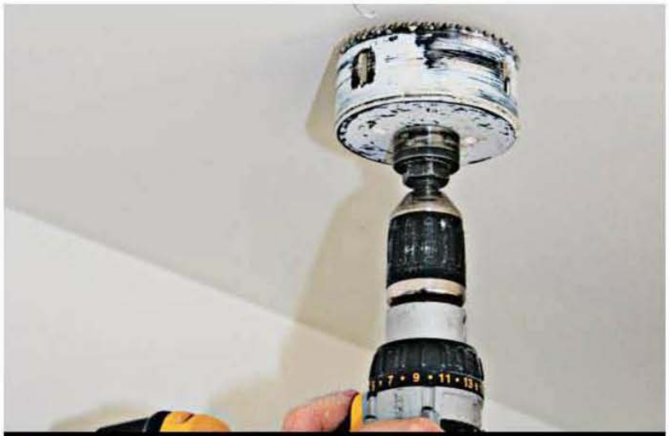

3 Keep the drill aligned vertically.

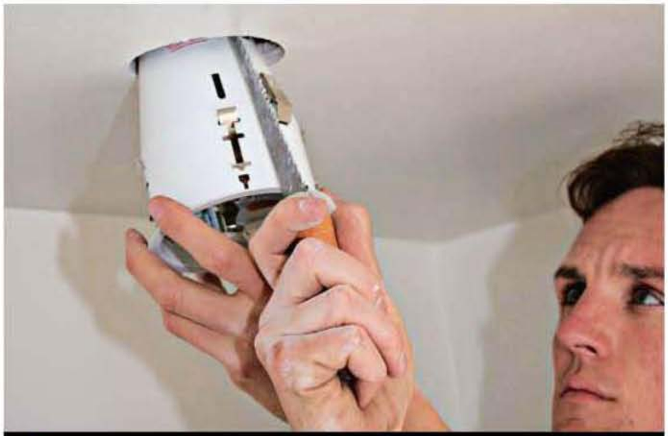

4 Enlarge the hole as needed.

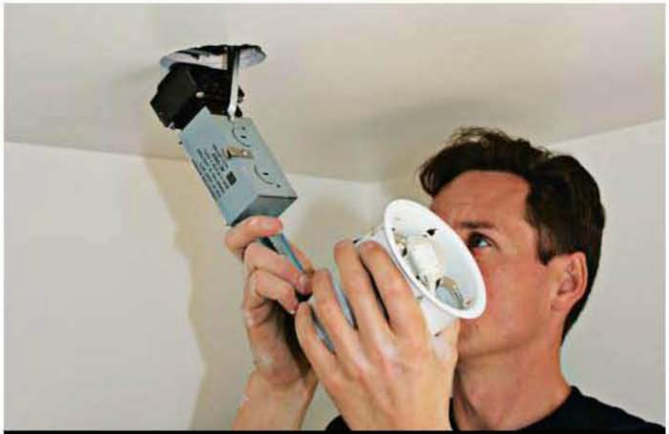

5 Test-fit the fixture.

TRADE SECRET

To minimize hitting pipes or cables when drilling into a ceiling, use a bit that's 1 in. long—just long enough to drill through the plaster or drywall. Use a cordless drill to further reduce the chances of shocks.

WIRING A RECESSED FIXTURE

To wire a recessed fixture, remove a Romex knockout from the unit's integral junction box ❶. Inside the knockout, there is a spring-loaded, strain-relief clamp that will grip the incoming cable, so you don't need to insert a Romex connector. Run a length of (unconnected) Romex cable from the nearest power source (power off, of course) and feed it into the knockout just removed. (To wire the box with AC or MC cable, remove one the circular ½-in. knockouts and insert an appropriate connector.)

➔ For more on connectors, see p. 36.

Inside the fixture's junction box will be two sets of wires that were spliced at the factory. They connect the secondary wires that run from the transformer to the socket. (At the transformer, the current is reduced from 120v to 12v, so polarity is no longer an issue.) There are also three unconnected fixture leads in the box, to which you'll splice the supply wires ❷. Using wire connectors, connect the incoming ground wire to the green fixture lead, the incoming neutral to the white lead, and the hot wire to the black fixture lead ❸.

Tuck the spliced wire groups into the fixture junction box ❹. At the right of the photo is a piece of threaded rod that can be adjusted to support the transformer at the correct height. Snap shut the junction box cover ❺. As with other outlet boxes, Code determines the number of wires you can splice in a fixture junction box, based on the cubic inches in the box.

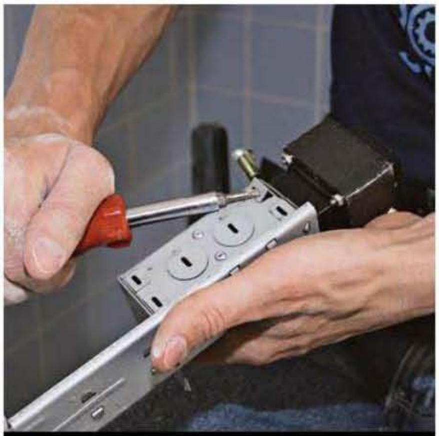

1 Remove a knockout from the junction box.

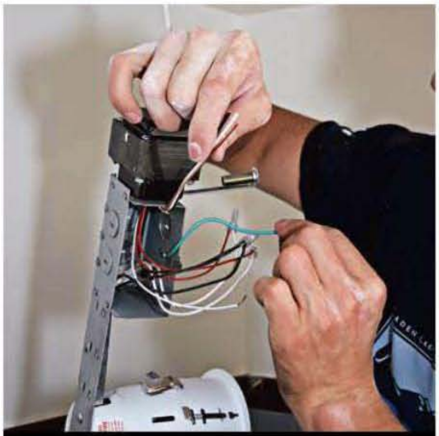

2 Splice the fixture wires to the supply wires.

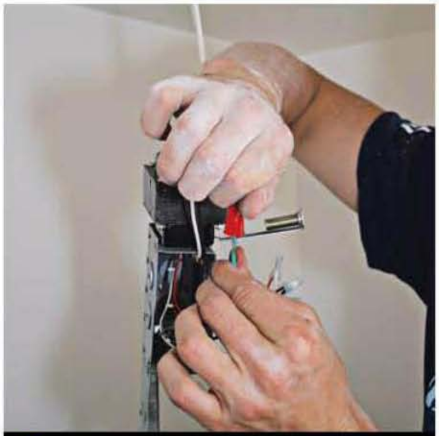

3 Connect the ground wires first.

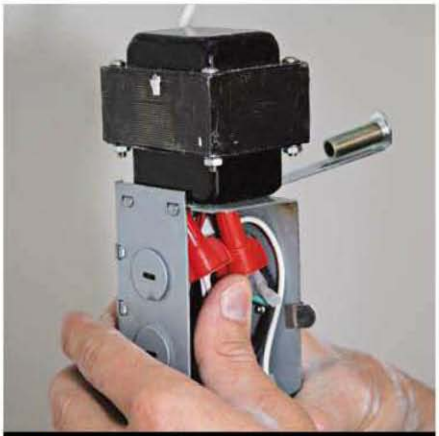

4 Tuck the spliced wires into the box.

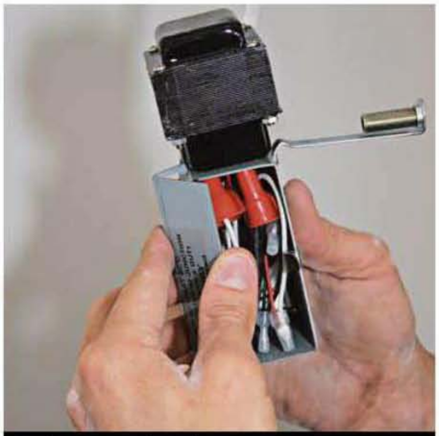

5 Close the junction box cover.

TRADE SECRET

If you have access to the space above, staple the Romex to a joist, leaving at least 12 in. of free cable so you can push the fixture components into the hole. A cable that's too short may prevent this.

SECURING THE CAN

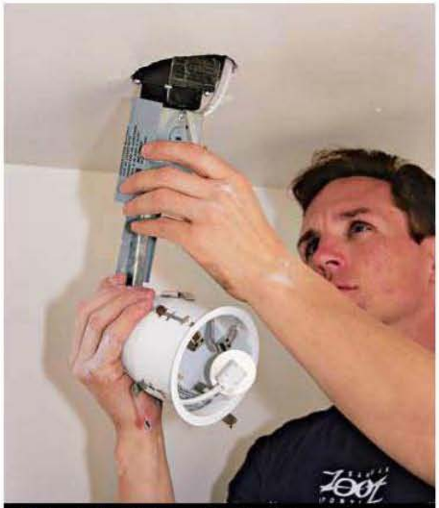

1 Push the fixture into the hole.

2 Use your fist to seat the can.

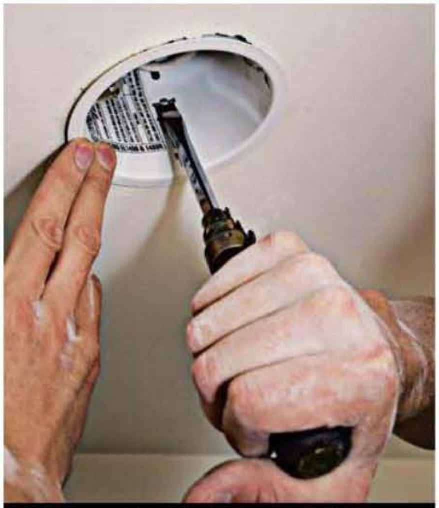

3 Engage the clips to secure the fixture.

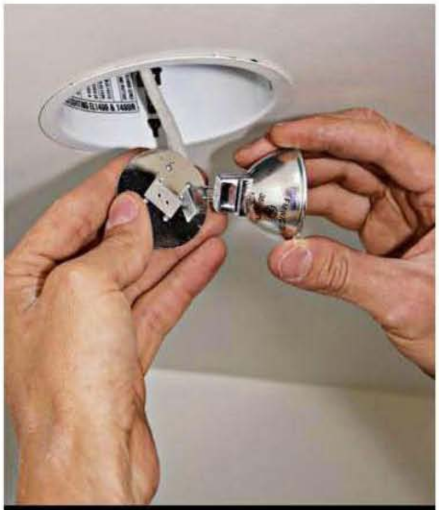

4 Insert the bulb with steady pressure.

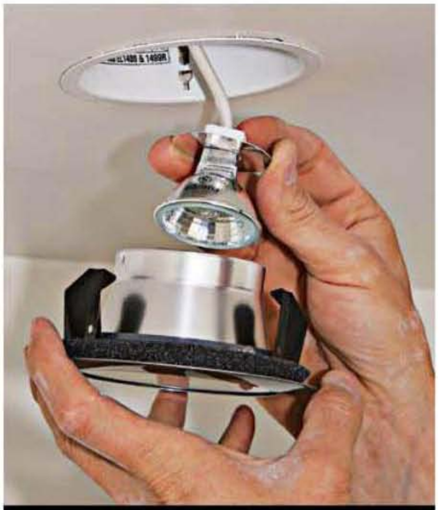

5 Install the trim assembly.

TRADE SECRET

The pros apply a tiny dab of antioxidant paste to the lamp pins before seating them in a fixture socket to prevent oxidation. A good practice, but not imperative.

Once the recessed lighting fixture has been wired, push the fixture into the hole, being careful not to bind the Romex cable as you do so ❶. If the fit is snug, use the side of your fist to seat the lip of the fixture flush to the ceiling ❷.

Use a screwdriver to push up the spring-loaded clips that pivot and press against the back face of the drywall to hold the fixture snugly in place ❸. To remove the fixture later, pop out the clips.

Insert the bulb (the fixture here uses an MR-16 bi-pin halogen bulb) into the fixture socket ❹. Here, the installer is gripping the lamp's reflector, not the bulb itself. The lamp pins should seat securely. Install the trim piece—this one has a watertight gasket. Snap in the lamp and socket and push the assembly up into the can. The three arms on the side of the assembly will grip the inside of the can ❺.

UNDERCABINET FLUORESCENT FIXTURES

Position undercabinet fluorescent fixtures to give you the countertop light coverage you want in each area. If you're running Romex cable to the fixture, you must protect it, either by placing the fixture flush against the back of the cabinet or by covering the Romex with a piece of trim. If you prefer to place the fixture out from the wall, you can also run MC cable, which can be left exposed.

→ **For more on MC cable, see pp. 31 and 204.**

Run the supply cable to the fixture location. Allow at least 12 in. of cable inside the fixture to facilitate wire splices. Remove the sheathing and then feed the supply wires through a knockout in back of the housing shell. Clamp the cable, then screw the housing to the underside of the cabinet ❶. Hold up the housing with one hand and use a cordless screw gun in the other.

Snap on the wiring compartment ❷, which contains numerous prewired connections and the ballast (the white rectangle seen at the right in the photos). Use wire connectors to splice the fixture leads to the supply wires ❸. (Extend the stranded wire ends beyond the ends of the solid wires so that the wire connectors engage the stranded wires first.)

→ **For more on splicing stranded wire, see p. 37 or the Trade Secret on p. 64.**

As you shut the wiring compartment, do your best to keep the wires neat and compact ❹. Messy wires are more likely to get pinched by the cover and short out—which means taking the unit apart and starting all over. Also, avoid locating the wire connectors too near the ballast, which may keep the cover from shutting. Screw in the cover screws that secure the wiring compartment.

As with most fluorescent tubes, insert their end pins into the slots in each keystone, then rotate the bulb clockwise to seat the pins in the sockets ❺. When turning fluorescent bulbs, grab them close to their ends and never force the lamps into place—they should seat easily. Finally, snap the plastic diffuser (lens) of the fixture into place. Be gentle because the diffusers break easily ❻.

UNDERCABINET FLUORESCENT FIXTURE

As with incandescent and halogen fixtures, there are also low-voltage fluorescent fixtures that have a transformer. Fixtures also vary to accommodate different tube shapes.

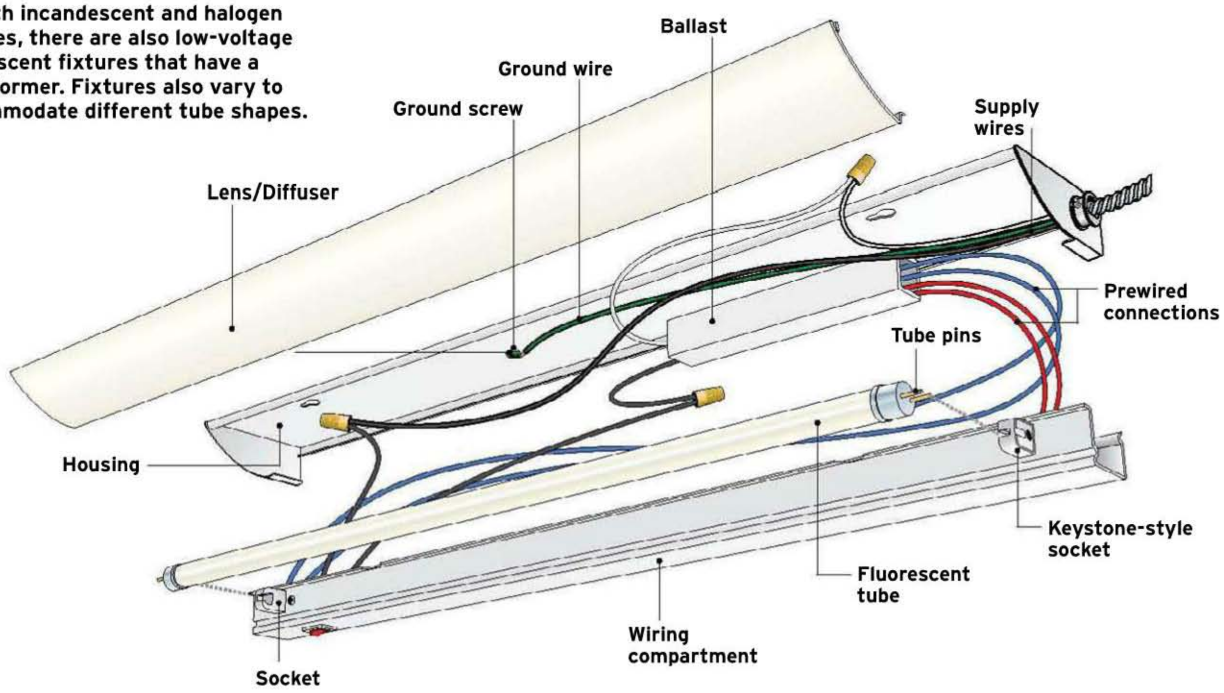

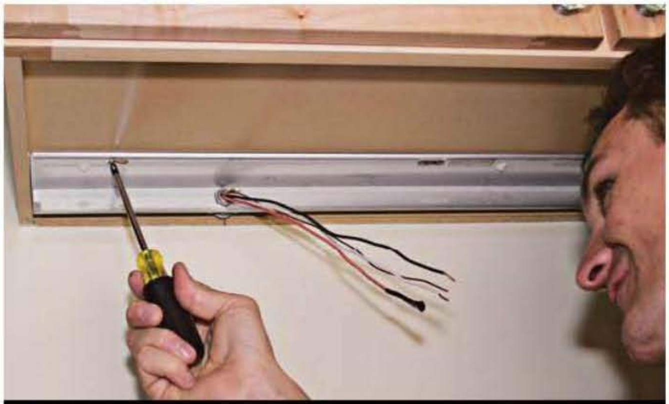

1 Feed in the wires, then screw the housing to the cabinet.

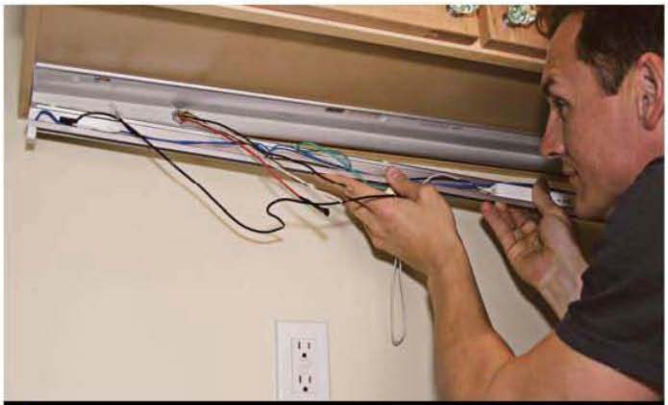

2 Snap on the wiring compartment.

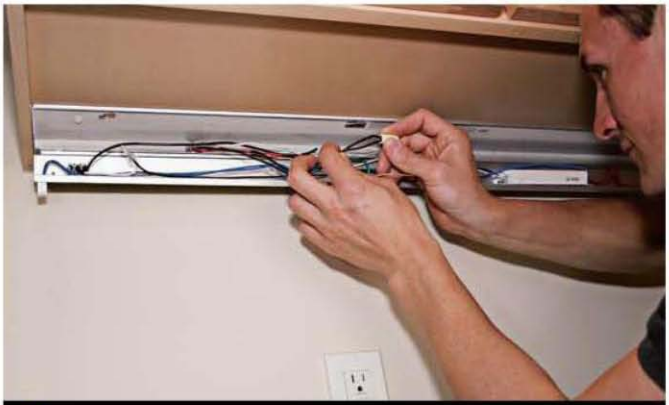

3 Splice the fixture and supply wires.

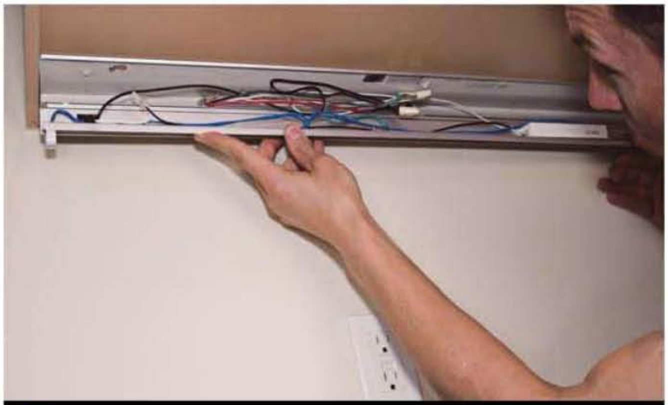

4 Close the wiring compartment.

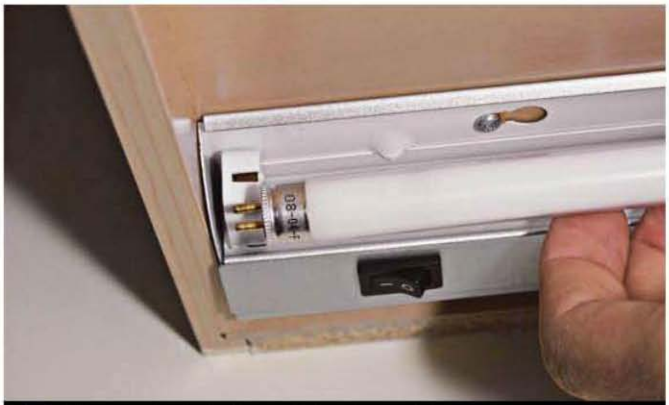

5 Slide the pins into the sockets, then rotate.

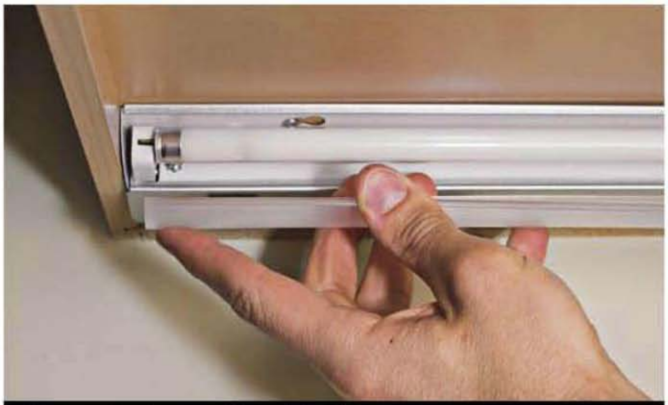

6 Snap the lens or diffuser into place.

LOW-VOLTAGE SYSTEMS

Low-voltage tracks are safe to touch, but you should shut off power to the circuit anyway.

Safety and low-voltage systems

Experienced electricians often handle low-voltage tracks while those parts are energized—with 12v of current—downstream from the transformer. If you install your lo-vo system correctly, and are careful *never* to touch the supply wires that run to the transformer, you can touch energized lo-vo tracks and install light fixtures without getting shocked. But if you have *any* confusion about which wires are 120v and which are 12v, shut the power off before doing any work on your lo-vo system. Best to be too cautious . . . and stay safe.

Low-voltage lighting systems are installed inside and outside houses and typically operate on 12v current, so they require a transformer to reduce standard house voltage from 120v. Transformers vary—some are coil-wound magnets, whereas others are electronic—so follow the installation instructions provided with your unit.

➔ **For more on low-voltage systems, see p. 233.**

Lo-vo systems have become extremely popular because they're safer (12v is roughly the same voltage as a car battery) and energy efficient; plus, their bulbs are easier to direct, enabling you to highlight a painting or a work area. Lo-vo bulbs come in a greater range of options than do incandescent bulbs; and, in general, lo-vo bulbs offer a better quality of light. Lo-vo systems can be controlled by standard switches or dimmers, but you should check the literature that comes with your system.

Though it's safe to touch the tracks of a lo-vo system, you should turn off the power when working on the system. The upstream part of the system (between the transformer and the power source) has 120v power, which could deliver a fatal shock. On the low-voltage side, there is a potential to short the system and damage the transformer.

After installing all the parts and doing a preliminary check of the system, it's safe to energize the system. (*Note:* Because any high-intensity bulb can get very hot, do not install any cable fixtures within 4 in. of a combustible surface.) Track-lighting systems are inherently complex, so read the instructions carefully before you begin the installation.

INSTALLING LOW-VOLTAGE CABLE LIGHTING

Track lighting systems may have a remote transformer or a surface-mounted transformer. There is also a range of cable standoff supports, both rigid and adjustable, mounting to walls and ceilings; they must be mounted solidly to framing if they are to keep position when the cable is tensioned. Cables are typically spaced 4½ in. or 8 in. apart. The illustration has been adapted from installation instructions for a product from Alfa Lighting Systems; your instructions may vary.

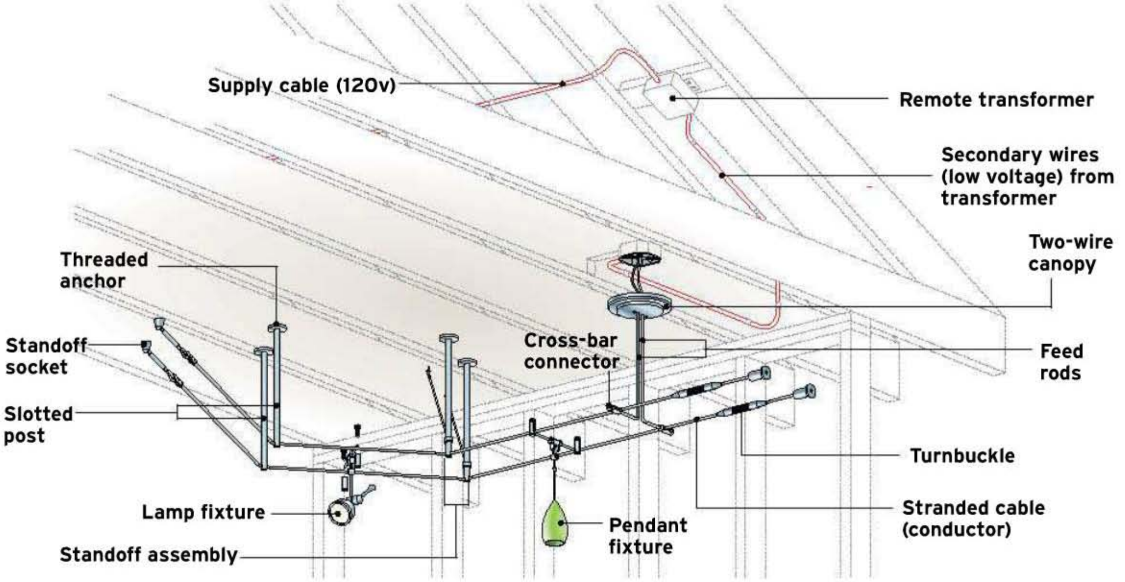
Supply cable (120v)
Remote transformer
Secondary wires (low voltage) from transformer
Two-wire canopy
Threaded anchor
Standoff socket
Slotted post
Cross-bar connector
Feed rods
Turnbuckle
Stranded cable (conductor)
Lamp fixture
Pendant fixture
Standoff assembly

ANCHORING LOW-VOLTAGE STANDOFFS

The key to a good-looking, correctly functioning cable system is getting the cables taut and level. Thus your first task in installing a lo-vo system is finding solid locations in which to anchor the cables. In this installation, the principle anchors were *standoffs* in each corner so the cable could be stretched around the perimeter of the room. Alternatively, you can anchor standoffs in a ceiling to support cables and bulbs. Standoffs are also called *rerouters* because cables often change direction as they emerge.

Use a laser level to establish level anchoring points around the room. Predrill holes in the plaster for anchor screws ❶. Plaster is harder than drywall and there may be lath nails in the way, so wear goggles and have extra drill bits on hand. >> >> >>

1 Predrill all anchor holes.

2 Sink the anchors into the framing.

ANCHORING LOW-VOLTAGE STANDOFFS (CONTINUED)

Because cables will be stretched taut, standoff anchors must be screwed to wood framing—in this case, into doubled studs in the corner ❷. Use screws at least 2 in. long to attach the anchors. After sliding a chrome base plate (washer) over the anchor, screw the standoff socket to the exterior threading of the anchor ❸. Insert the ball end of the fiberglass rod into the standoff socket ❹. This ball-and-socket assembly allows the standoff to swivel freely so you can fine-tune the cable positions. The cables will be spaced 4½ in. apart.

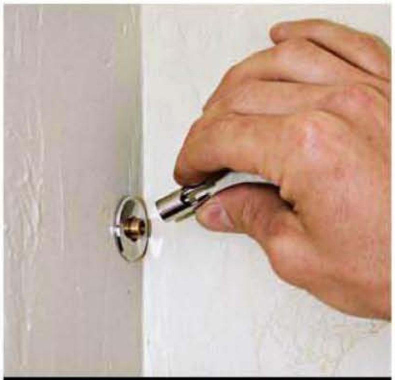

3 Screw on the standoff socket.

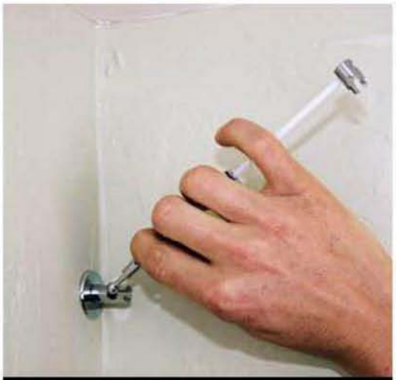

4 Insert the standoff into the socket.

RUNNING CABLE

The next step in installing this system is to run the cable. Note the standoff at the top of photo ❶. Because the walls in this room were only 12 ft. apart, both cables could be run through a single angled standoff with two slotted posts. If the walls are farther apart, you may need a cable support in the middle of the run.

Measure the cable length you need and cut it 3 in. to 4 in. longer than your measurement so you can insert the ends into turnbuckles without having to struggle. Once you tighten the turnbuckles and tension the cable you can snip off any excess cable. On the other hand, if you cut a cable too short, you'll have to discard it and start again with a new piece.

Place each cable into a slotted post, rather than pulling it through the slot, whose sharp edges can cut into stranded cable ❷. (This is a quirk of the particular system shown here; other standoff types allow you to pull cable more freely.)

After placing the cable into a slotted post, screw on the post cap to keep it from popping out when the cable is tensioned ❸. As you place the cable into the subsequent standoffs, loosely tension it to take up the slack ❹. With this system, the second cable will be about 4½ in. from the first.

1 Cut the cable a bit long.

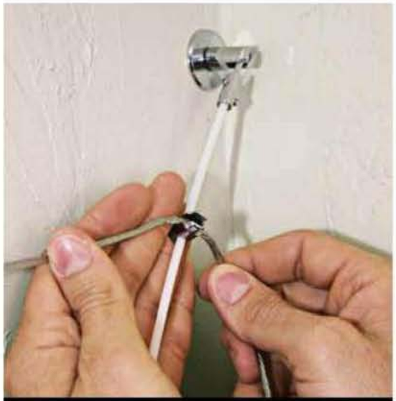

2 Place the cables carefully into the slots.

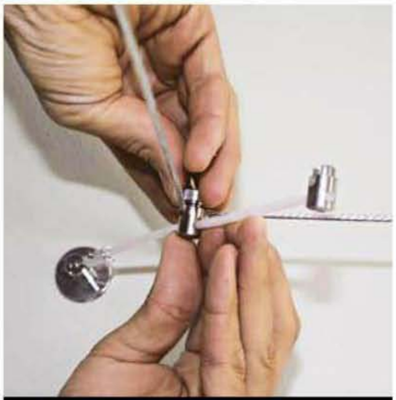

3 Screw on the post cap.

4 Tension the cable to take up the slack.

TENSIONING CABLES

Correctly installed, the cables of a low-voltage system should be more or less horizontal and equally spaced (parallel along their length). As with most systems, this installation uses turnbuckles to tighten the cables after they have been placed in the standoffs. The cables' Kevlar® core prevents stretching or sagging once the lightweight fixtures have been installed.

Once you've loosely run the cable, insert an end into a turnbuckle ❶. The cable end feeds in the end of the turnbuckle and exits in a slot in the middle. Tighten the setscrew(s) on the assembled turnbuckle to keep the cable from pulling out, then trim the excess cable sticking out ❷. Don't rush trimming the cable: Wait until you've made final adjustments to the whole layout before trimming.

A turnbuckle's center post has a thumbscrew with threads on both ends. As you turn the thumbscrew in one direction, it draws tight both ends of a cable; turn in the opposite direction to slacken cable tension ❸. With cable attached to both sides of a turnbuckle, the installer may struggle to draw the cable tight enough to join them ❹. This is a good reason not to trim the cable until the turnbuckle starts tightening both ends.

A certain amount of adjustment is necessary after both cables are taut. Here, the installer adjusts the cables so that the standoff comes out of the corner at a 45-degree angle, thus ensuring that the wires will be equidistant ❺.

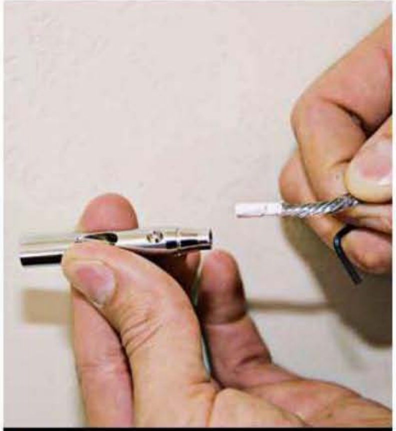

1 Insert the cable into the turnbuckle.

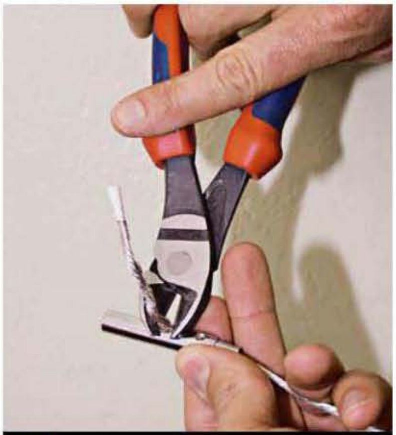

2 Trim and discard excess cable.

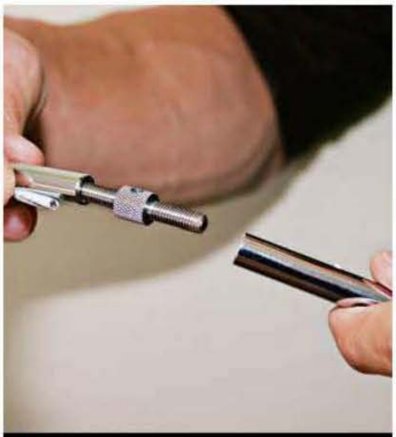

3 One section of the turnbuckle threads into the thumbscrew.

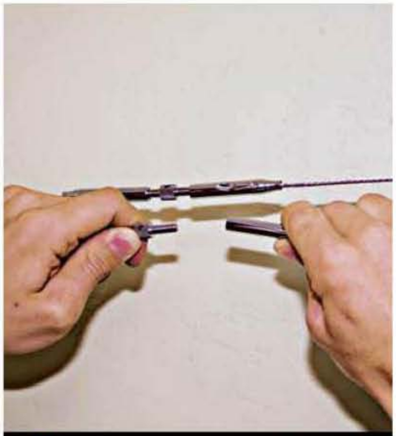

4 Draw the cable ends together.

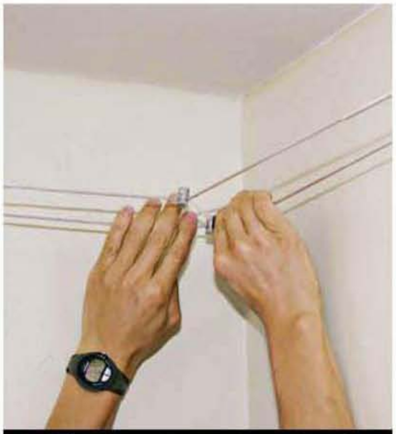

5 Fine-tune cable angles and spacing.

TRADE SECRET

When working with stranded cable, tape the cable ends and then cut in the middle of the tape to keep the strands from unraveling.

LOCATING THE CANOPY

Once the tensioned cable has been installed, you'll need to locate the canopy. The canopy (also called a *canopy feed*) receives low-voltage current from the transformer and delivers it to the cables. The canopy mounts to a junction box on the ceiling or wall and should fit flush to the finish surface.

Before installing the canopy, make its holes—or those of its mounting bracket—line up to the holes of the junction box you'll be installing in the wall or ceiling ❶. The slotted mounting bar on the back of this two-wire canopy can fit several box widths.

If the canopy will be ceiling mounted, drill a hole for it. To minimize the mess, use a hole-cutting tool with a dust cover ❷. *Note:* A screw gun with a ½-in. chuck will accept large-shank tools such as the one shown here. Set the hole-cutting tool's blade to the diameter of the junction box ❸. For large holes, this tool has a counterweight that attaches to the right side of the cutting bar to balance the torque of the blade.

Hold the cover of the tool snug against the ceiling so it can contain the dust ❹. Wait a few seconds for the dust to settle inside the cover before lowering the tool.

TRADE SECRET

Fixture canopies are often polished chrome, which is easy to cloud with fingerprints. Wearing lightweight plastic gloves solves the problem; nitrile plastic gloves are especially flexible.

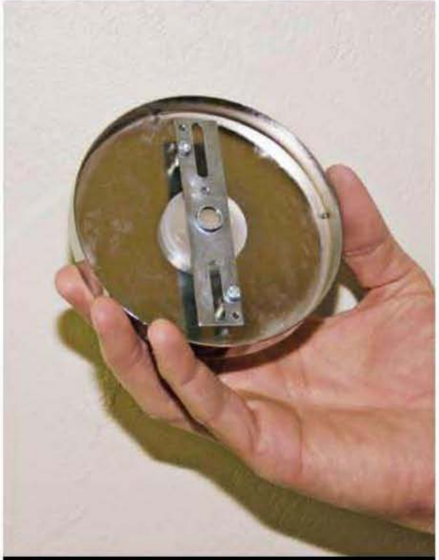

1 Make sure screw holes line up.

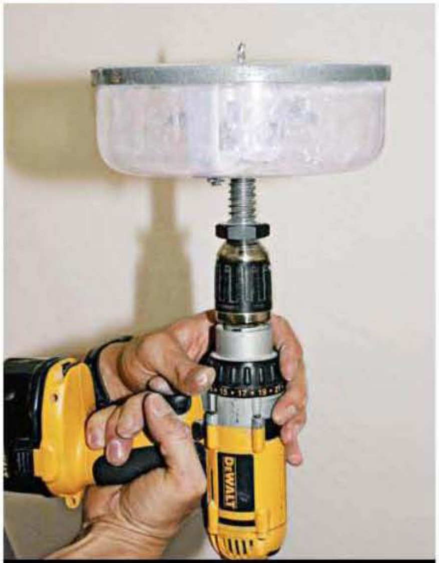

2 Use a dust cover to minimize the mess.

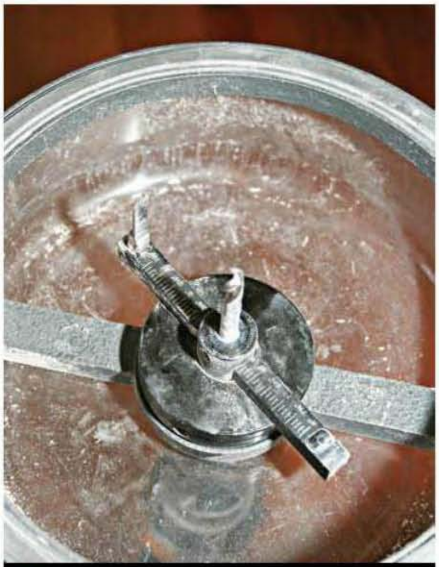

3 Set the cutter to the diameter of the box.

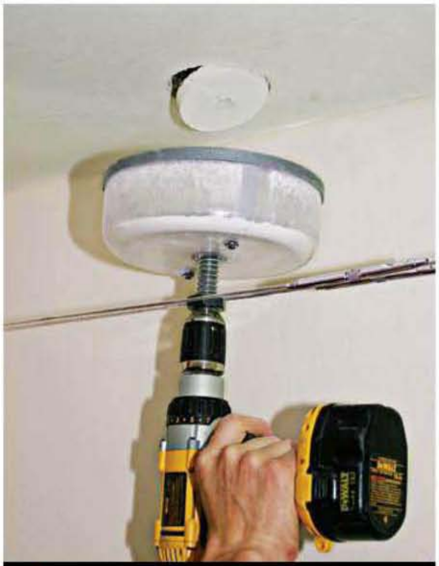

4 Hold the cover snug against ceiling as you drill.

MOUNTING THE BOX & WIRING THE TRANSFORMER

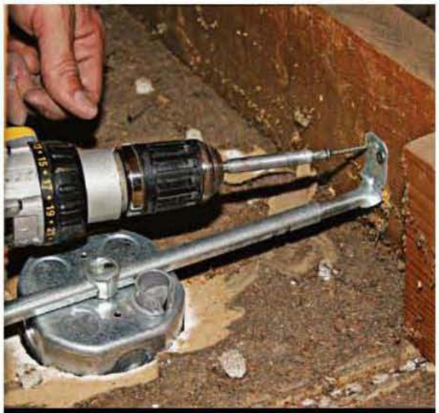

1 Mount the ceiling box.

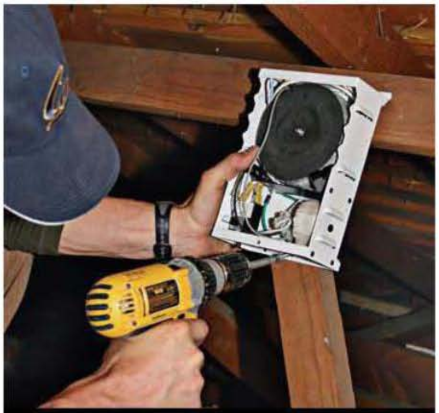

2 Place the transformer above the attic insulation.

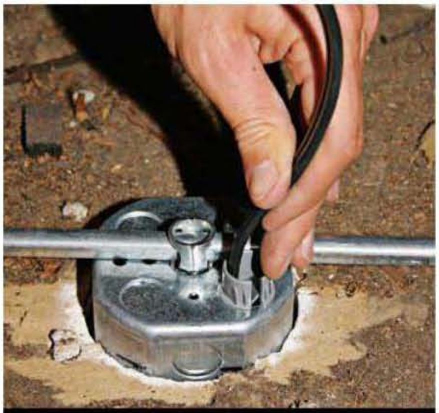

3 Insert the lo-vo cable into the ceiling box.

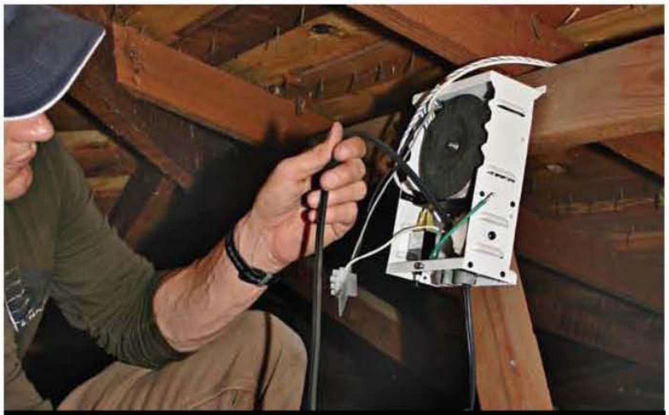

4 Insert the cable into the transformer.

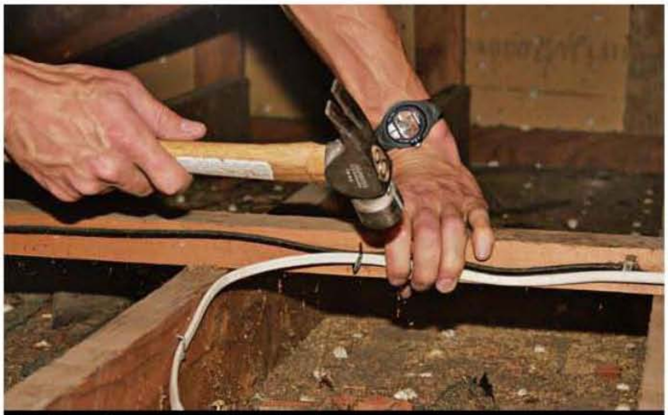

5 Run the supply cable to the transformer.

After the canopy has been located, it's time to mount the box and wire the transformer. The transformer shown here permits 12v or 24v wiring. Your installation may vary, so follow the instructions provided.

When retrofitting a ceiling box, bend up the bar tabs to make them easier to nail or screw. Extend the support bar until its tabs are snug against joists, adjust the height of the bar so the box is flush to the ceiling, and then screw the tabs to joists ❶.

→ **For more on installing ceiling boxes, see p. 85.**

Install the transformer above the insulation so that its vents work properly and the unit can be accessed easily ❷. Run the secondary (lo-vo) cable between the transformer and the ceiling box, stapling it within 12 in. of a box and every 4 ft. along its run ❸. Feed the other end of the lo-vo cable into the transformer ❹. The gray cable connectors are one-way clamps: easy to insert but difficult to pull out.

Run the supply cable (120v) from an existing outlet to the transformer ❺. Protect the cable by stapling it to the side of a joist or to a runner added for the purpose. >> >> >>

MOUNTING THE BOX & WIRING THE TRANSFORMER (CONTINUED)

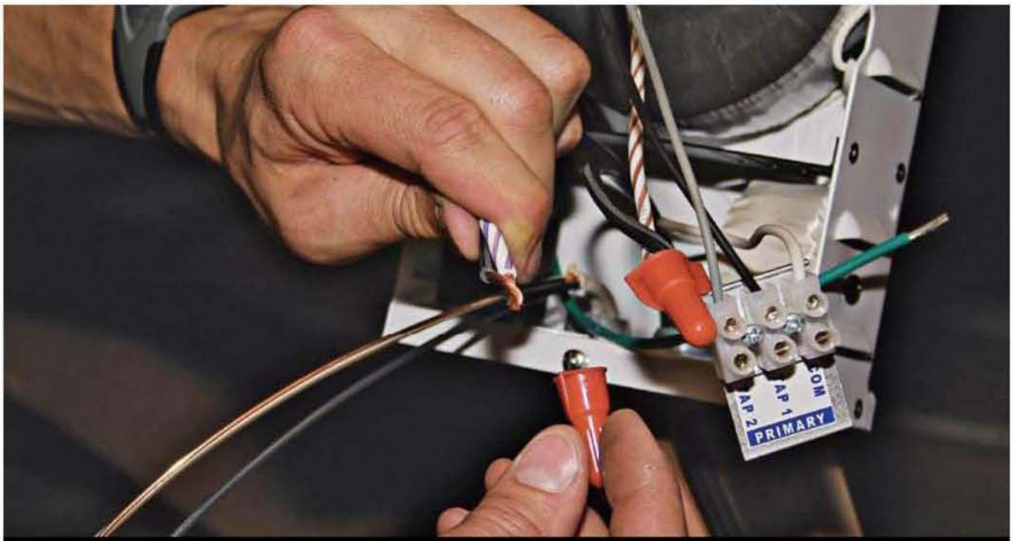

6 Splice the secondary wires.

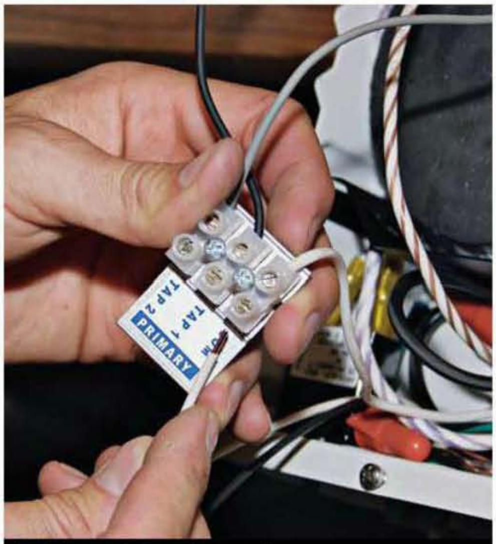

7 Connect the primary wires.

8 Close the cover on the transformer.

After stripping ½ in. of insulation off the wire ends, use wire connectors to splice the secondary wires, which run from the transformer to the ceiling box ❻. Next, splice the primary ground wires, using a wire connector. Then connect the primary neutral to the common (neutral) terminal, as shown, and the primary hot wire to one of the tap terminals. Tighten the terminal screws to grip the wires ❼.

Close the transformer cover to protect the connections inside ❽.

INSTALLING THE CANOPY

The canopy is installed after the transformer has been wired. Install the mounting bracket to the new ceiling box; it will support the canopy that supplies lo-vo power to the cables ❶. The wire hanging from the box is the secondary (lo-vo) cable from the transformer.

Separate and strip the two wires in the lo-vo cable and solder their ends. Soldering fine-strand wire makes it solid and unlikely to smash flat as you tighten down the set screws on the canopy terminals. Soldered wire is also less likely to arc and overheat ❷. (Note the tiny Allen wrench inserted into the setscrew on the right of the photo.)

Use mounting screws that are long enough to extend beyond the canopy face ❸; they're faster to install than short screws because they give you room to maneuver.

Slide the canopy over the mounting screws and turn the canopy cap nuts onto the mounting screws ❹. When the nuts bottom out on the screws, continue turning the cap nuts, which will turn the extra-long screws back up into the box. This will make the canopy nice and snug.

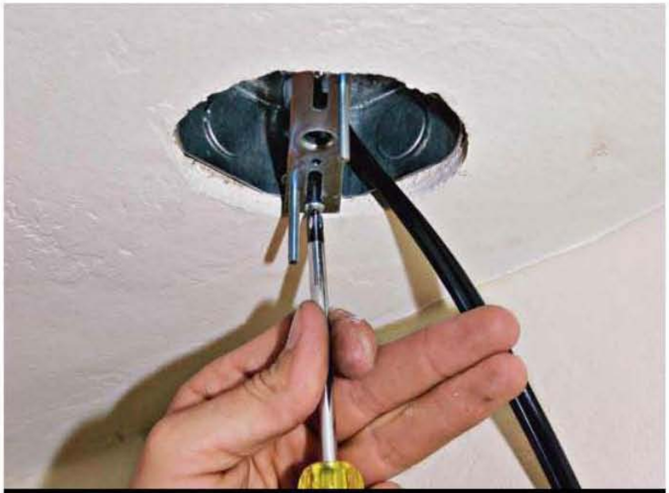

1 Attach the mounting bracket to the ceiling box.

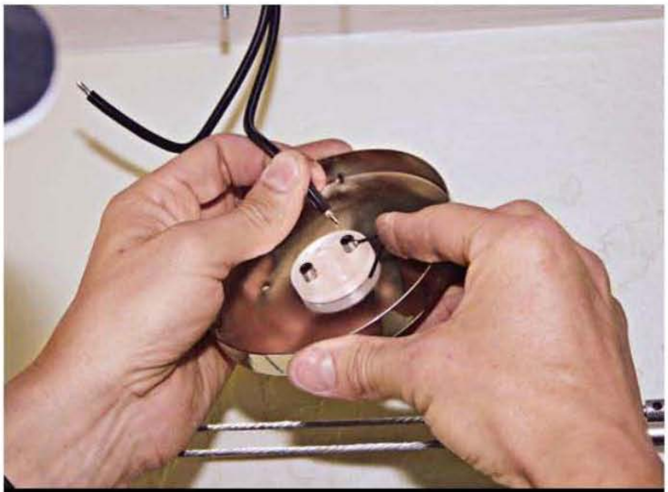

2 Insert lo-vo wires into canopy terminals and tighten.

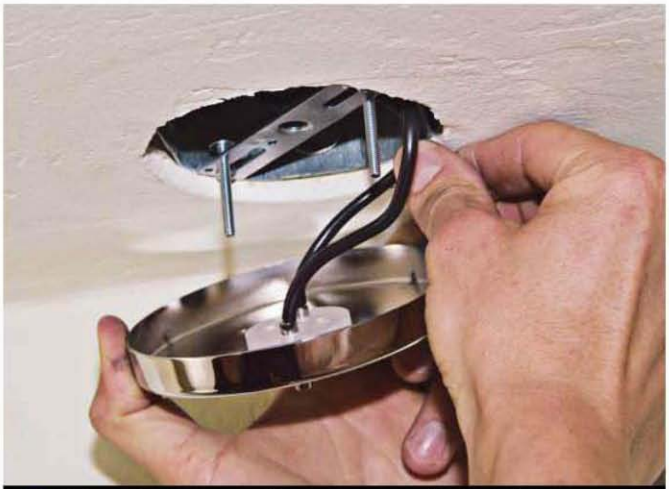

3 Using extra-long mounting screws eases the installation.

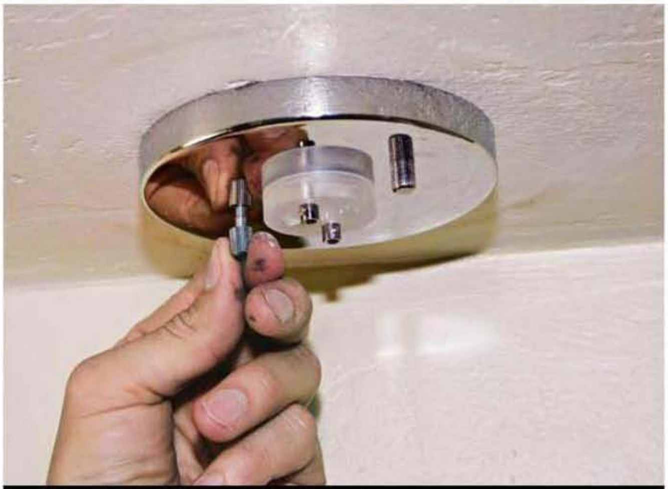

4 Install the cap nuts on the mounting screws.

ATTACHING THE FEED RODS & FIXTURES

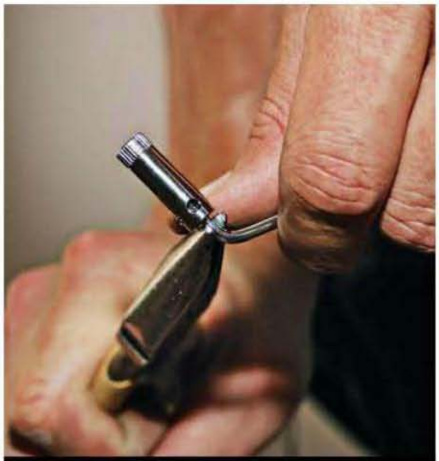

1 Bend the feed rods into position.

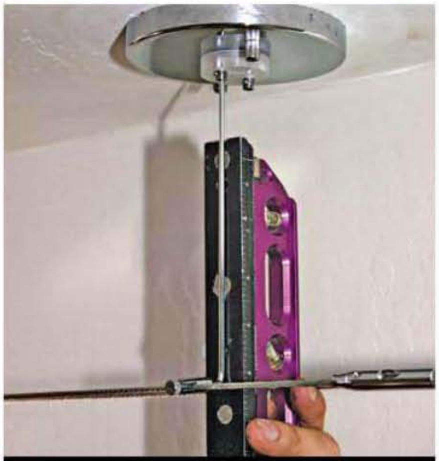

2 Test-fit and plumb each rod.

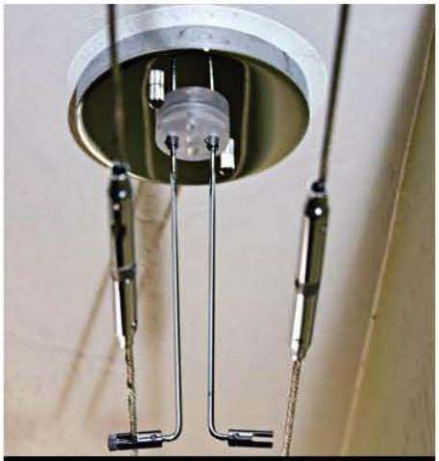

3 One rod has already been attached; the other has not.

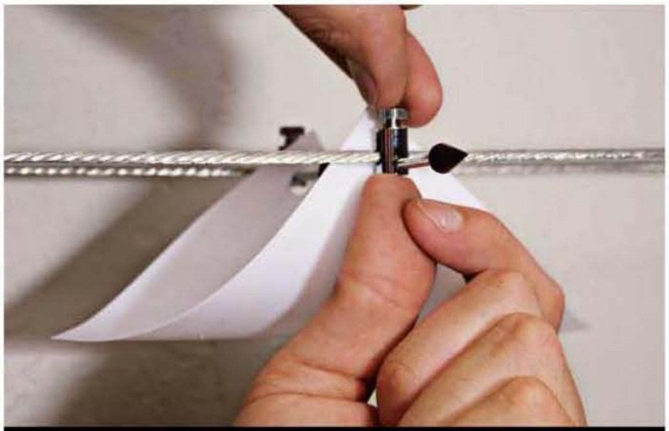

4 Attach the fixture crossbars to cables.

5 Once everything is installed, turn on the power.

The final task for installing a low-voltage system is adding the feed rods and fixtures. The feed rods transfer low-voltage current from the canopy terminals to the cables. Setscrews on the terminals secure the rods.

Because the canopy is centered over the two cables, you may need to bend the feed rods to bring them to the cables ❶. Secure the rods in a vise between two wood scraps (to minimize marring the finish) and bend them at the point equal to the distance from the canopy to the cables. Some systems will provide horizontal connectors for this purpose, eliminating the need for bending.

Bend the first rod and test-fit it, using a torpedo level to ensure that the rod is plumb and the cable is still level ❷. If the first rod fits well, use it as a template for the second. Repeat the process with the second feed rod. Note that rod ends are slotted like the standoff posts that anchor the cables in the corners of the room. A cap nut screws on to the slotted rod end to capture the cable ❸.

Insert the bulbs into the fixtures before you attach fixtures to the cables. If your system uses halogen bulbs, don't touch them with your bare hands because the oil in your skin will shorten the bulb life. In general, fixtures need to connect both cables to become energized, so each fixture has a crossbar that spans the cables. Hand-tighten the fixture connectors so they're snug ❹. After you'll installed all the fixtures and surveyed the system, turn on the power ❺.

INSTALLING LOW-VOLTAGE MONORAIL TRACK LIGHTING

In the installation shown, the transformer is surface-mounted in a circular housing. Although the transformer housing must be mounted to a ceiling box to be adequately supported, individual standoff supports can be anchored in drywall or plaster alone because the track and fixtures are lightweight. The illustrations have been adapted from installation instructions from a product by Tech Lighting^{SM}; your instructions may be different.

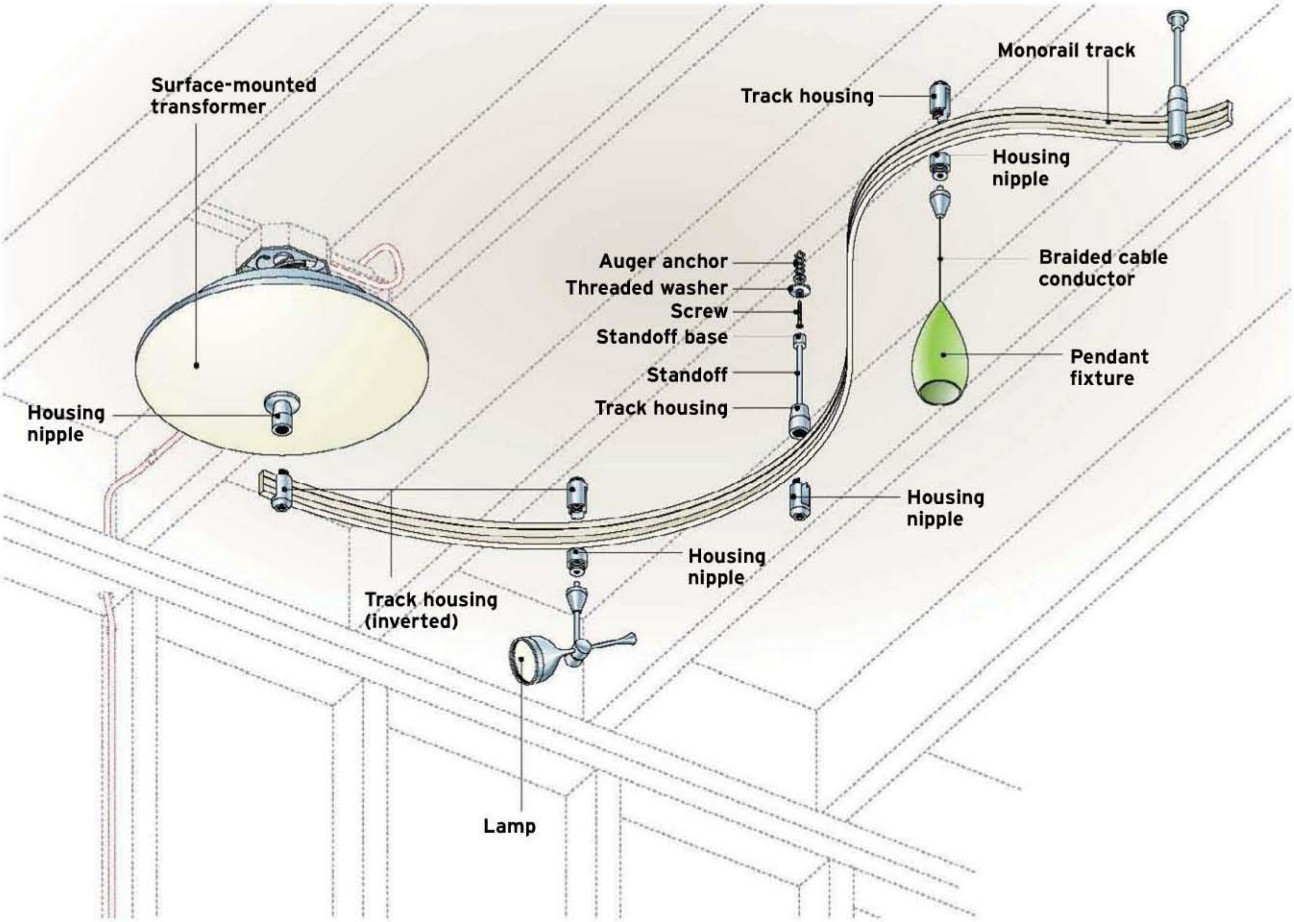
Surface-mounted transformer
Track housing
Monorail track
Housing nipple
Auger anchor
Threaded washer
Screw
Standoff base
Standoff
Track housing
Braided cable conductor
Pendant fixture
Housing nipple
Housing nipple
Housing nipple
Track housing (inverted)
Lamp

Before installing lo-vo monorail track lighting, turn off the power at the breaker panel or fuse box. As with any low-voltage system, a transformer will reduce the 120v current to 12v or 24v. Because monorail track systems have a lot of small parts that are easily lost, don't open the packages until you're ready to use the parts.

In cross-section, the monorail track is two square pieces of chrome-plated copper conductor sandwiched together with a clear (insulating) plastic piece. Although the track is sturdy, it can be bent freehand or shaped against a curved piece of plywood.

LOCATING THE TRACK & STANDOFFS

If possible, before working overhead, shape the track sections of your monorail system and preassemble them on the floor, then use a plumb bob or laser level to plumb up to the ceiling. If there's an obstruction in the room you'll have to improvise. First determine where you'll place the transformer box for the system, then mark off the standoffs that will mount the track to the ceiling.

With a helper holding one end, hold sections of the track against the ceiling and mark off standoff points at the track ends, where sections meet, and at intervals suggested by the fixture maker—usually, every 3 ft. ❶. Standoffs have multiple parts, which anchor them to the ceiling and support the track (See "Standoff Parts" below).

Fortunately, because the track weighs little, you can mount standoffs almost anywhere on a drywall or plaster ceiling, using *auger anchors* with wide threads. (That is, you don't need to mount standoffs to framing.) Sink the anchors flush, then screw a threaded washer to each anchor ❷. Then, using a setscrew, screw the standoff base to that threaded washer ❸.

Mount all the standoffs so that when you raise the monorail again, you can attach it quickly to the standoffs and to the bottom of the transformer housing. If you need to reposition a standoff or two, it's easy to patch the small holes left by misplaced standoff anchors.

1 Mark out the standoff locations on the ceiling.

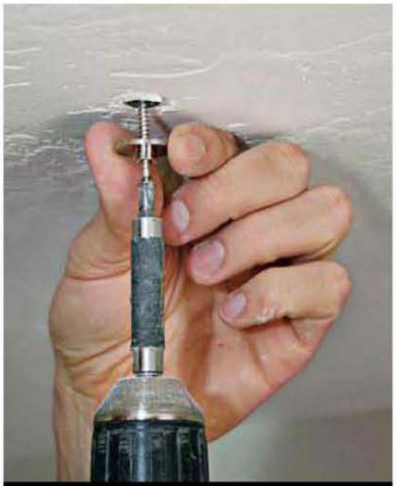

2 Screw threaded washers onto the anchors.

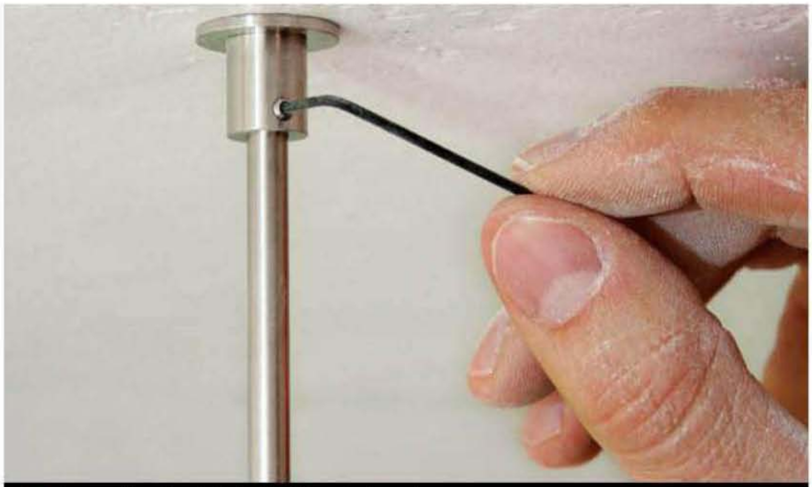

3 Screw the standoff to the threaded washers.

STANDOFF PARTS

Standoff supports for monorail track systems consist of many small parts.

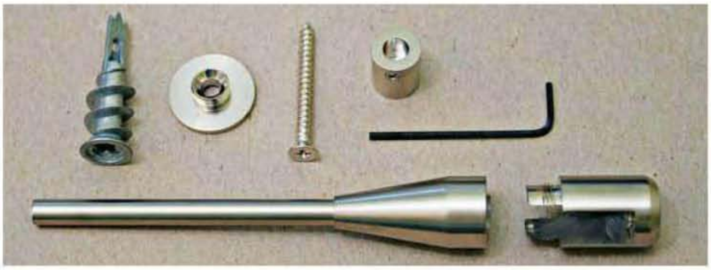

Clockwise, from top left: auger anchor, threaded washer, screw, standoff post, Allen wrench, housing, housing nipple.

MOUNTING & WIRING THE TRANSFORMER

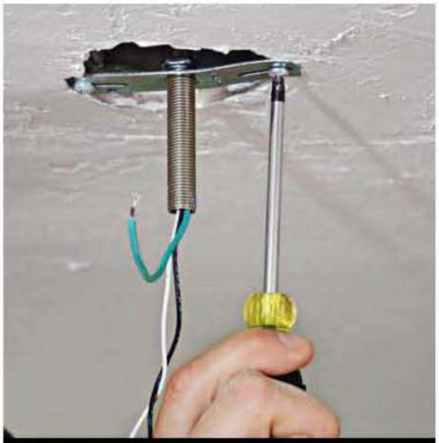

1 Attach the mounting bracket.

2 Feed the supply wires into the housing.

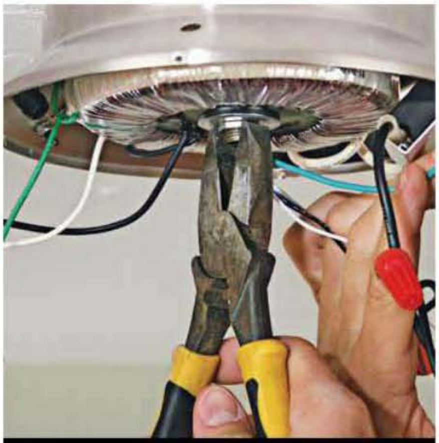

3 Secure the transformer housing.

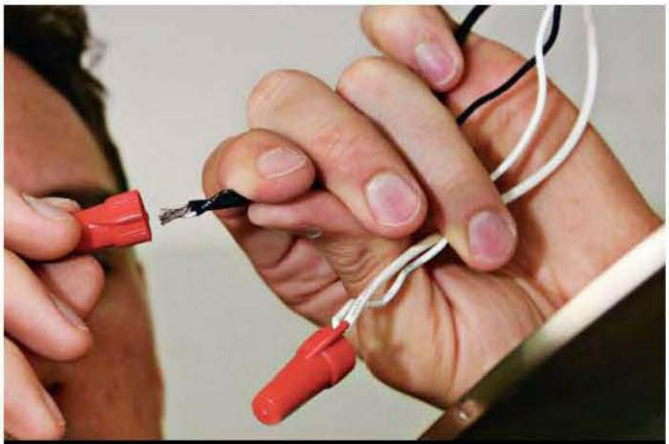

4 Splice the supply wires to the primary leads.

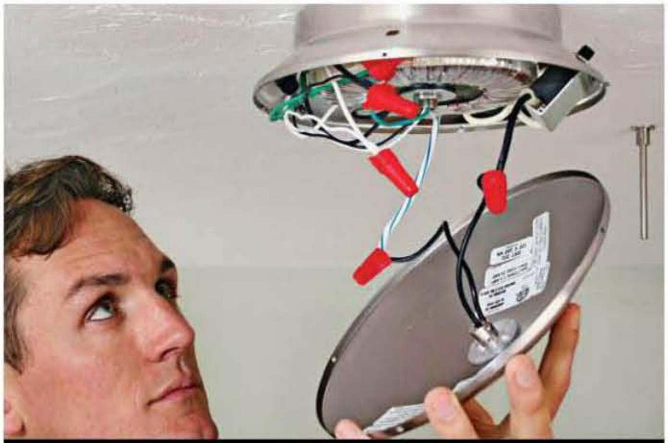

5 Test the connections and then close the housing.

Once the standoffs are installed, mount the ceiling box to a ceiling joist and attach the bracket that will support the transformer ❶. In this case, a nipple screws into the bracket and runs through the center of the transformer housing. The supply wires sticking out of the box will connect to the primary leads on the transformer.

→ **For more on mounting boxes, see p. 85.**

Feed the supply wires into the center of the transformer housing ❷, push the top piece of the housing snug against the ceiling box, slip a washer over the end of the nipple, and then tighten the inside nut that secures the transformer housing ❸. The circular mass inside the housing is a magnetic transformer, which has a series of copper coils.

Many transformers come with secondary leads preattached, so that the installer need only splice supply wires to the primary fixture leads. Using the wire connectors provided, splice the ground wires first, then the neutral leads, and finally hot wires ❹. When all the wire connections are snug—gently tug on spliced wires to be sure—close the transformer housing by snapping the bottom to the top and tightening the setscrews provided ❺. (By the way, the fat striped wires are secondary leads, which run from the transformer to the lo-vo power feed that energizes the track.)

HANGING THE TRACK

With the transformer secured and the standoffs installed in the ceiling, you're ready to hang the track. Get help supporting the track until you have two or three points secured. Place the track into the standoff housing—the lowest piece on that assembly—then screw on the housing nipple ❶. Once that's done, tighten the setscrew that holds the nipple to the support rail.

Where track sections meet, join them with conductive connectors and support the junction with a standoff ❷. Once the track is supported at several points, loosely attach the housing and nipple assemblies at several points, slide them beneath the support rails, and use an Allen wrench to tighten the setscrews that join the nipples to the rails ❸. Make sure the connections are tight.

Because the housing on the bottom of the transformer has setscrews on the side and the bottom, you'll need two different sizes of Allen wrenches. Loosen the small screw on the side of the housing nipple so you can turn it onto the threads of the housing. Once the nipple is tight on the housing, retighten that small screw. Then use a 3⁄16-in. Allen wrench to tighten the setscrew on the bottom of the housing ❹. Finally, retighten all connections and place end caps on the ends of the track ❺.

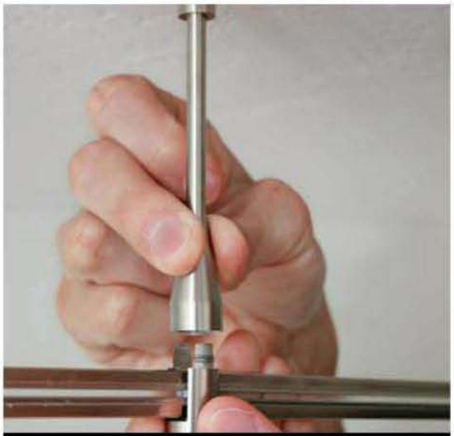
1 Place the track in the housing.

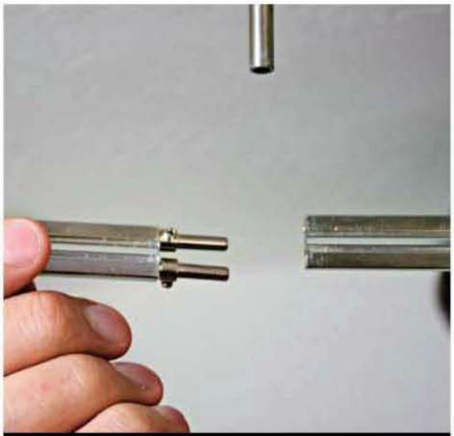
2 Use conductive connectors to join the sections.

3 Preattach the housing assemblies on the track.

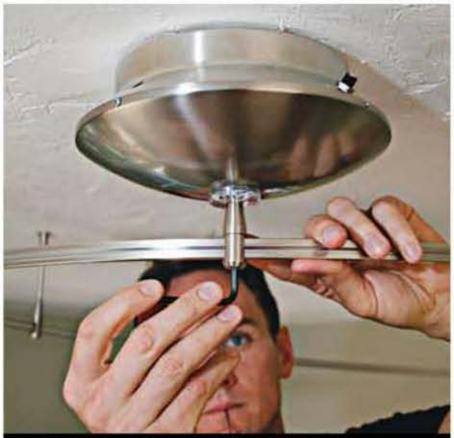
4 The transformer support requires two wrenches.

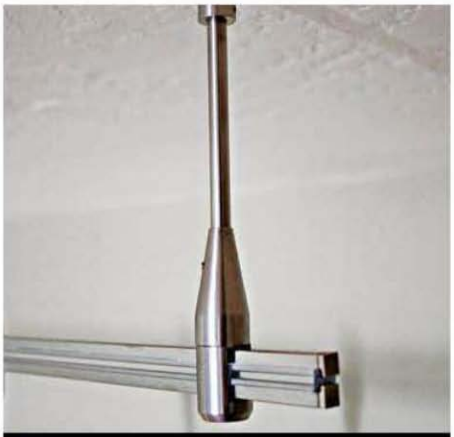
5 Place end caps at the end of each run.

TRADE SECRET

If you can't find anyone to help you support the track, bend short lengths of coat hanger into Z-shaped hanger brackets. Drill 1⁄4-in. holes into the ceiling and insert one end of the bracket; the other end supports the track. Once the track is up, remove the hangers and patch the holes with joint compound.

INSTALLING THE LIGHT FIXTURES

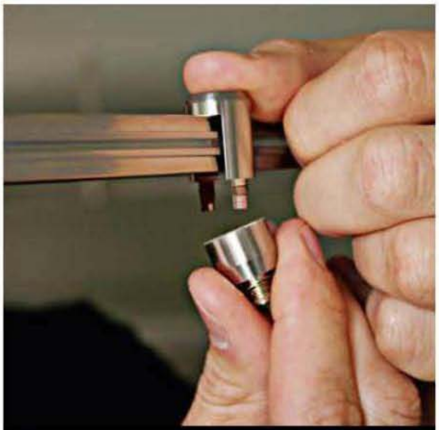

1 For monotrack lights, the housing goes on top.

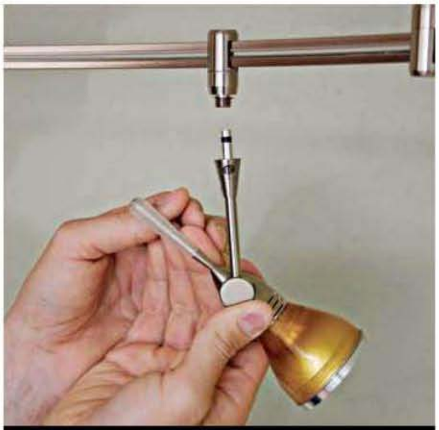

2 Insert the shaft of the light fixture.

3 Support pendant lights with braided cable.

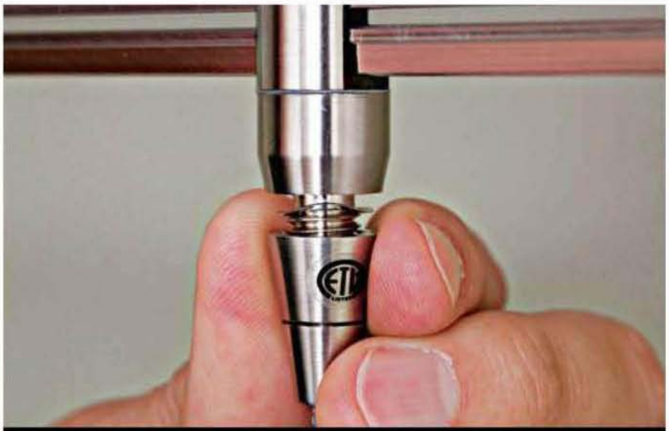

4 Tighten the light supports on the pendants.

5 Energize the system for 20 minutes, then make sure the connections aren't overheated.

Once the monorail tracks are installed, you can add the light fixtures. The fixtures require a two-part assembly that straddles the monorail track; note that, in this case, the housing and nipples are inverted. Insert the housing on top of the track and the nipple on the bottom ❶. Screw the pieces together, insert the shaft of the light fixture into the inverted nipple ❷, then screw the fixture nipple onto the threaded housing nipple and hand tighten it.

Attaching the pendant lights is similar, although the light pendants use braided cable rather than a solid shaft. Braided cable can be shortened if necessary, allowing you to install the pendant lights at the same or varying heights ❸. Once you've adjusted the pendant cables, tighten the fixture nipple to the housing nipple on the track ❹.

Most manufacturers recommend energizing the system and then turning the lights on for 20 minutes or 30 minutes before checking the monorail connections to see if any are hot to the touch. Warm is normal, but hot connections should be retightened ❺. Do not, however, touch any halogen lights; they are certain to be hot enough to burn you (which is normal).

TRADE SECRET

Retighten all connections after the first 10-20 hours of use.

REPAIRING LOW-VOLTAGE CONTROLS

REPAIRING LOW-VOLTAGE controls will be familiar territory if you've read the earlier chapters on switches and installing light fixtures. Basically, doorbell buttons and thermostats are switches. To repair or replace doorbell units, thermostats, and transformers, you'll need only basic tools such as those needed to install switches or lighting. And like low-voltage lighting, doorbell units and thermostats are energized by a transformer, which reduces house voltage. You can safely test or handle existing low-voltage wires without turning off the power. But because transformers are energized with 120v house current, you must always turn off the power before testing or repairing a transformer or wires upstream—those that run from the transformer to a power source. Thus an inductance tester and a multimeter are essential tools to keep you safe.

DOORBELLS

Troubleshooting a doorbell, p. 110
Replacing a doorbell switch, p. 113
Installing a chime unit, p. 114
Installing a doorbell transformer, p. 115

THERMOSTATS

Testing an old thermostat, p. 116
Installing a programmable thermostat, p. 117

FURNACES

Replacing a furnace transformer, p. 118

LONGER CYCLE
CAT III 600V~
100A~
FLUKE

TROUBLESHOOTING A DOORBELL

Troubleshooting a doorbell system takes a little detective work. On older systems, the problem is usually the chime or bell unit—we'll use the term *chimes* to denote either. In many cases, the plunger springs on the chimes become compressed, resulting in chimes that no longer ring predictably—or at all. Corrosion or dust buildup can also silence chimes. Try the tests given here to determine whether the problem is the switch, the chimes, the transformer, or the wiring between the transformer and the switch.

Start by testing the doorbell switch. Unscrew it and gently pull it out from the wall to expose its connections on the back side ❶. Disconnect one of the wires and, using a continuity tester, touch the tester clip to one switch terminal and the tester point to the other. Press the button: If the tester lights as you depress the button, the switch works ❷. If not, the switch is faulty and should be replaced.

A more thorough testing

Alternatively, you can detach both wires from the switch and touch their bare ends together to perform several tests. If the chimes sound when you join the wires, the switch is defective ❸. If the switch wires spark when you touch them but the chimes don't ring, test the chimes as shown in "Testing Chimes," on p. 112.

If there's no spark when you touch the switch wires, use a voltmeter to test the transformer output ❹.

If the transformer is working, you'll need to replace the wiring to the doorbell switch or install a wireless system.

TEST THE SWITCH

1 Unscrew the doorbell switch cover to expose the wiring.

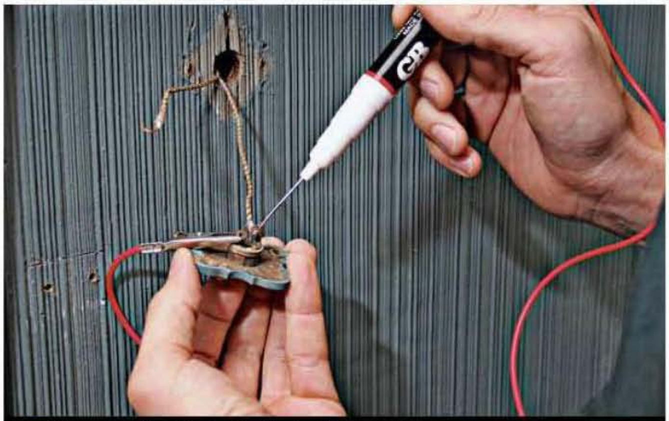

2 Use a continuity tester to test the switch.

TEST MULTIPLE COMPONENTS

3 Join the two bare ends of wire to test the system.

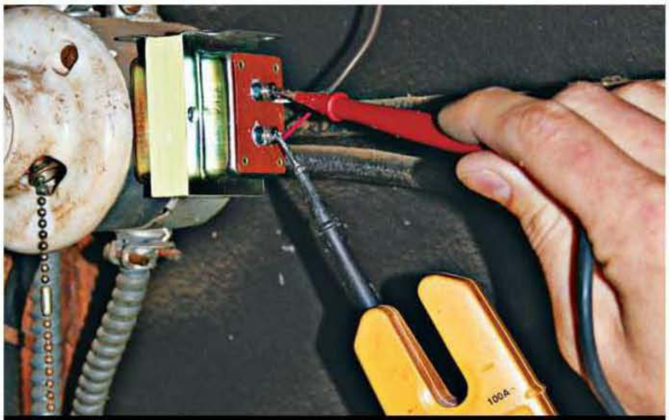

4 Testing the transformer is the last option.

Perhaps the trickiest part of low-voltage repairs is running the cable. If you're replacing dead lo-vo wires, such as those running to a doorbell, try to twist and tape new low-voltage wires to the old ones and pull them through the walls. Because low-voltage wire is thin, it's flexible and much easier to pull than Romex.

If you must drill through an exterior wall to bring low-voltage wiring into a room, avoid hitting cables that may be hidden in the walls. The safest place to drill is generally low on the wall, away from studs.

➔ **See "Fishing Cable," p. 191 and "Pulling Cable," p. 197.**

As shown in the rough-in sequences on p. 197, most house circuits are routed horizontally about 2 ft. high and secured to studs before entering outlet boxes. If you drill below that height and between studs, chances are good that you'll avoid hitting cables. To further reduce any chance of getting a shock, use a cordless drill for this task.

Once you've found the right location, drill from the interior of the house downward to keep water from running into the walls.

➔ **See photo 2, p. 125.**

After pulling the cable, fill the hole with siliconized latex caulk, which, unlike pure silicon, is paintable. When running low-voltage wire, staple it at least every foot to prevent sagging, strains on connections, and an unsightly appearance.

WARNING

It's safe to handle the energized low-voltage wires that run from the transformer to the doorbell switch or from the chimes to the switch. However, because 120v current can harm you, turn off the power before testing or repairing a transformer or the wires upstream—that is, toward the power source—as shown in the upper right drawing.

SINGLE DOORBELL WIRING

In a single doorbell system, the circuit runs from the transformer to the doorbell switch, from the switch to the chime unit, and then back to the transformer. By pressing the doorbell, you complete the circuit, and the chime unit rings.

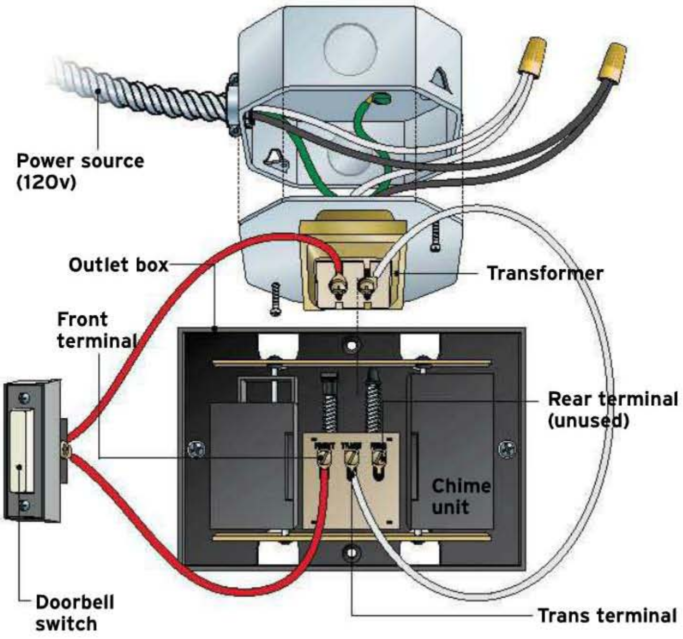

TROUBLESHOOTING A DOORBELL (CONTINUED)

DOUBLE DOORBELL WIRING

A double doorbell system has two circuits, each controlled by a doorbell switch, whose power comes from the transformer. Thus the chime unit for a double doorbell has three terminals. Typically, the chime unit has two different ring patterns (*ding* and *ding-dong*) so you can tell whether the visitor is at the front or rear door.

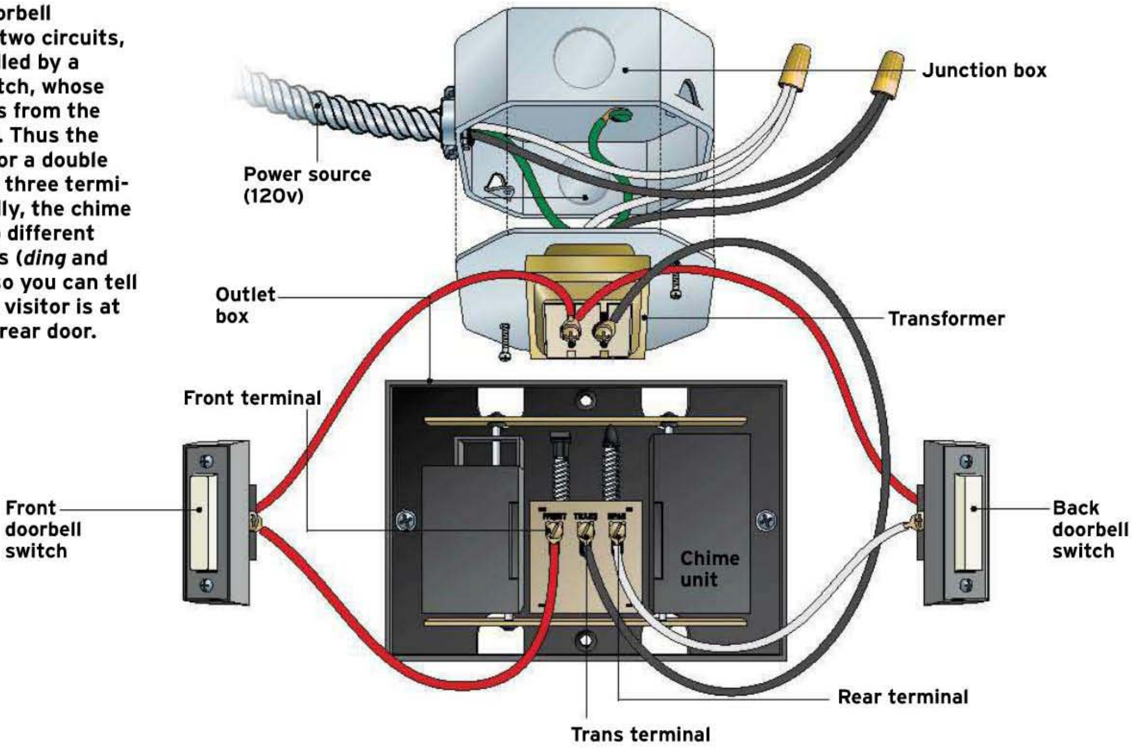
Junction box
Power source (120v)
Outlet box
Transformer
Front terminal
Front doorbell switch
Chime unit
Back doorbell switch
Rear terminal
Trans terminal

Testing chimes If the chimes don't ring when you touch the doorbell switch wires together, remove the chime cover and vacuum out accumulated crud. If the chime plungers are rusty or corroded, spray them with WD-40® or a similar lubricant and move them by hand to get them sliding freely. Rubbing screw terminals with steel wool may improve electrical contacts. But chances are the old unit is worn out and needs to be replaced. To find out, dial a multimeter AC setting that's close to the low-voltage rating on the chime unit, then touch the tester probes to the *trans* and *front* terminals and to the *trans* and *rear* terminals, as shown in the photo at right. If you get a reading close to the unit's rating but the chimes won't ring, the transformer is delivering power, but the chime unit is defective. Replace it.

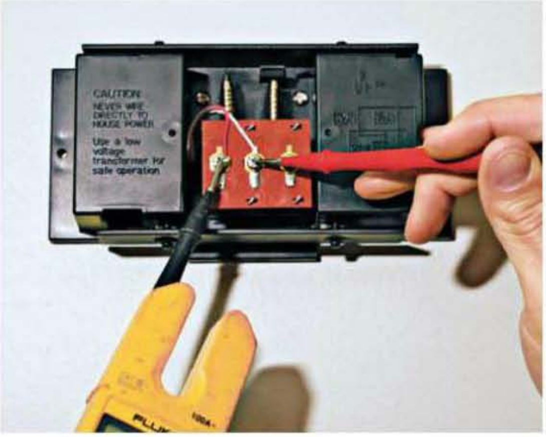
CAUTION
NEVER WIRE DIRECTLY TO HOUSE POWER
Use a low voltage transformer for safe operation

Test the chime terminals using a multimeter.

REPLACING A DOORBELL SWITCH

Replace a doorbell switch if it sticks, it is damaged, or continuity no longer exists between the contacts when the switch is depressed. You may also decide to upgrade simply because you want a newer style.

Unscrew the low-voltage wires and pull them out from the box for easy access. If they show corrosion, scrape or sand the wires lightly. Once the wires are clean and exposed, screw them to the terminals on the new doorbell switch. Then screw the new switch to the wall.

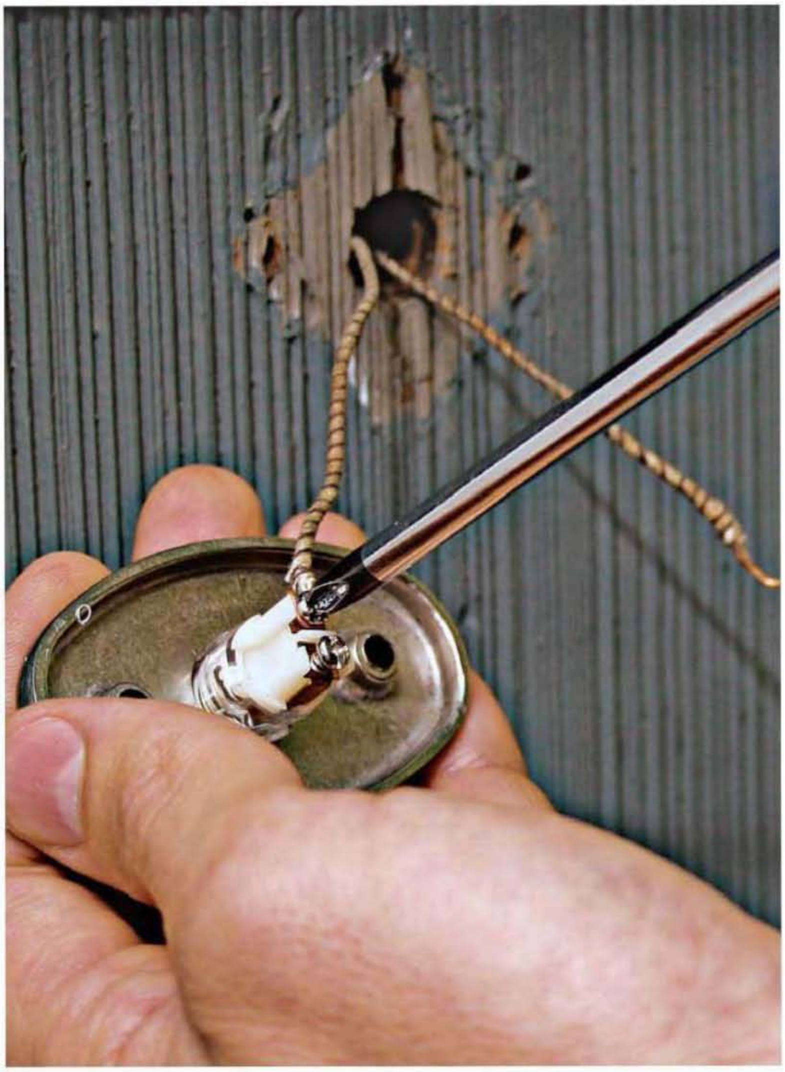

Screw the old wires to the terminals on the new switch.

Choosing a new doorbell When replacing a doorbell switch/button pay close attention to the existing mounting holes. Purchase a replacement button that has the same hole pattern or a larger hole pattern. Installing a replacement button that is smaller than the original will require patching and painting of the original holes, which can be difficult and, depending on the mounting surface, may add to the amount of work necessary to complete the project. Installing a button with a hole pattern that does not match the existing button, but is close, can lead to problems when trying to install new screws too close to the existing screw holes.

INSTALLING A CHIME UNIT

Once you've determined that the old chime is defective, unscrew the low-voltage wires from their terminals and unscrew the mounting screws that secure the unit to the wall. As you carefully remove the old unit, hold on to the low-voltage wires so they don't fall into the wall. If the exposed wires are especially short, it's a good idea to tape them to the wall with duct tape.

After making sure the new unit has the same voltage rating as the old unit, remove it from the packaging. Feed the low-voltage wires through the access hole in the new unit ❶. Level the unit's housing and then screw it to the wall. Because chime units are lightweight and installed in out-of-the-way places, it's seldom necessary to screw the chimes to the framing. It's usually sufficient simply to screw them to drywall or plaster ❷.

If the low-voltage wires look corroded or cracked and there's enough extra wire, snip their bare ends and restrip them. Then use needle-nose pliers to loop the ends clockwise ❸. Place the looped ends on the unit's terminals, exerting a slight pressure on the wires to keep them from slipping off as you tighten the screw terminals ❹. After you tighten the screw terminals, press the doorbell switch to test the new chimes. If they work correctly, snap on the unit's protective cover ❺.

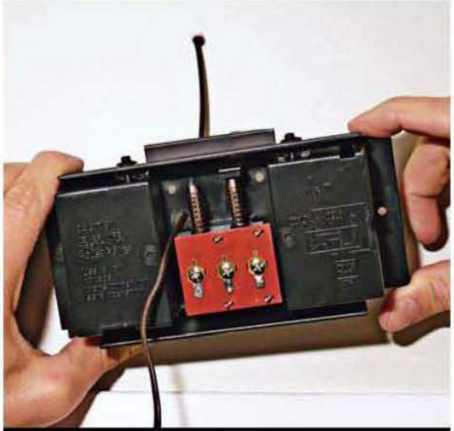

1 Feed the wires from the wall into the new unit.

2 Level and mount the chime unit on the wall.

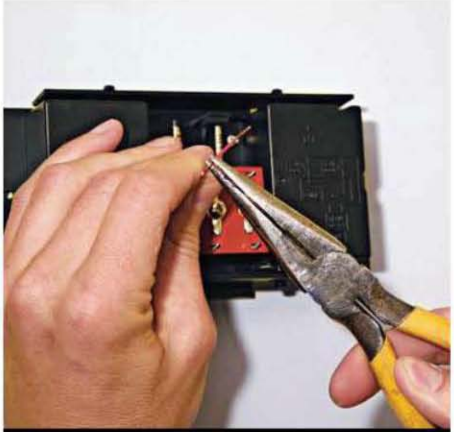

3 Strip and loop the low-voltage wires.

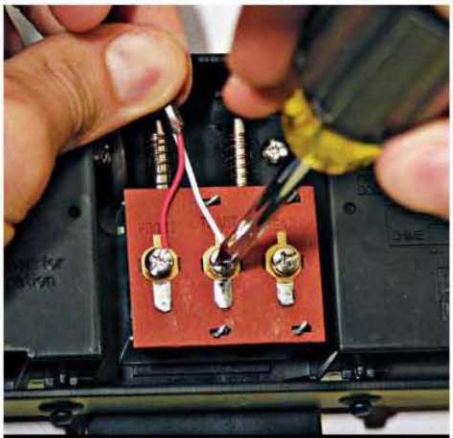

4 Attach the wires to the screw terminals.

5 Snap the cover into place and turn on the power.

TRADE SECRET

If doorbell wires are defective, try twisting new wires to the old ones, taping the splice, and pulling the new wires through the walls. If you don't succeed, install a wireless doorbell and chime system instead.

INSTALLING A DOORBELL TRANSFORMER

A doorbell transformer can be attached to any junction box that has enough room to accommodate the additional wire splices and is close enough to the doorbell switch that running wires to it is practical. But because junction boxes house wires with 120v voltage, you need to shut off the power to the junction box, carefully remove its cover, and then test to be sure the power is off ❶.

Once you've confirmed that no power is present, use a screwdriver or sturdy pliers to remove a knockout from the junction box ❷. Feed the new transformer's wires through the knockout opening ❸. Tighten the transformer's mounting screw to draw the unit tight to the junction box and cover the knockout opening ❹. (There is no need for Romex cable clamps or the like.)

If there is a green grounding wire from the transformer, screw it to the metal junction box, using a green grounding screw. If the box is plastic, splice the transformer's ground wire to the ground wire of the supply cable.

Strip the ends of the transformer wires and use twist-on wire connectors to splice them to the 120v supply wires. Splice like to like: Splice the hot (red or black) transformer wire to the hot supply wire; the neutral (white) transformer wire to the neutral supply wire ❺.

Once you have spliced the wires, tuck them into the junction box and replace the cover (the base of a porcelain light fixture is shown in the photo). Then connect the low-voltage wires to the screw terminals on the transformer ❻. Turn on the power and test the system.

TRADE SECRET

Before installing a new transformer, make sure it has the same voltage rating as the old one. The voltage rating should be stamped somewhere on the transformer housing.

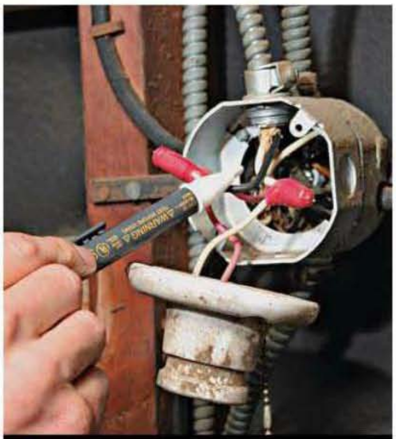
1 Turn off the power, then use an inductance tester to make sure it's off.

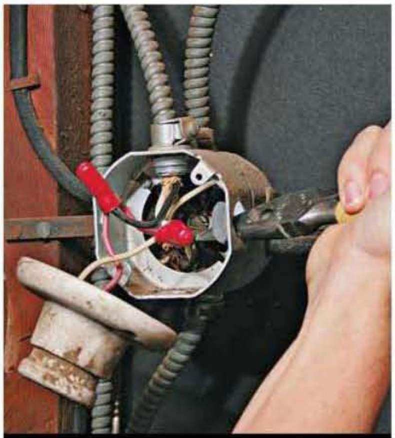
2 Use needle-nose pliers to remove a knockout.

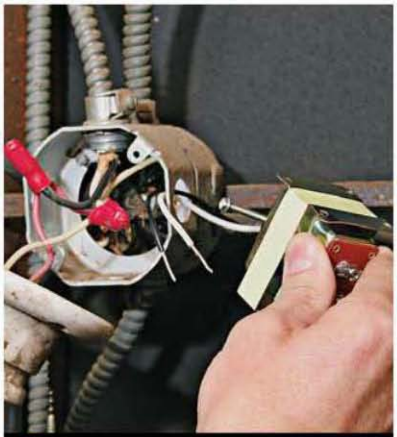
3 Pull the wires from the new transformer through the knockout.

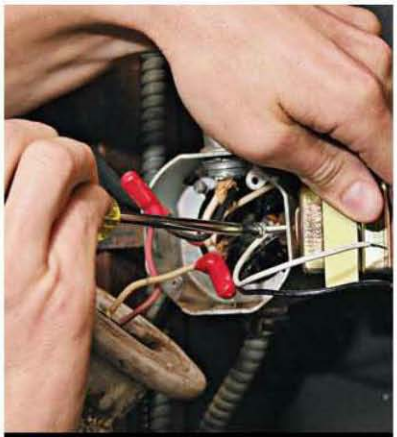
4 Mount the transformer to the box.

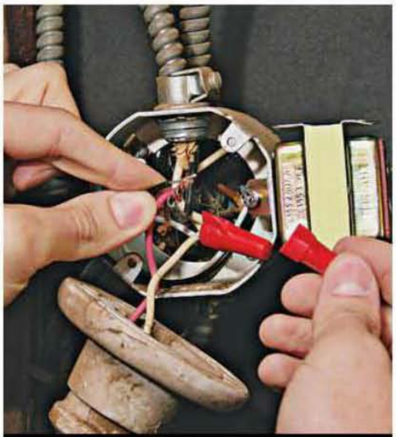
5 Splice the transformer wires to the supply wires.

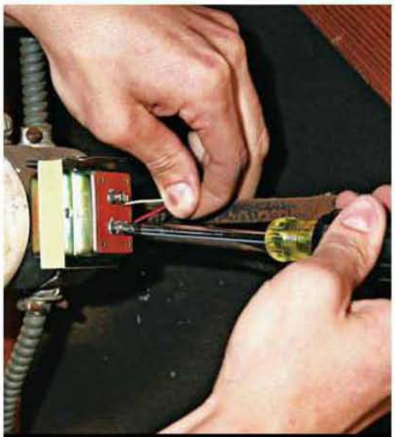
6 Attach the low-voltage wires to the transformer terminals.

TESTING AN OLD THERMOSTAT

If your thermostat is a low-voltage unit, it gets its reduced voltage from a transformer. If the thermostat doesn't turn the furnace on and off, remove its cover to see if it's clogged with dust or if the wires are loose ❶. Remove dust using a small painter's brush and reattach loose wires.

If the thermostat still doesn't work, use a continuity tester or a multimeter set to ohms (resistance) to test the unit. Unscrew the low-voltage wire attached to the R terminal and touch the tester probes to the R and W terminals. As you hold the probes to the terminals, flip or turn the setting lever from one side to the other ❷.

If the continuity tester lights up or the multimeter shows any reading at all, the thermostat works and the problem is the transformer. If there's no light or no meter reading, the thermostat is defective. Disconnect the remaining wire(s) and remove the unit ❸.

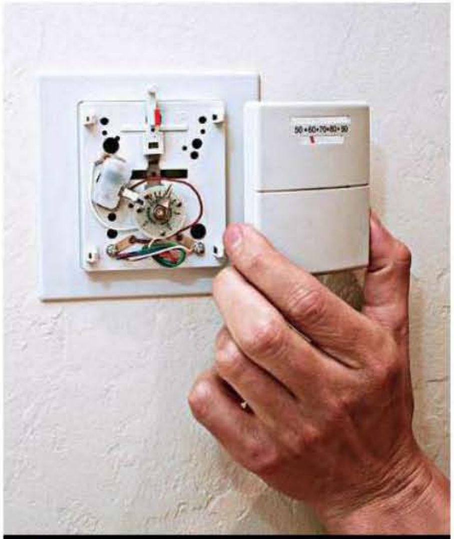

1 Remove the cover of the existing thermostat.

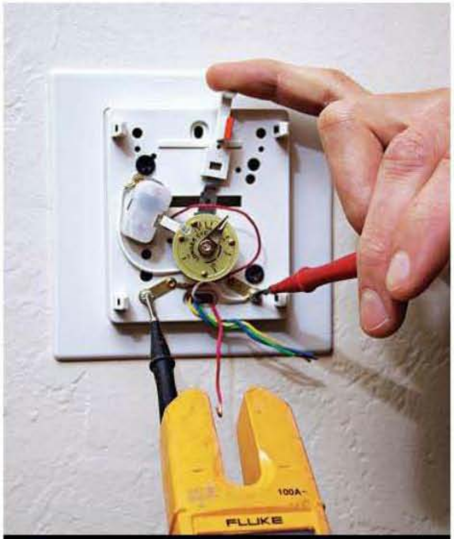

2 Test the terminals using a multimeter while adjusting the setting lever.

3 If defective, remove the existing unit.

WARNING Thermostats are specialized switches that open or close a circuit in response to temperature changes. Although most thermostats are low-voltage units that are safe to handle, your unit may be unsafe to handle if it mounts to an electrical box and is connected to house wiring. Shut off the power, test to be sure it's off, disconnect the thermostat, and have a heating professional assess it.

INSTALLING A PROGRAMMABLE THERMOSTAT

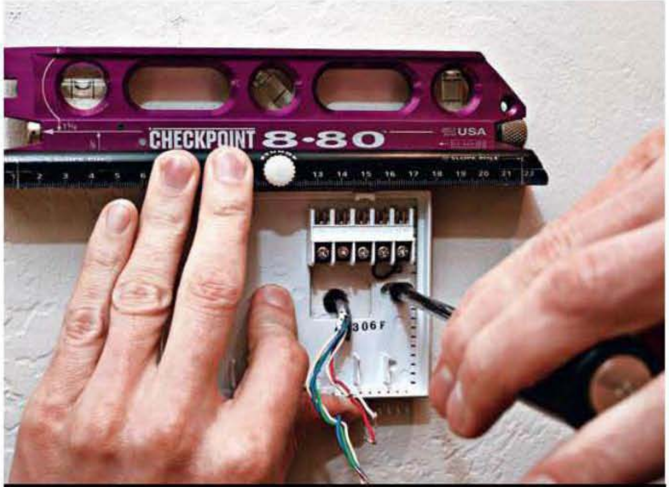
CHECKPOINT 8-80
USA
306F

1 Mount the new thermostat and make sure it's level.

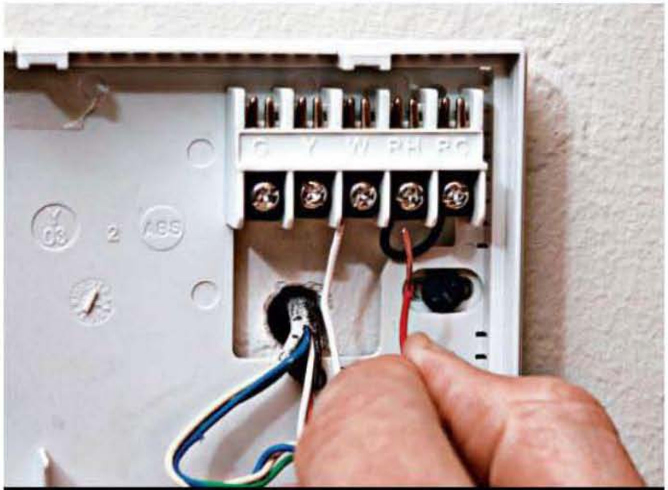

2 Attach the existing wires to the designated terminals on the thermostat.

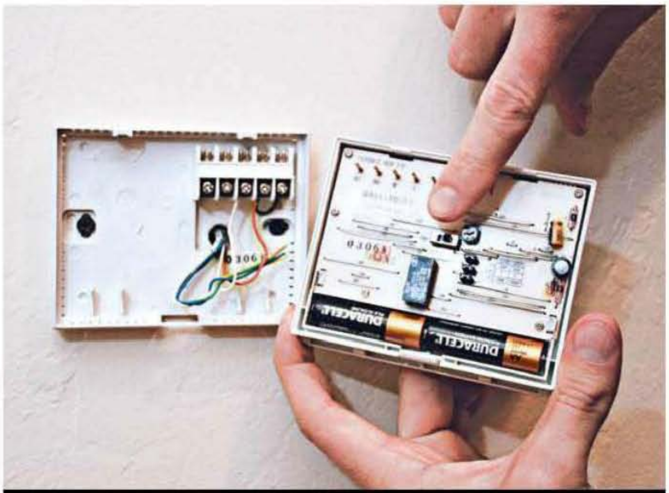
DURACELL
DURACELL
306

3 Insert batteries and select the appropriate heat source.

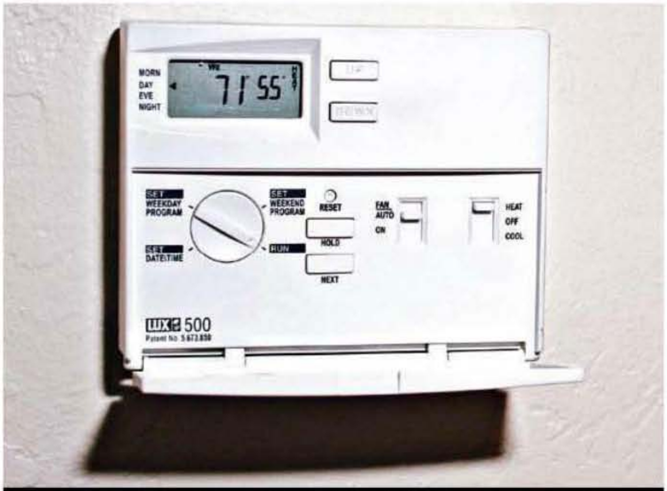
MORN
DAY
EVE
NIGHT
71 55
SET WEEKDAY PROGRAM
SET WEEKEND PROGRAM
RESET
SET DATE/TIME
RUN
HOLD
NEXT
FAN AUTO
ON
HEAT
OFF
COOL
LUX 500

4 Make sure the thermostat works and then program the unit.

Though most thermostats are installed in the same way, be sure to follow the installation instructions that come with your programmable unit. Most thermostats require leveling to function properly ❶. Once the unit is on the wall, attach the existing low-voltage wires to the designated terminals on the new unit ❷. Either insert the stripped wire ends into push-in clamps or loop them clockwise around screw terminals. Many modern units have batteries that need to be changed periodically. You may also need to set a switch on the back of the thermostat plate to select the type of heat (electric, oil, or gas) ❸. Replace the thermostat's cover, consult the instructions, and program the unit ❹.

TRADE SECRET

If the thermostat works but its temperature settings are too high or low, it may need to be plumbed or leveled. Remove its cover and look for a pair of plumb or leveling marks on the base.

REPLACING A FURNACE TRANSFORMER

If your thermostat isn't getting power, the furnace transformer may be defective. If your furnace is relatively new, the transformer will be inside the furnace housing—in which case, call a HVAC (heating, cooling, air-conditioning) specialist to assess and replace it. New HVAC units are sensitive and complicated. However, if the transformer is mounted to a junction box on the outside of the housing, it's relatively easy to replace.

Set your multimeter to a low-voltage AC setting and touch its probes to the two terminals of the transformer ①. If there's no reading, replace the transformer. Begin by shutting off the power and then using a voltage tester to be sure it's off. Then remove the old transformer and unscrew the cover on the junction box. (In the example shown here, the transformer and box cover are an integral unit.) After removing the cover, test the exposed house wires again to be sure the power's off.

Use a screwdriver or a sturdy pair of pliers to remove a knockout on the side of the box and insert the wire leads of the new transformer into the opening ②. Tighten the diagonal screw that mounts the transformer to the junction box, then screw the transformer's green grounding wire to the metal box using a green grounding screw. If the box is plastic, splice the transformer's ground wire to the ground wire of the supply cable. Use wire connectors to splice like wire groups: connect the transformer's hot lead (black) to the hot supply wire and the neutral lead (white) to the neutral supply wire ③.

After making these connections—but before covering the junction box—turn the power back on. Touch the probes on the multimeter to the transformer terminals to make sure the new unit functions normally ④. Then turn the power off again. Snip, strip, and loop the existing low-voltage wires as needed and attach them to the transformer terminals ⑤. Finally, gently push the wire splices back into the junction box and screw the cover in place ⑥.

WARNING

Make sure the new transformer is properly rated for the furnace. Check the voltage rating stamped on the old transformer or look inside the cover of the furnace, which will list the low-voltage requirement for the transformer.

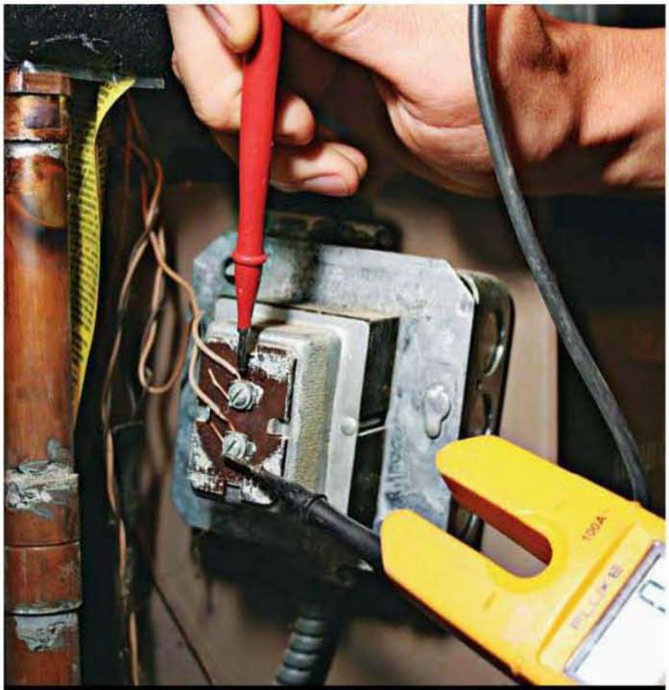

1 Touch the multimeter probes to the transformer terminals to test the unit.

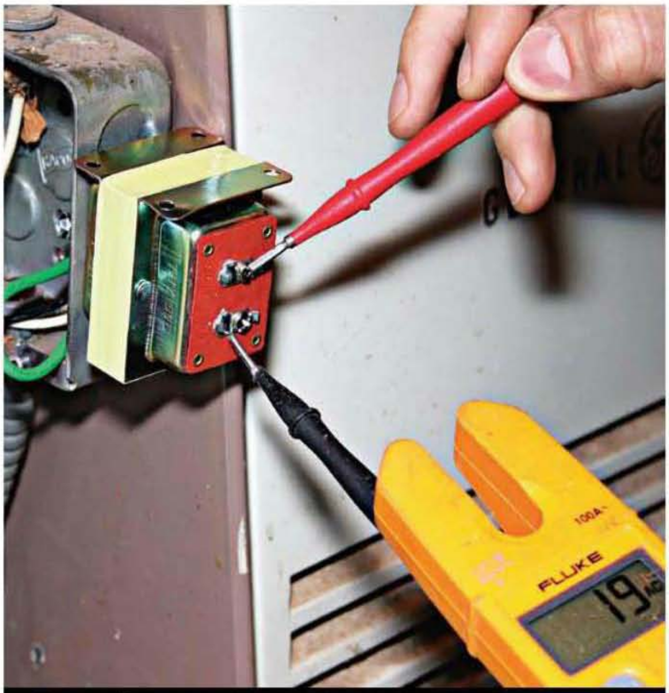

4 Turn off the power, then test the new transformer using a multimeter.

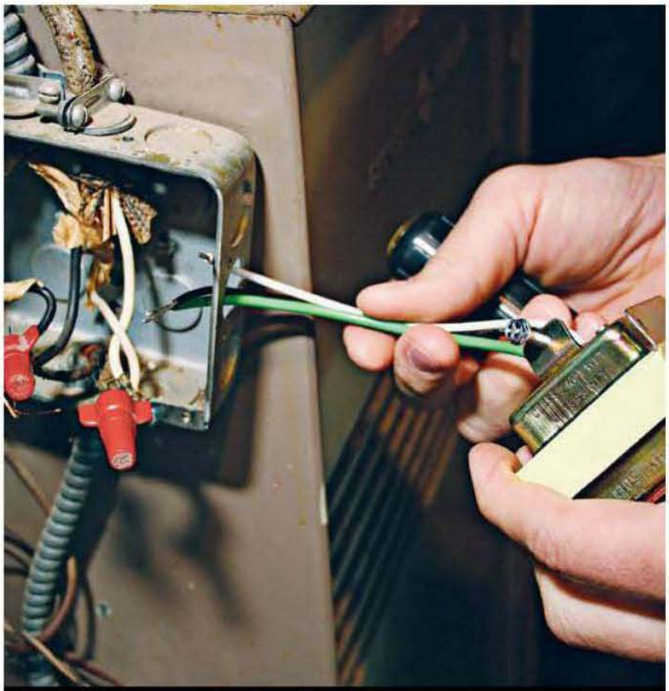

2 Turn off the power, then remove a knockout and route the transformer leads into the opening.

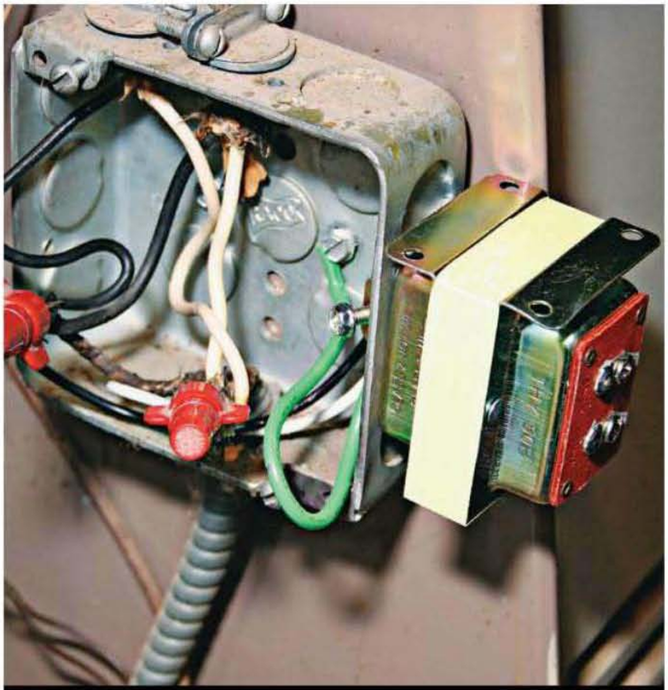

3 Attach the ground wire first, then splice like wire groups.

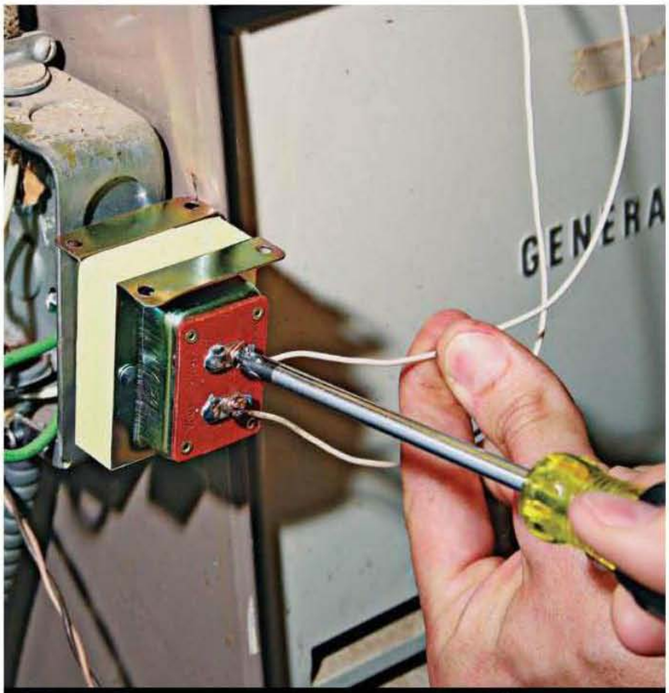

5 Turn off the power, then attach the low-voltage wires to the transformer.

GENERAL ELECTRIC

6 Tuck the wires back inside the junction box and install the cover.

MULTIMEDIA

MULTIMEDIA ENCOMPASSES various types of video, sound, voice, and data signals, and have multiple format options. Wiring multimedia outlets and distribution panels is somewhat like assembling a computer, though prefabricated modules and interchangeable parts make home wiring fairly straightforward.

The tools and methods used to install cables are similar to those used to install basic residential electrical circuits. Whether you're adding a phone jack to existing service or installing a media distribution panel, you'll need only a few specialized tools, such as a long, flexible drill bit if you'll be running wires through walls. Multimedia connections are extremely precise, so it's important to follow installation instructions exactly to ensure strong signals throughout your sound, video, phone, and data networks.

TOOLS & MATERIALS

Multimedia cable & connectors, p. 122
Stripping cable, p. 123

ADDING PHONE JACKS

Extending a phone line, p. 124
Mounting a new jack, p. 126

DISTRIBUTION PANELS

Mounting a media distribution panel, p. 127
Running cables to a media distribution panel, p. 130
Connecting cables at the panel, p. 131

INSTALLING A MULTIMEDIA OUTLET

Locating the multimedia outlet, p. 132
Installing cable connectors, p. 133
Connecting Panduit jacks, p. 134
Installing Leviton jacks, p. 135
Attaching jacks to an outlet plate, p. 136

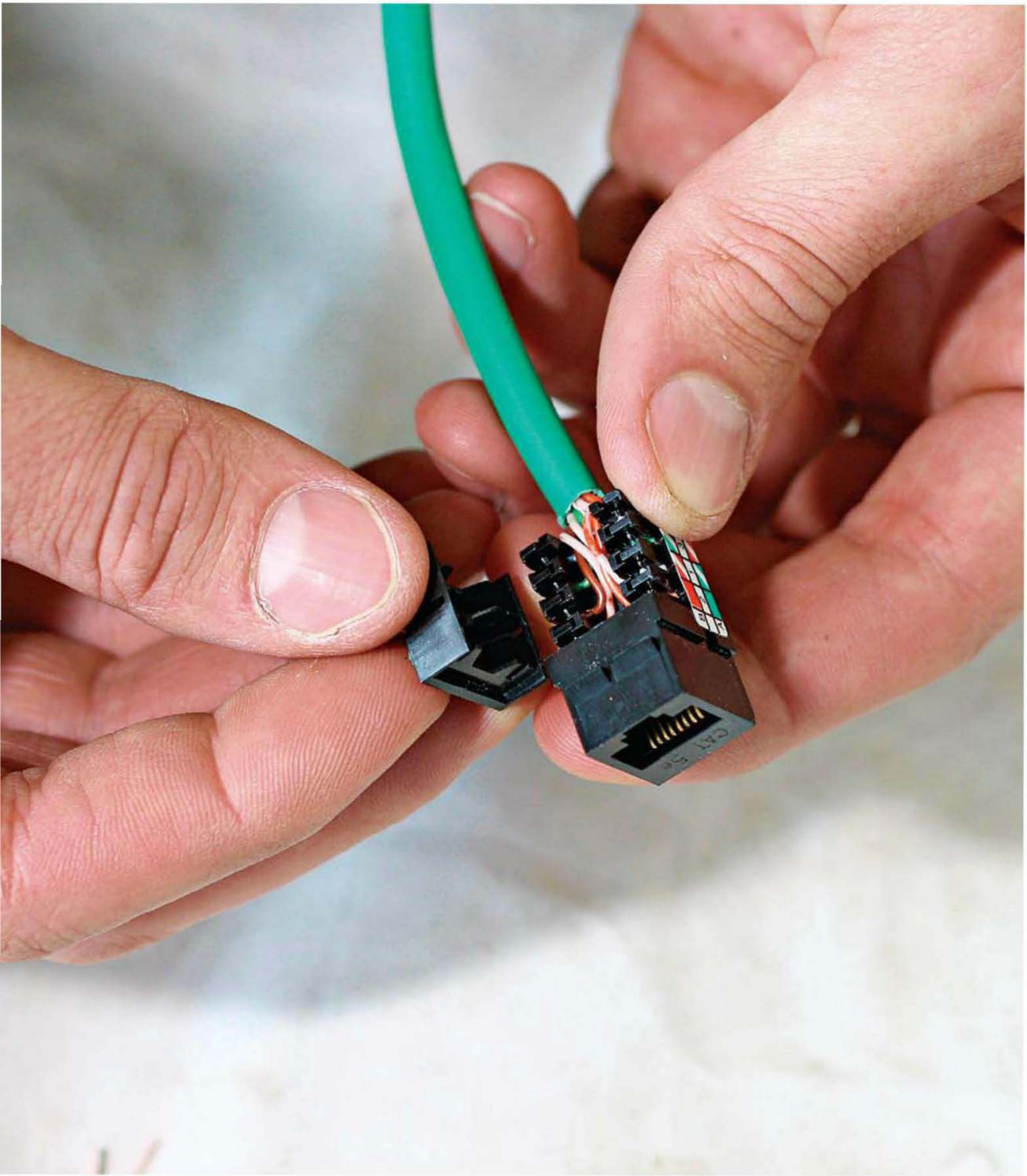

CAT 5e

MULTIMEDIA CABLE & CONNECTORS

Here's an overview of the cable and connectors you'll need to wire your house for sound, video, data, and telephone service:

Coaxial cable includes double-shielded RG6 and RG59 cable and single copper-shielded RG59. RG6 cable has a slightly thicker wire gauge than does GR59. Used for video distribution (cable TV), coaxial cable has been around for years. Video cable terminates in an F-connector such as the watertight RG6 connector shown in the photo at right. To simplify life and forestall callbacks, many professional electricians install watertight F-connectors inside and out.

Data cable can carry data or phone signals; it is typically 24-guage, solid-wire unshielded twisted pair (UTP) cable. Although Category (Cat) 5e cable meets present-day standards for data transmission, Cat 6 wire is likely to supplant it. In general, the higher cable numbers denote faster data-transmission capabilities. Data cable contains four pairs of wires, thus RJ-45 data jacks contain eight pins.

Telephone cable is usually Cat 3 cable, which is not twisted; its solid wire is considerably thinner than Cat 5e cable. RJ-11 phone jacks contain six pins. Although data cable can carry phone or data signals, Cat 3 phone cable is not rated for high-speed data transmission.

Audio (speaker) cable is usually 18-gauge to 12-gauge finely stranded wire. Speaker wire terminations vary from bare wires compressed between stacked washers to screw-on or crimped jacks that plug into speaker ports. Plug-in jacks are color striped to match speaker polarity: red-striped jacks for positive terminals and black-striped jacks for negative terminals.

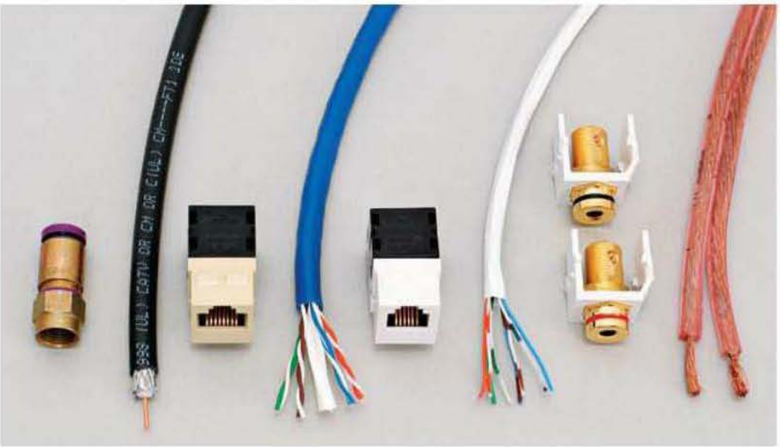

Each multimedia connector is to the left of the cable it terminates. From left: RG6 F-connector, dual-shielded RG6 coaxial cable; RJ-45 (eight-pin) jack, Cat 6 UTP data cable; RJ-11 (six-pin) jack, Cat 3 phone cable; two audio jacks (sometimes called *banana jacks*), 14-gauge low-loss audio cable.

Signal strength: coming in loud and clear

Although solid connections are as important to multimedia as they are to all electrical systems, signal strength—not voltage—is the objective when connecting data, sound, video, and phone components. In fact, most multimedia input is not impelled by AC, as is house wiring. Rather, video and phone signals are generated by cable or phone companies. Computers and routers amplify data signals, and stereo amplifiers boost sound signals; but although those devices run on house current, the signals themselves are not AC. Thus the cables that carry multimedia signals are dramatically different from, say, Romex cables, and require different connecting devices and a few specialized tools.

TRADE SECRET

Pros use watertight connectors on all coaxial cable connections, even those installed indoors. Watertight connectors such as the RG6s shown in the photo above don't cost much more and they always provide a solid connection.

STRIPPING CABLE

Electricians generally favor one type of stripper and use it to strip everything, thus reducing the number of tools in their belts.

Splicing scissors can trim tiny wires, but they can also score sheathing: Hold the cable in one hand, and, with the other, hold a scissors blade perpendicular to the cable and rotate it around the cable Ⓐ. It's not necessary to cut through the sheathing. Once scored, the sheathing will strip off when you pull on the scored sections Ⓑ. In fact, merely scoring the sheathing is less likely to damage individual wire insulation.

The cable scorer shown in photo Ⓒ is faster than using splicing scissors, but its razor blade is so sharp that it can easily nick wire insulation if you're not careful. Wire strippers are the most reliable way to strip insulation off individual wires because you can choose a stripper setting that matches the wire gauge Ⓓ. Finally, whether stripping wire, drilling holes, or doing any other wiring task, wear eye protection and sturdy gloves.

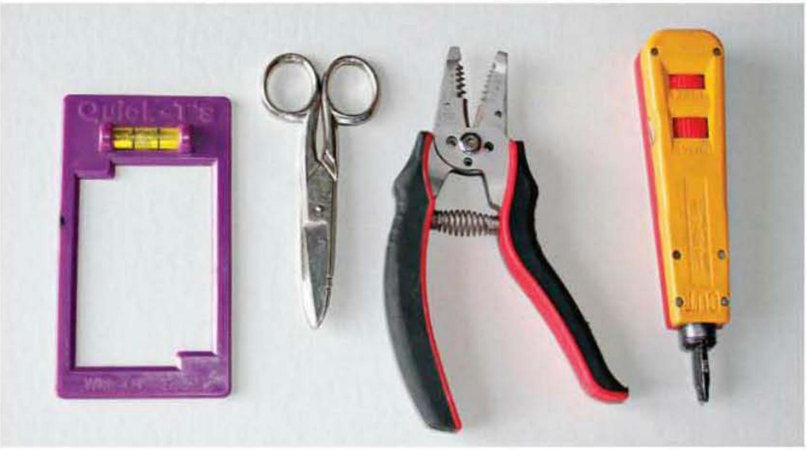

Multimedia installation tools. From left: low-voltage cut-in ring template, splicing scissors, wire strippers, and punch-down tool.

TWISTED WIRES

Remove the sheathing from Cat 5e or Cat 6 cable, and you'll discover twisted wires within. Twisting wires reduces the occurrence of *cross-talk*, in which electromagnetic signals jump from wire to wire. Although it is necessary to straighten some data cable wires to attach them to some jacks, the twisting must be maintained to within 1/4 in. of the termination on the jack.

Cat 5e and 6 cables are examples of UTP cable, which is twisted to prevent signals from jumping between wires.

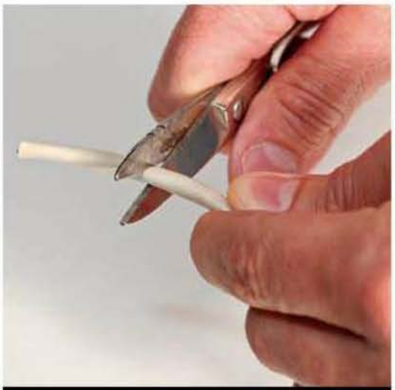

A **To score sheathing, lightly rotate splicing scissors around the cable.**

B **Properly scored, the sheathing should slide off with a gentle tug.**

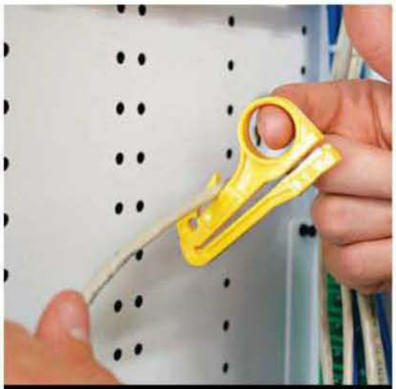

C **Lightly twirl a cable scoring tool around the sheathing.**

D **Wire strippers have settings that match wire gauges.**

EXTENDING A PHONE LINE

Thanks to cell phones and remotes, we're no longer tied to the room where the phone line ends. But for clear, reliable service you can't beat a hard-wired phone plugged into a nearby jack. Happily, almost anyone can run an extension from an existing jack, thereby saving a hefty installation fee from the phone company.

The only tricky part of the job is running the cable to the new jack. You can tuck it under carpets, tack it atop baseboards or run it around door jambs, but it won't look great. For the cleanest job, route that extension line into the attic or basement or—as shown here—drill through the wall and run it outside.

Before you start, find the shortest route between the existing jack and the new one. Measure that route carefully and add enough extra cable for drip loops and at least 1 ft. extra on each end for the thickness of walls and for stripping and connecting to the jacks. Buy Cat 3 or 5e cable, which contains four pairs of solid-core, 24-gauge wire. Because phone lines are low voltage, they're safe to handle. You can do the job with common tools.

Connecting to an existing jack

Start by unscrewing the cover on the jack and the mounting screws that hold the jack to the wall. Gently pull the jack out from the wall, being carefully not to pull loose the wires attached to the jack terminals ❶.

If the jack is surface mounted, drill a hole for the extension cable that will be covered by the jack. If the jack is flush mounted, drill anywhere inside the ring, because the hole will be covered by the jack cover. Drill at a downward angle ❷ so that exterior water will tend to run out of the

TRADE SECRET

Many electricians install Cat 5e cable for both data and phone lines to allow future expansion or modification. That is, Cat 3 is fine for phone service but inadequate for data, whereas Cat 5e can carry both signals.

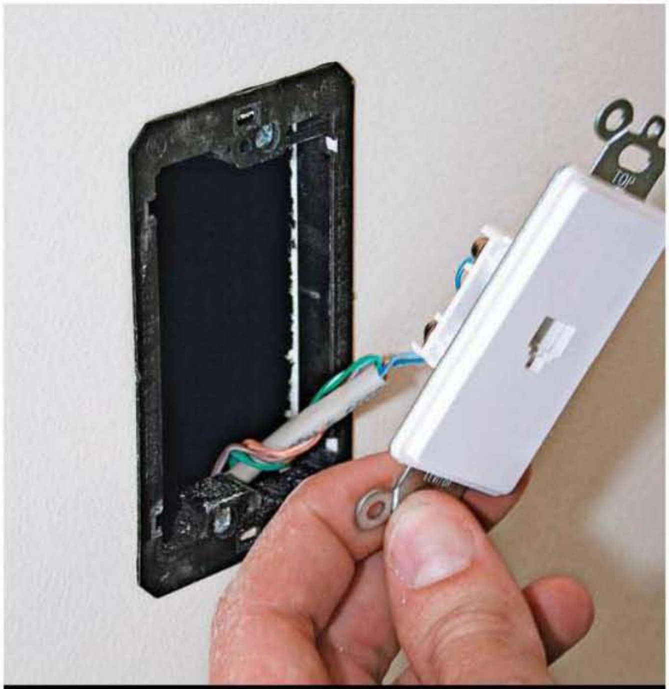

1 Gently pull the jack out from the wall.

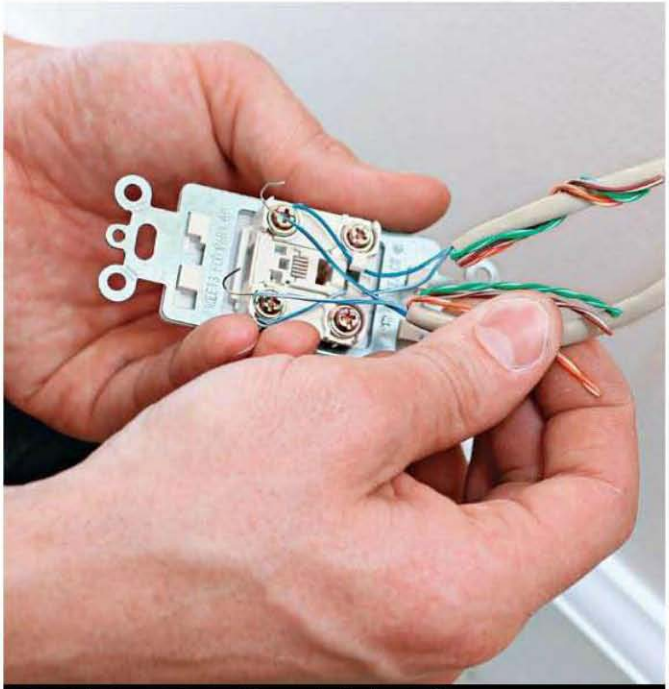

3 Strip the cable sheathing and wire insulation.

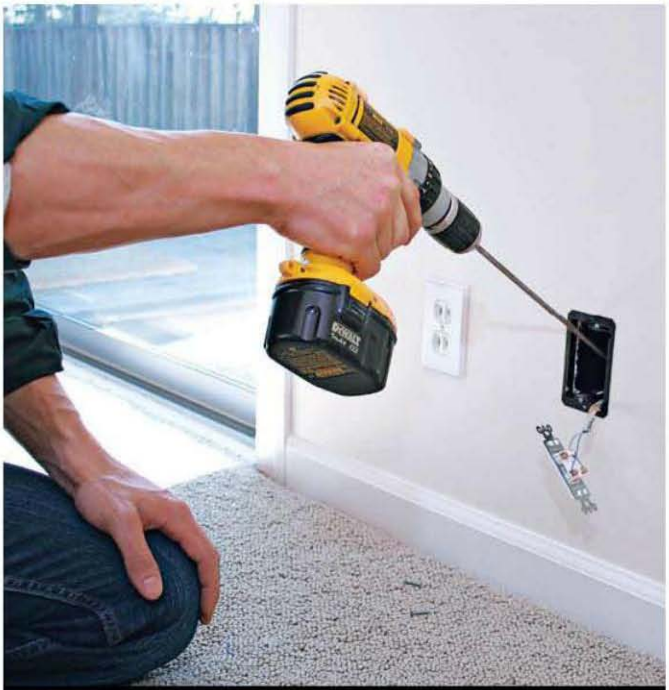

2 Drill downward to prevent water from traveling to the interior.

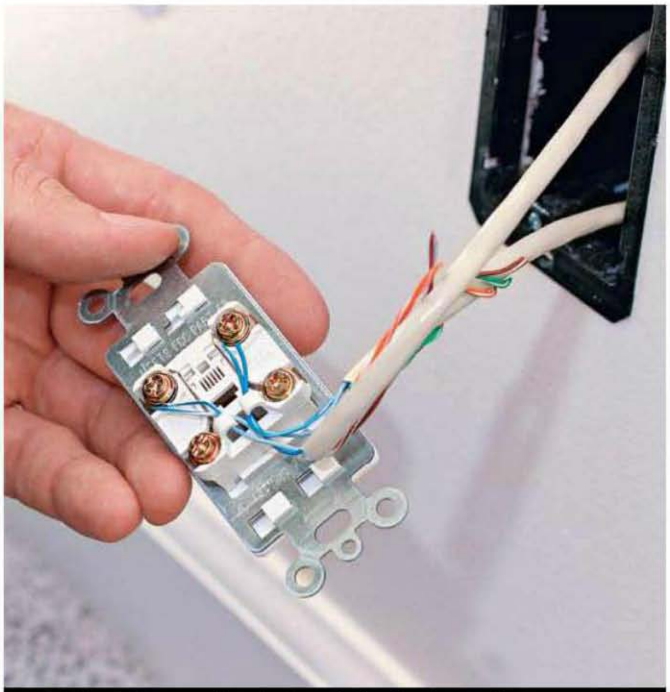

4 Attach the new wires to the existing jack's terminals.

hole. Look into the outlet opening before drilling to avoid electrical cables in the wall. Use a 1/4-in. extension bit in a cordless drill for the job.

After drilling through the wall, use duct tape to attach a piece of string to the bit and pull the bit back through the wall. Then tie the new Cat 3 or 5e phone cable to the string and pull it into the hole behind the jack. Once you've pulled the new cable to the existing jack, strip about 2 in. of the cable sheathing and separate the cable wires into pairs ❸.

Using splicing scissors or a wire stripper, strip about 1/2 in. of insulation from a wire pair (for example, a blue and a blue-white wire), loop the bare wire ends clockwise, and attach one wire to each of the two jack terminals that are presently wired ❹. Typically, terminals have stacked washers that hold several wires.

Wiring techniques

When connecting to an existing phone line, avoid bending existing wires repeatedly because they can become weak and break off. Screw terminals come with multiple washers and are intended to have one wire under each washer. It is not recommended to twist wires together and install them under one washer. Finally, always wrap wires around screw terminals in a clockwise direction.

MOUNTING A NEW JACK

Before pushing the cables into the wall and remounting the existing jack, staple the new cable so it can't move and stress electrical connections. If the cable runs outside, loop it downward so water will drip off, and staple it with insulated cable staples ❶. Then fill the hole in the siding with siliconized latex caulk.

Locate the new jack and drill through the wall to bring cable to the location ❷. As described earlier, tape a string to the end of the drill bit before withdrawing it, then tie the new cable to the string. Once you've pulled the new cable into the room, remove its sheathing, strip insulation from a pair of wire ends, feed the cable through the new jack, and screw the jack to the wall ❸.

Loop the bare wire ends clockwise, insert them between the stacked washers on the jack, and screw them tight ❹. Finally, tug gently on the wires to be sure they're well attached, tuck the wires neatly out of the way, and snap on the jack cover ❺.

TRADE SECRET

When running low-voltage wire or multimedia cable, staple it at least every foot to prevent sagging, strains on connections, and an unsightly appearance.

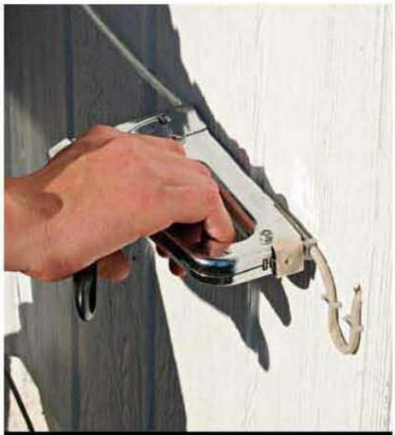
1 Form a drip loop, then staple the cable to the wall.

2 Drill a hole at the new jack location and pull in phone cable.

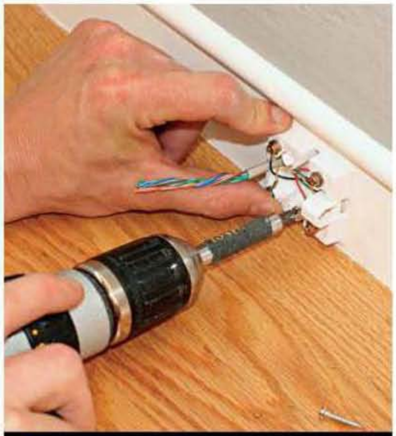
3 Mount the new jack in the opening.

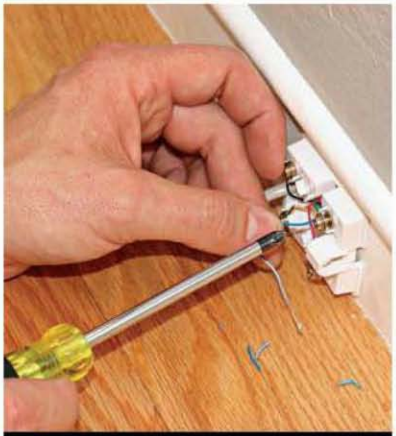
4 Screw the stripped wire ends to the jack.

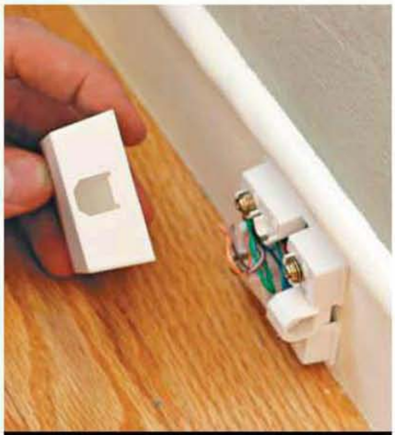
5 Snap on the jack cover and double-check to make sure it works.

TRADE SECRET

If it's necessary to drill through an exterior wall to bring lo-vo wiring or media cables into a room, drill downward to keep water from running into the walls. After inserting the wire or cable, fill the hole with siliconized latex caulk which, unlike pure silicon, is paintable.

MOUNTING A MEDIA DISTRIBUTION PANEL

A distribution panel is the heart of a home multimedia network because it receives incoming signals from phone, cable, and data companies and distributes them throughout a house. Panels can also distribute the signals from a stereo or sound system to speakers in many rooms. Many panel makers offer solutions with prewired modules, so a homeowner can configure a multimedia network today that can be easily changed tomorrow. To successfully install your system, closely follow the instructions provided, using the recommended tools.

As with other electrical installations, connect the power last, after wiring the panel and running cable to the outlets. Wear safety glasses at all times and work gloves to protect your hands. Panel edges can slice your hands in a flash; metal shards dislodged by hammers or drills are also dangerous. Finally, sturdy gloves enable you to grip and pull cables more easily.

Mounting the panel

Locate the distribution panel centrally so you'll have shorter cable runs: None should be longer than 300 ft. from the panel to an outlet. To minimize electrical interference, locate the multimedia panel away from the service entrance panel or a subpanel. Ideally, place the distribution panel on an interior wall to minimize temperature fluctuations—never locate it in an attic or unheated garage. You'll build in future flexibility if you run Cat 5e or Cat 6 cable for all data and telephone lines.

Install the rough-in box for the multimedia distribution center by screwing it to adjacent studs ❶. If you've already pulled most of cables to feed the panel, as shown here, tack them to one side so they'll be out of the way. Slide the panel in or out so its edges will be flush with the drywall ❷.

>> >> >>

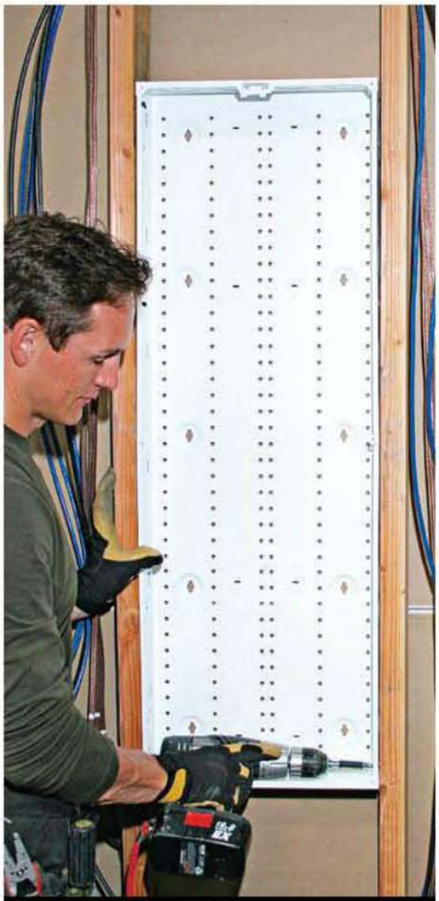

1 Screw the panel to the studs.

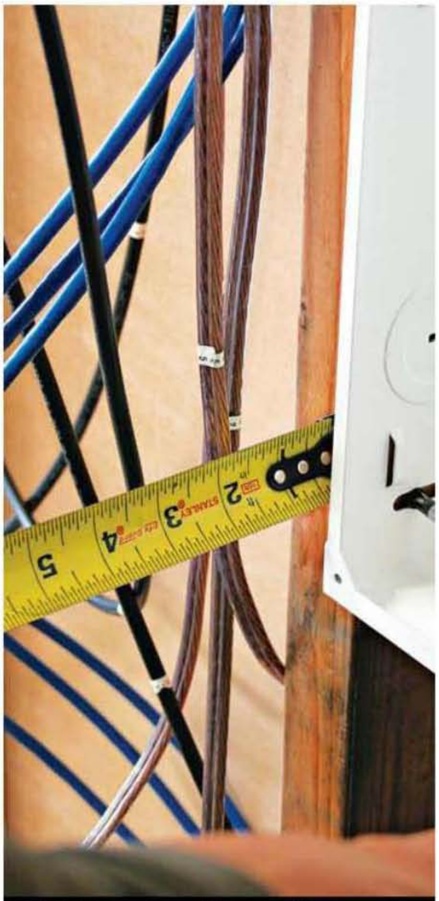

2 Set the edges of the panel flush with the drywall.

TRADE SECRET

When mounting panels, install screws at the top loosely to hold the panel in place. Set and secure the bottom, then reset the top screws as needed.

MOUNTING A MEDIA DISTRIBUTION PANEL (CONTINUED)

Remove the knockout(s) for the power supply ❸. If your panel will house only telephone and cable TV, you may not need a power supply, but routers require line voltage, so installing a power supply module will provide future flexibility. Drop the power-supply unit into the opening, run Romex cable to it, and screw the module to the panel housing ❹.

Install the final locknut on the utility conduit ❺. The conduit may run outside to the utility pole or to an underground pull box. Depending on your utility's main point of entry (MPOE) or demarcation, you may need an exterior box to which the utility will run the service. Or the utility may install the MPOE inside, in the distribution panel.

WARNING

Plan your interior layout before removing panel knockouts. That way you know where and how you want your cables to come in and you don't have to go back to reroute cables and seal unused knockouts.

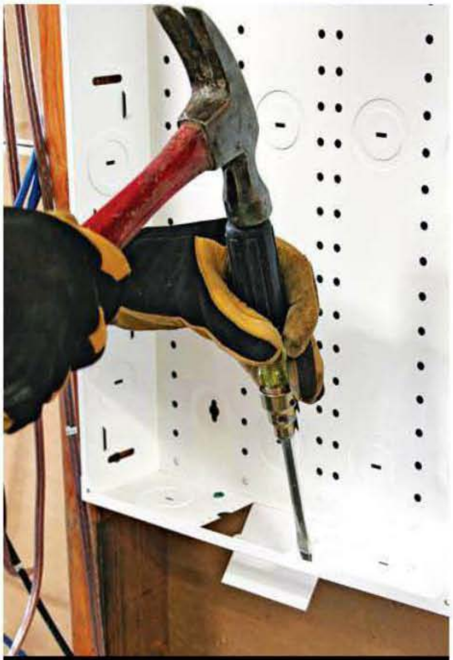
3 Remove a power-supply knockout.

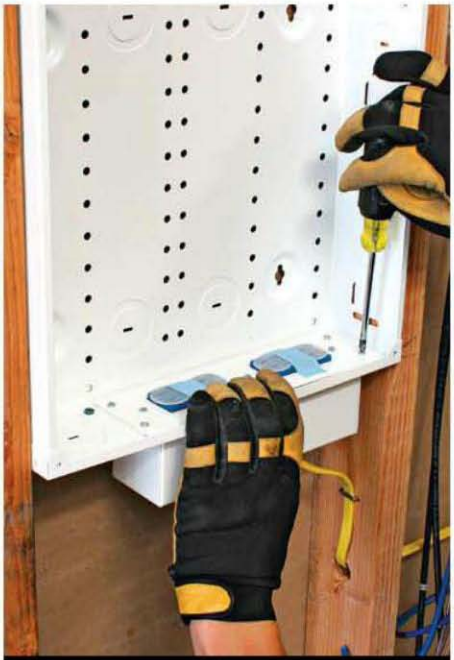
4 Attach a power-supply module to the panel.

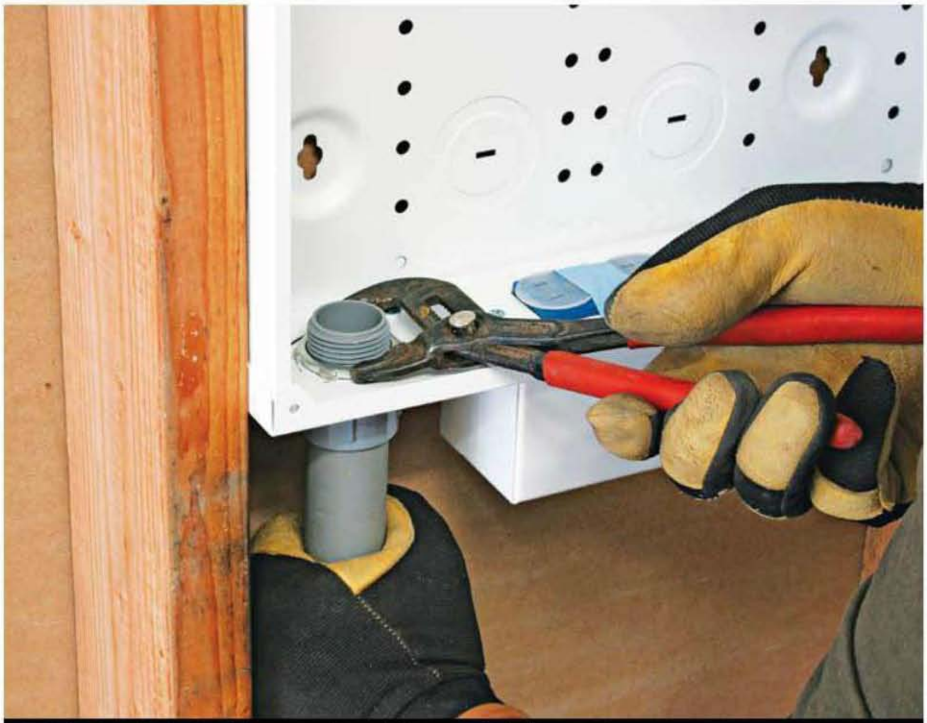
5 Tighten the locknut to secure the utility conduit to the panel.

A TYPICAL MULTIMEDIA PANEL

From this one multimedia panel (its outgoing low-voltage wiring has been completed), cables distribute video, sound, telephone, and data signals throughout the house. In this panel, the incoming utility and cable services are not yet installed—they will be pulled in through the conduit at the lower left.

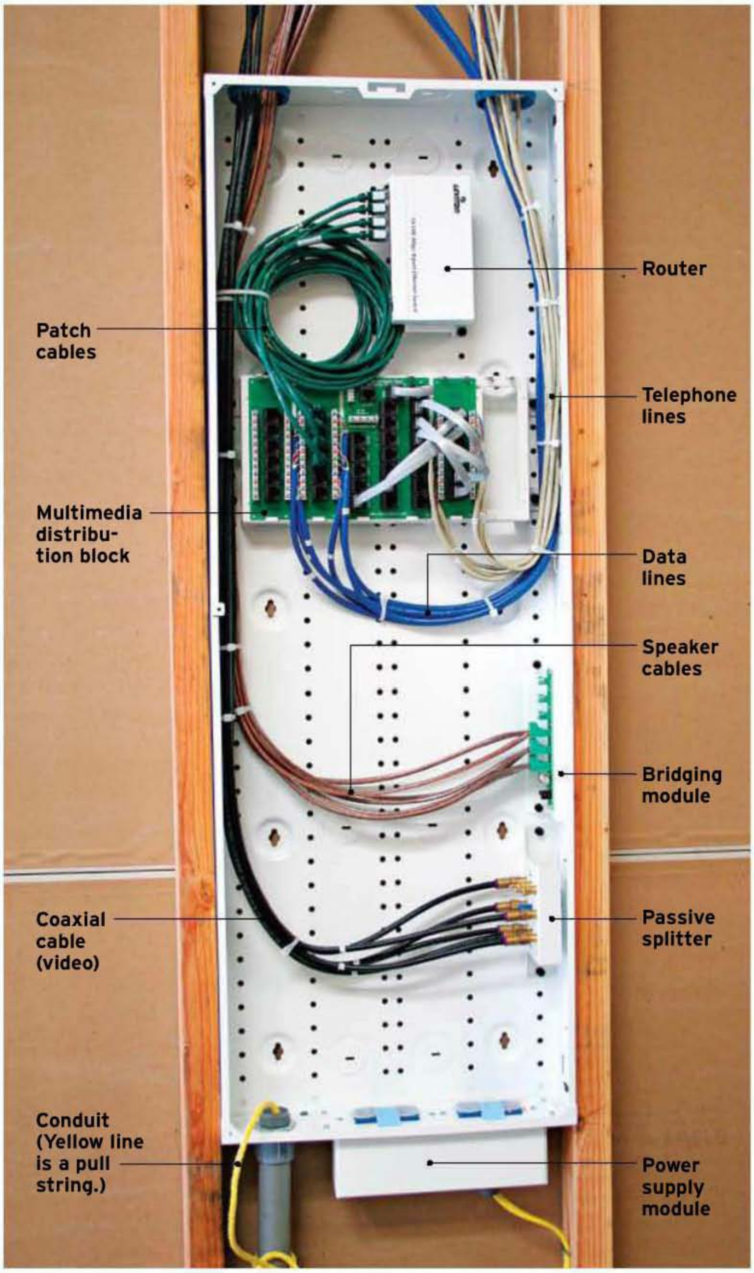
Router
Patch cables
Telephone lines
Multimedia distribution block
Data lines
Speaker cables
Bridging module
Coaxial cable (video)
Passive splitter
Conduit (Yellow line is a pull string.)
Power supply module

Panel basics A *router* distributes digital subscriber line (DSL) or broadband signals from a *modem* (not shown) throughout the house so you can have several people accessing the Internet at one time. *Patch cords* make connections from the router to jacks on the *data/phone distribution* block. The *multimedia distribution* block is a prebuilt unit: One side has banks of jacks and punch-down blocks that distribute data signals to outlets; its other side distributes telephone lines. A *passive splitter* distributes video signals from a cable company throughout the house; an *amplified splitter* boosts the video signal before distributing it. A *bridging module* distributes sound signals to speakers around the house. The *power supply module* is fed with 120v house current; it has receptacles that router or amplified splitter cords (not shown) can plug into.

RUNNING CABLES TO A MEDIA DISTRIBUTION PANEL

There's no one right way to run multimedia cable to a multimedia distribution panel. In new construction, electricians typically start from the distribution panel and pull cable outward. But remodel wiring is rarely predictable because you can't see obstructions hidden in walls and floors.

Before pulling cable into the panel, snap plastic bushings into the panel knockouts so sharp edges won't chafe the cable sheathing ❶. As you pull cable into the panel, loop it gently so that you don't crimp it ❷. If you have a lot of cables entering a panel, roughly divide them between two or more knockouts so the box will look neater.

Install prewired data and phone distribution boards and other modules. In photo ❸, the two columns on the left are jacks with 110-type punchdowns to which you'll attach data cables. Incoming phone lines terminate in the center of the module, which are then distributed to phone jacks throughout the house, permitting multiple lines at different locations.

As you route cable through the panel, tie-wrap cable bundles to free up workspace and enable you to see connections easily ❹. Labeling both ends of every cable is also essential, so you'll know which cable is which, should you need to repair or modify the multimedia wiring ❺. Finally, create a numbered house map to show the cable locations.

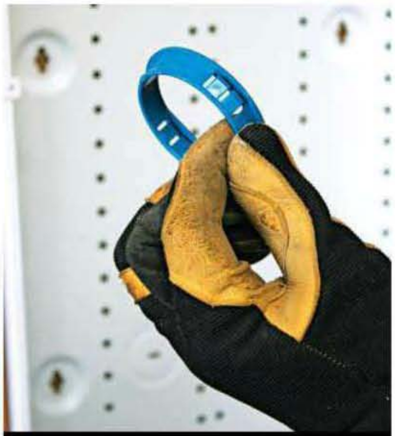

1 Use snap-in bushings to protect cables.

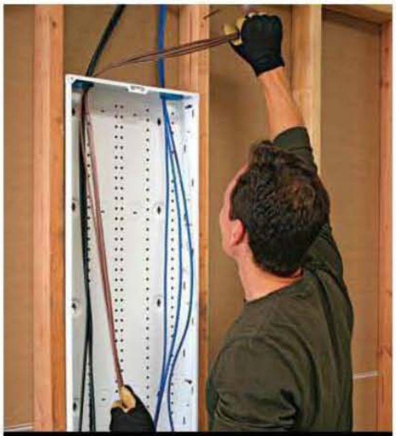

2 Loop the cable to avoid crimping it.

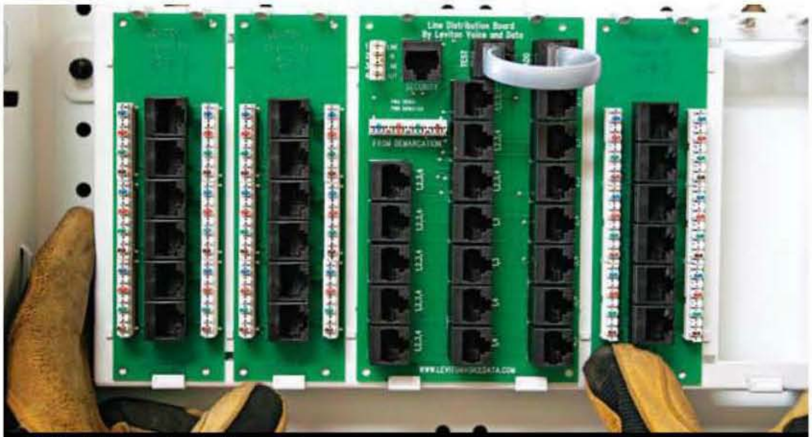

3 Install modules in the panel.

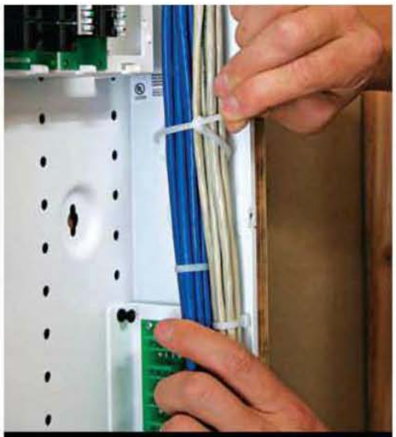

4 Bundle the cables to conserve space.

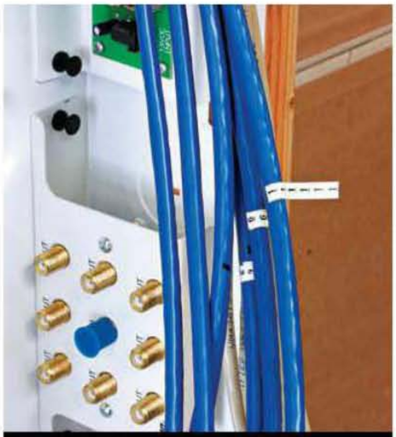

5 Label the cables at both panel and outlets.

CONNECTING CABLES AT THE PANEL

Because different types of cables converge at the distribution panel, there are, naturally, different connectors used to attach each one.

Speaker wires are typically stranded. Strip the ends as indicated by the panel maker—¼ in. is typical—insert them into the connectors' screw-down terminals and tighten them securely ①. Some panels will terminate speaker wires in a banana jack, as shown in the photo on p. 122, here, a plastic jack snaps to a bridging module ②.

Using a combination stripping and crimping tool, strip the coaxial cable that carries data or phone signals and crimp an F-connector into the cable ③. Plug coaxial cables into the passive splitter and tighten them snug ④. (*Note:* The capped blue terminal is for input from the cable company.)

Strip approximately 3 in. of sheathing from the ends of Cat 5e and Cat 6 cable (data and phone); separate the stranded wire pairs; and, using a punchdown tool, press the individual wires onto the insulation displacement connector (IDC) prongs on the data and phone board ⑤. The punchdown tool also removes excess wire. The blue cables in the foreground are data cables; the beige cables are phone lines ⑥. Finally, plug in patch cords that connect router input to jacks on the data and phone board ⑦.

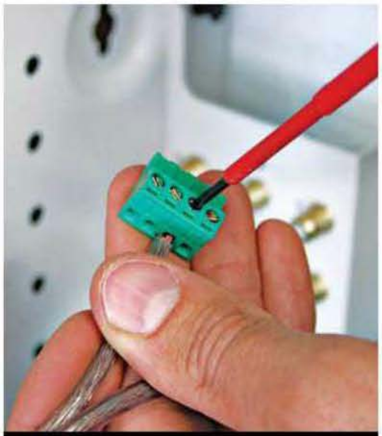

1 **Screw down the speaker wires in the terminal.**

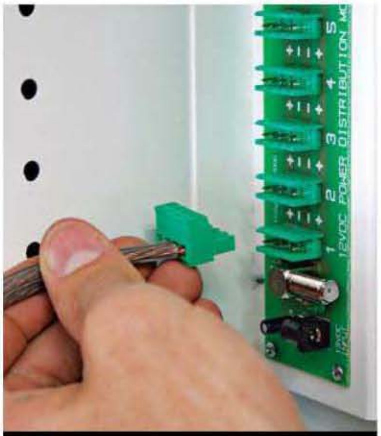

2 **Insert the speaker jack into the bridging module.**

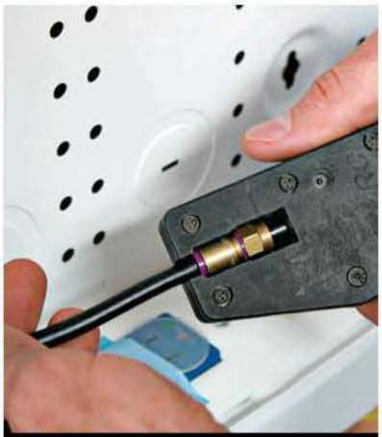

3 **Strip and crimp the coaxial cable.**

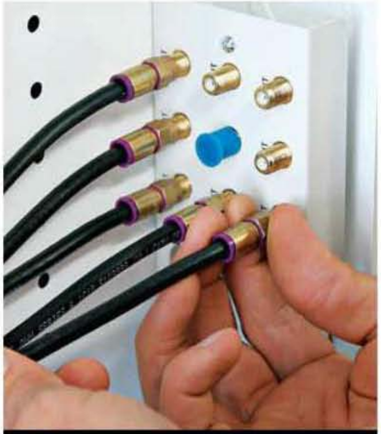

4 **Plug the coaxial cables onto the splitter terminals.**

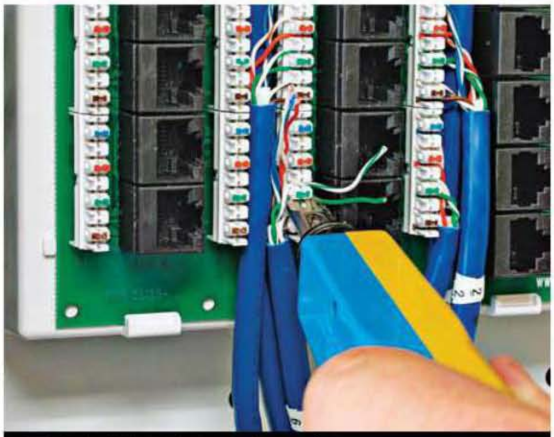

5 **Install the data and phone wires.**

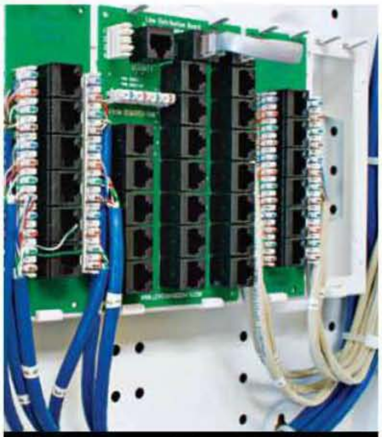

6 **Once connected, group and strap the cables to keep the panel neat.**

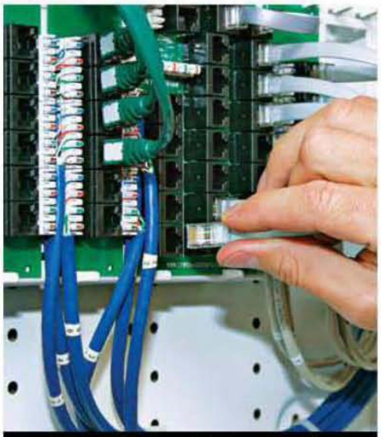

7 **Use patch cords to connect the router to the data.**

LOCATING THE MULTIMEDIA OUTLET

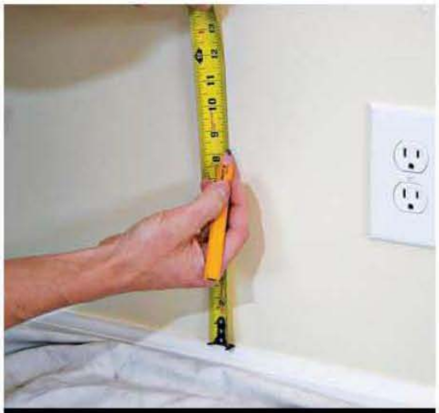

1 Align the new outlet with a nearby receptacle.

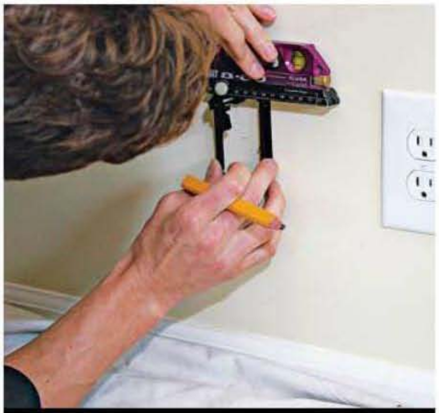

2 Level the cut-in ring and align it with nearby receptacles.

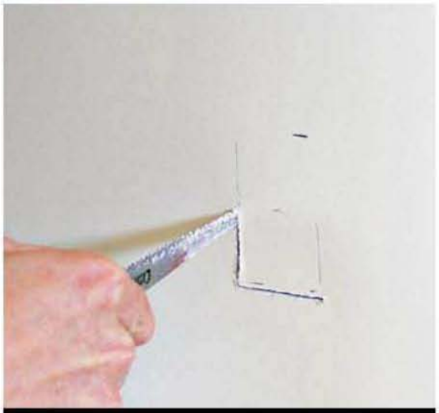

3 Cut on the waste side of the line.

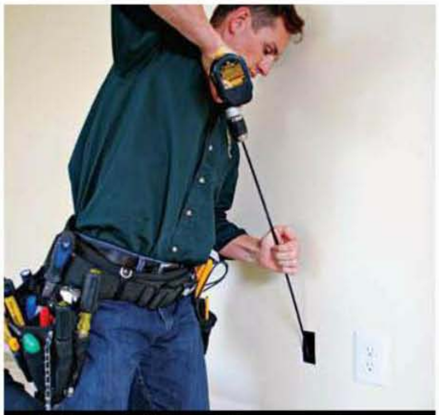

4 Use a flexibit to drill access holes for the cables.

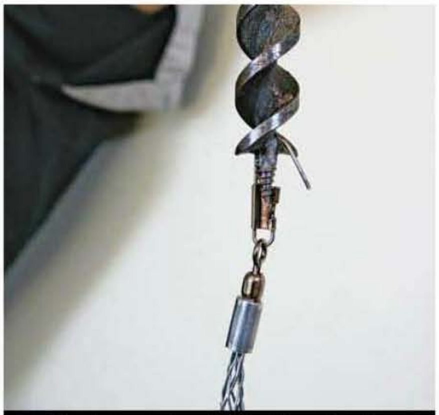

5 Hook the kellum to the bit and pull the cable to the outlet.

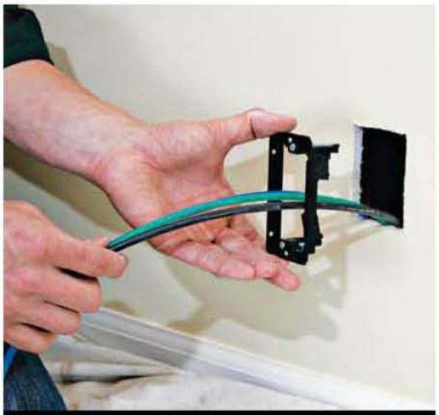

6 Route the cables through the cut-in ring and install it on the wall.

In general, electricians run one data, one phone, and one video cable to each outlet. In the example here, however, specs dictated two data lines (green and blue cables) and a coaxial cable for video.

When locating the cut-in ring for a new outlet set alongside an existing receptacle, always measure from the center of the existing box, so the center screws on the cover plates will line up ①. The two types of cover plates may be different sizes, so lining up their tops or bottoms won't look good. Level the cut-in ring and trace its outline on the wall. (The template shown in the top photo on p. 123 is another option.) Don't try to eyeball level: If you're even slightly off, the outlet will look cockeyed ②.

Use a drywall saw (also called a jab saw) to cut out the opening. To start the cut accurately, strike the heel of your hand against the saw handle. To avoid cutting too large a hole, cut on the waste side of the outline ③.

To bring cables from the distribution panel, drill holes into the wall plate using a flexibit ④. Wear gloves or use a drilling guide to protect your hands while guiding the bit. Flexibits have a hole in the point. After the bit emerges through the wall plate, have someone tape the cables together, slide a swivel kellum over the taped wires, and hook the kellum to the bit ⑤. The kellum swivels, so the wires won't twist up as you reverse the drill and pull them back up through the hole. Install the cut-in ring to provide a mounting surface for the outlet plate ⑥.

→ **For more on running wires, see pp. 188–197.**

INSTALLING CABLE CONNECTORS

In the next few pages, we'll discuss the installation of several different types of cable connectors: F-connectors used on coaxial cable (video), as well as two popular systems (from Leviton® and Panduit®) that connect to Cat 3 cable (phone), Cat 5e and Cat 6 cables (telephone or data). Each connector maker specifies tools and methods of assembly, so follow its installation instructions closely.

Use a combination stripper-crimper to strip insulation off the end of the coaxial cable. Stick the cable in the end of the tool, spin it, and peel off both types of insulation from the outer sheathing ❶. The tool leaves about ½ in. of bare copper and ¼ in. of white insulation with the shielding on it. Slide an F-connector over the stripped end of the coaxial ❷. Place an F-connector in the crimping bay, insert the stripped cable, and squeeze the tool to crimp the connector tight to the cable ❸. (*Note:* The F-connector shown here is a watertight variety.)

TRADE SECRET

It's not necessary to house multimedia connections in covered junction boxes (required for high-voltage splices). Install flush-mounted cut-in rings to provide a secure device to which you can attach outlet plates and insert jacks. Whether you're mounting doorbell chimes or a multimedia wall outlet, take the time to level and plumb the device and, if there's another outlet nearby, to align the height of the new outlet.

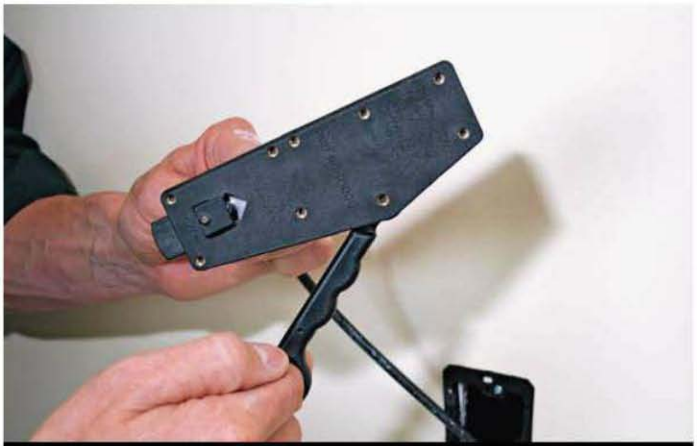

1 Spin the stripper-crimper on the coaxial cable.

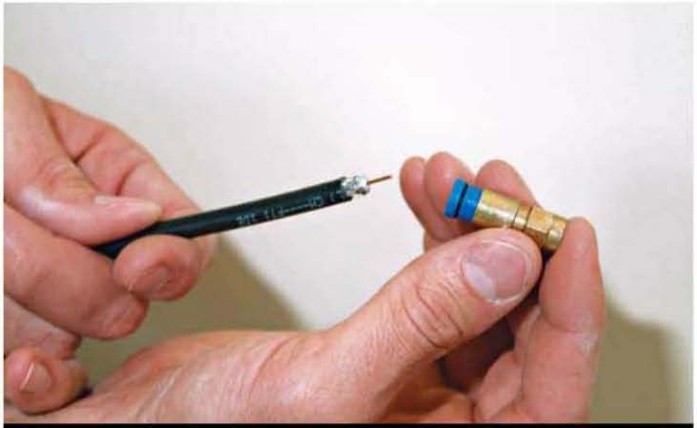

2 Use an F-connector to cap the coaxial cable.

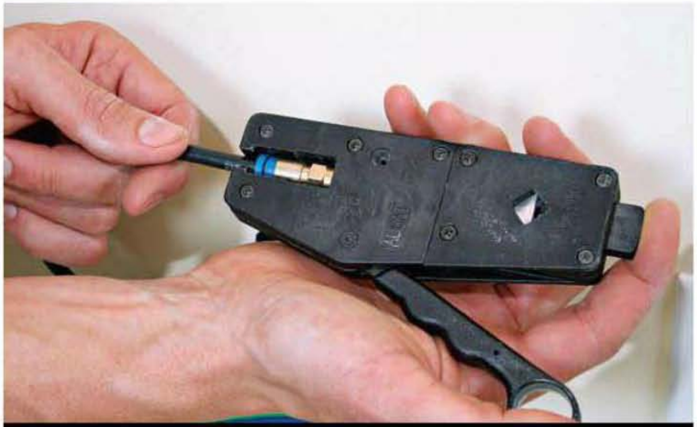

3 Crimp the F-connector to the cable.

CONNECTING PANDUIT JACKS

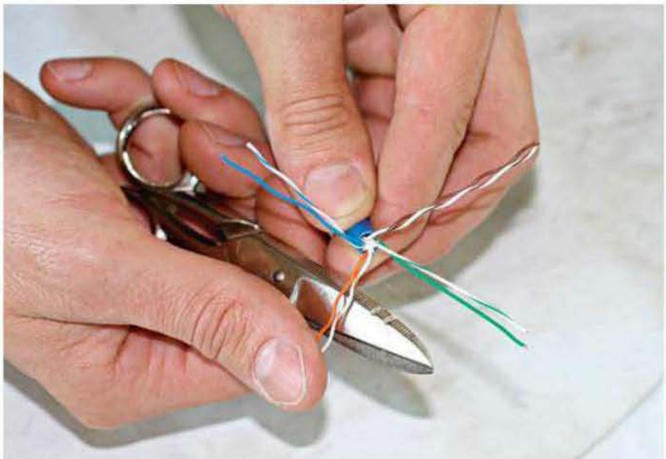

1 Separate and straighten the wire pairs.

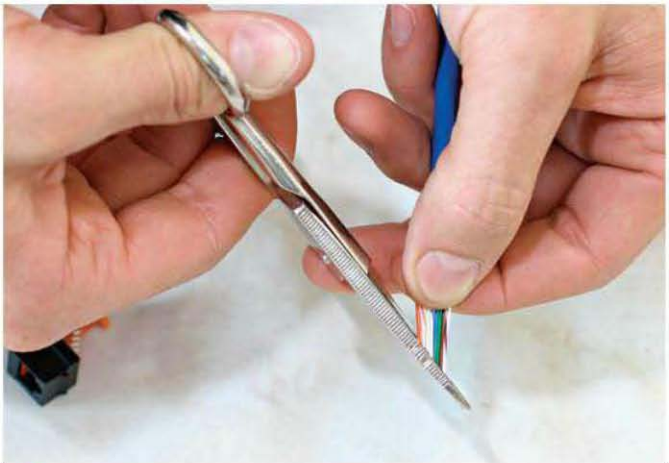

2 Cut the wire ends diagonally.

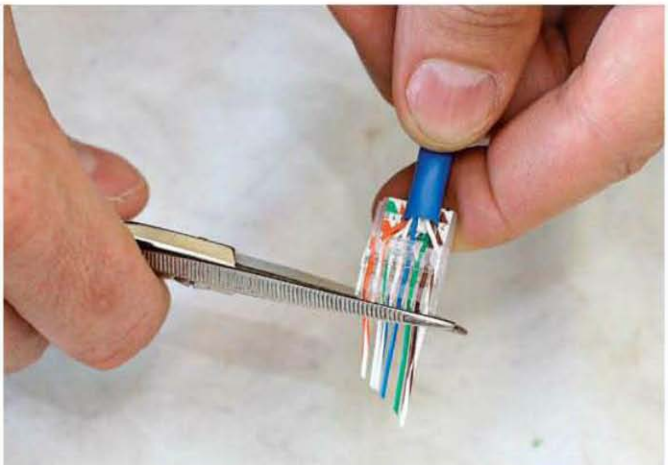

4 Trim the wire ends square.

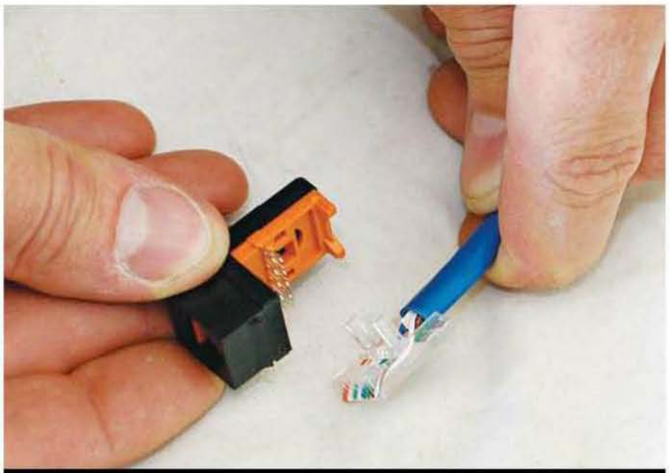

5 Pair a Panduit jack housing with a wired cap.

Connecting phone or data cable jacks to a multimedia outlet requires a few steps. Start by reviewing cable stripping (see p. 123). Cat 5e and Cat 6 cable are UTP cable, so after stripping about 3 in. of sheathing, separate the wire pairs before attaching them to a jack. If you use Panduit jacks (often referred to as *Pan jacks*), you'll also need to unwind (untwist) the wires. (Cat 3 phone wires are straight and so do not need untwisting.)

Use stripping scissors to unwind the individual cable wires before attaching them to a Pan jack ❶. Using light pressure, pull the wires across a scissor blade to straighten them. Next, slide the wires into a plastic cap, which is color coded to indicate where the wires go. Flatten the wires and cut their ends diagonally so they'll slide easily into the cap ❷. Push the wires all the way into the cap ❸, then trim the wire ends straight across ❹.

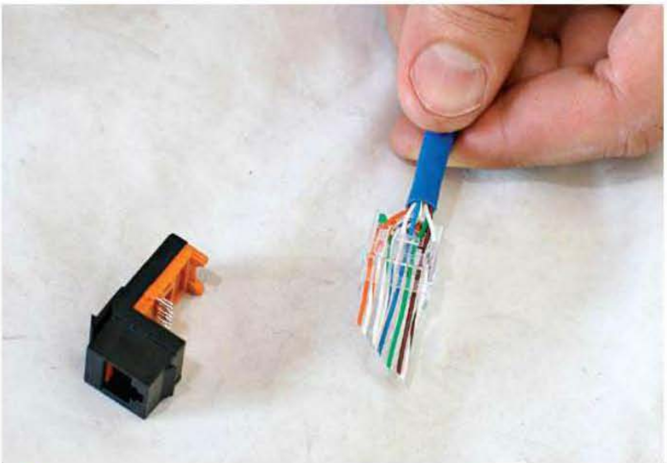

3 Push the wires in all the way.

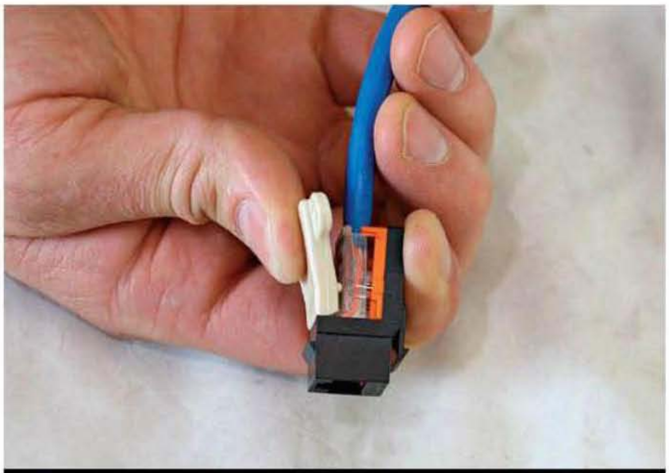

6 Snap the housing and cap together.

The wired cap snaps into a jack housing, forcing the wires into V-shaped IDC prongs that slice the wire insulation to make the electrical contact ❺. Place the wired cap onto the jack housing and use the small plastic lever provided to snap the assembly together ❻. Panduit jacks are reliable because they make secure connections.

INSTALLING LEVITON JACKS

The Leviton jack system uses the 110 punchdown tool required to punch down wires in the distribution panel (see photo 5 on p. 131). But the system takes practice to avoid loose wires, so amateur electricians might get more predictable results using Panduit jacks.

In the Leviton system, there's no need to untwist wires. Separate the wire pairs and punch them directly into the IDC prongs built into the jack. Use the plastic disc provided to back the jack as you punch down ❶. The punchdown tool comes with 110 and 66 blade sizes; each blade has a side that punches the wire down and a side that cuts off excess wire. For best results, work from back to front to avoid disturbing wires that are already down. Once you've connected the wires, snap on the jack's trim cap, which keeps wires in place and relieves strain on the connection ❷.

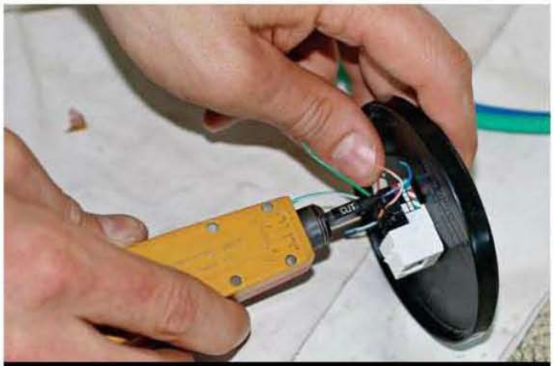

1 Punch down the wires to the Leviton jacks.

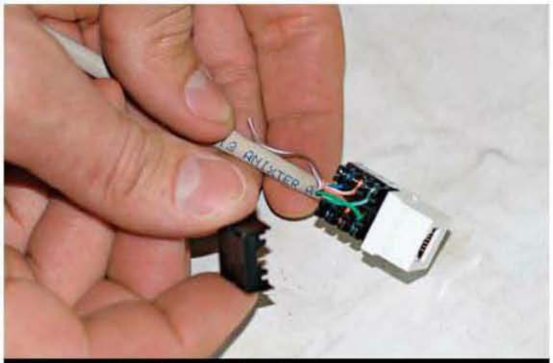

2 Snap the trim cap into place.

ATTACHING JACKS TO AN OUTLET PLATE

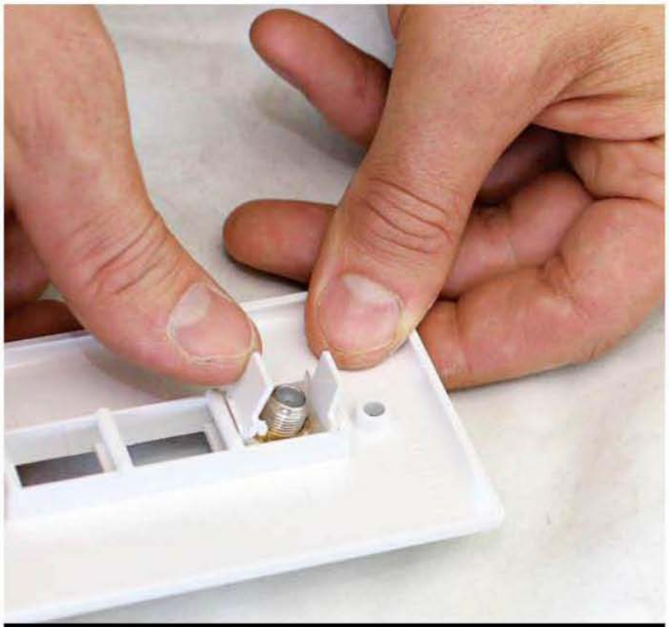

1 Snap on the screw-in coupling for coaxial cable.

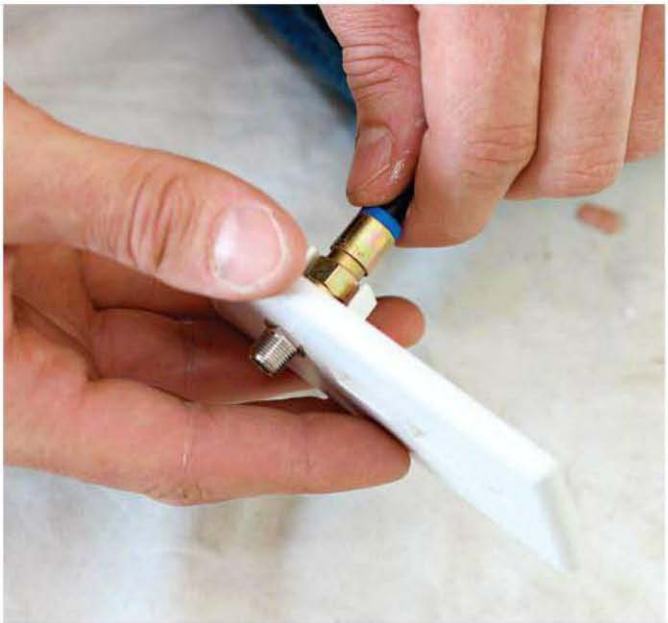

2 Screw an F-connector to the snap-in terminal.

In the project shown here, we're installing both Leviton and the Panduit jacks to show that despite variations in jack wiring, both snap into the most common type of face plate, the keystone style.

To install coaxial terminals, snap the threaded coupling into the face plate ❶. The coaxial cable's F-connector screws to the coupling, creating a clean termination instead of having cable hanging out of the wall ❷. Snap a RJ-45 (eight-pin) Leviton jack into the face plate ❸. Note the color coding on the side of the jack to indicate the order of wires you punched down earlier. Finally, snap a RJ-45 Panduit jack into the remaining port in the keystone plate ❹.

Carefully feed the cable into the wall, hold the face plate flush to the wall, and screw it to the cut-in ring ❺. The outlet shown uses phone and data jacks that are different colors so that users can quickly differentiate which jack is which. This differentiation is not an issue in residences, but it's imperative in business installations.

Faceplate options

There are many types of faceplates on the market for mounting telephone, data, coaxial, and specialty jacks. They come designed to match the adjacent electrical outlet (for example, duplex or Decora® style) and in various colors. Although it is beyond the scope of this book, there are also faceplates and adapters for an array of applications such as video (HDMI, S-video, component video, VGA), sound (binding posts, banana), and fiberoptics.

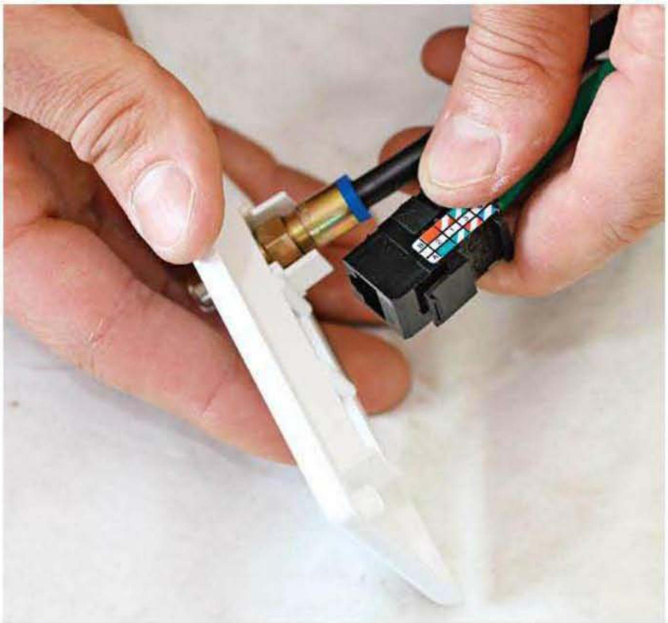

3 Snap in the Leviton data jack.

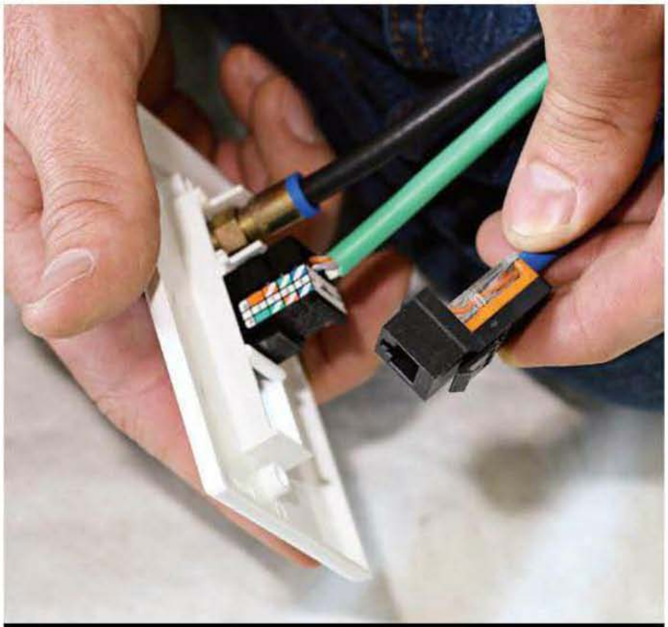

4 Snap in the Panduit jack.

5 Screw the face plate into place.

FANS

RESIDENTIAL FANS CAN BE divided into two groups. The first group might be better called exhaust or vent fans, because their primary function is to remove moist or odor-laden air from living spaces, route it through ducts, and expel it outside. Commonly installed bath fans and range hoods both fall into this category. By removing excessive interior moisture, these fans help forestall mold and other unhealthy conditions.

The second group, primarily ceiling fans, circulate still, summer air and thus increase comfort by evaporating moisture on your skin. In cold climates, these fans push warm air down from the ceiling where it collects, so it can once again warm the bodies below. Whole-house fans, installed in attics in warm regions, also circulate hot air, but because they typically push it out roof or gable-end vents, whole-house fans are more like exhaust fans without ducts.

WIRING TIPS

Before you begin, p. 140

BATHROOM VENT FANS

Choosing a bathroom fan, p. 141

Locating a bathroom fan, p. 142

Installing the roof vent, p. 143

Mounting the fan box, p. 145

Retrofitting a bathroom fan, p. 146

Wiring a bathroom fan, p. 147

RANGE HOODS

Mounting & ducting a range hood, p. 149

Wiring the range hood, p. 150

CEILING FANS

Installing the ceiling fan box, p. 152

Mounting the ceiling fan, p. 154

BEFORE YOU BEGIN

There are two important tips to note before wiring any type of fan. First, check out the wiring schematic that comes with the unit Ⓐ. In most cases, the schematic is affixed either to the fan housing or to the backside of a cover. Or it may be included in the installation instructions. You'll also find essential information such as the fan's rating, expressed in amperes.

WARNING

When installing or replacing a fan of any type, always turn off the electrical power to that location. Then test with a voltage tester to be sure the power is off.

Second, most units have an integral junction box, like the one seen in photo Ⓑ. The junction box contains wire leads that you'll splice to the incoming house wiring. All metal boxes—including fan housings—must be grounded, so if the fan maker doesn't include a grounding pigtail to the housing, add one.

TRADE SECRET

Never assume that all the parts you need for assembly and installation are in the box. Check the contents before you start by comparing the parts in the box to the inventory list included in the owner's manual. Smaller parts, such as screws, are typically packaged in clear plastic so you can count them without having to open the packet—a good idea because they're easy to lose.

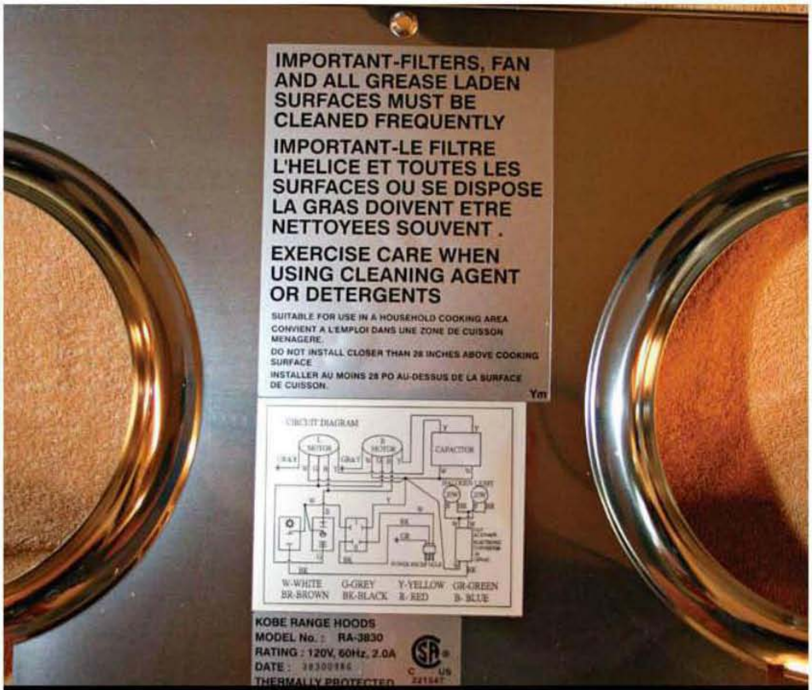
IMPORTANT-FILTERS, FAN AND ALL GREASE LADEN SURFACES MUST BE CLEANED FREQUENTLY
IMPORTANT-LE FILTRE L'HELICE ET TOUTES LES SURFACES OU SE DISPOSE LA GRAS DOIVENT ETRE NETTOYEES SOUVENT .
EXERCISE CARE WHEN USING CLEANING AGENT OR DETERGENTS
SUITABLE FOR USE IN A HOUSEHOLD COOKING AREA
CONVIENT A L'EMPLOI DANS UNE ZONE DE CUISSON MENAGERE.
DO NOT INSTALL CLOSER THAN 28 INCHES ABOVE COOKING SURFACE
INSTALLER AU MOINS 28 PO AU-DESSUS DE LA SURFACE DE CUISSON.
Ym
CIRCUIT DIAGRAM
MOTOR
MOTOR
CAPACITOR
W-WHITE G-GREY Y-YELLOW GR-GREEN
BR-BROWN BK-BLACK R-RED B-BLUE
KOBE RANGE HOODS
MODEL No. : RA-3830
RATING : 120V, 60Hz, 2.0A
DATE :
THERMALLY PROTECTED
C US

A The wiring diagram and fan ratings are found on the underside of this range hood.

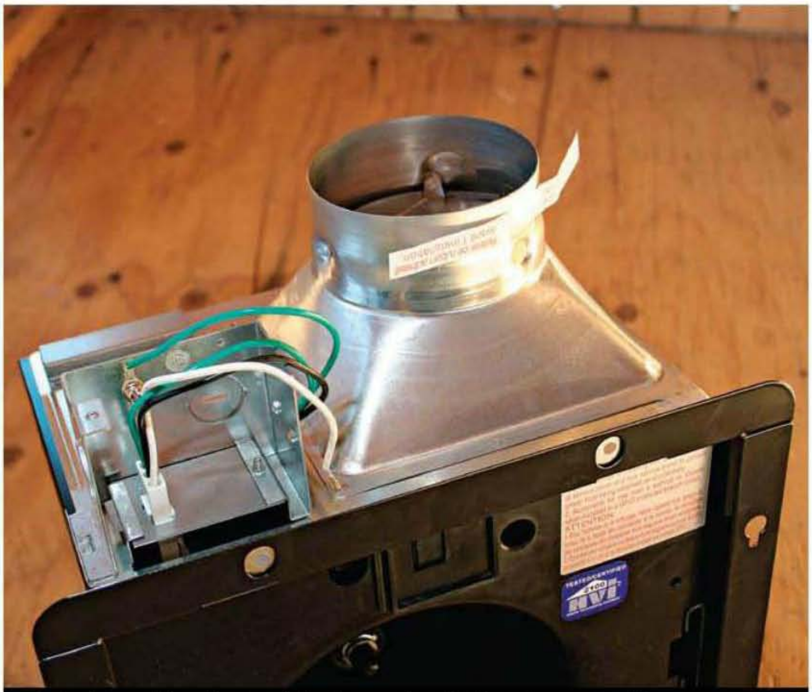

B An integral junction box is in the housing of this bathroom vent fan. The second green wire grounds the metal housing.

CHOOSING A BATHROOM FAN

Bathroom fans are increasingly powerful, quiet, and available with numerous bells and whistles. When considering all the extra features, however, remember that the primary function of a bathroom fan is to remove moisture.

There are complex formulas for sizing bath fans, but a good rule of thumb is 1 cfm (cubic foot per minute) per square foot for bathrooms 100 sq. ft. or smaller. For bathrooms larger than 100 sq. ft., allot 50 cfm for each fixture (toilet, lavatory, shower) and 100 cfm for hot tubs.

Also, get a quiet fan: a rating of 3 sones to 4 sones is tolerable, 1 sone is very quiet. And remote inline fans, typically installed in attics some distance above bathrooms, are quieter still. Consequently, you can install a larger fan inline and still have a quiet bathroom.

Next, consider switches. Fans usually need to continue venting after you leave the shower or use the toilet, so get an electronic switch with an integral timer so that the fan continues running after the light is turned off. If there's no integral timer on the switch, you can run the light and the fan on separate switches. You can also connect the fan to a *humidistat*, which is a moisture sensor that turns the fan off once a preset moisture level is attained.

WARNING

Never install a rheostat—commonly called a dimmer switch—to control an exhaust fan. Fan motors are designed to run on a fixed voltage, and if you reduce the voltage by using a rheostat, you may burn out the fan motor.

FAN COVERS

Most exhaust fans come with plastic covers, but they can be replaced with metal plates. Here, an antique wall register was plated to match the other fixtures in the room.

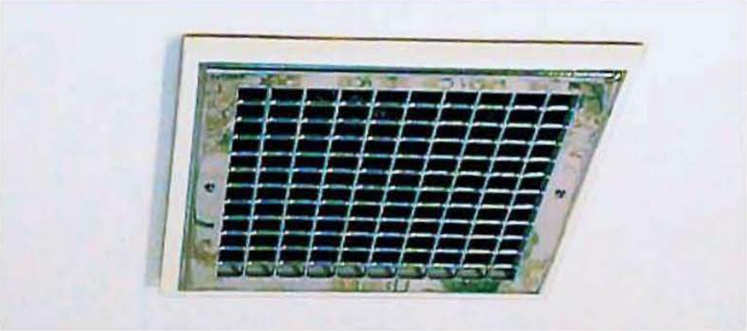

REMOTE INLINE FANS

Even a well-made bathroom fan will be relatively noisy if the motor is 2 ft. from your head in the bathroom ceiling. But if you install the fan some distance from the bathroom, you'll reduce the noise considerably. That remote location may mean that you have longer duct and wiring runs, but routing them is rarely a problem. In fact, with a large-enough fan motor and a duct Y-connector, you can vent two bathrooms with one fan. Because longer duct runs can mean greater air resistance, consider installing rigid-metal or polyvinyl chloride (PVC) ducts, whose smooth surfaces offer less resistance, rather than flexible metal ducts. Alternatively, you could oversize the fan slightly. Better fan makers, such as Fantech® and American Aldes™, offer acoustically insulated cases to deaden sound further.

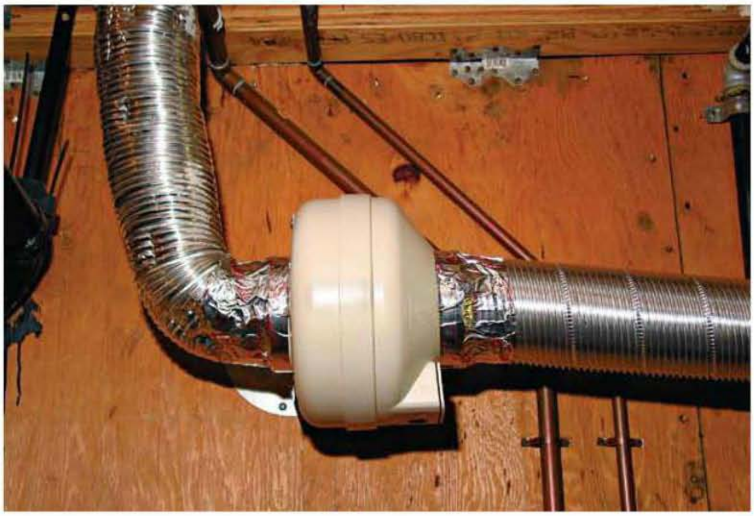

Remote inline fans require longer ducting, but they're quieter because they're located farther away from the bathroom.

LOCATING A BATHROOM FAN

There are three primary considerations when locating a bath fan. First, place the fan where it can expel moisture effectively—ideally, near the shower, where most of the moisture is generated.

Second, locate the fan so that its duct run is as short as possible and minimizes cutouts in the blocking or framing members—especially rafters. Vent the ductwork from the fan out the roof or through a gable-end wall. Avoid sidewall vents because moist air expelled by them could be drawn up into the attic by soffit vents in roof overhangs.

Third, locate roof vents away from problem areas such as skylight and valley flashing. Water usually dams up on the uphill side of a skylight, creating a leak-prone area that must be carefully flashed. Typically, skylight flashing consists of two pieces: a base flashing and a counterflashing that goes over it. Locate a roof vent near skylight flashing and you're inviting trouble. Valley flashing, on the other hand, may consist of a single broad piece of metal or elastomeric material folded up the middle. But because this flashing is located where roof planes converge, it channels an enormous amount of water during rainy seasons. So keep things simple: Locate roof vents away from obstructions in the roof or concentrations of water. Don't put a roof vent near an operable window, either.

A BATHROOM FAN

To keep moisture from leaking into the attic, apply silicon caulk between the fan box flanges and the ceiling drywall. Use metal foil tape to ensure airtight joints where ducting attaches to fan and vent takeoffs.

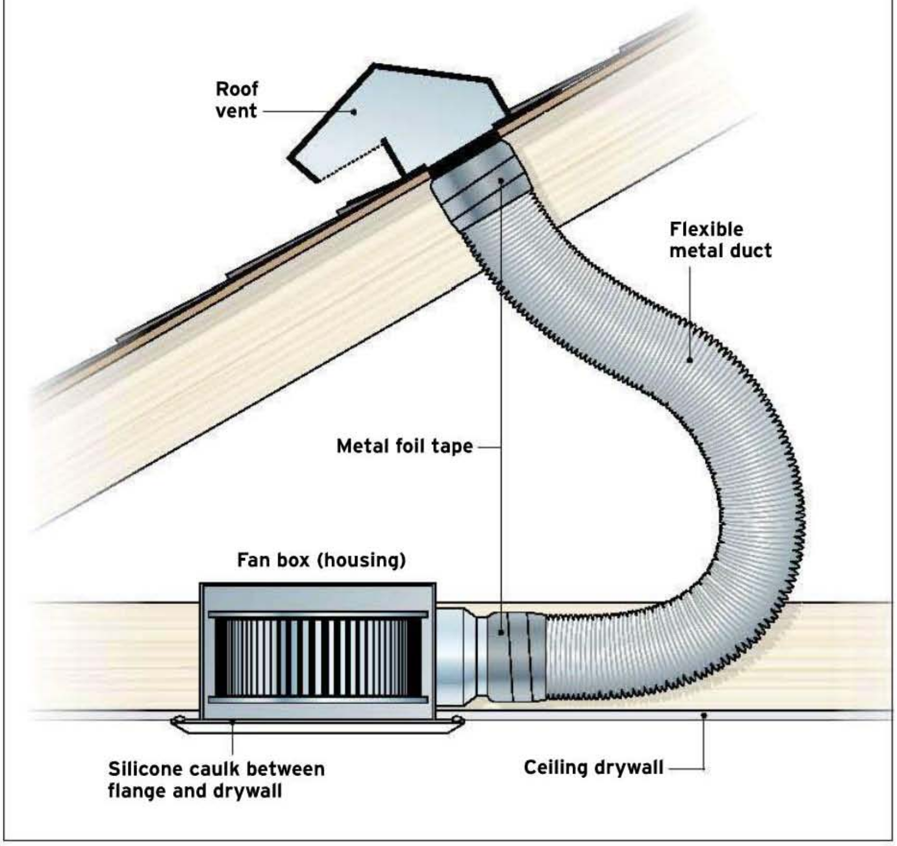

TRADE SECRET

Before you cut a hole in the ceiling, be sure there are no obstructions along the way. Tentatively locate the fan and use a 1/4 in. extension bit to drill along the proposed duct path. You'll also want to drill an exploratory hole up through the roofing.

INSTALLING THE ROOF VENT

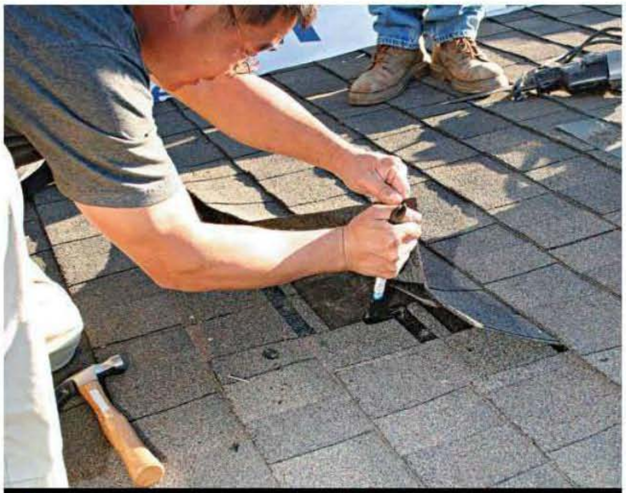
1 Lift up on the shingle to remove the roofing nails.

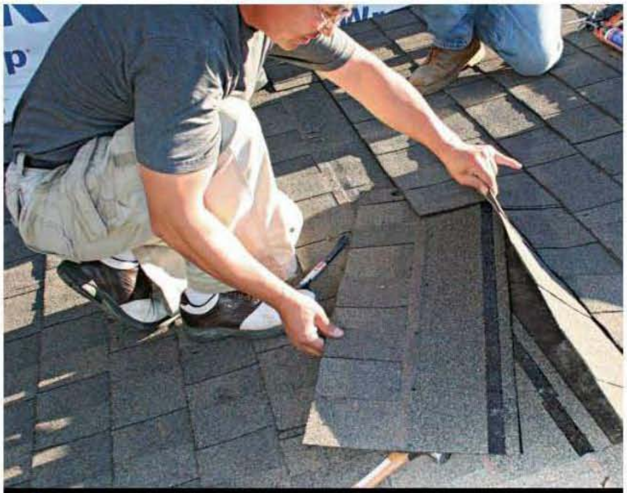
2 Remove the shingles over the entire section.

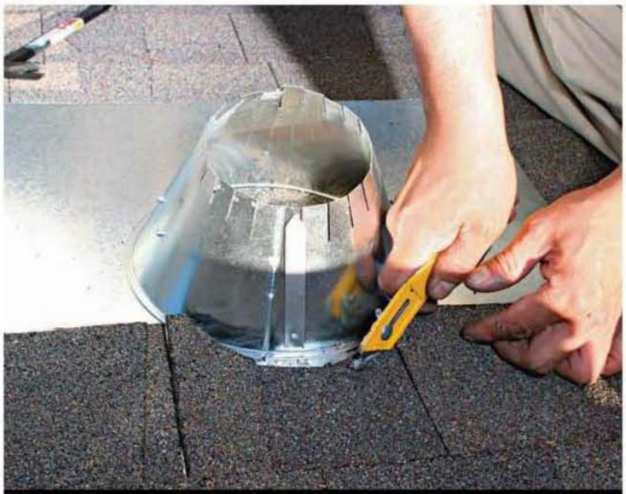
3 With the vent in place, trim the shingles to fit the upper arc of the vent.

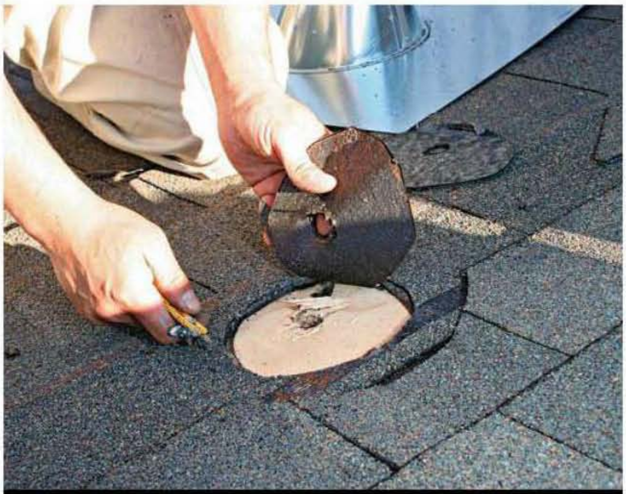
4 Use a utility knife to cut the shape of the vent hole through the roofing paper.

In the installation shown here, there was enough clearance around the exploratory hole and there happened to be a roofer on site, so the crew decided to install the roof vent first. (Roof vents vary; the model shown has a round stack and a weatherproof cap.)

Go onto the roof and find the ¼-in. exploratory hole drilled while locating the fan. The hole represents the center of the vent hole you'll need to cut.

Most ducting and roof-vent takeoffs are 4 in. diameter, so sketch that circle onto the roof. If the circle would cut into the tabs of any shingle—roughly the bottom half of a shingle strip—use a shingle ripper or a cat's paw to remove the nails holding those shingles in place before cutting the vent hole ❶. Be gentle when removing shingles so you can reuse them ❷.

Slide the upper flange of the roof vent under the shingle courses above and use a utility knife to trim its arc onto the shingles above ❸, then cut out the vent's circle into any remaining shingles and the roofing paper ❹. Next, use a

>> >> >>

INSTALLING THE ROOF VENT (CONTINUED)

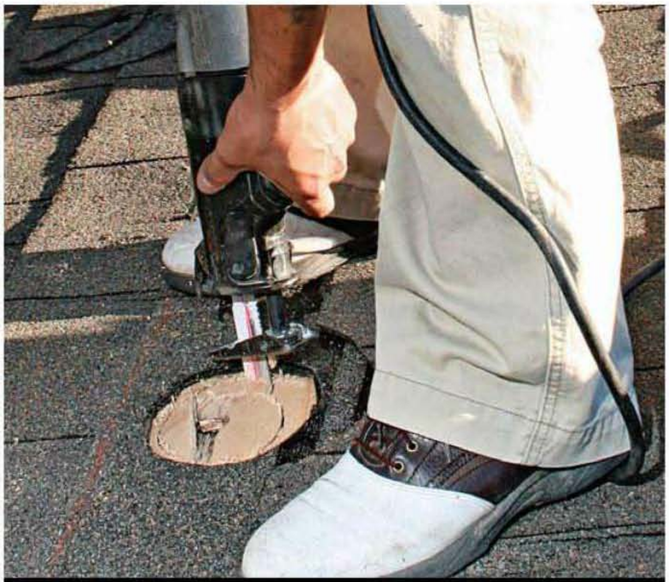

5 Cut the roof sheathing to allow for the vent.

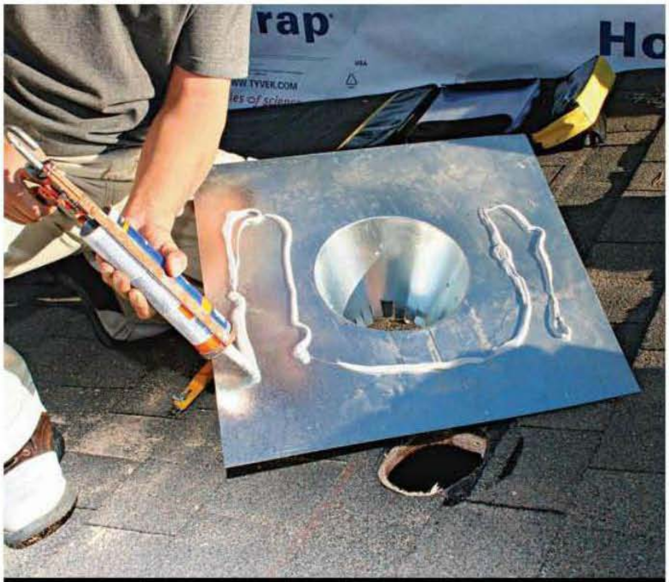

6 On low-pitch roofs, caulk the vents to help prevent leaks.

reciprocating saw to cut through the sheathing ❺.

If the roof pitch is 4:12 or greater, it usually isn't necessary to caulk the vent edges. Here, a 2:12 pitch required caulking to forestall leaks ❻. Carefully lift the shingle course above the vent and nail the two corners of the vent's upper flange into place ❼. Do not nail the lower corners of the vent: those nails would be exposed to weather and could leak. Slide in the surrounding shingles and renail them.

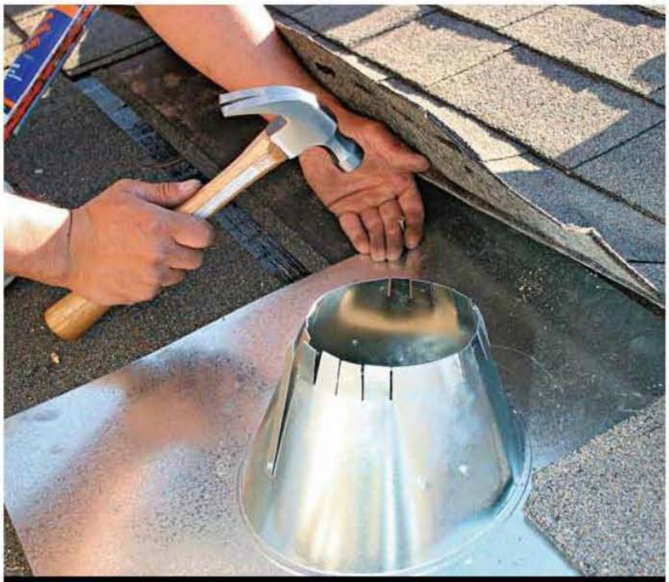

7 Nail only the top corners of the vent—shingles cover these nails.

TRADE SECRET

Keeping fan-expelled moisture out of attics and wall cavities is crucial, and the only way to do so is to create airtight connections: Caulk the fan housing to the ceiling and seal each duct joint with aluminum foil tape, not fabric duct tape.

MOUNTING THE FAN BOX

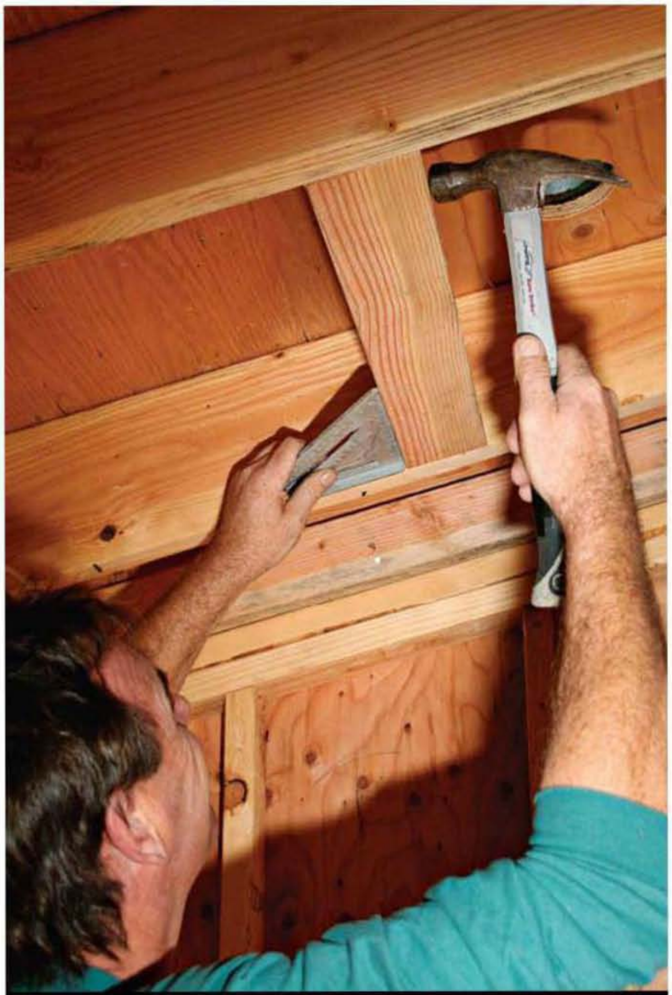

1 If needed, add blocking between the framing members.

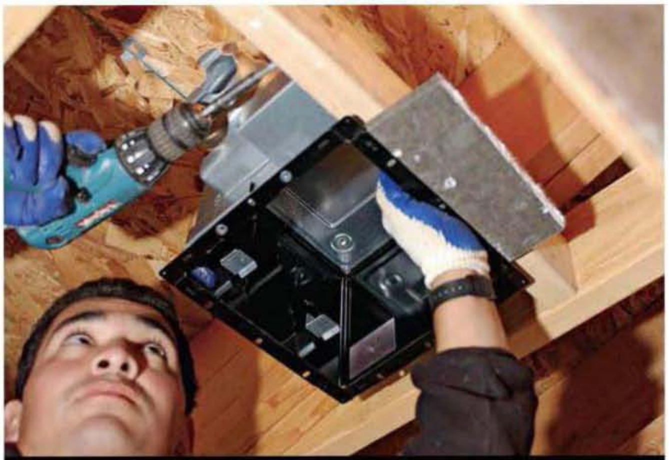

2 When positioning the fan, a scrap of drywall acts as a stand-in for the finished ceiling.

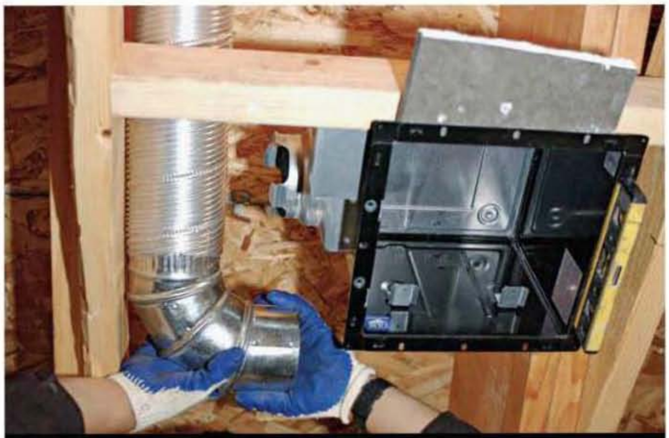

3 Slide on the flex duct and tape it in place.

If bathroom framing is exposed, mounting the fan is pretty straightforward. If you remove the fan assembly from the fan box, the box will be lighter and easier to hold in place one-handed while you use your other hand to screw the unit to the ceiling joists. Most fans have expandable brackets which extend between joists spaced 16 in. on-center (o.c.). But you should always screw at least one side of the fan box to a joist, to ensure that it's anchored securely. For ceiling joist (or rafter) spacing greater than 16 in. o.c., it's a good idea to add blocking ❶.

If the fan box flange mounts flush to the underside of the ceiling, use a piece of drywall scrap to gauge the depth of the unit relative to the finished ceiling ❷. Regardless of whether the box flange sits above or below the ceiling drywall, caulk the flange with polyurethane sealant to create an airtight seal between the two materials. If you removed the fan assembly earlier, reinstall it now.

Keep duct runs as short as possible to reduce air resistance. Slide the lower end of the flexible duct to the fan's exhaust port ❸ and seal the joint with metal duct tape, then attach the other end to the roof vent takeoff. Or, if you haven't yet cut the hole in the roof, hold the free end of the duct to the underside of the roof sheathing and trace its outline onto the surface.

RETROFITTING A BATHROOM FAN

If you are remodeling or installing a bath fan and the finished ceiling is already in place, begin by creating a cardboard template of the fan housing. Mark the approximate position of the fan by driving a screw or nail through the ceiling and then go up into the space above the bathroom and find the marker. If there is an insulated attic above, take along a dustpan to shovel loose insulation out of the way and be sure to wear a dust mask and gloves. After you've located the marker, place the fan template next to the nearest joist and trace around it. Most fan boxes mount to ceiling joists. If the fan box has an adjustable mounting bar, you have more latitude in placing the fan. Use a jigsaw or reciprocating saw to cut out the opening. To keep the drywall cutout from falling to the floor below, take a piece of scrap wood slightly longer than the cutout and screw it to the drywall.

Before placing the fan box on top of the ceiling drywall, caulk around the perimeter of the opening to create an airtight bond to the box flange ❶. Once the box is wired and the ceiling has been painted, install the trim piece to cover any gaps around the fan opening ❷.

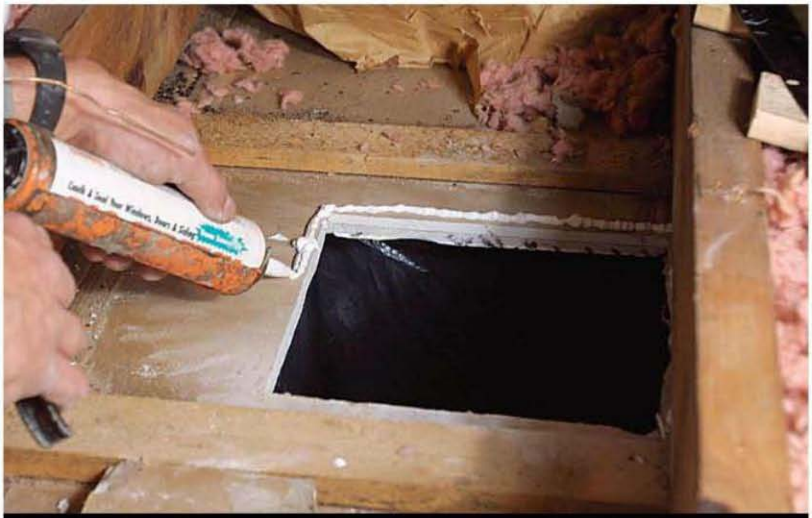

1 Apply a bead of caulk under the flanges before you set the fan box in place.

2 Install the trim pieces to cover the gaps between the fan and the ceiling cutout.

TRADE SECRET

If there's an existing ceiling light in the bathroom, a fan-and-light combination unit may simplify your remodel. If you want the fan and light to operate at the same time, you can reuse the two-wire cable that's presently controlling the light switch. But if you want to operate the fan and light separately—so that the fan continues running after the light is switched off—you'll have to install a length of 14/3 or 12/3 cable, as shown on p. 155.

WIRING A BATHROOM FAN

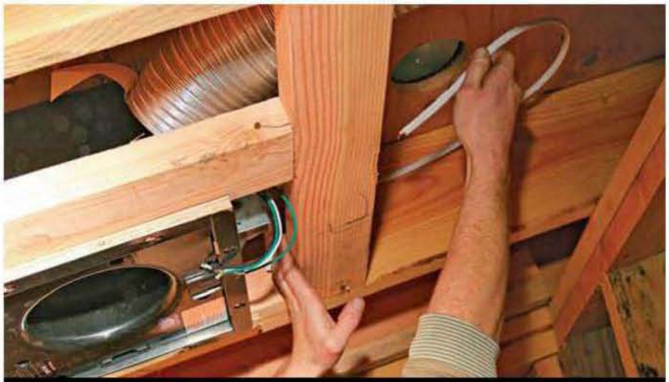

1 Staple cable no more than 1 ft. from the fan box.

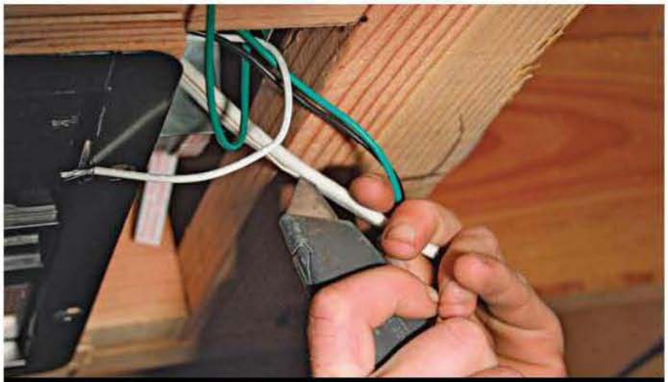

2 Strip the sheathing off the wires.

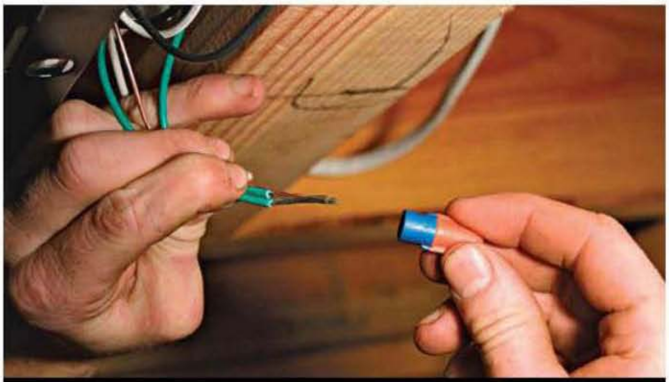

3 Splice the ground wires first.

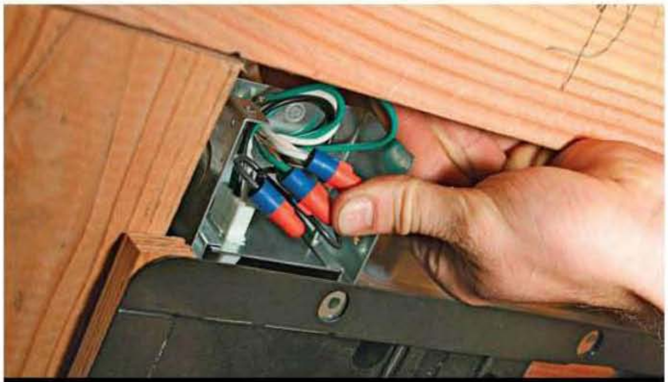

4 Fold the spliced wires into the junction box.

Before making any connections in the fan's junction box, make sure the power is off. Follow the wiring diagrams provided by the fan manufacturer. In general, it's easier if the incoming power runs through a switch box first; then you don't have to try to splice the switch legs in the fan junction box—junction boxes inside fan housings tend to be cramped. Bathroom fans should also have ground-fault circuit interrupter (GFCI) protection if installed in a shower whose ceilings are 8 ft. above the finish floor or lower.

If the duct space is tight, it's usually best to wire the fan box before installing the duct. When running cable to the fan box, allow a generous loop of cable, just in case. As with light fixtures and receptacles, staple the cable within 1 ft. of the fan box ❶. Insert a cable connector into a junction box knockout, feed the cable through it, and strip the cable sheathing ❷.

Using wire connectors, first splice the incoming ground wire to the fixture ground ❸. (If the fixture lead wires are stranded, extend them slightly beyond the solid wire so that the wire connector will engage them first.) Splice the neutral wires and then the hot wires. If the light and fan are wired separately, there will be two sets of hot wires.

Tug each wire group gently to be sure the splices are secure. When all groups are spliced, carefully fold the wires into the fan junction box and cover the box ❹.

TRADE SECRET

In tight spaces like a junction box, almost all the pros use a utility knife to remove sheathing from the wires.

RANGE HOODS

Choosing a Range Hood

All cooktops and stoves should be vented by a range hood. In addition to sucking up the smoke of a charred steak, range hoods exhaust airborne grease that might otherwise migrate to a cool corner and feed mold or adhere to woodwork and discolor its finish. Range hoods come in many configurations but basically there's a *hood* to collect smells and smoke, a *fan* to expel them, *ducts* to carry exhausts out, and *shrouds* and other trim pieces plus, of course, wires, switches, lights, and whatnot. Range hoods vary from low-powered and inexpensive (less than $50) to custom-designed units (which cost thousands).

Range hoods are most often *wall-mounted* directly over a range. Alternatively, there are *downdraft* and *side-draft* vents that pop up from a counter area to suck away fumes. Over island and peninsula ranges, you can install *chimney-type* vents. In general, install the type of vent that will carry exhaust gases outdoors with the shortest and straightest duct run possible. Because heated air rises, wall-mounted and chimney types are inherently efficient, whereas downdraft and side-draft vents pull heated gases in directions they wouldn't go naturally and can even pull burner flames sideways.

Ideally, a range hood should be slightly wider than the range, say, 3 in. wider on each end, and mounted 30 in. above the range, but follow the hood maker's suggested mounting height. More powerful hoods can be installed higher. Finally, buy a unit with a good-quality filter that can withstand regular washing with soap and water. Most filters are aluminum mesh, better ones are stainless steel; many can be popped into the dishwasher, which spares homeowners a very greasy and unpleasant task. In general, be skeptical of range hoods that recirculate air through a series of filters rather than venting it outside.

Range hoods can be handsome additions to kitchen decor. This one is a European-style wall-mounted hood from Kobe®.

The height of the fan and the fact that it is not against a wall mean that the hood needs to draw more air to be effective. This stylish design is set higher than a typical range hood to provide sight lines across the room.

SIZING THE HOOD

A 100-cfm wall-mounted hood should be adequate to vent the average four-burner, 30-in.-wide range. But if that same range is located on a kitchen island, its range vent should draw 125 cfm to 150 cfm. More is not better when sizing range hoods. For one thing, larger hoods are noisier. Midsize range hoods average 3 sones to 3.5 sones (a measure of noise), which is too noisy to have a conversation nearby; monster hoods can reach 8 sones. (In comparison, refrigerators register 1 sone.) Oversize hoods can also expel so much air that they create *back-drafting,* in which negative in-house air pressure draws furnace or fireplace exhaust gases back down the chimney.

MOUNTING & DUCTING A RANGE HOOD

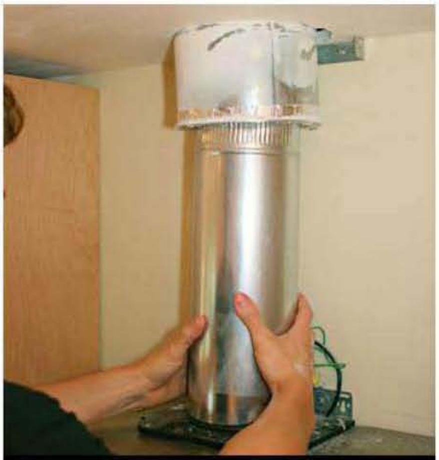

1 Slide the duct into the thimble.

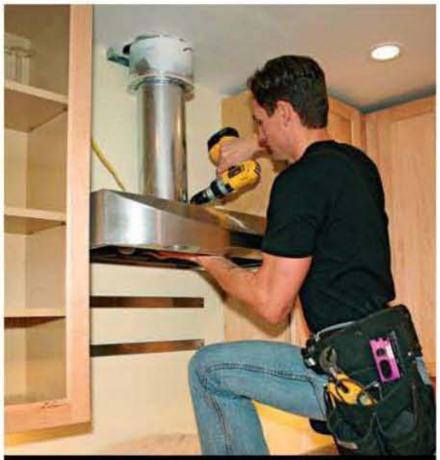

2 Level, center, and mount the hood.

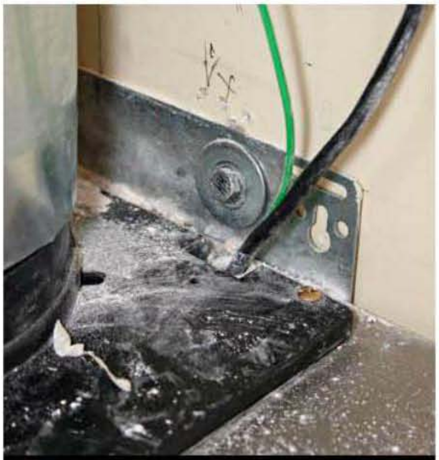

3 Bolt at least one side of the hood to a stud.

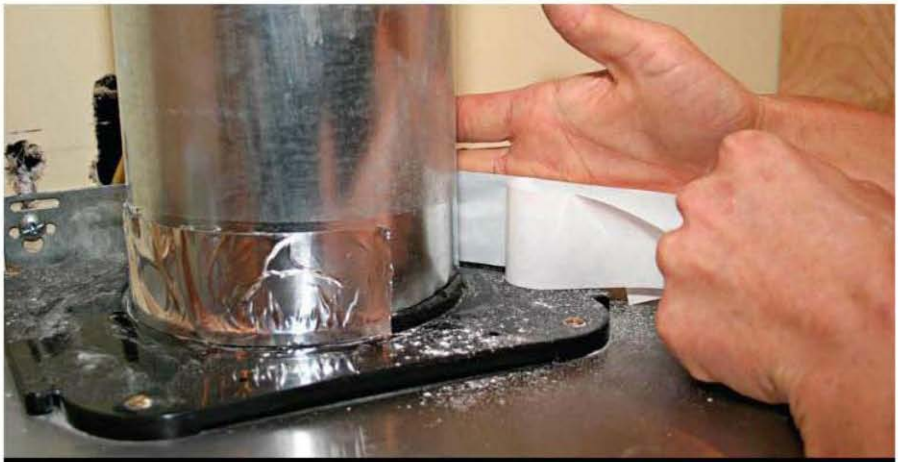

4 Lower the duct over the hood takeoff and seal the joint with metal tape.

TRADE SECRET

A well-sized hood should extend about 3 in. beyond the range on both ends.

Range hoods are typically screwed to the underside of a cabinet or mounted directly to a wall. A hood should be mounted about 30 in. above the range, and the ducting that vents it should exit the house as directly as possible. As you plan your duct route, use a stud finder to locate studs in the wall behind the stove or through ceiling joists above. The duct run should exit through the siding or the roofing. To locate the hole for the duct work, level the hood and center it over the range.

Cut a hole in the wall or ceiling and insert a thimble—a specialized fitting that creates an opening through a wall or ceiling into which the duct fits. If there already is a thimble in place, insert the duct from the range hood up into the thimble 3 in. to 4 in. above its final position ❶. Friction should hold the duct in place.

Level the hood, center it over the range, and bolt it to the wall ❷. Mount the hood in at least two points: If the studs behind the hood are 16 in. o.c., you may have to bolt one side of the hood to a stud ❸ and secure the other side of the hood with a toggle bolt. After the hood is mounted, slide the duct down over the takeoff atop the hood ❹. Tape all of the joints with self-adhering metal tape.

WIRING THE RANGE HOOD

Many range hoods have discrete electrical junction boxes with knockouts, into which you insert cable clamps and cables. In the example shown here, the shroud that encloses the duct doubles as a junction box. The lead wires from the fan and light emerge through a bushing on the top of the hood and attach to wires in a Romex supply cable, which emerges from the wall cavity.

Strip and splice like wire pairs using wire connectors. Splice the ground wires first—be sure there's a grounding pigtail to the hood itself—then neutral wires, and then hot wires ❶.

To minimize weight and avoid marring shiny trim pieces, many pros install the fan blades, filters, trim, and other trappings after the hood shell is mounted. Most of the parts simply snap into place—just follow the instructions provided ❷. Then, if your hood has a bottom casing, screw it into place ❸. Finally, fit the shroud over the ducting and the wire connection—they're usually held in place with one pair of screws at the top and another pair at the bottom ❹.

WARNING

Most hoods aren't heavy but they're unwieldy, so get help installing one.

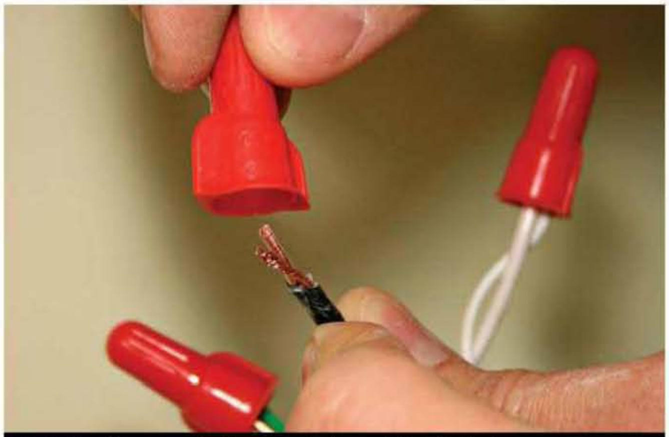

1 Splice the grounds, then neutrals, and then hot wires.

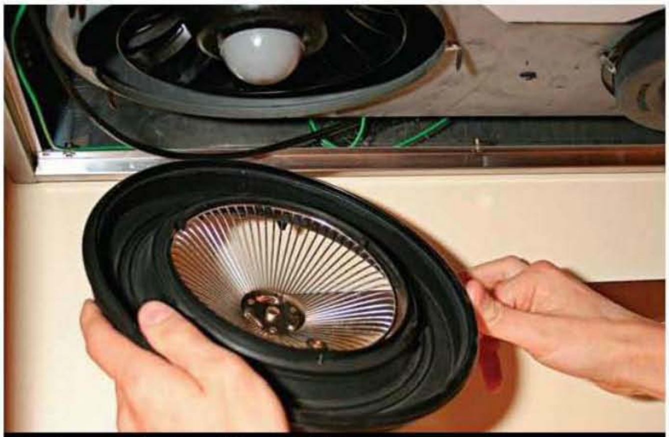

2 Snap on the fans, filters, and trim pieces.

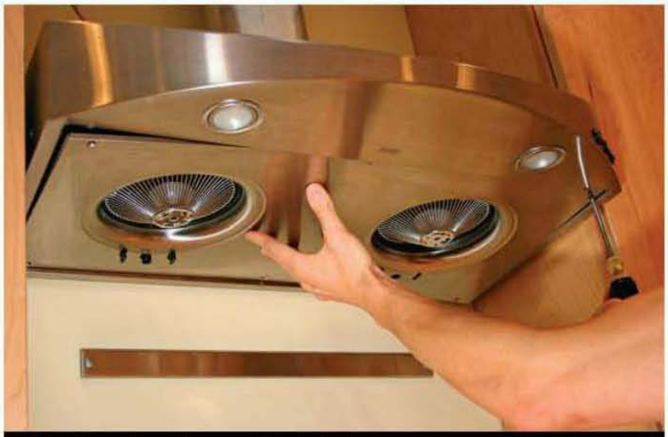

3 Attach the bottom casing onto the hood.

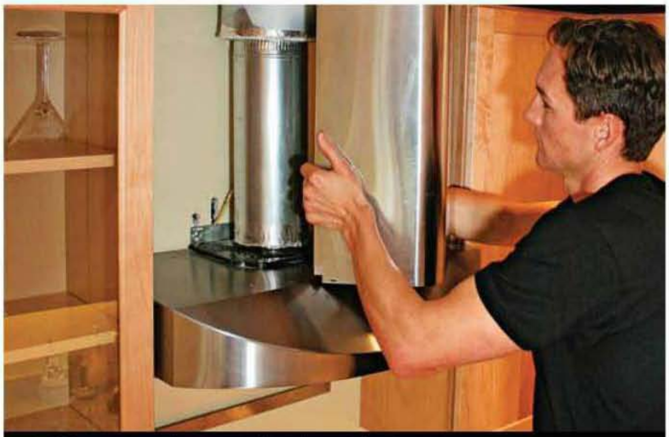

4 Use screws to secure the shroud in place.

CEILING FANS

Choosing a Ceiling Fan

These days, you can find fan styles that match almost any architectural style; the Internet is a good place to start looking. As you might expect, quality correlates closely to cost. Good-quality fans will be balanced, so they don't wobble, and will run quieter, thanks to features like sealed steel bearings.

Most fan companies offer three mounting profiles: *standard* (which employs extension downrods to situate fan blades at an optimal height), *flush* (for low ceilings), and *angled* (for sloping ceilings). To operate safely, fan blades must be located at least 7 ft. above the floor; optimally, fan blades should be 8 ft. to 10 ft. above the floor. To attain that height from ceilings higher than 9 ft., fan makers offer extension downrods of varying length. Most downrods are $^1\!/_2$ in. in diameter; better quality units use rods that are $^3\!/_4$ in. in diameter. For safety reasons, fan blades must be at least 18 in. away from the nearest wall or sloping ceiling.

When spinning, ceiling fan blade spans range from 32 in. to 60 in. There are, of course, complicated formulas for determining fan size, and ideally you should center any fan over the area in which people are most likely to congregate in a room. The chart on p. 155 will help you determine the right size.

In general, more fan blades won't move more air. More important is the blade angle. The steeper the pitch of the blade, the more air it will move. Less expensive fan blades will have a 10-degree pitch and spin faster; better-quality fans will have a 12-degree to 14-degree pitch.

Another important consideration is the switch that controls the fan. Most fans have a multiple-speed pull-chain switch built into the bottom of the housing. If the fan is too high for you to reach the chain comfortably or you just don't like the look of a pull chain, install a wall switch to control the fan. Among the more popular fan controls are three-speed wall switches and remote wireless controls. Avoid using dimmer switches (rheostats) to control the fan, however, as rheostats frequently cause fan motors to hum or, in some cases, burn out. Follow manufacturer's recommendations for appropriate switches.

This low-profile Nouvelle® fan comes with a bronze patina and walnut blades.

Remote controls, like this Adapt-Touch™, allow you to forgo long pull chains.

CEILING FAN LIGHTS

Many homes are designed with a single light outlet in the center of each room. If homeowners want both a fan and a ceiling light in a given room, installing a ceiling fan light is a straightforward and cost-effective solution. Thus many fan makers offer light kits, or fans with integral lights. Until recently, this all-in-one solution resembled a man wearing two hats because the light looked glued onto the fan. Fortunately, some fan makers are now designing combo units that look more integrated—search for them, they're out there. If you still don't find a combo fan and light that fits your taste, add a separate lighting outlet.

The integrated LED light fixture of this Scandia® fan can be modulated by a dimmer switch.

INSTALLING THE CEILING FAN BOX

CEILING FAN ELEMENTS

Mounting details and fan assemblies vary: Follow the installation instructions that come with your unit. Add down-rod extensions as needed to locate the fan blades 8 ft. to 10 ft. above the floor.

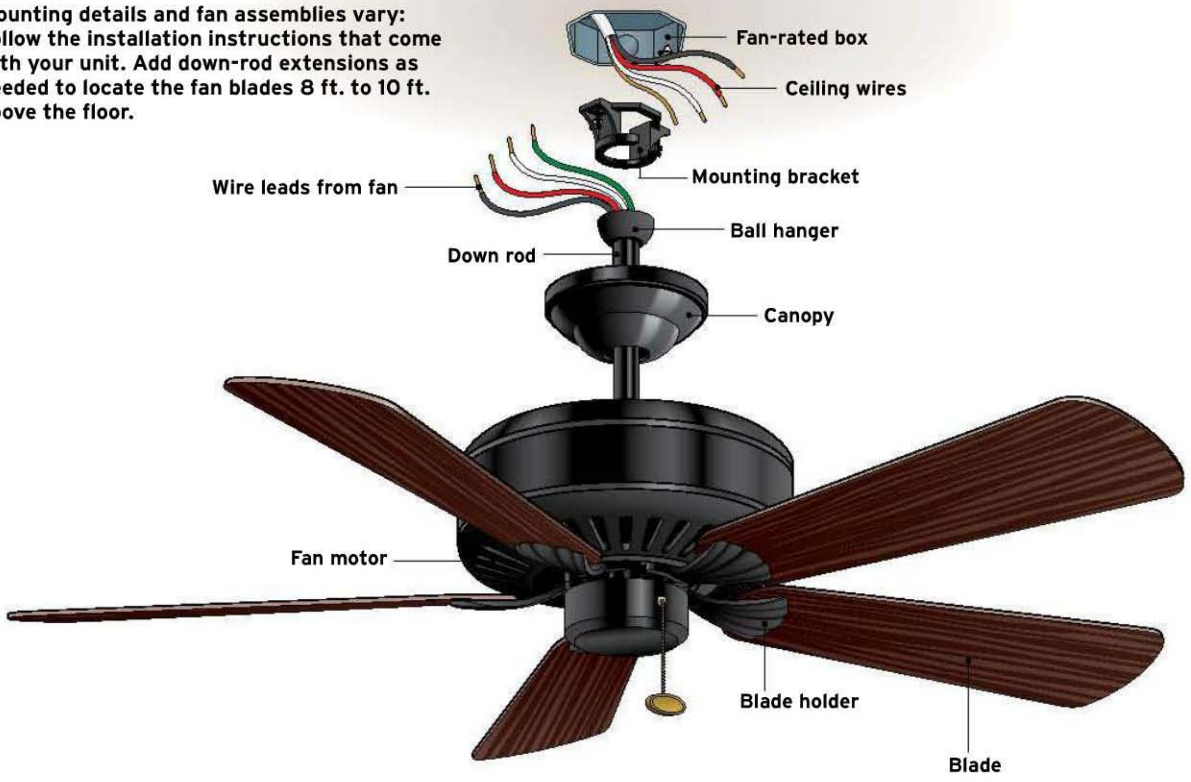

WARNING

Although there are plastic ceiling boxes rated for fans, many installers won't touch them. Spooked by the possibility of stripped screw holes and crashing fans, they insist on using a metal fan box.

All fans must be securely mounted to framing, but this is especially true of ceiling fans, which are weighty and subject to vibration and wobbling. All fans must be installed in a ceiling box rated for fan use—that is, capable of supporting 50 lb. or more. Check your fan's installation specs. If your fan is particularly heavy or complex, it may require a box with a higher weight rating or additional bracing.

If ceiling joists are exposed, you have several options. You can install a piece of 2x4 or 2x6 blocking to the ceiling joists (or rafters), then mount a 2-in.-deep octagonal metal box to the blocking. Mark the location of the blocking so the box will be flush to the finished ceiling surface. Then drive three 3-in. toenailed wood screws through each end of the blocking and into the joists, keeping the blocking on your marks.

If you must hang a fan from a ceiling joist's edge, screw a ½-in.-deep metal pancake box directly to the joist. Before doing so, however, be sure there is enough room inside the fan's canopy to hide electrical connections—because there's no room for them in the pancake box! If not, move the fan or modify your plans.

An alternative is to mount an adjustable hanger to the framing. Bar hangers are strong, easier than nailing up blocking, and easily positioned by sliding the box along its support bar. *Some types of hangers can also be used as remodel bars.*

➔ **For more on bar hangers, see p. 29.**

If joists are exposed, install the blocking and mount a fan box.

WARNING

If there's an existing fan box, turn off the power to it and test before proceeding. If the box is plastic, check the screw holes that the fan mounting bracket attaches to. If the screw holes are at all stripped, or the box is cracked or deformed, or if you have any doubts that it can support the new fan, replace the box.

If joists are not exposed, install a remodel bar (AKA a *braced box*) if there's a finish ceiling. Locate the fan, cut a 4-in.-diameter hole in the ceiling, insert the bar into the hole ❶, and maneuver it until its feet stand on top of the drywall. Then hand-turn the bar to expand it. When the bar touches a joist on both ends, turn the bar with a wrench to drive the bar points into the joists ❷. Finally, bolt the ceiling box to the remodel bar using the hardware provided ❸.

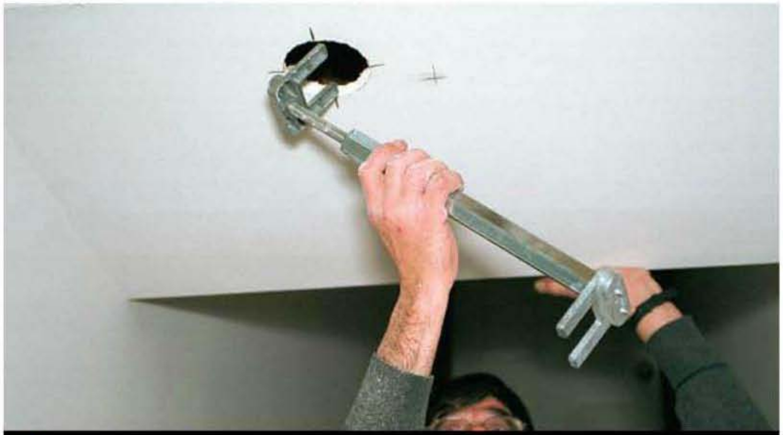

1 Insert a remodel bar through the ceiling cutout.

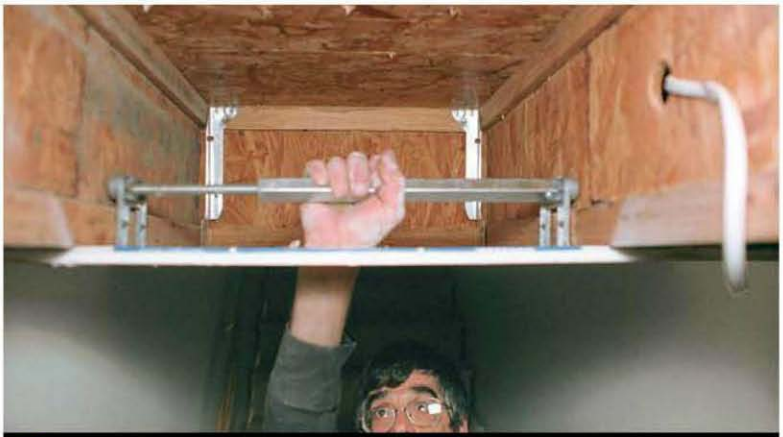

2 Expand the remodel bar by hand. Then force the pointed ends into the joists by turning the bar with a wrench.

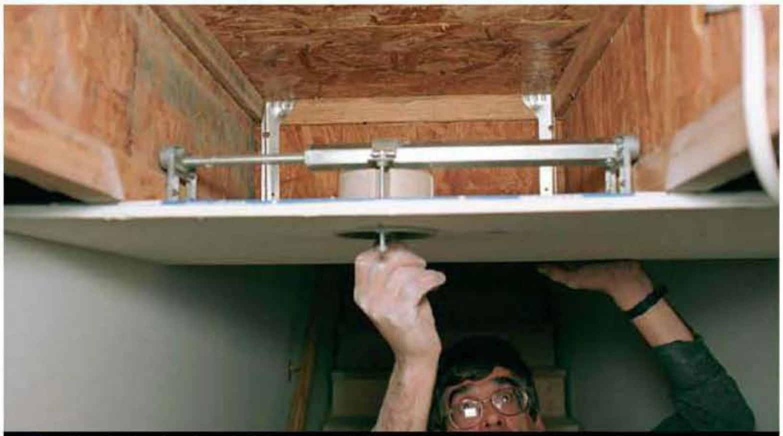

3 Attach the fan box to the remodel bar.

MOUNTING THE CEILING FAN

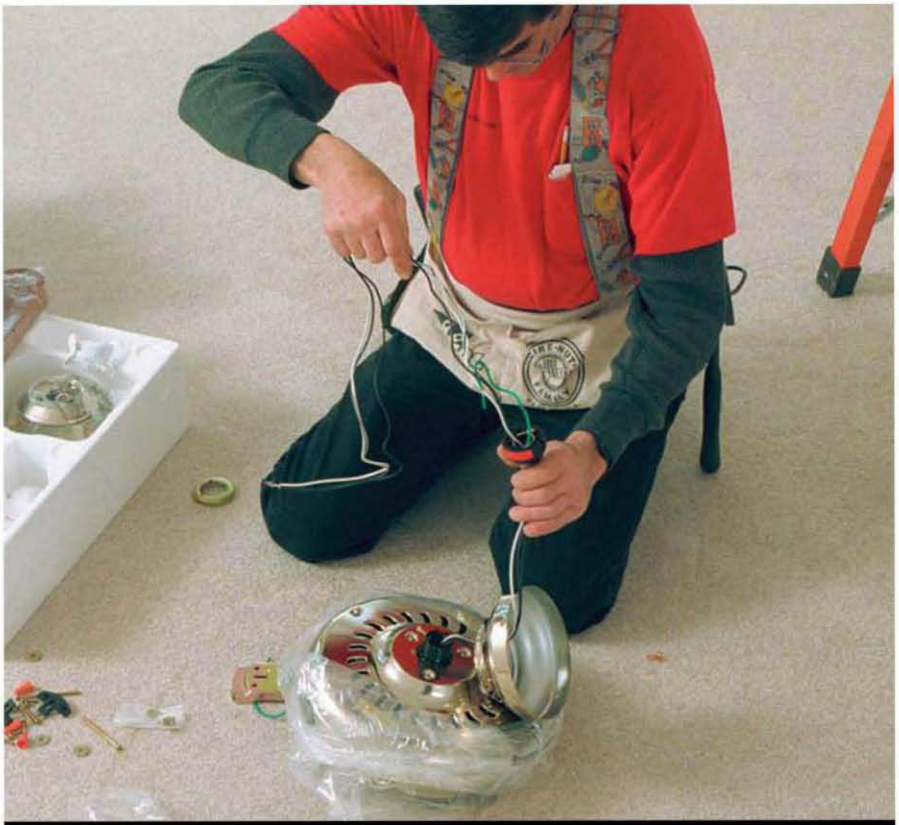

1 Preassemble the canopy and downrod.

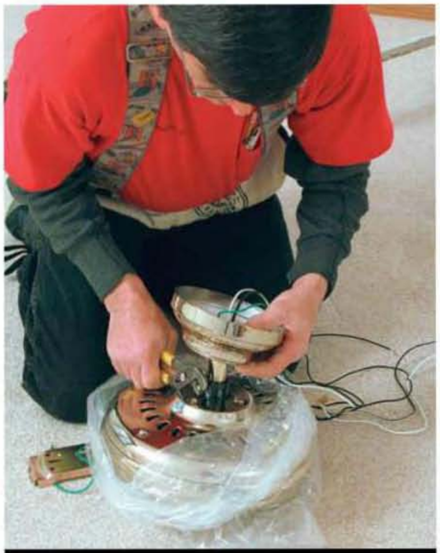

2 Insert the downrod into the bracket.

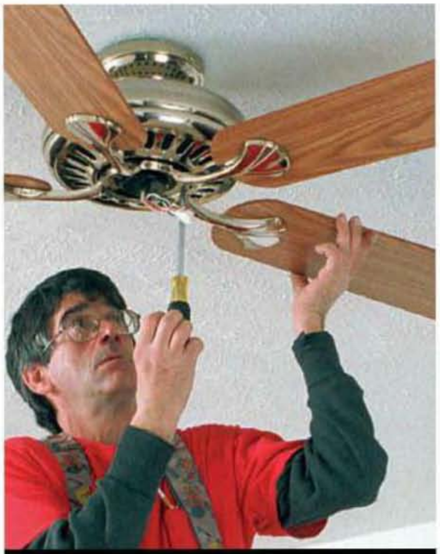

3 Bolt the blade holders to the fan motor.

Fan assemblies differ greatly, so be sure to follow the installation instructions that come with your fan. After testing to be sure the power is off, feed the ceiling wires through the opening in the mounting bracket and screw the bracket to the fan box. *Note:* For some assemblies, the mounting bracket is simply a flat crossbar that spans the fan box; for other types, the fan's weight is supported by a bracket that is somewhat like a slotted dome, into which a ball hanger fits.

Preassemble the parts necessary for hanging the fan motor and housing. Typically, you feed the fan's wire leads up through a canopy and a hollow downrod. The downrod has a ball head that slips into, and is supported by, a slotted mounting bracket screwed to a ceiling outlet box ❶.

Carefully lift the fan motor toward the mounting bracket in the ceiling. Being careful not to pinch the wires from the ceiling or those from the fan motor, insert the ball head of the downrod into the opening of the mounting bracket ❷. On some models, you may need to rotate the fan body until a slot in the ball head aligns with a pin in the bracket. If there is a locking bolt or hanging pin that secures the assembly, be sure to attach it or the unit may fall.

Now wire the unit. Because all metal boxes must be grounded, attach a grounding pigtail to the bracket. Use wire connectors to join like wire pairs, starting with grounding wires, then the neutral wires, and then hot wires. If there's a fan light, there will be two hot wire leads: The black lead typically connects to the fan and the blue or red lead controls the light. Tuck the spliced wires behind the canopy, slide the canopy up the downrod until the canopy is flush to the ceiling, then tighten the setscrew that holds it in place.

Attach the fan blades last. Typically, four machine screws attach a blade to each blade holder; each screw hole is lined with a rubber grommet that cushions the screw-and-nut assembly and forestalls vibration. Each blade holder then bolts to the underside of the fan motor ❸. Because fan blades are prebalanced at the factory, never use blades from other fans. Once you've tightened and checked all bolts and screws, turn the fan on its slowest speed to check for wobbling.

TRADE SECRET

The most common cause of wobbling is a blade holder that's become misaligned. To check for that, turn off the fan and use a yardstick to see if all blade holders are the same height from the ceiling. Most blade holders can be removed and bent slightly back into alignment; the store that sold you the fan may also have a technician. Fan makers also have websites that describe how to balance and troubleshoot their fans.

WIRING A FAN-LIGHT COMBINATION

The fan types discussed in this section are frequently configured with both fans and lights. Running a three-wire cable (12/3 with ground or 14/3 with ground) enables you to operate the fan and light separately. When a ceiling fan's junction box is closer to the power source than to the switch box, as shown in the drawing below right, the switch is called a "back-fed switch." When wiring a back-fed switch leg, tape the white wire black to indicate that it's being used as a hot wire.

INCOMING POWER THROUGH SWITCH BOX

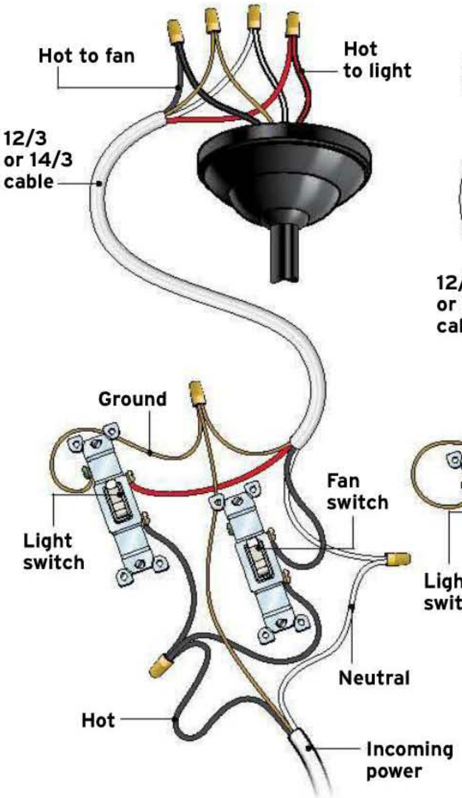
Hot to fan
Hot to light
12/3 or 14/3 cable
Ground
Fan switch
Light switch
Neutral
Hot
Incoming power

BACK-FED SWITCH

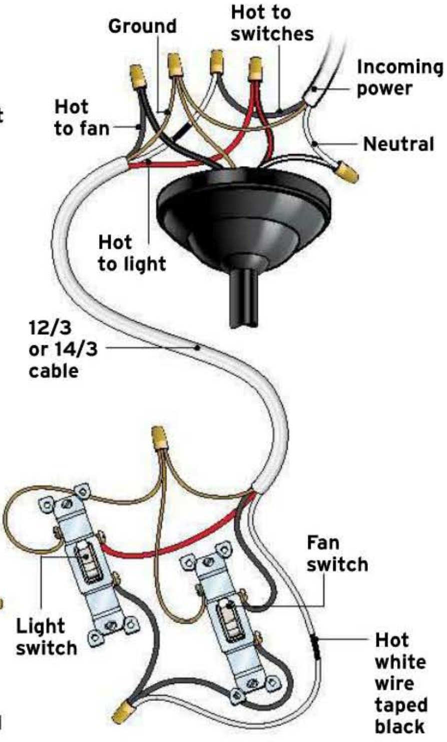
Ground
Hot to switches
Incoming power
Hot to fan
Neutral
Hot to light
12/3 or 14/3 cable
Fan switch
Light switch
Hot white wire taped black

FAN SIZES	
Blade Span (in.)	Room Size (sq. ft.), max*
32 or 36	100
42 or 44	144
48	225
50 or 52	400
54 to 60	> 400

* Output may also be expressed in cfm (cubic feet per minute). For example, a good-quality fan with a 36-in. span will move 2500 cfm to 4000 cfm.

PLANNING NEW WORK

BEING DETAIL ORIENTED IS AS important to planning as it is to installation. When you plan a wiring project, be methodical: Assess the existing system, calculate electrical loads, check local codes, and draw a wiring floor plan.

If you are only replacing existing devices—changing a light fixture, replacing a faulty switch, or upgrading a receptacle, for example—you seldom need a permit from the local building department. However, if you extend or add any circuit, you must pull (or file) a permit.

Most local electrical codes are based on the NEC. When it's necessary to pull a permit, local code authorities will want to approve your plans and later inspect the wiring to be sure it's correct. Don't short-circuit this process: Codes and inspections protect you and your home.

Whatever the scope of your project, if you work on existing circuits, first turn off the power and test to be sure it's off, as shown throughout this book.

BEFORE YOU BEGIN

Inspecting the fuse box or breaker panel, p. 158

Assessing wiring condition, p. 160

Is the system adequately sized?, p. 161

ACCORDING TO CODE

Common code requirements, p. 163

General-use circuit requirements, p. 164

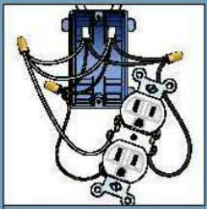

WORKING DRAWINGS

Developing a floor plan, p. 165

Electrical notation, p. 168

Receptacles, p. 170

Switches & lights, p. 172

Receptacles, switches & lights, p. 174

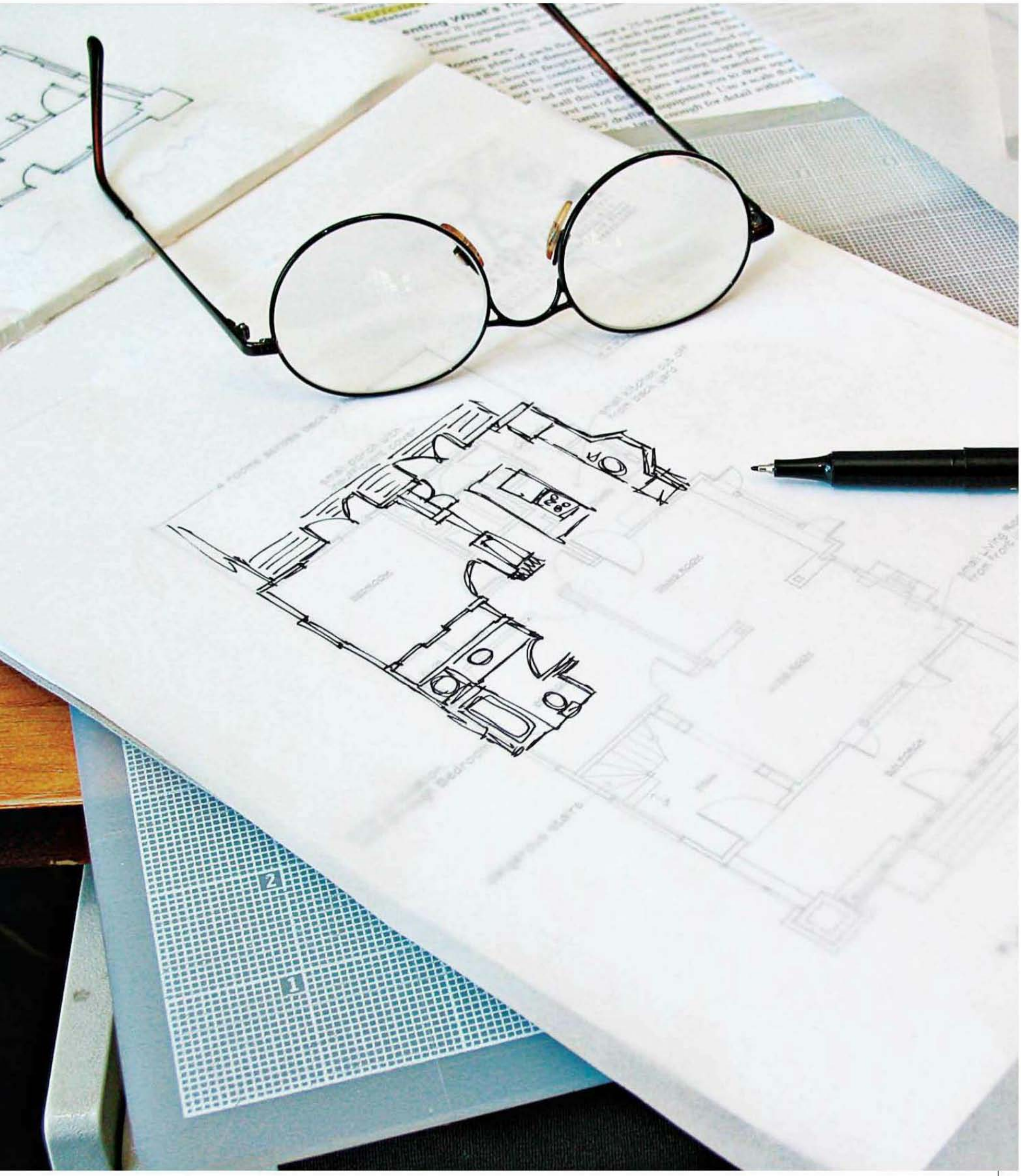

INSPECTING THE FUSE BOX OR BREAKER PANEL

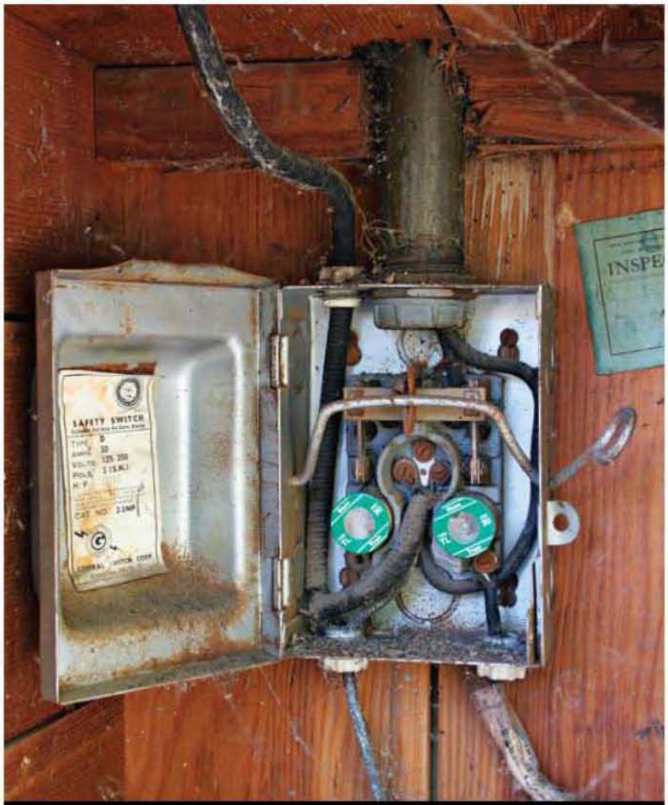
SAFETY SWITCH
INSPE

A This 30-amp main has seen better days. Even if it were safe, it would be dramatically undersize for today's electrical needs.

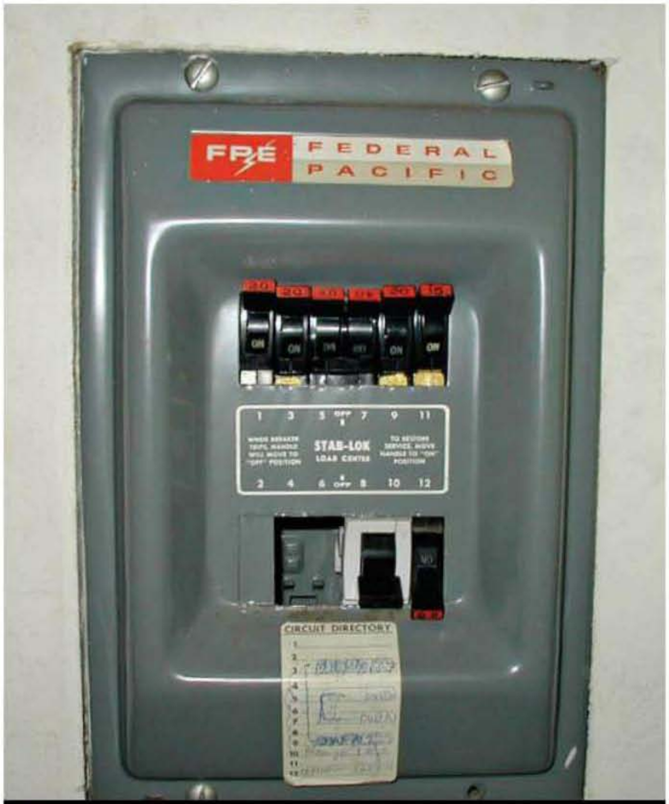
FPE
FEDERAL PACIFIC
STAB-LOK
LOAD CENTER
CIRCUIT DIRECTORY

B Replace this panel: Federal Pacific panels are notorious for arcing and panel fires and hot bus bars are exposed.

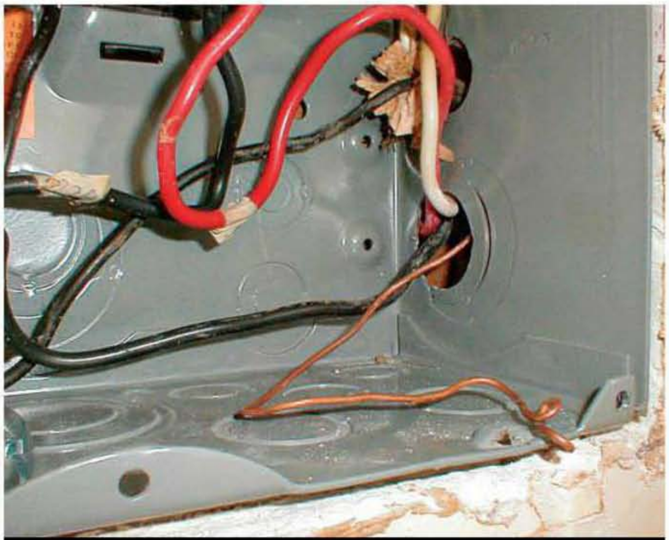

C All cables entering a panel must be protected by a cable clamp. Sharp panel edges could slice wire insulation.

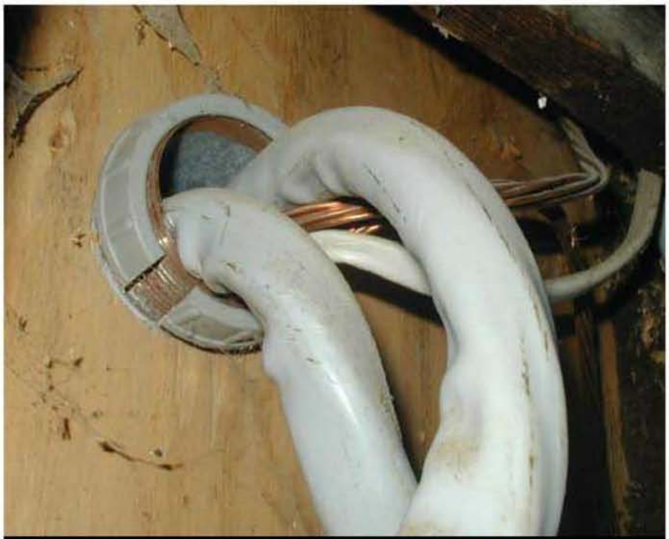

D Here, service cable enters a panel without a cable connector, and the box edge has nearly sliced through the sheathing.

By looking at the outside of the service panel and wiring that's exposed in the basement and attic, you can get a basic overview of the system's condition. If the wiring is in decent shape, you can probably continue using it and safely add an outlet or two. However, if the system seems unsafe or inadequate, hire a licensed electrician to open the panel and do a more thorough examination. Here's what to look for.

Start your investigation at the fuse box or breaker panel. You can learn a lot about the condition of the system by examining the outside of the service box. Examining the inside of a panel or fuse box is best left to a licensed electrician, however.

➔ **For more on inspecting the panel, see p. 14.**

Rust and corrosion on the outside of a service box or on the armored cable or conduit feeding it can indicate corroded connections inside Ⓐ. Such connections can lead to arcing and house fires, so have a licensed electrician replace the fuse box or panel. Likewise, if you see scorch marks on breakers or a panel, have a pro examine it.

Melted wire insulation is a sign either of an overheated circuit—usually caused by too many appliances in use at the same time—or of a poor wire connection in which arcing has occurred. In the first case, a homeowner typically installs an oversize fuse or breaker to keep an overloaded circuit from blowing so often; but this "remedy" exceeds the current carrying capacity of the wire. The wire overheats and melts its insulation, which can lead to arcing, house fires, or—if someone touches that bare copper wire—electrocution.

A fuse box full of oversized fuses may not melt wires where you can see them, but it may have damaged wire insulation somewhere you can't. Have an electrician inspect the electrical system. Installing type-S fuse socket inserts can prevent overfusing.

"Pennying" a fuse is another unsafe way to deal with an overloaded circuit that keeps blowing fuses. In this case, someone unscrews a fuse, inserts a penny or a blank metal slug into the bottom of the socket—a dangerous act in itself—and then reinstalls the fuse. The penny allows current to bypass the fuse and the protection it offers. Here, again, have an electrician examine the circuits for damage to the wire insulation.

Panel covers that don't fit, have gaps, or are missing are unsafe. Likewise, any installation in which panels, covers, and breakers are from different manufacturers is a Code violation; the NEC requires that panel components be listed (UL listed, for example) and come from a single manufacturer. So if you see covers that have been cut to fit a breaker, cover knockouts that are missing, bus bars that are visible when the panel cover is on, or mismatched components, hire a licensed electrician to correct those problems. Some older brands of breaker panels, such as Federal Pacific® and Zinsco®, have spotty reputations and should also be replaced Ⓑ.

Missing cable connectors or unfilled knockouts enable mice and vermin to enter the panel and nest in it, which can be a fire hazard. Missing connectors also allow cables to be yanked, stressing electrical connections inside the panel Ⓒ. A missing cable clamp may also allow the sharp edge of the panel to slice through thermoplastic cable sheathing, which could energize the panel and electrocute anyone who touches it Ⓓ.

An ungrounded service panel, a major hazard, is explained in greater length in Chapter 1. A properly grounded panel will have a large grounding wire running from the panel to a grounding electrode (rod). For the entire electrical system to be grounded, there must be a continuous ground wire or other effective grounding path running from each device or fixture to the service panel and, by extension, to the grounding electrode. Cold water and gas pipes must also be connected (bonded).

➔ **For more on grounding, see p. 12.**

ASSESSING WIRING CONDITION

Cables may be visible as they near the service panel and as they run through attics and basements. If there are covered junction boxes, carefully remove the covers and examine the wire splices inside—without touching them. You can also turn off power and pull a few receptacles out to better examine the wires.

Deteriorated sheathing is a potential shock hazard, so note brittle fiber insulation and bare wire, but avoid touching it. Sheathing that's been chewed on by mice, rats, or squirrels should be replaced.

NM cable must be stapled within 12 in. of boxes and every 4½ ft. Sagging wire is hazardous because it can get inadvertently strained, jeopardizing electrical connections. Likewise, all boxes must be securely mounted. All NM cable entering metal boxes must be gripped by cable clamps; whereas plastic boxes do not require the strain relief of cable clamps.

Aluminum circuit wiring is a fire hazard unless it is correctly terminated with a COPALUM connector or CO/ALR-rated outlets and switches. If it is incorrectly terminated in a copper device, the two metals will corrode and expand and contract at different rates each time the circuit is under load. This can lead to loose connections, arcing, overheating, and house fires.

Wire splices must be housed within a covered junction box, outlet box, or future box. Wires that are spliced outside a box or inside an uncovered box can be a fire hazard because of the dangers of arcing (sparks leaping gaps between wires). Loose connections not contained in a cover box can easily ignite combustibles nearby because arcs approach 2,000°F.

Knob-and-tube wiring, although outdated, is inherently safe unless individual wire insulation is deteriorated or splices are incorrectly made. Typically, splices that were part of the original installation will not be in a junction box, but must be wrapped with electrical (friction) tape and supported by porcelain knobs on both sides of each splice. Nonoriginal splices must be housed in covered boxes. Have knob-and-tube wiring assessed or modified by an electrician familiar with it; it's quirky stuff. The NEC does not allow knob-and-tube wiring to be buried in insulation, though some local inspectors are okay with the practice.

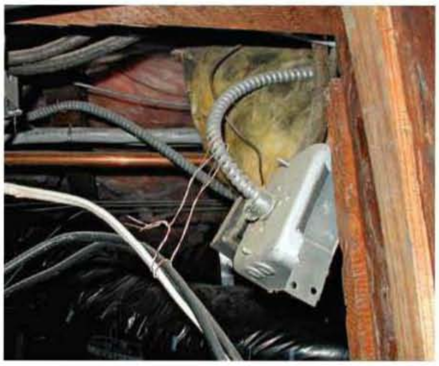

Unstapled cable can be **inadvertently yanked and stress electrical connections. If you see work this sloppy, suspect substandard wiring throughout the house.**

Ungrounded outlets

You can continue using two-slot, ungrounded receptacles on general-use circuits as long as the wires are in good condition and the connections are correctly made. Most new lamp and appliance plugs are double-insulated so there's almost no risk of getting a shock from the plug. However, if you want to use appliances or tools with three-pronged plugs or a surge protector at the location, you must upgrade to a grounded circuit. Putting a two-pronged adapter on a three-pronged plug is unsafe!

A PRO'S TAKE ON REWIRING

If an older home's wiring is in decent shape, it's probably okay to continue using it, even though it may not meet electrical code for a new installation. If you are planning to gut the house completely, it might make sense to rip out all the old wiring and completely rewire the house. But if you're remodeling only part of the house, leave most of the old wiring in place and spend your money rewiring the kitchen, baths, and laundry circuits. That will give you more bang for your buck.

However, you should replace old wiring that's unsafe. If you observe any of these conditions, the wiring should be replaced:

- Circuits that have been extended improperly, as evidenced by loose connections, unprotected splices, or arcing.
- Knob-and-tube wiring whose insulation has been damaged. Also, if knob-and-tube wiring in the attic has been covered with loose-fill insulation or insulation batts, that is a serious Code violation that could lead to overheating and fire danger—that wiring should be replaced.
- Circuits wired with unsheathed wires rather than with sheathed cable or conduit.

IS THE SYSTEM ADEQUATELY SIZED?

If receptacles in your house teem with multiplugs and extension cords, you may need to add more outlets. But there are also more subtle clues: If you blow fuses or trip breakers regularly, or if the lights brown out when you plug in a toaster or an electric hair dryer, you've got overloaded circuits and may need to add new circuits to relieve the load on existing circuits. This section will help you figure out whether your system has the capacity to add new circuits or add outlets to existing circuits.

Electrical service, revisited

Let's start with a recap of the electrical service running from the utility pole to your house. If there are only two large wires running from the utility pole to the house, they deliver only 120v service. A house with two-wire service probably has a 30-amp or 60-amp main fuse or breaker, which is inadequate for modern usage.

➔ **For more on how electrical service works, see p. 10.**

These days, three-wire service feeding a 100-amp service panel is considered minimal, and many electricians install 150-amp or 200-amp panels if the homeowners plan to enlarge the house at some point or acquire a lot of heavy energy-using appliances such as electric ranges and hot tubs.

RULES OF THUMB

The only sure way to know if you've got enough capacity to add an outlet or a circuit is to calculate electrical loads.

➔ **For more on calculating electrical loads, see p. 162.**

But for the benefit of those who wish that an electrician would just offer an offhand opinion of what works most of the time, the "rules of thumb" below continue on the next page.

Fuse box service If you've got a fuse box with a 30-amp or 60-amp main fuse, the best advice we can give is: Get rid of it or upgrade your service. Don't add outlets or circuits until you replace the outside service box with a breaker panel. A fused main should be replaced because it is frequently abused by people trying to bypass its protection, and it is hated by

>> >> >>

Three-wire service, made up of two large, insulated 120v wires and a bare ground/neutral wire, supplies enough power for modern needs.

VOLTS, AMPS, AND WATTS

At a power plant or a substation, electricity is multiplied (charged) and given pressure (*voltage*); in that form, electricity is potential energy, just like a charged battery. When electricity is put to work at an outlet, electrons flow through the wires, and power is delivered as heat or light. The *ampere* (amp) is a unit by which this current flow is measured. The amount of energy consumed at a given point—say, at a toaster or a light bulb—is measured in *watts*. Volts, amperes, and watts are thus interrelated:

- *Voltage:* the potential to do work (electrical pressure).
- *Amperes:* the rate of electrical flow.
- *Watts:* the rate at which energy is consumed.

Or, expressed as mathematical formulas:

- Watts = Voltage x Amperes
- Amperes = Watts ÷ Voltage

To reiterate briefly, electricity, impelled by voltage, flows from the power source. Along the way (at outlets), it encounters resistance and does work. It then returns to the power source, its voltage reduced or spent.

To determine the load-bearing capacity of a circuit you want to extend, identify the circuit breaker controlling the circuit and note the rating of the breaker. If it's a general-purpose circuit, the breaker will probably be 15 amp or 20 amp. A circuit controlled by a 15-amp breaker has a capacity of 1,800w (15 amp x 120v); one controlled by a 20-amp breaker, has 2,400w.

The total wattage of all energy users on the newly extended circuit must not exceed these capacities; otherwise, you risk overheating wires. To avoid overloading, actual load is calculated at 80 percent of capacity. For example, 80 percent of 1,800w is 1,440w for a 15-amp circuit; 80 percent of 2,400w equals 1,920w for a 20-amp circuit.

CIRCUIT CAPACITIES

Amperes x Volts*	Total Capacity (watts)	Safe Capacity† (watts)
15 x 120	1,800	1,440
20 x 120	2,400	1,920
25 x 120	3,000	2,400
30 x 120	3,600	2,990

*Amperes multiplied by volts equals watts. †Safe capacity = 80 percent of total capacity.

IS THE SYSTEM ADEQUATELY SIZED? (CONTINUED)

RULES OF THUMB, continued

insurance companies. Upgrade the service, and your insurance premiums may decrease. In many setups with fused mains there will also be a fuse box inside, which distributes power to house circuits.

Adding outlets If you've got a breaker panel, you can almost always add an outlet or two or more lights, for instance. If, for example, you have three-wire service and a 100-amp main, there's usually a lot of excess capacity.

Adding a circuit for general use

If there's space in the panel to remove a knockout and add a breaker, go ahead. This may be necessary if you're adding a home entertainment center or a computer.

Adding a kitchen or bath circuit

First, see if there's space in the panel to add a breaker. If you're adding a bath fan or some new light fixtures, no problem. If you're adding a 20-amp, small-appliance circuit to reduce the load on an existing circuit, you're probably okay.

Remodeling a kitchen Kitchens are complicated and often full of big energy users. Use the chart at right to help you add up the loads. If there aren't many open spaces for breakers, you may need to upgrade to a larger panel.

Adding dedicated circuits If you need to add dedicated circuits for heavy-use items such as an electric range (50 amps) or a hot tub (60 amps), get out the calculator and do the math.

EXAMPLE OF LOAD CALCULATION FOR SINGLE FAMILY DWELLING

CALCULATING GENERAL LIGHTING LOAD

Type of Load	NEC Reference	Calculation	Total VA
Lighting Load	Table 220-3 (b)	2000 sq.ft. X 3 VA	6000 VA
Small Appliance Load	Section 220-16 (a)	2 circuits X 1500 VA	3000 VA
Laundry Load	Section 220-16 (b)	1 circuit X 1500 VA	1500 VA
Total General Lighting			10500 VA*

Ⓐ CALCULATING DEMAND FOR GENERAL LIGHTING LOAD

Type of Load	Calculation	Demand Factor (DF)	Total VA
General Lighting	First 3000 VA X DF	100%	3000 VA
General Lighting	7500‡ X DF	35%	2625 VA
Total Lighting, Small Appliances & Laundry			Ⓐ 5625 VA

Ⓑ CALCULATING DEMAND FOR LARGE LOAD APPLIANCES

Type of Load	Nameplate Rating	Demand Factor (DF)	Total VA
Electric Range	Not Over 12KVA	Use 8KVA	8000 VA
Clothes Dryer	6600 VA X DF	100%	6600 VA
Water Heater	6600 VA X DF	100%	6600 VA
Other Fixed Appliances	0 VA X DF	100%	0 VA
Total Load for Large Appliances			Ⓑ 21200 VA

	Total VA (Ⓐ+Ⓑ)	26825 VA
Minimum Service Size	Total VA /240V	111.77 VA

*Use this to calculate Ⓐ
‡Total General Lighting Load 10,500 VA – First 3000 VA = 7500 VA

USING THIS TABLE

1. Square ft. for general lighting load is for the entire dwelling including habitable basements or attics.
2. NEC requires a minimum of 2 small appliance loads, but it is important to add small fixed kitchen appliances (microwaves, disposals, etc.).
3. Minimum of 1 laundry load is required for a single family dwelling.
4. The demand factor calculation is designed to take actual use into account (e.g. it is unlikely all lights and small appliances will be running at one time).
5. All large load appliances (high wattage) are added at 100%.
6. The final load calculation is the minimum. Often increasing capacity has little cost impact and is a good practice.

COMMON CODE REQUIREMENTS

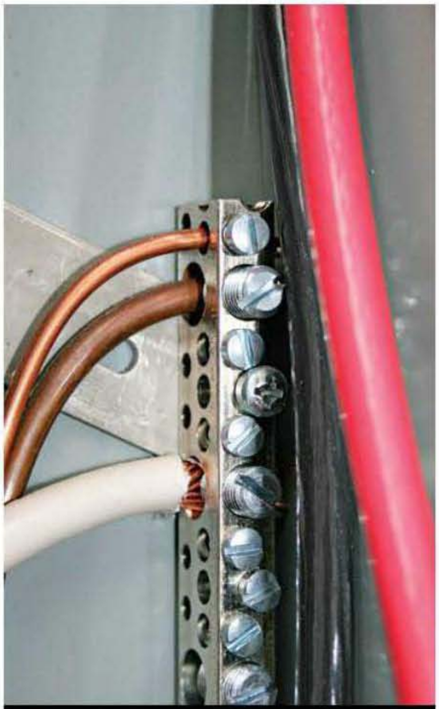
A Bare ground and neutral service wires connect to ground bus in main panel.

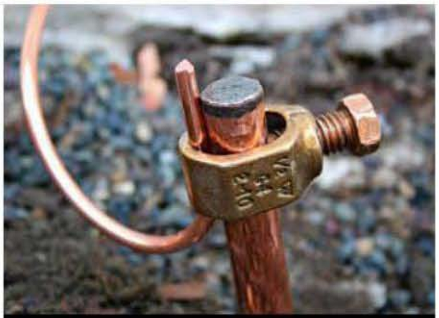
B Large copper ground wire is clamped to grounding electrode (rod).

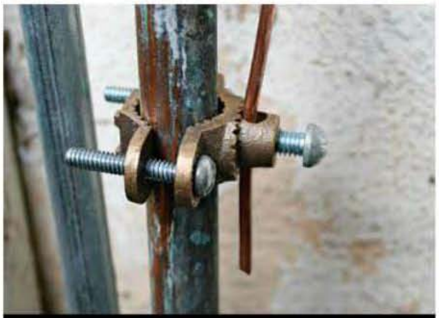
C The main water pipe should be properly bonded.

D All metal boxes and appliance housings require grounding.

E Ground pigtails to ground screws on devices such as receptacles and switches.

Electricians follow the NEC, which was compiled to promote safe practices and prevent house fires. Consider these requirements before you start drawing plans, but be sure to consult local electrical code—it is the final authority in your area.

The requirements given here apply to all circuits in the house, whether general lighting or heavy-use appliance circuits. Local codes rarely require you to change existing circuits—as long as they are safe—but new electrical work should reflect current electrical code.

Circuit wiring

Wire gauge must be large enough to carry the circuit load and be protected by a comparably sized breaker or fuse at the panel. General-use and lighting circuits are typically 14AWG wire, protected by a 15-amp breakers; kitchen, bath, and workshop circuits usually have 12AWG wire, protected by 20-amp breakers.

➔ **See the facing page for more on calculating circuit loads.**

Acceptable cable

Most circuits are wired with NM Romex cable because they are protected behind finish surfaces. When circuit wiring is to be left unprotected and exposed, it must be armored cable or conduit.

➔ **For more on cable and conduit, see p. 204.**

Grounding

All receptacles, appliances, and electrical equipment must be connected (bonded) to the service panel. NM cable contains a separate ground wire, whereas armored cable sheathing and metal conduit provide the path to ground Ⓐ-Ⓔ.

Boxes

All electrical connections must take place in covered boxes. Based on local code requirements, boxes may be plastic or metal. If metal, the box must also be connected (bonded). If NM cable is used, the ground wire must be connected to the box with either a ground screw or a ground clip. If AC cable or metal conduit is used, it must be properly attached to the box to ensure effective bonding. If the box is plastic, it does not need to be (and cannot be) grounded; run a ground wire to the device or fixture only.

GENERAL-USE CIRCUIT REQUIREMENTS

General-use circuits are intended primarily for lighting; but small users, such as televisions, fans, and vacuums, are allowed—as long as the loads they draw don't exceed the capacity of the circuit.

Lighting and small users Though 14AWG wire is sufficient for lighting and switch runs, electricians often run 12AWG wire on general-use circuits to accommodate future uses. Calculate lighting loads at 3w per square foot, or roughly one 15-amp circuit for every 500 sq. ft. of floor space. When laying out the lighting circuits, do not put all the lights on a floor on one circuit. Otherwise, should a breaker trip, the entire floor would be without lights.

Receptacles There must be a receptacle within 6 ft. of each doorway, and receptacles should be spaced at least every 12 ft. along a wall. (This is also stated as, "No space on a wall should be more than 6 ft. from a receptacle.") Any wall at least 2 ft. wide must have a receptacle; and a receptacle is required in hallway walls 10 ft. or longer.

Outlets The NEC does not specify a maximum number of outlets on a residential lighting or appliance circuit, though local jurisdictions may. Figure roughly 9 outlets per 15-amp circuit and 10 outlets per 20-amp circuit.

Light switches There must be at least one wall switch that controls lighting in habitable rooms, the garage, and storage areas (including attics and basements). There must be a switch near each outdoor entrance. Three-way switches are required at each end of corridors and at the top and bottom of stairs with six steps or more. When possible, put switches near the lights they control. It should be noted that the light switch can control a receptacle (considered "lighting") in habitable rooms.

AFCI protection All new 15-amp and 20-amp circuits in bedrooms must have arc-fault circuit interrupter (AFCI) breaker protection to guard against house fires.

KITCHEN LIGHTING BASICS

Kitchen lighting should be designed to utilize natural light during the day and achieve a balance of general light and task lighting at night. Do not be afraid of energy efficient lighting such as fluorescent. Today's energy efficient lighting is instant, dimmable, and available in colors that match incandescent light. Kitchen lighting is often highly regulated for energy efficiency. Check with local building officials before you begin your design.

GENERAL LIGHTING

General lighting is meant to illuminate the space generally and can come from recessed cans, surface mounted fixtures, track lighting, or cove uplighting. Consider cabinetry and appliances when laying out new light fixtures. A general rule is 2w incandescent or 1w fluorescent per square foot of kitchen area, but even illumination is the goal.

TASK LIGHTING

Task lighting is meant to provide a higher level of illumination at work areas (sinks, counter tops, and islands) and can be achieved with recessed cans, pendants, or undercabinet fixtures. If cabinets are over countertops, undercabinet fixtures (T5 fluorescent or halogen strips) are by far the best choice and should be spaced for even illumination of the counter surface. For islands and sinks choose a recessed can with a slightly higher wattage and narrower lamp beam spread, or install pendants with similar attributes.

An AFCI breaker can detect the minute current fluctuations associated with arcing. It then shuts off power to protect you from house fires.

DEVELOPING A FLOOR PLAN

Drawing a set of project plans can help you anticipate problems; find optimal routes for running cable; minimize mess and disruption; and in general, maximize your time and money. A carefully drawn set of plans is also an important part of the code compliance and inspection process.

If you're replacing only a receptacle, switch, or light fixture, you usually don't need to involve the local building department. But if you run cable to extend a circuit, add a new circuit, or plan extensive upgrades, visit the building department to learn local code requirements and take out a permit.

As noted throughout this book, the NEC is the foundation of national and local electrical codes for homes and businesses. Local code authorities have the final say. Your wiring plans should be approved by a local building inspector before you start the project.

Phone first

Call the building department and ask if local codes allow homeowners to do electrical work or if it must be done by a licensed electrician. You may be required to take a test to prove basic competency. This is also a good time to ask if the municipality has pamphlets that give an overview of local electrical code requirements. >> >> >>

REQUIREMENTS ROOM BY ROOM

Kitchen and bath appliances are heavy power users, so their circuits must be sized accordingly.

GFCI PROTECTION

The NEC requires GFCI protection for all bathroom receptacles; all receptacles serving kitchen counters; all outdoor receptacles; accessible basement or garage receptacles; and receptacles near pools, hot tubs, and the like. (Check the current NEC for a complete listing.)

BATHROOM CIRCUITS

Bathroom receptacles must be supplied by a 20-amp GFCI protected circuit. The NEC allows the 20-amp circuit to supply the receptacles of more than one bathroom or to supply the receptacles, lights, and fans (excluding heating fans) in one bathroom. New or remodeled bathrooms must have a vent fan.

SMALL-APPLIANCE CIRCUITS

There must be at least two 20-amp small-appliance circuits in the kitchen. No point along a kitchen countertop should be more than 2 ft. from an outlet—in other words, space countertop receptacles at least every 4 ft. Every counter at least 12 in. wide must have a receptacle.

KITCHEN LIGHTING

Adequate lighting is particularly important in kitchens so people can work safely and efficiently. Lay out a good balance of general and task lighting. Be aware that many jurisdictions have energy efficiency requirements for lighting in kitchens so check with your local building authority first.

BATHROOM LIGHTING

It is important to illuminate the face evenly in mirrors. Common practice is to place good quality light sources either above the vanity mirror or on either side of it. Be careful when using recessed cans over the vanity for they can leave shadows across the face. Many jurisdictions also have energy efficiency requirements for lighting in bathrooms including lighting and occupancy sensors.

DEDICATED CIRCUITS

All critical-use and fixed appliances must have their own dedicated (separate) circuits. These appliances include the water pump, freezer, refrigerator, oven, cooktop, microwave, furnace and/or whole-house air-conditioning unit, garbage disposal, window air conditioners, and water heater. A bathroom heater requires a dedicated circuit, whether it is separate unit or part of a light/fan. Laundry room receptacles must be on a dedicated circuit, as should an electric clothes dryer.

TRADE SECRET

Map your electrical system and place a copy near the panel so that you can quickly identify an outlet later if a circuit breaker trips or a fuse blows.

DEVELOPING A FLOOR PLAN (CONTINUED)

Read up

Make a rough sketch of the work you propose, develop a rudimentary materials list, and then apply for a permit. At the time you apply, the building department clerk may be able to answer questions generated by the legwork you've done thus far. This feedback often proves invaluable.

Inspectors inspect

Inspectors are not on staff to tell you how to plan or execute a job, so make your questions as specific as possible. Present your rough sketch, discuss the materials you intend to use, and ask if there are specific requirements for the room(s) you'll be rewiring. For example, must bedroom receptacles have AFCI protection? Must kitchen wall receptacles be GFCIs if they are not over a counter? Be specific.

Draw up plans

Based on the feedback you've gotten, draw detailed plans. They should include each switch, receptacle, and fixture as well as the paths between switches and the device(s) they control. From this drawing, you can develop your materials list. Number each circuit or, better yet, assign a different color to each circuit. When you feel the plans are complete, schedule an appointment with an inspector to review them.

MATERIALS LIST: LIGHTING AND SWITCHES

This materials list was derived from "Electrical Plan: Lighting and Switches Layer," on p. 169.

ROUGH

Fixture housings	Type 4, 6, and 5
4 ea.	3/0 or 4/0 metal boxes with bar hangars (verify fixture requirements)
3 ea.	4/0 NM boxes with bar hangars (wall fixtures only)
4 ea.	One-gang NM adjustable boxes
2 ea.	Two-gang NM adjustable boxes
4 ea.	Three-gang NM boxes
17 ea.	Romex connectors
500 ft.	14/2 NM Romex: 24 fixtures x 15 ft. ave. = 360 ft. (use excess for home runs, 3 x 50 = 150 ft.)
105 ft.	12/3 NM Romex (use remaining from power rough)

NOTE:
Staples, screws, and nail plates from power rough materials list.

Recommendations:
Verify all surface-mounted fixtures before rough. Some have very small canopies and must have a 3/0 or even a 1 gang box.

For undercabinet lights, do not install a box. Stub cable out of wall approximately 6 in. higher than the bottom of the upper cabinet. This way the drywall can be notched and the cable brought down to the perfect height by the cabinet installer. If undercabinet fixtures are to be installed at the front of the cabinet, some trim piece or metal sheath must be put over the NM cable to protect it from the wall to the fixture.

For recessed and other ceiling lighting, layout fixtures on the floor and use a plumb bob or laser to set fixtures on the ceiling.
To align straight rows of ceiling fixtures, use a taut line.

TRIM

Trims and lamps	Type 4, 6, 5
Fixtures and lamps	Type 1, 2, 3, 5, 8, 9,10
6 ea.	Single-pole Decora switches
5 ea.	Three-way Decora switches
2 ea.	Single-pole Decora dimmers
1 ea.	Four-way Decora dimmer
1 ea.	Decora timer
3 ea.	Three-gang plastic Decora plates
1 ea.	Two-gang Decora plate
4 ea.	One-gang plastic plates
4 ea.	Romex connectors

Listen well, take notes

Be low key and respectful when you meet with the inspector to review your plans. First, you're more likely to get your questions answered. Second, you'll begin to develop a personal rapport. Because one inspector will often track a project from start to finish, this is a person who can ease your way or make it much more difficult. So play it straight, ask questions, listen well, take notes, and—above all—don't argue or come in with an attitude.

On-site inspections

Once the inspector approves your plans, you can start working. In most cases, the inspector will visit your site when the wiring is roughed in and again when the wiring is finished. Don't call for an inspection until each stage is complete.

TRADE SECRET

If you're remodeling, keep in mind that every finish surface you drill or cut into is a surface that you'll have to patch later. So minimize cutting and drilling.

MATERIALS LIST: POWER

This materials list was determined from "Electrical Plan: The Power Layer," on p. 169.

ROUGH

18 ea.	One-gang NM adjustable boxes
1 ea.	One-gang metal cut-in box
1 ea.	Two-gang NM adjustable box (oven)
3 ea.	NM (Romex) connectors
1 ea.	Box (500) $^{3}/_{4}$-in. staples (to be used in rough lighting also)
1 ea.	Box (100) nail plates (to be used in rough lighting also)
500 ft.	12/2 NM Romex (2 x 250ft. rolls): 20 units @ 15 ft. ave. = 300 ft (4 home runs @ 50 ft. ave. = 200 ft.)
250 ft.	12/3 NM Romex (1 x 250ft. roll): 2 home runs @ 50 ft. ave. = 100 ft. (use remainder in rough lighting)
50 ft.	10/3 NM (purchase cut to length): 1 home run @ 50 ft.
1	Bag (250) red wire connectors (to be used in rough lighting also)
1	Container (500) wafer-head #10 x $^{3}/_{4}$-in. screws

TRIM

16 ea.	Duplex receptacles 15-amp or 20-amp rated
3 ea.	GFCI receptacles (15 amp with 20-amp feed through)
1 ea.	30-amp/220v receptacle (verify with range manufacturer)
16 ea.	1 gang plastic duplex plates
1 ea.	30-amp/220v plate

NOTE:

GFCIs are packaged with their plates.

ELECTRICAL NOTATION

Start by making an accurate floor plan of the room or rooms to be rewired using 1/4 in. = 1 ft. scale graph paper. Indicate walls and permanent fixtures such as countertops, kitchen islands, cabinets, and any large appliances. By photocopying this floor plan, you can quickly generate to-scale sketches of various wiring schemes.

Use the appropriate electrical symbols to indicate the locations of receptacles, switches, light fixtures, and appliances.

➔ For a key of electrical symbols, see "Common Electrical Symbols," at right.

Especially when drawing kitchens, which can be incredibly complex, use colored pencils to indicate different circuits. You can also number circuits, but colored circuits are distinguishable at a glance. Use solid lines to indicate cable runs between receptacles and switches and dotted lines to indicate the cables that run between switches and the light fixtures or receptacles they control.

The beauty of photocopies is that you can experiment with different options quickly. As you refine the drawings, refer back to the list of requirements given earlier to be sure that you have an adequate number of receptacles, that you have GFCI receptacles over kitchen counters, that there are switches near doorways, and so on. Ultimately, you'll need a final master drawing with everything on it. But you may also find it helpful to make individual drawings—say, one for lighting and one for receptacles—if the master drawing gets too busy to read.

If you have questions or want to highlight a fixture type, use callouts on the floor plan. As you decide which fixtures and devices you want to install, develop a separate materials list and use numbered keys to indicate where each piece goes on the master drawing. Finally, develop a list of all materials, so you'll also have enough boxes, cable connectors, wire connectors, staples, and so on. In short, list all you need to do the job.

➔ For more on creating a materials list, see the charts on pp. 166–167.

COMMON ELECTRICAL SYMBOLS

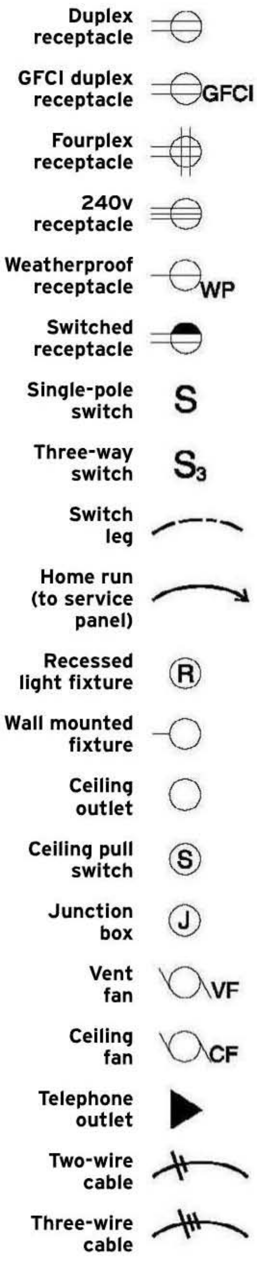
Duplex receptacle
GFCI duplex receptacle
GFCI
Fourplex receptacle
240v receptacle
Weatherproof receptacle
WP
Switched receptacle
Single-pole switch
S
Three-way switch
S3
Switch leg
Home run (to service panel)
Recessed light fixture
R
Wall mounted fixture
Ceiling outlet
Ceiling pull switch
S
Junction box
J
Vent fan
VF
Ceiling fan
CF
Telephone outlet
Two-wire cable
Three-wire cable

LIGHTING AND SWITCHES LAYER

Ⓐ Running 12/2 cable will accommodate the dishwasher circuit.

Ⓑ Use 14/2 cable for all general lighting home runs (cable runs back to the service panel).

Ⓒ Verify the dimmer load; dimmers must be de-rated when ganged together.

POWER LAYER

Ⓐ Feed-through wiring of GFCI receptacle at beginning of run so it affords protection to receptacles downstream (see the top right drawing on p. 171)

Ⓑ Dishwasher circuit. Install under sink in cabinet. Cut hot (gold) tab on receptacle to split receptacle for two circuits. Leave neutral (silver) tab intact. Be sure to install on two-pole breaker with handle tie.

Ⓒ Single-location GFCI protection. Do not feed through (see the bottom drawing on p. 171).

Ⓓ Stove is gas, so receptacle is only for igniter and clock and okay to run with hood. Leave NM cable stubbed at ceiling, leave 3 ft. to 4 ft. of slack for termination in hood/trim. (*NOTE:* Never run a stove igniter off GFCI-protected circuit as it will trip the GFCI every time the stove is turned on.)

Ⓔ Use 12/3 cable for home run, so a single cable takes care of dedicated refrigerator circuit and counter (small-appliance) circuit.

Ⓕ Home run for counter (small-appliance) circuit 2.

Ⓖ Oven outlet. Refer to unit specifications to verify receptacle or hard-wired connection.

ELECTRICAL PLAN: LIGHTING AND SWITCHES LAYER

A professional's electrical floor plan may be daunting at first, but it'll start to make sense as you become familiar with the symbols used. To make the plans easier to read, they have been divided into two layers: (1) lighting and switches and (2) power, which consists of receptacles and dedicated circuits. (There's some overlap.) The *circled letters* are callouts that indicate areas warranting special attention. The *circled numbers* correspond to a list (on p.166) that specifies the type of fixture. Drawing switch legs and circuits in different colors makes a plan much easier to read.

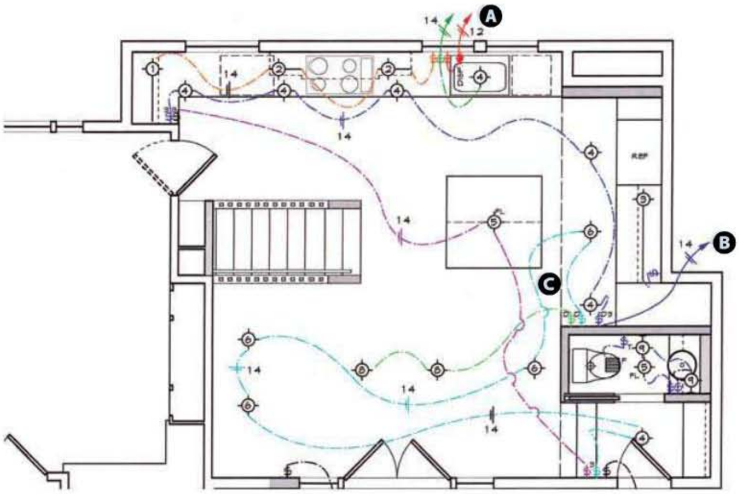
14
12
A
14
14
REF
14
B
14
C
14
14
14
14
1
2
3
4
5
6
9

Drop lights in the dining area are noted by a dotted green switch leg running from the two ceiling boxes to the single-pole switch on the wall.

All circuits for recessed, drop, and undercabinet lighting are noted in different colors. Be certain to track each circuit back to the appropriate wall switch.

ELECTRICAL PLAN: THE POWER LAYER

This kitchen remodel is typical in that it has many dedicated circuits (also called designated circuits) and, per Code, at least two 20-amp appliance circuits wired with #12 cable. *Circled letters* are callouts that correspond to the lettered notes.

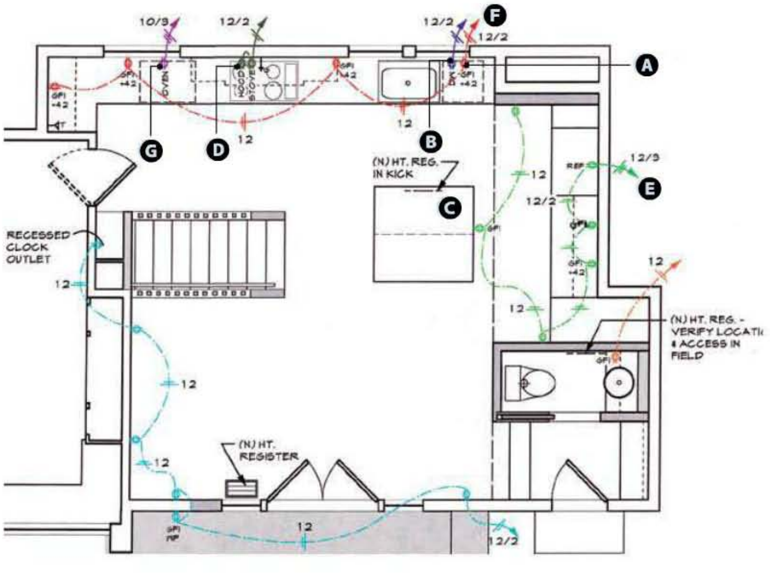
10/3
12/2
12/2
F
12/2
A
GFI +42
OVEN
HOOD STOVE
GFI +42
DW
GFI +42
12
12
G
D
B
(N) HT. REG. IN KICK
C
REF
12
12/3
E
12/2
GFI
GFI
GFI +42
12
RECESSED CLOCK OUTLET
12
12
(N) HT. REG. - VERIFY LOCATI & ACCESS IN FIELD
GFI
12
12
(N) HT. REGISTER
GFI
12
12/2

Whether for 110v wall outlets or 220v appliance outlets, each circuit should end with a home run back to the service panel.

Any receptacle that services the countertop must be GFCI protected. Refrigerators, however, should be run on a non-GFCI receptacle.

RECEPTACLES

RECEPTACLE IN MIDCIRCUIT

By splicing like wire groups and running *pigtails* (short wires) to the receptacle in this conventional method, you ensure continuous current downstream.

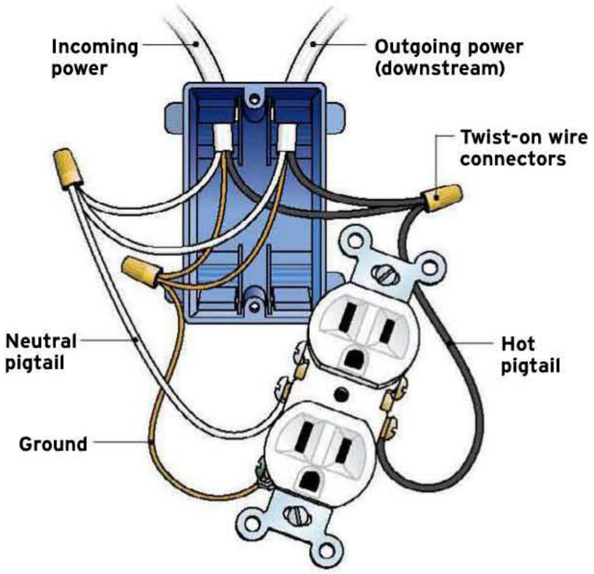
Incoming power
Outgoing power (downstream)
Twist-on wire connectors
Neutral pigtail
Hot pigtail
Ground

FEEDING WIRES THROUGH RECEPTACLE (not recommended)

Attaching hot and neutral wires directly to receptacle terminals is quicker and results in a less-crowded box. However, with this wiring method, if the receptacle fails, power can be disrupted to downstream outlets. *Note:* ground wires are *always* spliced to ensure continuity.

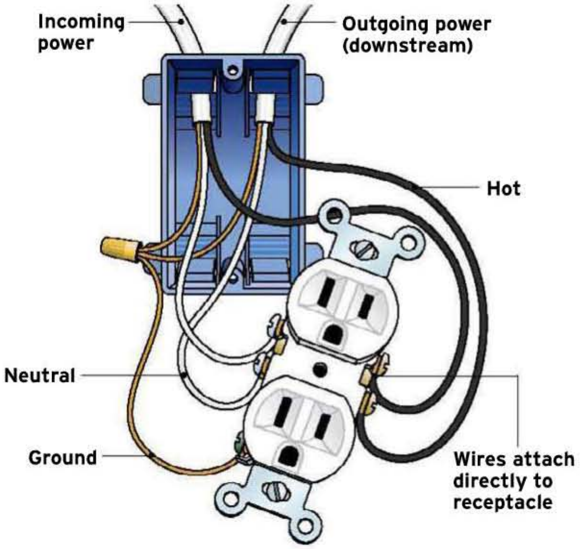
Incoming power
Outgoing power (downstream)
Hot
Neutral
Ground
Wires attach directly to receptacle

The diagrams in this section show most of the circuit wiring variations that you're likely to need when wiring receptacles, fixtures, and switches. Unless otherwise noted, assume that incoming cable (from the power source) and all others are two-wire cable with ground, such as 14/2 w/grd or 12/2 w/grd (#14 wire should be protected by 15-amp breakers or fuses; #12 wire should be protected by 20-amp breakers).

➔ See "Appliances" on p. 210 for more wiring schematics.

All metal boxes must be grounded. Assume that non-grounded boxes in the wiring diagrams are nonmetallic (plastic) unless otherwise specified. In sheathed cables, ground wires are bare copper. Black and red wires indicate hot conductors. (Some devices with multiple wire leads also use blue hot wires.) White wires indicate neutral conductors, unless *taped black* to indicate that the wire is being used as a hot conductor in a switch leg.

RECEPTACLE AT END OF CIRCUIT

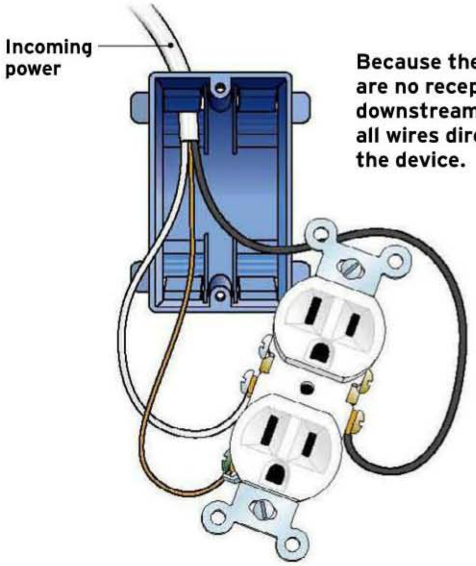
Incoming power

Because there are no receptacles downstream, attach all wires directly to the device.

GANGED RECEPTACLES IN METAL BOX

A two-gang box with fourplex (*double duplex*) receptacles will be crowded. If the box is metal, use insulated ground pigtails and ground the box.

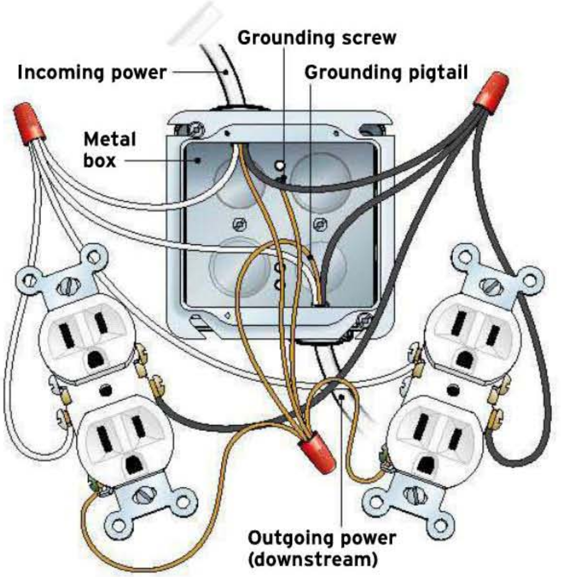
Grounding screw
Incoming power
Grounding pigtail
Metal box
Outgoing power (downstream)

GFCI RECEPTACLE, MULTIPLE-LOCATION PROTECTION

A GFCI receptacle can protect devices* downstream if wired as shown. Attach wires from the power source to terminals marked "line." Attach wires continuing downstream to terminals marked "load." As with any receptacle, attach hot wires to gold screws, white wires to silver screws, and a grounding pigtail to the ground screw. *Note:* Here, only ground wires are spliced; hot and neutral wires attach directly to screw terminals.

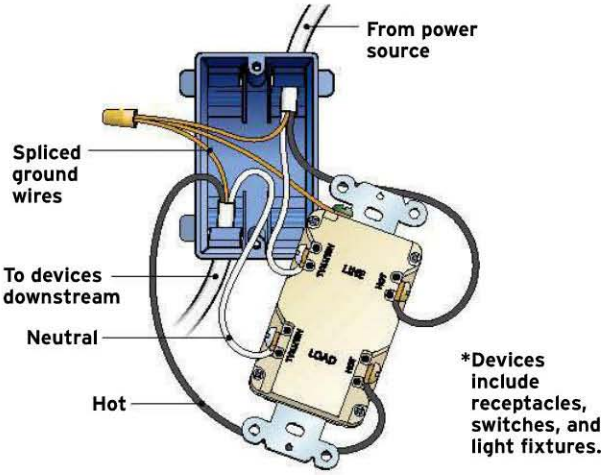
From power source
Spliced ground wires
To devices downstream
Neutral
Hot

***Devices include receptacles, switches, and light fixtures.**

GFCI RECEPTACLE, SINGLE-LOCATION PROTECTION

This configuration provides GFCI protection at one location—say, near a sink—while leaving devices downstream unprotected. Here, splice hot and neutral wires so the power downstream is continuous and attach pigtails to the GFCI's "line" screw terminals. With this setup, receptacle use downstream won't cause nuisance tripping of the GFCI receptacle.

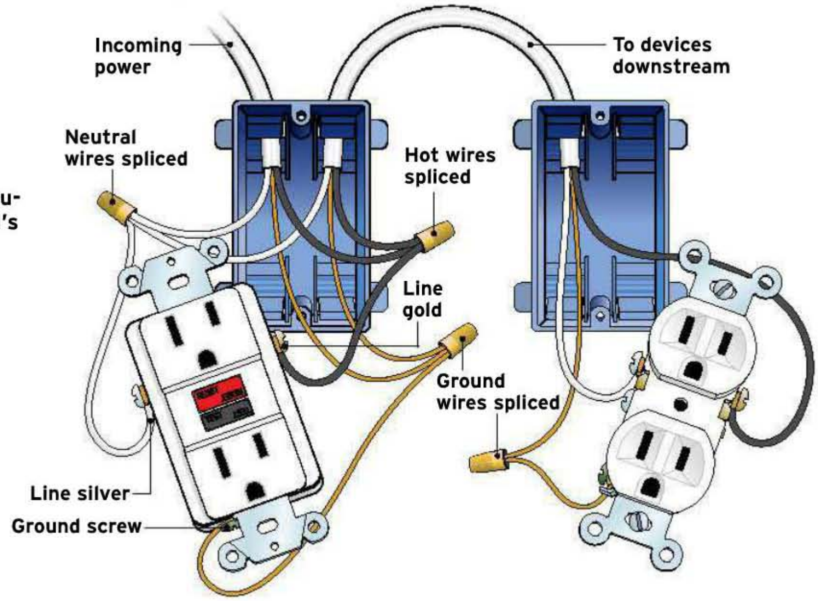
Incoming power
To devices downstream
Neutral wires spliced
Hot wires spliced
Line gold
Ground wires spliced
Line silver
Ground screw

SWITCHES & LIGHTS

LIGHT FIXTURE AT END OF CABLE RUN

Switch wiring at its simplest: Incoming and outgoing hot wires attach to the terminals of a single-pole switch. Neutrals and ground are continuous.

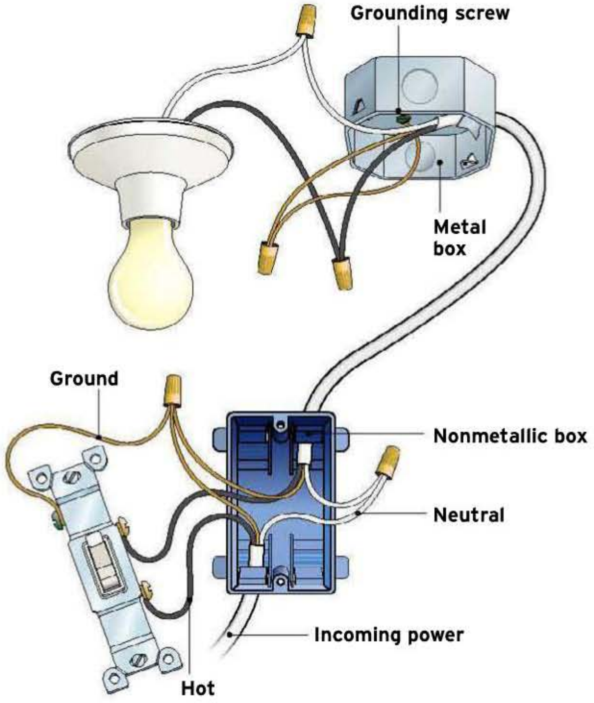
Grounding screw
Metal box
Ground
Nonmetallic box
Neutral
Incoming power
Hot

BACK-FED SWITCH AT END OF CABLE RUN

Attach the incoming neutral to a fixture lead; run the hot to a switch at the end of the cable. Use the white wire of a two-wire cable as one of the hot wires attaching to the switch—but *tape both ends of the white wire black* to show that it's hot.

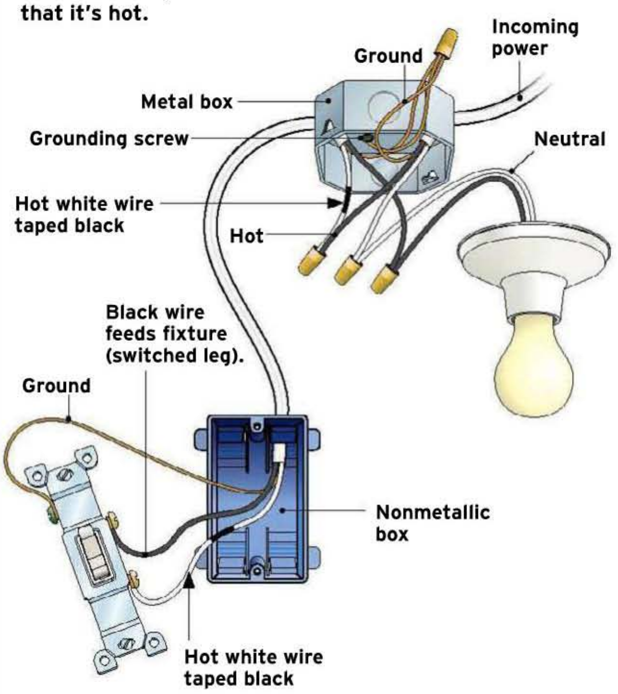
Incoming power
Ground
Metal box
Grounding screw
Neutral
Hot white wire taped black
Hot
Black wire feeds fixture (switched leg).
Ground
Nonmetallic box
Hot white wire taped black

GANGED SWITCHES, TWO FIXTURES

This is a typical setup for switches by exterior doors. For example, one single-pole switch controls an exterior light fixture, while the second switch controls an interior fixture.

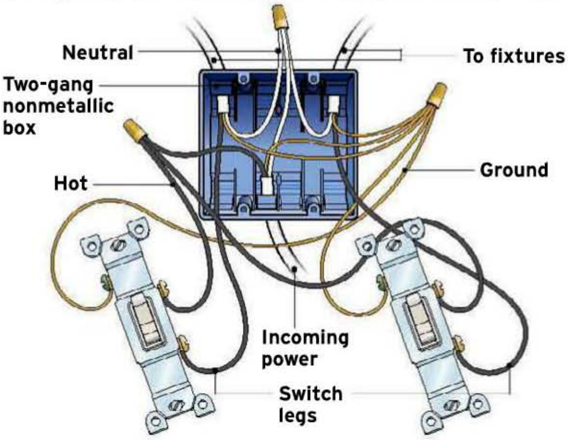
Neutral
To fixtures
Two-gang nonmetallic box
Hot
Ground
Incoming power
Switch legs

CLOSEUP: THREE-WAY SWITCH

Three-way switches control a fixture from two locations. Each switch has two gold screws and a black screw (*common terminal*). The hot wire from the source attaches to the common terminal of the first switch. *Traveler wires* between the switches attach to the gold screws. Finally, a wire runs from the common terminal of the second switch to the hot lead of the fixture.

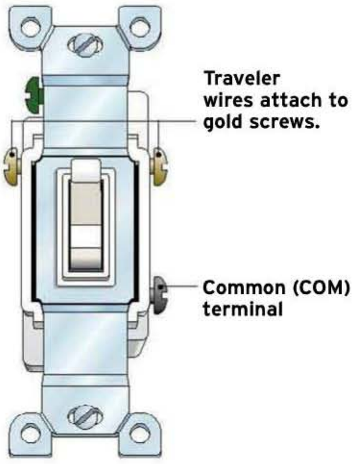
Traveler wires attach to gold screws.
Common (COM) terminal

THREE-WAY SWITCHES, LIGHT FIXTURE BETWEEN

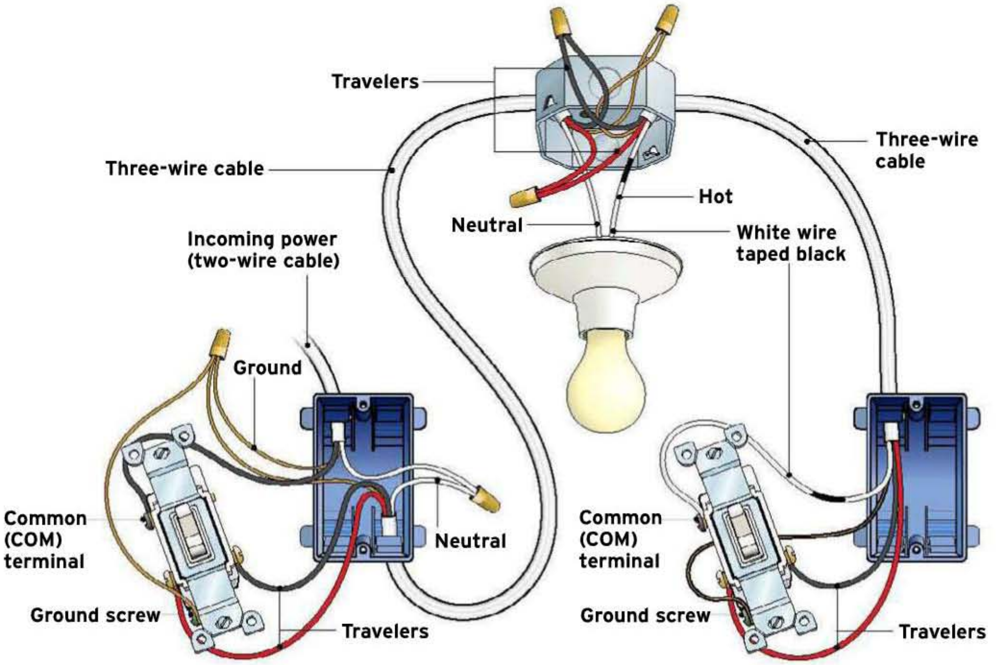
Travelers
Three-wire cable
Three-wire cable
Hot
Neutral
White wire taped black
Incoming power (two-wire cable)
Ground
Common (COM) terminal
Neutral
Ground screw
Travelers
Common (COM) terminal
Ground screw
Travelers

In this setup, two three-way switches control a light fixture placed between them. Run three-wire cables between each switch and the fixture. Whenever you use a white wire as a switch leg, *tape it black* to indicate that it's hot.

THREE-WAY SWITCHES, LIGHT FIXTURE AT START OF CABLE RUN

Here, incoming power enters through the fixture box.

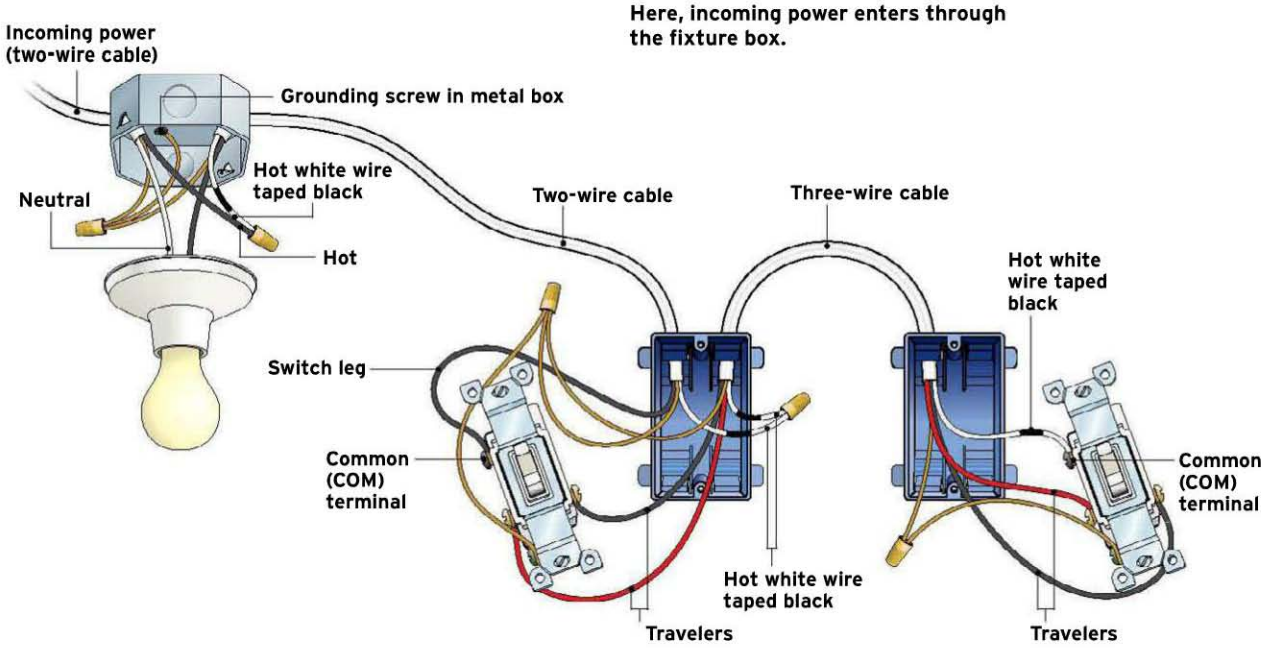
Incoming power (two-wire cable)
Grounding screw in metal box
Hot white wire taped black
Neutral
Two-wire cable
Three-wire cable
Hot
Hot white wire taped black
Switch leg
Common (COM) terminal
Common (COM) terminal
Hot white wire taped black
Travelers
Travelers

RECEPTACLES, SWITCHES & LIGHTS

THREE-WAY SWITCHES, LIGHT FIXTURE AT END OF CABLE RUN

Two three-way switches precede the fixture on the circuit.

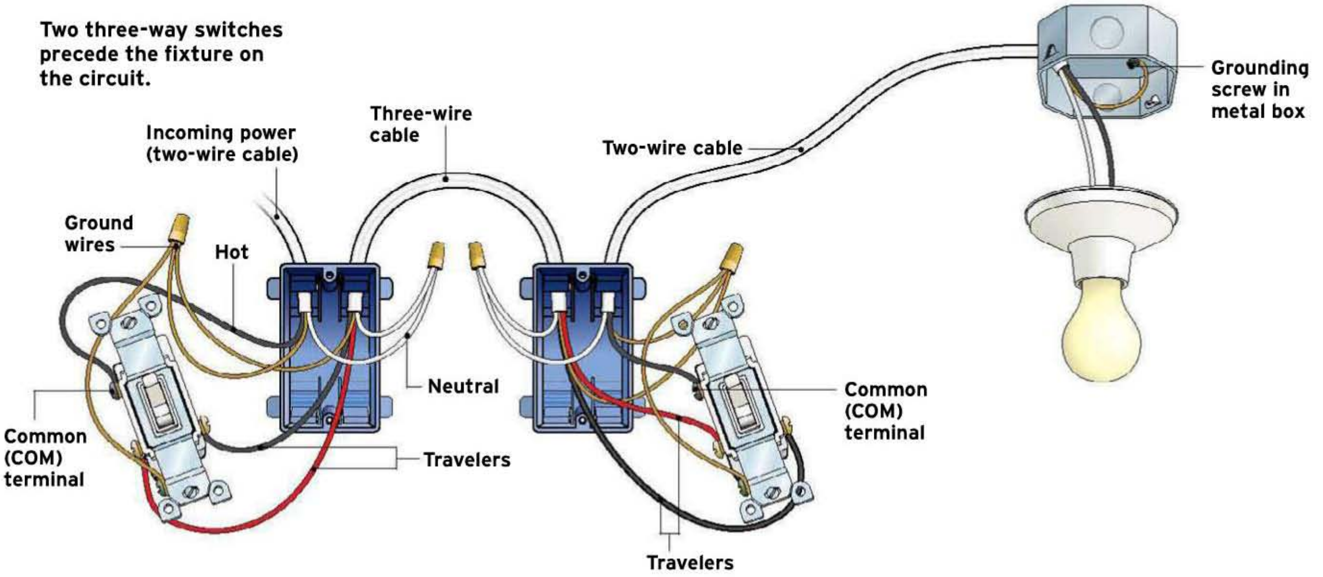
Incoming power (two-wire cable)
Three-wire cable
Two-wire cable
Grounding screw in metal box
Ground wires
Hot
Neutral
Common (COM) terminal
Travelers
Common (COM) terminal
Travelers

SPLIT-TAB RECEPTACLE CONTROLLED BY SWITCH, SWITCH UPSTREAM, REGULAR RECEPTACLE DOWNSTREAM

This setup is commonly used to meet NEC requirements if there is no switch-controlled ceiling fixture. As shown, the switch controls only the bottom half of the split-tab receptacle. The top half of the split-tab receptacle and all receptacles downstream are always hot. Removing the tab is shown in photo 1 on p. 49.

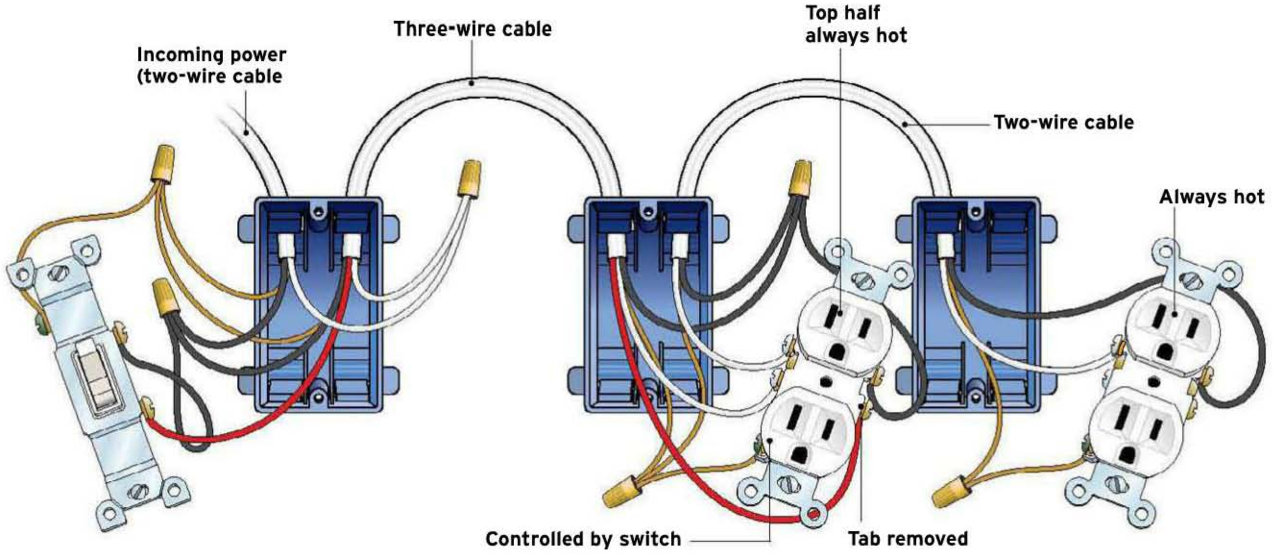
Incoming power (two-wire cable
Three-wire cable
Top half always hot
Two-wire cable
Always hot
Controlled by switch
Tab removed

SPLIT-TAB RECEPTACLE CONTROLLED BY SWITCH, SWITCH AT END OF CABLE RUN

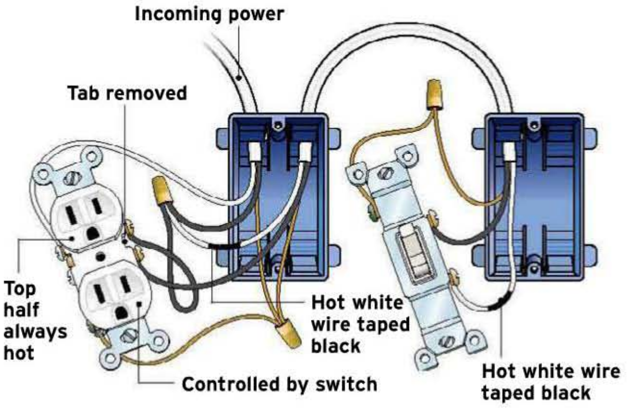
Incoming power
Tab removed
Top half always hot
Hot white wire taped black
Controlled by switch
Hot white wire taped black

The switch controls the bottom half of the split-tab receptacle. The white wire feeding the switch is taped black to indicate that it is being used as a hot wire.

SPLIT-TAB RECEPTACLE AT START OF CABLE RUN

The switch controls the bottom half of the split-tab receptacle on the left. The top half of the same receptacle stays hot, as does the regular duplex receptacle on the right. The white wire is taped black to show it is being used as a hot wire.

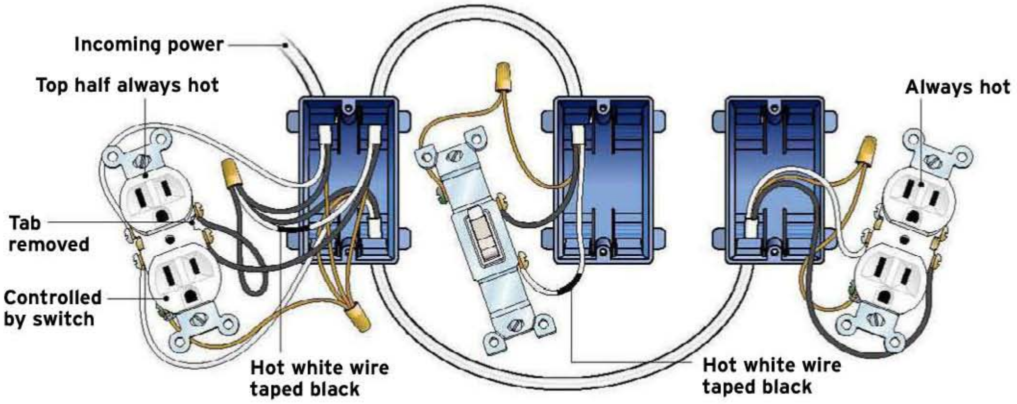
Incoming power
Top half always hot
Tab removed
Controlled by switch
Hot white wire taped black
Always hot
Hot white wire taped black

120/240V RANGE RECEPTACLE

This dedicated circuit requires an 10/3 cable with ground and a double-pole 30-amp breaker. Two 120v hot wires terminate to "hot" setscrews on the receptacle and the breaker poles; the neutral wire to the "neut" setscrew on the receptacle and the neutral/ground bus bar in the service panel.

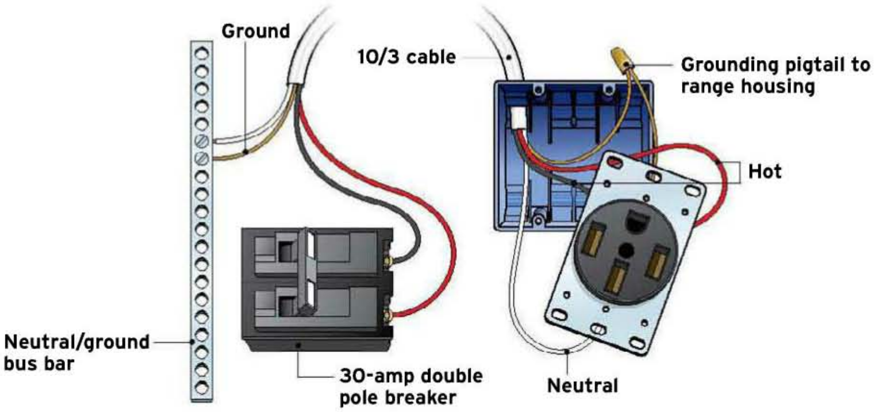
Ground
10/3 cable
Grounding pigtail to range housing
Hot
Neutral/ground bus bar
30-amp double pole breaker
Neutral

ROUGH-IN WIRING

R*OUGH-IN WIRING* REFERS TO THE first phase of a wiring installation. It is the stage at which you set outlet boxes and run electrical cable to them—as opposed to finish wiring, or connecting wires to devices.

Rough-in wiring is pretty straightforward when studs and joists are exposed. Whether a house is new or old, running wires through exposed framing is called *new work,* or new construction. If the framing is covered with finish surfaces such as plaster and drywall, however, the job is referred to as *remodel wiring.* Remodel wiring is almost always more complicated and costly because first you must drill through or cut into finish surfaces to install boxes and run cable, and later you need to patch the holes you made.

GETTING STARTED

Rough-in basics, p. 178
Tools for rough-in, p. 180
Materials for rough-in, p. 182
Laying out the job, p. 184

OUTLET BOXES

Installing wall boxes, p. 186
Installing ceiling boxes, p. 187

RUNNING CABLE

Drilling for cable, p. 188
Pulling cable, p. 189
Feeding cable through corners, p. 190
Fishing cable behind finish walls, p. 191
Fishing cable to a ceiling fixture, p. 192

RETROFIT WORK

Cutting a wall box into plaster, p. 194
Cutting a wall box into drywall, p. 194
Retrofitting a ceiling box, p. 195
Creating a wiring trench, p. 196
Retrofitting boxes & pulling cable, p. 197

MAKING UP BOXES

Making up an outlet box, p. 198
Making up a light can p. 199
Making up a single switch box, p. 200
Making up a multigang switch box, p. 201

METAL CABLE & CONDUIT

Flexible metal cable, p. 202
Installing AC cable, p. 203
Installing MC cable, p. 204
Working with EMT steel conduit, p. 205
Making turns with metal conduit, p. 206
Fishing cable & conduit, p. 207
Prepping receptacles for metal boxes, p. 208
Wiring outlets, p. 209

ROUGH-IN BASICS

Wait until rough carpentry is complete before you begin rough-in wiring. Part of an electrician's job is setting boxes so they'll be flush to finished surfaces. Thus, before an electrician starts working, modifications to the framing—such as furring out or planing down irregular studs and ceiling joists—must be complete. If you're not sure how well the studs are aligned, hold a long straightedge across their edges and look for high and low spots.

Where to Start

Wait until the plumbers are gone. Waste pipes are large and often difficult to locate, which usually means a lot of drilling and cutting into studs and joists. Once the plumbing pipes are in place, you'll clearly see what obstacles you face and will have more room to move around.

Check your plans often. If there's not a table on-site where you can roll out your electrical plans, staple them to a stud—preferably at eye level so you can read them easily. Checking and rechecking the plans is particularly important if you're not a professional electrician.

Be flexible. As you lay out devices, you'll realize that not everything specified on the plans is possible; most plans are developed without knowing exactly what the framing looks like or where obstructions are. Be flexible and choose a solution that makes sense.

Organize Your Work

Perform one task at a time. Each task—such as setting boxes or drilling—requires a different set of tools. So once you have the tools out to do a given task, go around the room and complete all similar tasks. You'll become more proficient as you go: You waste less time changing tools, and the job goes much faster. In general, the sequence of rough-in tasks looks like this:

1. Walk the room with plans, marking outlet locations on walls and floor.
2. Snap chalklines or shoot laser lines to pinpoint box elevations and so on.
3. Attach boxes to studs and ceiling joists.
4. Drill holes for cable runs.
5. Pull cable through holes and into boxes.
6. Make up boxes—strip wire ends, splice wires, attach grounds, attach mud rings, and push wires into boxes.
7. Rough-in inspection.

After the inspection, finish surfaces are installed. Then, at the trim-out or finish stage, wires are attached to the devices.

The Rough-in Inspection

At the rough-in stage, inspectors look for a few key signs of a job well done: cables properly sized for the loads they'll carry; the requisite number and type of outlets specified by Code; cables protected by nail plates as needed; neat, consistent work throughout the system; and, above all, ground wires spliced and, in metal boxes, secured to a

When running cable around doors and windows, find the easiest path. Consider drilling through wall plates and running cable in the floor above or below.

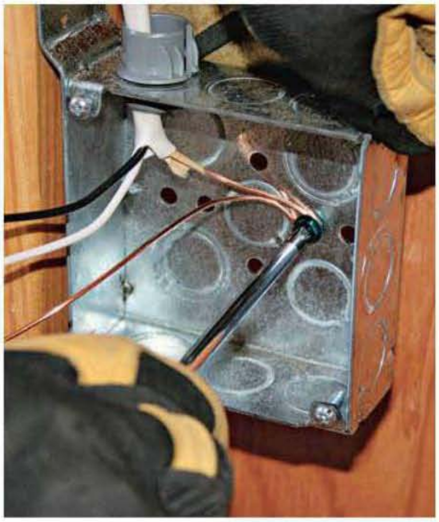

During the rough-in inspection, inspectors demand solid ground-wire splices and, in metal boxes, a ground screw or clip that secures the ground wire.

ground screw. If grounds aren't complete, you won't pass the rough-in inspection.

At this inspection, only grounds need to be spliced. But since you've got the tools out, it makes sense to splice neutrals and continuous hot wires (those not attached to switches).

For more on grounding, see p. 12.

When all splices are complete, carefully fold the wire groups into the box. When you come back to do the trim-out stage, simply pull the wires out of the box, connect wires to devices, and install devices and cover plates.

Connecting circuit wires to a main panel or subpanel is the very last step of an installation. As noted throughout this book, you should avoid handling energized cables or devices.

REMODEL WIRING SAFETY ESSENTIALS

Before removing box covers or handling wires, turn off the power to the area and use a voltage tester to be sure it's off.

First remove the fuse or flip off the circuit breaker controlling the circuit and post a sign on the main panel warning people of work in progress. Better yet, if you've got circuit breakers, do as the pros do and install a breaker lockout so it will be impossible for anyone to turn it on. Breaker lockouts are available at electrical supply houses and most home centers.

For more on lockouts, see p. 238.

Testing for power is particularly important in remodel wiring, because walls and ceilings often contain old cables that are energized. Here, an inductance tester is especially useful. Simply touch the tester tip to cable sheathing or wire insulation. That is, you don't have to touch the tester tip to bare wires to get a reading: If a cable, wire or electrical device is energized, the tip will glow. An inductance tester can detect current through cable sheathing.

Whatever tester you use, test it first on an outlet that you know is live to make sure the tester is working properly.

Always use a voltage tester to test for power before touching cables, devices, or fixtures.

WARNING

Make an emergency plan. It may be as simple as carrying a cell phone or having a friend close by—never do electrical work alone. Calling 911 is an obvious first step if an emergency occurs. On the job site, you should also post directions to the nearest hospital and a list of phone numbers of people to contact.

TOOLS FOR ROUGH-IN

PLS²
PALM LASER

A **Lasers make for quick work when setting boxes.**

B **A wire-connector driver speeds up wire splices.**

D **An 18-in. auger bit bores easily through several studs.**

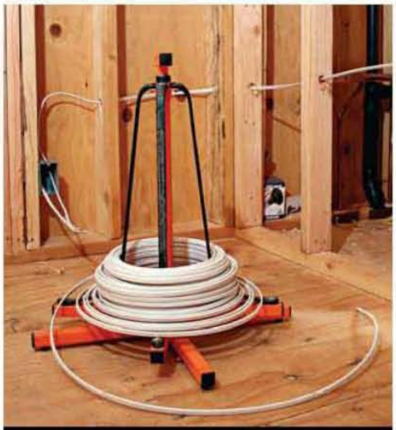

E **A wire reel prevents kinking cable as you pull it.**

Most of the tools you'll need for rough-in wiring are discussed in Chapter 2, so here we'll focus on tools that make rough-in easier and more productive.

Safety tools include voltage testers (especially an inductance tester), eye protection, work gloves, hard-soled shoes, kneepads, a dust mask, and a hard hat. Hard hats are clunky but essential when you're working in attics, basements or any other location with limited headroom. Every job site should also have a fire extinguisher and a first aid kit visibly stored in a central location.

Adequate lighting, whether drop lights or light stands, is essential to both job safety and accuracy. If a site is too dark to see what color wires you're working with, your chances of wrong connections increase.

Sturdy stepladders are a must. In the electrical industry, only fiberglass stepladders are Occupational Safety and Health Administration (OSHA) compliant because they're nonconductive. Wood ladders are usually nonconductive when dry, but if they get rained on or absorb ambient moisture, wood ladders can conduct electricity.

Shop vacuums, push brooms, dust pans, and garbage cans help keep the workplace clean. And a clean site is a safe site. Clean up whenever debris makes footing unsafe, especially if you're cutting into walls or ceilings. Plaster lath is especially dangerous because it's loaded with sharp little nails that can pierce the soles of your shoes.

Layout tools include tape measures, spirit levels, chalklines, and lasers. *Rough-in* is a misleading term because there's nothing rough or crude about locating fixtures or receptacles—layout is very exacting. For that reason, levels and plumb lasers are increasingly common when laying out kitchen and bathroom outlets Ⓐ.

The most used hand tools are the same pliers, cutters, strippers, levels, and screwdrivers mentioned throughout this book. It's also handy to have a wire-nut driver Ⓑ if you've got dozens of splices to make up. There's also a nut-driver bit that fits into a screw gun Ⓒ, but it tends to overtighten wire connectors if you're not experienced in using it. If you'll be cutting into and patching plaster or drywall, get a drywall saw and a taping knife. A flat bar (for prying) and an old, beat-up wood chisel are always busy in remodels, too.

Power tools help speed many tasks. Always wear eye protection when using them. A screw gun is preferable to nailing most of the time because you can easily remove a screw if you want to reposition, say, an outlet box.

A ½-in. right-angle drill is the workhorse of rough-in wiring because it has the muscle to drill through hard old lumber. If you use a standard 6-in.-long auger bit, the drill head and bit will fit between studs and joists spaced 16-in. on-center. It enables you to drill perpendicular to the framing to make wire pulling easier. Longer bits also have their merits: An 18-in. self-feeding auger bit Ⓓ bores easily through several studs or doubled wall plates.

To cut larger holes in plaster or drywall—say, to retrofit pancake boxes or recessed light cans—use a fine-toothed hole saw.

C A nut-driver bit can overtwist spliced wires.

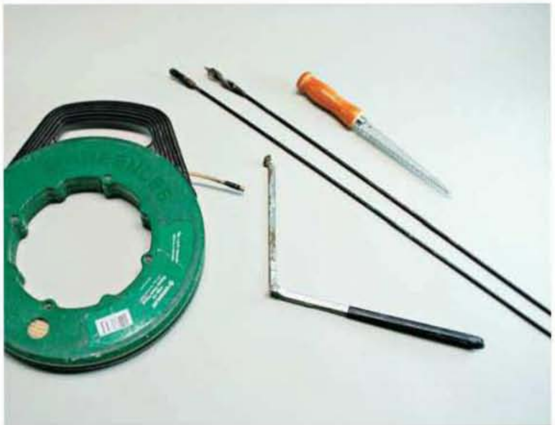

F Remodel wiring tools. From right: drywall saw, flexibit, extension, flexibit steering guide, reel of fishing tape.

A well-organized tool belt prevents you from looking for tools all day long.

Use a sabersaw or rotary cut-out tool to cut individual box openings into finish surfaces. When cutting through plaster lath, alternate cuts on each side of the opening rather than cutting one side completely. This will keep the lath from fluttering and cracking the plaster. Use a reciprocating saw to cut through framing or through plaster lath to create a cable trench.

➔ For more on creating a cable trench, see p. 196.

Use a demolition sawblade to cut through wood that contains nails or screws. It'll hold up to such heavy work.

There are a number of specialty tools designed to ease rough-in. A wire reel, a rotating dispenser that enables you to pull cable easily to distant points, is worth having. Reels hold 250 ft. of cable Ⓔ. A 25-ft. fishing tape—a flexible steel, fiberglass, or nylon tape—enables you to pull cable behind finish surfaces Ⓕ. In most cases, however, it's simpler to use a flexibit to drill through framing in one direction.

➔ For more on pulling cable, see p. 197.

A 48-in. drill extension will increase the effective drilling length of a flexibit. Use an insulated steering guide to keep the flexibit from bowing excessively. In a pinch, electricians wear a heavy work glove to guide a flexibit, but a glove may not protect you from shocks if you accidentally drill through an energized cable.

MATERIALS FOR ROUGH-IN

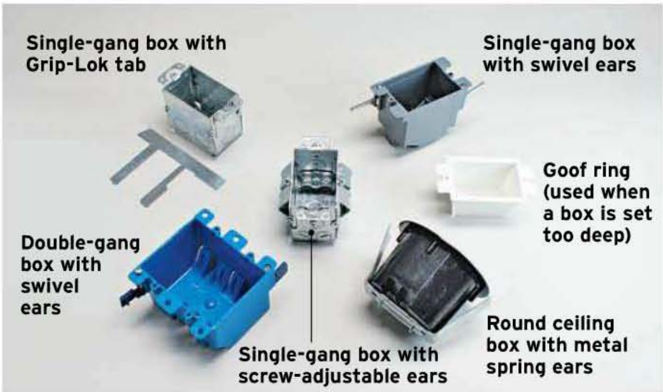
Single-gang box with Grip-Lok tab
Single-gang box with swivel ears
Goof ring (used when a box is set too deep)
Double-gang box with swivel ears
Single-gang box with screw-adjustable ears
Round ceiling box with metal spring ears

A Cut-in (remodel) boxes and accessories.

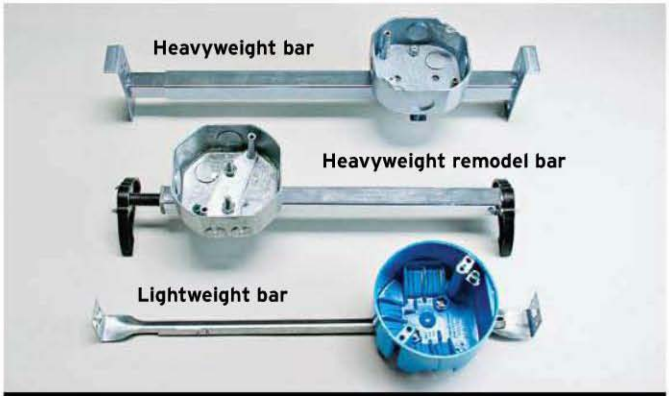
Heavyweight bar
Heavyweight remodel bar
Lightweight bar

B Adjustable bar hangers.

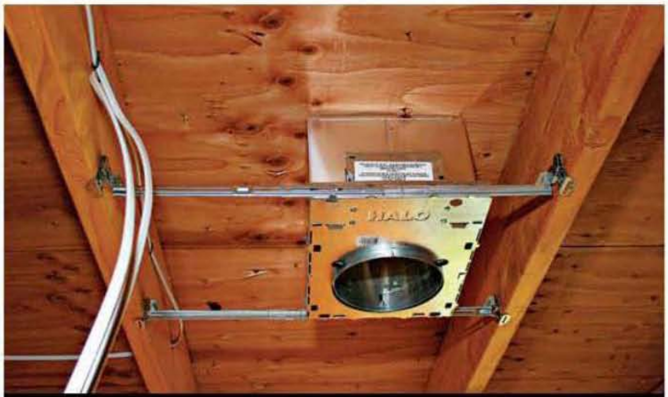

C A recessed light fixture with adjustable bar hangers.

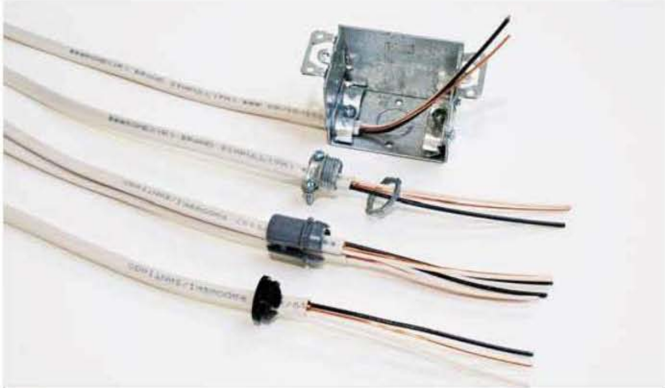

D Romex cable connectors.

As with tools, most of the materials installed during the rough-in phase were discussed in Section 2 and can be installed either in new construction or remodel wiring. There are, however, a number of specialized boxes, hanger bars, and other elements intended for remodel wiring that can be installed with minimal disruptions to existing finish surfaces.

Remodel (cut-in) boxes mount to existing finish surfaces—unlike new-work boxes, which attach to framing. Most cut-in boxes have small ears that rest on the plaster or drywall surface to keep boxes from falling into the wall or ceiling cavity. Spring clamps, folding tabs, or screw-adjustable wings on the box are then expanded to hold the box snug to the backside of the wall or ceiling. The devices that anchor boxes vary greatly **A**.

Code requires that exhaust fan boxes and ceiling boxes be mounted to framing. Expandable remodel bar hangers accommodate this requirement **B**, **C**.

Cable connectors (also called clamps) solidly connect cable to the box to prevent strain on electrical connections inside the box **D**. Cable clamps also prevent sheathing from being scraped or punctured by sharp box edges. Plastic boxes come with integral plastic spring clamps inside. If you use metal boxes, insert plastic push-in connectors into the box knockouts; no other cable connector is as quick or easy to install in tight spaces.

WHAT CAN GO WRONG

Don't forget to leave 12 in. to 18 in. of cable sticking out of each box for connecting devices later.

Ordering Materials

In general, order 10 percent extra of all boxes and cover plates (they crack easily) and the exact number of switches, receptacles, light fixtures, and other devices specified on the plans. It's okay to order one or two extra switches and receptacles, but because they're costly, pros try not order too many extras.

Cable is another matter altogether. Calculating the amount of cable can be tricky because there are infinite ways to route cable between two points. Electricians typically measure the running distances between several pairs of boxes to come up with an average length. They then use that average to calculate a total for each room. In new work, for example, boxes spaced 12 ft. apart (per Code) take 15 ft. to 20 ft. of cable to run about 2 ft. above the boxes and drop it down to each box. After you've calculated cable for the whole job, add 10 percent.

Cable for remodel jobs is tougher to calculate because it's impossible to know what obstructions hide behind finish surfaces. You may have to fish cable up to the top of wall plates, across an attic, and then down to each box. Do some exploring, measure that imaginary route and again create an average cable length to multiply. If it takes, say, 25 ft. for each pair of wall boxes and you have eight outlets to wire, then 8 outlets × 25 ft. = 200 ft. Add 10 percent, and your total is 220 ft. Because the average roll of wire sold at home center contains 250 ft., one roll should do it.

Have materials on hand when it's time to start installing boxes. Electricians often walk from room to room, dropping a box wherever floor plans indicate.

Rough-in recap: electrical code

- **Circuit breakers, wiring, and devices must be correctly sized for the loads they carry. For example, 20-amp circuits require 12AWG wire and receptacles rated for 20 amps. Mismatching circuit elements can lead to house fires.**
- **All wire connections must be good mechanical connections. There must be good pressure between the connectors you are joining, whether wires are spliced together or connected to a device. For that reason, buy devices with screw terminals rather than back-wired (stab-in) devices whose internal clamps can deform. Deformed clamps can lead to loose wires, arcing, and house fires.**
- **All wire connections must be housed in a covered box.**
- **Boxes must be securely attached to framing so that normal use will not loosen them.**
- **Box edges must be flush to finish surfaces. In noncombustible surfaces (drywall, plaster) there may be a 1/4-in. gap between the box edge and the surface. But in combustible surfaces, such as wood paneling, there must be no gap.**
- **All newly installed devices must be grounded. Code allows you to replace an existing two-prong receptacle or to replace a nongrounded box that has become damaged. However, if you install a new three-prong receptacle, it must be grounded. The only exception: You can install a three-prong receptacle into an ungrounded box if that new device is a ground-fault circuit interrupter (GFCI) receptacle. If you extend a circuit, the entire circuit must be upgraded to current Code.**
- **In new rough-in work, cable must be supported within 12 in. of any box and every 4 1/2 ft. thereafter.**

LAYING OUT THE JOB

With the electrical plans in hand, walk each room and mark box locations for receptacles, switches, and light fixtures. Each device must be mounted to a box which houses its wiring connections. The only exceptions are devices that come with an integral box, such as bath fans, recessed light cans, and undercabinet light fixtures ①.

Mark receptacles and switches on the walls first. Then mark ceiling fixtures. If studs and joists are exposed, use a vividly colored crayon that will show up on the wood. If there are finish surfaces, use a pencil to mark walls at a height where you can see the notations easily—these marks will be painted over later. Near each switch box, draw a letter or number to indicate which fixture the switch controls.

Once you've roughly located boxes on the walls, use a laser level ② to set exact box heights for each type of box.

➔For more on box heights, see "Rough-in Recap: Box Locations," on the facing page.

Use the laser to indicate the top, bottom, or center lines of the boxes ③. Many electricians prefer to determine level with the laser, then snap a permanent chalkline at that height so they can move the laser to other rooms ④.

To locate ceiling fixtures, mark them on the floor ⑤ and use a plumb laser ⑥ to transfer that mark up to the ceiling ⑦. This may seem counterintuitive, but it will save you a lot of time. Floors are flat, almost always the same size and shape as the ceilings above and—perhaps most important—accessible and easy to mark. In complex rooms, such as kitchens, draw cabinet and island outlines onto the floor as well. Those outlines will help you fine-tune ceiling light positions to optimally illuminate work areas.

1 Begin the lay out by walking the room and marking box locations.

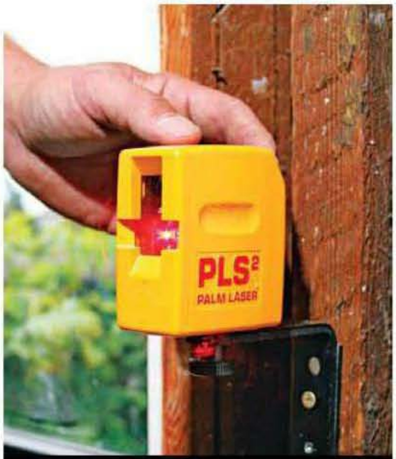

2 Use a laser level to set box heights.

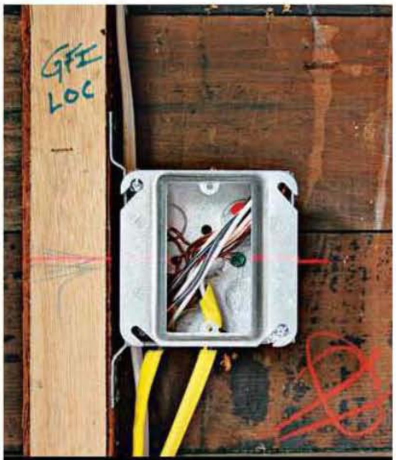

3 Here, the laser beam indicates the centerline of the box.

4 You can snap a level chalkline in addition to using a laser.

Avoiding hot wires In remodels, there may be live wires behind finish surfaces. Use an inductance tester to test receptacles, switches, fixtures, and any visible wires. The cables feeding those devices will be nearby. Wall receptacles are usually fed by cables running 1 ft. to 3 ft. above. Switches often have cable runs up to a top plate; each ceiling fixture has cable running to the switch(es) controlling it. Avoid drilling or cutting into those areas, and you'll minimize the risk of shock.

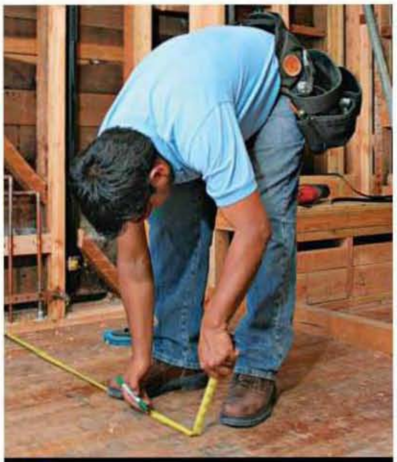

5 Many electricians prefer to mark ceiling fixture centers on the floor.

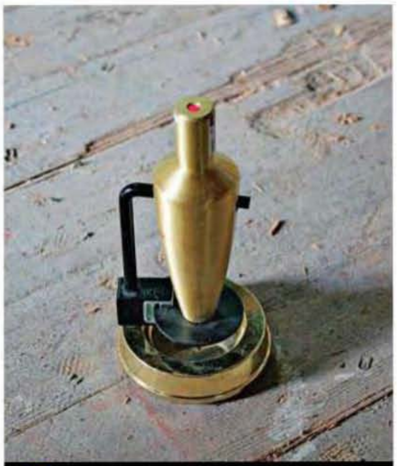

6 They then use a plumb laser to transfer marks to the ceiling.

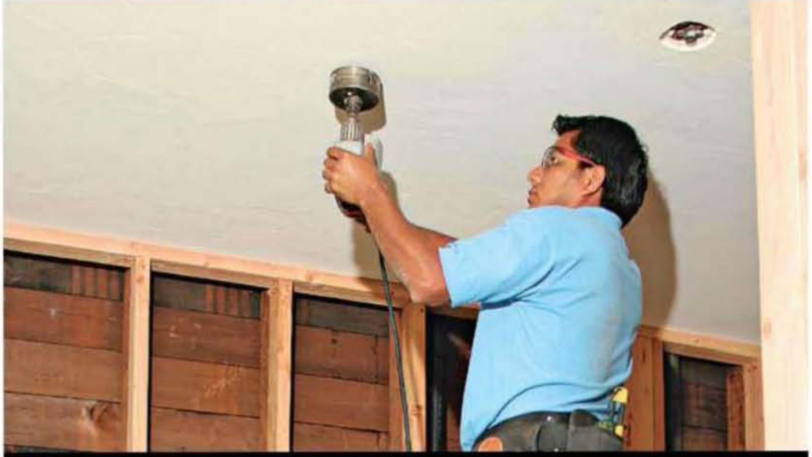

7 Center your hole saw on each ceiling mark.

WHAT CAN GO WRONG

When marking box locations on finish surfaces, use a pencil—never a crayon, grease pencil, or a felt-tipped marker. Pencil marks won't show through new paint. Also, grease pencils and crayons can prevent paint from sticking.

Rough-in recap: box locations

- Whatever heights you choose to set outlets and switches, be consistent.
- Code requires that no point along a wall may be more than 6 ft. from an outlet. Set the bottom of wall outlets 12 in. to 15 in. above the finished floor surface—or 18 in. above the finished floor surface to satisfy Americans with Disabilities Act (ADA) requirements.
- Place the top of switch boxes at 48 in., and they will line up with drywall seams (if sheets run horizontally), thus reducing the drywall cuts you must make.
- In kitchens and bathrooms, place the bottom of countertop receptacles 42 in. above the finished floor surface. This height ensures that each receptacle will clear the combined height of a standard countertop (36 in.) and the height of a backsplash (4 in.), with 2 in. extra to accommodate cover plates.

INSTALLING WALL BOXES

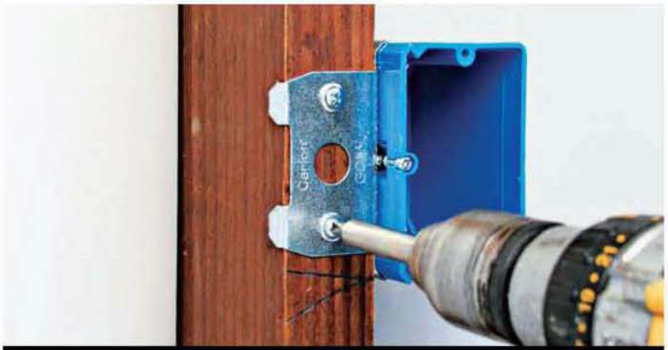

1 To install an adjustable box, press its bracket flush against the stud edge and screw the bracket to the edge or side of the stud.

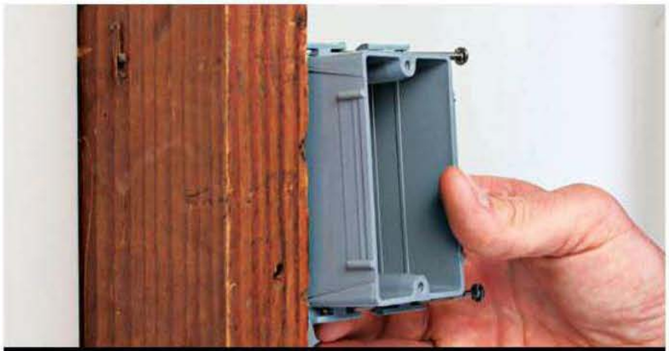

2 Nail-on boxes cannot be adjusted, so use the depth gauge on the side to ensure that box edges will be flush to finish surfaces.

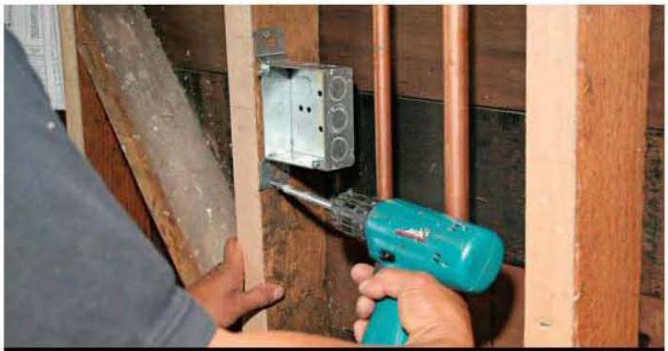

3 Boxes with integral brackets have small points at top and bottom that sink into the stud. A mud ring brings the box flush to the wall.

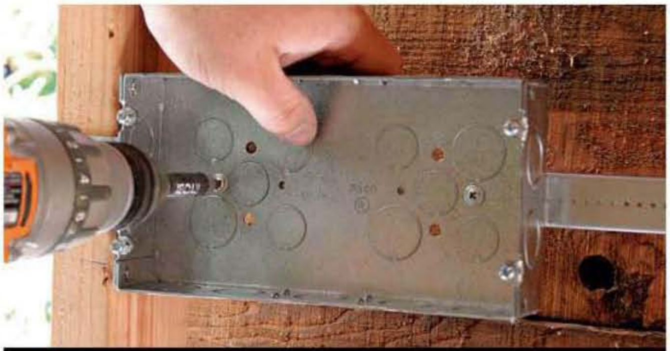

4 To provide solid support for wide multigang boxes, first install an adjustable box bracket that spans the studs.

Once you've established the heights of the switch and outlet boxes, installing them is pretty straightforward. Local electrical codes will dictate box capacity and composition.

➔ For more on box capacity and composition requirements, see p. 25.

In residences, 18-cu.-in. single-gang polyvinyl chloride (PVC) plastic boxes are by far the most commonly used. This size is large enough for a single outlet or a single switch and two cables. Otherwise, use a 22.5-cu.-in. single-gang box or a four-square box with a plaster ring.

Set each box to the correct height, then set its depth so that the box edge will be flush to the finish surface. If you use adjustable boxes, simply screw them to a stud ❶. To raise or lower the box depth, turn the adjusting screw. Side-nailing boxes typically have scales (gradated depth gauges) on the side ❷. If not, use a scrap of finish material (such as ½-in. drywall) as a depth gauge. Metal boxes frequently have brackets that mount the box flush to a stud edge ❸; after the box is wired, add a mud ring (plaster ring) to bring the box flush to the finish surface.

Multiple-gang boxes mount to studs in the same way. But if plans offset the box away from studs or a multigang box is particularly wide, nail blocking between the studs or install an adjustable box bracket and screw the box to it ❹. (The bracket is also called a *screw gun bracket*, because a screw gun is required to mount it and to attach boxes to it.) The more securely a box is supported, the more secure the electrical connections will be.

TRADE SECRET

Most electricians prefer to screw boxes than to nail them because screws allow them to reposition boxes easily.

INSTALLING CEILING BOXES

INSTALLING A RECESSED CAN

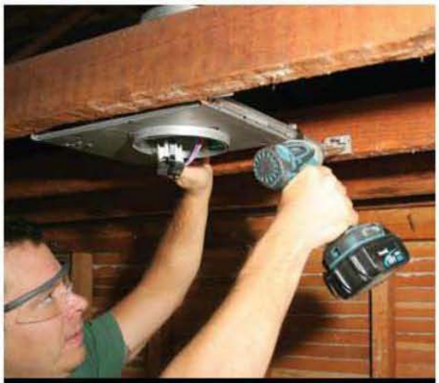

1 When the fixture is where you want it, fasten its hanger bars to the joists.

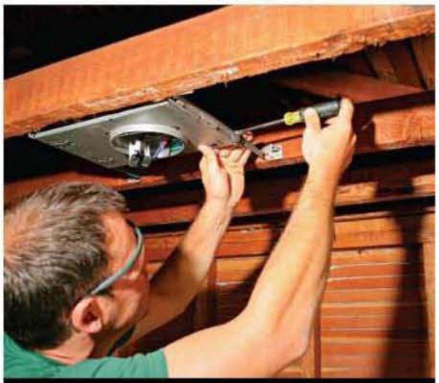

2 Slide the can to fine-tune its position. Tighten the screws to lock it in place.

INSTALLING A BOX TO THE JOIST

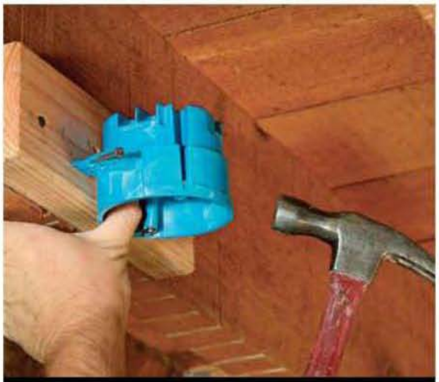

3 If a box position needs to be offset slightly from a joist, nail it to blocking.

INSTALLING A BOX BETWEEN JOISTS

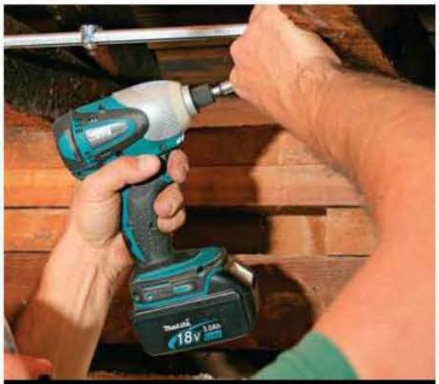

4 For boxes that occur midway between joists, use an adjustable bar hanger.

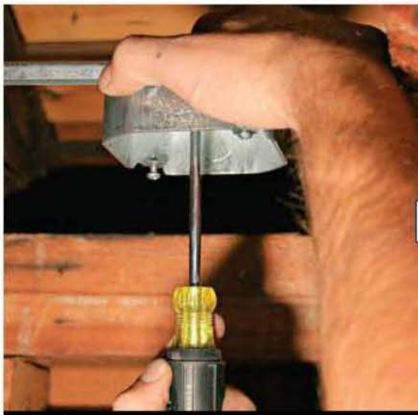

5 Screw a metal box to the hanger. Make sure the box edge is flush to the surface.

or

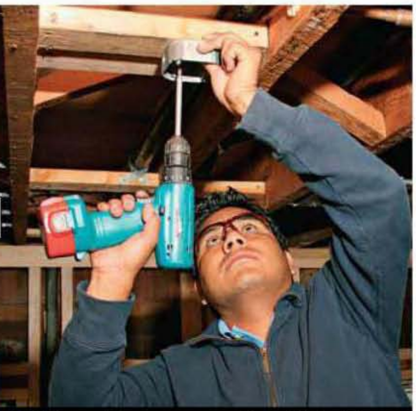

6 Alternately, you can nail 2x4 blocking between the joists.

Boxes for ceiling lights are most often 4-in. octagonal or round boxes or integral recessed light cans. Setting ceiling boxes in new work is similar to setting wall boxes, with the added concern that the ceiling box be strong enough to support the fixture weight. Many electricians prefer to use metal boxes for ceiling fixtures anyway. Ceiling fans require fan-rated boxes.

As noted earlier, you can mark ceiling box locations on the floor and use a laser to plumb up—or simply measure out from a wall.

→ **For more on laying out, see p. 184.**

In most cases, you'll need to reposition the box to avoid obstacles or line it up to other fixtures, but it's quick work if the box has an adjustable bar hanger. To install a recessed can, for example, extend its two bar hangers to adjacent ceiling joists. Then screw or nail the hangers to the joists ❶. Slide the can along the hangers until its opening (the light well) is where you want it and then tighten the setscrews on the side ❷.

To install a 4-in. box, simply nail or screw it to the side of a joist. If you need to install it slightly away from a joist, first nail 2x blocking to the joist, then attach the box to the blocking ❸. Remember: The box edge must be flush to finish surfaces.

To install a 4-in. box between joists, first screw an adjustable hanger bar to the joists ❹, then attach the box to it ❺. Alternately, you can insert 2x blocking between the joists and screw the box to it ❻.

→ **For a detailed look at installing hanger bars, see photo 2 on p. 85.**

DRILLING FOR CABLE

Once boxes are in place, you're ready to run cable to each of them. It's rather like connecting dots with a pencil line. To prevent screws or nails from puncturing cables, drill in the middle of studs or joists whenever possible. If the holes you drill are less than 1¼ in. from the edge of framing members, you must install steel nail-protection plates. Always wear eye protection when drilling.

Drill for cables running horizontally (through studs) first. It doesn't matter whether you start drilling at the outlet box closest to the panel or at the last outlet on the circuit. Just be methodical: Drill holes in one direction as you go from box to box. However, if you're drilling for an appliance that has a dedicated circuit—and thus only one outlet—it's usually less work to drill a hole through a top or bottom plate and then run cable through the attic or basement instead of drilling through numerous studs to reach the outlet.

If possible, drill holes thigh high ❶. Partially rest the drill on your thigh so your arms won't get as tired. This method also helps you drill holes that are roughly the same height—making cable-pulling much easier. Moreover, when you drill about 1 ft. above a box, you have enough room to bend the cable and staple it near the box without crimping the cable and damaging its insulation.

For most drilling, use a 6-in., ¾-in.-diameter bit. Use an 18-in. bit to drill lumber nailed together around windows, doorways, and the like ❷. Using an 18-in. bit is also safer because it enables you to drill through top plates without standing on a ladder. Bits that bind suddenly can throw you off a ladder ❸.

TRADE SECRET

Drilled holes don't need to be perfectly aligned, but the closer they line up, the easier it is to pull cable. Some electricians use a laser to line up drill holes.

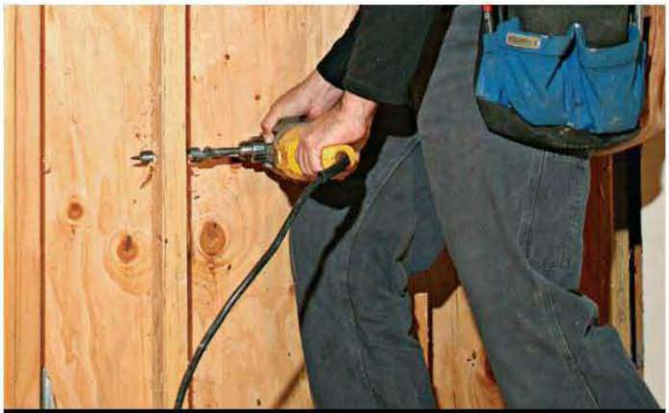

1 When drilling for cable runs, rest the drill on your thigh. This method eases drilling and places holes at a convenient height.

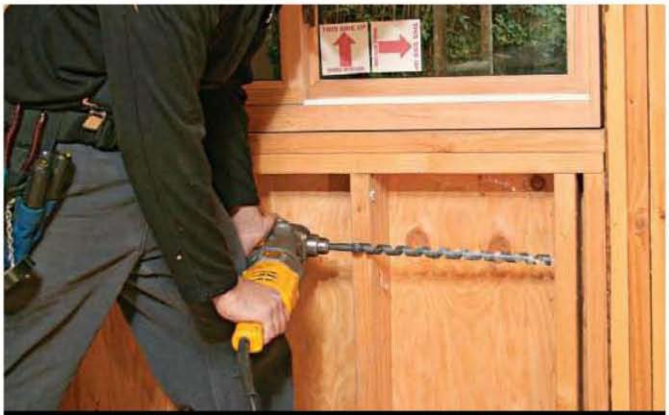

2 Use an 18-in. ship bit to drill through multiple studs.

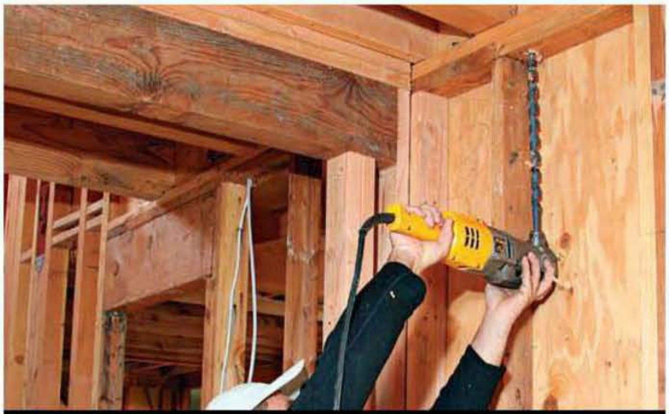

3 The 18-in. ship bit also allows you to drill through top wall plates while standing on the floor.

PULLING CABLE

For greatest efficiency, install cable in two steps: (1) Pull cable between outlets, leaving roughly 1 ft. extra beyond each box for future splices and (2) then retrace your steps, stapling cable to framing and installing nail-protection plates. As with drilling, it doesn't matter whether you start pulling cable from the first box of a circuit or from the last box. If there are several circuits in a room, start at one end and proceed along each circuit, pulling cable until all the boxes are wired. Don't jump around: you may become confused and miss a box.

In new construction, electricians usually place several wire reels by the panel and pull cables from there toward the first box of each circuit ❶. Once they've run cable to all these boxes, they move a reel next to each box and continue to pull cable outward until they reach the last (farthest) box ❷. When doing remodel wiring in a house with a crawl space, however, electricians often start at the last box and pull cables toward the panel. When they reach the first boxes of several circuits, they will move the wire reels to those locations. From there, they feed, say, three cables down to a helper in the crawl space. The helper can pull all the cables toward the panel at the same time. This method is much faster than pulling single cables three different times.

Staple cable along stud centers to prevent nail or screw punctures. It's acceptable to stack two cables under one staple ❸, but use standoffs ❹ to fasten three or more cables traveling along the same path. (Multigang boxes are fed by multiple cables, for example.) Standoffs and ties bundle cables loosely to prevent heat buildup. As you secure cable, install nail plates where needed ❺.

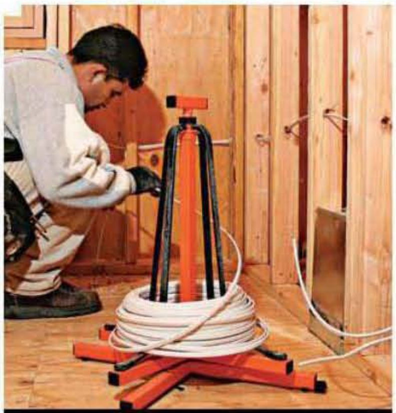

1 Place a wire wheel near the beginning or end of each circuit.

2 Avoid crimping cable as you pull it through narrow spaces.

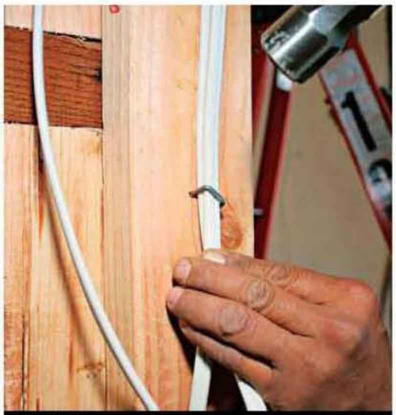

3 Drive staples just snug to the cable.

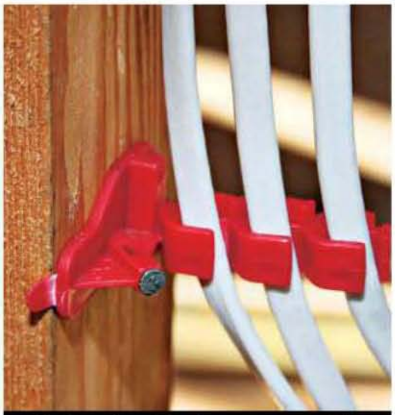

4 When bundling two or more cables, use standoffs.

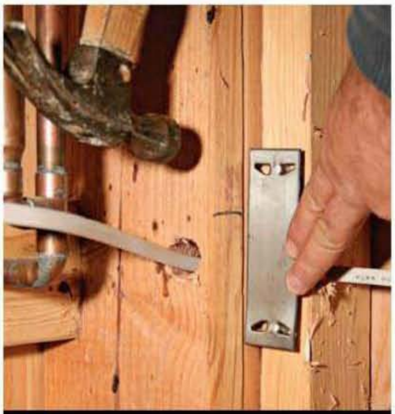

5 For cable within $1\frac{1}{4}$ in. of a stud edge, install steel plate.

TRADE SECRET

Put a piece of red tape on the first box in each circuit to ensure that you run cable from it to the panel. On a complex job with many circuits, you might run cable *between* all the outlets in a circuit but forget to install the home run cable that will energize the circuit. Not something you want to discover after the drywall's up.

FEEDING CABLE THROUGH CORNERS

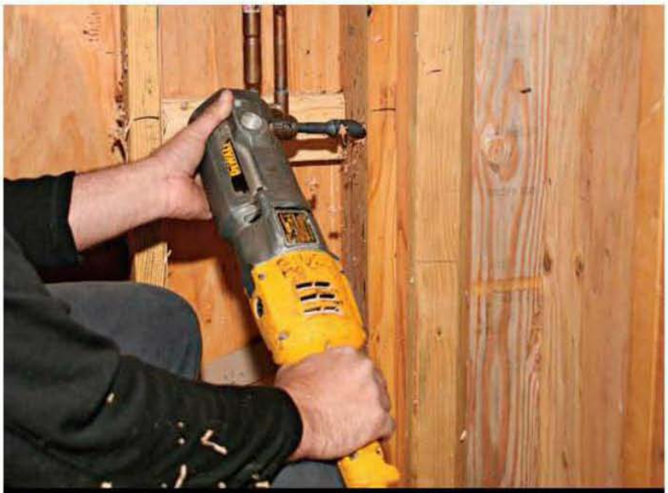

1 To feed cable through a corner, drill a level hole on one side of the built-up studs.

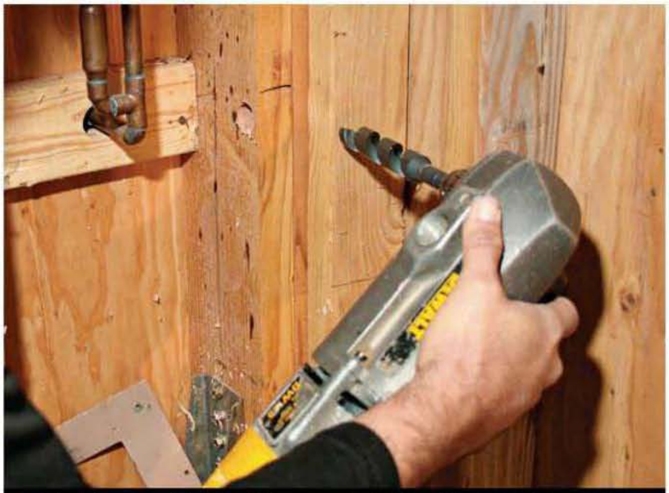

2 Drill a second hole, at the same height, so the holes intersect.

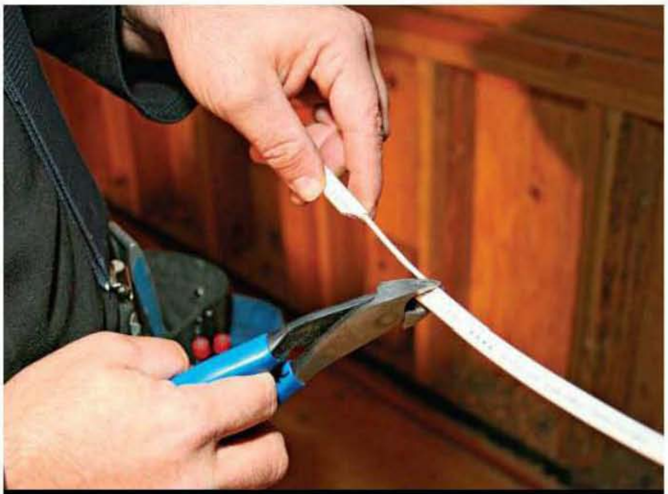

3 Use diagonal cutters to cut through the sheathing and two of the three wires in the cable, then loop the end of the remaining wire.

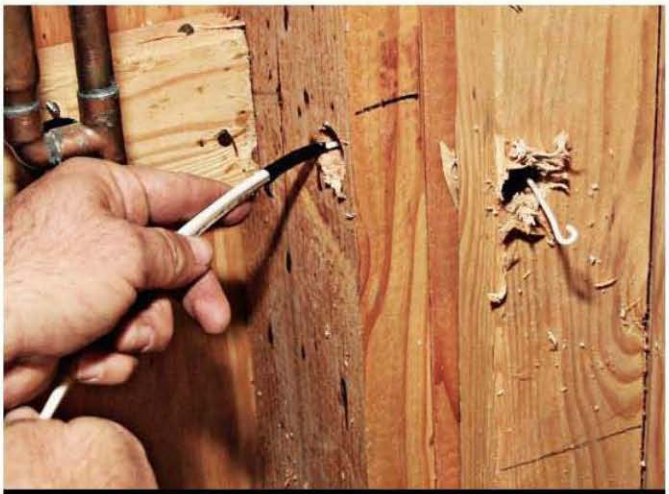

4 Wrap the end of the cable so it slides through the hole easier, then work the looped wire through the holes.

Corners are usually built by nailing three or four studs together, so feeding cable through them can be tricky. Drill intersecting holes at the same height, as shown in photos ❶ and ❷. Use diagonal cutters to cut through the sheathing and two of the wires, thus leaving a single wire protruding from the cable ❸.

Use pliers to loop the end of the wire. Then wrap tape around the end of the cable and onto the wire to create a tapering point that will slide easier through the hole. Push wire through the corner holes until it emerges from the other side ❹. If the wire gets hung up midway through, insert a finger from the other side to fish for its end.

FISHING CABLE BEHIND FINISH WALLS

Most electricians hate fishing wire behind existing walls. It can be tricky to find the cable and time consuming to patch the holes in plaster or drywall. If you're adding a box or two, try fishing cable behind the wall. But if you're rewiring an entire room, it's probably faster to cut a "wiring trench" in the wall. Before cutting into or drilling through a wall, however, turn off power to the area.

For more on cutting a wiring trench, see p. 196.

If you're adding an outlet over an unfinished basement, fishing cable can be straightforward. Outline and cut out an opening for the new box, insert a flexibit into the opening, then drill down through the wall's bottom plate ❶. When the bit emerges into the basement, a helper can insert one wire of the new cable into the small "fish hole" near the bit's point. As you slowly back the bit out of the box opening, you pull new cable into it ❷. No fish tape required! The only downside is that the reversing drill can twist the cable. This problem is easily avoided by sliding a swivel kellum over the cable end instead of inserting a cable wire into the flexibit hole ❸. Because the kellum turns, the cable doesn't.

Alternatively, you can start by removing a wall box. The closest power source is often an existing outlet. Cut power to that outlet and test to make sure it's off. The easiest way to access the cable is to disconnect the wires to the receptacle ❹ and remove it. Then remove the box ❺, which may require using a metal-cutting reciprocating-saw blade to cut through the nails holding the box to the stud. Fish a new cable leg to the location and insert the new and old cables into a new *cut-in box*. Secure the cut-in box to the finish surface, splice the cables inside the box, and connect pigtails from the splice to the new receptacle.

>> >> >>

RUNNING CABLE TO A NEW OUTLET

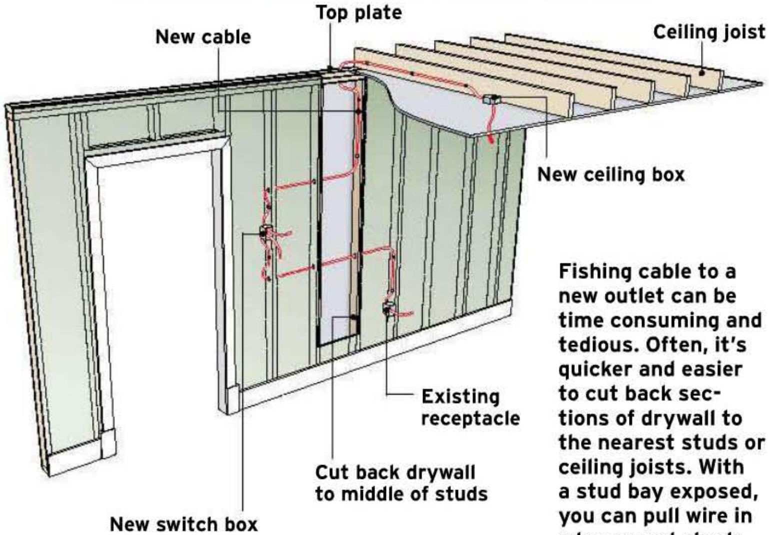

Fishing cable to a new outlet can be time consuming and tedious. Often, it's quicker and easier to cut back sections of drywall to the nearest studs or ceiling joists. With a stud bay exposed, you can pull wire in a hurry and staple cable to studs.

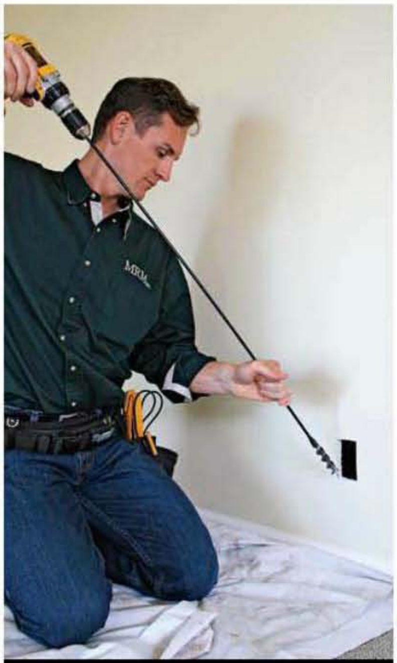

1 To fish cable to a new outlet, insert a flexibit into the new opening and drill.

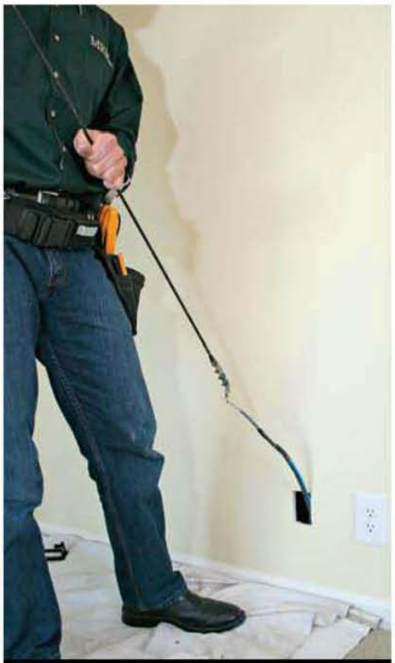

2 Pull back after a helper below the floor attaches the cable.

FISHING CABLE BEHIND FINISH WALLS (CONTINUED)

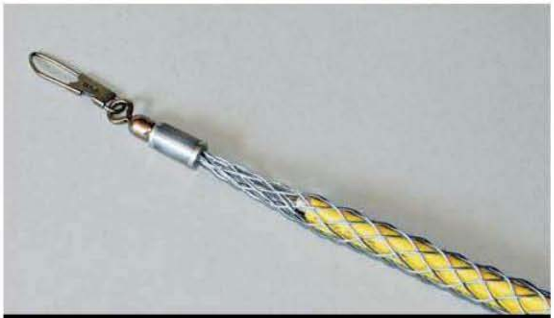

3 A swivel kellum slides over the end of the cable being fished so the cable doesn't twist.

or

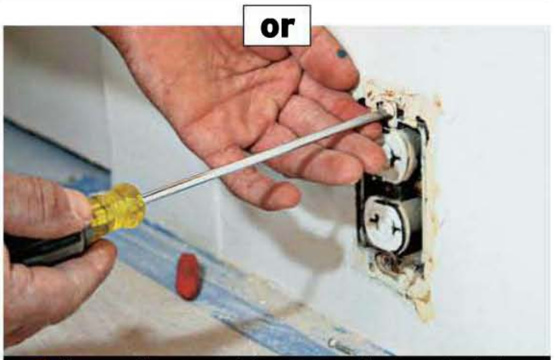

4 Remove the cover plate, unscrew the receptacle from the box, and pull it out. Then disconnect the wires.

5 Remove the old outlet box, fish new cable, and splice the old cable to the new. You'll replace the old box.

FISHING CABLE TO A CEILING FIXTURE

1 To run cable to a ceiling fixture, start by drilling an exploratory hole with a small-diameter bit.

Fishing cable to ceiling fixtures or wall switches is usually a bit complicated. If there is an unfinished attic above or a basement below, run the cable across it, then route the cable through a stud bay to the new box. To run cable to a ceiling light, for example, drill up through the fixture location using a 1⁄8-in. by 12-in. bit to minimize patching later ❶. Use a bit at least 6 in. long so a helper in the attic can see it—longer if the floor of the attic is covered with insulation. Measure the distance from the bit to the wall; a helper in the attic can use that measurement to locate the nearest stud bay to drill an access hole into. If you're working alone, loop the end of a stiff piece of wire about 1 ft. long ❷ and insert it in the drilled hole; friction will keep the wire upright in the hole until you can locate it in the attic.

If there's no access above the ceiling and/or cable must cross several ceiling joists to get from a switch to a light fixture, you'll have to cut into finish surfaces at several points ❸. To access cable in a stud bay, you'll need a cutout to expose the top plate. Using a flexibit may minimize the number of holes you must cut to drill across ceiling joists. But as noted earlier, it may ultimately take less time to cut and repair a single slot running across several joists than to patch a number of isolated holes. Whatever method you choose, make cuts cleanly to facilitate repairs. First outline all cuts with a utility knife.

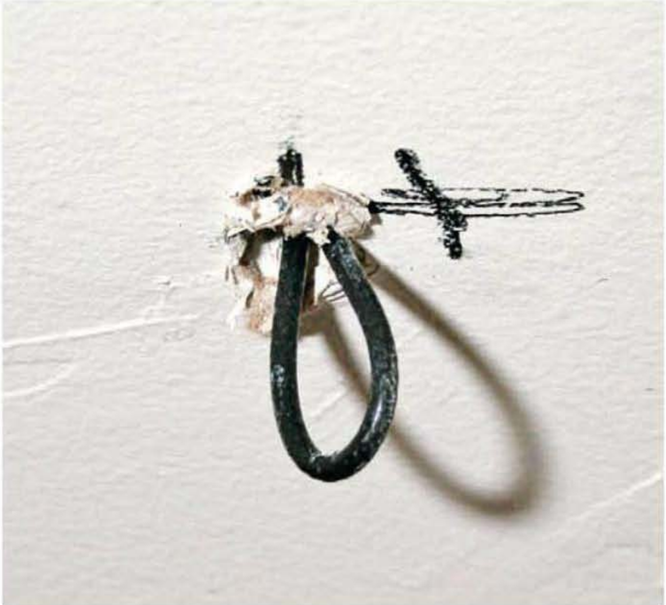

2 If you're working alone, jam a long, looped wire into the exploratory hole, then go aloft to look for it.

3 Running cable from a ceiling fixture to a wall switch means cutting access holes.

FISHING CABLE ACROSS CEILING JOISTS

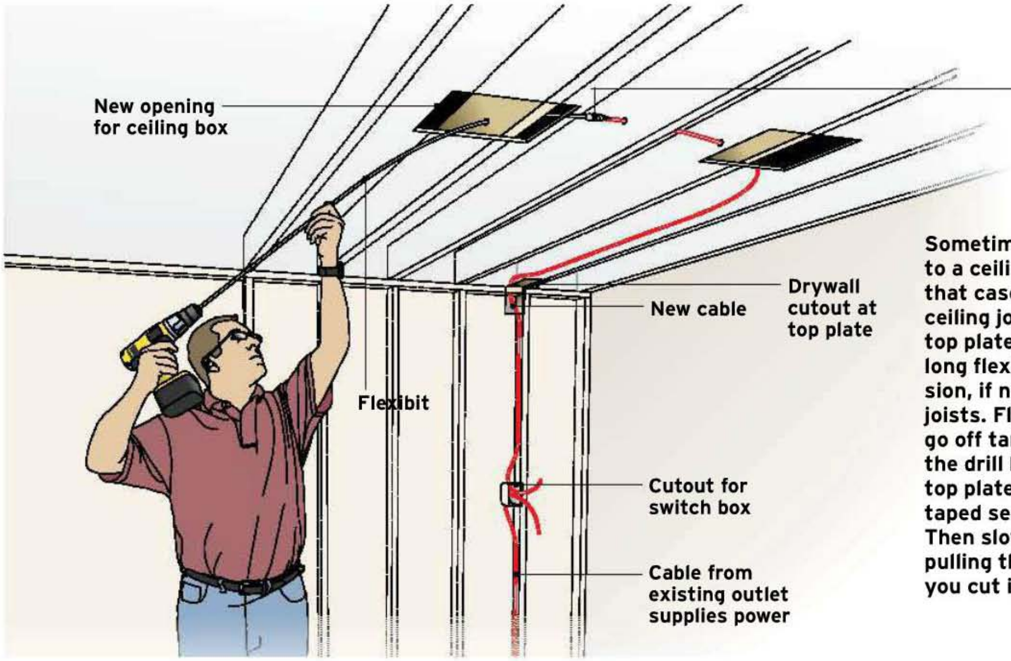
New opening for ceiling box
Drill across ceiling joists
Drywall cutout at top plate
New cable
Flexibit
Cutout for switch box
Cable from existing outlet supplies power

Sometimes there is no open bay to a ceiling fixture location. In that case, you'll have drill across ceiling joists to run cable to the top plate of the wall. Use a 6-ft.-long flexibit (and a 48-in. extension, if needed) to drill across joists. Flexibits can wander and go off target, so be patient. When the drill bit emerges above the top plate, attach a swivel kellum taped securely to the new cable. Then slowly back the drill bit out, pulling the cable to the opening you cut in the ceiling.

CUTTING A WALL BOX INTO PLASTER

If wiring is in good condition and an existing circuit has the capacity to add an outlet, turn off the power, cut a hole in the wall, fish cable to the location, and secure a cut-in box to the finish surface. This process is called *cutting-in* or retrofitting a box.

Hold the new box at the same height as other outlet or switch boxes and trace its outline onto the wall. Use a stud finder or drill a small exploratory hole to locate studs or wood lath behind. If you hit a stud, move the box. If you hit lath, keep drilling small holes within the opening to find the edges of the lath. If you position the box correctly, you'll need to remove only one lath section ❶.

Use a utility knife to score along the outline to minimize plaster fractures. Remove the plaster within the outline using a chisel. Then cut out the lath, using a cordless jigsaw ❷. As you cut through the lath strip, alternate partial cuts from one side to the other to avoid cracking the plaster. Then carefully remove the plaster beneath the box ears, so they can rest on lath. Before inserting cut-in boxes, remove box knockouts, insert cable clamps, strip sheathing off the ends of incoming cable, and feed cable into the cable clamps. If more than one cable enters the box, write the destination of each on the sheathing. Secure the box by screwing its ears to the lath ❸.

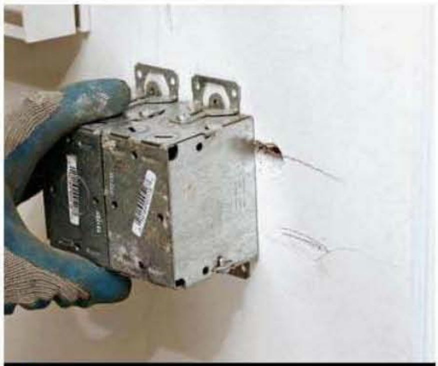

1 Position the box and then trace the box outline onto the wall.

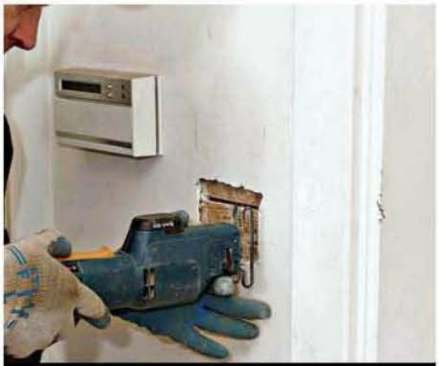

2 After you chisel out the plaster within the outline, use a sabersaw to cut lath.

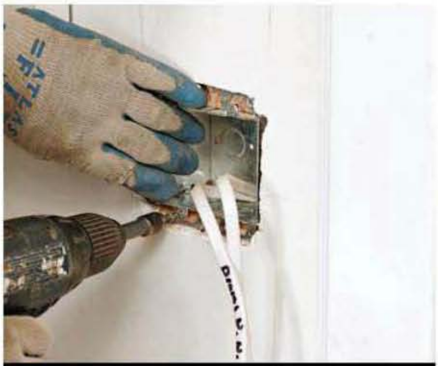

3 After fishing and pulling cable, secure the box to the wall.

CUTTING A WALL BOX INTO DRYWALL

Adding a cut-in box to drywall is essentially the same as adding one to plaster. Start by drilling a small exploratory hole near the proposed box location to make sure there's no stud in the way.

There are a number of cut-in boxes to choose from; most common is a type with side-mounted ears that swing out or expand as you turn its screws.

➔ **For more on box types, see p. 25.**

Hold the box against the drywall, plumb one side, then trace the outline of the box onto the wall ❶. Drywall is much easier to cut than is plaster: Simply align the blade of a drywall saw to the line you want to cut and hit the handle of the saw with the heel of your hand.

There is no one right way to cut out the box, but pros tend to cut one of the long vertical sides, then make three horizontal cuts across. Then score and snap the last cut ❷. Finish the cutout with the drywall saw and a utility knife.

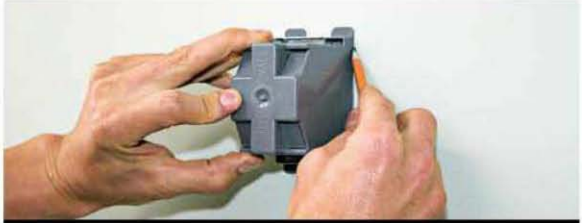

1 After drilling an exploratory hole and locating the box, trace the box outline onto the wall.

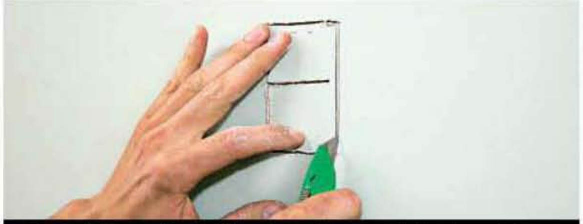

2 Cut three sides with a drywall saw; cut the last side with a utility knife.

RETROFITTING A CEILING BOX

As with all retrofits, turn off power to the area and explore first. Follow the mounting recommendations for your fixture. Attach the fixture box to framing.

If there's insulation in an attic above, remove it from the affected area. Be sure to wear eye protection and a dust mask when drilling through any ceiling—it's a dusty job.

Mark the box location and use a fine-tooth hole saw to cut through plaster or drywall ❶. Place the centering bit of the hole saw on the exact center of the box opening. Drill slowly so you don't damage adjoining surfaces—or fall off the ladder.

If the ceiling is drywall, you're ready to run cable through the hole in the ceiling. If the ceiling is plaster, cut through the lath or leave the lath intact and screw a pancake box through the lath and into the framing. Before attaching a pancake box, remove a knockout, test-fit the box in the hole, and trace the knockout hole onto the lath. Set the box aside, and drill through the lath, creating a hole through which you can run cable ❷.

Feed cable to the location and fit a cable connector into the box. Insert the cable into the connector, slide the box up to the ceiling ❸, and secure it ❹. Strip the cable sheathing and attach the ground wire to a ground screw in the box. Strip insulation from the wire ends and you're ready to connect the light fixture.

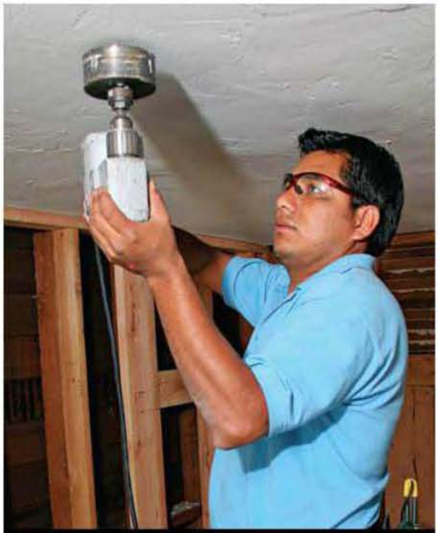

1 Mark the box center onto the ceiling, then use a fine-tooth hole saw.

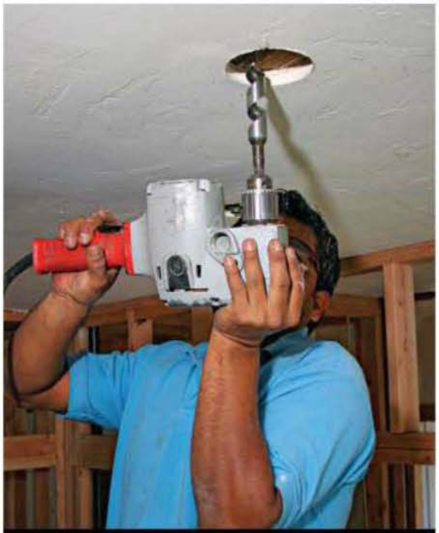

2 Use an auger bit to drill a hole through the lath so you can pull cable to the box.

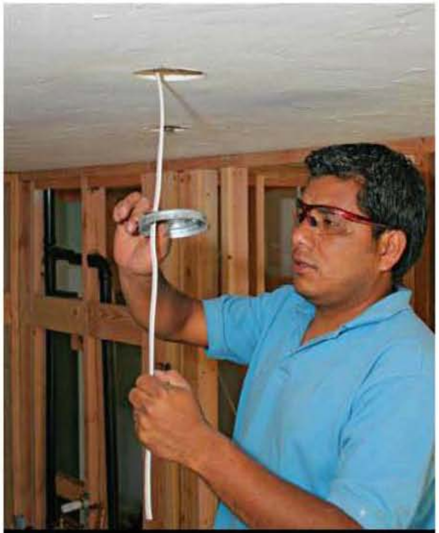

3 Fish cable through the lath and feed it through a connector in the box.

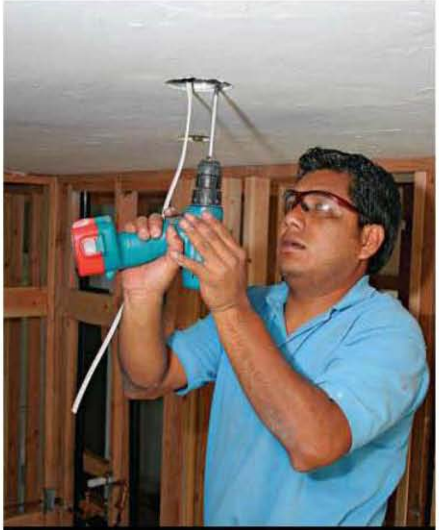

4 Screw the pancake box to the lath, strip the cable, and prep the wires.

TRADE SECRET

If you need to splice two or more cables, use a 4-in. octagonal or round box instead of a pancake box.

CREATING A WIRING TRENCH

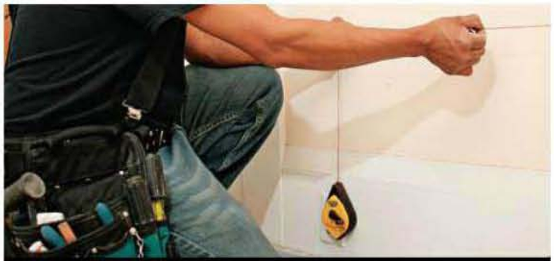

1 Use a chalkline to mark the top and bottom lines of the wiring trench.

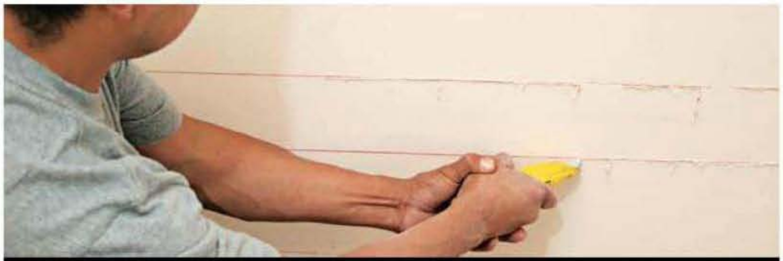

2 To ensure a clean cut to line and make patching easier when you're done, use a utility knife to score along the chalklines.

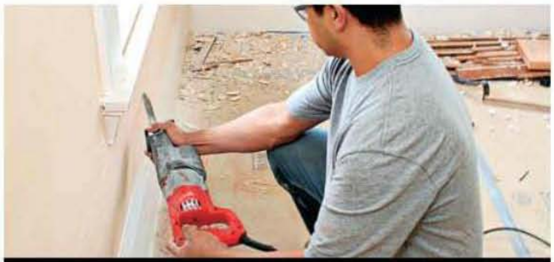

3 Cut along the chalklines, holding the reciprocating saw at a low angle. Do not cut the lath.

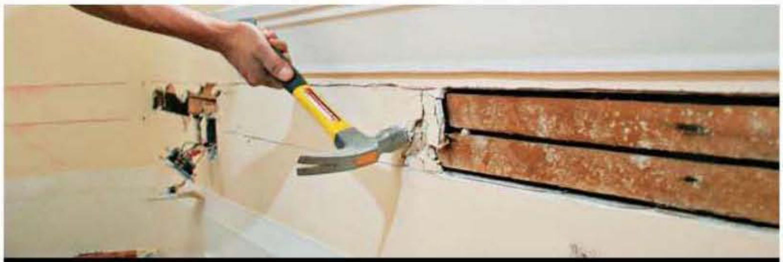

4 If walls are drywall, simply pull out the isolated strips. If plaster, use a hammer to remove the plaster and expose the lath.

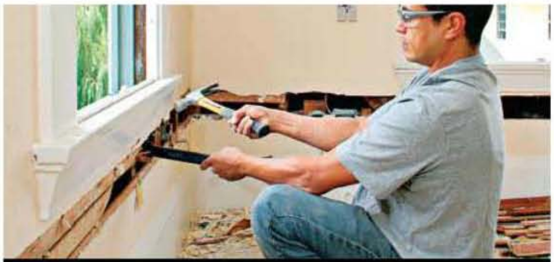

5 Use a hammer and a flat bar to pry the lath free from the studs. Work slowly to minimize damage.

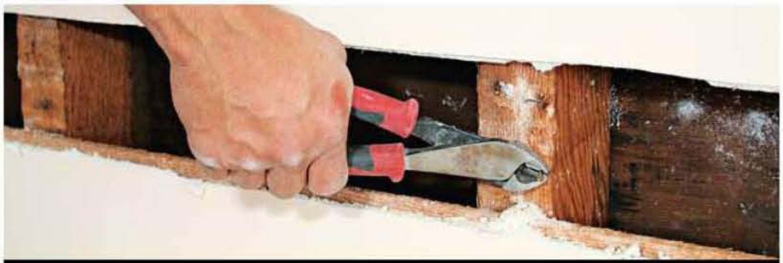

6 Pull out any nails still stuck in the studs; this will make later repairs easier.

When adding multiple outlets or rewiring an entire room, cutting a wiring trench in finish surfaces instead of fishing cable behind them is much faster. Before cutting or drilling, however, turn off the power to the areas affected. And be sure to wear eye protection and a dust mask.

If there are no windows in the walls to be rewired, cut the trench about 3 ft. above the floor so you won't have to kneel while working. If there are windows, cut the trench under the windows, leaving at least 1 in. of wall material under the windowsill to facilitate repairs. If there's plaster, make the trench as wide as two strips of lath. Snap parallel chalklines to indicate the width of the trench ①. Then use a utility knife to score along each line ②. Scoring lines first produces a cleaner cut and easier repairs.

Next use a reciprocating saw with a demolition blade to cut through the plaster or drywall. Hold the saw at a low angle: You'll be less likely to break blades or cut into studs ③. Using a hammer, gently crush the plaster between the lines ④. Use a utility bar (flat bar) to pry out the lath strips or drywall section. If you expose any cables in the walls, use an inductance tester to make sure they're not hot ⑤. Next, drill through the studs so you can run cable in the trench. Wherever there's an outlet indicated, expand the trench width to accommodate the boxes.

Finally, walk along the trench and pull any lath or drywall nails still sticking in stud edges ⑥. They're easy to overlook because they're small; if you pull them now, patching the trench will go smoothly.

RETROFITTING BOXES & PULLING CABLE

Once you've cut a wiring trench and drilled the holes, installing the boxes and pulling cable are straightforward and much like the basic sequences shown in Chapter 2.

➔ **For more on installing boxes, see p. 27.**

➔ **For more on pulling cables, see p. 189.**

Most electrical codes allow you to install either plastic or metal boxes in residences, but you must use metal boxes in commercial buildings.

If you're installing metal boxes, remove knockouts and insert cable connectors into their openings ❶. Then screw boxes to studs; screwing is less likely to damage nearby finish surfaces ❷. Be sure that the box will be flush to finish surfaces or, if you'll install plaster rings later, flush to the stud edge. Whenever you install boxes side by side—as with the outlet and low-voltage boxes shown here—install them plumb and at the same height ❸.

Installing cable in remodels can be tricky because space is tight and you must avoid bending cable sharply, which can damage wire insulation ❹. Install nail plates wherever the cable is less than 1¼ in. from the stud edges. Feed cable through the cable connectors into the boxes ❺. Finally, staple the cable to the framing within 12 in. of each box. If there's not room to loop the cable and staple it to a stud, it's acceptable to staple it to other solid framing, such as the underside of a sill ❻.

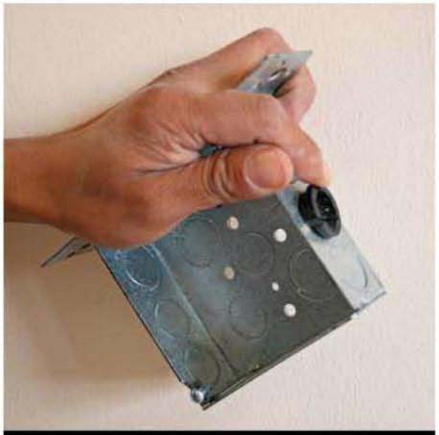
1 For metal boxes, start by removing knockouts and then insert a connector.

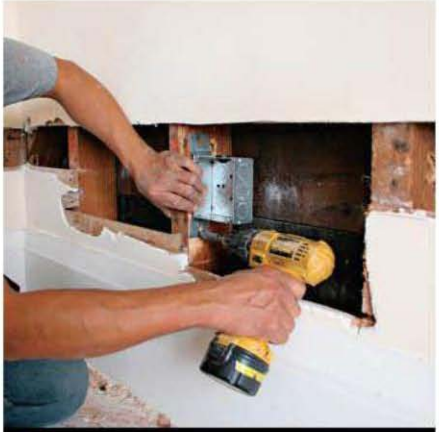
2 Screw the box to the side of a stud.

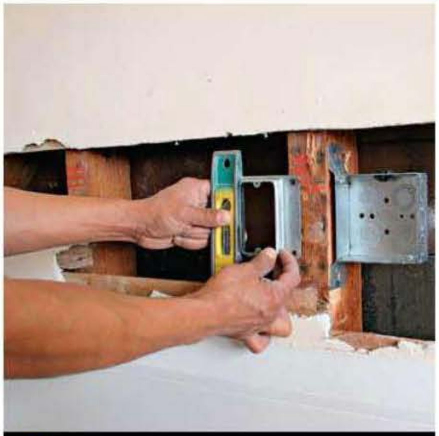
3 For the best appearance, install outlets that are side by side at the same height.

4 Run cable to each box.

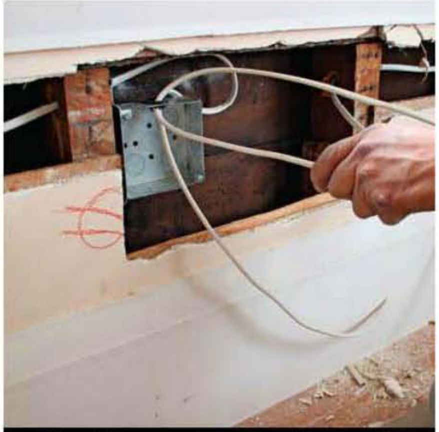
5 Feed the cable through cable connectors until you have about 1 ft.

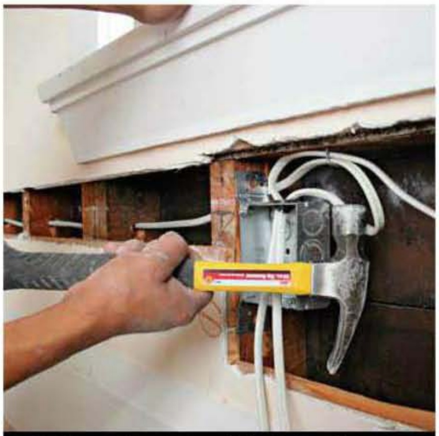
6 Staple the cable within 12 in. of each box.

MAKING UP AN OUTLET BOX

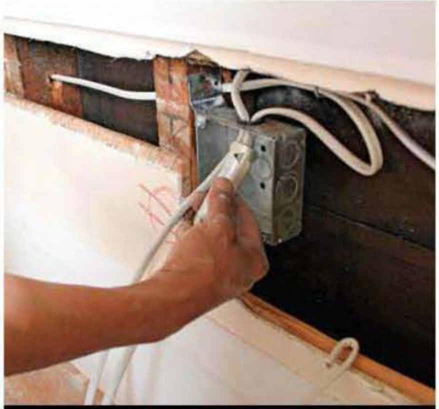

1 Use a cable ripper to slit sheathing, diagonal cutters to cut it.

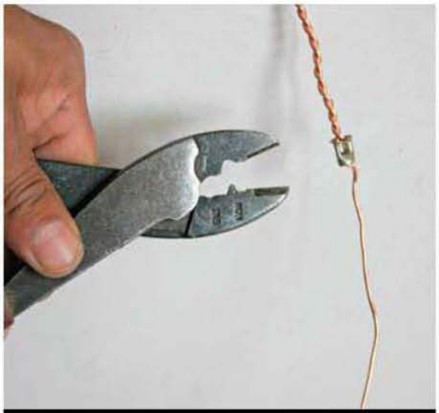

2 A crimp tool may be used instead of a wire nut for the ground.

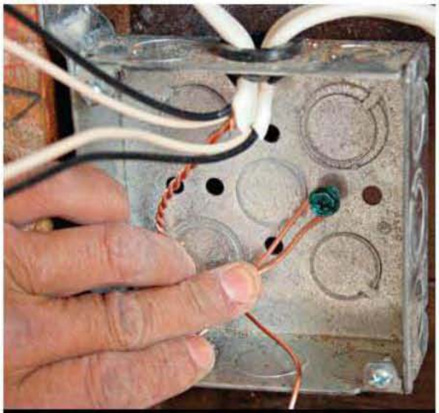

3 The long leg of the ground is looped beneath the grounding screw.

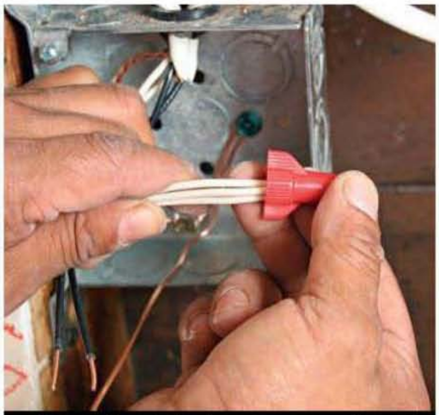

4 Strip insulation from the ends of individual wires and splice like wires together.

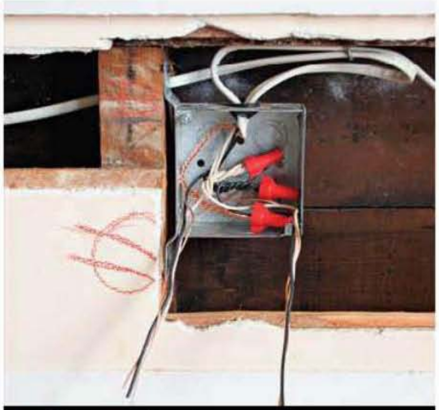

5 If there will be two receptacles in a box, make up two sets of pigtails.

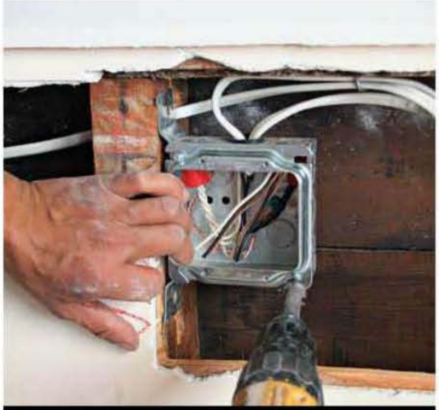

6 Fold the wires into the box and install a plaster ring.

Electricians call the last stage of rough wiring *making up a box*. Here, you'll remove sheathing from cables inside the boxes, group like wires, splice grounds, and—if the box is metal—attach ground wires to a ground screw. At this point, many electricians simply splice all wire groups because it will save time later.

First use a cable ripper ❶ to remove the cable sheathing. Leave a minimum of 1⁄4 in. of sheathing inside the box and a maximum of 1 in. Electricians favor utility knives to remove sheathing, but nonpros are less likely to nick wire insulation with a ripper. Once you've removed the sheathing, separate the insulated wires and bare ground wires.

There are several ways to splice grounds. You can use a twist-on wire connector (Wire-Nut is one brand) and run a pigtail to the ground screw. Or you can twist the bare wires and crimp them ❷. If you go the latter route, leave only one ground wire sticking out of the crimp, which you'll wrap around the ground screw ❸.

Next, use a wire stripper to remove 1⁄2 in. of insulation from the wire ends. Use wire connectors to splice all neutral wires together and all hot wires together ❹. As we recommended earlier, splice a short pigtail to each wire group as well. Attaching neutral and hot pigtails to the receptacle screw terminals ensures continuity power even if a receptacle fails. If the box will contain two receptacles, create two groups of pigtails ❺.

Accordion-fold the wire groups into the box and you're ready for inspection. If necessary, install a mud ring to bring the box opening flush to finish surfaces. The mud ring should be installed before you've repaired the plaster or drywall ❻.

MAKING UP A LIGHT CAN

Wiring a recessed light can or a ceiling box is essentially the same as making up an outlet box, although light cans frequently have stranded wire leads. Always follow the installation instructions supplied with your lighting unit.

Run cable to the light can, stapling it at least every 4½ ft. to the side of the ceiling joist and within 12 in. of the box. Also, make sure that the cable is at least 1¼ in. from the joist edge, so drywall screws or nails can't puncture it. Because you'll be working over your head and junction boxes are cramped, you may want to remove the sheathing before feeding cable into the box ❶. Tighten the cable connector to the small section of sheathing sticking into the box. Then strip insulation from individual wires ❷.

Group ground wires, neutral wires, and hot wires. Then splice each group using wire connectors. When wiring light cans, electricians typically start by splicing ground wires and running a ground pigtail to a ground screw on a metal box ❸. They then splice the neutral wires and, finally, the hot wires. When all wire groups are spliced, fold them carefully into the junction box and attach the cover.

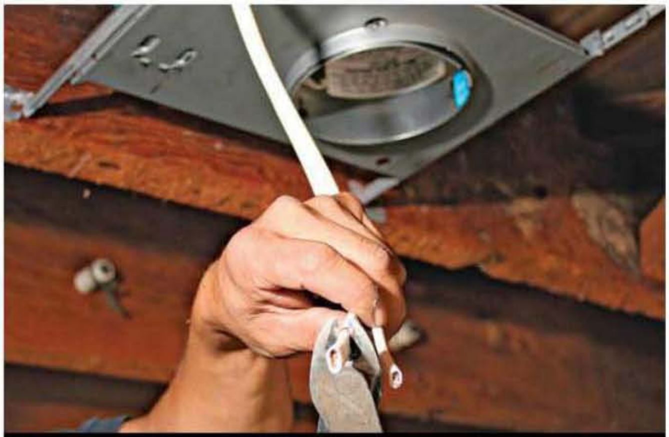

1 Run enough cable to the light fixture so that there's about 1 ft. sticking out of its junction box.

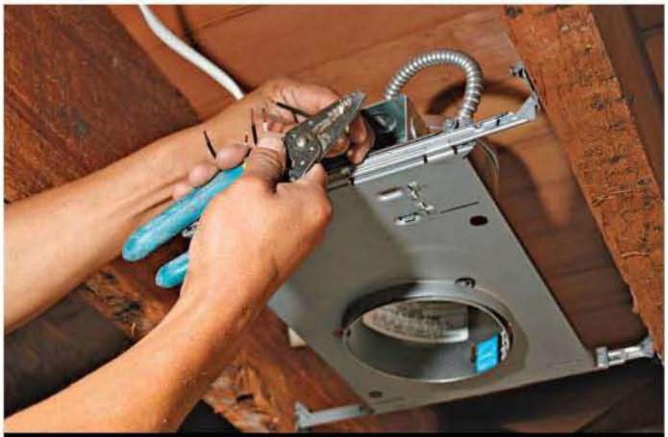

2 Feed the cable into the junction box, remove the sheathing, and strip the insulation from the ends of individual wires.

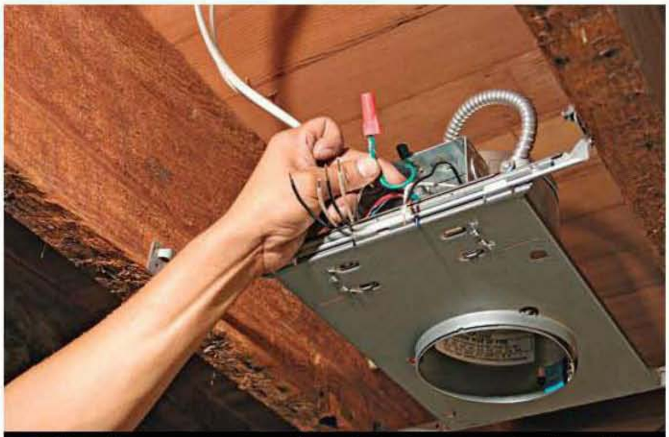

3 Use wire connectors to splice paired wire groups, starting with the ground wire group. Fold the spliced wires into the box.

TRADE SECRET

The ground screw must compress the ground wire evenly. Never cross the ground wire over itself because the screw would touch only that high spot.

MAKING UP A SINGLE SWITCH BOX

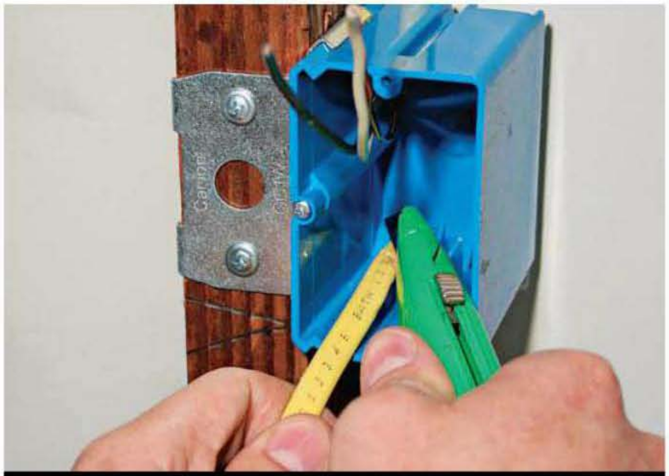

1 **Score the sheathing lightly on both sides and give it a quick tug to slide it off the wires.**

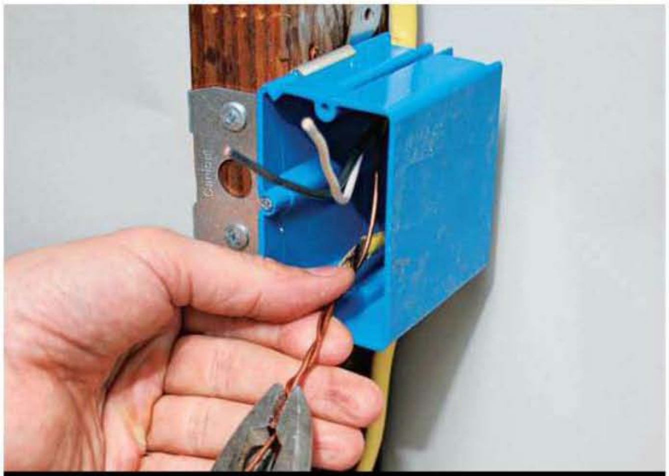

2 **Use lineman's pliers to twist the grounds together. Leave one ground longer to be attached to the device.**

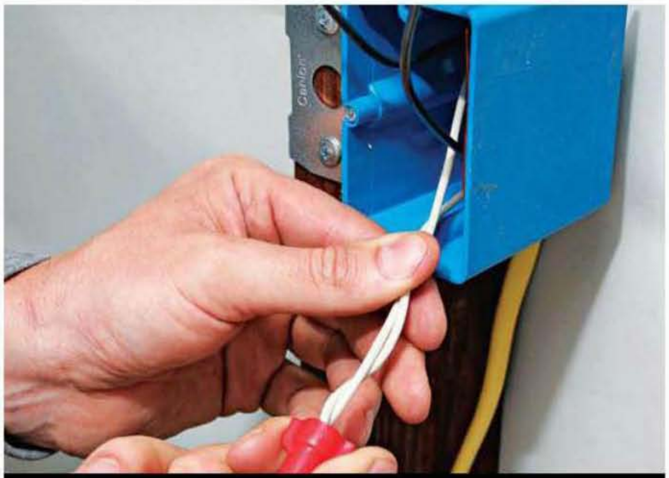

3 **Use a twist-on wire connector to splice the neutrals. Do not splice the hot wires; they are attached to the switch or receptacle later.**

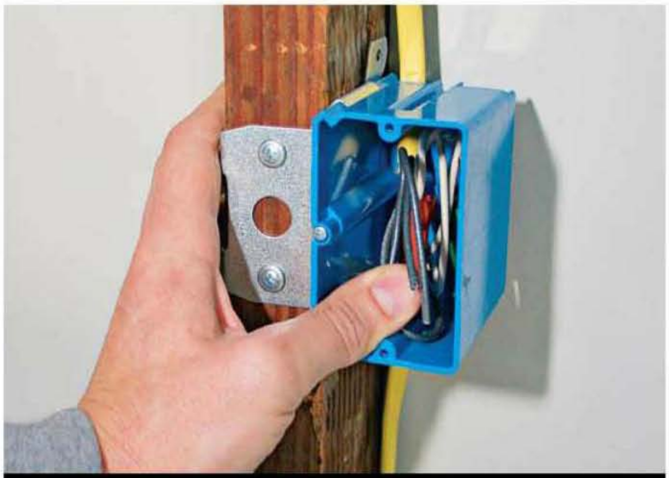

4 **Fold the wires in an accordion pattern and tuck them into the box.**

Switch boxes are most commonly plastic, so begin by using a screwdriver or needle-nose pliers to remove its knockouts. Feed cable (or cables) into the knockout and staple it within 12 in. of the box. There's no need to use cable connectors—plastic boxes have an integral spring clamp inside the knockout to prevent the cable from being yanked out. Leave about 1 ft. of extra cable sticking out of the box.

If there's a single cable entering the switch box, the switch is at the end of a switch leg and both insulated wires are hot wires. Remove the cable sheathing, strip insulation from the insulated wires, and wrap a piece of black electrical tape around the white wire to indicate that it is hot. Then tuck the wires into the box.

If there are two cables entering the switch box, one is incoming power and the other is a switch leg. Remove the cable sheathing ❶, then separate and twist the grounds together before splicing them with a wire connector ❷. Strip insulation from the ends of the neutral wires and splice them ❸. Fold the wires into the box ❹.

MAKING UP A MULTIGANG SWITCH BOX

The tricky part of making up a multigang switch box is keeping track of the wires. As you feed each cable into the box, use a felt-tipped marker to identify each cable's origin or destination. Write directly on the cable sheathing. Start by feeding the incoming cable into the box and marking it *hot*. This incoming cable will supply 110v from a panel or from an outlet upstream when the cable is connected later.

Next, nail a standoff to the side of the stud so you can secure each switch-leg cable as you feed it into the box ❶. Before inserting a cable into a knockout, however, note which fixture or device the switch leg controls ❷. Be specific and use labels such as *sconce*, *can ctr* (center can), and *perim* (perimeter fixture).

As you pull each cable into the box, strip the sheathing and cut off the small section that has writing on it ❸. When all the cables have been stripped, separate and twist the bare ground wires clockwise. Leave one ground about 6 in. longer than the others, then feed it through the hole in the end of a special ground wire connector ❹. This longer ground wire will run to a ground screw on each switch.

Strip the insulation from the ends of the neutral wires; then use a wire connector to splice them together ❺. Fold both ground- and neutral-wire groups into the box.

Cut a hot-wire pigtail 8 in. to 9 in. long for each switch and splice all pigtails to the hot wire of the incoming cable—the one that was earlier marked *110v* ❻. Finally, pair a hot pigtail with each switch leg ❼. Fold the wire pairs into the box. When it's time to wire the switches later, you'll connect a hot pigtail to one screw terminal and a switch-leg wire to the other screw terminal.

If a cable serves a three-way or four-way switch, indicate whether a wire is a "traveler" or "common."

For more on three- and four-way switches, see "Receptacles and Switches," p. 38.

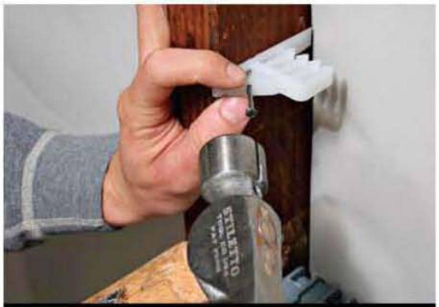

1 Nail a standoff to secure more cables.

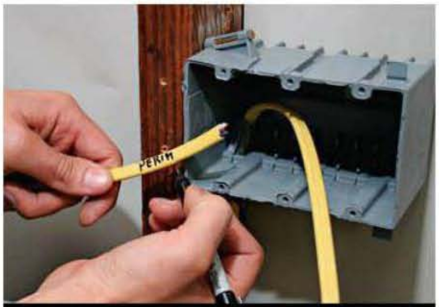

2 Before pulling each cable into the box, indicate which fixture it controls.

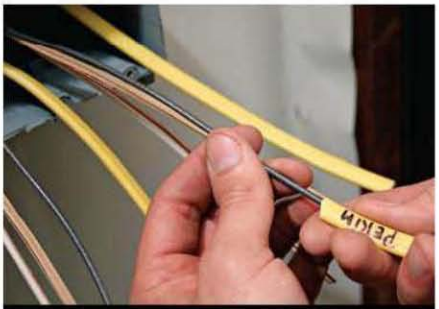

3 Score and remove the sheathing. Slide the label back on.

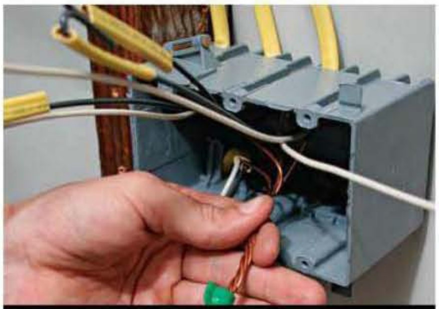

4 Feed the ground wire through the hole in the connector, then twist the connector.

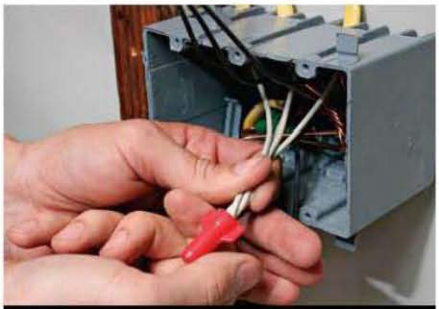

5 Twist on a wire connector to splice the neutral wire group.

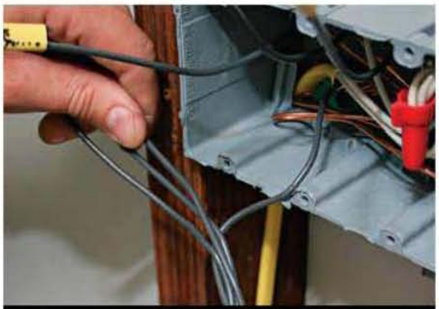

6 Cut a 8-in. hot-wire pigtail for each switch in the box. Use a wire connector.

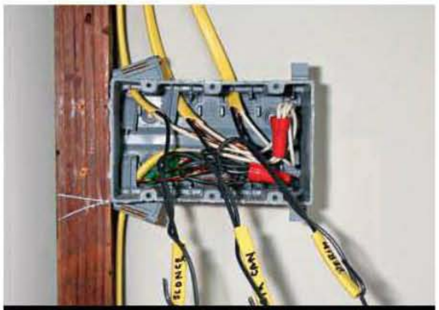

7 Loosely wrap a hot pigtail around each switch-leg wire. Then fold into the box.

FLEXIBLE METAL CABLE

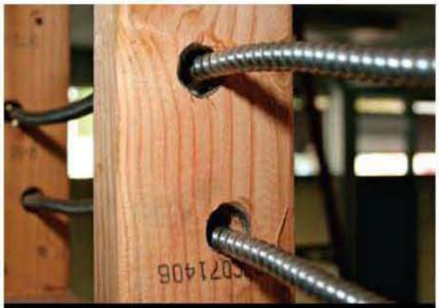

1 **Flexible metal cable must be at least 1¼ in. from stud edges.**

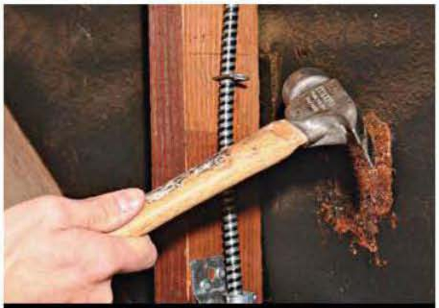

2 **Staple AC or MC cable within 12 in. of the box and 4½ ft. between boxes.**

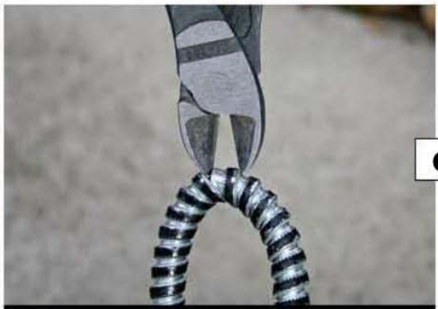

3 **To cut AC or MC cable, bend the cable till the jacket splits, then snip.**

or

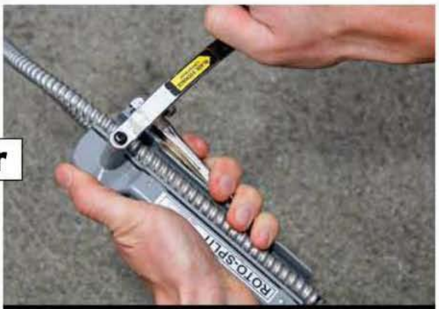

When using a Roto-Split, follow manufacturer's instructions.

In residences, MC and AC cable is most often used in short exposed runs—typically between a wall box and the junction box of an appliance. The cable's metal jacket protects the wiring inside, but you should still take pains to avoid puncturing MC or AC cable with a nail or a screw. For this reason, when flexible metal cable is run through stud walls, it's a good practice to run it through the center of the stud ❶. If several cables run through a stud, stack holes vertically to avoid compromising the strength of the stud. Use nail-protection plates if the cable is closer than 1¼ in. to a stud edge.

➔ For more on nail-protection plates, see p. 189.

Secure or support flexible metal cable within 12 in. of boxes ❷ and at least every 4½ ft. along the span of the run—electricians typically staple it every 3 ft. Flexible metal cable should also be secured to the underside of every joist it crosses or run through holes drilled through joists. Again, use nail-protection plates if the cable is less than 1¼ in. from a joist edge.

AC and MC cable require specialized connectors to secure them to boxes. You can use a hacksaw with a metal-cutting blade or diagonal cutters to cut through cable's metal jacket ❸, but a Roto-Split is the tool of choice. Whatever tool you use, cut through only one coil of the metal jacket to sever it. To prevent damaging the wire insulation within, cut no deeper than the thickness of the metal jacket. After cutting through the metal jacket, insert plastic antishort bushings to protect the wire insulation from the sharp edges of the metal jacket.

CABLE CONNECTORS

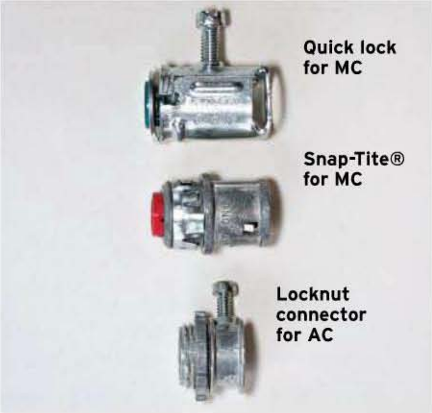

CABLE, CONDUIT, & MOISTURE

Moisture can short out electrical connections or, over time, degrade conductors and connectors. Thus it's important to use materials in appropriate locations.

- Nonmetallic (NM) sheathed cable (one brand is Romex): Dry locations only.
- Metal-clad (MC) or armored cable (AC): Dry locations; can be used in wet locations only if specifically listed for that use.
- Underground-feeder (UF) cable: Can be buried; when used in interiors, same rules as for NM cable.
- Electrical metallic tubing (EMT): Wet or dry locations; okay in wet locations if fittings (couplings and connectors) are listed for use in wet locations; may be in direct contact with earth, if suitable and approved by local code.
- Rigid metal conduit (RMC): Same rules as for EMT; may be buried in earth or embedded in concrete.
- PVC plastic conduit (schedule-40 PVC): Can be exposed or buried underground; fittings are inherently raintight, so can be installed outdoors, in damp crawl spaces, and so on.

INSTALLING AC CABLE

To install AC cable, start by marking the box heights onto the studs and installing the metal boxes flush to finish surfaces or flush to the stud edges. Remove the box knockouts you'll need. Then drill the studs and run the AC cable through each hole. Staple it every 4½ ft. along the run and within 12 in. of each box.

Use a Roto-Split to cut through a single coil of the cable's metal jacket, then slide off the severed jacket section to expose the wires inside. Use diagonal cutters to snip off the kraft paper covering the wires ❶. Slide a plastic antishort bushing between the wires and the metal jacket ❷. Next, wrap the silver bonding wire around the outside of the jacket ❸. Wrapping this wire bonds the jacket and creates a continuous ground path.

Slide a setscrew connector over the end of the AC cable and tighten the setscrew to the metal jacket ❹. The screw compresses the cable jacket and the bonding wire, holding them fast and ensuring a continuous ground. Insert the threaded end of the connector into a box knockout. Then tighten the locknut that secures the connector ❺.

Attach a mud ring to the box to bring it flush to finish surfaces, which will be installed later, and fold the wires into the box ❻.

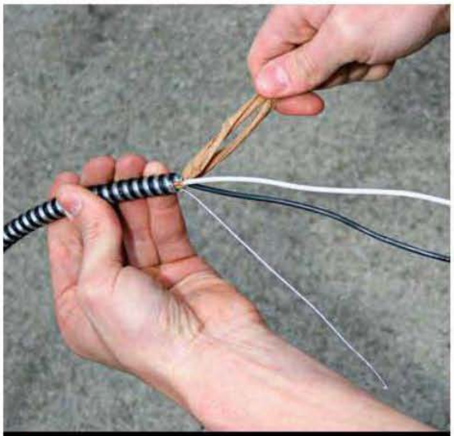

1 Pull the kraft paper to expose the insulated wires and bare bonding tracer.

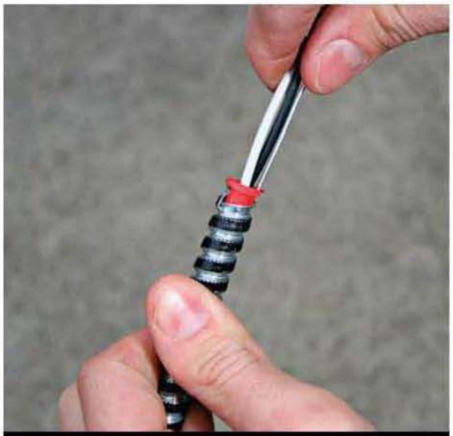

2 Insert an antishort bushing between the wires and the jacket to prevent shorts.

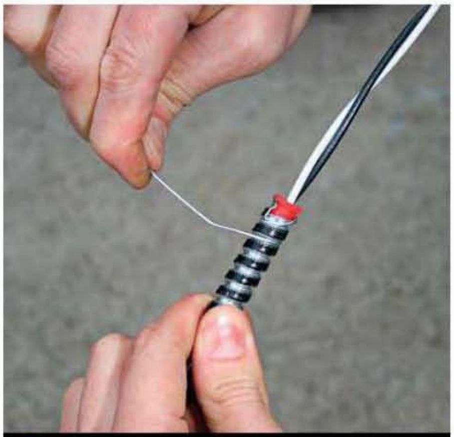

3 Wrap the bonding wire (tracer wire) around the outside of the jacket.

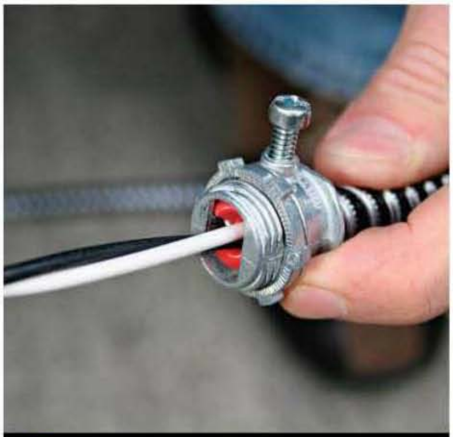

4 To ensure a solid ground path, slide a setscrew connector over the end of the cable.

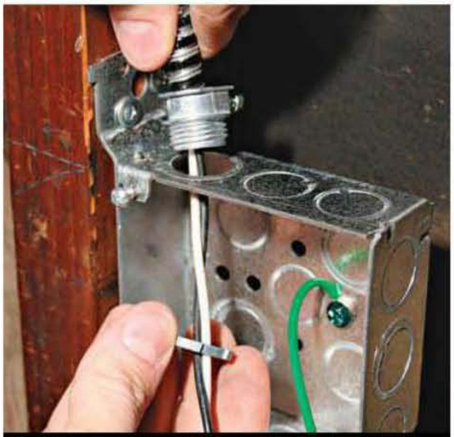

5 To secure the AC cable to the box, tighten a locknut onto the connector.

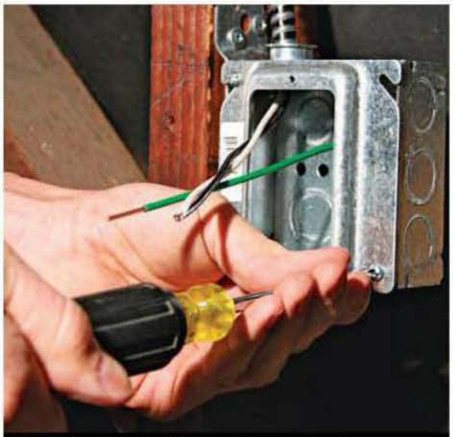

6 Attach a mud ring (here, a single-gang mud ring).

WORKING WITH EMT STEEL CONDUIT

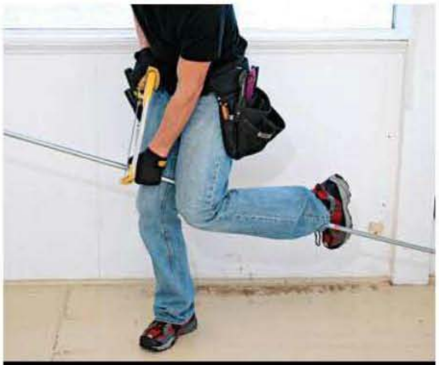

1 To steady the pipe as you cut it, brace it with your legs.

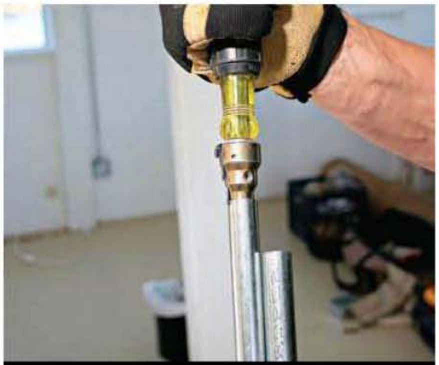

2 Use a reaming tool to remove burrs, which could nick insulation.

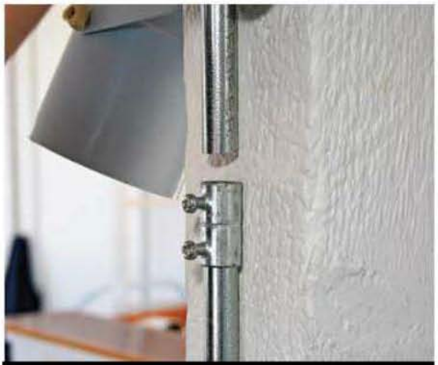

3 Use a setscrew coupling to join lengths of conduit. Be sure to tighten the screws.

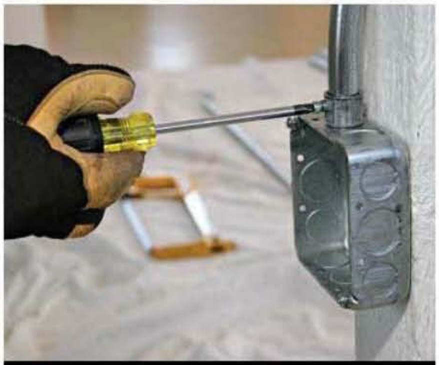

4 To secure pipe to a box, install a setscrew connector into a box knockout.

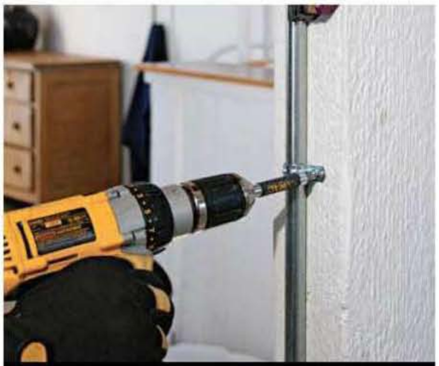

5 Use strap to secure conduit pipe.

TRADE SECRET

Use a hacksaw to cut metal conduit. Do not use the tubing cutters often used to cut plumbing pipe because they create a razor-sharp burr inside the conduit that is almost impossible to remove with a reaming tool.

Because EMT conduit is easy to work with, it's the rigid conduit type most commonly installed in residences. The main difference between conduit and flexible metal cable is that conduit comes without wire inside. Fortunately, pulling wire into conduit is a straightforward operation.

Assembling conduit is also straightforward. After cutting conduit pipe to length and deburring, insert the pipe ends into setscrew couplings and tighten the screws to secure the pipes. Threaded male fittings connect the pipe to boxes and Condulets®—covered fittings that facilitate wire-pulling or enable changes in pipe direction. You can buy prebent fittings such as capped-elbows (two-piece elbows that you can access to pull wires) and sweeps (single-piece elbows with a wide turning radius). Or you can rent a conduit bender to change pipe direction.

By Code, EMT conduit must be strapped within 3 ft. of every junction box and at least 10 ft. along runs. In actual practice, electricians strap pipe runs every 6 ft. to 8 ft. to prevent sagging.

Work from a layout sketch that indicates the power source, structural members, obstructions, existing outlets, and the locations of new outlets. Mount all the outlet boxes at the same height, and plumb each one. Plumbing boxes ensures that pipe connectors will line up with conduit pipes.

Use a hacksaw with a metal-cutting blade to cut EMT pipe; cuts need not be perfectly square. EMT pipe typically comes in 10-ft. lengths ❶. After cutting the pipe, use a reaming tool to remove the burrs inside and out ❷.

To join EMT pipe, insert pipe ends into setscrew couplings ❸. To secure pipe to a box, use a connector on the box knockout ❹. In general, plumb vertical conduit sections before strapping them into place ❺. On the other hand, you may want pipe to follow an angled architectural element such as a brace.

MAKING TURNS WITH METAL CONDUIT

To change directions in a metal conduit system, you can either bend the pipe or install directional fittings such as offset adaptors, elbows, or Condulets. EMT pipe is rigid, but its walls are thin enough to bend easily with a conduit bender.

Pros bend conduit whenever possible. Bending pipe reduces the number of specialty fittings to buy and enables pipe to follow the contours of surfaces and structural elements. It's not necessary or desirable for conduit to follow every last jog or bulge in a wall; the simpler you can make an installation, the better it will look and the faster it will progress.

To bend pipe, use a felt-tipped marker to mark the beginning of the bend on the pipe. Slide the pipe into the bender ❶. Gently step on one side of the bender and simultaneously pull on the lever bar ❷. The raised marks on the outer curve of the bender indicate the angle you're creating in the pipe—typically, 15, 22½, 30, 45, or 60 degrees.

After bending the conduit—but before cutting it to final length—test-fit the piece to see if it lines up with the connector on the box ❸ or to a coupling that joins two pipe sections. With practice, you can also offset pipe ❹. Offsetting creates two bends in opposite directions so a length of conduit can move from one plane to another.

There are a couple of rules to keep in mind as you bend conduit. First, there's a minimum requirement bend-radius for conduit: 10x the diameter. For ½-in.-diameter conduit, for example, the minimum bend radius is 5 in.

Second, each turn makes it harder to pull wire. So between each pair of boxes, you can have no more than 360 degrees of bends. In practice, every fourth turn should be a pulling point in which you can access and pull wire—in other words, the fourth turn should be a pulling elbow, a condulet, or a junction box. There is not enough room to splice wires in a pulling elbow or in a condulet. Splice wires only in a junction or outlet box.

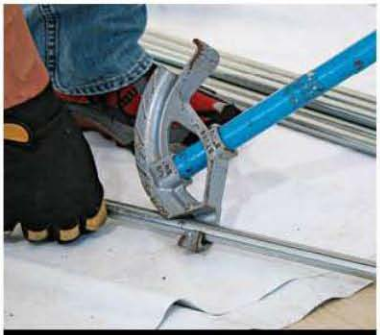

1 Mark the beginning of the bend on the pipe; then position the pipe.

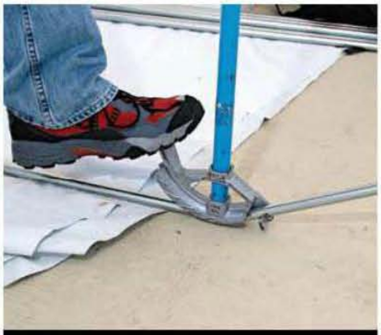

2 Pull and simultaneously step on one side of the bender.

3 Test-fit the piece to see if it lines up with the pipe connector.

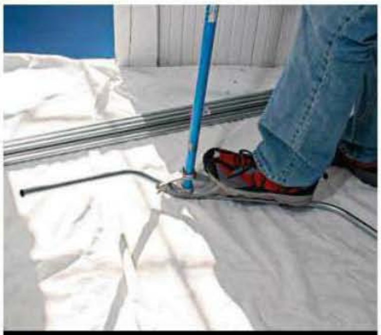

4 Conduit benders can also create multiple offset angles.

USING DIRECTIONAL FITTINGS

You can also use directional fittings to make turns. A T-condulet enables you to run wires in different directions and doubles as a pulling point when fishing wire. To attach pipe to a condulet, first screw a locknut onto the threaded shaft of a male adaptor. Turn the adaptor most of the way into the Condulet hub and turn the locknut clockwise until it seats against the hub. Back-tightening the locknut in this manner ensures grounding continuity.

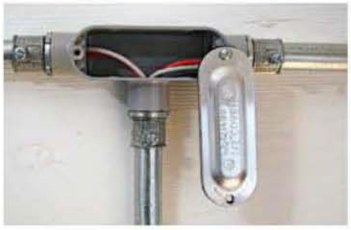

Use a T-condulet as a pulling point for fishing wires. Here, wires from the source diverge in two directions.

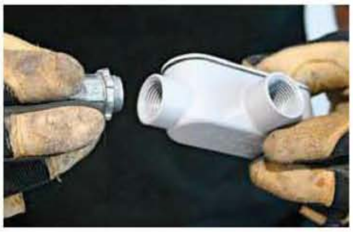

To attach pipe to a condulet, use an adapter and back-tighten the locknut until it lodges against the hub.

FISHING CABLE & CONDUIT

You can start fishing wire from either end of the circuit. If you're tapping into an existing outlet, it makes sense to fish from that outlet—after first turning off the power to the outlet and testing to make sure it's off.

In the installation shown here, a four-square extension box was mounted over an existing (recessed) box. Thus new wires can be pulled into the extension box and spliced to an existing cable to provide power for the circuit extension being added.

The fish tape can be fed easily into the conduit pipe ❶. At the other end of the conduit, tape wires to the fish tape ❷. To make wire pulling easier, leave the wire ends straight—do not bend them over the tape, but stagger them slightly so the bundled wires taper slightly. Wrap electrical tape tightly around the wire bundle so that wires stay together and won't snag as they're pulled. The pulling will also go easier if you pull stranded wire rather than solid wire, which is stiffer and less flexible.

Finally, use a wire caddy ❸, even if you have to build one out of scrap pipe and lumber. Using wire spools on a caddy helps minimize tangles.

TAPPING INTO AN EXISTING OUTLET

Tapping into an existing outlet is often a convenient way to supply power to a conduit extension. Remove the cover plate from the outlet, detach wires from the old receptacle, then use lineman's pliers to straighten the old wires so they'll be easier to splice to wires running to the new outlets.

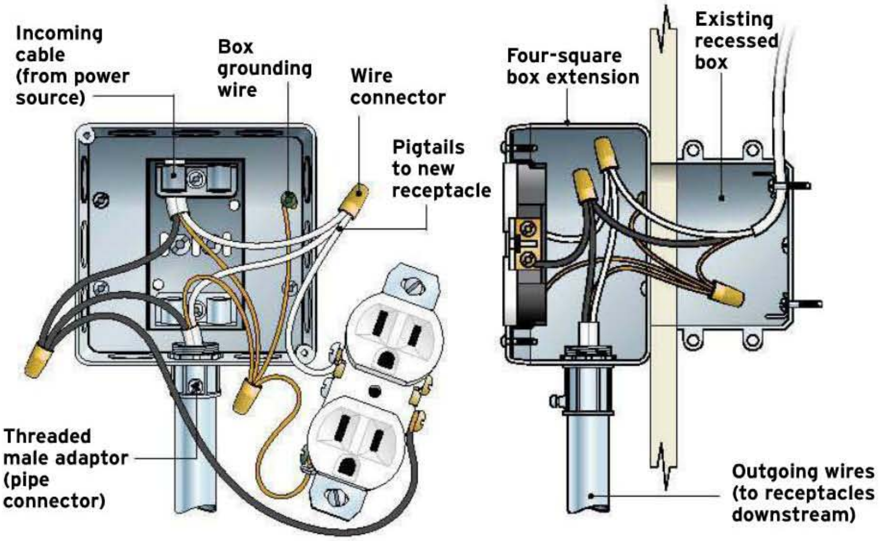
Incoming cable (from power source)
Box grounding wire
Wire connector
Pigtails to new receptacle
Four-square box extension
Existing recessed box
Threaded male adaptor (pipe connector)
Outgoing wires (to receptacles downstream)

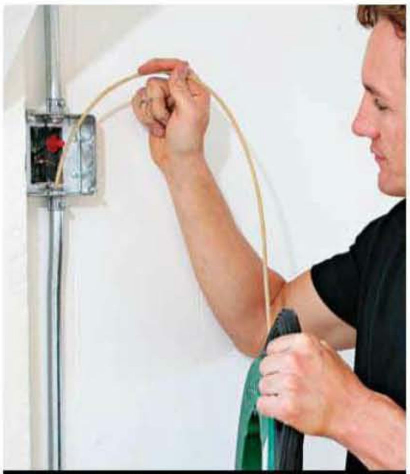

1 Push the fish tape down into the conduit.

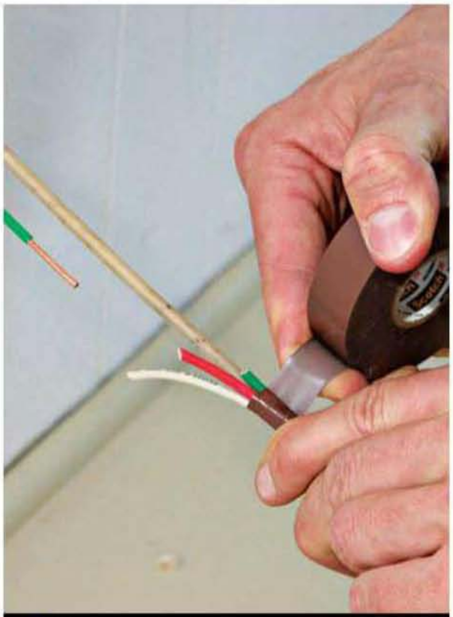

2 When the fish tape emerges from the conduit, attach the wires.

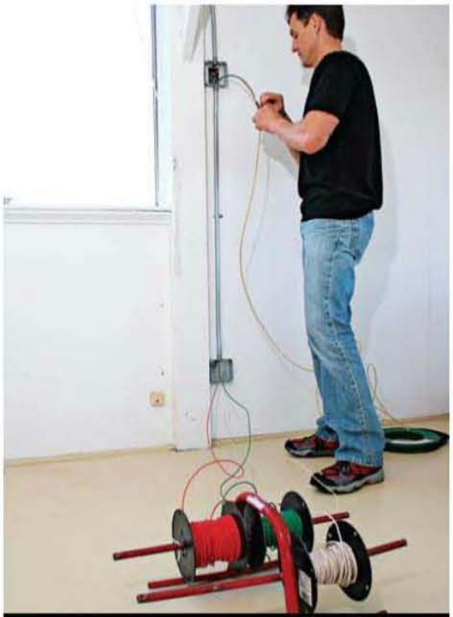

3 As you pull wires through the conduit, the spools on the wire caddy turn.

PREPPING RECEPTACLES FOR SURFACE METAL BOXES

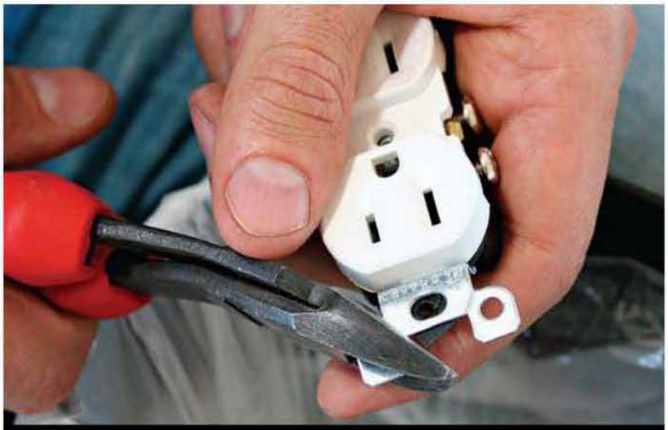

1 Use diagonal cutters to snip the plaster ears off the receptacles so they'll fit into industrial raised covers.

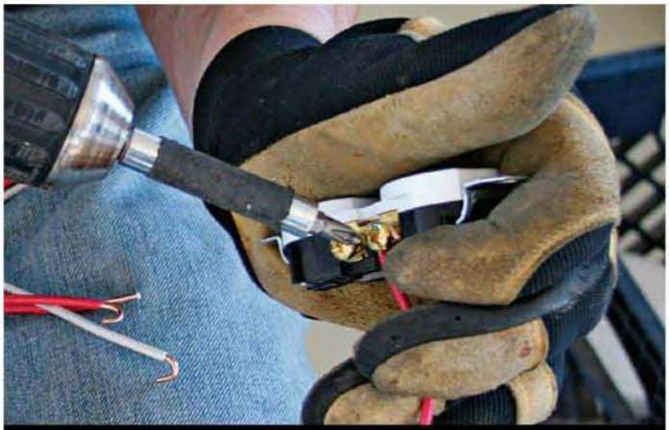

2 Attach solid pigtails to screw terminals.

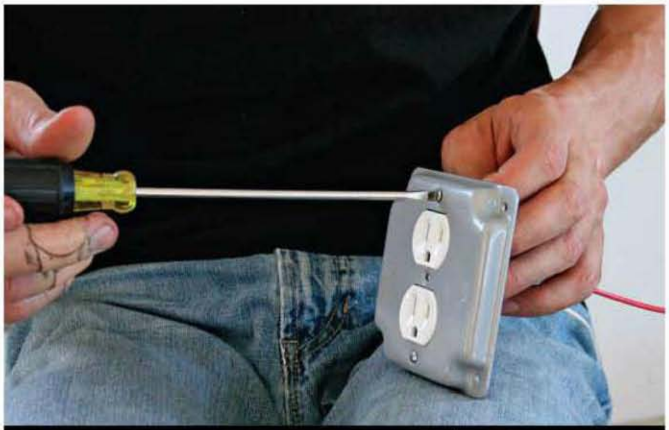

3 Screw the covers to the receptacle's mounting tabs.

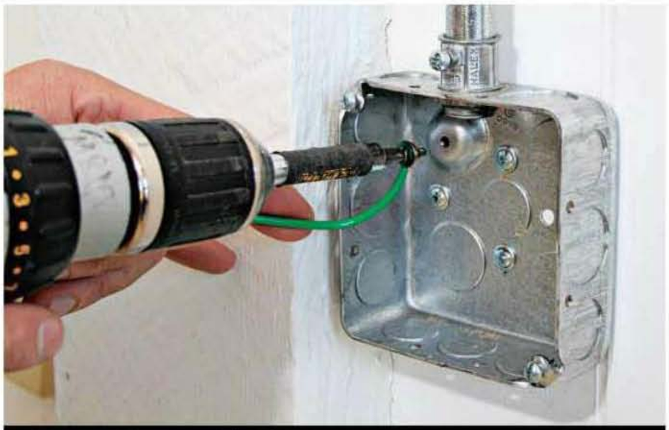

4 Screw the ground pigtail to the box.

To save time, professionals often divide receptacle wiring of a conduit system into several smaller tasks, each of which can be done with a single tool. Your conduit system may have slightly different materials, but these prep techniques should save you time.

Standard duplex receptacles come with plaster ears intended to seat against plaster or drywall surfaces. Consequently, the ears may not fit into metal boxes or behind the industrial raised covers often installed in conduit systems. Use diagonal cutters to snip off the ears ❶ so they fit inside such covers.

To speed the assembly, precut all the pigtails that you'll attach to receptacles. To loop pigtail ends, strip $1/2$ in. of insulation, insert the stripped ends into the little hole on the stripper jaw, and twist your wrist.

Use a screw gun to attach the looped ends to the receptacle screw terminals ❷. Remember to loop the pigtail ends clockwise so that when the screw tightens (clockwise), the loop stays on the screw.

Next, attach the prewired receptacles to the covers. Typically, there is a pair of machine screws (and nuts) that screw into the mounting tabs at top and bottom, and a single short screw for the center hole in the front of the receptacle ❸.

Finally, screw a ground pigtail to the box ❹. The threaded hole for the ground screw is raised slightly to facilitate surface mounting. If you surface-mounted a box on a concrete wall and the box didn't have this detail, the ground screw would hit concrete before it tightened all the way down—thus jeopardizing grounding continuity.

WIRING OUTLETS

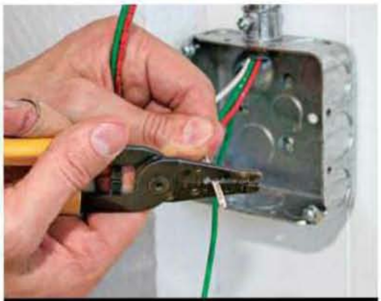

1 Strip ½ in. of insulation from the ends of the stranded wire.

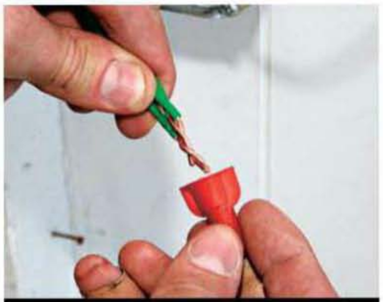

2 Splice the ground wires first.

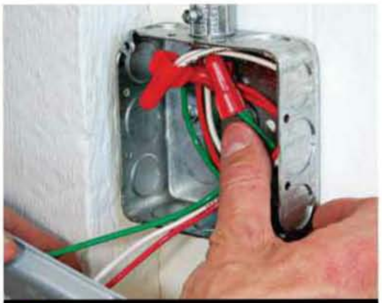

3 After splicing like-colored wire groups, fold them into place.

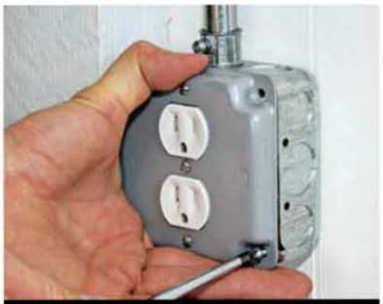

4 Then screw the receptacle cover to the box.

If you've already attached pigtails to the receptacle screw terminals, the connections at outlet boxes on a conduit system will go quickly. *Note:* In the project shown here, the red wire is the hot conductor, the white wire is neutral, and green is the ground conductor. Code requires that green or bare wires are always the designated ground.

Use wire strippers to strip ½ in. of insulation from the wires pulled through the conduit ❶. Splice the receptacle pigtails to like-colored circuit wires pulled through the conduit. Typically, electricians splice the ground wires first, which means splicing three wires: the receptacle pigtail, the grounding pigtail to the box, and the circuit ground ❷. If the metal conduit serves as the ground, there will be no circuit ground wire.

When splicing stranded wire to solid wire, strip the stranded wire slightly longer than the solid wire, so the stranded-wire end sticks out beyond the solid wire. By doing this, you force the stranded-wire end into the wire connector first, ensuring a solid connection. Use lineman's pliers to twist wires slightly before screwing on the connector.

Once you've spliced all wire groups, fold the wires into the box ❸ and attach the cover. Hold the cover tight against the box and attach it with the cover screws ❹.

GROUNDING CABLE & CONDUIT SYSTEMS

Every newly installed circuit must have continuous grounding. Steel conduit acts as its own ground path. MC cable contains an insulated ground because its metal jacket does *not* serve as a ground path; it exists solely to protect the wires inside. AC cable's metal jacket, on the other hand, *does* serve as a ground path. To ensure a continuous ground in AC cable runs, wrap and secure the cable's thin silver bonding wire.

If Code requires steel conduit, AC or MC, you must use steel boxes. Steel boxes must be grounded as well. In addition to tightening the cable or conduit couplings to the box knockouts, screw a grounding pigtail into the threaded hole in the box. The box pigtail is then spliced to circuit grounds and pigtails that run to the device.

Alternatively, if you're using solid wire, you can use a grounding clip to ground a metal box. (Code prohibits using a grounding clip with stranded wire.) If the incoming ground wire is insulated, strip approximately 6 in. of the insulation and slide the grounding clip onto a section of bare wire. This will allow enough wire beyond the clip so you can splice the bare wire end to other grounds or attach it directly to the grounding screw on a device.

Premade grounding pigtails come with one end looped around a green ground screw. If your job is large, using premade grounds can save a lot of time.

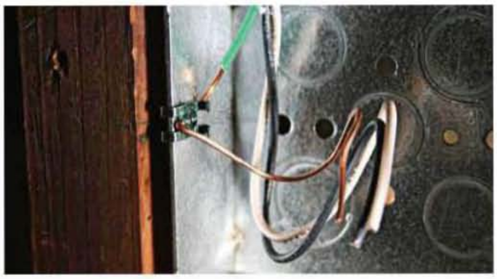

Using a ground clip to ground a metal box is suitable for solid conductors only.

APPLIANCES

WIRING APPLIANCES, LIKE WIRING general-use circuits, is primarily a matter of solidly connecting conductors and following the recommendations of the NEC and the manufacturer. Many appliances are big energy users, so it's particularly important to size circuit wires and breakers based on appliance loads. When your new appliances arrive, read—and save—the owner's manuals that come with them. Owner's manuals contain essential information on how to install the appliance, how to register the appliance and comply with its warranty, how to identify and order replacement parts, and other useful information. Increasingly, manufacturers offer owner's manuals online, so if your appliance lacks a manual, download a copy at once. There's no guarantee that manuals will stay in print—or online.

BASICS

A quick look at kitchen appliances, p. 212

Preparing an unfinished appliance cord, p. 213

GARBAGE DISPOSERS

Installing a garbage disposer, p. 214

OVENS

Roughing in an oven outlet, p. 216

Wiring a drop-in oven, p. 217

HOUSEHOLD APPLIANCES

Wiring laundry setups, p. 218

Installing baseboard heaters, p. 220

Electric water heaters, p. 221

A QUICK LOOK AT KITCHEN APPLIANCES

There's a lot going on in a modern kitchen. Typically it contains a refrigerator, a dishwasher, a garbage disposer, a range hood, a slew of small countertop appliances, and, in many cases, an electric range, cooktop, or oven.

→ **For more on wiring range hoods, see p. 150.**

Appliances have different wiring requirements. Smaller ones such as disposers, dishwashers, and range fans require 120v; other appliances require 240v; and others—such as electric ranges—require both 120v and 240v. Ranges burners and ovens use 240v, but a range's timer and clock use 120v.

Let's start with a handful of important concepts.

Covered junction boxes

All electrical connections must take place in a covered junction box—either an integral box inside the appliance or in a covered junction box secured to a framing member, such as a wall stud.

Equipment grounding

All appliances must have an equipment-grounding conductor that connects to both the appliance frame (or housing) and to the metal outlet box for the safe discharge of fault currents.

Accessible disconnect means

Appliances typically have an accessible "disconnect means" so you can cut the power. (This is in addition to the fuse or breaker controlling the circuit.) For appliances that slide out—such as a dishwasher or a refrigerator—the disconnect means is typically a cord and a receptacle plug. Appliances that are *hard wired*, such as drop-in cooktops and wall ovens, must have an access panel near the junction box where incoming circuit wires connect to *appliance whips*. (Whips are flexible cords—metal clad or plastic sheathed—that connect to appliance terminals.)

Dedicated circuits

Code requires that every fixed appliance must be served by a separate, *dedicated* circuit—a circuit that serves only that appliance. This includes appliances that must stay on, such as refrigerators and freezers, as well as large energy users, such as electric ranges, electric water heaters, and clothes dryers that require heavier wire and higher-rated breaker (or fuse) protection.

→ **For more on dedicated circuits, see p. 165.**

COMMON ELECTRICAL APPLIANCE REQUIREMENTS

User	Typical Wire Size*	Fuse or Breaker
Small-appliance circuit	12AWG	20 amps
Refrigerator (120v)[†,‡]	12AWG	20 amps
Stand-alone freezer (120v)[†,‡]	12AWG	20 amps
Dishwasher (120v)[†,‡]	12AWG	20 amps
Disposer (120v)[†,‡]	12AWG	20 amps
Microwave (120v)[†,§]	12AWG	20 amps
Range (120/240v)[†,‡]	10AWG	30 amps
Range (120/240v)[†,‡]	8AWG	40 amps
Range (120/240v)[†,‡]	6AWG	50 amps
General utility/workshop	12AWG	20 amps
Laundry circuit (washer)[D]	12AWG	20 amps
Clothes dryer (120/240v)[†,‡]	10AWG	30 amps
Water heater (240v)[†,‡]	10AWG	30 amps
Baseboard heater (120v)[†,‡]	12AWG	20 amps (max. 1,500w on circuit)
Baseboard heater (240v)[†,‡]	10AWG	30 amps (max. 5,760w on circuit)
Whole-house fan[†]	12AWG	20 amps
Window air-conditioner (240v)[†,‡]	10AWG	20 amps

[*] Ratings given for copper (CU) wiring.
[†] Requirements vary; check rating on appliance nameplate and follow manufacturer's specifications.
[‡] Requires dedicated circuit.
[§] Microwaves are typically rated 15 amps but are installed on 20-amp kitchen circuits.
[D] Requires designated circuit (see p. 219).

TRADE SECRET

Don't use ground-fault circuit interrupter (GFCI) receptacles with major appliances, such as refrigerators. The motors of such appliances have a high inductance load as they start up, which can cause a GFCI receptacle to trip unnecessarily—leaving you with a refrigerator full of spoiled food.

PREPARING AN UNFINISHED APPLIANCE CORD

There are many different types of appliance cords. Some cords come with a molded plug and precrimped connectors that attach to terminals on the appliance. Other cords have a molded plug but an unfinished end that you must strip and splice to the lead wires of an appliance. In most cases, the splice is housed in an integral junction box inside an appliance.

The cord seen here is a typical 120v cord that you might attach to a smaller fixed appliance, such as a garbage disposer. It contains a hot wire, a neutral wire and—in the center—a sheathed ground wire. Look closely at the cord and you'll see that its sheathing has a ribbed side and a smooth side. The ribbed side contains a neutral wire that must be spliced to the neutral wire of the appliance; the smooth side contains a hot wire ❶.

Start by snipping and separating the three stranded wires within the cord ❷. Using a utility knife, carefully slice and peel back the cord's outer (gray) insulation from the middle wire. As you do so, you'll expose the ground wire's green insulation. Only the ground wire has this additional layer of insulation ❸. Next use a wire stripper to remove ½ in. of insulation from the ends of all three cord wires ❹. Now you're ready to splice those wires to the lead wires on the appliance.

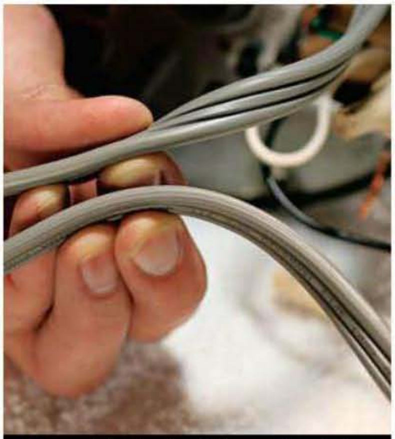

1 The ribbed side of the cord contains the neutral wire.

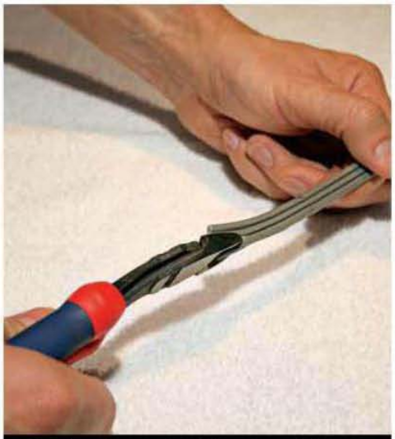

2 Use diagonal cutters to cut the cord and separate the three wires.

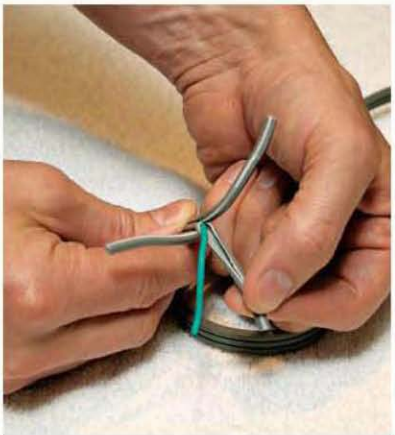

3 Peel back the gray insulation to expose the ground wire.

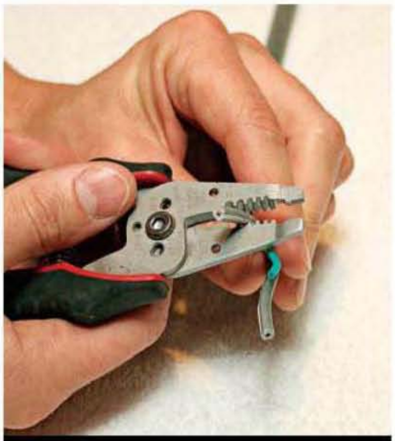

4 Use wire strippers to prepare the ends of all three cord wires.

WHIP IT!

For eons, how-to books had sections about replacing appliance plugs. Forget that. Today, you can easily find replacement cords—also called appliance whips—with molded plugs and precrimped connectors, which are far easier and safer to install. By the way, always grab the plug—not the cord—when unplugging an appliance.

Replacing the entire cord is often easier than replacing a broken plug.

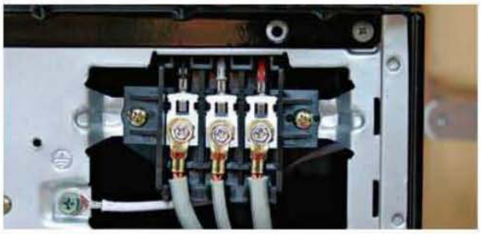

The connectors on the whip attach to terminals on the appliance.

INSTALLING A GARBAGE DISPOSER

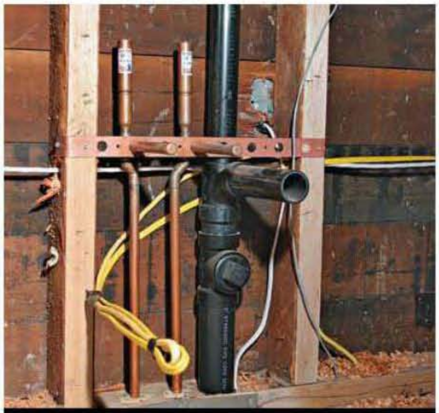

1 The bundled yellow wires supply the garbage disposer and dishwasher.

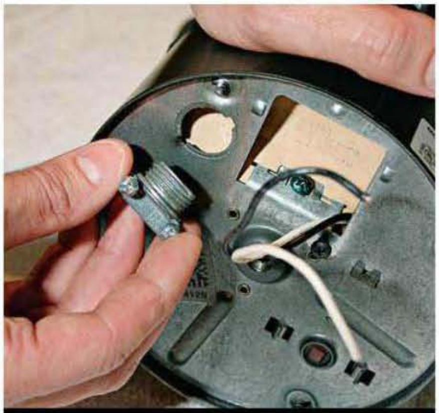

2 Remove the cover plate from the junction box and screw on a connector.

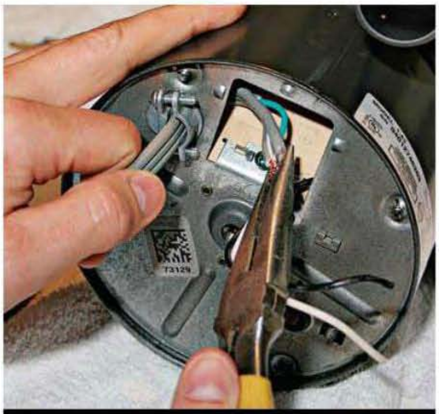

3 Feed the cord wires through the connector and junction box.

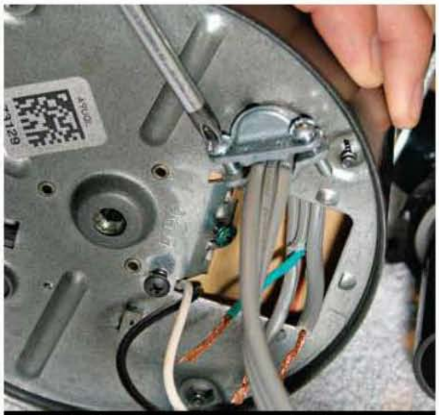

4 Tighten the cable connector to prevent stress on the wire connections.

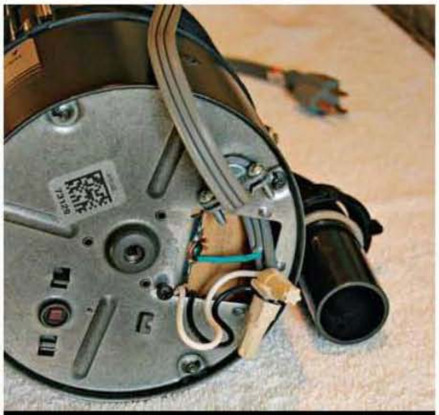

5 Connect the appropriate wires with wire connectors.

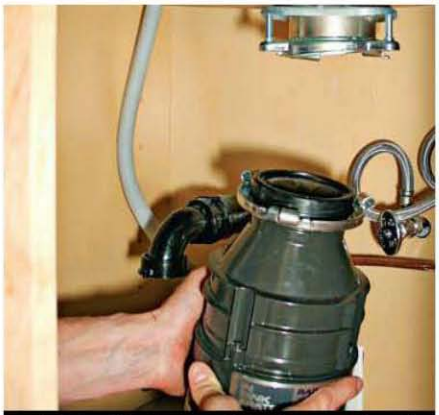

6 Raise the disposer into place and turn it until the mounting rings lock into place.

Though installation of a disposer is largely the same from brand to brand, be sure to follow the instructions that come with your unit. Typically, plumbing supply and waste pipes are stubbed out, and 12AWG cable is roughed-in before the finish walls and cabinets are installed ❶. After the base cabinet is in place, install a 20-amp duplex receptacle to supply power to the disposer and dishwasher. The receptacle for the disposer should be a switched receptacle.

→ **For more on wiring a split-tab receptacle, see p. 49.**

Install the sink and attach the disposer mounting assembly in the sink outlet. Then route the dishwasher overflow pipe into the cabinet under the sink.

With these prep steps done, you're ready to attach the appliance cord. Remove the cover plate from the bottom of the unit and pull its wire leads. Then screw a cable connector into the knockout in the bottom of the unit ❷. Separate and strip cord wires and feed them through the cable connector. Use needle-nose pliers to pull the wires through the junction box ❸ and then tighten the cable connector ❹.

Attach the green ground wire to the ground screw in the junction box. Then use wire connectors to splice like wires—neutral to neutral, hot to hot ❺. Fold all wires into the junction box and replace the cover plate. Lift the disposer until its mounting ring engages the mounting ring on the bottom of the sink. Turn the unit until the rings lock ❻.

Slide the tube from the dishwasher onto the dishwasher inlet stub and tighten its clamps ❼. Attach the P-trap to the discharge outlet on the disposer ❽. Plug the disposer plug into the undersink receptacle. The second plug runs to the dishwasher.

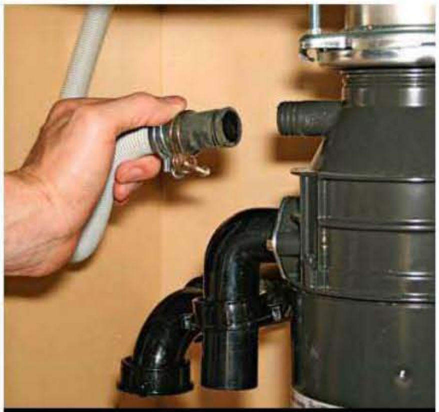

7 Slide the dishwasher drain onto the inlet stub and tighten its clamp.

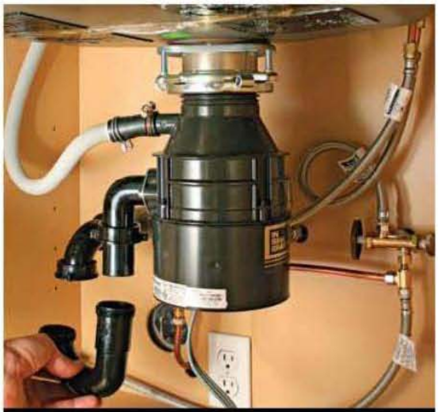

8 Attach the P-trap to the discharge outlet on the disposer.

WIRING A GARBAGE DISPOSER

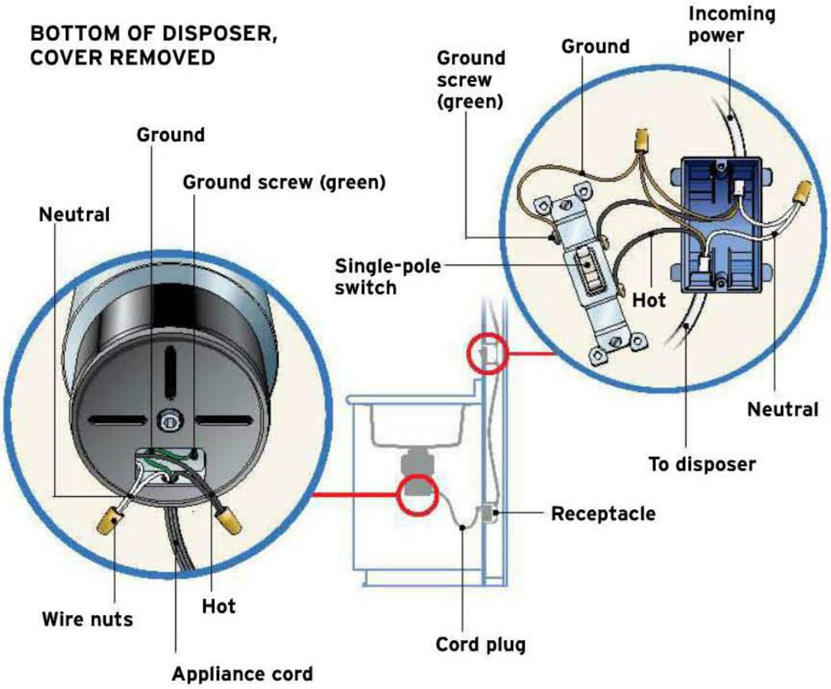

DISPOSERS & DISHWASHERS

Disposers and dishwashers are both 120v, 20-amp appliances, so both are supplied by 12AWG cable. Most garbage disposers have a covered junction box on the bottom, to which a plugged cord attaches. You can also hard wire a disposer, but having a receptacle in the cabinet under the sink enables a homeowner to quickly unplug the unit should he or she need to repair or replace it. Typically, an undersink outlet is controlled by a switch above the counter, as shown in "Wiring a Garbage Disposer," at left.

A dishwasher and a garbage disposer usually plug into a duplex receptacle in the cabinet under the sink—a split-tab receptacle. A split-tab receptacle is a standard duplex receptacle whose middle tab has been removed to create a duplex receptacle fed by two circuits—that is, by two hot wires.

Because they slide out for installation and maintenance, dishwashers are also installed with a cord and plug. Most of the time, the dishwasher junction box is located in the front of the unit, just behind the kick panel. After attaching cord wires to wire leads in the junction box, run the cord in the channel behind the dishwasher to reach to an outlet.

WARNING

The NEC specifies that you connect the two hot wires of a split-tab receptacle to a double-pole breaker. When you flip off the toggle for a double-pole breaker, you shut off both hot wires. If you instead connect the hot wires of a split-tab receptacle to separate, single-pole breakers and then flip off only one breaker, you might test the top half of the split-tab receptacle and conclude—mistakenly—that the bottom half was off, too. Attaching both hot wires to a double-pole breaker prevents such a potentially lethal mistake.

ROUGHING IN AN OVEN OUTLET

Roughing in an oven outlet is not that different from roughing in any other outlet. Remove the knockout from a four-square box, insert a plastic cable connector, and then feed in the 10/3 NM cable that will power the range. Staple the cable within 12 in. of the box; drive the staple just snug ①.

Although an experienced electrician can use a utility knife to strip sheathing from any cable, using a cable ripper that can accommodate large-gauge wire makes sense for nonprofessionals ②. Hold the ripper channel snug to the cable and pull it down the length of the cable to score the sheathing ③.

Pull back the sheathing to expose wires inside. Then use diagonal cutters to cut the sheathing free. Leave 1/2 in. to 1 in. of sheathing inside the box ④. Attach the ground wire to the box using a green ground screw in a threaded hole. Wrap the wire clockwise around the screw so it will stay in place as the screw tightens down on it ⑤.

Fold the wires neatly into the box so they can be easily pulled out during the trim-out phase, when they'll be attached to a 30-amp receptacle or hard wired directly to a metal-clad appliance whip. Finally, attach a two-gang mud ring to the outlet box so it will be flush to the finish surface ⑥.

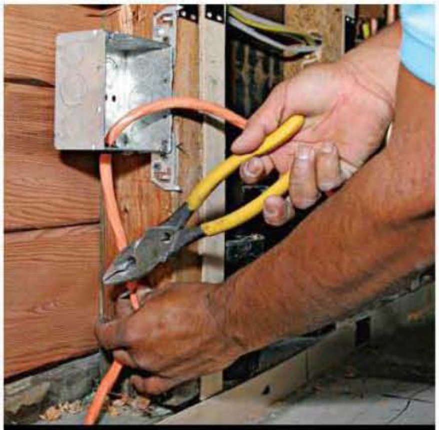

1 **10/3 NM cable is run to a four-square box.**

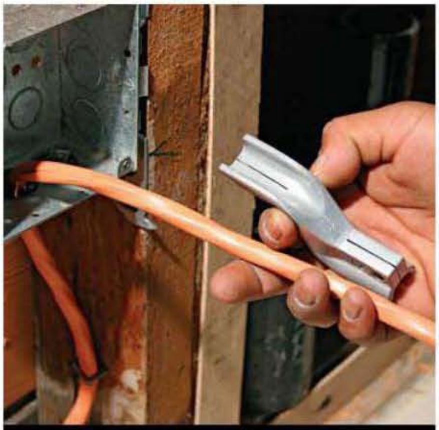

2 **Use a cable ripper that can accommodate larger gauge wire.**

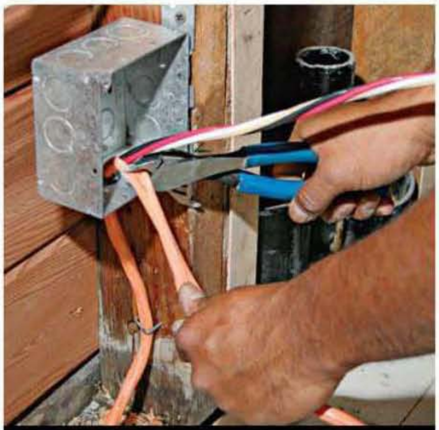

4 **With the wires exposed, use diagonal cutters to cut the sheathing loose.**

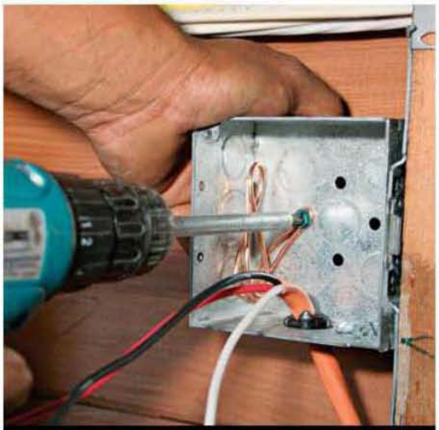

5 **Leave the ground wire long to be spliced to the ground in the appliance whip.**

ELECTRIC RANGES, OVENS, AND COOKTOPS

When discussing cooking appliances, you'll need to keep several terms straight: The enclosed cooking area in which you roast a turkey is an *oven*; you place pots and frying pans on *cooktop* burners. A *range* has both an oven and a cooktop.

In any case, the heating elements of ranges, ovens, and cooktops generally require 240v, but today's smart appliances come with a plethora of timers, clocks, sensors, buzzers. and other gizmos that use 120v. For this reason, many units require 120/240v wiring, with two hot wires, an insulated neutral, and an equipment ground wire.

As noted elsewhere, if the unit slides in and out for maintenance, it is usually installed with a plug inserted into a matched receptacle. The outlet box that contains that receptacle may be surface mounted or recessed (so that the receptacle can be flush mounted). If the unit drops in and stays put, it is typically hard wired to a junction box via an appliance whip.

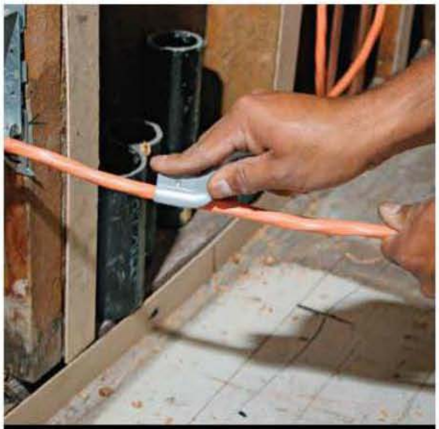

3 Pull the ripper to score the sheathing.

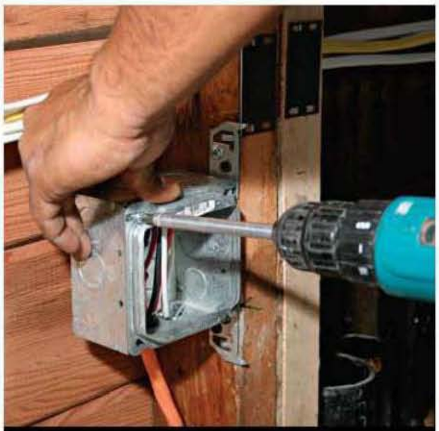

6 Attach a mud ring to bring the box flush.

TRADE SECRET

Appliances are usually the same depth as base cabinets (24 in.) so appliance faces will be even with cabinet faces. Because most appliances are installed against a wall, manufacturers often build a recessed area in the back of the appliance to accommodate electrical connections. Refer to the user's manual prior to rough-in.

WIRING A DROP-IN OVEN

Drop-in ovens and other stationary appliances must be hard wired. For the project shown here, an electric oven has been installed in a base cabinet, and the wires in its MC cable are ready to be connected to 10AWG wires, roughed in to a four-square box. The edge of the box must be flush to the cabinet back, because plywood is flammable. If the box or its mud ring is below the cabinet back, add a four-square box extension to bring it flush ❶.

Install a two-piece, right-angle cable connector (a flex-90) to the end of the cable whip. Slide the bottom of the connector onto the end of the whip and screw on the top of the cable connector ❷. Remove the stamped knockout on the four-square cover; then feed the whip wires and the connector end through the knockout ❸. Tighten a locknut onto the threaded connector end to lock the cable connector to the cover.

To ground the outlet box, loop the incoming ground wire under a green ground screw ❹. Splice that ground wire to the appliance ground. Use wire strippers to remove ½ in. of insulation from the neutral and hot wires on the incoming cable ❺; then use wire connectors to splice like-colored wires together: black to black, red to red, and white to white. To ensure that wire connectors grip the wires securely, use needle-nose pliers. (There are two hot wires—red and black—because the oven requires 240v.)

Once all the wire groups are spliced using connectors ❻, tuck them into the box. Then attach the cover to protect the connections within ❼.

>> >> >>

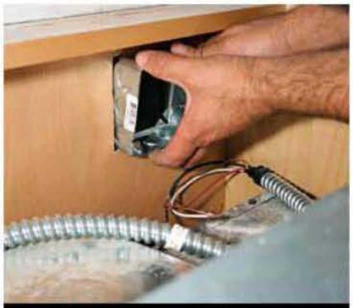

1 A four-square box extension brings the outlet flush.

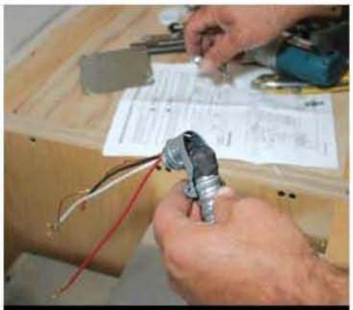

2 A right-angle cable connector is attached.

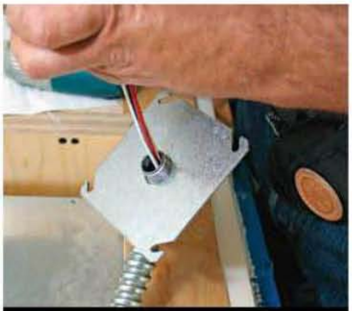

3 Remove a knockout and feed the wires through.

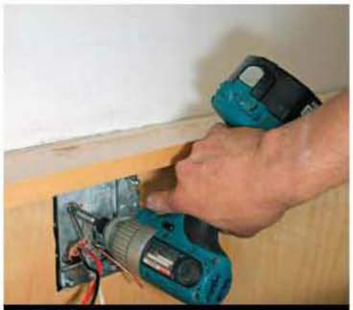

4 Wrap the incoming cable ground under a ground screw.

WIRING A DROP-IN OVEN (CONTINUED)

5 Remove ½ in. of insulation from the wires.

6 Fold the wires into the box.

7 Attach the cover to the box.

WIRING LAUNDRY SETUPS

Electricians sometimes run an insulated ground wire in conduit. In the setup shown here, the metal conduit is the system ground. There is, however, a green grounding pigtail that runs from each box grounding screw to a receptacle ground screw. In this project, wires are not attached to a power source, so they are safe to handle.

➔ For more on installing metal conduit, see pp. 205 and 206.

Start by fishing wire to the box nearest to the power source. Here, four wires had to be fished: two 10AWG wires to feed the dryer and two 12AWG wires to feed the washer. Untape the wire bundle and then trim each set of wires to rough length so that at least 6 in. of wire sticks out of the box ❶; many electricians leave 8 in. to 10 in. of wire sticking out.

Attach a green grounding pigtail to the box and strip ½ in. of insulation off the ends of the incoming wires. If the point of the grounding screw hits the masonry wall, you may need to shorten the grounding screw so that it doesn't run into the masonry. Insert the stripped wire ends into screw terminals on the dryer receptacle ❷. Twist the stranded wires beforehand so they will not spread excessively when you tighten the screws. After tightening the screws, gently tug each wire to be sure it's well attached. The green pigtail grounds the device to the box.

The receptacle for the washer is a standard 20-amp receptacle. Use a wire stripper to create loops that you can attach to the screw terminals on the receptacle ❸. Once both receptacles are wired, fold the wires into their respective boxes and attach the cover plates ❹, ❺. If you install an industrial raised cover, secure the receptacle to the cover, then screw the cover to the box. Once you've installed the covers, attach incoming the wires to breakers in the panel.

>> >> >>

MEASURING PIPE LENGTH

To determine the length of conduit pipe running between two outlet boxes: Measure from the centerline each box (A).

***Subtract* 2 in. from the centerline of each box (B).**

***Subtract* the distance each adaptor sticks out of the box (C).**

***Add* the distance that pipe ends fit into adaptor sockets (D).**

Pipe length = A − 2B − 2C + 2D

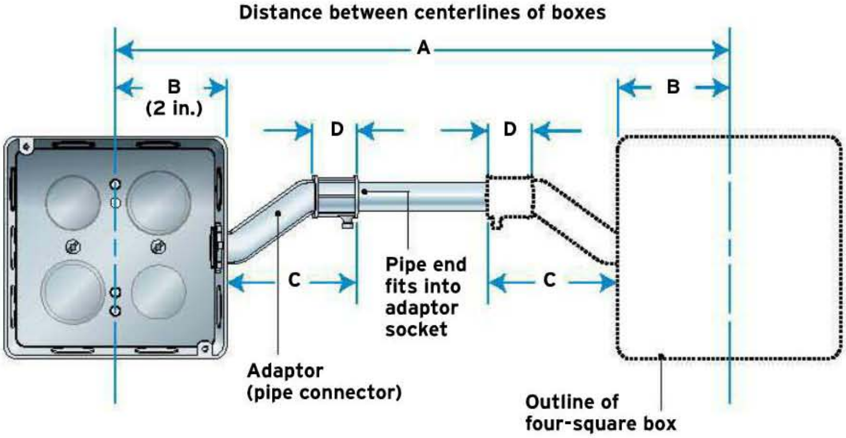
Distance between centerlines of boxes
A
B
(2 in.)
B
D
D
C
C
Pipe end fits into adaptor socket
Adaptor (pipe connector)
Outline of four-square box

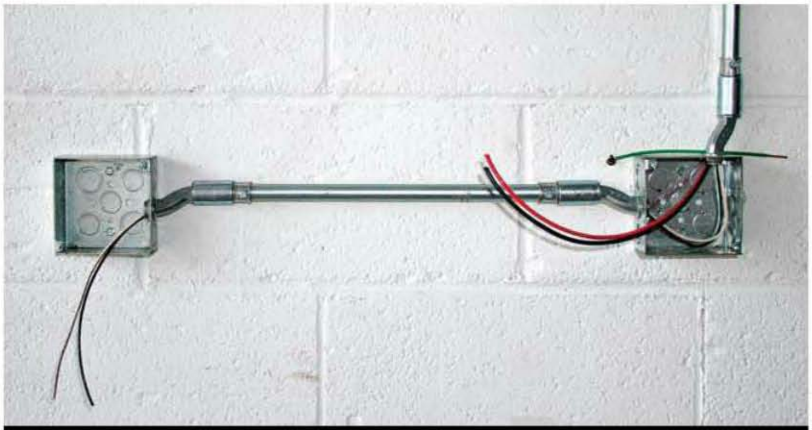

1 Trim the wires running to each box so that 6 in. to 8 in. of wire sticks out of the box.

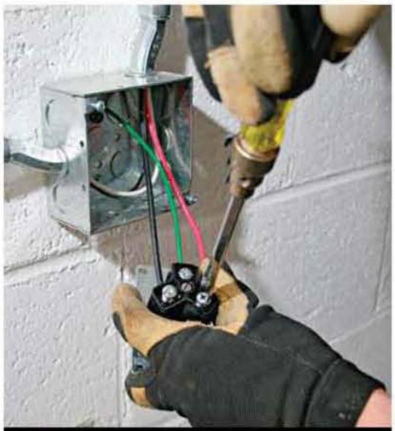

2 Twist the stranded wire ends, then insert them into screw terminals.

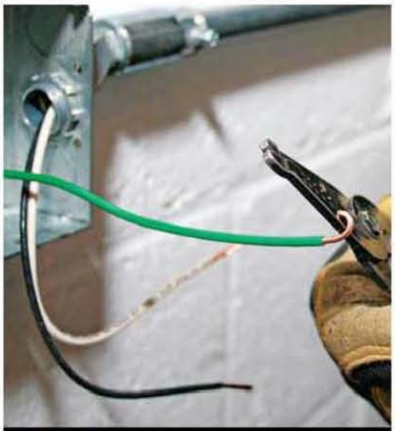

3 Use the hole on the handle of the wire stripper to create loops.

WIRING REQUIREMENTS FOR LAUNDRY SETUPS

Clothes washers and dryers are often wired with two surface-mounted receptacles.

As with electric ranges, electric dryers typically require 120/240v wiring because, in addition to their 240v heating elements, dryers are also equipped with several elements that use 120v—such as drum motors, timers, and buzzers. So dryer circuits include two hot wires, a neutral wire, and an equipment ground wire.

Equipment grounds on washers and dryers connect to appliance housings to provide a safe route for fault current, should a short circuit occur. Washer circuits are usually wired with 12AWG wire and protected by a 20-amp breaker or fuse; dryer circuits are wired with 10AWG wire are protected by a 30-amp breaker or fuse. But, as always, note the nameplate ratings on your appliances and wire them accordingly.

The type of appliances used determines how many wires you fish. Some dryers require three incoming wires (two hot, one neutral); whereas others require two hot wires only. For the washer and dryer installed in the projects shown here, we ran four wires: two #10 stranded THHN wires (two hots) for the dryer (240v) and two #12 stranded THHN wires (a hot and a neutral) for the washing machine (120v).

DEDICATED VS. DESIGNATED CIRCUITS

Electrical codes require that certain critical-use appliances, such as refrigerators and freezers, be installed on a *dedicated circuit*—that is, the circuit serves only one appliance. Less well known are *designated circuits*, such as the dryer receptacle shown in the photo. The receptacle is the only one on the circuit: The distinction is that the circuit is designated for one use (laundry) but not for one appliance. Thus, if you decided to switch to a gas clothes dryer, you plug both that dryer and a clothes washer into the receptacle.

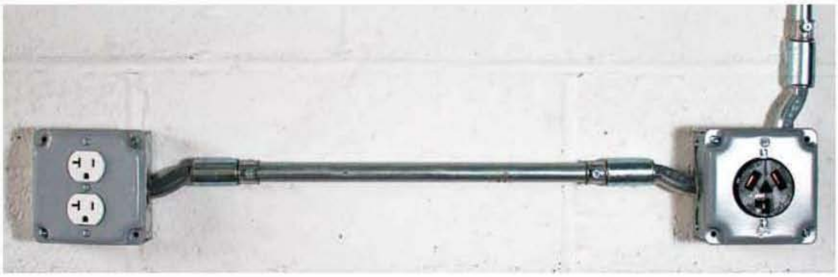

Typical laundry setup. A 30-amp dryer receptacle on a dedicated circuit (right) and a 20-amp washer receptacle on a designated circuit (left).

WIRING LAUNDRY SETUPS (CONTINUED)

4 Tuck the wires of each receptacle into the box.

5 Install the cover plates.

INSTALLING BASEBOARD HEATERS

Baseboard heaters are increasingly popular as a backup to a central heating system. Installing units with in-heater thermostats is a better choice than installing a central wall thermostat that controls all units. Baseboard units with in-heater thermostats are easier to install and more cost-effective to operate because they deliver heat to areas where it's needed most. Because most baseboard units are installed under windows, units with in-heater thermostats can respond faster to cold air as it enters.

Baseboard heaters are available in 120v and 240v models, but 240v models are generally more efficient. As a rule of thumb, you can connect several small heaters to one cable running from the main panel, as long as their combined continuous load doesn't exceed 80 percent of the cable's rating. In other words, if you run 12AWG cable with 20-amp protection, the continuous current should not exceed 16 amps. But because wattage varies from model to model, follow the installation instructions that come with your model.

Wiring an in-heater thermostat is pretty straightforward. Rough-in wiring to each heater location, remove a cover plate on one end of each unit, feed the incoming cable through a cable connector and splice the incoming circuit wires to the thermostat wires. Make sure to attach a grounding pigtail to the metal housing of the unit. Thermostats will be single-pole or double-pole switches. Double-pole thermostats are preferable because when you flip them off, you cut power to both legs of the unit, which is safer and more convenient.

WIRING BASEBOARD IN-HEATER THERMOSTATS

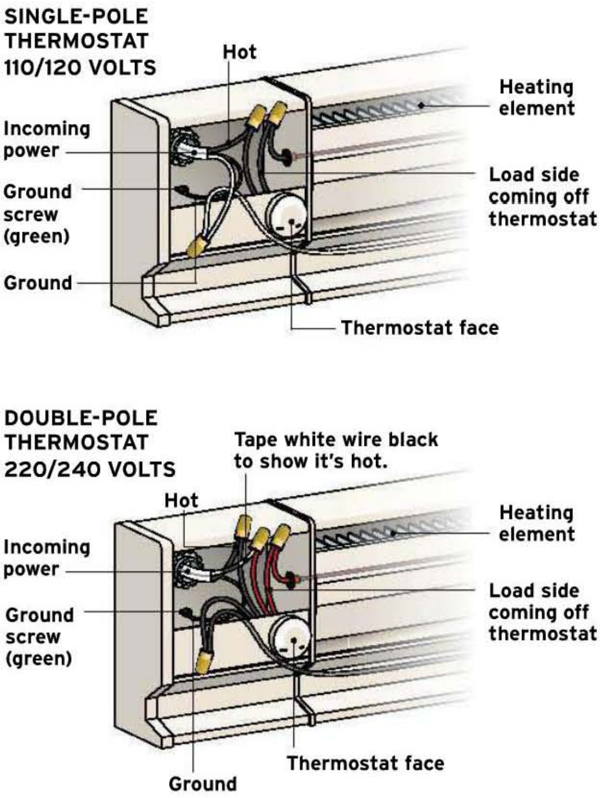
SINGLE-POLE THERMOSTAT 110/120 VOLTS
Hot
Heating element
Incoming power
Load side coming off thermostat
Ground screw (green)
Ground
Thermostat face
DOUBLE-POLE THERMOSTAT 220/240 VOLTS
Tape white wire black to show it's hot.
Hot
Heating element
Incoming power
Load side coming off thermostat
Ground screw (green)
Thermostat face
Ground

ELECTRIC WATER HEATERS

Replacing an electric water heater generally requires a permit, even if you attach the replacement to existing pipes. The reason is safety: Inspectors want to ensure that gas- and oil-fired units are properly vented and that electric heaters are correctly wired. Inspectors are particularly concerned that temperature and pressure relief (TPR) valves are correctly installed, because TPR valves keep water heaters from exploding in the event of a malfunction. For these reasons and for warranty issues, have a professional install your water heater. The drawing below is offered for information only.

TRADE SECRET

Electric tank heaters are becoming rare, given the greater cost-effectiveness of fuel-fired water heaters, especially tankless gas-fired water heaters.

WIRING A 240V WATER HEATER

Here's a look at a typical water heater installation. Three things to note: (1) The cutoff "switch" can be a circuit breaker or a fused switch rated for the load of the water heater, typically 30 amps. Place the cutoff switch close to the unit. (2) If you use two-wire cable to wire the water heater, tape the white wire with red or black tape on both ends to indicate that it is being used as a hot conductor. (3) Use flexible metallic cable (rather than rigid conduit) between the cutoff switch and the water heater for extra safety in earthquake regions.

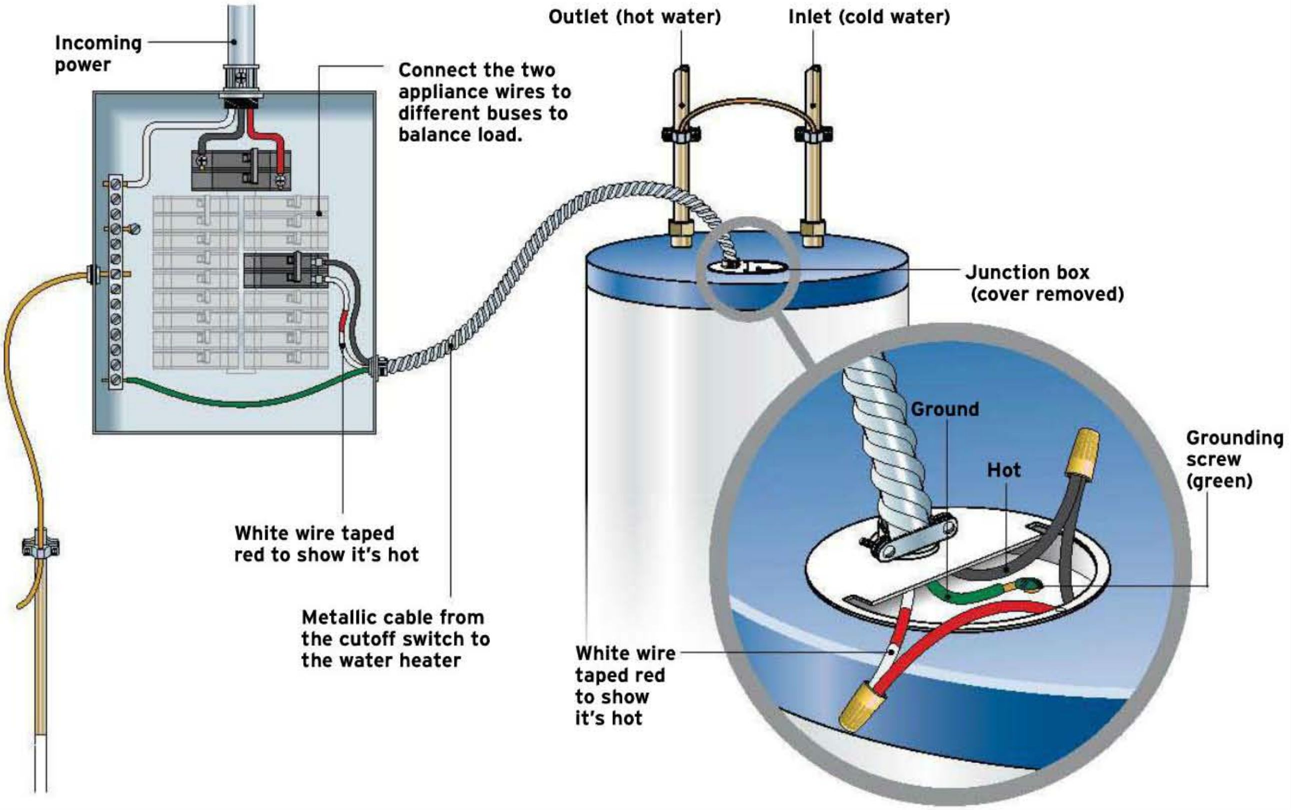
Incoming power
Connect the two appliance wires to different buses to balance load.
Outlet (hot water)
Inlet (cold water)
Junction box (cover removed)
Ground
Hot
Grounding screw (green)
White wire taped red to show it's hot
Metallic cable from the cutoff switch to the water heater
White wire taped red to show it's hot

OUTDOOR WIRING

ADDING AN OUTDOOR RECEPTACLE, a motion-sensor light, or low-voltage path lights can increase your security and safety and enable you to enjoy your property more fully. Because many outdoor devices are available as kits, they are usually easy to install and require few specialized tools. But before you start any project, review local building code requirements for outdoor installations. And remember: Before working on an existing outlet or fixture, be sure to turn off the power at the service panel or fuse box and test to be sure the power is off. Also, never operate power tools in damp or rainy conditions.

BEFORE YOU BEGIN
Planning your outdoor system, p. 224
Choosing outdoor lights, p. 225
Tools & materials, p. 226

OUTDOOR OUTLETS
Tapping into an existing outlet, p. 228
Mounting & wiring an exterior outlet, p. 230

OUTBUILDINGS
Running power to an outbuilding, p. 231

OUTDOOR PATH LIGHTS
Installing low-voltage path lights, p. 233

MOTION-SENSOR LIGHTS
Installing a motion-sensor light, p. 235

MRM Inc.
GREENLEE
438-20SS SILVER STEALTH

PLANNING YOUR OUTDOOR SYSTEM

Start by walking the property and noting where you'd like additional outlets, lights, and so on. Make a list. As you go, imagine activities that take place in different locations at night and day. Is there enough light along the path when you come home at night, enough outlets to entertain or to do chores on the weekend? At this stage, think big and anticipate future uses, especially if you must dig up the lawn to install your present project. With a little advance planning, you can avoid digging up the lawn a second time, later.

If your wish list is extensive, next make a scale drawing of the house and yard on graph paper. Note sidewalks, paths, and important landscape elements such as trees and large bushes. Note electrical devices that you'd like to add and existing ones that need upgrading. The drawing will be especially useful when it's time to calculate the number of fixtures and amount of cable you'll need. If you intend to take power from an existing outlet, note where the nearest outlet is, even if it's inside the house.

Getting power

Once you know roughly where you want to add exterior outlets or light fixtures, figure out how to get power to them. This will depend, in part, on how many devices you're adding.

If you're adding just one exterior receptacle, for example, find the interior receptacle closest to the one you want to add outside. If you position the new receptacle in the same stud bay as the interior receptacle, you'll simplify the task immensely because you won't have to drill through any studs.

Before cutting holes in anything, however, calculate the load on the circuit. Add up the wattage of all the lights and appliances presently in use and the wattage of the new outlet or light you want to add. If the total load on a 15-amp circuit exceeds 1,440w, run a new grounded circuit from the panel to the exterior device instead. Likewise, if you are running power to an outbuilding, run a new grounded circuit from the panel.

Install a 20-amp circuit if your plans call for electrical usage that exceeds the load limits of an existing circuit, includes a large number of receptacles, or contains three or more large flood lamps or other large lighting fixtures rated more than 300w each. If you will be using large, stationary tools in a workshop or heating the area with electricity, you may need to add several 120v and 240v circuits. Again, calculate the loads involved.

Two important points: First, the NEC allows you to tap into a general-use or lighting circuit only. You may not tap into circuits feeding kitchen-countertop receptacles or bathroom receptacles or into any dedicated circuits that supply power to air-conditioners, clothes washers or dryers, or other power-hungry appliances. Second, if you discover that the circuit you'd like to tap into is wired with nongrounded NM cable, BX armored cable, or knob-and-tube—don't tap into that circuit. The new extension may not be properly protected. Instead, run a new grounded circuit from the panel. In addition, don't forget weatherproof housings or covers and GFCI protection.

Finally, take your plans for new circuits or extensions to existing line-voltage circuits to the local building department and have them sign off before you start.

Outdoor safety: checking codes and utilities **Outdoor outlets and fixtures are exposed to weather; because moisture greatly increases the chance of electrical shocks, local codes are strict about what materials you can use and how they must be installed. In general, you don't need a permit to install low-voltage lights because the chance of shock is low; but if you want to add an outdoor receptacle or a light fixture that uses line voltage (standard house current of 110v or 220v), you'll need a permit.**

If you'll be running cable underground, check with local utilities before you dig. Call USA (Underground Service Alert) 800-227-2600. There may be water pipes, gas lines, telephone or cable lines, or electrical cables buried in the yard. Often, utilities will send out a technician to show you where such lines are located. If your lawn has a sprinkler system, note where sprinkler heads are and try to avoid the water pipes that feed them. Remember that only a licensed electrician should install hot tubs, swimming pools, and the like because such installations require special grounding methods.

CHOOSING OUTDOOR LIGHTS

There are a wide variety of light fixtures to choose from. For starters, choose line-voltage lights (120v) if you want to deter intruders, accent an architectural feature, or illuminate a work area such as an outdoor grill. To light up a walkway or add accent lights to the landscaping, however, low-voltage lights (12v or 24v) are usually a more economical choice and are generally easier to install. There are also lo-vo solar units (no wiring needed) that charge during day and glow softly all night.

In general, don't install more light than you need to serve the function for a given area. Outdoor lights that are too bright waste energy and will be too glaring for intimate dining or entertaining at night. Your neighbors will also thank you for not spot-lighting their house when they're trying to relax or sleep.

In addition to overhead lights, side-mounted lights, step-riser lights, in-ground fixtures, post-mounted lights, and stake-mounted lights, there are many switching options. You can control lights with standard on-off switches, timers, motion detectors, and photocells that turn lights on when the sun goes down. Most home centers and lighting stores have a dazzling array to choose from.

Outdoor lights are available in an array of styles. Here, Mission-style lights brighten a stone stairway.

ILLUMINATING OUTDOOR LIGHTS

As for your own safety, remember that intruders dislike being in the limelight. Install a security light, and they'll probably go elsewhere. A few tips to make your lights more effective:

- Install motion-sensor lights. Because they are dark until triggered by motion, motion-sensor lights startle intruders. Better-quality sensors can be calibrated so they are activated only when someone nears the house—not by every dog walker on the block.
- Put security lights high on a porch or under the eaves. Lights that can be reached without a ladder can be easily unscrewed. Most security lights have dual sockets so that if one bulb burns out, there's still one shining.
- Light walks and doors. It's easier for you to quickly approach and enter your house if sidewalks and entry doors are lit. Groping for your house keys in the dark is especially ill-advised if you live alone or along a dimly lit street.
- There are also security light kits that turn on lights inside the house if someone approaches or touches a window. And if you're leaving the house for an extended period, put lamps on timers to confuse would-be intruders.

Motion-sensor lights not only welcome you home but also help keep intruders away.

TOOLS & MATERIALS

A Use a square-nosed shovel to create shallow slots so you can tuck lo-vo cable into the ground.

B Rent a ditch-digging machine, also known as a trencher, to save time and avoid a sore back.

The tools you need to install outdoor wiring are pretty much the same ones needed to wire a house interior. The big exceptions, obviously, are digging and earth-moving tools. A *square-nose shovel* or *spade* is the most useful tool when you're making a shallow slot for low-voltage cable Ⓐ. If you're actually digging a trench, use a trenching shovel with a pointed nose and a reinforced shoulder that you can stamp on with a boot to drive it deep.

Digging trenches is hard work, however; so when the pros have to dig one of any length they rent a gas-powered ditch-digging machine, also known as a *trencher* Ⓑ. A trencher looks like a cross between a powered garden tiller and a chainsaw and typically cuts a trench 4 in. to 6 in. wide, and as deep as 24 in. Spread sheet plastic on either side of the trench so you can place the dirt from the trench nearby—which makes refilling the trench easier. Wear heavy boots, heavy gloves, safety glasses, and ear plugs when operating a trencher.

Materials

The NEC requires that all exterior outlets and circuits have ground-fault interrupter (GFCI) protection. This protection includes the following:

- A GFCI receptacle installed in a weatherproof box
- A standard receptacle installed in a weatherproof box protected by a GFCI receptacle installed *upstream* (toward the power source) or by a GFCI circuit breaker
- A circuit protected by a GFCI breaker.

Exterior light fixtures do not need GFCI protection. If the box is surface mounted, it needs to be raintight. If the box is recessed in the wall, it does not. The fixture must be listed for *damp locations* if under eaves and listed for *wet locations* if directly exposed to weather.

There are two common types of weatherproof covers (also called raintight covers) (see top photo on the facing page). A *weatherproof-while-in-use cover* has a plastic cover that shuts over an electrical cord, such as that used for holiday lights. A *weathertight* *cover* is a gasketed cover that shuts tight over the receptacles when not in use.

Aboveground, electrical cable must be housed in conduit with raintight fittings (see bottom photo at right). You can use polyvinyl chloride (PVC), thin-walled electrical metallic tubing (EMT) conduit, threaded intermediate metal conduit (IMC) or rigid steel conduit (RSC). PVC fittings are glued together to achieve a raintight fit; EMT conduit uses compression fittings; and IMC and RSC use threaded fittings. *Note:* EMT conduit fittings intended for interior use are not raintight and are *not* approved for exterior use.

Belowground, you can run flexible underground feeder (UF) cable at a depth acceptable to local codes—typically, 18 in. deep. Some codes allow you to dig a shallower trench if the cable runs in steel conduit but because threading steel conduit requires special equipment and advanced skills, it's not a reasonable option for most nonprofessionals.

WARNING

Avoid plugging or unplugging devices into exterior receptacles—or using corded power tools outside—when it's raining or snowing.

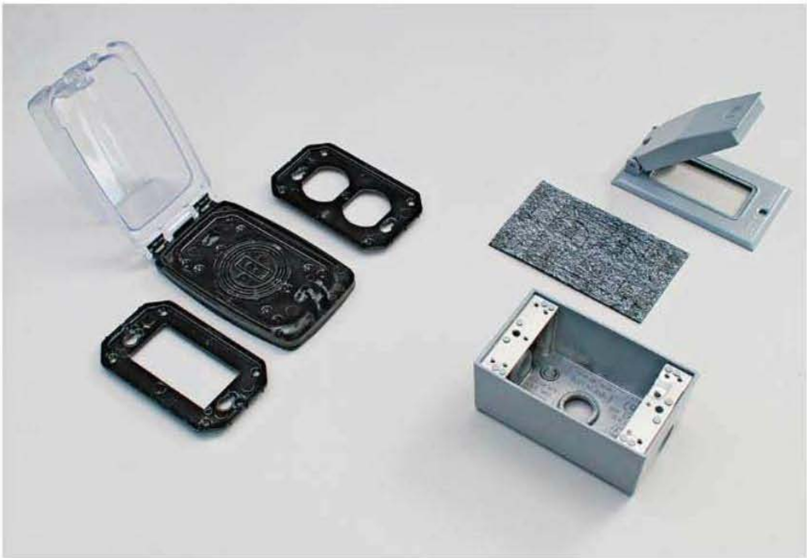

Raintight (weatherproof) covers include a weatherproof-in-use receptacle cover with adapter plates (left) and a weatherproof box with gasketed cover (right). Plastic adapter plates enable you to use the cover with a duplex or GFCI receptacle. The weatherproof box gasket is precut, so it accepts either a duplex or GFCI receptacle.

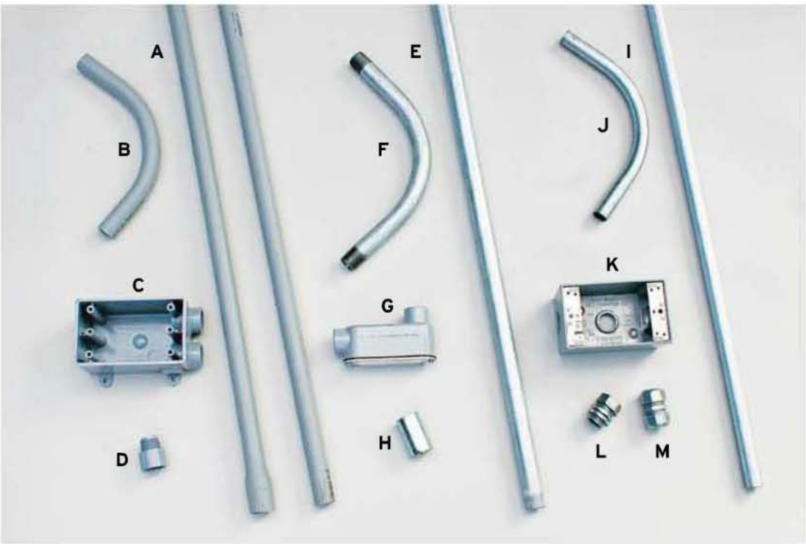

Overview of exterior conduit types, fittings, and boxes: A, ½-in. PVC pipe; B, PVC elbow; C, PVC box with unthreaded openings (for slip-in fittings); D, male adapter (MA) PVC fitting, which can be used with a threaded box or conduit; E, RSC pipe; F, RSC elbow; G, LB condulet; H, RSC coupling; I, EMT pipe; J, EMT elbow; K, bell box with three threaded holes; L, threaded EMT compression fitting; M, EMT compression coupling. Note: The LB conduit and bell box accept any ½-in. (trade size) threaded fitting—whether PVC adapter, EMT, or RSC.

TAPPING INTO AN EXISTING OUTLET

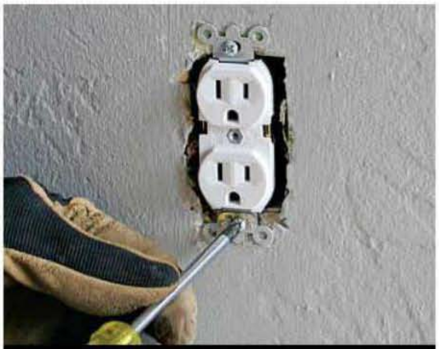

1 Remove the cover plate and unscrew the receptacle from the outlet box.

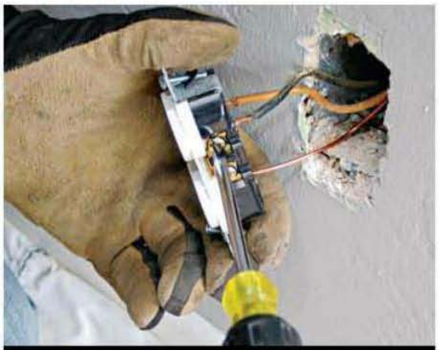

2 Pull the receptacle out of the box and remove the wires attached to its terminal.

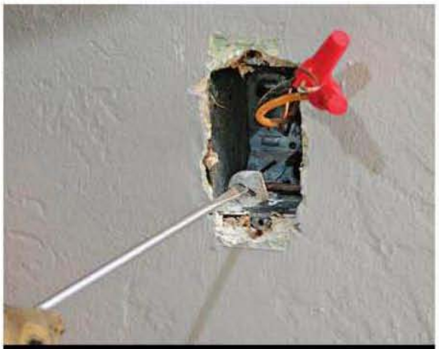

3 Use a screwdriver to pry out a knockout.

When adding a single outdoor receptacle, it's usually easiest to tap into an interior outlet within the same stud bay. In the installation shown here, the electrician solved the box-capacity problem by replacing the old receptacle with a commercial-grade duplex receptacle. Instead of splicing new and old wires and using pigtails—which would have required twist-on wire connectors and thus a larger box—he joined incoming and outgoing wires by inserting them into terminal holes in the back of the receptacle and then tightened down the terminal screws. *Note*: Wiring a commercial-grade receptacle in this manner is not the same as back-wiring a standard receptacle because standard receptacles have inferior spring clamps that are unreliable.

➔ **For information about back-wiring, see p. 42.**

Start by turning off the power to the receptacle at the panel or fuse box; then use an inductance tester to make sure the power is off. Remove the cover plate and unscrew the mounting screws securing the receptacle to the outlet box ❶.

Pull the receptacle out of the box, being careful not to touch the screw terminals; then test them again to make sure they're not energized ❷.

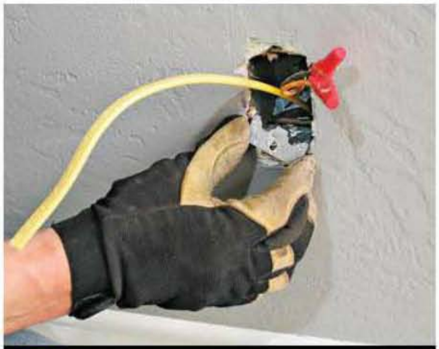

7 Pull the new cable through and attach a cable clamp.

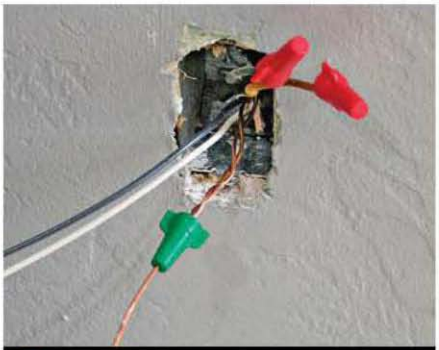

8 Leaving one leg long, use a green grounding nut.

Detach the wires from the terminals, fold them out of the way, and look for a knockout in the box to remove. Typically, a pair of knockouts will be hidden beneath an integral cable connector—unscrew it. Then remove a knockout from the bottom of the outlet box ❸. Drill through the exterior wall and into the same stud bay as the interior receptacle ❹. Then, inside the house, push fish tape through the knockout hole and into the stud bay ❺. Outside, a helper can catch the tape, pull it out of the hole you just drilled, and attach the new cable to it ❻.

Inside, pull the fish tape and the attached cable into the box. Pull about 1 ft. of new cable out of the box and then replace the integral cable clamp ❼ to secure the cable. Strip the cable sheathing and splice like wires together, starting with the ground wires. Use a special green grounding nut to splice the grounds ❽. Use wire strippers to remove 1/2 in. of the insulation from the ends of the hot and neutral wires ❾.

To save space in the old box, insert the stripped wire ends into terminal holes in the back of the commercial-grade receptacle ❿. Because commercial-grade receptacles solidly clamp wire ends, this connection is as solid as any splice. Fold the wires into the box as you push the receptacle into place until it is fully seated in the box—do not use mounting screws to pull the receptacle into the box because this could strip the screw threads. Then replace the cover and screw it into place.

➔ **See "Mounting and Wiring an Exterior Outlet," on p. 230.**

4 Outside, have a helper drill through the siding.

5 Push the fish tape through the knockout hole, into the stud bay.

6 Outside, retrieve the fish tape and attach the new cable.

9 Use wire strippers to remove $\frac{1}{2}$ in. of the insulation.

10 Gently push the wired receptacle into the box.

WARNING

To power holiday lights safely, plug them into a GFCI receptacle housed in a waterproof-while-in-use box cover. But before installing the lights, calculate the total wattage of all the bulbs so you don't overload the circuit.

ADDING AN OUTDOOR RECEPTACLE

If you tap into an existing receptacle at the end of a circuit, there should be enough room inside the box to bring a cable to feed the new outdoor receptacle. However, if there are already two cables in the box—incoming and outgoing—you may need to replace the existing box with a larger one.

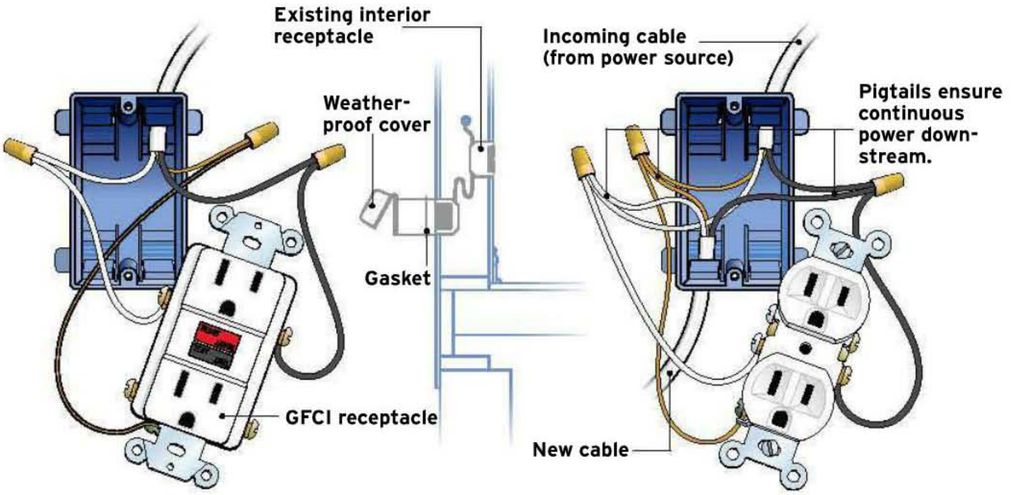
Existing interior receptacle
Weather-proof cover
Gasket
GFCI receptacle
Incoming cable (from power source)
Pigtails ensure continuous power down-stream.
New cable

MOUNTING & WIRING AN EXTERIOR OUTLET

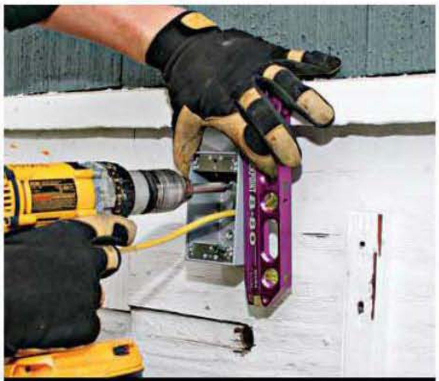

1 Feed the cable through the connector in the box and secure it to the building.

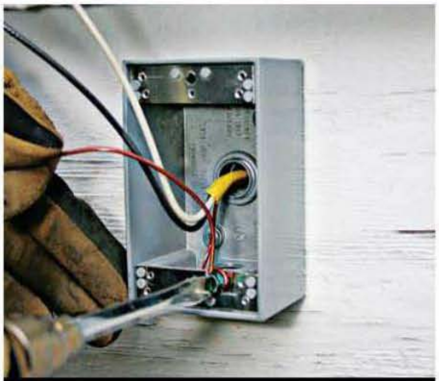

2 Loop the bare ground wire around the green grounding screw.

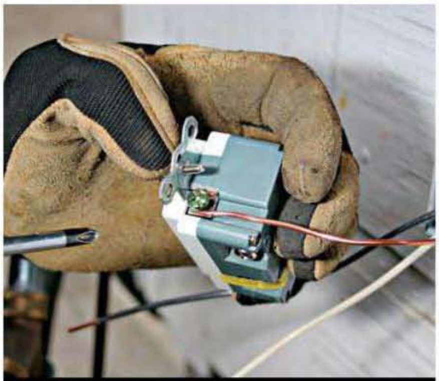

3 Attach the long ground wire to the green ground screw.

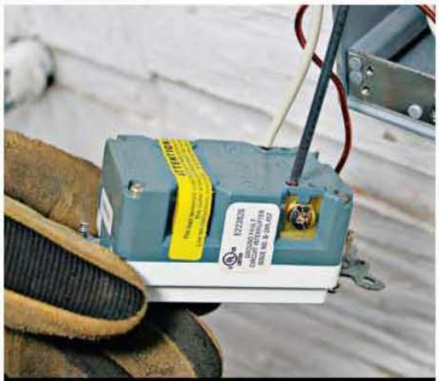

4 Insert wires into screw terminals. Gold screw for hot wire; silver screw, for neutral wire.

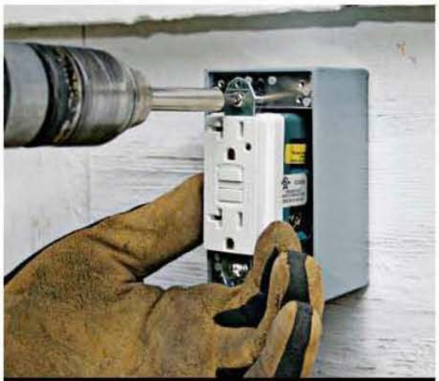

5 Fold and tuck the wires into the box. Then secure the receptacle to the box.

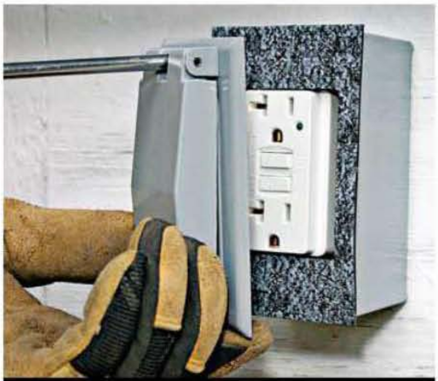

6 Fit the gasket around the receptacle and screw on the weatherproof cover.

Once the new cable is spliced to the cable in an existing receptacle, feed the cable through the cable connector in the back of the exterior box and mount the box. (The hole drilled in the exterior wall must be wide enough for the cable connector or the box won't sit flat to the wall.) To mount the box, use mounting ears or the small holes in the back. Apply siliconized caulking to the hole before attaching the box. Plumb the box and screw it to the outside of the building ❶.

Strip sheathing from the cable and ground the box by looping the bare ground wire around the green grounding screw. Leave this ground wire long ❷. Use a wire stripper to remove ½ in. of the insulation from the ends of the neutral and hot wires. Attach the ground wire that you earlier looped around the green ground screw to the GFCI receptacle ❸. Then connect the hot and neutral wires in their respective screw-terminal holes ❹.

If the GFCI receptacle has plaster ears, use diagonal cutters to remove them; otherwise, the receptacle may not fit into the box. Fold the wires and push the receptacle into the box; then screw down the mounting screws that secure the receptacle to the box ❺. A weatherproof gasket is used to keep water away from the wires—set it in place around the receptacle before you attach the cover ❻.

TRADE SECRET

GFCI receptacles are larger than standard duplex receptacles, so there will not be enough room in a single-gang box if you also need to splice an outgoing cable to feed another outdoor outlet, downstream. In that case, install a deep box or install an extension to the single-gang box.

RUNNING POWER TO AN OUTBUILDING

The first step to wiring an outbuilding is to figure out how many lights and outlets you need. If your needs are modest, you may be able to tap into an existing outlet in the main house and extend the circuit from it. Calculate the total loads for the existing circuit and the extension to see if the circuit has enough capacity. Otherwise, run a new circuit from the panel to the outbuilding.

If the outbuilding isn't more than 50 ft. from the house and has a few lights and outlets, it usually can be supplied by a 120v, 20-amp circuit and 12/2 w/grd UF cable. But check with local building authorities before you start. Get the necessary permit, Code requirements, and inspection schedules.

The hardest part of the job is usually digging the trench because local codes typically require it to be 18 in. deep. Fortunately, you can rent a gas-powered trencher to do the digging for you. After removing the dirt from the trench, pick out any rocks or debris that could damage the cable. Then lay the UF cable in the trench, flattening it as you go ❶.

>> >> >>

1 As you walk along the trench, unwind the cable. Ideally, it should lay flat in the bottom of the trench.

RUNNING POWER TO AN OUTBUILDING

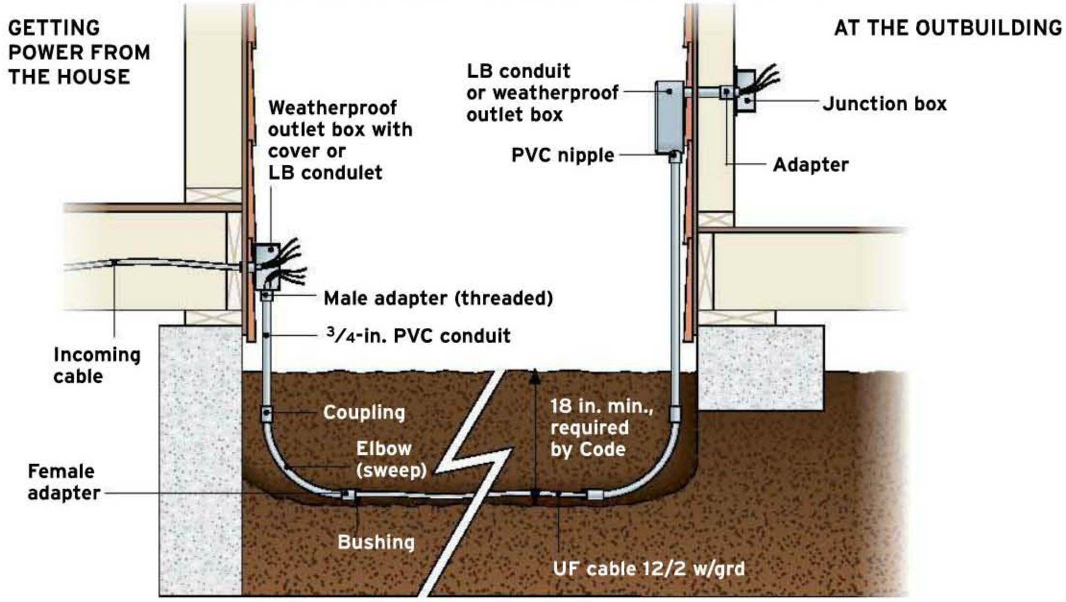
GETTING POWER FROM THE HOUSE
AT THE OUTBUILDING
Weatherproof outlet box with cover or LB condulet
LB conduit or weatherproof outlet box
Junction box
PVC nipple
Adapter
Male adapter (threaded)
3/4-in. PVC conduit
Incoming cable
Coupling
Elbow (sweep)
18 in. min., required by Code
Female adapter
Bushing
UF cable 12/2 w/grd

RUNNING POWER TO AN OUTBUILDING (CONTINUED)

Use electrical-grade PVC conduit, couplings, and adapters to protect the UF cable between the bottom of the trench and the outdoor boxes in which connections will be made. At the house, UF cable is typically spliced to an interior wire in a covered single-gang, weathertight box ❷. On the other end, the cable typically passes through an LB condulet and a short length of PVC pipe before terminating in a junction box inside the outbuilding ❸.

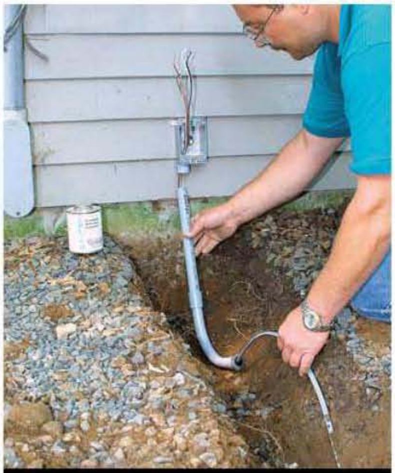

2 A PVC stub protects UF cable as it emerges from the ground.

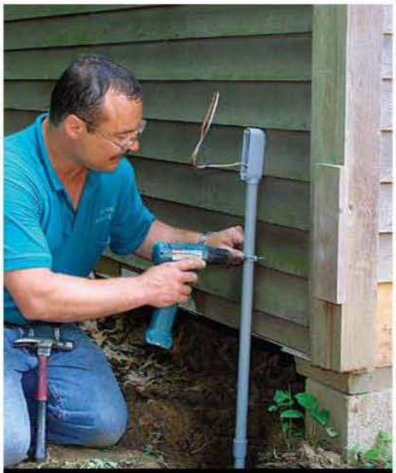

3 Strap rigid PVC pipe (with LB condulet) to the outbuilding.

WORKING WITH PVC CONDUIT

PVC conduit can be heated and bent, allowing you to change pipe direction with far less cutting and fewer fittings than you would need for rigid pipe. PVC conduit is intended to be used outside: It is flexible, durable, and waterproof; but its assemblies are not good-looking or as protective as metal pipe or cable. Be sure to use only the gray PVC conduit rated for electrical work (also called schedule-40 PVC conduit). White PVC plumbing pipe is unsuitable as electrical conduit; moreover, heating it can release noxious fumes. Wear work gloves to avoid burns when heating gray PVC pipe.

It's best to use a hacksaw with a metal-cutting blade to cut PVC pipe, though any fine-tooth saw will work. Make the cut as square as possible. Use a pocket knife or curved file to remove burrs from the pipe end, then wipe it with a rag. Apply an even coat of PVC cement to the outside of the pipe and the inside of the fitting. If you're bending pipe, screw the threaded adaptor into the outdoor box or conduit first, then glue the bent pipe to the adapter.

To bend PVC pipe, heat it with a MAP-gas torch; one common brand is the Bernz-Omatic™. You can also use a hot box or a PVC heating blanket (Greenlee). Keep the torch point moving constantly, rotate the pipe periodically, and tape the pipe ends to contain the heat and speed up the process. When the pipe droops, it's ready to shape. Place it against an irregular wall or foundation and it will conform to that contour. Strap the pipe close to the outdoor box and the foundation. To secure straps to the foundation, first predrill with a masonry bit. Then insert expandable plastic anchors into the holes and drive strap screws into the anchors.

To bend PVC pipe, heat it with a handheld MAP-gas torch. Rotate the pipe and keep the torch moving to heat the pipe evenly.

When the heated PVC pipe begins to droop, you know it's ready to bend. Shape it to fit the side of the building or foundation.

Use straps to secure the pipe to exterior walls and foundations.

INSTALLING LOW-VOLTAGE PATH LIGHTS

Low-voltage lights make nighttime paths and walkways safer, are easy to install, and pose almost no shock threat. Always follow the installation instructions that come with your low-voltage kit. Kits usually include light assembles, lo-vo cable, posts, ground stakes, and a timer-transformer power pack whose transformer reduces house current from 120v to 12v.

In the installation shown here, the power pack was mounted inside the garage, so PVC conduit was installed to protect the lo-vo cable as it traveled up the exterior wall into the garage. If you install the power pack outside, you probably won't need conduit. If you do install conduit, start by inserting a fish tape down the conduit ❶. Separate the two wires in the lo-vo cable, snip one, loop and tape the remaining wire to the fish tape, and pull the lo-vo cable through the conduit.

Place the lights where they'll best illuminate a walkway or highlight a landscape feature, then run the lo-vo cable to them ❷. Run cable along the ground and cover it with a few inches of mulch, or use a square-nosed shovel to create a shallow slot for the cable. Stomp on the shovel so it goes down 4 in., then rock the shovel from side to side to create a V-shaped slot ❸. Press the cable into the slot ❹; then stamp your feet to close the soil over the cable.

Each light fixture has two wire leads that terminate in sharp-pointed cable connec-

>> >> >>

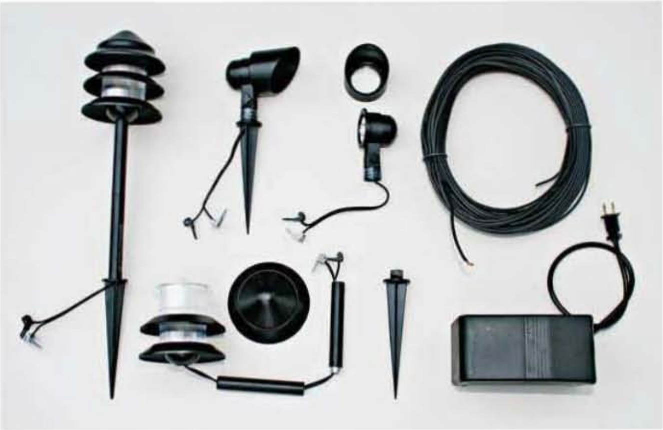

Lo-vo light kits typically contain screw-together parts: lamps, shades, riser posts, and ground stakes. At right: a coil of lo-vo cable and a power pack—a combination timer and transformer to step down voltage to 12V.

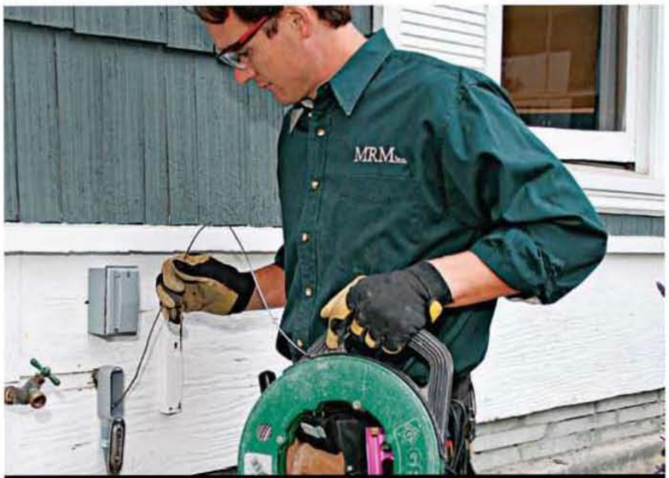

1 **After installing PVC conduit, insert a fish tape down the conduit and pull the wires through.**

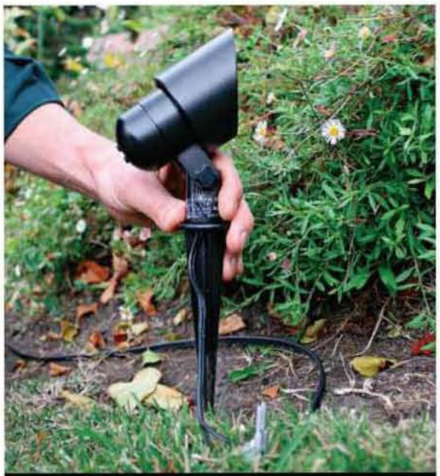

2 **Place the lights and run the conduit to each of them.**

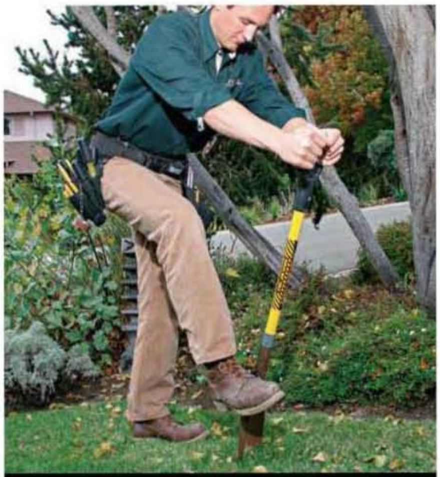

3 **Use a square-nosed shovel to create a small slot to bury the cable.**

4 **Press the lo-vo cable into the slot, making sure that it's fully covered.**

INSTALLING LOW-VOLTAGE PATH LIGHTS (CONTINUED)

tors. When snapped together, the connectors pierce the insulation of the lo-vo wires and supply each light with 12v current ❺. To power the system, attach lo-vo cable wires to the terminals on the power pack ❻. Mount the power pack securely to the wall and plug it into the outlet. Power packs are equipped with timers so that the lights come on and turn off whenever you like—whether you're home or away ❼.

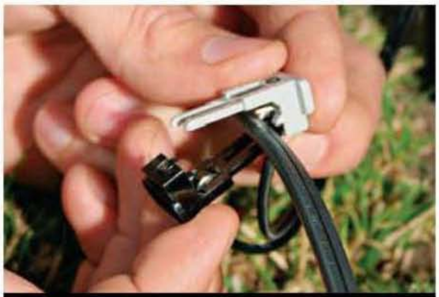

5 Connectors are used to join the cable in the ground to each light.

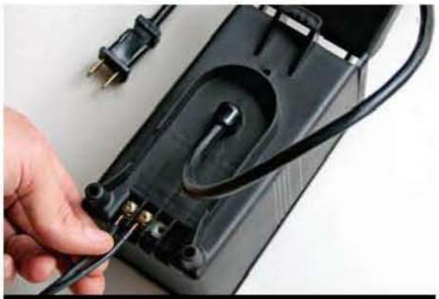

6 Attach the wires of the lo-vo cable to transformer terminals.

7 Once hung, plug the power pack into a nearby outlet and set the timer.

LOW-VOLTAGE LIGHT PARTS

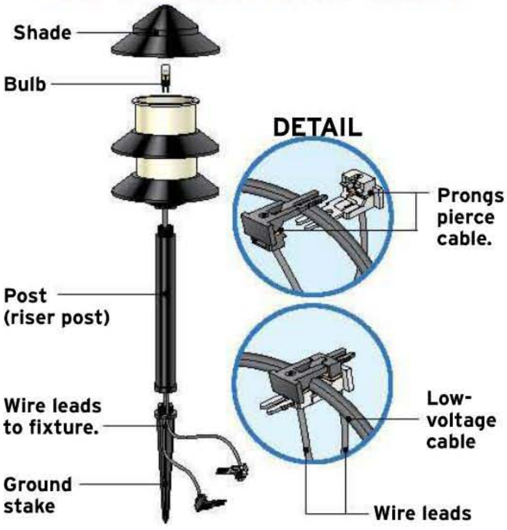
Shade
Bulb
DETAIL
Prongs pierce cable.
Post (riser post)
Low-voltage cable
Wire leads to fixture.
Ground stake
Wire leads

RUNNING LO-VO CABLE UNDER A SIDEWALK

The easiest way to run lo-vo cable under a sidewalk is to cut a piece of EMT pipe about 1 ft. longer than the sidewalk is wide. Hammer one end of the EMT shut so it creates a point, then use a hand sledge to drive the EMT under the sidewalk. To prevent closing the end of the pipe you hammer on, hold a scrap of 2x4 against the pipe end as you hit it. When the pointed end emerges from the other side of the walk, use a hacksaw to cut it off. Then feed the lo-vo cable through the EMT pipe. Be sure to wear safety goggles.

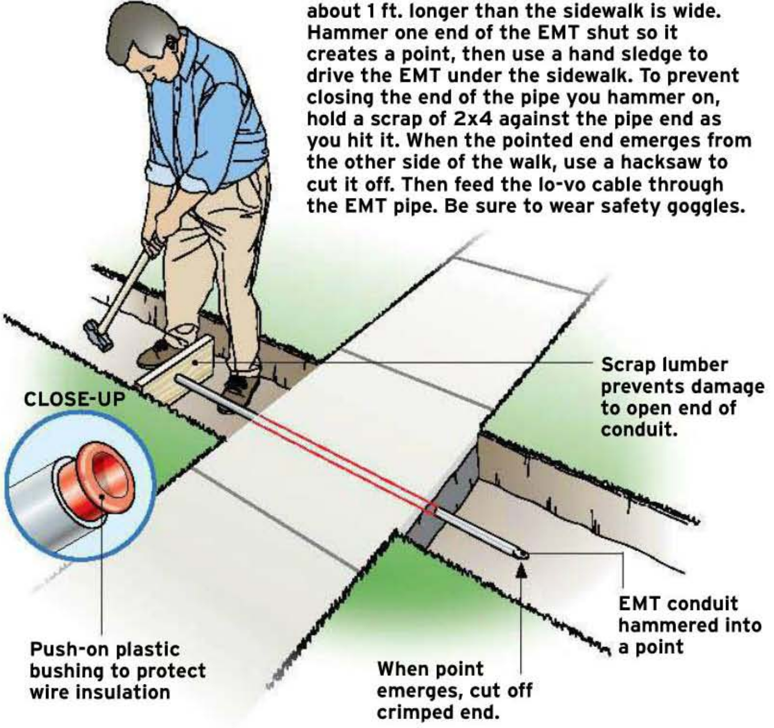
CLOSE-UP
Scrap lumber prevents damage to open end of conduit.
EMT conduit hammered into a point
Push-on plastic bushing to protect wire insulation
When point emerges, cut off crimped end.

INSTALLING A MOTION-SENSOR LIGHT

Motion-sensor lights require 120v, so they must be mounted on boxes rated for outdoor use. The fixture must be listed for damp locations if installed under eaves or must be listed for wet locations if directly exposed to weather. If you are replacing an existing light with a motion-sensor light, don't assume that the old box is raintight—examine it. Exterior boxes should have flexible gaskets between the box and the fixture base and threaded openings with closure plugs. If the box has only standard knockouts, it's not raintight.

Turn off power to the fixture and test to be sure it is off. Then unscrew the mounting screws securing it to the box. Holding the fixture in one hand, touch an inductance tester to the wire splices to be sure they are not energized. Disconnect the wire splices; if the existing box is damaged or inappropriate, remove it. The new box must be raintight ❶.

Feed the incoming wires through the cable connector in the back of the box and then attach the box to the wall ❷. Mounting screws should sink into sheathing or into a block screwed to the sheathing. Attach the mounting bar to the box ❸ and attach a ground wire to the ground screw on the box or mounting bar. Splice the ground wires from the box and the incoming cable ❹.

Fit a weatherproof gasket over the cable wires; then splice them to fixture leads: hot to hot, neutral to neutral ❺. Tuck the wires into the box and attach the fixture ❻. Before screwing the bulbs into the sockets, slide waterproof washers onto them ❼. Adjust the motion sensor so that it is triggered by people approaching the house—not by passing cars, dog walkers, and the like.

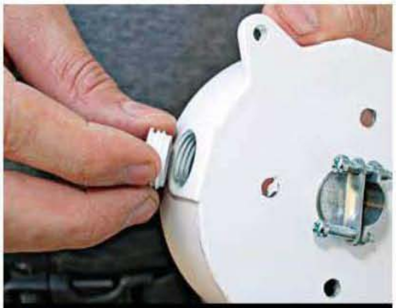

1 A Code-approved outdoor box must be used.

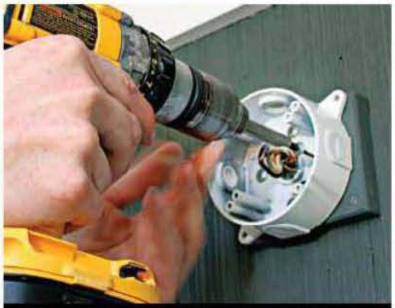

2 Feed incoming wires through the back of the box; then attach the box.

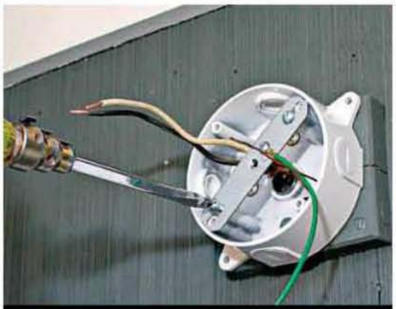

3 The green ground screw ensures a continuous ground path to the box.

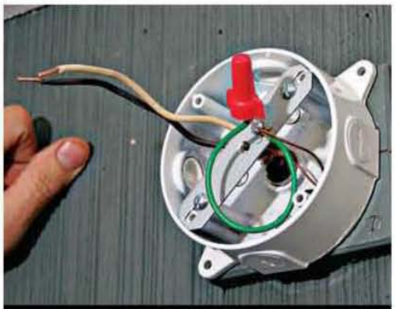

4 Splice all ground wires.

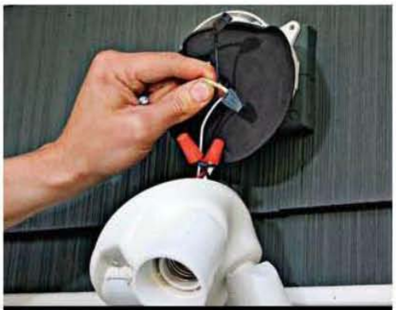

5 Slide a weatherproof gasket over the incoming cable wires.

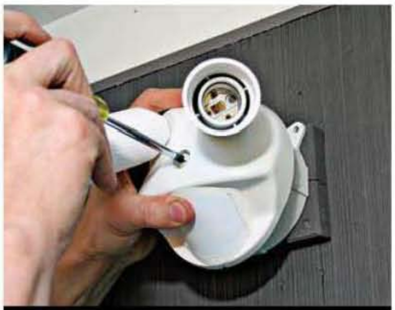

6 Fold all of the wires back into the box; attach the fixture.

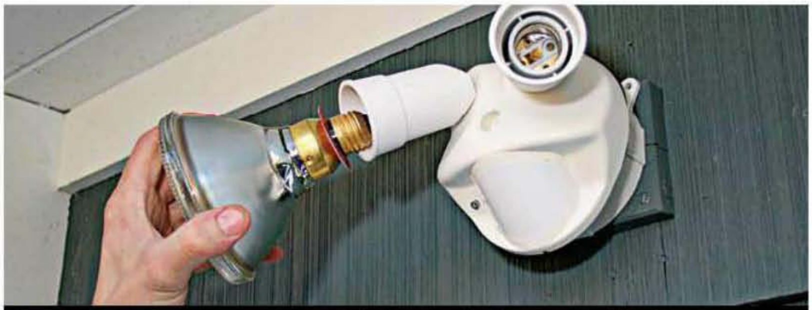

7 Waterproof washers around each bulb prevent water from getting into the sockets.

PANELS & SUBPANELS

THIS CHAPTER OFFERS AN overview for those who want to add a circuit to an existing panel, learn about installing subpanels, or understand the major elements of a service upgrade. Adding a circuit is not difficult, but we recommend having a licensed electrician do all work inside a panel. In particular, only a professional should upgrade electrical service. If you want more information about panel work or service upgrades, hire a pro and watch at a safe distance.

There are several reasons for this caution: First, there's no obvious way for a nonprofessional to be sure that an existing system is correctly grounded and free from shock dangers. Second, some panels are so old, poorly wired, or overloaded that simply removing or replacing a cover can be dangerous. Third, local building codes may require that major upgrades be done only by licensed electricians. Last, an inexpertly wired electrical system may nullify appliance or equipment warranties or even void your homeowner's insurance should an electrical fire occur.

BEFORE YOU BEGIN

Understanding service panels & subpanels, p. 238
Sizing panels, subpanels, & conductors, p. 240

CIRCUIT BREAKERS

Adding a circuit breaker, p. 242

SUBPANELS

Installing a subpanel, p. 244

UNDERSTANDING SERVICE PANELS & SUBPANELS

Electrical panels vary in size and configuration, contributing to the general confusion over how to wire them correctly. So let's first look at how utilities supply electricity (service entrances) and then consider the difference between *service panels* (also called main panels or main service panels) and *subpanels*.

For decades, overhead service entrances were the only type, with large service conductors running from a utility pole to a *service head* (or weather head) mounted on or near the roof. But large wires are unsightly, so buried service entrances have become more and more popular.

Typically, three service conductors run to a meter base. The two hot conductors attach to lugs on the supply side of the meter, and the neutral conductor attaches to a neutral bus bonded to the metal box. When a meter is inserted in the base, power flows from the house-side meter lugs to a main disconnect (main breaker) in a service panel. Important point: All service panels contain a main breaker. Subpanels typically do not, unless they are housed in a separate building from the main breaker.

In most setups the main breaker attaches directly to two hot buses. Turn off the main breaker and you disconnect power to all breakers and circuits energized by those buses and to any subpanels *downstream*—away from the power source. Nonetheless, always test to be sure the power is off after flipping off the main breaker.

To be safe, an electrical system must be grounded. At the transformer, a ground wire runs from the neutral conductor to a ground rod driven into the earth. At the service panel, a ground wire attaches to a ground buss,

WARNING

If you are at all uncertain whether a panel is energized, do not remove its cover. Call a licensed electrician. Do not take the risk!

WORKING SAFELY IN A SERVICE PANEL

Cut power inside the service panel by flipping the main breaker off or by removing the main fuse in a fuse box. This de-energizes the hot buses. But remember that large feeder wires may still be hot on the incoming side of the main breaker—avoid touching that area. Unscrew the panel cover and set it aside.

Carefully test feeders and breaker terminals for voltage. If voltage on any exposed part is still present, carefully replace the cover and call a licensed electrician. If no voltage is present you can proceed. If you have any doubt or uncertainty, call a licensed electrician. Do not take the risk!

Always turn off the power before working on an electrical system.

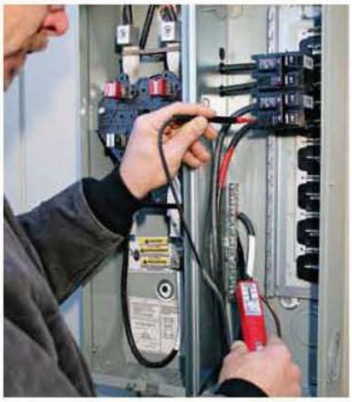

After removing the cover, touch one tester prong to the bare end of a feed wire and the other prong to the neutral bus.

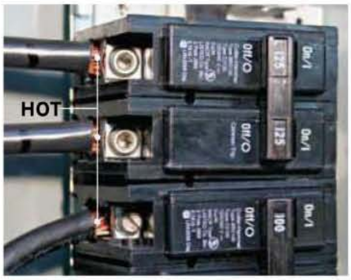

Avoid touching the feeder wires that attach to the main breaker—they stay hot even when the main breaker is turned off.

LOCK 'EM OUT!

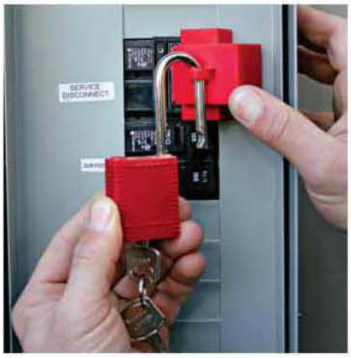

Once you've shut off power in a service panel, tape the panel shut and post a warning sign of work in progress as good first steps. But it's still possible for someone to remove the tape and reenergize the system. Pros prefer to use a *breaker lockout*, which limits panel access to the person holding the key. You can buy panel lockouts at electrical suppliers and most home centers.

exits through the bottom of the panel, and clamps to a ground rod. If there are separate neutral and ground buses in the service panel, both will be bonded to the metal body of the panel. In some service panels, both neutral and ground conductors attach to a common neutral/ground bus.

This is a second important distinction: Neutral and ground conductors are bonded at the service panel, but *never* in a subpanel. That is, in a subpanel there will always be a separate neutral bus and a ground bus. Typically, the ground bus bar in a subpanel will be bonded to the metal body of the panel, whereas the neutral bus bar in a subpanel will be mounted on nonconductive brackets.

Again, downstream from the service panel, grounds and neutrals are always separate. Thus, although three service conductors feed a service panel, four conductors feed a subpanel: two hot wires, one neutral, and one ground.

SERVICE ENTRANCE TO THE EAVES SIDE

When the service drop approaches over eaves, the service riser is sent up through the eaves and the point at which the riser emerges from the roof jack is caulked.

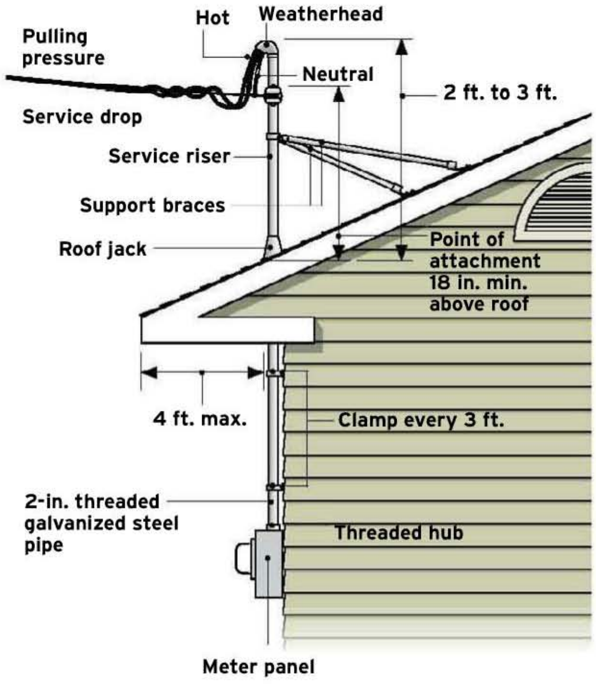
Hot
Weatherhead
Pulling pressure
Neutral
2 ft. to 3 ft.
Service drop
Service riser
Support braces
Roof jack
Point of attachment 18 in. min. above roof
4 ft. max.
Clamp every 3 ft.
2-in. threaded galvanized steel pipe
Threaded hub
Meter panel

SERVICE ENTRANCES

Though most older homes have overhead service, buried service is becoming increasingly common. The utility's responsibility ends where the service-drop cables are spliced to the service conductors running to the meter. In some areas, Code requires that underground conductors be housed in conduit.

AERIAL SERVICE ENTRANCE

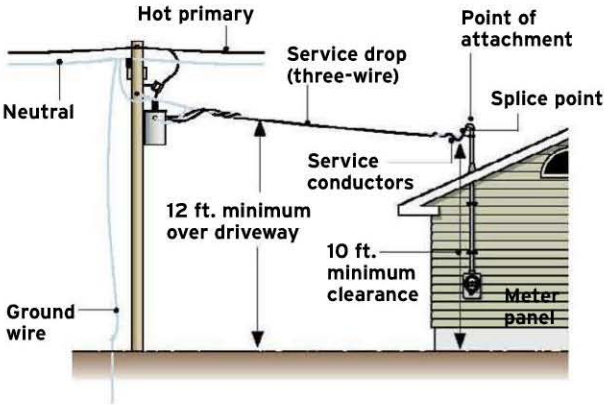
Hot primary
Point of attachment
Service drop (three-wire)
Splice point
Neutral
Service conductors
12 ft. minimum over driveway
10 ft. minimum clearance
Meter panel
Ground wire

BURIED SERVICE ENTRANCE

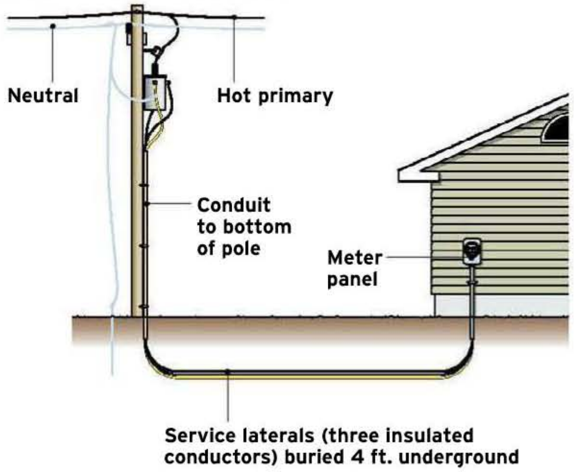
Neutral
Hot primary
Conduit to bottom of pole
Meter panel
Service laterals (three insulated conductors) buried 4 ft. underground

SIZING PANELS, SUBPANELS & CONDUCTORS

A METER-MAIN COMBO PANEL

Inside a meter-main combination panel, before the meter has been installed.

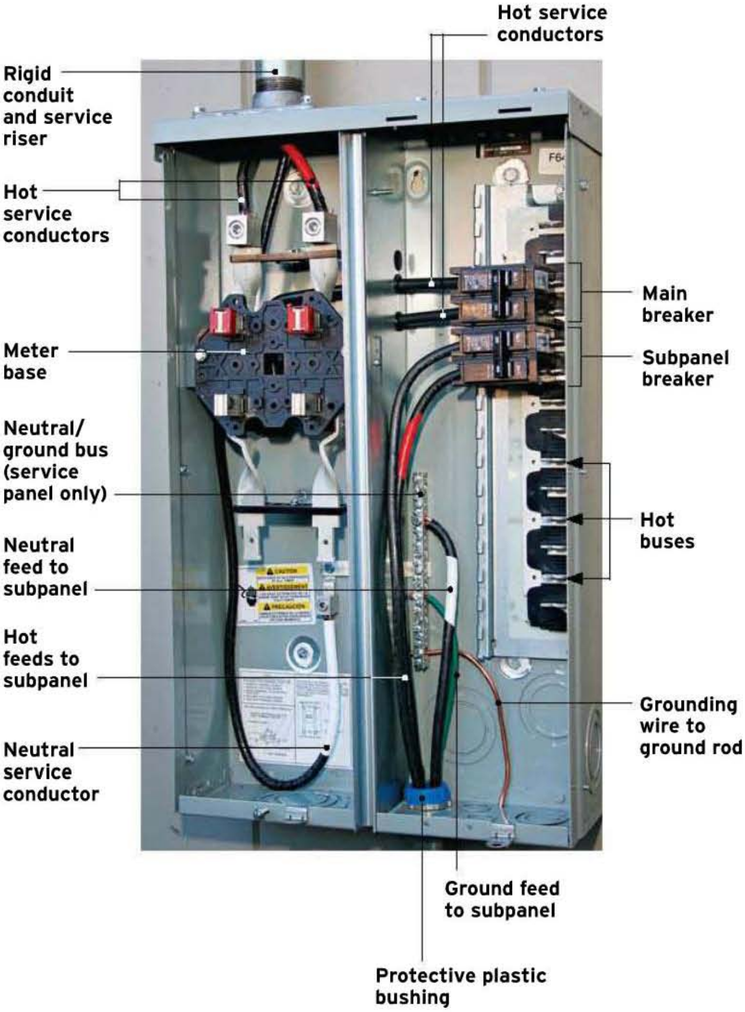
Hot service conductors
Rigid conduit and service riser
Hot service conductors
Main breaker
Meter base
Subpanel breaker
Neutral/ ground bus (service panel only)
Hot buses
Neutral feed to subpanel
Hot feeds to subpanel
Grounding wire to ground rod
Neutral service conductor
Ground feed to subpanel
Protective plastic bushing

A SUBPANEL WITH FEEDS TERMINATED AT LUGS

If there's no room left in the main panel, adding a subpanel is a good option.

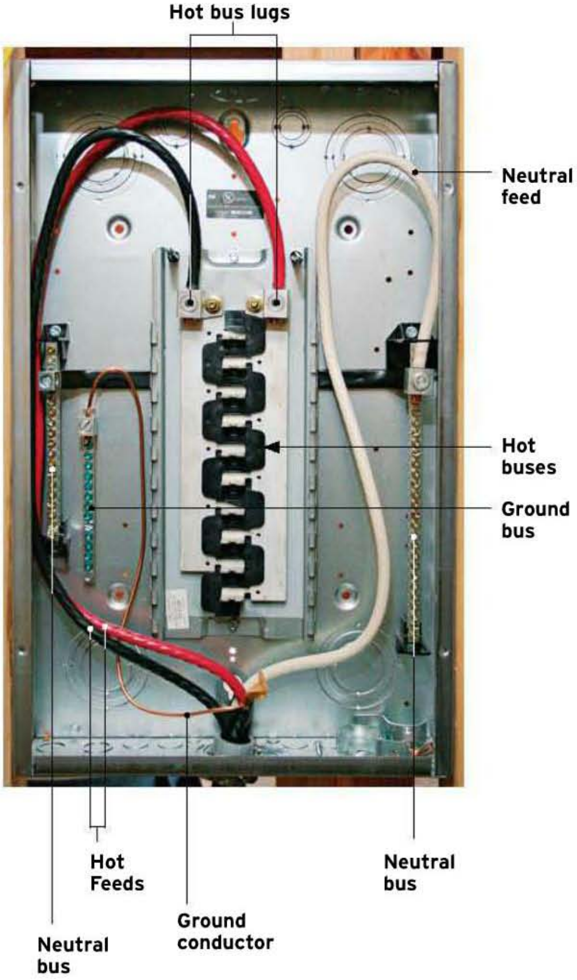
Hot bus lugs
Neutral feed
Hot buses
Ground bus
Hot Feeds
Neutral bus
Ground conductor
Neutral bus

To size the service panel, many electricians calculate household needs (ampacity), then add 20 percent to 25 percent for future needs.

For more on calculating household needs, see p. 164.

Because lifestyles and energy consumption vary widely, it's tough to generalize about service panel size. For a family in a 2,000-sq.-ft. house, a 125-amp service panel is probably adequate. But many electricians recommend a 200-amp panel for an "average" household in a home less than 3,500 sq. ft. that is not heated electrically.

The NEC allows a maximum of 42 breaker spaces in a single panel. If your system still has capacity but you've used up the available panel spaces, add a subpanel. If your needs exceed system ampacity, upgrade the service.

All conductors—whether they feed a service panel or energize wall outlets—must be sized according to the loads they carry. If you'd like more information on sizing conductors you can either get a copy of the most recent *National Electrical Code* or a copy of *Code Check: Electrical* (Taunton Press), which does a good job of summarizing the NEC tables.

To give one example of service conductor sizing: A home with a 200-amp service panel should be powered by three 2/0 copper THHN/THWN conductors. As noted earlier, two of the conductors are hot and the third is neutral (taped white). Each hot conductor carries 120v relative to the neutral, so together the two hot conductors are capable of delivering 240v relative to each other. Remember, only licensed electricians and utility company workers should upgrade a service entrance.

Key Terms

Conductors

Technically, anything that conducts electricity is a conductor, but the term most often denotes individual wires, regardless of size. Thus the large wires that run from a utility pole are service conductors. The fat wires that run down a service riser to a meter base are conductors, as are the slim circuit wires that energize outlets and light fixtures.

Wire

***Wire* is a generic term that refers to an individual conductor, and it is most often used when referring to the size or type of the conductor, for example 12AWG and 2/0 THHN/THWN stranded wires.**

Cable

Cable is an assembly of several conductors, usually in a plastic or metal sheathing. Hence, Romex cable, metal-clad (MC) cable, SE cable, and so on. SE cable, a frequently misused term, is a specific type of large, nonmetallic cable that is *not* an acceptable service conductor in all regions.

Feeds

Feeds, or feeders, are distinguished by function more than form, though typically they are larger conductors that supply power to an element downstream. Thus feeders might run from a meter base to a main breaker or from a service panel to a subpanel.

ADDING A CIRCUIT BREAKER

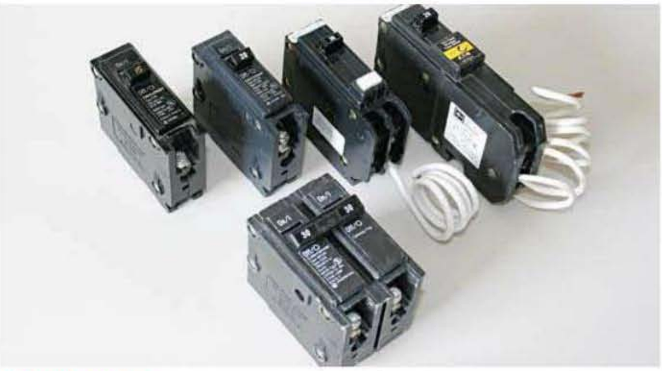

Various breakers. From upper left, clockwise: single pole 15A, single pole 20A, single pole 20A GFCI, single pole 20A AFCI, double pole 30A (for 220 V loads).

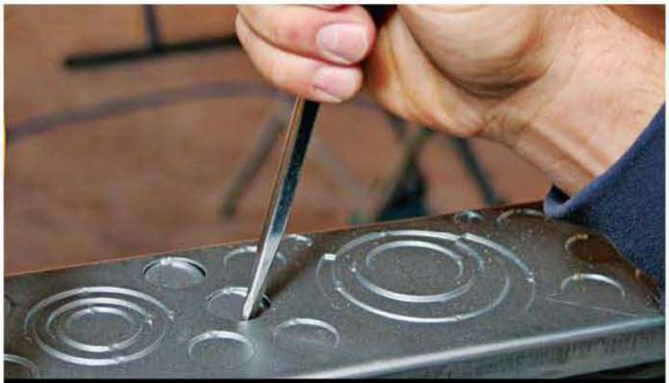

1 Remove a knockout from the existing panel.

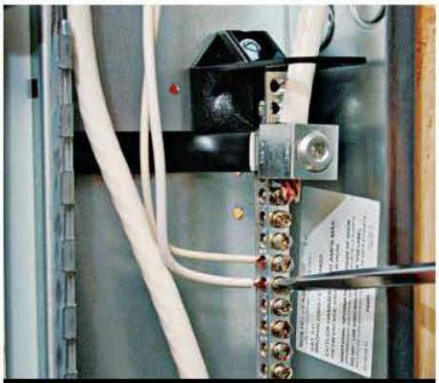

5 Connect the neutral wire to the neutral bus.

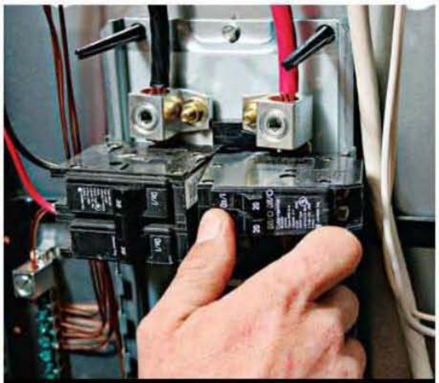

6 To install a breaker, place its heel on the mounting rail and snap it in.

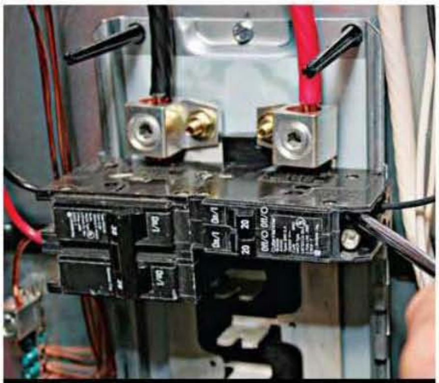